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)

Based on the boxplot analysis of categorical columns, we can draw several conclusions:

1. Fall season was the most popular, with a high number of bookings, while spring season had significantly fewer bookings compared to other seasons.
2. The booking amount generally increased from April to the middle of the year and then gradually decreased towards the end of the year. This suggests that booking rates are typically higher during the mid-year period.
3. There is a noticeable sharp decline in bookings as weather conditions worsen, indicating that people are less likely to rent bikes during unfavorable weather.
4. Mid-weekdays showed slightly higher bookings compared to other weekdays, but there were no significant variations observed across weekdays.
5. There was a relatively equal number of bookings on both working days and non-working days, suggesting that the rental demand remained consistent regardless of the day type.
6. The data showed a consistent linear increase in bike rentals from 2018 to 2019, indicating a positive yearly growth trend in the bike rental business.

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

When creating dummy variables for categorical variables, we represent the categories using binary variables (0/1). However, including all the created dummy variables in the model can lead to multicollinearity, where the predictor variables become highly correlated with each other.
To address this issue, we can set the parameter drop_first=True when creating the dummy variables. This means that for each categorical feature, we drop the first created dummy variable. By doing so, we ensure linear independence among the variables and prevent redundancy in the model.
Dropping the first dummy variable also serves another purpose. It designates the first category as the reference category, meaning it is not explicitly included as a dummy variable. Instead, its absence or presence is inferred from the presence or absence of the other categories. This choice helps us better understand the effects of the different categories by comparing them to the reference category. In simpler terms, setting drop_first=True allows us to discern how each category differs from the category that is not explicitly included.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
The variables "atemp" and "temp" show the strongest correlation, ranging from 62% to 65%, with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

I have confirmed the assumptions of the Linear Regression Model by conducting the following validations:

- The errors exhibit a nearly normal distribution, with the mean of residuals being close to zero.
- To ensure no significant multicollinearity among variables, I verified that all independent variables in the model have a Variance Inflation Factor (VIF) score below 5. High VIF scores indicate the presence of multicollinearity.
- I validated the presence of a linear relationship between predictors and the independent variable by examining an Actual vs Predicted plot, which demonstrated a rough diagonal line indicative of linearity.
- Homoscedasticity was verified by examining the Residual vs Fitted plot (Error vs Actual) for the absence of any discernible pattern in residual values. This ensures that the error variance remains constant throughout the plot.
- To confirm the independence of residuals, I performed the Durbin-Watson test, which yielded a score close to 2. This result indicates no autocorrelation between residuals.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for shared bikes? (2 marks)

The top three features that have the greatest impact on explaining the target variable are year, perceived temperature, and the weathersit_Light Snow.

General Subjective Questions

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

Linear regression is a commonly used statistical algorithm that models the relationship between a dependent variable and one or more independent variables. Its goal is to find the best-fitting straight line or hyperplane that minimizes the overall distance between observed data points and predicted values. Here's a simplified explanation of the linear regression algorithm:

Data preparation:

1. Gather the dataset with the dependent variable and independent variables.
2. Ensure the data is in numerical format and handle missing values or outliers.

Model representation:

1. Assume a linear relationship between the dependent and independent variables.
2. Represent the model as $Y = \beta0 + \beta1X + \varepsilon$ for single-variable regression or $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betapXp + \varepsilon$ for multiple-variable regression.

Model training:

1. Estimate the values of the coefficients ($\beta0, \beta1, \beta2, ..., \betap$) using ordinary least squares (OLS).
2. OLS minimizes the sum of squared differences between observed and predicted values.

Model evaluation:

1. Assess the quality and performance of the model.
2. Calculate metrics like R2, RMSE, or MAE to measure the accuracy and precision of predictions.

Model utilization:

1. Use the trained model to make predictions on new data by inputting values for the independent variables.
2. Conduct inference to determine the significance and impact of independent variables on the dependent variable.

Linear regression is widely applied in fields like economics, finance, and social sciences. It relies on assumptions:

1. Linearity: The relationship between variables is assumed to be linear.
2. Independence: Observations are independent, with no systematic relationship between residuals.
3. Homoscedasticity: Residual variance is constant across different values of independent variables.
4. Normality: Residuals follow a normal distribution.
5. No multicollinearity: Independent variables are not highly correlated with each other.

These assumptions ensure the validity and accuracy of the model's results.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a collection of four datasets that share similar statistical characteristics but display distinct patterns when plotted. Created by statistician Francis Anscombe in 1973, these datasets serve as a reminder of the importance of visualizing data and the limitations of relying solely on summary statistics. Each dataset in Anscombe's quartet consists of 11 (x, y) data points. Here are the characteristics of the four datasets:

Dataset I:

- Shows a relatively linear relationship, resembling a simple linear regression.
- Summary statistics (mean, variance, correlation coefficient) closely match those of Dataset II.

Dataset II:

- Exhibits a non-linear relationship following a quadratic curve when plotted.
- Despite the non-linearity, the summary statistics are similar to those of Dataset I.

Dataset III:

- Appears to have a linear relationship with an outlier that significantly influences the regression line and summary statistics.
- The outlier's impact makes the correlation coefficient similar to the previous datasets.

Dataset IV:

- Similar to Dataset III, it displays a strong linear relationship except for a different outlier at the end.
- The outlier dramatically alters the regression line and correlation coefficient, highlighting their sensitivity.

The purpose of Anscombe's quartet is to emphasize that visualizing data is essential alongside summary statistics. Despite having similar statistical properties, these datasets exhibit distinct patterns when visually examined. This demonstrates that relying solely on summary statistics can lead to misleading conclusions about the underlying relationships in the data.

Anscombe's quartet serves as a reminder that exploring and visualizing data are crucial steps in understanding the complexities that may not be evident from summary statistics alone.

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

Pearson's correlation coefficient, also known as Pearson's R, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It is represented by a

value between -1 and +1. A positive correlation coefficient indicates that the variables tend to move in the same direction, while a negative correlation coefficient suggests that the variables move in opposite directions, with low values of one variable associated with high values of the other.
The calculation of Pearson's R involves several steps:

1. Standardize the variables to make them comparable.
2. Calculate the covariance, which measures how the variables vary together.
3. Compute the correlation coefficient using the covariance and the standard deviations of the variables. Mathematically, Pearson's R is expressed as: R = Covariance(X, Y) / (Standard Deviation(X) * Standard Deviation(Y))

In summary, Pearson's correlation coefficient is a valuable tool for assessing the linear relationship between continuous variables. It helps determine the strength and direction of the relationship based on a range of -1 to +1, and its calculation involves standardizing the variables, calculating the covariance, and deriving the correlation coefficient using the standard deviations.

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

Scaling is a process that transforms numerical variables to a standardized range or distribution. Its purpose is to adjust the values of variables so that they fall within a specific scale, enabling direct comparison and avoiding biases in analysis.
There are two common types of scaling: normalized scaling (min-max scaling) and standardized scaling (z-score scaling). Here are their key differences:
Normalized scaling:

- Rescales variables to a specific range, typically between 0 and 1.
- The normalized value of each data point is calculated using the minimum and maximum values of the variable.
- Formula: normalized_value = (value - min) / (max - min).
- Maintains the original distribution shape while compressing the variable's range.

Standardized scaling:

- Transforms variables to have a mean of 0 and a standard deviation of 1.
- Each data point is subtracted by the mean value and divided by the standard deviation.
- Formula: standardized_value = (value - mean) / standard_deviation.
- Results in a distribution centered around 0 with a spread of 1.
- Maintains the shape of the distribution while changing the scale.

The choice between normalized scaling and standardized scaling depends on specific requirements and data characteristics. Normalized scaling is preferred when preserving the actual minimum and maximum values of the variable is important. Standardized scaling is commonly used when focusing on the relative position of each data point in the distribution or when algorithms or analyses require variables to have a mean of 0 and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

In certain cases, the Variance Inflation Factor (VIF) can take on an infinite value. This situation arises when there is perfect multicollinearity among the independent variables in a linear regression model. Perfect multicollinearity refers to a scenario where one or more independent variables can be precisely predicted by a linear combination of the other independent variables.

When perfect multicollinearity exists, the calculation of VIF breaks down as it involves dividing by zero. The VIF formula includes the computation of the variance of an independent variable when the model is fitted with that variable as the dependent variable, while considering all other independent variables as predictors. However, if perfect multicollinearity is present, the model cannot be properly fitted because one of the variables is a linear combination of the others. As a result, the variance becomes infinite, leading to an infinite VIF value.

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, or quantile-quantile plot, is a graphical tool used to assess whether a given dataset follows a specific probability distribution. In linear regression, a Q-Q plot is often used to examine the normality assumption of the residuals in the regression model.
The use of a Q-Q plot involves comparing the quantiles of the observed data with the quantiles expected from a theoretical distribution, typically the normal distribution. The observed data is sorted and corresponding quantiles are calculated. The expected quantiles are determined based on the probabilities in the theoretical distribution. The observed quantiles are plotted on the y-axis, while the expected quantiles are plotted on the x-axis.
In a Q-Q plot, if the observed data follows the theoretical distribution, the points on the plot should approximately fall along a straight line (45-degree reference line).
The importance of a Q-Q plot in linear regression lies in assessing the normality assumption of the residuals. Normality assumption is a key requirement in linear regression, and by visually examining the Q-Q plot of the residuals, one can identify departures from normality. If the points on the plot deviate from the 45-degree reference line systematically, it indicates a deviation from normality.
By analyzing the Q-Q plot of the residuals, researchers and analysts can gain insights into the normality assumption, detect potential issues, and take necessary actions to address them. This leads to more reliable and robust results in linear regression analysis.