Bike Sharing Assignment

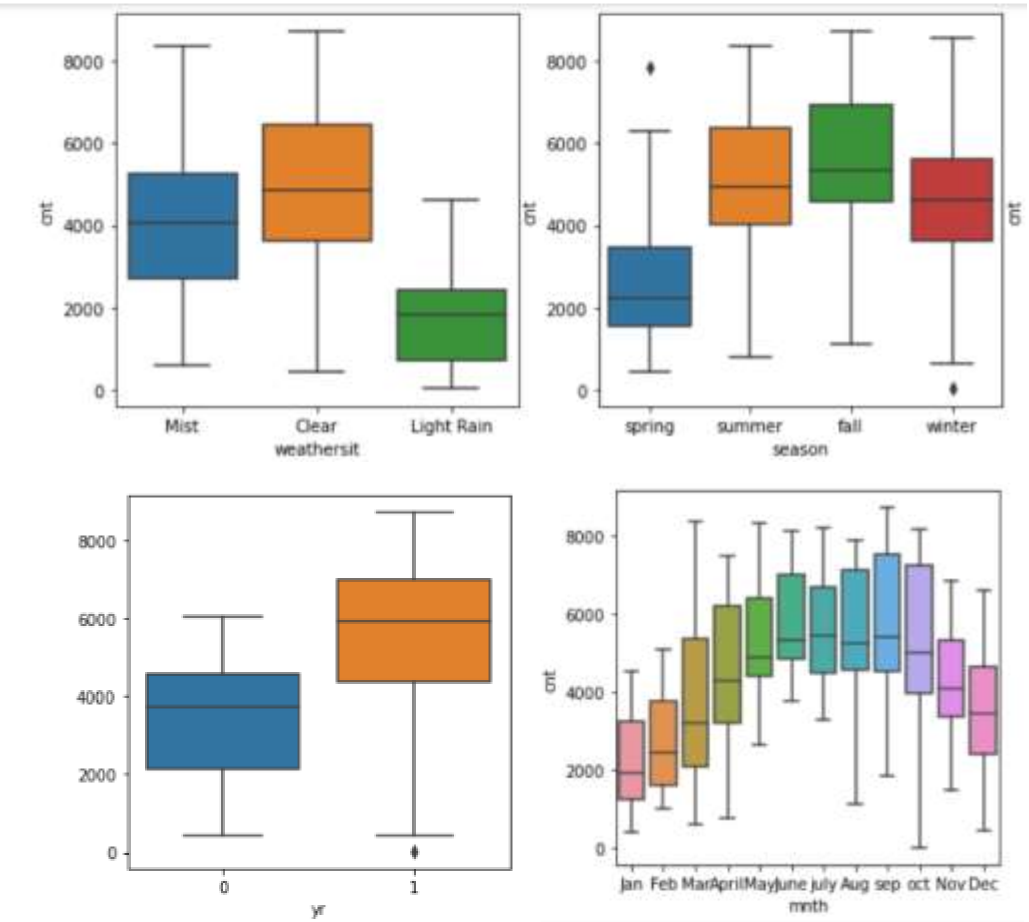
Linear Regression

NAMRATA SHIVTARKAR

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)

- Year 2019 shows more profit than 2018
- Clear and mist weather has strong effect on bookings,
- Demand increases gradually from jan to sep that is in fall and summer season
- During holiday the median is less, showing the demand going less
- Weekday and working day does not give any information since weekday has all values in same rage also same for working day



ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

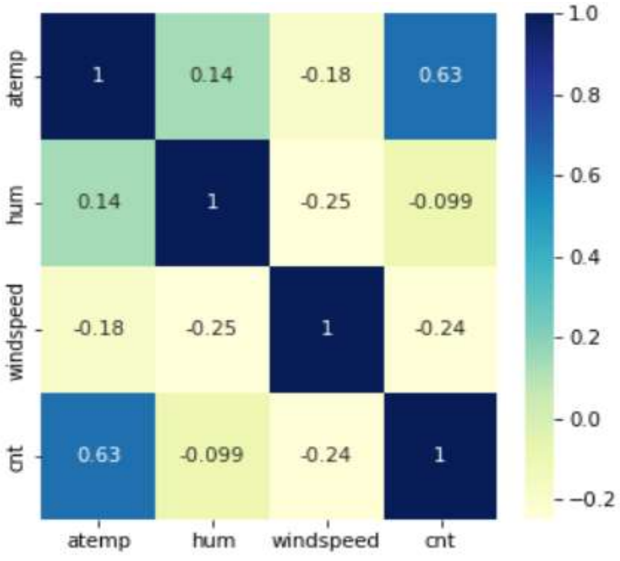
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

- suppose we have 3 columns furnished, semi-furnished, non-furnished. After creating dummy variables we get values in binary, so if semi-furnished and non-furnished are 0 that means its is obviously furnished, so we can delete 1st column furnished to remove redundancy

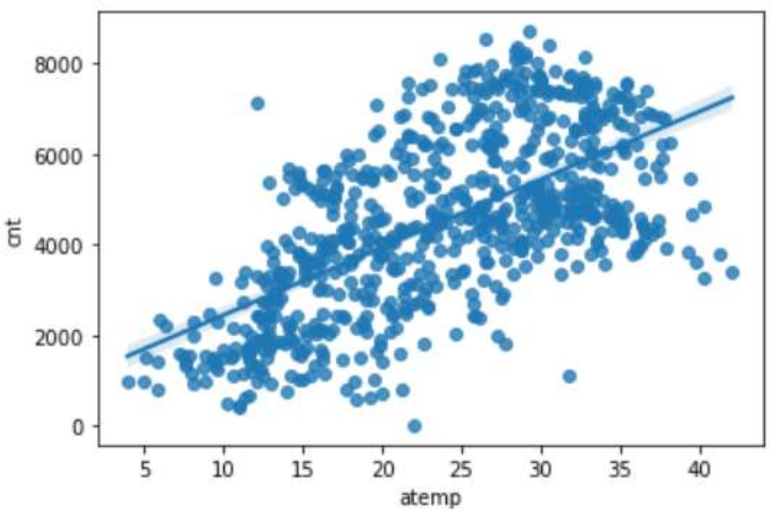
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

- atemp has the highest correlation with target variable



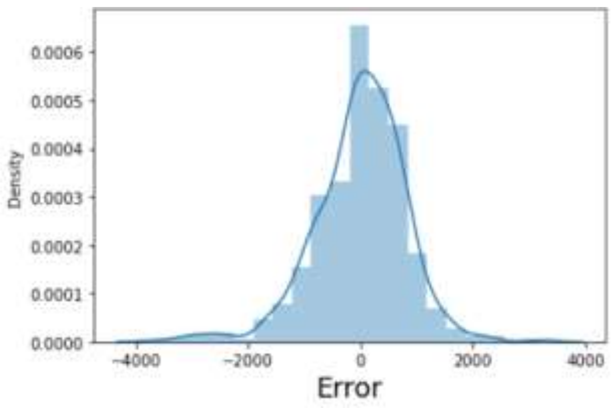
	atemp	hum	windspeed	cnt
atemp	1	0.14	-0.18	0.63
hum	0.14	1	-0.25	-0.099
windspeed	-0.18	-0.25	1	-0.24
cnt	0.63	-0.099	-0.24	1



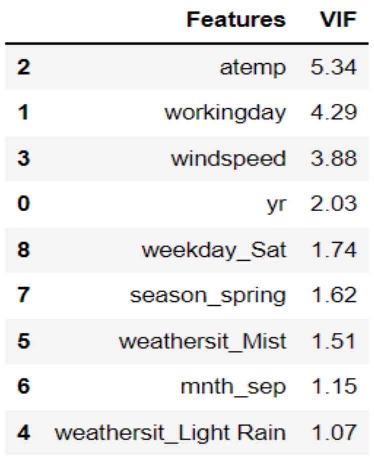
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

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

- Below distplot shows normal distribution of error terms with mean 0

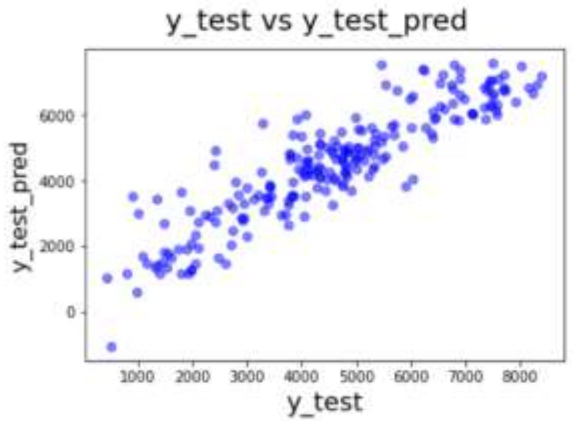


- Values of VIF apart from atemp are below 5



	Features	VIF
2	atemp	5.34
1	workingday	4.29
3	windspeed	3.88
0	yr	2.03
8	weekday_Sat	1.74
7	season_spring	1.62
5	weathersit_Mist	1.51
6	mnth_sep	1.15
4	weathersit_Light Rain	1.07

- Variance of residual is constant is seen below



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)

- Demand of bike increases with the increase in atemp by 3214 unit
- Demand increases every year by 2062 unit
- Demand of bike decreases when whether is bad by -2441 unit

GENERAL SUBJECTIVE QUESTIONS

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

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

Slope of given line is (y2-y1)/(x2-x1), x1y1 and x2y2 are any two points through given line passes.

. The line passes through (0,3)(2,4),so 4-3/2-0 = ½ = 0.5

m is the slope, C is the constant (interception) and X is the independent variable.

B1 is independent(predictor) variable

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.
Ex: x=2, y=4, but 7 is measured(actual yi), so 7-4=3 is residual
RSS = y1-B0-B1x)sq + yn-B0-B1xn)

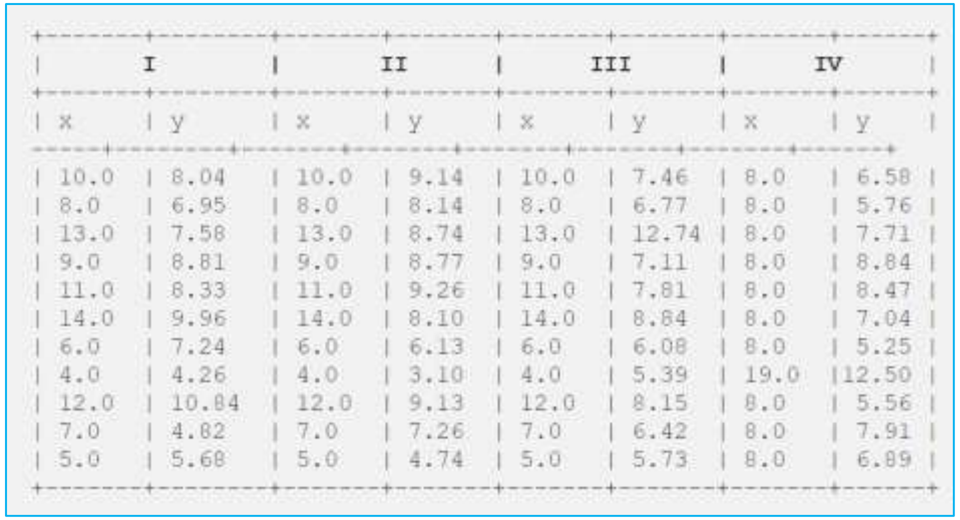
TSS : this is computed by considering the avg of all y-value data points, get diff, sq the diff, add them

R^2 = 1 - (RSS/TSS)

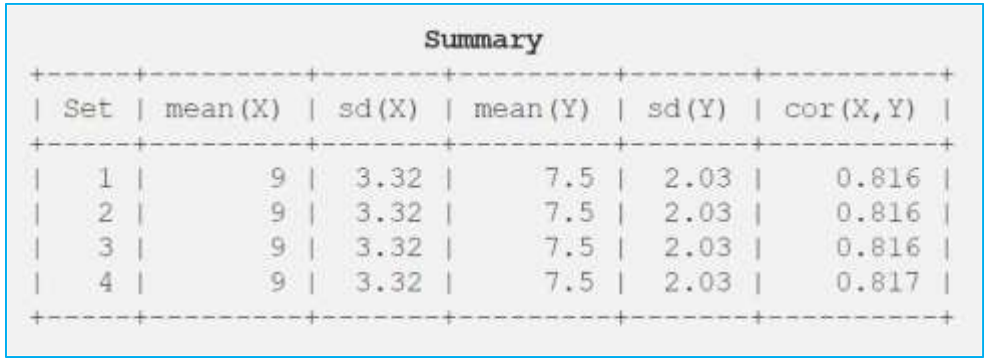
GENERAL SUBJECTIVE QUESTIONS

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

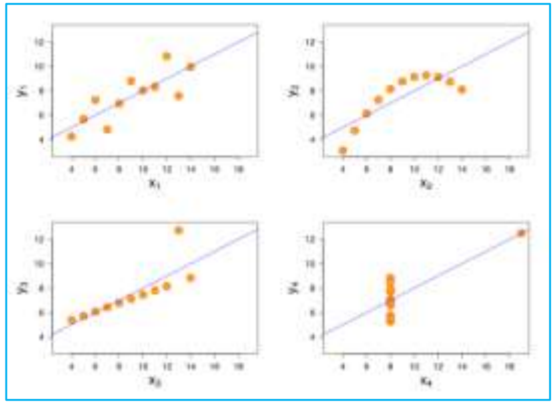


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

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



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 ≤ r ≤ 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.

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

GENERAL SUBJECTIVE QUESTIONS

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

- We have a multiple variables in our data, they may be in different ranges. Eg housing dataset, price, area, no of bedroom, no of bathrooms etc are in different scales. It is necessary to bring everthing between 0 to 1 build a model on same scale
- 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
  - Scaling is done using standard deviation
- 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)

- When two variables are perfectly corelated then the value of the VIF will be 1, so we drop one variable to see the changes of other variable

GENERAL SUBJECTIVE QUESTIONS

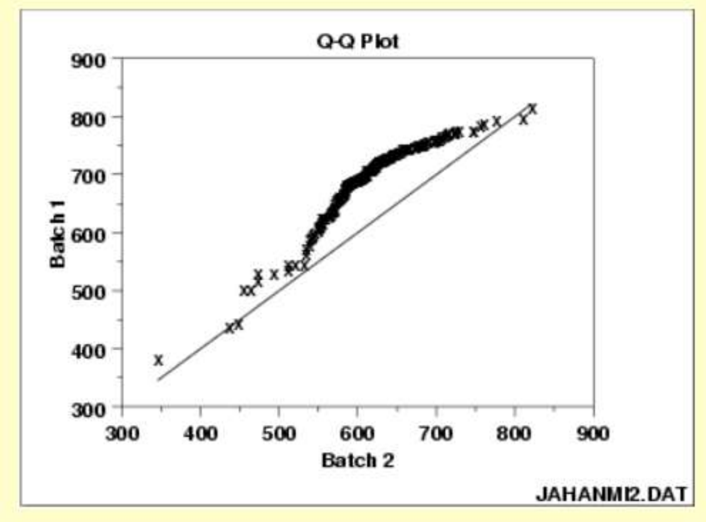
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

- use and importance of a Q-Q plot

1.The sample sizes do not need to be equal.

2.Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.


Q-Q Plot
Batch 1
Batch 2
JAHANMI2.DAT