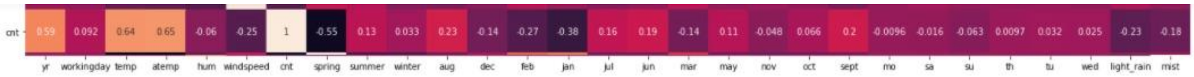
Assignment-based Subjective Questions

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: At first comparison of the dependent variable with categorical variables from the dataset, it seems that "*spring*" is highly correlated with *cnt* (negative correlation) also *Jan* and *Aug* were significantly correlated with the dependent variable *cnt*, while we created final regression model, found that *jan*, *feb* and *light_rain* is much correlated as *spring* was not having much significance as well as highly correlated with other independent variable.

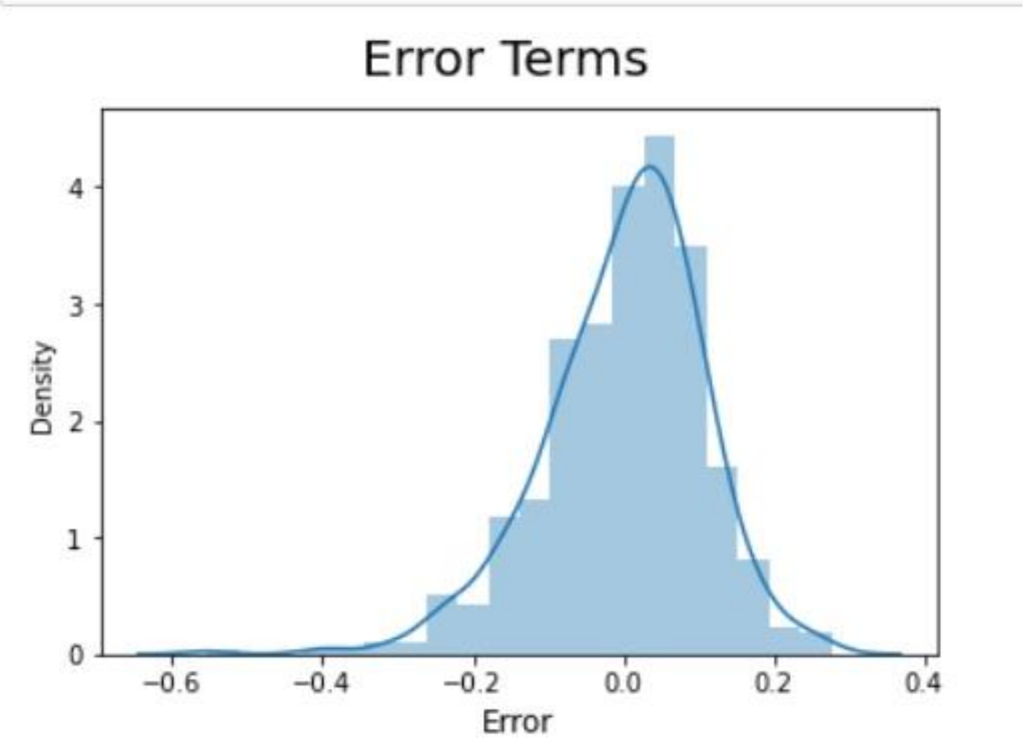


   By looking at the final model, we can also conclude that using dummy variable is a good idea as every categorical variable has different significance as well as correlation with the dependent variable *cnt.* For example, in season, spring had negative correlation with dependent variable while summer has positive. In months, j*an* high negative correlation while *sept* has good positive correlation.

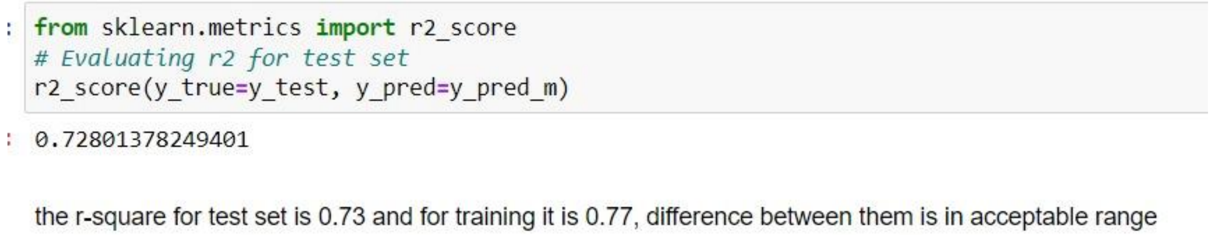
2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   Ans: While creating dummy variables from categorical variables, we only need n-1 variables to define n values. (One column can be defined with all other 0 values) that's why we use drop_first=True to drop first column.
   Now, why we need to delete one column? Because reducing columns which are not participating in increasing the significance of dependent column increases the complexity of the model by adding one additional coefficient constant.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   Ans: By looking the pair-plot, temp and atemp seems to be having highest correlation with the target variable, while we plot heatmap for that with correlation values, we came to know that atemp has slightly more correlation value with the target variable.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   Ans: First we need to validate the training set with its prediction values using model and plot the error terms plot between the predicted values and actual values (constant or non-constant variance).



After that make a prediction using final model for test set values and then calculate the r-squared value for test set original values and predicted values using final model. It should be in same range (acceptable range +-5%)
In this case, r-squared for training set was 77% and for test set 73% which is acceptable.



```
from sklearn.metrics import r2_score
# Evaluating r2 for test set
r2_score(y_true=y_test, y_pred=y_pred_m)
```

```
0.72801378249401
```

the r-square for test set is 0.73 and for training it is 0.77, difference between them is in acceptable range

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:

- light_rain (negatively correlated; corr coef: -0.3149)
- jan (negatively correlated; corr coef: -0.2782)
- yr (positively correlated; corr coef: +0.2462)

General Subjective Questions

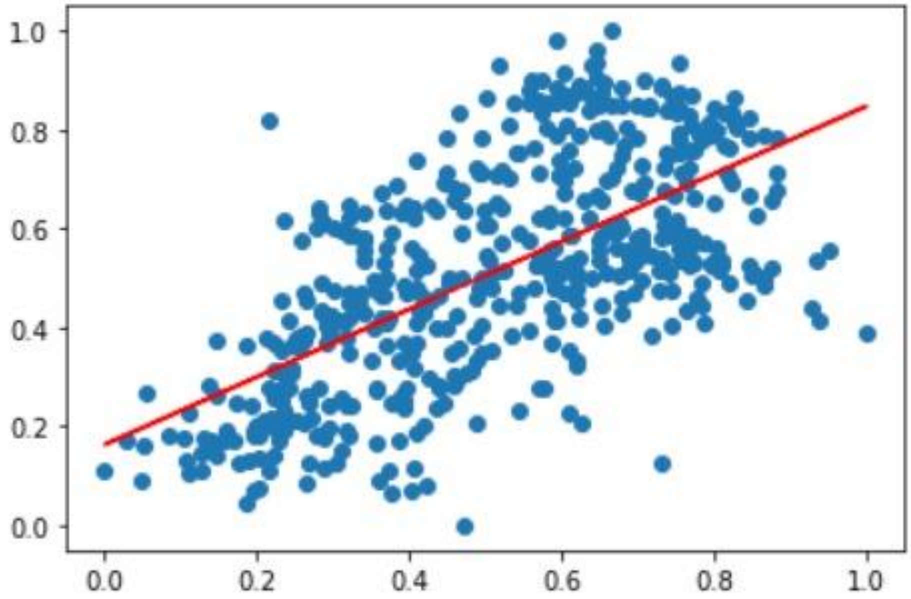
1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear Regression is a machine learning algorithm based on supervised learning which performs a regression task. 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.

There are 2 types of Linear Regression models:

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.



In case of multiple linear regression more than one independent variable shows relationship with dependent variable.

Steps to follow for Linear Regression model:

i) Reading and understanding the data
   (1) Load the data
   (2) Remove the unnecessary fields like id field and the fields that are not known in case of future forecasting.
   (3) Remove duplicate columns (Columns that represents same type of data).
ii) Visualizing the data
   (1) Visualize all numeric data using pair plot.
   (2) Create derived variables in case they define model better.
   (3) Visualize all categorical data using boxplot/barplot.
   (4) Try to visualize more than one variable together if there is any pattern.
iii) Data Preparation
   (1) Create Dummy variables for categorical data
   (2) Labalize column name of dummy variables in case needed.
iv) Split data into training and testing sets
   (1) Divide data randomly into training and testing sets. (usually 70-30 ratio).

    (2) Scale the data using standardizing or MinMax approach.
    (3) Check the correlation coefficients using values or heatmap.
    (4) Divide the dependent and independent variables for model building.

v) Build the linear model
    (1) Try to visualize the data using graph for highest correlated variable.
    (2) Try to build OLS Regression model.
    (3) Check the significance of the model then add new variable.
    (4) Try to add all the variables in the model and then check its significance and VIF values.
    (5) Remove low significant and high VIF values variables.
    (6) Build the final model.

vi) Residual Analysis of train data
    (1) Predict the train data set values using final model.
    (2) Plot Error Terms graph with predicted and actual values.

vii) Make Predictions of test data set using final model
    (1) Divide test data set into 2 parts.
    (2) Drop all the columns that are not present in training set.
    (3) Scale the continuous variables.
    (4) Predict the values using final model.

viii) Model Evaluation
    (1) Check r-squared value of the test model.
    (2) Plot actual test value vs predicted value graph to see the pattern.

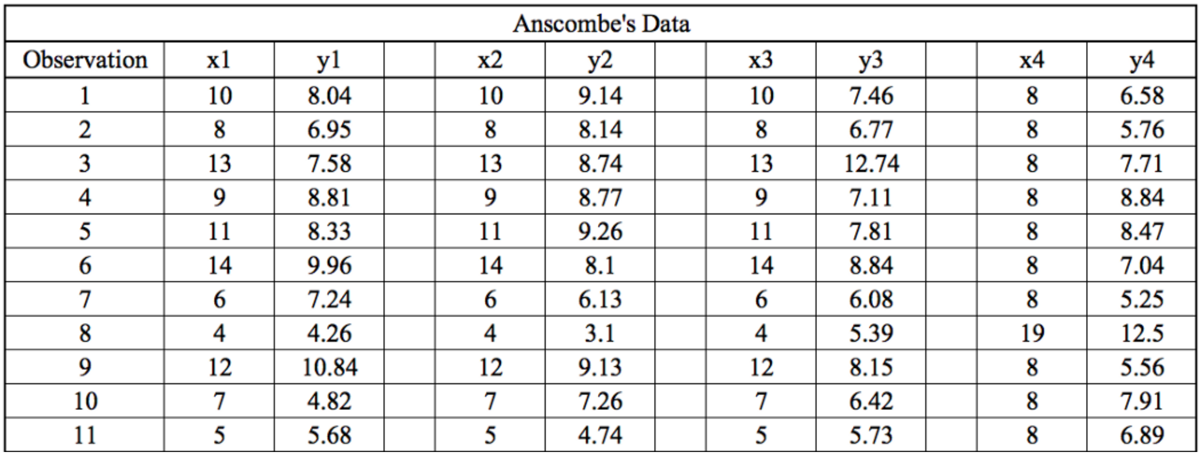
ix) Equate the final equation for best fitted line.

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

Ans: 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.

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

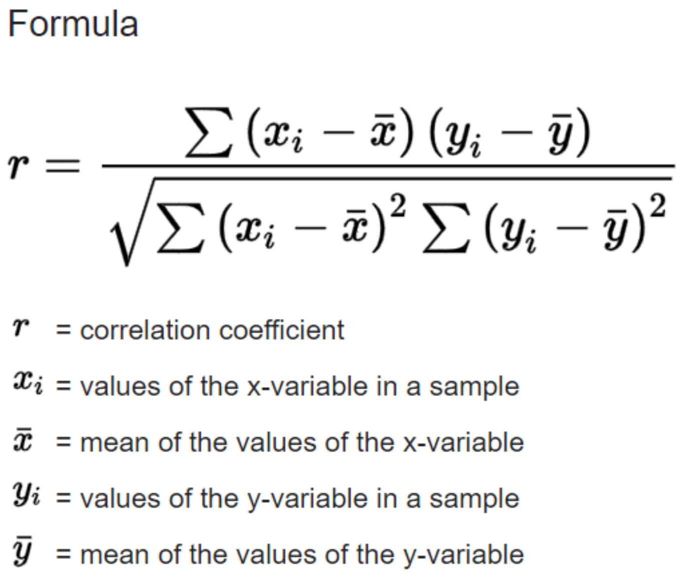


Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

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

Ans: Pearson correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and 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. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).



Formula

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

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

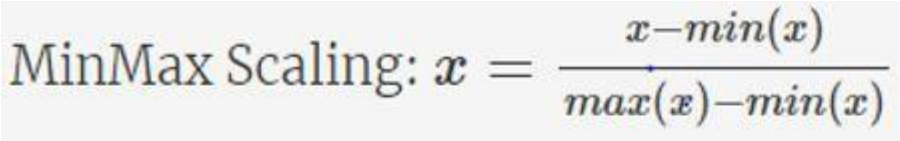
$\bar{y}$ = mean of the values of the y-variable

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

Ans: Scaling is a method used to normalize the range of independent variables of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

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 must do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/MinMax Scaling brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



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

standardized scaling 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).

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

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

(3 marks)

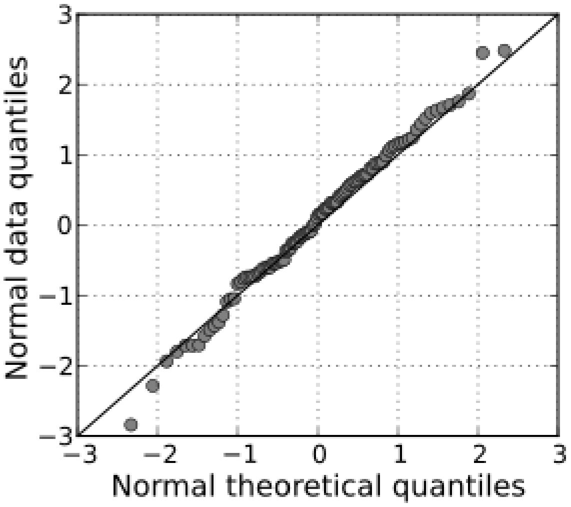
Ans: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing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

Ans: A Q–Q plot (*Quantile-Quantile plot*) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.



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.

Advantages:

1) It can be used with sample sizes also
2) 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 —

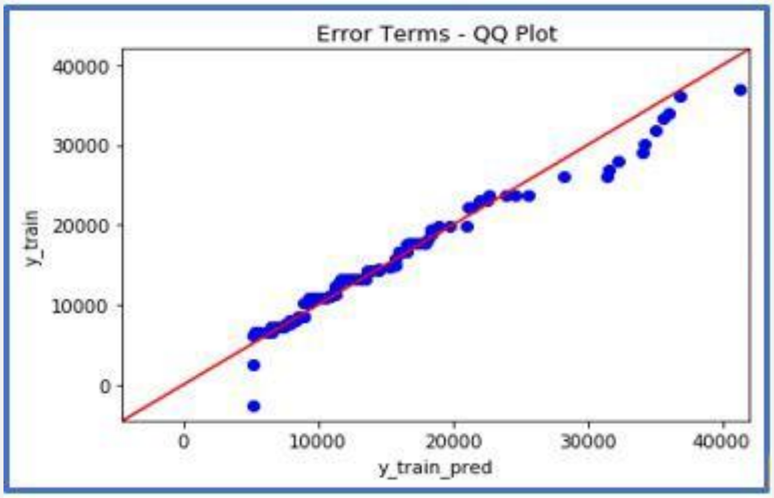
a) come from populations with a common distribution
b) have common location and scale
c) have similar distributional shapes
d) have similar tail behavior

Interpretation:

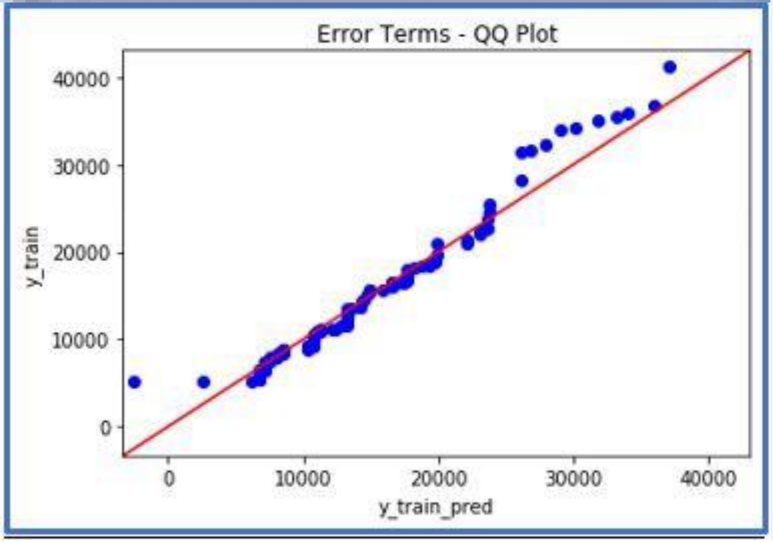
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

I. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
II. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



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



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