CPSC 392: ASSIGNMENT 1

DIYA KULKARNI

INTRODUCTION

In this model, I am utilizing clothing store data, “boutique.csv”, to predict the average amount that customers will spend with the company per year. In the data, I was given the categorical variable of the self-disclosed gender identity of customers, “gender”, along with numeric data of their current age (“age”), self-reported height (“height_cm”), self-reported waist size (“waist_size_cm”), self-reported inseam (“inseam_cm”), whether or not the customer is in an experimental test group that gets special coupons once a month (“test_group”), self-reported salary (“salary_self_report_in_k”), number of months customer has been part of the clothing store's preferred rewards program (“months_active”), number of purchases the customer has made (“num_purchases”), and the year the data was collected. (“year”).

If this model is successful, it could be incredibly useful for the company to evaluate what the key variables are in maximizing its revenue through customer sales. It can also help them determine problem areas within these variables and implement solutions to ensure a higher level of customer satisfaction.

METHODS

After loading the data and all the necessary imports and checking for missing values, the first thing I did was utilize dummy variables to convert gender, a categorical variable, to a numeric variable that could be used in my model. Following this, I split the data into 80% training data and 20% testing data. Finally, I Z-scored all of the continuous/interval variables, also converting them to a format that would work with my model.

After the pre-processing, I fit my linear regression model using the data from both the train and test sets, and found the mean squared error, mean absolute error, and r-squared values for both the train and test sets. Comparing the two is a way for us to determine the accuracy of the model.

Following the linear regression model, I created a polynomial regression model, which is a good way to model data that isn’t linear. By using the PolynomialFeatures() function, the model is able to consider the polynomial functions of inputs, allowing it to find a curve that best fits the data. Similar to the linear regression model, I used my training data and testing data (now a bit different due to the incorporation of the PolynomialFeatures()), and found the mean squared error, mean absolute error, and r-squared values.

RESULTS

I believe that both models performed incredibly well.

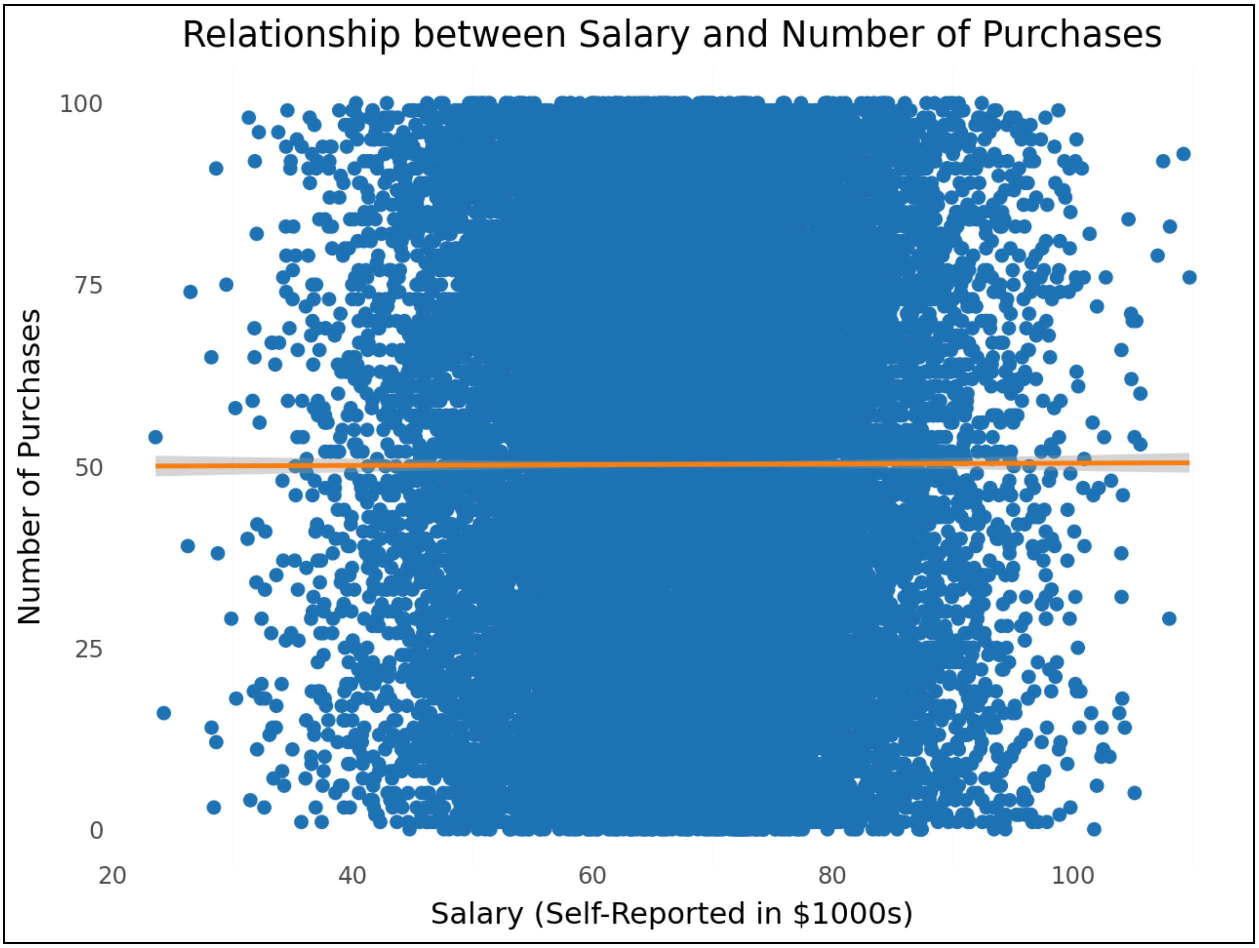
First, looking at the **linear regression model**, the testing mean squared error has a fairly large difference from the training mean squared error. The testing mean absolute error also varies a good amount from that of the training. The r-squared values for both the training and testing data are pretty small, indicating that only around 28% of the variance in the actual dataset is explained by both the training and testing model. This indicates that the model does not do a very good job of explaining the variance in the data. It would need more complexity to be a better fit.

	MSE	MAE	R^2
Train	19321.169092703483	110.41923767775656	0.2863640452148456
Test	19899.74546347162	112.00428344705955	0.2748757565434625
Difference (Train - Test)	-578.576371	-1.58504577	0.0114882887
Difference in %	-2.99%	-1.44%	4.01%

Next, looking at the **polynomial regression model**, the training mean squared error and mean absolute error were a much smaller margin lower than the testing when compared to the linear regression model. Similarly, the mean absolute error had little difference in its training and testing data. The r-squared values of around 64% indicates that the model did a much better job at predicting the data when compared to the linear regression model. It is definitely beneficial to have been able to carry out the Polynomial Regression Model, as the PolynomialFeatures() function allowed for more complexity, ensuring the model was a better fit for the data.

	MSE	MAE	R^2
Train	9736.676500265015	80.1616634396498	0.6403715325215578
Test	9958.863242250442	80.60596733646962	0.6371102742253791
Difference (Train - Test)	-222.186742	-0.444303897	0.0032612583
Difference in %	-2.28%	-0.55%	0.51%

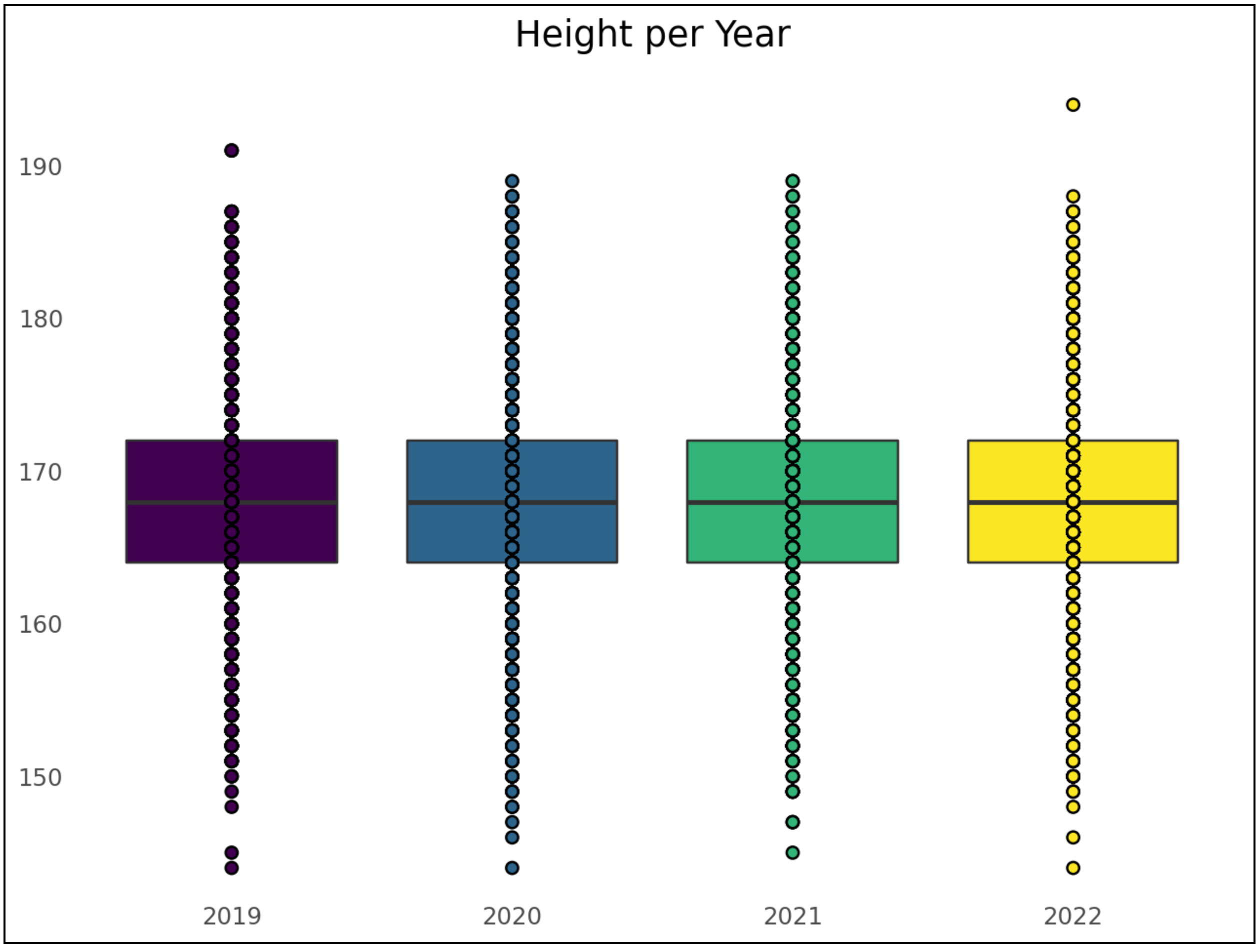
Question 1: Does making more money (salary) tend to increase the number of purchases someone makes? Does it increase the total amount spent?



This model demonstrates the relationship between the self reported salary of customers (in 1000s) and the number of purchases they made yearly.

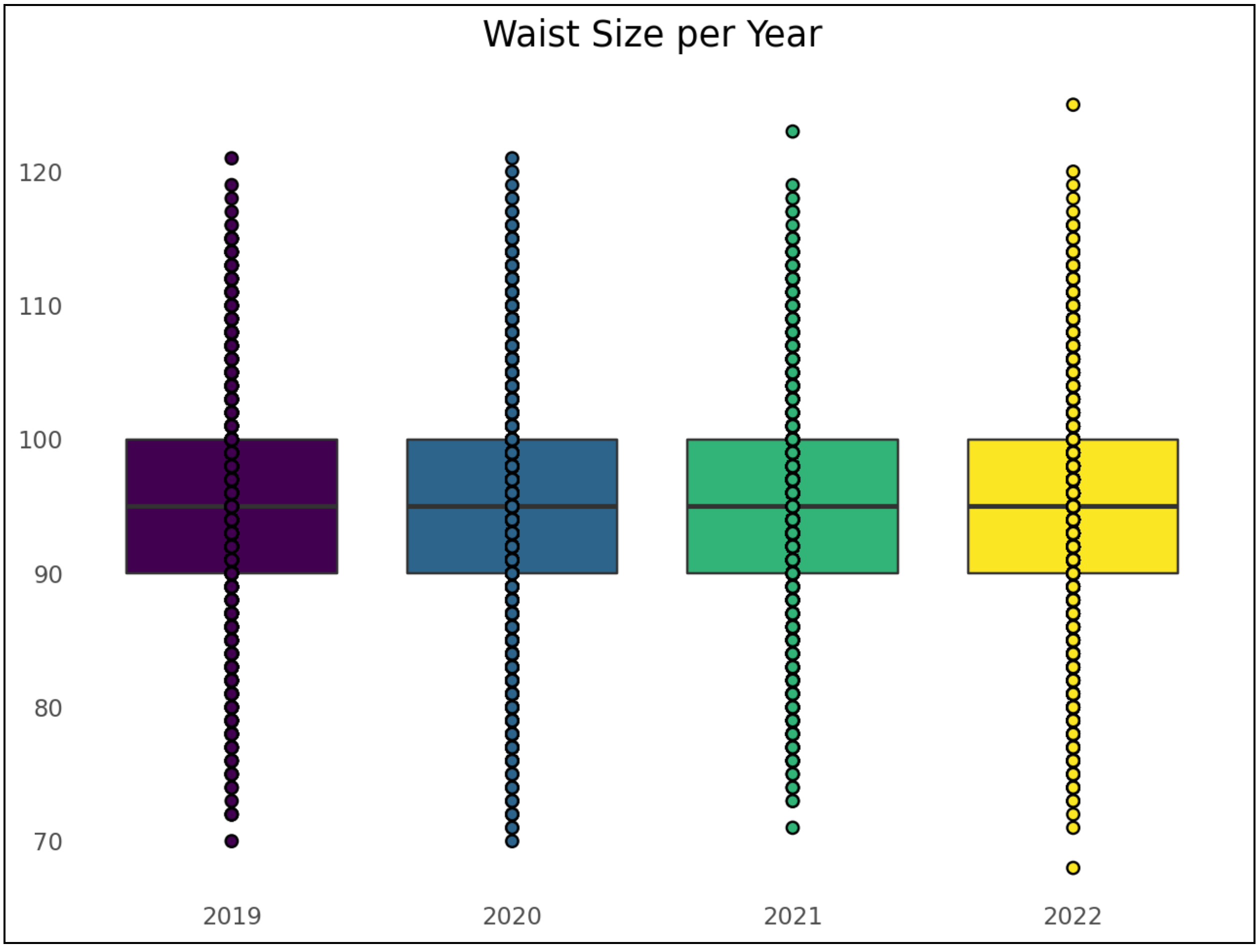
According to the graph depicted above, there is no evident relationship between a customer's self-reported salary and the number of purchases they made. This is surprising because I would have expected those with a higher salary to have the budget for more purchases, but there does not seem to be a relationship at all. Most of the purchases are clustered around those who make 40k to 95k, but even those range from 0 to 100 purchases with no clear distinction based on the salary.

Question 2: Has the store's customer base changed over time (based on the height, waist size, and inseam per year)?



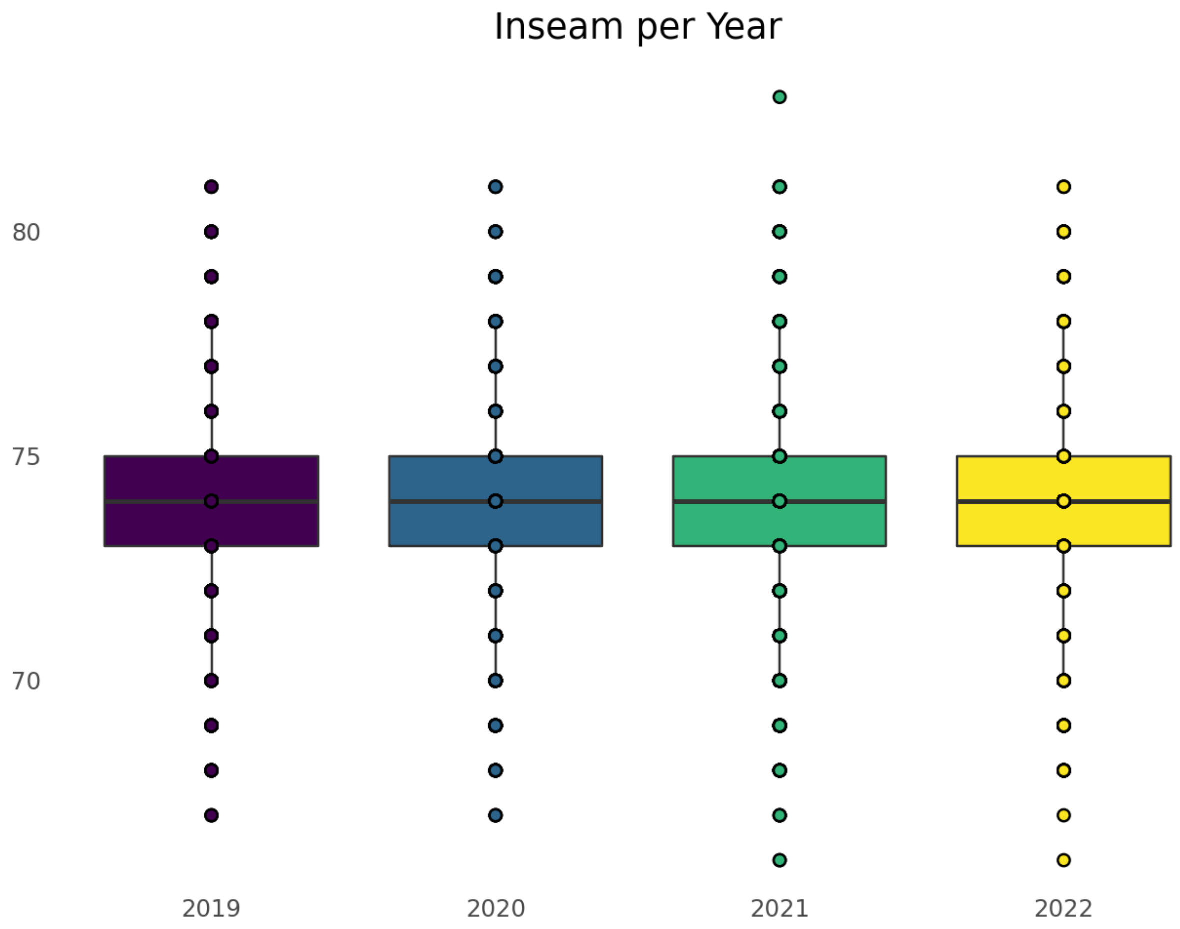
This model demonstrates the distribution of customers' self reported heights throughout 2019 to 2022

According to the graph depicted above, the minimum, maximum, and average heights of customers stayed constant throughout the years. There were a few data points that lie outside the typical range throughout the year, such as 2019 and 2022. However, for the most part, the customer base in terms of heights stayed constant.



This model demonstrates the distribution of customers' self reported waist sizes throughout 2019 to 2022.

According to the graph depicted above, the minimum, maximum, and average waist sizes of customers stayed relatively constant throughout the years. There are a few data points that imply a slight increase in waist sizes as the years progressed, but seeing as it was only really one data point out of many, it is not enough to draw a general conclusion.



\This model demonstrates the distribution of customers' self reported inseams throughout 2019 to 2022.

According to the graph depicted above, the minimum, maximum, and average inseams of customers varied slightly throughout the years. It stayed relatively constant between 2019 and 2020, then had a slight outlier in 2021, before reverting back to its consistent state in 2022.

DISCUSSION/REFLECTION

This assignment was incredibly helpful in letting me apply the techniques learned in class to create an effective data visualization. By immersing myself in the dataset, I was able to put myself in the company's perspective and understand what would be beneficial to analyze, and what that meant in terms of the company's success. I did run into some confusion when creating the models, but working through it allowed me to better understand how to carry out data visualization in Python.