**Report on Approach used to Predict Traffic Volume including Insights about the Data**

We are trying to predict the metro traffic volume based on the weather conditions at different hours of time. We look at each attribute and determine which particular variable is important to our target variable(Traffic volume).  
  
**1.) Date And Time:**

We have data and time column, which has date(YYYY-MM-DD and hh-mm-ss) , so we will extract year month day and hour data in separate columns, and we will drop the data and time column. As we are trying to predict the metro traffic which merely depends on the hour at which people try to board the metro. Just to check if the seasonality matters in the metro traffic volume, there is an test of mean for more than two samples(ANOVA) for each year and the values of traffic in that particular year so that we can check if the mean’s of the metro traffic volume is same or different. And we can see that the values are different which means as the time goes there is a significant difference in the metro traffic.

**2.) Air Pollution Index:**

In Scatter plot between Air Pollution and Traffic Volume that there is no relationship between the air pollution index and the metro traffic volume. Also, as we hover over the graph between traffic volume and the hour column in the tableau file, it is observed that there is no constant rise or fall in air pollution index with increase in the metro traffic volume.

**3.) Humidity:**

In Scatter plot between Humidity and Traffic Volume that there is no relationship between the Humidity and the metro traffic volume. Also, as we hover over the graph between traffic volume and the hour column in the tableau file, it is observed that there is no constant rise or fall in Humidity with increase in the metro traffic volume.

**4.) Wind Speed:**

We have 16 different wind speeds , assuming each value is an category of wind speed, as we can see in the value counts that there are less number of rows for wind speeds > 8 when compared to the other data points, may be we need more data to understand the variable as in how it changes with the metro traffic volume, but as of now we are binning the rows which have wind speeds >8 as a single variable and doing the statistical test of mean for greater than 2 sample(ANOVA) and checking if there is any significant difference in the means of Traffic Volume against Wind Speed. From the statistical test, it can be confirmed there is no Significant difference in Traffic Volume with respect to Wind Speed. We can also check in the tableau file that the average wind speed is almost the same throughout the data irrespective of time, which stays around 3.5, so we can disregard the wind speed column.  
  
**5.) Wind Direction:**

Looking at the values of the column, I don’t understand how direction is a numerical column, so maybe we need more domain knowledge here or we can just stay that when the average wind speed through out the day is same, then it doesn’t matter in which direction in wind is blowing , so we can disregard that particular column as well.

**6.) Visibility & Dew point:**

We look at the data and see that the values for both these columns are same, we need a little domain knowledge as how these 2 are related, but one thing is that visibility is always a priority when you travel so we will dis regard the dew point column and also as the values are same for both the columns one of them is redundant data.

**7.) Temperature:**

We need temperature column in the model, as it also people may also tend to look at the temperature to go out, as in if its too hot or too cool or is it normal temperature to travel, but there were some outliers in the temperature column, which were removed keeping the fact that weather temperature has no specific extreme values, the temperature always lies in one particular range in a particular type of weather, I used the IQR method to remove the outliers.

**8.) Rain\_p\_h & Snow\_p\_h:**

We have a total of 34000 values of which around 30000 data points have zero rain\_p\_h and around 32000 values have zero snow\_p\_h, maybe it is true or maybe we need to change the data accordingly with the remaining columns look where the rain\_ph and snow\_ph values cant be zero and change the values accordingly.

**9.) Clouds\_all:**

There is a little ambiguity about what this column tells us and what it means we need a little domain knowledge , but looking at the current important columns which are useful for modelling it has no co relation with other variables even the target but just to be sure, I also checked the R2\_Score, and there was just a 0.01 difference in the score, as I have no knowledge about that particular variable, I am dis regarding the variable as of now, but maybe in the future if we get some information about it and how useful it is with respect to other variables and also our target, maybe we can include it.

**10.)Weather\_Type & Weather\_Description:**

Looking at the weather description column, we can say that it describes about the particular weather in depth, maybe we can do some NLP and learn if we can get other features out of that column, other than already present in the weather type column, in the weather type column, as each weather type is particularly important, we have included that column and did one-hot encoding on those columns while modelling.