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:

From the analysis of categorical variables from the dataset, it is inferred as the bike rentals are likely to be higher in summer and winter seasons, prominent in months of August and September then trend started decreasing in end of year [in the months of Nov, Dec] till Feb. More over during Sundays there is a decline in rentals. Clear weather attracts more rentals indicating Cloudy mist weather and Light rain thunder [Bad weather days] has less rentals in the year 2019. When it is a holiday rentals are quite low probably indicating people wants to spend time at home.

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:

Drop_first= True is important to use, as it helps in reducing the extra columns created during dummy variable creation. And also it avoids the redundancy and correlation among dummy 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:

'Temp' variable has the highest correlation with target variable

---

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:

I have validated the assumptions of Linear regression based on

1. Normality of error [Errors are normally distributed]
2. Multicollinearity [desired VIF < 5]
3. Linear relationship between dependent and feature variables
4. Statistical significance [P-value < 0.05]
5. Compared adjusted R-squared of test and train datasets

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:

top 3 features contributing significantly towards demand of shared bikes –

1. Temp
2. Summer
3. September

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer:

Linear Regression is a statistical method used to model the relationship between a dependent variable y and one or more independent variables x. The goal is to find a line (or hyperplane in multiple dimensions) that best fits the data by minimizing the error between the predicted and actual values.

Key Steps:

1. **Model**: $y=\beta 0+\beta 1x+C$
2. **Optimization**: The algorithm minimizes the **sum of squared errors** (residuals) to find the best values for $\beta 0$ and $\beta 1$.
3. **Prediction**: Once the coefficients are determined, the model can predict y for new values of x.

Linear regression is widely used for predicting continuous outcomes based on linear relationships between variables.

Assumptions of linear regression:

1. **Linearity**: The relationship between the dependent and independent variables is linear.
2. **Independence**: The residuals (errors) are independent of each other.
3. **Homoscedasticity**: The variance of the residuals is constant across all levels of the independent variable.
4. **Normality of Errors**: The residuals are normally distributed.
5. **No Multicollinearity** (for multiple regression): The independent variables are not highly correlated with each other.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)

Answer:

Anscombe's Quartet emphasizes the importance of **data visualization** in understanding the underlying patterns and potential issues (like outliers) in the data, which summary statistics alone might miss.

It is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but appear very differently when graphed.

Datasets:

The quartet consists of four datasets, each with:

- The same **mean** of x and y,
- The same **variance** for both x and y,
- The same **correlation** between x and y,
- The same **regression line**.

However, despite these similarities, the datasets have very different underlying patterns when visualized.

Visualization:

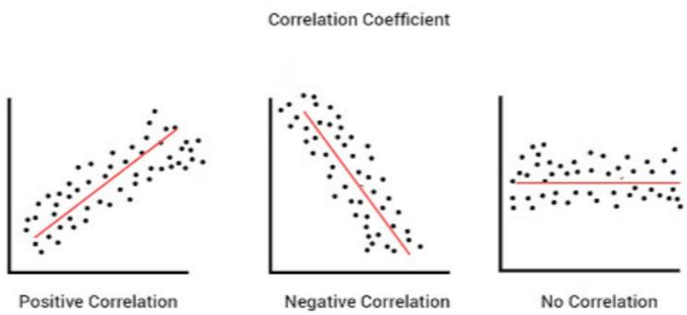
1. **Dataset 1**: A linear relationship between x and y (straight line).
2. **Dataset 2**: A perfect quadratic relationship (parabolic curve).
3. **Dataset 3**: A linear relationship with one extreme outlier that greatly influences the regression line.
4. **Dataset 4**: A strong linear relationship with one vertical line of points and an outlier that distorts the regression line.

Question 8. What is Pearson’s R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

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

- **r=+1**: Perfect positive linear correlation — as one variable increases, the other increases in a perfectly linear fashion.
- **r = -1**: Perfect negative linear correlation — as one variable increases, the other decreases in a perfectly linear fashion.
- **r=0**: No linear correlation — the variables do not have a linear relationship.
- **0<r<1**: Positive linear correlation — as one variable increases, the other tends to increase.
- **−1<r<0**: Negative linear correlation — as one variable increases, the other tends to decrease.

Pearson’s R is sensitive to outliers, which can significantly affect the correlation value.



Example of Pearson's R:

Consider a study examining the relationship between **hours studied** and **exam scores** for a group of students.

- If **Pearson's R = 0.9**, it indicates a strong positive correlation: as the number of hours studied increases, the exam scores tend to increase as well.
- If **Pearson's R = -0.8**, it indicates a strong negative correlation: as hours spent on social media increase, exam scores decrease.
- If **Pearson's R = 0**, it suggests no linear relationship: hours studied and exam scores may not be related in a linear way.

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:

Scaling refers to the process of adjusting the range or distribution of feature values in a dataset, so that they have similar scales or magnitudes. This is particularly important for machine learning models that are sensitive to the scale of the data, such as those using distance calculations or gradient-based optimization.

Why is Scaling Performed?

1. **Improves model performance**: Some algorithms, like K-Nearest Neighbors or gradient descent, perform better and converge faster when features are on a similar scale.
2. **Prevents dominance**: Features with larger ranges can dominate the model, leading to biased results.
3. **Ensures equal contribution**: All features contribute equally to the model when they are scaled.

Difference Between Normalized Scaling and Standardized Scaling:

1. **Normalized Scaling (Min-Max Scaling)**:
   - Rescales data to a fixed range, typically [0, 1].
   - **Formula**: $Xnorm=(X-Xmin)/(Xmax-Xmin)$

2. **Standardized Scaling (Z-score Scaling)**:
   - Centers data around the mean (0) and scales based on the standard deviation (1).
   - **Formula**: Xstd=(X−μ)/σ
   - **Use**: Best when data follows a normal distribution or when dealing with algorithms that rely on the distance between data points.

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:
A **VIF (Variance Inflation Factor)** becomes **infinite** when there is **perfect multicollinearity** between two or more independent variables.
This means one variable is a perfect linear function of another (e.g., x2=2x1) or **nearly perfectly linearly** related to other predictors (Even if the relationship is not exact, if the correlation between two predictors is very close to 1 or -1, it can cause the VIF to approach infinity).

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:
A **Q-Q plot** (Quantile-Quantile plot) compares the distribution of a dataset (e.g., residuals in linear regression) to a theoretical distribution, such as the normal distribution. If the points lie along a straight line, the data follows the normal distribution.

Use in Linear Regression:

- **Check normality of residuals**: Ensures that residuals are normally distributed, which is an assumption for valid hypothesis testing and inference in linear regression.
- **Detect deviations**: Helps identify outliers or non-normality (skewness, heavy tails).

Importance:

- Validates regression assumptions.
- Supports reliable p-values and confidence intervals.
- Identifies potential outliers.
