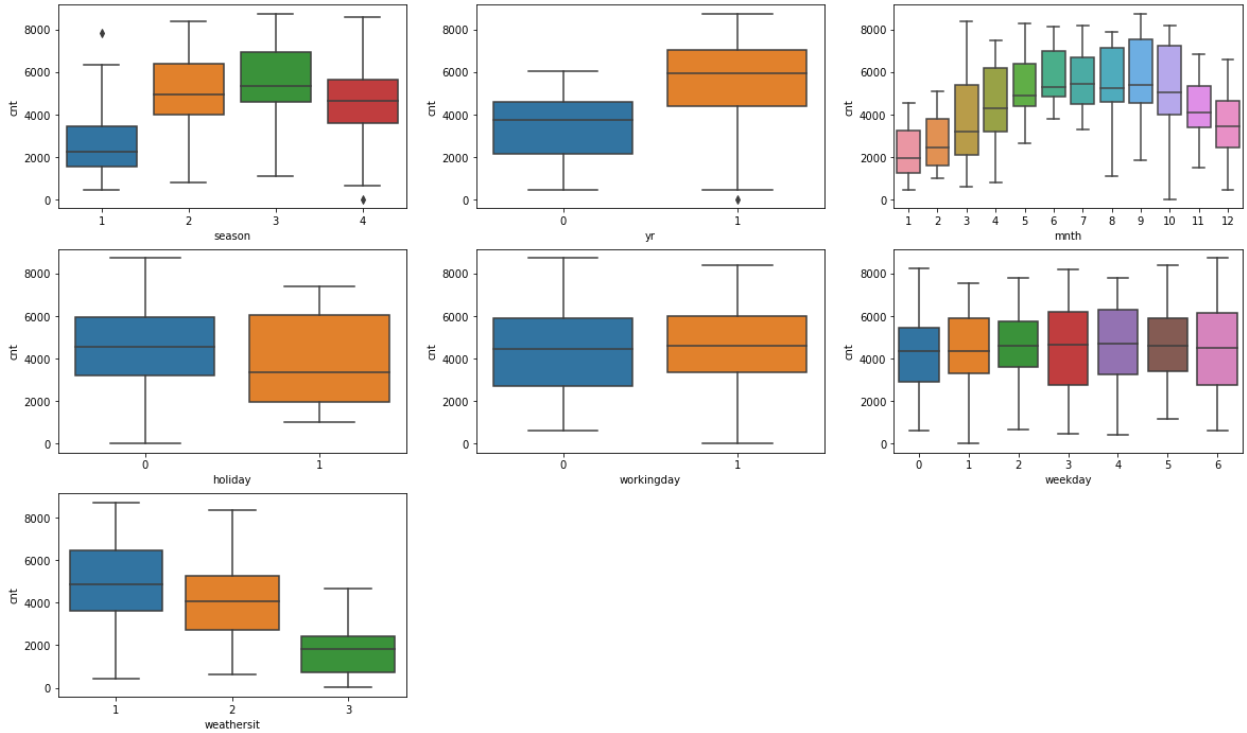
Bike Sharing Assignment

Nikhil Barigidad

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)



As we can see above totally there are seven categorical variables in Bike Sharing Dataset,

- **Season:** Most bookings are done in season 3 with a median of more than 5000, followed by seasons 2, 3 and 4
- **yr:** Bookings have been increased drastically over years (>45%)
- **mnth:** Bookings are increasing from month 1 to 6, but then from 7 to 12 they tend to decrease
- **holiday:** Bookings are more in when there is no holiday, the distribution is as biased as almost 97.6% of bikes were booked when there is no holiday
- **workingday:** Most bikes were booked on a working day
- **weekday:** Trend between all weekdays is very close, So this I think won't be influencing much in prediction

- **weathersit:** As we can see weathersit 1 has the highest cnt, During weathersit 1, 67% of the bikes were booked, followed by weathersit 2 with only 30% bikes booked.

However, the categorical variables that are contributing to the final model are

- season_2
- season_4
- yr_1
- mnth_8
- mnth_9
- mnth_10
- weekday_6
- workingday_1
- weathersit_2
- weathersit_3

As we can see holiday is not contributing to the final model, and out of all the values of categorical variables(all dummy variables) we can infer that not all of them are considered in the final model

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

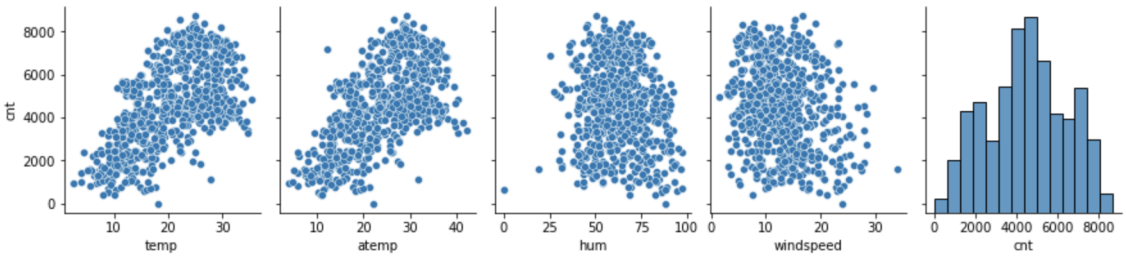
When creating dummy variables using *get_dummies* we pass an argument *drop_first=True* as shown below

```
pd.get_dummies(bikeSharing_cleaned, drop_first=True)
```

Passing this in get_dummies helps us removes an extra column created during dummy creation, From a categorical variable with 'n' unique values, (n-1) dummy values are required, but with the argument drop_first = False, n variable will be created so to avoid the creation of an extra column this argument is used.

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

Following are the pair plots for all the numerical variables with cnt(dependent variable)



From the above pair plots, temp and atemp seems to be highly correlated to the cnt variable

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

Assumptions of Linear Regression:

"The Dependent variable and Independent variable must have a linear relationship"

We can see in the above pair plot(from the previous question) that the variables show a somewhat linear relationship with the cnt(dependent) variable. Although hum and windspeed may show linear relation as good as temp and atemp

"No Perfect Multicollinearity."

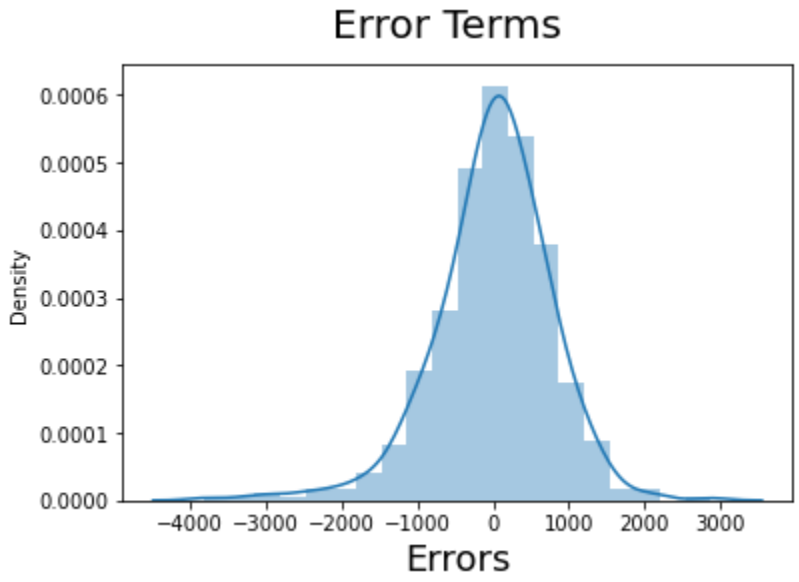
	Variables	VIF
0	temp	7.41
1	windspeed	4.74
9	workingday_1	4.30
4	yr_1	2.04
3	season_4	2.01
2	season_2	1.83
8	weekday_6	1.75
7	mnth_10	1.61
5	mnth_8	1.58
10	weathersit_2	1.55
6	mnth_9	1.33
11	weathersit_3	1.10

Multicollinearity, Dependency of Independent variable on other independent variables can be calculated through VIF(Variance Inflation Factor)

- If VIF=1, Very Less Multicollinearity
- VIF<5, Moderate Multicollinearity
- VIF>5, Extreme Multicollinearity (This is what we have to avoid)

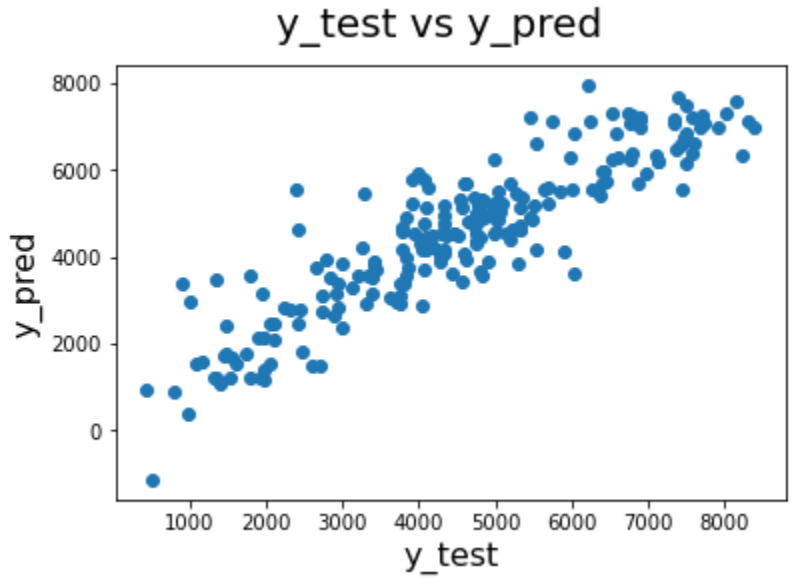
Although the `temp` variable has VIF > 5, intuitively temp variable should affect the count of bike bookings, and also temp has a very high correlation with the dependent variable cnt, thus it has not been removed from the list.

"Error terms are normally distributed with mean zero"



As we can see in the distribution plot on the left, Error terms are normally distributed with a mean of 0, for the Bike Sharing dataset

"Error terms have constant variance (homoscedasticity)"

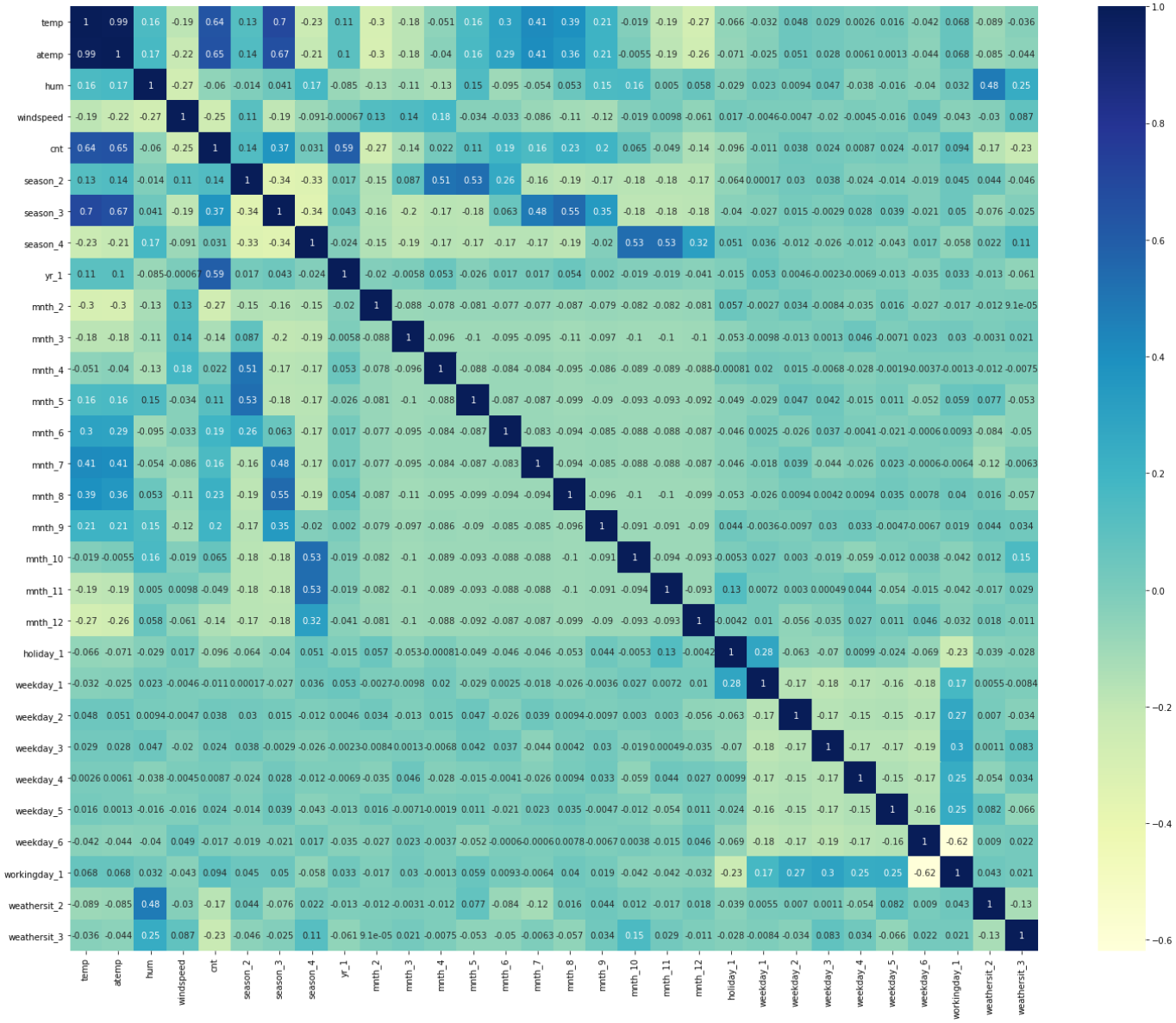


As we can see error term is constant throughout the distribution, hence this validates the Homoscedasticity

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

Based on the heatmap and the final features selected based on p-value and VIF, Top 3 features of the model are **temp, yr_1, windspeed**

Reference heatmap of all variables



Correlation of top 3 features are

temp	0.64
yr_1	0.59
windspeed	-0.25

General Subjective Questions

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

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task, i.e prediction based on the independent variables present in the dataset.

Mathematically, Simple Linear Regression can be represented as shown below

$$y = C + \beta * X$$

Multiple Linear Regression can be represented as shown below

$$y = C + \beta_0 * X_0 + \beta_1 * X_1 + \ldots\ldots + \beta_n * X_n$$

Here X and y are two variables of linear regression,

X is an independent variable from the database

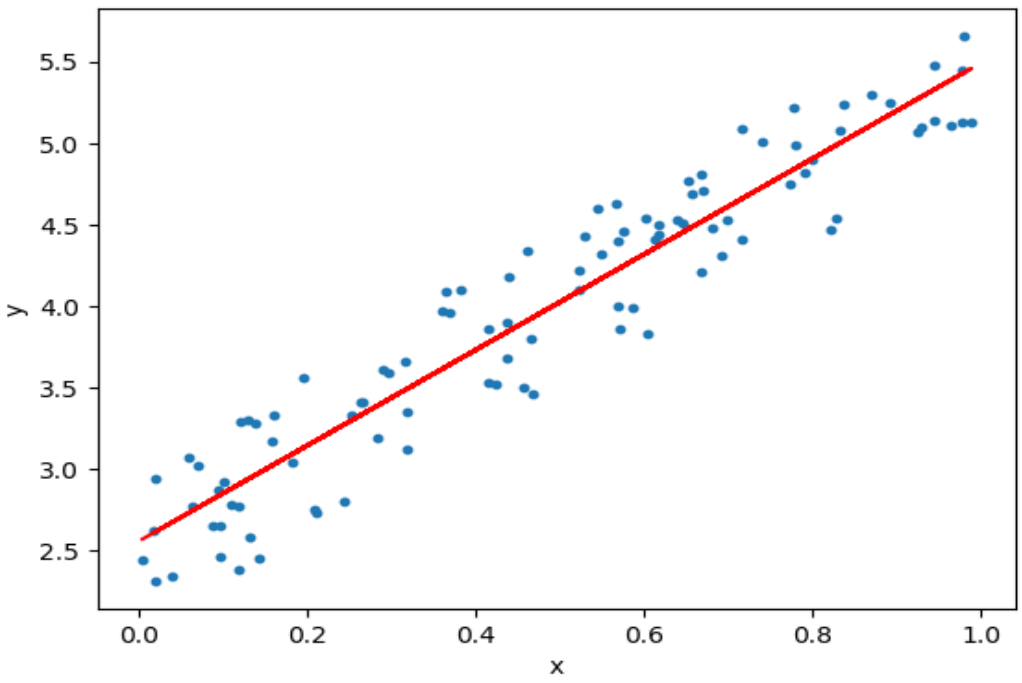
y is a dependent variable from the database

C is the intercept of the predictor line

βis the coefficient of the independent variable

Through the best fit line, we can describe the impact of change in independent variables on the dependent variable.

The red line below is the best fit line for the dataset.



Hypothesis Testing in Linear Regression:

- Null Hypothesis: (H_0): Coefficients of Linear Equation are equal to zero, i.e coefficients are not significant
- Alternate Hypothesis: (H_1): At Least one coefficient of Linear Equation should not be equal to zero

Assumptions associated with Linear Regression model:

1. Linearity: There is a linear relationship between X and Y
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of an independent variable, the dependent variable is normally distributed.

When moving from Simple Linear Regression to Multiple Linear Regression, additional aspects have to be considered

1. Overfitting: When adding more variables to the relation, the model becomes too complex and ends up memorizing the data resulting in very high accuracy on the training set and very low on the test set
2. Multicollinearity: Association between predictor variables, This is measured using VIF(Variance Inflation Factor)
3. Feature Selection: Among all the features provided in the dataset, Selecting optimal features that highly impact the model to obtain better accuracy.

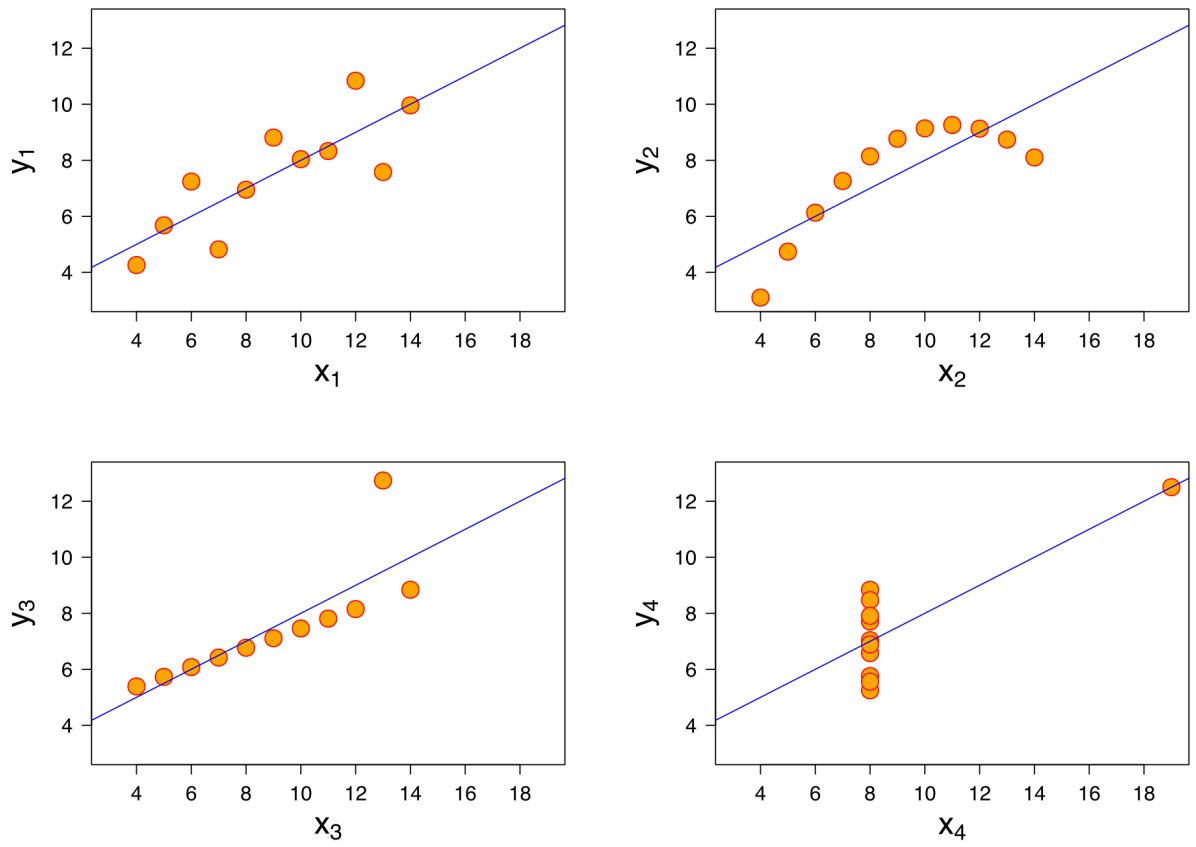
Model Evaluation of Linear Equation:

A model can be evaluated using the below methods

- Mean Absolute Error
- Mean Squared Error
- Root Mean Squared Error

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

According to Wikipedia, **Anscombe's quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modeled as gaussian with mean linearly dependent on *x*.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is used to illustrate the importance of visualizing datasets before starting to analyze for a particular type of relationship.
The dataset comprised in **Anscombe's quartet** are:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89

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

Correlation between different data sets is measured by how well they are related. The most commonly used correlational measure is Pearson's R correlation. It is a static that is used to measure the linear correlation between two variables. The value of Pearson's R is between -1 and 1.

The value of Pearson's R is calculated using the formula(shown below) proposed by Karl Pearson

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

Where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores

Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

The positive value of the Pearson correlation implies that if we change either of these variables, there will be a positive effect on the other. For example, if a person's age is increased, income will also increase. Likewise, when Pearson correlation is negative there will be a negative effect on the other variable.

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

Scaling is a technique to standardize the independent features present in the dataset in a fixed range. If scaling is not performed then the machine learning model will tend to weigh variables with higher values and the model will ignore variables with smaller values.

Techniques to perform feature scaling

1. **Min-Max Normalization**: This technique re-scales the feature values with distribution between 0 and 1

$$X_{new} = \frac{X_i - min(X)}{max(X) - min(X)}$$

2. **Standardization:** This technique re-scales the feature values with a mean of distribution equal to 0 and standard deviation as 1

$$X_{new} = \frac{X_i - X_{mean}}{Standard\ Deviation}$$

The data in Min-Max scalar will be distributed within 0 and 1, whereas in the case of standardization the distribution will be around the mean of the distribution.

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

A Large value of VIF indicates there is a high correlation between the independent variables.

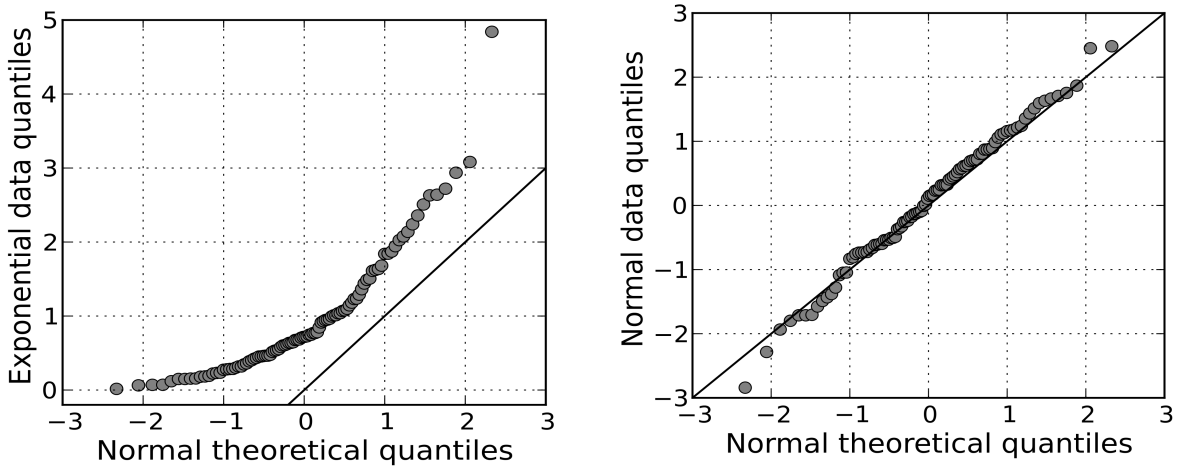
When there is a perfect correlation between variables then the VIF value will tend to infinity. When there is a perfect correlation between two variables we get $R^2 = 1$, thus $1/(1 - R^2)$ tending to infinity. To solve this we drop one of the variables from the dataset.

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

Q-Q plots(Quantile-Quantile) are plots of two quantiles against each other. It is a graphical method for comparing two probability distributions by plotting their quantiles against each other. First, the set of intervals for the quantiles is chosen. A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the point will lie on the line y=x, If the two distributions are not similar then the distribution will be scattered away from the line y=x

Example Q-Q plots



- The points on the graph on the left follow a strictly nonlinear pattern, suggesting that the data are not distributed as standard normal. The offset of the points in the graph suggests that the mean of the data is not equal to 0
- The points on the graph on the right lie on the line y=x, following a linear pattern, suggesting that the data are distributed as standard normal