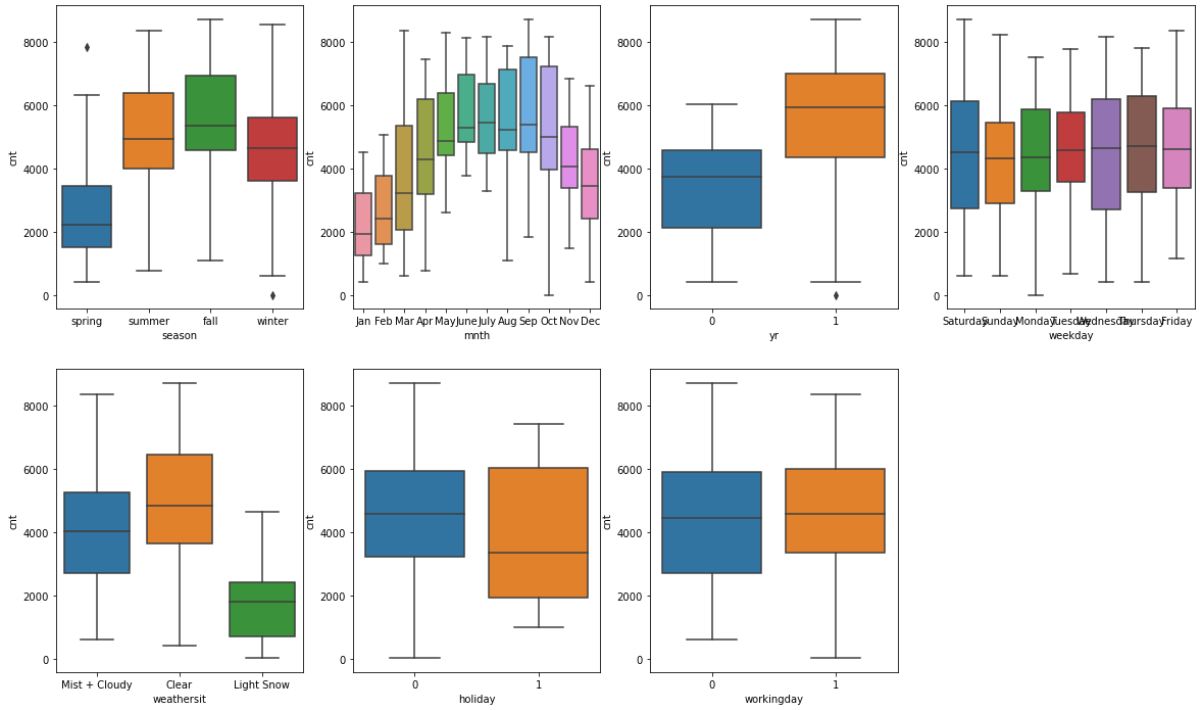
Assignment-based Subjective Questions

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



Observations from above boxplots for categorical variables:

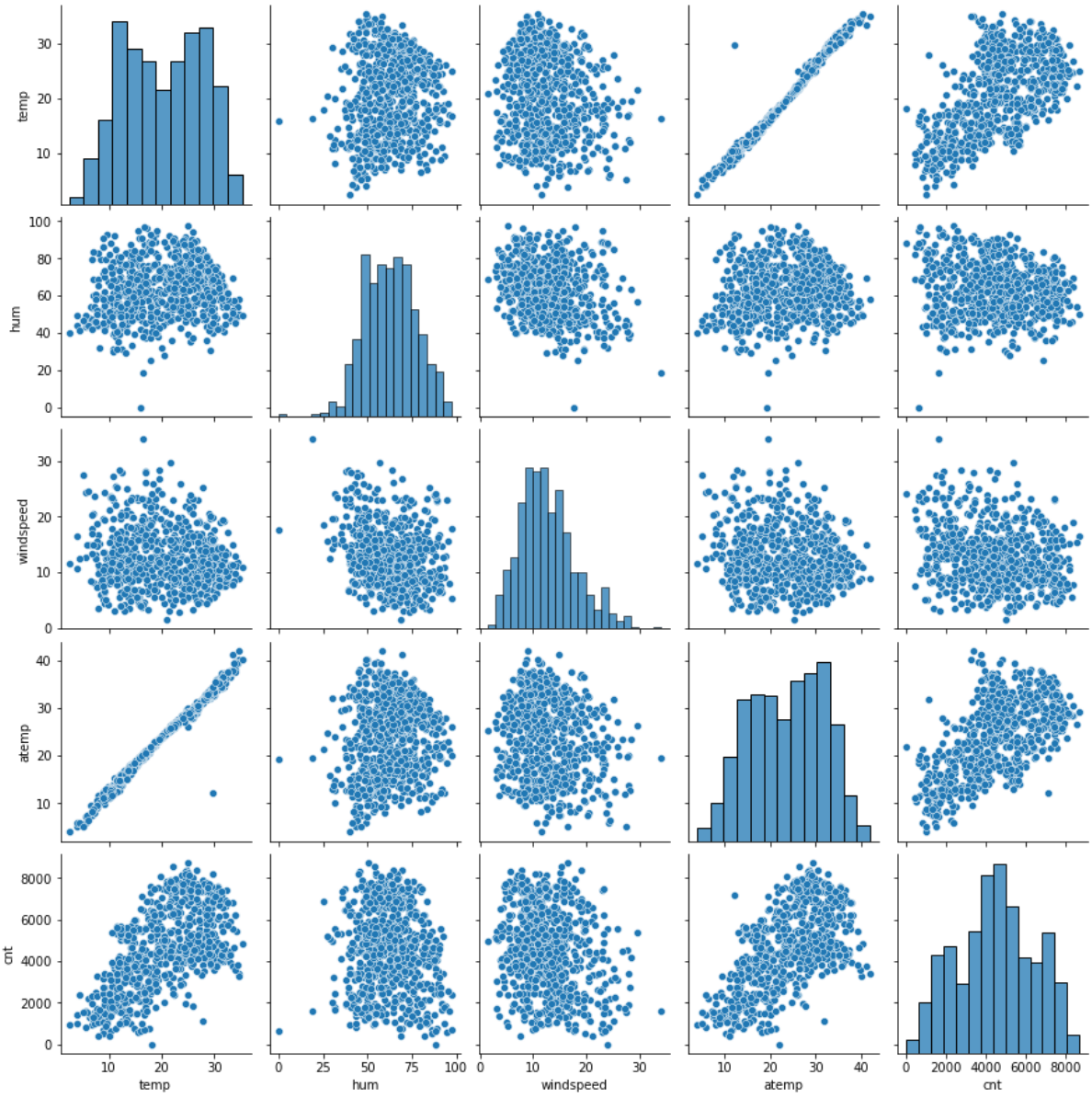
- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during September month.
- The weekday box plots indicates that more bikes are rent during saturday.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

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

The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.

Eg: If there are 3 levels, the drop_first will drop the first column.

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



The 'temp' and 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'.

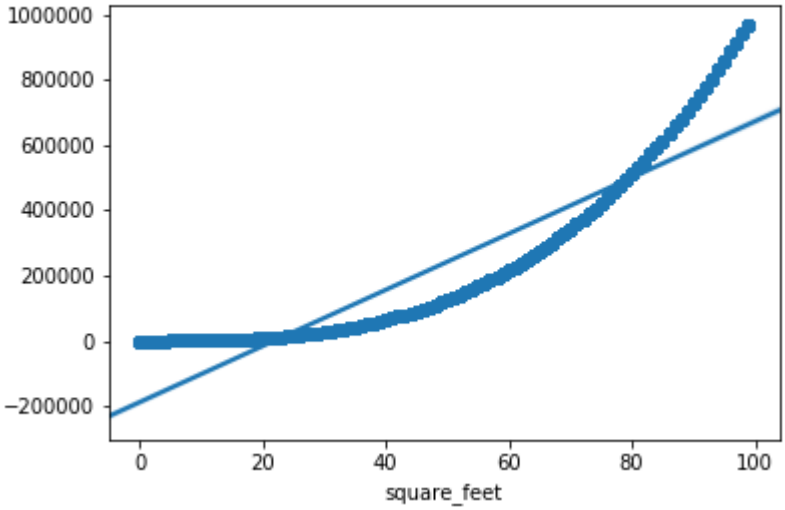
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1. Linear Relationship

As obvious as this may seem, linear regression assumes that there exists a linear relationship between the dependent variable and the predictors.

How can it be verified?

Pair-wise scatterplots may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.



In the example above, the relationship between both variables is clearly not linear

In addition and similarly, a partial residual plot that represents the relationship between a predictor and the dependent variable while taking into account all the other variables may help visualize the "true nature of the relationship" between variables.

$$PartialResidual = Residual + \hat{\beta}_i X_i$$

The formula behind a Partial Residual Plot

What could it mean for the model if it is not respected?

If linearity is not respected, the regression will underfit and will not accurately model the relationship between the dependent and the independent variables.

What could be done?

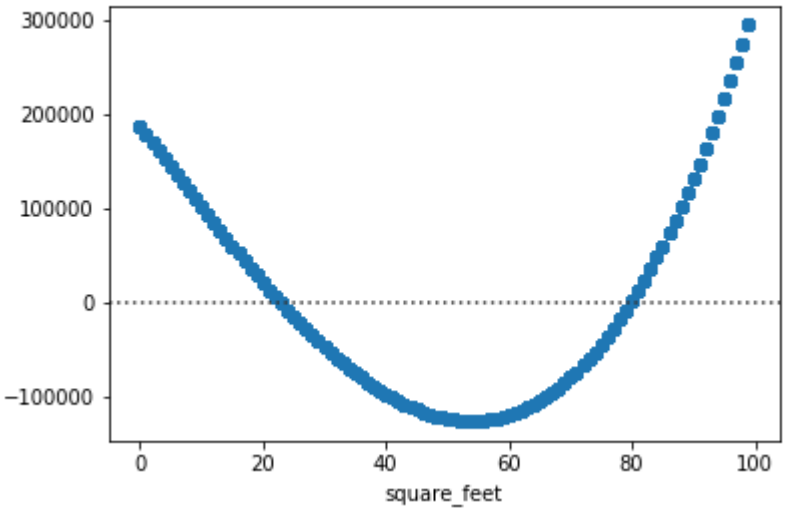
Independent variables and the dependent variables could be transformed so that the relationship between them is linear. For instance, you could find that the relationship is linear between the *log* of the dependent variables and some of the independent variables *squared* (c.f. *Polynomial Regression* and *Generalized Additive Models* (GAM) for an interesting generalization of this).

2. Homoscedasticity

Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.

How can it be verified?

To verify homoscedasticity, one may look at the residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.



In the example above, there is heteroscedasticity as the variance of the residual is not constant

What could it mean for the model if it is not respected?

In the case of heteroscedasticity, the model will not fit all parts of the model equally and it will lead to biased predictions. It also often means that confounding variables, important predictors, have been omitted (it could also be due to the fact that the linearity assumption is not respected). While for the predictive context of data science, this may not be of utmost importance, heteroscedasticity is relatively more important in the context of inference, regarding the interpretability of the coefficients.

What could be done?

As heteroscedasticity generally reflects the absence of confounding variables, it can be tackled by reviewing the predictors and providing additional independent variables (and maybe even check that the linearity assumption is respected as well).

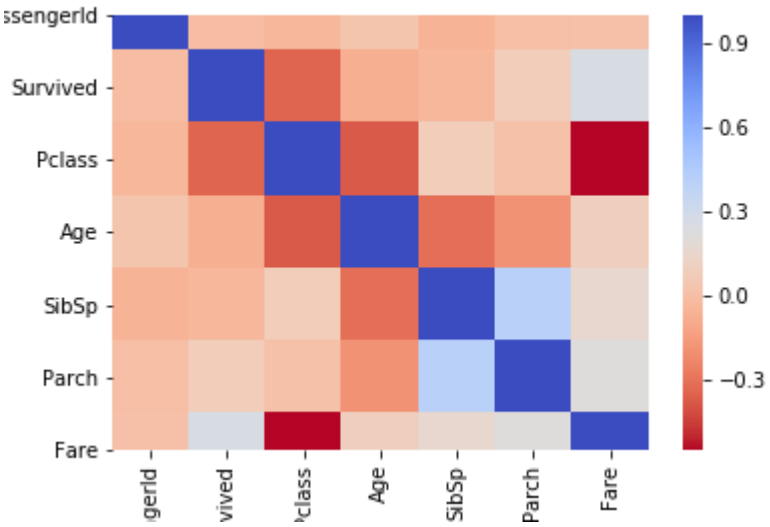
3. Absence of Multicollinearity

Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case). While it may not be important for non-parametric methods, it is primordial for parametric models such as linear regression.

How can it be verified?

Often, a tell-tale sign of multicollinearity is the fact that some of the estimated coefficients have the "wrong" sign (i.e. the coefficient related to the size of a being negative in a model attempting to predict house prices).

Pairwise correlations could be the first step to identify potential relationships between various independent variables.



A correlation heatmaps may allow to quickly notice pair-wise correlations

A more thorough method, however, would be to look at the Variance Inflation Factors (VIF). It is calculated by regressing each independent variable on all the others and calculating a score as follows:

$$VIF = \frac{1}{1 - R^2}$$

Formula to calculate the VIF of an independent variable

Hence, if there exists a linear relationship between an independent variable and the others, it will imply a large *R-squared* for the regression and thus a larger VIF. As a rule of thumb, VIFs scores above 5 are generally indicators of multicollinearity (above 10 it can definitely be considered an issue).

What could it mean for the model if it is not respected?

The model may be producing inaccurate coefficient estimates that could thus not be interpreted. It may thus hurt inference power and possibly predictive performance.

In the presence of multicollinearity, the regression's results may also become unstable and vary tremendously depending on the training data.

What could be done?

Multicollinearity can be fixed by performing feature selection: deleting one or more independent variables.

A common approach is to use backward subset-regression: start by building a regression with all the potential independent variables and iteratively remove variables with high VIF and using domain-specific knowledge.

Another method could be to isolate and keep only the interaction effects between multiple independent variables (using intuition or regularization generally).

As multicollinearity is reduced, the model will become more stable and the coefficients' interpretability will be improved.

4. Independence of residuals (absence of auto-correlation)

Autocorrelation refers to the fact that observations' errors are correlated.

How can it be verified?

To verify that the observations are not auto-correlated, we can use the **Durbin-Watson test**. The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables (0–2: positive auto-correlation, 2–4: negative auto-correlation).

What could it mean for the model if it is not respected?

Auto-correlation could mean that the linearity of the relationship is not respected or that variables may have been omitted.

Auto-correlation would lead to spurious relationships between the independent variables and the dependent variable.

What could be done?

For time-series, one could add a lag variable. Another potential way to tackle this is to modify the variables from absolute value to relative change (i.e. instead of a stock price, it could be the change percentage from one period to the next).

More generally, variables should be further fine-tuned and added to the model.

5. Normality of Errors

If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased.

How can it be verified?

To verify the normality of error, an easy way is to draw the distribution of residuals against levels of the dependent variable. One can use a QQ-plot and measure the divergence of the residuals from a normal distribution. If the resulting curve is not normal (i.e. is skewed), it may highlight a problem.

What could it mean for the model if it is not respected?

If it is not respected, it may highlight the presence of large outliers or highlight other assumptions being violated (i.e. linearity, homoscedasticity). As a result, calculating t-statistics and confidence intervals with the standard methodologies will become biased.

What could be done?

In the case where errors are not normally distributed, one could verify that the other assumptions are respected (i.e. homoscedasticity, linearity), as it may often be a tell-tale sign of such a violation, and fine-tune the model accordingly.

Otherwise, one should also attempt to treat the large outliers in the data and check if the data could not be separate subsets using different models.

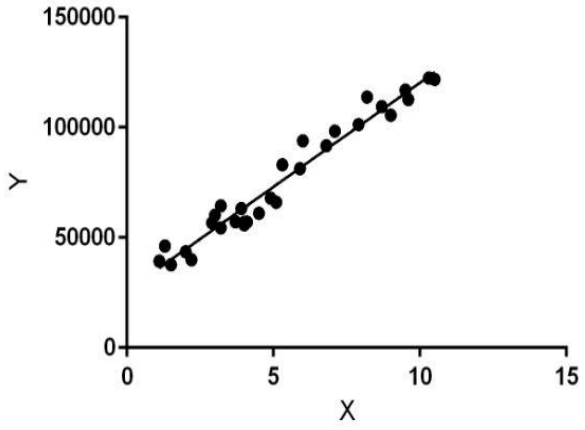
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1. Temperature (0.6009)
2. weathersit : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (-0.2504)
3. year (0.2328)

General Subjective Question

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**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

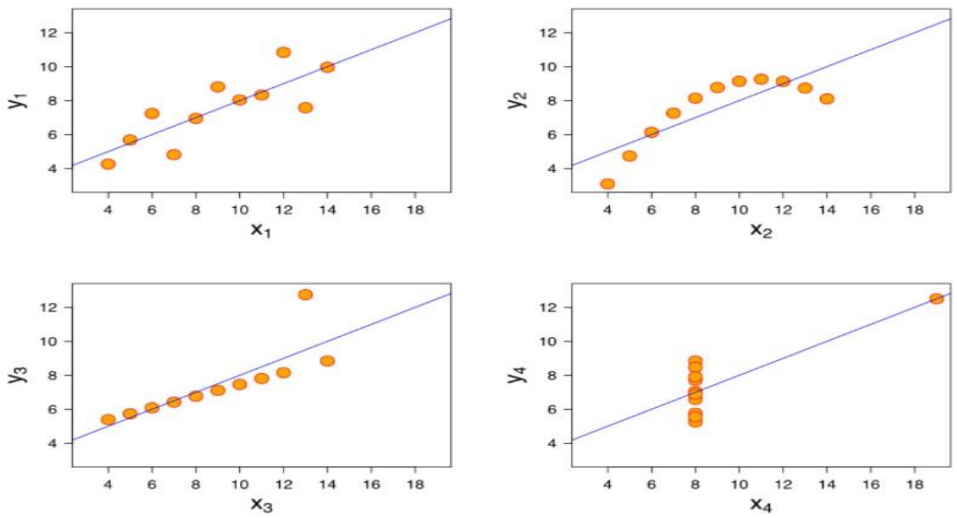
Hypothesis function for Linear Regression:

$$y = \theta_1 + \theta_2.x$$

While training the model we are given:
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x
Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly 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 was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



- 1 st data set fits linear regression model as it seems to be linear relationship

between X and y

- 2 nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
- 3 rd data set shows some outliers present in the dataset which can't be handled by a linear regression model.
- 4 th data set has a high leverage point means it produces a high correlation coeff.

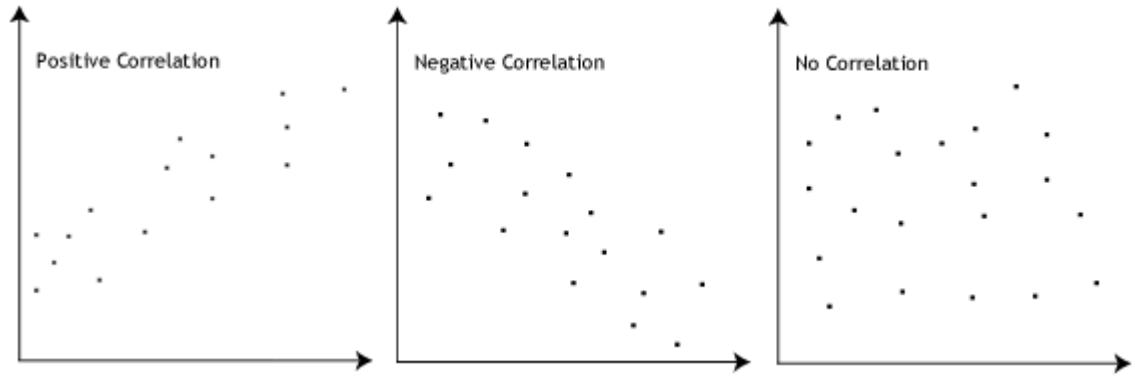
Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

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

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



Pearson r Formula

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

Here,

- 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

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

What?
It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why?
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.

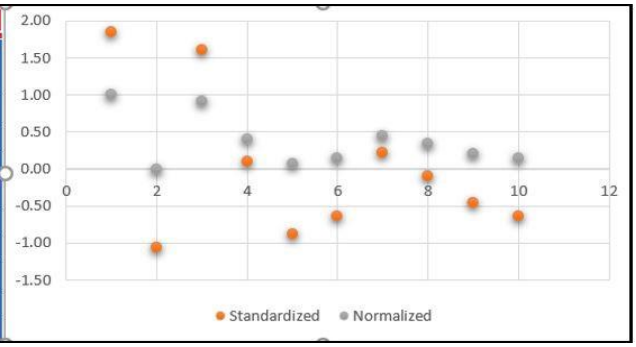
Normalization/Min-Max Scaling:
It brings all of the data in the range of 0 and 1. sklearn. preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.
Example:

Below shows example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

The value of VIF is calculated by the below formula:

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

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the

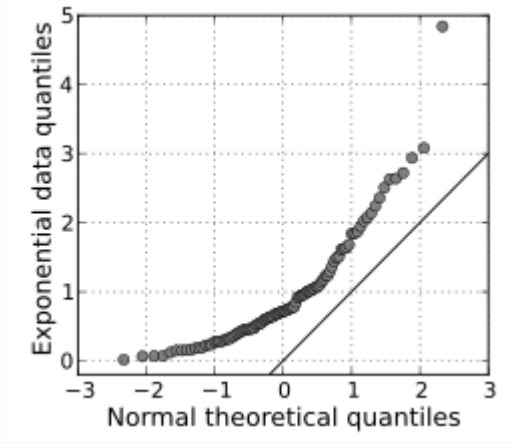
overall value become infinite. It denotes perfect correlation in variables.
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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 plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45-degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.