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?

a. The median of Clear weather box plot is higher than the cloudy followed by raining season . The rentals were more when the weather was clear.

b. In the Fall season we have seen High count of bike rentals.

c. In the year 2019 as we move away from covid, we have seen High count of bike rentals.

d. In the medians are almost same for Jun – Oct month. We see Sept – Oct has the high count of rentals.

e. Medians were almost same for all the days – whether weekday or weekend (Mon - Sun)

f. Holiday box plot – has large area - means more no. of people rented bikes.

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

When we create dummies, for 3 values in one categorical variable 3 dummies are created.

Eg Refer below table

High	Medium	Low
0	0	1
0	1	0
1	0	0

So when the High and medium are 0 – then definitely – low would be One. So we don't need extra colum to depict this and this will also avoid multicollinearity.

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

Temperature, Atemp (feeling of the temperature) – as we can see – it almost forms a straight line.

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

a. By creating residual analysis of errors – it was normally distributed, the centre at 0.0

b. We are able to create a linear relationship between x & Y variables and define the line equation.

c. Constant Variance of error terms were seen by plotting scatter plot.

d. Independence of error terms.

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

a. Temperature, Winter season, September month

General Subjective Questions

1. Explain the linear regression algorithm in detail.

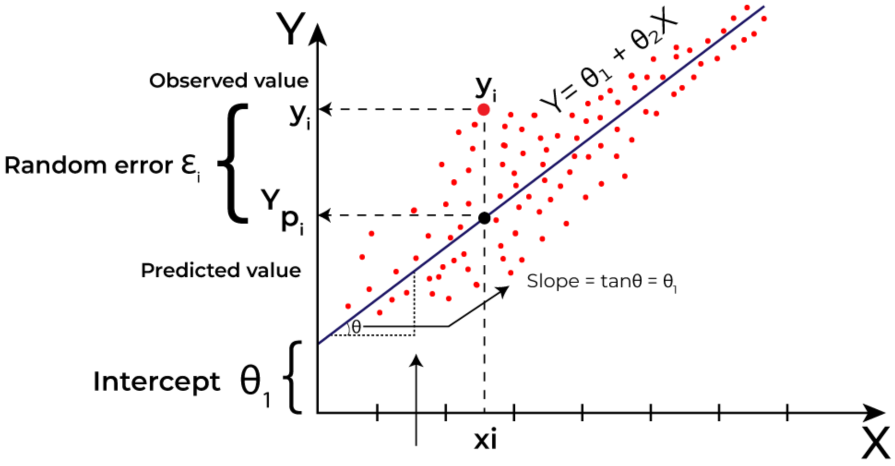
Linear regression is a type of supervised machine learning algorithm that calculates the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to the observed data.

If there is only one Independent variable then it is called simple linear regression, represented by below equation:

$Z= \theta1 + \theta2X$

If it has more independent variables, then called multiple linear regression.

$Z= \theta1+ \theta2X+ \theta3X+\dots\dots\dots \theta NX$


Y
Observed value
y_i
Random error $\mathcal{E}_i$
Y_{p_i}
Predicted value
Intercept θ_1
y_i
$Y= \theta_1 + \theta_2X$
Slope = tanθ = θ_1
θ
xi
X

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

To achieve the best-fit regression line, the model aims to predict the target value such that the error difference between the predicted value and the true value Y is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimizes the error between the predicted y value (pred) and the true y value (y).

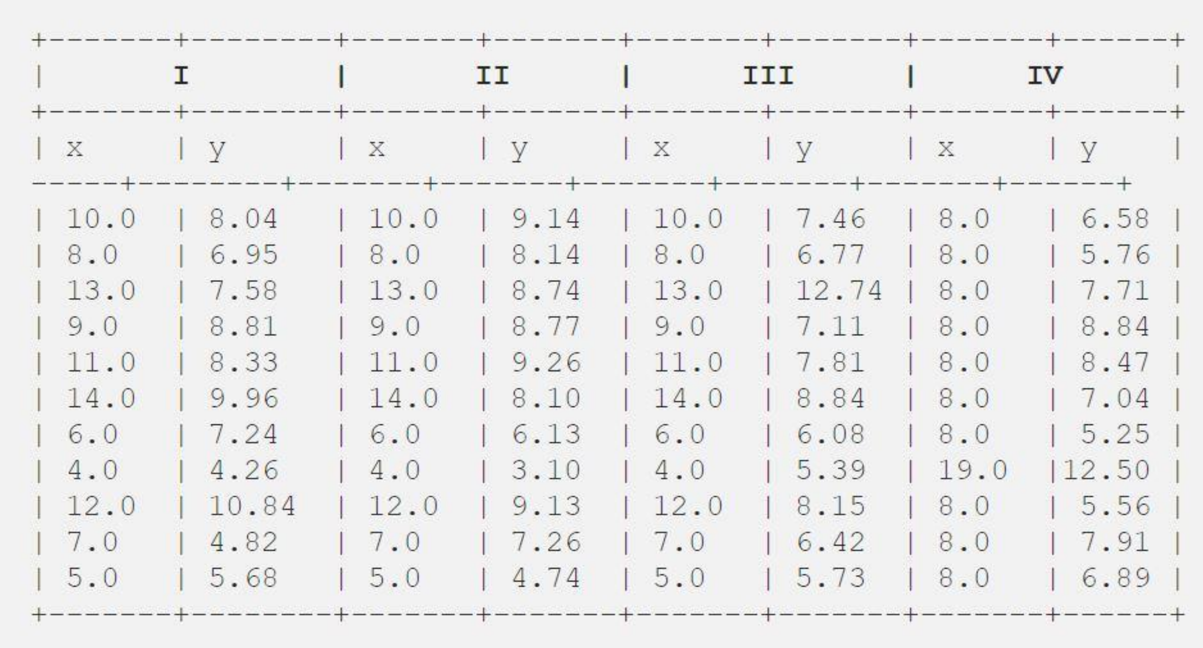
Assumptions of simple linear regression

- **a.** Linear relationship between X and y.
- **b.** Normal distribution of error terms.
- **c.** Independence of error terms.
- **d.** Constant variance of error terms.

2. Explain the Anscombe’s quartet in detail.

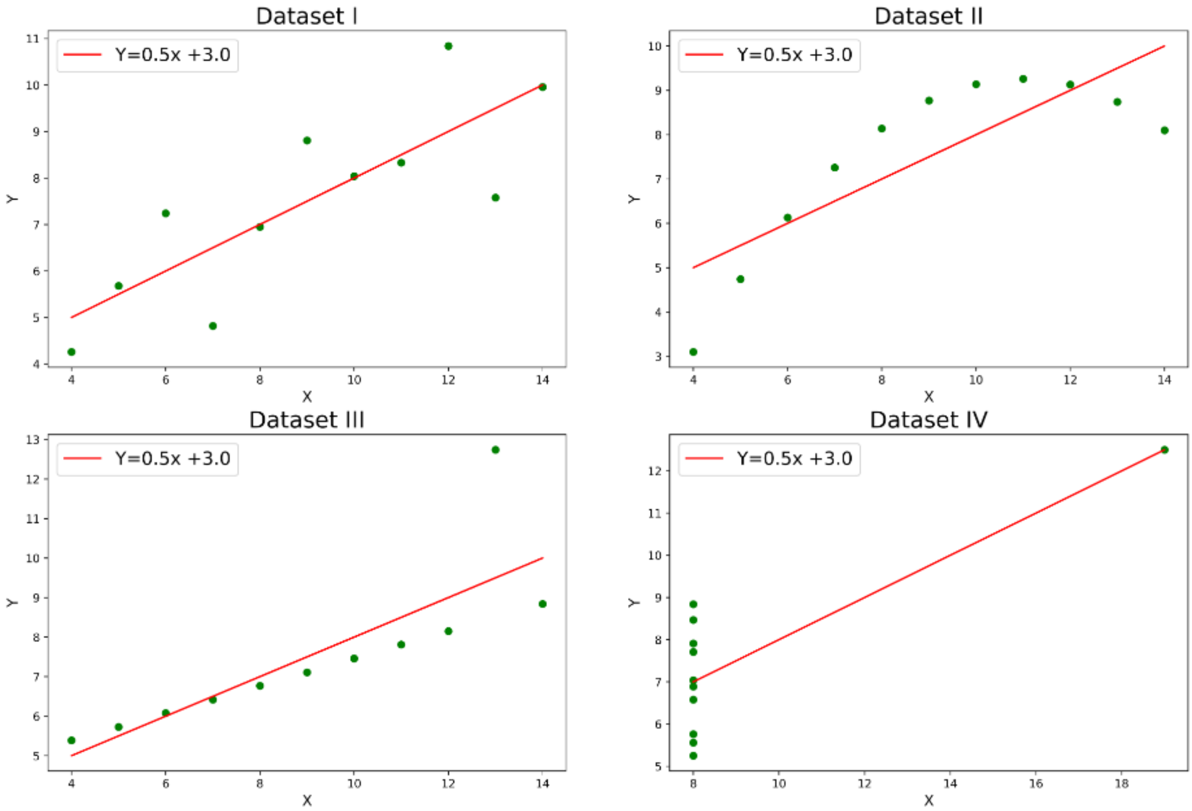
Anscombe's quartet consists of a set of four data sets, which have identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines, but have different representations when we scatter the plots on the histogram.

Each of the four data sets that make up Anscombe's quartet includes 11 pairs of data. When plotted, each data set appears to have a unique connection between x and y, with unique patterns of variation and distinct correlation strengths. Despite these differences, each data set contains the same summary statistics, such as the same mean and variance x and y, correlation coefficient x and y, and linear regression line.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's quad is used to illustrate the importance of exploratory data analysis and the disadvantages of relying solely on summary statistics. It also emphasizes the importance of using data visualization to identify trends, outliers, and other important details that may not be apparent from summary statistics alone.



Explanation of this output:

- In the first figure (top left) if you look at the scatter plot, you will see that there appears to be a linear relationship between x and y.
- In the second image (top right) If you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third point (bottom left) you can say when there is a perfect linear relationship for all data points except a point that appears to be an outlier which is referred to as being far from this line.
- Finally, the fourth (bottom right) shows an example when a high leverage point is sufficient to produce a high correlation coefficient.

Conclusion

While the descriptive statistics of Anscombe's quadrilateral may appear uniform, the accompanying visualizations reveal distinct patterns, and showcase the necessity of combining statistical analysis with graphical exploration for robust interpretation of the data.

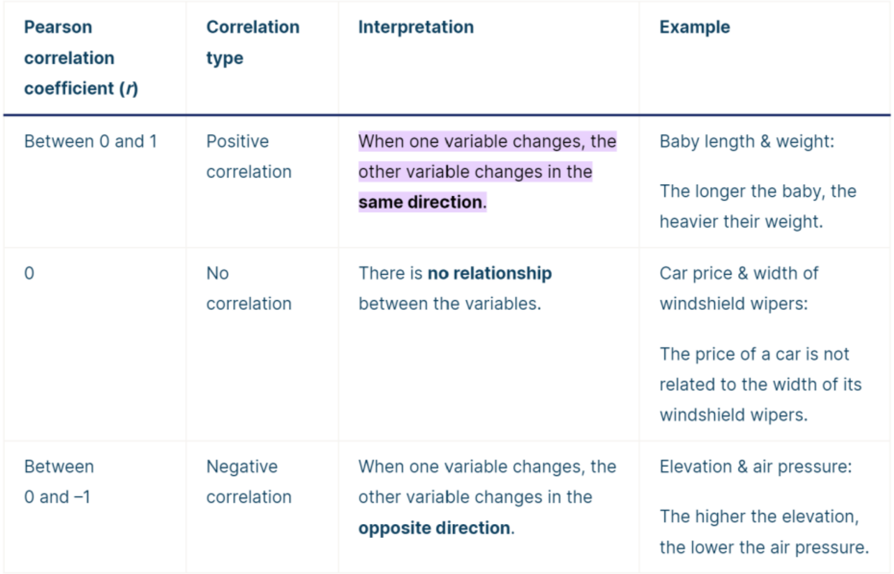
3. What is Pearson’s R?

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

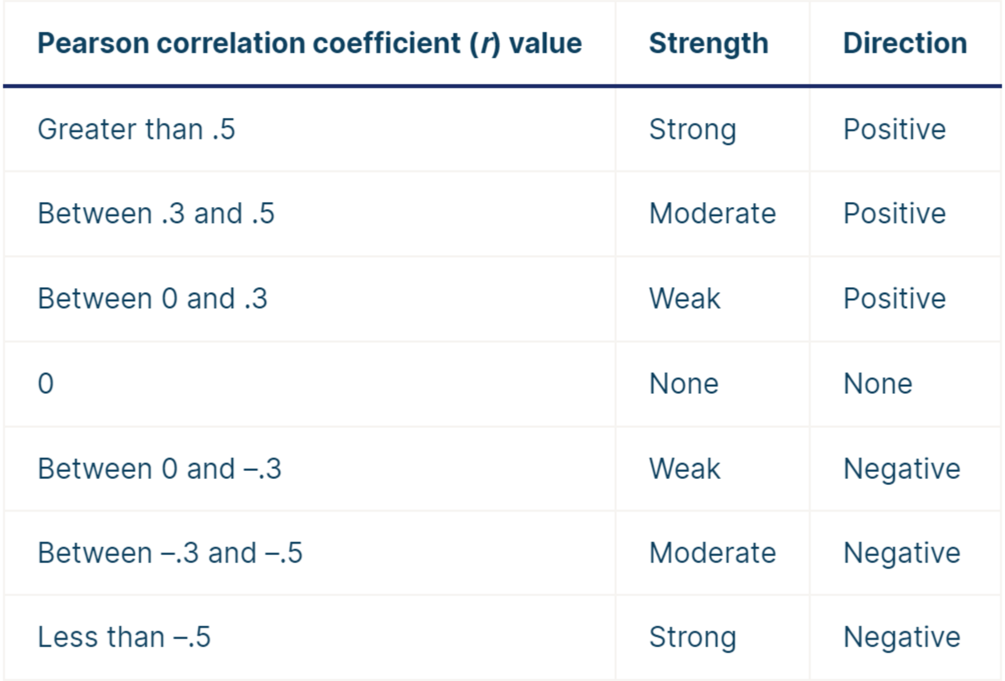
When one variable changes, the other variable changes in the same direction

Formula for Pearson coefficient

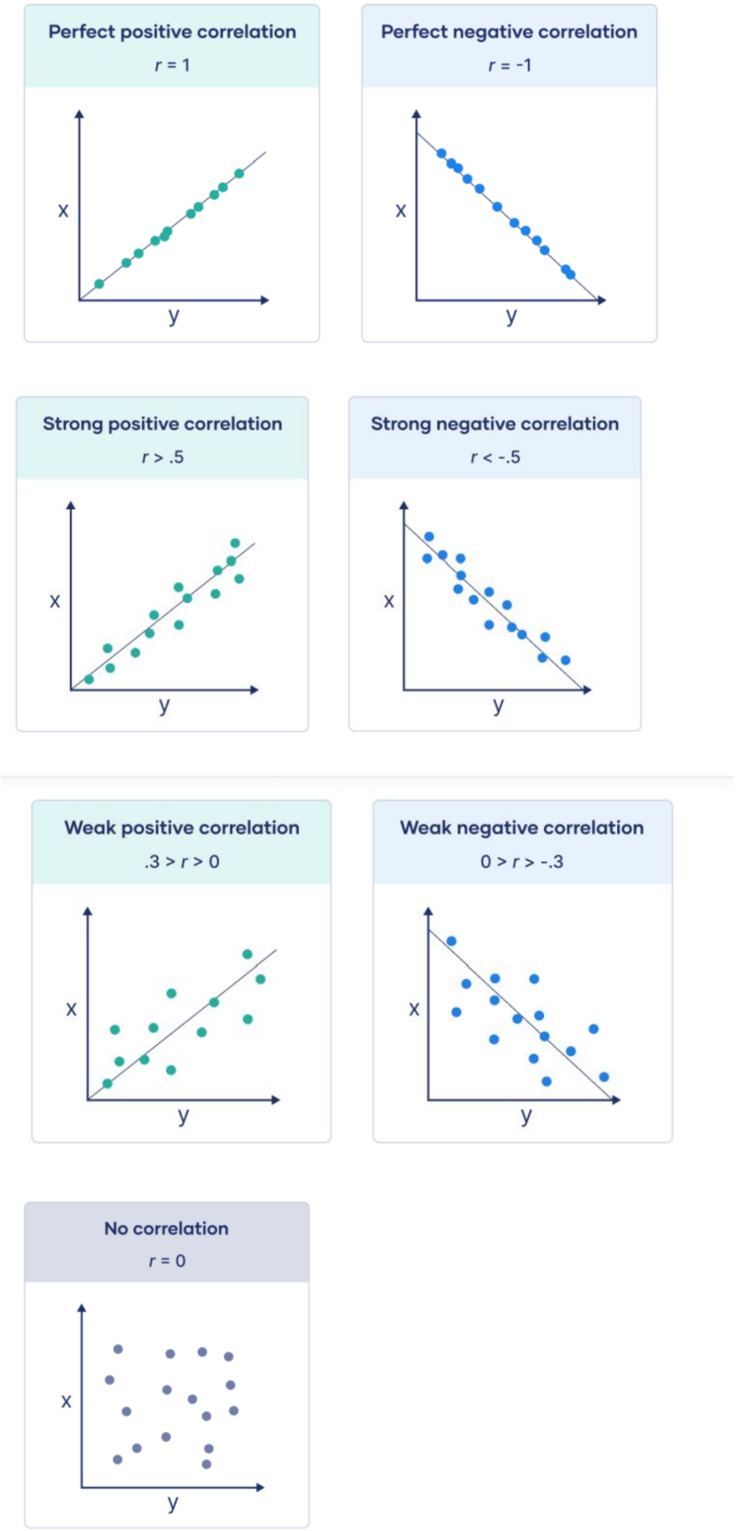
$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$



Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.



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
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative



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

The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.

Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance.

The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

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

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

$$VIF = \frac{1}{1 - R^2}$$

The VIF formula clearly signifies when the VIF will be infinite. If the R^2 is 1 then the VIF is infinite. The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

- Interpretations
  - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
  - Y values < X values: If y-values quantiles are lower than x-values quantiles.
  - X values < Y values: If x-values quantiles are lower than y-values quantiles.
  - Different distributions – If all the data points are lying away from the straight line.
- Advantages
  - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
  - The plot has a provision to mention the sample size as well.