**Online Grocery Shopping Orders Data Analysis**

by Beh Jie Shen (a1834032)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Abstract**

More than 3 million official Instacart orders were statistically analyzed, leading to two data-driven recommendations to boost the multi-billion grocery shopping industry.

**Introduction**

Grocery shopping is an activity that is familiar to most people regardless of their profession. With the rise of the Internet, online grocery shopping has become prevalent in our society. Australia’s revenue for the grocery industry in 2019 was worth more than AUD109 billion (Hinton, 2021).

The online grocery shopping experience provides a seamless and convenient unique proposition for their users. A user is only required to browse through the various grocery selections from the comfort of their homes and clicking the checkout button.

**Motivation**

As a mature industry that possesses multitudes of collected data across the years, the research was motivated to unlock hidden insights, provide data-driven business recommendations and boost revenue. In order to achieve this, the focus area of the rest of this research was to analyze the grocery reorders of past order data to improve the reorder rates among users.

**Methodologies**

**I. Dataset**

The dataset utilized in this research originated from Instacart and was compiled by Kaggle, a popular data science platform (Instacart, 2017). Instacart is a USD39 billion online grocery shopping app primarily in the United States of America and Canada (Beale, 2021).

The Instacart dataset has approximately 3 million official Instacart orders data made in 2017, amounting to more than 4GB of data. In the dataset, there were five different comma-separated values (CSV) files that are aggregated by their respective id. There was a total of 15 data columns after combining them.

For each of the CSV files, data nullity checks were performed. There were more than 206,209 missing values in the orders data at the days\_since\_prior\_order column. For the other files, no null values were found. Nonetheless, these null values were disregarded as some orders were not classified as reorders, and hence the null values were justified.

**II. Technologies Used**

The research was conducted entirely in a Jupyter Notebook 1.0.0 environment that was powered by Python 3.8.8. Additional data science and analysis tools used include Pandas 1.2.4, NumPy 1.20.1, Seaborn 0.11.1, Matplotlib 3.3.4, SciPy 1.6.2, Plotly 5.5.2 and Statsmodels 0.12.2.

All visualizations shown can be found in the reproducible GitHub repository link in the references section. Labels such as Code 1.1 provide quick references in the Jupyter Notebook code.

**Initial Visualizations**

Preliminary exploratory data analysis was performed to exhibit initial initializations.

**I. When do people reorder products on a typical day?**

We utilized Pandas to calculate the mean of reorders for each hour and day and created a pivot table (Code 1.1.1). The visualization was made using a heatmap (Code 1.1.2).

**A picture containing background pattern

Description automatically generated**

Figure 1.1: A heatmap of the reorder ratio against the order day and hour (Code 1.1.2).

According to Figure 1.1, most people reorder their groceries in the early mornings. The highest reorder ratio was at 7 a.m. on a Sunday morning. Most people do not reorder groceries after midnight.

**II. How does the cart position correlate with reorders?**

A copy function from Pandas was used to copy a column called add\_to\_cart\_order (Code 1.2.1). After a basic visualization to visualize the frequency of reorders against cart position, it was evident that the reorders above the 60th cart position was small. Therefore filtration of cart position was performed until the 60th position (Code 1.2.2). After computing the ratio of reorders against the total orders for each cart position, the reorder ratio was obtained and visualized (Code 1.2.3).

Chart, line chart

Description automatically generated

Figure 1.2: A line chart that shows the relationship between reorder ratio and cart position (Code 1.2.3).

According to Figure 1.2, there was a strong negative correlation of reorder ratio as cart position increases, which suggested that the first cart position has the highest reorder ratio. A Pearson correlation coefficient of -0.8718 was further computed to support the observed negative correlation. (Code 1.2.4)

**III. How many days do people wait before reordering?**

By using Pandas to filter reorders, a stacked bar plot was visualized (Code 1.3.1)

Chart, bar chart, histogram

Description automatically generated

Figure 1.3: A stacked bar plot to show the orders and reorders against days since prior order (Code 1.3.1).

According to Figure 1.3, most people reorder after approximately seven days or 30 days, suggesting that most people have a weekly or monthly grocery shopping habit.

**Experiments & Discussions**

**Claim I. Most people do not reorder on Monday or Tuesday.**

In the dataset, Saturdays were encoded as 0, while Sundays were encoded as 1. This encoding pattern was utilized across all days that are given in a week.

Pandas, a data manipulation software, was used to obtain the reorders and their corresponding order day (Code 2.1.1). The current population mean for reorders was 2.7252 days, which was between Mondays and Tuesdays (Code 2.1.2). The first research claim, motivated by Figure 1.1, was that most people do not reorder on Mondays and Tuesdays because users now have a grocery shopping application that is available at every hour and any day throughout the week.

H0: Most reorders occur on Mondays and Tuesdays (µ 2.7252).

H1: Most reorders do not occur on Mondays and Tuesdays (µ 2.7252).

A sample of 500 random reorders were taken for the hypothesis testing (Code 2.1.3). Statistical z-test analysis was performed with a test statistic of -2.7039 (Code 2.1.4). Therefore, we rejected the null hypothesis as -2.7039 < -1.96 at a 5% significant level. There was sufficient evidence to suggest that most reorders do not occur on Mondays and Tuesdays.

**Claim II. The mean cart position for reorders is more than 7.5731.**

Pandas was utilized to extract the reorders with their cart position (Code 2.2.1). The current population mean for cart position is 7.5731 (Code 2.2.2). The second research claim, motivated by Figure 1.2, was that there should be more reordered groceries because most households have an abundance of regularly used products, such as milk, shampoos and detergents.

H0: The mean cart position for reorders is 7.5731 (µ = 7.5731).

H1: The mean cart position for reorders is more than 7.5731 (µ > 7.5731).

For hypothesis testing, a random sample of 500 was obtained using NumPy (Code 2.2.3). Another statistical z-test was performed, and the obtained test statistic was -1.250 (Code 2.2.4). Therefore, we do not reject the null hypothesis as -1.250 > - 1.96 at a 5% significant level. There was insufficient evidence to suggest that the mean cart position for reorders is more than 7.5731.

**Claim III. The mean days before most people reorder is more than 10.4386.**

We utilized pandas to obtain the days\_since\_prior\_order column for each reorder (Code 2.3.1). The current population mean days before more people reorder again was 10.4386 (Code 2.3.2). The third research claim, motivated by Figure 1.3, is that most people reorder their groceries weekly. Therefore, it should be lesser than the population mean.

H0: The mean days before most people reorder is 10.4386 (µ = 10.4386).

H1: The mean days before most people reorder is more than 10.4386 (µ > 10.4386).

We used a random sample of 500 reorders chosen by NumPy for hypothesis testing (Code 2.3.3). A test statistic of 0.7449 was derived from a z-test (Code 2.3.4). Therefore, we do not reject the null hypothesis as 0.7449 < 1.96 at a 5% significant level. There was insufficient evidence to suggest that the mean days before most people reorder is more than 10.4386.

**Recommendations**

Based on the results obtained above, Instacart can take specific actions to boost its revenues. Firstly, Instacart has to not only increase their delivery riders during the peak timing as according to Figure 1.1., but to also ensure that there are substantial delivery riders to cater for all days as according to Claim I. Secondly, Instacart can try to introduce promotions for products to increase reorders. For example, after 21 days of reordering a specific product, a promotional discount can be given to a customer to encourage reordering.

**Conclusions**

In order to boost the revenue of Instacart, several data-driven recommendations were provided after data analysis. However, there is a need for more visualizations and research claims to be analyzed, as most of the analysis made in this report is related to reorders.

**References**

All code references can be found at

Hinton, T 2021, Customer penetration in the grocery market in Australia in 2019, by store, viewed 3 September 2021, <https://www.statista.com/statistics/1106566/australia-customer-penetration-in-the-grocery-market-by-store/>.‌

Beale, M 2021, Instacart Doubles Valuation With $265 Million Funding Round, viewed 18 August 2021, <https://www.forbes.com/sites/margheritabeale/2021/03/02/instacart-doubles-valuation-with-265-million-funding-round/>.

Instacart, 2017, Instacart Market Basket Analysis, viewed 1 September 2021, <https://www.kaggle.com/c/instacart-market-basket-analysis/overview/description/>.