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)

   The data clearly illustrates a growth in customer numbers from 2018 to 2019, as evidenced by the graphical representations.
   Across both years, the Fall Season emerges as the peak period for bike rentals, consistently outperforming other seasons.
   The company experiences robust growth up to September, marked by a notable surge in bike rental orders during this month. However, there's a subsequent decline in rentals during the following three months.
   When analyzing rental patterns on a daily basis, it's observed that while the median rental values remain fairly consistent across all days, Fridays exhibit slightly higher rental activity compared to other days.
   Notably, favorable weather conditions such as Clear skies, Few clouds, Partly cloudy, and Partly cloudy are associated with increased bike rental bookings. Conversely, there tends to be a downtrend in bike rental orders during periods of less favorable weather conditions.

---

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

   It helps avoiding Multi-Collinearity issue when we make Dummy Variables from Categorical Variables. As Multicollinearity occurs when two or more predictor variables are highly correlated, which can lead to unstable coefficient estimates and difficulties in interpreting the model.
   So Dropping First variables we prevent multi collinearity between dummy variables.
   Ex we have weather_status as column which have 3 types of weather i.e. good, average, worst then we can make dummy variables out from it
   - 0 0 will corresponds to good
   - 0 1 will corresponds to average
   - 1 0 will corresponds to worst

So we do not need all three. We can conclude with 2 dummy variables too .

---

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

   I can see “temp” & “atemp” is highly co-related with “count_Total”

---

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

   We can validate it by checking
   Error terms are normally distributed with mean 0
   It does not follow any pattern
   Linearity Check
   And by checking VIF values

---

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

   I got 3 Features as :-
   Temp
   Windspeed
   Good Weather (i.e. in weathersit – 1. -- Clear, Few clouds, Partly cloudy, Partly cloudy)

General Subjective Questions:

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

Linear regression is a widely used statistical technique for modeling the relationship between a dependent variable (often denoted as *Y*) and one or more independent variables (often denoted as *X*). It assumes a linear relationship between the independent variables and the dependent variable. The basic form of linear regression can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_N X_N + \epsilon$$

Y is the dependent variable
β_0 is the intercept
β_1, β_2 ... are the coefficients
X_1 , X_2 ... are the independent variables
The goal of linear regression is to estimate the coefficients $(\beta_1, \beta_2 \ldots, \beta_n)$ that minimize the difference between the observed and predicted values of the dependent variable. This is typically done using the method of least squares, where the sum of the squared differences between the observed and predicted values is minimized.
Steps involved:
1). Data Preparation
2). Model Building
3). Model Training
4). Model Evaluation
5). Model Interpretation
Overall, linear regression is a powerful and widely used technique for modeling the relationship between variables and making predictions based on that relationship. It is relatively simple, interpretable, and can be applied to a wide range of problems in various fields such as finance, economics, healthcare, and social sciences.

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 when analyzed using common statistical measures such as mean, variance, correlation, and linear regression coefficients. Despite their statistical similarity, the datasets differ significantly when visualized, highlighting the importance of data visualization in understanding and interpreting data.

The quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the limitations of relying solely on numerical summaries without visualizing the data. Here's an overview of the four datasets in Anscombe's quartet:

1. **Dataset I**: This dataset consists of linearly related variables with a strong positive correlation. When plotted, it resembles a simple linear relationship.
2. **Dataset II**: Similar to Dataset I, this dataset also has a linear relationship between variables but with an outlier that significantly influences the regression line. Despite the outlier, the statistical summary remains similar to Dataset I.
3. **Dataset III**: In contrast to the first two datasets, Dataset III shows a non-linear relationship between variables. It follows a quadratic curve pattern.
4. **Dataset IV**: Dataset IV demonstrates the influence of an outlier on correlation coefficient and linear regression. It consists of three clusters of data points with nearly identical statistical properties, but the presence of an outlier creates a misleading linear relationship.

Key points to note about Anscombe's quartet:

The quartet illustrates that relying solely on numerical summaries like mean, variance, and correlation can be misleading and may not provide a complete understanding of the data.

Visualizing data through plots such as scatter plots, histograms, and box plots is crucial for gaining insights and identifying patterns or anomalies.

Anscombe's quartet emphasizes the importance of exploring data visually before drawing conclusions or making decisions based on statistical summaries alone.

The quartet is often used in statistical education to teach the importance of data visualization and the limitations of summary statistics.

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

Pearson's *r*, commonly known as Pearson correlation coefficient, is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson in the late 19th century and is one of the most widely used measures of correlation.

Pearson's correlation coefficient ranges from -1 to 1:

- A value of -1 indicates a perfect negative linear relationship, where one variable increases as the other decreases.
- A value of 1 indicates a perfect positive linear relationship, where both variables increase or decrease together.
- A value of 0 indicates no linear relationship between the variables.

Formula:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

- $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y respectively.
- The numerator represents the covariance between X and Y, while the denominator represents the product of the standard deviations of X and Y.

---

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

Scaling is a preprocessing technique used in machine learning and data analysis to standardize or normalize the features (independent variables) of a dataset. Scaling ensures that all features have similar scales or ranges, which can be crucial for certain algorithms to perform effectively.

<u>Why Scaling is Performed:</u>

- Improved Model Performance
- Faster Convergence

- Interpretability

Difference between Both:

1. Normalized Scaling (Min-Max Scaling): Normalized scaling transforms the values of features to fit within a specified range, typically between 0 and 1. It is performed using the formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

X_{max} and X_{min} are the minimum and maximum values of the feature respectively.

2. Standardized Scaling (Z-score Standardization): Standardized scaling transforms the values of features to have a mean of 0 and a standard deviation of 1. It is performed using the formula:

$$X_{std} = \frac{X - \mu}{\sigma}$$

μ and σ are mean and standard deviation.

Normalized scaling transforms the features to a specific range, usually between 0 and 1, preserving the original distribution of the data. Standardized scaling centers the data around zero with a standard deviation of 1, making it more suitable for algorithms that assume a Gaussian distribution or have regularization components.

---

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

The occurrence of infinite VIF values typically arises due to perfect multicollinearity among the independent variables in the regression model. Perfect multicollinearity occurs when one or more independent variables can be exactly predicted by a linear combination of other variables in the model. In such cases, the VIF calculation involves division by zero, leading to an infinite VIF value.

Perfect multicollinearity can manifest in various scenarios, such as:

- Redundant Variables
- Dummy Variable Trap
- Linear Dependence

In practice, infinite VIF values indicate a severe problem in the regression model and must be addressed. Solutions may involve:

Variable Removal: Identifying and removing one or more variables that cause perfect multicollinearity.

Feature Engineering: Transforming or combining variables to resolve multicollinearity issues.

Regularization: Applying regularization techniques such as ridge regression or Lasso regression to mitigate multicollinearity.

Addressing infinite VIF values is crucial for ensuring the stability and reliability of the regression model, as it indicates that the model's estimates are unstable and may not be valid for inference or prediction.

---

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, short for Quantile-Quantile plot, is a graphical tool used to assess whether a given set of data follows a specific probability distribution, such as the normal distribution. It compares the quantiles of the empirical data distribution against the quantiles of a theoretical distribution (e.g., normal distribution) expected if the data were normally distributed.

   Use and Importance of Q-Q plot in Linear Regression:
   Assessment of Normality Assumption: In linear regression, it is often assumed that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots provide a visual means to assess this assumption by plotting the observed quantiles of the residuals against the expected quantiles of a normal distribution. If the points fall approximately

along the diagonal line, it suggests that the residuals are normally distributed, supporting the assumption of normality.

Identification of Outliers and Skewness: Deviations from the diagonal line in a Q-Q plot can indicate departures from normality, such as skewness or the presence of outliers. Outliers may appear as points that deviate significantly from the expected line, while skewness may manifest as curvature or non-linearity in the plot.

Model Diagnostics: Q-Q plots serve as a diagnostic tool to identify potential issues with the linear regression model. If the residuals violate the normality assumption, it can affect the validity of statistical inferences and predictions made by the model. By examining the Q-Q plot, researchers can detect such violations and take appropriate steps to address them, such as transforming variables or using alternative modeling techniques.

In summary, Q-Q plots play a crucial role in linear regression analysis by helping to assess the normality assumption of residuals, identify outliers and skewness, and diagnose potential problems with the regression model. They provide a visual aid that aids in the interpretation and validation of the regression results, ultimately leading to more accurate and reliable conclusions.