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: - cnt (count) is the target variable or the dependent variable. The categorical variables from the dataset are:

- season: Almost 32% of the bike booking were happening in fall with a median of over 5000 booking (for the period of 2 years). This was followed by summer & winter with 28% & 26% of total booking. This indicates, season can be a good predictor for the dependent variable.

- month: Almost 10% of the bike booking were happening in the months August, June, September, July and May with a median of over 4000 booking per month. This indicates, month has some trend for bookings and can be a good predictor for the dependent variable.

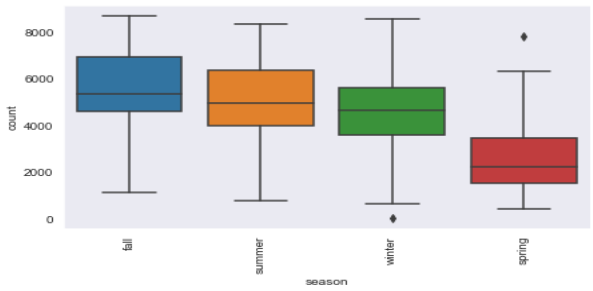
- weather: Almost 68% of the bike booking were happening during clear or partly cloudy with a median of close to 5000 booking (for the period of 2 years). This was followed by Mist and cloudy with 30% of total booking. This indicates, weather does show some trend towards the bike bookings can be a good predictor for the dependent variable.

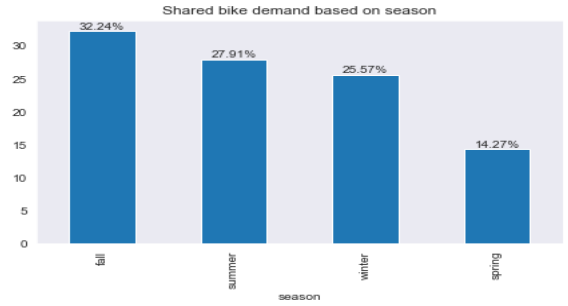
- holiday: Almost 98% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

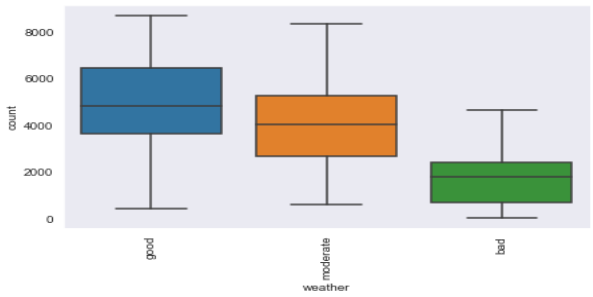
- weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.

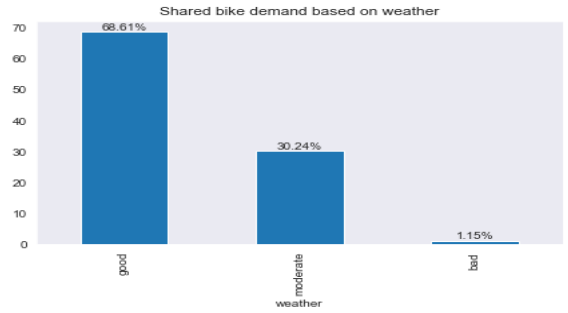
- workingday: Almost 69% of the bike booking were happening on a workingday with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

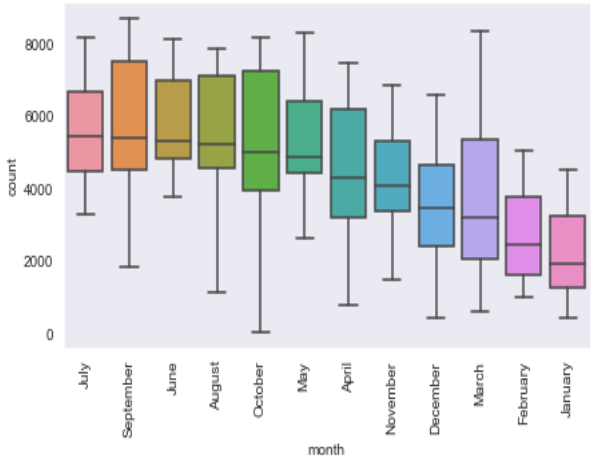
- year: this shows almost 25% increase in the bike bookings from 2018 to 2019. Due to pandemic, the sales might have dropped but this trend indicates that post covid the sales will increase year on year.

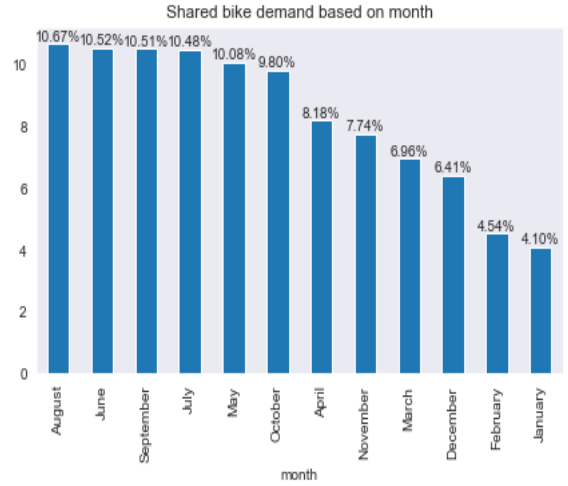

count
season
fall
summer
winter
spring

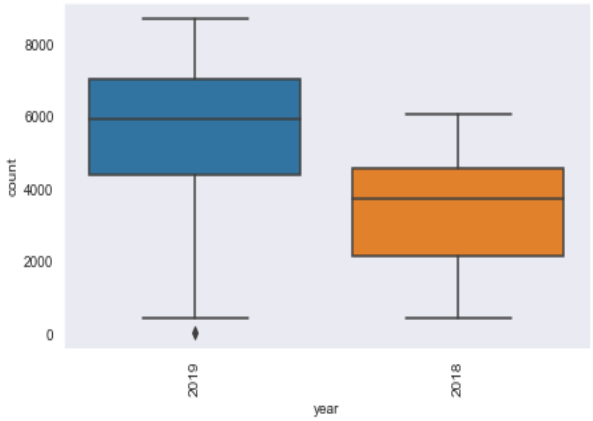

Shared bike demand based on season
32.24%
27.91%
25.57%
14.27%
season
fall
summer
winter
spring

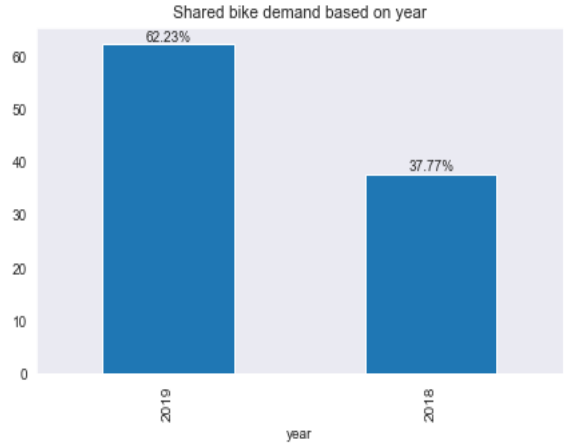

count
weather
good
moderate
bad

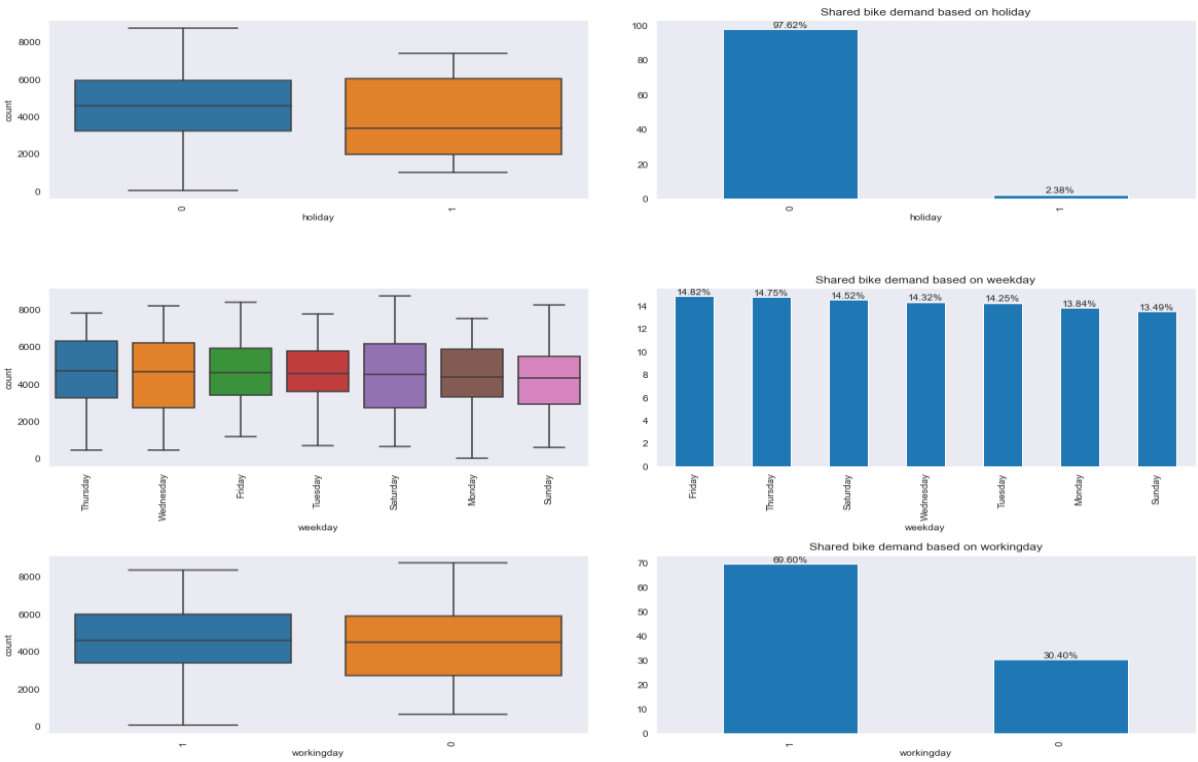

Shared bike demand based on weather
68.61%
30.24%
1.15%
weather
good
moderate
bad


count
month
July
September
June
August
October
May
April
November
December
March
February
January


Shared bike demand based on month
10.67%
10.52%
10.51%
10.48%
10.08%
9.80%
8.18%
7.74%
6.96%
6.41%
4.54%
4.10%
month
August
June
September
July
May
October
April
November
March
December
February
January


count
year
2019
2018


Shared bike demand based on year
62.23%
37.77%
year
2019
2018



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

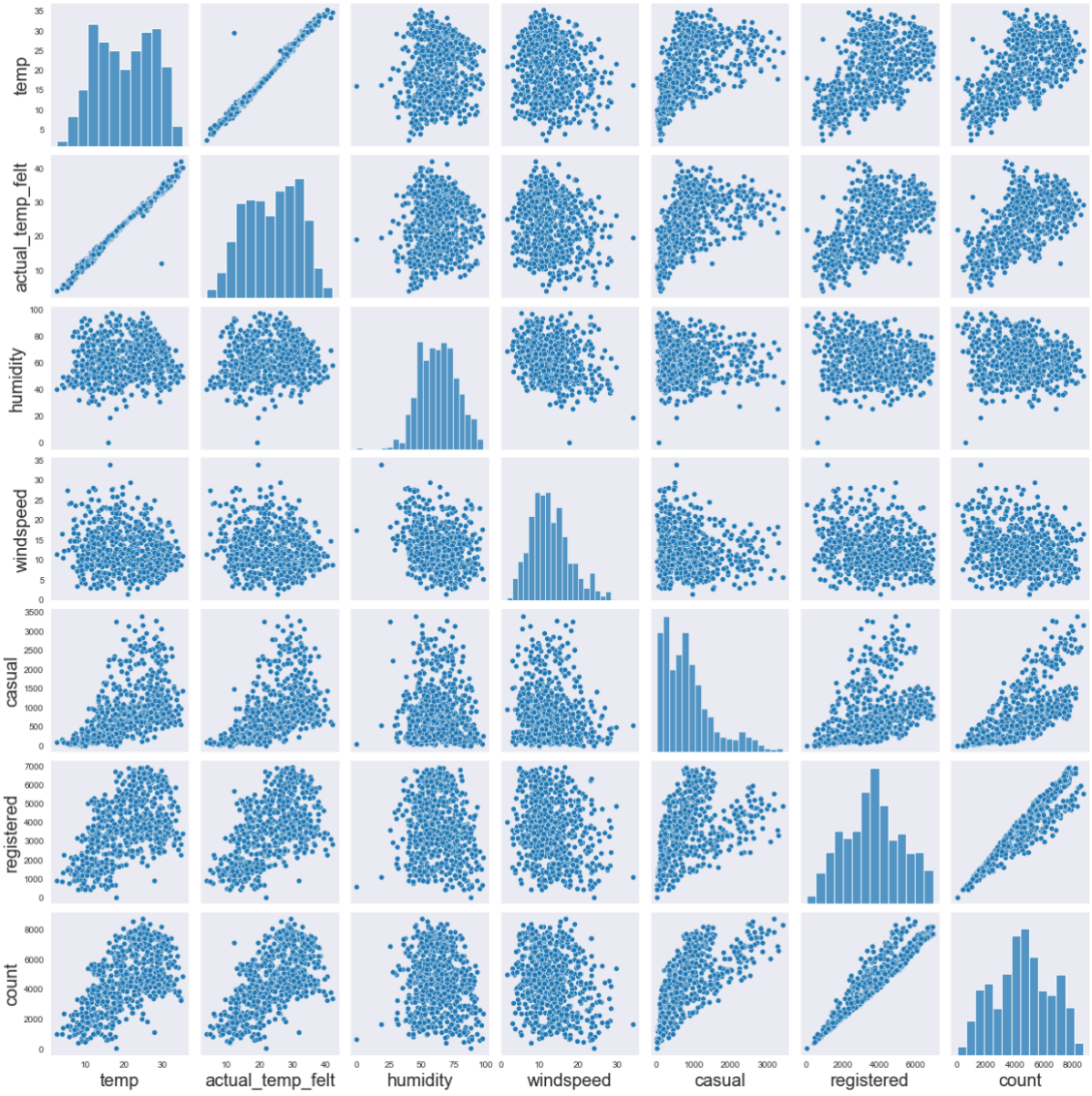
Ans: During one hot encoding we create dummy variables from categorical data into a form which is understood by ML. If there are n categorical levels for a variable then ideally there will be n columns created. However we only need n-1 columns to represent all n levels because if all columns values are 0 then that can represent one of the levels.

So to reduce the number of columns and in turn reduce the multicollinearity with linear regression, it is important to use drop_first = True that will drop the first column and identifies it if values in all other columns related to that category is 0.

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

Ans: Based on the pair-plot, the registered variable have the highest correlation with the target variable cnt (count).

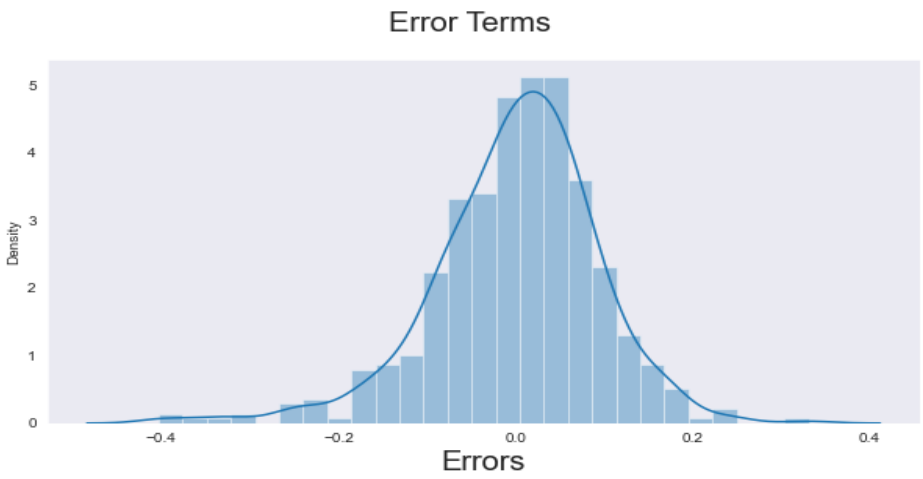
Since the registered variable is part of total count, if we have to consider other variables then atemp and temp have next highest correlation to the target variable.



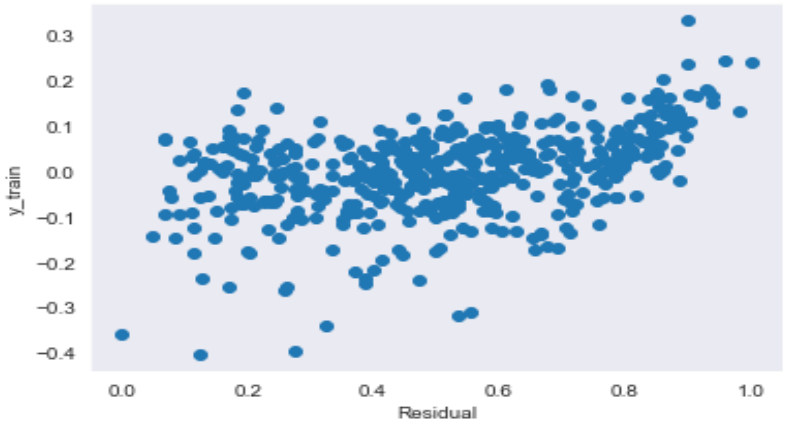
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: -

1. We can validate the first assumption of Linear Regression that residual should follow normal distribution around mean 0. We can check this by plotting distribution of the errors (i.e actual – predicted)

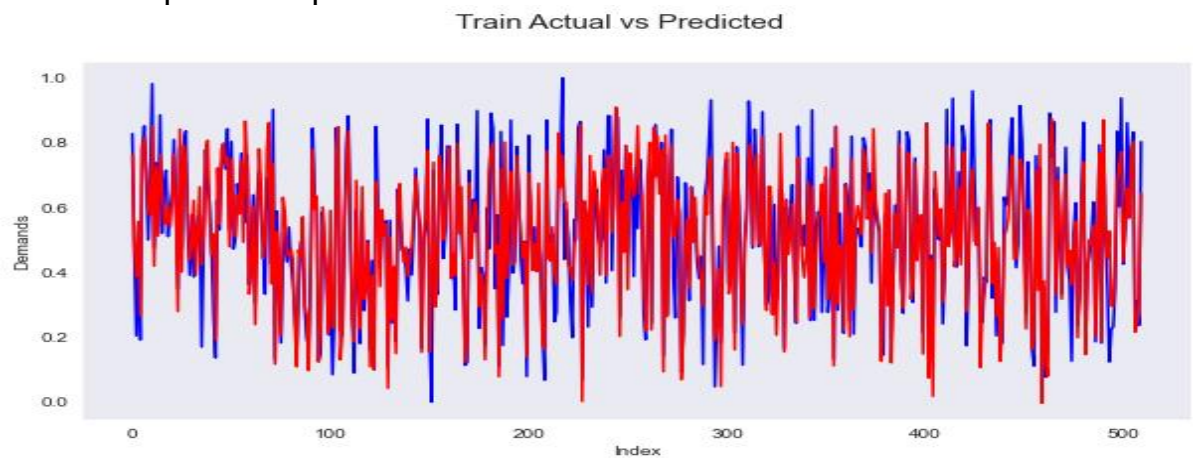


2. The error terms are randomly distributed and not dependant on each other. We can check this by drawing a scatter plot between y_train and residuals.



3. Then we can check that actual and predicted patterns fit well. Sometimes just having high adjusted R^2 alone doesn't guarantee that model is good.

   We can check this by line plotting actual vs predicted and check if the actual and predicted patterns fit well.



4. Then we can also use the statistical information to ensure the model is good.

   As shown below,

- Adj-R^2 is 0.823. 0 is worst and 1 is best fit. 0.82 is a good fit.
- F statistic is high and Prob (F-statistic) is ~ 0 which indicates the model is more significant
- t statistic – helps in determining if the selected variables are significant or not based on p values.
- all coefficients are non -zero
- Hence we can infer that the model is significant.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  count   R-squared:                       0.826
Model:                            OLS   Adj. R-squared:                  0.823
Method:                 Least Squares   F-statistic:                     296.5
Date:                Wed, 09 Feb 2022   Prob (F-statistic):          1.53e-184
Time:                        14:45:43   Log-Likelihood:                484.24
No. Observations:                 510   AIC:                            -950.5
Df Residuals:                     501   BIC:                            -912.4
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0442      0.017      2.601      0.010       0.011       0.078
month_September  0.0978      0.016      6.052      0.000       0.066       0.130
season_summer  0.0894      0.011      8.460      0.000       0.069       0.110
season_winter  0.1281      0.011     12.051      0.000       0.107       0.149
temp           0.5527      0.020     27.295      0.000       0.513       0.592
weather_bad   -0.2019      0.026     -7.839      0.000      -0.252      -0.151
weather_good   0.0767      0.009      8.553      0.000       0.059       0.094
windspeed     -0.1552      0.026     -6.041      0.000      -0.206      -0.105
year           0.2332      0.008     27.645      0.000       0.217       0.250
==============================================================================
Omnibus:                       65.957   Durbin-Watson:                   2.042
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              141.455
Skew:                          -0.715   Prob(JB):                     1.92e-31
Kurtosis:                       5.148   Cond. No.                         10.9
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
          Features   VIF
3             temp  4.37
6        windspeed  3.16
5     weather_good  2.66
7             year  2.00
1    season_summer  1.55
2    season_winter  1.35
0  month_September  1.20
4      weather_bad  1.11
```

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:- Based on the final model, the top 3 features contributing significantly towards explaining the demand of shared bikes are:

1. Temp: with a co-efficient of + 0.552655 indicating that the demand will increase if the temperature increases.
2. Year: with a co-efficient of + 0.233163 indicating that the demand will increase year on year.
3. Weather_bad: with a negative co-efficient of – 0.201859 indicating that the demand will decrease if the weather is bad.

```
const              0.044248
month_September    0.097843
season_summer      0.089399
season_winter      0.128144
temp               0.552655
weather_bad       -0.201859
weather_good       0.076670
windspeed         -0.155224
year               0.233163
```

Model Interpretation

y = B0 + B1X1 + B2X2 ...BnXn

y = (const * 0.044248) + (temp * 0.552655) + (year * 0.233163) + (season_Winter * 0.128144) + (month_September * 0.097843) + (season_Summer * 0.089399) + (weather_good * 0.076670) -(windspeed * 0.155224) -(weather_bad * 0.201859)

General Subjective Questions

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

   Regression is a method to model a target value based on independent predictors. This is mainly used in forecasting and finding the relation between dependent and independent variables. Regression techniques vary based on number of independent variables and type of relation between dependent and independent variables.

   Linear regression algorithm is a simple type of regression analysis where there is a linear relationship between independent and dependent variables.

   Simple Linear regression will have one independent variable and multiple linear regression will have more than one independent variables.

   Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task** and is performed on continuous variables.

   Simple Linear regression performs the task to predict a dependent variable (y) based on a given independent variable (x) having a linear relationship.

   The basic equation for simple linear regression is:

   $y = mX + c$
   where m is slope i.e change in y value for change in x value.
   $m = (yi-y)/(xi-x)$
   c = intercept were y meets x i.e. value of y when x = 0

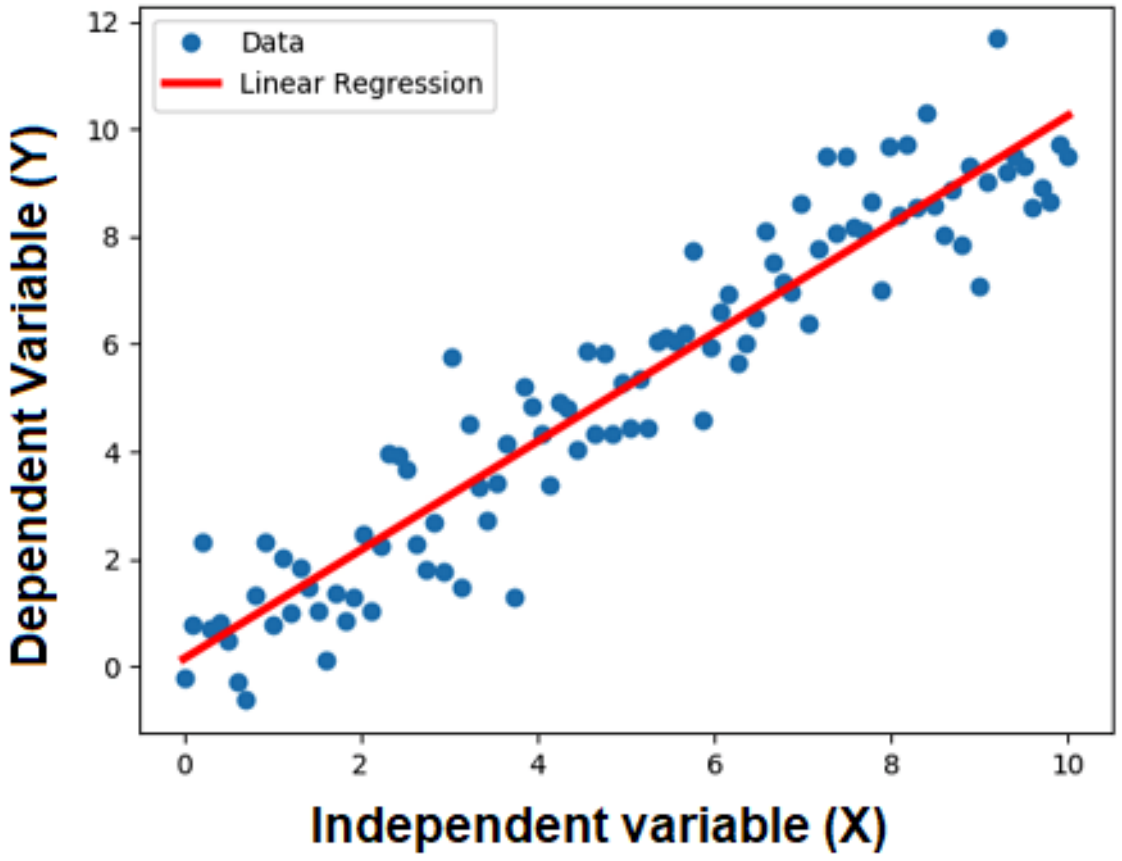
   Now we can extend the simple linear regression to multiple independent variables (X1, X2….) and that is called multiple linear regression.

   $y = b0 + b1X1 + b2X2 + ……bnXn$

   The idea is to fit a line passing through the majority of the data points in such a way that distance between the predicted and actual value is minimum. For this we use LMSE (lease mean square error)

i.e. error i.e. actual – predicted value is our cost function and objective is to minimize this cost function and find the best fit line using methods like Gradient descent etc..

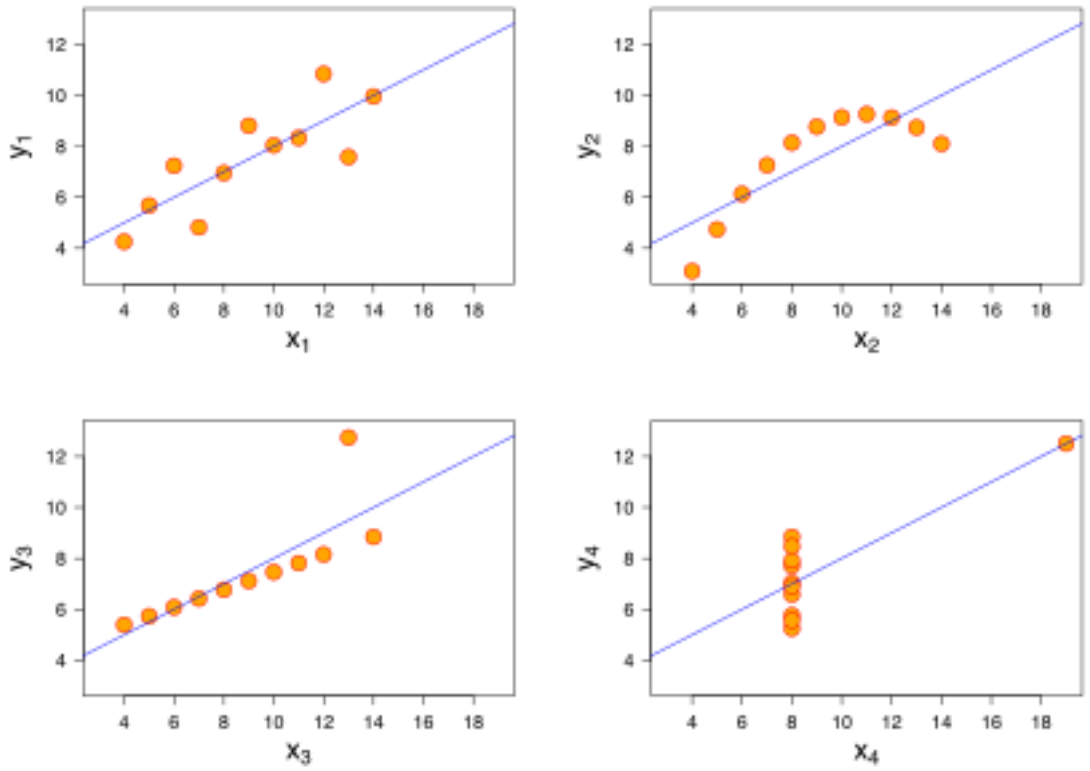
The objective of linear regression model is to predict these co-efficients b0, b1, b2…bn in order to generate best fit line and predict target variable for given independent variable values.



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

Anscombe's quartet way is to illustrate the importance of visualization of data. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

In 1973 statistician Francis Anscombe constructed 4 data sets to show the importance of visualization as these had similar statistical information.



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	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 : R^2	0.67	to 2 decimal places

As you can see all 4 had same statistical information but the visualization shows that they have completely different patterns.

The question is then what is the application of this quartet ?

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

Even in our assignment we have done lot of visualization and then confirmed the inference we got from the statistical data.

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

In linear regression we mainly deal with more than one variable and try to understand the relation between the variables. So in statistical terms a statistic that measures the relationship between two variables in called correlation. It not only shows how strong the relationship is but also gives the direction in a numerical way. The numerical value of the correlation co-efficient lies between -1 to +1.
+1 indicates a strong positive correlation meaning if one value changes the other value also changes in same proportion in the same direction.

-1 indicates a strong negative correlation meaning if one value changes the other value also changes in same proportion but opposite direction i.e. inversely proportional

0 indicates that there is correlation between the two variables.

Now that we have understood correlation what is Pearson's R?

It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

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

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

In short, Pearson's co-efficient denoted by r calculates the Linear relationship between two variables.

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 which is applied to independent variables to normalize the data within a particular range.

   Why scaling?
   Say for. e.g. if the dataset has 10 variables and say some variable is in the range of 1 to 10 and some in 0.1 to 0.8 and some in the range of 10000 to 20000.
   If we represent this dataset visually, the ones in the smaller range will get masked and it becomes difficult to explain the data.

   Is that all? No – it also helps in speeding the calculations in an algorithm.

   That's great then how to perform scaling. The simple idea is to bring all the variable values into the same range. So there are 2 types of scaling techniques.

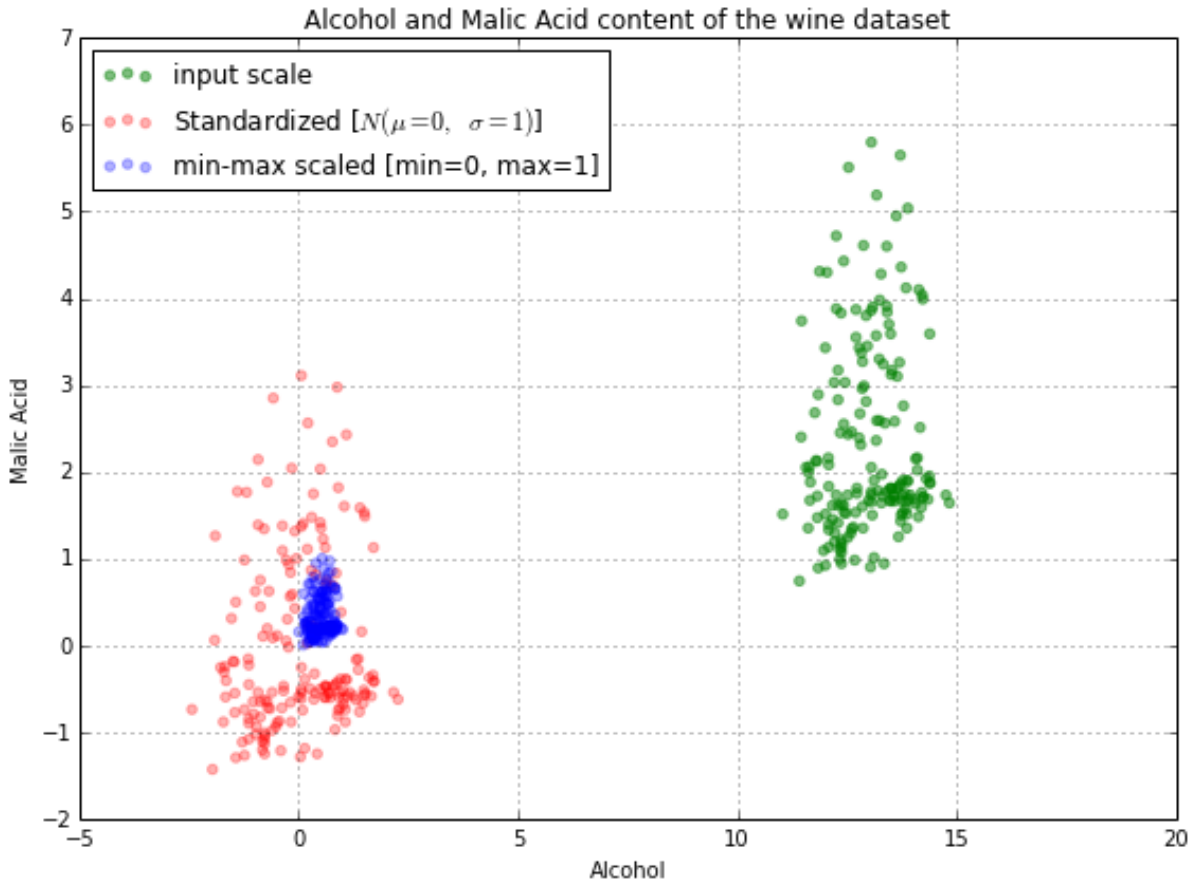
   1. Normalized scaling also called min-max scaling.

      It brings all the data in the range of 0 and 1.

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

   2. Standardized scaling
      It brings all the data into standard normal distribution which has a mean zero and standard deviation of one.

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



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

   VIF represents the correlation between two independent variables.

   VIF = 1 / 1-R2

   If VIF is infinite then R2 should be equal to 1 as per the above equation. R2 = 1 in case of perfect correlation. So the value of VIF being infinite indicates a perfect correlation between two independent variables. This indicates multicollinearity.

   To solve the problem of multicollinearity, we should drop one of the variables from the dataset which is causing this 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 line y = x will be plotted at 45 degrees and the if the two distributions are similar then the points in the Q-Q plot will approximately lie on the line y = x

What is the importance of Q-Q plot in linear regression then:
If the two distributions are linearly related, the points in the Q-Q plot will approximately lie on the line but not exactly on the line.

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.

We have used the Q-Q plot in the assignment to see how well. If the data points didn't fit on the line, we would then use logarithmic scale etc.. to convert the target variable to fit well on the line.