FINAL PROJECT ON US ACCIDENTS ANLYSIS

BANA 6760 E01: Data Visualization

Instructor: Dr Marcus Ellis

University of Colorado, Denver

Business School

Date: 10th May 2022

CREATED BY:
Urvish Patel

Abstract

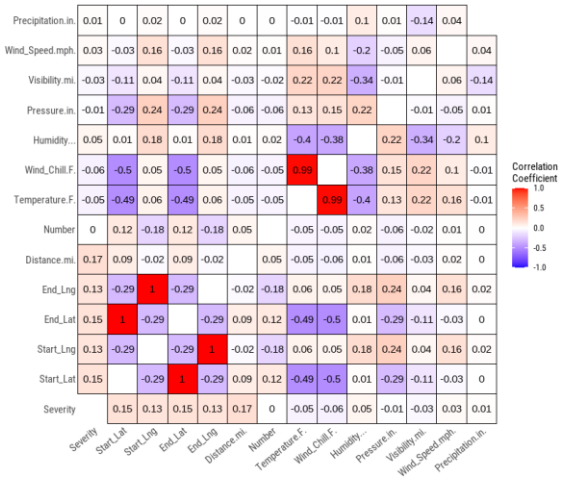
This document offers an analytical view on road accidents which helps the Department of Transportation, Road safety department, Hospitals & various response teams as they will be prepared for such accidents and will be knowing the exact measures to take beforehand an accident occurs. Specifically, this document is a summary of findings through R-code, Tableau, & Power BI which tells us statistics of the dataset through various visualizations which answer questions such as which State/Zip-code/County has the highest number of accidents? At what time/day/ do accidents usually occur in the US? What are the factors which cause road accidents? Predictions of accidents in the future. This document lays out everything which might be helpful for the safety of travelers.

Keywords: Time Series Forecasting, Predictions, Tableau Story, Geospatial visualization, Weather dependency, R-Script, Key-influencers, Decomposition Tree, Top Segments, Predictions, Exploratory Data Analysis, Regression Model, Geospatial Visualization

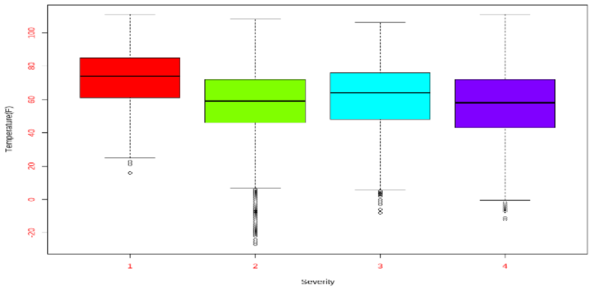
Understanding and cleaning Data

The US Accidents dataset is data reported by the relevant police department, US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks whenever an accident occurs. The dataset used covers 49 states of the USA. The accident data are collected from February 2016 to Aug 2019. They note down the exact location of the accident and the distance of traffic which is affected after an incident occurs. The very first thing which the Department of Transportation wants to know is the severity of an accident and how frequently an accident occurs during the day/week/time. How response teamwork can has efficiently given these circumstances? Weather plays a severe role in any type of accident. This analysis discloses the facts and relations about accidents that might be directly linked to the type of weather condition during an accident.

The very basic steps of EDA analysis are to first approach the data and analyze categorical, numerical, & both at the same time. After loading the dataset into R, all the values with “NA” are removed. As this may become outliers in future analysis. After removing NA from the original dataset which consisted of 1516064 rows shorted to 335552 rows. Now, there are instances where your data rows are repeated, the reason might be a machine error. So we check if we have any repeated rows. Understanding the summary of the data frame helps us to know which columns are of what data type. Describing the function of R gave me the statistics of each column. It shows mean, standard deviation, skewness, Kurtosis & Inter Quartile Range. R has this amazing function by which we can generate reports. After doing all the things mentioned above, I generated a normality report, which shows the normality plot and along with the plots if we transform the columns into a log or squared value. The correlation matrix is the best way to find out which columns are dependent in pairs. The below image shows us the same.



Now let’s see if each severity is having a normal distribution or not. To prove this I created a Box Plot which is shown below.



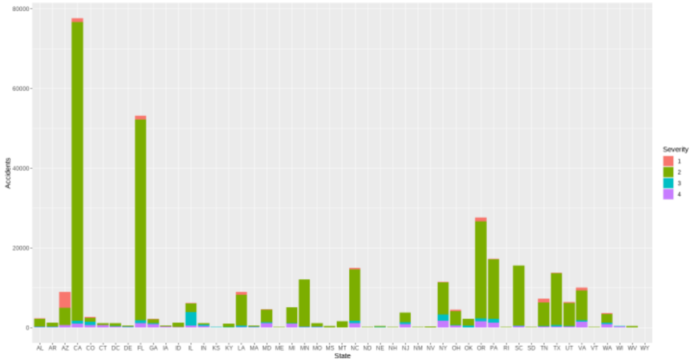
From this, we can conclude that the Severities of 1, 2, and 3 & 4ares normally distributed. As I can see there is hardly any skewness. The points shown are the outliers but these may not be considered outliers as these can be useful in future analysis.

After plotting a correlation matrix I converted Boolean data types (True, False) to 1 and 0 so that I could build a Logistic regression model for knowing the factors which cause accidents.

Common Visualizations

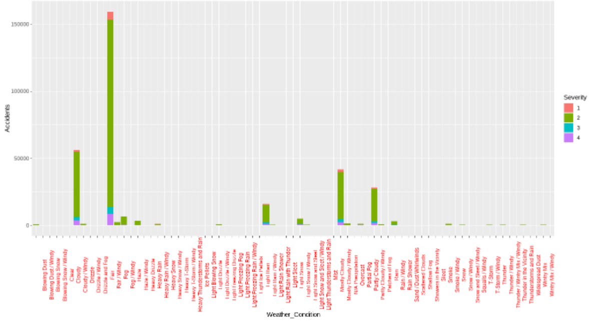
After going through the cleaning part I created some visualizations which answer the following questions.

1. What is the count of accidents along with their severity in each state?



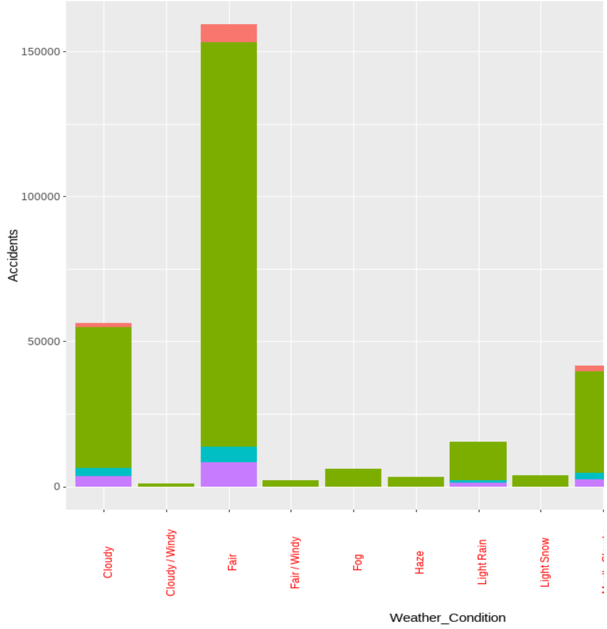
This shows CA with the highest number of road accidents with a severity of 2. We can also conclude that most of the accidents are of severity 2.

2. In which Weather Condition do accidents occur most?



Most of the accidents occur when the weather is fair. Apart from that accidents also occur when the weather is cloudy, windy, and having light rain. Now, this doesn't give us a clear view, this visualization is not good enough, so for that reason, I created a plot with the same weather conditions on the x-axis but where the accidents are greater than 1000. This answer's the following question.

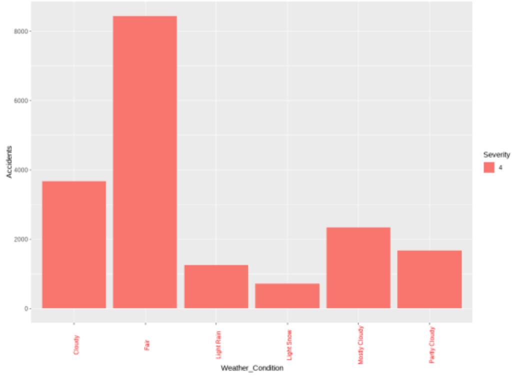
3. In which Weather Condition do accidents occur most where accidents are greater than 1000?



Now this shows us a clear picture. Sometimes it is hard to read any visualization when it contains more space.

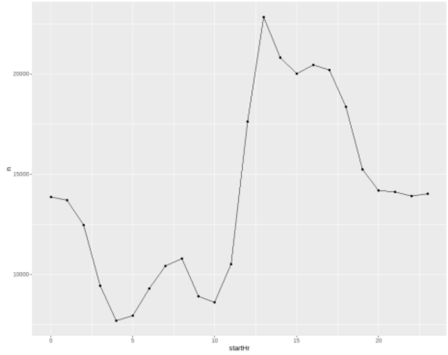
Now the area of focus is Severity 4, as these are the most severe accidents. This answers the following question.

4. In which Weather Condition do accidents occur most where accidents are greater than 1000 and have a severity of 4?



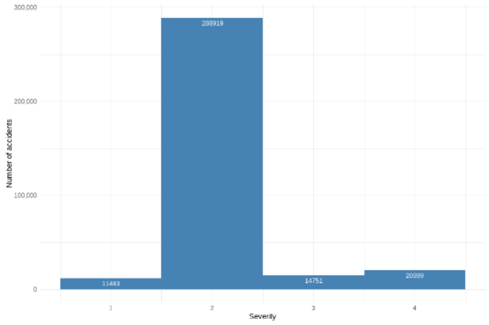
This concludes most accidents occur during fair weather conditions. But the total sum of all other accidents with weather conditions surpasses i.e. Accidents occur most when the weather is Cloudy, have a light rain & snow.

5. Usually at what hour do most accidents occur in the US?



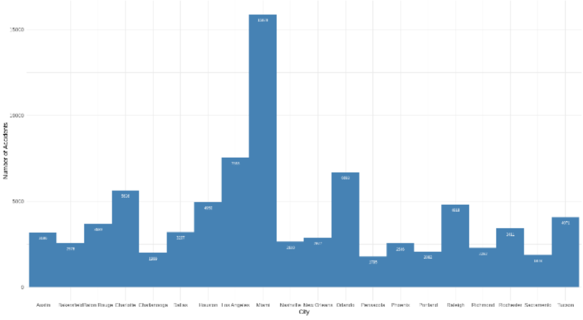
According to this plot, we can conclude that most accidents occur between 11 hrs. To 15 hrs.

6. What type of severity accidents mostly occur?

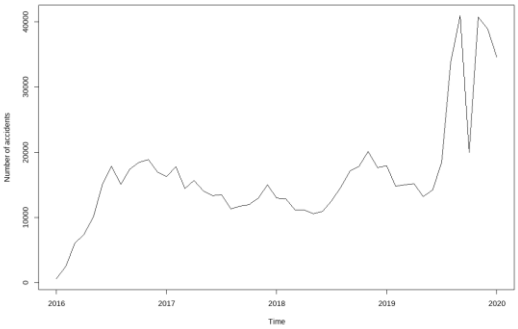


The severity of type 2 has the highest number followed by severity 4, 3, & 1.

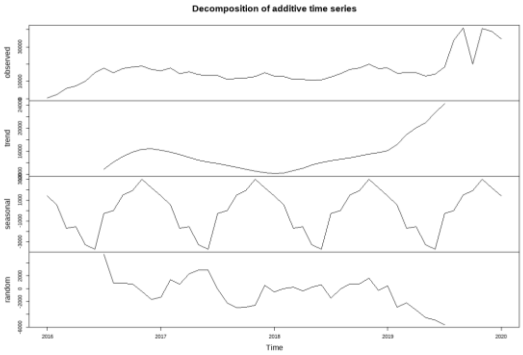
7. What city has the highest number of accidents?



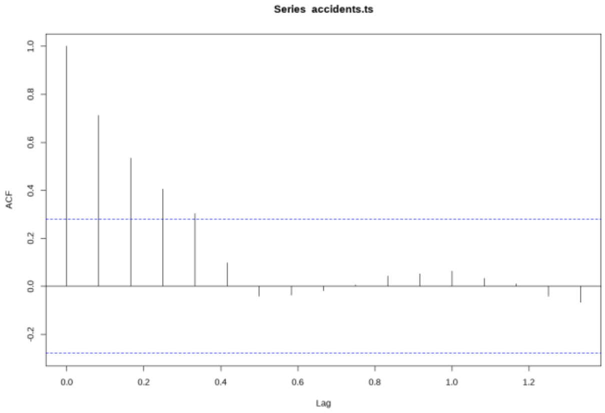
Miami has the highest number of accidents followed by Los Angeles, Orlando, and many others.

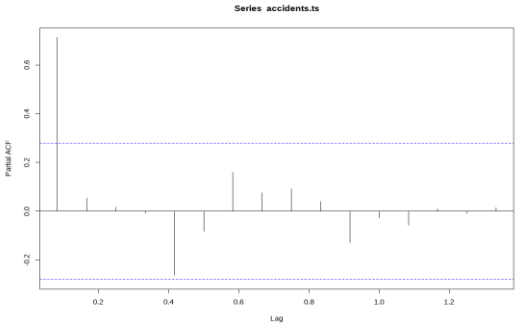


8. What would be a time series plot for several accidents?
   We can see that the trend number of accidents that occur each year is a positive trend. To see clearly what type of trend is there both seasonally and randomly I plotted a decomposition plot, which looks like the image below.

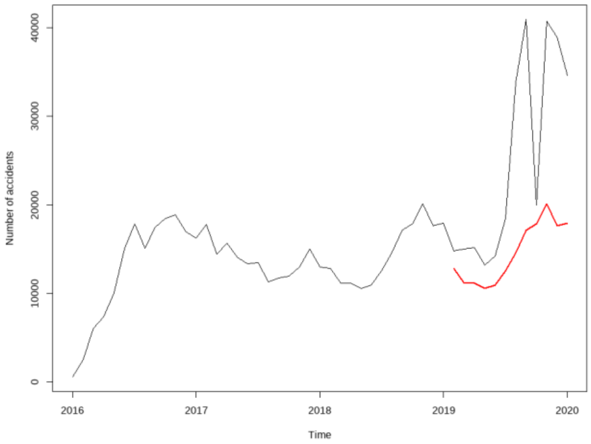


From this plot, we can conclude that the trend is positive if we see the trend seasonally it is also a positive trend, but the randomly selected data is showing a negative trend. I have also created ACF & PACF plots for knowing the autocorrelation and the type of trend.



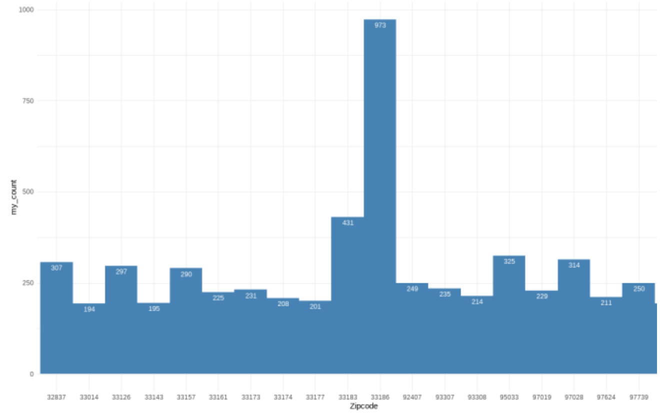


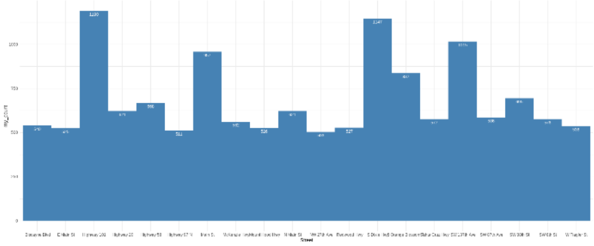
9. What are the predictions if we use the naïve model if we divide the dataset into train and test data?



From this, we can conclude how good our naïve model is we can see the predictions in the red line and can compare them with the actual dataset.

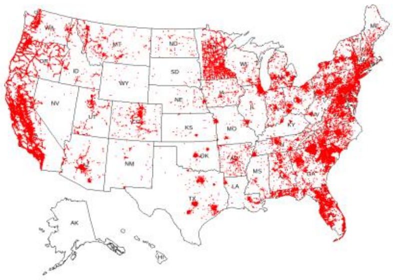
10. Which Zip code and street in the US has the highest number of accidents?





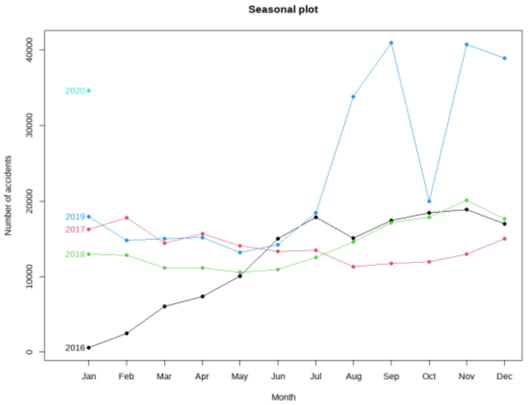
From this, we can conclude that Zipcode with 33183 & Highway 101 has the highest number of accidents. This area can be declared an accident-prone zone and we can provide a better response team. We can improve the roads and Lights.

11. General view on the number of accidents on the US map.



From this, we can conclude that most of the accidents reported are from either the east side or the west side.

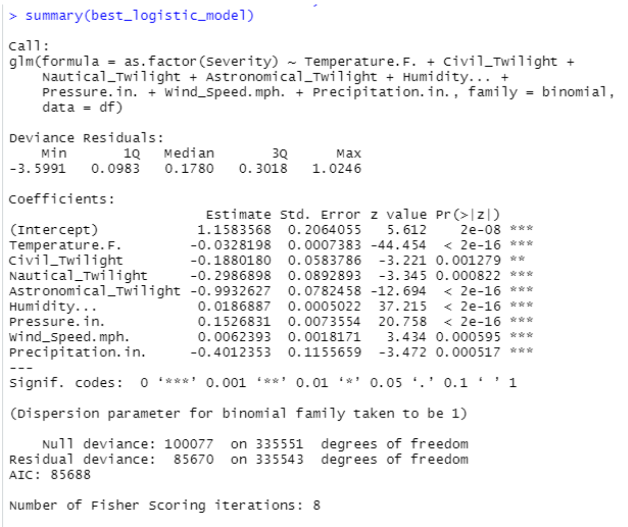
12. What is the comparison between different years concerning the count of accidents?



This shows the monthly number of accidents concerning the year. We can see the trends in these lines. Surely, in 2016 there was a positive trend. The year 2017 may be concluded as a negative trend as the count of accidents this year is in decreasing numbers.

Logistic Regression Model

From this model, we will be able to identify the actual causes of road accidents. The sole objective of this model is to find out factors that affect the severity of road accidents. Upon using and trying many models I found one bet model with a good P-value. Following is the image of the Logistic Regression model.



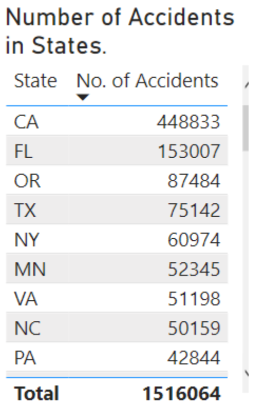
```
> summary(best_logistic_model)

Call:
glm(formula = as.factor(Severity) ~ Temperature.F. + Civil_Twilight +
    Nautical_Twilight + Astronomical_Twilight + Humidity... +
    Pressure.in. + Wind_Speed.mph. + Precipitation.in., family = binomial,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5991   0.0983   0.1780   0.3018   1.0246

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.1583568  0.2064055   5.612    2e-08 ***
Temperature.F.        -0.0328198  0.0007383 -44.454  < 2e-16 ***
Civil_Twilight        -0.1880180  0.0583786  -3.221 0.001279 **
Nautical_Twilight     -0.2986898  0.0892893  -3.345 0.000822 ***
Astronomical_Twilight -0.9932627  0.0782458 -12.694  < 2e-16 ***
Humidity...            0.0186887  0.0005022  37.215  < 2e-16 ***
Pressure.in.           0.1526831  0.0073554  20.758  < 2e-16 ***
Wind_Speed.mph.        0.0062393  0.0018171   3.434 0.000595 ***
Precipitation.in.     -0.4012353  0.1155659  -3.472 0.000517 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 100077  on 335551  degrees of freedom
Residual deviance:  85670  on 335543  degrees of freedom
AIC: 85688

Number of Fisher Scoring iterations: 8
```

From this we can conclude that the Severity of an accident mostly depends on Temperature, Civil Twilight (Beginning in the morning or ending in the evening), Humidity, Pressure, Wind speed, and precipitation. So it is most like that an accident might occur if these conditions are fulfilled.

Visualizations through PowerBi

Visualizations that answer the following questions.

1. What is the count of accidents in each state?

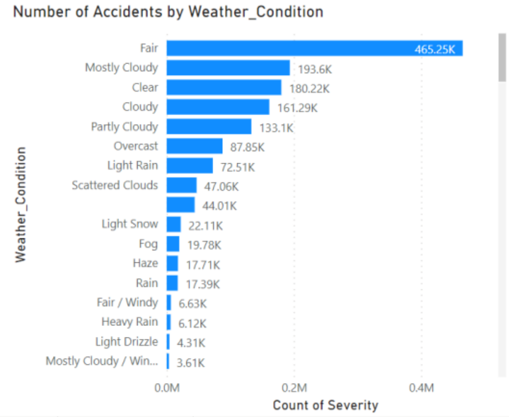


Number of Accidents in States.

State	No. of Accidents
CA	448833
FL	153007
OR	87484
TX	75142
NY	60974
MN	52345
VA	51198
NC	50159
PA	42844
Total	**1516064**

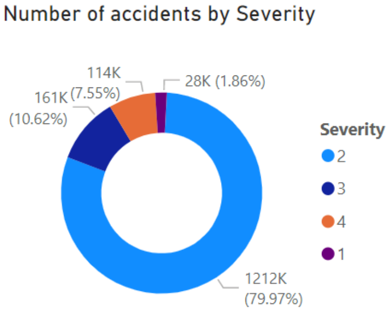
This table represents the number of accidents that happened in each state from highest to lowest. Sometimes a simple table is enough for giving an insight. After all the understanding of data is dependent on these numbers.

2. In which Weather Condition do accidents occur most?



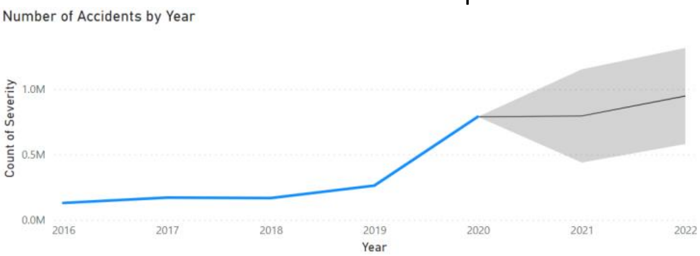
This is a bar graph that represents that most accidents occur when the weather is fair. Apart from that, accidents also occur when the weather is mostly cloudy, windy, and rainy. This is a very simple way to represent the number of numerical columns of a dataset.

3. Usually at what hour do most accidents occur in the US?
   According to this, the plot we can conclude that most accidents occur between 15 hrs. To 17 hrs. This visualization is generated through the R script. Now I had to use R script because it was the easy way as power BI did not provide me ease to build this particular visualization. The column of Start_time which had both date and time did not generate a time hierarchy, but the tableau does this with ease.
4. What type of severity accidents mostly occur?



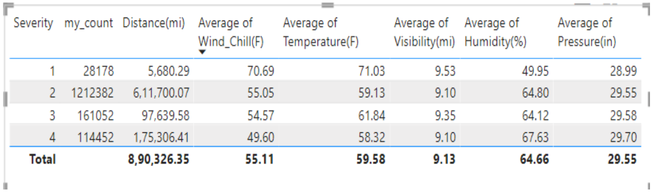
This is a pie chart that represents the number of accidents by Severity. We can conclude from this visual that, Severity of type 2 has the highest number followed by severity 4, 3, & 1.

5. What are the predictions for the future? What number of accidents we can expect?



This is a time series forecasting graph. From this, we can say that the expected number of accidents in the future is likely to increase according to the trend. This prediction is generated not by year but by considering month-wise.

6. What are the number of accidents that happened according to severity while considering the average temperate, humidity, pressure, Visibility, and wind chill?

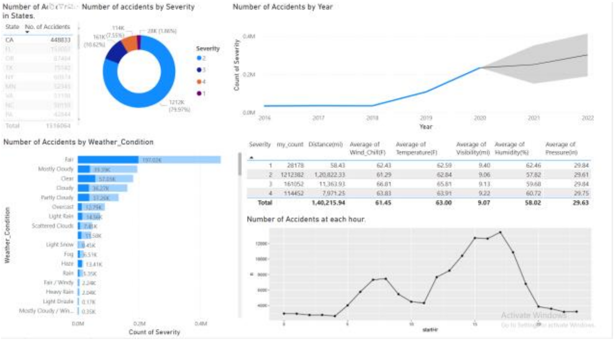


Severity	my_count	Distance(mi)	Average of Wind_Chill(F)	Average of Temperature(F)	Average of Visibility(mi)	Average of Humidity(%)	Average of Pressure(in)
1	28178	5,680.29	70.69	71.03	9.53	49.95	28.99
2	1212382	6,11,700.07	55.05	59.13	9.10	64.80	29.55
3	161052	97,639.58	54.57	61.84	9.35	64.12	29.58
4	114452	1,75,306.41	49.60	58.32	9.10	67.63	29.70
Total		**8,90,326.35**	**55.11**	**59.58**	**9.13**	**64.66**	**29.55**

From this, we can conclude that most accidents occur with a severity of 2 with an average of respective weather statistics. This also gives us information on how much the total distance of traffic is affected by their respective severity number. A total of 890326.35 Miles of traffic is affected due to all the accidents.

Now, these same visualizations can also be made in R as I did before. But the usage of tools like power BI is it makes everything easy. Now power Bi gives the option to select and filter data. This can help us to answer the following questions.

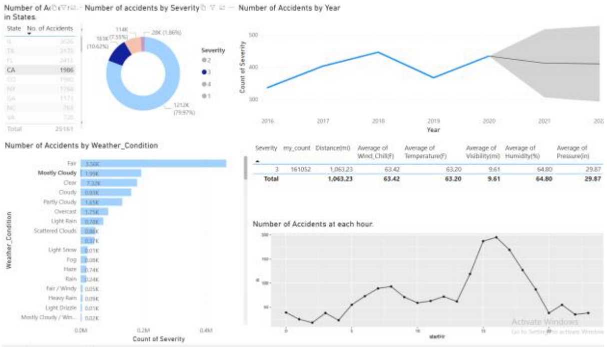
7. In which weather condition do accidents occur most given that a particular state with the highest accident count & what would be the average weather statistics according to the Severity? What will be the prediction o the number of accidents in the future for that particular state?



Severity	my_count	Distance(mi)	Average of Wind_Chill(F)	Average of Temperature(F)	Average of Visibility(mi)	Average of Humidity(%)	Average of Pressure(in)
1	28178	58.43	62.43	62.59	9.40	62.46	29.84
2	1212382	1,20,822.33	61.29	62.84	9.06	57.82	29.61
3	161052	11,363.93	66.81	65.81	9.13	59.68	29.84
4	114452	7,971.25	63.83	63.91	9.22	60.72	29.75
Total		**1,40,215.94**	**61.45**	**63.00**	**9.07**	**58.02**	**29.63**

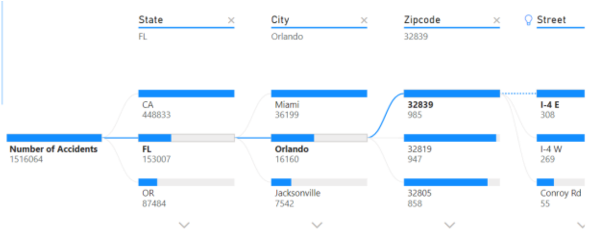
We can see that State CA is selected and all the values in visualizations are changed according to the selected data.

8. Now, What if I wanted to know the number of accidents at each hour or severity of 3 where the weather is Mostly Cloudy? And what would be the prediction for future years?


Number of Accidents in States
Number of accidents by Severity
Number of Accidents by Year
Number of Accidents by Weather_Condition
Number of Accidents at each hour
Severity
Count of Severity
Year
Weather_Condition

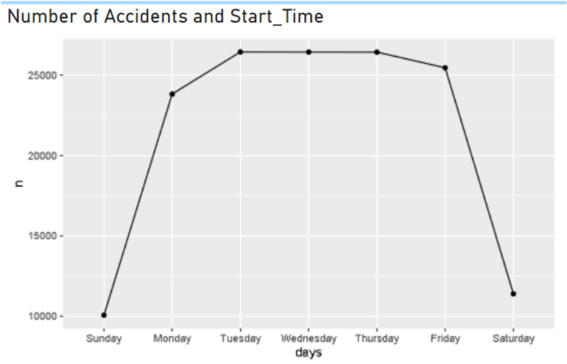
We can see that for the following conditions how our visualizations changed. We can conclude that under the same conditions the number of accidents will decrease in near future.

9. What is the number of accidents according to the state, city, zip co, de, and street?



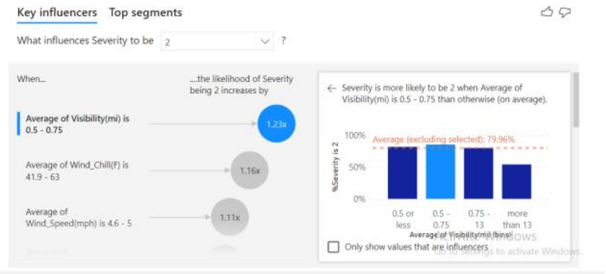
This is a decomposition tree. This represents the number of accidents in a hierarchal order. Suppose, you want to know the statewide h highest number of accidents, just click on Number of Accidents, then you want to know which city from that state has the high number of accidents. You click on the state CA, you get, sorted cities, and so on till you want to know which street has the highest number of accidents.

10. During weekday days accidents occur most?



This is also an R script. This tells us that there is less number of accidents on weekends as compared to the regular weekdays.

11. On what weather factor does the severity of the accident depends?


Key influencers Top segments

What influences Severity to be 2 ?

When... ...the likelihood of Severity being 2 increases by

Average of Visibility(mi) is 0.5 - 0.75 1.23x

Average of Wind_Chill(F) is 41.9 - 63 1.16x

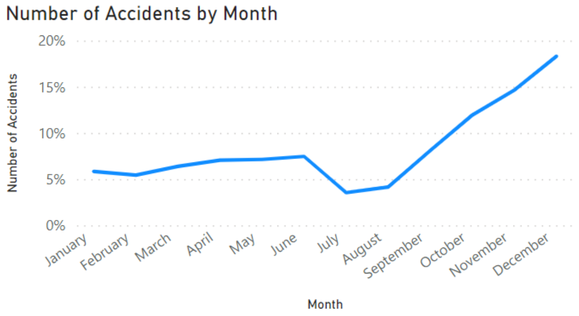
Average of Wind_Speed(mph) is 4.6 - 5 1.11x

Severity is more likely to be 2 when Average of Visibility(mi) is 0.5 - 0.75 than otherwise (on average).

□ Only show values that are influencers
100%
50%
0%
%Severity is 2
Average (excluding selected): 79.96%
0.5 or less
0.5 - 0.75
0.75 - 13
more than 13
Average of Visibility(mi) (bins)

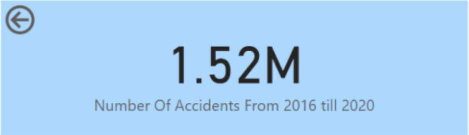
This visualization is known as Key influencers. The key influencers' visual helps us to understand the factors that drive a metric we're interested in. It analyzes our data, ranks the factors that matter, and displays them as key influencers. For example, the above image shows what influences the Accident with a Severity of 2? The answer to that is. When the visibility is between 0.5 and 0.75 the likelihood of having an accident with a severity of 2 increases by 1.23 X. The same type of interpretation can be done for all the influencers mentioned in the visual box.

12. In what month did the most accidents occur?



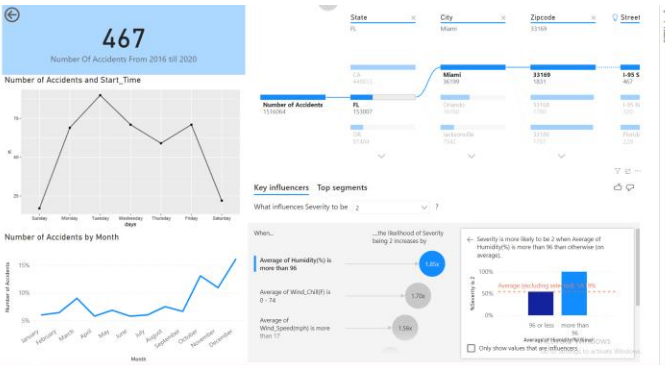
This is a line chart that shows most of the accidents occur in October, November, & December.

13. What is the Total number of accidents for the dataset?



This is a card that holds theatrical value. In this case, it shows the number of accidents.

14. In this too we can filter the data if we select something for example if I wanted to know what weather factors affect the severity of 2 for the highest number of accidents on the street of Florida State, and at what month and day most accidents occur, for the same state?

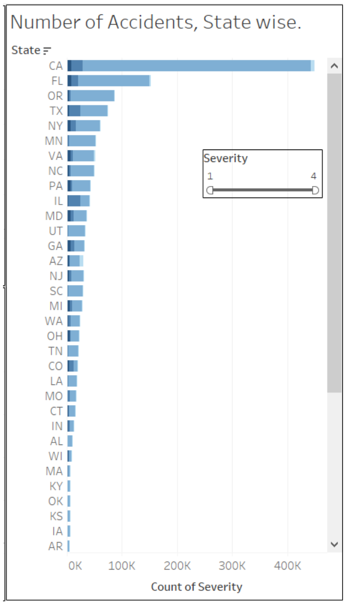

467
Number Of Accidents From 2016 till 2020
Number of Accidents and Start_Time
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
days
Number of Accidents by Month
15%
10%
5%
Number of Accidents
January
February
March
April
May
June
July
August
September
October
November
December
Month
State
City
Zipcode
Street
Number of Accidents
1516064
FL
153007
Miami
36199
33169
1831
I-95 S
467
Key influencers
Top segments
What influences Severity to be 2
When...
...the likelihood of Severity being 2 increases by
Average of Humidity(%) is more than 96
1.86x
Average of Wind_Chill(F) is 0 - 74
1.70x
Average of Wind_Speed(mph) is more than 17
1.56x
Severity is more likely to be 2 when Average of Humidity(%) is more than 96 than otherwise (on average).
100%
50%
0%
%Severity is 2
96 or less
more than 96
Average of Humidity(%) (bins)
Only show values that are influencers

So according to the selected decomposition tree of Florida State, until 2020 there were 467 accidents and the most accident occurred on Tuesdays, and in December. Now the above image also shows what affect the accident severity of 2. When humidity is more than 96, the chances of having an accident with severity 2 increases by 1.86x.

Tableau Visualizations

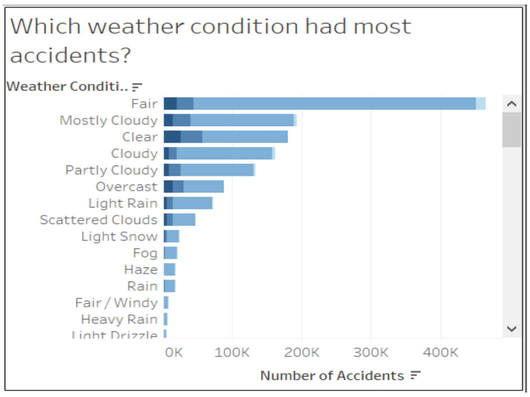
Visualizations that answer the following questions.

What is the count of accidents in each state?



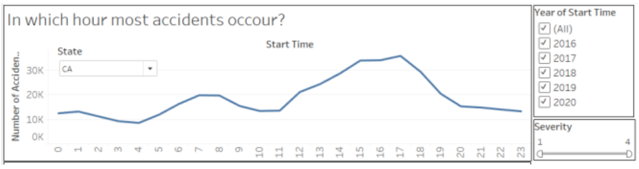
This bar graph represents the number of accidents that happened in each state from highest to lowest. Sometimes a simple graph is enough for giving an insight. You can also note that this graph also determines the number of accidents according to the level of severity. The lightest color is blue on level one & darkest on level 4.

1. In which Weather Condition do accidents occur most?



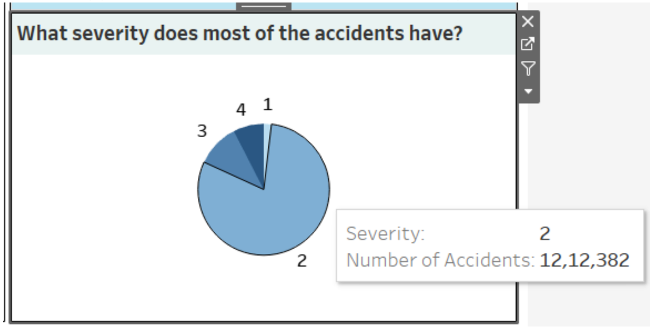
This is a bar graph that represents that most accidents occur when the weather is fair. Apart from that, accidents also occur when the weather is mostly cloudy, windy, and light rain. This is a very simple way to represent the number of categorical columns of a dataset. You can also not that this graph also determines the number of accidents according to the level of severity lightest test color blue is on level one & the darkest on level 4.

2. Usually at what hour do most accidents occur in the US?



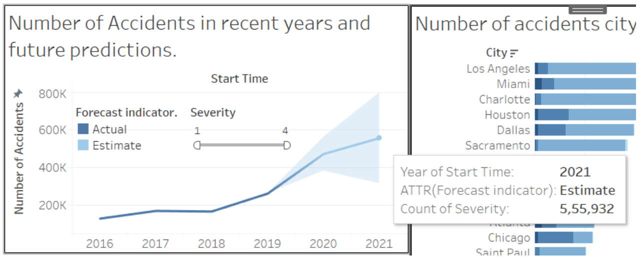
According to this plot, we can conclude that most accidents occur between 15 hrs. To 17 hrs. The column of Start_time which had both date and time did not generate a time hierarchy in power bi, but tableau does this with ease. You can also filter data according to severity & year.

3. What type of severity accidents mostly occur?

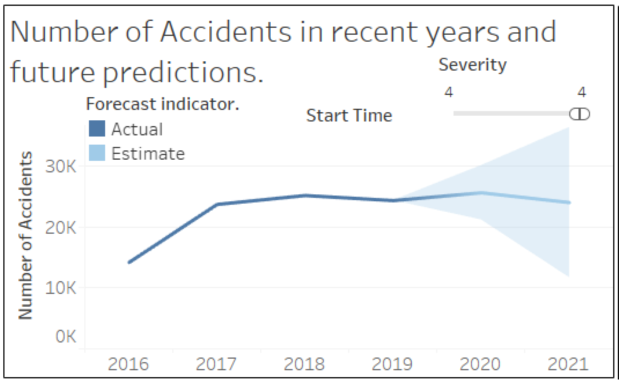


This is a pie chart that represents the number of accidents by Severity. We can conclude from this visual that, Severity of type 2 has the highest number followed by severity 4, 3, & 1.

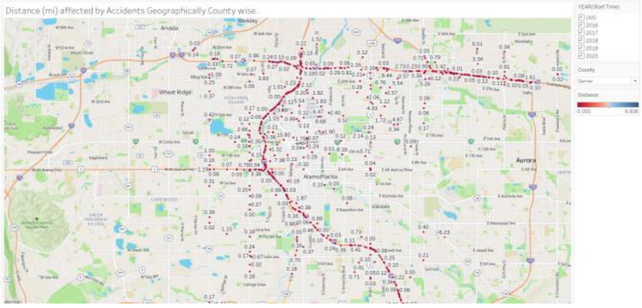
4. What are the predictions for the future? What number of accidents we can expect?



This is a time series forecasting graph. From this, we can say that the expected number of accidents in the future is likely to increase according to the trend. Now we can also filter the data according to the severity. If you select a severity of 4 the forecast will predict the accident count for severity of 4. Below is the image for the same.

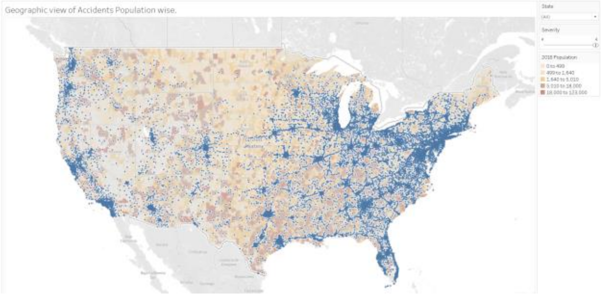


5. What is the Distance (mi) affected by Accidents Geographically?


Distance (mi) affected by Accidents Geographically County wise.
YEAR(Start Time)
County
Denver
Distance
Arvada
Wheat Ridge
Aurora

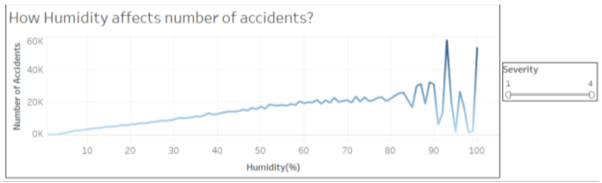
The above image shows what and where the most distance is affected when an accident occurred in the county of Denver. We can also do it for another county by using filters right beside it. This visualization was generated by making a new measure named "Distance" which had a calculated field that determined the distance between 2 longitude and latitude values in miles.

6. What is the overview? In which region does r of accidents occur in the USA?



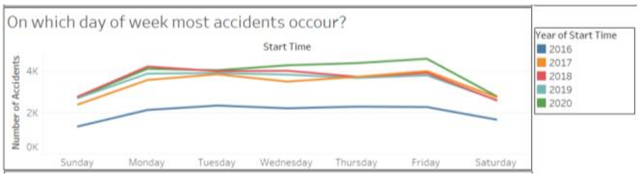
The above image gives you an overview of where most accidents occur? This also gives you an insight on population, does the number of accidents depend on the population of a county? We can sure compare that using this graph.

7. How the weather conditions like humidity affect the number of accidents?



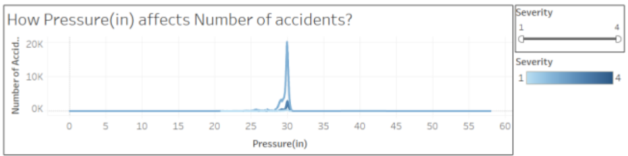
This image answers the question. You can also try changing the severity using the filter given on the right-hand side.

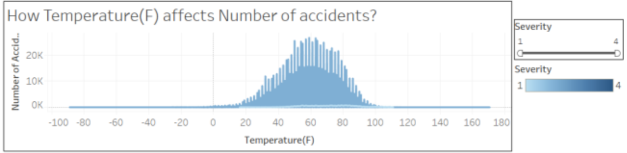
8. On which day of the week do most accidents occur?

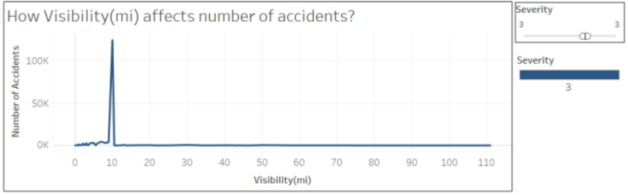


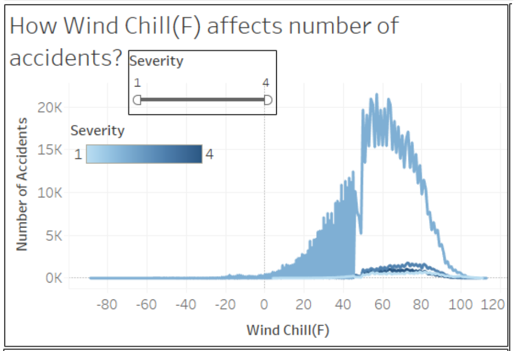
This image shows the number of accidents that occurred on each day of the week. You can compare this year to year. Try hovering on one of the lines to see more details.

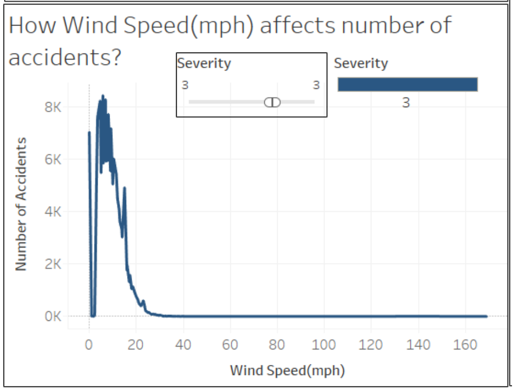
9. How the weather conditions like pressure, temperature, visibility, wind chill, & wind speed affect the number of accidents?





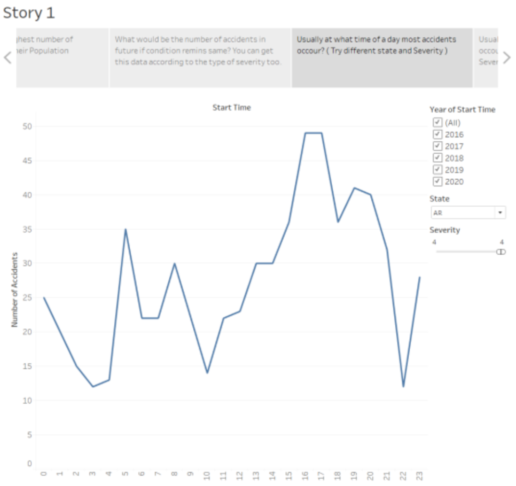






This is dashboard 5, this answers all the questions mentioned above. We can also filter the data using filters given on the right-hand side. From this, we can directly compare and predict the number of accidents that might occur if we know these weather conditions.

Now Tableau has several features and one of them is the story. I have created a story that shows the same as a dashboard but it makes it easy for the user to understand. Below is just one example, you can have a look at the tableau file for more insights.



Story 1

ghest number of
eir Population

What would be the number of accidents in future if condition remins same? You can get this data according to the type of severity too.

Usually at what time of a day most accidents occour? (Try different state and Severity)

Usual
occou
Sever

Year of Start Time
☑ (All)
☑ 2016
☑ 2017
☑ 2018
☑ 2019
☑ 2020

State: AR

Severity: 4 – 4

Factors Affecting Road Traffic Accidents

- Driver Behavior. Alcohol and drug use, reckless operation of the vehicle, failure to properly use occupant protection devices, the use of cell phones or texting, and fatigue.
- Roadway characteristics. Road geometries and roadside conditions, such as well-designed curves and grades, wide lanes, adequate sight distance, clearly visible striping, flared guardrails, good quality shoulders, roadsides free of obstacles, well-located crash attenuation devices, and well-planned use of traffic signals.
- Vehicle factors. Vehicle type, and the engineering and the safety design standards for vehicle performance. For example, the design of windshield glass and the location and durability of gas tanks can increase safety. Passenger protection systems in vehicles (i.e. airbags, safety belts), if used, can eliminate injuries or reduce their severity.
- Traffic volumes. Average annual daily traffic (AADT) or the vehicle miles traveled (VMT). AADT is the average number of vehicles passing a point along a particular road section each day. Thus, AADT represents the vehicle flow over a road section on an average day of the year. VMT refers to the distance traveled by vehicles on roads. It is often used as an indicator of traffic demand and is commonly applied to evaluate mobility patterns and travel trends.
- Time factors. The season of the year, the month of the year, weekdays, and the hour of crash occurrence.

Possible Solutions to the Problems.

- The response teams and hospitals must be given special provisions in the hours in which most accidents occur.
- Warning signs about speed limits are to be put on the accident-prone streets.
- The state with the highest accidents must be provided with better resources and budget plans to avoid accidents and rescue the victims.
- Warnings are to be put depending on the weather conditions which cause accidents.
- A mandate for vehicles to have a first aid kit should be passed.
- Online surveillance for a prompt response from emergency services should be implemented.
- To have enough response teams to rescue in accident-prone locations

References

https://zebrabi.com/top-power-bi-dashboard-tips-and-tricks/

https://databear.com/3-hidden-tricks-in-power-bi/

https://docs.microsoft.com/en-us/power-bi/create-reports/desktop-tips-and-tricks-for-creating-reports

https://docs.microsoft.com/en-us/power-bi/visuals/power-bi-visualization-influencers

https://www.youtube.com/watch?v=i1kCHZEhnEY

https://www.youtube.com/watch?v=DWjbd7NPUCM

https://powerbi.microsoft.com/en-in/partner-showcase/bizone-traffic-accident-analytics/

https://www.averanalytics.com/accident-dashboard

https://medium.com/analytics-vidhya/visual-outputs-in-r-example-1-d71cb4be50eb

https://youtu.be/HPJn1CMvtmI

https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html

http://r-statistics.co/ggplot2-Tutorial-With-R.html

https://towardsdatascience.com/exploratory-data-analysis-in-r-for-beginners-fe031add7072

https://blog.datascienceheroes.com/exploratory-data-analysis-in-r-intro/

Dhakal, 2018. *A Naïve Approach for Comparing a Forecast Model*.

Paul S.P. Cowperthwait, 2009. Introductory Time Series with R (Use R!)

Hrishi V. Mittal, 2011. R Graphs Cookbook

https://www.flerlagetwins.com/2020/07/obscure-tips.html

https://www.tableau.com/about/blog/7-tips-and-tricks-dashboard-experts

https://www.youtube.com/watch?v=CAZ3IAJEuCI

https://www.youtube.com/watch?v=rnP7zP8J_7g

https://www.youtube.com/watch?v=aHaOIvR00So

https://youtu.be/gWZtNdMko1k