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

Answer:

- The demad of bike is less in the month of spring when compared with other seasons
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non-working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Light snow and light rainfall. We do not have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we cannot derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

2.Why is it important to use Drop_first true during dummy variable creation?

Answer:

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.

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

Answer:

temp has highest positive correlation with target variable cnt.

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

Asnwer:

We can validate the assumptions of Linear Regression after building the model on the following training set by below method:

1)Fitted regression line is linear.

2)Error terms came out normally distributed with mean as 0.

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

Asnwer:

The changes of increasing the number of bikes being rented increases during the working day. The demand for bikes on rent if negatively affected by windspeed The demand for bikes on rent is high in Fall season The demand of bikes on rent is high in clear weather

General Subjective Questions

1.**Explain the linear regression algorithm in detail.**

Answer:

In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Residual Sum of Squares Method.

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It’s used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog). There are two main types:

Simple regression: - Simple linear regression uses traditional slope-intercept form, where m and b are the variables our algorithm will try to “learn” to produce the most accurate predictions. x represents our input data and y represents our prediction. y=mx+by=mx+b

Multivariable regression: - A more complex, multi-variable linear equation might look like this, where w represents the coefficients, or weights, our model will try to learn. f (x, y, z) =w1x+w2y+w3z

The variables x,y,z represent the attributes, or distinct pieces of information, we have about each observation. For sales predictions, these attributes might include a company's advertising spend on radio, TV, and newspapers. Sales=w1Radio+w2TV+w3News

Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties,
yet appear 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.

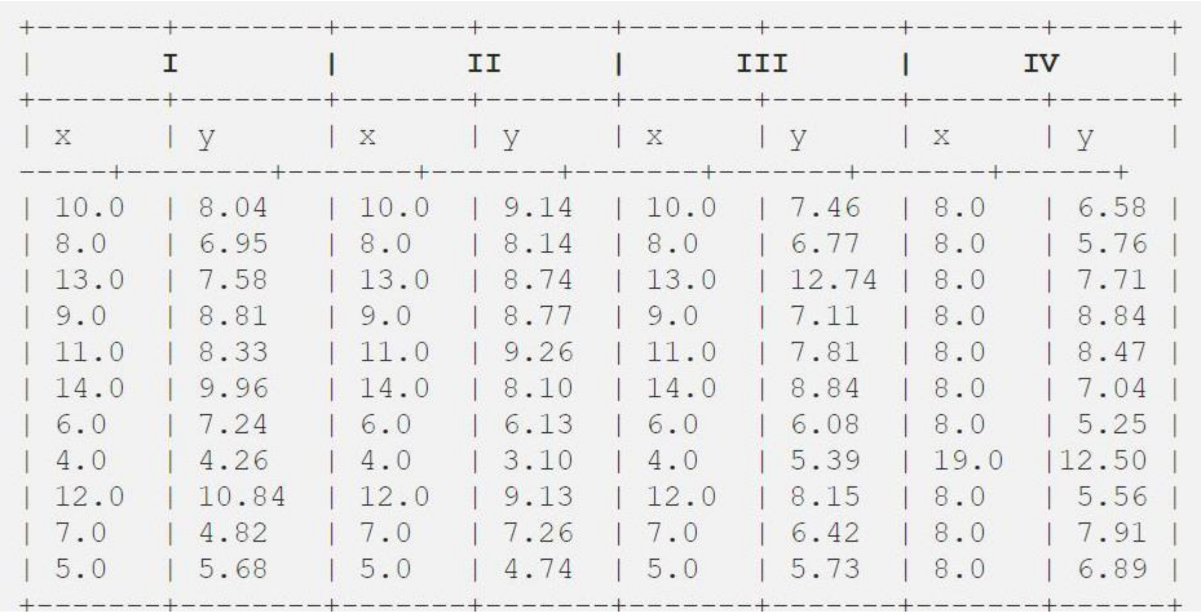
Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 datapoints

in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11

data-points are given below.

After that, the council analyzed them using only descriptive statistics and found the mean,standard deviation, and correlation between x and y.



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

What is Pearson’s R?

The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

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

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values
or units. If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Techniques to perform Feature Scaling:
1.Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.
2.Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

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

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between
the independent variables. If all the independent variables are orthogonal to each other, then VIF=1.0.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to 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.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:
a) It can be used with sample sizes also
b) 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.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior