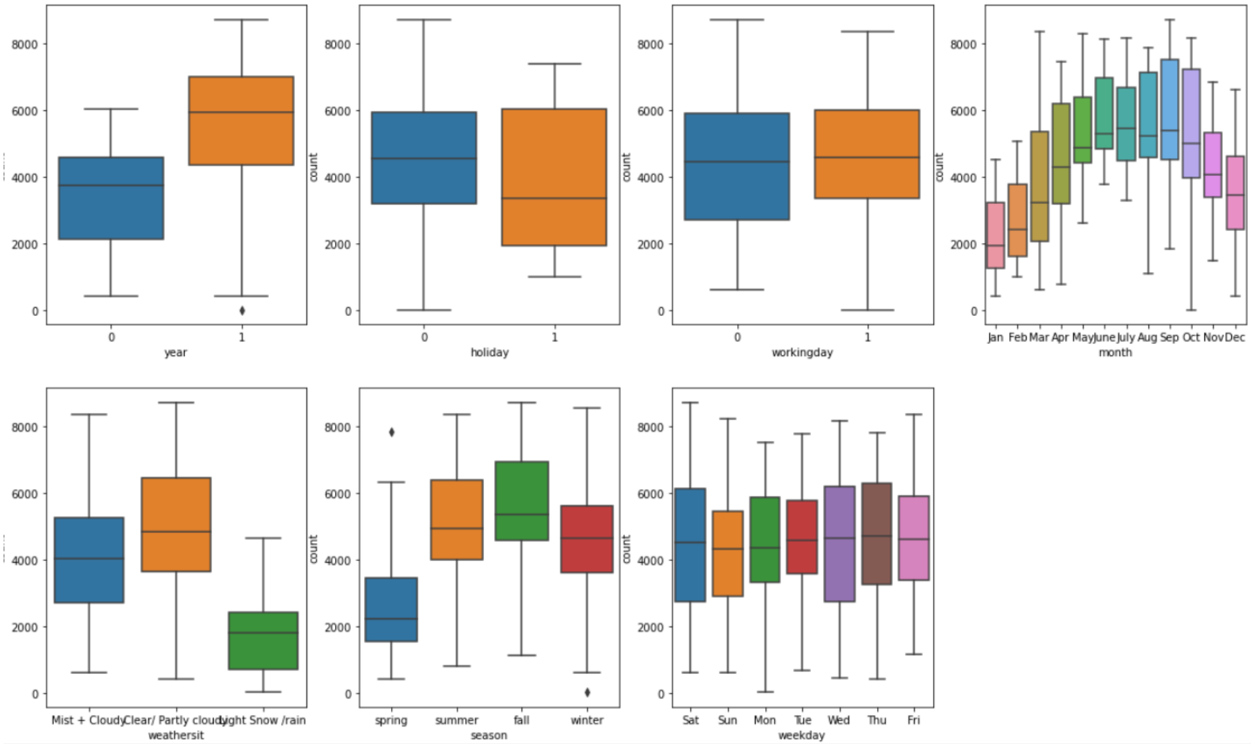
Assignment-based Subjective Questions

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

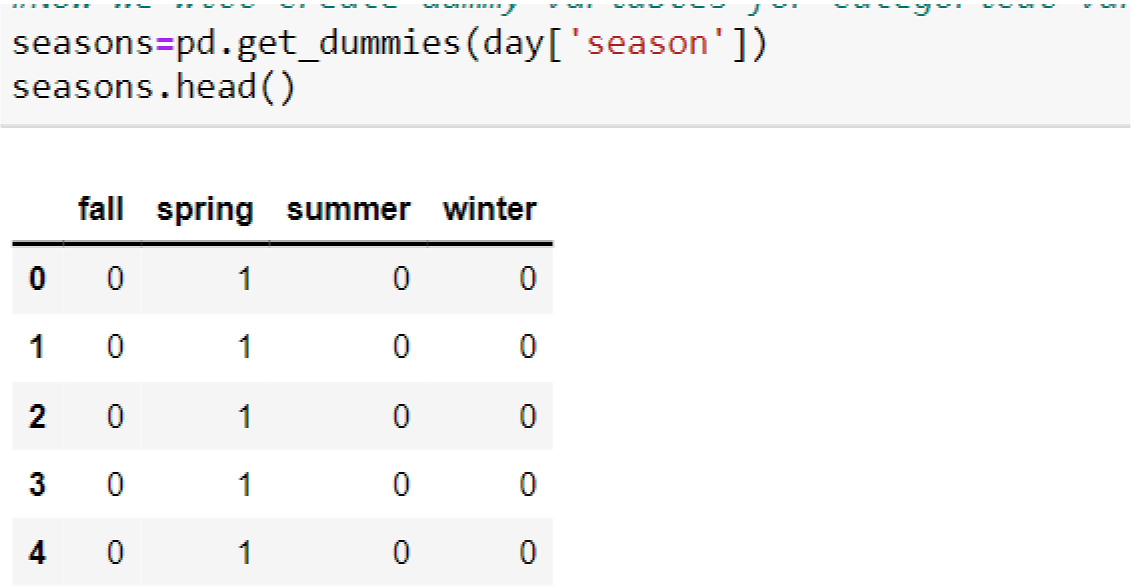
Ans: The dependent variable was found to be higher in the following cases-

- In the year 2019
- On a non-holiday
- In the month of July and September
- On partly cloudy day
- In fall season
- On Saturday



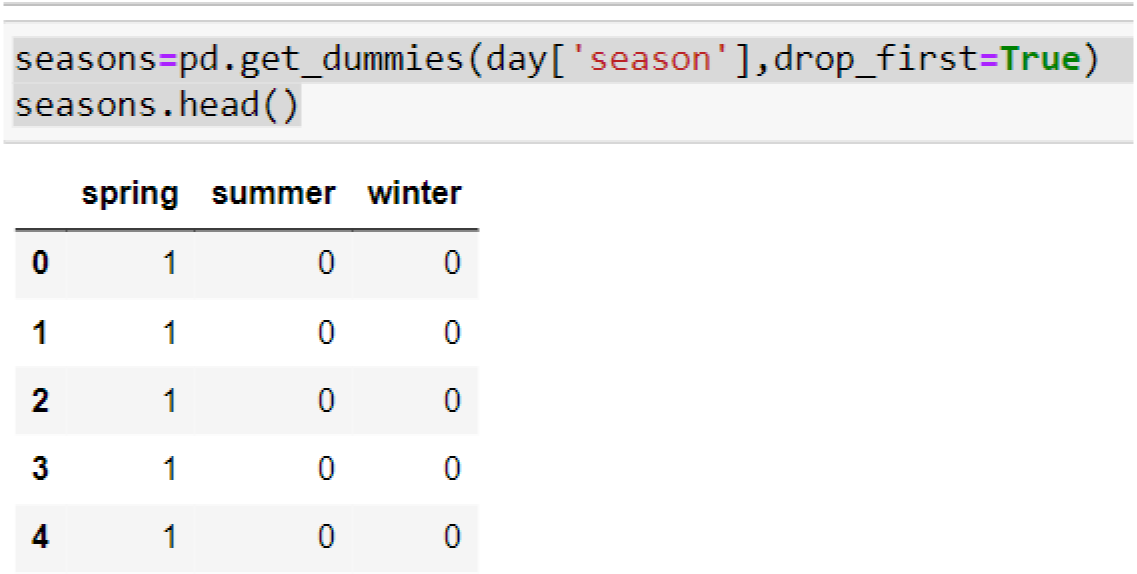
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans – The drop first=True command is used during dummy variable creation to delete the extra column. If we have 4 unique entries for which we need to create dummy variables, it creates 3 dummy columns instead of 4.



```
#Now we will create dummy variables for categorical var
seasons=pd.get_dummies(day['season'])
seasons.head()
```

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

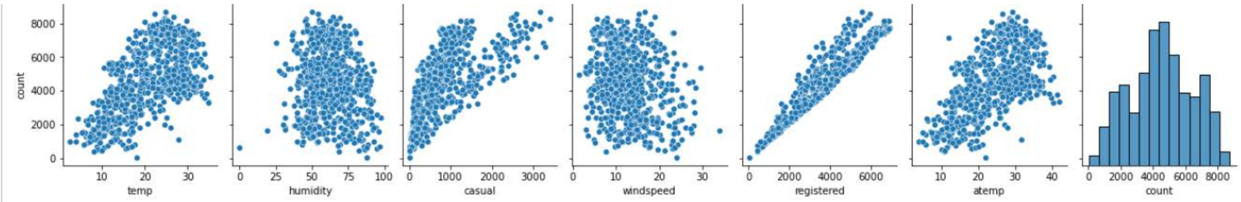


```
seasons=pd.get_dummies(day['season'],drop_first=True)
seasons.head()
```

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

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

Ans. Registered



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

Ans We see the summary of the model created to check if the model created is a good fit or not.

- The R- squared should be close to 1
- None of the varibles should have p value greater than 0.05
- None of the variables should have VIF greater than 5

If these 3 conditions are met then we can say the model is a good fit

After this, we need to run the test data model and compare the R-squared. If they are have a minor difference then we can say our model is a good fir.

To choose variables to eliminate, we have 3 options-

High VIF, High p value - remove

Low VIF, low p value - keep

High-Low – in this case we prefer removing the variables with high p value

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

Ans: We can get this by seeing the coefficients of these variables

1. Temperature

2. Light snow/Rain

3.Year

General Subjective Questions

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

Ans: Linear regression is of 2 types

-Single Linear regression

-Multiple Linear regression

Basic formula- y= b0 +b1X

Algorithm is as follows-

Read and understand data- Use the data dictionary provided with the data set and also try to infer from your understanding.

Visualize the data- Create pairplot for numeric variables and boxplot for categorical variables to understand the correlation

Make data easily readable- create dummy variables whenever required to make the data comparison easy or map to binary variables

Train Test Split- A very important step to divide the given data into 2 models-1 to train and other to test. It is important to provide a random state variable so that the data split remains constant every time we run the model

Normalization-This is done to bring the values between 0-1 for easy comparison

Choose X and Y variables for our rain dataset – We store our target variable in y and other independent variables in X

Add constant to variable in X as X_train does not automatically add it

Create the model – Formulation of the basic equation of our model

Fit the model – Use our dataset to find values of coefficients of our variables

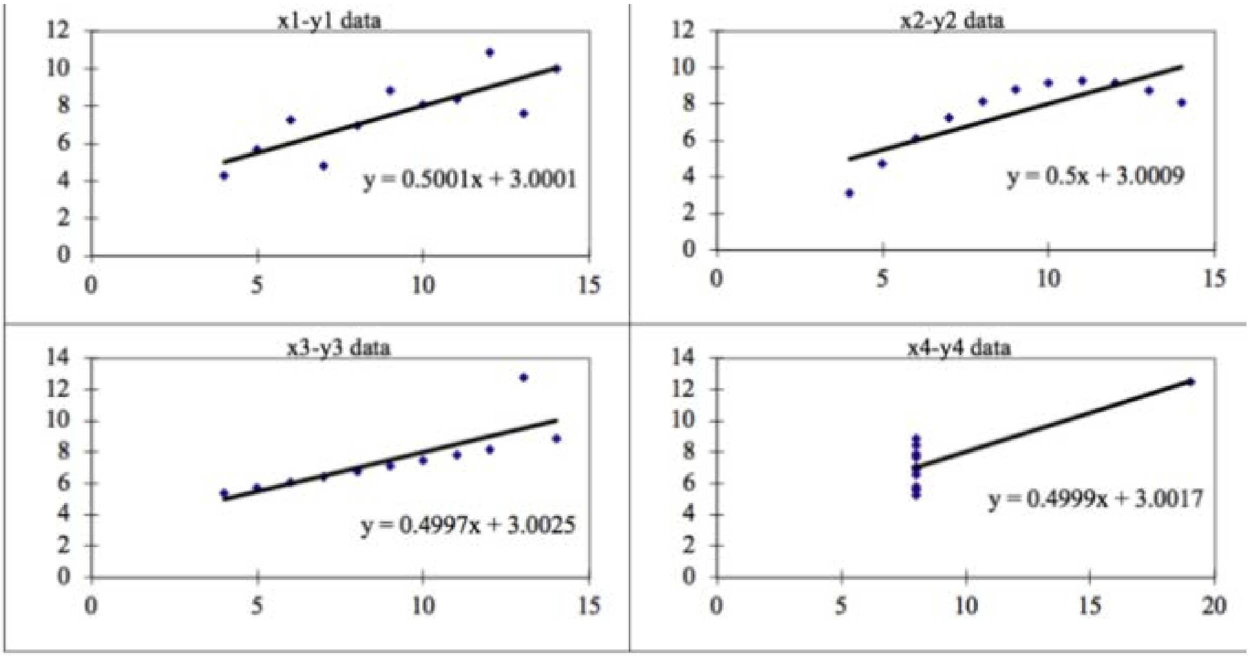
Variable selection-Appropriate variables need to be selected manually or by automated methods

Evaluate R-squared – Do it on training set first, then test set and see if they are compareable.

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

Ans: According to Anscombe's Quartet, a group of four data sets which might look identical in simple descriptive statistics, have some features that interrupt the model building if not taken care of. They look very different when plotted on scatter plots but if you have just an overview, they might look same.

Hence, all the important features in the dataset must be visualized atleast once before creating a machine learning algo on them. This will help us make a good fit model.



The data for these models looks almost the same when not analyzed

Here is when Anscombe's Quartet comes into play

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

Ans: According to Pearson correlation ,we assigns a value between − 1 and 1 to a numeric variable, where 0 means no correlation, 1 means total positive correlation, and − 1 means total negative correlation. This is interpreted as follows: a correlation value of 0.8 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will increase, but if the value of the correlation is negative, then if A increases, B will decrease.

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

Ans: Whenever we look at a dataset, it has all kind of values from categorical to numeric. In numeric also they may range from negative to thousands and lakhs.

Scaling is a method of bringing all values at one scale for easy understanding and comparison.

In Normalized scaling (also called min max scaling),we bring the values between 0 and 1 by applying the given formula –

MinMax Scaling x= (x-min(x))/(max(x)-min(x))

In Standardized Scaling, we replace the value by z-score so that the mean of the values becomes 0. Formula is as follows-

Standardized x= (x-mean(x))/sd(x)

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

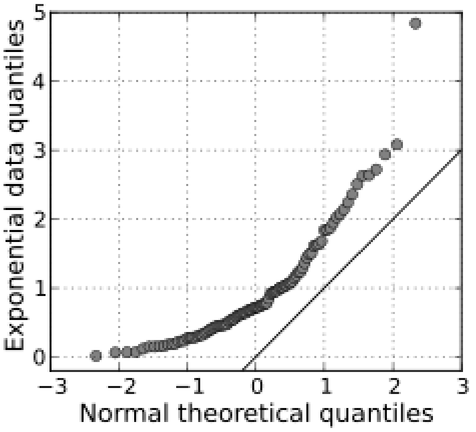
Ans: VIF=1/(1-R^2)

So if VIF is infinite, then 1-R^2=0 which means R^2=1 .

If we remove the variable which is perfectly collinear with the this variable then the value will change

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

Ans:Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line**.**


Exponential data quantiles
Normal theoretical quantiles