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)

Answer: I have done analysis on categorical columns using boxplot on the dependent variable “cnt”.

1. “Fall” and “Spring” season has the highest and lowest median.
2. Bike booking got increased drastically from 2018 to 2019.
3. Most of the bookings were done during the month of may, june, july, aug, sep and oct and then it decreased again
4. Most of the bookings were done when it is not holiday
5. Most of the bookings were done when weather was clear
6. Median is same for working and non-working days but the spread is bigger in non-workingdays

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

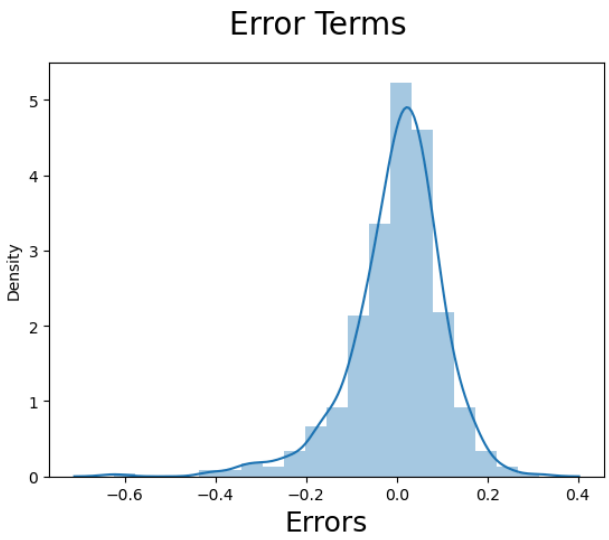
Answer: A categorical variable with n levels can be represented with n-1 levels (n columns will created during dummy variable creation).So we can remove the redundant column. Other n-1 variables will be able to retain all the information.

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

Answer: ‘atemp’ variable has the highest correlation with the target variable.

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

Answer: I have validated the assumptions of Linear Regression by plotting a histogram of residuals and found that it has normal distribution.



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

Answer: Below are the top 3 features contributing significantly towards explaining the demand of the:

- temp (0.549011 - coeff)

- yr (0.238506 - coeff)

- windspeed (0.182087 - coeff)

General Subjective Questions

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

Answer: Linear regression is a machine learning algorithm based on supervised learning.

1. Here dependent variable Y is dependent on independent variables X1,X2.. etc and only in base format i.e, $Y = β0 + β_1 * X_1 + β_2 * X_2 +..+ βn * Xn$
2. It is all about finding betas
3. Target variable is always continuous

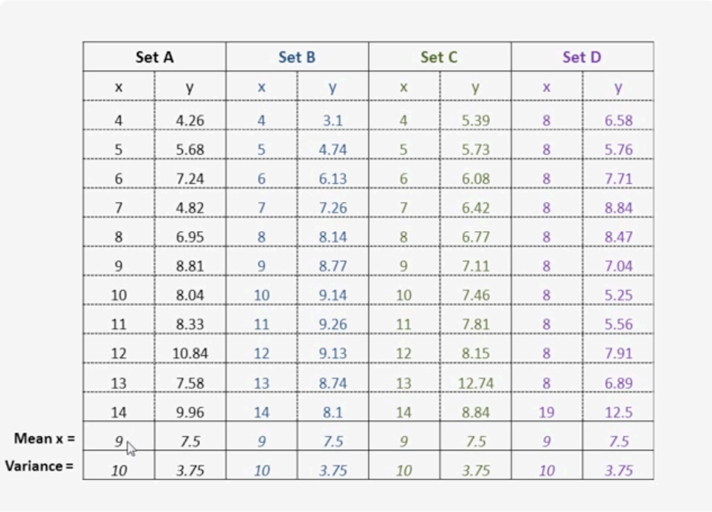
Assumption in Linear Regression:

1. Adding more attribute is not always helpful
2. Error should be normally distribute
3. Multicollinearity check

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

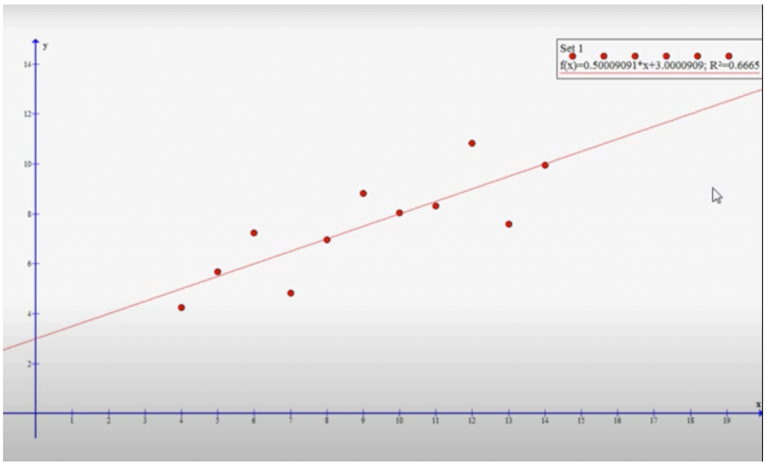
Answer: Anscombe's quartet serves as a powerful reminder that statistics are tools for analysis, they do not replace common sense and should be supported by anecdotal analysis and visualisation.

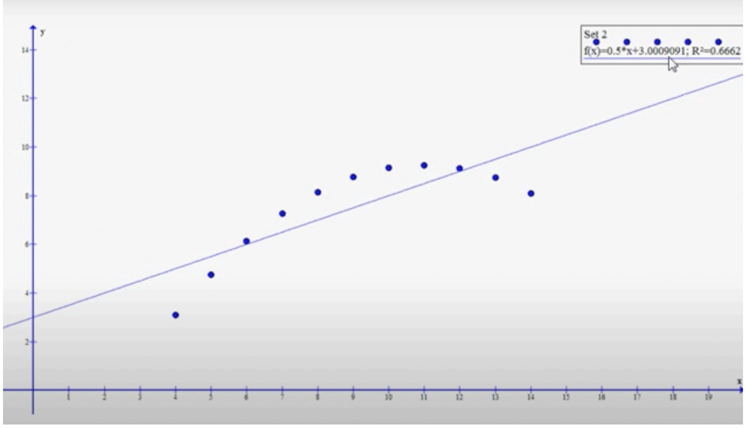
Below are the 4 data sets with same mean and variance:

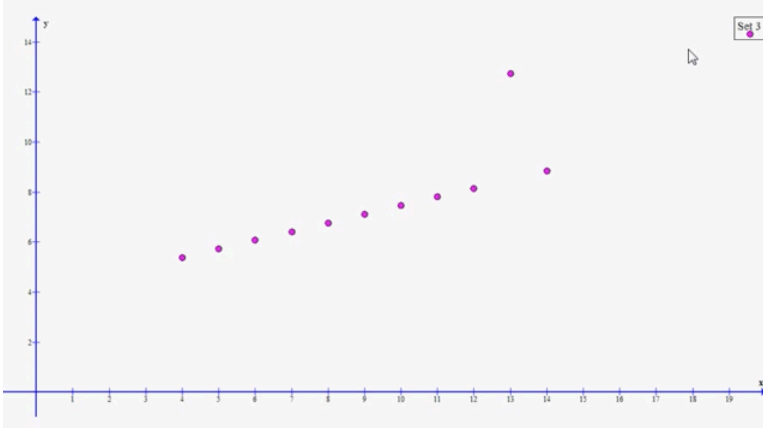


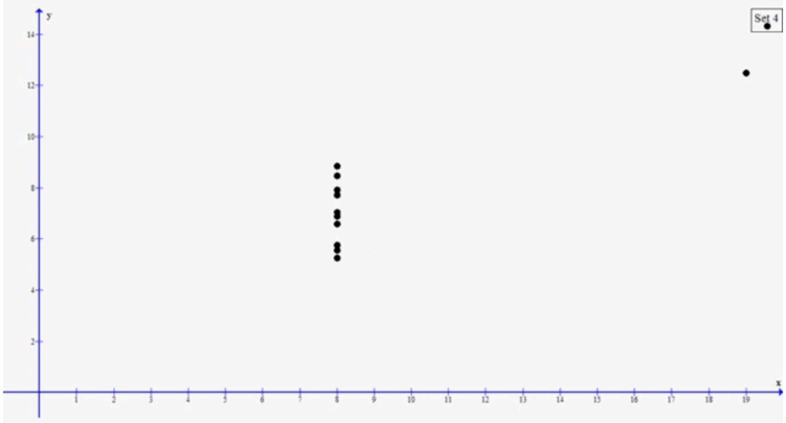
	Set A		Set B		Set C		Set D	
	x	y	x	y	x	y	x	y
	4	4.26	4	3.1	4	5.39	8	6.58
	5	5.68	5	4.74	5	5.73	8	5.76
	6	7.24	6	6.13	6	6.08	8	7.71
	7	4.82	7	7.26	7	6.42	8	8.84
	8	6.95	8	8.14	8	6.77	8	8.47
	9	8.81	9	8.77	9	7.11	8	7.04
	10	8.04	10	9.14	10	7.46	8	5.25
	11	8.33	11	9.26	11	7.81	8	5.56
	12	10.84	12	9.13	12	8.15	8	7.91
	13	7.58	13	8.74	13	12.74	8	6.89
	14	9.96	14	8.1	14	8.84	19	12.5
Mean x =	9	7.5	9	7.5	9	7.5	9	7.5
Variance =	10	3.75	10	3.75	10	3.75	10	3.75

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:

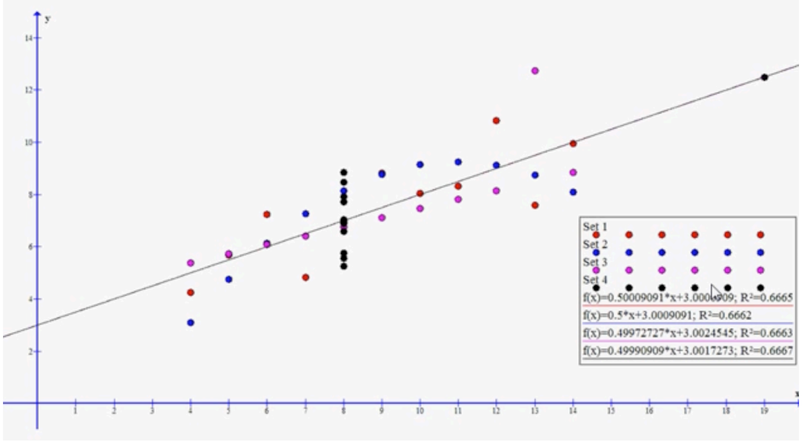








Summary of the data sets:



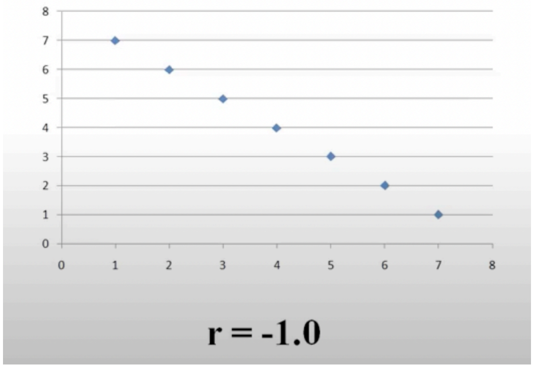
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data

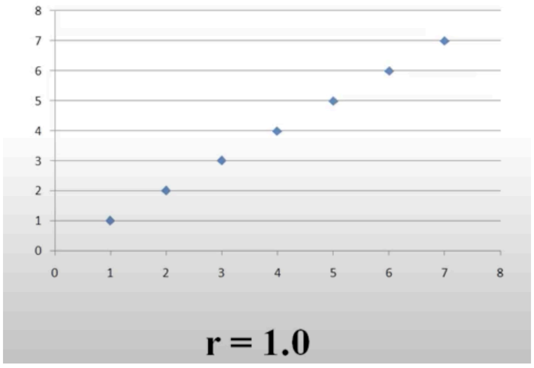
reveals a lot of the structure and a clear picture of the dataset.

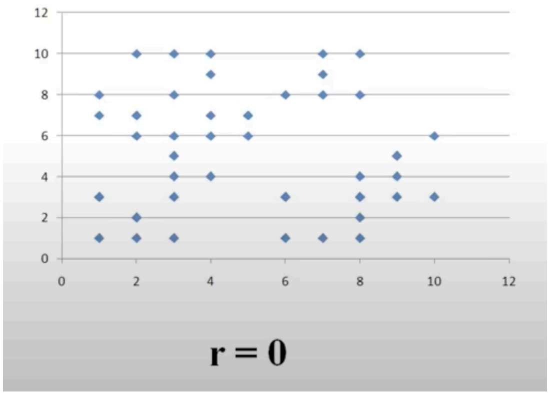
3. What is Pearson's R? (3 marks)

Answer: Pearson's R measures the strength of the linear relation ship between two variables.

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.







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

Answer: Feature scaling is a fundamental preprocessing step in machine learning aimed at ensuring that values of numerical features are within a certain range. It makes computation of gradient descent faster

Difference between normalized scaling and standardized scaling:

1. Normalization involves scaling data values in a range between [0,1] or [-1,1], and is best for unknown or non-normal distributions. Data standardization involves scaling data values so that they have a mean of 0 and standard deviation of 1, and is best for normal distributions
2. 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)

Answer: Formula of VIF is : 1/(1 - R2) , R2 => R squared

If R2 is 1 then it can become infinite. It means perfect correlation present between the features

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

Answer: A Q-Q (quantile-quantile) plot shows how two distributions' quantiles line up. We can verify that the distribution of the data set

For example, We can plot sample quantiles against theoretical quantiles. If it gives straight line then we can say that sample is following normal distribution.