Exploratory Data Analysis for Revenue Predicting Service - *by Jan Korinek*

Deliverable goals

> (1) Assimilate the business scenario and articulate testable hypotheses.

Answer:

The main goal is to create service capable to predict company revenue for following month. Projection has to be able to estimate separated revenues for predefined countries. Predictive performance has to achieve sufficient accuracy to have positive impacet on manager decision making.

Revenue estimates will be based on charging for combination of services to which is each customer subscribed.

From business perspective is expected to increase company revenue by well projected budged and staffing allocation. This is dependent on executive decisions based on more accurate predictions. Therefore can be defined business metric as function of revenue generated by more accurate predictions.

Null Hypothesis:

Revenue of the company is not affected by increase of the prediction accuracy.

In order to reject a null hypothesis, it may be proceed to testing.

> (2) State the ideal data to address the business opportunity and clarify the rationale for needing specific data.

Answer:

Based on defined business scenario, obtained customer data should be ideally at transaction level and has to be time-dependent in order to create supervised prediction model. Data should cover customer payment history, country of subscription, profile information and names of subscribed services.

> (3.) Create a python script to extract relevant data from multiple data sources, automating the process of data ingestion.

In [1]:

```
# Load and extract data from raw json into dictionary of dataframes for top 10 countries by revenue
%run cslib.py

# Show selected df
print(ts_all['all'])

# Update df about 'year' column
for key, df in ts_all.items():
    df['year'] = df.year_month.str[:4]
```

```
...fetching data
... loading ts data from files
load time: 0:00:00
all (607, 7)
eire (607, 7)
france (607, 7)
germany (607, 7)
hong_kong (426, 7)
netherlands (607, 7)
norway (577, 7)
portugal (607, 7)
singapore (456, 7)
spain (607, 7)
united_kingdom (607, 7)

           date  purchases  unique_invoices  unique_streams  total_views  \
0    2017-11-01          0                0               0            0
1    2017-11-02          0                0               0            0
2    2017-11-03          0                0               0            0
3    2017-11-04          0                0               0            0
4    2017-11-05          0                0               0            0
..          ...        ...              ...             ...          ...
602  2019-06-26       1358               67             999         6420
603  2019-06-27       1620               80             944         9435
604  2019-06-28       1027               70             607         5539
605  2019-06-29          0                0               0            0
606  2019-06-30        602               27             423         2534

    year_month  revenue
0      2017-11     0.00
1      2017-11     0.00
2      2017-11     0.00
3      2017-11     0.00
4      2017-11     0.00
..         ...      ...
602    2019-06  4903.17
603    2019-06  5499.38
604    2019-06  3570.60
605    2019-06     0.00
606    2019-06  1793.98

[607 rows x 7 columns]
```

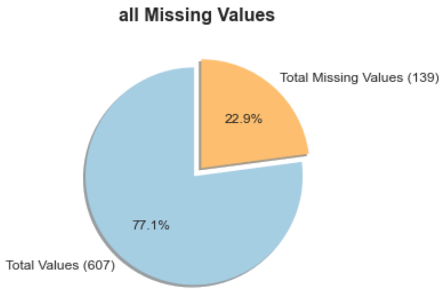
> (4.) Investigate the relationship between the relevant data, the target and the business metric.

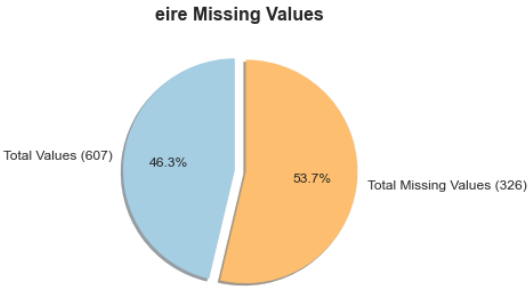
Missing Values Summary and Visualization

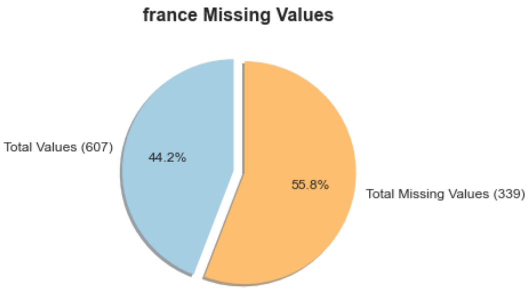
```
all Missing Value Summary:
----------------------
date                0
purchases         139
unique_invoices   139
unique_streams    139
total_views       139
year_month          0
```

```
revenue            139
year                 0
dtype: int64

eire Missing Value Summary:
---------------------
date                 0
purchases          326
unique_invoices    326
unique_streams     326
total_views        328
year_month           0
revenue            326
year                 0
dtype: int64

france Missing Value Summary:
---------------------
date                 0
purchases          339
unique_invoices    339
unique_streams     339
total_views        341
year_month           0
revenue            339
year                 0
dtype: int64

germany Missing Value Summary:
---------------------
date                 0
purchases          269
unique_invoices    269
unique_streams     269
total_views        269
year_month           0
revenue            269
year                 0
dtype: int64

hong_kong Missing Value Summary:
---------------------
date                 0
purchases          418
unique_invoices    418
unique_streams     418
total_views        418
year_month           0
revenue            418
year                 0
dtype: int64

netherlands Missing Value Summary:
---------------------
date                 0
purchases          475
unique_invoices    475
unique_streams     475
total_views        475
year_month           0
revenue            475
year                 0
dtype: int64

norway Missing Value Summary:
---------------------
date                 0
purchases          559
unique_invoices    559
unique_streams     559
total_views        559
year_month           0
revenue            559
year                 0
dtype: int64

portugal Missing Value Summary:
---------------------
date                 0
purchases          538
unique_invoices    538
unique_streams     538
total_views        539
year_month           0
revenue            538
year                 0
dtype: int64

singapore Missing Value Summary:
---------------------
date                 0
purchases          451
unique_invoices    451
unique_streams     451
total_views        451
year_month           0
revenue            451
year                 0
dtype: int64

spain Missing Value Summary:
---------------------
date                 0
purchases          510
unique_invoices    510
unique_streams     510
total_views        510
year_month           0
revenue            510
year                 0
dtype: int64

united_kingdom Missing Value Summary:
---------------------
date                 0
purchases          139
unique_invoices    139
unique_streams     139
total_views        139
year_month           0
revenue            139
year                 0
```

dtype: int64


all Missing Values
Total Missing Values (139)
22.9%
77.1%
Total Values (607)

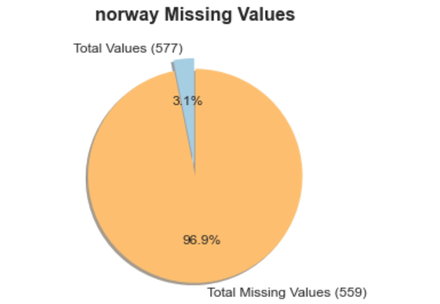

eire Missing Values
Total Values (607)
46.3%
53.7%
Total Missing Values (326)

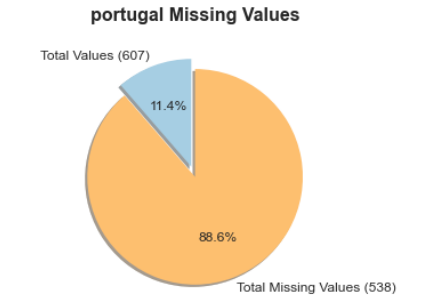

france Missing Values
Total Values (607)
44.2%
55.8%
Total Missing Values (339)

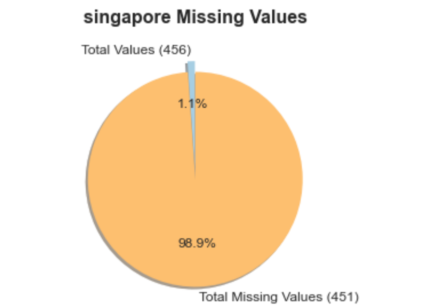

germany Missing Values
Total Missing Values (269)
44.3%
55.7%
Total Values (607)

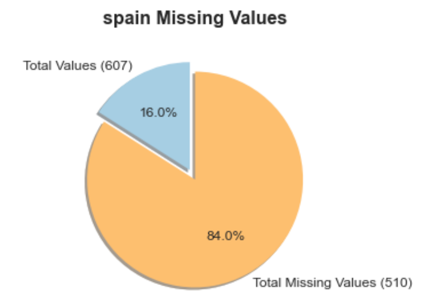

hong_kong Missing Values
Total Values (426)
1.9%
98.1%
Total Missing Values (418)

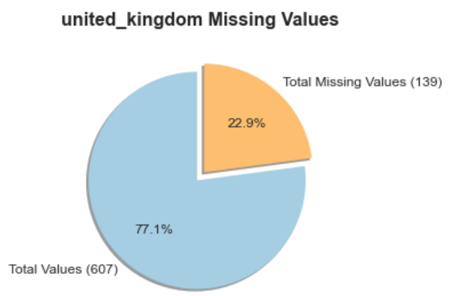

netherlands Missing Values
Total Values (607)
21.7%
78.3%
Total Missing Values (475)





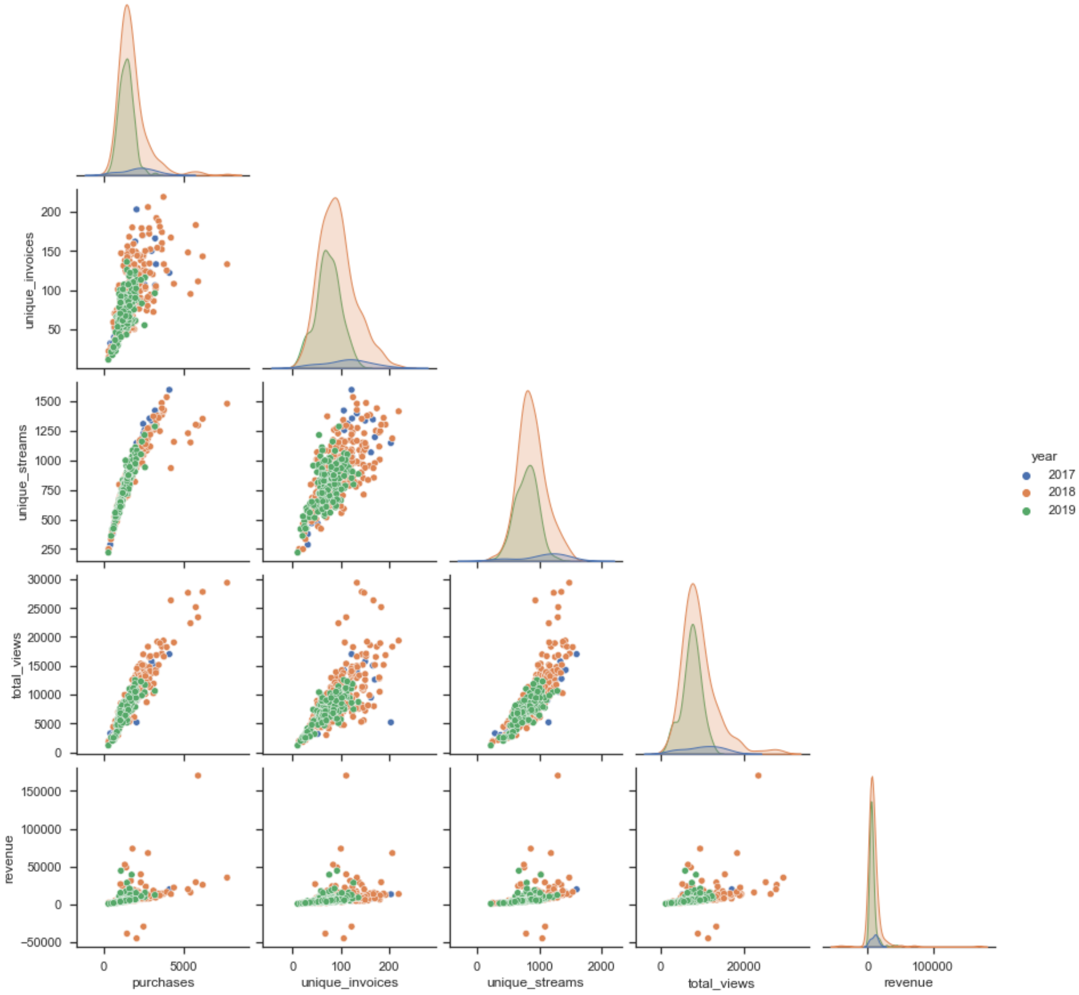






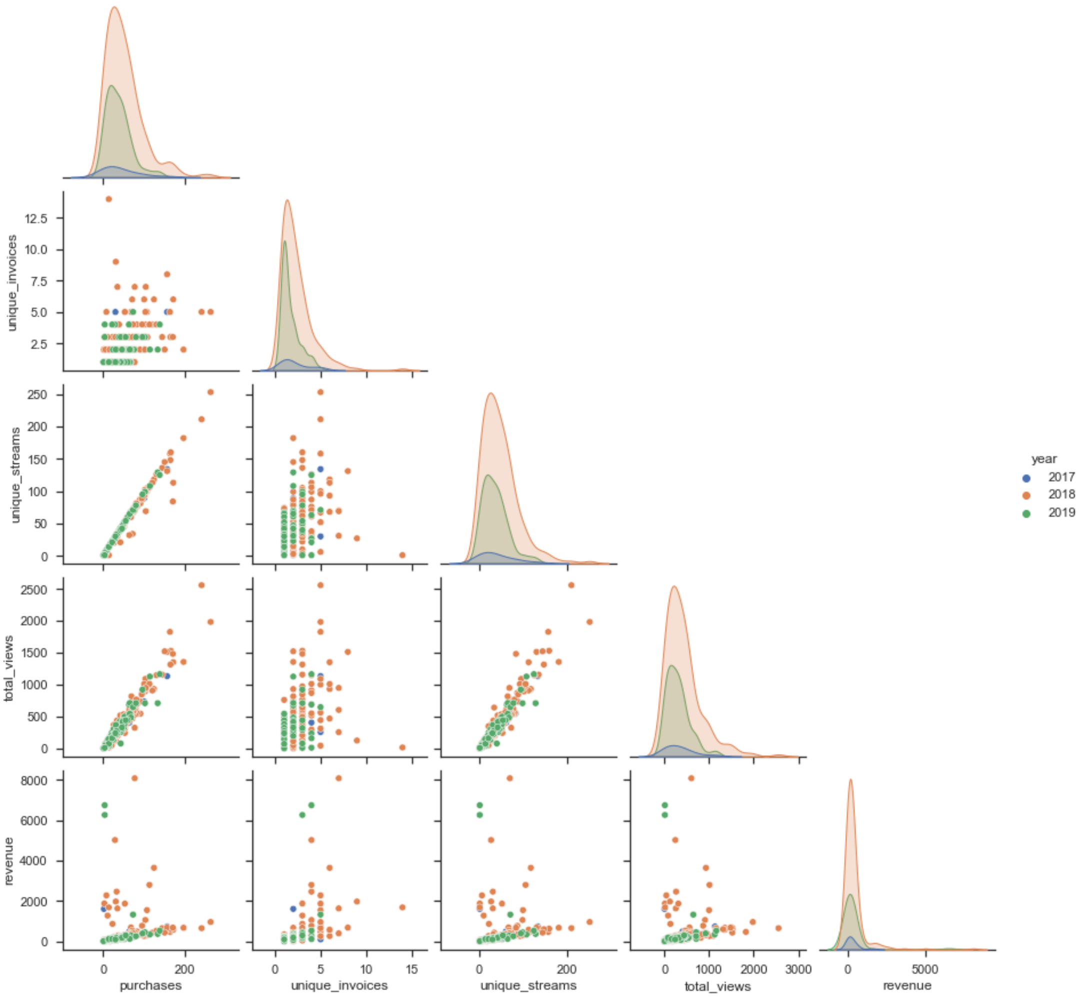
Pair Plot Visualization

all



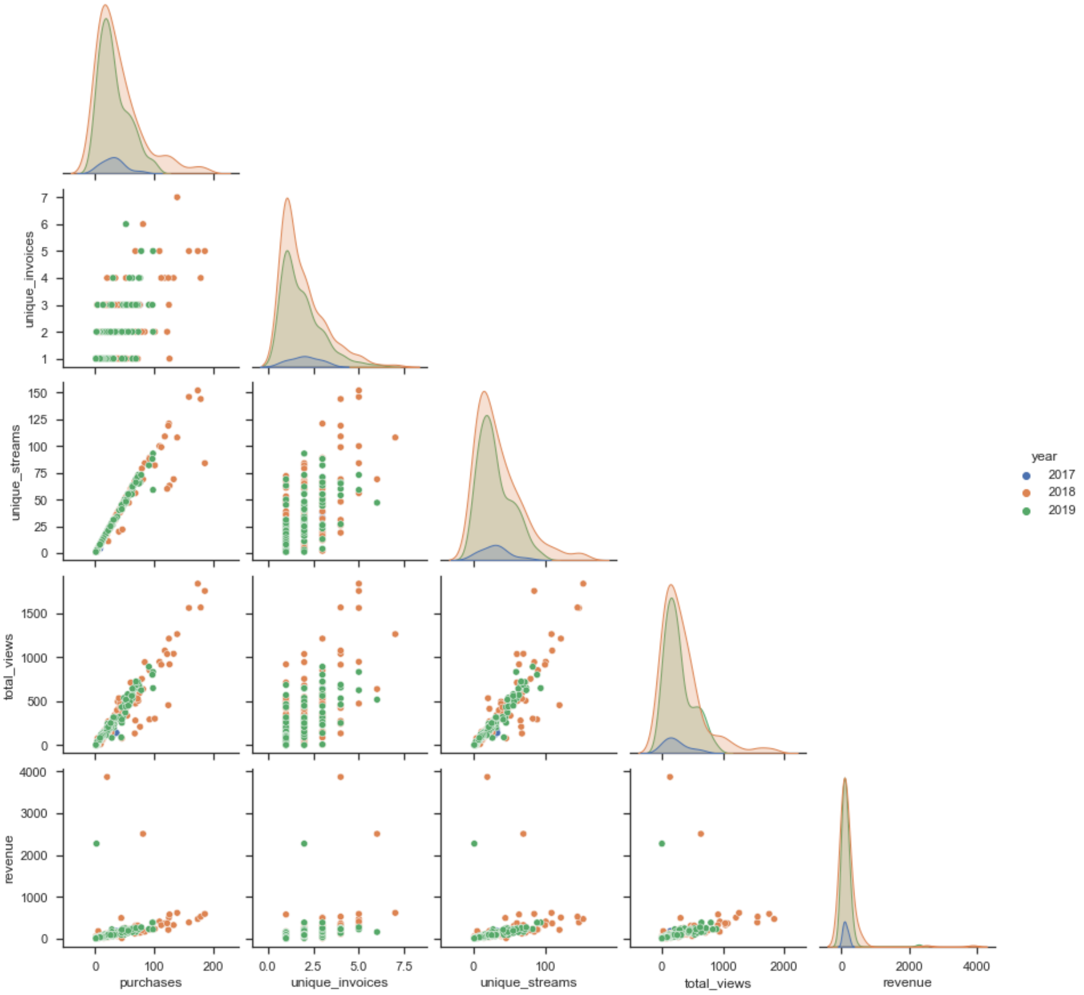
unique_invoices
unique_streams
total_views
revenue
purchases
unique_invoices
unique_streams
total_views
revenue
year
2017
2018
2019

eire

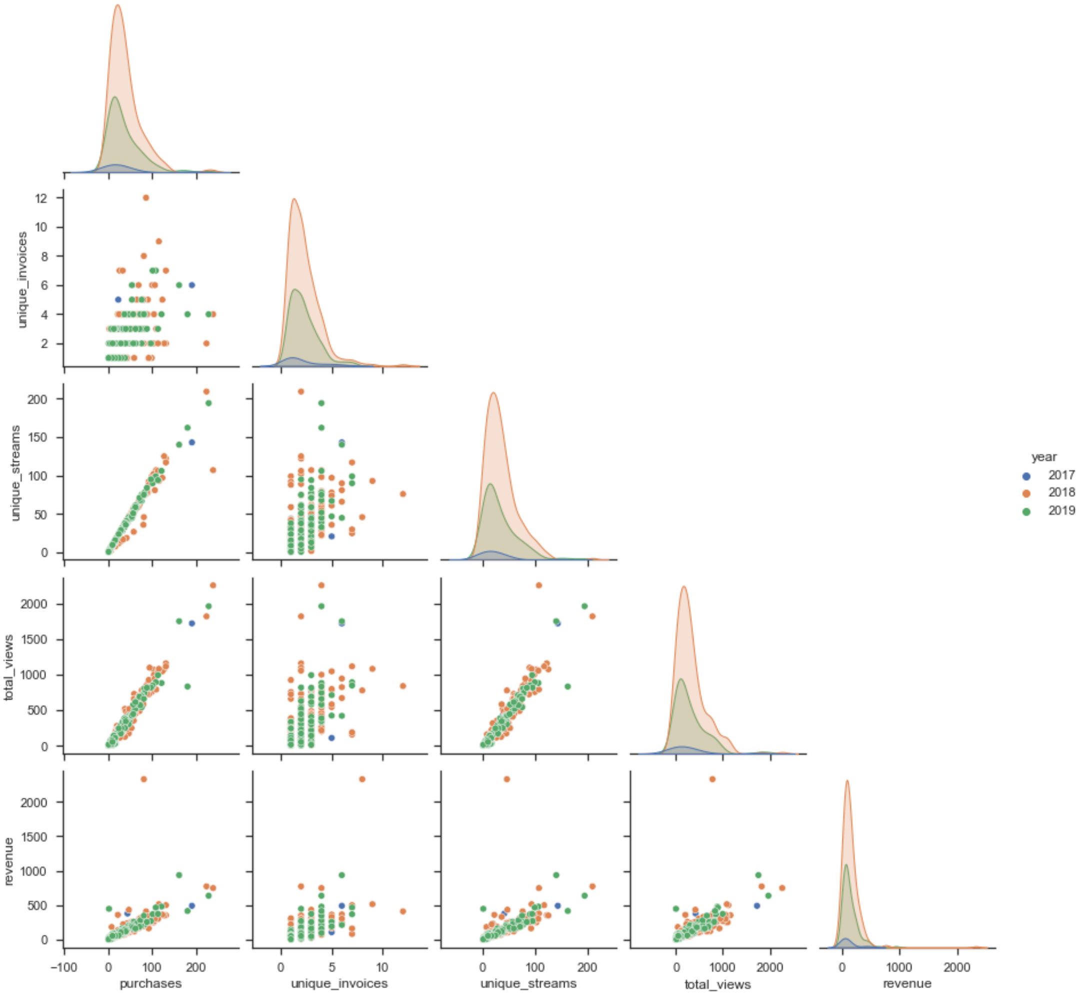


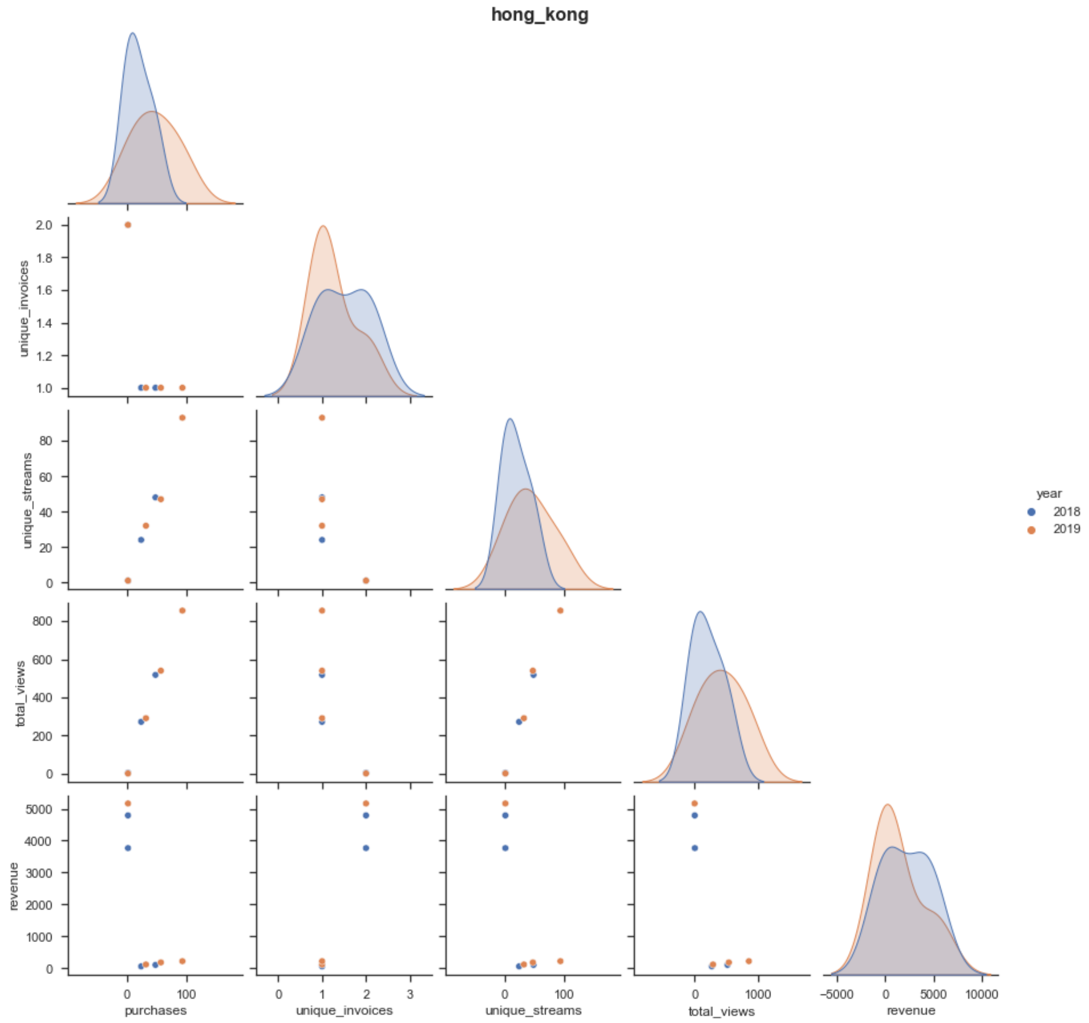
unique_invoices
unique_streams
total_views
revenue
purchases
unique_invoices
unique_streams
total_views
revenue
year
2017
2018
2019

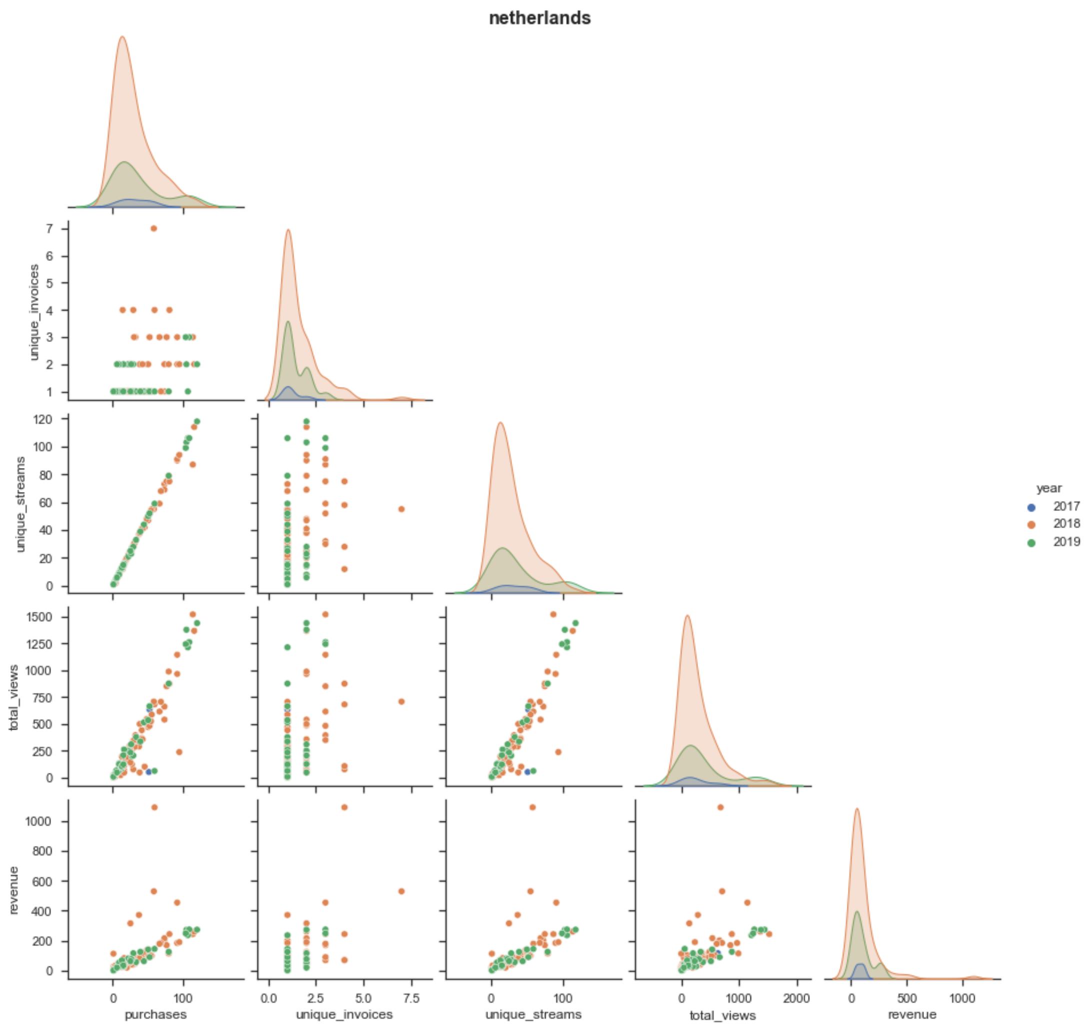
france

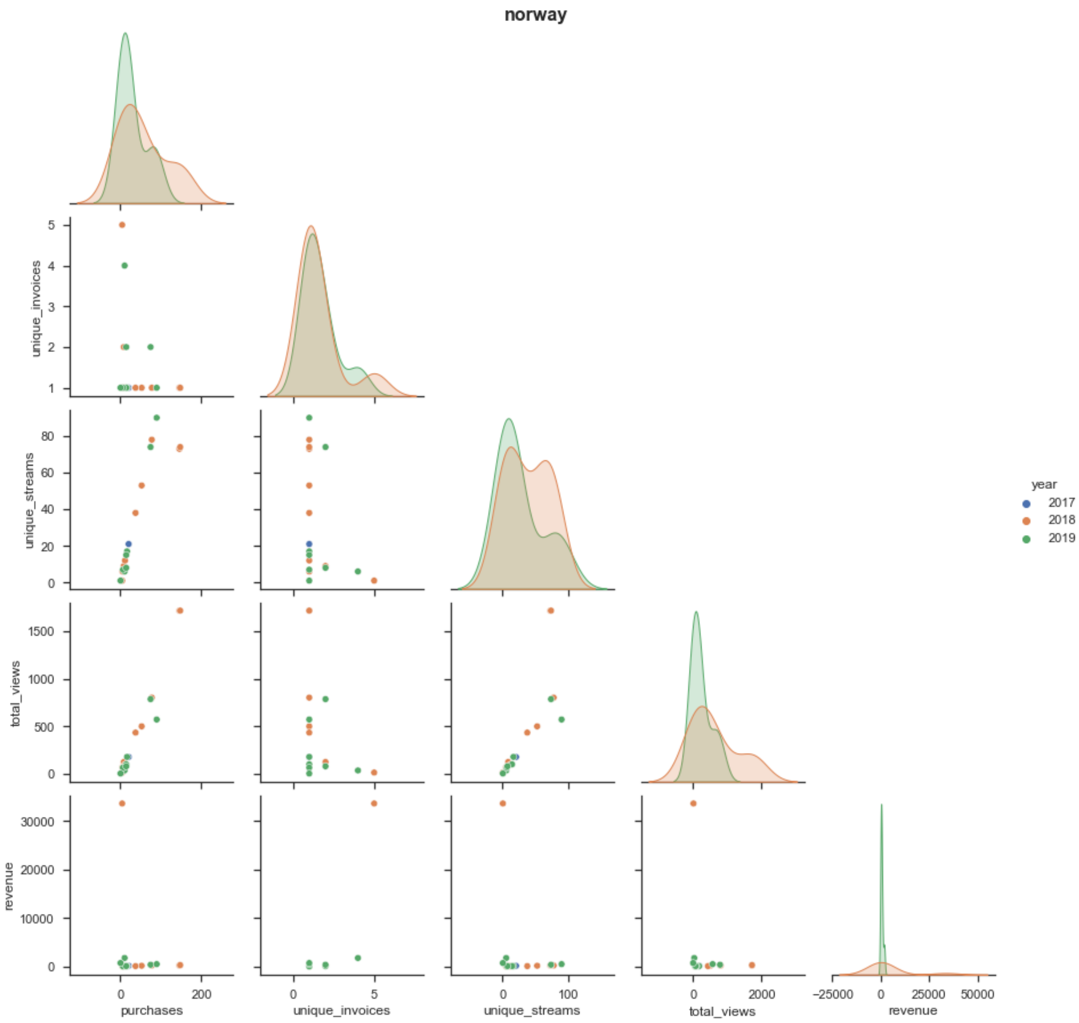


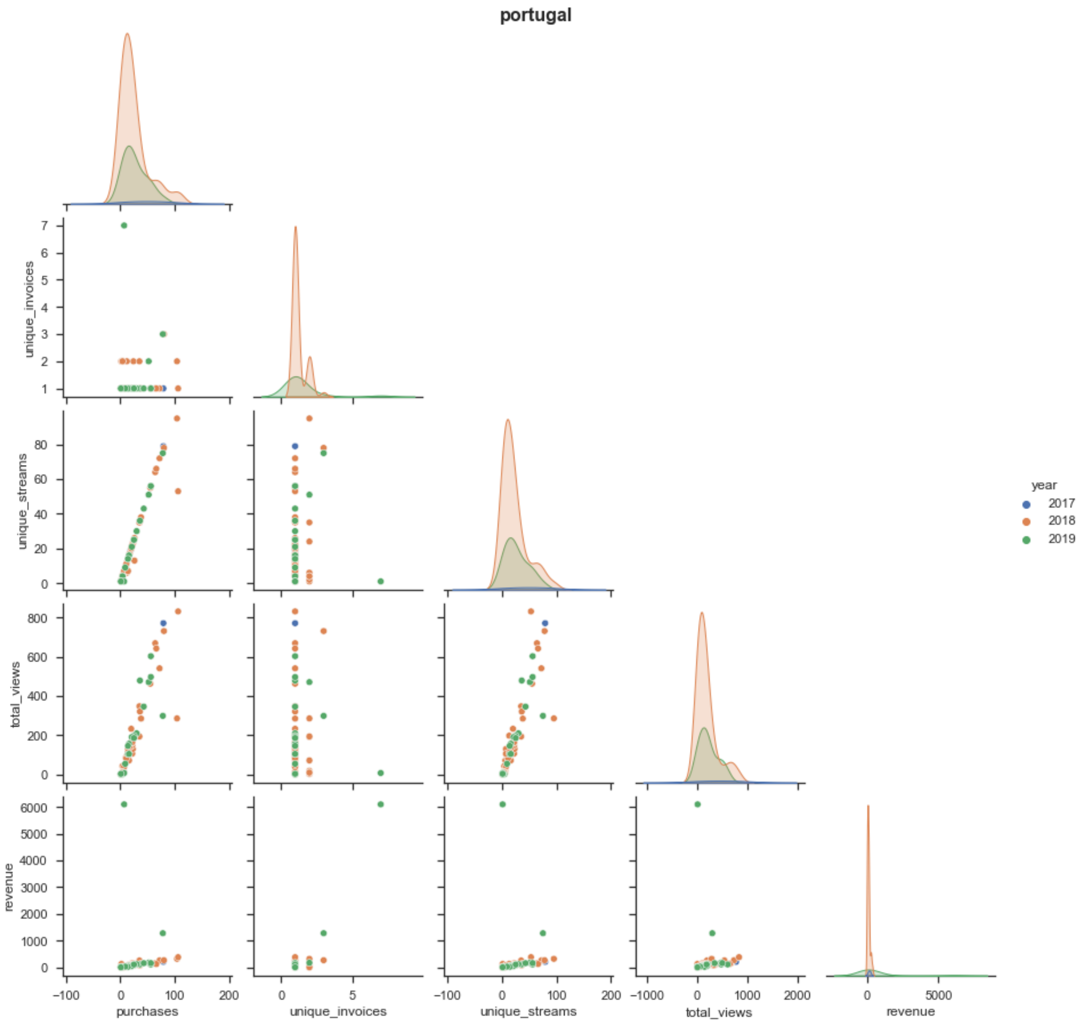
germany



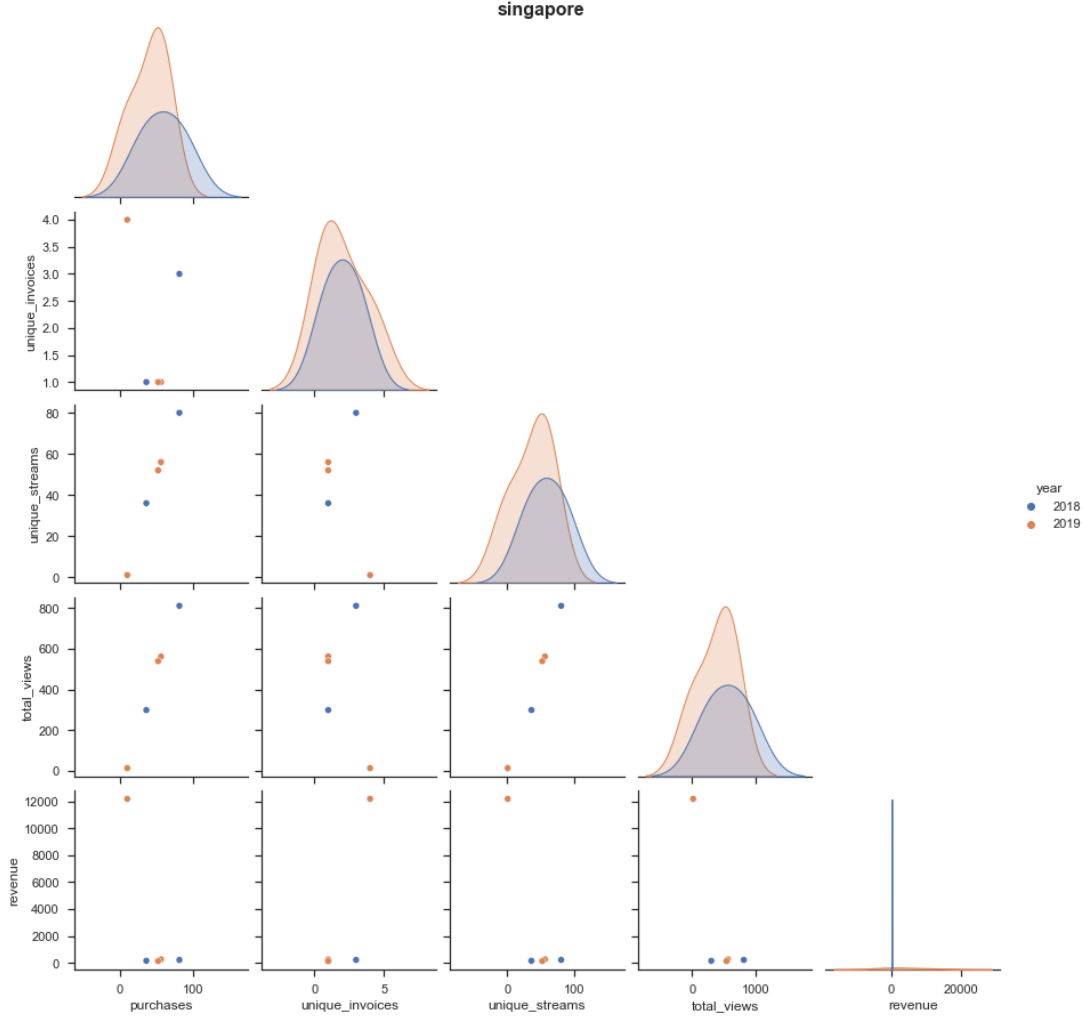

hong_kong
unique_invoices
unique_streams
total_views
revenue
purchases
unique_invoices
unique_streams
total_views
revenue
year
2018
2019


netherlands
unique_invoices
unique_streams
total_views
revenue
purchases
unique_invoices
unique_streams
total_views
revenue
year
2017
2018
2019

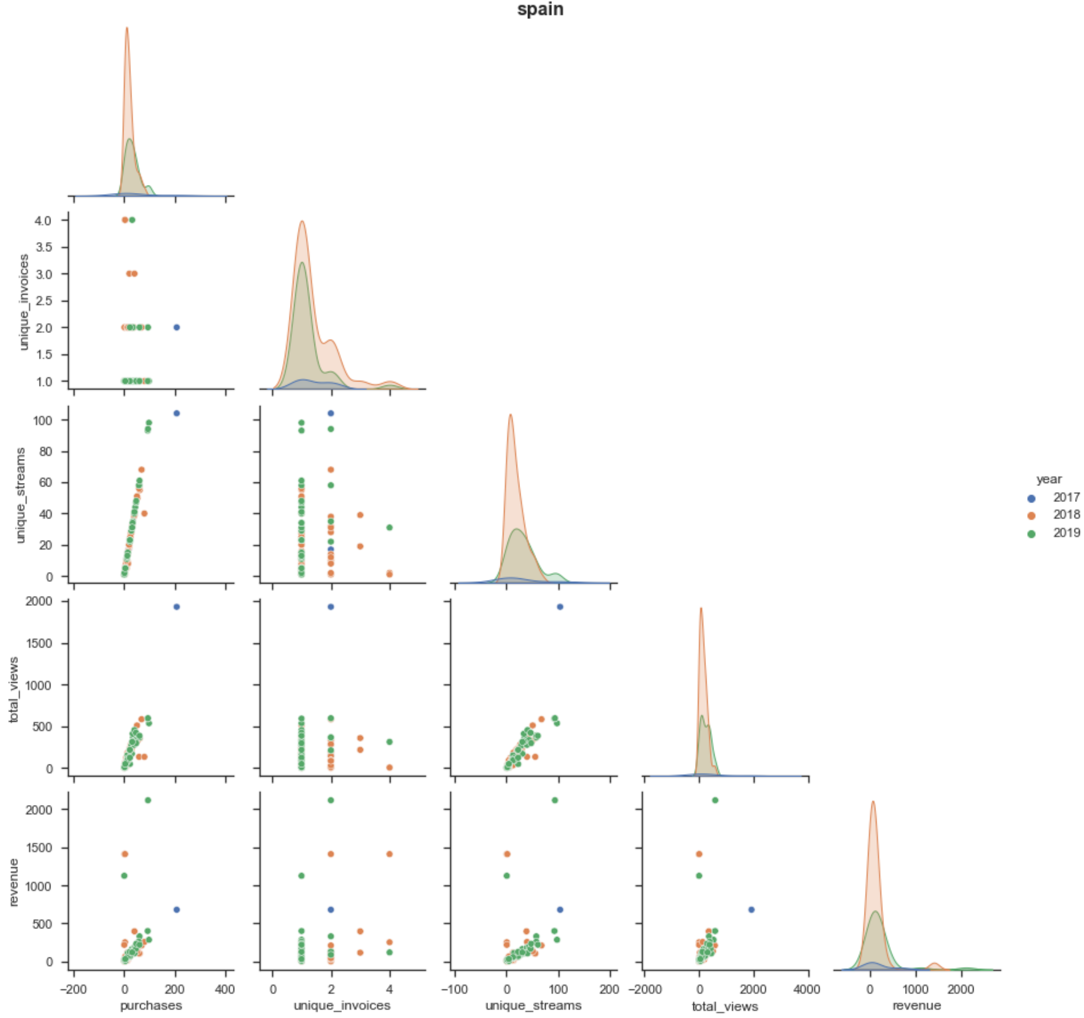

norway
unique_invoices
unique_streams
total_views
revenue
purchases
unique_invoices
unique_streams
total_views
revenue
year
2017
2018
2019

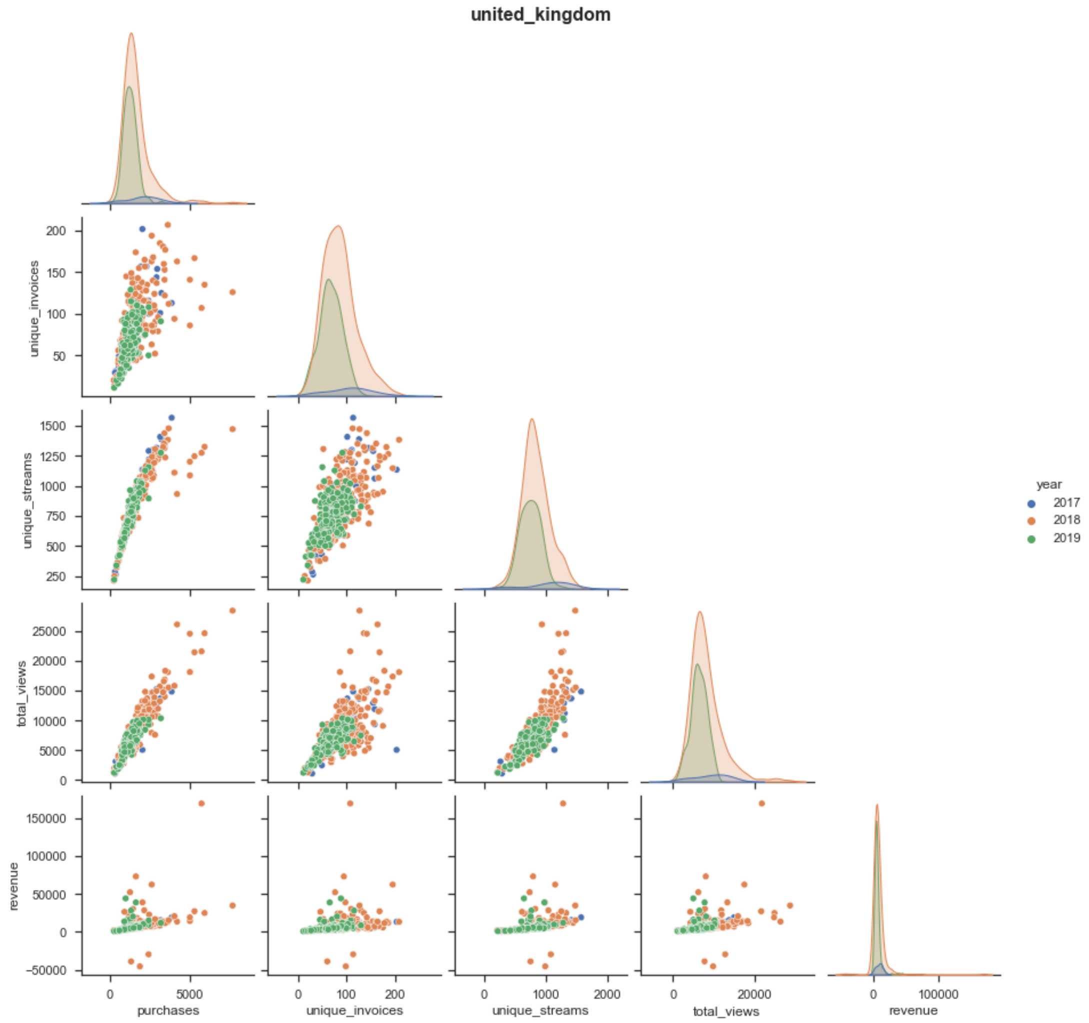

portugal
unique_invoices
unique_streams
total_views
revenue
purchases
unique_invoices
unique_streams
total_views
revenue
year
2017
2018
2019

singapore



spain

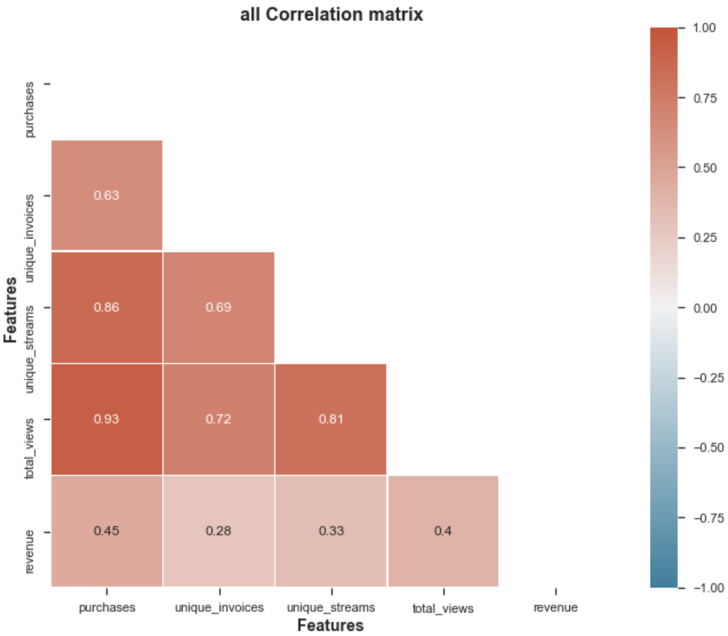


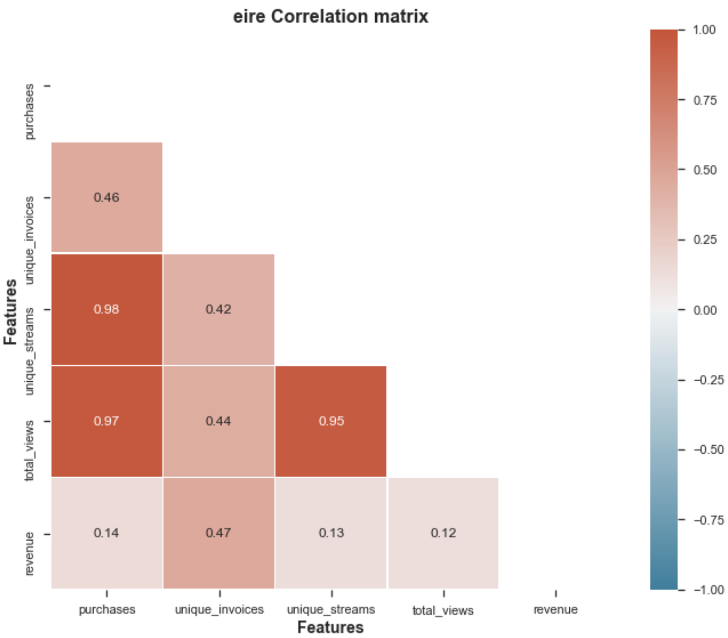


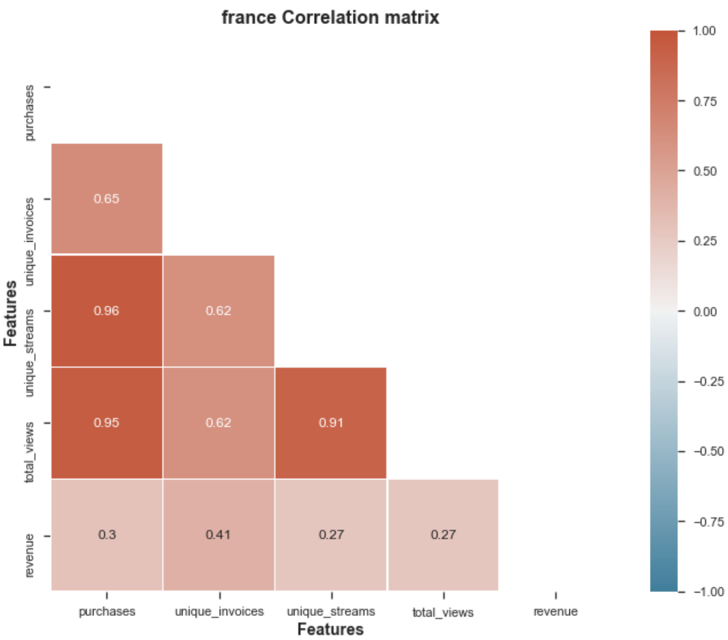
Correlation Matrix Visualization

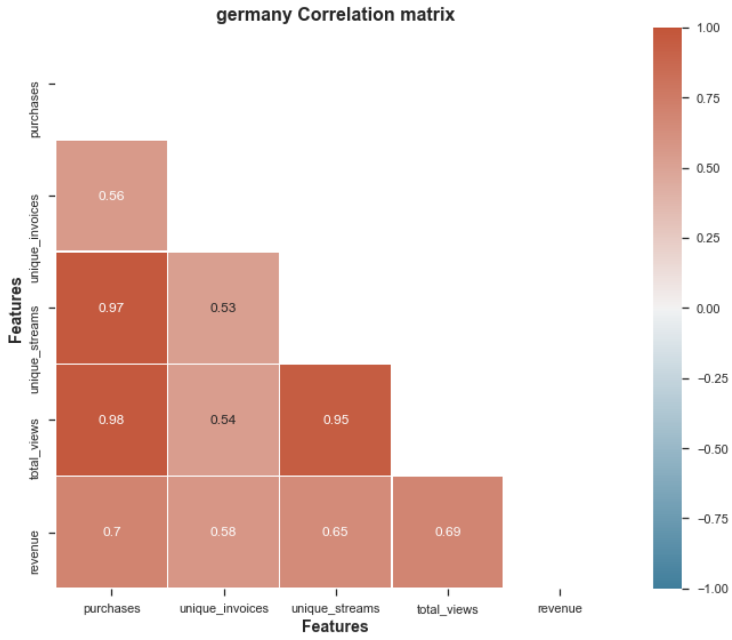
```
----- all Strong Positive Pairs -----
revenue         unique_streams     0.331129
total_views     revenue            0.399065
revenue         purchases          0.454676
purchases       unique_invoices    0.634036
unique_invoices unique_streams     0.691142
                total_views        0.718337
total_views     unique_streams     0.814562
unique_streams  purchases          0.863381
purchases       total_views        0.931435
                purchases          1.000000
dtype: float64

----- all Strong Negative Pairs -----
Series([], dtype: float64)


----- eire Strong Positive Pairs -----
unique_streams  unique_invoices    0.418989
unique_invoices total_views        0.441921
                purchases          0.455005
                revenue            0.466120
unique_streams  total_views        0.954755
purchases       total_views        0.970074
unique_streams  purchases          0.984630
purchases       purchases          1.000000
dtype: float64

----- eire Strong Negative Pairs -----
Series([], dtype: float64)


----- france Strong Positive Pairs -----
revenue         unique_invoices    0.413157
unique_invoices total_views        0.616620
                unique_streams     0.624344
                purchases          0.647822
unique_streams  total_views        0.908442
purchases       total_views        0.946012
unique_streams  purchases          0.961219
purchases       purchases          1.000000
dtype: float64

----- france Strong Negative Pairs -----
Series([], dtype: float64)


----- germany Strong Positive Pairs -----
unique_invoices unique_streams     0.525786
                total_views        0.537952
purchases       unique_invoices    0.555734
revenue         unique_invoices    0.582065
                unique_streams     0.652187
total_views     revenue            0.693439
revenue         purchases          0.702357
total_views     unique_streams     0.947758
unique_streams  purchases          0.974790
purchases       total_views        0.975122
```

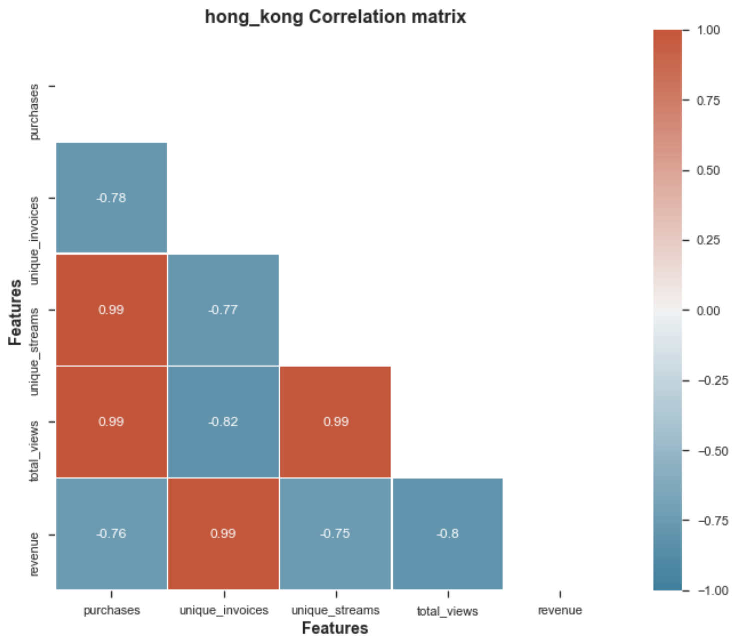
```
                 purchases          1.000000
dtype: float64

----- germany Strong Negative Pairs -----
Series([], dtype: float64)


----- hong_kong Strong Positive Pairs -----
unique_invoices  revenue        0.985877
unique_streams   total_views    0.990302
purchases        total_views    0.994452
unique_streams   purchases      0.994474
purchases        purchases      1.000000
dtype: float64

----- hong_kong Strong Negative Pairs -----
unique_invoices  total_views       -0.818670
revenue          total_views       -0.797153
purchases        unique_invoices   -0.778688
unique_streams   unique_invoices   -0.774664
revenue          purchases         -0.755547
unique_streams   revenue           -0.752270
dtype: float64


----- netherlands Strong Positive Pairs -----
unique_invoices  unique_streams     0.402331
                 total_views        0.421418
purchases        unique_invoices    0.422446
revenue          unique_invoices    0.571453
                 total_views        0.587631
unique_streams   revenue            0.611357
revenue          purchases          0.613001
unique_streams   total_views        0.924023
purchases        total_views        0.932564
unique_streams   purchases          0.996788
purchases        purchases          1.000000
dtype: float64

----- netherlands Strong Negative Pairs -----
Series([], dtype: float64)


----- norway Strong Positive Pairs -----
revenue         unique_invoices    0.774076
total_views     unique_streams     0.816240
unique_streams  purchases          0.895052
purchases       total_views        0.979470
                purchases          1.000000
dtype: float64

----- norway Strong Negative Pairs -----
unique_invoices  unique_streams   -0.339396
dtype: float64


----- portugal Strong Positive Pairs -----
unique_invoices  revenue        0.870625
unique_streams   total_views    0.878131
purchases        total_views    0.898735
unique_streams   purchases      0.964373
purchases        purchases      1.000000
dtype: float64

----- portugal Strong Negative Pairs -----
Series([], dtype: float64)


----- singapore Strong Positive Pairs -----
revenue         unique_invoices    0.792429
total_views     unique_streams     0.994686
unique_streams  purchases          0.995666
purchases       total_views        0.996495
                purchases          1.000000
dtype: float64

----- singapore Strong Negative Pairs -----
revenue          unique_streams   -0.839410
total_views      revenue          -0.796255
revenue          purchases        -0.785431
unique_invoices  unique_streams   -0.375269
                 total_views      -0.328951
dtype: float64


----- spain Strong Positive Pairs -----
revenue          purchases      0.341833
unique_streams   revenue        0.345967
unique_invoices  revenue        0.365380
unique_streams   total_views    0.834256
purchases        total_views    0.927247
unique_streams   purchases      0.933247
purchases        purchases      1.000000
dtype: float64

----- spain Strong Negative Pairs -----
Series([], dtype: float64)


----- united_kingdom Strong Positive Pairs -----
revenue          unique_streams     0.330754
total_views      revenue            0.400748
revenue          purchases          0.457264
purchases        unique_invoices    0.619657
unique_invoices  unique_streams     0.673479
                 total_views        0.711250
total_views      unique_streams     0.809947
unique_streams   purchases          0.865378
purchases        total_views        0.927356
                 purchases          1.000000
dtype: float64

----- united_kingdom Strong Negative Pairs -----
Series([], dtype: float64)
```

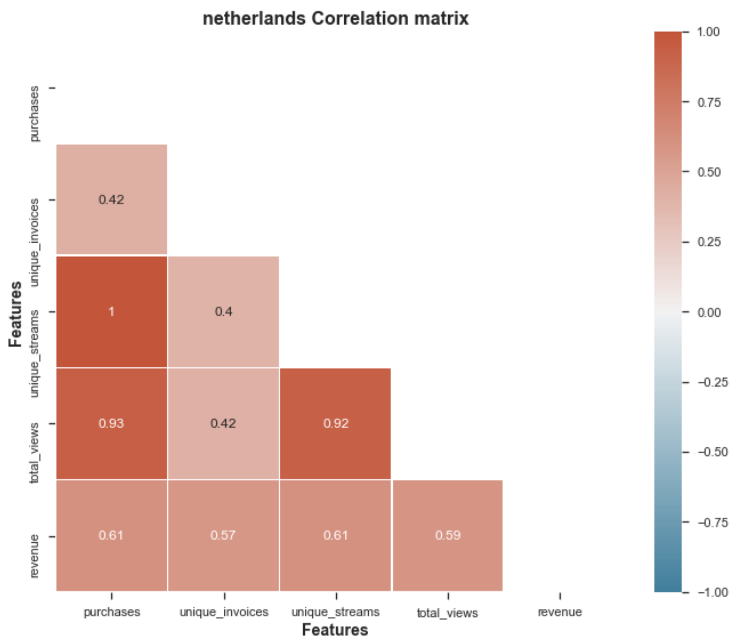

all Correlation matrix
Features
purchases
unique_invoices
unique_streams
total_views
revenue
0.63
0.86
0.69
0.93
0.72
0.81
0.45
0.28
0.33
0.4
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

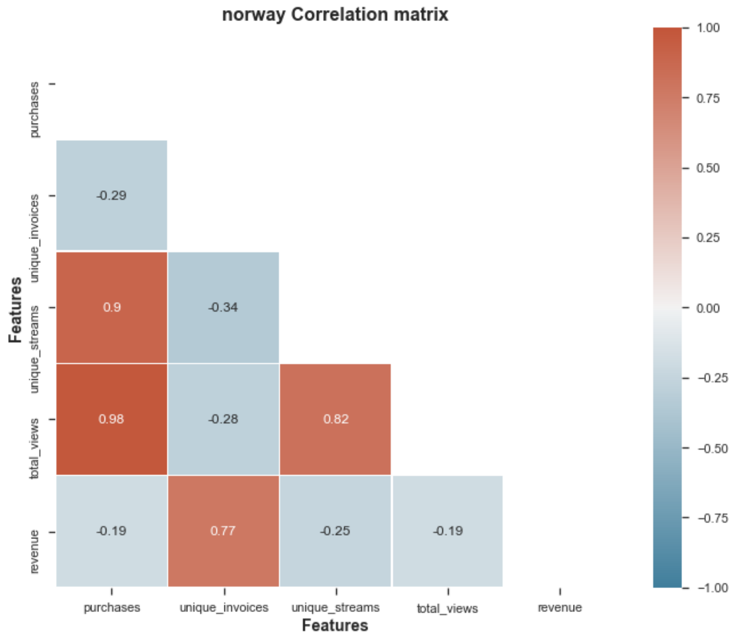

eire Correlation matrix
Features
purchases
unique_invoices
unique_streams
total_views
revenue
0.46
0.98
0.42
0.97
0.44
0.95
0.14
0.47
0.13
0.12
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

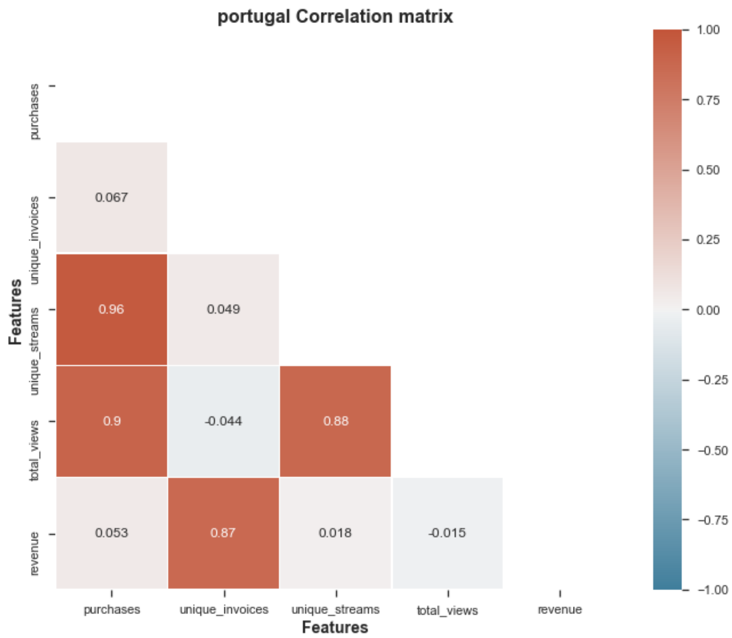

france Correlation matrix
Features
purchases
unique_invoices
unique_streams
total_views
revenue
0.65
0.96
0.62
0.95
0.62
0.91
0.3
0.41
0.27
0.27
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

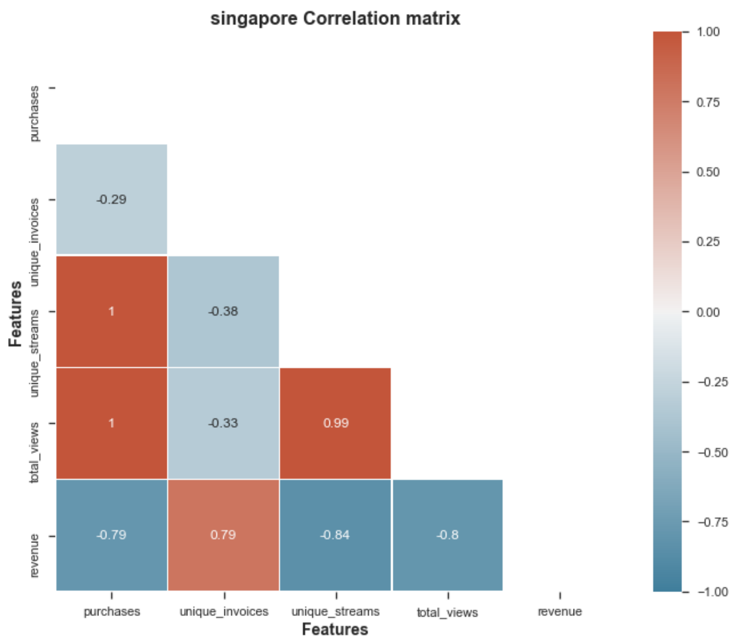

germany Correlation matrix
Features
purchases
unique_invoices
unique_streams
total_views
revenue
0.56
0.97
0.53
0.98
0.54
0.95
0.7
0.58
0.65
0.69
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

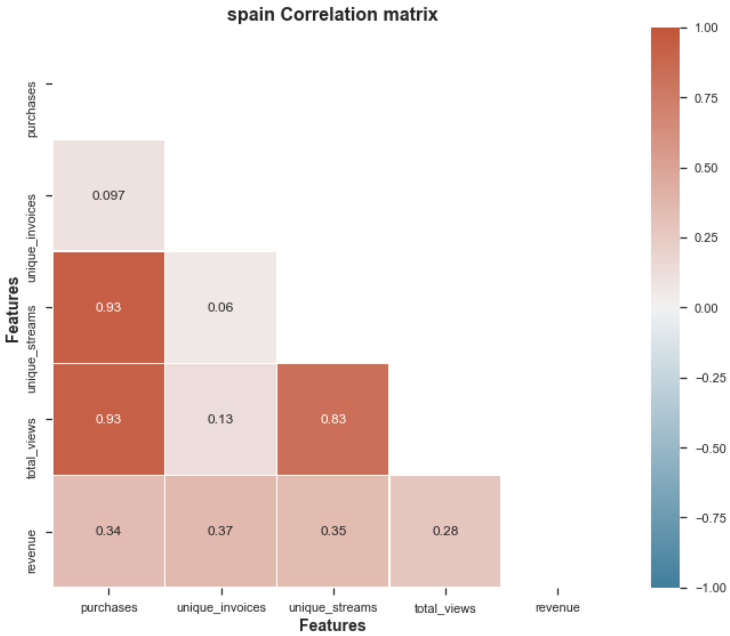

hong_kong Correlation matrix
Features
purchases
unique_invoices
unique_streams
total_views
revenue
-0.78
0.99
-0.77
0.99
-0.82
0.99
-0.76
0.99
-0.75
-0.8
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

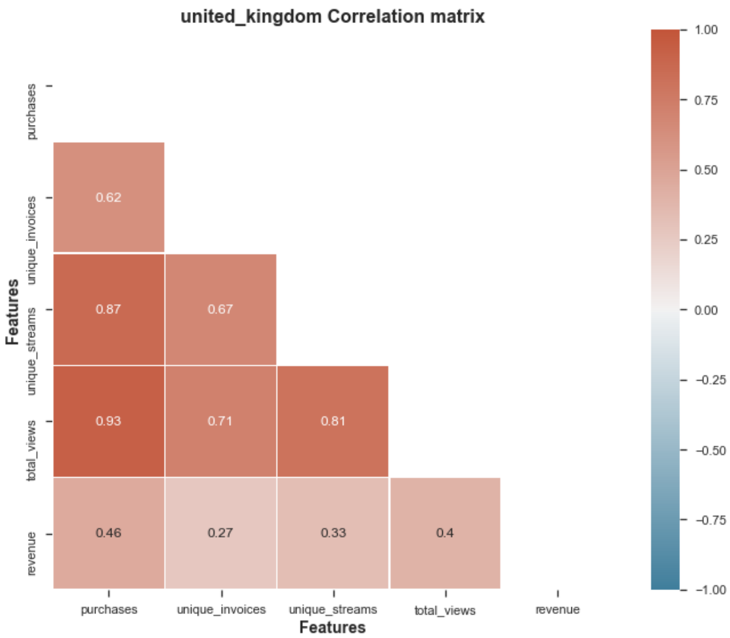

netherlands Correlation matrix
Features
purchases
unique_invoices
unique_streams
total_views
revenue
0.42
1
0.4
0.93
0.42
0.92
0.61
0.57
0.61
0.59
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


norway Correlation matrix
purchases
unique_invoices
unique_streams
total_views
revenue
-0.29
0.9
-0.34
0.98
-0.28
0.82
-0.19
0.77
-0.25
-0.19
Features
Features
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

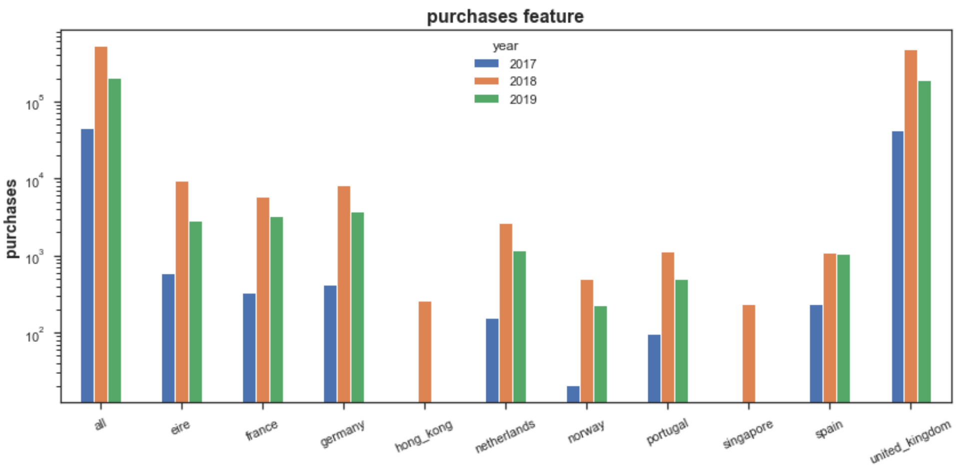

portugal Correlation matrix
purchases
unique_invoices
unique_streams
total_views
revenue
0.067
0.96
0.049
0.9
-0.044
0.88
0.053
0.87
0.018
-0.015
Features
Features
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

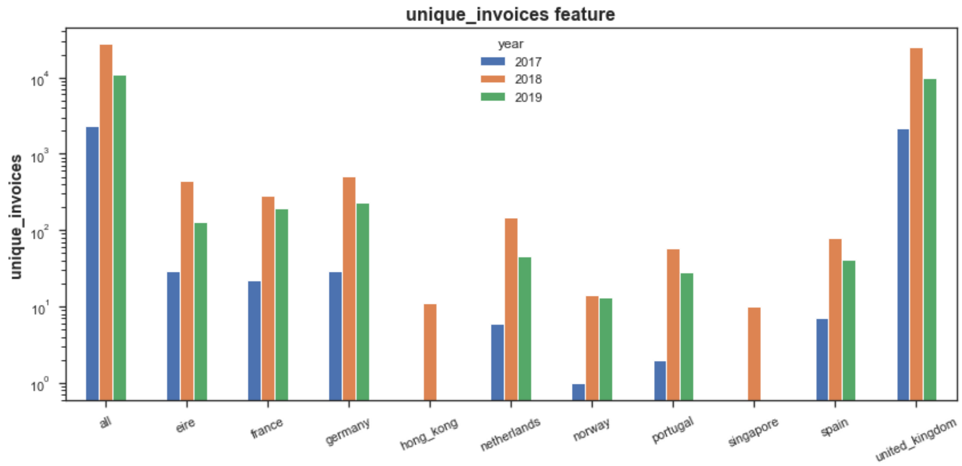

singapore Correlation matrix
purchases
unique_invoices
unique_streams
total_views
revenue
-0.29
1
-0.38
1
-0.33
0.99
-0.79
0.79
-0.84
-0.8
Features
Features
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

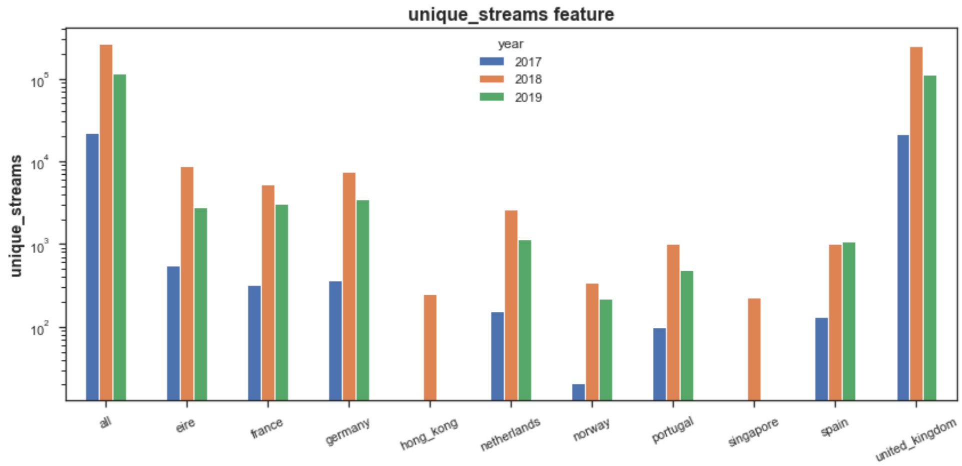


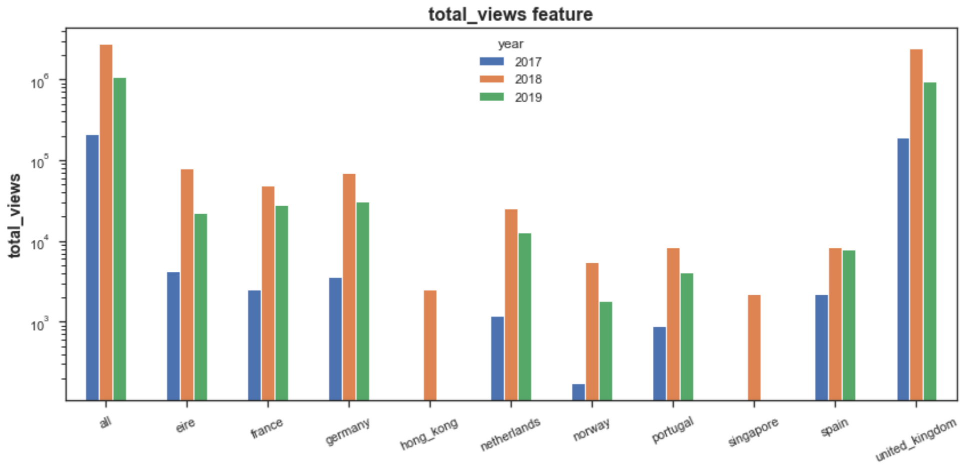


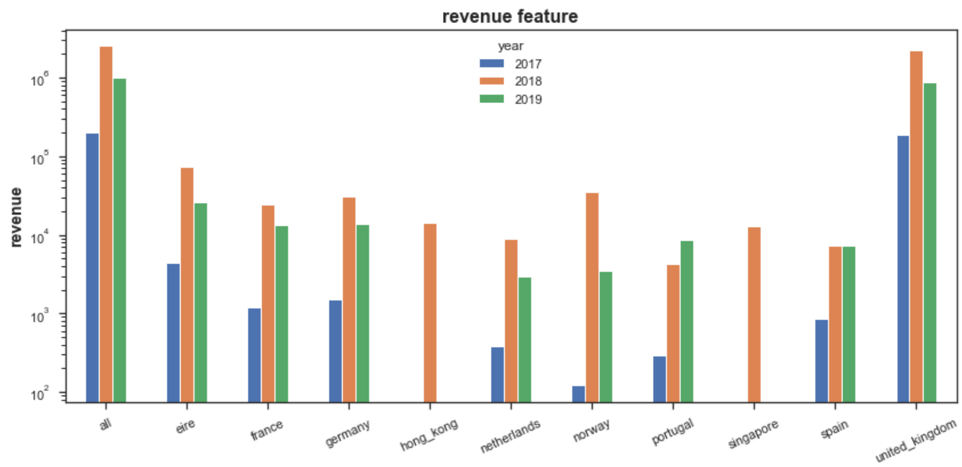
Features Visualization Over Countries



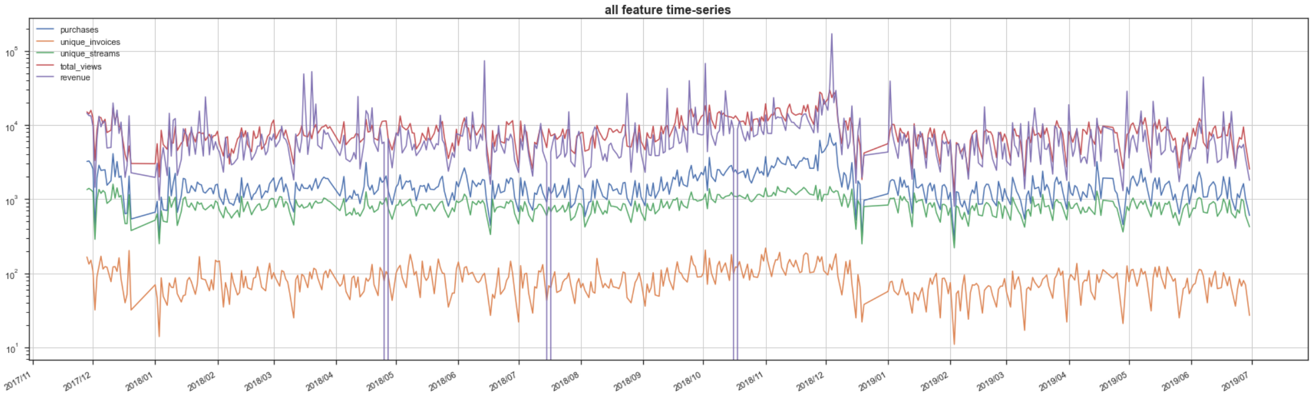


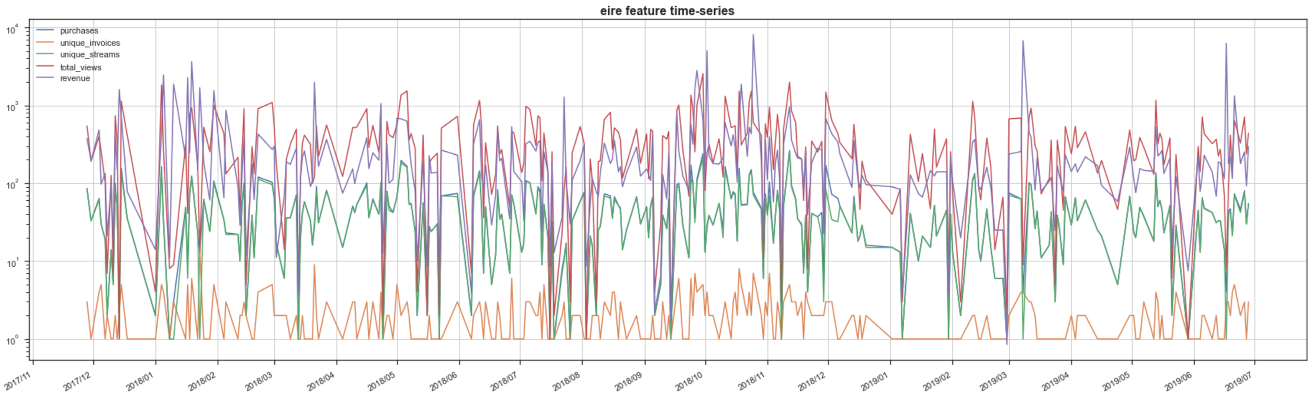


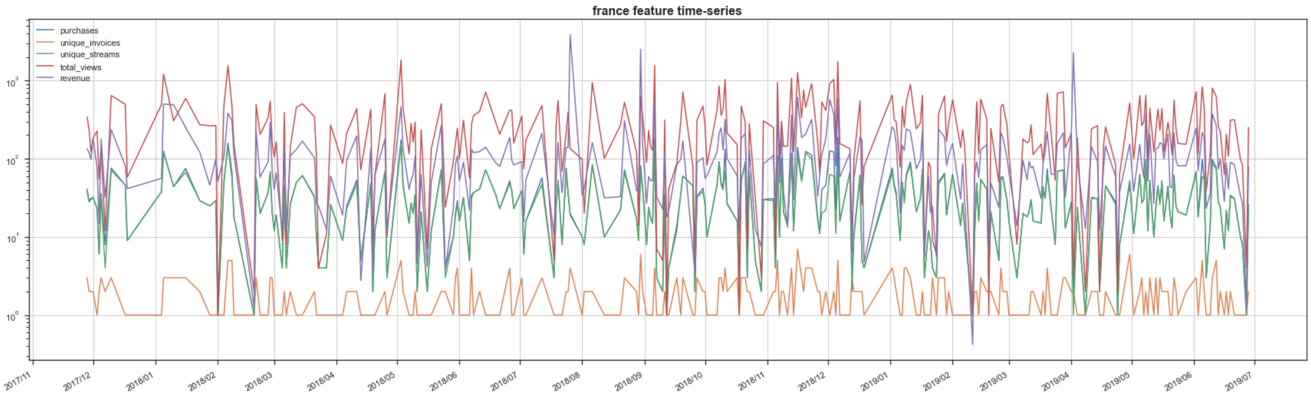


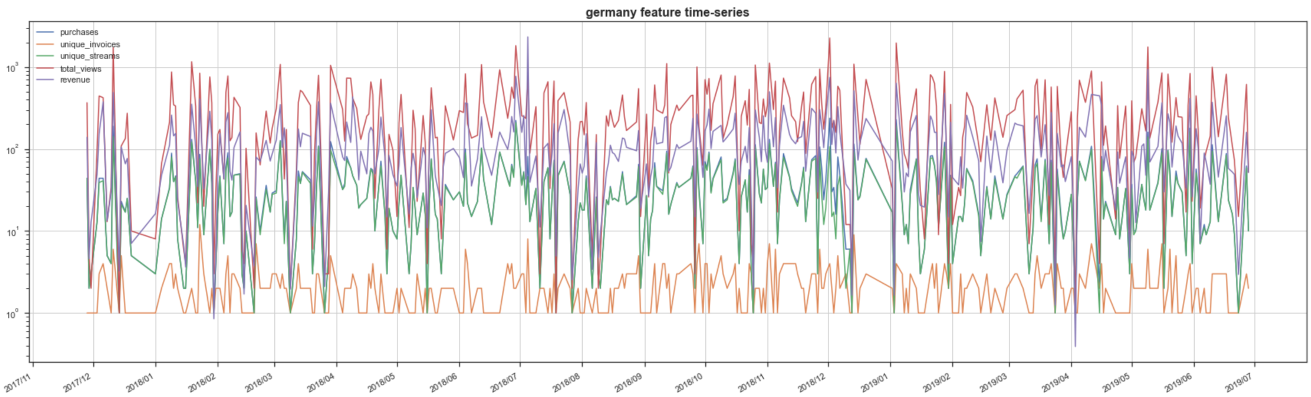


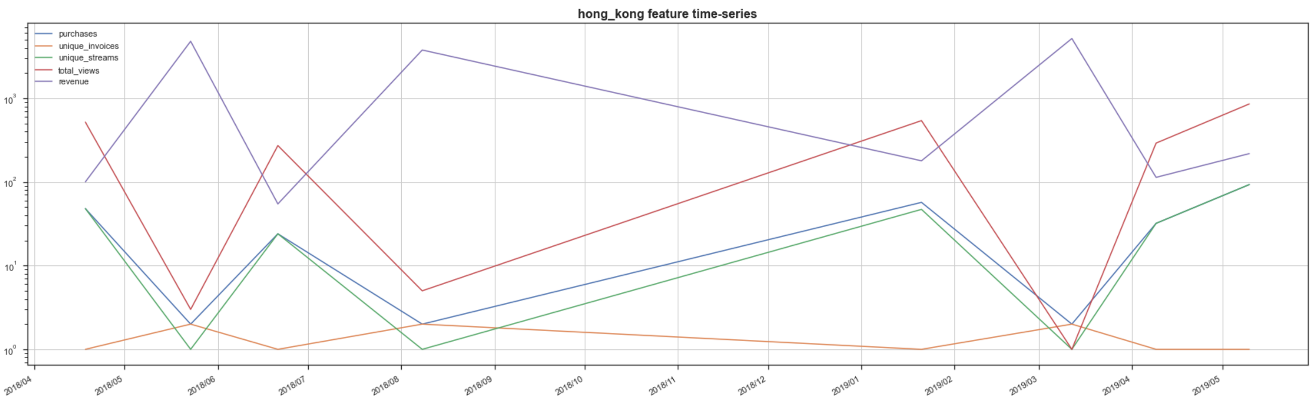
Time-series plotting

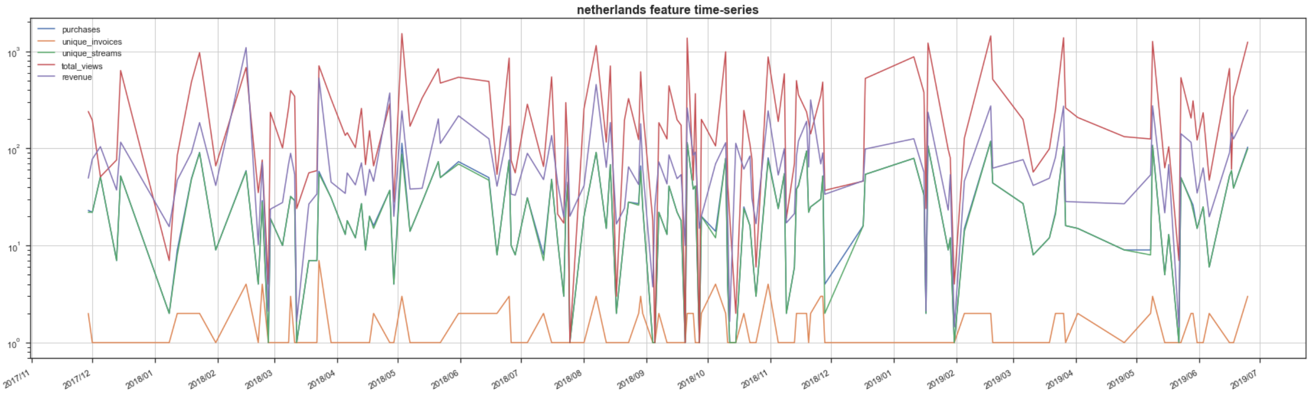

all feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue

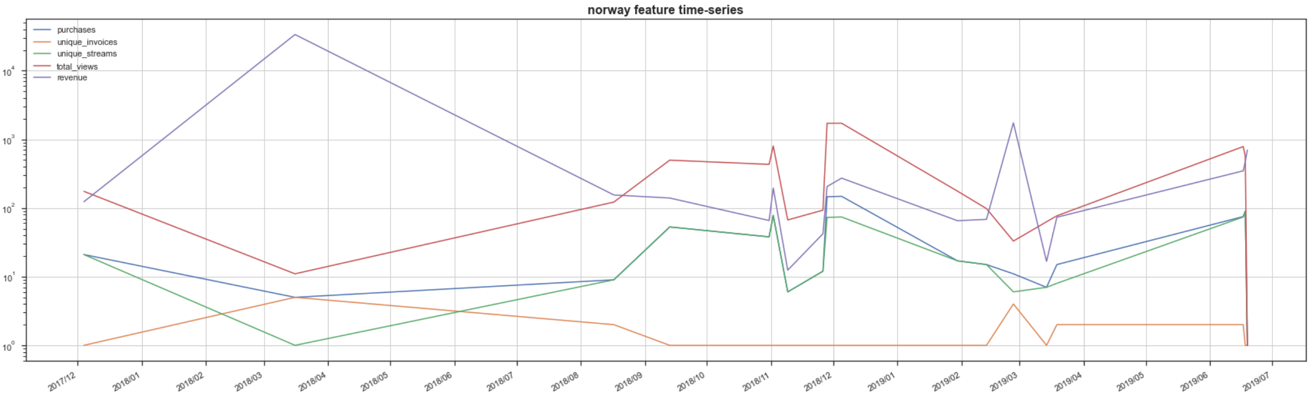

eire feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue

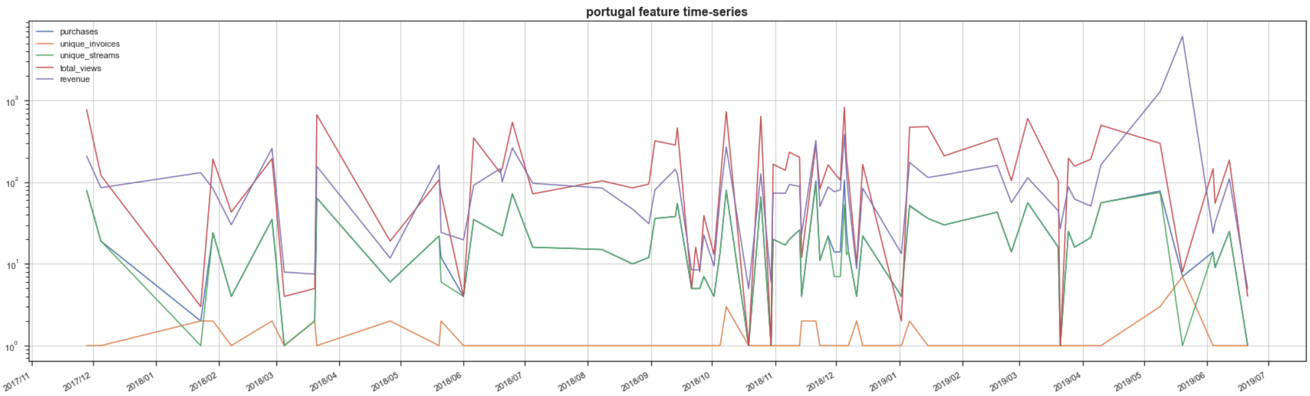

france feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue

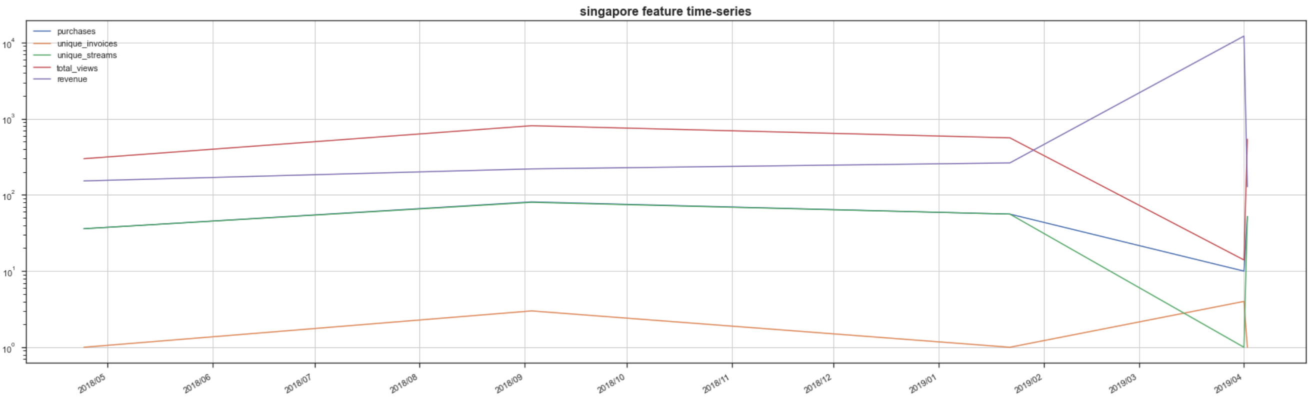

germany feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue

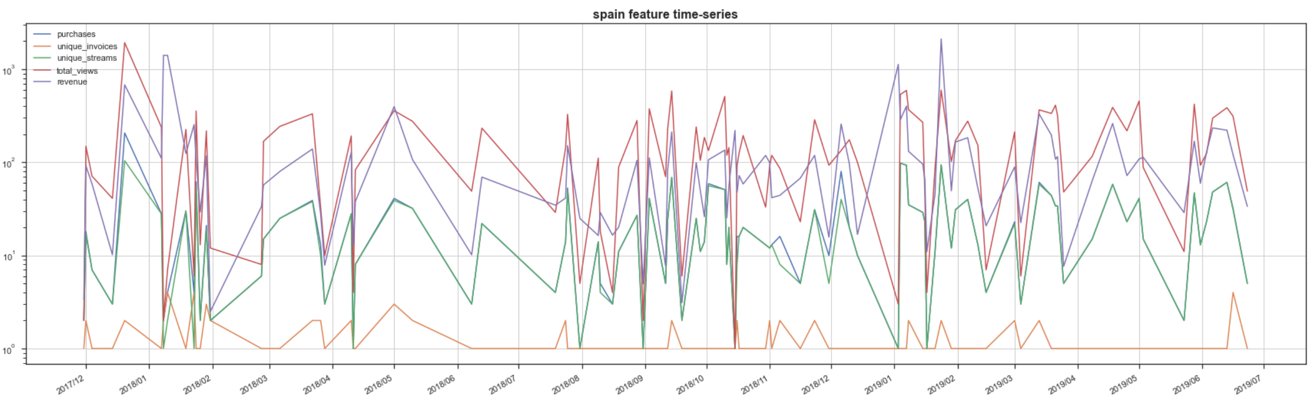

hong_kong feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue

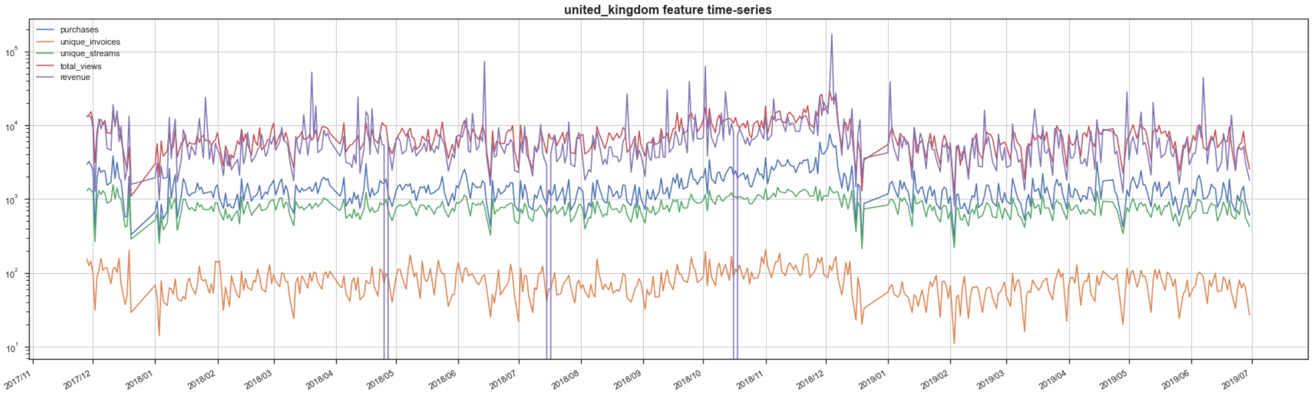

netherlands feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue


norway feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue


portugal feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue


singapore feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue


spain feature time-series
purchases
unique_invoices
unique_streams
total_views
revenue



> (5.) Articulate your findings using a deliverable with visualizations.

Summary:

Missing Data:

- Overall is missing 22.9% values in dataset from total of 607. When consider segmentation for separate country markets, It can be divided into three categories based on amount of missing data. Into **first category** falls only UK market where is missing 22.9% of the values. **Second category** is consist of markets where missing data are in range from 44 to 56%. Those countries are Eire, France and Germany. The rest of the countries (Hong Kong, Netherlands, Norway, Portugal, Singapore, Spain) falls into **third category** where missing values ranges from 78 to 99%.
- Based on fact that significant proportion of data is not available for the most of the markets, It was decided to drop missing columns and not to use any imputation technique to not impact data bias in larger scale.

Features and Business Metric Correlation

- There has been created pair plots and feature correlation matrices overall and for particular markets. From perspective of gathered data of all markets, there is only positive and quite strong correlation between features. The strongest correlation (0.93) is observed between feature ***purchases*** and ***total_views*** followed by ***purchases*** and ***unique_streams*** at 0.86. Our business metric (***revenue***) correlates the most to ***purchases*** at 0.45 and ***total_views*** at 0.4. Based on that can be assumed that focus on globally leveraging ***total_views*** and ***purchases*** might have positive impact on overall revenue.
- Similar behavior can be observed on the markets which falls into **first** and **second category** missing of data. Additionally, it can be observed also strong correlation between ***revenue*** and ***unique_invoices***.
- In case of countries like Hong Kong, Singapore, Norway and Portugal is possible to see significant negative or no correlation between most of the features. All of them falls into **category three** where the most of the data missing. Although this can point for local market specifics, it's necessary to take into account that noticeable lack of data can create significant bias. Therefore, no conclusion is provided for this category at the moment.

Features Visualization Over Countries

- When total sum of each feature is compared over the years 2018 and 2019, it is possible to notice stagnating trend. It is necessary to take into account that data for year 2019 are available only up to 07/2019, and so they cannot be compared with the year 2018 in full extent. However, it's possible to state that global ***revenue*** for the year 2019 at approx. 60% timeframe of the whole year reach only 39,5% of ***revenue*** generated during the year 2018. This can be seen as a decreasing trend but that's not exactly true when the time-series characteristic of the revenue is taken into account. On that, there is possible to see from 09/2018 to 12/2018 there was noticeable rise in sales on overall revenue. If we compare the same periods for both years (01/2018-07/2018 and 01/2019-07/2019) it can be seen that revenue trend doesn't change.
- The trend discussed above is more or less visible in the case of countries which falls into **first** and **second category** with similar proportions in revenue generation. Exception is visible on Portugal and Spain market where ***revenue*** for 2019 already exceeds those from 2018. This can be explained by proportionally significant lack of data up to 09/2018 which causes tracking lower amount of revenue. (see relevant time-series plots)
- For Hong Kong and Singapore counties there is no conclusion provided at the moment due to lack of data.

Conclusion & Recommendations

To use data in supervised modeling pipeline, it is vital to take into account significant lack od data for particular markets. Prediction performance can be negatively impacted and biased for those markets where majority of data is missing. On the other hand, there were already detected features (***purchases***, ***total_views***, ***unique_streams*** and ***unique_invoices***) on "data-rich" markets which can have positive impact to generating revenue if proper strategy for their stimulation is implemented. Such a strategy can be formulated with help of prediction tool which can be trained on data from "data-rich" countries. Subsequently, this strategy (with relevant modifications) can be also applied to markets where data is lacking at the moment.