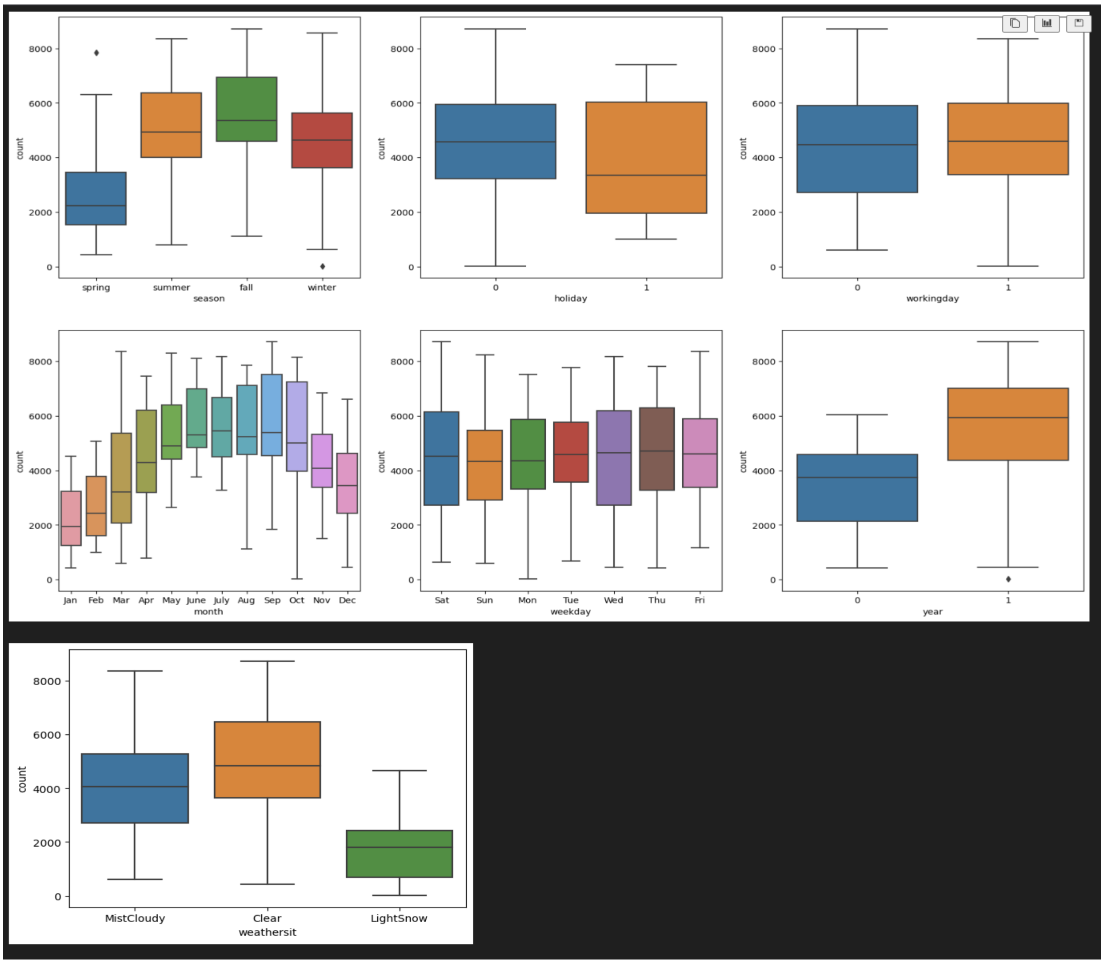
Assignment-based Subjective Questions

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

- the demand is more in the fall and summer seasons. Owing to the pleasant weather
- the demand curve against the months closely follows seasons since the months are a subset of seasons. So the high demand in May to October
- demand is more on holidays because people do recreational activities like bicycling more on holidays
- workingdays dont infuence the demand much
- weekdays also dont influence the demand much
- demand goes high on days with nice weather. Snowy days dont have much takers

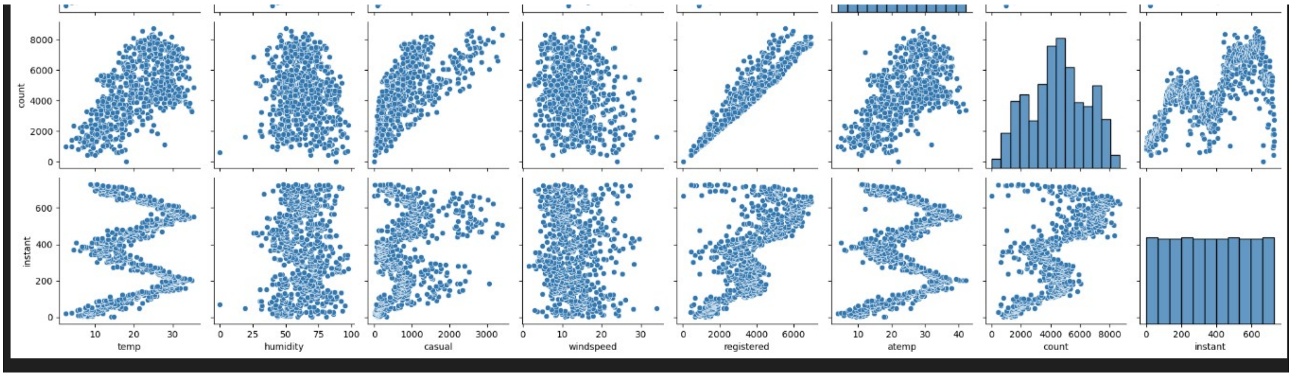


2. Why is it important to use drop_first=True during dummy variable creation?

Only n-1 levels are necessary to represent n number of states since the n-1 th state is implied. This helps reduce an extra column and also cuts the chance of corelation between the new dummy variables.

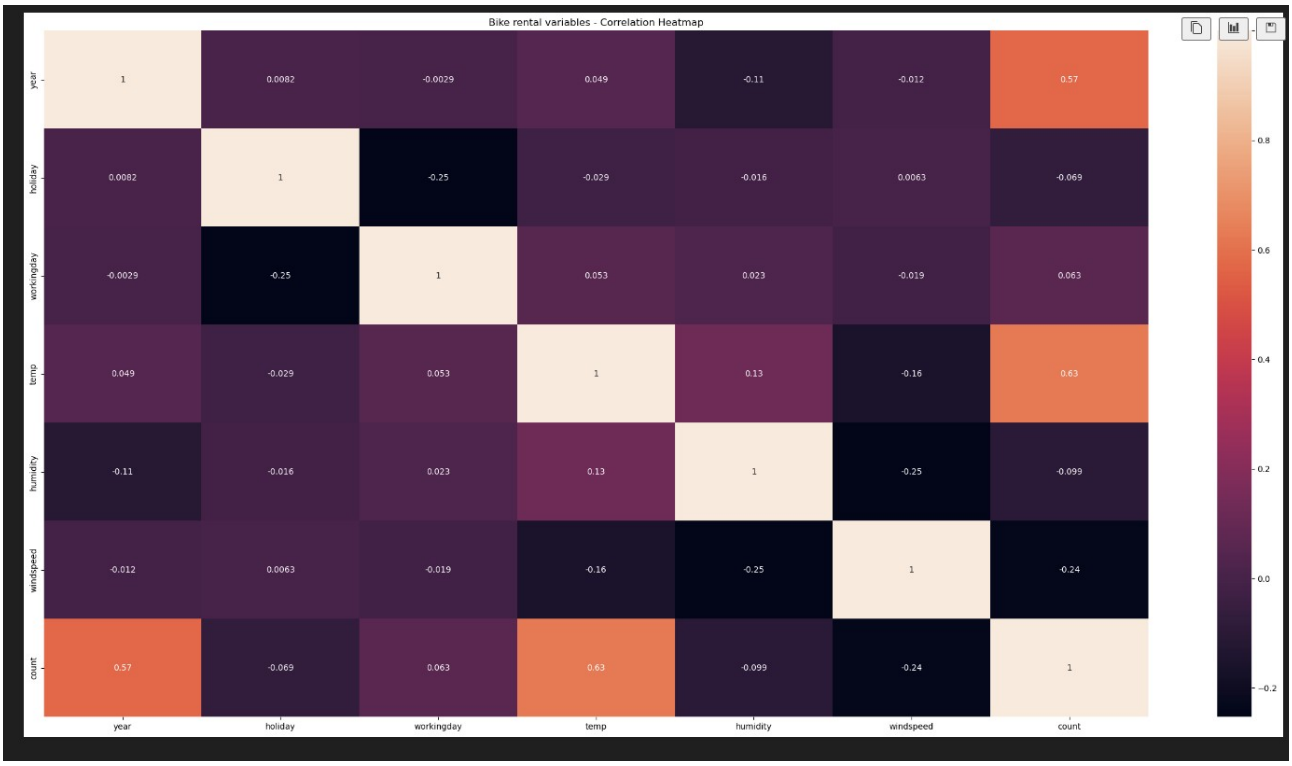
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Count and temp have a good correlation. Likewise, the atemp and temp have a strong correlation, which is why the apparent temperature field was dropped



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

- The VIF values are <5 implying that they are not multicollinear
- Normally distributed error terms
- linear relationship of variables
- close to zero multicollinearity in feature variables



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

- temperature
- windspeed
- year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression algorithm tries to establish a relationship between a few independant variables and a dependent variable. This is done by fitting a straight line in a scatterplot of independent vs dependent variable. The characteristics of the fitted line, like the slope and intercept(beta 0 and beta 1) are used to predict the dependent variable for the given independent variables.

There are univariate and Multivariate Linear regressions involving single or multiple independent variables, respectively.

The strength of the linear regression model is measured through

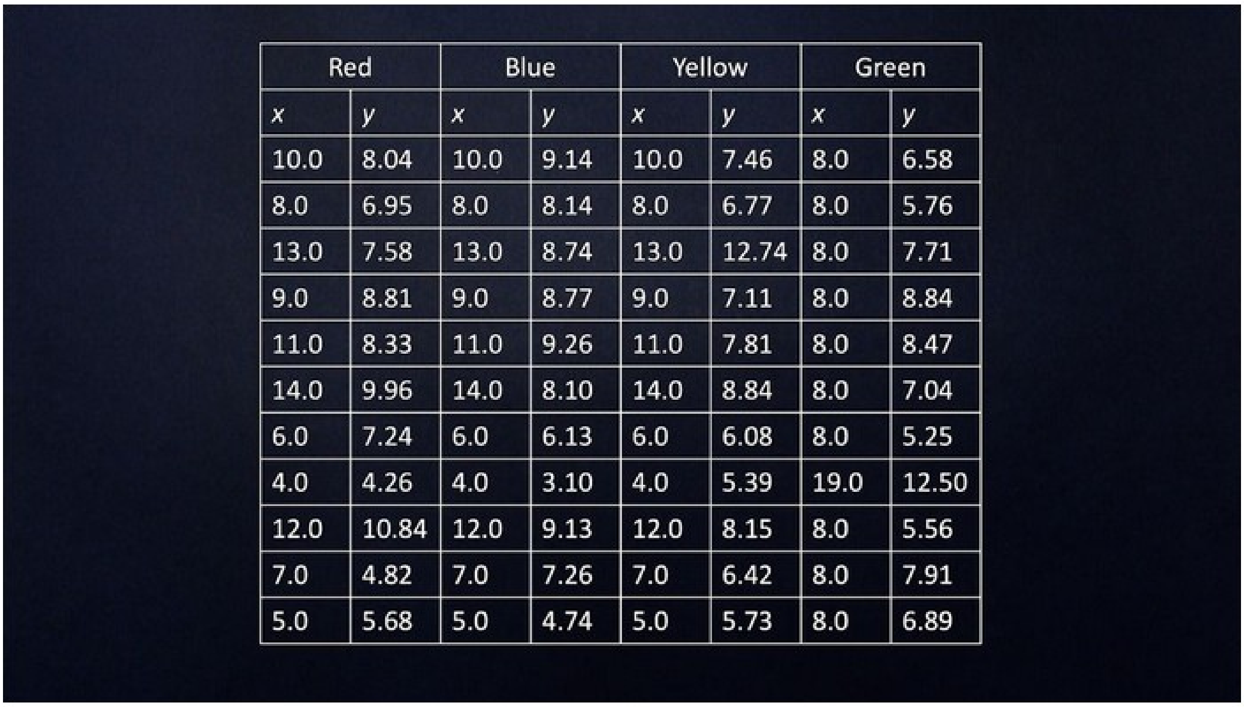
1. R2 or Coefficient of Determination
2. Residual Standard Error (RSE)

2. Explain the Anscombe's quartet in detail.

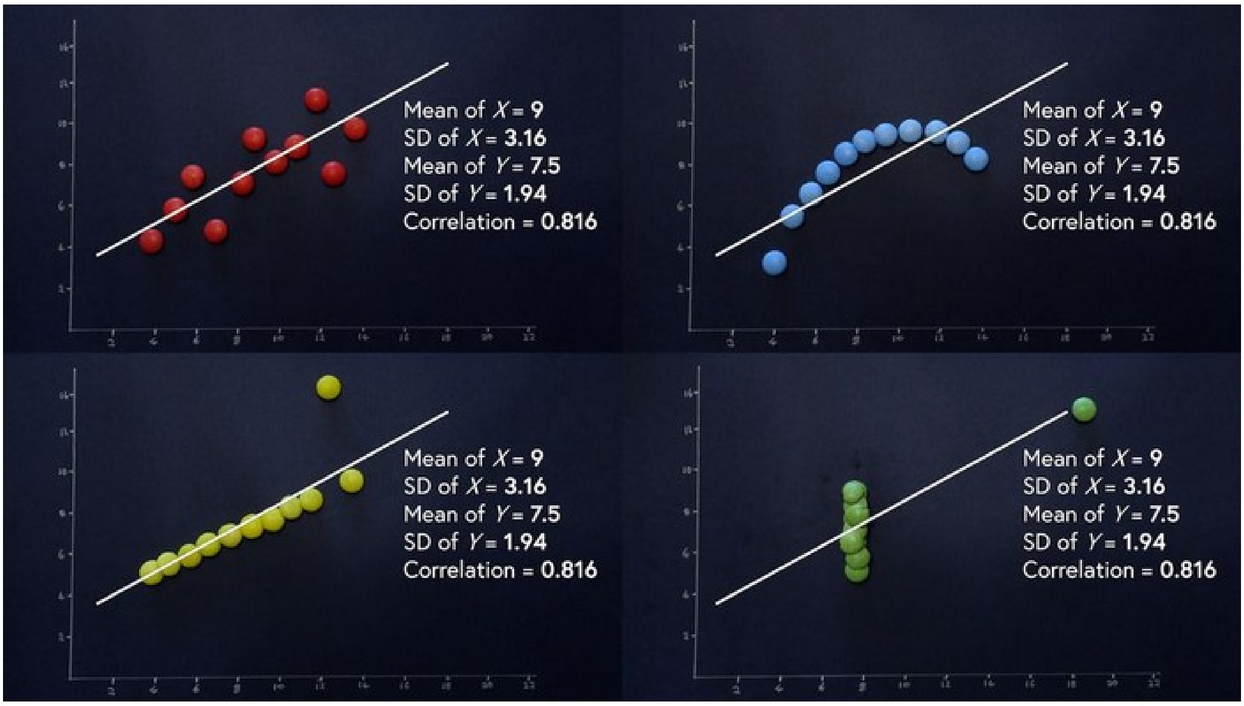
Anscombe's Quartet shows how four entirely different data sets can be reduced down to the same summary metrics.

This proves that summary stats can be misleading and we need to visualize and scrutinize data before we move on to build modes.

For this below example set of data



Red		Blue		Yellow		Green	
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



We find that the Mean, SD and correlation are the same despite the data being very differnt.

3. What is Pearson’s R?

Pearson’s R is a value between -1 and +1 which is a way of measuring the strength and direction of linear correlation between 2 continuous variables.

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

Pearson correlation coefficient (*r*) value	**Strength**	**Direction**
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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

Scaling reduces the values of the continuous feature variables in a model so that the values are normalized. This is necessary for the ease of interpretation of the model. There are two kinds of scaling viz.,

- Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

- Standardization is a scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

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

When VIF is infinite, it means that there is a perfect correlation between two predictor variables. Which means that the R squared value would be 1 between those variables.

This generally cannot happen in real life and our model will be considered over fit.

This can be addressed by dropping one of the features that have infinite VIF

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

QQ plot is used to find if the datasets came from the same generative theoretical distribution like normal or exponential. It is a subjective visual check to see if our assumptions are plausible. When two sets of quantiles are plotted against each other, if they are from the same dataset, the points form a straight line. If they diverge, they are from a different datset.