SOULPAGE

IT SOLUTIONS

Washington Metropolitan Area Transit Authority Analysis

Washington Metropolitan Area Transit Authority Analysis

Problem statement:-

The Washington Metropolitan Area Transit Authority (Metro) was created by an interstate compact in 1967 to plan, develop, build, finance, and operate a balanced regional transportation system in the national capital area. Metro began building its rail system in 1969, acquired four regional bus systems in 1973, and began operating the first phase of Metrorail in 1976. Today, Metrorail serves 91 stations and has 117 miles of track. Metrobus serves the nation's capital 24 hours a day, seven days a week with 1,500 buses. Metrorail and Metrobus serve a population of approximately 4 million within a 1,500-square mile jurisdiction. Metro began its paratransit service, MetroAccess, in 1994; it provides about 2.3 million trips per year.

The Washington Metropolitan Area Transit Authority wants to know some insights from its old data recorded. We have data from 2012 to 2016 recorded with seven features and 26785 records of problems faced in different tracks due to different problems. Each problem has a different time durations to clear it.

Opportunity:-

We have several records of data available to perform some analysis and few questions can be answered such as

- In which direction most of the metro moves in?
- What are the available tracks, which track is the busiest one?
- In which location we had more problems?
- What kind of problems does WMATA metro face?
- Which problem does occur most of the time?
- How much delay does it take for each problem?
- What is the amount of time does most of the problems get resolved?

Tools used:-

1. Python 3.6
2. Numpy
3. Pandas
4. Matplotlib
5. Seaborn
6. Time Series Forecasting

Dataset:-

We have a dataset from WMATA here. This data set contains several records with features like ID, Time, location, direction, color, problem and delay. We have another file with scraped metro disruptions of size 1.4MB.

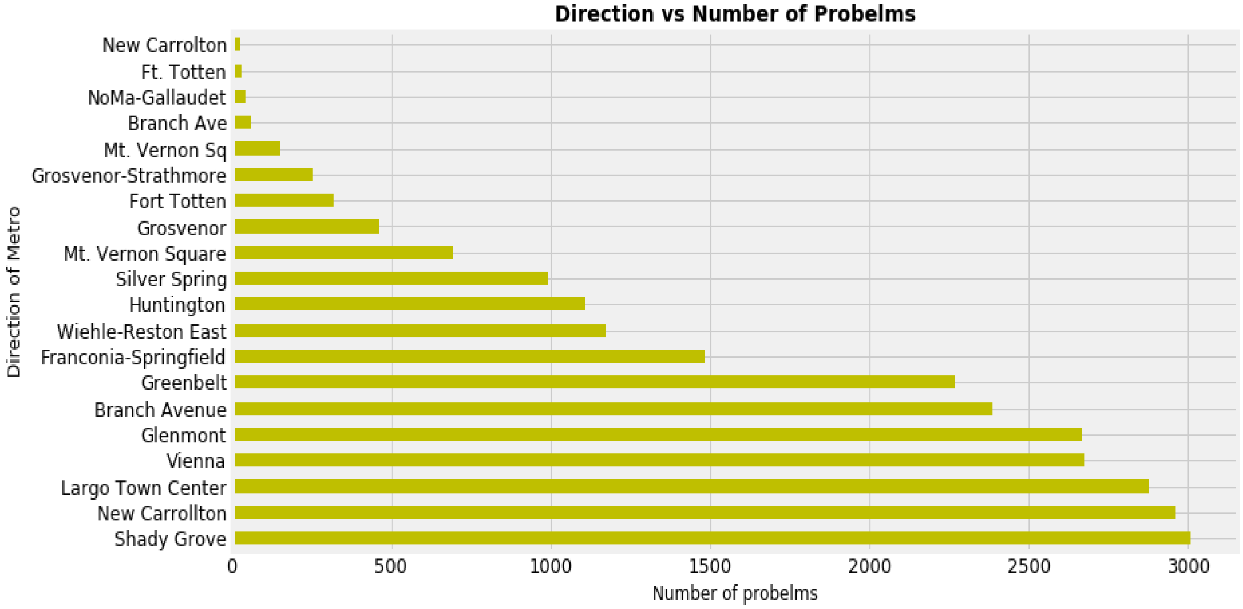
Process:-

Use the tools that we have like Numpy, Pandas, Matplotlib, Seaborn and Python to get insights as asked above from the data. Let's answer each and every point one by one.

Let's explore each and every column that can contribute to our analysis.

Direction:-

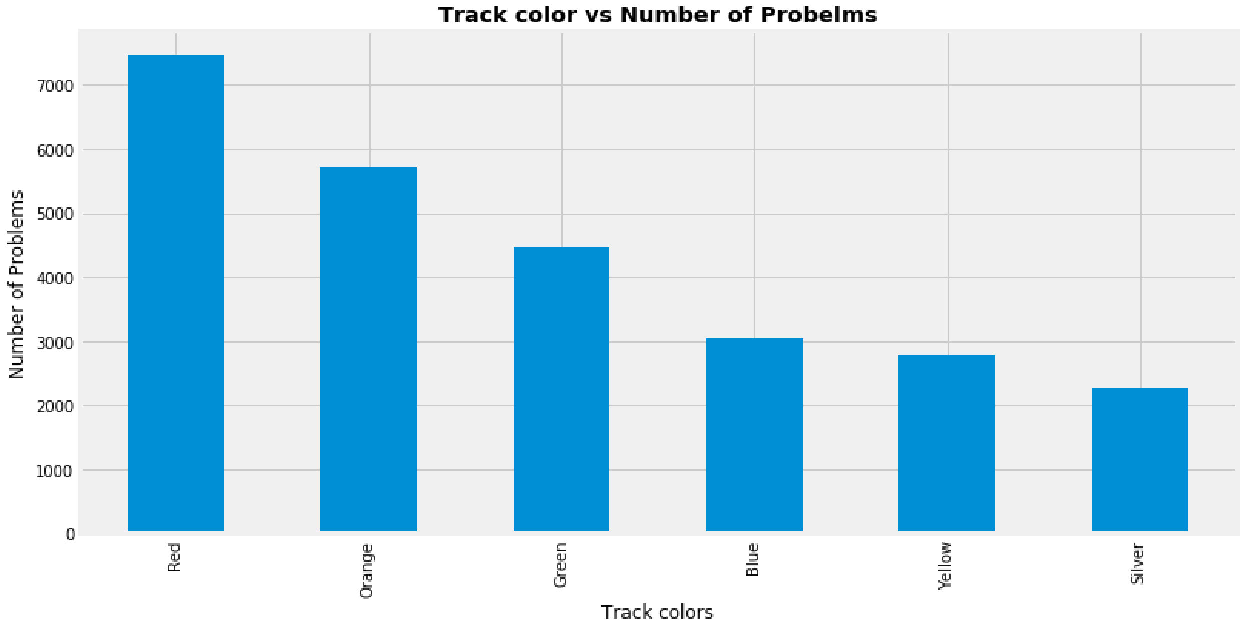
We have a feature called Direction that Metro moves towards.



From the graph, you can see that we have most of the metros that go in the direction of Shady grove faces more problems.

Track Color:-

This color feature tells us about the track that the metro travels. We have several tracks in the Washington metro station like blue, green, silver, orange, red, yellow, Greed.

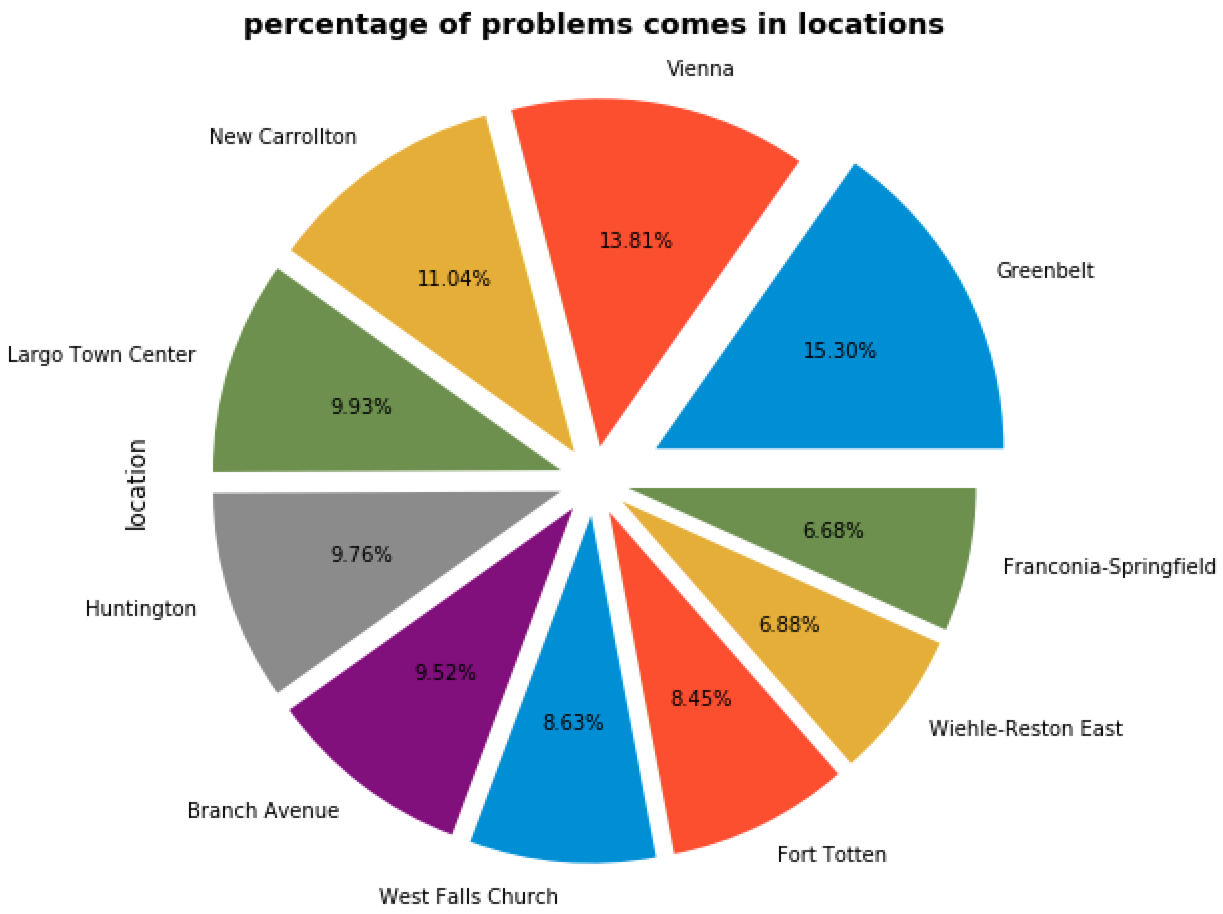


You can see from the graph that we have most of the problems on the Red track.

Location:-

This feature tells us that the problem occurring locations. We have locations such as Greenbelt , Vienna, New Carrollton, Largo Town Center, Huntington, Branch Avenue, West Falls Church, Fort Totten, Wiehle-Reston East, Franconia-Springfield, Rhode Island Avenue, King Street, Silver Spring, Metro Center, Farragut North, Stadium-Armory, Gallery Place, College Park, Takoma, Rosslyn and so on….

Let's explore this feature and try to find out the location that faces more problems.

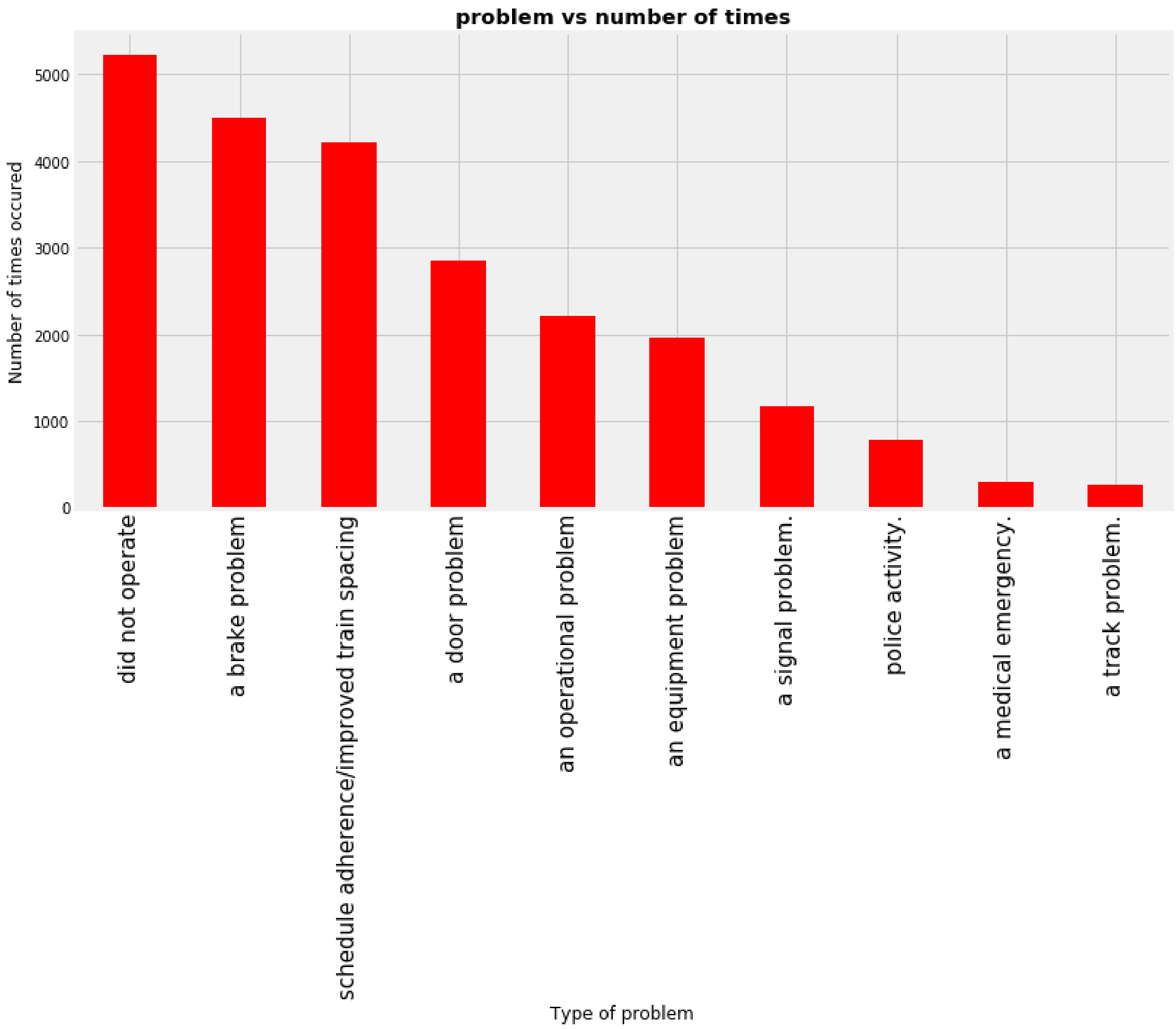


From the graphs, you can see that we have more problems in locations Greenbelt with Vienna in second place.

Problem:-

So we are going to explore what kind of problems do Washington metro faces and what is the most occurring problem. There are several problems like a brake problem, operational problem, door problem, equipment problem, track problem, police activity, and emergency problem.

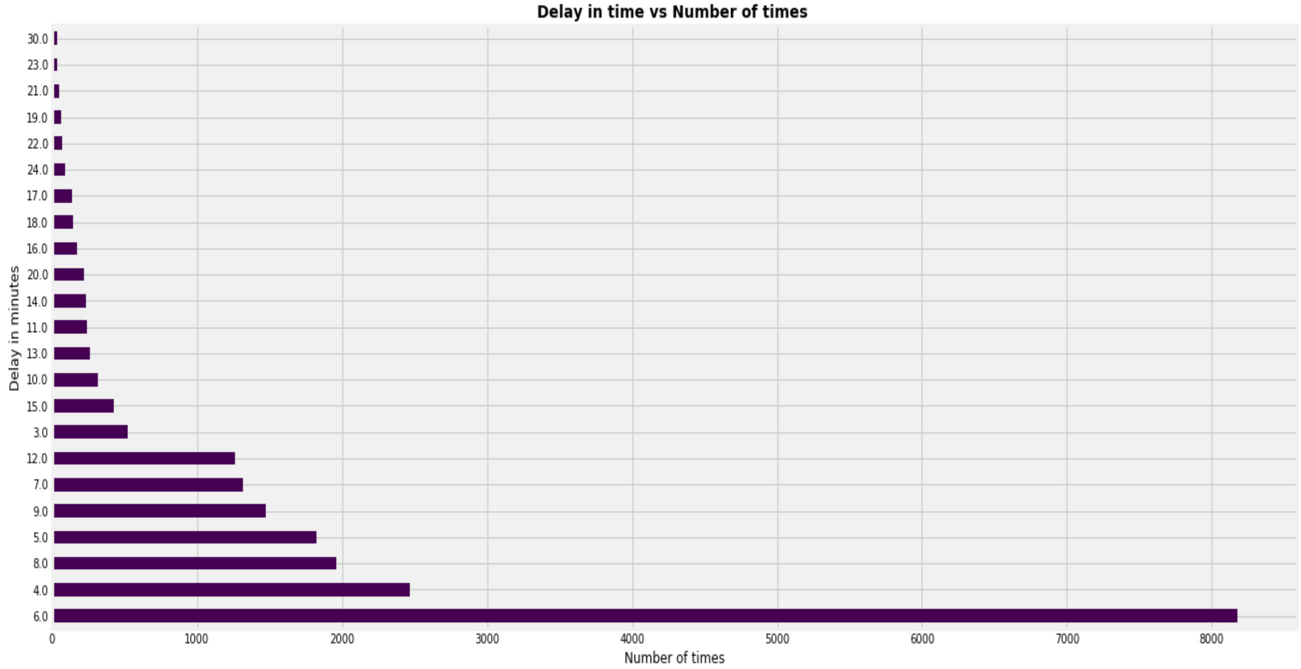
Let's plot the graph and see which problem caused a number of times.



From the graph, you can see that we have more problems because they did not operate. The most caused second problem is a brake problem, which is dangerous to the lives of passengers travelling in Wmata if not identified earlier.

Delay in minutes:-

We have a feature called delay in our dataset that is caused due to different problems. Each problem has a different time duration delay. Let's see the amount of delay that occurred most times due to problems.

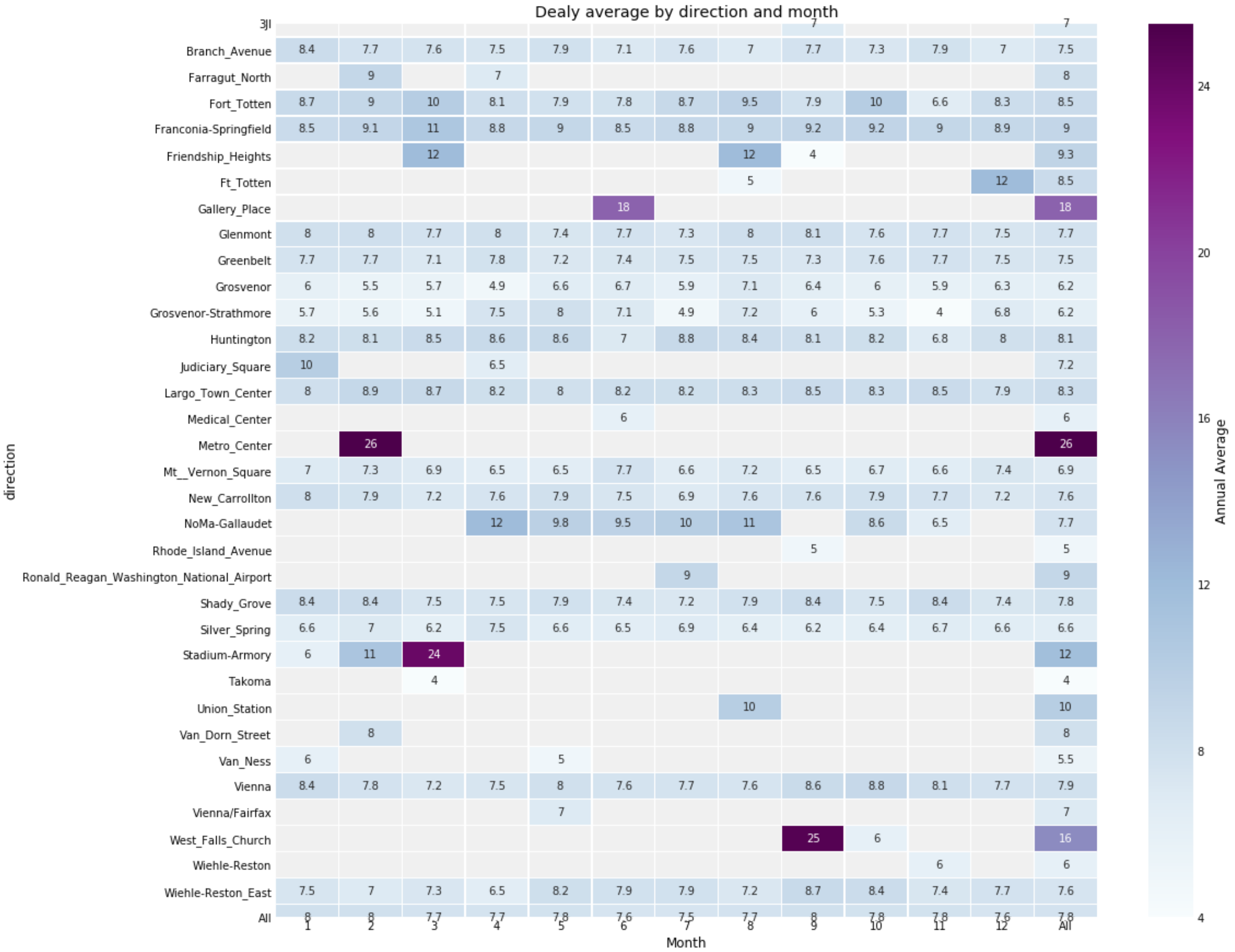


From the graph, it is observed that for most of the problems the maximum time delay is just 6 minutes.

Month:-

We have Monthly Average of Delay in minutes from the years 2012 to 2016. From the graph below you can see that we have 26 Minutes delay in the direction of Metro_center in February month. Also, we can see that we have 24 Minutes delay in the direction of the stadium Armory in the Month of March.

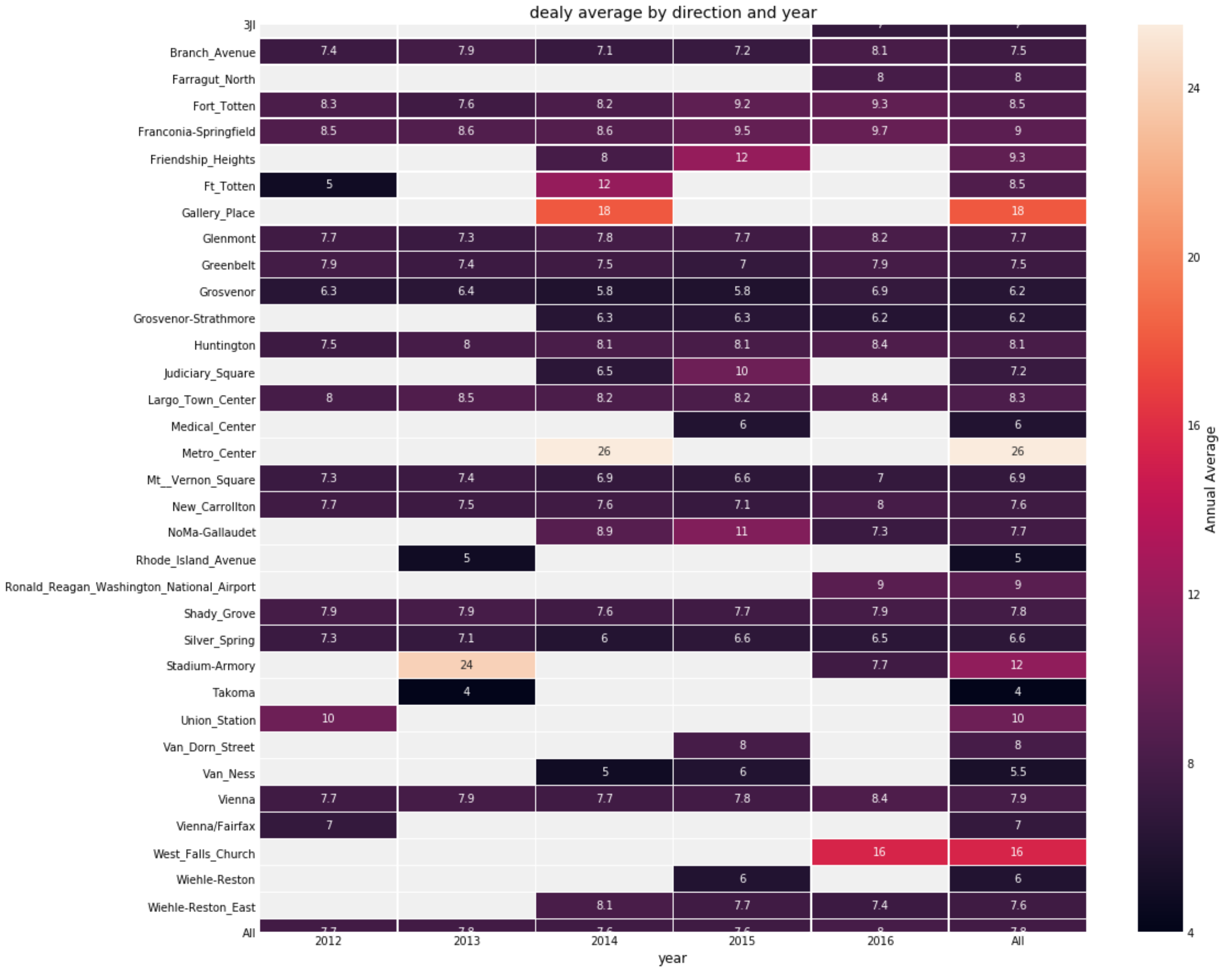
We have Average Monthly delay in minutes from 4 minutes to 24 minutes. Least Monthly average delay occurred in September month in the direction of Francois to Springfield and Also in the direction of Takoma.



Year:-

We have most of the problems that occurred in the year 2015 compared to others in the time duration of 2012 to 2016.

Down below we have Yearly Average Delay in minutes to each and every direction we have. We have the Highest yearly Average delay in the year of 2014 in the direction of Metro center.



dealy average by direction and year

Final observations:-

- most of the problems are coming to the trains that travel to the direction of shady grove
- we have most of the problems on the Redline.
- Most of the problems are coming at the location Greenbelt.
- at the location, Vienna comes more problems after Greenbelt.
- most metros get the problem of did not operate.
- The second-most problem is a Brake problem.

this "Brake problem" is dangerous and risk to people's lives.

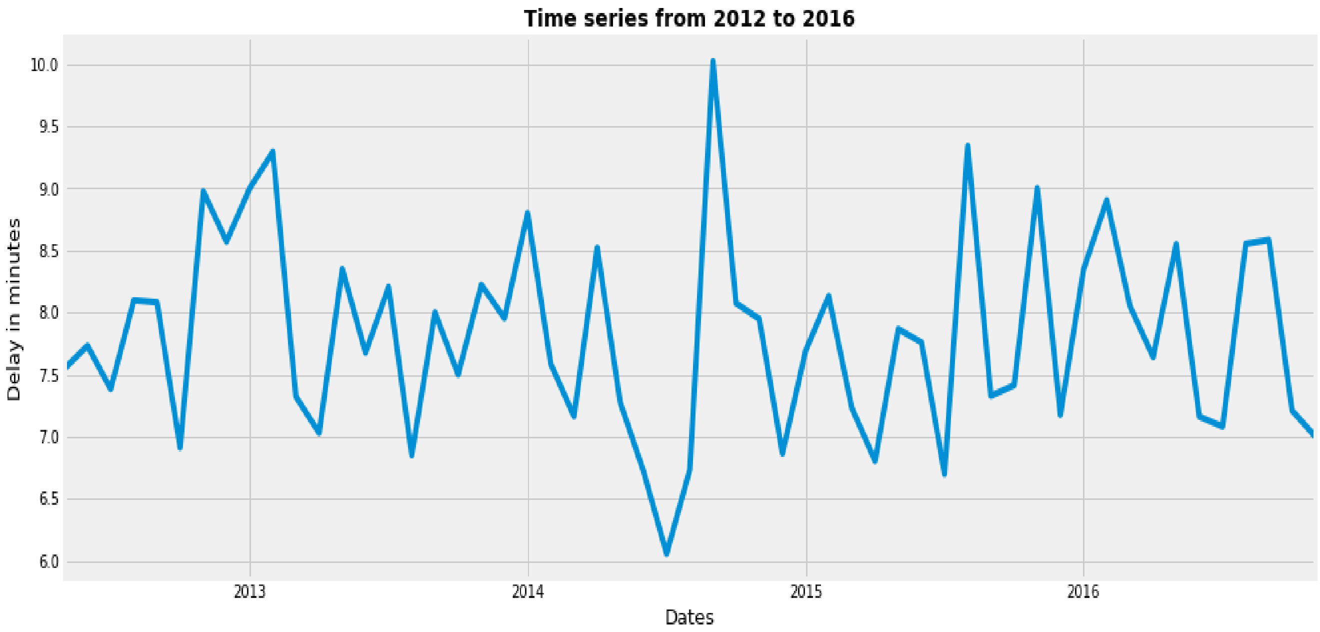
- most of the trains get delayed for 6 min with the 4min second place.
- we have most of the problems in the month of 7,8 i.e July, august.
- we had 62highest problems in 30- august- 2016.
- we had the highest problems in the year 2015.

- we got Most of the problems in the office hours which said to be the busiest hours.
  - 7 and 8 am in the morning.
  - 4 and 5 pm in the evening.

Time series forecasting:-

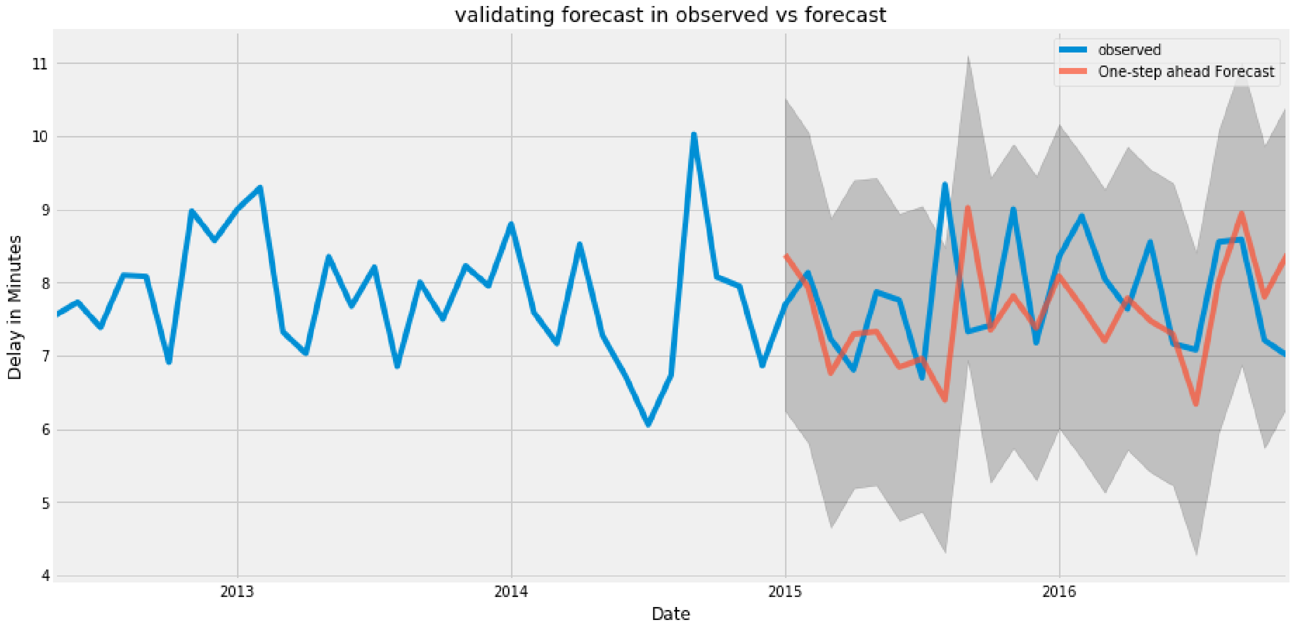
Now we have all the data cleaned for time series forecasting. As we have several directions, take one direction and predict the average delay that is going to occur for every month for 2017 to 2025. We have the data available from 2012 to 2016.

Lets plot the data we have from 2012 to 2016 with its average delay per month in minutes.



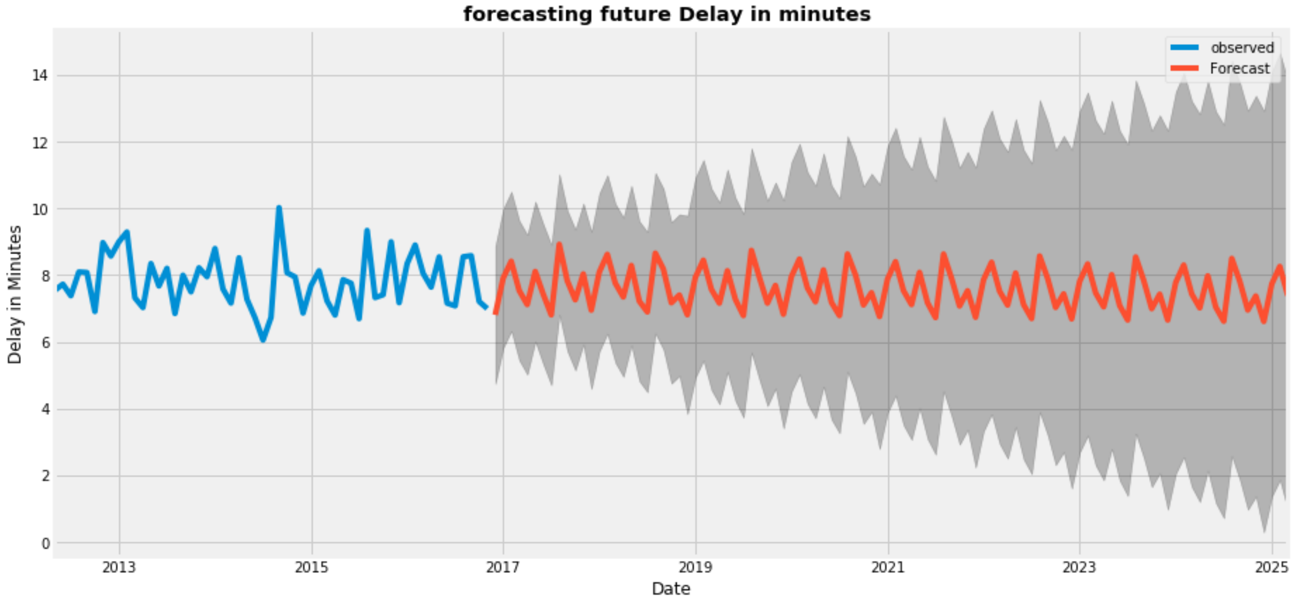
This is how it looks once you plot the data. You can see that we have an increase in delay from the second part of the year.

Let's use ARIMA - Auto-Regressive Integrated Moving Average to validate the forecast and see the observed vs forecasted.



Our error metric Mean Absolute Percentage Error gives only 9.91 which means our model is predicting to 91% accuracy here.

Now let's forecast the future average delay per month in minutes from 2017 to 2025.



Check out the predicted average delay in minutes per month here below.

Month	Predicted Average Delay
2019-02-01	8.450399
2019-03-01	7.577566
2019-04-01	7.160183
2019-05-01	8.133552
2019-06-01	7.274166
2019-07-01	6.788310
2019-08-01	8.745031
2019-09-01	7.930906
2019-10-01	7.160700