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?

Categorical variables (season, holiday, weekday, workingday, weathersit, and month) were visualized using boxplots. Their effects on the dependent variable are summarized as follows:

- Season: Demand was highest in fall and lowest in spring.
- Holiday: Rentals decreased during holidays.
- Weekday: Bike demand remained fairly consistent throughout the week.
- Workingday: Maximum bookings occurred between 4000 and 6000, with consistent median user counts throughout the week.
- Weathersit: No users during heavy rain/snow, while clear and partly cloudy weather conditions had the highest counts.

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

Using drop_first=True during dummy variable creation helps prevent multicollinearity issues in regression models by creating $k-1$ dummy variables for a categorical feature with k levels, thus ensuring model stability and reliability.

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

“temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

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

In linear regression analysis:

a. Verify the existence of a linear relationship between independent and dependent variables by visualizing numeric variables using a pairplot.
b. Ensure that the residuals distribution follows a normal distribution centered around 0 by plotting a distplot of residuals.
c. Check for multicollinearity by calculating the Variance Inflation Factor (VIF) to quantify the correlation between feature variables in the model.

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

Temp coefficient 0.5032
Year coefficient 0.2390
Season_winter coefficient 0.0718

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical technique used to understand and quantify the relationship between two or more variables. It aims to find the best-fitting straight line that describes the relationship between the independent variable(s) and the dependent variable. The primary goal is to predict the value of the dependent variable based on the values of the independent variable(s).

In practical terms, linear regression allows us to answer questions such as:

- How does a change in one variable affect another variable?
- Can we predict the value of a dependent variable given certain values of independent variables?

The process of linear regression involves:

Data Collection: Gathering data on the variables of interest.
Visualization: Exploring the relationship between variables using scatter plots or other visualizations.
Model Fitting: Identifying the best-fitting straight line (linear equation) that represents the relationship between the variables. This is done by minimizing the differences between the observed data points and the values predicted by the model.
Evaluation: Assessing the goodness of fit of the model to the data using metrics such as the coefficient of determination (R-squared), mean squared error (MSE), or other relevant measures.
Interpretation: Interpreting the coefficients of the linear equation to understand the strength and direction of the relationship between variables.
Prediction: Using the fitted model to make predictions on new or unseen data.

Linear regression is widely used in various fields, including economics, finance, social sciences, engineering, and more. It provides valuable insights into the relationships between variables and helps in making informed decisions based on data analysis.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous example in statistics that demonstrates the importance of visualizing data and the potential pitfalls of relying solely on summary statistics. It consists of four datasets, each containing 11 (x, y) points, which have nearly identical statistical properties but vastly different visual appearances when plotted.

The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the limitations of relying solely on summary statistics (such as means, variances, and correlation coefficients) to describe data. Despite having the same mean, variance, correlation coefficient, and linear regression line, the datasets in Anscombe's quartet exhibit different shapes, patterns, and relationships when plotted.

By examining Anscombe's quartet, one can understand the importance of visualizing data and how different datasets can lead to different conclusions if not carefully analyzed. It highlights the potential dangers of relying solely on summary statistics without examining the underlying data distribution and patterns.

In summary, Anscombe's quartet serves as a cautionary example in statistics, emphasizing the importance of exploring and visualizing data thoroughly to gain a comprehensive understanding of its characteristics and relationships.

3. What is Pearson’s R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear rela□onship between two sets of data. In simple terms, it tells us “can we draw a line graph to represent the data?”

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.
● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normaliza□on, standardization does not have a bounding range. So, even if you have outliers in your data, they will not

be affected by standardization.

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

The occurrence of infinite values for the Variance Inflation Factor (VIF) typically indicates perfect multicollinearity among the predictor variables in a regression model. Perfect multicollinearity means that one or more independent variables in the model can be perfectly predicted from a linear combination of the other independent variables.

Perfect multicollinearity leads to issues when calculating the inverse of the matrix used in the computation of VIF. Mathematically, when perfect multicollinearity exists, one or more of the variables will have a zero determinant, resulting in an infinite VIF value.

Perfect multicollinearity can occur due to various reasons, such as:

Data duplication or repetition: Having duplicate or nearly identical variables in the dataset can lead to perfect multicollinearity.
Linear dependencies: When one variable can be expressed as a linear combination of other variables, it leads to perfect multicollinearity.
Incorrect model specification: Including variables that are not meaningful or relevant to the model can introduce multicollinearity issues.

To address infinite VIF values, it's essential to identify and resolve the root cause of perfect multicollinearity. This may involve:

- Removing redundant or highly correlated variables from the model.
- Transforming variables or creating new variables to reduce multicollinearity.
- Re-evaluating the model specification and ensuring that only meaningful variables are included in the analysis.

Handling multicollinearity effectively is crucial for obtaining reliable and interpretable results from regression analysis.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions .A Q-Q plot is a scaterplot created by ploting two sets of quantiles against one another. If both sets of quantiles came from the same distribu□on, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

Do two data sets come from populations with a common distribution?

Do two data sets have common location and scale?

Do two data sets have similar distributional shapes?

Do two data sets have similar tail behavior?