Data Science Capstone Project Proposal: ‘Data.gov’ Trips Data

Trevor McCormick

Tmccormick2@bellarmine.edu

January 12, 2023

Data Science Capstone Project Proposal: TripAdvisor

Executive Summary:

In this project, I am going to be using Trip data from data.gov in order to predict when and how often people are leaving (trips) their homes. This will be done by moving the data through many applications for cleaning, storage, visualization, analysis, and modeling. All of these are necessary for the end goal of making a useful and working program. Most of the project will be done in Python using the Pandas package. The storage method will be SQLite due to that being what I am semi-familiar with. Then, the visualization will be done in Tableau because I am both familiar with and enjoy using it. Finally, everything will go back into Python for the analysis and modelling. While these are my ideas as of writing this, I am aware and sure that many things are likely to change through the process of trial and error.

Project Idea:

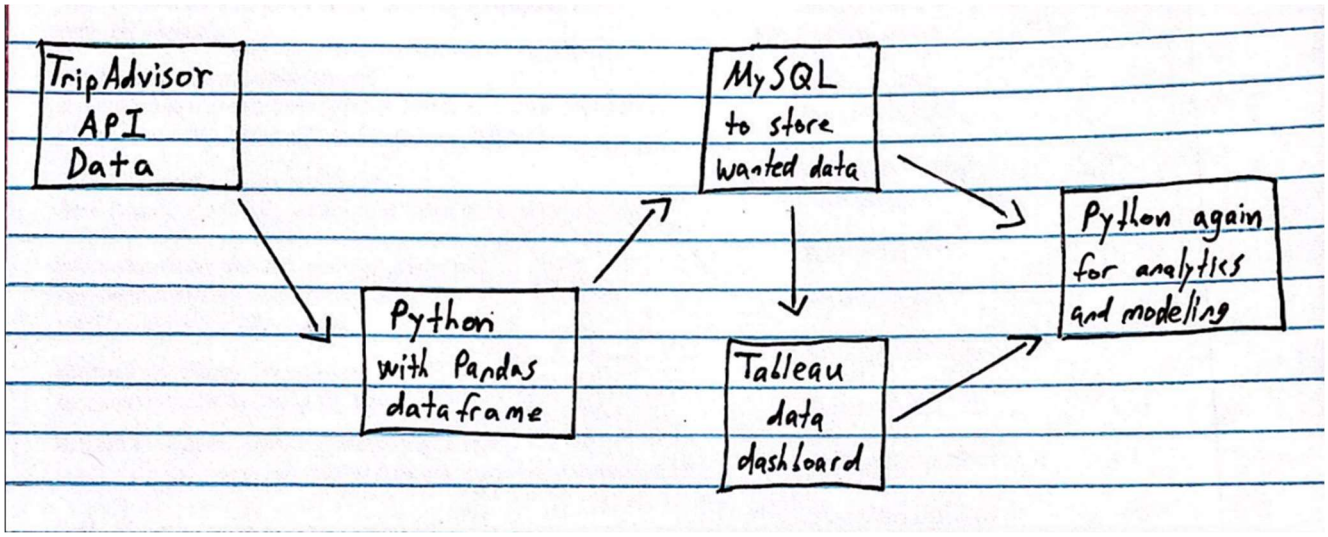
The topic that I have decided to pursue for the data science capstone project is a dive into the open data from the data.gov website which has tons of readily available data for use. This project will have many steps that will all cumulate to achieve a final goal at the end. The trips data gives users (like myself) access to numerous aspects of information regarding trips taken per state on given dates. Some of the columns in this data include dates, states, many columns of how many trips were taken, and a few more. As you can see, all of these options provide potential ideas and goals that could be used for a data science project. One of the hardest decisions is deciding what aspect to work on and where to execute all of the ideas. After considering all of the possible options that I could pursue for my project, I had to pick the one that seemed both the most interesting to me and actually doable for a class project. The end goal for this project as of right now (it may change if the need arises) is to use the trips data to predict how often people will be leaving their homes on given dates. Ideally, you would be able to look at a certain location (state) and the program will tell you how many people and how often people are leaving at different points of the year (dates). The ways the program could decide this would be by looking at all of the columns and data given in this dataset. These are my ideas for the project at this time.

Background:

My main drive for picking the trips data as my project area is due to always being interested in travel and understanding how people function and move (another reason is because of my initial idea falling through). So, I figured a project that enables me to reflect on and build off of that would be a perfect fit. A project should always be focused on something that you are passionate about or at the very least, interested in. My data for this project is going to come from the trips data that is available for free use from data.gov. Data.gov always has constant new data and datasets flowing through it and being added to it. This means that their information and options are constantly updating and growing. This makes it a perfect choice for this project. New data means more current and accurate information that we can work with. Data.gov has this description for their Trips dataset, "How many people are staying at home? How far are people traveling when they don't stay home? Which states and counties have more people taking trips? The Bureau of Transportation Statistics (BTS) now provides answers to those questions through our new mobility statistics" (Data.gov)

Preliminary Architecture:

The architecture of a data science project like this includes the data, how to collect the data, where to store the data, cleaning the data, and finally, using the data. As stated many times above, the data for this project will come from the data.gov trips data that is available for use in projects like this. As of right now, I am planning to store the data in SQLite. SQLite is often used as a data store, and it is one that I am at least somewhat familiar with. It should do the job for what I need it to accomplish do to being a local data store. I will use Python (or maybe R) for the data cleaning and analysis. These are the two programs that I am the most familiar with as of now. Python and R are both excellent recourses for what I need to do with this project. For the data dashboard I am planning on using Tableau since it is something that I am familiar with due to using it in a course my junior year. It seems like it would be a good fit for what I need, bit if it doesn't end up working properly there are other options out there such as Power BI. Considering these things, Python and Tableau/Power BI will most likely be the most important programs that I will be using during this project. However, every application is necessary to reach the final goal of completion. All of this can be seen in the graph below:



Modeling:

The modeling section of this project is the main area where I am not quite sure which idea to settle on. There are so many options for what to do here and so many to choose from. I do know that with what my project goal is, I am going to want to use predictive analytics. Predictive analytics uses the data to learn and make predictions based on past data. I think the main methods that I'll be using are Scikit-learn and decision trees. However, these may need to chance once I get further in the project and see how thing work and inevitably don't work.

Conclusion:

In conclusion, I will be using the trips data available from data.gov to conduct and data science project which will let us see how often, when, and how much people are leaving their homes. The data which comes from this will then be moved into Python for cleaning and transfer into a data frame. Next it will go to SQLite for storage. Then I will use Tableau as a dashboard for visualization. Finally, I will go back to Python for analysis and modeling. I am excited to begin working on this project and seeing how close things end up going to plan. Even though some things with most likely need to change.

References

"Trips by Distance." *Catalog*, Publisher Bureau of Transportation Statistics, 1 Feb. 2023, https://catalog.data.gov/dataset/trips-by-distance.