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?

Below are observations of categorical variables from the dataset.

- `Season fall` has the highest cnt followed by `summer`, `winter` and `spring`
- `2019` year growth is higher than the `2018` growth
- non-holiday days contain most of the data (more than 90%).
- non-working day has more data but there is not much difference in the mean
- bike demand when the weather is `clear` is highest followed by `mist` and `light snow`
- No bike demand found for `snow + fog`
- Bike demand during `all weekdays` is almost same
- Bike demand remains unchanged whether it is a working day or not
- Bike demand is highest between the months of `May to Oct`

Q2. Why is it important to use drop_first=True during dummy variable creation?

When we create dummy variable, we drop the first because of below reasons:

- It helps to reduce the extra column created during variable creation.
- It reduces the correlation created among dummy variables
- Eg: We have three variables: Furnished, Semi-furnished and unfurnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don't need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

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

temp and atemp has the highest correlation value. I.e 0.63

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

Below are the parameters that helped to validate the assumption of Linear Regression

- Less Multi-collinearity between features (Low VIF)
- Normal distribution of error terms
- Constant variance of the error
- R^2 of training model and test model are close to each other
- Predictor variable and response is in linear relationship

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

As per our final model, the top 3 predictor variables are

- **Temperature (temp):** Coefficient value of 0.346 indicates that a unit increase in temperature will increase bike hire units by 0.346
- **Year(yr):** Coefficient of 0.237 indicates that increase in year will increase the bike hire units by 0.237
- **weathersit_Light Snow:** negative Coefficient of 0.280 indicates that a unit increase in weather_situation Light Snow will decrease the bike hire by 0.28 units

General Subjective Questions

Q1. Explain the linear regression algorithm in detail?

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Y = mX + C

It can be further divided into two algorithms:

- **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Q2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

The linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Q3. What is Pearson's R?

Pearson's correlation coefficient is 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.

The Pearson correlation coefficient (PCC) is a correlation coefficient that measures 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

What is scaling? It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is scaling performed? Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

- If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.
- A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.
- Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.
- A large VIF on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

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

Q–Q plot (quantile–quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.
Q–Q plots are commonly used to compare a data set to a theoretical model. This can provide an assessment of goodness of fit that is graphical, rather than reducing to a numerical summary statistic. Q–Q plots are also used to compare two theoretical distributions to each other.

The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous

cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF.

Importance: A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals