International Institute of Information Technology Bangalore

upGrad

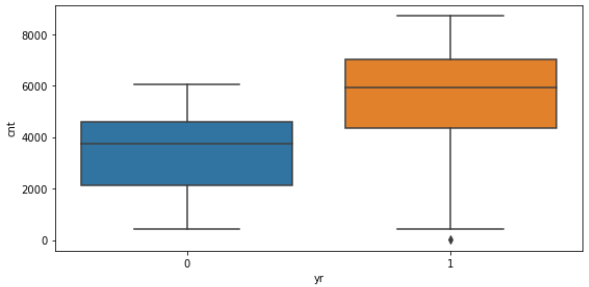
Bike sharing :
Assignment Submission

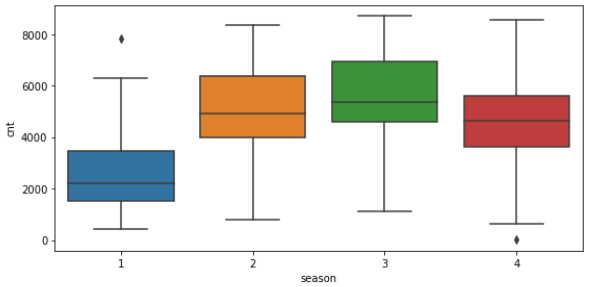
Prepared and submitted by

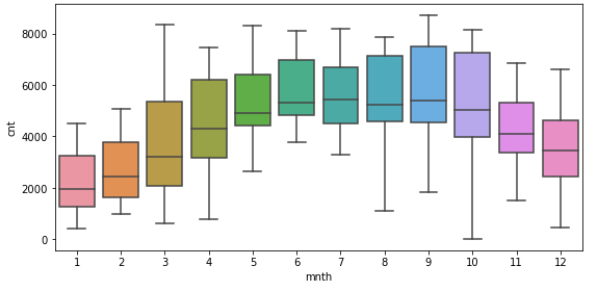
Krishna Murthi B

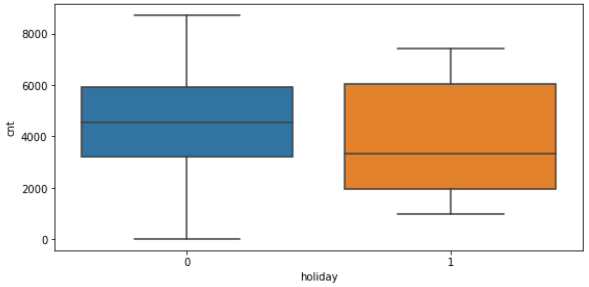
Course : Executive PG Programme in Machine Learning & AI Feb 2022
Date : 09-May-2022

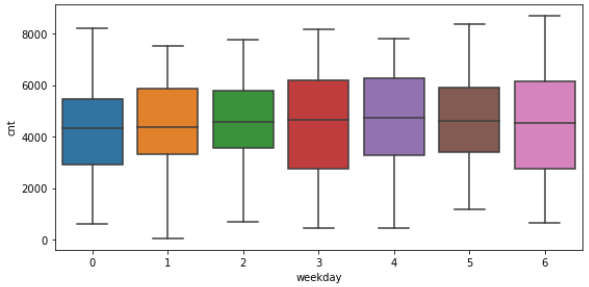
Assignment-based Subjective Questions

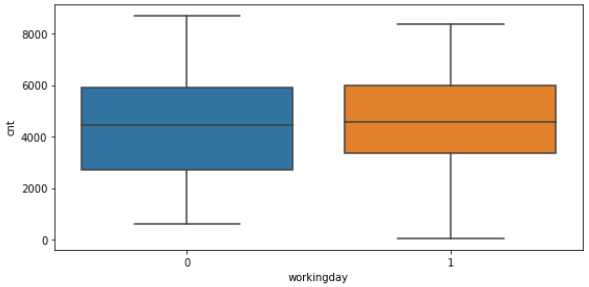


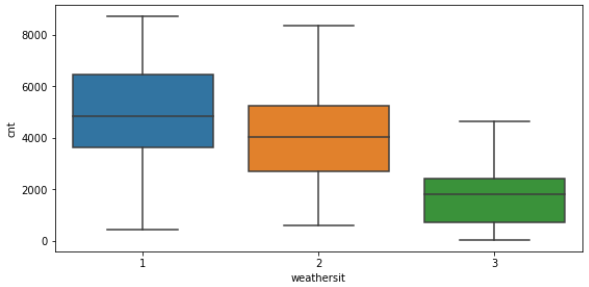












1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Observation

a) year 1 has more demand than year 0. Shows a growth in demand.

b) Weather situation is showing strong correlation with booking

b) month wise good demand is between apr to oct months

c) During holiday the median is less, showing the demand going less

d) The following fields are not good indigators for demand. All categories are showing similar demand.

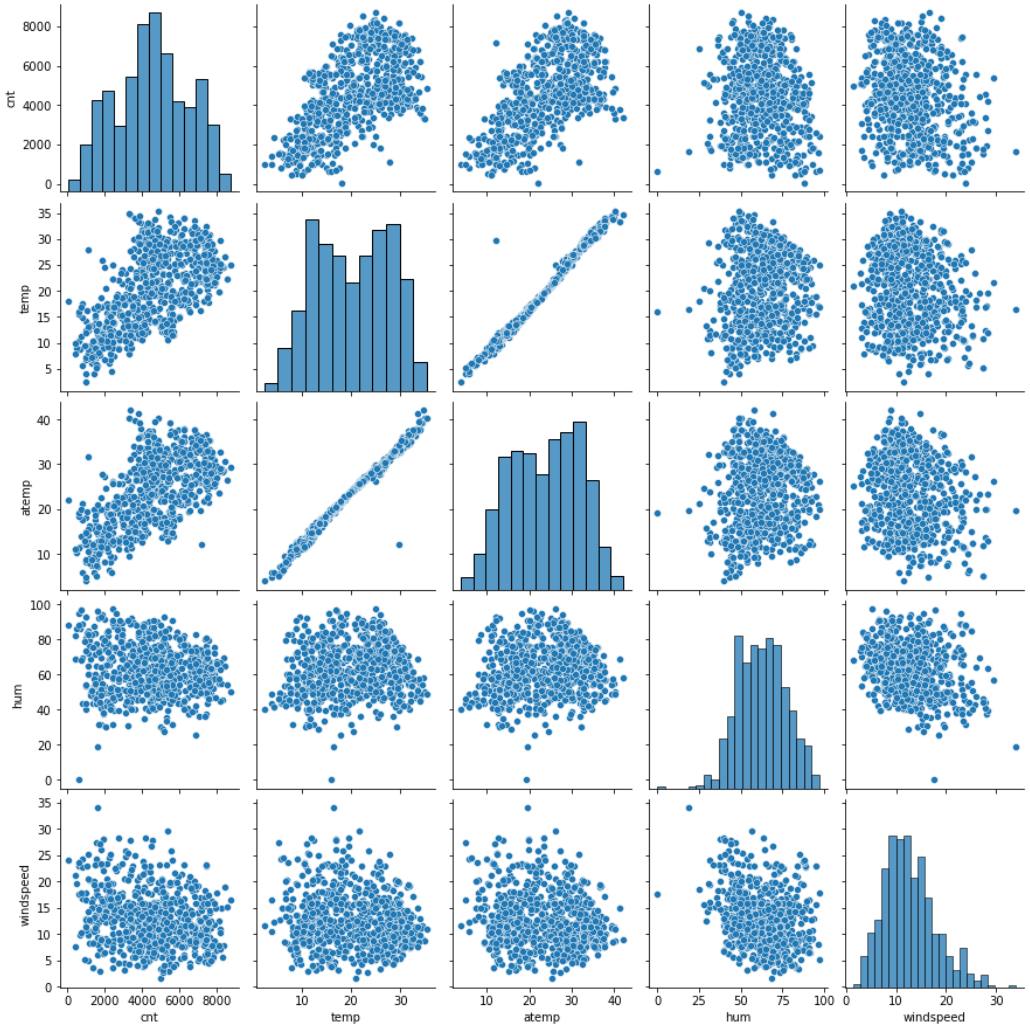
- Working
- Weekday

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

If a field has three values, say "Red", "amber", "Green", when we do dummy, it creates three columns. Now statistically one of the column is redundant. Because we can arrive to the value using k-1 values itself.

Since the set of all k dummies creates multicollinear, we need to drop one.

Assignment-based Subjective Questions



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

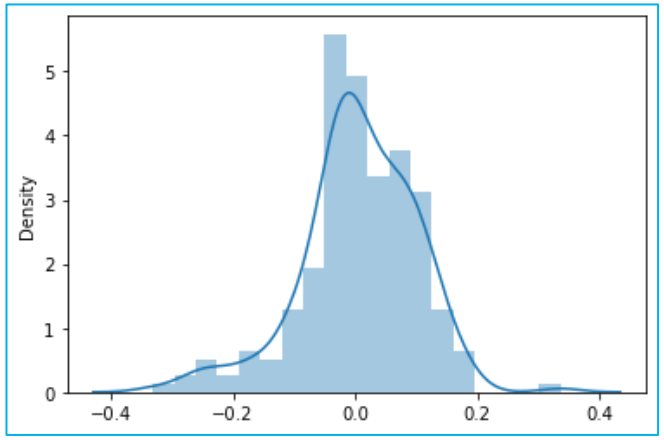
Observation

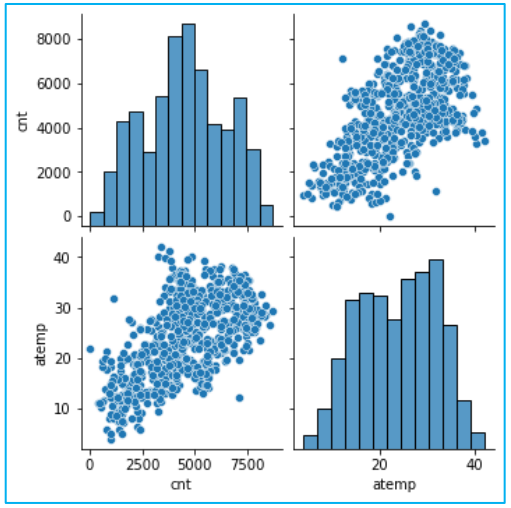
a) Looking at the pair-plot the atemp has the highest correlation

Assignment-based Subjective Questions

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

I validated the following assumptions






```
--- Comparing VIF values ---

           Features   VIF
2             atemp  5.19
3         windspeed  3.94
9   weathersit_good  2.84
0                yr  2.05
4     season_spring  1.68
5     season_winter  1.40
6          mnth_jul  1.33
7         mnth_sept  1.20
8    weathersit_bad  1.11
1           holiday  1.04
```

Error terms are normally distributed with mean zero.
This is proved by plotting a normal curve. error terms are also normally distributed

There is a linear relationship between X and Y
This is satisfied by having linear relationship between atemp and cnt

No Multicollinearity exists between the predictor variables
The values (other than the atemp) are well below 5.

Assignment-based Subjective Questions

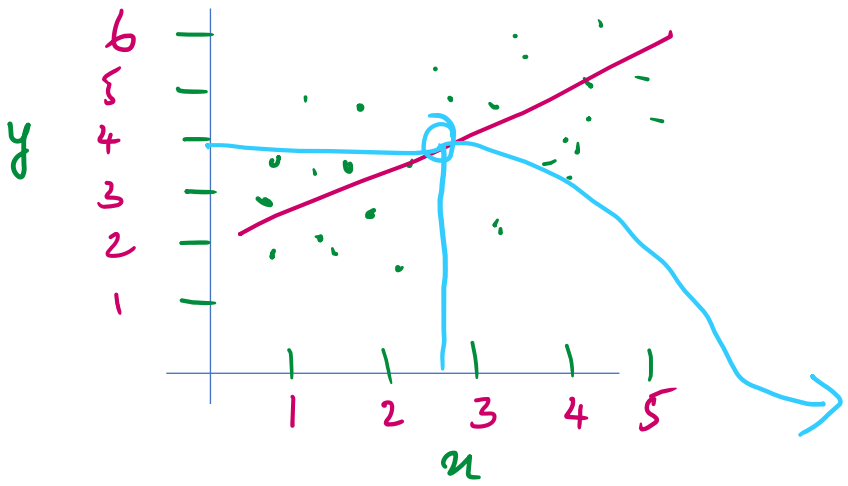
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

1. Demand of bike increases with the increase in atemp
2. Demand increases year on year
3. Demand of bike decreases whenthe weather situation is bad

Variable	Coefficient	Description
atemp	0.4642	Indicates that a unit increase in temp variable increases the bike hire numbers by this many units
yr	0.2351	Indicates that a unit increase in yr variable increases the bike hire numbers by this many units
weathersit_bad	-0.2002	Indicates that a unit increase in weathersituation bad variable **decreases** the bike hire numbers by this many units
windspeed	-0.1256	Indicates that a unit increase in windspeed variable **decreases** the bike hire numbers by this many units
season_spring	-0.1174	Indicates that a unit increase in season_spring variable **decreases** the bike hire numbers by this many units

General Subjective Questions

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



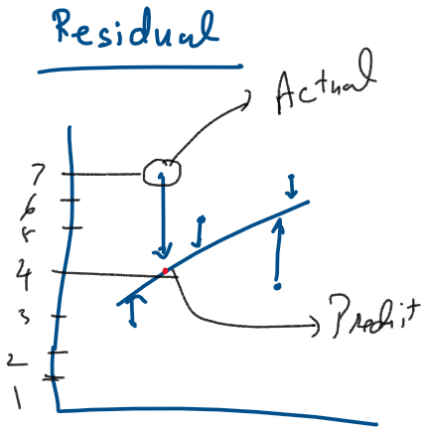
Green dots : They are the existing values
Straight line : That is the line we fit
Purpose : The purpose of fitting a straight line is to help us predict “y” value for any “x” value.

Y = mx + C is the formula for the straight line also written as Y = B0 + B1 x

C is the constant (intercept)
m is the slope
X is the independent variable

For eg, we may have actual data for x =2 and x=3. if we need to forecast for 2.5, using the y=mx+c we can get the y.

Linear Regression is an algorithm based on supervised learning.



Once the line is fit, we have to find out whether the line is the best fit line using the RSS and TSS.

RSS : this is computed by considering the straight line, difference between the line and the actual data point, sq the diff and add them

TSS : this is computed by considering the avg of all data points, get diff, sq the diff, add them

R^2 = 1 - (RSS/TSS)

General Subjective Questions

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

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

Summary statistics

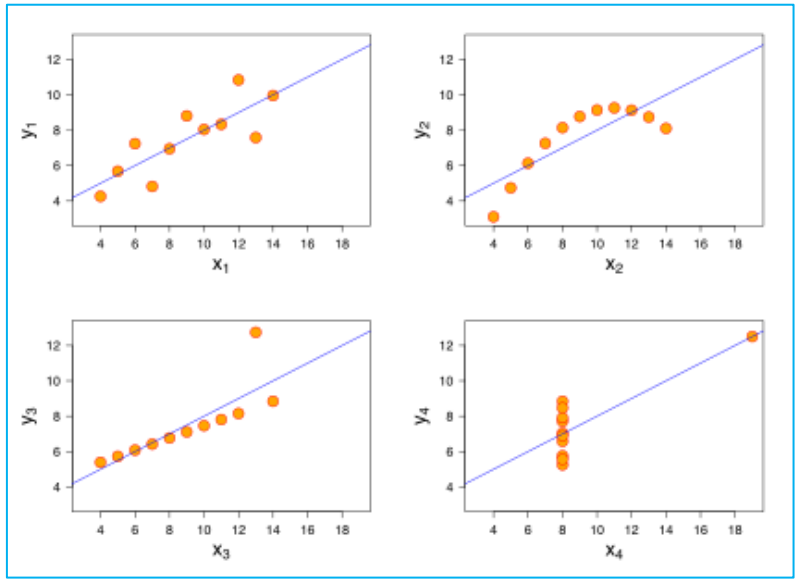
Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

My key takeaway from Anscombe's quartet is that we have to visualize data using graphs, it is important to plot our data. Summary statistics alone is not sufficient.

Now looking at the data in the table, the summary statistics for all 4 sets it looks the same.
But when we plot them as graphs, its all look totally different. So it proves how much Anscombe has visualized this data in his dream!

Plot the graph!



General Subjective Questions

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

Pearsons's R measures the strength of the linear relationship between two variables.

Pearson's R is always between -1 and +1

- The correlation coefficient lies between -1 and +1. *i.e.* $-1 \leq r \leq 1$
- A positive value of '*r*' indicates positive correlation.
- A negative value of '*r*' indicates negative correlation
- If $r = +1$, then the correlation is perfect positive
- If $r = -1$, then the correlation is perfect negative.
- If $r = 0$, then the variables are uncorrelated.

The formula

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{(\sum x_i^2 - \frac{(\sum x_i)^2}{n})}\sqrt{(\sum y_i^2 - \frac{(\sum y_i)^2}{n})}}$$

General Subjective Questions

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

When we have a multiple variables in our data, they may be in different scales. Eg from our housing dataset, price, area, no of bedroom, no of bathrooms etc are in different scales. It is necessary to bring everying between 0 to 1 so that we can effectively put them in perspective and build a model.

When every predictive variables are in same scale, it is easy to interpret them and easy to use in model.

- We fit&transform for the train data.
- We only transform the test data.

Standardizing

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

Standardizing

Scaling is done using standard deviation

Min-Max

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Min max

Using the maximum and minimum of the data, values are set between zero and one.

General Subjective Questions

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

$$VIF_i = \frac{1}{1 - R_i^2}$$

> 10 high co relation

> 5 need to check

< 5 we can have this variable

When two variables are perfectly corelated then the value of the VIF is very high.

In such cases, we need to drop one variable and see how the other one behaves (and try vice-versa as well.)

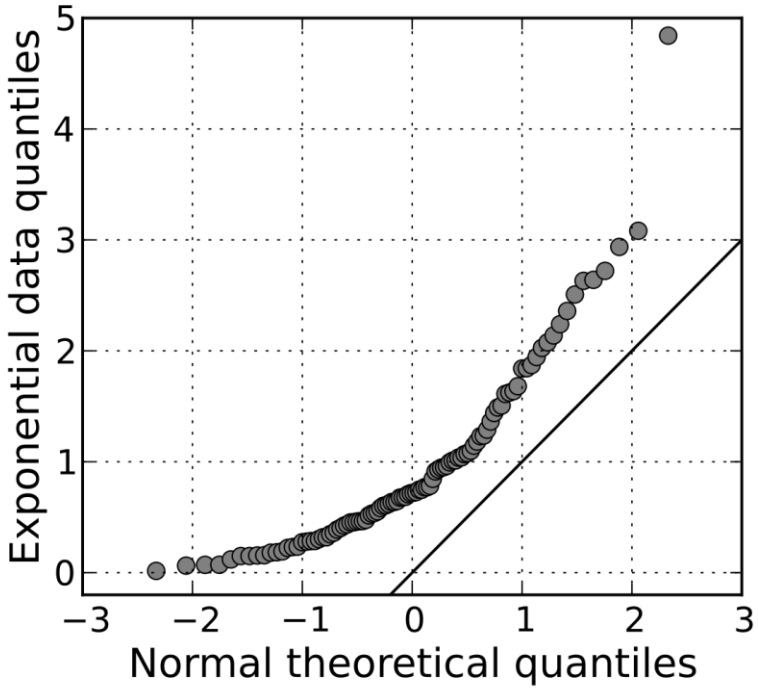
By having high correlated variables we wont get desirable results in model.

General Subjective Questions

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

According to Wikipedia, QQ plot is
"In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other."

By plotting two sets of quantiles against one another, we can build a QQ plot. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.



<u>Usage of Q-Q plot to name a few</u>

- Can be used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions
- Is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples

https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot