UNIVERSIDAD
SERGIO ARBOLEDA

Técnicas Avanzadas de Minería de Datos y Machine Learning

Profesor: Luz Estela Gómez , Ph. D

Taller 5 – 30 Marzo de 2021

Técnicas Avanzadas de Minería de Datos y Machine Learning

Presentado Por:

Larry Prentt
Diógenes Barreto

Presentado a:
Luz Stella Gómez Fajardo, Ph. D.

UNIVERSIDAD SERGIO ARBOLEDA

Escuela de Ciencias Exactas e Ingeniería - Maestría en Matemáticas Aplicadas
Técnicas Avanzadas de Minería de Datos y Machine Learning
Taller grupal (2 personas máx.)
2021-I

Realizar un informe analítico que detalle cada uno de los siguientes puntos:

I. Utilizando la Base de Datos Universo (CrunchBase):

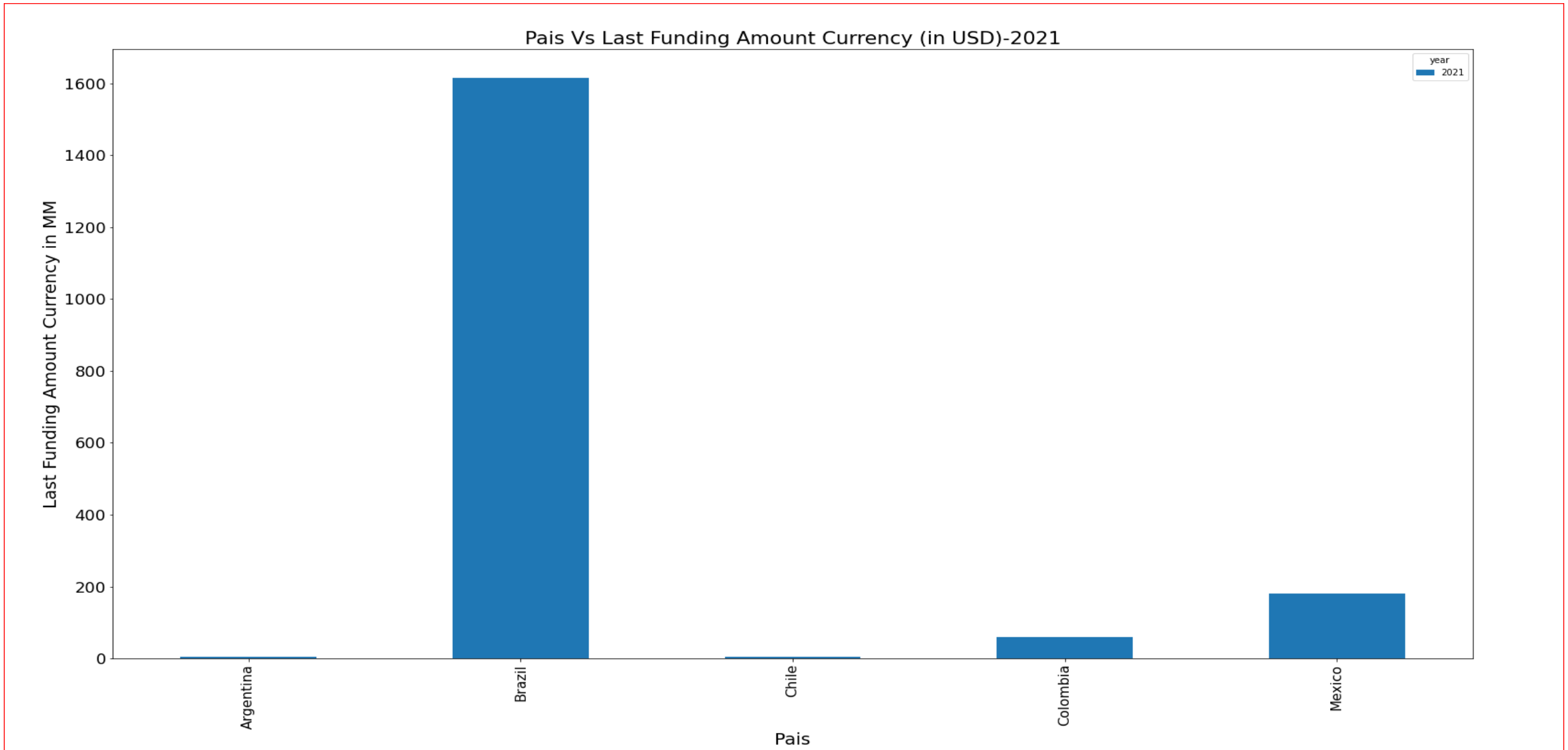
1. Cuánto capital se ha invertido en LaTAM durante el último año. Desagregue gráficamente por país.
2. Haga una comparación entre Colombia con cada uno de los otros países. Analice.
3. ¿Cuáles son los fondos que más invierten en Colombia? Haga un análisis descriptivo de cada uno de ellos.
   3.1 ¿Cuál es la tesis de inversión de cada uno de estos fondos?
4. Muestre gráficamente los exits de capital privado en Colombia por deal size.
5. Muestre el crecimiento porcentual mensual de ingresos por inversión en Colombia en comparación con los demás países.
6. ¿De acuerdo con los hallazgos, qué le hace falta a Colombia para lograr más inversión?

II. Con la unión de las bases de datos, luego de etiquetar con 1 para coincidencias y 0 en caso contrario:

1. Construir el data warehouse
2. Validar gráficamente y eliminar aquellas variables que no afectan la variable de respuesta.
3. Realizar una regresión logística para determinar las características que hacen exitosa una startup para obtener inversión. Hacer el análisis correspondiente.

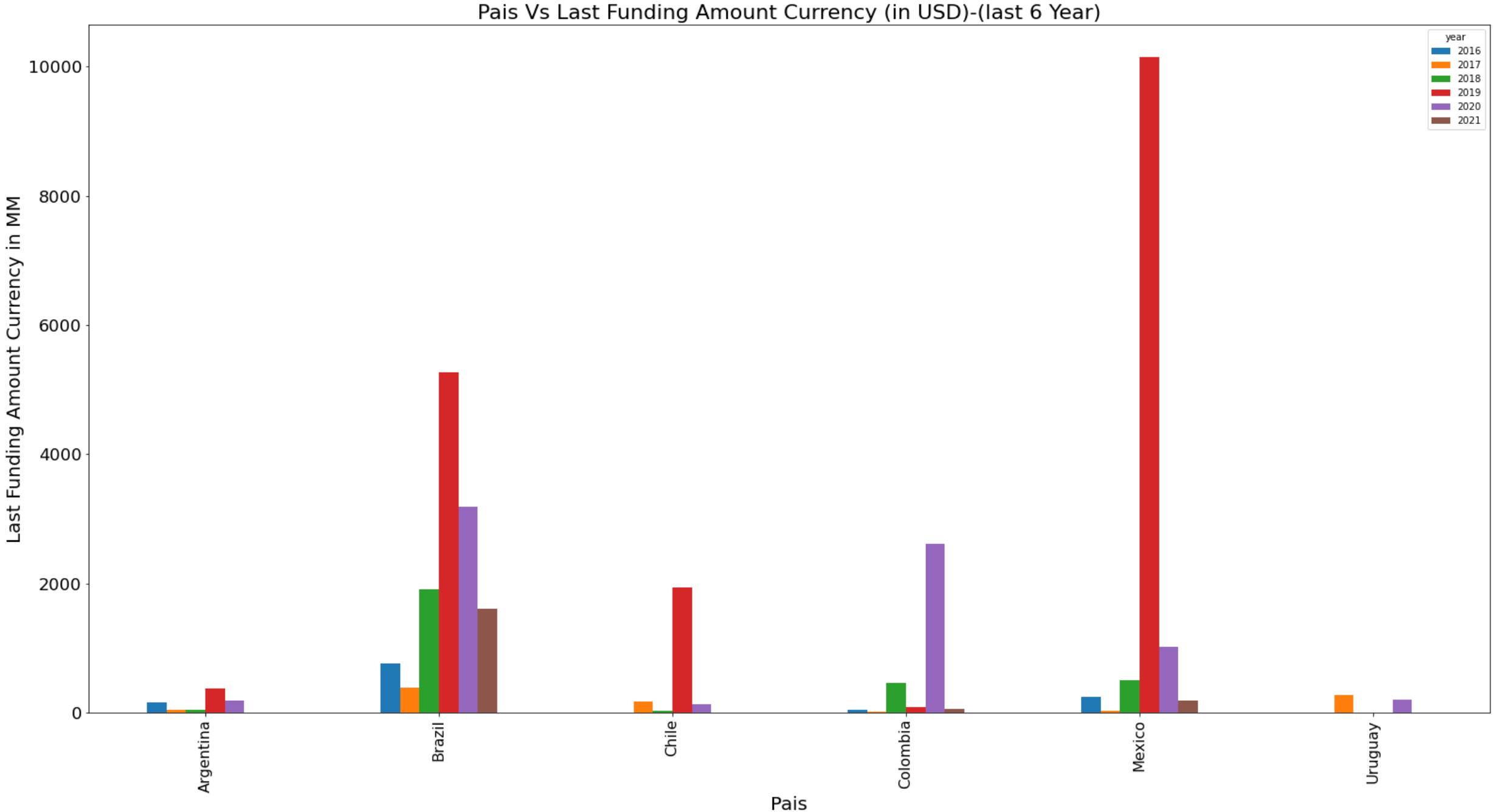
Nota: Para cada uno de los numerales agregar análisis gráfico y análisis descriptivo. Subir los códigos y gráficos a GitHub. El informe debe estar PDF. Fecha de entrega: 10 Abril 2021 – 7:00 am.

1.Cuánto capital se ha invertido en LaTAM durante el último año. Desagregue gráficamente por país

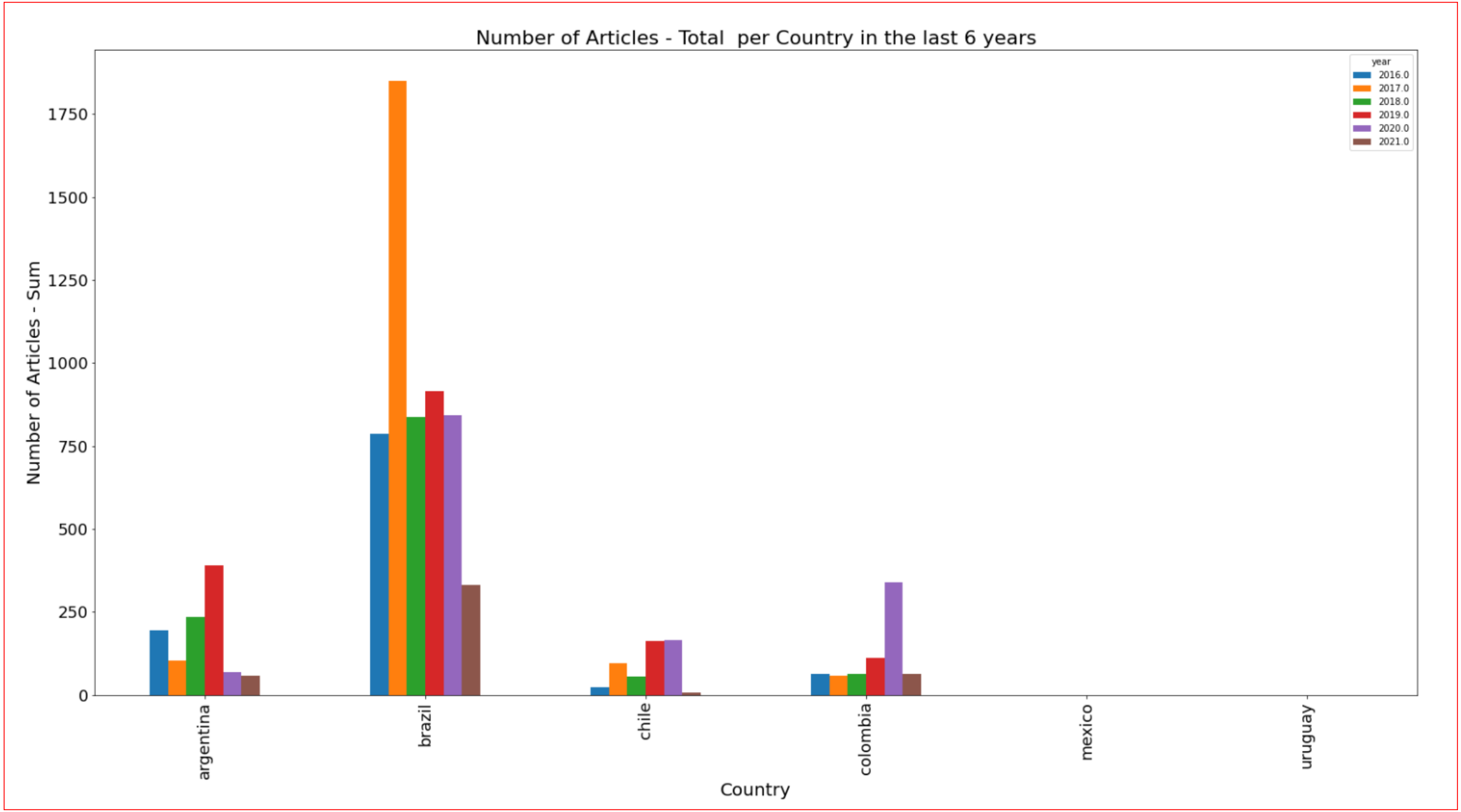

Pais Vs Last Funding Amount Currency (in USD)-2021
year
2021
Last Funding Amount Currency in MM
0
200
400
600
800
1000
1200
1400
1600
Argentina
Brazil
Chile
Colombia
Mexico
Pais

Análisis: De acuerdo al grafico , en lo que ha corrido del 2021 notamos que la mayor inversión en la región de Latinoamérica durante el 2021 corresponde a Brasil, seguido de México y Colombia. Países como Argentina y Chile la inversión es mínima , posiblemente afectada por problemas sociales.

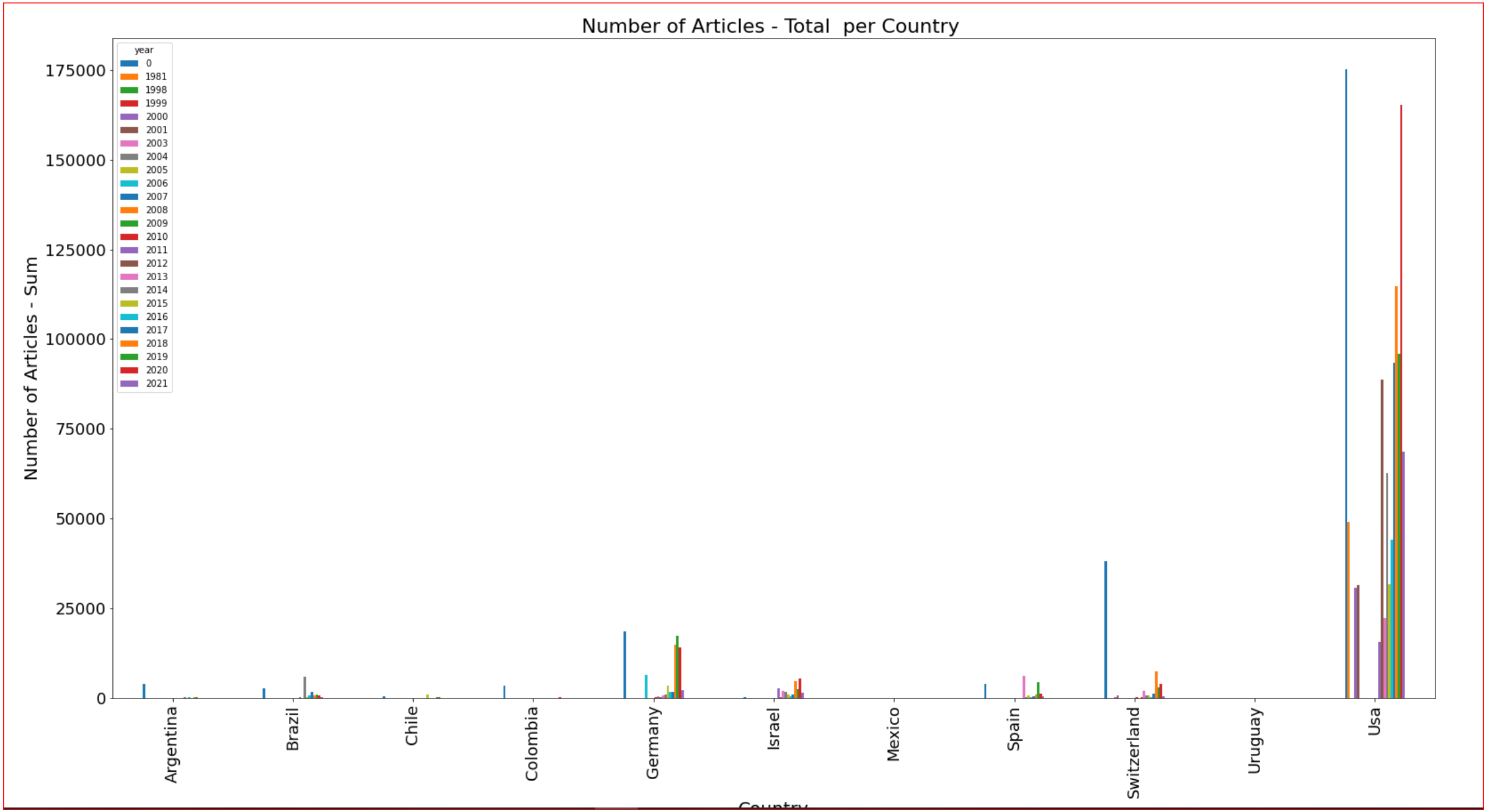
2. Haga una comparación entre Colombia con cada uno de los otros países


Pais Vs Last Funding Amount Currency (in USD)-(last 6 Year)
Last Funding Amount Currency in MM
Pais
year
2016
2017
2018
2019
2020
2021
0
2000
4000
6000
8000
10000
Argentina
Brazil
Chile
Colombia
Mexico
Uruguay

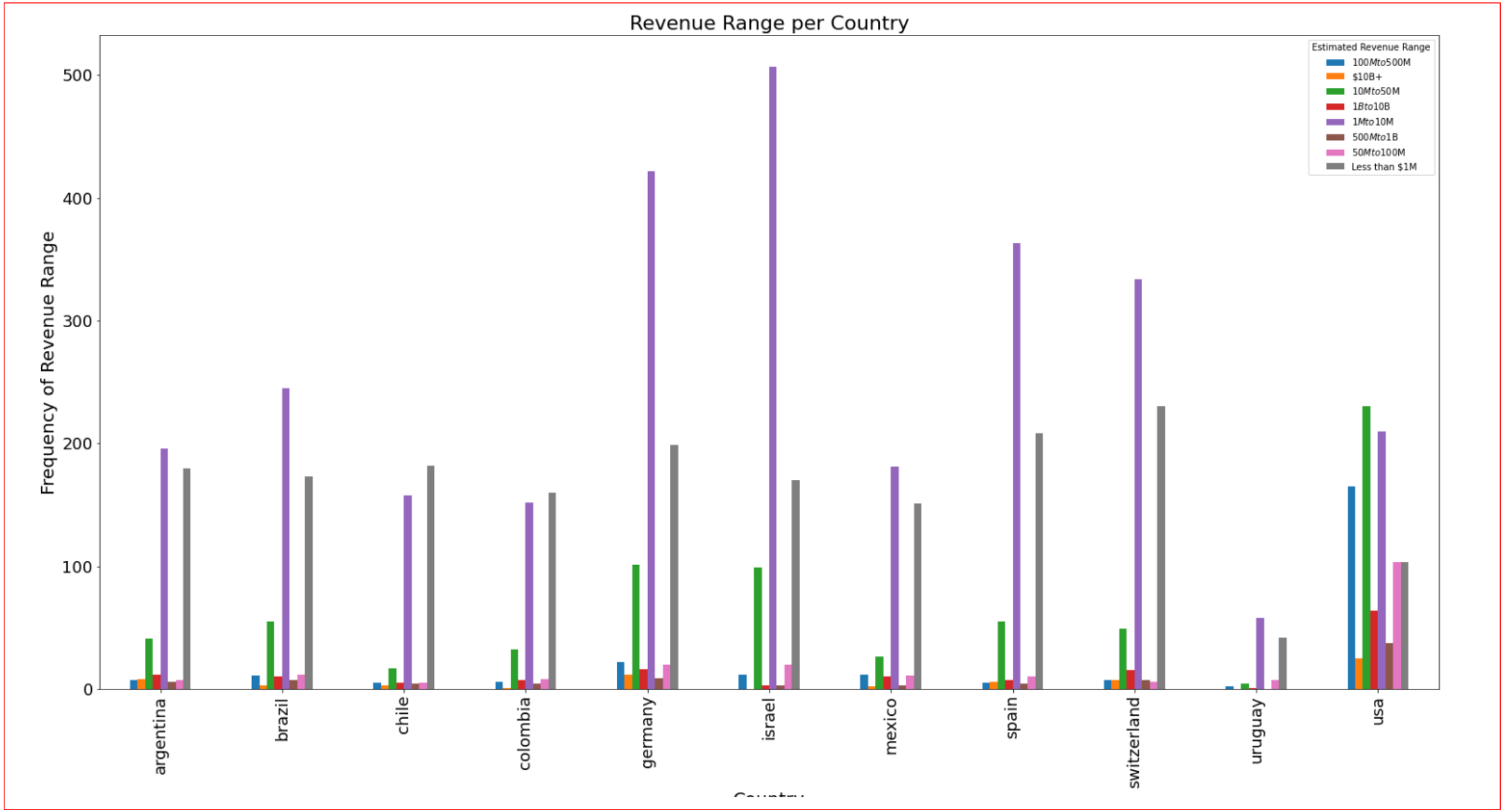
Análisis: Al comparar a Colombia con los países con la región de Latinoamérica, los países México y Brasil han tenido la mayor inversión en los ultimo 6 años , seguido de Colombia. Uruguay y Argentina, los países con menor inversión.


Number of Articles - Total per Country in the last 6 years
year
2016.0
2017.0
2018.0
2019.0
2020.0
2021.0
Number of Articles - Sum
0
250
500
750
1000
1250
1500
1750
argentina
brazil
chile
colombia
mexico
uruguay
Country

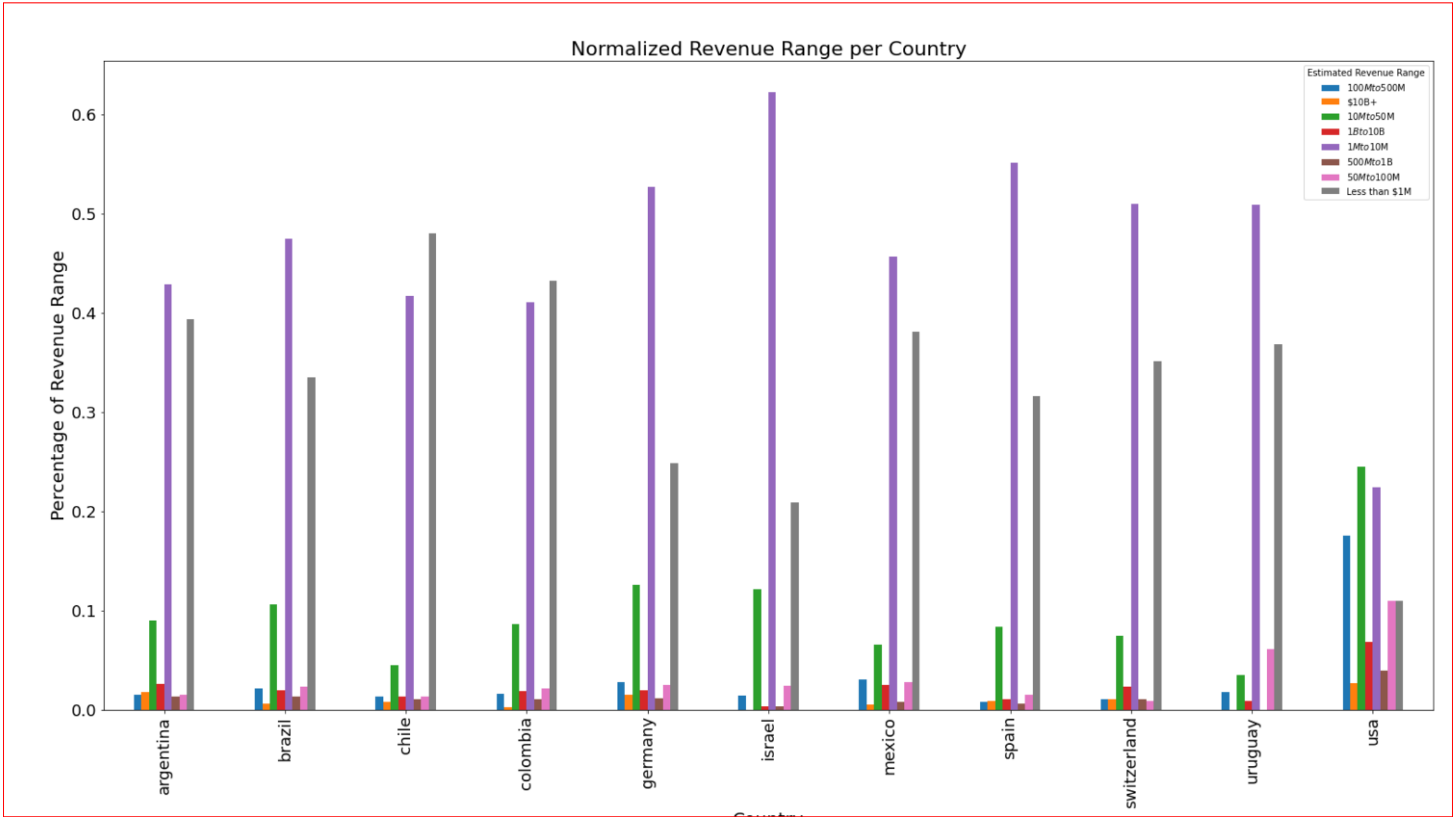
Análisis: Al comparar a Colombia con los países de la región de Latinoamérica, en ventas de articulo, Brasil lidera las ventas , seguido de argentina y Colombia.


Number of Articles - Total per Country
year
0
1981
1998
1999
2000
2001
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Number of Articles - Sum
175000
150000
125000
100000
75000
50000
25000
0
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Country

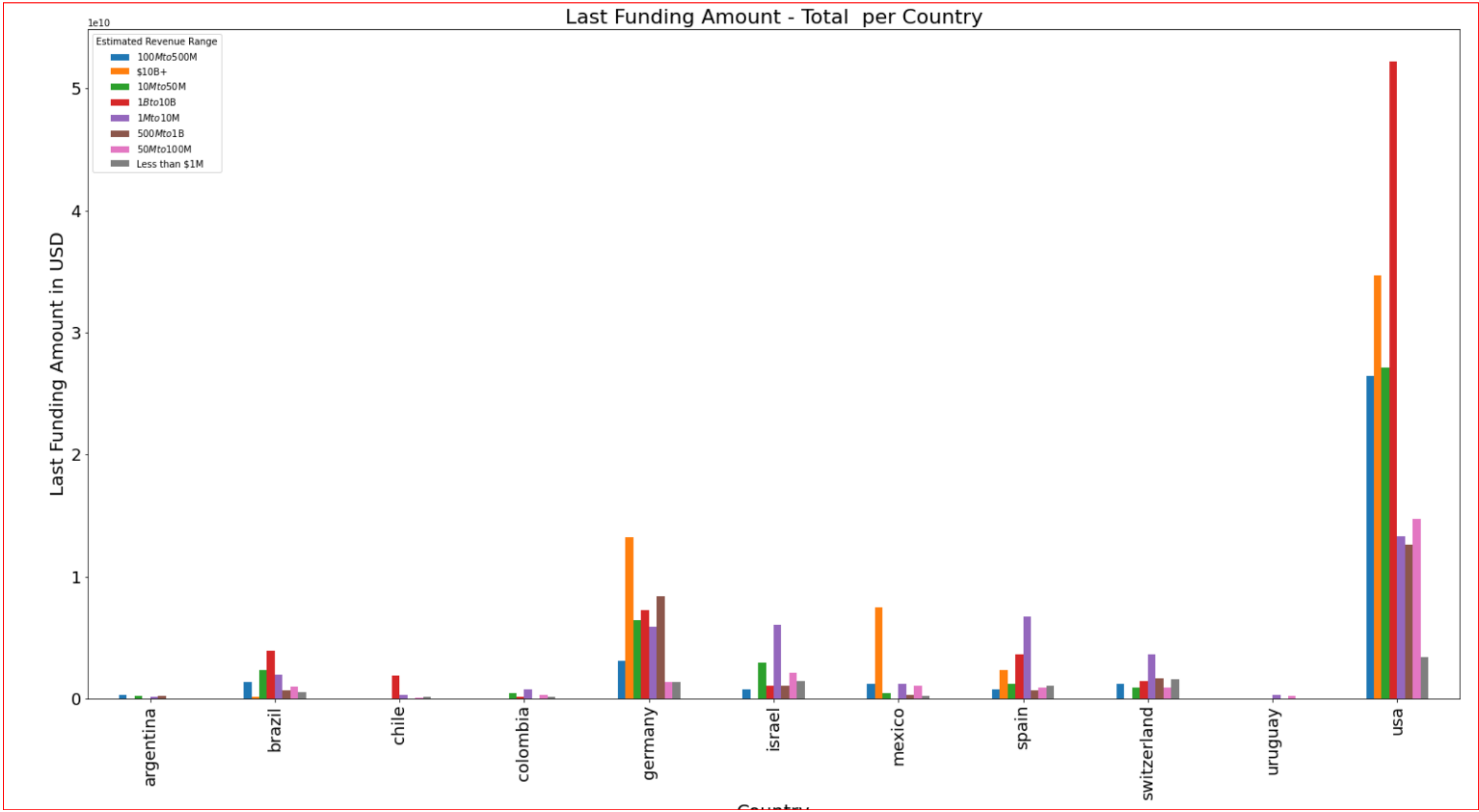
Análisis: Al comparar a Colombia con el resto de países, se determina que estados unidos es el principal país en ventas de artículos, seguida de Alemania.

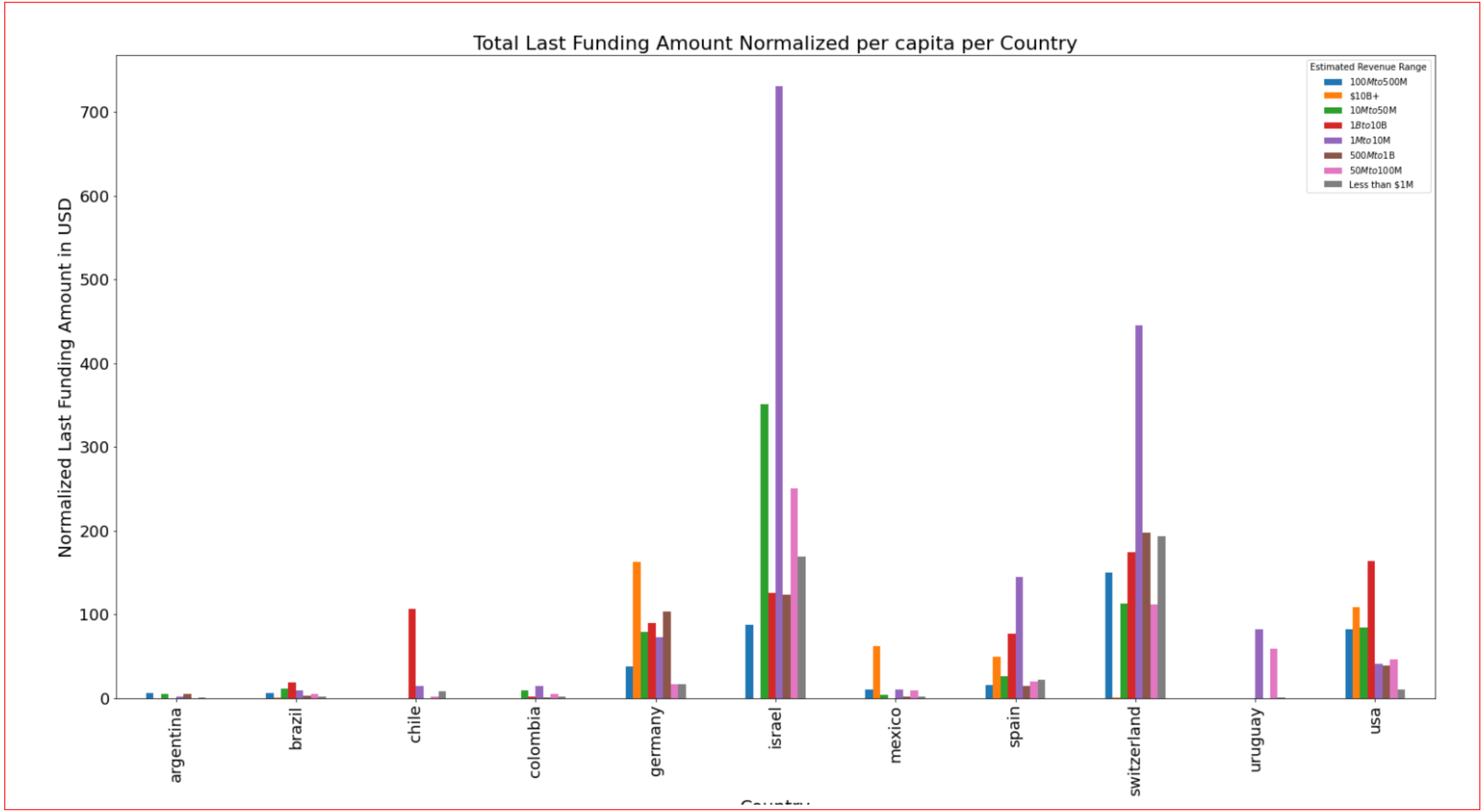

Revenue Range per Country
Estimated Revenue Range
100Mto500M
$10B+
10Mto50M
1Bto10B
1Mto10M
500Mto1B
50Mto100M
Less than $1M
Frequency of Revenue Range
0
100
200
300
400
500
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa

Análisis: En la grafica se compara el numero de compañía por rango de ingresos en los diferentes países , en Colombia predomina las compañías con rango de ingreso entre 1-10 M (USD) y menor de 1 M(USD). Estados unidos es el país con mayor numero de compañías con ingresos altos.


Normalized Revenue Range per Country
Estimated Revenue Range
100Mto500M
$10B+
10Mto50M
1Bto10B
1Mto10M
500Mto1B
50Mto100M
Less than $1M
Percentage of Revenue Range
0.0
0.1
0.2
0.3
0.4
0.5
0.6
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa

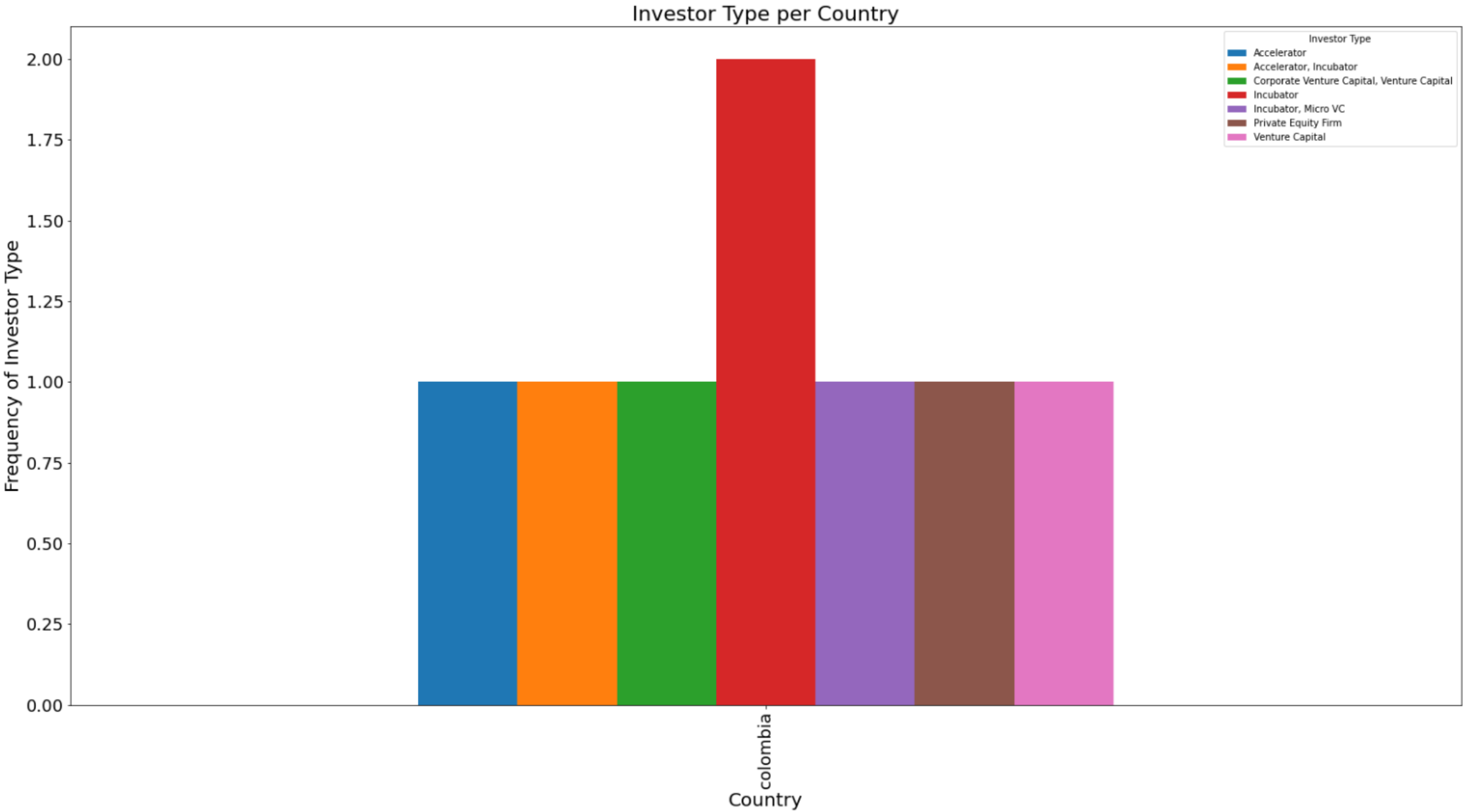
Análisis: En la grafica se compara el porcentaje de compañía por rango de ingresos en los diferentes países. Israel , corresponde al país con un 60% de compañías en un rango de ingreso de 1-10M (USD). Las startups más numerosas son las que ganan entre 1 y 10 MMUSD y están entre el 40 y 60% de las startups de cada país.


Last Funding Amount - Total per Country
1e10
Estimated Revenue Range
100Mto500M
$10B+
10Mto50M
1Bto10B
1Mto10M
500Mto1B
50Mto100M
Less than $1M
Last Funding Amount in USD
0
1
2
3
4
5
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa

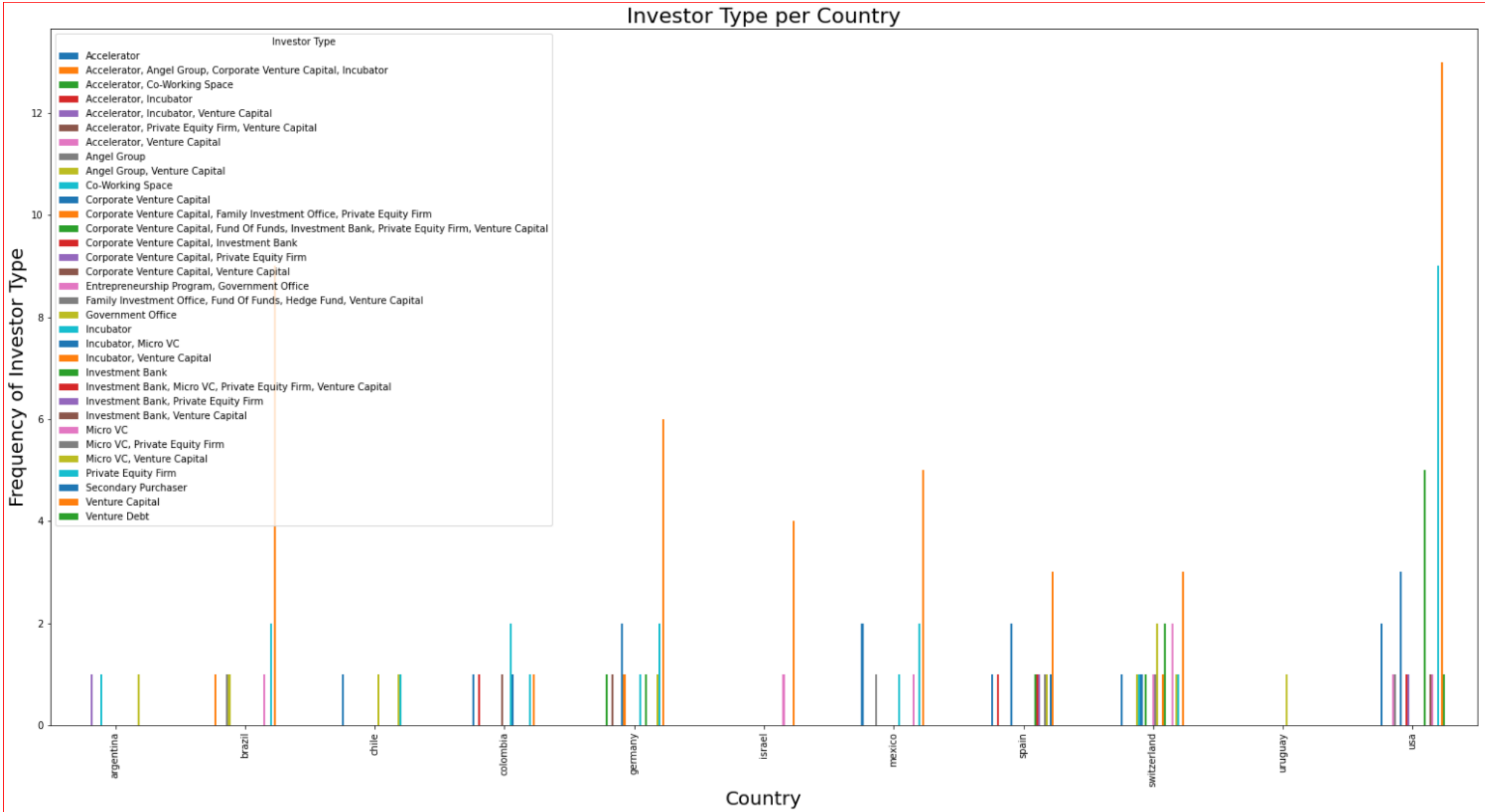

Total Last Funding Amount Normalized per capita per Country
Normalized Last Funding Amount in USD
0
100
200
300
400
500
600
700
Estimated Revenue Range
100Mto500M
$10B+
10Mto50M
1Bto10B
1Mto10M
500Mto1B
50Mto100M
Less than $1M
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa
Country

Análisis: En esta grafica comparamos el ingreso per capital por país de acuerdo a su población. Israel , es el país con mejor ingreso per capital acorde a su población, supera a países como estados unidos y Alemania. En Colombia coincide con la realidad que atraviesa al país, ingresos muy bajos.

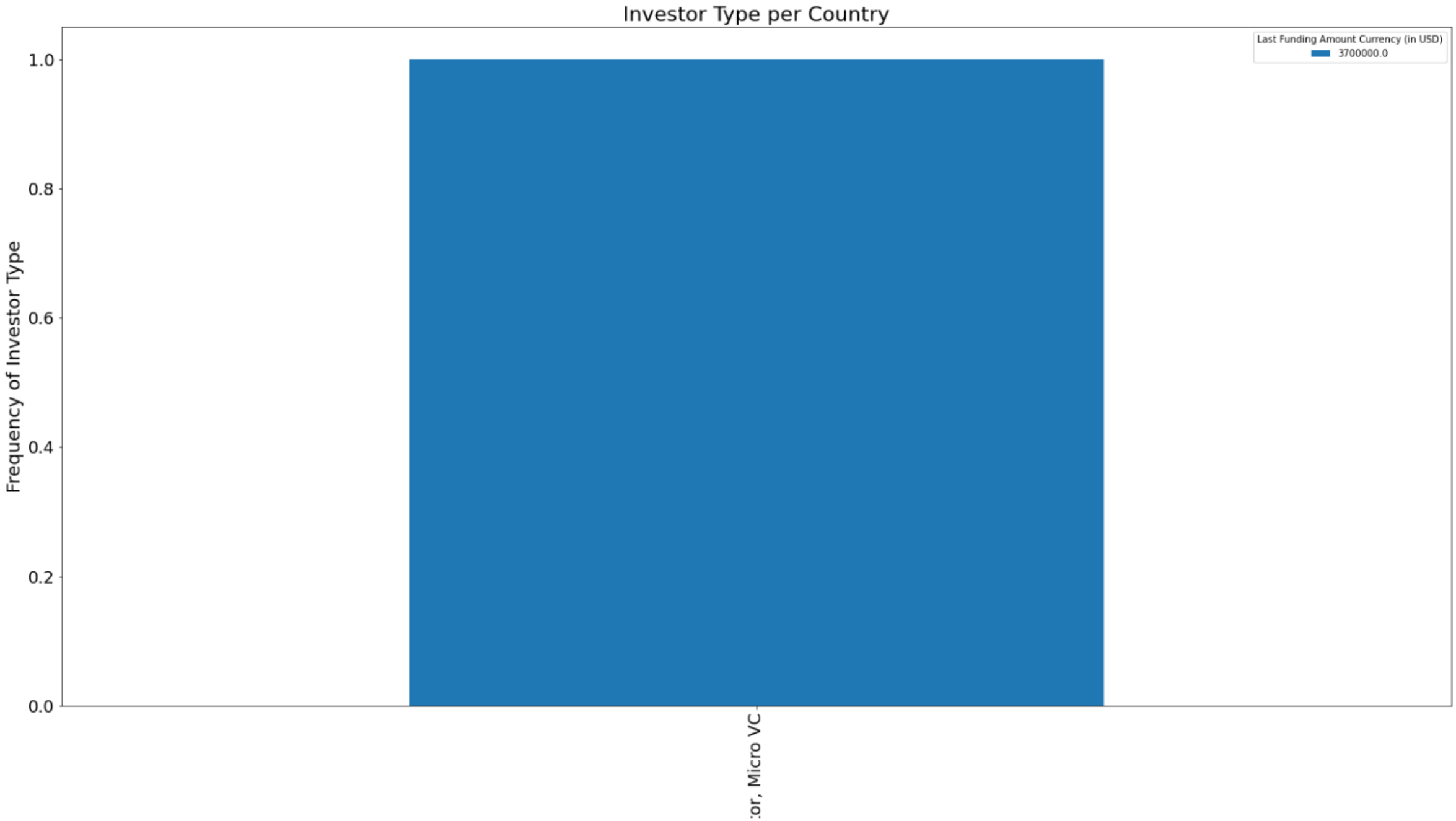
3. ¿Cuáles son los fondos que más invierten en Colombia? Haga un análisis descriptivo de cada uno de ellos

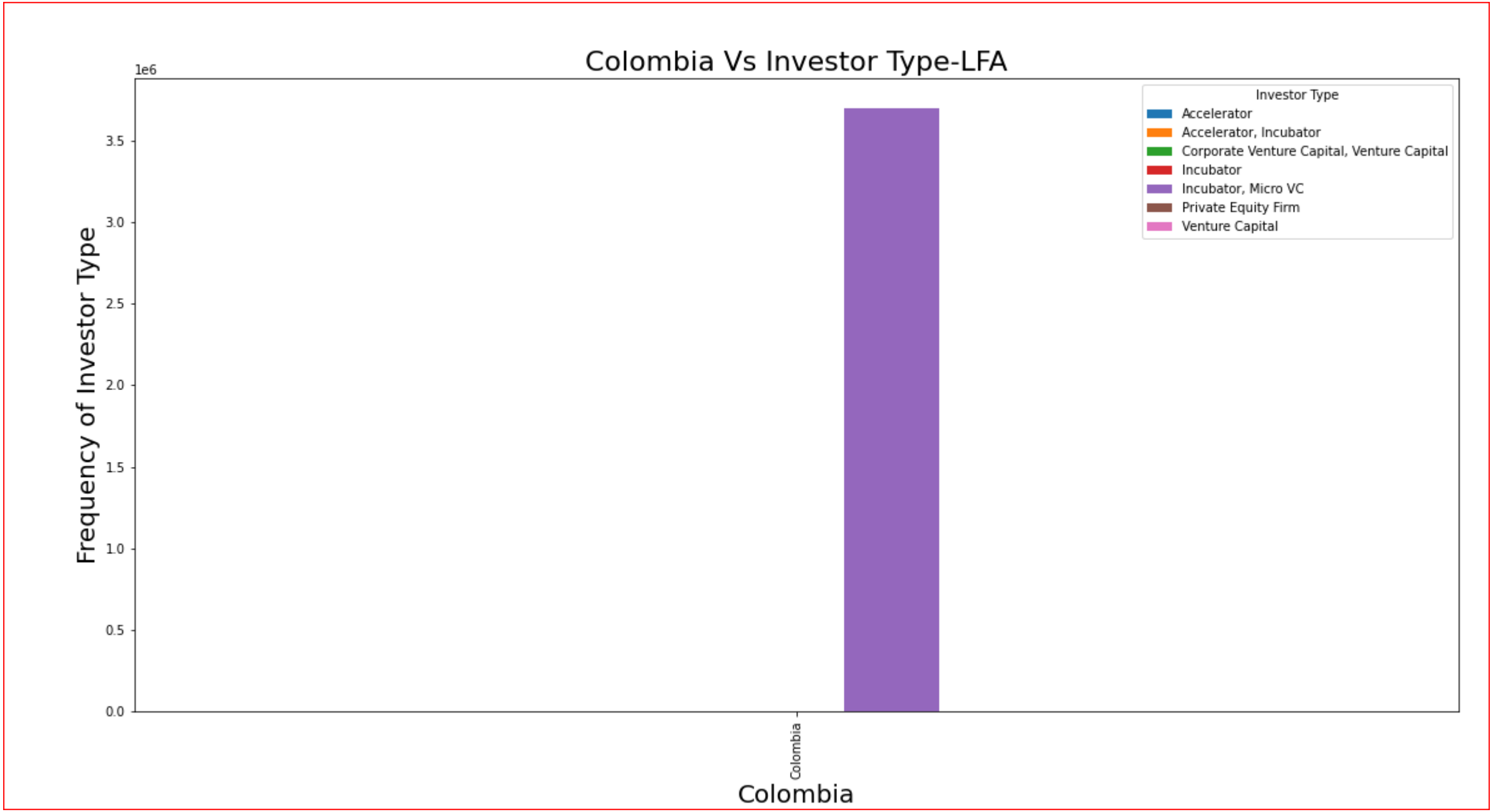

Investor Type per Country
Investor Type
Accelerator
Accelerator, Incubator
Corporate Venture Capital, Venture Capital
Incubator
Incubator, Micro VC
Private Equity Firm
Venture Capital
Frequency of Investor Type
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
colombia
Country

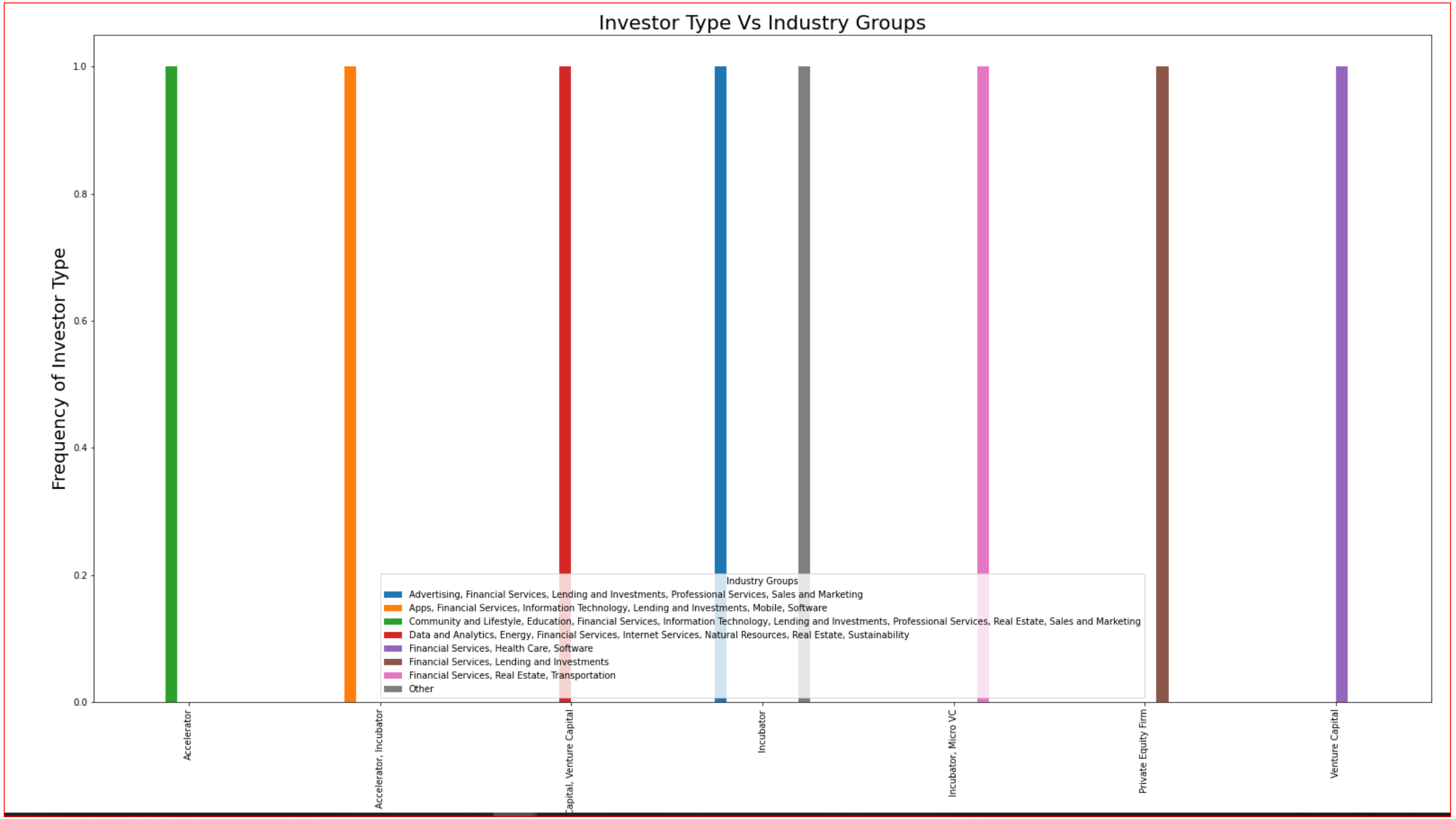
Análisis: Al comparar los fondos inversores que mas invierten en Colombia, se obtiene solo información relacionada a 8 fondos inversores, siendo el fondo “incubator” el de mayor frecuencia. Adicional, la información que se tiene de los fondos es muy poca para poder un análisis mas detallado. De acuerdo a la base de datos los tipos de inversores en Colombia son: Accelerator, incubator, Venture Capital, Micro VC y Private Equity Firm. **Tesis de inversión:** Revisando dataframe estos inversionistas apoyaron principalmente empresas del sector financiero y una sola empresa de tecnología (data análisis)


Investor Type per Country
Investor Type
Accelerator
Accelerator, Angel Group, Corporate Venture Capital, Incubator
Accelerator, Co-Working Space
Accelerator, Incubator
Accelerator, Incubator, Venture Capital
Accelerator, Private Equity Firm, Venture Capital
Accelerator, Venture Capital
Angel Group
Angel Group, Venture Capital
Co-Working Space
Corporate Venture Capital
Corporate Venture Capital, Family Investment Office, Private Equity Firm
Corporate Venture Capital, Fund Of Funds, Investment Bank, Private Equity Firm, Venture Capital
Corporate Venture Capital, Investment Bank
Corporate Venture Capital, Private Equity Firm
Corporate Venture Capital, Venture Capital
Entrepreneurship Program, Government Office
Family Investment Office, Fund Of Funds, Hedge Fund, Venture Capital
Government Office
Incubator
Incubator, Micro VC
Incubator, Venture Capital
Investment Bank
Investment Bank, Micro VC, Private Equity Firm, Venture Capital
Investment Bank, Private Equity Firm
Investment Bank, Venture Capital
Micro VC
Micro VC, Private Equity Firm
Micro VC, Venture Capital
Private Equity Firm
Secondary Purchaser
Venture Capital
Venture Debt
Frequency of Investor Type
0
2
4
6
8
10
12
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa
Country

3.1 ¿Cuál es la tesis de inversión de cada uno de estos fondos?

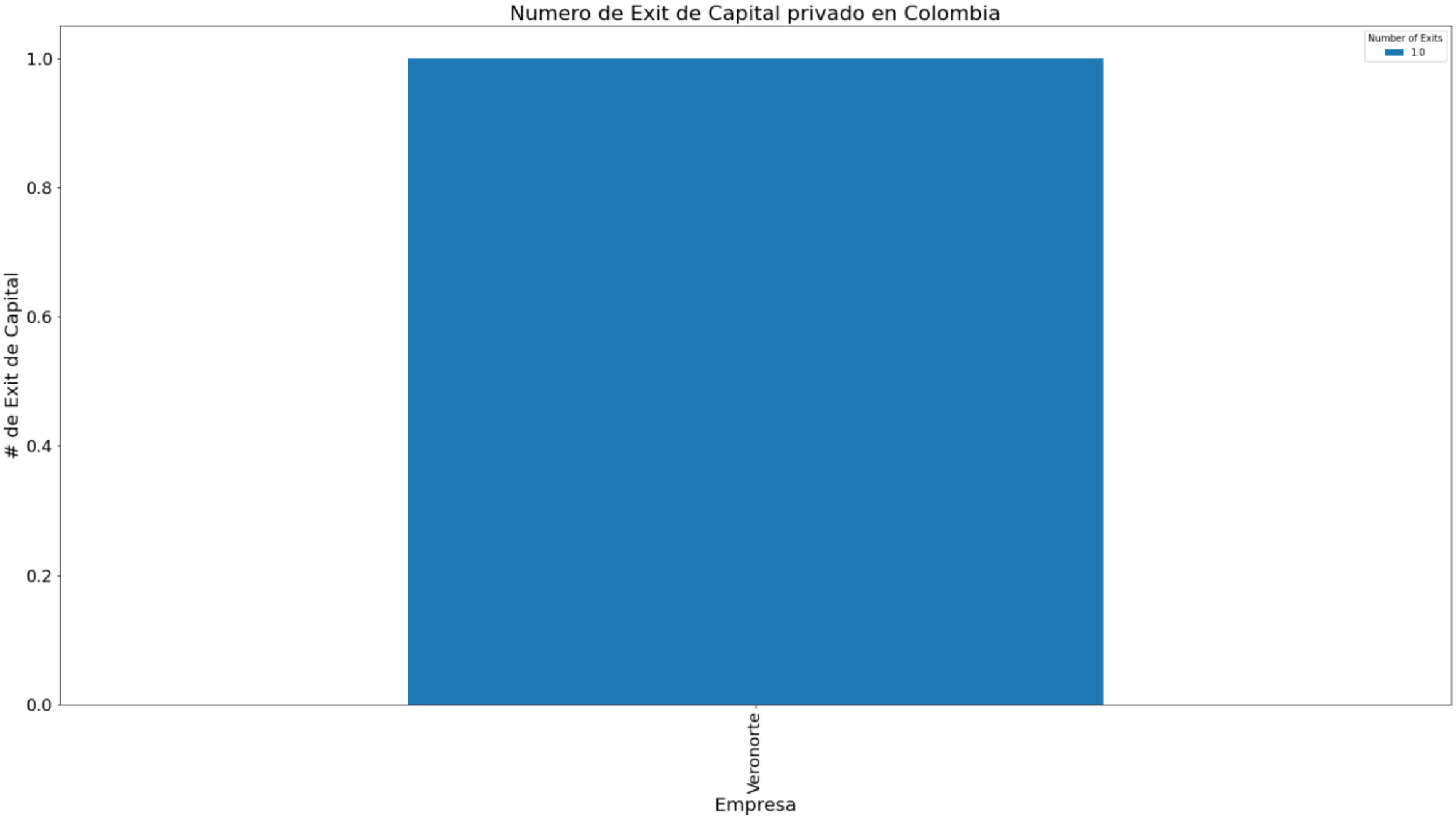

Investor Type per Country
Last Funding Amount Currency (in USD)
3700000.0
Frequency of Investor Type
1.0
0.8
0.6
0.4
0.2
0.0
:or, Micro VC


Colombia Vs Investor Type-LFA
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Frequency of Investor Type
Investor Type
Accelerator
Accelerator, Incubator
Corporate Venture Capital, Venture Capital
Incubator
Incubator, Micro VC
Private Equity Firm
Venture Capital
Colombia
Colombia


Investor Type Vs Industry Groups
Frequency of Investor Type
1.0
0.8
0.6
0.4
0.2
0.0
Industry Groups
Advertising, Financial Services, Lending and Investments, Professional Services, Sales and Marketing
Apps, Financial Services, Information Technology, Lending and Investments, Mobile, Software
Community and Lifestyle, Education, Financial Services, Information Technology, Lending and Investments, Professional Services, Real Estate, Sales and Marketing
Data and Analytics, Energy, Financial Services, Internet Services, Natural Resources, Real Estate, Sustainability
Financial Services, Health Care, Software
Financial Services, Lending and Investments
Financial Services, Real Estate, Transportation
Other
Accelerator
Accelerator, Incubator
Capital, Venture Capital
Incubator
Incubator, Micro VC
Private Equity Firm
Venture Capital

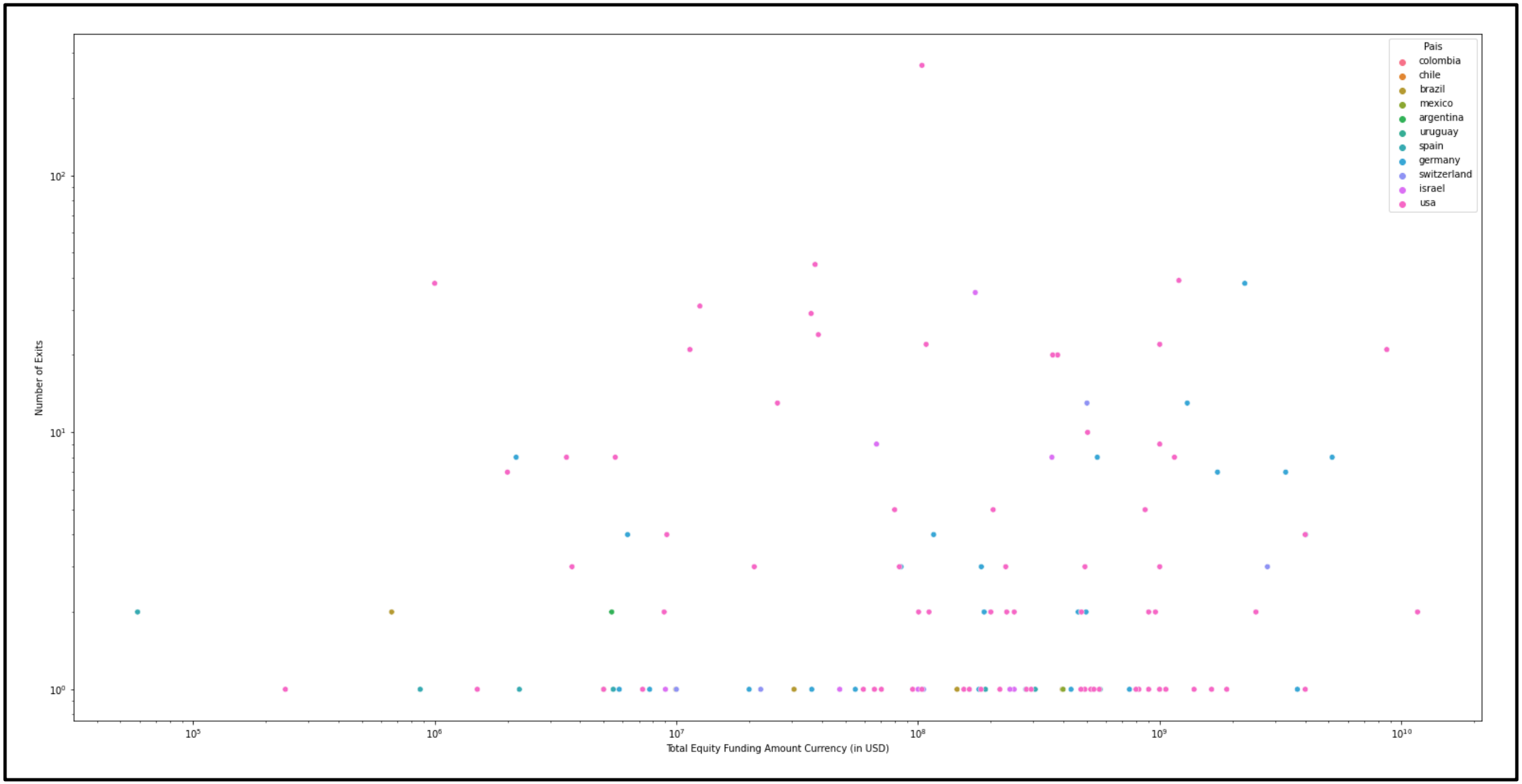
Análisis: En data solo se tiene información del monto de inversión por parte del fondo “Incubator, Micro VC”, por 3 .7 M (USD). Los fondos tienen una preferencia de inversión en compañías del sector financiero y sector de tecnologías.

4. Muestre gráficamente los exits de capital privado en Colombia por deal size.

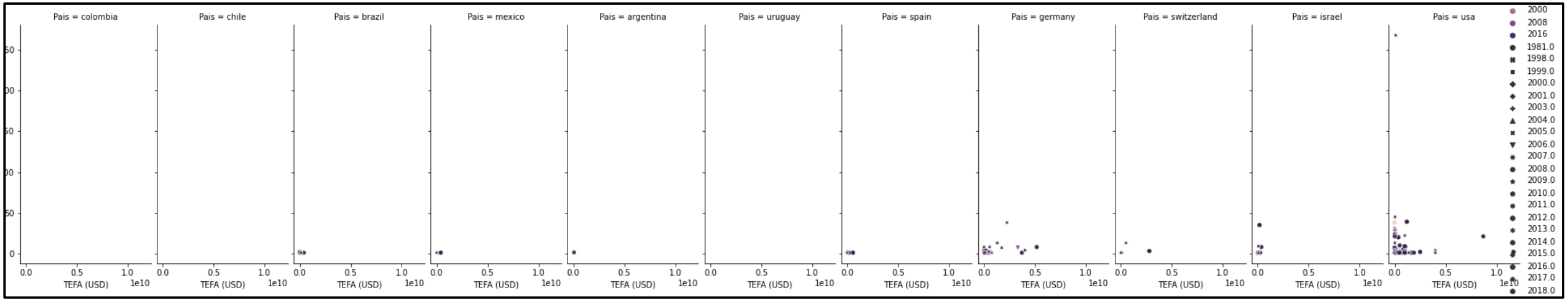

Numero de Exit de Capital privado en Colombia
Number of Exits
1.0
1.0
0.8
0.6
0.4
0.2
0.0
de Exit de Capital
Veronorte
Empresa

Análisis: De acuerdo a la grafica , solo se tiene información de una sola compañía de capital privado que ha salido del país, “Veronorte”.

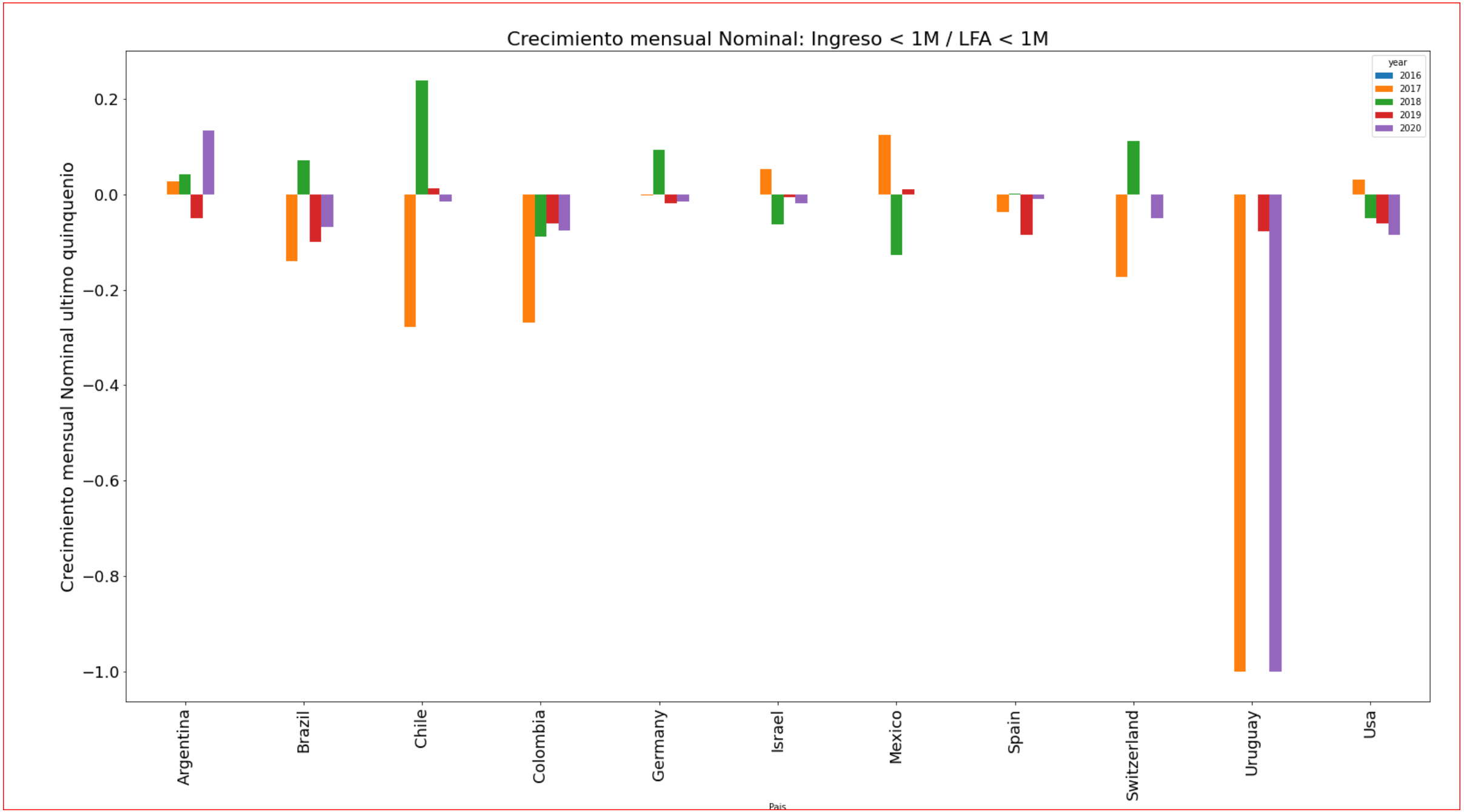
Se intentó correlacionar number of exits con total Equity Funding amount con data de otros paises



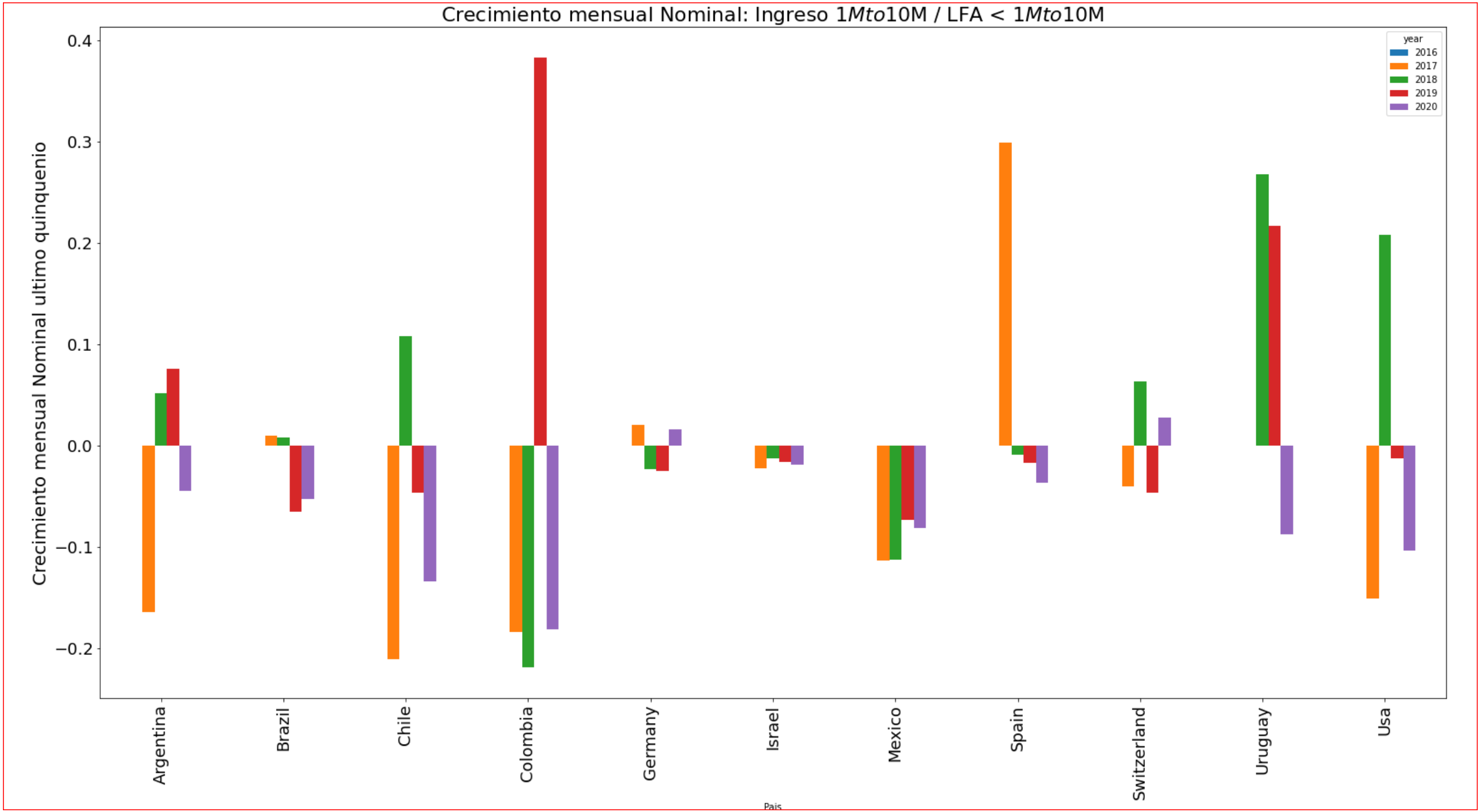
Se intentó correlacionar number of exits con total Equity Funding amount con data de otros paises



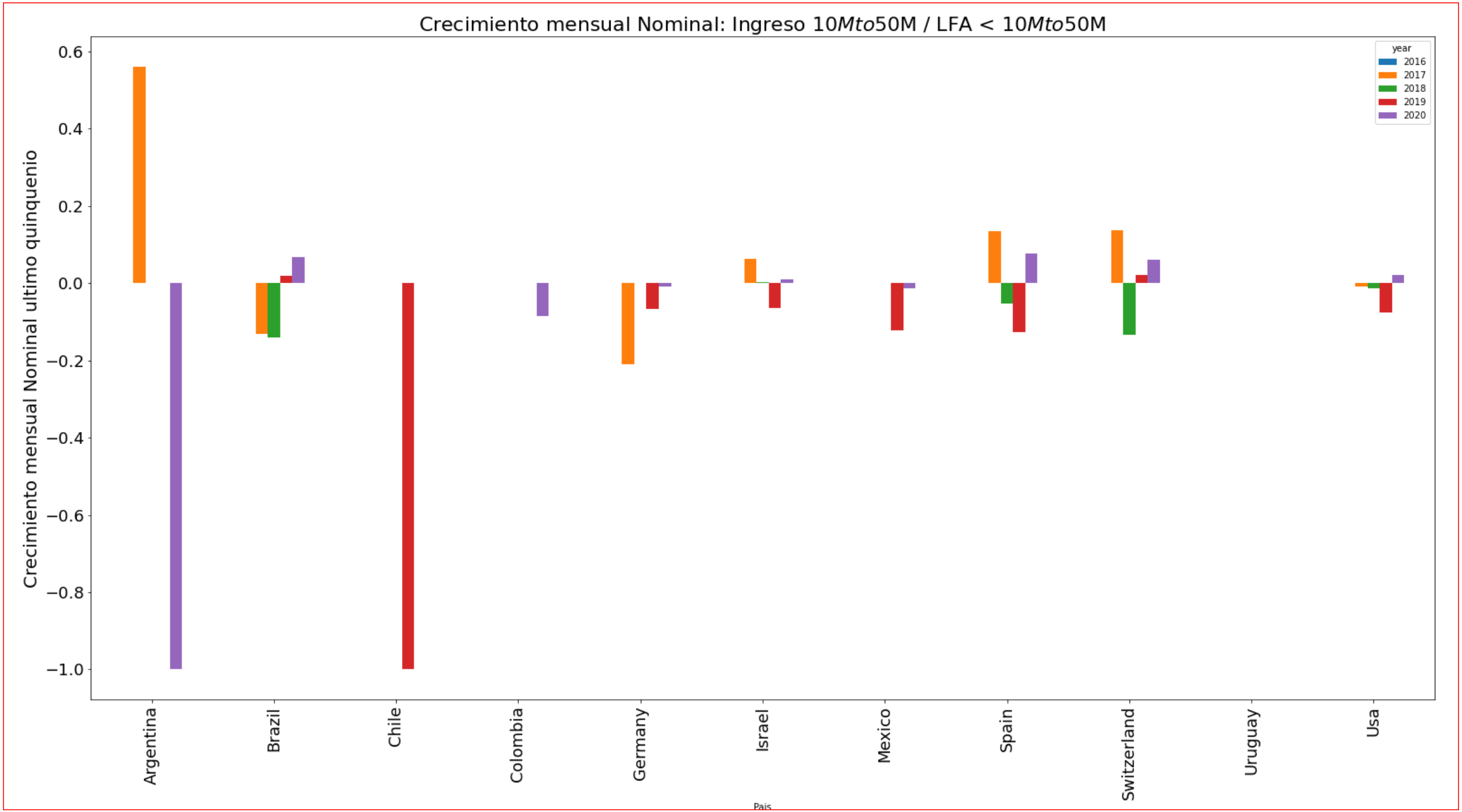
5. Muestre el crecimiento porcentual mensual de ingresos por inversión en Colombia en comparación con los demás países


Crecimiento mensual Nominal: Ingreso < 1M / LFA < 1M
Crecimiento mensual Nominal ultimo quinquenio
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
year
2016
2017
2018
2019
2020
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Pais

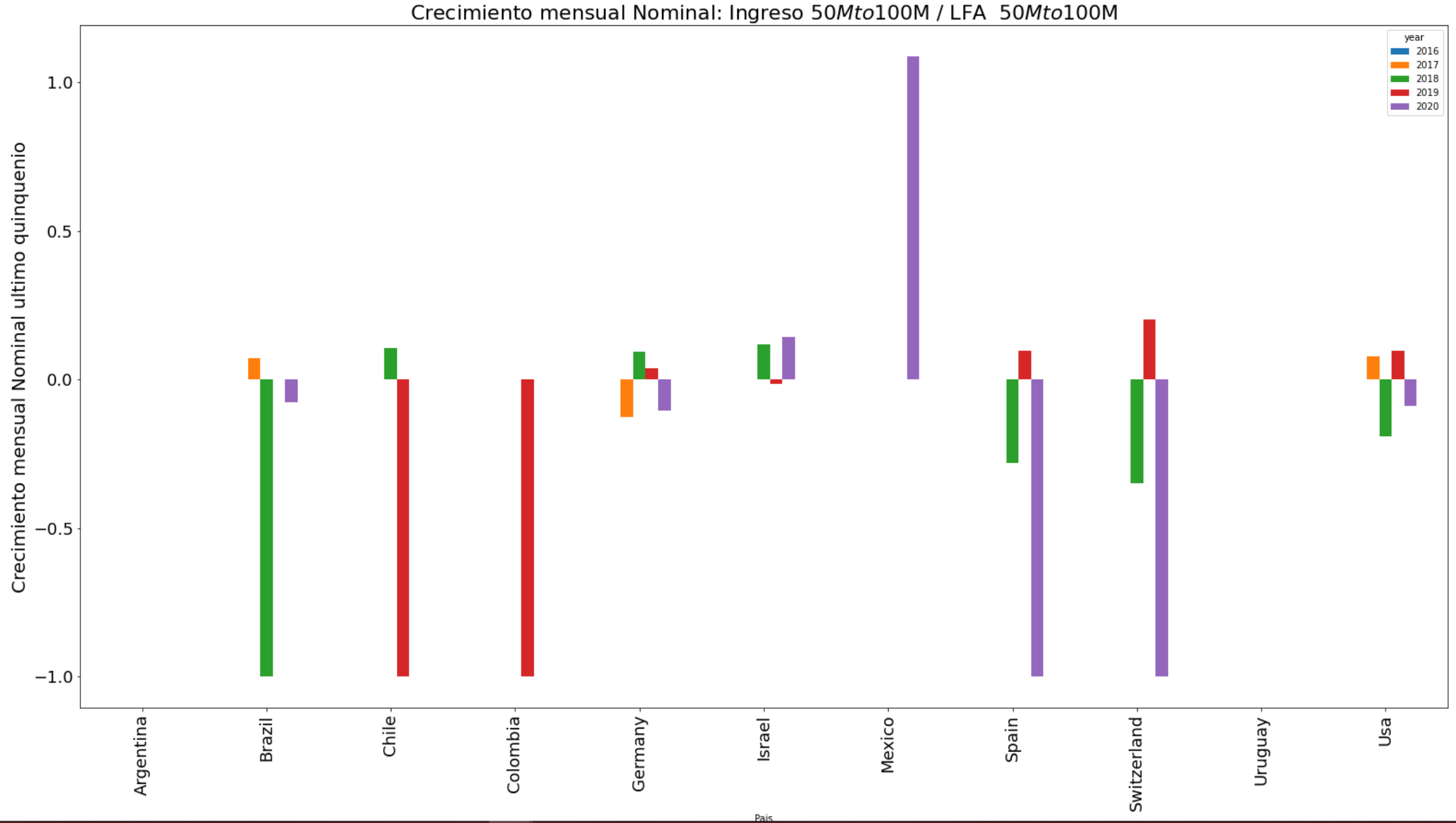
Análisis: De la grafica , notamos que las compañías que tiene ingresos menor a un 1M de dólares, tienen perdidas. Adicional, los inversores posiblemente no logran recuperar su inversión. En Colombia, las compañías con un ingreso de 1 M, mostraron perdidas en los últimos 5 años.


Crecimiento mensual Nominal: Ingreso 1Mto10M / LFA < 1Mto10M
Crecimiento mensual Nominal ultimo quinquenio
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
year
2016
2017
2018
2019
2020
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Pais

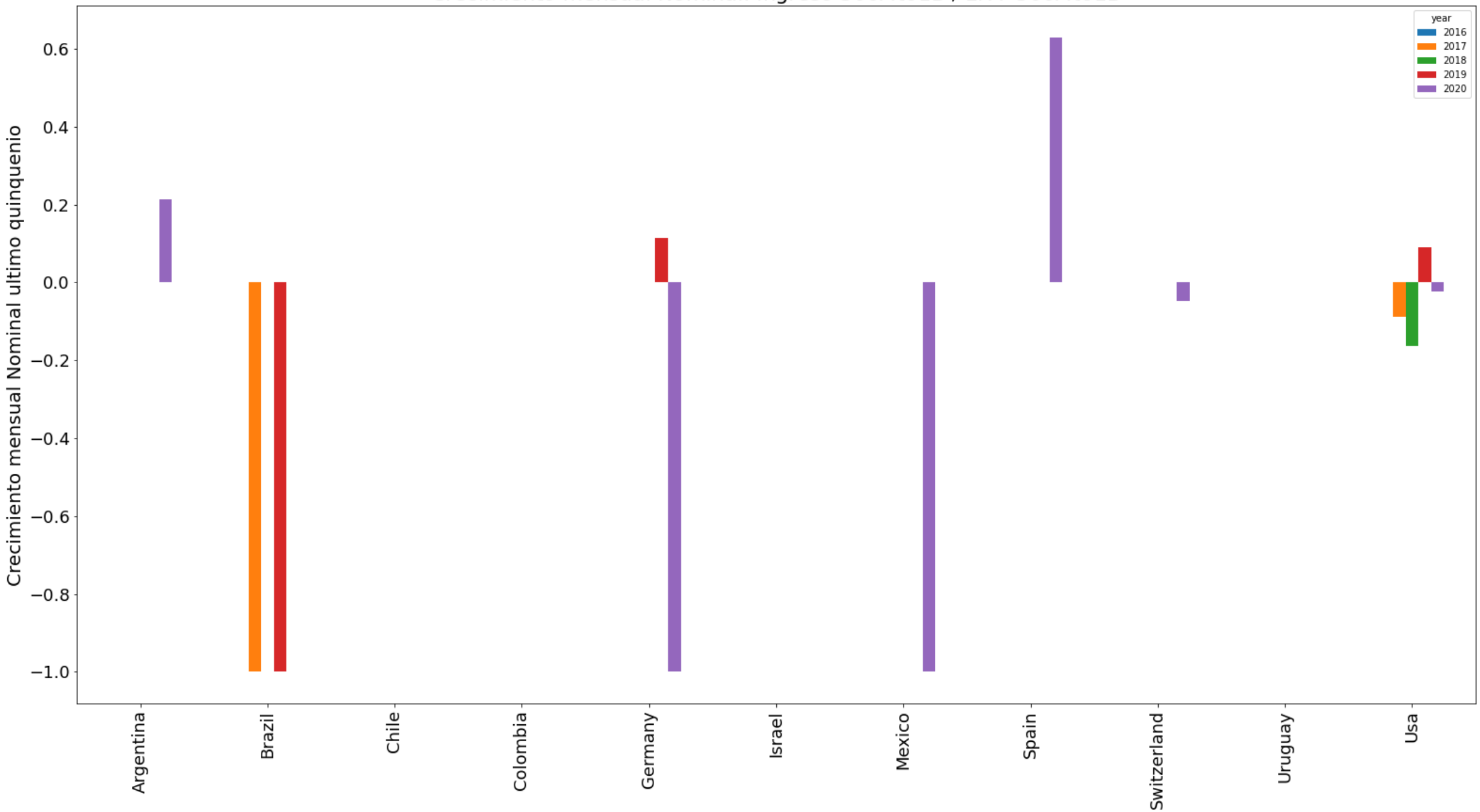
Análisis: De la grafica , notamos que las compañías que tiene ingresos entre 1 – 10M de dólares, algunas empresas tiene guanacias y otras perdidas en los últimos 5 años. En países como Israel, las compañía no muestran crecimiento, pero las perdidas son leves, muy cercano al breakeven. En Colombia, solo se tiene información de una sola compañía con un crecimiento cercano al 38%.


Crecimiento mensual Nominal: Ingreso 10*Mto*50M / LFA < 10*Mto*50M
Crecimiento mensual Nominal ultimo quinquenio
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
year
2016
2017
2018
2019
2020
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Pais

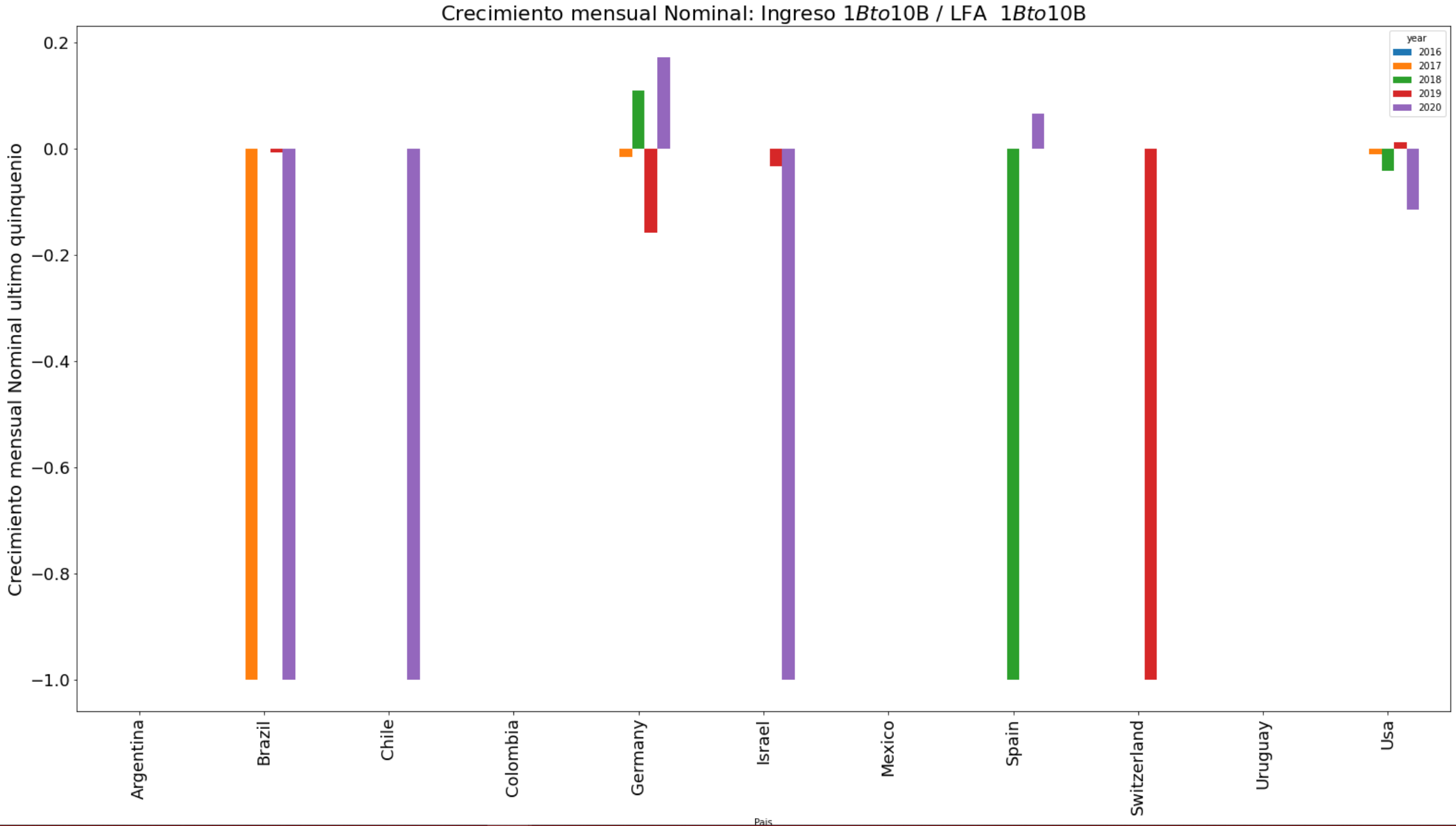
Análisis: De la grafica , notamos que las compañías que tiene ingresos entre 10 – 50M de dólares, son pocas las que tiene un crecimiento positivo. En Colombia solo se tiene registro de compañías con esté rango de ingresos durante el 2020 y su crecimiento fue negativo.


Crecimiento mensual Nominal: Ingreso 50Mto100M / LFA 50Mto100M
Crecimiento mensual Nominal ultimo quinquenio
year
2016
2017
2018
2019
2020
1.0
0.5
0.0
−0.5
−1.0
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Pais

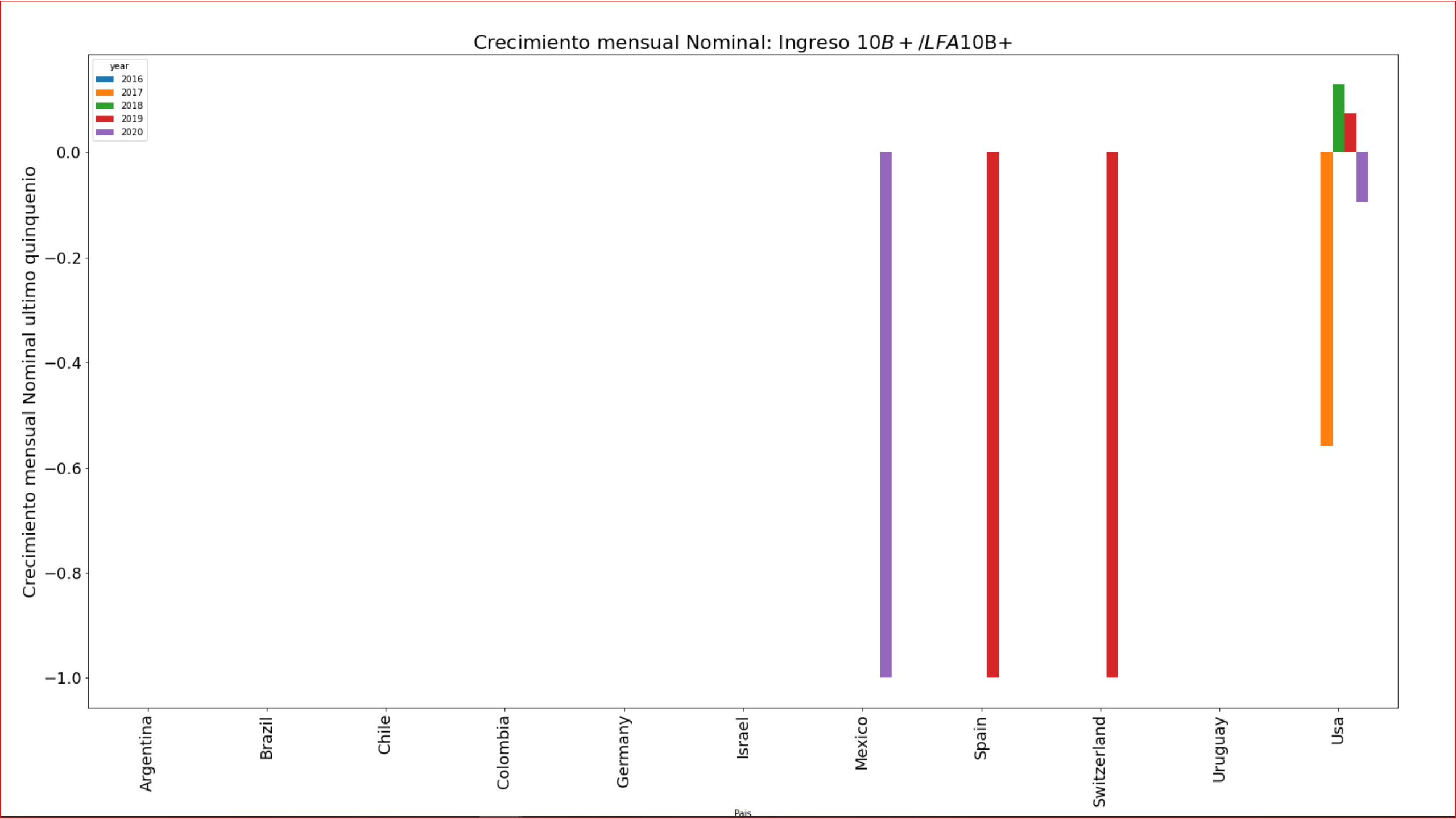
Análisis: De la grafica , notamos que las compañías que tiene ingresos entre 500 – 100M de dólares, se nota que el un gran numero de compañías en diferentes países presentaron un crecimiento pequeño en los últimos 5años. En Colombia se tiene un registro de compañía durante el 2019 , el cual fue negativo. México, único país con una compañía lo suficiente rentable durante el 2020.


Crecimiento mensual Nominal: Ingreso 500Mto1B / LFA 500Mto1B
Crecimiento mensual Nominal ultimo quinquenio
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
year
2016
2017
2018
2019
2020
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Pais

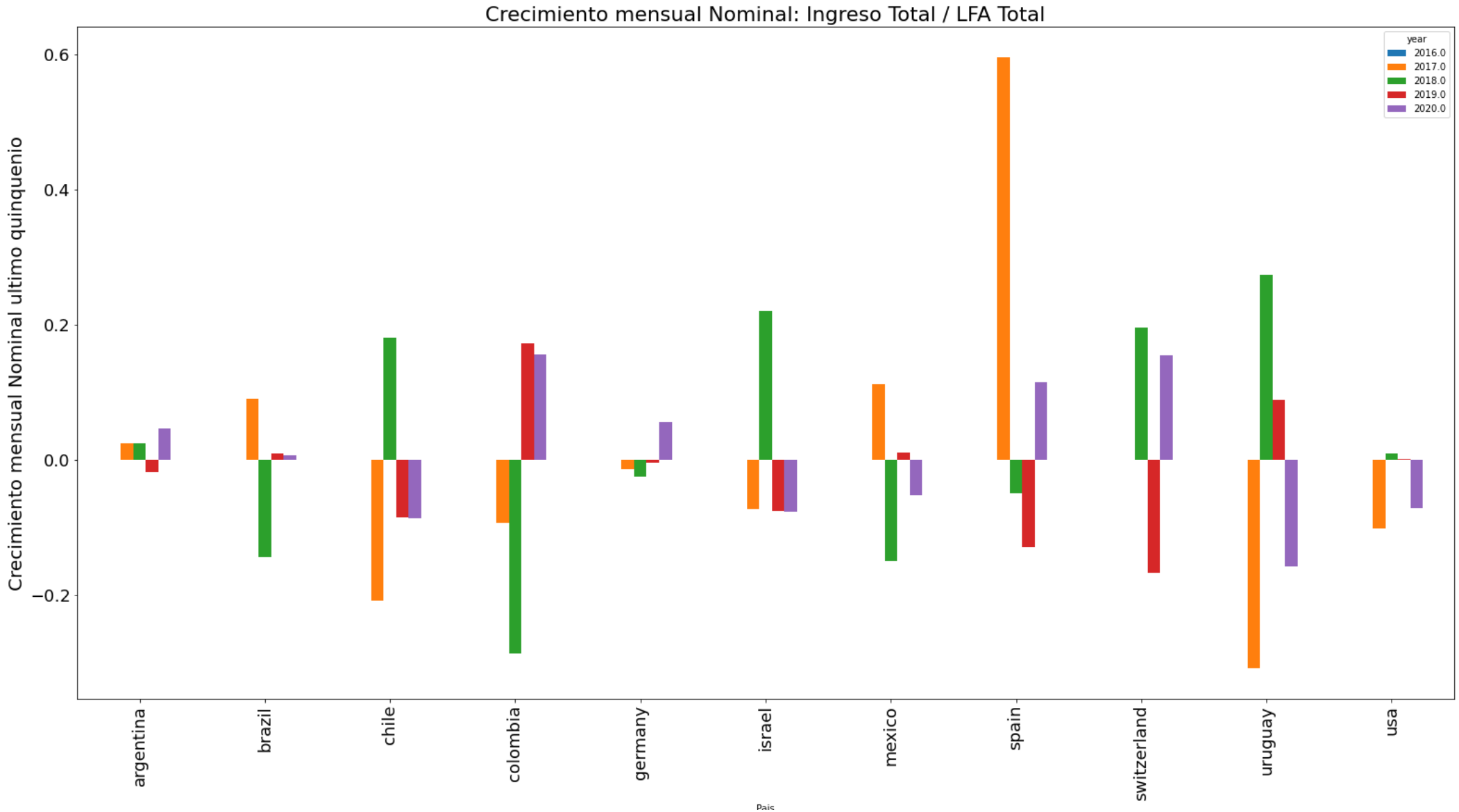
Análisis: De la grafica , notamos que muy pocos países tienen compañías con ingresos entre 500 MM – 1B de dólares. Colombia no registra compañías con estos ingresos. Las compañías en la mayoría de los países, presentaron perdidas durante el 2020.


Crecimiento mensual Nominal: Ingreso 1Bto10B / LFA 1Bto10B
Crecimiento mensual Nominal ultimo quinquenio
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
year
2016
2017
2018
2019
2020
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Pais

Análisis: De la grafica , notamos que muy pocos países tienen compañías con ingresos entre 1 – 10B de dólares. Colombia no registra compañías con estos ingresos. En países como Brasil, Chile, Israel, usa, las compañía presentaron perdidas durante el 2020. Alemania, las compañías presentaron un mejor crecimiento durante los últimos 5 años.


Crecimiento mensual Nominal: Ingreso 10B+/LFA10B+
year
2016
2017
2018
2019
2020
Crecimiento mensual Nominal ultimo quinquenio
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
Argentina
Brazil
Chile
Colombia
Germany
Israel
Mexico
Spain
Switzerland
Uruguay
Usa
Pais

Análisis: De la grafica , notamos que muy pocos países tienen compañías con ingresos 10B+ de dólares. Colombia no registra compañías con estos ingresos. USA, país con mayor numero de compañías en el rango de ingreso analizados, el crecimiento de las compañías fue positivo en los años 2018 y 2019. En el 2020 generaron perdidas.


Crecimiento mensual Nominal: Ingreso Total / LFA Total
Crecimiento mensual Nominal ultimo quinquenio
0.6
0.4
0.2
0.0
−0.2
year
2016.0
2017.0
2018.0
2019.0
2020.0
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa
Pais

6. ¿De acuerdo con los hallazgos, qué le hace falta a Colombia para lograr más inversión?

De la información, se puede determinar que Colombia es el tercer-cuarto país de la región Latam con mayor inversión. Sin embargo, es necesario generar mas incentivo para compañías del sector tecnológico (e.g. Colombia respecto a Israel esta muy atrás en el sector tecnológico). En este momento la mayoría de startups en Colombia están enfocadas en el sector financiero.

II. Con la unión de las bases de datos, luego de etiquetar con 1 para coincidencias y 0 en caso contrario:

Parte II

Base de datos Colombia Crunchbase (Marzo 21-2021) vs Top100Startups- Colombia.xlsx y Empresas Unicorn - Contactos.xlsx.

Data Warehouse

No realizó merge de los dataframe = ColombiaCB-5March21.csv, Top100Startups- Colombia.xlsx y Empresas Unicorn - Contactos.xlsx. El análisis se hace sólo con archivo **ColombiaCB-5March21.csv**

Se Convirtieron los nombres de las empresas en los 3 dataframe a minúsculas (lowercase) y realizó 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. A partir de triple intersección se creó la **variable objetivo "y"**

Se elimina columna Organization Name URL
df = df.drop(['Organization Name URL'], axis=1)

Se renombra nombre columna CB Rank (Company) a CBRank
df.rename(columns={"CB Rank (Company)": 'CBRank'}, inplace=True)

Data Warehouse

```
######## Función para determinar el porcentaje de datos nulos

def cols_90per_nulls(data, perc):
    count = 0
    cols_to_drop = {}
    for col in data.columns:
        per_nulls = data[col].isna().sum()/len(data[col])
        if per_nulls >= perc:
            cols_to_drop[col] = per_nulls
            # print(col, per_nulls)
            count+=1
        else:
            None

    print('Number of cols with > ', perc*100, '% nulls:', count)
    return cols_to_drop

# cols_to_drop es un diccionario que tiene los nombres de las Columnas del dataframe con mas del 80% de datos nulos

dict_col_nul=cols_90per_nulls(df, 0.80)

# Dataframe data mantiene el dataframe original
data = df

# Dataframe con las columnas eliminadas, en este caso 56 columnas
# con mas del 80% de datos nulos.  Dataframe df pasó de 102 a 46 Columnas
df = df.drop(columns=dict_col_nul)
df.shape
```

Data Warehouse

```
# Junta de Andalucía y Courbox son de Andalucía España.  Se eliminan
df.drop(df[df['Headquarters Location']=='Andalucía, Valle del Cauca, Colombia'].index, inplace =True)

# Neoalgae compañia española, se elimina ciudad Asturias
# Suaval Group compañia española, se elimina ciudad Asturias
df.drop(df[df['Headquarters Location']=='Asturias, Cundinamarca, Colombia'].index, inplace =True)

# Aspyre Solutions (hay una compañia en USA y otra en UK), aparece con dirección en Barrios
# unidos. Barrios Unidos, Distrito Especial, Colombia Se elimina
df.drop(df[df['Headquarters Location']=='Barrios Unidos, Distrito Especial, Colombia'].index, inplace =True)

# BD Sensors, Pinnery, Ratiokontakt y Atlantik Bruecke son compañías Alemana,
# aparecen en Bavaria, Cundinamarca, Colombia. se eliminan
# LocalXXL empresa alemana, aparece en Bavaria, Magdalena, Colombia se elimina
df = df[~df['Headquarters Location'].str.contains('Bavaria(?!$)')]

# iFut, Inova House 3D, MÃ¡quina de Vendas, Capital Empreendedora y Acceleratus aparecen
# en Brasilia, Distrito Especial, Colombia.  son compañías de Brasil, se eliminan
df = df[~df['Headquarters Location'].str.contains('Brasilia')]

#####################################################
# Biometric Update, Boardwalk REIT, ChartBRAIN, LIFT Partners, canada Toronto, calgary
# Pulse Software, compañia australiana, pero aparece en Canada colombia

# se eliminan todas las Empresas de Canada, Cundinamarca, Colombia,
# excepto Qinaya y Partsium
# Ejemplo de Drop con 3 condiciones !!!
df=df.drop(df[(df['Headquarters Location'] == 'Canadá, Cundinamarca, Colombia')
        & ((df['Organization Name'] != 'Qinaya')  & (df['Organization Name'] != 'Partsium'))].index)
#####################################################
```

Data Warehouse

```
# mRisk, BattleBit, Cencosud Shopping Centers compañías de Chile # Chile, Huila, Colombia  se Eliminan
df.drop(df[df['Headquarters Location']=='Chile, Huila, Colombia'].index, inplace =True)

# Cocodrilo Dog es de Melbourne y aparece en Cundinamarca, Distrito Especial, Colombia # se elimina
df.drop(df[df['Organization Name']=='Cocodrilo Dog'].index, inplace =True)

# Homeland Security Careers es de USA y esta en El Paso, Cesar, Colombia
df.drop(df[df['Headquarters Location']=='El Paso, Cesar, Colombia'].index, inplace =True)

# WindoTrader USA, aparece como Las Vegas, Sucre, Colombia # se elimina
df.drop(df[df['Headquarters Location']=='Las Vegas, Sucre, Colombia'].index, inplace =True)

# Onyx, Elm, Photogramy, Ferrisland, BeyondROI sede en Los Angeles, Huila, Colombia # se eliminan
df.drop(df[df['Headquarters Location']=='Los Angeles, Huila, Colombia'].index, inplace =True)

# Peris Costumes, Cositas de España, Esri, Pirsonal, Barrabes, Carousel Group, WORLD COMPLIANCE ASSOCIATION
# Clupik, Mobile Dreams ltd., Acqualia, LoAlkilo, Codekai, Vitriovr, El inmobiliario mes a mes
# Core Business Consulting, Renewable Energy Magazine, datosmacro, 1001talleres, GGBOX
# Puravida Software y Consultia IT con sede en Madrid, Distrito Especial, Colombia
df.drop(df[df['Headquarters Location']=='Madrid, Distrito Especial, Colombia'].index, inplace =True)

# Advanet (Japon) y Truland Service Corporation en USA. aparecen Maryland, Cundinamarca, Colombia
df.drop(df[df['Headquarters Location']=='Maryland, Cundinamarca, Colombia'].index, inplace =True)

# POC Network Technologies (TransactRx), Alert Global Media.  sede Miami, Magdalena, Colombia
df.drop(df[df['Headquarters Location']=='Miami, Magdalena, Colombia'].index, inplace =True)
```

Data Warehouse

```
# Sicartsa sede en Mexico, Huila, Colombia
df.drop(df[df['Headquarters Location']=='México, Huila, Colombia'].index, inplace =True)

# 24marine, Merkadoo sede en Panama, Magdalena, Colombia
df.drop(df[df['Headquarters Location']=='Panamá, Magdalena, Colombia'].index, inplace =True)

# Agros, Downloadperu.com, Mesa 24/7, Dconfianza, Caja Los Andes, Snacks America Latina Peru S.R.L.
# Pandup, Apprende sede en Peru, Valle del Cauca, Colombia
df.drop(df[df['Headquarters Location']=='Perú, Valle del Cauca, Colombia'].index, inplace =True)

# Big Picture Solutions en Florida, Santander, Colombia
df.drop(df[df['Headquarters Location']=='Florida, Santander, Colombia'].index, inplace =True)
```

Data Warehouse

```
##########################################################################
################## Depuración de nombres de ciudad incorrectos a correctos
##########################################################################

# La compañia Savy tiene sede Usaquen, se cambia a Bogota
df= df.replace({"Usaquén, Distrito Especial, Colombia":'Bogotá, Distrito Especial, Colombia'})

# Se cambia El Herald por El Heraldo
# Tiene Sede Atlantico, Magdalena, Colombia se cambia a Barranquilla, Atlantico, Colombia
df= df.replace({"El Herald":'El Heraldo'})
df= df.replace({"Atlántico, Magdalena, Colombia":'Barranquilla, Atlantico, Colombia'})

# compañia Monolegal es de tunja y aparece Boyaca, Boyaca, Colombia
df= df.replace({"Boyacá, Boyaca, Colombia":'Tunja, Boyacá, Colombia'})

# Celotor es de cali aparece como Colombiano, Magdalena, Colombia
df= df.replace({"Colombiano, Magdalena, Colombia":'Cali, Valle del Cauca, Colombia'})

# Santiago De Cali, Valle del Cauca, Colombia por Cali, Valle del Cauca, Colombia
df= df.replace({"Santiago De Cali, Valle del Cauca, Colombia":'Cali, Valle del Cauca, Colombia'})

# Qinaya, compañia colombiana
#https://www.wradio.com.co/noticias/tecnologia/qinaya-el-emprendimiento-que-convierte-cualquier-televisor-en-un
computador/20210301/nota/4113498.aspx
# https://www.youtube.com/watch?v=XBgbwUxkatc
# Canada, Cundinamarca, Colombia vs Bogota, Distrito Especial, Colombia
# Replace with condition
df.loc[(df['Organization Name'] == 'Qinaya'),'Headquarters Location']='Bogotá, Distrito Especial, Colombia'
```

Data Warehouse

```
#########################################################################
################## Depuración de nombres de ciudad incorrectos a correctos
#########################################################################

# Partsium
# Partsium, bogota.  El Sitio pone a disposición de los Usuarios un espacio virtual que les permite
# comunicarse mediante el uso de Internet para encontrar una forma de vender o comprar productos y
# servicios. PARTSIUM no es el propietario de los artículos ofrecidos, no tiene posesión de ellos ni
# los ofrece en venta. Los precios de los productos y servicios están sujetos a cambios sin previo aviso.
# website rental to do business
df.loc[(df['Organization Name'] == 'Partsium'),'Headquarters Location']='Bogotá, Distrito Especial, Colombia'
df.loc[(df['Organization Name'] == 'Partsium'),'Industries']='Website rental, Doing business'

# Chiper, SkyFunders, Plastic Surgery Colombia Cias de Bogota y aparecen
# en Cundinamarca, Distrito Especial, Colombia
df= df.replace({"Cundinamarca, Distrito Especial, Colombia":'Bogotá, Distrito Especial, Colombia'})



df= df.replace({"Antioquia, Antioquia, Colombia":'Medellín, Antioquia, Colombia'})
df= df.replace({"Bucaramanga, Cundinamarca, Colombia":'Bucaramanga, Santander, Colombia'})
df= df.replace({"Santander, Bolivar, Colombia":'Bucaramanga, Santander, Colombia'})
df= df.replace({"Cúcuta, Antioquia, Colombia":'Cucuta, Norte de Santander, Colombia'})
df= df.replace({"Popayán, Cordoba, Colombia":'Popayán, Cauca, Colombia'})
```

Data Warehouse

```
##########################################################################
######  Separando columna Headquarters Location en 3 columnas
###### Headquarters Location (igual a Ciudad para no crear otra columna), Departamento y Pais

df3=df["Headquarters Location"].str.split(",", n = 2, expand = True)

####  adicionando las nuevas columnas a df original
df["Headquarters Location"]= df3[0]
df["Departamento"]= df3[1]
df["Pais"]= df3[2]
##########################################################################

# Se elimina espacio al inicio y al final de cada string
df.columns = df.columns.str.strip()

# se renombra nombre columna Organization Name a Organization
df.rename(columns={"Organization Name": 'Organization'}, inplace=True)

# con estos cambios hasta aquí, se genera dataframe dfx = df
dfx = df.copy()

# Se eliminan las columnas Website, Twitter, Facebook y Linkedin
del dfx["Website"]
del dfx["Twitter"]
del dfx["Facebook"]
del dfx["LinkedIn"]
```

Data Warehouse

```
##########################################################################

# CBrank tiene datos no numericos, se convierten a numeros
dfx["CBRank"] = dfx["CBRank"].str.replace(r'\D', '')
dfx["CBRank"] = pd.to_numeric(dfx["CBRank"])

# Number of Articles tiene datos no numericos, se convierten a numeros
dfx[["Number of Articles"]] = dfx[["Number of Articles"]].fillna('')
dfx["Number of Articles"] = dfx["Number of Articles"].str.replace(r'\D', '')
dfx["Number of Articles"] = pd.to_numeric(dfx["Number of Articles"])

# CBrank Organization tiene datos no numericos, se convierten a numeros
dfx[["CB Rank (Organization)"]] = dfx[["CB Rank (Organization)"]].fillna('')
dfx["CB Rank (Organization)"] = dfx["CB Rank (Organization)"].str.replace(r'\D', '')
dfx["CB Rank (Organization)"] = pd.to_numeric(dfx["CB Rank (Organization)"])

dfx = dfx.drop(['Contact Email'], axis=1) # elimina columna Contact Email
dfx = dfx.drop(['Phone Number'], axis=1)  # elimina columna Phone Number
dfx = dfx.drop(['Full Description'], axis=1) # elimina columna Full Description
```

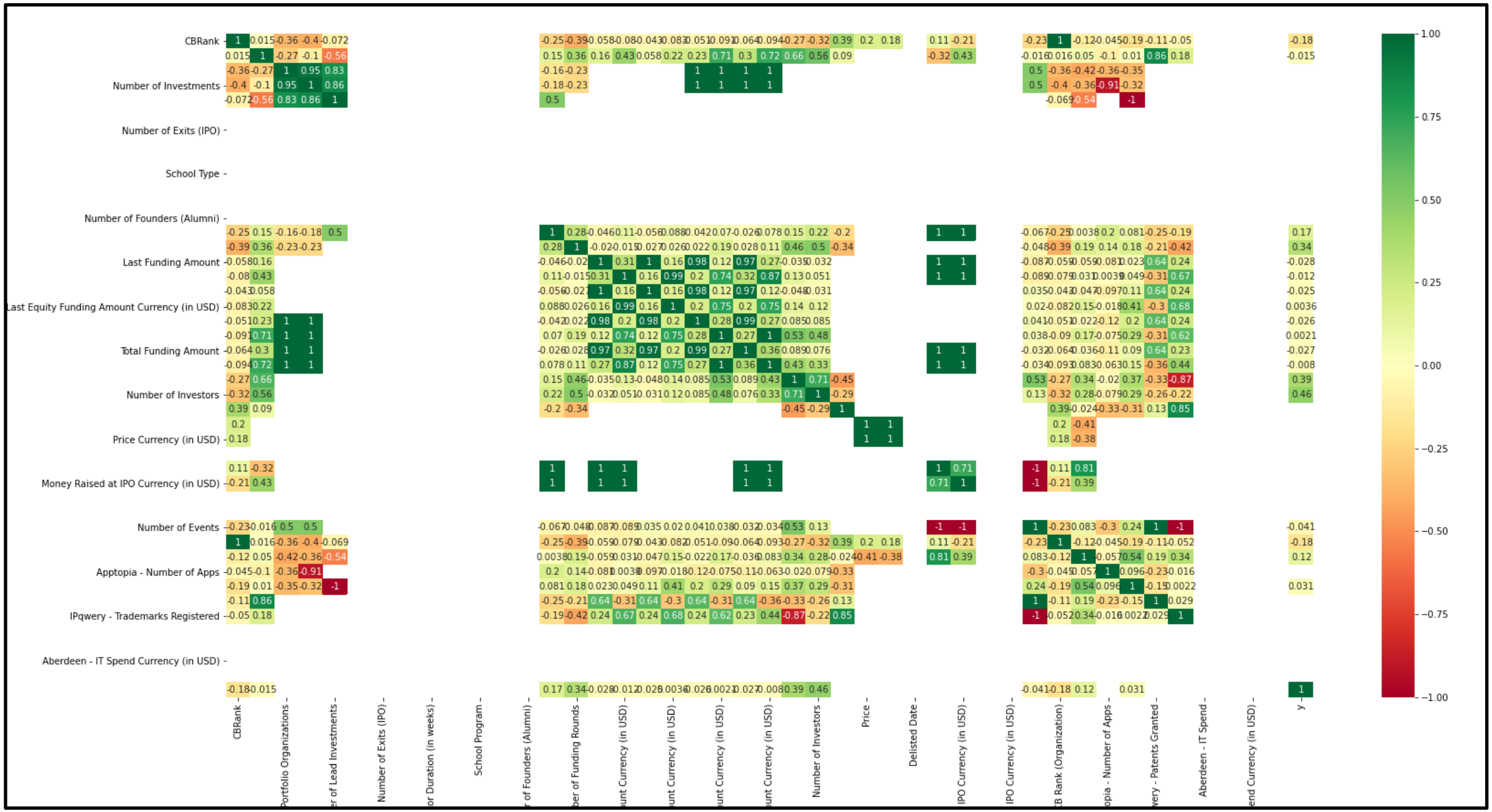
Data Warehouse

```
##########################################################################

#Funcion que corrige espacios
def correct_word(word):

    new_word = word.split()[0]
    return new_word

#Aplicandon la funcion para la columna Departamento
dfx['Departamento'] = dfx['Departamento'].apply(correct_word)

#Aplicandon la funcion para la columna Pais
dfx['Pais'] = dfx['Pais'].apply(correct_word)

#Cambiar las siguientes 2 variable a formato fecha
dfx['Last Funding Date'] = pd.to_datetime(dfx['Last Funding Date'])
dfx['Founded Date'] = pd.to_datetime(dfx['Founded Date'])

##########################################################################
df2 = dfx.copy()

################## CrunchBase vs Unicorn vs Top 100
df2["y"]=0
intersect3=set(df['Organization']).intersection(set(dftop['Organization'])).intersection(set(dfunicorn['Name']))
len(intersect3)

for l in intersect3:
    df2.loc[df2['Organization'] == l, ['y']] = 1

df2.to_excel("BD_Final.xlsx")
```

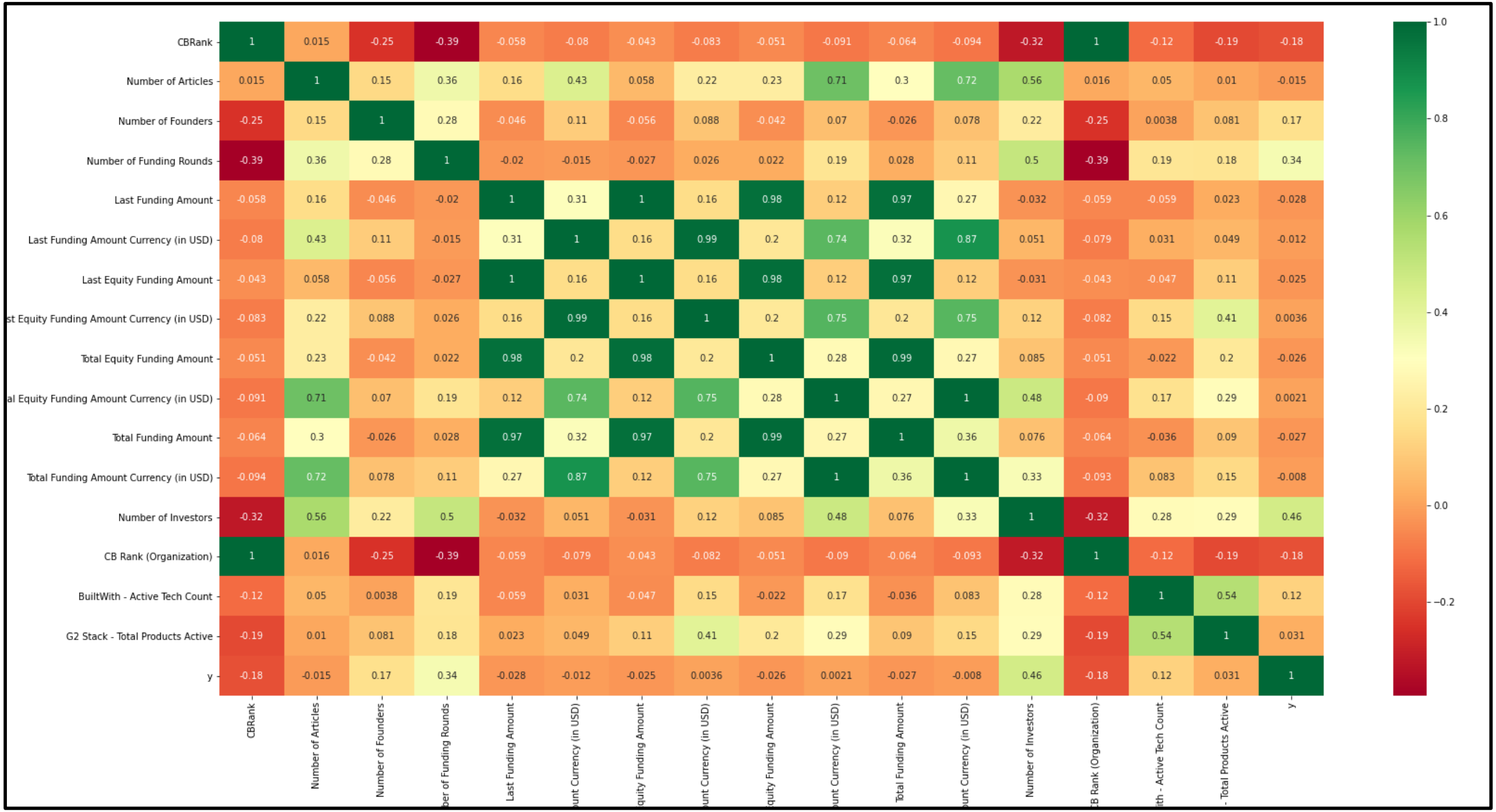
La base de datos para empezar a trabajar en el análisis de datos es el dataframe **df2**

Análisis de Variables

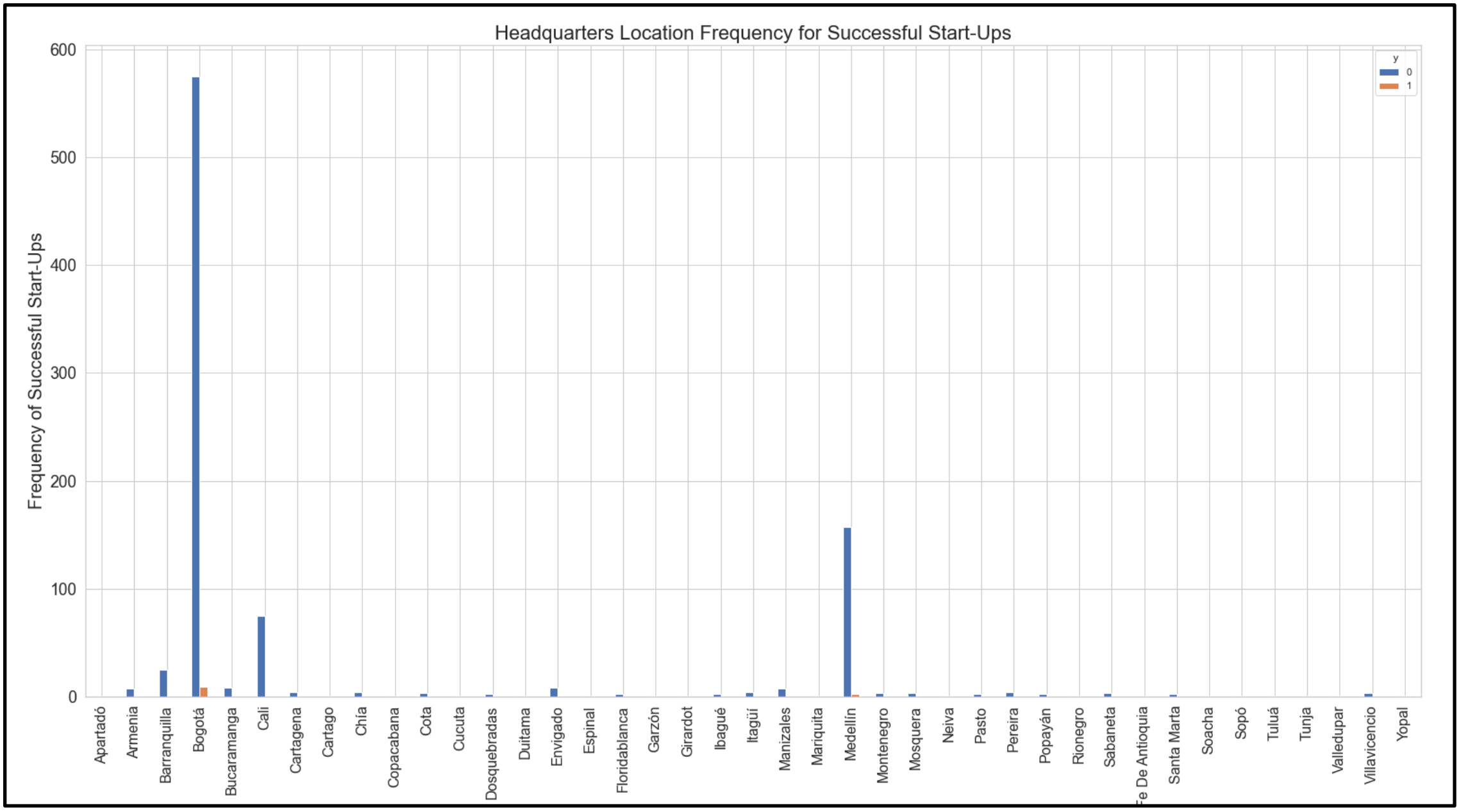
Matriz de Correlación, sin filtro de columnas con mas 80% de nan



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

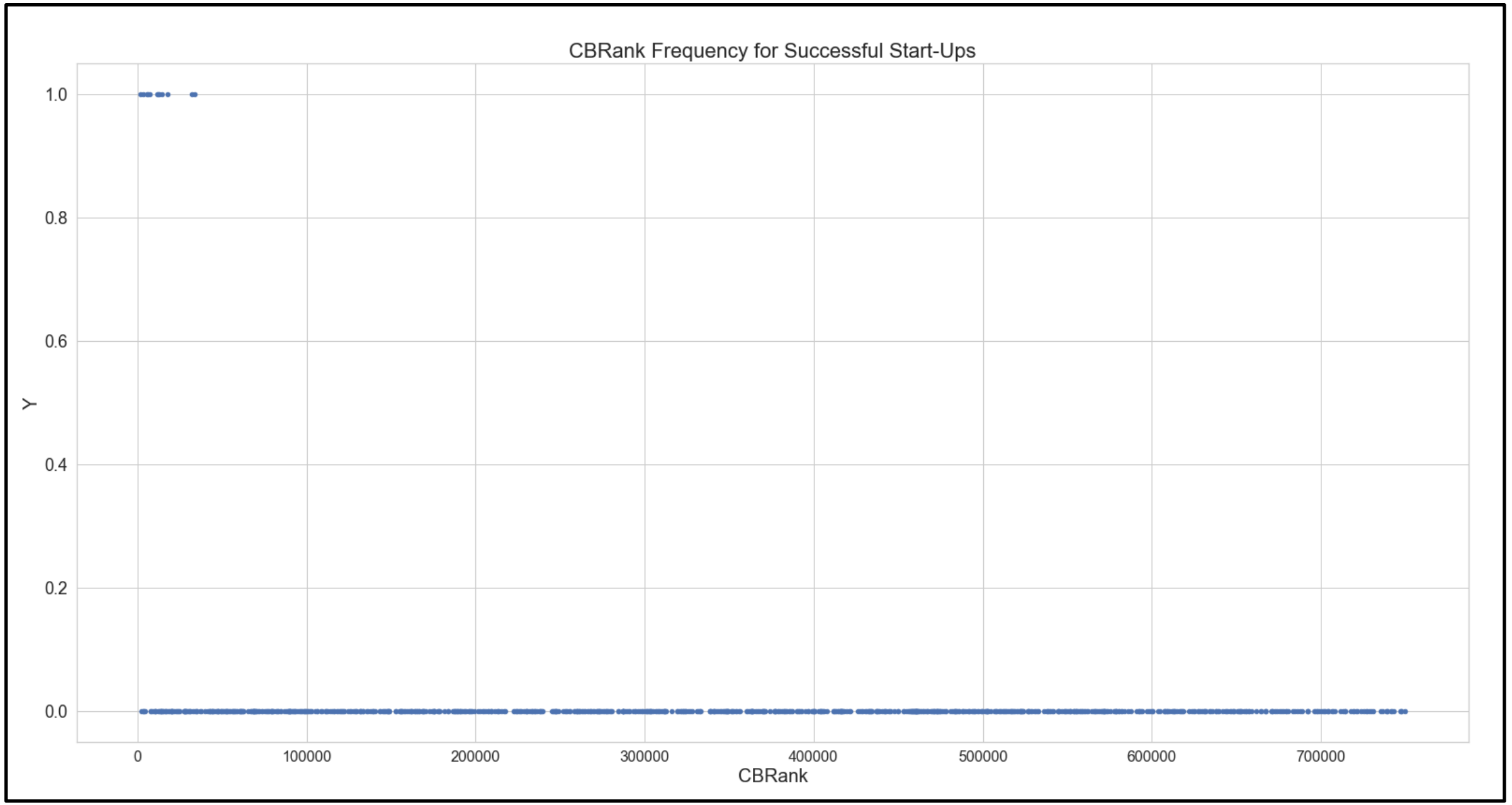


Headquarters Location Frequency for Successful Start-Ups



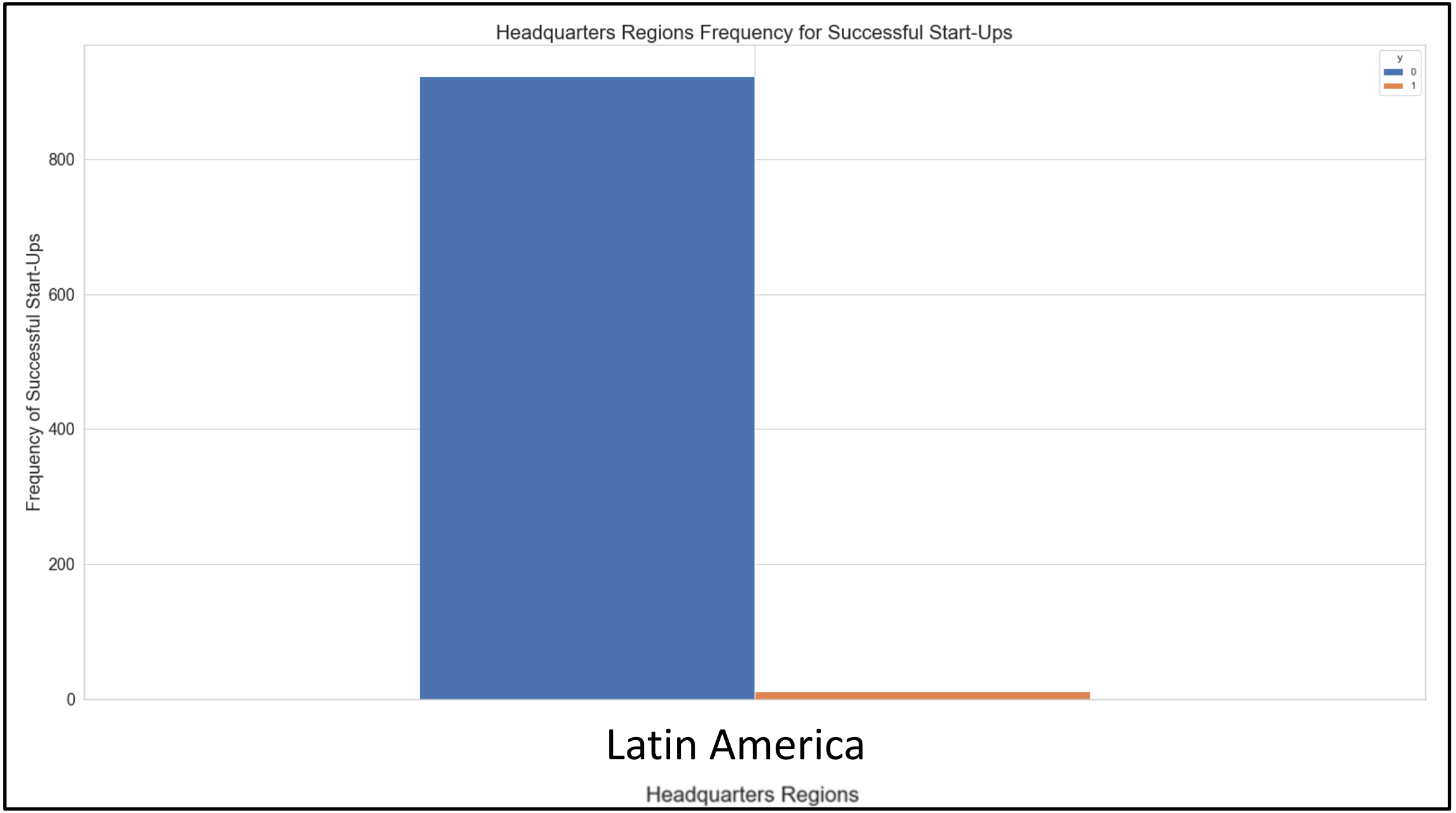
Si Importa la localización de la startup para su éxito. Bogotá y Medellín tienen las Start-ups con éxito, sin embargo **se elimina** para evitar sesgo de que sólo hay éxito en Bogotá y Medellín

CBRank vs Successful Start-Ups



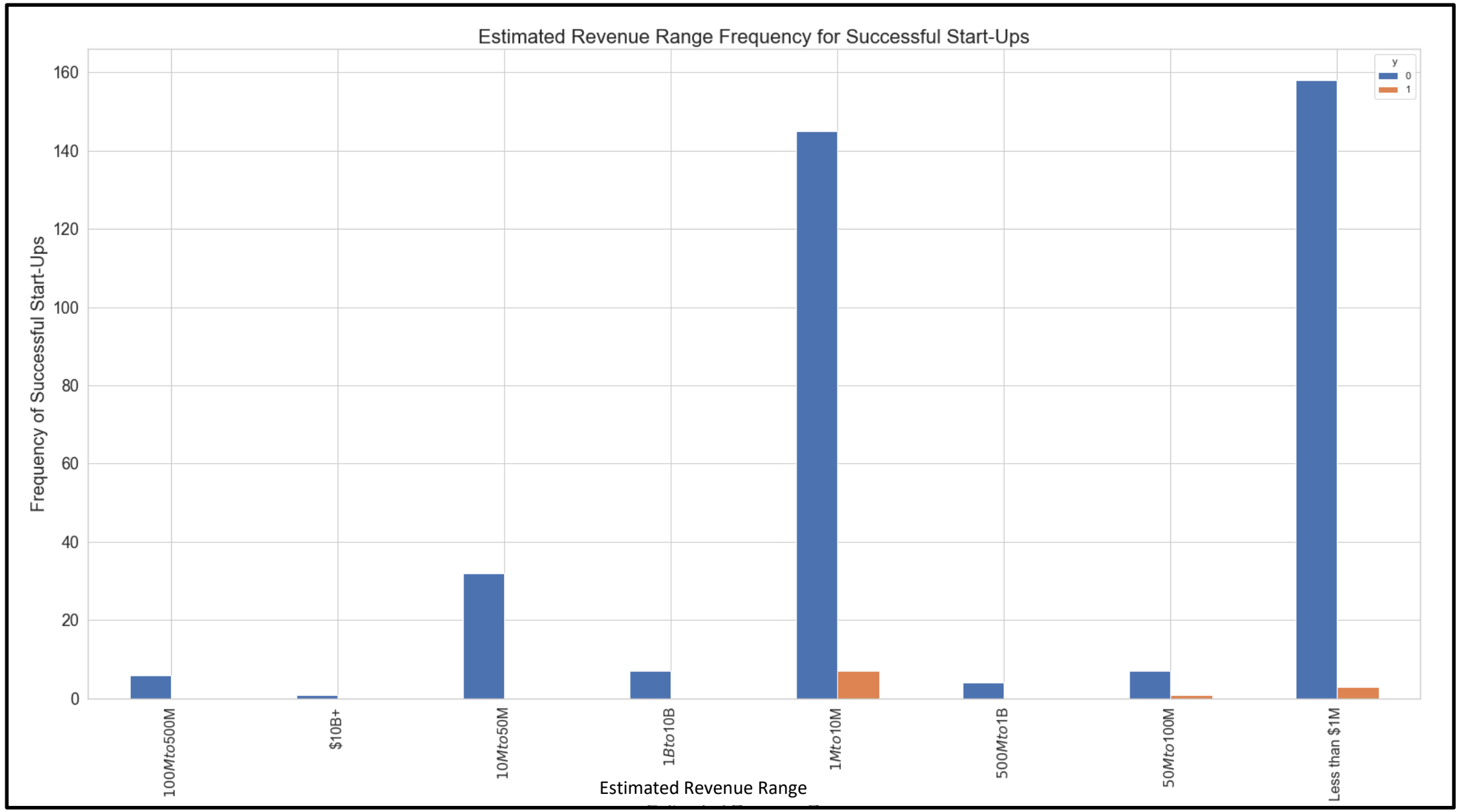
Si Importa Cbrank para éxito de Startups.

Headquarters Regions vs Successful Start-Ups



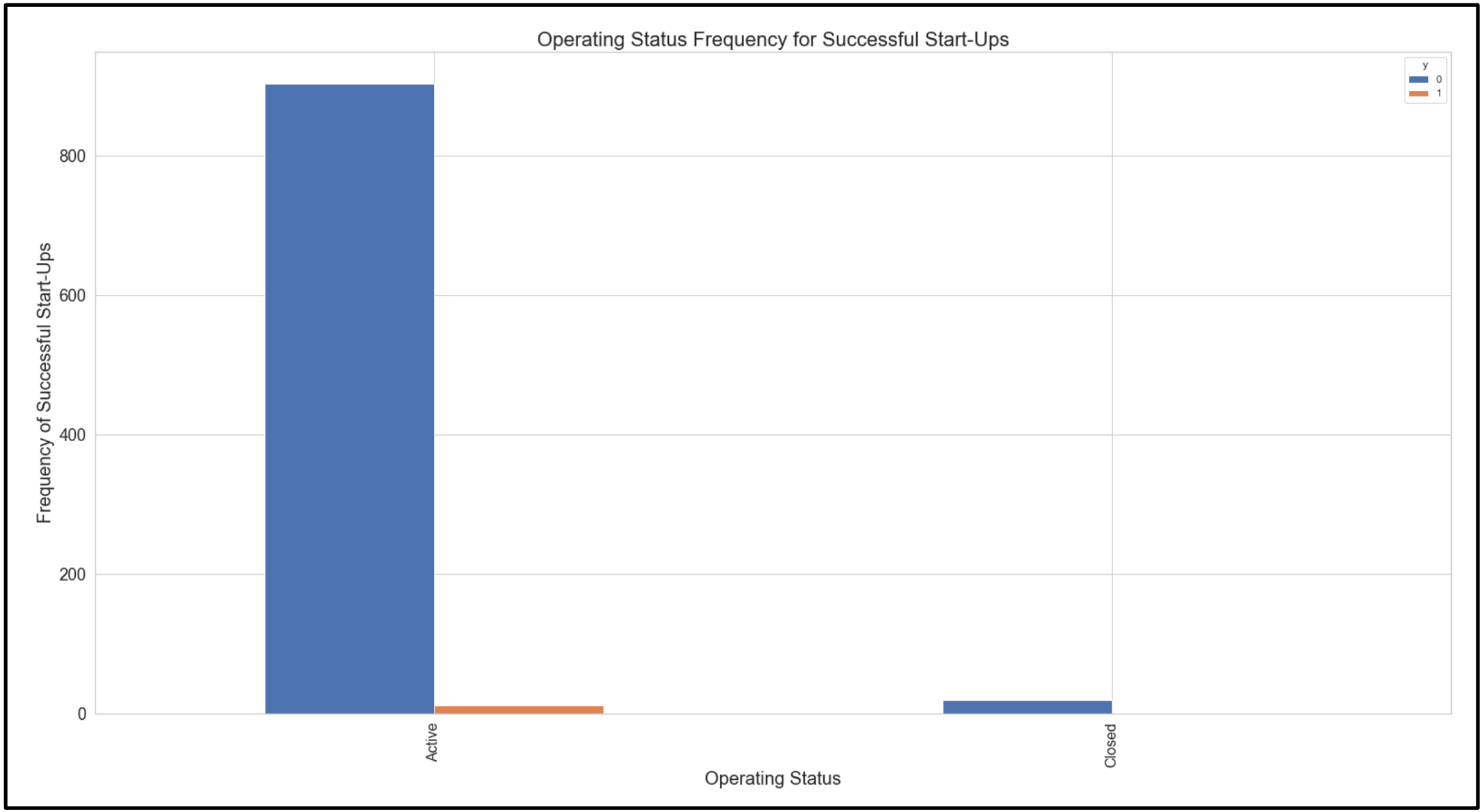
No importa Headquarters Regions para éxito de Startups

Estimated Revenue Range vs Successful Start-Ups



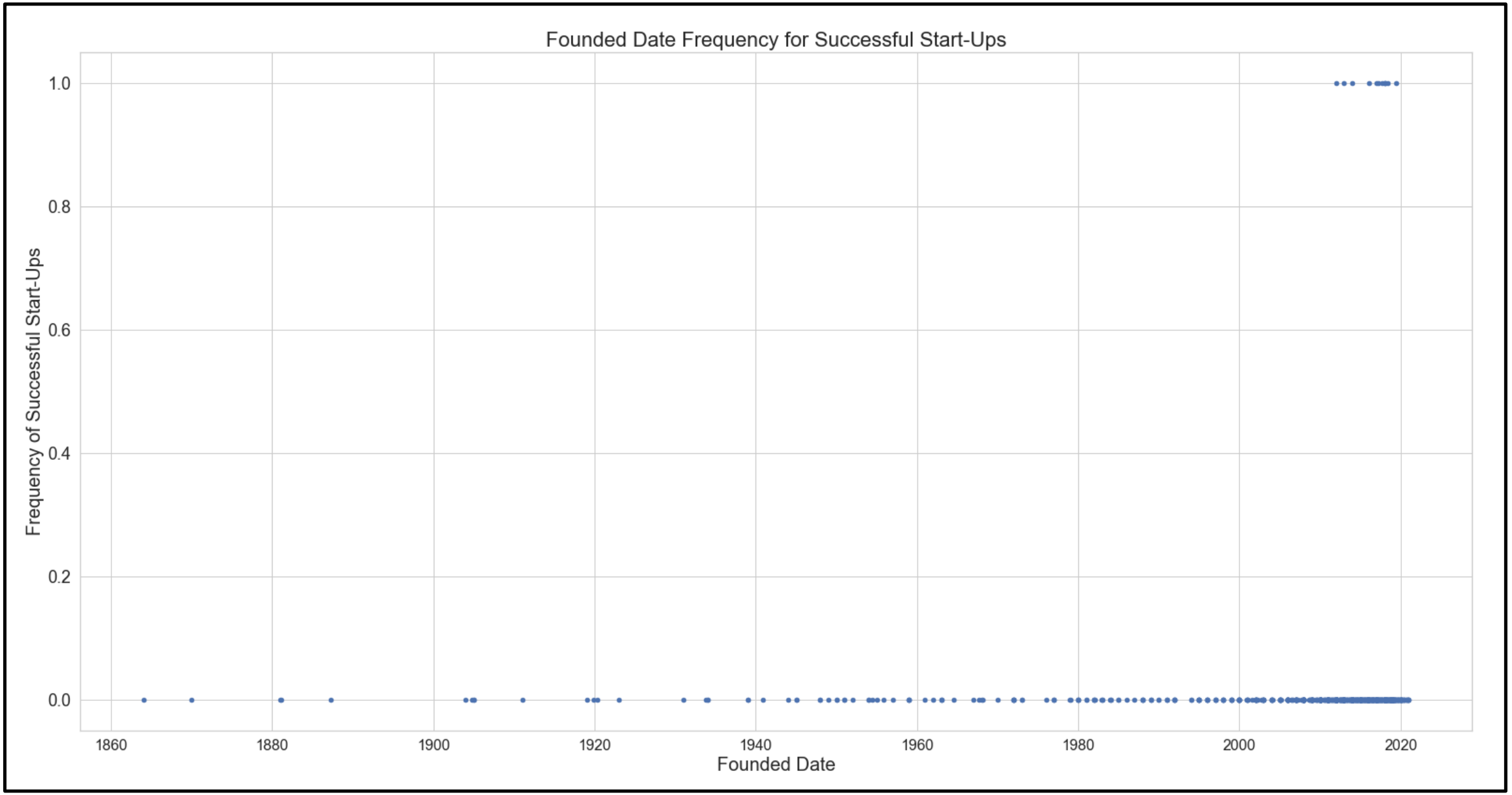
Si Importa Estimated Revenue Range para éxito de Startups.

Operating Status vs Successful Start-Ups



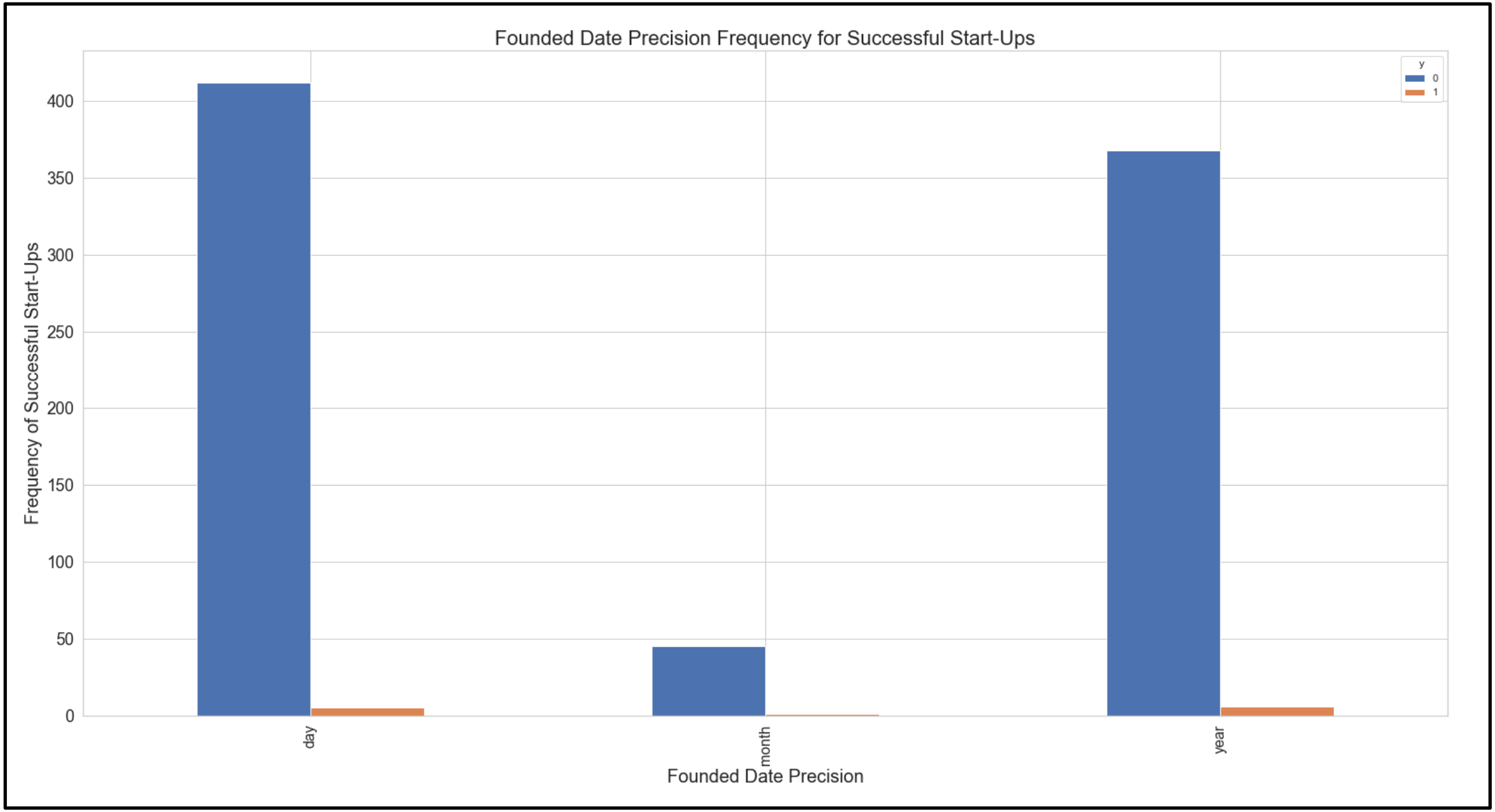
No importa Operating Status para éxito de Startups

Founded Date vs Successful Start-Ups



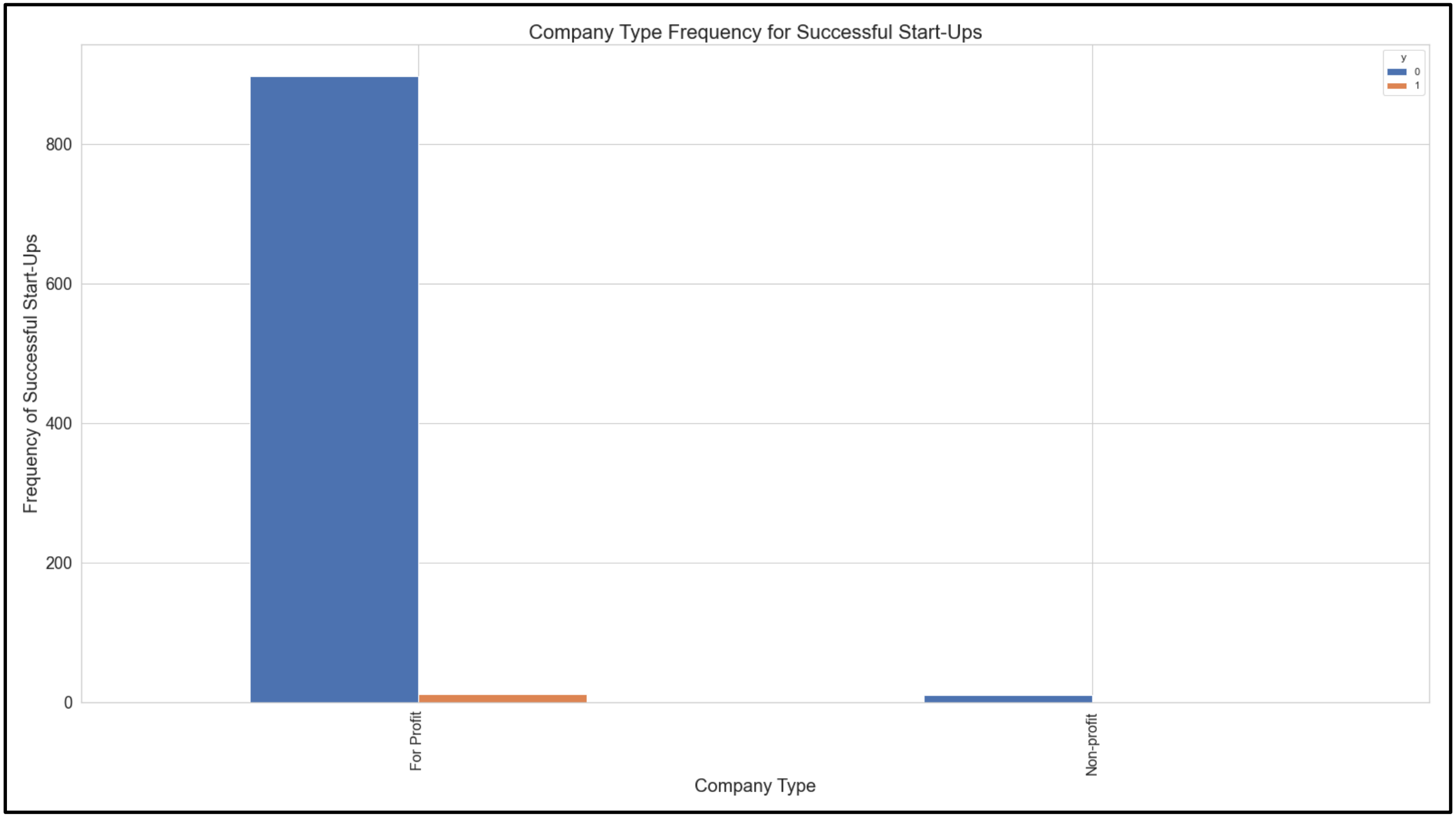
No importa Founded Date para éxito de Startups

Founded Date Precision vs Successful Start-Ups



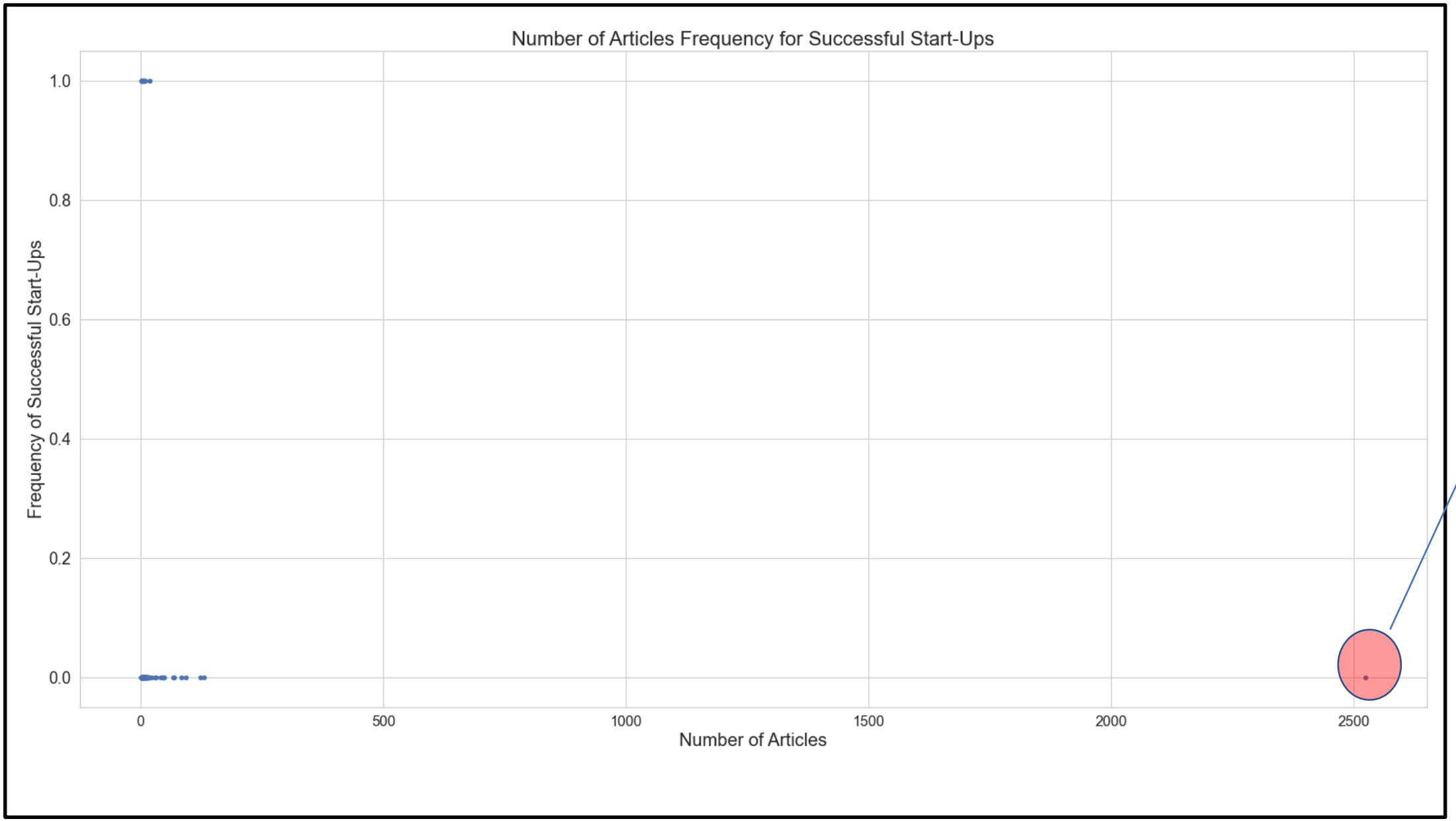
No importa Founded Date Precision para éxito de Startups. Founded date Precision tiene valores de día, mes o año.

Company Type vs Successful Start-Ups



No importa Company Type para éxito de Startups

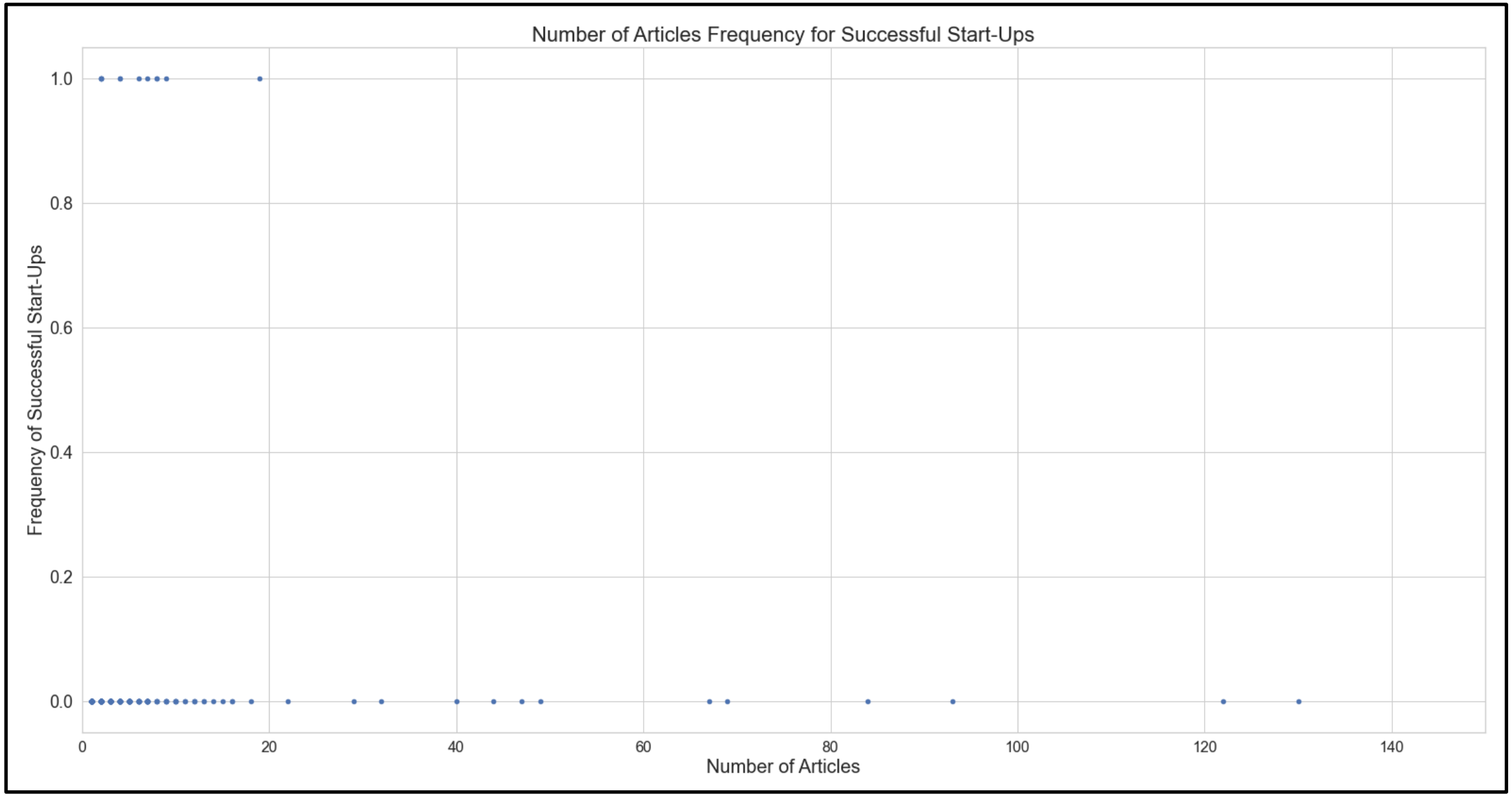
Number of Articles vs Successful Start-Ups



Outlier
o
Dato mal
Introducido
En BD

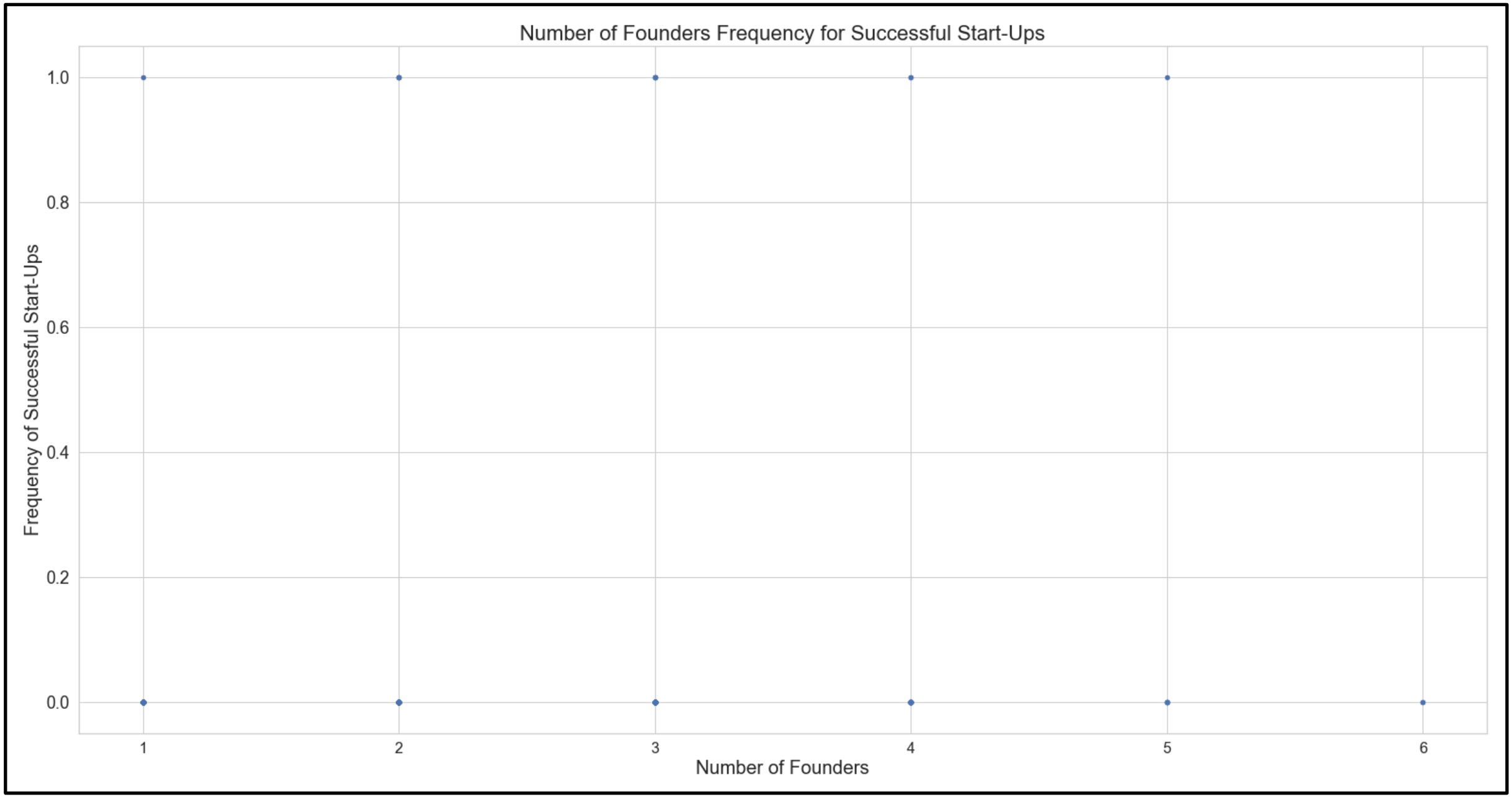
Si importa Number of Articles para éxito de Startups

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



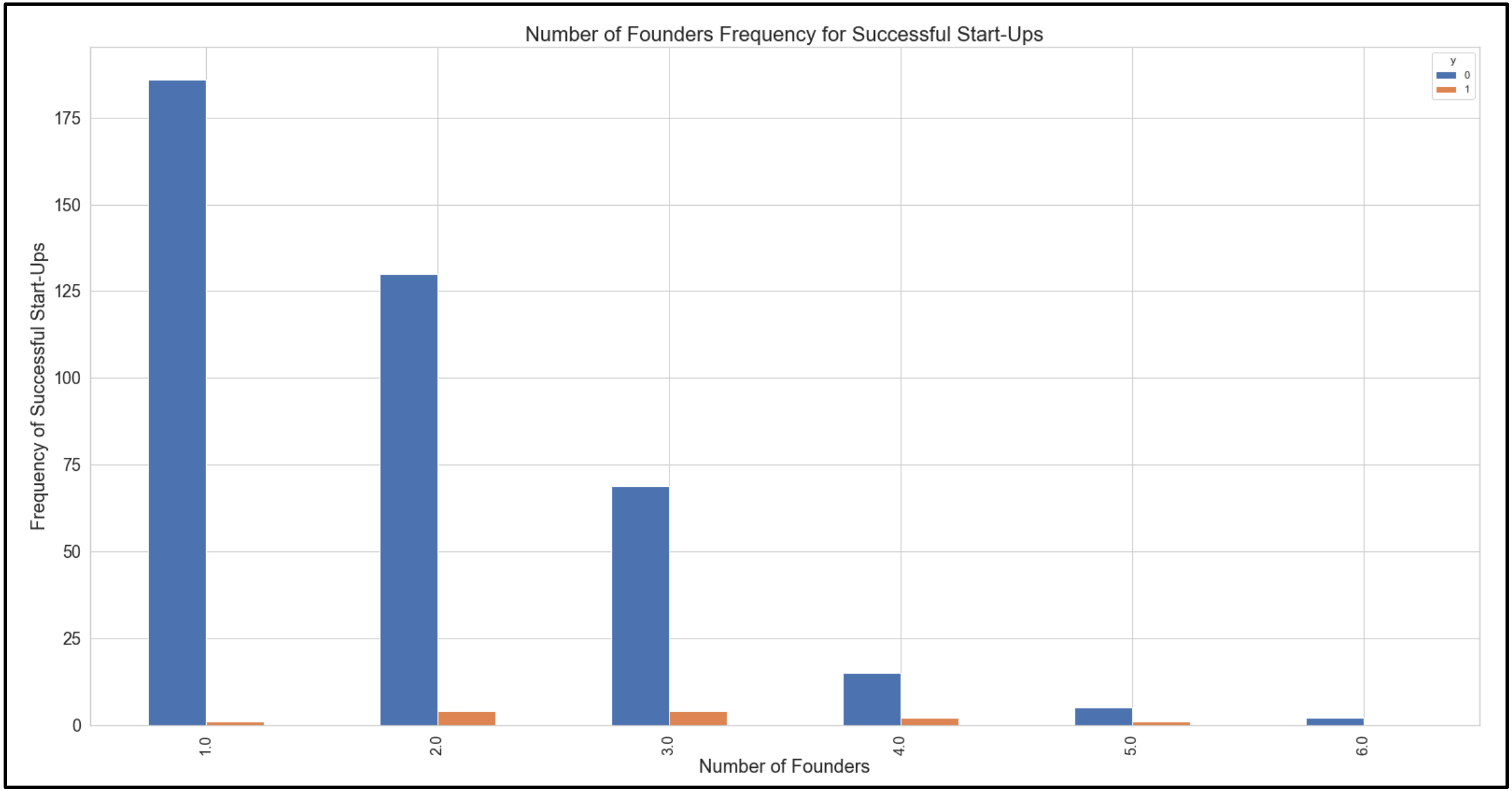
Si importa Number of Articles para éxito de Startups

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



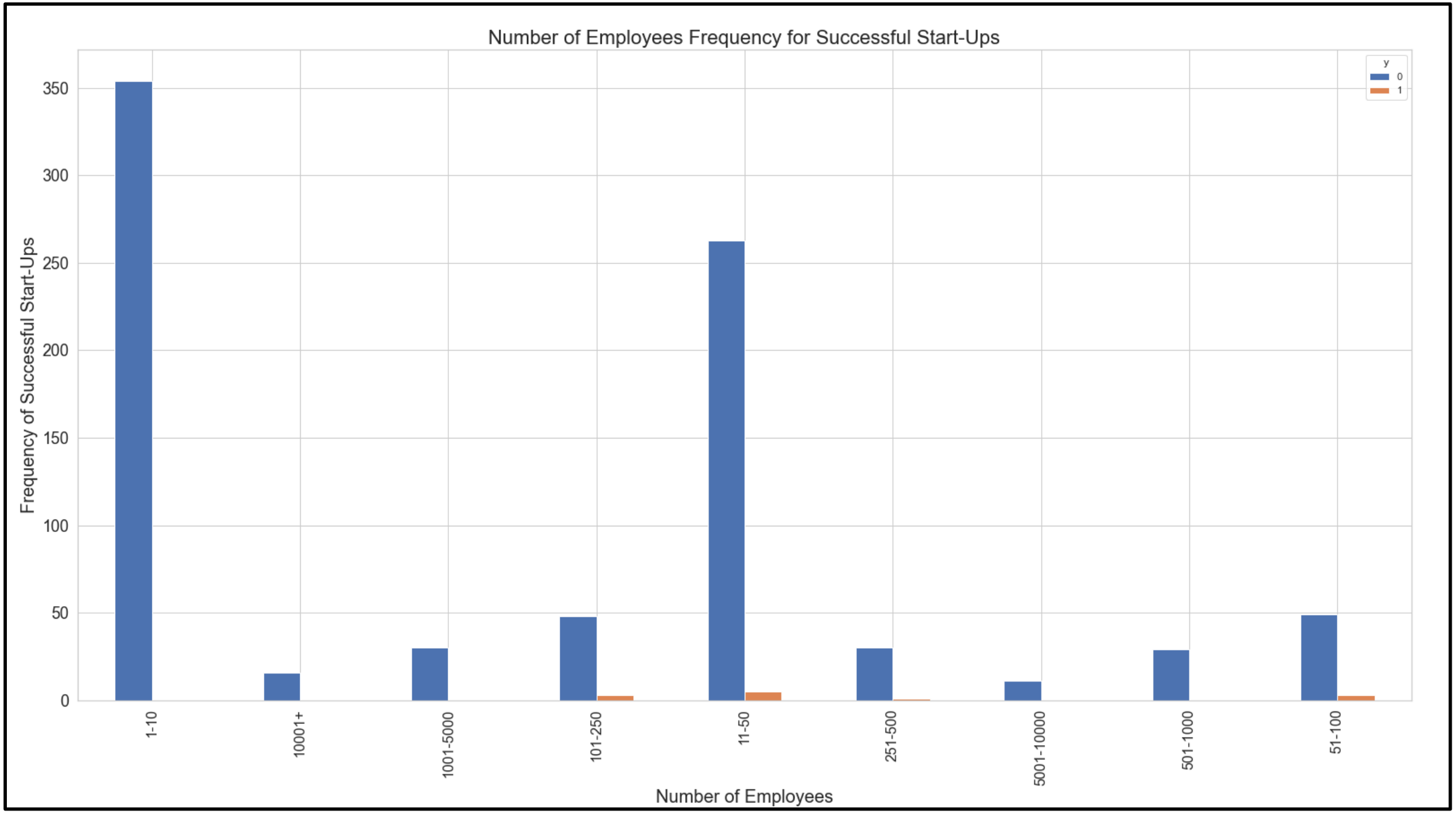
Si importa Number of Founders para éxito de Startups

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



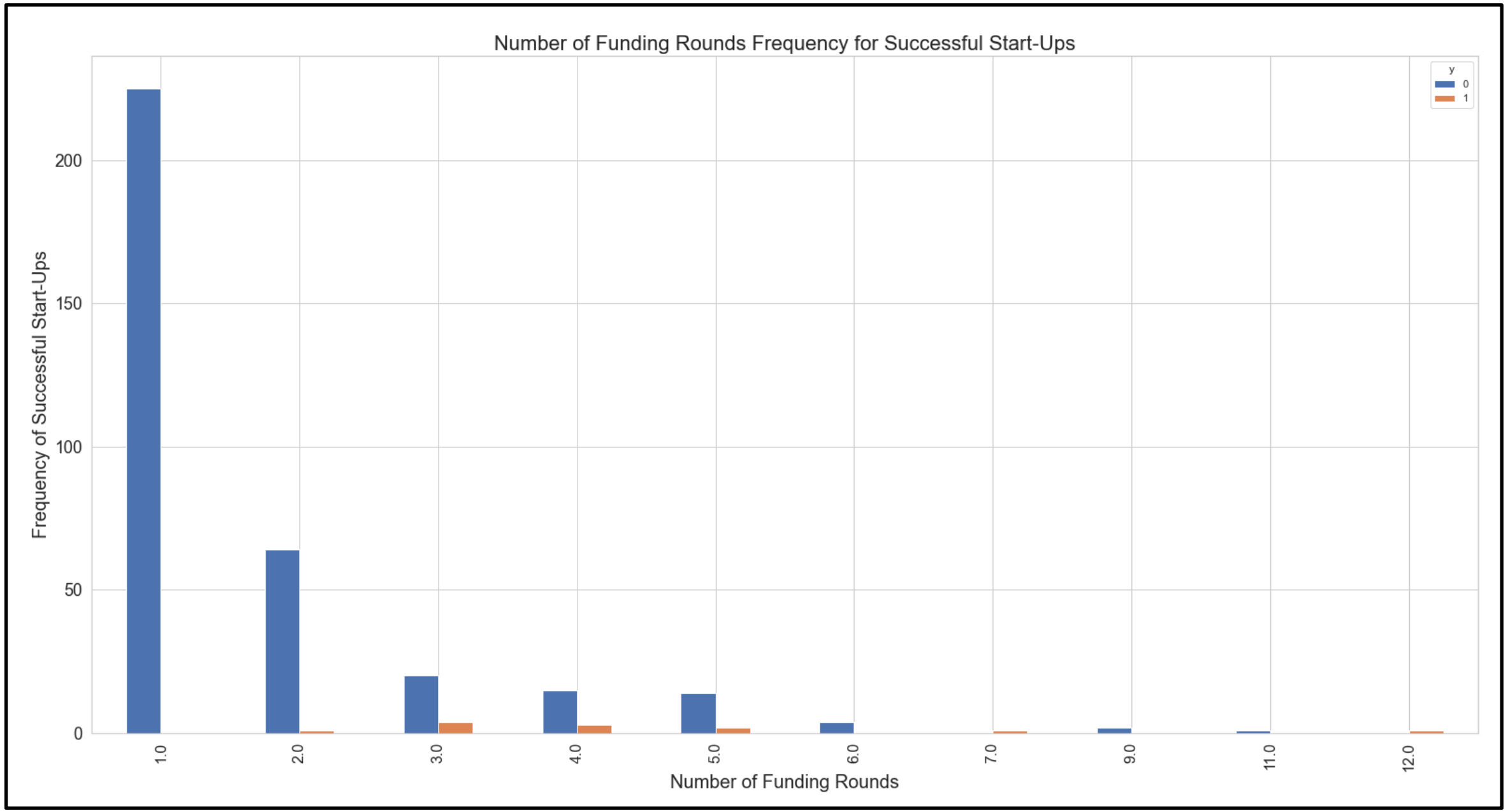
Si importa Number of Founders para éxito de Startups

Number of Employees vs Successful Start-Ups



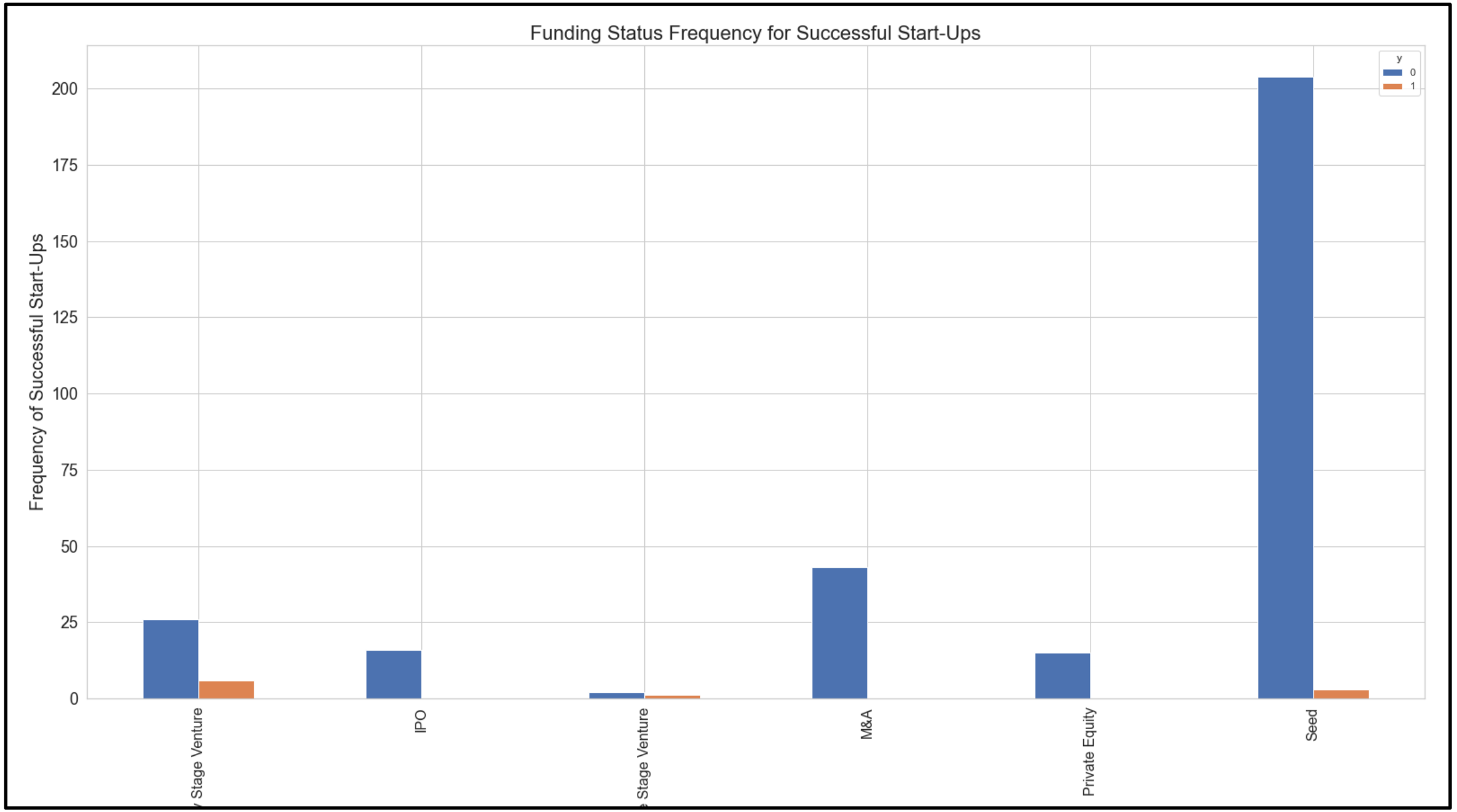
Si importa Number of Employees para éxito de Startups

Number of Funding Rounds vs Successful Start-Ups



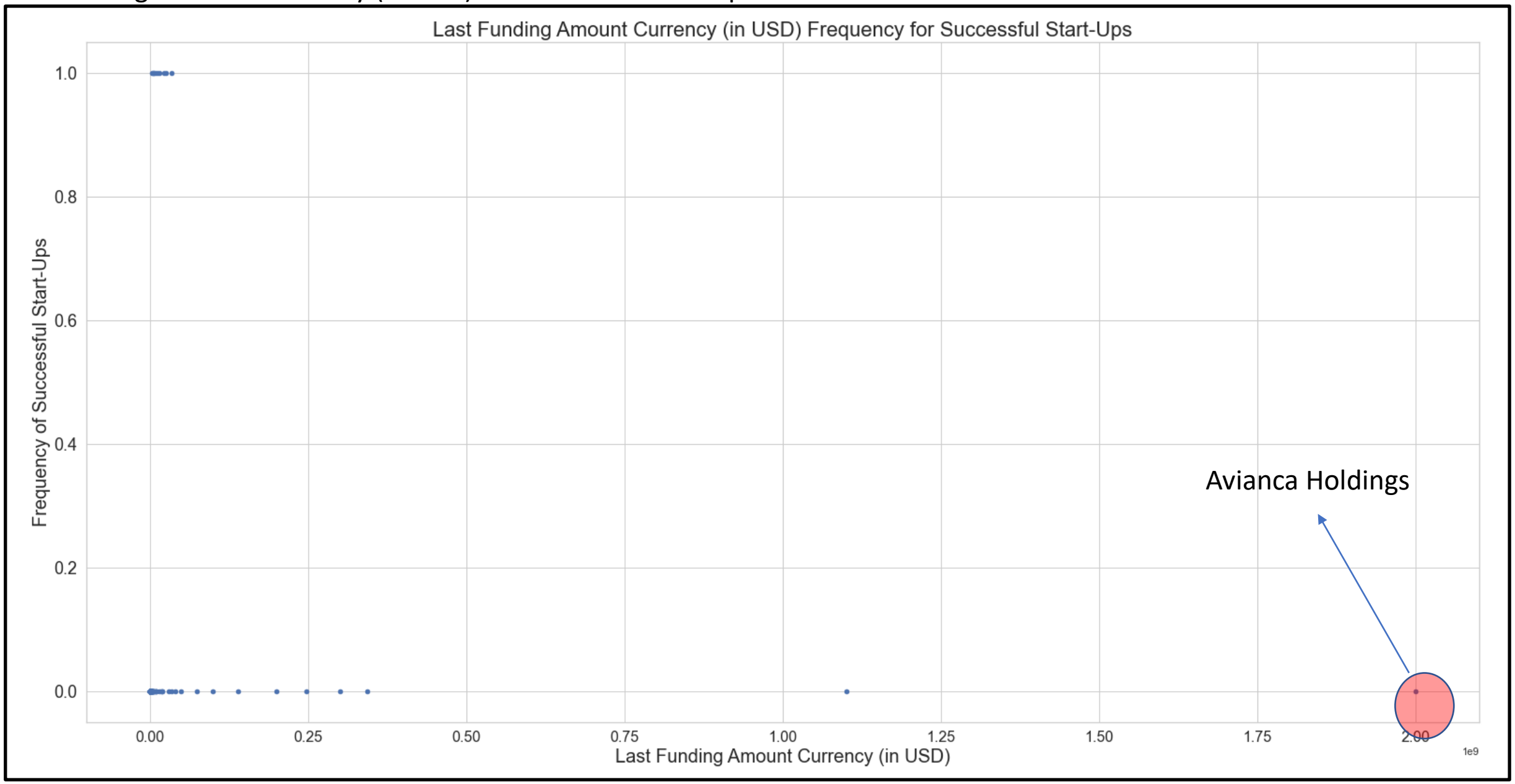
Si importa Number of Funding Rounds para éxito de Startups

Funding Status vs Successful Start-Ups



Si importa Funding Status para éxito de Startups

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

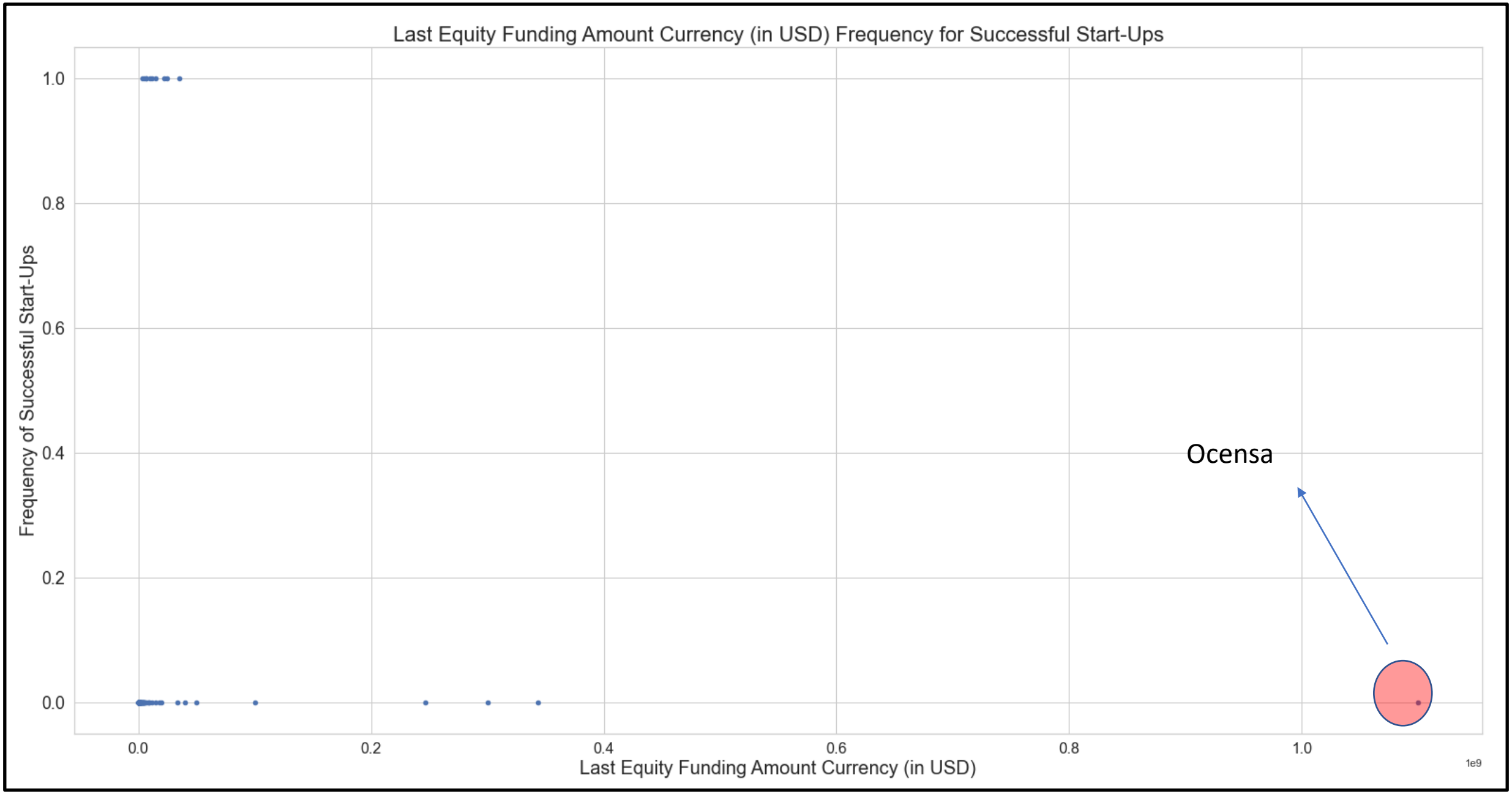


Si importa **Last Funding Amount Currency (in USD)** para éxito de Startups

La variable **Last Funding Amount <u>NO</u> se analiza porque tiene valores numéricos de dinero en diferentes monedas**

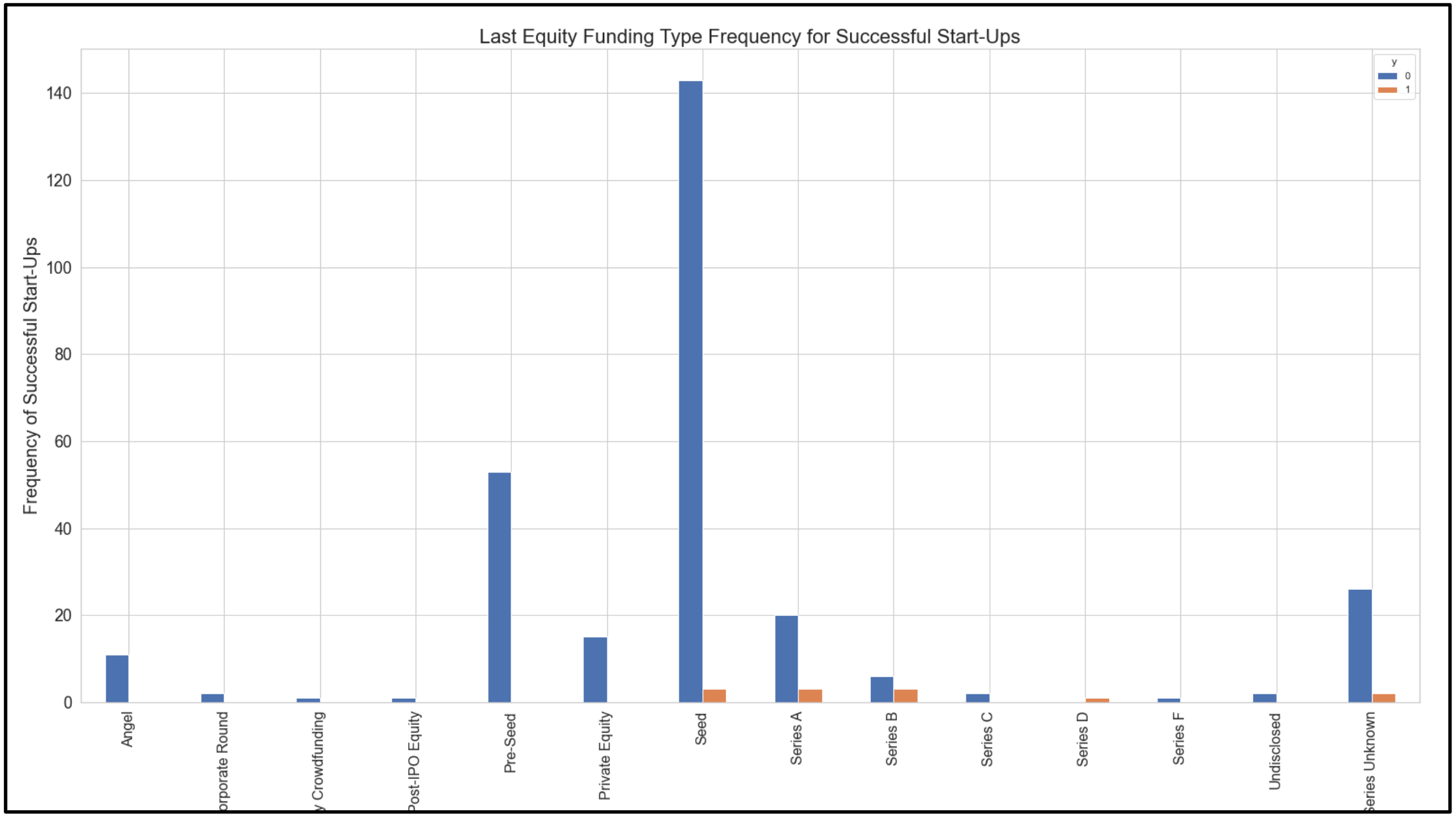
Se analiza Last Funding Amount Currency (in USD)

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



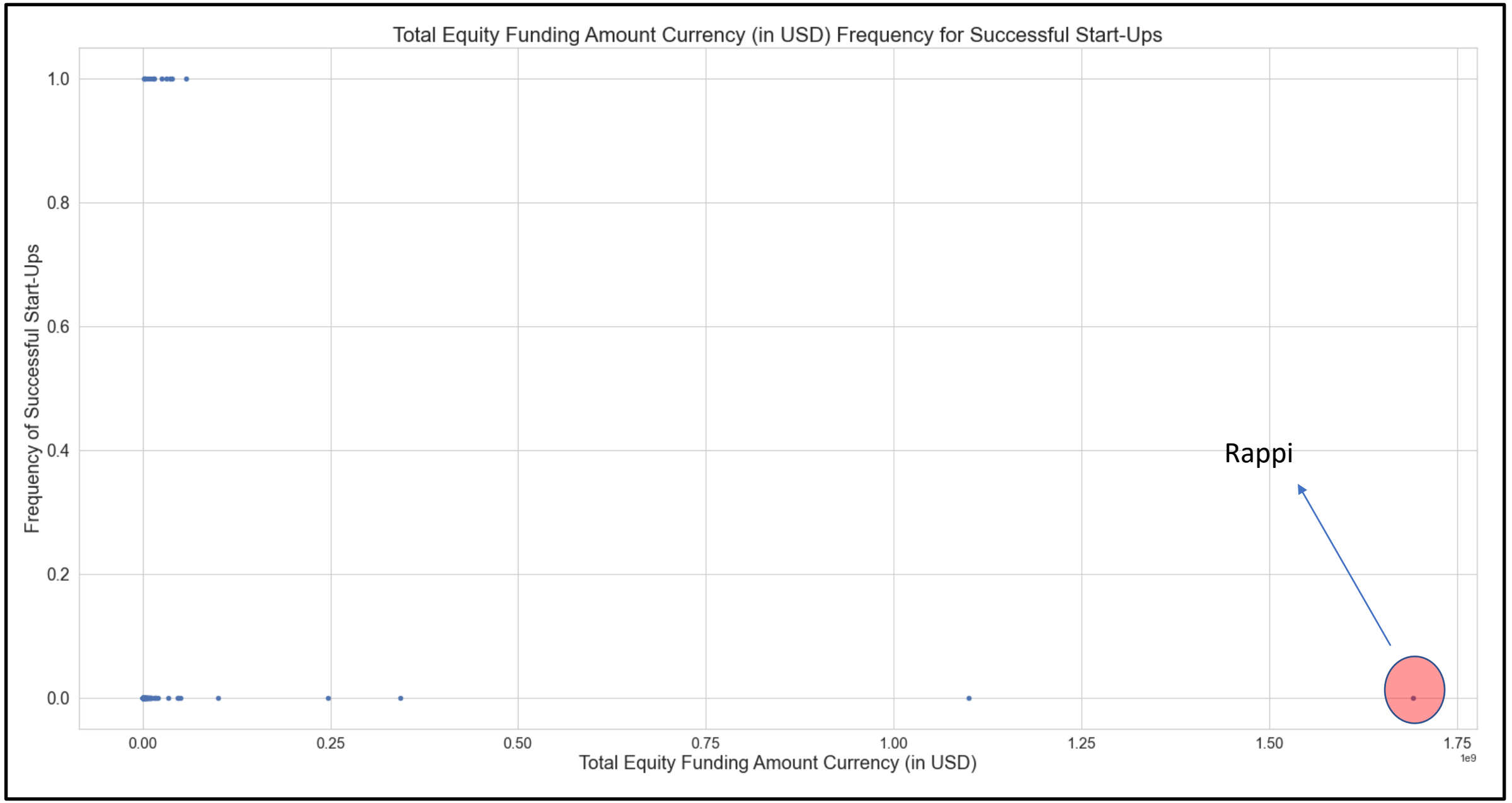
Si importa Last Equity Funding Amount Currency (in USD) para éxito de Startups

Last Equity Funding Type vs Successful Start-Ups



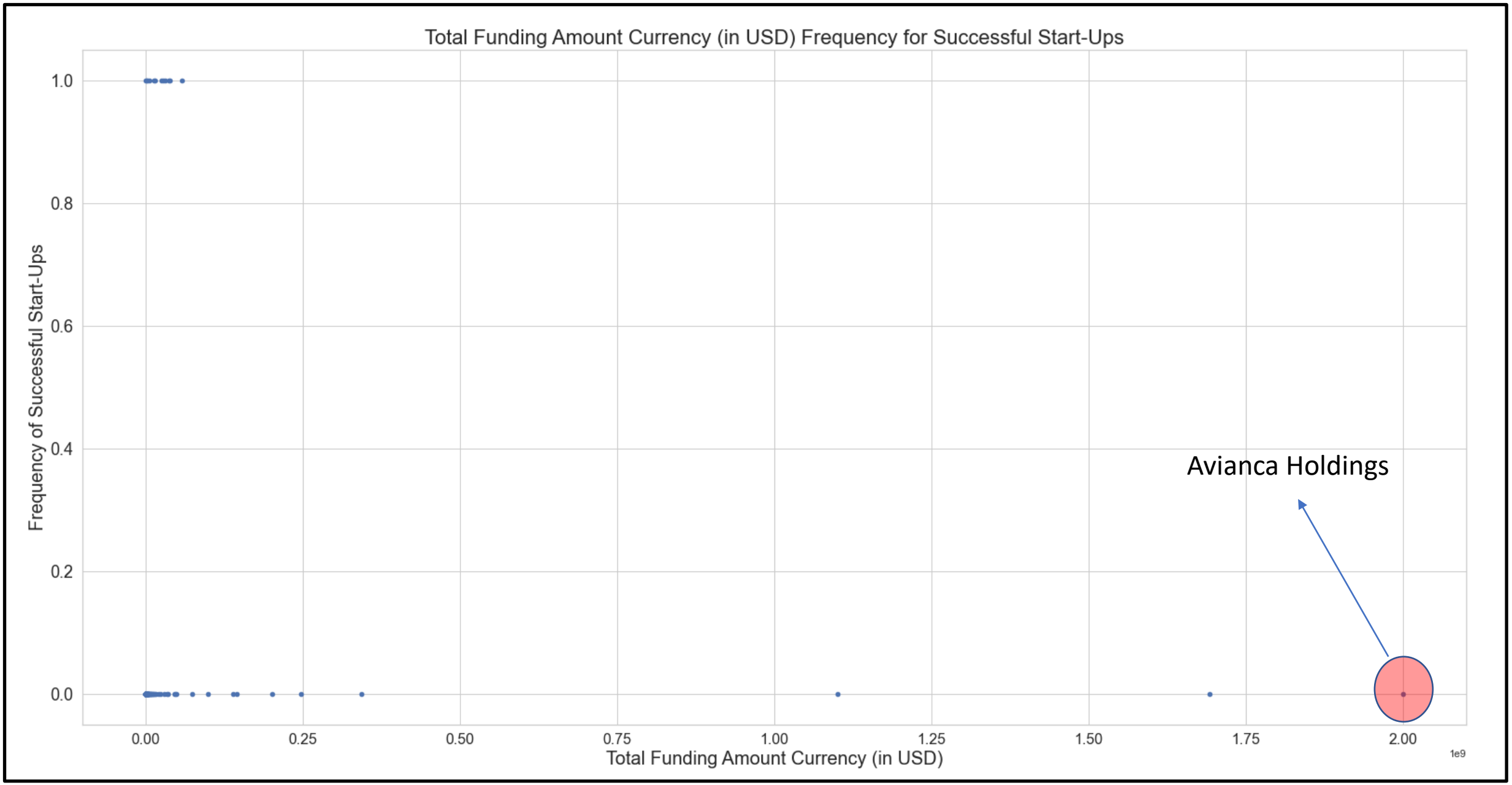
Si importa Last Equity Funding Type para éxito de Startups

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



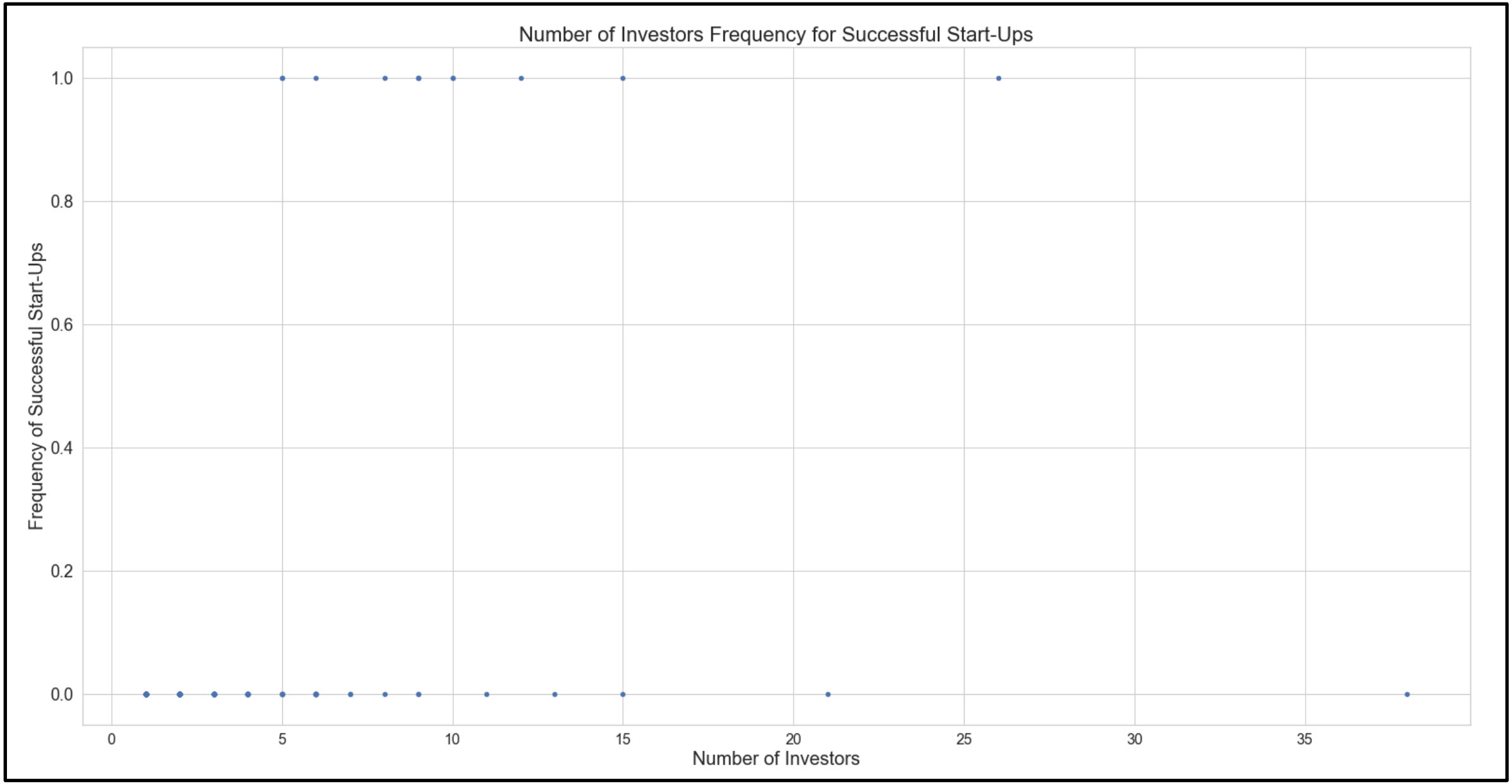
Si importa Total Equity Funding Amount Currency (in USD) para éxito de Startups

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



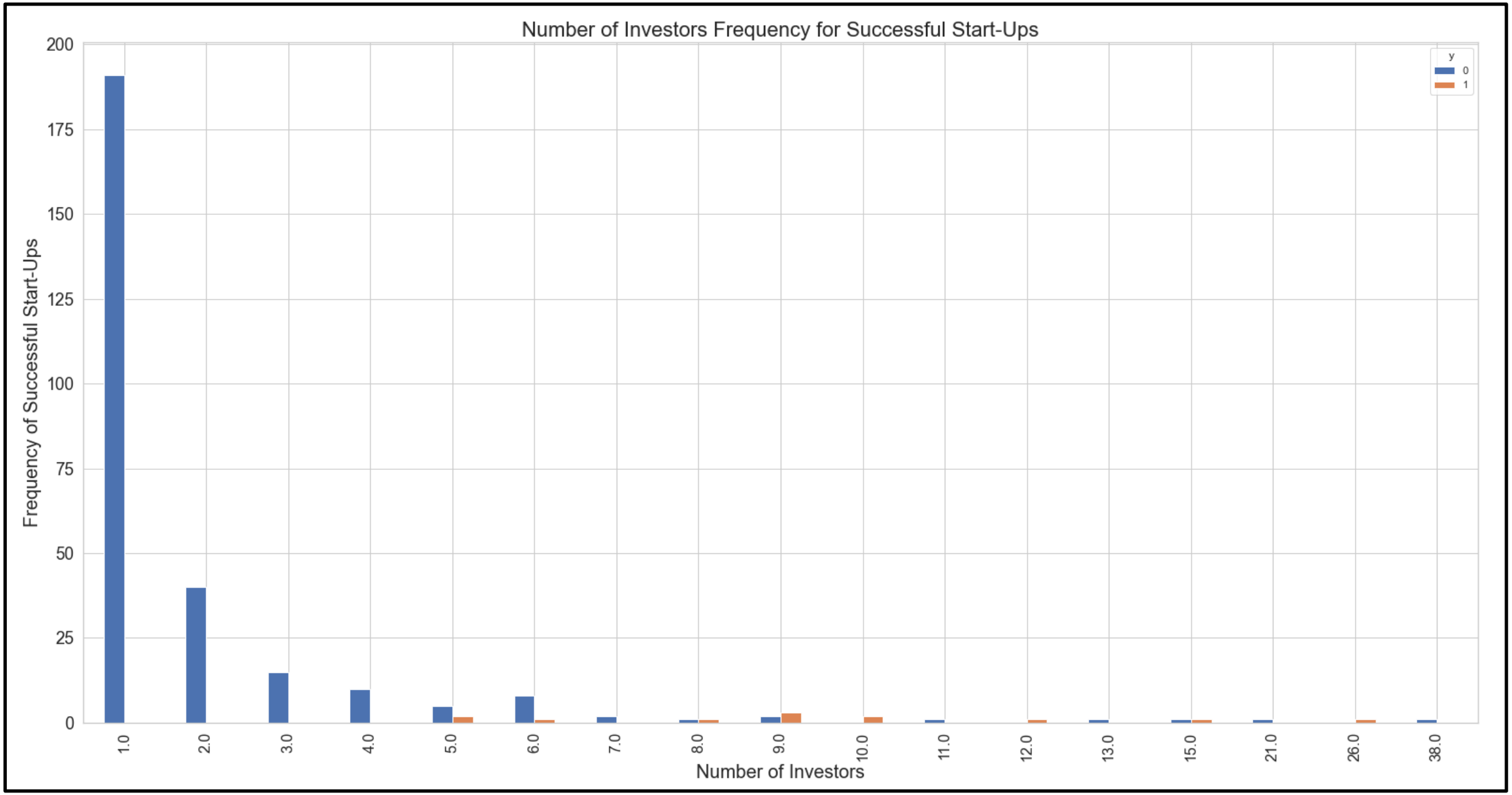
Si importa Total Funding Amount Currency (in USD) para éxito de Startups

Number of Investors vs Successful Start-Ups (Númerica)



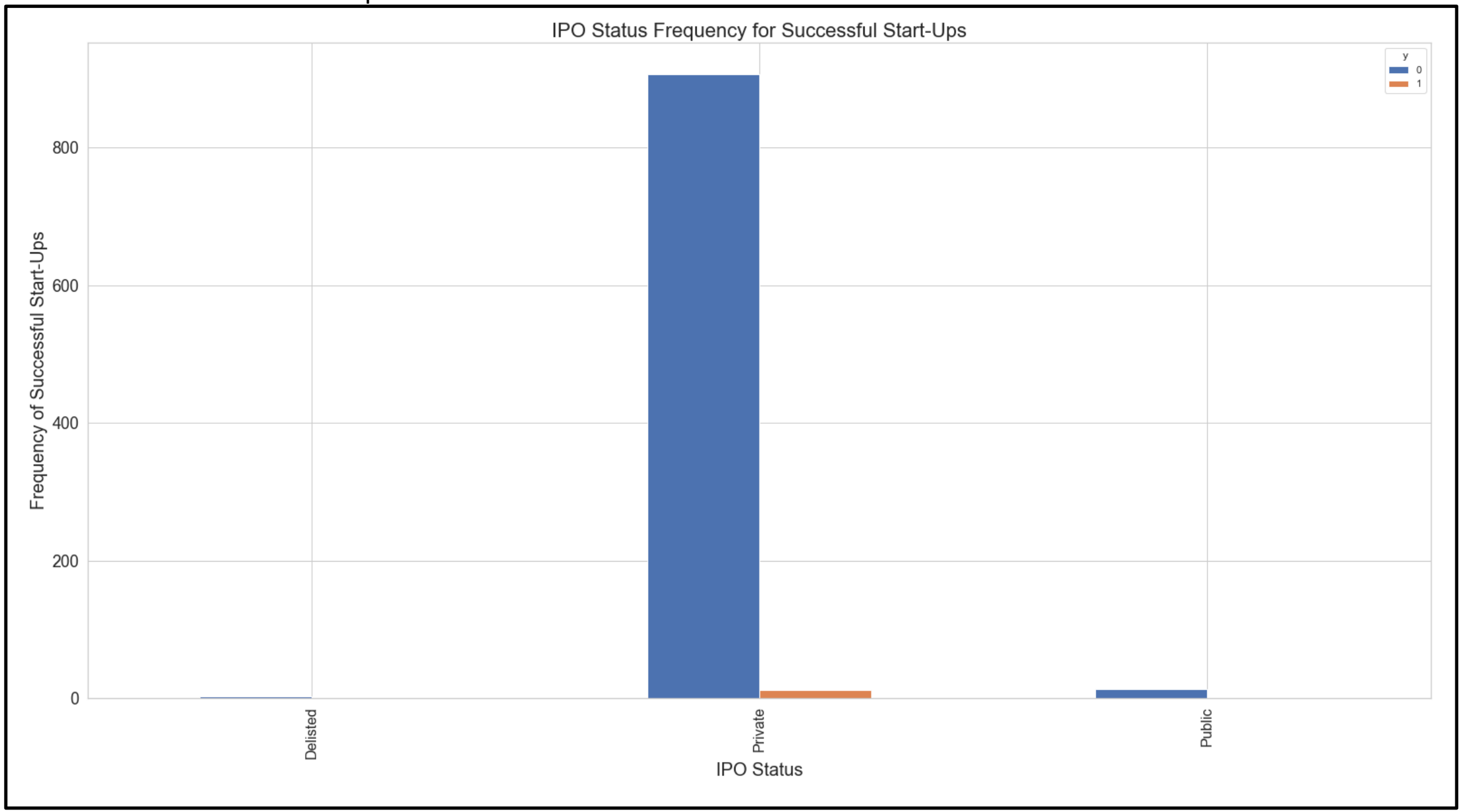
Si importa Number of Investors para éxito de Startups

Number of Investors vs Successful Start-Ups (Categórica)



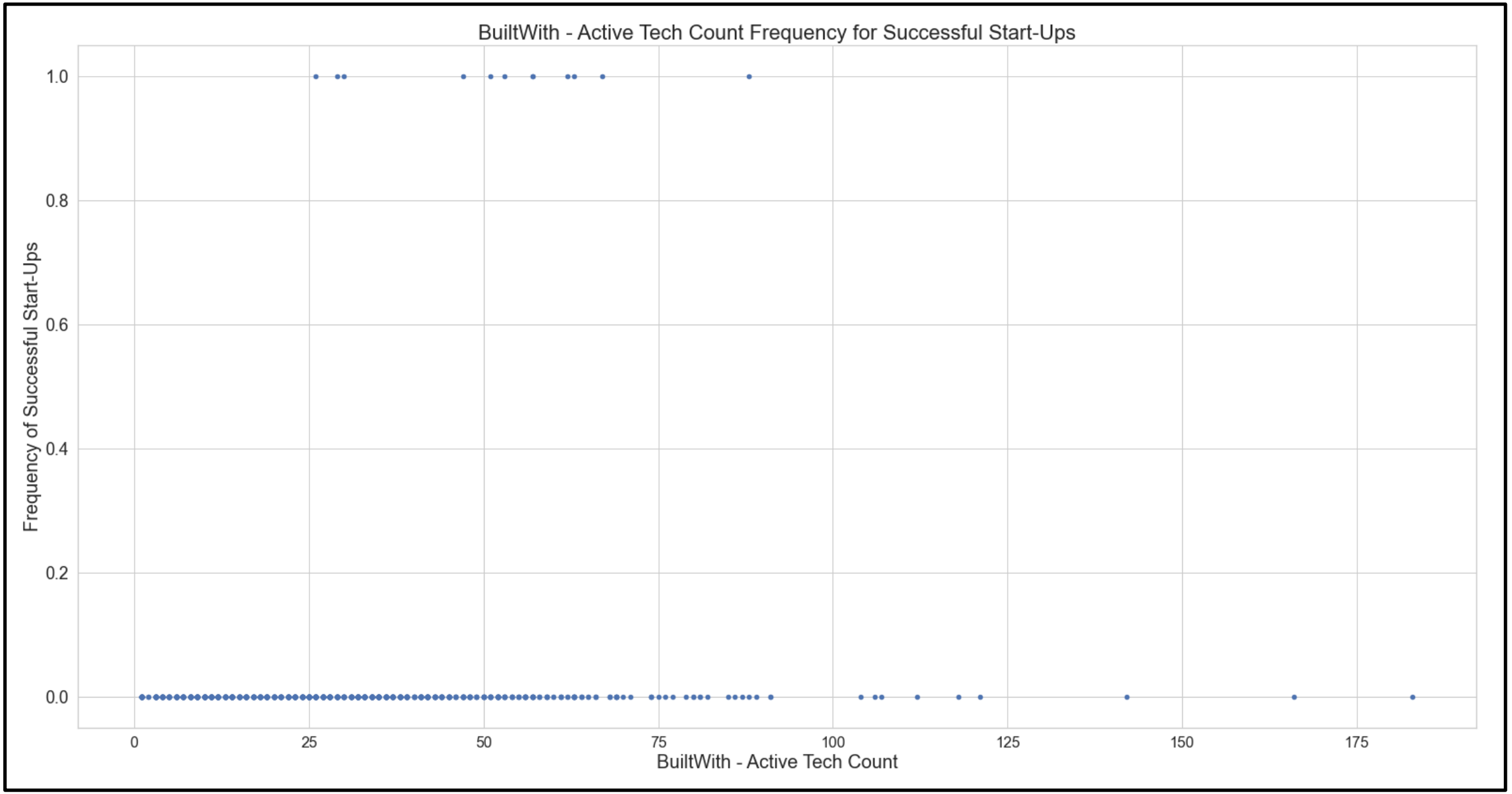
Si importa Number of Investors para éxito de Startups

IPO Status vs Successful Start-Ups



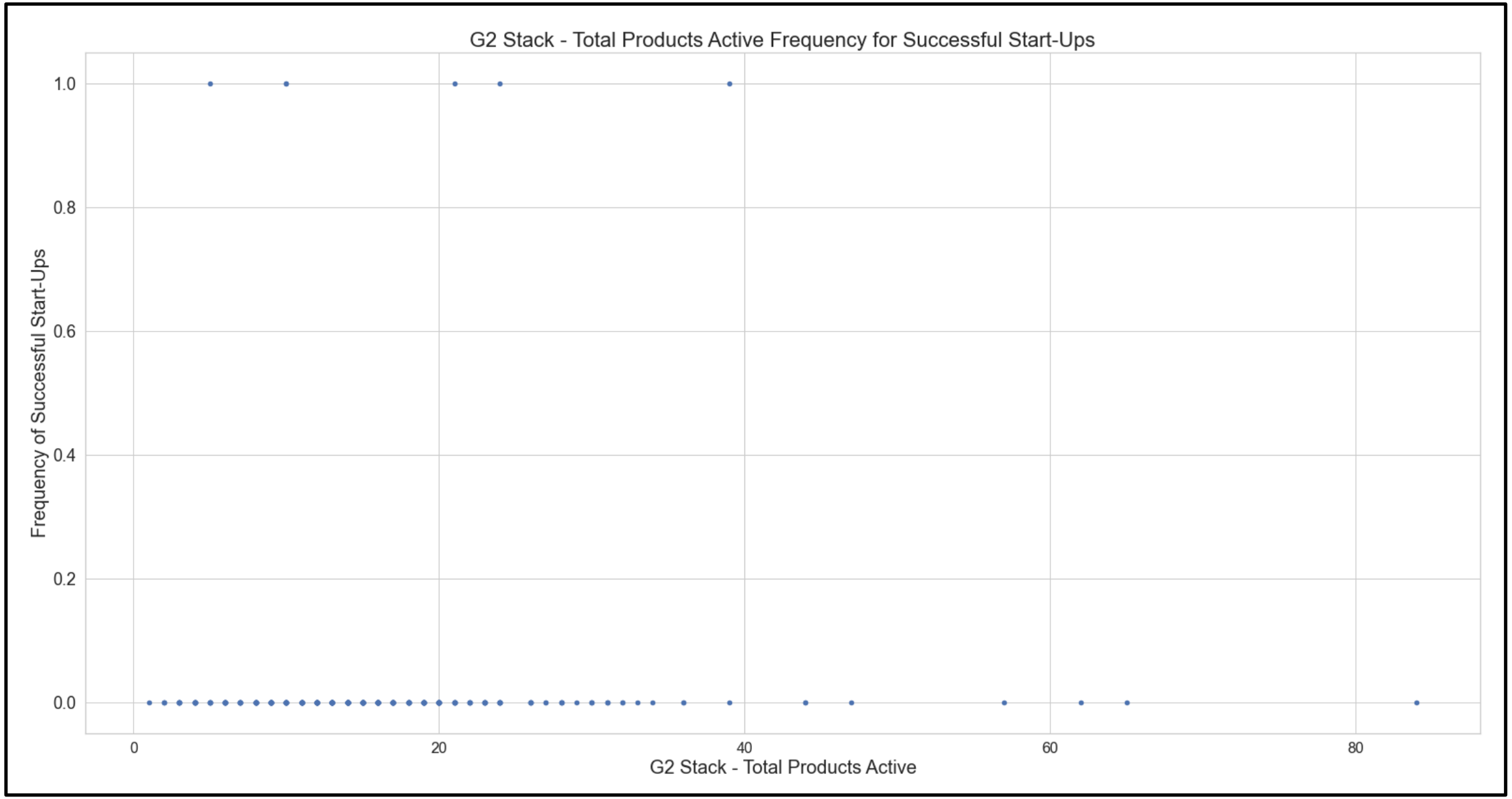
No importa IPO Status para éxito de Startups

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

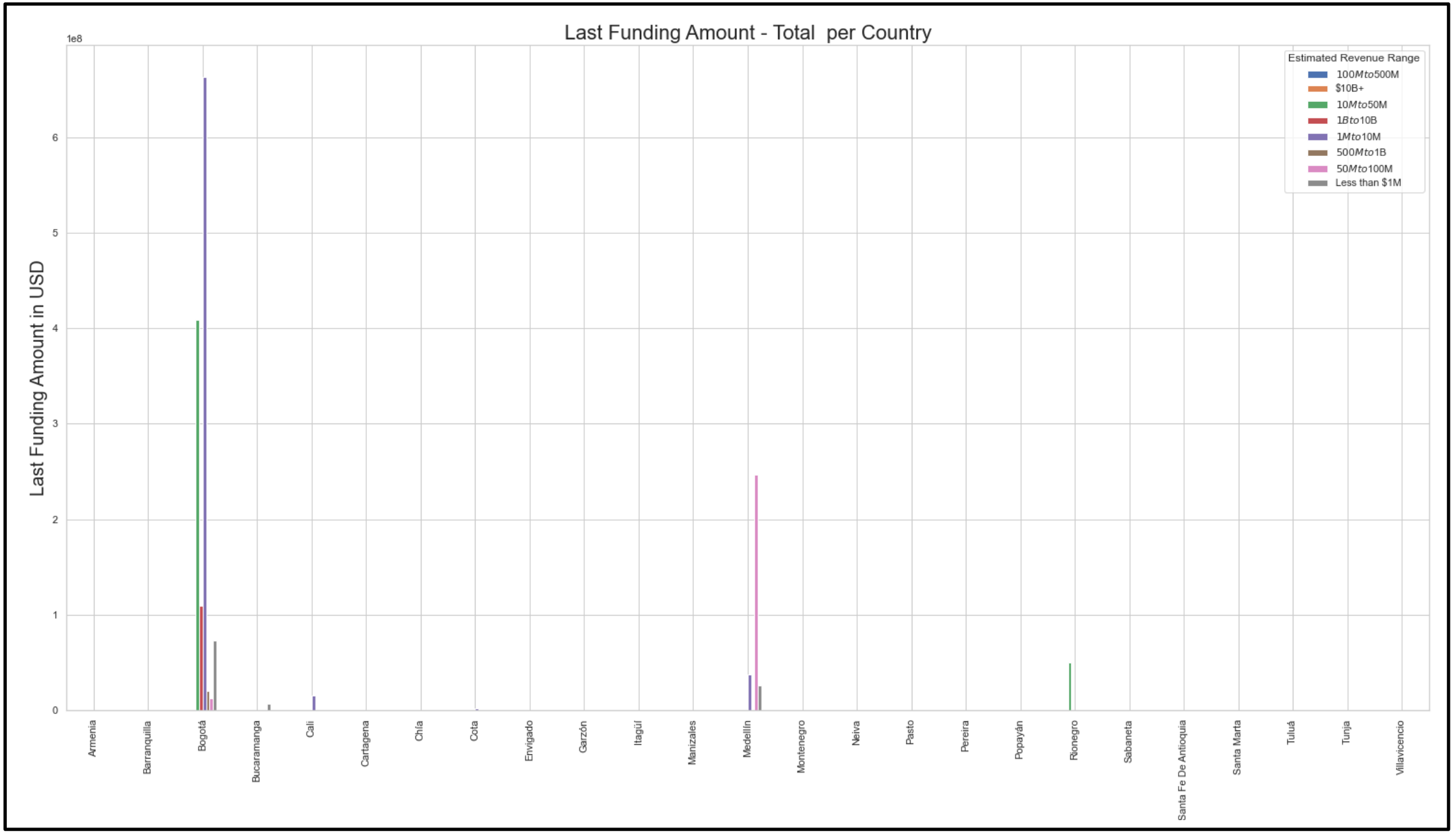


Si importa BuiltWith - Active Tech Count para éxito de Startups

G2 Stack - Total Products Active vs Successful Start-Ups



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



Bogota y Medellín tienen las start-ups con mayor inversión por parte de terceros

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

Se deja Dataframe con las 15 variables de entrada y la variable objetivo “y”
Se realiza get dummies a las variables 'Estimated Revenue Range', 'Number of Employees', 'Funding Status', 'Last Equity Funding Type’ quedando un **dataframe con 935 filas y 50 variables**

Regresión Logística

Regresión Logística

Se generó Variable Xtrain con 48 Variables y Ytrain con variable objetivo “y” y se obtuvo matriz singular.

Se analizó la base de datos encontrándose que las variables get dummies generadas donde el 95% o mas de los datos son ceros son los causantes de la matriz singular.

Se generó function para determiner columnas con mas del 95% de sus datos con ceros y se eliminaron 28 variables de entrada, quedando 20 variables de entrada.

```
def cols_90per_zeros(data, perc):
    count = 0
    cols_to_drop = {}
    for col in data.columns:
        per_zeros = data[col][data[col]==0].count()/len(data[col])
        if per_zeros >= perc:
            cols_to_drop[col] = per_zeros
            # print(col, per_nulls)
            count+=1
        else:
            None

    print('Number of cols with > ', perc*100, '% Zeros:', count)
    return cols_to_drop

# este diccionario tiene los nombres de Columnas con mas del 80% de datos con ceros

dict_col_zeros=cols_90per_zeros(Xtrain, 0.95)
```

Regresión Logística

Se corrió modelo de regression logística con statsmodels.api y no se obtuvo convergencia con 2000 iteraciones. Nuevamente se revisó la base de datos y se eliminaron las siguientes variables de entrada con alto volume de ceros en sus datos:

Number of Employees_1-10], Funding Status_Seed, Last Equity Funding Type_Pre-Seed y Last Equity Funding Type_Seed

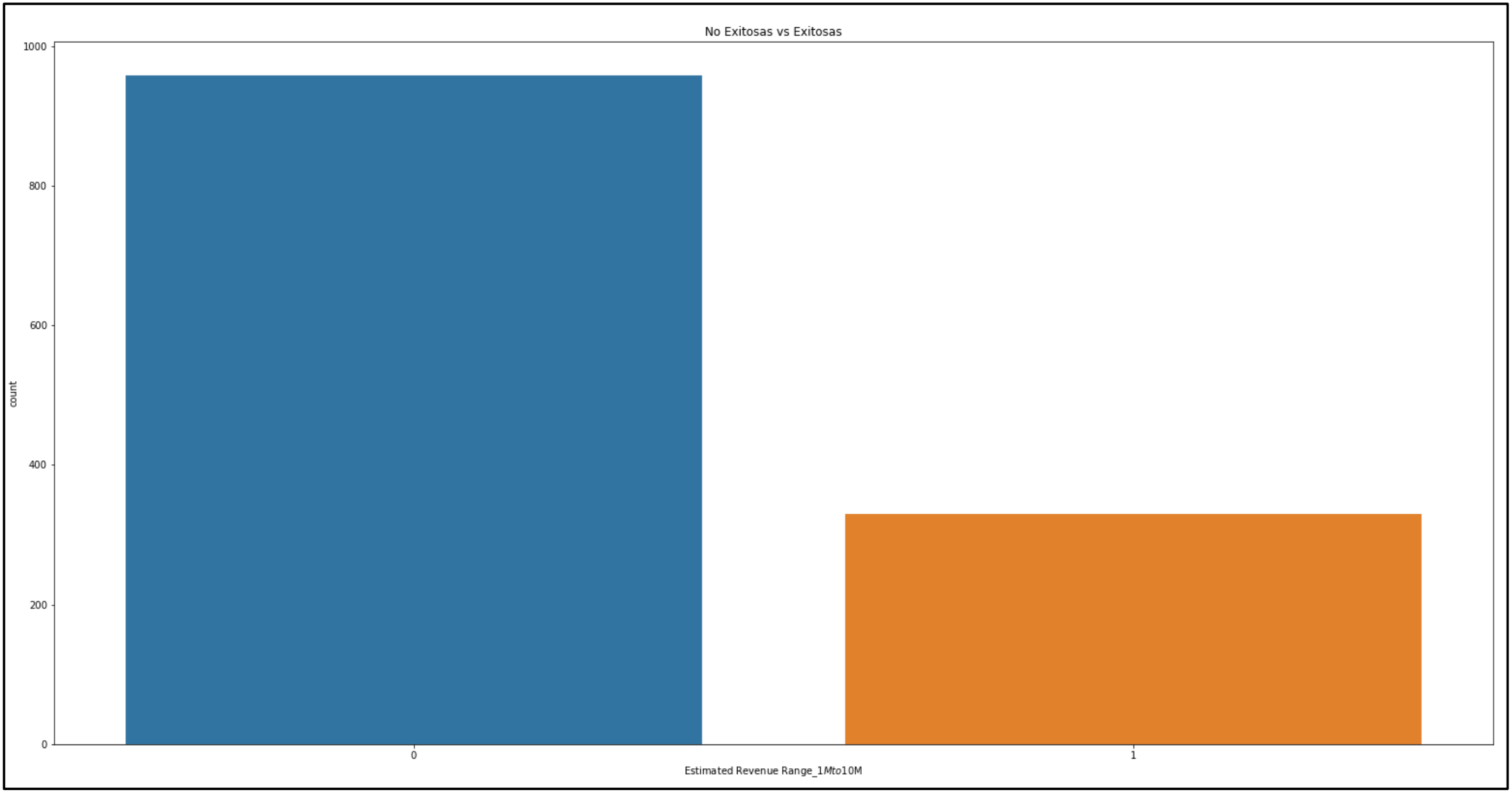
Se corrió modelo de regression logística con statsmodels.api con 16 variables de entrada y en 15 iteraciones se obtuvo convergencia

Previamente se realizo SMOTE de los datos y se generó conjunto de training con 70% y conjunto de prueba con 30% de los datos:

```
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

smt = SMOTE(random_state=0)
data_X,data_y=smt.fit_sample(X_train, y_train)
```

Al realizar SMOTE los datos de training subieron a 1288 y los de testing a 281. Sin el SMOTE el pseudo R-Squared da entre 0.74 – 0.78

Ejemplo de SMOTE con Count-Plot



Regresión Logística

summary = <class 'statsmodels.iolib.summary.Summary'>

"""

Logit Regression Results

Dep. Variable:	y	No. Observations:	1288
Model:	Logit	Df Residuals:	1272
Method:	MLE	Df Model:	15
Date:	Sun, 04 Apr 2021	Pseudo R-squ.:	0.9735
Time:	19:36:05	Log-Likelihood:	-23.681
converged:	True	LL-Null:	-892.77
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
CBRank	-0.0002	6.78e-05	-3.680	0.000	-0.000	-0.000
Number of Articles	-0.5773	0.307	-1.883	0.060	-1.178	0.024
Number of Founders	1.3145	0.605	2.173	0.030	0.129	2.500
Number of Funding Rounds	0.3160	0.717	0.441	0.659	-1.090	1.722
Last Funding Amount Currency (in USD)	1.21e-07	4.5e-07	0.269	0.788	-7.61e-07	1e-06
Last Equity Funding Amount Currency (in USD)	-1.342e-07	4.6e-07	-0.291	0.771	-1.04e-06	7.68e-07
Total Equity Funding Amount Currency (in USD)	2.374e-07	4.02e-07	0.590	0.555	-5.51e-07	1.03e-06
Total Funding Amount Currency (in USD)	-2.228e-07	3.91e-07	-0.570	0.569	-9.89e-07	5.43e-07
Number of Investors	0.5132	0.154	3.341	0.001	0.212	0.814
BuiltWith - Active Tech Count	0.1812	0.052	3.500	0.000	0.080	0.283
G2 Stack - Total Products Active	-0.1750	0.067	-2.622	0.009	-0.306	-0.044
Estimated Revenue Range_$1M to $10M	-1.3017	1.243	-1.047	0.295	-3.739	1.135
Estimated Revenue Range_Less than $1M	-4.2922	1.922	-2.234	0.026	-8.058	-0.526
Number of Employees_101-250	-3.2557	1.528	-2.130	0.033	-6.251	-0.261
Number of Employees_11-50	-6.5091	2.267	-2.871	0.004	-10.952	-2.066
Number of Employees_51-100	-1.8427	1.653	-1.115	0.265	-5.082	1.397

Matriz de confusión y accuracy Test
Para datos de Training

```
Confusion Matrix :
 [[638   6]
 [  1 643]]
Test accuracy =  0.9945652173913043
```

Matriz de confusión y accuracy Test
Para datos de Testing

```
Confusion Matrix :
 [[270   9]
 [  0   2]]
Test accuracy =  0.9679715302491103
```

Conclusión

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

13 de las 15 variables originales son representativas para el modelo que identifica las startups exitosas. Las que están en rojo fueron descartadas finalmente