*Correlation and regression*

**R**

**Practice**

Rp

ALY6010 Probability Theory and Introductory Statistics

Module 6 R Practice

**PREPERATION:**

By: John DiSessa

For: Professor Goulding

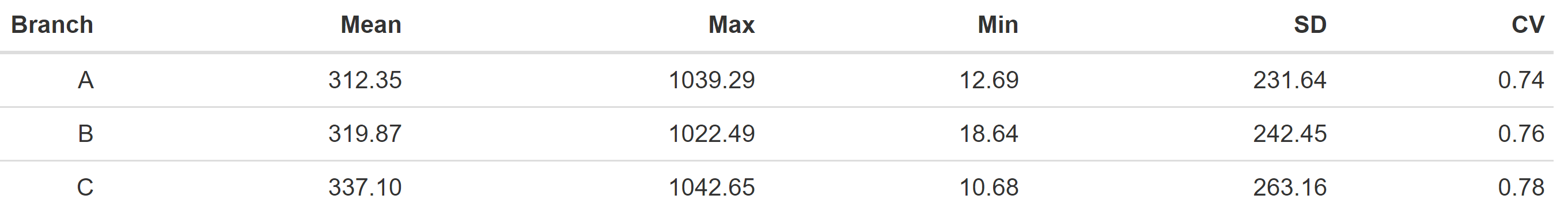
On: April 11th,2021

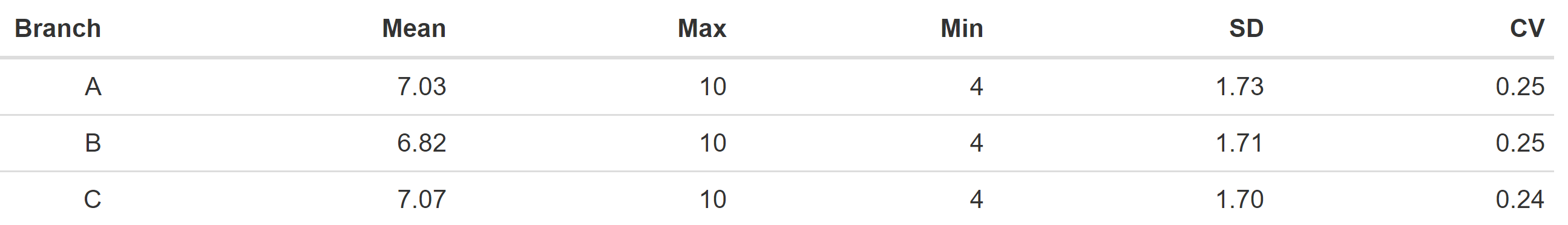
Introduction

I decided to analyze the sales of three different hardware store branches as a way of practicing using dummy variables, scatterplots, and regression lines. My goal is to represent the differences in total sales by each branch as a function of customer satisfaction rating as well as quantity of items.

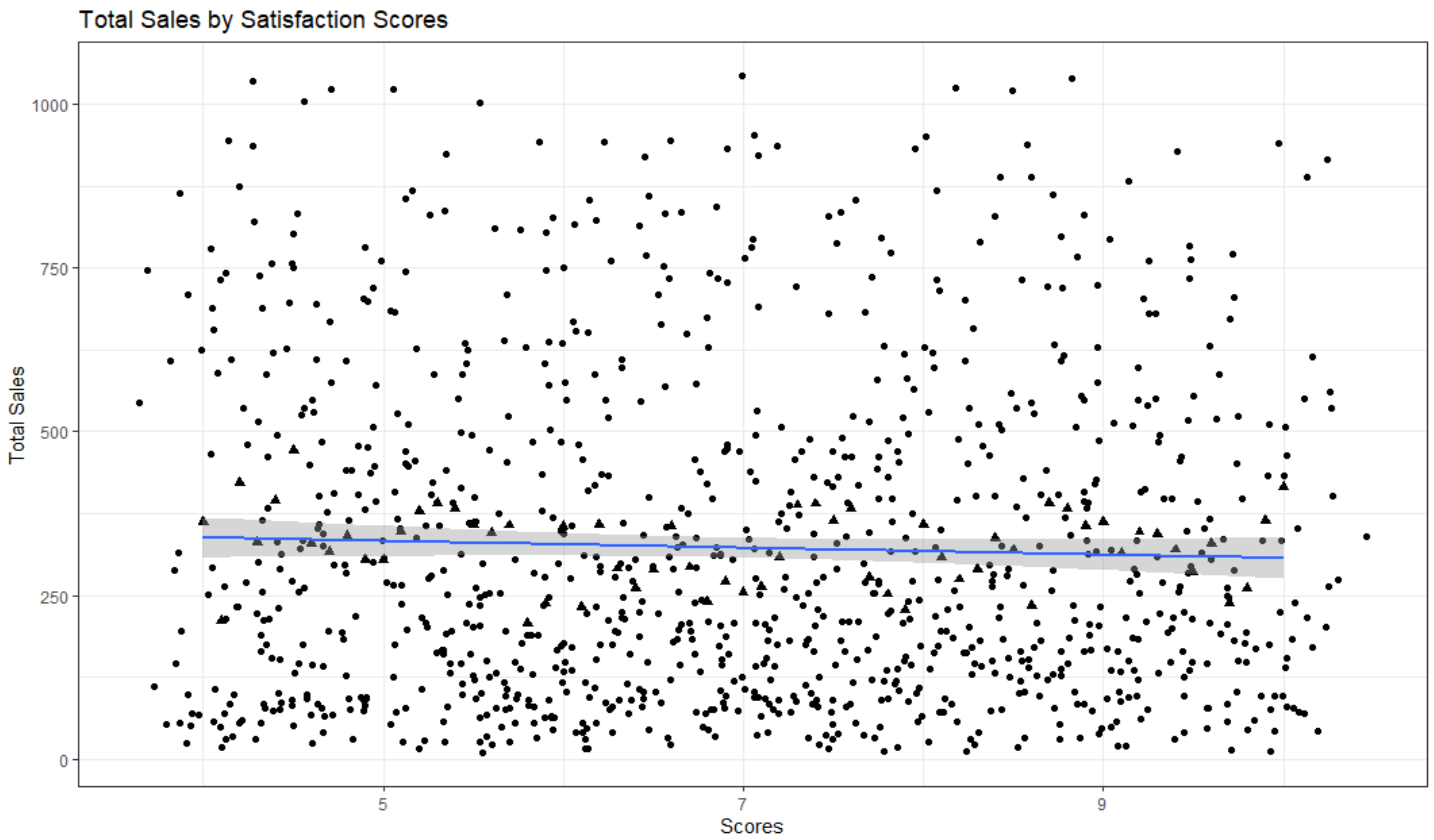
Analysis

I created the following summary statistics tables for the sales (table 1) and customer satisfaction scores (table 2) of each branch. Each branch seems to have similar sales data but it will be interesting to compare this data against customer satisfaction scores and sales quantity to see if higher scores mean more sales or if more items sold means higher sales. The important takeaway from the following tables are the columns “CV”. The coefficient of variations (sd/mean) normalizes the standard deviations so that anything below 1 means low variation, which we have here.





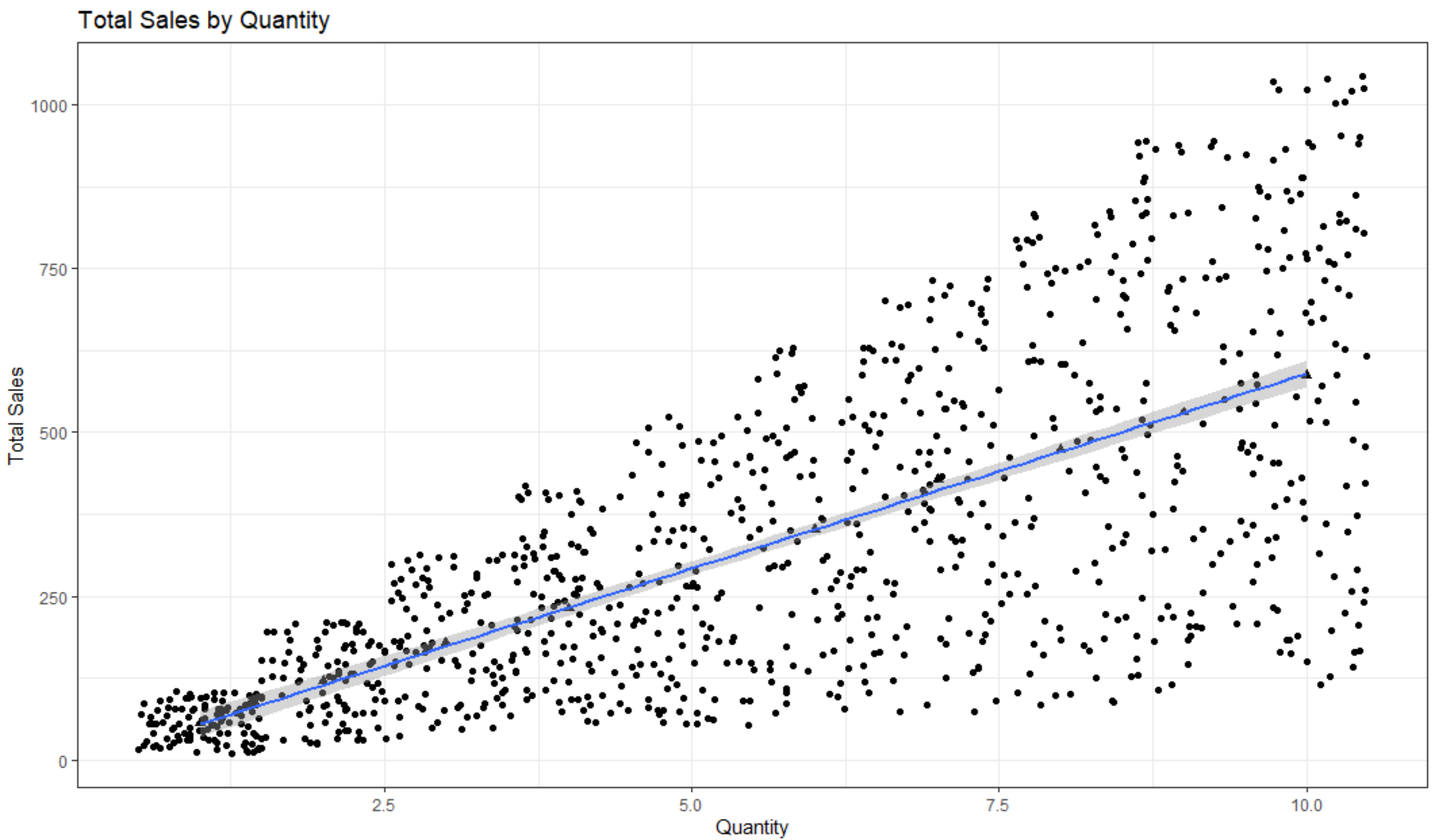
First, we will look at sales as a function of customer satisfaction scores. As you can see in the graph below, sales seem to slightly go down as customers are more satisfied. The correlation coefficient of sales and scores is -.03 so it is basically negligible. Even though there wasn’t a correlation, it is still useful to know that customer satisfaction scores don’t impact sales.



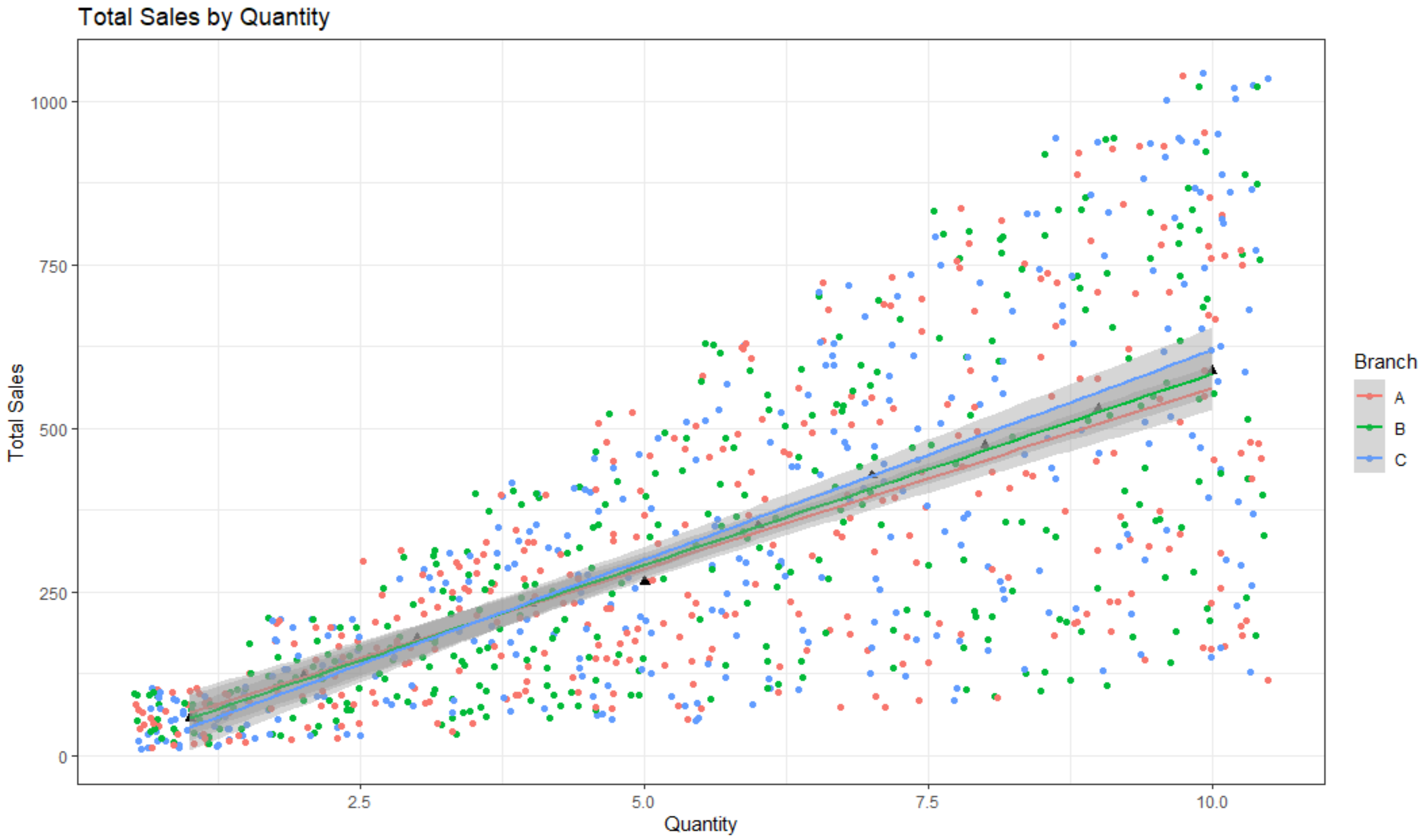
Next, we will look at sales compared to items sold in order to see if these stores sell more lower priced items or fewer higher priced items. Sales and quantity are strongly correlated with an r value of .71. As you can see from the graph below this chain’s business model is geared towards selling many lower priced items with smaller margins rather than large margin items. I also conducted simple linear regression to create a model that can predict sales based on quantity of items. With a p-value of 0, the following model is statistically significant.

Sales = -4 + 59.3\*(Quantity) ± 13.7

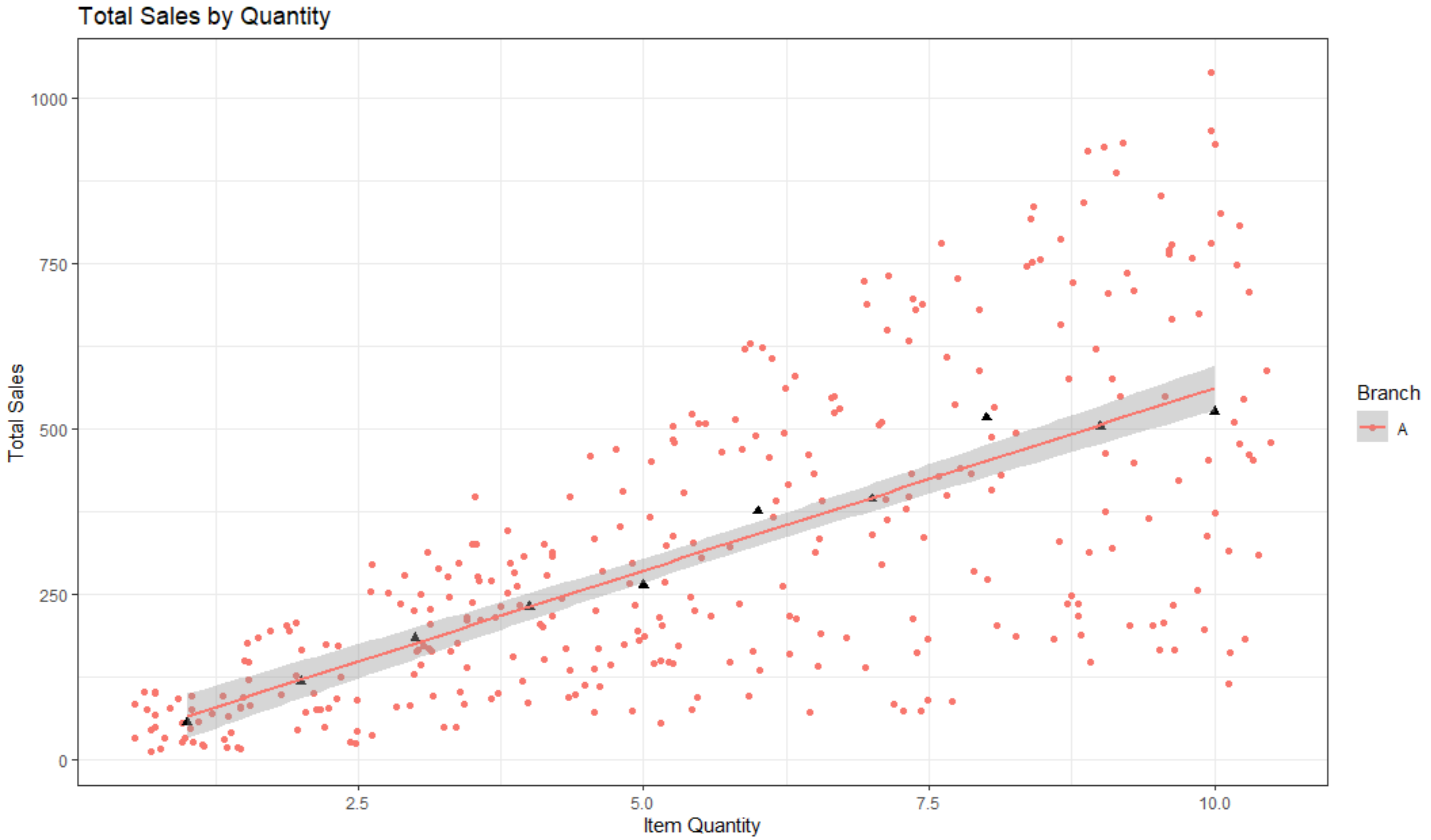
The adjusted R2 value is .4972 so 49.72% of the variation in Sales is explained by Quantity.

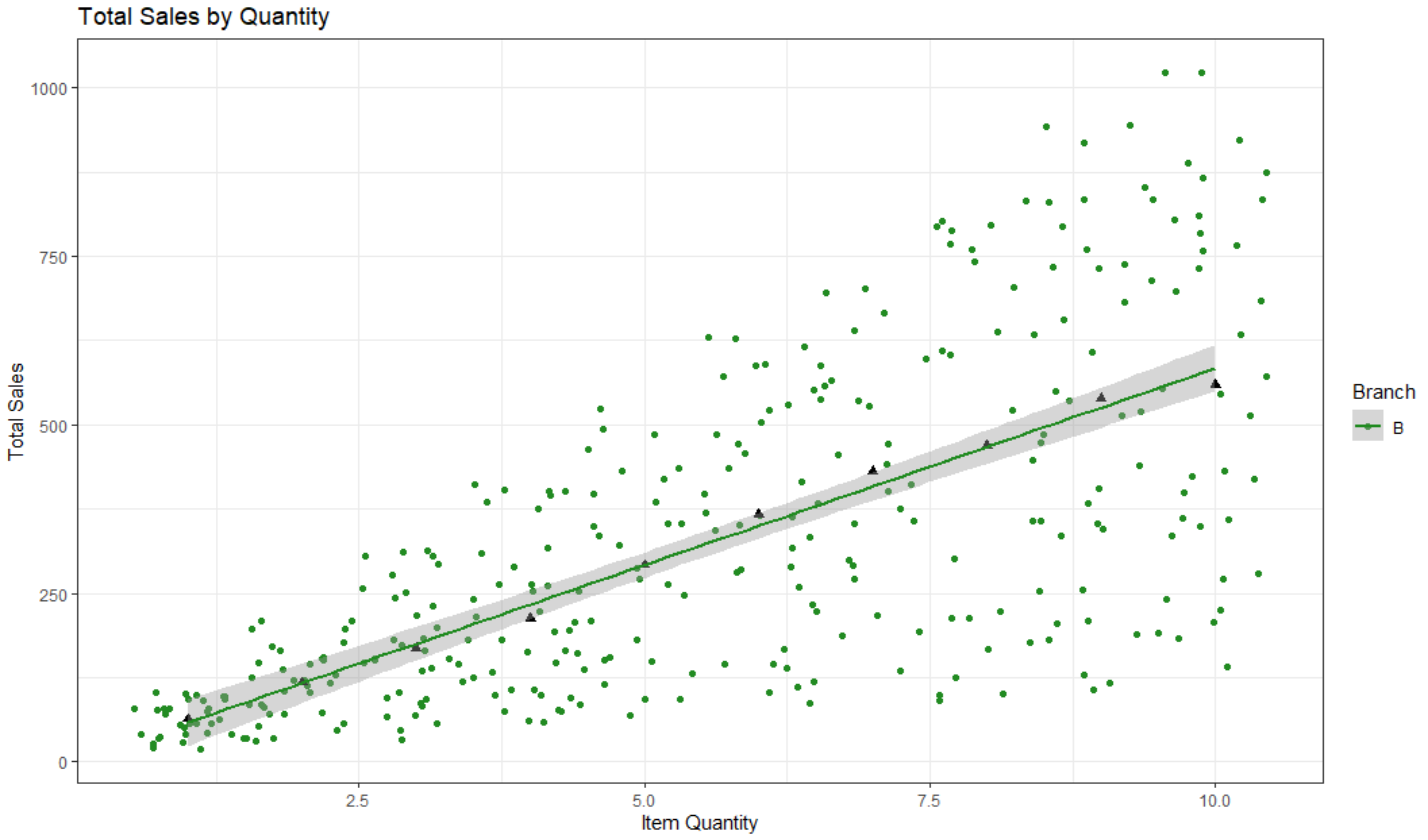


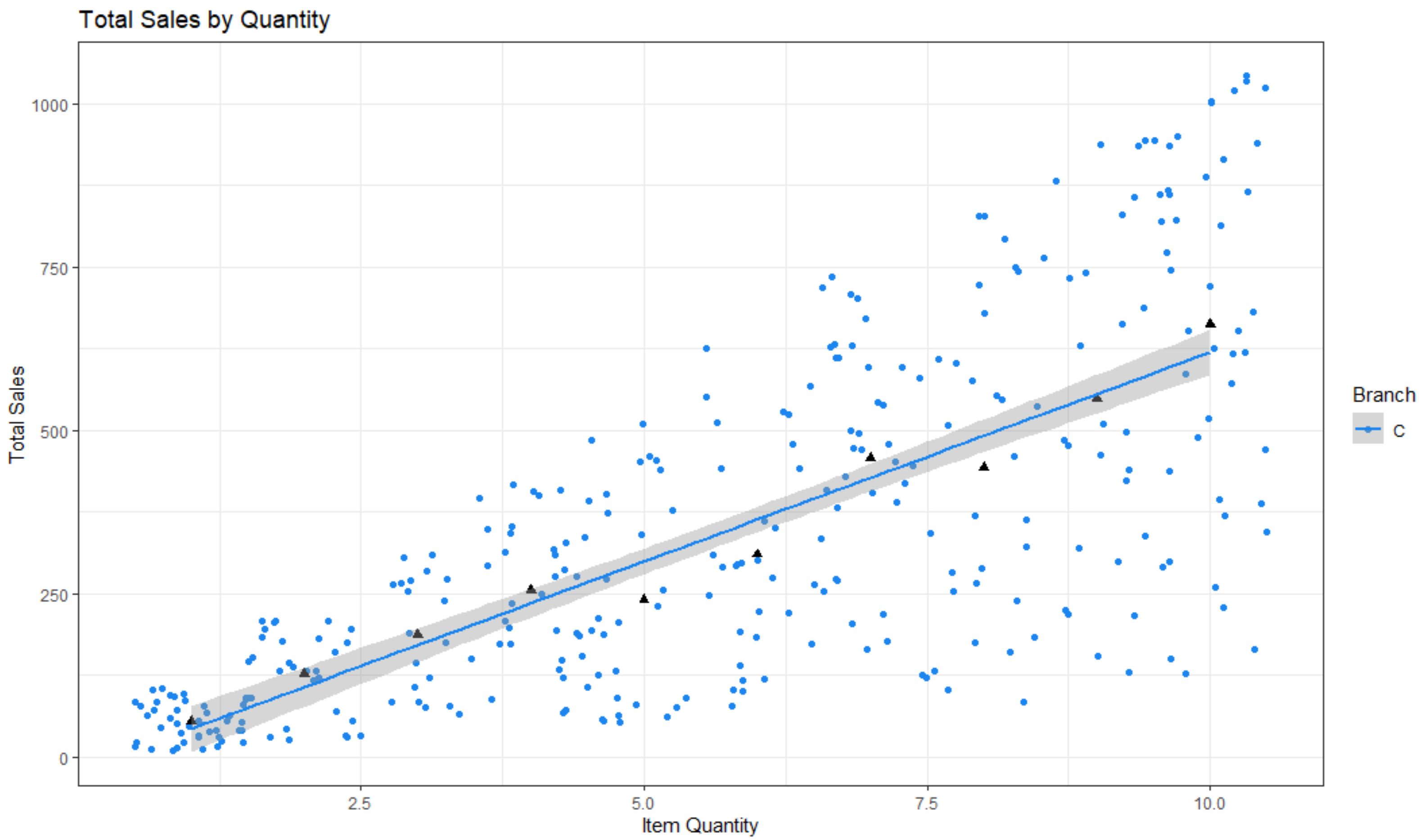
I then wanted to see if there were any noticeable differences between the branches, so I created the graph below.



Even though the points on the scatter plot seem equally scattered, the regression lines have noticeable differences between each branch. The respective correlation coefficients between Sales and Quantity for branches A, B, and C are .68, .70, and .74 so each has strong and positive correlation. I created subsets for each branch and plotted their sales data in order to highlight any differences between. As seen in the graphs below each branch has almost identical sales patterns.







I then conducted three independent two sample t-tests (95% confidence levels) to see if the differences in mean sales data are significant or not. I compared the mean sales data of branch A to B, B to C, and A to C with the following null and alternative hypotheses.

H0: µ1 = µ2

Ha: µ1 ≠ µ2

All of them had p-values greater than .05 so I failed to reject the nulls and conclude that there are no statistically significant differences in sales data between the branches.

Summary

Most of my work for ALY6010 have had goals to find statistically significant differences. However in the real-world, finding out there are not statistically significant differences can be just as telling. Through my analysis of three different branches of hardware stores, I found out they have basically identical sales. If I were an executive for the corporate chain I would be very happy to see that independently operated branches all follow the same strategy and achieve the same results. I would also be intrigued that customer satisfaction ratings do not impact sales. If these hardware stores are located in populated areas, customers may shop out of convenience or price instead of brand loyalty. This is very useful information for the executives to focus their business model on.

Citations

Pyae, Aung. “Supermarket Sales.” Kaggle, 27 May 2019, www.kaggle.com/aungpyaeap/supermarket-sales.

Kassambara, et al. “Regression with Categorical Variables: Dummy Coding Essentials in R.” STHDA, 11 Mar. 2018, www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/.

Moon, Keon-Woong. GgPredict() - Visualize Multiple Regression Model, 6 Oct. 2020, cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html.

Kassambara, et al. “Multiple Linear Regression in R.” STHDA, 10 Mar. 2018, www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/.

“Adding a Regression Line on a Ggplot.” Stack Overflow, 1 Jan. 1962, stackoverflow.com/questions/15633714/adding-a-regression-line-on-a-ggplot.

“How to Change Default Color of Points in ggplot2 Conditional Aes?” Stack Overflow, 1 Nov. 1967, stackoverflow.com/questions/54208043/how-to-change-default-color-of-points-in-ggplot2-conditional-aes.