KKBOX:
Let's Analyze!

By: Carlene Williams

Project Background

MAJOR TECHNOLOGIES USED:

- SQL (Postgres)
- Python (Pandas, StatsModel)
- Tableau

DATA/INFORMATION FROM:

- Kaggle.com
- KKBOX (Company)

More detailed information can be found at:
https://github.com/Carlene/KKBox_Analysis

1

KKBOX

A short history of the company

Over 10,000,000 users

Since 2005

50,000,000+ tracks

Largest database of Chinese Music

5 different countries

Taiwan, Hong Kong, Japan, Singapore and Malaysia


Microsoft
KKbox
GROUP
微軟與 KKBOX 集團
啟動全球
合作
KKbox
KKStream
KKbox
GROUP
Microsoft
Let's Go Global
Azure NO.1

Goal

Look at sampled customers and figure out the characteristics of churned users and higher payers.

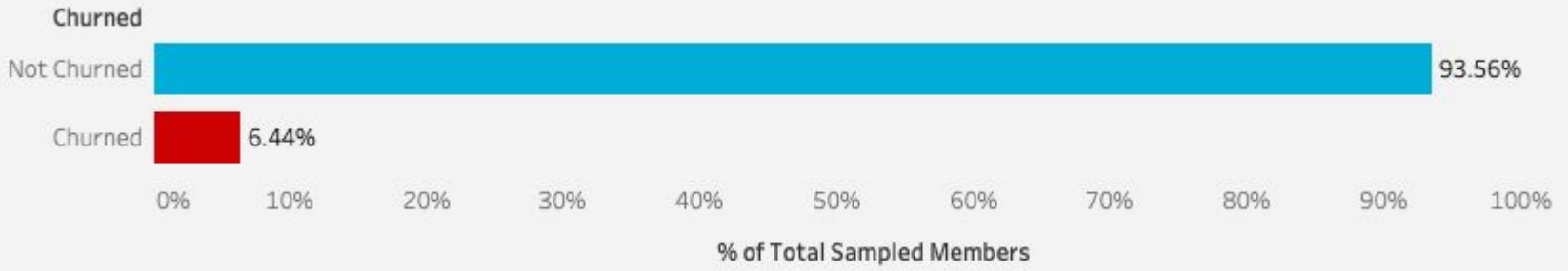
2

Who are KKBOX's customers?

Study of demographics

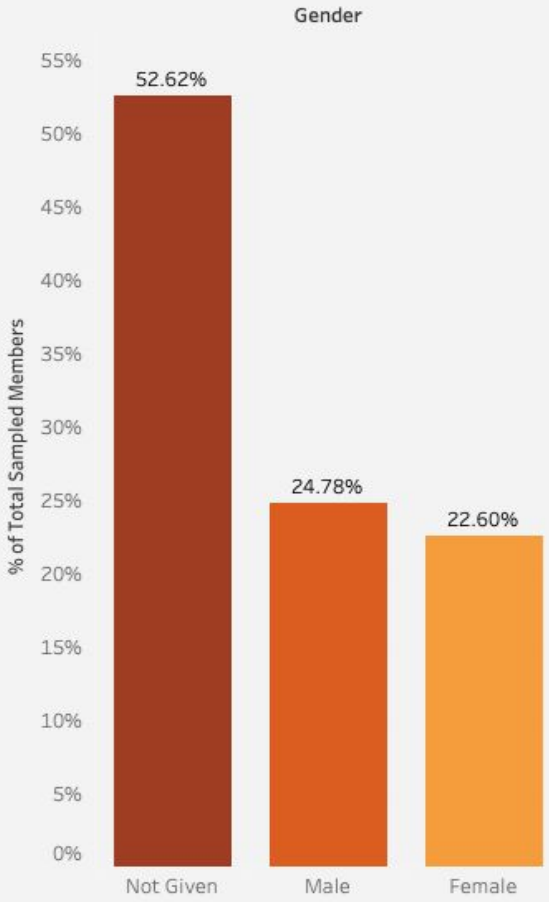
Customer Churn

- Sample of 665k subscribers in March 2017



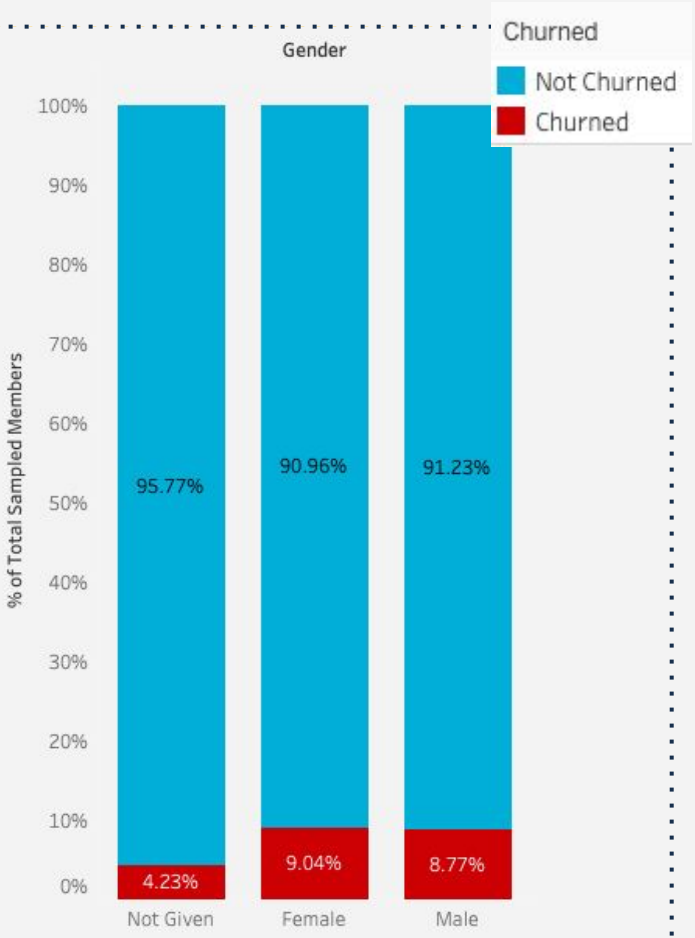
Gender

- 2% difference male and female subscribers



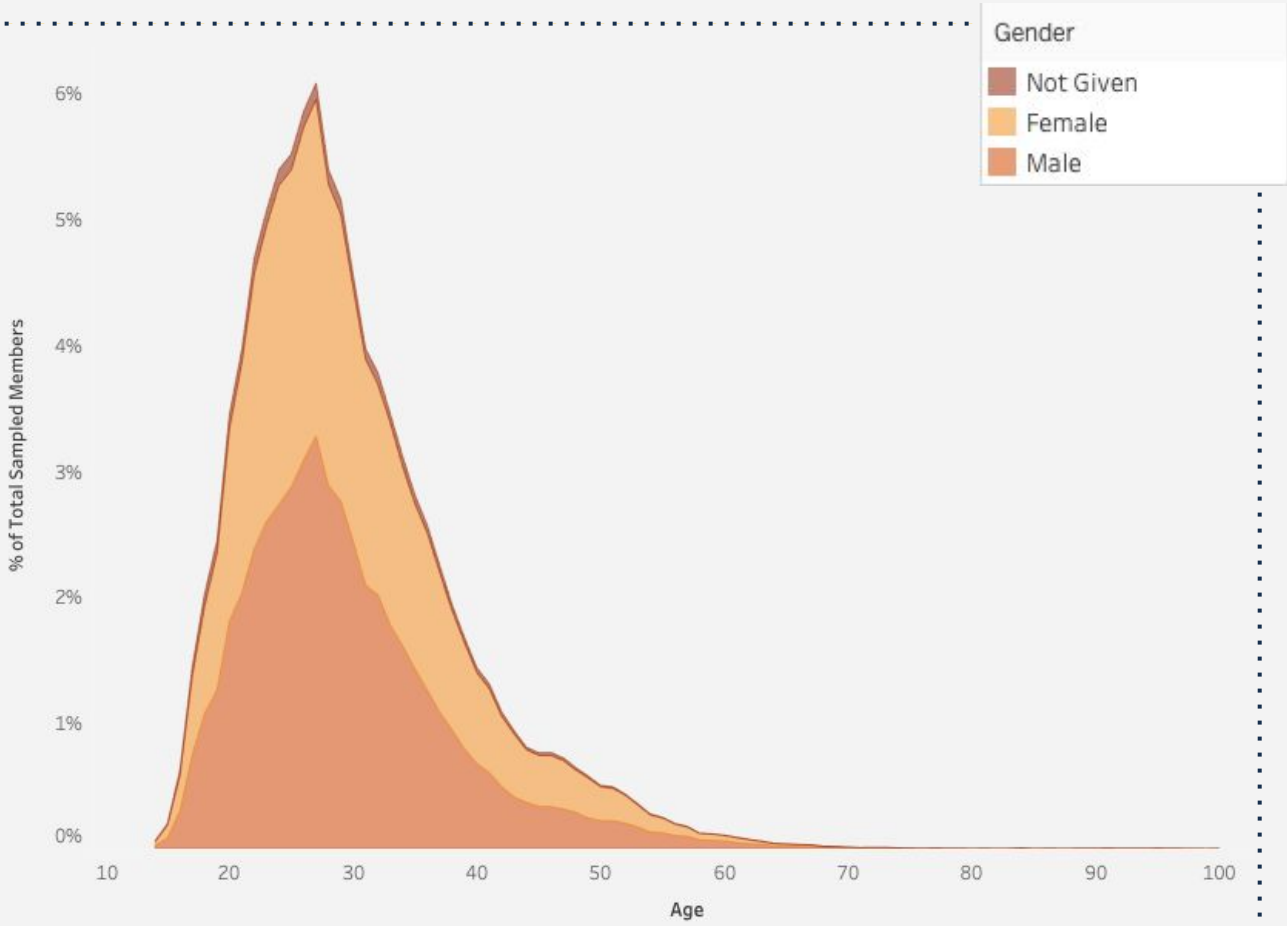
Gender

- Women churn slightly more
- Male and female subscribers churn more than those who declined to self-identify



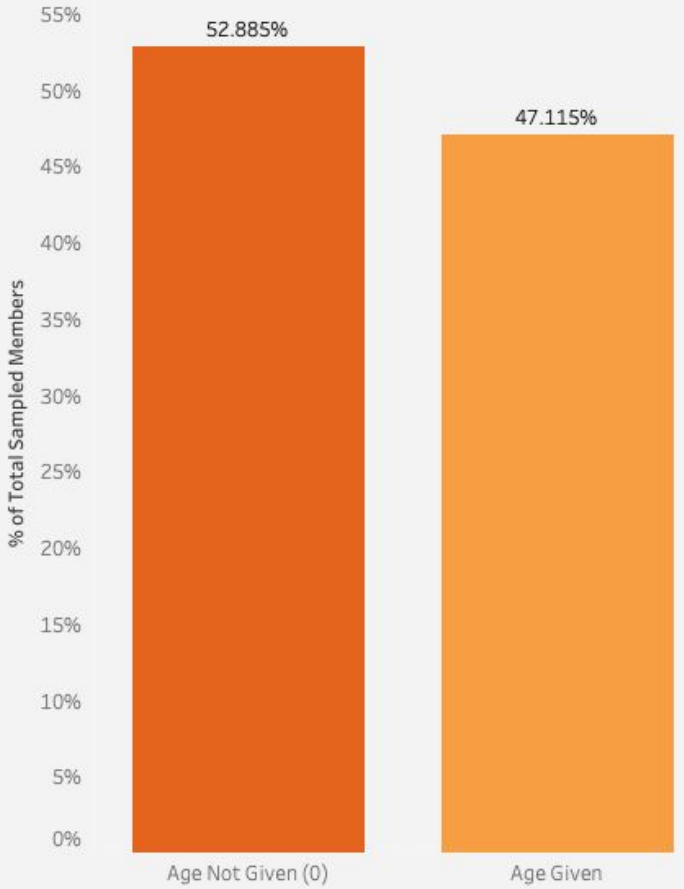
Age

- Only 1% of users give an age with no gender



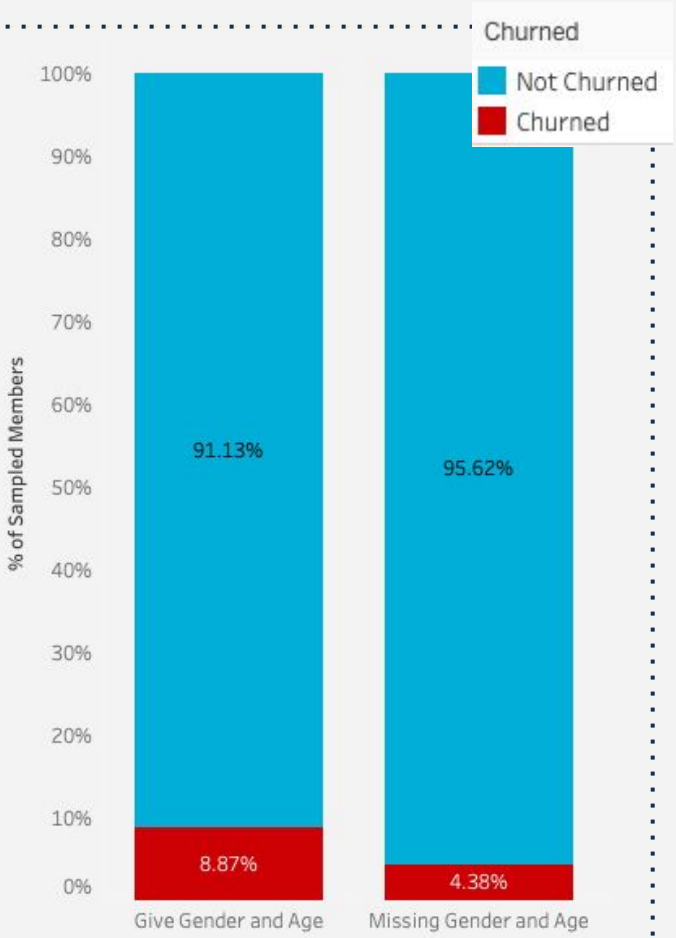
Age

- 5% more subscribers who choose not to give their age



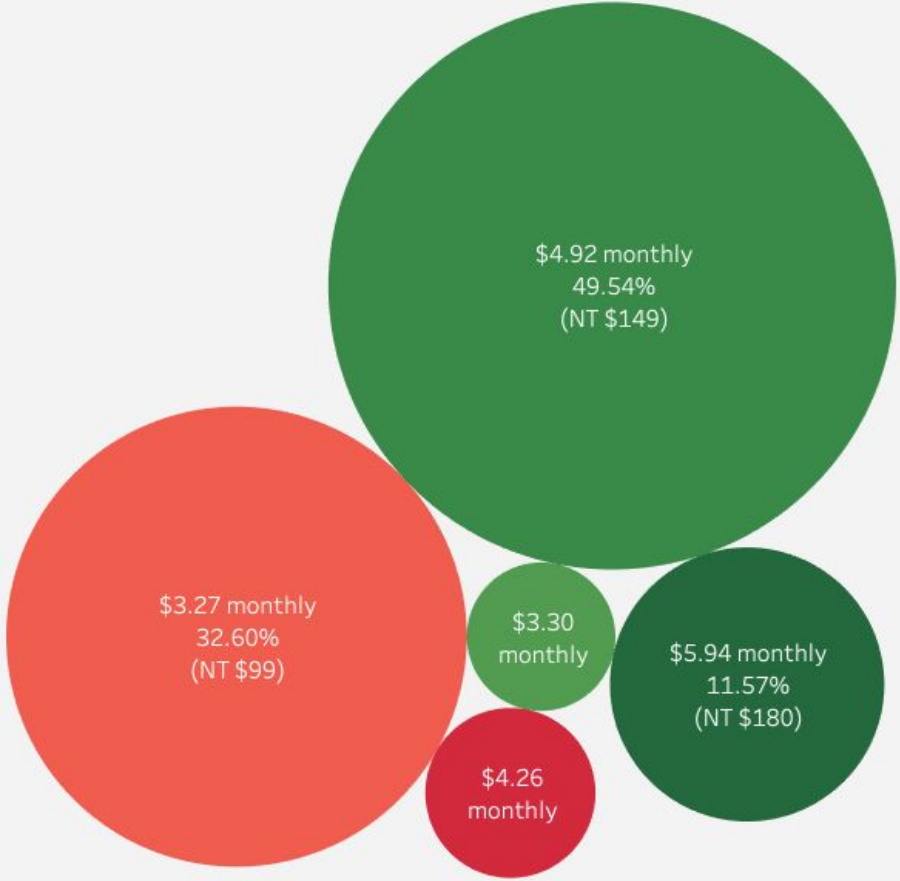
Age

- 5% more subscribers who choose not to give their age
- Users who self-identify both gender and age tend to churn more often



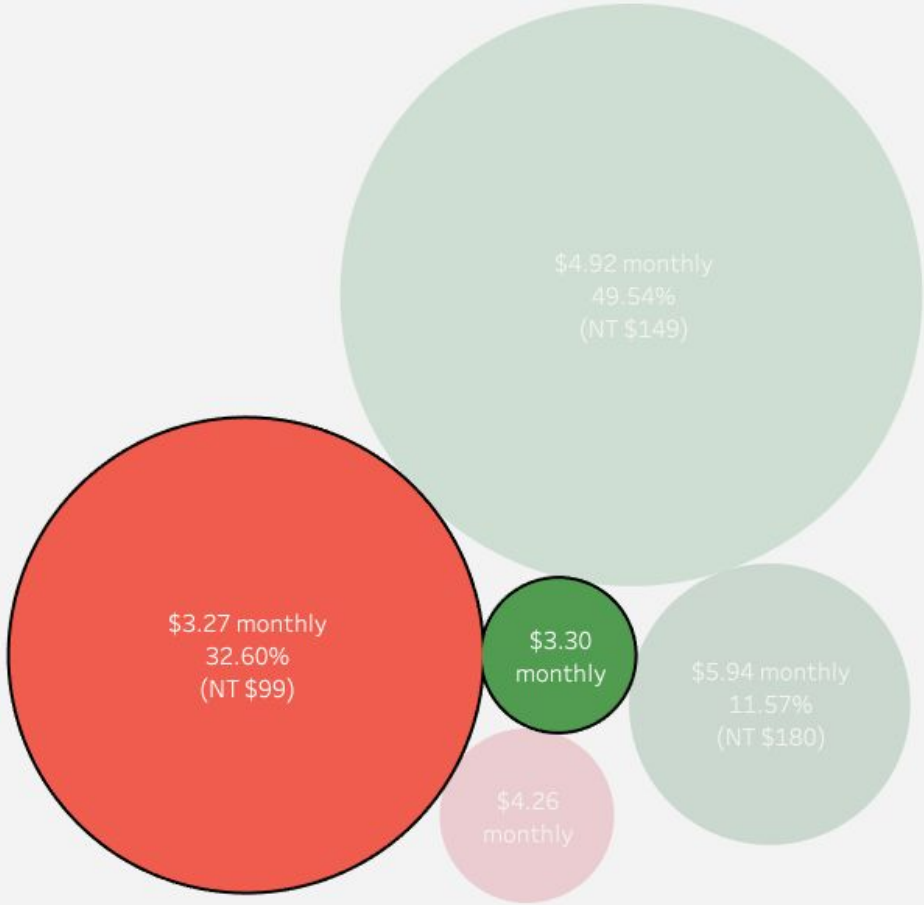
Payment Plans

- 42 Different Plans
- Top 5 plans by popularity (~98% of Subscribers)

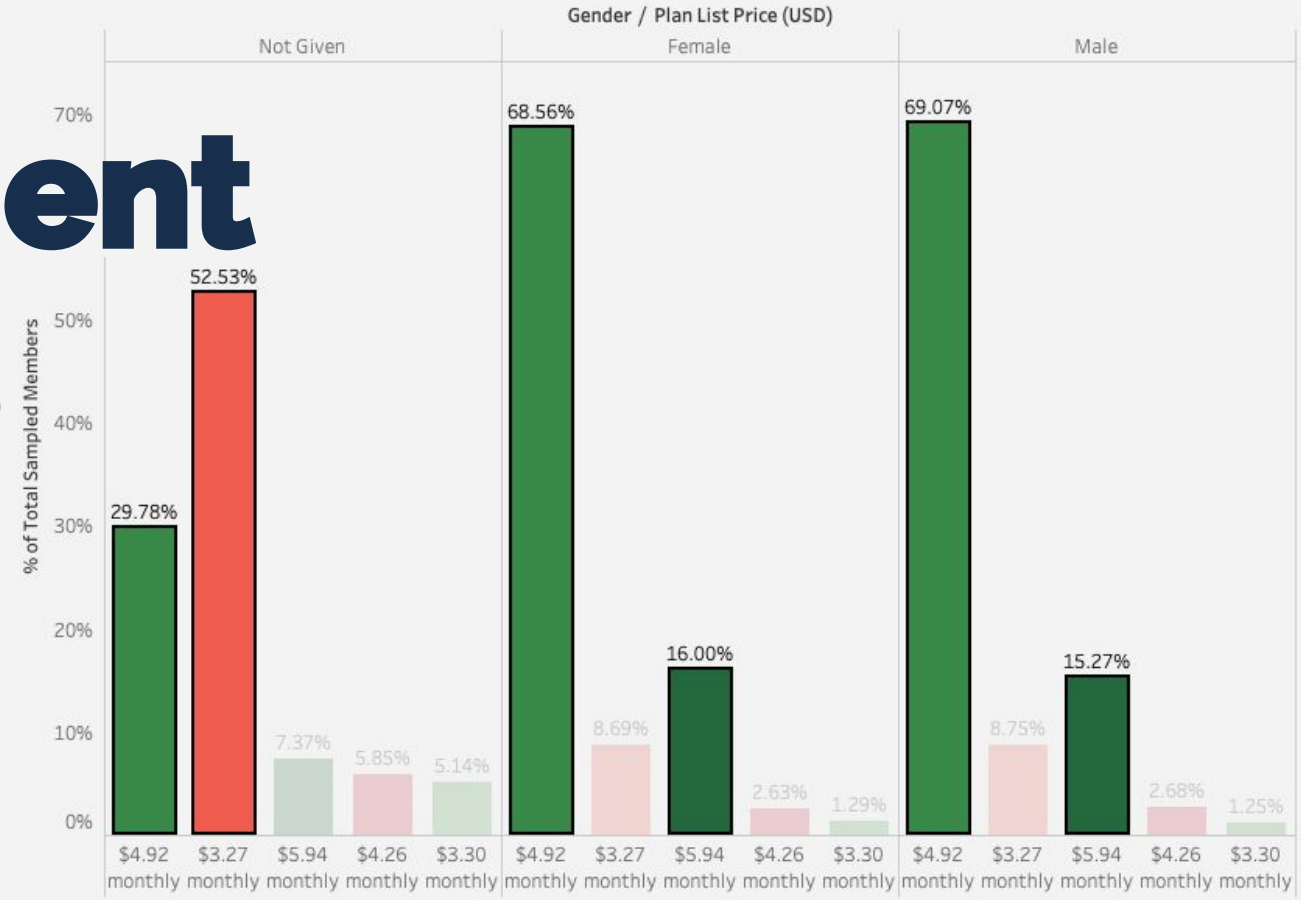


Payment Plans

- 42 Different Plans
- Top 5 plans by popularity (~98% of Subscribers)



Payment Plans

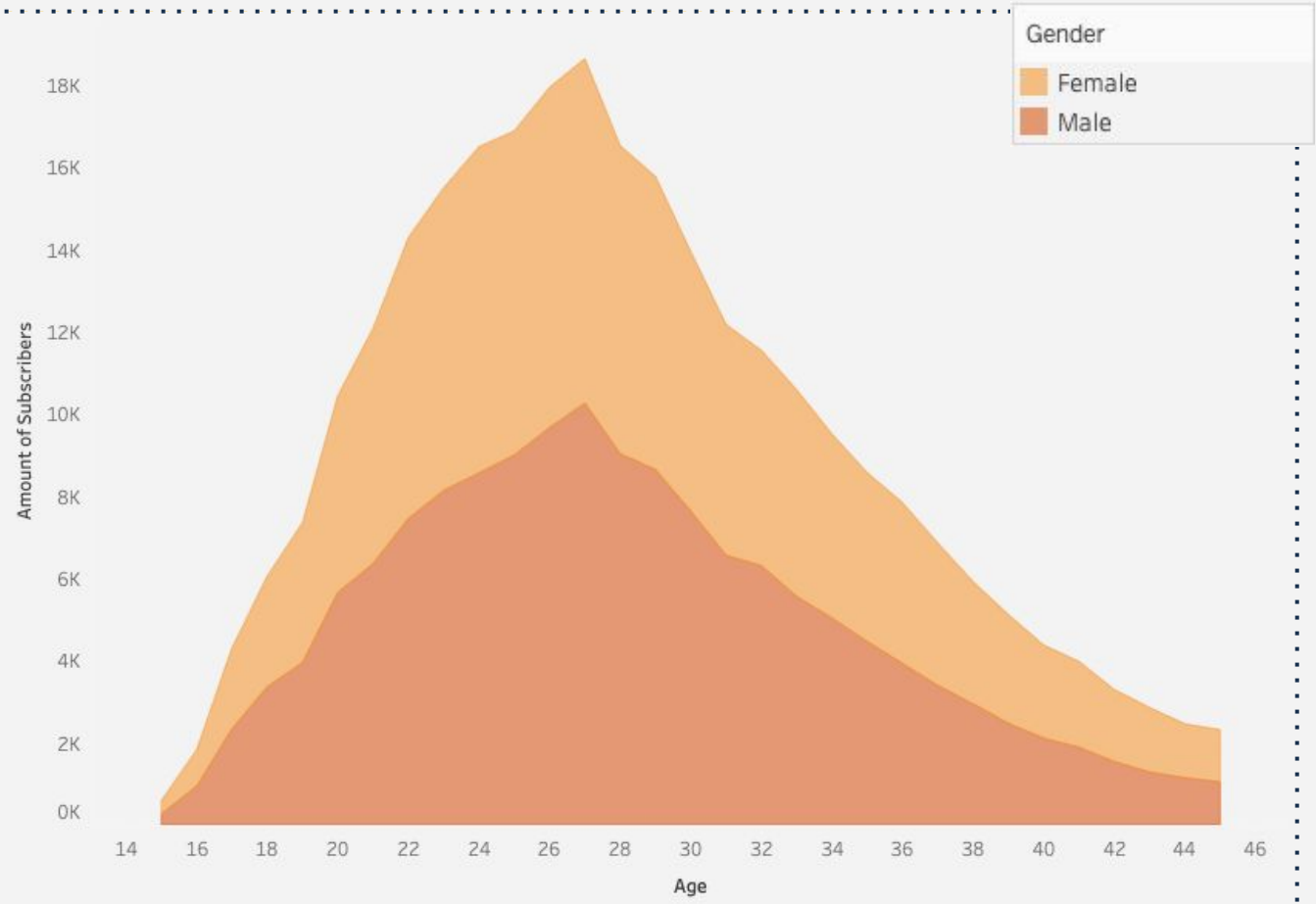


3

Hypothesis Testing

Questions/Answers

Age


Gender
Female
Male
Amount of Subscribers
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
Age

20-26 vs. 27-33

- Do not tend to pay the for the same plans
- The 20-26 tends to pay for **more** expensive plans

```
from scipy.stats import ttest_ind
stat, p = ttest_ind(adult_dist['plan_list_price'], ya_dist['plan_list_price'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

```
stat=-10.785, p=0.000
Probably different distributions
```

20-26 vs. 27-33

- Do not tend to churn at the same rate
- The 20-26 group tends to churn **more** often

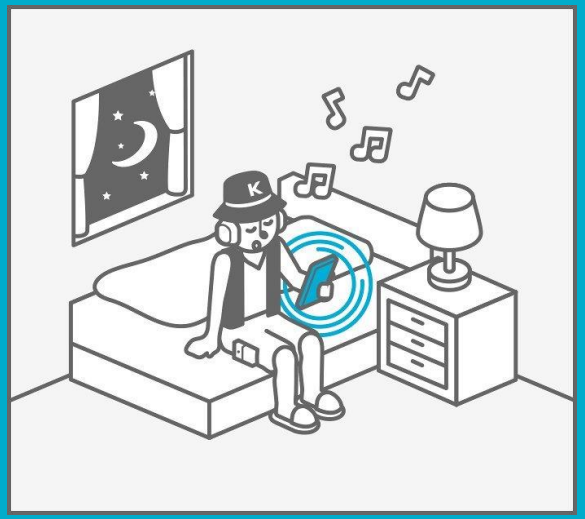
```
from scipy.stats import ttest_ind
stat, p = ttest_ind(adult_dist['is_churn'], ya_dist['is_churn'])
print('stat=%.3f, p=%.3f' % (stat, p))
if p > 0.05:
    print('Probably the same distribution')
else:
    print('Probably different distributions')
```

```
stat=-28.100, p=0.000
Probably different distributions
```

4

Conclusions

Questions/Answers



Results

- Over 50% of subscribers do not self-identify
- Self-identified users pay more than those who don't
- They also churn slightly more than those who don't

Next Steps

- Finding less anonymized data for more detailed customer study
- Figure out which payment plan will keep the most users while allowing for maximum revenue
- A/B testing for plan pricing

Thanks!

Any questions?