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)

1. Season (season)
As mentioned earlier, the season variable shows clear seasonal patterns:

- Spring (1) and Summer (2) are associated with higher bike rentals, as people prefer biking in warmer, pleasant weather.
- Fall (3) and Winter (4) typically show lower demand due to cooler temperatures and less favorable weather conditions for biking.

2. Weather Situation (weathersit)

- Clear weather (1) correlates with the highest bike rental demand.
- Partly cloudy (2) and cloudy weather (3) lead to moderate demand, as biking is still feasible.
- Rainy weather (4) has a significant negative impact, as fewer people opt to rent bikes during rainy conditions.

3. Year (yr)
The yr variable indicates a growing demand for bike rentals from 2018 (0) to 2019 (1), reflecting an upward trend in bike-sharing popularity.

4. Working Day (workingday)
The workingday variable, which takes values 0 (non-working day) or 1 (working day), has a stronger predictive powercompared to weekday:

- Working days (1) generally see higher bike demand, especially for commuting purposes. People tend to rent bikes to go to work or run errands.
- Non-working days (0) (weekends, holidays) might show lower demand, as people have more time to engage in leisure activities but might not need bikes for commuting.

Conclusion:

1. Season has a significant effect on bike demand, with higher demand in Spring and Summer and lower demand in Fall and Winter.
2. Weather conditions also heavily influence demand, with clear weather encouraging more bike rentals and rainy weather deterring bike use.
3. The year (yr) variable shows a positive trend in bike rentals, suggesting the business is growing over time.
4. The workingday variable is a strong predictor of bike demand, as demand is typically higher on working days due to commuting needs.

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)

When creating dummy variables (also called one-hot encoding) for categorical variables, drop_first=True is important because it helps to avoid the dummy variable trap, which can lead to issues in regression models, especially multicollinearity.

When we include all the dummy variables for a categorical feature with k categories, we may

introduce multicollinearity into our model. This happens because the sum of all the dummy variables for each row will always equal 1. For example:

- If Spring is encoded as (1, 0, 0, 0), it implies that the row represents Spring.
- If Summer is encoded as (0, 1, 0, 0), it implies that the row represents Summer, and so on.

Since the sum of all the dummies equals 1, one column (for example, Winter) can be perfectly predicted from the others. This introduces perfect collinearity and makes it impossible to uniquely estimate the coefficients in a regression model, as the model cannot distinguish between these perfectly correlated columns.
This is known as the dummy variable trap, where the regression model fails to find a unique solution due to high multicollinearity among the predictor variables.

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)

Registered feature.

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)

After building a multiple linear regression model, it's crucial to validate the assumptions of linear regression to ensure that the model is valid, reliable, and provides meaningful results. There are several key assumptions that need to be checked when using linear regression:

1. Linearity: The relationship between the dependent and independent variables should be linear.
2. Independence: The residuals (errors) should be independent of each other.
3. Homoscedasticity: The residuals should have constant variance across all levels of the independent variables.
4. Normality of residuals: The residuals should be normally distributed.
5. No multicollinearity: The independent variables should not be highly correlated with each other.

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)

Windspeed, casual, workingday

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>

Linear Regression is one of the most commonly used statistical methods for predictive modeling, especially for estimating the relationship between one or more independent (predictor) variables and a dependent (target) variable. It is used to predict continuous values (such as sales, price, or temperature) based on the relationship between the variables.

In a **simple linear regression** model, the relationship between a single independent variable (X) and the dependent variable (Y) is modeled as a straight line. In **multiple linear regression**, there are two or more independent variables.

1. Simple Linear Regression:

In simple linear regression, the model assumes that there is a linear relationship between the dependent variable.

2. Multiple Linear Regression:

In multiple linear regression, we extend the concept to multiple predictors (independent variables).

3. The Objective of Linear Regression:

The goal of the linear regression algorithm is to find the values of the coefficients.

4. Finding the Best Coefficients:

Linear regression typically uses a technique called **Ordinary Least Squares (OLS)** to estimate the coefficients. The idea is to find the values of the coefficients that minimize the **residual sum of squares (RSS)**.

5. Gradient Descent Algorithm (Alternative to OLS):

While OLS provides an exact solution, it may not always be efficient for large datasets with many predictors. An alternative method to solve for the coefficients is **Gradient Descent**.

Gradient Descent is an optimization algorithm used to minimize the loss function iteratively by updating the coefficients in the direction of the steepest descent of the loss function.

The algorithm continues to update the coefficients until the loss function converges to a minimum value.

6. Assumptions of Linear Regression:

Linear regression makes several assumptions about the data that must be validated for the model to provide reliable results:

1. **Linearity**: There should be a linear relationship between the predictors and the dependent variable.
2. **Independence of errors**: The residuals should be independent of each other (no autocorrelation).
3. **Homoscedasticity**: The residuals should have constant variance across all levels of the independent variables.
4. **Normality of errors**: The residuals should be normally distributed.
5. **No multicollinearity**: The predictors should not be highly correlated with each other.

Violating these assumptions can lead to biased or unreliable estimates of the coefficients.

7. Evaluation Metrics for Linear Regression:

To evaluate the performance of a linear regression model, we often use the following metrics:

- **R-squared (R2)**: It measures the proportion of the variance in the dependent variable that is explained by the independent variables.
- An R2 value close to 1 indicates that the model explains most of the variance in the target variable, while an R2 value close to 0 suggests a poor fit.
- **Mean Squared Error (MSE)**: It measures the average squared difference between the predicted and actual values. A lower MSE indicates better model performance.
- **Root Mean Squared Error (RMSE)**: It is the square root of the MSE and gives a more interpretable result in the original units of the dependent variable.
- **Mean Absolute Error (MAE)**: It measures the average absolute difference between the predicted and actual values. It is less sensitive to outliers than MSE.

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 collection of four datasets that have nearly identical simple descriptive statistics but very different distributions and appearances when plotted. It was created by the statistician **Francis Anscombe** in 1973 to demonstrate the importance of graphing data and the potential pitfalls of relying solely on summary statistics.

1. Purpose of Anscombe's Quartet

Anscombe's Quartet serves as a demonstration that relying only on numerical summaries like mean, variance, or correlation can be misleading. Even when these statistical measures are similar, the underlying data distributions might be vastly different, and this can have a

significant impact on the conclusions drawn from the data. The Quartet encourages us to visualize the data to gain insights into its true nature.

2. Key Takeaways from Anscombe's Quartet

- **Visualizing the Data**: The most important takeaway from Anscombe's Quartet is that visualizing the data can reveal insights that summary statistics alone cannot. Although the summary statistics (mean, variance, and correlation) are identical for all four datasets, the actual distributions of the data are very different.

- **Outliers**: Even a single outlier can dramatically affect the results of regression analysis. Dataset 3 shows that an outlier can drastically skew the results, even if most of the data follows a linear pattern.

- **Modeling Considerations**: The quartet demonstrates that choosing the correct model (e.g., linear vs. quadratic) and understanding the nature of the data (e.g., the effect of outliers) are crucial for building meaningful models. Relying on summary statistics like correlation can sometimes lead to incorrect assumptions and poor model choices.

3. Conclusion

Anscombe's Quartet serves as an excellent illustration of why it's important to not just rely on statistical summaries but also visualize the data. It teaches that the same correlation coefficient, mean, and variance can describe very different underlying data patterns, and different models might be required depending on the data structure. The quartet encourages analysts to always visualize their data before drawing conclusions or making predictions.

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 correlation coefficient (also known as **Pearson's r)** is a statistical measure that quantifies the linear relationship between two continuous variables. It is a value between -1 and +1 that indicates the strength and direction of the relationship.

Pearson's r is the most commonly used method to measure the linear correlation between two variables. The closer the value of Pearson's r is to 1 or -1, the stronger the linear relationship between the two variables. A value close to 0 suggests that there is little to no linear relationship between the variables.

Assumptions for Pearson's Correlation

For Pearson's r to give meaningful results, several assumptions should be met:

1. **Linearity**: The relationship between the two variables should be linear. This is why Pearson's correlation is best suited for linear relationships.
2. **Normality**: Both variables should ideally follow a normal distribution, especially when performing hypothesis testing or making inferences.
3. **Homoscedasticity**: The variability of one variable should be consistent across the values of the other variable (i.e., the spread of the data points should be roughly the same across all values of X).

If these assumptions are violated, the Pearson correlation may not accurately represent the relationship between the variables, and other correlation methods (like **Spearman's rank correlation**) might be more appropriate.

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 refers to the process of transforming data to fit a specific range or distribution, often to improve the performance of machine learning algorithms. It adjusts the range or distribution of numerical features so that they are on a similar scale. This is important when different features have different units of measurement (e.g., age in years, income in dollars) or vastly different ranges (e.g., one feature may range from 0 to 1000, while another might range from 0 to 1).

Scaling ensures that each feature contributes equally to the model and prevents certain features with larger ranges from disproportionately affecting the model's results.

2. Why is Scaling Performed?

Scaling is performed for several reasons:

- **Improved Convergence in Gradient-Based Optimization**: Many machine learning algorithms, especially those that use gradient descent (e.g., linear regression, logistic regression, neural networks), benefit from scaling because it helps the optimization process converge faster.
- **Equal Weight to Features**: If features are on different scales, algorithms that calculate distances (e.g., k-NN, k-means, SVM) might give more weight to features with larger values. Scaling ensures all features are treated equally.
- **Avoiding Bias in Regularization**: Regularization methods (like Ridge and Lasso) can be biased towards features with larger numerical values. Scaling ensures that regularization penalizes each feature equally.
- **Ensuring Fairness**: For algorithms like Principal Component Analysis (PCA), scaling ensures that each feature contributes equally to the analysis and avoids features with larger values dominating the results.

3. Types of Scaling

There are two common types of scaling techniques used in machine learning:

a. Normalization (Min-Max Scaling)

Normalization refers to rescaling the feature values so that they lie within a specific range, typically between 0 and 1.

b. Standardization (Z-score Scaling)

Standardization, also known as Z-score scaling, refers to transforming the feature values so that they have a **mean of 0**and a **standard deviation of 1**.

4. Differences Between Normalized Scaling and Standardized Scaling

Feature	Normalization (Min-Max Scaling)	Standardization (Z-score Scaling)
Range of values	Fixed, typically [0, 1]	Unbounded, values can range from $-\infty$ to $+\infty$
Assumptions	Assumes data is bounded and has a fixed range	Assumes data is normally distributed
Sensitivity to Outliers	Highly sensitive to outliers, can distort the data if there are extreme values	Less sensitive to outliers
Use cases	Suitable when data needs to be transformed into a specific range or when features have known boundaries (e.g., pixel values in images)	Suitable when data has varying units and is not bounded, or when you need to compare features with different units
Effect on Mean and Standard Deviation	Does not affect the mean or standard deviation directly	Changes the mean to 0 and the standard deviation to 1

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)** is a measure of how much the variance of a regression coefficient is inflated due to multicollinearity with other independent variables in the model. It quantifies how much the variance of a regression coefficient is increased because of correlations with other predictors in the model.

The VIF of a feature becomes **infinite** when the **R-squared value (R2) is equal to 1**. This happens in the following situations:

Perfect Multicollinearity:

- When one predictor is **perfectly correlated** with another predictor (or a set of predictors), the R2 value becomes 1.
- In this case, one feature can be **exactly predicted** by other feature(s) in the model. This is known as **perfect multicollinearity**.

Redundant Features:

- If a feature is **redundant** or an **exact duplicate** of another feature, the R-squared between the two will be 1. For example, if you have two features in the dataset where one is just a linear transformation or a duplicate of the other (e.g., feature1 = feature2), the VIF for either of them will be infinite.

Linear Dependence:

- If one feature is a **linear combination** of other features, then it creates a dependency where its variance cannot be explained without the others. This leads to an R2 value of 1, and thus an infinite VIF.

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 assess if a dataset follows a particular theoretical distribution, most commonly the **normal distribution**. The plot compares the quantiles (or percentiles) of the data against the quantiles of a specific distribution.

- **Q-Q Plot Definition**: A Q-Q plot is a scatter plot of the quantiles of the sample data against the quantiles of a specified reference distribution (e.g., normal distribution). If the data follows the reference distribution closely, the points will lie approximately along a straight line.

Q-Q Plot in Linear Regression

In linear regression, **residuals** (the differences between the observed and predicted values) should ideally be **normally distributed** if the assumptions of linear regression are to hold. This is because many of the statistical tests (such as hypothesis tests on the coefficients) rely on the assumption that the errors (residuals) are normally distributed.

Thus, a **Q-Q plot of residuals** is used to assess the **normality** of residuals in a linear regression model.

The Q-Q Plot is Important in Linear Regression because:

- **Normality of Residuals**: As mentioned, the residuals of a linear regression model should be normally distributed to ensure that the model's inferences are valid (e.g., p-values for hypothesis tests). A Q-Q plot helps verify this assumption.
- **Visualizing Deviations**: While other tests (like the Shapiro-Wilk test) can formally test for normality, a Q-Q plot provides a visual representation that can help in **detecting deviations** from normality, such as skewness, kurtosis, or outliers.
- **Improving Model Accuracy**: By checking if the residuals are normally distributed, the Q-Q plot helps in improving the model's accuracy. If the residuals do not follow a normal distribution, you can apply various techniques (e.g., transformations) to improve the model.