M Case study

Aziz Mamatov, 12/4/2019

Deliverables and reference materials

Analyze churn as shown in the dataset, and provide a set of recommendations as to how M might approach this issue.

- How do different groups of users behave?
- What factors appear to be strongly correlated with churn?
- How much revenue could we make if we prevented churn?
- What are some ways we could capture that revenue?

Reference materials:

https://app.mode.com/editor/amamatov/reports/a7cb4172597c/notebook

Python file will be sent separately

Executive summary

Sample consists of transactional data for the last 30 days plus account age and integrations columns. Data is not normally distributed and as such, applying arithmetic means will not accurately represent the data, i.e. Excel won't help much

It is possible to heuristically segment the customers based on age, number of seats and transactions, however algorithmical approach (k-means) may be more favorable due to inclusion of all available features and automation. **K-means algorithm** was applied and five clusters identified which cannot easily characterized by any single feature. *See the slide.*

We also can calculate churned % for each cluster and calculate Kohonen cart. Segments are useful in determining overall customer service approach to the population divided by the segments

Based on **logistic regression**, two features strongly correlated with the churn: number of seats (negative correlation) and DAU per seat (positive correlation). *See the slide*

If we prevented the churn we could possible make at least as much revenue - *see the slide.* As the revenue number itself is not provided, it is not possible to determine the exact amount

To recapture lost revenue, we need to try to re-engage the lost customers by extending discounts, specific training, and possibly - specific integrations (however, it seems that number of integrations is the same for both types). It seems that the large customers (with many seats) and old customers (by age) are amongst those churned and such accounts should be paid careful attention as to why they churned   .

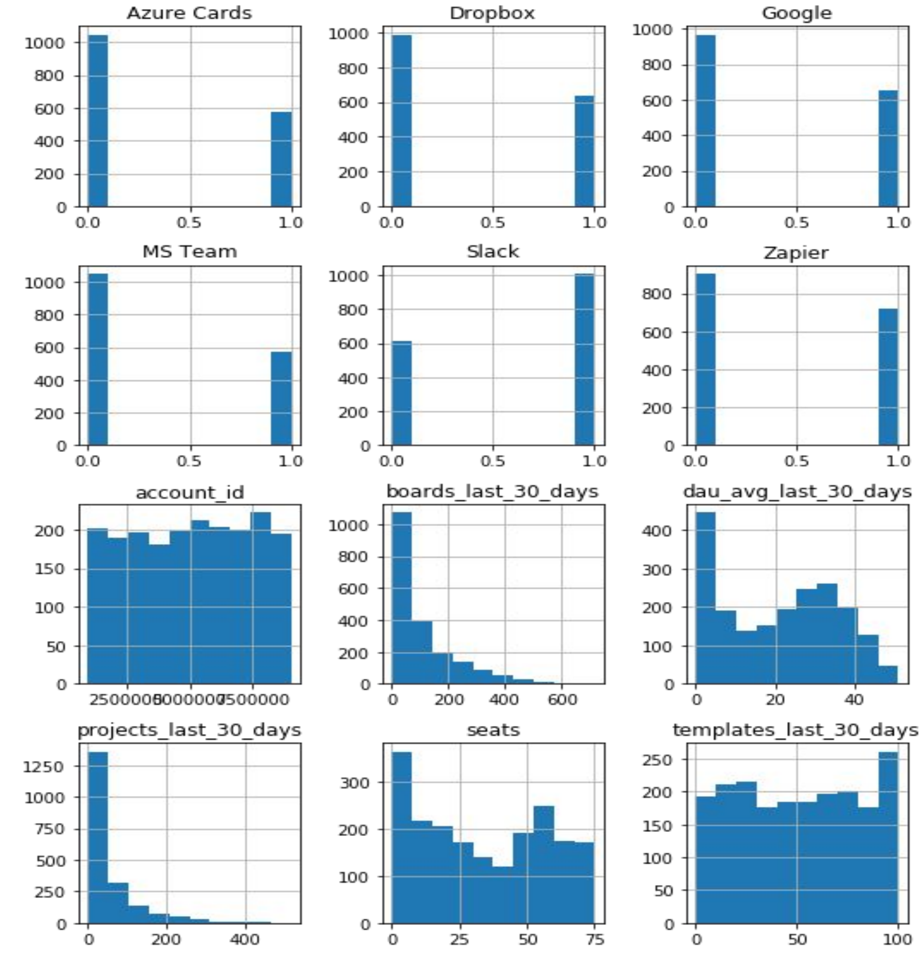
DAU/Seats (utilization ratio) under 60% has the highest probability of churn

Summary statistics:

- 2K unique ID
- 2 categorical features, 7 numerical features, one index (account ID)
- Numerical features are <u>not</u> distributed normally and as such arithmetic means should not be used
- Churned and paying customers are distributed 49.4% vs 50.6%
- Sample size seems to be sufficient to extrapolate to the population for the majority of features except the rare ones (outliers or rare events)
- However, it is important to ensure that the sample is truly random before extrapolation

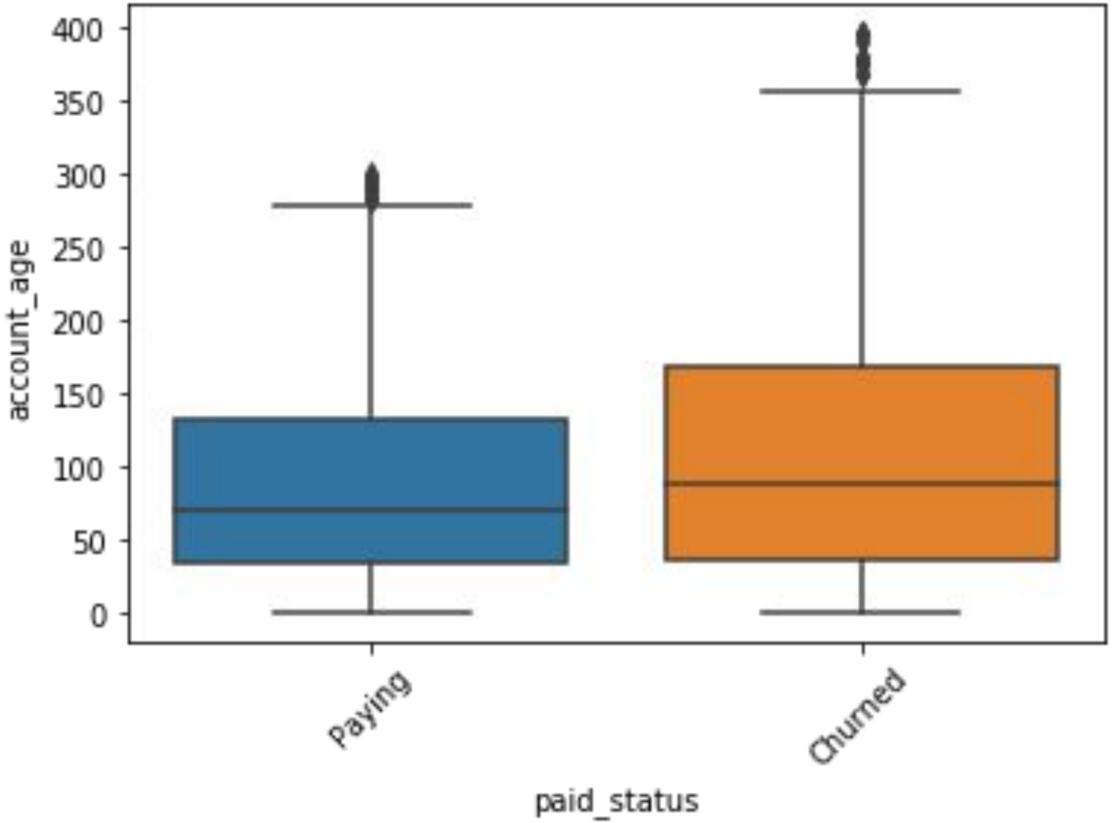
index	account_id	account_age	seats	boards_last_30_days	projects_last_30_days	dau_avg_last_30_days	integrations_count	templates_last_30_days
count	2.000000e+03	2000.000000	2000.00000	2000.000000	2000.000000	2000.000000	2000.000000	2000.000000
mean	5.057934e+06	100.195000	33.83000	107.242000	52.810500	21.144500	2.437500	50.453000
std	2.314520e+06	82.610455	23.32248	119.667391	74.486393	14.261603	1.743449	29.765093
min	1.001950e+06	1.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.025739e+06	35.000000	12.00000	19.000000	4.000000	7.000000	1.000000	24.000000
50%	5.146980e+06	78.000000	32.00000	62.000000	24.000000	22.000000	2.000000	51.000000
75%	7.050892e+06	145.000000	55.00000	156.000000	71.000000	33.000000	4.000000	77.000000
max	8.997315e+06	395.000000	75.00000	717.000000	518.000000	51.000000	5.000000	100.000000

Distribution of data (*dummy variables were created for integrations column*)



Demographic data

- the only available feature pertaining to the age of the customer
- Significant outliers for Paying and Churned with the similar averages



Lost revenue

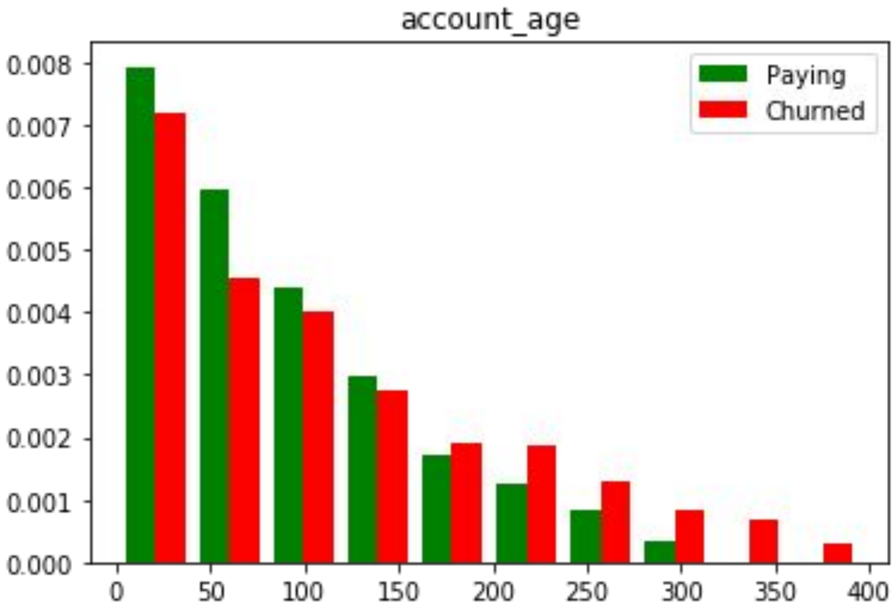
- Significant amount of lost revenue judging by the number of seats.
- Number of churned accounts is a bit larger than paying but number of churned seats is much higher.
- However, pricing may be the same for the range of the seats and as I don’t know about the brackets it is hard to estimate the lost revenue
- However, it is still significant and most probably equals the revenue from paying accounts

```
Number of churned seats:  36963.0
Number of paying seats:  30697.0
Mean of churned seats:  37.411943
Number of paying seats:  30.33300
Count of churned seats:  988
Count of paying seats:  1012
```

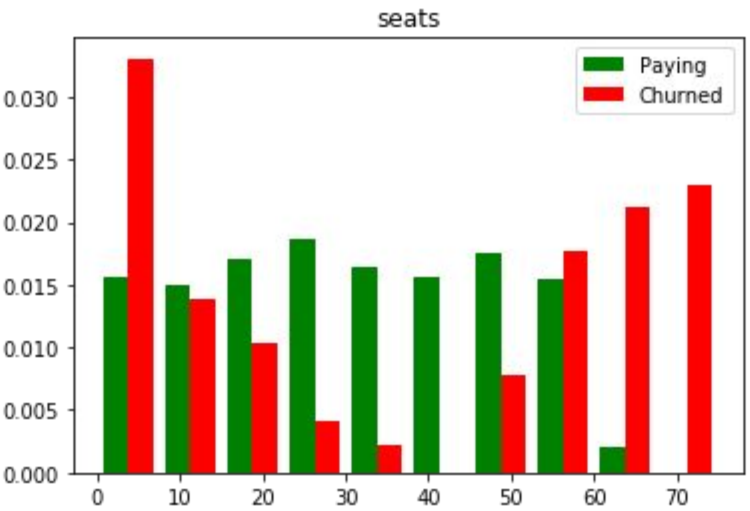
Features’ histograms - Paying vs Churned

We can use this data to infer dependencies

All histograms are normalized, i.e. area under curve equals 1 (integral)



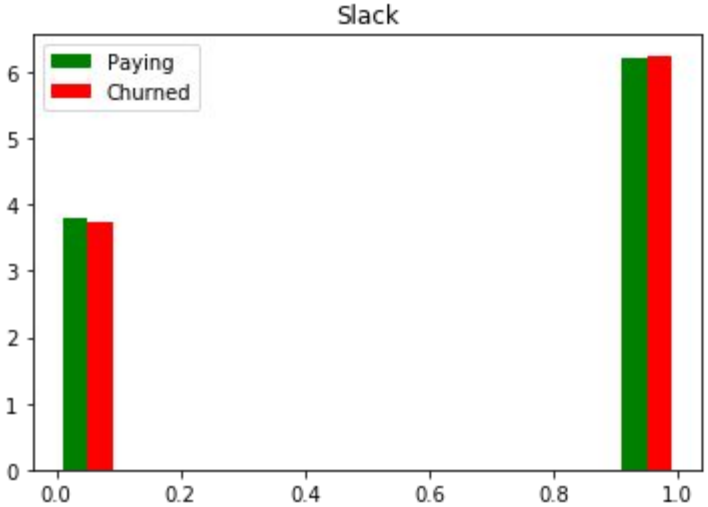
With the age, number of churned exceeds paying ones.
Especially true for the ‘eldest’ customers



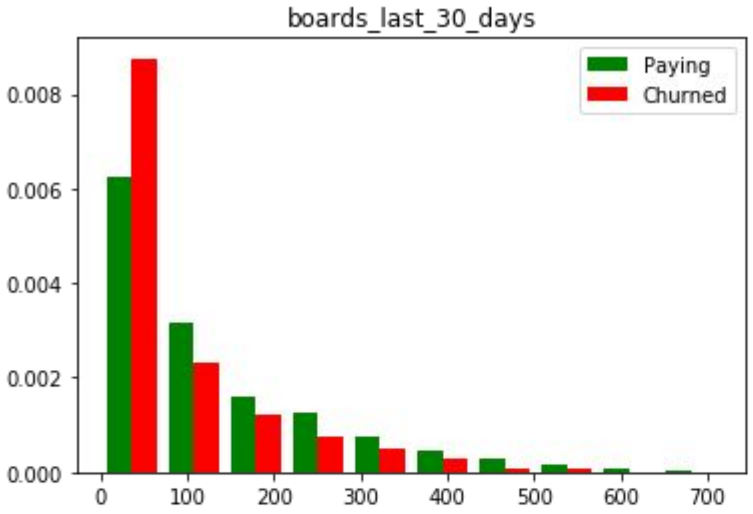
Customers are churned either if the number of their seats close to min or max.

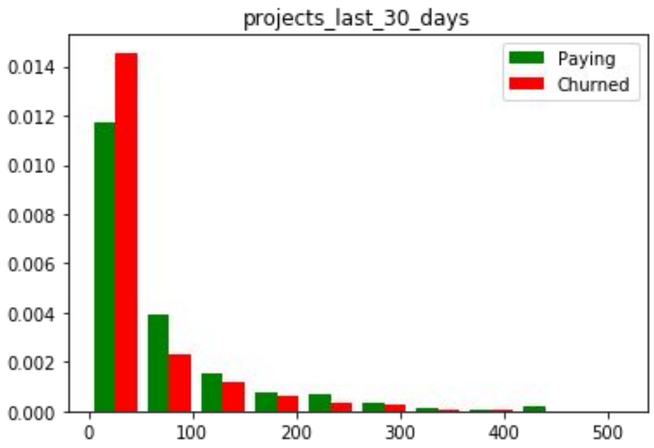
There are significant number of accounts with 0 seats, does it mean no onboarding and/or no use of product?

Largest accounts (by seat no.) churn, so we may need to pay attention to both accounts



Among integrations there was no meaningful correlation with churn. Only note is that Slack was the most popular integration and the majority of users had it

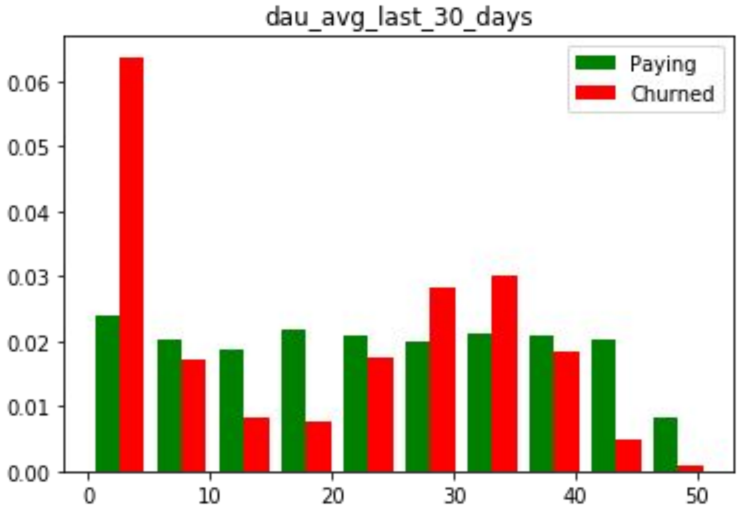


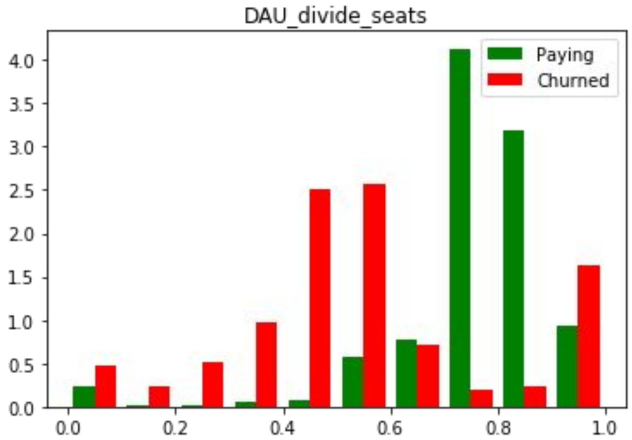


Largest churn at 0 boards (i.e. no activity) but there are some paying customers with no activity during the last 30 days of activity as well

Obviously engagement should be increased, as it is important for the users to actually use the product

It is possible that customers with zero engagement during the last 30 days will cancel their subscription as they may have forgotten to cancel it



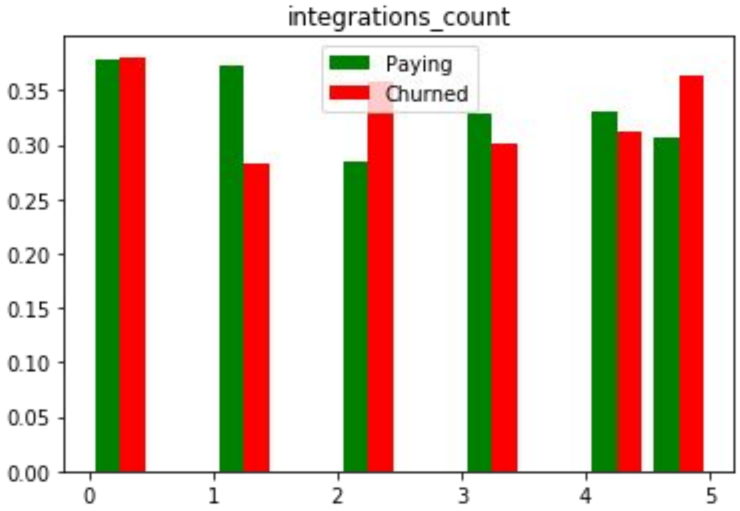


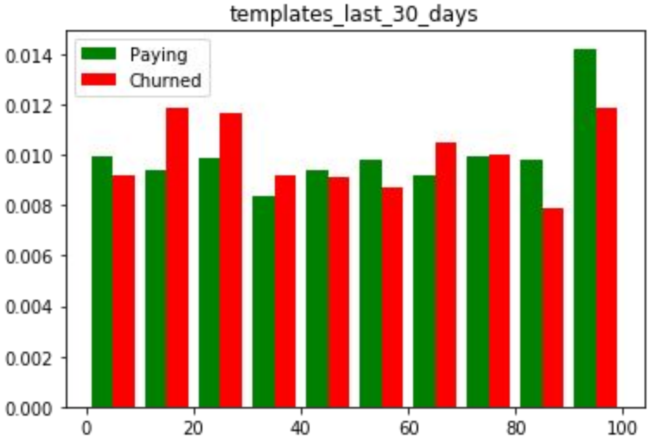
If zero number of DAU, then big chance of churn, however, DAU of 30 to 40 also have the highest churn

So, someone may decide to churn even if there are lots of DAU like 30-40.

I created a new feature: DAU/seats (engagement ratio) and we can see that the highest churn is up to 60% mark of DAU/seats, HOWEVER, there are anomalies in 60% and near 100%.

Even large accounts may churn and we need to pay special attention to them





Number of integration does not seem to matter.

Similarly number of templates used also does not seem to matter much

Logistic regression and churn

Logistic regression have been performed on the data (after creating dummy features from categorical data and normalizing numerical data) and the following accuracy appeared:

- *ROC_AUC_train: 0.956 ROC_AUC_test: 0.968*

The following features had the highest coefficients:

- *Seats: -7.03 (the more seats the higher chance of churn)*
- *DAU_avg_last_30_days: 6.88 (the higher DAU the lesser chance of churn)*

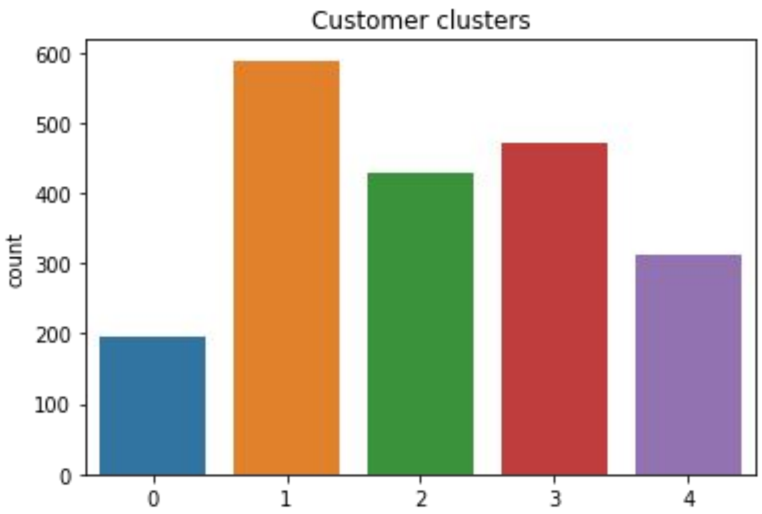
Customer segmentation and k-means

Customer segmentation can be created using the transactional, demographic, geographic and psychographic information.

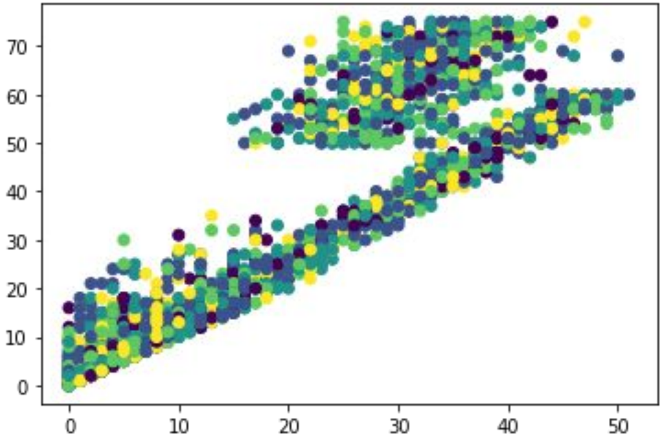
In our case, we have transactional information together with sort of demographic information (account_age)

As such, the segmentation can be used for determining a probability to churn, i.e. whether the customer is active, less active, dormant or churned

I will use k-means algorithm for segmentation, by first estimating the number of clusters and second by applying the algorithm itself (on data with dummies and normalized)



Here is the countplot of segments created by k-means



Also, visualization of scatter plot, seats vs dau_avg_last_30_days

As it can be seen, clustering is multidimensional and as such is not visible clearly on 2D plot

Stat summary of segments 4 and 0: not much visible difference in selected features and as such, it will be hard to describe the segments heuristically. Kohonen carts can be used in this case

```python
df3[['account_age','seats','dau_avg_last_30_days']][df3['label']==4].describe()
```

Py

index	account_age	seats	dau_avg_last_30_days
count	313.000000	313.000000	313.000000
mean	83.370607	33.479233	21.031949
std	64.892688	23.247009	14.336577
min	2.000000	0.000000	0.000000
25%	34.000000	11.000000	7.000000
50%	70.000000	31.000000	22.000000
75%	117.000000	54.000000	32.000000
max	314.000000	75.000000	49.000000

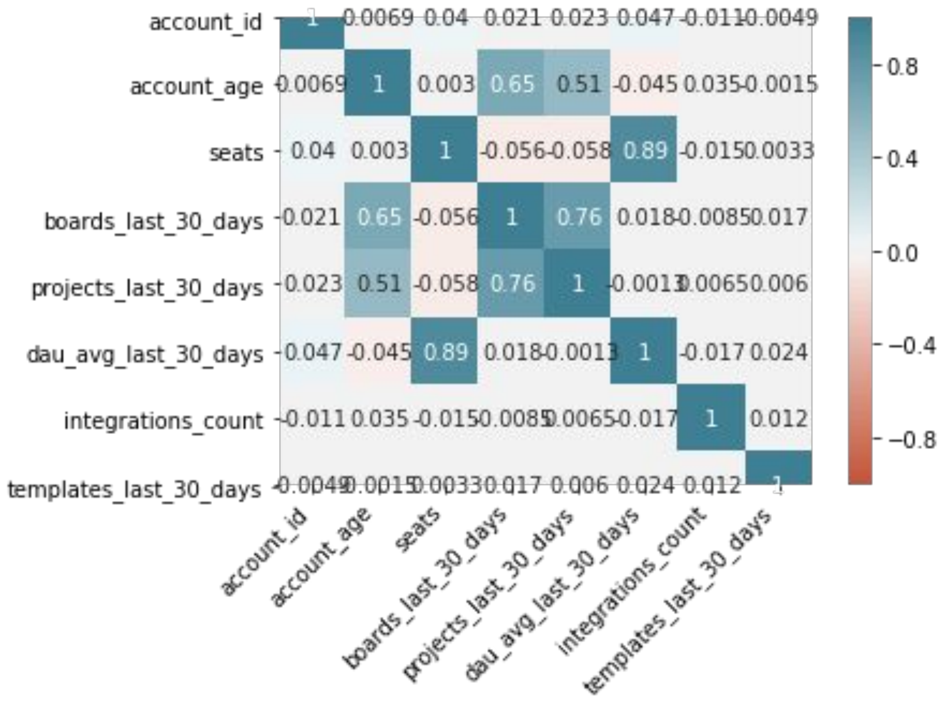
```python
df3[['account_age','seats','dau_avg_last_30_days']][df3['label']==0].describe()
```

Py

index	account_age	seats	dau_avg_last_30_days
count	197.000000	197.000000	197.000000
mean	232.218274	31.979695	20.730964
std	65.919231	23.161580	14.941509
min	87.000000	0.000000	0.000000
25%	179.000000	10.000000	5.000000
50%	236.000000	28.000000	20.000000
75%	277.000000	52.000000	[illegible]
max	395.000000	75.000000	49.000000

Ready

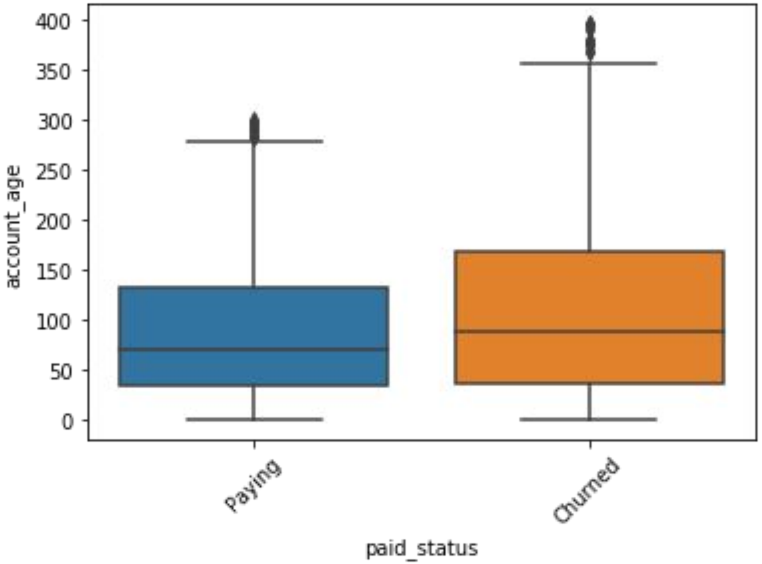
Additional slides



Correlation between the features

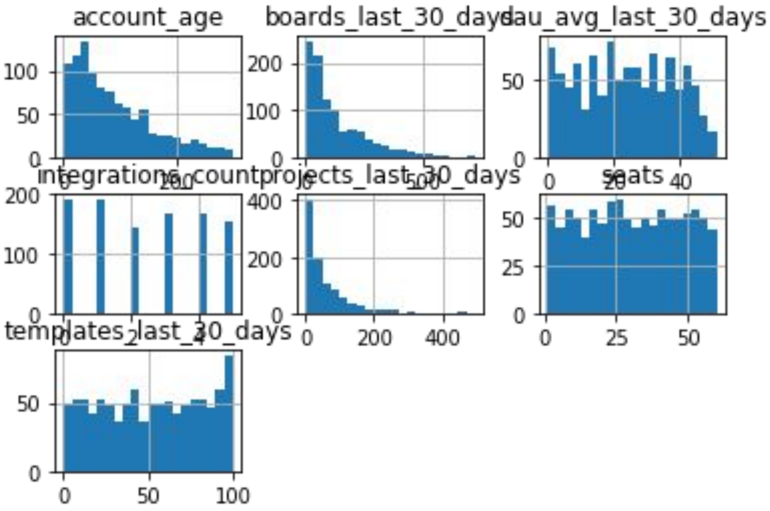
Lets figure out which features are correlated.

- 'Seats' correlate with 'dau_average_30_days': the more seats the more DAU, which is obvious.
- 'boards_last_30_days' correlates with 'projects_last_30_days': also once onboarding is done you can expect the surge in projects
- 'boards_last_30_days' correlates less strongly with 'account_age': the older the account the larger number of recent onboardings

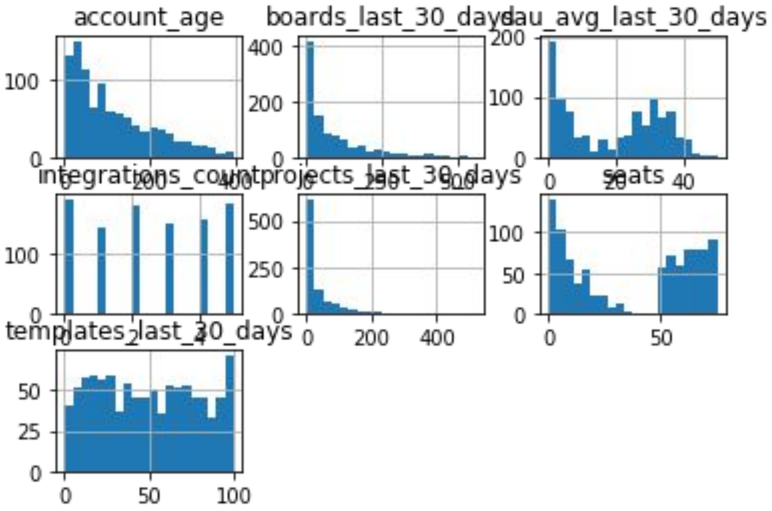


Distribution of Account_age by paying status

Paid



Churned



Visible distribution differences in Boards_last_30_days, Avg_last_30_days, seats.