Ahmed El-Ghazouani
Advanced Programming
Professor Hashmi
March 10, 2022

Midterm Project

Introduction

As COVID-19 hit, there has been a shortage of microchips worldwide, and this has caused many shortages in different sectors. One industry that has been heavily hit is the automotive industry, especially in the U.S. This has led to a shortage in the supply of cars, which has caused a great increase in the demand for used cars, which ultimately led to a huge jump in the price of used cars. In this analysis, I aim to analyze the used car market between April 2021 and May 2021, specifically in the U.S.

Questions Posed

The questions posed in this analysis were mainly in regard to the number of cars used in different states and from different brands. This allowed me to better understand what brands were most in demand in this crisis, and also which states were listing the most cars.

Dataset Used

The data used is from kaggle, and it is a dataset that was compiled by an individual who scraped craigslist for any used car post. This resulted in a dataset with many variables:

id	price	condition	**Title status**	size	description
url	year	cylinder	transmission	type	**county**
region	maunfacturer	fuel	**VIN**	Paint color	state
Region url	model	odometer	drive	**Image_url**	Posting date

*The bolded values are the variables I deemed irrelevant in my analysis so I removed them

Since I am not conducting any supervised learning, I was able to clean all the data by simply removing rows that included NULL values. This allowed me to ensure all the rows I used had complete data if I needed it. This made me go from an initial number of observations of ~400k to 79k.

Challenges

The first task and challenge I encountered was attempting to clean the data by removing rows with null values. For some reason, every time I ran a line with my original data frame to remove the rows pertaining to a specific variable, it would work, but when I did the same thing for the next variable, the old variable I removed would reappear. To go around this, I decided to create a new data frame each time I removed rows from a variable that contained a null value. In the end, I ended up with a df_final, which I used for the rest of the analysis. The second challenge I encountered was attempting to create a map based on the latitude and longitude variables that were provided in the dataset. With these, I thought it would be interesting to visualize where in the US were located all these used cars. However, my attempt did not work as planned because some values were not accurate and resulted in inaccurate

locations across the globe. The process of downloading a map from a website that could be read by Python and then filtering it to only include the US was also a challenge.

Libraries Used

Data Analysis

```
#import all packages
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import geopandas as gpd
from shapely.geometry import Point, Polygon
from geopandas import GeoDataFrame
from sklearn.model_selection import train_test_split
# Import Statsmodel functions:
import statsmodels.formula.api as smf

%matplotlib inline
```

News API and Twitter API Analysis

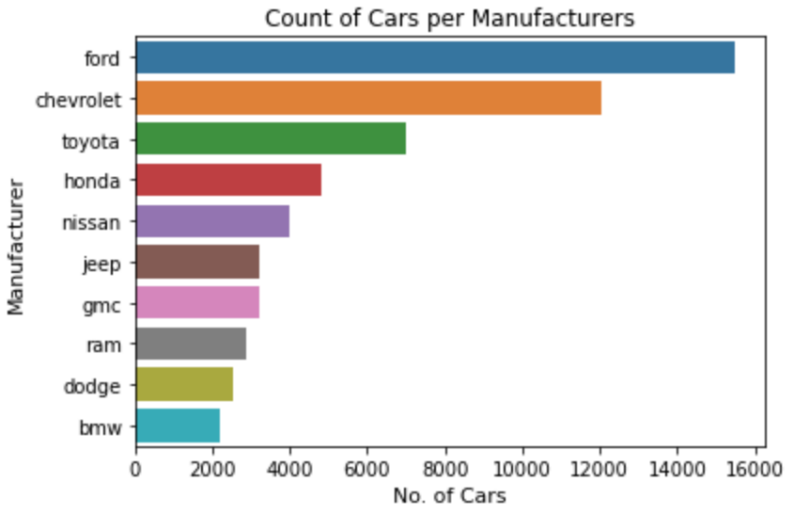
```
from GoogleNews import GoogleNews
from newspaper import Article
import pandas as pd
import numpy as np
from sklearn.datasets import load_iris
import pandas as pd
# Import modules
import matplotlib.pyplot as plt
import re
import string
import collections
from nltk.corpus import stopwords
import nltk
#Word Cloud
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from PIL import Image
#twitter sentiment analysis
import tweepy
from textblob import TextBlob
import seaborn as sns
```

Results

I. Data Analysis

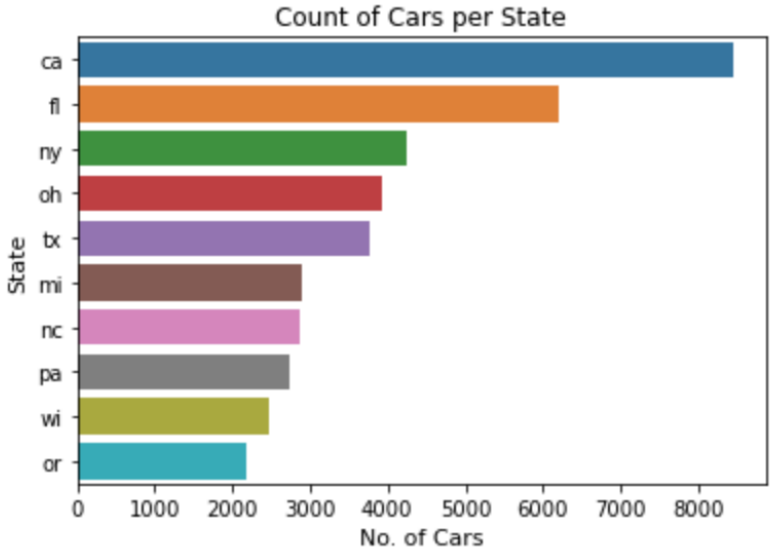
In regards to the data analysis portion of this project, the questions posed were to get a better understanding of what the industry was looking like.

1. Which manufacturer has the most used cars for sale?



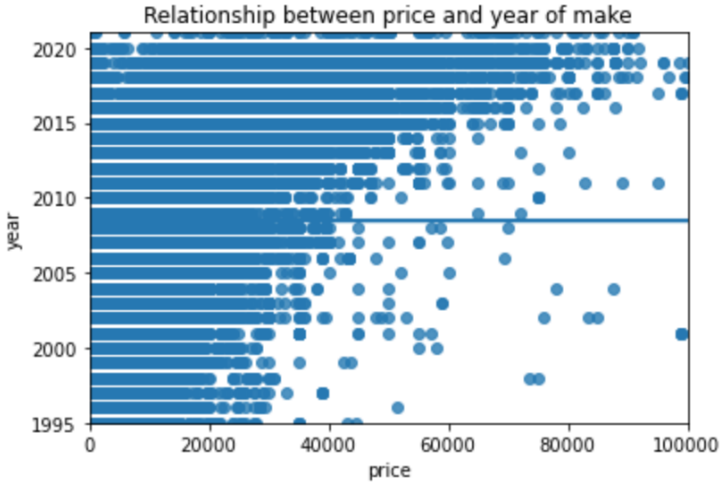
This graph shows us the top 10 manufacturers ranked by the number of cars that are currently on sale on Craigslist. Here we can see that Ford and Chevrolet account for the main sales. Another interesting insight is that American made cars are the most popular.

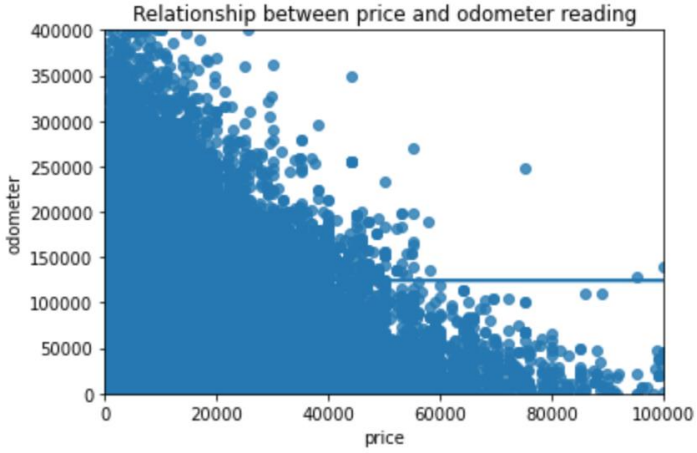
2. **Which states have the most used cars for sale?**



This graph shows the top 10 states where there are the most used cars. California, Florida, and NY were not surprising; however, Ohio was ranked 4th, ahead of Texas, and that was an interesting piece of information.

3. **Relationship between price and year, and price and odometer**



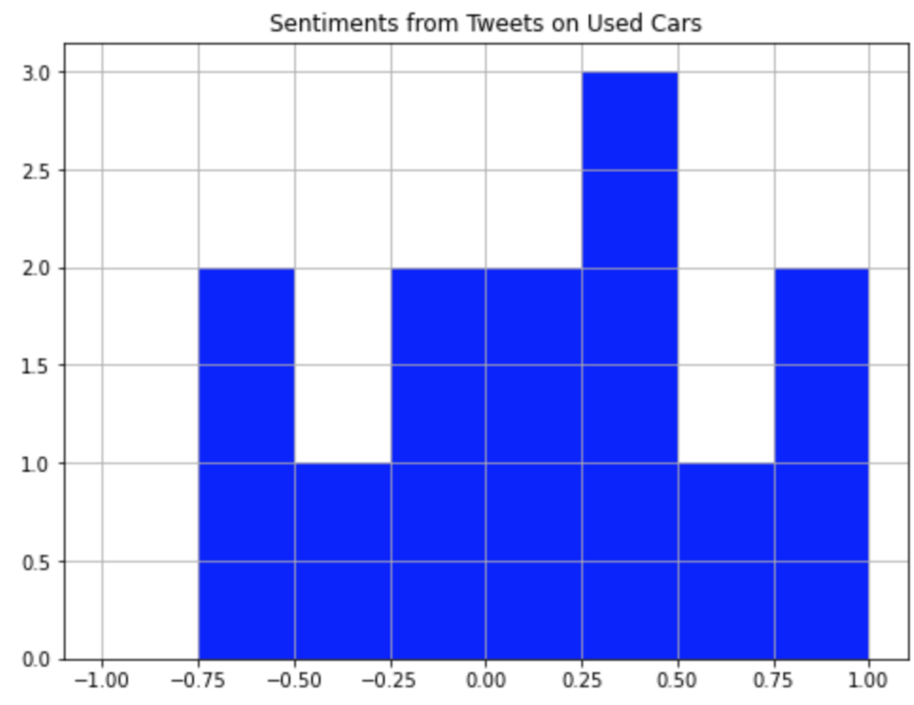


The number of observations made the reference line irrelevant in this case, however, in both plots we can see that there is a certain relationship between the year a car was made and the price of that car. The same goes for the number of miles a car has and the price.

II. Sentiment Analysis

The first step was to get all the news related to 'used cars' from Google News. From that, I was able to separate each title, and then each word. This allowed me to get the word cloud below, which I put into the shape of a car for presentation purposes. This word cloud gives us interesting insights into what is said about the used car market. A word that sticks out is 'shortage' which gives us hints about the state of the market. Inflation is another interesting word which we may interpret to justify higher prices. This word cloud gives us a good idea of the state of the market.

Next, I ran a sentiment analysis on tweets about used cars. This analysis was not very conclusive, although there were more tweets deemed positive in this scenario. My hypothesis would have been that there should be a negative sentiment regarding used cars, however this was wrong.



Conclusion

In conclusion, the used car market has been through a lot in the past couple of years. It was interesting to see which brands are being sold the most, and in which states. It was also interesting to see what the news was saying about it, and how people on Twitter were reacting.

Citations

I. Used Cars Dataset - Kaggle Website

https://www.kaggle.com/austinreese/craigslist-carstrucks-data

II. Helpful Websites

https://jcutrer.com/python/learn-geopandas-plotting-usmaps

https://www.census.gov/cgi-bin/geo/shapefiles/index.php

https://stackoverflow.com/questions/53233228/plot-latitude-longitude-from-csv-in-python-3-6

https://stackoverflow.com/questions/46623583/seaborn-countplot-order-categories-by-count

https://pypi.org/project/GoogleNews/

https://www.statsmodels.org/devel/generated/statsmodels.regression.linear_model.OLS.html

III. Helpful people

Omercan Polat

Aayush Tripathi