**Logo, company name

Description automatically generated**

**Assessing Retail Sales for an Anonymous Company and The External Factors That Affect It**

**Analytical Report**

**Prepared By:**

Abdullah Majali 20200378

Tamara Otaibi 20200961

Abdelrahman Elayyan 20200441

Dana Twal 20200959

**Course Name**: Data Visualization **Course Instructor:** Prof. Moh’d Azzeh

**Team Contributions:**

* Abdullah Majali
  + Feature engineering: Creation of new columns
  + Exploratory analysis: Pie charts, Bar Plots
  + Prepared Presentation Slides
  + Data cleaning: scanning for data discrepancies
* Abdelrahman Elayyan
  + Exploratory analysis: Bar charts, box plots
  + Data Exploration: looking for any discrepancies in the data
  + Feature engineering: Creation of new columns
* Tamara Otaibi
  + Exploratory analysis: individual findings in R.
  + Explanatory analysis: Consolidation of findings.
  + Prepared project presentation slides.
* Dana Twal
  + Exploratory analysis: Correlation heatmaps
  + Exploratory analysis: Computing P-value measures
  + Exploratory analysis: Computing measures of central tendency/symmetry

**1. Executive Summary**

**1.1 Aim**

This report aims to gain insight into external factors that may have affected the sales revenue on a variety of branches for a chain store and recommendations for increasing sales revenue. Data was collected from cash registers as customers purchased products from different departments. External factors that might have impacted the sales are the holiday seasons, promotions, customer membership, customer age group, employee tenure, and employee type. Giveaways had the highest sales compared to sales promotions having the lowest sales. Also, the average amount spent in all departments with promotions is greater than without promotions. During the holiday season, departments that had higher sales were gifts, activewear, and sales. Customer membership and ages have affected the sales as it shows that the 39-59 age group had the highest sales, especially female customers in that age group. Also, customers with memberships have spent more money in the store. Logically, this is explained by members receiving access to promotions and discounts. Employees, such as their employment length and types, with fixed-term contracts and 25+ months of employment, have impacted the sales revenue significantly. The results of this study display that the data collection method is limited as the other external factors have shown no significant difference or no correlation. For future research, variables such as stock count, older dates for sales revenue data, and customer satisfaction surveys can be considered.

**1.2 Methods**

From the dataset consisting of 4000 records with 33 features, we performed EDA and feature engineering. The data was then investigated for any relationships, outliers, and inconsistencies by using statistical tests and measures along with visualization techniques including but not limited to, box plots, bar charts, and pie charts, to draw our conclusions.

**2. Introduction**

In the high-pressure quest to make a sale, acquire a good profit, and beat out other bidders, short-term strategies like cutting prices, offering discounts, or making other concessions are frequently resorted to. These short-term strategies are destructive to the long-term sustainability of a business. A sales analysis report helps to rethink the approaches to sales goals so that businesses not only sell a greater quantity but also sell with the bottom line in mind.

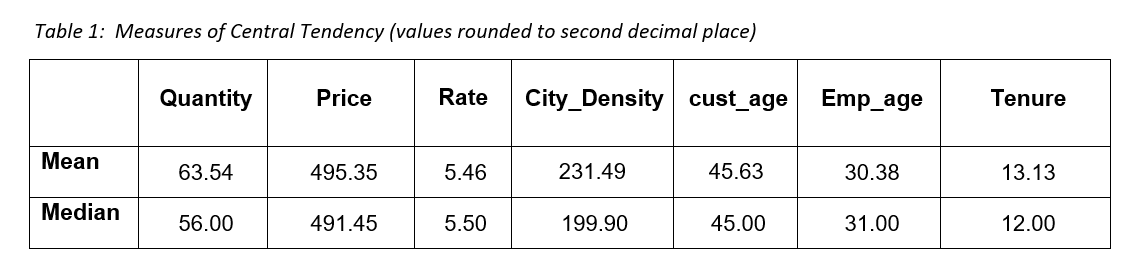
The dataset at hand shows the company’s sales broken down by geography, product department, customer group, membership, and date of items bought. It provides attributes that may have influenced sales in the years (2020-2022), some of these attributes include information about the customers; their birthdate, their cities, and whether they have a membership or not. Some provided customer variables such as their names and emails, however, were excluded. It also provides attributes about the employees such as the hire date, birthdate, working hours, etc. What we perceived to be the most influential factors were promotions, holidays, customer membership, customer age group, and employee tenure.

According to our analysis, higher revenues were accredited to both customers and employees, we saw that on average customers were willing to spend more on holidays and promotions; memberships also had an effect since members have spent more overall in the store than non-members. The majority of the sales also came from adult customers and that can be attributed to the fact that customers aged 20-59 had more money to spend than retirees ( 60+) and teenagers (11-20). In addition, employees with the longest tenures tended to complete more sales orders and were handling a majority of online orders which netted higher sales than their less experienced colleagues.

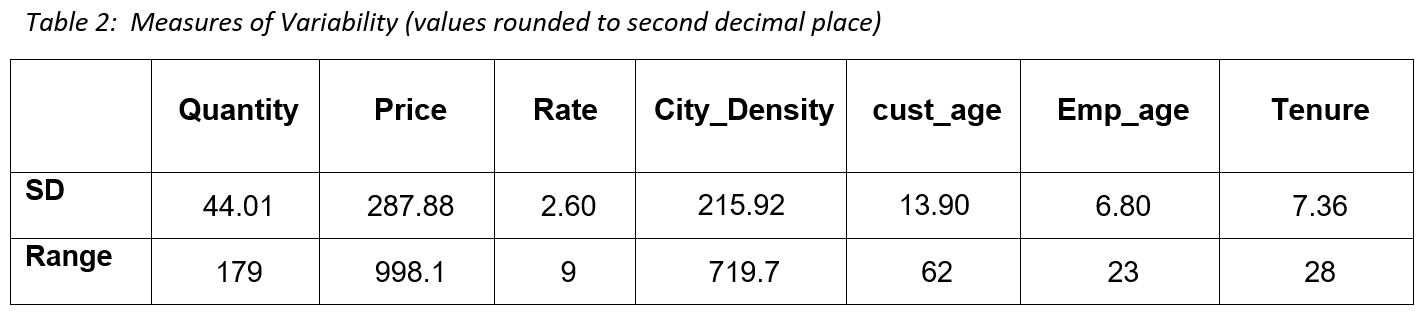
**3. Discussion**

**3.1 Main statistical features**

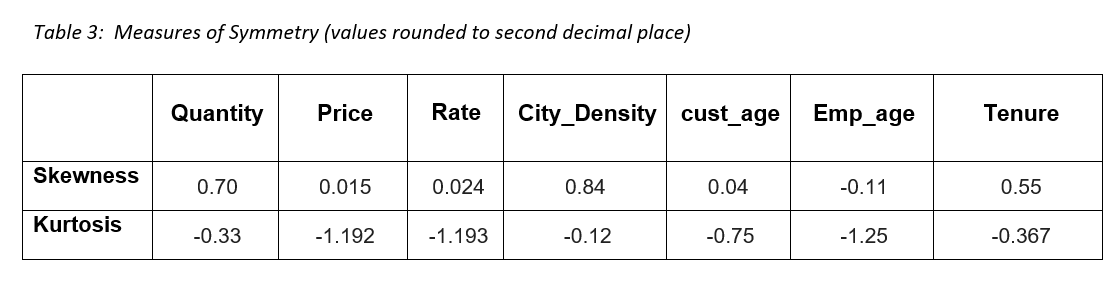
**3.1.1 Measures of Central Tendency**

****

**3.1.2 Measures of Variability**

****

**3.1.3 Measures of Symmetry**

****

To examine the symmetry of the continuous features, we plotted their distributions using boxplots and discovered that the variables had almost normal distributions. To validate this further, we performed skewness and kurtosis tests.

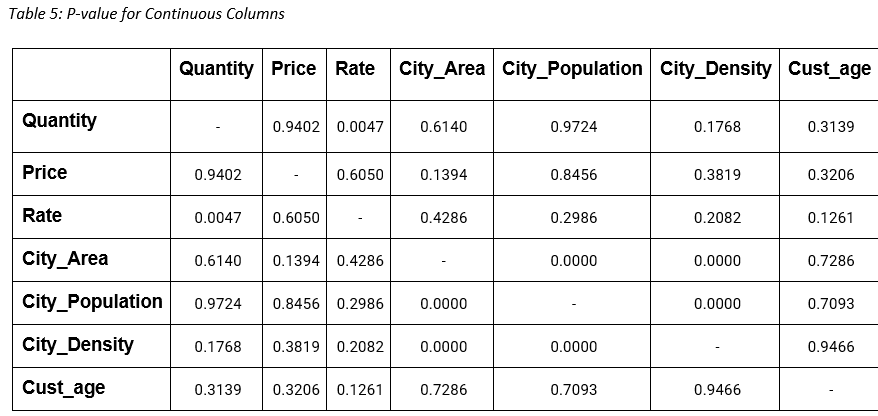
Skewness measures the deviation of the given distribution of a variable from a symmetric distribution, such as normal distribution.

For interpretation of the skewness results we used the scale in table4, from it we can interpret that City\_Density, Quantity, and Tenure are moderately symmetrical while the rest of the features are fairly symmetrical.

The kurtosis statistic measures how heavy the tails of a distribution are, for interpretation of the results of kurtosis we used the scale in table5.

The kurtosis values of the features imply that all features' tails' weights are heavier than a normal distribution.

**3.1.4 P-values**

****

**3.2 Central statistical relationships and correlations**

When identifying a relationship between two variables, there is always a possibility that this correlation might be a coincidence (data randomly generated), a p-value calculation helps determine if the observed relationship could arise as a result of chance. Common practice is to set a p-value of 0.05, this value is called a threshold. If the p-value is less than the threshold that means that the relationship is significant and therefore not produced by coincidence, if it was higher than the threshold, however, the relationship is insignificant, produced by coincidence. From Table 7, we can interpret that the data is indeed randomly generated since the p-value for all continuous columns is greater than 0.05.

We decided to use Pearson’s correlation matrix to inspect the correlations and the results showed that almost all feature correlations were negligible, we found no significant correlation with our target column, Price, however, we managed to find low correlations between Quantity and Online, and Quantity and Promotion as well as Online and employee\_tenure, which helped us in analyzing and understanding the dataset more. Quantity and Promotion had a low positive correlation of 0.18 which can be interpreted as customers tend to buy more products if there are promotions. Quantity and Online had a low negative correlation of 0.13 which we can interpret as items Online sell less, and finally, employee\_tenure and Online had a low positive correlation of 0.32 which suggests that employees with higher experience sell online the best.

**3.3 Data Visuals**

**3.3.1 Box Plots**

Box plots were used in our investigations to visually relay the central tendencies of any given numerical variable. When applied to the Price column in figure 1 it was clear that the variable is near normal distributions and no outliers were initially detected. When applied to the quantity variable in figure 2, however, the distribution was more positively skewed. Plots of city population and city areas in figure 9 showed severe skewness and revealed the existence of outliers, though, the existence of these data points was attributed to Amman which had the biggest population, and Ma’an which had the biggest area.

**3.3.2 Pie Charts**

Pie charts were employed in our investigation to visualize the percentage of sales a variable had contributed. Our pie charts in figure 7 showed that 91.5% of revenue was recorded in 2021 and only 8.5% in 2022. This is justified by the incompleteness of the data and imbalance of records taken from those years, not particularly showing that 2021 was a better year for business. It came as no surprise that promotions in figure 8 contributed to 88.9% of all revenue, with giveaways being the most popular and being responsible for 17.4% of the overall revenue. This goes to show that running promotions are a fantastic strategy for increasing sales and attracting customers.

**3.3.3 Bar Plots**

In our investigation, bar plots were heavily used to show sums and counts of different categorical variables. Our most prominent graphs include figure 3 which showed the total revenue collected from each promotion category and proved that giveaways bring in the largest overall sum with over 30000 in sales alone. We also used bar plots in figure 4 to measure the amount spent by male and female customers which showed an even split between the genders which led us to eliminate gender as a deciding factor when gauging customer spending. A bar plot of the amount of money spent split on age and gender in figure 5 clarified that adults (20- 59) in general spent more money in stores than teenagers (11-20) and seniors (59+) which can be due to them having jobs and a stable income stream. In addition, we investigated how holidays could affect purchasing tendencies across different departments in figure 6 which showed that consumers bought more activewear and gifts during holidays; it aligns with our expectations.

**4. Limitations and recommendations**

The results of this study aim to highlight the potential factors that could have affected the sales revenue in the chain store in Jordan. However, the study does have limitations that should be mentioned. The revenue dates have ranged from January 2021 to February 2022. This might have impacted the sales revenue as it is not reliable to depend on only two years of ordered dates. This is called sample size bias and imbalance. This will increase the margin of error.

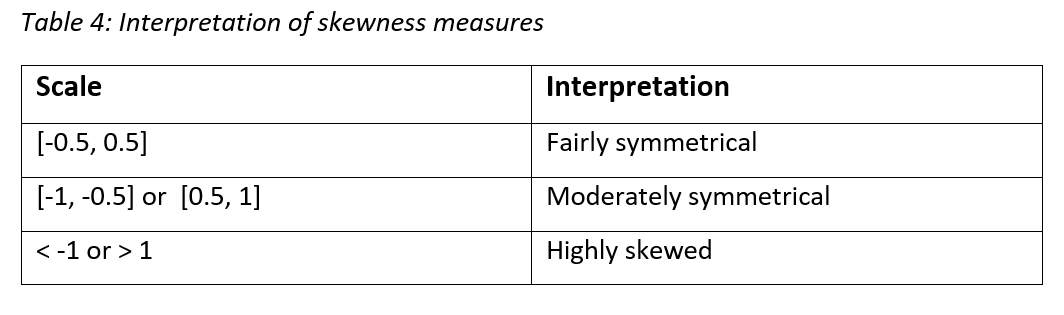
To improve the research, older ordered dates should be collected to confirm the effects of the external factors on the sales revenue. Furthermore, the stock count of the items might help to decide on the liability that each department has an equal number of items; this will reduce bias. Also, customer satisfaction surveys that can be collected in numerical forms can lead to further insights into the service quality in departments. Service quality can be determined from the survey; concluding if it affected sales. According to PWC, “73% of all people point to customer experience as an important factor when it comes to their purchasing decisions”. Therefore, it is recommended to add more dates, customer experience data, and stock count to the dataset provided to test if these factors are plausible to explain the sales revenue.

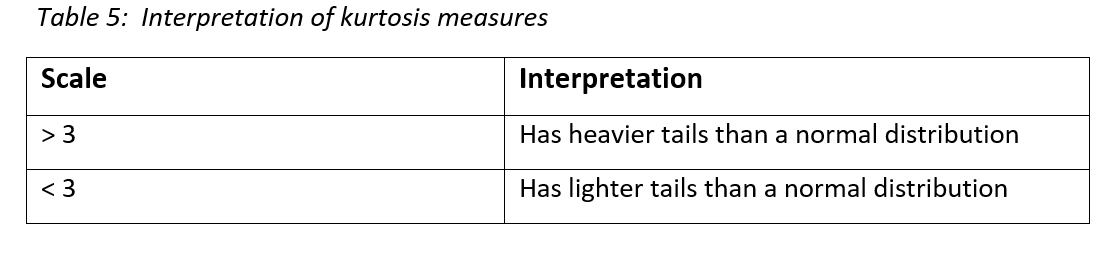
**References:**

*PWC: Audit and Assurance, consulting, and Tax Services*. (n.d.). Retrieved August 14, 2022, from https://www.pwc.com/us/en/advisory-services/publications/consumer-intelligence-series/pwc-consumer-intelligence-series-customer-experience.pdf

**5. Appendix**

**5.1 Tables**

****

****

**5.2 Figures**

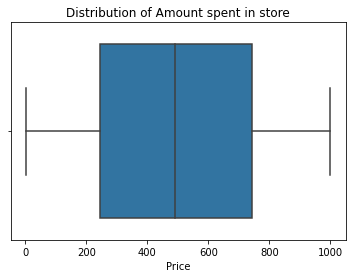


Fig.1 Boxplot of Price column

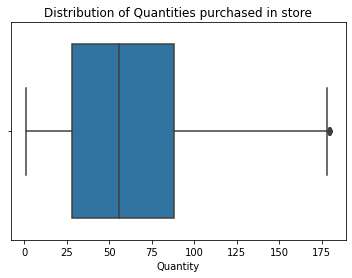
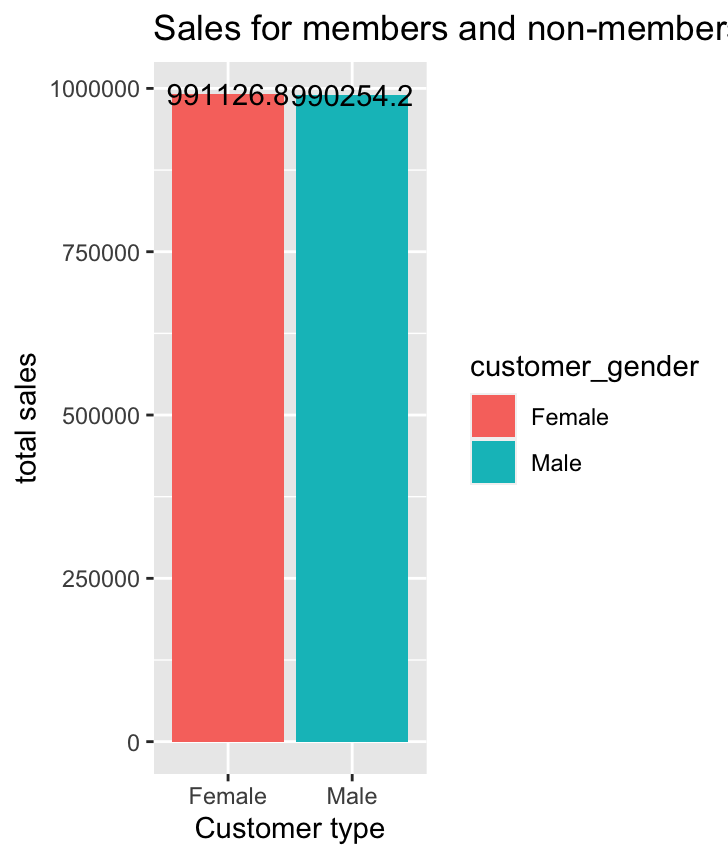


Fig.2 Boxplot of Quantity column



Fig.3 Sum of sales with each promotion type

Fig.4 Total sales among customer genders

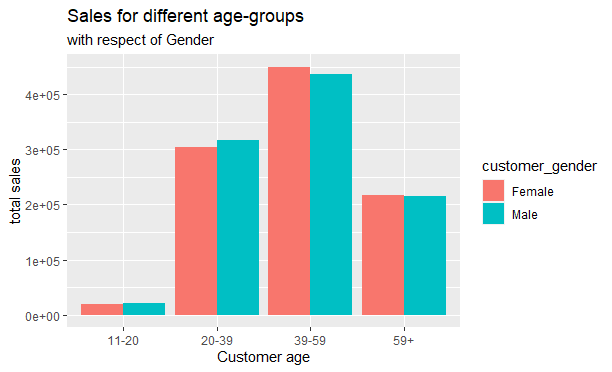


Fig.5 Sales among customer ages and gender

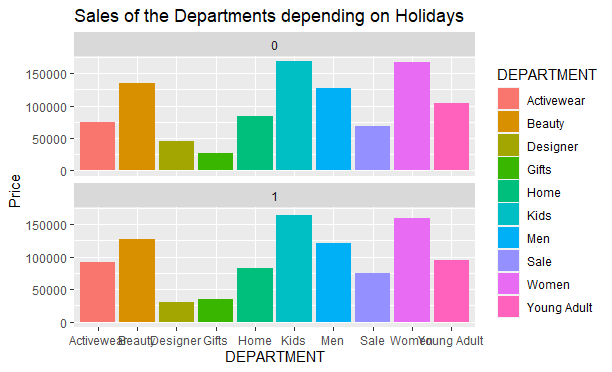


Fig.6 Sales per department depending on holidays

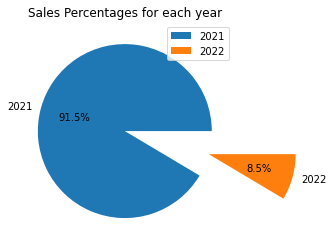


Fig.7 Percentages of Sales per year



Fig 8. Percentages of sales per promotion type



Fig.9 Boxplots of numerical variables

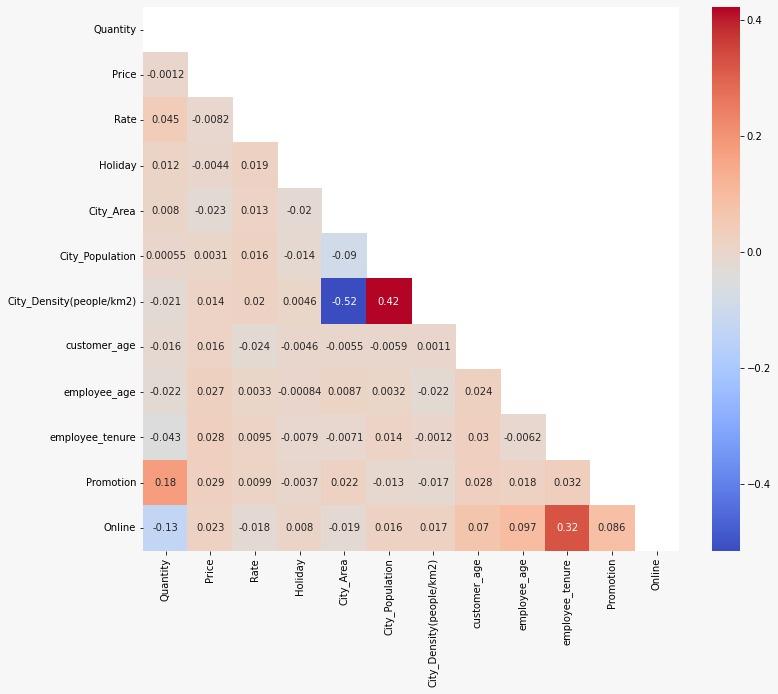
******

Fig. 10 Correlation heatmap

hf

d