Journal of Information Systems Education

Volume 37
Issue 3
Summer 2026

Teaching Tip

A Data Preparation and Enrichment Exercise for Location-Aware Analysis of University Innovation

Andrés Díaz López, Saisrinivas Mamunuru, and Joseph Cazier

Recommended Citation: Díaz López, A., Mamunuru, S., & Cazier, J. (2026). Teaching Tip: A Data Preparation and Enrichment Exercise for Location-Aware Analysis of University Innovation. *Journal of Information Systems Education*, 37(3), 360-396. https://doi.org/10.62273/AWXK7246

Article Link: https://jise.org/Volume37/n3/JISE2026v37n3pp360-396.html

Received:	April 1, 2025
First Decision:	June 29, 2025
Accepted:	May 22, 2026
Published:	September 15, 2026

ISSN: 2574-3872 (Online) 1055-3096 (Print)

Journal of Information Systems Education, 37(3), 360-396, Summer 2026
https://doi.org/10.62273/AWXK7246

Teaching Tip
A Data Preparation and Enrichment Exercise for Location-Aware Analysis of University Innovation

Andrés Díaz López
W. P. Carey School of Business
Arizona State University
Tempe, AZ 85281, USA
adiazlop@asu.edu

Saisrinivas Mamunuru
Ira A. Fulton School of Engineering
Arizona State University
Tempe, AZ 85281, USA
smamunur@asu.edu

Joseph Cazier
W. P. Carey School of Business
Arizona State University
Tempe, AZ 85281, USA
joseph.cazier@asu.edu

ABSTRACT

Prior research highlights the growing importance of location-based data for business decision-making, while also documenting its limited integration into information systems curricula. Although students are routinely exposed to data analytics techniques, they rarely gain hands-on experience preparing and enriching datasets with location-based data. This teaching tip addresses that gap with a modular, instructor-ready exercise that introduces geospatial analytics through the data preparation process. Using a publicly available university innovation dataset derived from the Association of University Technology Managers (AUTM) survey data, enriched with geospatial data, the exercise guides students through a structured workflow that includes data cleaning, identification of location attributes, dataset enrichment using external sources, and transformation from panel to cross-sectional formats. The emphasis is on practical data-preparation decisions—such as handling missing values, resolving inconsistencies across datasets, selecting appropriate units of analysis, and integrating location-based variables—rather than on specialized geospatial functionality. Designed for use in information systems and analytics courses, the exercise supports core IS competencies in data management, analytics, and decision support while remaining adaptable across commonly used tools. By focusing on geospatial data preparation as an entry point, the teaching tip lowers the adoption barrier for instructors and provides students with foundational skills to incorporate location into business analysis.

Keywords: Geospatial analytics, Location analytics, Data cleansing, Geographic information system (GIS), Generative AI in education, University innovation

1. INTRODUCTION

According to a U.S. Department of Commerce report, innovation—defined as the process of creating and implementing unique products, services, or practices that generate value—accounts for between one-third and one-half of economic growth (North Carolina Department of Commerce, 2016). Universities, colleges, and research institutions are among the most important sources of innovation in the United States. These institutions not only develop new products and processes that drive societal progress but also cultivate human capital, which plays a crucial role in fostering innovation within both public and private organizations.

We use data on multiple indicators of university innovation, including filed and issued patents, startups formed and closed, license revenue, research expenditures, and institutional characteristics, from more than 200 universities. These data are collected annually by the Association of University Technology Managers (AUTM) and are widely used to support institutional benchmarking and strategic decision-making related to technology transfer and innovation outcomes.

In its original form, the AUTM dataset supports conventional descriptive and trend-based analyses commonly taught in analytics courses. However, its analytical potential is substantially expanded when enriched with location-based and geospatial information. In this teaching exercise, the original innovation data are augmented with additional attributes related to the communities in which universities are located, including city identifiers, demographic characteristics, and other publicly available location-based measures. This enrichment enables analyses that consider how outcomes vary across space and context, rather than treating location as a purely categorical attribute.

Prior research suggests that location and proximity influence social interaction and innovation processes. For example, Wineman et al. (2009) argue that space structures patterns of circulation, copresence, coawareness, and encounters within organizations, which are fundamental to the development of social networks critical to innovation. Similarly, McPherson et al. (2001) note that perhaps "the most basic source of homophily is space: we are more likely to have contact with those who are closer to us in geographic location than those who are distant," a principle Logan (2012) characterizes as reasoning about where things are in relation to one another. Whittington et al. (2009) further show that innovation, measured by patent output, is associated with physical proximity and location within key geographic regions in the biotechnology industry (as cited in Logan, 2012).

Despite the richness of the AUTM data, it has rarely been examined from a location analytics perspective. Except for Malik et al. (2023), few studies have attempted to integrate geospatial attributes into analyses of university innovation outcomes. This gap presents an opportunity to demonstrate how location-based data can be used to enrich an existing dataset and expand the range of analytical questions that can be addressed, particularly in instructional settings.

Location analytics encompasses a broad set of methods for incorporating spatial and location-based information into analysis, and it is not limited to specialized geospatial datasets or tools. Much of the data used in this exercise—such as demographic, community, and institutional information—is publicly available and commonly encountered in business analytics contexts. While GIS platforms can support advanced spatial analysis and visualization, meaningful geospatial insights can also be generated using general-purpose analytics tools by integrating location attributes into standard data preparation and analysis workflows.

Many contemporary business models—including those of firms such as Airbnb, Uber, and DoorDash—depend heavily on location-based information, while established organizations such as Walmart, McDonald's, and Maersk rely on spatial data to support operations and strategy. Nevertheless, prior research indicates that business schools have been slow to integrate location analytics into their

curricula. King and Arnette (2011) report that, at the time of their study, only one in thirty business schools offered a course related to geographic information systems or location intelligence.

More recently, Erskine et al. (2024) document the limited presence of location analytics in information systems curricula and argue that this gap constrains students' ability to address real-world, data-driven problems. They conclude that many organizations fail to fully capitalize on spatial data to enhance decision-making and gain competitive advantage. Similarly, Ritchie et al. (2023) recommend that business schools incorporate location analytics not only in supply chain courses but also across analytics, finance, human resources, and cybersecurity. Despite these recommendations, few teaching materials provide step-by-step guidance for enriching non-spatial datasets with location-based information in a form that is accessible to IS instructors.

The purpose of this teaching tip is to provide a practical, instructor-ready exercise for information systems and analytics courses that introduces geospatial analytics through the data preparation process. Rather than focusing on theoretical models or advanced geospatial functionality, the exercise guides students through the steps required to clean, enrich, and restructure an existing dataset so that location-based attributes can be incorporated into analysis. In doing so, it addresses a common gap in analytics instruction, where datasets are often presented as already prepared and ready for analysis (Ritchie et al., 2023).

Although specific tools and platforms are referenced, the exercise is designed to remain software-agnostic. The tools selected were chosen based on their accessibility to instructors and students, and alternative tools can be substituted as appropriate. The emphasis throughout is on practical data preparation decisions—such as handling missing values, resolving inconsistencies across datasets, selecting appropriate units of analysis, and integrating location-based variables—rather than on mastery of any single technology. It is recommended that students be familiar with basic data types, which can be covered in an introductory data analytics or statistics course, or prior to starting the exercise.

In the following sections, we describe the data used in the exercise (Section 2), outline the methodological structure of the teaching case (Section 3), detail the sequence of modules and associated instructor guidance (Section 4), report the implementation and feedback (Section 5), discuss limitations (Section 6), offer suggestions for future extensions (Section 7), and conclude with implications for information systems education (Section 8).

2. THE DATA

Our experience in data analytics has shown that, despite accounting for a substantial proportion of effort in real-world analysis, data preparation is often underemphasized in classroom settings. Instruction frequently focuses on teaching a specific technique or software application using datasets that are already cleaned and structured for immediate use. This approach rarely reflects practice, in which data are often incomplete, inconsistent, outdated, or improperly structured due to granularity issues, missing values, or reporting inconsistencies.

Location-based analysis is particularly effective at revealing such issues, especially those related to granularity and unit of analysis. Even when data initially appear suitable, careful analytical inspection often raises concerns about validity and comparability, suggesting the need to enrich the dataset, integrate additional sources, or restructure the data to better address the analytical task. This exercise illustrates how incorporating location-based attributes during data preparation influences these decisions and expands the range of analytical questions that can be explored.

The data used in this exercise are based on a publicly available, open-source dataset published by Malik et al. (2023). That dataset draws on university innovation data collected through the Association of University Technology Managers (AUTM) annual survey for the years 2010 to 2015 and was enriched with additional university-, community-, and location-based attributes. More recent versions of the data are not publicly available. However, the dataset used in this teaching tip is well suited to its purpose, which is not to evaluate the data itself but to demonstrate what can be accomplished with a dataset that exhibits these characteristics.

The dataset enrichments include city-level identifiers, demographic and economic characteristics, measures of community attractiveness and accessibility, and other geospatially relevant variables derived from public sources. For use in this teaching exercise, the dataset was further updated, cleaned, modified, and modestly enriched by the authors to improve internal consistency and suitability for instructional analysis. This structure allows students to work with realistic innovation data while observing how integrating location-based information extends analytic possibilities beyond those available from the original survey data alone.

As with many longitudinal and self-reported datasets, the data exhibit limitations commonly encountered in practice. Collected over multiple years, the dataset reflects changes in reporting practices and data definitions as the survey evolved. Certain variables available in later years were not included in earlier collections, and not all institutions report data consistently from year to year. In addition, institutions may report results at different organizational levels, such as a university system, a main campus, or a research foundation managing multiple campuses. These characteristics introduce common data preparation challenges related to the unit of analysis, record matching, and comparability, making the dataset particularly well-suited for instructional use.

The Malik et al. (2023) dataset further incorporates supplementary data describing community demographics, economic conditions, startup support, and other location-related characteristics relevant to innovation outcomes. In its raw form, the dataset exhibits many of the same challenges encountered in real analytics projects, requiring attention to structure, completeness, and suitability for analysis. At the same time, these characteristics provide a strong foundation for a teaching exercise that demonstrates how analysts can enrich existing datasets with geospatial information to improve analytic insight.

A data dictionary for the final version of the dataset is provided in Appendix A. An extended version of the data dictionary, along with datasets corresponding to the different stages of data preparation (AUTM_0.xlsx through AUTM_10.xlsx, as well as additional community and university datasets), is available at https://github.com/GitHubAndres/UniversityInnovation. Updated versions of the AUTM datasets are not publicly available but can be requested directly from AUTM.

3. METHODOLOGY

The dataset served as the starting point for the design of this teaching exercise. As the materials were developed, we recognized that the dataset presents a realistic scenario for guiding students through a hands-on application of geospatial analytics focused on data preparation and enrichment. The exercise follows the sequence of analytical decisions involved in preparing location-aware data, allowing students to refine, enrich, and clean the dataset while encountering the same types of challenges that commonly arise in practice.

The exercise is grounded in a real-world context, incorporates structured questions with instructor guidance, and includes supporting materials that can be shared at different stages of instruction. The questions and challenges vary in scope and difficulty, allowing instructors to adjust time allocation and instructional format based on course level and audience characteristics (e.g., undergraduate versus graduate, technical versus non-technical).

For challenges involving a new tool, we suggest estimated time ranges based on our experience with student groups of 30-45 individuals and on whether the tool is already familiar to students or needs to be introduced. Instructors should adjust these times to their specific circumstances. All tools used in the exercise are either free or available through educational licenses, including Excel, Tableau, Python, Google Colab, ArcGIS, and ChatGPT.

The exercise is organized into a sequence of modules, each consisting of targeted questions and hands-on tasks that can be completed individually or in groups. Materials are provided so that each module can be used independently, regardless of whether prior modules have been completed. This modular design gives instructors flexibility to adapt the exercise to different instructional contexts, use selected modules, or integrate them into existing courses. After students present their responses, the instructor facilitates discussion and shares one possible analytical approach before students continue refining their datasets.

To complete the exercise, students draw on foundational data analytics concepts while applying them in a location-aware context. Tasks emphasize practical decisions related to data structure, unit of analysis, granularity, integration of external data sources, and preparation of datasets for subsequent analysis. Through repeated refinement and discussion, students gain experience working with enriched datasets that incorporate location-based attributes and support more context-sensitive analysis.

The exercise is designed as an active learning process that allows for multiple valid outcomes. Similar to case-based instruction, it does not aim to produce a single correct solution but instead provides a structured framework for exploring alternative data preparation strategies. Students are encouraged to experiment with different data structures and enrichment choices, while instructors ensure that the activity remains aligned with its instructional purpose: providing accessible, hands-on experience with geospatial data preparation in information systems and analytics courses.

4. LOCATION ANALYTICS EXERCISE DATA PREPARATION

To facilitate following the exercise, we provide the following flowchart that maps each module to the type of feedback the instructor should provide, the hands-on challenges, and the visualizations produced by completing the steps. This can be found in Appendix B. The duration of each module depends on students' familiarity with the concepts discussed and the amount of time the instructor devotes to class discussion. However, the estimated time required for the hands-on challenges can help instructors decide whether to use one or multiple class sessions to cover individual modules.

4.1 Module 1 (M1): Initial Dataset Exploration

In the first stage of the exercise, all students receive the AUTM_0.xlsx dataset, which contains self-reported innovation data from 2010 to 2015, along with its corresponding data dictionary. Their first challenge is to respond to the following questions:

Module 1 Question 1 (M1-Q1): What does each of the dataset rows represent?

M1-Q2: The structure of this dataset corresponds to:

a) Cross-sectional data
b) Time-Series data
c) Panel data

Note: Students can use search engines or Generative AI applications to learn about these data types.

M1-Q3: What are the average, minimum, and maximum values for Startups Formed (St-Ups Formed)?

M1-Q4: How many universities are in the dataset?

M1-Q5: What anomalies do you notice in this dataset?

M1-Q6: What columns represent location-based information?

4.1.1 M1 Feedback. Once the responses are collected, the instructor should discuss them and summarize the initial data exploration as follows:

Answer to M1-Q1 (M1-A1): The dataset rows represent university innovation and demographics data per year. However, some rows present subtotals that will affect the use of the data in the current state.

M1-A2: The dataset contains panel or longitudinal data that tracks the same university over time. It is not cross-sectional data because the subjects are observed at more than one time point. It is not time series data because the time field does not make each record unique.

M1-A3: The average Startups Formed is 4.4, the minimum is zero, and the maximum is 83 for the University of California System. Instructors can demonstrate this by performing an Average, Minimum, or Maximum aggregation of St-Ups Formed in Tableau or by creating a pivot table in Excel—placing the ID field in the Rows area and St-Ups Formed in the Values area—to find the average. INSTITUTION must be added to filters to remove Confidentials and Totals.

M1-A4: There are 231 unique institution IDs in the dataset. Instructors can demonstrate this by performing a COUNT(DISTINCT ID) aggregation in Tableau or by creating a pivot table in Excel—placing

the ID field in both the Rows and Values areas—to count the number of unique entries. If needed, instructors should clarify that the total number of rows in the AUTM_0 dataset includes repeated records and subtotals, which can inflate the count. Additionally, counting institutions by name can introduce errors due to inconsistent naming conventions across records, where the same institution may appear under different variations.

M1-A5: At least the following are clear anomalies:

1) There are missing values.
2) Some columns have no values at all.
3) Medschl values have inconsistent representations, e.g., N and no.
4) Some rows are used for subtotals, not for institutions.
5) The number of IDs and Institutions do not match.

Missing values are generally acceptable in datasets, provided that key identifiers—such as ID and INSTITUTION in this case—are present in every row. However, when too many values are missing, certain variables may become unreliable or unusable for analysis.

It is possible that students detect other problematic characteristics in the dataset. Most likely, these will be addressed later in the process. The instructor can acknowledge this.

M1-A6: State is clearly location-based information. St-Ups in Home St can be considered location-based information, as they are associated with the state. This initial data exploration reflects a standard step in any data analysis project. Question 6, however, shifts the focus to the State column to raise awareness of the unique nature of a variable that might otherwise be treated as just another categorical attribute. By explicitly prompting students to consider location-based information, the exercise highlights how location-related attributes differ from standard categorical variables. Students are encouraged to reflect on how place-related variables can influence interpretation and analysis.

4.1.2 M1 Hands-on Challenge (Estimated time: 10 to 20 minutes). Students are prompted to clean the data to conform with the previous findings. The safest way to do this is to create a copy of the dataset to make the changes. Students should:

1) Consolidate the values of Medschl as Yes or No.
2) Remove empty columns, i.e., the last columns on the right and the Fed Disclosure and INVDISRES columns.
3) Identify fields as measures or dimensions. In the dataset, dimensions' labels are in capital letters. Using the same style, State becomes STATE and NEWCOMPSUPP becomes New Comp Supp in both the dataset and data dictionary.
4) Remove rows with subtotals at the end of each year, i.e., those with the values Confidentials and Totals in the INSTITUTION column.

4.2 Module 2 (M2): Exploring the Spatial Component

The instructor requests that the new dataset be named AUTM_1.xlsx. A copy of this cleaner version will be available if the instructor prefers to ensure that students continue working with the exact same file. This is the case for subsequent stages.

From the AUTM_1 dataset, the first spatial explorations can be performed because the universities' locations are given—if only at the state level. The instructor can prompt students to consider how location-based attributes affect analysis.

M2-Q1: What interesting location-related questions can be responded to using this dataset?

4.2.1 M2 Feedback. This could be a challenging question for students, given that they are constrained to making associations with only one location-based variable: the state in which an institution is located. The instructor's role is to guide discussions around students' proposals, verifying that:

1) They are genuinely location-based questions.
2) They can be responded to using the available data.

3) The potential finding is relevant and meaningful.

This conversation helps illustrate how incorporating location-based attributes can expand the analytical possibilities of a constrained dataset, even in the early stages of data exploration.

M2-A1: Many answers are possible because all measures can be aggregated by state. In addition, we can count the number of institutions, their types, or the number of medical schools per state. Other combinations of those variables are also possible.

4.2.2 M2 Hands-on Challenge (Estimated time: 15 to 30 minutes). Students can now visualize many of their proposed findings. For the sake of consistency, we propose a common challenge: they must create a single chart showing the number of institutions per state, the gross license income per state, or both.

Two possible options are:

1) A bar graph sorted by institutions per state with colors indicating gross license income:

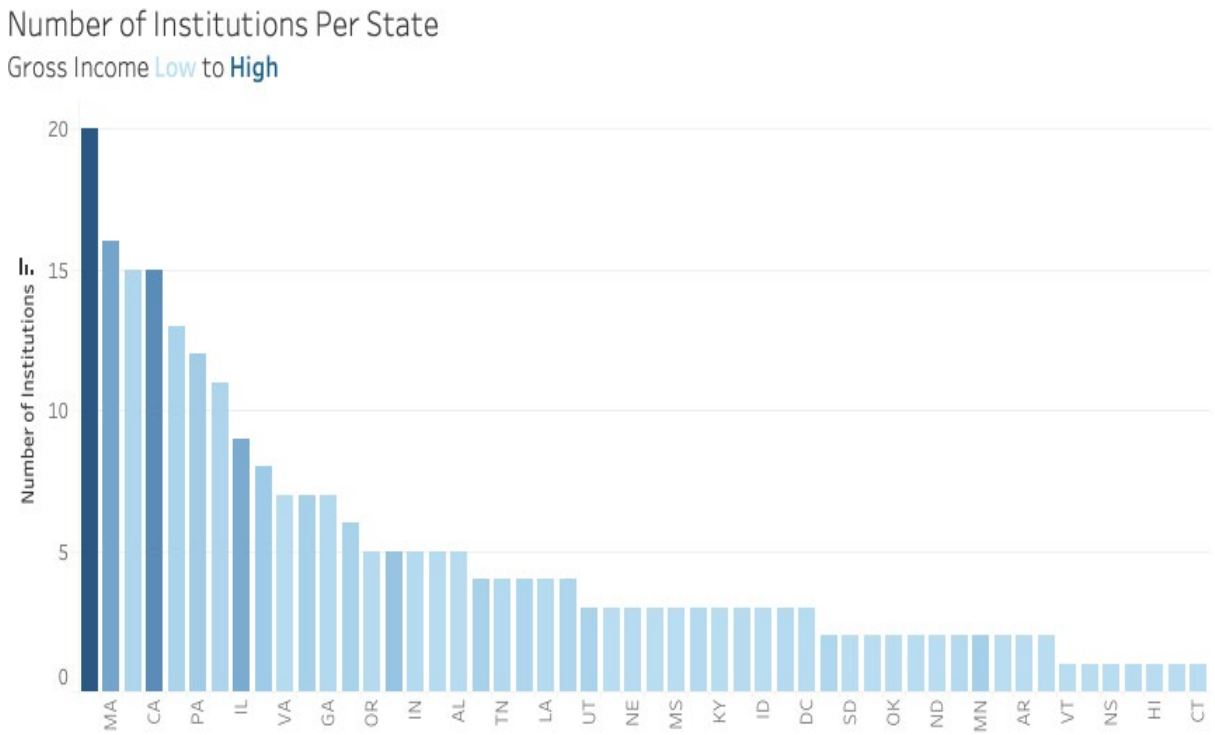


Figure 1. Bar Chart of Institutions and Gross License Income per State

An interesting observation from this type of graph is that visual representations in location-based analysis are not necessarily tied to geographic maps. Using this example, the instructor can emphasize that location analytics can be conducted even without mapping software. Moreover, it shows that other dimensions of location, such as relative position, connectivity, or functional proximity, are reflected in how analysts interpret relationships among locations.

2) A map with the number of institutions per state and darker colors indicating gross income:

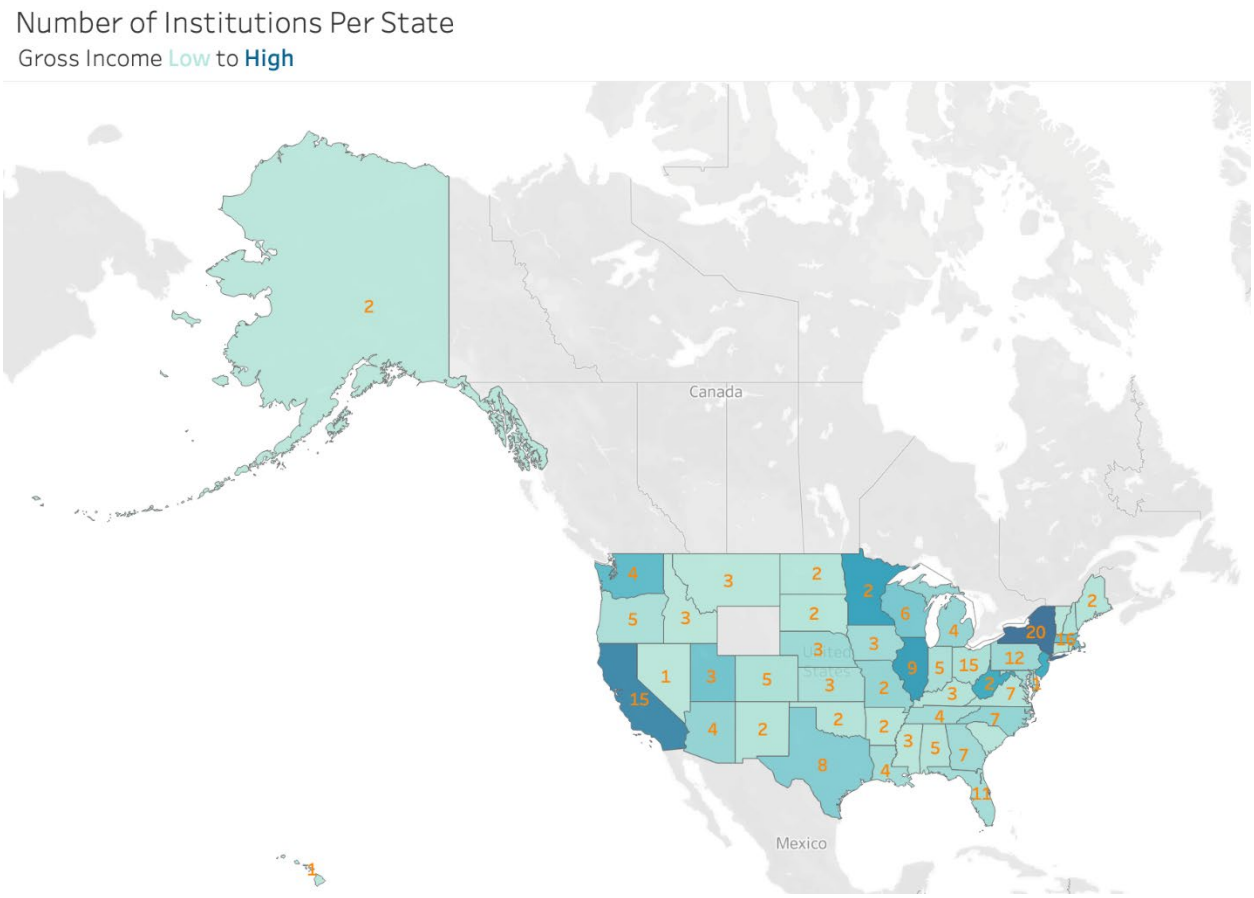


Figure 2. Map of Institutions and Gross License Income per State

Unlike bar charts, this representation leverages the traditional geographic display of location-based data and demonstrates the visual power of maps. Although not all location analyses produce geographic maps, map-based visualizations, when grounded in accurate geographic data, can convey information more clearly and memorably than other types of visualizations (Saket et al., 2015; Xin et al., 2018). For this reason, encouraging students to incorporate geographic views into their analysis is recommended. The instructor should take time to clarify this distinction at this point.

Both charts were created using Tableau, which offers free one-year licenses to students of academic institutions. Tableau also provides plenty of clear videos to help students learn to use the software, including how to create maps like the one in Figure 2. Other visualization options are possible. However, it is important to note that:

1) Institutions can show up in different years. Therefore, the count of institutions per state must be distinct.
2) Some records are missing values for STATE. Therefore, some counts may be incorrect. If students did not detect this issue before creating their charts, they should go and enter the appropriate values in the dataset and recreate their charts. It is advised that they consult the two-letter notation of states because not all are intuitively derived from the state's name. These are the missing values completed:

YEAR	[ID]	INSTITUTION	MEDSCHL	STATE
2015	72399	University of New Hampshire	no	NH
2014	71194	Auburn University	no	AL
2014	92614	California Institute of Technology	no	CA
2014	71372	Cold Spring Harbor Laboratory	no	NY
2014	80678	Louisiana State University System	yes	LA
2014	72140	San Diego State University	no	CA
2014	72222	Stanford University	yes	CA
2014	72342	University of Alabama in Huntsville	no	AL
2014	94410	University of Arkansas for Medical Sciences	yes	AR
2014	109486	University of Arkansas Fayetteville	no	AR
2014	95188	University of California System	yes	CA
2014	72399	University of New Hampshire	no	NH
2014	72432	University of Southern California	yes	CA
2014	95274	Washington State University Research Foundation	no	WA
2013	72399	University of New Hampshire	No	NH
2012	72399	University of New Hampshire	No	NH
2011	72399	University of New Hampshire	No	NH
2010	72399	University of New Hampshire	No	NH

Table 1. State Values for Institutions Missing Them

Students must ensure that their AUTM_1 dataset contains these values.

4.3 Module 3 (M3): Analytical Opportunities – Predictors and Targets

Researchers are often interested in identifying causal relationships. When examining datasets, they typically ask whether certain variables might influence or cause changes in others. Students should begin to recognize attributes as potential variables whose values may change relative to one another. The instructor should include the concept of predictor and target variables. The initial exploration should have introduced this concept through practical application; however, it is now essential to pose this question explicitly.

M3-Q1: What variables in the AUTM_1 dataset could be the cause (predictors) of others (targets)?

4.3.1 M3 Feedback. This is one of the most challenging questions to answer, and instructors should carefully consider students' justifications.

M3-A1: Our perspective is that variables capturing licenses, options, inventions, patents, and startups primarily represent downstream outcomes of research expenditures. Staffing levels in Technology Transfer Offices (TTOs)—administrative units responsible for managing intellectual property, licensing, and commercialization activities—are measured in Full-Time Equivalents (FTEs), where one FTE corresponds to the workload of a full-time employee. TTO staffing may increase in response to higher research funding, but it may also facilitate greater research commercialization and output, implying a potentially mediating or moderating role within a conceptual model. A formal assessment of these causal mechanisms, however, lies beyond the scope of this exercise. Additional institutional characteristics, including institution type and the presence of a medical school, may also systematically influence research output. Since the exercise focuses on incorporating location-based variables, we now turn to evaluating the role of the state variable.

4.4 Module 4 (M4): Enriching the Spatial Component

Our attention now shifts to spatial data. Although the AUTM dataset includes the state in which each institution is located, state-level data lacks the granularity needed to investigate innovation at the university level, which is our primary unit of analysis. This stage provides an excellent opportunity for the instructor

to introduce or reinforce key analytical concepts such as granularity and unit of analysis. To explore the potential influence of location on innovation outcomes—such as licenses, inventions, patents, and startups—we need more detailed information about the specific locational characteristics of each university.

Fortunately, we have access to three additional datasets: University Location, Community Attractiveness, and Community Demographics. The University Location dataset includes the Institution ID, the university, the state, and the city where each university is located, providing a more granular representation of location. It is also more reasonable to associate certain city-level characteristics—such as weather or air quality—with the university itself. In our dataset, Community Attractiveness is measured using biking and transit scores, the number of nearby airports, the air quality index, and the crime rate. The Community Demographics dataset includes variables such as median age, male population, female population, total population, number of veterans, and foreign-born in the city. For the purposes of this exercise, we made slight modifications to the original datasets.

Ideally, we should consolidate the available information into a single dataset representing location characteristics. We can then challenge students to make important decisions:

M4-Q1: We now have information about the state, the city, and the university. Which one is our unit of analysis?

1. Cities
2. Universities
3. States

M4-Q2: What operation must we implement to merge these datasets:

1. Unions
2. Inner Joins
3. Right Outer Joins
4. Left Outer Joins
5. Full Outer Joins
6. Cross Joins

4.4.1 M4 Feedback. Our unit of analysis is the university.

M4-A1: This question prompts students to think carefully about unit of analysis and provides a foundation for applying location-based reasoning to subsequent data preparation tasks. It also offers instructors an opportunity to distinguish between an entity and its surrounding geographic context. In this context, it is important to clarify that our unit of analysis is the university—a place—even though we incorporate characteristics of the surrounding city, which represents its space. This distinction reflects the assumption that universities are not only geographic points on a map but also entities that inherit attributes from their broader urban environments.

Furthermore, this question provides an opportunity to discuss how location-related information can be represented using both discrete attributes and continuous measures. Given the varying sizes of universities and cities—and the challenges of defining their boundaries for the purpose of measuring community-level effects on innovation outcomes—the instructor should lead a conversation on the assumptions necessary to move forward with most research processes.

M4-A2: Left Outer Joins are a good choice. This operation allows us to combine datasets horizontally, preserving all records from the primary table (e.g., university data) while adding attributes from the secondary dataset (e.g., city or state-level data) without losing variables.

4.4.2 M4 Hands-on Challenge 1 (Estimated time: 30 to 60 minutes). Students must perform the Left Outer Joins to merge all relevant tables. Python code is provided to help them complete this task. To begin, students should connect to Google Colab and upload the required datasets by following the instructions in Appendix C: How to Use Google Colab and Upload a Dataset.

Once connected, students must upload the files University Location.xlsx and Community Attractiveness.xlsx and choose the appropriate field(s) to merge them using the code found in Appendix D:

Merging University Location and Community Attractiveness Tables. *Note:* All code must be transcribed exactly as written in Python, except for the placeholder "Common Field(s)," which students must replace with the actual field name(s) they believe are appropriate for joining the two tables.

Joining the tables using only "**City**" creates duplicate records for institutions located in cities with the same name in different states (e.g., Newark, Delaware and Newark, New Jersey). On the other hand, joining the tables using only "**State**" creates cross-product matches between all cities within the same state. Therefore, the correct approach is to merge the tables using both "City" and "State" as join keys.

After completing the merge, students can convert the UniversityAttractiveness_0_df DataFrame into a CSV file by entering the following code:

UniversityAttractiveness_0_df.to_csv('UniversityAttractiveness_0.csv', index=False). Then, they can download the CSV file using the following command: from google.colab import files files.download('UniversityAttractiveness_0.csv').

Unfortunately, the merged dataset is missing attractiveness values for several institutions. This provides an opportunity to challenge students to identify and resolve the issue. Upon inspection, they will notice that in the CommunityAttractiveness_0.xlsx table, Boston was misspelled as Bostom. Students must correct the spelling and save the revised table as CommunityAttractiveness_1.xlsx. After repeating the merge using the corrected table, the join should succeed, and they can use the same code structure to merge the resulting dataset with CommunityDemographics_0.xlsx. The merge code is provided in Appendix E: Merging University Attractiveness and Community Demographics Tables.

4.4.3 M4 Hands-on Challenge 2 (Estimated time: 30 to 60 minutes). Finally, students must merge the resulting table with the AUTM_1.xlsx dataset. However, this time the join is performed using the institution ID as the key and applying an Outer Join. This type of join retains all records from both tables, including those without a match in the other. As a result, the merged table will include institutions from the AUTM dataset that lack attractiveness or demographics data, as well as locations with attractiveness and demographics data but no corresponding innovation data. The code can be found in Appendix F: Merging Autm1 and Autm2 Tables.

This task helps students understand how different join operations work and when each should be applied. The outcome of this module is a dataset that combines institutions' innovation data with characteristics of their respective locations.

4.5 Module 5 (M5): Cleaning Mismatches

One issue arises after the final merge. The outer join reveals that some institutions have no associated attractiveness or demographic data, while some locations with such data are not linked to any institution. To raise students' awareness of the mismatch, the instructor can ask:

M5-Q1: Which institutions with location data lack innovation data?

M5-Q2: What should we do with these institutions?

a) Remove them from analysis.

b) Search for additional information to complete the data.

4.5.1 M5 Feedback. Students should be able to identify some of the institutions missing innovation data by visually exploring the missing values for certain IDs in the AUTM_2 dataset.

M5-A1: Based on their IDs, the following institutions have location data but lack innovation data:

ID	INSTITUTION_y	City	State
71180	Arizona State University	Tempe	AZ
71361	Children's Hospital Cincinnati	Cincinnati	OH
71367	Cleveland Clinic	Cleveland	OH
71402	Dana-Farber Cancer Institution	Boston	MA
71510	Fred Hutchinson Cancer Research Center	Seattle	WA
71580	H. Lee Moffitt Cancer Center & Research Institute	Tampa	FL
71789	Massachusetts Institute of Technology	Cambridge	MA
71822	Medical College of Wisconsin Research Foundation	Milwaukee	WI
72031	Pennsylvania State University	State College	PA
72134	Salk Institute for Biological Studies	San Diego	CA
72299	Thomas Jefferson University	Philadelphia	PA
72513	Washington University in St. Louis	St. Louis	MO
72514	Wayne State University	Detroit	MI
85917	Nationwide Children's Hospital	Columbus	OH
92821	Georgia Institute of Technology	Atlanta	GA
94788	Mount Sinai School of Medicine of NYU	New York	NY
20160310	Temple University	Philadelphia	PA

Table 2. Institutions With Mismatched IDs

M5-A2: Both options are valid. If the number of mismatched institutions is small and the missing data is difficult to obtain, those records can be removed from the analysis. However, all the institutions in the list do appear in the dataset—just under a different ID. In such cases, students can reconcile the records by linking the location data to the corresponding institution and retaining the ID from the location dataset, as it reflects the most recent update. We also recommend spelling out the names as they appear in Table 2.

4.5.2 M5 Hands-on Challenge (Estimated time: 15 to 30 minutes). Beyond simple observation, students must be able to recreate the exact list of institutions with location data but lacking innovation data. This can be done in Excel by applying filters to the columns, selecting the blank cells under the newly created INSTITUTION_x column, and hiding all empty columns between the ID and INSTITUTION_y columns to simplify the view. Once the complete list is generated, students should identify institutions that appear elsewhere in the dataset under a different ID and manually correct the information. We recommend that instructors advise students to make these changes by opening the AUTM_1 table, saving a new version as AUTM_3.csv, and editing the institution IDs for each year in which they appear. Students can then upload AUTM_3.csv to Colab and perform the final merge, naming the resulting dataset AUTM_4.csv. The code can be found in Appendix G: Merging Autm3 and Autm4 Tables.

4.6 Module 6 (M6): Creating A Cross-Sectional Dataset

We are also working with an unbalanced panel dataset. While we have six years of data, not all institutions report data for every year. In addition, some values exhibit large year-to-year variations. Although the time component adds analytical value during advanced stages, our current focus is on exploratory analysis, where researchers are primarily testing correlations. To ensure more consistent comparisons among institutions with varying reporting frequencies, we decided to aggregate the values of all numeric variables

across the available years. This introduces an additional challenge: deciding which values to retain across dimensions. For example, what year should be kept, or which version of the institution name should be used when variations appear across the dataset? Students must answer the following questions:

M6-Q1: What is the appropriate aggregation method for measures?

a) Average
b) Median
c) Sum
d) Other

M6-Q2: What value should be retained across dimensions such as YEAR or INSTITUTION?

4.6.1 M6 Feedback. Options vary.

M6-A1: We recommend using the average for all numeric measures, given the high variability observed across years. However, instructors may choose alternative approaches and are encouraged to discuss their rationale with students.

M6-A2: We retained the most recent value for categorical dimensions.

4.6.2 M6 Hands-on Challenge (Estimated time: 15 to 30 minutes). Our AUTM_4.csv dataset now contains all available information in a consolidated table. Students can execute the Python code in Appendix H: Average and Recent Values. Then edit the dataset in Excel.

Step 1: Python. To transform AUTM_4 into a cross-sectional dataset, the code performs two key actions based on the decisions previously made:

1) Computes the average for all numeric measures.
2) Retains the most recent value for non-numeric dimensions.

Step 2: Excel. The AUTM_5.csv file from the previous step can be opened in Excel to rearrange the order of columns and adjust column data types. Excel may automatically convert some numeric values into scientific notation; therefore, students should change the formatting of relevant columns to Number or Currency, as appropriate. Once the necessary edits are completed, we recommend saving the file as an Excel document to preserve the formatting and naming it AUTM_6.xlsx.

4.7 Module 7 (M7): Reviewing the Unit of Analysis

Because some universities reported data at the system level rather than at the campus level—making it difficult to associate them with a single campus located in a specific city—students must respond to the following question:

M7-Q1: Given that some records represent university systems, and our unit of analysis is the university campus, what should we do with those system-level records?

4.7.1 M7 Feedback. The discussion centers on the trade-off between data cleansing and sample size.

M7-A1: Records that represent more than one campus should be removed from the dataset. The characteristics of a single location cannot be assumed to apply uniformly across all campuses within a university system, and the aggregate innovation data reported at the system level cannot be reliably disaggregated to individual institutions. This becomes evident when considering location-related data and unit-of-analysis decisions. The instructor must clarify that losing records from a dataset is generally undesirable, as larger sample sizes are typically preferred in research. However, including observations that represent a different unit of analysis can introduce significant bias and lead to erroneous conclusions.

4.7.2 M7 Hands-on Challenge (Estimated time: 10 to 20 minutes). Students should identify records that clearly represent data from more than one campus and remove them from the dataset. The revised version should be saved as AUTM_7.xlsx. In our analysis, we identified and removed the following institutions:

Texas A&M University System

Texas Tech University System
Louisiana State University System
University System of Maryland
University of Missouri all campuses
University of Oklahoma All Campuses
University of California System
University of Texas System
Children's National Health System

It is possible that additional records fall into this category, which requires an individual review of each institution. The primary goal of this exercise is to emphasize the importance of considering the unit of analysis when preparing data. In addition to the institutions removed for system-level reporting, some records lack city-level information because no demographic or attractiveness data were available. While researchers would typically attempt to complete missing data through supplemental sources, for the purposes of this exercise, we remove these records to simplify the analysis. These additional institutions, which do not contribute to the spatial component of the study, are as follows:

Washington State University
Research Foundation of State University of New York
University of Alaska
California State University Institution

4.8 Module 8 (M8): Using Web Scraping for Data

Through our investigation, we identified additional location-based factors that may influence our target variables—for example, walking scores and prevalent occupations in the area (CARE, 2017). Although this information is not included in the original dataset, it is publicly available online. To help students become familiar with the process of retrieving external data, we introduce a hands-on challenge in which they crawl the web to obtain relevant information. In this case, students focus on collecting walking scores for cities.

4.8.1 M8 Hands-on Challenge (Estimated time: 30 to 60 minutes). Using the code provided in Appendix I: Setting Up the Walk Score Scraper in Google Colab, students can scrape the web to obtain additional walking score data for cities and incorporate it into the dataset. The resulting file, AUTM_8.csv, includes a new column, "WalkScore", appended to the end of the table. The scraper does not retrieve scores for all cities, which is expected due to data availability and web limitations. Nevertheless, this process demonstrates the value of web scraping as a practical tool for enriching datasets with external information.

4.9 Module 9 (M9): Preliminary Analyses

At this stage, we have expanded our university innovation dataset by incorporating additional location-based attributes through table merges and web-based data enrichment. To illustrate the value of this data preparation process, we invite students to conduct basic correlation analyses using findings reported by CARE (2017). Specifically, students can examine whether walking scores are correlated with key innovation indicators, such as gross license income, startups formed, patent applications filed, and licenses/options executed. The corresponding Python code is provided in Appendix J: Running Correlations in Google Colab.

The output, shown in Table 3, displays statistically significant correlations (*) between walking scores and gross license income, startups formed, and patent applications filed. However, the correlation between walking scores and licenses/options executed is not significant.

Input Column	Output Column	Correlation	P-Value	Rows Used
WalkScore	Gross Lic Inc	0.377679	0.000002*	152
WalkScore	St-Ups Formed	0.208941	0.014640*	136
WalkScore	Tot Pat App Fld	0.189745	0.016947*	158
WalkScore	Tot L&O Exec	0.063736	0.429258	156

Table 3. Preliminary Correlation Analysis

These results introduce students to key analytical concepts such as correlation, p-value, and sample size. This creates an opportunity for the instructor to initiate a discussion around these topics and transition into the analysis stage. Our primary objective, however, was to guide students through a step-by-step process of building a dataset that enables such analyses.

Before concluding the exercise, we introduce an effective method for integrating artificial intelligence into the data preparation process. This is described in the following module.

4.10 Module 10 (M10): Using GenAI

Despite having a dataset enriched with location-based characteristics, one crucial element was still missing: the precise geographic location of each institution. Traditionally, collecting this information would require manually searching for each institution—an effort that could take several hours. To streamline this process, we employed Generative AI (GenAI) to retrieve the latitude and longitude of each university. While GenAI is known for occasional hallucinations, it tends to be reliable when asked for factual data—particularly when clear sources are defined and publicly accessible. We used the following prompt with ChatGPT:

> *"Please, give me a CSV table with the following columns: INSTITUTION, Lat, Long. Where: INSTITUTION is the name of the institution that I will give you, Lat is the latitude of the location of that institution, Long is the longitude of the location of that institution. Your first source of reference must be Wikipedia and Google Maps but use alternative sources if needed and indicate what source you used with a link to the source. All data must be real and verifiable."*

After the prompt, we provided the list of all universities from the AUTM_8 dataset. As expected, the results were accurate. ChatGPT returned a CSV-like list, which we imported into Excel, parsed into columns, and then appended to the existing dataset. The resulting file, which now includes Latitude and Longitude columns, was saved as AUTM_9.csv, as shown in Appendix K: Enriched Autm9 Dataset.

Having precise geographic coordinates for each institution significantly enhances the ability to use maps and other location-based visualizations. The visualization in Figure 3 provides a more detailed and nuanced view than Figure 1, where the analysis was constrained to the state level.

Figure 3. Geospatial Location of Institutions

Figure 3 was produced using ArcGIS, a platform developed by Esri Inc., which is available to students and educators under free academic licenses. ArcGIS enables users to conduct a wide range of advanced spatial analyses and further enrich datasets by leveraging resources such as U.S. Census data, the World Living Atlas, various base maps, and analytical tools for proximity, clustering, routing, and more. To support instructors introducing students to this software, we propose the following challenge:

4.10.1 M10 Hands-on Challenge (Estimated time: 15 to 30 minutes). Recreate the following chart in ArcGIS:

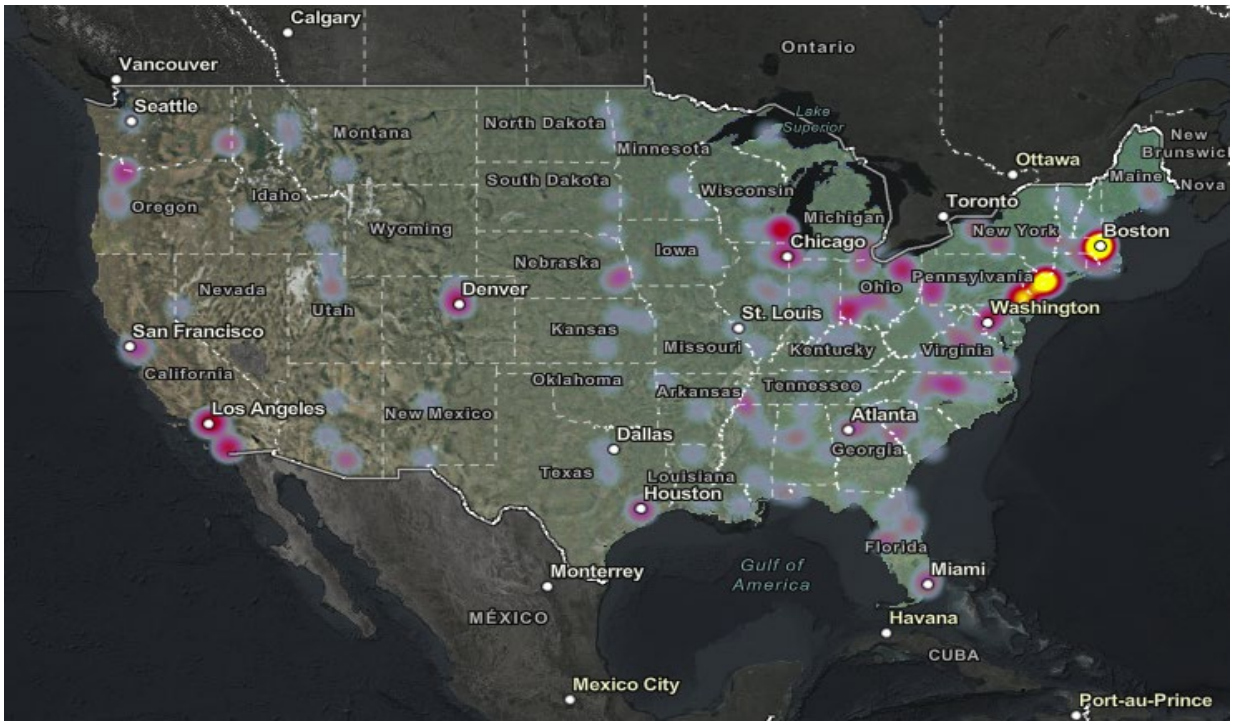


Figure 4. Heat Map Over NAIP Imagery Hybrid

We recommend creating a new dataset, AUTM_10.xlsx, to complete this task. Unlike the CSV format, Excel supports the easier assignment of appropriate data types—such as numeric values to measures—which facilitates more accurate data recognition by ArcGIS when uploading files. The process can be summarized in the following steps:

1. Log in to arcgis.com.
2. Select Map from the top menu.
3. Click Add and choose Add Layer from File.
4. Upload the AUTM_10.xlsx dataset.
5. Choose the NAIP Imagery Hybrid basemap.
6. From the Settings toolbar on the right, apply the Heat Map style.

The analytical path from this point—whether pursued in Python or through a GIS platform—will present many additional challenges. For example, substantial differences remain between well-funded research institutions and smaller regional universities. To make meaningful comparisons, students will need to learn about normalization and other key analytical concepts that account for such disparities. We are satisfied to have guided learners to this stage of enhanced readiness for deeper analytical work.

This concludes the exercise.

5. CLASSROOM IMPLEMENTATION AND STUDENT FEEDBACK

The exercise was implemented in an undergraduate *Introduction to Analytics* course at a large public university. Students completed selected modules of the exercise as part of regular course activities and subsequently provided feedback regarding their experience. A total of 199 complete student responses were collected and analyzed descriptively.

Overall, student feedback indicated strong engagement with the exercise and positive perceptions of its usefulness. The majority of students reported interest in using the tools introduced in the exercise and in further developing skills related to location analytics. Specifically, 163 students indicated that they would

like to use Tableau and ArcGIS tools introduced in the exercise, 35 expressed interest in using one of the tools, and only one student reported no interest in using either tool. In addition, 158 students stated that they would be likely or very likely to incorporate location analytics into their future academic or professional work.

Students also reported that the exercise was accessible and manageable within the course context. Feedback suggested that the step-by-step structure, modular design, and hands-on data preparation tasks supported student participation, even for those with limited prior exposure to geospatial or location-based analysis. The ability to work incrementally with the dataset and observe how data preparation and enrichment decisions affected subsequent analysis was frequently cited as a valuable aspect of the exercise.

Taken together, these observations suggest that the exercise is feasible to implement in analytics courses and that students respond positively to its focus on practical data preparation and location-based enrichment. While this feedback is not intended as a formal evaluation of learning outcomes, it provides evidence that the exercise can be successfully integrated into classroom instruction and that students perceive it as useful and relevant.

6. LIMITATIONS

This teaching tip illustrates a specific data preparation and enrichment workflow using a particular dataset and set of tools. While the exercise reflects common challenges encountered in real-world analytics projects, it is not intended to be comprehensive or exhaustive. Other data preparation scenarios may involve different data structures, sources, or constraints that require alternative approaches.

Although most of the tools used in this exercise are accessible through free or academic licenses, instructors and students may not have access to the same software in all instructional contexts. The exercise is intentionally designed to be software-agnostic, and instructors are encouraged to adapt the workflow to alternative tools that align with their institutional resources and course objectives.

The modular structure of the exercise allows instructors to select and sequence activities based on course level, student background, and available instructional time. As a result, not all students will necessarily complete every module, and learning experiences may vary depending on how the materials are implemented. The exercise is best viewed as a flexible instructional resource rather than a prescriptive curriculum.

Finally, this teaching tip focuses on data preparation and geospatial enrichment as a foundation for analysis. While preliminary analytical tasks are included to demonstrate how enriched data can be used, the exercise does not aim to assess learning outcomes or compare instructional effectiveness across pedagogical approaches. Its primary contribution is to provide a practical, reusable framework for incorporating location-based data preparation into analytics instruction.

7. NEXT STEPS

Data preparation is a natural precursor to data analysis, and the workflow presented in this teaching tip represents only one possible path. Even when working with the same dataset, instructors and students may choose different strategies for cleaning, enrichment, and restructuring based on analytical goals, data availability, and instructional context. These alternative paths can serve as valuable discussion points in the classroom and reinforce the judgment-driven nature of data preparation.

The exercise is intentionally designed to be adaptable beyond the specific dataset used here. Instructors may substitute alternative innovation, organizational, or publicly available datasets while preserving the core emphasis on location-based enrichment, unit-of-analysis decisions, and data integration. Similarly, individual modules can be extended to incorporate additional external data sources, alternative enrichment techniques, or more advanced analytical methods as appropriate for course level and student background.

Future instructional extensions may also explore emerging tools to support data preparation and enrichment, including analytics platforms and generative AI technologies. As tools and data sources continue to change, the underlying skills emphasized in this exercise—careful data preparation, thoughtful

integration of location-based attributes, and critical evaluation of analytical assumptions—remain relevant across analytics curricula.

8. CONCLUSIONS

This teaching tip presents a hands-on, modular exercise that introduces location-informed analytics through the data preparation process. Using university innovation data enriched with location-based attributes, the exercise guides students through a realistic sequence of data cleaning, enrichment, integration, and restructuring decisions commonly encountered in analytics projects. By working incrementally with the dataset, students gain practical experience incorporating location into analysis without relying on specialized geospatial training.

The primary contribution of this work is instructional. The exercise provides instructors with a flexible, instructor-ready resource that can be adapted to different courses, tools, and student backgrounds. Its modular structure allows activities to be selected and sequenced based on instructional goals, while the accompanying guidance supports classroom discussion around key analytical decisions such as unit of analysis, data granularity, and the integration of external data sources.

By focusing on geospatial data preparation as an entry point, the teaching tip lowers barriers to incorporating location analytics into information systems and analytics curricula. The exercise demonstrates how location-based enrichment expands analytical possibilities and provides a practical foundation for subsequent analysis. Instructors seeking to introduce or expand coverage of location analytics will find the materials readily adaptable and aligned with common learning objectives in analytics education.

9. REFERENCES

Center for Analytics Research and Education (CARE). (2017, June 15). Technology Transfer Policy Analysis. *Chief Research Officer Meeting, UNC General Administration, Appalachian State University*.

Erskine, M. A., Pick, J. B., Satpathy, A., Díaz López, A., Sarkar, A., & Shin, N. (2024). Location Analytics in Information Systems: Opportunities for Research and Teaching. *Communications of the Association for Information Systems*, 55(1), 21. https://doi.org/10.17705/1CAIS.05521

King, M. A., & Arnette, A. N. (2011). Integrating Geographic Information Systems in Business School Curriculum: An Initial Example. *Decision Sciences Journal of Innovative Education*, 9(3), 325-347. https://doi.org/10.1111/j.1540-4609.2011.00318.x

Logan, J. R. (2012). Making a Place for Space: Spatial Thinking in Social Science. *Annual Review of Sociology*, 38(1), 507-524. https://doi.org/10.1146/annurev-soc-071811-145531

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1), 415-444. https://doi.org/10.1146/annurev.soc.27.1.415

Malik, G., Markworth, N., Nomula, A., Ostadi, P., Volstad, S., Hancock, S. W., Brown, C. S., & Cazier, J. A. (2023). University Technology Transfer: A Data Set Showing How University Policies, Demographics and Location Influence Success (Version 4) [Data set]. *Data & Analytics for Good*.

North Carolina Department of Commerce. (2016). *Recommendations of the Governor's University Innovation Council*. https://www.commerce.nc.gov/governors-university-innovation-council-recommendations/download?attachment

Ritchie, W. J., Kerski, J., Novoa, L. J., & Tokman, M. (2023). Bridging the Gap Between Supply Chain Management Practice and Curriculum: A Location Analytics Exercise. *Decision Sciences Journal of Innovative Education*, 21(2), 83-94. https://doi.org/10.1111/dsji.12286

Saket, B., Scheidegger, C., Kobourov, S. G., & Börner, K. (2015, June). Map-Based Visualizations Increase Recall Accuracy of Data. *Computer Graphics Forum*, 34(3), 441-450. https://doi.org/10.1111/cgf.12656

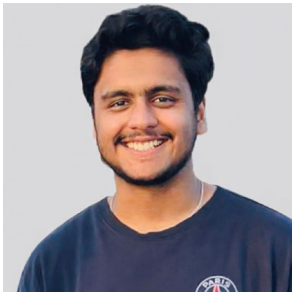
Whittington, K. B., Owen-Smith, J., & Powell, W. W. (2009). Networks, Propinquity and Innovation in Technological Communities. *Administrative Science Quarterly*, 54, 90-122. https://doi.org/10.2189/asqu.2009.54.1.90

Wineman, J. D., Kabo, F. W., & Davis, G. F. (2009). Spatial and Social Networks in Organizational Innovation. *Environment and Behavior*, 41(3), 427-442. https://doi.org/10.1177/0013916508314854

Xin, R., Ai, T., & Ai, B. (2018). Metaphor Representation and Analysis of Non-Spatial Data in Map-Like Visualizations. *ISPRS International Journal of Geo-Information*, 7(6), 225. https://doi.org/10.3390/ijgi7060225

AUTHOR BIOGRAPHIES

Andrés Díaz López is a Clinical Associate Professor at Arizona State University (ASU). He holds degrees in Industrial Engineering, Statistics, Marketing, and Information Systems. He has taught courses in information systems, technology management, data visualization, databases, case analysis, cyber risk management, privacy, ethics and compliance, and digital transformation at institutions in the United States, Colombia, China, and Chile. He co-created the Location Analytics in Business course at ASU and, in 2022, co-founded the Location Analytics in Business Symposium, which brings together experts annually to discuss cutting-edge applications in the field. His research interests include data visualization, location analytics, cybersecurity, artificial intelligence, innovation, pedagogy, smart systems, and sustainability.



Saisrinivas Mamunuru holds an M.S. in Data Science, Analytics and Engineering from the Ira A. Fulton School of Engineering at Arizona State University and a B.E. (Hons.) in Computer Science from BITS Pilani, Goa, India. He is currently a Software Engineer at Aetna, a CVS Health company, where he works on data-driven and AI interoperability initiatives. His research interests include data-driven decision making, user and customer behavior analytics, deep learning, and LLM inference optimization.

Joseph Cazier is Associate Director of the Center for AI and Data Analytics and a Clinical Full Professor at Arizona State University who coaches students and business leaders on how to succeed in making better decisions with data. He previously served as the Dean's Club Professor, Associate Dean, and the Founding Executive Director of the Center for Analytics Research and Education at Appalachian State University, where he designed and directed analytics programs and worked with industry and governments in finding ways to make better decisions with data, and retains the title of Full Emeritus Professor.

APPENDICES

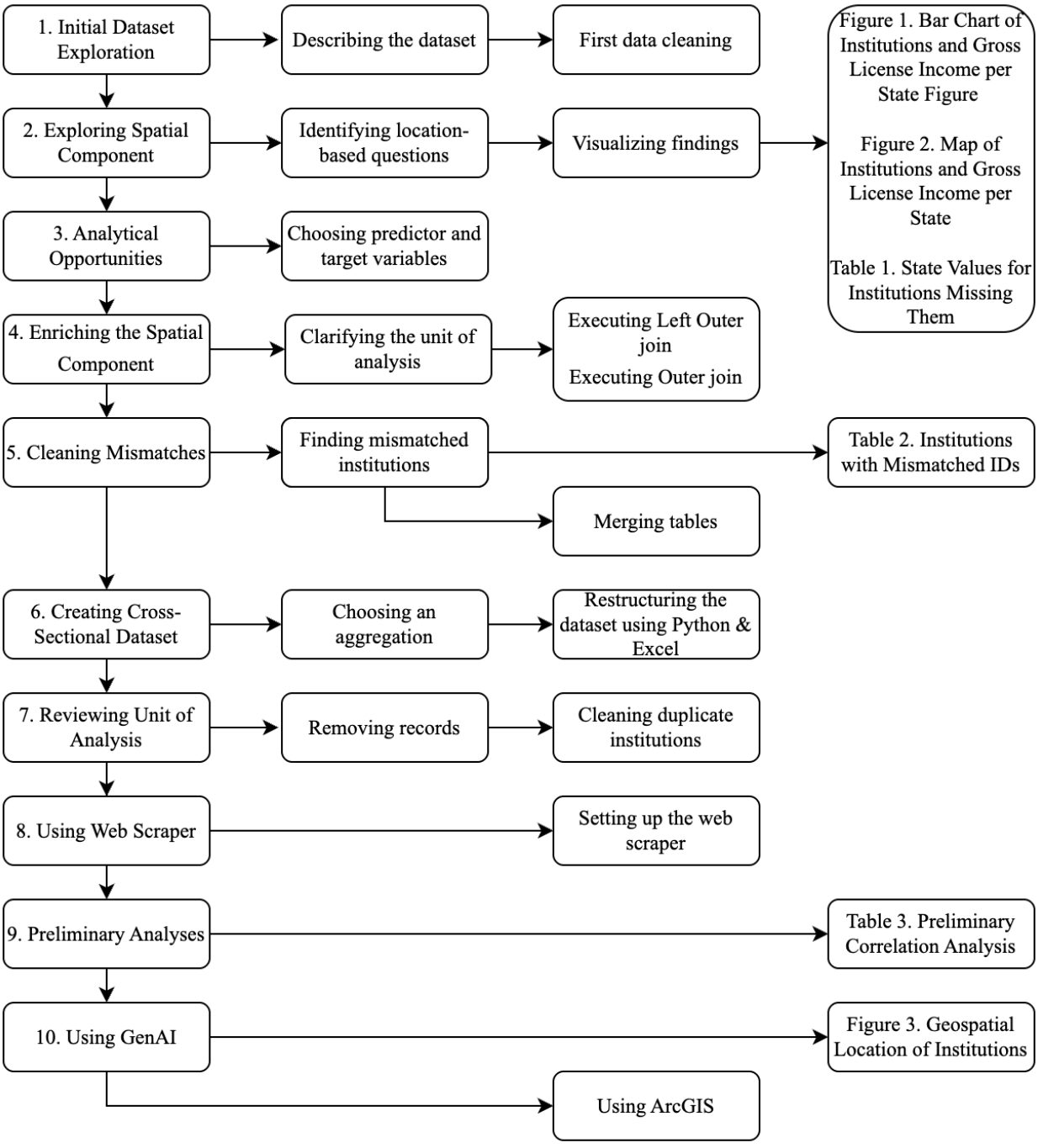
Appendix A. Data Dictionary

Column	Type	Description
WalkScore	Int	Walk Score measures the walkability of any address based on the distance to nearby places and pedestrian friendliness.
Transit Score	Int	Transit Score measures how well a location is served by public transit based on the distance and type of nearby transit lines.
Bike Score	Int	Bike Score measures whether an area is good for biking based on bike lanes and trails, hills, road connectivity, and destinations.
AirQualityIndex	Int	Air Quality according to https://www.airnow.gov/
Number of Airports Nearby	Int	Number of google search results for "Number of Airports nearby" + "city" + "state"
Median Age	Float	The median age of the population in the city, representing the age at which half the population is older and half is younger.
Male Population	Int	The total number of male residents in the city.
Female Population	Int	The total number of female residents in the city.
Total Population	Int	The total number of residents in the city, including all age groups and genders.
Number of Veterans	Int	The number of individuals in the city who have served in the U.S. military.
Foreign-born	Int	The total number of residents in the city who were born outside the United States.
Crime Rate	Float	The number of reported crimes per 1,000 residents per year in a given area. This includes both violent and property crimes, providing an overall measure of crime risk in the neighborhood.
[ID]	Int	Unique ID identifying institution
INSTITUTION	String	Name of the University (Systems excluded)
City	String	City in which the University is located
State	String	State in which the University is located
MEDSCHL	Bool	Does university have a medical school
INSTTYPE	Enum	Institution classification - Research University/Graduate Level University etc.
Tot Res Exp	Float	Total Research Expenditure
Tot Lic/Opt Exe	Float	Total Licenses and Options Executed
Fed Res Exp	Float	Federal Research Expenditure
Ind Res Exp	Float	Industrial Research Expenditure
Leg Fees	Float	Legal Fees Reimbursements include the amount reimbursed by licensees to the institution for LEGAL FEES EXPENDITURES
Reimb Leg Fee	Float	Legal Fees Reimbursements include the amount reimbursed by licensees to the institution for LEGAL FEES EXPENDITURES
Act Lic	Float	The cumulative number of LICENSES/OPTIONS, over all years, that had not terminated by the end of the Survey's year requested.
Excl Lic/Opt.	Int	The reporting of a license as exclusive or non-exclusive should follow the terms of the license agreement. If a license is designated as exclusive in the license agreement, it should be reported as an exclusive license to this Survey. Exclusive licenses include licenses that are designated as exclusive by field of use, territory, or otherwise but excludes co-exclusive licenses, which are reported as NON-EXCLUSIVE LICENSES.
Lic w Equ	Int	The number of LICENSES/OPTIONS that were executed in the year surveyed that included EQUITY, where EQUITY is defined as an institution acquiring an ownership interest in a company.
Tot Lic Lg Co	Float	Companies that had more than 500 employees at the time the license/option was signed.

Tot Lic Sm Co	Float	Companies that had 500 or fewer employees at the time the license/option was signed, but, for the purposes of this Survey, not including START-UP COMPANIES initiated by an institution.
Tot Lic St-Ups	Float	The number of LICENSES/OPTIONS from start up companies
Tot Discl Lic	Float	Disclosures include the number of disclosures, no matter how comprehensive, that are submitted during the survey year requested and are counted as received by the institution.
Non-Excl Lic/Opt	Int	The reporting of a license as exclusive or non-exclusive should adhere to the terms of the license agreement. If a license is designated as non-exclusive or co-exclusive in the license agreement, it should be reported under non-exclusive licenses to this Survey.
Lic Iss	Int	Licenses issued
Opt Iss	Int	Options issued
NEWCOMPSUPP	Float	Number of New Companies Formed with Support from the Institution
Lic Inc Equ	Float	The number of LICENSES/OPTIONS that were executed in the year surveyed that included EQUITY, where EQUITY is defined as an institution acquiring ownership interest in a company.
Lic Gen Inc	Float	The number of LICENSES/OPTIONS that generated LICENSE INCOME RECEIVED in the year requested.
Lic Gen Run Roy	Float	The number of LICENSES/OPTIONS that generated RUNNING ROYALTIES in the year requested.
Lic Inc Other	Float	The number of LICENSES/OPTIONS that were executed in the year surveyed that included other income
Lic Inc Pd Oth	Float	License income paid to other institutions under inter-institutional agreements.
Gross Lic Inc	Float	License Income Received includes: license issue fees, payments under options, annual minimums, running royalties, termination payments, the amount of equity received when cashed-in, and software and biological material end-user license fees equal to $1,000 or more, but not research funding, patent expense reimbursement, a valuation of equity not cashed-in, software and biological material end-user license fees less than $1,000, or trademark licensing royalties from university insignia. License Income also does not include income received in support of the cost to make and transfer materials under Material Transfer Agreements.
Lic Inc Run Roy	Float	The number of LICENSES/OPTIONS that generated RUNNING ROYALTIES in the year requested.
Inv Dis Rec	Float	Invention Disclosures Received
New Pat App Fld	Float	New Patent Applications Filed are the first filing of the patentable subject matter. NEW PATENT APPLICATIONS FILED do not include continuations, divisionals, or reissues, and typically do not include CIPs.
For. Pat App Fld	Float	License income paid to other institutions under inter-institutional agreements.
Pro Pat App Fld	Float	U.S. Provisional Application became available June 8, 1995 as a new type of US patent application that can be used to obtain a filing date, and is less formal than a “regular” application. It may be filed without claims or named inventors.
Util Pat App Fld	Float	US Utility Patent Applications is the “regular”, non-provisional, U.S. patent application filed during the survey year for new, useful and nonobvious machines, manufactures, compositions of matter, processes or any new or useful improvements of the above.
Tot Pat App Fld	Float	Total Patent Applications Filed
Iss US Pat	Float	U.S. Patents Issued includes the number of U.S. patents issued or reissued to a institution in the year requested.
New Prod	Float	Number of new products created from startups
Cum Op St-Ups	Float	Cumulative operating Startup companies
St-Ups Cld	Float	Startup companies closed
St-Ups in Home St	Float	Startup companies created within state in which university is located
St-Ups Formed	Int	Number of Startups formed

St-Ups w Equ	Float	Startup companies created with Equity
Lic FTEs	Int	Person(s) employed in the Technology Transfer Office whose duties are specifically involved with the licensing and patenting processes as either full or fractional FTE allocations. Licensing examples include licensee solicitation, technology valuation, marketing of technology, license agreement drafting and negotiation, and start-up activity efforts.
Oth FTEs	Float	The number of non-licensing staff working in the TTO or equivalent innovation-related unit, measured in full-time equivalents (FTEs).

Appendix B. Flowchart of the Exercise



Detailed Description of Activities:

4.1 Module 1 (M1). Initial Dataset Exploration
4.1.1 M1 Feedback: Describing the dataset.
4.1.2 M1 Hands-on Challenge: First data cleaning
4.2 Module 2 (M2). Exploring the Spatial Component

Appendix C. How to Use Google Colab and Upload a Dataset

Google Colaboratory (Colab) is a free cloud-based platform that allows you to write and execute Python code in a Jupyter Notebook environment.

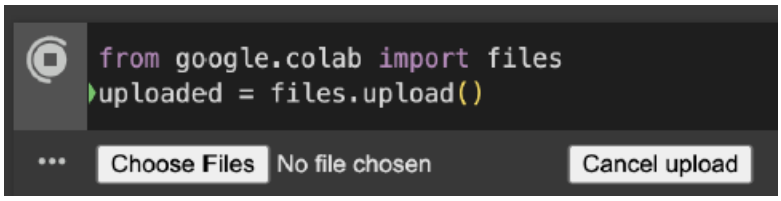
Part 1: Accessing Google Colab

1. Open a web browser (Google Chrome is recommended).
2. Go to Google Colab at https://colab.research.google.com/.
3. Sign in with your Google account if you haven't already.
4. Click on File > New Notebook to create a new notebook.

Part 2: Uploading a Dataset

1. Run the following code in a Colab cell and click on the Choose Files button that appears.

```
from google.colab import files
uploaded = files.upload()
```



2. Select your dataset file (CSV, EXCEL, etc.) from your computer. The file will be uploaded, and you can access it using Pandas or any other method:

```
import pandas as pd
df = pd.read_excel('University Location.xlsx')
print(df.head())
```

```
   [ID]                                        INSTITUTION        City State
0  95180                                University of Akron       Akron    OH
1  81632  The Research Foundation for The State Universi...      Albany    NY
2  95228       University of New Mexico/Sci. & Tech. Corp.  Albuquerque    NM
3  71681                            Iowa State University        Ames    IA
4  95213                      University of Massachusetts     Amherst    MA
```

Appendix D. Merging University Location and Community Attractiveness Tables

```
import pandas as pd
UniversityLocation = pd.read_excel('/content/UniversityLocation_0.xlsx') CommunityAttractiveness =
pd.read_excel('/content/CommunityAttractiveness_1.xlsx') UniversityAttractiveness_0_df =
pd.merge(UniversityLocation, CommunityAttractiveness, on=['City','State'], how='left')
UniversityAttractiveness_0_df.head()
```

Appendix E. Merging University Attractiveness and Community Demographics Tables

```
UniversityAttractiveness = pd.read_csv('/content/UniversityAttractiveness_0.csv')
CommunityDemographics = pd.read_excel('/content/CommunityDemographics_0.xlsx')
UniversityAttractivenessDemographics_0_df = pd.merge(UniversityAttractiveness,
CommunityDemographics, on=['City','State'], how='left')
UniversityAttractivenessDemographics_0_df.head()
# Convert it to CSV
UniversityAttractivenessDemographics_0_df.to_csv('UniversityAttractivenessDemographics_0.csv',
index=False)
# Download it
from google.colab import files files.download('UniversityAttractivenessDemographics_0.csv')
```

Appendix F. Merging Autm1 and Autm2 Tables

```
Autm1 = pd.read_excel('/content/AUTM_1.xlsx') UniversityAttractivenessDemographics =
pd.read_csv('/content/UniversityAttractivenessDemographics_0.csv')
Autm_2_df = pd.merge(Autm1, UniversityAttractivenessDemographics, on=['[ID]'], how='outer')
Autm_2_df.head()
# Convert it to CSV
Autm_2_df.to_csv('AUTM_2.csv', index=False)
# Download it
from google.colab import files files.download('AUTM_2.csv')
```

Appendix G. Merging Autm3 and Autm4 Tables

```
Autm3 = pd.read_excel('/content/AUTM_3.xlsx')
UniversityAttractivenessDemographics =
pd.read_csv('/content/UniversityAttractivenessDemographics_0.csv')
Autm_4_df = pd.merge(Autm3, UniversityAttractivenessDemographics.drop(columns=['INSTITUTION',
'State']), on=['[ID]'], how='outer')
# Move 'City' to column E (index 4) cols = Autm_4_df.columns.tolist()
if 'City' in cols:
# Remove 'City' and reinsert it at index 4 cols.remove('City')
cols.insert(4, 'City')
Autm_4_df = Autm_4_df[cols]
else:
print("Column 'City' not found in merged DataFrame.")
Autm_4_df.head()
# Create the CSV file:
Autm_4_df.to_csv('AUTM_4.csv', index=False)
#Download it:
from google.colab import files files.download('AUTM_4.csv')
```

Appendix H. Average and Recent Values

```
# Ensure Year is numeric
Autm_4_df['YEAR'] = pd.to_numeric(Autm_4_df['YEAR'], errors='coerce')
# Identify key columns group_col = '[ID]' time_col = 'YEAR'
# Split numeric and non-numeric
numeric_cols = Autm_4_df.select_dtypes(include='number').columns.difference([time_col])
non_numeric_cols = Autm_4_df.select_dtypes(exclude='number').columns.difference([group_col])
numeric_avg = Autm_4_df.groupby(group_col)[numeric_cols].mean()
# The reset_index() method moves the index back to columns to allow further calculations
# If the name of the index is the same as another column, a ValueError will be raised
# to prevent this error, provide a name using the index parameter in the rename method.
numeric_avg = numeric_avg.rename_axis(index={group_col: 'ID_renamed'}).reset_index()
# Sort by Year in descending order, so the most recent comes first
Autm_4_df_sorted = Autm_4_df.sort_values(by=time_col, ascending=False)
# Drop duplicates to keep only the most recent entry per institution
non_numeric_latest = Autm_4_df_sorted.drop_duplicates(subset=group_col)[[group_col] +
non_numeric_cols.tolist()]
cross_sectional_df = pd.merge(numeric_avg, non_numeric_latest, on=group_col, how='left')
# Create the CSV file:
cross_sectional_df.to_csv('AUTM_5.csv', index=False)
#Download it:
from google.colab import files files.download('AUTM_5.csv')
```

Appendix I. Setting Up the Walk Score Scraper in Google Colab

Part 1: Create a New Colab Notebook

1. Go to Google Colab
2. Sign in with your Google account
3. Click on "File" > "New notebook"

Part 2: Install Required Dependencies

The required libraries are mostly pre-installed in Colab, but let's make sure. Run this cell to install or update required packages:

```
!pip install pandas requests beautifulsoup4
```

Part 3: Upload Your CSV File

1. Click on the file icon in the left sidebar
2. Click on the "Upload" button
3. Select your "AUTM_7.xlsx" file (A file which has city and state data) from your computer.
4. Wait for the upload to complete
5. Alternatively, you can use this code to upload from your local machine:

```
from google.colab import files
uploaded = files.upload() # This will prompt you to select and upload a file
```

Part 4: Add the Scraper Code

Create a new code cell and paste your entire Walk Score scraper code:

```
import pandas as pd
import requests
from bs4 import BeautifulSoup
import time
import random
import re

def get_walkscore(city, state):
"""
Fetches the walk score for a given city and state from walkscore.com

Args:
    city (str): Name of the city
    state (str): State abbreviation or full name Returns:
    int or None: Walk score if found, None otherwise """
# Format the city and state for the URL
formatted_city = city.replace(' ', '-')
formatted_state = state.replace(' ', '-')

# The correct URL format for walkscore.com
url = f"https://www.walkscore.com/{formatted_state}/{formatted_city}"
print(f"Trying URL: {url}")

# Add headers to mimic a browser request
```

```
    headers = { 'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36
    (KHTML, like Gecko) Chrome/91.0.4472.124 Safari/537.36', 'Accept-Language': 'en-US,en;q=0.9',
    }

    try:
        # Make the request
        response = requests.get(url, headers=headers) response.raise_for_status() # Raise an exception for
        bad status codes

        # Parse the HTML
        soup = BeautifulSoup(response.text, 'html.parser')

        # Look for the text that contains "has an average Walk Score of X"
        # This targets the text below the bubbles as shown in the screenshot

        text_elements = soup.find_all(string=re.compile(r'has an average Walk Score of \d+'))

        if text_elements:
            # Extract the walk score from the text
            for text in text_elements:
                match = re.search(r'Walk Score of (\d+)', text)
                    if match:
                    return int(match.group(1))

                # If the above didn't work, try finding paragraphs that mention Walk Score
        paragraphs = soup.find_all('p')
        for p in paragraphs:
            if p.text and 'Walk Score' in p.text:
                match = re.search(r'Walk Score of (\d+)', p.text)
                if match:
                    return int(match.group(1))

        # Try another approach - looking for the description text
        description_elements = soup.find_all(lambda tag: tag.name and re.search(r'has an? (?:average
        |)Walk Score of \d+', tag.text))
        for element in description_elements:
            match = re.search(r'Walk Score of (\d+)', element.text)
            if match:
                return int(match.group(1))

        # If we couldn't find the walk score, print a message and return None
        print(f"Walk score not found for {city}, {state}") return None
    except Exception as e:
        print(f"Error fetching walk score for {city}, {state}: {e}") return None
def add_walkscores_to_df(df, city_col='City', state_col='STATE', output_col='WalkScore'):
    """
    Adds walk scores to a dataframe based on city and state columns

    Args:
        df (pd.DataFrame): Dataframe containing city and state information
        city_col (str): Name of the column containing city names
```

```
        state_col (str): Name of the column containing state names
        output_col (str): Name of the column to store walk scores
    Returns:
        pd.DataFrame: Dataframe with added walk score column
    """
    df[output_col] = None

    for index, row in df.iterrows():
        city = row[city_col]
        state = row[state_col]

        # Get the walk score
        walk_score = get_walkscore(city, state) df.at[index, output_col] = walk_score

        # Add random delay to avoid being blocked time.sleep(random.uniform(2, 5))

        # Print progress
            if index % 10 == 0 and index > 0: print(f"Processed {index} records")
    return df

# Modified main function for Colab
# Load your dataframe - make sure the path matches your uploaded file
df = pd.read_excel("AUTM_7.xlsx")
print(f"Successfully loaded Excel file with {len(df)} rows") print(f"Columns: {df.columns.tolist()}")

# Let's look at the first few rows
print(df.head())

# Add walk scores
df = add_walkscores_to_df(df)

print(df)

# Save the result
df.to_csv('AUTM_8.csv', index=False)
print("Saved results to data_with_walkscores.csv")

# To download the results file
from google.colab import files files.download('AUTM_8.csv')
```

Part 5: Run the Code

1. Click the play button next to the cell to run it
2. The script will process each city/state pair in your CSV
3. When complete, it will automatically download the results file to your computer

Expected Fallback:
Missing City/State Data on WalkScore.com
Issue: Some cities or states might not have direct pages or data on WalkScore.com.

- For small towns/suburbs, use the walk score of the nearest major city, but note it as an approximation.
- Look for walk scores of specific neighborhoods or zip codes within that city.

Appendix J. Running Correlations in Google Colab

Open a new notebook in Google Colab and enter the following code:

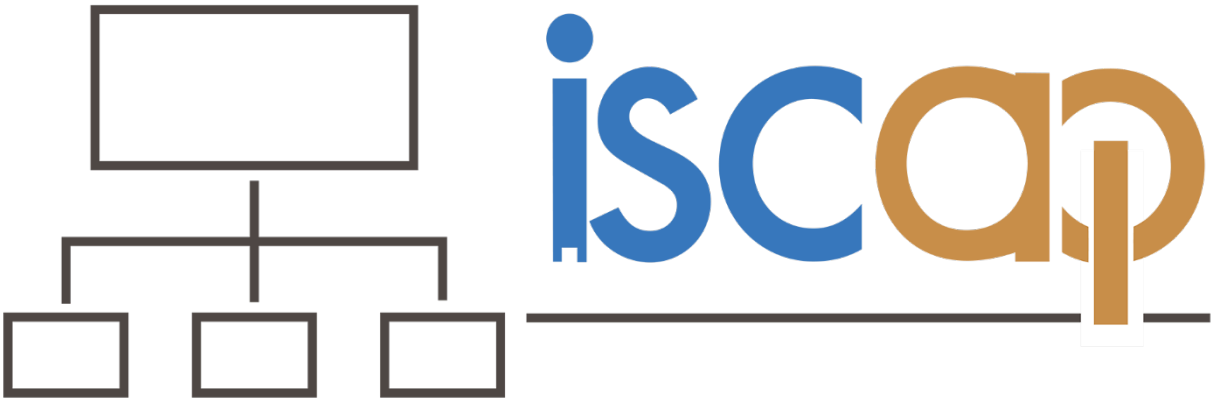
```
# Upload AUTM_8.csv
from google.colab import files uploaded = files.upload()

# import needed packages
import pandas as pd
import numpy as np
Autm_8 = pd.read_csv('/content/AUTM_8.csv') Autm_8.head()

Autm_8.to_csv('Autm_8_df.csv', index=False)

import pandas as pd
from scipy.stats import pearsonr

def calculate_correlations(df, input_column, output_columns):
"""
Calculate the correlation and p-value between an input column and output columns.
Parameters:
- df: The DataFrame containing the data.
- input_column: The input column to correlate with the output columns.
- output_columns: List of output columns to test correlation with the input column.
Returns:
- A DataFrame with the correlation results, styled for display.
"""
# Ensure column names are standardized
df.columns = df.columns.str.strip()  # Remove leading/trailing spaces

# Store results
correlation_results = []

# Compute correlation for each output column
for col in output_columns:
    if col in df.columns and input_column in df.columns:  # Ensure both columns exist
    # Convert columns to numeric, coerce errors to NaN
    df[col] = pd.to_numeric(df[col], errors='coerce')
    df[input_column] = pd.to_numeric(df[input_column], errors='coerce')
    valid_data = df[[col, input_column]].dropna()  # Drop NaNs together
    num_rows = len(valid_data)  # Count valid rows

    if num_rows > 1:  # Ensure at least two rows remain
        corr, p_value = pearsonr(valid_data[col], valid_data[input_column])
        correlation_results.append({
            "Input Column": input_column,
            "Output Column": col,
            "Correlation": corr,
            "P-Value": p_value,
            "Rows Used": num_rows
        })
```

```
# Convert results into DataFrame
correlation_df = pd.DataFrame(correlation_results)

# Sort by correlation in descending order
correlation_df = correlation_df.sort_values(by="Correlation", ascending=False)

# Define function for styling: color rows with p-value < 0.05 green
def style_rows(row):
    color = 'background-color: green' if row['P-Value'] < 0.05 else ''
    return [color] * len(row)

# Apply style
styled_df = correlation_df.style.apply(style_rows, axis=1)
return styled_df

input_column = "WalkScore"
output_columns = ["Tot Pat App Fld", "St-Ups Formed", "Tot Lic/Opt Exe", "Gross Lic Inc"]
styled_correlation_df = calculate_correlations(Autm_8, input_column, output_columns)

styled_correlation_df
```

Appendix K. Enriched Autm9 Dataset

```
import pandas as pd
Autm_8 = pd.read_csv('/content/AUTM_8.csv') UniversityLatLong =
pd.read_excel('/content/UniversityLatLong_0.xlsx') Autm_9_df = pd.merge(Autm_8, UniversityLatLong,
on=['INSTITUTION'], how='left')
Autm_9_df.head()
Autm_9_df.to_csv('AUTM_9.csv', index=False)
from google.colab import files files.download('AUTM_9.csv')
```

INFORMATION SYSTEMS & COMPUTING ACADEMIC PROFESSIONALS



STATEMENT OF PEER REVIEW INTEGRITY

All papers published in the ***Journal of Information Systems Education*** have undergone rigorous peer review. This includes an initial editor screening and double-blind refereeing by three or more expert referees.

ISSN: 2574-3872 (Online) 1055-3096 (Print)