**The need for adoption and integration of big data analytics Agron**

**Written by**

* Oluwatobi Oluseyi Akanbi - 20023021
* Hiaka Bibi - 19021980
* Joris Brendeke - 22030767
* Alkestid Sterjo - 18021753
* Sheila Aladejana - 22010494
* Saba Majeed - 21020845

13-01-2023

**BDA REPORT**

**RT**

# Glossary

**α (alpha):** shows the significance level. A percentage of confidence will be set to determine how sure the researcher is about the findings of the analysis. For this project the α = 0.05.

**Big data analytics**: the science of drawing insights from large quantities of raw data to aid informed decision-making.

**Correlation:** the statistical relationship between two values.

**Enterprise resource planning (ERP)**: a software application that helps companies manage and automate core business processes.

**F-value**: the variation of the means of samples.

**Hypothesis (H):** an assumption made based on former knowledge.

**Multiple linear regression (MLR):** distinguishes relationships between one dependent factor and multiple independent factors.

**R2:** displays the amount of variance in the dependent variable explained by the independent variable.

**R-value:** correlation coefficient ranging from -1 to 0 or from 0 to 1. With 0 being no correlation and (-) 1 being a perfect positive (or negative) correlation.

**Simple linear regression (SLR):** estimates the relationship between two factors to what extent.

**T-test:** To test whether a strategy/ process can work or statistically find a difference between two factors.

**Variable:** means the same as data.

# Table of content

[Glossary 2](#_Toc124512055)

[Table of content 3](#_Toc124512056)

[Introduction 4](#_Toc124512057)

[Literature review 4](#_Toc124512058)

[Methodology and Justification by Saba 6](#_Toc124512059)

[Analysis and Findings 7](#_Toc124512060)

[Summary statistics 7](#_Toc124512061)

[One sample t-test 9](#_Toc124512062)

[Two sample t-test 11](#_Toc124512063)

[One Way ANOVA 13](#_Toc124512064)

[Correlation test 14](#_Toc124512065)

[Simple Linear Regression (SLR) 15](#_Toc124512066)

[Multiple Linear Regression (MLR) 17](#_Toc124512067)

[Data visualisation 18](#_Toc124512068)

[Data visualisation 19](#_Toc124512069)

[Big Data Analytics in Agron 24](#_Toc124512070)

[Implications of findings 24](#_Toc124512071)

[Recommendations for the way forward 26](#_Toc124512072)

[Conclusion 27](#_Toc124512073)

[References 28](#_Toc124512074)

[Appendix 29](#_Toc124512075)

# Introduction

This report presents why Agron should adopt and use big data and analytics to gain a competitive edge and technological advantage in the retail industry. For context the company vision is stated below:

“…to create a world where individuals and businesses can access stylish and durable workwear produced in a way that respects the environment and the people who make it. We believe that everyone has the right to work in clothing that is safe, comfortable, and environmentally friendly, and we are committed to making that a reality through our products and practices”

Big data analytics (BDA) comes with many challenges, which, if appropriately managed, would ultimately result in success. However, the advantages are innumerable, as demonstrated through the quick analysis conducted using SAS. The findings showed what stores were doing well and which needed to be improved; the relationship between the sales performance and the other independent factors which an average individual would rarely consider when making critical business decisions. Thus, Agron can also apply these techniques to identify trends, issues, and opportunities.

BDA will contribute to this vision and also help Agron gain a competitive over Caterpillar Workwear and 3M (Agron’s competitors for the past ten years). Thus, enabling Agron to achieve its objective “to become the leading provider of high-quality, organic, and sustainable workwear for individuals and businesses who value sustainability and ethical production practices.”

# Literature review

The use of analytics and big data in influencing organisational decision-making has received much attention in the new age (Mikalef et al. 2018), made possible by the emergence of powerful software (Galetti et al. 2019). Chen et al. (2012) explored how business analytics and related technologies would enable businesses to understand their company and market; companies can take advantage of these opportunities discovered through the copious data and domain-specific analytics. The use of BDA has significantly transformed the competitive landscape of every industry, and any organisation that utilises this massive current information flow can significantly boost its performance (McAfee and Brynjolfsson no date).

The sheer volume of data generated globally daily gave rise to various trends geared towards the efficient analysis of big datasets. Khan et al. (2018) discussed the biggest trends in big data analytics for businesses to be the increased use of cloud-based solutions, which provides a cost-efficient platform for storing and analysing datasets; the integration of machine learning and artificial intelligence; the use of big data technologies such as Hadoop and Spark.

Fountaine et al. (2019) discovered that despite the enormous potential of adopting big data and analytics, many businesses are afraid of adopting it or have failed in their attempts to use it; business leaders make the mistake of establishing the technical aspect without budgeting and investing in the integration and adoption of this modern technology system.

Subsequently, Galetsi et al. (2019) analysed the challenges of processing large volumes of data, and a key challenge revolves around data management and security. Lazer et al. (2014) found that the trend in data analytics is the integration of IoT data. This proliferation of connected devices has led to a deluge of real-time data, which can cause ‘Analysis Paralysis’ according to Dekimpe (2020), where a company struggles to make data-driven decisions as there is too much data. Marr (2018) recommended adhering to data minimisation, to limit the collection of personal identifiable information to only what is necessary for accomplishing a specified purpose as a step to resolving this problem.

Blackman (2022) reviewed the need for establishment of an ethical committee to encourage ethical use and creation of technology. This, he proposed would help with privacy and governance issues, another set of challenges for businesses; complying with regulations governing personal data collection, ownership, and its use - like the General Data Protection Regulation (GDPR) is also encouraged with severe penalties stated for non-compliance (UK Government 2018).

Korherr et al. (2022) proposed a paradigm shift for firms, especially their top executives, to embrace analytics in corporate decision-making processes – moving from intuitive to data-driven decision-making.  Elfindah Princess (2021), on the other hand, highlights how crucial it is to meld the use of intuition and data in commercial decision-making, mainly to use intuition in situations where it excels.

Currently, the retail industry is maturing in BDA. Seetharaman et al. (2016), in their studies have shown that retail organisations that practice the use of data-driven decisions have higher output and productivity. The impact of BDA in retail is discussed to be evident in certain aspects - “the utilisation of inventory, customer engagement amends market value in the retail industry”. Phani and Venkatesham (2021) discussed how BDA guides retailers in deciding on investment in ERP systems to better the quality of data and gain more accurate insights into customers.

Chang and Wang (2021) posit that there will be a trend of “new retail” and E-Commerce (EC) based on BDA; three aspects of this new retail: evolving technologies, social engagement, and becoming business & socially mobile with the community. Revamping Agron to have a footing in these three aspects would put the company forward in customer understanding and improve its efficiency in operations management & supply management, as well as drive an increase in profit and organisation expansion.

# Methodology and Justification by Saba

The sample data is obtained from Kaggle, an online platform for data analytics and machine learning, and covers sales from 5th February 2010 to 1st November 2012. The dataset is from the retail industry titled ‘Walmart Dataset’; see bibliography for details.

For this report, the data is revised to cover the sales data of 12 stores. The following fields include store number, date, weekly sales, consumer price index (CPI), cost of fuel, weather temperature, and unemployment rate in the region. The data types in the dataset are continuous and categorical variables (quantitative and qualitative data).

The data was chosen because it allows for effortless application of the data analytics techniques using SAS –a powerful software package that provides a wide range of analytical techniques. SAS is used by many organisations to analyse data and make decisions. The data analytics techniques were applied to the data to test for various assumptions or expectations.

The first technique, summary statistics (descriptive statistics), was applied to summarise and describe the dataset characteristics. One-sample t-test, two-sample t-test and the one-way ANOVA (ANOVA post-hoc test) are applied to determine differences between groups by comparing means, as the data set had categorical data and continuous data.

Correlation and regression are used to examine and model the relationship between variables. Where: dependent data (weekly sales) compared to independent data (the CPI, temperature, etc.) together and individually. These techniques require continuous data.

Analysts commonly use these techniques to ascertain findings giving insights into operations. Using this demonstrates to Agron examples of insights drawn from analysing data.

# Analysis and Findings

## Summary statistics

Figure 1: summary stores 1-12 stores in one table; looking at the standard deviation shows weekly sales have a bigger spread at 53,8% compared to others. Figure 2 is more specific as it shows the 143 observations of each store. Both results show the average, standard deviation, minimum value and maximum value in the dataset, total number of values in the dataset – 1716, middle value in the dataset – median, and the last two are the lowest and highest value where 95% of the data is.

Figure 3: Box and Whisker Plots: spread of weekly sales. Showing the range of the outlying values.

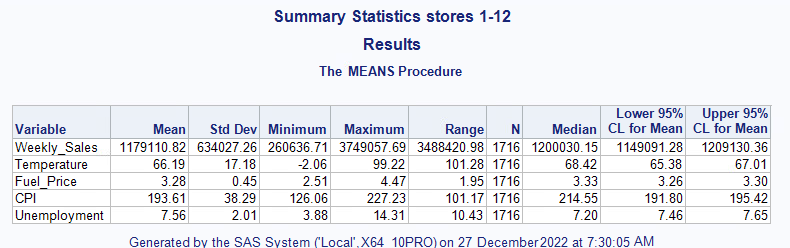
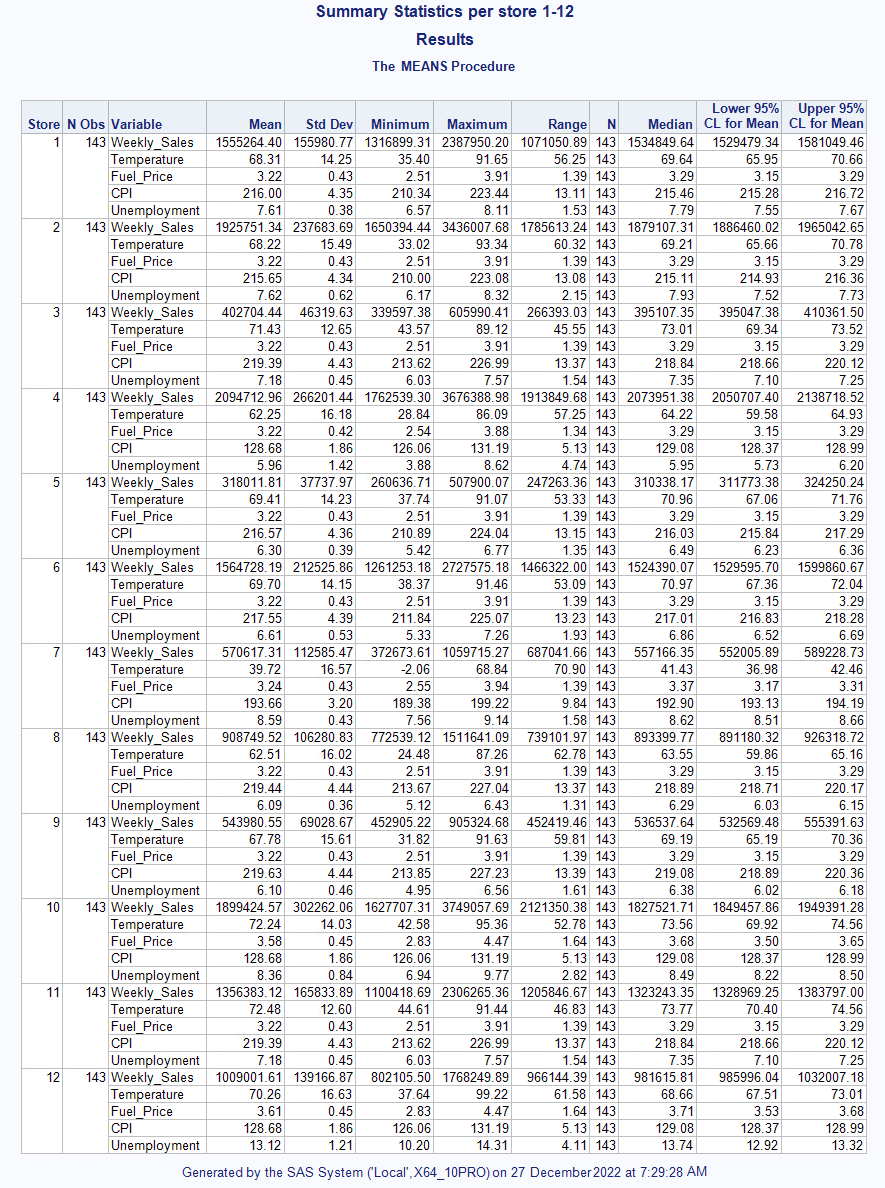


Figure 1: summary stores 1-12



Figure

2: statistics per store

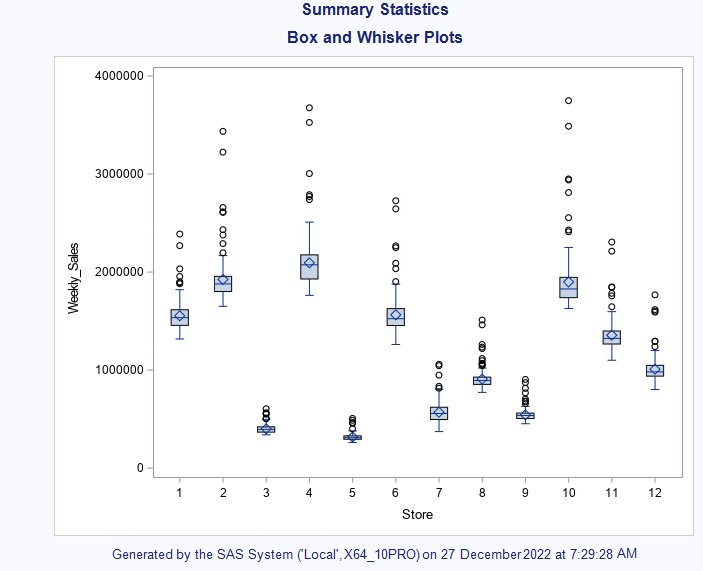


Figure 3: Box and Whisker Plots

## One sample t-test

Hypothesis is that the average weekly sale of all stores is 1,000,000. Thus, H0 = 1,000,000. H1 > 1,000,000, based upon the observations of Figure 3: Box and Whisker Plots. α = 0.05. In Figure 4 the spread of the data is visualised in red.

For the one sample t-test, the H1 uses an equation with >, the test is constructed like a one-tailed test with an upper tail focus, meaning the critical region is at the upper end of the data. The H1 has been accepted for the one sample t-test instead of the H0, there were no implications in these findings.

The results are p-value < 0.0001 which is less than α, and t-value is 11.70; the results fall within the critical region and are statistically significant. This means a rejection of the H0. The spread of the mean of weekly sales is displayed in Figure 5. Figure 5: t-test mean distribution and Q-Q plot where the mean is not near H0 = 1,000,000. H0 can be rejected successfully since there is a significant difference between H0 and actual average weekly sales.

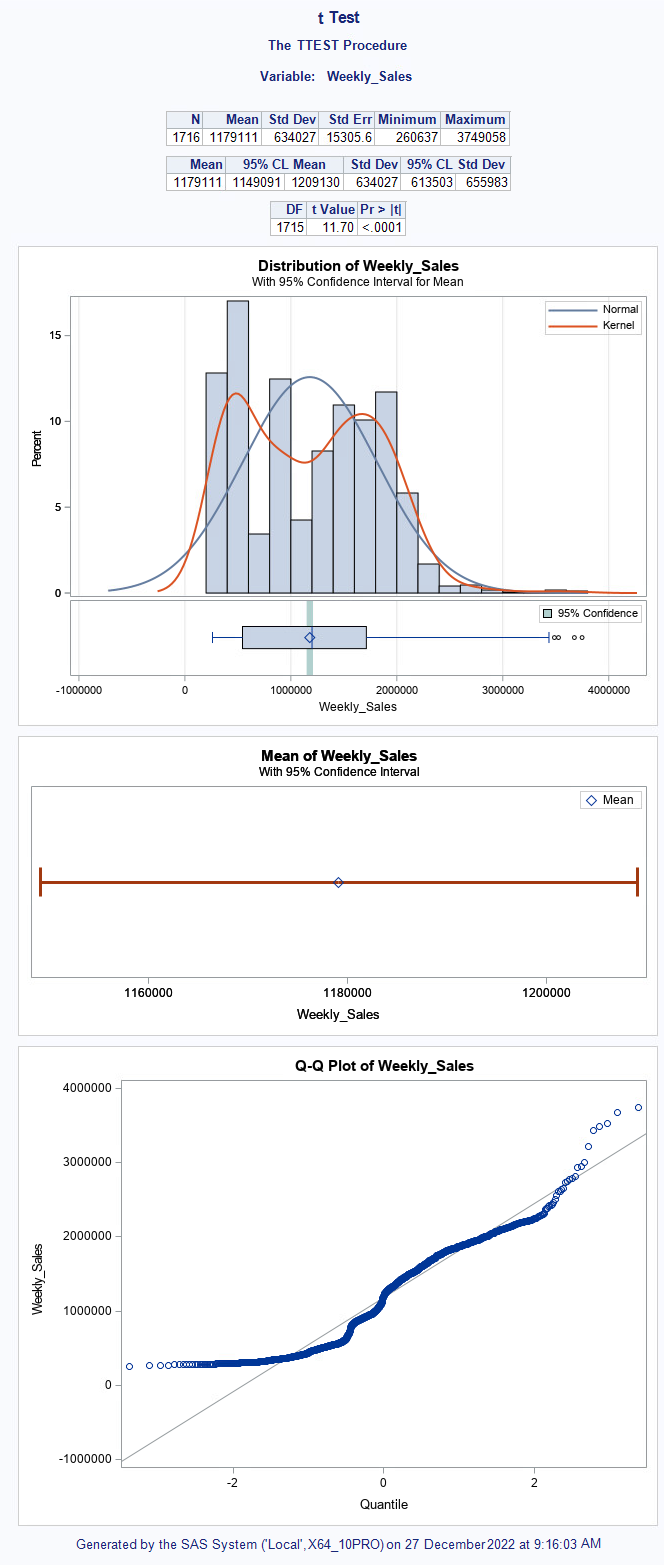
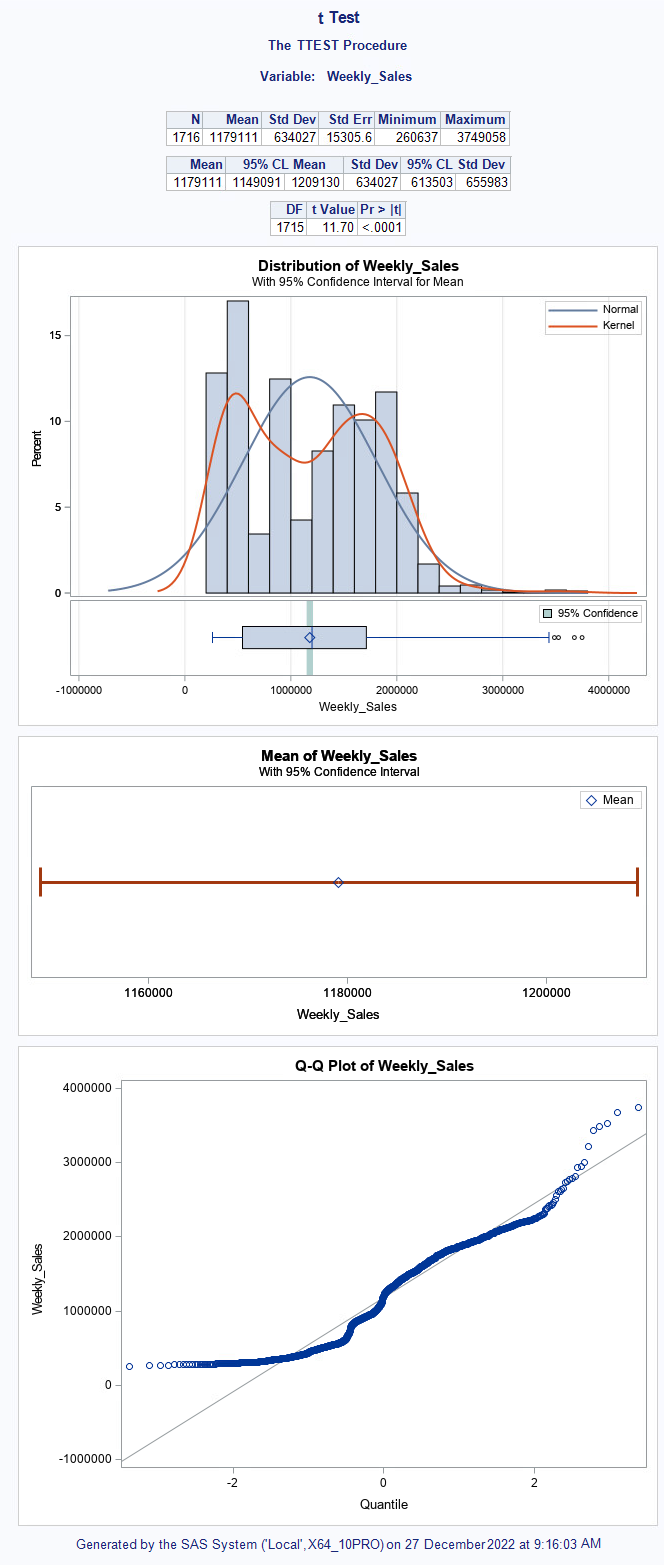
 

Figure : t-test results and distribution

Figure : t-test mean distribution and Q-Q plot

## Two sample t-test

The H0 is that there is no significant difference in average weekly sales of the two regions California and Texas stores.

H0 = p-value > alpha (no difference between weekly sales in California and Texas).

H1 = p-value <= alpha (significant difference between weekly sales in California and Texas).

The F-test for equal variances has a p-value of < 0.0001, because this value is smaller than the alpha, the H0 can be rejected. Using the equal variance t-test, the H0 is also rejected. The mean difference between Texas and California’s stores is ‑984647. The p-value is smaller than 0.05 (Pr > t = 0.0001); concluding that there is a significant difference in the between the average weekly sales in California and Texas stores.

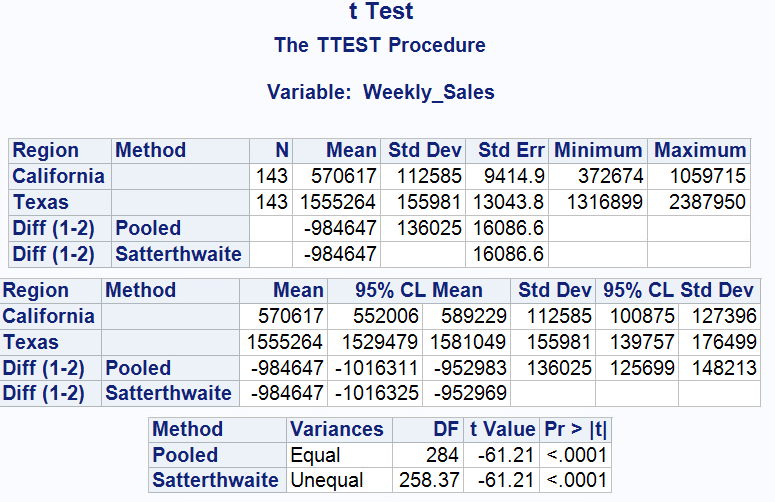


Figure : two sample t-test

The X-axis for the Q-Q plot is scaled as quantiles, rather than probabilities. For each group it seems that the data approximate an abnormal distribution, and the alternate hypothesis is accepted.

The confidence intervals between California and Texas stores are similar as the pooled and Satterthwaite intervals and p-values are remarkably similar. The lower bound of the pooled interval does not extend zero.

A picture containing chart

Description automatically generated

Figure : weekly sales difference

Chart, line chart, histogram

Description automatically generated

Figure : Distribution of weekly sales

Chart, line chart

Description automatically generated

Figure : Q-Q plots of weekly sales

## One Way ANOVA

This is like a two-sample t-test but should be applied to two or more groups. The dataset has only two groups; thus, the results are like the results of the two-sample t-test above. Figure 10 shows the values and the number of observations used in the test.

Graphical user interface, text, application

Description automatically generated

Figure : ANOVA overview

The results as seen in Figure 11 and Figure 12 establish there is a statistically significant difference in sales performance between California and Texas.

Graphical user interface, text, application

Description automatically generated

Figure

Graphical user interface, text, application

Description automatically generated

Figure

ANOVA analysis does not consider the regions within the state, thus a post-hoc analysis is required. However, because at least 3 groups are required to conduct ANOVA post-hoc, the following table does not truly reflect the real result.

A picture containing text

Description automatically generated

Figure

## Correlation test

Graphical user interface, text, application

Description automatically generated

Figure

Table

Description automatically generated

Figure : Texas

Table

Description automatically generated

Figure : California

Hypothesis:

* H0 = There is no significant relationship between weekly sales and CPI, fuel price, temperature, unemployment rate in California and Texas.
* H1 = There is a significant relationship between weekly sales and CPI, fuel price, temperature, unemployment rate in California and Texas.

Figure 15: Texas result showed there is a negative relationship between CPI and weekly sales (R = -0.50), which is statistically significant (p-value < 0.0001); a strong correlation between the two variables in both regions [Texas (r = -0.49, p < 0.0001), California (r = -0.50, p < 0.0001)]. This means a rejection of the null hypothesis.

Weather temperature on the day of sales showed positive but moderate correlation with weekly sales, r = 0.26, p = 0.026 in California. However, there is a negative correlation between temperature and weekly sales in Texas (r = -0.17, p < 0.0001) with statistical significance. The cost of fuel and unemployment rate have also shown a significant positive correlation with the weekly sale in Figure 16: California, with statistically significant difference. No statistically significant correlation was evident for the cost of fuel and unemployment rate with weekly sales, as α > 5% in Texas.

## Simple Linear Regression (SLR)

SLR is used to examine the level of relationship between an independent variable (IV) - example temperature, CPI, unemployment, and fuel price - on the dependent variable (DV) - weekly sales different regions.

The applicable SLR equation for prediction is:

Y^ = Bo + B1X (Where Y = predicted weekly sale, Bo = intercept, B1 = slope, X = independent variable). The baseline model uses averages and shows the best predictive model with no slope. The hypothesis being tested is as follows:

* H0 B1 = 0 no slope
* H1 where B1 ≠ 0, i.e., there is a slope.

The following shows the prediction comparing SLR and baseline model:

Where p-value < 0.05, reject H0.

Table

Description automatically generated

Figure : SLR with Temperature as Predictor and Weekly Sales as the Response variable

From the results, the F-value = 4.91 and p-value = 0.0268 is small, which means the model has made significant improvement from the baseline model hence, the H0 is rejected.

R2 of 0.0029 shows there is a 0.29% change in predicted weekly sales (Y^) can be attributed or explained by the Weather temperature - the IV (X).

Table

Description automatically generated with medium confidence

Figure : Result of SLR with CPI as Independent Variable

Graphical user interface

Description automatically generated

Figure : SLR with CPI as Predictor and Weekly Sales as the Response variable

The F-value of 345.15 shows the extent the model has improved on previous models. It’s validated by the small p-value of < 0.0001, meaning, the model has significantly improved from the baseline model; H0 is rejected.

An R2 figure of 0.1676 shows that 16.76 % of the change in predicted weekly sales (Y^) can be attributed to CPI (X). A negative relationship between CPI and weekly sales as shown in the graph.

## Multiple Linear Regression (MLR)

Table

Description automatically generated

Figure : MLR results

This technique used the continuous data from the dataset and created a model to depict the assumed linear relationship between these variables.

For this analysis, α = 0.05. The H0 is that none of the predictor variables fuel cost, CPI, unemployment rate and temperature has had an impact on the weekly sales. The H1 posits that at least one of the predictor variables has an impact on the weekly sales. The model is shown in Figure 17 to have a high F-value at 136.48. Thus, the null hypothesis must be rejected as each predictor has a statistically significant impact on weekly sales based on the p‑value of the model which is <0.0001. The highest impact is made by the predictor CPI, as it has the highest t-value compared to others.

The Adjusted R-square at 0.24 shows how much variability in the dependent variable can be explained by the model.

## Data visualisation

This briefly shows the insight from the data; it makes the data easier to read and understand.

Chart, bar chart

Description automatically generated

Figure : total weekly sales by region

## Data visualisation

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Figure : Percentage total weekly sales by stores

Figure 23: sales by region

Figure 21: total weekly sales by region and Figure 2323: sales by region displays total number of sales recorded within the period by two regions Texas and Calfornia.

Figure 22 shows Texas region performing better than California region at 55.56% to 44.44% respectively.

The Figure 24: total weekly sales by store shows the cumulative percentage performance of the twelve stores in California and Texas. The highest performing store in both region is store 4 with 14.8% while the least performing store is store 7 with a 4.07% contribution to total weekly sales.

Chart, line chart

Description automatically generated

Figure : total weekly sales by store

**Chart, bar chart

Description automatically generated**

Figure : bar charts showing total weekly sales across each store

Figure 25 shows the chart showing the distribution of total weekly sales in each store in its domiciled region.

Table

Description automatically generated

Figure : year on year performance of stores

Figure 26 above depicts the year-on-year sales made in each of the store for three (3) years.

Chart, pie chart

Description automatically generated

Figure : Percentage representation of sales per store per region

Figure 27 depict the total volume of weekly sales in the six stores in each of the two regions. At glance, the percentage performance of each store in a particular region can be deduced for proper management. Store 10 leads in the California region with 30.21% and is closely followed by store 11 with 21.57%. This information can be useful in inventory management. Store 5 is the least performing in the Texas region, with 4.05%.

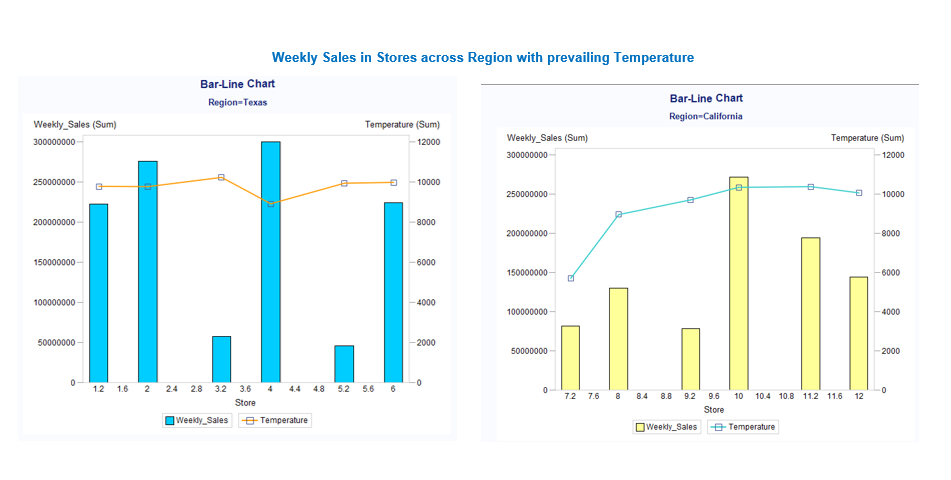
****

Figure : Bar- Line chart of Total sales across states against weather temperature

Figures 28 and 29 reveal the prevailing temperature and CPI against sales in each store across California and Texas, respectively. Store 4 Texas has the highest weekly sales but lowest temperature and CPI. Store 10 in the California region has highest sales, highest temperature, and lowest CPI in California. A lower CPI translates to more weekly sales in stores 4 and 10.

Graphical user interface, chart

Description automatically generated

Figure : Weekly sales and CPI across regions per store

# Big Data Analytics in Agron

Agron must take advantage of the new industrial era by adopting and using BDA. Using BDA is recommended because the business has a constant stream of data that meets the six V’s – volume, variety, veracity, velocity, value, and variability, from its daily operations of distribution and sale of workwear and equipment.

The findings from the application of the techniques above are related to the objectives of Agron as by understanding the patterns in the data, a lean approach to retail operations management can be employed effectively. The findings from BDA would improve customer identification, optimise its logistics through predictions from big data analysis of seasonal effects, monitor its suppliers, and enhance its inventory levels at its various branches – ultimately leading to improved supply chain management.

Good use of BDA can provide insight into customer behaviour and preferences, enabling Agron to develop and refine its offerings to meet the needs of its customers, leading to efficient customer relationship management. Additionally, BDA can be used to track customer engagement and provide valuable feedback on how to improve Agron's marketing and sales strategy. With these analytical findings, the company can identify and address discrepancies in sales between various locations, make predictions with accuracy, revise pricing strategies, and optimise the use of resources to become more sustainable.

With BDA, Agron can identify trends using semantic analysis and more complex techniques; use this information to tailor its business and social activities to meet the needs of its clientele and to create a strategy to grow specific stores or to make the smaller, less profitable ones redundant. Thus, becoming more efficient and effective in achieving the set business objectives and goals.

## Implications of findings

The cross-industry standard process for data mining was followed during the analysis conducted – a systematic way for mining data that is widely used by various industries.

To start, the summary statistics gives a surface summary of the data before it is analysed. This would enable the analysts at Agron to understand the data and determine what analytical techniques can be applied to the dataset. The box & whisker plots show the spread of the dependent data being analysed. Showing the range of these extremities would aid Agron in looking into the highest-performing store and use this to explore issues in underperforming locations. Basic statistics is the first step to using the data. This is supported by two-sample t-test, where it compares the means of two groups. This will aid Agron by comparing weekly sales averages, contributing to making unknown differences apparent through BDA.

The SLR analysis can be used externally to support sales in Agron, to be competitive through its diverse customer base. SLR evaluates customers’ behaviours through B2B or B2C, to gain a significant insight into their purchasing habits. For example, SLR may find through an evaluation that Agron customers’ orders more equipment towards the end of the year. Agron can further explore this through the MLR, where it considers multiple factors (externally) to know why business customers (BC) order more at the end of the year. These factors could include increased demand for products (in business customers) during that time of year and increased recruitment in those businesses. MLR can help allude to this being Christmas time, where there is an increase product demand in Agron’s business customers and a big need for employees. Resulting in Agron’s increased orders of equipment and workwear, leading to higher sales and profits. This aids Agron’s inventory management and leads to a more efficient operations department. As complex as this is, Agron's efficiency and improved preparedness will have massive benefits.

The correlation will help make more informed decisions in strategic planning, resource allocation and product development. Analysing the relationship between external elements and their performance can also identify cost savings and increased efficiency opportunities. Companies can better anticipate changes in the market and adjust their strategies, accordingly, leading to increased customer satisfaction, customer loyalty, and profitability.

The correlation findings are supported by ANOVA analysis, it’s helpful for Agron’s internal operations, such as managing the stock of materials to produce the workwear and equipment. ANOVA analyses the patterns in inventory, such as establishing when to order specific materials in advance. For instance, the analysis may show that when there are adverse weather conditions, delivery of materials is far and few between, therefore, the stock control software and operations managers know to order certain materials in advance. resulting in smoother and more efficient operations, decreasing downtime and operations costs.

Data Visualization proves the saying that a picture is worth a thousand words. A chart or graph conveys briefly what is being said. In this report, the graphs and charts were used to convey a description of the data as it is; they show the proportion or percentage of sales in various stores and states without need for complex words. The best performing state or store can be easily grasped at first look at the charts and graphs. SAS was used to create the visuals above but other software like Tableau and Power BI can be employed by Agron to see a visual representation of the position of the company and make decisions accordingly.

## Recommendations for the way forward

Based on the findings discussed, Agron can take a step towards adopting BDA by:

1. Hiring data scientists, domain experts and other talents to ensure the data is properly analysed and understood.
2. Creating and implementing a data capturing and warehousing system to collect and store all relevant data securely. Data can be messy and building great architecture would help with this challenge. A 3-tier architecture is recommended as this involves the use of Online Analytical Processing.
3. Ensuring legal, ethical, and regulatory compliance by establishing departments for this purpose and implementing policies that ensure that the technological talents hired mine data in line with the cross-industry standard process for data mining – CRISP-DM.
4. Setting up an Enterprise Resource Planning (ERP) system for Agron to ensure that the data is organised and easily accessible, as well as big data technologies like Hadoop.
5. Creating a data-driven culture within the firm by educating employees about the importance of big data analytics and encouraging them to use it to make decisions.
6. Leveraging the benefits of cloud computing as it’s cheaper and easier to maintain. For example, the data warehouse can be established using Amazon Redshift. That way, Agron does not need to build physical infrastructure for storing large amounts of data collected.



Figure : A broad roadmap that Agron can follow

It is important to remember that analytical work requires a good time investment, from establishing the necessary infrastructure to its full cultural adoption by the organisation. Every time one explores a dataset, one cannot always expect to learn something useful.

# Conclusion

Data is only as good as the infrastructure it sits in, it is considered the new oil, but that is only true when it is put to work and how effectively it is used. To get the data to work, Agron must establish an ecosystem to enable this to succeed, meaning establishing a solid technological infrastructure as well as a plan for its integration and adoption; company culture must be improved to be more data-driven as this is the only way for the organisation to grow in this new age and finally, invest time.

Word count: 4014

# References

1. Chang. J., and Wang, Y. (2021) Future Development Trend of “New Retail” and E-Commerce on Big Data. *Journal of Physics: Conference Series*  <https://www-sciencedirect-com.brad.idm.oclc.org/science/article/pii/S016781161930062X?via%3Dihub> Accessed 28 December 2022.
2. Chen, H., Chiang, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *Management Information Systems Quarterly* 36 (4) 1165–1188.
3. Princess, E. (2021) Data-driven analytics in the decision-making process: do we still need intuition? *Journal of Southwest* 56 (4).
4. Fountaine, T., McCarthy, B. and Saleh, T. (2019) Building the AI-Powered Organization. *Harvard Business Review* <https://hbr.org/2019/07/building-the-ai-powered-organization> Accessed 22 December 2022.
5. Dekimpe, M. G. (2020) Retailing and Retailing Research in the Age of Big Data Analytics. *International Journal of Research in Marketing* 37 (1) 3-14. <https://www-sciencedirect-com.brad.idm.oclc.org/science/article/pii/S016781161930062X?via%3Dihub> Accessed 28 December 2022.
6. Phani, J., and Venkatesham, T., (2021) Big Data Analytics- In Retail Sector. *International Journal of Computer Science and Mobile Computing* 10 (7), 34-38.
7. Korherr, P., Kanbach, D.K., Kraus, S. and Mikalef, P. (2022) From intuitive to data-driven decision-making in digital transformation: A framework of prevalent managerial archetypes. *Digital Business* 2 (2), 1-11.
8. McAfee, A. and Brynjolfsson, E. (2012) Big Data: The Management Revolution. *Harvard Business Review* <https://hbr.org/2012/10/big-data-the-management-revolution> Accessed 25 December 2022.
9. Mikalef, P., Pappas, I.O., Krogstie, J. and Giannakos, M. (2018) Big data analytics capabilities: a systematic literature review and research agenda. *Information Systems and e-Business Management* 16 (3), 547–578.
10. UK Government (2018) *The Data Protection Act*. <https://www.gov.uk/data-protection#:~:text=The%20Data%20Protection%20Act%202018%20is%20the%20UK's%20implementation%20of,used%20fairly%2C%20lawfully%20and%20transparently> Accessed 28 December 2022
11. Zaitsava, M., Marku, E. and Di Guardo, M.C. (2022) Is data-driven decision-making driven only by data? When cognition meets data. *European Management Journal* 40 (5), 656–670.
12. Griffiths, D., (2008). *Head first statistics.* O'Reilly Media Inc. <https://www-vlebooks-com.brad.idm.oclc.org/Product/Index/420200?page=0>
13. te Grotenhuis, M. & Matthijssen, A., 2016. *Basic SPSS Tutorial.* SAGE.

# Appendix

Chart, scatter chart

Description automatically generated

The intercept Bo is 1048541, where temperature (X) is zero. If the temperature is increased by 1°, the weekly sales should increase by 197200%. Based on the regression model, Weekly Sales = 1048541 + 1972.56 \* Temperature

The 95% confidence interval reveals that the regression line is slightly accurate

Graphical user interface, application, Word

Description automatically generated

Chart

Description automatically generated

The data assumptions that variances are equally and randomly spread are not fully fulfilled as the data was not scattered equally randomly across the line.

**Simple Linear regression with Fuel price**

The fuel price is used as the independent variable, and weekly sales as the dependent variable as shown in the table below

Table

Description automatically generated

A small degree improvement from the baseline model in terms of prediction with a F‑value of 4.17 is observed on analysis. Also, the p-value 0.0413 is small (i.e., P < 0.05), meaning the model has made significant improvement from the baseline model, hence H0 is rejected.

Table

Description automatically generated

R2 of 0.0024 shows that 0.24 % of the change in the predicted weekly sales (Y^) can be attributed or explained by prevailing fuel price (X).

Table

Description automatically generated

The intercept Bo is 951375 where Fuel (X) is capped at a minimum $2.5. If the price is increased by $1 the weekly sales should increase by 6936400%.

Graphical user interface, chart, application, scatter chart

Description automatically generated

The Simple Linear Regression Line to help prediction is Weekly Sales = 951375 + (69364 \* Fuel\_ Price). Based on 95% confidence interval the regression line is slightly accurate.

Chart, line chart

Description automatically generatedChart

Description automatically generated

The data assumptions that variances are equal are not fully fulfilled as the data was not scattered equally randomly across the line. The assumption of variances been normally distributed is partially fulfilled.

**SIMPLE LINEAR REGRESSION WITH CPI AS PREDICTOR AND WEEKLY SALES AS REPONSE VARIABLE**

Graphical user interface

Description automatically generated

**Simple Linear Regression with Unemployment as Predictor and Weekly Sales – Response Variable**

Table

Description automatically generated

A small F value of 0.36 and large P value of 0.5484 (I.e., P > 0.05) is observed with unemployment as predictor. No significant improvement has been made from the baseline model hence, we accept the Ho (null hypothesis). The mean predictor is better than the simple linear regression model. This is illustrated in the graph below where little slope can be observed.

Text, table

Description automatically generated

R2 figure of 0.0002 shows that 0.002 % of the change in the predicted weekly sales (Y^) can be attributed or explained by Unemployment (X). (Very negligible)

Text, table

Description automatically generated

The intercept Bo is 1213754, where Unemployment is infinite. If the CPI is increased by 1 the weekly sales should decrease by 458287.47%.

Chart

Description automatically generated

