Homework 1

Graedon Beeler

Introduction

A clothing company would like to be able to predict how much their customers will spend at their store in a given year. To determine the answer to this question, they have tracked the following variables that they believe play a factor in an accurate prediction: gender, age, height, waist size, inseam, membership of a special coupon group, salary, months in the store's rewards program, total number of purchases, year, and of course, the past annual purchase totals. This model would be useful for the store to be able to prepare for the future and make goals. Knowing how much money from customers they might bring in for the coming years could help them plan where to allocate money for expansion, upgrades, employment, etc.

Methods

The two models made will take in the predictors, or the data points we were given about the customers, and find the relation between them and the final data field, the total amount spent. The first model will try to create a linear relationship between the predictors and the total amount spent, so for the test, or prediction of the future for the model, it will continue predicting the amount spent based on a linear slope made by the linear line of best fit from the inputted data. For the second model, it will do the same thing, but based on a polynomial model, in this case, of degree 2.

To allow for the model to understand the data that it has been given, some preprocessing steps were needed. Data needs to be normalized for it to be comparable, as the model needs to know which variables play the biggest role in affecting the amount spent in order to make an accurate prediction. First, numeric and categorical variables need to be separated, because they are processed differently. Categorical variables in this model are gender, test group, and year. These need to be "One Hot Encoded" so that the computer recognizes them as whole values or groups. The rest of the variables are numeric, but for them to be compared to each other, they need to be z scored, which normalizes the number in terms of standard deviations from the mean of that field.

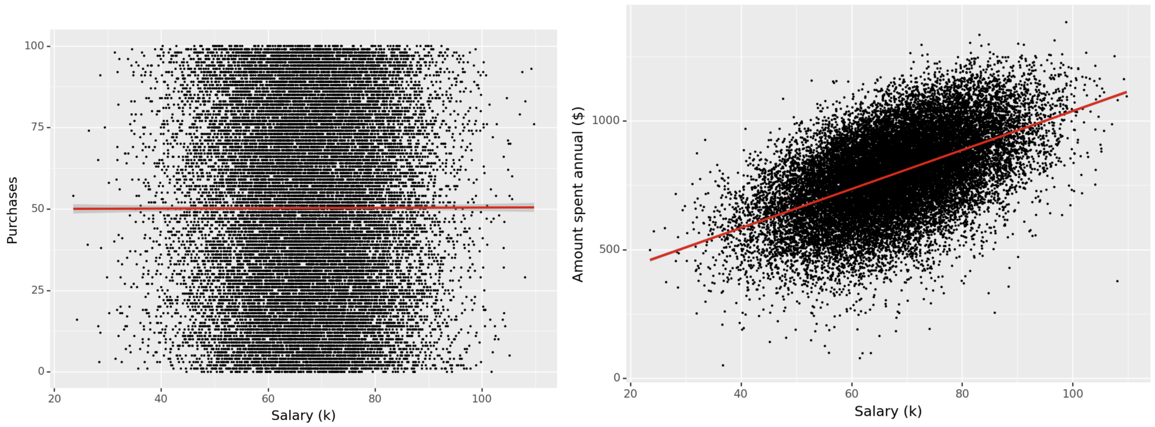
Results

The linear regression did moderately well with the data, but was underfit. We know this because both the train and test data had similar accuracy, but the MSE and MAE were relatively high (compared to the next model), the mean absolute percentage error was above 10%, and the R^2 was only around 50.

The polynomial regression performed much better with a degree 2. It doesn’t seem very overfit because although the accuracy of the train was better than the test, a slightly worse performance is expected, and the test metrics were still pretty close. The R^2 for the test was .88, meaning that 88% of the variability of the amount spent variable in relation to the independent variables can be explained by the model. Overall, the PolynomialFeatures were needed to explain an accurate relationship between the variables. I would recommend this model to the company, because the absolute percentage error is low, and the R^2 is high for both the seen and unseen data. A caveat would be from the interpretation of the MSE vs the MAE. Because the MSE is so much higher than the MAE, it suggests that there are some very large outliers that are being squared to create the large MSE. Therefore, dropping some outliers which represent customers with abnormal shopping behavior, could create a more accurate model on average, however, it would be ignoring the fact that this store does have a lot of customers that shop atypically compared to their peers, which would be helpful to know.

Question 1: Does making more money (salary) tend to increase the number of purchases someone makes? Does it increase the total amount spent?

Add an image:



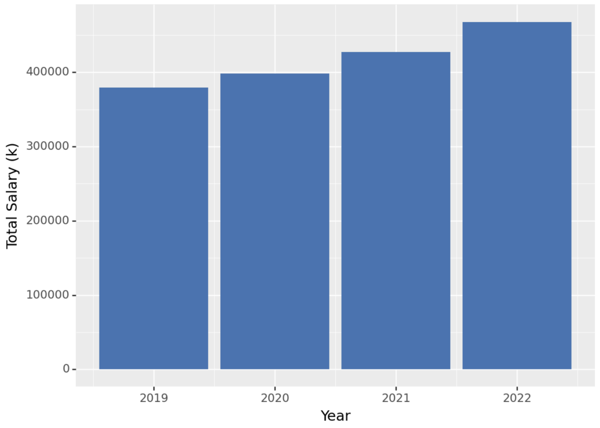
Caption : Salary vs purchases dot plot (left) and Salary vs amount spent annually (right)

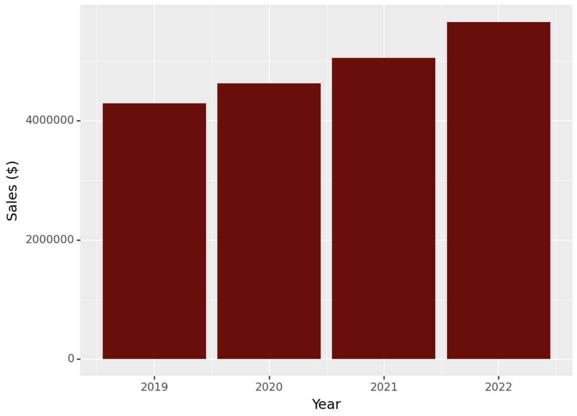
There does not seem to be a relationship between salary and purchases. Increasing salary (moving right on the graph) does not increase the number of dots higher up on the y axis (# of purchases). However, we see that for the salary vs amount spent graph, higher x values generally correspond to higher y values, which indicates a linear relationship. The line of best fit for the graph on the right shows that on average, making more money results in a higher total amount spent at the store.

Because with a raise in salary, the number of purchases stay the same but total purchase cost increases, we conclude that with more money, customers will buy more expensive products.

Question 2: In which year did the store's customers make the most money? Were the store's sales highest in those years?

Add an image:





Caption : Sum of customer salary per year (left) and total store sales per year (right)

The customers made more money each year, meaning they made the most money in 2022. Likewise, the sales went up each year, with the highest revenue in 2022.

Discussion/Reflection

Needing to make this model and analyze it on my own rather than watching the professor come up with the answers, I gained a better understanding of every topic that was part of this assignment, including data visualization, linear regression, and calculations with data frames and python packages such as pandas. In the future, I would work harder on the classwork assignments so that more of this assignment would have been intuitive from the start.