UNIVERSIDAD
SERGIO ARBOLEDA

Técnicas Avanzadas de Minería de Datos y Machine Learning

Profesor: Luz Estela Gómez , Ph. D

Taller 5 – 10 Abril 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.

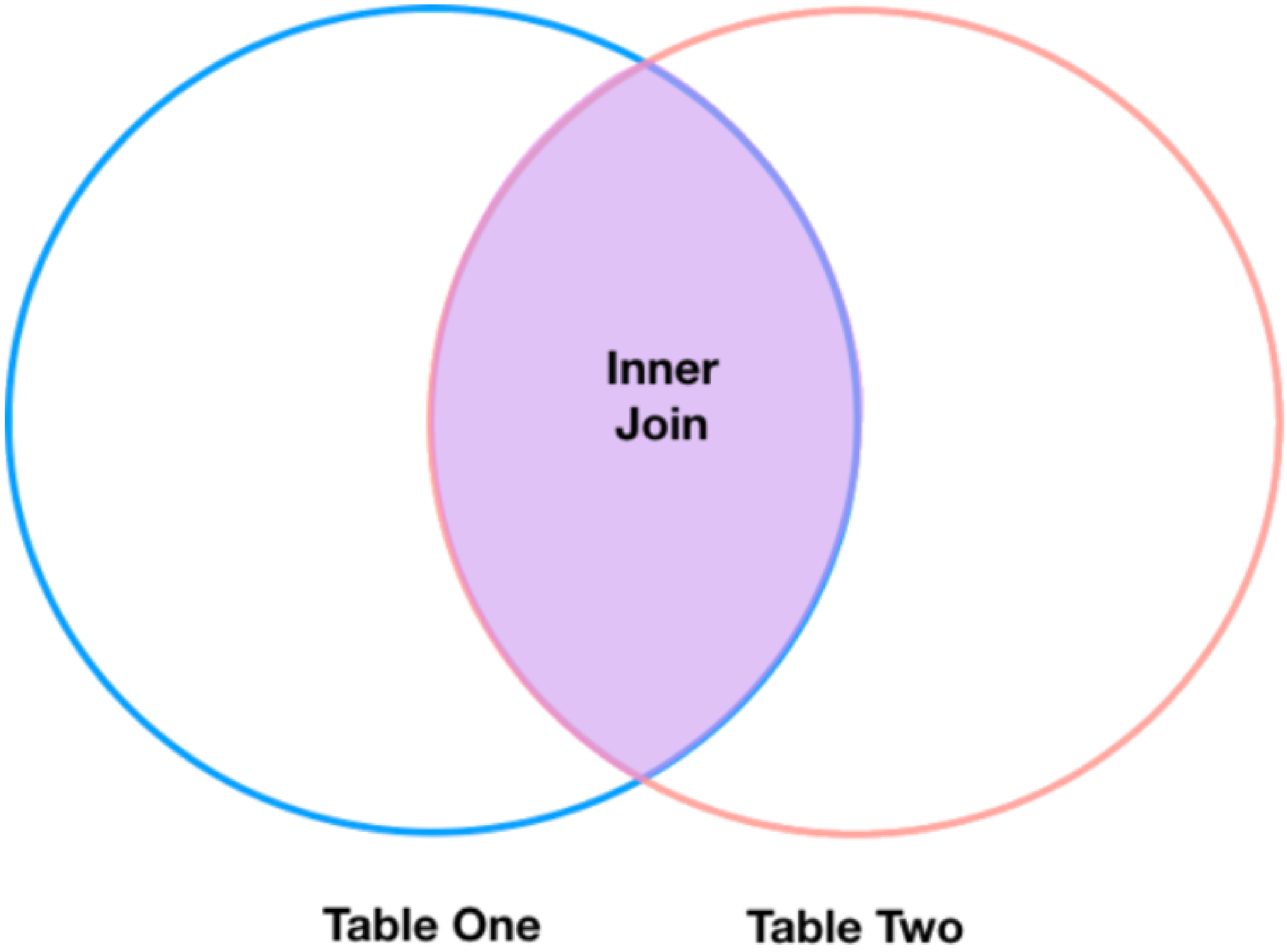
1. Integración de 2 dataframes

Intersección de 2 dataframes a través de valores de una columna que comparten en común

Inner joins

The most common type of join is called an *inner join*. An inner join combines two DataFrames based on a join key and returns a new DataFrame that contains **only** those rows that have matching values in *both* of the original DataFrames.

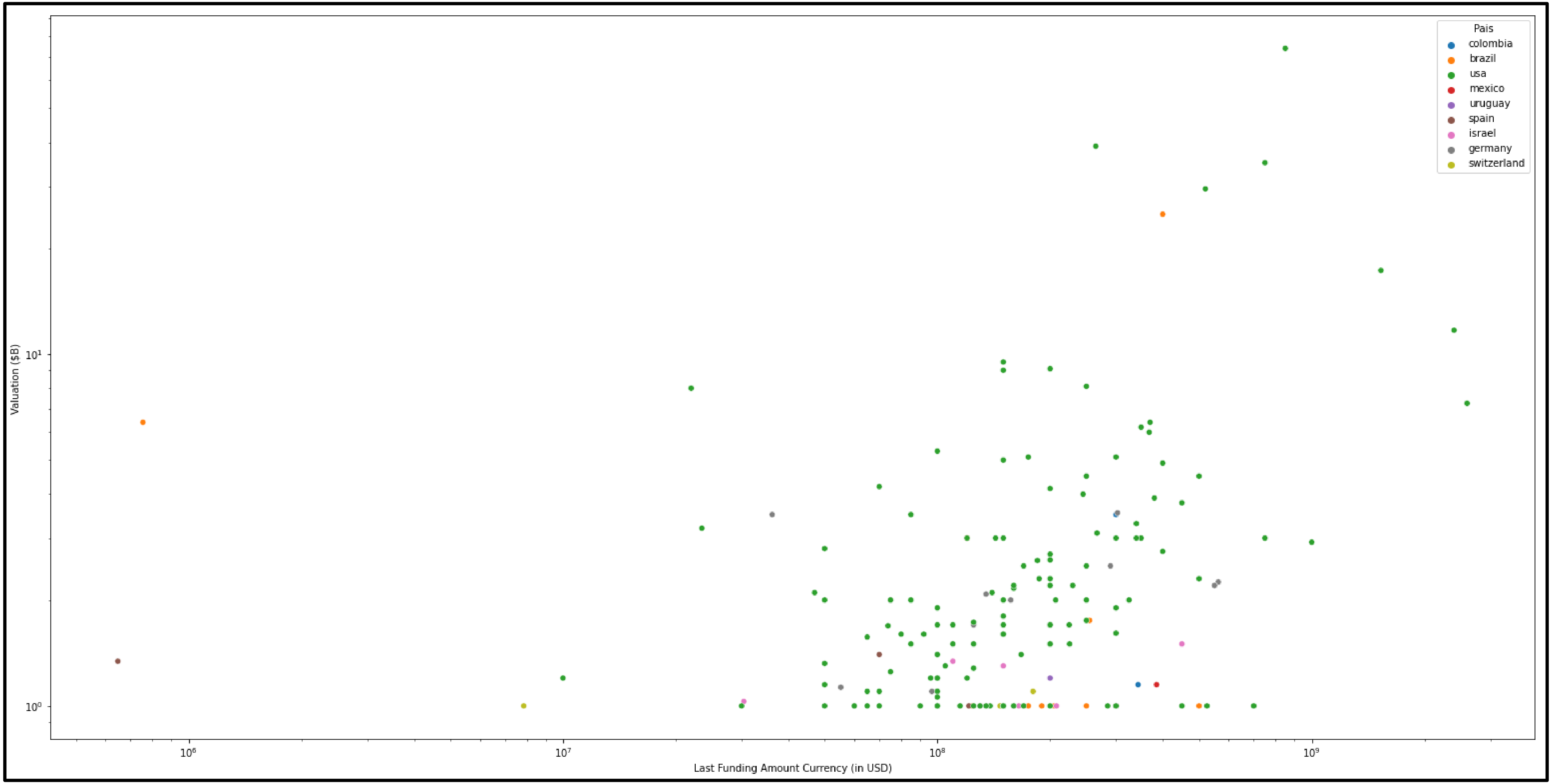
Inner joins yield a DataFrame that contains only rows where the value being joined exists in BOTH tables. An example of an inner join, adapted from Jeff Atwood's blogpost about SQL joins is below:



Tomado de: https://datacarpentry.org/python-ecology-lesson/05-merging-data/

```python
##############################################################################
# Integracion de 2 dataframes
##  dfx y df12
# dfx es el dataframe con informacion de los 11 paises
# df12 = Global Unicorn Club: Private Companies Valued at $1B+ (as of March 8th, 2021)

df12=pd.read_excel('CB-Insights_Global-Unicorn-Club_2021.xlsx')

# cambiando valores de celdas de columna Company del dataframe df12 a minusculas
df12["Company"]=df12["Company"].str.lower()

# cambiando valores de celdas de columna Organization del dataframe dfx a minusculas
dfx["Organization"]=dfx["Organization"].str.lower()

# interseccion de 2 dataframes a traves de columnas Organization y Company
merged_inner = pd.merge(left=dfx, right=df12, left_on='Organization', right_on='Company')
```

df12	DataFrame	(591, 5)	Column names: Company, Valuation ($B) , Country, Category, Select Inve ...
dfx	DataFrame	(10195, 47)	Column names: Organization, Industries, Headquarters Location, Descrip ...
dict_col_nul	dict	51	{'Exit Date':0.8727807748896518, 'Exit Date Precision': 0.8727807748896 ...
i	DataFrame	(1000, 103)	Column names: Organization Name, Organization Name URL, Industries, He ...
Lista	list	11	[Dataframe, Dataframe, Dataframe, Dataframe, Dataframe, Dataframe, Dat ...
merged_inner	DataFrame	(206, 52)	Column names: Organization, Industries, Headquarters Location, Descrip ...

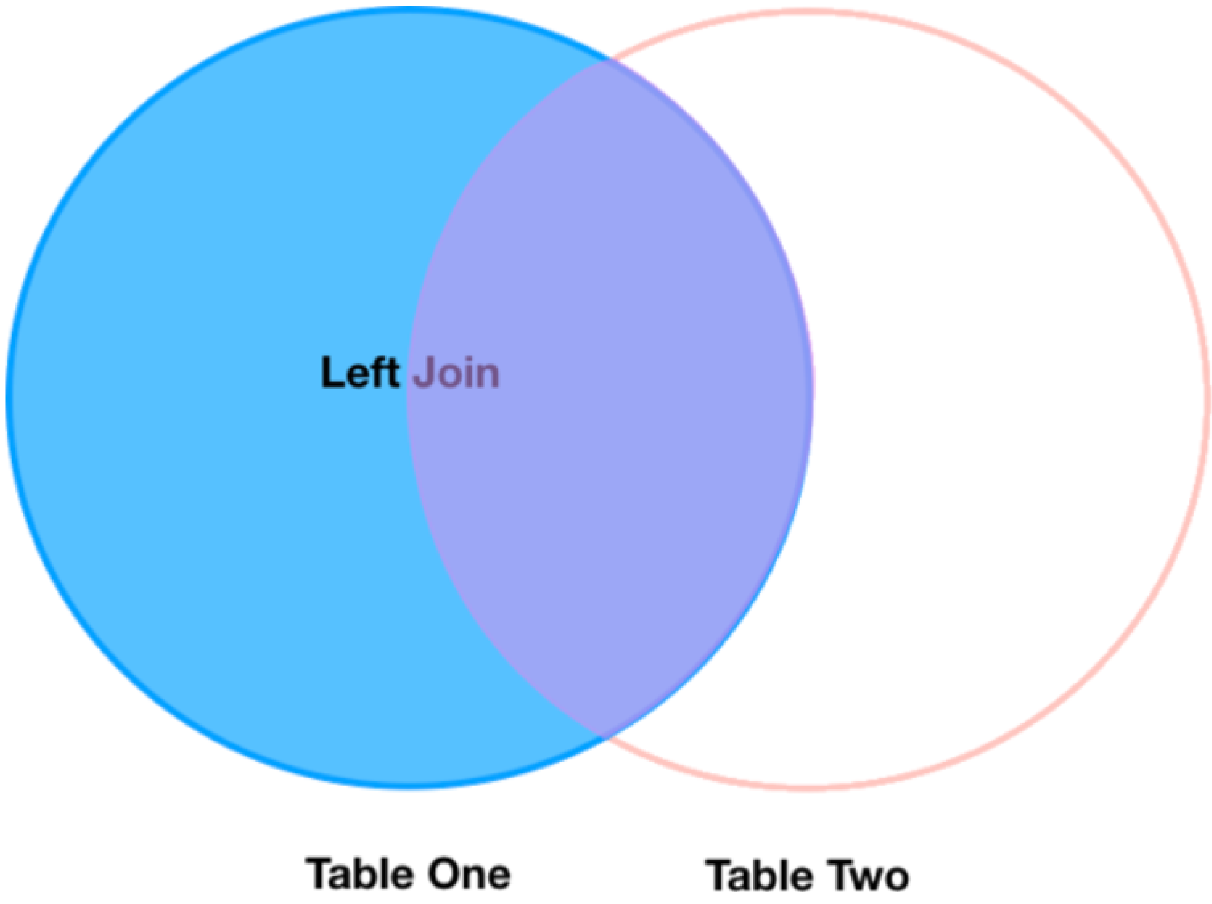

Pais
colombia
brazil
usa
mexico
uruguay
spain
israel
germany
switzerland
Valuation ($B)
10^1
10^0
10^6
10^7
10^8
10^9
Last Funding Amount Currency (in USD)

Left joins

What if we want to add information from `species_sub` to `survey_sub` without losing any of the information from `survey_sub`? In this case, we use a different type of join called a "left outer join", or a "left join".

Like an inner join, a left join uses join keys to combine two DataFrames. Unlike an inner join, a left join will return *all* of the rows from the `left` DataFrame, even those rows whose join key(s) do not have values in the `right` DataFrame. Rows in the `left` DataFrame that are missing values for the join key(s) in the `right` DataFrame will simply have null (i.e., NaN or None) values for those columns in the resulting joined DataFrame.

Note: a left join will still discard rows from the `right` DataFrame that do not have values for the join key(s) in the `left` DataFrame.



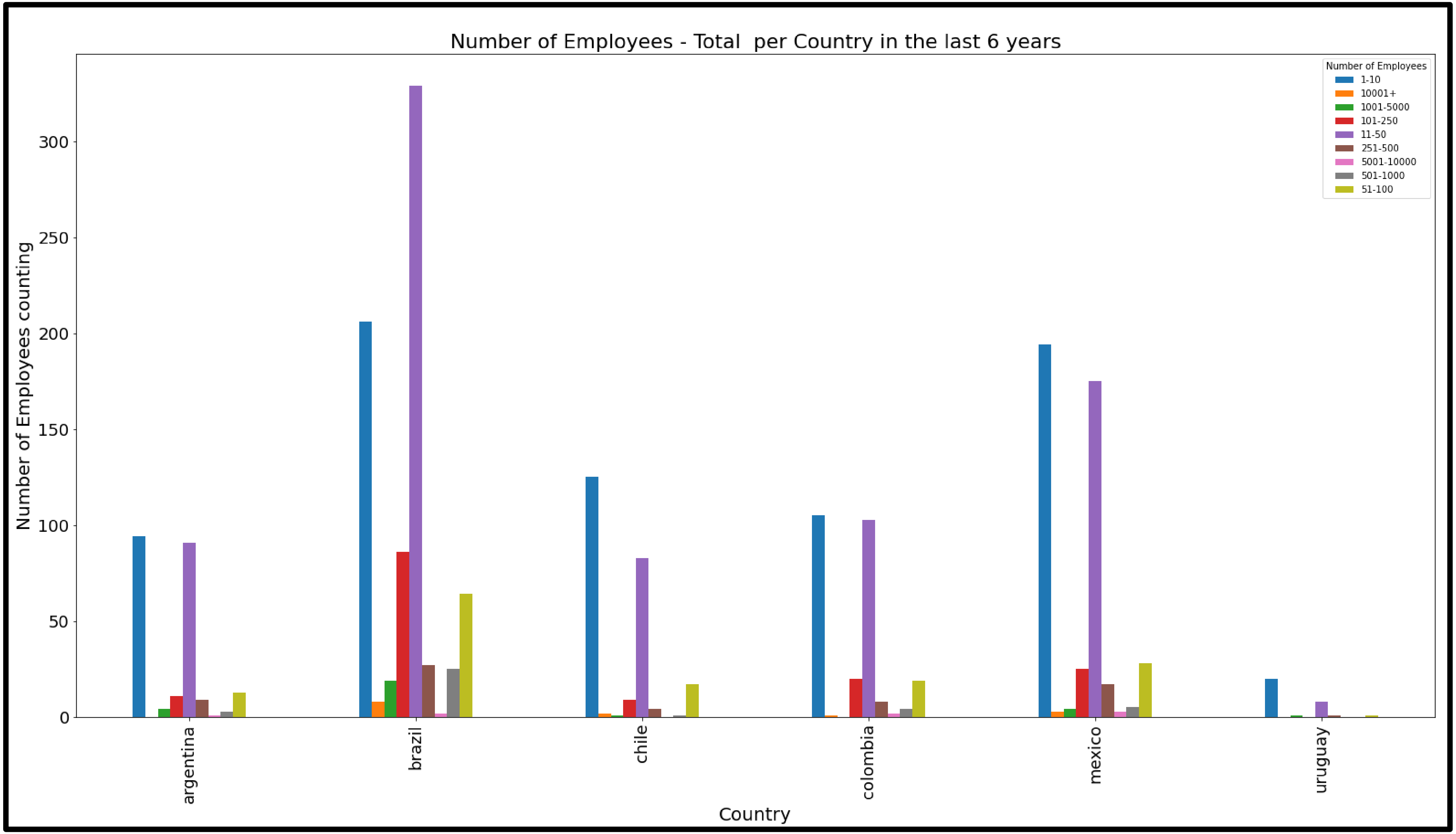
Tomado de: https://datacarpentry.org/python-ecology-lesson/05-merging-data/

A left join is performed in pandas by calling the same `merge` function used for inner join, but using the `how='left'` argument:

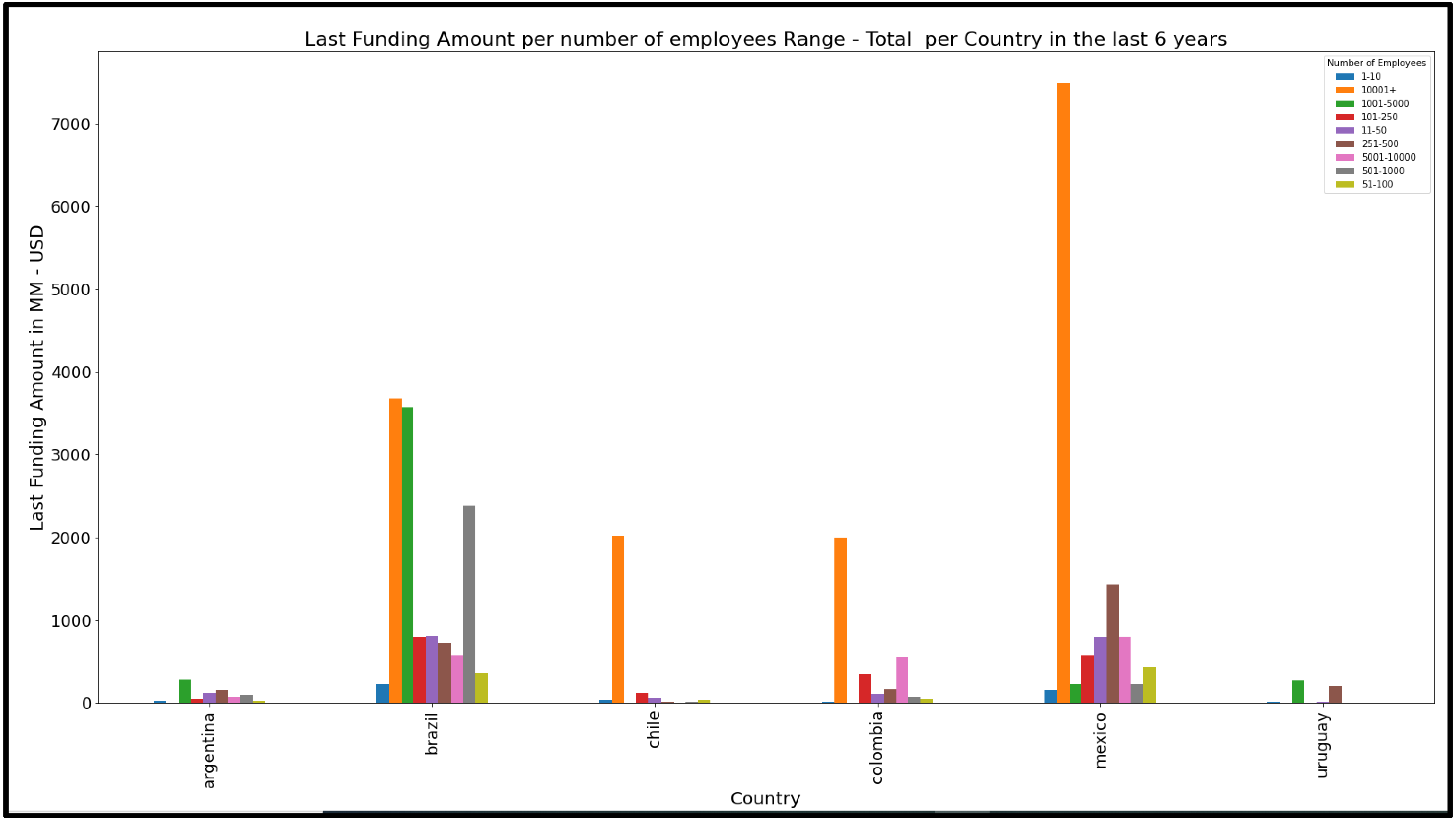
```python
##############################################################################
# Integracion de 2 dataframes
##  dfx y df12
# dfx es el dataframe con informacion de los 11 paises
# df12 = Global Unicorn Club: Private Companies Valued at $1B+ (as of March 8th, 2021)

df12=pd.read_excel('CB-Insights_Global-Unicorn-Club_2021.xlsx')

# cambiando valores de celdas de columna Company del dataframe df12 a minusculas
df12["Company"]=df12["Company"].str.lower()

# cambiando valores de celdas de columna Organization del dataframe dfx a minusculas
dfx["Organization"]=dfx["Organization"].str.lower()

# interseccion de 2 dataframes a traves de columnas Organization y Company
merged_inner = pd.merge(left=dfx, right=df12, left_on='Organization', right_on='Company')

# interseccion de 2 dataframes a traves de columnas Organization y Company
merged_inner = pd.merge(left=dfx, right=df12, how = 'left', left_on='Organization', right_on='Company')
```

df12	DataFrame	(591, 5)	Column names: Company, Valuation ($B) , Country, Category, Select Inve ...
dfx	DataFrame	(10195, 47)	Column names: Organization, Industries, Headquarters Location, Descrip ...
dict_col_nul	dict	51	{'Exit Date':0.8727807748896518, 'Exit Date Precision': 0.8727807748896 ...
i	DataFrame	(1000, 103)	Column names: Organization Name, Organization Name URL, Industries, He ...
Lista	list	11	[Dataframe, Dataframe, Dataframe, Dataframe, Dataframe, Dataframe, Dat ...
merged_inner	DataFrame	(10195, 52)	Column names: Organization, Industries, Headquarters Location, Descrip ...

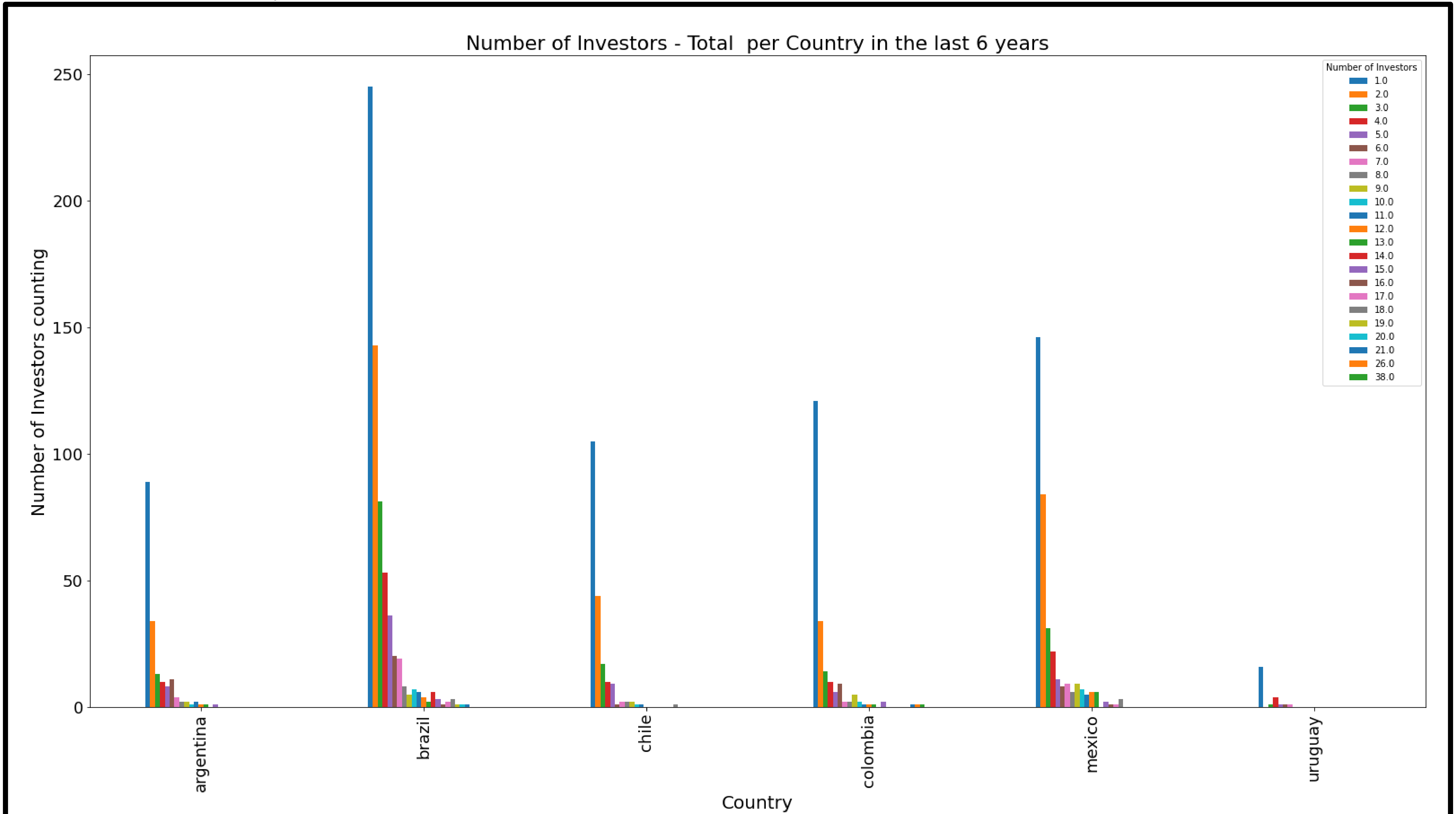
Graficas adicionales en el punto 2



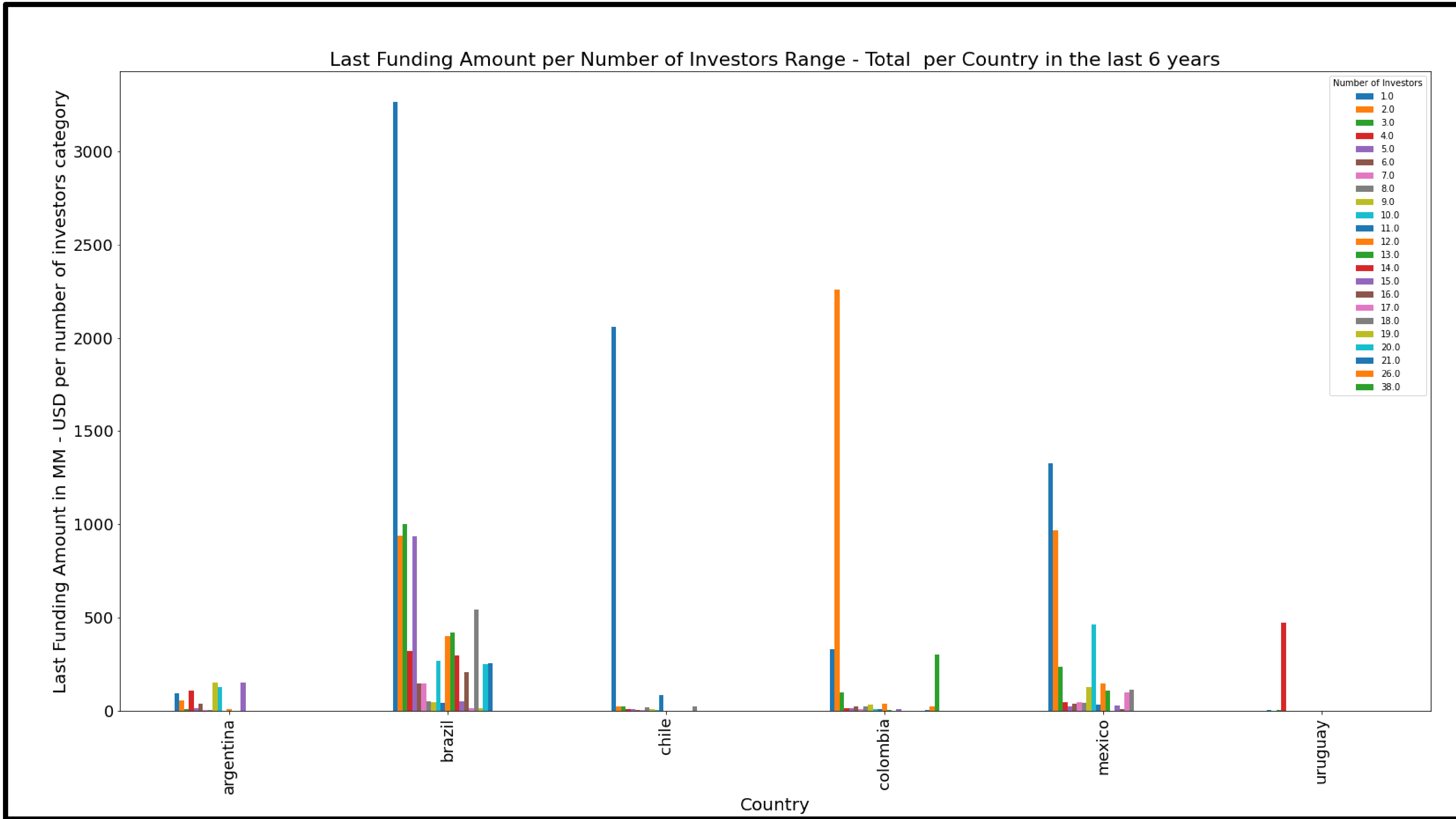
Graficas adicionales en el punto 2



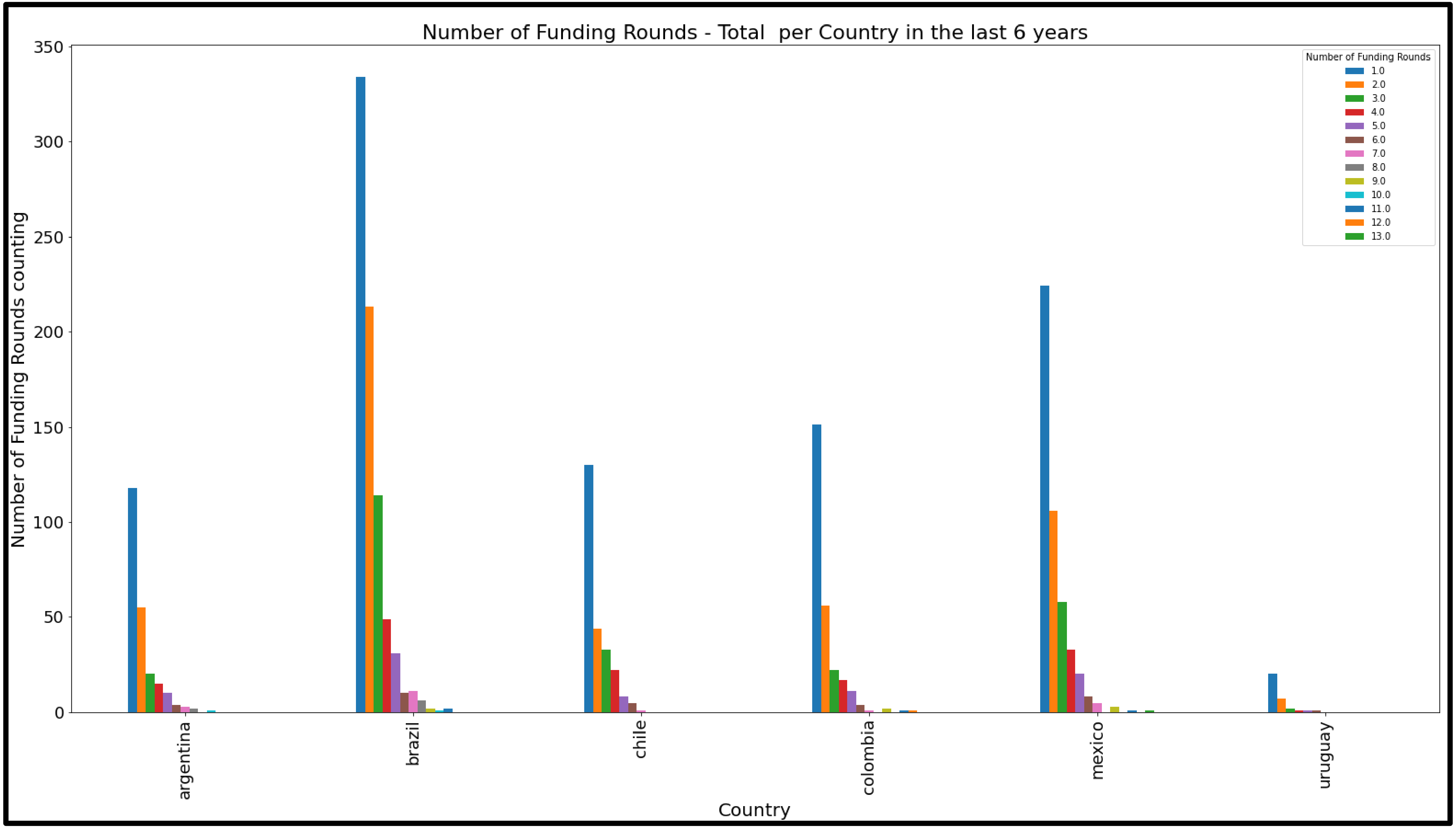
Graficas adicionales en el punto 2



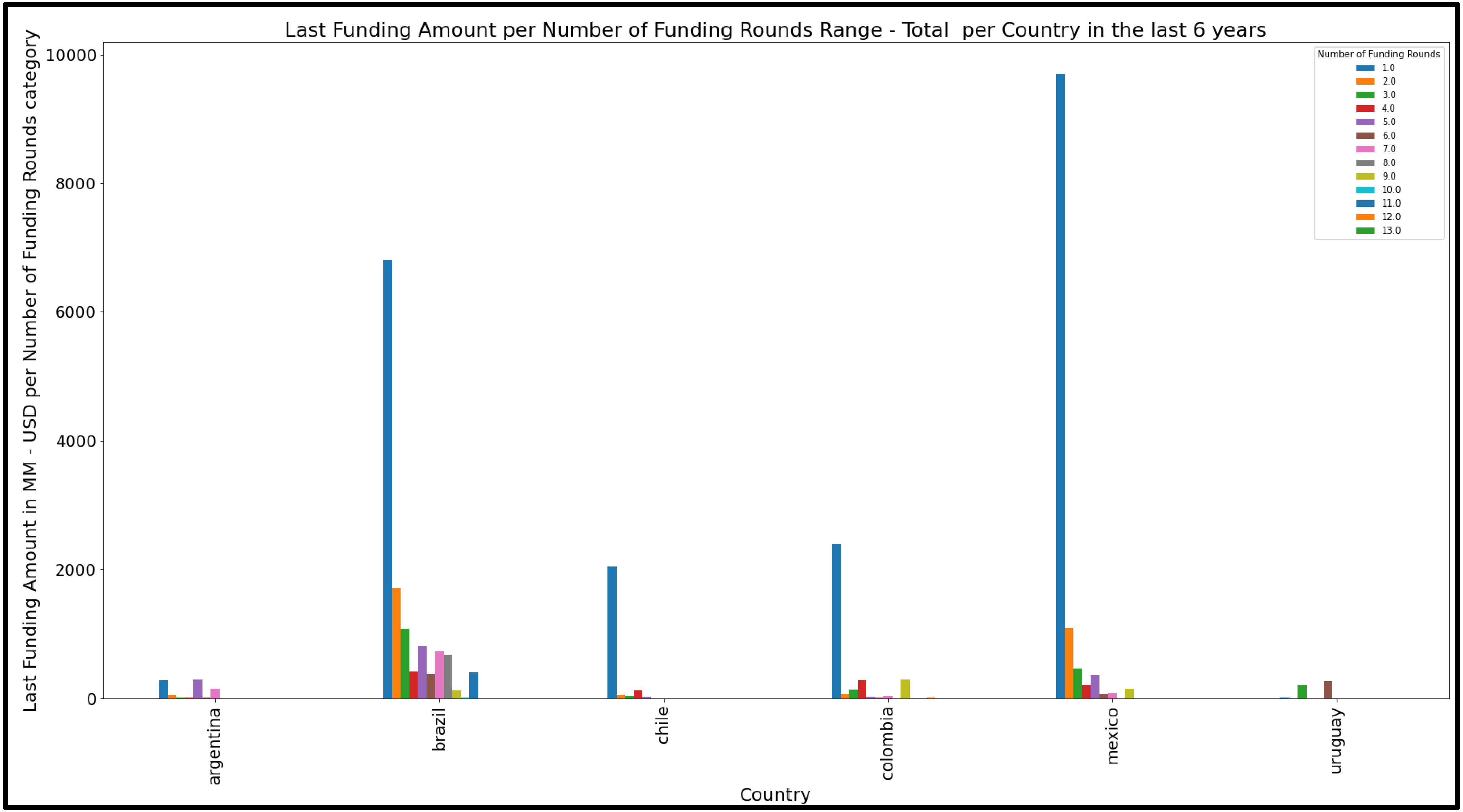
Graficas adicionales en el punto 2

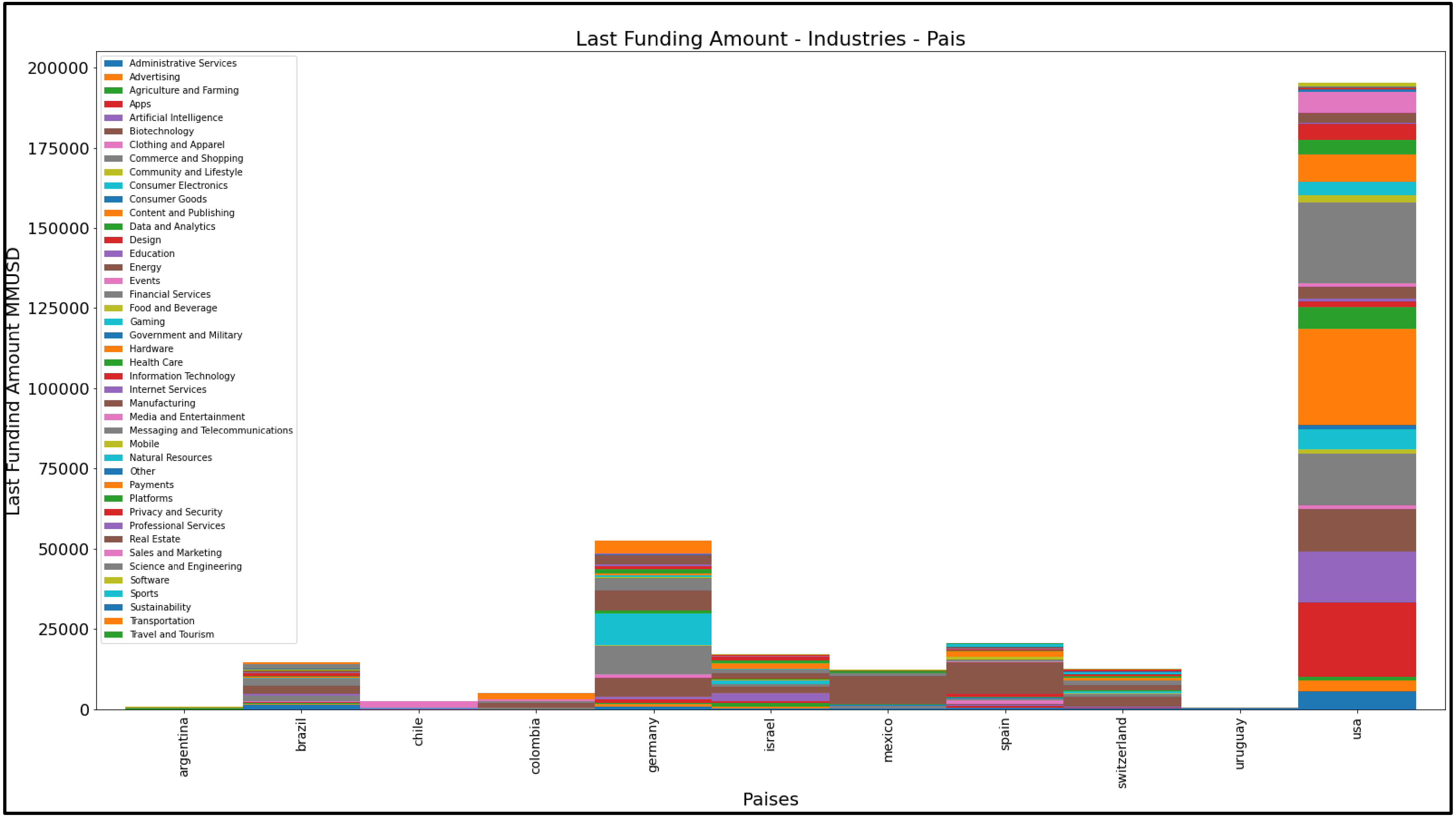


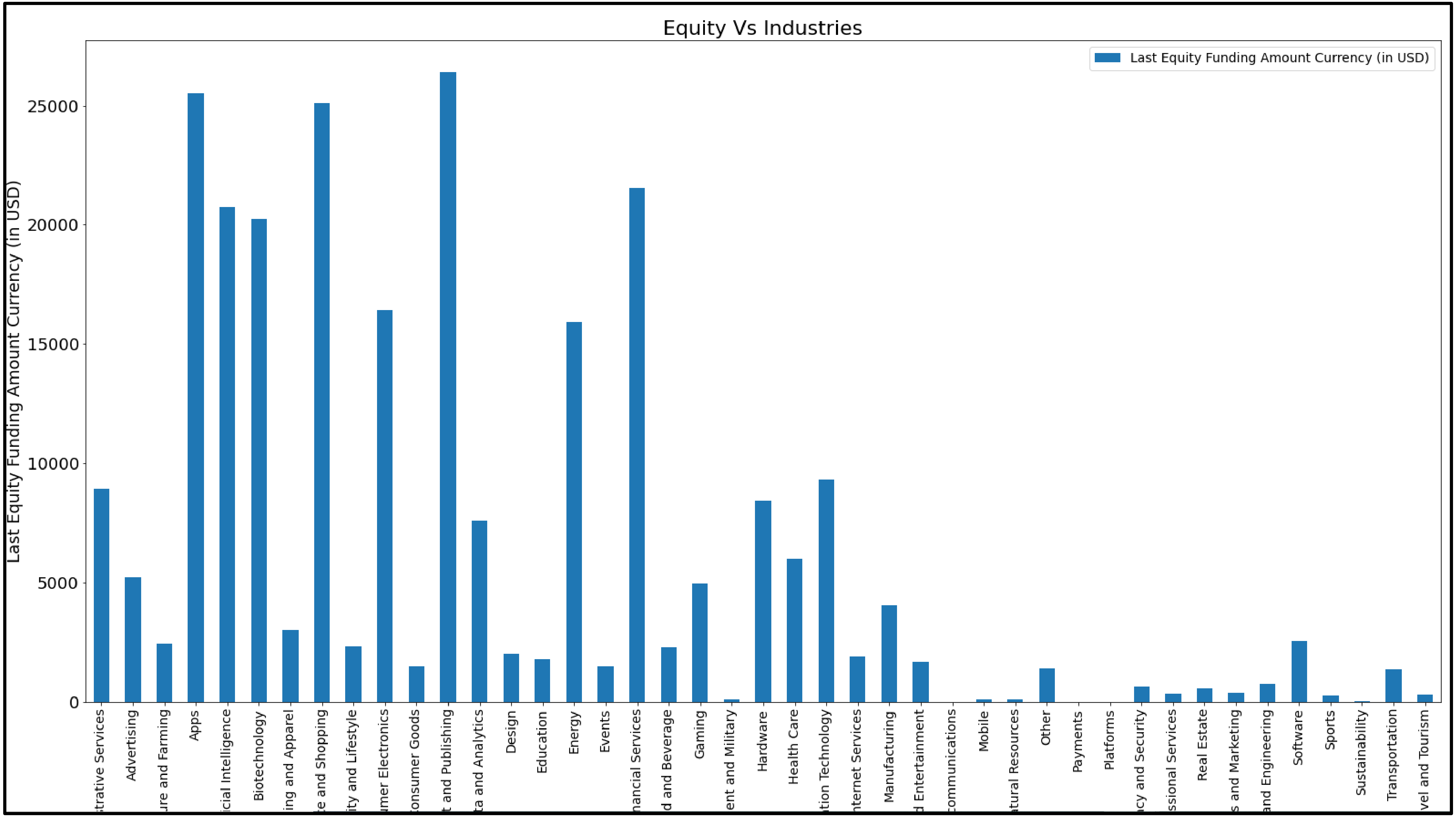
Graficas adicionales en el punto 2



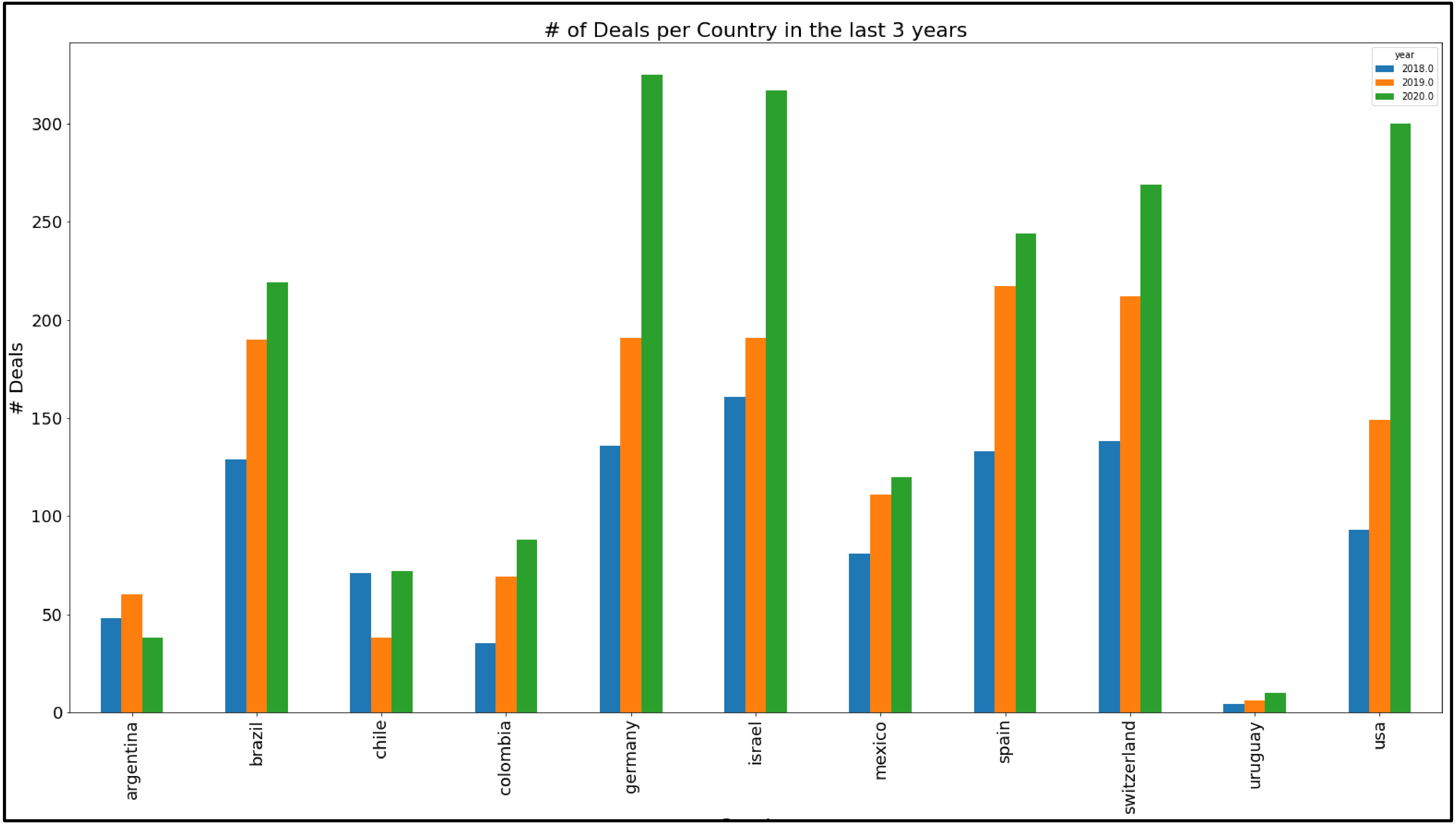
Graficas adicionales en el punto 2



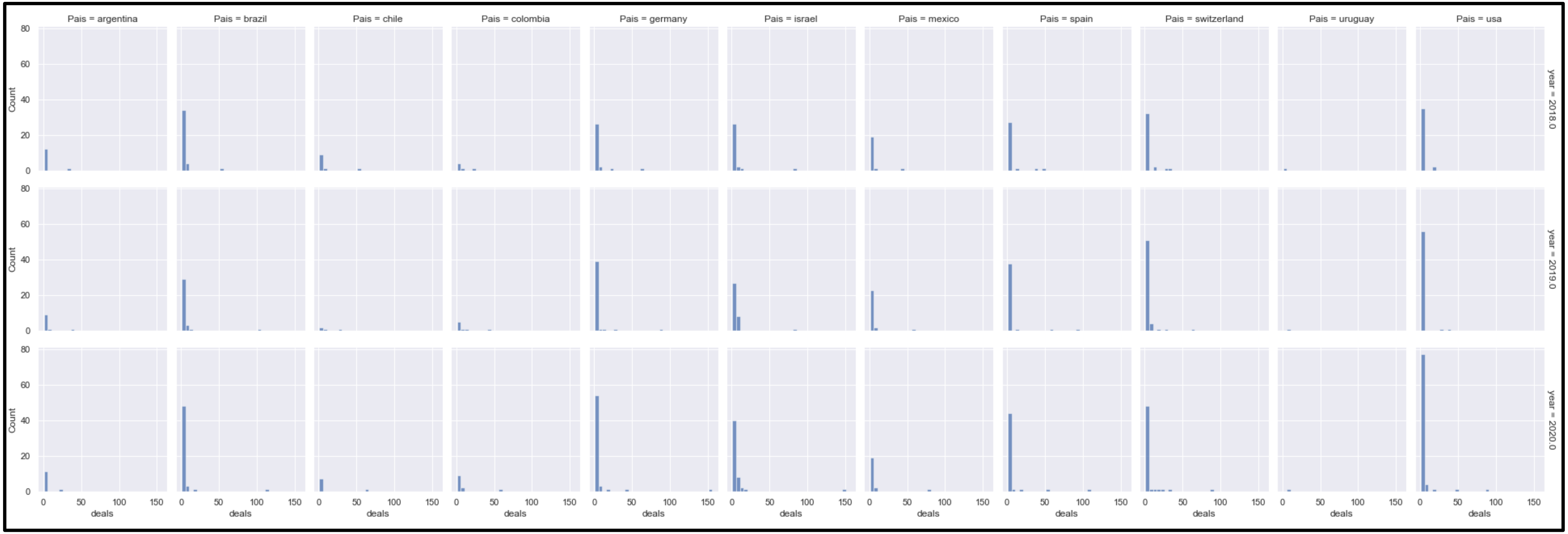

Last Funding Amount - Industries - Pais
Last Fundind Amount MMUSD
Paises
200000
175000
150000
125000
100000
75000
50000
25000
0
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa
Administrative Services
Advertising
Agriculture and Farming
Apps
Artificial Intelligence
Biotechnology
Clothing and Apparel
Commerce and Shopping
Community and Lifestyle
Consumer Electronics
Consumer Goods
Content and Publishing
Data and Analytics
Design
Education
Energy
Events
Financial Services
Food and Beverage
Gaming
Government and Military
Hardware
Health Care
Information Technology
Internet Services
Manufacturing
Media and Entertainment
Messaging and Telecommunications
Mobile
Natural Resources
Other
Payments
Platforms
Privacy and Security
Professional Services
Real Estate
Sales and Marketing
Science and Engineering
Software
Sports
Sustainability
Transportation
Travel and Tourism



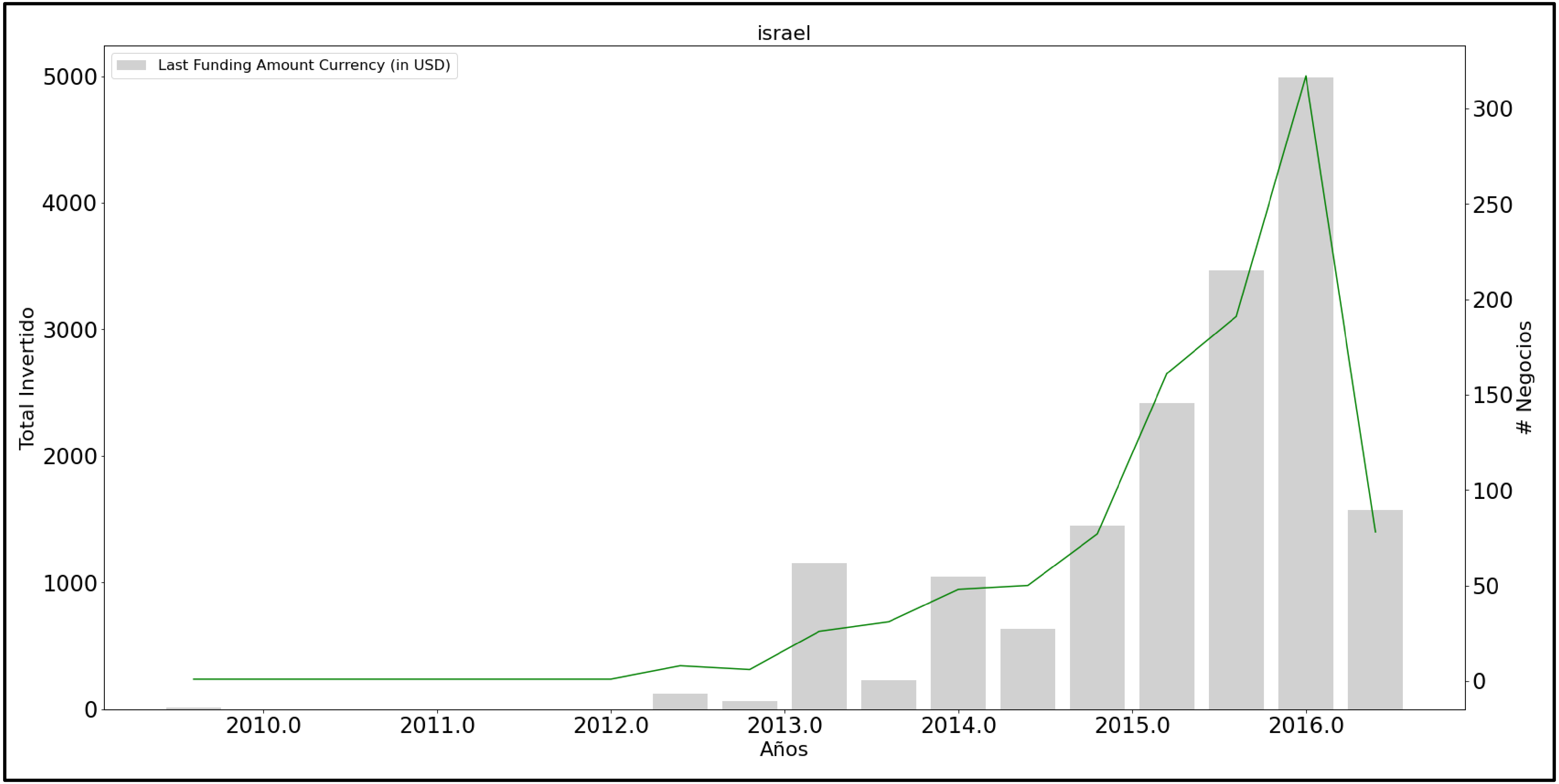
Equity Vs Industries
Last Equity Funding Amount Currency (in USD)
Last Equity Funding Amount Currency (in USD)
25000
20000
15000
10000
5000
0
Advertising
Apps
Biotechnology
Design
Education
Energy
Events
Gaming
Hardware
Health Care
Manufacturing
Mobile
Other
Payments
Platforms
Real Estate
Software
Sports
Sustainability
Transportation



# of Deals per Country in the last 3 years
year
2018.0
2019.0
2020.0
Deals
0
50
100
150
200
250
300
argentina
brazil
chile
colombia
germany
israel
mexico
spain
switzerland
uruguay
usa



Pais = argentina
Pais = brazil
Pais = chile
Pais = colombia
Pais = germany
Pais = israel
Pais = mexico
Pais = spain
Pais = switzerland
Pais = uruguay
Pais = usa
year = 2018.0
year = 2019.0
year = 2020.0
Count
deals
0
20
40
60
80
50
100
150


israel
Last Funding Amount Currency (in USD)
Total Invertido
Negocios
Años
0
1000
2000
3000
4000
5000
0
50
100
150
200
250
300
2010.0
2011.0
2012.0
2013.0
2014.0
2015.0
2016.0

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.

Regresión Logística Tarea 5

Se muestra resumen de la tarea 5 con objeto de tener punto de comparación con los resultados de Tarea 6.

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
=============================================================================================================================
```

P-valores altos mayores al 5%

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

Regresión Logística Tarea 6

	NaN	NaN After removing Last Equity Funding Amount Currency (in USD) Rows
CBRank	0	0
Number of Articles	685	109
Number of Founders	516	31
Number of Funding Rounds	578	0
Last Funding Amount Currency (in USD)	712	0
Last Equity Funding Amount Currency (in USD)	734	0
Total Equity Funding Amount Currency (in USD)	715	0
Total Funding Amount Currency (in USD)	692	0
Number of Investors	644	47
BuiltWith - Active Tech Count	39	12
G2 Stack - Total Products Active	661	139
y	0	0
Estimated Revenue Range_$100M to $500M	0	0
Estimated Revenue Range_$10B+	0	0
Estimated Revenue Range_$10M to $50M	0	0
Estimated Revenue Range_$1B to $10B	0	0
Estimated Revenue Range_$1M to $10M	0	0
Estimated Revenue Range_$500M to $1B	0	0
Estimated Revenue Range_$50M to $100M	0	0
Estimated Revenue Range_Less than $1M	0	0
Number of Employees_1-10	0	0
Number of Employees_10001+	0	0
Number of Employees_1001-5000	0	0
Number of Employees_101-250	0	0
Number of Employees_11-50	0	0
Number of Employees_251-500	0	0
Number of Employees_5001-10000	0	0
Number of Employees_501-1000	0	0
Number of Employees_51-100	0	0
Funding Status_Early Stage Venture	0	0
Funding Status_IPO	0	0
Funding Status_Late Stage Venture	0	0
Funding Status_M&A	0	0
Funding Status_Private Equity	0	0
Funding Status_Seed	0	0
Last Equity Funding Type_Angel	0	0
Last Equity Funding Type_Corporate Round	0	0
Last Equity Funding Type_Equity Crowdfunding	0	0
Last Equity Funding Type_Post-IPO Equity	0	0
Last Equity Funding Type_Pre-Seed	0	0
Last Equity Funding Type_Private Equity	0	0
Last Equity Funding Type_Seed	0	0
Last Equity Funding Type_Series A	0	0
Last Equity Funding Type_Series B	0	0
Last Equity Funding Type_Series C	0	0
Last Equity Funding Type_Series D	0	0
Last Equity Funding Type_Series F	0	0
Last Equity Funding Type_Undisclosed	0	0
Last Equity Funding Type_Venture - Series Unknown	0	0

Se eliminaron todas las filas donde la variable:

Last Equity Funding Amount Currency in USD, quitando 734 filas

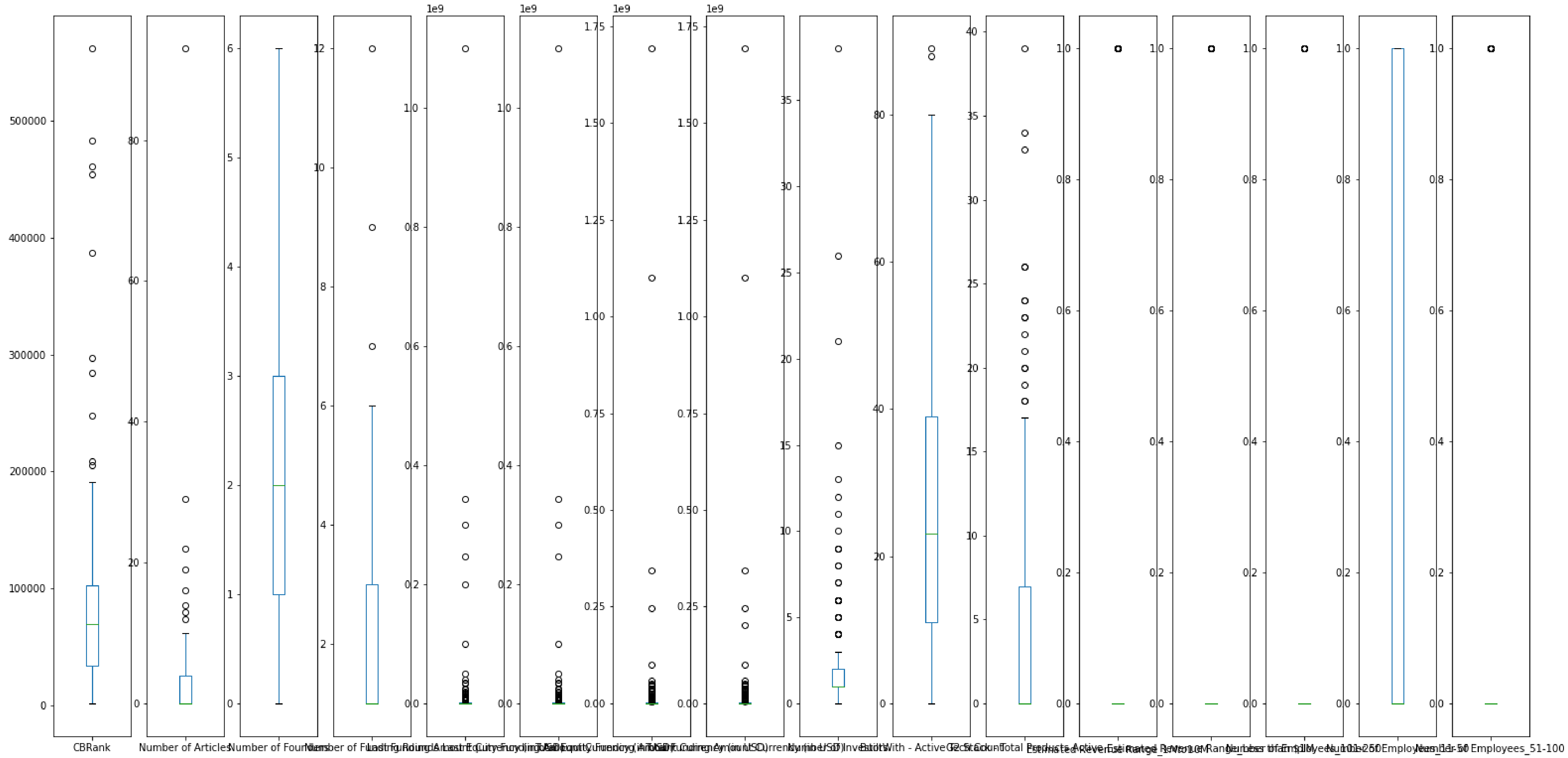
Las demás columnas que quedaron con NaNs no se eliminaron y se volvieron ceros.

La base de datos quedó con 201 filas y 49 columnas.

Se obtuvo matriz singular, para solucionar esto se eliminaron las variables con mas del 80% con valores de cero.

Quedando 201 filas y 18 Variables.

Obteniéndose el siguiente modelo de regression logística



Con base en resultados de box-plots se observa que hay que reescalar variables, especialmente las numéricas que implican cantidades en USD

```
summary = <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  266
Model:                          Logit   Df Residuals:                      248
Method:                           MLE   Df Model:                           17
Date:                Mon, 12 Apr 2021   Pseudo R-squ.:                  0.9457
Time:                        23:28:04   Log-Likelihood:                -10.016
converged:                       True   LL-Null:                       -184.38
Covariance Type:            nonrobust   LLR p-value:                 9.091e-64
==================================================================================================================
                                                coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
CBRank                                       -0.0005      0.000     -2.102      0.036      -0.001   -3.49e-05
Number of Articles                           -0.2281      0.536     -0.425      0.671      -1.279       0.823
Number of Founders                            0.3643      3.096      0.118      0.906      -5.703       6.432
Number of Funding Rounds                      0.0635      1.700      0.037      0.970      -3.268       3.394
Last Funding Amount Currency (in USD)     -5.844e-06   8.72e-06     -0.670      0.503   -2.29e-05    1.12e-05
Last Equity Funding Amount Currency (in USD) 5.864e-06   8.73e-06      0.671      0.502   -1.13e-05     2.3e-05
Total Equity Funding Amount Currency (in USD) 4.763e-08   5.34e-07      0.089      0.929      -1e-06     1.1e-06
Total Funding Amount Currency (in USD)    -6.109e-08    5.3e-07     -0.115      0.908    -1.1e-06    9.78e-07
Number of Investors                           0.5609      0.466      1.203      0.229      -0.353       1.475
BuiltWith - Active Tech Count                 0.3629      0.193      1.880      0.060      -0.015       0.741
G2 Stack - Total Products Active             -0.2280      0.532     -0.429      0.668      -1.270       0.814
Estimated Revenue Range_$1M to $10M          -4.8686      3.181     -1.531      0.126     -11.102       1.365
Estimated Revenue Range_Less than $1M       -11.0845      5.617     -1.973      0.048     -22.094      -0.075
Number of Employees_101-250                  -1.9546      7.430     -0.263      0.792     -16.516      12.607
Number of Employees_11-50                    -7.5532      9.177     -0.823      0.410     -25.539      10.433
Number of Employees_51-100                   -8.2092     10.831     -0.758      0.448     -29.437      13.018
Funding Status_Early Stage Venture           14.0958      7.541      1.869      0.062      -0.683      28.875
Last Equity Funding Type_Series A           -17.2616     11.275     -1.531      0.126     -39.361       4.838
==================================================================================================================
```

Regresión sin
Re escalamiento.

Nótese p-valores altos

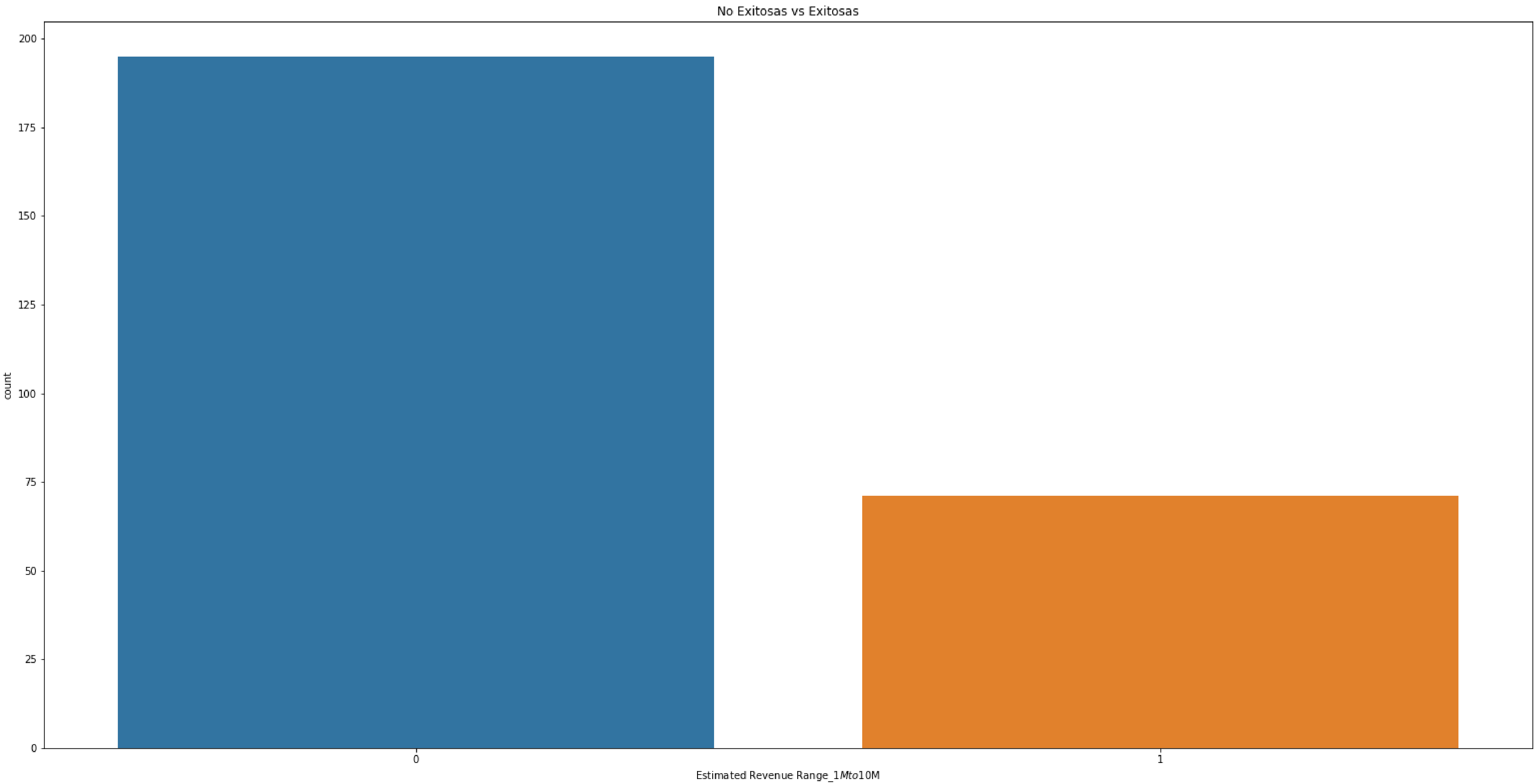
Se re-escalaron las variables:

Last Equity Funding Amount Currency (in USD)
Total Equity Funding Amount Currency (in USD)
Total Funding Amount Currency (in USD)
Last Funding Amount Currency (in USD)
G2 Stack - Total Products

Usando:

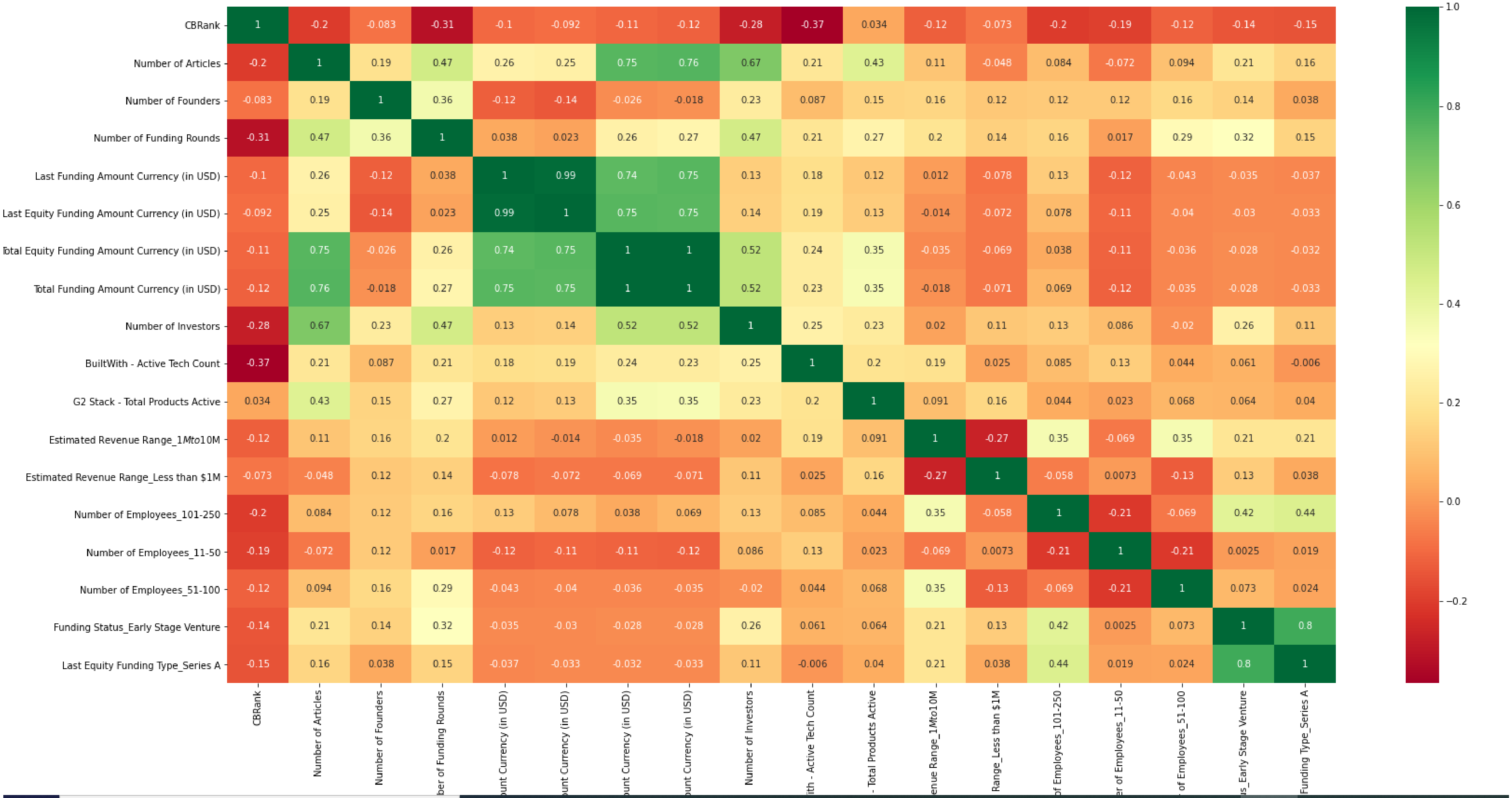
```
# Importing libraries for scaling
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train[['Last Equity Funding Amount Currency (in USD)','Total Equity Funding Amount Currency (in USD)','Total Funding Amount Currency (in USD)','Last Funding Amount Currency (in USD)','G2 Stack - Total Products Active']]=scaler.fit_transform(X_train[['Last Equity Funding Amount Currency (in USD)','Total Equity Funding Amount Currency (in USD)','Total Funding Amount Currency (in USD)','Last Funding Amount Currency (in USD)','G2 Stack - Total Products Active']])
```

SMOTE


No Exitosas vs Exitosas
count
200
175
150
125
100
75
50
25
0
0
1
Estimated Revenue Range_1Mto10M

```
summary = <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  266
Model:                          Logit   Df Residuals:                      248
Method:                           MLE   Df Model:                           17
Date:                Wed, 14 Apr 2021   Pseudo R-squ.:                  0.9433
Time:                        20:58:48   Log-Likelihood:                -10.448
converged:                       True   LL-Null:                       -184.38
Covariance Type:            nonrobust   LLR p-value:                 1.375e-63
=================================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
CBRank                                     -0.0004      0.000     -2.399      0.016      -0.001   -6.73e-05
Number of Articles                         -0.4292      0.296     -1.451      0.147      -1.009       0.151
Number of Founders                          1.8296      1.573      1.163      0.245      -1.254       4.913
Number of Funding Rounds                   -0.0479      0.776     -0.062      0.951      -1.569       1.473
Last Funding Amount Currency (in USD)      34.9748     30.905      1.132      0.258     -25.598      95.548
Last Equity Funding Amount Currency (in USD) -35.8705  30.524     -1.175      0.240     -95.697      23.956
Total Equity Funding Amount Currency (in USD) 55.2821  50.152      1.102      0.270     -43.015     153.579
Total Funding Amount Currency (in USD)    -57.1160     50.995     -1.120      0.263    -157.064      42.832
Number of Investors                         0.5567      0.368      1.514      0.130      -0.164       1.277
BuiltWith - Active Tech Count               0.2756      0.129      2.136      0.033       0.023       0.528
G2 Stack - Total Products Active            2.9337      2.266      1.295      0.195      -1.507       7.374
Estimated Revenue Range_$1M to $10M        -3.5159      2.271     -1.548      0.122      -7.966       0.935
Estimated Revenue Range_Less than $1M     -10.6711      5.965     -1.789      0.074     -22.362       1.020
Number of Employees_101-250                -4.4662      4.193     -1.065      0.287     -12.685       3.752
Number of Employees_11-50                  -9.5738      6.228     -1.537      0.124     -21.780       2.633
Number of Employees_51-100                 -7.0526      4.358     -1.618      0.106     -15.594       1.489
Funding Status_Early Stage Venture         12.4522      6.975      1.785      0.074      -1.219      26.123
Last Equity Funding Type_Series A         -11.7307      6.324     -1.855      0.064     -24.125       0.664
=================================================================================================================
```

Mejoraron los p-valores con el re-escalamiento de variables


CBRank
Number of Articles
Number of Founders
Number of Funding Rounds
Last Funding Amount Currency (in USD)
Last Equity Funding Amount Currency (in USD)
Total Equity Funding Amount Currency (in USD)
Total Funding Amount Currency (in USD)
Number of Investors
BuiltWith - Active Tech Count
G2 Stack - Total Products Active
Estimated Revenue Range_1Mto10M
Estimated Revenue Range_Less than $1M
Number of Employees_101-250
Number of Employees_11-50
Number of Employees_51-100
Funding Status_Early Stage Venture
Last Equity Funding Type_Series A
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Se eliminan las variables dummies con alta correlación (80%) entre ellas:

Last Equity Funding Type_Series A

Funding Status_Late Stage Venture

No re-escalamiento

```
summary = <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  266
Model:                          Logit   Df Residuals:                      250
Method:                           MLE   Df Model:                           15
Date:                Wed, 14 Apr 2021   Pseudo R-squ.:                  0.9141
Time:                        21:29:26   Log-Likelihood:                -15.845
converged:                       True   LL-Null:                       -184.38
Covariance Type:            nonrobust   LLR p-value:                 1.061e-62
==========================================================================================================
                                              coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------
CBRank                                     -0.0002   6.83e-05     -3.397      0.001      -0.000   -9.82e-05
Number of Articles                         -0.1477      0.146     -1.013      0.311      -0.433       0.138
Number of Founders                          1.0235      0.736      1.391      0.164      -0.419       2.466
Number of Funding Rounds                    0.1105      0.502      0.220      0.826      -0.874       1.095
Last Funding Amount Currency (in USD)      12.3391     25.088      0.492      0.623     -36.831      61.510
Last Equity Funding Amount Currency (in USD) -11.5111  24.737     -0.465      0.642     -59.995      36.973
Total Equity Funding Amount Currency (in USD) 20.2923  41.582      0.488      0.626     -61.207     101.792
Total Funding Amount Currency (in USD)     -22.4111    41.704     -0.537      0.591    -104.150      59.327
Number of Investors                         0.3258      0.201      1.618      0.106      -0.069       0.721
BuiltWith - Active Tech Count               0.1685      0.056      2.985      0.003       0.058       0.279
G2 Stack - Total Products Active            0.5470      1.122      0.488      0.626      -1.652       2.746
Estimated Revenue Range_$1M to $10M        -1.9363      1.536     -1.260      0.208      -4.948       1.075
Estimated Revenue Range_Less than $1M      -3.0145      1.968     -1.532      0.126      -6.871       0.842
Number of Employees_101-250                -3.3459      2.186     -1.530      0.126      -7.631       0.939
Number of Employees_11-50                  -6.8778      3.085     -2.230      0.026     -12.923      -0.832
Number of Employees_51-100                 -5.6869      3.476     -1.636      0.102     -12.500       1.126
==========================================================================================================
```

Matriz de confusión y accuracy Test
Para datos de Testing

```
Confusion Matrix :
 [[49  9]
 [ 0  3]]
Test accuracy =  0.8524590163934426
```

Matriz de confusión y accuracy Test
Para datos de Training

```
Confusion Matrix :
 [[128   5]
 [  1 132]]
Test accuracy =  0.9774436090225563
```

Reescalando con método de centering

$$X_c = X - \bar{X}$$

```
##############################################################################
# Centering method to variables with high p-values

# 'Last Funding Amount Currency (in USD)'
# 'Last Equity Funding Amount Currency (in USD)'
# 'Total Equity Funding Amount Currency (in USD)'
# 'Total Funding Amount Currency (in USD)'
# 'Number of Funding Rounds'
# 'Estimated Revenue Range_$1B to $10B'
# 'Number of Employees_51-100'
```

```
summary = <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  266
Model:                          Logit   Df Residuals:                      250
Method:                           MLE   Df Model:                           15
Date:                Fri, 16 Apr 2021   Pseudo R-squ.:                  0.9303
Time:                        23:11:04   Log-Likelihood:                -12.846
converged:                       True   LL-Null:                       -184.38
Covariance Type:            nonrobust   LLR p-value:                 5.927e-64
==================================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------------
CBRank                                           -0.0003      0.000     -1.053      0.292      -0.001       0.000
Number of Articles                                0.3792      0.786      0.483      0.629      -1.161       1.919
Number of Founders                               -0.8138      1.939     -0.420      0.675      -4.614       2.987
Number of Funding Rounds                          6.4576      7.250      0.891      0.373      -7.752      20.667
Last Funding Amount Currency (in USD)          1.128e-06   1.17e-06      0.960      0.337   -1.17e-06    3.43e-06
Last Equity Funding Amount Currency (in USD)  -1.073e-06   1.11e-06     -0.967      0.334   -3.25e-06     1.1e-06
Total Equity Funding Amount Currency (in USD)  1.148e-06   1.18e-06      0.974      0.330   -1.16e-06    3.46e-06
Total Funding Amount Currency (in USD)        -1.186e-06   1.22e-06     -0.969      0.333   -3.58e-06    1.21e-06
Number of Investors                               0.8610      0.747      1.152      0.249      -0.604       2.326
BuiltWith - Active Tech Count                    -0.0228      0.148     -0.153      0.878      -0.314       0.268
G2 Stack - Total Products Active                 -1.6235      1.962     -0.828      0.408      -5.469       2.222
Estimated Revenue Range_$1M to $10M               9.8058     11.509      0.852      0.394     -12.751      32.363
Estimated Revenue Range_Less than $1M            -3.0442      2.043     -1.490      0.136      -7.049       0.960
Number of Employees_101-250                     -13.6156     12.386     -1.099      0.272     -37.891      10.660
Number of Employees_11-50                       -11.8867     11.470     -1.036      0.300     -34.368      10.594
Number of Employees_51-100                      -34.6962     40.802     -0.850      0.395    -114.666      45.273
==================================================================================================================
```

Mejoran los p-valores
y R-Cuadrado

Reescalando con método de Estandarización

$$X_{std} = \frac{X - \bar{X}}{s_X}$$

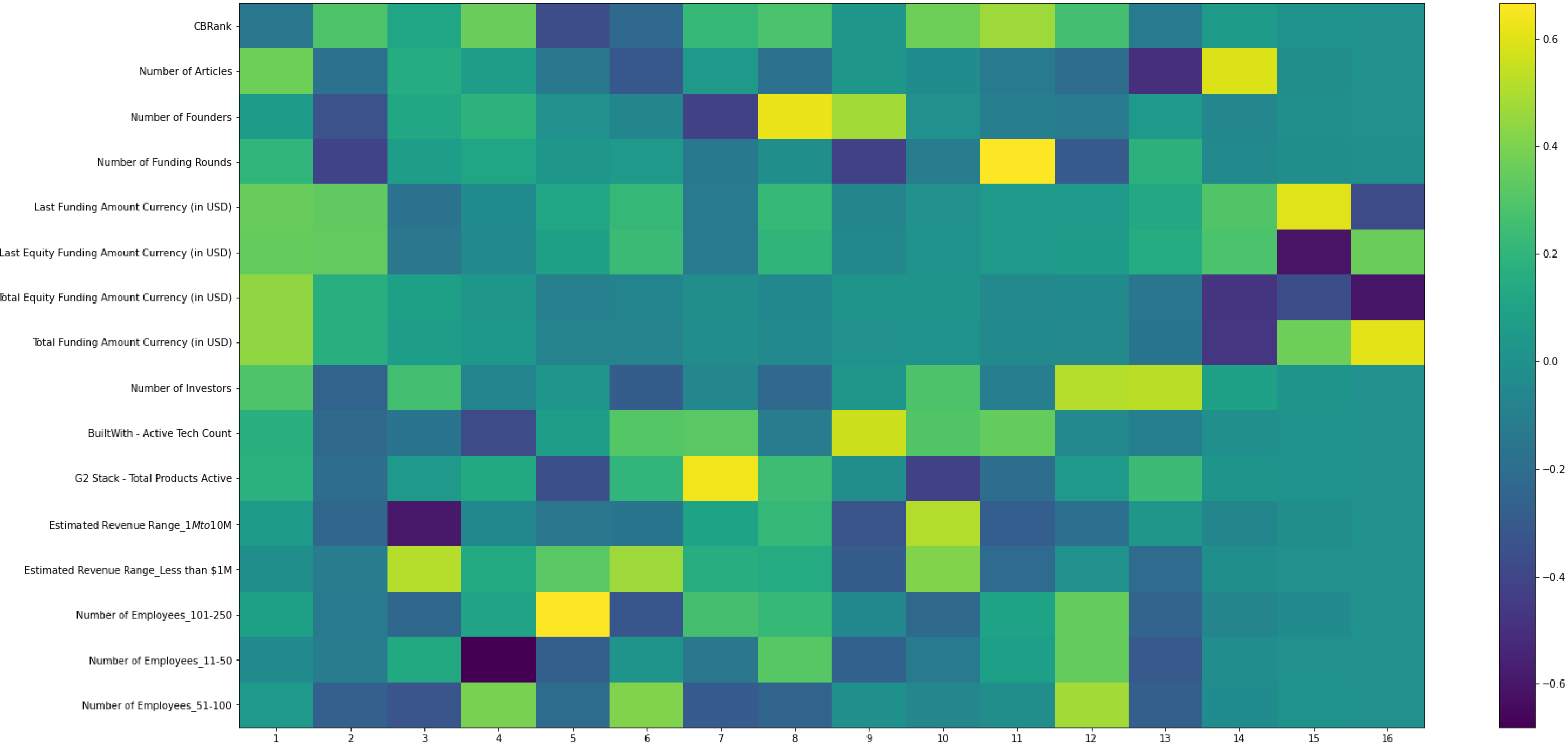
```
##############################################################################
# Centering method to variables with high p-values

# 'Last Funding Amount Currency (in USD)'
# 'Last Equity Funding Amount Currency (in USD)'
# 'Total Equity Funding Amount Currency (in USD)'
# 'Total Funding Amount Currency (in USD)'
# 'Number of Funding Rounds'
# 'Estimated Revenue Range_$1B to $10B'
# 'Number of Employees_51-100'
```

```
summary = <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  266
Model:                          Logit   Df Residuals:                      250
Method:                           MLE   Df Model:                           15
Date:                Fri, 16 Apr 2021   Pseudo R-squ.:                  0.9315
Time:                        23:12:52   Log-Likelihood:                -12.622
converged:                       True   LL-Null:                       -184.38
Covariance Type:            nonrobust   LLR p-value:                 4.778e-64
=================================================================================================================
                                                    coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------------
CBRank                                           -0.0005      0.001     -0.763      0.446      -0.002       0.001
Number of Articles                                0.7988      1.507      0.530      0.596      -2.155       3.753
Number of Founders                               -1.5615      2.486     -0.628      0.530      -6.434       3.311
Number of Funding Rounds                         16.1836     21.074      0.768      0.443     -25.121      57.488
Last Funding Amount Currency (in USD)           147.0147    165.626      0.888      0.375    -177.606     471.636
Last Equity Funding Amount Currency (in USD)   -137.5199    153.384     -0.897      0.370    -438.146     163.106
Total Equity Funding Amount Currency (in USD)   249.4732    280.279      0.890      0.373    -299.863     798.809
Total Funding Amount Currency (in USD)         -259.5286    293.149     -0.885      0.376    -834.091     315.033
Number of Investors                               1.1946      1.074      1.112      0.266      -0.910       3.299
BuiltWith - Active Tech Count                    -0.0737      0.188     -0.393      0.694      -0.441       0.294
G2 Stack - Total Products Active                 -2.6219      3.751     -0.699      0.485      -9.974       4.730
Estimated Revenue Range_$1M to $10M               6.1228      7.607      0.805      0.421      -8.786      21.031
Estimated Revenue Range_Less than $1M            -2.9936      2.066     -1.449      0.147      -7.042       1.055
Number of Employees_101-250                     -19.1707     21.586     -0.888      0.374     -61.478      23.137
Number of Employees_11-50                       -17.6609     21.515     -0.821      0.412     -59.829      24.507
Number of Employees_51-100                      -13.6096     18.453     -0.738      0.461     -49.777      22.558
=================================================================================================================

Possibly complete quasi-separation: A fraction 0.79 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.
"""
```

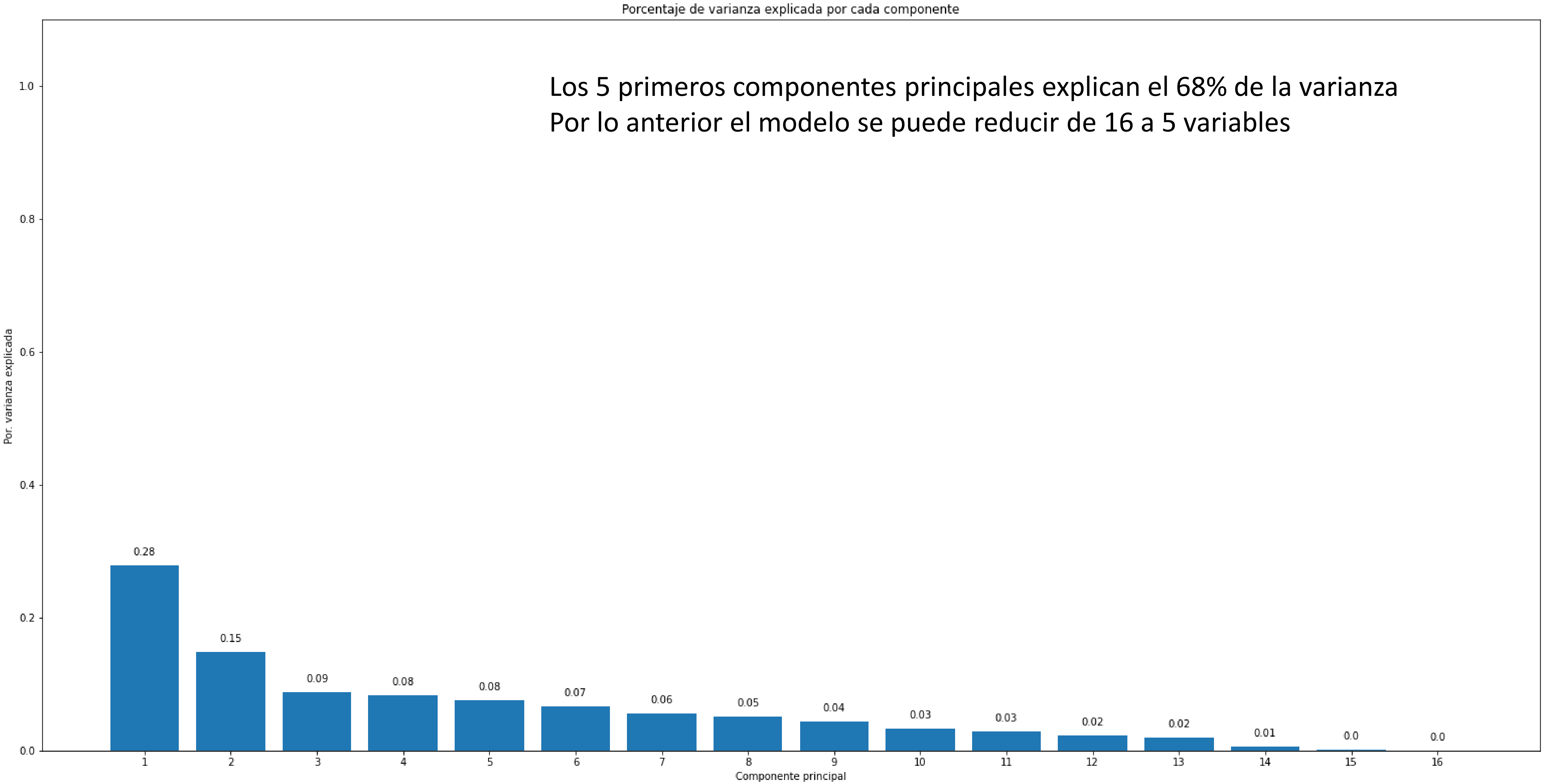
Mejoran los p-valores
y R-Cuadrado

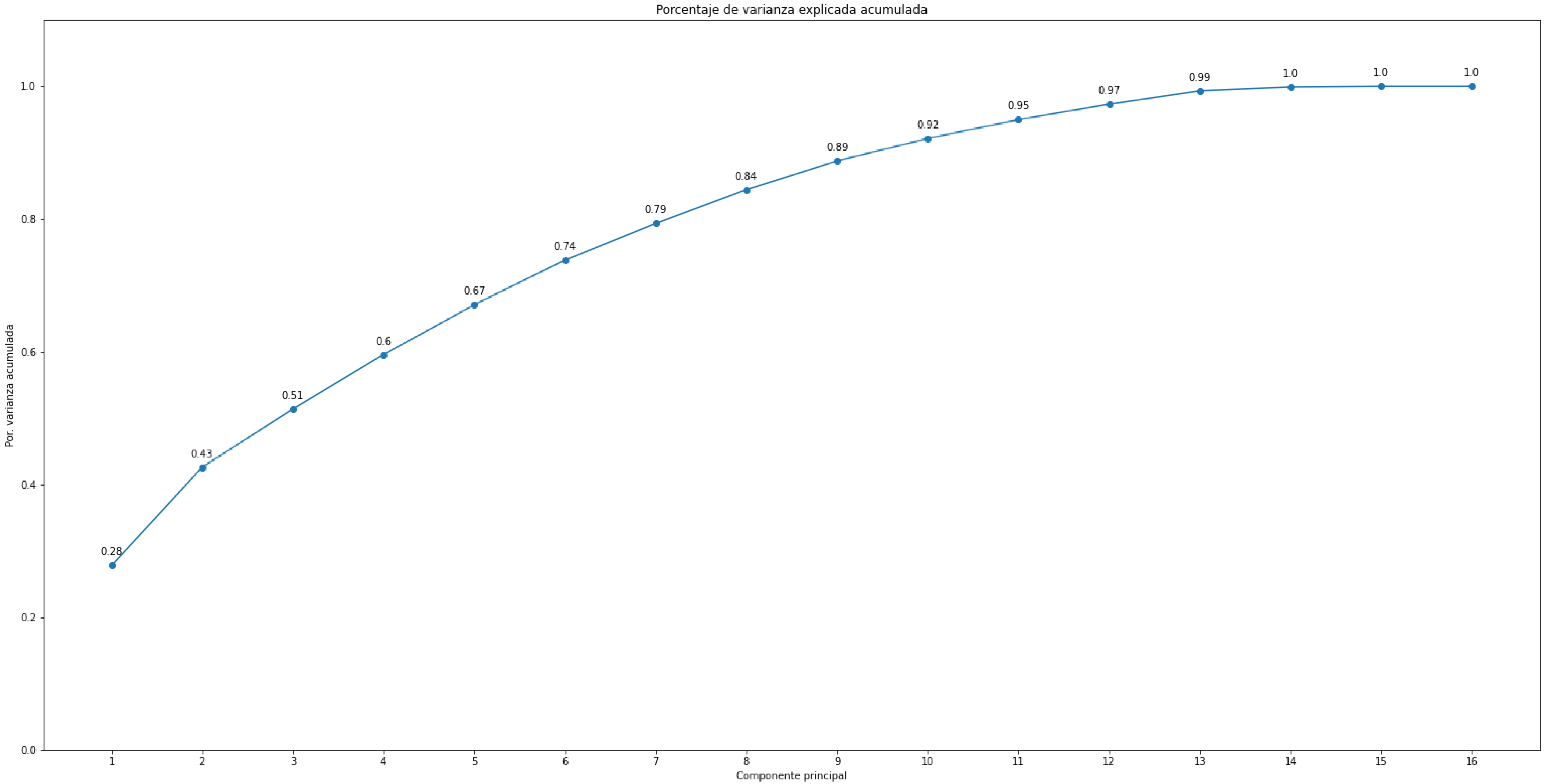
PCA

Análisis de Componente Principal con variables originales sin re-escalar



Los 5 primeros componentes principales explican el 68% de la varianza
Por lo anterior el modelo se puede reducir de 16 a 5 variables




Porcentaje de varianza explicada acumulada
Por. varianza acumulada
Componente principal
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
0.28
0.43
0.51
0.6
0.67
0.74
0.79
0.84
0.89
0.92
0.95
0.97
0.99
1.0
1.0
1.0

Creación de DataFrame de Componentes Principales para conjunto de datos de entrenamiento

```
        PC1       PC2       PC3  ...      PC14      PC15      PC16
0  3.124126 -4.077189 -1.432224  ... -0.141630 -0.146836 -0.009238
1  0.398145 -0.738557  0.626767  ...  0.374750 -0.006924 -0.003259
2  0.840432 -1.370550  1.478440  ...  0.874954 -0.005814  0.001624
4  2.213896 -4.784303 -2.253603  ... -0.360776 -0.006310  0.044059
5  1.510527 -3.687383  0.533599  ... -0.106914  0.064659 -0.005581

[5 rows x 16 columns]
```

Modelo de Regresión con Componentes Principales

```
summary = <class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  266
Model:                          Logit   Df Residuals:                      261
Method:                           MLE   Df Model:                            4
Date:                Wed, 14 Apr 2021   Pseudo R-squ.:                  0.4011
Time:                        22:37:37   Log-Likelihood:                -110.43
converged:                       True   LL-Null:                       -184.38
Covariance Type:            nonrobust   LLR p-value:                 5.755e-31
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
PC1            0.1122      0.094      1.196      0.232      -0.072       0.296
PC2           -0.6993      0.129     -5.420      0.000      -0.952      -0.446
PC3           -0.0393      0.136     -0.288      0.773      -0.306       0.228
PC4           -0.3711      0.152     -2.446      0.014      -0.668      -0.074
PC5            0.5151      0.185      2.783      0.005       0.152       0.878
==============================================================================
"""
```

Matriz de confusión y accuracy Test
Para datos de Testing

```
Confusion Matrix :
 [[37 21]
 [ 0  3]]
Test accuracy =  0.6557377049180327
```

Matriz de confusión y accuracy Test
Para datos de Training

```
Confusion Matrix :
 [[ 79  54]
 [  0 133]]
Test accuracy =  0.7969924812030075
```

Con este método se gana mejora el uso de memoria del PC, se pierde R2, sin embargo la exactitud del modelo esta entre 65 y 80 %, lo cual es muy razonable y lógico