Linear Regression Subjective Question

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?

-->

Bike demand in the fall is the highest.

Bike demand takes a dip in spring.

Bike demand in year 2019 is higher as compared to 2018.

Bike demand is high in the months from May to October.

Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.

The demand of bike is almost similar throughout the weekdays.

Bike demand doesn’t change whether day is working day or not.

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

-->

It is important to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don’t need unfurnished as we know 0-0 will indicate un-furnished. So, we can remove it

It is also used to reduce the collinearity between dummy-variables.

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

In fact in pandas.get_dummies there is a parameter i.e. drop_first allows you whether to keep or remove the reference (whether to keep k or k-1 dummies out of k categorical levels).

The drop_first = False meaning that the reference is not dropped, and k dummies created out of k categorical levels! You set drop_first = True, then it will drop the reference column after encoding.

Beacause, Keeping k dummies for k levels of a categorical variable is good idea, but there is a redundancy of one level, which is here in separate column. This is not needed since one of the combination will be uniquely representing this redundant column. Hence, its better to drop one of the column and just have k-1 dummies(columns) to represent k levels.

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

--> atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.

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

--> Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.

Also, Normality of the error distribution (Normal distribution of error terms). And constant variance of the errors or Homoscedasticity with less multi-collinearity between features (Low VIF)

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

--> The top 3 features contributing significantly towards the demand of the shared bikes are the temperature,

the year and the holiday variables.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

--> In Machine learning modelling, in which the nature of the output is a continuous variable and task is to predict the continuous variable is classified into Regression model e.g. score of students.

And in linear regression model method is used to find-out cause and effect relation between variables.

Also, the regression come under the supervised learning algorithms, in which you have the previous year's datasets with labels and use for building the models.

A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship

In linear regression algorithm, explains the relationship between a dependent variable and one or many independent variables using a straight-line. The straight line is plotted on the scatter plot of these two points. The objective of linear regression is to find a linear equation that can best determine the value of dependent variable Y for one or many values of independent variables X.

There 3 type of linear regression algorithms:

- Simple linear regression: Simple linear regression is defined by the linear function:

$$Y= \beta0*X + \beta1 + \varepsilon$$

β0 and β1 - are two unknown constants representing the regression slope

ε (epsilon) - error term

e.g. Rainfall and crop yield

- Multiple linear regression: In multiple linear regression analysis, the dataset contains one dependent variable and multiple independent variables. The linear regression line function changes to include more factors as follows:

Y= β0*X0 + β1X1 + β2X2+…… βnXn+ ε

As the number of predictor variables increases, the β constants also increase correspondingly.

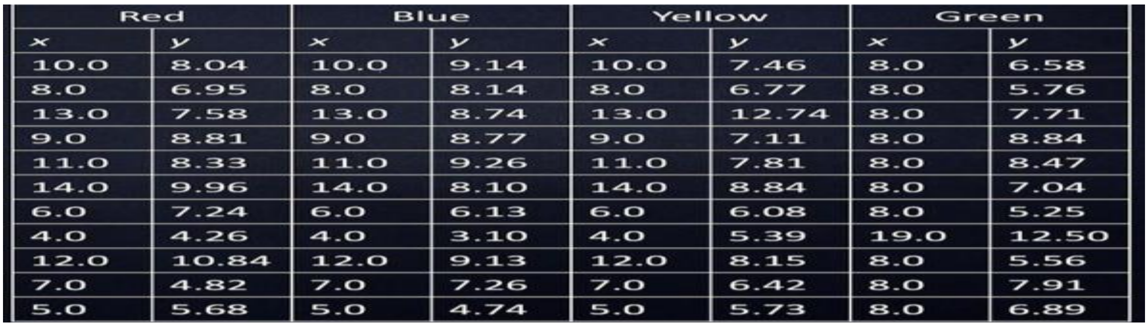
e.g. Rainfall, temperature, and fertilizer use on crop yield.

● Logistic Regression: Data scientists use logistic regression to measure the probability of an event occurring. The prediction is a value between 0 and 1, where 0 indicates an event that is unlikely to happen, and 1 indicates a maximum likelihood that it will happen. Logistic equations use logarithmic functions to compute the regression line. e.g. The probability of passing or failing a test

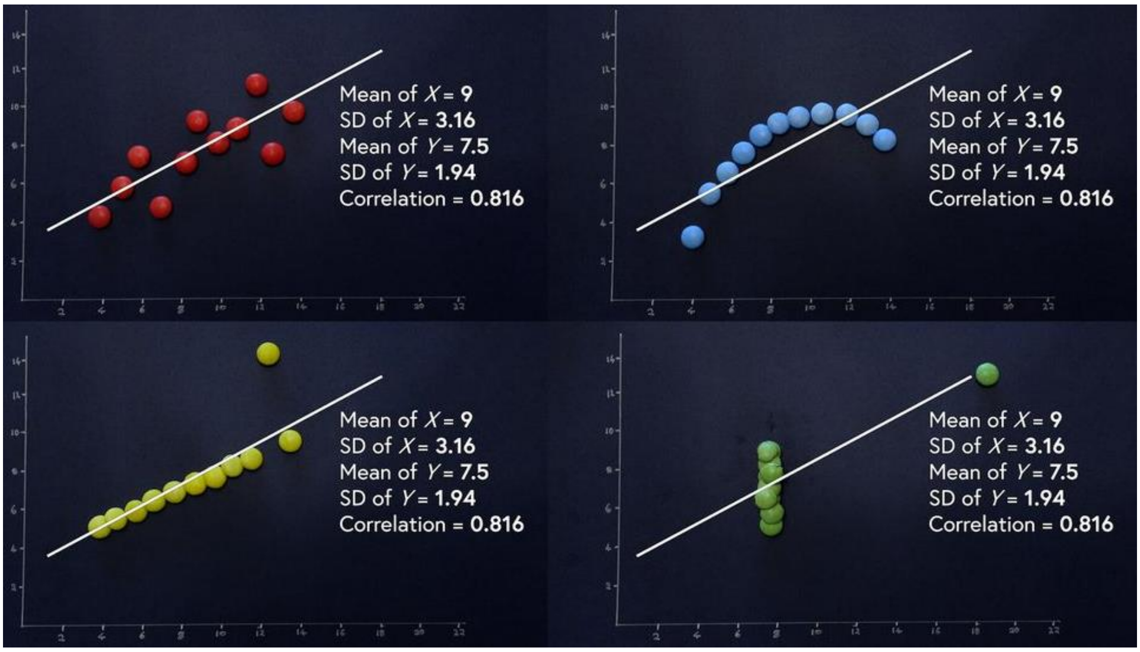
2. Explain the Anscombe’s quartet in detail.

--> Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading. Anscombe's Quartet shows how four entirely different data sets can be reduced down to the same summary metrics. Here are the data sets from Anscombe’s Quartet – both as raw data, and plotted on a chart:



Red		Blue		Yellow		Green	
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



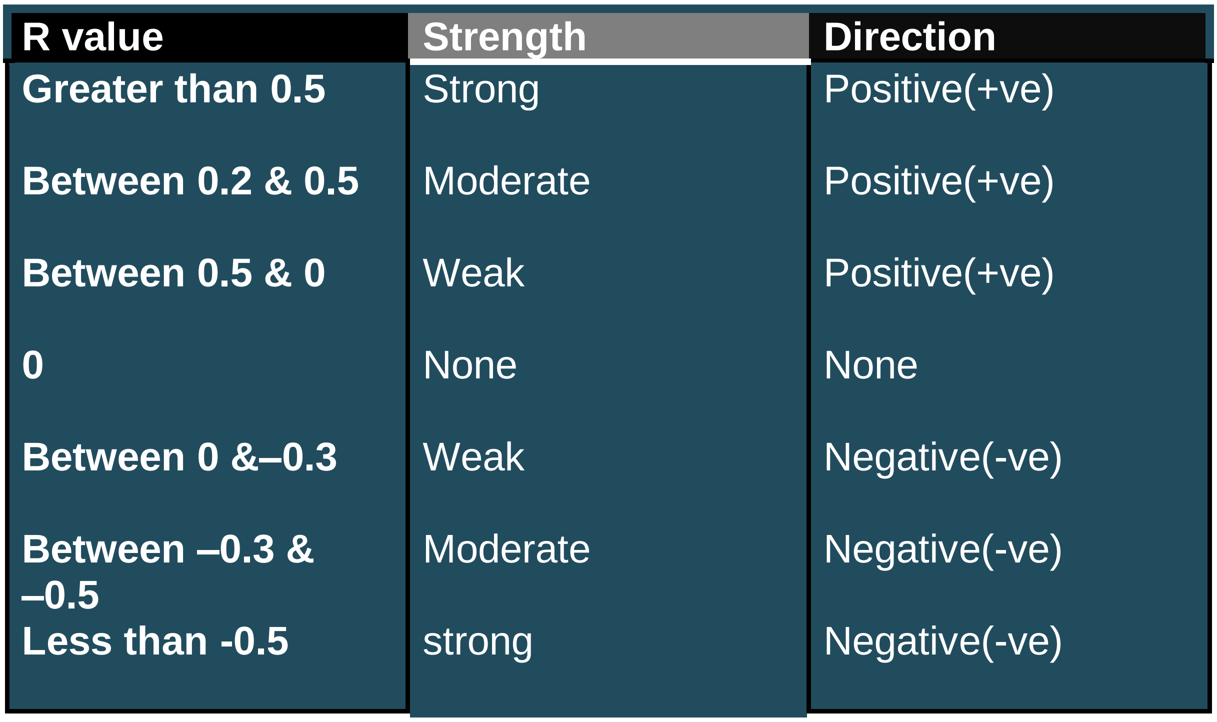
Anscombe's Quartet shows us that we should not blindly trust summary statistics or standard methods of analysis. It tells us to look closely at our data, question our assumptions, and use a variety of analytical tools to get a full picture. This concept emphasizes the importance of visualizing data, as graphs can reveal patterns and outliers that summary statistics alone may overlook.

3. What is Pearson’s R?

--> Correlation measures the strength of association between two variables as well as the direction. The correlation is measured by One significant type i.e. Pearson's correlation coefficient. This type of correlation is used to measure the relationship between two continuous variables.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

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.



R value	Strength	Direction
Greater than 0.5	Strong	Positive(+ve)
Between 0.2 & 0.5	Moderate	Positive(+ve)
Between 0.5 & 0	Weak	Positive(+ve)
0	None	None
Between 0 &–0.3	Weak	Negative(-ve)
Between –0.3 & –0.5	Moderate	Negative(-ve)
Less than -0.5	strong	Negative(-ve)

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

--> Scaling: It is a geometric change that linearly enlarges or reduces things. A property of objects or rules known as scale invariance is that they remain unchanged when scales of length, energy, or other variables are multiplied by a common factor.

Scaling law, a law that explains how many natural phenomena exhibit scale invariance.

scaling performed because:

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.

Normalized scaling and Standardized scaling:

(i) The values of a normalized dataset will always fall between 0 and 1.

(ii) A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

(iii) In standardized, the variables are scaled in such a way that their mean is zero and standard deviation is one. And equation is:

$$X = x - \text{mean}(x) / \text{SD}(x)$$

(iv)In Normalized, the variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. And the equation is:

$$X = x - \min(x) / \max(x) - \min(x)$$

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

--> Let First look at Equation for the VIF:

$$\mathbf{VIF_i} = \frac{1}{1 - R_i^2}$$

The common heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

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) infinity. To solve this problem, 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.

--> In statistics, a Q–Q plot (quantile–quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.

A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.

QQ plots is very useful to determine:

- If two populations are of the same distribution.
- 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.
- Skewness of distribution.

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

As you build your machine learning model, ensure you check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, you might want to check the distribution of your feature variable and consider transforming them into a normal shape