

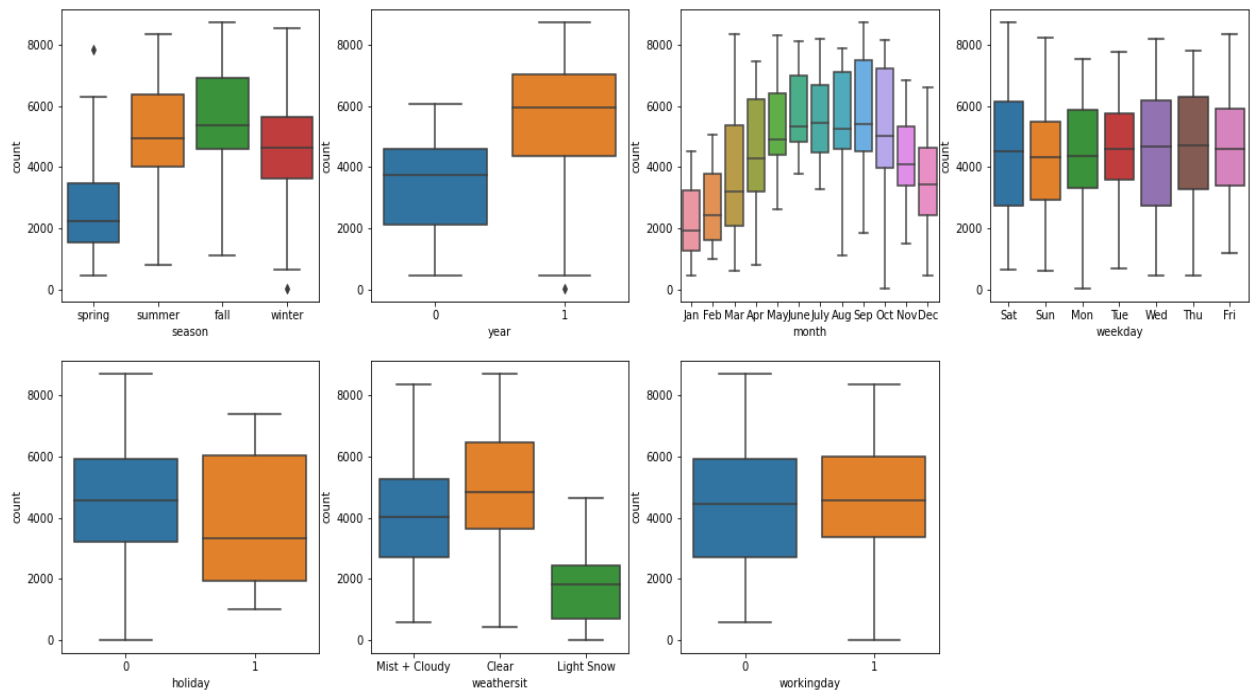
Assignment-based Subjective Questions

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: Categorical variables in the dataset are

1. season
2. year
3. month
4. weekday
5. holiday
6. weathersit
7. workingday

Box plots between different categorical variables vs count (dependent/target variabel)



Inference:

1. Bike rentals are more during fall season and then in summer.
2. Bike rentals in 2019 is more than in 2018. This might be due to bike sharing system is gaining popularity year after year.
3. Bike rentals are high in May to October which again falls under Fall and summer season.
4. Bike rentals demand is more on non-holidays. It might be due to people will commute more on non-holidays to offices, schools, universities etc.
5. Bike rentals is more on Clear, Few clouds, Partly cloudy, Partly cloudy days. And as expected there are no bikerentals on Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog days.
6. Working day or holiday doesn't seem to have much effect on bike rentals.

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

Ans: pandas.get_dummies() is used for data manipulation. It is used to convert categorical data into dummy variables or indicator columns (columns of 0's and 1's). For `n` categorical levels we need n-1 dummy variables.

For example in our dataset, we have `season` categorical variable which can take 4 values i.e. `spring`, `summer`, `fall` and `winter`.

Without setting drop_first to True while creting dummies, we will be getting four dummy columns each for `spring`, `summer`, `fall` and `winter` respectively.

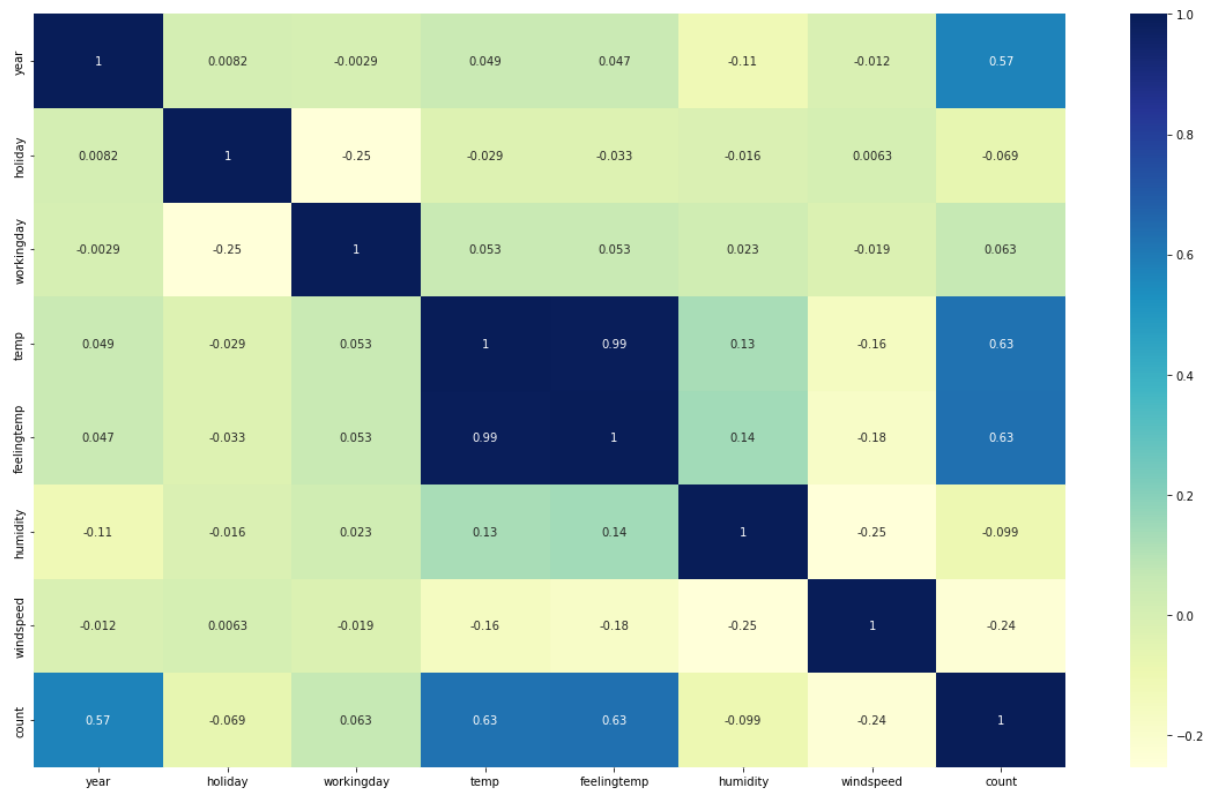
Now, if we take an example row where `spring` has value 1 and all other dummy columns has value `0`.

spring	summer	fall	winter
1	0	0	0

Here we can clearly notice that we don't need `spring` column since when all other column values are 0, the value of `spring` should be 1. Likewise, this column is a redundant and in a way adds introduces correlation with other columns. So in order to avoid this, setting drop_first = True, removes first level to get n-1 dummies out of n categorical levels.

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

Ans:

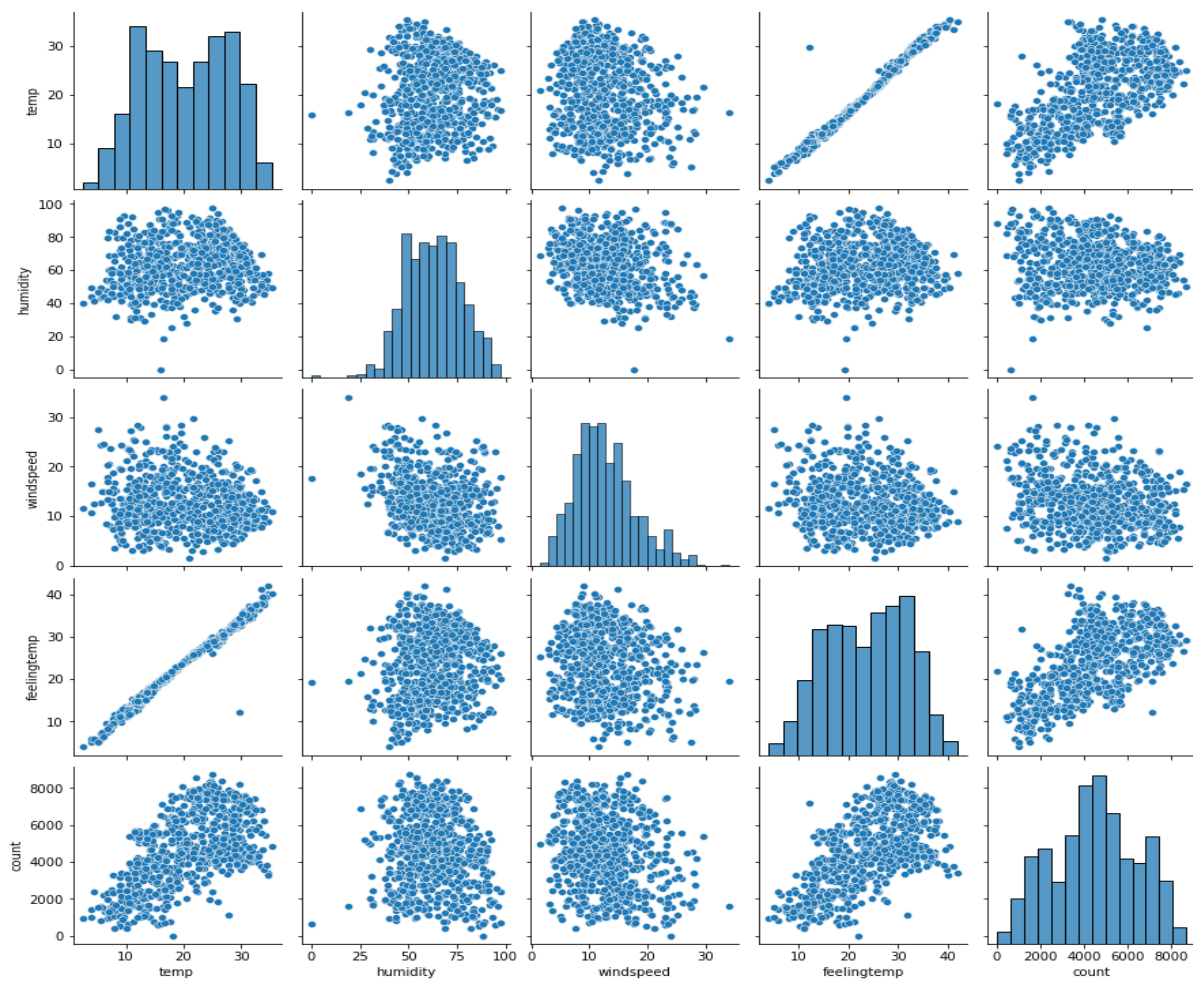


`temp` and `atemp` (which is rename to feelingtemp) has very strong correlation with `count` which is a target variable.

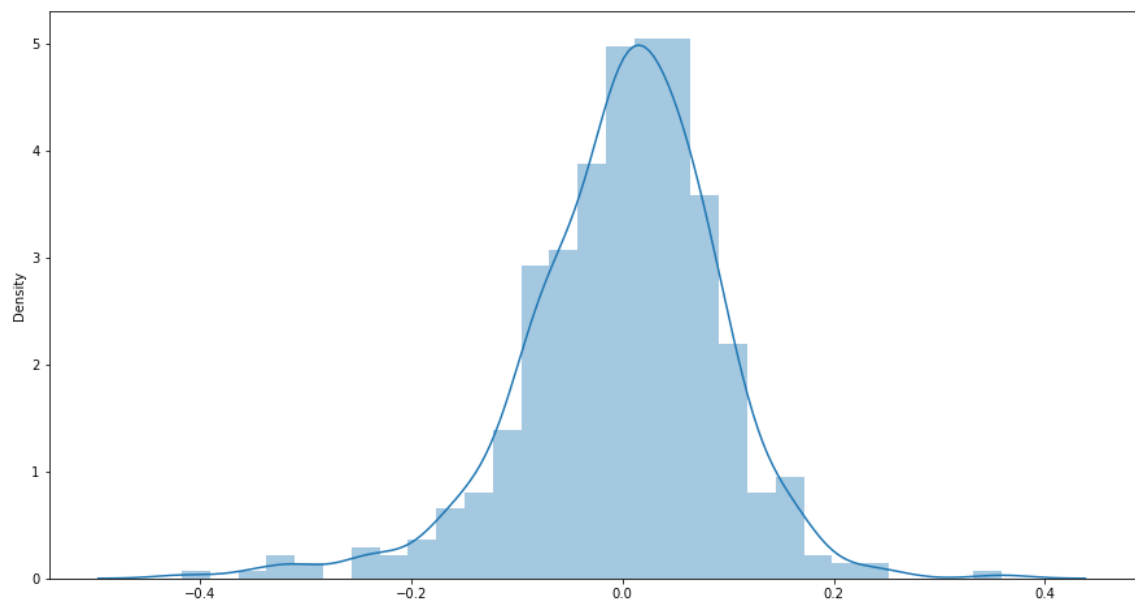
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: There are 5 assumptions in Linear regression

1. Linear relationship: Linear assumption validation is done through scatter plots between independent and dependent variables. From the below graph we can notice that there is clear linear relationship between dependent variables `temp`, `feelingtemp` and `windspeed` with the target variable.



2. Normality of the Error Terms: From the below plot we can clearly notice that error terms are normally distributed with mean centered at 0.

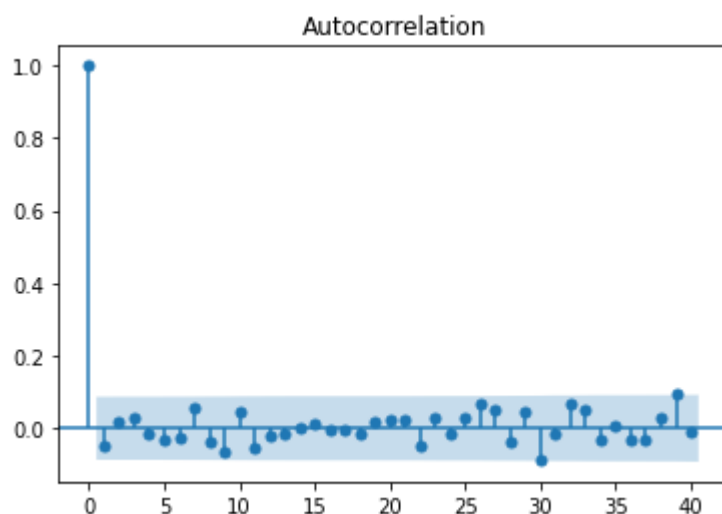


3. Zero or little multicollinearity: There should be zero or very negligible collinearity between predictor variables. This can be checked using Correlation coefficient or by checking VIF of predictor variables.

	Features	VIF
9	windspeed	4.60
8	temp	3.84
6	year	2.07
3	spring	1.99
4	summer	1.90
5	winter	1.63
2	Mist + Cloudy	1.55
0	Sep	1.23
1	Light Snow	1.08
7	holiday	1.04

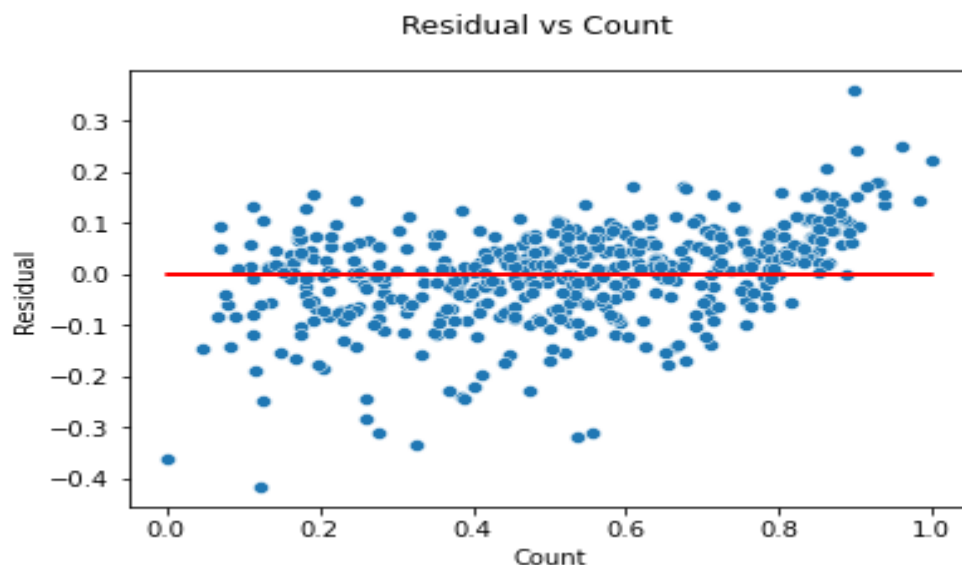
Since VIF of all final predictor variables is less than 5, we can say that there is very less or negligibly zero multicollinearity between variables.

4. No auto-correlation: Auto-correlation will check whether there is any patterns between the errors or not i.e error depending on y value or previous error values or not. Since there are no much error components crossing the greyed out area (confidence interval) and hence we can say that there is no pattern in the error and hence we can say No auto-correlation assumption has been preserved.



5. Homoscedasticity: Homoscedasticity means that the error is constant along the values of the dependent variable. Here, we have created scatterplot with the

residuals against the dependent variable. We can clearly observe that there is a constant deviation from the zero line and hence we can conclude our assumption of Homoscedasticity valid true.



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: Top 3 features contributing significantly for demand of shared bikes are:

1. `temp` with coefficient of .4777
2. `weathersit - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds` with coefficient of -0.2850.
3. `yr` with coefficient of 0.2341

Linear Regression Assignment Subjective Questions

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

Ans: Linear regression is a Supervised Machine Learning algorithm that finds the best linear-fit relationship between independent and dependent variables on a given data.

Target variable is known as independent variable or label and input features are known as dependent variables. Linear regression is used for prediction and forecasting.

Assumptions in a Linear Regression:

- It is assumed that there is a linear relationship between dependent and independent variables.
- It is assumed that the error terms are normally distributed i.e. residuals have a mean value of zero.

- It is assumed that residuals have the constant variance at each level of the predicted values. This assumption is called homogeneity or homoscedasticity.
- It is assumed that residuals terms are independent of each other which means co-variance between residuals terms is 0.
- It is assumed that independent variables are independent of each other. There is no multicollinearity in the data.

There are two types of Linear Regression models:

1. Simple Linear Regression: Here we aim to reveal relationship between single independent variable and corresponding dependent/target variable. This can be expressed in a string line as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- a. Y represents the output or dependent variable.
 - b. β_0 and β_1 are two unknown constants that represent the intercept and coefficient (slope) respectively.
 - c. ϵ (Epsilon) is the error term.
2. Multiple Linear Regression: Here we aim to reveal relationship between 2 or more independent variables and corresponding dependent/target variable. The independent variables can be continuous or categorical.

The diagram shows the equation $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$ with the following annotations:

- A red circle around Y is labeled "response, dependent variable, observation, 'y-variable'" with a red arrow pointing to it.
- A green circle around x_1 is labeled "predictor, 'x-variable', independent variable, explanatory variable" with a green arrow pointing to it.
- An orange circle around β_2 is labeled "coefficient" with an orange arrow pointing to it.
- A blue bracket under the terms $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$ is labeled "linear predictor" with a blue arrow pointing to it.
- A purple circle around ϵ is labeled "random error, 'noise'" with a purple arrow pointing to it.

Source:

<https://medium.datadriveninvestor.com/types-of-linear-regression-89f3bef3a0c7>

Shortcomings of linear regression:

1. Linear Regression is sensitive to outliers.
2. It models only linear relationships.
3. Few assumptions are required in order to make an inference.

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

Ans: Anscombe's quartet is defined as a group of 4 datasets which has nearly identical simple statistical properties. Anscombe's quartet emphasises the importance of plotting and visualising data before building a model.

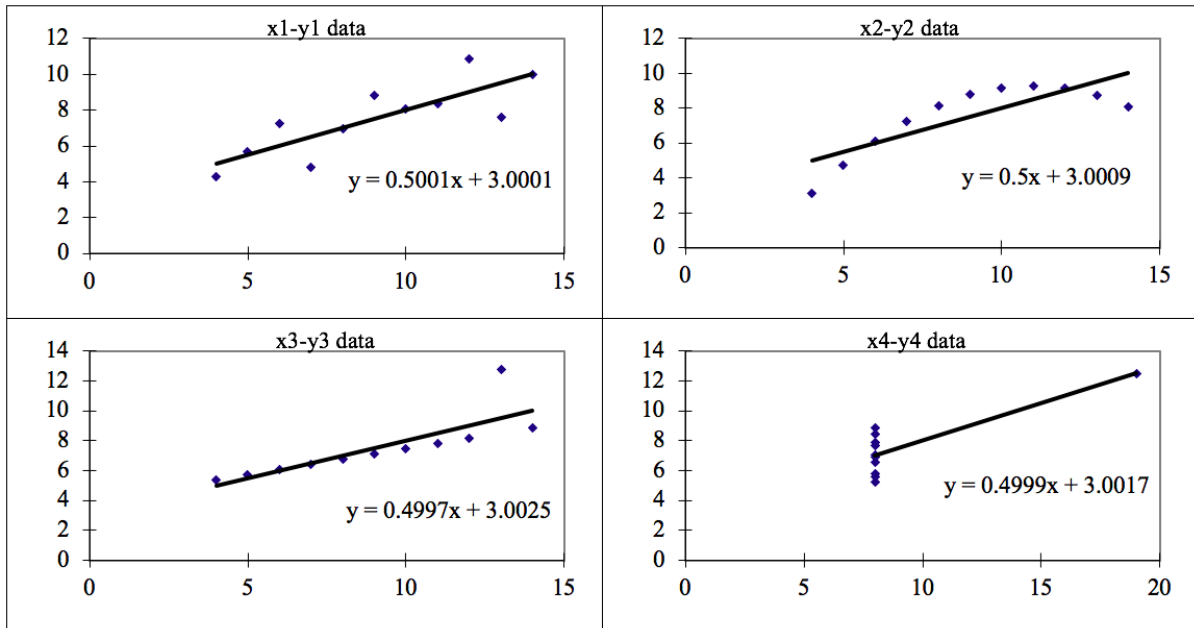
Below are the 4 datasets and its statistics.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary						
Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)	
1	9	3.32	7.5	2.03	0.816	
2	9	3.32	7.5	2.03	0.816	
3	9	3.32	7.5	2.03	0.816	
4	9	3.32	7.5	2.03	0.817	

Source: <https://www.geeksforgeeks.org/anscombes-quartet/>

But when these datasets are plotted on a scatter plot, they generate different kind of plots which cannot be interpreted by any regression algorithms.



Source:

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

1. 1st dataset (top left) fits the linear regression model as we it seems to be clear linear relationship exists between x and y
2. In 2nd dataset (top right), there is a non-linear relationship between x and y and hence this could not fit linear regression model.
3. In 3rd dataset (bottom left), there is a outlier involved in the dataset which cannot be handled by the linear regression model.
4. In 4th dataset (bottom right), we can observe that high leverage point is enough to produce a high correlation coefficient.

From this we can understand the importance of data visualisation and how regression algorithm can be fooled by the same. It is important to visualise data before building any ML model.

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

Ans: Correlation coefficient formulas are used to find strength of a relationship is between data. There are several types of correlation coefficients and Pearson's correlation coefficient is the most commonly used one.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

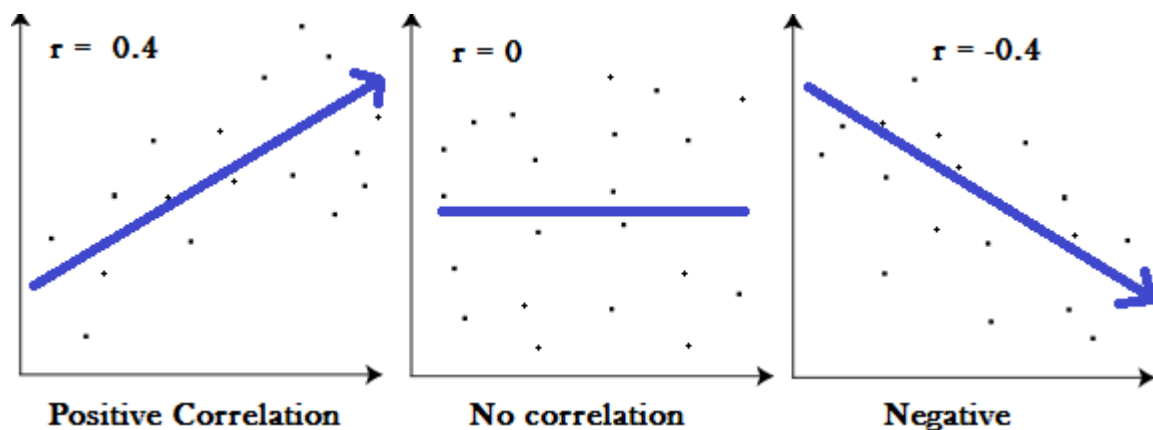
x_i = values of the x-variable in a sample

\bar{x} = mean of the values of the x-variable

y_i = values of the y-variable in a sample

\bar{y} = mean of the values of the y-variable

Pearson's R coefficient value ranges between -1 and 1. Negative value indicates negative correlation i.e. increase in one variable will cause decrease in other variable and vice versa. Positive value indicates positive correlation i.e. increase in one variable will increase the other variable and vice versa. 0 indicates there is no relation between variables and +/-1 indicates the perfect correlation.



Source:

<https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/>

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

Ans: It is one of the step in data pre-processing which is applied to independent variables to normalize the data within a particular range. This helps in speeding up the some calculations in algorithms like gradient descent. Some distance algorithms like K-Means, KNN and SVM are most effected by the range of features.

In general in datasets we will have different featuring highly varying in magnitudes and ranges. Since algorithms consider just magnitudes, we may end up in constructing incorrect model when scaling is not performed. And hence, scaling is performed in order to bring all the features to the same level of magnitude.

And also scaling doesn't effect any of the statistical properties, it just affects the coefficients.

Difference between Normalized scaling and Standardized scaling:

Normalized Scaling / Min-Max Scaling:

This is used to bring all the data in the range of 0 and 1 using the maximum and minimum value of the feature.

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

Standardized scaling:

Standardized scaling replaces the values by their z-scores. The data is scaled in such a way that its mean is zero and standard deviation is 1.

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

The advantage of standardisation over other is it doesn't compress the data to a particular range like min-max scaling. Hence standardisation is preferred over normalization when there are outliers in the data.

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

Ans: Variance Inflation Factor (VIF) is used to calculate how well one independent variable can be explained by all other independent variables combined.

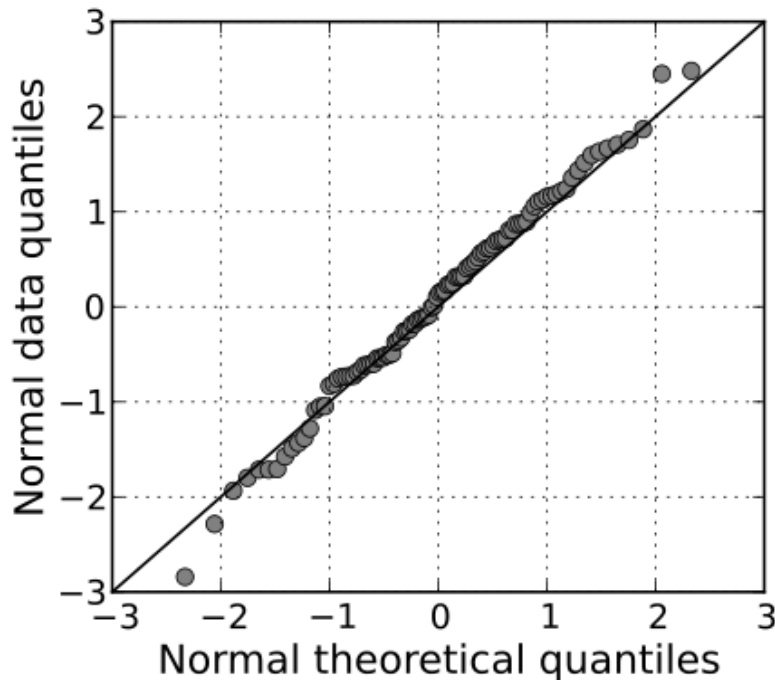
$$VIF_i = \frac{1}{1 - R_i^2}$$

Where i refers to the ith variable which is represented as a linear combination of rest of the independent variables.

When VIF is infinity, it means $1 - R_i^2 = 0 \Rightarrow R_i^2 = 1$. It means there is a perfect correlation and this variable can be expressed as a linear combination of other independent variables. In this case, we need to drop this variable in order to avoid multicollinearity.

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

Ans: Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. This plays a vital role in comparing two probability distributions. If the distributions of the plots that are in comparison are exactly equal then points on the Q-Q plot will perfectly lie on $y=x$ (45-degree reference line). Q-Q plot for normal distribution.



Source:

https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot#/media/File:Normal_normal_qq.svg

Importance of Q-Q plot in Linear regression:

In linear regression, when we have training and test dataset received separately from different sources then from the Q-Q plot, we can confirm that both the datasets are from the populations with same distributions.

Advantages:

It is used to check following scenarios:

1. If the datasets come from the populations with a common distribution.
2. If the datasets have common location and scale.
3. If the datasets have similar distribution scales.
4. If the datasets have similar tail behaviour.
5. It can be used in varying sample sizes also.