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: The analysis on the categorical variables was done through bar and box plots of which below were found:

- Clear weather attracted more bookings
- Year on year increase in bike bookings can be predicted and seen
- Apparently on weekdays, Wednesday, Thursday and Saturday more bookings were identified.
- In the season of "fall" the bookings were more comparatively other seasons.
- In a year uniformly bookings were seen with peaking during September month.
- Almost equal bookings were visible on working and non-working days.

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

Ans: During the dummification process to perform linear regression, to avoid an extra column from getting created we use drop_first=True.

For example, if I have a column named "color" holding only values like "Red","Green" & "Yellow". I can simply have two color columns for Red & Green, skipping Yellow as if none is red and none is green i.e. it implies it's yellow.

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

Ans: The variable "temp" has the highest correlation with the target variable count ("cnt").

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

Ans: I have validated the assumption of Linear Regression Model based on below 5 assumptions:

- Normality of error terms
    - Error terms should be normally distributed
- Multicollinearity check
    - There should be insignificant multicollinearity among variables.
- Linear relationship validation
    - Linearity should be visible among variables
- Homoscedasticity
    - There should be no visible pattern in residual values.

- Independence of residuals
    - No autocorrelation

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

Ans: Based on the analysis the top 3 items explaining or contributing towards the model is:

- Winter
- Temp
- Sep

General Subjective Questions

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

Ans: Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the independent variables (also known as predictors, features, or regressors) and the dependent variable (also known as the target variable or response).

In linear regression, we model the relationship between the independent variable's 'X' and the dependent variable 'y' using a linear equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \Omega$$

- y is the dependent variable.
- $X_1, X_2, \ldots X_n$ are the independent variables.
- $\beta_0, \beta_1, \beta_2 \ldots \beta_n$ are the coefficients (also known as weights or parameters) that represent the slope of the relationship between each independent variable and the dependent variable.
- Ω is the error term, representing the difference between the observed values and the values predicted by the model.

The objective of linear regression is to estimate the coefficients $\beta_0, \beta_1, \beta_2 \ldots \beta_n$ that minimize the sum of squared differences between the observed and predicted values of the dependent variable.

Linear regression estimates the coefficients using the method of ordinary least squares (OLS). OLS finds the values of the coefficients that minimize the sum of the squared residuals (the vertical distance between the observed and predicted values) for the given set of data points.

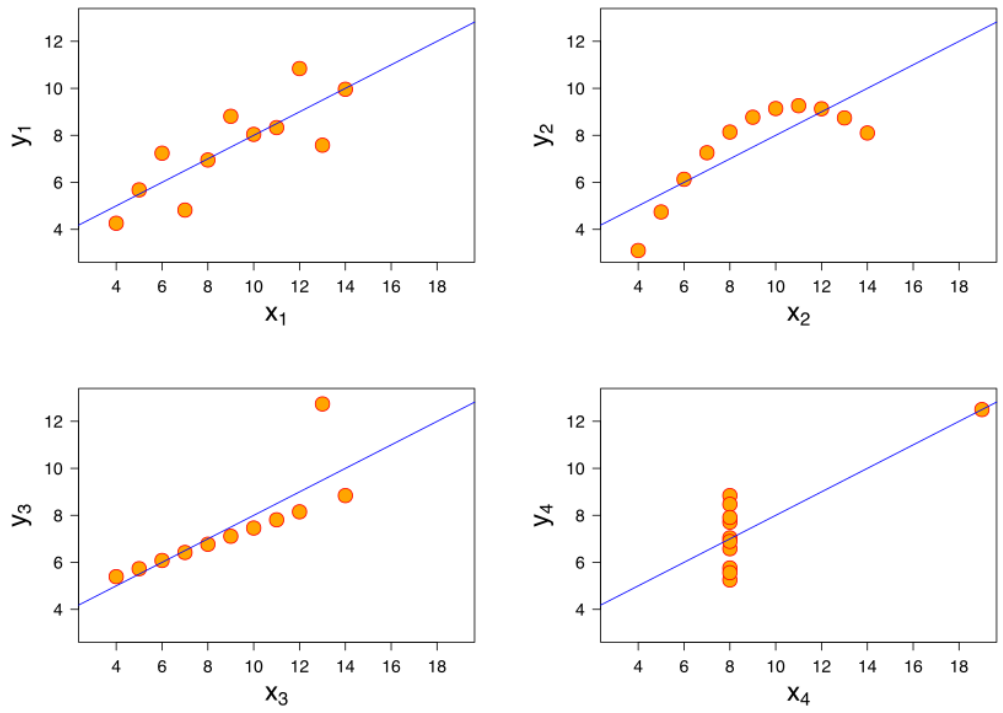
After fitting the linear regression model, it's essential to evaluate its performance. Common metrics for evaluating linear regression models include:

- R-squared
- Adjusted R-squared

- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Residual Analysis

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

Ans: **Anscombe's quartet** comprises four data sets that have nearly identical simple descriptive statistics yet are very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where *y* could be modelled as gaussian with mean linearly dependent on *x*.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust generation would have been called for). The

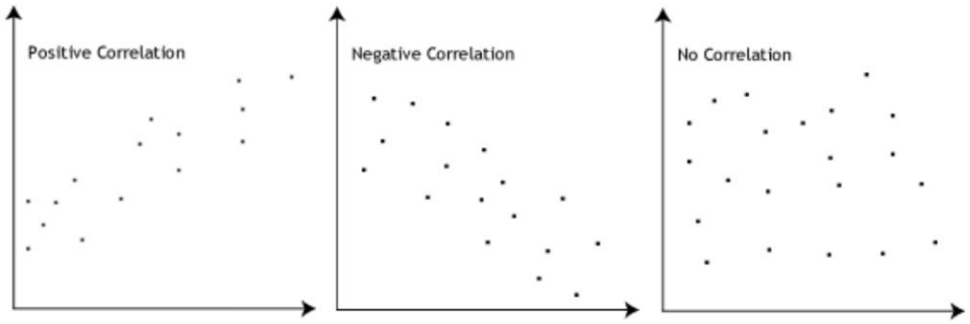
calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-level point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

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

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with the high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

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 3000 meter (about 1.86 mi) to be greater than 5 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.

Some of the differences between Normalized Scaling and Standardized Scaling are:

Normalized Scaling	Standardized Scaling
Minimum and maximum value of features are used for scaling.	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
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? (3 marks)

Ans: VIF goes to infinite if there is perfect correlation. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which leads to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

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.