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 used the boxplot and bar plot to analyze categorical columns.
   The following are a few conclusions that may be drawn from the visualization:

   1. The autumn appears to have drawn more reservations. And, from 2018 to 2019, the number of bookings in each season significantly grew.
   2. Most reservations were made in the months of May, June, July, August, September, and October. Beginning in January and continuing through mid-year, the trend grew before beginning to decline as the year came to a close.
   3. It appears that more bookings were made during clear weather.
   4. There are more reservations on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.
   5. The number of reservations for non-holiday are more than holiday

   6. The number of reservations for 2019 increased over the prior year

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

   Answer:

   Use of drop_first = True is crucial since it helps in eliminating the excess column produced when a dummy variable is formed. As a result, it lessens the connections that dummy variables cause.

   Drop_first: bool, defaulting to False, indicates whether to remove the first level from the k category levels in order to obtain k-1 dummies.

   Let's imagine we want to build a dummy variable for a categorical column that has three different types of data. If one factor is neither A nor B, then it is clear that C. Thus, we do not require the third variable to locate the C.

Variable 1	Variable 2	Info
0	1	A is require
1	0	B is require
0	0	Neither A nor B require so C reqiure

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

   Answer:

   ‘temp’ variable has the highest correlation with the target variable.

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

   Answer:

   I have validated the assumption of Linear Regression Model based on below 5 assumptions -

   1. Normality of error terms
      - Error terms should be normally distributed
   2. Multicollinearity check
      - There should be insignificant multicollinearity among variables.
   3. Linear relationship validation
      - Linearity should be visible among variables
   4. Homoscedasticity
      - There should be no visible pattern in residual values.
   5. Independence of residuals
      - No auto-correlation

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

   Answer:

   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

   - temp
   - winter
   - sep

General Subjective Questions

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

   Answer:

   Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   Mathematically the relationship can be represented with the help of following equation −

   Y = mX + c

   Here, Y is the dependent variable we are trying to predict.

   X is the independent variable we are using to make predictions.

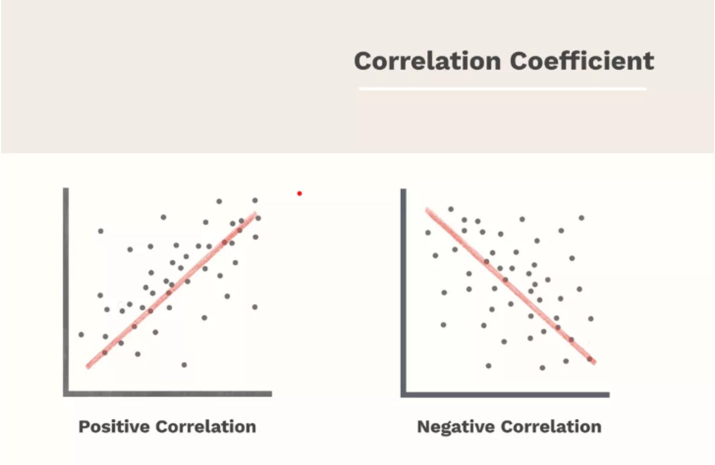
   m is the slope of the regression line which represents the effect X has on Y

   c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

   Furthermore, the linear relationship can be positive or negative in nature as explained below−

   - Positive Linear Relationship:
     - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of

following graph –



- Negative Linear relationship:
    - A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

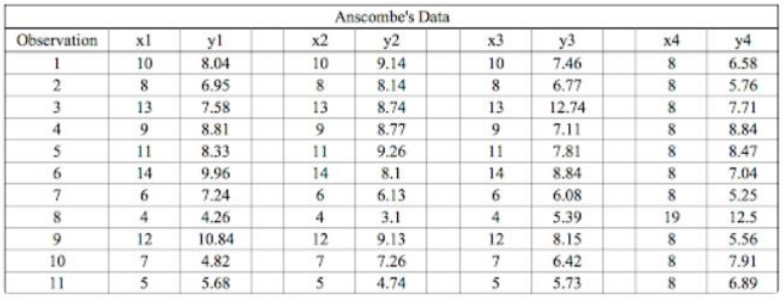
- Multi-collinearity –
    - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation –
    - Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables –
    - Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms –
    - Error terms should be normally distributed
- Homoscedasticity –
    - There should be no visible pattern in residual values.

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

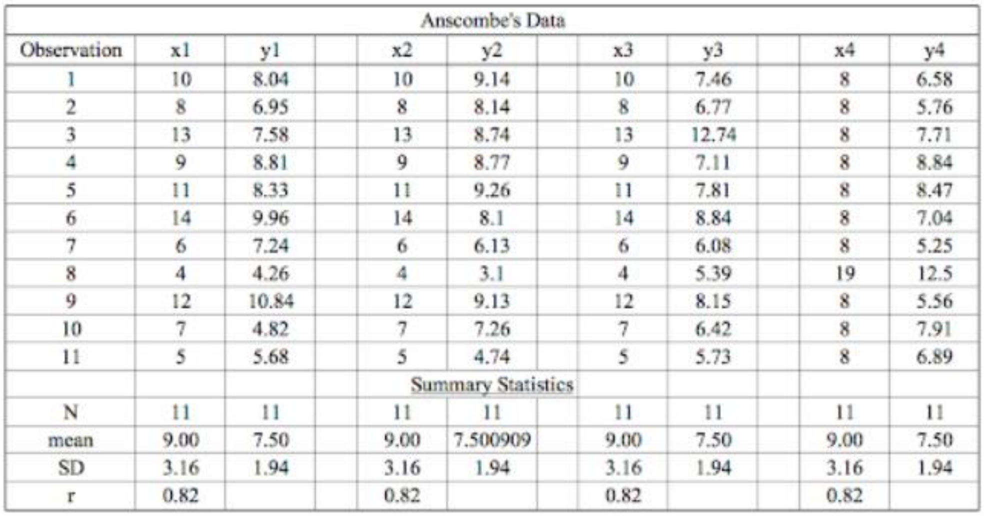
Answer:

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph The summary statistics show that the means and the variances were identical for x and y across the groups:

We can define these four plots as follows:

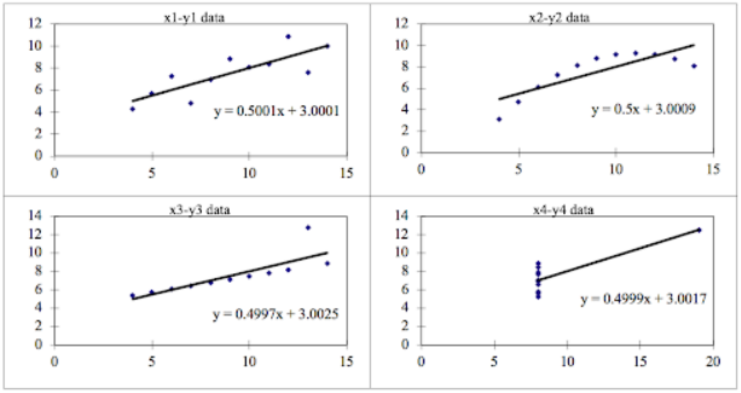


Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		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



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	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



ANSCOMBE’S QUARTET FOUR DATASETS

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

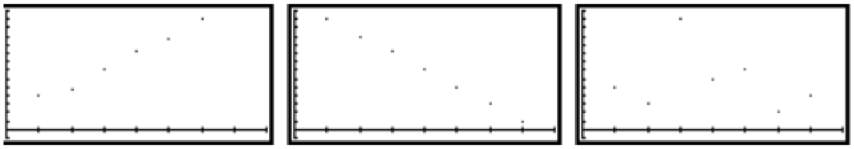
As we can see, Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. What is Pearson’s R? (3 marks)

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:
r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association
The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233



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

Answer:

It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations.

The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and range. If scaling is not done, the algorithm will only consider magnitude and not units, which will result in inaccurate models. We must scale all the variables to the same degree of magnitude in order to resolve this problem.

The t-statistic, F-statistic, p-values, R-squared, etc. are unaffected by scaling, which is significant because they are all dependent on the coefficients.

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:

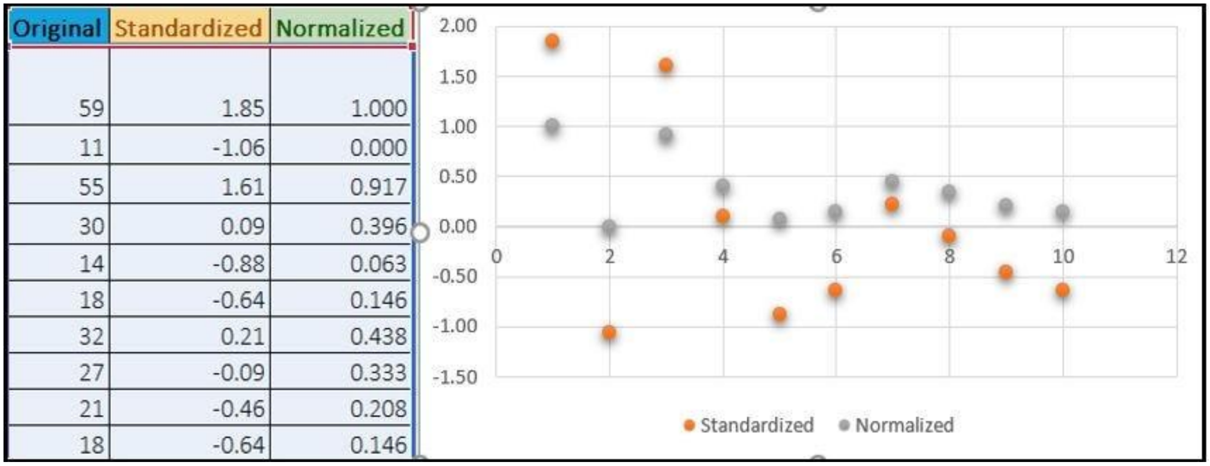
- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).*

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

- ***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? (3 marks)

Answer:

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 5, this means that the variance of the model coefficient is inflated by a factor of 5 due to the presence of multicollinearity

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing thisperfect 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 quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:
When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.