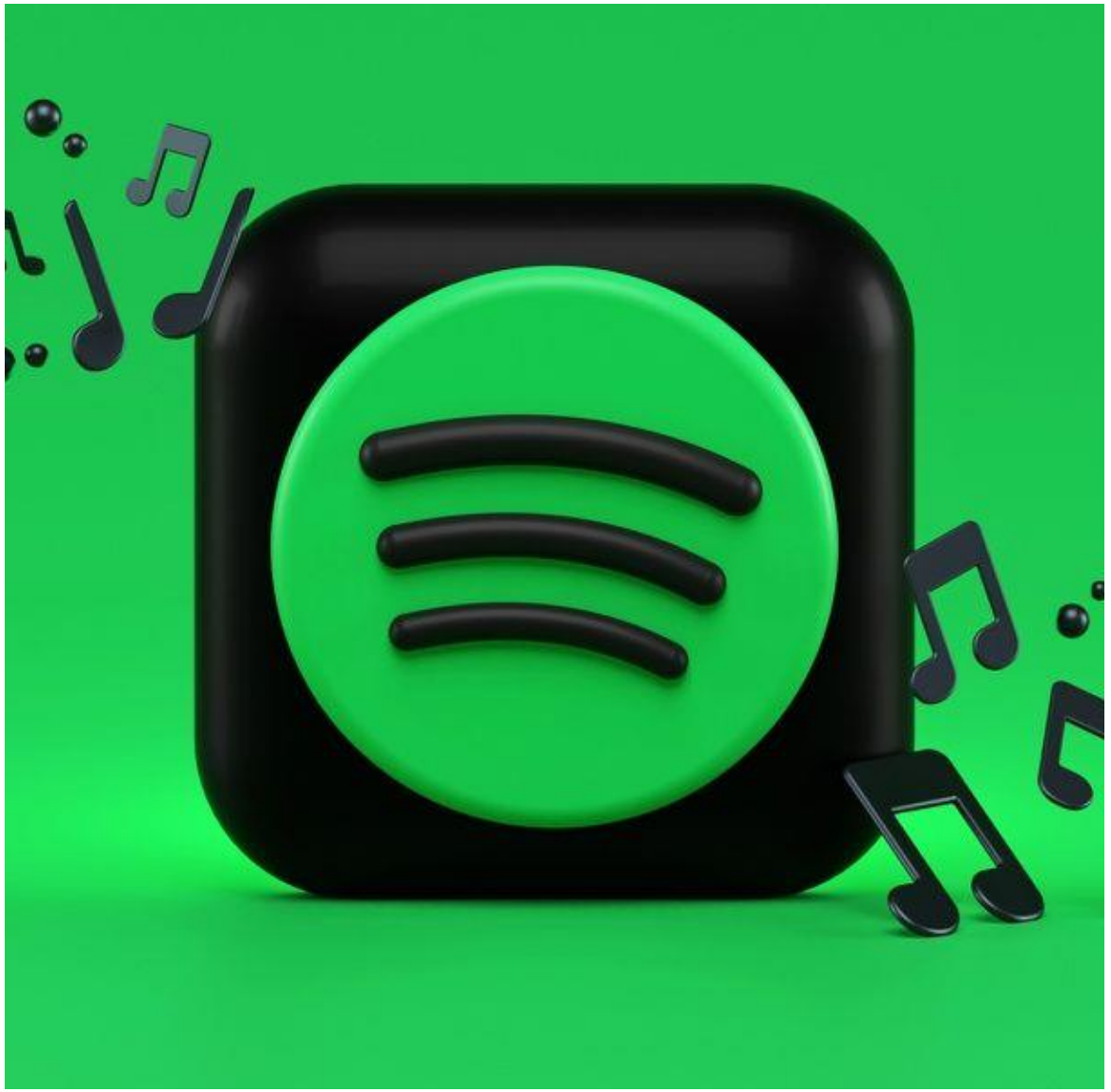


Project Marketing Analytics

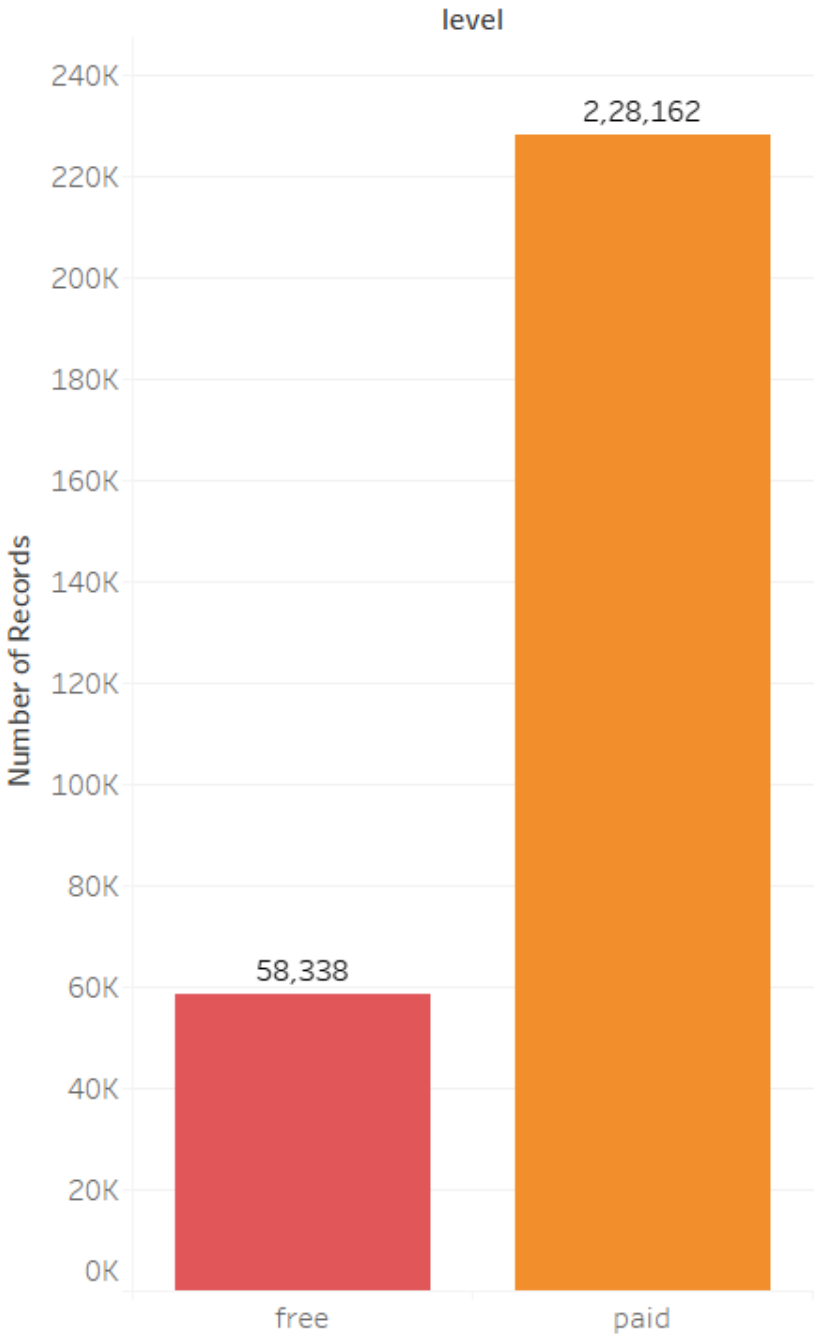
SPOTIFY

DATA SUMMARY

- Spotify is a music streaming company delivering services to 75 million active users globally.
- Spotify is currently offering two type of subscriptions: a free and a premium subscription.
- A huge milestone has Spotify has its presence only in 60 nation in the world.
- This marketing plan is about Spotify extending its services to highly potential country.

DATA COLUMNS WITH DESCRIPTION

Column	Type	Description
userId	string	Unique identifier of the user, the event is related to
artist	string	Name of the artist related to the song related to the event
auth	string	"Logged in" or "Cancelled"
firstName	string	First name of the user
gender	string	Gender of the user, "F" or "M"
itemInSession	bigint	Item in session
lastName	string	Last name of the user
length	double	Length of the song related to the event
level	string	Level of the user's subscription, "free" or "paid". User can change the level, so events for the same user can have different levels
location	string	Location of the user at the time of the event
method	string	"GET" or "PUT"
page	string	Type of action: "NextSong", "Login", "Thumbs Up" etc.
registration	bigint	Registration
sessionId	bigint	Session id
song	string	Name of the song related to the event
status	bigint	Response status: 200, 404, 307
ts	bigint	Timestamp of the event
userAgent	string	Agent, which user used for the event, for example, "Mozilla/5.0"

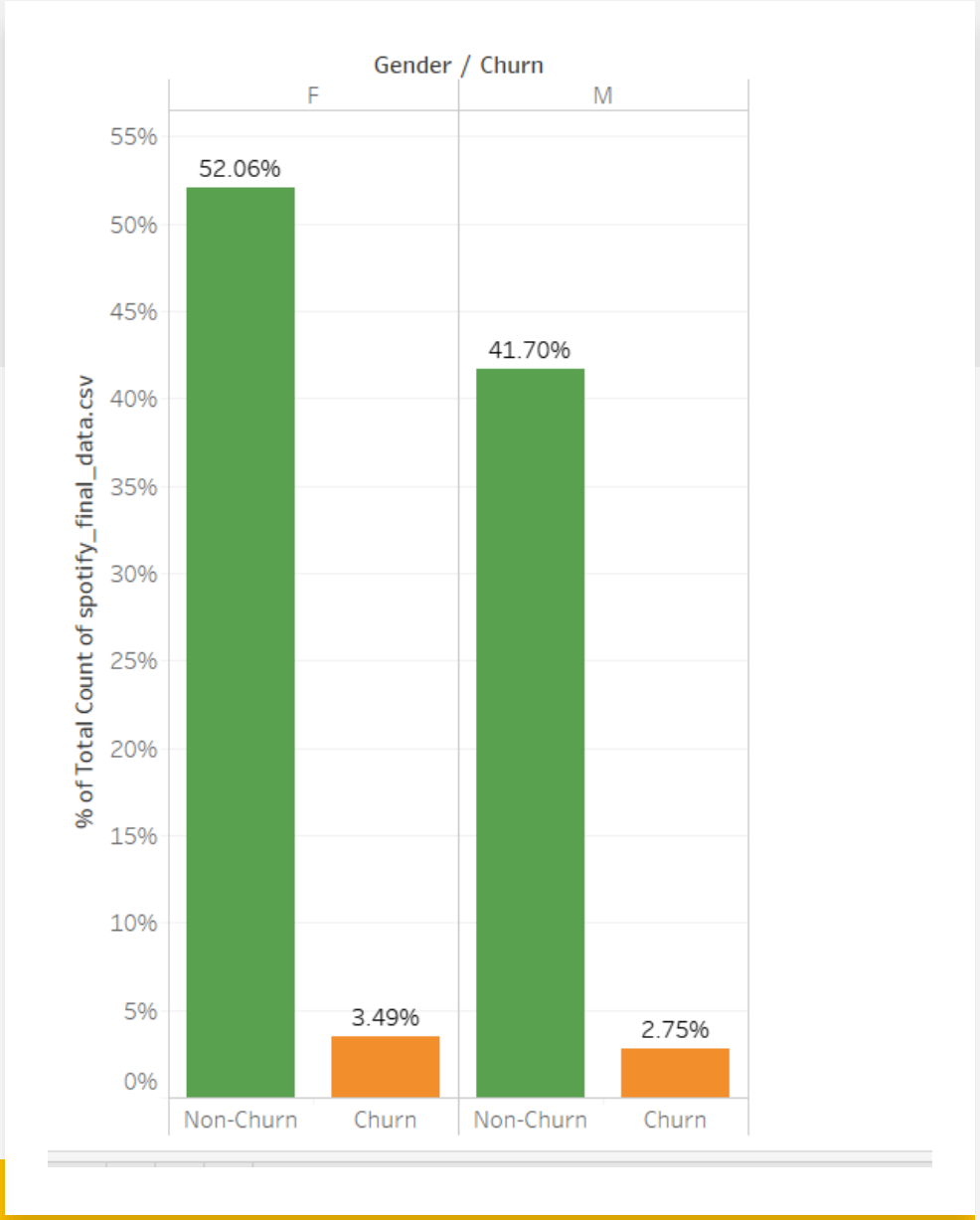


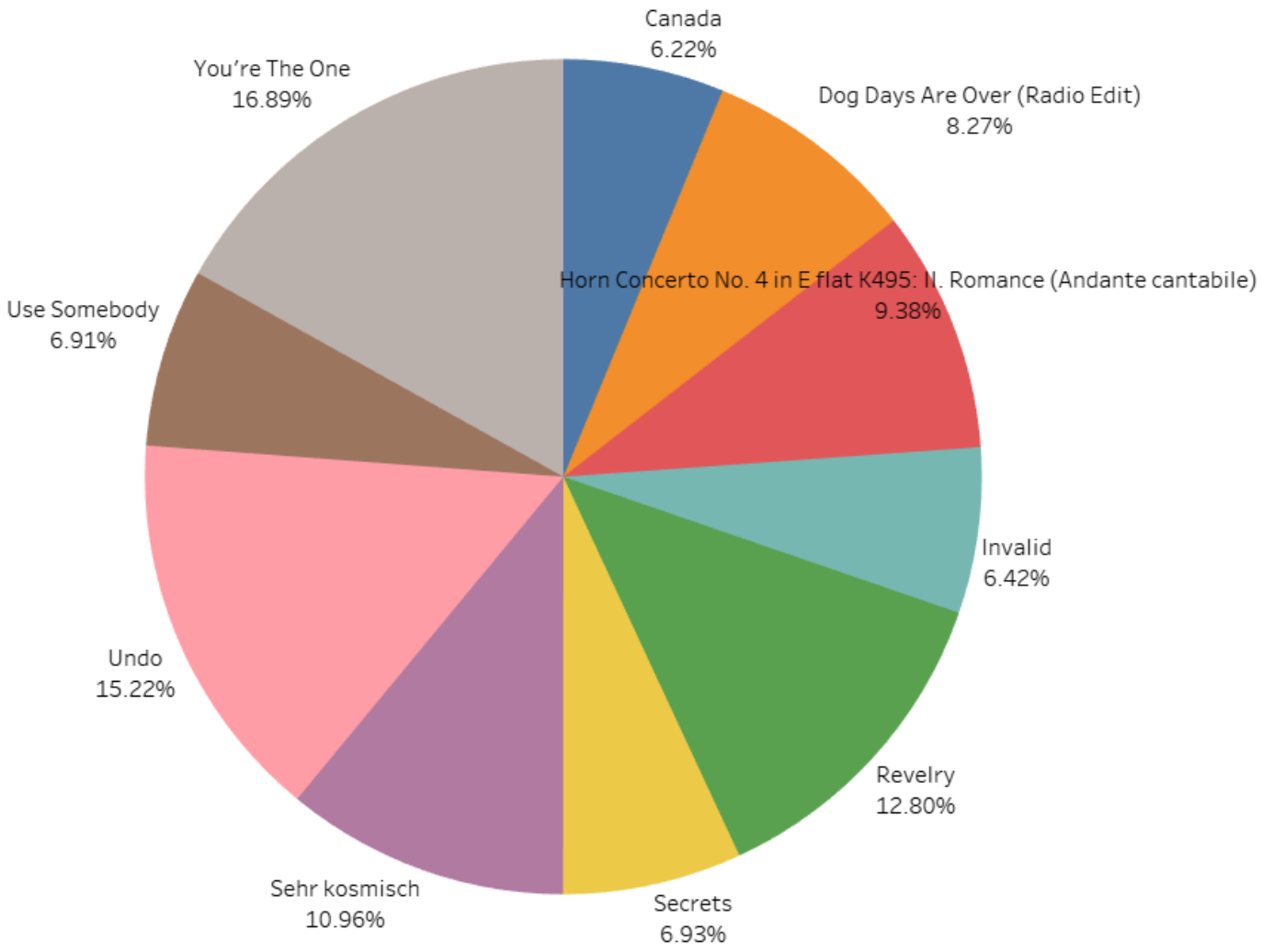
What we got more subscribers or free users?

- We have more Paid users than Free users.
- Number of Paid users are 2,28,162.
- Number of Free users are 58,338.

Which gender group churn more??

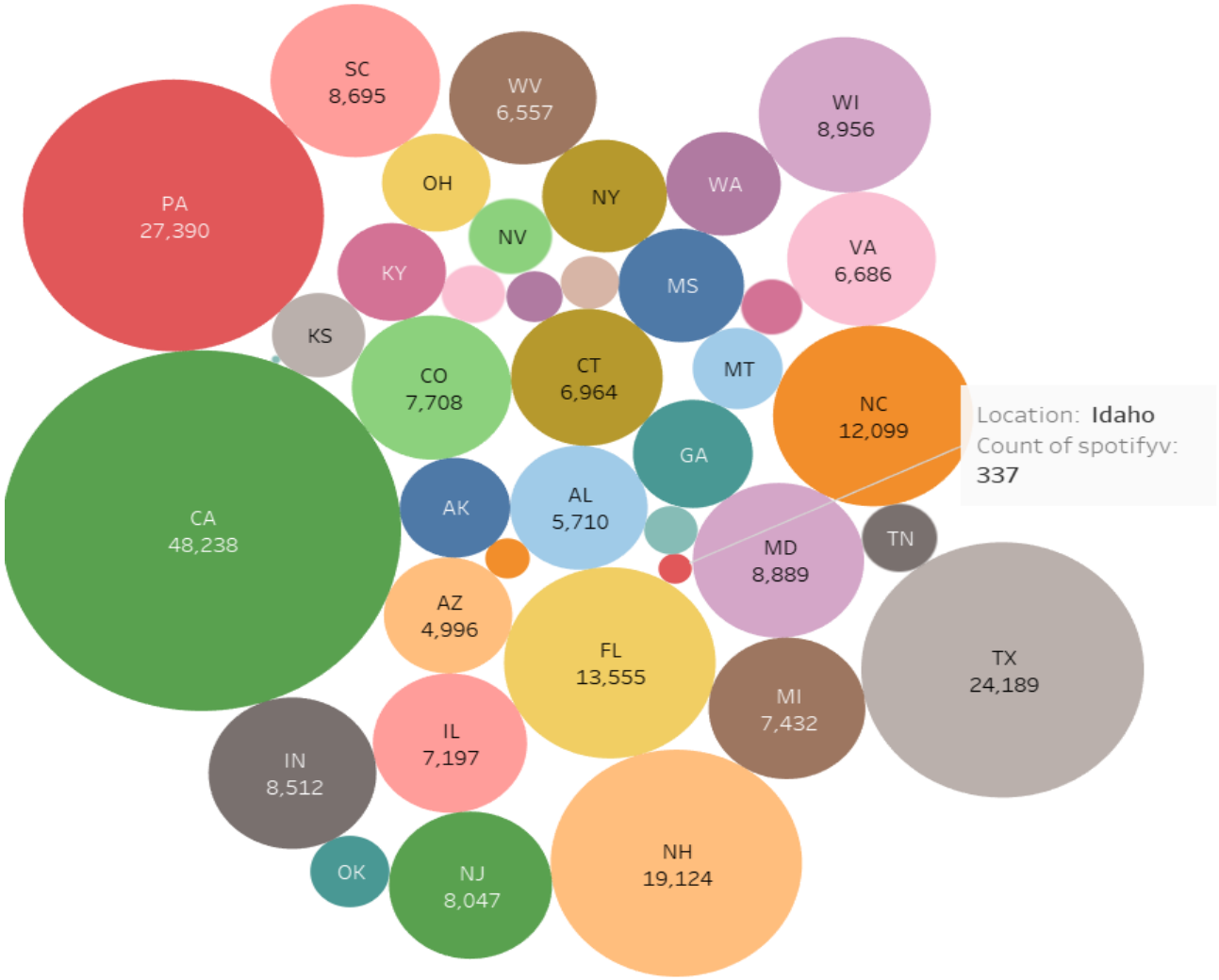
- Comparing churn with overall percentage for male and female.
- As result we got:
- Female users are more than male and likewise their churn ratio is also more.





WHICH ARE TOP 10 SONGS?

- A percentage wise ratio can be seen in this pie chart.
- "**You're the One**" seems to be more popular than any other song with 16.89%, followed by **"Undo"** at 15.22% among top 10 songs.

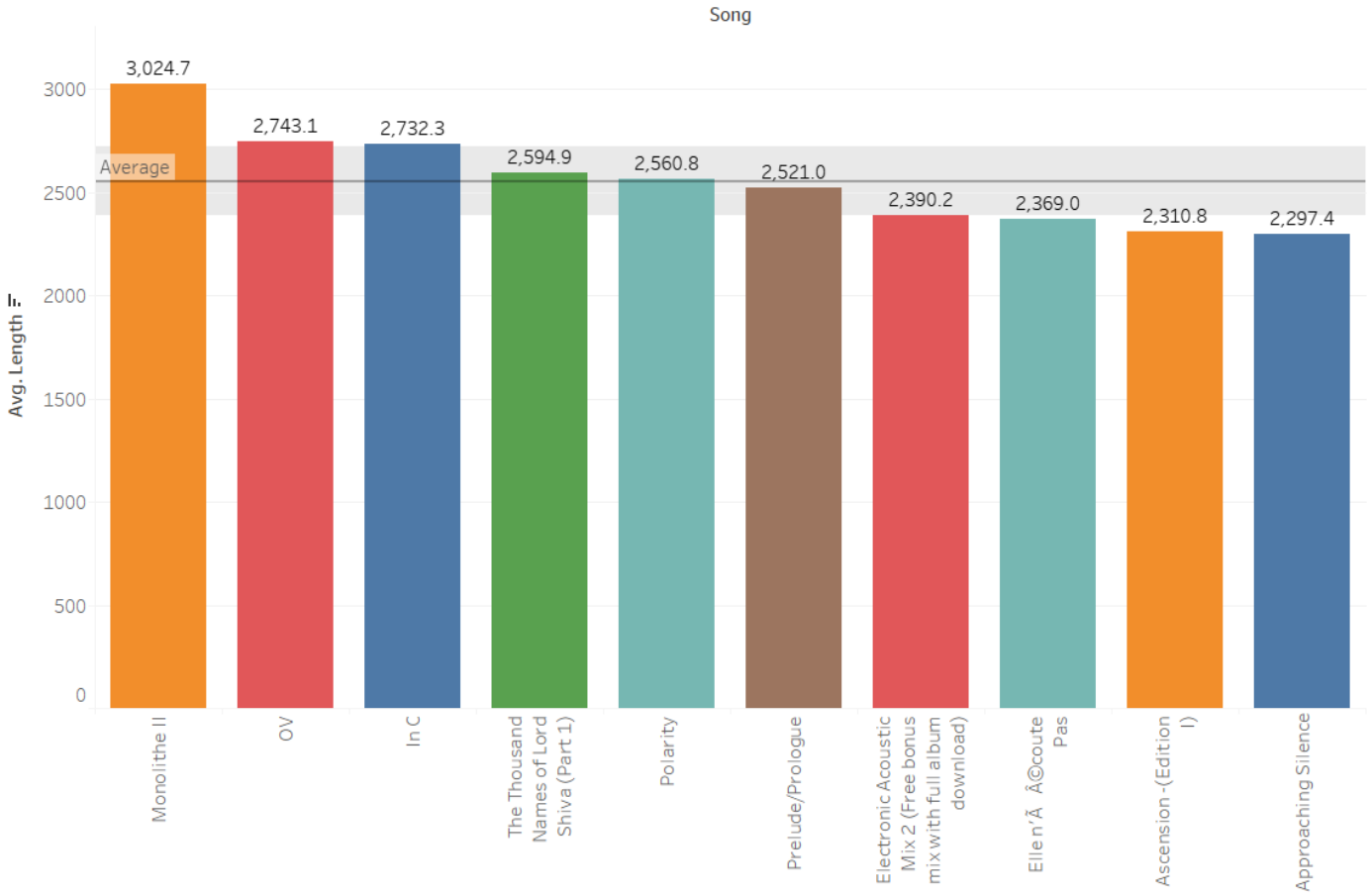


FROM WHICH STATE DO SPOTIFY GET MORE LISTENERS?

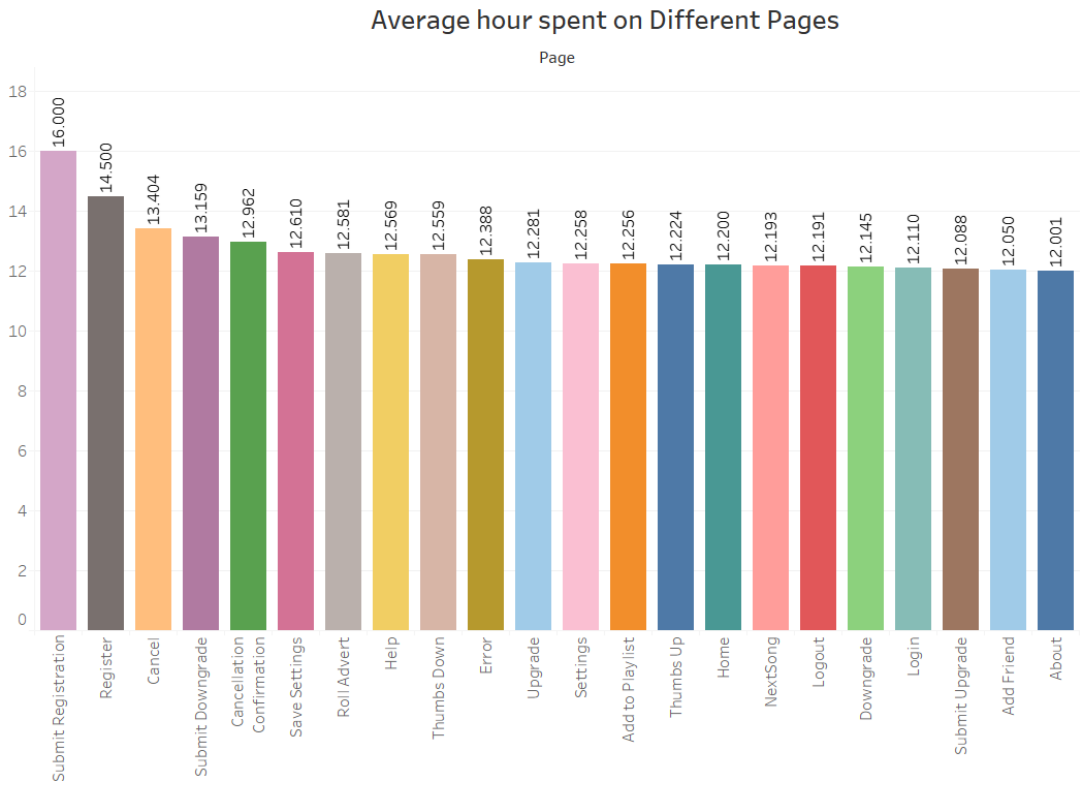
- This bubble chart gives a glimpse at what area Spotify got more listeners and where less.
- California tops it with 48,238 songs played and Idaho has least with just 337 songs listened.

AVERAGE TIME SPEND ON SONGS

- We can see there are people who love to listen songs on repeat mode.
- Following Analysis shows which song has the highest average and what is the average of songs listened.
- The average marked is 2554.4 times.



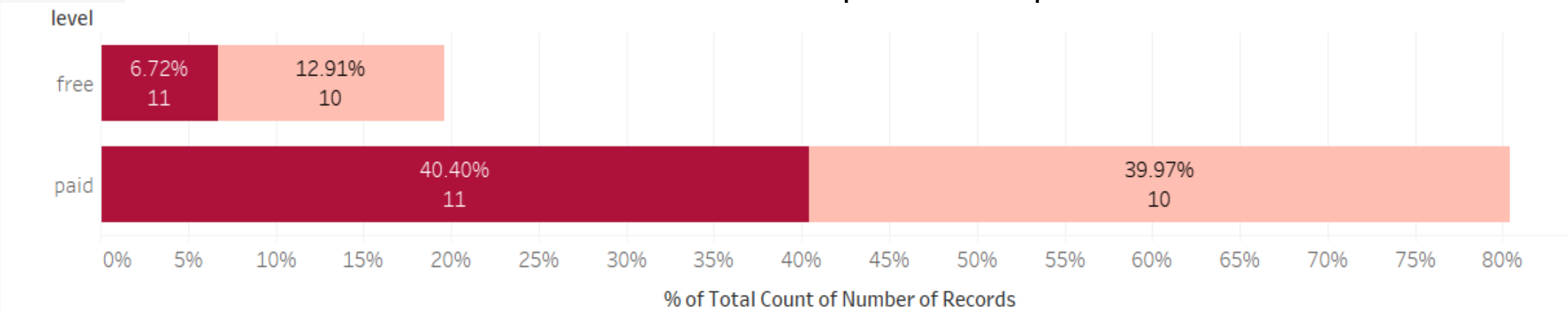
TIME SPENT IN APPLICATION FOR OTHER CAUSE.



- As noticed 16 hours are spent for the registration process only.
- Nearly 12 hours are spent just to know about the application

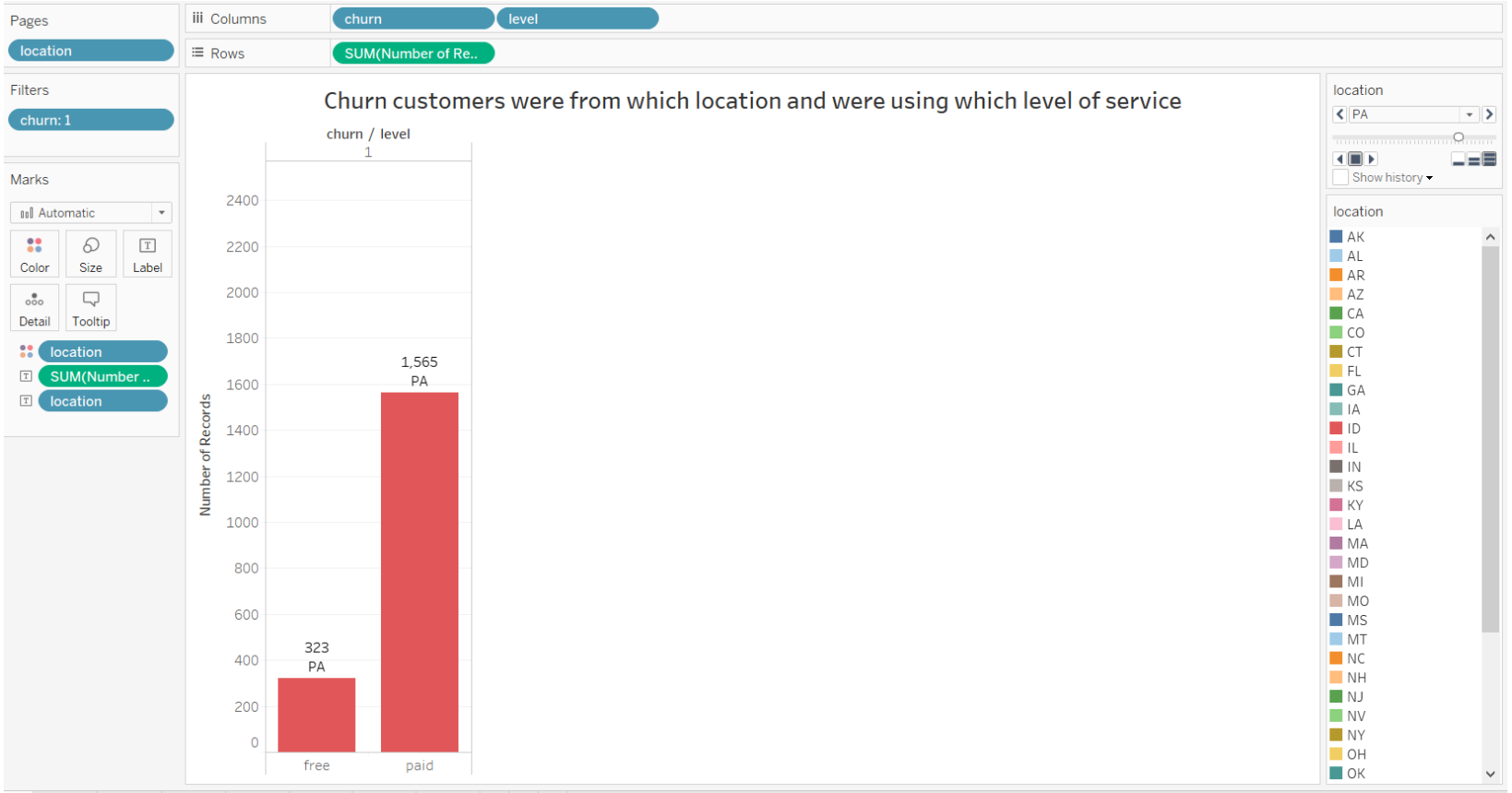
TYPE OF SERVICES USED BY CHURN PEOPLE

- Are the people using free services or paid service?
- Which month is affected more?
- From the analysis we can say that paid subscribers are getting more churn compare to free users in the month of November.
- Where dark pink is for October and light pink color represents November

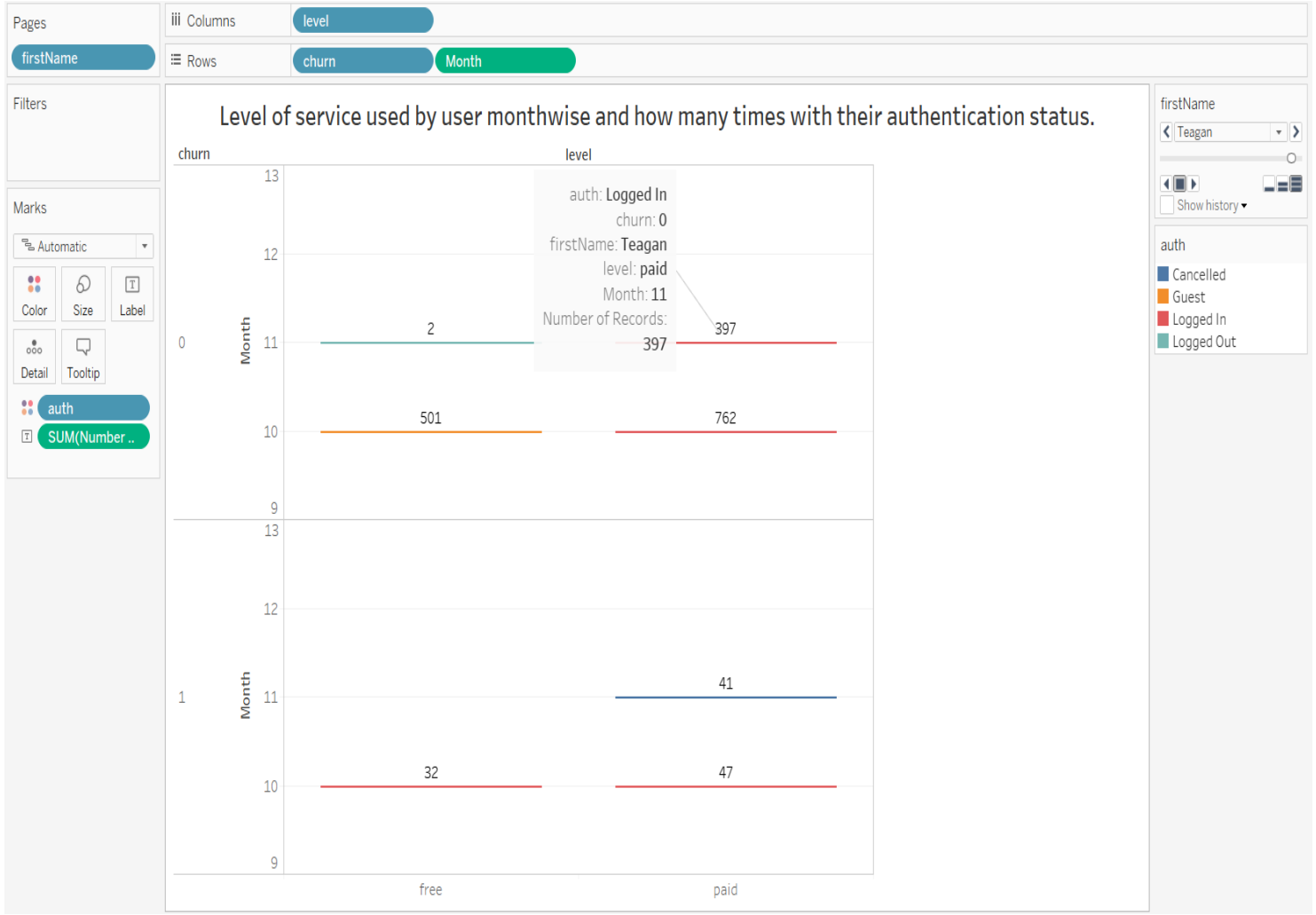


LOCATION WHERE PEOPLE GOT MORE CHURNED.

- This graph gives us knowledge about the churn customers, were they using free services or paid services.



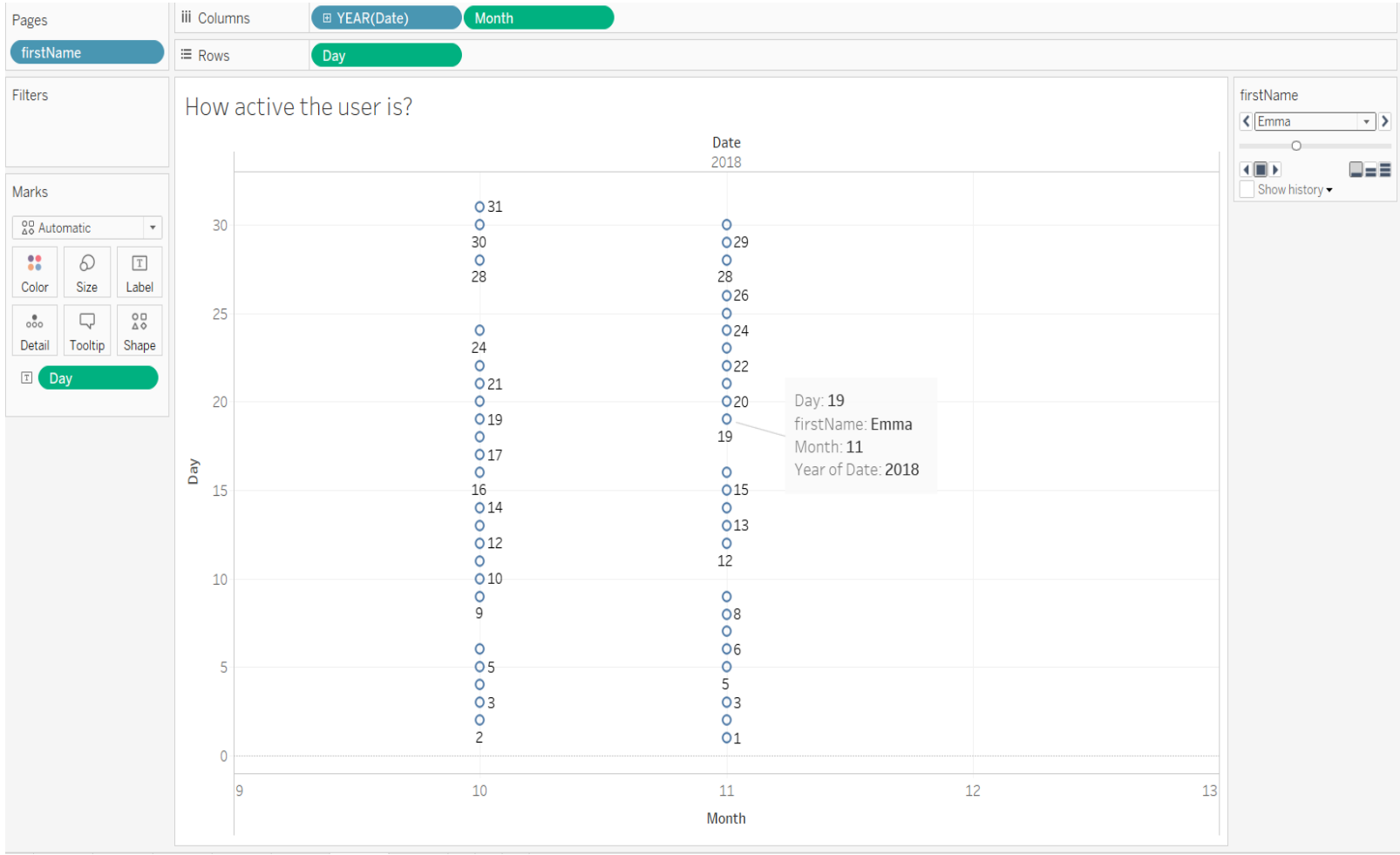
LEVEL OF SERVICE, AUTHENTICATION STATUS AND NO. OF TIME USED.



- We have made a survey where we can have all these information at once by monthly basis.

ACTIVE USERS

- The data shows activity of the user each day of the months.
- The data currently shows the activity of Emma

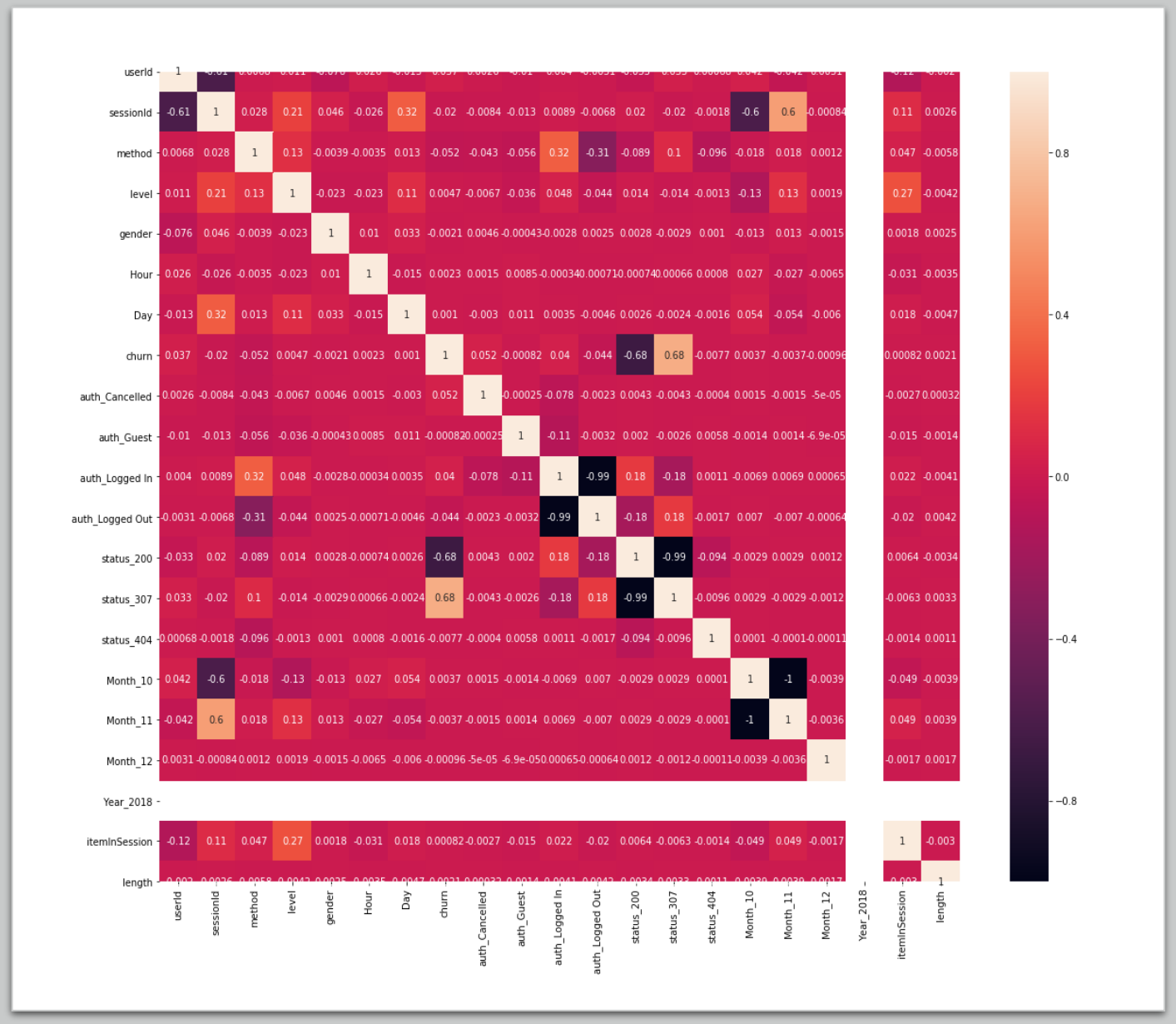


DATA MODELING

Data modeling is a process used to define and analyze **data** requirements needed to support the business processes within the scope of corresponding information systems in organizations.

We have used following models so that we can determine which model is best.

We have also used correlation matrix heatmap to show the relation between the data.



CORRELATION MATRIX HEATMAP

MODELS

- Logistic Regression
- Decision tree classifier
- Random forest Classifier.
- Linear SVC
- Navie Bayes

LOGISTIC REGRESSION

- The accuracy is 94%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	53716
1	0.00	0.00	0.00	3584
accuracy			0.94	57300
macro avg	0.47	0.50	0.48	57300
weighted avg	0.88	0.94	0.91	57300

DECISION TREE CLASSIFIER

- The accuracy is 94%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	53716
1	0.00	0.00	0.00	3584
accuracy			0.94	57300
macro avg	0.47	0.50	0.48	57300
weighted avg	0.88	0.94	0.91	57300

RANDOM FOREST CLASSIFIER

- The accuracy is 96%.

	precision	recall	f1-score	support
0	0.98	0.98	0.98	53716
1	0.70	0.66	0.68	3584
accuracy			0.96	57300
macro avg	0.84	0.82	0.83	57300
weighted avg	0.96	0.96	0.96	57300

LINEAR SVC

- The accuracy is 73%

	precision	recall	f1-score	support
0	0.95	0.75	0.84	53716
1	0.11	0.45	0.17	3584
accuracy			0.73	57300
macro avg	0.53	0.60	0.50	57300
weighted avg	0.90	0.73	0.80	57300

NAIVE BAYES

- The accuracy is 94%.

	precision	recall	f1-score	support
0	0.94	1.00	0.97	53716
1	0.00	0.00	0.00	3584
accuracy			0.94	57300
macro avg	0.47	0.50	0.48	57300
weighted avg	0.88	0.94	0.91	57300

MODEL PREFERRED.

- The analysis shows Random forest classifier has the best accuracy which is 96%.

Thank You

- Group 25:
  - Anant Patel (0756333)
    Darshan Patel (0753849)
    Parth Sutariya (0756042)
    Rushil Patel (0755980)
    Ruchita Tamboli (0756325)