Course 1 Task 2 Analysis - Keegan Gelvoria

* Correlations
  + Positive correlation between region and amount purchased
  + Negative correlation between age and amount
    - Negative covariance between age and amount
    - As age increases amount purchased decreases
      * See graph of average purchase amount by age
* Are there differences in the age of customers between regions?
  + Average age by region displays that Region 4 had the youngest average age of consumer and Region 2 had the highest average age of consumer
* Can we predict the age of a customer in a region based on the demographic data?
  + After discretizing age into decades and fitting into three different algorithms, the accuracy of predicting age based on demographic data was very low
  + Algorithms used
    - Decision Tree, Random Forest, and Gradient Boosting
    - Cross Validation scores of 0.22 which were all low
  + With low evaluation scores, we cannot predict age of a customer using any of those algorithms
* Can we predict the amount of purchase based on the demographic data?
  + After discretizing purchase amount into three buckets and then fitting into three different algorithms, the accuracy (0.71) of predicting amount was decent enough to utilize one of the algorithms for prediction
  + Algorithms used
    - Decision Tree, Random Forest, and Gradient Boosting
    - Cross validation scores of 0.71
    - Utilized the decision tree algorithm as it was the simplest to use
    - Prediction accuracy was 71% using the Decision Tree
  + No valuable insight from this predictive model
* Can we predict if the purchase was made in-store or not?
  + Fitting the data into the three different algorithms showed the highest accuracy of 0.84
  + Algorithms used
    - Decision Tree, Random Forest, and Gradient Boosting
    - Cross validation scores of 0.88
    - Utilized the decision tree algorithm for ease of use
    - Prediction accuracy was 84% using the decision tree model
  + All purchases in region 1 were done in-store
  + For region 3 and 4, purchase amounts of 2000 or less were mainly done in-store
    - Higher purchase amounts were done online
    - Highlighted in the scatter plot of region and amount, based on in-store
* Martin’s hypotheses
  + Customers who shop in-store are older than customers who shop online
    - Average age of in-store shoppers are younger than average age of shoppers online
    - This hypothesis is refuted
  + Older customers spend more money (on electronics)
    - Graph of average purchase amount by age refutes this hypothesis as average amount decreases as age increases
  + Do any other factors predict if a customer will buy online or in our stores?
    - Region had a slightly stronger correlation to in-store than when compared to amount but weaker correlation than age
  + Is there a correlation between number of items purchased and amount spent?
    - Number of items did not have any strong correlations to any other variable
    - Number of items and amount spent was the second weakest correlation among possible relations with number of items
  + Strategic options
    - Market the in-store designs for younger audiences
    - Market online designs for older shoppers