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)
   Seasons, weather situations and year has considerable impact on the target variable as noticed in the box plot analysis where medians are considerably different for various categories. The other categorical variables such as holiday, individual_days and working day do not have as much of an impact on the target variable.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   To keep the model simple. Since *n* different variables can be represented by *n-1* columns we use drop_first=True.
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   ***atemp* has the highest correlation with the target variable with a value of 0.646475.**
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   a. **By making sure the R2 value is reasonably high thereby indicating that the model is account a good percentage of the variance of the data.**
   b. **By making sure the adjusted R2 is close/similar to the R2 value to make sure the R2 is not caused by simply adding more and more features.**
   c. **Making sure the VIF is small for all the used feature variables indicating low multicollinearity between the independent variables.**
   d. **Making sure the R2 score on the test set is close to the R2 score on the training set to make sure there is no overfitting.**

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

   i. **Temp -> positive correlated.**
   ii. **Rain -> negatively correlated.**
   iii. **Year -> positive correlated.**

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Linear regression is a fundamental and widely used statistical method for modeling the relationship between a dependent variable and one or more independent variables. It's used to predict a continuous outcome based on the input features. The goal of linear

regression is to find the best-fitting linear equation that describes the relationship between the variables.

The main assumptions for linear regression include:

1. **Linearity: The relationship between the independent variables (X) and the dependent variable (Y) should be approximately linear. This means that the change in Y for a unit change in X should be constant across all levels of X.**
2. **Independence: The residuals (the differences between observed and predicted Y values) should be independent of each other. This assumption is often violated in time series data or when the order of data points matters.**
3. **Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables. In other words, the spread of residuals should be roughly the same across the range of predicted values.**
4. **Normality of Residuals: The residuals should follow a normal distribution. This assumption is important for hypothesis testing and constructing confidence intervals.**
5. **No or Little Multicollinearity: Independent variables should not be highly correlated with each other. High multicollinearity can make it difficult to determine the individual effect of each independent variable on the dependent variable.**

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

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, yet they look quite different when graphed. It emphasizes the importance of visualizing data and not relying solely on summary statistics.

Each dataset in Anscombe's quartet consists of 11 data points with two variables: x and y. Here's a brief description of the four datasets:

1. **Dataset I: Linear Relationship**

This dataset shows a perfect linear relationship between x and y. The linear regression line fits the data well, and the correlation coefficient is close to 1.

It is intended to highlight that even when the relationship appears strong, it's essential to visualize the data to confirm the accuracy of the model.

2. **Dataset II: Non-Linear Relationship**

This dataset exhibits a clear curvilinear relationship between x and y. The linear regression line is a poor fit for this data, yet the summary statistics might still suggest a linear relationship.

The dataset illustrates the significance of examining the data visually to identify nonlinear patterns that might be missed by relying solely on numerical metrics.

3. **Dataset III: Outlier Impact**

Most of the data points in this dataset lie along a linear relationship, except for one significant outlier. This outlier heavily influences the linear regression line.

Anscombe's third dataset emphasizes the importance of identifying and addressing outliers, as they can significantly impact statistical summaries and modeling.

4. Dataset IV: Influential Point

The fourth dataset consists of mostly linear data points except for one influential point that creates a different regression line when included or excluded.

This dataset illustrates the concept of influential points, which can substantially alter the results of regression analysis.

The key takeaway from Anscombe's quartet is that relying solely on numerical summary statistics can lead to incorrect assumptions about the relationships within the data. Visualization is essential for revealing patterns, trends, outliers, and potential issues that might not be evident from summary metrics alone. This concept highlights the need to use both quantitative analysis and visualizations to gain a comprehensive understanding of the data and make informed decisions in statistical analysis and modelling.

3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient, often denoted as Pearson's "r," is a statistical measure that quantifies the linear relationship between two continuous variables. It measures the strength and direction of the linear association between the two variables. Pearson's correlation coefficient ranges between -1 and +1, where:

A positive value close to +1 indicates a strong positive linear relationship, meaning that as one variable increases, the other tends to increase as well.

A negative value close to -1 indicates a strong negative linear relationship, meaning that as one variable increases, the other tends to decrease.

A value close to 0 indicates a weak or no linear relationship between the variables.

Pearson's correlation coefficient has some assumptions:

Linearity: It measures only linear relationships between variables. Nonlinear relationships may not be accurately captured by Pearson's r.

Homoscedasticity: The variability of the data should be roughly constant across all levels of the variables.

Independence: The data points should be independent of each other.

Normally Distributed Data: The variables should follow a normal distribution.

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

Scaling is an important preprocessing step to ensure that the features are on a similar scale, which helps improve the performance and stability of various machine learning algorithms. For example, if 2 variables are on very different scale, their coefficients would also be drastically different and may not be truly representing the variable's impact on the dependant variable comparatively. The choice between normalized and standardized scaling depends on the nature of the data and the requirements of the algorithm being used.

<u>Normalized Scaling (Min-Max Scaling):</u>

Normalized scaling transforms the data to a specific range, usually between 0 and 1.

This method is sensitive to outliers, as outliers can disproportionately affect the scaling of the data.

<u>Standardized Scaling (Standardization):</u>

Standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1. This method is robust to outliers, as outliers have a limited impact on the mean and standard deviation.

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

The formula for VIF is $1/(1-R^2)$. When R^2 equals 1, VIF becomes infinite. This happens when the variable in question can be completely explained by all other independent variables, i.e., perfect multicollinearity.

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 (Quantile-Quantile) plot is a graphical tool used to assess whether a given dataset follows a particular distribution. It compares the quantiles of the observed data against the quantiles of a distribution, typically a normal distribution. This plot helps in visually understanding the degree of similarity between the observed data distribution and the theoretical distribution. A Q-Q plot is a valuable diagnostic tool in linear regression analysis that assess whether the residuals of the model adhere to the normality assumption, which is essential for making accurate and valid statistical inferences.