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)

   Many of the categorical variables fall among the most important variables for the final model.For example Yr,holiday,workingday these 3 are the top important variables of our model and hold immense importance. We needed to convert the categoprical variables to their dummy though so that some categories do not get unfairly advantaged because of a higher numerical value by linear regression.

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

   This is essential since otherwise we create a redundant variable which could have already been represented by the other categories being 0. So to avoid this redundancy and multicollinearity issue we use this

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

   This will be either temp and atemp. Even their correlation with cnt from correlation matrix is same.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   - There is linear relationship between predictors and predicted. Picked the numerical variables and plotted their scatterplot with target variable. Straight line in plot indicated good linear relationship of predictor and target.
   - No multicollinearity among the independent variables – All variables VIF score below 5
   - Errors are normally distributed with mean 0. – Plotted histogram of errors and found a normal distribution curve with mean around 0.
   - Errors exhibit constant variance (homoscedacity property) – Plotted scatterplot of important numerical predictors like atemp and humidity with residual values and found no definite pattern existing.Thus errors are random and show constant variance and thus homoscedacity.

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

   -yr
   -weathersit_3
   -month_10

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for predicting a continuous target variable based on one or more input features. It aims to find the best-fitting straight line (or hyperplane in higher dimensions) that minimizes the difference between the actual target values and the predicted values. This is achieved by adjusting the coefficients (weights) associated with each input feature.

The linear regression model is represented as:

y = b0 + b1x1 + b2x2 + ... + bn*xn

where:

y is the predicted value (target variable).

b0 is the intercept term (the value of y when all input features are zero).

b1, b2, ..., bn are the coefficients (weights) corresponding to the input features x1, x2, ..., xn.

The goal is to find the optimal values for the coefficients that minimize the Mean Squared Error (MSE) between the actual and predicted target values. This process is typically performed using optimization techniques like Gradient Descent or Least Squares.

Once the model is trained, it can make predictions on new data by plugging the input features into the learned linear equation. Linear regression is widely used due to its simplicity and interpretability, especially when the relationship between the features and the target variable appears to be linear. However, it may not be suitable for non-linear relationships, for which other regression algorithms or non-linear models should be considered.

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

Anscombe's quartet is a set of four datasets in statistics that have nearly identical simple descriptive statistics but exhibit significantly different relationships when visualized. It was introduced by the statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization and the limitations of relying solely on summary statistics.

The four datasets in Anscombe's quartet are as follows:

Dataset I:

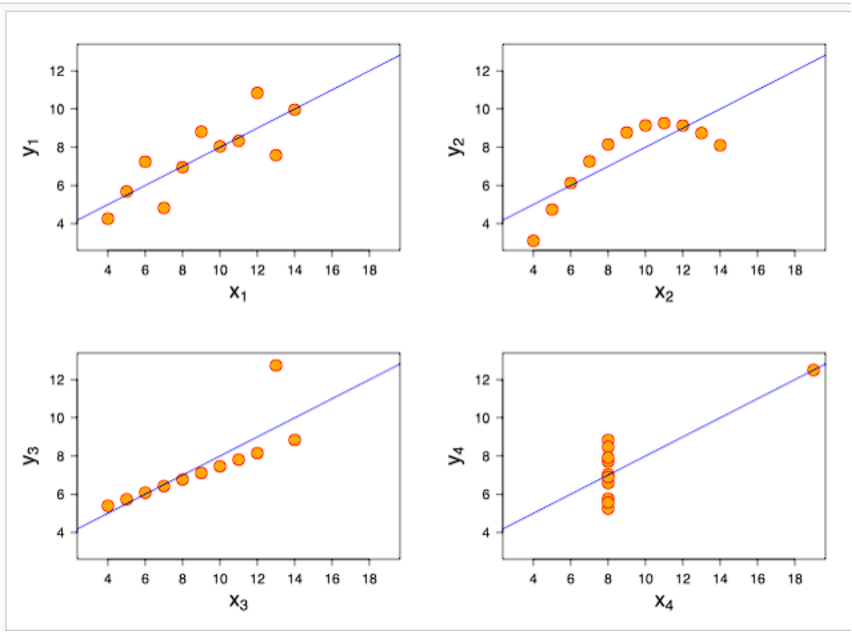
x: 10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0
y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
Dataset II:

x: 10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0

y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
Dataset III:

x: 10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0
y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
Dataset IV:

x: 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 19.0, 8.0, 8.0, 8.0
y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
Each dataset contains 11 (x, y) data points. The interesting thing about these datasets is that despite having similar means, variances, correlation coefficients, and regression lines, they look entirely different when plotted.



When visualized, Dataset I shows a clear linear relationship between x and y, Dataset II shows a linear relationship with an outlier, Dataset III shows a nonlinear relationship, and Dataset IV has a completely different pattern.

Anscombe's quartet highlights the importance of visualizing data and not relying solely on summary statistics. It serves as a reminder that datasets with similar summary statistics can exhibit vastly different underlying relationships, which can lead to different interpretations and conclusions. Data visualization allows us to gain deeper insights and uncover patterns that may not be apparent from summary statistics alone.

3. What is Pearson’s R? (3 marks)

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

The Pearson correlation coefficient is denoted by the symbol "r" and takes values between -1 and +1:

If "r" is close to +1, it indicates a strong positive linear relationship, meaning that as one variable increases, the other tends to increase proportionally.
If "r" is close to -1, it indicates a strong negative linear relationship, meaning that as one variable increases, the other tends to decrease proportionally.
If "r" is close to 0, it indicates a weak or no linear relationship between the two variables.
The formula to calculate Pearson's correlation coefficient between two variables X and Y is as follows:

r = (Σ((X_i - X_mean) * (Y_i - Y_mean))) / (sqrt(Σ(X_i - X_mean)^2) * sqrt(Σ(Y_i - Y_mean)^2))

where:

X_i and Y_i are individual data points for variables X and Y, respectively.
X_mean and Y_mean are the mean values of variables X and Y, respectively.
Σ represents the summation (sum over all data points).

Pearson's correlation coefficient has several important properties:

It only measures linear relationships. Non-linear relationships may not be adequately captured by "r."
It is sensitive to outliers, meaning that extreme values can heavily influence the value of "r."
It is symmetric, meaning that the correlation coefficient between X and Y is the same as the correlation coefficient between Y and X.
Pearson's R is widely used in various fields, including statistics, data analysis, machine learning, and social sciences, to understand the relationship between two continuous variables and to make inferences about the strength and direction of that relationship.

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

Scaling, in the context of data preprocessing in machine learning, refers to the process of transforming numerical features to a specific range or distribution. It involves changing the scale of the features so that they all have a similar magnitude. Scaling is often applied to ensure that no single feature dominates the learning process or negatively impacts the performance of machine learning algorithms.

Scaling is performed for several reasons:

- Equalizing feature magnitudes: Many machine learning algorithms use distance-based calculations (e.g., Euclidean distance, K-nearest neighbors). If features have different magnitudes, those with larger values can overshadow others, leading to biased model training.
- Faster convergence: Scaling helps algorithms converge faster during optimization processes like gradient descent by providing a more balanced landscape in the feature space.
- Regularization: Some regularization techniques (e.g., L1 or L2 regularization) are sensitive to feature magnitudes, and scaling can ensure a fair impact on all features.
- Certain algorithms require scaling: Some algorithms, like support vector machines

(SVM), are sensitive to feature scaling and may not perform well without it.

Normalized Scaling (Min-Max Scaling):

Formula: X_normalized = (X - X_min) / (X_max - X_min)
The range of the transformed data is mapped to a specified range, usually between 0 and 1.
It preserves the shape of the original distribution but constrains the feature values within the specified range.
Normalization is sensitive to outliers since it uses the minimum and maximum values.
Standardized Scaling (Z-score Scaling or Standardization):

Formula: X_standardized = (X - X_mean) / X_std
It transforms the data to have a mean of 0 and a standard deviation of 1.
Standardization centers the data around the mean, making the distribution symmetric.
Standardization is robust to outliers because it uses the mean and standard deviation, which are not affected by extreme values.

The choice between normalized scaling and standardized scaling depends on the specific requirements of the machine learning algorithm and the characteristics of the data. Normalization is suitable when you want to confine the data to a specific range, while standardization is preferable when you want to center the data and have a more Gaussian-like distribution.

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

The VIF (Variance Inflation Factor) is a metric used to detect multicollinearity in a multiple linear regression model. It quantifies how much the variance of the estimated regression coefficients increases due to multicollinearity, which occurs when two or more predictor variables are highly correlated with each other.

The VIF for a particular predictor variable is calculated as follows:

VIF = 1 / (1 - R^2)

where R^2 is the coefficient of determination obtained by regressing the predictor variable against all other predictor variables in the model.

When the VIF is infinite, it means that there is perfect multicollinearity between the predictor variable and other variables in the model. Perfect multicollinearity occurs when one predictor variable can be exactly predicted by a linear combination of other predictor variables.

This situation arises when, for example:

One predictor variable is a constant multiple of another predictor variable, such as:

X1 = 2 * X2
In this case, the VIF for X1 would be infinite because X1 can be perfectly predicted using X2.
Two or more predictor variables are linearly dependent, leading to a perfect correlation:

X1 = 3 * X2 + 5 * X3

In this case, the VIF for X1 would be infinite because X1 can be expressed as a linear combination of X2 and X3.

In these case as R squared would near 1 due to perfect explanability so denominator in above formula will tend to 0 and thus the VIF value will tend to infinity.

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 dataset follows a specific theoretical distribution. It is particularly useful for examining whether the data approximates a normal (Gaussian) distribution. The Q-Q plot compares the quantiles of the dataset against the quantiles of the theoretical distribution.

Here's how a Q-Q plot is constructed:

The dataset is sorted in ascending order.
The quantiles of the dataset are calculated.
The quantiles are plotted against the quantiles expected from the theoretical distribution.
If the data closely follows the theoretical distribution, the points in the Q-Q plot will fall approximately along a straight line. Deviations from the straight line indicate departures from the theoretical distribution.

Use and Importance of a Q-Q Plot in Linear Regression:

Checking Normality Assumption: One of the key assumptions of linear regression is that the residuals (the differences between the actual and predicted values) are normally distributed. Q-Q plots are useful for examining the normality of the residuals. If the residuals follow a straight line on the Q-Q plot, it suggests that the normality assumption is likely met. If the points deviate from the line, it indicates non-normality, which may require further investigation or data transformations.

Identifying Skewness and Outliers: Q-Q plots can help identify skewness in the data. If the data points deviate from the straight line in the tails of the plot, it indicates the presence of skewness. Additionally, outliers may be detected if the data points significantly deviate from the line in the plot.

Comparing Distributions: Q-Q plots can be used to compare the distribution of two datasets. By plotting their quantiles against each other, it becomes easier to determine if they follow the same distribution or if there are differences.

Model Validation: Q-Q plots are also helpful for validating the assumptions of other statistical models, not just linear regression. They can be used to assess the goodness of fit of any model by comparing its residuals against a theoretical distribution.

In summary, Q-Q plots are a valuable tool for examining the distributional characteristics of a dataset and assessing whether it approximates a theoretical distribution, such as the normal distribution. They are particularly important in linear regression to validate assumptions and ensure the reliability of the model's results.