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?

Ans: season, yr, mnth, holiday, weekday, weathersit are categorical variables in the dataset. From the analysis, it can be inferred that

- Fall is the season to get maximum active customers (September being the month). 2019 observed more sale than 2018.
- Holidays affect the active count which drops.
- During heavy rain, there are no users whereas partly cloudy/clear sky saw the maximum count.

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

Ans : Removes the first column since its an extra column. We can use n-1 columns to derive values of n columns. It also helps in reducing collinearity between dummy variables.

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

Ans: atemp has it; I also believe temp would have the same (but I deleted it) since it was a duplicate variable.

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

Ans. One of the fundamental assumptions of a linear regression model is that the error terms should correspond to a normal curve, when plotted on histogram.

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

Ans. Yr. and aTemp. positively affect the model while windspeed -vely affects the same.

General Subjective Questions

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

Ans. An interpolation technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s), is linear regression.
Ideally the below 4 steps are used to do a liner regression analysis and create a model.

1. Reading and Understanding the Data.
2. Training the model.
3. Residual Analysis.
4. Predicting evaluating the model on the test set.

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

Ans. Anscombe's quartet is a famous example in statistics used to highlight the importance of data visualization. It consists of four sets of data, each containing 11 data points (x, y pairs).

Here's the key point: all four sets have nearly identical summary statistics like mean, variance, correlation coefficient, and even linear regression lines.

However, when you plot these data sets, they reveal very different underlying relationships between the x and y variables.

Importance of visualization: Summary statistics alone can be misleading. Visualizing the data with scatter plots allows you to identify trends, outliers, and non-linear relationships that wouldn't be evident from just numbers.
Impact of outliers: A single outlier can significantly affect summary statistics like correlation, but its influence might not be obvious without visualization.
Anscombe's quartet is a reminder that before diving into statistical analysis, it's crucial to explore your data graphically. This can help you choose the most appropriate analysis methods and avoid misinterpretations.

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

Ans. Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure used to assess the strength and direction of a linear relationship between two continuous variables. It's denoted by the symbol "r".

Strength: The value of r ranges from -1 to +1.

A value close to +1 indicates a strong positive correlation. As one variable increases, the other tends to increase as well. (Example: Height and weight)
A value close to -1 indicates a strong negative correlation. As one variable increases, the other tends to decrease. (Example: Study time and exam scores - assuming more study leads to better scores)
A value close to 0 indicates a weak or no correlation. There's no clear linear relationship between the variables.

Direction: The positive or negative sign of r indicates the direction of the correlation.

It only measures linear relationships. Non-linear relationships won't be captured by Pearson's R. (Imagine a scatter plot that curves instead of a straight line)

It assumes both variables are continuous. It's not suitable for analysing categorical data.

The interpretation of a "strong" correlation can vary depending on the field of study.

It helps researchers understand how changes in one variable might be associated with changes in another.

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

Ans.

Scaling in statistics refers to the process of transforming your data values to a common scale. This is often done to improve the performance of machine learning algorithms or statistical models. There are several reasons why scaling is important:

- **Feature comparability:** Different features (variables) in your data may be measured on different scales. For example, income might be in thousands of dollars while age is in years. Scaling helps put all features on a similar scale, allowing the model to treat them more equally.

- **Gradient descent optimization:** Many machine learning algorithms rely on optimization techniques like gradient descent. Scaling can improve the convergence of these algorithms, making them learn faster and potentially reach better solutions.

- **Normalization vs. Standardization:** There are two main types of scaling commonly used:
    - **Normalization:** This technique scales the data to a specific range, typically between 0 and 1 (or -1 and 1). There are different normalization methods, but a common one is Min-Max scaling, which subtracts the minimum value from each data point and then divides by the difference between the maximum and minimum values.
    - **Standardization:** This technique transforms the data to have a zero mean and a unit standard deviation. This is achieved by subtracting the mean value from each data point and then dividing by the standard deviation.

Here's a table summarizing the key differences:

Feature	Normalization	Standardization
Target range	User-specified range (e.g., 0-1)	Mean = 0, Standard deviation = 1
Outlier impact	Can be sensitive to outliers	Less sensitive to outliers
Interpretation	Scaled values don't directly reflect original unit	Scaled values represent units of standard deviation from the mean

Choosing between normalization and standardization depends on specific needs and the algorithm:

- **Normalization:** Use this if the absolute values of your features are important, or if your algorithm is sensitive to the range of data values.
- **Standardization:** Use this if the distribution of your data is important, or if your algorithm relies on assumptions about normality (like linear regression).

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

Ans.

An infinite VIF (Variance Inflation Factor) in regression analysis indicates a scenario where a predictor variable (independent variable) is perfectly collinear with a linear combination of other predictor variables in the model.

Here's a breakdown of what it means:

- **Multicollinearity:** VIF measures the inflation of a variable's variance due to multicollinearity. Multicollinearity occurs when there's a high degree of correlation between two or more independent variables.
- **Perfect Collinearity:** An infinite VIF signifies a more extreme case - **perfect collinearity**. This means one predictor variable can be entirely expressed as a linear combination of the other predictor variables.

In simpler terms, the information contained in the variable with the infinite VIF is completely redundant because it can be perfectly predicted by the other variables. This redundancy creates several problems for your regression analysis:

- **Unreliable Coefficients:** The regression coefficients associated with highly collinear variables become unstable and unreliable. Even small changes in the data can lead to significant changes in these coefficients.
- **Difficult Interpretation:** It becomes challenging to interpret the individual effect of a variable with an infinite VIF on the target variable because its influence is entangled with the collinear variables.
- **Inaccurate Model:** The overall model's accuracy and reliability can be compromised due to the inflated variances and unstable coefficients.

Here's what to do if you encounter an infinite VIF:

- **Identify the Culprit:** Examine the correlation matrix or other diagnostics to pinpoint the variables with high correlations (often exceeding 0.8 or 0.9). The variable with the infinite VIF is likely part of this group.
- **Remove or Combine:** There are several approaches to address multicollinearity, and the best course of action depends on your specific situation.Here are some options:
    - **Remove** the variable with the infinite VIF if it's not theoretically important to your model.
    - **Combine** highly correlated variables into a single new variable if they represent a similar underlying concept.
    - **Regularization techniques** like ridge regression or LASSO regression can be applied to handle multicollinearity while keeping all variables. However, these techniques introduce additional complexity and require careful tuning.

By addressing multicollinearity and removing variables with infinite VIF, you can improve the stability, interpretability, and overall accuracy of your regression model.

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

Ans.

A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool used in linear regression to assess the normality of the residuals (errors) in your model. Here's why it's important:
Normality Assumption: Linear regression often relies on the assumption that the errors (the difference between actual and predicted values) are normally distributed. This assumption is crucial for the validity of certain statistical tests and the interpretation of the model's coefficients.
Q-Q Plot to the Rescue: The Q-Q plot helps you visually check if the normality assumption holds. It plots the quantiles of your standardized residuals against the quantiles of a standard normal distribution (the bell curve).
Interpreting the Q-Q Plot:

- **Straight Line:** If the points in the Q-Q plot fall roughly along a straight diagonal line, it suggests that the residuals are likely normally distributed. This is a good sign for your model.
- **Deviations from the Line:** If the points deviate significantly from the line, especially in the tails (far left or right), it indicates that the residuals might not be normally distributed. This could be due to factors like outliers, skewness, or kurtosis.

Importance in Linear Regression:

- **Reliable P-Values:** When the normality assumption is met, the p-values associated with the regression coefficients become more reliable. P-values help you assess the significance of each variable in the model.
- **Confidence Intervals:** Normality also allows for the construction of valid confidence intervals for the regression coefficients. These intervals provide an estimate of the range within which the true coefficient values likely lie.
- **Model Diagnostics:** Even if the normality assumption isn't perfectly met, the Q-Q plot can still be a valuable diagnostic tool. It can help you identify potential problems with your data or model that need further investigation.

Addressing Non-Normality:

If the Q-Q plot reveals non-normality, there are several approaches you can take:

- **Transform the Data:** You can try transforming your target variable (dependent variable) using techniques like log transformation or square root transformation to achieve normality.
- **Robust Regression Methods:** Consider using robust regression methods that are less sensitive to violations of the normality assumption.
- **Large Enough Sample Size:** Sometimes, with a sufficiently large sample size, the normality assumption becomes less critical, and the model might still be reliable.

By using Q-Q plots and addressing potential non-normality issues, you can ensure that your linear regression model is built on solid statistical ground and produces more interpretable and reliable results.