Trips Data
Exploratory Analysis

Trevor McCormick - tmccormick2@bellarmine.edu

I. INTRODUCTION

This dataset, called Trips Data, is a dataset containing many different variables and data about the number of trips people take from their homes on a given date. It gets much more descriptive than that, but we will dive more into that later. The Trips dataset was found and pulled from data.gov which is a website for open data available from the United States government. Data of all different kinds and sizes can be found from this website, which is why I turned to it in order to find my dataset after my initial idea for the project was no longer viable. This dataset was not my original dataset or idea for this project, but it ended up being the one that worked the most and had the most usable/useful information available. In this analysis, I will be looking at many aspects of the data such as statistics, frequencies, proportions, correlations, graphs, plots, and more. I expect to be able to pull out a decent amount of information through this analysis.

II. TRIPS DATA DESCRIPTION

The Trips dataset contains 45,267 samples with 20 columns with various data types. A complete listing of these columns and samples is shown in **Table 1**. This is by far the largest dataset that I have ever worked with, which is both exciting and challenging. A large dataset enables you to learn things at a very specific level due to all of the information that can be pulled and learned from it. The data types are indicated by two things (nominal, ordinal, interval, or ratio) and the Pandas data type. In this dataset, there happens to be 0 missing data present. This is not normally the case, but this dataset happened to be fully complete. The full list everything discussed here can be found below:

Table 1: Data Types and Missing Data

Variable Name	*Data Type*	*Missing Data (%)*
Level	Nominal/Object	0%
Date	DateTime64 [ns]	0%
State FIPS	Nominal/Int64	0%
State Postal Code	Nominal/Object	0%
Population Staying at Home	Ratio/Int64	0%
Population Not Staying at Home	Ratio/Int64	0%
Number of Trips	Ratio/Int64	0%
Number of Trips <1	Ratio/Int64	0%
Number of Trips 1-3	Ratio/Int64	0%
Number of Trips 3-5	Ratio/Int64	0%
Number of Trips 5-10	Ratio/Int64	0%
Number of Trips 10-25	Ratio/Int64	0%
Number of Trips 25-50	Ratio/Int64	0%

Number of Trips 50-100	Ratio/Int64	0%
Number of Trips 100-250	Ratio/Int64	0%
Number of Trips 250-500	Ratio/Int64	0%
Number of Trips >=500	Ratio/Int64	0%
Row ID	Nominal/Object	0%
Week	Nominal/Int64	0%
Month	Nominal/Int64	0%

III. Data Set Summary Statistics

The following two tables show the various summary statistics that describe the Trips data set. **Table 2** shows statistics such as the mean, standard deviations, and min/max. **Table 3** shows the proportions for the categorical variables in the dataset. Finally, **Table 4** is the correlation table which shows how the variables are correlated and related to each other.

Table 2: Summary Statistics for Trips Data

Variable Name	*Count*	*Mean*	*Standard Deviation*	*Min*	*25th*	*50th*	*75th*	*Max*
Population Staying at Home	45,267	1.428632e+06	1,735,886.64	571	376,033.5	900,199.0	1,738,435	15,719,267
Population Not Staying at Home	45,267	4.976276e+06	5,619,811.18	4,618	1,366,908	3,504,438	5,785,856.5	33,675,879
Number of Trips	45,267	2.317937e+07	26,085,108.91	17,543	5,799,036.5	15,685,029	28,944,518	206864081
Number of Trips <1	45,267	5.752116e+06	6,818,360.27	2,166	1,453,253.5	3,670,768	7,033,357.5	58535102
Number of Trips 1-3	45,267	5.777840e+06	6,555,675.74	2,168	1,485,978	3,904,398	7,214,650.5	57222317
Number of Trips 3-5	45,267	2.839236e+06	3,138,751.07	948	725,626	1,934,135	3,575,975	26937995
Number of Trips 5-10	45,267	3.584870e+06	3,972,561.82	1,558	893,777.5	2,412,702	4,561,396	30720927
Number of Trips 10-25	45,267	3.504542e+06	3,904,027.99	2,768	891,253	2,379,404	4,623,131	30774902
Number of Trips 25-50	45,267	1.134852e+06	1,267,035.75	656	306,717	792,821	1,443,285.5	9756023
Number of Trips 50-100	45,267	3.713002e+05	405,818.24	530	120,522.5	274,656	446,826.5	3330076
Number of Trips 100-250	45,267	1.526553e+05	165,613.63	336	47,933.5	108,868	190,219	1863349
Number of Trips 250-500	45,267	3.391814e+04	38,632.73	12	8,769	23,063	42,797.5	365897
Number of Trips >=500	45,267	2.803428e+04	42,262.65	0	5,298	13,540	33,016	606342

Table 3: Proportions and Frequencies

State FIPS		
	Frequency	Proportion
1	887	0.019595
2	887	0.019595
3	887	0.019595
4	887	0.019595
5	887	0.019595
6	887	0.019595
7	887	0.019595
8	887	0.019595
9	887	0.019595
10	887	0.019595
11	887	0.019595
12	887	0.019595
13	887	0.019595
14	887	0.019595
15	887	0.019595
16	887	0.019595
17	887	0.019595
18	887	0.019595
19	887	0.019595
20	887	0.019595
21	887	0.019595
22	887	0.019595
23	887	0.019595
24	887	0.019595
25	887	0.019595
26	887	0.019595
27	887	0.019595
28	887	0.019595
29	887	0.019595
30	887	0.019595
31	887	0.019595
32	887	0.019595
33	887	0.019595
34	887	0.019595
35	887	0.019595
36	887	0.019595
37	887	0.019595
38	887	0.019595
39	887	0.019595

40	887	0.019595
41	887	0.019595
42	887	0.019595
43	887	0.019595
44	887	0.019595
45	887	0.019595
46	887	0.019595
47	887	0.019595
48	887	0.019595
49	887	0.019595
50	887	0.019595
51	888	0.019595
State Postal Code		
	Frequency	**Proportion**
AK	887	0.019595
AL	887	0.019595
AR	887	0.019595
AZ	887	0.019595
CA	887	0.019595
CO	887	0.019595
CT	887	0.019595
DE	887	0.019595
FL	887	0.019595
GA	887	0.019595
HI	887	0.019595
IA	887	0.019595
ID	887	0.019595
IL	887	0.019595
IN	887	0.019595
KS	887	0.019595
KY	887	0.019595
LA	887	0.019595
MA	887	0.019595
MD	887	0.019595
ME	887	0.019595
MI	887	0.019595
MN	887	0.019595
MO	887	0.019595
MS	887	0.019595
MT	887	0.019595
NC	887	0.019595

ND	887	0.019595
NE	887	0.019595
NH	887	0.019595
NJ	887	0.019595
NM	887	0.019595
NV	887	0.019595
NY	887	0.019595
OH	887	0.019595
OK	887	0.019595
OR	887	0.019595
PA	887	0.019595
RI	887	0.019595
SC	887	0.019595
SD	887	0.019595
TN	887	0.019595
TX	887	0.019595
UT	887	0.019595
VA	887	0.019595
VT	887	0.019595
WA	887	0.019595
WI	887	0.019595
WV	887	0.019595
WY	888	0.019617
Week		
	Frequency	**Proportion**
1	561	0.012393
2	1070	0.02366
3	1070	0.02366
4	1070	0.02366
5	1070	0.02366
6	1070	0.02366
7	1070	0.02366
8	1070	0.02366
9	1070	0.02366
10	1070	0.02366
11	1070	0.02366
12	1070	0.02366
13	1070	0.02366
14	1070	0.02366
15	1070	0.02366
16	1070	0.02366
17	1070	0.02366

18	1070	0.02366
19	1070	0.02366
20	1070	0.02366
21	1070	0.02366
22	1070	0.02366
23	744	0.016436
24	714	0.015773
25	714	0.015773
26	714	0.015773
27	714	0.015773
28	714	0.015773
29	714	0.015773
30	714	0.015773
31	714	0.015773
32	714	0.015773
33	714	0.015773
34	714	0.015773
35	714	0.015773
36	714	0.015773
37	714	0.015773
38	714	0.015773
39	714	0.015773
40	714	0.015773
41	714	0.015773
42	714	0.015773
43	714	0.015773
44	714	0.015773
45	714	0.015773
46	714	0.015773
47	714	0.015773
48	714	0.015773
49	714	0.015773
50	714	0.015773
51	714	0.015773
52	408	0.009013
Month		
	Frequency	**Proportion**
1	4743	0.104778
2	4335	0.095765
3	4743	0.104778
4	4590	0.101398
5	4743	0.104778

6	3345	0.073895
7	3162	0.069852
8	3162	0.069852
9	3060	0.067599
10	3162	0.069852
11	3060	0.067599
12	3162	0.069852

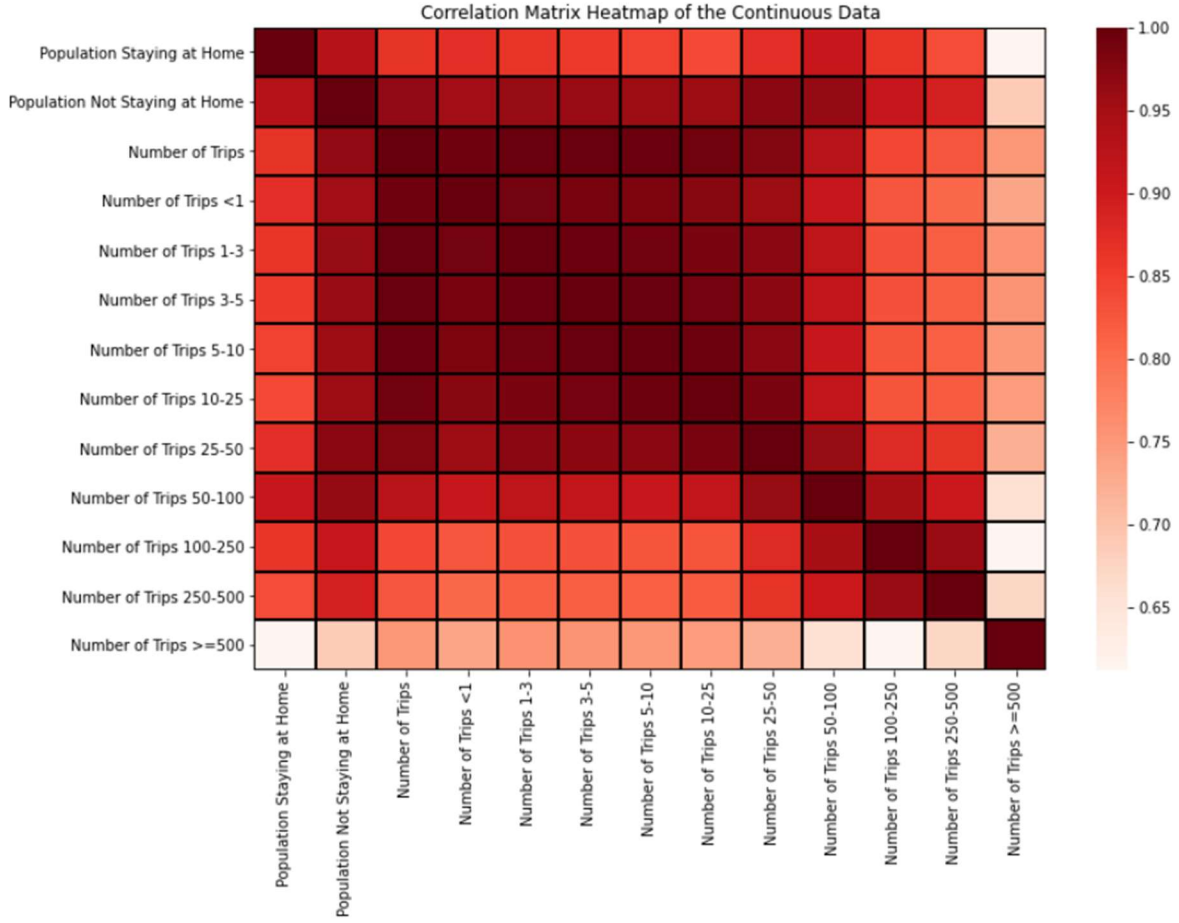
Table 4: Proportions and Frequencies for Categorical Columns

	count	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
State FIPS	45267	28.94603	15.68405	1	16	29	42	56
Week	45267	23.39433	15.0149	0	11	21	36	52
Month	45267	5.920008	3.446575	1	3	5	9	12

Table 5: Correlation Table

	Population Staying at Home	Population Not Staying at Home	Number of Trips	Number of Trips <1	Number of Trips 1-3	Number of Trips 3-5	Number of Trips 5-10	Number of Trips 10-25	Number of Trips 25-50	Number of Trips 50-100	Number of Trips 100-250	Number of Trips 250-500	Number of Trips >=500
Population Staying at Home	1.000000	0.930640	0.866426	0.870476	0.862630	0.856471	0.848241	0.840316	0.869959	0.908069	0.862407	0.837184	0.614280
Population Not Staying at Home	0.930640	1.000000	0.966007	0.954270	0.960839	0.960157	0.957780	0.957357	0.972361	0.962598	0.909205	0.891764	0.689142
Number of Trips	0.866426	0.966007	1.000000	0.992446	0.997492	0.997691	0.996568	0.991525	0.978067	0.924869	0.841624	0.826423	0.751561
Number of Trips <1	0.870476	0.954270	0.992446	1.000000	0.989457	0.987018	0.983245	0.972763	0.957603	0.909241	0.826884	0.806172	0.734393
Number of Trips 1-3	0.862630	0.960839	0.997492	0.989457	1.000000	0.996233	0.991904	0.984596	0.971696	0.920524	0.832802	0.817790	0.758633
Number of Trips 3-5	0.856471	0.960157	0.997691	0.987018	0.996233	1.000000	0.997675	0.988938	0.970066	0.913359	0.832318	0.817867	0.755333
Number of Trips 5-10	0.848241	0.957780	0.996568	0.983245	0.991904	0.997675	1.000000	0.994674	0.972514	0.909192	0.827901	0.816195	0.752626
Number of Trips 10-25	0.840316	0.957357	0.991525	0.972763	0.984596	0.988938	0.994674	1.000000	0.983363	0.914355	0.827401	0.819883	0.745478
Number of Trips 25-50	0.869959	0.972361	0.978067	0.957603	0.971696	0.970066	0.972514	0.983363	1.000000	0.960916	0.877843	0.865236	0.722601
Number of Trips 50-100	0.908069	0.962598	0.924869	0.909241	0.920524	0.913359	0.909192	0.914355	0.960916	1.000000	0.946176	0.903956	0.659594
Number of Trips 100-250	0.862407	0.909205	0.841624	0.826884	0.832802	0.832318	0.827901	0.827401	0.877843	0.946176	1.000000	0.959441	0.611861
Number of Trips 250-500	0.837184	0.891764	0.826423	0.806172	0.817790	0.817867	0.816195	0.819883	0.865236	0.903956	0.959441	1.000000	0.672852
Number of Trips >=500	0.614280	0.689142	0.751561	0.734393	0.758633	0.755333	0.752626	0.745478	0.722601	0.659594	0.611861	0.672852	1.000000

Table 6: Correlation Matrix



IV. DATA SET GRAPHICAL EXPLORATION

In this section, I will be generating graphs and plots to attempt to explore the data in this way, which is different than the methods explored in the sections above. This section will have a couple of different kinds of charts, which are listed below (A-D). Due to the Trips dataset being so large, many of these graphs and plots have been generated in the Power BI application, which is able to run graphs with large datasets much faster and more efficiently than Python (more on that in the final section).

A. Distributions

B. Scatter Plots / Pairwise Plots

C. Bar Charts

D. Decomposition Chart

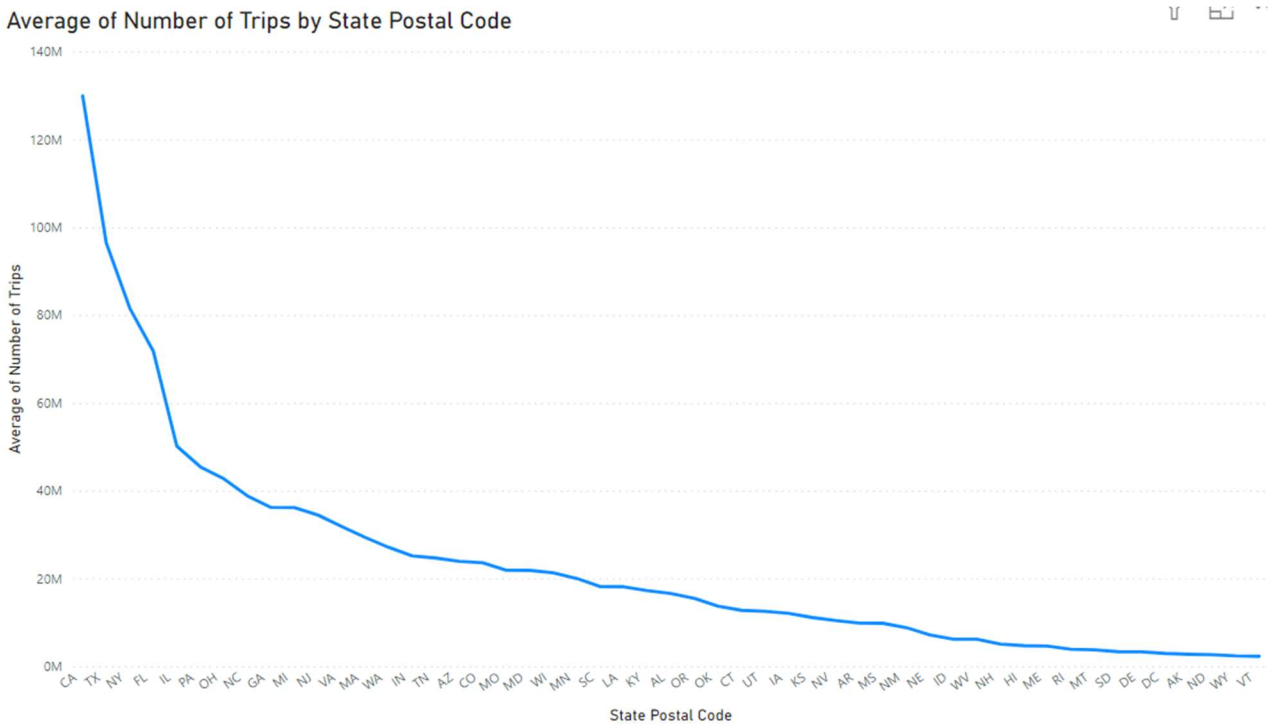


Figure 1: Comparison of AVG Number of Trips/State Postal Code from Trips dataset

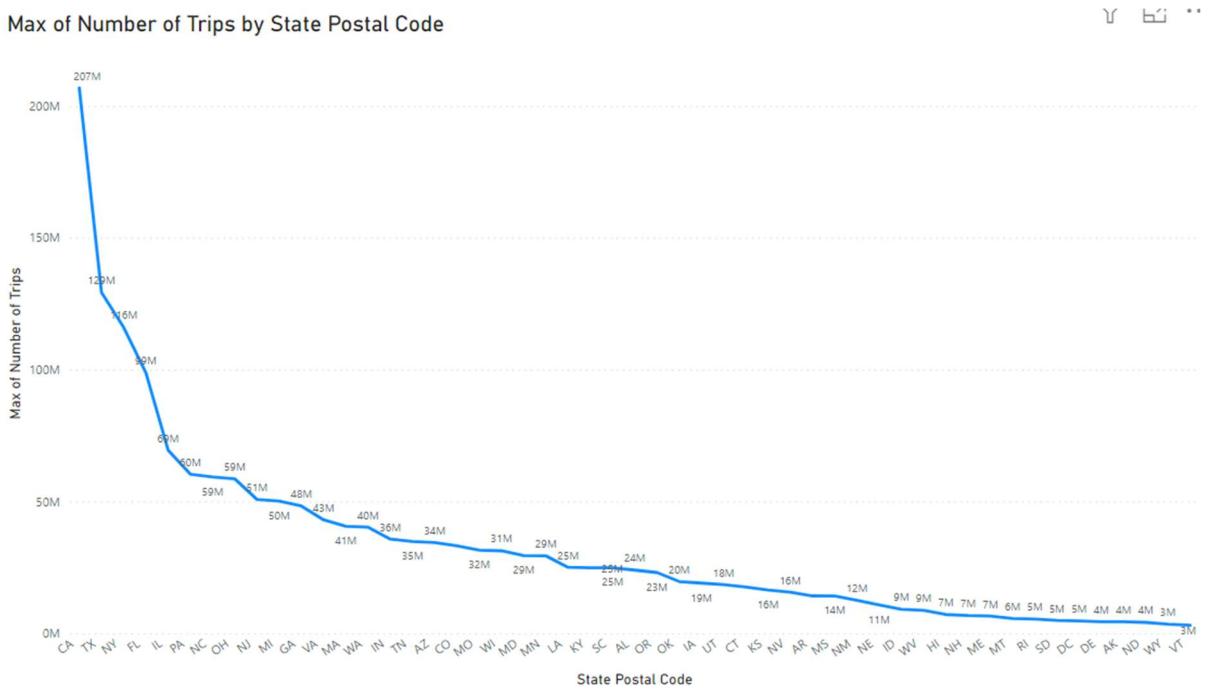


Figure 2: Comparison of MAX Number of Trips/State Postal Code from Trips dataset

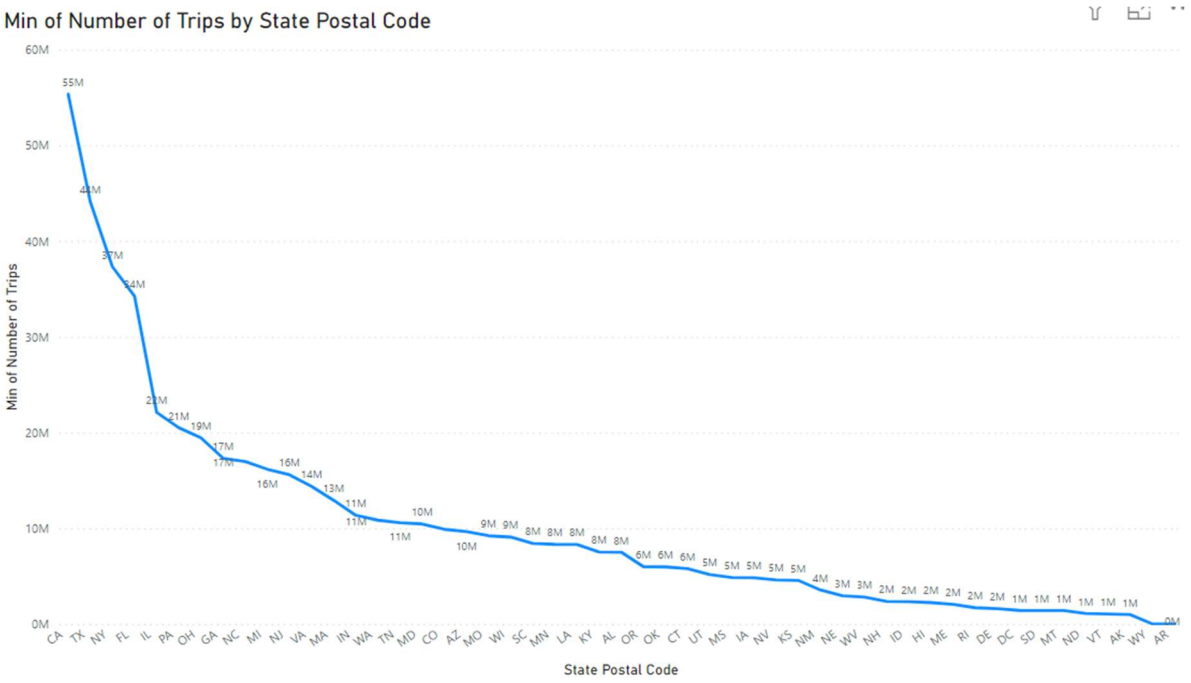


Figure 3: Comparison of MIN Number of Trips/State Postal Code from Trips dataset

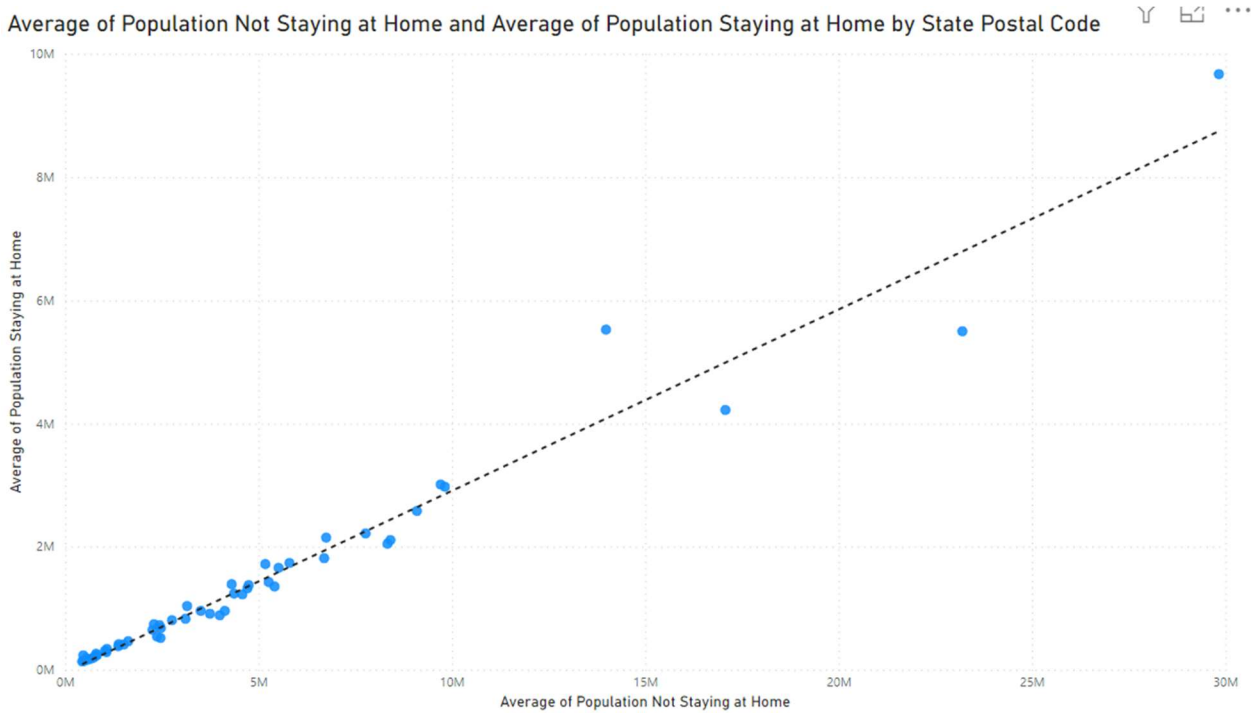


Figure 4: Scatterplot of AVG Population Staying Home against AVG Population Not Staying Home

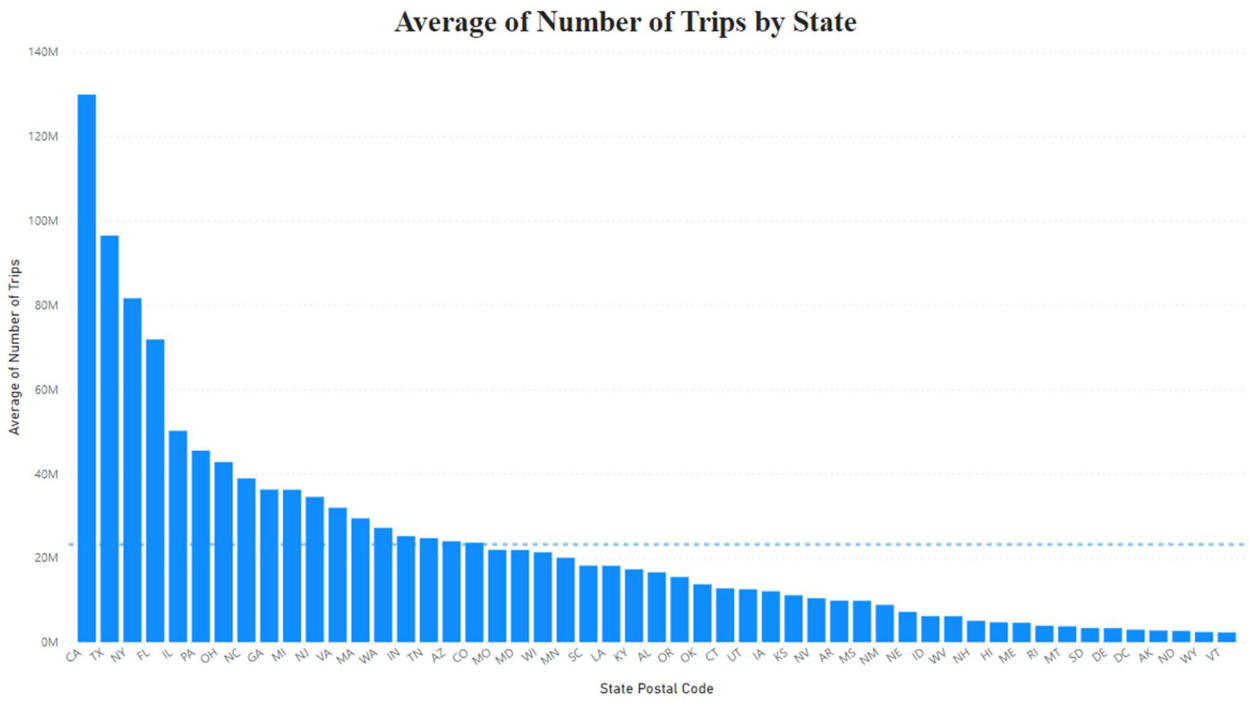


Figure 5: Bar Graph of AVG Number of Trips by State Postal Code

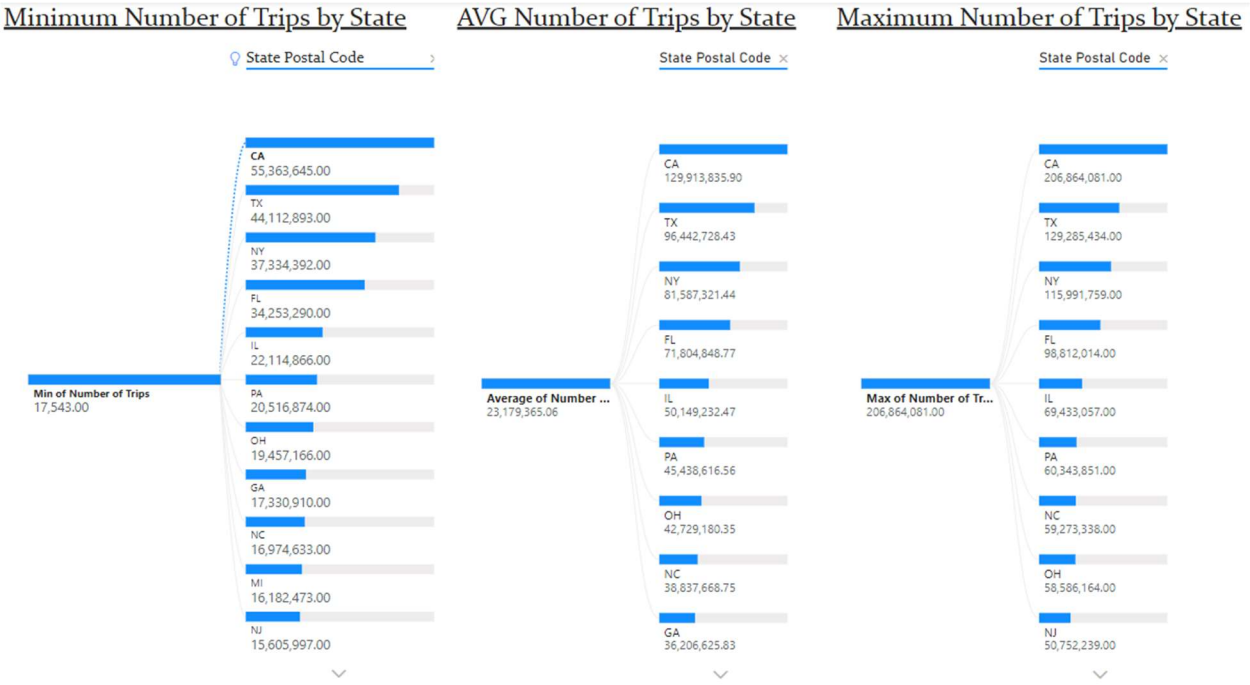


Figure 6: Decomposition Tree of MIN, AVG, and MAX values of Number of Trips by State Postal Code

V. SUMMARY OF FINDINGS

Throughout this analysis, I have had one main thought the entire time. This is a very large dataset. While it is manageable and usable for what this project demands, it does make certain aspects of this project/analysis more difficult. The main difficulty being my hardware having trouble loading graphs for columns with over 45,000 rows. I found that oftentimes Python struggles with this as well. Given the struggles of Python with the data set, I found that using the dashboard Power BI, I was able to generate graphs much more rapidly than with Python. While that was not the goal, it was how it had to be with my limited hardware and time constraints. I discovered that with the amount of data and columns in this dataset, using Power BI was actually very beneficial in the graphical section. Other than the graphs, in the earlier sections, I discovered that some of these columns are much more useful for most aspects of analysis. Some of the columns in the dataset go into very specific detail, which is great for the machine learning aspect, but can make the analysis more complex than it needs to be. Overall, I think this is a complex yet still usable and viable for analysis and machine learning.