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 categorical variables in the dataset are: *"season", "workingday", "year", "weathersit", "weekday", "year", "holiday" and "month"*
    - More number of bikes were rented in year 2019 as compared to 2018
    - Overall, bikes rented during working days is almost the same as on non-working days
    - Overall, there's not much of a significant pattern over rented bikes count with the weekdays
    - Bike were rented less in spring
    - Bikes were rented more during the clear weather/ partly cloudy weather situation
    - Bikes were rented more during the fall season followed by the summer season

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

    Dummy variables are created to cover the range of categorical values with values 0 and 1, 0 for absence and 1 for presence of the respective category.
    With a categorical variable with 'n' levels, the idea behind dummy variable creation is to create 'n-1' variables that represent the levels, and remove the first level (base category) using *"drop_first=True".*
    It is important to use to avoid multi-collinearity being added into the model which may arise if all dummy variables are included.

    Example: If we have a variable '*blood type'* with four levels *'A','B','AB','O'*, we do not need to define four levels. We could drop one level, for instance *'A'*, so in case when all the other three variables are denoted by 0 it means that the blood type is *'A'*.

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

    Looking at the pair-plot among the numerical variables, '*temperature'* has the highest correlation with the target variable '*count'*.

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

    Linear relationship between independent and dependent variables: We can observe that there is linearity between dependent and independent variables using the pairplot.

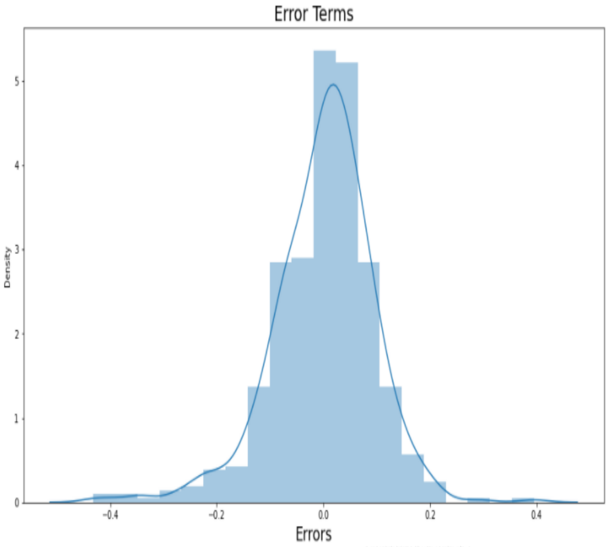
    - We must first ensure that the model is accurate before making predictions. To do so, we must conduct a residual analysis of the error terms. Error terms/residuals is the difference between the observed and predicted values.

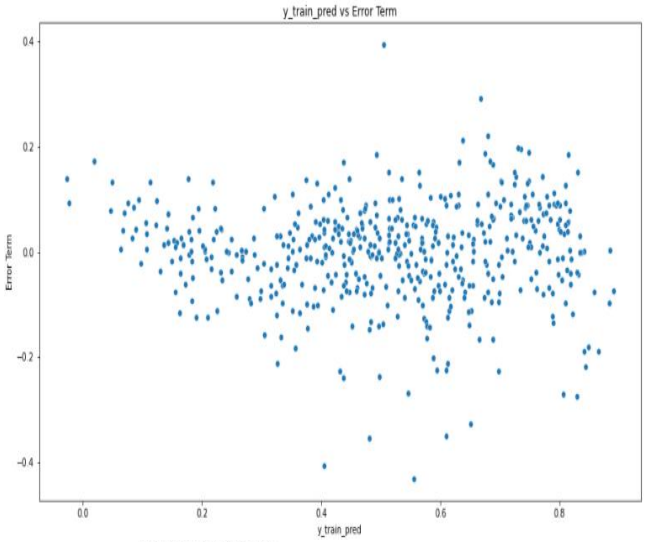
    - We observe by seeing the histogram and distribution plot that shows Error Terms are normally distributed with mean zero, otherwise there could be repercussions like p-values which are obtained during the hypothesis test to determine the significance of the coefficients become unreliable.

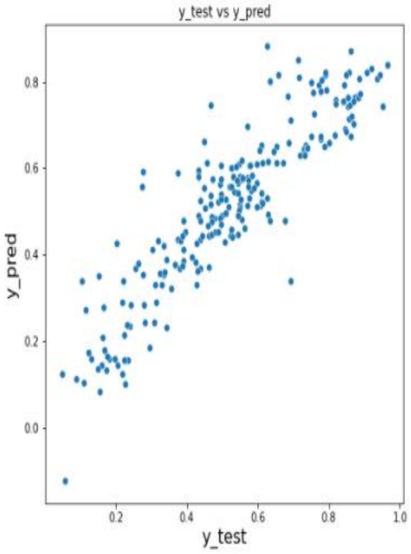
- Error terms have approximately same variance, so this assumes homoscedasticity which is the violation of heteroscadasticity that occurs when the variance is not same across the error terms.

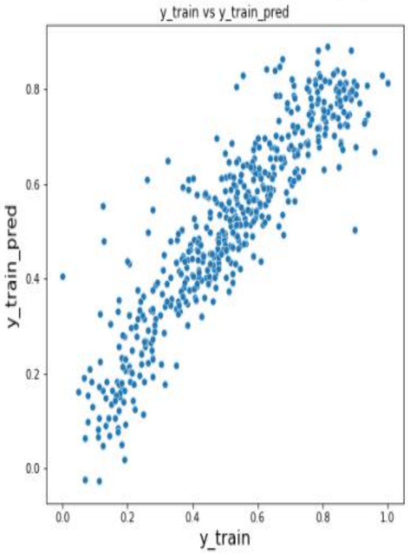
- We observe that residuals are independent of each other as there is no specific pattern between the residual (error) terms and they distribute around zero.

- After that prediction on the test data is made and model is evaluated based on the predictions.









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

We built the model, selected features using RFE, dropped the features that were least helpful in prediction (high p-value), dropped the redundant features (using correlations and VIF) to reach the final model.
Based on the final model, the top 3 features contributing towards explaining the demand of the shared bikes are:

- 'temperature' with coefficient 0.4695, which implies that a 1 unit increase in temperature, increase the rented bikes count by 0.4695.
- 'year' with coefficient 0.2332 also seems an organic feature in explaining the demand of bikes rented.
- Weather situation- 'Light snow/ Light rain' with coefficient −0.2993, also plays a crucial role as it affects the rented bike business negatively.

General Subjective Questions

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

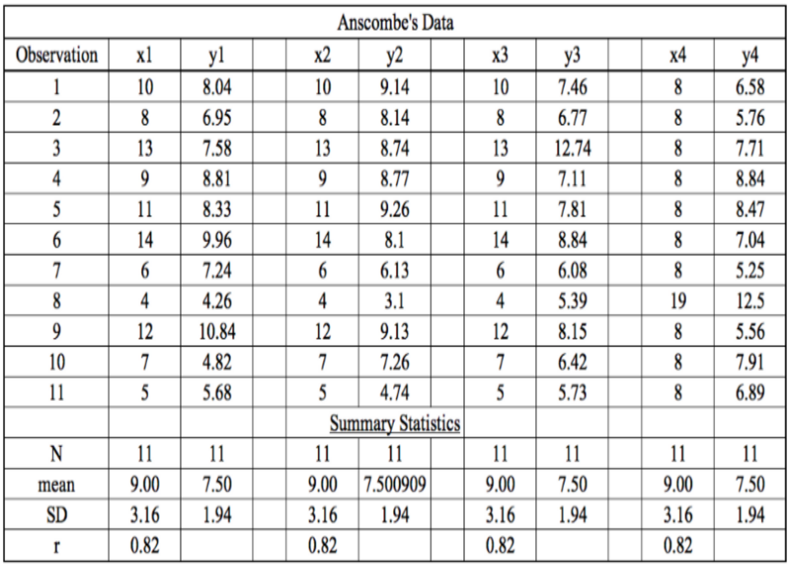
- Linear regression attempts to model the best linear relationship between two variables by fitting a linear equation to observed data.
- The algorithm employs the best-fitting line to map the association between two variables, one of which is considered independent variable and the other a dependent variable.
- There are two types of linear regressions:
  a) Simple Linear Regression (SLR): where model is built using one independent variable only
  $Y = \beta_0 + \beta_1 X$ is the line equation for SLR.
  b) Multiple Linear Regression (MLR): represents the relationship between two or more independent input variables and an output variable
  $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.

$$where\ \beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$$
$$\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient$$

- The cost functions is used to determine the best possible values for $\beta_0, \beta_1, \ldots, \beta_p$, which is used in predicting the probability of the response variable. The equation of the best fit regression can be found by minimising the cost function to predict the dependent variable. There are two ways of cost function minimization: unconstrained and constrained.
- Sum of squares of the difference between the observed and dependent variables is used as a cost function to identify the best fit line. The unconstrained minimization are done using 2 methods
  a) Closed form
  b) Gradient descent
- Gradient descent is the algorithm for finding a local minimum of the differentiable cost function. The data is sliced and split into test and train to build the linear model. We might run into errors while mapping the actual values to the predictors while looking for the best fit line. These errors are known as the residuals. The OLS (Ordinary least square) method is used to estimate these residuals with the goal of minimising the sum of squares and estimate the beta coefficients to achieve a relationship between them.
- We also check the summary statistics including F-statistic, R-squared, coefficients and their p-values to check the significance.

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

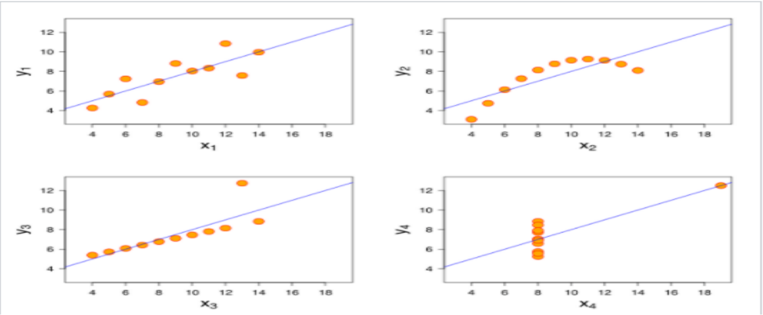
- Descriptive statistics is commonly used to understand the variation of data without having to look at each individual data point. Statistics is useful for describing general trends and aspects of data, but it cannot fully depict any data set on its own. In 1973, Francis Anscombe realised this and created four data sets with a nearly identical statistical properties to demonstrate the point.



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

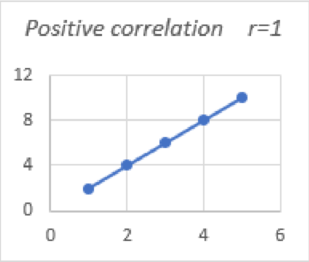
The data sets that make up "Anscombe's Quartet" along with its statistical information are displayed above.

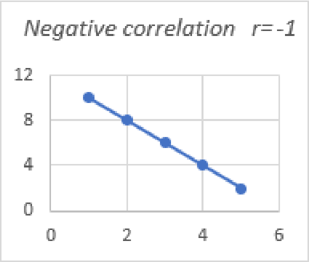


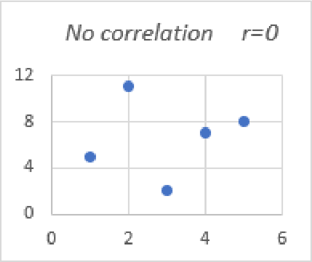
- However, as shown on the right, when plotted on a scatter plot, each dataset produces a different type of plot that was not accurately portrayed even by any regression algorithm in its native form. It demonstrates how, when graphed, multiple data sets with many similar statistical properties can still be vastly different.
- Anscombe's Quartet also warns of outliers present in data sets. The bottom two graphs clearly show the outliers in the datasets that cannot be handled by the linear regression model; otherwise, the descriptive statistics would have been completely different.
- Anscombe's Quartet is a great example of the importance of graphing data in order to visually analyse it before applying any machine learning algorithm to it in order to create a good fit model.

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

- Correlation is a method for determining the strength and direction of a relationship between two variables.
- Pearson's correlation coefficient (r), commonly used in linear regression is a measure of the strength of the linear relationship between two variables. Pearson's R yields values ranging from -1 to 1.







Where:

a) Coefficient r = -1: indicates perfect negative correlation, with data points lying on a perfectly straight line with a negative slope, i.e., strong inversely proportional relationship.

b) Coefficient r = 0: indicates a zero correlation, i.e., no linear relationship.

c) Coefficient r = 1: indicates perfect positive correlation, with data points lying on a perfectly straight line with a positive slope, i.e., strong proportional relationship.

- Positive correlation means that both variables change (increase or decrease) together in the same direction, whereas negative correlation means that as one variable increases, the other decreases, and vice versa. Example: Generally the height of a person increases with the age. The more a person works, the less free time she has in her hands.
  The formula for the coefficient is as follows:

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where : n = sample size
Σxy = the sum of the products first and second value
Σx = the sum of first variable value
Σy = the sum of second variable value
Σx^2 = the sum of squared of first variable value
Σy^2 = the sum of squared of second variable value

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

- Scaling is a step in the data preparation process for a regression model. It normalises/standardizes the various independent features present in data within a given range. It is used during data pre-processing step to deal with highly varying values, and to speed up the process.
- When a model contains a large number of independent variables, many of them may be on very different scales and units, resulting in a model with very strange coefficients that are difficult to interpret, hence incorrect modelling. As a result, we must scale all of the variables to bring them to the same magnitude. We must therefore scale features for two reasons:
  1. Simplicity of interpretation
  2. Faster convergence time

  It should be noted that scaling only affects the coefficients and not the other parameters such as the t-statistic, F-statistic, p-values, R-squared, etc.
- There are two popular methods for scaling:
  (a) Normalization/Min-Max Scaling: It is a scaling technique in which values are shifted and rescaled to a range between min and max, i.e, 0 and 1 respectively. The Min max scaling also aids in normalising outliers.
  Formula:

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

  (b) Standardization scaling: This scaling technique replaces the values by their Z scores. It brings all of the data into a standard normal distribution where the values are centered around the mean (μ) zero with a unit standard deviation (σ).
  Formula:

$$Standardization: x' = \frac{x - \mu}{\sigma}$$

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

- The Variance inflation factor (VIF) index quantifies the degree of multicollinearity in a set of multiple regression variables. To calculate VIF, we fit a regression model between the independent variables.
- Detecting multicollinearity is critical because, while it in no way reduces the model's explanatory power, but it does reduce the statistical significance of the independent features. By a general rule of thumb the VIF should be <10 to avoid multicollinearity.
  Formula:

$$VIF = \frac{1}{1 - R^2}$$

- We can see from the above formula that VIF is infinite when we get $R^2 = 1$, which is the case of perfect fit. It leads to $1/(1- R^2)$ infinity. Therefore, VIF = infinity, if there is perfect correlation between the two independent variables.
- Hence, an infinite VIF value suggests that a linear combination of other variables can exactly express the corresponding variable.

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

- Quantile-quantile plot or Q-Q plot are plots quantiles of two datas against each other. It is a graphical tool for determining whether two data sets are from the common distribution. Theoretical distributions can be normal, exponential, or uniform in nature. In linear regression, Q-Q plots are useful for identifying whether the train and test data sets come from populations with same distributions.
- The interpretations can be made looking at the distribution of the data sets in a straight line as follows:
    a) Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from the x-axis.
    b) Y values < X values: If y-values quantiles are lower than x- quantiles.
    c) X values < Y values: If x-values quantiles are lower than y- quantiles.
    d) Different distributions – If all the data points lie away from the straight line at an angle of 45 degree from the x-axis.
- There are some advantages of using Q-Q plot as well. The sample sizes do not need to be equal and the plot has also a provision to mention the sample size. Using this plot many distributional aspects can be simultaneously tested, for example, shifts in location, scale, and the plot can also used to detect the presence of outliers.