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)

- The booking count is higher in Fall season
- The booking count is highest when the weather is Clear. The booking count decreases when it is raining
- Booking count is higher for 2019 compared to 2018. This could be indicating that the count increases over time.
- The counts are slightly higher on working days compared to holidays/weekdays. This could be because people rent bikes to go to office during workdays.
- There is no clear trend on count base on just the day of the week
- The count is increasing till the second quarter of the year and decreases on the last quarter.

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

If a categorical variable have k values, using drop_first=True while creating dummy variables creates only k-1 dummy variables.If you include all k dummy variables in the regression model without dropping one, it creates perfect multicollinearity. This is known as the "dummy variable trap".

For example, if you have a categorical variable "Color" with three levels (Red, Green, Blue), creating dummy variables would result in two binary variables: "Color_Red" and "Color_Green". Including both of these in the model would be problematic because if both dummy variables are 0, it automatically means that the third category "Blue" is present. So, one of the dummy variables should be omitted to avoid multicollinearity.

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

Temperature (temp variable) has the highest correlation with cnt. This can be seen from the correlation matrix also.

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

1. Normality of error terms: This is verified by plotting the residuals using a distplot. We can see that the residuals form a normal distribution.
2. Multicollinearity check: This is done by calculating the VIF using variance_inflation_factor function from statsmodel library

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

1. Temperature (Coeff = 0.477737): As temperature increases, the model shows the rentals increases.

2. Weather (weathersit_LightFalls coeff = -0.285031) If the weather is raining, the number of rentals drops.
3. Year (coeff = 0.234132) If the year increases, the rental seems to be increasing. We generally see the renting increases.

General Subjective Questions

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

Linear regression supervised machine learning algorithm which is used to analyse and predict the relationship between one or more independent variables and a dependent variable. The goal of the algorithm is to find the best fit line through the data points which minimizes the difference between predicted and actual values.

Linear regression is of two types:

1. Simple linear regression: In simple linear regression there is only one independent variable (X) and only one dependent variable Y. The relationship between both are represented as:

   $Y = \beta0 + \beta1 * X$

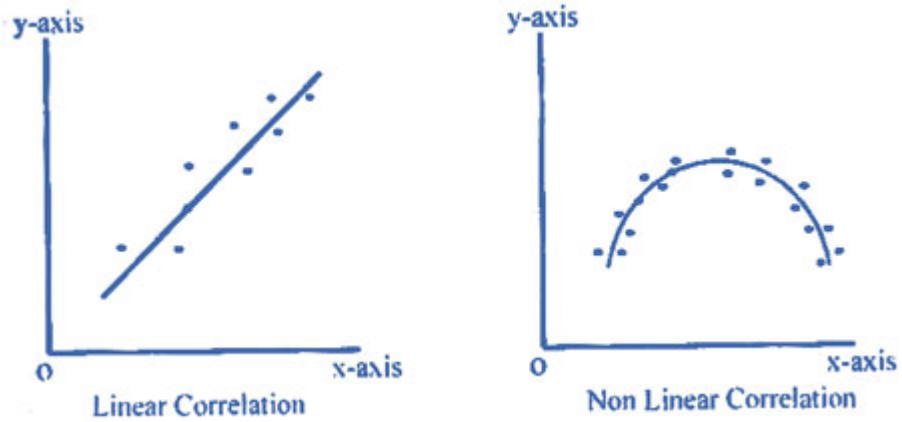
   where β0 is the intercept, β1 is the slope (also known as the coefficient), and X is the input variable.
2. Multiple linear regression: There is more than one independent variable and one dependent variable in Multiple linear regression. The relationship is expressed as:

   $Y = \beta0 + \beta1 * X1 + \beta2 * X2 + ... + \beta n * Xn$

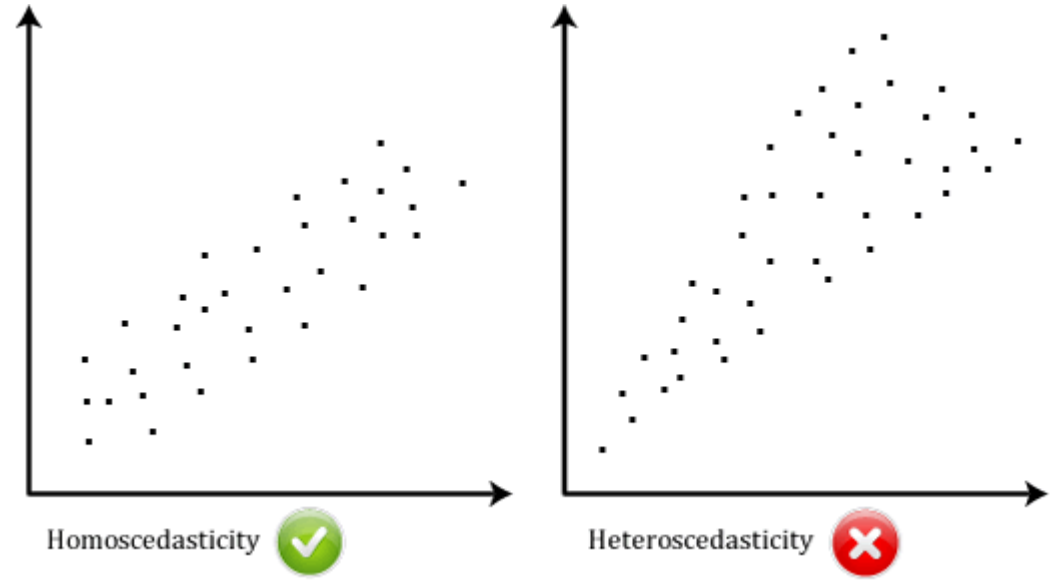
   where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, and β0, β1, β2, ..., βn are the corresponding coefficients (slopes) for each independent variable.

There are few assumptions while doing linear regression:

1. Linearity: Dependent variable Y is linearly dependent on independent variables.



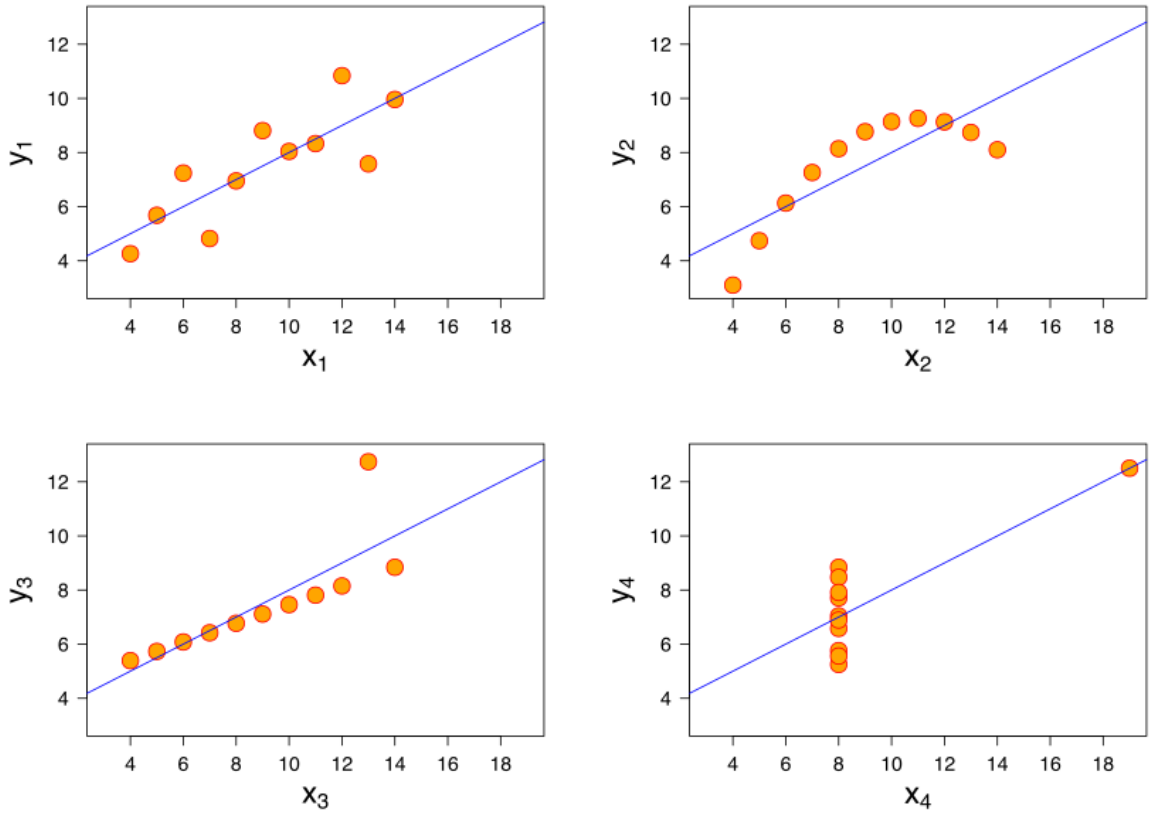
2. Homoscedasticity or constant variance: The variance of the error terms should be constant.



3. No multicollinearity: The independent variables should not be correlated to each other. This can be checked using the Variance Inflation Factor or correlation matrix.
4. Normality: Error terms must be normally distributed
5. No Autocorrelation: Autocorrelation occurs when the residuals exhibit a pattern and are not independent of each other. In linear regression, the residuals should be independent of each other and show no correlation.

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

Anscombe's quartet comprises 4 datasets that have nearly identical simple descriptive statistics, but look completely different when graphed.



Data Points (X, Y):

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

Properties

1. Mean of X and Y: For all four datasets, the mean of X is approximately 9, and the mean of Y is approximately 7.5.
2. Variance of X and Y: For all four datasets, the variance of X is approximately 11, and the variance of Y is approximately 4.12.

3. Correlation coefficient (r) between X and Y: For all four datasets, the correlation coefficient between X and Y is approximately 0.816.

When plotted:

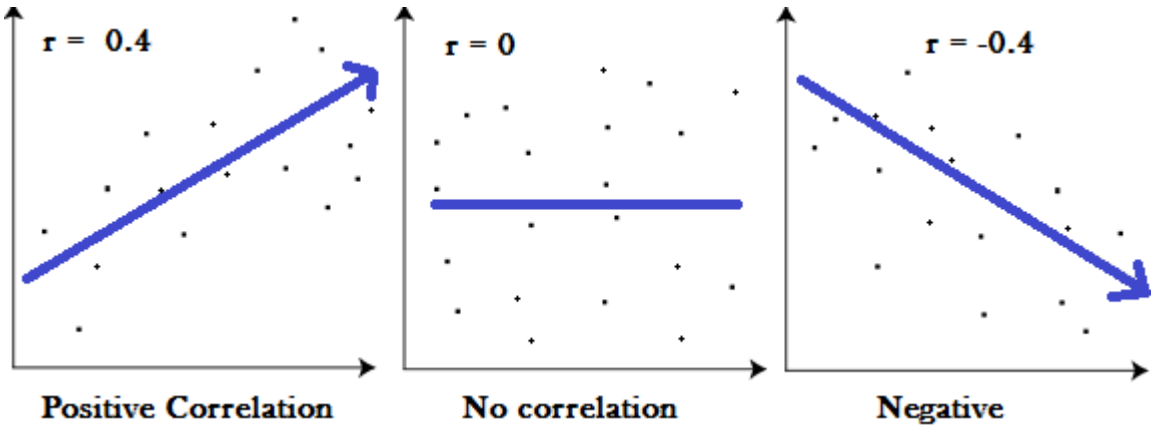
- Dataset I and Dataset II form an approximately linear relationship.
- Dataset III has an outlier that significantly influences the linear regression line.
- Dataset IV seems to have a quadratic relationship between X and Y.

The main lesson from Anscombe's quartet is that relying solely on summary statistics can be misleading.

3. What is Pearson's R? (3 marks)
Pearson's R, or Pearson's correlation coefficient is a statistical measure used to quantify the relationship between two continuous variables. It is used to assess the strength and direction of linear association between two continuous variables.

Its value ranges from -1 to 1:

- r = 1 => Perfect correlation. When one variable increases the other increases proportionately.
- r = -1 => Perfect negative correlation. When one variable increases the other decreases proportionately.
- r = 0 => No correlation between two variables.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Scaling is the process of transforming the values of different variables to a common scale. This is done to ensure that all variables have comparable ranges and to prevent certain variables from dominating the analysis.

Scaling is performed to achieve:

1. Better convergence: Scaling helps in faster convergence and more stable training. Many machine learning algorithms are optimized for variables in similar scales.

2. Improved performance: Some machine learning algorithms, such as distance-based algorithms (e.g., k-nearest neighbors) and gradient-based methods (e.g., gradient descent), are sensitive to the scale of features. Scaling can improve the performance of these algorithms.
3. Feature importance: Scaling ensures that all features are equally treated in terms of importance during the analysis. Without scaling, features with larger ranges might dominate the results.

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 assess multicollinearity in a multiple linear regression model. It quantifies how much the variance of the estimated coefficients increases due to collinearity among the independent variables.

The formula for VIF is:

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where R2i is the coefficient of determination of the regression equation in step one, with Xi on the left hand side, and all other predictor variables (all the other X variables) on the right hand side.

The above formula would reach infinity if denominator is closing to zero. This would happen if Ri^2 is high. Ri^2 is high when the ith variable have high correlation with other variables in the model.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used to evaluate how well a dataset follows a specific theoretical distribution, like the normal distribution. It helps in understanding if the data behaves as expected and aids in identifying deviations from the assumed distribution.

In linear regression, Q-Q plots are essential for various purposes:

1. Assessing Normality: Linear regression assumes that the residuals (the differences between actual and predicted values) have a normal distribution. A Q-Q plot of the residuals can verify if this assumption holds true. If the points on the Q-Q plot significantly depart from a straight line, it suggests that the residuals are not normally distributed, which may violate the model's assumptions.
2. Detecting Outliers: Outliers in the data can have a significant impact on regression results. Q-Q plots help identify outliers by highlighting data points that substantially

deviate from the expected pattern based on the assumed distribution. These outliers can be detected by points that fall far away from the straight line on the Q-Q plot.

3. Assessing Residuals: A Q-Q plot of the residuals allows for visual inspection of their symmetry around zero and adherence to a normal distribution. If the residuals deviate from normality, it might be necessary to transform the dependent variable or employ other techniques to enhance the model's performance.
4. Model Validity: By examining the normality of residuals through a Q-Q plot, the validity of the linear regression model and the accuracy of associated statistical tests (e.g., hypothesis testing, confidence intervals) can be ensured.