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)

>> Below conclusions can be drawn about the effect of Categorical Variables on the Target Variable 'cnt' :

- The demand of bikes increased in the year of 2019 compared to 2018.
- Bike demands increase in the Mid-year months and is lower towards the start and end of the year.
- Demand of Bikes is almost similar in all the Weekdays.
- Less number of people use BoomBikes on a Holiday.
- There are no significant Changes observed in demand based on Working/Non-working days.

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

>> **drop_first=True** helps in reducing the extra column created during dummy variable creation thereby reducing correlations created among the dummy variables.

Example: Column 'weathersit' had 3 unique values. By using "**drop_first = True"** , we get 2 new dummy variables (3-1=2)

This helps in reducing the one extra variable and its correlation with other two. However, we can estimate this value using 2 dummy variables created when both has '0' as a flag

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

>> “Temp” and “atemp” has the highest positive correlation with the Target Variable

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

>> Assumptions are validated using Residual Analysis . We use training dataset for the prediction and to calculate the residual values. Then we create a Distribution plot to check if the errors are normally distributed with the mean value at 0. Additionally , a check for Homoscedasticity is carried out , where we compare the Residuals data points on a scatter plot with respect to the mean line at 0. This shows the variance of Residual points across the Model created using Training Dataset.

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

>> Top three features contributing to bike demands are : (Using Best Fit line Equation of the selected model)

Temp > 0.5173

2019 (Year) > 0.2326

weather_light_snow_rain (Weather Condition) > 0.2819

General Subjective Questions

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

>>Linear regression is a Supervised machine learning algorithm used for predicting a continuous target variable based on one or more input features.

- The main objective of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted values and the actual values of the target variable.
- It represents the relationship between the input features (often denoted as X) and the target variable (often denoted as y) using a linear equation of the form: y = mx + b. Here, 'm' represents the slope of the line, 'b' is the y-intercept, and 'x' is the input feature.
- Linear regression can handle multiple input features by using a multi-dimensional equation: y = b + m1*x1 + m2*x2 + ... + mn*xn. Here, 'n' represents the number of input features, and 'xi' are the individual feature values.
- **Cost Function:** The algorithm uses a cost function, often the Mean Squared Error (MSE), to measure the average squared difference between the predicted values and the actual values. The goal is to minimize this cost function to achieve the best-fitting line.
- **Assumptions**: Linear regression assumes that there is a linear relationship between the input features and the target variable. It also assumes that the errors are normally distributed and have constant variance.
- **Evaluation:** After training, the model's performance is evaluated on a separate validation or test dataset using metrics like R-squared to assess how well it generalizes to new data

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

>> Anscombe's quartet is a set of four datasets that have the same summary statistics but exhibit significantly different distributions and patterns when plotted. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data and not solely relying on summary statistics. Let's explore each dataset in detail:

1. **Dataset I:**
   - x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   - y-values: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82
- This dataset forms an approximately linear relationship between x and y. When plotted, it appears to have a clear linear trend, well-suited for linear regression analysis.

2. **Dataset II:**
   - x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   - y-values: 9.14, 8.14, 8.74, 8.77, 9.26, 8.1, 6.13, 3.1, 9.13, 7.26
- Similar to Dataset I, this dataset also forms a linear relationship between x and y, but with a slight deviation from the trend observed in Dataset I.

3. **Dataset III:**
   - x-values: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7
   - y-values: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42
- Dataset III appears to have a curved relationship when plotted. It is better described by a quadratic function rather than a linear one.

4. **Dataset IV:**
   - x-values: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8
   - y-values: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.5, 5.56, 7.91
- Dataset IV has an outlier at one end that significantly impacts the linear regression line and the correlation coefficient.

- The key takeaway from Anscombe's quartet is that visualizing data is crucial for understanding its underlying patterns and relationships. While summary statistics like mean, variance, and correlation can provide useful information, they might not reveal the whole picture. Different datasets can have the same summary statistics but behave very differently when visualized, leading to different conclusions and insights. Therefore, data visualization remains an essential tool in the data analysis process.

- 3. What is Pearson’s R? (3 marks)

>>Person's R, also known as Pearson correlation coefficient or correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It is denoted by the symbol "r.“

- The Pearson correlation coefficient takes values between -1 and +1, where:

  r = +1 indicates a perfect positive linear relationship between the variables.
  r = -1 indicates a perfect negative linear relationship between the variables.
  r = 0 indicates no linear relationship between the variables.

- The formula for Pearson correlation coefficient (r) is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

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

>> Scaling is a preprocessing technique used in data preparation for machine learning and statistical analysis. It involves transforming the features or variables of a dataset to a specific range or distribution. The primary goal of scaling is to bring all the features to a common scale, so they contribute equally to the model's training process or analysis.

- Reasons for performing scaling:

1. **Equalizing Magnitudes:** Different features in a dataset can have different scales. For example, one feature may have values ranging from 0 to 100, while another has values ranging from 0 to 10000. Scaling ensures that all features have a similar scale, preventing any one feature from dominating the others during modeling.
2. **Feature Importance:** Feature importance analysis becomes more meaningful when all features are on the same scale, as it allows fair comparisons of their contributions to the model's predictions.

- The difference between normalized scaling and standardized scaling:

Normalized Scaling (Min-Max Scaling) :

- It transforms the data to a specified range, usually between 0 and 1.
- Normalization is sensitive to outliers since it scales the data based on the minimum and maximum values.
- Formula :

$$X_{normalized} = \frac{(X - X_{minimum})}{(X_{maximum} - X_{minimum})}$$

Standardized Scaling (Z-Score Scaling/ Standardization):

- It transforms the data to have zero mean and unit variance.
- Standardization is less sensitive to outliers compared to normalization since it uses the mean and standard deviation.
- Formula:

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean
σ = Standard Deviation

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

>> VIF stands for Variance Inflation Factor, and it is used to assess multicollinearity between predictor variables in regression models. Multicollinearity occurs when two or more predictor variables are highly correlated, which can lead to unstable coefficient estimates and difficulties in interpreting the model.

- When the VIF value shows as infinite, it indicates that there is perfect multicollinearity between at least one pair of predictor variables. Perfect multicollinearity occurs when one predictor variable can be perfectly predicted by a linear combination of other predictor variables in the model.

- VIF Formula: $$VIF_i = \frac{1}{1 - R_i^2}$$

VIFi is the Variance Inflation Factor for the ith predictor variable.

*Ri*2 is the coefficient of determination (R-squared) obtained when regressing the ith predictor variable against all other predictor variables in the model.

- If the R-squared value obtained from this regression is equal to 1, it means that the ith predictor variable is a linear combination of other predictor variables, leading to a division by zero in the VIF formula, which results in an infinite VIF value.

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

>> A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a specific probability distribution, such as the normal distribution. It helps to compare the quantiles of the dataset to the quantiles of the theoretical distribution to check for similarity.

- Use and Importance of Q-Q plot in Linear Regression:

1. **Checking Normality Assumption:** In linear regression, one of the key assumptions is that the residuals (the differences between the observed values and the predicted values) are normally distributed. A Q-Q plot is a useful tool to visually assess whether the residuals follow a normal distribution.
2. **Detecting Departures from Normality:** If the data points in the Q-Q plot deviate significantly from the diagonal line, it suggests that the residuals might not be normally distributed. This departure from normality can impact the accuracy and reliability of the linear regression model's predictions.
3. **Validating Model Assumptions:** Assessing the normality of residuals is crucial for validating the assumptions of linear regression. If the residuals are not normally distributed, it might indicate that the model's assumptions are not met, and further investigation or modifications to the model might be necessary.
4. **Model Improvement:** If the Q-Q plot reveals substantial deviations from normality, it might be an indication of potential outliers or heteroscedasticity (unequal variance). Identifying and addressing these issues can lead to improved model performance.

- A Q-Q plot is a powerful tool to visually assess the normality assumption and check for the appropriateness of linear regression models. It helps data analysts and researchers gain insights into the distribution of residuals and make informed decisions about the validity and reliability of the linear regression model.