

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 bookings show a trend from months 5 to 9, with over 4000 bookings per month. Clear weather and working days have close to 5000 bookings, indicating they are good predictors. Season 3 has the most bookings, followed by summer and winter, suggesting season can also be a good predictor.

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

Creating dummy variables without `drop_first=True` generates dummy variables for all categories in the column. For example, the Weekday column has seven values (0 to 6). With `drop_first=True`, only six columns are created as `WeekDay_Sunday` is unnecessary. The absence of this column allows Sunday to be identified by observing zero values in the remaining columns.

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

Temperature

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

- Assumptions of linear regression were validated post-model building on the training set through residual analysis.
- Linearity was verified via pair plots.
- Multicollinearity was assessed via VIF values (VIFs above 5 indicate multicollinearity)
- Normal distribution of error terms was checked using QQ or distribution plots.
- Homoscedasticity was evaluated by inspecting the residual plot for constant variance across dependent variable values.

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

Feature	Explanation
Temperature	Temperature has a coefficient of 0.5499, meaning a unit increase in temperature results in a 0.5499 unit increase in bike hires.
Season	The Winter season has a coefficient of 0.1318, indicating that a unit increase in winter season results in a 0.1318 unit increase in bike hires.
year	The year variable (yr) has a coefficient of 0.2331, showing that a unit increase in year results in a 0.2331 unit increase in bike hires.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is finding the best linear relationship between the independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Linear regression is one of the very basic forms of machine learning algorithms where we can train a model to predict the behavior of the data based on some variables.

In Linear Regression, whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X. Mathematically, we can write a linear regression equation be like below:

$$Y = mx + c$$

m is slope of the line,

c is the y intercept of the line

Assumption of Linear Regression is:

- There is a linear Relationship between dependant and independent variables.
- Error values (ϵ) are normally distributed for any given value of X that means
- Error terms are normally distributed around zero.
- Constant variance assumption: It is assumed that the residual terms have the variance, σ^2 , this assumption is also known as the assumption of homogeneity or homoscedasticity.
- Independent error assumption: residual terms are independent of each

other, i.e. their pair-wise covariance is zero.

- The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

Use Cases of Linear Regression:

- **Prediction of trends and Sales targets:** To predict how industry is performing or how many sales targets industry may achieve in the future.
- **Price Prediction:** LR is used in price prediction – we can predict the change in price of a product.
- **Risk Management:** LR is used in Risk Management in the financial and insurance sector.

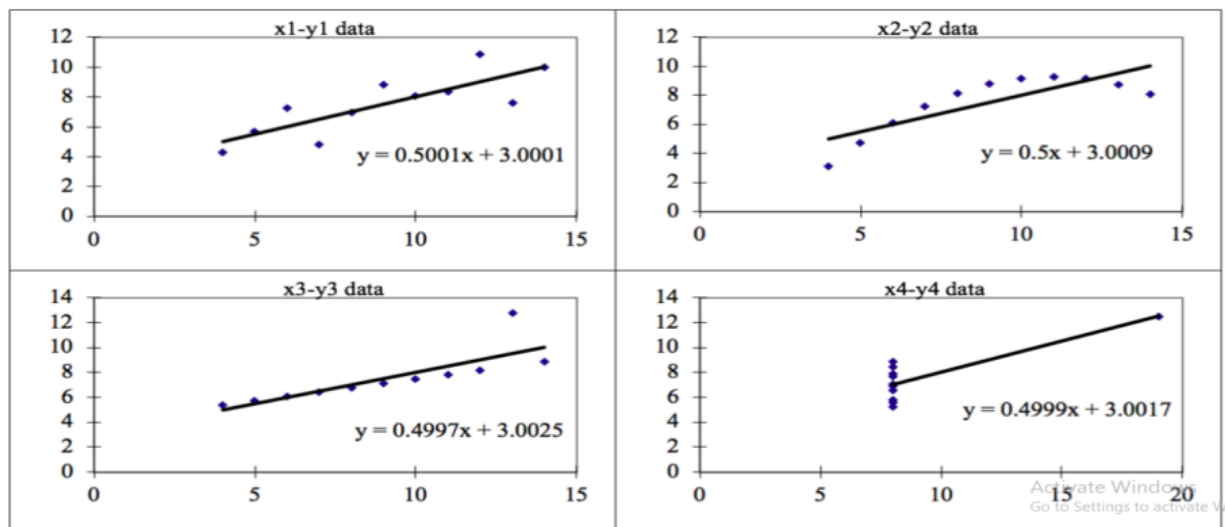
2. Explain the Anscombe's quartet in detail?

Anscombe's Quartet tells us the importance of data visualization before applying any algorithm to the data set. This can be defined by four data sets having 11 data points each, which are nearly identical. Even though these data points look identical, they appear differently when plot scatter plots.

The Four Datasets looks like:

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

The plots are:



When we observe the four data set scatter plots, Data set 1 – fits Data set 1 the Linear Regression model pretty well, Data Set 2 – could not fits the LR model as the data is non Linear. Data set 3 and 4 have outliers involved in the data set which cannot be handled by Linear Regression Model These datasets are created intentionally to describe the importance of data visualisation and how any regression algorithm can be fooled by the same. Hence, 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?

Pearson correlation coefficient will calculates the strength of a linear association between two variables that means change in the one variable where the other variable changes and this is denoted by r .

It seeks to draw a line through the data of two variables to show their relationship. The relationship can be measured with the calculator of Pearson correlation coefficient. This linear relationship can either be positive or negative.

For example:

Positive linear relationship: The income of a person increases as his/her experience increases.

Negative linear relationship: If the vehicle speed increases, then time taken to travel decreases.

Pearson correlation coefficient formula:

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

Where:

- N = the number of pairs of scores
- $\sum xy$ = the sum of the products of paired scores
- $\sum x$ = the sum of x scores
- $\sum y$ = the sum of y scores
- $\sum x^2$ = the sum of squared x scores
- $\sum y^2$ = the sum of squared y scores

It is the covariance of two variables, divided by the product of their standard deviations and the result always has a value between -1 and 1 . This measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

Correlation coefficient formulas are used to find how strong a relationship is between data.

The formulas return a value between -1 and 1 , where:

1 indicates a strong positive relationship. -1 indicates a strong negative relationship. A result of zero indicates no relationship at all.

Pearson's correlation coefficient formula, when applied to a population is:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

cov is the covariance, other terms are standard deviations of X and Y.

The formula for p can be expressed in terms of mean and expectation.

$$\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

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

Scaling: Scaling is the pre-processing applied to independent variables to normalize the data within a particular range and also helps in speeding up the calculations.

For Ex: one feature is in KG's and the other is in grams, the other one is in liters etc. Here scaling is performed to normalize all the units.

Often, the data set contains the most varying units, magnitudes and range. If we have such data and if we have not done scaling then it will only take the magnitude in account and not units hence the incorrect model will build. To resolve this we will use scaling to bring all the variables to the same level of magnitude and units.

Scaling only affects the coefficients but not the other parameters.

Normalization/Min-Max Scaling: It is a technique where the values are shifted or rescaled and is ranging between 0 and 1 known as min – max scaling. That means it brings all the data within the range of 0 and 1.

sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling: It is another technique in which the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

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).

sklearn.preprocessing.scale helps to implement standardization in python.

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

Disadvantage of Min-max scaling over standardization: Normalization loses the information of outliers.

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

- VIF = infinity indicates that there is perfect correlation between the two independent variables.
- If such cases are arrived, we get $R^2 = 1$, which leads to $1/(1-R^2) = 1/0 = \text{infinity}$.
- To solve this problem we need to drop one of the variable which causes the perfect correlation i.e. multicollinearity
- This infinite value of VIF indicates that the corresponding variable can be expressed by Linear Combination of the other variables.

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

Q Q Plots (Quantile-Quantile plots) are the plots of two quantiles against each other.

A quantile is a fraction where certain values fall below that quantile

For ex: The median is a quantile where 50% of the data fall below that point and 50% lie above it.

The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

This helps a scenario of linear regression we can analyze that error terms are normally distributed

IQQ plots are used to check the following scenarios:

- If two data sets, come from populations with a common distribution
- If two data sets have common location and scale
- If two data sets have similar distributional shapes
- If two data sets have similar tail behavior