Final Report of Traineeship Program 2024

On

"VISUALIZING EXECUTIVE SALES"

MEDTOUREASY

27th Feb 2024

ACKNOWLDEGMENTS

The traineeship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies of the subject of Data Visualizations in Business Intelligence; and also, for personal as well as professional development. I am very obligedfor having a chance to interact with so many professionals who guided me throughoutthe traineeship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training & Development Team of MedTourEasy who gave me an opportunity to carry out my traineeship at their esteemed organization. Also, I express my thanks to the team for making me understand the details of the Business Intelligence Analyst profile and training me in the same so that I can carry out the project properly and with maximum client satisfactionand also for spearing his valuable time in spite of his busy schedule.

I would also like to thank the team of MedTourEasy and my colleagues who made the working environment productive and very conducive.

TABLE OF CONTENTS

ABSTRACT

This project focuses on the development of an executive sales dashboard using Tableau, aimed at providing key performance indicators (KPIs) crucial for decision-making. The dashboard includes metrics such as total sales, total quantity, overall profit, and average discount, offering a comprehensive view of the sales landscape.

Initially, Python's Pandas library was utilized for basic data cleaning to ensure data integrity and accuracy. The cleaned dataset was then imported into Tableau for analysis and visualization.

The dashboard presents two main graphical representations: one illustrating sales trends over months across multiple years, and the other depicting profit trends over the same period. These visualizations provide insights into seasonal trends, identifying peaks and troughs in sales and profit.

Additionally, a filter pane is incorporated into the dashboard, allowing users to dynamically interact with the data. Filters such as category, subcategory, region, and segment empower executives to drill down into specific subsets of the data, enabling targeted analysis and decision-making.

This project aims to enhance strategic decision-making processes by offering a user-friendly and visually intuitive platform for exploring and understanding sales performance metrics.

1.1 About the Company

MedTourEasy, a global healthcare company, provides you the informationalresources needed to evaluate your global options. MedTourEasy provides analytical solutions to our partner healthcare providers globally.

1.2 About the Project

This project solely focuses on the performance of the sales and profit based on various criteria such as category, sub-category, region and segment by providing an interactive dashboard through which user can easily focus on the specific area of the need. This dashboard contains high level KPI’s to quickly assess the performance followed by monthly broken-down chart both for sales and profit to analyze the trend over the years and finally user can filter the data by provided filter options.

1.3 Deliverables and Requirements:

This project focuses on creating easily understandable, interactive and dynamic dashboards by gathering data of the Sales provided by the senior team member by using Python's pandas library to perform basic data cleaning like checking the data for any duplicates, and removing duplicates, finding outliers, renaming columns etc. and choosing the appropriate subset of the data for performing the analysis as asked in stakeholder requirements to visualize which will enable the firm to analyze the situation and draw conclusions regarding the sales. The prototype for all the dashboards will be created using Tableau (primarily to create dynamic visualizations like area charts for sales, profit, slicers etc.)

The dashboard requirements are as follows:

- The dashboard size is 1250px wide and 750px tall.
- Prefer using containers.
- The dashboard has total of 5 containers.
- The filter-pane must have four filters of category, sub-category, region and segment.
- Each filter has some padding.
- The dashboard should have the title 'Executive Sales'.
- The first chart should have the title "YTS KPIs" and consists four elements like Total Discount, Overall profit, Total Quantity and Total Sales,
- The second chart should have the title as 'Sales' and should show monthly sales per year.
- The third chart should have the title as 'Profit' and should show monthly profit per year.
- Do proper formatting for professional look.

The project consists of one dashboard which has 3 chart elements as detailed as follows:

a. Analysis of the Key Performance Indicators (KPI's): This visual gives the overall idea about Overall Profit, Total Quantity, Total Sales, Total Orders and Avg. Discount.

b. Analysis of the monthly sales: This visual focuses on analyzing the sales on monthly level. The requirement for this visual is as follows:

    - Make sure it's an Area chart with proper padding both inside and outside and also has a slightly dark grey border

c. Analysis of monthly profit: This visual focuses on analyzing the profit over the months to find any historical patterns. The requirement for the visual is as follows:

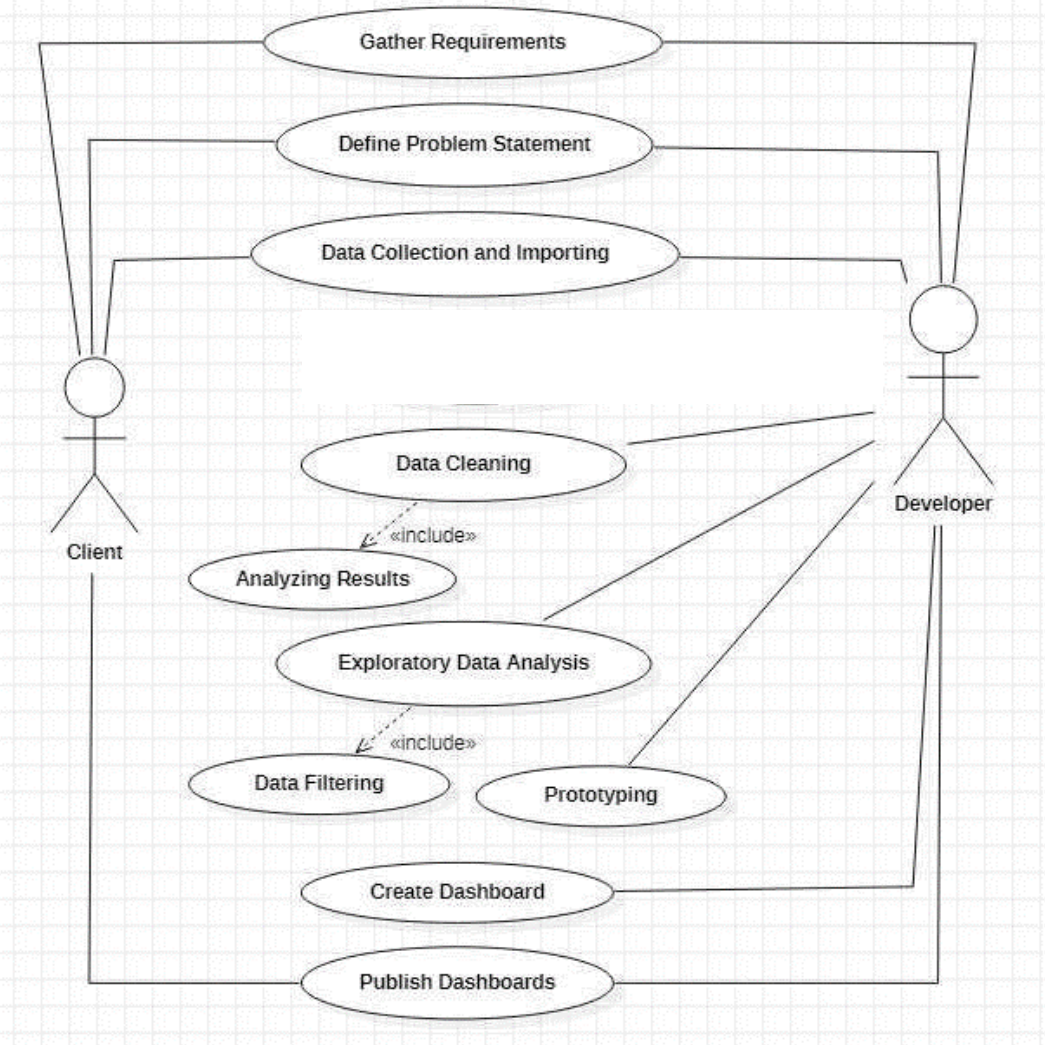
    - Make sure it is an Area chart with proper padding on both inside and outside and also has a slightly dark grey border.

I. METHODOLOGY

2.1 Flow of the Project

The project followed the following steps to accomplish the desired objectives and deliverables. Each step has been explained in detail in the following section.

2.2 Use Case Diagram



Above figure shows the use case of the project. There are two main actors in the same: TheClient and Developer. The developer will first gather requirements and define the problem-statement then collecting the required data and importing it. Then the developer will designdatabases so as to identify various constraints and relations in the data. Next step is to cleanthe data to remove irregular values, blank values etc. Next, exploratory data analysis is conducted to filter the data according to the requirements of the project. Then a prototype of the dashboards is created using Tableau to get a clear view of the visualizations to be developed. Finally, dashboard is developed and analyzed to publish the results to the client.

2.3 Language and Platform Used

2.3.1 Language: Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application - Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

The important features of python are:

- Free and Open Source.
- Easy to code.
- Object oriented language.
- GUI programming support.
- Large standard libraries.

2.3.2 IDE: JupyterNotebook

The Jupyter Notebook is the original web application for creating and sharing computational documents. It offers a simple, streamlined, document-centric experience.

Major features are:

- Language of choice: Jupyter supports over 40 programming languages, including Python, R, Julia, and Scala.
- Share notebooks: Notebooks can be shared with others using email, Dropbox, GitHub and the Jupyter Notebook Viewer.
- Interactive output: Your code can produce rich, interactive output: HTML, images, videos, LaTeX, and custom MIME types.
- Big data integration: Leverage big data tools, such as Apache Spark, from Python, R, and Scala. Explore that same data with pandas, scikit-learn, ggplot2, and TensorFlow.

2.3.3 Packages:

1. Numpy:

It is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. At the core of the NumPy package, is the ndarray object. This encapsulates *n*-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance.

2. Pandas:

It is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis/manipulation tool available in any language. It is already well on its way toward this goal.

- Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet.
- Ordered and unordered (not necessarily fixed-frequency) time series data.
- Arbitrary matrix data (homogeneously typed or heterogeneous) with row and column labels.
- Any other form of observational / statistical data sets. The data need not be labeled at all to be placed into a pandas data structure.

3. Plotly Express:

The plotly.express module (usually imported as px) contains functions that can create entire figures at once, and is referred to as Plotly Express or px. Plotly Express is a built-in part of the plotly library, and is the recommended starting point for creating most common figures. Every Plotly Express function uses graph objects internally and returns a plotly graph objects. Figure instance. Throughout the plotly documentation, you will find the Plotly Express way of building figures at the top of any applicable page, followed by a section on how to use graph objects to build similar figures. Any figure created in a single function call with Plotly Express could be created using graph objects alone, but with between 5 and 100 times more code.

II. IMPLEMENTATION

3.1 Gathering Requirements and Defining Problem Statement

This is the first step wherein the requirements are collected from the clients to understand the deliverables and goals to be achieved after which a problem statement is defined which has to be adhered to while development of the project.

3.2 Data Collection and Importing

Data collection is a systematic approach for gathering and measuring information from a variety of sources (in this case only one) in order to obtain a complete and accurate picture of an interest area. It helps an individual or organization to address specific questions, determine outcomes and forecast future probabilities and patterns.

In this case the data was already provided by the senior team member in form of the Excel workbook containing 3 worksheets as below:

- Orders
- Person
- Returns

Data importing is referred to as uploading the required data into the coding environment from internal sources (computer) or external sources (online websites and data repositories). This data can then be manipulated, aggregated, filtered according to the requirements and needs of the project.

3.3 Database Schema:

Once the data is collected and imported into the Tableau environment, it is important to design the structure of the database tables so as to identify the constraints in the data, keys, dependencies and relations between various tables.

The data contains of three Excel sheets like Orders, People and Returns. Only the Orders sheet is used for analysis purpose. The schema of the table is as mentioned below:

Attribute	Data type	Extra
Row_ID	VARCHAR	Primary Key
Procedure_ID	VARCHAR	Not Null, Unique
Procedure Date (Order Date)	Date	Not Null
Discharge Date	Date	Not Null
Ship Mode	CHAR	
Patient ID	VARCHAR	Not Null
Patient Name	VARCHAR	
Segment	CHAR	
Country	CHAR	
City	CHAR	
State	CHAR	
Postal Code	INT	
Region	CHAR	
Transaction ID	VARCHAR	Not Null
Category	CHAR	
Sub Category	CHAR	
Sales	FLOAT	
Days (Quantity)	INT	
Discount	FLOAT	
Profit	Float	

3.4 Data Cleaning

"Quality data beats fancy algorithms"

Data is the most imperative aspect of Analytics and Machine Learning. Everywhere in computing or business, data is required. But many a times, the data may be incomplete, inconsistent or may contain missing values when it comes to the real world. If the data is corrupted then the process may be impeded or inaccurate results may be provided. Hence, Data cleaning is considered a foundational element of the basic data science.

Data Cleaning means the process by which the incorrect, incomplete, inaccurate, irrelevant or missing part of the data is identified and then modified, replaced or deleted as needed.

With reference to the Executive Sales dataset, it may contain many null values or incorrect value simply because of inconsistency in sales and profit. Hence various functions/methods are used to clean this data.

Approach:

The need of data cleaning in this case is a minimal, since the data provided is almost cleaned and formatted correctly. Nevertheless, it is essential to double check the data before using it for the analysis.

1. Checking for the duplicates in the data using pandas df.duplicated() and then dropping that duplicated values using df.drop_duplicates().

```
df[df.duplicated()]
```

Row ID	Procedure ID	Procedure Date (Order Date)	Discharge Date	Ship Mode	Patient ID	Patient Name	Segment	Country	City	State	Postal Code	Region	Transaction ID	Category	Sub-Category	
3407	US-2015-150119	2015-04-23	2015-04-27	Standard	LB-16795	Laurel Beltran	Corporate	United States	Columbus	Ohio	43229.0	East	FUR-CH-10002965	Cardiology	Coronary Angioplasty	2

```
df = df.drop_duplicates()
```

2. Formatting the data in desired format: This is done using df.dtypes method as shown below.

```
df = pd.read_excel('Patients - United States.xlsx')
df = df.set_index('Row ID')
df.dtypes
```

```
Procedure ID                          object
Procedure Date (Order Date)    datetime64[ns]
Discharge Date                 datetime64[ns]
Ship Mode                             object
Patient ID                            object
Patient Name                          object
Segment                               object
Country                               object
City                                  object
State                                 object
Postal Code                          float64
Region                                object
Transaction ID                        object
Category                              object
Sub-Category                          object
Sales                                float64
Days (Quantity)                        int64
Discount                             float64
Profit                               float64
dtype: object
```

3. Finding Null/Nan values in the data: There are some null/Nan values in the 'Postal Code' column but since the weight of that column is negligible in the analysis, we can ignore those null/Nan values.

```
df.isnull().sum()
```

```
Procedure ID                    0
Procedure Date (Order Date)     0
Discharge Date                  0
Ship Mode                       0
Patient ID                      0
Patient Name                    0
Segment                         0
Country                         0
City                            0
State                           0
Postal Code                    11
Region                          0
Transaction ID                  0
Category                        0
Sub-Category                    0
Sales                           0
Days (Quantity)                 0
Discount                        0
Profit                          0
dtype: int64
```

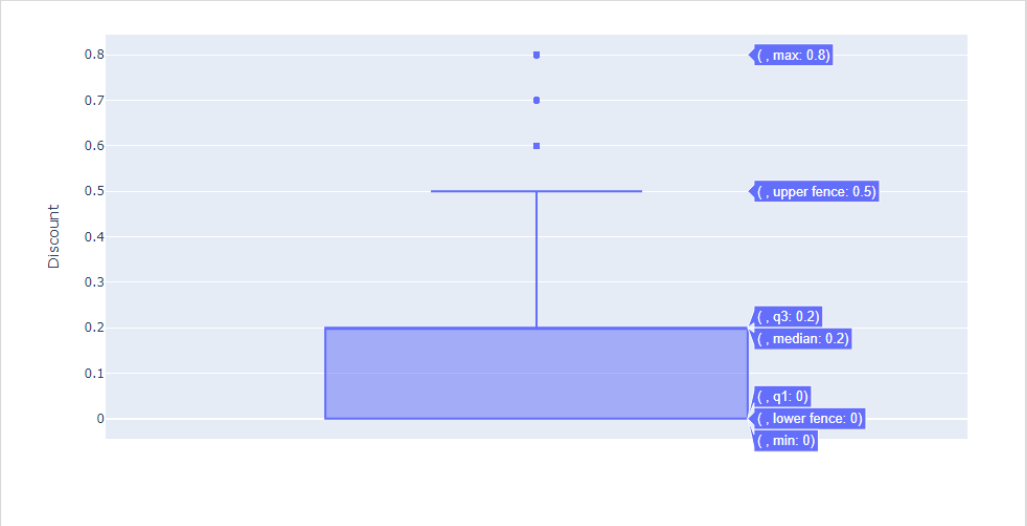
4. Renaming the dataframe columns using pandas df.rename() for consistency across the project.

```
columns = {'Procedure ID':'Procedure_ID','Procedure Date (Order Date)':'Procedure_Date',
           'Discharge Date':'Discharge_Date','Ship Mode':'Ship_Mode',
           'Patient ID':'Patient_ID','Patient Name':'Patient_Name',
           'Postal Code':'Postal_Code','Transaction ID':'Transaction_ID',
           'Sub-Category':'Sub_Category','Days (Quantity)':'Days_Quantity'}
df = df.rename(columns = columns)
```

5. Getting the overview of the data distribution by pandas df.describe(). The results are as below:

	Sales	Days_Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000
mean	229.857989	3.789574	0.156203	28.656783
std	623.245113	2.225110	0.206452	234.260148
min	0.440000	1.000000	0.000000	-6599.980000
25%	17.280000	2.000000	0.000000	1.730000
50%	54.490000	3.000000	0.200000	8.665000
75%	209.940000	5.000000	0.200000	29.360000
max	22638.480000	14.000000	0.800000	8399.980000

6. Visualizing the distribution by means of the box-plots. The outlier detection for one of the parameter Discount is as shown below, for other parameters like Sales, Profit and Days_Quantity approach is the same:



Python Code:

```
box_plot = px.box(data_frame = df,
                  y = 'Days_Quantity')
box_plot
```

3.5 Data Filtering

Data filtering is the method of choosing a smaller portion of the data set and using that subset to view, analyze and evaluate data. Generally, filtering is temporary – the entire data set is retained, but only part of it is used for calculation. It is also called sub setting or drill down data wherein data is extracted with respect to certaindefined logical conditions. Filtering is used for the following tasks:

- Analyzing results for a particular period of time.
- Calculating results for particular groups of interest.
- Exclude erroneous or "bad" observations from an analysis.

With respect to Executive Sales dataset, the data needs not to be filtered, since we are performing the analysis over the whole period provided which is between 2015-2018.

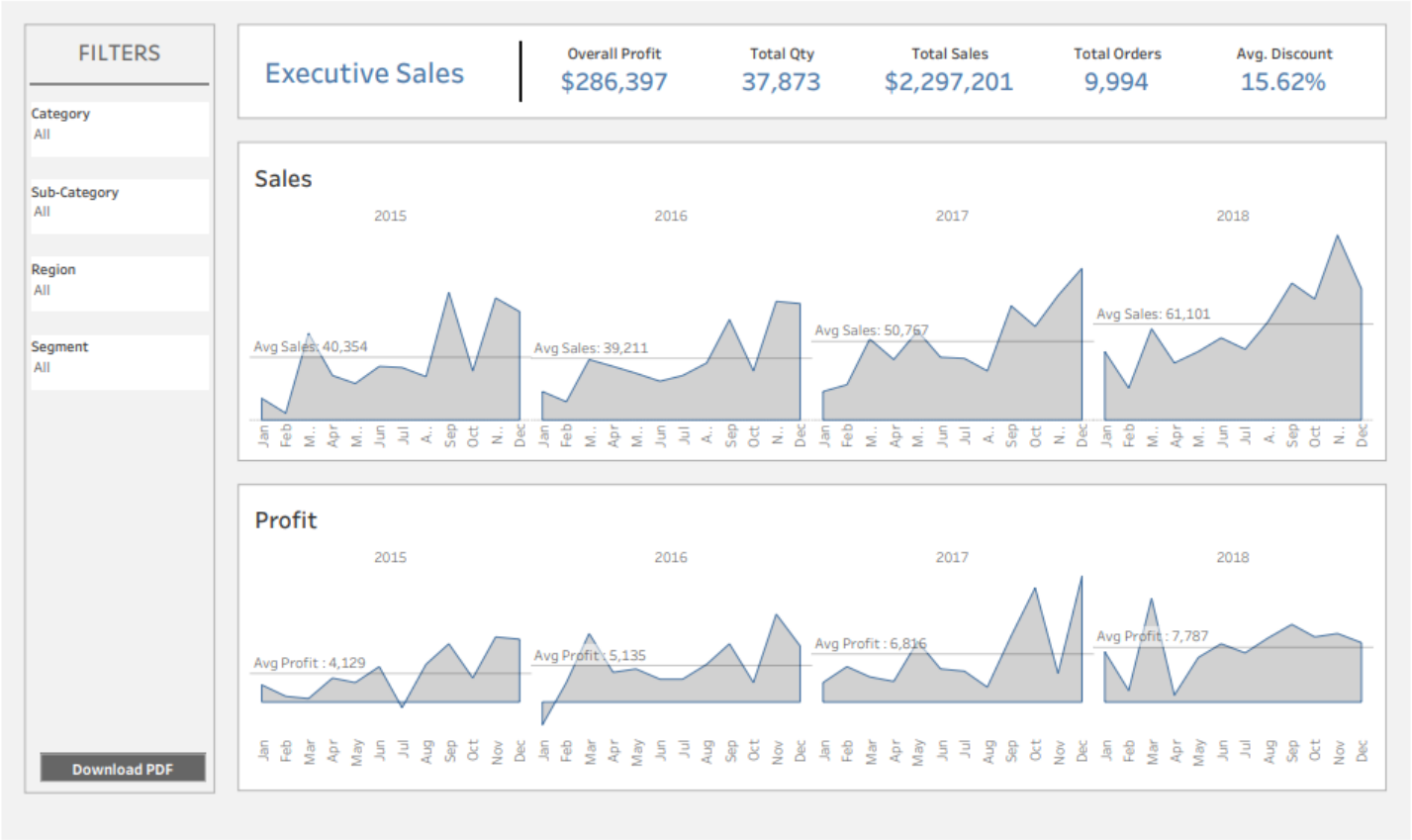
3.6 Prototyping – Tableau

A prototype is an early version, model, or release of a product that is constructed to test a design or process. It is generally used by system analysts and users to assess a new design to enhance precision. Prototyping serves to specify a real, working system rather than a theoretical one. Creation of a prototype in some design workflow models is the step between formalizing and testing an idea.

Tableau is Salesforce's business analytics software. It aims to provide interactive visualizations and business intelligence capabilities with an interface that is easy enough to create your own reports and dashboards for end users. It provides capabilities for data planning, data discovery and interactive dashboards. It has the following features:

- Easy to connect, model, and visualize data, creating memorable reports personalized with KPIs and brand.
- Data is better secured across Tableau dashboards, and data sets with persistent protection that keeps working even when shared outside theorganization or exported to other formats such as Excel, PowerPoint, and PDF.
- Input: In the form of Excel, CSV, text, SQL and other formats
- Visualizations: Wide variety of graphs, infographics, KPIs, Filters, Slicers, etc.
- Output: Easily publishable reports and dashboard

Screenshot of the prototype is as follows:



The prototype provides three main visuals as mentioned in the business-requirement Monthly Sales, Monthly Profit and filter pane for filtering the data as needed and finally a download PDF button to quickly share report as PDF

3.7 Development of Dashboards

The dashboard has been created using the containers of two kinds mainly vertical and horizontal. There are 2 main sections to focus on Monthly Sales and Monthly Profit across the years. It gives the user idea about how sales are going from month to month for the given year and the average lines shows the Avg performance over year to quickly compare it to the previous year.

3.7.1 Defining Visuals

Data visualization is presenting data in a graphical or pictorial format. It allows decision-makers to see visually presented analytics, so that they can grasp difficult concepts or identify new patterns. In interactive visualizations, technology can be used to dig in charts and graphs for more detail, interactively modifying what data one can see and how it works.

Because of the way in which the human brain processes information, it is easier to visualize large amounts of complex data using charts or graphs than to poring over spreadsheets or reports. Data visualization is a quick, easy and universal way of conveying concepts. Data visualization can also:

- Identify areas that need attention or improvement.
- Clarify which factors influence customer behavior.
- Help you understand which products to place where.
- Predict sales volumes.

In Tableau each visualization is made from one or more sheets (Ideally one) and then Each visualization is formatted as per the requirement. After that dashboard making is started in which each visual made previously is inside the containers and final layout is being created. Which may contain visuals, KPI's and the filter pane. It contains the following types of containers:

1. Vertical Container: It allows the users to put visuals one above the other (stacked format).
2. Horizontal Container: It allows the users to put visuals side by side.

3. Blank: It is used in the layout design section and acts as a placeholder for the visuals and allows user to visually interpret the layout of the dashboard. Mainly used in creating the mockup for the dashboard

III. SAMPLE SCREENSHOTS AND OBSERVATIONS

4.1 Executive Sales Final Dashboard

4.1.1 YTS KPI's:

Top KPI's are shown at the top of the dashboard. Which consists of Overall Profit, Total Qty, Total Sales, Total orders, and Avg.Discount. Slight increment in the discounts had big impact on the sales and profit.

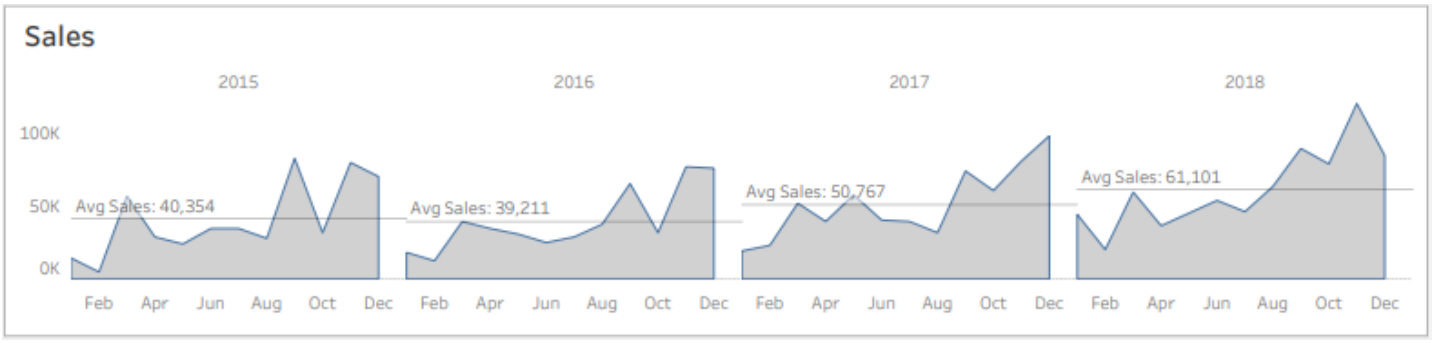
Observations:

- Overall Profit: YoY Change 2017-18 (**14.23%)**
- Total Qty: YoY Change 2017-18 (**26.82%**)
- Total Sales: YoY Change 2017-18 (**20.35%**)
- Count of Orders: YoY Change 2017-18 (**28.02%**)
- Avg.Discount: YoY Change 2017-18 (**0.18%**)

4.1.2 Sales

The chart below shows the sales over months across the years. Since the fluctuation of the sales over months is high the average lines are also provided to compare YoY change more easily.

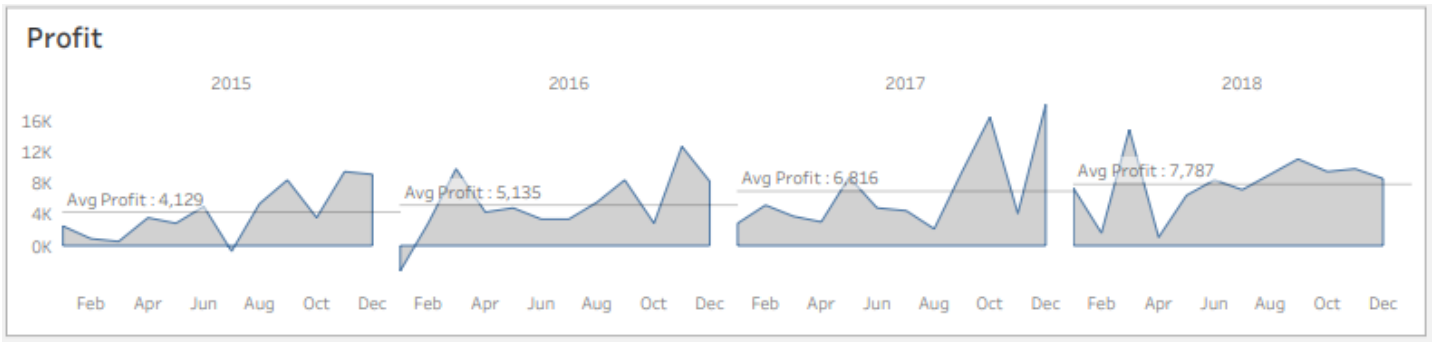
Sales Observations: The only year where YoY change was negative was for year 2015-16 (-2.83%). The highest YoY change was for year 2016-17 (**29.47%**). From last 3 years the YoY change is positive.



4.1.3 Profit

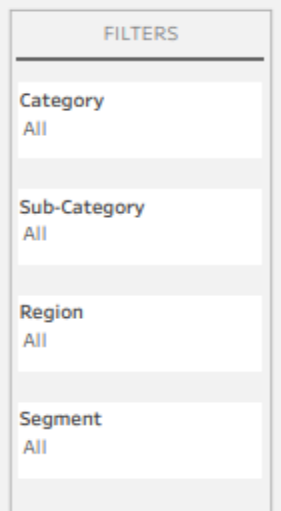
The chart below shows the profit over months across the years. Since the fluctuation of the sales over months is high the average lines are also provided to compare YoY change more easily.

Profit Observations: The Avg profit of the company is always growing over the years. The highest YoY change was from years 2016-17(**32.73%**). The only time company faced losses are July-2015 ($ **-841**) and Jan-2016 ($ **-3281**)



4.1.4 Filters

The filter pane consists of main 4 filters, i. Category (Single select), ii. Sub-category (Multi select), iii. Region (Single select) and iv. Segment (Single select)



IV. CONCLUSION AND FUTURE SCOPE

The executive sales dashboard developed using Tableau has successfully provided a comprehensive overview of key performance indicators (KPIs) crucial for decision-making in sales, highlighting insights into total sales, total quantity, overall profit, and average discount. This user-friendly platform empowers executives to make informed decisions, optimizing strategies for enhanced profitability. Despite encountering limitations in data granularity, particularly in certain regions and segments, the dashboard adds tangible value by improving efficiency in accessing and interpreting sales data.

Recommendations for future enhancements include expanding KPIs to include customer acquisition cost and sales conversion rates, integrating predictive analytics for forecasting, and enhancing visualization with interactive features. The future scope involves integrating real-time data updates, implementing machine learning algorithms for demand forecasting, and establishing regular user training sessions with feedback mechanisms.

With regards to the future work, the firm aims at regularly updating the dashboards with time and integrating it with their systems so as to continually draw conclusions and analyze the results. This will enable them to predict future business opportunities and provide a basis on which they can plan on increasing their market presence and capacity planning.

V. REFERENCES

Data Collection

The following websites have been referred to obtain the input data and statistics:

https://drive.google.com/file/d/1XsNol-NdHhudke8RjIaOd2Ws_yctniA3/view?usp=

Programming References

The following websites/videos have been referred for Tableau dashboard making:

a. https://youtu.be/mAYvayW77Vw?si=edGNq6B5hw7AGV1Ahttps://rstudio.com/resources/webinars/
b. https://youtu.be/3w4s_6r3B6A?si=5gOr71_VSKDqVL7C
c. https://youtu.be/r-75D9JzVTI?si=DacNyxSap1oQUFBb
d. https://jupyter.org/
e. https://pandas.pydata.org/pandas-docs/stable/getting_started/overview.html
f. https://numpy.org/doc/stable/user/whatisnumpy.html