Intersecciones de Bases de Datos
Archivo: Taller-2_LPrentt-20-03-2021.py **Sólo para análisis de intersección. No hay limpieza de base de datos**

No realicé merge de los dataframe = ColombiaCB-5March21.csv, Top100Startups- Colombia.xlsx y Empresas Unicorn - Contactos.xlsx.

Convertí los nombres de las empresas en los 3 dataframe a minúsculas (lowercase) y realice 3 intersecciones:

1. CrunchBase vs Top100, salieron 60 coincidencias
2. CrunchBase vs Unicorn, salieron 19 coincidencias
3. CrunchBase vs Top100 vs Unicorn, salieron 12 coincidencias

Intersecciones de Bases de Datos

Código para opción 1

1. CrunchBase vs Top100, salieron 60 coincidencias. Lo implemente con 2 For

```
df["y"]=0
df["Flag"]=0

intersect=set(df['Organization Name']).intersection(set(dftop['Organization']))
len(intersect)

######## Forma  L. Prentt
k = df.shape[1]
for i in range(df.shape[0]):
    for j in range(dftop.shape[0]):
        if df["Organization Name"][i]==dftop["Organization"][j]:
            df.iloc[i:i+1, k-2:k-1] = 1
            break

###### Forma de Miller Quiroga
for j in intersect:
    df.loc[df['Organization Name'] == j, ['Flag']] = 1
```

Las 2 formas producen los mismos resultados, sin embargo la Miller Quiroga es mas eficiente

Intersecciones de Bases de Datos

Código para opción 2

2. CrunchBase vs Unicorn, salieron 19 coincidencias. Lo implemente con 2 For

```
df1["y"]=0
df1["Flag"]=0

intersect2=set(df1['Organization Name']).intersection(set(dfunicorn['Name']))
len(intersect2)

######## Forma L. Prentt
k = df1.shape[1]
for i in range(df1.shape[0]):
    for j in range(dfunicorn.shape[0]):
        if df1["Organization Name"][i]==dfunicorn["Name"][j]:
            df1.iloc[i:i+1, k-2:k-1] = 1
            break

######## Forma Miller Quiroga
for t in intersect2:
    df1.loc[df1['Organization Name'] == t, ['Flag']] = 1
```

Las 2 formas producen los mismos resultados, sin embargo la Miller Quiroga es mas eficiente

Intersecciones de Bases de Datos

Código para opción 3

3. CrunchBase vs Top100 vs Unicorn, salieron 12 coincidencias. Lo implemente con 2 For

```
df2["y"]=0
df2["Flag"]=0

intersect3=set(df['Organization Name']).intersection(set(dftop['Organization'])).intersection(set(dfunicorn['Name']))
len(intersect3)

# conversión del conjunto intersect3 a numpy.array
int3=np.array(list(intersect3))

######## Forma L. Prentt
k = df2.shape[1]
for i in range(df2.shape[0]):
    for j in range(len(intersect3)):
        if df2["Organization Name"][i]==int3[j]:
            df2.iloc[i:i+1, k-2:k-1] = 1
            break

######## Forma Miller Quiroga
for l in intersect3:
    df2.loc[df2['Organization Name'] == l, ['Flag']] = 1
```

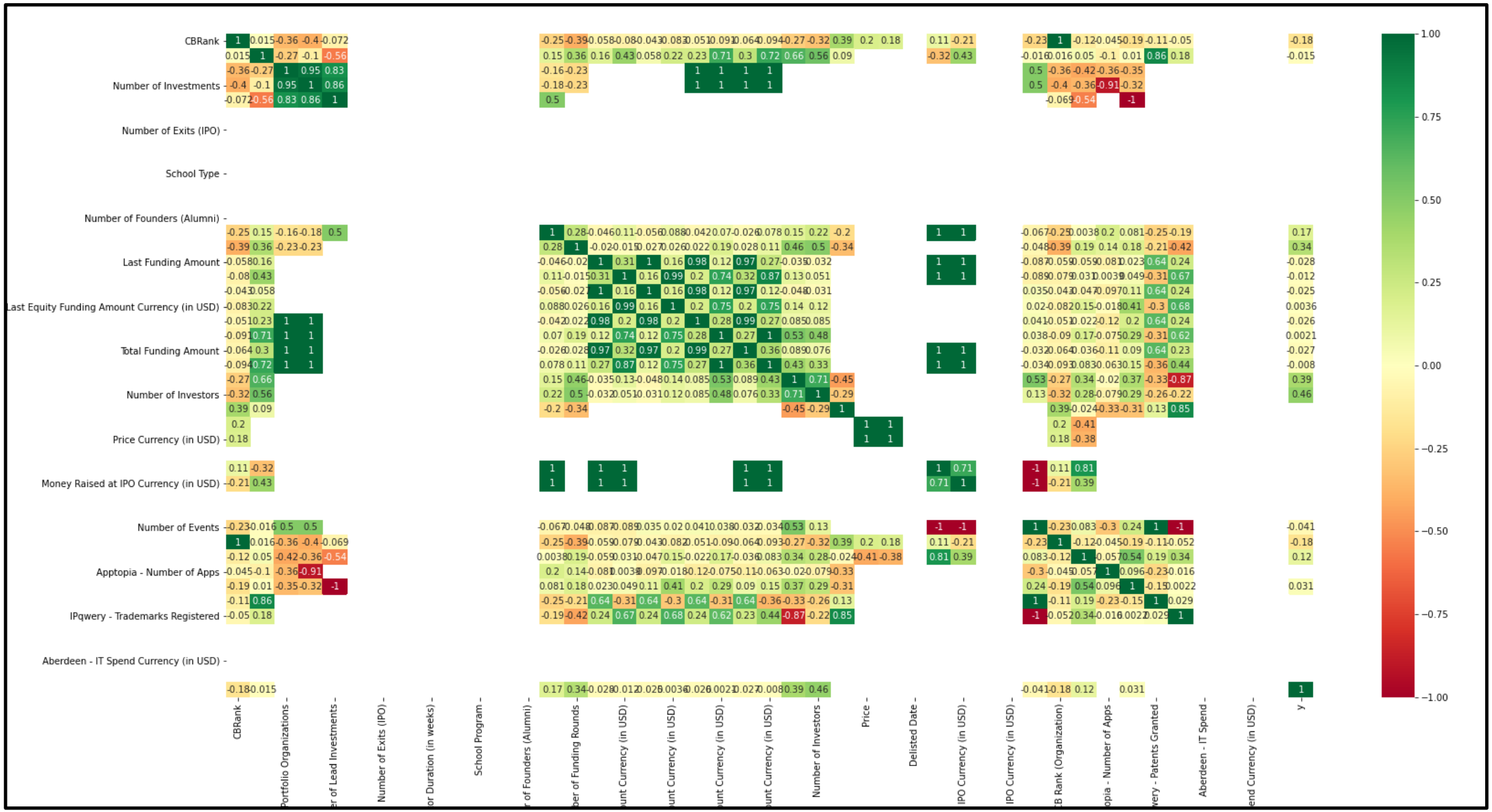
Para Análisis de Regresión se usa el archivo **Taller-2_LPrentt.20A-03-2021.py :**

Se eliminaron todas las empresas que no pertenecen a Colombia

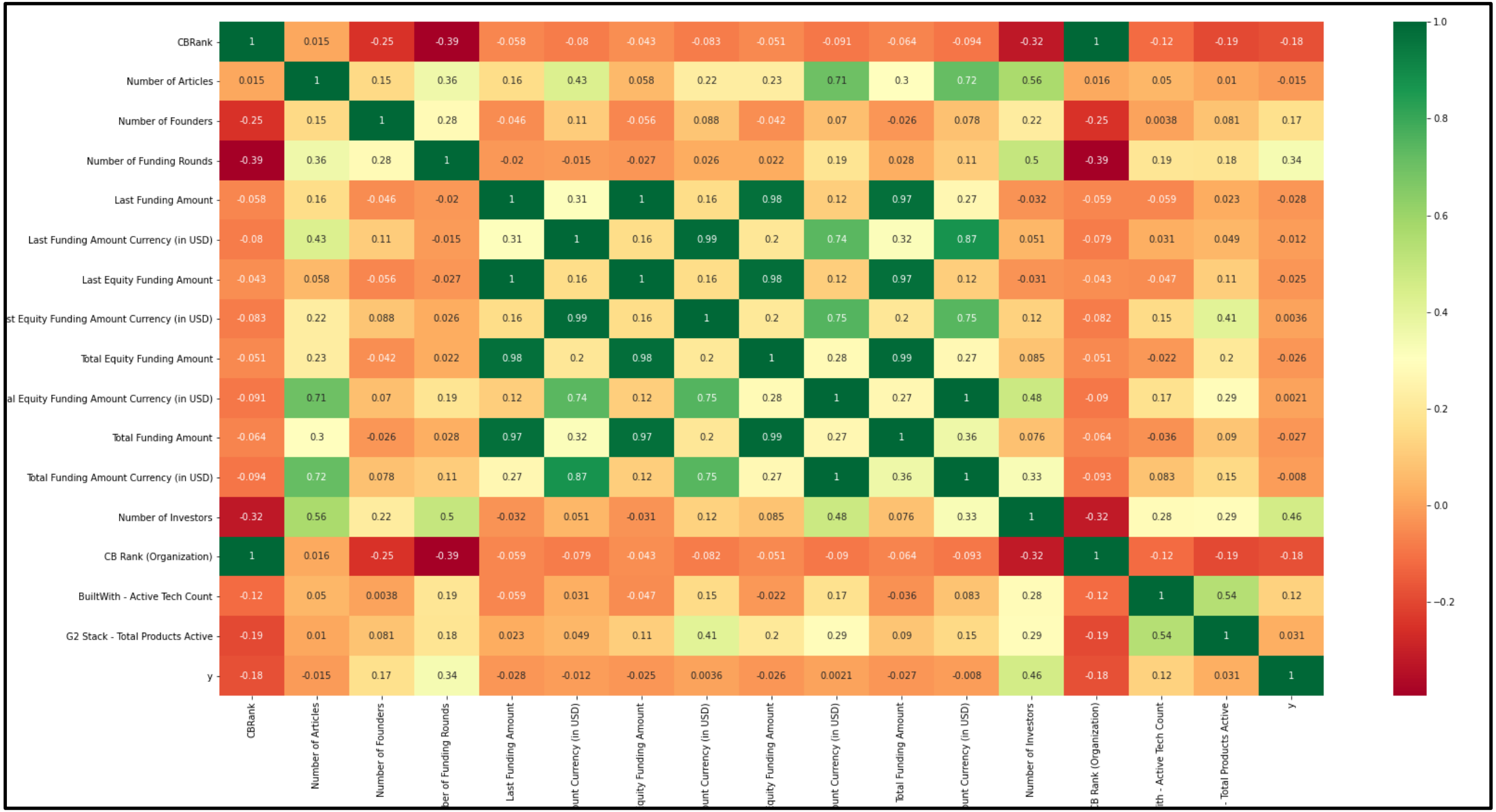
El cual realiza las intersecciones con el código de Miller Quiroga y realiza limpieza de base de datos crunchbase. La variable objetivo a modelar que contiene las intersecciones es “y”

La base de datos final para trabajar es el dataframe **df2**

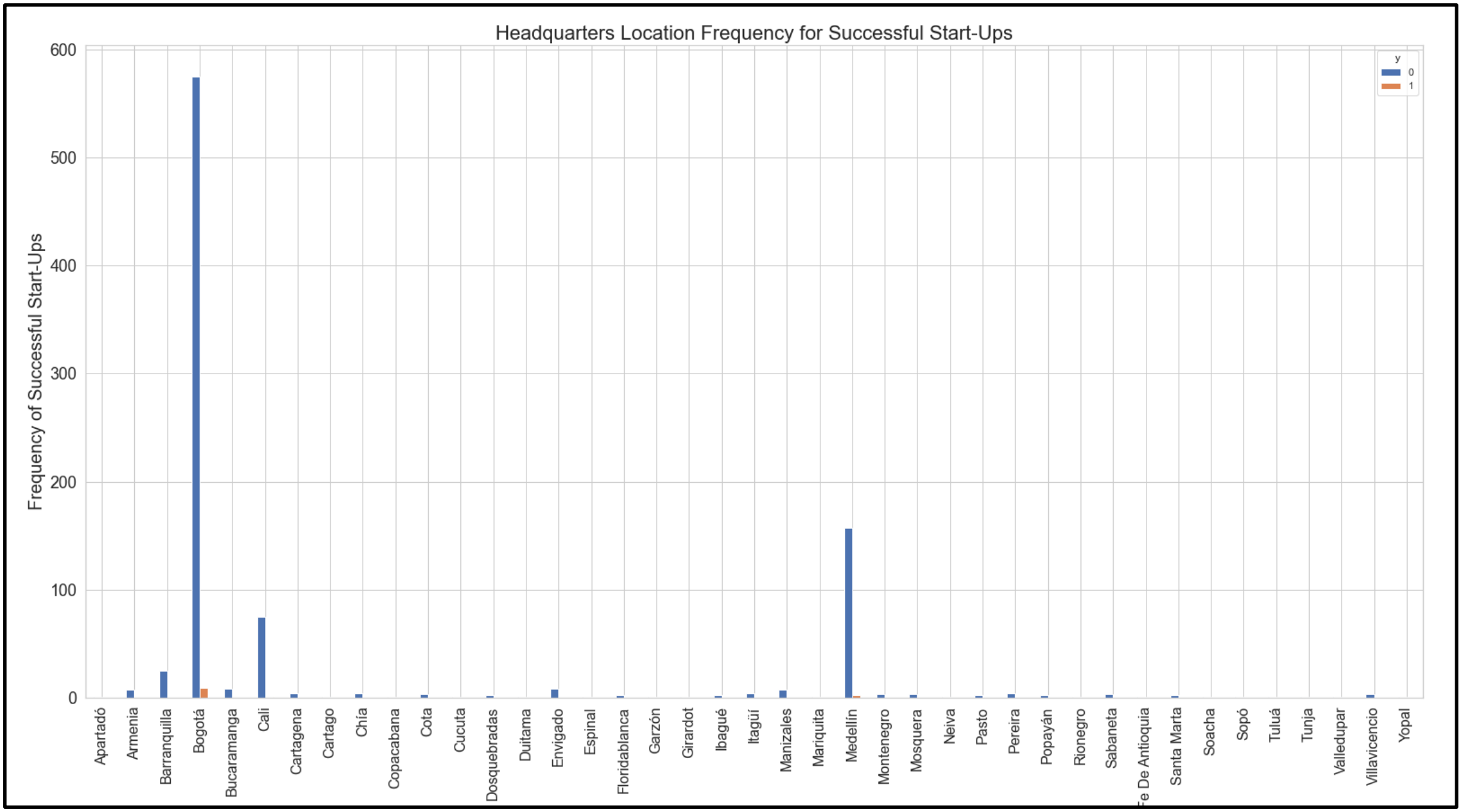
Matriz de Correlación, sin filtro de columnas con mas 80% de nan



Matriz de Correlación, con filtro de columnas con mas 80% de nan

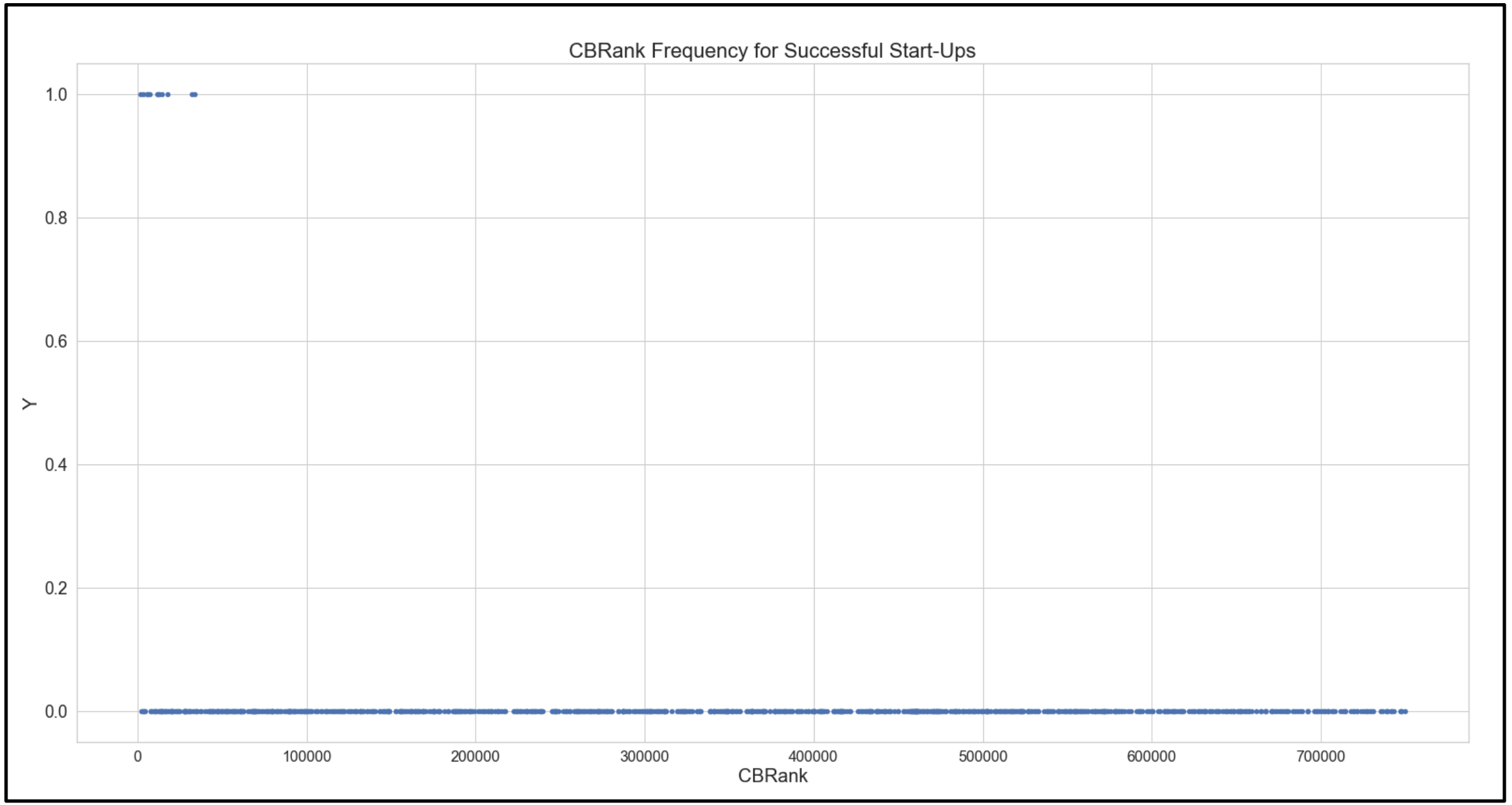


Headquarters Location Frequency for Successful Start-Ups



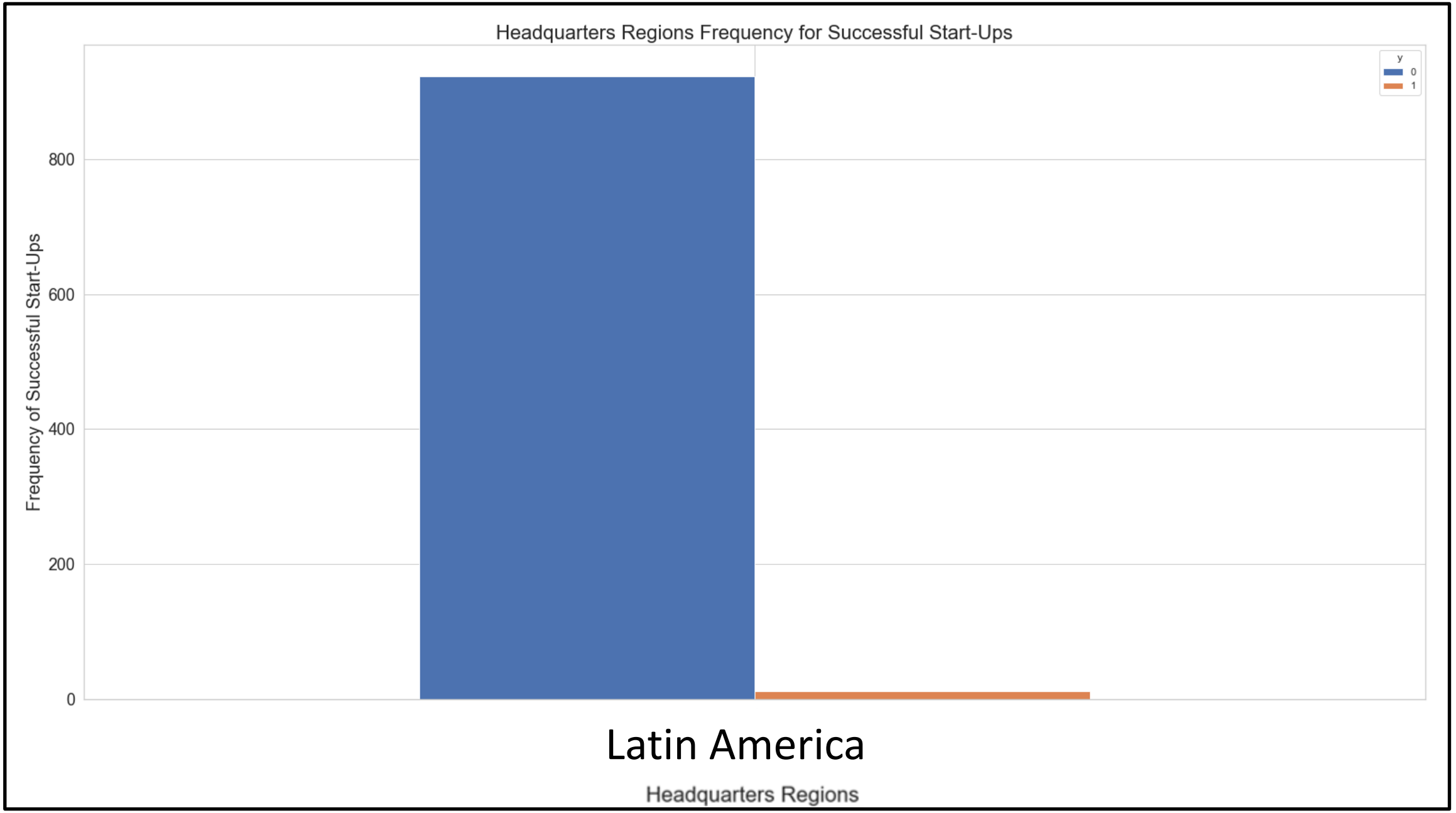
Si Importa la localización de la startup para su éxito. Bogotá y Medellín tienen las Start-ups con éxito, sin embargo también se puede eliminar para evitar sesgo de que sólo hay éxito en Bogotá y Medellín

CBRank vs Successful Start-Ups



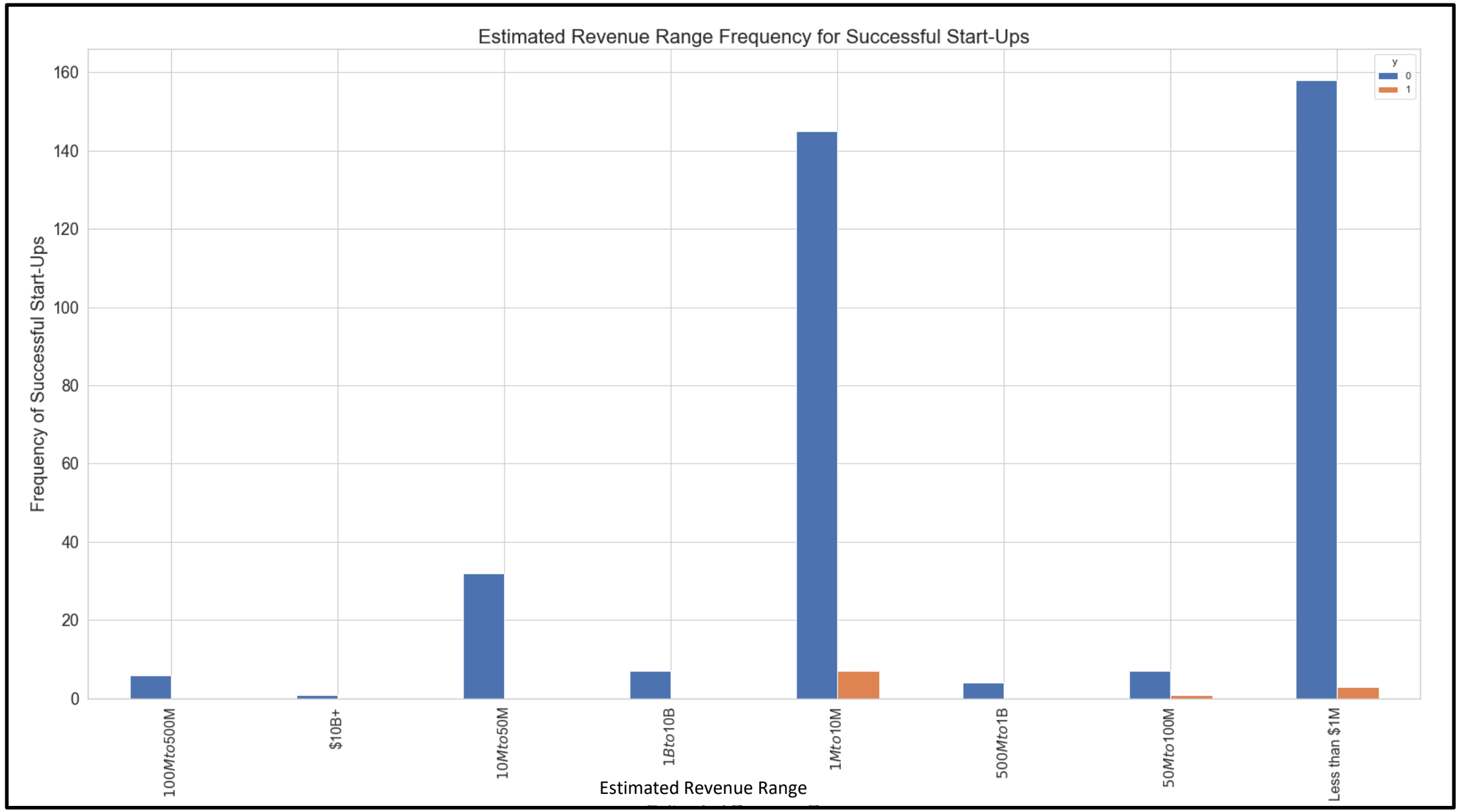
Si/No Importa Cbrank para éxito de Startups. (Tengo duda)

Headquarters Regions vs Successful Start-Ups



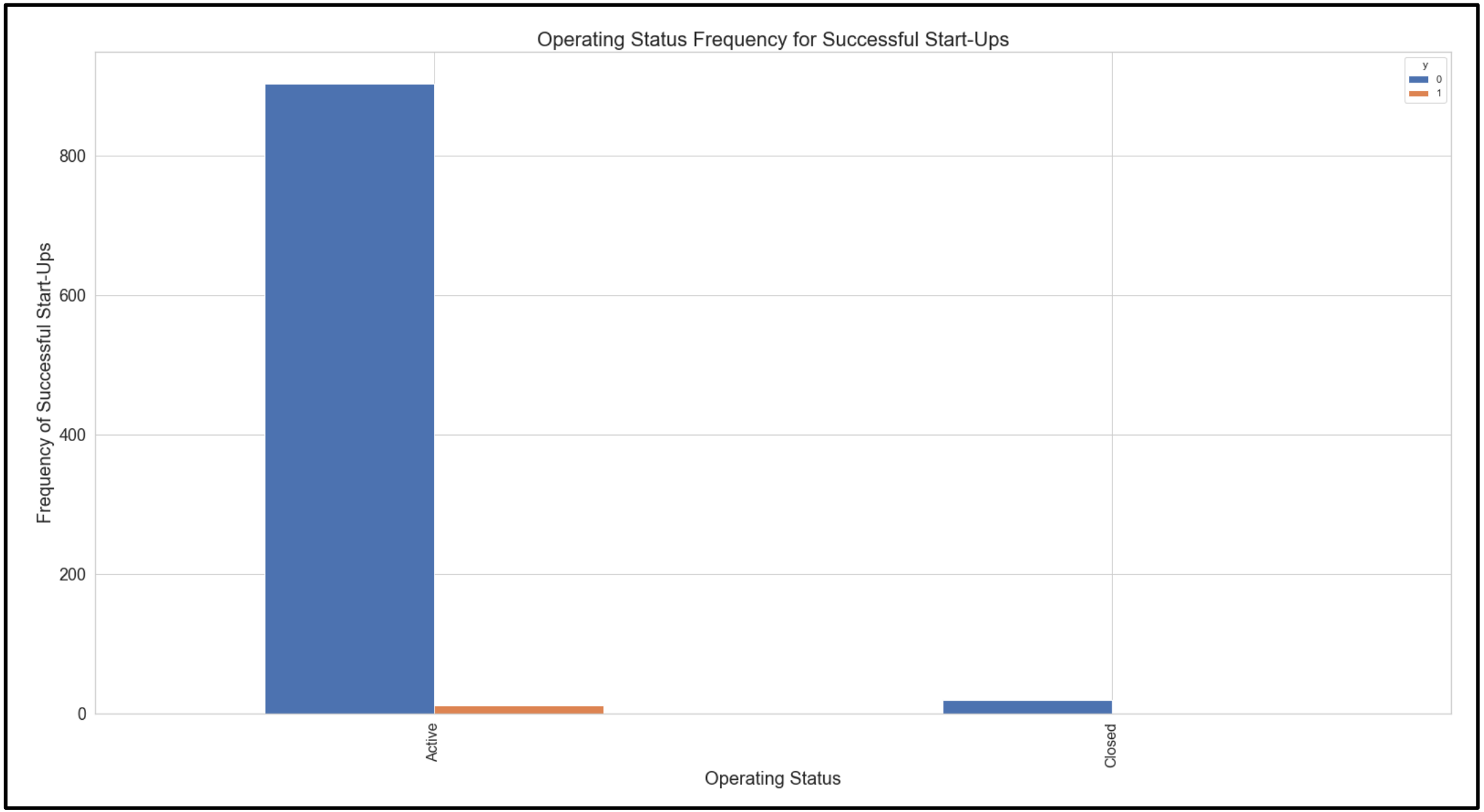
No importa Headquarters Regions para éxito de Startups

Estimated Revenue Range vs Successful Start-Ups



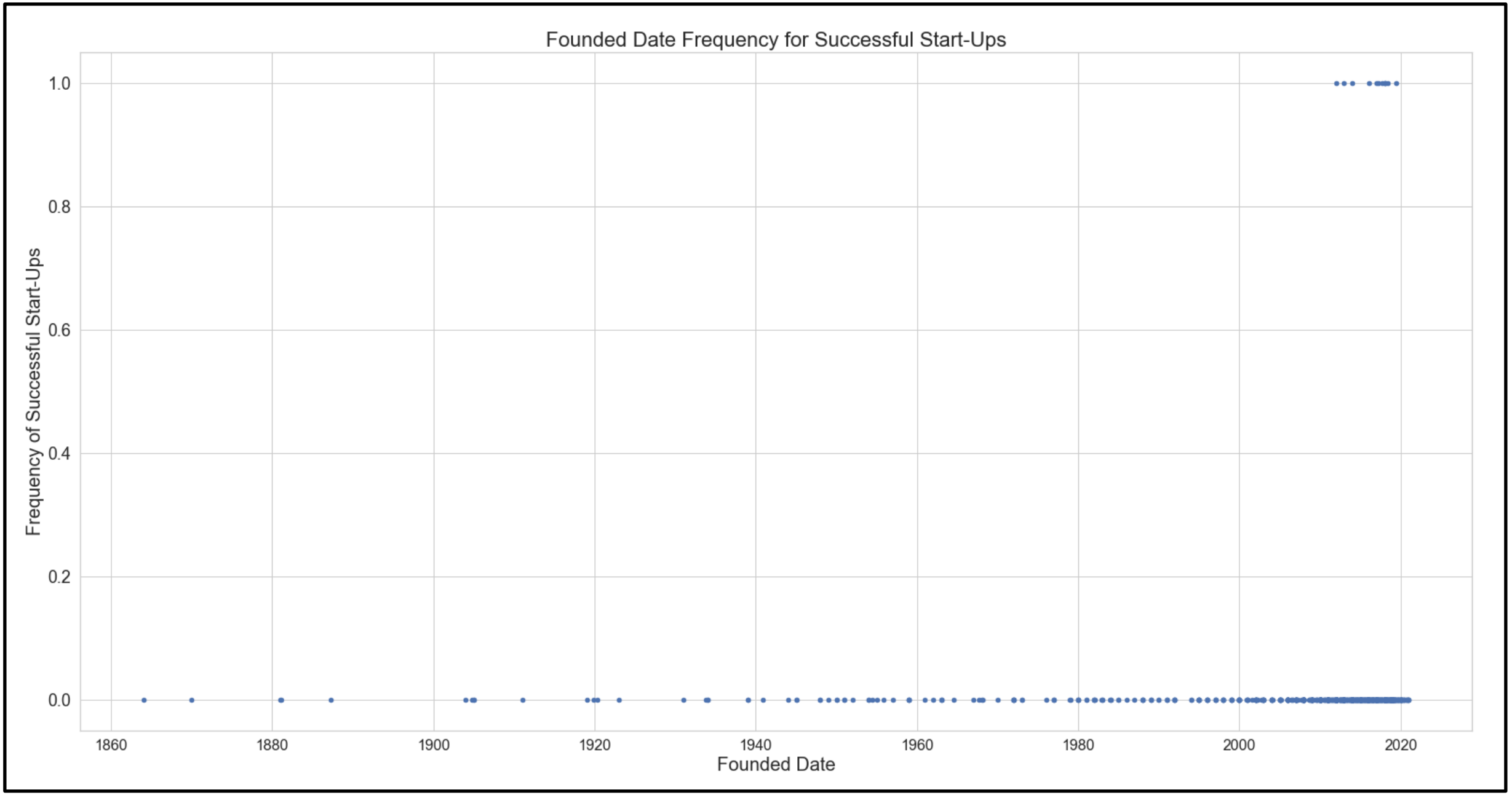
Si Importa Estimated Revenue Range para éxito de Startups.

Operating Status vs Successful Start-Ups



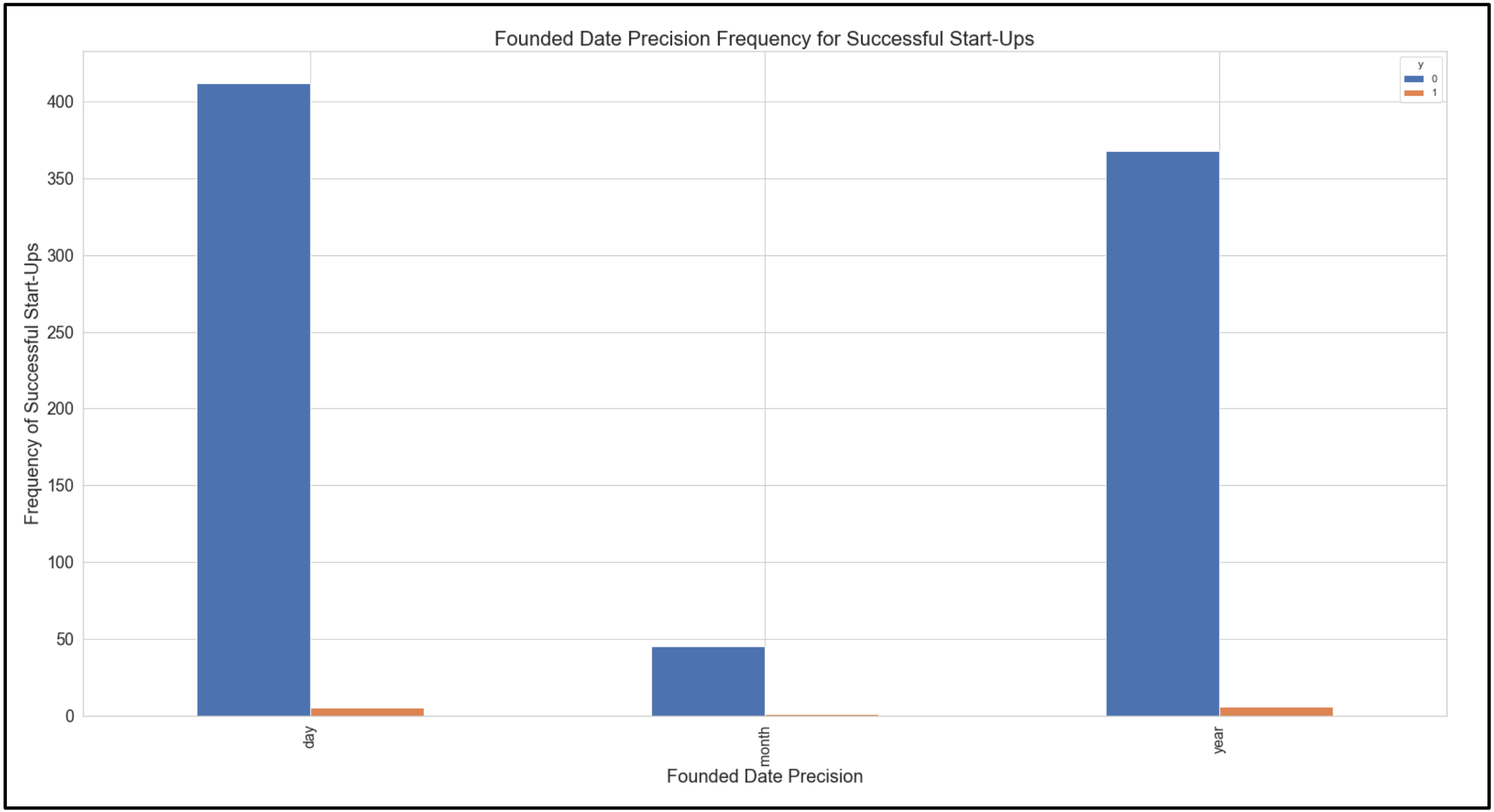
No importa Operating Status para éxito de Startups

Founded Date vs Successful Start-Ups



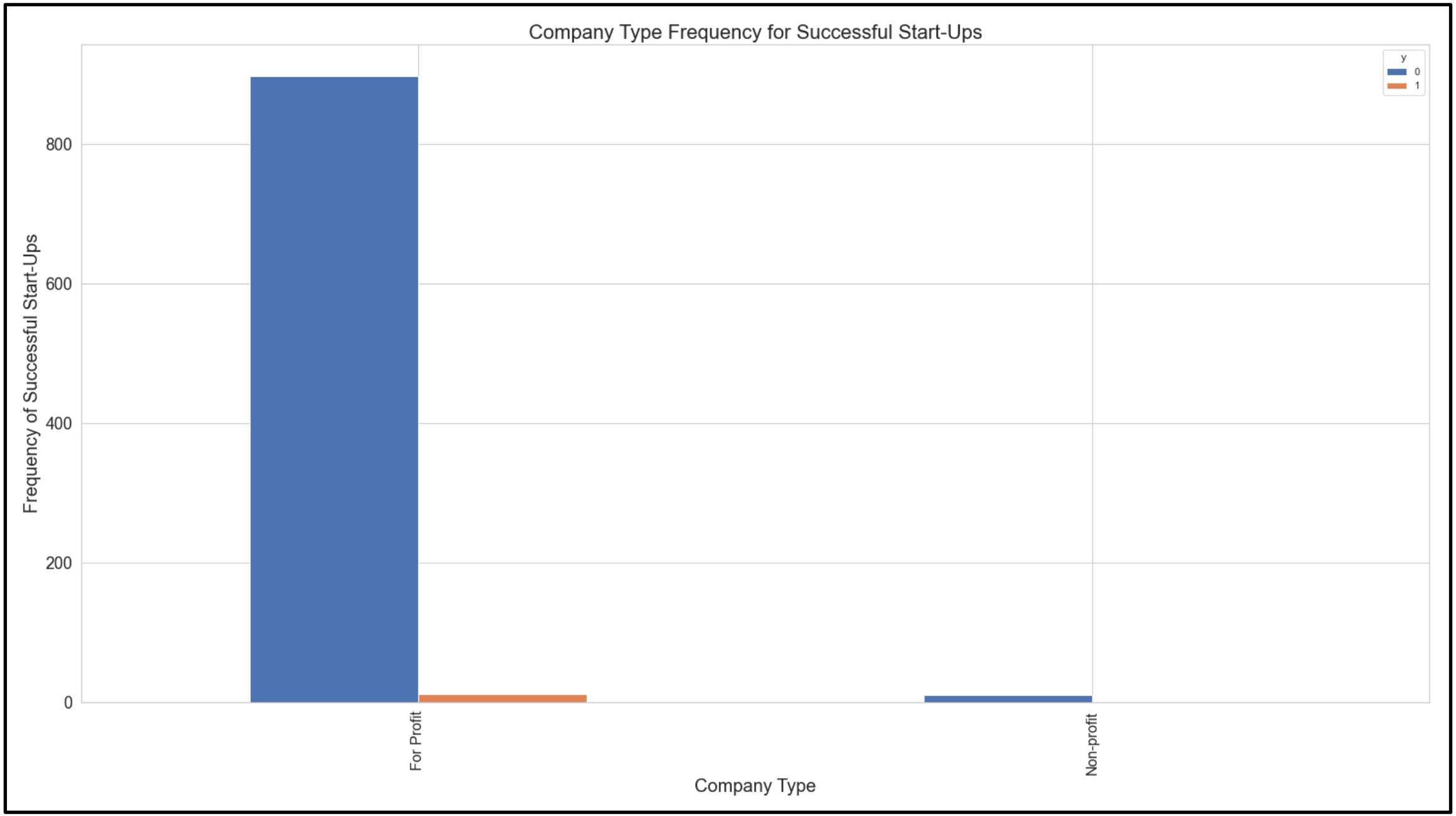
No importa Founded Date para éxito de Startups

Founded Date Precision vs Successful Start-Ups



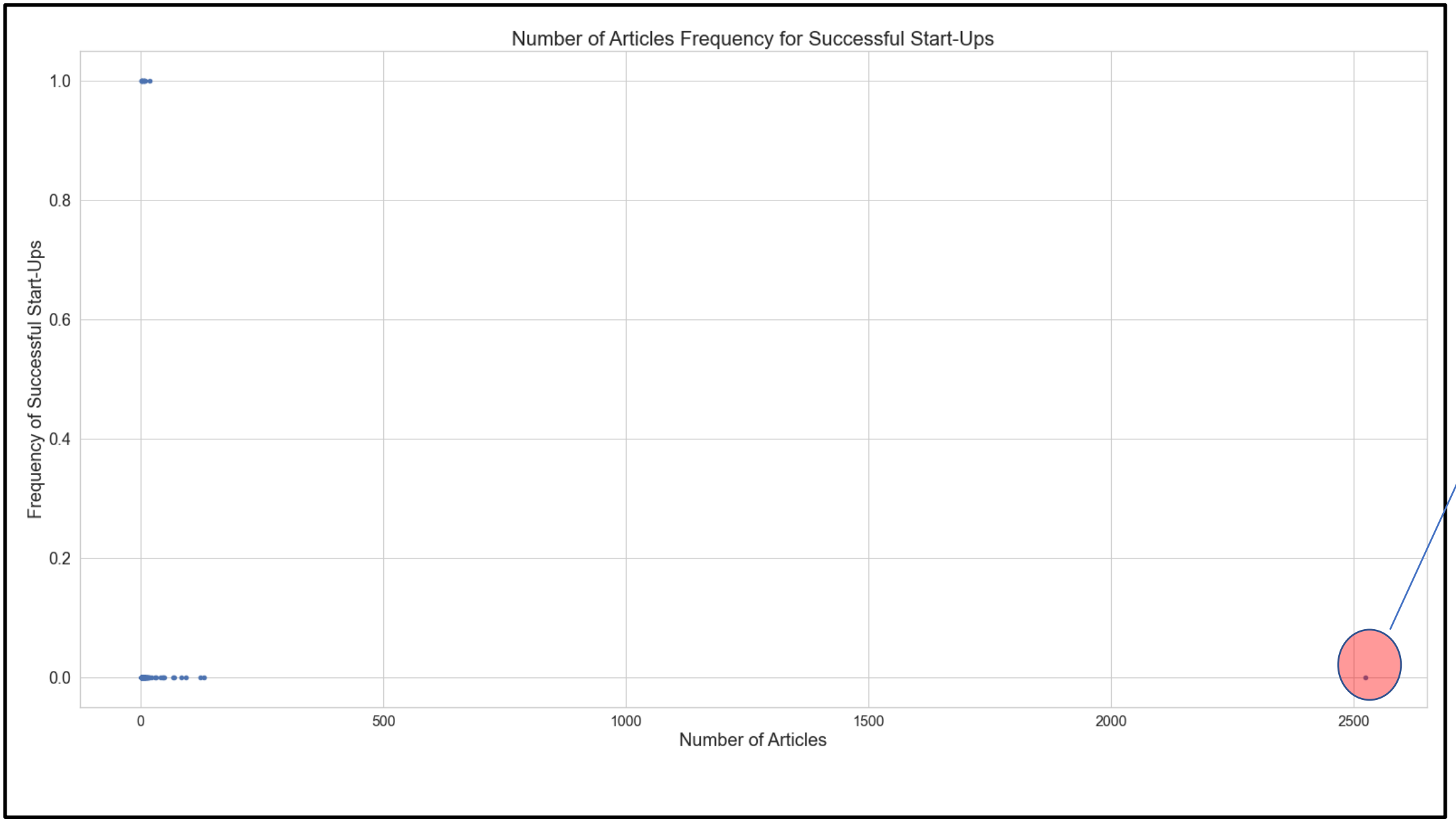
Parece importar Founded Date Precision para éxito de Startups, sin embargo no entiendo significado de la variable. La saco

Company Type vs Successful Start-Ups



No importa Company Type para éxito de Startups

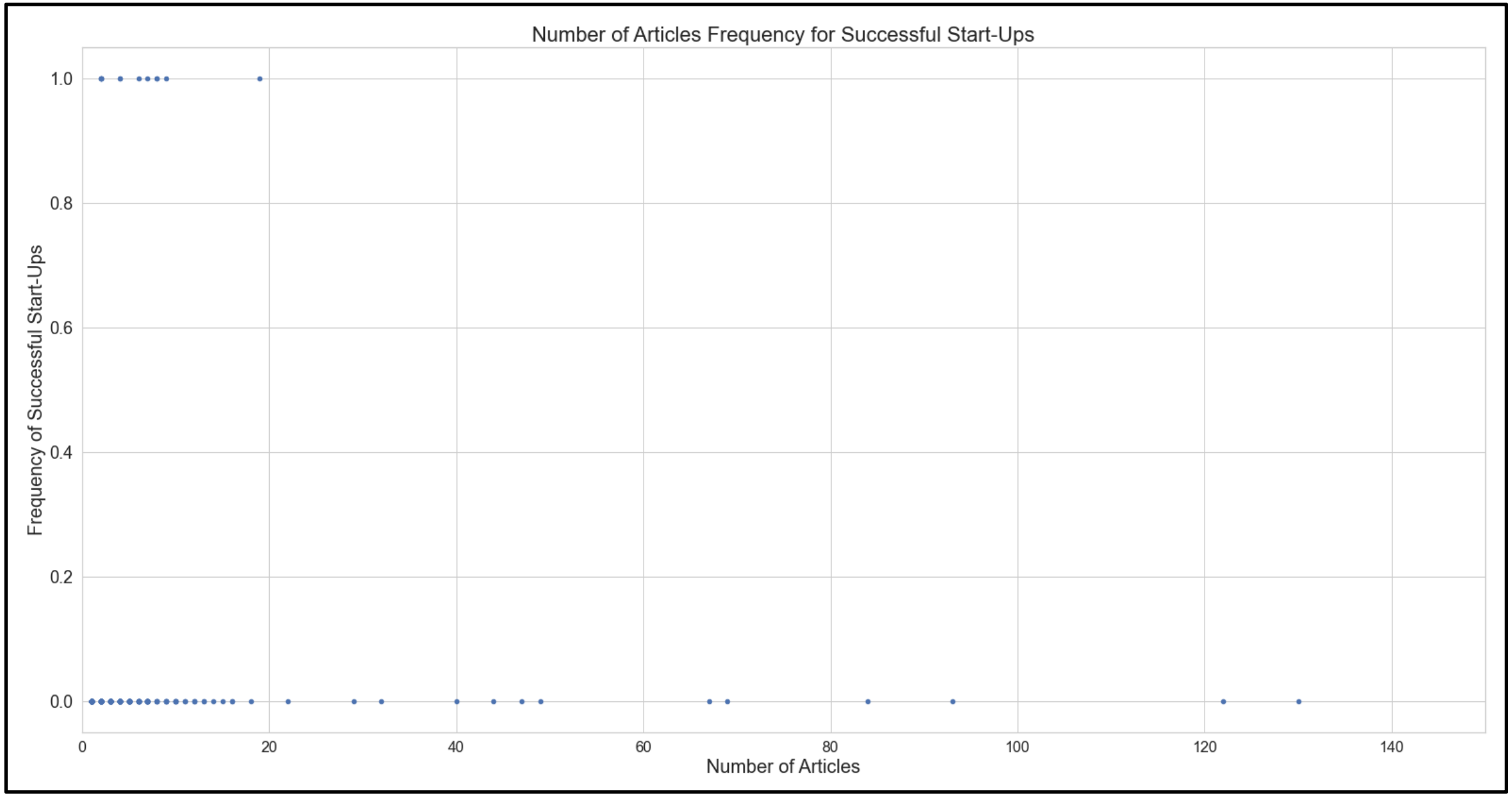
Number of Articles vs Successful Start-Ups



Outlier
o
Dato mal
Introducido
En BD

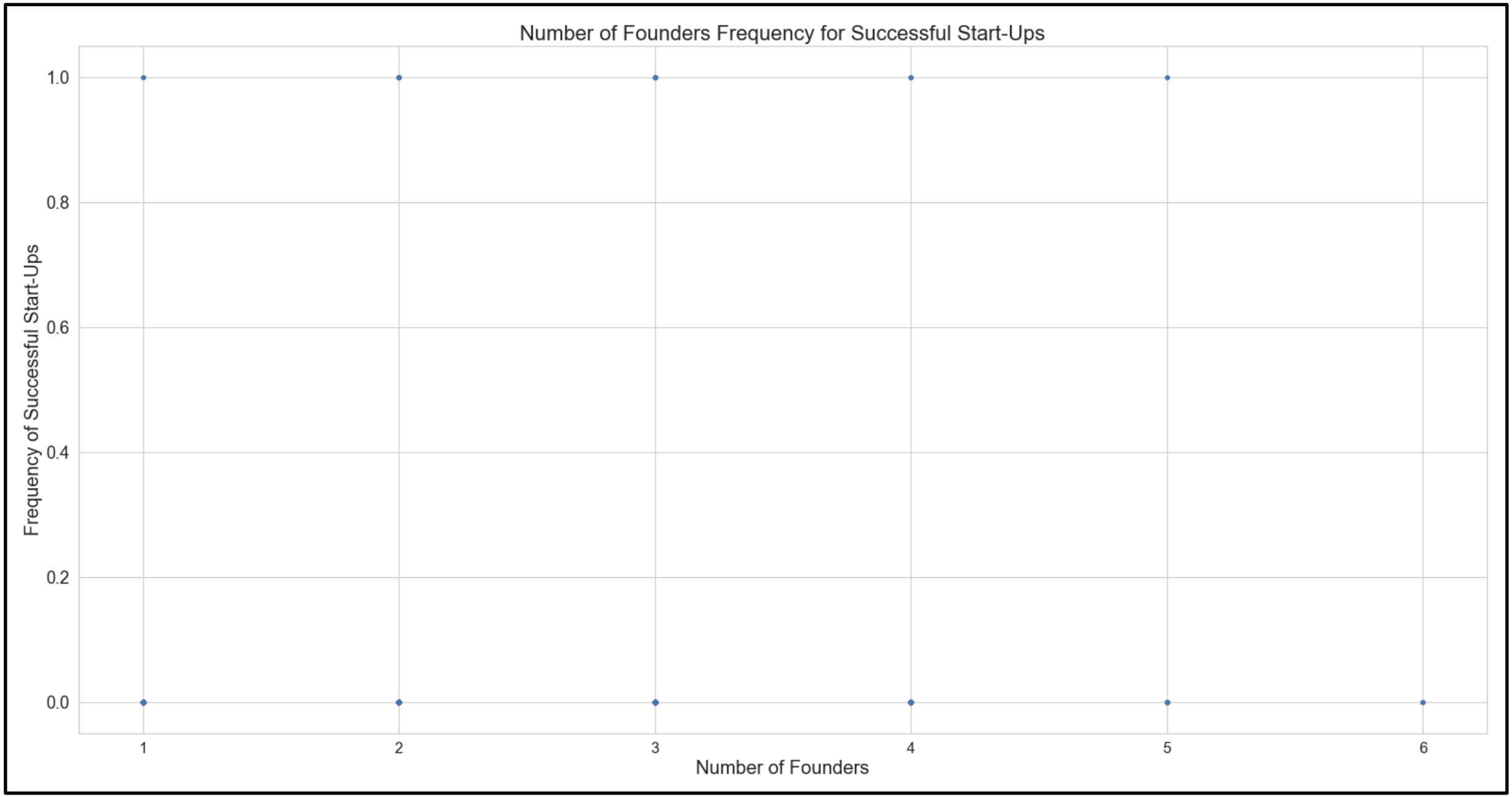
Si importa Number of Articles para éxito de Startups

Number of Articles vs Successful Start-Ups **PLOT ZOOM**



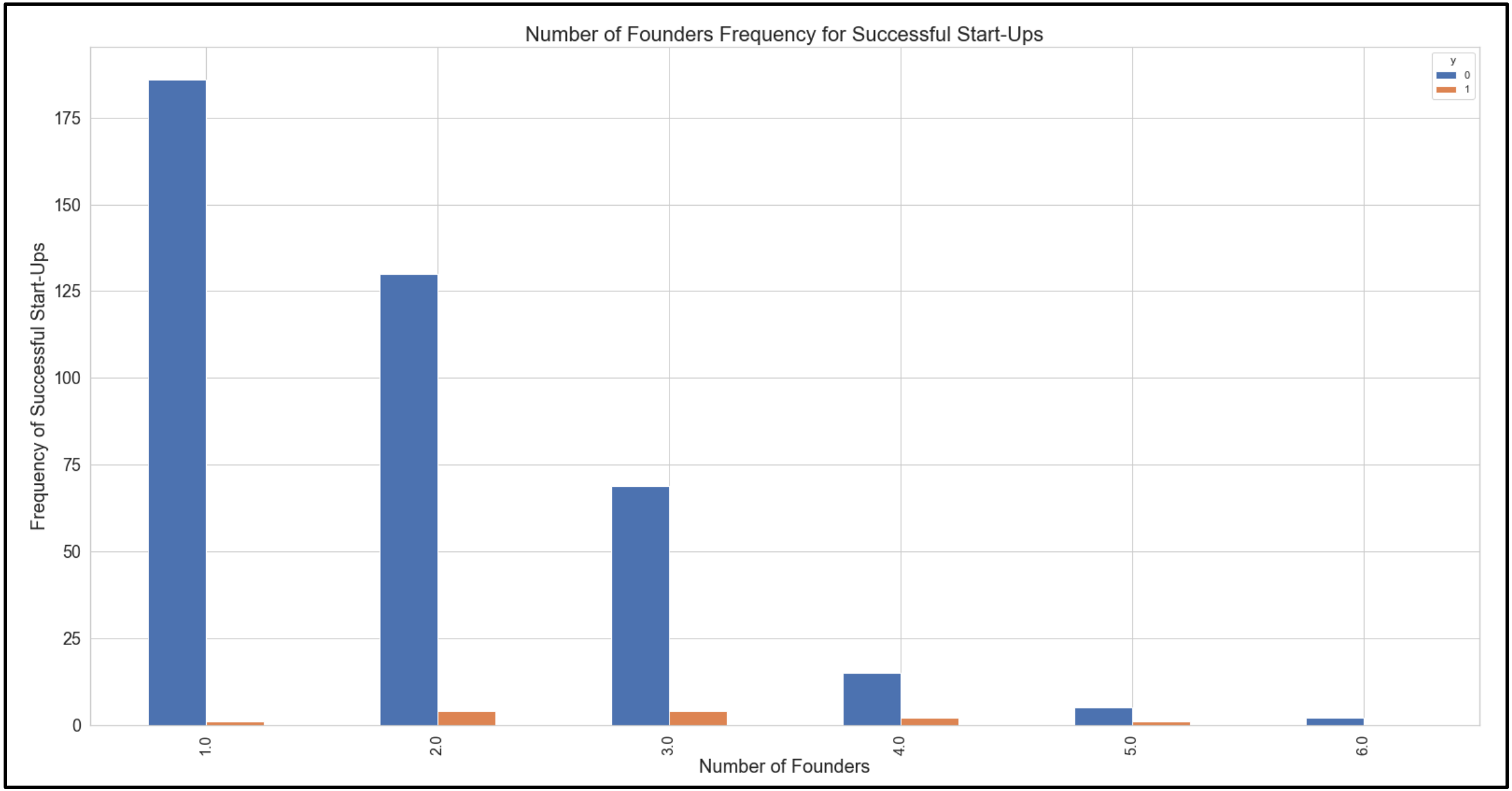
Si importa Number of Articles para éxito de Startups

Number of Founders vs Successful Start-Ups (Variable Númerica)



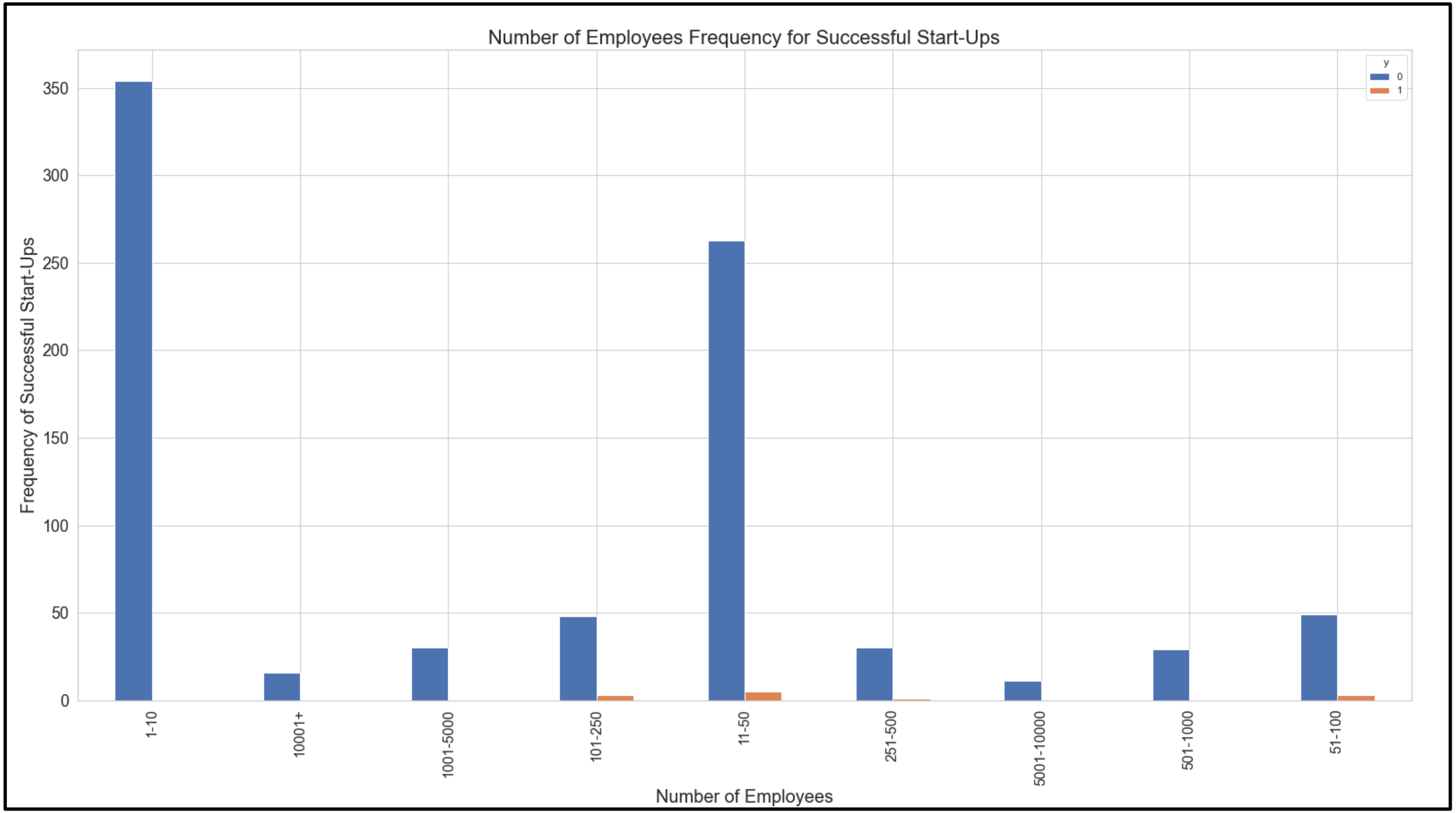
Si importa Number of Founders para éxito de Startups

Number of Founders vs Successful Start-Ups (Variable Categórica)



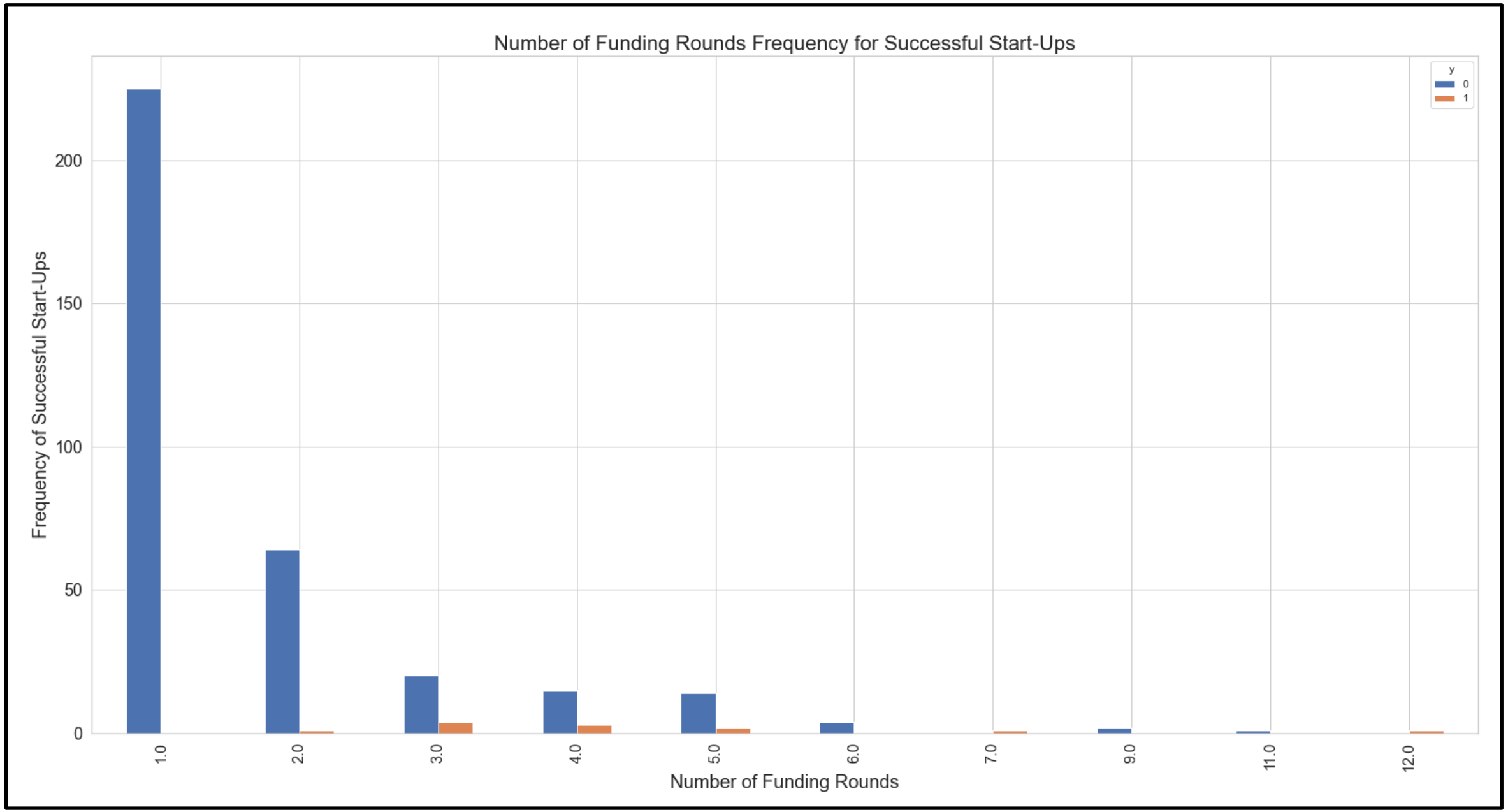
Si importa Number of Founders para éxito de Startups

Number of Employees vs Successful Start-Ups



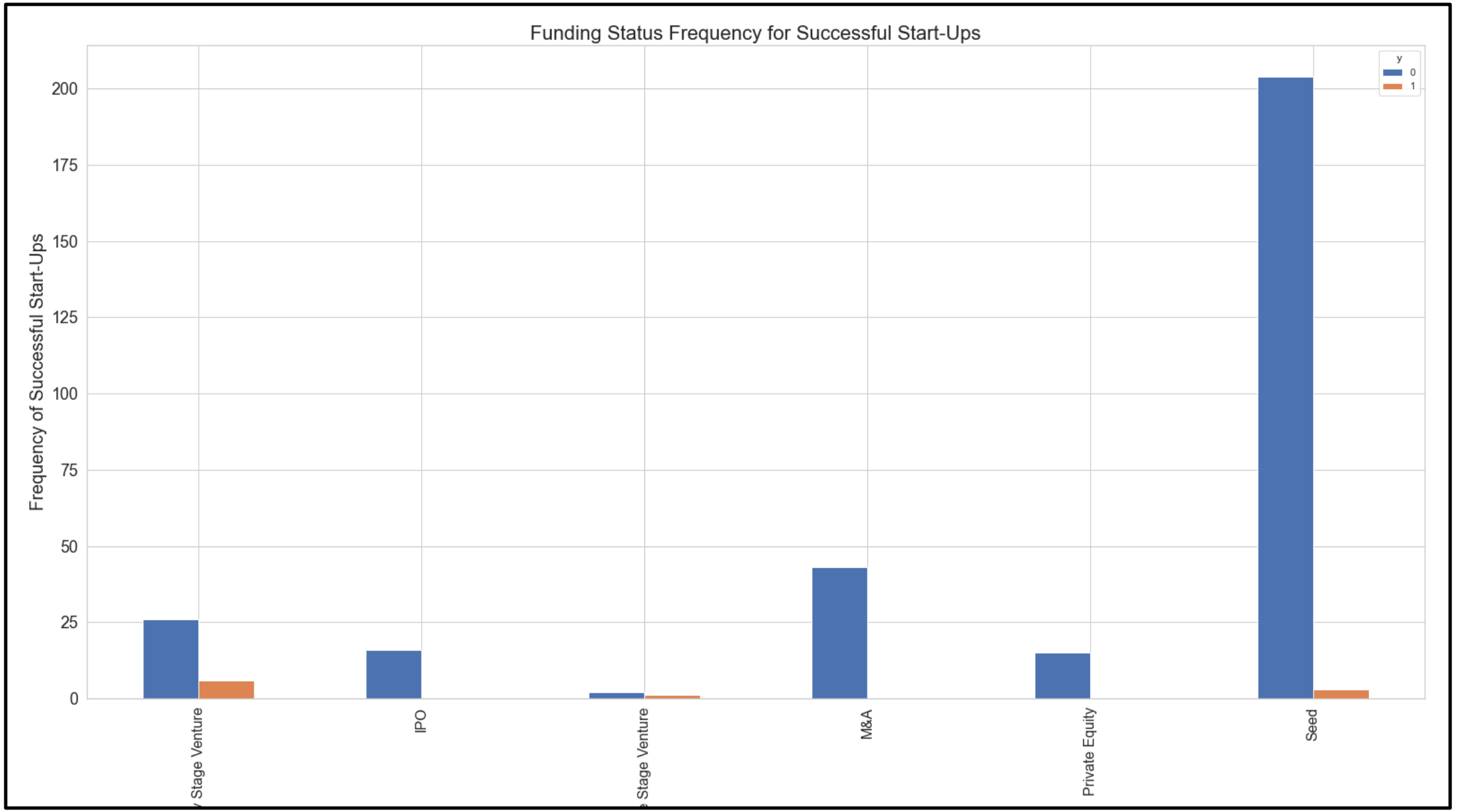
Si importa Number of Employees para éxito de Startups

Number of Funding Rounds vs Successful Start-Ups



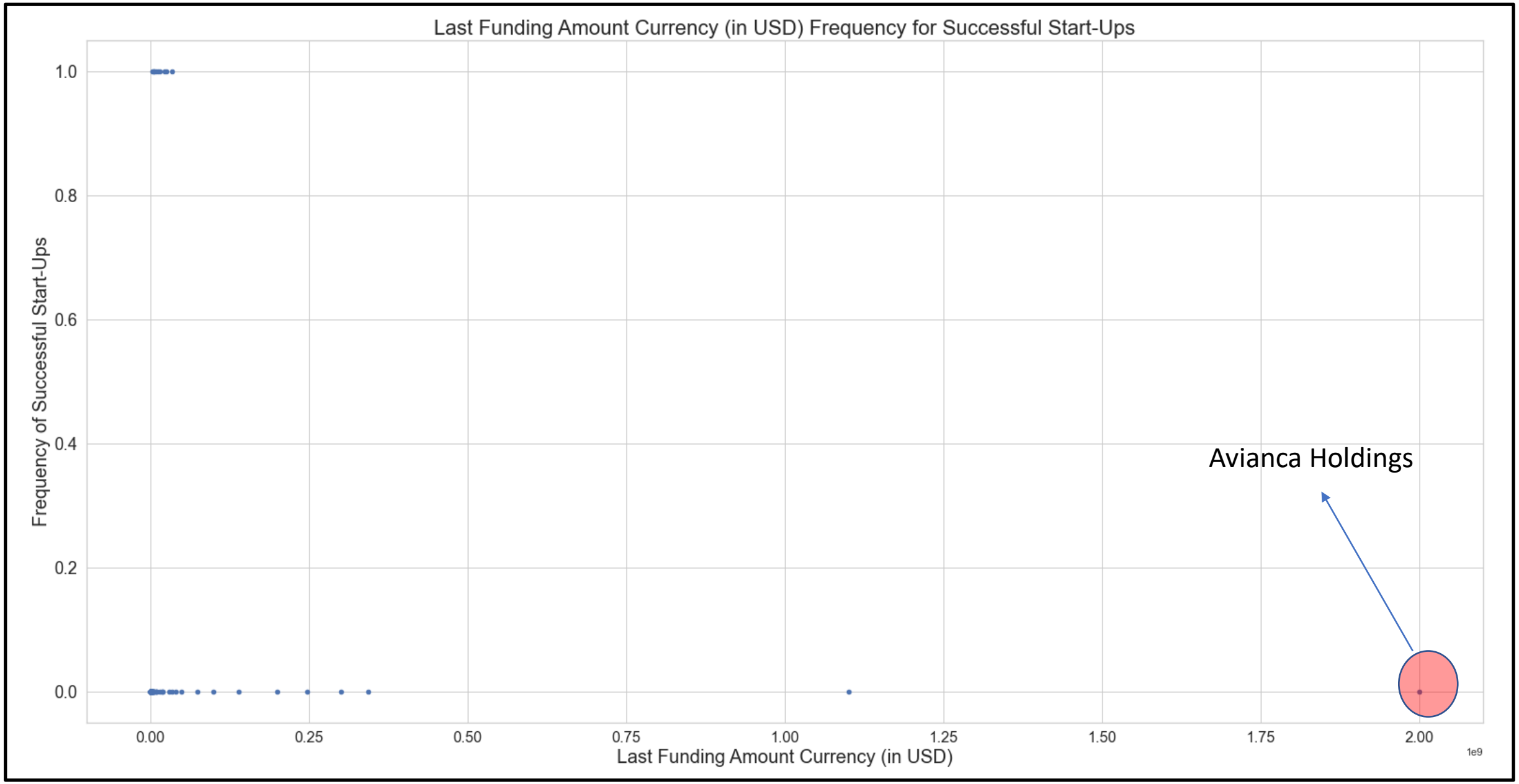
Si importa Number of Funding Rounds para éxito de Startups

Funding Status vs Successful Start-Ups



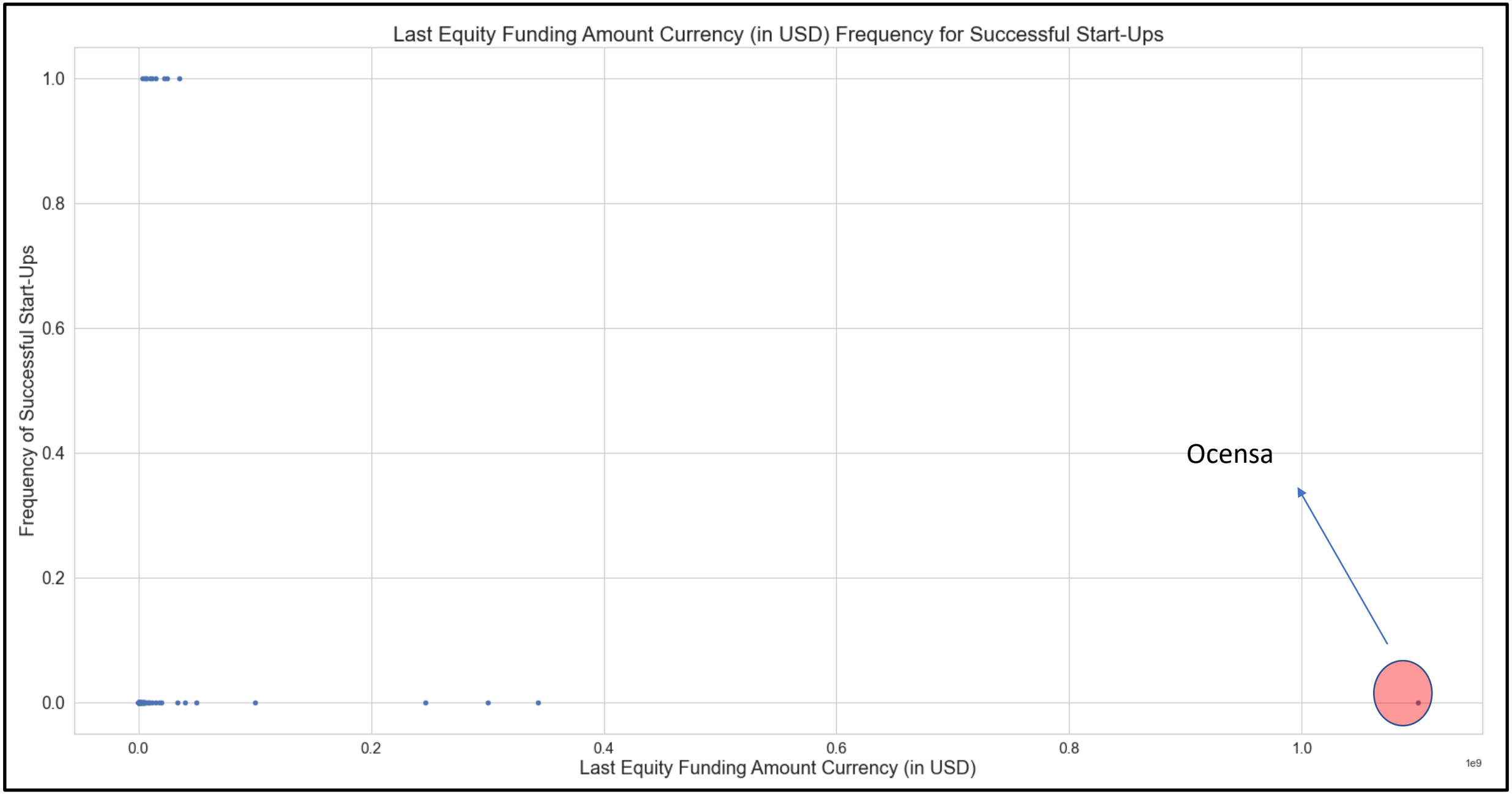
Si importa Funding Status para éxito de Startups

Last Funding Amount Currency (in USD) vs Successful Start-Ups



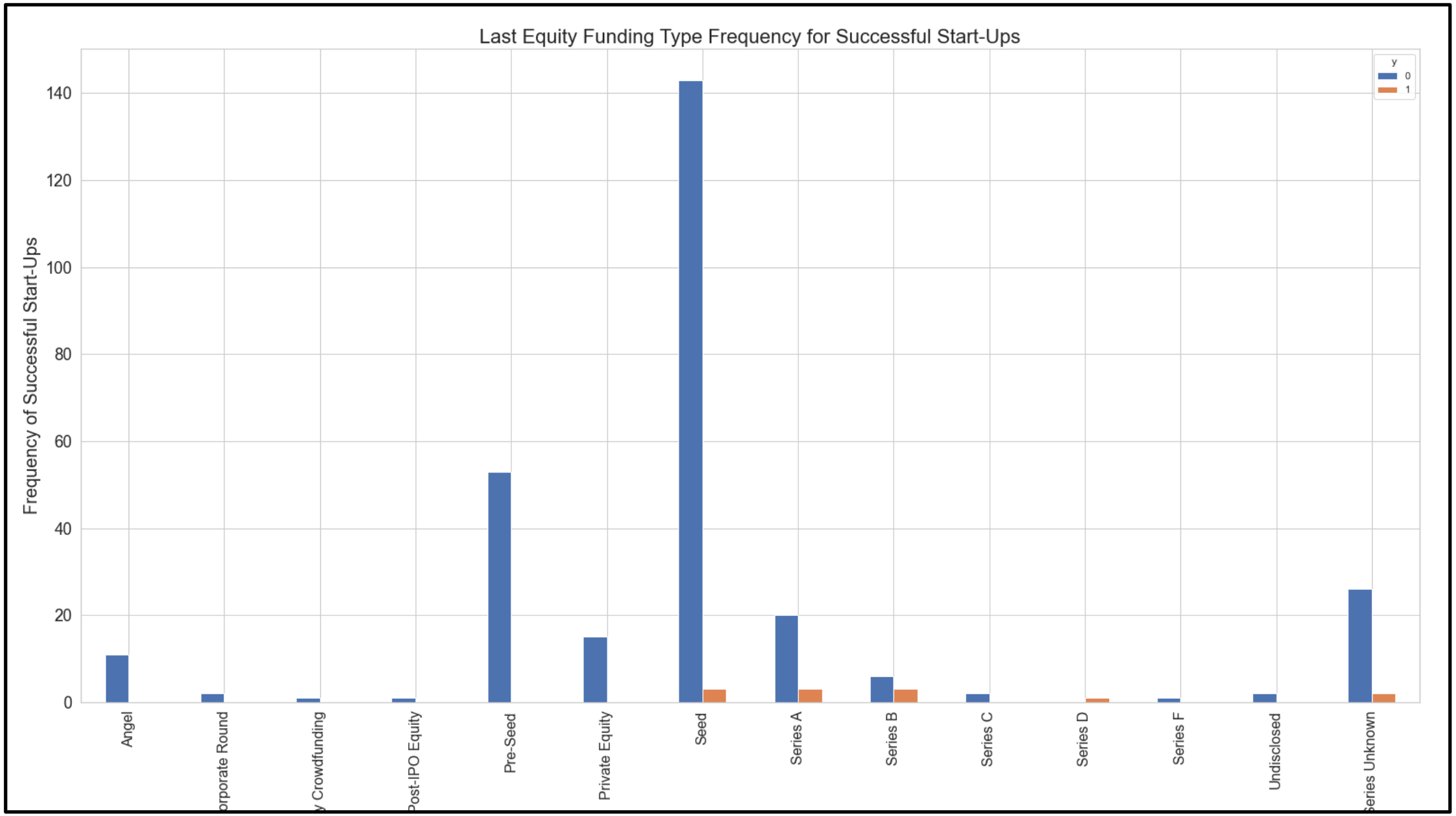
Si importa Last Funding Amount Currency (in USD) para éxito de Startups

Last Equity Funding Amount Currency (in USD) vs Successful Start-Ups



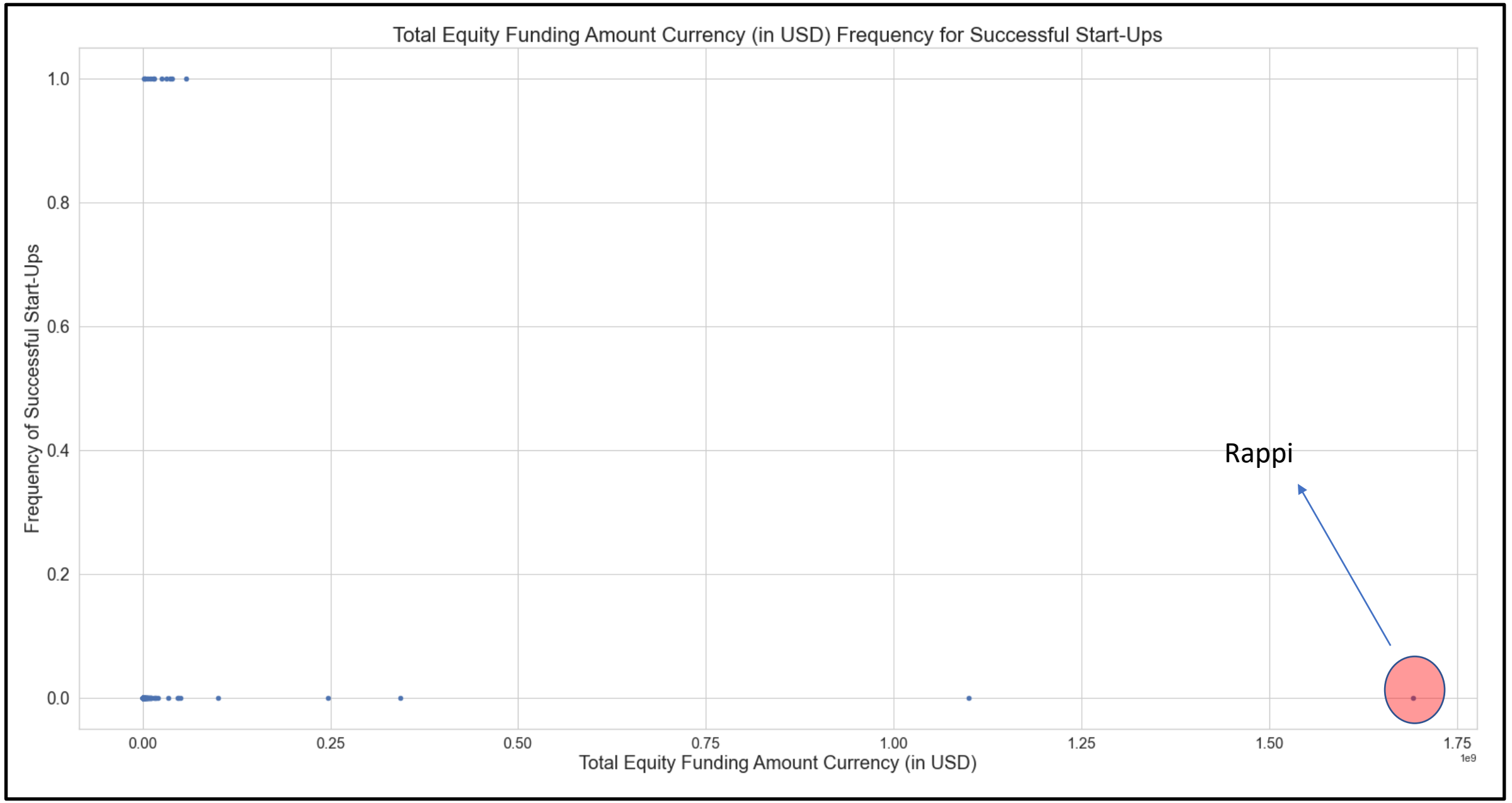
Si importa Last Equity Funding Amount Currency (in USD) para éxito de Startups

Last Equity Funding Type vs Successful Start-Ups



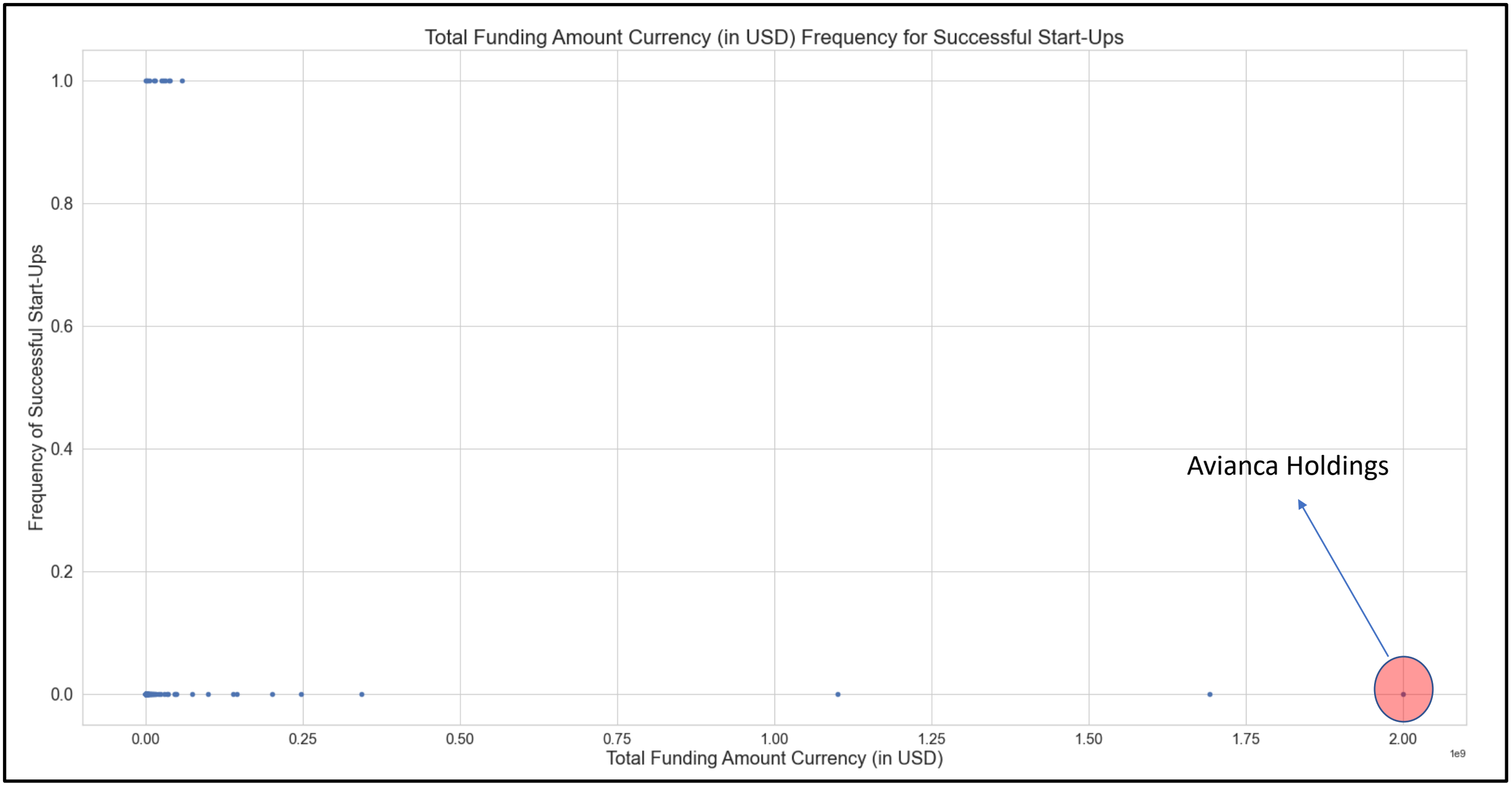
Si importa Last Equity Funding Type para éxito de Startups

Total Equity Funding Amount Currency (in USD) vs Successful Start-Ups



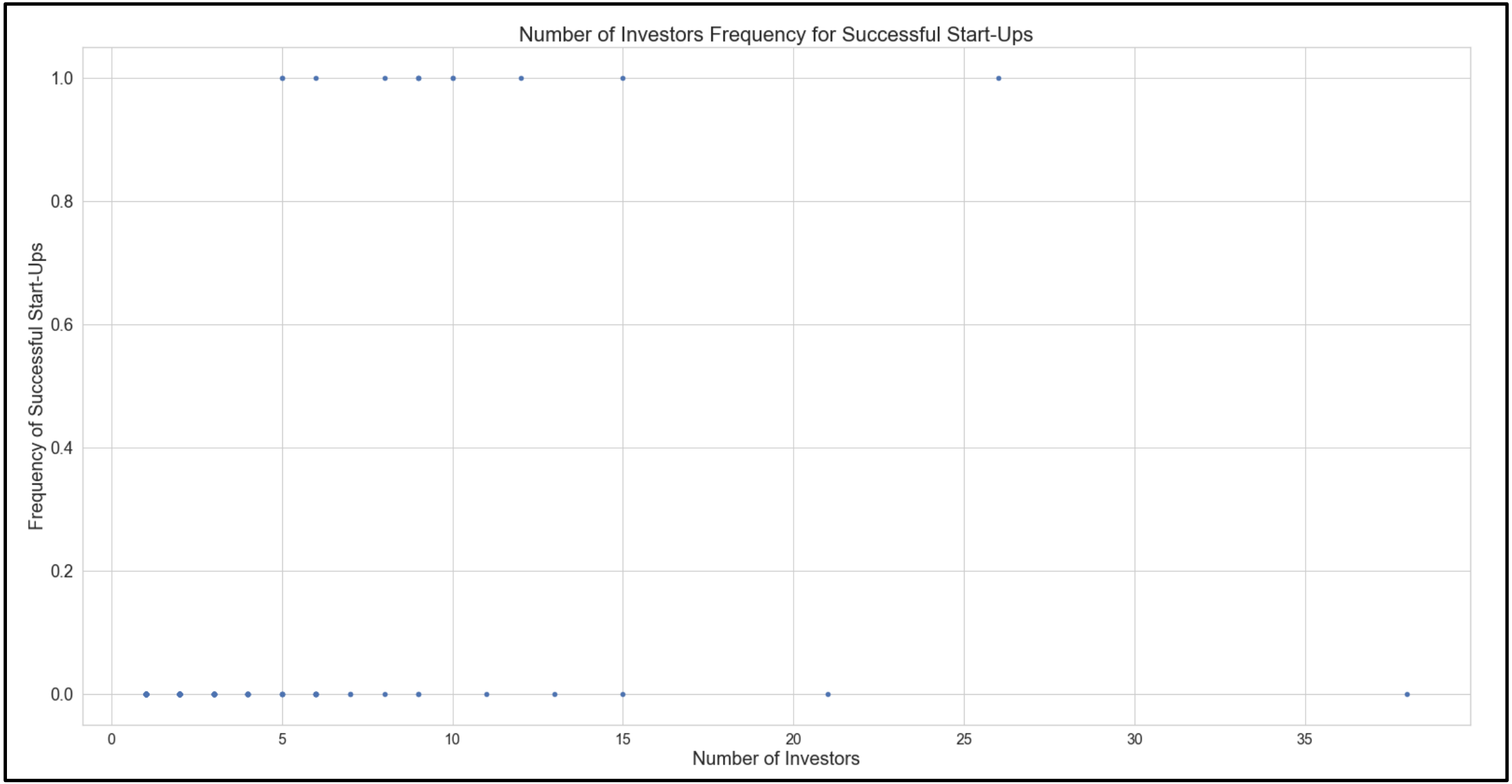
Si importa Total Equity Funding Amount Currency (in USD) para éxito de Startups

Total Funding Amount Currency (in USD) vs Successful Start-Ups



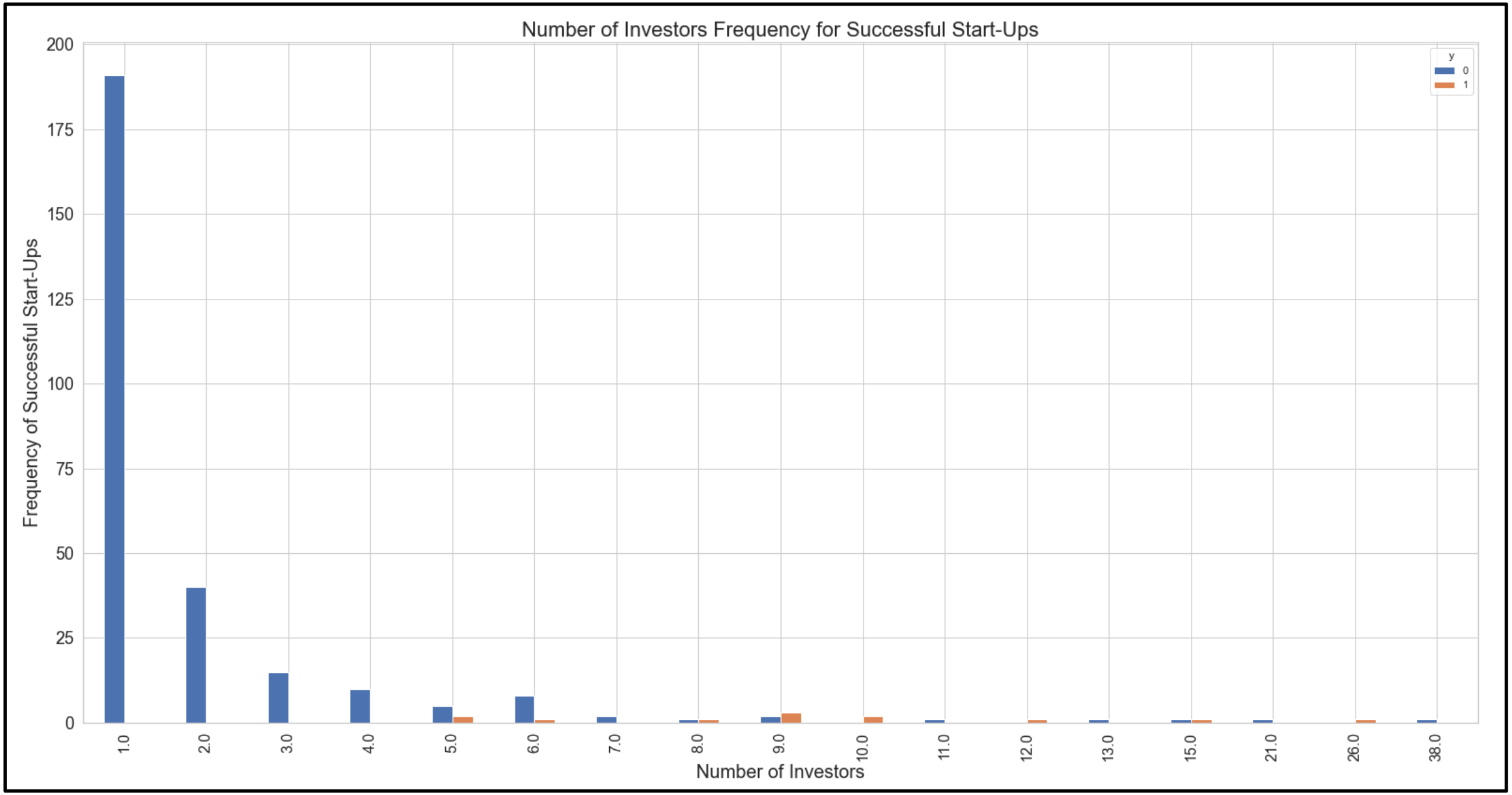
Si importa Total Funding Amount Currency (in USD) para éxito de Startups

Number of Investors vs Successful Start-Ups (Númerica)



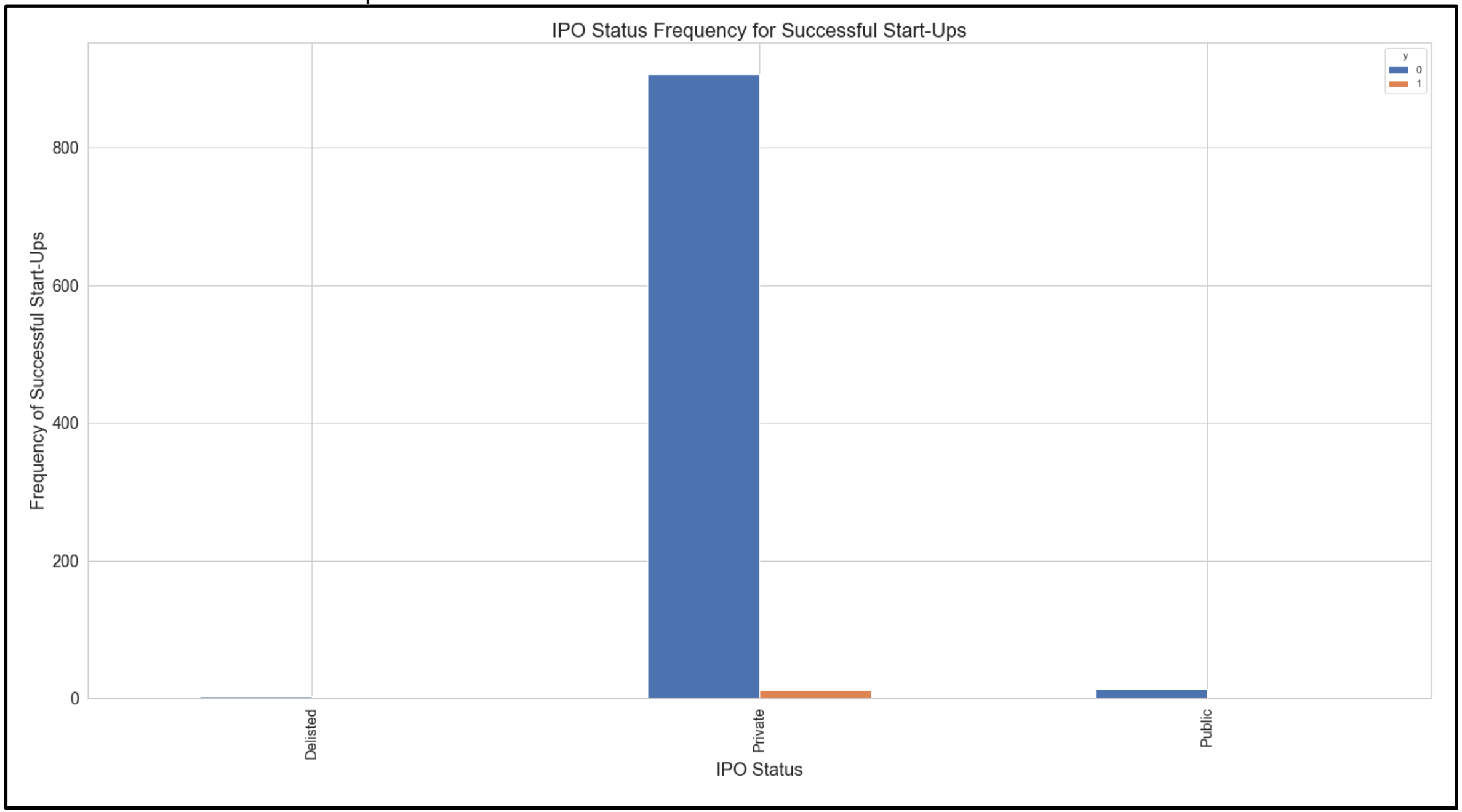
Si importa Number of Investors para éxito de Startups

Number of Investors vs Successful Start-Ups (Categórica)



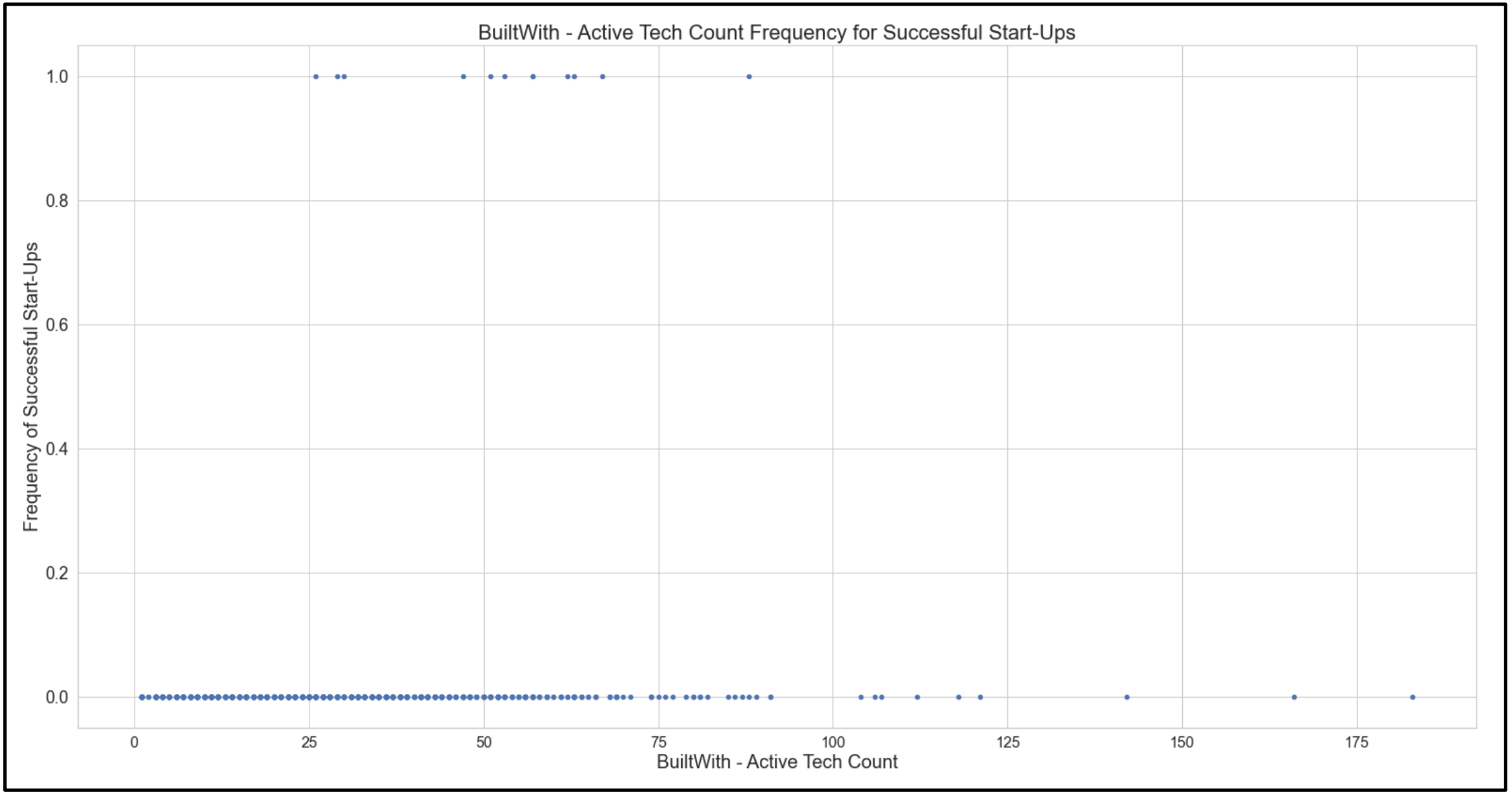
Si importa Number of Investors para éxito de Startups

IPO Status vs Successful Start-Ups



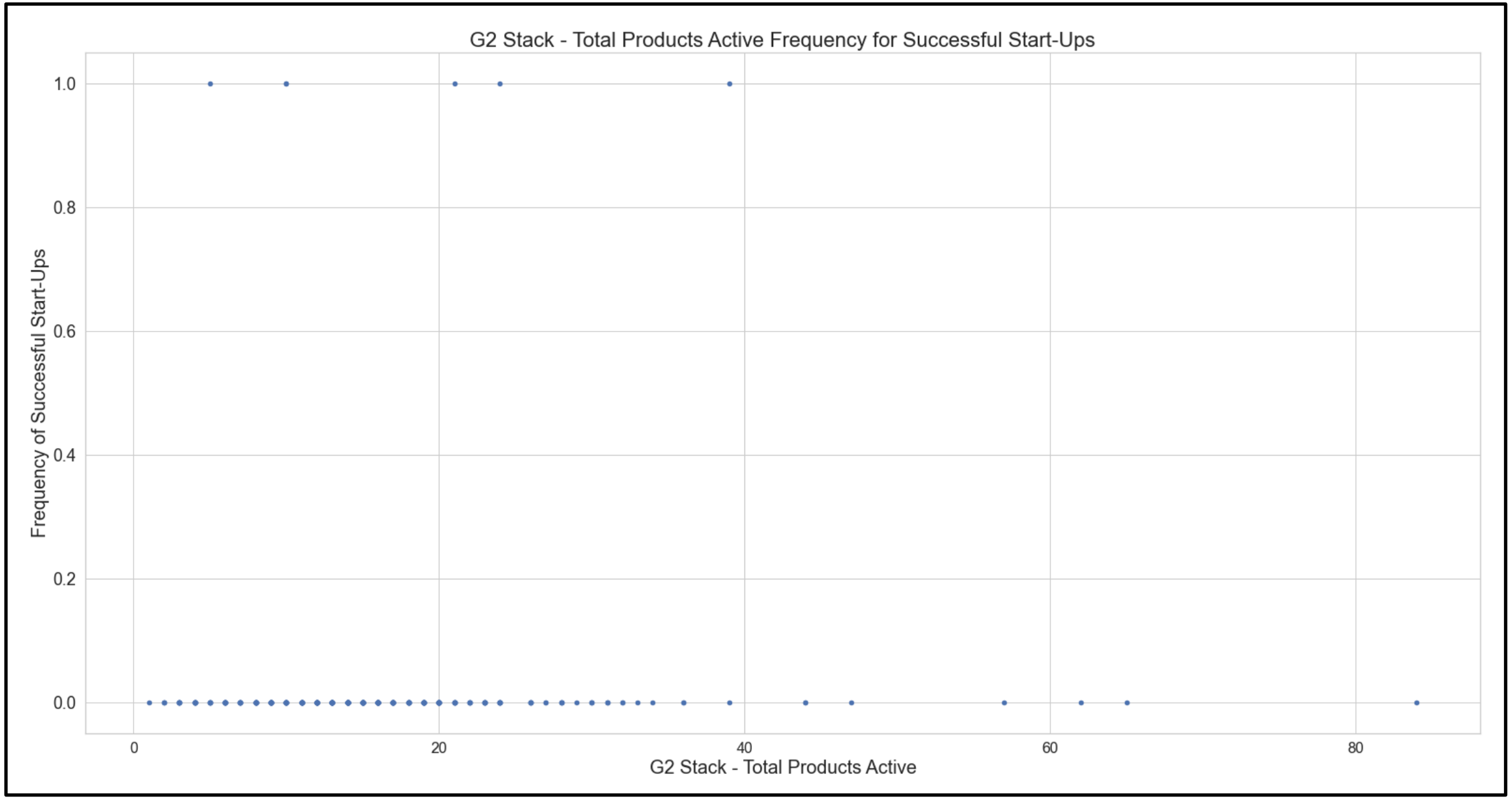
No importa IPO Status para éxito de Startups

BuiltWith - Active Tech Count vs Successful Start-Ups (Categórica)

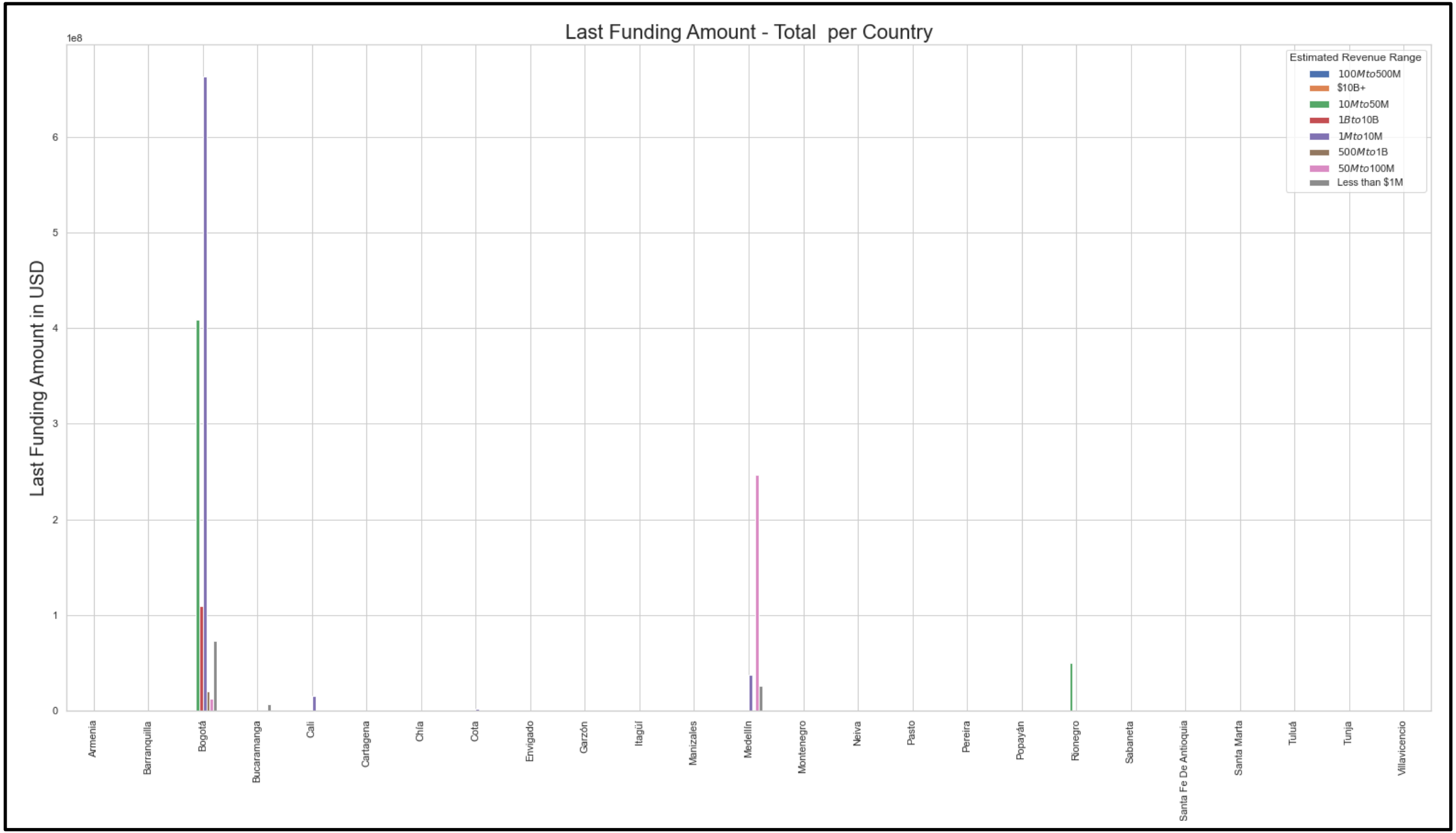


Si importa BuiltWith - Active Tech Count para éxito de Startups

G2 Stack - Total Products Active vs Successful Start-Ups



Si importa G2 Stack - Total Products Active para éxito de Startups


Last Funding Amount - Total per Country
1e8
Last Funding Amount in USD
0
1
2
3
4
5
6
Estimated Revenue Range
100Mto500M
$10B+
10Mto50M
1Bto10B
1Mto10M
500Mto1B
50Mto100M
Less than $1M
Armenia
Barranquilla
Bogotá
Bucaramanga
Cali
Cartagena
Chía
Cota
Envigado
Garzón
Itagüí
Manizales
Medellín
Montenegro
Neiva
Pasto
Pereira
Popayán
Rionegro
Sabaneta
Santa Fe De Antioquia
Santa Marta
Tuluá
Tunja
Villavicencio

Set of variables

CBRank	Númerica
Estimated Revenue Range	Categórica
Number of Articles	Númerica
Number of Founders	Númerica
Number of Employees	Categórica
Number of Funding Rounds	Númerica
Funding Status	Categórica
Last Funding Amount Currency (in USD)	Númerica
Last Equity Funding Amount Currency (in USD)	Númerica
Last Equity Funding Type	Categórica
Total Equity Funding Amount Currency (in USD)	Númerica
Total Funding Amount Currency (in USD)	Númerica
Number of Investors	Númerica
BuiltWith - Active Tech Count	Númerica
G2 Stack - Total Products Active	Númerica