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?

   Answer: Categorical variable analysis

   - Season: Season 3 has highest demand for rental bikes
   - Year: Bike sharing demand grows next year.
   - Month: Bike demand grows until June, drops in July, august and again increases in September & oct, followed by a drop in last quarter. It could be due to bad weather conditions.
   - Holiday: In holiday period demand increases
   - Weekday & working day: It has almost no impact on bike sharing
   - Weather: The weathersit 1(Good) has highest demand followed by 2 (Moderate) & 3 (Bad)

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

   Answer:

   - In regression model, categorical variables are represented as dummy variables. If a categorical variable has n levels, it is represented as n-1 level i.e. one of the column values can be derived from others.
   - Dropping the first variable helps in reducing the extra column created during dummy variable creation, which further reduces the correlation created among dummy variables.

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

   Answer: “temp” has the highest correlation with the target variable cnt.

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

   Answer: After evaluating the final model following characteristics can be seen:

   - Residual analysis i.e. no relationship between the residual (error) terms.
   - Error terms have constant variance and normally distributed around mean.
   - No multi-collinearity between the independent variables and all variables have variance inflation factor i.e. (VIF) < 5
   - A linear relationship between independent and dependent variables can be seen by plotting the graph.

   Above characteristics points that linear regression assumptions hold true

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

   Answer: Top 3 features contributing towards the final model are:

   1. **Temperature (temp):** With coefficient value of 0.5981, it indicates that a unit increase in temp increases the bike sharing by 0.5981.

   2. **Year (yr):** With coefficient value of 0.2391, it indicates that a unit increase in year increases the bike sharing by 0.2391.

   3. **Weather situation 3:** With coefficient value of (- 0.2586), it indicates that a unit decrease in temperate will increase the bike demand by 0.2586

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression falls under one of the machine learning method “supervised learning” where a output variable is predicted in form of continuous variable. Linear regression can be defined as:

Simple linear regression: It explains the relationship between a dependent variable(Y) and one independent variable(X) using a straight line on scatter plot. It can be described as:

Y = C+ m*X (where c is intercept when X = 0, and m is the slope of the line)

Multiple linear regression: It explains the linear relationship between one dependent variable(y) and multiple independent variables (X) on a scatter plot. It can be described as:

Y = C + m1*X1 + m2*X2 + m3*X3 + ….+ C (

where c is intercept on Y axis and m1,m2….. are the coefficients obtained by keeping other variables constant which describes a one change in unit may increase/decrease the value in Y by keeping other variables constant.

In linear regression, a best fit line is drawn on a scatter plot between data points. For ex: consider a below graph.

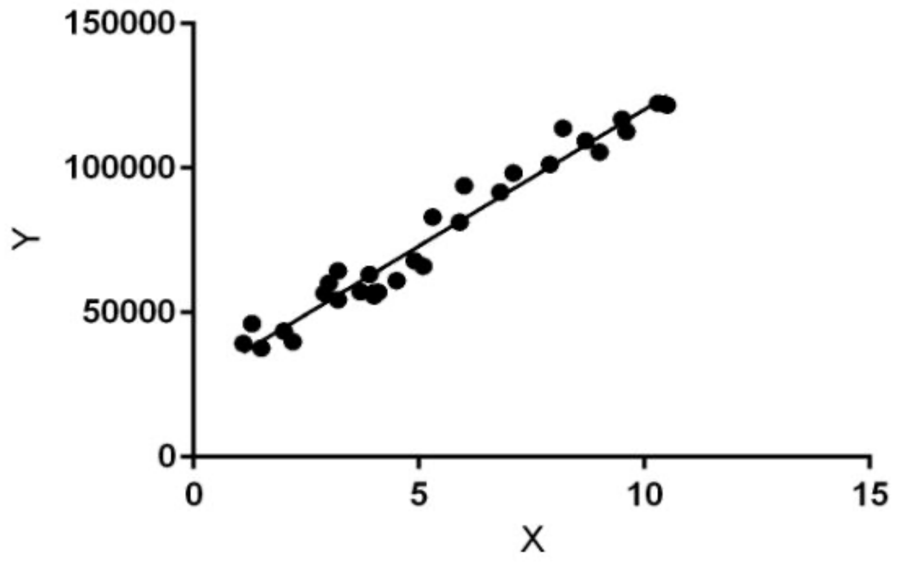
A best fit line can be described in terms of equation as mentioned above, where the strength of a model is represented by R2 (r-square) and it can be described as:

R^2 = 1 - (RSS / TSS) where

RSS (residual sum of squares) = sum of (Y_{actual} – $Y_{predicted}$)^2

TSS (total sum of squares) = sum of (Y_{actual} – Y mean)^2

In linear regression model when a best fit line is drawn, our objective is to minimize the cost function which is described below.



$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Error terms or cost function.

2. Explain the Anscombe’s quartet in detail.

Answer:

Anscombe’s quartet emphasizes on the importance of data visualisation before analysing and model building. A number of datasets may exhibits the same behaviour i.e. variance, and mean of all X,Y points in our data set.

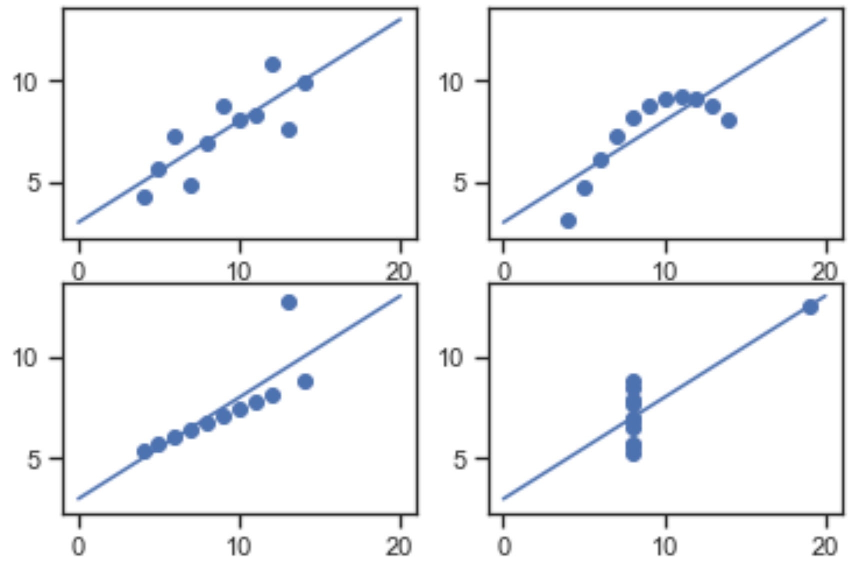
For ex: consider four data-sets represent below:

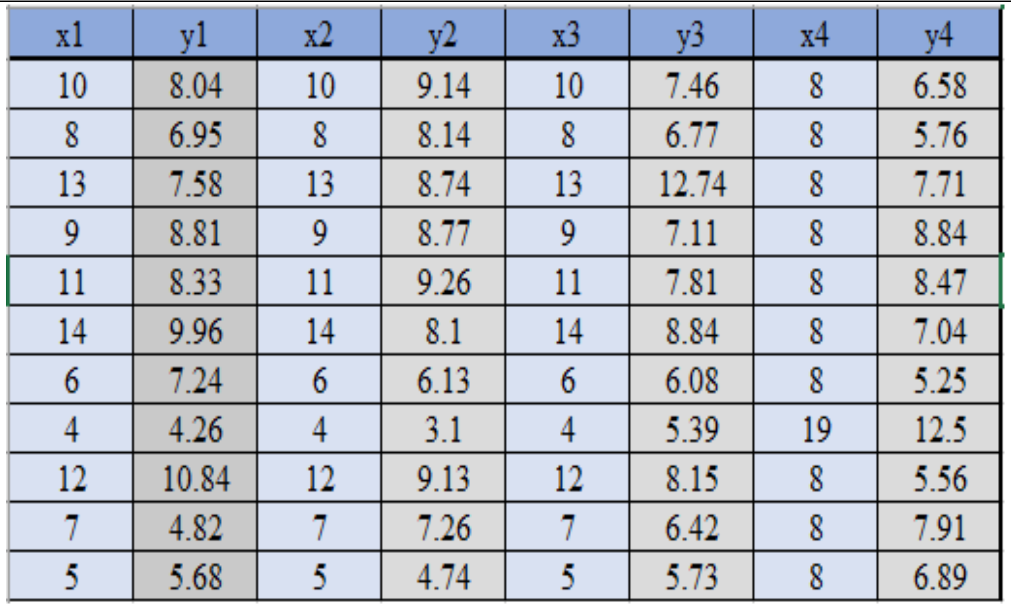
- There average value of X = 9
- Variance (X)= 11
- Variance (y) = 7.50
- Correlation coefficient = 0.816

The statistical analysis of these datasets are approximately similar, but when we plot these four data-sets they exhibit different behaviour where:

1. Data-set 1: Exhibits a liner relationship between datapoints
2. Data-set 2: Exhibits a non-linear relationship between data points
3. Data-set 3: Exhibits a linear relationship with outliers which cannot be handled by linear regression model.
4. Data-set 4: represents the outliers which cannot be explained by linear regression model as well.

To summarize before building a machine learning model datasets must be visualized to understand the relationship between different variable.





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

3. What is Pearson’s R?

Answer:

In statistics, Pearson’s correlation coefficient or Pearson’s r measures the strength of the relationship between two variables and their association with each other.

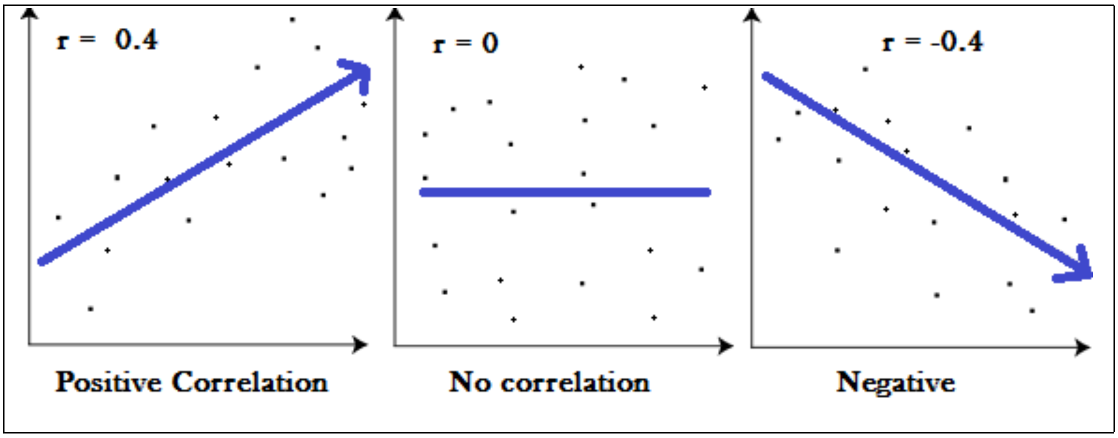
The Pearson’s r can take the range of values from +1 to -1, where

.

Pearson’s r	Description
Zero (0)	indicates that there is no association between the two variables
< 0	Shows a negative relationship i.e. variable value increases w.r.t to decrease in value of other variable
> 0	Shows a positive relationship i.e. variable value increases w.r.t to increase in value of other variable

Pearson’s R is only valid if following criteria’s meet else it may provide invalid result.

1. Two variables should be measured on continuous scale
2. Continuous variables should be paired i.e. each case has two values which is referred as data points.
3. There should be a relationship between two continuous variables.
4. Both the variables should follow a normal distribution.
5. Homoscedasticity i.e. variances along the line of best fit remain similar as you move along the line.
6. There should be no univariate or multivariate outliers i.e. an observation within the sample that does not follow a similar pattern to the res



Positive Correlation

No correlation

Negative

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

Answer:

In multiple linear regression model, a lot of independent variables are analysed which may be in a different scale which may lead a model with very weird Coefficients that might be very difficult to interpret. Scaling is required for primarily two reasons

1. Ease of interpretation
2. Faster convergence for gradient descent methods

Scaling only affects coefficients and interpretation not the model predictions and parameters like (t-statistics, f-statistics, p-values, r-squared).

Scaling is performed in two ways:

1. **Standardized scaling:** The variables are scaled in such a way that their mean is zero and standard deviation is one.

   X = x – mean(x) / sd(x)

2. **Min max scaling/Normalization:** Variables are scaled in such a way that it lies in range of zero and one.

   X = x - min(x) / max(x) – min(x)

Normalization is preferred over standardization because it loses some information in the data, especially about outliers.

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

Answer:

Variance inflation factor (VIF) measures the collinearity between predictor variables within a multiple regression model. The higher the value, the greater is the correlation between variables.

VIF = infinity, shows a perfect correlation between two independent variables. In case of a perfect correlation (i.e. R2=1) leads to 1/(1-R2) infinity. To solve the problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer:

Q-Q plot is also known as quantile-quantile plot; It’s a graphical tool which helps to determine if two possible sets of data came from same theoretical distribution such as normal/exponential or uniform distribution.

In linear regression, if training data sets and test data-sets received separately, we can confirm by using Q-Q plot i.e. both are from populations with same distributions.

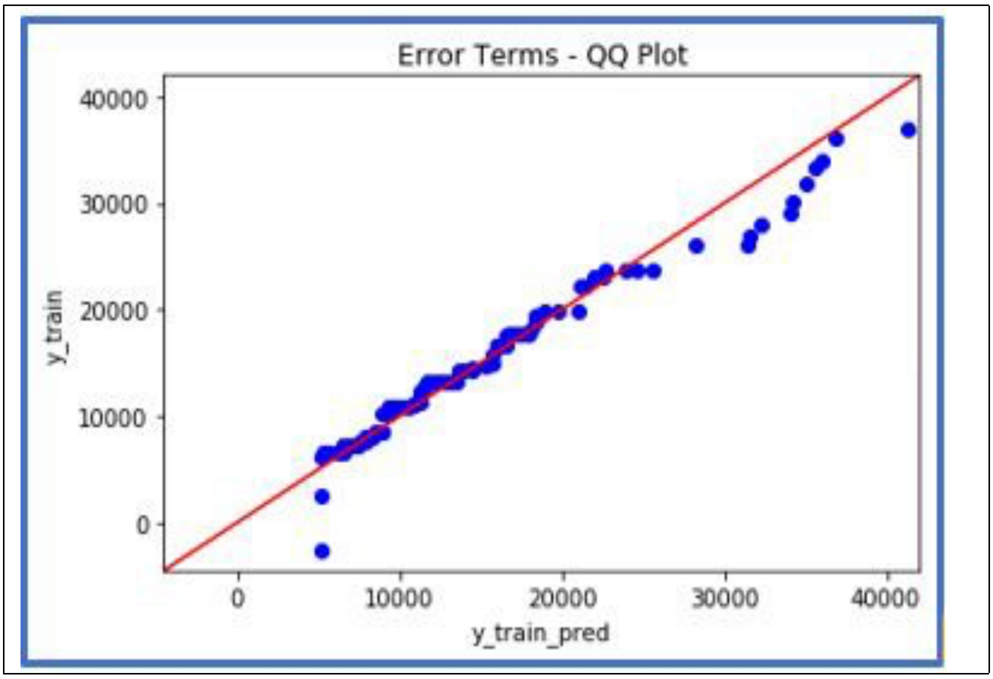
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Below are possible interpretation for two data sets:

a) Similar distributions:
   - If all point of quantiles lies on or close to straight line which is at 45 degree from x-axis.

b) Y-values < X-values:
   - Y-quantiles are lower than X-quantiles.

c) Y-values > X-values:
   - X-quantiles are lower than 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.


Error Terms - QQ Plot
y_train
y_train_pred