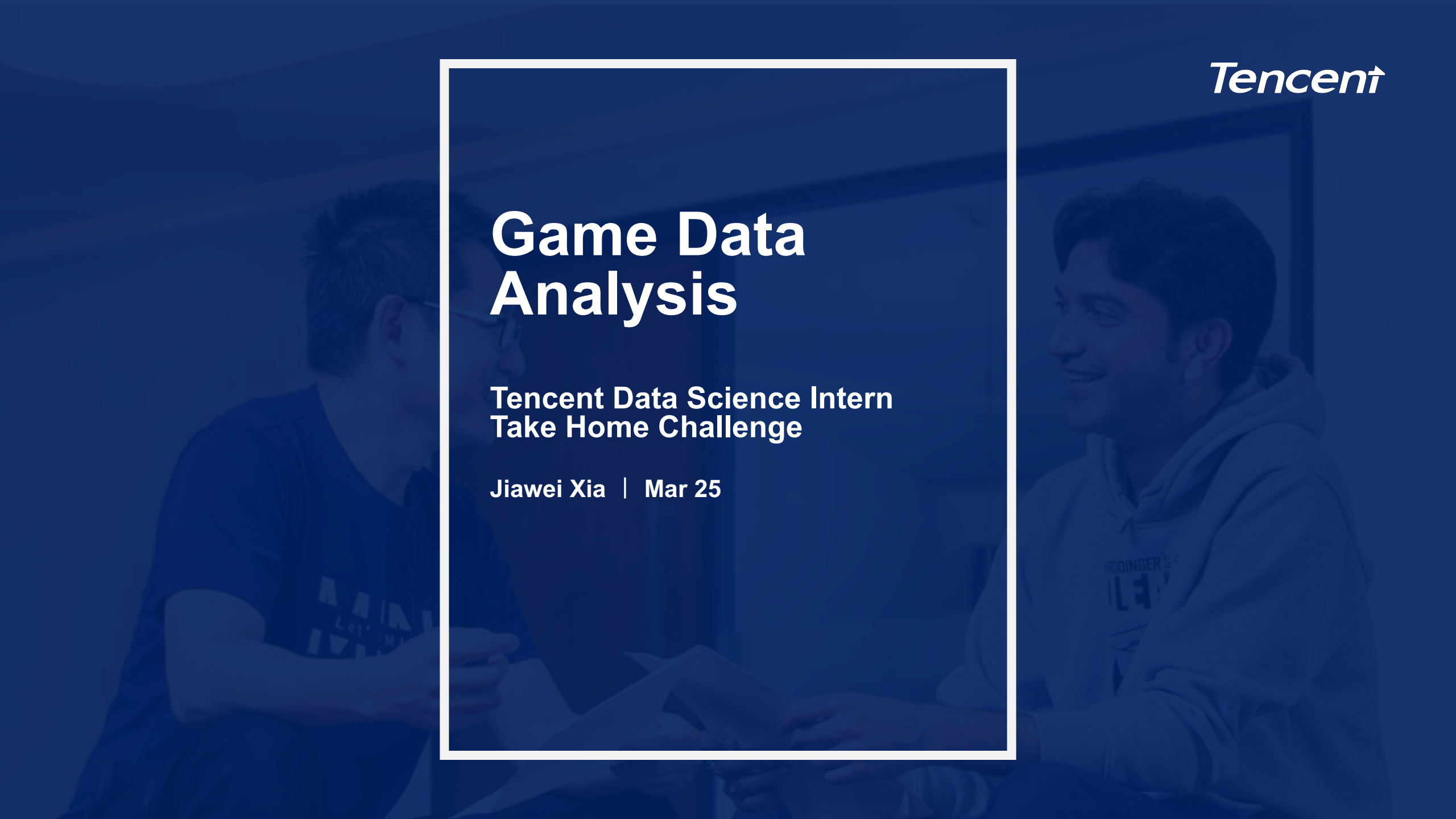


Game Data Analysis

Tencent Data Science Intern Take Home Challenge

Jiawei Xia | Mar 25

Content

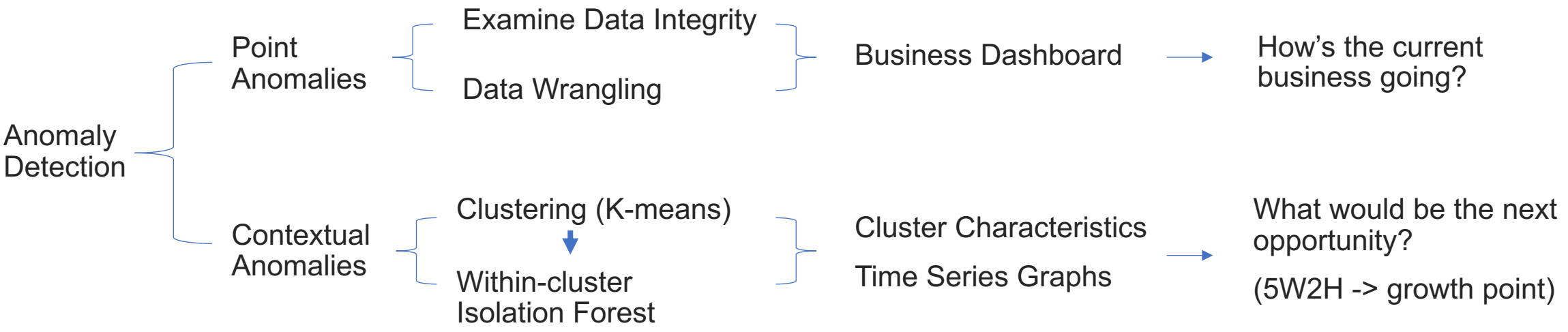
Tencent

01 Anomaly Detection

02 Data Science (Clustering/Prediction)

03 Visualization/ Time Series

04 Questions Answered



Point Anomalies

Examine Data Integrity ✔

Data Wrangling

Data Frame

- Head sample data
- Shape (748264 x 14)
- Data types
- Missing values (0)

Actions:

- Transform str object into datetime for “Date”

Columns

- Date range (from 2018-03-15 to 2020-03-09)
- Countries (239)
- Platforms (3)
- Key stats (min, max, etc)
- Duplicated columns (1)

Actions:

- Remove anomalies disobeying business rules
- Drop the second “Date”

Rows

- Duplicated rows (~50%)
- Conflicting rows (in country ‘Dominica’)
- Inconsistent country format

Actions:

- Drop repeated rows
- Drop conflicting rows
- Translate zh-cn to en

Pairwise

- Diagonal: concentrated data
- Line shape: possible linear relationship
- L shape & U shape: 2 possible clusters
- E shape: 3 possible clusters

Actions:

- Assumption for clustering k=2 or 3
- Linear correlation

Point Anomalies

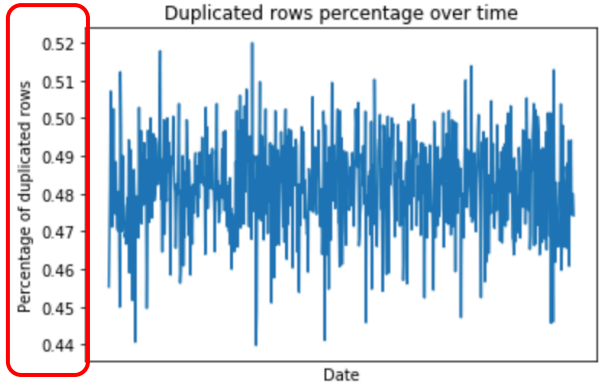
Examine Data Integrity

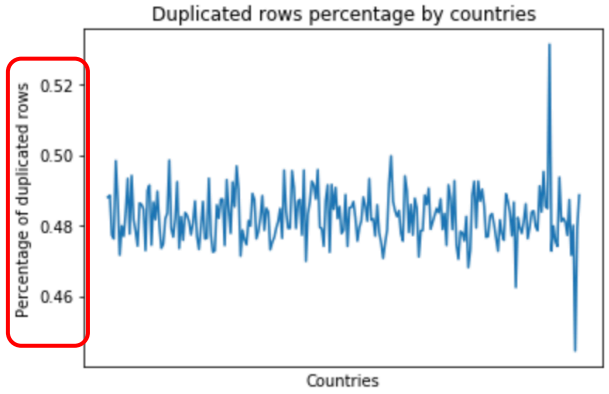
Data Wrangling

Date	object
Date.1	object
Platform	object
Country	object
TRU	float64
DAU	float64
Items	float64
Trans	float64
Items Per Trans	float64
Items per DAU	float64
Conversion	float64
Cash Flow	float64
Return Customer	float64
Time Spend Per Day(seconds)	float64

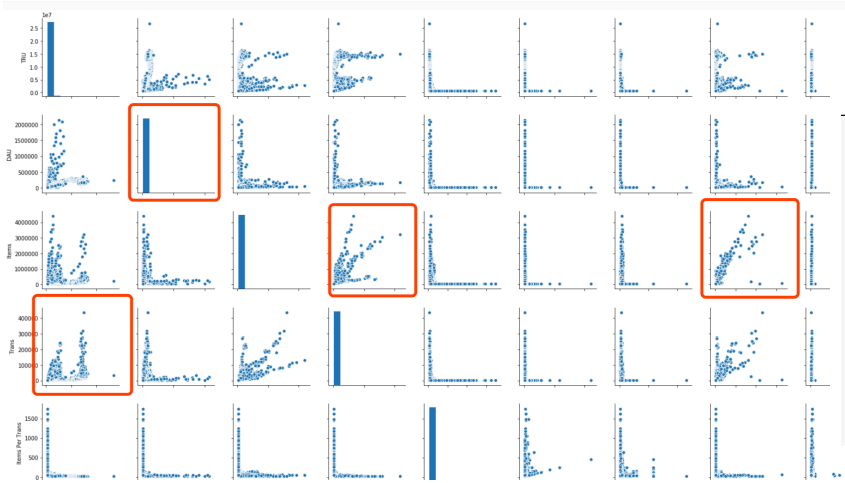
dtype: object

Iran(伊朗)	3441
Australia(澳大利亚)	3434
圣卢西亚	3366
Israel(以色列)	3365
Iraq(伊拉克)	3363
...	
厄立特里亚	529
托克劳	496
圣巴泰勒米	239
unknown(未知)	31
蒙塞拉特岛	18





	TRU	DAU	Items	Trans	Items Per Trans	Items per DAU	Conversion	Cash Flow	Return Customer	Time Spend Per Day(seconds)
count	7.482640e+05	7.482640e+05	7.482640e+05	748264.00000	748264.000000	748264.000000	748264.000000	748264.000000	748264.000000	748264.000000
mean	7.399568e+05	1.816949e+04	5.089120e+04	3958.49364	31.569433	1.414330	2.902036	5093.453512	39.795054	122.605716
std	6.776595e+05	1.801799e+04	4.610241e+04	3655.62504	23.629177	1.208879	2.591666	4712.790645	24.796009	43.331318
min	6.286478e+05	1.529217e+04	4.342911e+04	3428.09000	17.580000	1.300000	2.300000	4527.010000	12.210000	-9.990000
25%	6.288102e+05	1.529677e+04	4.342911e+04	3428.09000	17.580000	1.300000	2.300000	4527.010000	28.660000	98.950000
50%	6.314452e+05	1.537393e+04	4.352124e+04	3433.85000	28.090000	1.320000	2.560000	4532.770000	39.430000	118.180000
75%	6.574735e+05	1.617663e+04	4.466021e+04	3522.53000	36.080000	1.390000	3.000000	4611.080000	48.920000	139.370000
max	2.677675e+07	2.141935e+06	4.387412e+06	434846.58000	1744.870000	444.240000	232.630000	819825.280000	3912.090000	1390.060000



Country	Date	Platform	TRU	DAU	Items
	2018-03-16	4	4	4	4
	2018-03-19	5	5	5	5
	2018-03-20	5	5	5	5
	2018-03-22	5	5	5	5
	2018-03-23	5	5	5	5
Dominica	...	...	...	...	...
	2020-03-05	6	6	6	6

Point Anomalies

Examine Data Integrity ✔

Data Wrangling ✔

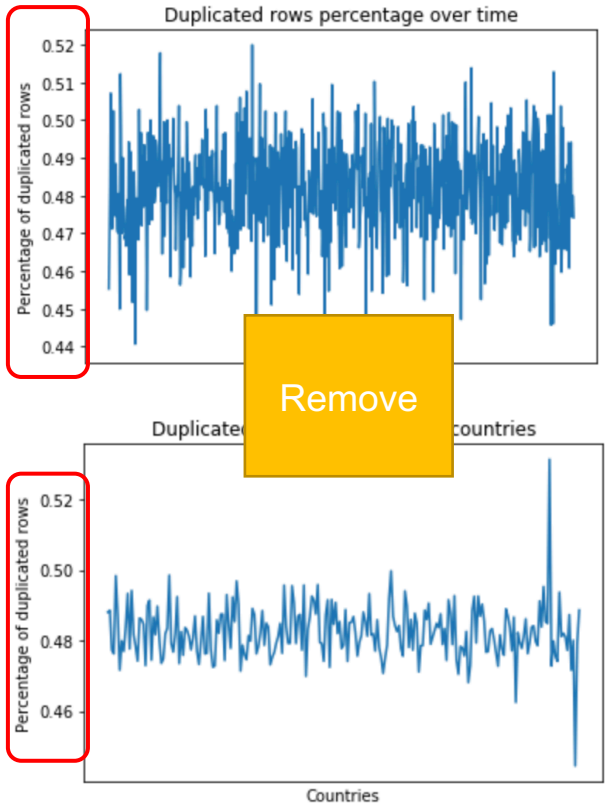
```
Date                          object
Date.1                        object
Platform                      object
Country                       object
TRU                          float64
DAU                          float64
Items                        float64
Trans                        float64
Items Per Trans              float64
Items per DAU                float64
Conversion                   float64
Cash Flow                    float64
Return Customer              float64
Time Spend Per Day(seconds)  float64
dtype: object
```

Drop

Transform

```
Iran(伊朗)                    3441
Australia(澳大利亚)           3434
圣卢西亚                      3366
Israel(以色列
Iraq(伊拉克)
                              ...
厄立特里亚                     529
托克劳                         496
圣巴泰勒米                     239
unknown(未知)                   31
蒙塞拉特岛                      18
```

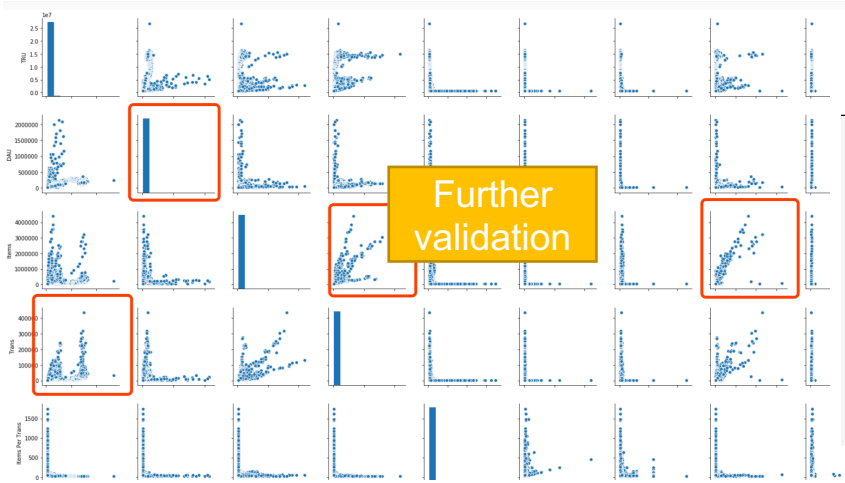
Translate



	TRU	DAU	Items	Trans	Items Per Trans	Items per DAU	Conversion	Cash Flow	Return Customer	Time Spend Per Day(seconds)
count	7.482640e+05	7.482640e+05	7.482640e+05	748264.00000	748264.000000	748264.000000	748264.000000	748264.000000	748264.000000	748264.000000
mean	7.399568e+05	1.816949e+04	5.089120e+04	3958.49364	31.569433	1.414330	2.902036	5093.453512	39.795054	122.605716
std	6.776595e+05	1.801799e+04	4.610241e+04	3655.62504	23.629177	1.208879	2.591666	4712.790645	24.796009	43.331318
min	6.286478e+05	1.529217e+04	4.342911e+04	3428.09000	17.580000	1.300000	2.300000	4537.010000	12.310000	-9.990000
25%	6.288102e+05	1.529677e+04	4.342911e+04	3428.09000	17.580000	1.300000	2.300000	4537.010000	12.310000	98.950000
50%	6.314452e+05	1.537393e+04	4.352124e+04	3433.85000	28.090000	1.320000	2.560000	4532.770000	39.430000	118.180000
75%	6.574735e+05	1		3522.53000	36.080000	1.390000	3.000000	4611.080000	48.920000	139.370000
max	2.677675e+07	2		34846.58000	1744.870000	444.240000	232.630000	819825.280000	3912.090000	1390.060000

Remove

Further validation



Further validation

Country	Date	Platform	TRU	DAU	Items
	2018-03-16	4	4	4	4
	2018-03-19	5	5	5	5
	2018-03-2			5	5
	2018-0 22			5	5
	2018-03-23	5	5	5	5
Dominica	...	...	...	...	...
	2020-03-05	6	6	6	6

Drop

Business Dashboard

Tencent

Business Dashboard

Tencent

Business Dashboard

Tencent

Contextual Anomalies

K-Means Clustering ✔

Within-cluster Isolation Forest

Data Wrangling

- **Datetime**: dayofweek, dayofmonth, month, year
- **Categorical**: “Country” into dummies and drop 1 column
- **Categorical**: “Platform” into dummies and drop 1 column

Result:

- 252 columns in total

Find the Best K

- **Assumption** from previous steps
- **Elbow method**: compute within-Cluster-Sum of Squared Errors (WSS)

Result:

- Selected k=3 validated from both methods

Fit and Evaluate Model

- Kmeans(n_clusters=3, random_state=0).fit(X)
- Evaluation: compute silhouette_score

Result:

- Sihouette score > 0.91, excellent performance

Examine Clusters

- Attached labels to each records
- Subset clusters for later analysis

Result:

- # Cluster 1: 377101
- # Cluster 2: 774
- # Cluster 3: 5390

Contextual Anomalies

K-Means Clustering ✔

Within-cluster Isolation Forest

Data Wrangling

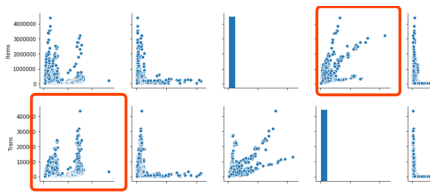
```
game_data.shape
```

```
(383265, 13)
```

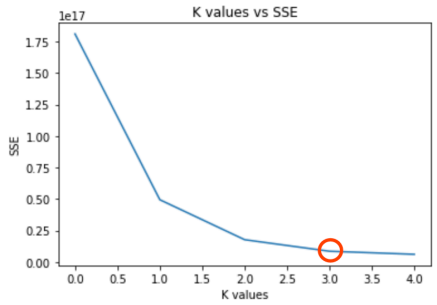
```
X.shape
```

```
(383265, 252)
```

Find the Best K



k=2/3



k=3

Fit and Evaluate Model

```
silhouette_score(X,
```

```
0.9194291375101088
```

Examine Clusters

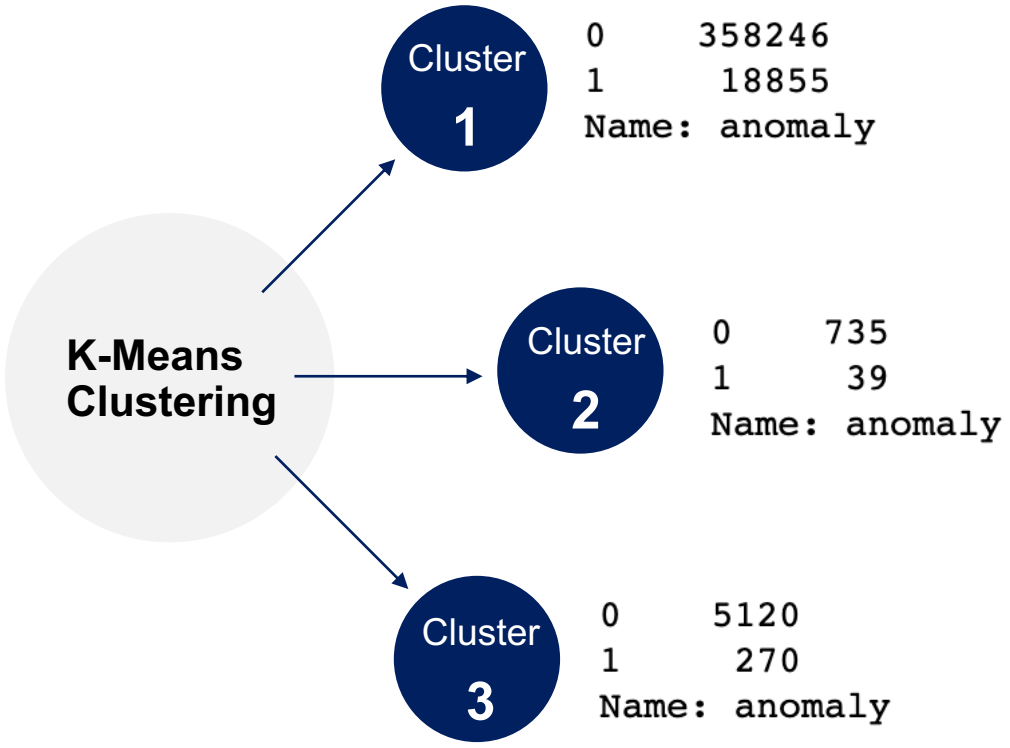
```
0    377101
2      5390
1       774
Name: label, dtype: int64
```

```
cluster1
cluster2
cluster3
```

Contextual Anomalies

K-Means Clustering ✔

Within-cluster Isolation Forest ✔



Within-cluster Isolation Forest

```
# input: 3 parameters
def isolation_forest(df, columns, outliers_fraction):
    data = df[columns] # extract data
    scaler = StandardScaler()
    np_scaled = scaler.fit_transform(data)
    data = pd.DataFrame(np_scaled) # scaled data

    model =  IsolationForest(contamination=outliers_fraction)
    model.fit(data) # train isolation forest

    # attach anomaly label
    # 0: normal, 1: anomaly
    df['anomaly'] = np.array(pd.Series(model.predict(data)))
    df['anomaly'] = df['anomaly'].map( {1: 0, -1: 1} )

    # return dataframe with anomaly label
    return df
```

Cluster Characteriestics

Tencent

Cluster Time Series Analysis

Tencent

Anamoly Time Series Analysis

Tencent

Conclusion

Tencent

1st Issues

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.

2nd Issues

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.

3rd Issues

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.


Tencent
THANK YOU

Content

Tencent

01

Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

02

Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

03

Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

04

Lorem ipsum dolor sit amet, consectetuer adipiscing elit.

Project status report

Tencent

Status 1

- Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.
- Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Status 2

- Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.
- Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

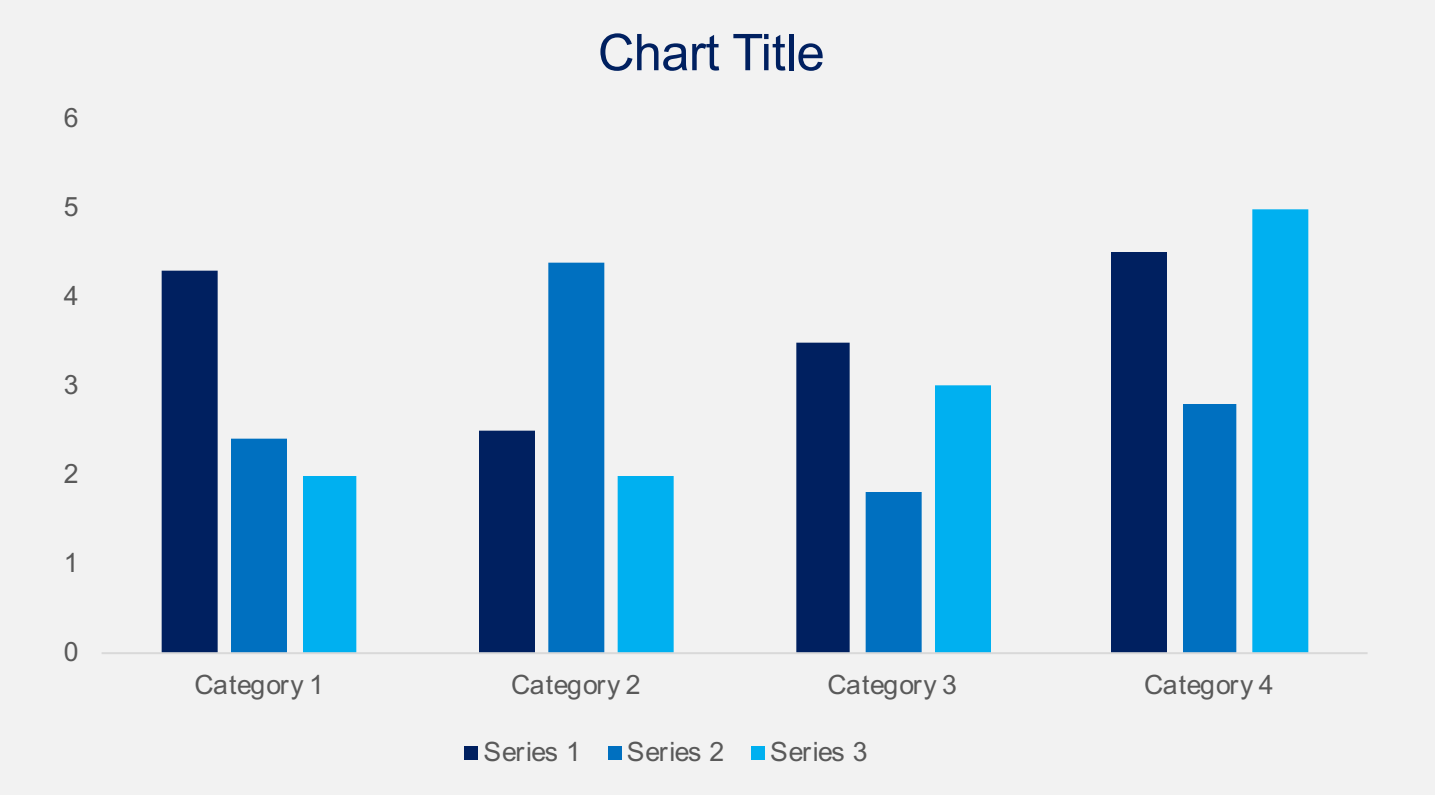
Status 3

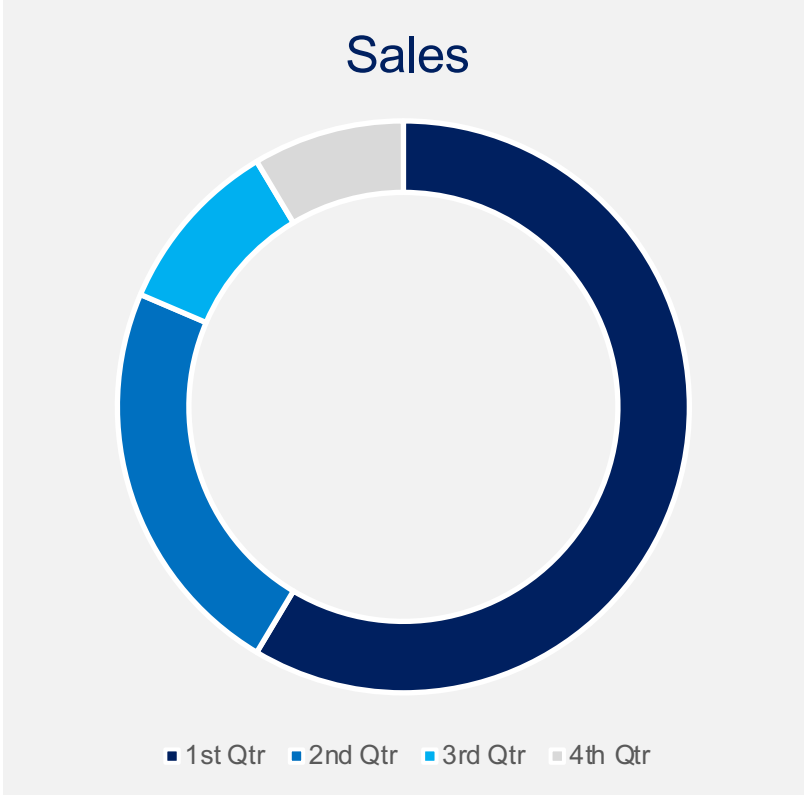
- Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.
- Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Status 4

- Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.
- Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Data report





Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa.

Risks

Risks 1

- Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.
- Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Risks 2

- Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.
- Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Risks 3

- Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.
- Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.
- Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Proin pharetra nonummy pede. Mauris et orci.

Summary

Project	Progress	Status	Team	Comments
Project 1	Lorem ipsum	Lorem ipsum	Lorem ipsum	Lorem ipsum dolor sit amet.
Project 2	Lorem ipsum	Lorem ipsum	Lorem ipsum	Consectetuer adipiscing elit.
Project 3	Lorem ipsum	Lorem ipsum	Lorem ipsum	Lorem ipsum dolor sit amet.
Project 4	Lorem ipsum	Lorem ipsum	Lorem ipsum	Lorem ipsum dolor sit amet.
Project 5	Lorem ipsum	Lorem ipsum	Lorem ipsum	Lorem ipsum dolor sit amet.
Project 6	Lorem ipsum	Lorem ipsum	Lorem ipsum	Lorem ipsum dolor sit amet.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.

Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Solutions offer

Tencent

Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies.

Solution #1

Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies.

Solution #2

Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies.

Solution #3

Solution #4

Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies.

Solution #5

Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies.

Solution #6

Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies.