Project

August 14, 2021

1 Project Analysis - Cohort Analysis & Layer Cake Chart

1.1 1.Proposal

Aim:

To conduct a cohort study of the paid and refunded amount of customers to monitor how acquired customers continue the product or service on a monthly or yearly basis and what does the montly revenue is highly dependent on.

Objectives:

Primary Objectives:

- To figure out the top 3 countries with the most customers

- To determine the user retention for each of the top 3 country with the use of a heatmap

- To determine the trend of customers that may have churned and acquired new customers over a fixed time period with the help of a layer cake chart

- To figure out if there was any activities contributed to the monthly revenue during the fixed time period

Secondary Objectives:

- To compare the difference of user retention between 3 countries that how the amount of customers may affected

- To provide the basis for future research to improve products and service

Steps involved in Cohort Analysis:

1. Data loading and cleaning

2. Assigned the cohort and calculate

Step 2.1

- Truncate data object into needed one (which is the first purchasing month)

- Create groupby object with target of unique customer id

- Transform with a min() function to figure out the smallest purchasing month

Step 2.2

- Create a groupby objet with CohortMonth and Cohort Index

- Count the number of customers in each group by applying the pandas nunique() function, also with num of paid and refunded

- Reset the index and create a pandas aggregate with paid, refunded and customer

Step 3

Calculate matrices: Retention rate and transform it into heatmap.

Step 4

Generate layer cake chart based on customers and first purchasing date

1.2 2. Introduction

In the 21st century, customers' data are becoming more and more essential, SaaS companies will be able to provide the most appropriate service to customers based on them. For this report, we are going to figure out whether the amount of revenue (paid - refunded) is related to the country that they come from. We assume that the more developed the country is, the more people are likely to spend money.

As we all know that US is one of the most powerful countries around the world, so our expected analysis outcome based on what we assume is citizens from US tend to spend more money. As cohort analysis gives the user an understanding of the why, how, and when of customer's actions that helps the user to make decisions, by conducting cohort analysis, we can conclude whether there is a positive correlation that can fit our assumption.

To avoid bias and any other side effects, we will conduct cohort analysis on the other two top countries from the list to ensure our conclusion is evidential and reliable.

For further progress, we will conduct a cohort layer cake chart with no geographic separation. Layer cake chart is another type of chart used by SaaS(Software as a service). As they need to monitor monthly recurring revenue and annual recurring revenue to make decisions, it is essential to use a layer cake chart to monitor customers on a monthly or yearly basis.

1.3 3. Retrieve Dataset

Datasets: customer details in 2020: https://github.com/shy1213/Project/blob/main/query_result_2021-05-18T03%2046%2057.095586Z%20(1).csv

```
[4]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import matplotlib as mpl
     import seaborn as sns
     import warnings
     warnings.filterwarnings('ignore')

     pd.set_option('max_columns', 50)
     mpl.rcParams['lines.linewidth'] = 2
```

```
%matplotlib inline
print('Libraries imported')
```

Libraries imported

```
[5]: df = pd.read_csv('/Users/shy/Downloads/customer.csv')
     df['card__country'] = df.groupby(['customer'])['card__country'].ffill()
     df['created'] = pd.to_datetime(df['created'], format = '%Y-%m-%d %H:%M:%S')
     df.groupby(['card__country'])['customer'].count().sort_values().tail(3)
```

```
[5]: card__country
     CA       256
     AU       307
     US      1685
     Name: customer, dtype: int64
```

```
[6]: df.set_index('customer', inplace = True)

     df['CohortGroup'] = df.groupby(level = 0)['created'].min().apply(lambda x: x.
      ↪strftime('%Y-%m'))
     df.reset_index(inplace = True)

     def cohort_period(df):
         df['CohortPeriod'] = np.arange(len(df)) + 1
         return df
```

```
[7]: import seaborn as sns
     def heat_map(cohort_analysis):
         sns.set(style = 'white')

         plt.figure(figsize = (12, 8))
         plt.title('Cohorts: User Retention')
         sns.heatmap(cohort_analysis.T, mask = cohort_analysis.T.isnull(), annot =␣
      ↪True, fmt = '.0%')
```

```
[8]: def Cohort_Analysis(country):
         data_set_country = df[df['card__country'] == country]
         data_set_country['Period'] = data_set_country['created'].apply(lambda x: x.
      ↪strftime('%Y-%m'))
         data_set_country.set_index('customer', inplace = True)
         data_set_country['CohortGroup'] = data_set_country.groupby(level =␣
      ↪0)['created'].min().apply(lambda x: x.strftime('%Y-%m'))
         data_set_country.reset_index(inplace = True)


         grouped = data_set_country.groupby(['CohortGroup', 'Period'])
```

```
cohorts = grouped.agg({'paid': np.sum,
                        'refunded': np.sum,
                        'customer': pd.Series.nunique})

cohorts = cohorts.groupby(level = 0).apply(cohort_period)
cohorts.reset_index(inplace = True)
cohorts.set_index(['CohortGroup', 'CohortPeriod'], inplace = True)

cohort_group_size = cohorts['customer'].groupby(level = 0).first()
user_retention = cohorts['customer'].unstack(0).divide(cohort_group_size,␣
↪axis = 1)

return user_retention
```
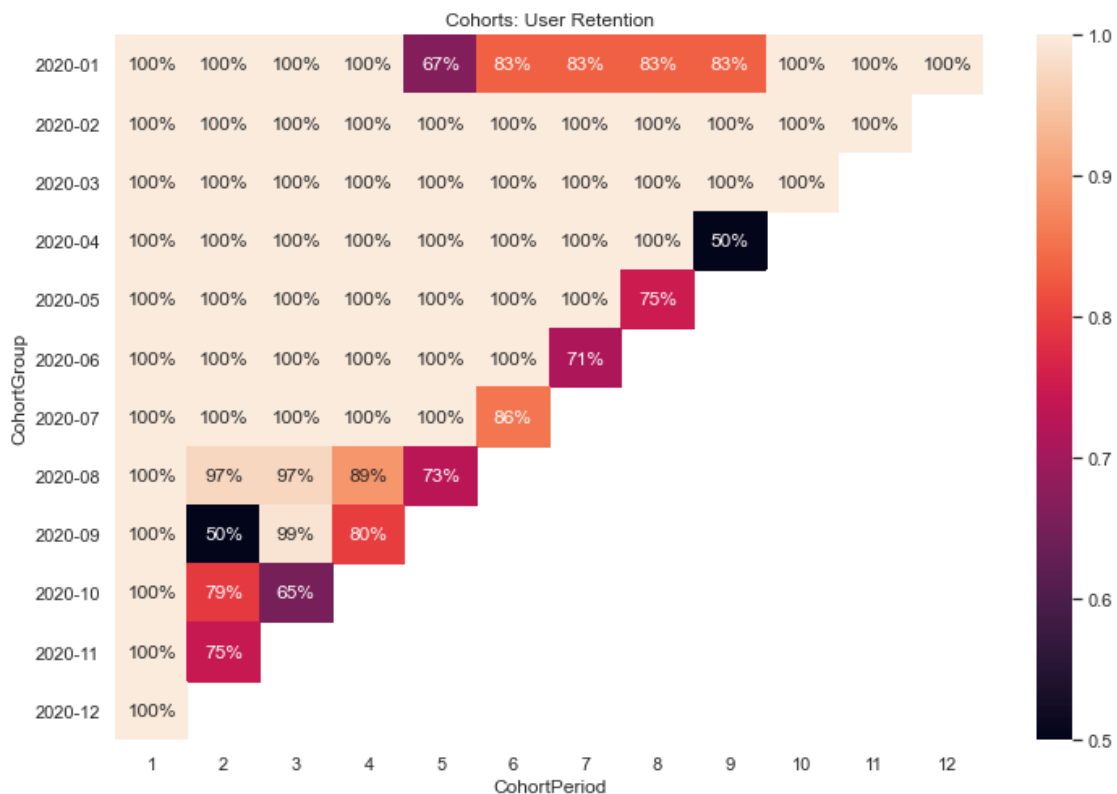
1.3.1   3.1 US Cohort Analysis

```
[9]: heat_map(Cohort_Analysis('US'))
```
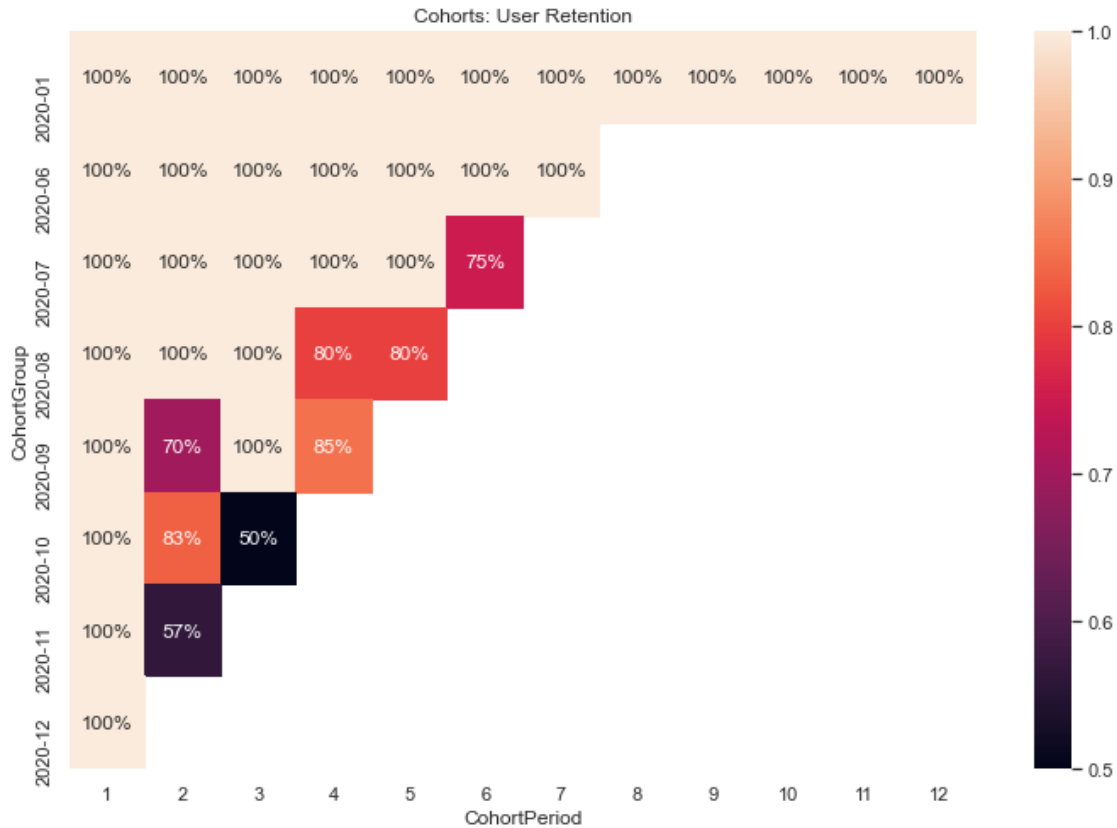

We are able to see the first period (column) of each cohort is 100% as we assume each customer purchased in month 1. We can see from the above chart that fewer users tend to purchase as time goes on with purchaseing month other than 2020-02 and 2020-03. Over the rest month, we see that 2020-09 cohort month in 3rd cohort index with 99%, which means that 99% of customer purchased

in Sep 2020 and again 3 months later. The lowerst is 2020-04 cohort month in 9th cohort index with 50%, only 50% of customers returned to purchase 9 months after April 2020.
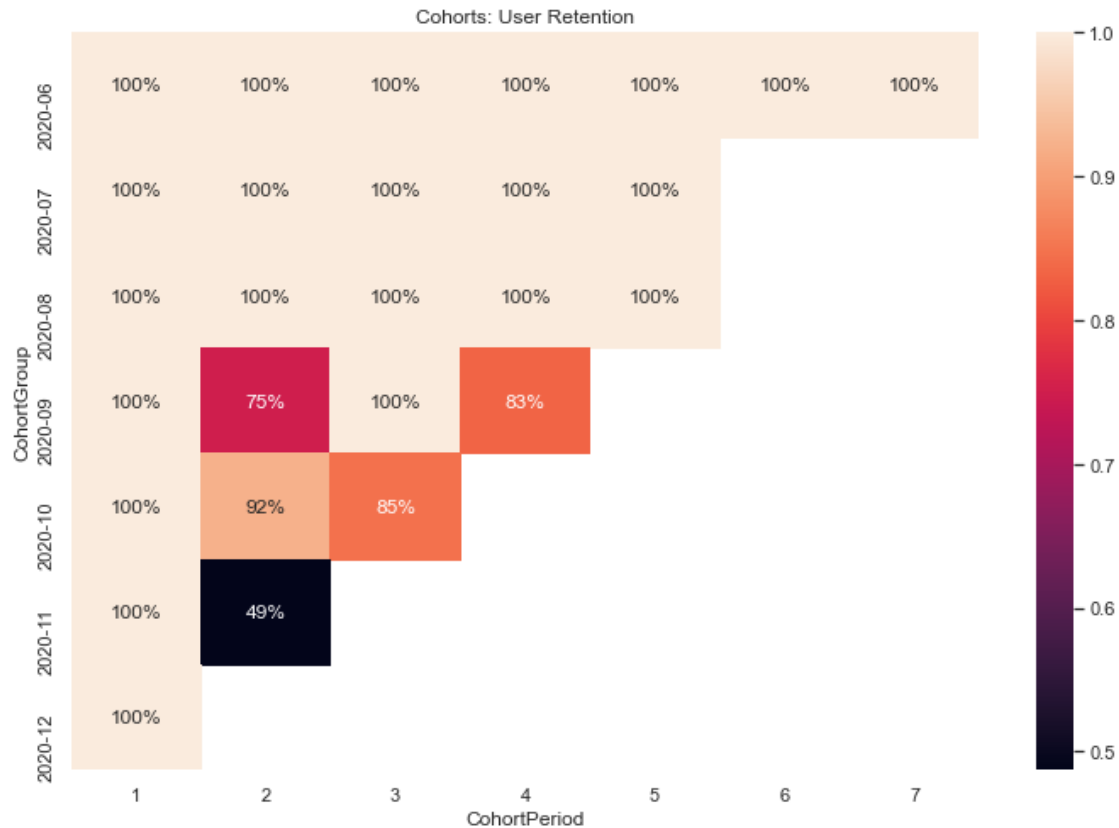
1.3.2 3.2 AU Cohort Analysis

```
[10]: heat_map(Cohort_Analysis('AU'))
```


We are able to see the first period (column) of each cohort is 100% as we assume each customer purchased in month 1. We can see from the above chart that fewer users tend to purchase as time goes on with purchaseing month other than 2020-01 and 2020-06. Over the rest month, we see that 2020-09 cohort month in 4th cohort index with 85%, which means that 85% of customer purchased in Sep 2020 and again 4 months later. The lowerst is 2020-10 cohort month in 3rd cohort index with 50%, only 50% of customers returned to purchase 3 months after Oct 2020.

1.3.3 3.3 CA Cohort Analysis

```
[11]: heat_map(Cohort_Analysis('CA'))
```
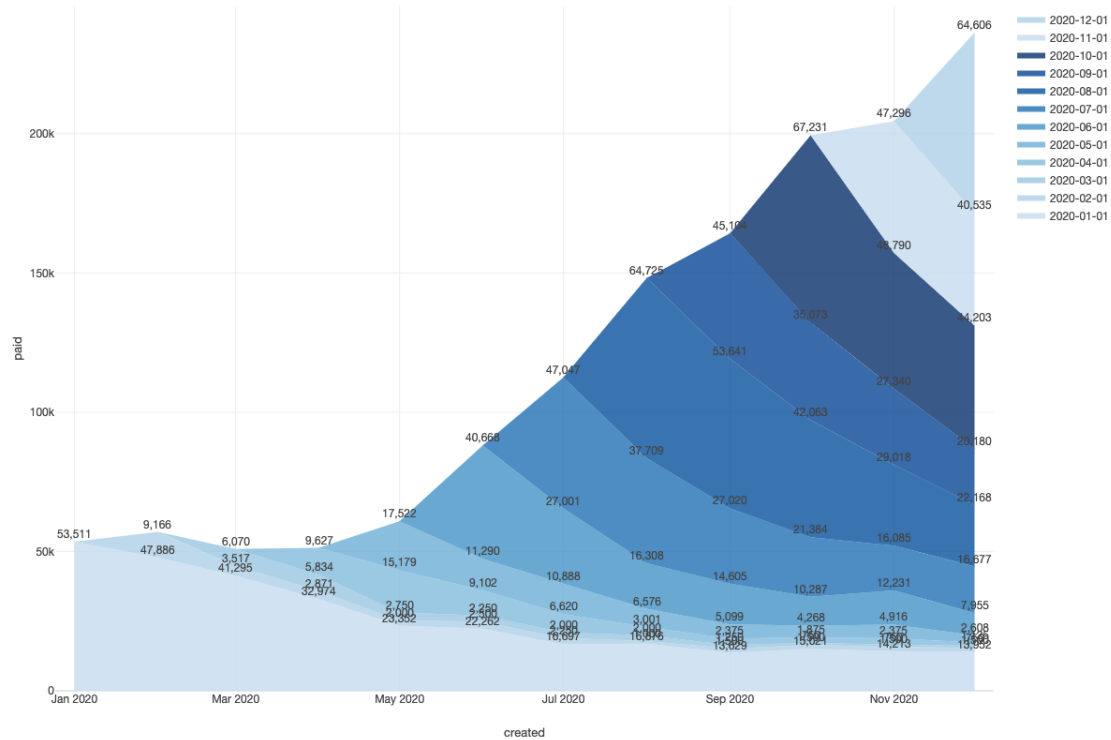
Cohorts: User Retention

We are able to see the first period (column) of each cohort is 100% as we assume each customer purchased in month 1. We can see from the above chart that fewer users tend to purchase as time goes on with purchaseing month other that 2020-06, 2020-07 and 2020-08. Over the rest month, we see that 2020-10 cohort month in 2nd cohort index with 92%, which means that 92% of customer purchased in Oct 2020 and again 2 months later. The lowerst is 2020-11 cohort month in 2nd cohort index with 49%, only 49% of customers returned to purchase 2 months after Nov 2020.

1.4 4.Layer Cake Chart

```
[12]: from IPython import display
      display.Image("/Users/shy/Downloads/Cake_Chart.png")
```

[12]:

paid

200k

150k

100k

50k

0

Jan 2020   Mar 2020   May 2020   Jul 2020   Sep 2020   Nov 2020

created

53,511   9,166   6,070   9,627   17,522   11,290   27,001   27,020   45,184   47,296   64,606

47,886   41,295   3,517   5,834   15,179   9,102   10,888   16,308   64,725   43,790   40,535

32,974   2,871   2,750   3,250   6,620   6,576   14,605   53,641   35,073   44,203

23,352   22,262   2,900   3,001   5,099   42,063   27,340   26,180

16,697   16,676   2,375   21,384   29,018   22,168

13,629   1,558   10,287   16,085   16,677

15,021   14,213   12,231   7,955

1,875   4,268   4,916   2,608

67,231   47,047   40,666   37,709

From the graph, it could be seen that monthy revenue is highly dependent on new customers. However, the company accumulates several layers of incomes from existing customers and reduced dependence. Further more, it seems like there was activities like promo happended in Nov 2020 and a few cohorts responded.

1.5   5. Discussion

From my previous work, I didn't focus carefully on the missing values that many customers don't have their corresponding card_countries. As a result of that, all three of my cohort analysis charts look absurd and this kind of issue happens a lot in real life, that data might not be arranged nice and neat. Therefore, I filled in those missing values and my charts start to seem reasonable and can be analyzed. Also, I used several for loops in my draft work as I was not familiar with pandas. They can be time-wasting when dealing with huge datasets and I turned to use built-in functions to do data cleaning and analysis. Cohort analysis is useful and essential to business companies as it's a kind of descriptive analytics that able to reflect user retention correctly and clearly.