Assignment-based Subjective Questions

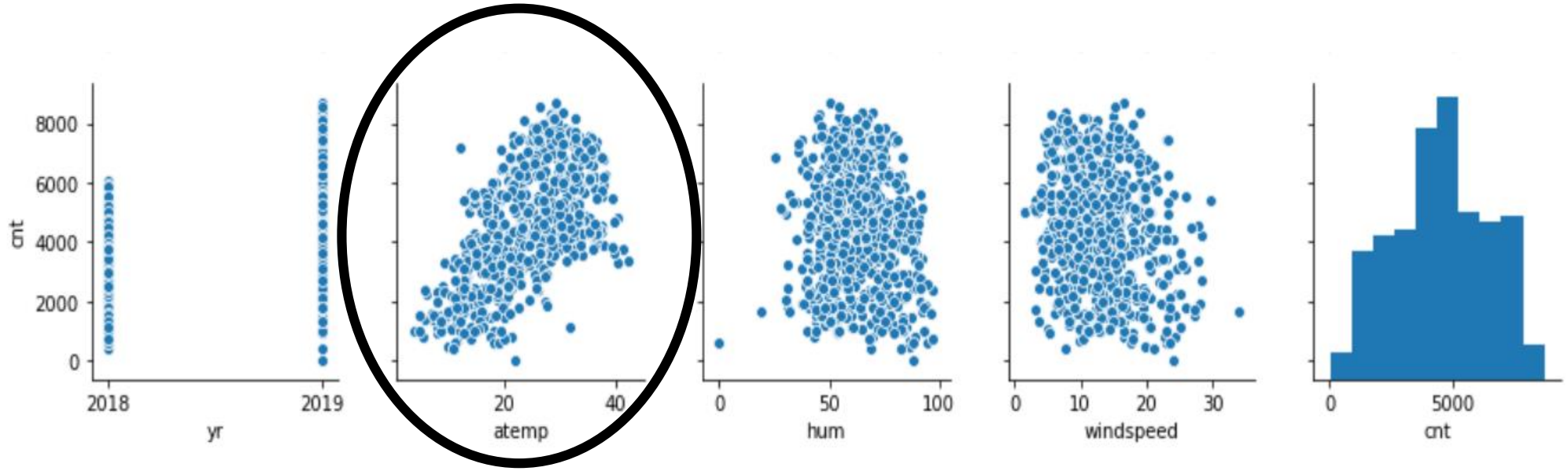
2. Why is it important to use drop_first=True during dummy variable creation?

It is depends on the model. If we don't drop the first column then the dummy variables will be correlated (redundant as Dimitra shows in the post below). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted.

Assignment-based Subjective Questions

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

atemp plot is looking the highest correlation with the target variable because the values is increasing as comparing to the “cnt” variable values.



Assignment-based Subjective Questions

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

- The process consists of 4 simple steps:
  - Identify and categorize the assumptions made about initiative.
  - Vote for the assumptions that agree and apply to the initiative.
  - Rate each assumption based on its impact and confidence level.
  - Discuss results, view alignment, and finalize an action plan to validate the assumptions.

Assignment-based Subjective Questions

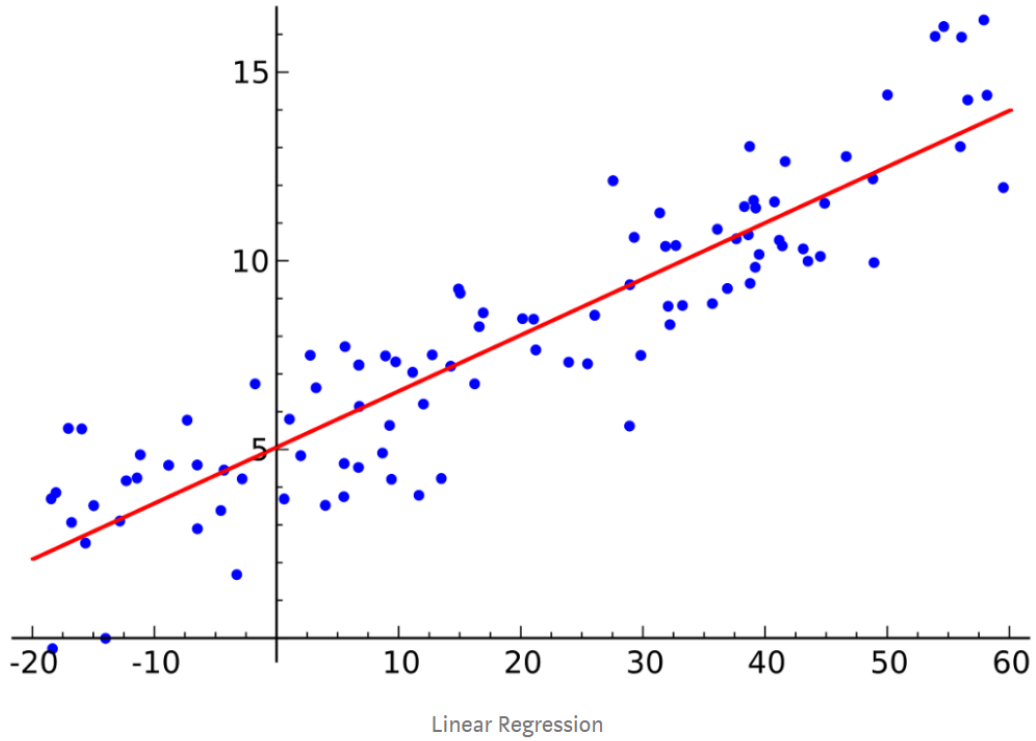
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- The line of best fit states the demand for shared bikes is affected by the following equation:
- cnt= 0.247*yr+0.055*workingday-0.21*windspeed-0.178*spring-0.121*dec-0.098*feb-0.167*jan-0.097*nov+0.072*sep+0.063*saturday-0.313*light_drizzle-0.09*misty

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



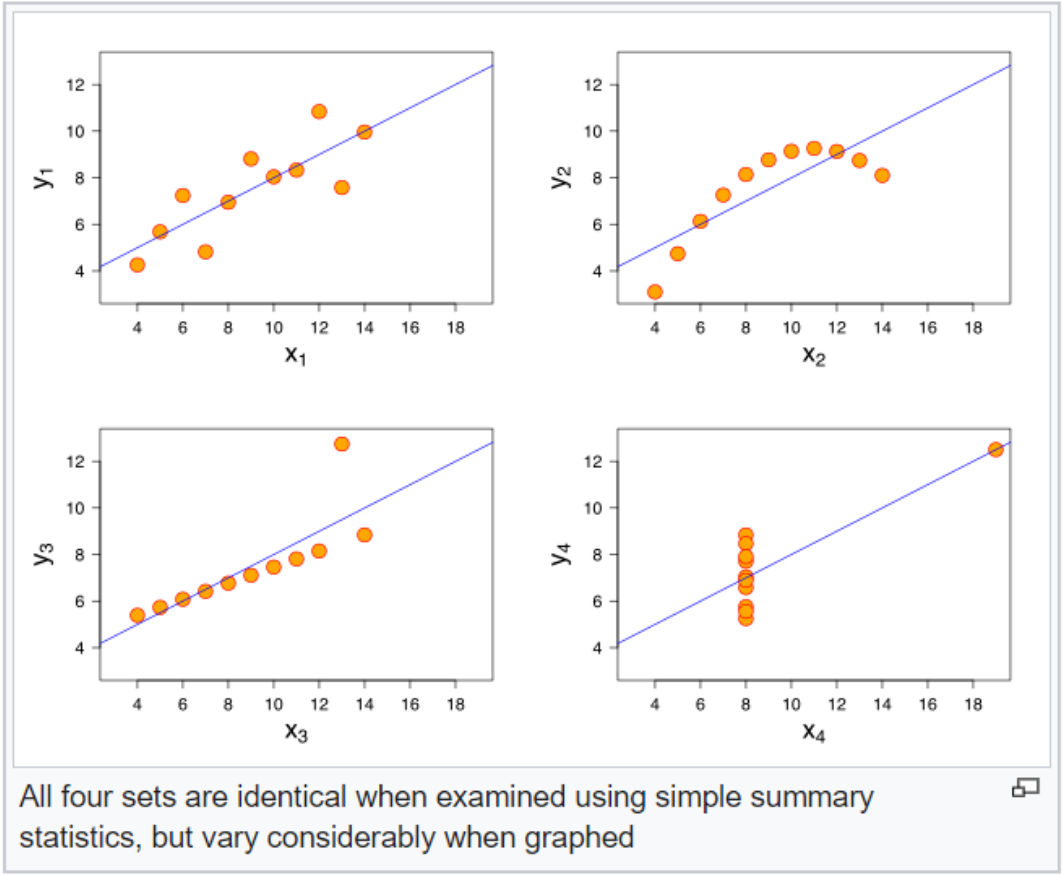
Linear Regression

General Subjective Questions

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

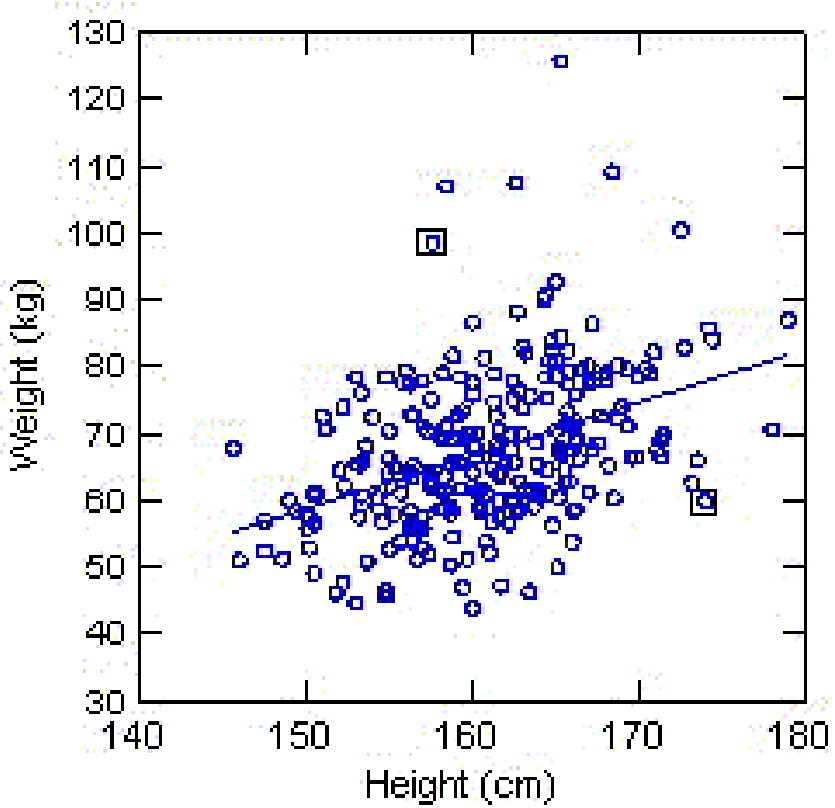


All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

General Subjective Questions

3. What is Pearson’s R ?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.



General Subjective Questions

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

A. scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

B. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

C. 1. Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

C. 2. Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

General Subjective Questions

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

General Subjective Questions

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

A. In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Another way the Q-Q plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.

B. Q-Q Plot answer the following statements –
   i. If 02 data sets from populations with a common distribution
   ii. If the distribution have a similar shape
   iii. If they have common location and scale
   iv. If they have similar tail behavior