part1-metrics

December 13, 2019

1 Part 1: Know Your Metrics

In terms of growth we could want more customers, more orders, more revenue, more signups, more efficiency…

Before going into coding, we need to understand what exactly is our metric.

> *The North Star Metric is the single metric that best captures the core value that your product delivers to customers.*

This metric depends on company's product, position, targets & more. Airbnb's North Star Metric is nights booked whereas for Facebook, it is daily active users.

We will use a sample dataset of an online retail. For an online retail, we can choose our North Star Metric as **Monthly Revenue**. In addition to Monthly Revenue we will also calculate following metrics: * **Monthly Active Customer** * **Monthly Order Count** * **Average Revenue per Order** * **New Customer Ratio** * **Monthly Retention Rate** * **Cohort Based Retention Rate**

1.1 Load the dataset

Let's first load the dataset. This is how our data looks like.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
InvoiceNo      541909 non-null object
StockCode      541909 non-null object
Description    540455 non-null object
Quantity       541909 non-null int64
InvoiceDate    541909 non-null datetime64[ns]
UnitPrice      541909 non-null float64
CustomerID     406829 non-null float64
Country        541909 non-null object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
None
```

```
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
```

```
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6
5    536365     22752         SET 7 BABUSHKA NESTING BOXES         2
6    536365     21730   GLASS STAR FROSTED T-LIGHT HOLDER         6
7    536366     22633                HAND WARMER UNION JACK         6
8    536366     22632             HAND WARMER RED POLKA DOT         6
9    536367     84879        ASSORTED COLOUR BIRD ORNAMENT        32

          InvoiceDate  UnitPrice  CustomerID         Country
0 2010-12-01 08:26:00       2.55     17850.0  United Kingdom
1 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
2 2010-12-01 08:26:00       2.75     17850.0  United Kingdom
3 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
4 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
5 2010-12-01 08:26:00       7.65     17850.0  United Kingdom
6 2010-12-01 08:26:00       4.25     17850.0  United Kingdom
7 2010-12-01 08:28:00       1.85     17850.0  United Kingdom
8 2010-12-01 08:28:00       1.85     17850.0  United Kingdom
9 2010-12-01 08:34:00       1.69     13047.0  United Kingdom

(541909, 8)
```

We have all the crucial information we need: * Customer ID * Unit Price * Quantity * Invoice Date

With all these features, we can build our North Star Metric equation: > *Revenue = Active Customer Count * Order Count * Average Revenue per Order*

1.2 Revenue

We want to see monthly revenue. So let's calculate revenue of each order first.

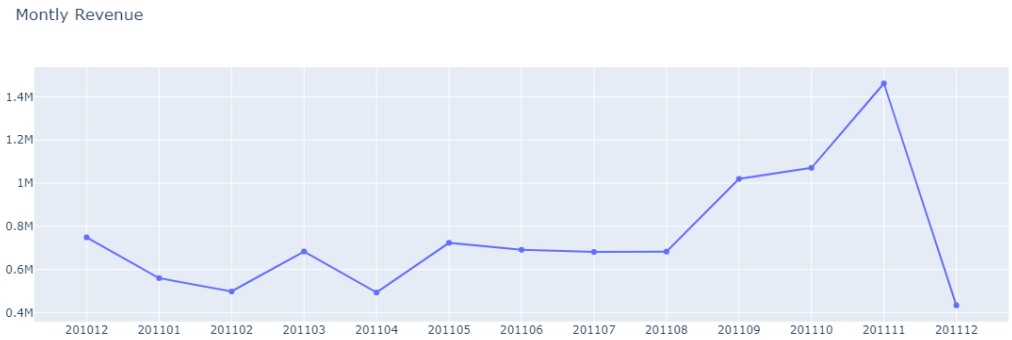
```
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

          InvoiceDate  UnitPrice  CustomerID         Country  \
0 2010-12-01 08:26:00       2.55     17850.0  United Kingdom
1 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
2 2010-12-01 08:26:00       2.75     17850.0  United Kingdom
3 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
4 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
```

	InvoiceYearMonth	Revenue
0	201012	15.30
1	201012	20.34
2	201012	22.00
3	201012	20.34
4	201012	20.34

Summing up for every month and we have montly revenue.

	InvoiceYearMonth	Revenue
0	201012	748957.020
1	201101	560000.260
2	201102	498062.650
3	201103	683267.080
4	201104	493207.121
5	201105	723333.510
6	201106	691123.120
7	201107	681300.111
8	201108	682680.510
9	201109	1019687.622
10	201110	1070704.670
11	201111	1461756.250
12	201112	433668.010

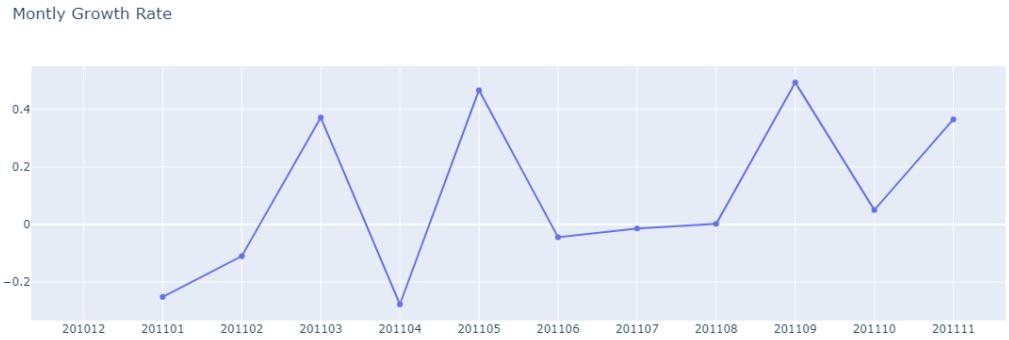
We can also visualize **Monthly Revenue**.



This clearly shows our revenue is growing especially Aug '11 onwards (and our data in December is incomplete). Absolute numbers are fine, let's figure out what is our **Monthly Revenue Growth Rate**:

	InvoiceYearMonth	Revenue	MonthlyGrowth
0	201012	748957.020	NaN
1	201101	560000.260	-0.252293

```
2            201102  498062.650      -0.110603
3            201103  683267.080       0.371850
4            201104  493207.121      -0.278163
```



Everything looks good, we saw 36.5% growth previous month (December is excluded in the code since it hasn't been completed yet). But we need to identify what exactly happened on April. Was it due to less active customers or our customers did less orders? Maybe they just started to buy cheaper products? We can't say anything without doing a deep-dive analysis.

1.3 Monthly Active Customers

To see the details Monthly Active Customers, we will follow the steps we exactly did for Monthly Revenue. Starting from this part, we will be focusing on UK data only (which has the most records). We can get the monthly active customers by counting unique `CustomerID`s.

```
Country
United Kingdom    8187806
Netherlands        284661
EIRE               263276
Germany            221698
France             197403
Name: Revenue, dtype: int32
```

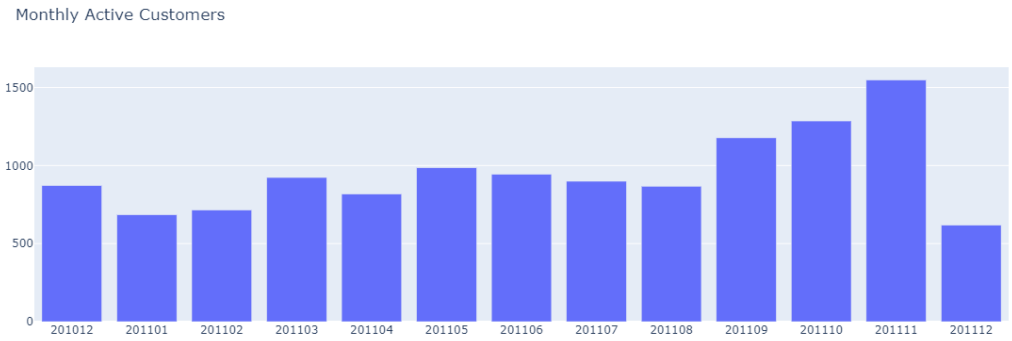
```
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

          InvoiceDate  UnitPrice  CustomerID         Country  \
0 2010-12-01 08:26:00       2.55     17850.0  United Kingdom
1 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
```

```
2 2010-12-01 08:26:00       2.75     17850.0  United Kingdom
3 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
4 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
```

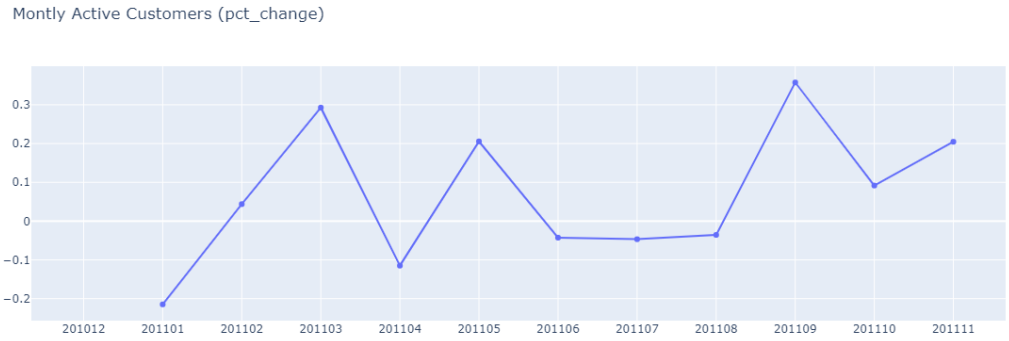
	InvoiceYearMonth	Revenue
0	201012	15.30
1	201012	20.34
2	201012	22.00
3	201012	20.34
4	201012	20.34

Number of active customers per month and its bar plot:

	InvoiceYearMonth	CustomerID	pct_change
0	201012	871	NaN
1	201101	684	-0.214696
2	201102	714	0.043860
3	201103	923	0.292717
4	201104	817	-0.114843
5	201105	985	0.205630
6	201106	943	-0.042640
7	201107	899	-0.046660
8	201108	867	-0.035595
9	201109	1177	0.357555
10	201110	1285	0.091759
11	201111	1548	0.204669
12	201112	617	-0.601421



In April, Monthly Active Customer number dropped to 817 from 923 (-11.5%).

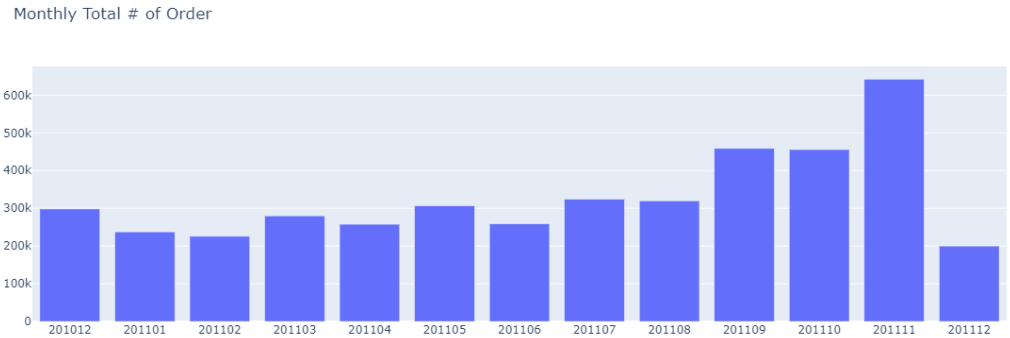


We will see the same trend for number of orders as well.

1.4 Monthly Order Count

We will apply the same steps this time using Quantity field:

	InvoiceYearMonth	Quantity
0	201012	298101
1	201101	237381
2	201102	225641
3	201103	279843
4	201104	257666
5	201105	306452
6	201106	258522
7	201107	324129
8	201108	319804
9	201109	458490
10	201110	455612
11	201111	642281
12	201112	199907

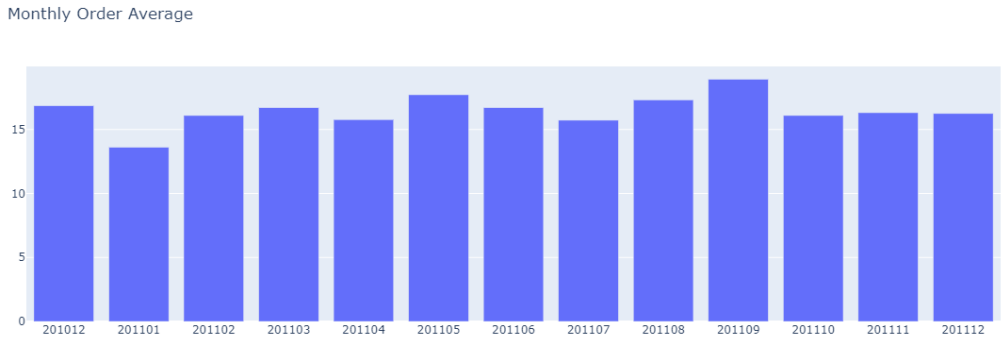


As we expected, Order Count is also declined in April (279k to 257k, -8%) We know that Active Customer Count directly affected Order Count decrease. At the end, we should definitely check our **Average Revenue per Order** as well.

1.5 Average Revenue per Order

To get this data, we need to calculate the average of revenue for each month:

	InvoiceYearMonth	Revenue
0	201012	16.865860
1	201101	13.614680
2	201102	16.093027
3	201103	16.716166
4	201104	15.773380
5	201105	17.713823
6	201106	16.714748
7	201107	15.723497
8	201108	17.315899
9	201109	18.931723
10	201110	16.093582
11	201111	16.312383
12	201112	16.247406



Even the monthly order average dropped for April (16.7 to 15.8). We observed slow-down in every metric affecting our North Star.

We have looked at our major metrics. Of course there are many more and it varies across industries. Let's continue investigating some other important metrics: * **New Customer Ratio**: a good indicator of if we are losing our existing customers or unable to attract new ones * **Retention Rate**: King of the metrics. Indicates how many customers we retain over specific time window. We will be showing examples for monthly retention rate and cohort based retention rate.

1.6 New Customer Ratio

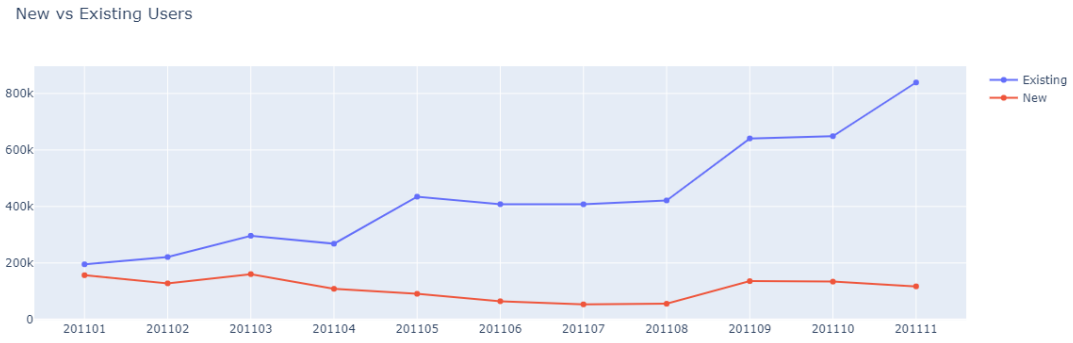
First we should define what is a new customer. In our dataset, we can assume a new customer is whoever did his/her first purchase in the time window we defined. We will do it monthly for this example.

We will be using `.min()` function to find our first purchase date for each customer and define new customers based on that. The code below will apply this function and show us the revenue breakdown for each group monthly.

```
   CustomerID     MinPurchaseDate  MinPurchaseYearMonth
0     12346.0 2011-01-18 10:01:00                201101
1     12747.0 2010-12-05 15:38:00                201012
2     12748.0 2010-12-01 12:48:00                201012
3     12749.0 2011-05-10 15:25:00                201105
4     12820.0 2011-01-17 12:34:00                201101

Existing    256114
New         105764
Name: UserType, dtype: int64

  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

          InvoiceDate  UnitPrice  CustomerID         Country  \
0 2010-12-01 08:26:00       2.55     17850.0  United Kingdom
1 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
2 2010-12-01 08:26:00       2.75     17850.0  United Kingdom
3 2010-12-01 08:26:00       3.39     17850.0  United Kingdom
4 2010-12-01 08:26:00       3.39     17850.0  United Kingdom

   InvoiceYearMonth  Revenue     MinPurchaseDate  MinPurchaseYearMonth  \
0            201012    15.30 2010-12-01 08:26:00                201012
1            201012    20.34 2010-12-01 08:26:00                201012
2            201012    22.00 2010-12-01 08:26:00                201012
3            201012    20.34 2010-12-01 08:26:00                201012
4            201012    20.34 2010-12-01 08:26:00                201012

  UserType
0      New
1      New
2      New
3      New
4      New
```

Lets calculate the Revenue per month for each user type.

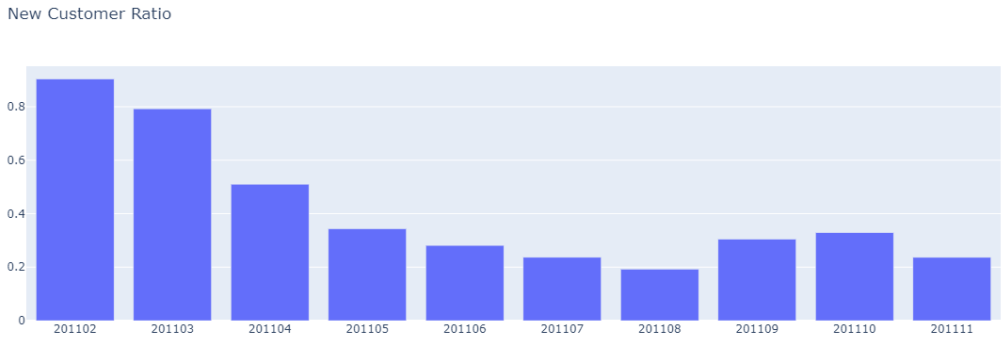
	InvoiceYearMonth	UserType	Revenue
0	201012	New	483799.740
1	201101	Existing	195275.510
2	201101	New	156705.770
3	201102	Existing	220994.630
4	201102	New	127859.000
5	201103	Existing	296350.030
6	201103	New	160567.840
7	201104	Existing	268226.660
8	201104	New	108517.751
9	201105	Existing	434725.860
10	201105	New	90847.490
11	201106	Existing	408030.060
12	201106	New	64479.190
13	201107	Existing	407693.610
14	201107	New	53453.991
15	201108	Existing	421388.930
16	201108	New	55619.480
17	201109	Existing	640861.901
18	201109	New	135667.941
19	201110	Existing	648837.600
20	201110	New	133940.280
21	201111	Existing	838955.910
22	201111	New	117153.750
23	201112	Existing	273472.660
24	201112	New	24447.810



Existing customers are showing a positive trend and tell us that our customer base is growing but new customers have a slight negative trend.

Let's have a better view by looking at the New Customer Ratio:

```
InvoiceYearMonth
201012    871
201101    362
201102    339
201103    408
201104    276
201105    252
201106    207
201107    172
201108    140
201109    275
201110    318
201111    296
201112     34
Name: CustomerID, dtype: int64

InvoiceYearMonth
201101     322
201102     375
201103     515
201104     541
201105     733
201106     736
201107     727
201108     727
201109     902
201110     967
201111    1252
201112     583
Name: CustomerID, dtype: int64

    InvoiceYearMonth  CustomerID
1             201101    1.124224
2             201102    0.904000
3             201103    0.792233
4             201104    0.510166
5             201105    0.343793
6             201106    0.281250
7             201107    0.236589
8             201108    0.192572
9             201109    0.304878
10            201110    0.328852
11            201111    0.236422
12            201112    0.058319
```



New Customer Ratio has declined as expected (we assumed on Feb, all customers were New) and running around 20%.

1.7 Generate Sign Up Data

We want to inspect channel the customers came from when they made first purchase.

Imagine we have data on the date cusotmer SignedUp and date when customer installed the mobile app. We don't so we'll simulate this data. This is what fucntion `generate_signup_data` does.

```
   CustomerID     MinPurchaseDate  MinPurchaseYearMonth
0     12346.0 2011-01-18 10:01:00                201101
1     12747.0 2010-12-05 15:38:00                201012
2     12748.0 2010-12-01 12:48:00                201012
3     12749.0 2011-05-10 15:25:00                201105
4     12820.0 2011-01-17 12:34:00                201101

array([201101, 201012, 201105, 201109, 201102, 201110, 201108, 201106,
       201103, 201107, 201104, 201111, 201112], dtype=int64)

   CustomerID     MinPurchaseDate  MinPurchaseYearMonth  SignupYearMonth  \
0     12346.0 2011-01-18 10:01:00                201101           201101
1     12747.0 2010-12-05 15:38:00                201012           201012
2     12748.0 2010-12-01 12:48:00                201012           201012
3     12749.0 2011-05-10 15:25:00                201105           201104
4     12820.0 2011-01-17 12:34:00                201101           201101

   InstallYearMonth
0            201101
1            201012
2            201012
3            201101
4            201101
```

Simulate the channel from which the customer came.

1.8 Activation Rate

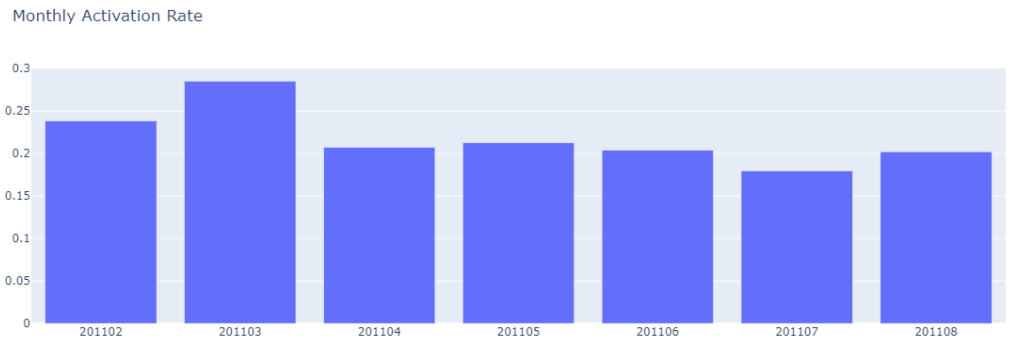
Retention rate should be monitored very closely because it indicates how effective is your campaign.

Activation rate indicates the percentage of people who upon signing-up your campaign actually engage with by making a purchase.

For making Monthly Activation Rate visualized, we need to calculate how many customers were made purchase in the month they signed up.

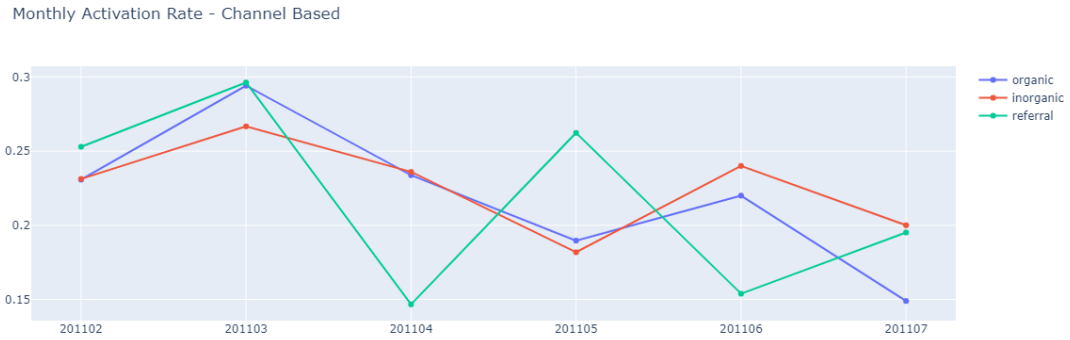
> ***Monthly Activation Rate*** = *Signed Up and Engaged Customers / Total Signed Up Customers*

	SignupYearMonth	CustomerID
0	201012	0.577203
1	201101	0.274268
2	201102	0.238477
3	201103	0.285303
4	201104	0.207469
5	201105	0.212644
6	201106	0.203947
7	201107	0.179688
8	201108	0.202128
9	201109	0.295775
10	201110	0.491525
11	201111	0.846154
12	201112	1.000000



Let's also check activation rates by acquisition channel. We do it by counting all customers that purchased and signed-up in the selected month vs. total signed-up customers and grouping by month and channel.

	SignupYearMonth	AcqChannel	CustomerID
0	201012	inorganic	0.604207
1	201012	organic	0.537155
2	201012	referral	0.586408
3	201101	inorganic	0.301508
4	201101	organic	0.267606
5	201101	referral	0.257384



Since it is synteticaly generated data we won't take it into consideration for the analysis.

1.9 Monthly Retention Rate

Retention rate should be monitored very closely because it indicates how sticky is your service and how well your product fits the market. For making Monthly Retention Rate visualized, we need to calculate how many customers were retained from previous month.

> ***Monthly Retention Rate*** = *Retained Customers From Prev. Month / Active Customers Total*

We will be using `crosstab()` function of pandas which to calculate Retention Rate.

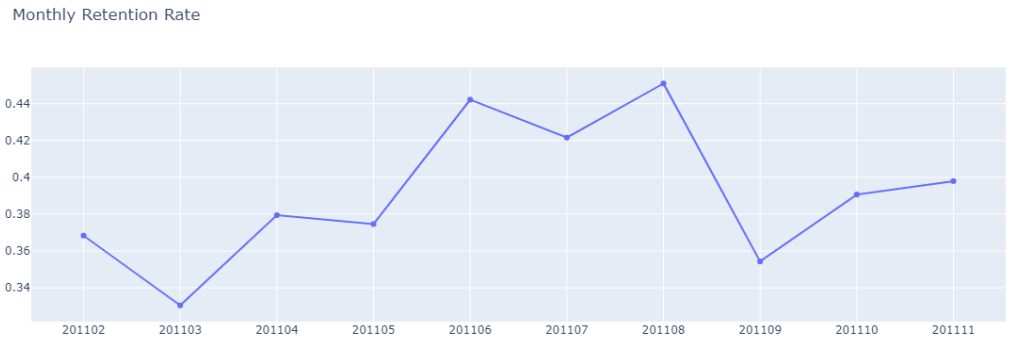
InvoiceYearMonth	CustomerID	201012	201101	201102	201103	201104	201105	\
0	12346.0	0	1	0	0	0	0	
1	12747.0	1	1	0	1	0	1	
2	12748.0	1	1	1	1	1	1	
3	12749.0	0	0	0	0	0	1	
4	12820.0	0	1	0	0	0	0	

InvoiceYearMonth	201106	201107	201108	201109	201110	201111	201112
0	0	0	0	0	0	0	0
1	1	0	1	0	1	1	1
2	1	1	1	1	1	1	1
3	0	0	1	0	0	1	1
4	0	0	0	1	1	0	1

Retention table shows us which customers are active on each month (1 stands for active).

In the for loop, for each month we calculate Retained Customer Count from previous month and Total Customer Count.

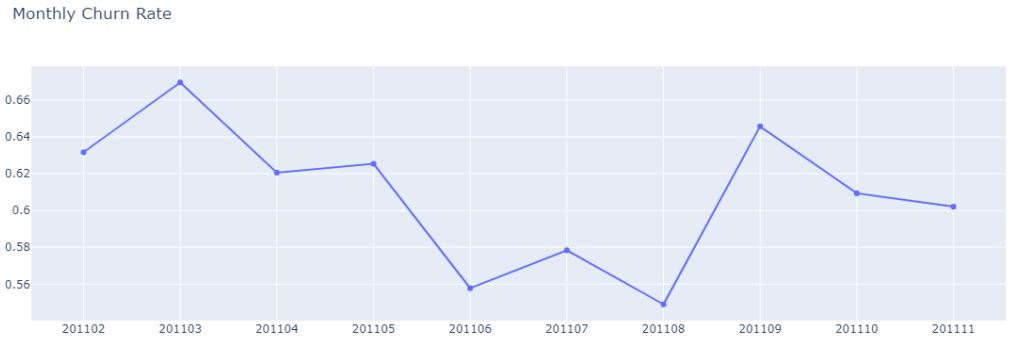
	InvoiceYearMonth	RetainedUserCount	TotalUserCount	RetentionRate
0	201102	263	714	0.368347
1	201103	305	923	0.330444
2	201104	310	817	0.379437
3	201105	369	985	0.374619
4	201106	417	943	0.442206
5	201107	379	899	0.421580
6	201108	391	867	0.450980
7	201109	417	1177	0.354291
8	201110	502	1285	0.390661
9	201111	616	1548	0.397933
10	201112	402	617	0.651540



Monthly Retention Rate significantly jumped from June to August and went back to previous levels afterwards.

1.10 Churn Rate

Chrun rate is opposite of retention rate. It is the percentage of customer that were active on the previous month but did not buy anyting in the current month.



1.11 Cohort Based Retention Rate

There is another way of measuring Retention Rate which allows you to see Retention Rate for each cohort. Cohorts are determined as first purchase year-month of the customers. We will be measuring what percentage of the customers retained after their first purchase in each month.

This view will help us to see how recent and old cohorts differ regarding retention rate and if recent changes in customer experience affected new customer's retention or not.

InvoiceYearMonth	TotalUserCount	201101	201102	201103	201104	201105	\
201101	684	1.0	0.38	0.26	0.18	0.15	
201102	714	NaN	1.00	0.43	0.23	0.19	
201103	923	NaN	NaN	1.00	0.34	0.23	
201104	817	NaN	NaN	NaN	1.00	0.45	
201105	985	NaN	NaN	NaN	NaN	1.00	
201106	943	NaN	NaN	NaN	NaN	NaN	
201107	899	NaN	NaN	NaN	NaN	NaN	
201108	867	NaN	NaN	NaN	NaN	NaN	
201109	1177	NaN	NaN	NaN	NaN	NaN	
201110	1285	NaN	NaN	NaN	NaN	NaN	
201111	1548	NaN	NaN	NaN	NaN	NaN	
201112	617	NaN	NaN	NaN	NaN	NaN	

InvoiceYearMonth	201106	201107	201108	201109	201110	201111	201112
201101	0.13	0.12	0.11	0.10	0.08	0.08	0.07
201102	0.16	0.14	0.12	0.11	0.10	0.09	0.07
201103	0.17	0.13	0.11	0.11	0.09	0.09	0.06
201104	0.28	0.20	0.16	0.15	0.12	0.11	0.08
201105	0.42	0.25	0.19	0.16	0.13	0.12	0.08
201106	1.00	0.40	0.25	0.19	0.15	0.13	0.09
201107	NaN	1.00	0.43	0.27	0.19	0.17	0.11

201108	NaN	NaN	1.00	0.48	0.28	0.23	0.14
201109	NaN	NaN	NaN	1.00	0.43	0.29	0.15
201110	NaN	NaN	NaN	NaN	1.00	0.48	0.19
201111	NaN	NaN	NaN	NaN	NaN	1.00	0.26
201112	NaN	NaN	NaN	NaN	NaN	NaN	1.00

We can see that first month retention rate became better recently (don't take Dec '11 into account) and in almost 1 year, only 7% of our customers retain with us.

2 Summary

In this notebook defined important metrics for the company and calculated / analysed them with Python. * Monthly Revenue * Monthly Active Customer * Monthly Order Count * Average Revenue per Order * New Customer Ratio * Monthly Retention Rate * Cohort Based Retention Rate

In next part we'll try to segment our base to see who are our best customers.