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? (3 marks)

Analysis on Categorical Variables using countplot, boxplot & barplot.

Barplot used for Categorical variables against count to determine Mean/Med, identify trends and comparison for year 2018, 2019. Also Predictor for Dependent variables.

Season - Most of the bookings do happen in season '3-fall' when compared to others

month - 6-Jun, 9-Sep & 8- Aug do have the highest overall bookings

Weather - This variable plays a major role in determining dependent variable as the bookings have increased significantly when Weather is good and dry. The same subdues if the condition is wet.

Holiday - Most of the bookings were done on a weekday

weekday - All the days in the weekday show similar trend and this variable could influence prediction

workingday - Higher # of bookings were made on a working day and this could be a candidate for determining if this variable to determine the bookings

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

When creating dummy variables from categorical variables, the drop_first=True parameter is used to avoid multicollinearity in the regression model.

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

with the target variable? (1 mark)

(temp, atemp) Continuous variable has the highest correlation with the target variable ‘cnt’

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

training set? (3 marks)
Normality of Error Terms

Multicollinearity check

Independence of residuals

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

- Temp
- Sep
- Windspeed

General Subjective Questions

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

Linear regression is a popular algorithm used in supervised machine learning for predicting continuous numeric values. It models the relationship between one or more independent variables (predictors) and a dependent variable (target) by fitting a linear equation to the observed data.

The basic idea behind linear regression is to find the best-fitting line that minimizes the difference between the predicted values and the actual values of the target variable. The line is defined by the equation:

$y = b0 + b1x1 + b2x2 + ... + bn*xn$

where:

y is the predicted value of the target variable

b0 is the intercept (the value of y when all predictors are zero)

b1, b2, ..., bn are the coefficients (slopes) associated with each predictor variable (x1, x2, ..., xn)

The goal of linear regression is to estimate the values of the coefficients that minimize the sum of squared errors between the predicted and actual values. This process is often done using the method of least squares.

Linear regression assumes a linear relationship between the predictors and the target variable. It also makes several assumptions about the data, such as linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors.

Once the model is trained using the available data, it can be used to make predictions on new data by substituting the predictor values into the equation.

2. Explain the Anscombe’s quartet in detail. (3 marks)
Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but exhibit different patterns when visualized. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization in understanding and analyzing data.

Each dataset in Anscombe's quartet consists of two variables, X and Y, with 11 data points. When looking at the summary statistics of the datasets, such as means, variances, and correlation coefficients, they appear to be very similar. However, when plotted, they reveal different patterns and relationships.

The first dataset is a simple linear relationship with no outliers. The second dataset is also a linear relationship but with one outlier, which has a strong influence on the correlation coefficient. The third dataset appears to follow a non-linear pattern with a clear relationship between X and Y. Finally, the fourth dataset is a perfect fit to a quadratic equation but is often overlooked due to relying solely on summary statistics.

3. What is Pearson's R? (3 marks)

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling refers to the process of transforming numerical variables in a dataset to a specific range or distribution. It is performed to ensure that all variables are on a comparable scale, which helps in improving the performance and stability of machine learning models.

The primary reasons for scaling are as follows:

Feature Comparison: Scaling allows for fair and meaningful comparisons between different features or variables. If the variables are on different scales, the one with a larger scale may dominate the learning process.

Gradient Descent Optimization: Many machine learning algorithms, such as gradient descent-based algorithms, converge faster when the features are on a similar scale. Scaling helps in achieving faster convergence and better optimization.

Distance-Based Algorithms: Algorithms that rely on distance calculations, such as k-nearest neighbors and support vector machines, are sensitive to the scale of the variables. Scaling ensures that the distances are computed accurately and fairly.

There are two commonly used scaling techniques: normalized scaling (also known as min-max scaling) and standardized scaling (also known as z-score scaling).

Normalized Scaling: In normalized scaling, the values of the variables are transformed to a specific range, typically between 0 and 1. The formula for normalized scaling is:

$$X' = (X - X_{min}) / (X_{max} - X_{min})$$

Here, X' is the scaled value, X is the original value, X_min is the minimum value in the dataset, and X_max is the maximum value in the dataset.

Normalized scaling preserves the relative relationships between the variables and maps the minimum value to 0 and the maximum value to 1.

Standardized Scaling: In standardized scaling, the values of the variables are transformed to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

X' = (X - mean) / standard deviation

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

(3 marks)

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

(3 marks)
A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess the distributional similarity between two datasets. In linear regression, a Q-Q plot is commonly used to evaluate the assumption of normality of the residuals.