Course Project

# **Name**

# 

# **UNIVERSITY**

# 

# **Subject**

# 

# **Professor**

1. **Alteryx – The data mining tool**

Alteryx is a data analytics and business intelligence software that allows users to connect to, prepare, blend, and analyze data from multiple sources. It is a platform for self-service data analytics that offers a drag-and-drop user interface for workflows related to blending, prepping, and analyzing data.

Its capacity to execute complicated data preparation and blending, link to a variety of data sources, and perform advanced analytics, such as statistical and geographic analysis, predictive modelling, and machine learning, make Alteryx a well-known product. Alteryx additionally offers a vast selection of pre-built data functions and tools for data visualization, making it simple for users to derive insights from their data. Professionals in business intelligence, data analysis, and data science frequently use Alteryx to examine,

It is a robust data mining tool with a user-friendly drag-and-drop interface that is available to both technical and non-technical users. It also includes a large range of pre-built data transformation and analysis features. The program can be connected with other tools and platforms to broaden its capabilities. It is made to assist users in deriving insights and knowledge from huge and complicated datasets.





Fig.1. Alteryx

**Advantages:**

* **Drag-and-drop user interface**: Alteryx offers a straightforward and intuitive user interface that enables users to carry out data manipulations and analytics fast and easily without having to write code.
* **A comprehensive range of pre-built data transformation and analysis functions:**- Alteryx, making it simple for users to conduct complicated data manipulations without writing code. Alteryx also contains a wide range of pre-built functions for data cleaning, transformation, and analysis.
* **Advanced data visualisation features**: Alteryx comes with several tools for data visualisation and data exploration, making it simple for users to comprehend and convey the insights they get from their data.
* **Integrates with other tools:** To expand its functionality and enable more sophisticated data analysis, Alteryx may be integrated with a range of other tools and platforms, including Tableau, R, and Python.
* **Easy team collaboration**: Alteryx is a wonderful option for enterprises with numerous stakeholders engaged in data analysis since it enables teams to collaborate and share their work easily and effectively.
* **Data preparation and quality**: Only timely and accurate data are valuable. Bad judgements can be made as a result of poor data, which can hurt an organization's bottom line. To guarantee that your data is prepared for analysis, the Alteryx platform offers extensive data quality capabilities.
* **Data Science and Decisions**: The Alteryx platform provides many data science capabilities, such as ML and AI. Without having any coding or analytics experience, you may use these tools to develop models that make predictions or suggestions or uncover trends in your data.
* **Outcome Automations**: After creating your models, you must put them into use so that they can begin making decisions on your behalf. There are various tools on the Alteryx platform for automating the deployment and execution of analytics.

1. **Why Alteryx is known as data mining tool?**

Because it has many features and functionalities that are especially made to assist users in deriving knowledge and insights from sizable and complex datasets, Alteryx is referred to as a data mining tool.

These features include:

1. **Data cleaning and preparation:** Alteryx comes with a number of tools for cleaning and preparing data, including removing duplicates, filling in blanks, and standardising data formats. Users now find it simple to work with big, disorganised datasets.
2. **Data exploration and visualization**: Histograms, scatter plots, and heat maps are just a few of the tools that Alteryx provides for studying and visualising data. Users can easily comprehend and share the insights they derive from their data thanks to these tools.
3. **Data modelling and analytics**: Alteryx comes with a large selection of pre-built data modelling and analytics features, including decision trees, logistic and linear regression, and k-means clustering. Users can easily perform complex data analysis using these services without having to write any code.
4. **Automation:** Alteryx also provides the option to automate the process of data processing, transformation, and analysis, which can reduce errors and save time.
5. **Support for vast and varied data**: Alteryx supports a wide range of data sources, including databases, spreadsheets, and APIs. It can manage massive amounts of data. Due to its adaptability, it can be used for data analysis and mining regardless of the nature or complexity of the data.

With all of these features and capabilities, Alteryx is a strong data mining and analytics solution that enables users to quickly and simply extract information and insights from huge and complicated datasets.

1. **Business reasons for selecting Alteryx as a data mining tool**

Alteryx was chosen as a data mining tool for a number of commercial reasons, including:

* **Better decision-making**: Alteryx makes it simple for users to extract knowledge and insights from significant and complicated datasets, which can aid in better strategic planning and decision-making within a company.
* **Efficiency:** Alteryx may assist boost efficiency and eliminate errors by automating the data manipulation, transformation, and analysis process, which can eventually save time and money.
* **Better data visualization**: Alteryx offers a number of tools for data exploration and visualization that can improve users' understanding of and ability to convey the insights they get from their data. Through better communication and teamwork, an organization may benefit.
* **Greater scalability:** Alteryx can manage enormous amounts of data and it scales quickly as data volume and analysis complexity increase, enabling businesses to stay up with data expansion and avoid being constrained by a solution that can't handle their data size.
* **Better understanding of the customer**: Alteryx enables organizations to better comprehend their customers by analyzing customer data from a range of sources, including social media, CRM systems, and purchase history. This might assist companies in enhancing client engagement and boosting sales.
* **Improved risk management**: Alteryx may assist companies in identifying potential hazards by analyzing data from many sources, including financial data, market data, and social media data and make better decisions.

1. **Data mining project topic - Predicting Catalog Demand**

In this project, I'll examine a commercial issue in the mail-order catalogue industry. We have to estimate the revenue our business will generate from distributing a catalogue to new clients. In order to make a suggestion to management, this activity will require developing the model and putting the findings to use.

**The Business Problem**

Assume My new job is with a business that produces and distributes high-end home items. The business published its first print catalogue last year, and it is getting ready to publish this year's catalogue in the coming months. The business wants to send the catalogue to 250 new clients on its mailing list.

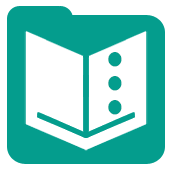
My Manager has been tasked with estimating the amount of revenue the business can generate from mailing a catalogue to these clients. The business analyst is charged with assisting our manager with the calculations. Our manager is fairly competent about data analysis, but less so about predictive models.

I've been asked to estimate the anticipated revenue from these 250 new clients. The catalogue won't be distributed to these new clients unless the expected profit contribution is greater than $10,000, according to management.

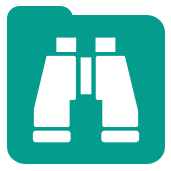
***Details***

Each catalogue has a $6.50 printing and distribution expense.

For all items sold through the catalogue, the average gross margin (price minus cost) is 50%.

1. **Alteryx Tools used in the Project**
2. **The Input Tool** 

Data is added to the workflow using the Input tool from a variety of sources, including files, databases, and APIs. You can use this tool to provide your data source's location, connection information, and any required credentials. Once linked, the input tool will read the data and make it accessible for additional workflow processing.

1. **The Browse Tool **

You can see and examine the data in the workflow using the Browse tool without actually changing it. You can use this tool to view the data in a tabular style and to filter and sort it for easier comprehension.

1. **The Select Tool **

Data rows can be filtered using the Select tool according to various criteria. With the help of this tool, you may define one or more requirements that a row must meet in order to appear in the output. This is helpful for eliminating extraneous or pointless data from your analysis.

1. **The Unique Tool **

Duplicate rows of data are removed using the Unique tool. This tool eliminates any identical rows by comparing each row to the one before it. This might be helpful for cleaning up data that has been manually entered or imported from different sources.

1. **The Interactive Charts** 

You may generate interactive visualizations of your data, such as bar charts, line charts, and scatter plots, using the Interactive Chart tool. This tool can be used to examine your data and look for trends and connections.

1. **The Summarize Tool **

To combine data and generate summary statistics like sums, averages, and counts, utilise the Summarize tool. With the help of this tool, you may group data according to one or more fields and then compute various summary statistics for each group.

1. **The Formula Tool **

You can add new fields to your data collection and conduct calculations on your data using the Formula tool. With the help of this tool, you may build calculated fields that can be utilised for further in-depth analysis using a wide range of functions and operators.

1. **The Linear Regression Tool **

You can anticipate future values based on historical data by using the Linear Regression tool to do a linear regression study on your data. This tool fits a linear equation to the data using the least squares method, and it also computes numerous statistics and metrics to assist you in assessing the model's fit.

1. **Steps for Predicting Catalog Demand**

**Step 1: Business and Data Understanding**

**Decisions needs to be made –**

The company wants to give some prospective clients a catalogue, but they will only do so if they would contribute more than $10,000 in profits. To accomplish this, the business must forecast the anticipated revenue from these new clients.

**Data is needed to inform those decisions –**

We have been given two datasets, p1-customers.xlxs and p1-mailinglist.xlxs, to help us make those decisions. This indicates that the Company already has some data in the first dataset for more than 2,000 consumers. The p1-mailinglist dataset's linear regression model may be used to estimate the profit from these new customers since this information, which is the Sales made by the Company for the Company's current Customers, can be thought of as numerical data. While the mailing dataset was used to generate predictions and validate the model, the customers dataset was used to create or train the model.

**Step 2: Analysis, Modeling, and Validation**

In your model, how and why did I choose the predictor variables?

Avg Num Products Purchased and Customer segment had p-values greater than 0.05, indicating there was no true association between these predictor factors and the observation when I ran the model with practically all of the predictor variables. In their place, I chose Customer segment and Avg Num Products Purchased, which have extremely high predictability for the target variable. I then eliminated irrelevant variables that had less predictive potential. The degree of correlation between the predictor variables and the target variable is depicted in the visualizations below.

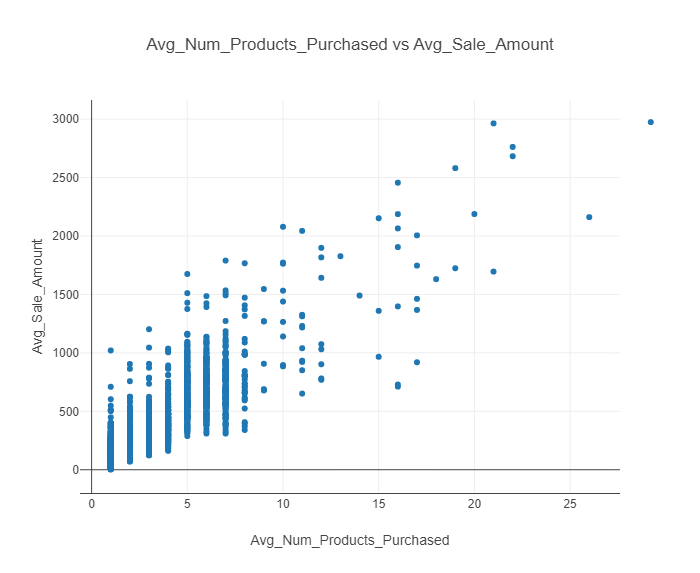
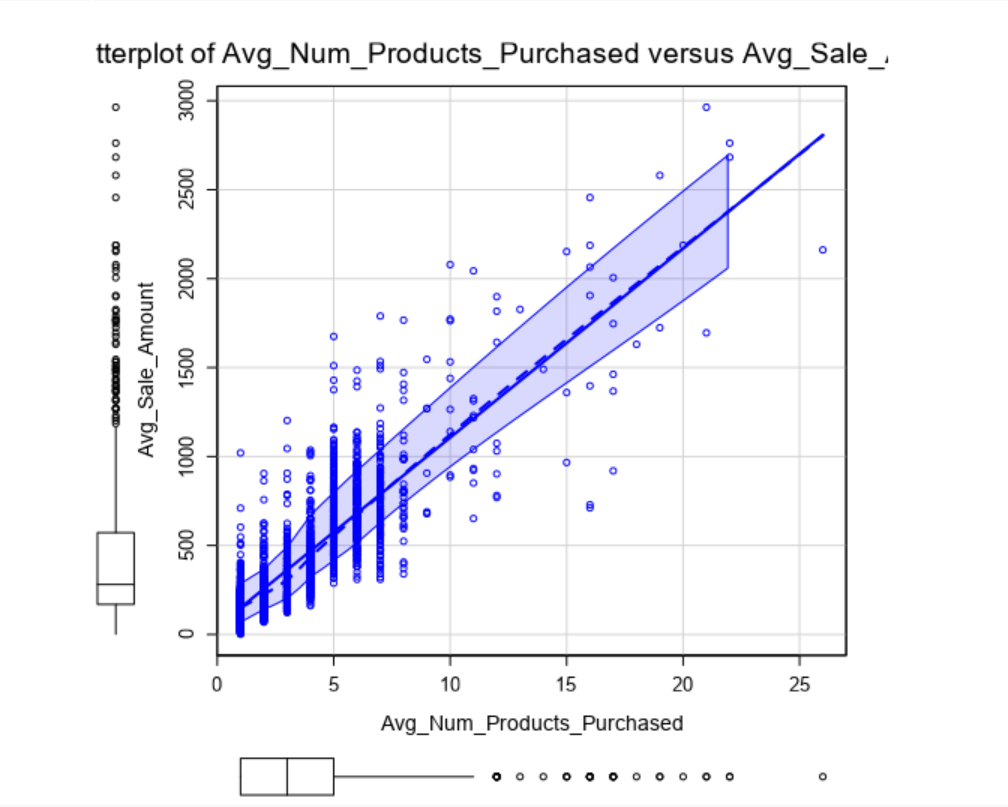
Fig .2. Graph between Average Sales Amount and Average Number of Products Purchased 

Fig.3. Graph between Average Sales Amount and Average Number of Products Purchased

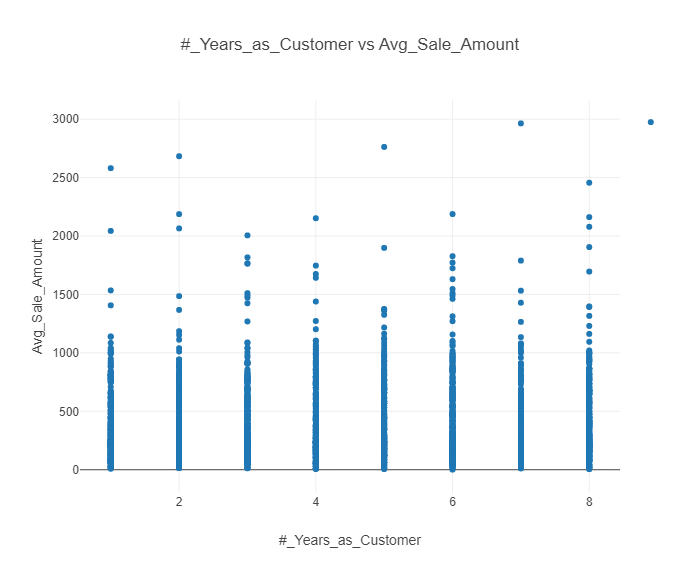


Fig .4. Graph between Average Sales Amount and Years as Customers

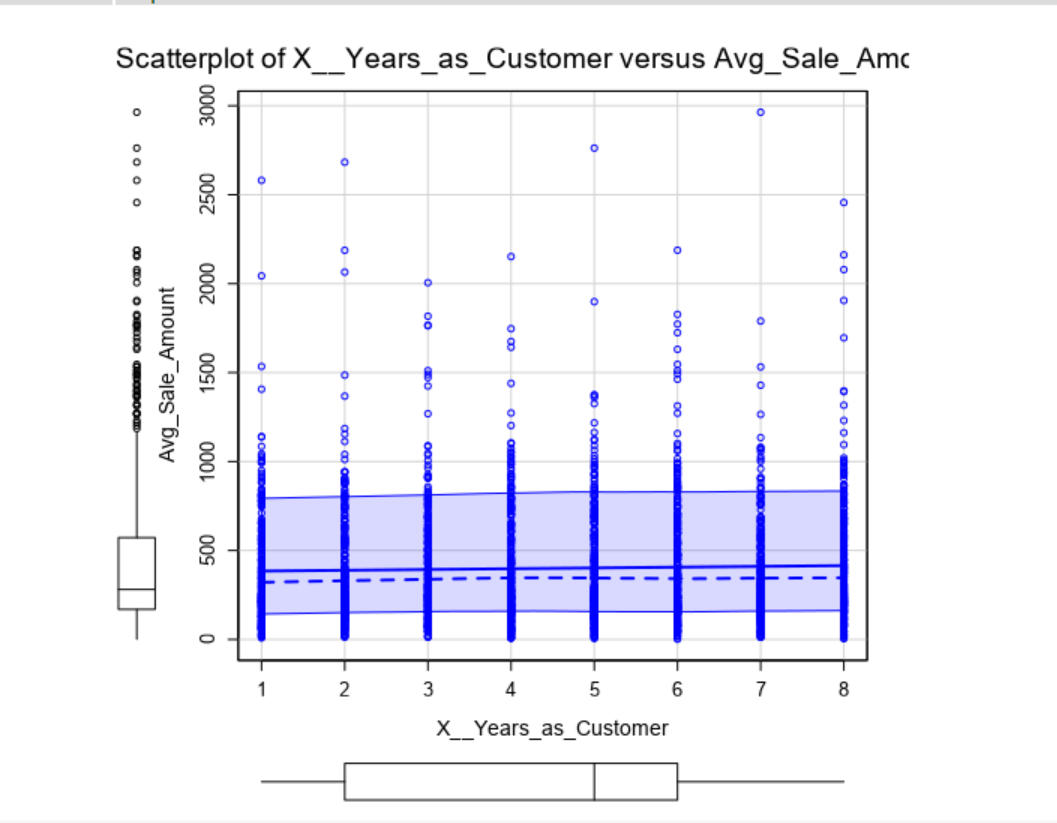
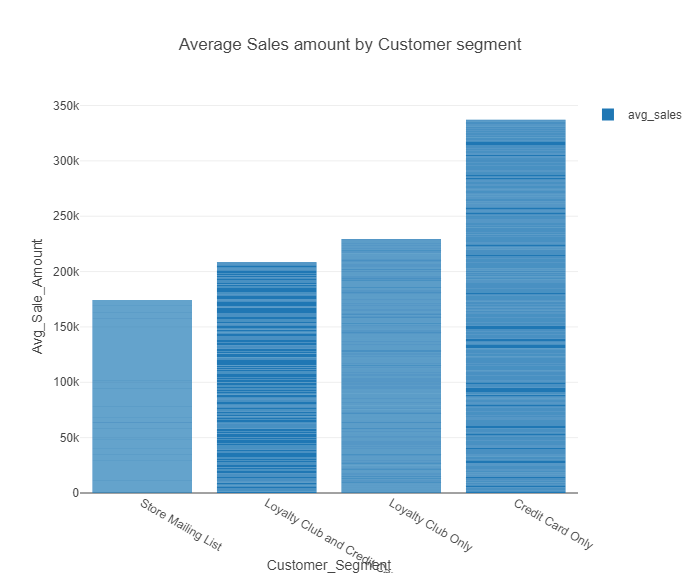


Fig.5. Graph between Average Sales Amount and Years as Customers



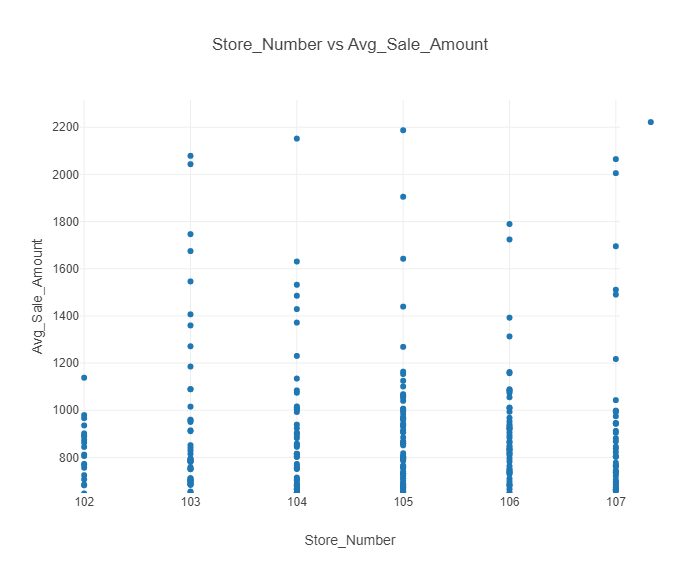
Fig.6. Graph between Average Sale Amount and Customer Segment

Fig.7. Graph between Average Sales Amount and Store Number

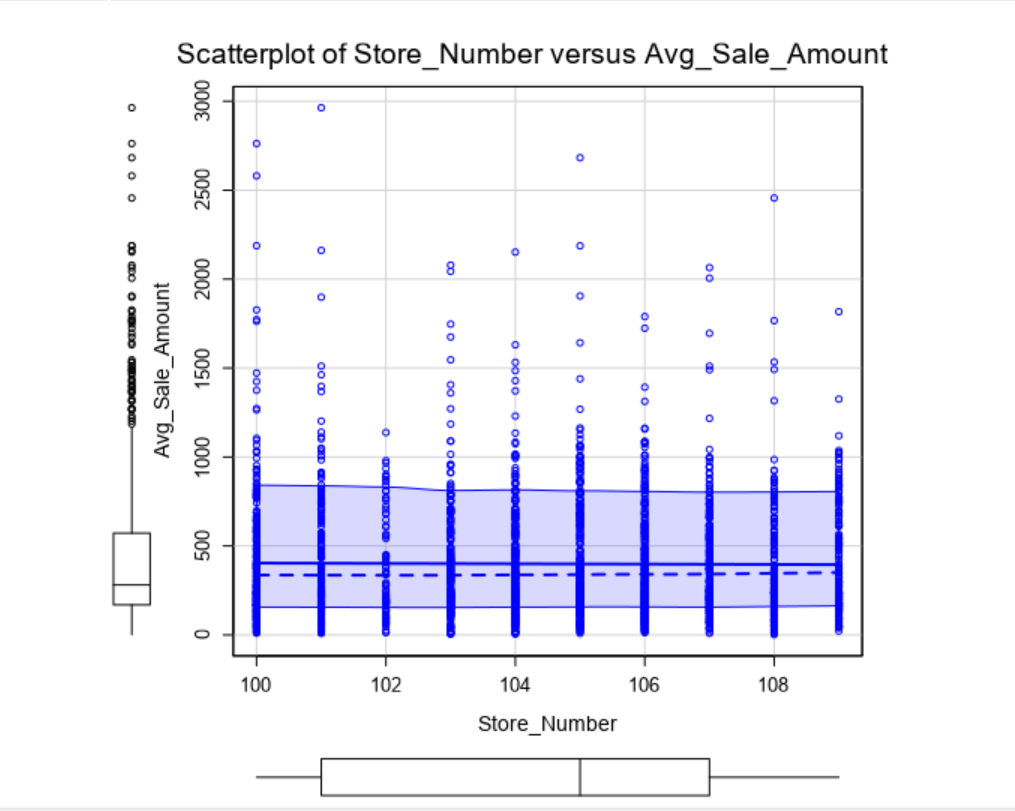


Fig.8. Graph between Average Sales Amount and Store Number

The correlations between the predictor variables and the target variables were depicted in the graphs above. Since Customer segment is a categorical variable, I used a bar chart to visualize the contributions, and Figure 1 demonstrates a very significant connection between Customer segment and the goal variable, which is the average number of sales.

1. **Why linear Regression model is a good model for this business problem?**

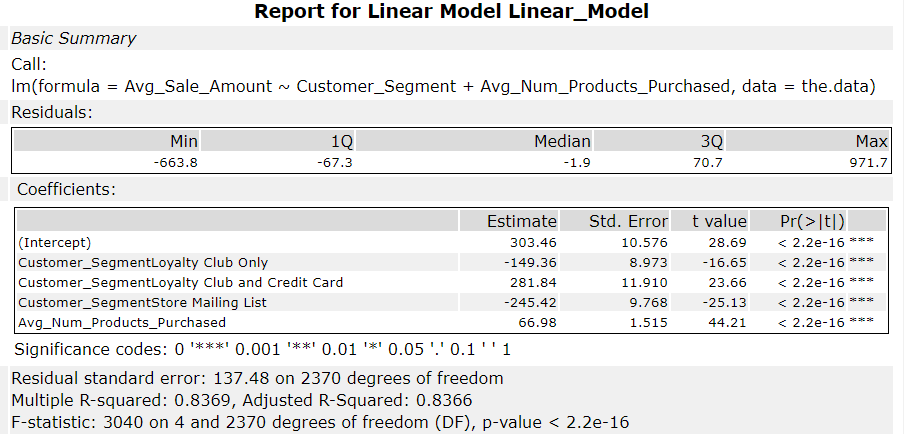
The report that results from running the model is shown below. I'll utilize this report to justify the merits of my linear model.

Fig.9. Report for Linear Regression Model

Multiple linear regression has an R-Squared value of "0.8369," whereas the adjusted R-Squared was "0.8366." The decline demonstrates the presence of additional variables with lower predictive value. The P-value being much less than 0.05 confirmed that the values from my model are good. The P-value indicates that the model has a higher than 95% confidence level that the predictors and the target variable are related.

Based on the facts at hand, what linear regression equation is the best? There shouldn't be more than 2 digits following the decimal for each coefficient (ex: 1.28)

Based on the information available, the optimal linear regression equation is shown below.

Avg\_Sales\_Amount = 303.46 + 66.98 \* (Avg\_Num\_Products\_Purchased) - 149.36 \* (Customer\_Segment: Loyalty Club Only) + 281.84 \* (Customer Segment: Loyalty Club and Credit Card) - 245.52 \* (Customer\_Segment: Store Mailing List) + 0 \* (Customer\_Segment: Credict Card Only)

**Step 3: Presentation/Visualization**

**The view of the Alteryx workflow**

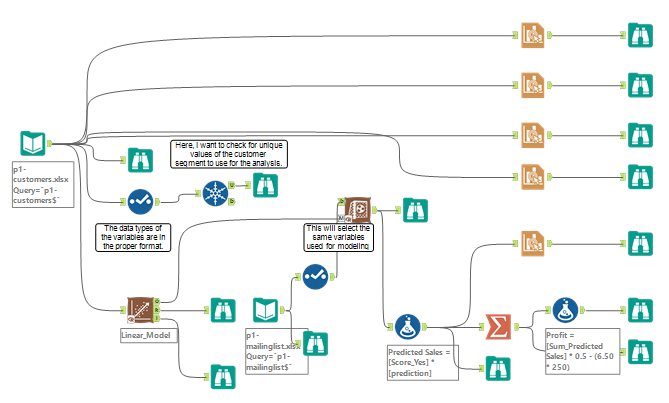


Fig.10. Alteryx Workflow\_1

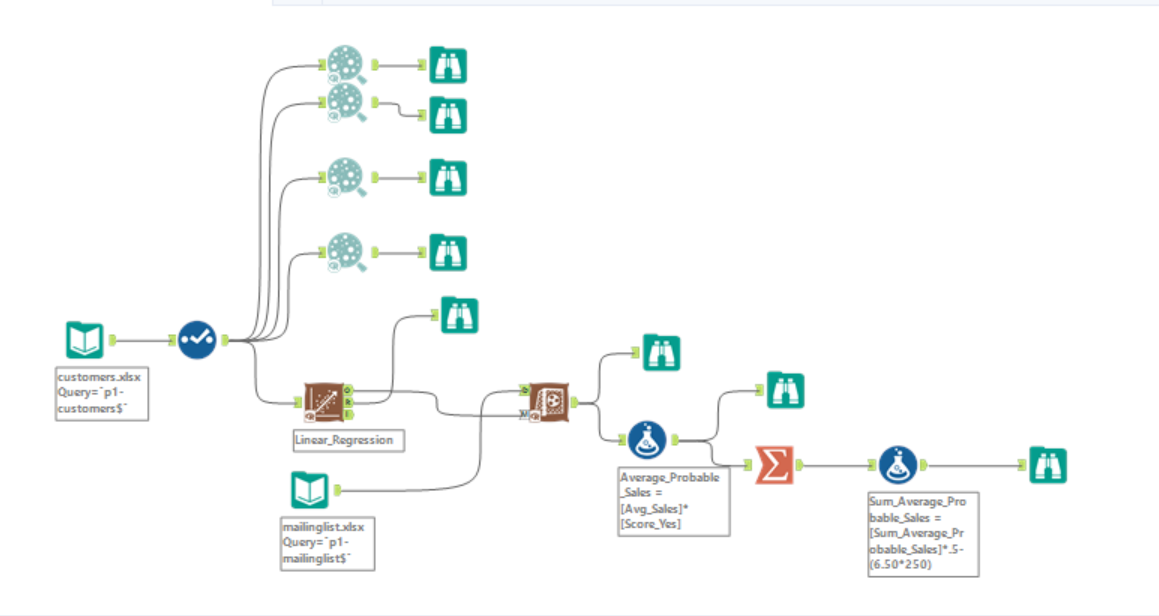


Fig.11. Alteryx Workflow\_2

1. **Output**

**Recommendation**

I advise the business to distribute the catalogue to these 250 clients since they will generate more revenue than anticipated. The outcome of the predictive analysis is as follows.

**Explanation:-**

According to my research, the company will make roughly $21,000 after accounting for the likelihood that these 250 clients will respond to the catalogue that was delivered to them. The company had anticipated making a profit contribution of $10,000. This profit contribution was calculated by deducting $6.50 from each catalogue given to these consumers, or 50% of the total revenue.

Profits anticipated with the new catalogue (assuming the catalogue is sent to these 250 customers)

Here is how to calculate estimated profit:

Profit = [Sum\_Predicted Sales] \* 0.5 - (cost of each catalog \* 250)

= $21, 987.44

Assuming the catalog was sent to these customers, a profit contribution of $21, 987.44 is expected from them.

1. **Conclusion**

Conclude with the 3 W’s (What Went Well, What Did NOT go Well, What Would you do Differently Next Time):

**What went well:**

* The linear regression model from Alteryx was used to estimate the anticipated revenue from the 250 new clients.
* This model is a powerful statistical tool that can be used to predict a continuous outcome variable based on one or more predictor variables.
* In this case, the predictor variables could include the number of catalogues distributed, the average gross margin, and the printing and distribution expenses.
* The model was able to take into account the printing and distribution expenses and the average gross margin of the items sold through the catalogue.
* These variables are important factors that can impact the overall revenue generated from distributing the catalogues, and the model was able to incorporate them into its predictions.

**What did not go well:**

* It was not mentioned in the problem statement whether the team had enough data for the model, If the team did not have enough data, it could have affected the accuracy of the model's predictions.
* It was not mentioned if the team had verified the assumptions of linear regression model which are linearity, independence of errors, homoscedasticity, and normality of errors, if these assumptions were not met, it could have affected the validity of the model's predictions.

**What would you do differently next time:**

* If more data is available and it is relevant, it could be used to improve the model's predictions.
* I would verify the assumptions of linear regression model before building the model and make sure that the assumptions are met.
* I would also perform cross validation to check the robustness of the model.
* I would also explore other models that may be more appropriate for this problem, such as decision trees or random forests.

1. **References**

**Data Sources:**

1. https://www.kaggle.com/datasets/shreydevil/productprdictionusingcustomermail
2. *Datameer 4.0 enables instant visualization in big data analytics*, April 2014, [online] Available: http://www.techrepublic.com/article/datameer-4-0-enables-instant-visualization-in-big-data-analytics/.
3. *Alteryx Designer*, [online] Available: http://www.alteryx.com/products/alteryx-designer.
4. *Alteryx*, [online] Available: http://in.teradata.com/partners/Alteryx/?LangType=16393&LangSelect-true.
5. *Alteryx: The Definitive Guide to Data Blending*, [online] Available: http://pages.alteryx.com/rs/alteryx/images/ALT\_WPDefGuideDataBlending-WithGraphics38.pdf.
6. "Alteryx Analytics Gallery", *Data Sheet*, [online] Available: http://www.alteryx.com/sites/default/files/downloads/alt-gallery-ds-v3a.pdf.
7. U. Chandrasekhar, A. Reddy and R. Rath, "A Comparative Study of Enterprise and Open Source Big Data Analytical Tools", Proceedings of IEEE Conference on Information and Communication Technologies, pp. 1-3, April 11–12, 2013.