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)

[Answer]

From the analysis of the categorical variables in the dataset, we can infer the following about their effect on the dependent variable (bike demand or cnt):

- ***Season (season)***:
  - The categorical variable for season was split into different categories like spring, summer, fall, and winter.
  - Among these, spring showed a significant negative impact on bike demand. This suggests that people tend to rent fewer bikes during spring compared to other seasons.
- **Year (yr):**
  - The yr variable had a strong positive coefficient, suggesting that bike demand increased in 2019 compared to 2018. This could be due to the growing popularity of the bike-sharing service over time.
- **Holiday (holiday):**
  - The variable holiday was found to be statistically insignificant, meaning that whether it is a holiday or not doesn't have a major impact on bike demand in this context.
- **Working Day (workingday):**
  - The variable workingday had a significant positive impact, indicating that bike rentals increase on working days. This aligns with the expectation that people use bike-sharing services more during weekdays for commuting.
- **Weekday (weekday):**
  - Specific days like Saturday and Sunday had a significant effect.
- **Weather Situation (weathersit):**
  - Different weather situations, like clear weather, misty weather, or light snow, were included as categorical variables.
  - Clear weather positively impacts bike demand, as expected, since people are more likely to rent bikes in favorable weather conditions.
  - Light snow or rain has a negative impact, indicating that bad weather reduces bike demand.

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

[Answer]

Using 'drop_first=True' during dummy variable creation is important for preventing multicollinearity and ensuring that the regression model remains interpretable and stable. It simplifies the model and makes the coefficients more meaningful.

Example:

- If we have a categorical variable Color with three categories: Red, Green, and Blue
- We create dummy variables without drop_first=True then we will end up with three dummies: Color_Red, Color_Green, Color_Blue.
- We create dummy variables with drop_first=True then all three in a regression model would cause multicollinearity because if you know the values of two, you can determine the third.
- With drop_first=True, we would drop one dummy (e.g., Color_Red), leaving Color_Green and Color_Blue. In this case, the category "Red" becomes the reference, and the coefficients for Color_Green and Color_Blue would tell you how those categories differ from "Red".

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

[Answer]

‘temp’ variable has the highest correlation with the target variable `cnt`

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

[Answer]

To validate the assumptions of Linear Regression after building the model on the training set, you typically follow these steps:

- Linearity:
  - Residual plots can reveal whether the relationship between the independent and dependent variables is truly linear.
  - The residuals should be randomly distributed around zero, indicating a linear relationship.
- Multicollinearity:
  - Multicollinearity is checked using the Variance Inflation Factor (VIF) for each predictor variable.
  - A VIF value greater than 5 indicates significant multicollinearity, and steps should be taken to address it.
  - Multicollinearity can inflate the variance of coefficient estimates, making the model unstable.
- Homoscedasticity (Constant Variance of Errors):
  - Check for homoscedasticity by plotting the residuals against the predicted values.
  - If the variance of the residuals is constant across all levels of the predicted values, the plot will show a random scatter.
- Normality of Errors:
  - The residuals should be normally distributed.
  - If the residuals fall along the 45-degree line, they are normally distributed.

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

[Answer]

Based on the final model analysis, the top 3 features contributing significantly towards explaining the demand for shared bikes are likely to be:

- Year (yr):
  - The year variable typically shows a strong positive correlation with bike demand, reflecting an overall increase in bike rentals over time.
- Temperature (temp):
  - Temperature is usually a key predictor, as higher temperatures generally encourage more people to rent bikes, making it a significant factor in explaining bike demand.
- Working Day (workingday):
  - Whether it is a working day or not significantly influences bike demand, with higher rentals typically observed on working days, particularly due to commuting.

General Subjective Questions

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

[Answer]

Linear Regression is a fundamental statistical and machine learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal of linear regression is to find the best-fitting line (or hyperplane in the case of multiple variables) that describes the relationship between the dependent and independent variables. This is achieved by minimizing the difference between the actual data points and the predicted values on this line.

There are two main types of linear regression:

1. **Simple Linear Regression:**
   - This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.
   - The equation for simple linear regression is: $y=\beta0+\beta1x$
     where:
       Y is the dependent variable
       X is the independent variable
       $\beta0$ is the intercept
       $\beta1$ is the slope

2. **Multiple Linear Regression**
    - This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is: $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn$
      where:
      Y is the dependent variable
      X1, X2, …, Xn are the independent variables
      β0 is the intercept
      β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

Linear Regression is a powerful, interpretable algorithm that serves as the foundation for many other more complex models. Its effectiveness depends on the assumptions being met and the careful consideration of the model's limitations. Regularization and careful model evaluation are key steps in ensuring that the model performs well on unseen data.

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

[Answer]

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. The quartet was created by the statistician **Francis Anscombe** in 1973 to illustrate the importance of data visualization in statistical analysis and to show that relying solely on summary statistics can be misleading.

The Four Datasets

Here are the four datasets in Anscombe's Quartet:

Dataset I

- This is a classic linear relationship, where the points closely follow a straight line.
- A linear regression model is appropriate, and the summary statistics accurately reflect the underlying data.

Dataset II

- The x-values are constant except for one point, and the y-values vary around a horizontal line.
- Despite the summary statistics suggesting a relationship, the data actually form a horizontal line with one outlier. The apparent correlation is due to the single outlier.

Dataset III

- Similar to Dataset I, but with an outlier.
- The summary statistics suggest a linear relationship, but the plot shows that one outlier is distorting the relationship. The other points form a vertical line, meaning no real linear relationship exists.

Dataset IV

- The x-values are almost identical, except for one extreme value.
- The y-values are nearly constant, except for one outlier.
- The linear regression line is influenced by the single outlier, and the plot reveals that the data do not follow a linear relationship.

Visual Representation

The power of Anscombe's Quartet lies in the visual representation of the data. When you plot each dataset:

- **Dataset I** shows a linear trend, and the summary statistics make sense.
- **Dataset II** shows that the correlation is driven by a single outlier, which is misleading.
- **Dataset III** demonstrates how an outlier can skew the perception of a relationship.
- **Dataset IV** shows how extreme values can influence summary statistics, even when there is no meaningful relationship in the data.

Importance of Anscombe’s Quartet

Anscombe's Quartet serves as a cautionary tale in statistical analysis. It highlights several important lessons:

1. **Always Visualize Your Data**: Summary statistics alone can be misleading. Plotting the data helps you see the true nature of the relationships.
2. **Be Wary of Outliers**: Outliers can heavily influence statistical measures like correlation and regression coefficients. Visualizing the data helps identify these outliers.
3. **Context Matters**: Understanding the context and the story behind the data is crucial. Even if two datasets have the same summary statistics, they may tell very different stories.
4. **Avoid Over-reliance on Statistics**: While summary statistics are useful, they should not be the only tool used for data analysis. Visualization and deeper exploratory analysis are necessary for a complete understanding.

Purpose of Anscombe’s Quartet:
Anscombe’s quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson’s R? (3 marks)

[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 one of the most commonly used correlation coefficients in statistics.

Key Characteristics of Pearson's R:

1. **Range**:
   - Pearson's R ranges from **-1 to 1**.
   - A value of **1** indicates a perfect positive linear relationship, meaning as one variable increases, the other variable also increases proportionally.
   - A value of **-1** indicates a perfect negative linear relationship, meaning as one variable increases, the other decreases proportionally.
   - A value of **0** indicates no linear relationship between the variables.
2. **Interpretation**:
   - **Positive R**: Indicates that as one variable increases, the other tends to increase.
   - **Negative R**: Indicates that as one variable increases, the other tends to decrease.
   - **Magnitude**: The closer the value is to 1 or -1, the stronger the linear relationship.
3. **Assumptions**:
   - **Linearity**: Pearson's R measures only the strength of a linear relationship. Non-linear relationships may not be accurately captured by this coefficient.
   - **Continuous Variables**: Both variables should be continuous and measured on an interval or ratio scale.
   - **Normality**: It assumes that both variables are normally distributed.
   - **Homogeneity of Variance**: The spread of data points should be roughly equal across the range of values.

Applications:

- Pearson's R is widely used in fields like finance, medicine, social sciences, and more to understand the relationship between variables, such as height and weight, or the relationship between stock returns.
- In summary, Pearson's R is a valuable tool for understanding the linear association between two variables, but it should be used with an understanding of its limitations and assumptions.

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

[Answer]
Scaling is the process of adjusting the range of features in your data so that they fit within a specific scale. This is typically done to ensure that all features contribute equally to the model, especially in algorithms that rely on distance calculations (e.g., k-nearest neighbors, support vector machines, and gradient descent-based methods like linear regression and neural networks).

Scaling is performed for several reasons:

1. **Improves Model Performance**: Many machine learning algorithms perform better when features are on a similar scale. For example, in gradient descent optimization, features with larger scales can dominate the objective function, leading to suboptimal learning.
2. **Faster Convergence**: Scaling can speed up convergence in gradient-based algorithms by ensuring that all features contribute similarly to the cost function.
3. **Reduces Bias**: In algorithms like k-nearest neighbors and support vector machines, scaling ensures that no feature disproportionately influences the distance calculation due to its scale.
4. **Stability**: Scaling can make the computation more numerically stable, particularly in models that involve matrix inversion or other linear algebra operations.

Key Differences Between Normalization and Standardization:

- **Range**:
  - **Normalization**: Scales data to a specific range (e.g., [0, 1]).
  - **Standardization**: Scales data to have a mean of 0 and a standard deviation of 1.
- **Impact of Outliers**:
  - **Normalization**: Sensitive to outliers because the min and max values can be drastically affected by extreme values.
  - **Standardization**: Less sensitive to outliers, but extreme outliers can still impact the mean and standard deviation.
- **When to Use**:
  - **Normalization**: Useful when the features are on different scales and when a specific range is required.
  - **Standardization**: Preferred when data follows a normal distribution or when you need features to have similar distributions.

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

[Answer]
The value of the **Variance Inflation Factor (VIF)** becomes infinite when there is **perfect multicollinearity** among the independent variables in a regression model. Perfect multicollinearity means that one or more independent variables are exact linear combinations of

others, which makes it impossible to estimate their unique contributions to the dependent variable.

Causes of Infinite VIF:

1. **Duplicate Variables**:
   - If you accidentally include the same variable more than once or include variables that are exact linear transformations of each other, VIF will be infinite.
2. **Highly Correlated Variables**:
   - When two or more variables are highly correlated (nearly perfect multicollinearity), their VIF values can become extremely large or infinite, indicating that one variable can be almost exactly predicted by the others.
3. **Dummy Variable Trap**:
   - When using dummy variables to represent categorical data, if you include all categories without dropping one (e.g., using drop_first=True), the model can suffer from perfect multicollinearity, leading to infinite VIF values.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

[Answer]

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. The plot displays the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the theoretical distribution closely, the points on the Q-Q plot will fall approximately along a straight line.

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

In the context of linear regression, a Q-Q plot is particularly important for assessing one of the key assumptions: **normality of the residuals**.

1. Assessing Normality of Residuals:

- Linear regression assumes that the residuals (the differences between the observed and predicted values) are normally distributed. This assumption is important because many inferential statistics, such as confidence intervals and hypothesis tests for coefficients, rely on the normality assumption.
- A Q-Q plot of the residuals can be used to check this assumption. If the residuals are normally distributed, the points on the Q-Q plot will follow the straight diagonal line.

2. Detecting Deviations from Normality:

- **Heavy Tails**: If the points on the Q-Q plot form a curve that bows away from the line, it may indicate that the residuals have heavier tails than a normal distribution (i.e., more extreme values).
- **Skewness**: If the points form an S-shape, it indicates that the residuals are skewed, either to the left or right, meaning the model might be biased or that a transformation of the data might be necessary.

3. Model Diagnostics:

- The Q-Q plot is a vital diagnostic tool in linear regression because it can reveal whether the model's assumptions hold. If the residuals deviate significantly from normality, it suggests that the model may not be appropriate or that certain assumptions are violated. This could lead to unreliable predictions and inferential statistics.

4. Corrective Actions:

- If the Q-Q plot shows significant deviations from normality, you might consider:
  - **Transforming the Variables**: Applying a transformation (e.g., log, square root) to the dependent variable or predictors to achieve normality.
  - **Using a Different Model**: If normality is a severe issue, consider using a different modeling approach that does not require normally distributed residuals (e.g., generalized linear models).