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?
   - season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
   - mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
   - Weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
   - holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
   - weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
   - workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years).
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   - drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   - Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - The Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - Error terms are normally distributed with mean zero (not X, Y) , Residual Analysis Of Training Data
   - There is a linear relationship between X and Y
   - There is No Multicollinearity between the predictor variables
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

1) Temperature (temp) ,Weather Situation ,Year (yr)

General Subjective Questions

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

1. Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.Linear regression is of the 2 types:

- **Simple Linear Regression:** It explains the relationship between a dependent variable and only one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points .**Formula for the Simple Linear Regression:**$Y=\beta0+\beta1X1 +\epsilon$

- **ii. Multiple Linear Regression:** It shows the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It fits a ‘hyperplane’ instead of a straight line.**Formula for the Multiple Linear Regression:**$Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$

2. The equation of the best fit regression line $Y = \beta_o + \beta_1X$ can be found by the following two methods: Differentiation-· Gradient descent

3. We can use statsmodels or SKLearn libraries in python for the linear regression.

4. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s Quartet was developed by statistician Francis Anscombe. This is a method which keeps four datasets, each containing eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. Each graph tells a different story irrespective of their similar summary statistics. Below is the glimpse of the statistics of the 4 datasets:

5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process to normalize the data within a particular range. Many times, in our dataset we see that multiple variables are in different ranges. So, scaling is required to bring them all in a single range.The two most discussed scaling methods are Normalization and Standardization. Normalization typically scales the values into a range of [0,1]. Standardization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

6. What is Pearson’s R? (3 marks)

Pearson’s R was developed by Karl Pearson and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by ‘r’. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.Mathematically, Pearson’s correlation coefficient is denoted as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a “product moment”, that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

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

The value of VIF is calculated by the below formula:

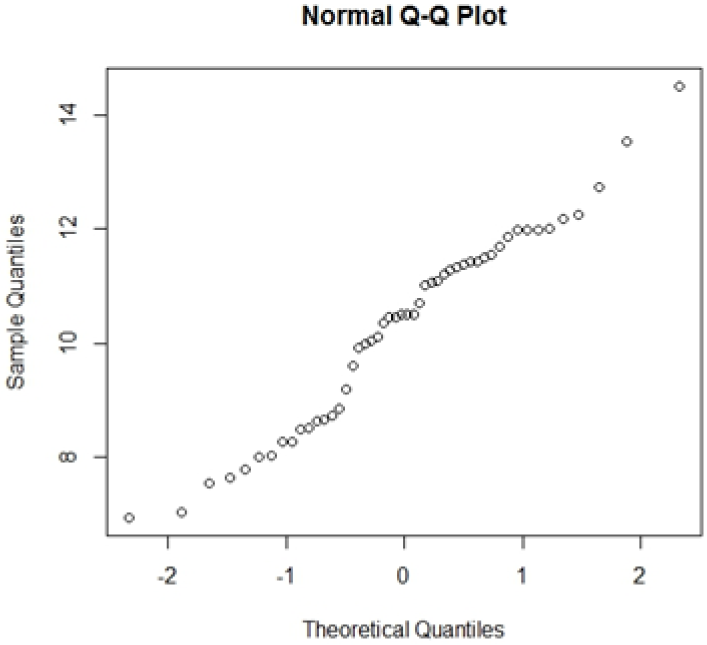
$VIF_i = \frac{1}{1-{R_i}^2}$ Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables.

7. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.