CPSC 392

Homework 1

Ethan E. Lopez

Introduction

description of the problem (e.g. what are you predicting? what variables do you have available? How might this model be useful if you are successful). You should end with a sentence or two about what the impact of these models could be.

A clothing store aims to predict how much their customers will spend with them annually using previous records and demographics. Of the information in the data, there are nine independent variables with yearly labels geared towards predicting the amount_spent_annual for future customers.

Two variables are inherently categorical in this dataset. Gender distinguishes the spending differences between males, females, nonbinaries, and others with test_group identifying if they were part of an experimental program receiving monthly coupons or not. Age, height_cm, waist_size_cm, inseam_cm, salary_self_report_in_k, months_active, and num_purchases are the remaining numerical factors determining traits such as the customer's clothing size, monetary wealth, and level of store involvement.

Given the model is proven successful after testing, it will be able to determine a person's yearly average spending given certain characteristics in their background. Allowing the store to identify incoming customers with greater spending traits, the company will be able maximize their profit through specialized advertising, deals, and other offers extended to these customers to engage them to buy more products.

Methods

Describe your models in detail (as if explaining them to the store's CEO), as well as any pre-processing you had to do to the data.

Prior to building the models, the data's contents had to be evaluated for suspicious or misaligned values. Looking at the head, shape, and info, this was important to confirm the data's size, variable types, and descriptive stats as sensible.

Checking null and missing values, I further had to count how many were present to ensure that when I deleted them, a good portion of the data wouldn't be lost. Columns for age, inseam_cm, and salary_self_report_in_k together containing 235 missing values, this resulted in 233 rows being dropped. Out of the 25000 rows total, however, a majority 99% of the data was still there, which meant I was able to proceed with building my models.

Three models in total were created using the nine variables (excluding year) to predict the amount_spent_annual. The first model a line and our second and third models curves, these were each generated using a method called train-test-split, which took our data and split it into two groups. The first group being 80% of the data, this portion was used for building the actual model during training with the remaining 20% used for testing the model's accuracy by predicting values. Each of the three models having different interpretations and regression types, comparisons and differences were then charted using various error measurements and R^2 values.

The first model being a linear regression, this assumes that there is a constant rate of change determined by the nine independent variables. With every one unit increase in a factor, this results in an increase or decrease in the customer's amount spent by the same value every time given all the other variables don't change. This model works for simplicity but also shows weaknesses in high bias, which can result in misleading conclusions.

The second model is a polynomial regression of degree 2, forming a U-shaped quadratic curve to predict annual amount spent. This model is more complex and not as simple as the line in the first model, introducing non-linear features that suggest a shifting rate of change. With this model, however, there's less bias and more variance, allowing it to capture relationships between data points in more detail. At the same time, coefficients become harder to interpret as the degree increases.

The third model out of the three is the most complex, a polynomial degree of 3 showing a curve with one bump and one dip in the middle before continuing. Where in the case a simple U doesn't capture the trend of the points well, this more elaborate curve is meant to perform better by adding more edges to the graph. Having a higher complexity and shape, on the other hand, makes this one of the hardest models to interpret and could miss identifying simple relationships between variables. With less bias, this also makes the model weaker in generalizing as more variance applied to training doesn't always mean better predictions in unseen data.

Results

How well did your model perform according to the various metrics, was the model overfit (how can you tell)? What do those performance metrics tell you about the model? Did you need Polynomial Features (which includes both polynomial features and interactions)? How much do you trust the results of your model (in other words, would you be confident telling the store that they should use the model? Why or why not? Are there any caveats you'd give them?) Also answer the two questions you chose from part 2 above. Include the image, a caption as well as your written answer.

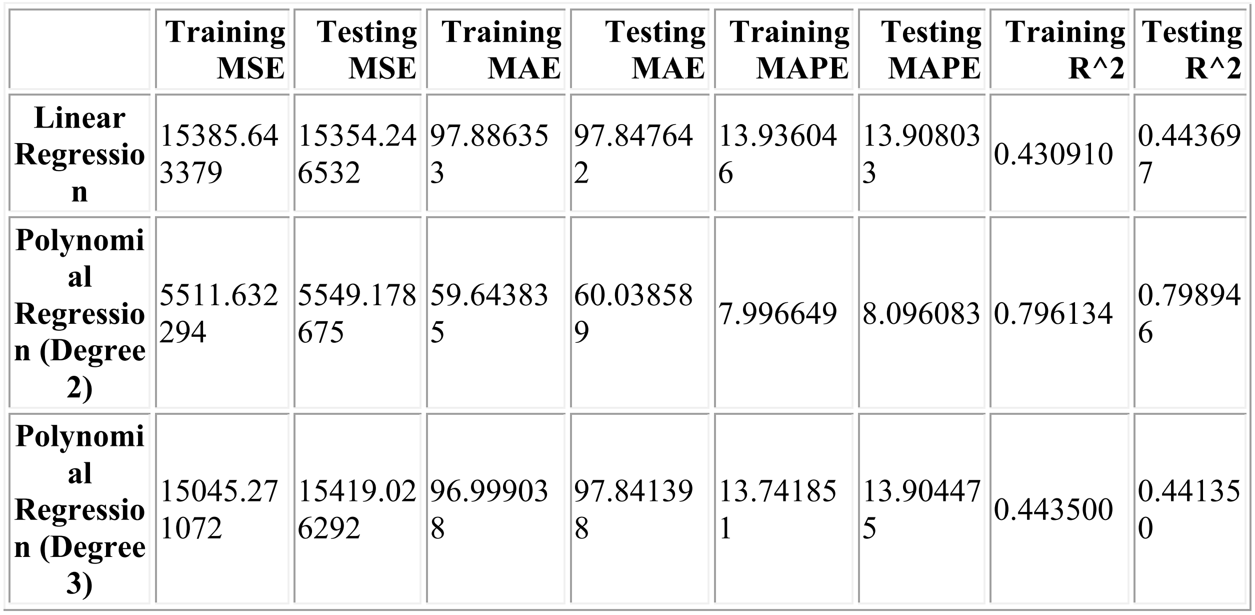
In the first and third models validating the predicting power of a line vs. a 3rd degree curve, there were similar patterns observed. The R^2's in both were found to be moderate, where around 44% of the variance in average annual spending was explained by all other variables in the models. All errors across MSE, MAE, and MAPE were also standard, showcasing optimally average predicting power for the store.

The second model, on the other hand, stood out the most in terms of its performance. With a strong R^2 of approximately 80% and errors that were significantly less than the ones in the previous models, this quadratic regression was found to be the best fit and superior in predicting the average annual amount spent for customers in this dataset. Displaying that polynomial features of degree 2 were in fact necessary to guide along the relationships between the variables, this conveys a moderately complex association in customer background vs. their spending habits.

Now, as good as this sounds, I personally wouldn't trust the results of the degree 2 model, nor would I be confident in telling the store that they should use it to predict the future data. The overwhelming optimism in the measurements alone shows weakness in the model being prone to slight overfitting, where its strength in explaining sampling data doesn't always mean the regression will perform well with the unseen, unpredictable data. Involving social/demographic trends especially is where we should aim to be less complex as certain environments and events can influence people's stylistic decisions, dramatically altering the fashion world to make it less foreseeable in the future. With this, a brief overview in my opinion is better for the store's situation, focusing on less noise in the data as this noise is likely to change not now, but perhaps in the coming years following it.

Supporting my previous claim, in fact, is where I believe the linear regression model to be the most appropriate out of the three. The line simplifying variables to interpretable factors with a constant rate of change, despite having more bias, could possibly perform better with unseen data as it isn't specifically tailored to fit the flow of this dataset only. Not only so, but this also makes the model easier to communicate to the store's stakeholders and clients who might not understand non-linear patterns.

I'd still advise using the linear model with caution, however, as it assumes that with every independent variable, there is a fixed change in spending when it increases by 1 unit. This is rarely ever the case, so it's important not to get to hung up on coefficients showing accurate rates of change when testing the model in future years.



	Training MSE	Testing MSE	Training MAE	Testing MAE	Training MAPE	Testing MAPE	Training R^2	Testing R^2
Linear Regression	15385.643379	15354.246532	97.886353	97.847642	13.936046	13.908033	0.430910	0.443697
Polynomial Regression (Degree 2)	5511.632294	5549.178675	59.643835	60.038589	7.996649	8.096083	0.796134	0.798946
Polynomial Regression (Degree 3)	15045.271072	15419.026292	96.999038	97.841398	13.741851	13.904475	0.443500	0.441350

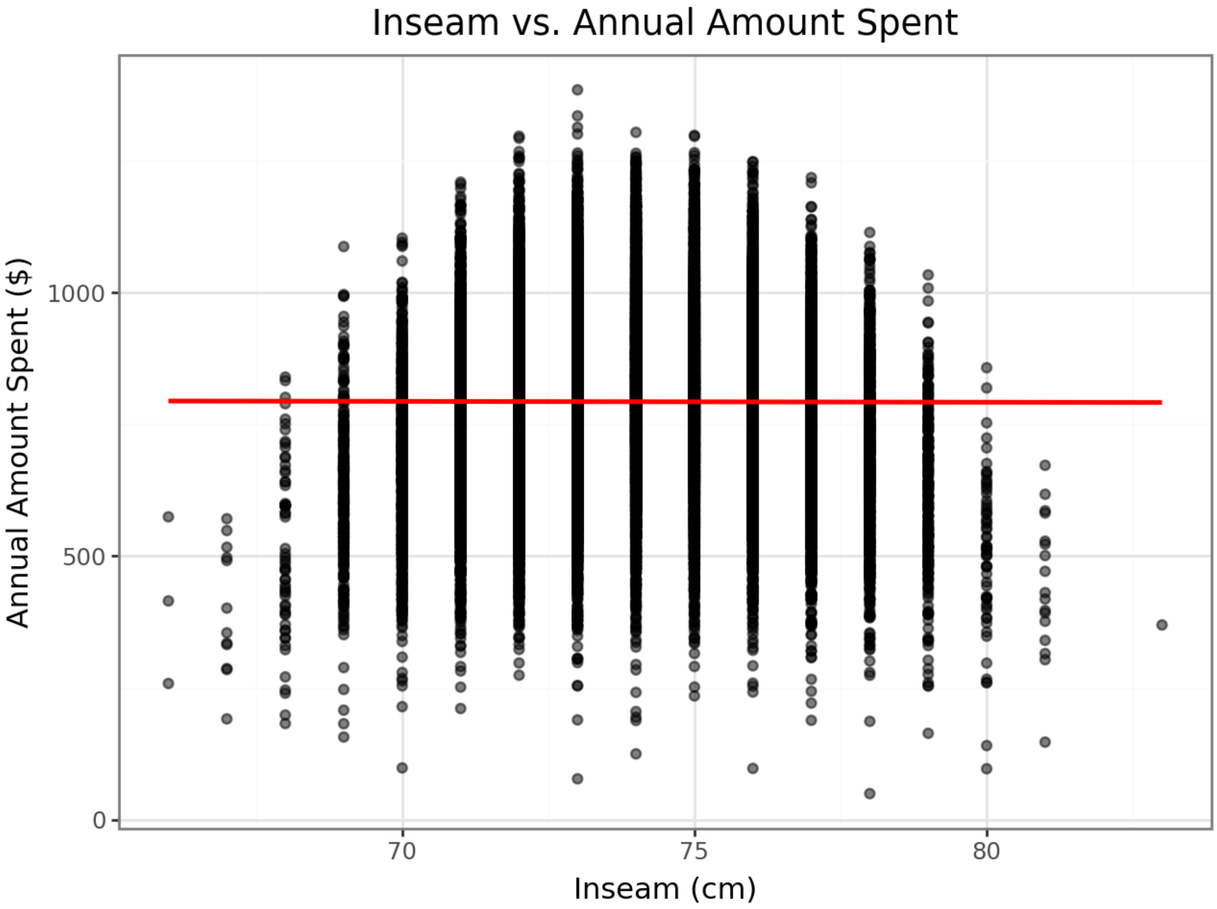
Also: answer the two questions you chose from part 2 above. Include the image, a caption as well as your written answer.

Question 4:

People who are not your "average" size often find it difficult to buy clothes in traditional stores. Is there a relationship between inseam and amount spent in the

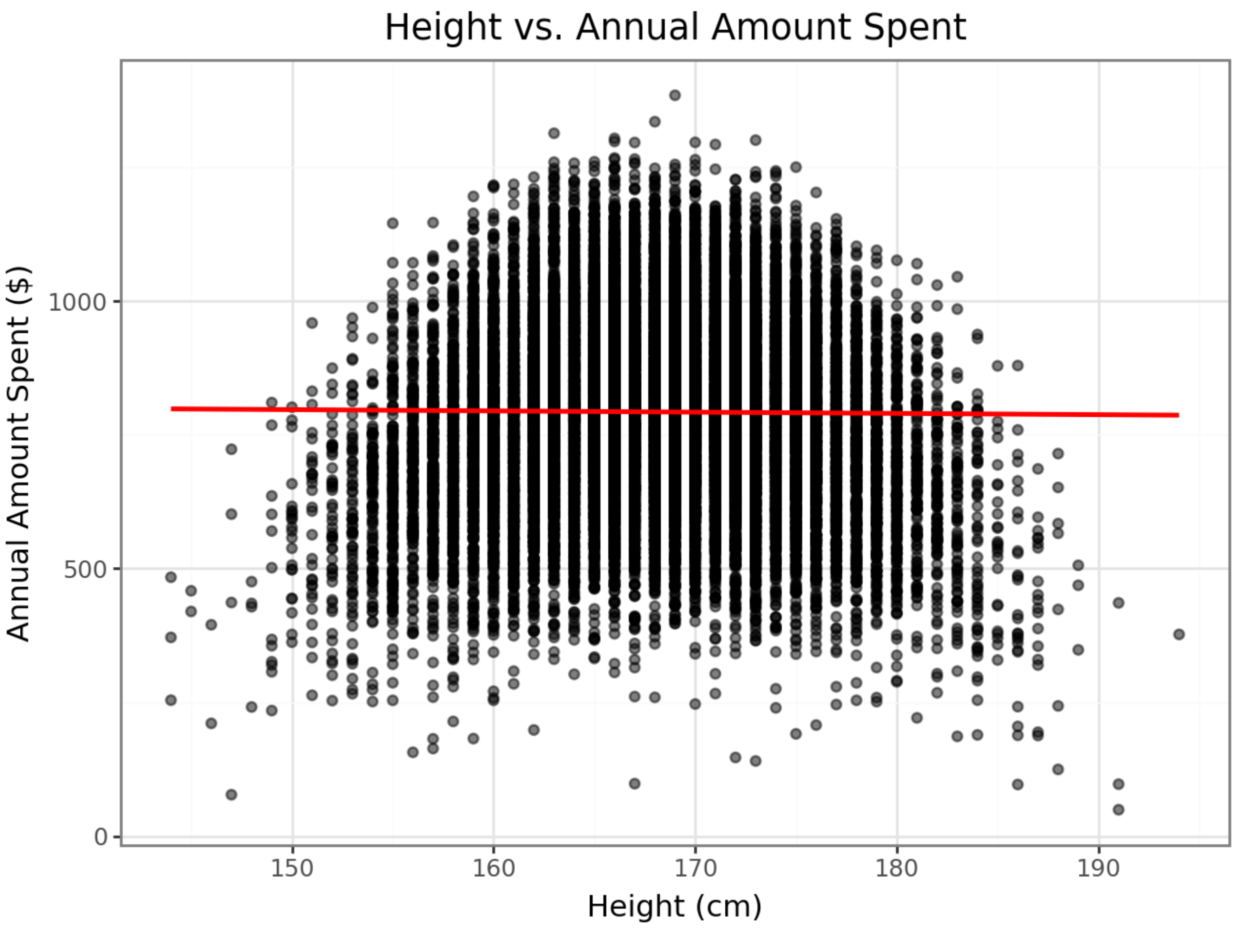
store annually? Is there a relationship between height and amount spent in the store annually?

Figure 1:



A scatterplot demonstrating the relationship between inseam length in cm and the average annual amount spent in dollars / red line shows overall trend by linear smoothing

Figure 2:



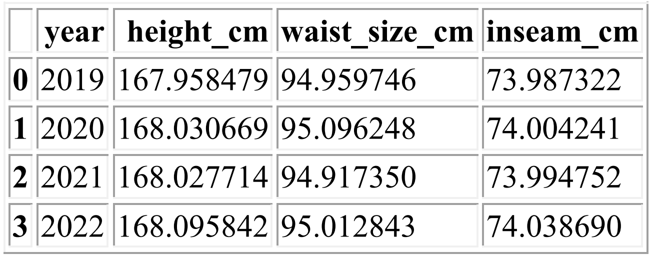
A scatterplot demonstrating the relationship between customer height in cm and the average annual amount spent in dollars / red line shows overall trend by linear smoothing

Displayed in the scatterplots above, there is no evident relationship between average annual amount spent vs. a person's inseam and height. With linear smoothing applied, the trends in both diagrams amount to a zero slope, indicating no significant correlation in one variable affecting the other. This shows in the case of the clothing store that a person's size generally doesn't affect their spending habits.

Question 5:

The store is interested in whether their customer base has changed over time. Present the minimum, maximum, and average height, waist size, and inseam for each year.

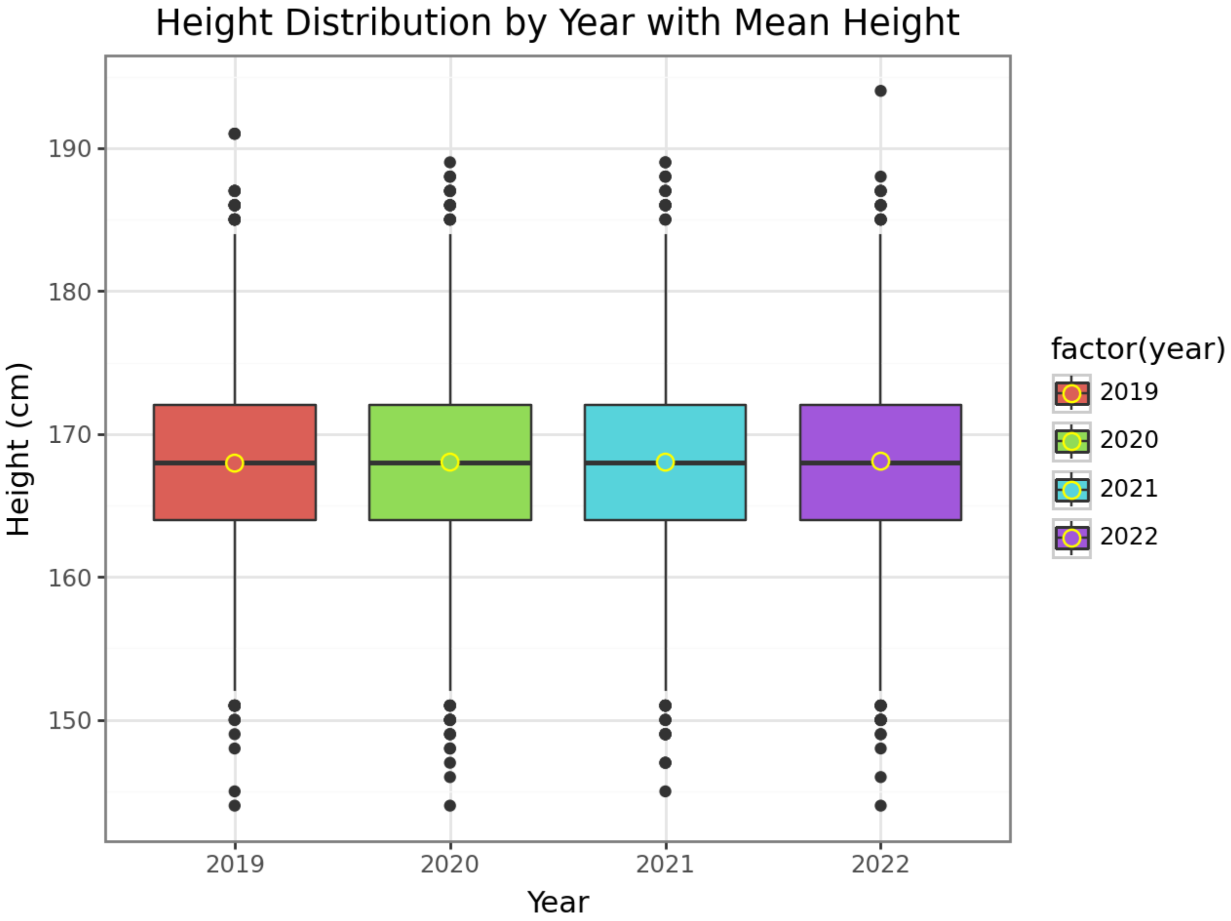
Figure 1:



	year	height_cm	waist_size_cm	inseam_cm
0	2019	167.958479	94.959746	73.987322
1	2020	168.030669	95.096248	74.004241
2	2021	168.027714	94.917350	73.994752
3	2022	168.095842	95.012843	74.038690

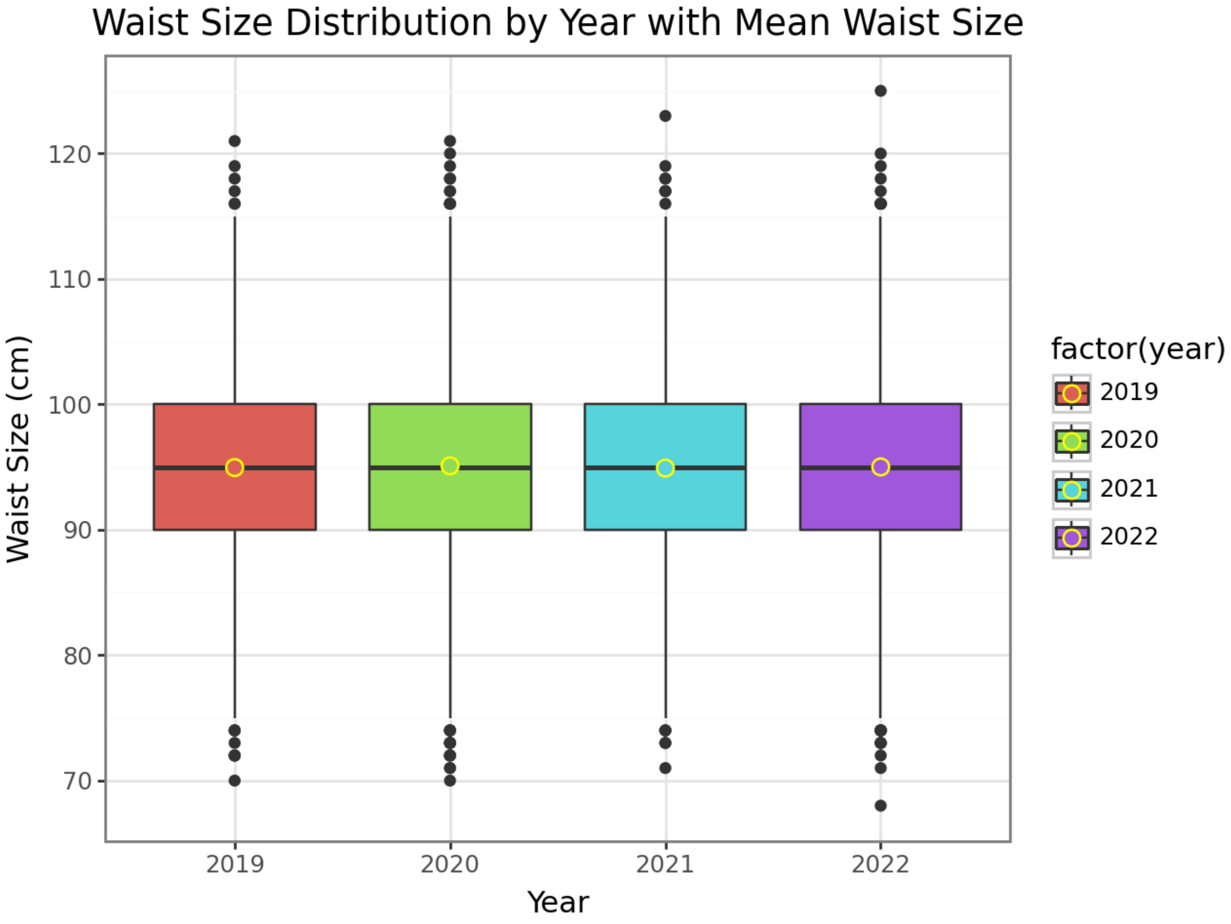
A table displaying averages for height, waist size, and inseam within years 2019-2022

Figure 2:



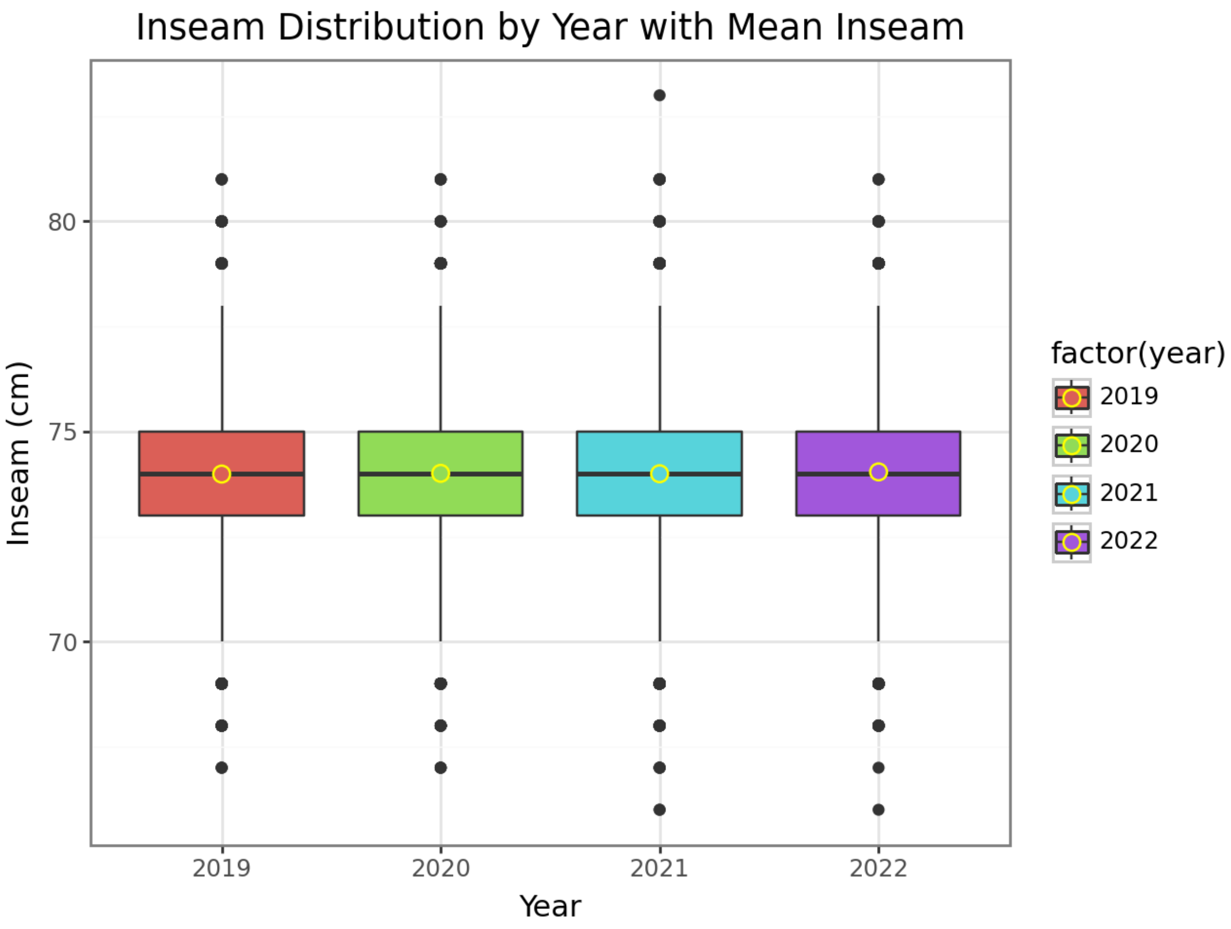
A series of boxplots displaying customer height in cm from years 2019-2022 / including minimums, maximums, and average points (yellow)

Figure 3:



A series of boxplots displaying customer waist size in cm from years 2019-2022 / including minimums, maximums, and average points (yellow)

Figure 4:



A series of boxplots displaying customer inseams in cm from years 2019-2022 / including minimums, maximums, and average points (yellow)

Observing the widths of each boxplot, there appear to be no visible changes in the company's overall customer base. Where minimums and maximums shift minimally in each diagram, averages, medians, and quartiles are seen at similar levels throughout the years. This shows us that despite time moving forward, customer traits in waist size, height, and inseam length have remained consistent.

Discussion/Reflection

A few sentences about what you learned from performing these analyses, and at least one suggestion for what you'd add or do differently if you were to perform this analysis again in the future.

Performing this analysis, I observed the clothing store’s most “optimal” model for predicting customer spending habits to be a quadratic regression, though conceptually a linear model fit for better simplicity. A few important techniques I used to help me reach this point included pre-processing the data by checking null values and using train-test-split to evaluate each model’s performance. Values of R^2 and error measurements helped identify strengths in explanation for the three degrees, and though degree 2 was the best fit, it wasn’t completely trustworthy because of its powerfully high percentage and the idea of future data possibly not adhering to the same curve.

Viewing the simplicity of the assignment, however, I realized there were other features that could’ve been used to validate my model’s strength further. Performing this analysis in the future, I’ll likely add additional graphs including residual plots to determine homoscedasticity and use Lasso and Ridge regularizations to prevent final model overfitting. Moreover, I would also look at coefficient p-values and more correlation measurements like r to evaluate strengths between variables and overall relationships before deciding the most optimal model.