Assignment-based Subjective Questions

Q- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

There is some correlation between the weathersit like bike demand increase in FALL season

Bike demands show some pattern while anaysis monthly demands. The demands increasing from Jan to June and then some decline and then highest in *September* and then demand continues to decrease

There are random demands in weekdays

There is highest demand for bike when weathersit is 1 means clear sky or Good weather and demands tend to decrease as weather tends to detoriate and not demands when weathersit is 4 or very bad weather.

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

Answer

The drop_first is just that we can get rid of one extra column and can still be able identify p levels with (p-1) level

If you have category having 3 levels they dumm variale of length 2 can define them.

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

Answer:

cnt vs temp or atemp

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

Answer:

By looking the coefficients of model features we can reject the NULL Hypothesis.

Also the acquired p-values are Zeros 0 and VIF are less than 5

Higher F-statistics more significant the model is.

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

Answer:

Year

Workingday

Temperature

General Subjective Questions

Q- Explain the linear regression algorithm in detail. (4 marks)

Answer:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

The function for linear regression is
y = mX + c

Where

X – Input data provided

Y – predicted data

m – coeffiecents of independent variables

c – intercept

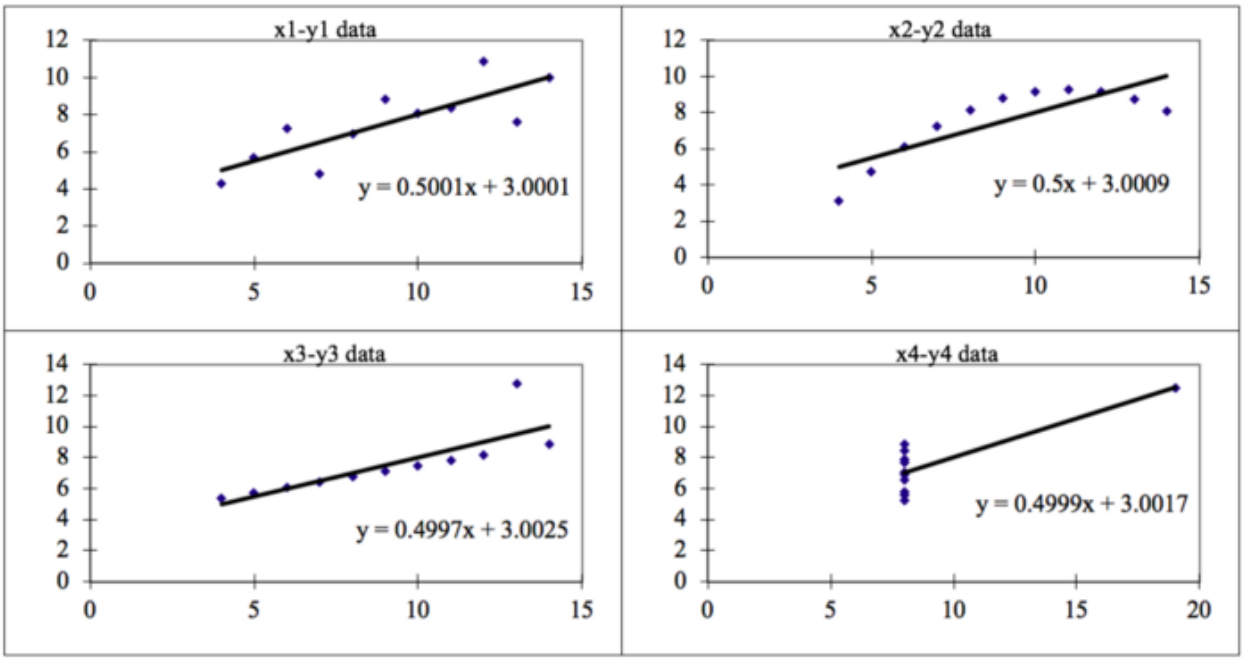
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the m and c values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

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

Answer:

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

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	



The four datasets can be described as:

Dataset 1: this **fits** the linear regression model pretty well.

Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

Q- What is Pearson's R? (3 marks)

Answer:

The pearson Correlation Cofficient(r) is the most common way of measuring a linear correlation. It is a number between -1 to 1 that measures the strength and direction of the relationship between two variables.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

Q- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Feature Scaling is a technique to standardize the independent features in the data in a fixed range.

Feature scaling is done because the coefficients of independent variables may vary in range and algorithm may bias towards variable having high values. Also feature scaling helps in quickly achieving the minima of the cost function.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale. The new point is calculated as:

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

Standardization or Z-Score Normalization 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

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

Answer:

If VIF is infinite which means there is perfect correlation between the variables

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

Answer:

The Q-Q plot as the name suggest that it is the graphical plotting of the quantities of the two distributions with respect to each other. In simple words it is the plot quantiles against quantiles. Whenever we concentrate on Q-Q plot we should concentrate on y=x. It is sometimes also called 45-degree line in statistics.

Advantages :

It can be used with sample sizes also

Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.