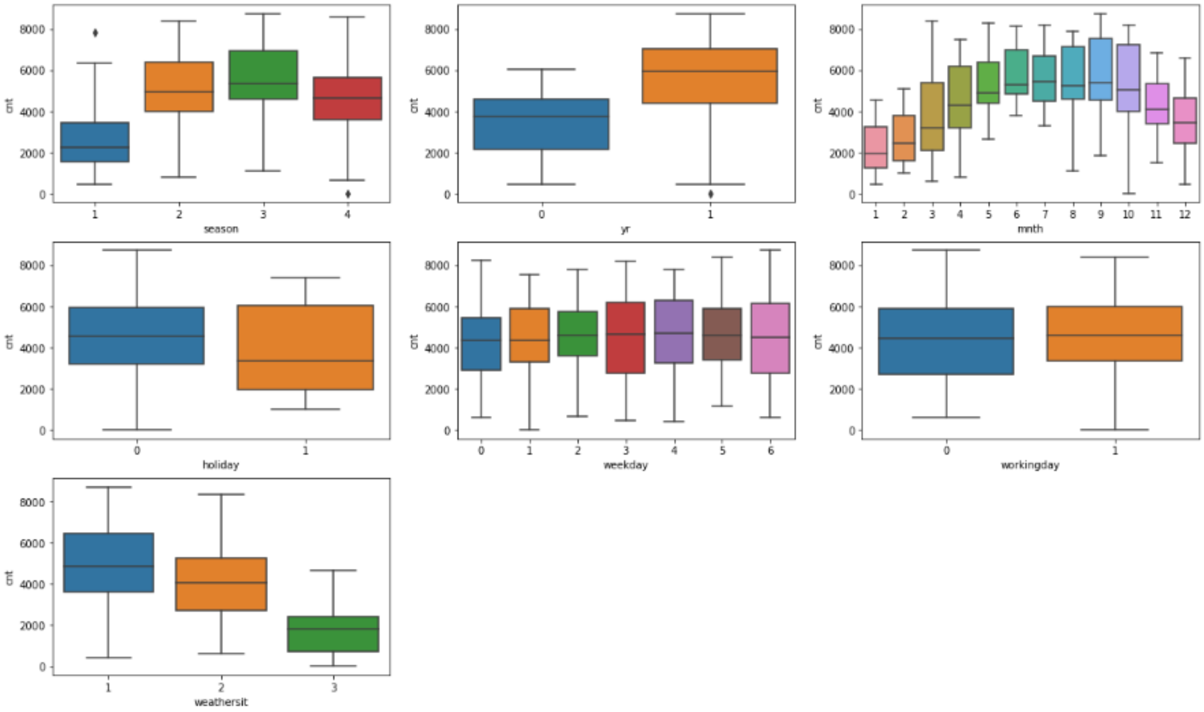
Assignment-based Subjective Questions

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

We have done analysis on the relationship between the categorical variables and the dependent variables in the EDA step of this assignment. Categorical features in the dataset are:

- **season** : season (1:spring, 2:summer, 3:fall, 4:winter)
- **yr** : year (0: 2018, 1:2019)
- **mnth** : month (1 to 12)
- **holiday** : weather day is a holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- **weekday** : day of the week
- **workingday** : if day is neither weekend nor holiday is 1, otherwise is 0.
- **weathersit** :
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

We plotted the boxplots for each of the categorical variables against the target variable to observe their relationship.



Observations from the above plots:
Cnt vs season - We can see that in summer and fall, the number of rider are higher compared to spring and winter

Cnt vs year - Since the data is only for 2 years, and it is a startup company, the number of rides are growing so we have higher number of rides in 2019 compared with 2018

Cnt vs month - This plot also shows somewhat same trend as the season plot, the number of rides start increasing in the month of March April, peaks around Aug-Sep and then dips dowards the end of the year, it shows some kind of seasonal trend for the bike demand, something that we also saw in the temperature charts also.

Cnt vs Holiday - The overall spread of the data is higher on holidays, but the median requests on non-holidays days are higher.

Cnt vs weekday - Again there is no specific trend, but weekday 3 (Wednesday) and weekday 6 (Saturday) have higher spread.

Cnt vs Workingday - Rides are slightly higher on working days.

Cnt vs weathersit - Clear trend that the demand is higher when the skies are clear, no rain or snow (weathersit =1), demand goes very low when their is mist, rain or snowfall around (weathersit = 2 and 3)

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

For a feature with n levels of categorical values, we need n-1 number of columns to convert it into dummy variable. Now when we use drop_first = True, then we are saying that the first column of the dummy variables can be ignored, as when all the other columns of the dummy variables are 0, we can infer that the first columns is 1. So in order to keep the optimal number of features in the model and not to unnecessarily keep columns that are not needed, we use drop_first = True.
For example: If we have columns called color, and the values in the column are Red, Blue, and Green as shown below, we can represent them using three columns (Table A) as shown below, but even if we drop the column R, still we can represent the data as shown in the table B.

	R	G	B
Red	1	0	0
Green	0	1	0
Blue	0	0	1

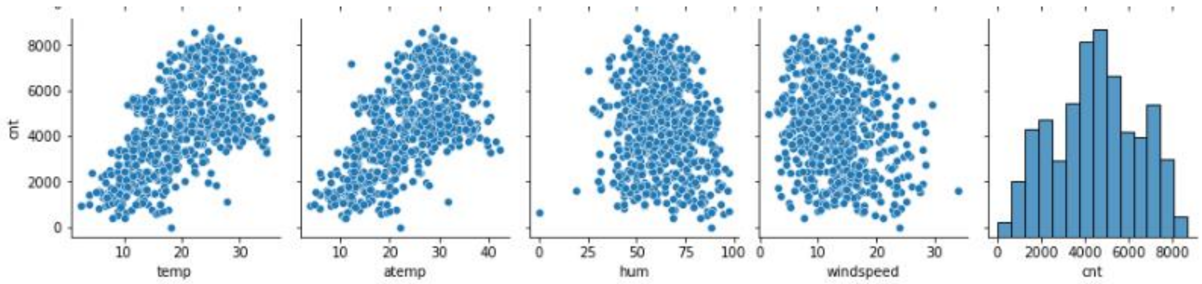
Table A

	G	B
Red	0	0
Green	1	0
Blue	0	1

Table B

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

Pair plot of the numerical variables and the target variable is given below:

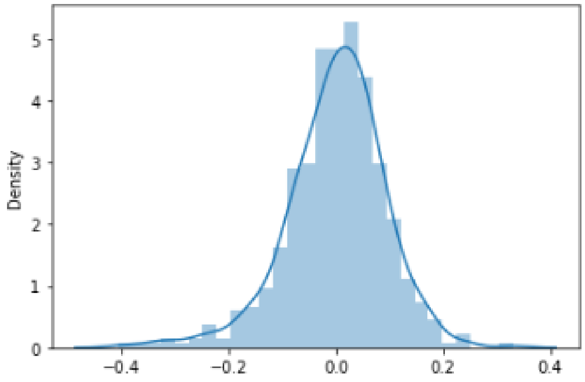


Looking at the pair plot we can say that ***temp*** and ***atemp*** has highest correlation with target variable ***cnt***

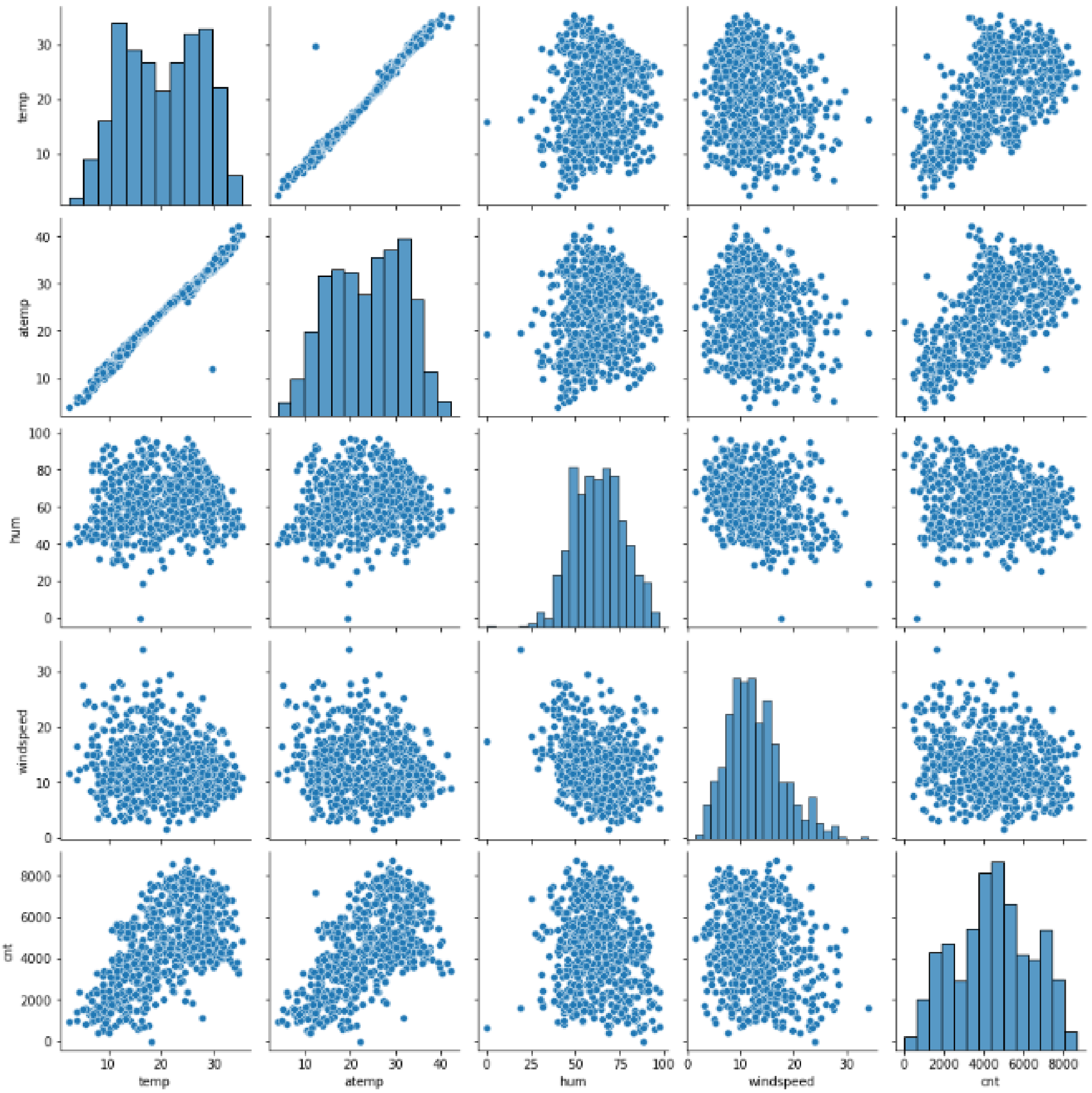
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Assumptions around the linear regression were validated as below:

- **Normally distributed error terms with zero mean** – by plotting a distplot of the residuals i.e. y_train-y_pred as shown in the below image that the residuals are normally distributed with zero mean.



- **Linear relationship** between a number of numeric variables (predictors) and cnt variable (Predicted), which we saw in the pairplot during the EDA. We plotted a pair plot during the EDA step of the assignment to find some linear relationship between some of the predictor variables and the predicted variables.



- **Multicollinearlity between predictor variables:** No multicollinearity between the predictor variables in the final model, checked using the VIF table (all less than 5) in the final model

```
        feature   VIF
0            yr  2.00
1       holiday  1.04
2          temp  3.68
3     windspeed  3.06
4      season_2  1.57
5      season_4  1.37
6        mnth_9  1.20
7  weathersit_2  1.48
8  weathersit_3  1.08
```

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

Below is our final model:
cnt = 0.1259 +(yr * 0.2329) +(holiday * -0.0987)+(temp * 0.5480)+(windspeed * -0.1532)+(season_2 * 0.0881)+(season_4 * 0.1293)+(mnth_9 * 0.1012)+(weathersit_2 * -0.0784)+(weathersit_3 * -0.2829)

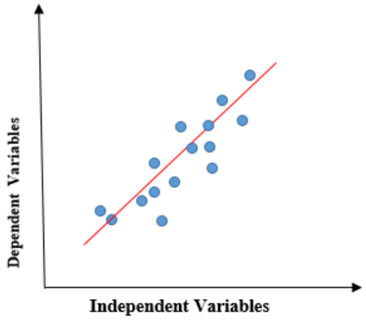
As per this model, the top three features impacting the demand are:

- Temperature – A beta coefficient of 0.5480 which means that for each unit rise in temperature the demand of the bikes increases by 0.5480 given that all the other factors remain same.
- Year – Year is the second most significant feature impacting the demand, with each unit rise, the demand is expected to go up by 0.2329.
- Windspeed – This is the third most significant feature, but the difference from the above thwo variables is that it is negatively correlated with the demand i.e. with each unit rise in the windspeed the demand is expected to go down by 0.1532, assuming everything else remains constant.

General Subjective Questions

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

Linear regression is the most basic and widely used statistical regression method used for predictive analysis and shows the relationship between the continuous variables. 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.



Types of Linear Regression
Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

- $$Y = \underset{\text{Intercept}}{\underset{\downarrow}{\beta_0}} + \underset{\text{Slope}}{\underset{\downarrow}{\beta_1}}\ X$$
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression. Formula of multiple linear regression will be –

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Algorithm:

Linear regression algorithm works by trying to find out something known as the best fit line. To find the best fit line, we use the concept of residuals which is nothing but error term i.e. y_actual - y_predicted
For every data point we have e_i = y_i – y_pred, to find the best fit line we use Ordinary Least Squares method, we square the errors for each data point i.e. e1^2 + e2^2 + ….+en^2
We want to minimize the above since they are nothing but errors.

It is also called RSS (Residual Sum of squares) = e1^2 + e2^2 + ….+en^2

$$RSS = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_2 - \beta_0 - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

So we are trying to minimize the RSS to get the best fit line.

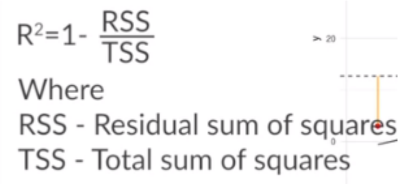
However we find that RSS is an absolute quantity and will change if we change the units of measurement, so we need to convert it into relative measurement, we do that using TSS i.e. Total Sum of Squares. Formula for TSS is given below:

$$TSS = (Y_1 - \bar{Y})^2 + \ldots.. + (Y_n - \bar{Y})^2$$

$$\text{Or} \quad \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Ratio of TSS and RSS is used as normalized quantity used to measure the goodness of a model.

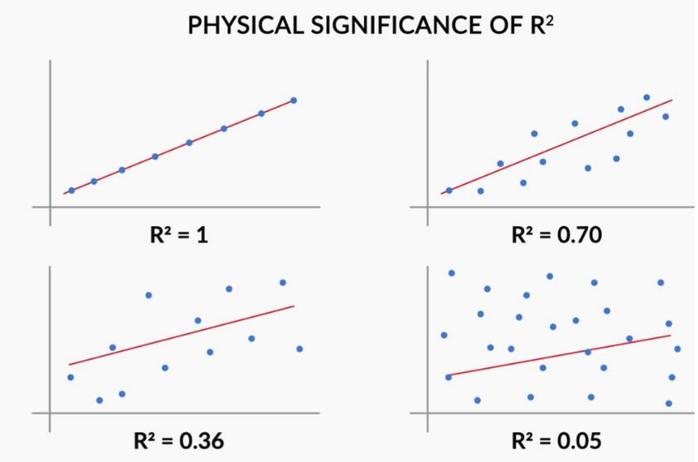
R(squared) is typically used to measure the goodness of a linear regression model given by the below formula. Higher the R(Squared), better is the model fit.



$$R^2=1-\frac{RSS}{TSS}$$

Where
RSS - Residual sum of squares
TSS - Total sum of squares

Physical significance of R(squared) is given by the below diagram:



We can see that more the data points are clustered towards the straight line, better is the R(squared) value i.e. model is a better fit as the R(squared) value approaches towards 1.

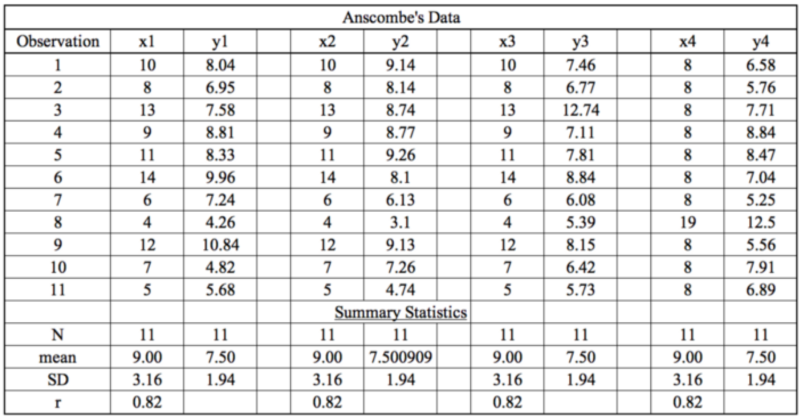
Gradient descent Gradient Descent is an optimization algorithm which optimizes the objective function (for linear regression it's cost function) to reach to the optimal solution. In case of linear regression the Cost function is the RSS which we want to minimize to find out the optimal values of the slope and intercept of the straight line.
Ways to minimize cost function:

- Differentiation – Differentiate the cost function -> equate it to zero -> Solve the equations and get the slope and intercept.
- Iterative method i.e. gradient descent – Start with an initial values of the slope and intercept and then iteratively improve on the values till we reach an optimal value.

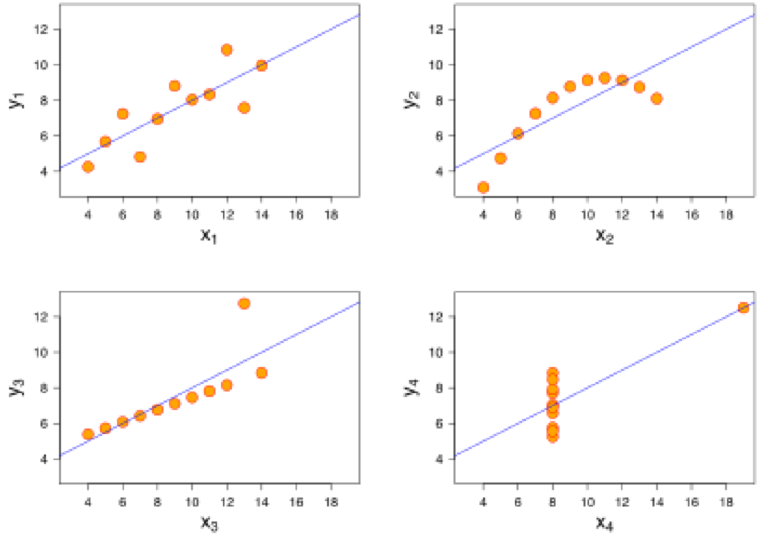
2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s Quartet can be defined as a group of four intentionally prepared data sets to explain the importance of data visualization. The data in these datasets are very much identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It tells us about the significance of visualizing the data before applying various algorithms 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. The dataset and the plots are given below:



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	

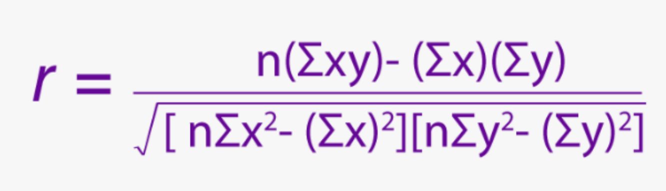


The four datasets can be described as:

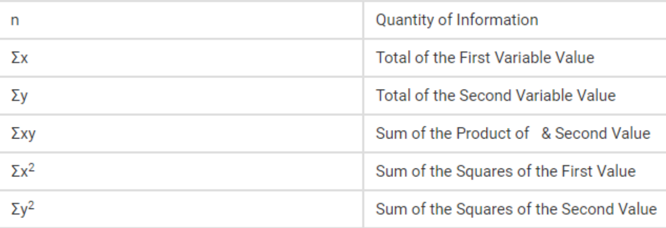
- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson’s R? (3 marks)

Pearson’s R is the correlation coefficient which is used to measure correlation between two numerical variables i.e. how strongly the two numerical variables are related. The formula for finding the Pearson coefficient is:



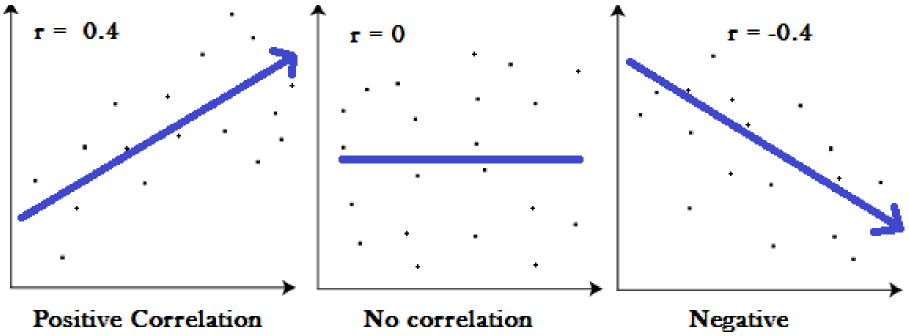
$$r = \frac{n(\Sigma xy)- (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2- (\Sigma x)^2][n\Sigma y^2- (\Sigma y)^2]}}$$



n	Quantity of Information
Σx	Total of the First Variable Value
Σy	Total of the Second Variable Value
Σxy	Sum of the Product of & Second Value
Σx^2	Sum of the Squares of the First Value
Σy^2	Sum of the Squares of the Second Value

It returns a value between -1 and +1, the interpretation of which is as follows:

- 1 indicates a strong positive relationship; a positive value indicates that as one variable increases the other will also increase by a fixed proportion.
- -1 indicates a strong negative relationship; a negative value indicates that as one variable increases the other will also decrease by a fixed proportion.
- A result of zero indicates no relationship at all.



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

Scaling is a data pre-processing step in the linear regression to ensure all the variables are on the same scale. It helps to converge the gradient descent function at the backend of the linear regression algorithm faster, thereby helping to improve the processing speed. If the variables are not scaled to a comparable scale then the interpretation of the coefficients would be very difficult i.e. variables having large range may have small coefficients while variables having small range will have large coeffcients which can result in incorrect interpretation of the coefficients.

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.

Difference between Min-max scaling vs Standardized scaling

Min-max (Normalization)	**Standardized scaling**
Compress all the data points in the range 0 and 1	Replaces the data points with their Z scores, with zero mean and Standard deviation of one
MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$	Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
	Doesn't compress the data between a range which is sometimes useful if there are outliers

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

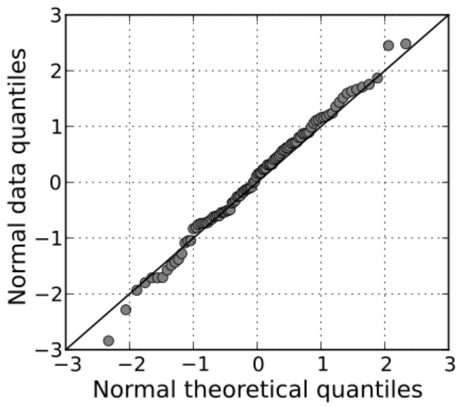
We know that VIF formula is = 1/(1-R2) which means VIF is Infinity when the R2 = 1, i.e. the variable (in question) has a perfect R2, and the variance in the concerned variable is 100% explained by the other variables in the model. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

PS – The variable is not the target variable, but the variable for which the VIF is being calculated.

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

A Q-Q (Quantile –Quantile) plot in statistics is a graph where we plot theoretical quantiles i.e. standard normal values (a normal distribution with mean = 0 and std deviation =1) on X axis and actual data points in question on Y axis. If the curve results in perfct straight line Y=X, then we can say that the data in question is normally distributed or not. Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.



Use and importance in linear regression:

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.