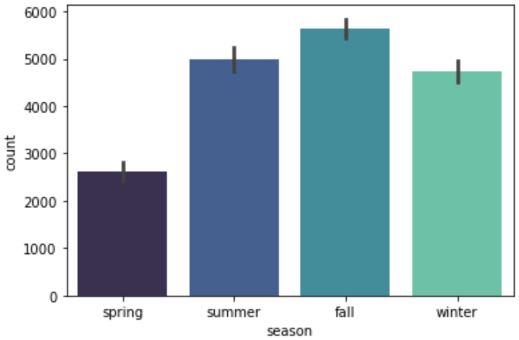
Linear Regression Assignment Subjective answer_Thao Tran Chi

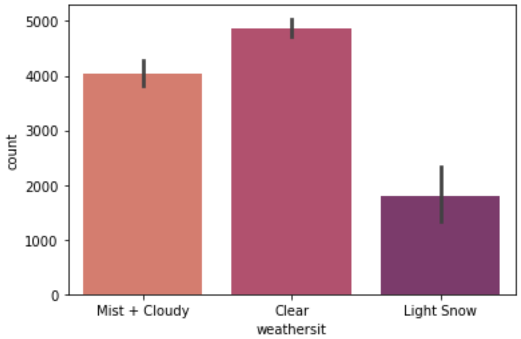
Assignment-based Subjective Questions

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

- Bike rentals in `spring` is the less than other seasons.

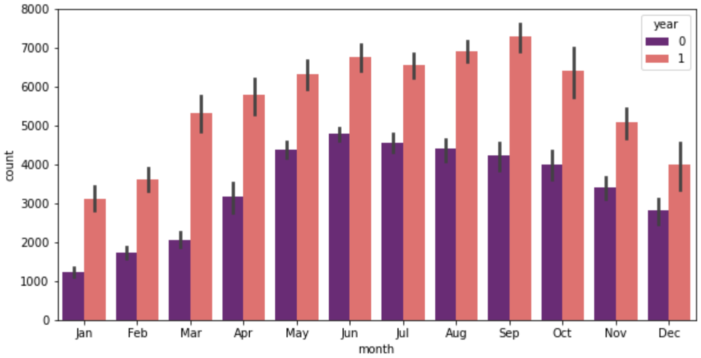


- The bike rental demand is high when weather is Clear, Few clouds, Partly cloudy, Partly cloudy, however demand is less in case of Lightsnow and light rainfall. We do not have any dat for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog , so we can not derive any conclusion.

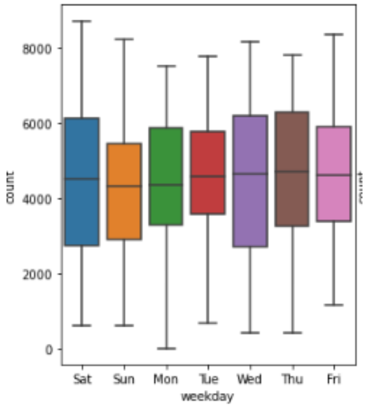


- The demand bike rental increased in the year 2019 when compared with year 2018.

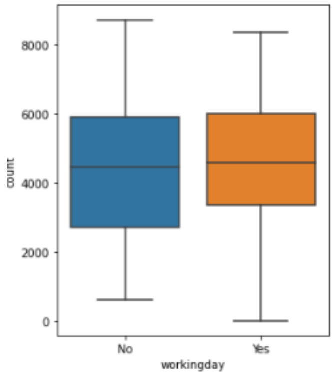
Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.



- The demand of bike is almost similar throughout the weekdays.

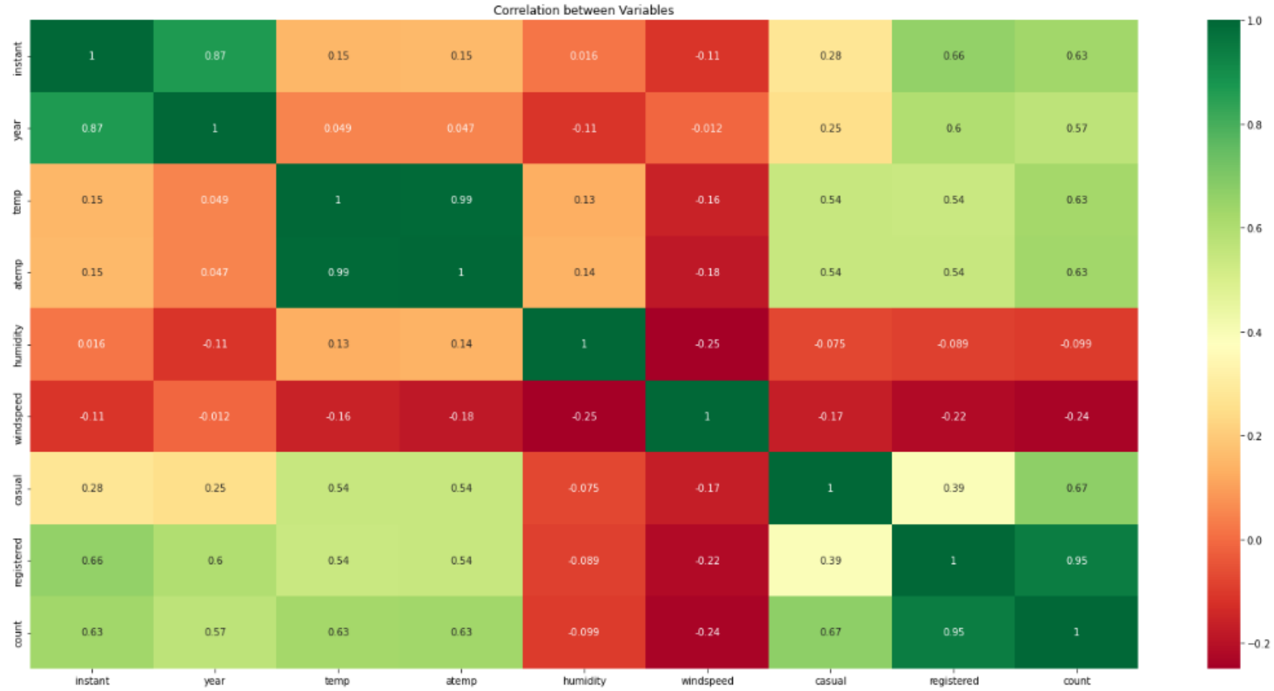


- There is no significant change in bike demand with working day and non working day.



2. Why is it important to use drop_first=True during dummy variable creation?
   - drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'count'.

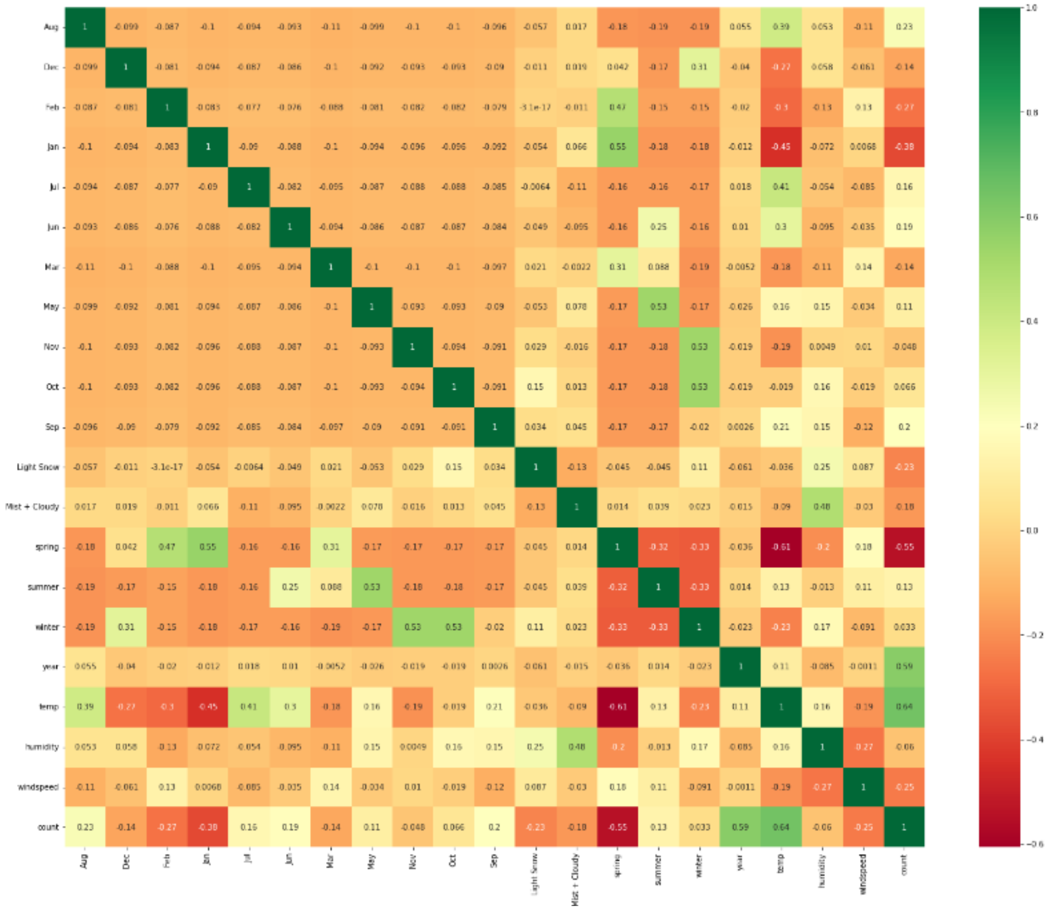


4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   The simple way to determine if this assumption is met or not is by creating a scatter plot x vs y. If the data points fall on a straight line in the graph, there is a linear relationship between the dependent and the independent variables, and the assumption holds.

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

The top 3 features contributing significantly towards the demands of share bikes are:

- season (‘spring’ Negative correlation).
- year (‘2019’ Positive correlation).
- temp (Positive correlation).



General Subjective Questions

1. Explain the linear regression algorithm in detail

The standard equation of the regression line is given by the following expression:

$$y = \beta_0 + \beta_1 x$$

Where:

y *is outcome variables or depedent variable.*

β_0 *is the intercept, the default value when* x *equals* 0.

β_1 *is the slope of the line.*

x *is the depedent variable.*

The general form of linear regression is

$$y = \beta_0 + \sum_1^n \beta_i x_i$$

In this equation, the change of y is defined by the any changes of xi given the slope of alpha i.

Linear regression is the most basic machine learning algorithm where we train a model to predict the outcome of continuous variable (depedent variable) of our data based on some other independent variables.

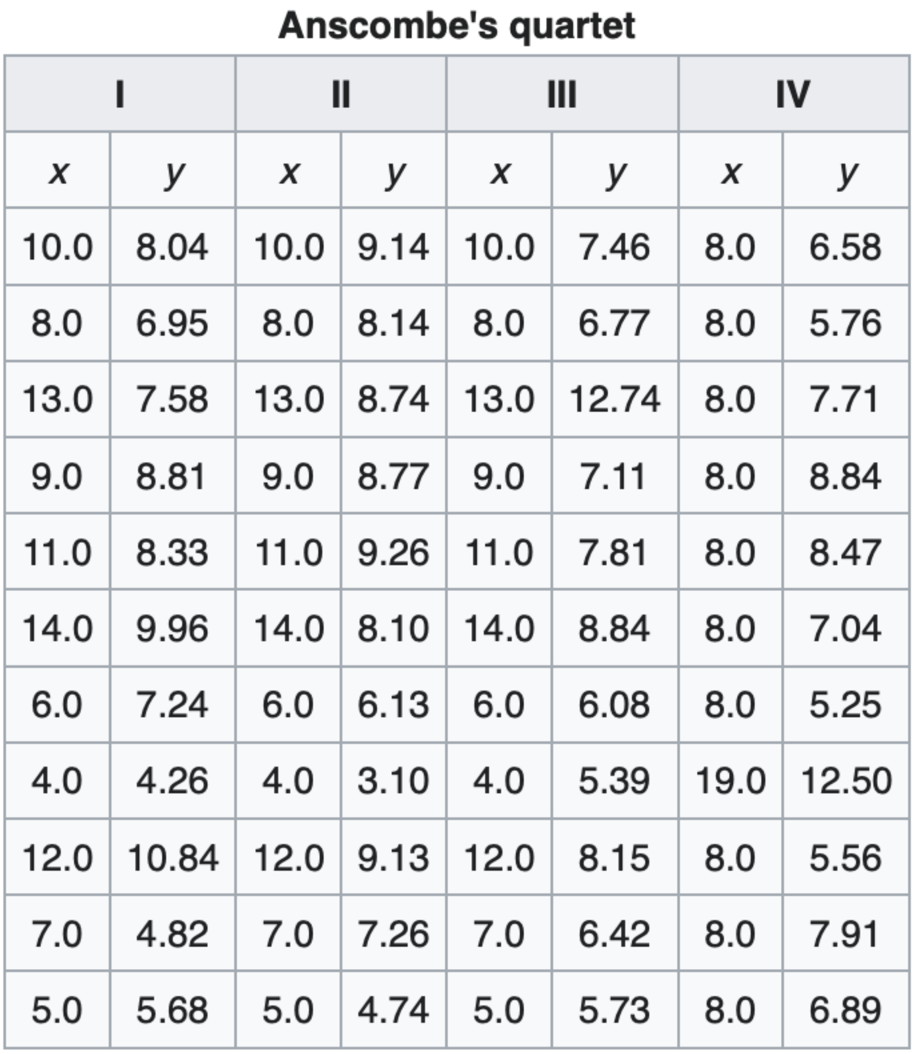
For example, in the linear regression model assignment, we need to predict the demand of bike shares (unknown data) using data such as date, season, number of registered customers, other weather conditions (known data). Some changes in known data doesn’t reflect in the changes of other known data. That’s why we call them independent variables. For example, changes in season doesn’t result in changes in working day or holiday. In contrast, changes in independent variables (such as season, weather condition) may lead to a change in number of bike sharing demands.

2. Explain the Anscombe’s quartet in detail

Anscombe’s quartet are four datasets that almost identically equal in statisticals summary (such as mean, standard deviation, correlation, etc), yet differently visualise on graphs.

In another words, if we visualise them on charts, we will see completely different 4 charts, but when we examine statistical summary (mean, standard deviation, correlation), we have identical result.

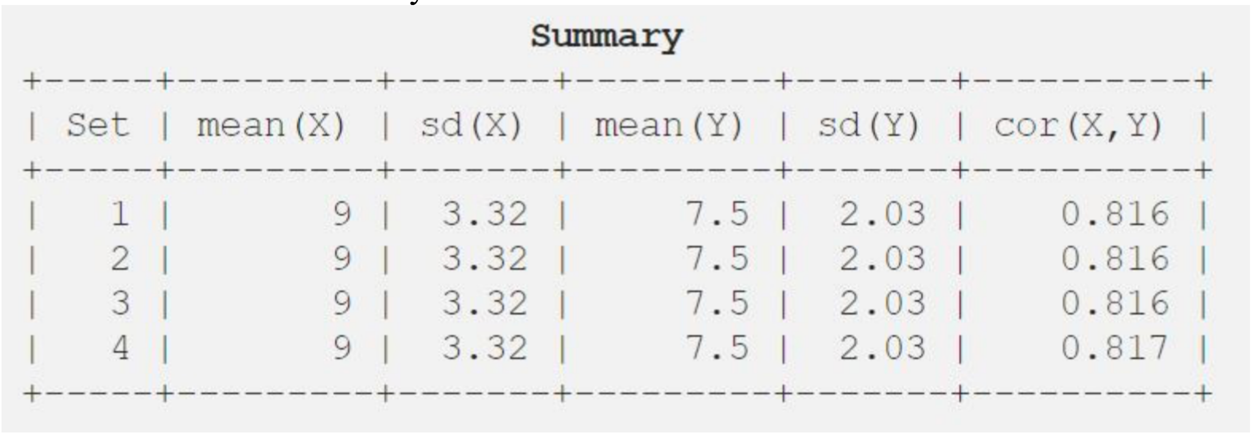
Following example is copied from wikipedia. The table below shows the example of Anscombe’s quartet.



Anscombe's quartet

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

Here is the statistical summary.



```
                         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 |
+-----+---------+-------+---------+-------+----------+
```

As we can see, all datasets have idential mean, standard deviation, and correlation.

3. What is Pearson’s R

Pearson's R represents the linear correlation betweens variables. There are three possible scenario of correlation. Possitive correlation is when variables tend to go up and down together. Whereas if variables go in different direction it is negative correlation. Zero correlation means there is no linear correlate at all.

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

During steps of data processng, sometimes we need to scale independent variables. It normalised the data within a certain range. Besides, it’s usefl to boost calculations in a algorithm.

Scaling data is essential important, because data collection sometimes contains features highly varying in weights, units and range. Scaling data helps to get all the variables to the same level of scale. If this is not done properly, algorithm only considers magnitude in account and not units so it leads to inapproriate result.

Normalisation transform all of the data in the range of 0 and 1, in which 0 equals the minimum values and 1 are the maximum values in the original data. On the other hand, standardisation replaces original values with their Z-score. The new dataset has mean zero and standard deviation one.

Imagine, the original dataset may varry a lot in range. The standardisation compresses it a bit, but it’s not bounded to a certain range. While the normalisation compresses it way better in range of 0 to 1. However, the normalisation shouldn’t be applied if there is outliers as a matter of fact, the extrem outliers are always 1.

Normalisation is usually used when we don’t know about the distribution while standardisation is useful for normal distribution.

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

If there is a perfect correlation amongs variables, the R^2 is equal 1. The VIF = $\frac{1}{1-R^2}$ so if $R^2 = 1$ the VIF returns infinity. In my asssignment, it happens when I didn’t drop first columns of months, wheathersit during dummy variables creation.

VIF is useful to detect multicollinearity, and without dropping first columns, strong multicollinearity happens. After dropping first columns during dummy variables creation, infinity VIF vanished.

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

   Quantile-Quantile (Q-Q) plot, is useful to detect if a dataset is normal distribution. Besides, it also helps to determine if two dataset comes from same population with a common distribution.

   In linear regression model, Q-Q plots is useful for model evaluation in term of residual analysis. Residual is the difference between actual values and predicted values. Given train and test dataset, if we can conclude that both set of residuals come from the same population, we are pretty sure that our model perform on test dataset as well as on train dataset.