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 categorical variables are season,weathersit,yr,mnth,holiday,weekday,workingday
- Among these variables, while month might seem like a variable which can affect the cnt, it had a high VIF and P-Value, so it was not a good column for prediction
- Season and weathersit had an impact over cnt, but since the values were 1,2,3,4. It might have lead to a confusion that there is an hierarchy, so we pivoted these columns based on values. And looking on the box plot, we can say that summer and fall had higher counts
- Weather did not have as strong an impact as season, but there was some correlation b\w weather and cnt.And looking at the boxplot among the weather, it was clear weather which had the highest counts
- Yr also had a strong correlation with count, with 2019 having higher counts than 2018
- With holiday, weekday and working day, it was the holiday(working day =0 and weekday=0 and 6) which had greater counts

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

- Generally when we create dummy variables, we pivot the column based on values
- If a column has 3 values, we create 3 columns based on the values, and populate true/false
- Now, if the two of the values are false, it would mean that 3rd is true, which is why we don't need 3 columns, 2 coulmns suffice
- Hence for N values, we create N-1 columns, and specifying drop_first=true implements the same

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

with the target variable? (1 mark)

- Among the numerical variables, it's the Registered variable which has the highest correlation

4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

- After making predictions on the training set, we can validate the results by comparing the predicted value and the actual y_train value
- We can do residual analysis on the predicted values
- We checked the distribution of the residuals, it was normally distributed with mean as 0
- We also plotted a qq plot which also suggested normality
- We also did durbin Watson test, which gave a value of around 2

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes?

- Based on final model, yr,temp and season seems like the top 3 features contributing significantly towards explaining the demand of the shared bikes

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, meaning that the relationship can be represented by a straight line.

The basic equation for linear regression is:

Y = β0 + β1*X1 + β2*X2 + ... + βn*Xn + ε

Where:

- **Y:** is the dependent variable (what we want to predict)
- **β0:** is the intercept (the value of Y when all independent variables are 0)
- **β1, β2, ..., βn:** are the coefficients (weights) assigned to each independent variable (X1, X2, ..., Xn)
- **X1, X2, ..., Xn:** are the independent variables (predictors)
- **ε:** is the error term (residuals) representing the difference between the actual value of Y and the predicted value

The goal of linear regression is to find the values of the coefficients (β0, β1, ..., βn) that minimize the sum of squared errors (SSE). This is achieved using a technique called **ordinary least squares (OLS)**.

Steps Involved in Linear Regression:

1. **Data Preparation:**
   - Collect and clean the data, ensuring there are no missing values or outliers.
   - Scale numerical features if necessary to improve model performance.
2. **Model Specification:**
   - Define the dependent variable (Y) and independent variables (X1, X2, ..., Xn).
   - Choose the appropriate linear regression model (simple or multiple).
3. **Model Fitting:**
   - Use an optimization algorithm (like OLS) to estimate the coefficients (β0, β1, ..., βn) that minimize the SSE.

4. **Model Evaluation:**
    - Assess the model's performance using metrics like R-squared, adjusted R-squared, mean squared error (MSE), and root mean squared error (RMSE).
    - Check for model assumptions (linearity, homoscedasticity, normality, and independence of residuals).
5. **Interpretation:**
    - Interpret the coefficients to understand the relationship between the independent variables and the dependent variable.
    - Use the model to make predictions for new data.

Types of Linear Regression:

- **Simple Linear Regression:** Involves only one independent variable.
- **Multiple Linear Regression:** Involves multiple independent variables.

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a famous set of four datasets, each with the same mean, variance, correlation, and linear regression line. Despite these identical statistical properties, the datasets have vastly different visual representations and underlying patterns.

This quartet demonstrates the importance of **visualizing data** before drawing conclusions solely based on statistical summaries. It highlights the limitations of relying solely on numerical measures and emphasizes the value of exploring data graphically.

The four datasets are:

1. **Dataset 1:** A typical linear relationship with a positive correlation.
2. **Dataset 2:** A quadratic relationship with a positive correlation, but the points are scattered around the line.
3. **Dataset 3:** A linear relationship with a positive correlation, but with one outlier that significantly affects the regression line.
4. **Dataset 4:** A horizontal line with no correlation between the variables, but the mean and variance are the same as the other datasets.

Key takeaways from Anscombe's quartet:

- **Visual exploration is essential:** Always visualize your data to identify patterns, outliers, and non-linear relationships that might not be apparent from numerical summaries.
- **Correlation does not imply causation:** Correlation is a measure of association between variables, but it doesn't necessarily mean that one variable causes changes in the other.
- **Outliers can have a significant impact:** Even a single outlier can dramatically affect the results of a statistical analysis.

- **Statistical summaries alone are insufficient:** While statistical summaries are useful, they should be complemented with visual analysis to get a complete picture of the data.

3. What is Pearson’s R?

Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1:

- **-1:** Perfect negative correlation: As one variable increases, the other decreases perfectly.
- **0:** No correlation: There is no linear relationship between the variables.
- **1:** Perfect positive correlation: As one variable increases, the other increases perfectly.

Formula:

r = (nΣxy - ΣxΣy) / sqrt((nΣx^2 - (Σx)^2)(nΣy^2 - (Σy)^2))

Where:

- n is the number of data points
- x and y are the individual data points for the two variables
- Σ denotes the summation

Interpreting Pearson's r:

- **Strength:** The absolute value of r indicates the strength of the linear relationship. A value closer to 1 or -1 indicates a stronger relationship.
- **Direction:** The sign of r indicates the direction of the relationship:
    - Positive (r > 0): As one variable increases, the other increases.
    - Negative (r < 0): As one variable increases, the other decreases.

Assumptions:

- **Linearity:** The relationship between the two variables is linear.
- **Normality:** Both variables are normally distributed.
- **Homoscedasticity:** The variance of the residuals (errors) is constant

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

and standardized scaling?

Scaling is a technique used in data preprocessing to transform numerical features to a common scale. This is often necessary for machine learning algorithms, as they may perform better when features are on a similar scale.

Why is scaling performed?

- **Improved algorithm performance:** Many algorithms rely on distance calculations. Scaling ensures that features contribute equally to these calculations, preventing features with larger magnitudes from dominating the results.
- **Enhanced interpretability:** Scaled features can make it easier to interpret the coefficients of a model. For example, if features are on a similar scale, a coefficient of 2 for one feature might be considered more important than a coefficient of 0.1 for another feature.

Normalized Scaling vs. Standardized Scaling

1. **Normalized Scaling (Min-Max Scaling):**
    - Rescales features to a specific range, typically between 0 and 1.
    - Formula: X_scaled = (X - X_min) / (X_max - X_min)
    - Preserves relative differences between data points.
    - Suitable when you want to maintain the original distribution of the data.
2. **Standardized Scaling (Z-score Standardization):**
    - Centers the data around the mean and scales it to have a standard deviation of 1.
    - Formula: X_scaled = (X - mean(X)) / std(X)
    - Removes the influence of outliers and ensures that features have a mean of 0 and a standard deviation of 1.
    - Suitable when you want to remove the effects of scale and m

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

VIF (Variance Inflation Factor) becoming infinite typically indicates a severe case of multicollinearity. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other.

Reasons for Infinite VIF:

1. **Perfect Multicollinearity:** If two or more independent variables are perfectly correlated (e.g., one variable is a linear combination of the others), the VIF will be infinite. This means that one variable can be perfectly predicted from the others, making it redundant in the model.
2. **Numerical Instability:** In some cases, due to numerical issues or the nature of the data, the VIF calculation might become unstable, leading to an infinite value.

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

Q-Q plot (Quantile-Quantile plot) is a graphical technique used to compare the distribution of a sample data set against a theoretical distribution, typically a normal distribution. It's a valuable tool in statistical analysis, especially when assessing the normality assumption in linear regression models.

How a Q-Q plot works:

1. **Quantiles:** Both the sample data and the theoretical distribution are sorted in ascending order.
2. **Matching Quantiles:** Corresponding quantiles from the two distributions are plotted against each other.
3. **Visual Interpretation:** If the points in the Q-Q plot fall close to a straight line, it suggests that the sample data follows the theoretical distribution. Deviations from the line indicate departures from normality.

Use of Q-Q plots in linear regression:

In linear regression, the assumption of normality is crucial for the validity of statistical tests and the interpretation of model results. A Q-Q plot can help assess this assumption by comparing the residuals (the differences between the actual and predicted values) to a normal distribution.

Importance of Q-Q plots in linear regression:

- **Detecting non-normality:** Q-Q plots can reveal patterns of non-normality, such as skewness or heavy-tailed distributions.
- **Identifying outliers:** Outliers may appear as points that deviate significantly from the diagonal line in the Q-Q plot.
- **Assessing model fit:** If the residuals follow a normal distribution, it suggests that the linear regression model is a good fit for the data.
- **Guiding transformations:** If the Q-Q plot shows a clear deviation from normality, transforming the data (e.g., log transformation) might help improve the fit.