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?

Answer:

Observations below:

   a. We see that the demand cycle sharing is higher during fall and lower in spring
   b. The demand trend aligns with season vs cnt observation. Demands are higher between May to Oct.
   c. the demand doesn't vary based on the days of the week
   d. people don't prefer to buy during Rainy season
   e. Demand is way higher in 2019 when compared to 2018. So it is a positive trend.
   f. Demand is higher in working day. So people prefer to avoid using the shared cycles on holidays.

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

Answer:

It is used to avoid the creation of extra columns while using the dummy variable concept. If there n unique values representing the categorical variable, then dummy variable concept can be used for generating n different variable from the n unique values, but we can actually drop the first unique value from n and have only n-1 variables. If the values against n-1 variable is false for a dependent variable then it must be true for the n^{th} variable.

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

Answer:

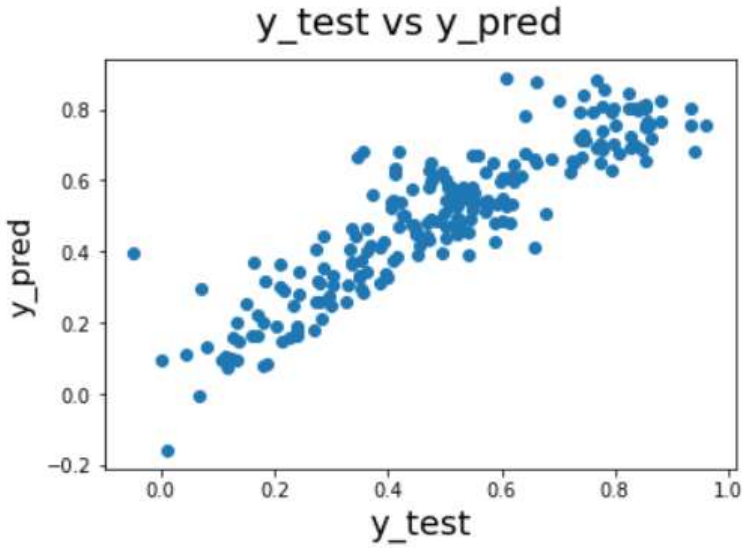
By looking at the pair-plot I think atemp variable (actual temperature) has the highest correlation of 0.63 with target variable cnt

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

Answer:

A scatter plot between the dependent and independent variable can be used to validate the assumptions of linear regression.

Based on the analysis on the boombikes dataset we can see that the predicted values based on the test data using the final linear regression object form straight line. This means that the assumptions based on linear regression holds good. Image below is taken from python notebook.



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

Answer:

3 features are:

1. Atemp (Actual Temperature) (Positive coefficient)
2. Weathersit : (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) – (Negative coefficient)
3. Yr (Year) has positive coefficient

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

It is a machine learning algorithm. It falls under supervised learning methods. It can be used for predicting the future outcomes. There are two types of linear regression:

- Simple linear regression
- Multiple linear regression

Simple Linear regression:

Expression:

$$Y = \beta 0 + \beta 1 X$$

$\beta 0$ is called as intercept and $\beta 1$ is called as slope.

This explains the relationship between a dependent variable and one independent variable.

Multiple linear regression

Expression:

$$Y=\beta 0+\beta 1x1+\beta 2x2+\beta 3x3$$

This technique is used to understand the relationship between one dependent variable and several independent variables.

In a real world scenario for a business case most of the time there are multiple variables which affects the value of the target variable. By using the multiple regression technique we find the linear equation which help us to predict the value of dependent variable using the independent variable.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet has 4 data sets with 11 x and y points. All the 4 data sets have same statistical properties i.e. mean, standard deviations and correlations are same , but when they are plotted in the graph using scatter plot , the graph appears different.

The four data sets are below:

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

Mean and SD data below:

```
In [3]: anscombe.describe()
```

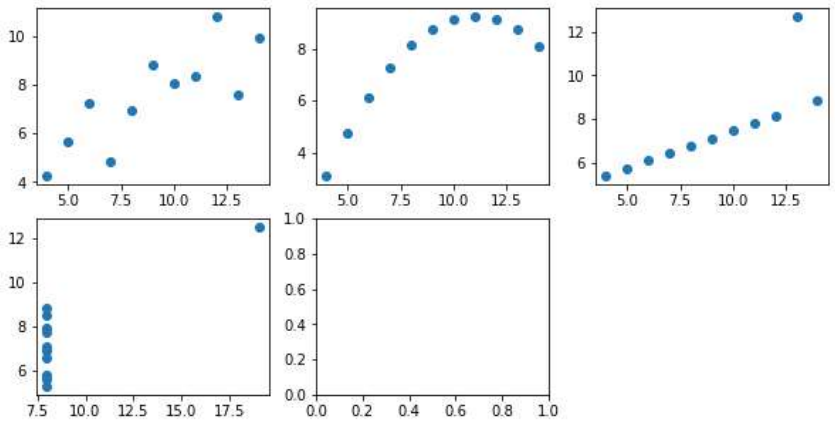
Out[3]:

	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

Scatter plot below:

```
In [14]: plt.figure(figsize=(10, 5))
         plt.subplot(2,3,1)
         plt.scatter(anscombe[['x1']], anscombe[['y1']])
         plt.subplot(2,3,2)
         plt.scatter(anscombe[['x2']], anscombe[['y2']])
         plt.subplot(2,3,3)
         plt.scatter(anscombe[['x3']], anscombe[['y3']])
         plt.subplot(2,3,4)
         plt.scatter(anscombe[['x4']], anscombe[['y4']])
         plt.subplot(2,3,5)
```

```
Out[14]: <AxesSubplot:>
```



We can see the scatter plot are different for 4 sets. This basically tells us that we must first look at the graphically.

3. What is Pearson's R?

   Answer:

   It is a measure of correlation between two variables. It is the most common way used for measuring the linear correlation.

   The value of Pearson R always falls within the range of -1 and 1.

If the value of pearson R is between 0 and 1 then it is positive correlation , if the value is 0 then there is no correlation and if it is between 0 and -1 then there is negative correlation.

Expression:

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer:
   Algorithms used for ML are affected by higher values of numerical features. So to avoid such case the values of the variables are re-scaled. Also the re-scaling is done to improve the prediction as the algorithms doesn't deal with higher numerical data.

Normalized Scaling:	Standardized scaling:
This technique is used for re-scaling the data so that the values are within 0 and 1. X' = X – X-min/ Xmax – Xmin.	In this method after the application of the method, the values subtracted from the mean and divided by standard deviation. X' = (X – mean)/ standard-deviation. There are not specific ranges.
It is affected by outliers	It is not much affected by outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer:
   If there is perfect correlation of the variable , then the VIF tends to have infinite values.

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

   Answer:

   Quantile-quantile plot is a graphical method used for comparing two probability distributions.

   If the 2 distributions are equal then the Q-Q plot will be a straight line. It is used to find if the distribution of the variable.

   Q-Q plot can be used to find if the test data and train data are derived from the same populations with same distributions.