



CUSTOMER SEGMENTATION IN PYTHON

Customer Segmentation in Python

Karolis Urbonas

Head of Data Science, Amazon



About me



- Head of Data Science at Amazon
- 10+ years experience with analytics and ML
- Worked in eCommerce, banking, consulting, finance and other industries



Prerequisites

- `pandas` **library**
- `datetime` **objects**
- **basic plotting with** `matplotlib` **or** `seaborn`
- basic knowledge of k-means clustering



What is Cohort Analysis?

- Mutually exclusive segments - cohorts
- Compare metrics across **product** lifecycle
- Compare metrics across **customer** lifecycle



Types of cohorts

- Time cohorts
- Behavior cohorts
- Size cohorts

Elements of cohort analysis

- Pivot table

[illegible]

Elements of cohort analysis

- Pivot table
- Assigned cohort in **rows**

[illegible]

Elements of cohort analysis

- Pivot table
- Assigned cohort in **rows**
- Cohort Index in **columns**

[illegible]

Elements of cohort analysis

- Pivot table
- Assigned cohort in **rows**
- Cohort Index in **columns**
- Metrics in the **table**

[illegible]

Elements of cohort analysis

- First cohort was acquired in December 2010

[illegible]

Elements of cohort analysis

- First cohort was acquired in December 2010
- Last cohort was acquired in December 2011

[illegible]



CUSTOMER SEGMENTATION IN PYTHON

Explore the cohort table



CUSTOMER SEGMENTATION IN PYTHON

Time cohorts

Karolis Urbonas

Head of Data Science, Amazon

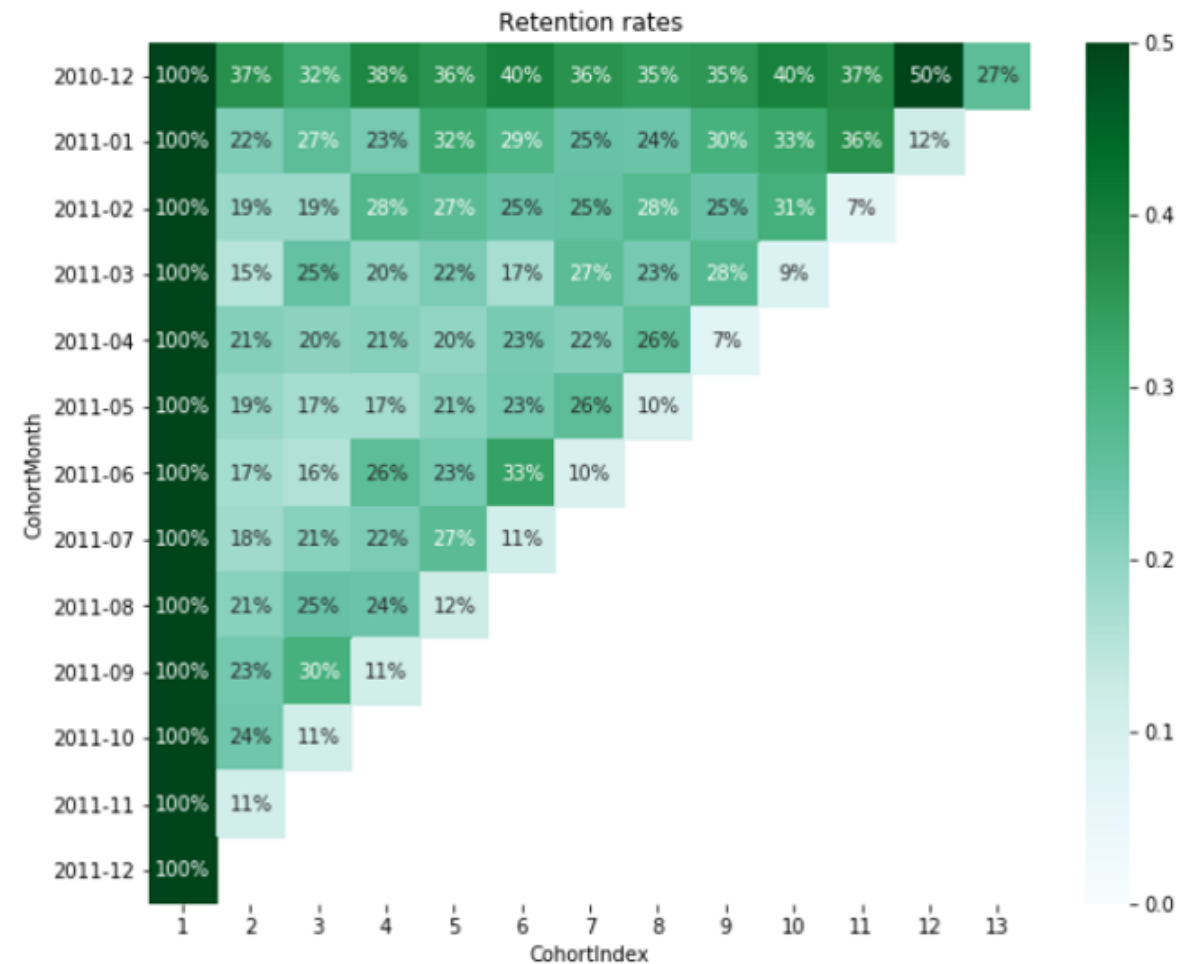
Cohort analysis heatmap

Rows:

- First activity
- Here - month of acquisition

Columns:

- Time since first activity
- Here - months since acquisition



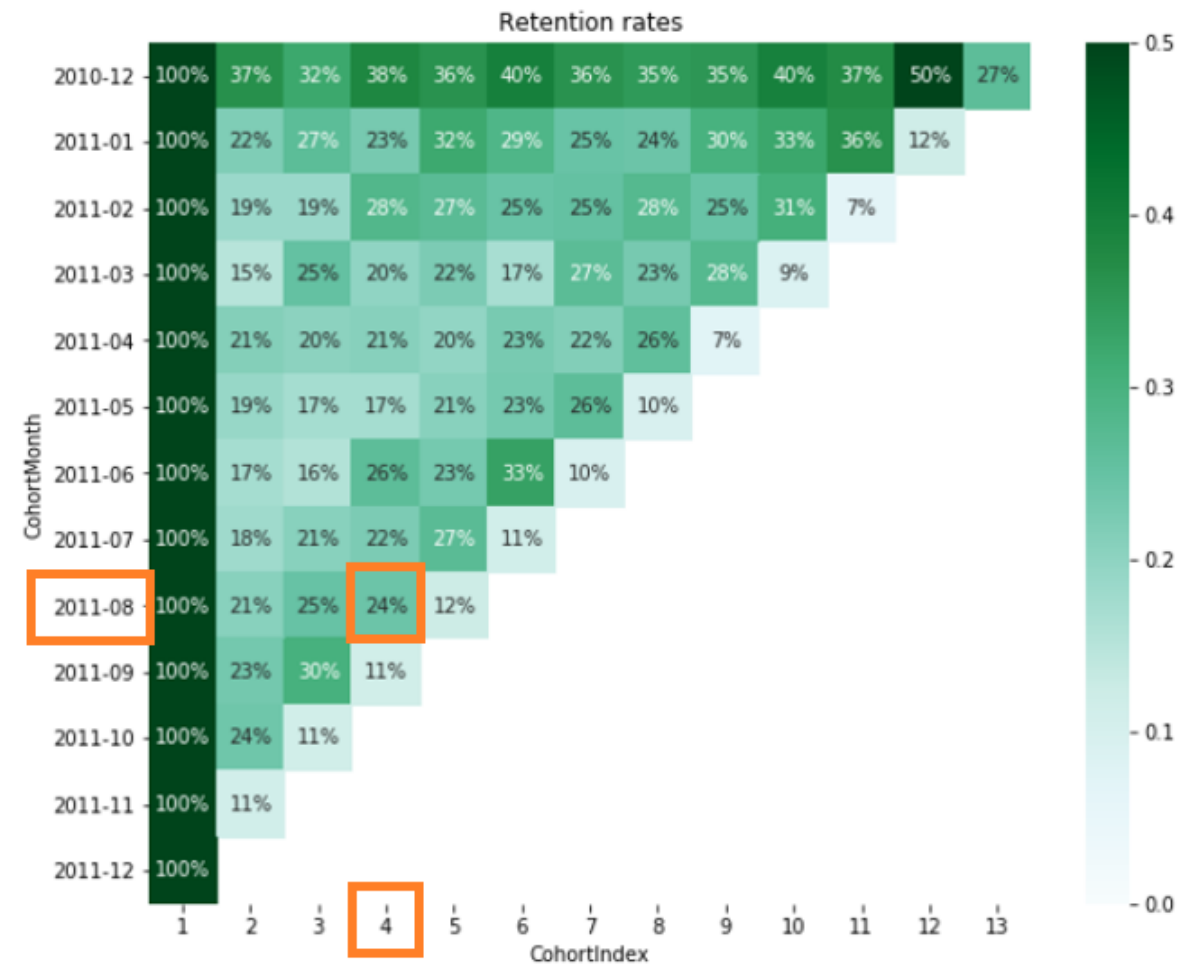
Cohort analysis heatmap

Rows:

- First activity
- Here - month of acquisition

Columns:

- Time since first activity
- Here - months since acquisition



Online Retail data

Over 0.5 million transactions from a UK-based online retail store.

We will use a randomly sampled 20% subset of this dataset throughout the course.



Online Retail Data Set

Download: [Data Folder](#), [Data Set Description](#)



Top 5 rows of data

```
online.head()
```

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25 08:26:00	2.10	14286	United Kingdom
577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20 11:56:00	1.45	16360	United Kingdom
560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14 13:35:00	3.75	13933	United Kingdom
578307	72349B	SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23 15:53:00	2.10	17290	United Kingdom
554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25 13:36:00	5.95	17663	United Kingdom



Assign acquisition month cohort

```
def get_month(x): return dt.datetime(x.year, x.month, 1)
```

```
online['InvoiceMonth'] = online['InvoiceDate'].apply(get_month)
```

```
grouping = online.groupby('CustomerID')['InvoiceMonth']
```

```
online['CohortMonth'] = grouping.transform('min')
```

```
online.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	InvoiceMonth	CohortMonth
416792	572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25 08:26:00	2.10	14286.0	United Kingdom	2011-10-01	2011-04-01
482904	577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20 11:56:00	1.45	16360.0	United Kingdom	2011-11-01	2011-09-01
263743	560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14 13:35:00	3.75	13933.0	United Kingdom	2011-07-01	2011-07-01
495549	578307	72349B	SET/6 PURPLE BUTTERFLY T- LIGHTS	1	2011-11-23 15:53:00	2.10	17290.0	United Kingdom	2011-11-01	2011-11-01
204384	554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25 13:36:00	5.95	17663.0	United Kingdom	2011-05-01	2011-02-01



Extract integer values from data

Define function to extract `year`, `month` and `day` integer values.

We will use it throughout the course.

```
def get_date_int(df, column):  
    year = df[column].dt.year  
    month = df[column].dt.month  
    day = df[column].dt.day  
    return year, month, day
```



Assign time offset value

```
invoice_year, invoice_month, _ = get_date_int(online, 'InvoiceMonth')
cohort_year, cohort_month, _ = get_date_int(online, 'CohortMonth')
```

```
years_diff = invoice_year - cohort_year
months_diff = invoice_month - cohort_month
```

```
online['CohortIndex'] = years_diff * 12 + months_diff + 1
online.head()
```

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	InvoiceMonth	CohortMonth	CohortIndex
416792	572558	22745 POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25 08:26:00	2.10	14286.0	United Kingdom	2011-10-01	2011-04-01	7
482904	577485	23196 VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20 11:56:00	1.45	16360.0	United Kingdom	2011-11-01	2011-09-01	3
263743	560034	23299 FOOD COVER WITH BEADS SET 2	6	2011-07-14 13:35:00	3.75	13933.0	United Kingdom	2011-07-01	2011-07-01	1
495549	578307	72349B SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23 15:53:00	2.10	17290.0	United Kingdom	2011-11-01	2011-11-01	1
204384	554656	21756 BATH BUILDING BLOCK WORD	3	2011-05-25 13:36:00	5.95	17663.0	United Kingdom	2011-05-01	2011-02-01	4

Count monthly active customers from each cohort

```
grouping = online.groupby(['CohortMonth', 'CohortIndex'])
```

```
cohort_data = grouping['CustomerID'].apply(pd.Series.nunique)
```

```
cohort_data = cohort_data.reset_index()
```

```
cohort_counts = cohort_data.pivot(index='CohortMonth',  
                                   columns='CohortIndex',  
                                   values='CustomerID')
```

```
print(cohort_counts)
```

[illegible]



CUSTOMER SEGMENTATION IN PYTHON

Your turn to build some cohorts!



CUSTOMER SEGMENTATION IN PYTHON

Calculate cohort metrics

Karolis Urbonas

Head of Data Science, Amazon

Customer retention: cohort_counts table

- How many customers originally in each cohort in the `cohort_counts` table?

[illegible]

Customer retention: cohort_counts table

- How many customers originally in each cohort?
- How many of them were active in following months?

[illegible]

Calculate Retention rate

1. Store the first column as `cohort_sizes`

```
cohort_sizes = cohort_counts.iloc[:,0]
```

2. Divide all values in the `cohort_counts` table by `cohort_sizes`

```
retention = cohort_counts.divide(cohort_sizes, axis=0)
```

3. Review the retention table

```
retention.round(3) * 100
```

Retention table

[illegible]



Other metrics

```
grouping = online.groupby(['CohortMonth', 'CohortIndex'])

cohort_data = grouping['Quantity'].mean()

cohort_data = cohort_data.reset_index()

average_quantity = cohort_data.pivot(index='CohortMonth',
                                      columns='CohortIndex',
                                      values='Quantity')

average_quantity.round(1)
```

Average quantity for each cohort

[illegible]



CUSTOMER SEGMENTATION IN PYTHON

**Let's practice on other
cohort metrics!**



CUSTOMER SEGMENTATION IN PYTHON

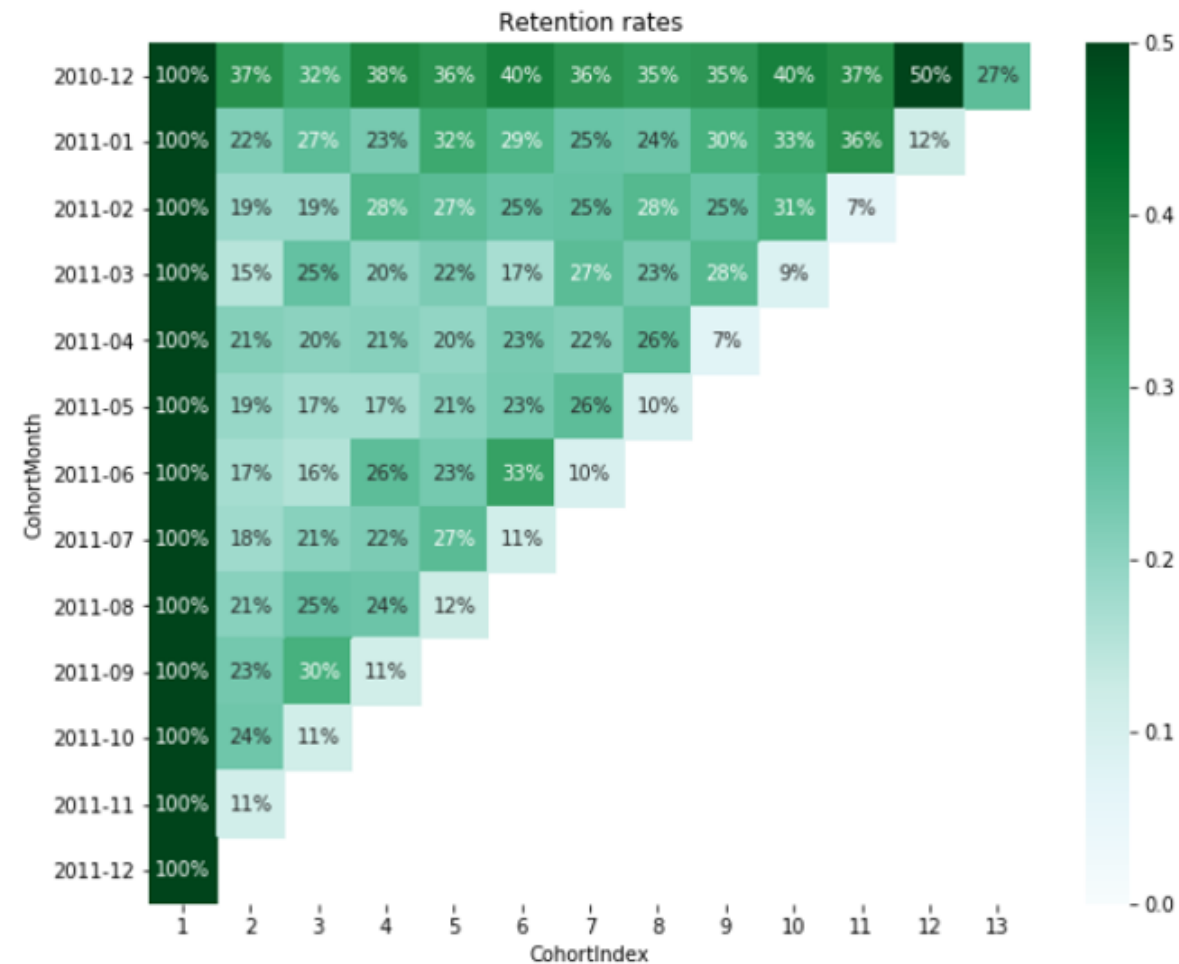
Cohort analysis visualization

Karolis Urbonas

Head of Data Science, Amazon

Heatmap

- Easiest way to visualize cohort analysis
- Includes both data and visuals
- Only few lines of code with `seaborn`





Build the heatmap

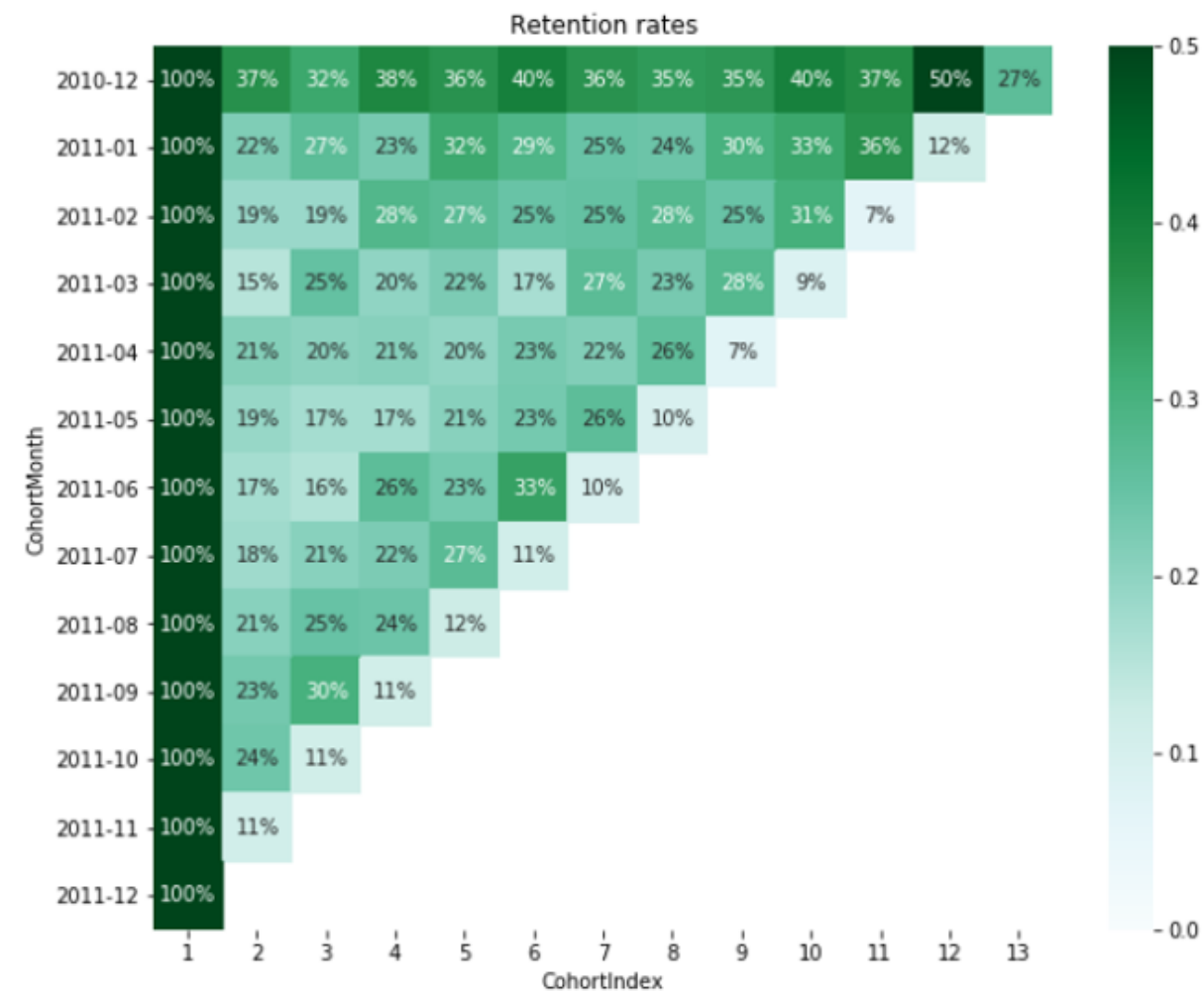
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 8))

plt.title('Retention rates')

sns.heatmap(data = retention,
            annot = True,
            fmt = '.0%',
            vmin = 0.0,
            vmax = 0.5,
            cmap = 'BuGn')

plt.show()
```





CUSTOMER SEGMENTATION IN PYTHON

Practice visualizing cohorts