Members:

Aquino, Kerwin Dominique Aquino

Lagazo, John Louise

Manlapig, Ralph Miguel

Tadeo, Lorenz Christian

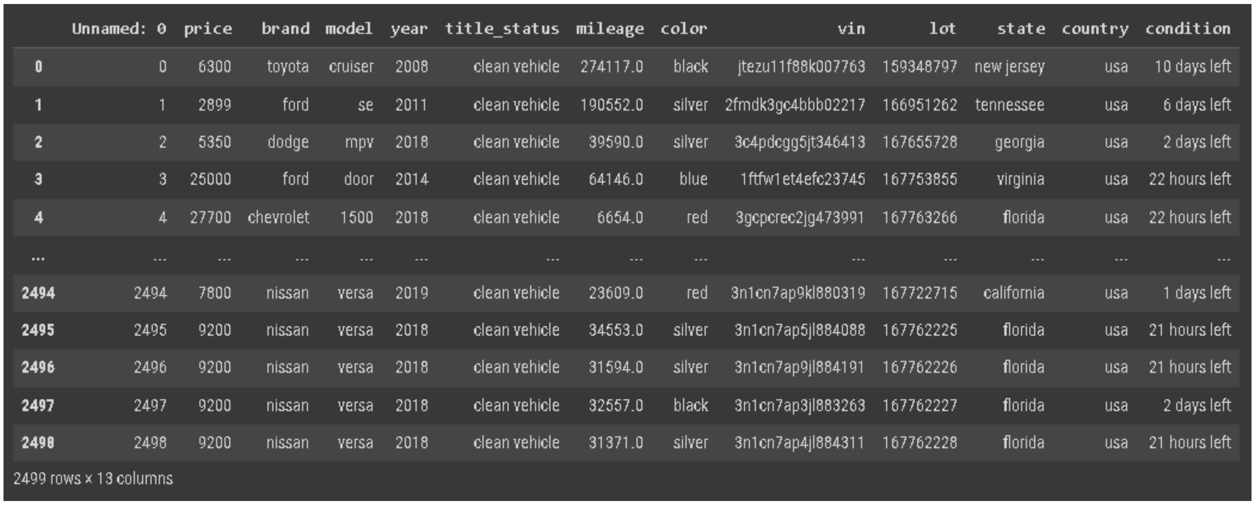
Colaboratory File Link:

https://colab.research.google.com/drive/1AiXAkrv5gLUuziVLgboz6IB60OWa9bow?usp=sharing

Introduction

This project was done to showcase the group's ability to use different visualization techniques in order to analyze and explain a dataset. The group initially chose a dataset from Kaggle about cars in the USA that contained a car's information such as its sale price, registration year, brand and model. Since this initial dataset is not able to provide all the necessary information for the visualization techniques that will be used, additional datasets related to the initial dataset were used for those techniques.

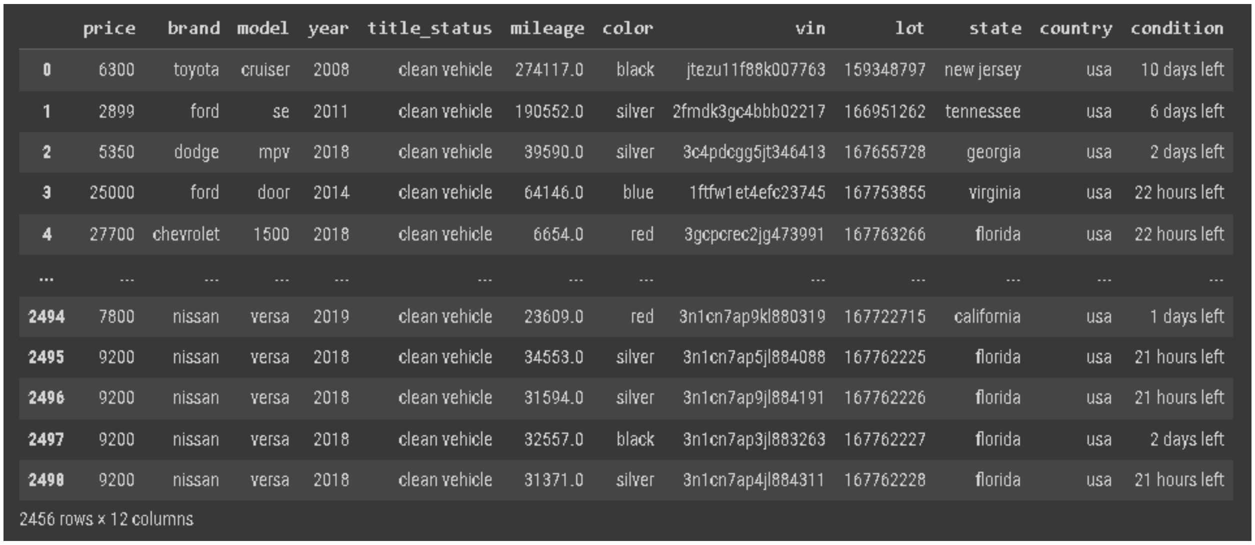
The initial dataset contained 13 features that described 2499 records in the USA, after removing cars that had a price of 0 (since it skews our visualizations to some degree), and a column that served as the index of the dataset, the group was left with a dataset that contained 12 features with 2456 records.



	Unnamed: 0	price	brand	model	year	title_status	mileage	color	vin	lot	state	country	condition
0	0	6300	toyota	cruiser	2008	clean vehicle	274117.0	black	jtezu11f88k007763	159348797	new jersey	usa	10 days left
1	1	2899	ford	se	2011	clean vehicle	190552.0	silver	2fmdk3gc4bbb02217	166951262	tennessee	usa	6 days left
2	2	5350	dodge	mpv	2018	clean vehicle	39590.0	silver	3c4pdcgg5jt346413	167655728	georgia	usa	2 days left
3	3	25000	ford	door	2014	clean vehicle	64146.0	blue	1ftfw1et4efc23745	167753855	virginia	usa	22 hours left
4	4	27700	chevrolet	1500	2018	clean vehicle	6654.0	red	3gcpcrec2jg473991	167763266	florida	usa	22 hours left
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2494	2494	7800	nissan	versa	2019	clean vehicle	23609.0	red	3n1cn7ap9kl880319	167722715	california	usa	1 days left
2495	2495	9200	nissan	versa	2018	clean vehicle	34553.0	silver	3n1cn7ap5jl884088	167762225	florida	usa	21 hours left
2496	2496	9200	nissan	versa	2018	clean vehicle	31594.0	silver	3n1cn7ap9jl884191	167762226	florida	usa	21 hours left
2497	2497	9200	nissan	versa	2018	clean vehicle	32557.0	black	3n1cn7ap3jl883263	167762227	florida	usa	2 days left
2498	2498	9200	nissan	versa	2018	clean vehicle	31371.0	silver	3n1cn7ap4jl884311	167762228	florida	usa	21 hours left

2499 rows × 13 columns

Figure 1: Initial Dataset Before Editing



	price	brand	model	year	title_status	mileage	color	vin	lot	state	country	condition
0	6300	toyota	cruiser	2008	clean vehicle	274117.0	black	jtezu11f88k007763	159348797	new jersey	usa	10 days left
1	2899	ford	se	2011	clean vehicle	190552.0	silver	2fmdk3gc4bbb02217	166951262	tennessee	usa	6 days left
2	5350	dodge	mpv	2018	clean vehicle	39590.0	silver	3c4pdcgg5jt346413	167655728	georgia	usa	2 days left
3	25000	ford	door	2014	clean vehicle	64146.0	blue	1ftfw1et4efc23745	167753855	virginia	usa	22 hours left
4	27700	chevrolet	1500	2018	clean vehicle	6654.0	red	3gcpcrec2jg473991	167763266	florida	usa	22 hours left
...	...	...	...	...	...	...	...	...	...	...	...	...
2494	7800	nissan	versa	2019	clean vehicle	23609.0	red	3n1cn7ap9kl880319	167722715	california	usa	1 days left
2495	9200	nissan	versa	2018	clean vehicle	34553.0	silver	3n1cn7ap5jl884088	167762225	florida	usa	21 hours left
2496	9200	nissan	versa	2018	clean vehicle	31594.0	silver	3n1cn7ap9jl884191	167762226	florida	usa	21 hours left
2497	9200	nissan	versa	2018	clean vehicle	32557.0	black	3n1cn7ap3jl883263	167762227	florida	usa	2 days left
2498	9200	nissan	versa	2018	clean vehicle	31371.0	silver	3n1cn7ap4jl884311	167762228	florida	usa	21 hours left

2456 rows × 12 columns

Figure 2: Initial Dataset After Editing

Temporal Visualization

For the Temporal Visualization, the group made use of the “year”, “price”, and “brand” features to answer the question, “What were the prices of cars of each brand throughout the years?”. The group’s hypothesis is that the prices of cars in the earlier years would be considerably high since it was harder to manufacture cars back then, and then the price would go down as the process of manufacturing cars becomes more common, and then rise up again because modern cars are more efficient and would then need better parts.

The Figures 3 and 4 below presents the price of cars of each brand over time, Figure 3 presents a lineplot that visualizes all of the brands at once, while Figure 4 presents a relationship plot that shows a lineplot for each brand. From these figures we can see the following, Chevrolet had the earliest record of having a car for sale in 1973 with a price of 29800 and has the longest record among the 28 brands in the dataset, Ford, Jeep, Chrysler, and GMC, had the most consistent rise in price, BMW had the most abrupt increase in its price around the middle of the 2010s into the early 2020s, Mercedes-Benz had the most inconsistent prices but also had the highest priced car back in 2017 with 84900, and that there are brands such as Lincoln, Jaguar and Toyota that looked like they had no results in the second figure because of how miniscule their price was compared to the other brands.

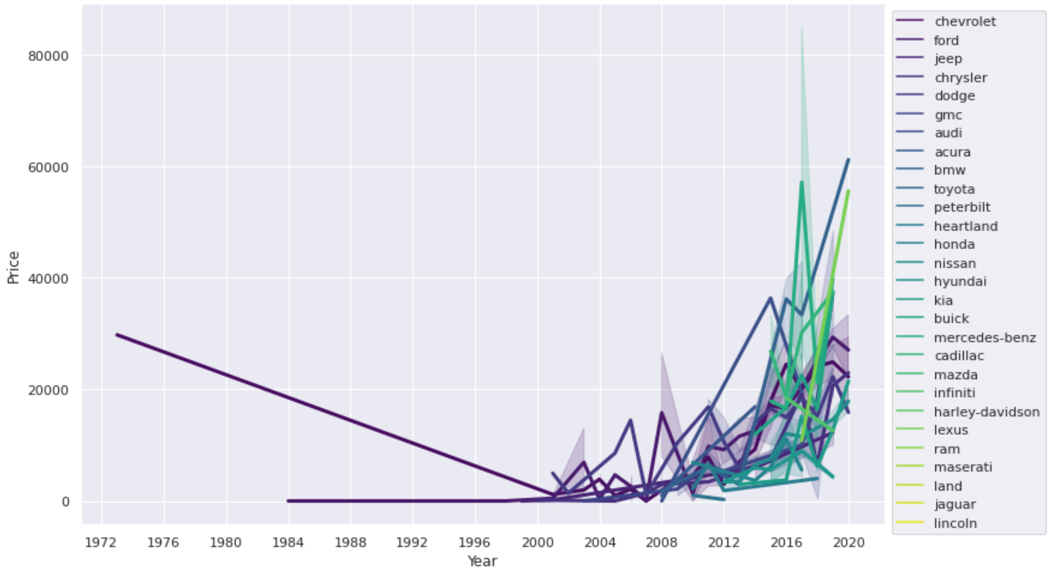


Figure 3: Price of Cars per Brand over Time

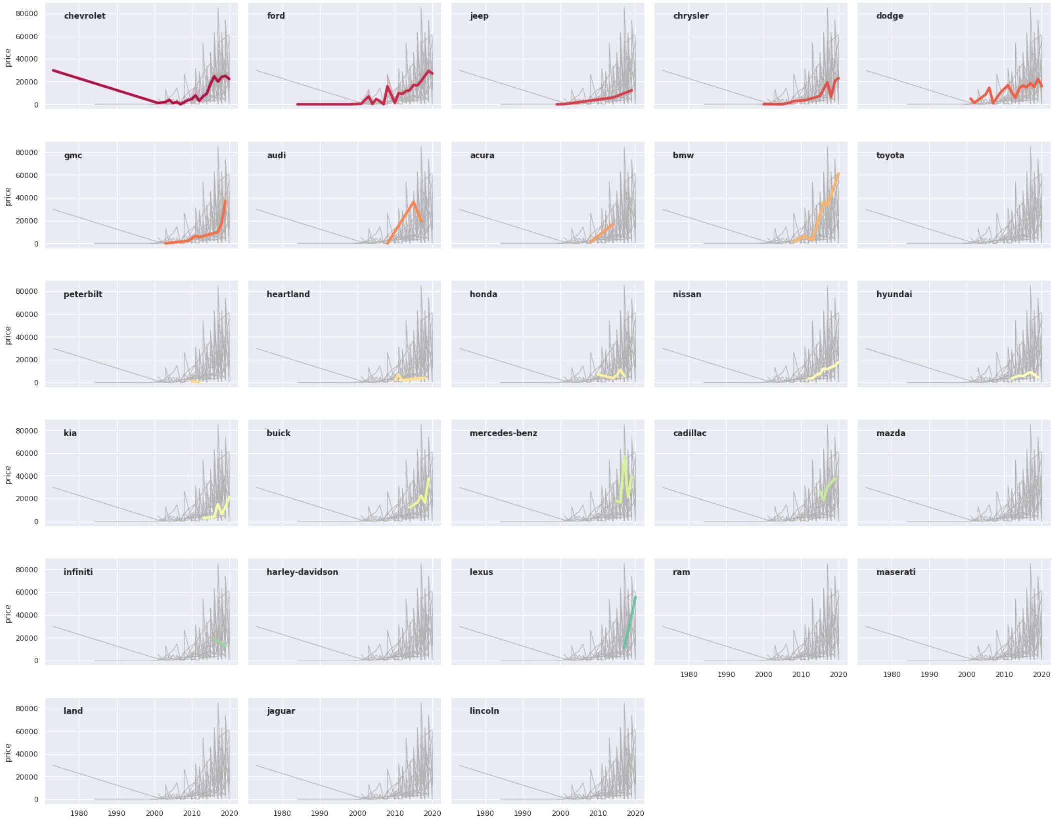
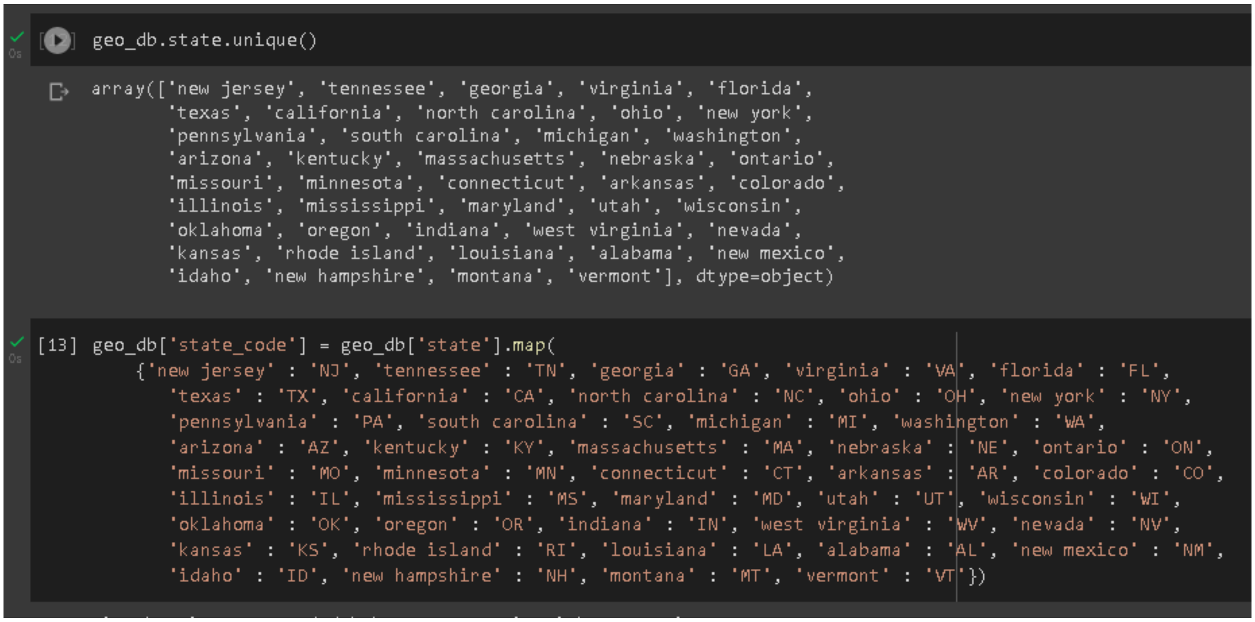


Figure 4: Price of Cars per Brand over Time 2

Geospatial Visualization

The initial dataset lacked the state code that is required to make use of the choropleth and scatter_geo functions of the plotly express, in order to do this the group had to input the equivalent state code of the states in the dataset. Figure 5 below shows the method in which the group achieved this.



```
[ ] geo_db.state.unique()

array(['new jersey', 'tennessee', 'georgia', 'virginia', 'florida',
       'texas', 'california', 'north carolina', 'ohio', 'new york',
       'pennsylvania', 'south carolina', 'michigan', 'washington',
       'arizona', 'kentucky', 'massachusetts', 'nebraska', 'ontario',
       'missouri', 'minnesota', 'connecticut', 'arkansas', 'colorado',
       'illinois', 'mississippi', 'maryland', 'utah', 'wisconsin',
       'oklahoma', 'oregon', 'indiana', 'west virginia', 'nevada',
       'kansas', 'rhode island', 'louisiana', 'alabama', 'new mexico',
       'idaho', 'new hampshire', 'montana', 'vermont'], dtype=object)
```

```
[13] geo_db['state_code'] = geo_db['state'].map(
         {'new jersey' : 'NJ', 'tennessee' : 'TN', 'georgia' : 'GA', 'virginia' : 'VA', 'florida' : 'FL',
          'texas' : 'TX', 'california' : 'CA', 'north carolina' : 'NC', 'ohio' : 'OH', 'new york' : 'NY',
          'pennsylvania' : 'PA', 'south carolina' : 'SC', 'michigan' : 'MI', 'washington' : 'WA',
          'arizona' : 'AZ', 'kentucky' : 'KY', 'massachusetts' : 'MA', 'nebraska' : 'NE', 'ontario' : 'ON',
          'missouri' : 'MO', 'minnesota' : 'MN', 'connecticut' : 'CT', 'arkansas' : 'AR', 'colorado' : 'CO',
          'illinois' : 'IL', 'mississippi' : 'MS', 'maryland' : 'MD', 'utah' : 'UT', 'wisconsin' : 'WI',
          'oklahoma' : 'OK', 'oregon' : 'OR', 'indiana' : 'IN', 'west virginia' : 'WV', 'nevada' : 'NV',
          'kansas' : 'KS', 'rhode island' : 'RI', 'louisiana' : 'LA', 'alabama' : 'AL', 'new mexico' : 'NM',
          'idaho' : 'ID', 'new hampshire' : 'NH', 'montana' : 'MT', 'vermont' : 'VT'})
```

Figure 5: State Code equivalents of the States

This new dataset is then used for the aforementioned functions. Figure 6 presents a choropleth that answers the questions "What brands are the most popular in each state?". From what we can see, the most popular brand is Nissan, followed by Chevrolet, and the least popular brands are Mazda and Peterbilt.

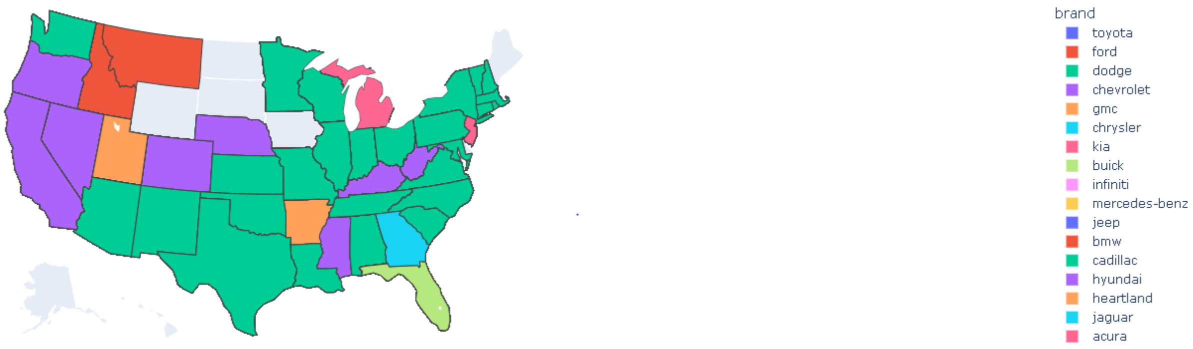


Figure 6: Most Popular Brands in each State

Figure 7 below presents a proportional symbol map answering the question, "What are the average prices of cars in each state?". From the figure, we can notice that most prices range from 15000 to 35000 for the majority of the states, we can also see that the states of Florida, Washington, and Kentucky had prices that reached past 50000.

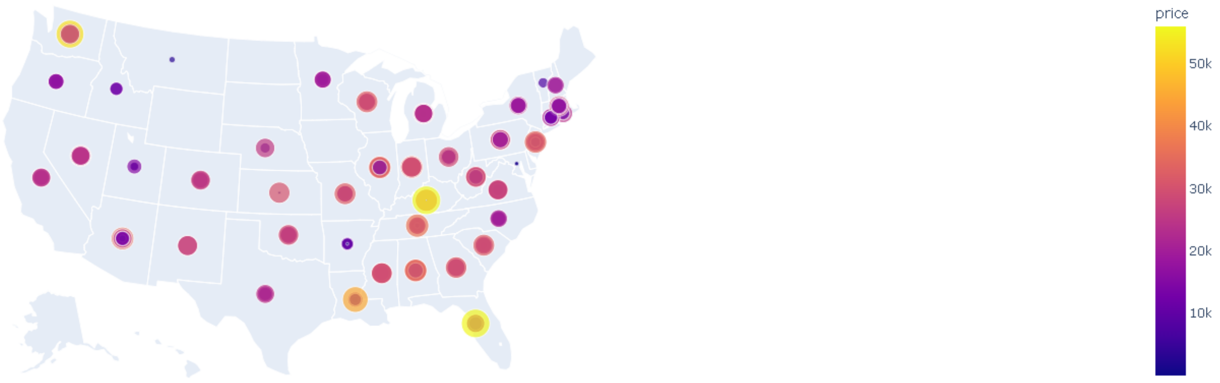
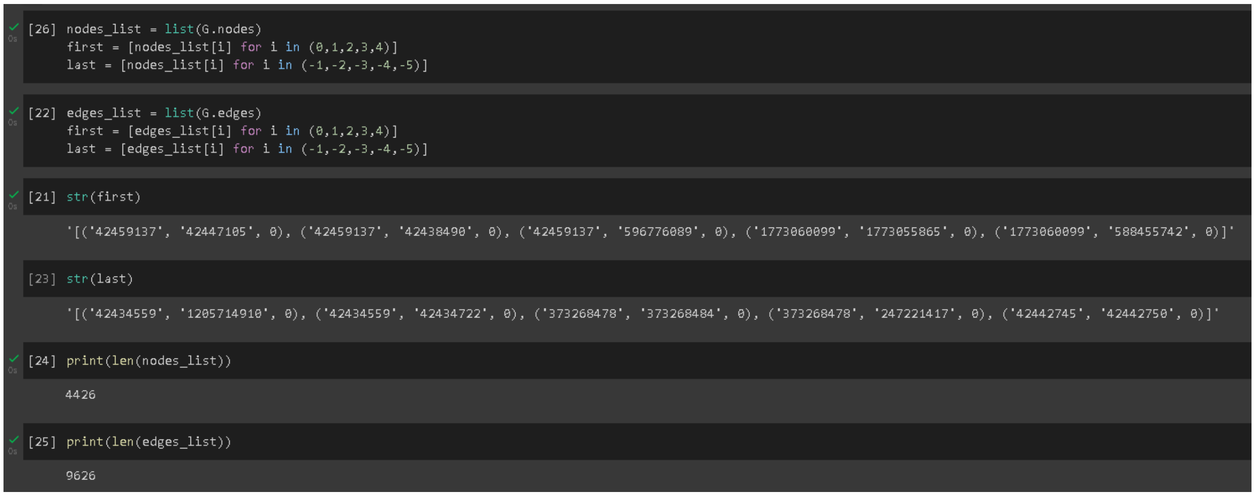


Figure 7: Average Price of Cars in each State

Network Visualization

The initial dataset did not have the necessary data for use to make use of it in the network visualization, therefore, we made use of a dataset loosely related to the initial dataset that is about the Street Network of Manhattan, a borough in the City of New York. The dataset came from Kaggle in a graphml file. Figure 7 shows the first and last five elements of both the nodes and edges as well as their respective lengths.



```
[26] nodes_list = list(G.nodes)
     first = [nodes_list[i] for i in (0,1,2,3,4)]
     last = [nodes_list[i] for i in (-1,-2,-3,-4,-5)]

[22] edges_list = list(G.edges)
     first = [edges_list[i] for i in (0,1,2,3,4)]
     last = [edges_list[i] for i in (-1,-2,-3,-4,-5)]

[21] str(first)

     '[('42459137', '42447105', 0), ('42459137', '42438490', 0), ('42459137', '596776089', 0), ('1773060099', '1773055865', 0), ('1773060099', '588455742', 0)]'

[23] str(last)

     '[('42434559', '1205714910', 0), ('42434559', '42434722', 0), ('373268478', '373268484', 0), ('373268478', '247221417', 0), ('42442745', '42442750', 0)]'

[24] print(len(nodes_list))

     4426

[25] print(len(edges_list))

     9626
```

Figure 7: Details of the Street Network of Manhattan Dataset

Figure 8 below presents a circular diagram presenting the street network of Manhattan. The figure looks congested because the number of nodes and edges in the dataset is larger than what we have practiced with. From what we can see, the nodes in this network are very connected to each, almost every node is connected to at least two other nodes and that there some nodes connected to each other.

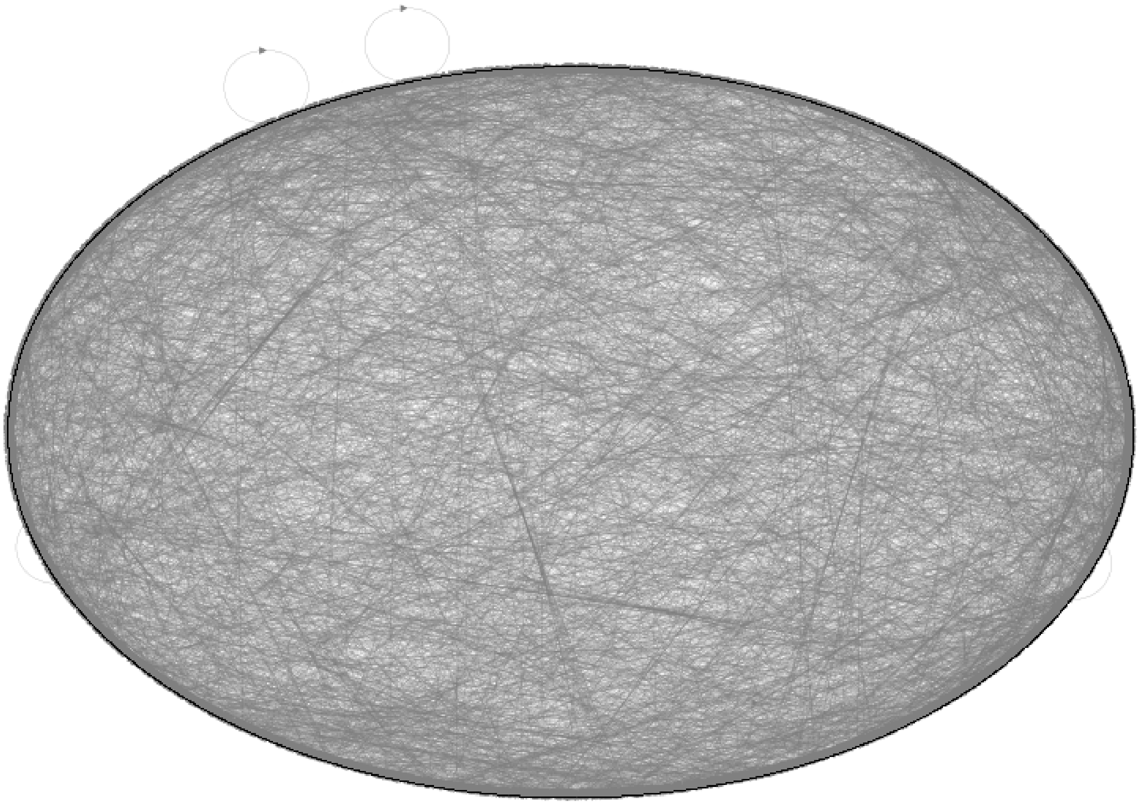


Figure 8: Circular Diagram of the Manhattan Street Network

Hierarchical Visualization

The initial dataset is able to provide us with a hierarchy of the different brands and the model/s they manufacture. From the figures below, we can answer the question, “What are the models available in each brand?”. We can see that Ford has the most model with 40 different models, followed by Dodge with 13 different models. We can also see that there are brands such as Toyota and Lincoln that only have 1 model.

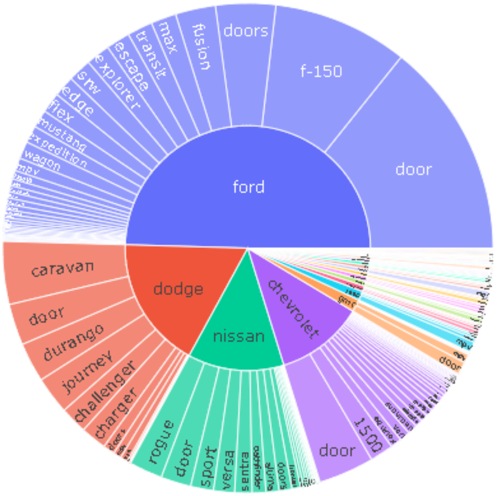


Figure 9: Hierarchy of all Brands and their Models

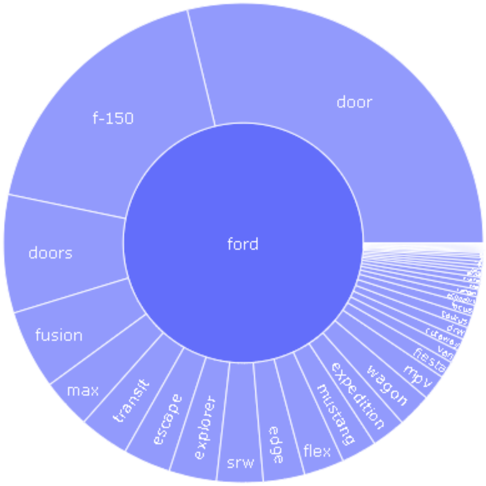


Figure 10: Hierarchy of Ford and its Models

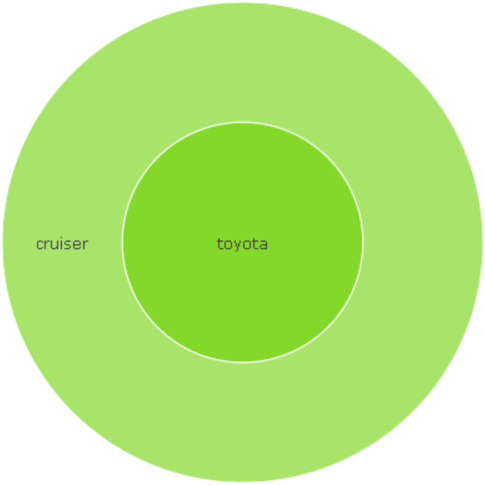


Figure 11: Hierarchy of Toyota and its Model

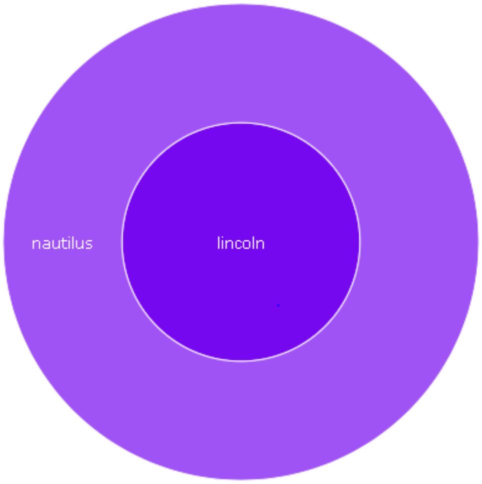


Figure 12: Hierarchy of Lincoln and its Model

Combining the results of the Temporal Visualization and the Hierarchical Visualization, we can conclude that the price of a brand is directly proportional to the amount of model they provide. In other words, if a brand has a lot of car models in their name, the amount of sales they can potentially have also increases, this is acceptable as more models means there are more variations to the cars a brand can provide and this increases the chance that a specific car can suite a buyer's preferences.

Sources of the Dataset:

Street Network of New York in GraphML:
https://www.kaggle.com/datasets/crailtap/street-network-of-new-york-in-graphml?select=manhatten.graphml
US Cars Dataset:
https://www.kaggle.com/datasets/doaaalsenani/usa-cers-dataset