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)

   Ans:

   From the univariate analysis:

   1) There is no variation in demand for weekdays or weekends.

   2)The demand for bikes is lower at the start of the year, increases, and reaches its peak in the middle of the year, reducing again towards the end of the year.

   3) Bike demand is lower in winter compared to other seasons.

   From the bivariate analysis:

   4) It appears that temp and atemp have a similar relationship with the count. Therefore, atemp is being removed from the analysis.

   From the multivariate analysis:

   5) The count has a positive correlation with temperature, but a negative correlation with wind speed. Other variables can be ignored due to their very low correlation values.

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

Ans:

drop_first=True serves to make the reference category clear in the interpretation of the model and avoids statistical issue of Multicollinearity and help in Reducing Complexity.

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

Ans:

"temp" variable has highest 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)

   Ans:

   1> R-Squared on test is 0.826. The value of 82.6%, therefore, suggests that this model is relatively effective - but not perfect - at predicting bike rental demand based on the given factors. It means there is still 17.4% of the variance that is unexplained by the model, which could be due to other variables not included in the model due to randomness or errors in the data.

   2> The relationship between the independent variables and the dependent variable is linear. Inspected using scatter plots of the predicted values against actual values and looked for linear patterns.

   3> Normality of Residuals : I plotted a histogram with residuals resulting in a bell-shaped curve around zero, indicating normally distributed residuals. I used a Q-Q plot to compare residual distribution with a normal distribution. Points along the diagonal line suggest normal distribution of residuals.

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:

A) temp (Temperature):
Coefficient: 3892.2865
P-value: Significantly low ($p < 0.001$)
Interpretation: Temperature is the most significant predictor of bike demand. The positive coefficient suggests that higher temperatures are associated with increased bike rentals, making it the top contributing feature.
B) yr_1 (Year: 2019):

Coefficient: 1917.1324
P-value: Significantly low ($p < 0.001$)
Interpretation: The year 2019 (encoded as yr_1) significantly impacts bike rental demand, indicating a notable increase in demand from 2018 to 2019. This reflects the growing popularity or expansion of the bike-sharing service.
C) weathersit_Light_Snow_Rain (Weather Situation: Light Snow or Rain):

Coefficient: -1434.9070

P-value: Significantly low ($p < 0.001$)

Interpretation: This feature has a substantial negative impact on bike demand. It suggests that light snow or rain significantly reduces the number of bike rentals, making it a critical factor affecting demand, albeit in a negative way.

General Subjective Questions

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

   Ans:

   Linear regression is a statistical algorithm used to predict a dependent variable based on independent variables. It establishes a relationship between a dependent variable (Y) and one or more independent variables (X) using a best fit straight line known as the regression line. It uses the method of least squares to minimize the errors and finds the best fit line. It's mostly used for forecasting, time series modelling and finding the cause-effect relationship between variables.

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

Anscombe's quartet comprises four distinct datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. This quartet was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. It serves as a powerful reminder that statistical summaries can sometimes be misleading, and visualizing data is a crucial step in data analysis.

Key Properties:

When analyzed statistically, the four datasets in Anscombe's quartet have the same or very similar values for the following properties:

➔Mean of both x and y variables.

➔Variance of x and y variables.

➔Correlation between x and y variables.

➔Linear regression line ($y = mx + c$) that best fits the data, including the slope (m) and y-intercept (c).

➔Coefficient of determination (R^2), which measures the proportion of the variance in the dependent variable that is predictable from the independent variable.

Despite these similarities in statistical summaries, the datasets have very different distributions and appear distinct when graphed. Each set illustrates a different case or problem in regression analysis.

The Four Datasets:

Dataset I: Follows a pattern that appears to be a simple linear relationship, corresponding closely to the assumptions of linear regression.
Dataset II: Demonstrates a curve (quadratic relationship) rather than a linear relationship. Linear regression is not appropriate here, but the statistical properties mirror those of the first dataset.
Dataset III: Contains an outlier that influences the slope of the regression line. Without this outlier, the dataset would have a very different linear relationship.
Dataset IV: Shows a case where one outlier is driving the entire correlation. Without the outlier, there would be little to no correlation between x and y variables.
Importance:
Anscombe's quartet is a foundational lesson in data analysis, emphasizing the following points:
Visualize Data: Always graph your data before starting the analysis. Visual inspection can reveal data properties and structures that summary statistics cannot.
Beware of Outliers: Outliers can significantly affect the results of your analysis, leading to misleading conclusions.
Understand the Data: Knowing the underlying assumptions of statistical methods is crucial. For instance, linear regression assumes a linear relationship between variables, which might not always hold true.
The quartet serves as a cautionary tale, reminding analysts and statisticians to look beyond numerical summaries and to use visualization tools as an integral part of their data analysis workflow.

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

Pearson's R, also known as the Pearson correlation coefficient or Pearson's product-moment correlation coefficient (PPMCC), is a measure of the linear correlation between two variables X and Y. It gives a value between +1 and -1 inclusive, where:
+1 indicates a perfect positive linear relationship,
-1 indicates a perfect negative linear relationship, and
0 indicates no linear correlation between the variables.

Formula:

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

Scaling is a data preprocessing technique used to standardize the range of independent variables or features of data. In the context of machine learning and data analysis, scaling can be crucial because the range of data values can vary widely. If one feature has a range from 0 to 1, while another has a range from 0 to 1000, the model might unduly weigh the latter feature more heavily than the former, potentially leading to inaccurate predictions or analyses.

Why is Scaling Performed?

Improves Algorithm Performance: Many machine learning algorithms that use distance calculations (e.g., k-nearest neighbors, k-means clustering) or optimization (e.g., gradient descent in neural networks, support vector machines) perform better when the data is scaled.

Increases Computational Efficiency: Algorithms converge faster when features are on similar scales, particularly in optimization algorithms.

Prevents Skewed Influence: Prevents features with larger scales from overshadowing those with smaller scales in models where feature weighting is important.

Required by Some Models: Some algorithms, like Support Vector Machines (SVM) and Principal Component Analysis (PCA), explicitly require scaling for correct execution.

Normalized Scaling vs. Standardized Scaling:

1. **Normalized Scaling (Min-Max Scaling):**

Normalization adjusts the data values to a specific scale, typically 0 to 1, without distorting differences in the ranges of values or losing information. It is performed using the formula:

Normalization is useful when you need to bound your values between two numbers, e.g., 0 and 1.

2. **Standardized Scaling (Z-score Normalization):**

Standardization transforms the data to have a mean of zero and a standard deviation of one. The formula for standardization is:

σ is the standard deviation of the feature values. Standardization does not bound values to a specific range, which may be a problem for certain algorithms (e.g., neural networks often expect an input value bounded between 0 and 1).

Choosing between normalization and standardization depends on the specific requirements of the model, the data distribution, and the presence of outliers.

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

The Variance Inflation Factor (VIF) is a measure used to detect the presence of multicollinearity among independent variables in a regression model. Multicollinearity occurs when two or more independent variables are highly correlated, meaning they contain similar information about the variance. VIF quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated.

A VIF value of 1 indicates no correlation between a given independent variable and any other independent variables. VIF values between 1 and 5 suggest moderate correlation, but they are often considered acceptable. Values greater than 5 or 10 indicate potentially problematic levels of multicollinearity, depending on the context and the source.

Reasons for an Infinite VIF:

An infinite VIF (or a VIF value that is exceedingly large) can occur for several reasons, primarily related to perfect or near-perfect multicollinearity:

Perfect Multicollinearity: This happens when an independent variable is an exact linear combination of one or more independent variables. For example, if one variable is the sum or difference of two others in your dataset, it will lead to a situation where the VIF for at least one of those variables could be infinite.

Redundant Variables: Including variables in the model that are redundant (duplicate information in another format) can cause infinite VIFs. A common scenario is when dummy variables are created from a categorical variable but all categories are included without dropping one as a reference. This creates a perfect linear relationship among the dummy variables due to the "dummy variable trap."

Highly Correlated Variables: Even if not perfectly linear, variables that are highly correlated can produce very large VIF values, approaching infinity as the correlation approaches perfect linearity.

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

The Q-Q (quantile-quantile) plot is a graphical tool primarily used to assess similarities between two probability distributions by depicting their quantiles. In case these distributions are similar, the points on the Q-Q plot will lie approximately on the line y = x. In the field of linear regression, a Q-Q plot is routinely used to compare the distribution of residuals or errors in the model with a theoretical normal distribution. This process is instrumental for assessing the presumption of normality in linear regression models.

Implementation of the Q-Q plot is pinnacle in gauging the normality of residuals, which are differences between observed and predicted values in linear regression. The Q-Q plot provides a visual demonstration, portraying how closely the model's residuals match the anticipated distribution (generally normal distribution). This plot allows for an easy detection of deviations from normality. The points should form a straight line if the residuals are normally distributed. However, any curvature indicates non-normality due to skewness or kurtosis. Additionally, significant outliers could be recognized as points diverging considerably from this straight line.

Moreover, Q-Q plots are fundamental for the validity of statistical tests. Many of these tests assume a normality of residuals, including those used in calculating the significance of regression coefficients. Any deviation from this assumption could result in inaccurate inferences. Further, a Q-Q plot is a non-parametric method for validating the normality presumption and can suggest necessary transformations of variables or consideration of an alternate modeling approach. It also helps to identify outliers, which may unduly impact the regression model. Ensuring a normal distribution of residuals can enhance the accuracy and interpretability of the model.