

# Dependence of risks on product logic

## (5. Audience structure, last year)

Problem statement:

**Hypothesis:** Locations with a low proportion of active women have a higher risk level.

**What to check:**

- Build a map of the proportion of active women in locations.
- Correlation between the proportion of active women and fraud metrics in a location.

To check whether there is a correlation between the proportion of active women and fraud metrics (we will consider `fraud_rate_usd`, `vamp_rate_orders`, `chargeback_rate_usd/orders`, `decline_rate_orders`, `alert_rate_usd/orders`, `complaints_rate_usd/orders`, and the percentage of unsubscribes immediately after subscriptions) in different countries, we first need to collect data on active women.

We will take data for the last year and review it on the Web, as it has the highest risk of fraud, within two product groups - A and B. We will consider those users who perform at least one activity consistently at least once a week to be active. Such generous limits for defining active users may inflate their actual number in some way, but it is important for us to study general trends, so we will accept this for now.

We calculate the main metrics as follows:

*`Fraud_rate_usd`: Calculate the sum of all payments where `fraud = 1` divided by the sum of all payments.*

*`Vamp_rate_orders`: Calculate the number of orders from `a.ov` divided by the number of all orders in `a.opdp`.*

*`Chargeback_rate_usd`: Calculate the sum of all payments where `chb = 1` divided by the sum of all payments.*

*`Chargeback_rate_orders`: Calculate the number of orders where `chb = 1` divided by the number of all orders.*

*`Decline_rate_orders`: We calculate the number of orders where `o_status = 'decline'` divided by the number of all orders (we do not calculate `Decline_rate_usd` because in all payments where `o_status = 'decline'`, `g_usd` is always equal to 0, so `decline_rate_usd = 0` is always true).*

*`Alert_rate_usd`: We calculate the sum of all payments where `chb_p_dt` IS NOT NULL divided by the sum of all payments.*

*`Alert_rate_orders`: We calculate the number of all payments where `chb_p_dt` IS NOT NULL divided by the number of all payments.*

*`Complaints_rate_usd`: We calculate the sum of all payments whose users are in the `a.ccbd` table, divided by the sum of all payments.*

*`Complaints_rate_orders`: We calculate the number of all payments whose users are in the `a.ccbd` table, divided by the number of all payments.*

*Unsubs\_rate\_usd: We calculate the sum of all subscription purchases that the user canceled within 24 hours after purchase, divided by the sum of all payments.*

*Unsubs\_rate\_orders: We calculate the number of all subscription purchases that the user canceled within 24 hours after purchase, divided by the number of all payments.*

*We take the unsubscription data from the p.uf table.*

We will limit the data for voters as follows:

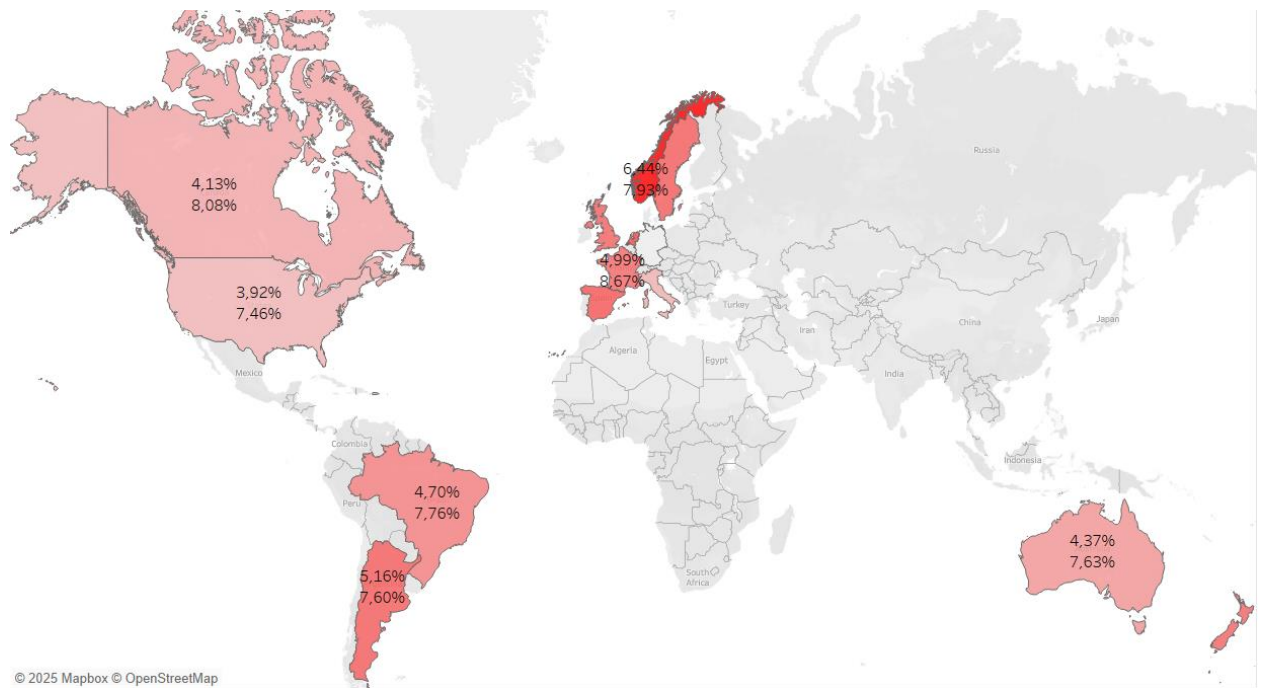
- 1) We will remove all test users
- 2) When analyzing countries, we will exclude those where the number of registered users does not exceed 10,000, because a smaller number does not allow us to objectively assess the proportion of active women in that location.

Having dealt with all the formalities, we move on to the research itself.

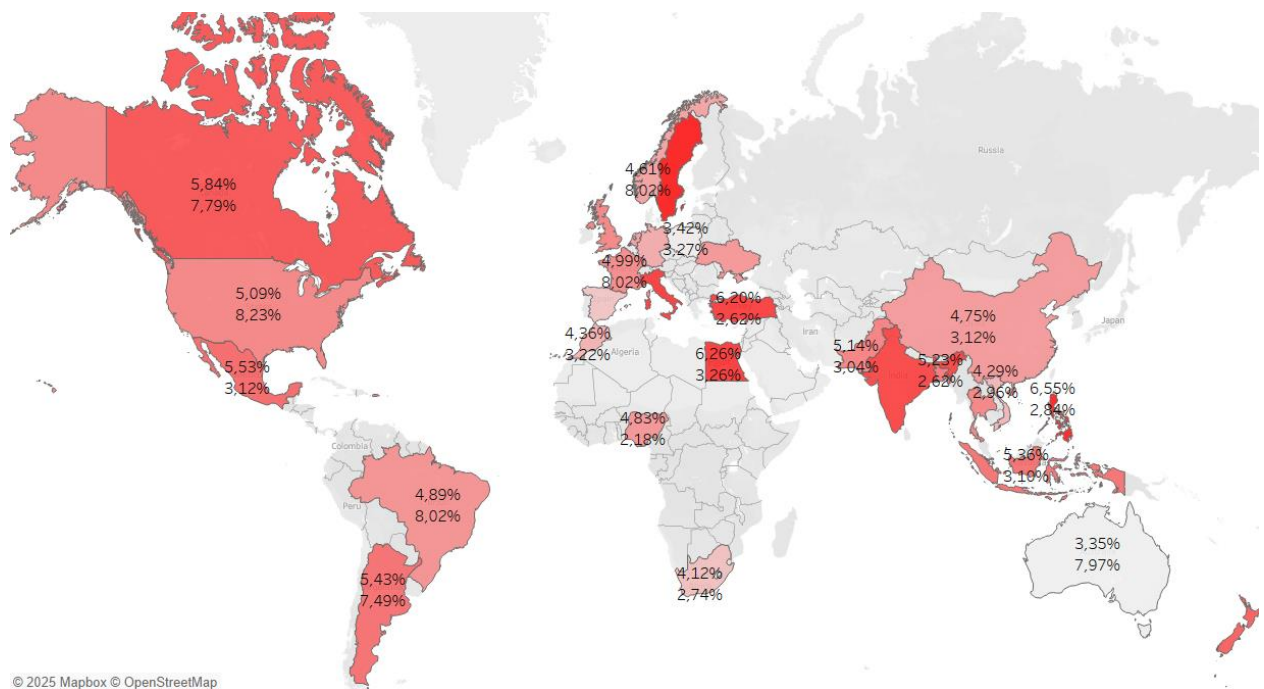
Fraud Rate	Vamp Rate (Orders)	Alert Rate (USD / Orders)	Chargeback Rate (USD / Orders)	Active Women Rate	Total Number of Users	Total Number of Orders	Active Women	Gross
2,5%	0,3%	3,5%	1,5%	2,5%	43,564,543	15,235,353	935,254	\$31 435 473
Alerts	Chargebacks	Fraud	Vamps	Complaints		Declines	Unsubscriptions (In 24 hours after subscription)	
54,653	26,214	60,146	24,643	285,464		9,224,463	2,853,575	
product_group_name (Multiple values)					Complaints Rate (USD / Orders)		Declines Rate (Orders)	
Type of Rate USD							Unsubs Rate (In 24 hours after subscription, USD / Orders)	
app Web					2%		55,76%	
							25,6%	

The average indicators for A (Web) and B (Web) (we will look at each product group separately below) show that over the year, our fraud rate is 2.5%, and the percentage of active women is 2.5% of all nearly 44 million users (for now, we will analyze the relationship between active\_women\_rate and fraud\_rate only, as these are our two main metrics; we will return to other indicators later). The indicators are close to each other, but it is not yet clear how they are related. To find out, I suggest referring to the following visualizations:

A (Web)



B (Web)



*Note: All the real values are hidden for the public share. Values on the screenshots are randomly generated for the presentation*

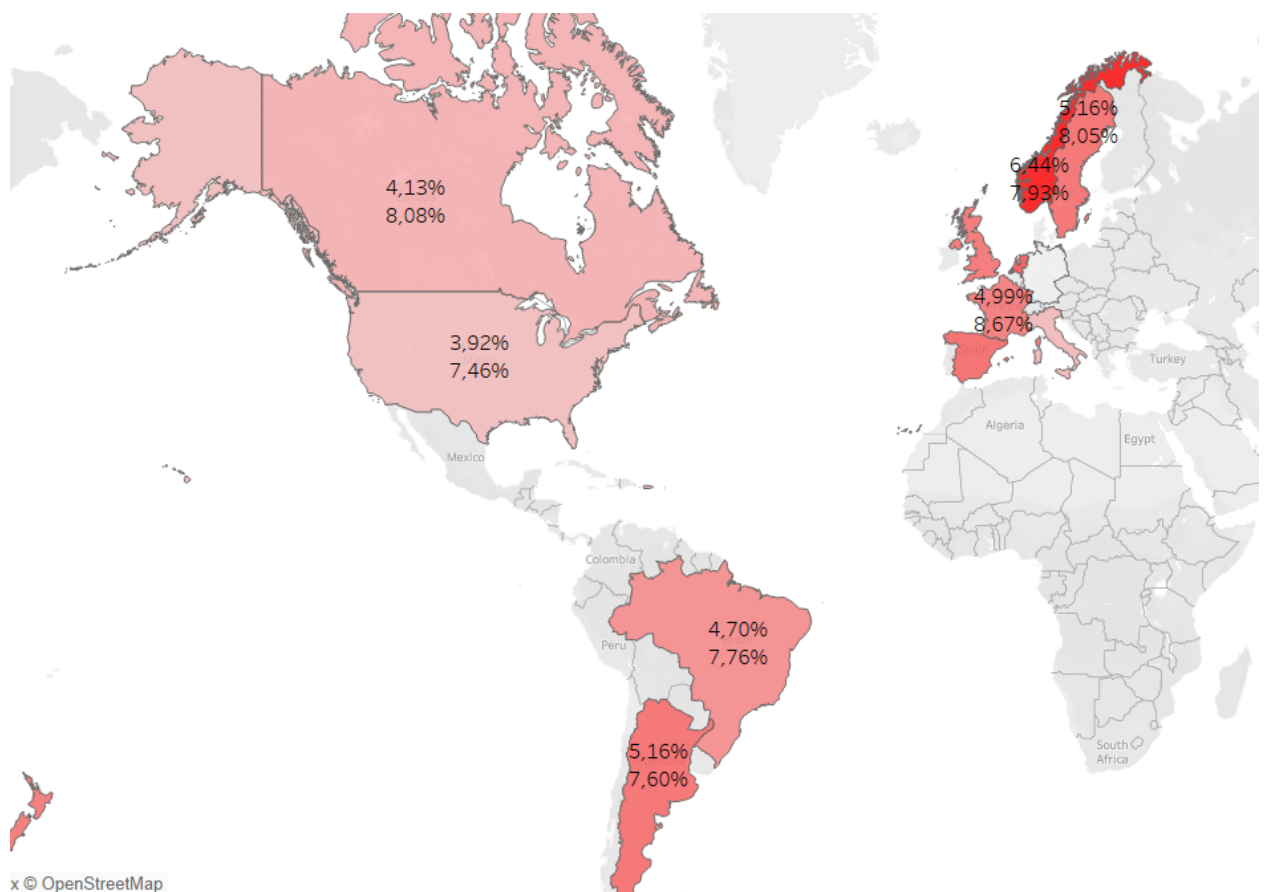
This map is actually what is required of us in the first sub-item of our task. On it, we see a map of countries colored in different shades of red depending on the level of active women there: the lower the level, the darker the country; the higher the level, the lighter the country (remember that countries with fewer than 10,000 users were not included in

the sample). The percentage value we see in the label is not the percentage of active women, oddly enough, but the level of fraud.

In fact, thanks to this map, we can conduct our first visual study. So, if our hypothesis is as follows: “Locations with a low proportion of active women have a higher level of risk,” this means that darker countries — where the proportion of active women is lower — should have a higher fraud rate, and vice versa, lighter countries — where the proportion of active women is higher — should have a lower fraud rate. Let’s take a closer look at how this corresponds to reality.

It is important to note that since the proportion of women in our services is generally low compared to men (specifically, in the sample from our study, 541,783 women out of 3,856,650 users, or 14%), we will consider 10% of active women to be a very high number.

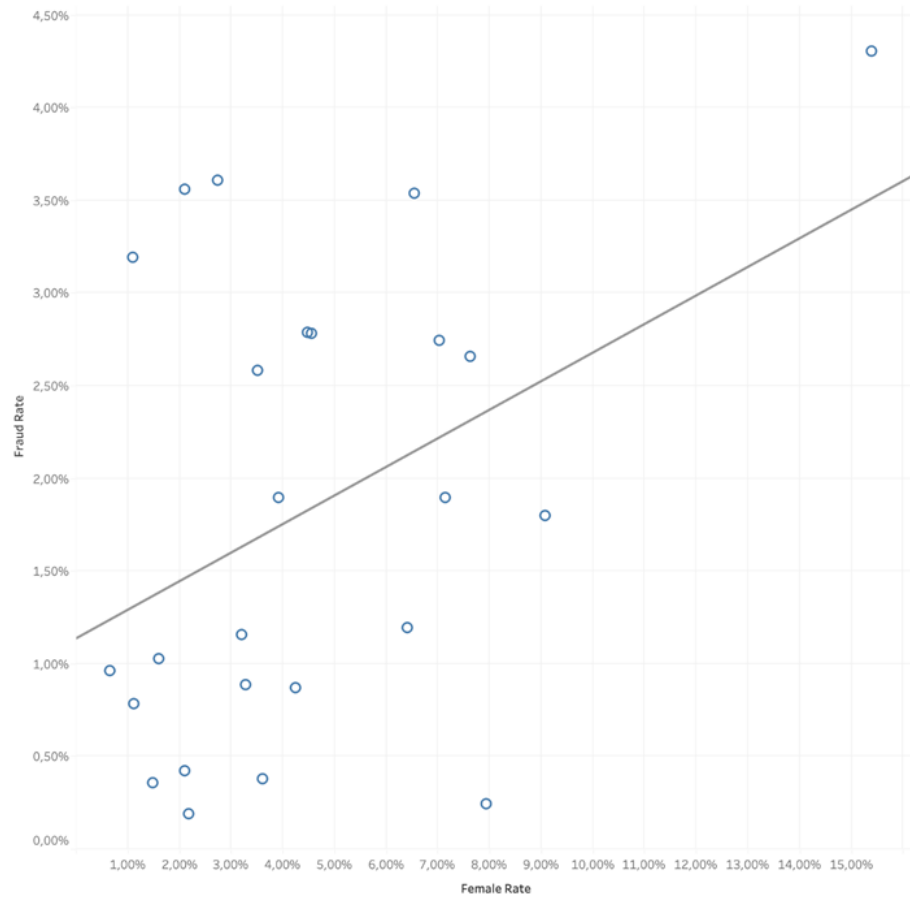
#### A (Web)



#### B (Web)

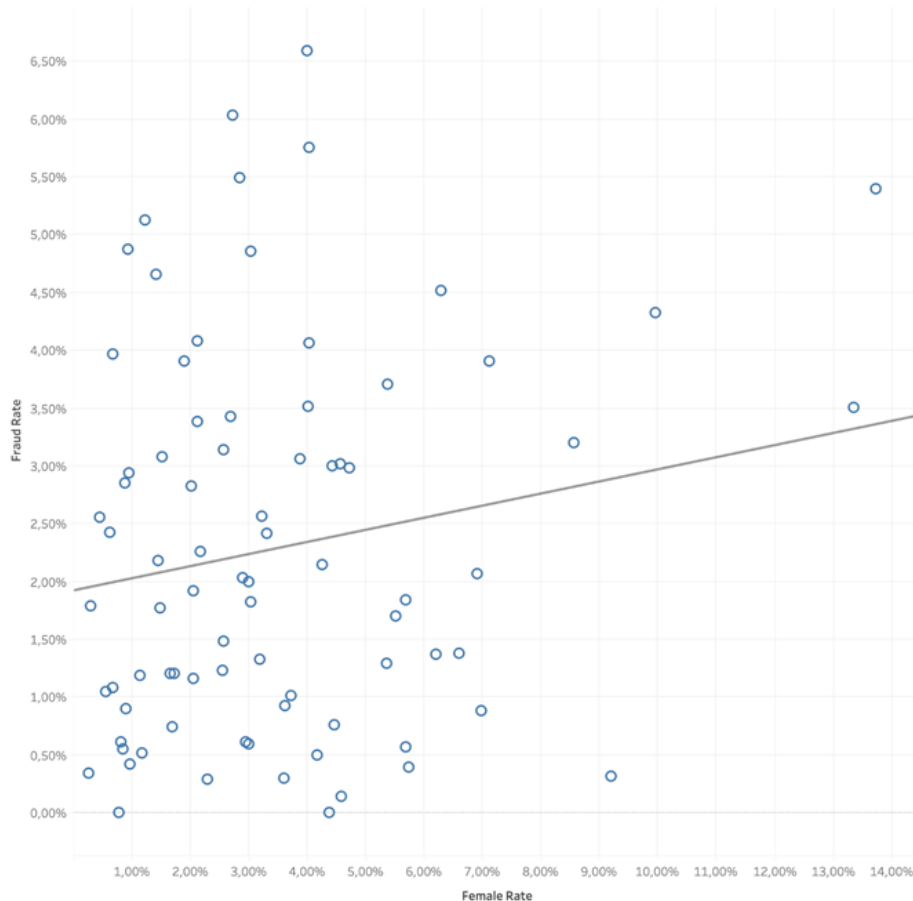


Scatter Plot



B (Web)

Scatter Plot



In this case, we have a scatter plot graph of countries by `activity_women_rate` and `fraud_rate`. As we can see, in the case of A and B, the countries are distributed quite chaotically across the coordinate plane. We see that we have countries with a small proportion of active women and low fraud, countries with a large proportion of women and high fraud, countries with a small proportion of active women and high fraud, and countries with a large proportion of active women and low fraud.

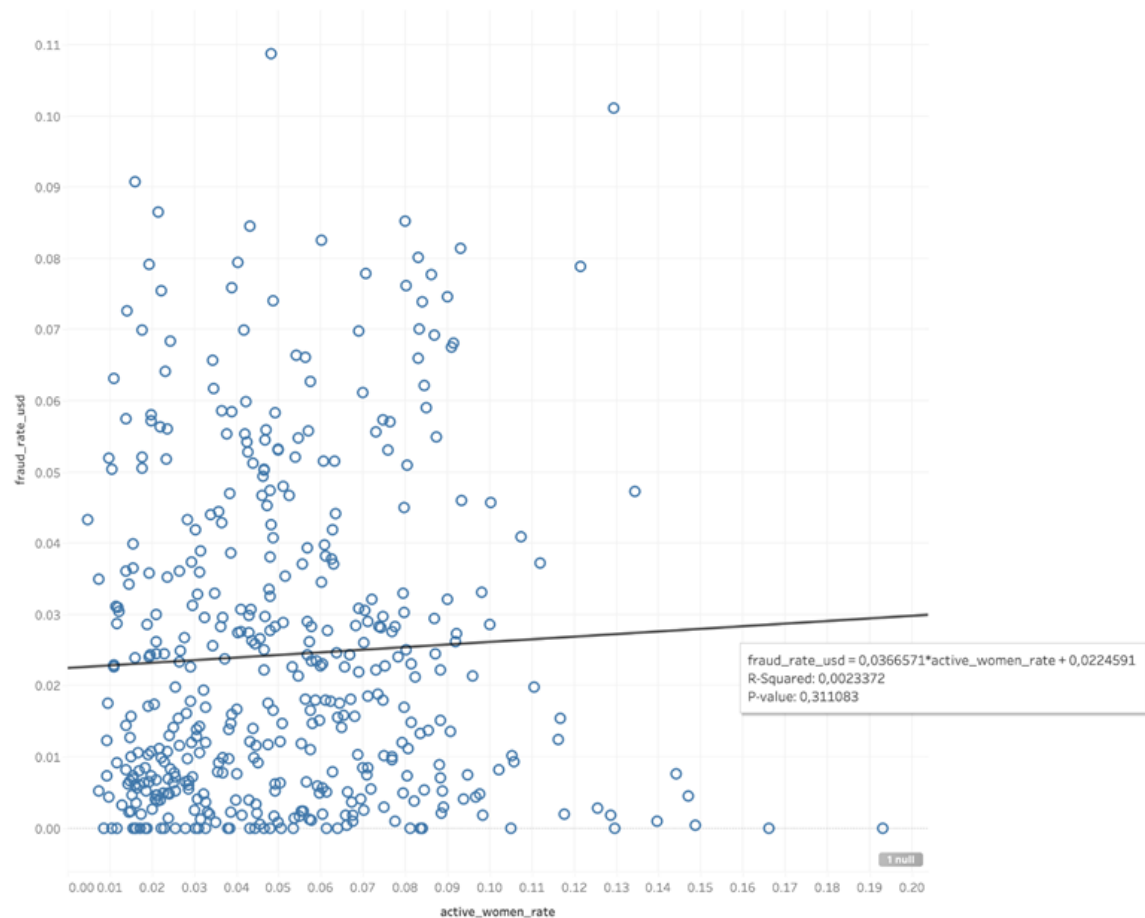
However, while in the case of B the trend line is quite flat, indicating that if there is any correlation, it is very insignificant, in A it is directed upwards at a much greater angle, indicating a greater direct dependence. So let's explore this in more detail.

The trend line indicates that between `active_women_rate` and `fraud_rate` for B,  $R^2$  is approximately 0.0297, and  $p\text{-value} = 0.1235$ .  $P\text{-value} > 0.05$ , so there is no statistically significant evidence that `active_women_rate` affects `fraud_rate`, i.e., there are no grounds for rejecting the null hypothesis. Pearson's coefficient square is 3%, meaning that the relationship between `active_women_rate` and `fraud_rate` is quite small.

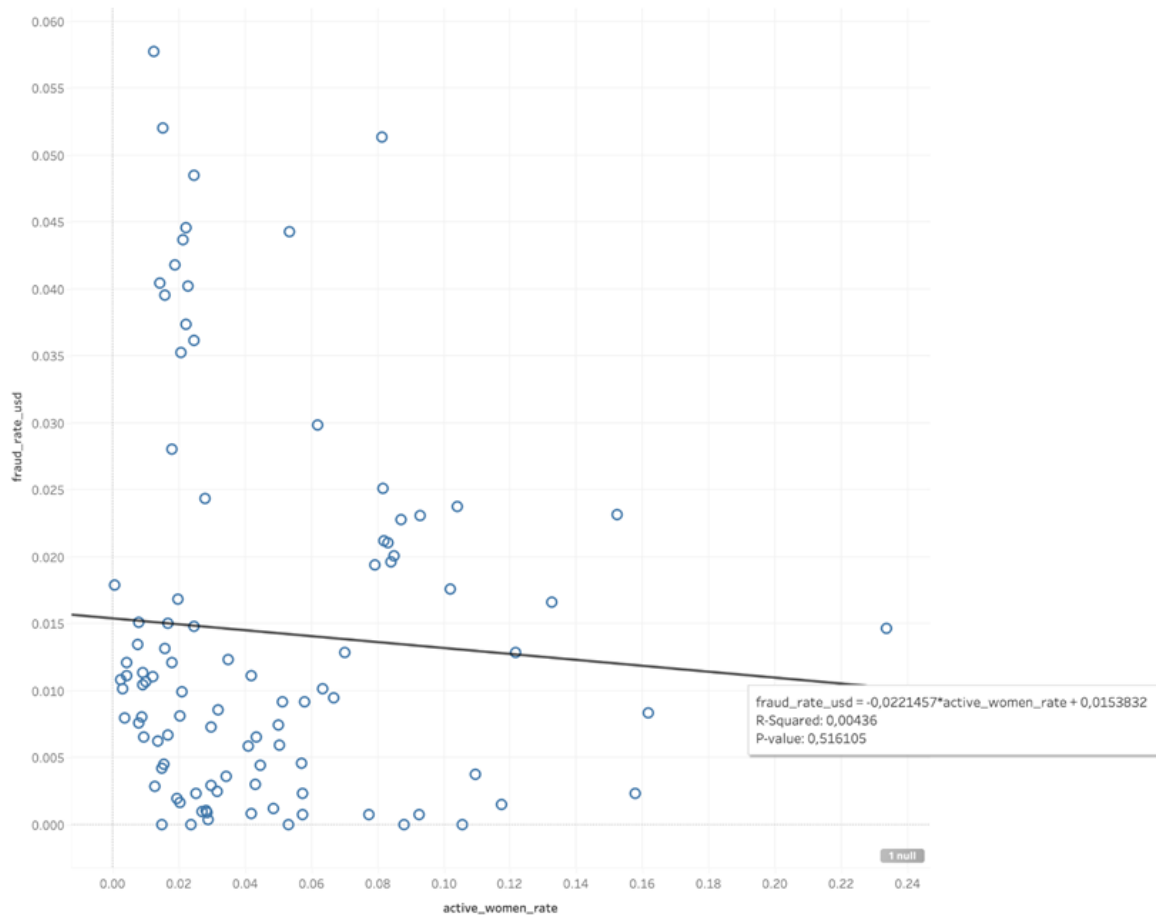
The situation is different at A:  $R^2 = 0.1663$  (i.e., 16%, which is still not a high correlation between the two metrics, but significantly higher than in the case of B), and  $p\text{-value} = 0.0429$ , i.e., less than 5%  $\Rightarrow$  there are signs of rejection of the null hypothesis. Does this mean that the proportion of active women in A affects fraud? Here, it is worth paying attention to the data we have. After filtering out the noise, only 25 countries remained in A, while in B the sample is much larger, which helps to see the trend better. The smaller number of countries distorts the data in A, so the result is incorrect.

In order to make the sample more representative, I suggest looking at the distribution not only of countries by active\_women\_rate and fraud\_rate, but also of countries by month. This will help us see the bigger picture in case the fraud dynamics of some countries have changed significantly. Let's create a Country + Month field and make a new scatter plot based on it. Also, since a single point on the graph is now not just a country, but a country for a specific month, we will filter out noise based on 10,000 users and 833 users (10,000 / 12) as the lower limit.

B (Web)



A (Web)

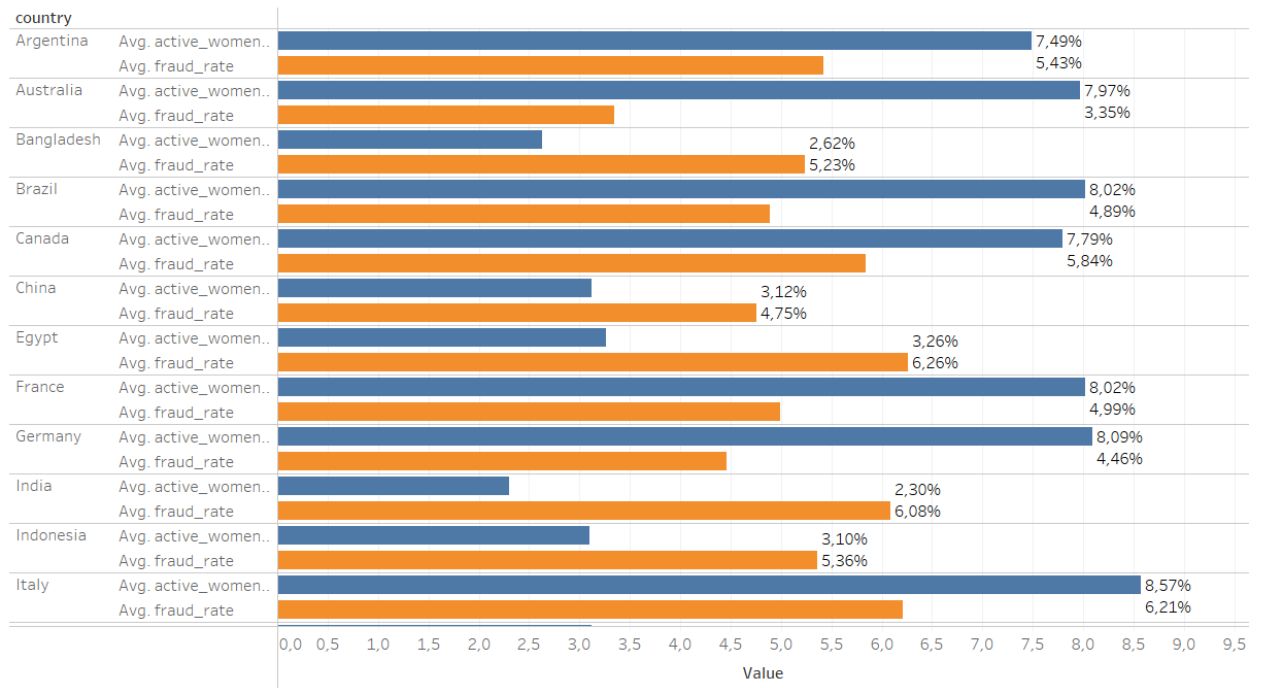


As we can see,  $R^2 = 0.0023$ ,  $p\text{-value} = 0.311$ , i.e., the  $p\text{-value}$  is still greater than 0.05, and  $R^2$  is still very small, everything converges, and in A  $R^2 = 0.0043$ ,  $p\text{-value} = 0.5161$   $\Rightarrow$  as a result, in A the relationship between `active_women_rate` and `fraud_rate` is very low, and there are no signs of rejecting the null hypothesis.

To make sure that our data is not distorted, I suggest calculating  $R^2$  and  $p\text{-value}$  for the top 15 countries by number of active women.

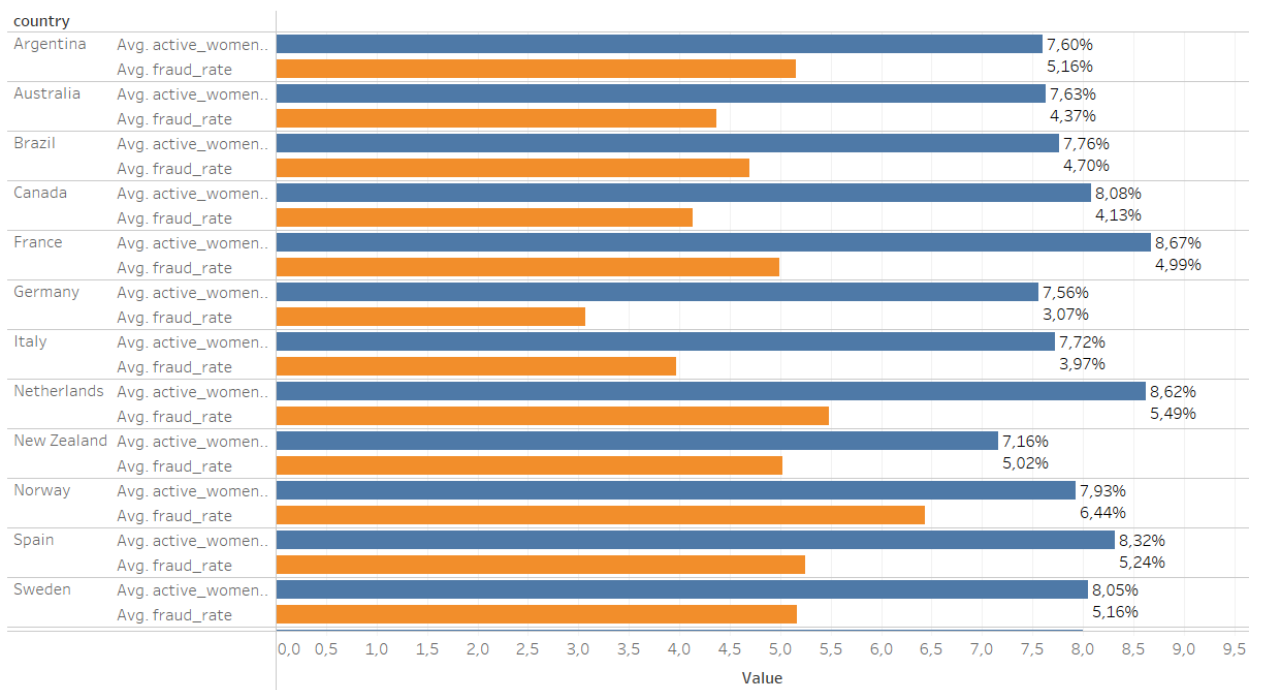
B (Web)

Female Rate Bar Chart



A (Web)

Female Rate Bar Chart



As we can see, the list of top countries includes quite diverse states in terms of the number of fraud cases, the proportion of women, region, culture, population, etc. This means that even if we exclude a significant number of countries from this study, our sample will still be very diverse and will allow us to objectively look at trends specific to the countries that are most important for our products. This will be an additional study that will allow us to confirm that there is no direct connection between the two metrics we are interested in.

For A, we will take data from Country A, B, C, D, E, F, G, H, I, J, K, L, M, N, O.

```
'female_rate': [4.48, 9.08, 3.29, 1.61, 7.93, 2.12, 0.67, 2.18, 3.61, 6.41, 6.54, 7.63, 2.75, 7.03, 15.4],  
'fraud_rate': [2.79, 1.8, 0.89, 1.03, 0.24, 3.56, 0.96, 0.19, 0.38, 1.19, 3.53, 2.66, 3.6, 2.74, 4.3]
```

For top 15 countries of B group we have these values:

```
'female_rate': [5.37, 5.68, 2.54, 1.68, 2.85, 4.03, 0.9, 3.59, 2.56, 7.11, 2.29, 0.81, 9.21, 4.25, 3.04],  
'fraud_rate': [3.7, 1.84, 1.23, 0.74, 5.49, 4.06, 0.9, 0.3, 3.14, 3.91, 0.29, 0.61, 0.31, 2.15, 4.86]
```

Let's execute the following code:

```
import pandas as pd  
from scipy.stats import pearsonr  
  
data = {  
    'country': ['Country A', 'Country B', 'Country C', 'Country D',  
               'Country E', 'Country F', 'Country G', 'Country H',  
               'Country I', 'Country J', 'Country K', 'Country L',  
               'Country M', 'Country N', 'Country O'],  
    'female_rate': [4.48, 9.08, 3.29, 1.61, 7.93, 2.12, 0.67, 2.18, 3.61, 6.41, 6.54, 7.63, 2.75, 7.03, 15.4],  
    'fraud_rate': [2.79, 1.8, 0.89, 1.03, 0.24, 3.56, 0.96, 0.19, 0.38, 1.19, 3.53, 2.66, 3.6, 2.74, 4.3]  
}  
  
df = pd.DataFrame(data)  
  
r, p_value = pearsonr(df['female_rate'], df['fraud_rate'])  
  
print("Pearson r:", r)  
print("p-value:", p_value)  
  
data = {  
    'country': ['Country 1', 'Country 2', 'Country 3', 'Country 4',  
               'Country 5', 'Country 6', 'Country 7', 'Country 8',  
               'Country 9', 'Country 10', 'Country 11', 'Country 12',  
               'Country 13', 'Country 14', 'Country 15'],
```

```

    'female_rate': [5.37, 5.68, 2.54, 1.68, 2.85, 4.03, 0.9, 3.59, 2.56, 7.11,
2.29, 0.81, 9.21, 4.25, 3.04],

    'fraud_rate': [3.7, 1.84, 1.23, 0.74, 5.49, 4.06, 0.9, 0.3, 3.14, 3.91,
0.29, 0.61, 0.31, 2.15, 4.86]
}

df = pd.DataFrame(data)

r, p_value = pearsonr(df['female_rate'], df['fraud_rate'])

print("Pearson r:", r)
print("p-value:", p_value)

```

Result:

A: p-value = 0.1220, r = 0.4169  $\Rightarrow$  R2 = 0.1738

B: p-value = 0.655, r = 0.1257  $\Rightarrow$  R2 = 0.0158

The result is the same — there are no signs of deviation from the null hypothesis, and the correlation is insignificant.