Stevens Institute of Technology

School of Systems and Enterprises
EM624 - Informatics of Engineering Management
FALL 2019 Final Project On

"NJ Transit Trains Delay Visualization"

Submitted by- Soham Sanjay Shinde
CWID- 10452638
Guided by- Dr. Carlo Lipizzi

Stevens Institute of Technology,1 Castle Point on Hudson, Hoboken NJ 07030, USA

What will be covered in this report for delay analysis?

I will write the report according to the CRISP DA format:

1. **Understanding Project**
2. **Data Preparation**
3. **Data Understanding**
4. **Methodology/ Evaluation**
5. **Results**
6. **Conclusions**

1. Understanding Project:

NJ transit has the second largest number of ridership in United States, and it connects the New Jersey and New York city. Amtrak and NJ Transit operates total 750 trains across NJ transit rail network. NJ Transit **serves more than 3,00,000 commuters** on average weekday, so most of the peoples depend on the NJ Transit and it is the vital and most mode of commuting for most of the people. As of 2012, NJ Transit's commuter rail network consists of 11 lines and 164 stations, primarily concentrated in northern New Jersey, with one line running between Atlantic City and Philadelphia.

This project can help NJ Transit to get some insights from analysis and lots of interesting, high-impact projects could be driven by this data.

By analysis of NJ Transit data we can analyze vital information given below –

- To **analyze which line has the maximum delay** to get some reasons of delays to help increasing productivity of NJ Transit.
- To **analyze the statistical data** from all the all data sets, for example what is the mean, standard deviation, maximum, minimum of delayed trains in minutes.
- To **analyze the delay from the March 2018 to December 2018** so that NJ transit can analyze that which season has more delays of trains on 11 lines, so that they can withdraw insights and plot a pie chart.
- To **analyze how many trains in July got delayed**, **on-time or cancelled** and plotted on pie chart.
- To **provide intelligent, targeted advance warnings of delays** or cancellations for millions of riders.

2. Data Preparation:

URL -“https://www.kaggle.com/pranavbadami/nj-transit-amtrak-nec-performance”

Source Website – Kaggle.com

Data Collection -

I have taken data from March 2018 to December 2018 for my analysis. Most of the data I downloaded had 2,44,850 rows. While collecting the data I edited some data and had to see whether it has ‘scheduled time’ of departure and ‘actual time’ of departure so that the difference between them can give me delay in minutes.

After collecting data, I have cleaned the data which had some blank rows in it for smooth running of program.

3. Data Understanding:

To understand data set I have downloaded all comma separated value(csv) files from above link. This dataset contains monthly arrival and departure of every train of NJ Transit. This data set has 287,000+ train trips data (2428,000+ NJ Transit trips, 38,000+ Amtrak trips) in total of 14 csv files. The data has covered from March 1, 2018 to April 30, 2018. From the total of the following trips, 97.5% train trips were correctly captured in the data set.

There are following columns in the csv file that are listed below:

- **date -** Date is considered in the form 04/09/18 4:00 to 04/09/18 27:00
- **train_Id -** These are unique on a daily basis and correspond to the same scheduled train across multiple days

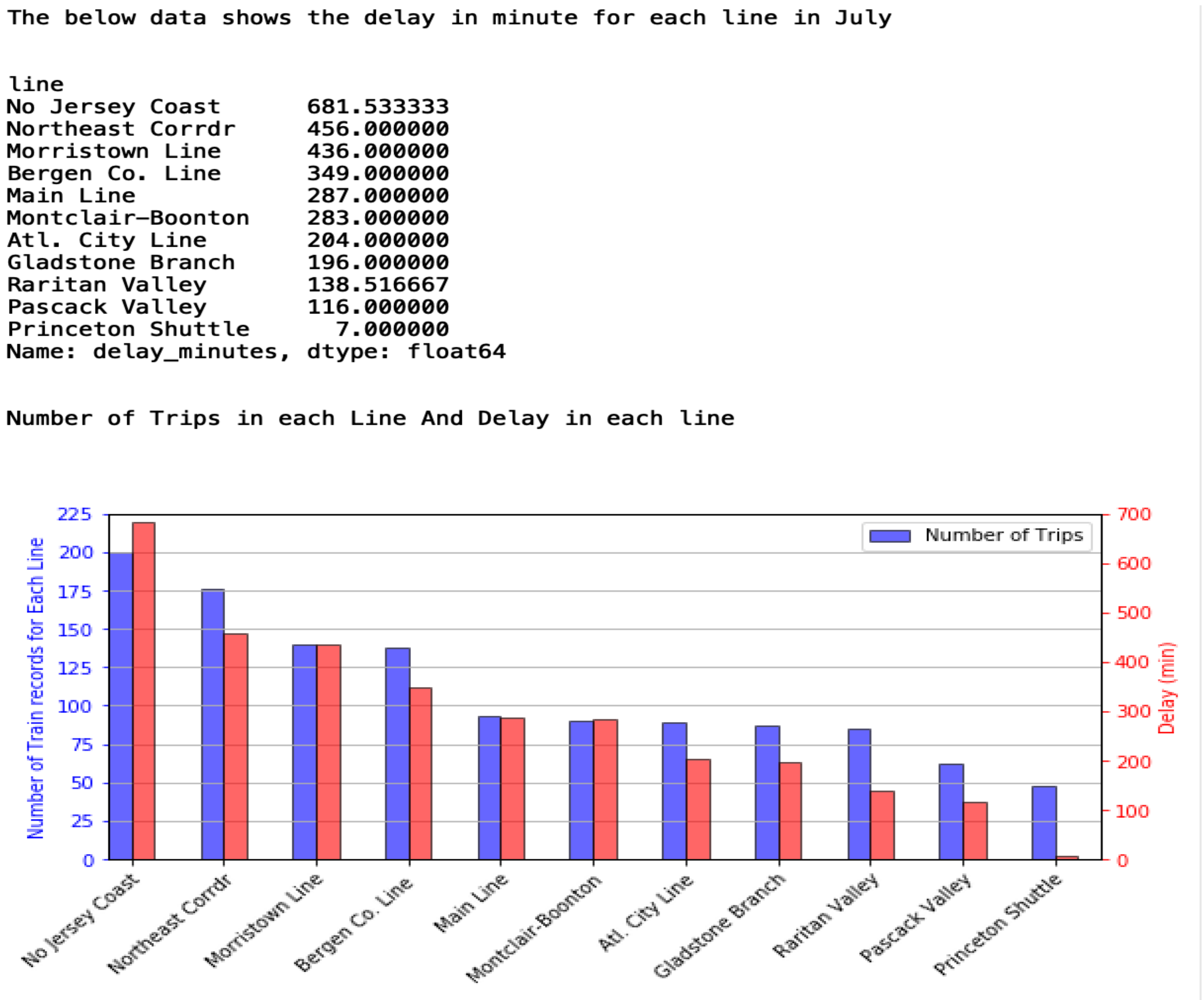
- **stop_sequence** - Scheduled stop number (e.g. 1st stop, 2nd stop) for the stop in the current row
- **from -** Station the train is traveling from
- **from_id** - Station id for the "from" station
- **to** - Station the train is arriving “to” for the stop in the current row
- **to_id** - Station id for the "to" station
- **scheduled_time -** The time that the train was supposed to depart in minutes
- **actual_time –** The time at which the train leaved the specific station in minutes
- **delay_minutes -** The difference between actual_time and scheduled_time, in minutes
- **status –** The status of train whether train has departed or cancelled or on-time
- **lines –** There are total of 11 lines operated by NJ Transit and in this column they have provided the data of which train travels on which line
- **type –** Either ‘NJ Transit’ or ‘Amtrak’

4. Methodology/ Evaluation:

1) **Importing Libraries –** In this step I have imported the libraries required to run the program like numpy, matplotlib, pandas, datetime
2) **Read csv file -** First using pandas, I read the .csv file and dropped the row which had Amtrak data in last column ‘type’
3) **Display Statistical data -**Displayed statistics of dataset like mean, max, standard deviation, minimum, maximum and also converted the “scheduled_time" and "actual_time" columns to datetimes
4) **Using group by -** Grouping July data according to train Id and date
5) **Function creation -** Function to produce pie charts with matplotlib
6) **Reading of csv file and dropping the Amtrak values** -Drop all the Amtrak values in each month
7) **Grouping the data from march to december** - Group by train_Id and date for all months
8) **Counting the delay minutes from data which are greater than zero**
9) **Converting the delay minutes into list** – The values that we get from each month can be converted to list to show list for graph
10) Display data for each month delay and for July delays using function that we have defined above
11) **Status of trains in july**
12) **Most Delayed Lines in July**
13) **Lines Vs delays –** In this part I have plotted a graph for number of departures, cancelled and on-time lines out of one
14) **Most delayed lines in July –** I have plotted a graph for delay(Y-axis) in each line (X-axis) in the month of July

5. Results:

a. Line Vs Number of Trips and Line Vs Delay

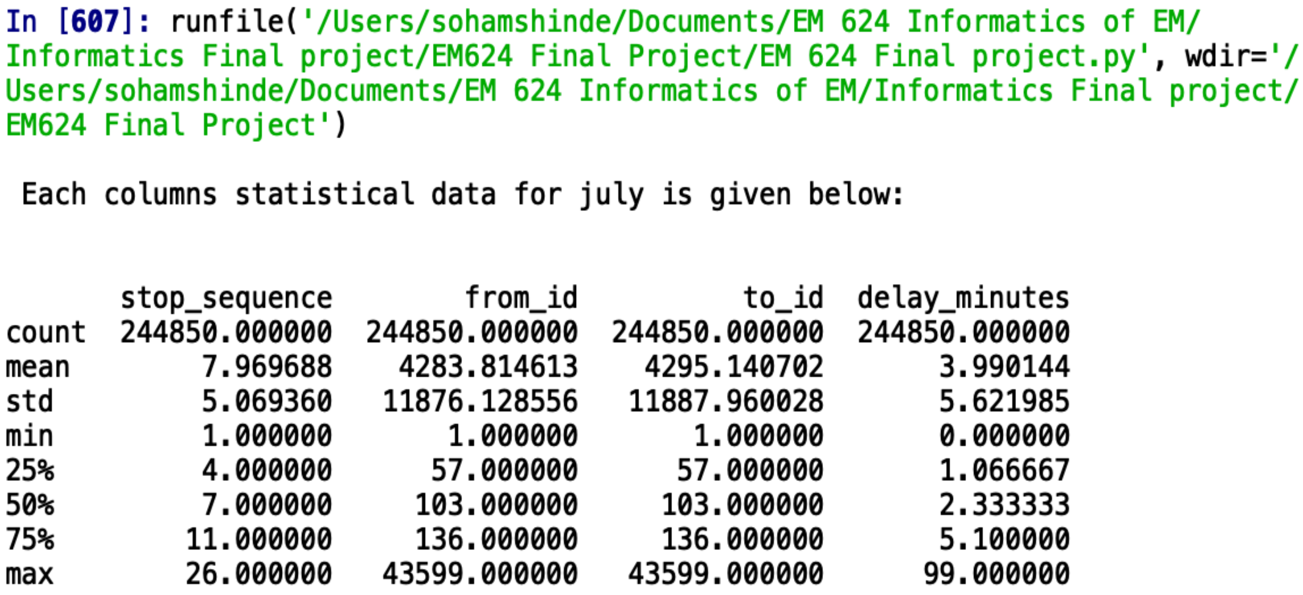


```
The below data shows the delay in minute for each line in July


line
No Jersey Coast       681.533333
Northeast Corrdr      456.000000
Morristown Line       436.000000
Bergen Co. Line       349.000000
Main Line             287.000000
Montclair-Boonton     283.000000
Atl. City Line        204.000000
Gladstone Branch      196.000000
Raritan Valley        138.516667
Pascack Valley        116.000000
Princeton Shuttle       7.000000
Name: delay_minutes, dtype: float64


Number of Trips in each Line And Delay in each line
```

Figure 1. Bar chart for Number of Trains on each line Vs Delay per minute

The above bar graph shows that highest number of train trips on specific line, has the most delays on that line. We can see, if the number of trains reduces then delay in minute also reduces for given line of train. So, if they have to reduce the delay on the No Jersey coast then they might have to reduce the number of train record on that line. The bar chart clearly specify number of trips is directly proportional to delay on that line.

b. I have got the following results by running the program:



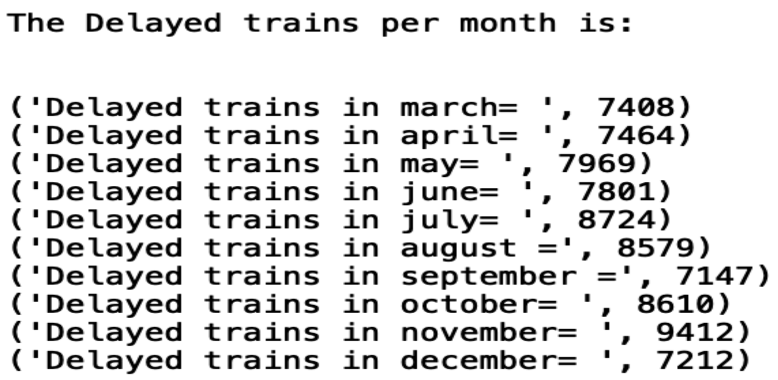
```
In [607]: runfile('/Users/sohamshinde/Documents/EM 624 Informatics of EM/
Informatics Final project/EM624 Final Project/EM 624 Final project.py', wdir='/
Users/sohamshinde/Documents/EM 624 Informatics of EM/Informatics Final project/
EM624 Final Project')

 Each columns statistical data for july is given below:


       stop_sequence        from_id          to_id  delay_minutes
count  244850.000000  244850.000000  244850.000000  244850.000000
mean        7.969688    4283.814613    4295.140702       3.990144
std         5.069360   11876.128556   11887.960028       5.621985
min         1.000000       1.000000       1.000000       0.000000
25%         4.000000      57.000000      57.000000       1.066667
50%         7.000000     103.000000     103.000000       2.333333
75%        11.000000     136.000000     136.000000       5.100000
max        26.000000   43599.000000   43599.000000      99.000000
```

Figure 2. Table for statistical data

The above output displays table which has data for stop_sequence, from_id, to_id, delay_minutes. The data is of the mean, count, mean, standard deviation, minimum value, maximum value etc. We will only concentrate on delay_minutes because we are only focusing on delay of trains.

From this we can infer, the mean delay is 3.99 minutes and maximum delay of train is 99 minutes. Standard deviation is 5.621 minutes and minimum delay is 0.00 minutes. We also come to know that 25 % of the trains had 1minute delay, 50 % of all trains from march to December has 2.3 minutes delay and 75% of total trains had more than 5 minutes delay.

c. Delayed trains per month from March 2018 to December 2018:



```
The Delayed trains per month is:

('Delayed trains in march= ', 7408)
('Delayed trains in april= ', 7464)
('Delayed trains in may= ', 7969)
('Delayed trains in june= ', 7801)
('Delayed trains in july= ', 8724)
('Delayed trains in august =', 8579)
('Delayed trains in september =', 7147)
('Delayed trains in october= ', 8610)
('Delayed trains in november= ', 9412)
('Delayed trains in december= ', 7212)
```

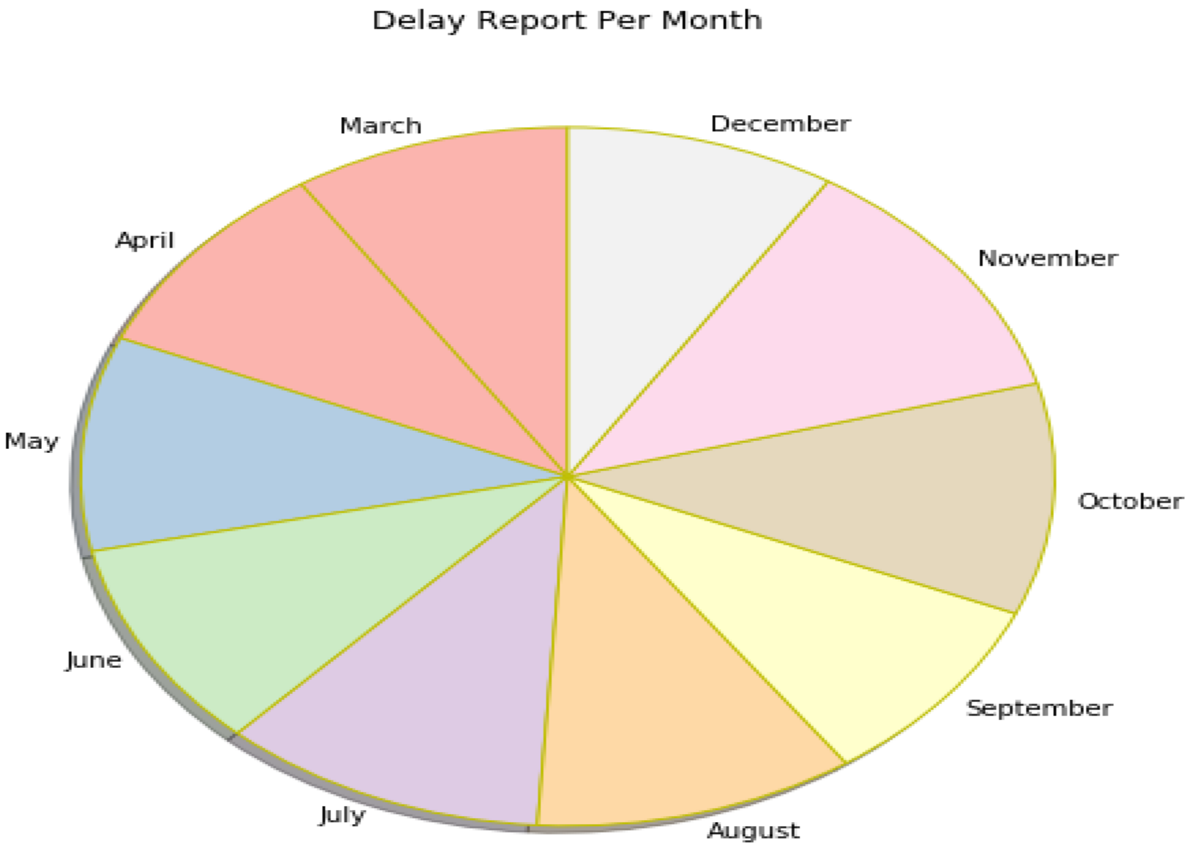
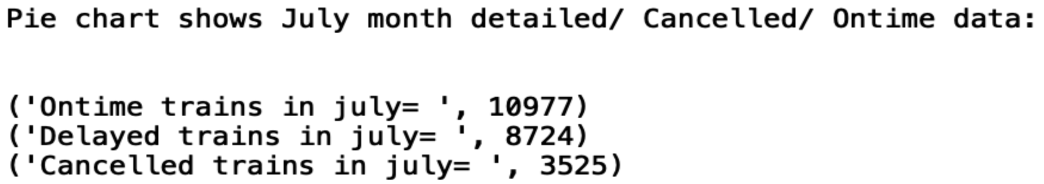


Figure 3. Pie chart for delay per month

We can see that the delay of trains is maximum in the month of November and minimum is the month of September. And we can see a pie chart showing graphical representation of delay from March to December.

6. Delay/ Cancelled/ On-time trains for July:



```
Pie chart shows July month detailed/ Cancelled/ Ontime data:


('Ontime trains in july= ', 10977)
('Delayed trains in july= ', 8724)
('Cancelled trains in july= ', 3525)
```

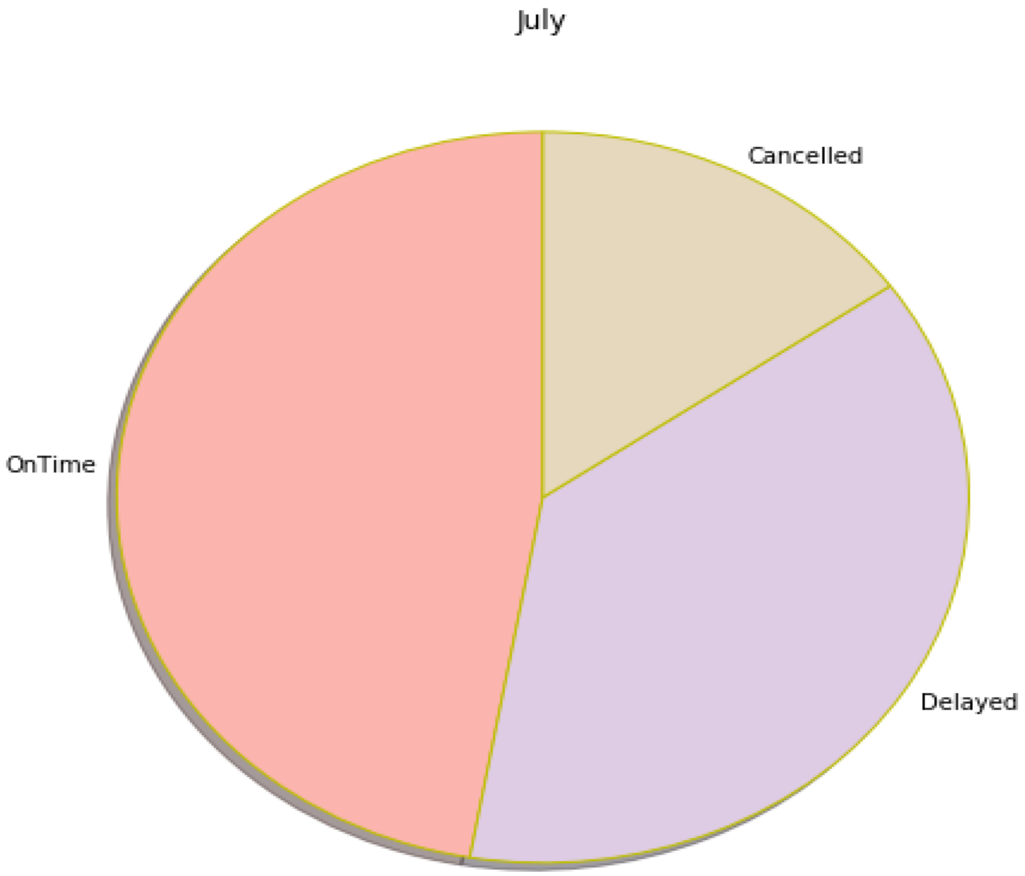
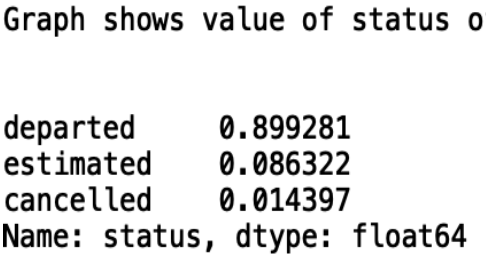


Figure 4. Pie chart for train status in July

The above graph shows the status of the trains which were On-time, delayed or cancelled. On-time trains were 10977 and 8724 trains were delayed and 3525 trains were cancelled in the month of July which would be because July has the second largest delay trains. The pie chart given in figure 3 shows that the same value for delay as delay for month of July in figure 4.

7. Train status in July:



```
Graph shows value of status of lines in july


departed     0.899281
estimated    0.086322
cancelled    0.014397
Name: status, dtype: float64
```

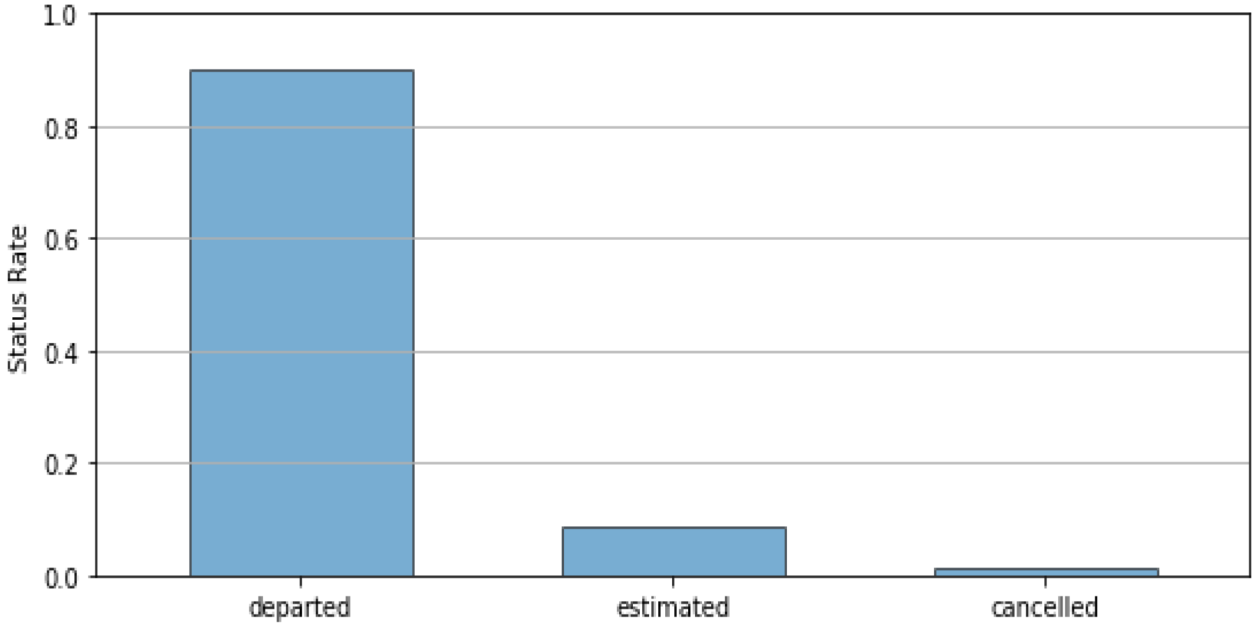


Figure 5. Bar chart for status of train in July

This graph shows that if we consider total number of trains as 100% then 89.9281% trains are departed on time. Cancelled trains in July were 1.4%. For displaying status rate, I have plotted in bar chart with X- axis has status and Y- axis has status rate. Basically the above values of figure 4 has been taken and plotted in the form of bar chart.

8. Conclusions:

From above results we can conclude that:

1. Maximum delay is in the North Jersey coast line of NJ Transit, so NJ Transit has to focus more on this line because it has highest number of trips. NJ Transit has to reduce train trips for delay to help minimize delay of train on North Jersey coast line.

2. From statistical analysis done in figure 2 we can infer that the mean delay is 3.99 minute. We can decrease this mean delay by implementing some discussed techniques in point 1, to increase the productivity.

3. From figure 3 we can conclude, November has highest delayed trains and July has second highest month for train delay. In this month NJ Transit can has to be more precise in terms of their management.

4. NJ Transit cancelled most of the trains in the month of July.