Daily Trips: How often do Americans Leave Home?

April 25th, 2023

Trevor McCormick

Trevormccormick01@gmail.com

Executive Summary

In this project, I used the Trips dataset from data.gov in order to predict when and how often people are leaving their homes. This was 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 was done in Python using Pandas and other various packages that are discussed more later on. The storage method used was SQLite due to being semi-familiar with SQL and similar applications. Then, the visualization was done in PowerBI because I am both familiar with and enjoy using it, plus it connects to SQLite smoother than Tableau does. Finally, everything went back to Python for analysis and modeling. The analysis was a fairly standard exploratory data analysis which will be explored later in this paper. The final modeling section of the project used Random Forest methods along with Multiple Linear Regression. In the end, I concluded that Linear Regression was the best fitting model for this dataset.

Introduction

This paper is the final part of the semester long final project for the Bellarmine University Data Science Major. Initially, I was going to do a dive into the TripAdvisor API that is free of use. For the first week or so of this project, that is exactly what I did. Unfortunately, it did not end up containing the information that I needed for this kind of analysis. Due to this, I had to make a change of data source.

For this project, I decided to do a dive into open data from the data.gov website, which has tons of readily available data for use, in order to find a dataset that would work well for this project. The data that I ended up selecting from data.gov is the trips dataset. The trips data gives users 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. 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 was to find a predictive model that would best predict how often people will be leaving their homes on given dates, based on the other variables and information in the dataset. The way the program could decide this would be by looking at all of the columns and data given in this dataset using various predictive models in Python.

Project Details

My data for this project came 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 can be worked 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 now provides answers to those questions through our new mobility statistics" (Data.gov)

The architecture of a data science project includes the data, how to collect the data, where to store the data, cleaning the data, and finally, using the data. After gathering the data from data.gov, the next step was to store the data in SQLite. SQLite is often used as a data store, and it is one that I was at least somewhat familiar with before starting this project. It ended up doing the job perfectly for what I need it to accomplish, due to being a local data store. I coded the transfer from an excel file to SQLite in Python. Once the SQLite database was created, I installed an application called DB Browser which enables SQLite data to be seen and analyzed.

I then used python for the data cleaning and analysis. Python is an excellent recourse for what I need to do with this project. In python, I used multiple packages for the analysis of the trips data. Pandas was the main package used due to how useful it is for creating data frames for the datasets. This is also where the exploratory data analysis (EDA) mainly took place. In the EDA, I discussed the intricate details of the trips dataset. Things like data types, missing data, summary statistics, correlations, and various graphs. The details of this EDA can be found in the exploratory data analysis tab of the final project repository.

Next up in the project comes the data dashboard. For the dashboard, I originally planned on using Tableau, but Microsoft Power BI ended up working much better for me. I was able to connect Power BI with SQLite far easier and smoother than I was able to with Tableau. It was also a good chance to familiarize myself with Power BI due to it being something new to me. This application enabled me to make multiple graphs that showed various aspects of the dataset. They showed correlations, relationships, amounts, and more.

Lastly, came the final part of this project which is the predictive analysis code. In order to execute predictive analysis on the trips dataset, all of the previous steps had to first be completed. The cleaning, analysis, and visualization steps were all extremely important for being able to make a proper model for prediction. In this step, I tried many different methods. Some which worked, and most which didn't. The two methods that I ended up getting to work smoothly and properly were Random Forest and Multiple Linear Regression. The Random Forest Model yielded a predictive score of 98%, which is suspiciously high. The Linear Regression Model yielded a score of 94%, which is still high, but less likely to be flawed than 98%. Due to this, I think the Multiple Linear Regression Model is the best model to use for this dataset. These findings were then added to the poster presentation that was presented at the celebration of student research at Bellarmine University.

Reflection

During this project, I had many things that went well and many that did not. The nature of a data science project is trial and error, so this was at least expected to happen. We will first start with the bad in order to get that out of the way. The main error of this project occurred at the very first step. When I was choosing what topic to pursue for this project, I picked one that ended up not being viable for it. Thankfully, it was only around two weeks' worth of work on the failed topic, but it was still a fair amount of lost time and effort. I was fortunate to be able to find a replacement topic rather quickly, which enabled me to get right back to working on it without wasting more time than I already had. In reflection, realizing that my original idea wasn't going to work was a fairly stressful time, I'm just glad that it ended up working for the better. There were other aspects of trial and error in this project, but I cannot really count those as negatives due to them being fully expected before even starting.

On the positive side, there was a lot that went very well with this project overall. Once the second idea was found and I decided that it was going to be viable, it was mostly uphill and positive from there. The next step, which was the data collection and storage system, went extremely smoothly. The discovery that SQLite was going to work perfectly for my data was a breath of fresh air, and the importation to it went very well. Once the code for that was written in Python, it made that part of the project as well as the next parts very smooth. The EDA took a significantly longer amount of time than the storage step, but it was still a fairly simple and smooth step. Most of the code I used and wrote worked well which made it a quite enjoyable step, and writing the write-up section of the EDA was easy after everything in Python was working as it should. The visualization step also worked extremely

well with how SQLite was able to connect with Power BI. That might have been the most enjoyable step overall. Lastly, the modeling section took a bit more time to get working right and accurately, but it was still a net positive experience overall.

If I had to change anything, it would be to thoroughly research and prepare for each step of the project before I actually got to them. With the due dates for each section being spread out, it was nice to be able to take them one step at a time, but I didn't prepare for each step very far in advance which led to some steps being harder or taking longer than they probably should have. Although, this is just a minor inconvenience overall. The greatest impact this made was on the modelling step. It took me a while to refamiliarize myself with all of the various models we had used in the past. If I could go back and prepare more for this, that would be the main thing I would change.

If there was one thing that I would recommend to others doing a similar project, it would be to make sure your dataset is viable before you put too much work into it. As stated many times above, this was easily the biggest misstep of this whole project for me. Avoiding this would be the best advice I could give. Time is essential, and it is important not to waste much of it if at all possible.

It is not putting it lightly to say that I learned so much from doing this entire project. Doing the entire thing from scratch was something that I had never done before, but I am very thankful that I had to do it. Gathering the data by myself was new to me for a project like this, and I think it was really beneficial that I learned how to do so. Even if I did have to do it twice. I think that doing the project in separate steps of various difficulty was a good way to show how real-world data science projects truly work. Nothing comes easily, and everything has to be executed precisely and in certain ways in order to be done correctly. While the project definitely was a lot of work and trial and error at times, I am very glad that I did it. I think I gained a lot of experience from completing it.

Conclusions and Future Work

In summary, this was a four-month project that was completed in many different steps. As previously stated, much was learned over the course of the entire project, and there are things I did that worked well and things that didn't. The main conclusion of the final project came from the predictive modelling section, which was the last step of the overall project. Before I got to this, many steps were first completed such as the importation to SQLite,

the exploratory data analysis, the dashboard, and more. The predictive analysis coding was probably the toughest part of the project overall, but even that was accomplished in the end. As discussed earlier, I am happy with what I learned and accomplished through this final project. If I was to return to this project in the future, I would like to look at the updated data to see how things have changed. The data will change over the years, and I expect the numbers to increase, but it would be nice to see exactly how things change and update. I'm sure the models will yield similar yet slightly different results in the future.