**Sales Analysis and Insights for US Superstore**

Group Member

# Background and Motivation

Under the impact of the global pandemic COVID-19 in 2020, online shopping is becoming more and more popular not only in the US but also all over the world because of its speed and convenience. According to Statista – a famous German company specializing in market and consumer data - estimated that 1.92 billion people purchased goods or services online in 2019. I believe this figure is much bigger in 2020. Because online shopping is so popular, many brick-and-mortar retail-stores are forced to do digital transformation. Therefore, I believe that understanding sale and profit dynamic of current online stores is of great importance for those intent-to-change retail stores to avoid loss in the fast-growing market.

# Project Objectives

In this report, I assume I am a business analyst works for this online store and reports to the sales to the head sales manager. Therefore, I will try to answer the following questions to help the sales manager to make sales decisions.

## Geographical Analysis

1. Which state in the US has the greatest need for online shopping in 2018, does the need change when compare to past years?
2. the Revenue-profit and customer-order pattern across states in the US
3. the distribution of loyal customers across states in the US

## Product Analysis

1. which product category is popular among US customers? what’s the best-selling ranking?
2. What’s the sales and profit relationship for each product category

## Customer Analysis

1. What kind of customers can be treated as loyal customers?
2. What’s the geographical distribution for loyal customers?

## Salesperson Analysis

1. What’s the performance for salespersons? What’s the ranking for best salesperson?
2. Do they reach their own KPIs in 2017?

# Data

The dataset will be used in this report is from Kaggle (https://www.kaggle.com/juhi1994/superstore), which is basically sales data of an online store called superstore in the US from The dataset can answer the analytical questions listed in the last section because it has information as following lists shows:

1. General information, including, order date and receive date for each order, sales, profit and quantity etc.
2. Geographical information, including city, state and postal code for each order etc.
3. Product-related information, including product name, product category, product sub-category
4. Customer-related information, including customer name, customer ID etc.
5. Salesperson-related information, including sales name for each order etc.

There are 23 variables in total in the dataset, the description is shown as following table:

A picture containing text, newspaper

Description automatically generated

Figure Data Description

# ETL

## Import and merge data

Because the dataset includes 3 sheets, so first I have to merge all the data sheets that I can do the analysis based on the full dataset.

Text

Description automatically generated

Figure import data and merge data

## Handle missing values

Graphical user interface, text, application

Description automatically generated

Figure detect missing values

First, I check how many missing values in total and I found there are 9194 missing values. When I want to handle the missing values, I have to understand the nature of missing values, from the figure 3, I can observe that all missing values are from the variable ‘Returned’. Because I did merge in the last session, I should understand missing values in ‘Returned’ column means some orders do not have return information, and those missing values cannot be handled by simply remove them, therefore I can replace the n/a values with the tag ‘0’ as figure 4 shows: all 9194 missing values are replaced with 0.

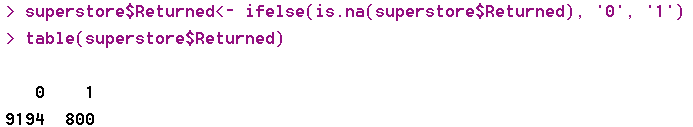


Figure handle missing values for column 'Returned'

Finally, the plot of distribution for returned items is shown as follows:

Graphical user interface

Description automatically generated

Figure Returned item distribution

## Check other abnormal values

After I handled all missing values, I start checking other abnormal values. for this checking I use summary function, which is easy for us to detect abnormal.

As figure 6 shows, the summary of all variables is shown, from which, I can observe that:

1. All date variables are in proper format
2. All qualitive variables are in character format
3. All quantitative variables are in numeric format

|  |  |
| --- | --- |
| Text  Description automatically generated | Table  Description automatically generated |
|  | |

Figure summary of all variables

Therefore, I can conclude that this dataset is a pretty clean dataset, which means it does not need much more transformation to make it more structured.

## Write cleansed dataset to csv

Final step for my ETL session is write the cleansed dataset as csv file to local:



Figure write csv file to local

This csv is file what I will use to do visualization in tableau.

# Design and Implementation of Visualization

## Audience

For this dashboard, I assume I am the business analyst who have to report monthly/quarterly/monthly to the sales manager. Therefore, the whole dashboard design will be catered for sales related need of sales manager

## Dashboard overview

The dashboard can be found in this link: <https://public.tableau.com/profile/yilang.shen#!/vizhome/SalesReportingDashboard_16073496079740/Dashboard3?publish=yes>

The overview of the dashboard is shown as following:

Graphical user interface, application

Description automatically generated

Figure 8 dashboard overview

## Dashboard structure

|  |  |  |  |
| --- | --- | --- | --- |
| Sections | Section names | Functions | Descriptions |
| Main Section | Current state of the business | An overview of the current business performance compared to performance of  1) Last Period  2) Same Period of Last year  on the reporting date | Six key performance factors:   1. Total Revenue 2. Total Orders 3. Total Customers 4. Avg. Revenue per Order 5. Avg. Revenue per Customer 6. Total Profit |
| Sub-Section 1 | Product Analysis | Help sales manager to understand the best-selling product | The top 7 best-selling product group and the relationship between revenue and profit in those products |
| Sub-Section 2 | Customer Analysis | Help sales manager to gain insights of loyal customers | The 1) number 2) details (names, orders etc.) 3) geographical distribution of loyal customers |
| Sub-Section 3 | Salesperson Analysis  - KPI Check | Help sales manager to check if 4 salespeople reach the KPI and decide who can get the bonus | The 1) overall performance 2) KPI performance of 4 salespeople |

Figure Dashboard structure

## Implementation

for the whole dashboard, I implement a main filter. As for each section, I have filters for them as well.

1. Main filter

The main filter I implement for the dashboard is reporting scope – monthly, quarterly and yearly – in which audience can easily choose the reporting time period he or she wants to view, and it satisfied all kinds of viewing needs related to reporting scope.

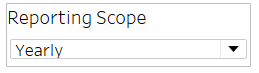


Figure design for main filter

The main filter is for the whole dashboard, which means it applies for the whole dashboard.

1. Sub-filters

For each section, I also have filters for them, and descriptions of sub-filters for each section is shown as following.

|  |  |  |  |
| --- | --- | --- | --- |
| Section names | Filter design | Filter values | Filter description |
| Current state of the business | Graphical user interface, application, table  Description automatically generated | Comparison (2):   * Current vs last period * Current vs same period last year   Region (4)   * Central * East * South * West | 1. It helps the audience to compare the current figures with last period or same period last year 2. It helps audience to choose to view the business status of a specific region |
| Product Analysis | Table  Description automatically generated | State   * All states in the US   Type-in filter   * Audience self-defined filter, can key-in any numbers and related graph will show the top N ranking | 1. Help audience to understand product-related info in state level 2. Show top-N ranking, which is user friendly |
| Customer Analysis | Graphical user interface, application  Description automatically generated | # of orders in period   * Drag filter, audience can self-define loyal customers, when choose 2 means the customers make 2 orders in the current period   Region (4)   * Same as section 1 | 1. Help audience to define loyal customers (who make at least N orders) |
| Salesperson Analysis  - KPI Check | A picture containing text  Description automatically generated | Type-in KPI   * A dynamic filter, audience can self-define the KPI target | 1. Help audience to set a KPI target |

Figure filter description of the dashboard

# Results and conclusions

By implementing the ETL and data visualization, I am able to answer questions that I am interested in:

1. Superstore has top sales in the following 5 states: California, New York, Washington, Texas and Pennsylvania
2. In the Sales and profit prospect:
3. When comes to yearly comparison, the sales and profit both increase
4. When comes to quarterly comparison, both sales and profit increased when compared to last quarter while profit decreased when compared to same quarter last year
5. When comes to monthly comparison, both sales and profit show a decreasing trend no matter compare to last month or same month last period
6. In the product analysis, I can observe that:
7. the best-selling product category is chairs, storage, tables, phones and appliances
8. some best-selling product category have negative profits, for example binders and tables, which means the store should review the pricing strategy of these products
9. for the customer analysis, I can conclude that:
10. there are in total 493 loyal customers in the yearly basis, and most of them are in California, Texas and New York. This loyal customer related finding is aligned with the sales distribution.
11. For the salesperson analysis, I found that:
12. Salesperson Anna Andreadi contributes the most the yearly sales – which is 34%, while the worst salesperson in the 4 is Cassandra brandow who only contributes 17% to the sales.
13. If I set the KPI target as 50000, all the four salesperson reached the target; however when I raise the KPI target to 15000, only Anna Andreadi and Chunk Magee reached the target