Assignment-based Subjective Questions

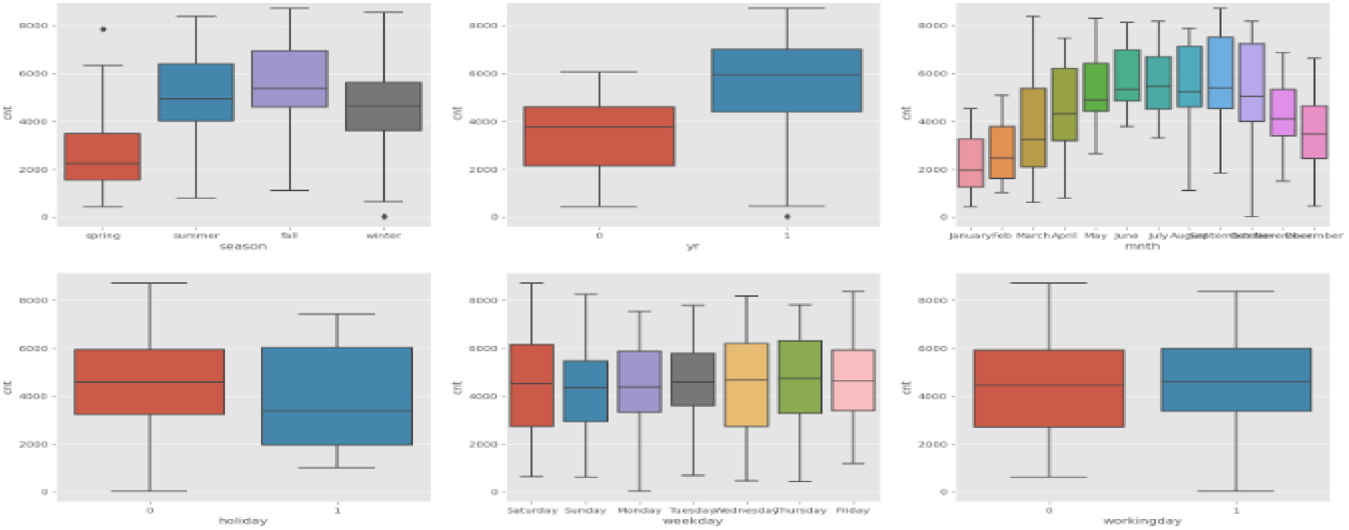
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

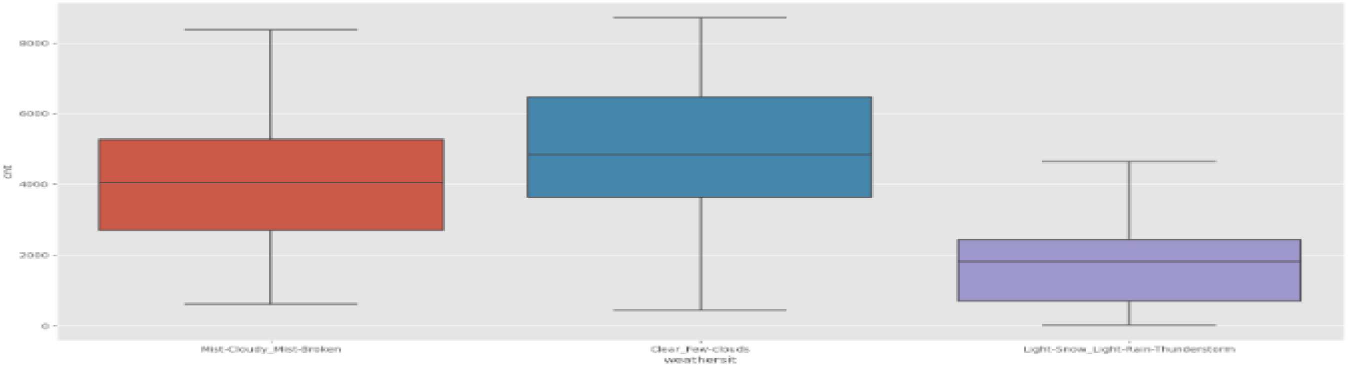
Ans:

These are the categorical values - Season, Weathersit, Year, Month, Holiday, Weekday and based on the analysis below is the effect on the dependent variable

- Season: Fall Season has the highest demand but Spring Season has the lowest demand for bikes
- Weathersit:
  - "Clear, Few clouds" weather Situation has highest demand
  - "Light Snow, Light Rain" weather situation has lowest demand
  - "Mist-Cloudy_Mist-Broken" weather situation demand is in medium
- Year: The year 2018 had low demand but there was high demand in the year 2019
- Month: September month has the highest demand followed by October, August and June but the lowest demand is in January.
- Holiday: There is high demand in holidays compared to a non- holiday
- Weekday: There is not high variation in demand in the weekdays

Below is the boxplot of the categorical variables:





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

Ans:

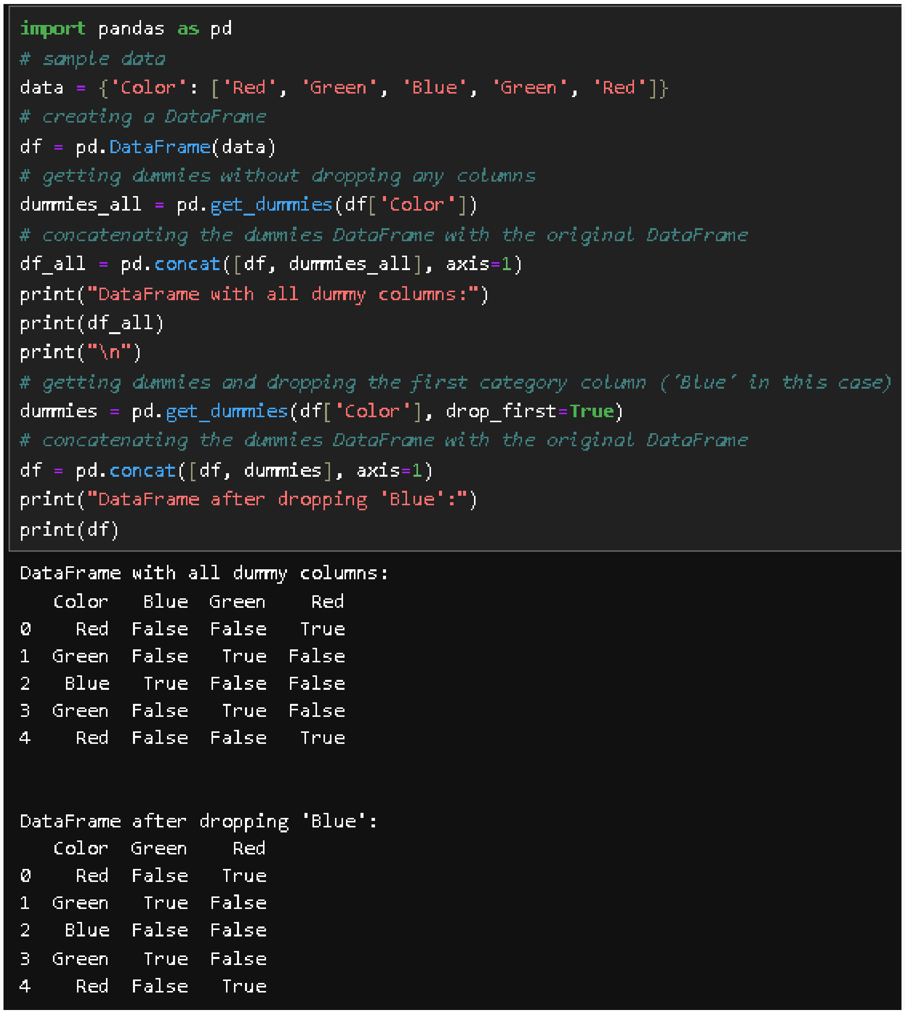
In Pandas ‘drop_first=True’ helps in reducing the extra column created during dummy variable creation.Hence it reduces the correlations created among dummy variables.

For example if we have 3 types of values in Categorical column with Blue, Green, Red as its values and we want to create dummy variable for that column.

If one variable is not Blue and Red, then It is obvious it is Green. So we do not need 3rd variable to identify the Green Value.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

Example:



```
import pandas as pd
# sample data
data = {'Color': ['Red', 'Green', 'Blue', 'Green', 'Red']}
# creating a DataFrame
df = pd.DataFrame(data)
# getting dummies without dropping any columns
dummies_all = pd.get_dummies(df['Color'])
# concatenating the dummies DataFrame with the original DataFrame
df_all = pd.concat([df, dummies_all], axis=1)
print("DataFrame with all dummy columns:")
print(df_all)
print("\n")
# getting dummies and dropping the first category column ('Blue' in this case)
dummies = pd.get_dummies(df['Color'], drop_first=True)
# concatenating the dummies DataFrame with the original DataFrame
df = pd.concat([df, dummies], axis=1)
print("DataFrame after dropping 'Blue':")
print(df)
```

```
DataFrame with all dummy columns:
   Color   Blue  Green    Red
0    Red  False  False   True
1  Green  False   True  False
2   Blue   True  False  False
3  Green  False   True  False
4    Red  False  False   True


DataFrame after dropping 'Blue':
   Color  Green    Red
0    Red  False   True
1  Green   True  False
2   Blue  False  False
3  Green   True  False
4    Red  False   True
```

As can be seen the **drop_first=True** argument is passed to **get_dummies()** to indicate that the first category should be dropped.
Hence the resulting DataFrame contains two columns Green and Red. The category named Blue is not represented in these columns because it was dropped.

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

Ans: Both **temp** and **atemp** have high correlation with **cnt** variable as can be seen in the pair plot.

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

Ans:

Below are the assumptions for simple linear regression:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

With these assumptions we can go ahead and make inferences about the model which, otherwise, we wouldn't have been able to.
Also note that, there is NO assumption on the distribution of Xand Y, just that the error terms have to have a normal distribution.
To meet above assumptions , we ensured to build the model on training set with below values;

- Co-Efficient Values – Non-Zero Co-Efficient indicate that all there is relationship between independent and dependent variables.
- P-Value – P-values are less than 0.05 which indicate they are significant to model.
- VIF – VIF is less than 5, so that there is no multicollinearity between predictor variables.
- F-statistic and Prob(F-statistic) – F-statistic high and low Prob(F-statistic) indicates that overall model fit is significant and not
  just by chance or only predictor variables are significant.

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: These are the top 3 features contributing significantly towards the demand of the shared bikes;

1. **atemp**
2. **year**
3. **Weathersit**

General Subjective Questions

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

In statistics, linear regression is a statistical model which estimates the linear relationship between a scalar response (dependent variable) and one or more explanatory variables (regressor or independent variable).
The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.
Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.
This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.
Linear regression is also a type of machine-learning algorithm, more specifically a supervised machine-learning algorithm, that learns from the labelled datasets and maps the data points to the most optimized linear functions which can be used for prediction on new datasets. This supervised machine-learning model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).
Linear regression models can be classified into two types depending upon the number of independent variables:

1. Simple linear regression: When the number of independent variables is 1

2. Multiple linear regression: When the number of independent variables is more than 1

The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimizing the cost function (RSS in this case, using the Ordinary Least Squares method) which is done using the following two methods:

1. Differentiation
2. Gradient descent method
3. The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$
4. RSS: Residual Sum of Squares
5. TSS: Total Sum of Squares

<u>Equation of Simple Linear Regression</u>, where b_o is the intercept, b_1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y= b_o + b_1 x$$

<u>Equation of Multiple Linear Regression</u>, where bo is the intercept, b_1,b_2,b_3,b_4…,b_n are coefficients or slopes of the independent variables x_1,x_2,x_3,x_4…,x_n and y is the dependent variable.

$$y= b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \; \dots . + b_n x_n$$

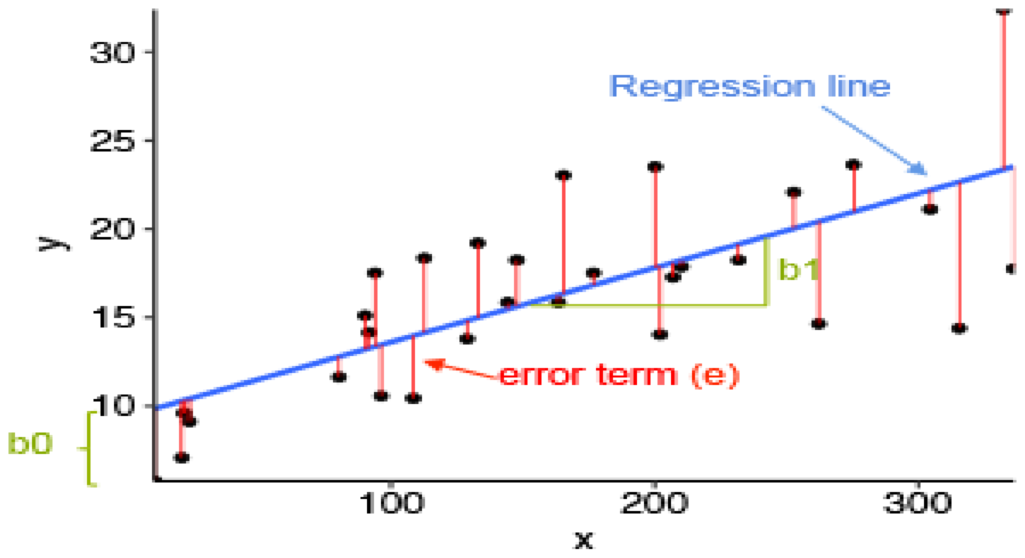


Image Source: Statistical tools for high-throughput data analysis

In the above diagram,

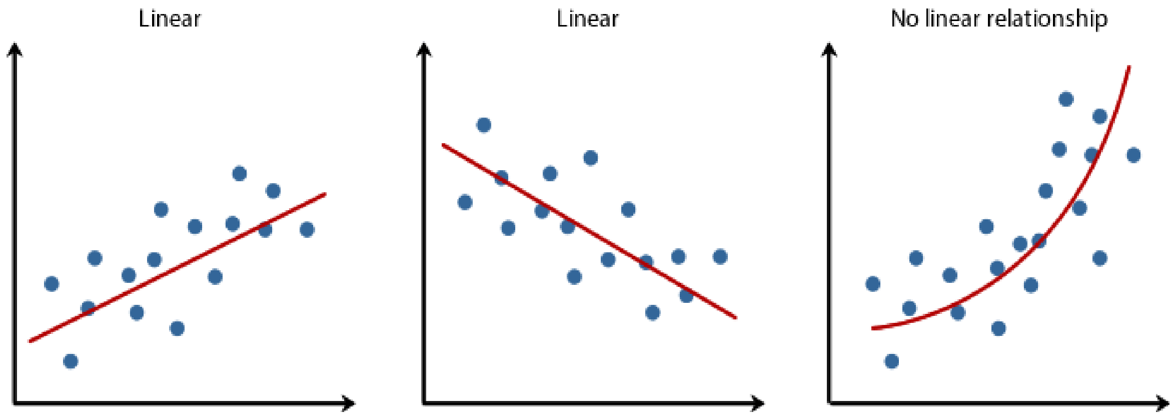
- x is our independent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.

- Black dots are the data points i.e the actual values.
- b_o is the intercept which is 10 and b_1 is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.
- The vertical distance between the data point and the regression line is known as error or residual. Each data point has one
  residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Mathematical Approach:
Residual/Error = Actual values – Predicted Values
Sum of Residuals/Errors = Sum(Actual- Predicted Values)
Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2

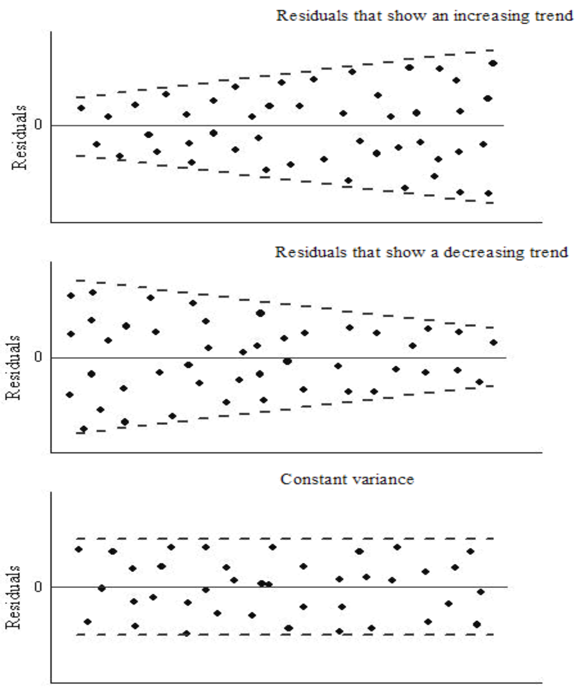
$$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

Rsq, AdjRsq, MSE, RMSE, MAE – 5 evaluation metrics
Assumptions of Linear Regression –
The basic assumptions of Linear Regression are as follows:
Linearity: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by
plotting a scatter plot between both variables.



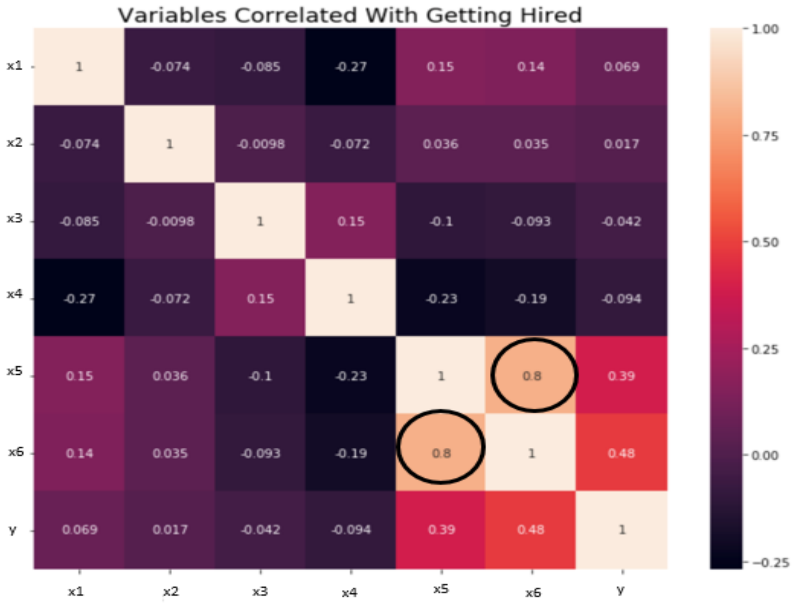
Copyright 2014. Laerd Statistics.

3. Homoscedasticity: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X.
   This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.

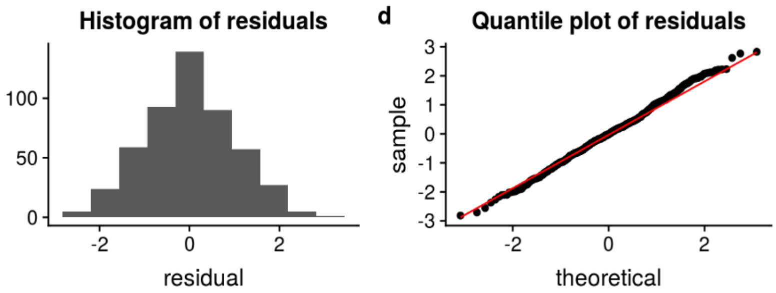


Error Term : y act – y pred

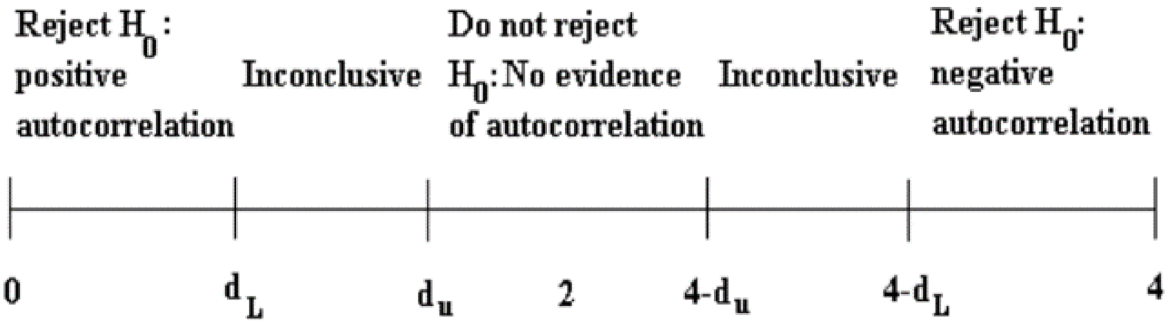
4. **Independence/No Multicollinearity**: The variables should be independent of each other i.e no correlation should be there between the independent variables.
   To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.
   In the below image, a high correlation is present between x5 and x6 variables.



4. The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.



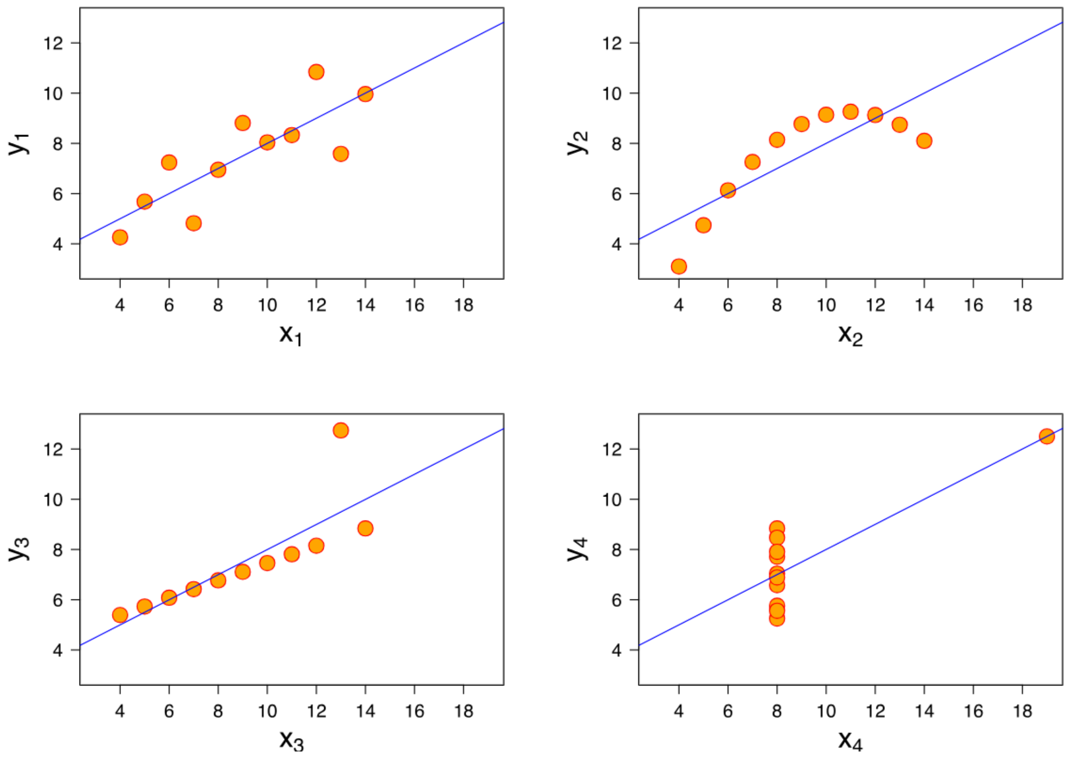
5. **No Autocorrelation**: The error terms (yact – ypred) should be independent of each other. Autocorrelation can be tested using the **Durbin Watson test**.
   The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.


Reject H_0 : positive autocorrelation
Inconclusive
Do not reject H_0: No evidence of autocorrelation
Inconclusive
Reject H_0: negative autocorrelation
0
d_L
d_u
2
$4\text{-}d_u$
$4\text{-}d_L$
4

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

Ans:

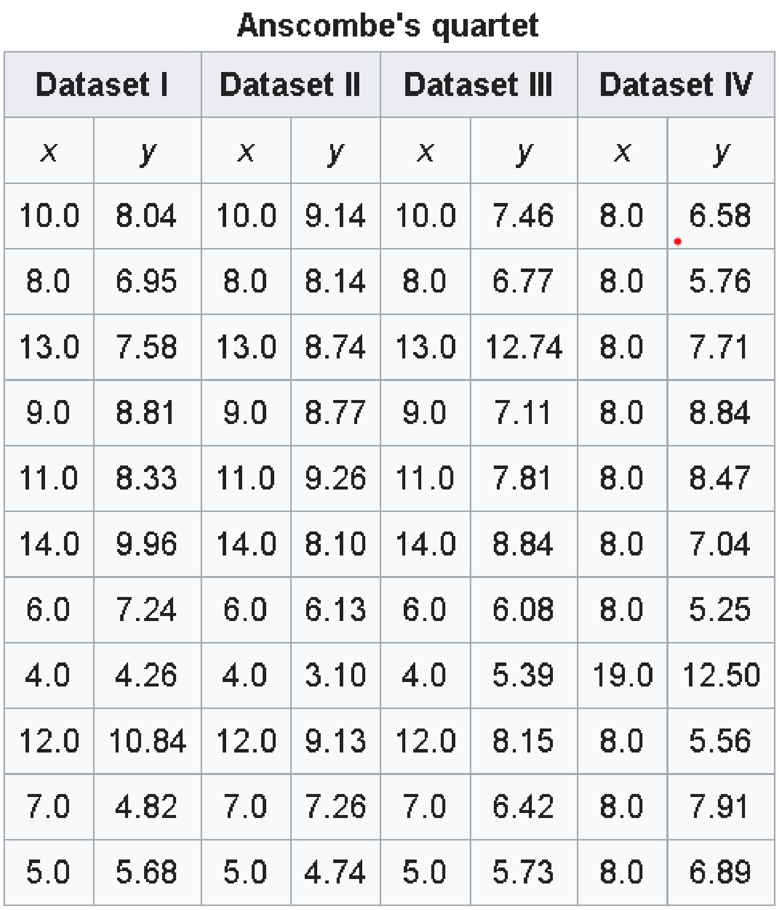
Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear
very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to
demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties.
He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".
As can be seen below, the four datasets composing Anscombe's quartet.
All four sets have identical statistical parameters, but the graphs show them to be considerably different



For all four datasets:

Property	**Value**	**Accuracy**
Mean of x	9	exact
Sample variance of x: s2x	11	exact
Mean of y	7.5	to 2 decimal places
Sample variance of y: s2y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R2	0.67	to 2 decimal places

The datasets are as follows. The x values are the same for the first three datasets.



Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset 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

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

Ans:

Pearson's correlation helps us understand the relationship between two quantitative variables when the relationship
between them is assumed to take a linear pattern. The relationship between two quantitative variables (also known as continuous variables),
can be visualized using a **scatter plot**, and a straight line can be drawn through them. The closeness with which the points lie along this line is
measured by Pearson's correlation coefficient, also often denoted as Pearson's R, and sometimes referred to as Pearson's product moment
correlation coefficient or simply the correlation coefficient. Pearson's R can be thought of not just as a descriptive statistic but also an inferential
statistic because, as with other statistical tests, a hypothesis test can be performed to make inferences and draw conclusions from the data.

Pearson correlation coefficient formula

The formula for Pearson's correlation coefficient, r, relates to how closely a line of best fit, or how well a linear regression, predicts the relationship
between the two variables. It is presented as follows:

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

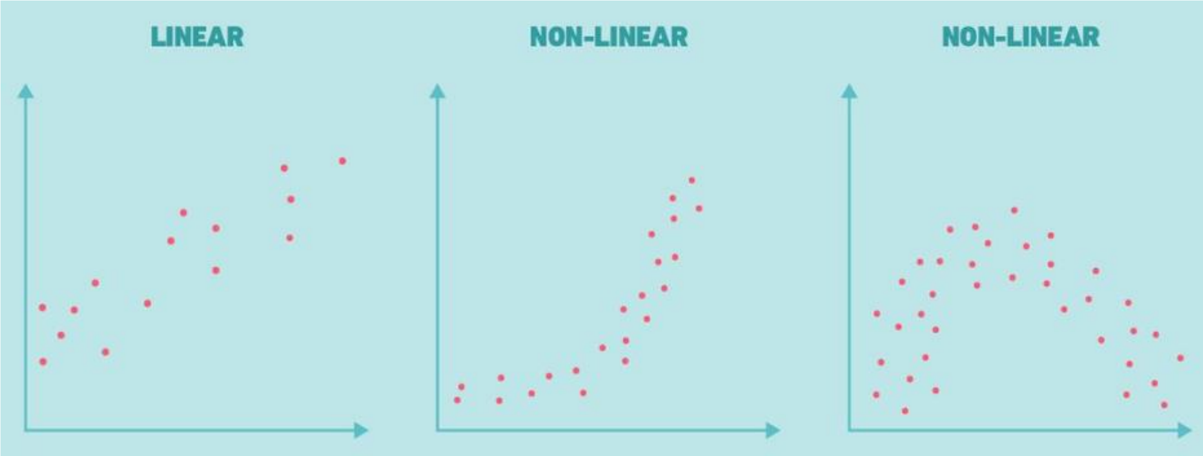
where x_i and y_i represent the values of the exposure variable and outcome variable for each individual respectively, and $\bar{x}$ and $\bar{y}$ represent the mean of the values of the exposure and outcome variables in the dataset.

This formula works by numerically expressing the variation in the outcome variable “explained” by the exposure variable. In essence, we are calculating the sum of products (multiplying corresponding values from pairs and then adding together the resulting values) about the mean of x and y divided by the square root of the sum of squares about the mean of x multiplied by the sum of squares about the mean of y. The variation is expressed using the sum of the squared distances (e.g. $(x_i - \bar{x})^2$) of the values from the mean of y and x. This captures the variation in the values around the line of best fit and also ensures the correlation coefficient lies between − 1 and + 1.

The assumptions of Pearson's correlation test:

There are some key assumptions that should be met before Pearson’s correlation can be used to give valid results:

- The data should be on a **continuous** scale. Examples of continuous variables include age in years, height in centimeters and temperature in degrees Celsius. Sometimes continuous variables are referred to as quantitative variables, although, it’s important to remember that, while all continuous variables are quantitative, not all quantitative variables are continuous.
- The variables should take a **Normal distribution**. This assumption means the Pearson’s correlation test is a parametric test. If the data in the variables of interest take some other distribution, then a non-parametric test for correlation such as **Spearman’s rank correlation** should be used. This assumption can be checked using a histogram.
- There should be **no outliers** in the dataset. Outliers are values that are notably different from the pattern of the rest of the data and may influence the line of best fit and warp the correlation coefficient.
- The relationship between the two variables is assumed to be **linear**. This assumption is related to the “no outliers” assumption, in that the relationship should be able to be described by a straight line relatively well. These assumptions can be checked using scatter plots as can be seen below



Scatter plots showing examples of linear and non-linear relationships. *Credit: Technology Networks.*

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

Ans:

What is scaling?

Scaling 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 is scaling performed?
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.
What is the difference between normalized scaling and standardized scaling?

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.

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

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).

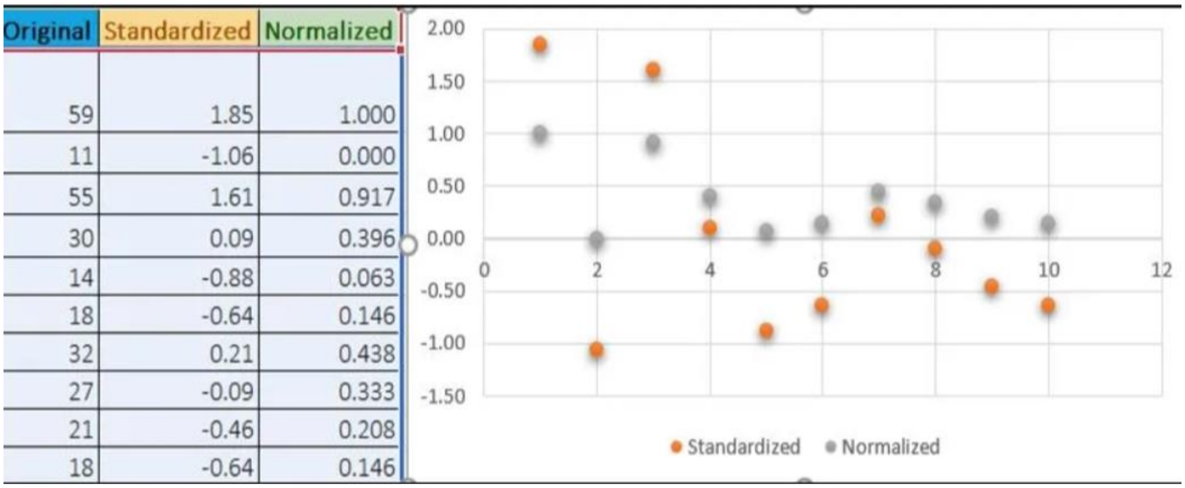
$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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?**

IF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1=C+ α_2 X_2+α_3 X_3+⋯

〖VIF〗_1=1/(1-R_1^2)

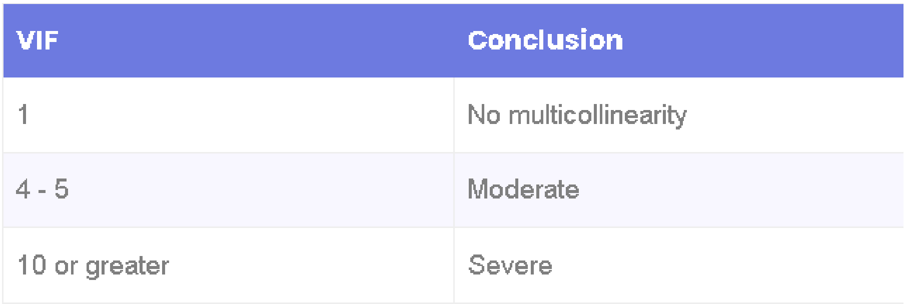
Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

X_2=C+ α_1 X_1+α_3 X_3+⋯

〖VIF〗_2=1/(1-R_2^2)

If all the independent variables are orthogonal to each other, then VIF = 1.0. **If there is perfect correlation, then VIF = infinity**. orthogonality can refer to the independence of predictors: **Independent Variables**: If two independent variables are orthogonal, it means they do not share any variance. This can reduce multicollinearity, leading to more stable and interpretable models.

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. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation). The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.



VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
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.
Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

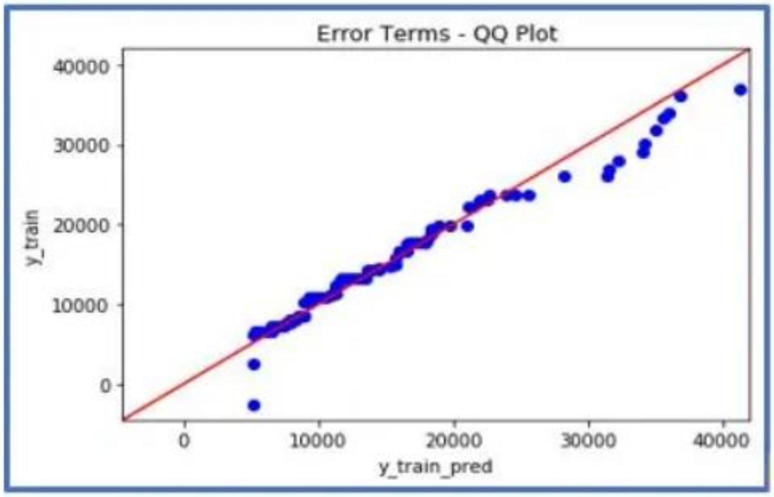
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

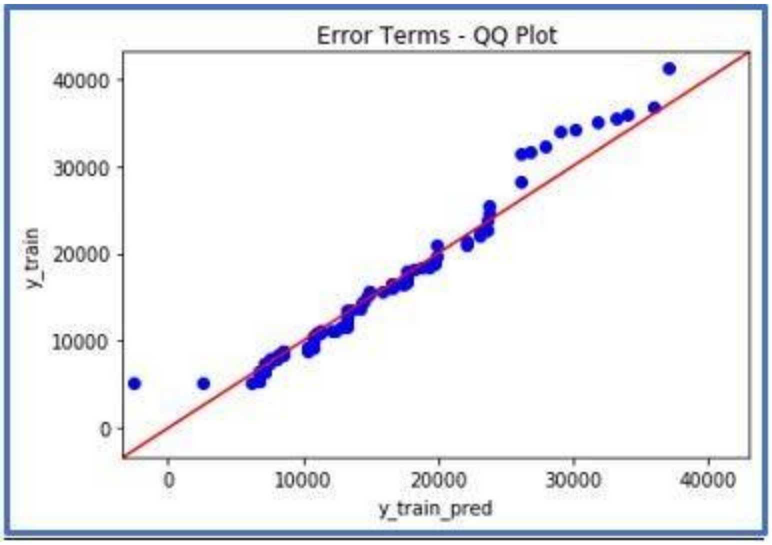
Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis

Python: statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.