Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:- We inferred the following points from the categorical variable analysis

i. Almost 10% of bookings happened per month in the months from May-Sep. More bikes were rented in 2019 as compared to 2018.

ii. Almost 32% of bookings happened in the Fall season, followed by summer and winter

iii. Bike bookings are more in 'Clear' weather (~68%)

iv. 97% of bookings happened during non-holiday days

v. Roughly 70% of the bookings happened on working days

Q2: Why is it important to use drop_first=Trueduring dummy variable creation?

Ans:- Because it helps in reducing the extra column created during dummy variable creation which reduces the correlations created among dummy variables. And we can still be able to convey the same information with fewer variables, which also improves the efficiency of the model building process.

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans:- In the raw data, 'registered' has the highest correlation of 0.95, whereas if we just talk about the final model, the 'temp' variable has the highest correlation of 0.63

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:- We plotted the distribution of residual (Residual analysis), the plot was centered around zero, hence we can say that means our linear regression model is giving the correct result.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes?

Ans:- The top 3 variables contributing significantly towards explaining the demand for shared bikes along with their coefficients are as follows (not considering the constant):

1. temp : 0.5029
2. yr : 0.2326
3. winter : 0.0829

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans:- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Q2: Explain Anscombe's quartet in detail.

Ans:- Anscombe's Quartet can be defined as a group of four data sets that are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots that have nearly the same statistical observations, which provide the same statistical information that involves variance and mean of all x,y points in all four datasets. This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of dataset.

Q3: What is Pearson's R?

Ans:- Pearson's R OR the Pearson product-moment correlation coefficient (PPMCC), is a measure of the linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

For a population: Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables (X, Y), the formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

cov, is the covariance
sigma x is the standard deviation of X
sigma y is the standard deviation of Y

For a sample:

Pearson's correlation coefficient, when applied to a sample, is commonly represented by r(xy) and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:- It is a step of data Pre-Processing that is applied to independent variables & dependent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is Scaling Performed:

i. Ease of model interpretation

ii. Helps in faster convergence in gradient descent methods

Scaling doesn't affect p-value for a feature and doesn't affect Model accuracy.

There are 2 ways we scale a variable:

i. Standardization: In this, we subtract mean from every value and divide that with Standard Deviation (sigma), such that it is centered at 0 and has a S.D. of 1.

ii. Normalization: It is also called as Min-Max scaling, in this we subtract the min value and then divide by (max(x) - min(x)), where x is the feature variable. The resulting value will always lies between 0 and 1.

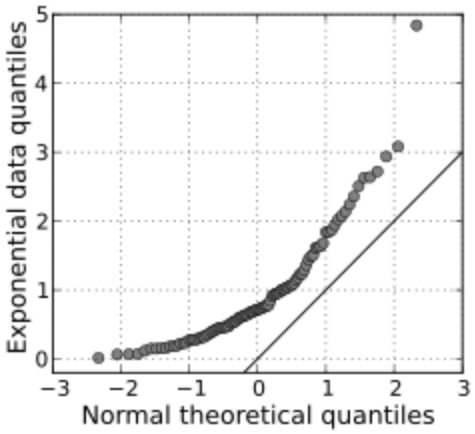
Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans:- If there is a perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 = 1, which leads to 1/(1-R2) becoming infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans:- Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.



A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.