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)**

   Answer :

   I have done analysis on following categorical variables :

   1. season
   2. mnth
   3. weathersit
   4. holiday
   5. weekday
   6. workingday
   7. year

   Their effect on dependent variable “cnt” are as follows :

   1. Bulk of bike booking happening in “**Summer**” and “**Fall**” season with a median of over 5000 bike booking.
   2. Most of the bike booking happening in the months of “**May**”, “**June**” “**July**”, “**August**”, “**September**” and “**October**” with a median of over 4000 bike bookings.
   3. “**Clear**” weather situation having most of the bike booking with a median over 5000 followed by “**Misty**” weather situation.
   4. Large bike booking happening when it is **“Not a holiday”**.
   5. “**Thu**”, “**Fri**”, “**Sat**” & “**Sun**” have more number of bookings as compared to other weekdays with a median between 4000 to 5000 bookings.
   6. Bike booking seeming to be same for “**Working day”** or “**Non-working day”**.
   7. Boxplot shows better bike bookings happening in **“Year 2019”** than previous year 2018.

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

   Answer :
   The feature drop_first = True,We use while creating dummy variables, It helps in reducing the extra column created during dummy variable creation. It reduces the correlations created among dummy variables.

   It drops the first column during dummy variable creation. The goal is to reduce the number of columns by dropping the column that is not necessary. However, it is not always true. For some situations, we need to keep the first column.
   If we have categorical variable with k-levels, then we need to use k-1 columns to represent the dummy variables.

Example : Suppose we have a column for gender that contains 4 variables – “Male”, “Female”, “Other” and “Unknown”. So a person is either “Male” or “Female” or “Other”. If they are not either of these 3 then their gender is “Unknown”.

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

   Answer :

   From the pair-plot of the numerical variables, “**temp**” variable 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? (3marks)**

   Answer :

   I have validated the assumptions of Linear Regression after building the model based on below 5 assumptions :

   1. **Linear relationship between dependent and independent variable :**
      - I have plotted a simple “pair-plot” and checked linear relationships.
   2. **Multicollinearity :**
      - I have checked that “VIF values” of all predictor variables should be less than 5.
   3. **Homoscedasticity :**
      - I have used “scatterplot” to check any visible pattern present for residuals.
   4. **Error terms should be normally distributed :**
      - To validate this assumption, I have plotted “distribution plot” and check that residuals are normally distributed with mean 0.
   5. **Independence of residuals :**
      - I have checked that there is “Auto-correlation” for residuals.

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

   Answer :

   The top 3 features that contributing significantly towards explaining the demand of the shared bikes are:

   1. temp
   2. year
   3. sept

General Subjective Questions

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

Answer :

Linear regression algorithm is a machine learning algorithm based on supervised learning where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range.

Linear regression analysis is a technique of predictive modeling that helps to find out the relationship between input and the target variable.

Linear regression analysis is used for :

a. Finding out the effect of Input variables on Target variable.

b. Finding out the change in Target variable with respect to one or more input variable.

c. To find out upcoming trends.

There are main two types of Linear Regression :

1) Simple Linear Regression : It uses traditional slope-intercept form, where we train a model to predict the behaviour of our data based on some variables. Simple linear regression used to predict a quantitative response “y” from the predictor variable “x”.

Equation :

y = mx + b

Where,

m = slope of the line

x = Independent variable from dataset

b = Intercept of the line

y = Dependent variable from dataset

2) Multiple Linear Regression : It’s used to estimate the relationship between two or more independent variables and one dependent variable.

Equation :

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

Where,

y = The predicted value of the dependent variable

B_0 = The y- intercept

B_1X_1 = The regression coefficient B_1 of the independent variable X_1

B_nX_n = The regression coefficient of thr last independent variable

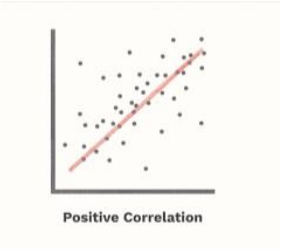
e = Model error

Assumptions for Linear Regression :

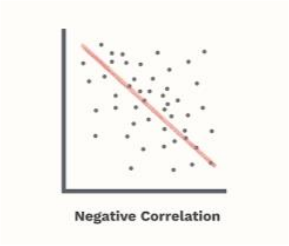
1. Data should be linear
2. No Multicollinearity
3. No auto-correlation
4. Homoscedasticity should be there means there should be no visible patterns in rediduals.
5. Error terms should be normally distributed to mean 0.

Types of Linear Relationship :

1. **Positive linear relationship** : A linear relationship will be called Positive if both independent and dependent variable increases.



2. **Negative linear relationship** : A linear relationship will be called Negative if independent variable increase and dependent variable decreases.



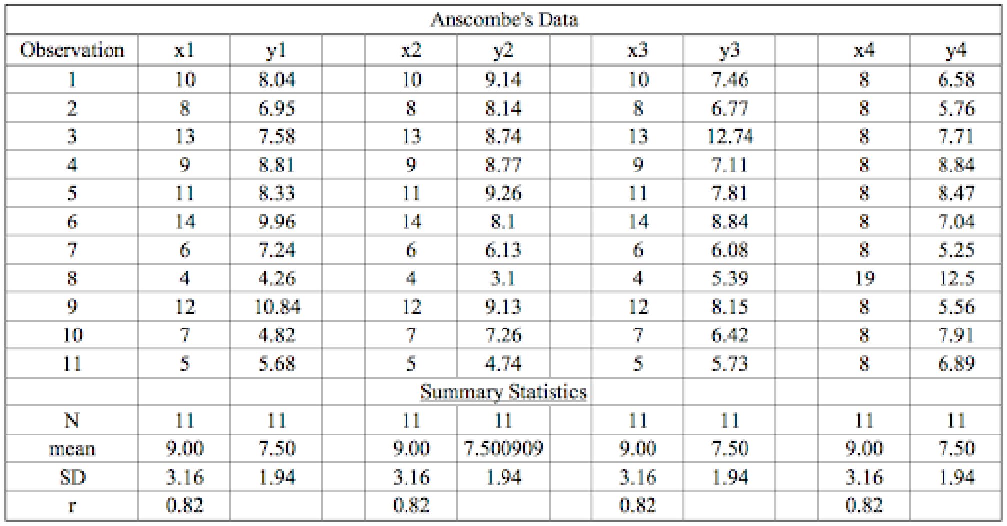
2. **Explain the Anscombe's quartet in detail.** **(3 marks)**

Answer :

Anscombe quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but there are peculiarities that fool the regression model once you plot each data set.

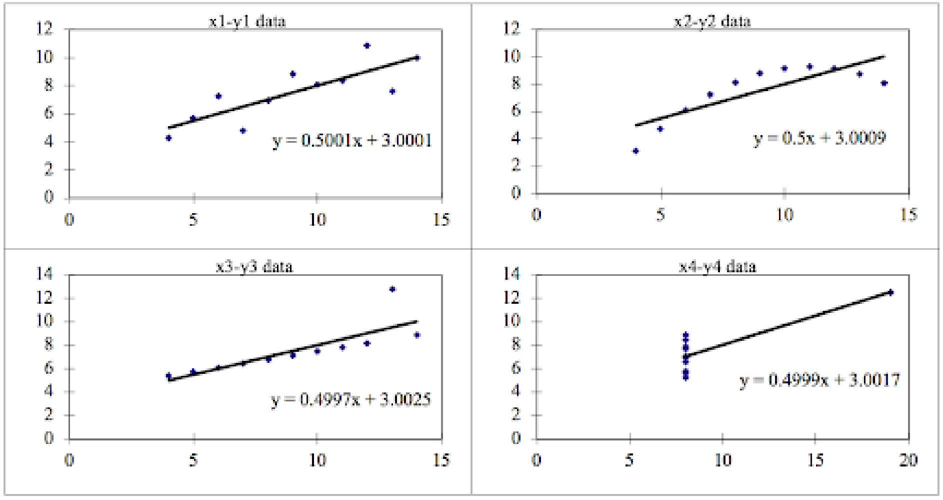
Anscombe quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. It comprises four datasets, each containing eleven (x,y) pairs. These datasets share the same descriptive statistics but, when they graphed each graph tells a different story irrespective of their similar summary statistics.

The statistical information of these four data set are as follows :



					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
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

These four data sets have similar statistical information, however if these models are plotted on scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm.



We can describe the four data sets as :

- Data set 1 : Fits the linear regression model pretty well.
- Data set2 : Can't fir the linear regression model because the data is non-linear.
- Data set 3: Outliers involved in the data set, which can't be handled by the linear regression model.
- Data set 4: Outliers involved in the data set, which can't be handled by the linear regression model.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the data set.

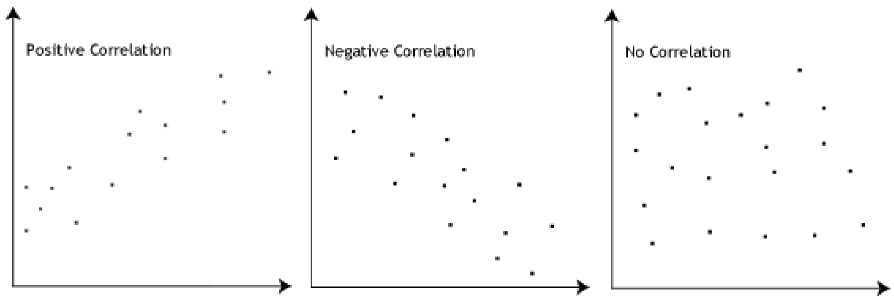
3. **What is Pearson's R? (3 marks)**

Answer :

The Pearson's Correlation Coefficient is also referred as Pearson's R, Bivariate correlation or Pearson's product-moment correlation coefficient (PPMCC). It is a statistic that measures the linear correlation between teo variables. Pearson's R is the covariance of the two variables divided by the product of their standard deviations.

- The Pearson's R can take a range of values from + 1 to -1. A value of 0 indicates that there is No association between the two variables.
- A value greater than 0 indicates Positive association i.e As the value of one variable increase, so does the value of the other variable.
- A value less than 0 indicates a Negative association i.e As the value of one variable increase, the value of the other variable.

These three associations shown in the diagram below :



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

Answer :

Scaling is a step of Data-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculation in a an algorithm. Scaling is performed because most of the time 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 it leads to incorrect modeling. To solve this issue, we have to do scaling to bring all the variable to the same level of magnitude.

1. **Normalized Scaling/MinMax Scaling :**

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

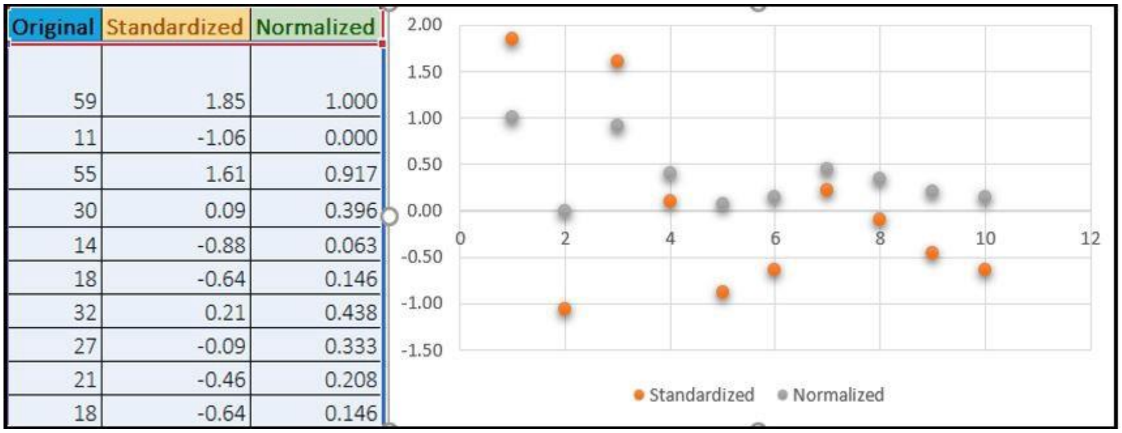
$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

2. **Standardization Scaling :**

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

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

Representation of Normalized & Standardized Scaling :



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

Difference between Normalized Scaling & Standardized Scaling :

Sr.No	Normalized Scaling	Standardized Scaling
1.	Minimum and Maximum values of the features are used for scaling.	Mean and Standard deviation is used for scaling.
2.	It is used when feature are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between (0, 1) (-1,1)	It is not bounded to a certain range.
4.	It is rarely affected by outliers.	It is much less affected by outliers.
5.	Scikit-learn provides a transformer called MinMaxScaler for Normalization	Scikit-learn provides a transformer called StandardScaler for Standardization.

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

Answer :

If there is perfect correlation between the variables, 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).

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 presence of multicollinearity.

To get rid of this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

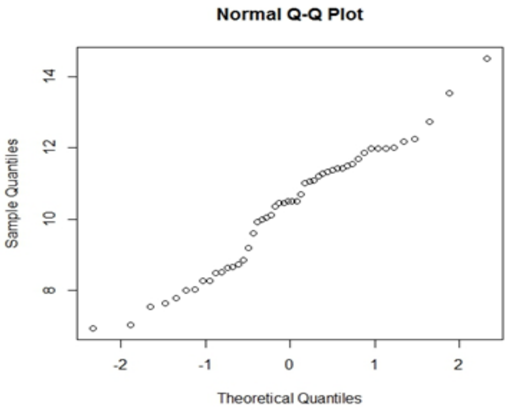
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer :

The Q-Q plot or Quantile-Quantile plot is a graphical technique for determining if two datasets are come from populations with a common distribution.

A Q-Q plot is a scatter plot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughlt straight.

Example of Q-Q plot when both sets of quantiles truly come from Normal Distribution.



Use of Q-Q plot in Linear Regression :

The Q-Q plot is used to see if the points lies approximately on the line. If they don't, it means, our residuals aren't Normal and thus, our errors also not Normal.

- **Importance of Q-Q Plot :**

  1. The sample sizes do not need to be equal.
  2. Many distributional aspect can be simultaneously tested. e.g Shifts in location, shifts in scale, changes in symmetry and presence of outliers.
  3. The Q-Q plot can provide more insight into the nature of the difference than analytical methods.