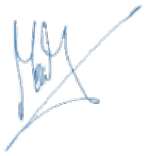
EUROPEAN INTERNATIONAL UNIVERSITY

ASIC ACCREDITED

COVER PAGE AND DECLARATION

	Master of Science (MS) in IT & Data Science
Specialisation:	
Module Code & Module Title:	MSITDS460 Data Analytics and Visualization
Students' Full Name:	Mony Ho
Word Count:	2820
Date of Submission:	25 July 2024

I confirm that this assignment is my own work, is not copied from any other person's work (published/unpublished), and has not been previously submitted for assessment elsewhere.

E-SIGNATURE: 

DATE: 25 July 2024

EIU Paris City Campus
Address: 59 Rue Lamarck, 75018 Paris, France | **Tel:** +33188320435 | **Mobile/WhatsApp:** +33607591197 | **Email:** paris@eiu.ac

EIU Operations HQ
Address: 16th Fl. Amarin Tower, 496-502, Room No.2.1 Ploenchit Rd., Bangkok 10330, Thailand | **Tel:** +66(2)256923 & +66(2)2569908 | **Mobile/WhatsApp:** +33607591197 | **Email:** info@eiu.ac

Table of Content

Introduction ---------- 1

Main Body ---------- 1

Data Cleaning and Preparation ---------- 1

Exploratory Data Analysis ---------- 7

Data Analytics ---------- 16

Conclusion ---------- 18

References ---------- 19

Introduction

According to Hashemi-Pour et al. (2024), “Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets. The term is often used interchangeably with information graphics, information visualization and statistical graphics.” Atlan (2023) stated about the benefits of data visualization that “Data visualization is no longer a nice to have but a must-have. Here are its biggest benefits of data visualization: simplifies complex data, reveals patterns and trends, aids in decision making, improves retention and engagement, increases accessibility, real-time monitoring, identify areas that need attention or improvement, predictive analysis, enhances storytelling, increases productivity, and risk management.”

Data visualization is the one of the easiest ways to catch attention and make the best communication to deliver information from a presenter to audiences. Charts, graphs, maps, or dashboard can describe data into information with thousands meaning that make audiences are easy to understand more than thousand paragraph of texts that make the audiences hard to understand (GeeksforGeeks, 2024). There are many popular tools and languages such as Excel, Tableau, Power BI, R, or Python to analyze and visualize data, and may also data modeling and prediction. Microsoft Power BI is one of several data analysis and visualization application tools will be used in this report to analyze a real-world dataset using data analytics and visualization techniques, prepare a report and presentation summarizing the findings. Datasets are from a company namely, **PTC WORKSHOPS** locates at Kampot province, Cambodia, will be used for data clean and preparation, exploratory data analysis, data analytics, and data visualization.

Main Body

Data Cleaning and Preparation

There are many tables in the company’s dataset but for working with data analysis and visualization, some tables need to be used only and extract from database of the company’s management system include **tableInvoice.csv**, **tableStaffs.csv**, **tableProducts.csv**, and

tableInvoiceDetails.csv. And **tableInvoice.csv** is the most important dataset that need to use for data analysis and visualization because the owner of the company needs to understand current situation of the company's income and predict the future of company's income.

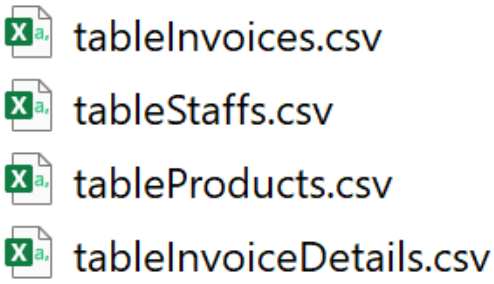
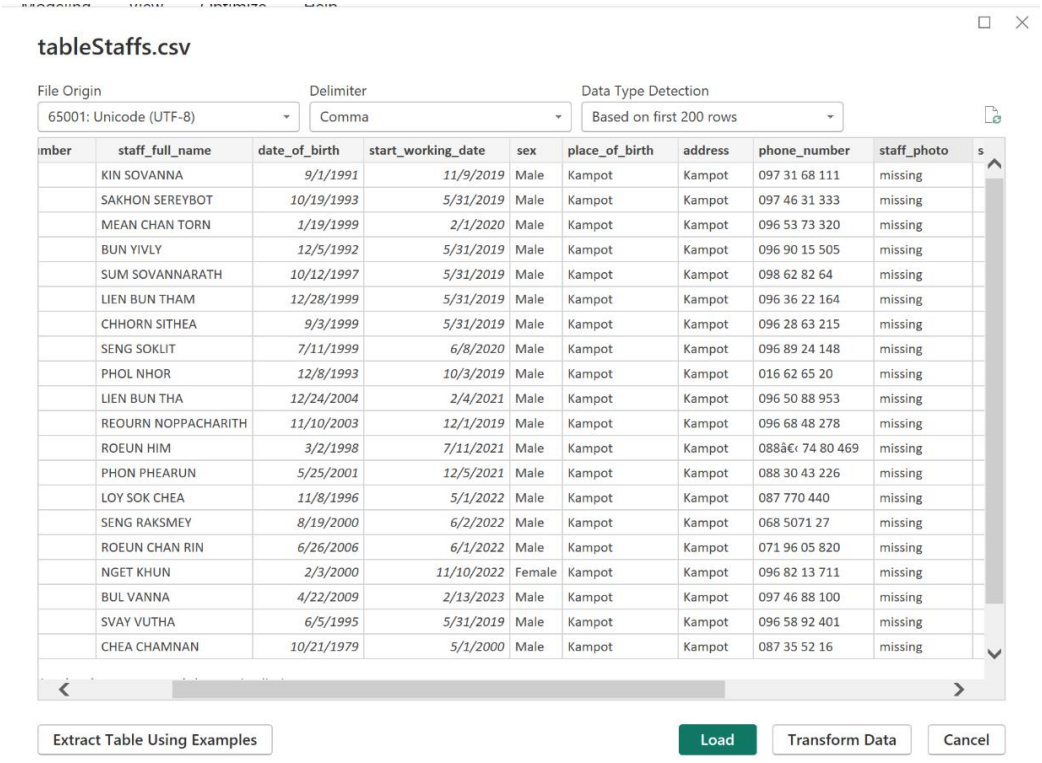


Figure 1. List of tables for data analysis and visualization

Handling of Missing Data

After importing a dataset namely, **tableStaff.csv**, there is a column namely, **staff_photo** does not store any single value for the whole column.



mber	staff_full_name	date_of_birth	start_working_date	sex	place_of_birth	address	phone_number	staff_photo	s
	KIN SOVANNA	9/1/1991	11/9/2019	Male	Kampot	Kampot	097 31 68 111	missing	
	SAKHON SEREYBOT	10/19/1993	5/31/2019	Male	Kampot	Kampot	097 46 31 333	missing	
	MEAN CHAN TORN	1/19/1999	2/1/2020	Male	Kampot	Kampot	096 53 73 320	missing	
	BUN YIVLY	12/5/1992	5/31/2019	Male	Kampot	Kampot	096 90 15 505	missing	
	SUM SOVANNARATH	10/12/1997	5/31/2019	Male	Kampot	Kampot	098 62 82 64	missing	
	LIEN BUN THAM	12/28/1999	5/31/2019	Male	Kampot	Kampot	096 36 22 164	missing	
	CHHORN SITHEA	9/3/1999	5/31/2019	Male	Kampot	Kampot	096 28 63 215	missing	
	SENG SOKLIT	7/11/1999	6/8/2020	Male	Kampot	Kampot	096 89 24 148	missing	
	PHOL NHOR	12/8/1993	10/3/2019	Male	Kampot	Kampot	016 62 65 20	missing	
	LIEN BUN THA	12/24/2004	2/4/2021	Male	Kampot	Kampot	096 50 88 953	missing	
	REOURN NOPPACHARITH	11/10/2003	12/1/2019	Male	Kampot	Kampot	096 68 48 278	missing	
	ROEUN HIM	3/2/1998	7/11/2021	Male	Kampot	Kampot	088â€‹ 74 80 469	missing	
	PHON PHEARUN	5/25/2001	12/5/2021	Male	Kampot	Kampot	088 30 43 226	missing	
	LOY SOK CHEA	11/8/1996	5/1/2022	Male	Kampot	Kampot	087 770 440	missing	
	SENG RAKSMEY	8/19/2000	6/2/2022	Male	Kampot	Kampot	068 5071 27	missing	
	ROEUN CHAN RIN	6/26/2006	6/1/2022	Male	Kampot	Kampot	071 96 05 820	missing	
	NGET KHUN	2/3/2000	11/10/2022	Female	Kampot	Kampot	096 82 13 711	missing	
	BUL VANNA	4/22/2009	2/13/2023	Male	Kampot	Kampot	097 46 88 100	missing	
	SVAY VUTHA	6/5/1995	5/31/2019	Male	Kampot	Kampot	096 58 92 401	missing	
	CHEA CHAMNAN	10/21/1979	5/1/2000	Male	Kampot	Kampot	087 35 52 16	missing	

Figure 2. Import tableStaff.csv to Power BI

Because of this column is optional, it is removed from the table by using Power Query Editor in Power BI.

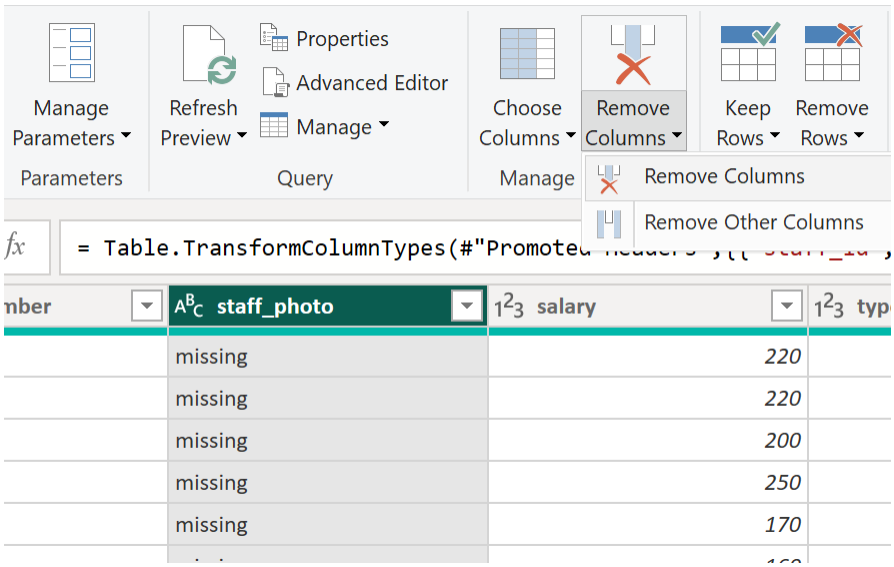


Figure 3. Remove **staff_photo** column from **tableStaff.csv** in Power BI

There is also a column namely, **product_image** in **tableProduct.csv** dataset that need to remove because it has no any value.

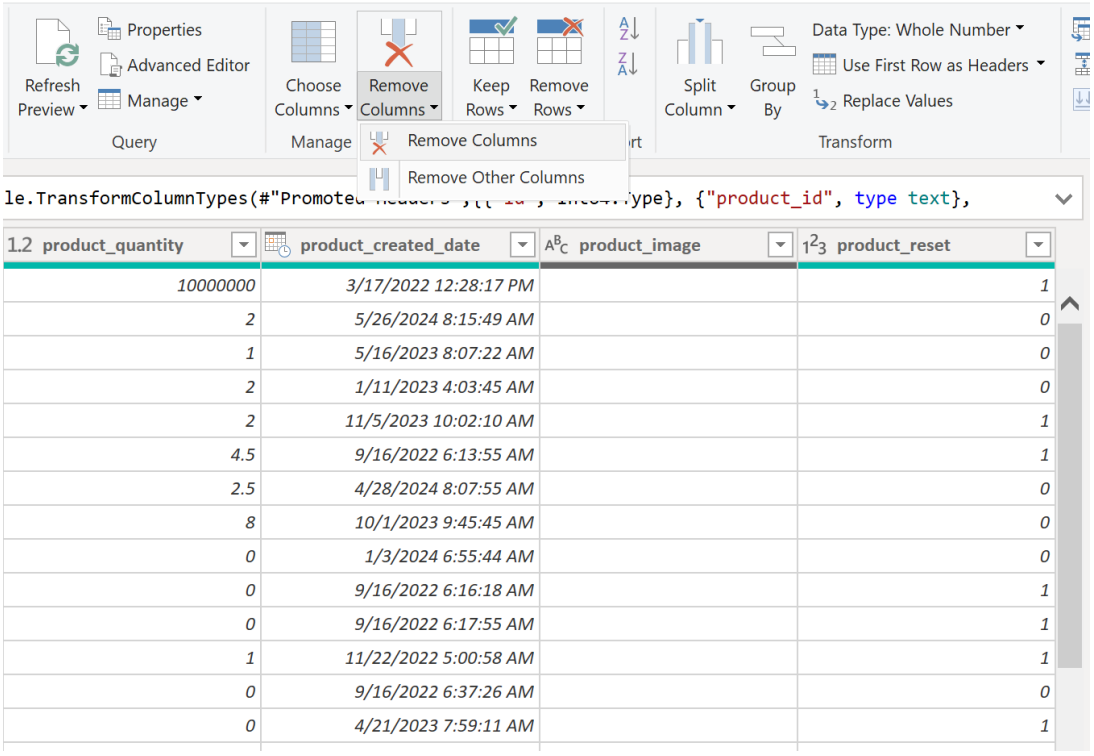
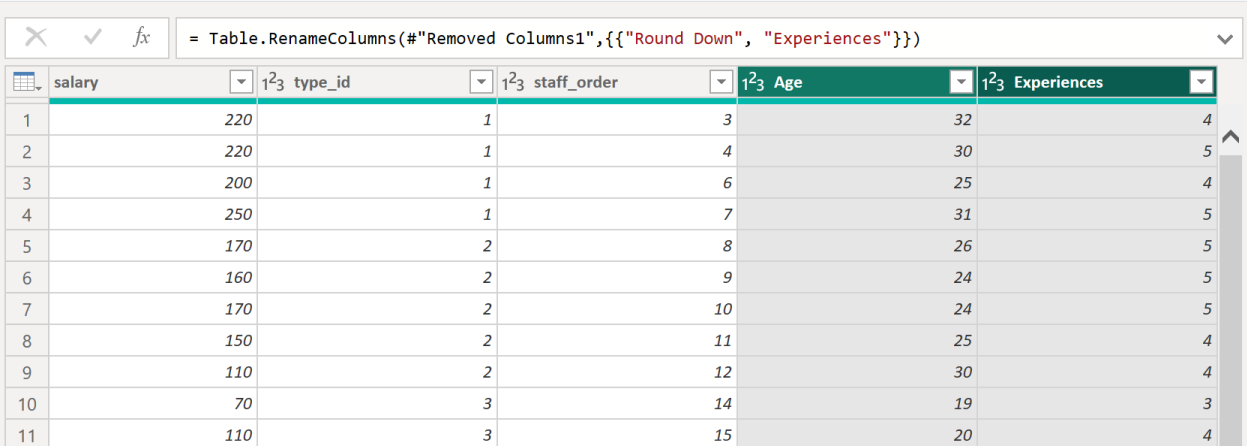


Figure 4. Remove **product_image** column from **tableStaff.csv** in Power BI

Data Transformation

Age and **Experiences** columns have been created in **tableStaffs.csv** follow the request of the owner of the company to understand staff's age based on their birth date and experiences based on their start working date (Rad, 2020).



	salary	type_id	staff_order	Age	Experiences
1	220	1	3	32	4
2	220	1	4	30	5
3	200	1	6	25	4
4	250	1	7	31	5
5	170	2	8	26	5
6	160	2	9	24	5
7	170	2	10	24	5
8	150	2	11	25	4
9	110	2	12	30	4
10	70	3	14	19	3
11	110	3	15	20	4

Figure 5. Create **Age** and **Experiences** columns in Power BI

The four tables are connected relationship by using one-to-many relationship. For symbol "1" refers to primary table. On the other hand, for symbol "*" (asterisk symbol) refers to related table.

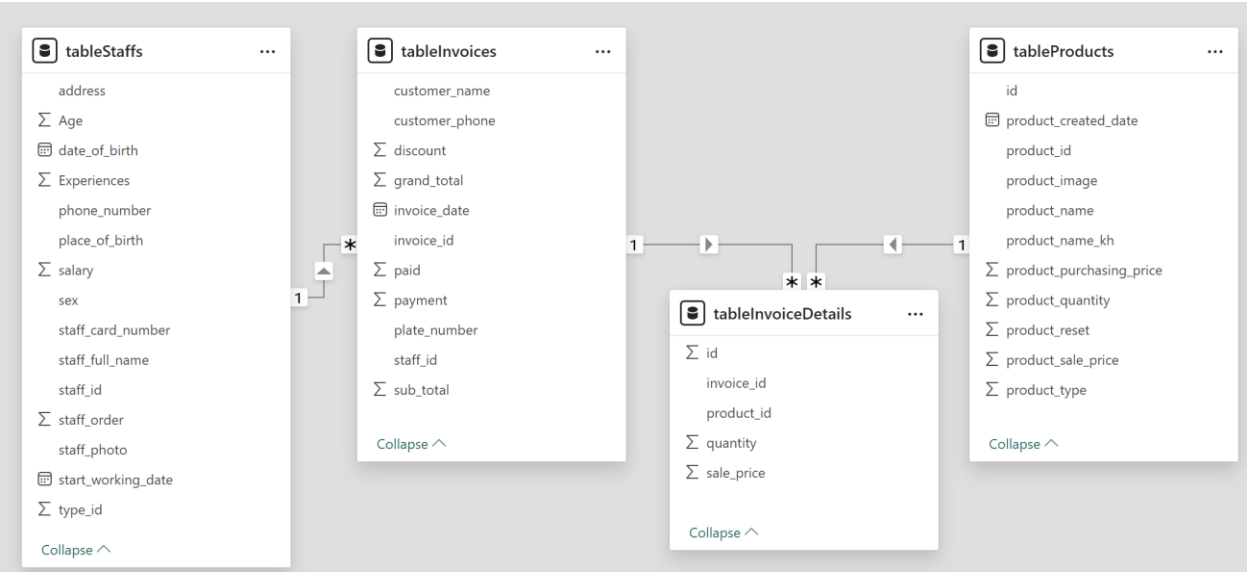


Figure 6. Relationship between tables in Model View of Power BI

Data Normalization

After calculating in **tableInvoice.csv** dataset to find the result of minimum and maximum value. The result found that minimum value is 0.00 USD but maximum value is 3.21 K USD. This result makes a wide range of the dataset and can make the prediction produce incorrect result.



Figure 7. Minimum and maximum value of **tableInvoice.csv**

Percentile is one of many ways for data normalization by cutting data in 100 pieces. In Power BI, it has a function called PERCENTILE.EXC to work with percentile. DAX function below is used with **grand_total** column in **tableInvoice.csv** in Power BI to find where is the 5% lowest value and 95% highest value (How to Find Percentile Using Power Bi DAX -, 2023).

Low05 = PERCENTILE.EXC([grand_total],0.05)

High95 = PERCENTILE.EXC([grand_total],0.95)

After completing finding, the result found that the lowest 5% of **grand_total** column is 5 USD and the highest column of 95% of **grand_total** column is 218 USD. This way help reduce the outlier data from the dataset and make the analysis is more accuracy than the original data.

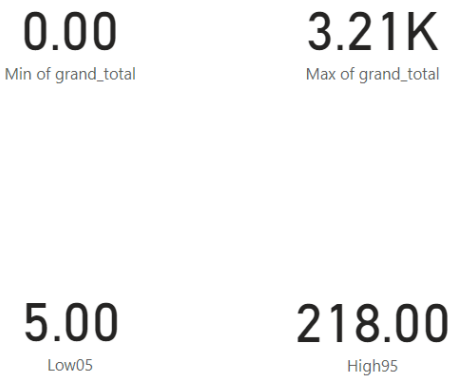


Figure 8. The result of 5% and 95% percentile

After finding the percentile, the **Number filter** menu item can be used to filter data of **grand_total** column by using **Between item**.

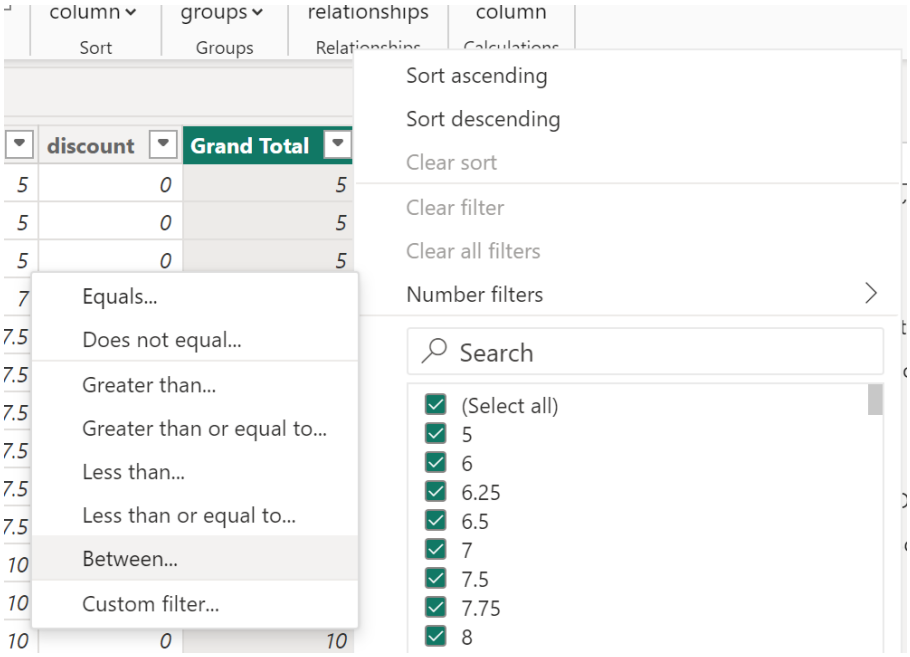


Figure 9. Using filter in Power BI

In **Number filters** windows, first 5% percentile is used with '**is greater than or equal to**' condition. To filter from low value to high value, '**And**' logical operator is used to compare with second condition is '**is less than or equal to'**.

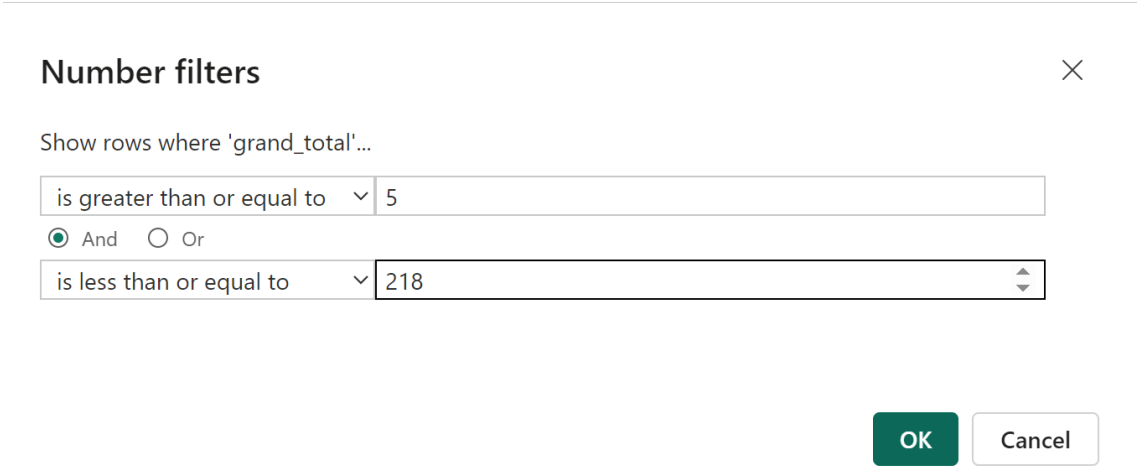
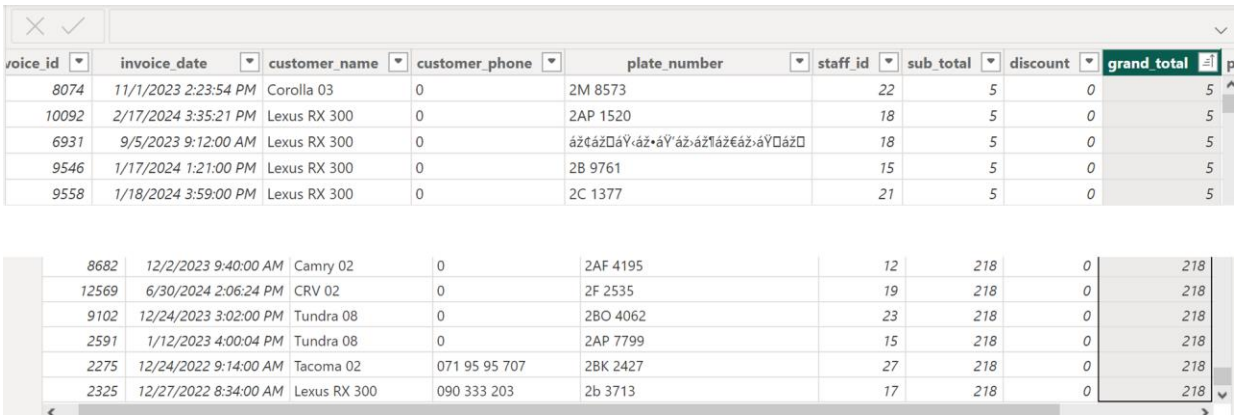


Figure 10. Filter **grand_total** column with percentile in Power BI

Next, **grand_total** can be sorted to verify the result of percentile and filtering. The first row of **grand_total** is 5 and the last row of **grand_total** is 218, match to the percentile result finding.



voice_id	invoice_date	customer_name	customer_phone	plate_number	staff_id	sub_total	discount	grand_total
8074	11/1/2023 2:23:54 PM	Corolla 03	0	2M 8573	22	5	0	5
10092	2/17/2024 3:35:21 PM	Lexus RX 300	0	2AP 1520	18	5	0	5
6931	9/5/2023 9:12:00 AM	Lexus RX 300	0	áž¢áž□áŸ‹áž•áŸ'áž›áž¶áž€áž›áŸ□áž□	18	5	0	5
9546	1/17/2024 1:21:00 PM	Lexus RX 300	0	2B 9761	15	5	0	5
9558	1/18/2024 3:59:00 PM	Lexus RX 300	0	2C 1377	21	5	0	5
8682	12/2/2023 9:40:00 AM	Camry 02	0	2AF 4195	12	218	0	218
12569	6/30/2024 2:06:24 PM	CRV 02	0	2F 2535	19	218	0	218
9102	12/24/2023 3:02:00 PM	Tundra 08	0	2BO 4062	23	218	0	218
2591	1/12/2023 4:00:04 PM	Tundra 08	0	2AP 7799	15	218	0	218
2275	12/24/2022 9:14:00 AM	Tacoma 02	071 95 95 707	2BK 2427	27	218	0	218
2325	12/27/2022 8:34:00 AM	Lexus RX 300	090 333 203	2b 3713	17	218	0	218

Figure 11. Data in **grand_total** column after percentile filtering.

After complting filtering, clicking on **Close & Apply** button is required to save everything, especially saved the filtering data from value 5 to 218 in **grand_total** column. After clicking **Close & Apply** button, the Power Query Editor is closed and redirect to Power BI Editor.

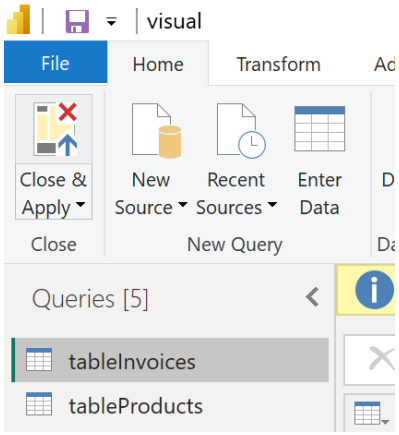


Figure 12. **Close & Apply** button in Power Query Editor

Exploratory Data Analysis

Descriptive Statistics and Visualization

Some of descriptive statistics are used in Power BI such as sum, mean, count, maximum, minimum, standard deviation, variance, and median to summarize data by displaying in **card** on

the dashboard of report view in Power BI. A **slicer** is also used to filter all value in **card** based on years. This is the result of **Page 1** on the dashboard in Report view.

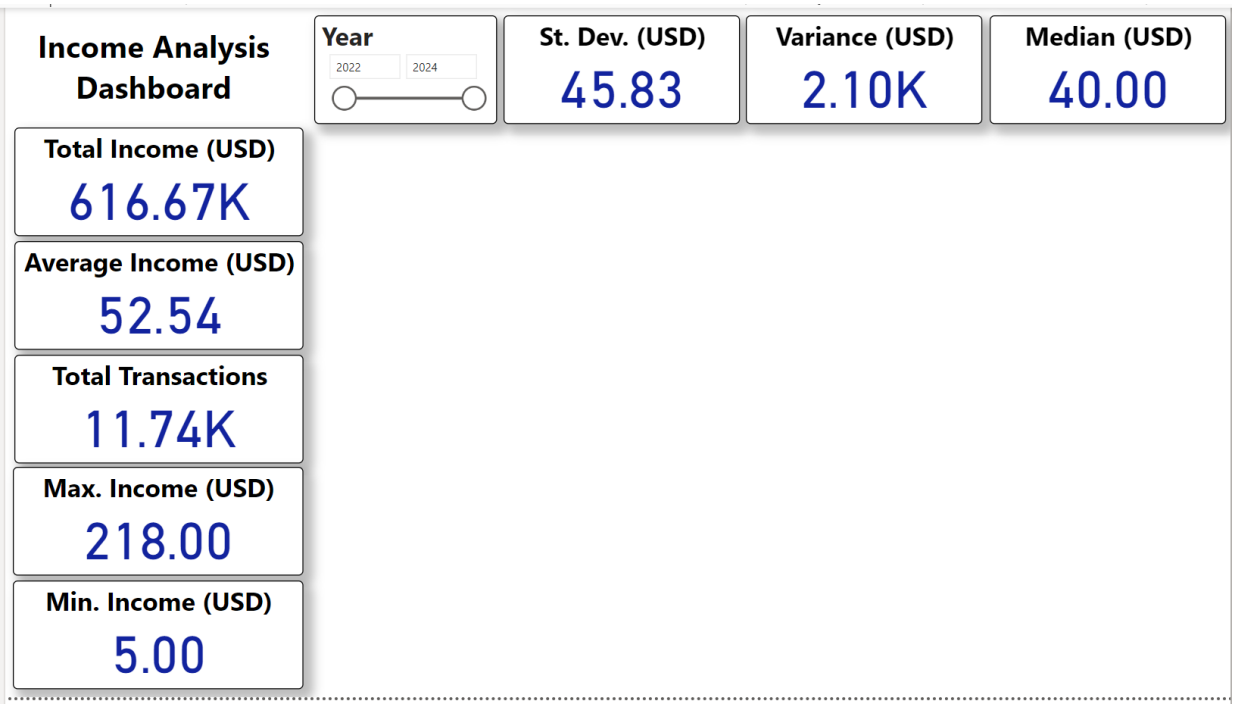


Figure 13. Income Analysis Dashboard in Power BI

Summarize the Data

Sum: In the Total Income (USD) card displays the value of summary of **grand_total** column of **tableInvoice.csv**.

Mean: In the Average Income (USD) card displays the value of mean of **grand_total** column of **tableInvoice.csv**.

Count: In the Total Transactions card displays the value of count of **grand_total** column of **tableInvoice.csv**.

Maximum: In the Max. Income card displays the value of maximum of **grand_total** column of **tableInvoice.csv**.

Minimum: In the Min. Income card displays the value of minimum of **grand_total** column of **tableInvoice.csv**.

Standard Deviation: In the St. Dev. (USD) card displays the value of standard deviation of **grand_total** column of **tableInvoice.csv**.

Variance In the Variance (USD) card displays the value of variance of **grand_total** column of **tableInvoice.csv**.

Median: In the St. Dev. (USD) card display the value of standard deviation of **grand_total** column of **tableInvoice.csv**.

Frequency Table: In **Matrix**, there is a way to displays frequency table group by **years** with total to **grand total**. And the result that found in matrix can be export to CSV file type for using later. In this case the frequency table is exported to **frequencyTable.csv**. After that, the **frequencyTable.csv** will be imported back to Power BI for using analysis later. This is the result of **Page 1** on the dashboard in Report view.

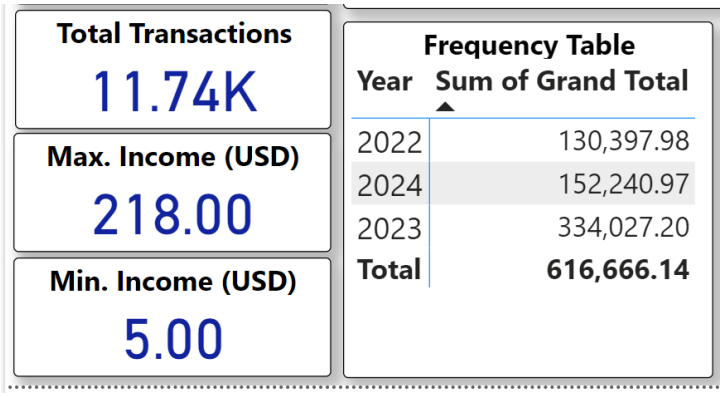


Figure 14. Frequency table in dashboard

Data Analytics

Selection of Appropriateness Methods

Regression analysis is a technique for predictive modeling that measures the correlation between an independent and dependent variable. It is used to predict how the dependent variable changes follow with its independent variable. This modeling is used like a tool to predict values in

numeric type based on a variety of inputs. There are many types of regression models such as Simple Regression, Multiple Regression, Linear Regression, Multiple Linear Regression, logistic Regression and so on. Linear regression is one of many regression types that used to create a hypothetical line that best connects all data points. It is the simplest kind of regression and it is used to draw a hypothetical line that most closely connects each data point.

Interpretation of results

Linear regression is applied in Power BI to predict the income of **tableInvoice.csv** dataset between year of 2022 to 2042 from the actual data of **tableInvoice.csv** if between 2022 to 2024 only. After completing the linear analysis, a dashboard will be created in Power BI to interpret the results by displaying frequency table of **total income** group by **years**. And the **Income Estimation card** will be calculated and displays the estimate income each year based on the **slicer** value (Data Analytics Central, 2023). The steps below describe of using linear regression formula with Power BI application to analyze and interpret the results. The result will be placed on **Page 2** of the dashboard in Report view.

Step 1: Import **frequencyTable.csv** to Power BI.

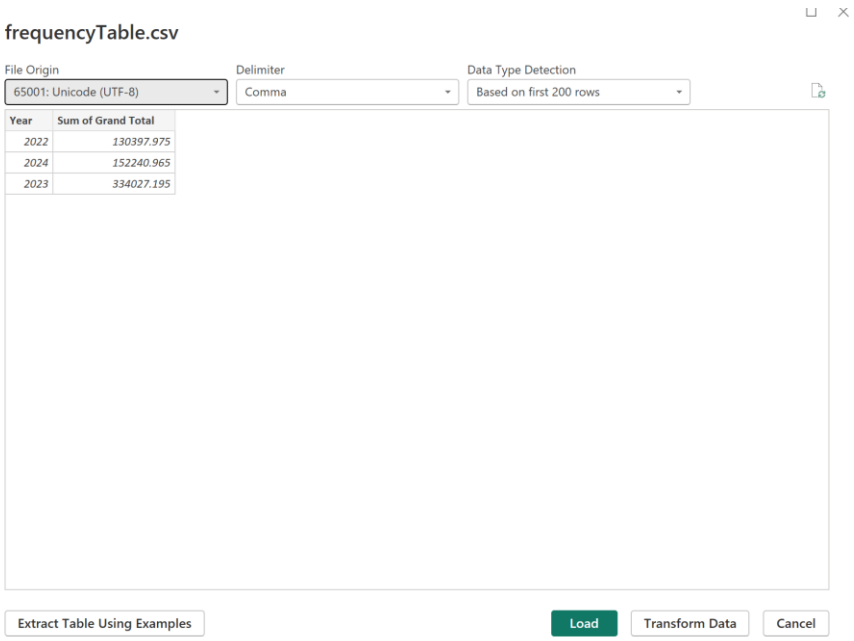


Figure 15. Import frequencyTable.csv to Power BI

Step 2: Create a **New table** from **Modeling Tab** and write the following calculation by using LINEST function of DAX functions in **fx textbox** and click on **green check icon** to commit.

LinearRegression = LINEST(frequencyTable[Sum of Grand Total],frequencyTable[Year])

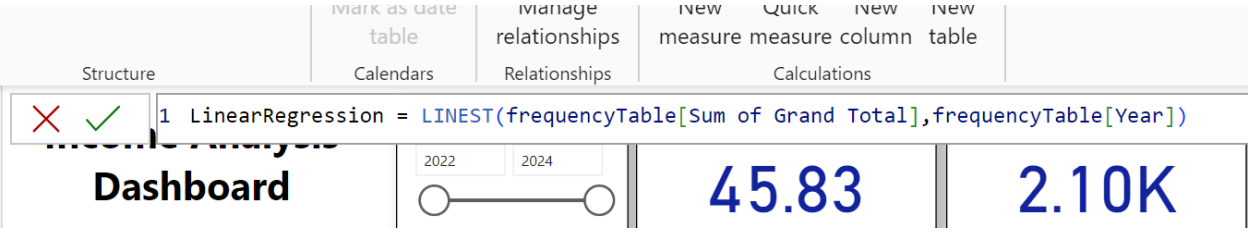


Figure 16. Using LINEST function to create **LinearRegression** table

Power BI created **LinearRegression** table with many columns as the following figure.

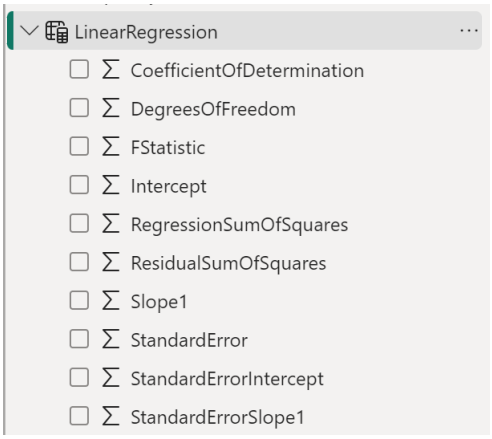


Figure 17. **LinearRegression** table after using LINEST function

Step 3: Add a **New measure** to **frequencyTable** and write the following calculation.

IncomeEST = sum(LinearRegression[Intercept]) + sum(LinearRegression[Slope1]) * sum(frequencyTable[Sum of Grand Total])

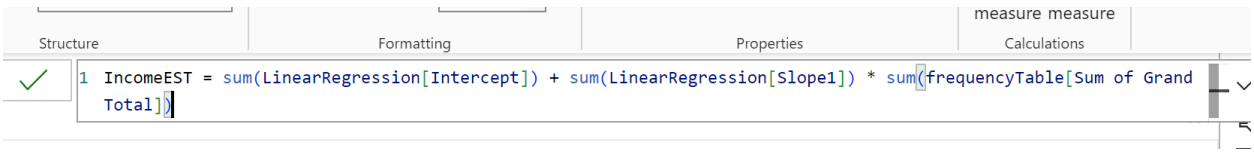
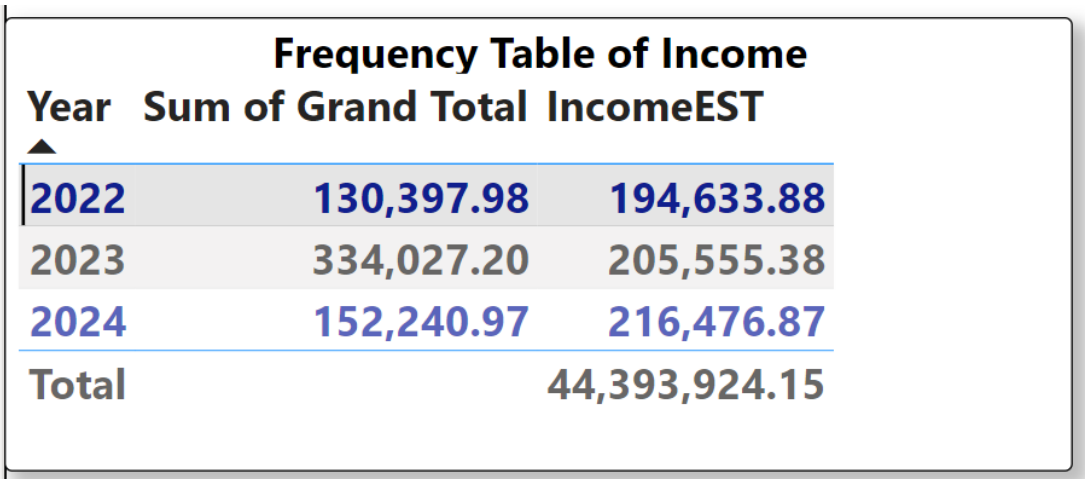


Figure 18. Calculate **IncomeEST** in **frequencyTable**

Step 4: Create a **table visual** by add 3 columns from **frequencyTable** such as **Year**, **Sum of Grand Total**, and **IncomeEST**.



Frequency Table of Income

Year	Sum of Grand Total	IncomeEST
2022	130,397.98	194,633.88
2023	334,027.20	205,555.38
2024	152,240.97	216,476.87
Total		44,393,924.15

Figure 19. Create frequency table on the dashboard

Step 5: Add **Numeric range** of **New Parameter** in **Modeling Tab**

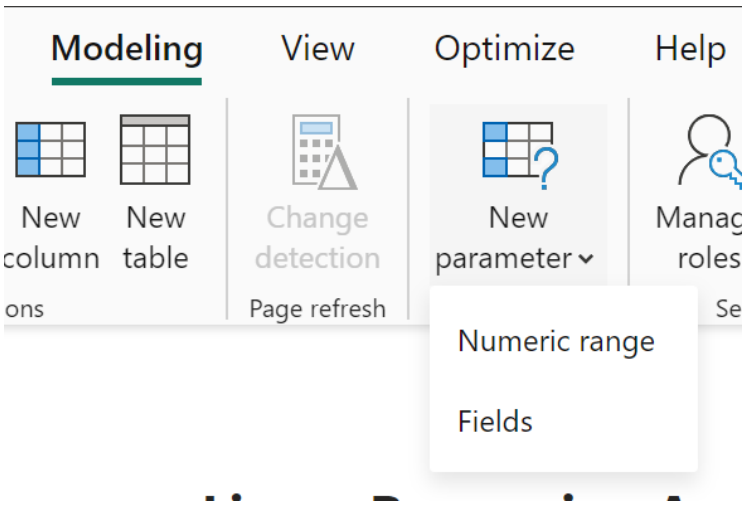


Figure 20. New parameter in Modeling tab

Step 6: Set the following values in their textbox to create a slicer with value of years from 2022 to 2042. This slicer is used to set or change the value of year then give to the formula to calculate for the estimation.

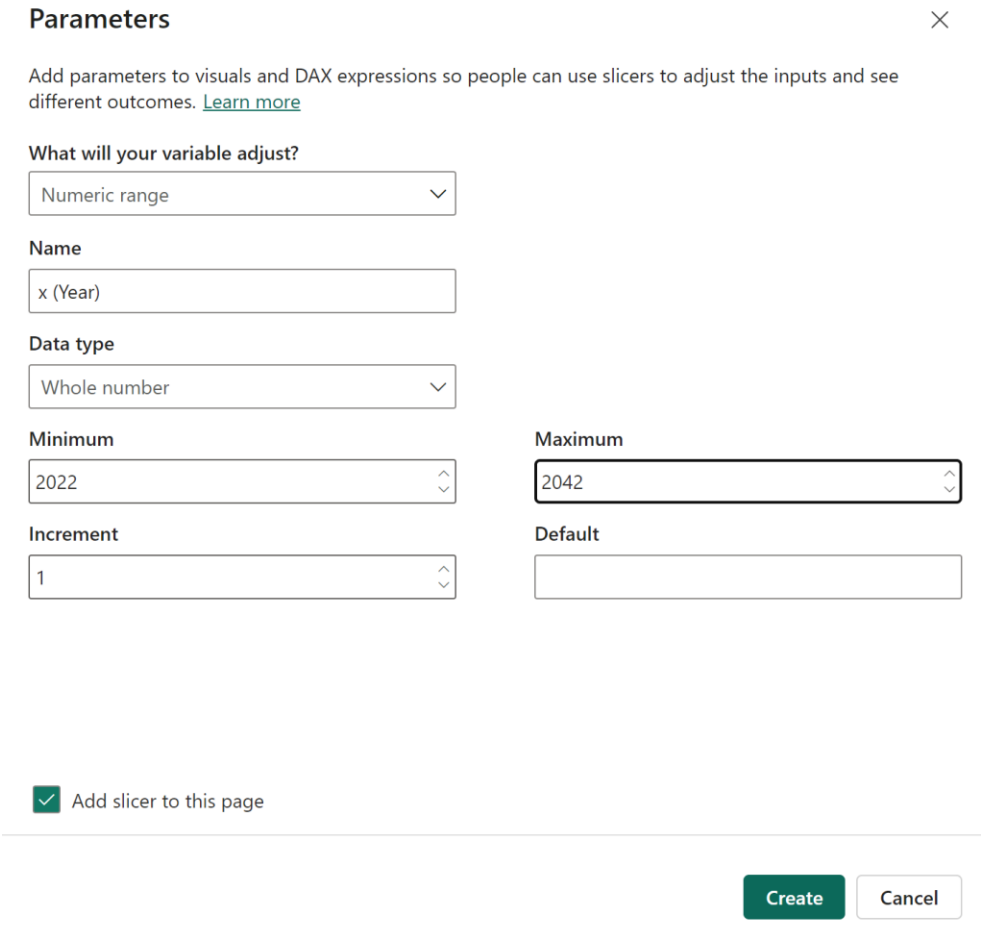


Figure 21. Parameter window

The **x (Year) table** is created with 2 columns such as **x (Year)** and **x (Year) Value**.

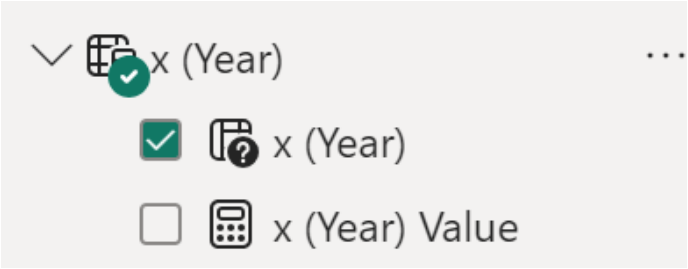


Figure 22. **x (Year) table** after using Numeric range parameter

Year Selected slicer and table are placed and designed in Page 2 of the dashboard. And the title of the **Page 2** is also set to **Linear Regression Analysis Dashboard**.

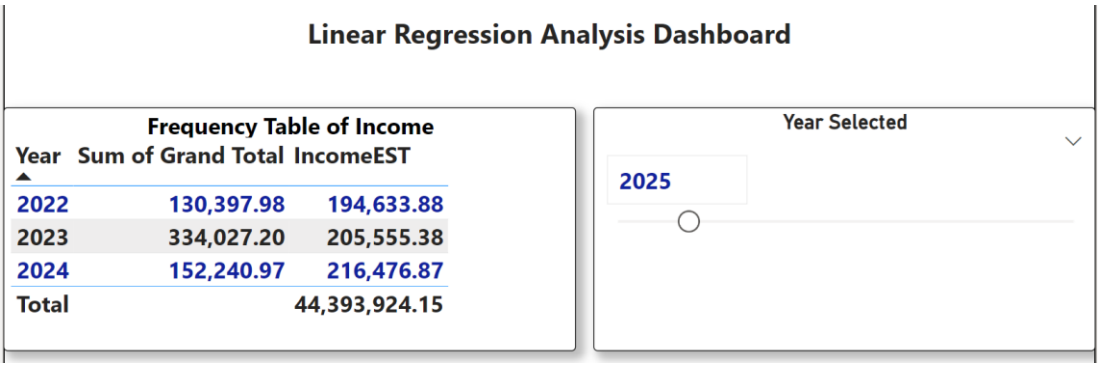


Figure 22. Slicer designed on the dashboard

Step 7: Create **New measure** in **x (Year)** table and write the following function.

YearSelected = SELECTEDVALUE('x (Year) '[x (Year)])



Figure 23. Create New measure for **YearSelected**

The result of this step for **x (Year)** s has a new column namely, **YearSelected** that it stores value of the years from 2022 to 2024.

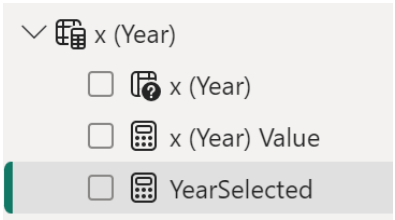


Figure 24. **YearSelected** column after created

Step 8: Add 4 cards such as **Income Estimation**, **Sum of Intercept**, **Sum of Slop 1** and **Year Selected**. The **Year Selected** card is from **YearSelected** column in **x (Year)** table. The **Sum of Slop 1** card is from **Slop1** column in **LinearRegression** table. The **Sum of Intercept** card is from **intercept** column in **LinearRegression** table. And The **Income Estimation** card uses the

following functions to generate **Income Estimation** result based on the calculation of **Sum of Intercept + Sum of Slop 1 * Year Selected**. The value of **Year Selected** card can be changed based on the slicer handle that can be dragged between value of year of 2022 to 2042.

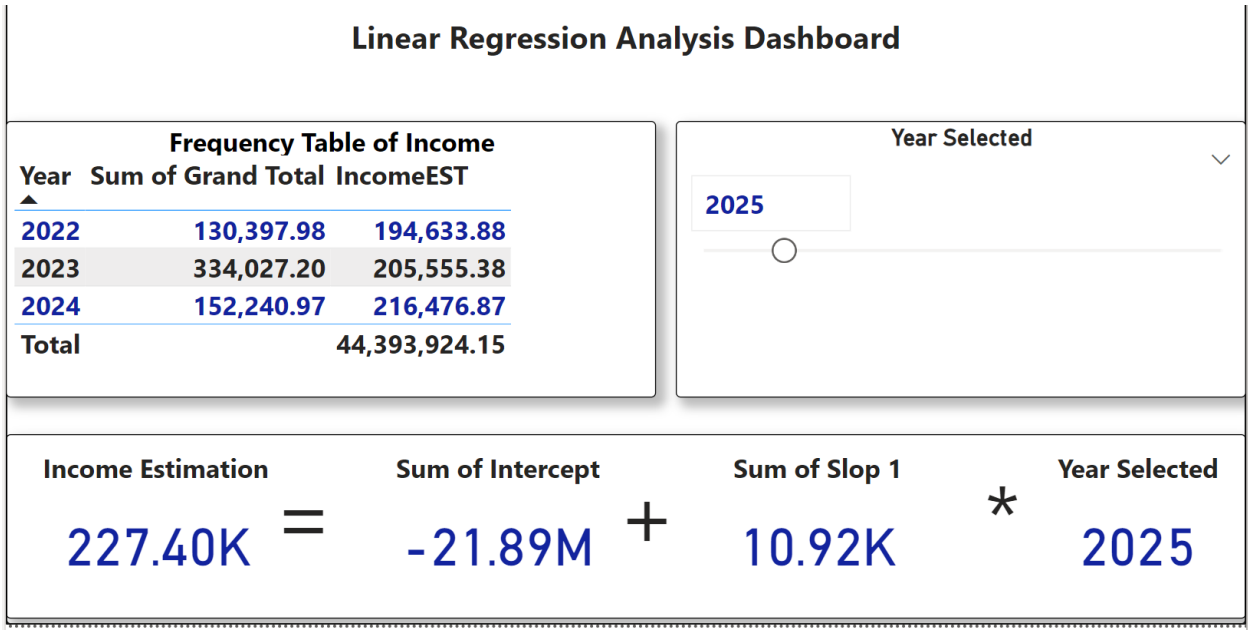


Figure 25. Income Estimation calculation

Step 9: The final result on **Page 2** of the dashboard design look like the following figure.

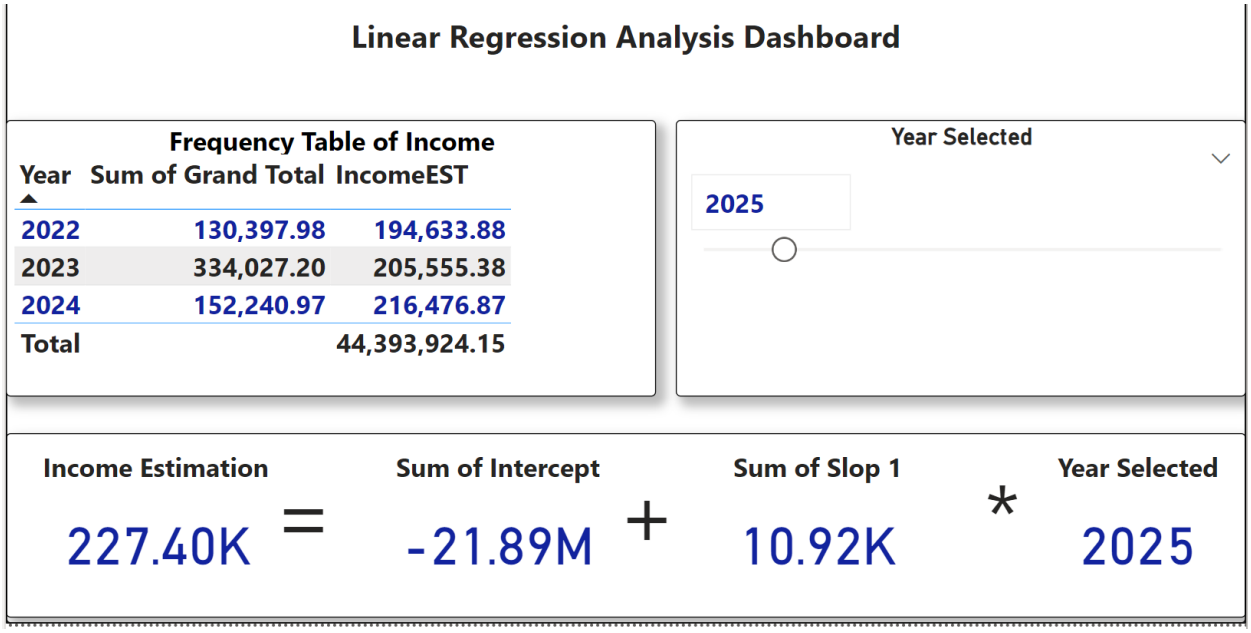


Figure 26. The final result of **Page 2** in the dashboard

Quality of Insights Obtained

From laco.be, there are 4 steps in strong data visualization such as step 1 is collecting high-quality data, step 2 is aligning data visualization with the end user, step 3 is getting the intention of data visualization right, and step 4 is selecting the right data visualization method. In summary, step 1 determines about checking the data sources or removing duplicate data. Step 2 prepares visualization to makes it looks creativity and skills. Step 3, the purpose of the visualization can be anything and make sure it is very clear. And step 4 uses the appropriate visual for the appropriate case (Jolijn, 2021).

Data Visualization

Selection of Appropriate Visualizations

Page 1 of the dashboard that is completed design in Report view is used to add 3 more charts include **Total Income By Years Pie Chart**, **Customer Numbers By Years Stacked Column Chart**, and **Scatter Plot**.

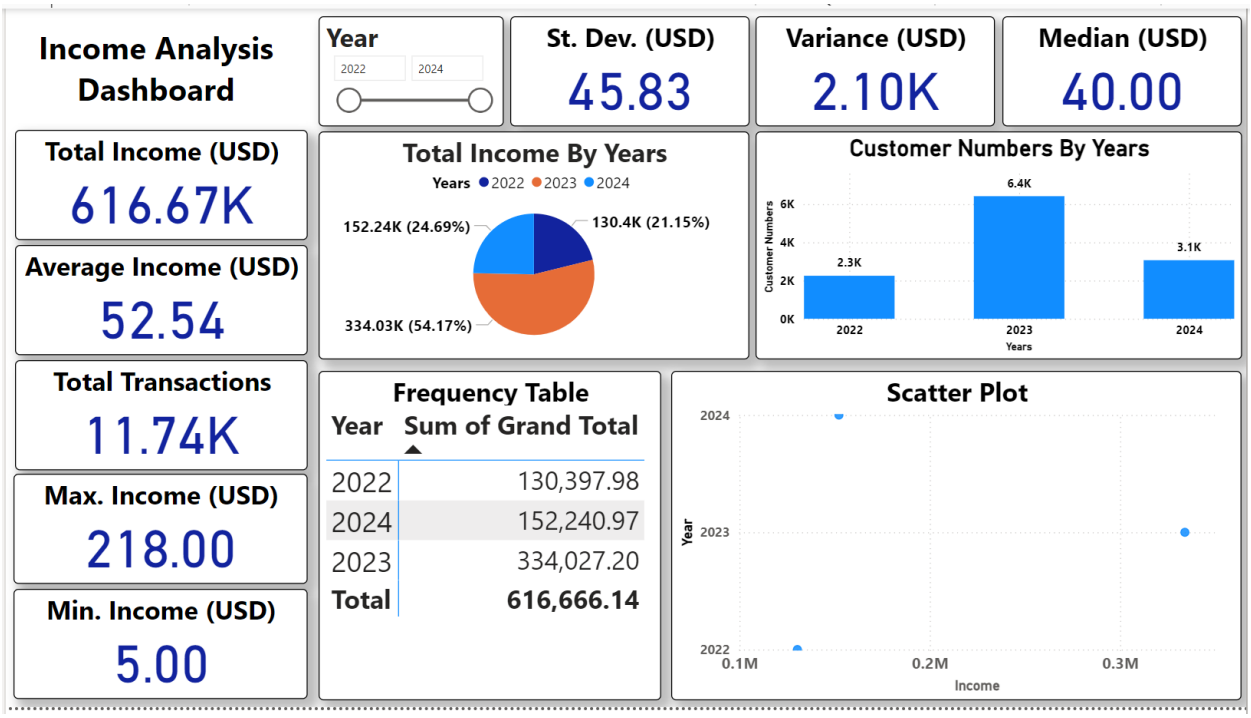


Figure 27. Completed dashboard design on **Page 1** of Report view in Power BI

Page 2 of the dashboard that is completed design in Report view has four sections such as **dashboard's title** is at the top, frequ**ency table of income** is at the left, **Year Selected slicer** is at the right, and linear regression estimation calculation is at the bottom. The actual data is between year of 2022 and 2024 only. But after using linear regression in Power BI, this dashboard can predict the **Income Estimation** from the year of 2022 to 2042.

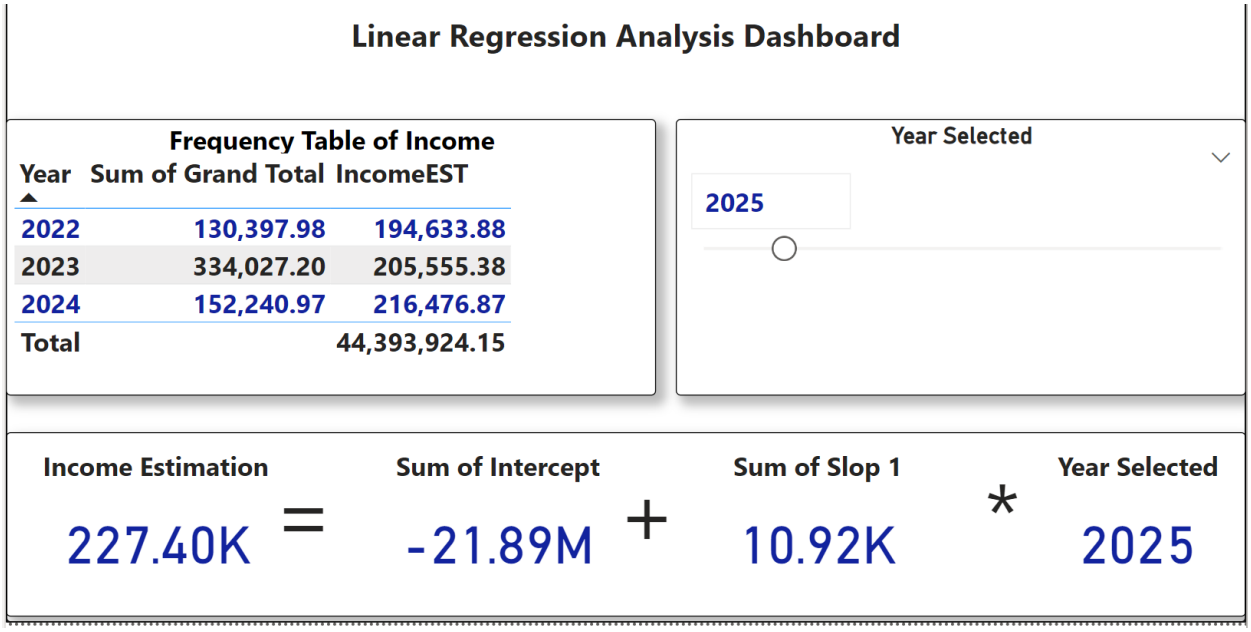


Figure 28. Completed dashboard design on **Page 2** of Report view in Power BI

Effective Communication of Insights and Best Practice in Data Visualization

The main objective of data visualization is to simplify large and complex information to make it easy to understand. There are 5 best practices of effective communication of insights such as understand the audience, choose the right visualization tool, clear the clutter, use less fonts, and use colors creatively (Quach & Quach, 2024). Understanding the audience is very important at the first step, stack holder's objective, dataset, and business problem that need to solved by data analysis and visualization. Power BI is one of the best visualization tools that used to visualize data in many types include bar chart, pie chart, scatter plot, and so on. All visualization charts and graphs on the dashboard must be simple and clean. Use a few font names and color to make eye-catching. In the **Page 1** and **Page 2** on the dashboard used only black and dark blue color on the write background. This technic makes the texts and numbers look very clear to see. Appreciating and carefully considering the technical components of data presentation is

necessary for good data visualization. However, there is also a creative component to it. Decisions are based on the story that wish to tell, and design choices are motivated by the desire to present that story to our audience as effectively as possible. Use default settings for most graphical elements in software systems (Best Practices for Data Visualisation, n.d.).

Conclusion

Data analytics and visualization is one of the most important parts in information technology or data science field. Its concepts and best practices can be applied in data collection and transformation, data analytics, and data visualization. Many applications such as Tableau or Power BI can work with data analytics and visualization smoothly. Moreover, some programming languages include Python or R are also can be worked with data analytics and visualization. Power BI is the one of them that can perform tasks in data visualization. Data importing and cleaning, data preparation and transformation, and data visualization are the core concepts and practices that come from data analytics and visualization plays as an important role to help many analysts or researchers success in their careers as Data Scientist.

References

Atlan, T. (2023, October 6). 11 Benefits of data visualization you can't ignore in 2024. *Atlan*. https://atlan.com/benefits-of-data-visualization/

Best practices for data visualisation. (n.d.). Best Practices for Data Visualisation. https://royal-statistical-society.github.io/datavisguide/

Data Analytics Central. (2023, February 18). *The New LINEST() DAX Function in Power BI - Regression Analysis Made Easy #regressionanalysis* [Video]. YouTube. https://www.youtube.com/watch?v=R2M_UBqaagg

GeeksforGeeks. (2024, June 11). *What is Data Visualization and Why is It Important?* GeeksforGeeks. https://www.geeksforgeeks.org/data-visualization-and-its-importance/

Hashemi-Pour, C., Brush, K., & Burns, E. (2024, April 17). *data visualization*. Business Analytics. https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

How to find Percentile using power bi DAX -. (2023, March 30). ProjectPro. https://www.projectpro.io/recipes/find-percentile-power-bi-dax

Jolijn. (2021, February 12). *Obtain insights using data visualization: 4 steps to take - LACO*. LACO. https://www.laco.be/obtain-insights-using-data-visualization-four-steps/

Quach, S., & Quach, S. (2024, April 11). *Leverage Data Visualization for effective communication*. Knowi. https://www.knowi.com/blog/data-visualization-for-effective-communication/

Rad, R. (2020, January 17). *Age Calculation in Power BI using Power Query*. RADACAD. https://radacad.com/age-calculation-in-power-bi-using-power-query

ITDS460 - Assignment Rubric

Student's Full Name: Mony Ho

EIU ID: EIU95527

Category	Marks Allocation	Marks Attained	Comments
1. Data Cleaning and Preparation	15 marks	13	
2. Exploratory Data Analysis	15 marks	12	
3. Data Analytics	30 marks	23	
4. Data Visualization	20 marks	17	
5. Report and Presentation	20 marks	17	
*** Penalty *** This penalty is for papers that do not meet the assignment requirements. (i.e. word count, format, etc.)	(-5)		
Total marks allocated and achieved:	100 marks	82	