STATISTICAL ANALYSIS OF THE STORE SALES DATA

Filza Aqeel	30134	2:30		F.Aqeel.30134@khi.iba.edu.pk
Hania Raza	30208	2:30		H.Raza.30208@khi.iba.edu.pk

Contents

1. Introduction

1.1 Store Profit Defined

Store profit refers to the financial gain a retail outlet earns after subtracting the total costs of its operations from its revenue. In the context of this research, profit is analyzed as a function of two key business indicators—sales and quantity of items sold—within the consumer segment of retail stores across various U.S. states. Understanding the distribution and drivers of profit is essential for businesses to make informed decisions about sales strategies, inventory management, and resource allocation. Through statistical and visual tools, this report aims to explore how profit varies across different regions and how it can be predicted using store-level sales and quantity data.

1.2 Research Report Outline

This research report focuses on analyzing and comparing store profits, specifically filtered for the consumer segment, across five U.S. States:

1. Alabama
2. Arizona
3. Arkansas
4. California
5. Colorado

The two key methodologies employed in this report are:

Comparative Descriptive Analysis: This section presents descriptive statistics, histograms, and boxplots to explore and compare the distribution of store profits across the selected five states. This provides insights into the average performance, variation, and outliers in store-level profit within each state.

Regression Analysis: This section conducts two separate regression analyses to study the relationship between profit (Y) and two independent variables: sales (X1) and quantity (X2). Predictive models are developed using a training dataset of selected stores, and predictions are then tested on out-of-sample data (i.e., stores not included in the training set). Results from these models are compared with actual profits to evaluate the effectiveness of each predictive variable.

2. Comparative Descriptive Analysis of Store Profit

	Alabama	*Arizona*	*Arkansas*	*California*	*Colorado*
Mean	68.443756	-7.31009	66.96345	22.1838	18.13736
Standard Error	18.807041	14.50772	17.80461	5.318481	13.53719
Median	19.8072	-1.35	18.7812	11.7741	9.2386
Standard Deviation	94.035204	72.53862	89.02305	26.59241	67.68594
Sample Variance	8842.6197	5261.851	7925.103	707.1561	4581.386
Kurtosis	1.0256032	4.84525	1.502319	1.768931	6.559041
Skewness	1.490143	-0.99717	1.55793	1.713946	0.928779
Range	306.4894	383.6346	316.5115	88.7496	409.871
Minimum	0	-204.446	1.7901	1.9656	-161.875
Maximum	306.4894	179.1888	318.3016	90.7152	247.996
Sum	1711.0939	-182.752	1674.086	554.595	453.434
Count	25	25	25	25	25
Coefficient of Variation	137%	-992%	133%	120%	373%

The Descriptive Statistics table compares values of store-level profits for five U.S. states: Alabama, Arizona, Arkansas, California, and Colorado. Of all the measures presented in the table, our analysis focuses on comparing the values of:

1. Mean
2. Median
3. Standard Deviation
4. Skewness
5. Coefficient of Variation

2.1.1 Mean of Store Profit

The mean represents the average value and is a fundamental measure of central tendency. According to Table 1, Alabama has the highest mean (68.443756), followed closely by Arkansas (66.96345). Arizona displays a highly unusual negative mean (-7.310009), possibly due to outliers or data anomalies. California (22.1838) and Colorado (18.13736) fall in the lower middle range.

2.1.2 Median of Store Profit

Unlike the mean, the median offers a center value that is not skewed by extreme data points. Table 1 shows that Alabama has the highest median (19.8072), followed by Arkansas (18.7812) and California (11.7741). Arizona's median is the lowest (-1.35), reflecting its negative distribution. Colorado's median (9.2386) also falls on the lower end.

2.1.3 Standard Deviation of Profit

Standard deviation captures the spread of data around the mean. Alabama exhibits the highest standard deviation (94.035204), indicating wide variability in its dataset. Arkansas (89.02305) and Arizona (72.53862) also show considerable dispersion. California has the lowest variability (26.59241), while Colorado (67.68594) shows a relatively high level of spread, though not as extreme as Alabama or Arkansas.

2.1.4 Skewness of Profit Distribution

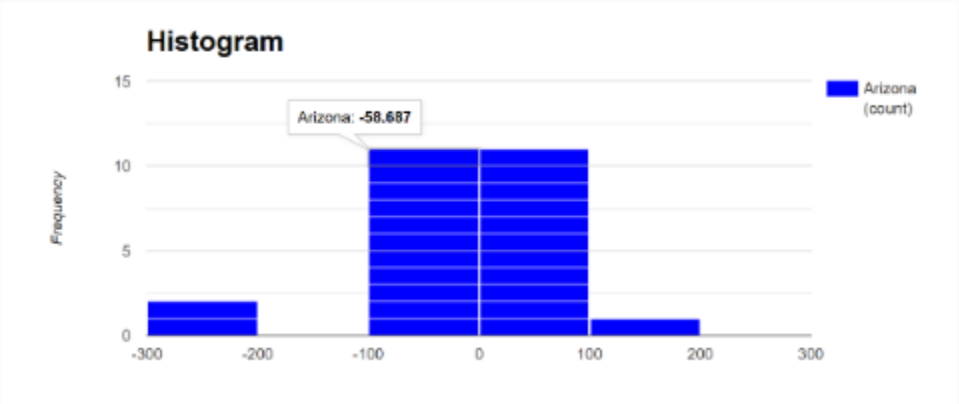
Skewness measures the asymmetry of data. A value above +0.5 indicates positive skew, while below -0.5 indicates negative skew. Arizona's data is negatively skewed (-0.99717), suggesting a longer left tail. All other states show positive skewness, with California (1.713946) and Arkansas (1.55793) displaying the strongest right-skew, followed by Alabama (1.490143) and Colorado (0.928779).

2.1.5 Coefficient of Variation

CV compares variability relative to the mean. Colorado has the highest CV (373%), followed by Alabama (137%) and Arkansas (133%), indicating substantial inconsistency. Among states with valid positive means, California (120%) exhibits the least variation. Arizona's CV is -992%, which is mathematically invalid due to its negative mean and should be interpreted with caution.

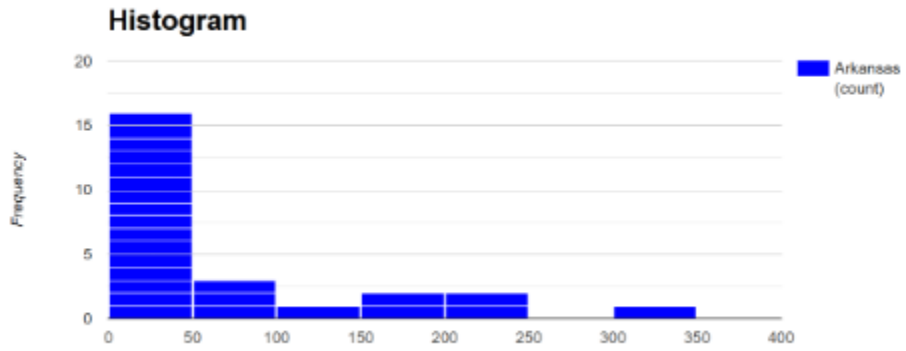
2.2 Histograms

2.2.1 Arizona



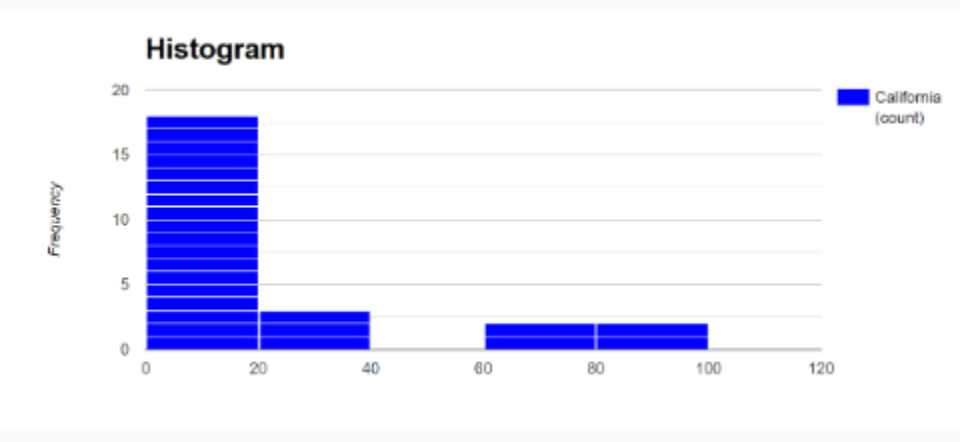
The histogram for Arizona displays a roughly symmetrical distribution with a mean close to zero, ranging approximately from -250 to 250. The highest frequency of data points falls between -100 and 100, suggesting that most transactions result in moderate profits or losses. However, the presence of a few extreme negative values indicates occasional significant losses. This indicates instability in profit margins across Arizona, potentially due to fluctuating costs or inconsistent demand.

2.2.2 Arkansas



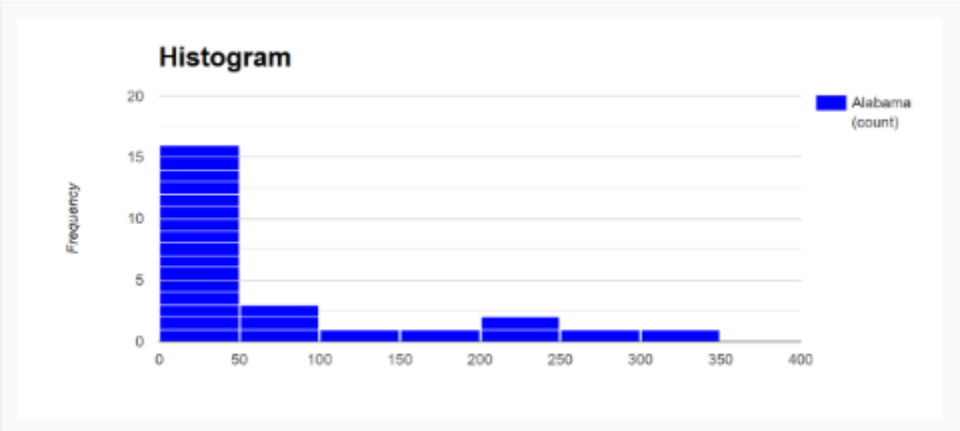
Arkansas exhibits a positively skewed distribution, where the bulk of profits are concentrated in the lower range (0–50), with fewer but significant occurrences stretching up to around 350. This suggests that while most transactions yield small profits, there are occasional large gains, likely driven by high-value products or successful campaigns. The skewness implies that performance is driven by a few high-profit sales, making the overall revenue somewhat dependent on outliers.

2.2.3 California



California's histogram is heavily right-skewed, indicating that the majority of profit values lie between 0 and 20, with a sharp drop-off beyond that. The extremely high frequency in the first bin suggests that most transactions yield minimal profit, and only a small number result in moderate to high gains. This pattern reflects a high volume, low-margin business model, possibly suggesting operational efficiency but limited pricing power or competitive pressure.

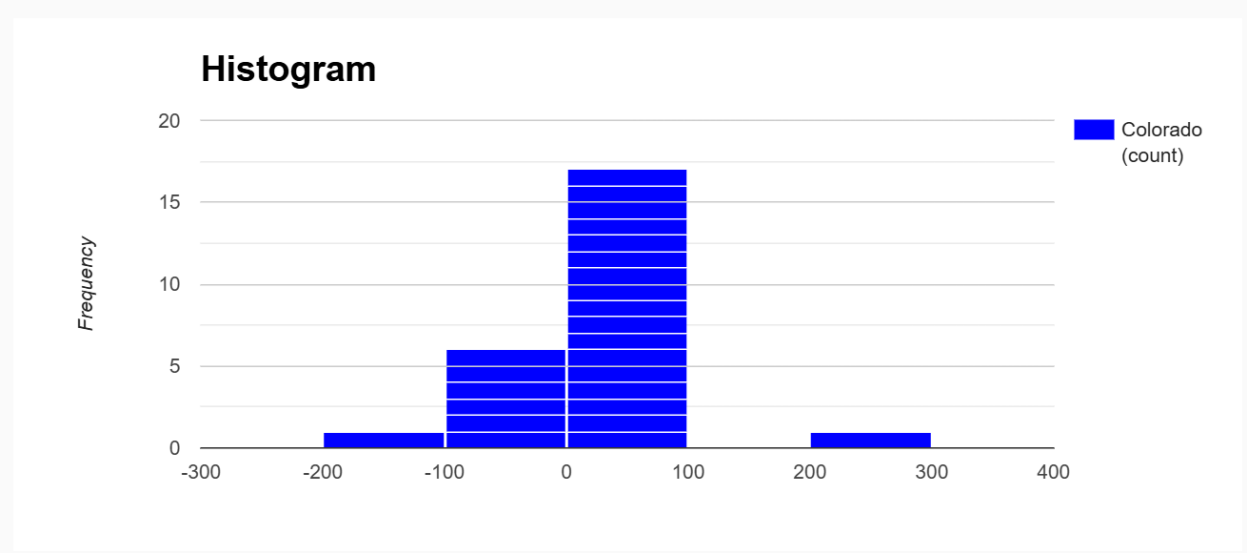
2.2.4 Alabama



Similar to Arkansas and California, Alabama shows a positively skewed distribution. Most profits fall below 50, but the tail extends to 350, pointing to occasional high-profit transactions. This implies that while day-to-day sales remain modest, there are rare instances of exceptional

profit. The spread hints at potential opportunity, but the business may not be consistently capitalizing on high-margin sales.

2.2.5 Colorado

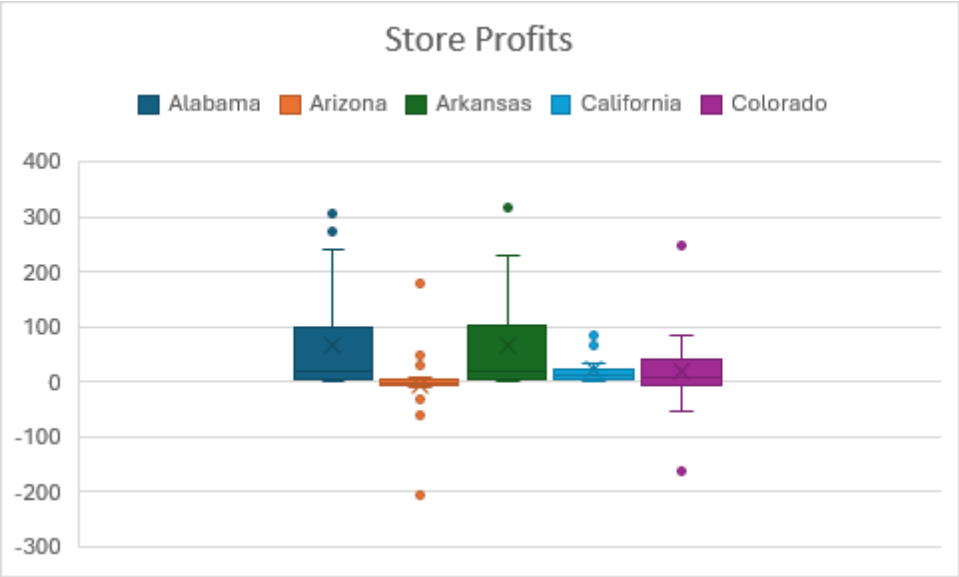


The histogram above illustrates the frequency distribution of store profits in Colorado. Most values fall between 0 and 100, indicating that the majority of stores in Colorado generate moderate profits. The distribution is positively skewed, with a longer tail on the right side due to a few stores achieving profits well above 200 units, while a smaller number of stores fall into the negative profit range below -100.

This positively skewed pattern suggests that while many stores perform reasonably well, a few high-performing stores significantly boost the upper end of the profit range. However, the presence of negative profit values also highlights the risk of losses, indicating that store performance is not uniformly stable across the state.

Overall, Colorado's histogram reflects a concentration of mid-level profits, with some extremes on both ends of the spectrum.

2.3 Box and Whiskers



The box and whisker plot above visually compares the distribution of store profits across five states: Alabama, Arizona, Arkansas, California, and Colorado. From the plot, it is evident that Alabama and Arkansas have the highest upper outliers, reaching above 300 units, indicating potential for extremely high-performing stores. Arizona, on the other hand, stands out for its negative outliers as low as -250, reflecting significant losses in some stores.

The interquartile range (IQR)—represented by the width of each box—is largest for Alabama and Arkansas, signaling greater variability in store performance within these states. California shows the smallest IQR, suggesting more consistent store profit outcomes.

Arizona's data distribution is negatively skewed, with the median closer to the upper quartile and multiple outliers on the lower end. In contrast, Arkansas and Alabama show positive skewness, with long upper whiskers and higher upper outliers. California and Colorado appear more symmetrical, though Colorado still displays a few high and low outliers.

Overall, the plot highlights substantial variation in store profitability across states, with Alabama and Arkansas offering both high opportunity and high risk, while California presents the most stable performance pattern.

2.4 Comparative Analysis of Profit: Insights from Descriptive Statistics

The comparative analysis of the five states shows that Alabama and Arkansas have the highest means, indicating better performance overall, while Arizona has a negative mean and median, suggesting poor outcomes across the data points. California and Colorado fall in the mid-range but differ in consistency—California is more stable with the lowest standard deviation, while Colorado shows high variability.

Skewness and kurtosis indicate that most states are positively skewed with heavy tails, especially Colorado, which also has the highest kurtosis (6.56), suggesting frequent outliers. Arizona is negatively skewed, which aligns with its low central tendency values.

The coefficient of variation highlights Colorado's extreme inconsistency (373%), whereas California has the most compact and reliable dataset. Overall, Alabama and Arkansas show strong but volatile trends, while Arizona clearly underperforms with high volatility in negative values.

3. REGRESSION ANALYSIS

This section of the report investigates the relationship between a store's profit and two key variables: Sales and Quantity using separate simple regression models. The data consists of selected consumer-segment stores across five US states from the Sample Superstore dataset.

The analysis aims to understand whether higher sales or a greater number of items sold contribute significantly to store profitability. To assess this, two linear regression models are estimated independently: one with Sales as the predictor of Profit, and the other with Quantity as the predictor.

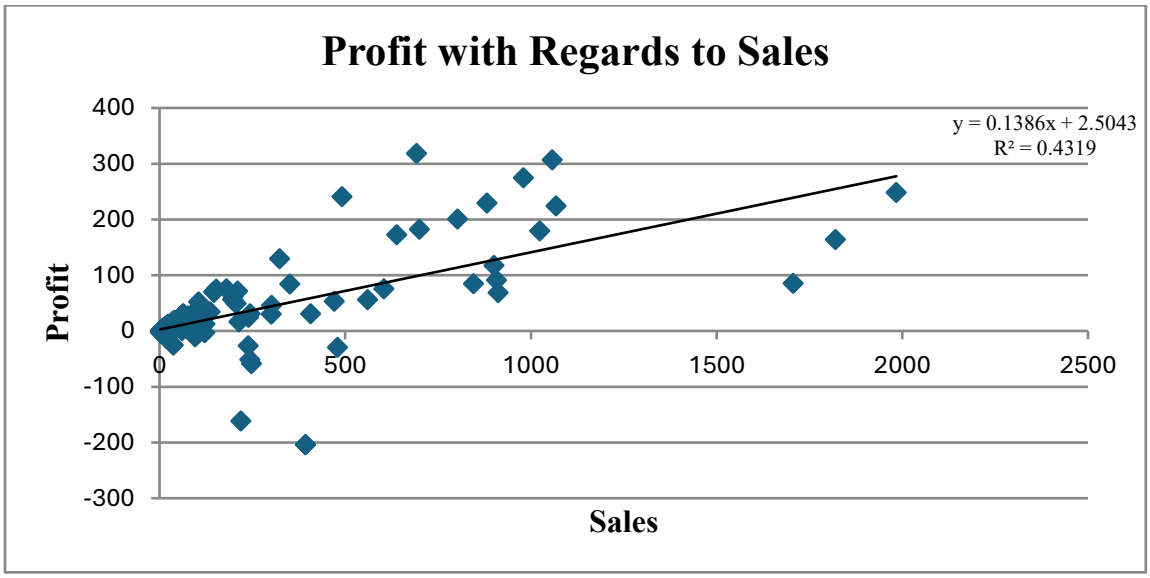
In both cases, scatter plots are used to visualize patterns, and the strength of each relationship is assessed through regression coefficients, R-squared values, and correlation coefficients. Each model is further tested by making predictions for stores excluded from the regression estimation process, helping evaluate how well the model performs in practical forecasting scenarios.

These regressions will help determine which factor: sales or quantity is a more reliable predictor of profit in the selected consumer-segment stores.

3.1 PROFIT AND SALES

Sales refer to the total dollar value of products sold by a store. This section develops a regression model to explore the relationship between sales revenue and profit earned by consumer-segment stores. The analysis aims to determine whether generating more sales leads to higher profits, and how strong that relationship is. The model will also be used to predict profit values for stores not included in the estimation sample, providing an opportunity to evaluate the predictive power of the relationship. Based on economic intuition, a positive relationship is expected, as greater sales activity is typically linked to higher levels of profitability.

3.1.1 Scatterplot:



3.1.2 Interpretation of Scatterplot:

The scatterplot illustrates a positive linear relationship between Sales and Profit, indicating that as sales increase, profits also tends to rise. However, the relationship is weak, showing a lot of variability in the data. There are 2 potential outliers, one at profit 318.3016 and the other at profit -204.4458.

3.1.3 Estimates of the Regression Model:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.657178321
R Square	0.431883346
Adjusted R Square	0.427264511
Standard Error	59.43822731
Observations	125

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	330343.4843	330343.5	93.50483	8.54813E-17
Residual	123	434547.0525	3532.903		
Total	124	764890.5367			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.504317064	6.217718555	0.402771	0.687816	-9.803275577	14.8119097	-9.803275577	14.8119097
Sales	0.138605031	0.014333821	9.66979	8.55E-17	0.110232112	0.166977951	0.110232112	0.166977951

3.1.4 Regression Equation:

From the regression table above, the following regression equation can be established:

$$\text{Profit} = 2.5043 + 0.1386\ \text{Sales}$$

The estimates for the regression equation can be interpreted as follows:

The estimated slope is 0.1386 which implies that for every 1 unit increase in sales, the profit increases by approximately 0.14 units, on average. The estimated intercept is 2.5043. It suggests that if sales were 0, the model predicts a base profit of about 2.50 units. While this might not be realistic in a business context, it helps define the starting point of the regression line.

The coefficient of determination (R-squared) has a value of 0.4319. This suggests that approximately 43.19% of the variation in profit can be explained by changes in sales. While the model captures some of the variability, more than half (56.81%) of the variation in profit remains unexplained, suggesting that other factors beyond sales also influence profit significantly. The correlation coefficient r = +0.6571 indicates a strong positive linear relationship between sales and profit.

We can calculate the elasticity of Profit with respect to Sales at their mean values, using the formula:

Mean value of Profit = $ 33.68366

Mean value of Sales = $ 224.951

Elasticity = $b_1 \times \frac{\bar{x}}{\bar{y}}$

$= 0.1386 \times \frac{224.951}{33.68366}$

= 0.9256%

This means that a 1% increase in Sales is associated with an estimated 0.9256% increase in Profit, on average.

3.1.5 Test of Model (Predictions from the Model)

In order to test the model, it is essential to apply it to data points that were not part of the original regression analysis. This helps evaluate the predictive accuracy of the regression equation. For this purpose, five U.S. states were selected: Arizona, Arkansas, California, Colorado, and Connecticut.

The sales figures for each state were substituted into the regression equation:

$$\text{Profit} = 2.5043 + 0.1386\ \text{Sales}$$

This provided estimated profit values, which were then compared with the actual profit figures to compute the residuals (Residual = Actual Profit – Predicted Profit). The residual indicates the extent to which the model overestimated or underestimated the profit.

For example, Arizona had total sales of $8.376. Substituting into the regression equation:

$\hat{Y} = 2.5043 + 0.1386(8.376)$

$\hat{Y} = 3.665$

The actual profit was $2.7222, leading to the following residual:

Residual = 2.7222 – 3.665 = - 0.94301

This shows that the model overestimated the profit for Arizona by approximately $0.94.

A similar approach was used for the remaining states, and the results are shown in the table below:

State	Sales	Actual Profit	Predicted Profit	Residual
Arizona	8.376	2.7222	3.6652136	-0.94301
Arkansas	19.89	5.3703	5.261054	0.109246

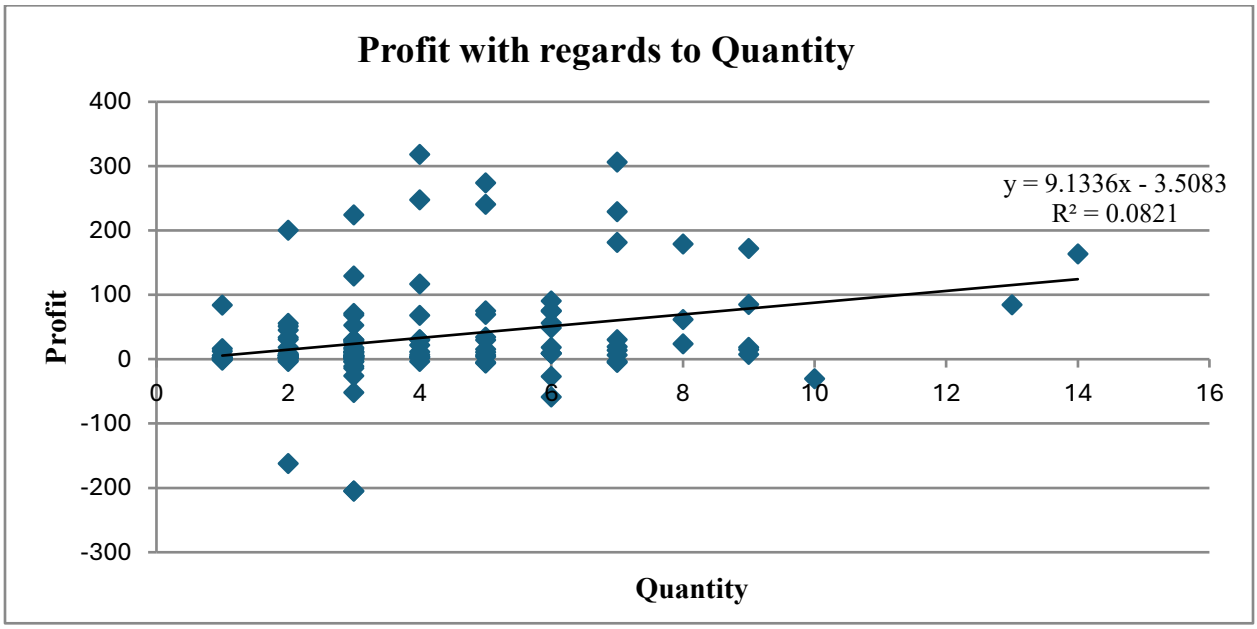
California	110.96	53.2608	17.883356	35.37744
Colorado	662.88	74.574	94.379468	-19.8055
Connecticut	79.92	37.5624	13.581212	23.98119

From the residual values, it is evident that the model predicts profits fairly accurately for states like Arizona and Arkansas, where residuals are relatively small. However, there are significant discrepancies in states like California and Connecticut, where the model underestimates profit substantially. In Colorado, the model overestimates profit by a large margin. These variations indicate that while sales have some predictive power, the model likely omits other important variables influencing profit, limiting its overall accuracy.

3.2 PROFIT AND QUANTITY

Quantity refers to the number of units of products sold by a store. This section creates a regression model to examine the relationship between the quantity of items sold and the profit earned by a store. The model investigates whether selling more items translates into higher profit margins for stores operating in the consumer segment. The theoretical relationship between these two variables is then used to make predictions and assess the strength and nature of the association.

3.2.1 Scatterplot (in Excel):



3.2.2 Interpretation of Scatterplot:

The scatterplot illustrates a positive linear relationship between Quantity and Profit, indicating that as Quantity increase, profits also tends to rise. However, the relationship is weak, showing a lot of variability in the data. There are 13 outliers that fall outside the range of -61.746 to 109.518. The outlier values are: 274.386, 306.4894, 318.3016, 247.996, 240.8595, 229.3018, 224.2674, 200.49,-204.4458, -161.875, -58.6872, -51.7191, and -29.94.

3.2.3 Estimates of the Regression Model:

Regression Statistics								
Multiple R	0.286451908							
R Square	0.082054696							
Adjusted R Square	0.074591726							
Standard Error	75.55365705							
Observations	125							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	62762.86024	62762.86	10.99491	0.001201177			
Residual	123	702127.6765	5708.355					
Total	124	764890.5367						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3.508264283	13.09480906	-0.26791	0.789215	-29.42863591	22.41210734	-29.42863591	22.41210734
Quantity	9.133575511	2.754513788	3.315858	0.001201	3.681184361	14.58596666	3.681184361	14.58596666

3.2.4 Regression Equation:

From the regression table above, the following regression equation can be established:

$$Profit = 9.1336\ Quantity - 3.5083$$

The estimates for the regression equation can be interpreted as follows:

The estimated slope is 9.1336, which implies that an increase in quantity sold by one unit is associated with an expected increase in profit by approximately $9.1336. The estimated intercept is -3.508. This suggests that when no units are sold, the expected loss is 3.508. While selling zero units is not a practical scenario, the intercept may reflect minimal fixed gains or rounding errors, serving more as a technical element in the regression.

The coefficient of determination (R-squared) has a value of 0.0821, which indicates that only 8.21% of the variation in profit can be explained by changes in quantity sold. The correlation coefficient (r) is +0.2865, indicating a weak positive correlation between Quantity and Profit. This means that while profit tends to increase as quantity increases, the relationship is not strong.

We can calculate the elasticity of Profit with respect to Quantity at their mean values using the formula:
Mean value of Profit = **$** 33.68366

Mean value of Quantity = 4.072

Elasticity = $b_1 \times \frac{\bar{x}}{\bar{y}}$

$= 9.1336 \times \left(\frac{4.072}{33.684}\right)$

$= 1.104\%$

This means that a 1% increase in quantity sold is associated with an estimated 1.104 % increase in profit, on average.

3.2.5 Test of Model (Predictions from the Model)

To assess the predictive accuracy of the Profit vs. Quantity regression model, data from five U.S. states not used in the original regression were selected: Arizona, Arkansas, California, Colorado, and Connecticut. This helps determine how well the model generalizes to new observations.

The regression equation used for prediction is:

$$Profit = 9.1336\ Quantity - 3.5083$$

Each state's quantity value was substituted into the equation to obtain the predicted profit. The difference between actual and predicted profits was then calculated to find the residual (Residual = Actual Profit – Predicted Profit), which shows the prediction error.

For instance, Arizona sold **3** units. Using the equation:

$\hat{Y} = 9.1336(3) - 3.5083$

$\hat{Y} = 23.8925$

The actual profit was $2.7222, so:

Residual = 2.7222 – 23.8925

= – 21.1703

This negative residual indicates that the model overestimated profit for Arizona by approximately $21.1703.

The same process was applied to all five states, summarized in the table below:

State	Quantity	Actual Profit	Predicted Profit	Residual
Arizona	3	2.7222	23.8925	-21.1703
Arkansas	9	5.3703	78.6941	-73.3238
California	2	53.2608	14.7589	38.5019
Colorado	3	74.574	23.8925	50.6815
Connecticut	4	37.5624	33.0261	4.5363

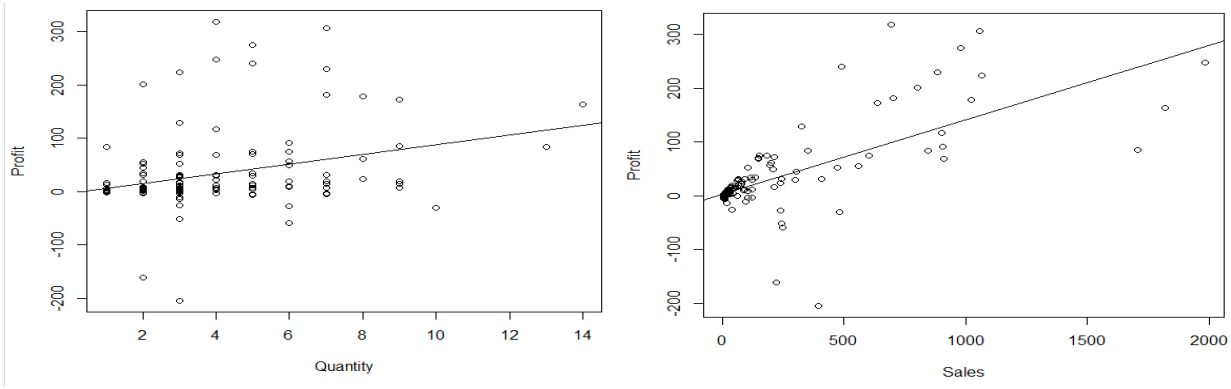
The residuals in the table highlight notable differences between actual and predicted profits, indicating inconsistencies in the model's predictions. For instance, Arkansas shows a significant overestimation, with a residual of -73.32, while Colorado exhibits a strong underestimation of 50.68. California also reflects a notable underestimation, whereas Connecticut's prediction is relatively close to the actual value. These wide fluctuations suggest that the model struggles to accurately capture the relationship between Quantity and Profit, reinforcing the earlier conclusion that Quantity alone is a weak predictor—reflected by the low R-squared value of 0.0821.

3.3 Comparing the Models

Of the two regression models created, Sales appears to be the better predictor of Profit based on the tests conducted. The Sales model has a coefficient of determination (R-squared) of 0.4319, whereas the Quantity model has an R-squared value of only 0.0821. Since the Sales model has a significantly higher R-squared value, it explains a greater proportion of the variation in profit across the selected states. This makes Sales a more reliable and effective predictor of profit than Quantity, which shows a weak and inconsistent relationship. Additionally, the residuals for the Sales model are generally smaller and more stable in states like Arizona and Arkansas, while the Quantity model displays much larger residuals, indicating poorer predictive accuracy.

4. Scatterplot (in R):

Codes:



5. Conclusion

The analysis of profit across the five selected states reveals notable variation when compared to overall averages. While some states report profits above the mean and median values, others fall significantly below, reflecting an uneven distribution. The skewness in the data suggests that profits are not symmetrically distributed, with some extreme values influencing the overall picture. This variability highlights that profit levels differ widely across regions, which may be influenced by local market conditions or other external factors.

The regression analysis further clarifies the relationship between profit and its potential predictors. Profit shows a stronger and more consistent positive association with sales, as indicated by a higher coefficient of determination, suggesting that sales value is a more reliable predictor of profit. In contrast, the relationship between profit and quantity sold is weaker and less consistent, reflected by a lower R-squared value and larger prediction errors. These findings suggest that while both sales and quantity influence profit, sales provide a clearer understanding and better forecasting ability for profit outcomes.

6. References

Definitions of Descriptive Statistics taken from:

https://datavizcatalogue.com/methods/box_plot.html

https://www.cuemath.com/data/descriptive-statistics/

https://www.khanacademy.org/test-prep/v2-sat-math/x0fcc98a58ba3bea7:problem-solving-and-data-analysis-harder/x0fcc98a58ba3bea7:data-representations-harder/a/v2-sat-lesson-data-representations

Data taken from:

https://community.tableau.com/s/question/0D54T00000CWeX8SAL/sample-superstore-sales-excelxls

Histogram created on:

https://www.statskingdom.com/

Scatterplot created on: R