CAR CRASH DATA 2016-2020

Group One
4/24/2021

Car Crash Data in Texas from 2016-2020

INFORMATION TAKEN FROM KAGGLE.COM

DATA CROPPED FROM 4.2 MILLION RECORDS TO ONLY INCLUDE TEXAS

ANTHONY JONES, BRITTANY JOHNSON, CARLOS VILLANUEVA, JARED SANDERSON, MATT TERHUNE, AGUST ERLINGSSON

Rationale

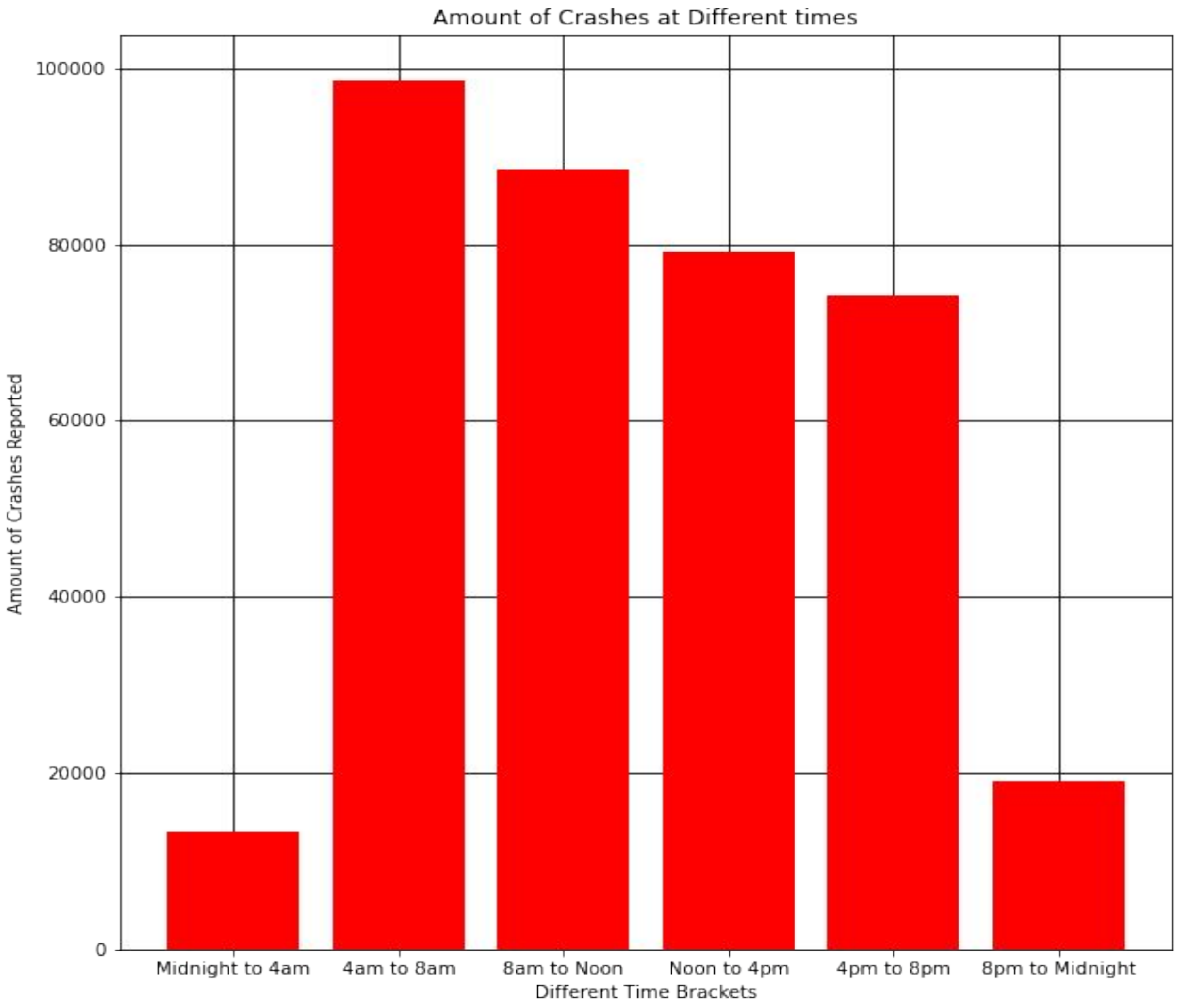
The purpose of this project is to identify whether three factors have a correlation to the severity of a car accident in the state of Texas.

The three variables we will be looking at are the ***weather conditions (precipitation)***, *the* ***time of the day***, and the ***visibility***.

Hypothesis and Questions Asked

Question: Do the weather conditions (precipitation), the time of the day, and the visibility affect the severity of a car accident in the state of Texas?

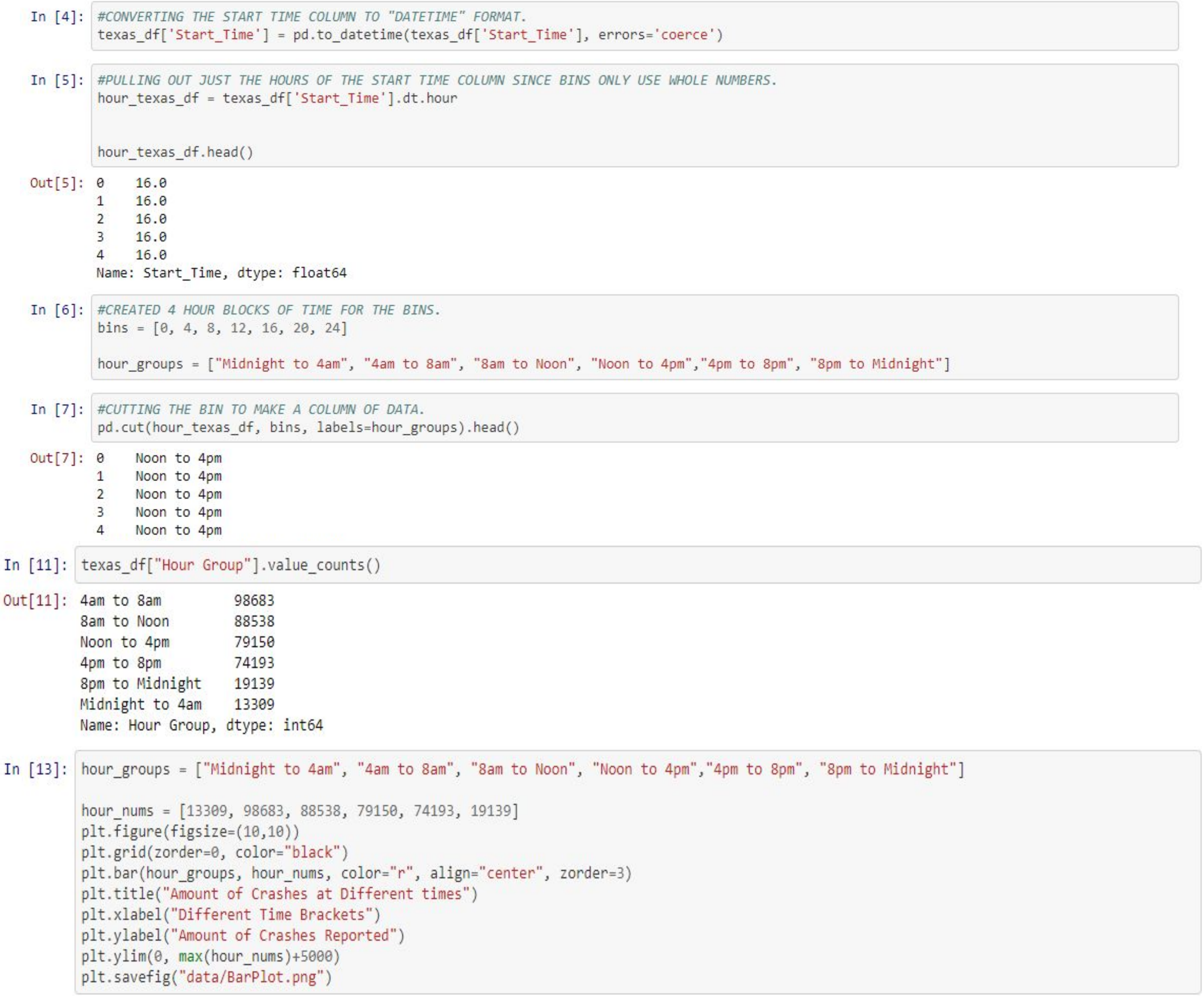
Hypothesis: The weather conditions (precipitation), the time of the day, and the visibility do affect the severity of a car accident in the state of Texas.



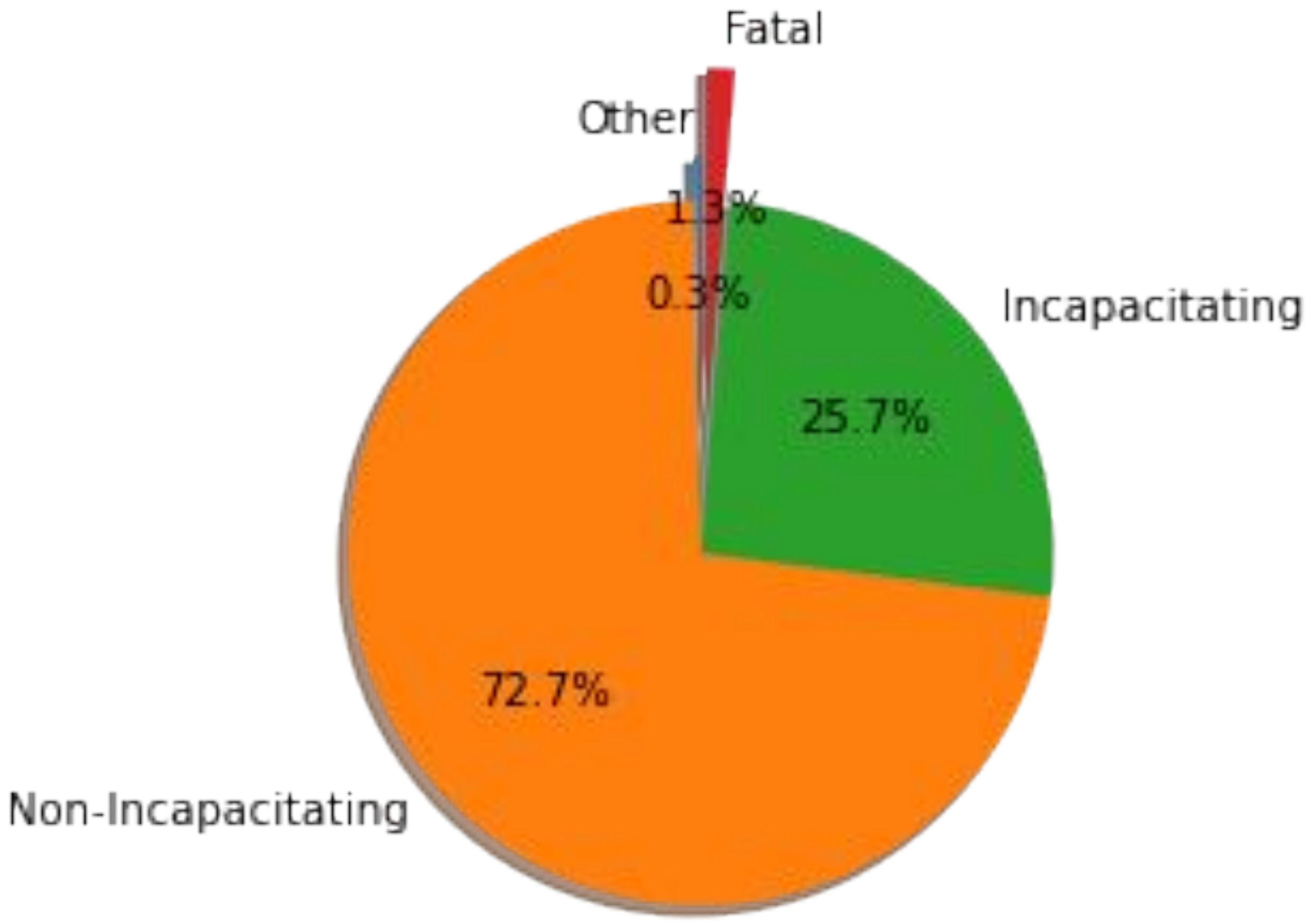
Crash Gross Total versus Time of Day

Two variable were compared which were the time of day and the amount of accidents. By setting six different time blocks we could show which times when accidents occur more often.

Crash Gross Total versus Time of Day cont.



```
In [4]:  #CONVERTING THE START TIME COLUMN TO "DATETIME" FORMAT.
         texas_df['Start_Time'] = pd.to_datetime(texas_df['Start_Time'], errors='coerce')

In [5]:  #PULLING OUT JUST THE HOURS OF THE START TIME COLUMN SINCE BINS ONLY USE WHOLE NUMBERS.
         hour_texas_df = texas_df['Start_Time'].dt.hour


         hour_texas_df.head()

Out[5]:  0    16.0
         1    16.0
         2    16.0
         3    16.0
         4    16.0
         Name: Start_Time, dtype: float64

In [6]:  #CREATED 4 HOUR BLOCKS OF TIME FOR THE BINS.
         bins = [0, 4, 8, 12, 16, 20, 24]

         hour_groups = ["Midnight to 4am", "4am to 8am", "8am to Noon", "Noon to 4pm","4pm to 8pm", "8pm to Midnight"]

In [7]:  #CUTTING THE BIN TO MAKE A COLUMN OF DATA.
         pd.cut(hour_texas_df, bins, labels=hour_groups).head()

Out[7]:  0    Noon to 4pm
         1    Noon to 4pm
         2    Noon to 4pm
         3    Noon to 4pm
         4    Noon to 4pm

In [11]: texas_df["Hour Group"].value_counts()

Out[11]: 4am to 8am         98683
         8am to Noon        88538
         Noon to 4pm        79150
         4pm to 8pm         74193
         8pm to Midnight    19139
         Midnight to 4am    13309
         Name: Hour Group, dtype: int64

In [13]: hour_groups = ["Midnight to 4am", "4am to 8am", "8am to Noon", "Noon to 4pm","4pm to 8pm", "8pm to Midnight"]

         hour_nums = [13309, 98683, 88538, 79150, 74193, 19139]
         plt.figure(figsize=(10,10))
         plt.grid(zorder=0, color="black")
         plt.bar(hour_groups, hour_nums, color="r", align="center", zorder=3)
         plt.title("Amount of Crashes at Different times")
         plt.xlabel("Different Time Brackets")
         plt.ylabel("Amount of Crashes Reported")
         plt.ylim(0, max(hour_nums)+5000)
         plt.savefig("data/BarPlot.png")
```



Severity Breakdown of Accidents

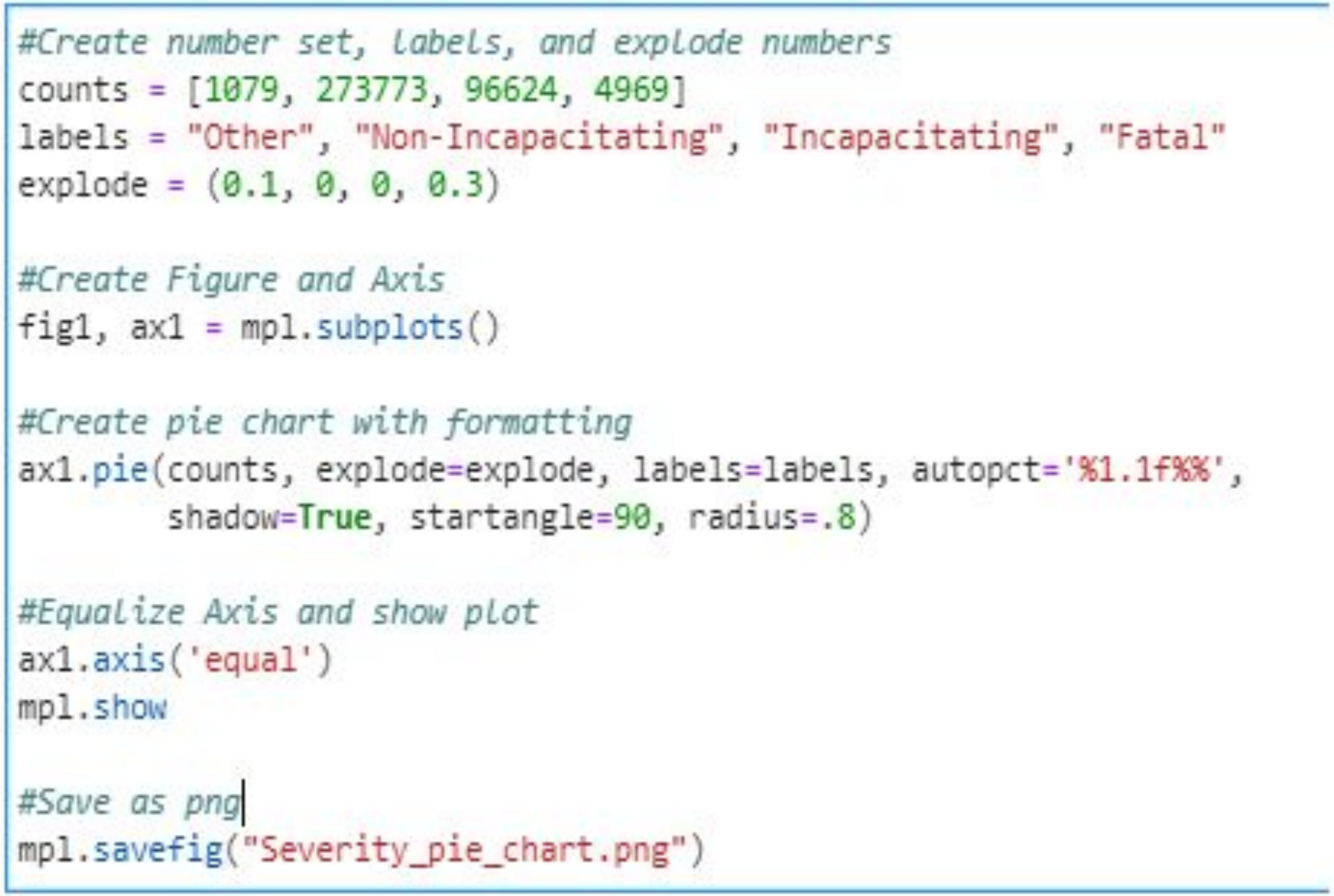
Other - Either not documented or not applicable to the other categories

Non-Incapacitating - Crash resulted in minor injuries to all or some passengers - they could still walk

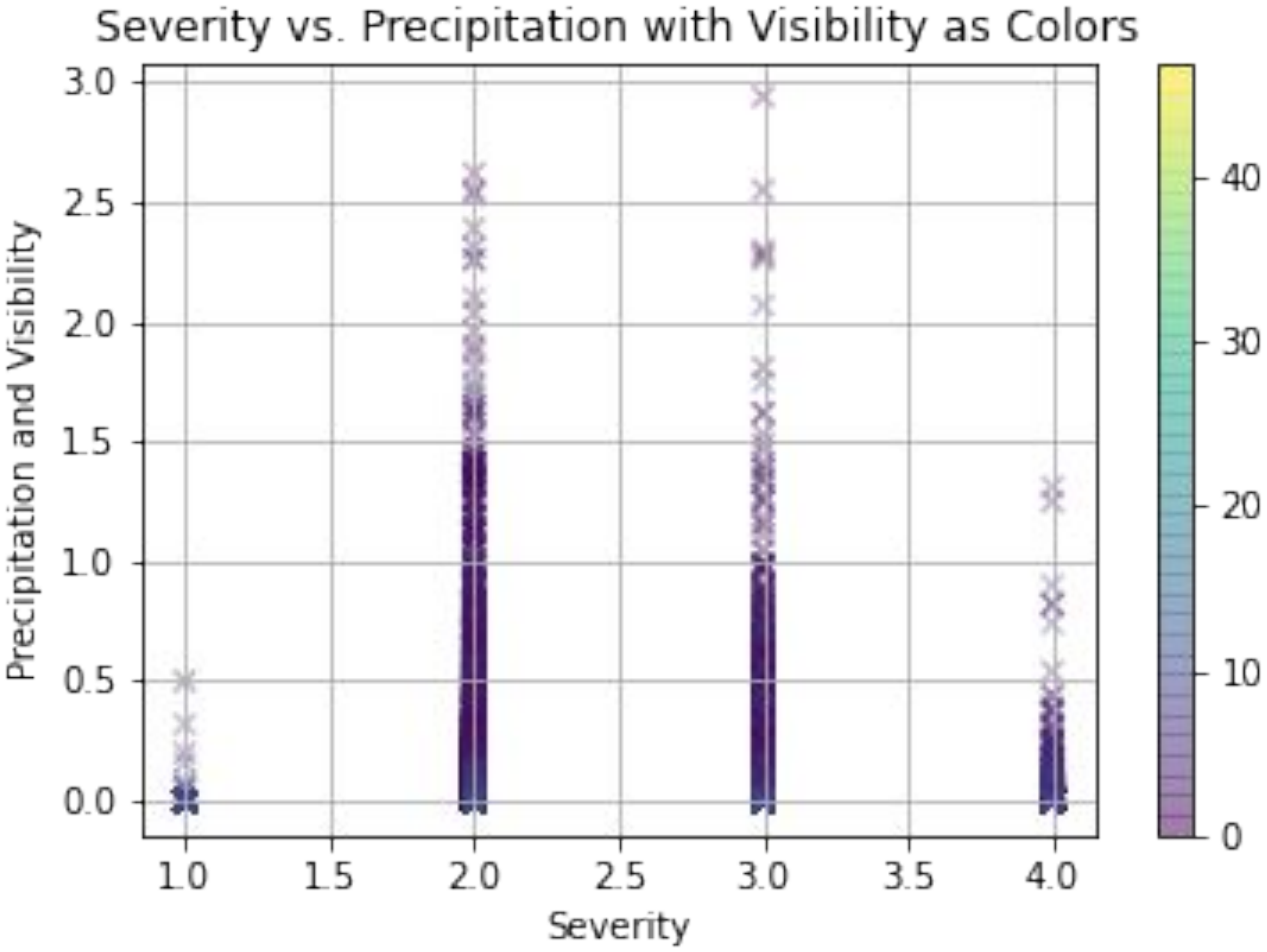
Incapacitating - Crash resulted in passengers needing major attention at hospital - could not walk post-crash

Fatal - One or more person from the crash died as a result of injuries sustained

Code for Severity Pie Plot

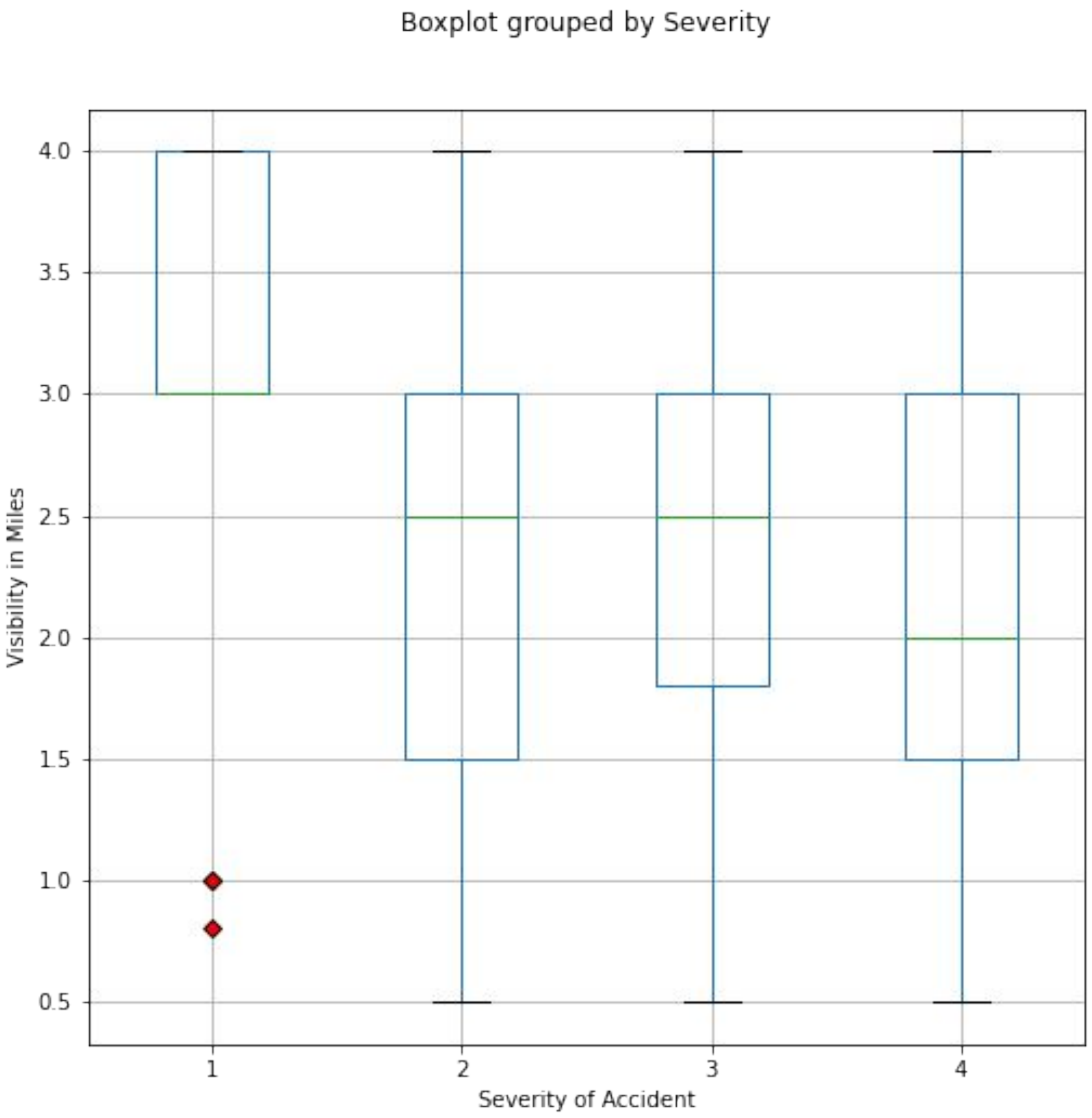


```
#Create number set, labels, and explode numbers
counts = [1079, 273773, 96624, 4969]
labels = "Other", "Non-Incapacitating", "Incapacitating", "Fatal"
explode = (0.1, 0, 0, 0.3)

#Create Figure and Axis
fig1, ax1 = mpl.subplots()

#Create pie chart with formatting
ax1.pie(counts, explode=explode, labels=labels, autopct='%1.1f%%',
        shadow=True, startangle=90, radius=.8)

#Equalize Axis and show plot
ax1.axis('equal')
mpl.show

#Save as png
mpl.savefig("Severity_pie_chart.png")
```



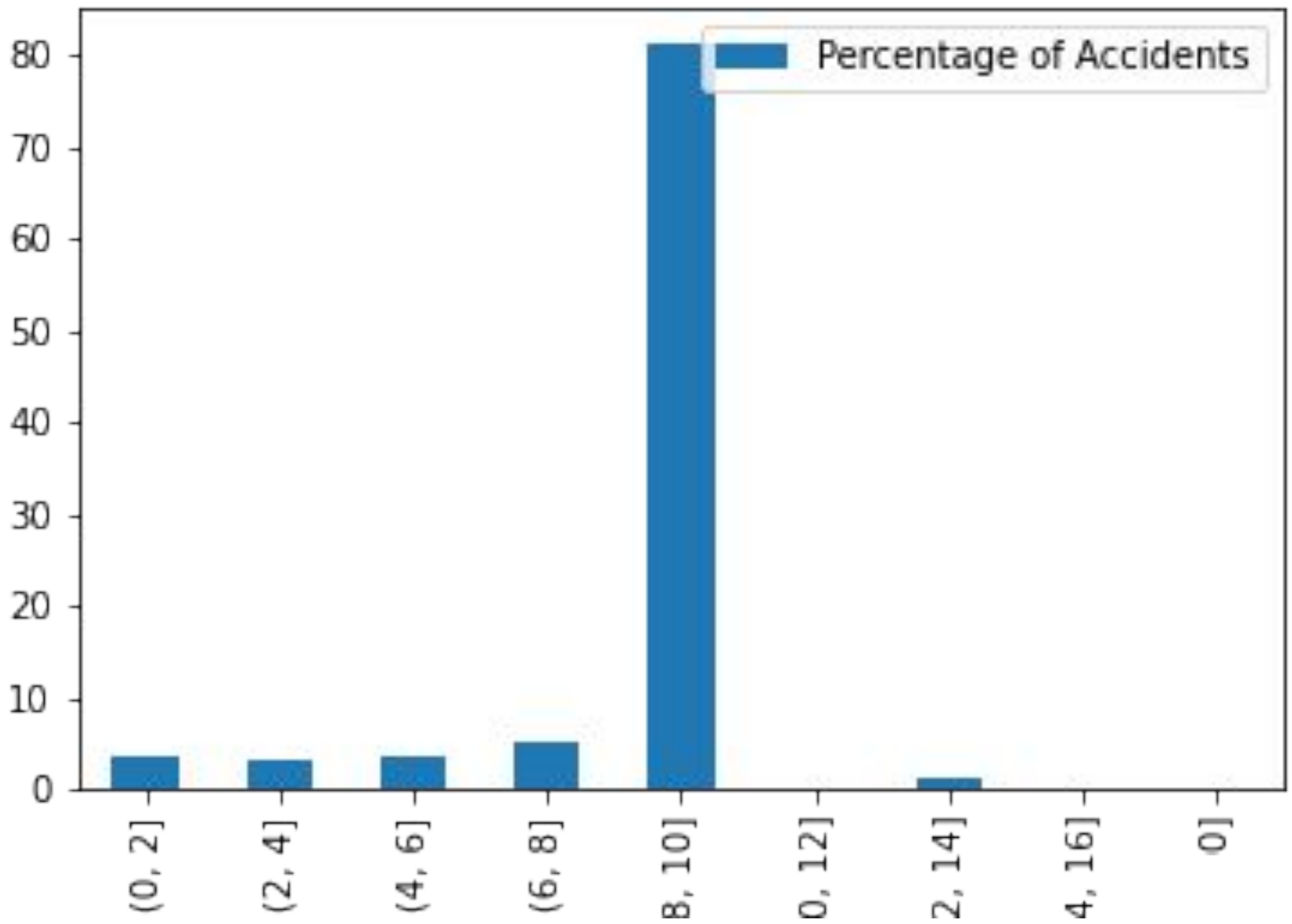
Severity versus Various Weather Factors

Three variables were considered. Severity, shown on the X-axis, Precipitation, shown on the Y-axis, and Visibility, shown as color. Most of the 2-level crashes occurred at below 1.5in of rain, 3-level below 1in and 4-level below .5in. The majority of color, visibility, was from 8-10 miles, shown as a light blue. The density of crashes makes the color more saturated.



Visibility vs. Severity

Does visibility affect the severity of car accidents in Texas?



Percentage of Accidents vs Visibility

The majority of accidents happen around a visibility of 8-10 miles

201: Accidents

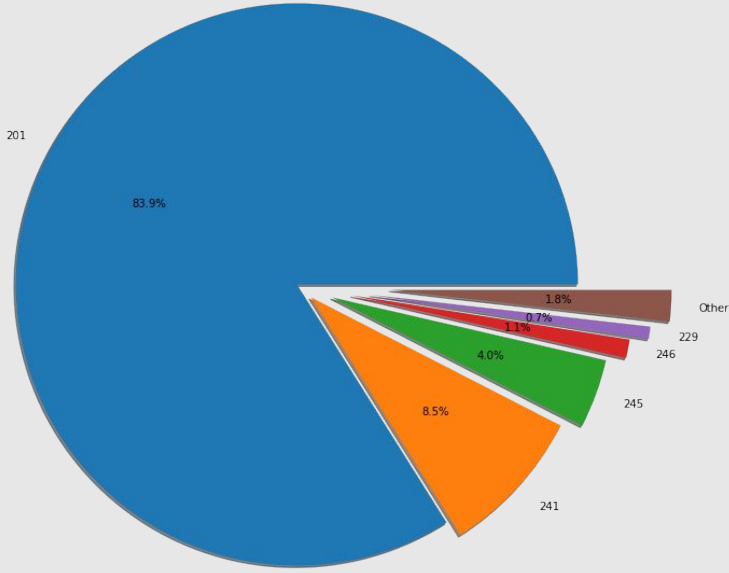
241: Accidents - Right lane blocked

245: Accidents - Two lanes blocked

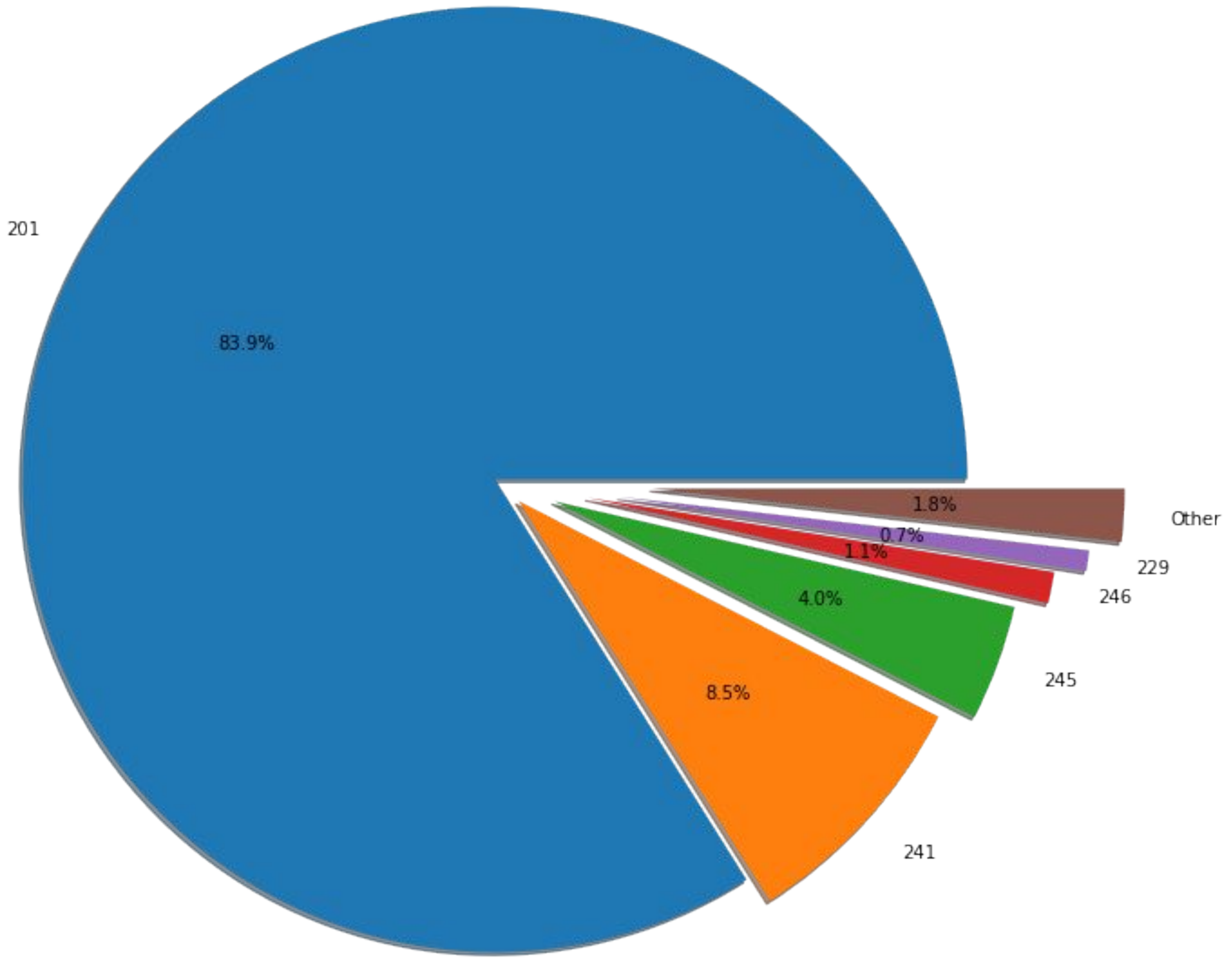
246: Accidents - Three lanes blocked

229: Accident - Slow traffic

- There are a total of 20 TMC codes in the dataset
- The top 5 TMC codes constitute 98.8% of all reports
- Traffic Message Channel is a technology for delivering traffic and travel information to motor vehicle drivers
- In the United States, XM Satellite Radio and Sirius Satellite Radio provide TMC service all over the US

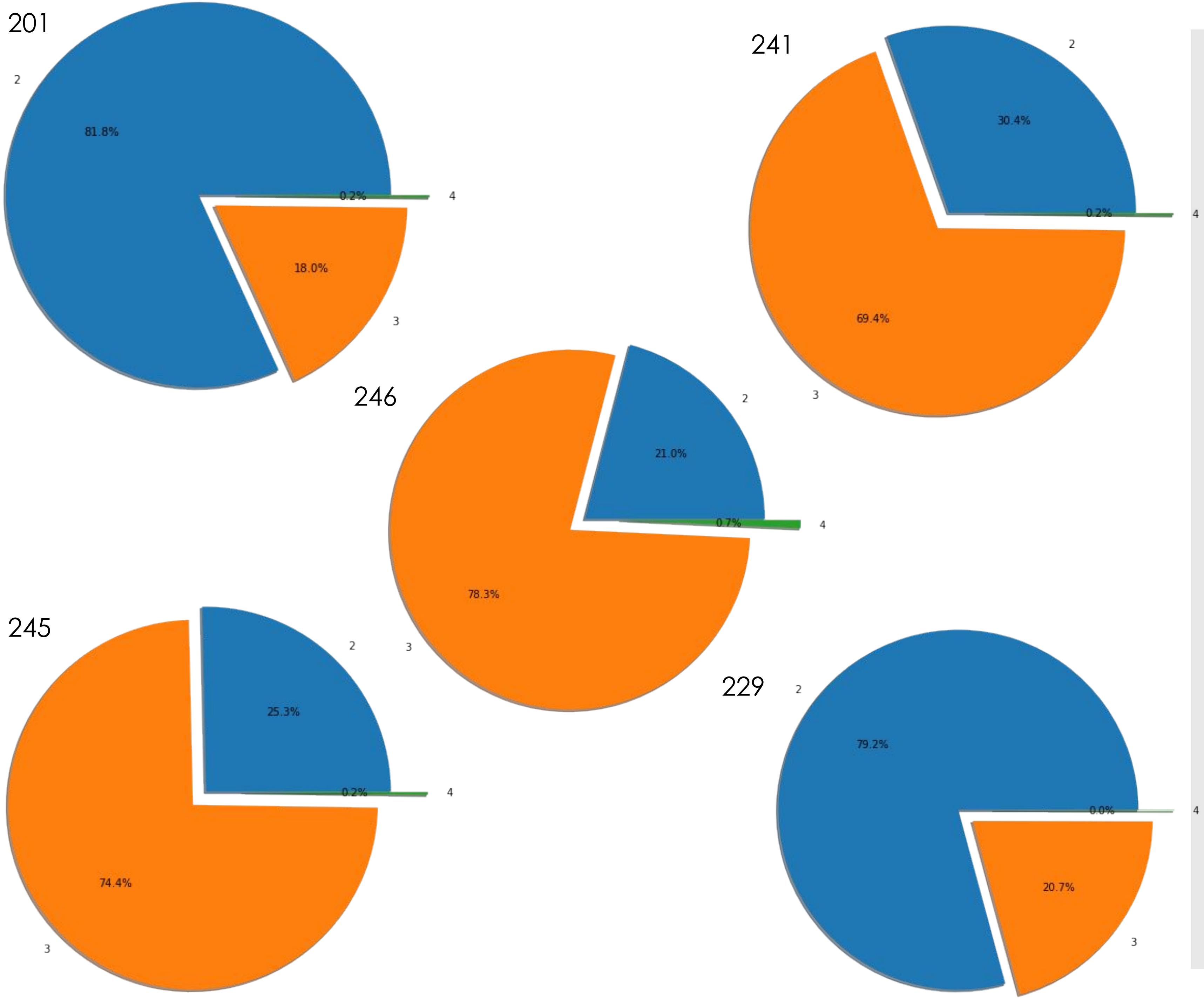


Top 5 traffic message channel (TMC) Codes



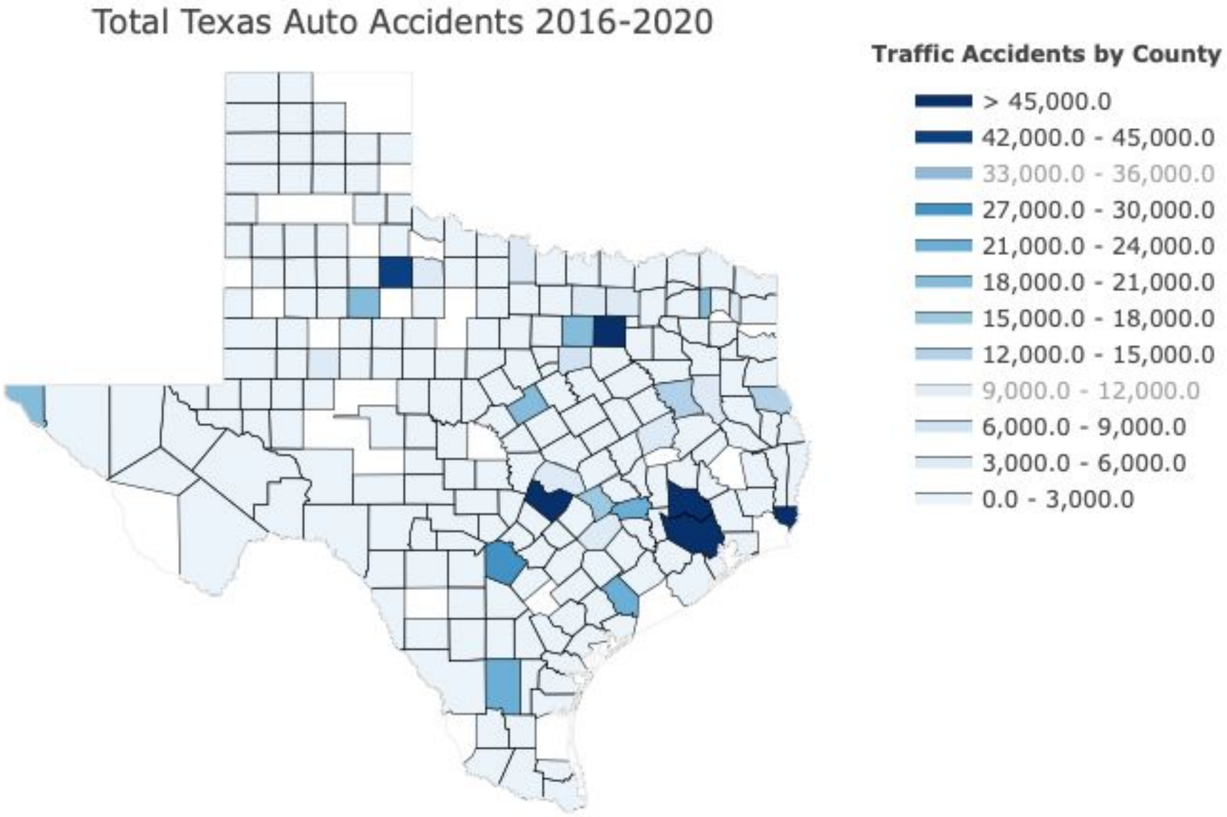
Traffic Message Channel (TMC)

Percentages of TMC codes based on number of occurence



TMC Vs Severity

We learn from the data that the majority of accidents (80% of them) are not severe (201 no disruptions, and 229 slow traffic), and this comes with the combination of little to no traffic disruption. The more an accident disrupts traffic (241 one lane, 245 two lane, and 246 three lanes blocked, the more severe the accident is.



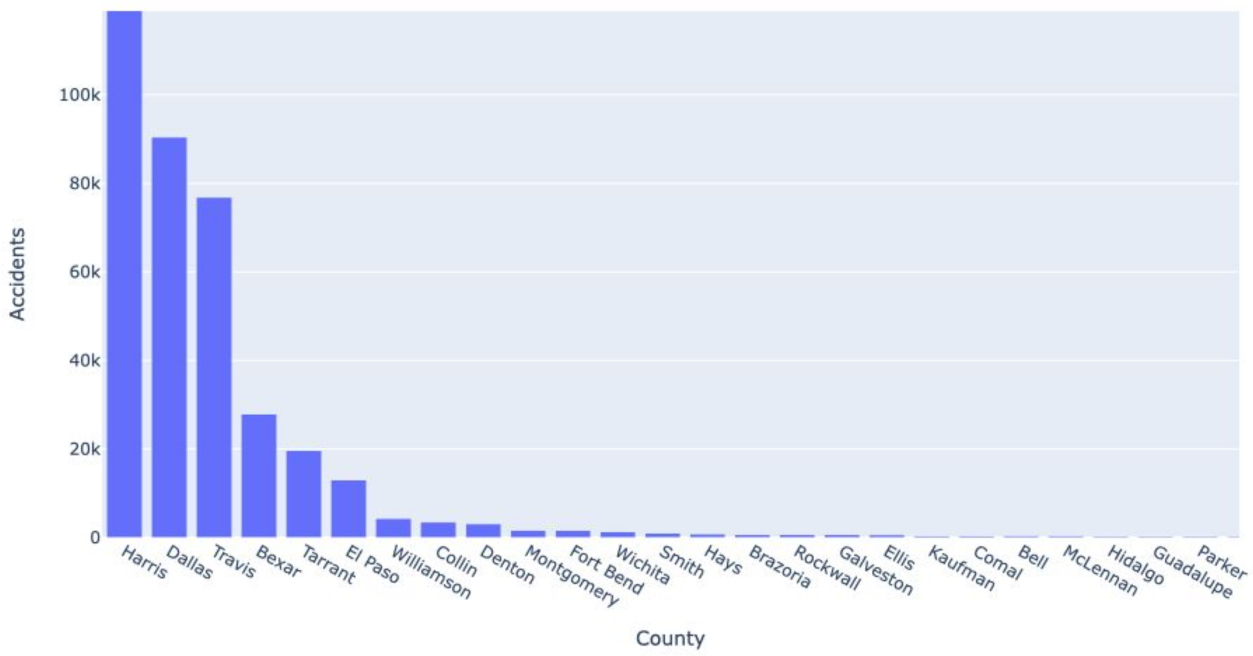
Total Texas Auto Accidents 2016-2020

Texas Choropleth

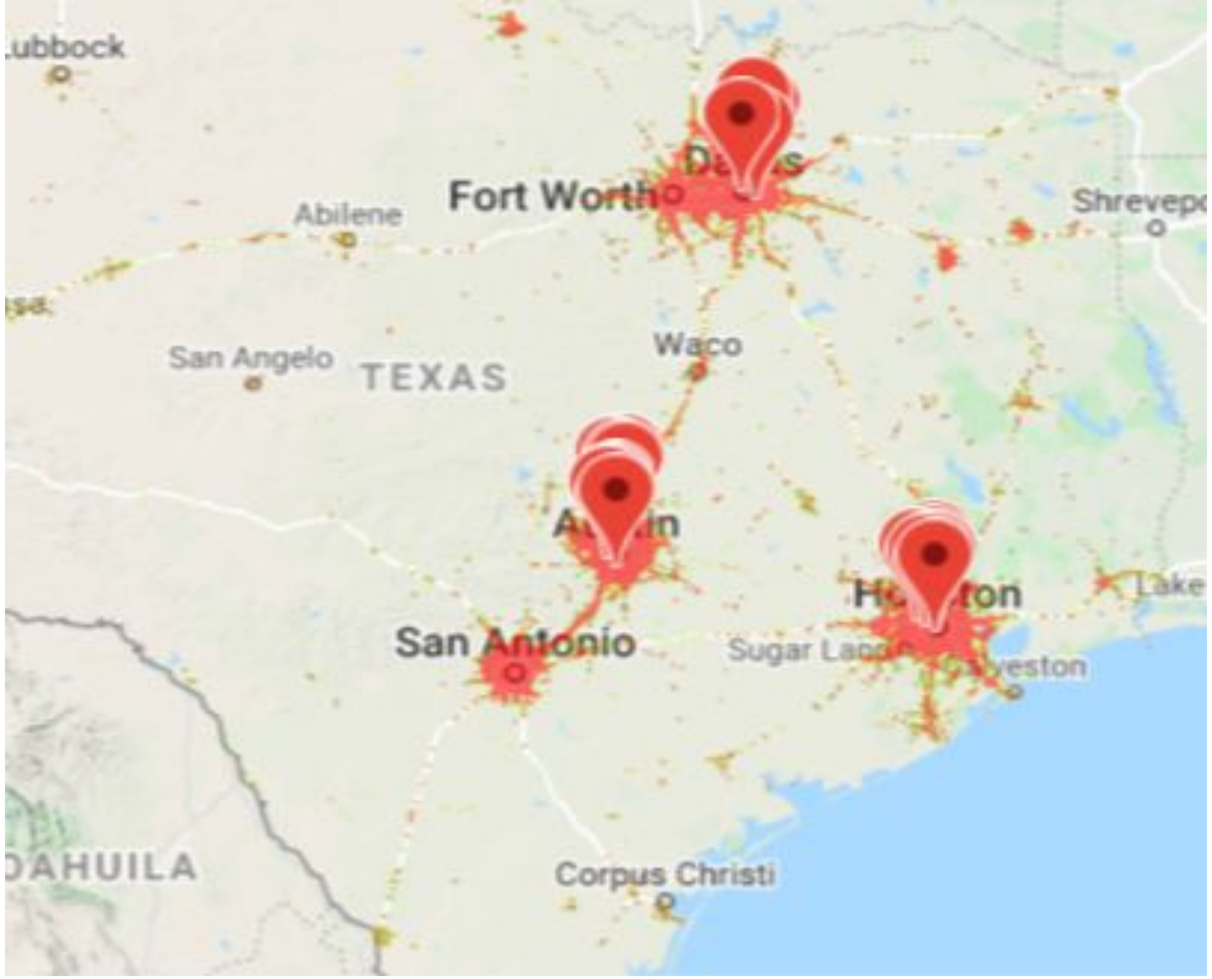
Texas urban centers represent the majority of accidents specifically along I-20, I-30 and I-45.

With such a large portion of accidents in the urban areas, population is a leading factor as chance of having an accident is increase simply because of the amount of individuals on the road

Total Accidents 2016-2020 by Texas' County



Different representation of accidents by county, showing the urban centers far outnumber the rural areas



Population Heat Map

Heat is a representation of population in texas in 2016-2020.

The markers show the highest accident rate by zip codes.

78753 or Austin Tx had the most accidents.

If the area is highly populated then accidents are more likely to occur.

Summary

Neither the weather conditions (precipitation), the time of the day, or the visibility are the main contributors to the severity of car accidents in Texas.

There is a weak correlation between precipitation and severity of accidents. Overall, accidents of all types happen at all rainfall numbers. We think traffic and population density has a higher correlation. The overwhelming majority of accidents are non-incapacitating, and less than 2% are fatal in Texas.

There is a strong correlation between population size and accident occurrence.

Most of the accidents happen in the morning to midday. But most coming from the morning time block. You could conclude that when there is more people on the roads then more accidents will occur.

The more open the space is between vehicles, and the less there is traffic, the lower the severity is. The bigger traffic becomes when an accident occurs, the more severity it will have.

Citations

https://www.kaggle.com/sobhanmoosavi/us-accidents - full data set

api census - https://www.census.gov/data/developers.html - Heat map