Paciolan: Custom Fields

Spring 2021 Presentation 1
Presented By: Yihan Wang, Sirui Hu, Yadi Yang

Overview

- **Background**
  - **Paciolan** is a company aiming to serve the **business clients**, like the university sports and venues, and provide a better platform for them to serve their **patrons and donors (users)**.

- **Goal**
  - To derive business insights in helping define user categories and user potential interests from the **5 client-defined major custom fields** that were assigned to each user.

- **Objectives**
  - To find out similar characteristics among users who have been assigned common custom keywords/tags.
    - Features will be processed in a **data analysis pipeline**

Data Description

Response/Target Data

5 Custom Fields

- Keywords
- Tags
- User Defined Fields
- Donor Categories
- Motives

Features

Clients' & Users' Demographics

- Donor Membership
- Allocation of Donation
- User Email

Action Data

- Donation History
- Ticket Purchasing History
- All Transaction History

Cleaned Data Sets (Snowflake DB & S3)

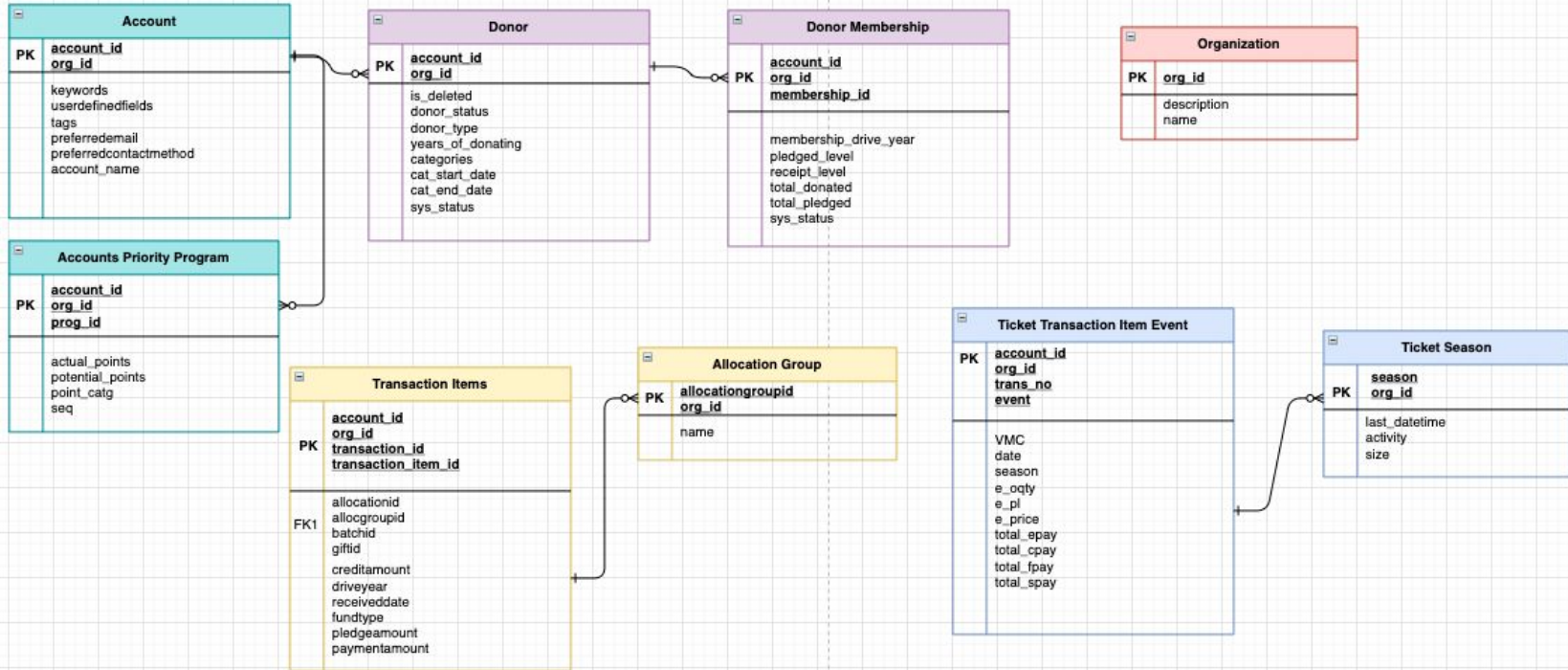*** We use AWS S3 bucket to store and manage our cleaned data.**

Major Datasets

- **Account Data**
- **Donor Data**
- **Transaction Items Data**
- **Ticket Season Data**
- **Ticket Transaction Item Event Data**

Other Usable Datasets

- **Account Priority Points Program Details**
- **Donor Membership**
- **Allocations**
- **Allocation group**
- **Organizations**
- **Gifts**

ER Diagram for Cleaned Dataset

Account Data (2018-2021)

account_id	org_id	keywords	userdefinedfields	tags	preferredemail	preferredcontactmethod	account_name
138788	MSSTATE	NaN	NaN	NaN	{'mr.erise@gmail.com'}	NaN	{'Erise Wilson'}
106626	MSSTATE	{'PRITCC', '0'}	NaN	{'NONSTMNOND17'}	NaN	{'N'}	{'Pritchard Engineering'}
119515	MSSTATE	{'P6625156412', 'DAVIDJ', 'PJW6412@GMAIL.COM'}	NaN	{'NONSTMNOND17'}	{'jw6412@gmail.com'}	{'N'}	{'James Davidson'}
137104	MSSTATE	NaN	NaN	NaN	{'richlopez70@gmail.com'}	NaN	{'Rich Lopez'}
132816	MSSTATE	NaN	NaN	NaN	{'john.harden@gmail.com'}	NaN	{'John Harden'}
...	...	...	...	...	...	...	...
445879	ARKANSAS	NaN	NaN	NaN	{'kellyrussell712@gmail.com'}	{'N'}	{'kellly russell'}
53939559	TAM	NaN	NaN	NaN	{'hwoolwine@bigcompany.com'}	{'N'}	{'Heather Woolwine'}
371009425771	FSU	NaN	{'STID': '200311128'}	NaN	{'jhh19d@my.fsu.edu'}	{'N'}	{'john hunter harris'}
386122	TTU	NaN	NaN	NaN	{'Trejomike18@gmail.com'}	{'N'}	{'Michael Trejo'}
635515	NEBRASKA	NaN	NaN	NaN	{'ajlempke@gmail.com'}	{'N'}	{'Austin Lempke'}

10,868,781 rows

Donor Data

	org_id	account_id	is_deleted	donor_type	donor_status	years_of_donating	categories	cat_start_dates	cat_end_dates	sys_status
0	NEBRASKA	634566	False	"Individual"	"Prospect"	0.0	NaN	None	None	1.0
1	ARKANSAS	443476	False	"Individual"	"Active"	0.0	NaN	None	None	1.0
2	ARKANSAS	443495	False	"Individual"	"Active"	0.0	NaN	None	None	1.0
3	ARKANSAS	443579	False	"Individual"	"Active"	0.0	NaN	None	None	1.0
4	ARKANSAS	443554	False	"Individual"	"Active"	0.0	NaN	None	None	1.0
...	...	...	...	...	...	...	...	...	...	...
5259325	OKLAHOMA	204432	False	"Individual"	"New"	NaN	NaN	NaN	NaN	1.0
5259326	OKLAHOMA	572635	False	"Individual"	"New"	0.0	NaN	None	None	1.0
5259327	OKLAHOMA	161379	False	"Individual"	"New"	NaN	NaN	NaN	NaN	1.0
5259328	OKLAHOMA	809988	False	"Individual"	"New"	2.0	"TFPP": "61"	None	None	1.0
5259329	OKLAHOMA	537414	False	"Individual"	"New"	NaN	NaN	NaN	NaN	1.0

5259330 rows × 10 columns

5,259,330 rows

Transaction Item Data

	ORG_ID		ACCOUNT_ID	ALLOCATIONID	ALLOCGROUPID	TRANSACTIONID	TRANSACTION_ITEM_ID	BATCHID
0	TAM		10039	PWFBEC	12MF	18375	1	2
1	TAM		14100	WFBEC	12MF	18603	1	2
2	TAM		15698457	KFC-ADA_NUAfromNUB		18355	1	2
	TAM		121008594	ANF		18107	1	2
	TAM		95788	KFC-ADA_NUAfromNUB		18351	1	2
...	...		...	...		...	...	...
	MSON			ISF		2964011	1	1561
10878583	TAM			SMC		1891484	1	6859
10878584	FSU			BAF		1379631	1	5750
10878585	ARKANSAS		6620	AF-One	ANNUAL	1537523	1	1486
10878586	FSU		5258196	BAF	BAF	1379600	1	5750

Who received the money

Who sent the money

Where the money will be allocated

	CREDITAMOUNT	DRIVEYEAR	RECEIVEDDATE	FUNDTYPE	GIFTID	PLEDGEAMOUNT	PAYMENTAMOUNT
	800000	2016	2016-06-09 05:00:00	C	52972	0	0
	600000	2016	2016-08-03 05:00:00	C	371731	0	0
	-190000	0	2016-06-03 05:00:00	C	246008	0	0
	15000	2016	2016-05-13 05:00:00	C	236778	0	0
	-95000	0	2016-06-03 05:00:00	C	203817	0	0
	...	...	...	...	...	...	...
	0	2021	2021-04-09 23:03:52	S	None	45000	0
	0	2021	2021-04-09 23:26:40	S	None	0	0
	0	2021	2021-04-10 03:24:05	S	None	0	130000
	0	2021	2021-04-10 00:38:29	S	None	110000	0
	0	2021	2021-04-10 01:09:01	S	None	0	33000

How much is paid (divided by 100)

[10,878,587 rows x 20 columns]

Ticket Season Data

	activity	org_id	last_datetime	name	season	size
0	VB	BAYLOR	2013-02-26 09:15:15	2012 Acrobatics & Tumbling	AT12	M
1	VB	BAYLOR	2013-04-30 15:20:03	Acrobatics & Tumbling 2013	AT13	T
2	VB	BAYLOR	2014-01-23 14:03:05	Acrobatics & Tumbling 2014	AT14	T
3	VB	BAYLOR	2015-01-27 08:17:17	Acrobatics & Tumbling 2015	AT15	T
4	VB	BAYLOR	2017-01-09 11:29:14	2016 Acrobatics and Tumbling	AT16	T
5	VB	BAYLOR	2017-12-11 14:25:46	2017 Acrobatics and Tumbling	AT17	T
6	VB	BAYLOR	2018-10-02 10:24:10	2019 Acrobatics & Tumbling	AT19	M
7	VB	BAYLOR	2019-12-20 11:19:14	2020 Acrobatics & Tumbling	AT20	M
8	B	BAYLOR	2015-06-19 11:52:38	MEN'S BASKETBALL 2000/01	BB01	L
9	BB	BAYLOR	2013-02-19 17:14:04	MEN'S BASKETBALL 2001/2002	BB02	M
10	SA	BAYLOR	2013-02-19 17:14:14	BDSC TICKET OFFICE	BDSC12	S

3,643 rows

Ticket Transaction Item Event Data

	account_id	org_id	trans_no	vmc	date	event	season	e_oqty	e_pl	e_price	total_epay	total_cpay	total_fpay	total_spay
0	2282.0	BAYLOR	32193	2	2018-07-09	F04	F18	0	7.0	58.33	0.0	0.0	0.0	0
1	2282.0	BAYLOR	32193	2	2018-07-09	F05	F18	0	7.0	58.33	0.0	0.0	0.0	0
2	2282.0	BAYLOR	32193	2	2018-07-09	F06	F18	0	7.0	58.35	0.0	0.0	0.0	0
3	2282.0	BAYLOR	32193	3	2018-07-09	F01P	F18	0	2.0	75.00	0.0	0.0	0.0	0
4	2282.0	BAYLOR	32193	3	2018-07-09	F02P	F18	0	2.0	75.00	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
78234969	463205.0	KSU	16459	1	2020-02-17	B27	KS19	0	1.0	5.44	0.0	0.0	0.0	0
78234970	191150.0	KSU	16460	1	2020-02-17	B01	KS19	0	1.0	5.56	0.0	0.0	0.0	0
78234971	191150.0	KSU	16460	1	2020-02-17	B02	KS19	0	1.0	5.56	0.0	0.0	0.0	0
78234972	191150.0	KSU	16460	1	2020-02-17	B03	KS19	0	1.0	5.56	0.0	0.0	0.0	0
78234973	191150.0	KSU	16460	1	2020-02-17	B04	KS19	0	1.0	5.56	0.0	0.0	0.0	0

78,234,974 rows

Other Usable Data

Account Priority Program Data

	account_id	actual_points	org_id		points_catg	potential_points	prog_id	seq
0	14926389.0	72502000	TAM	{ "1": 45302000, "Total Giving","10": 0, "Texa...	0.0	TAMA	2.0
1	20212.0	23031600	TAM	{ "1": 12181600, "Total Giving","10": 0, "Texa...	0.0	TAMA	6.0
2	20333.0	14601680	TAM	{ "1": 3151680, "Total Giving","10": 0, "Texas...	0.0	TAMA	13.0
3	602584.0	15729000	TAM	{ "1": 9479000, "Total Giving","10": 0, "Texas...	0.0	TAMA	18.0
4	20696.0	93336000	TAM	{ "1": 72036000, "Total Giving","10": 0, "Texa...	0.0	TAMA	19.0
...	...	...	...	...	...	...	...
770548	565129.0	220000	OKLAHOMA	{ "1": 20000, "Current Seating Gifts","10": 0,...	0.0	CP	312.0
770549	579356.0	500000	OKLAHOMA	{ "1": 100000, "Current Seating Points","10": ...	0.0	SP	319.0
770550	273945.0	69680000	OKLAHOMA	{ "1": 1500000, "Current Seating Points","10":...	0.0	SP	322.0
770551	807691.0	1215000	OKLAHOMA	{ "1": 0, "Current Seating Points","10": 0, "S...	0.0	SP	331.0
770552	26887.0	2269600	OKLAHOMA	{ "1": 40000, "Current Seating Gifts","10": 17...	150000.0	CP	214.0

770553 rows × 7 columns

Organization

	ORG_ID	DESCRIPTION	NAME
0	SMU	Southern Methodist University	SMU
1	DUQUESNE	Duquesne University	DUQUESNE
2	UNCW	University of North Carolina Wilmington	UNCW
3	NET-PT04	Portland 5 Centers for the Arts (NET-PT04)	Portland 5 Centers for the Arts
4	GLOBAL-EVERETT	Everett Box Office	Global-Everett
...	...	...	...
284	CLEMSON	Clemson University	CLEMSON
285	PITT	University of Pittsburgh	PITT
286	UCI	UC Irvine	UCI
287	MIAMIOH		MIAMIOH
288	NET-PT04		NET-PT04

Donor Membership Data

	org_id	account_id	membership_id	membership_drive_year	pledged_level	receipt_level	total_donated	total_pledged	sys_status
0	ARKANSAS	119362	RF	2011	Big Red	Big Red	200.0	200.0	1
1	ARKANSAS	119362	RF	2012	Big Red	Big Red	200.0	200.0	1
2	ARKANSAS	119362	RF	2013	Big Red	Big Red	200.0	200.0	1
3	ARKANSAS	119362	RF	2014	Big Red	Big Red	200.0	200.0	1
4	ARKANSAS	119362	RF	2015	Big Red	Big Red	200.0	200.0	1
...	...	...	...	...	...	...	...	...	...
1174115	USC	330282	TAF17	2017	Trojan Club Starter	NaN	0.0	200.0	1
1174116	USC	131049	TAF18	2018	Trojan Club Starter	Trojan Club Starter	200.0	200.0	1
1174117	USC	131049	TAF19	2019	Trojan Club Starter	Trojan Club Starter	200.0	200.0	1
1174118	USC	131049	TAF21	2021	Trojan Club Starter	Trojan Club Starter	200.0	200.0	1
1174119	USC	9061463568	TAF21	2021	Trojan Club Starter	Trojan Club Starter	200.0	200.0	1

1174120 rows × 9 columns

Allocation group

	ORG_ID	ALLOCATIONGROUPID	NAME
0	ARKANSAS	OTH	Other
1	WASHINGTON	ENDOW	Endowed Scholarships
2	USC	SWM	Swim With Mike
3	KANSAS	CC	Capital Campaign
4	KANSAS	GIK	GIK
...	...	...	...
209	UT-A	SG	Special Gifts
210	UT-A	LP	Loyalty Points
211	FSU	CR	Campaign Restricted
212	CFL	DEVEL	Development
213	CFL	OUTRE	Outreach

Data Visualization - Word Clouds
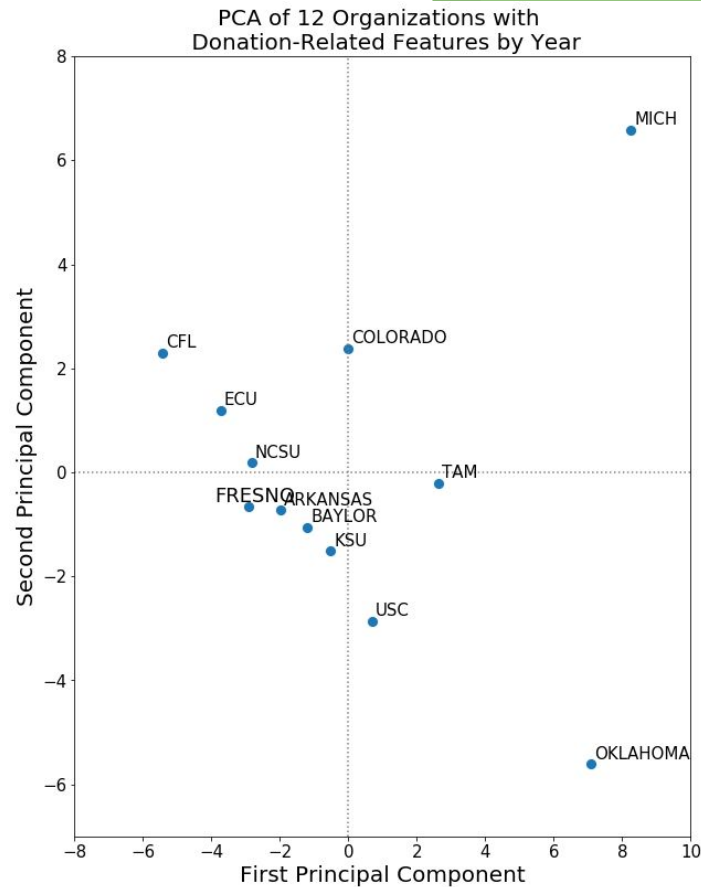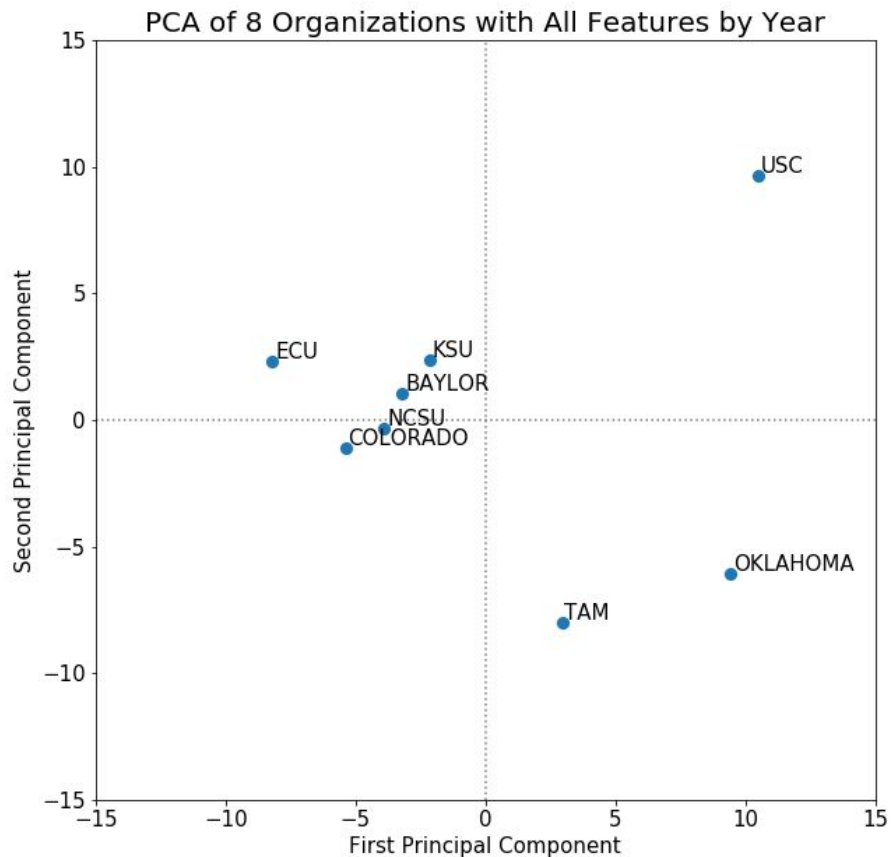
21,015,018 Keywords

134,480 Tags
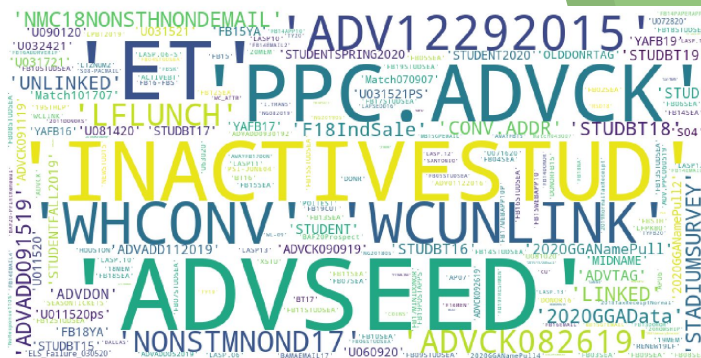
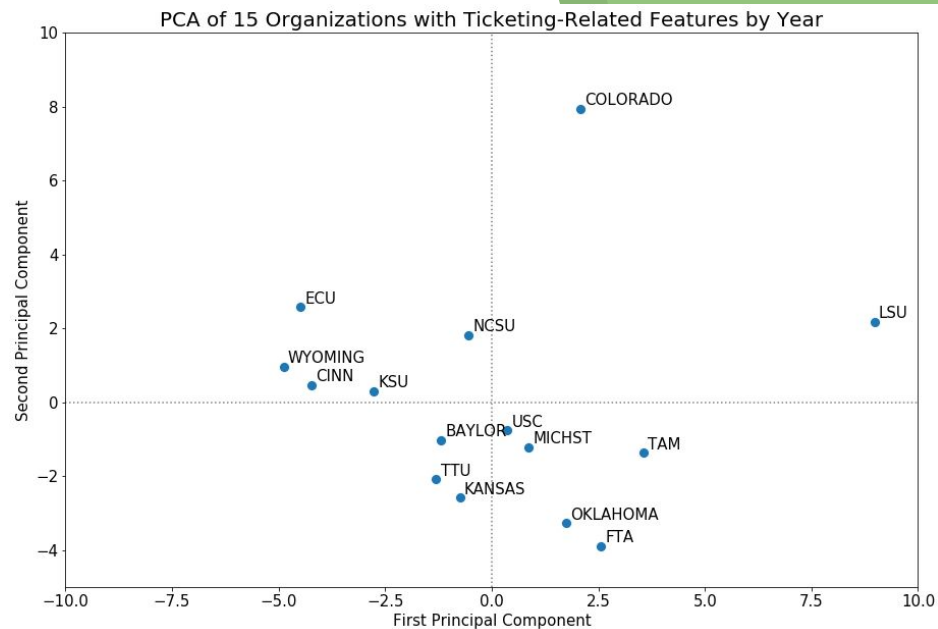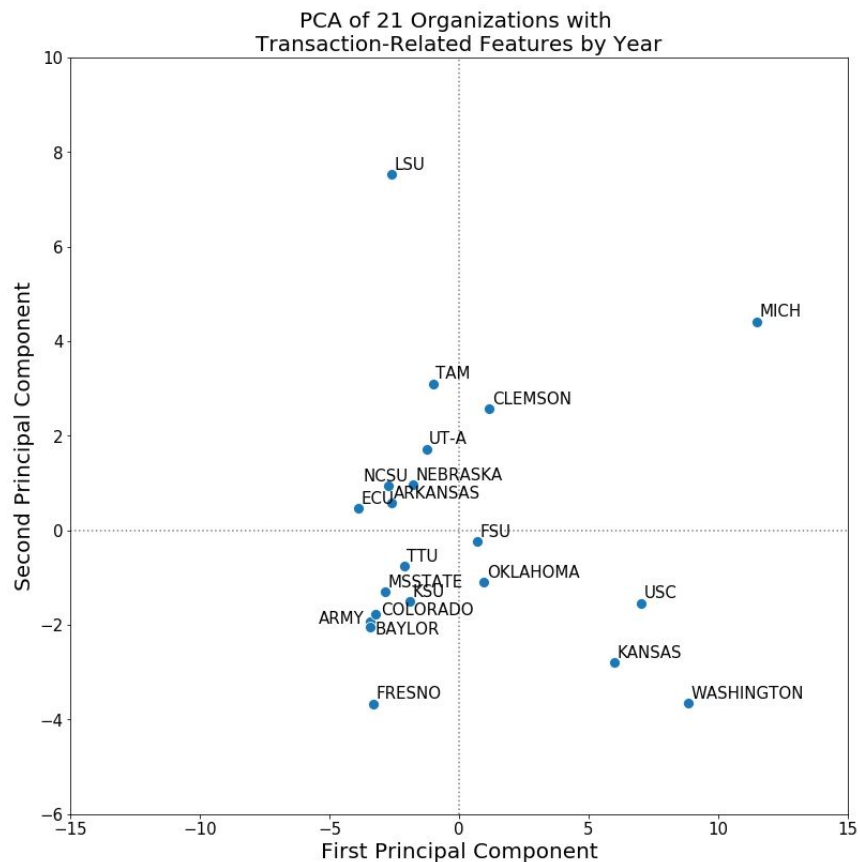

134 User Defined Fields

PCA Analysis Summary

- Duration: 3 consecutive years (2018-2020)

- Features: values in each column were aggregated in:
  - sum, mean, median, min, max, var, size

- 4 PCA plots:
  - Total ticket payment features by year
  - Total pledged and received donation features by year
  - Total transaction amount aggregate values by year
  - Total payment features with all features in the above 3 tables, excluding the NULL values

- Method:
  - Step 1: Group by org_id, account_id and year:
    - total amount in each year for each organization
  - Step 2: Group by org_id and year:
    - aggregate values in each year for each organization

Data Visualization - PCA

PCA of 8 Organizations with All Features by Year

PCA of 12 Organizations with Donation-Related Features by Year

Data Visualization - PCA Cont.

PCA of 21 Organizations with Transaction-Related Features by Year

PCA of 15 Organizations with Ticketing-Related Features by Year

Milestones

- ☑️ Set up accounts
- ☑️ Explore the raw data
- ☑️ Draft the proposal presentation.
- Research relevant literature and methods to come up with some initial ideas.
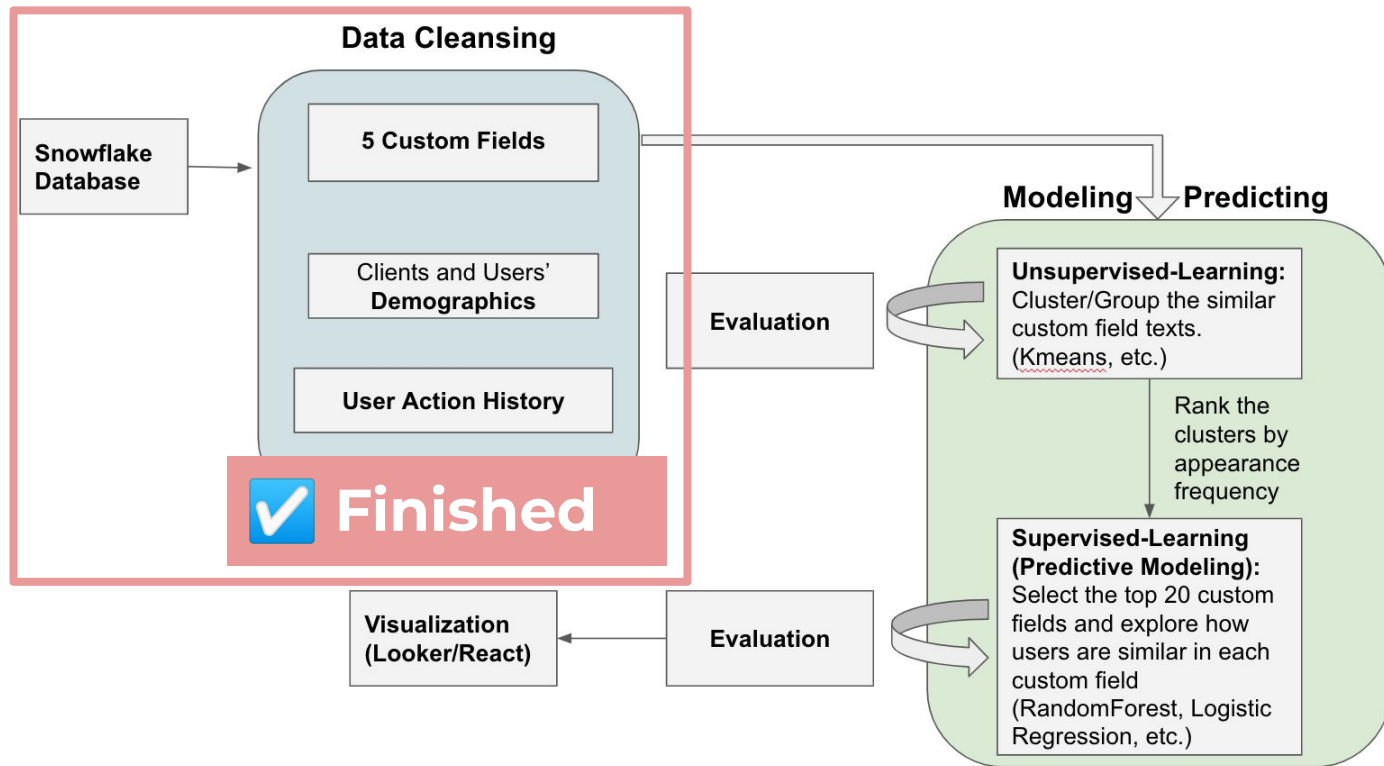
- Evaluate the results of experiments.
- Use Looker/React to build data visualization for the deliverable results
- Automate the process of assigning tags and gain insights from the results
- Start writing the final report

MARCH	APRIL	MAY	JUNE

- ☑️ Set up a weekly agenda and recurrent zoom meetings with the Paciolan team.
- ☑️ Preprocess the raw data and start cleaning and organizing the data for future use.
- ☑️ Discuss with the sponsor and know their preference
- Continue brainstorming and experimenting on any possible approaches.

- Optimize results and prepare for the final presentation
- Finalize the final report

Data Analysis Pipeline

Data Cleansing

Snowflake Database

5 Custom Fields

Clients and Users' **Demographics**

User Action History

✅ **Finished**

Evaluation

Modeling ⬇ **Predicting**

Unsupervised-Learning: Cluster/Group the similar custom field texts. (Kmeans, etc.)

Rank the clusters by appearance frequency

Supervised-Learning (Predictive Modeling): Select the top 20 custom fields and explore how users are similar in each custom field (RandomForest, Logistic Regression, etc.)

Evaluation

Visualization (Looker/React)

THANKS!

☺

Any questions?