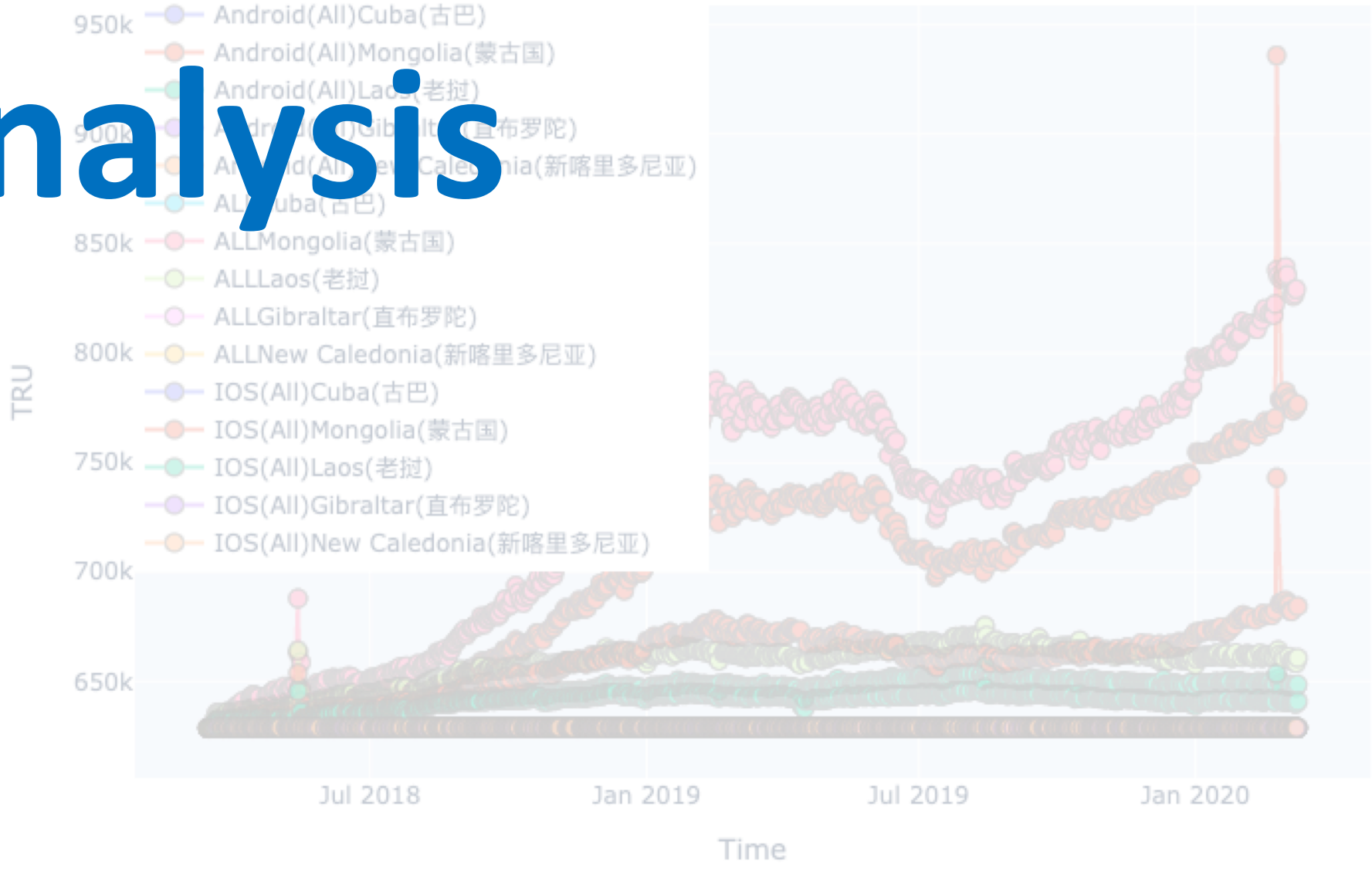
Topline Metrics Analysis



Outlines

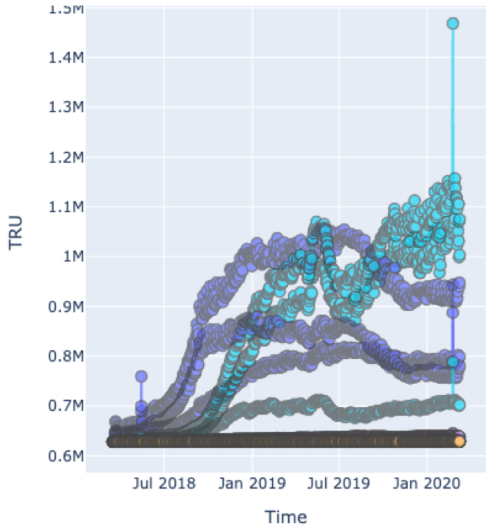
- ❑ Overview and Workflow
- ❑ Data Explorations
- ❑ Dashboard
- ❑ Time Series Prediction/Clustering
- ❑ Anomaly Detections
- ❑ TODO

Overall Workflow

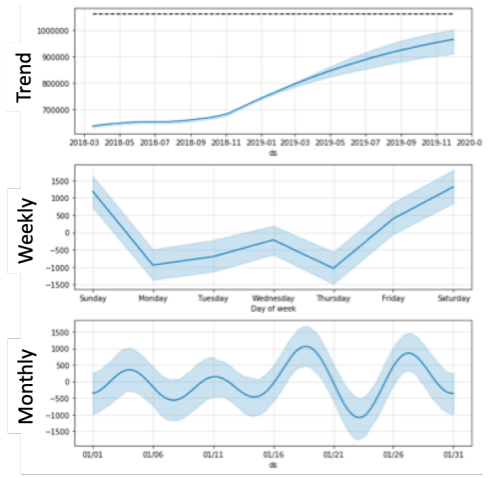
Initial Exploration to Identify Potential Problem → Interactive Visualization: Dashboard to Better Investigations → Time Series Analysis: Cleaning, Prediction, Clustering → Anomaly Detection: Within Single Time Series/Between Multiple Time Series

	TRU	DAU	
count	3.872870e+05	3.872870e+05	3.8
mean	7.405406e+05	1.814749e+04	5.0
std	6.836457e+05	1.706030e+04	4.0
min	6.286478e+05	1.529217e+04	4.3
25%	6.288102e+05	1.529677e+04	4.3
50%	6.314659e+05	1.537393e+04	4.3
75%	6.574378e+05	1.617778e+04	4.4
max	2.677675e+07	2.141935e+06	4.3

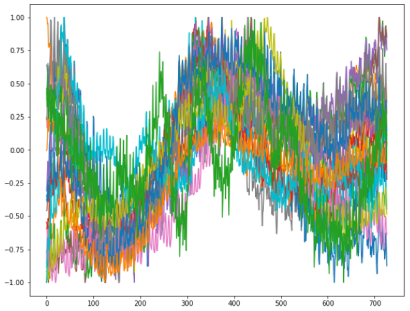
Stats Check



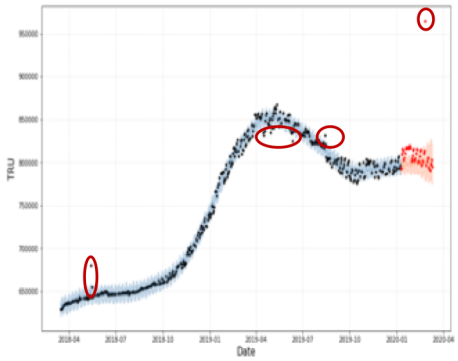
Visual Check

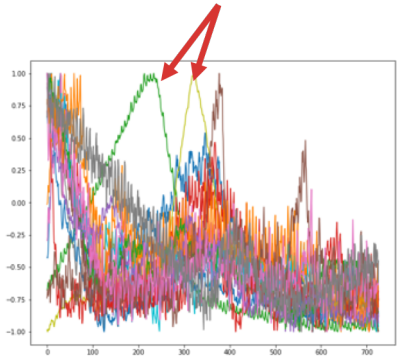


Decomposed Data



Prediction and Clustering





Anomaly Detection

Outlines

- ❑ Overview and Workflow
- ❑ Data Explorations
- ❑ Dashboard
- ❑ Time Series Prediction/Clustering
- ❑ Anomaly Detections
- ❑ TODO

Data Explorations

In this part, some basic investigations will be performed to get a initial idea about our data.

- Sanity check: with same platform, same country and same date, the data should be the same.
- Missing data check: check if any missing value, e.g. NAN, NULL.
- Duplication removal: remove duplicates
- Attribution checks: timespan, # of countries, # platform
- Outlier checks: identify abnormal data points

Remove Duplications

Origin Data
Rows: 748264

	TRU	DAU	Items	Trans
count	7.482640e+05	7.482640e+05	7.482640e+05	748264.00000
mean	7.399568e+05	1.816949e+04	5.089120e+04	3958.49364
std	6.776595e+05	1.801799e+04	4.610241e+04	3655.62504
min	6.286478e+05	1.529217e+04	4.342911e+04	3428.09000
25%	6.288102e+05	1.529677e+04	4.342911e+04	3428.09000
50%	6.314452e+05	1.537393e+04	4.352124e+04	3433.85000

After Remove Duplications
Rows: 387287
More than 50% duplications

	TRU	DAU	Items	Trans
count	3.872870e+05	3.872870e+05	3.872870e+05	387287.000000
mean	7.405406e+05	1.814749e+04	5.090994e+04	3958.958757
std	6.836457e+05	1.706030e+04	4.659028e+04	3653.633879
min	6.286478e+05	1.529217e+04	4.342911e+04	3428.090000
25%	6.288102e+05	1.529677e+04	4.342911e+04	3428.090000

Erroneous Data1

	TRU	DAU	Items	Trans	Items Per Trans	Items per DAU	Conversion	Cash Flow
count	3.872870e+05	3.872870e+05	3.872870e+05	387287.000000	387287.000000	387287.000000	387287.000000	387287.000000
mean	7.405406e+05	1.814749e+04	5.090994e+04	3958.958757	31.588136	1.415199	2.900178	5092.220355
std	6.836457e+05	1.706030e+04	4.659028e+04	3653.633879	23.878126	1.344340	2.579061	4656.117813
min	6.286478e+05	1.529217e+04	4.342911e+04	3428.090000	17.580000	1.300000	2.300000	4527.010000
25%	6.288102e+05	1.529677e+04	4.342911e+04	3428.090000	17.580000	1.300000	2.300000	4527.010000
50%	6.314659e+05	1.537393e+04	4.352009e+04	3433.850000	28.070000	1.320000	2.560000	4532.770000
75%	6.574378e+05	1.617778e+04	4.466252e+04	3522.530000	36.070000	1.390000	3.000000	4611.080000
max	2.677675e+07	2.141935e+06	4.387412e+06	434846.580000	1744.870000	444.240000	232.630000	819825.280000

!!! Min == 25% Quantiles

From the above summary of the table, we found that there are lots of erroneous data for many columns. Specifically, we can see that the *minimal value == 25% Quantiles* for many columns (except TRU and DAU, Time Spend, Return Customer).

*It is better to view these statistic by platform and country

Erroneous Data2

Negative Time Spend Per Day

	TRU	DAU	Items	Trans	Items Per Trans	Items per DAU	Conversion	Cash Flow	Return Customer	Time Spend Per Day(seconds)
count	3.872870e+05	3.872870e+05	3.872870e+05	387287.000000	387287.000000	387287.000000	387287.000000	387287.000000	387287.000000	387287.000000
mean	7.405406e+05	1.814749e+04	5.090994e+04	3958.958757	31.588136	1.415199	2.900178	5092.220355	39.803183	122.565397
std	6.836457e+05	1.706030e+04	4.659028e+04	3653.633879	23.878126	1.344340	2.579061	4656.117813	24.101990	372413
min	6.286478e+05	1.529217e+04	4.342911e+04	3428.090000	17.580000	1.300000	2.300000	4527.010000	12.210000	-9.990000
25%	6.288102e+05	1.529677e+04	4.342911e+04	3428.090000	17.580000	1.300000	2.300000	4527.010000	28.660000	98.840000
50%	6.314659e+05	1.537393e+04	4.352009e+04	3433.850000	28.070000	1.320000	2.560000	4532.770000	39.440000	118.180000
75%	6.574378e+05	1.617778e+04	4.466252e+04	3522.530000	36.070000	1.390000	3.000000	4611.080000	48.915000	139.370000
max	2.677675e+07	2.141935e+06	4.387412e+06	434846.580000	1744.870000	444.240000	232.630000	819825.280000	3912.090000	1390.060000

Extreme Values: Max > 100X Mean

*It is better to view these statistic by platform and country

Outlines

- ❑ Overview and Workflow
- ❑ Data Explorations
- ❑ Dashboard
- ❑ Time Series Prediction/Clustering
- ❑ Anomaly Detections
- ❑ TODO

Dashboard : Multiple Filters for Interactive Visualization

Filters:

Platform, Metrics, and Date Range and Continents/Countries

Dashboard for Topline Metrics

Filters

Platforms

- Android(All)
- ALL
- IOS(All)

Metrics

TRU | DAU | Items | Trans | Items Per Trans | Items per DAU | Conversion | Cash Flow | Return Customer | Time Spend Per Day(seconds)

Visulizations

Date Range: 03/15/2018 → 03/09/2020

You have selected: Start Date: March 15, 2018 | End Date: March 09, 2020

Y-scale: Linear | Log

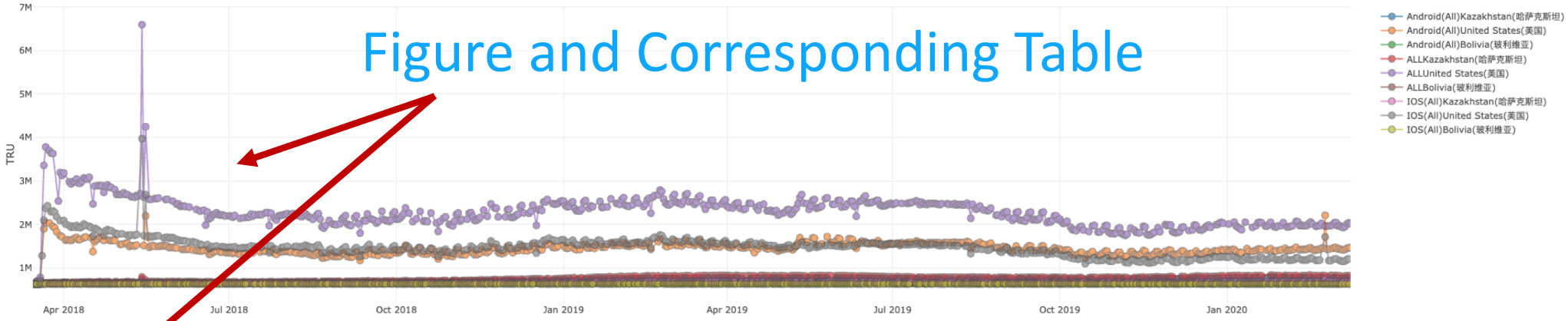
Scatter-Line Plot

Continents

- North America
- South America
- Asia
- Australia
- Africa
- Europe
- Antarctica

Countries: Kazakhstan(哈萨克斯坦) | United States(美国) | Bolivia(玻利维亚)

Figure and Corresponding Table

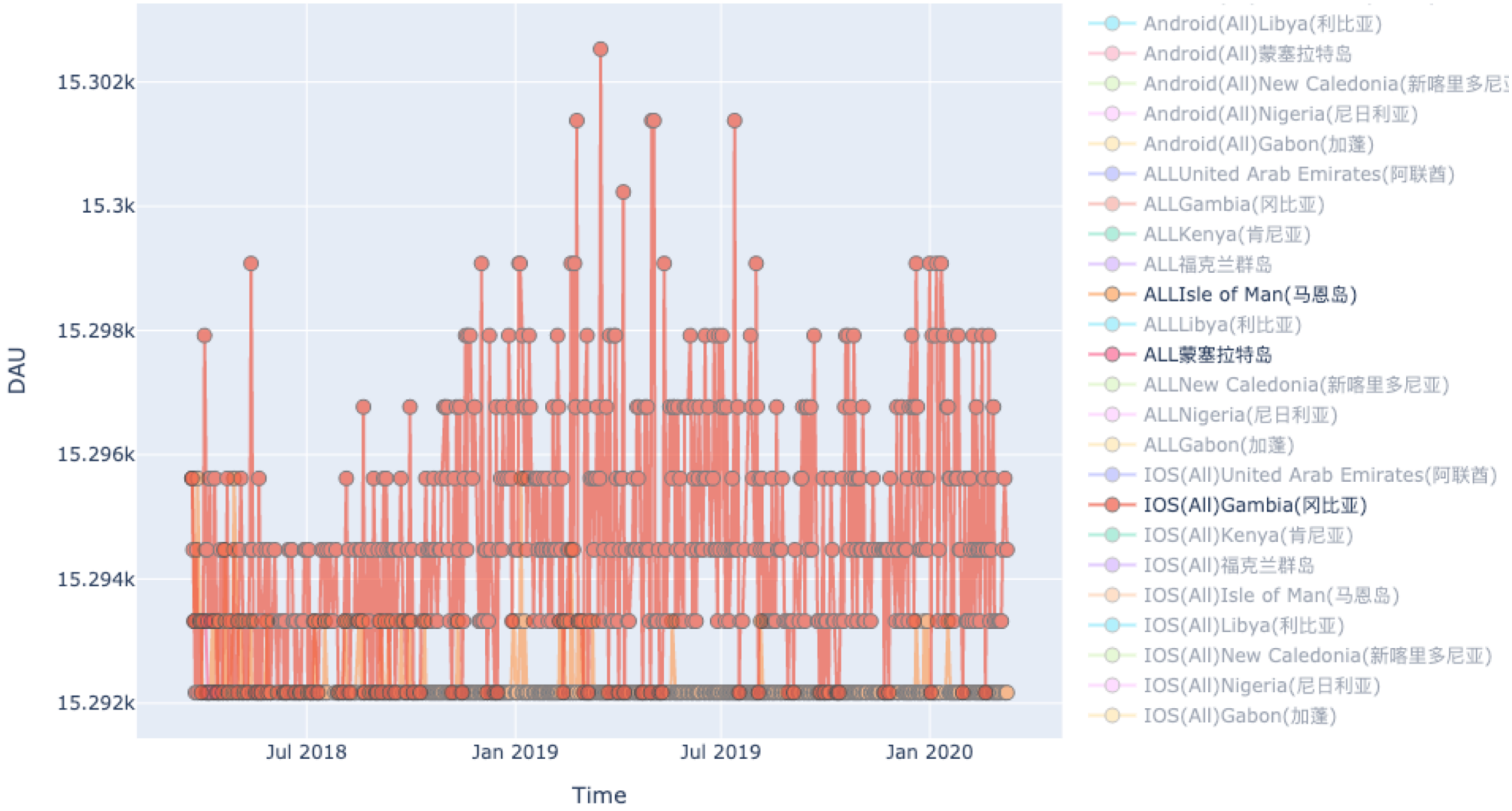


Tables

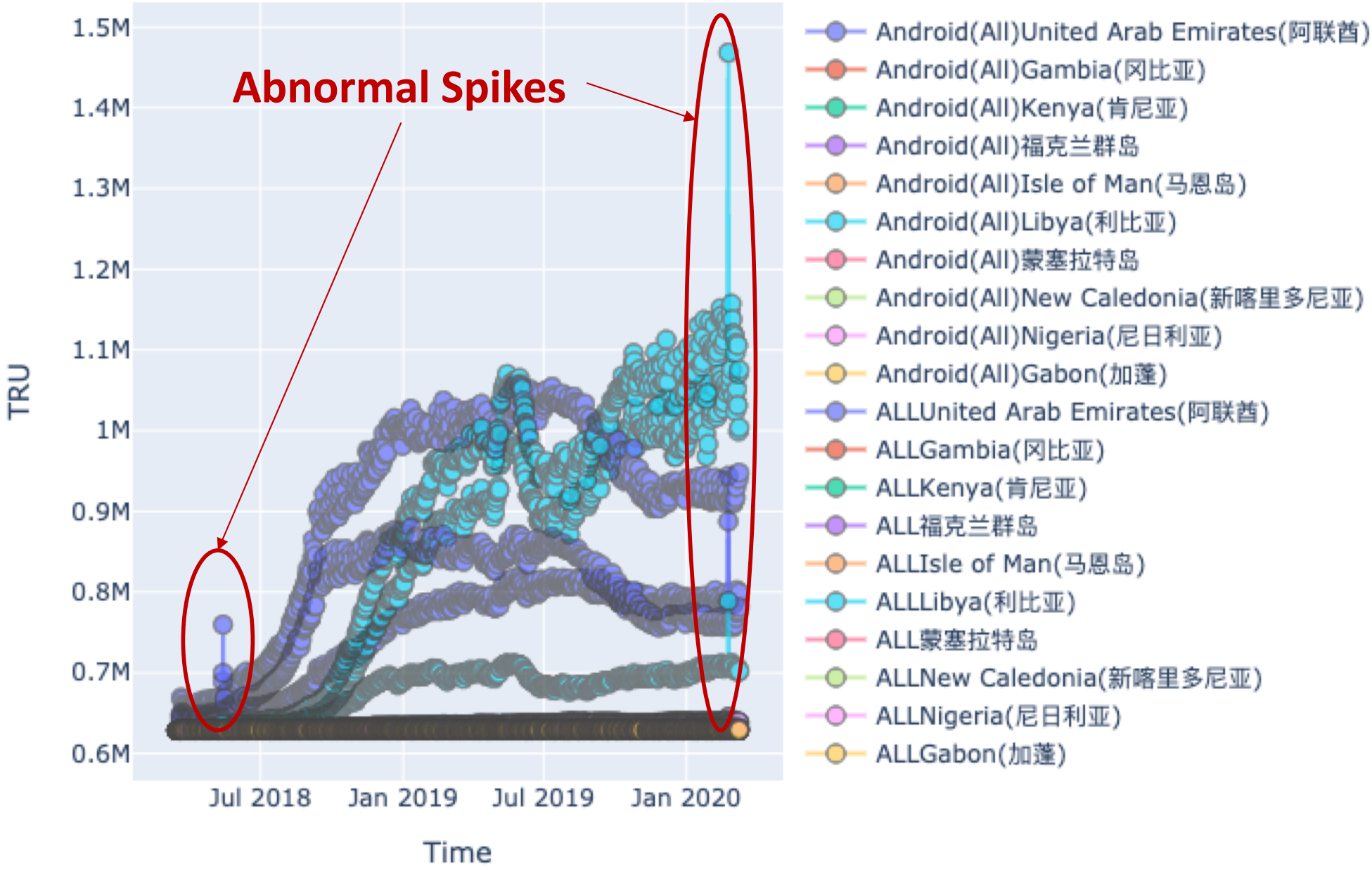
Selected Rows

Date.1	Platform	Country	TRU	DAU	Items	Trans	Items Per Trans	Items per DAU	Conversion	Cash Flow	Return Customer	Time Spend Per Day(seconds)
2018-03-15	Android(All)	Kazakhstan(哈萨克斯坦)	628852.81	15498.31	43429.11	3428.09	17.58	1.3	2.3	4527.01	79.12	58.18

Visual Check: Lower Bounded DAU



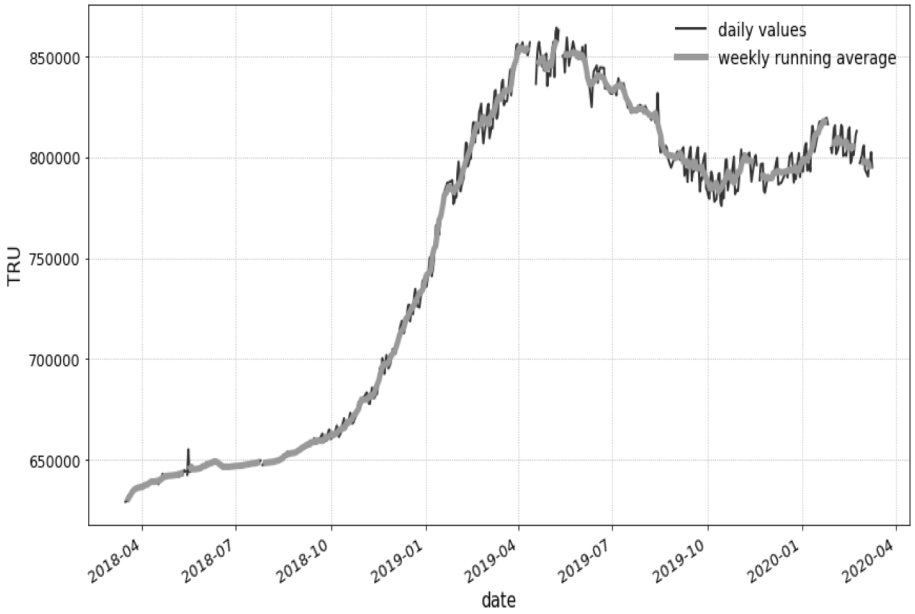
Visual Check: Abnormal Spikes

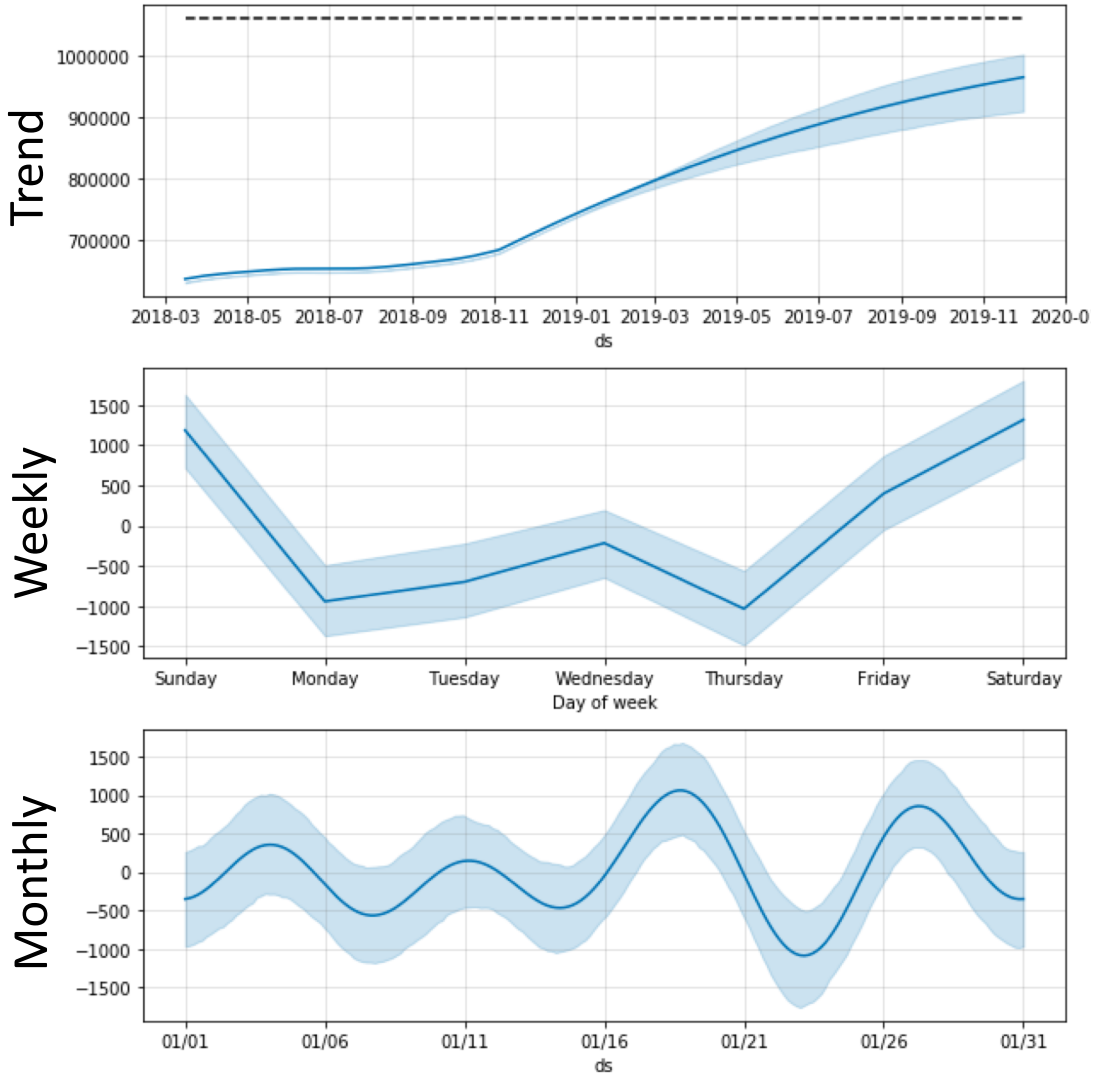


Outlines

- ❑ Overview and Workflow
- ❑ Data Explorations
- ❑ Dashboard
- ❑ Time Series Prediction/Clustering
- ❑ Anomaly Detections
- ❑ TODO

Key Idea: Decomposition



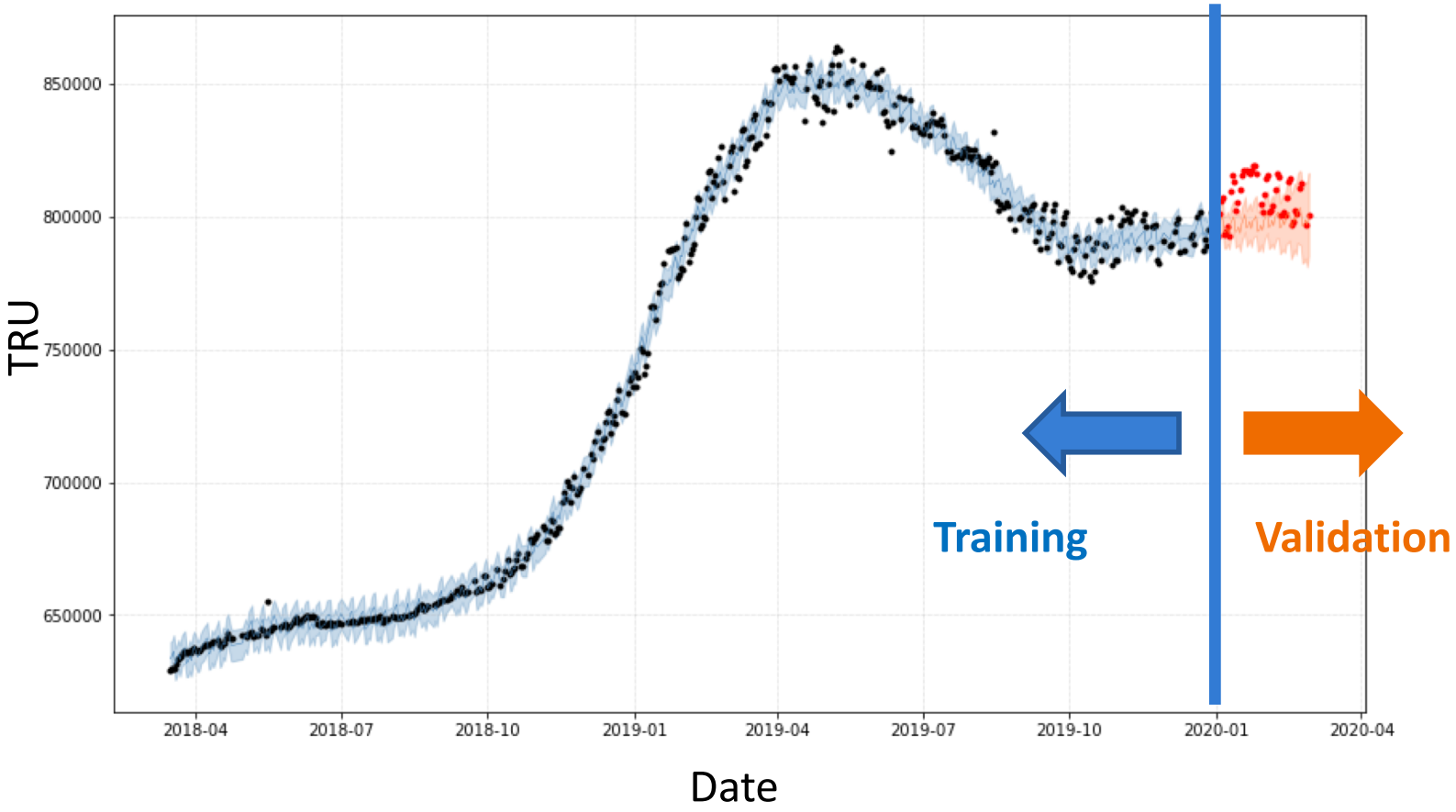


Original Data

Decomposed Data

* Prophet is used for periodical decompositions, which is based on FFT method.

Exemplary Predictions



Split the historical data into training and validation sets

Expanding Window for Hyperparameter Tuning

Illustration of expanding window for hyperparameter tuning

Specifically, last last three month's data are used for validation/forecasting sets to tune model's parameters

Evaluation Metrics

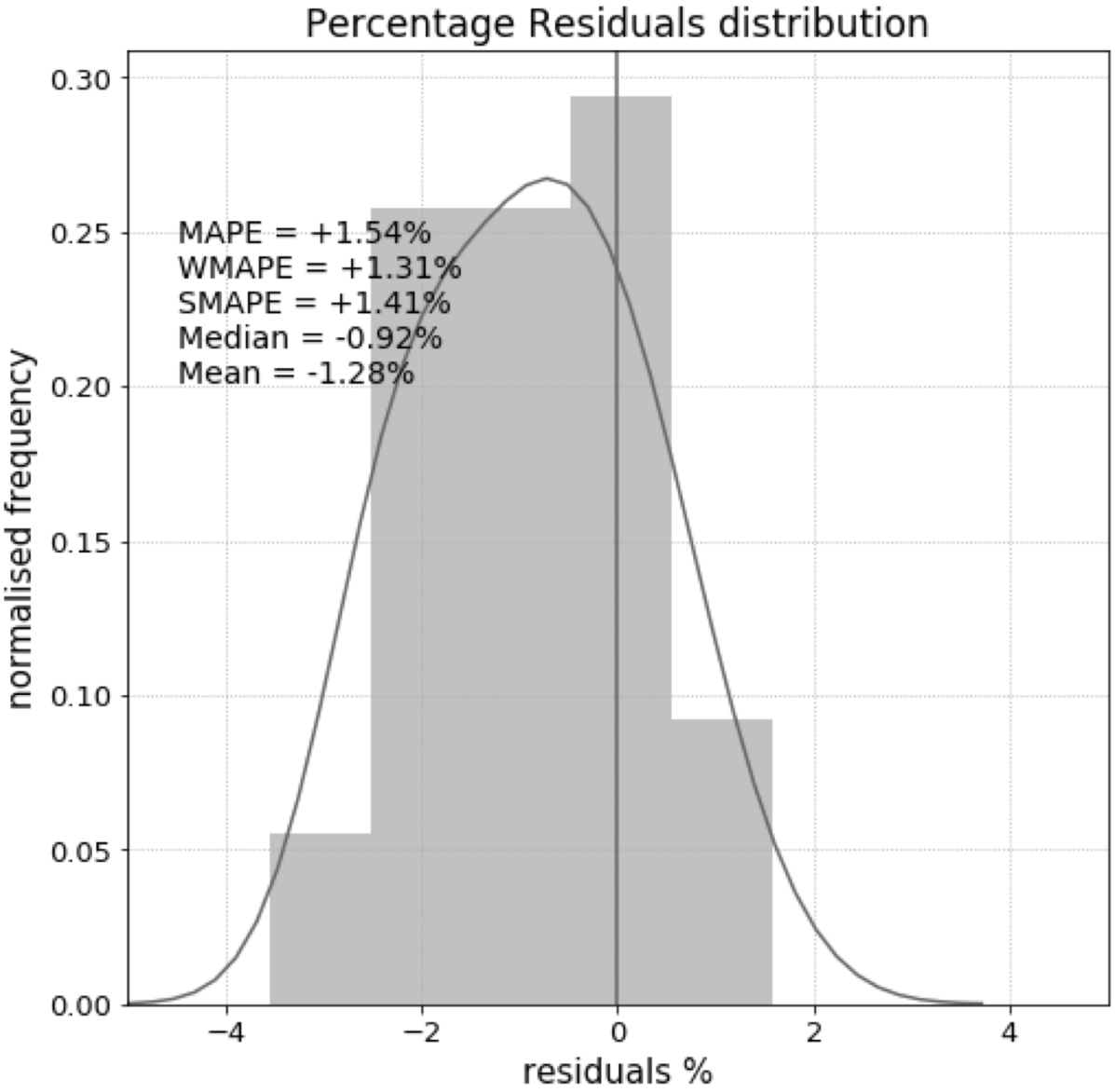
F: Forecasting A: Actual value

$$MAPE = Mean(\frac{100 * |F - A|}{A})$$

$$WMAPE = \frac{\sum 100 * |F - A|}{\sum A}$$

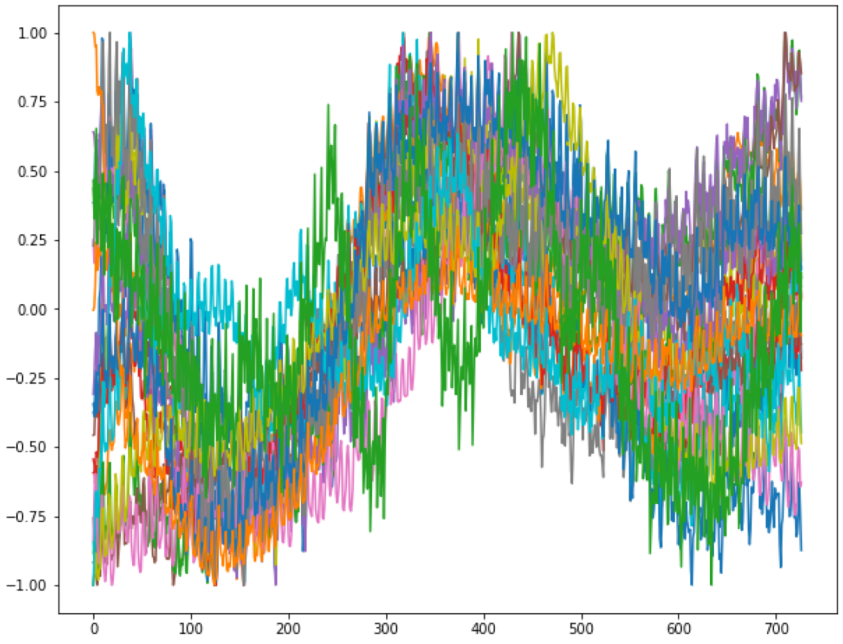
$$SMAPE = \frac{\sum 100 * 2 * |F - A|}{\sum (A + F)}$$

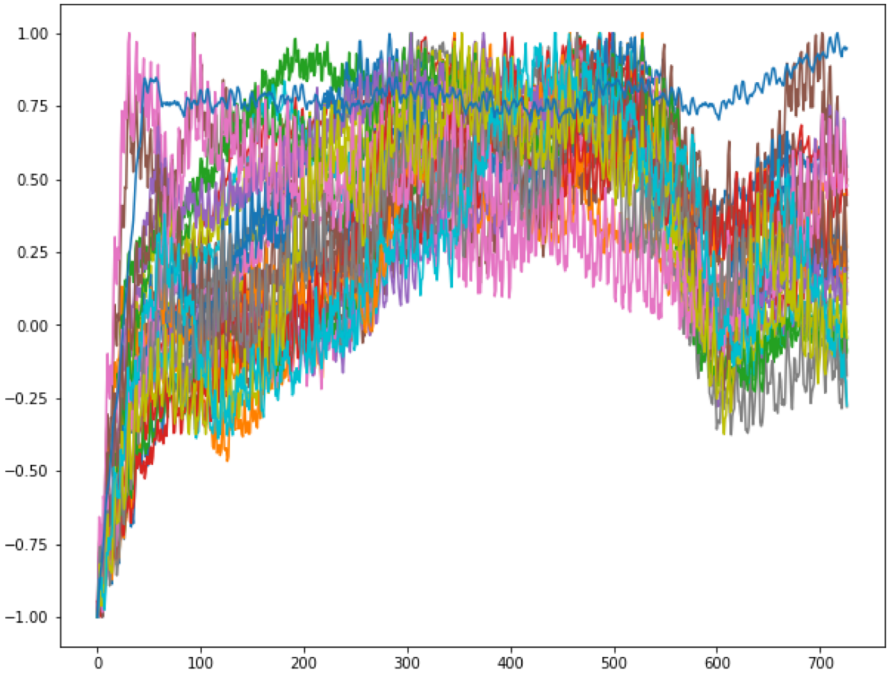
MAPE **is very sensitive to small A**, which will lead to biased result. WMAPE and SWMAPE is a more robust metrics. Thus, in this study **WMAPE and SMAPE** will be used for model evaluation.



Clustering Using Prediction and DCT as Features

Original Time Series → Time Series Prediction to Remove Outliers and Filling Missing Data → Discrete Cosine Transform to Get Features → Clustering Algorithms





Exemplary Clustered Time Series: Showing Similar Patterns

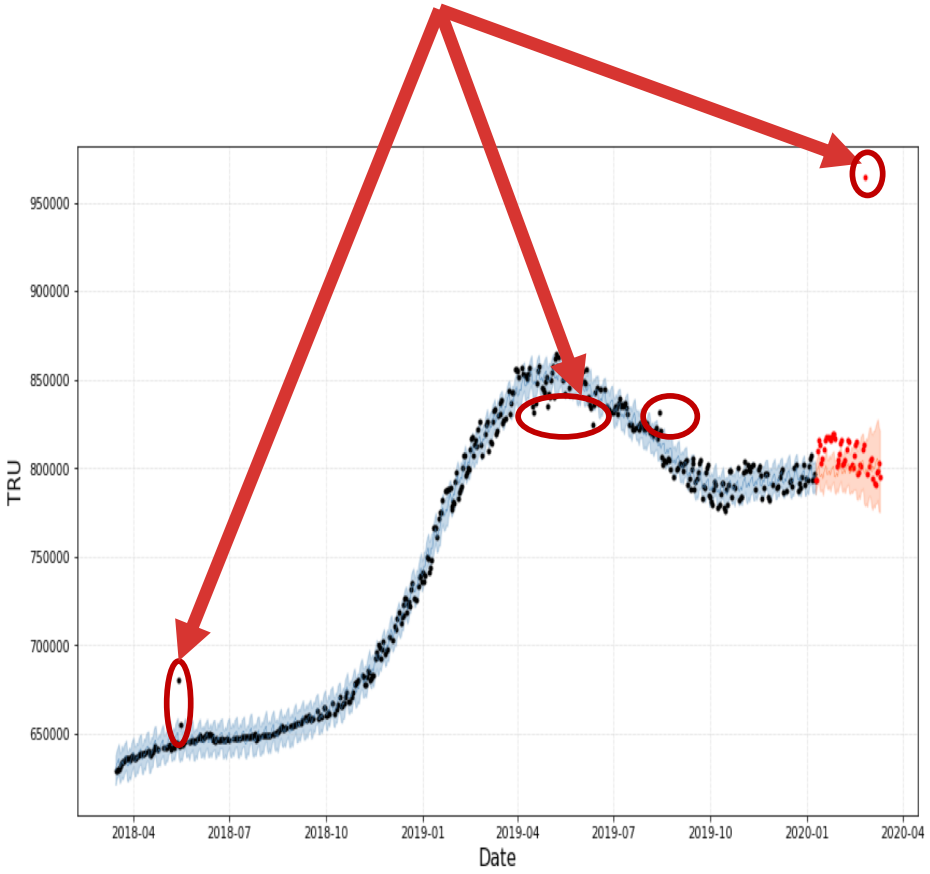
* All time series are scaled to [-1,1]

Outlines

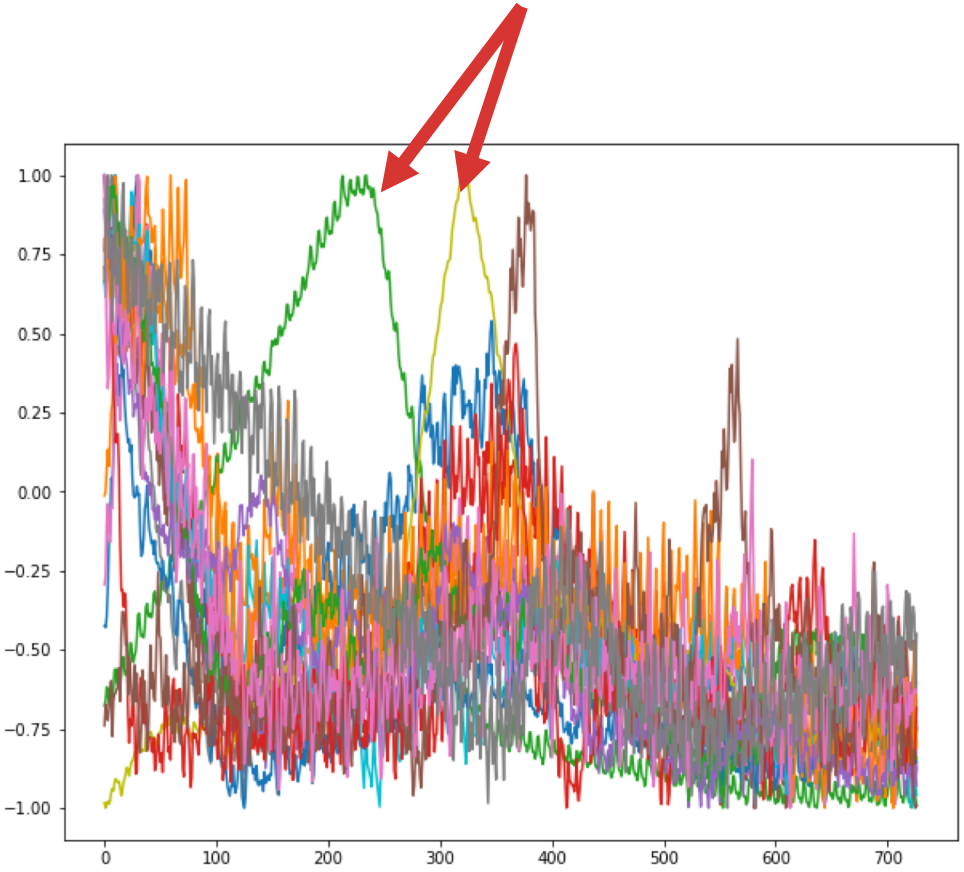
- ❑ Overview and Workflow
- ❑ Data Explorations
- ❑ Dashboard
- ❑ Time Series Prediction/Clustering
- ❑ Anomaly Detections
- ❑ TODO

Types of Anomaly

- Within Single Time Series Anomaly:
  => Using Prediction Based Anomaly Detections



- Between Multiple Time Series Anomaly:
  => Using Clustering Based Anomaly Detections



Outlines

- ❑ Overview and Workflow
- ❑ Data Explorations
- ❑ Dashboard
- ❑ Time Series Prediction/Clustering
- ❑ Anomaly Detections
- ❑ TODO

Feature Work

TO DO :

- More feature engineering
- More advanced hyper parameter tuning: Bayesian Optimization
- Try time more series predictions (SARIMA, STL, ES, Holter-Winter, Tbats, Thetams (M3 winning solution), LSTM/RNN(M4 winning solution)) to better capture seasonality
- Try more anomaly detection models: PyOD, Isolation Forest, SVM.
- Try ensemble with multiple models to reduce variance and improve generality.

Related Materials

- Traditional Methods: KNN, Isolation Forest, SVM: https://towardsdatascience.com/time-series-of-price-anomaly-detection-13586cd5ff46
- Traditional Methods: PyOD https://towardsdatascience.com/anomaly-detection-with-pyod-b523fc47db9
- Time Series: SARIMA, STL, ES, Holter-Winter https://blog.statsbot.co/time-series-anomaly-detection-algorithms-1cef5519aef2
- Time Series: SARIMA, STL, ES, Holter-Winter https://towardsdatascience.com/anomaly-detection-with-time-series-forecasting-c34c6d04b24a
- Time Series: AutoARIMA https://towardsdatascience.com/anomaly-detection-with-time-series-forecasting-c34c6d04b24a
- Google Paper: https://arxiv.org/pdf/1708.03665.pdf
- IOT Anomaly: https://arxiv.org/pdf/1812.00890.pdf
- Pinterest: https://medium.com/pinterest-engineering/building-a-real-time-anomaly-detection-system-for-time-series-at-pinterest-a833e6856ddd
- Uber: https://www.youtube.com/watch?v=VYpAodcdFfA&list=PL-pMQDw5Rryn37-msYa_ILELAmmwUrw3_