# Financial Data Analysis Visualization

Mayra Vazquez-Sanchez, Michael Mayaguari

**Abstract:** On a day-to-day basis, people spend money often without even realizing how much they're actually spending. This was our case. We wanted to analyze our spending as college students. Our goal was to identify the categories, time periods, and seasons in which we spent the most/ least money as two separate individuals, and as two commuter students attending the same school, in the same age range. Our project aims to answer the following questions: (1) How do the months/seasons affect our spending? (2) Which categories do we spend the most/least on? (3) How does gender affect our spending? While we were able to successfully answer 1 and 2, we later determined 3 wouldn't be accurate since we were only analyzing data from 2 people - this wouldn't be a fair representation for the population. In this report, we discuss our process from beginning to end, as well as our findings using visualizations created with Vega and Vega-lite.

## 1 INTRODUCTION

Bank statements and transaction CSV files only tell us one thing, that we're spending money. We sought to find out more information as to where exactly our money was going. With the help of Vega and Vega-lite, we created visualizations that answered not only our initial research questions but also provided us with additional insight into our findings. Analyzing our financial data by simply downloading our bank/transaction statements only provided us with minimal information. We were able to see the dates, amounts, and descriptions of a transaction, but spending patterns were not easily identifiable.

We chose to analyze our own financial data as opposed to finding a dataset online because we wanted this project to be more personal to us. We plan to later evolve this project so that users can import their own data to be analyzed. The process of how we created our own dataset can be found in the next section where we describe our entire procedure.

Not only did we want to analyze our individual spending habits, but we also wanted to take into account how our common factors affected our spending. Our initial research questions were the following:

1. How do the months/seasons affect our spending?
2. Which categories do we spend the most/least on?
3. How does gender affect our spending?

Apart from these questions, we also wanted to see how the fact that we were both college students around the same age affected our spending. We discuss our results towards the end of this report, as well as discuss some of the issues we faced while completing this project. These findings will help us understand our spending patterns, which can be taken into account to budget more effectively.

## 2 PROCEDURE

### 2.1 Data

We created our own dataset by first downloading our respective CSV transaction files from our bank accounts for 2019 and 2020. Since 2020 is not over yet, our data ranges from January to mid-April. After having done this, we cleaned our data to remove unwanted attributes, as well as add needed attributes. To refine our data and format it to meet Vega/Vega-lite criterias we used Python.[1] We imported the CSV files and converted them into data frames with the help of the Pandas library available by Python. This helped us work with our data and modify it as we needed easily. Next, we combined both data frames holding our individual data into a single one. We changed the labels to keep consistency and removed columns that were not being used. Lastly, we exported the data frame into a CSV. For the monthly spending, we were required to have a JSON file, so we had to modify the previously created dataset to calculate and keep track of the specific data we needed in a python dictionary. When the process was completed, we exported the dictionary into a JSON file. Some of the additional attributes we added to our data consisted of category, as well as season. The list of attributes is as follows: **account_number, transaction_date, transaction_amount, transaction_description, category, and season**.[2]

In terms of the categories in which we were to divide our transactions, we chose 9 significant ones. The following is the list of categories and how we defined them in our data-cleaning process:

1. **Education:** this consisted of any money spent on our tuition, school fees, textbooks, and any other academic sources.
2. **Entertainment:** this consisted of any money spent on activities such as going to the movie theater, bowling, etc., as well as subscriptions to music/movie streaming services.
3. **Food:** this consisted of any money spent on groceries, snacks, or outings to restaurants. This also included take-out/delivery food purchases.
4. **Health:** this consisted of any money spent on health services, as well as pharmacy-related purchases and gym memberships.
5. **Investments:** this consisted of any money dedicated to stocks or investing accounts.
6. **Shopping:** this consisted of any purchases made at retail-companies for clothing, shoes, etc.
7. **Transportation:** this consisted of any money spent on public transportation, gas money, and taxi/cab services.
8. **Travel:** this consisted of any money spent on air travel and hotel accomodations.

[1] For access to the code for this process check our GitHub repo at the end of this report. Code related to data can be found under the "Data" folder.

[2] Our finalized CSV files are labeled "/debit_combined_2019.csv", and "/debit_combined_2020.csv" for the respective years.

9. **Other:** lastly, this consisted of any money spent on miscellaneous items/services that don’t fit into the other categories.

For the season attribute, we divided our transactions based on the dates in which the seasons began. Additionally, we created JSON files for each year when one of our visualizations required it. More will be explained on this issue in section 2.3.[3] Ultimately, these JSON files defined each category as an attribute, as opposed to a data value like in the CSV files. This made rendering the data on the parallel coordinates plot visualization easier.

In the next section we will discuss how we decided on the visualizations used, as well as our initial ideas.

## 2.2 Visualizations

Upon cleaning our data, we decided on choosing Vega/Vega-lite as our visualization tool. In total we created four visualizations with Vega-lite, and one with Vega. We determined that we needed one visualization to effectively show our monthly spending patterns. For this we chose a line chart and parallel coordinate plots.

We wanted a visualization in which we could easily compare both of our monthly spendings. With the line chart we plotted the sum of all transactions that took place in each month for each person. This allowed us to easily compare both accounts, as well as identify the months in which we spent the most/least in.

We also decided to go with parallel coordinates plots to analyze our monthly spending in each individual category. We wanted to see if we could identify any clusters in our monthly spending data and this visualization could easily point this out. We thought about using a scatter plot for this as well, but certain outliers in our data would cause all the points to gather at the bottom, making everything seem like a big cluster. Another issue with using a scatter plot for this was encoding the marks (points) to the category attribute. Nine categories would mean there would be nine different colors, not to mention how we would differentiate between accounts -- all of this would have resulted in chart junk.

Next, we wanted visualizations which would allow us to compare how much we spent in each category, according to a selected time frame. Vega-lite had a perfect example for what we wanted in their seattle-weather visualization. They used an interactive scatter plot which allowed the user to select an interval, changing the adjacent bar chart. We concluded that this was our best option. The scatter plot shows all the transactions that took place in the span of one year. The interactive feature allows for the selection of a time interval, as well as the selection of data points to focus on. For our data, we used a grouped bar chart to easily compare both accounts and their spending according to the categories defined. Our visualization also provides the option to look at both accounts or individual accounts. For individual accounts, we used a classic bar chart as opposed to a grouped bar chart. With these two graphs, we could easily select a time frame, and compare the total amount spent in each category.

[3] Our finalized JSON files are labeled “/monthly_expenses_2019.json” and “/monthly_expenses_2020.json” according to their respective years.

Lastly, we wanted to analyze our seasonal spending. Although you can select a time interval in the previous visualizations, we wanted something easier to interpret. We decided on a radial plot that we could break down into the four seasons. We provided three different scenarios -- small multiples if you will -- as well as the option to analyze the seasonal spending for each category. This radial plot allowed us to demonstrate which season we spent the most/least in, as well as determine how much we spent in a certain category for each season. We placed radial plots for both accounts and individual accounts side-by-side for easy comparison.

Our visualizations are far from perfect. We faced a multitude of issues which caused us to change our plans and layouts for these visualizations. These issues will be discussed in the next section.

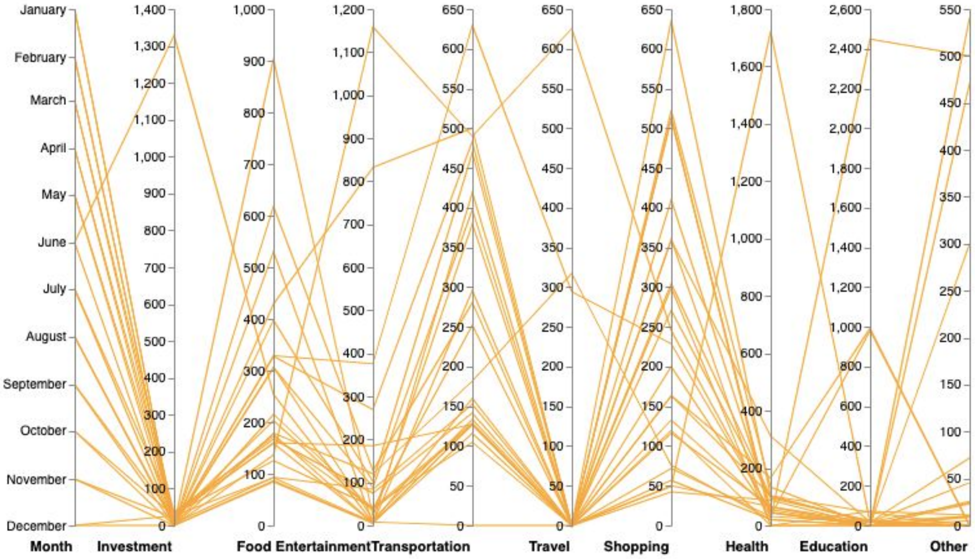


Fig. 1. Parallel coordinates plot using Vega - no color differentiation between accounts. This issue led us to resort to using a parallel coordinates plot for each account and place them side-by-side for easy comparison.

## 2.3 Issues

During the coding process we started facing a number of issues. Beginning with the parallel coordinates plot, we were unable to render this visualization through Vega-lite. We attempted to follow their Iris dataset example, but because one of our attributes was a ‘date’ datatype, Vega-lite was unable to properly label the y-axis. We looked to Vega to attempt to solve this issue. Vega used the cars dataset for the parallel coordinates example and as seen in Fig. 1, we were successfully able to label the months. Unfortunately, we were faced with another issue. We wanted 2 sets of different colored lines to compare monthly expenses for each account, but Vega was having trouble recognizing the two different accounts on one graph. We resorted to using a parallel coordinates plot for each account in our final version. This will be shown in the next section when discussing our results.

Another issue we faced had to deal with the radial plots. Initially, we wanted the radial plot to be a part of the scatter plot and bar chart interaction. Upon selecting a category in the bar chart, the radial plot would change accordingly. Due to limitations in the language and in time, we were unable to sync the radial plot to the other graphs. We opted to make the radial plot a standalone visualization, which in the end makes it easier to compare different scenarios. Our initial plan can be seen in Fig. 2.

One last issue faced can be seen in our final version of the radial plot. Due to scaling conflicts, when looking at the individual accounts' radial plots, the charts are incomplete. They're able to successfully display the correct data, but for some categories it is difficult to see certain sections of the pie because they appear really small.

Regardless, we found ways around these issues and attempted our best to still successfully present our data. In the following sections we present our final visualizations, as well as our results and their analysis.

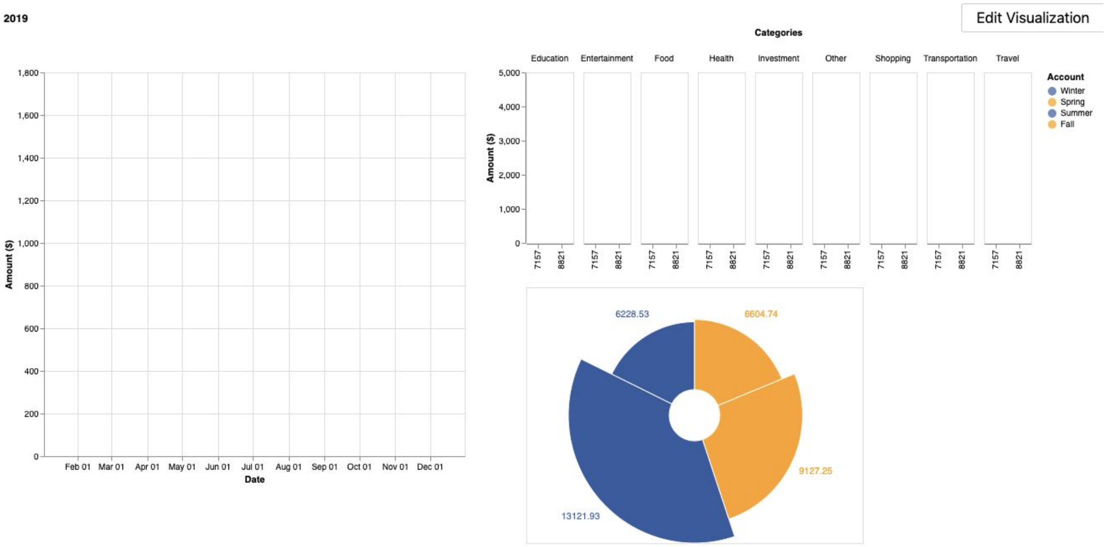


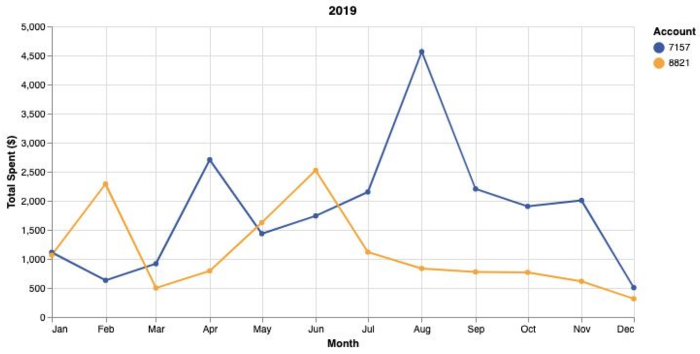
Fig. 2. Our original plan was to include the radial plot along with the scatter plot and bar chart. As you can see, all the graphs fail to display properly and there are merging conflicts in the legend. This led us to resort to excluding the radial plot from this layout.

## 3 RESULTS

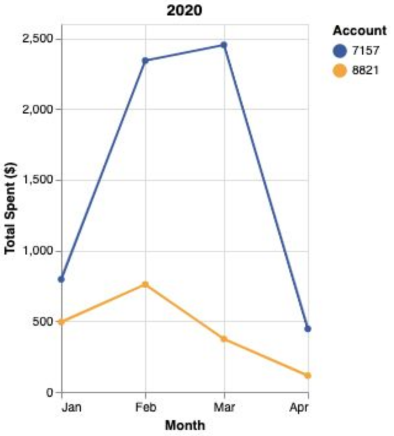
In this section we present our final visualizations, as well as discuss our findings.

### 3.1 Monthly Spending

In 2019, for both accounts from Fig. 3 (a), we can see that our spending increases towards the middle of the year, then decreases towards the end of the year. Compared to 2019 in Fig. 3 (b), we can see a significant decrease in spending after March. Since our data for 2020 only includes four months, the visualization is smaller than that of 2019.



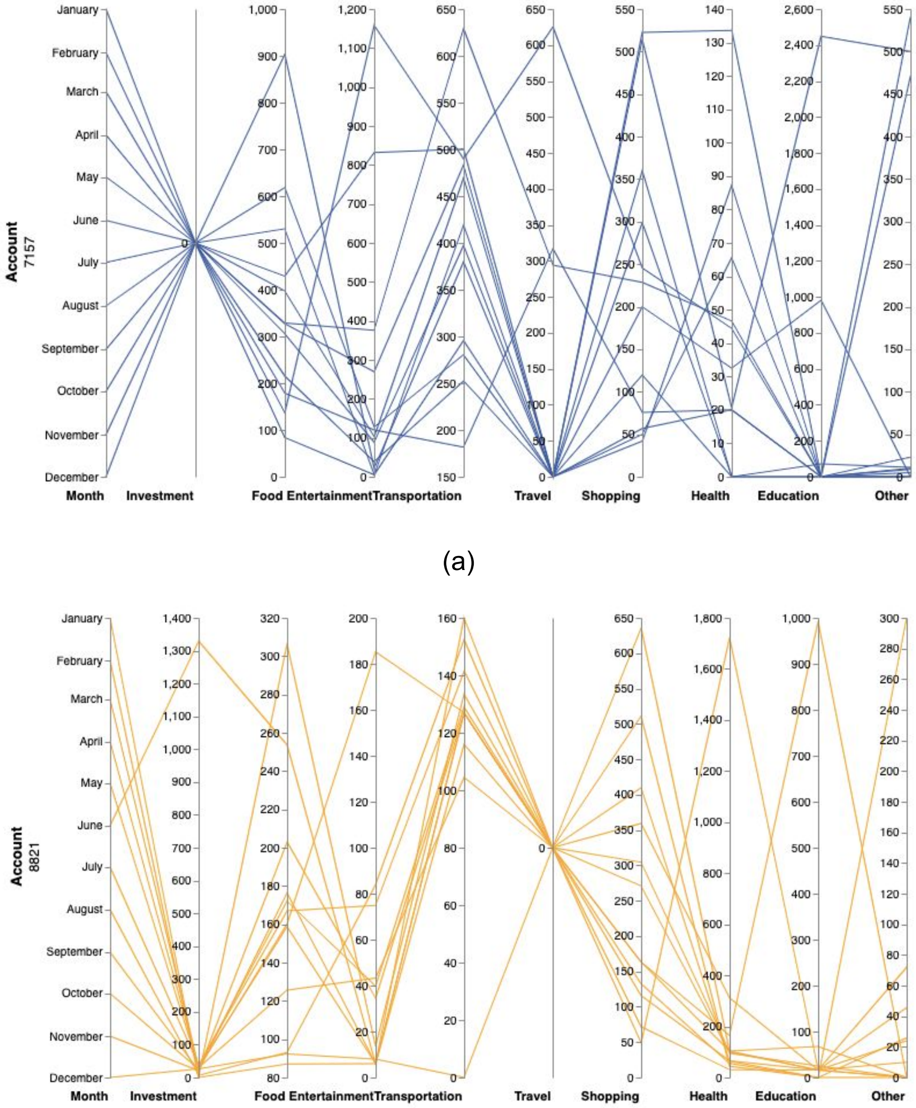
(a)



(b)

Fig. 3. Line charts for 2019 and 2020. Account 7157 spent the most in August 2019, and account 8821 spent the most in June 2019. Compare to 2020 where account 7157 spent the most in March, and account 8821 spent the most in February.

Our parallel coordinate plots helped us find patterns in our spendings based on the different categories. For account 7157 in 2019, as seen in Fig. 4 (a),  the deviation of their spendings was very spread out from most of the categories. For example food, shopping and health. Similarly, for account



(a)

(b)

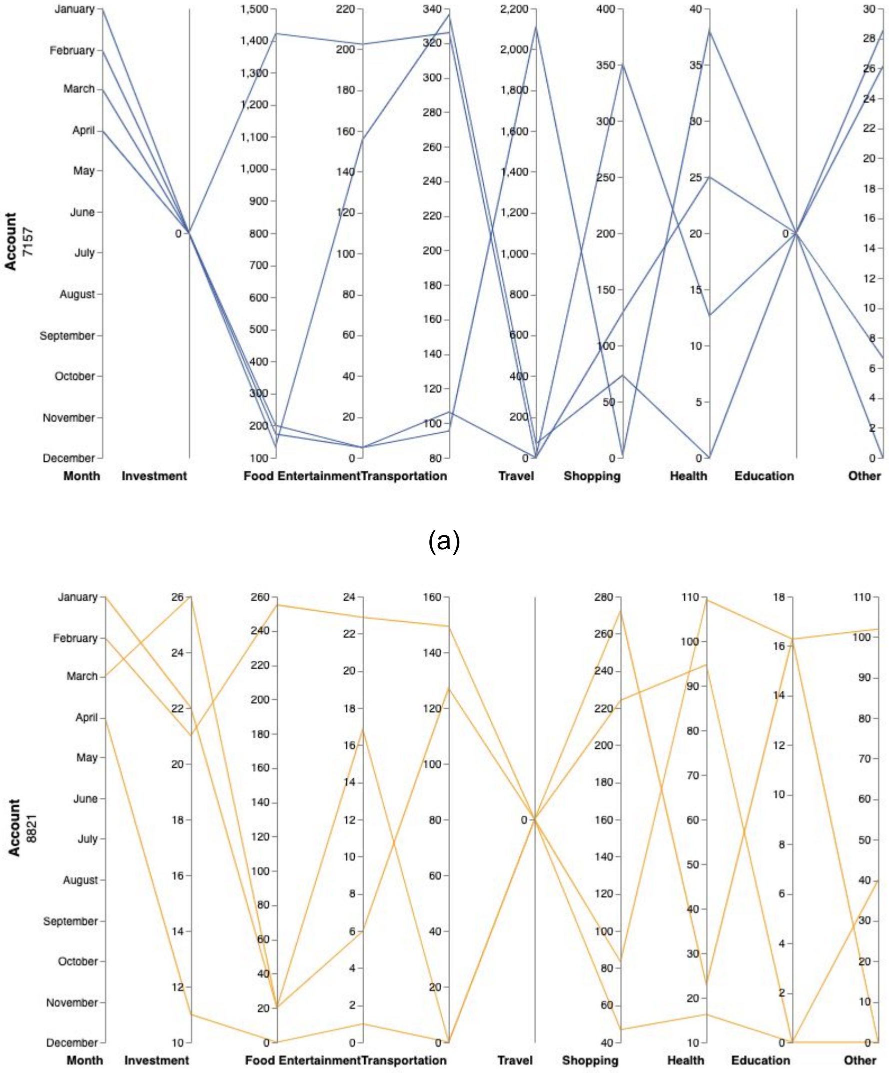
Fig. 4. Parallel Coordinates graphs for account 7157 and 8821 in 2019. Account 8821 displays more clusters and predictable behavior compared to account 7157.

8821, as seen in Fig. 4 (b), we can clearly identify that this user has a more predictable pattern to their spendings on each category per month with a few exceptions.

For the year 2020, by only looking at the range of deviation, we can conclude that the range of spending has reduced and therefore the cluster is only at a certain range. it is hard to see a clear pattern for both accounts as our spendings per category had reduced. This can be seen in Fig. 5 (a) and (b).

### 3.2 Annual Spending

Looking at the entire year of 2019, we can see from Fig. 6 (a), the outstanding amount spent was by account 7157 on transportation. Overall, we can see that account 7157 generally spends more on most categories than account 8821. Account 7157 spent the most on transportation, with an amount of $4778.56, whereas



(a)

(b)

Fig. 5. Parallel Coordinates graphs for account 7157 and 8821 in 2020. It is unclear to determine a pattern due to the limited data for 2020 available.

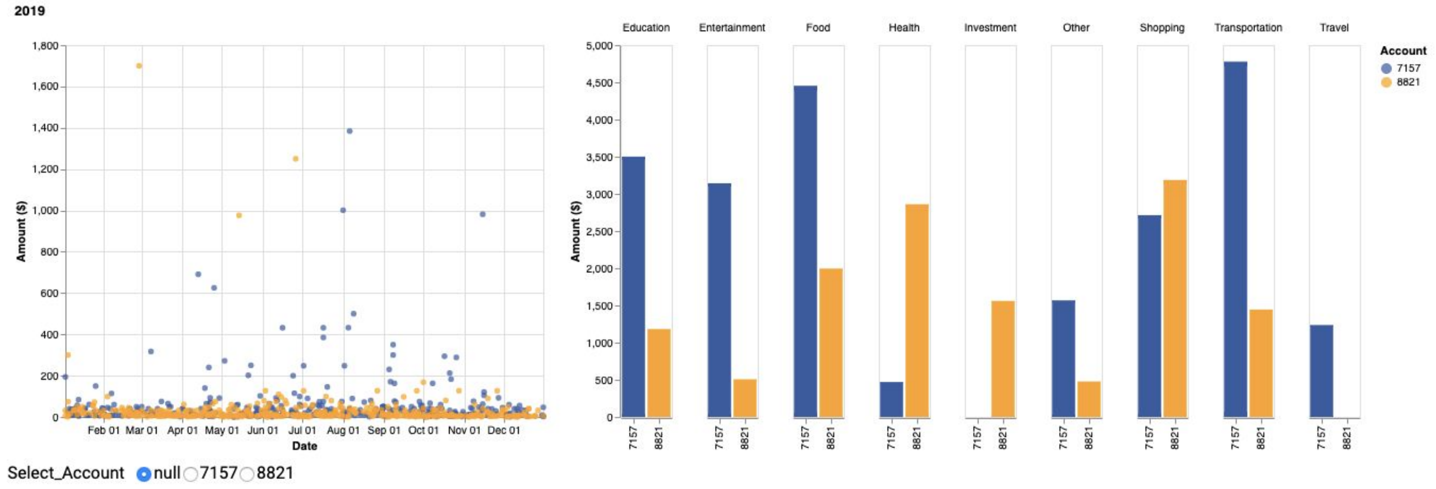
account 8821 spent the most on shopping with an amount of $3187.13. From this visualization, we can also see that account 7157 didn't spend any money on investments, and account 8821 didn't spend any money on travel.

In comparison to 2019, we can see from Fig. 6 (b), less spending activity from both accounts. Account 7157 previously spent the most on transportation, but in 2020 so far has spent the most on travel with an amount of $2181.10. Account 8821 spent the most on shopping, similar to 2019, with an amount of $626.43. The same trend holds true so far when it comes to each account's spending habits for travel and investments: account 7157 doesn't spend money on investments and account 8821 doesn't spend money on travel.

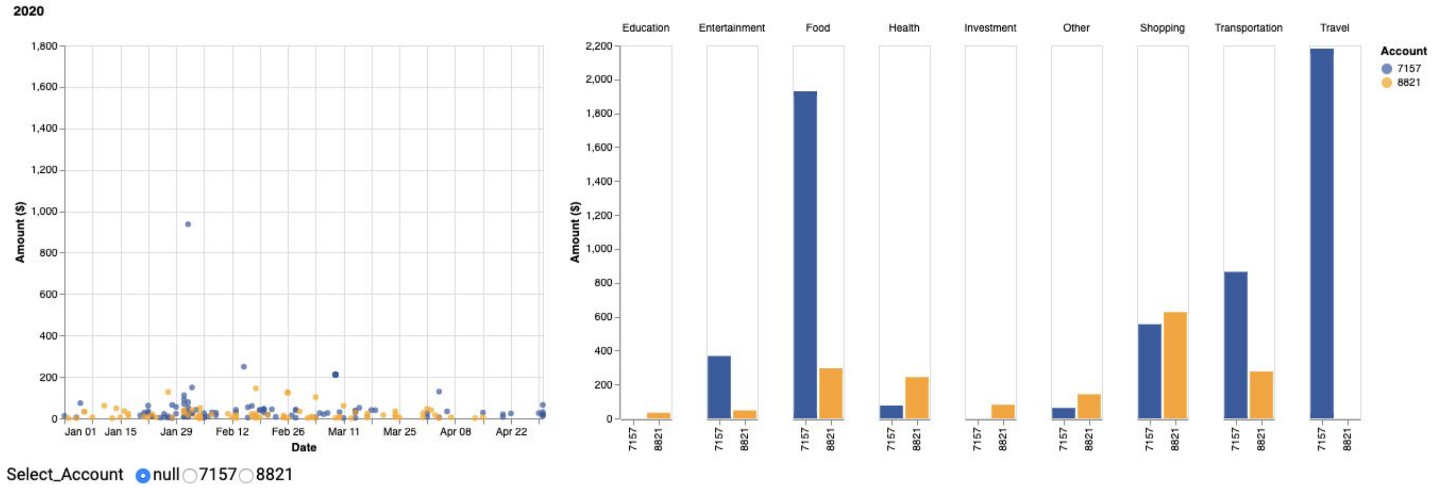
### 3.3 Seasonal Spending

When looking at the overall year for 2019 in Fig. 7 (a), we can clearly see that when added together, both accounts spend the most money in the summer. For account 7157, it is also apparent that more money is spent in the summer. For account 8821, the amount spent for summer, winter, and spring are very close. Still, summer wins by a little over $100. For some categories like travel and investment, there are missing radial plots - this is explained in the above section.

For 2020, since there have only been two seasons so far, we can see only two colors on our radial plots as seen in Fig. 7 (b). On the "Both Accounts" radial plot, we can see that there was more spending done in the winter than in the spring.

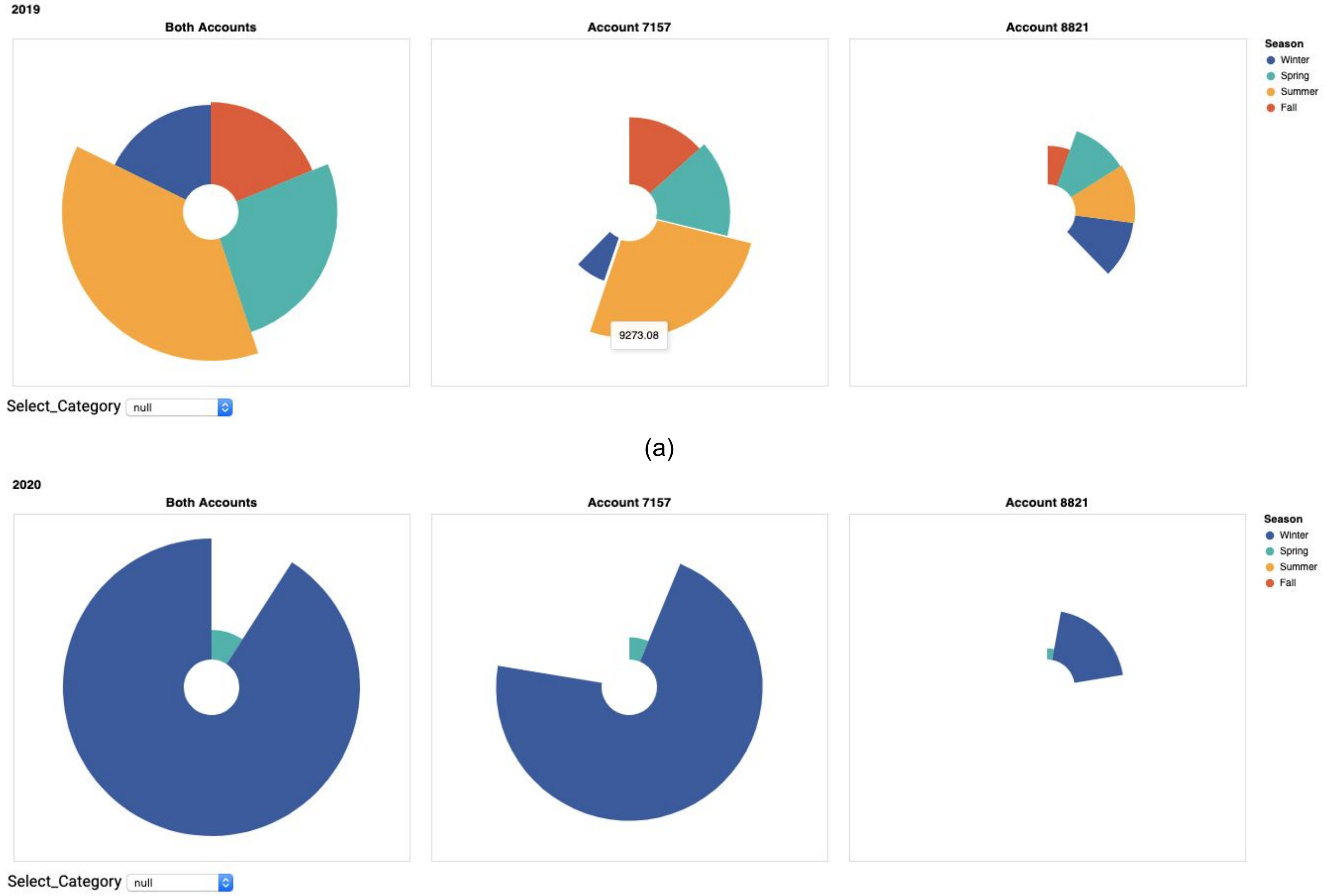


(a)



(b)

Fig. 6. Annual spending for 2019 and 2020 for both accounts. Account 7157 spent more in transportation and travel in 2019 and 2020, respectively, while account 8821 spent the most in shopping for both years. Account 7157 spent the least on health in 2019, and account 8821 spent the least on other.



(a)

(b)

Fig. 7. Seasonal spending for 2019 and 2020 for both accounts, as well as individual accounts. Radial plots for 2020 contain only two colors due to the limited data available so far.

## 4 DISCUSSION

In this section we discuss the results obtained in the section above and attempt to draw conclusions, as well as explain why certain patterns took place.

From our monthly spending visualizations we can conclude that during the months in which we have school breaks, for example winter, spring, and summer breaks, we tend to spend more. This can be due to the fact that we have more time and we shop more or go on vacations.

This is also supported on the radial plots for 2019. We tend to spend more in the warmer seasons like summer and spring. Which again, can be attributed to our school breaks.

When looking at the bar charts in the annual spendings, it is clear both accounts spend a lot of money on transportation, especially during the school months/seasons as the parallel coordinates plots and the radial plots also show. As commuter students, we expected this result.

For all visualizations, when it came to our 2020 data, we noticed a drastic decrease in our spending. This can definitely be attributed to the current health crisis we're facing. As seen on the line chart, our spendings greatly decreased after March, which is when social-distancing measures started taking place. This allows us to answer a question we hadn't initially thought about answering: How does a global pandemic affect our spending? In the future we hope to analyze the entire year of 2020 to further determine the long term effects of this lifestyle transition.

One of our initial questions had also been whether our gender affected our spending habits. As we began rendering our visualizations, we realized it wouldn't be accurate to determine patterns in spendings based on gender with only a sample size of one for each group. In the future we also hope to be able to analyze this by taking in larger samples of groups.

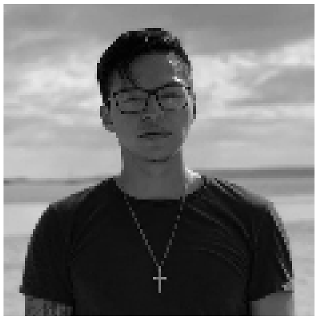
## 5 CONCLUSION

Analyzing our financial data was definitely an eye-opening experience. We live our lives day-by-day spending money on necessities as well as non-necessities, not really paying attention sometimes to how much we're really spending.

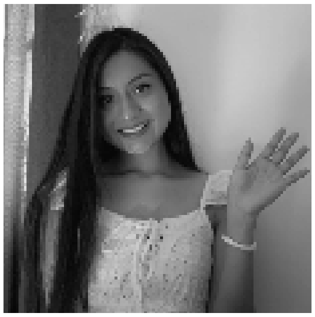
The first time we rendered our graphs we were astonished at how much we actually spent on certain categories. High amounts in categories like transportation confirmed that being commuter students can get expensive on a yearly basis. We also confirmed that most of our spending happens during the months in which we're on breaks from school. Initially, we thought about taking into account our genders to see how that affected our spending. We later determined that we would need a larger sample of each gender to make more accurate claims. Some of our visualizations did not go as originally planned. This was due to some limitations in Vega/ Vega-lite.

For now, we analyzed our own data, but we plan to allow other users to import their own CSV files containing their transaction statements so that they too can analyze their spending.

## AUTHORS



**Michael Mayaguari**



**Mayra Vazquez-Sanchez**

**For access to our code and visualizations visit our github repository at:**

**https://github.com/mgmayagu/Visualization**