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 following inferences can be drawn from the analysis of categorical variables using box plots:
1. **'Fall'** season has highest demand for rental bikes.
2. **'Friday'** followed by **'Thursday'** has the highest demand.
3. **1:'good'** - Clear, few clouds, partly cloudy, partly cloudy weather situation has the highest demand for rental bikes.
4. **Continuous growth** in demand of bikes until **'June'**. **'Sept'** month has the highest demand for rental bikes.
5. **Demand** has **increased** from **0:2018 to 1:2019** for rental bikes.

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

Ans: It is important to use drop_first=True during dummy variable creation to reduce the collinearity of the variables. Too many variables of the same nature may lead to redundancy.

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 the highest correlation with the target variable 'cnt'.

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

Ans: The assumptions of Linear Regression are as follows:
a. Linear relationship between X & Y.
b. Error terms are normally distributed.
c. Error terms are independent of each other.
d. Error terms have constant variance (Homoscedasticity).

- After the model is built, the variables are added or dropped using the RFE (Recursive feature elimination) or Manual elimination method.
- The variables shall be dropped in the following order:
- High P-value, high VIF – Definitely drop the variable.
- High P-value, low VIF – 1st
- Low P, high VIF – 2nd
- Low P, low VIF – Keep the variable
- Residual analysis to check if the error terms are normally distributed.
- Model evaluation using scatter plot to check Homoscedasticity.

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: Positive impact

1. Month of 'September'

2. 'Summer' season.

3. 'Winter' season.

4. 'moderate' weather situation

5. 'Temp' - temperature in Celsius.

The above variables can be considered by the company to come up with a business plan to increase the numbers of rental of bikes.

Negative impact

1. 'holiday', 'hum', 'jul', 'windspeed' variables have a negative impact and will result in decrease in rental of bikes.

General Subjective Questions

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

Ans: Linear regression is a type of supervised learning method where past data is used for building the model. Regression is the most commonly used predictive analysis model.

- A simple linear regression model attempts to explain the relationship between a dependent variable and an independent one using a straight line.
- The independent variable is also known as the predictor variable, and the dependent variables are also known as the output variables.
- The standard equation of the regression line is given by the following expression: $Y = \beta_o + \beta_1 X$ where β_o = intercept parameter & β_1= slope parameter.
- The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable. (Figure 1)
- The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination

2. Residual Standard Error (RSE)

- $R^2 = 1 - (RSS / TSS)$ where TSS = Total sum of squares

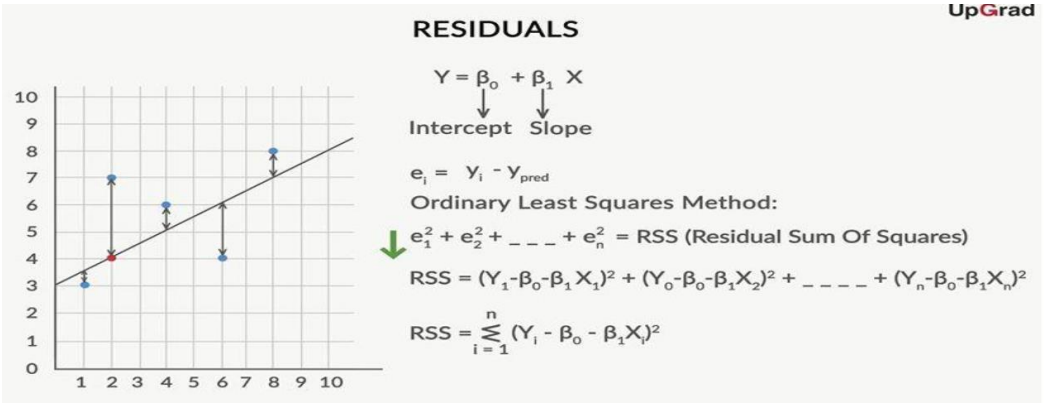


Figure 1 : Residuals
Source: Linear+Regression+Lecture+notes - Upgrad

- TSS (Total sum of squares): It is the sum of errors of the data points from mean of response variable.

Source: Course 5 – Module 1 – Linear Regression

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Source: Linear+Regression+Lecture+notes - Upgrad

Example : Stock market forecast, Weather forecast, Trade forecast etc

2. Explain the Anscombe's quartet in detail.

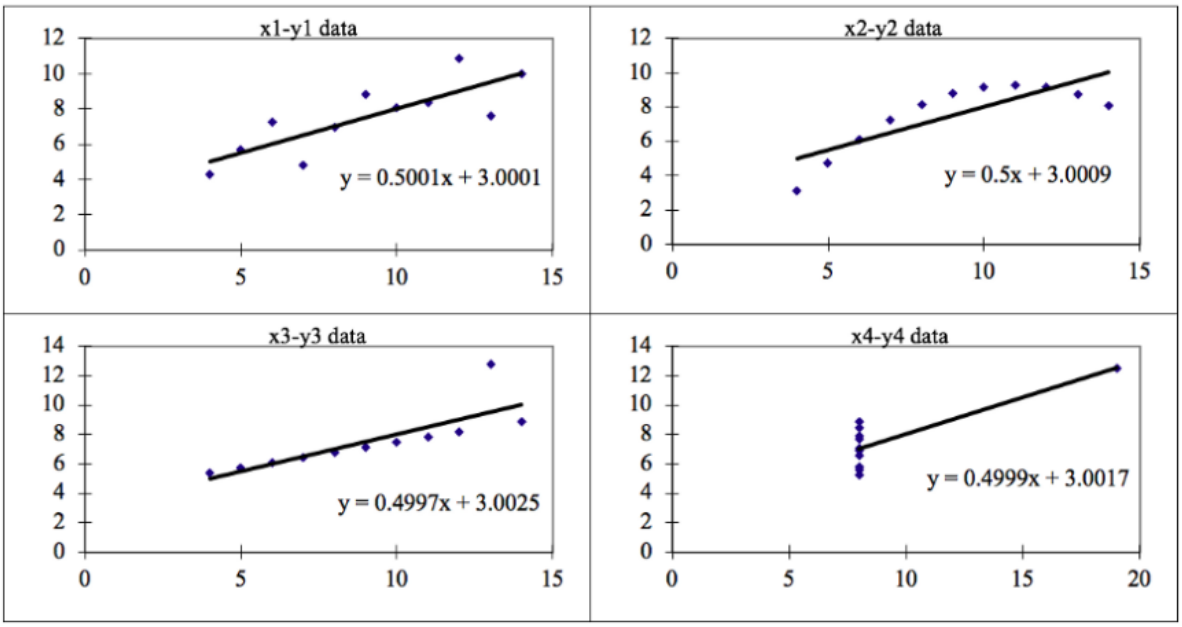
Ans: **Anscombe's quartet** was constructed in 1973 by statistician Francis Anscombe. It comprises of four datasets which have nearly identical simple statistical properties but they look completely different from one another when you visualize the data on scatter plots.

The plotted data helps us in identifying various anomalies present in the data.

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	

Source : *https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2*

When these data sets are plotted on the scatter plot one can observe each data set generates a different kind of plot which is difficult to comprehend by using simple statistical analysis.



Source : *https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2*

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.

Source : https://towardsdatascience.com/importance-of-data-visualization-anscombes-quartet-way-a325148b9fd2

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

Ans: Pearson's R was formulated by Karl Pearson**.** Pearson's R or Pearson's corelation coefficient **explains the linear relationship/strength between two variables. The value of the coefficient lies between -1 & +1.** -1 indicates there is negative corelation between the variables. +1 indicates positive corelation between the variables. 0 indicates no corelation between the variables. To calculate Pearson's R the following conditions should be met:

1. The relation should be linear.
2. Variables should be normally distributed.
3. Absence of outliers in the data.
4. Rescale of variables to have a comparable scale or same scale.

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

Ans: Scaling refers to normalization of data which is applied to independent variables in linear regression.

Scaling is performed so that all the variables are in the same range and can be compared easily**.**

Normalized scaling: Min-max scaling:

All the data is normalized in the range of 0 to 1.

Normalization = $x - x_{min} / x_{max} - x_{min}$

standardized scaling:

All the data is transformed into a mean of '0 'and standard deviation of '1'.

Standardisation = x- mean(x) / standard deviation(x)

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

Ans: **VIF value is infinite when VIF = 1.0 that means there is a perfect corelation between two independent variables.**

VIF or Variance Inflation Factor gives us an idea on how the feature variables are corelated with each other. VIF is calculated by:

$VIF_i = 1/1\text{-}R^2$

VIF or variables greater than 5.00 should definitely be dropped since they are highly corelated. VIF of variables less than 5.00 should be compared with P-value whether to drop the variable or not.

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

   Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

   Source : https://medium.com/@premal.matalia/q-q-plot-in-linear-regression-explained-ab040567d86f#:~:text=Quantile%2DQuantile%20(Q%2DQ)%20plot,populations%20with%20a%20common%20distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Source : https://medium.com/@premal.matalia/q-q-plot-in-linear-regression-explained-ab040567d86f#:~:text=Quantile%2DQuantile%20(Q%2DQ)%20plot,populations%20with%20a%20common%20distribution.