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)

   Following are the points we can infer from the analysis:

   a. Fall has highest demand for bike rental

   b. Demand for rental bike is increasing till June

   c. During weekdays and workingdays demands don't have that much of variation

   d. Clear weathersit has highest demand

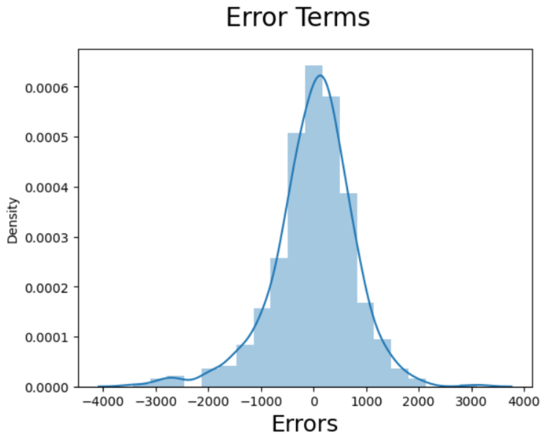
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

   It will helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations 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)

   ‘temp’ and ‘atemp’ 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)



Residuals distribution should follow normal distribution and centered around 0.(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not. The above diagram shows that the residuals are distributed about mean = 0

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

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

a. Temp
b. Month
c. Weather

General Subjective Questions

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

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, rather than trying to classify them into categories. It is a part of regression analysis. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression

Simple Linear Regression - Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to “learn” to produce the most accurate predictions. x represents our input data and y represents our prediction

$Y=mx+b$

Multiple Linear Regression - A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn

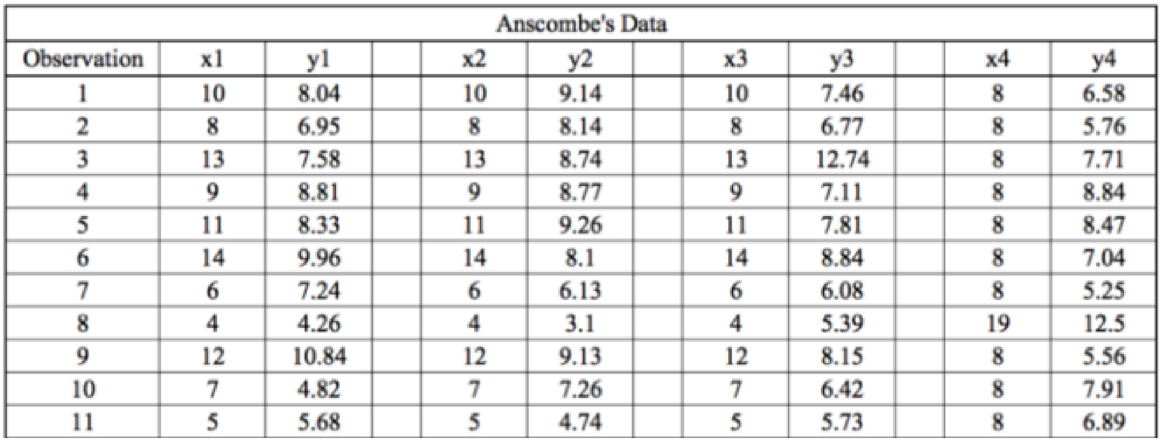
$f(x,y,z)=w1x + w2y + w3z$

Multiple linear regression analysis makes five key assumptions: 1. Linear relationship: There exists a linear relationship between each predictor variable and the response variable. 2. No Multicollinearity: None of the predictor variables are highly correlated with each other. 3. Independence: The observations are independent. 4. Homoscedasticity: The residuals have constant variance at every point in the linear model. 5. Multivariate Normality: The residuals of the model are normally distributed

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

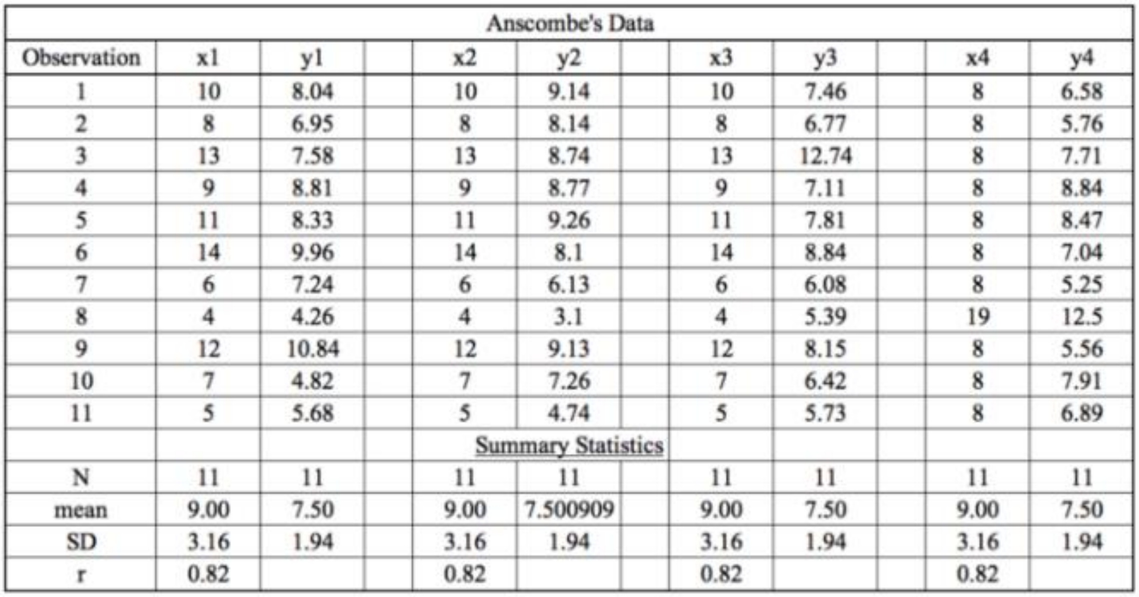
Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. Anscombe’s quartet intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

These four plots can be defined as below:



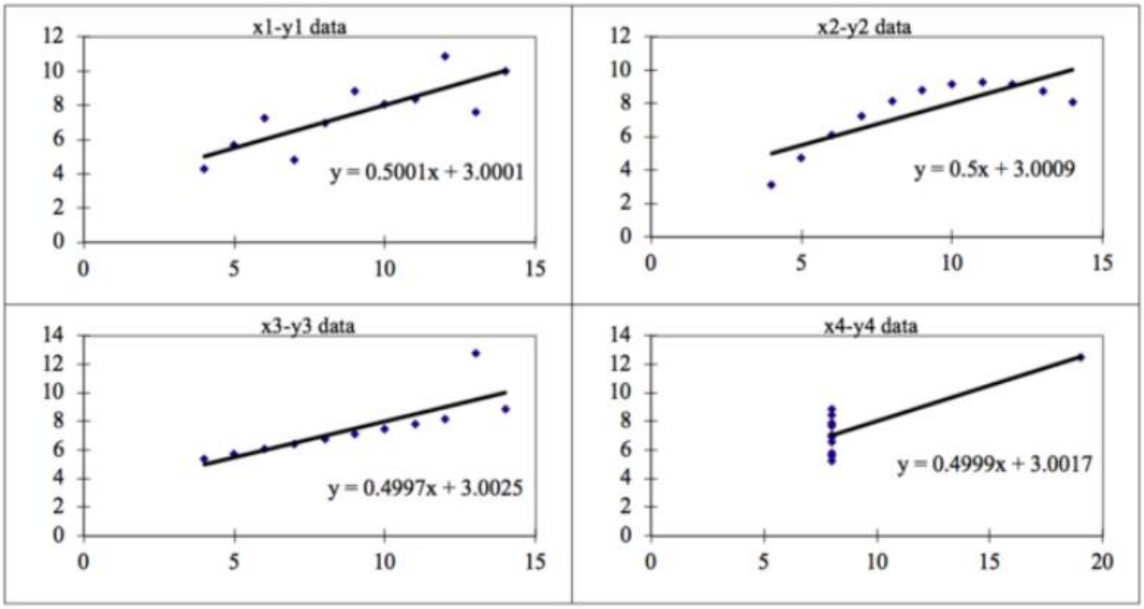
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08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

The statistical information for all these four datasets are approximately similar and can be computed as below:



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	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as below:



The four datasets can be described as below:

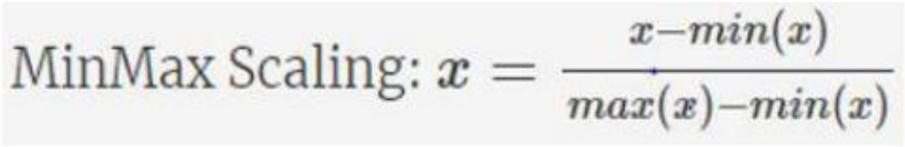
1. The first scatter plot (top left) appears to be a simple linear relationship.
2. The second graph (top right); cannot fit the linear regression model because the data is non-linear.
3. In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. It shows the outliers involved in the dataset which cannot be handled by linear regression model.
4. The fourth graph (bottom right) shows the outliers involved in the dataset which cannot be handled by linear regression model. It shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables. It shows the outliers involved in the dataset which cannot be handled by linear regression model.

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

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 correlation coefficient, 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 a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

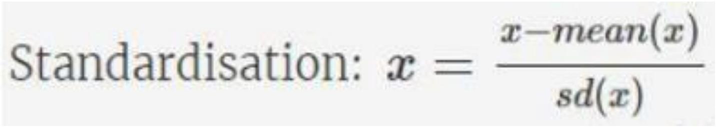
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. In simpler terms, in machine learning algorithms we need to bring all features in the same standing, so that one significant number doesn't impact the model just because of their large magnitude. This is called scaling or Feature scaling. Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. 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 which results in an incorrect model. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t statistic, F-statistic, p-values, R-squared, etc.
Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks. It brings all the data in the range of 0 and 1.



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

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization. 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).



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

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

   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 4, this means that the variance of the model coefficient is inflated by a factor of 4 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 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)

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.

Advantages:

a. The sample sizes do not need to be equal.
b. Many distributional aspects can be simultaneously tested

Usage:

a. If two populations are of the same distribution
b. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it’s met using this.
c. Skewness of distribution