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?

Answer → there is no significant difference in the mean bike bookings on weekdays and workingdays.

In fall we saw maximum number of bike hires, when there is clear sky, during the months of july and September. All this was highest in the year 2019.

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

Answer → 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.

For example - we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then it is obviously unfurnished. Hence we don't need 3rd variable to identify the unfurnished.

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

Answer → 'temp' and 'atemp' have almost similar correlation with 'cnt'.

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

Answer → after building model on train dataset, we applied the same model on the test set and found similar results, with R^2 value of train set being ~84% and for test set being ~82%.

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

Answer → Temperature, year and weather_sit3 have highest significance on cnt of bike hires.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer → Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

Y = mX + b

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b.

The linear relationship can be positive or negative in nature.

Simple Linear Regression (SLR)

It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

Multiple Linear Regression (MLR)

It is the extension of simple linear regression that predicts a response using two or more features

2. Explain the Anscombe's quartet in detail.

Answer → 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.

3. What is Pearson's R?

Answer → In statistics, the **Pearson correlation coefficient** also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data.

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

Answer → scaling is performed to evenly distribute the values in any numerical variable, especially in cases where the values are far apart from each other.

2 types of scaling –

- MaxScaling (aka normalization) – values lie between 0 and 1
- standardisation → values lie between -1 and 1

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

Answer → If there is 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 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.

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).

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

Answer → 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.