Assignment-based Subjective Questions

Q 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
Ans 1:

1. Fall season has highest demand for rental bikes
2. Demand for next year (2019) has grown
3. Demand is continuously increasing each month until September. September month has highest demand. After September, the demand is decreasing
4. When there is a holiday, demand has decreased
5. Weekdays show consistent demand all days
6. No changes in demand with respect to Working days
7. The good weathersit has highest demand. Demand decreases as the weather goes from good to bad. There are no demand when it Rains (severe weather)
8. During September, bike sharing is more. During the end and beginning of year, it is less.

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

Ans 2: drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation. It reduces the correlations created among dummy variables. (p-1) dummies can explain p categories .

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

Ans 3: temp and atemp have the highest correlation with the target variable (cnt).

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

1. Residual Analysis: Errors are normally distributed with mean 0
2. Homoscedacity: We observed that variance of the residuals (error terms) is constant across predictions. i.e error term does not vary much as the value of the predictor variable changes
3. R2 value for predictions on test data (0.779) is almost same as R2 value of train data(0.788). This is a good R-squared value, hence we can see our model is performing good even on unseen data (test data)

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

1. Temp
2. Season
3. Weather situation

General Subjective Questions

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

Ans 1: Linear regression is a fundamental algorithm in machine learning and statistics used for modelling the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting linear relationship between the dependent variable and the independent variables.

Key Concepts of Linear Regression

1. **Model Representation**
   - For a simple linear regression (one predictor), the model is represented as:

     $y=\beta_0+\beta_1 x+\epsilon$

     where:

     - y is the dependent variable.
     - x is the independent variable.
     - β_0 is the intercept.
     - β_1 is the slope (coefficient) of the independent variable.
     - ϵ is the error term (residual).
   - For multiple linear regression (multiple predictors), the model is:

     $y=\beta_0+\beta_1+\beta_2+\cdots+\beta_n+\epsilon$

     where:

     - $x_1, x_2, \ldots, x_n$ are the independent variables.
     - $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients for each independent variable.

2. **Assumptions of Linear Regression**
   - **Linearity**: The relationship between the dependent and independent variables is linear.
   - **Independence**: The residuals are independent.
   - **Homoscedasticity**: The residuals have constant variance at every level of the predictor.

- **Normality**: The residuals are normally distributed (particularly important for hypothesis testing).
- **No multicollinearity**: The predictors are not too highly correlated with each other.

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

Ans 2: Anscombe's quartet is a set of four datasets that are used to illustrate the importance of graphing data before analysing it with statistical methods. The quartet was constructed by the statistician Francis Anscombe in 1973 to demonstrate how datasets that have nearly identical simple descriptive statistics can have very different distributions and appearances when graphed. This example underscores the need for visual data analysis and not solely relying on summary statistics.

The Four Datasets

Each of the four datasets in Anscombe's quartet contains 11 pairs of (x,y) values. Despite having nearly identical summary statistics, the datasets exhibit very different characteristics when plotted. Here's a detailed look at these datasets:

Summary Statistics

- **Mean of x**: Each dataset has a mean x of 9.
- **Mean of y**: Each dataset has a mean y of approximately 7.50.
- **Variance of x**: Each dataset has a variance x of 11.
- **Variance of y**: Each dataset has a variance y of approximately 4.12.
- **Correlation between x and y**: Each dataset has a correlation coefficient of approximately 0.816.
- **Linear regression line**: For each dataset, the linear regression line is y=3+0.5x.

Dataset 1

This dataset appears to be a simple linear relationship between x and y with some random noise.

Dataset 2

This dataset consists of a quadratic relationship where y increases as x increases, but only one point is causing a slight distortion from a perfect quadratic fit.

Dataset 3

In this dataset, all x values are the same except for one point. This single point drives the relationship seen in the regression statistics.

Dataset 4

This dataset has a vertical arrangement with an outlier that drives the linear regression results. The outlier significantly impacts the slope and intercept of the regression line.

Visual Analysis

Plotting each dataset reveals their distinct characteristics, despite identical summary statistics:

1. **Scatter Plots**: The scatter plots for each dataset show very different distributions and patterns.
2. **Linear Fit**: Each dataset's regression line is almost identical, even though the data points' distribution around the line varies greatly.

Importance of Graphing Data

Anscombe's quartet demonstrates that relying on summary statistics alone can be misleading because different datasets can produce similar statistical properties but have fundamentally different distributions and relationships. Visualizing the data allows for a deeper understanding of the underlying patterns, relationships, and potential anomalies.

Q 3: What is Pearson’s R? (3 marks)

Ans 3: Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol r and its value ranges from -1 to 1.

Key Features of Pearson's R

1. **Range**: The coefficient ranges from -1 to 1.
    - $r=1$: Perfect positive linear correlation.
    - $r=-1$: Perfect negative linear correlation.
    - $r=0$: No linear correlation.
2. **Direction**: Indicates whether the relationship between the variables is positive or negative.
    - Positive correlation ($r>0$): As one variable increases, the other variable also increases.
    - Negative correlation ($r<0$): As one variable increases, the other variable decreases.
3. **Magnitude**: Reflects the strength of the correlation.
    - Closer to 1 or -1: Strong correlation.
    - Closer to 0: Weak correlation.

Formula for Pearson's R

The formula for calculating Pearson's R between two variables X and Y is:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$

Where

- X_i and Y_i are the individual data points of variables X and Y.
- X bar and Y bar are the mean values of X and Y.

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

Ans 4: Scaling is a preprocessing step in data analysis and machine learning that involves adjusting the range and distribution of data features so that they can be analysed and processed more effectively. This step is particularly important for algorithms that are sensitive to the scale of data, such as gradient descent-based methods (e.g., linear regression, logistic regression) and distance-based methods (e.g., k-nearest neighbours, support vector machines).

Why Scaling is Performed

1. **Improves Convergence Speed**: For optimization algorithms like gradient descent, scaling can significantly speed up convergence by ensuring that features contribute equally to the calculation of the gradients.
2. **Improves Accuracy**: Many machine learning algorithms assume that features are centered around zero and have the same variance. Without scaling, some features might dominate others simply due to their larger magnitude.
3. **Equal Contribution**: Ensures that all features contribute equally to the model, avoiding bias towards features with larger ranges.
4. **Distance-Based Algorithms**: Algorithms like k-nearest neighbors, k-means clustering, and SVMs calculate distances between data points. Scaling ensures that all features influence the distance calculations equally.

Differences Between Normalized Scaling and Standardized Scaling

1. **Range**:
    - **Normalization**: Transforms data to a fixed range, typically [0, 1].
    - **Standardization**: Transforms data to have a mean of zero and a standard deviation of one, without a fixed range.
2. **Impact on Data**:
    - **Normalization**: Can affect the magnitude of outliers since they will be scaled to the fixed range, possibly reducing their impact.

- **Standardization**: Preserves the relative distances between values but transforms the distribution to have a specific mean and variance.

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

Ans 5: The VIF is calculated as: $VIF = 1/(1 - R_i^2)$

VIF becomes infinite when $R_i^2 = 1$. This situation occurs when there is perfect multicollinearity, meaning that the predictor is a perfect linear combination of one or more of the other predictors in the model. In other words, one predictor is completely predictable from the others, which leads to the denominator in the VIF formula becoming zero, and thus VIF approaches infinity.

Causes of Infinite VIF

1. **Perfect Multicollinearity**: This is the primary cause. Perfect multicollinearity means that there is an exact linear relationship between the predictor variables. For example, if X_1 and X_2 are linearly dependent, such that $X_1 = aX_2 + b$ for some constants a and b, the VIF for these variables will be infinite.
2. **Redundant Variables**: Including a variable more than once in the dataset, either directly or as a linear combination of other variables, will lead to infinite VIF values. For example, if you accidentally include the same column twice or if one column is a perfect sum of two other columns.
3. **Data Issues**: Errors in the dataset, such as duplicate columns or exact copies of one variable, can result in perfect multicollinearity.

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

Ans 6: A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps to visually assess if the data follows a specified distribution by plotting the quantiles of the dataset against the quantiles of the theoretical distribution.

Use and Importance of Q-Q Plot in Linear Regression

In the context of linear regression, Q-Q plots are crucial for validating the assumptions of the regression model, particularly the assumption that the residuals (errors) are normally distributed. This is important for:

1. **Inference Validity**: Many inferential statistics, such as confidence intervals and hypothesis tests, rely on the assumption of normality. Non-normal residuals can lead to invalid inferences.
2. **Model Diagnostics**: Helps in identifying issues such as outliers, skewness, and kurtosis in the residuals, which may affect model performance and prediction accuracy.

3. **Transformations**: Suggests the need for transformations of the response variable or predictors to achieve normality and improve model fit.