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?

In season column, most of the booking were happen in season 2 and 3 over 5000, and in season 4 above 4000.

In month column, May to October had booking over 4000 which was also a good sign.

In Weathersit column, weathersit_1 Clear, Few clouds, partly cloudy, partly cloudy had more booking compared to other situation.

In holiday column most of bike booking happen when it's not holiday. Weekday variables value were same in all days so it has no significant effect. Year 2019 had more booking then 2018. In Working day column, high booking percentage had 2 years of data

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

In more than two variables we have N numbers of dummy variables which is increase model redundancy, less efficient, less effectiveness in model, and if we can intercept data using N-1 dummy variables why we need N variables, that is why we need to use drop_first=True.

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

Temp and atemp has highest correlation between target variable

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

Based on P-value (0.00) which show all significant variable.

F-statistic value is high (272.9) which show best fit model. Adjusted R2 and R2 value has less than 5 % gap and shows above 80% accuracy. No sign of multicollinearity between variables. AIC value is less which shows best fit for model. Durban-Watson value is 2.0 which shows no autocorrelation between variables. Based on these factors I validated the assumptions.

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

Temp, yr shows positive significantly change on demanding bikes means increase those variables increase bike demand while weathersit_3 shows show negative sign when increase this situation decrease bike demand. Thus Temp, yr, weathersit_3 shows significantly change in bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a machine learning algorithm based on supervised learning that models a target prediction value based on independent variables. It is used for finding out the relationship between variables for prediction.

The algorithm works by finding the best fit line that describes the relationship between the independent and dependent variables.

The line is represented by an equation of the form y = mx + b where y is the dependent variable, x is the independent variable, m is the slope of the line and b is the y-intercept

The slope of the line represents how much we expect y to change as x increases3. The y-intercept represents the predicted value of y when x is 0.

The algorithm can be used for both simple linear regression (one independent variable) and multiple linear regression (more than one independent variable). It has many applications such as predicting stock prices, sales forecasting, weather forecasting, etc.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly the same statistical properties. The datasets demonstrate both graphical data before analyzing and statistical property of data.

With help of these datasets, we can find mean, standard deviation, correlation between x and y. It also shows various anomalies of data like diversity, linear separability of data, data is fit for linear relation with other data or its incapable of handling any other datasets.

3. What is Pearson's R?

Pearson's R correlation coefficient referred to as Pearson's R is a linear correlation between two sets of data, bivariate analysis.

It is the covariance of two variables divided by the product of their standard deviation. Covariance has normalized measurement such result value has -1 to 1. This pcc varies between -1 to 1.

r = 1 means data is perfectly positive linear relationship with slope.

r = -1 means data is negative linear relationship with slope.

r = 0 means there is no relationship.

r > 0 < 5 means there is weak association.

r > 5 < 8 means there is moderate association.

r > 8 means there is a strong association.

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

Scaling is a technique used to standardize the range of independent variables or features of data. It is performed to ensure that all features are uniformly evaluated by the model.

Normalization typically means rescaling the values into a range of [0,1].

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1.

The difference between normalization and standardization is that when scaling, you change the range of your data, while in normalization, you change the shape of the distribution of your data.

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If there is perfect correlation between two independent variables, then VIF = infinity.

In the case of perfect correlation, we get R2 = 1, which lead to 1/ (1-R2) infinity.

To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

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

A Q-Q plot (quantile-quantile plot) is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

The Q-Q plot compares the quantiles of our data against the quantiles of the desired distribution (defaults to the normal distribution, but it can be other distributions too if we supply the proper quantiles).

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.