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? (3 marks)**

Ans

- yr and Light_rain are in the top 3 contributors to the dependent variable count
- People are using bikes more in summer and fall ,
- In Aug, Sep, and Oct, the demand is high for the bikes compared to other months.
- In case of high humidity, wind speed and rain, the demand is low as expected
- Bike riders increased in 2019 compared to 2018.
- Month_July,Mist_or_cloud, windspeed, hum, Light_rain has negative coefficient which shows if weather is bad; less users use bikes.

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

Ans

drop_first=True helps in reducing the extra column created during dummy variable creation. As a result, it reduces the correlations created among dummy variables.

Example - Say we have 3 types of values in Categorical column Weather i.e. ‘Summer’, ‘Winter’ and ‘Rainy’’ and we want to create dummy variable for that column. If one variable is not ‘Winter’ or ‘Rainy’, it means it’s obviously ‘Summer’. So we do not need 3rd variable to identify the ‘Summer’

Eg.

Winter	Rainy

0	1
1	0
0	0

.

00 means it's summer.

Hence if we have categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Ans

-Temp and atemp has the highest positive correlation with the target variable 'cnt'. . This is evident from the scatterplot where we can see the linear relation . The same can be confirmed in the heatmap too which shows these 2 variables have a high positive correlation of 0.62.

Also, temp and atemp are highly correlated to each other too as 'atemp' is feeling temperature in Celsius which will be related to actual temperature in celsius.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Ans

-Linear Regression assumptions were verified in following way:--

a. **Linearity** - Plotted y_test vs y_pred and observed the relationship. Relationship in our case is linear
b. **Normality** - Plotted Error distribution and Q-Q plot(between fitted and predicted values) to observe the normality. A light fatter tail was observed in the left side in the Error distribution and a similar observation was made in the Q-Q plot. Higher omnibus value and lower probability(omnibus) showed it's not perfectly normal but it was close to Normal.

c. **Multi-collinearity** -Higher VIF values(> 5 needs attention >10 is multicollinearity) shows multicollinearity . In my model, I was able to get all the VIF values <5 making sure there is no correlation between the predictors.
d. **Independence of Residuals** - Durbin-Watson value ~2 (2.081 precisely) shows no autocorrelation in the sample.
e. **Homoscedasticity** - This was checked using breuschpagan and white-Test. This is a quantitative method which give p-value. If p-value < 0.05 it means Hetroscedasticity is present and it's not completely homoscedastic. In our model too, we got p value < 0.05. To confirm, I further plotted residual vs fitted and found some hetroscedasticity is present as the variance increased a bit towards right.

5. **Based on the final model, which are the top 3 features contributing significantly toward explaining the demand of the shared bikes?(2 marks)**

Ans

Modal parameters

temp	0.57
yr	0.23
const	0.15
winter	0.14
Month_Sep	0.09
summer	0.08
Day_Mon	0.08
workingday	0.05
Month_July	-0.04
Mist_or_cloud	-0.05
windspeed	-0.10
hum	-0.16
Light_rain	-0.21

Top 3 contributors :-

1.**temp -** coeff**: 0.57** - impacts in a positive way. It it increases, count/demand increases for the bike

2. **yr** - coeff : **0.23,** impacts in a positive way. when yr is '1' i.e. year 2019 contributes more to the demand of bikes.

3. **Light_rain** coeff: **-0.21**, negatively impacts the demand. When there is Light rain, demand decreases as expected.

General Subjective Questions

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

Ans

Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis. Dependent variable is the target variable which we aim to predict and the independent variables are the variables which are used to predict the target variable.

Equation -

Simple linear regression

Y = β*x +b where β is the slope and b is the bias(y-axis intercept)

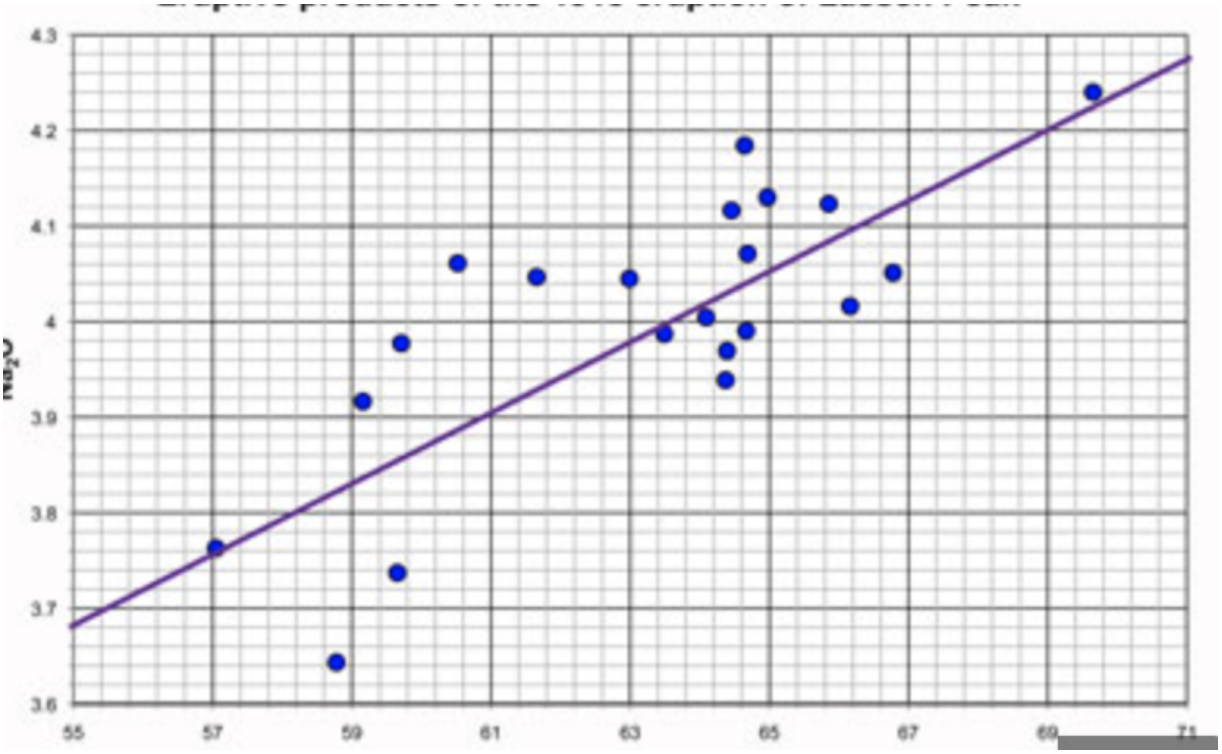
There will be multiple $\beta's$ in case of multiple linear regression

Y = $\beta 0x1 + \beta 1x2 + \beta 2x3 + \beta 3x4 + b\ where\ x1, x2, x3, x4\ are\ different\ independent\ variables$)

Idea is to find the best fitted line which can explain the relationship between the independent variables and the target. Methods like RSS are used to find which is the best fit where all points are plotted and perpendicular distance of the points form line is calculated. These numbers and squared and sum and whichever line has the lowest is the best fit line.

Linear regression can be applied only to predict continuous target variables. There are different linear regression assumptions which has been explained before in this exercise.

Below diagram shows a fitted linear regression line:--

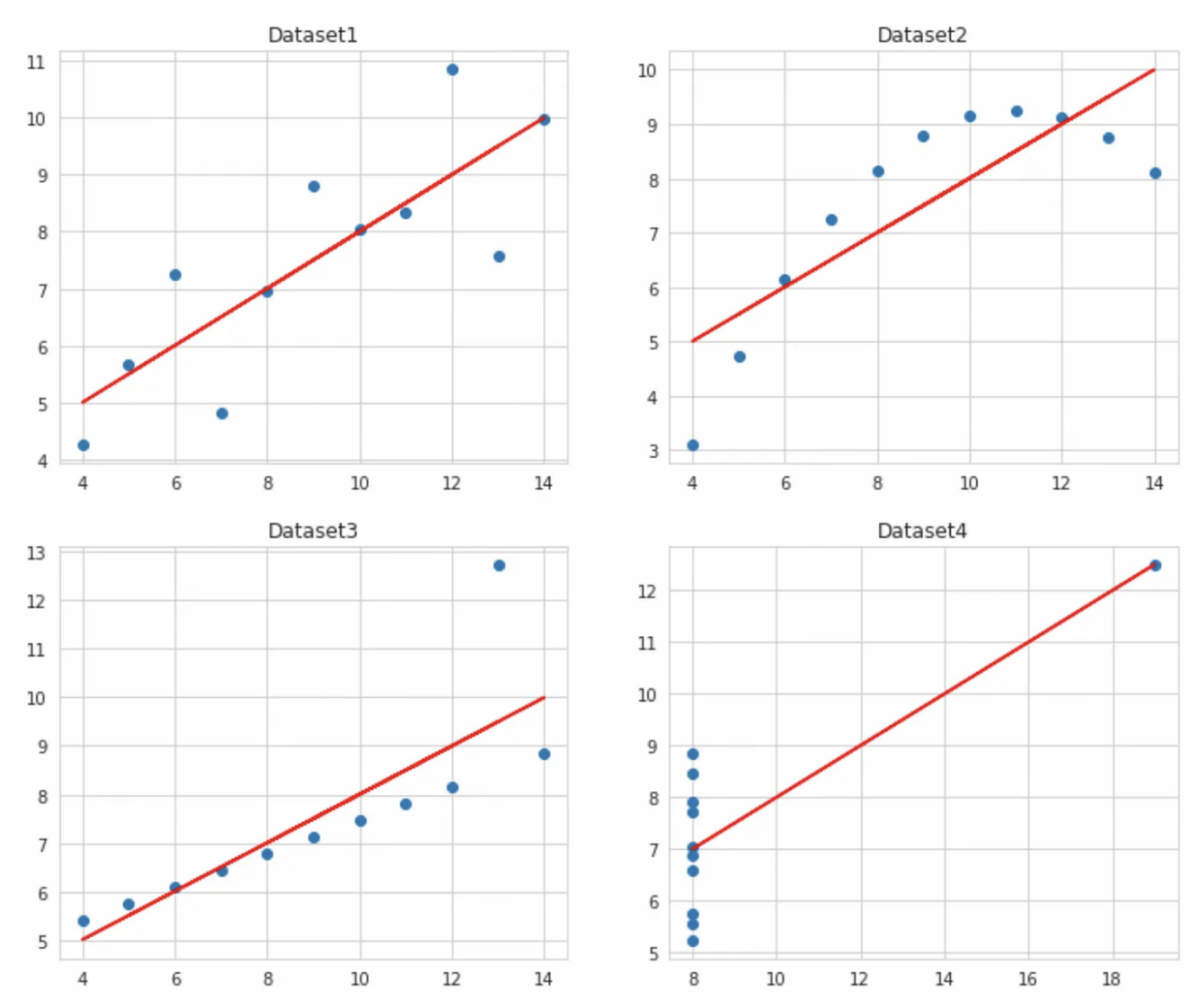


2. Explain the Anscombe's quartet in detail. (3 marks)

Ans

- It brings in the interesting concept why Data visualization is necessary before analyzing the data. Anscombe's quartet comprises four different datasets that have very identical simple statistical properties. When this dataset was graphed it looked very different when graphed. Each dataset consists of eleven (x,y) points.
- They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

- When graphs were plotted it looked like below :-



- Observations were:--
-
- 1. Dataset1 shows linear relation(positive correlation between x & y)
- 2. Linear relationship not present in Dataset2
- 3. Dataset3 has linear relationship with outlier
- 4. Dataset4 shows high correlation due to high leverage(how far away one of the independent value is from other observations)

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

It's a measure of linear correlation between 2 sets of data. It varies between -1 to 1. It depicts the linear relationship between variables. r=1 means data is completely linear with positive slope, r=-1 means data is completely linear with negative slope . r= 0 means no linear association.

One example is age and height data of individuals, as age increases height will increase to a certain extent. It shows positive correlation. So, Pearson's R will be > 0 and less than 1 in this case.

Mathematically: -

$\rho x, y = COV(x, y)/\sigma x \sigma y \; where \; COV \; is \; covariance \; and \; \sigma x \; and \; \sigma y \; are \; the \; std \; deviation \; for \; x \; 7 \; y$

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

Ans

It is a mechanism of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It is very helpful in making the calculations fast for an algorithm. Scaling is required in most of the linear regression problems in the industry.

It's needed because most of the time the data collected has a different value range for different independent variables. The problem arises when we try modeling without scaling, the algorithm only takes magnitude of the values into account and not units and since the magnitude ranges are different, results can't be deterministic. Scaling helps in bringing all variables to the same level of magnitude,

Scaling only affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Difference between Normalized and Standardized scaling:--

Normalized scaling uses min and max value of features for scaling. It's applicable when features are on different scales. It scales values between -1 to 1. It is affected by the outliers. MinMaxScaler in Scikit learn can be used for Normalized scaling.

Standardized scaling uses mean and standard deviation for scaling. It is used when we want to ensure 0 mean and unit std deviation. There is no range for this type of scaling. Effect of outlier is minimal. StandardScaler can be used in Scikit learn to achieve this.

5.You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans

VIF also known as variance inflation factor is mathematically 1/(1-R-square). So if R-square =1, VIF will become infinite. Now the question is when will R-square become 1? This will happen in case of perfect correlation.

To fix this, one needs to drop a variable which is causing the perfect multicollinearity. Infinite VIF shows the variable can be expressed by combination of other variables and it's not needed..

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

Ans

Q-Q plots are also called Quantile-Quantile plots which is the plot of two quantiles against each other. Quantile is nothing but a fraction where certain values fall below the quantile level. Median is considered as a quantile where 50% of the data are below that point which means 50% are above it.

The idea of Q-Q plot is to find if the two sets of data came from the same distribution. Few intuitions:--

a. If two distributions are similar the Q-Q plot will lie on the y=x line.
b. If two distributions have linear relation, Q-Q plot will lie on a line but not necessarily the y=x line.
c. if both tails twist counterclockwise there are heavy tails.
d. if both tails twist clockwise, there are light tails
e. if the right tail twists counterclockwise and the left tail twists clockwise, you it's right skewed
f. if the left tail twists counterclockwise and the right tail twists clockwise, it's left skewed

Q-Q plot also answers if 2 datasets have common location and scale, have same distributional shapes, and have similar tales. Python's scipy library can be used to make Q-Q plot