Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

Categorical variables and dependent variables taken to consideration while calculating the infer.
Categorical Variables:
season: This represents the season of the year, with values:

1: Spring
2: Summer
3: Fall
4: Winter
yr: This indicates the year:

0: 2018
1: 2019
mnth: This indicates the month of the year, with values:

1 to 12, corresponding to each month (January to December).
holiday: This binary variable indicates whether it is a holiday or not:

0: Not a holiday
1: Holiday
weekday: This represents the day of the week, with values:

0: Sunday, 1: Monday, ..., 6: Saturday.
workingday: This binary variable indicates whether the day is a working day (neither a weekend nor a holiday):

0: Not a working day (weekend or holiday)
1: Working day
weathersit: This represents the weather situation on that day, with values:

1: Clear, Few clouds, partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Dependent Variable:
cnt: This is the total number of bike rentals, which is the sum of both casual and registered users.

Inference from Categorical Variables on the Dependent Variable:
Each of these categorical variables potentially plays a significant role in influencing bike demand.
1. The demand for bikes is likely to vary with the seasons. For example, warmer months like summer (season 2) may see higher bike demand.

2. Bike demand may have increased over time, suggesting that as bike-sharing systems gain popularity, demand tends to rise each year. Year 2019 (yr = 1) might show higher demand compared to 2018 (yr = 0)

3. weekday variable, working days (working day = 1) might see more bike rentals compared to non-working days.
4. weekdays may see a steady demand, while weekends may experience higher leisure-based demand.
5. On days with clear weather (weathersit = 1), bike usage may be higher, while adverse weather conditions such as mist or heavy rain (weathersit = 2 or 4) will likely reduce demand.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

1. When creating dummy variables from categorical features, the argument drop_first=True is used to avoid the "dummy variable trap," which occurs when one of the dummy variables is highly correlated with the others.
2. Dropping one category also makes the model more interpretable. With drop_first=True, the model coefficients for the remaining dummy variables.
3. Using drop_first=True during dummy variable creation is crucial to avoid multicollinearity by ensuring that the model does not suffer from redundant predictors.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Given the nature of the dataset, **registered** users are likely to have the highest correlation with the target variable **cnt**. This is because the **cnt** variable includes both casual and **registered** users, and **registered** users generally contribute a large portion of the total rentals. Therefore, the correlation between **cnt** and **registered** is expected to be very strong.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

To validate the assumptions of Linear Regression after building the model on the training set, several key assumptions of Linear Regression need to be checked. These assumptions ensure that the model is a good fit for the data and that the results from the model can be interpreted confidently.

Steps:

1. **Linearity:** Check the residual plot for random distribution.
   Assumption: There is a linear relationship between the independent variables (predictors) and the dependent variable (target)
2. **Homoscedasticity:** Check the residuals vs. fitted values plot.
   Assumption: The variance of the residuals is constant across all levels of the independent variables (predictors).
3. **Normality of residuals**: Check the Q-Q plot and histogram of residuals.
   Assumption: The residuals (errors) should be normally distributed.
4. **Independence**: Use the Durbin-Watson test to check for autocorrelation.
   Assumption: The errors are independent of each other, meaning there's no autocorrelation (correlation between residuals).
5. **No Multicollinearity**: Calculate the Variance Inflation Factor (VIF) for each independent variable.
   Assumption: The independent variables are not highly correlated with each other.

By performing these diagnostic tests and validating the assumptions, you ensure the linear regression model is reliable and the results are trustworthy.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The following features might contribute significantly towards the demand for shared bikes.

1. **Temperature (temp):** This variable likely has a strong influence on bike demand, as people are more likely to rent bikes during warmer weather. A positive coefficient would suggest that as temperature increases, the demand for bikes increases.
2. **Season (season):** Different seasons (spring, summer, fall, winter) could affect the demand due to varying weather conditions and seasonal activities.

3. **Month (mnth):** The time of year could also be significant, with higher bike demand expected in certain months due to better weather conditions. For example, spring and summer months may see higher demand than winter months.
4. **Holiday (holiday):** Whether the day is a holiday or not could influence bike demand. People might rent more bikes on holidays when they have more leisure time.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Overview:

1. Linear regression is one of the simplest and most widely used algorithms for predictive modeling. It assumes a linear relationship between the independent variables (predictors) and the dependent variable (target).

2. Linear regression is a foundational technique in machine learning and statistics. It's used to model relationships between a dependent variable and one or more independent variables.

3. The objective of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the error between predicted and actual values.

$y=\beta0+\beta1\cdot X1+\beta2\cdot X2+\cdots+\beta n\cdot Xn+\epsilon$

ϵ represents the error term

Goal of Linear Regression:

The goal is to estimate the coefficients that minimize the difference between the predicted and actual values of y. This difference is quantified using a loss function, which measures how well the model fits the data. The most common loss function used in linear regression is Mean Squared Error

Types of Linear Regression

1) Simple Linear Regression:

This involves only one independent variable (feature) and is represented by a straight line.

$y=\beta0+\beta1\cdot X1+\epsilon$

The model tries to find the best fit line that minimizes the error between the observed values and the line.

2) **Multiple Linear Regression:**

This involves two or more independent variables and is represented by a hyperplane. The equation is:

$y=\beta0+\beta1\cdot X1+\beta2\cdot X2+\cdots+\beta n\cdot Xn+\epsilon$

3) **Ridge Regression (L2 Regularization):**

A variant of linear regression that adds a penalty to the coefficients to prevent overfitting. It adds a regularization term to the loss function:

$MSE + \lambda i = 1\sum n\ \beta i2$

4) **Lasso Regression (L1 Regularization):**

Lasso can also help in feature selection by forcing some coefficients to become exactly zero.

Assumptions of Linear Regression

For linear regression to produce valid results, several assumptions need to be satisfied:

Linearity: There should be a linear relationship between the independent and dependent variables.

Independence of Errors: The residuals (errors) should not be correlated.

Homoscedasticity: The variance of the errors should be constant across all levels of the independent variables.

Normality of Errors: The residuals should be approximately normally distributed.

No Multicollinearity: The independent variables should not be highly correlated with each other.

Advantages of Linear Regression

Simplicity: It's easy to implement and interpret.

Fast to Train: Linear regression has a low computational cost, making it suitable for large datasets.

Transparency: The model is interpretable, as you can directly observe the effect of each feature on the target variable.

Disadvantages of Linear Regression

Assumes Linearity: It only works well if the relationship between the variables is linear. If this assumption is violated, the model will perform poorly.

Sensitive to Outliers: Linear regression is highly sensitive to outliers, which can disproportionately affect the model.

Multicollinearity: If there is high correlation between independent variables, the model may become unstable and produce misleading results.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a set of four datasets that have nearly identical statistical properties, yet they differ significantly when visualized.

Key Statistical Properties

Each dataset in the quartet has:

Mean of x: Approximately 9.
Mean of y: Approximately 7.5.
Variance of x: Approximately 11.
Variance of y: Approximately 4.12.
Correlation (r): Approximately 0.816.
Regression line: **y=3+0.5x**

Datasets Overview

Dataset 1(A Linear Relationship)

1. Represents a typical linear relationship.
2. The data points are scattered around a straight line with minor noise.

Dataset 2(A Non-linear Relationship)

1. All data points lie on a curve, but the correlation and regression line give the misleading impression of a linear relationship.
2. This demonstrates how non-linear relationships can lead to misinterpretation when relying only on linear regression.

Dataset 3(Presence of Outliers)

1. Most data points are tightly clustered, except for one significant outlier.
2. The outlier heavily influences the regression and correlation values, highlighting the impact of outliers on statistical summaries.

Dataset 4(Extreme Leverage Point)

1. Almost all points have the same x-value except for one.
2. The single differing point gives a false correlation, emphasizing that a single unusual point can distort statistical results.

Advantages:

1. The advantages of Anscombe's quartet lie in its ability to underscore the importance of data visualization and reveal the pitfalls of relying solely on statistical summaries.

2. It remains a powerful educational tool for demonstrating the importance of comprehensive data analysis.

Limitation:

1. its limitations include a narrow focus on specific issues and the artificial nature of its datasets.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R (Pearson Correlation Coefficient) r is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is one of the most widely used correlation coefficients in statistics.

Formula:

the formula can be rewritten using covariance and standard deviations:

r=Cov(x,y) / σx σy

Where:

- Cov(x,y) : Covariance between x and y
- σx : Standard deviation of x
- σy : Standard deviation of y

Key Characteristics:

Range:

r ranges from −1 to +1

r = +1: Perfect positive linear correlation.

R = −1: Perfect negative linear correlation.

R = 0: No linear correlation.

Sign:

The sign of r indicates the direction of the relationship.

+: Positive correlation

−: Negative correlation

Example 1: Perfect Positive Correlation

X = [1,2,3,4,5]

Y = [2,4,6,8,10]

R = +1: Both variables increase proportionally.

Example 2: Perfect Negative Correlation

X = [1,2,3,4,5]

Y = [10,8,6,4,2]

R = −1: One variable increase as the other decreases proportionally.

Example 3: No Correlation

X = [1,2,3,4,5]

Y = [7,7,7,7,7]

r=0: No relationship between xxx and yyy

Advantages

1. Easy to compute and interpret.
2. Useful in a variety of domains, including economics, biology, and social sciences.
3. Independent of the scale of measurement due to normalization.

Disadvantages

1. Captures only linear correlations and may miss non-linear relationships.

Area Of Application:

Finance and Healthcare.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a data preprocessing technique used to adjust the range of features in a dataset. It transforms the data to ensure that different features contribute equally to the model's learning process.

Scaling Performed

1. **Improves Model Performance:**

Algorithms like gradient descent perform better when features are scaled because they converge faster.

2. **Avoids Bias Towards Larger Magnitudes:**

Models that use distance-based metrics (e.g., K-Nearest Neighbors, Support Vector Machines) or gradient-based learning are sensitive to the scale of features.

3. **Prevents Numerical Instability**:

Features with large values may cause computational difficulties or dominate during optimization.

4. **Improves Interpretability:**

Scaled data is easier to interpret and visualize in some cases.

5. **Maintains Uniform Contribution:**

Ensures that features with larger magnitudes don’t disproportionately influence the model's predictions.

Difference Between Normalized Scaling and Standardized Scaling:

Definition:

Normalized scaling adjusts data to fit within a specific range, typically between 0 and 1.
Standardized scaling transforms data to have a mean of 0 and a standard deviation of 1.

Output Range:

Normalized scaling results in data values within a fixed range, such as [0,1] [0, 1] [0,1] or [−1,1] [-1, 1] [−1,1].

Standardized scaling does not constrain data to a specific range but ensures the data is centered around 0 with unit variance.

Effect on Distribution:
Normalized scaling preserves the original shape of the data distribution.
Standardized scaling alters the data to follow a standard normal distribution**.**

Use case:
Normalized scaling is best when features have known bounds or need proportional comparisons.
Standardized scaling is used when data needs to be centered and scaled for algorithms that assume normally distributed inputs.

Sensitivity to Outliers:
Normalized scaling is highly sensitive to outliers.
Standardized scaling is less sensitive to outliers but can still be influenced due to its reliance on the mean and variance.

Algo:
Normalized scaling is effective for algorithms that depend on distance calculations, such as K-Nearest Neighbours (KNN) and Neural Networks.
Standardized scaling is more appropriate for gradient-based algorithms like Logistic Regression, Linear Regression.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity between one independent variable and one or more other independent variables in a regression model.

Formula for VIF:
For a given independent variable Xi, the VIF is calculated as:
VIF(Xi) = 1/(1-R^2)
where R^2 is the coefficient of determination for the regression of Xi on all other independent variables in the model.

Reason for Infinite VIF:

1. **Perfect Multicollinearity:**
   If Xi is a perfect linear combination of other independent variables (e.g Xi = c1X1+ c2X2+…ckXk) the regression of Xi on the other variables will have an R^2 value of 1.
   Substituting R2=1 R^2 = 1 R2=1 into the VIF formula results in division by zero
   VIF(Xi) = 1/(1-1) = Infinity
2. **Dependency in the Design Matrix:**

In matrix terms, perfect multicollinearity means that the design matrix (matrix of independent variables) is not full rank. This causes issues in calculating the regression coefficients and, by extension, the VIF.

Implications:

No Unique Solution:

Perfect multicollinearity means the regression coefficients cannot be uniquely determined.

Corrective Measures: In order to address this.

1. Drop one of the collinear variables
2. Combine collinear variables into a single composite variable.
3. Use regularization techniques (like Ridge or Lasso regression) to handle multicollinearity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

A Q-Q plot (quantile-quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, most commonly a normal distribution. It helps visualize whether a dataset follows a specified distribution or not.

Key Components of a Q-Q Plot:

Quantiles of the Sample Data:

These are the sorted values of the dataset, scaled to align with the quantiles of the theoretical distribution.

Quantiles of the Theoretical Distribution:

These are the corresponding quantiles from the theoretical distribution (e.g., normal distribution).

Diagonal Line:

The line y=x represents the scenario where the sample quantiles perfectly match the theoretical quantiles.

Importance of Q-Q Plots in Linear Regression:

Validates Model Assumptions:

Ensuring residuals are normal is critical for valid p-values, confidence intervals, and hypothesis tests.

Helps Detect Model Misspecifications:

Deviations from normality in residuals may suggest that the model is not appropriately capturing the relationship between the variables (e.g., missing interactions or nonlinear terms).

Improves Interpretation and Prediction:

Normality ensures the model's outputs are reliable and interpretable, particularly when making predictions or drawing statistical inferences.

By examining the Q-Q plot, a modeler can diagnose and address issues in the regression model to improve its performance and reliability.

Use of a Q-Q Plot in Linear Regression:

1**. Checking the Normality Assumption**

Linear regression assumes that the residuals (errors) of the model are normally distributed. A Q-Q plot of the residuals is used to:

1. Assess if the residuals follow a normal distribution.
2. Identify deviations like skewness or heavy tails.

Interpretation:

Points close to the diagonal line: Residuals are approximately normal.

Points deviating systematically: Suggests issues like skewness, kurtosis, or other departures from normality.

2. **Detecting Outliers**

Points far away from the diagonal line in the Q-Q plot indicate potential outliers or extreme values.

3. **Identifying Nonlinear Patterns**

Curved patterns in the Q-Q plot can indicate that the residuals are not normally distributed and might require data transformation or adjustments to the model.