**Sales Data Technical Report**

**Sana Mourad, Rand Al-Shoubaki, Leen Daqa, Noor Yazeed**

A picture containing logo

Description automatically generated

**Professor Mohammed Azzeh, Data Visualization**

**Summer Semester 2022**

Abstract

During the data visualization process of the given data set we found that there were 4000 rows and 30 columns. We also looked at the meta data and made sure to understand the significance of the variables, also tried to find all the relationships between all the data attributes them by plotting to be able to improve the sales. This study aims to determine how this company can increase its profits. We did data analysis on sales raw data, started filtering to focus on the important data, then mapped the data. The results of our analysis will also be shown in the report.

Keywords:

* Quantity: the quantity of the product sold
* Rate: the tax rate applied to each purchase
* Holiday: whether there is a holiday at that time (0: no holiday, 1: holiday)
* Customer Type: whether they are members or not of the rewards program
* Employee Role Description: whether they are a cashier or captain
* For branch location and customer city the following replacements were used

(online was used as it is).

|  |  |  |  |
| --- | --- | --- | --- |
| North | East | South | West |
| Irbid | Mafraq | Ma’an | Balqa |
| Ajloun | Zarqa | Aqaba | Madaba |
| Jerash | Amman |  | Karak  Tafilah |

* EDA: exploratory data analysis

Sales Data Technical Report

The data displays information about the sales of a certain retail department store in different cities of Jordan. It has many departments such as Kids, Women, Home, Beauty …etc. They also have an online store where the role description of the employees is Captain. The store also runs different promotions throughout the year in hopes of attracting customers and therefore increasing profits. Data is captured at two main points the store’s back door where deliveries are made and at the POS (point of-sale) system where products are checked out of the system and then leave the store. We found that we had to use more than one platform to perform EDA and visualization, the main platform used was RStudio.

# Methodology

### Using Python (on Google Colab)

Since the data was really big we decided to start with python since it is the easiest and fastest option available. We removed unnecessary columns such as customers and employees personal data (first name, last name, and email) and the city area. We also replaced the customer city, and branch location values according to the table in the keywords while keeping the online branches save as ‘Online’.

#### Using Microsoft Excel Sheet

We found that it was easier to use the date/time data on excel than other methods since we had previous deep knowledge of excel. It was used to calculate the ages of customers and employees, also to calculate how many months an employee has worked for the company. Another benefit of excel was that it made it easier for us to calculate in which quarter orders were made and promotions were run.

* 1: January-March
* 2: April-June
* 3: July-September
* 4: October-December

##### Using RStudio

RStudio was used to explore the data. For categorical variables we checked the unique values to see if they are useful in the visualization, we also checked the skewness and sharpness of the peak of the distribution of the numerical variables. Numeric variables that we found are better used as categorical, such as Holiday, orderDate\_quarter and promotionDate\_quarter, were transformed using as.factor(). A long data frame was desperately needed for some plots of only numeric variables. A lot of graphs such as bar, scatter, pie, line and histogram plots using many customizations such as aesthetics, facet\_wrap, position, themes and colors to make their appearance more pleasing to the eye of the viewer.

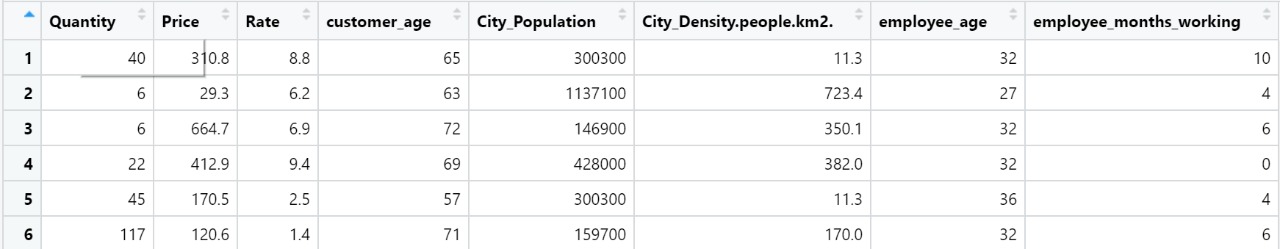
# Results

In conclusion, we saw that the online branch is the main option for customers and the most selling departments are Kids, Women and Beauty, most of the customers are between 20 and 60 years old. Prices are the highest in the West, and lowest in the South. The retailer should do promotions to increase the number of people using credit cards as the payment method as it is safer and easier. We also recommend the store improve their rewards program to encourage members to spend more at the store.

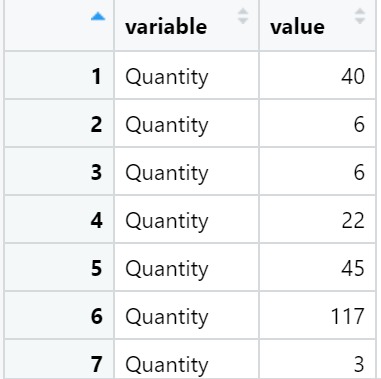
# Appendix

|  |  |  |  |
| --- | --- | --- | --- |
| North | East | South | West |
| Irbid | Mafraq | Ma’an | Balqa |
| Ajloun | Zarqa | Aqaba | Madaba |
| Jerash | Amman |  | Karak  Tafilah |

1. This table explains in which category each of the cities found in the data set were put:
2. The table of numeric variables created in RStudio



1. The numeric table as long data instead of wides



# Footnotes

The work done on the project was not clearly split among the group members. Instead, we were all working on one device with one person typing and everyone providing input.