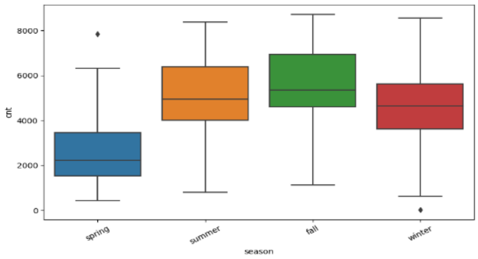
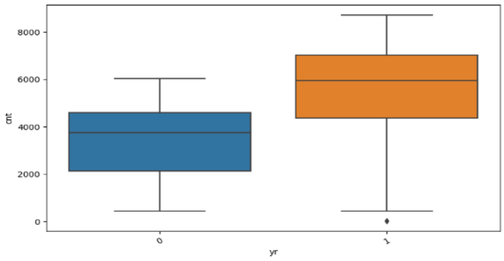
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Season: Fall season has higher demand for bikes.

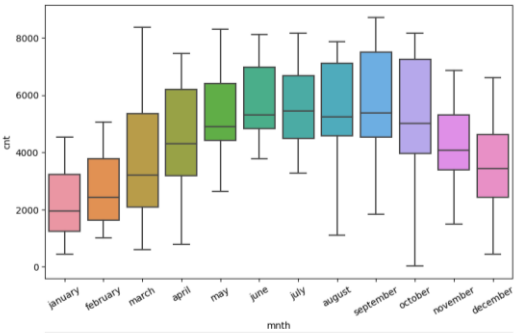
Spring season has less demand.

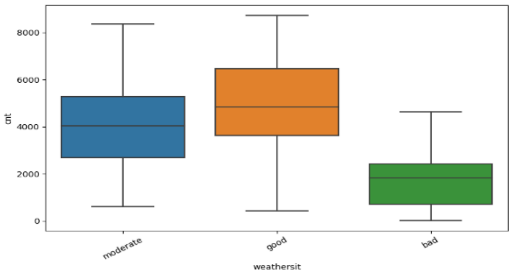




Year: 2019 has higher demand than 2018.

Month: From January to June, we can see the increasing pattern in demand, from June to September has the highest steady demand and after September there is decreasing in demand for bikes.





Weather sit: Good weather condition has highest demand followed by moderate and bad weather condition, there is not data on severe weather condition.

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

Since one of the columns can be generated completely from the others and hence retaining this extra column does not add any new information for the modelling process.

If ‘K’ dummy variables present in data, we should create ‘(K-1)’ dummy variable columns to represent ‘K’ levels.

Hence it is the good practice to drop the first column dummy variable creation.

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

'temp' and 'atemp' feature has highly correlated to target feature 'cnt'

'temp' - temperature in Celsius

'atemp' - feeling temperature in Celsius

'cnt' - count of total rental bikes including both casual and registered

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

To validate the assumptions of linear regression after building the model on the training set step to follows are:

- ***Linearity check:***
  The relationship between the independent and dependent variables should be linear.
- ***Normality of Residuals:***
  The residual should be normally distributed.
- ***Homoscedasticity:***
  Homoscedasticity means that the variance of residuals should be constant across all levels of the independent variable.
- ***Multicollinearity:***
  There should be very little or no multicollinearity among the variables.

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- ✓ '**temp**' - temperature in Celsius
- ✓ **'weathersit_bad'** - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- ✓ **'yr'** – Year

General Subjective Questions

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

Linear regression is machine learning algorithm based on supervised learning used to model the linear relationship between a dependent variable and one or more independent variables.

The general form of a linear regression equation,

For single independent variable, **Y = β0 + β1 * X + ε**

For multiple independent variable, **Y = β0 + β1 * X1 + β2 * X2 + ... + βn * Xn + ε**

where,

Y: Dependent variable (target)

X: Independent variable (feature)

β0: Y-intercept (constant term)

β1: Coefficient of the independent variable (slope)

ε: Error term (residual), accounting for unexplained variability

In linear regression, the commonly used cost function is the Mean Squared Error (MSE)

MSE = (1/n) * Σ(actual Y - predicted Y)^2

Where, n – number of data points in the data.

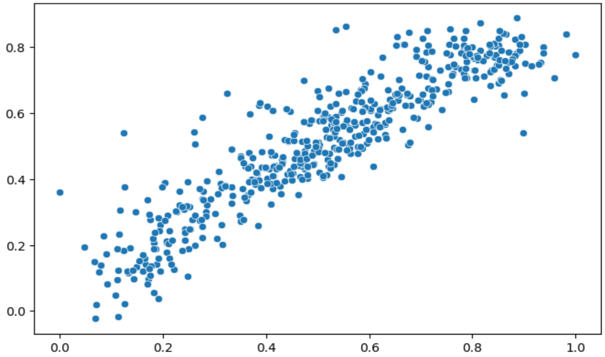
The goal is to determine the values of β0, β1, β2, ..., βn this can be done using various optimization techniques, with the most common being the least squares method that minimize the Residual sum of squared (RSS). Residual = (actual Y – predicted Y)

Once the coefficients are determined, you can use the linear equation to make predictions on new data. We can evaluate the model performance using **R-squared** value to measure how well the model fits the data.

Assumptions:

Linear regression assumes certain conditions, such as

- ***Linearity check:***

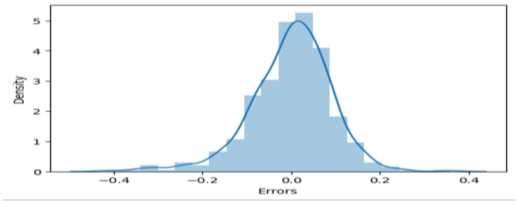


✓ The relationship between the independent and dependent variables should be linear either positive or negative.

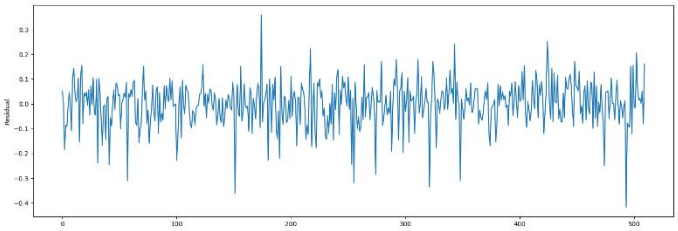
✓ Here the graph shows the positive linear relationship.

- ***Normality of Residuals:***

  - ✓ The residual should be normally distributed.



- ***Homoscedasticity:***



✓ Homoscedasticity means that the variance of residuals should be constant across all levels of the independent variable.

- ***Multicollinearity:***

  There should be very little or no multicollinearity among the variables.

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

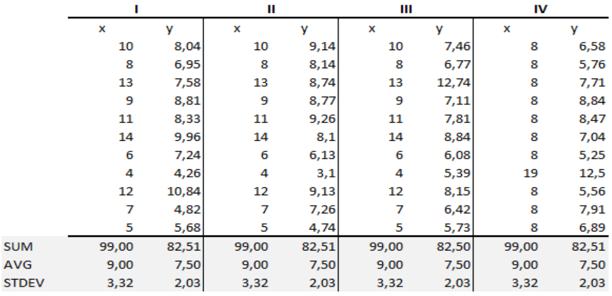
In 1973 Anscombe’s quartet were constructed by the **statistician Francis Anscombe** to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.

We should never just run the regression without having a good look because simple linear regression has quite a few shortcomings.

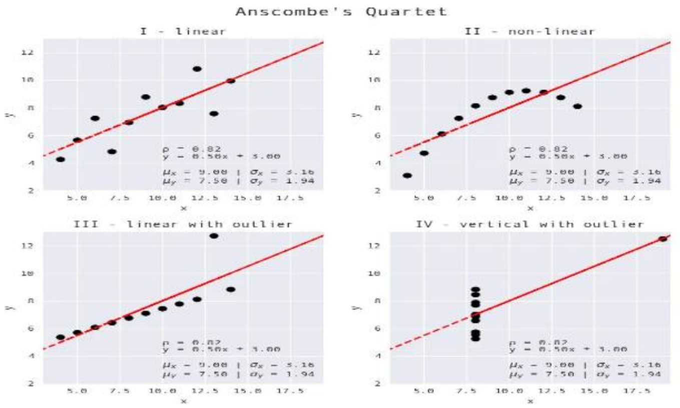
- ✓ sensitive to outlier
- ✓ Models the linear relationships only
- ✓ Few assumptions are required to make inferences.

Let’s comprise four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics.

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.



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03



- ➢ Dataset I appear to have clean and well-fitting linear models.
- ➢ Dataset II is not distributed normally.
- ➢ In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- ➢ Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

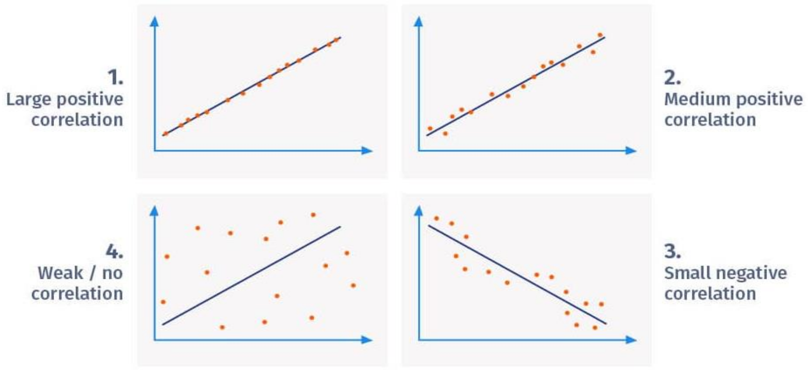
As shown all the four linear regressions are exactly the same. But there are some peculiarities in datasets which have fooled the regression line.

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)

Pearson's correlation coefficient, often denoted as "r" or Pearson's "r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is used to assess how closely the data points of two variables cluster around a straight line.

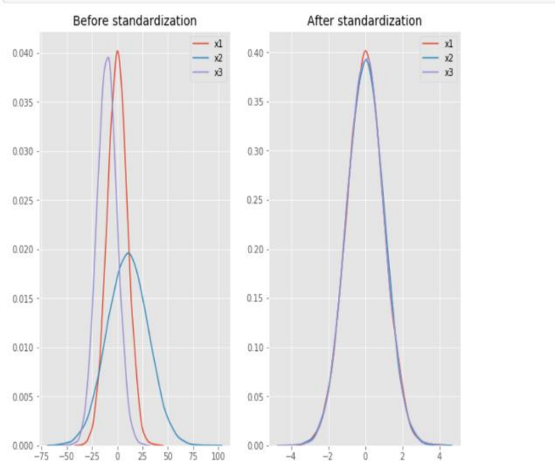
The value of Pearson's r ranges from -1 to +1.



- ✓ A positive value close to +1 indicates a strong positive linear correlation, meaning that as one variable increases, the other variable also tends to increase.
- ✓ A negative value close to -1 indicates a strong negative linear correlation, meaning that as one variable increases, the other variable tends to decrease.
- ✓ A value close to 0 indicates a weak or no linear correlation between the variables.

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

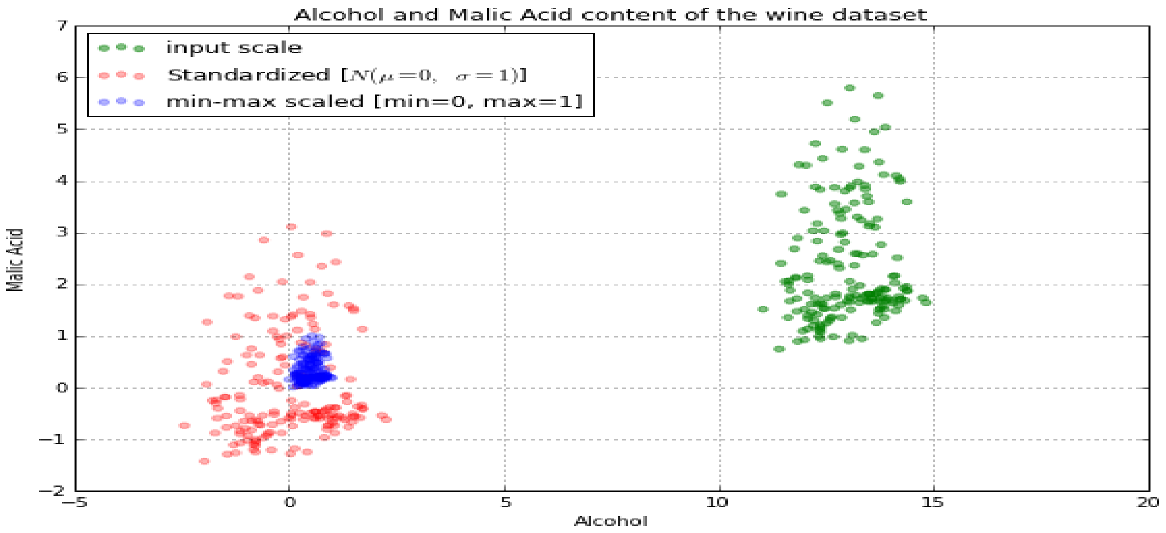
Scaling refers to the process of transforming the features of a dataset to a common scale. The purpose of scaling is to ensure that all features contribute equally to the analysis or modelling process and to improve the performance of certain algorithms that are sensitive to the scale of the input features.



Scaling is performed for several reasons:

- ***Algorithm Sensitivity***: With features having different scales, algorithms might assign disproportionate importance to features with larger scales, leading to biased results.
- ***Convergence Speed***: Speed up the training process and lead to more efficient model convergence.
- ***Distance Metrics****:* Scaling ensures that distances are computed accurately and are not dominated by features with larger scales.

S.No	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization



The above plot shows the distribution of values of through different scales and input scale.

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

It is a statistical measure used to assess multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to determine the individual effects of these variables on the dependent variable. VIF is used to quantify the extent to which the variance of the estimated regression coefficients is increased due to multicollinearity.

If there is perfect correlation, then VIF = infinity. 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.

In case of perfect correlation, we get r-squared (R^2) = 1
then **VIF = 1 / (1 – R^2)** leads to infinity.

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

A Quantile-Quantile (Q-Q) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution or to compare the distribution of two datasets. It is particularly useful for evaluating the assumption of normality in statistical analysis, including linear regression.

The Q-Q plot compares the quantiles of the dataset to the quantiles of a theoretical distribution to determine if the observed data deviate from the expected distribution.

The use and importance of a Q-Q plot in linear regression are as follows:

- ✓ **Checking Normality Assumption:** By creating a Q-Q plot of the residuals, you can visually assess whether the residuals follow a normal distribution.

- ✓ **Detecting Outliers:** Outliers can cause issues in linear regression by disproportionately influencing the regression line and affecting the overall fit of the model. Outliers might cause the Q-Q plot to deviate from a straight line at the tails.

- ✓ **Model Evaluation:** If the Q-Q plot indicates that the residuals deviate significantly from normality, it suggests that the linear regression assumptions might not be fully met. This might lead you to consider alternative regression techniques or transformations of the data.

- ✓ **Data Transformation:** If the Q-Q plot shows significant departures from normality, you might consider transforming the data to better meet the assumption of normality. Common transformations include logarithmic, square root, or Box-Cox transformations.

In summary, a Q-Q plot is a valuable tool for assessing the normality of residuals and identifying potential outliers in linear regression analysis. It helps you make informed decisions about the appropriateness of the linear regression model and guides you in addressing any issues that might arise due to deviations from normality.