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)

ans) I conducted an analysis of the categorical variables using boxplots and bar plots. From the visualizations, we can infer the following insights:

- The fall season appears to have witnessed higher booking activity, with a significant increase in booking counts from 2018 to 2019 across all seasons.
- A majority of the bookings occurred during the months of May, June, July, August, September, and October. The trend showed a rise in bookings from the beginning of the year, peaking mid-year, before declining toward the year-end.
- Clear weather conditions attracted a higher volume of bookings, which aligns with general expectations.
- Thursdays, Fridays, Saturdays, and Sundays saw more bookings compared to the beginning of the week.
- Bookings tended to be lower on non-holidays, which is understandable as people likely prefer staying home and spending time with family on holidays.
- There was little difference in the number of bookings between working days and non-working days.
- The year 2019 saw a noticeable increase in booking activity compared to 2018.

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

Ans) Using `drop_first=True` during dummy variable creation is important because it helps prevent multicollinearity by reducing the number of dummy variables. It removes one category from the dummy variables, ensuring only `k-1` dummies are created for `k` categories, which avoids redundancy and correlation among the dummy variables.

Syntax Explanation:
`drop_first`: A boolean, defaulting to `False`. When set to `True`, it creates `k-1` dummies out of `k` categorical levels by dropping the first level.

For example, if a categorical column has three categories (A, B, C), creating dummy variables for all three would be unnecessary. If a value is neither A nor B, it must be C by default, so we don't need a dummy variable for C.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

ans) ‘temp’ variable has the highest correlation with the target variable.

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

Ans) I validated the assumptions of the Linear Regression model based on the following five key assumptions:

- **Normality of Error Terms**:
  The residuals should follow a normal distribution.
- **Multicollinearity Check**:
  There should be minimal multicollinearity among the independent variables to ensure reliable coefficient estimates.
- **Linearity Validation**:
  A linear relationship should exist between the independent and dependent variables.
- **Homoscedasticity**:
  The residuals should exhibit constant variance, with no discernible pattern in the residual plot.
- **Independence of Residuals**:
  The residuals should be independent, with no signs of autocorrelation.

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

ans) Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –
temp
winter
sep

General Subjective Questions

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

Ans) Linear regression is a statistical model used to analyze the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship, where changes in the independent variable(s) directly affect the dependent variable. The relationship can be expressed by the equation:

$Y=mX+c$

Where:

- `Y` is the dependent variable (the output we predict),
- `X` is the independent variable (input data),
- `m` is the slope (representing the impact of X on Y),
- `c` is the intercept (the value of Y when X is zero).

There are two types of linear regression:

- **Simple Linear Regression** (involving one independent variable),
- **Multiple Linear Regression** (involving multiple independent variables).

Assumptions:

- **Linearity**: A linear relationship between the dependent and independent variables.
- **No Multicollinearity**: Independent variables should not be highly correlated.
- **Homoscedasticity**: Constant variance of the error terms.
- **No Autocorrelation**: No correlation between residuals.

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

Ans)**Anscombe's Quartet** is a collection of four datasets. These datasets have nearly identical summary statistics (such as mean, variance, and correlation) but reveal vastly different relationships when visualized through scatter plots. The purpose of the quartet is to emphasize the importance of data visualization in statistical analysis, showing that relying solely on numerical summaries can be misleading.

Key Points:
- Each dataset contains 11 (x, y) data points.
- The descriptive statistics for all four datasets are almost identical:
  - Mean of `x` is 9 for all datasets.
  - Mean of `y` is 7.5 for all datasets.
  - Variance of `x` is 11, and variance of `y` is 4.12 for all datasets.
  - Correlation between `x` and `y` is 0.816 for all datasets.
  - The linear regression line for each dataset is nearly identical.

Differences Revealed Through Visualization:
1. **Dataset I**: Displays a linear relationship, where the points are scattered fairly evenly around the regression line, showing a proper linear trend.

2. **Dataset II**: Although the summary statistics suggest a linear trend, the scatter plot reveals a curvilinear relationship, indicating that the data do not fit well with a straight line.

3. **Dataset III**: The plot shows a linear trend but is heavily influenced by a single outlier, which dramatically alters the regression line.

4. **Dataset IV**: The scatter plot shows that most of the data points have the same x-value, with one significant outlier. The correlation is driven by this outlier, although the dataset doesn't show a meaningful linear relationship.

Importance of Anscombe's Quartet:
- **Highlight the Limitations of Descriptive Statistics**: Despite similar statistical measures, the visualizations show distinct patterns in each dataset, demonstrating that summary statistics alone do not capture the full story of the data.
- **Emphasizes the Need for Visualization**: Anscombe's Quartet underscores the importance of graphing data before making assumptions or drawing conclusions. Visualizations help identify outliers, nonlinear relationships, and other underlying patterns that might not be evident from summary statistics.

In conclusion, Anscombe's Quartet illustrates the importance of combining numerical analysis with graphical exploration to gain a complete understanding of the data.

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

Ans)**Pearson's R**, also known as the **Pearson correlation coefficient**, measures the strength and direction of a linear relationship between two continuous variables. Its value ranges from **-1 to +1**, where:
- **+1** indicates a perfect positive correlation, meaning as one variable increases, the other also increases in a perfectly linear fashion.
- **-1** indicates a perfect negative correlation, meaning as one variable increases, the other decreases.
- **0** indicates no linear relationship between the variables.

Pearson's R is calculated using the covariance of the variables divided by the product of their standard deviations. It helps determine whether a change in one variable predicts a corresponding change in another. For example, a Pearson's R of **0.85** between height and weight suggests a strong positive correlation—taller individuals tend to weigh more.

However, Pearson's R only captures **linear relationships**. Non-linear relationships between variables may not be well-represented by this coefficient. Additionally, it is **sensitive to outliers**, which can distort the true relationship.

Pearson's R is widely used in data analysis to quantify associations between variables, but it should be complemented by visualizations, such as scatter plots, to better understand the data and its potential patterns.

Example:
If the Pearson's R value between height and weight is **0.85**, it indicates a strong positive linear relationship, meaning taller individuals tend to weigh more.

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

Scaling is the process of standardizing the range of independent variables in a dataset. It is important because models like linear regression and gradient descent algorithms can be sensitive to the scale of input features.

Why scaling is performed:

- To prevent larger magnitude features from dominating the model.
- To make the model training more efficient.

Difference between normalized and standardized scaling:

normalized	**standardized**
Minimum and maximum value of features are used for scaling	**Mean and standard deviation is used for scaling.**
It is used when features are of different scales.	**It is used when we want to ensure zero mean and unit standard deviation.**
Scales values between [0, 1] or [-1, 1].	**It is not bounded to a certain range.**
It is really affected by outliers.	**It is much less affected by outliers.**

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Ans)
The value of **Variance Inflation Factor (VIF)** becomes infinite when there is perfect multicollinearity between two or more independent variables. In this case, the R-squared value for a regression of one variable on others is 1, causing VIF to approach infinity.

This happens because one variable can be perfectly predicted by another. To resolve this, you can drop one of the highly correlated variables to reduce multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans)
Use of Q-Q Plot:
A Q-Q plot (quantile-quantile plot) visualizes the quantiles of one dataset against the quantiles of another. A quantile represents the fraction (or percentage) of data points below a specific value. For instance, the 0.3 (or 30%) quantile indicates that 30% of the data points are below this value, while 70% are above it. The plot typically includes a 45-degree reference line. If both datasets originate from populations with the same distribution, the points will roughly align along this reference line. A significant deviation from this line suggests that the two datasets are likely from different distributions.

Importance of Q-Q Plot:
When comparing two data samples, it is often essential to assess whether the assumption of a shared distribution holds. If the assumption is valid, location and scale estimators can combine both datasets to produce estimates for the common location and scale. Conversely, if the samples are different, understanding these differences becomes important. A Q-Q plot offers insights into the nature of the differences that may not be as clear through analytical methods such as the chi-square test or the Kolmogorov-Smirnov two-sample test.