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?

1. Bikes are most used in clear weather and mist than light snow
2. Spring is the season with least bikes usage.
3. Jan,Feb, Nov,Dec have least usage due to chilled weather.
4. Year 2019 had more sales than 2018.
5. Non-holidays have more sales than holiday days.

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

1. If we do not use drop_first then we might face correlations among the varibales.
2. It would also easily be possible to get data understanding when only 2 varibales.
3. if we have n varibales then n-1 varibales are sufficient to do predictions.
4. when all varibales are 0 then it indicates the final n.

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

1. atemp and temp

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

1. We do so by plotting a histogram of the residuals and if that shows mean at zero it indicates a good model.
2. We also draw the residuals and see if there is any pattern followed or are they evenly distributed,
3. Y_pred v/s Yactuals is also plotted and the lineraility is seen.

4. Error terms must be nearly constant

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

1. temp, yr and weathersit_Mist-type 3

General Subjective Question

1. Explain the linear regression algorithm in detail.

The aim is to predict the value of a continuous varibale. It could be a simple linera regression where we have just one dependent and one independednt varibale. It could also be multiple linera regression where we could have multiple independednt varibales and one dependednt varibale.

it falls under supervised learning as the lablels are known.

The formulas used to verify the valididty of the model include below, Higher R-sqaure is better. Higher Adjusted R-square is better R-square must be close to adjusted r-square. Varibales with p-values very close to zero and having positive co-efficients would bring positive corelations. Varibales with p-values very close to zero and having negative co-efficients would bring positive corelations.

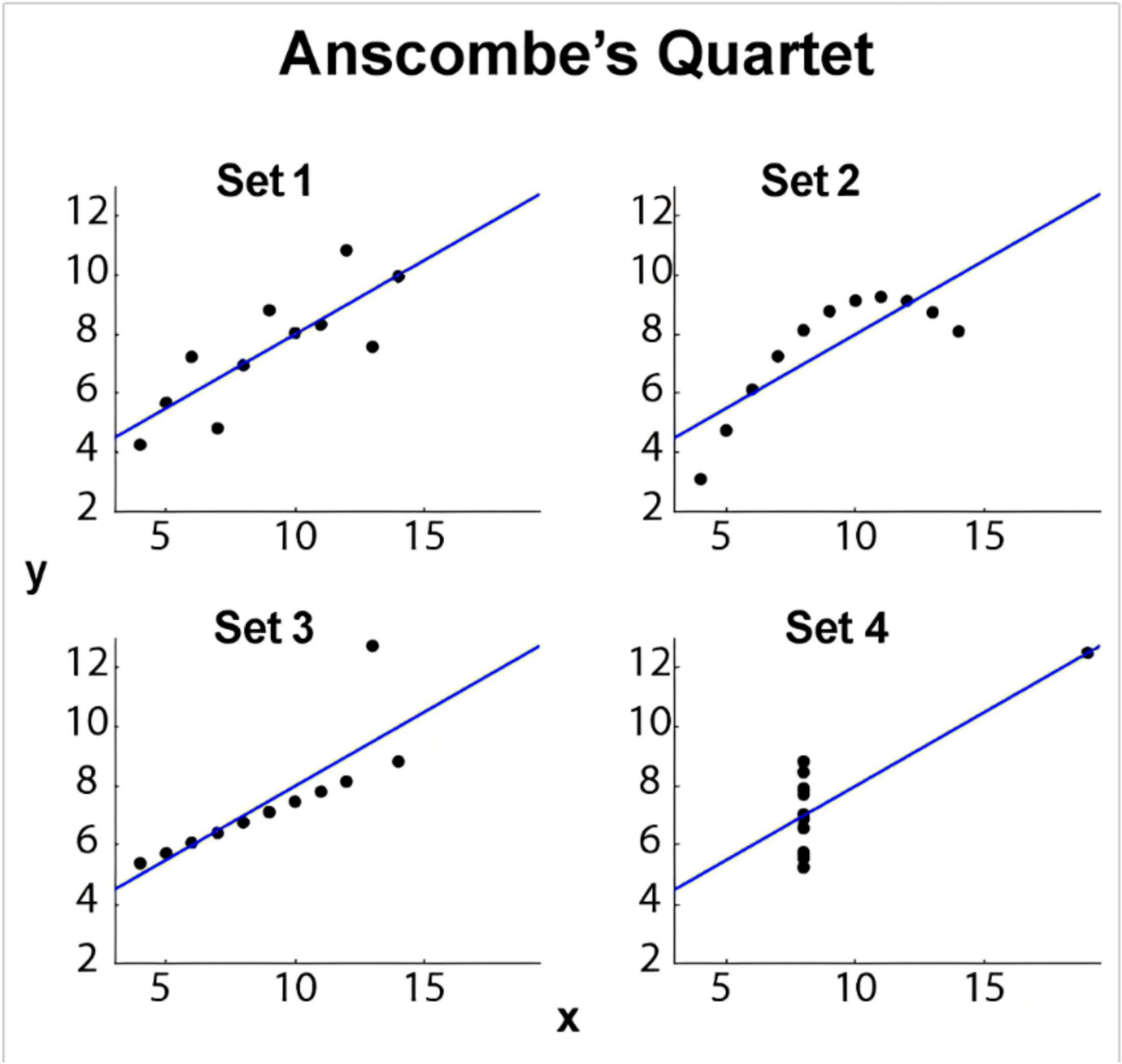
Error terms must be normally distributed MSE would also tell model strength.

In case of multiple linera regression VIF can be used to reduce multicolinear varibales

2. Explain the Anscombe’s quartet in detail

1. It is a collection of 4 datasets
2. The main aim is to find out if the linera model can be applied on different datasets.
3. Linear model best fits dataset-1.
4. Dataset-4 has a lot of oultliers hence linerar model will not fit. In such case we will have to use SVM or polynomial regressions.

Image shown below for reference,



3. What is Pearson's R

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Pearson correlation coefficient (r) Correlation type Interpretation

Example Between 0 and 1 Positive correlation When one variable changes, the other variable changes in the same direction. Baby length & weight: The longer the baby, the heavier their weight.

0 No correlation, There is no relationship between the variables. Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.

Between 0 and –1 Negative correlation When one variable changes, the other variable changes in the opposite direction. Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

- Scaliing is way to represent all the varibales at the same scale. Eg- 100 can be divided by 10 and be represented as 10 and 10 itself as 1. This helps us to reduce the time taken for the gradient descent process.
- Scaled varibales also help us in evaluating co-efficants else one co-efficient might look very huge but impact actually might be lower than expected.
- Mi max scaling reduces everything between 0 and 1 . Standard scaling takes care of outliers better than min max.

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

This means that varibale is highly dependeent on other varibales and can be perfectly predicted

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

The purpose of Q Q plots is to find out if two sets of data come from the same distribution

Double-click (or enter) to edit

Colab paid products - Cancel contracts here