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?**

Ans: From the study and model, I could see the following.

- Some categorical variables which are included in the final model, where quite significant and to understand the variation in the dependent variable, these variables were significantly import both quantitatively and quantitatively (e.g., Variables like year, spring and lightshow are having the highest coefficient, therefore the dependency on these variables is more as compared to numeric variable like windspeed in the final model.

- Moreover, these variables have shown their significance using plots as well.

Examples:

1. More people use the bike during the non-holiday time and clear weather.
2. Fall season has the most bike share count, whereas the Spring has minimum count.
3. Number of counts in 2019 are significantly (33.33%) more than that of 2018.
4. Number of counts in the month of Jun, July, Aug, and Sep are the most.
5. Count is more when the weather is clear.

---

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

Ans:

- Using `drop_first=True` during dummy variable creation is important to avoid multicollinearity and improve the interpretability of the model when using categorical variables in regression analysis, including logistic regression.

- Multicollinearity refers to the presence of strong correlations between independent variables, which can lead to unstable and unreliable estimates of the model coefficients. Therefore, it is important to use drop_first = True.

---

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

Ans: Looking at the pair-plot among the numerical variables**, 'temp'** has the highest correlation with the target variable.

---

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Ans: Following were the assumptions:

- Error terms are normally distributed with the mean zero.
- There should not be multicollinearity exists between the variables.
- Linear relationship should exist between the variable.
- No visible pattern should be observed in residuals.
- No auto correlation.

---

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

Ans: Based on the final model, the top three features are:

- Year
- Spring
- Summer

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans.

Linear regression is a widely used supervised machine learning algorithm used for modelling the relationship between a dependent variable (also called the target) and one or more independent variables (also called features or predictors). The algorithm assumes a linear relationship between the variables and aims to find the best-fitting linear equation that predicts the target variable based on the given features. In simple terms, linear regression tries to find a straight line that best fits the data points.

Mathematically the relationship can be represented with the help of following equation.

Y = mX + c

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Here's a detailed explanation of the linear regression algorithm:

a. **Objective:** The primary goal of linear regression is to find the coefficients (weights) of the linear equation that minimizes the difference between the predicted values and the actual target values.
b. **Linear Equation:** In the case of simple linear regression (one independent variable), the linear equation is given by:
c. **Error Metric:** The algorithm aims to minimize the error between the predicted values and the actual target values. A common error metric used in linear regression is the Mean Squared Error (MSE), which calculates the average squared difference between the predicted and actual values.

d. **Finding Coefficients:** The coefficients are determined through a process called "ordinary least squares" (OLS) regression. The goal is to find the coefficients that minimize the sum of squared residuals, which are the differences between the actual and predicted values.
e. **Training the Model:** To train the linear regression model, you provide it with a labelled dataset where you have the values of the target variable and the corresponding feature values. The algorithm iteratively adjusts the coefficients to minimize the error, usually using optimization techniques like gradient descent.
f. **Prediction:** Once the model is trained, you can use it to make predictions on new, unseen data. You plug in the feature values into the linear equation to calculate the predicted target values.

Assumptions of Linear Regression:

- Linearity: The relationship between the variables is assumed to be linear.
- Independence: The residuals (differences between predicted and actual values) should be independent and not exhibit any pattern.
- Homoscedasticity: The residuals should have constant variance across all levels of the independent variables.
- Normality: The residuals should follow a normal distribution.

Variants of Linear Regression:

- Simple Linear Regression: One dependent variable and one independent variable.
- Multiple Linear Regression: Multiple independent variables.
- Polynomial Regression: Incorporates polynomial terms of the independent variables to model more complex relationships.
- Ridge and Lasso Regression: Introduce regularization to prevent overfitting by adding penalty terms to the coefficients.

Linear regression is a foundational algorithm in statistics and machine learning, and its simplicity and interpretability make it a powerful tool for understanding and predicting relationships between variables.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but are graphically very different. These datasets were introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics.

The quartet consists of four sets of x and y values, each containing 11 data points:

Dataset I:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset II:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26, 4.74

Dataset III:

- x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
- y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset IV:

- x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
- y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91, 6.89

Key points about Anscombe's quartet:

Descriptive Statistics Paradox: Despite having nearly identical means, variances, linear regression coefficients, and correlation coefficients, these datasets look very different when plotted. This highlights the danger of relying solely on summary statistics to understand data.

Graphical Exploration: Anscombe's quartet illustrates the importance of visually exploring data using graphs and plots. Different datasets can have the same basic statistics but convey entirely different relationships when plotted.

Outliers and Influential Points: In Dataset IV, a single outlier significantly affects the linear regression line. This demonstrates the impact of outliers and influential points on regression analysis.

Assumptions of Linearity: Datasets I, II, and III show linear relationships, while Dataset IV does not. This underlines the importance of checking the linearity assumption before applying linear regression.

Statistical Analysis vs. Data Visualization: Anscombe's quartet emphasizes that statistical analysis should be complemented by effective data visualization to gain a more comprehensive understanding of the data.

---

3. What is Pearson's R?

Ans:

Pearson's correlation coefficient, often denoted as \(r\) or Pearson's \(r\), is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is a value between -1 and 1, where:

- r = 1, indicates a perfect positive linear correlation, meaning that as one variable increases, the other variable also increases proportionally.
- r = -1, indicates a perfect negative linear correlation, meaning that as one variable increases, the other variable decreases proportionally.
- r = 0, indicates no linear correlation between the variables.

Key property:

Linearity: It specifically measures linear relationships. It may not capture nonlinear associations between variables.

---

4. . What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans:

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Example:
- If an algorithm is not using feature scaling method, then it can consider the value 1000 meter to be greater than 10 km but that's not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.
- **Difference:**

Normalized scaling	Standardized scaling
Normalized scaling, also known as Min-Max scaling, transforms the data to a specific range, typically between 0 and 1. The formula for normalized scaling is:	Standardized scaling, also known as Z-score scaling, transforms the data to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:
Normalized scaling preserves the original distribution of the data and is suitable when the data does not have	Standardized scaling is suitable when the data has outliers or when you want to ensure that the scaled data has a

outliers that could disproportionately affect the scaling.	standard normal distribution (mean of 0 and standard deviation of 1).
Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans: VIF (Variance Inflation Factor) is a measure used to assess multicollinearity in a regression analysis. It quantifies how much the variance of the estimated regression coefficient is increased due to collinearity among the predictor variables. An infinite VIF value can occur for a specific predictor variable due to perfect multicollinearity. Perfect multicollinearity arises when one predictor variable is a linear combination of other predictor variables in the model. This means that one predictor variable can be exactly predicted from a linear combination of the others, resulting in a mathematical singularity.

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

Ans:

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

- A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted.
- If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.

- The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

- When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale.
- If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.