Answers for subjective questions based on Boombikes case study.

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

- From season plot we can see that fall has the highest demand for boombikes and 2019 was the year with highest count with median of 2019 almost equal to the highest count of 2018.

- Month wise also there is increase from May till September which is basically a summer season and then decreasing, this means in general summer season has an increase in demand in USA.

- Holidays show lower count as compared to when it’s not holiday.

- There is so such significant difference in weekday’s demands and also on working days.

- Clear weather shows highest demand in bikes which justifies that it’s easy to drive during clear weather.

- Also you can see that there is no data for 4th category of weather i.e. "Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog".

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 dropping that extra variable which is created and of no use e.g. the get_dummies() creates the dummy variables equal to the count of categories present in that particular variable and we need 1 column less as it is of no use because if all other variables are 0 then but obvious the last variable will be 1.

- Also 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?**

- temp (temp) feature had the highest correlation with the target variable totalcount (cnt).

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

- I first build the model with all the variable in it and decided to drop variables one by one by validating 2 conditions.

1. Significance (p-value) → Drop features having significance values more than **0.05**.

2. VIF (variation inflation factor) → Drop features having VIF values more than **5**.

- Drop values: High p-value, High VIF.

- Keep values: Low p-value, Low VIF.

- Take a decision on:

    - High p-value, Low VIF → Drop them first.

    - Low p-value, High VIF → Drop them after dropping on the basis of p-value.

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

– top 3 features:

1. Year

2. Light Snow

3. Spring

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

- Linear regression is a machine learning algorithm and is used in predictive analysis of continuous variables such as score, sales, price etc.

- Types of Linear regression:

1. Simple linear regression: here single independent variable is used to predict target variable.

2. Multiple linear regression: here more than 1 independent variables are used to predict target variable.

- It shows a relationship between one dependent/output/target variable(y) and one or more independent/predictor variables(X) hence it is called linear regression.

- It predicts the target variable using predictor variable.

- Equation for a linear regression can be written as (**y = ßX + c**) where ß is coefficient (slope) and c is the intercept (constant). There are 2 types of relationships:

1. Positive linear relationship (+ß).

2. Negative linear relationship (-ß).

- This regression line is the best fit for building our model where we have training data where we have to train our model with the training data and that model should predict values of y so the model build should be a perfect fit.

- The model gets the best fit line by finding the best ß and c.

- The different values of coefficient's gives different line of regression and the cost function is used to get the values of coefficient for the best fit line.

- The cost function(J) in linear regression is Root mean squared error between predicted y value and true y value.

$$\mathrm{MSE}= 1\frac{1}{N}\sum_{i=1}^{n}(y_i \;-\; (a_1x_i + a_0))^2$$

Where

n = total number of observations.

yi = actual value of i.

$(a1x_i+a_0)$ = predicted value of y.

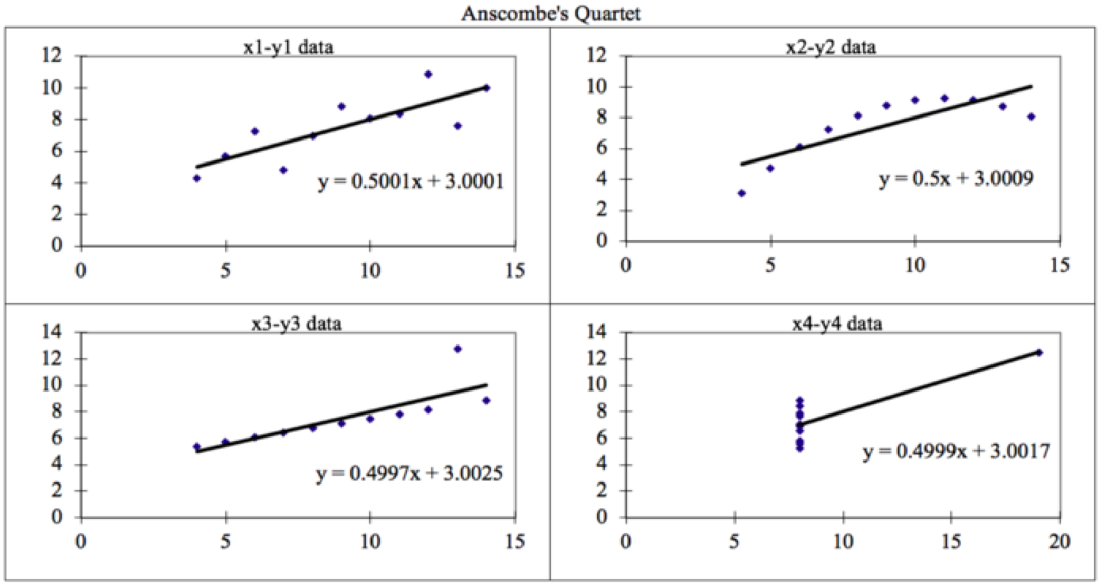
- Residuals: The difference between actual value and predicted value is called residual.

2. **Explain the Anscombe's quartet in detail.**

- Anscombe's quartet is a group of 4 data sets which has identical simple statistical properties but when plotted on graph or when a model is built it shows true insights and is different when viewed without a graph.

- It was constructed by a statistician named francis anscombe in 1973 to demonstrate the importance of visualizing the data.

- The statistics data has same mean, standard deviation and correlation between X and y but in graph whne plotted they appear to have this output.



- You can see that in data set 1 there is a linear relationship and the line fits properly, in data set 2 there is no linear relationship, in data set 3 there is a perfect linear relation except one outlier and in data set 4 one high outlier without any linear relation.

3. **What is Pearson's R?**

- Pearson correlation coefficient(R) measures the strength of linear relationship between 2 variables and it always lies between 1 and -1.

- It can be positive or negative or no relationship so r can be R=1 or R=-1 or R=0 respectively.

- It is one of many correlation coefficients to choose from when you want to measure a correlation.

- The formula to calculate Pearson R is:

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

Where x and y are variables and n is sample size.

- You can also use Pearson correlation coefficient to check the significance of relationship between 2 variables by using hypothesis testing.

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

- Scaling is a method of transforming the data present in features into a specific range the range can be 0-100 or 0-1.

- Scaling is performed because suppose the different features is having different units and data can be of different magnitude so for machine learning models to interpret this data faster and train the model better on same scale we are using scaling. So if the data points are far away from each other than scaling brings them closer and thus helping the machine learning models to train better.

- Algorithms that calculate the distance between features are affected more by larger values if data is not scaled.

Normalisation: (x - xmin)/(xmax - xmin)

- Normalized scaling also known as Min-Max scaling is a technique where the data points in a particular feature is scaled between a fixed range of 0 and 1. Also normalization works better where there are no outliers.

Standardisation: (x - mean)/sigma

- Standardized scaling also known as z-score normalization is a technique where we try to make data points centered around the mean of all the data points that are present in the feature with a standard deviation of 1. This means the mean of the data points will be 0 and standard deviation will be 1. This technique also tries to scales the data points between 0 and 1 but we don't use max or min.

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

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Now here if R-squared becomes 1 then VIF becomes infinity which shows a perfect correlation between 2 independent variables so when R-squared value is 1 which indicates that all the variance in dependent variable is explained by independent variable i.e. when residual sum of squares (RSS) becomes 0 and this happens when there is no difference between actual and the predicted values of target variable.

- It also means that the corresponding feature is explained perfectly by linear combinations of other features where the VIF becomes infinity.

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

- Q-Q plot also known as quantile-quantile plot is graphical method for comparing 2 probability distributions by plotting their quantiles against each other where the quantiles are calculated.

- Quantile means where your data is divided into equal sized, subgroups. It defines a particular part of a data set i.e. a quantile determines how many values are above or below a data set.

- Then we see how the points fit the straight line in that graph and if the data is normally distributed then most of the points will be on the line then you can check the same with other distributions also.

- A Q-Q plot is used to compare the shape of distributions.

- Q-Q plot can be used to check if the data is normally distributed or not and also which distribution matches your data.