# Clothing Retail Sales Analysis

Emmy Son

# Introduction

This project aims to predict how much a customer spends annually at a clothing store using a variety of demographic, body measurement, and behavioral variables. The dataset includes self-disclosed gender identity (male, female, nonbinary, or other), current age, height (in cm), waist size (in cm), inseam length (in cm), and self-reported salary (in thousands). It also tracks whether the customer is part of a test group that receives monthly coupons, as well as how many months they’ve been active in the store’s preferred rewards program, the number of purchases they’ve made, and the year of data collection.

By building linear and polynomial regression models using these variables, we aim to uncover patterns that influence customer spending. This model can help the store personalize marketing campaigns, optimize promotional strategies, and better understand how customer traits relate to spending habits. If successful, the model could lead to more data-driven decisions and improved customer engagement across different identities and income levels.

# Methods

To better understand the factors influencing annual customer spending, I began by carefully examining the available data. This included customer demographics (such as gender and age), physical attributes (height, waist size, inseam length), self-reported salary, participation in a promotional test group, and engagement metrics like months active in the rewards program and number of purchases. Initial exploratory analysis, using visualizations like scatter plots and box plots, helped us identify early patterns, such as differences in spending behavior by gender and income level.

Prior to modeling, we performed standard data cleaning to remove incomplete records and formatted variables appropriately for analysis. I converted gender into multiple binary variables (e.g., male, female, nonbinary) so that the models could interpret them effectively. The dataset was then split into training and testing subsets to ensure our models would generalize well to new customer data.

We developed two predictive models. The first was a linear regression model, which provides a straightforward way to quantify how individual features—such as salary or time in the rewards program—are associated with spending. This model is interpretable and helps identify the most direct drivers of customer value.

The second was a polynomial regression model, which allows for more nuanced relationships, including nonlinear effects and interactions between variables. This model is useful for uncovering subtler patterns—for example, whether certain combinations of physical attributes correspond to higher spending. We evaluated both models using standard performance metrics like Mean Squared Error and R-squared to assess how accurately they predict annual spending. These insights provide valuable guidance for refining customer segmentation, optimizing marketing efforts, and increasing overall revenue.

# Results

Our analysis compared two models: a simpler linear regression and a more complex polynomial regression. The polynomial model performed substantially better across all key metrics. For example, it explained approximately 89% of the variation in annual spending, compared to just 52% for the linear model. The polynomial model also showed lower errors in predicting spending amounts on both the training and testing data, indicating that it captures customer behavior more accurately without overfitting—meaning it generalizes well to new customers.

The improved performance of the polynomial model suggests that spending patterns are influenced by more complex relationships between factors such as salary, age, and physical measurements. Therefore, including polynomial features and interactions was necessary to capture these subtleties. However, it’s important to note that the models still showed high percentage errors when customers spent very little, so predictions for low spenders should be interpreted with caution.

**Evaluation Metrics Tables:**

Linear Regression Model:

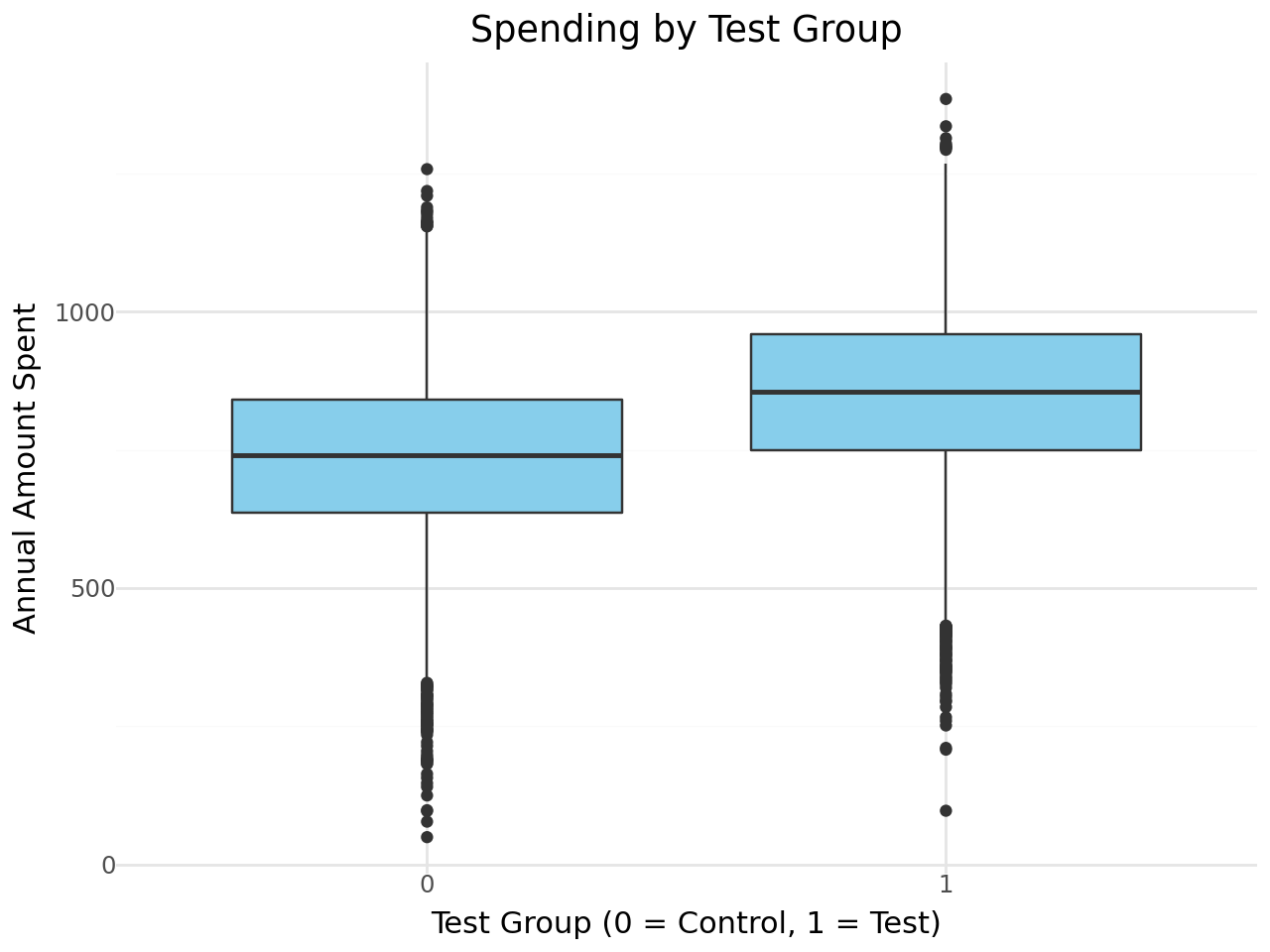
|  | Mean Squared Error | Mean Absolute Error | Mean Absolute Percentage Error | R² |
| --- | --- | --- | --- | --- |
| Training Set | 0.4765 | 0.5453 | 327.65% | 0.5207 |
| Testing Set | 0.4818 | 0.5457 | 643.97% | 0.5292 |

Polynomial Regression Model:

|  | Mean Squared Error | Mean Absolute Error | Mean Absolute Percentage Error | R² |
| --- | --- | --- | --- | --- |
| Training Set | 0.1153 | 0.2711 | 239.40% | 0.8840 |
| Testing Set | 0.1118 | 0.2661 | 202.92% | 0.8907 |

**Exploratory Data Analysis:**

## Question 1: Does being in the experimental “test\_group” increase the amount a customer spends at the store? Is this relationship different for the different genders?

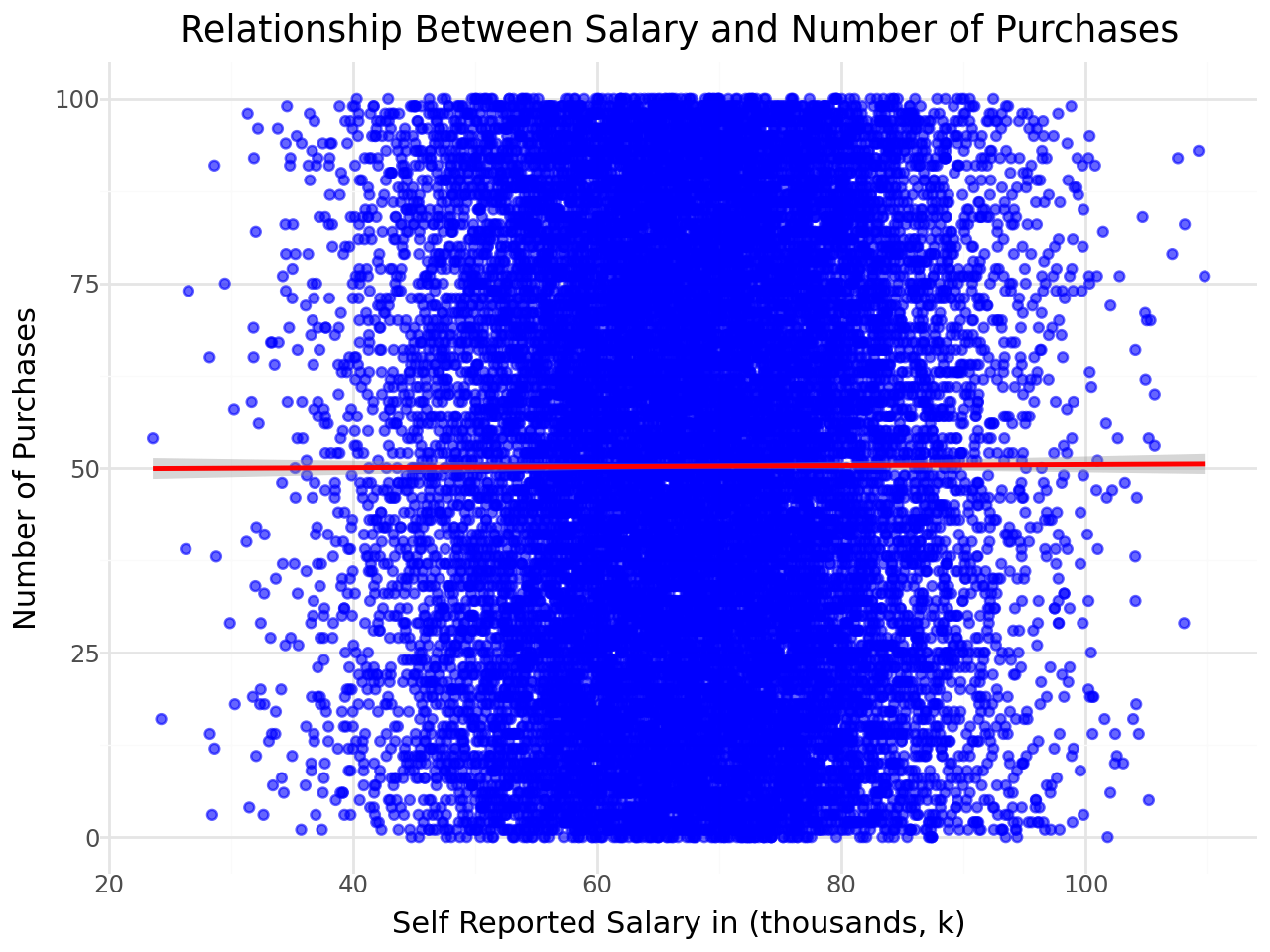


**Figure 1:** Comparison of Annual Spending Between Test Group and Control Group Customers.

This boxplot illustrates that customers who received monthly coupons as part of the experimental test group tend to spend more annually at the store compared to those in the control group, indicating the effectiveness of the promotional program.

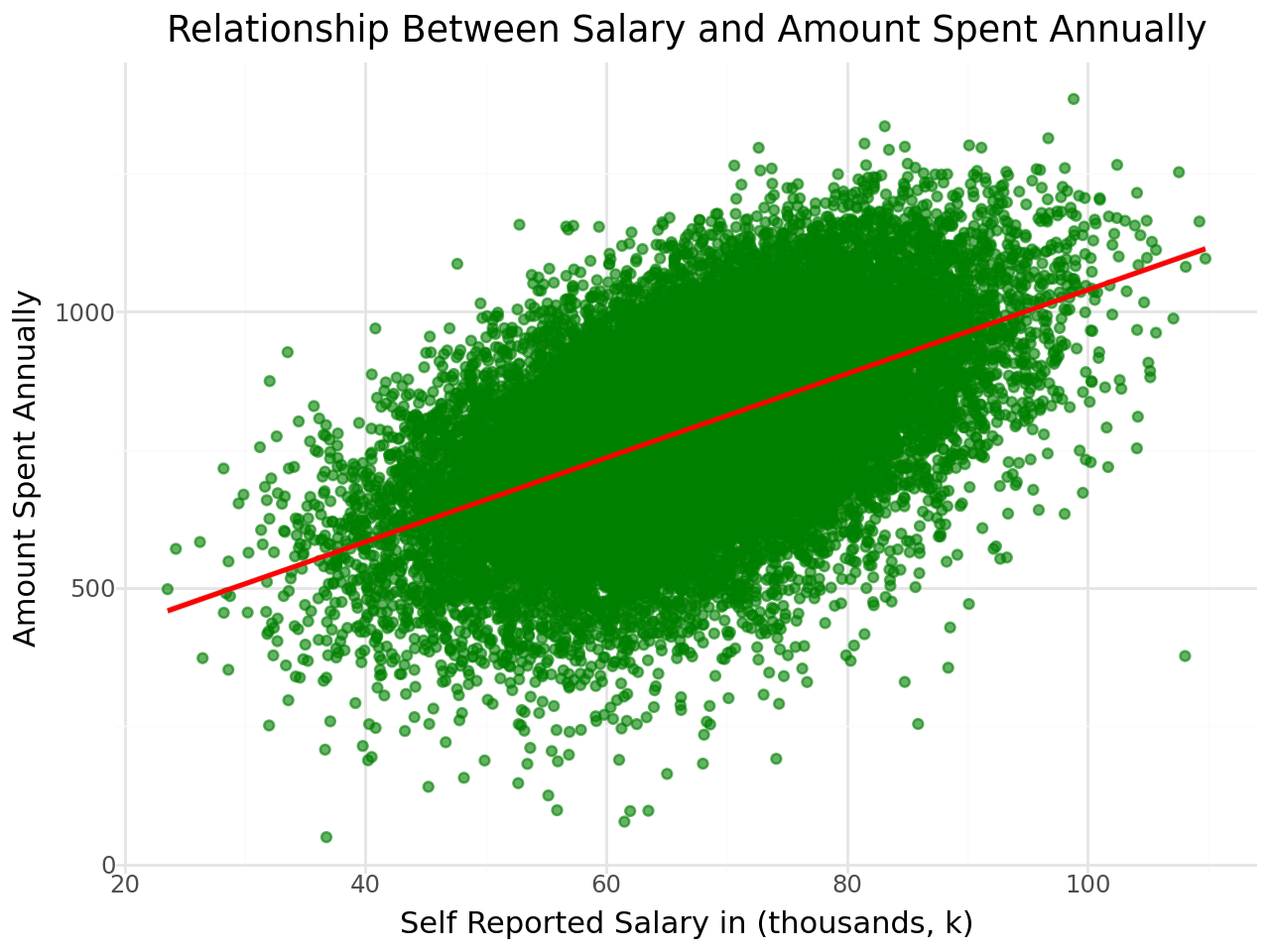
## 

## Question 2: Does having a higher salary increase the number of purchases made and total amount spent?



**Figure 2a:** Relationship Between Salary and Number of Purchases

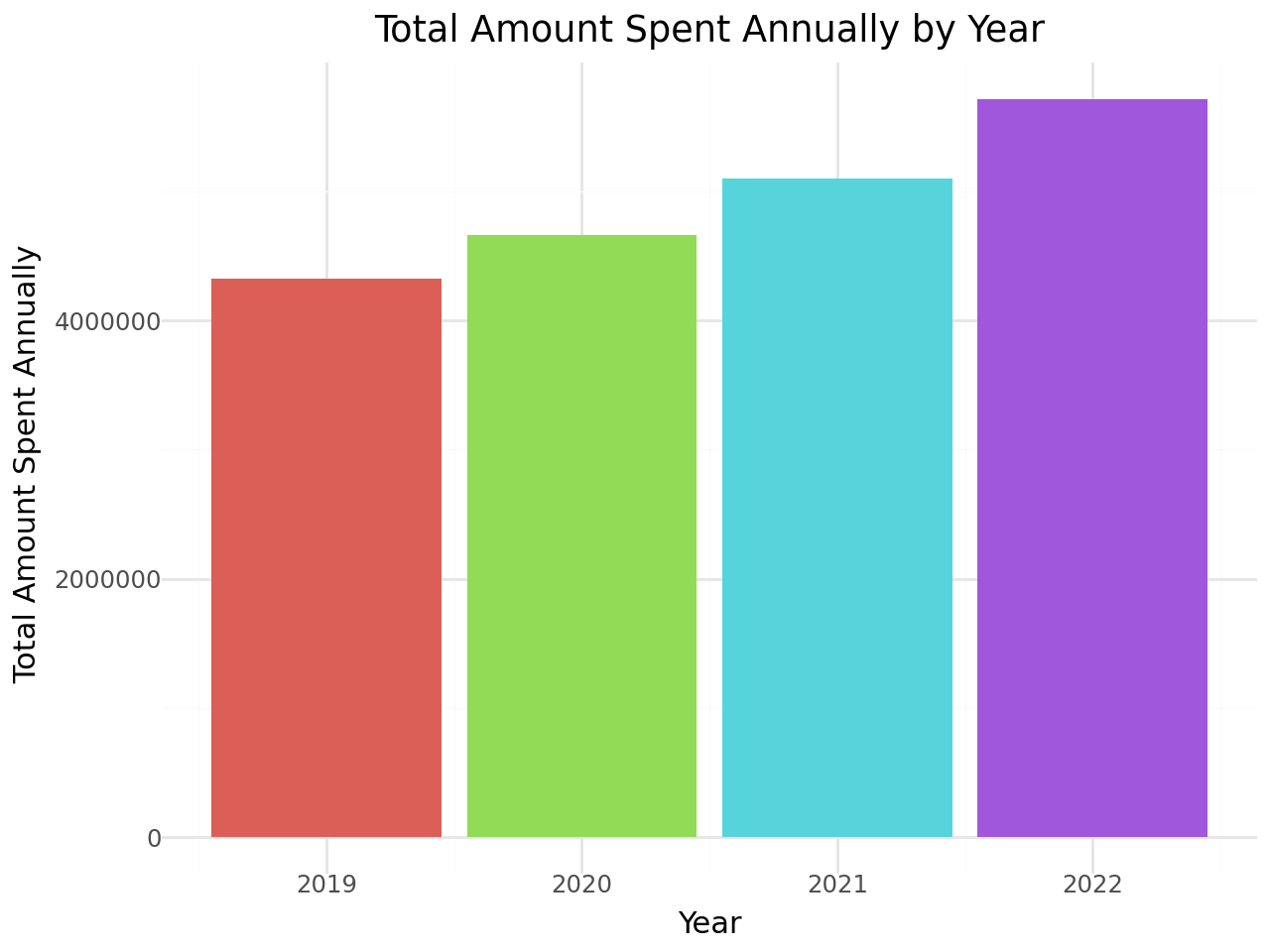
This scatter plot shows no clear linear trend between customers’ self-reported salary and the number of purchases they make, suggesting that higher income does not necessarily lead to more frequent transactions at the store.



**Figure 2b**: Relationship Between Salary and Annual Amount Spent

This scatter plot demonstrates a positive linear relationship between customers’ self-reported salary and the amount they spend annually at the store. The upward trend suggests that higher earners tend to spend more on clothing purchases each year.

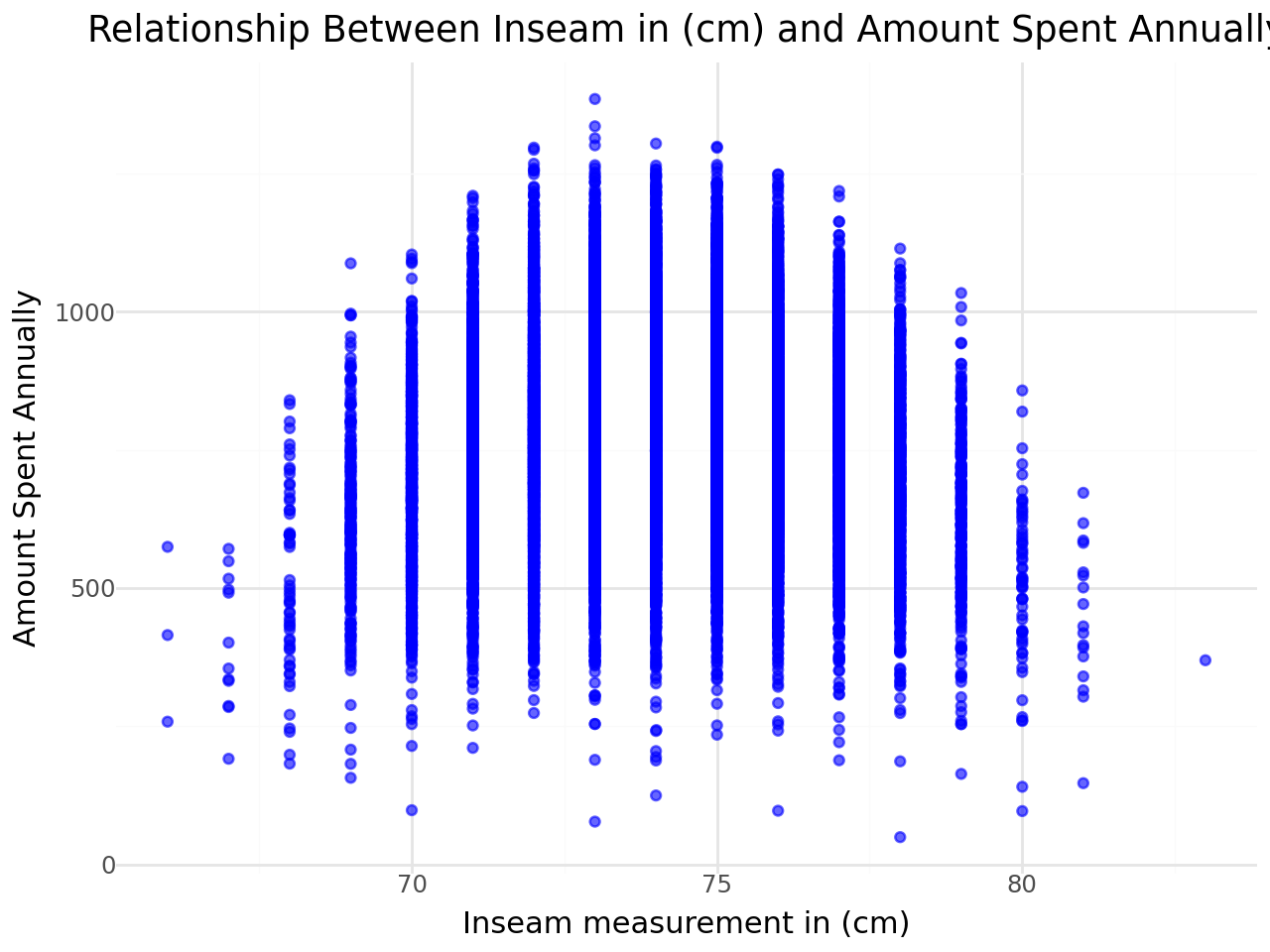
## Question 3: In which year did the store's customers make the most money? Were the store's sales highest in those years?



**Figure 3:** Total Amount Spent Annually by Year.

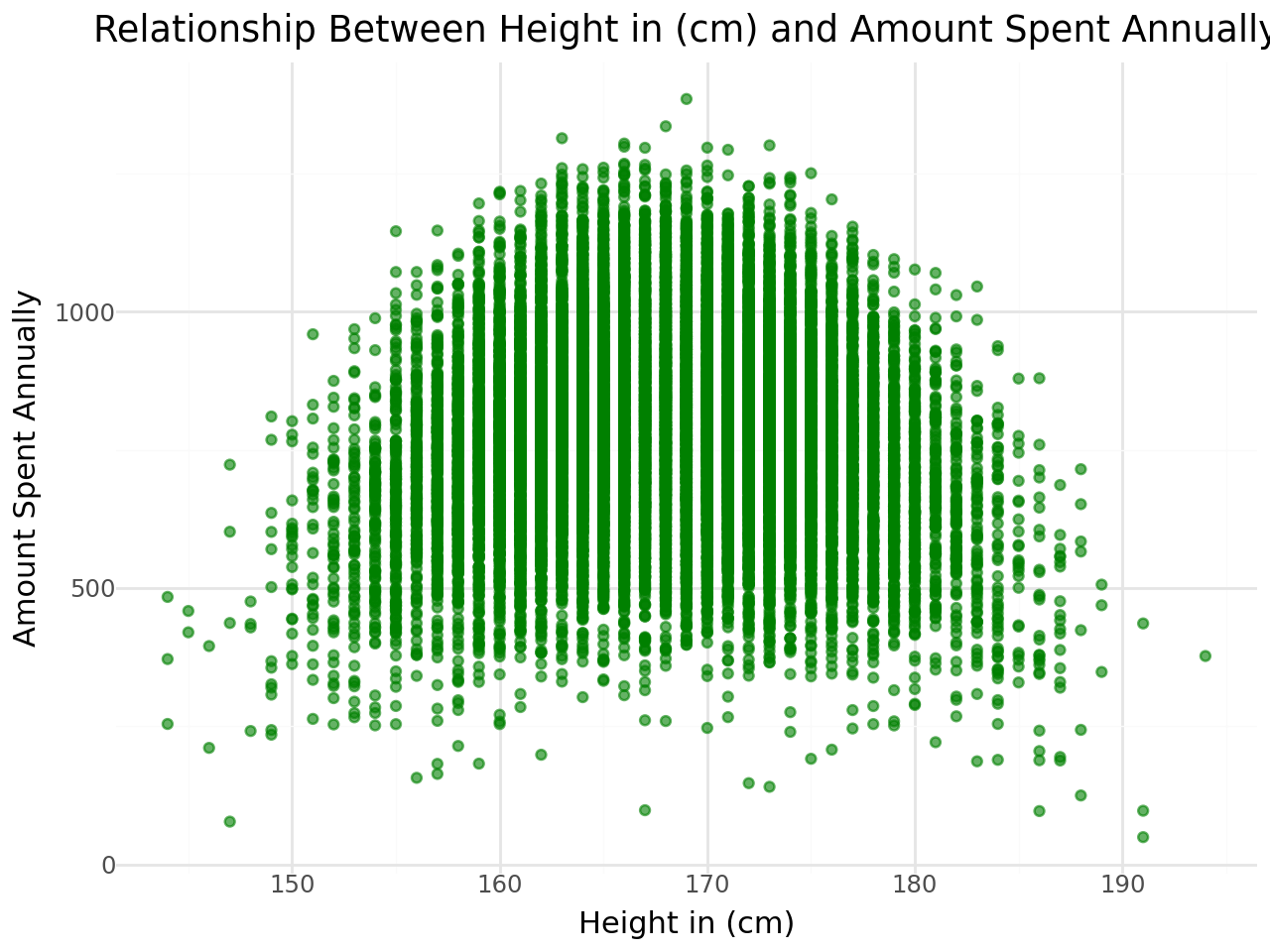
This bar chart shows the total annual spending by all customers for each year in the dataset. The data indicates that sales peaked in 2022, suggesting that this was the most profitable year for the store during the recorded period.

## Question 4: People who are not your “average” size find it difficult to buy clothes in traditional stores. Is there a relationship between inseam and amount spent? Is there a relationship between height and amount spent?



**Figure 4a:** Relationship Between Inseam Length and Annual Spending

This scatter plot shows that customers with inseam measurements near the average tend to spend more annually. Those with inseam lengths farther from the average generally spend less, suggesting potential difficulty in finding clothing that fits well.

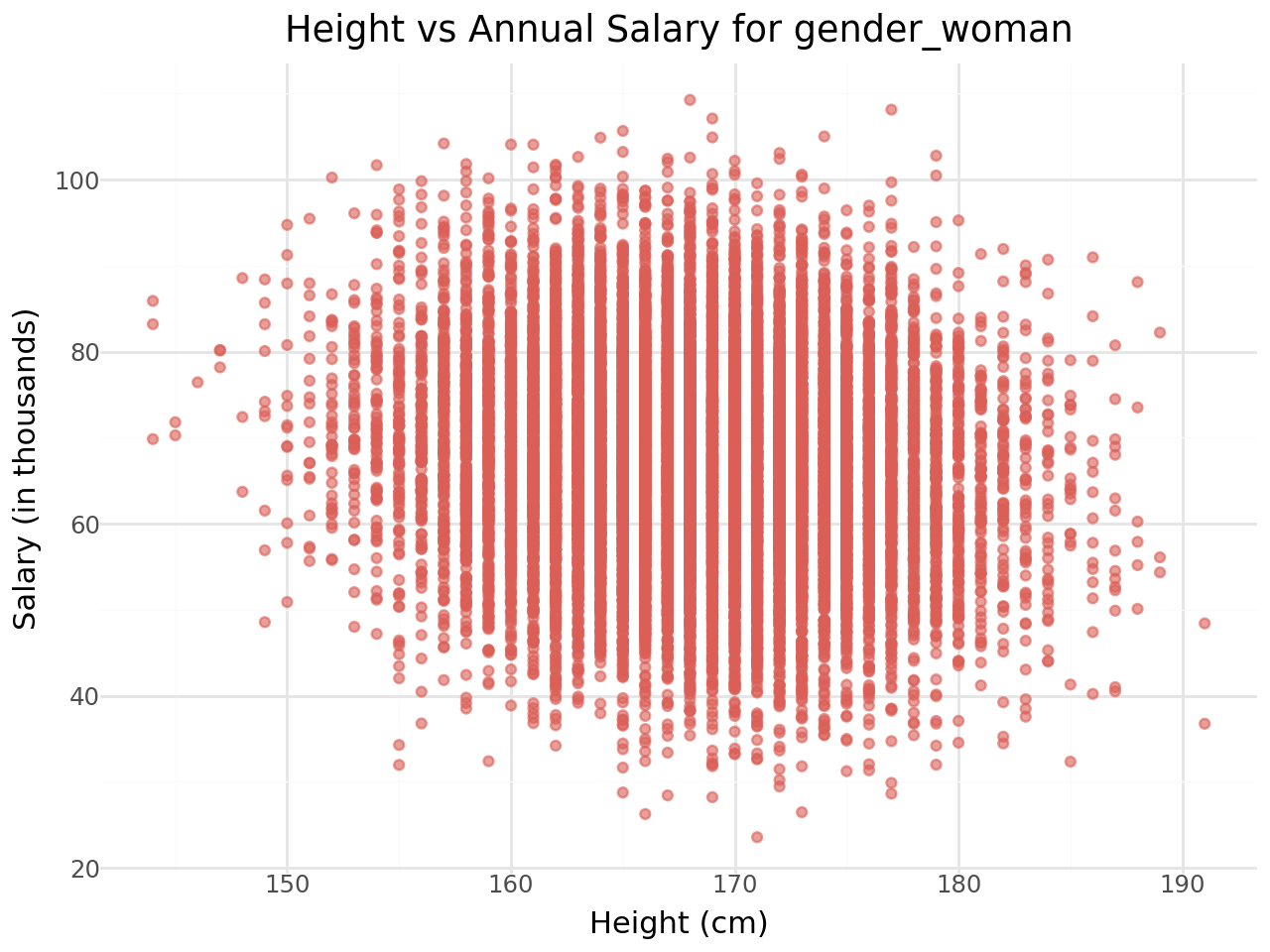
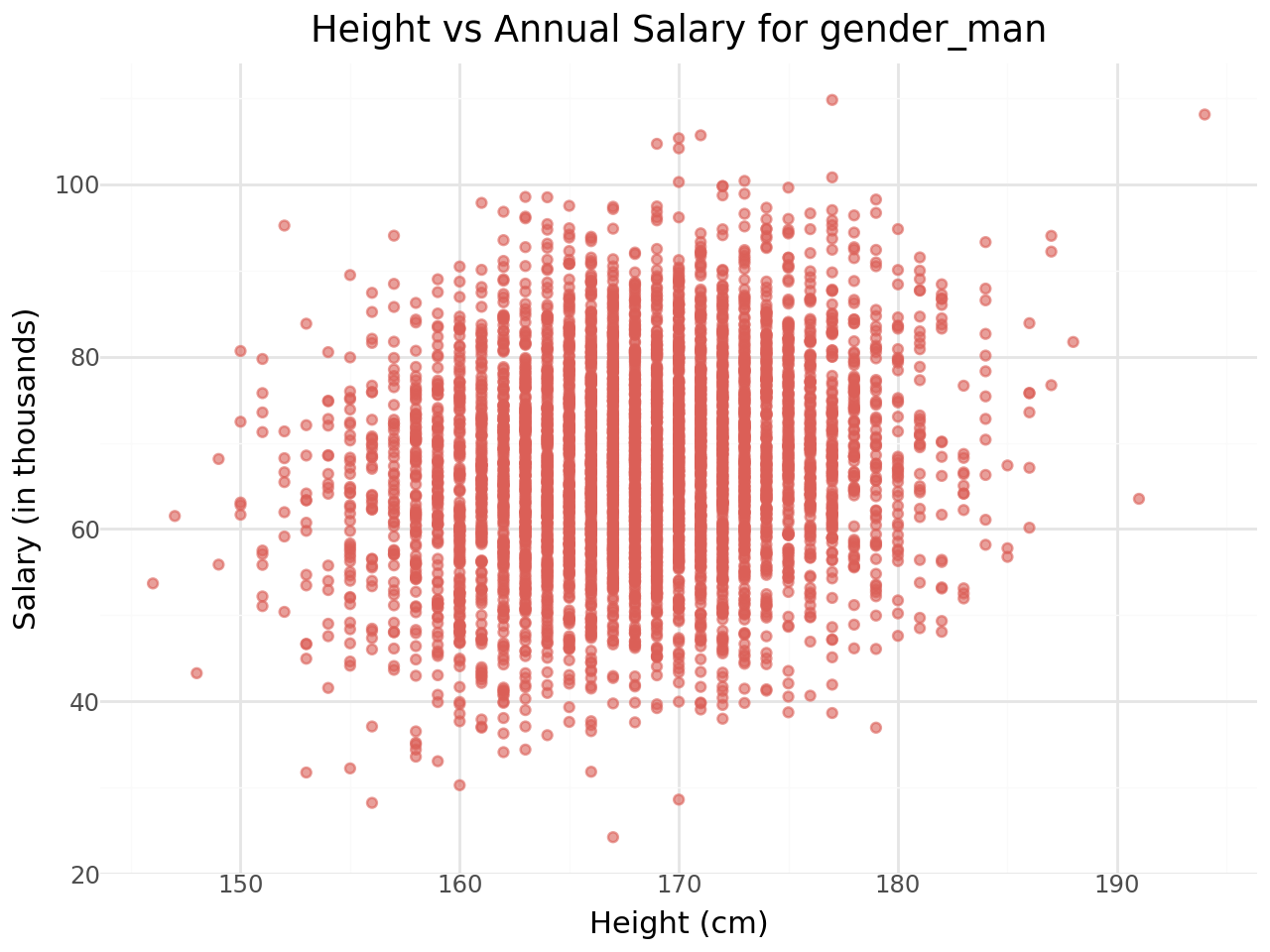


**Figure 4b:** Relationship Between Height and Annual Spending

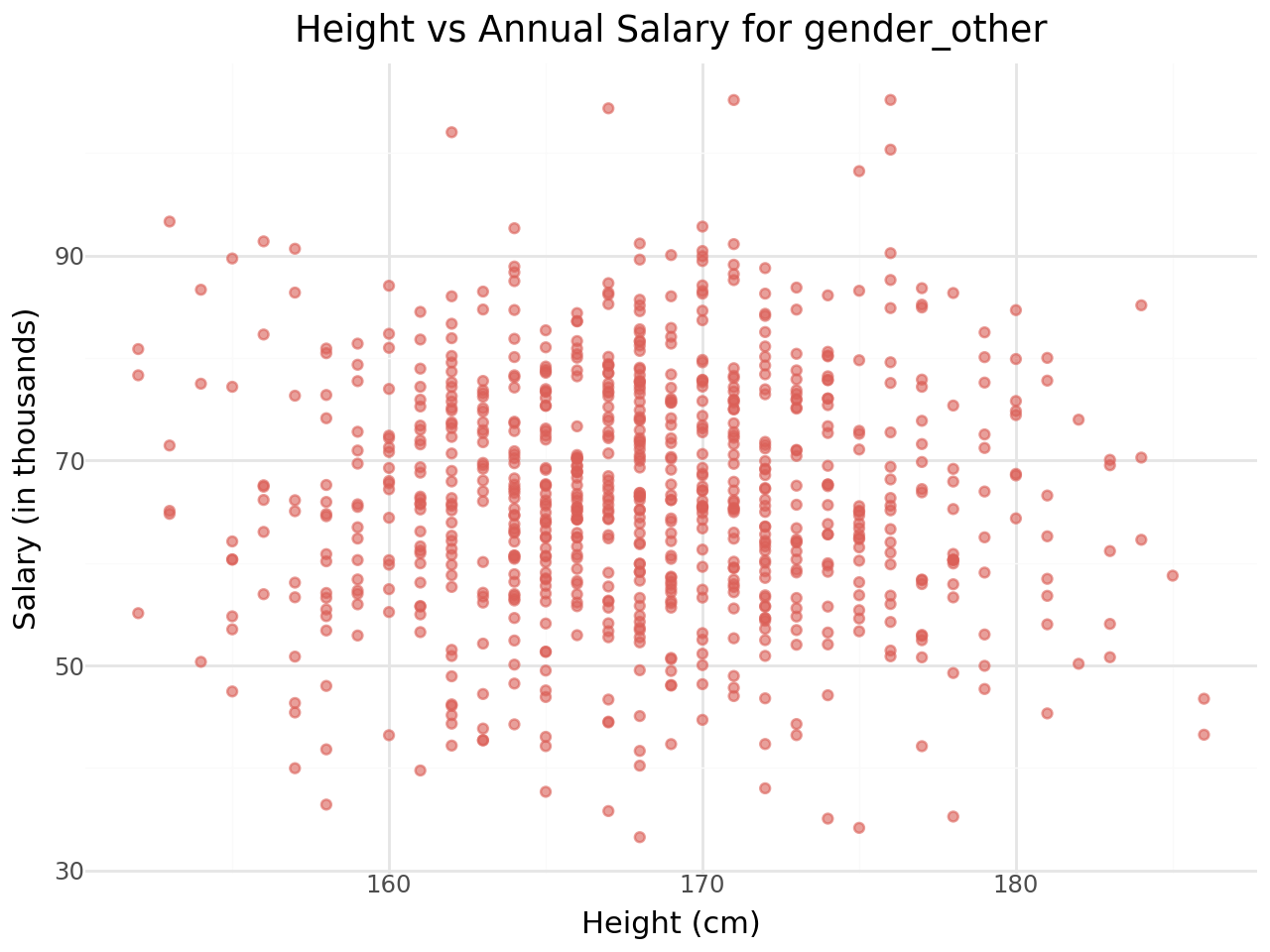
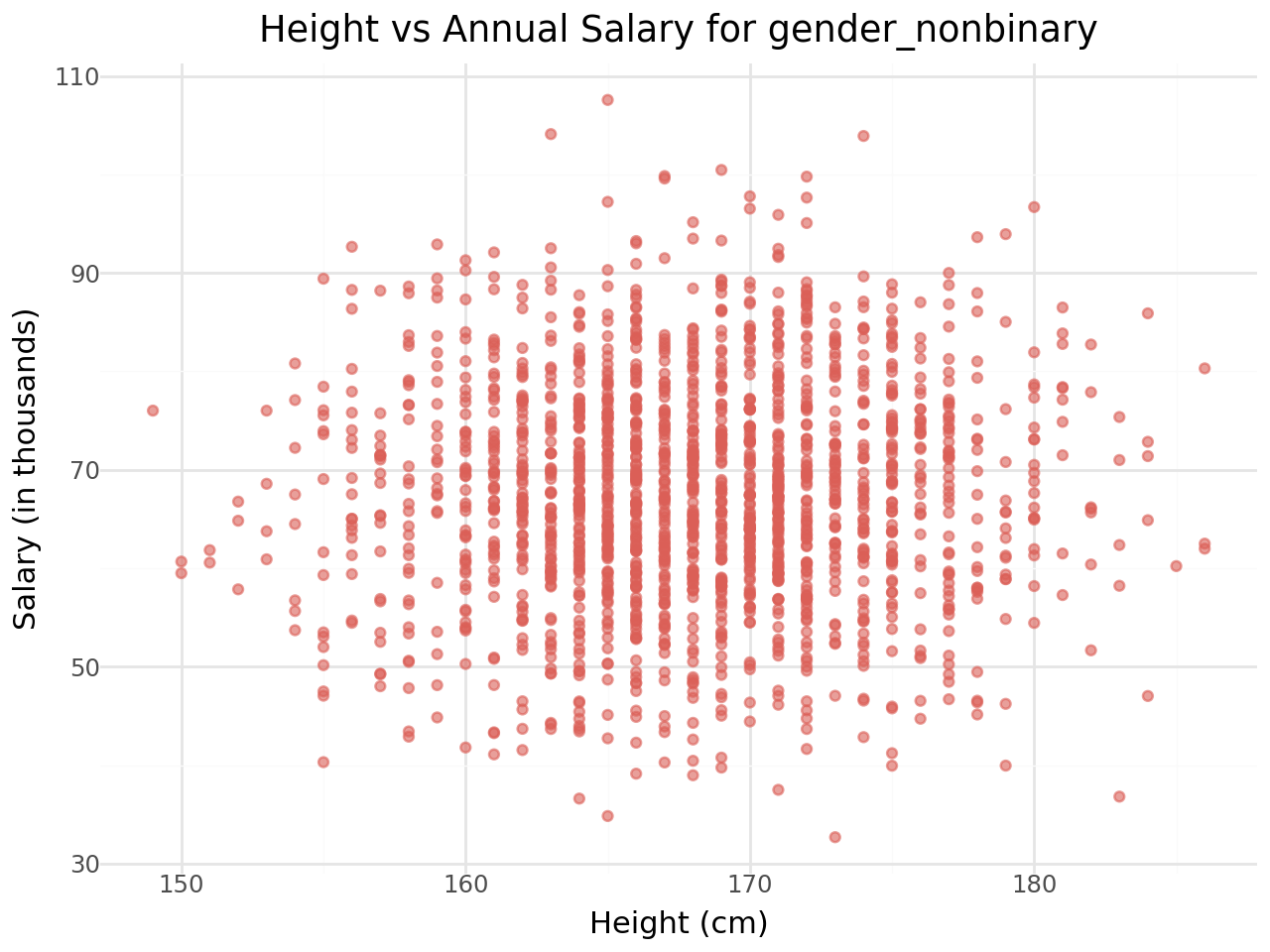
This plot indicates that customers with heights closer to the average spend more each year. Spending tends to decrease as height moves further from the average, highlighting possible challenges for customers outside typical height ranges when shopping.

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

**Figure 5a:** Height vs Salary by Men **Figure 5b**: Height vs Salary by Women



**Figure 5c:** Height vs Salary by Nonbinary **Figure 5d:** Height vs Salary by Other



Across all gender groups, the plots show no clear linear relationship between height and self-reported salary. The data points appear scattered without a consistent trend, indicating that height does not significantly impact salary within each gender category.

## Question 6: 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.

| **Year** | **Height Min (cm)** | **Height Max (cm)** | **Height Mean (cm)** | **Waist Min (cm)** | **Waist Max (cm)** | **Waist Mean (cm)** | **Inseam Min (cm)** | **Inseam Max (cm)** | **Inseam Mean (cm)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2019 | 144.0 | 191.0 | 167.96 | 63.0 | 121.0 | 80.50 | 67.0 | 81.0 | 73.99 |
| 2020 | 144.0 | 189.0 | 168.02 | 63.0 | 121.0 | 80.51 | 67.0 | 81.0 | 74.00 |
| 2021 | 145.0 | 189.0 | 168.03 | 64.0 | 123.0 | 80.52 | 66.0 | 83.0 | 73.99 |
| 2022 | 144.0 | 194.0 | 168.10 | 65.0 | 125.0 | 80.57 | 66.0 | 81.0 | 74.04 |

# Discussion/Reflection

From performing these analyses, I learned how consumer behavior at a clothing store may be more complex than simple linear associations. While initial assumptions might suggest that physical attributes such as height, waist size, or inseam would strongly predict annual spending (due to fit or sizing needs), the data revealed otherwise, salary stood out as the most informative predictor of spending. Interestingly, there was no consistent linear relationship between height and salary across gender groups, suggesting that socioeconomic patterns related to physical attributes may not generalize well across demographics.

This exercise also deepened my appreciation for exploratory data analysis (EDA) as a critical step before modeling. Patterns that appeared noisy or ambiguous in scatterplots were validated by weak performance in regression models, reminding me that model quality hinges on signal clarity in the data.

If I were to extend this work, I would consider creating derived features, such as a “fit ratio” combining height, waist, and inseam, or segmenting customers by body type or price sensitivity using clustering. I’d also bring in external socioeconomic data (e.g., occupation or zip code) to enrich our understanding of purchasing behavior and improve model performance.