Data Visualization and Exploratory Data Analysis on US Accident Dataset

Shivang Bavda
shivangbavda@gmail.com

ABSTRACT

Huge amount of data is generated daily and processing the same is the ongoing challenge we are facing in the field of analytics. Generating insight from Big Data becomes difficult using standard python libraries. One of the libraries to handle the Big Data is Pyspark which helps in increasing the computational time for processing big data. We will use SQL context of Pyspark to extract useful data from spark data frame. Data Visualization with interactive control will help us explore the data in graphical format with variable inputs.

KEYWORDS

Data Visualization, Big Data Analysis, US Accident Analysis, Exploratory Data Analysis.

1 INTRODUCTION:

The Dataset US Accidents contains 4,232,541 records with 49 features. Features consist of different datatypes like string, double, int, timestamp and boolean. Since the datafile is quite big (around 1.5GB), it becomes convenient to use spark data frame for analysis and exploratory purpose. In this exploratory data analysis, we would be looking at the starting hour of the accident in a day i.e., when is the accident likely to happen over different years. We will also see how many accidents were noted per year based on the severity index. Lastly, we will find top 10 states where accident took place based on year and severity index.

The ipywidget library in python will help us to have interactive component of Jupyter Notebook for our Data visualization. Although there are many components for use in interactive notebook, we will be using slider and dropdown widget to give variable input for our graphs.

2 DATA EXAMINING AND CLEANING:

There are around 49 features to consider for the analysis. Not all features are important in answering our analysis question. We will consider only following features into consideration:

- Severity
- Start_Time
- End_Time
- Side
- City
- State
- Temperature(F)
- Visibility(mi)
- Sunrise_Sunset

Severity is integer variable which ranges from 1 to 4. Start_time and End_Time are the start and end timestamp of the accident. (US accidents data taken here for analysis ranges from 2016 to 2020). Side represents the side of the street where the accident took place (L or R). State and City are Categorical variable representing the State and City in US. Sunrise_Sunset denotes the nature of day i.e., Day or Night. Visibility and Temperature are continuous variable denoting the weather attributes at the time of accident.

Since we need some more data to answer our questions, we extracted Start_Hour from the Start_Time to denote at what hour of day the accident took place. Along with it we extracted Year from Start_Time to use it as categorical variable for some of our analysis.

When checking the NULL values in each column, we found around 137 NULL city values, 141 NULL Sunrise_Sunset values, 89900 NULL Temperature values, 98668 NULL Visibility values. Since City and Sunrise_Sunset have quite less NULL values compared to total records in dataset, we can consider removing the records from the dataset as it will not impact the analysis.

For Temperature and Visibility, we will keep the records as it is as removing those values will largely impact our analysis.

While looking at the descriptive statistics of all this variable, we got to know that Temperature is showing minimum value as -89 F and maximum value touches 203 F. We know that temperature does not go as high as 203 F or as low as -89 F in US, so we will check for outliers in the data.

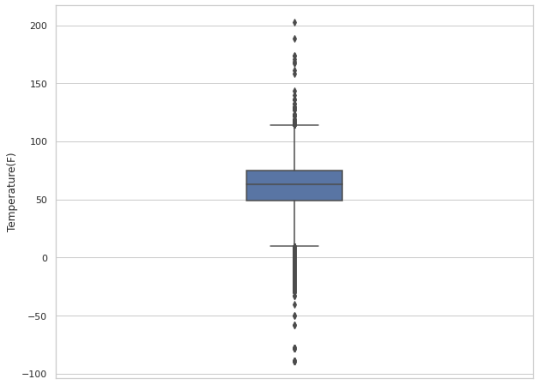

Figure 1 Outlier Detection for Temperature (F)

Since there are many outliers seen above, it becomes necessary to remove the outliers or replace them with them with median temperature. In practice, temperature imputation or modification is done by checking the temperature for a day before or after and take average of the same, but we are not sure if we have consecutive missing data or continuous obscure data. Thus, we will replace all NULL values and outliers to median value. Similarly for Visibility we will impute NULL values with the median value.

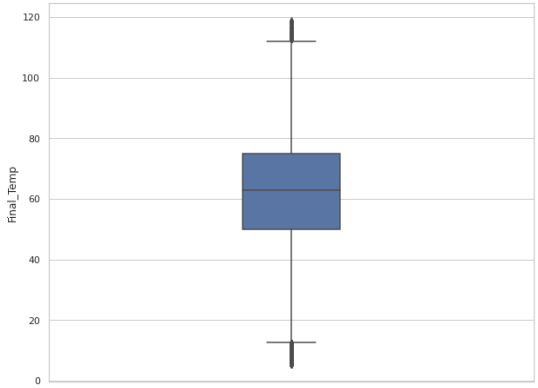

Figure 2 After Removal of Outliers

After replacing outliers to the median value, we can see in Figure 2 that the Temperature boxplot seems under acceptable temperature range in US. Let us move ahead with the exploration part with this clean data.

3 EXPLORATORY DATA ANALYSIS:

Firstly, we will plot the graph for no. of accidents taking place based on the start time of the day i.e., at what time of the day does the accidents take place each year. This graph will also give a bifurcation of the accidents based on Side of the road.

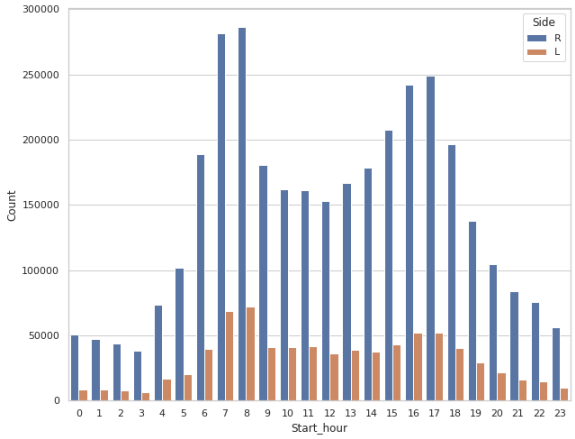

Figure 3 Overall US Accident hourly Analysis

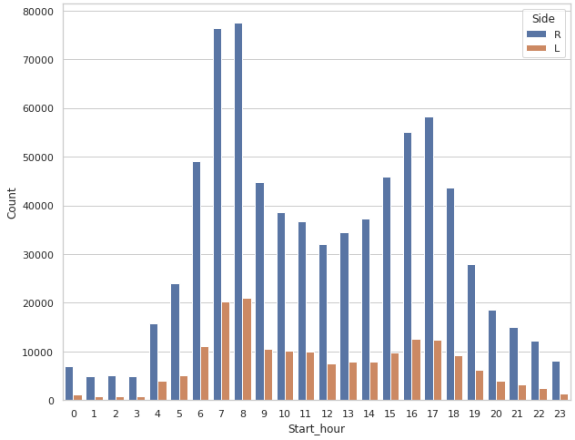

Figure 4 2019 US Accident hourly Analysis

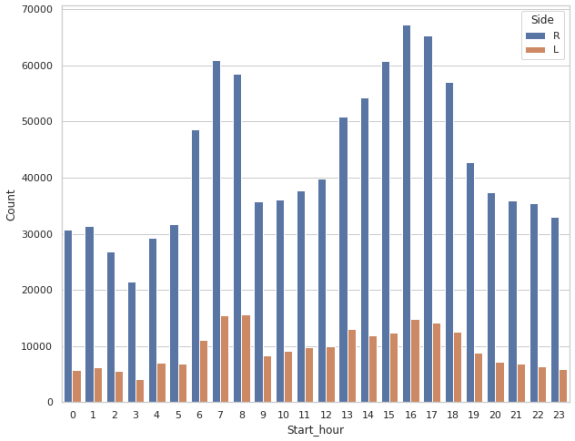

Figure 5 2020 US Accident hourly analysis

As we can see in figure 3, we can say that in all years from 2016 to 2020, most accidents took place during the start and end of office hours i.e., from 6 am to 10 am and from 3 pm to 6 pm. I think this proves the fact that the busier the road gets, the higher the chance of accidents. Also, it is interesting to note that most of the accidents take place on the right side of the road. We do not know the exact reason from the data, but we can assume that the accidents with less severity are shifted to the right side of the track so that the ongoing traffic can pass through the left lane.

When we compare 2019 and 2020 hourly analysis for US accidents, we see a pattern shift in the accident taking place in 2020. In 2020, peak of accidents still lies during the start and end of office hours, but accidents have increased drastically during nighttime compared to 2019. The reason for the same can be that due to pandemic shutdown, people were out of job and/or working from home unless essential and that led to less travelling during start and end of office hours. Another reason for increase in nighttime accidents can be that people have started driving during late hours due to frustration and depression. Mental health seems to be quite a culprit in the cause of this sudden peak.

Now let us examine the accidents taking place over year based on the severity index. As we can see in figure 6 that for Severity 2 and 4, we have increasing trend of accidents from 2016 to 2020 while for Severity 3, the no of accidents peaks in year 2018 and then decreases from there.

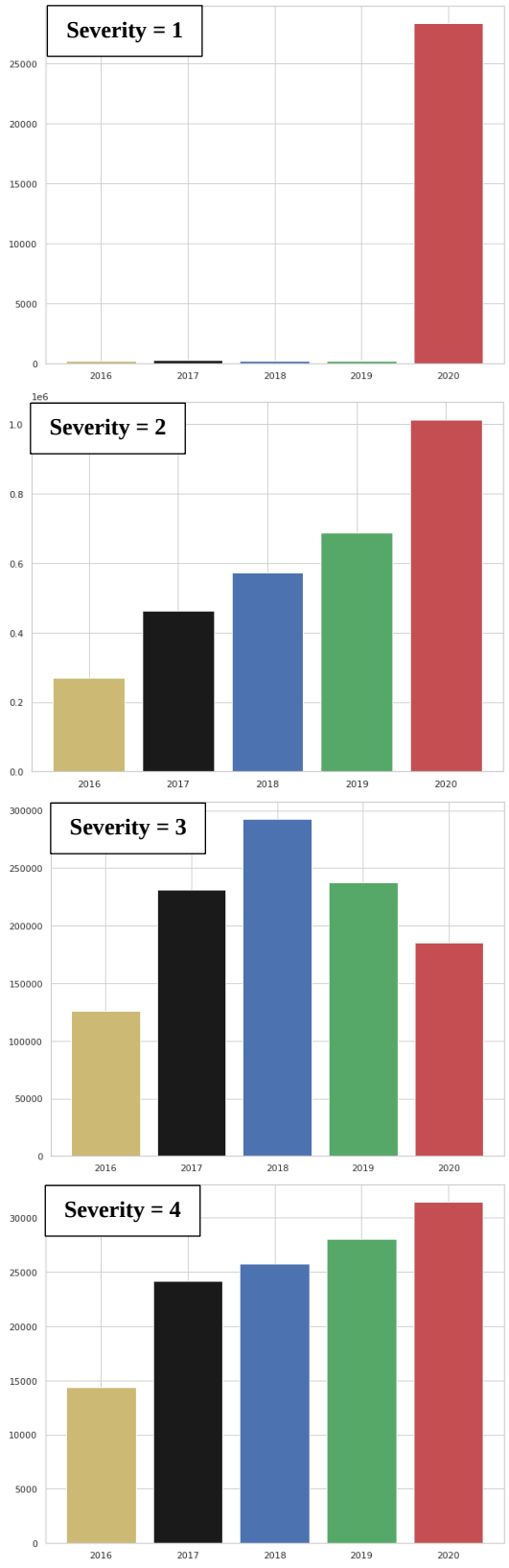

Figure 6 US Accident Severity Yearly Analysis

As per the figure 6, for we have very high no of accidents with Severity 1 (around 28K) for 2020 year while for other years it is in low 1000's. The main question which rises here is "what made the sudden rise in 2020 for Severity 1 accidents?"

Let us tackle this question by looking at what time of the day these accidents take place for the year 2020.

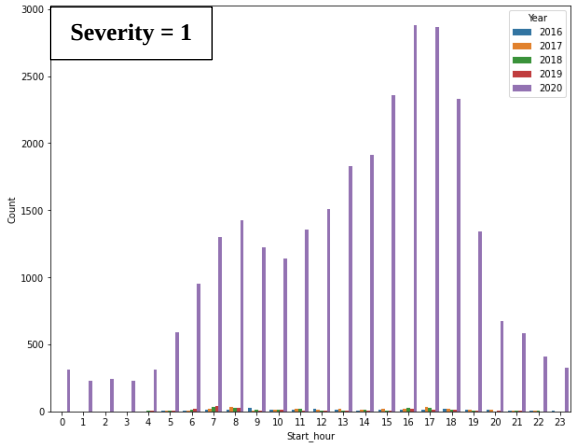

Figure 7 US Accident Hourly Analysis ALL Years

As we can see from figure 7, we can see that for the year of 2020, the no of accidents has increased drastically round the clock peaking at evening hours from 3 pm to 7 pm. While when we compare to other years, there are less than 20 accidents per hour throughout the day for 2016 to 2019.

Lastly, we will try to find top 10 states which have highest no of accidents based on severity and year. In Figure 8, we can see top 10 states with highest accident for the year 2020 and varying Severity Index. For Severity 1, we can see that Arizona(AZ) marks the highest number of accidents followed by California(CA) and Florida(FL). These 3 states are the major contributors for Severity 1 accidents in US for 2020.

For Severity 2 and 3, we see that California(CA) is showing at least 25% higher no of accidents than the second highest state in the respective Severity index. This implies that California is the highest contributors for the year of 2020 accidents.

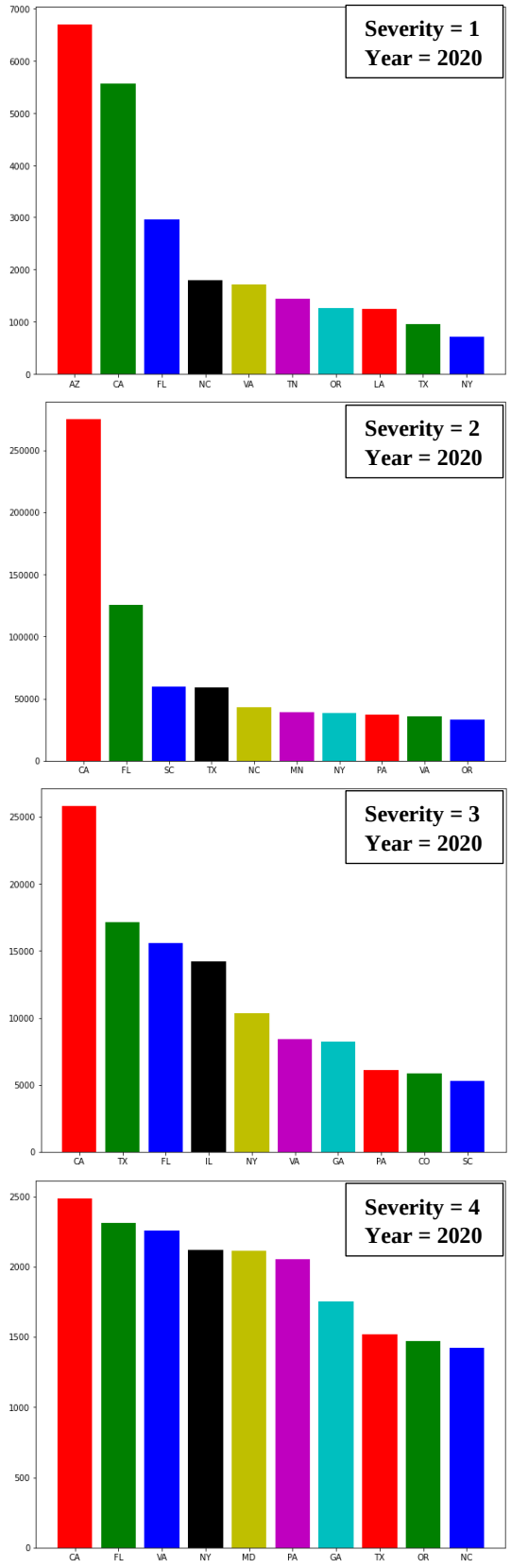

Figure 8 US Top 10 Sates with highest accident

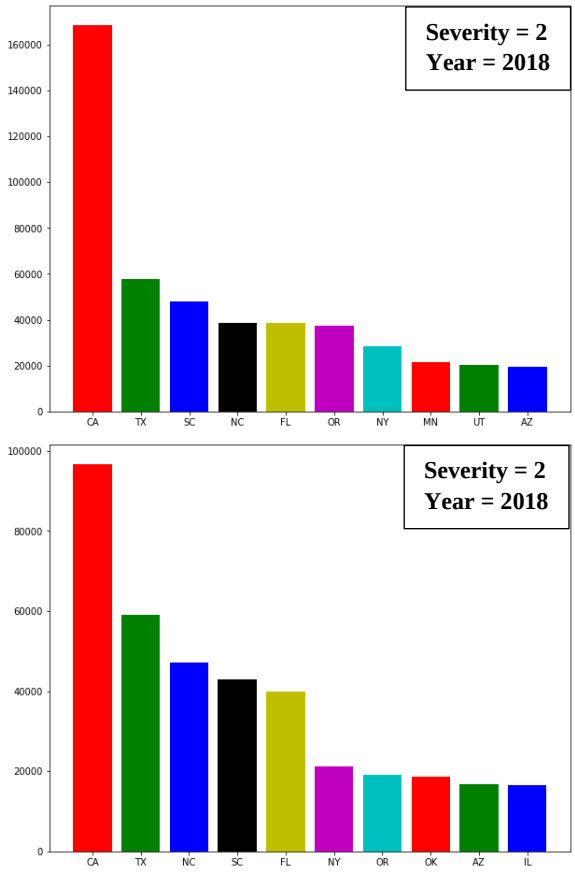

Figure 9 US Top 10 states with highest Severity 2 Accidents

We know that Severity 2 and Severity 3 accidents have far higher no of data points than Severity 1 and 4. So it is valid to look at these two severity accidents to make a general deduction. As seen in figure 8 and 9, California (CA) is having the highest no of accidents since 2018 for severity 2. Same trend has been shown for California for Severity 3.

It is interesting to note that Texas has been the second highest for Severity 2 Accidents in 2018 and 2019 but for 2020, Florida has overtaken Texas with vast amount and Texas stands fourth in the list after California, Florida, and South Carolina.

4 CONCLUSIVE INSIGHTS:

We have learned the ratio of accidents occurring on the right side of the street compared to left side of the street is high for all year and during all time of the day.

Also, we found that before 2020, most accidents took place during start and end of office hours but in 2020, due to pandemic, it seems that no of accidents have increased late at night along with daytime accidents.

We compared different severity accidents over the years and found that number of accidents have increased in 2020 for all Severity index except 2. It was also interesting to note that in 2020, Severity 1 Accidents skyrocketed compared to previous years. For Severity, we compared accident time of the day over all years and found that accident numbers for 2020 was far higher round the clock compared to previous years.

Lasty, we checked the top states in US where highest accidents took place and without any doubt, California topped the same for all severity index over the years. We did some in-depth study by breaking Severity and Year to find some more insights.

REFERENCES

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[3] Data Cleaning, https://medium.com/@aieeshashafique/exploratory-data-analysis-using-pyspark-dataframe-in-python-bd55c02a2852

[4] Interactive Notebook, https://towardsdatascience.com/interactive-controls-for-jupyter-notebooks-f5c94829aee6