SUBJECTIVE QUESTIONS

Answers for all subjective questions

Bharath Veer K S
bharathsanjai@gmail.com

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?

- Fall has high range of bike sharing with 75 percentile around 6k and the trend is consistently reaching high count of 6k.
- Year 2019 has high demand across all months than year 2018 which proves post pandemic people are adopted to exploring.
- Clear weather has high trend with median of 5k whereas Light_snow doesn’t encourage people to book bikes.
- Month of April – October has high demand of bikes than the months from December to March. This could be due to Fall/Summer holidays.
- Working day doesn’t seem to have much difference in bike booking trend
- In the month of July holiday have good reasonably high trend in bike bookings

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

- Drop first parameter explicitly specifies if you want to drop the first column when you do the encoding
- it is a technique which converts categorical data into a form which is understandable by ml model

Documentation:
drop_firstbool, default False

Whether to get k-1 dummies out of k categorical levels by removing the first level.

Example:

Let’s us take a small example of categorical column ‘Furnished_status’ which has three values

a. Furnished
b. Semi Furnished
c. Unfurnished

- By default there will be three dummy columns one column for each unique value of our original column and wherever this value is true for a row it is indicated as 1 else 0.

- In our case, When we specify Drop first=True then it will be 2 columns not 3 and it is useful because it reduces number of columns

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
    - Temp variable has highest correlation which is same as atemp but atemp seem to be derived variable hence it is temp

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
    a. Linearity
        i. There must be a linear relationship between predictor and the target variable
    b. Multicollinearity
        i. Predictor variables are not highly correlated to each other which is derived by VIF value
        ii. VIF must be < 5
    c. Homoscedasticity
        i. The residuals have constant variance at every point in the linear model.
    d. Independence

        i. Observations are independent to each other

   e. Error terms distribution

        i. The residuals of the model are normally distributed.

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

   Below are the top variables contributed to the model

   a. Temp
   b. Year
   c. Winter

General Subjective Questions

1. Explain the linear regression algorithm in detail?

   Linear Regression is a machine learning algorithm based on **supervised learning.** It tries to apply relations that will predict the outcome of an event based on the independent variable data points. The relation is usually a straight line that best fits the different data points as close as possible.

   The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analysed or studied.

   ### Equation of linear regression

   $$y= \beta 0+ \beta 1x+ \varepsilon$$

   - Y= Dependent Variable

- X= Independent Variable
- β 0= intercept of the line
- β1 = Linear regression coefficient (slope of the line)
- ε = random error

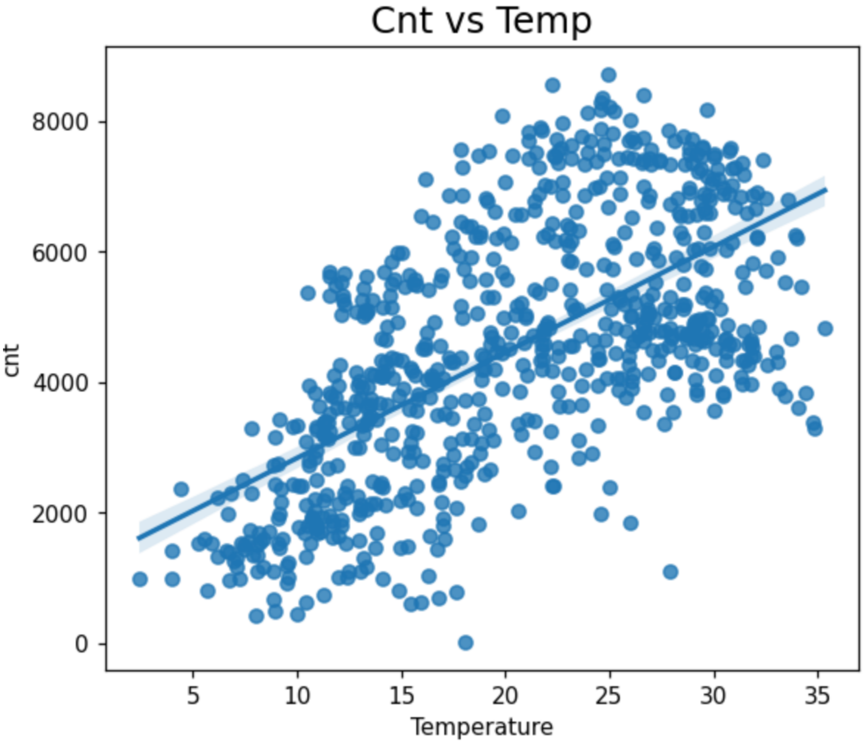
Types of Linear Regression

Linear Regression can be broadly classified into two types of algorithms:

Simple Linear Regression

Simple linear regression reveals the correlation between a dependent variable (input) and an independent variable (output). Primarily, this regression type describes the relationship between the given variables

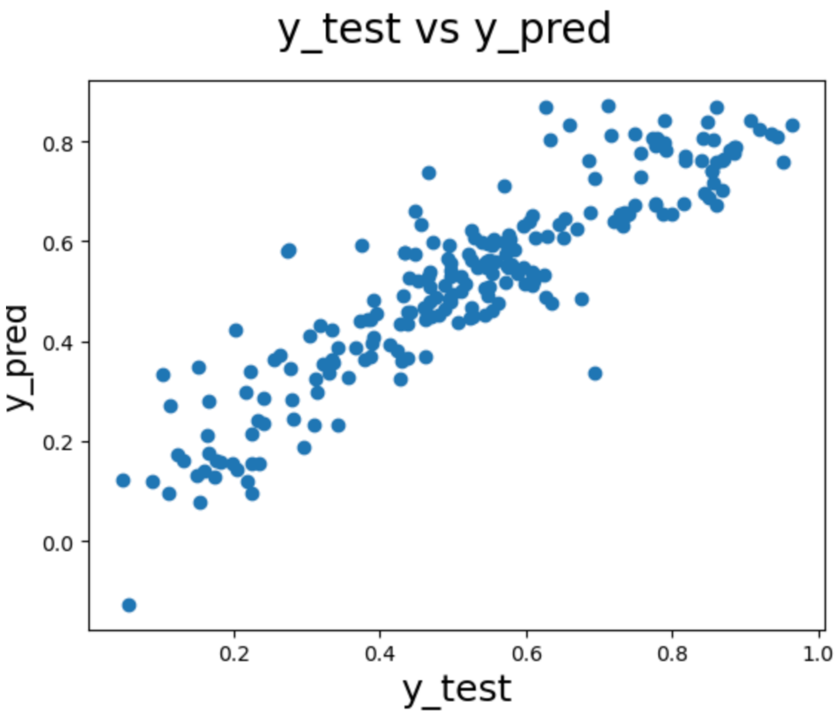
Example: From our bike sharing assignment temp and count have strong linear relationship.



Multiple Linear Regression

Multiple linear regression establishes the relationship between independent variables (two or more) and the corresponding dependent variable. Here, the independent variables can be either continuous or categorical.

Example: From our bike sharing assignment the multiple linear model based scatter plot will depict the linear relationship of multiple variables to the target



Cost Function

The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as **the Hypothesis function**.

In Linear Regression, Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Gradient descent

Another important concept in Linear Regression is Gradient Descent. It is a popular optimization approach employed in training machine learning models by reducing errors between actual and predicted outcomes.

2. Explain the Anscombe's quartet in detail?

Definition

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

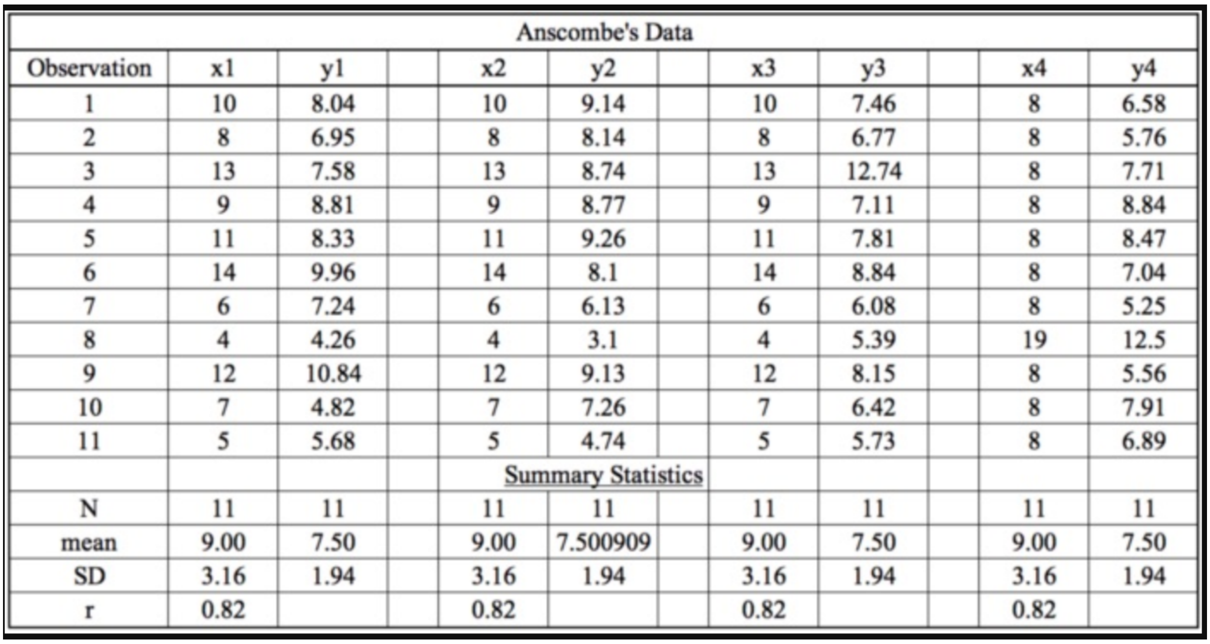
Explanation

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

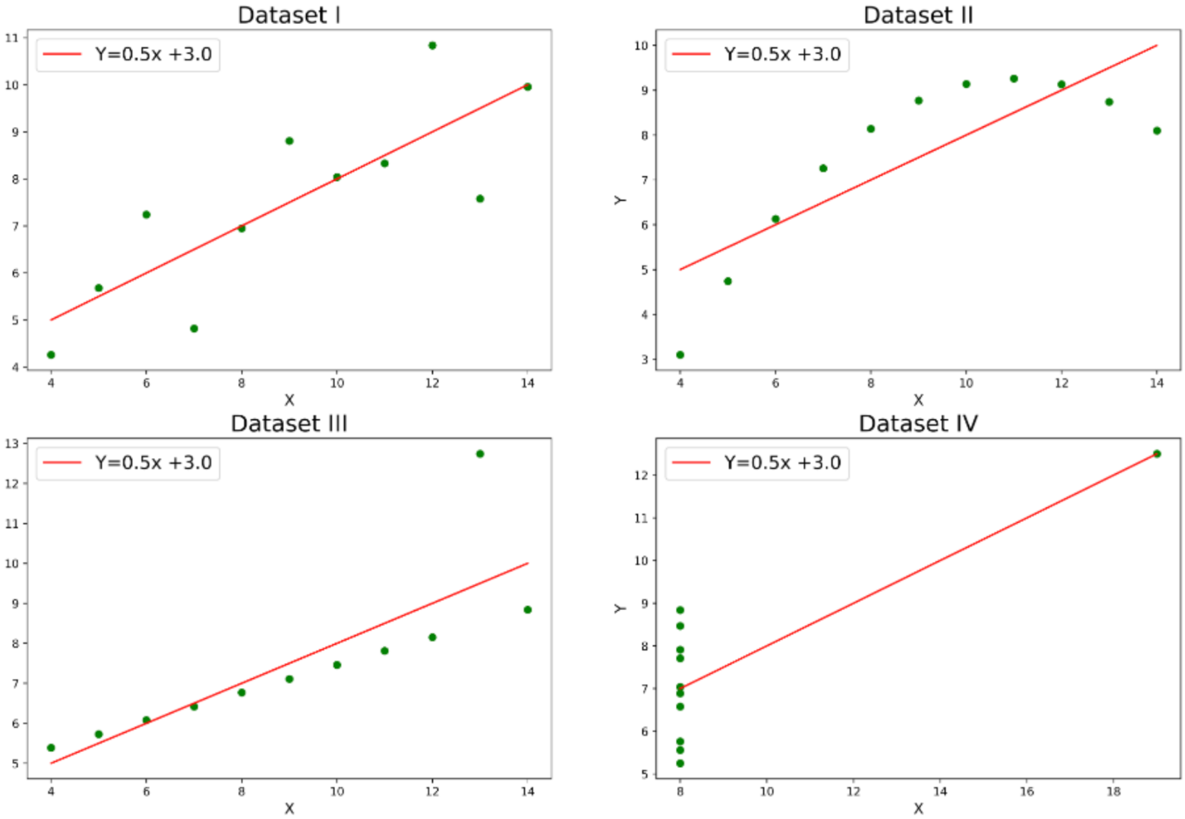
Example

Let's take the below four plots



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	

Plotting the above data



Analysis:

- **First data set:** Depicts a good linear regression model.
- **Second data set:** Depicts a clear non-linear model
- **Third Dataset :** This dataset has good linear regression but outliers are present which doesn't fit into a linear regression model
- **Fourth Dataset:** Clearly tell us that one leverage point with high value is good enough for deriving high correlation coefficient

Conclusion

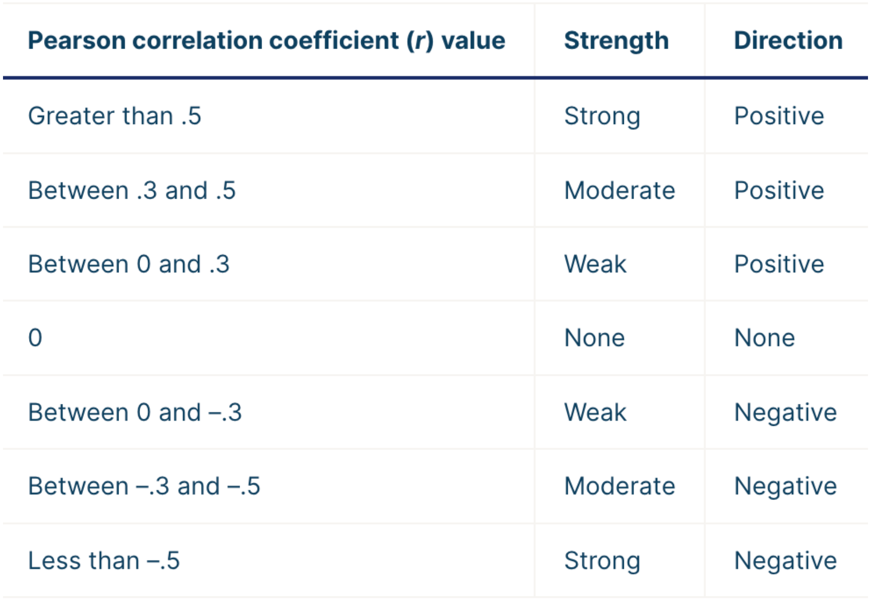
All the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

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.

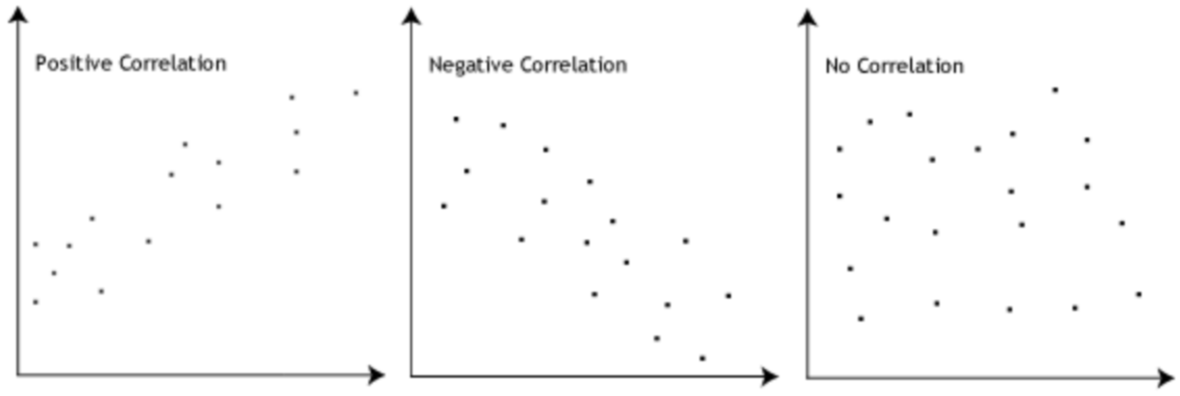
The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line.

Below table depicts the strength and direction



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

Visualising Pearson correlation coefficient



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

Explanation

Feature scaling is a data preprocessing technique that involves transforming the values of features or variables in a dataset to a similar scale. Feature scaling is essential when working with datasets where the features have different ranges, units of measurement, or orders of magnitude.

Standardization or Z-Score Normalization

It is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

X_new = (X - mean)/Std

Normalization or Min-Max Scaling

This is used to transform features to be on a similar scale. This scales the range to [0, 1] or sometimes [-1, 1].

X_new = (X - X_min)/(X_max - X_min)

Differences:

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

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

A large value of VIF indicates that there is a correlation between the variable but if VIF = infinity then there is perfect correlation of one independent variable over other.

$$VIF_1 = \frac{1}{1 - R^2}$$

This happens when R^2 approaches 1 in above equation.

Recommendation

One approach is to review your independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model.

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

Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against.

Why QQ Plot?

It help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or

Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Key benefits of QQ Plot?

- If two populations are of the same distribution
- have common location and scale
- have similar distributional shapes
- Skewness of distribution

Summary

1. Q-Q plot helps you compare the sample distribution of the variable at hand against any other possible distributions graphically.
2. Q-Q plot can also be used to test distribution amongst 2 different datasets.