

University of Wales
Trinity Saint David

MODULE CODE: ACUA6002
Big Data

MODULE TUTOR: Dr Terry Walcott

Assignment Title: Big Data model and warehouse for Pestron
Foods

Antoci Bianca
1817784

Table of Contents

Abstract	2
Introduction	3
Literature Review	4
Pestron Food data analysis	6
Pestron Food data breakdown	7
Data model	9
Data Warehouse	10
Conclusion	12
References	14
Appendices	16

Abstract

In recent years Big Data applications have gotten progressively significant especially with the rise of Cloud Computing. Many companies need progressively more detailed information extricated from enormous volumes of data generated by their business. Big Data applications can assist organizations with settling better business choices by investigating enormous volumes of information and finding hidden connections [1]. In our study first, we are going to understand the role of Big Data in small companies. It will continue by examining Pestron Foods requirements using the appropriate tools and suggest improvements on various operational areas of the company [2]. The last part of the report will be concluded with a summary and future recommendations.

Introduction

The overall volume of data has risen significantly in recent years as a result of the rise of social networks, multimedia, and the Internet of Things. Data has spread throughout the industry and is now regarded as a critical component in determining the manufacturing process, as well as capital and labour. Additionally, as the amount of unstructured data grows, conventional data analysis technologies are unable to adequately handle it.

All large and small companies produce this big amount of data. With all this data collected, it is good that the companies will use this information to better understand their customers and their conduct. This is the motivation behind why the prevalence of Data Science has developed complex throughout the most recent couple of years [3]. The Big Data set is a statistical data set, which can have any form and type and be of enormous size. Due to these factors, traditional processing software and hardware cannot handle such a task in a meaningful time.

More companies like Pestron Foods are considering implementing a Big Data analytics solution because they generate multiple complex data types from their various stores located across the country. With the help of data analytics solutions, business organizations can gain deep insights, which can improve performance through better management decisions.

Additionally, manufacturers can improve forecast accuracy through big data analysis, thus minimizing inventory costs.

In our study, the CEO of Pestron Foods would like to introduce Big Data in the company to better predict sales, overall costs and make an educated decision on expanding in new markets, like to different locations across the country with potentially different and more complex requirements [4].

Pestron Foods is a mini-supermarket that collects detailed data for each individual store on customer purchases and financial data like taxes and profits. There are several options for analyzing large amounts of structured data and unstructured data. The two potential solutions are Google DataStudio and Knime Analytics. DataStudio is a cloud-based solution that integrates with BigQuery(a data warehouse solution from Google) and Knime is software that runs locally, equally flexible for handling large amounts of data but requires local computationally powerful hardware.

Literature Review

Big Data is the term used to depict enormous volumes of information that a business makes as a piece of its ordinary working methodology. Big Data is a digital phenomenon that allows large quantities of data produced from both man and machine to be collected and used. The amount, diversity, velocity, authenticity, variability, and complexity of this data are all factors to consider [5]. Organizational executives want to utilise big data analytics to establish competitive difference and make the most of the data they already have.

Organizations hope to attain excellence and greater efficiency in their business by using big data analytics, which is one of today's growing information systems.

[6].

It is very important for a business to use big data because it reduces costs, increases efficiency, optimizes pricing strategies, identifies and prevents fraud and makes smarter business decisions.

Big Data can benefit from appropriate models and storage settings in the following ways:

- Cost: good data models may considerably minimise needless data redundancy, reuse computing results, and lower the big data system's storage and computing expenses.
- Efficient: good data models may substantially improve user experience and data utilisation efficiency.
- Consistency: good data models ensure that data statistics are consistent and that compute mistakes are minimised.
- Traffic: good data models can help us query the needed data fast and decrease I/O throughput.

When bringing big data into a firm, there are a few challenges to consider:

1. Data Security: data security is routinely placed to the back burner as businesses focus on understanding, conserving, and analysing their data collections.
2. Lack of knowledge Professionals: to operate this cutting-edge technology and massive data tools, businesses demand qualified data specialists.
3. Lack of support: data analytics will be ineffective without organisational support, both at the top and at the bottom.
4. Visual representation of data: data is frequently displayed graphically in graphs or charts in order to be understood and effective. While these tools are really valuable, manually creating them is tough.

Big Data isn't too massive for little businesses. In reality, it's necessary to uncover hidden patterns, market trends, customer preferences, and alternative pregnant perspectives. With Big Data, small businesses will build the good selections they have to form to urge prior to their competition and increase their gain [7].

Many small firms think that they are too small to benefit from Big Data. This is frequently not the case, as even small firms, like larger corporations, want Big Data to flourish. This sort of intelligence gives organisations the insights they need to become more cost-effective and successful.

Fortunately, there are Big Data technologies and solutions on the market that are affordable and manageable for small organisations with limited budgets and resources. A number of these technologies track and analyse website user behaviour; others will aid in the monitoring and analysis of my client conversations, revealing internal career trends, caller demographics, and typical choice difficulties. The good news is that these technologies don't generally need a highly technical staff or a large sum of money in order to benefit from Big Data [8].

The two main concepts that need to be very well understood when processing this type of data are dimensions and metrics. A dimension is, in essence, a data type; for example, user clicks, page visits or items purchased. A metric is the actual data represented by the dimension; for example how many users have clicked or how many users have purchased an item. The combination of these two concepts is very powerful and it can generate very interesting insights that can help companies to make informed decisions in a very short amount of time; a very important characteristic for small businesses looking to expand rapidly.

A typical business, in today's age, hosts most of its data in the cloud. All of the largest three Cloud providers(Amazon, Azure and Google Cloud) offer Big Data solutions. For example, Google Cloud offers BigQuery and DataStudio; BigQuery being a data warehouse solution that offers unlimited storage and DataStudio, a powerful data visualisation tool.

Amazon Web Services offers Athena and EMR. Athena is a tool that allows querying large files from a storage bucket, useful when checking through large log files. Amazon EMR is a hosted Hadoop framework.

Apache Hadoop is a very popular open-source framework used by both small and large companies and was developed by Apache Software Foundation. It has been around since 2006 and it is specifically designed to scale on multiple computers and to process almost any kind of structured and unstructured data. Hadoop can easily be run on a local computer or on a cluster in the cloud [9].

Microsoft Azure offers a suite of tools called Azure Big Data that allows, similar to its competitors, to analyze large amounts of data. It is specifically designed to run on Azure Cloud datacenters.

Pestron Food data analysis

The data provided by Pestron Foods is divided into two main data sets:

- Sales figures - provided as a CSV file
- Food Store list - provided as a CSV file

CSV files are a very common way to pass data between different non-compatible systems. The data is basically stored as comma-separated values in a plain text file.

The sales figure dimensions are represented by the list of invoices from different branches in different locations. Other financial information such as unit prices, payment type, tax percentages is also presented in the same structured format.

The food store list file contains the list of stores with the exact physical location as well as the area within the city. This should help us to easily identify, for example, store performance and predict future sales based on not only location but product type, customer type or day of the week.

Other useful insights can be generated by comparing each store's dimensions over time. This type of data analysis is important as Pestron Foods benefits from a seasonal boost in sales in certain periods of the year [10].

To get started and analyse the data provided by the company, it is recommended to use Google Cloud's BigQuery, DataStudio and KNIME Analytics. The data from Pestron Foods servers could be directly ingested by BigQuery in real-time but for this initial analysis, multiple CSV files will be used as seen in figure 1. It is recommended that for security reasons the sensitive data should not be transferred using CSV in the future.

Every data project involves a data visualization phase. Usually, this phase is just temporary, it is just an investigation about the structure and property of the data we are dealing with [11].

Now that we have the actual data ready we can start the analysis process by generating some visual representations. We can use elements like charts, graphs and maps to provide an accessible way to see and understand trends, outliers and patterns in data. Other useful visual elements are time series, tables and comparison tables.

In our case, we choose to use DataStudio to visualize the data of Pestron Food. Most of the visualizations are an interactive view, some produce an image, some produce data and some produce a combination of all those outputs.

Pestron Food data breakdown

We will now analyse Pestron Food from different angles. We are going to look in detail at the different dimensions and extract any relevant information.

As displayed in figure 2 we can see the list of invoices using a simple table view. Each invoice has details about the location of the transaction, customer details, product details and information about total price, quantity and tax. The product line is also present which should help to better understand the best performing store and city. The table provided in figure 2 can easily be sorted by any of the dimensions and a total amount can also be generated. Notice the flexibility that DataStudio provides in terms of UI, which allows us to easily modify the table to better suit our objectives.

Analysing the store related data we can see that, as displayed in figure 3, all the physical stores are located in different cities. Each store has a phone number and an exact address which should help us better understand any relevant patterns. A simple table was used to display the data that can be filtered and sorted as needed.

Analysing the data related to invoices, as seen in figure 4, we can see that invoices were generated from three different cities: Naypyidaw, Yangon and Mandalay. Also, it looks like each city has one branch(A, B or C). We deduce that Branch C (Naypyidaw), has the highest gross sale at \$5,265.18 which represents 34.2% of the total gross income. A combination of a pie chart and a bar chart was used to represent percentages and actual values. This gives us a clear picture of the best-performing branch. This type of analysis is very important when making future decisions on business expansion or reduction as less performing stores in a particular area can be changed or closed.

Geospatial analysis is required to better understand the territory coverage. This is particularly important as the distance between our customer and a store has a direct impact on sales. E.g. stores that are closer to our customers perform better.

Plotting the stores on a map as seen in figure 5, we can see that the majority of stores(58.8%) are on the westside of Columbus. Notice that each blue dot represents a store. Having the stores plotted on a map helps us identify multiple insights:

- The routes that a customer has to take to reach a store;
- The distance between each store;
- A store closeness to a motorway or relevant road;
- Public transport routes;
- A store closeness to the city centre;

It is advised to further investigate if there is a particular reason why the majority of stores are on the west side of the city. Looking at the sales data there seems not to be a particular increase in sales in the area.

Payment types are particularly relevant to a business that has physical stores. Recent trends in payment types have changed and are now heading towards online payments and blockchain type payments.

Analyzing the stores by EBT (figure 6), we can see that 94.1% of stores accept EBT and only 5.9% don't. In regards to WIC, we can see that 76.5% of stores accept WIC and 23.5% don't. We used a pie chart to represent this data and we can clearly see the percentages differences.

Members customers are on average more loyal and nonmembers. There is an overall global trend for brands to develop a fan base using social media that really helps to convert customers from non-members to members. This should be a priority as it has a direct impact on revenue.

From figure 7 we can visualise the two types of customers: members and non-members. The members total sales is \$7.820,16 and non-members total sales is \$7.559,21. In conclusion, the member sales are higher by £260.95. This is a very important insight as it should allow us to predict that increasing the membership number should have a direct impact on sales.

Customer retention is often impacted by the company's ability to innovate and keep up with the industry. Diverse and frequently updated product lines are usually a good practice in order to improve customer satisfaction and retention.

Pestron Food has six product lines (as seen in figure 8): health and beauty, electronic accessories, home and lifestyle, sport and travel, food and beverages, fashion accessories. The Food and beverage line is the top seller product type generating \$2,673.56. After analysing the data more in detail it seems like all product types perform almost the same with very small differences. This should be further investigated as the data provided might be incomplete. Data from previous years will be required to further understand the patterns in product line sales over time.

In regards to customer payment type, the data shows a few available options: cash, e-wallet and credit card. All three payment methods generate a similar amount of sales but cash seems to account for 34.7% of the total. There is a global payments trend that indicates that there is an increase in card payments (in-person and online). We recommend all stores to accept all three payment types and for the company to further investigate alternative payment options like blockchain currencies that are nowadays more and more common.

In sales, another very important type of analysis is the 'time of day' breakdown. This provides insights into how customers behave by looking at the time of the day when most sales happen.

Plotting the sales by time of day as seen in figure 9, we can see that 7 PM is the highest-grossing hour of the day. In the afternoon there is a small decline between 4 and 5 PM and 8 PM is the time of the day with the lowest amount of sales. This implies that customers shop more after work, as the sales are increasing after 5 PM.

This pattern is quite common for this type of business. Further analysis by the day of the week and day of the month is required to understand more if there is a pattern in sales by month of the year (e.g. customers buy more in August than September).

Data model

After all of the data has been analysed, a big data model will be created. The Decision Tree Model is a typical approach that may help a firm like Pestron Food comprehend their data better. A decision tree is a diagram or chart that helps in decision-making by illustrating statistical likelihood. The chart is known as a decision tree because of its resemblance to the eponymous plant, and it's often displayed as an upright or horizontal arrangement with branches. Each "branch" in the decision tree represents a possible conclusion, consequence, or response, starting with the choice itself (called a "node"). The tree's furthest branches are the "leaves," which reflect the final consequences of a given decision-making process. Businesses utilise decision trees to explain, map out, and find a solution to a complex problem. Decision trees are widely used to determine a course of action in banking, investing, and business. A decision tree is a graphical depiction of a choice and all of its potential outcomes or repercussions [12].

Knime Analytics Platform will be used to construct the decision tree model for Pestron Food. The first step, as shown in figures 10 and 11, is to import the data from KNIME Analytics. The identical data format as in the CSV will be used to construct two tables. The File Reader node is the first one used, and it reads the most popular text files. The Partitioning Node is linked to the File Reader. The input table is divided into two parts: train data and test data. Figure 15, figure 16, figure 17, and figure 18 show two partitions accessible at the two output ports: First Partition (as defined in dialogue) and Second Partition (remaining rows).

The two partitions output ports are connected with the decision tree learner and decision tree predictor node.

In the main memory, the decision tree learner node creates a categorization decision tree. As the target attribute, the nominal attribute must be utilised. Other decision-making characteristics might be nominal or numerical. Numerical splits are always binary at a given split point, dividing the domain into two parts. Nominal splits can be binary or have any number of outcomes. The Pestron Food Decision Tree Model may be seen in Figures 19 and 20.

The decision tree predictor node, on the other hand, uses an existing decision tree (provided by the model port) to forecast the class value for new patterns (figure 21 and 22).

The last node connected with the Decision Tree Predictor is a node called Scorer. Scorer compares two columns based on their attribute-value pairs and displays the confusion matrix (figure 23 and 24), which indicates how many rows of each attribute and classification match.

Decision tree learner is connected as well with decision tree view. A decision tree with a number of nodes is displayed in this view. The class label and class counts presented inside the nodes match the training data exactly (Figures 25 and 26). Figure 27 shows all the nodes connected.

The decision tree predictor node, on the other hand, uses an existing decision tree (provided by the model port) to forecast the class value for new patterns (figure 21 and 22).

Data Warehouse

For a small company like Petron Foods, a data warehouse solution must be first affordable and scalable. A data warehouse is a sort of data management system that is meant to make corporate intelligence and analytics operations more convenient and efficient. Data warehouses are dedicated to accessing and analysing large volumes of historical data, and they generally keep a lot of it. A data warehouse is a system for storing and organising huge amounts of data from many sources. Its analytical skills help organisations obtain key business insights from their data, allowing them to make better decisions [13].

A Cloud solution is recommended as there is no need for a physical server and for a small company this is ideal. Also, because of the simplicity of the approach, there is no need for a software engineer to conduct the job.

Cloud-based data warehouses offer the same features and benefits as on-premises data warehouses, plus the advantages of cloud computing, such as flexibility, scalability, agility, security, and cost savings. Instead of having to create and operate the data warehouse's hardware and software infrastructure, cloud data warehouses allow organisations to focus solely on extracting value from their data [14].

One of the biggest Cloud providers is Google Cloud. Google Cloud offers a very scalable data warehouse called BigQuery [15].

BigQuery has a few advantages over its competitors:

- It has the ability to scale on Google's global infrastructure, which will help the business in case of rapid expansion;
- The amount of data stored can be virtually unlimited;
- Data hosting is free, as the business grows this is will be really helpful;
- Only pay for query result data size, which can be really an advantage for a small business like Pestron Foods;
- It uses SQL query language, which is a very popular way to query data;
- Its really fast, allowing to process petabytes of data with zero operational headaches;

The first step in uploading Pestron Foods on BigQuery is to create a Google Cloud account. To do so we follow a simple registration process as seen in figure 28 and figure 29, assuming we already have a Gmail account. Once a project has been created we can head to BigQuery and start uploading our data.

To start uploading our data we need to create a dataset first. A dataset can be hosted in one of the many data centres owned by Google. As seen in figure 30 we created a dataset called `Pestron_Foods_Tables` and uploaded our first CSV file with the sales data. Note the "Schema and input parameters" option should be selected so Biguery can automatically detect our dimensions in the CSV file. We can also modify the "Partition and Cluster settings" but for the time being this can be skipped and modified at some later stage once more data is available.

Once the table has been created we can visually see the new SQL table (figure 31). A simple count query can be executed to prove that the data has been successfully loaded:

```
select count(*) from Pestron_Foods_Tables.Sales
```

This should give us the total number of records this new table has. This proves that we successfully loaded all sales data on BigQuery.

The second table will be uploaded in the same way as seen in figure 32.

As seen, both tables are now uploaded and stored for free in the Cloud. This enables incredible flexibility to analyze, process and visualise the data.

BigQuery allows to connect of data from multiple data sources:

- Google Analytics, Pestron Foods website can be linked to cross-reference data and understand if there is any correlation between sales data and website traffic;
- Adsense, Pestron Foods marketing (pay per click) data can be linked to understand the performance of the ads and calculate customer lifetime value;
- Other marketing tools like Facebook Ads, Instagram Ads, LinkedIn Ad can also be connected as a data source with BigQuery to have an overall view of spending on the marketing side of the business;

Uploading CSV files in BigQuery is fast and easy but a very inefficient and not a secure process.

A better approach is recommended:

1. Pestron Foods website data can be imported automatically from a backup in the Cloud, subsequently, no human interaction should be required;
2. Store and personnel data can be uploaded from the internal software system used by the business weekly using Google Cloud Storage as a buffer, subsequently, no human interaction should be required;
3. An overall data visualization dashboard is recommended and an automatic daily email to all stakeholders with the updates over the last 24 hours. This can be achieved using a tool called DataStudio;

Conclusion

As a fast-growing business, Pestron Foods can benefit from a Cloud-based Big Data data solution. It is a smart strategic decision that will really help the growth KPIs set by the director and the CEO.

When setting company KPIs a data visualisation tool can be used. This relates directly to the proposed Cloud-based Data Warehouse solution. Using a Cloud solution means that the KPIs are updated in almost real-time, reducing the time needed for the company to react to sudden market changes. This is particularly important for a business like Pestron Foods as sales can be influenced by many external factors.

The data extraction and storage process can be improved and migration from a CSV approach is necessary. Most modern businesses have their Big Data processes fully integrated and automated. Simple actions like reading backups, linking all the marketing and sales data can really make a difference when setting new KPIs for the business quarter.

The Big Data model will improve and get more accurate as more data is provided. More data means more flexibility, more insights and better decisions by the company. It is recommended that more than sales and stores data should be provided to more accurately predict trends????

Plotting and understanding geospatial data is a crucial step that each business that has a physical presence should do. It allows us to link distance to many factors and to understand the circumstances that can influence the performance of a particular store. E.g. store X performs better than store Y because of how close it is to a bus station. Or how much the distance to a train or tube station influences the ability to recruit personnel. This particular type of insight can be used to strategically plan further expansion in different cities or countries.

Potential fraud can also be detected by looking at trends. Once a particular trend has been identified, for example, sales figures, Pestron Foods can see any anomalies and take action. This can also be applied to other types of financial data relevant to the company. It is strongly recommended to have automatic triggers and notifications when anomalies occur.

Publicly available data on competitors and overall industry should be studied. This should allow a small business to quickly act based on external factors.

Current data on stores doesn't have a unique identifier. This could be a potential issue as the business grows. It is recommended in the future to assign a unique code to each store that will allow us to analyze data across platforms. Most other types should have a unique identifier assigned at the source of generation of the data. For example, each person should have an internal ID, each store incident should have an ID or each product line should have a unique ID.

All significant data available should be aggregated and visualized on a daily basis. Pestron Foods can benefit from a further proposed collection of views:

- A graph view displaying the list of stores ranked by performance
- A chart displaying the total sales in current and past quarter

- A map view visualized all stores by physical location as well as distances to relevant key points(bus stops, train stops, other shops)
- A view displaying all company KPIs

Some of the views are implemented and available on DataStudio at this URL https://datastudio.google.com/reporting/7aa8f892-91a5-4ba1-95ff-62fbda055c6b/page/p_ggdp4fuxlc

The DataStudio report contains various useful insights on Pestron Foods data.

- Sales overview
- Stores overview
- Gross income by city
- Stores by geographical area
- Stores by other criteria(e.g. payment types)
- Store by customer types
- Store by product line
- Sales by the time of day

The provided DataStudio reports are a good starting point for Pestron Foods and are an example of the flexibility of a Cloud-based Big Data approach. Each report can be easily modified to reflect new business requirements. Because all the reports are connected with BigQuery, another benefit of using this approach is that each report updates automatically when the new data is provided and no human interaction is required in the whole process.

Pestron Foods can leverage the latest technologies to transform itself from a small business to a large organization that can truly make a difference to its customers.

Big Data and Cloud Computing can truly make a big difference to any type of business that embraces “change” as part of its core values and understands that technology can help to analyse and simplify large amounts of data faster and cheaper than ever before.

References

- [1] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431–448, Oct. 2018, DOI: 10.1016/j.jksuci.2017.06.001.
- [2] “UWTSD Moodle 2020/21:Log in,” [moodle.uwtsd.ac.uk. https://moodle.uwtsd.ac.uk/pluginfile.php/48872/mod_resource/content/13/ACUA6002_Assessment_June2021%20-%20New%20template.pdf](https://moodle.uwtsd.ac.uk/pluginfile.php/48872/mod_resource/content/13/ACUA6002_Assessment_June2021%20-%20New%20template.pdf) (accessed Jul. 18, 2021).
- [3] Vishesh Arora, “What is Big Data | Introduction and Application Of Big data,” *Analytics Vidhya*, May 20, 2021. <https://www.analyticsvidhya.com/blog/2021/05/what-is-big-data-introduction-uses-and-applications/> (accessed Jul. 13, 2021).
- [4] M. Sultan Mahmud, J. Zhexue Huang, S. Salloum, T. Z. Emara, and K. Sadatdiynov, *A Survey of Data Partitioning and Sampling Methods to Support Big Data Analysis*, vol. 3. TUP, 2020, pp. 85–101.
- [5] R. Herschel and V. M. Miori, “Ethics & Big Data,” *Technology in Society*, vol. 49, pp. 31–36, May 2017, DOI: 10.1016/j.techsoc.2017.03.003.
- [6] L. Hansol, K. Eunkyung, K. Minkyun, and C. Sangmi, “Does Implementation of Big Data Analytics Improve Firms’ Market Value? Investors’ Reaction in Stock Market,” *Uwtsd.ac.UK*, 2021. <https://www-proquest-com.ezproxy.uwtsd.ac.uk/docview/1944303020?accountid=130472&pq-origsite=primo> (accessed Jul. 19, 2021).
- [7] “Big Data for Small Business: A Complete Guide - Talend,” Talend Real-Time Open Source Data Integration Software. <https://www.talend.com/resources/big-data-small-business/#:~:text=Big%20data%20allows%20small%20businesses>. (accessed Jul. 13, 2021)

- [8] C. Jia, D. Jing, Y. Yang, P. Fan, W. Sun, and Y. Feng, "Engineering Big Data to Small Businesses: Lessons Learned from a Case Study," *Uwtsd.ac.uk*, 2021. <https://ieeexplore-ieee-org.ezproxy.uwtsd.ac.uk/document/8632796> (accessed Aug. 13, 2021).
- [9] "Amazon Web Services (AWS) - Cloud Computing Services," Amazon Web Services, Inc., 2018. <https://aws.amazon.com>. (accessed Aug. 13, 2021)
- [10] "UWTSD Moodle 2020/21:Log in," moodle.uwtsd.ac.uk. <https://moodle.uwtsd.ac.uk/course/view.php?id=3765>.
- [11] "Visualization Process - an overview | ScienceDirect Topics," [Sciencedirect.com](https://www.sciencedirect.com), 2017. <https://www.sciencedirect.com/topics/computer-science/visualization-process> (accessed Jan. 28, 2020).
- [12] Y.-Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, pp. 130–5, 2017, DOI: 10.11919/j.issn.1002-0829.215044.
- [13] I. Ahmed, "What is Data Warehousing? Concepts, Tools, Examples," *Astera*, Nov. 06, 2020. <https://www.astera.com/type/blog/what-is-data-warehousing/>. (accessed Aug. 13, 2021)
- [14] "Data Warehouse Concepts: Traditional vs. Cloud," *Panoply*, 2019. <https://panoply.io/data-warehouse-guide/data-warehouse-concepts-traditional-vs-cloud/>. (accessed Aug. 10, 2021)
- [15] "When to Use Google BigQuery," *Toptal Engineering Blog*. <https://www.toptal.com/database/google-bigquery-tutorial>. (accessed Aug. 18, 2021)

Appendices

Create table

Source
Create table from: Upload Select file: ? pestron_sales.csv Browse File format: CSV

Destination
☒ Search for a project ☐ Enter a project name

Project name Pestron Foods **Dataset name** Preston_Foods_Tables **Table type** Native table

Table name
Sales

Schema
Auto detect
☒ Schema and input parameters
Schema will be automatically generated.

Partition and cluster settings
Partitioning No partitioning

Clustering order (optional): ?
Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

Create table Cancel

Figure-1

1	Sales Overview														Report by Bianca Antoci			
2	Invoice_ID	Branch	City	Customer_type	Gender	Product_Line	Unit_price	Quantity	Tax_5_	Total	Date	Time	Payment	cogs	gross_margin_percentage	gross_income	Rating	
	1. 898-04-2717	A	Yangon	Normal	Male	Fashion accessories	76.4	9	34.38	721.98	Mar 19, 2019	15:49	Ewallet	687.6	4.761904762	34.38	7.5	
	2. 896-34-0956	A	Yangon	Normal	Male	Fashion accessories	21.32	1	1.066	22.386	Jan 26, 2019	12:43	Cash	21.32	4.761904762	1.066	5.9	
	3. 895-66-0685	B	Mandalay	Member	Male	Food and beverages	18.08	3	2.712	56.952	Mar 5, 2019	19:46	Ewallet	54.24	4.761904762	2.712	8	
	4. 895-03-6665	B	Mandalay	Normal	Female	Fashion accessories	36.51	9	16.4295	345.0195	Feb 16, 2019	10:52	Cash	328.59	4.761904762	16.4295	4.2	
	5. 894-41-5205	C	Naypyitaw	Normal	Female	Food and beverages	43.18	8	17.272	362.712	Jan 19, 2019	19:39	Credit card	345.44	4.761904762	17.272	8.3	
	6. 892-05-6689	A	Yangon	Normal	Female	Home and lifestyle	28.32	5	7.08	148.68	Mar 11, 2019	13:28	Ewallet	141.6	4.761904762	7.08	6.2	
	7. 891-58-8335	B	Mandalay	Member	Female	Sports and travel	29.61	7	10.3635	217.6335	Mar 11, 2019	15:53	Cash	207.27	4.761904762	10.3635	6.5	
	8. 891-01-7034	B	Mandalay	Normal	Female	Electronic accessories	74.71	6	22.413	470.673	Jan 1, 2019	19:07	Cash	448.26	4.761904762	22.413	6.7	
	9. 889-04-9723	B	Mandalay	Member	Female	Food and beverages	89.14	4	17.828	374.388	Jan 7, 2019	12:20	Credit card	356.56	4.761904762	17.828	7.8	
	1. 888-02-0338	A	Yangon	Normal	Male	Electronic accessories	26.23	9	11.8035	247.8735	Jan 25, 2019	20:24	Ewallet	236.07	4.761904762	11.8035	5.9	
	1. 887-42-0517	C	Naypyitaw	Normal	Female	Sports and travel	83.14	7	29.099	611.079	Jan 10, 2019	10:31	Credit card	581.98	4.761904762	29.099	6.6	
	1. 886-77-9084	C	Naypyitaw	Normal	Male	Electronic accessories	71.89	8	28.756	603.876	Feb 19, 2019	11:33	Ewallet	575.12	4.761904762	28.756	5.5	
	1. 886-54-6089	A	Yangon	Normal	Female	Home and lifestyle	11.43	6	3.429	72.009	Jan 15, 2019	17:24	Cash	68.58	4.761904762	3.429	7.7	
	1. 886-18-2897	A	Yangon	Normal	Female	Food and beverages	56.56	5	14.14	296.94	Mar 22, 2019	19:06	Credit card	282.8	4.761904762	14.14	4.5	
>	1. 885-56-0389	C	Naypyitaw	Member	Male	Fashion accessories	52.35	1	2.6175	54.9675	Feb 12, 2019	17:49	Cash	52.35	4.761904762	2.6175	4	

Figure - 2 Invoices overview

1	Stores Overview									Report by Bianca Antoci	
2	city	state	columbus_area	name	street_address	wic	zip_code	columbus_area	ebt	phone_number	
3	1. Columbus	OH	east	Steward's Foodmart	334 N. 20th Ave	false	43203	east	false	(614) 258-4604	
4	2. Columbus	OH	west	Lena's Market	2305 W. Broad St	false	43204	west	true	(614) 274-6846	
5	3. Columbus	OH	south	Little's Village IGA	704 Parsons Ave	true	43206	south	true	(614) 449-7703	
	4. Columbus	OH	south	Mae's Produce	440 Marion Rd	false	43207	south	true	(614) 445-7236	
	5. Columbus	OH	north	North Linden Market	1580 Joyce Ave	false	43219	north	true	(614) 804-5380	
	6. Columbus	OH	west	Mound St. Market	1690 W. Mound St	false	43223	west	true	(614) 947-0958	
	7. Columbus	OH	west	Broad St. Beverage	2403 W. Broad St	true	43223	west	true	(614) 276-8155	
	8. Columbus	OH	west	Belvidere Mini-Mart	126 Belvidere Ave	false	43223	west	true	(614) 279-2342	
	9. Columbus	OH	west	Herbert's Market	1121 Sullivant Ave	false	43223	west	true	(614) 224-8874	
	10. Columbus	OH	west	M&J Carryout	1958 Sullivant Ave	true	43223	west	true	(614) 351-1922	
	11. Columbus	OH	west	N&N Supermarket	2240 Sullivant Ave	false	43223	west	true	(614) 272-1670	
	12. Columbus	OH	west	Matan Grocery	699H Harrisburg Pike	true	43223	west	true	(614) 279-3224	
	13. Columbus	OH	west	Rosco's Market	2500 Sullivant Ave	false	43223	west	true	(614) 725-3424	
	14. Columbus	OH	north	Stafford's Market	2204 Parkwood Ave	false	43211	north	true	(614) 476-6400	
	15. Columbus	OH	west	Pat's Hague and Steele	154 N. Hague Ave	false	43204	west	true	(614) 279-2342	
	16. Columbus	OH	east	Five Brothers Carryout	1230 E. Long St	false	43203	east	true	(614) 252-1106	
	17. Columbus	OH	east	Long St. Carryout	1032 E. Long St	false	43203	east	true	(614) 258-5306	

Figure-3 Stores overview

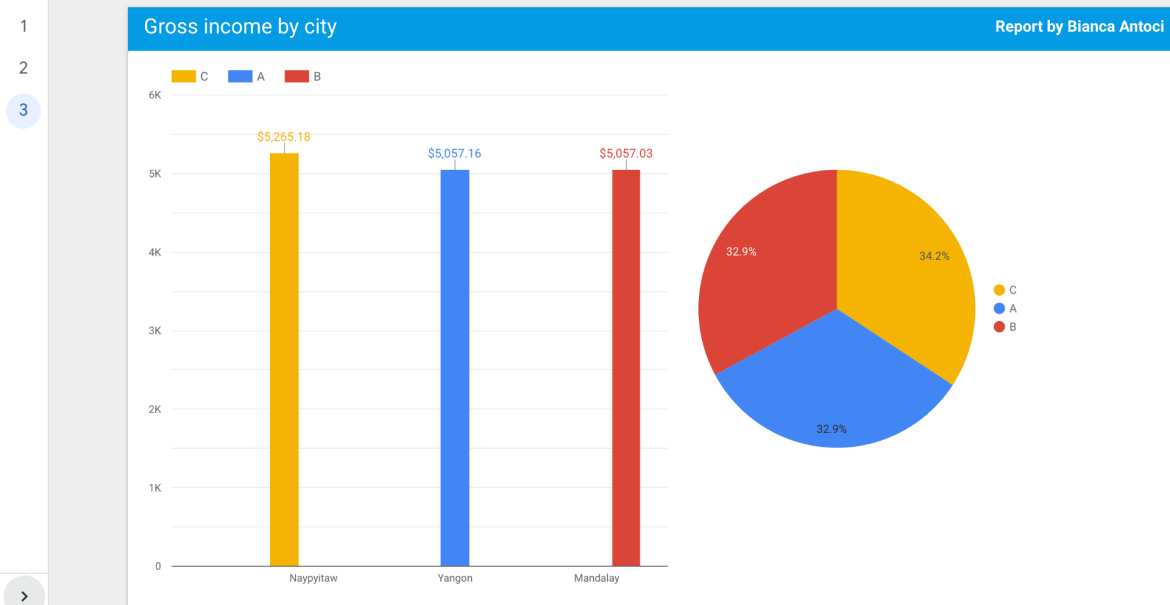


Figure-4 Gross income by city

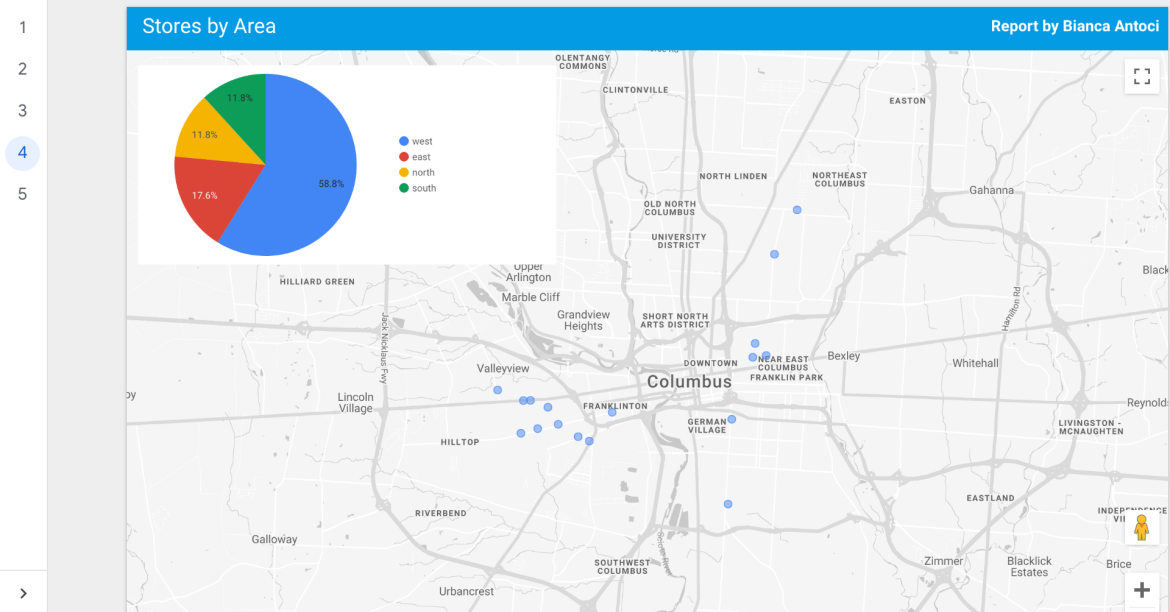


Figure-5 Stores plotted on a map

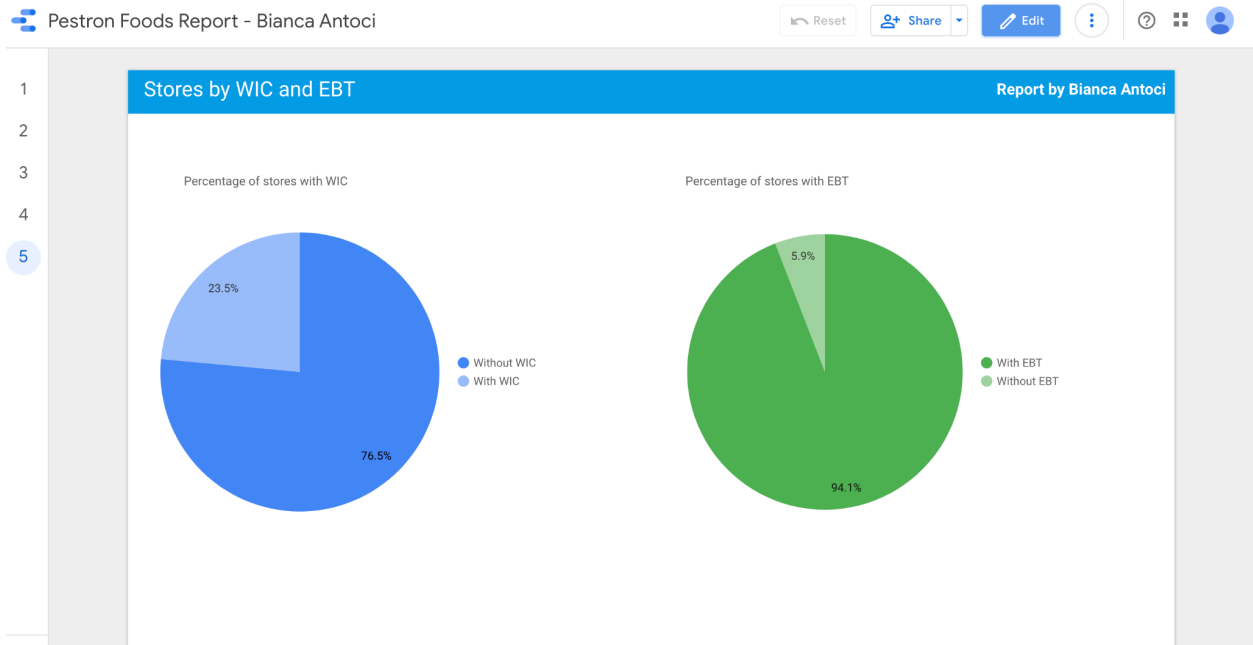


Figure-6 Stores WIC and EBT payments

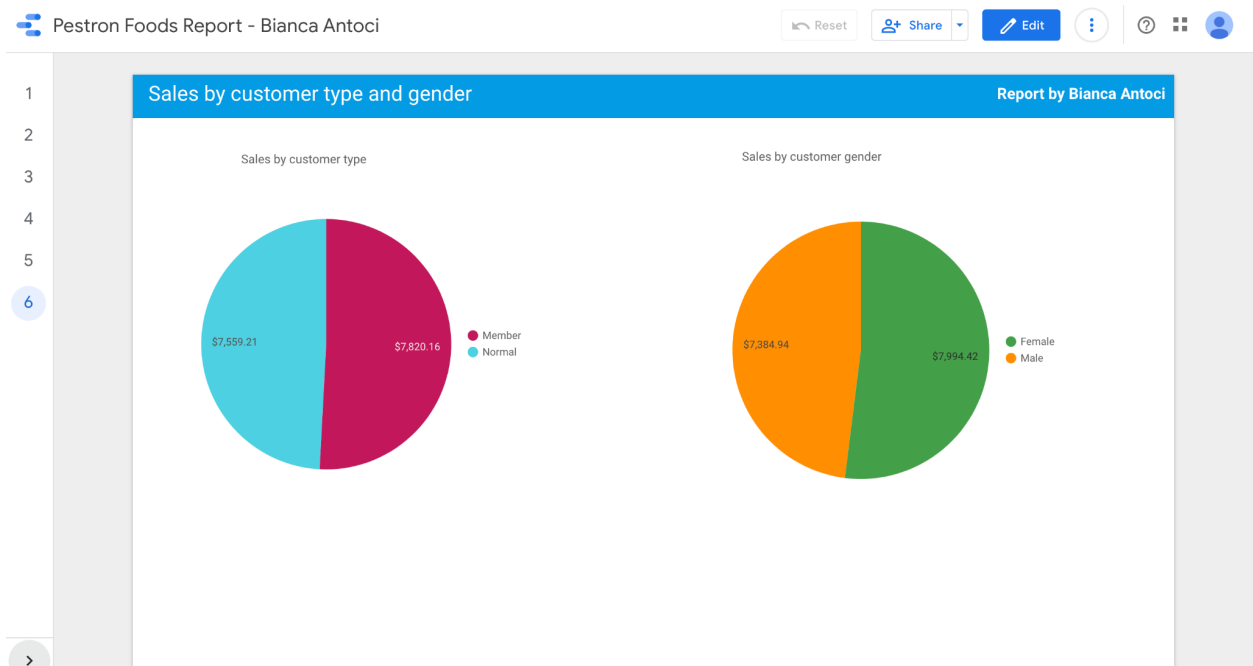


Figure-7 Pestron Food customer types

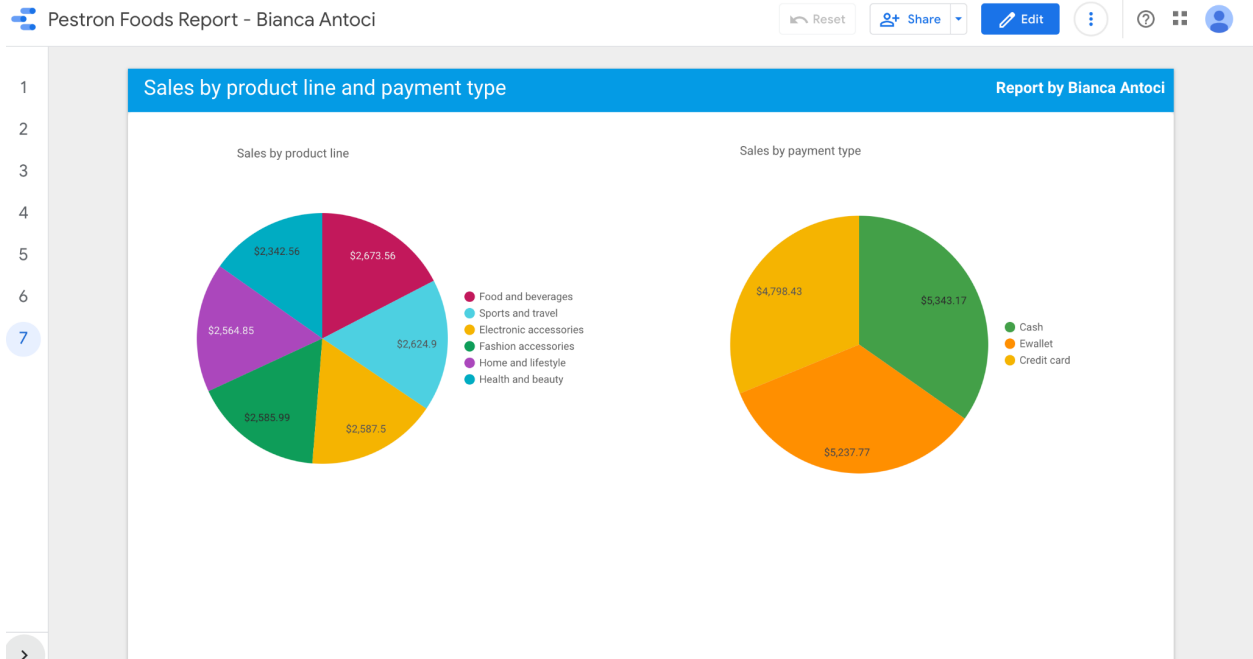


Figure-8 Stores product lines and payment types

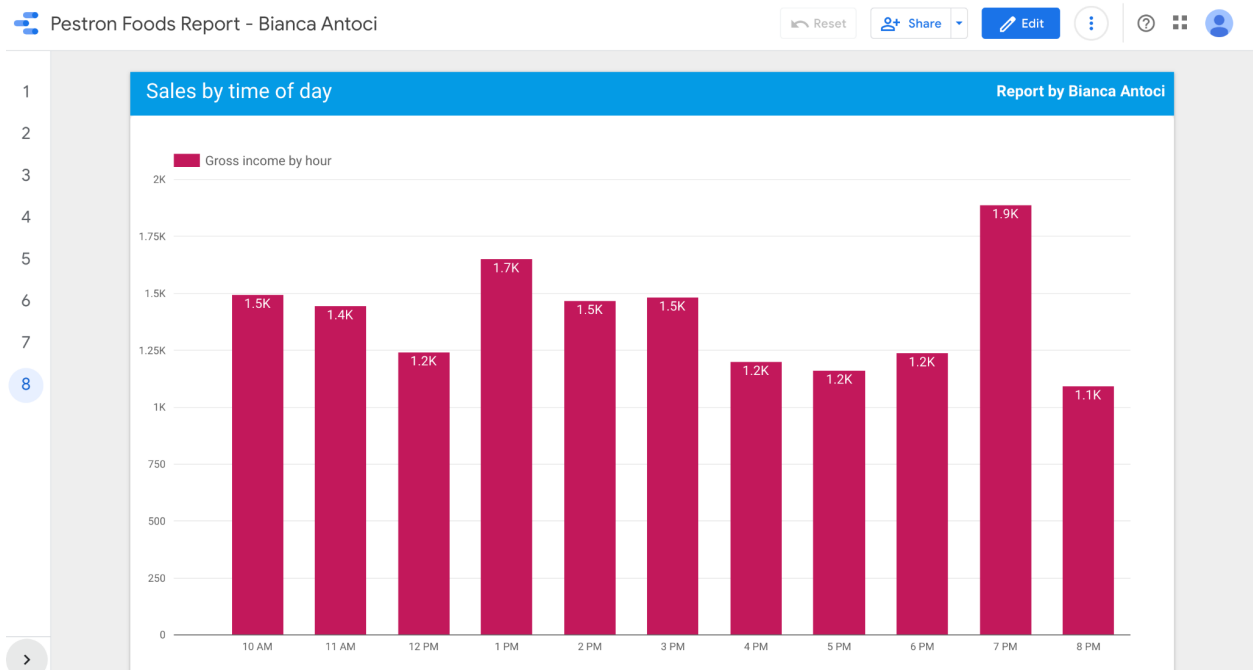


Figure-9 Time of day breakdown

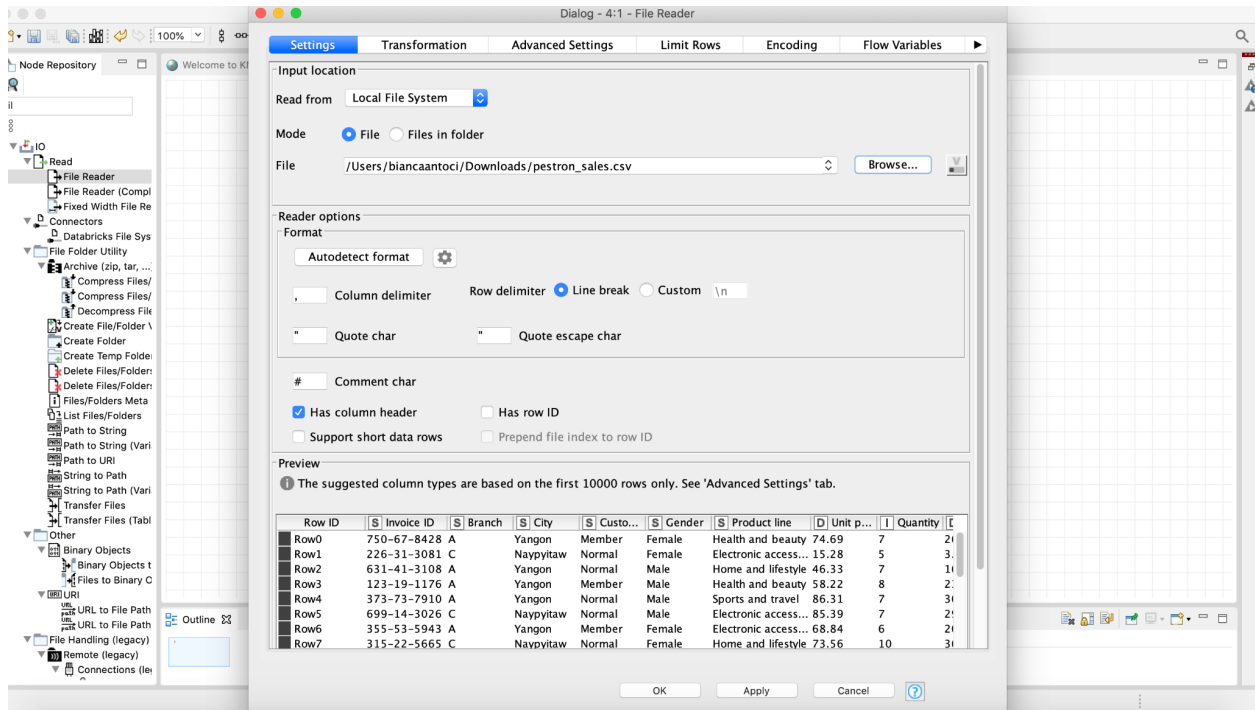


Figure-10 File Reader Pestrone Food Sales

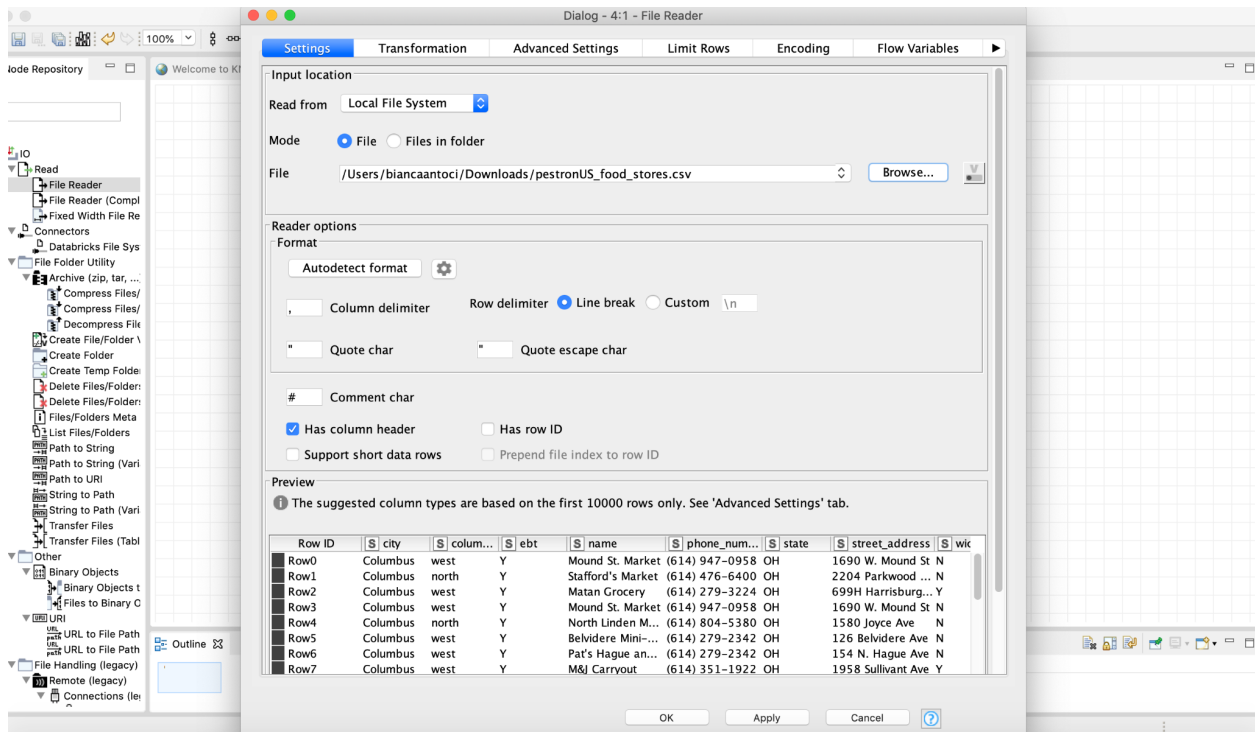


Figure-11 File Reader Pestrone Food Stores

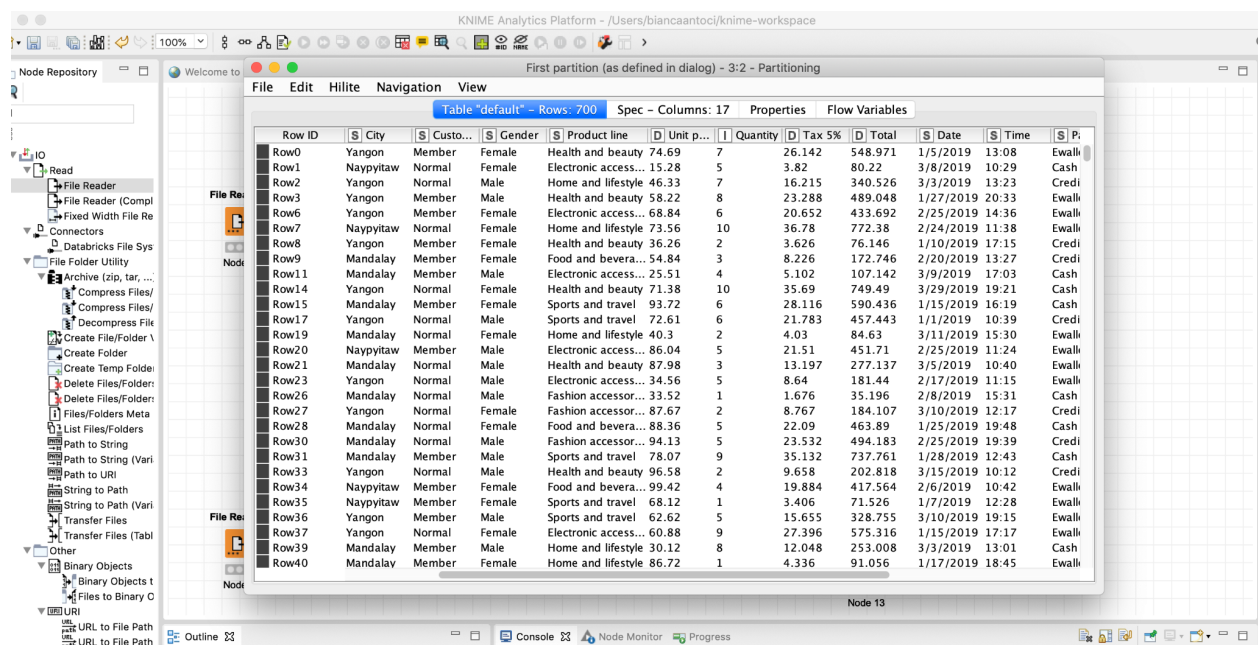


Figure-12 First Partition Pestron Food Sales

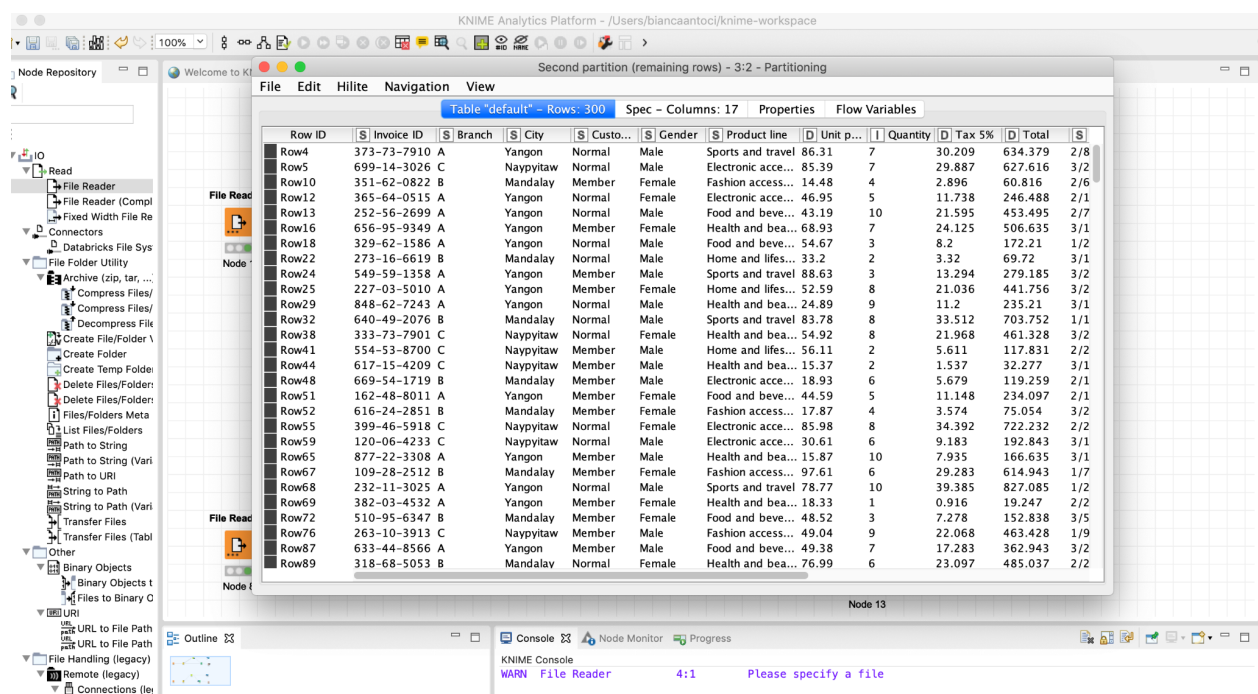


Figure-13 Second Partition Pestron Food Sales

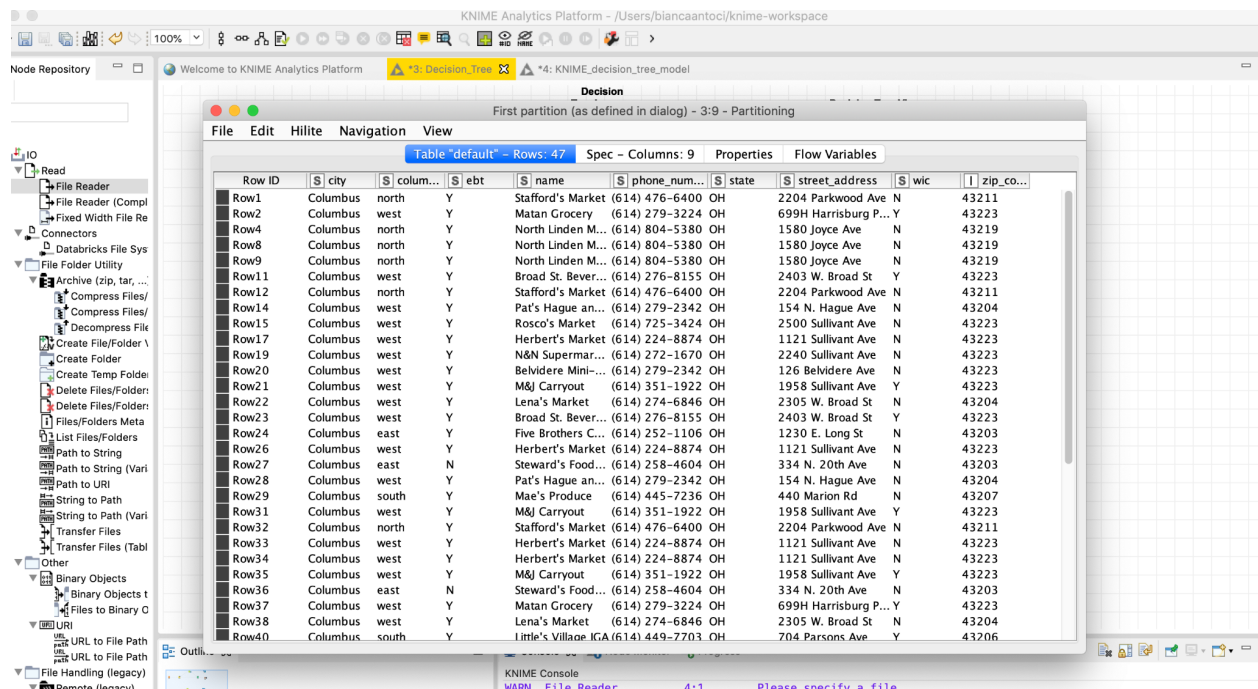


Figure-14 First Partition Pestron Food Stores

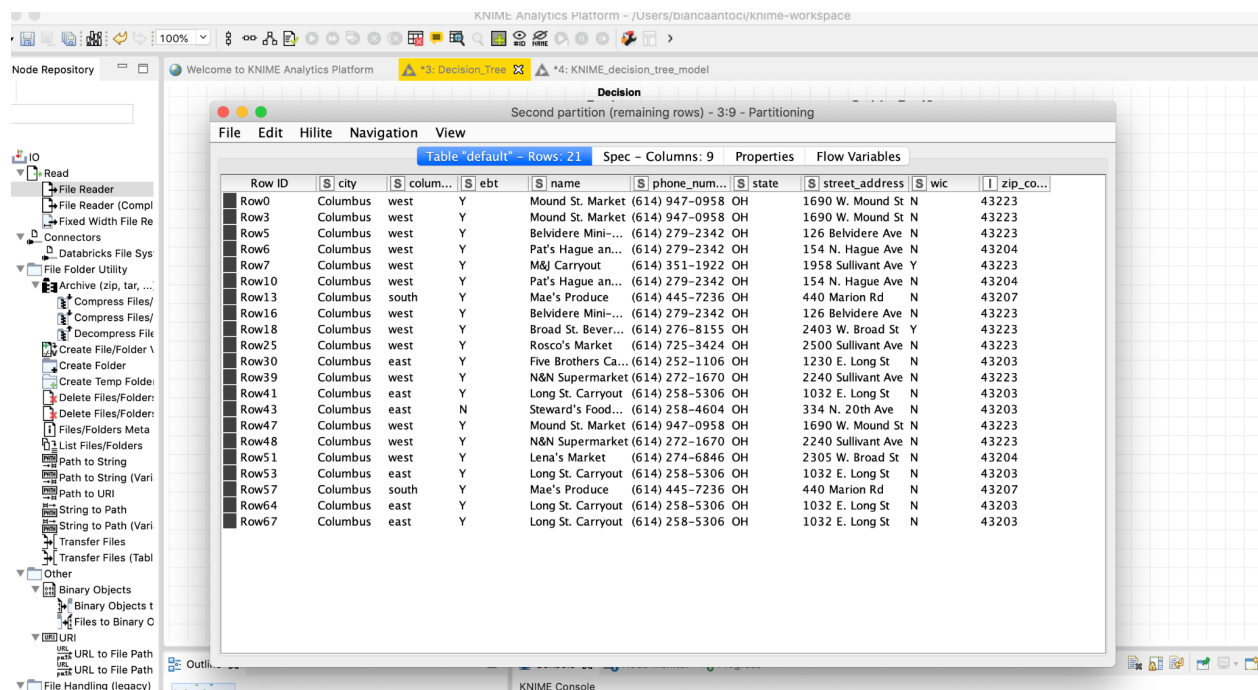


Figure-15 Second Partition Pestron Food Stores

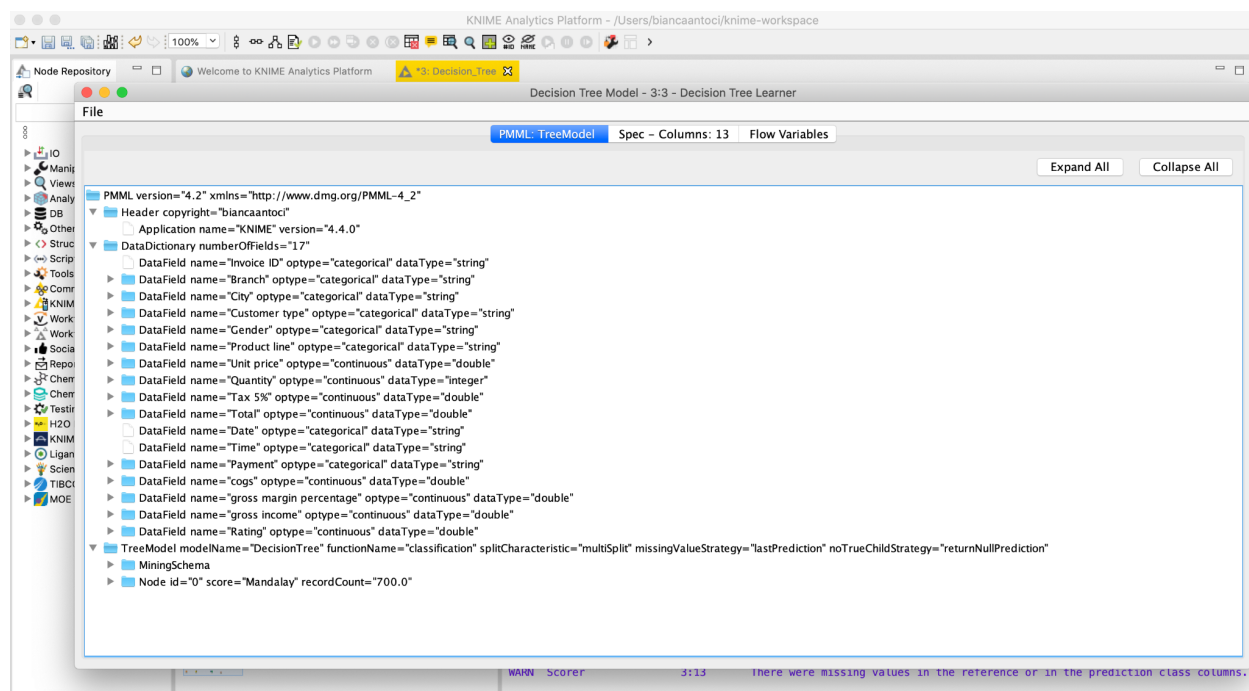


Fig-16 Decision tree model Pestron Food Sales

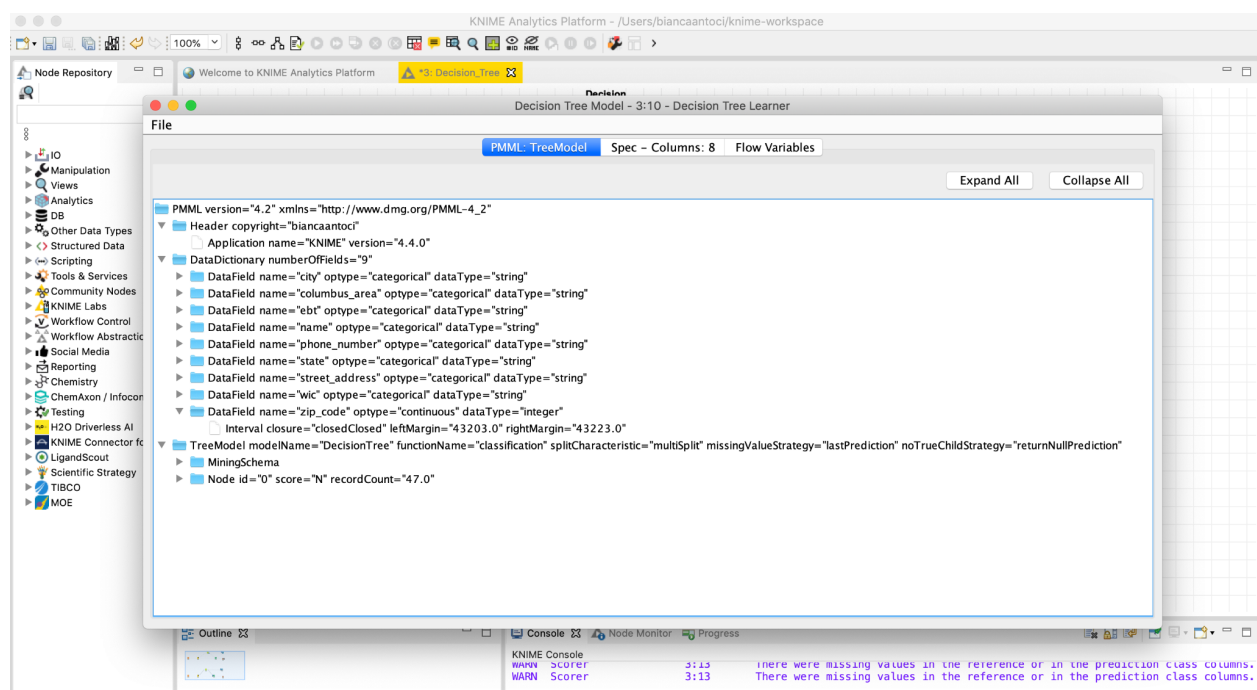


Figure-17 Decision tree model Pestron Food Stores

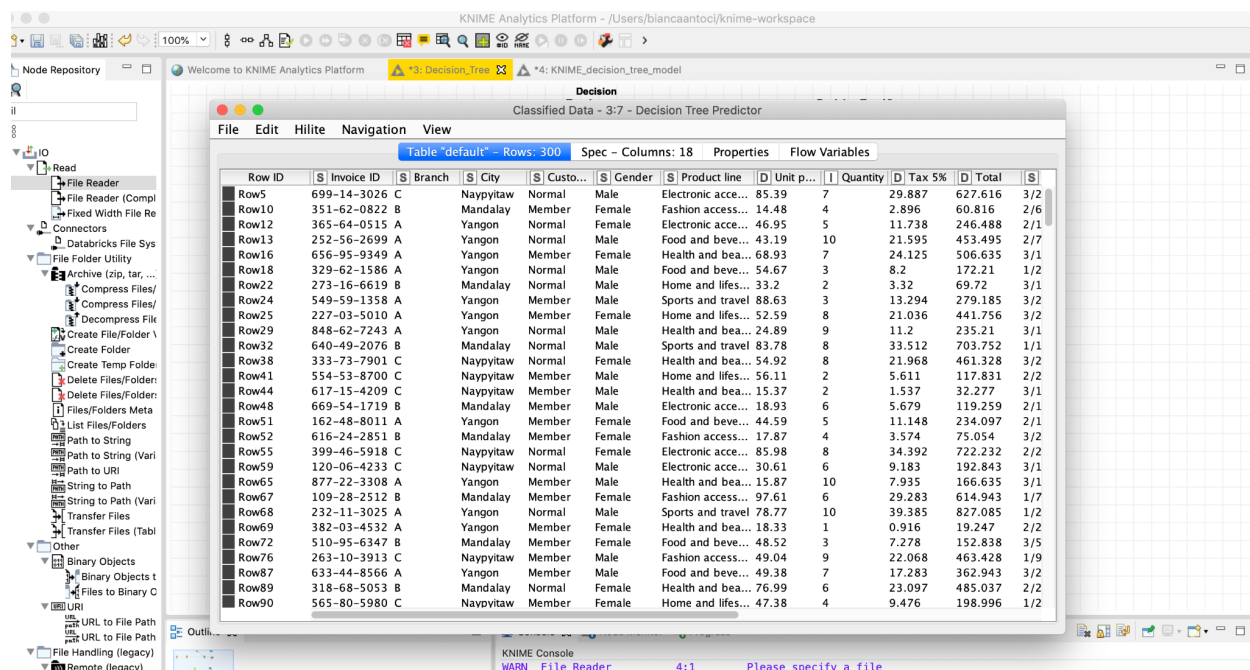


Figure-18 Decision Tree Predictor Pestron Food Sales

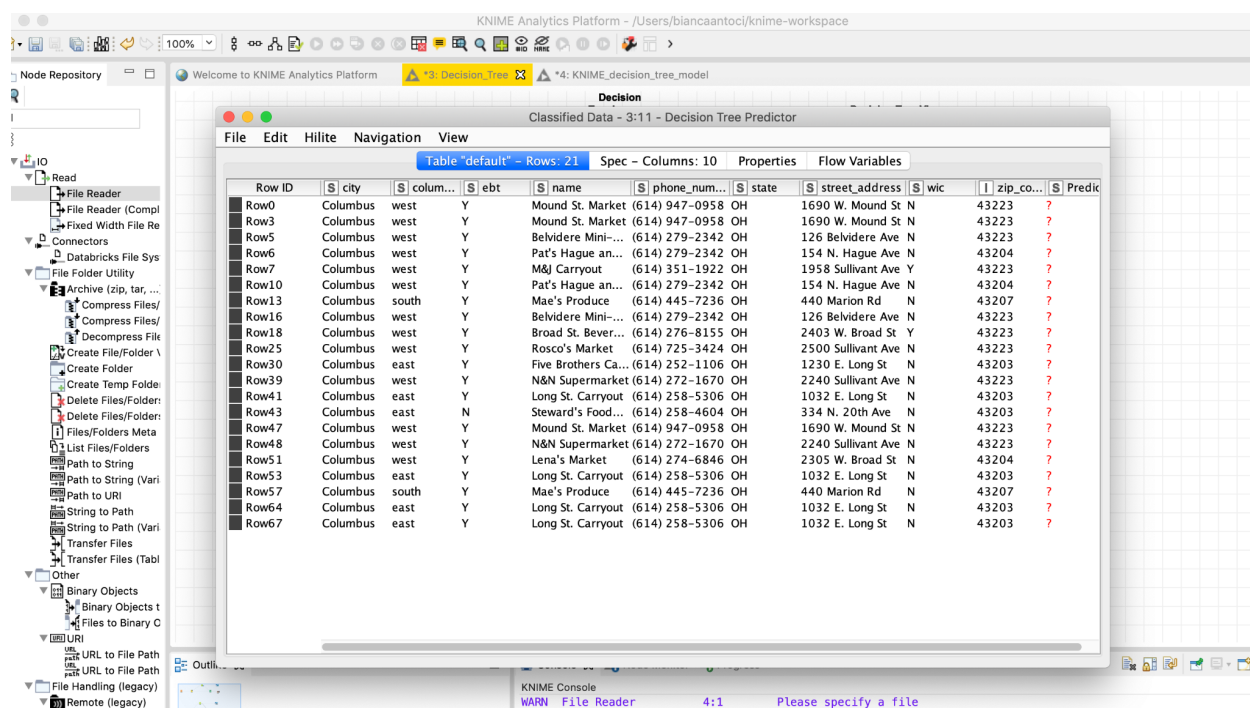


Figure-19 Decision Tree Predictor Pestron Food Stores

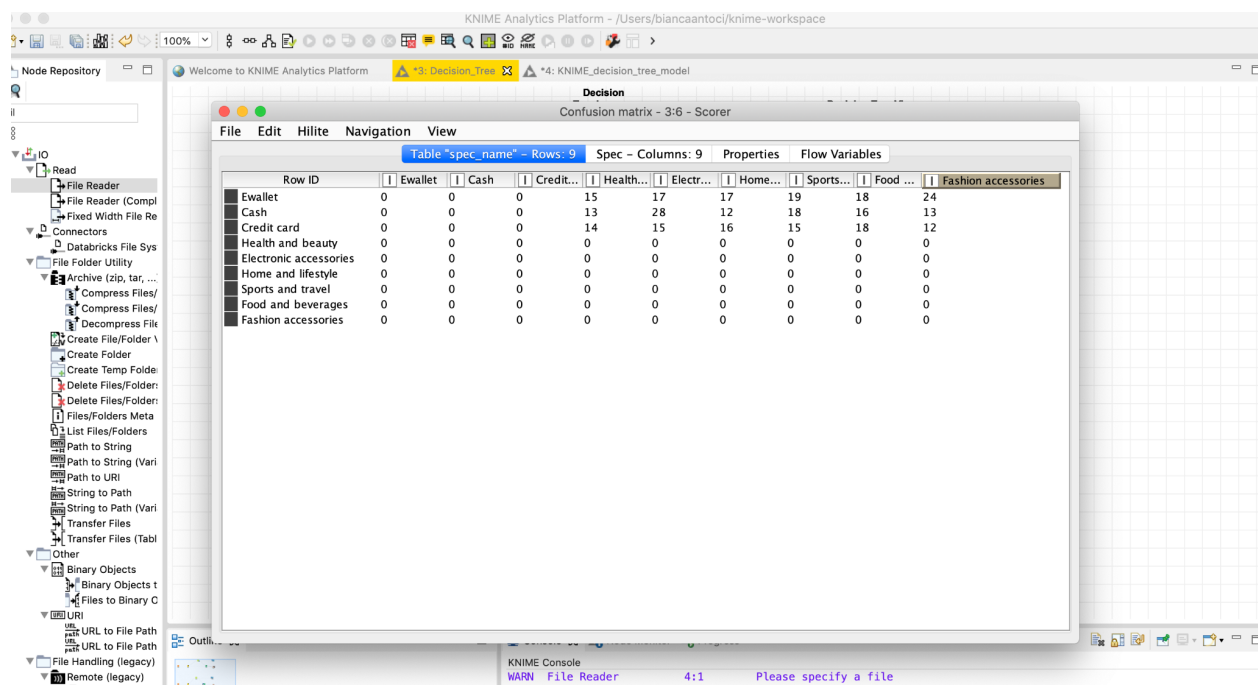


Figure-23 Confusion matrix Pestron Food Sales

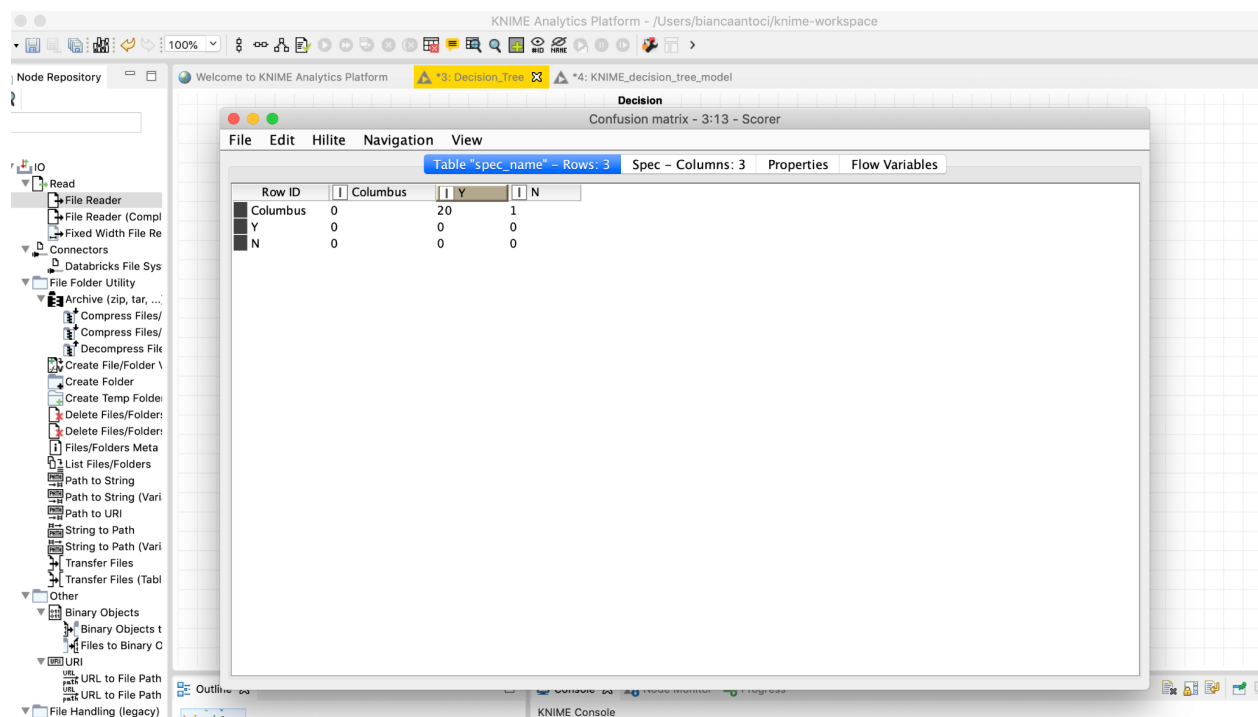


Figure-24 Confusion matrix Pestron Food Stores

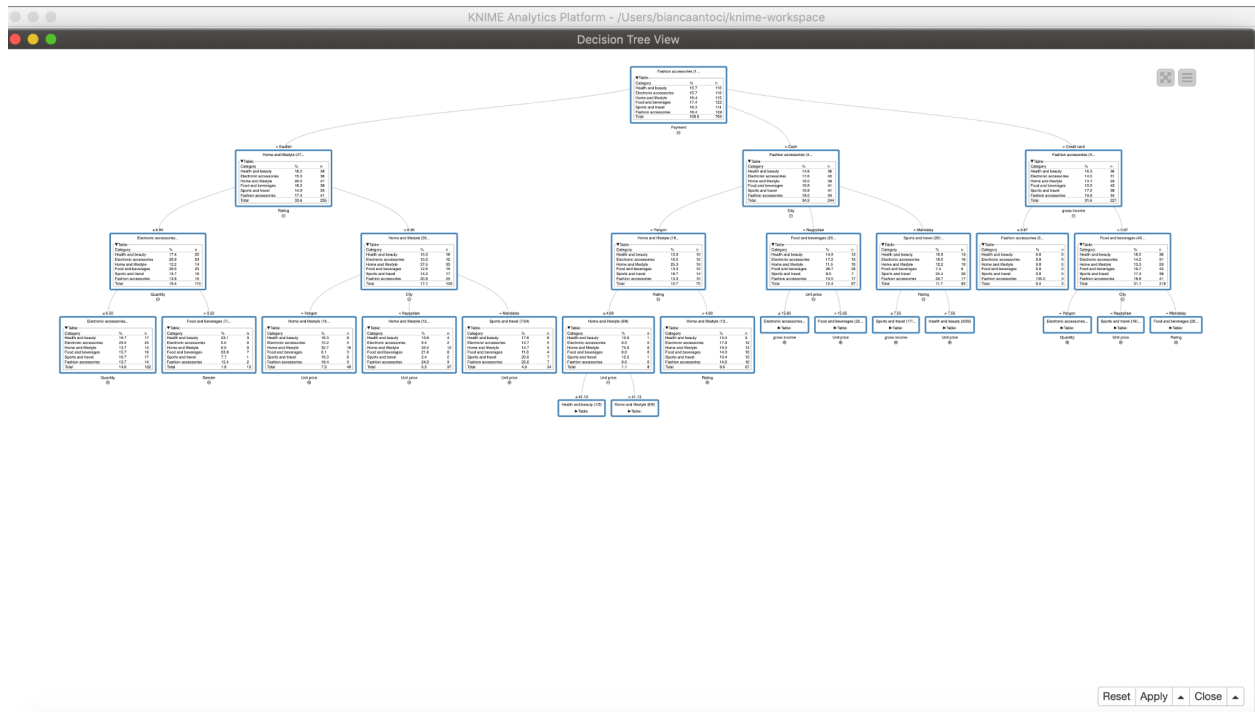


Figure-25 Decision Tree view Pestron Food Sales

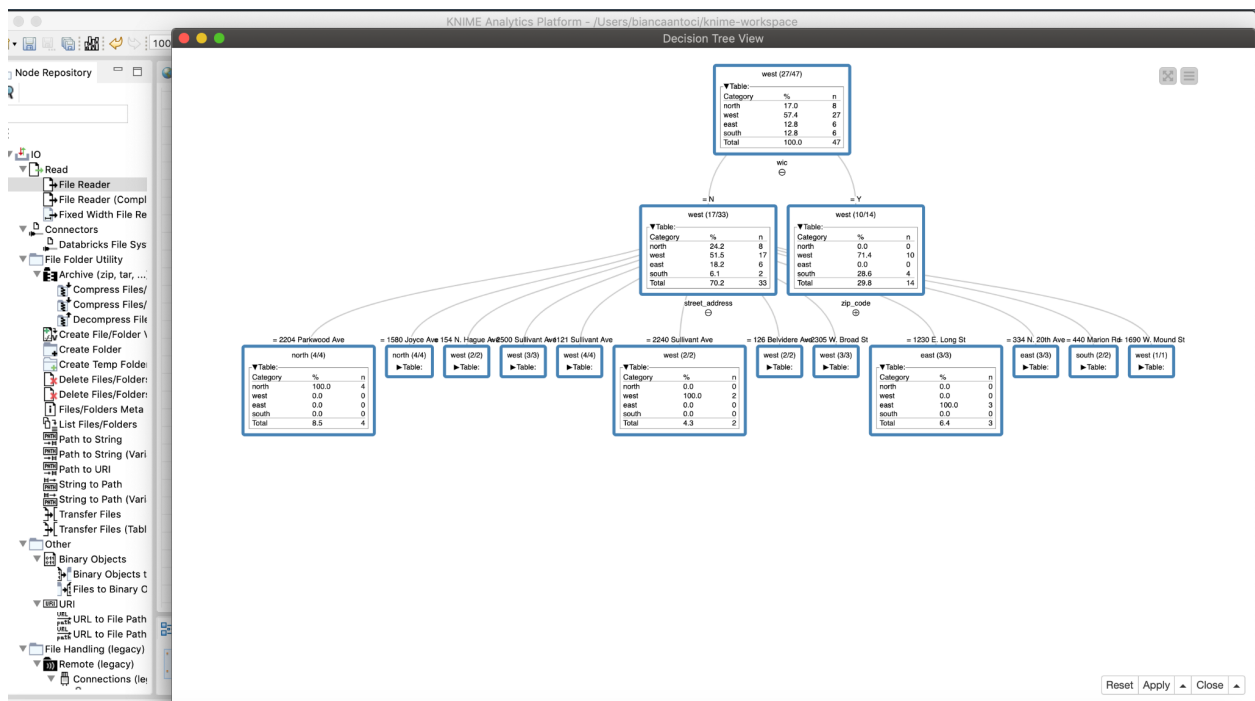


Figure-25 Decision Tree view Pestron Food Stores

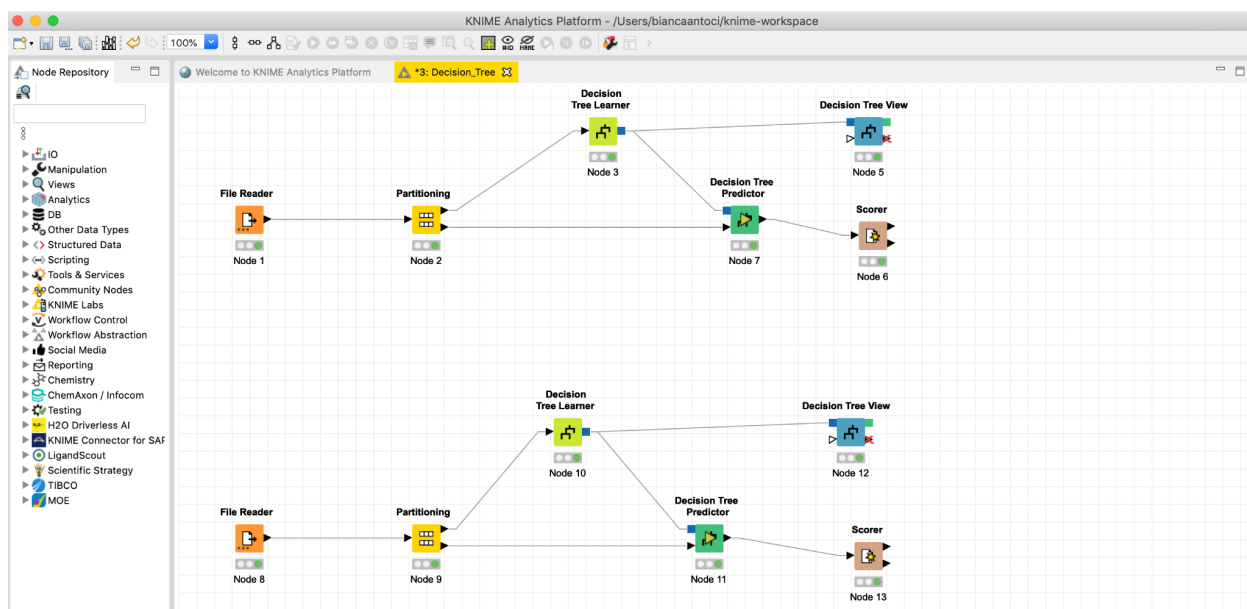


Figure-27 Decision Tree nodes

Google Cloud

Why Google Solutions Products Pricing Getting Started [Contact Us](#) [Docs](#) [Support](#) [Language](#) [Sign in](#) [Get started for free](#)

New customers get \$300 in free credits to spend on Google Cloud. All customers get free usage of 20+ products. [See offer details.](#)

Accelerate your transformation with Google Cloud

Build apps faster, make smarter business decisions, and connect people anywhere.

[Get started for free](#)

Google Cloud Next '21

Join our global digital experience

Streaming live • October 12-14

[Register now](#)

Choose from [100+ free sessions, demos, and moments to learn](#) from Google experts.

Figure-28 Google Cloud registration

The screenshot shows the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'Pestron Foods', and a search bar. The left sidebar contains the 'Explorer' panel with a search bar and a list of projects. The main area displays a SQL query editor with a query that selects data from the 'Preston_Foods_Tables.Sales' table, limited to 5 rows. Below the query editor, the 'Query results' section shows the execution status and a table of results.

Query complete (0.3 sec elapsed, 145.9 KB processed)

Job information Results JSON Execution details

Row	Invoice_ID	Branch	City	Customer_type	Gender	Product_line	Unit_price	Quantity	Tax_5_	Total	Date	Time	Payment	cogs	gross_margin_
1	382-03-4532	A	Yangon	Member	Female	Health and beauty	18.33	1	0.9165	19.2465	2019-02-02	18:50	Cash	18.33	4.0
2	504-35-8843	A	Yangon	Normal	Male	Sports and travel	42.47	1	2.1235	44.5935	2019-01-02	16:57	Cash	42.47	4.0
3	594-34-4444	A	Yangon	Normal	Male	Electronic accessories	97.16	1	4.858	102.018	2019-03-08	20:38	Ewallet	97.16	4.0
4	275-28-0149	A	Yangon	Normal	Male	Sports and travel	63.69	1	3.1845	66.8745	2019-02-25	16:21	Cash	63.69	4.0
5	400-60-7251	A	Yangon	Normal	Male	Home and lifestyle	74.07	1	3.7035	77.7735	2019-02-10	12:50	Ewallet	74.07	4.0

Figure-31 BigQuery SQL query result

The screenshot shows the 'Create table' dialog in the Google Cloud Platform BigQuery interface. The dialog is divided into several sections: Source, Destination, Schema, and Partition and cluster settings. The 'Source' section shows the file 'pestronUS_food_stores.csv' selected for upload. The 'Destination' section shows the project 'Pestron Foods' and dataset 'Preston_Foods_Tables' selected. The 'Schema' section shows the 'Table type' set to 'Native table'. The 'Partition and cluster settings' section shows 'No partitioning' selected. The dialog includes a 'Create table' button and a 'Cancel' button.

Create table

Source

Create table from: Select file: File format:

Destination

☒ Search for a project ☐ Enter a project name

Project name: Dataset name: Table type:

Table name:

Schema

Auto detect ☒ Schema and input parameters

Schema will be automatically generated.

Partition and cluster settings

Partitioning:

Figure-32 BigQuery new table Stores

