Vaibhav Jain

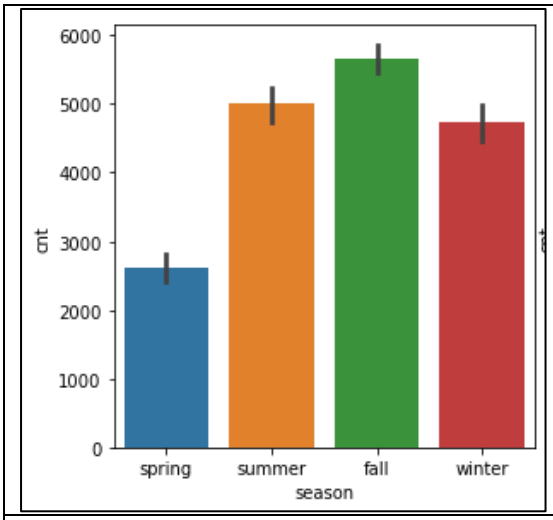
Linear Regression Assignment

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?

Ans:

Lets look at each category and its effect on the dependent variable 'cnt'

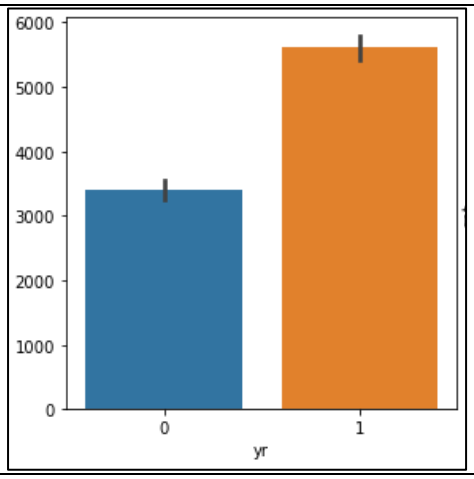


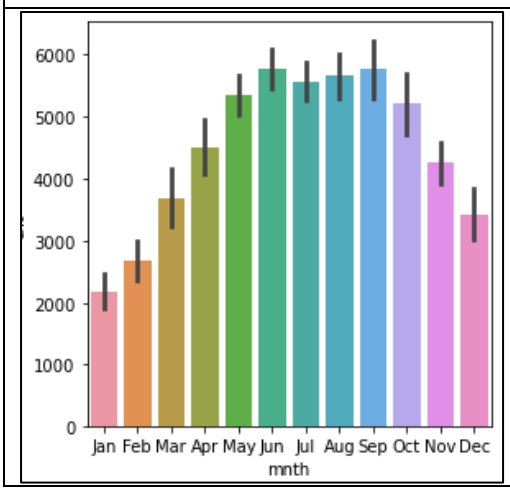
Season Vs Count

We can see bike demand is highest during fall season, and lowest in spring season

Year vs Count

Rental Bike demand has significantly increased in year 2019 (represented by 1 in graph)
This is inline with the comments mentioned in problem statement, that business has grown over years



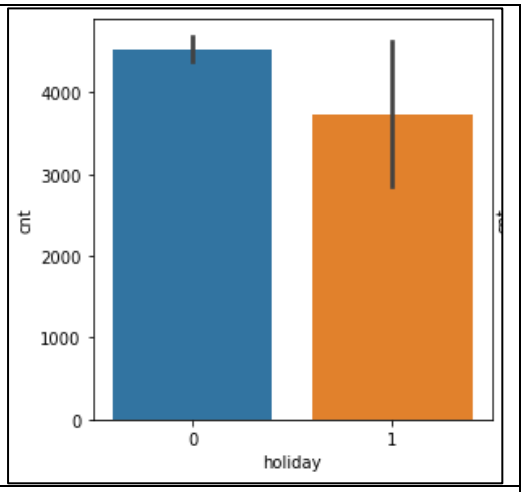


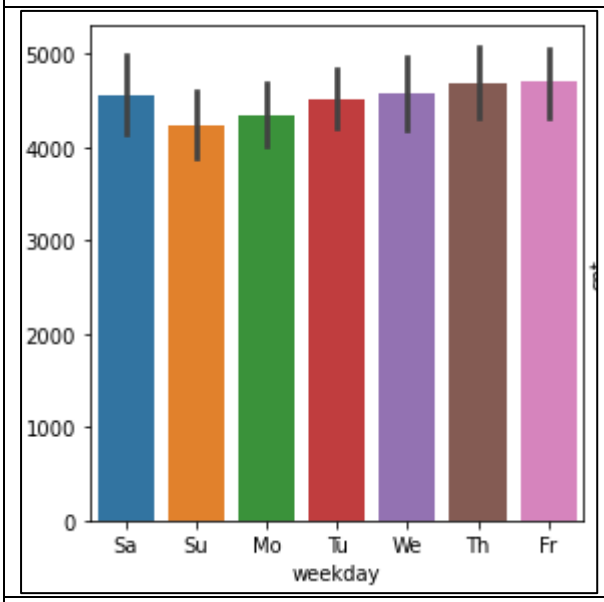
Month vs Count

Bike rental demand grows continuously from Jan to June, and then starts declining from Sep till Dec

Holiday vs Count

When its not a holiday people prefer to rent bike more.



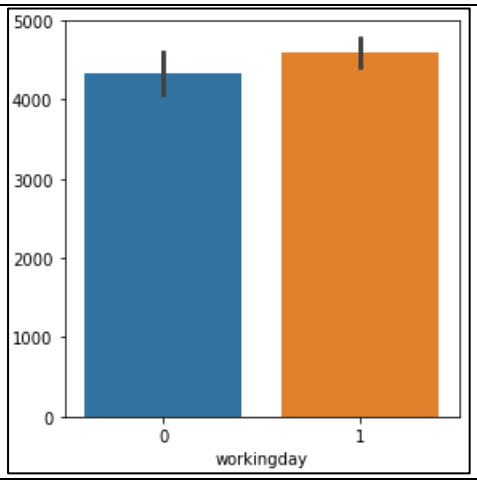


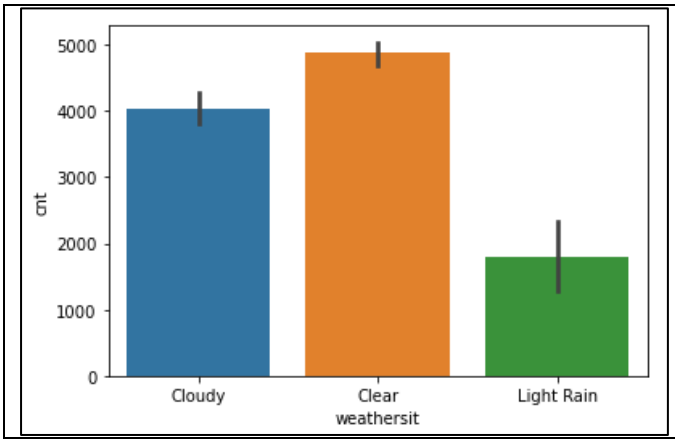
Weekday vs Count

There is no significant pattern with respect to weekday. Bike rental demand is pretty much same over the week, with very slight dip observed on Sundays.

Working Day vs Count

There is no significant change in number of users whether it's a working day or not.





Weather Vs Count

Weather shows a clear graph that bike users are most when weather is clear. Whenever it rains the demand is effected signigicantly.

Based on above data visualizations, I can infer that bike rental demand is effected by season, year, month, holiday and weather, and no significant change based on weekday and working day.

---

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

Ans:

Dummy variable is used to create columns for categorical variables when encoding the dataset before building the model. By default a dummy variable is created for each category level by the method 'pd. get_dummies'. For e.g.

If we have a categorical variable weathersit, and it has 4 levels, Given dataset stores these levels as values of 1,2,3,4, each representing a different weather.

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy
- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog,

So get_dummies will convert 1 column to 4 separate columns.

Weathersit	Clear(1)	Mist(2)	Light Snow(3)	Heavy Rain(4)
Clear(1)	1	0	0	0
Mist(2)	0	1	0	0
Light Snow(3)	0	0	1	0
Heavy Rain(4)	0	0	0	1

Where as actually we can represent all above information with only 3 columns, and deduce when all 3 column values are 0 means that it would be fourth category. So applying **drop_first=True** will result in below (one less column)

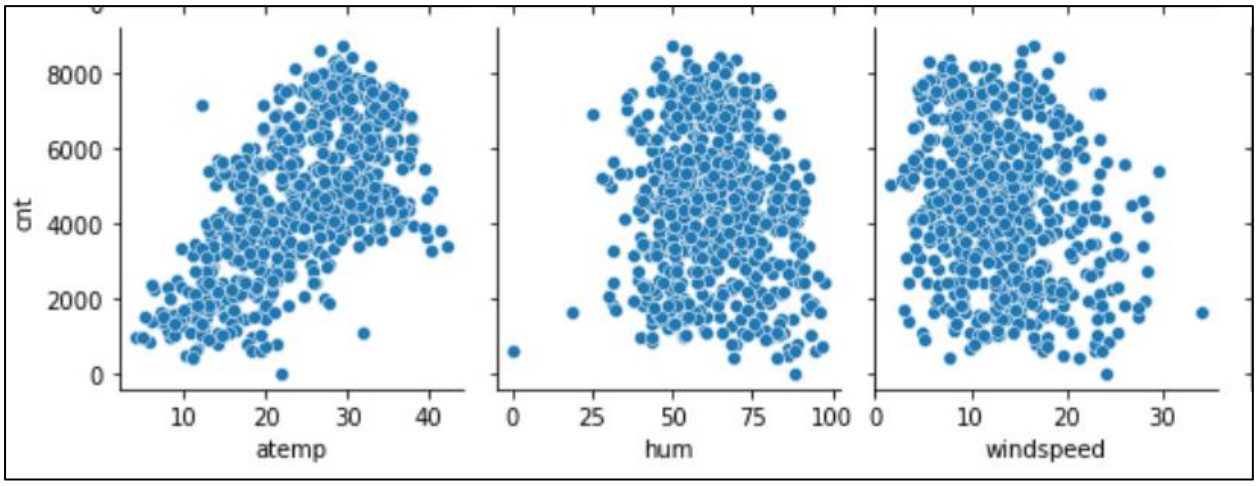
Weathersit	Mist(2)	Light Snow(3)	Heavy Rain(4)
Clear(1)	0	0	0
Mist(2)	1	0	0

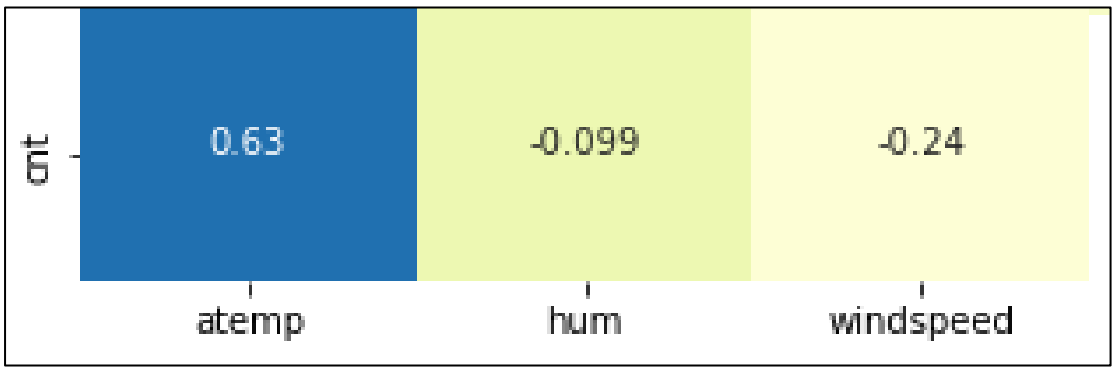
Light Snow(3)	0	1	0
Heavy Rain(4)	0	0	1

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

Ans:

Below is pair plot of numerical variables with 'Count' target variable:





Both pair-plot and correlation matrix shows atemp has the highest correlation coefficient.

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

Ans:

- Residual analysis plot shows the residuals follow normalization and are centered around 0
- Make sure all p-values are < 0.05
- All categorical variables are converted to dummy variables
- Find out independent variables must not be correlated, like temp and atemp

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

Ans:

```
#These params shows how each variable affects the demand of rental bike
lr_model.params
```

```
const          0.264917
yr             0.238661
workingday     0.053781
atemp          0.327716
windspeed     -0.140634
spring        -0.078302
winter         0.079503
Aug            0.036025
Dec           -0.072384
Feb           -0.052565
Jan           -0.093779
Jun            0.045175
May            0.049406
Nov           -0.062330
Sep            0.081433
Sa             0.064598
Cloudy        -0.082147
Light Rain    -0.294868
dtype: float64
```

Based on the model outcome, top 3 features contributing to count target variable are:

- atemp
- yr
- Light Rain (weather type category 3)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a statistical method that is used for predictive analysis. This is used to make predictions for continuous numerical variables for e.g. price of house, salary of person, number or users etc.

Linear regression predicts a dependant variable based on given independent variables, and builds a linear equation.

y = constant + B1x1 + B2x2 + B3x3 + … + Bnxn

once the constants are known, a value of target variable y can be predicted based on input variables x.

→Simple Linear Regression

When we have a single input, we can use statistics to estimate the coefficients, this is represented by linear equation of

y = mx + c

we can predict y based on any value of x when we have constnat c and cofficient m known.

For e.g. predict weight based on single input variable height

→Multiple Linear Regression

This involves multiple input variables and each having its own coefficient:

y = constant + B1x1 + B2x2 + B3x3 + … + Bnxn

here analysis becomes more complex and needs to satisfy various assumptions before building the model.

For e.g. predict weight based on input variables such as foodhabits, exercise habits, dna, geography, income level, etc..

---

2. Explain the Anscombe's quartet in detail.

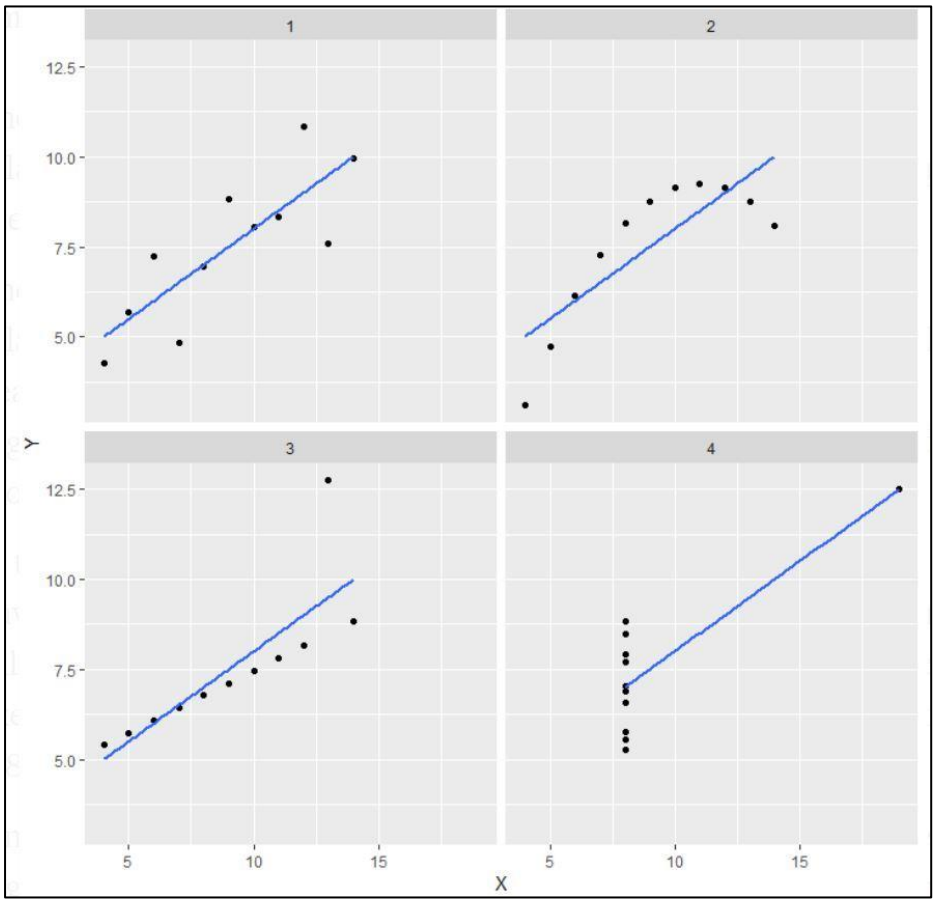
Ans:

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.

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



Each given dataset has same mean, standard deviation for x and y, where as when represented in graph they all look totally different.

The objective was to emphasize the importance of visualization before starting any data analysis.

---

3. What is Pearson's R?

Ans:

Correlation measures the strength of association between two variables. It is always between -1 and 1. Pearson's R is one of the type of correlation.

Pearson'r R is used to measure relationship between two continuous variables.

It is calculated with this formula:

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

Where,

N = the number of pairs of scores

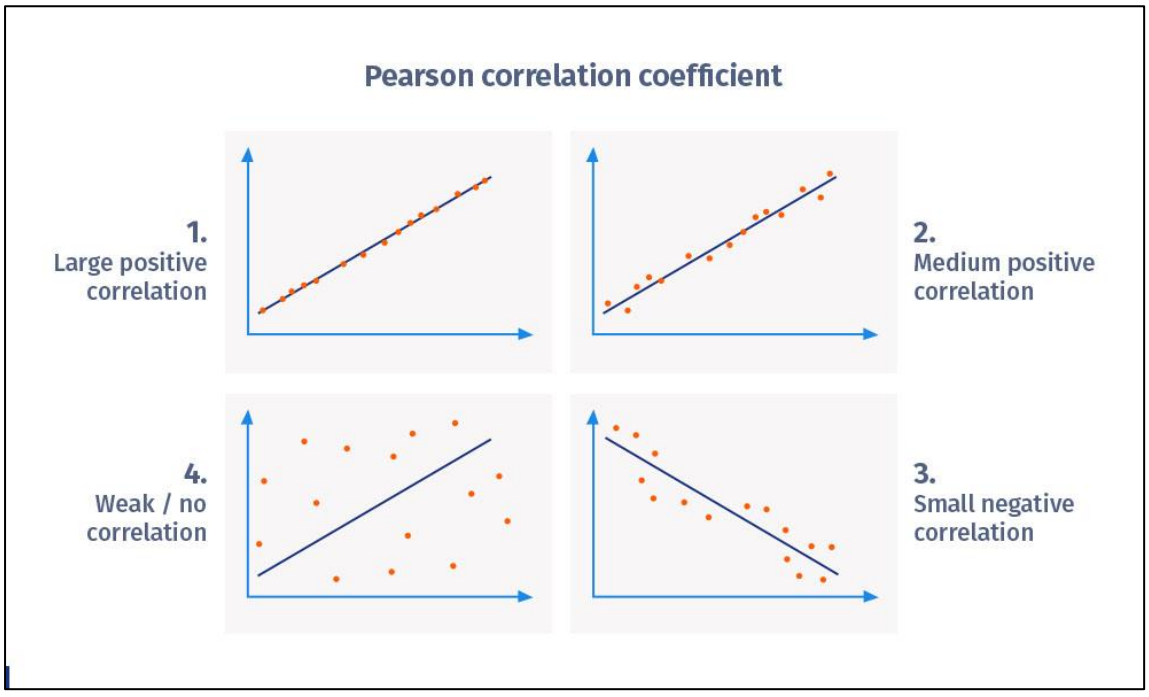
Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores



---

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

Ans:

Scaling is a technique to standardize the independent variables present in the data in a fixed range. For e.g. we can have temperature in range of -10 to 45, where as humidity in 0 to 80, or rain in 0 to 50mm, so scaling brings all these ranges in standard range irrespective of the unit used to represent the data.

Scaling is performed to bring all variable values to same magnitude, otherwise 4000 mililitre will be considered as greater than 4 litre where as infact those are same.

Normalized scaling and standardized scaling – both uses different formula to perform the scaling.

Normalization brings all values between 0 and 1. Where as standardization brings data mean to 0 and variance to 1

Below image shows formula for calculating each one:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **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.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

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

Ans:

The formula to calculate VIF is

$$VIF = \frac{1}{(1-R^2)}$$

This will become infinite, when R-Square is 1. R-Square only becomes 1 rarely when model is 100% perfect in predicting the output or in other words 100% of variation in data is explained by the model.

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

Ans:

Q-Q Plots are plots of two quantiles against each other.

Q-Q Plot determines if a curve is normally distributed. Q-Q plot is used to find the type of distribution for a random variable. Distribution type can easily be identified just by looking at Q-Q plot.