Exploratory Data Analysis and Modelling for:

US Traffic Data, 2015 (Kaggle)

The dataset consisted of two .txt files in their archived format as dot_traffic_2015.txt.gz & dot_traffic_station_2015.txt.gz. Firstly, I extracted those files and then loaded the text files as pandas dataframe into the notebook. Naming conventions used in the Jupyter notebook for the two text files are as follows: traffic_df & traffic_station_df.

Details of the text files loaded as pandas dataframe are as follows

1. **Traffic_df:**
   Shape: (7140391,38)
   It consisted of traffic count of each hour of each day of year 2015 for various states in US represented with their fips code, station ids, travel lane, functional classes, direction of travel etc.
2. **Traffic_station_df: (cross-referenced by station ids)**
   Shape: (28466,55)
   It consisted of deeper insights for each station their geographical coordinates, historical data, lanes & vehicle monitored for traffic, sensors involved, county fips codes, data retrieval methods, vehicle classification algorithms etc. This data was rather more typical to deal with as it contained too many null values, repeated columns and large number of columns.

Data Cleaning & Handling

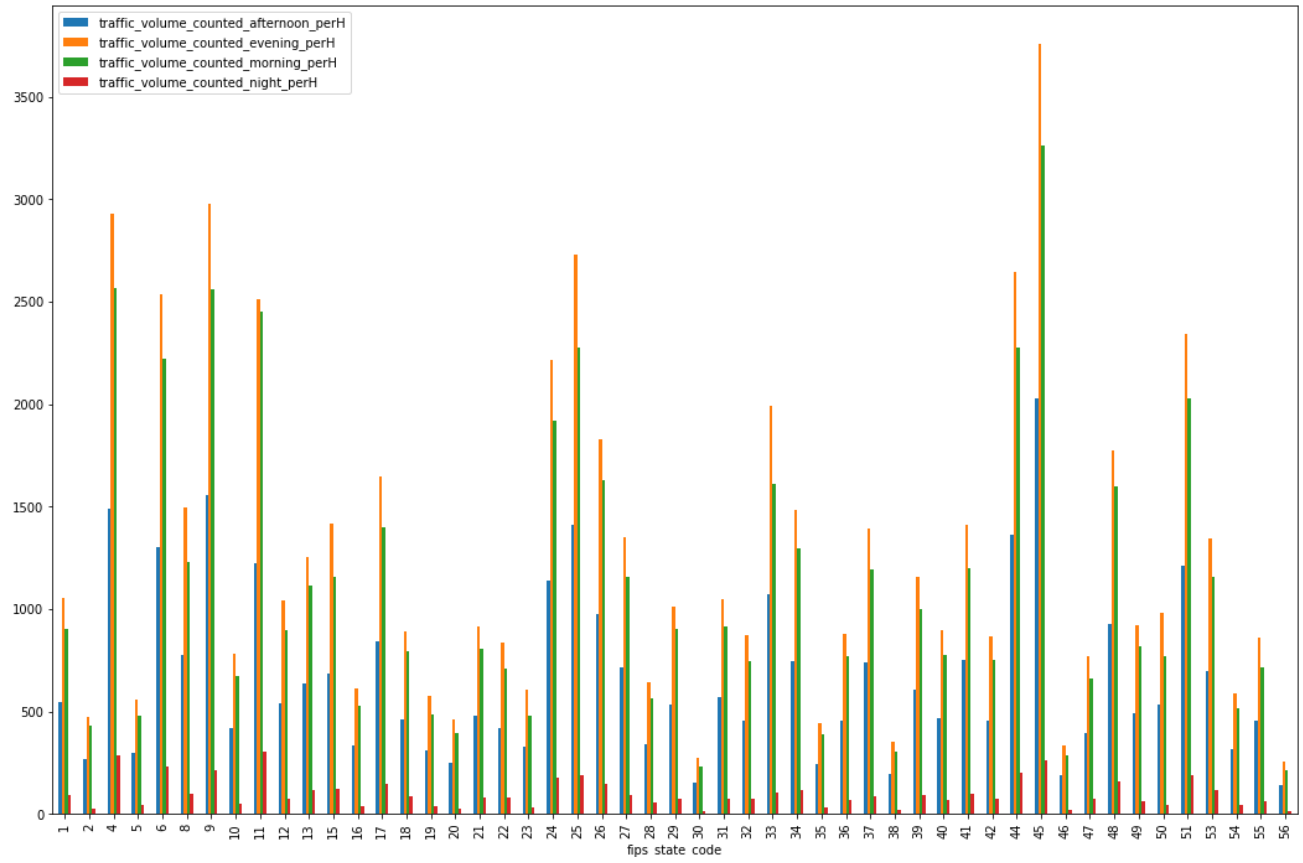
1. Traffic_df:
   - Firstly, I assigned the names for the week days & months in a dictionary for better understanding for the data.
   - Dropped the unwanted (like restrictions, had all NAN values), repeated (like direction of travel & direction of travel name) and constant columns (like year is same for all 7.1M rows).
   - Converted the 24 hour formatted traffic count columns to 4 columns as:
     NMAE:
     Night (00-06), Morning (06-12), Afternoon (12-18) and Evening (18-24). [as int type]
     And dropped the 24 hour format traffic count columns.
   - The four section binned format (NMAE) contained outliers as 75% of the data was about approx. 1/1000th time of the maximum traffic count. Hence, rows with NMAE values greater than 10k were dropped.
     New shape of traffic_station: (7140391,12)
2. Traffic_station_df:
   - Firstly, dropped the unwanted, repeated and constant columns from dataframe.
   - Dropped the columns which contained null values more than 20k as the dataset contains 28k rows. Hence removed columns with 70% above null values.
   - Only taken station ids which are present in traffic_df and removed station ids which are not in traffic_df.
   - **Binary Encoded** the following columns as 0 and 1:
     hpms sample type, national highway system and sample type for vehicle classification
     New shape of traffic_station_df: (24275,15)

Patterns/Trends

1. **45, 9 and 4** are the **heavy traffic state**s in US for the year 2015, whereas **56 and 30 are the states where there's no traffic** at all *(Numbers are fips state codes).* The **station** with **017200, 119780 and**

10093 ids have recorded **highest mean traffic in US for the year 2015**, whereas stations ids **00041 and 075040 appears to be have no habit at all.**

Fips state code wise mean traffic count per hour for all four time interval



Station ids wise mean traffic count per hour for all four time interval (top 10 and min 10 stations)

station_id	traffic_volume_counted_afternoon_perH	traffic_volume_counted_evening_perH	traffic_volume_counted_morning_perH	traffic_volume_counted_night_perH
071200	5470.000000	9934.000000	9720.000000	1002.000000
119780	5434.000000	9776.500000	9253.500000	908.500000
100093	5106.400000	9571.900000	9797.500000	1365.400000
070300	5060.714286	9536.000000	9669.142857	1028.714286
100117	5284.172414	9439.551724	9421.103448	1003.275862
100118	5234.984615	9334.538462	9153.907692	946.492308
100105	3842.000000	9318.000000	7619.000000	1677.000000
071280	5068.083333	9310.500000	9494.333333	924.083333
101879	4993.482759	9281.344828	9028.068966	1005.068966
101881	4906.137931	9261.206897	8789.620690	966.724138

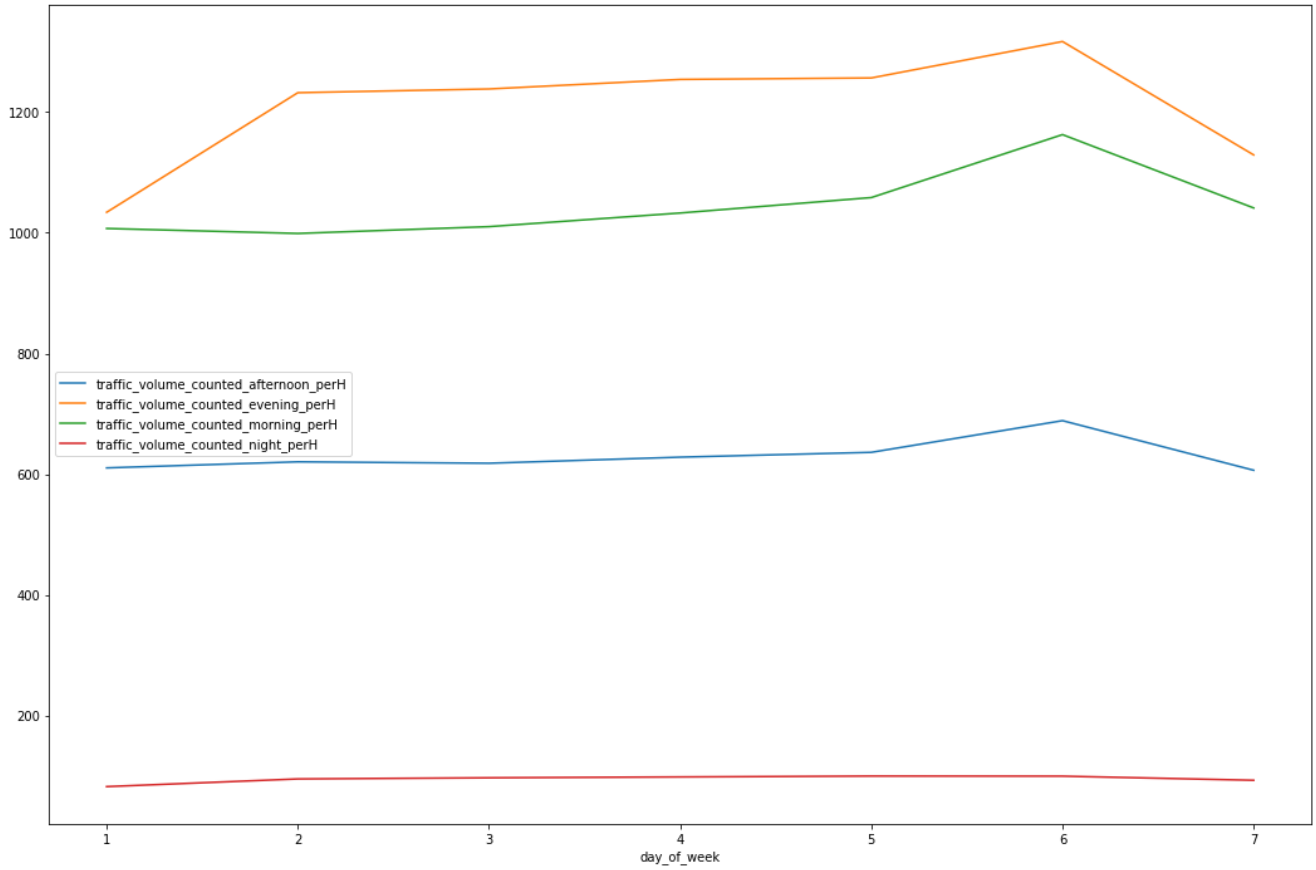
station_id	traffic_volume_counted_afternoon_perH	traffic_volume_counted_evening_perH	traffic_volume_counted_morning_perH	traffic_volume_counted_night_perH
001007	8.425714	15.674286	12.180000	0.492857
011004	6.972603	12.916438	9.094521	0.060274
781230	5.648199	11.185596	10.518006	1.481994
0131SE	4.627273	9.424242	6.995455	0.637879
K10500	3.702703	7.162162	4.993994	0.064565
K20800	3.344681	6.451064	4.227660	0.236170
P32AAA	1.483562	3.012329	1.928767	0.173973
000599	0.208448	0.359963	0.275023	0.042700
000441	0.000000	0.000000	0.000000	0.000000
075040	0.000000	0.000000	0.000000	0.000000

Name: Sanskar Sharma **E-mail: sansskarsharma@gmail.com** **Contact: +91 6261614589**

2. It is observed that US have heaviest traffic in the evening (18-24 Hr) and lightest traffic is obtained at night (00-06 Hr) for the year 2015. Morning mean traffic is second heaviest among four.
   General trend in mean traffic reduction of a day for year 2015:
   Evening (18-24) > Morning (06-12) > Afternoon (12-18)> Night (00-06)
   It can be justified as Office workers, students, daily wage workers etc. returns home in the evening and usually plans to go out then, resulting in heavy evening traffic, morning traffic is accounted by the rush to go to school, offices and for work and usually its bedtime at night resulting in least traffic due to 24x7 transportation services in US.

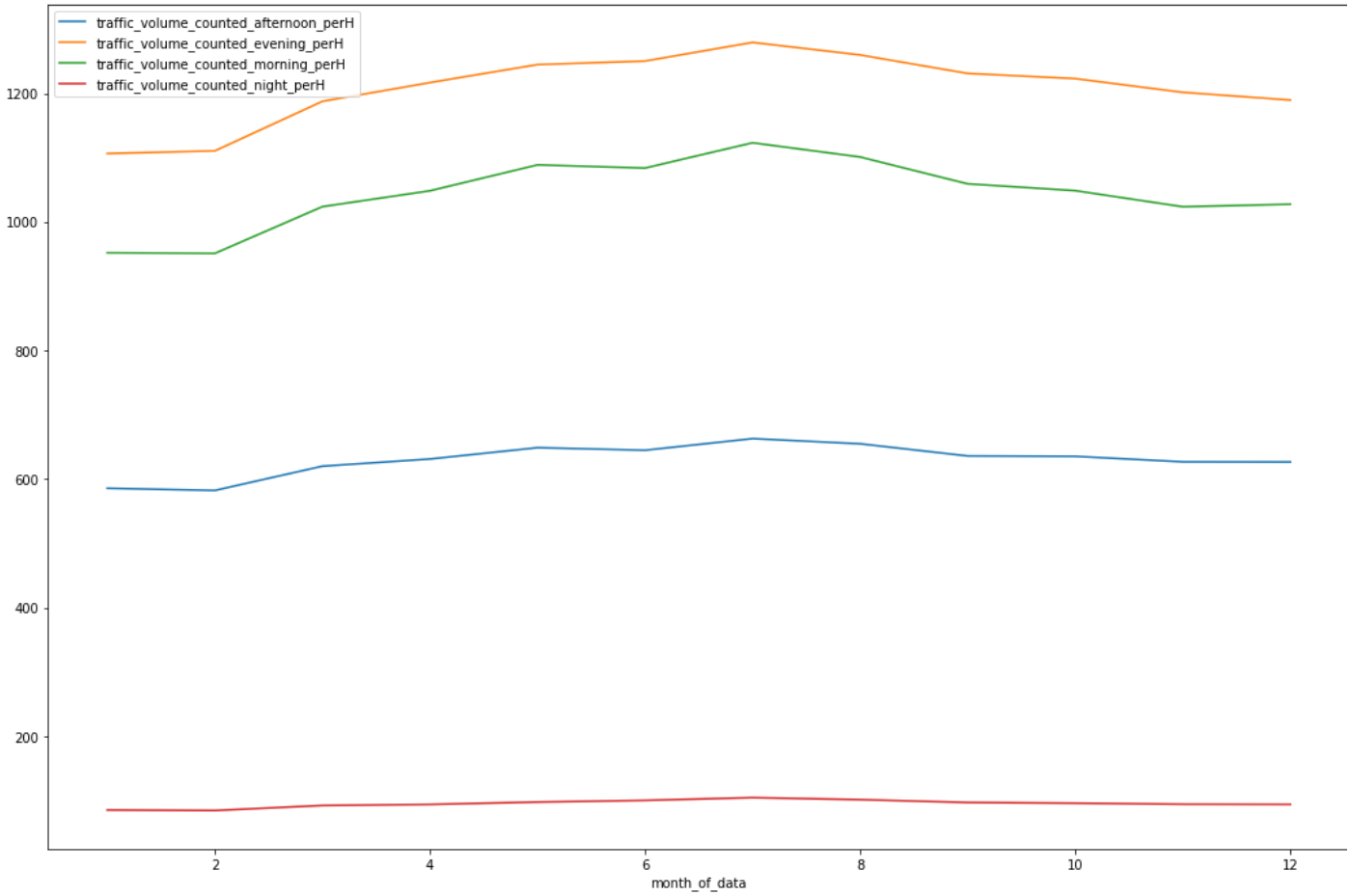
```
#Mean traffic count per hour over the year 2015 at night, morning, afternoon and evening
print("Night traffic per hour over the complete year: ",round(traffic_df["traffic_volume_counted_night_perH"].mean()))
print("Morning traffic per hour over the complete year: ",round(traffic_df["traffic_volume_counted_morning_perH"].mean()))
print("Afternoon traffic per hour over the complete year: ",round(traffic_df["traffic_volume_counted_afternoon_perH"].mean()))
print("Evening traffic per hour over the complete year: ",round(traffic_df["traffic_volume_counted_evening_perH"].mean()))

Night traffic per hour over the complete year:  95
Morning traffic per hour over the complete year:  1045
Afternoon traffic per hour over the complete year:  630
Evening traffic per hour over the complete year:  1209
```

3. **Saturdays** in US are found to be extreme heavy traffic day than other days ***except for night*** time.
   General trend in traffic for a week is:
   It increases from Monday reaches its extreme count on Saturday and then drops nearer to Monday's traffic on Sunday.
   It can be explained as Saturday and Sunday are the **weekends** and as Saturday being the first people generally plan this day for some outing, shopping and personal amusements resulting in more traffic while they in general choose to rest at home on Sunday.

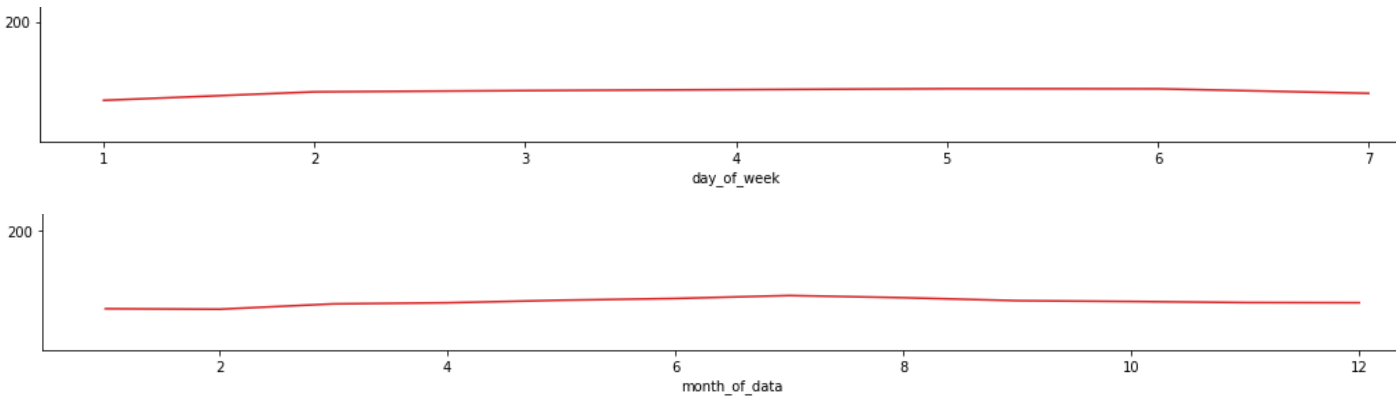


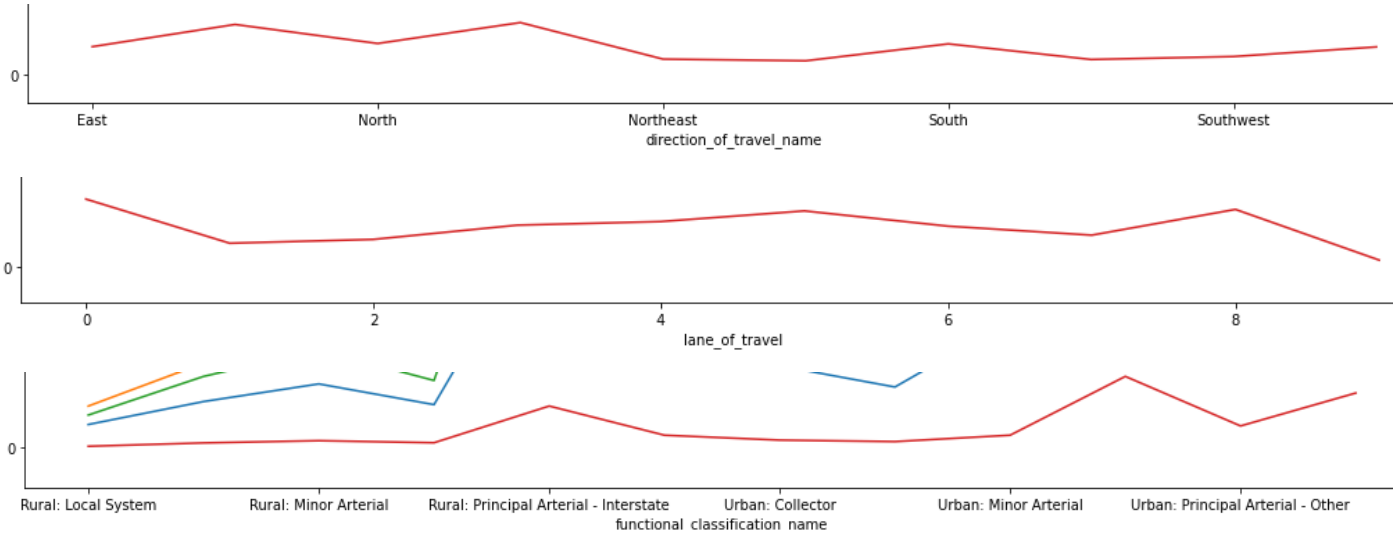
4. There's **a monthly parabolic trend observed in traffic for the complete year 2015 in US**. **January and February are months with lightest and almost equal traffic**. The traffic increases after February and Reaches its extreme count in month of **July (being heaviest traffic month)** and then **starts decreasing gradually till December *except for night time*** (least noticeable parabola)

It can be justified as in US Jan and Feb are the months when its snows blocking almost most of the lanes and routes. Also, people avoids travelling in the **snowy season** which explains the least traffic in these months. **The parabola is due to the seasoning in US.** As it is a colder place with its summer season close to June and July and people there prefer this weather to travel and celebrate and what not resulting it to be heaviest traffic month.



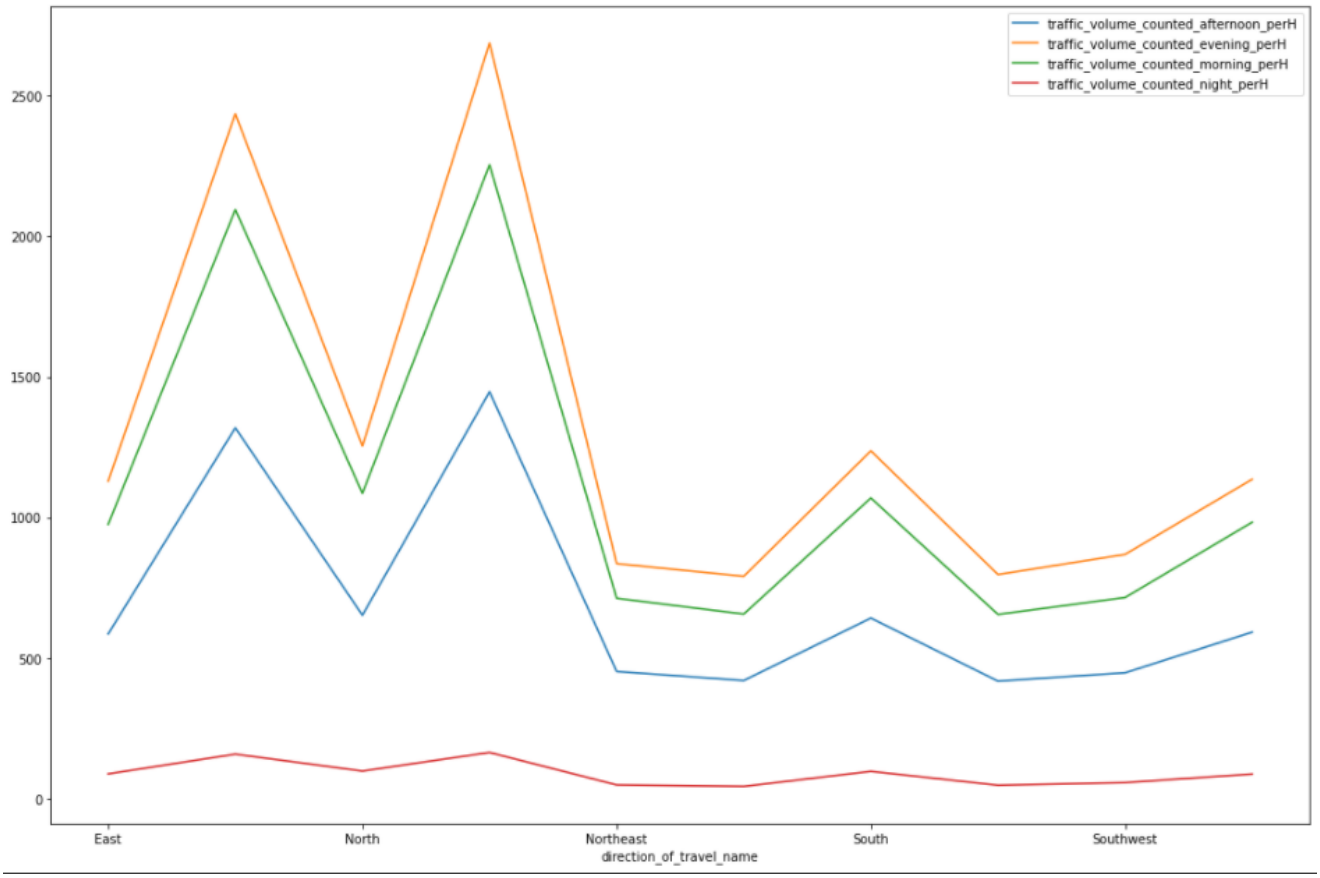
5. From all the inferences and observations made from various visualizations and analysis. It is found that for various states and station ids in US on all weekdays and months mean traffic on various directions of travel, lanes of travel and functional classifications at the **night time is constant and doesn't change much.**
   In general, across the United States night traffic is almost same, which can be visualized by the approximately straight lines in the various lines plots and also seen by the constant colour in heat maps for night time.
   Below are the instances of constant/straight line mean traffic per hour at night with different scenarios.





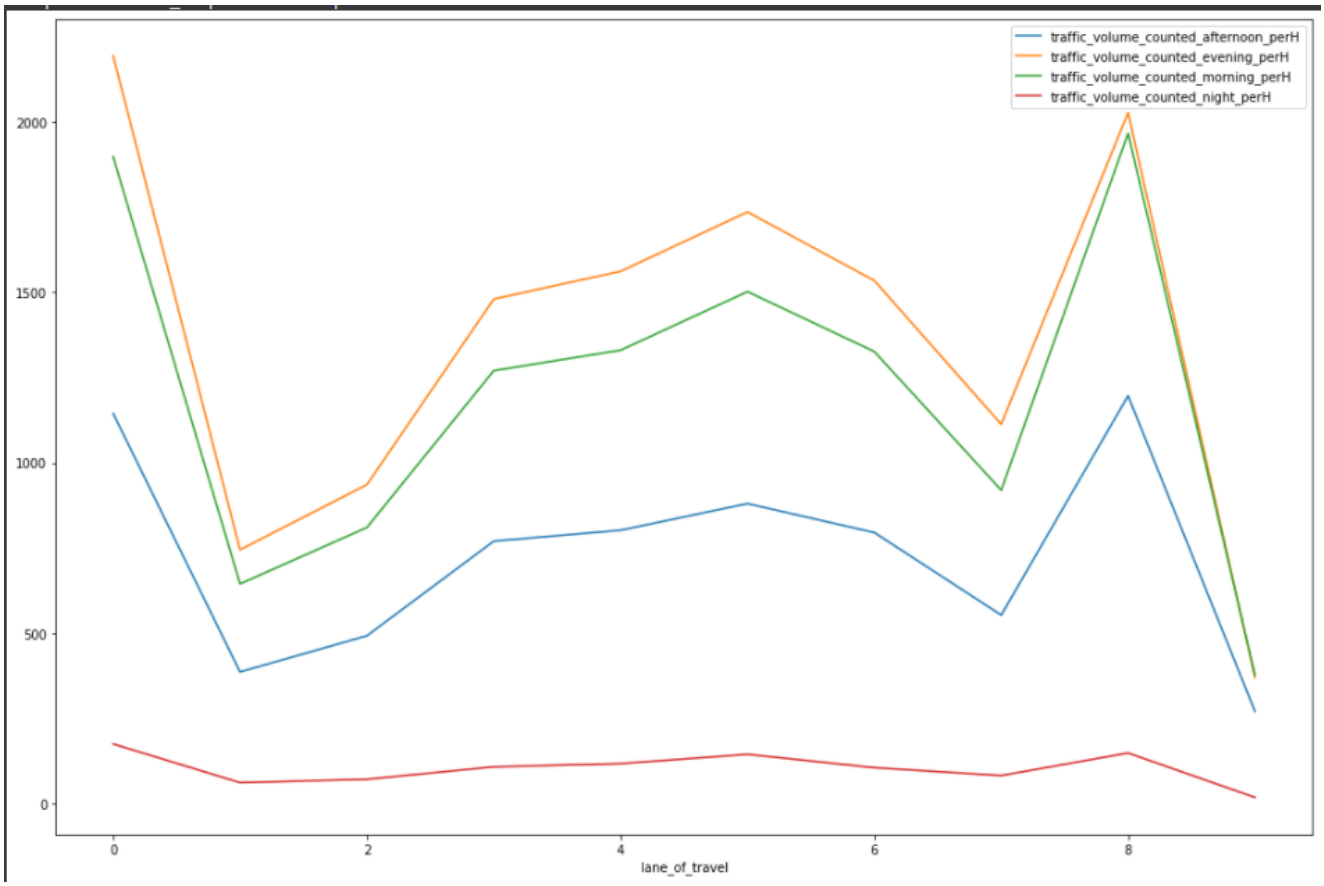
Additional Patterns/Trends

1. **Direction of travel and traffic: (10 travel directions)**
   Conventions: N-North, E-East, W-West, S-South
   - Observation:
     Pivot table, plot and heatmap on directions of travel and NMAE columns resulted that the **maximum mean traffic directions** are **North-South or Northeast-Southwest (N-S or NE-SW) > East-West or Southeast-Northwest (E-W or SE-NW)** with **maximum mean traffic in evening (2.7k)>morning (2.3k)>afternoon (1.4k) > night (170).**
     Whereas, **Northeast** and **Northwest are the directions with least mean traffic**.
   - **Inference:**
     The N-S or NE-SW, E-W or SE-NW directions of various stations and states are found to have heavy traffic in US whereas NE and NW are the lowest traffic directions in general.

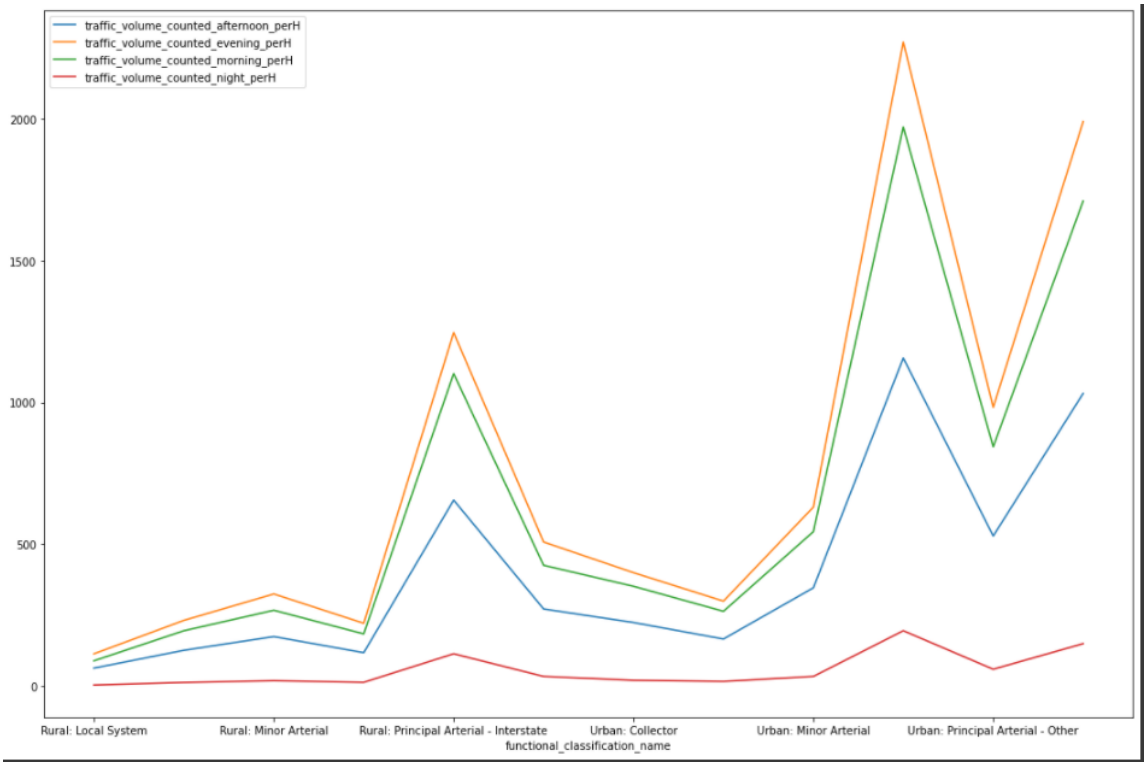


Name: Sanskar Sharma **E-mail: sansskarsharma@gmail.com** **Contact: +91 6261614589**

2. **Lane of travel and traffic: (10 travel lanes)**
    - Observation:
      Pivot table, plot and heatmap on lanes of travel and NMAE columns resulted that the **maximum mean traffic lanes** are **0^{th} >8^{th} >5^{th}** with **maximum mean traffic in evening (2.2k)>morning (1.9k)>afternoon (1.1k) > night (170) for the 0^{th} lane.**
      Whereas, **minimum mean traffic lanes are 9^{th} <1^{st} <2^{nd}** with **minimum mean traffic in night (18)< afternoon (270)< evening (370)< morning (380) for 9^{th} lane.**
    - **Inference:**
      The 0^{th}, 8^{th} and 5^{th} lanes for travelling are heavy traffic lanes whereas the 9^{th}, 1^{st} and 2^{nd} lanes are better for a traveller to take as they have least traffic on them across US.



3. **Functional Classification and traffic: (12 functional classes)**
   These are the various urban and rural area divisions in US.
    - Observation:
      Pivot table, plot and heatmap on functional classes and NMAE columns resulted that the **maximum mean traffic regions** are **Urban: Principal Arterial(PA) - Interstate > Urban: Principal Arterial(PA) – Other Freeways and Expressways** with **maximum mean traffic in evening (2.3k)>morning (2k)>afternoon (1.2k) > night (200) for Urban: PA-Interstate.**
      Whereas, **minimum mean traffic regions are Rural: Local System (LS) < Rural: Major Collector (MC) < Rural: Minor Arterial (MA).**
    - **Inference:**
      The Urban regions are found to have generally more traffic than the Rural ones **except for the Rural: Principal Arterial – Interstate which seems to have more traffic than some minor urban areas.** Urban Principal Arterial (PA) – Interstate and Other Freeways & Expressways have most traffic in US than any other regions whereas Rural: Local System, Major Collector and Minor Arterial seems to be the least traffic regions.



Data Modelling

Simple Route Prediction Model:

- **User input:** Weekday of travel, Month of Travel, Fips state code (US)
- **Output:** 5 Most and 5 least traffic scenarios to keep in mind like direction of travel, functional class and lane of travel.

Complex Evening Traffic Prediction Model for UK state with maximum dispersed traffic count:

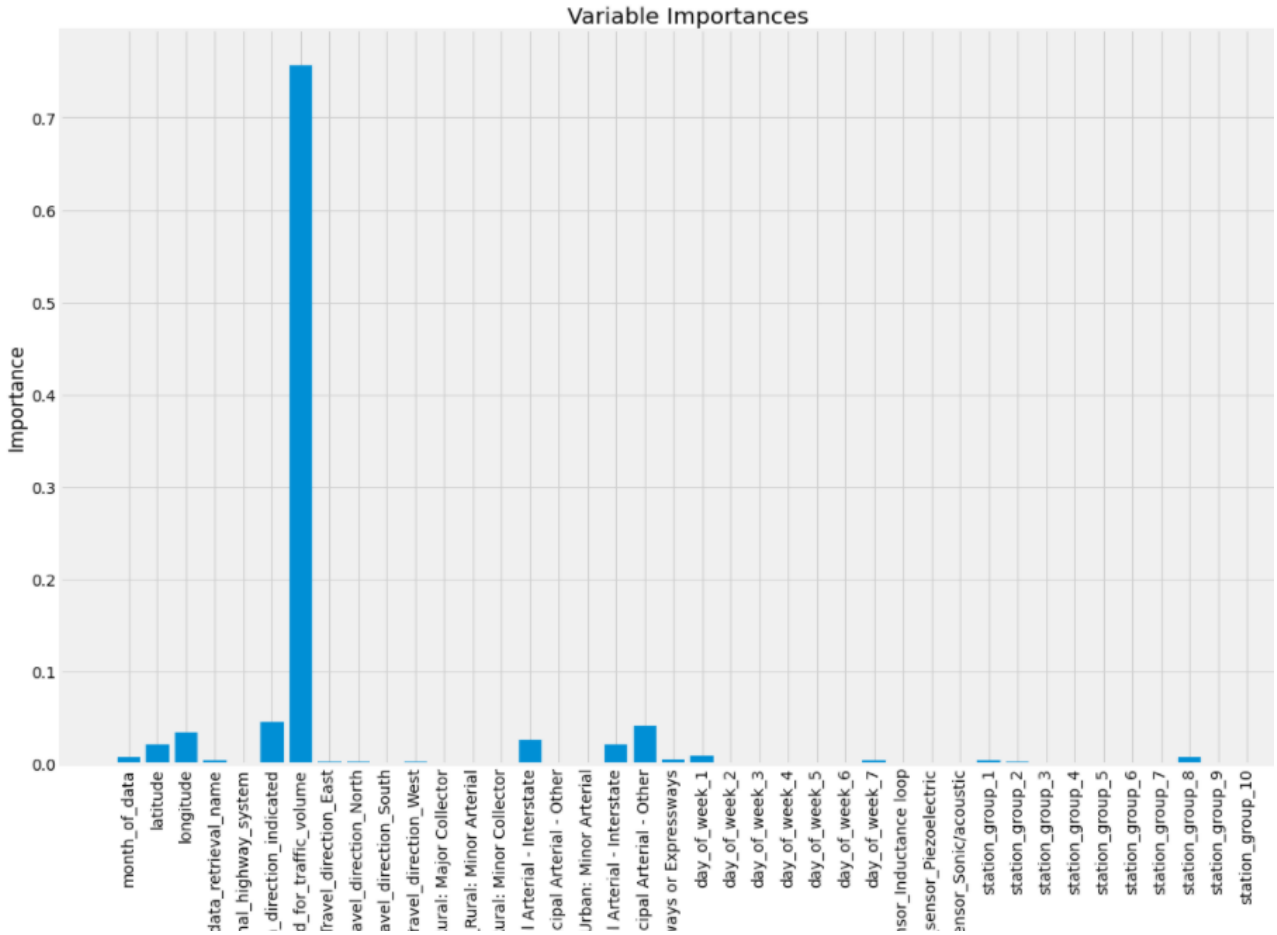
Points to Remember:

- I have predicted only evening (18-24 Hr) traffic as it is maximum compared to others across all states for the year 2015
- Due to large training set and training time for all the states, this prediction model is built only for one UK state (i.e with fips 4)
- State with fips 4 has been selected because of its most dispersed traffic data (maximum variance/ standard deviation) making it difficult for normal humans to predict.
- Also, the dataset used is a combination of both datasets traffic_df and traffic_station_df with suitable columns

1. **Data Preparation:**
   (Note: Data cleaning and handling for both traffic_df and traffic_station_df has been done already)
   - Using pivot table and standard deviation found the US state with most dispersed data i.e. state with "4" as fips id. Filtered both traffic_df and traffic_station_df for fips id "4".
   - Inner Join both the dataset using station_id as the common key and create a new_df which will be treated as final traffic_df.
   - Drop the repeated and unwanted columns like station_location, index_x, index_y etc.
2. **Feature Engineering**:
   As all the features were categorical except for **longitude, latitude, number of lanes in direction indicated and number of lanes monitored for traffic volume.**
   Depending on the type of feature, type of categorical features and number of categories in categorical features different encoding techniques were used.

- Binary Encoding:
  Method of data retrieval contained two categories as automated and manual which were encoded as 1 and 0 respectively.
  National highway system contained two categories as Yes and No which were encoded as 1 and 0 respectively.
- Label Encoding:
  Month is a qualitative nominal variable but as the data is of year 2015 only, there are only 12 months of 2015 which can be ordered in a sequence and can be considered as ordinal.
  1-Jan, 2-Feb….12-Dec.
- One-Hot Encoding:
  Categories of day of week, direction of travel, functional classes, lane of travel and type of sensor were one-hot encoded as columns.
- Feature Hash Encoding:
  The fips state with id "4" contained near about 269 different station ids. This feature is a nominal categorical feature hence cannot be label encoded and being large in number it can't be one-hot encoded as well.
  The station ids categorical sequences of symbolic feature names (strings) into SciPy sparse matrices, using a hash function to compute the matrix column corresponding to a name.
  Hence, 269 categorical features were converted to 10 station groups columns.

3. **Training & Predictions:**
   Sklearn: Random Forest Regressor Class
   - It uses **Standard deviation reduction algorithm** for regression based decision tree models.
   - The object of following Random Forest in-built class in sklearn python library uses 500 decision trees with random state as 42 to training the model.
   - The **Mean Absolute Error** in evening traffic prediction is **140.24 degrees**
   - Calculated the **feature importance** for each feature and observed that out of 40 only 10 features were contributing for the accuracy of the model. Hence, in future models and trainings other non-contributing features can be dropped.



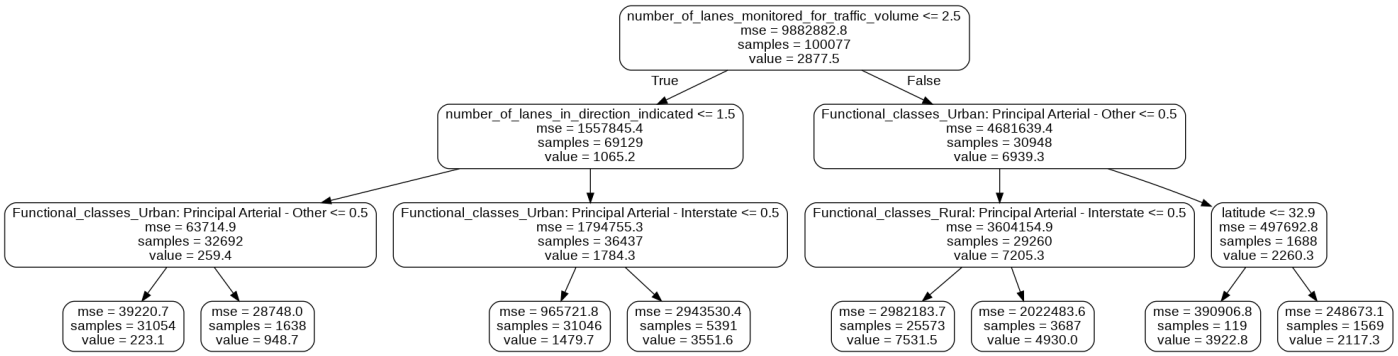
Name: Sanskar Sharma **E-mail: sansskarsharma@gmail.com** **Contact: +91 6261614589**

Most important features were: (Feature importance out of 1)

Number of lanes monitored for traffic volume (0.76), number of lanes in direction indicated (0.05), Functional class_Urban: Principal Arterial – Other (0.04), longitude (0.03), Functional class_Rural: Principal Arterial – Interstate (0.03), latitude (0.02), Functional class_Urban: Principal Arterial – Interstate (0.02), month (0.01), day of week Monday (0.01) and station group 8 (0.01).

Accuracy of the model: 93.26%

Data Model: (A Sample Decision Tree with 3 layer Depth)



Importance of the above Evening Traffic Prediction Model

This model can be used for any of the fips state code as desired by the trainer. As the data set contained 7.1M rows which couldn't be trained completely on my system configurations hence I chose to build it for only a particular state of US.

A user have to enter certain details of his travel route, station ids, day & month of travel and the geographical coordinate of that area and in turn the model will return him the traffic conditions in evening for the same with an **accuracy of 93.26%.** So that he can accordingly revise his travel details as per his/her interests.

Name: Sanskar Sharma **E-mail: sansskarsharma@gmail.com** **Contact: +91 6261614589**