Project: Capstone Project 2: Milestone Report

Customer Purchasing Behavior

During the course of our lives most of us one point or another have had to buy at least one product from Walmart. However, for the lucky few who haven't heard of Walmart; Walmart is the poster child for multi-national big-box retail store that stock and sell over 75 million different products, items ranging from paint to plushies, from guns to guacamole. With over 11,300 stores globally, employing 2.2 million people, and with 275 million weekly customers, Walmart is easily one of the largest franchises in the world. Walmart's fiscal revenue is something to the tune of $514 billion per year making it the most lucrative company in the world. With nearly every Walmart store selling grocery items and produce, what many may not have considered is that Walmart is also one of the world's largest U.S. grocery retailers.

https://en.wikipedia.org/wiki/Walmart

https://expandedramblings.com/index.php/walmart-statistics/

Problem Statement:

Customer purchasing optimization is a common problem for a lot of companies, both large and small. Luckily most large companies should have a wealth of data available to them in order to predict the possible purchasing behavior for many of their clients so they can make informed decisions about stocking their shelves, providing customer recommendations on items, customer promotions, and predicting when a customer is losing interest in their service and planning to buy from a competitor. My proposal is to analyse and demonstrate some of these metrics on public data found on the internet as well as provide an algorithm that can help to predict future long-term purchasing in the form of Customer Lifetime Values.

Dataset:

Inspired by the paper, Customer Shopping Pattern Prediction: A Recurrent Neural Network Approach by Hojjat Salehinejad and Shahryar Rahnamayan, the authors used a Recurrent Neural Network to predict customer loyalty values (R,F,M) using the **Ta Feng Grocery Dataset**. After doing some internet searching I managed to find a copy or sample of the Ta Feng Grocery dataset hosted on Kaggle but not part of a competition. The Ta Feng Dataset is a Supermarket Dataset containing 817741 transactions from November 2000 until the end of February 2001. The dataset contains information about 119578 shopping baskets, belonging to 32266 users, where 1129939 items were purchased from a range of 23812 products.

Summary:

Exploratory Data Analysis

After exploring and investigating each column of the dataset independently I find explore and find out some valuable insights about 1 or two of those columns; namely, discovering the unexpected behavior regarding the 'Amount' and 'Sales Price' columns and their effect on one another. I found that these two important columns have non-normally distributed data and seem to have an unclear relationship. After investigating that relationship I surmise that the 'Sales Price' column is probably the total transaction price as opposed to the price per unit. Using this assumption would explain the relationship between

'Amount' and 'Sales Price'. After further exploring the data by performing some statistical tests on the data I provide some valuable business metrics based on cohort analysis.

Cohort Analysis

I first separate the cohorts into weekly intervals by the first week that the customers purchased an item. After which I provide tables based on cohort counts, such as customer retention rate, average number of items purchase per cohort, the average transaction cost per cohort, and the average unit price (*see below*) per cohort. After doing some basic cohort analysis I then start to classify customer ids based on their Recency, Freqency, and Monetary values. In doing so, I make it easier to identify and classify customers whom are most important to target with special promotions and offers.

Customer Segmentation

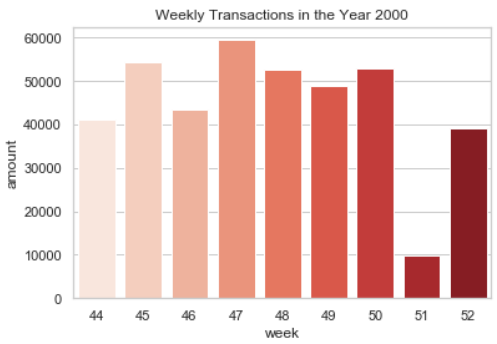
The first form of customer segmentation I employ is based on percentiles (*quartiles*) of RFM purchasing behavior. I further make these RFM findings easier to understand by providing an 'RFM Score' which shows the relative customer value of each customer id. This is further simplified by classifying customer ids into groups such as gold, silver, and bronze level shoppers. Finally, I finish off my customer segmentation by using K-means to automatically classify similar behavior customers which is a convenient method for simplicity and unique ability to work with especially large datasets.

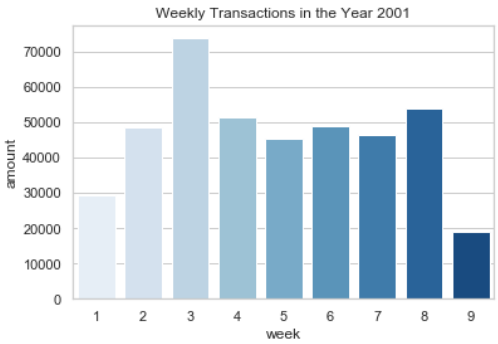
Data Wrangling:

The data was separated into four separate datasets each representing a one month of data. For the most part the data was in a fairly clean format, the data was semi-colon separated and the first row of the dataset had garbage in it. Likely the first row was column names written in Chinese characters but obviously jupyter notebook and notepad couldn't make that distinction. I had to remove a number of white-spaces within the dataset in order to get everything to align correctly, but that was pretty easy to do, after the data was clean enough I was able to merge the data into one larger dataframe so that I could do some more work on the entire set of data. The first thing I did was I created some dictionaries with label mappings so that I could change the format of the 'age_group' column and 'pin_code' column easily without spending too much time on it. I also created the inverse of those dictionaries in order to reverse those changes easily. Lastly, I created some operations to be able to encode dummy values on columns of my choosing, followed by changing as many of the columns into integer values as I could; since I knew that I would need to use the data for modeling or summarizing.

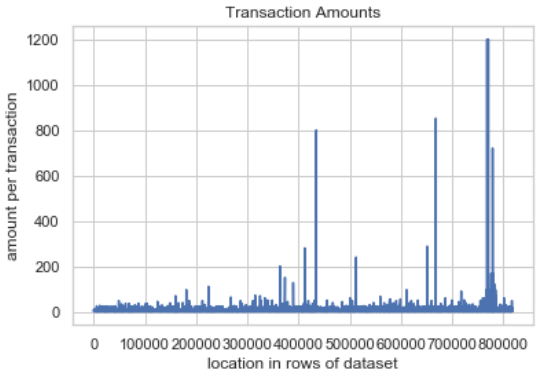
Exploratory Analysis:

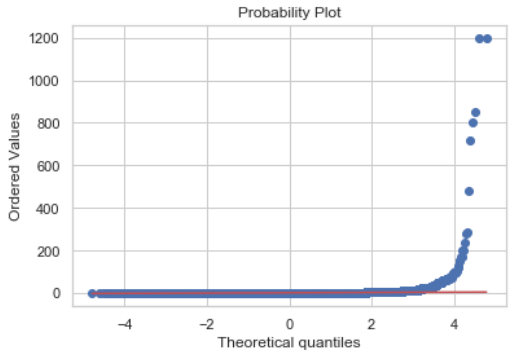
Once I had my data in a format that I was comfortable with I went about exploring the data itself. I started by graphing the columns and looking at their column statistics (mean, count, max, and percentiles) I created graphs where I could to help visualize the data.

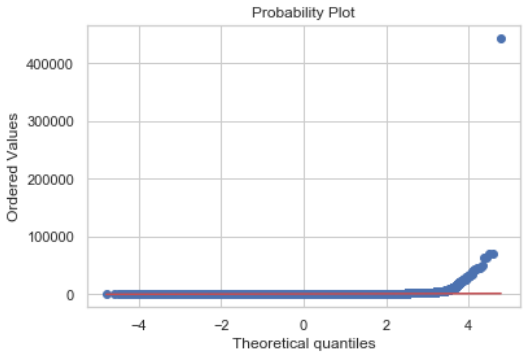




After looking at the differences in transaction counts by year I tested the hypothesis that perhaps the means of each year could be different from each other and found the means to not be statistically different from each other. I further looked at the distributions of 'Amount' and 'Sales Price' and found they had particularly odd behavior where there were dramatic spikes in activity for both purchasing 'Amounts' and 'Sales Prices'.



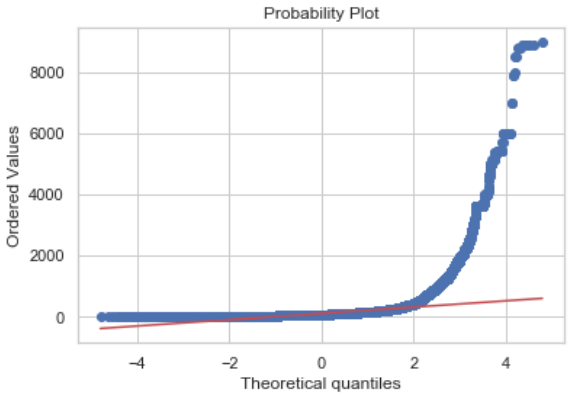




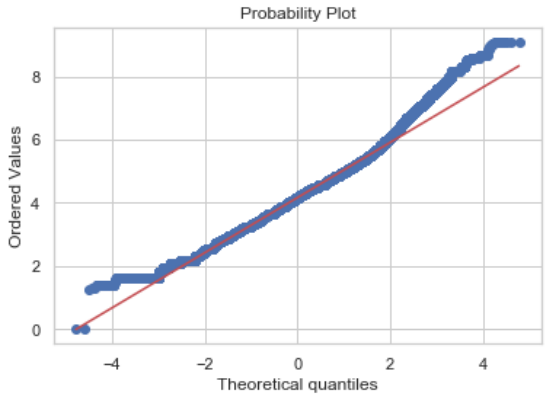
After further exploring these behaviors I found that there appeared to be a non-linear relationship with the two variables and decided that perhaps 'Sales Price' reflects the total transaction cost and not the individual 'Unit Price' that I initially believed.

I tested both the 'Amount' variable and the 'Sales Price' variable and found neither variable were normally distributed; thereby, making it prohibitive to do a statistical test on their relationship. I finally finished off with dropping the extreme values from the data that I felt were unnecessarily large and created a new variable in the dataframe 'Unit Price' that I felt most likely reflected the true cost of the items in the dataset.

Below: Unit Price



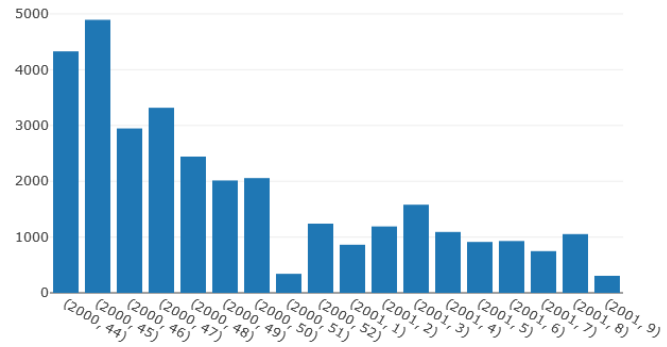
Log Transformed Unit Price

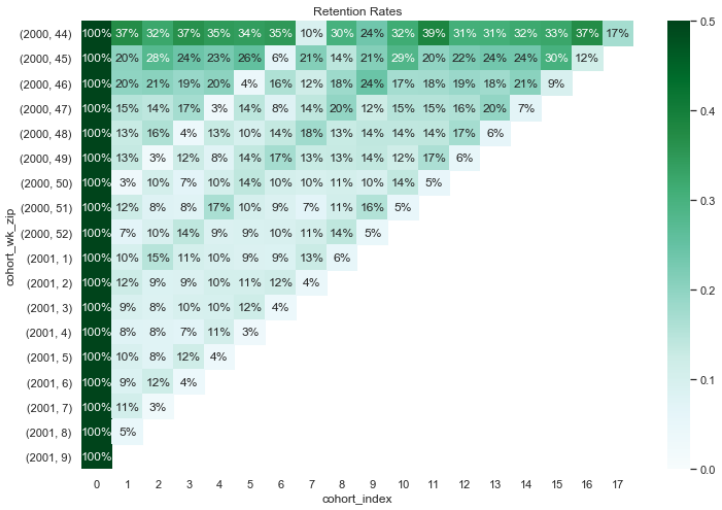


Metrics

With my data exploration complete to my satisfaction my next goal was to help provide some valuable metrics about the customer purchasing behaviors in order to provide promotions and identify customers whom would be most valuable to the business or organization. In order to do so I grouped customer ids into cohorts of customers who made purchases around the same period of time and try to provide some valuable metrics for predicting purchasing behavior among those cohorts. Typically time cohorts would be separated by a yearly, monthly, or even daily basis for online customer groups. However, the nature of the data that I am working with doesn't lend itself to these former options very well because it is too short a time to separate the cohorts into yearly and monthly cohorts, and too long a time to separate into daily cohorts. So I chose to use a weekly length of time. Doing so added a certain amount of complication since a lot of time operations work well on the former time periods rather than the latter. I also chose to identify cohorts by the week that they enrolled from the start of the study to the end of the study (i.e. week 0 is November 1, 2000 and week 17 is February).

Below is a plot representing counts of customers for their first week purchasing an item. The X-axis is each year and week within the year, so (2000, 52) would be the last week of the year 2000.





Once I had counts of cohorts I was able to create a retention table, or the counts of active customers whom purchased from one week to the next. I also created similar heat-maps for average transaction price as well as unit price. I found that typically customer cohorts bought anywhere from 1 to 1.5 items with the average cost being about 100 – 150 yuan. After creating the cohort groups the RFM metrics were my primary focal point for the project.

Many behavioral customer segmentation is based on these three foundation metrics:

1. Recency (R)
   - How many days since customer's last purchase (*the lower the better*) until the present
2. Frequency (F)
   - How many purchases the customer has done since their start of the time period
3. Monetary Value (M)
   - Measures how much the customer has spent since the start of the time period

With the RFM metrics calculated I was able to use the quantiles of each of these values and simply add them up in order to create their *RFM score*. The RFM score is a simplified way of segmenting customers based on their behavior in order to decide which customers are most valuable to the business. I further simplify this segmentation by classifying customers into only three metrics, gold, silver, and bronze level customers. In this way decision makers can draw conclusions about customers without having to get into the numbers. Finally, rather than using percentiles and quantiles to separate customers into recency, frequency, and monetary values I used K-means to separate them into semi-unsupervised groups. This has the added bonus of being easier to execute and letting the data decide the groups of customers. The downside of using this method is not specifically choosing the groups based on a simple metric that can be easily explained, and can be a bit more difficult to describe if needing to provide the solution to business partners.