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:

There were 6 categorical variables in the dataset. I used Box plot to study their effect on the dependent variable count of bikes that is ('cnt') .

The inference derived are:

- season: Spring season has lowest number of bike demand ('cnt'), and Fall had highest.

- mnth: September had most bike rentals and December had least.

- weathersit: in rain and snowy weather the demand is low, highest is usually in clear weather conditions.

- holiday: Bike rentals are lower on Holidays.

- weekday: weekday variable shows very close trend

- workingday: It had little effect on bike rental counts.

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

Answer:

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obviously unfurnished. So we do not need 3rd variable to identify the unfurnished.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Answer: There is linear relationship between temp and atemp. Both of the parameters cannot be used in the model due to multicollinearity .

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

Answer: The distribution of residual (by plotting distplot )should be normal and centered around Zero.

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 - As per our final Model, the top 3 predictor variables that influences the bike booking are:
Temperature (temp) - A coefficient value of '0.5682' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5682 units.

Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2535' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2535 units.

Year (yr) - A coefficient value of '0.2334' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2334 units.

**General Subjective Questions:**

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

Answer:
Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. The fundamental idea behind linear regression is to find the best-fitting linear relationship that predicts the values of the dependent variable based on the values of the independent variables.
Here is a detailed explanation of the linear regression algorithm:

**1. Assumptions:**

Linear regression relies on several assumptions, including:

- **Linearity:** Assumes that the relationship between the independent and dependent variables is linear.
- **Independence:** Assumes that the residuals (the differences between observed and predicted values) are independent of each other.
- **Homoscedasticity:** Assumes that the variance of the residuals is constant across all levels of the independent variables.
- **Normality:** Assumes that the residuals are normally distributed.

**2. Simple Linear Regression:**

- **Model:** The simple linear regression model is represented as $Y=\beta0+\beta1X+\varepsilon$
    - Y is the dependent variable.
    - X is the independent variable.
    - $\beta0$ is the y-intercept (constant term).
    - $\beta1$ is the slope of the line.
    - $\varepsilon$ is the error term.

**3. Multiple Linear Regression:**

- **Model:** Extends simple linear regression to multiple independent variables: $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn+\varepsilon$
    - Y*Y* is the dependent variable.
    - X1,X2,…,Xn*X*1,*X*2,…,*Xn* are the independent variables.
    - β0 is the y-intercept.
    - β1,β2,…, are the coefficients for the independent variables.
    - ε is the error term.

**4. Objective Function (Cost Function):**

- The goal is to find the values of β0,β1,…, that minimize the sum of squared differences between the observed and predicted values.
- The objective function, also known as the cost function, is often expressed as the Mean Squared Error (MSE) or the Mean Absolute Error (MAE).

**5. Ordinary Least Squares (OLS):**

- The most common method for estimating the coefficients is the Ordinary Least Squares approach.
- OLS minimizes the sum of squared differences between observed and predicted values.

**6. Gradient Descent (Alternative Optimization):**

- An alternative method for finding the optimal coefficients is gradient descent.
- It iteratively adjusts the coefficients to minimize the cost function.

**7. Model Evaluation:**

- Once the model is trained, it needs to be evaluated on new data.
- Common metrics for evaluation include Mean Squared Error (MSE), Mean Absolute Error (MAE), and R2 (coefficient of determination).

**8. Interpretation of Coefficients:**

- The coefficients (β0,β1,…,βn) represent the impact of each independent variable on the dependent variable, holding other variables constant.

**9. Regularization (Optional):**

- Regularization techniques like Ridge Regression and Lasso Regression can be applied to prevent overfitting and improve the model's generalization.

**10. Applications:**

- Linear regression is widely used in various fields for tasks such as predicting sales, analysing the impact of variables on an outcome, and understanding relationships between variables.

Despite its simplicity, linear regression remains a powerful and interpretable tool in statistical modelling. It provides insights into the relationships between variables and serves as a foundation for more advanced regression techniques.

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 but differ significantly when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analysing it and the limitations of relying solely on summary statistics. Each dataset consists of 11 (x, y) points.

**Characteristics of Anscombe's Quartet:**

1. **Dataset I:**

- **Descriptive Statistics:**
  - Mean of x: 9.0
  - Mean of y: 7.5
  - Variance of x: 11.0
  - Variance of y: 4.125
  - Correlation between x and y: 0.816
  - Linear regression: y=3.0+0.5x$y$=3.0+0.5$x$
- **Graphical Representation:**
  - Scatter plot: A clear linear relationship.

2. **Dataset II:**
   - **Descriptive Statistics:**
     - Mean of x: 9.0
     - Mean of y: 7.5
     - Variance of x: 11.0
     - Variance of y: 4.125
     - Correlation between x and y: 0.816
     - Linear regression: y=3.0+0.5x$y$=3.0+0.5$x$
   - **Graphical Representation:**
     - Scatter plot: A clear curved relationship (non-linear).
3. **Dataset III:**
   - **Descriptive Statistics:**
     - Mean of x: 9.0
     - Mean of y: 7.5
     - Variance of x: 11.0
     - Variance of y: 4.125
     - Correlation between x and y: 0.816
     - Linear regression: y=3.0+0.5x$y$=3.0+0.5$x$
   - **Graphical Representation:**
     - Scatter plot: A clear linear relationship except for one outlier.
4. **Dataset IV:**
   - **Descriptive Statistics:**
     - Mean of x: 9.0
     - Mean of y: 7.5
     - Variance of x: 11.0
     - Variance of y: 4.125
     - Correlation between x and y: 0.816
     - Linear regression: y=3.0+0.5x$y$=3.0+0.5$x$ (except for one outlier)
   - **Graphical Representation:**
     - Scatter plot: A clear linear relationship except for one influential outlier that affects the regression line.

**Key points:**

1. **Same Descriptive Statistics:**
   - All four datasets have the same means, variances, and correlation coefficients, which makes them indistinguishable when looking only at summary statistics.
2. **Graphical Differences:**

    - Despite having identical summary statistics, the datasets exhibit different patterns when visualized.
    - The importance of visualizing data is emphasized by Anscombe's quartet because summary statistics alone might not reveal the true nature of the data.
3. **Statistical Caution:**
    - It highlights the importance of considering both graphical and statistical methods in data analysis.
    - Relying solely on summary statistics can lead to misinterpretation.
4. **Educational Tool:**
    - Anscombe's quartet is often used as an educational tool in statistics courses to demonstrate the concept of variability and the limitations of relying on summary statistics.

In summary, Anscombe's quartet serves as a powerful illustration of the need for graphical exploration of data alongside numerical summaries to gain a comprehensive understanding of datasets.

3. What is Pearson's R? (3 marks)
Answer:
Pearson's correlation coefficient, often denoted as r, is a measure of the linear relationship between two variables. It quantifies the strength and direction of a linear association between two continuous variables. The coefficient ranges from -1 to 1, where:

- r=1 indicates a perfect positive linear relationship.
- r=−1 indicates a perfect negative linear relationship.
- r=0indicates no linear relationship.

Key points about Pearson's correlation coefficient:

1. **Direction of Relationship:**
    - If r>0, it indicates a positive correlation (as one variable increases, the other tends to increase).
    - If r<0, it indicates a negative correlation (as one variable increases, the other tends to decrease).
2. **Strength of Relationship:**
    - The closer |r| is to 1, the stronger the linear relationship.
    - r=0 implies no linear relationship, but it does not rule out the possibility of other types of relationships.
3. **Assumption:**
    - Pearson's correlation assumes a linear relationship. If the relationship is non-linear, r may not accurately represent the association.
4. **Sensitive to Outliers:**
    - Pearson's correlation coefficient can be influenced by outliers, particularly if they have a substantial impact on the means.
5. **Not a Measure of Causation:**
    - Correlation does not imply causation. Even if two variables are strongly correlated, it doesn't mean that changes in one variable cause changes in the other.
6. **Range:**

- The range of r is from -1 to 1. A value of 0 indicates no linear relationship, while -1 and 1 indicate perfect negative and positive linear relationships, respectively.

Pearson's correlation is commonly used in various fields, including statistics, finance, biology, and social sciences, to examine relationships between variables. It's a widely used and important tool for understanding associations in quantitative data.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:
Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

1- Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
2- Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ). sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?(3 marks)
Answer:
The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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, which stands for Quantile-Quantile plot, is a graphical tool used to assess whether a given set of data follows a specific theoretical distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically

the normal distribution. In the context of linear regression, Q-Q plots are often used to check the normality of residuals, which is an important assumption of linear regression models.

**Components of a Q-Q Plot:**

1. **Quantiles:**
    - The x-axis of a Q-Q plot represents the theoretical quantiles of a chosen distribution (e.g., normal distribution).
2. **Sample Quantiles:**
    - The y-axis represents the sample quantiles of the observed data.
3. **Line of Equality:**
    - A reference line (usually a diagonal line) is drawn on the plot, representing the line of equality. If the points on the Q-Q plot lie close to this line, it suggests that the observed data follows the chosen theoretical distribution.

**Use and Importance in Linear Regression:**

1. **Normality Check:**
    - Q-Q plots are commonly used to assess whether the residuals (the differences between observed and predicted values) in a linear regression model are normally distributed. Normality of residuals is a key assumption for valid hypothesis testing and confidence intervals in linear regression.
2. **Identification of Outliers:**
    - Outliers in the residuals can be identified through deviations from the expected pattern in the Q-Q plot. Outliers may indicate issues with the model or the presence of influential data points.
3. **Model Assumption Validation:**
    - Linear regression assumes that the residuals are normally distributed. A well-behaved Q-Q plot provides evidence that this assumption holds, enhancing the validity of statistical inferences drawn from the model.
4. **Interpretation:**
    - Q-Q plots are easy to interpret. If the points on the plot closely follow the line of equality, it suggests that the residuals are approximately normally distributed. Deviations from the line may indicate departures from normality.
5. **Diagnosis of Distributional Assumptions:**
    - While normality is a common assumption, Q-Q plots can also be used to check the distributional assumptions of other types of models. For example, they can be used to assess whether residuals from a logistic regression model follow a logistic distribution.
6. **Decision Making:**
    - The results of Q-Q plots can influence decisions about whether transformations or adjustments to the model are needed to meet the assumptions of linear regression.