



# Anomaly Detection on Procurement Data

Penn Data Science Group

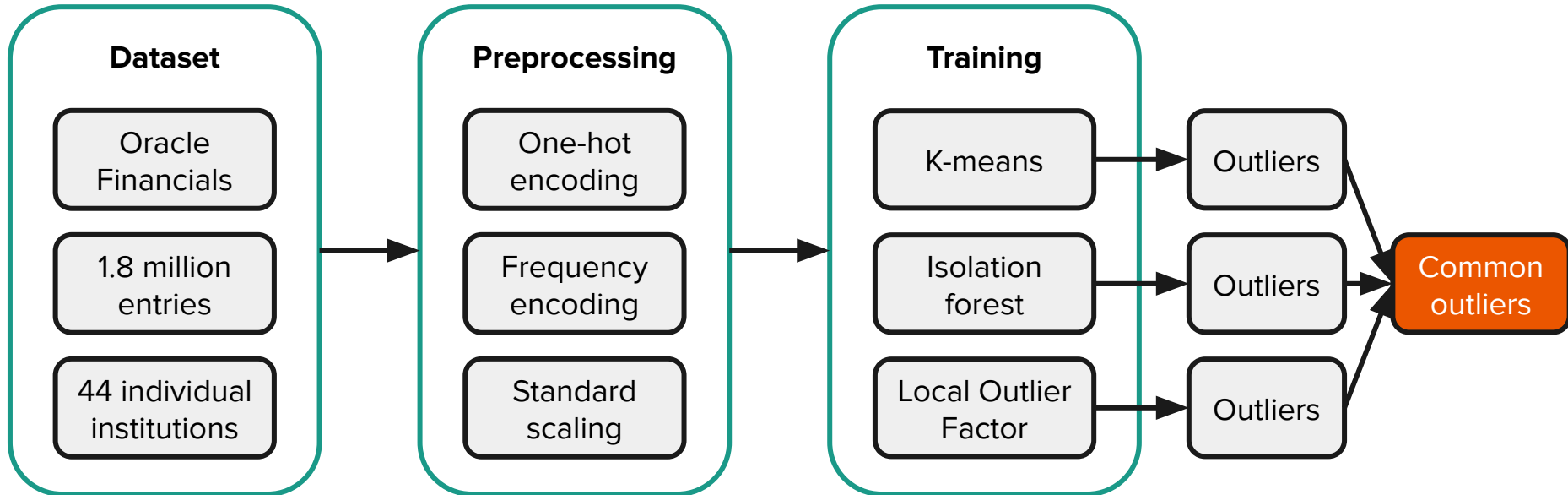


# BEN Dataset

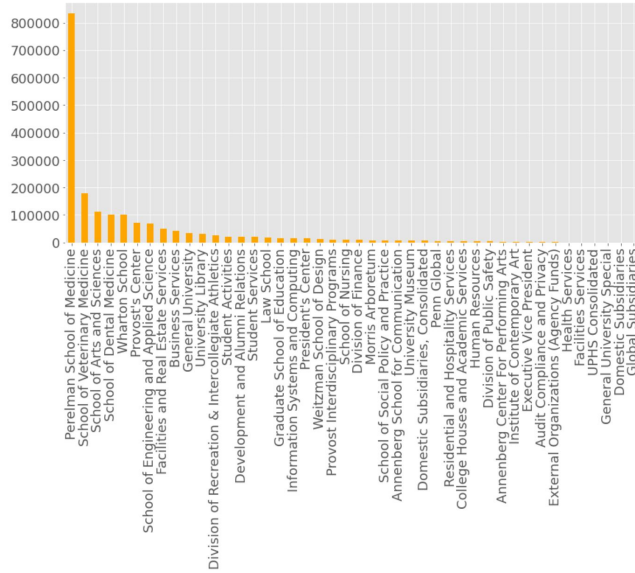
Jasper Huang and Arth Talati

# General Methodology

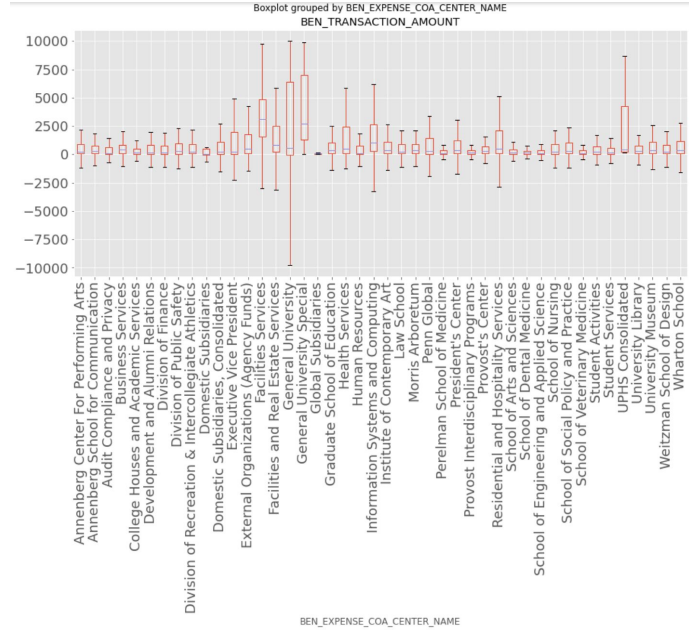
- **Objective:** find anomalous transactions, or “outliers”
- Apply three techniques and find common outliers



# Exploratory Data Analysis - BEN Dataset



Transactions per center



Distribution of transaction amounts grouped by center (<\$10,000)

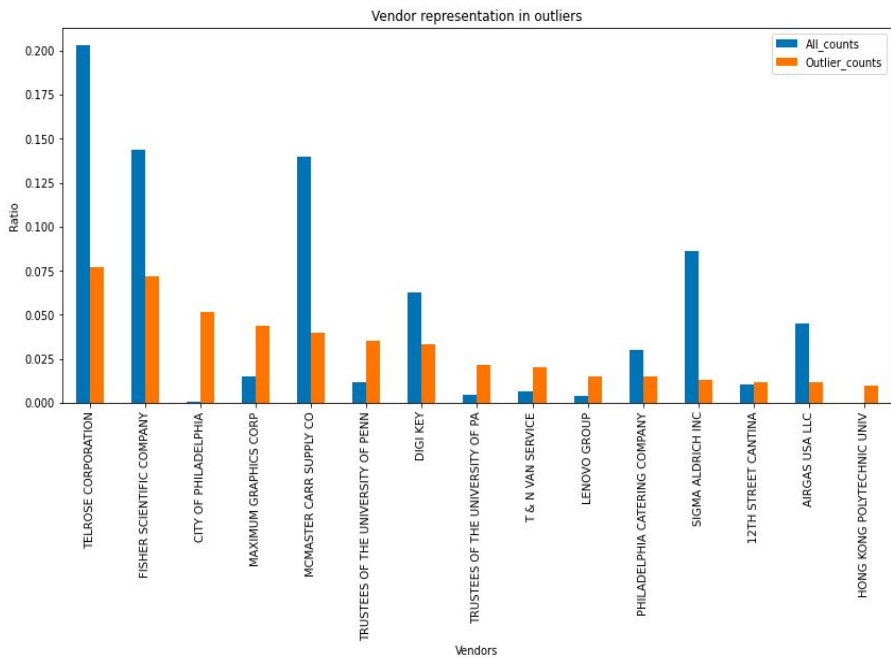
1. Transactions per center
  - a. School of Medicine dominates in number of transactions
2. Distribution of transaction amounts per center
  - a. Differing distributions

Rationale for doing analyses by individual centers

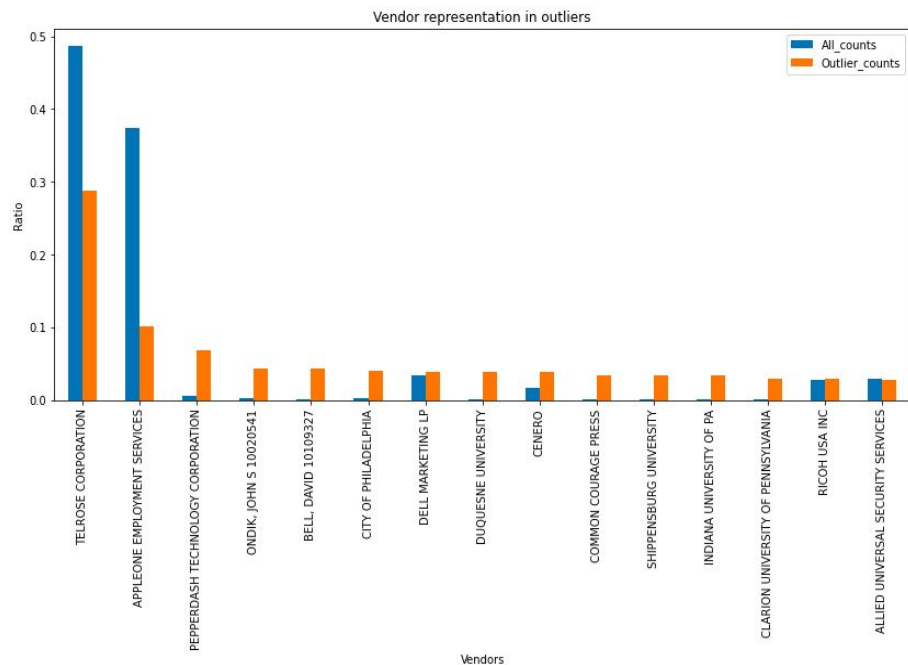
# Distribution of Outliers



SEAS



Wharton



# Rankings - Vendor name



SEAS		Wharton	
Vendor	% Outlier	Vendor	% Outlier
TELROSE CORPORATION	46/6900	APPLEONE EMPLOYMENT SERVICES	47/10226
HONG KONG POLYTECHNIC UNIV	6/7	ALLIED UNIVERSAL SECURITY SERVICES	13/796
CITY OF PHILADELPHIA	31/31	BUCKNELL UNIVERSITY	12/45

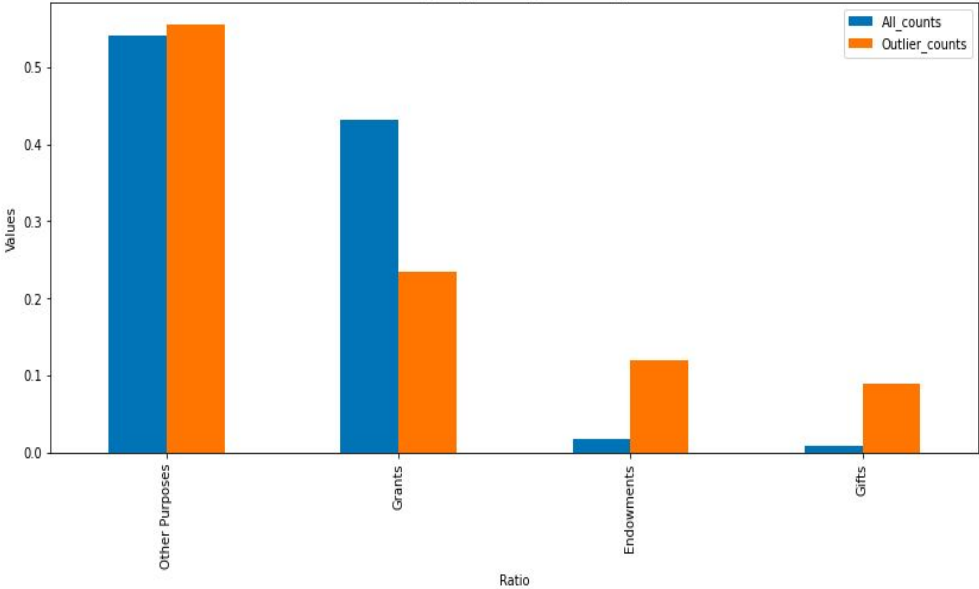
# Rankings - Fund type



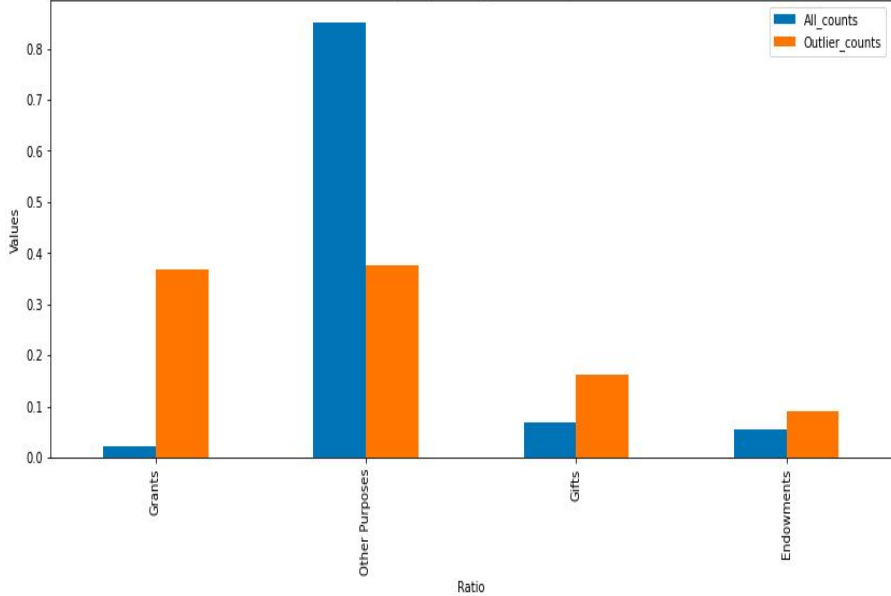
SEAS

Wharton

Comparing Fund Type Percentage



Comparing Fund Type Percentage





# Understanding What Contributes to Outliers

- We want to see which attributes **contribute most** to whether a transaction is flagged as an outlier or not
- Methodology:
  - **Label** all transactions as 0 or 1 based on anomaly detection output
  - **Train** another classifier on the labeled dataset
  - **Rank** feature importances using the classifier

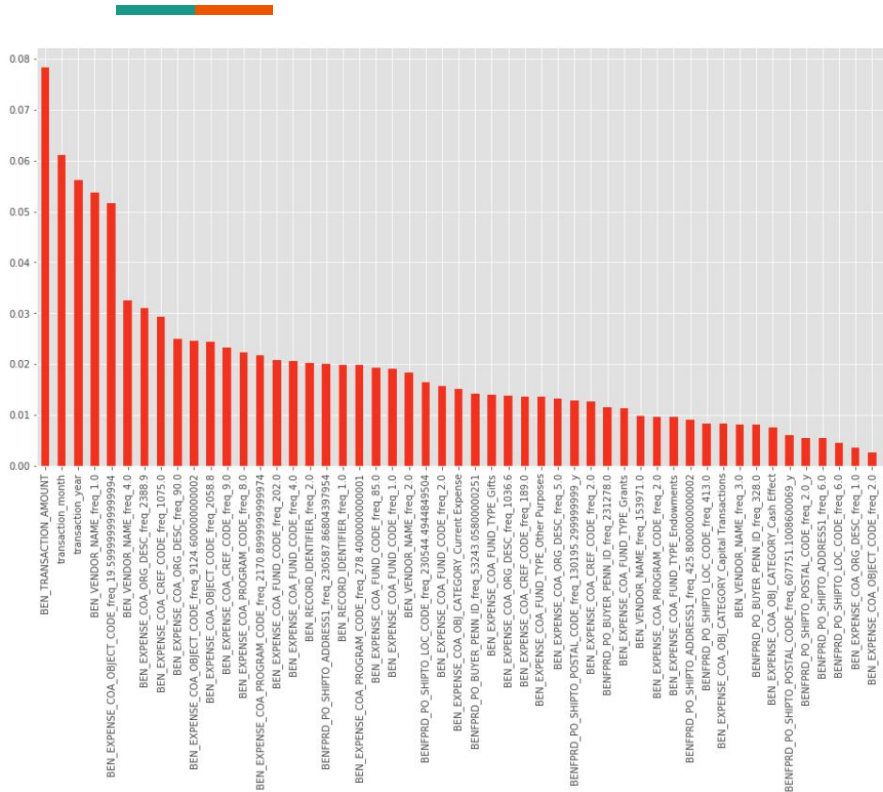




# Random Forest for Feature Importance Ranking

- Random forest classifier
  - Default 100 estimators
- Non-outliers = 0, outliers = 1
- Cross-validation
  - Scoring: **ROC\_AUC** and **weighted F1** due to imbalanced dataset

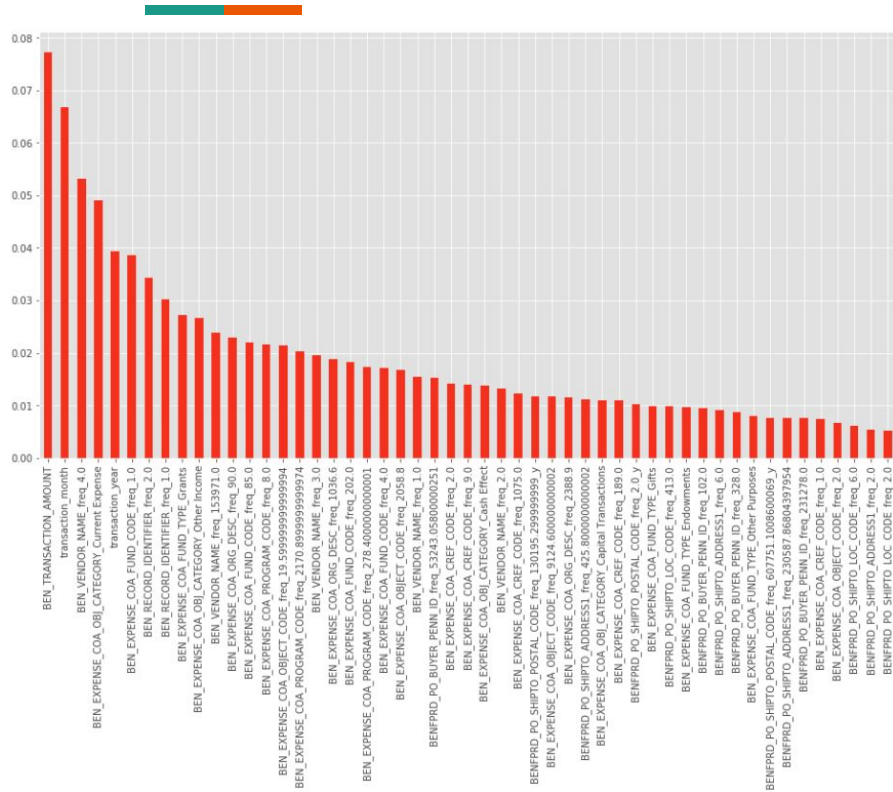
# Feature Importance Ranking - SEAS



	importance
<b>BEN_TRANSACTION_AMOUNT</b>	<b>0.078397</b>
<b>transaction_month</b>	<b>0.061150</b>
<b>transaction_year</b>	<b>0.056133</b>
<b>BEN_VENDOR_NAME_freq_1.0</b>	<b>0.053746</b>
<b>BEN_EXPENSE_COA_OBJECT_CODE_freq_19.599999999999994</b>	<b>0.051705</b>
<b>BEN_VENDOR_NAME_freq_4.0</b>	<b>0.032488</b>
<b>BEN_EXPENSE_COA_ORG_DESC_freq_2388.9</b>	<b>0.030926</b>
<b>BEN_EXPENSE_COA_CREF_CODE_freq_1075.0</b>	<b>0.029371</b>
<b>BEN_EXPENSE_COA_ORG_DESC_freq_90.0</b>	<b>0.024852</b>
<b>BEN_EXPENSE_COA_OBJECT_CODE_freq_9124.600000000002</b>	<b>0.024539</b>

- **Top Contributing Features:**
  - Transaction Amount (dollars)
  - Vendor names with freq. 0-1
  - COA expenses with freq. 2-19
  - Vendor names with freq. 3-4
  - Organisations with freq. 1000-2000

# Feature Importance Ranking - Wharton



	importance
BEN_TRANSACTION_AMOUNT	0.077344
transaction_month	0.066721
BEN_VENDOR_NAME_freq_4.0	0.053132
BEN_EXPENSE_COA_OBJ_CATEGORY_Current Expense	0.049009
transaction_year	0.039365
BEN_EXPENSE_COA_FUND_CODE_freq_1.0	0.038596
BEN_RECORD_IDENTIFIER_freq_2.0	0.034249
BEN_RECORD_IDENTIFIER_freq_1.0	0.030254
BEN_EXPENSE_COA_FUND_TYPE_Grants	0.027271
BEN_EXPENSE_COA_OBJ_CATEGORY_Other Income	0.026559

- Top Contributing Features:
  - Transaction Amount (dollars)
  - Vendor names with freq. 3-4
  - COA current expense with freq. 0-1
  - COA Fund codes with freq. 0-1



# Miscellaneous observations

- Louis Vuitton Transaction

PEW RESEARCH CENTER	1	180	1	2699	2519
SONKUSALE, SAMEER R	1	182	1	2699	2517
<b>LOUIS VUITTON</b>	1	183	1	2699	2516
KITANI, KRIS	1	184	1	2699	2515
ALSOTN, TREVOR	1	185	1	2699	2514



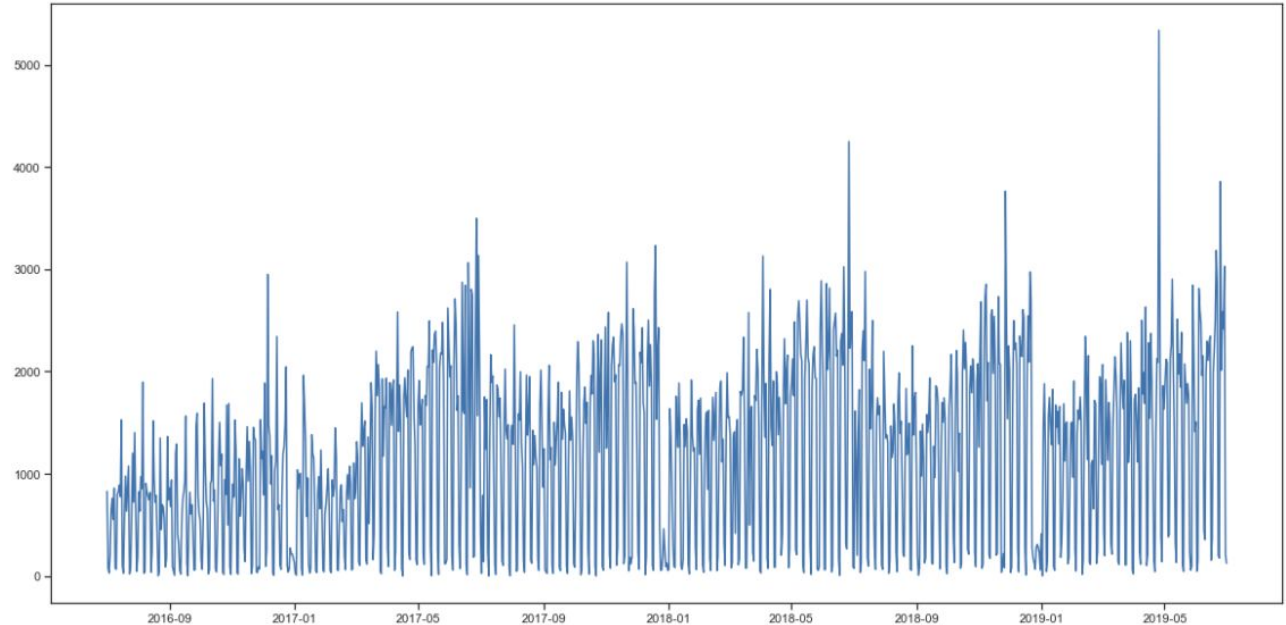
# Concur Dataset

Megha Mishra, Xuanyi Zhao, Yuhong Qin



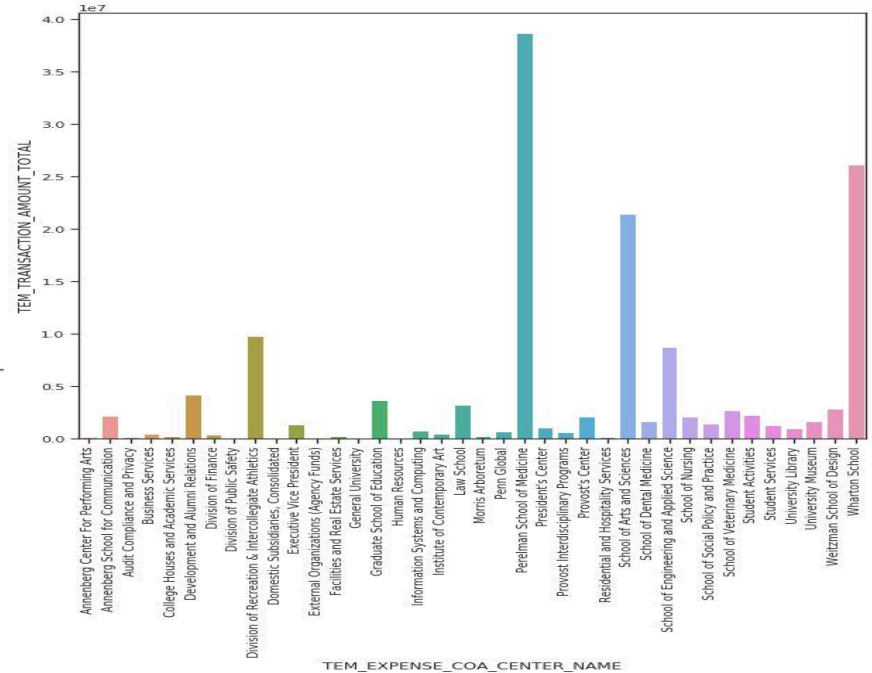
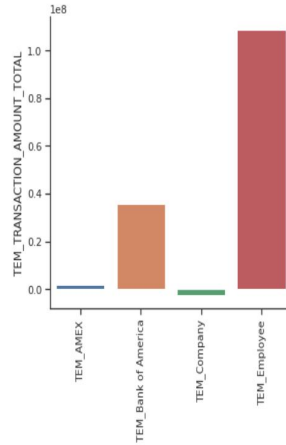
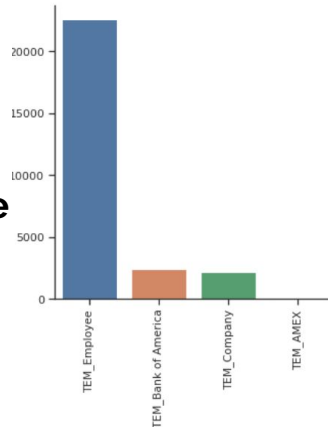
# Exploratory Data Analysis - Concur Dataset


**Seasonality of  
Transaction Frequency**



# Exploratory Data Analysis - Concur Dataset

Class Imbalance



Transaction Source 

School Distribution 

TEM\_EXPENSE\_COA\_CENTER\_NAME



# Stats for continuous columns outliers

	TEM_TRANSACTION_AMOUNT	weekly_rolling_mean	biweekly_rolling_mean	monthly_rolling_mean	TEM_EXPENSE_COA_OBJECT_DESC_quantile
count	12.000000	12.000000	12.000000	12.000000	12.000000
mean	129.884167	84512.877262	85036.996488	89004.587798	181.262917
std	315.635589	28746.530746	33534.600994	33004.866920	40.362682
min	9.880000	63387.337143	54652.061429	61343.387857	148.602500
25%	17.500000	63608.801429	54652.061429	61343.387857	148.602500
50%	33.200000	66158.925000	70972.286786	70900.710357	148.602500
75%	66.930000	108423.878571	116871.312857	126473.228214	226.987500
max	1128.000000	143374.225714	141122.460714	141409.707500	226.987500





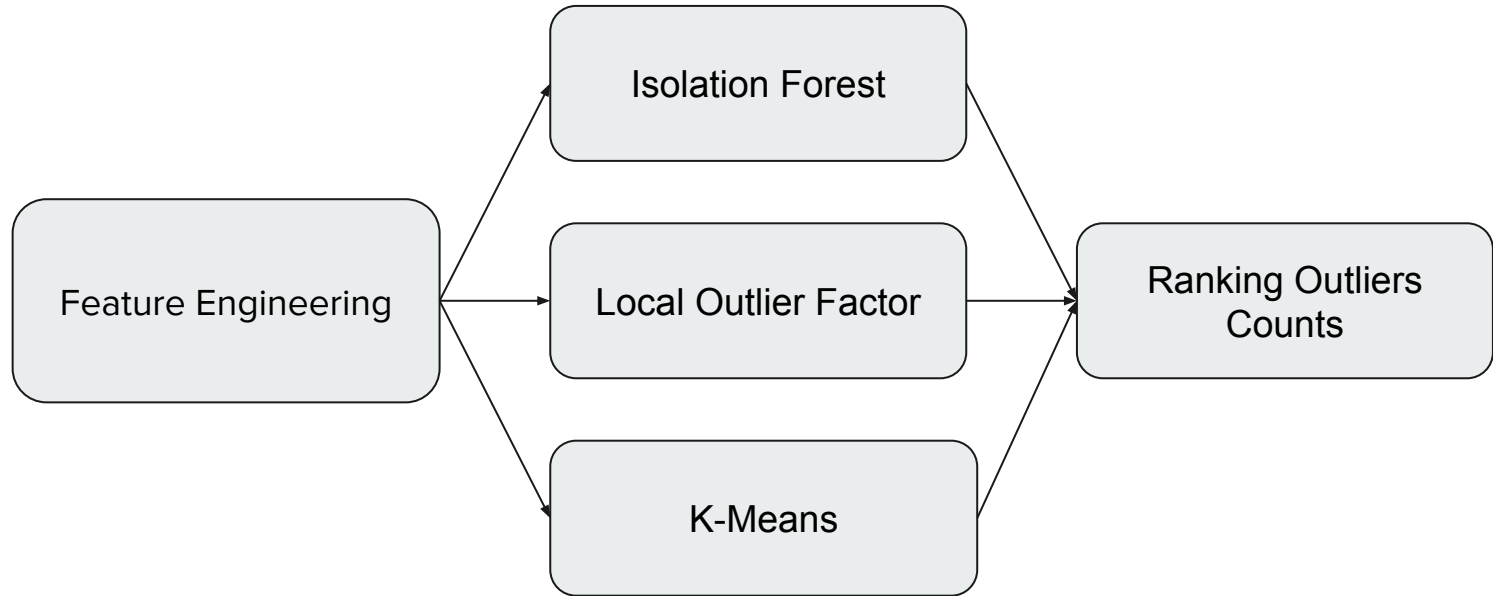
## Percentiles for continuous features of outliers

TRANSACTION_AMOUNT	Percentage(%)	Weekly_Rolling_Mean	Percentage(%)
60.22	2.7323	63387.34	5.5585
66.1	13.5903	63387.34	5.2001
26.06	7.2285	68709.05	1.6573
98.59	5.0587	108423.88	35.2372
69.42	49.9580	108423.88	35.2372
9.88	80.4547	143374.23	54.4075

Biweekly_Rolling_Mean	Percentage(%)	Monthly_Rolling_Mean	Percentage(%)
64523.7	3.7895	69071.13	3.8013
64523.7	4.6114	69071.13	4.2451
77420.87357	2.9385	72730.29	0.6000
116871.3129	45.7674	126473.23	35.6075
116871.3129	45.7674	126473.23	35.6075
141122.4607	51.5516	136388.95	46.9714



# Methods



# Results: Common outliers

VENDOR_NAME	Percentage(%)	EXPENSE_COA_CENTER_NAME	Percentage(%)	EXPENSE_COA_ORG_DESC	Percentage(%)
WHITE DOG	0.0537	School of Arts and Sciences	16.7100	History of Art	0.5572
WHITE DOG	0.0537	School of Arts and Sciences	16.7100	History of Art	0.5572
WORLD CAFE LIVE	0.0031	President's Center	0.8434	WXPN Surrogate	0.2309
CHARLIES BAKERY CAFE	0.0001	Weitzman School of Design	1.8582	Historic Preservation	0.1656
CITY TAP HOUSE PHILLY	0.0312	School of Arts and Sciences	16.7100	Center for Programs in Contemporary Writing	0.1470
SQ *ELIXR COFFEE LL	0.0009	Institute of Contemporary Art	0.6318	ICA Exhibitions	0.3767

EXPENSE_COA_FUND_TYPE	Percentage(%)	PERSON_IDENTIFIER	Percentage(%)
Other Purposes	71.7326	16334725	0.0564
Other Purposes	71.7326	16334725	0.0564
Other Purposes	71.7326	73377171	0.0020
Grants	18.0958	10008000	0.0463
Gifts	4.4302	10087803	0.0240
Endowments	5.7415	56038183	0.0193

\* For the top outliers, the transaction sources are all **TEM\_Bank of America** or **TEM\_EMPLOYEE**, the expense type are all **Business Meal (attendees) - Breakfast/Lunch or Parking**, the expense\_coa\_object\_desc are all **BUS MEALS** or **DT/DF CURR EXP**,



# Methods: Ranking difference

The common outliers number for all three models is too small (just 12). Therefore, we used ranking method to describe results.

**Rank increase = Counts Rank in Outliers - Counts Rank in whole data**

We ranked a particular column based on its count in the original dataset and on the dataset having outliers predicted by the algorithm. Then we compared the increases in rank which told us about occurrence of particular item in the predicted outliers, and found:

- **Main categories with high risk** [ordered by outlier counts and picked high rank increase]
- **Rare categories** [ordered by rank increase]



# Main Categories of Vendor

LOF		Isolation Forest		K-Means	
Vendor Name	%outlier	Vendor Name	%outlier	Vendor Name	%outlier
HOMEWOOD SUITES LEXINGTON	40/162	AXIS PIZZA	152/248	Club Quarters - Philad	70/111
HOLIDAY INN EXPRESS & SU	52/221	WHITE DOG	108/676	Ritz-Carlton	37/40
CASH	36/138	SQ *UNITED BY BLUE	70/222		

\*All of these vendor names have relatively high rank increase in the top of outlier vendor names.



# Rare Categories of Vendor

LOF		Isolation Forest		K-Means	
Vendor Name	%outlier	Vendor Name	%outlier	Vendor Name	%outlier
Victoria Guesthouse	11/11	CAMPO'S	12/13	Delta Dallas Convention Center	12/12
EXPEDIA 7437382010565	11/11	MCO Congres	15/15	NJ-Cherry	12/12
Hotel San Felipe	15/15	THE AMERICAN BOARD OF RA	15/16	Adobes Websales	11/13
AIRBNB * HMAF2C5MYR	20/21	Cafe Pan	15/17		



# Main Categories of Organization

LOF		Isolation Forest		K-Means	
School Name	%outlier	School Name	%outlier	School Name	%outlier
Lightweight Crew Team	124/1503	Institute of Contemporary Art	144/1144	Greenfield Intercultural Center Operating	385/1238
Baseball Team	120/2036	Museum Director's Office	151/2314	Weitzman School of Design	288/2755
Women's Track Team	103/2082	OT-Otorhinolaryngology	220/4623	Exhibits	90/210
Graduate Office	181/3570				

\*All of these organizations have relatively high rank increase in the top of outlier vendor names.



## Main Categories - School

LOF		Isolation Forest		K-Means	
School Name	%outlier	School Name	%outlier	School Name	%outlier
President's Center	126/8766	Institute of Contemporary Art	1439/7809	Institute of Contemporary Art	488/7809
University Museum	97/8116	University Museum	271/8116	Annenberg School for Communication	537/16573
Penn Global	47/5313				





# Main Categories of Expense Type

LOF		Isolation Forest		K-Means	
Expense Type	%outlier	Expense Type	%outlier	Expense Type	%outlier
Airfare	295/18336	Subscriptions	48/4634	Currency Exchange Fees	358/16552
Miscellaneous	299/23883	Bus	19/34	Intl Project/Program Costs	213/1560
Tolls	111/6682	Business Meal (attendees) - Breakfast/Lunch	4255/33446		



# Feature Importance Ranking

Top Categorical Features	Top Numerical Features
Fund type - Gifts	Transaction amount
Expense type - Intl Project/Program Cost	Weekly/biweekly/monthly rolling mean
Expense type - Hotel Tax	Organization frequency
School - Perelman School of Medicine	Vendor name frequency



# Concur Dataset Summary

1. Transaction amount related features were found to be important in deciding the outliers
2. Pay more attention on abnormal high frequency for some features, like rare vendor/organization/expense\_type but have high transaction frequency in specific months/weeks
3. Some extremely rare vendor names might be abnormal transactions

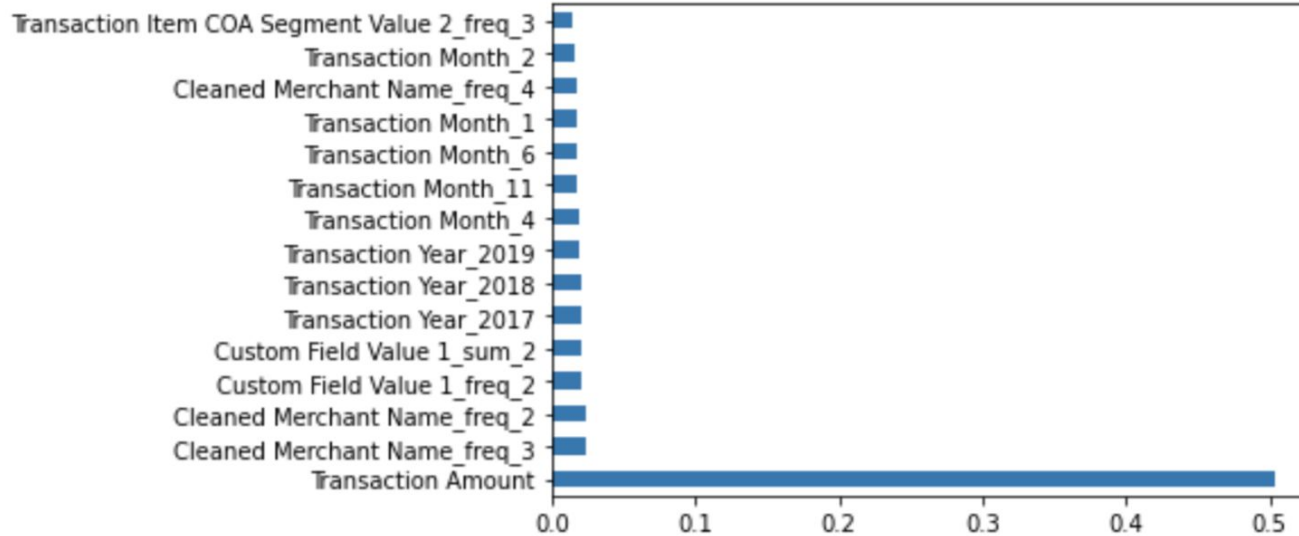


# PCard Dataset

Siyun Hu and Jessie Dong



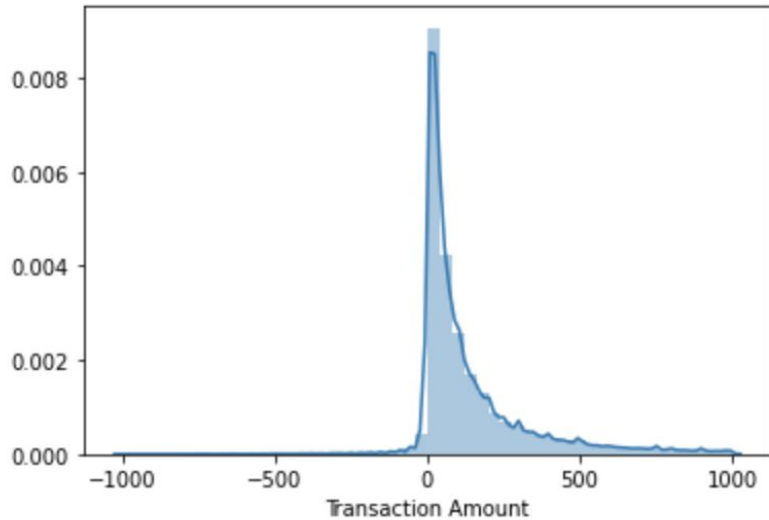
# Feature Importance



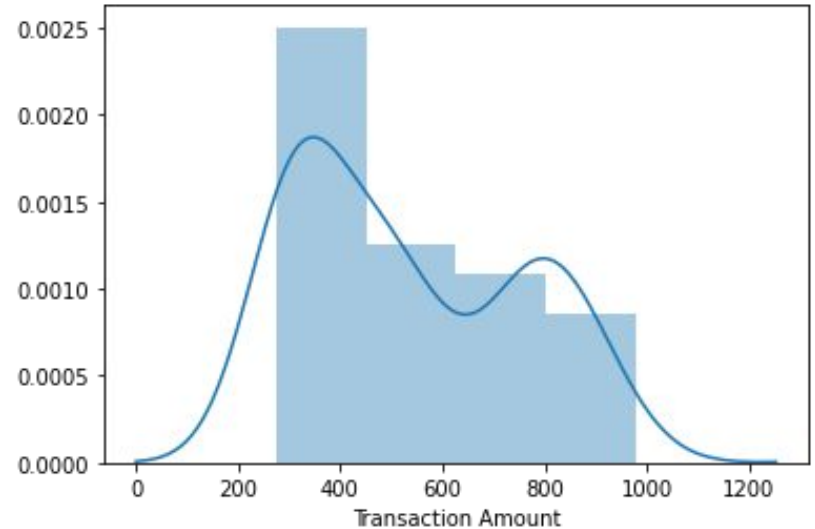


# Insight 1 - transaction amount

Transaction amount of whole dataset

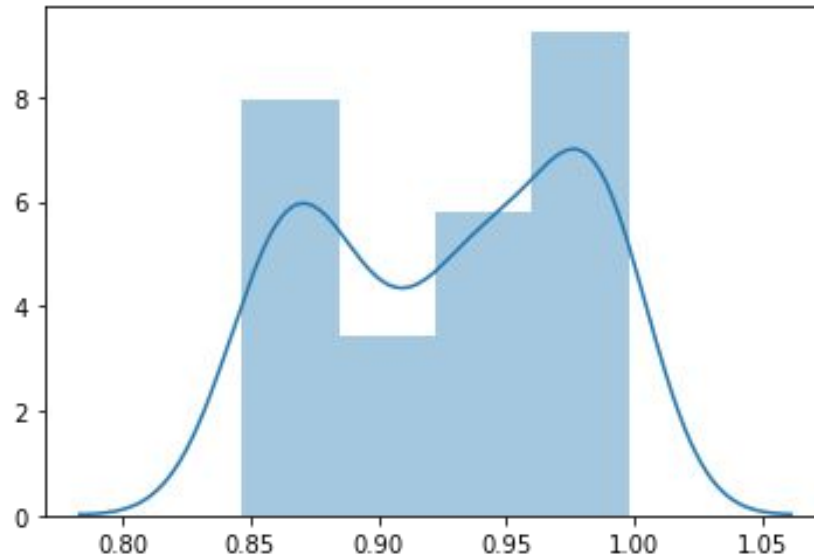


Transaction amount of outliers



# Insight 1 - continue

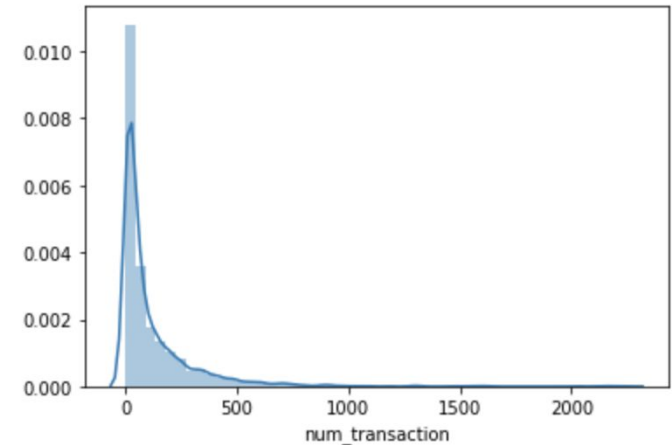
The percentile of transaction amount in outliers



## Insight 2- # transactions

	appearance	percentile
<b>count</b>	2331.000000	2331.000000
<b>mean</b>	124.126555	0.496505
<b>std</b>	200.176906	0.291412
<b>min</b>	1.000000	0.000000
<b>25%</b>	18.000000	0.243243
<b>50%</b>	50.000000	0.498069
<b>75%</b>	149.000000	0.749464
<b>max</b>	2254.000000	0.999571

Number of transactions per customer





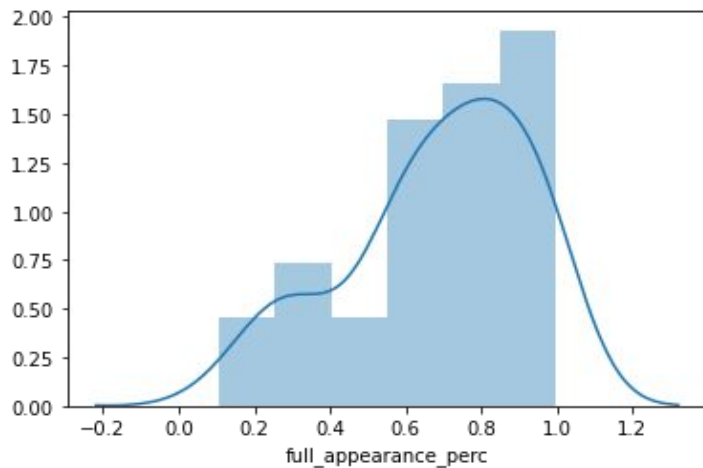


Customer_ID	#purchase in outlier	#purchase in full dataset
10052017	2	2115
63045049	2	2033
71361679	1	1601
45334879	1	1573
61555096	1	1462
39804367	1	847
10014797	1	791
10024452	1	659
10170786	1	514
10001760	1	465

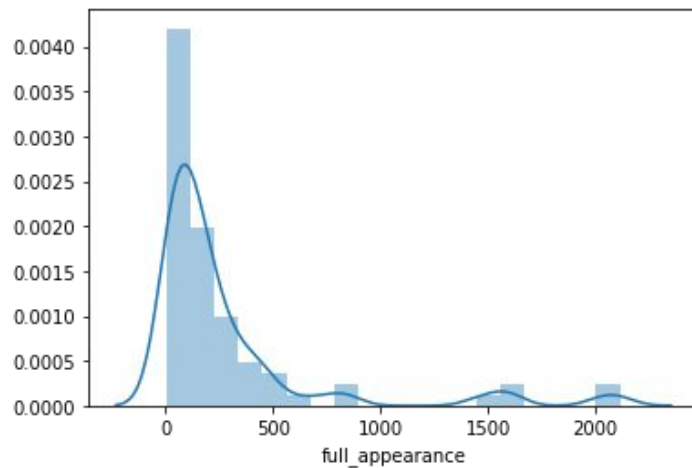


## Insight 2 continue

Percentile number of transactions per customer in outliers



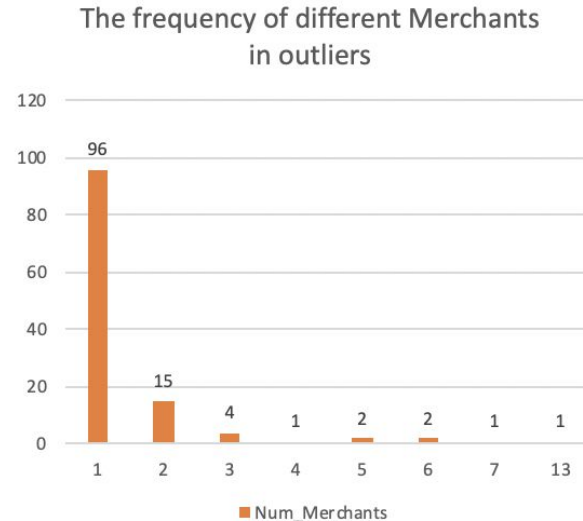
Number of transaction per customer in outliers



## Insight 3 - Merchant Name

The distribution of abnormal transactions across different Merchants

Appearance	Cleaned Merchant Name
1	96
2	15
3	4
4	1
5	2
6	2
7	1
13	1





## Insight 3 - Continue

Table: Top 10 merchants in abnormal transactions

Cleaned Merchant Name	Appearance	Original Rank	Current Ranking	Ranking Increase
PAYPAL	13	5	1	4
FACEBOOK	7	11	2	9
HILTON	6	252	3	249
SHERATON	6	266	3	263
AMAZON	5	1	5	-4
USDA	5	180	5	175
AT AND T	4	10	7	3
AAVMC	3	829	8	821
R AND K TOWING AND AUTO	3	72	8	64
VERIZON	3	4	8	-4



## Insight 3 - Continue

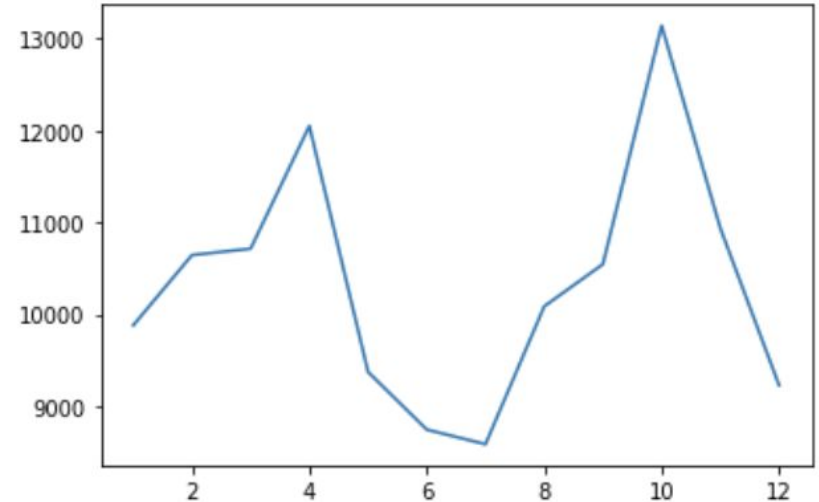
Table: The proportion of abnormal transactions in different Merchants

Cleaned Merchant Name	Appearance(Outliers)	Appearance(Complete)	Proportion
BIO-SYNTHESIS	1	1	100.0%
STATISTICAL INNOVATION	1	1	100.0%
X AND O EVENTS L	1	1	100.0%
CAN AM PIZZA	1	1	100.0%
GOOFY PHOTO BOOTH	1	1	100.0%
MCC	1	2	50.0%
SH SHANG CHANG BO TE M	1	2	50.0%
RACEPARTNER	1	2	50.0%
TOTAL VIDEO PRODUCTS I	1	2	50.0%
PAYCHEX	1	2	50.0%
COLTS MANUFACTURING	1	2	50.0%
NORMAN CARPET CO	1	2	50.0%
CJM ENGINEERING.COM	2	7	28.6%
GREEN GUARD FIRST AID	2	9	22.2%
SAGE SCIENCE	1	5	20.0%
AVIXA	1	5	20.0%
NATIONAL ORAL HEALTH C	1	5	20.0%
ANIMAL REPRODUCTION SY	1	6	16.7%
ICONOSQUARE	1	6	16.7%
SIGNS OF ART GRAPHIC D	1	6	16.7%

## Insight 4 - Months

Table 1: The number of abnormal transactions over months    Graph 1: The total transaction amount over months (2018)

Month	Appearance	Original Rank	Current Ranking	Ranking Increase
Jan.	35	8	1	7
Jun.	32	11	2	9
Apr.	29	2	3	-1
Feb.	22	5	4	1
May	19	9	5	4
Nov.	18	3	6	-3
Dec.	9	10	7	3
Mar.	8	4	8	-4
Oct.	5	1	9	-8
Jul.	4	12	10	2
Aug.	0	7	11	-4
Sep.	0	6	11	-5





## P-card results

By analyzing the 181 overlapping samples we found by isolation forest, LoF and K-Means algorithms, we conclude following possible insights:

- Transaction amount is the one of the key differences we found between normal and abnormal transactions. Abnormal transactions usually have excessively high transaction amount.
- The customers who have tremendous transactions are more likely to have abnormal transactions.
- The abnormal transactions we found happen to concentrate on several merchants, so that we could pay special attention to them in the future.
- The abnormal transactions we found concentrate on January, June, April, February.