Assignment 1

Jaden Johnson

Introduction

With the models I have created, we are attempting to predict the annual spending for customers of a boutique store. The main goal is to have a model that can accurately predict the spending habits of a customer based on multiple factors. These factors include gender, age, height in cm, waist size in cm, inseam in cm, whether or not they were in a test group, self-reported salary, months in the company rewards program, the number of purchases they've made, the amount they spend annually at the store, and the year the data was collected.

If the model is successful, the store can use data to predict the spending behavior behind their customers and therefore tailor their store and products towards those customers in order to make more money. For example, if the data proves that they have many customers that aren't "average" size, the store may adjust their clothing to fit those of their customers, therefore causing the customers to buy more and make the store more money.

Methods

The first implementation I had was a linear regression model, which assumed a linear one-to-one relationship between the various factors that we were using in this model. As a form of preprocessing, I handled the possible missing data by removing rows with any missing values. I also coded categories such as gender and test group into numerical data with ones and zeroes so that it was readable by the graph. Lastly, I used StandarScalar to normalize the data so that major outliers didn't throw off the data too hard.

The second implementation that I had was a polynomial regression model using a degree of 2. This essentially means that instead of a linear regression model, I now have a polynomial model assuming the data will be more in the shape of a parabola (U) instead of a line. My feeling is that this would be better for the data because I feel as though a linear one-to-one relationship is not complex enough to describe the data relationships that we are working with here.

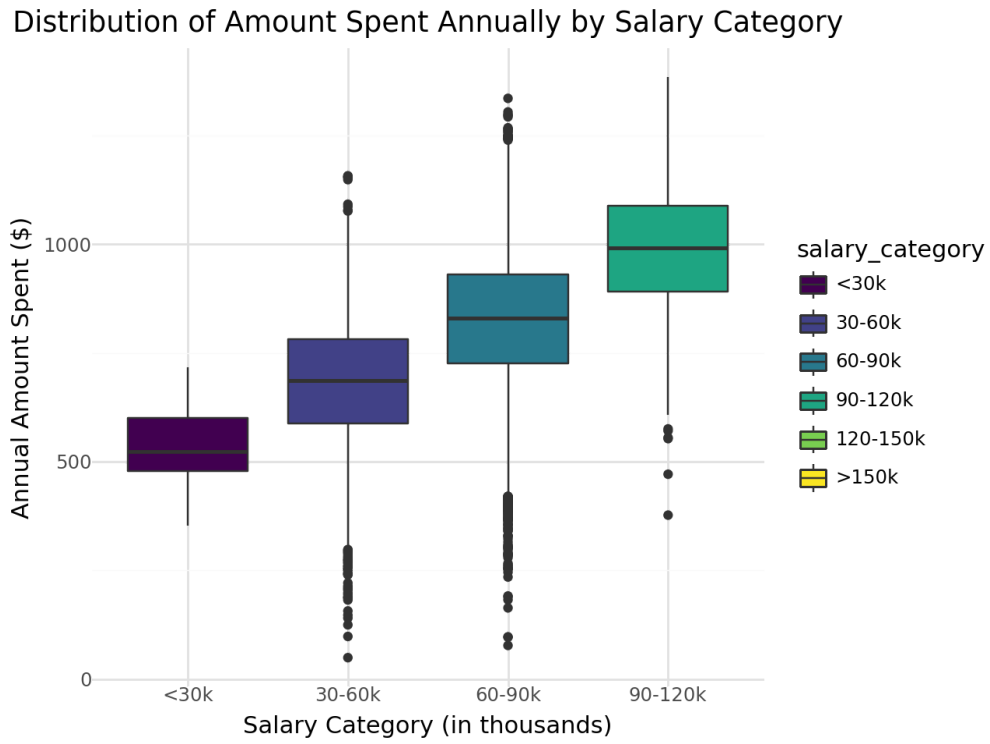
Results

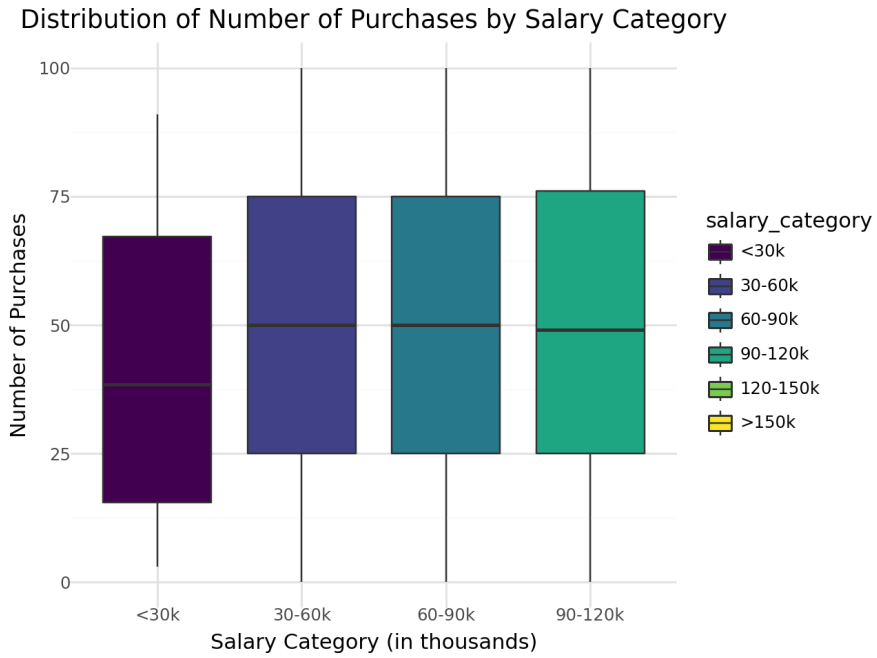
Overall, my polynomial regression model did much better than my linear regression model. My Train MSE & Test MSE are much lower than that of my polynomial regression, as well as the R^2 values. I think this makes a lot of sense, given that it's hard to state that these complex relationships can be dimmed down to a linear relationship. From this data, the performance metrics indicate that a polynomial regression model is much more accurate and fitting than a linear one.

In terms of whether or not my models are over/underfit, I would say that the polynomial model is doing pretty well since both its test and training R^2 values are about 80%. Since both the testing and training R^2 values in the linear model are super low around 44%, it's safe to say that the linear model is underfit for the data. In all, the linear model does not do a gerat job of representing this far more complex dataset. PolynomialFeatures were helpful in this case when switching from linear to polynomial regression models, which makes a lot of sense given that a linear model was poor for interpreting this data, while the polynomial model was much better. As explained earlier, this is most likely due to the fact that relationships such as height and salary are much more complex than just a one-to-one relationship.

Overall, I would feel pretty confident in my polynomial regression model since it had great training and testing datasets, meaning that it represented the data well. However, the linear model I would not be comfortable with giving out, as it's doing a very poor job and not representing the data sets in either the training or testing sets. I would also say that there is a caveat for both in that the relationships between the data I feel are much more complex than just a linear or polynomial model, so that it should need some further form of validation in order to truly prove that either model, especially the polynomial one, are valid to use. This could include some form of cross validation or feature programming to truly make sure the model we are using is good to use to predict data from customers.

Question 2: Does making more money (salary) tend to increase the number of purchases someone makes? Does it increase the total amount spent?

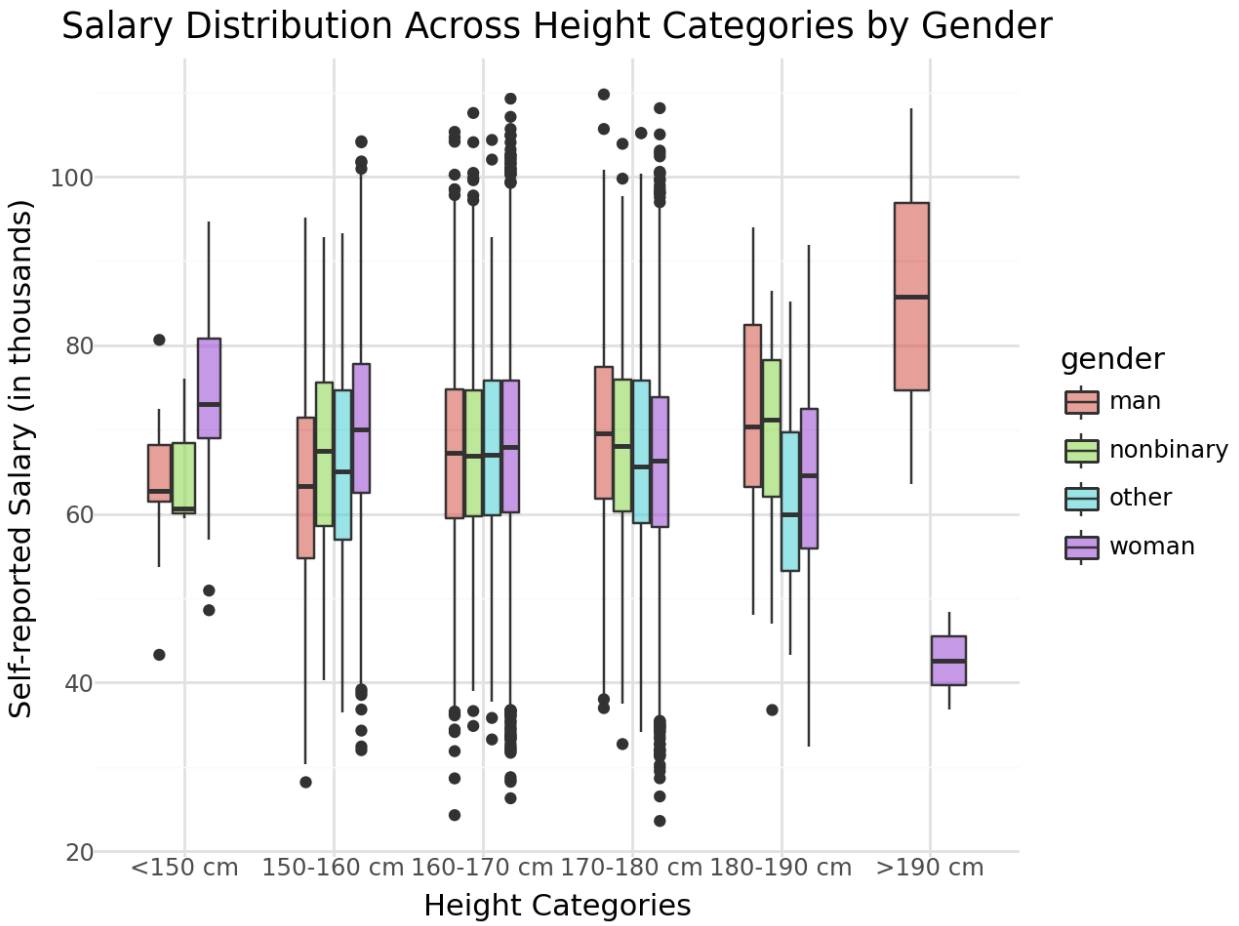




These two graphs are used to illustrate the relationship between customer salary, and either the amount of money spent by each customer, or the total number of purchases that each customer has made.

According to the graphs, salary doesn't necessarily determine the number of purchases that customers make. Given that the 30k -120k salary range for customers relatively have the same number of purchases, I think that this is a safe assumption to say that there's no tmuch of a relationship here. However, according to the first graph of salary vs total money spent, those who have a higher salary definitely have a larger annual spending amount at the boutique than those who have a lower salary. We can tell this because as the salary increases, there is also a tendency for the annual spending amount for customers to also increase, hinting at a possibly one-to-one relationship. This would mean that the boutique might want to focus on their customers with higher salaries, as they may possibly make more money due to the fact that those people tend to spend more.

Question 5: In this dataset, is there a relationship between salary and height? Is it different for the different genders?



Caption 2:

This graph is used to illustrate the relationship between the height of customers of all genders, and their self-reported salaries. Each different color bar represents a different gender, and the black dots are outliers amongst their respective data points.

There is a very clear difference between men and women above 190 cm, with men having an overall higher salary than women. However overall, I would say that there is actually a linear relationship here in terms of height vs self-reported salary. At < 150cm, women had a higher salary, but gradually decreased as the height decreased. For men however, they had a lower salary at shorter heights, but then increased as they grew in height. For nonbinary and others though, they stayed pretty relatively neutral throughout the entirety of the heights. Given this information, the boutique might want to look into making clothes tailored for men of taller builds as they have higher self-reported salaries. Since this group has a higher salary, given our last graph, they may also be willing to spend more.

Discussion/Reflection

I now can see a clear difference in the modeling approaches between linear and polynomial aggression. Given how complex the relationships between data are here, it makes sense that polynomial regression can better match the data than a simpler linear model. Overall, this shows that polynomial regression can offer better benefits when the data relationships are more complex. This is especially true when we are working with data from real-life scenarios, as these relationships are more complicated than just one choice or the other, further proving why the polynomial model is better than the linear one.

In the future, I would probably try to utilize more ridge or lasso techniques in order to help control the complexity of the data and generalize it into more simple numbers. This is especially true here because there were some data points that were major outliers and could've definitely skewed the data. I would also try and find better strategies to fit the data if possible to the linear model so that it's does better than just a R^2 value of 44%.