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:
Based on the analysis done on categorical columns below are some points we can infer using various plots–

- End of the year months from November to February has lowest no of users. However most of the bookings has been done during the month of May, June, July, Aug, Sep and Oct.
- Working day or Non-Working day doesn't make much difference to Number of users.
- Fall season has max users followed by summer and winter has least users.
- There are more number of booking on Holidays as compared to Working days which makes sense.
- Year 2019 had more users as compared to year 2018.

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

Answer:
drop_first = True helps in reducing the extra column created during dummy variable creation. It also reduces the correlations created among dummy variables.

For example we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not X and Y, then It is obvious Z. So we do not need 3rd variable to identify the Z.

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:
Linear Regression Model was validated based on below assumptions –

- Residual Analysis
  Difference between the actual values and the model predicted values should be centred on 0 mean.

- Multicollinearity check
  Multiple dependent columns should not be present in the model. It was validated using VIF values.

- Linear relationship
  Linear relationship was measured using the pair plots.

- R Square and P value

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
- Saturday

General Subjective Questions

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

Answer:
Linear regression is a statistical model which is used to analyse 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 representation −
Y = mX + c

- ➢ 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.

Linear regression is of the following two types −

- ➢ Simple Linear Regression
- ➢ Multiple Linear Regression

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

✓ Multi-collinearity –
Multi-collinearity means Associations between predictor variables. In 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.

✓ Relationship between variables –
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.

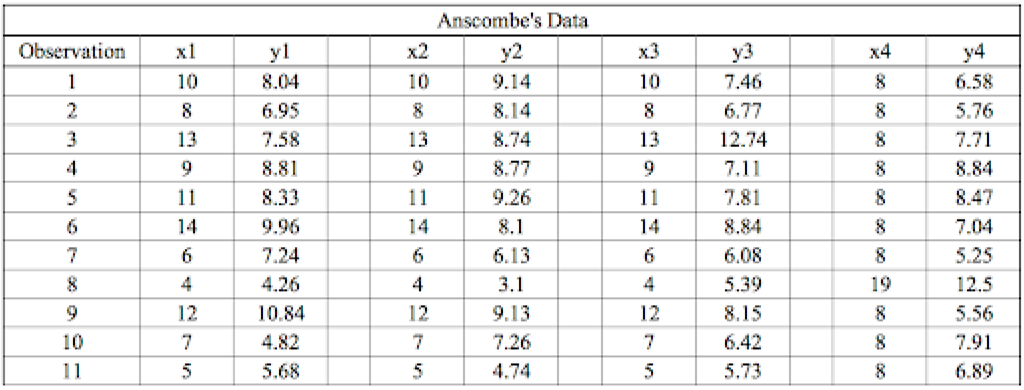
2. **Explain the Anscombe's quartet in detail. (3 marks)**

Answer:

Anscombe's quartet 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.

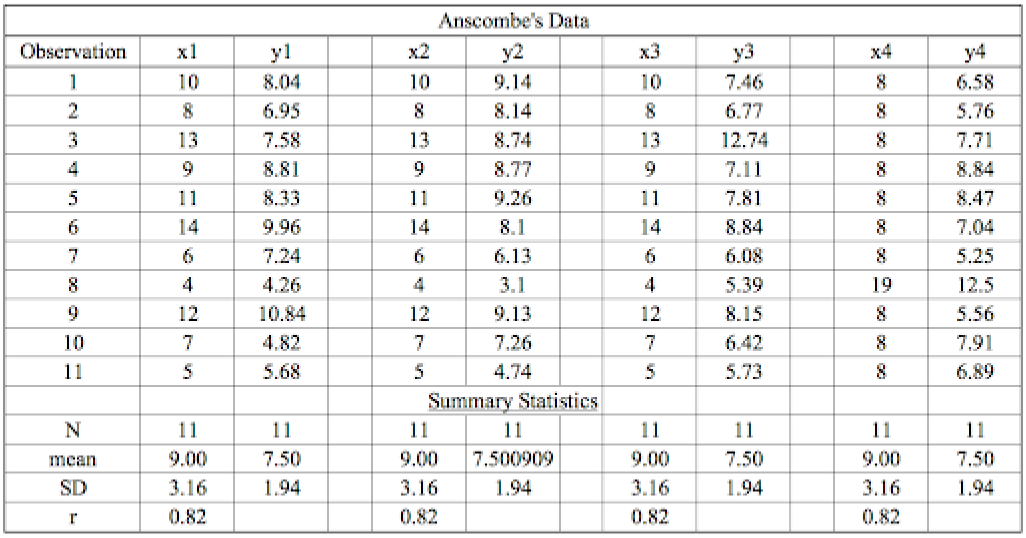
Purpose:
Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

We can define these four plots as follows:



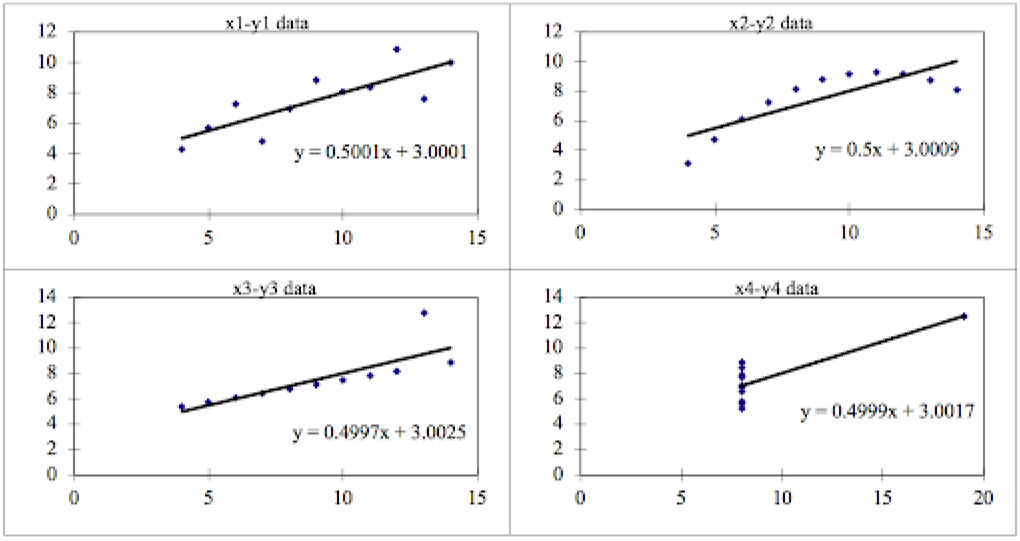
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08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

The statistical information for these four data sets are approximately similar. We can compute them 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	

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.

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

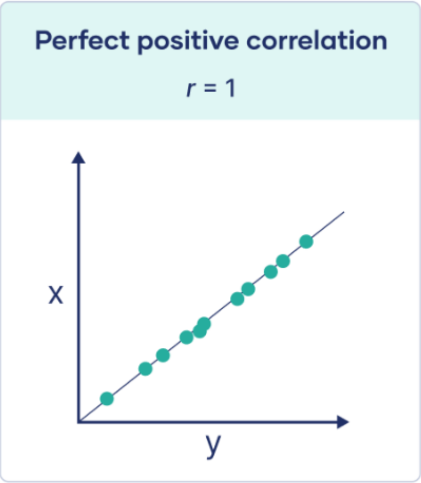
Answer:

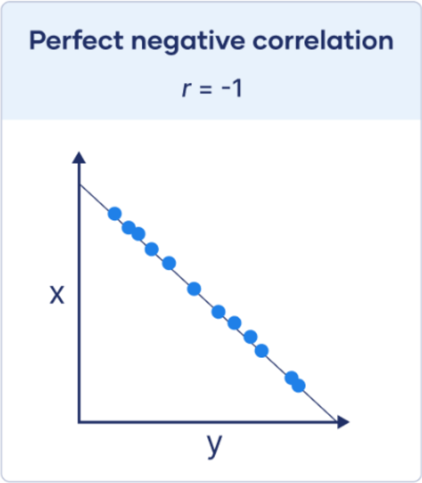
The Pearson correlation coefficient (r) is a way of measuring a linear correlation. The value lies between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None

Pearson correlation coefficient (r) value	Strength	Direction
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

Visualizing Pearson Coefficient:





Formula: -

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y}^2 - (\sum \mathbf{y})^2]}}$$

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

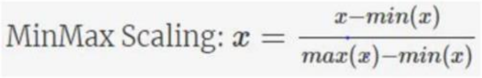
What is scaling - Scaling is a process of putting all the independent variables in the linear regression in the defined range so that it is easy to make correct calculations. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

Why is scaling performed -In case scaling is not done, then a machine learning algorithm considers greater values, higher and consider smaller values as the lower values, irrespective of the unit of the values.

Example:- If an algorithm is not using scaling method then it can consider the value of 50 inches to be greater than 20ft.

Normalized Scaling: It brings the data in the range of 0 to 1.

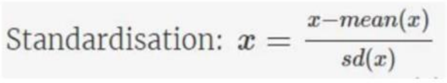
Formula:



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

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).

Formula:



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

Standardization Scaling vs Normalized Scaling:

Normalized Scaling	**Standardization Scaling**
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization