Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: The impact of categorical variables on counts are as follows:

Month: May to October(~10% of total counts) the counts are higher than rest of the months.

Season: Fall,Summer,Winter have higher counts (ranges from 26 to 32%) than Spring.

Weathersit:Clear (69 %)and Misty(30%) Weather have high impact than others.

Weekday: are uniformly distributed across all the months

Workingday: Working day has doubled the counts than non working day.

Holiday: 97% of the counts are on non holidays.

Year: 2019 has higher count (62%) than 2018(38%) – year on year nice growth.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Ans: While creating dummy variable drop_first=True is used to create the first occurrence of the categorical vars in the data as the reference of the dummy var(all zeros).so that it can create dummy vars for rest of the categories of that var.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: variable 'registered' with correlation of .945 with the target variable.

3. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: The assumptions of Linear Regression are:

1. predictors are linearly related with target variable.
2. Errors are normally distributed
3. Errors are having constant variances
4. Errors of the predictors are not correlated with each other.

I have used QQ plot and normal probability of the residual to validates this.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: Weather_Light_Rainsnow : - .3219, Season_Spring : -0.2828, yr: 0.2470,

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear Regression is linear regression is a method of finding the best linear relationship between the independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

There are assumption of the linear regressions:

1. Predictors are linearly related with dependent variable
2. Residual which is the difference between actual observed value and predicted values are normally distributed.
3. Residual are having constant variance which is called homoscedasticity.

4. There is no auto correlation(changes over time) among the residuals means errors are independent.
   The H0: all the predictors coefficients are zero.
   H1 : At least one of the predictor's coefficient is not zero.

Multicollinearity: One of the predictor explain by other predictors, there is a linear relationship among the predictors which explain one of the predictors. In such cases the variable with high Variance Inflation Factor can make the model unstable if we keep that in the model So that variable needs to be dropped to bring stability in the model.

Model with high VIF >5 should be dropped, but it depends on business cases too.

The metrics used to evaluate the models are:

F Stats: Over all significance of the model, p- value of the F stat: significance of the F test.

R squared, adjusted r squared for the explanatory power of the model.

Significance variables based on the trade off between p value (<0.05) and VIF.

Low vif and low p values are acceptable.

2. Explain the Anscombe's quartet in detail. (3 marks)

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Application:

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R? (3 marks)

Ans: The Pearson correlation coefficient (r) is used to identify patterns in relationship between numeric variables.The values are ranges from -1 to 1.

-1 means highly negatively correlated means when one var increase other decreases.

1 means highly positively correlated which means when one var increase other also increase.

It is important to remember that a highly significant value of r does not mean the vars are highly correlated as correlation not always causation.

For example: ice cream sale and drown cases during summer both are high. But they are not related even though correlation could be very high.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Since all variables have different units so it is very difficult for the model to interpret the value and do the convergence, so in order to maintain the interpretability and faster convergence scaling is very much important.

Normalized scaling : all values are normalized min-max scale which is min=0 and ,max=1.

Formula: (x-xmin)/(xmax-xmin)

Standardized Scaling: All the variables are scale with mean zero and stand dev 1.
Formula: (x- mean)/std dev

Standardized scaling are more useful as the denominator is always 1 and scaling value never be 0 or infinity, while in normalized scaling that possibilities are exist.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)
Ans: VIF stands for Variance Inflation Factor, which is basically the variation explained by of one of predictor, considering s target variable and rest of the predictors are independent variable.

VIF=1/1-Ri^2
If the VIF is high which indicates the denominator is low, that is R^2 is higher.
If the VIF is low which indicates the denominator high what is R^ is lower.

A lower VIF is acceptable to avoid multi-collinearity and instability of the model.
So VIF is infinite when one of the predictor explain 100% variability which shows that predictor is linear combination of rest of the predictors in multiple liner regression set up.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Ans: Quantile Quantile plot is used to validate errors are normally distributed or not. If the QQ plot passes through the 45 degree of the line then they are normally distributed.