Boom Bikes Case Study

Submitted By: Harjaspreet Singh

Batch: ML- C43

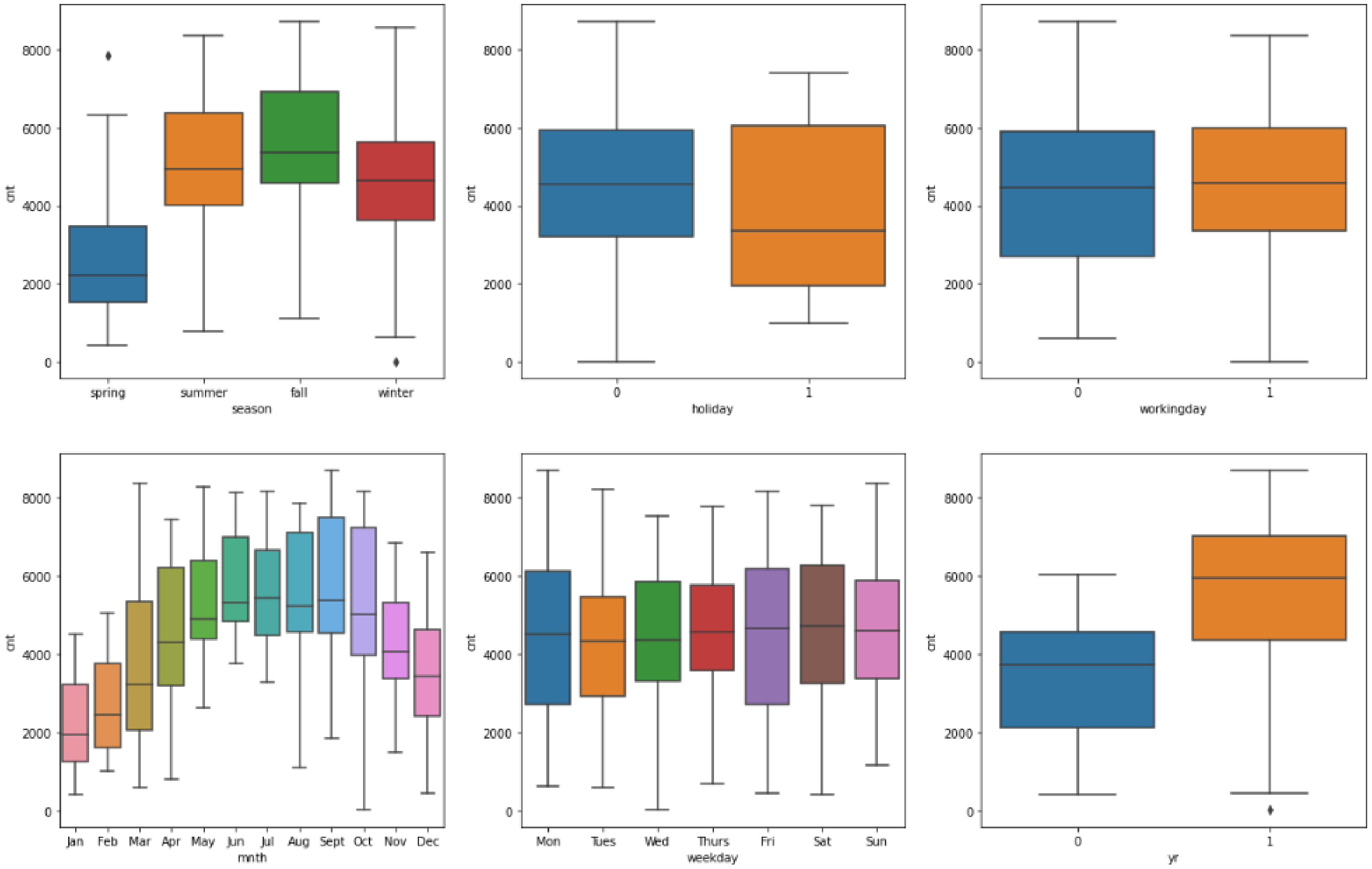
Institute: Post Graduate Diploma in Machine Learning and AI - IIIT, Bangalore

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)
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

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 demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall.
- We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.



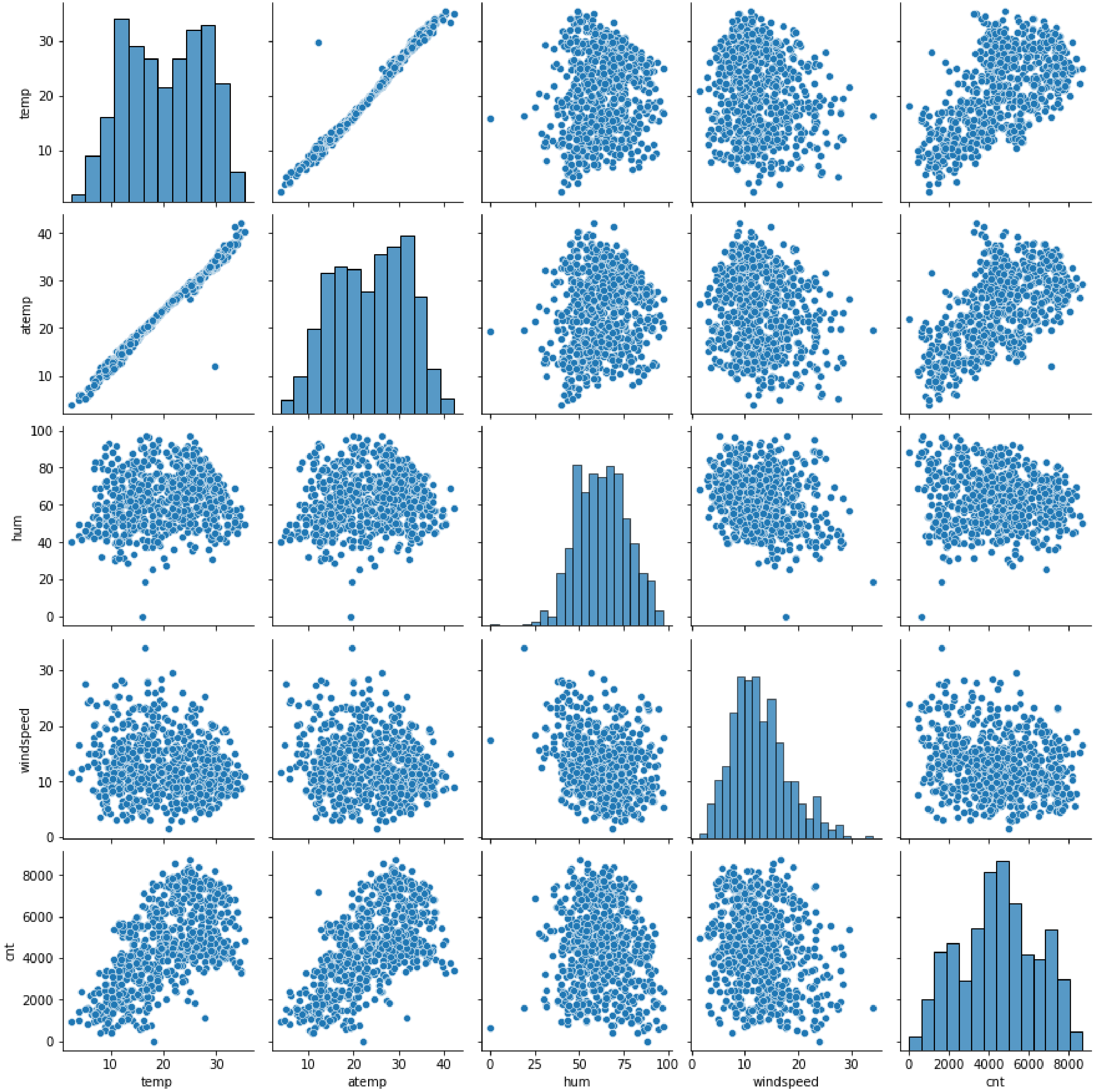
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

- **drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation**. Hence it reduces the correlations created among dummy variables.

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

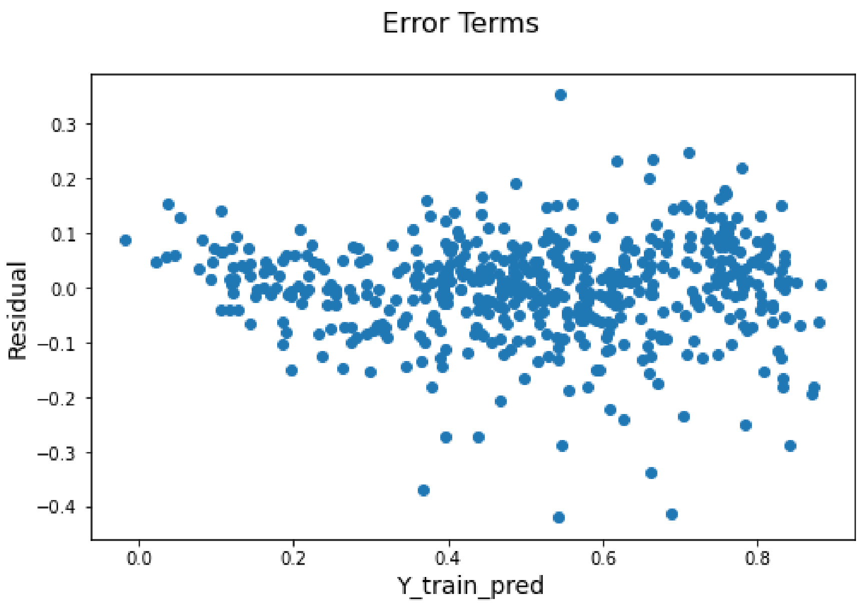
❑ From the Pairplots among the numerical variables, it is quiet visible that – ‘temp’ has the highest correlation with the Target variable (cnt).

❑ It is also evident that both ‘temp’ and ’atemp’ have very strong correlation with each other, and that’s why we have dropped the variable ‘atemp’ from the Training Dataset.



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

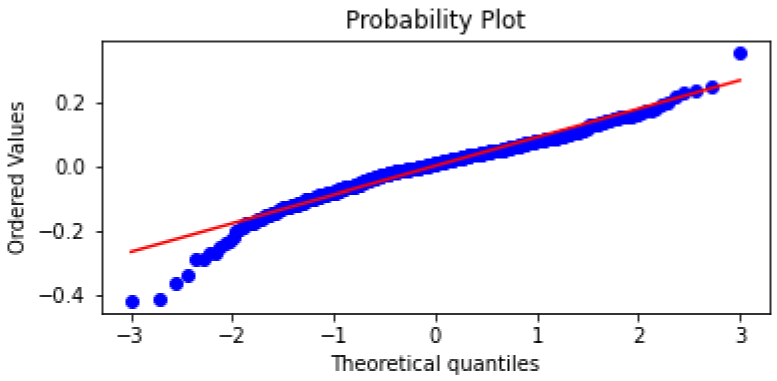
❑ The Assumptions of Linear Regression were tested under ‘ Residual Analysis’

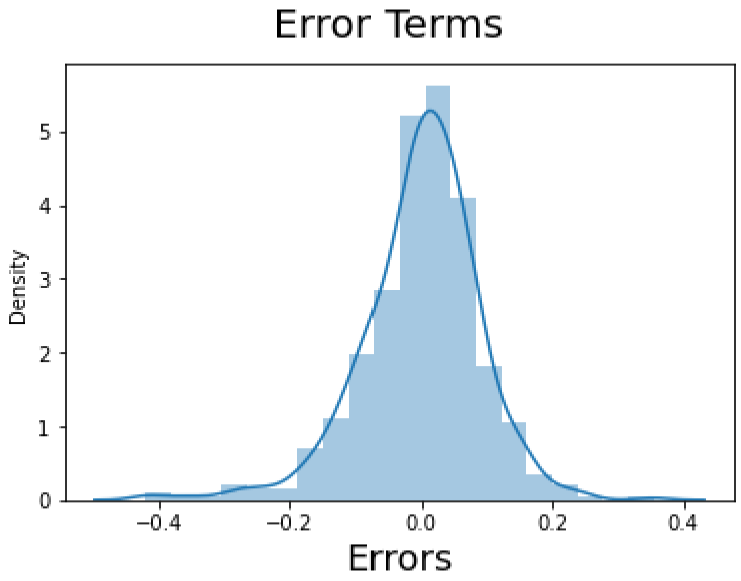


❑ It seems like the corresponding residual plot is reasonably random.

❑ Also the error terms satisfies to have reasonably constant variance (homoscedasticity)

Q-Q Plot





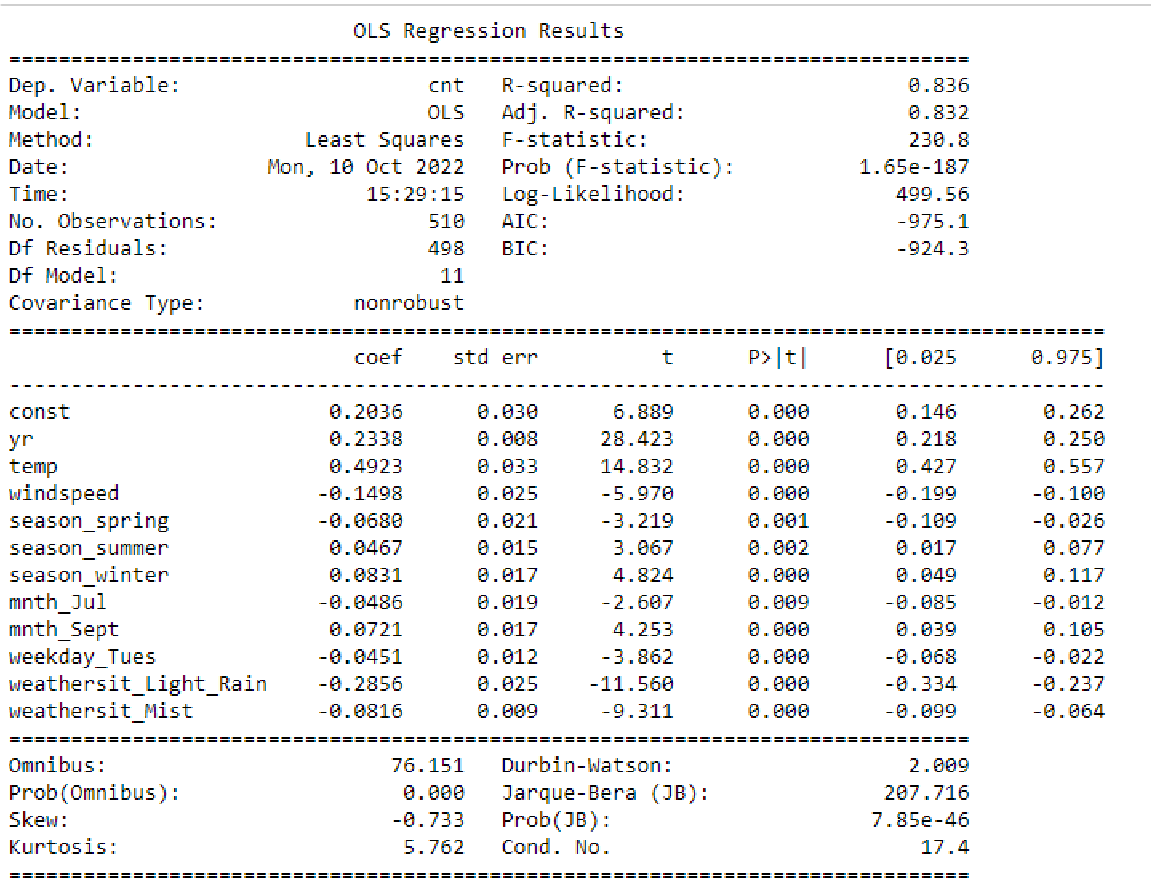
❑ The distribution plot of error term shows the normal distribution with mean at Zero.

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

- The equation of our best fitted line is:

cnt=0.2036
+(0.2338×yr)
+(0.4923×temp)
−(0.1498×windspeed)
−(0.0680×SeasonSpring)
+(0.0467×SeasonSummer)
+(0.0831×SeasonWinter)
−(0.0486×MonthJuly)
+(0.0721×MonthSep)
−(0.0451×WeekdayTuesday)
−(0.0816×WeathersitMistCloudy)
−(0.2856×weathersitLightSnowLightRain)

- From R-Sqaured and adj R-Sqaured value of both train and test dataset we could conclude that the above variables can well explain more than 80% of bike demand.
- Coefficients of the variables explains the factors effecting the bike demand
- Based on final model top three features contributing significantly towards explaining the demand are:
    - Temperature (0.4923)
    - weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.2856)
    - year (0.2338)
- So it recommended to give these variables utmost importance while planning to achieve maximum demand.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.836
Model:                            OLS   Adj. R-squared:                  0.832
Method:                 Least Squares   F-statistic:                     230.8
Date:                Mon, 10 Oct 2022   Prob (F-statistic):          1.65e-187
Time:                        15:29:15   Log-Likelihood:                 499.56
No. Observations:                 510   AIC:                            -975.1
Df Residuals:                     498   BIC:                            -924.3
Df Model:                          11
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 0.2036      0.030      6.889      0.000       0.146       0.262
yr                    0.2338      0.008     28.423      0.000       0.218       0.250
temp                  0.4923      0.033     14.832      0.000       0.427       0.557
windspeed            -0.1498      0.025     -5.970      0.000      -0.199      -0.100
season_spring        -0.0680      0.021     -3.219      0.001      -0.109      -0.026
season_summer         0.0467      0.015      3.067      0.002       0.017       0.077
season_winter         0.0831      0.017      4.824      0.000       0.049       0.117
mnth_Jul             -0.0486      0.019     -2.607      0.009      -0.085      -0.012
mnth_Sept             0.0721      0.017      4.253      0.000       0.039       0.105
weekday_Tues         -0.0451      0.012     -3.862      0.000      -0.068      -0.022
weathersit_Light_Rain -0.2856     0.025    -11.560      0.000      -0.334      -0.237
weathersit_Mist      -0.0816      0.009     -9.311      0.000      -0.099      -0.064
==============================================================================
Omnibus:                       76.151   Durbin-Watson:                   2.009
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              207.716
Skew:                          -0.733   Prob(JB):                     7.85e-46
Kurtosis:                       5.762   Cond. No.                         17.4
==============================================================================
```

General Subjective Questions

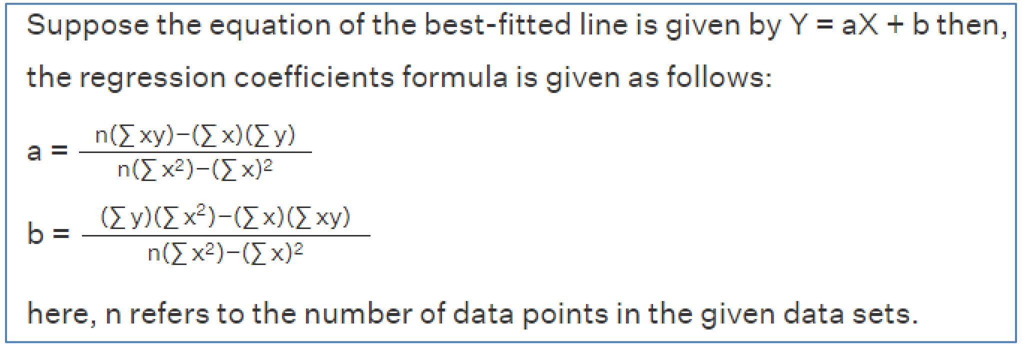
1. Explain the linear regression algorithm in detail. (4 marks)
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

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

- ❑ Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

- ❑ Simple Linear Regression/ Univariate Linear Regression : When we try to find out a relationship between a dependent variable (Y) and one independent (X) then it is known as Simple Linear Regression/ Univariate Linear regression.
  - ❑ The mathematical equation can be given as:
  - ❑ Y = β0 + β1*x
  - ❑ Where
  - ❑ Y is the response or the target variable ,x is the independent feature
  - ❑ β1 is the coefficient of x ;β0 is the intercept ; β0 and β1 are the model coefficients (or weights).

  We can find the values of β0 and β1 using OLS Method or Gradient Descent Method
  The values of Beta coefficients are determined BY MINIMISING THE SSR (Cost Function) in OLS Method



Suppose the equation of the best-fitted line is given by Y = aX + b then,

the regression coefficients formula is given as follows:

$$a = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2}$$

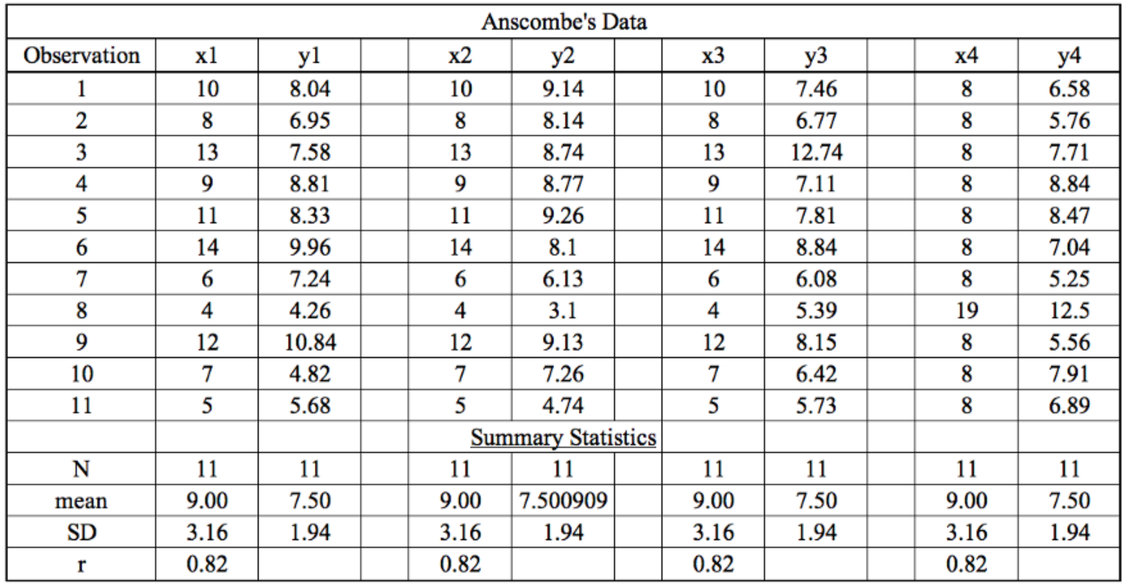
$$b = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2}$$

here, n refers to the number of data points in the given data sets.

- ❑ Multiple Linear Regression : When we try to find out a relationship between a dependent variable (Y) and many independent (X) then it is known as Multiple Linear regression.

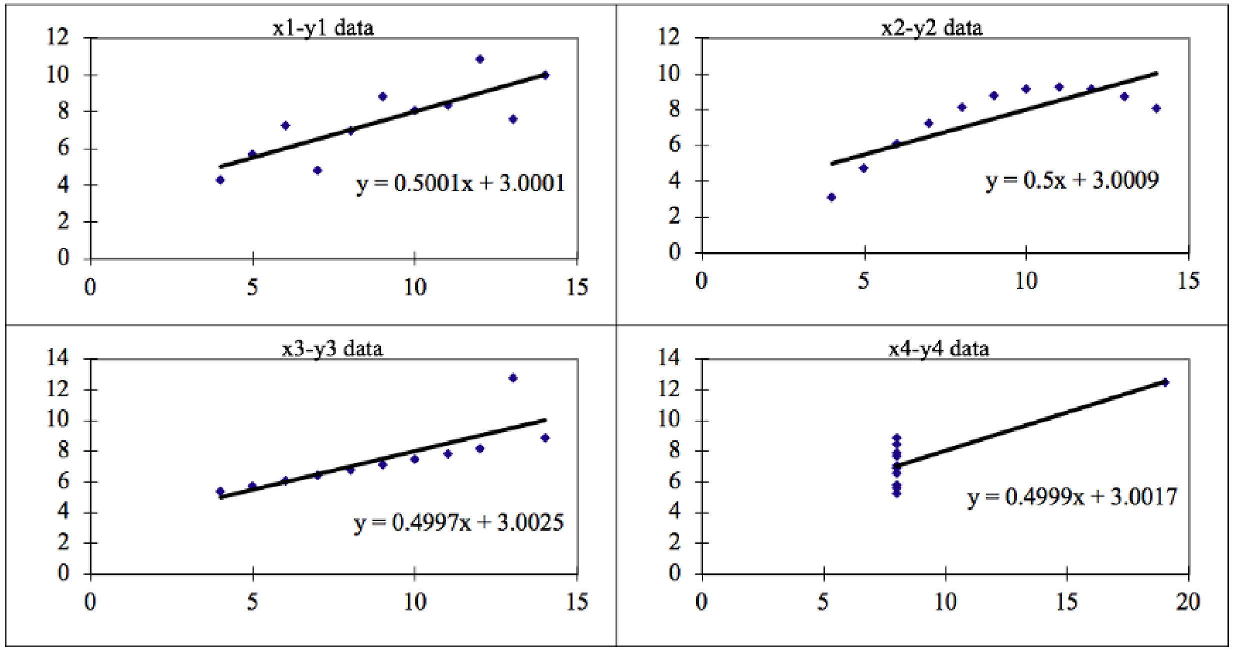
2. Explain the Anscombe’s quartet in detail. (3 marks)

- **Anscombe’s Quartet** was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analyzing and model building, and the effect of other **observations on statistical properties**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets.
- This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.
- Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined as follows:



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 four datasets can be described as:

1.**Dataset 1:** this **fits** the linear regression model pretty well.

2.**Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

3.**Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

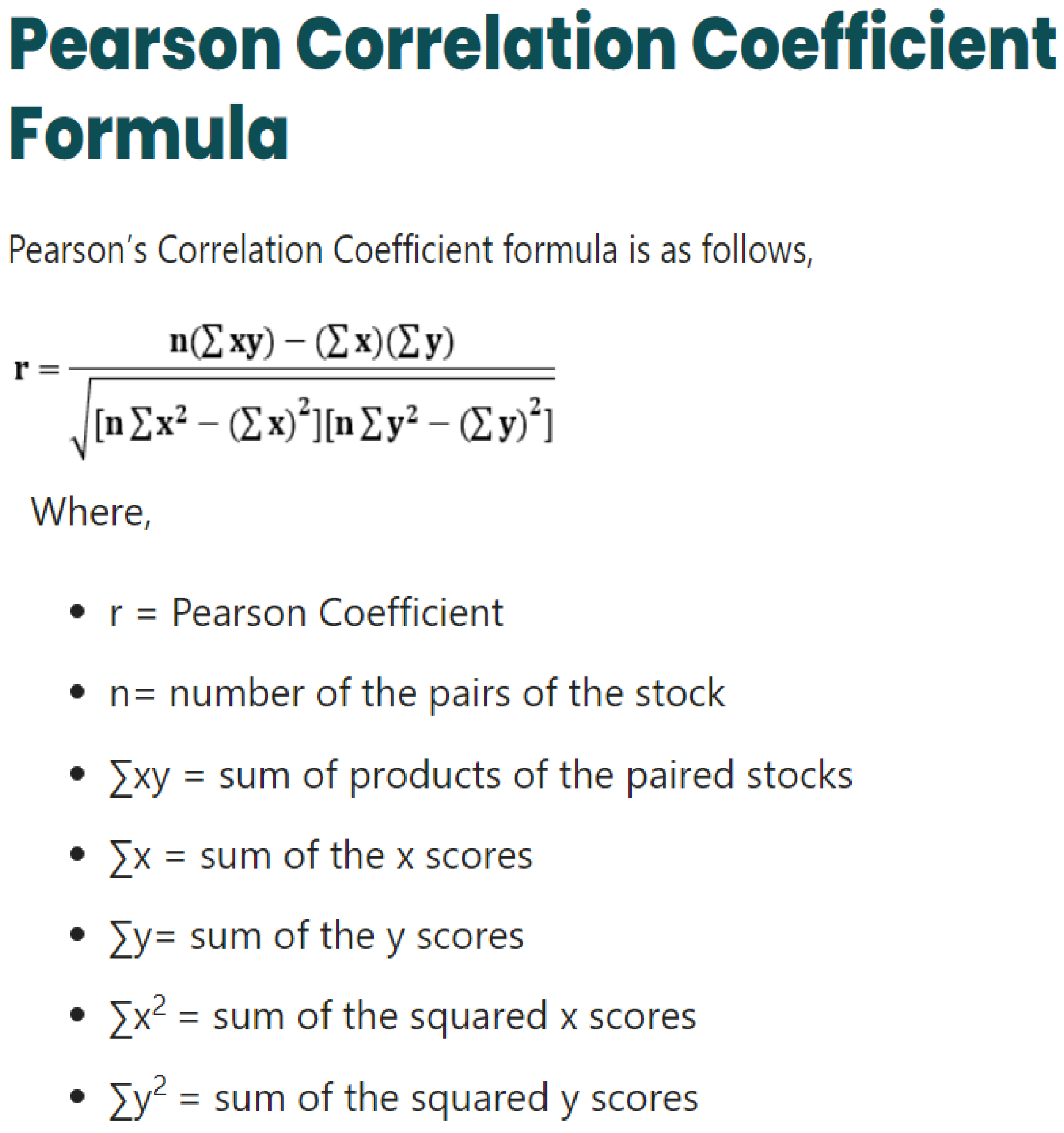
4.**Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

3. What is Pearson’s R? (3 marks)

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter “r” for a sample.

Pearson’s correlation coefficient returns a value between -1 and 1. The interpretation of the correlation coefficient is as under:

•If the correlation coefficient is -1, it indicates a strong negative relationship. It implies a perfect negative relationship between the variables.
•If the correlation coefficient is 0, it indicates no relationship.
•If the correlation coefficient is 1, it indicates a strong positive relationship. It implies a perfect positive relationship between the variables.



Pearson Correlation Coefficient Formula

Pearson’s Correlation Coefficient formula is as follows,

$$\mathbf{r} = \frac{\mathbf{n(\sum xy) - (\sum x)(\sum y)}}{\sqrt{[\mathbf{n\sum x^2 - (\sum x)^2}][\mathbf{n\sum y^2 - (\sum y)^2}]}}$$

Where,

- r = Pearson Coefficient
- n= number of the pairs of the stock
- ∑xy = sum of products of the paired stocks
- ∑x = sum of the x scores
- ∑y= sum of the y scores
- $\sum x^2$ = sum of the squared x scores
- $\sum y^2$ = sum of the squared y scores

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

- ❑ What is scaling?
  - ❑ It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

- ❑ Why is scaling performed?
  - ❑ Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
  - ❑ It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- ❑ What is the difference between normalized scaling and standardized scaling?
  - ❑ *Normalization/Min-Max Scaling:*
    - ❑ *It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

  - ❑ *Standardization Scaling:*
    - ❑ *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).*

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

❑ If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

❑ An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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 is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

❑ For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s)

How is it generated?

Below are the steps to generate a Q-Q plot for team members age to test for normality

1.Take your variable of interest (team member age in this scenario) and sort it from smallest to largest value. Let's say you have 19 team members in this scenario.

2.Take a normal curve and divide it into 20 equal segments (n+1; where n=#data points)

3.Compute z score for each of these points

4.Plot the z-score obtained against the sorted variables. Usually, the z-scores are in the x-axis (also called theoretical quantiles since we are using this as a base for comparison) and the variable quantiles are in the y-axis (also called ordered values)

5.Observe if data points align closely in a straight 45-degree line

6.If it does, the age is normally distributed. If it is not, you might want to check it against other possible distributions

❑ **Importance of Q-Q Plot in Linear Regression : One of the assumption of Linear Regression is that Error Terms follow a Normal Distribution. Hence, we plot a Q-Q Plot to confirm if the assumption is being met**