NHTSA Complaints

January 20, 2021

1 Introduction

Cars are an integral and vital part of our everyday lives, with riding in motor vehicles being the main way Americans get around. In 2016, a study showed that over 75% of Americans drive to work everyday while 9% of Americans carpool. While cars are an essential method to get around, they pose their own risks when operating when going from place to place.

According to statistics, in 2016 there were over seven million crashes and around 102 people had a fatal crash a day which equates to a person being involved in a fatal crash every 14 minutes. While we can only do our best individually to take the human error factor out of crashes, there is the factor of car defects/part failures to account for that also attribute to crashes.

By analyzing statistics and data, we can ask ourselves important questions that can help us make informed decisions on what cars are reliable and which ones we should purchase. Some of these questions are which car brand has the lowest amount of defects over the course of history, if luxury cars are safer and less prone to defects than economy cars, and what type of car is more/least prone to defects or problems.

The economy car brands we will be looking at are Nissan, Toyota, Honda, Ford, Dodge. While the luxury car brands we will be looking at are, Lexus, Acura, Infiniti, Cadillac, and BMW. When observing the types of cars, the distinction of car types will be sedans, sport utility vehicles (SUVs), and minivans as they account for the majority of widely used cars.

I think it would be helpful to mention the actual data source here in the Introduction.

2 Preparing to Obtain Data

In order to obtain data to analyze, we refer to the National Highway Traffic Safety Association's complaint database. We can make requests to their database, store our findings, and make a chart to visualize the data.

[1]:
```
# Code provided by Dr. John Ringland
# Import all the necessary libraries
import json
import requests
from os.path import exists,join
from os import makedirs

#NHTSA's api link
url0 = 'http://www.nhtsa.gov/webapi/api/Complaints/vehicle/modelyear/{}/make/{}/
 ↪model/{}?format=json'
```

```
#This function creates a local copy of the data we search for
def get_complaints(make,model,year):
    folder = 'NHTSAcomplaints'
    if not exists(folder): makedirs(folder) #creates a folder to store our␣
 ↪results
    filename = make + '_' + model + '_' + str(year) + '.json'
    path = join(folder,filename)
    if exists( path ): # if an existing stored data for the search is found, we␣
 ↪load the data
        #print('Have previously downloaded this: getting from local file')
        with open(path) as f:
            results = json.load(f)
    else: # if not, we search for the data then we store it
        #print('Haven\'t downloaded this one yet: getting from NHTSA and will␣
 ↪save to local file')
        url = url0.format(year,make,model)
        results = json.loads(requests.get(url).text)
        with open(path,'w') as f:
            json.dump(results,f,indent=3)
    return results
```

From the function above, we are now able to access the NHTSA's database and search for information we need, the function stores the data we search for so we can save time in doing repeated searches in the future. We will be searching in the year range of (2000-2019)

3 Economy Manufactured Cars

These are the models of the brands that we will be using for each type of car made by economy manufacturers. The models were picked to be around the same sizes and price range. Why is your table not rendering properly? https://stackoverflow.com/questions/48655801/tables-in-markdown-in-jupyter

	Car Brands					
		Toyota	Nissan	Honda	Ford	Dodge
	Sedan	Camry	Altima	Accord	Taurus	Charger
Car Types	SUV	Rav4	Rogue	CR-V	Edge	Journey
	Minivan	Sienna	Quest	Pilot	Flex	Grand Caravan

4 Luxury Manufactured Cars

These are the models of the brands that we will be using for each of car made by luxury manufacturers. The models were picked to be around the same sizes and price range.

		Car Brands				

		Car Brands				
Car Types	Sedan SUV	Lexus GS 350 RX350	Acura RLX MDX	Infiniti Q70 QX80	Cadillac CT6 XT6	BMW 540i X5

5 Obtaining Our Data

We can make a list of the names of the models and use that list to search through the database.

[3]:
```
def complaintCount(make,model,fyear,lyear):
    count = 0
    for i in range(fyear,lyear+1):
        count += get_complaints(make,model,i)["Count"]
    return count
```

This function makes it easy for us to get the count of complaints for each car model we decide to search for across a time span.

[4]:
```
carNamesE = {"Toyota":
 ↪[["Camry","Economy","Sedan"],["Rav4","Economy","SUV"],["Sienna","Economy","Minivan"]]
             ,"Nissan":
 ↪[["Altima","Economy","Sedan"],["Rogue","Economy","SUV"],["Quest","Economy","Minivan"]]
             ,"Honda":
 ↪[["Accord","Economy","Sedan"],["CR-V","Economy","SUV"],["Pilot","Economy","Minivan"]]
             ,"Ford":
 ↪[["Taurus","Economy","Sedan"],["Edge","Economy","SUV"],["Flex","Economy","Minivan"]]
             ,"Dodge":
 ↪[["Charger","Economy","Sedan"],["Journey","Economy","SUV"],["Grand␣
 ↪Caravan","Economy","Minivan"]]}

carNamesL = {"Lexus":[["GS 350","Luxury","Sedan"],["RX350","Luxury","SUV"]]
             ,"Acura":[["RLX","Luxury","Sedan"],["MDX","Luxury","SUV"]]
             ,"Infiniti":[["Q70","Luxury","Sedan"],["QX80","Luxury","SUV"]]
             ,"Cadillac":[["CT6","Luxury","Sedan"],["XT6","Luxury","SUV"]]
             ,"BMW":[["540i","Luxury","Sedan"],["X5","Luxury","SUV"]]}

ncarNames = {"Make":[],"Model":[],"Class":[],"Type":[],"Count":[]}
```

In this code block, we are just setting up the relevant searching information and data frame structure for later use.

[5]:
```
for makes in carNamesE:
    for models in carNamesE[makes]:
        ncarNames["Make"].append(makes)
        ncarNames["Model"].append(models[0])
        ncarNames["Class"].append(models[1])
```

```
        ncarNames["Type"].append(models[2])
        ncarNames["Count"].append(complaintCount(makes,models[0],2000,2019))

for makes in carNamesL:
    for models in carNamesL[makes]:
        ncarNames["Make"].append(makes)
        ncarNames["Model"].append(models[0])
        ncarNames["Class"].append(models[1])
        ncarNames["Type"].append(models[2])
        ncarNames["Count"].append(complaintCount(makes,models[0],2000,2019))
```

A lot of data. I wonder how long that took!

6 Organizing Our Data

We should now make a data frame to store the information we will search from the NHTSA database.

[7]:
```
import pandas as pd

data = ncarNames
df = pd.DataFrame(data) #creating a dataframe from ncarNames
```

We now have a data frame that looks like this:

[8]:
```
df
```

[8]:

	Make	Model	Class	Type	Count
0	Toyota	Camry	Economy	Sedan	11123
1	Toyota	Rav4	Economy	SUV	5395
2	Toyota	Sienna	Economy	Minivan	7262
3	Nissan	Altima	Economy	Sedan	10681
4	Nissan	Rogue	Economy	SUV	2741
5	Nissan	Quest	Economy	Minivan	1174
6	Honda	Accord	Economy	Sedan	12805
7	Honda	CR-V	Economy	SUV	8294
8	Honda	Pilot	Economy	Minivan	3744
9	Ford	Taurus	Economy	Sedan	7557
10	Ford	Edge	Economy	SUV	7472
11	Ford	Flex	Economy	Minivan	987
12	Dodge	Charger	Economy	Sedan	4548
13	Dodge	Journey	Economy	SUV	3609
14	Dodge	Grand Caravan	Economy	Minivan	6965
15	Lexus	GS 350	Luxury	Sedan	139
16	Lexus	RX350	Luxury	SUV	679
17	Acura	RLX	Luxury	Sedan	36
18	Acura	MDX	Luxury	SUV	1653
19	Infiniti	Q70	Luxury	Sedan	7
20	Infiniti	QX80	Luxury	SUV	27
21	Cadillac	CT6	Luxury	Sedan	10

```
22  Cadillac            XT6   Luxury      SUV      0
23       BMW           540i   Luxury    Sedan     64
24       BMW             X5   Luxury      SUV   1748
```

We can now manipulate this data frame into groups of data that we want to test and observe.

[18]:
```
byMake = (df.groupby("Make")
           .agg({"Count": "sum"})
           .rename(columns = {"Count":"Number of Complaints"})
           .sort_values("Number of Complaints",ascending = False)
           .reset_index())
byMake
```

[18]:
```
       Make  Number of Complaints
0     Honda                 24843
1    Toyota                 23780
2      Ford                 16016
3     Dodge                 15122
4    Nissan                 14596
5       BMW                  1812
6     Acura                  1689
7     Lexus                   818
8  Infiniti                    34
9  Cadillac                    10
```

[19]:
```
byClass = (df.groupby("Class")
            .agg({"Count": "sum"})
            .rename(columns = {"Count":"Number of Complaints"})
            .sort_values("Number of Complaints",ascending = False)
            .reset_index())
byClass
```

[19]:
```
     Class  Number of Complaints
0  Economy                 94357
1   Luxury                  4363
```

[20]:
```
byType = (df.groupby("Type")
           .agg({"Count": "sum"})
           .rename(columns = {"Count":"Number of Complaints"})
           .sort_values("Number of Complaints",ascending = False)
           .reset_index())
byType
```

[20]:
```
      Type  Number of Complaints
0    Sedan                 46970
1      SUV                 31618
2  Minivan                 20132
```

Happy to see you're already using the aggregation we discussed last week!

7 Visualization of Data

Now that we have the data sorted into appropriate data frames, we can use the Altair library to create visual graphs to better understand the data that we have.

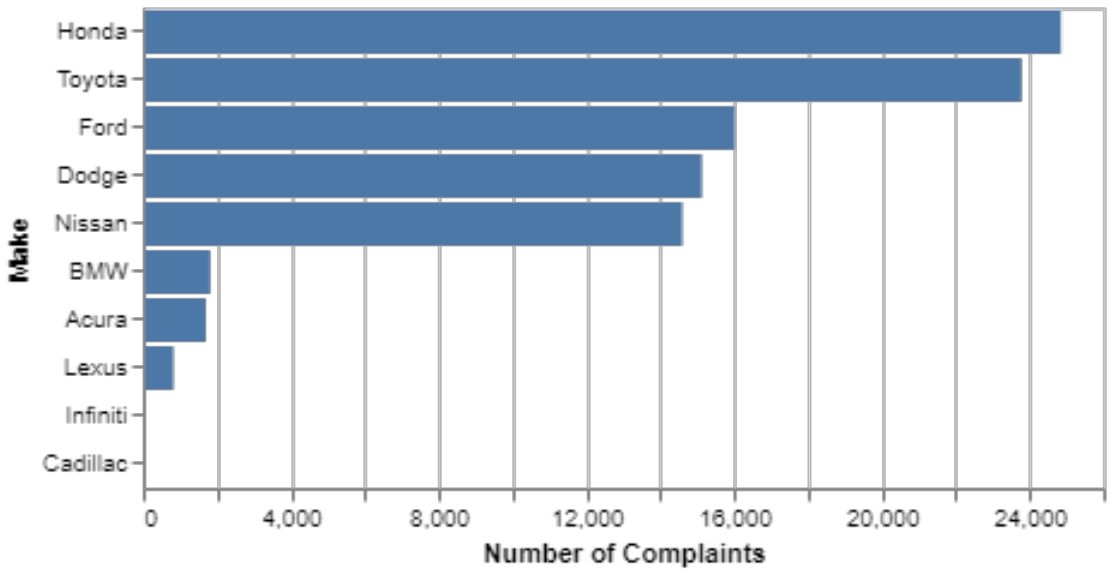
8 Number of Defects by Make

[25]:
```
import altair as alt
alt.renderers.enable("notebook")

alt.Chart(byMake).mark_bar().encode(x = "Number of Complaints",y = alt.
 ↪Y("Make",sort = None))
```

```
<vega.vegalite.VegaLite at 0x220931a7948>
```

[25]:



In future, please suppress "vega commercials" with a semicolon (;).

From this graph we can see that out of all our searched manufacturer's Honda's car had the most amount of complaints against their vehicles.

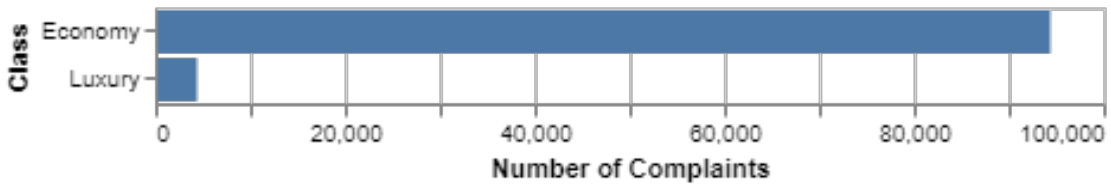
Yes, but for this to be meaningful, we would need to compare to the number of units sold. Data is available here: http://carsalesbase.com/category/car-sales-us/.

9 Number of Defects by Class

[23]:
```
alt.Chart(byClass).mark_bar().encode(x = "Number of Complaints",y = alt.
↪Y("Class",sort = None))
```

```
<vega.vegalite.VegaLite at 0x22093147688>
```

[23]:



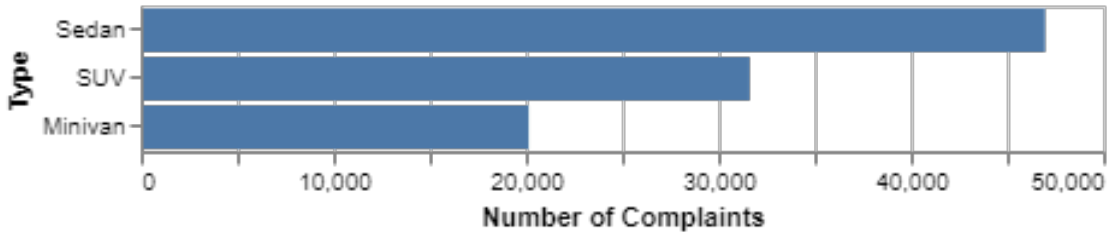
From this graph we can see that between economic and luxury cars, there is a stark difference in the number of complaints. But surely this is largely a reflection of the smaller number of luxury cars sold?

10 Number of Defects by Type

[24]:
```
alt.Chart(byType).mark_bar().encode(x = "Number of Complaints",y = alt.
↪Y("Type",sort = None))
```

```
<vega.vegalite.VegaLite at 0x220931789c8>
```

[24]:



From this graph we can see that out of the car types that we have searched, there seems to be a larger amount of complaints against sedans than other types of cars. Again, to draw any conclusion, we'd really need to divide by the number of units sold.

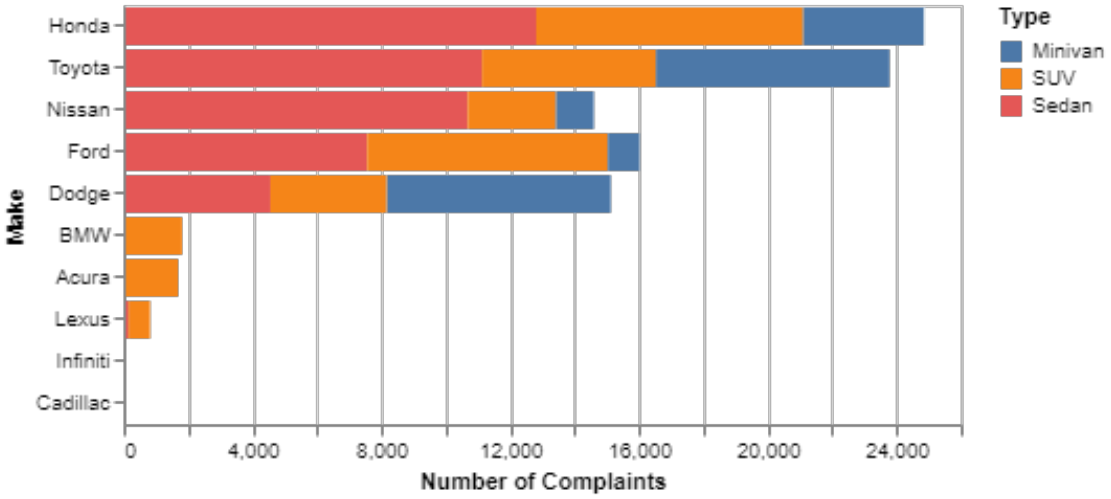
[32]:
```
import warnings
warnings.filterwarnings('ignore')

byMakeType = (df.groupby(("Make","Type"))
              .agg({"Count": "sum"})
              .rename(columns = {"Count":"Number of Complaints"})
                      .sort_values("Number of Complaints",ascending = False)
                      .reset_index())
```

[31]:
```
alt.Chart(byMakeType).mark_bar().encode(x = "Number of Complaints", y = alt.
↪Y("Make", sort = None), color = "Type")
```

```
<vega.vegalite.VegaLite at 0x220931b9f08>
```

[31]:



By filtering the data of complaints by make with types, we can see what percentage of the defects are from what type.

11 Conclusion

From our data, we can see that from the manufacturers that we have searched, the luxury cars overall seem to have less defects reported than their consumer counterparts. This may be due to a few reasons. The first reason being that it really is true that luxury cars are built better and in result have less defects. The second reason could be that our searching of the NHTSA database is flawed due to luxury cars having different series and then in effect having multiple different entries in their database that we have to search for in order to find the true amount of defects for that model. Another reason maybe the fact that luxury car models are not always in production the same amount of time span as other economy models are in production. The searches are tested

across a span of 19 years and while the economy models may be in production for the majority of that time period, the luxury cars may not be. The last reason being that due to the high cost of luxury vehicles, there are simply less people buying luxury cars as there are buying economy cars. I think the last reason is the big one.

Furthermore, it is interesting that the highest percentage of complaints among economy cars is from sedans, while the highest percentage of complaints among luxury cars is from SUVs. This could be indicative of luxury SUVs not being built as well as luxury sedans possibly because of the fact that when people think of luxury cars, they imagine a sleek, high-end, race car resembling vehicle. This might incentivize luxury manufacturers to prioritize building their sedans with as little defects as possible.

12 Sources

https://www.brookings.edu/blog/the-avenue/2017/10/03/americans-commuting-choices-5-major-takeaways-from-2016-census-data/#targetText=Over%2076%20percent%20of%20Americans,hitting%20A

https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812580#targetText=In%202016%20there%20

https://www.consumerreports.org/cars-who-owns-which-car-brands/