Detailed Analysis on Personal Transactions

Haochong(Rogers) YANG

13/07/2022

INTRODUCTION

Using data retrieved from Scotiabank, I've put together an analysis report for my transactions from August 21, 2021 to July 7, 2022. I used wrangling skills when processing the data and used several analyzing skills while looking at the patterns of data. This project not only gave me a chance to practice and learn more new R programming skills, but also helped me keep track of my financial condition. Thanks for your time:)

First, let's get a quick glimpse into the data set converted from the csv file provided by Scotiabank.

```
## Rows: 663
## Columns: 5
## $ `7/12/2022`              <chr> "7/12/2022", "7/11/2022", "7/11/2022", "7/9/~
## $ `-35`                    <dbl> 200.00, -67.00, -18.00, -60.03, 41.00, 41.51~
## $ `-`                      <chr> "-", "-", "-", "-", "-", "-", "-", "-", "-",~
## $ WITHDRAWAL               <chr> "Customer Transfer Cr.", "WITHDRAWAL", "WITH~
## $ `FREE INTERAC E-TRANSFER` <chr> "MB-TRANSFER", "FREE INTERAC E-TRANSFER", "F~
```

DATA PROCESSING

After taking a glimpse into the raw data, we can find out that the raw data is missing the header, i.e. the titles for each columns, and there is also a meaningless column that only serves as the separator in the raw data.

So first we should get rid of the column '-' and add a header to the data set.

```
##        Date  Amount       Method                            Detail
## 1 7/11/2022  -67.00   WITHDRAWAL          FREE INTERAC E-TRANSFER
## 2 7/11/2022  -18.00   WITHDRAWAL          FREE INTERAC E-TRANSFER
## 3  7/9/2022  -60.03 POS Purchase   APOS GYUBEE JAPANESE     TORON
## 4  7/9/2022   41.00      DEPOSIT          FREE INTERAC E-TRANSFER
## 5  7/9/2022   41.51      DEPOSIT          FREE INTERAC E-TRANSFER
## 6  7/9/2022 -124.52 Bill Payment TS-TORONTO HYDRO-ELECTRIC SYSTEM
```

Now we can see that the columns names were added and the useless column "-" is removed. To make it easier to examine, let's turn all the values in the data frame lowercase and reduce all the extra spaces. Then I format the date from "DD/MM/YYYY" into "YYYYMMDD" and rearranged date from least recent to most recent.

```
##         Date  Amount                Method                          Detail
## 1 2021-08-21 2000.00 customer transfer cr.           investment proceeds
```

```
## 2 2021-08-23   -6.77          pos purchase          opos uber* eats uber.
## 3 2021-08-23  -64.80          pos purchase  opos uber* eats pending uber.
## 4 2021-08-23   -0.95          pos purchase fpos kitchen food fair#28bramp
## 5 2021-08-23  -20.44          pos purchase fpos kitchen food fair#28bramp
## 6 2021-08-23  -70.69          pos purchase           opos instacart 88824
```

Next is to filter out transactions via my credit card, since the payments were covered by my savings account on debit card.

Now the data frame looks perfect! We can start analyzing the data! My first step is to categorize the transactions, namely into "transport", "food", "grocery", "e-transfer", "entertainment", "shopping", and "billing".

```
##         Date  Amount               Method                          Detail
## 1 2021-08-21 2000.00 customer transfer cr.            investment proceeds
## 2 2021-08-23   -6.77          pos purchase          opos uber* eats uber.
## 3 2021-08-23  -64.80          pos purchase  opos uber* eats pending uber.
## 4 2021-08-23   -0.95          pos purchase fpos kitchen food fair#28bramp
## 5 2021-08-23  -20.44          pos purchase fpos kitchen food fair#28bramp
## 6 2021-08-23  -70.69          pos purchase           opos instacart 88824
##   Category
## 1    other
## 2     food
## 3     food
## 4     food
## 5     food
## 6     food
```

DATA ANALYSIS

After sorting all the data with respect to the category, I summarized the total amount and total number of transactions in each category, which is shown in the following chart.

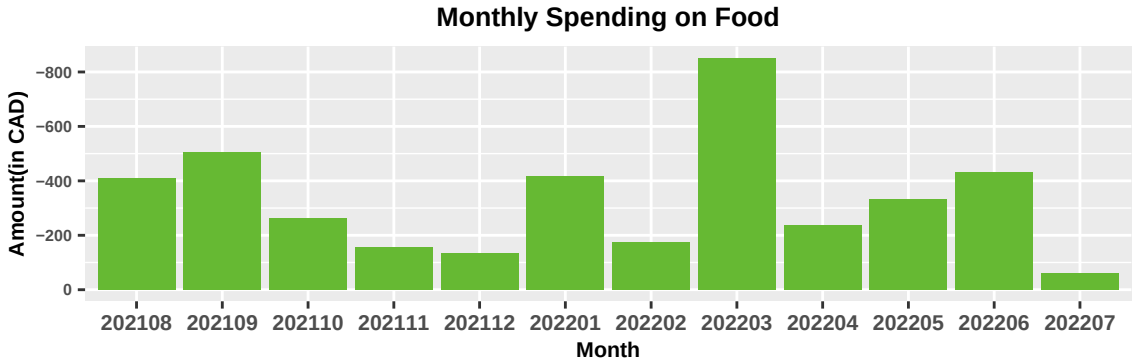
```
## # A tibble: 9 x 3
##   Category      Total_number Total_Amount
##   <chr>                <int>        <dbl>
## 1 billing                  9       -8965.
## 2 e-transfer             143       -8934.
## 3 entertainment           10       -1171.
## 4 food                   125       -3972.
## 5 grocery                 36       -1276.
## 6 other                   86       30453.
## 7 salary                   3        2686.
## 8 shopping                33       -4569.
## 9 transport              172       -1933.
```

Now let's sort all the transactions into months to see in which month I spent the most and in which I spent the least.

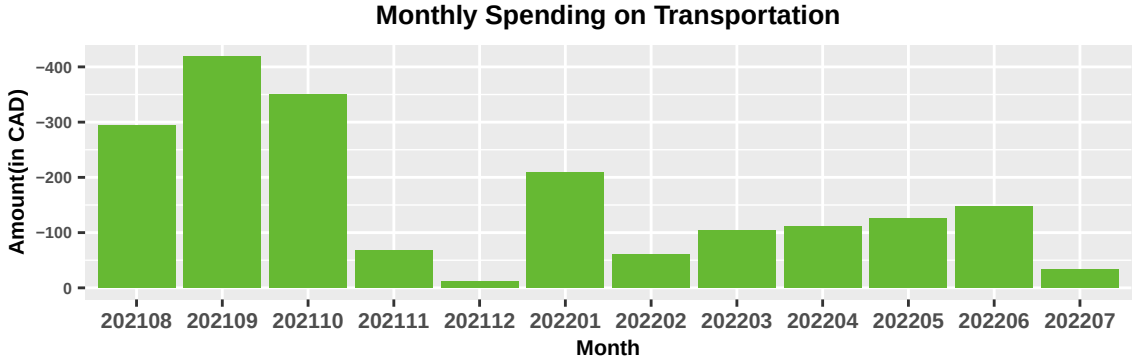
```
## # A tibble: 12 x 3
##    Year   Month Monthly_Spending
##    <chr>  <chr>            <dbl>
```

```
##  1 202108 08             1676
##  2 202109 09            -2153.
##  3 202110 10             -498.
##  4 202111 11               18.5
##  5 202112 12             383.
##  6 202201 01             290.
##  7 202202 02             -166.
##  8 202203 03             5279.
##  9 202204 04            -3831.
## 10 202205 05            -1146.
## 11 202206 06             8274.
## 12 202207 07            -5808.
```

From the monthly summary we can see that I spent the most in July 2022, because I transferred all my rent to my landlord. In March and June of 2022, I received significantly amount of money compared with my spending. This is because I received money for house renting from my dad and also received my salary from the company I'm currently doing my internship in. However, this chart is extremely misleading because the amount of rent is too large compared with my spending. So in order to get a better look at my spending pattern, I need to filter out my rent first.

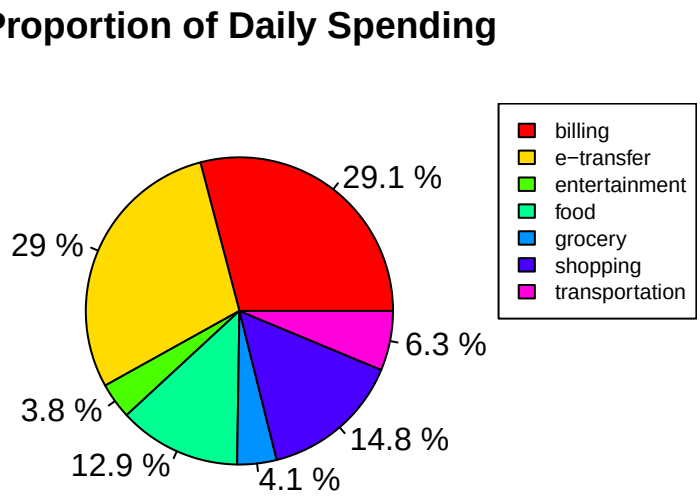


As can be seen from the graph, the distribution of my spending on food is bimodal, with two peaks at September 2021 and June 2022, which are two months that burden on school work is not heavy. But there are two exceptions, including January and March of 2022. I spent a lot on food in January 2022 because it was the time my first semester in university ended, so I spent more money beyond my budget. I also spent money "abnormally" in March 2022, which is because I brought my best friend who came to Toronto to visit me to some fancy restaurants, so I spent the most monthly speaking.



My monthly spending on transportation varied throughout the year. I spent average about 350 dollars per month on transportation from August 2021 to October 2021, because that was the time I'm exploring Toronto. As approaching to midterms and finals, I don't have enough time to go out, so my spending on transportation dropped accordingly. Later as I got used to the balance between study and living, I have time to go out more often so my spending on transportation gradually increased in the second semester, which is from February 2022 to May 2022.

Looking at the proportions of my daily spending is also important, since I can know my expenditure structure.

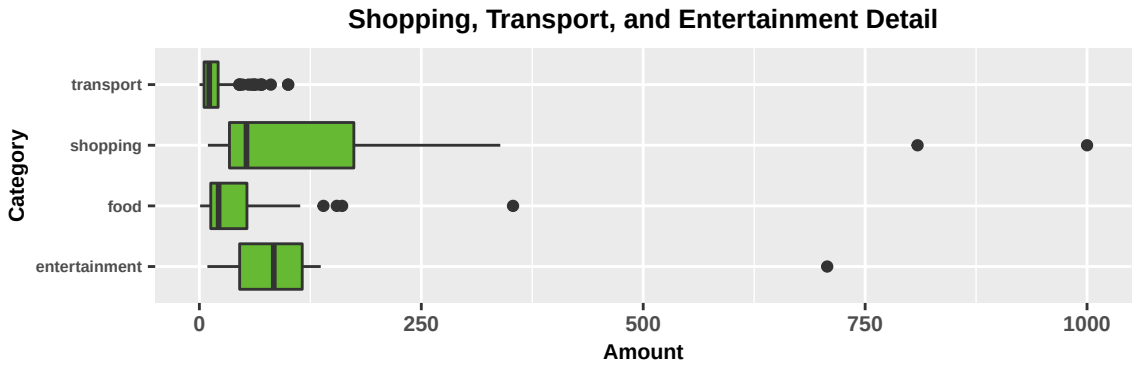


From the pie chart above I can directly tell that the largest portion of my transaction is on "billing", which included my rent, hydro, and internet fees. Billing took approximately 29% of my overall transaction. The smallest portion is entertainment, which only took about 3.8% of my total amount of transaction. This is because I barely go to play or relax after school started. The surprising part is the transportation. I actually spent more money on transit than grocery. This might because I took a lot of Uber trips in the city before I got used to the public transportation system in Toronto.

Next, I want to take a look at my shopping transactions.

```
##                          Amount(in CAD)
## Number_of_Transaction                33
## Total_Shopping_Amount            -4568.90
## Mean_Shopping_Amount              -138.45
## Median_Shopping_Amount             -50.29
## Largest_Transaction              -1000.00
```

In the summary table we can see that I've made 34 transactions in shopping. The total amount of shopping transaction is 4704.5 CAD and the largest transaction is 1000 CAD (because it is my debit card daily limit). The average amount in each transaction is 138.37 CAD, and the median amount is 51.69 CAD.



From the box plot above I can see that three types of transactions are all right-skewed. This shows that most of the transactions are relative small, concentrating on the interval of amount less than 100 CAD per transaction.

Now let's find out which specific kind of transaction I have made the most in the past year. As we can see from the following chart, the most transfer I've made is e-transfer to my friends, which might be caused by spliting the bill or other kind of cases. The second most number of transfer is Uber trip and the third is Uber eats. This fall within my expectation because these spendings are really common in my daily life.

```
## # A tibble: 10 x 2
##    Detail                          number_of_transaction
##    <chr>                                           <int>
##  1 free interac e-transfer                           142
##  2 opos uber canada toron                            126
##  3 opos uber* eats pending uber.                      22
##  4 opos uber* pending uber.                           22
##  5 apos shoppers drug m toron                         15
##  6 scotiawire/mts fee                                 13
##  7 yang xing wen                                      12
##  8 investment proceeds                                11
##  9 opos uber* eats uber.                              10
## 10 opos instacart 88824                                9
```

What if I extend my exploration from the existing data to predicted data? This is how it does. I want to predict what range will my next transaction's price fell in, so I've made 10000 stimulations based on the existing data. The result is shown as follows:

```
##     Amount(in CAD)
## 5%       -39.06876
## 95%      -25.27323
```

The 5th percentile among all the 10000 values generated in the stimulation is -39.07, which is the largest value that is smaller or equal to 5% of all values. Similarly, the 95th percentile is -25.27.

So, the 90% confidence interval is from -39.07 stars to -25.27 stars on average. If I repeated this sampling procedure many times, 90% of the confidence intervals would include the true mean amount for the next transaction.

Can say: I am 90% confident that the mean amount for my next transaction based on my transaction history is between 39.07 CAD to 25.27, which is an interval of amount for a regular dinner in the restaurant.

Conclusion

The transactions in shopping appeared in my payment history varied much more than other category, which suggests that I should pay more attention on shopping, trying to avoid impolsive purchases on shopping. The transaction amount varied a lot between months in school and months in school breaks, which suggests that I should keep my purchasing habit during school time the same, and maybe try to control myself buying more stuff when school ends. These suggestions I concluded from my analysis on the transaction history will help me improve my daily spending pattern and hence help me control myself financially.