CONTRIBUTING INSIGHTS TO OBTAIN A SERIES B FUNDING ROUND FOR ARCUS FINANCIAL INTELLIGENCE

Abstract

The company Arcus Financial Intelligence seeks to expand its services and grow its presence in the United States and Latin America, but for that they need to obtain financing in a Series B funding round. Arcus wants through this study to improve their understanding of the socio economic components that enable Fintech emergence in Latin America and to have powerful insights as to how the market trends are moving in recent years for B2B and B2C Fintech companies in Latin America. The aim of this project is to provide the company in question powerful insights through multiple tested hypotheses, visualizations and forecasts that provide a better understanding of the Fintech market in Latin America.

Context

This work combines exploratory analysis through summary statistics tables and data visualizations and explanatory analysis by hypothesis testing for multiple geographic region although it is focused in Latin America.[1] The dissertation finishes with a time series analysis to forecast Fintech emergence in the next three years in Latin America.[2]

Workflow

This implementation of this dissertation consists of three parts:

part I: International case
The main objective of this section is to explore the significance of six hypotheses explored by Christian Haddad in a different time period (2009-2009) for a different sample of countries (17) with a self developed technique (weighted variable testing Pearson correlation one tailed test).

part II: Latin American case
Part I will be contrasted with a sample of Latin American countries. In particular the significance level of the hypotheses will be compared against the international case (part I) and the overall shift of components in recent years will be visualized and tested.

part III: Fintech B2B-B2C market trend analysis
A set of new hypotheses personally assigned will be tested with a Student's T test to gain powerful insights of Fintech market trends of recent years in Latin America. Times series analysis will be applied to forecast new Fintech company emergence in the next three years.

Technical Applications

Programming Language: Python
Programming Environment: Google Colab
Data Visualization: Matplotlib*[3] , seaborn* , Google Sheets graphics.
Data Engineering: Google sheets, pandas* , fancyimpute* , numpy*, sklearn*
Data extraction: Crunchbase Pro, World Bank, and World Economic Forum.
Hypothesis testing: scipy.stats*
Machine Learning: sklearn*, statsmodels*
Data Science techniques and domains: imputation with predictive algorithms, data pre processing, hypothesis testing, time series analysis, variable architecture, data visualization, machine learning, exploratory analysis, storytelling.

[1] Specifically the sample frequently used contains all Southamerican countries and Mexico.
[2] I have a personal motivation to explore for the first time time series.
[3] * *package from Python*

Implementation

Part I.

Socioeconomic component analysis International case

Dataset

"Mastersheet_part1.csv": shape 21 x 171

	Countries fintechs	Country	Year	1A	1B	1C	2A	2B	2C
0	4.0	Argentina	2018.0	2.447560	2.805120	4.061309	3.430802	4.195683	3.930331
1	1.0	Bangladesh	2018.0	2.578552	3.683762	4.117619	3.058403	3.355734	3.716894
2	32.0	Brazil	2018.0	2.459301	3.302393	4.566283	3.238514	4.296651	4.089989
3	30.0	Canada	2018.0	3.689212	4.582466	5.105235	4.659581	4.638758	4.652790
4	1.0	Chile	2018.0	3.504282	4.101409	3.883189	3.885442	3.263582	3.997062

Sources

Crunchbase *, World Economic Forum "World Competitiveness report" **, World Bank***

Target variable

Countries Fintechs* : the number of founded Fintechs in a particular year for a specific country.

Independent variables used

Venture Capital availability** → renamed as "1A"
Availability of financial services** → renamed as "1B"
Affordability of financial services**→ renamed as "1C"
Quality of education** → renamed as "2A"
Internet access in schools** → renamed as "2B"
Availability of latest technologies** → renamed as "2C"
FDI on technology** → renamed as "2D"
Individuals using internet** → renamed as "2E"
Availability of financial services** → renamed as "3A"
Affordability of financial services** → renamed as "3B"
Access to loans** → renamed as "3C"
Soundness of financial sector** → renamed as "3D"
Mobile broadband subscriptions** → renamed as "4A"
Mobile telephone subscriptions ratio** → renamed as "4B"
Domestic market size index** → renamed as "5A"
Foreign market size index** → renamed as "5B"
Unemployment rate*** → renamed as "6A"

Time period

2009 - 2019

Countries

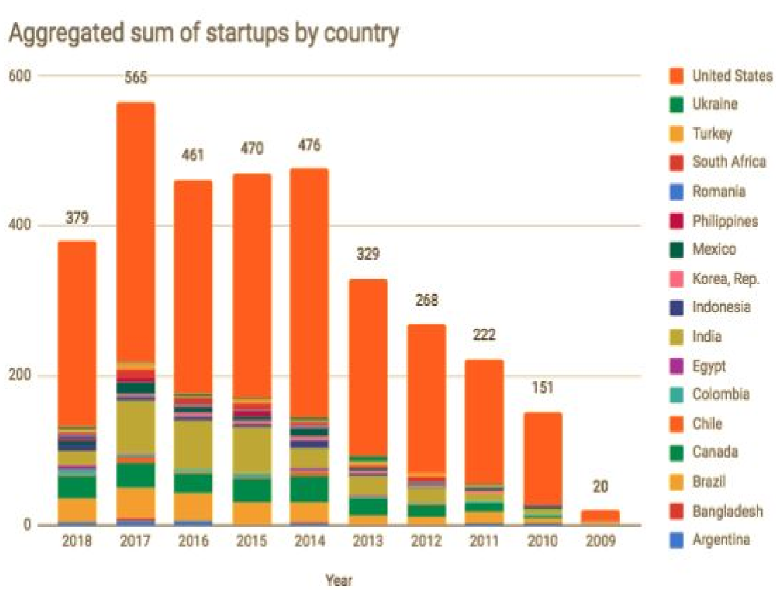
Two criteria is to get a (1) balanced set of countries in terms of economic power and development including 5 highest GDP economies in LatAm, (2) Information completeness.

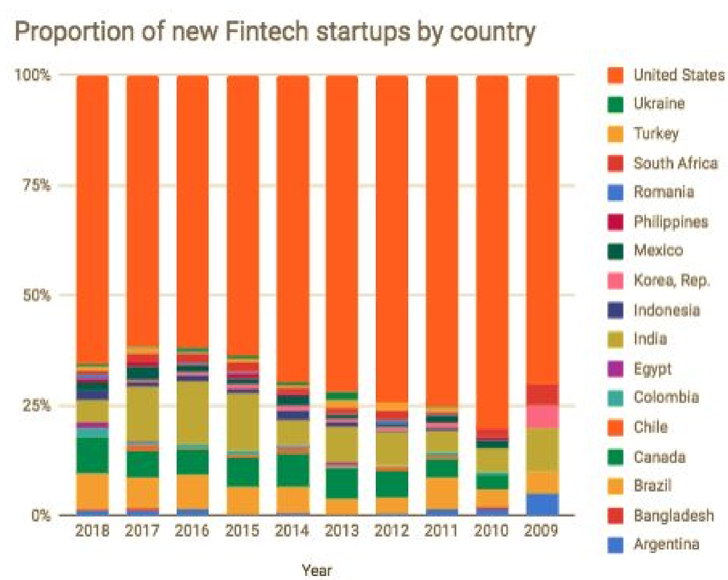
5 highest GDP in Latin America: Brazil, Mexico, Colombia, Chile, Argentina
3 High income United States, Rep. of Korea, Canada
3 middle high income Turkey, South Africa, Romania
5 lower middle emerging emerging: India, Egypt, Indonesia, Bangladesh, Philippines
1 lower middle income: Ukraine

Completeness
The following columns contained missing values: ["1B", "1C", "2B", "2D", "2E", "3A","3B","3C","4A","5B","5C"] **deterministic linear model** was used to estimate the missing values.

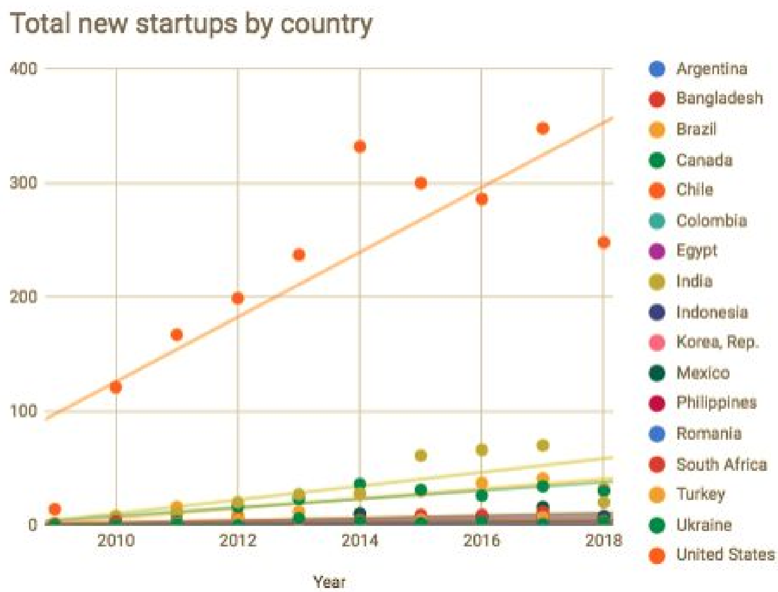
Exploratory Analysis part I

The following graphs explore new Fintech startups emergence in the sample. The first graph shows the aggregate sum of new companies, being 2018 the year with most new startups (565) and 2009 year with least (20). The second visual illustrates that the United States concentrates more than Fintechs followed the sample followed by India in every .

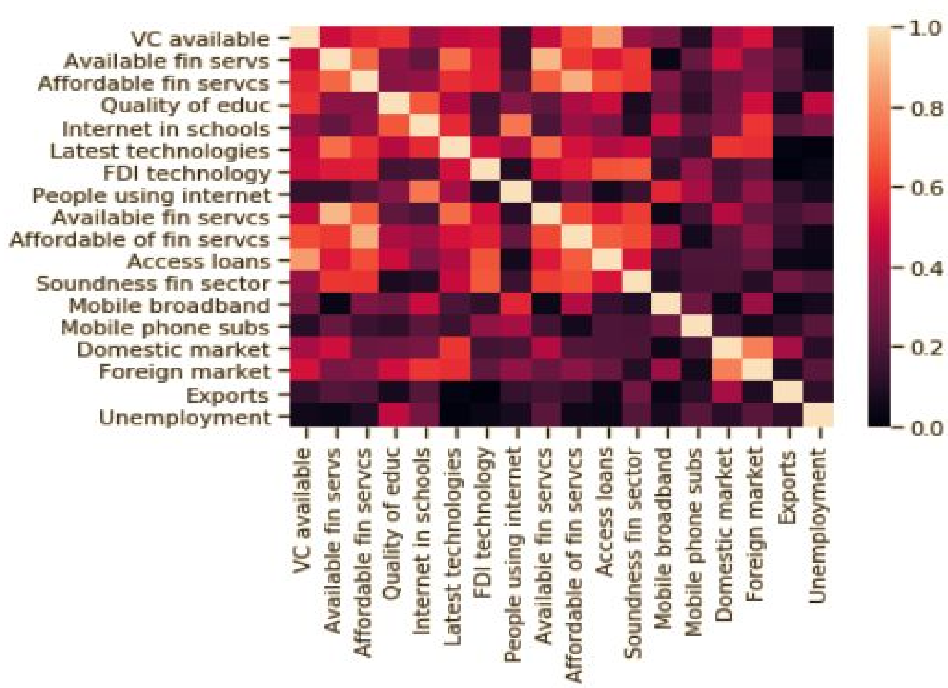




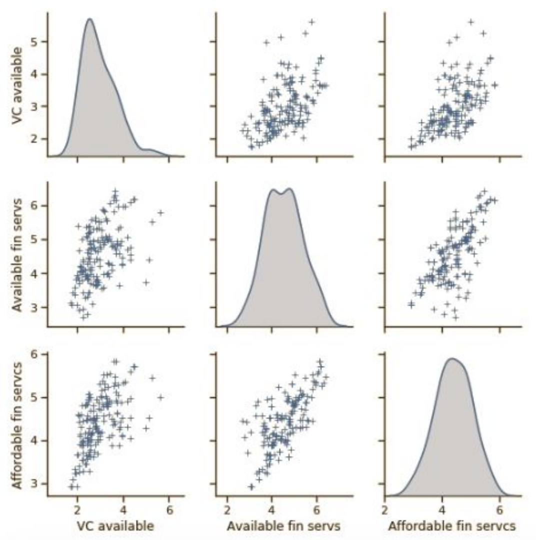
There is a clear growing tendency of emergence of new companies illustrated by the tendency lines.

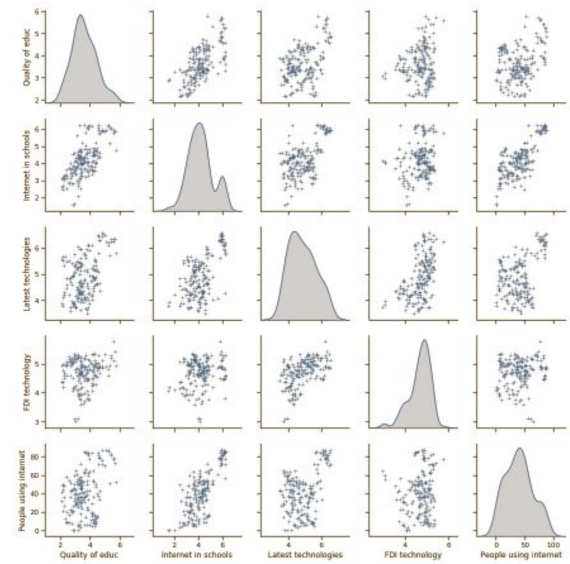


The next graph shows correlation in a heatmap fashion. Lighter colors demonstrate high correlation in absolute number, while dark colours show the opposite.

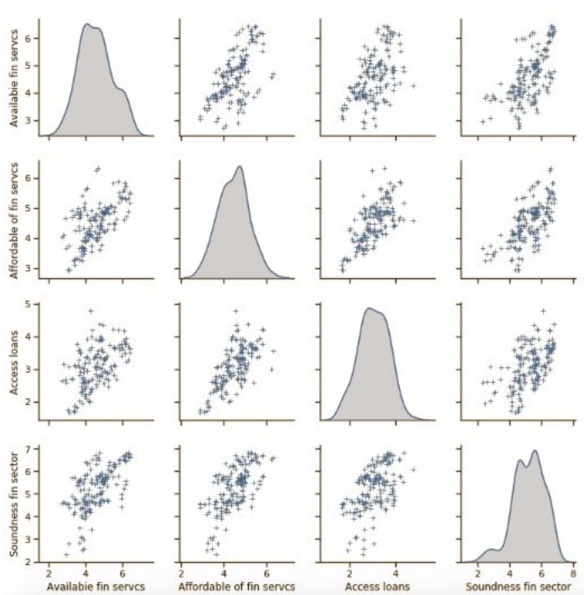


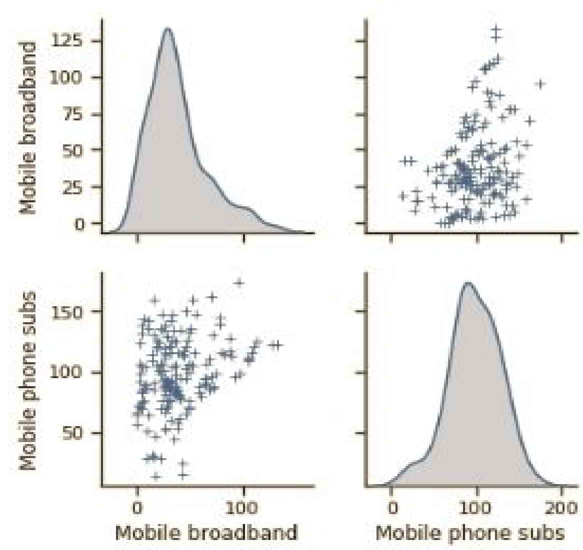
The graph on the left shows contain distribution and pairplots for the following capital development related variables in the following order: Venture capital availability, Availability of financial services, Affordability of financial services. The second one contains the same information for technology available variables: Quality of education, Internet access in schools, Availability of latest technologies, FDI on technology, and Individuals using internet.



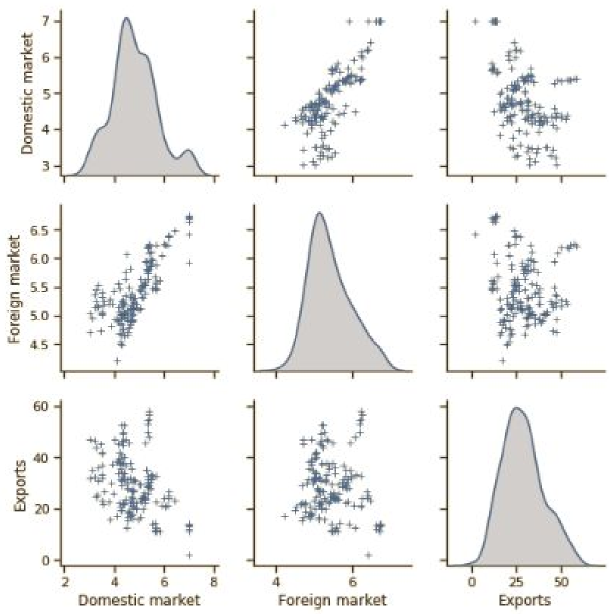


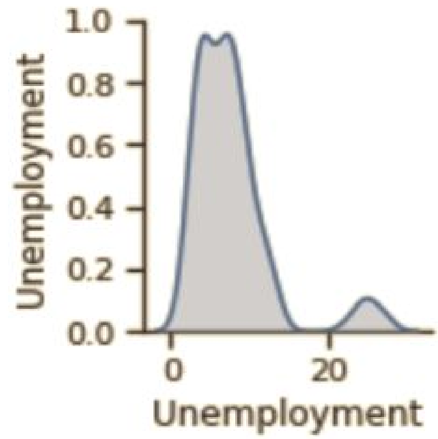
Below is the normality and pairplot relation among financial soundness variables (5) and mobile phone use variables (2).



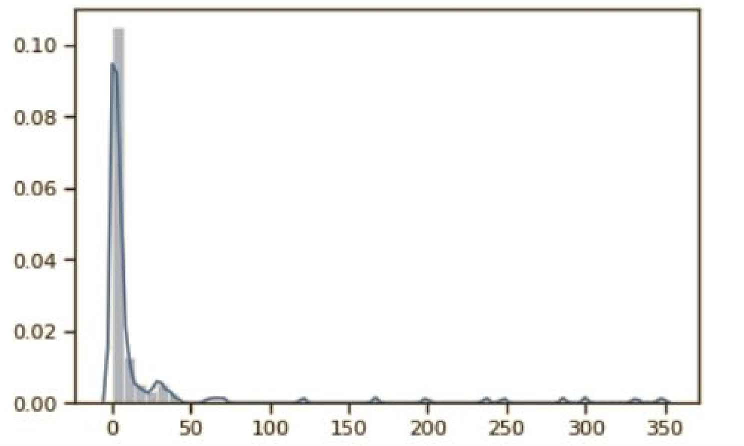


The next visuals is the normality and pairplot relation among financial market size variables (3) and the unemployment variable, which was extracted from the World Bank data.





The next visual illustrates the target variable which contains the total new Fintech startups in periods of one year from 2009 to 2018. It demonstrates that the data does have a normal distribution since most data points concentrate from zero to 50, but there is a major outlier, which is the United States which consistently has more than 100 new Fintech startups per year.

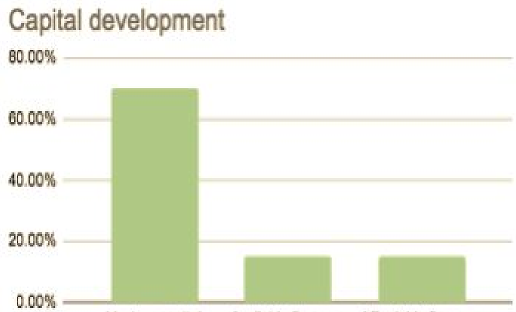


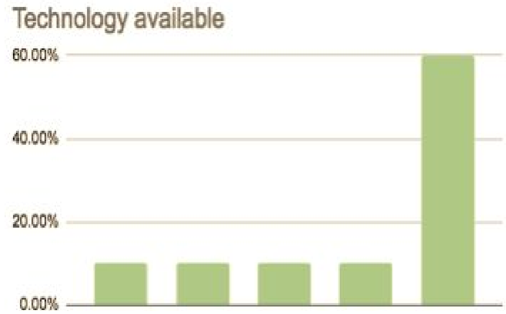
Weighted variables

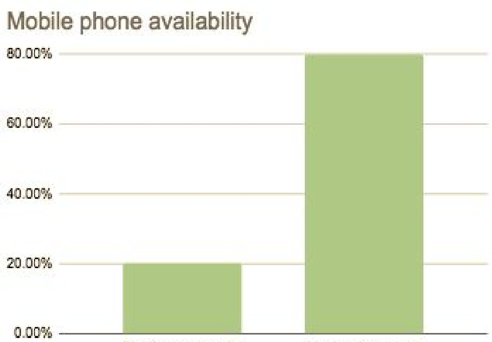
As seen in the previous section, each variable used for this part of the dissertation belongs to a category which can be, capital, technology, financial, mobile telephone, market size, or unemployment. Each category has a called *master variable* that is composed of the weighted values of all the variables that are contained in that category and they all sum 100% of the variable. The master variable contains a percentage of the contingent variables like follows:

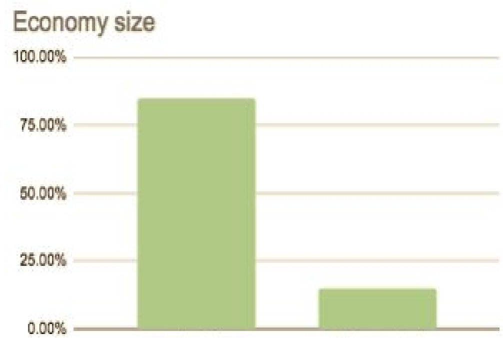
Master capital = (1A*0.7)+(1B*0.15)+(1C*0.15)
Master technology = (2A*0.1)+(2B*0.1)+(2C*0.6)+(2D*0.1)+(2E*0.1)
Master financial = (3A*0.05)+(3B*0.05)+(3C*0.05)+(3D*0.85)
Master mobile phone = (4A*0.2)+(4B*0.8)
Master economy size = (5A*0.9)+(5B*0.1)
Master unemployment = 6A

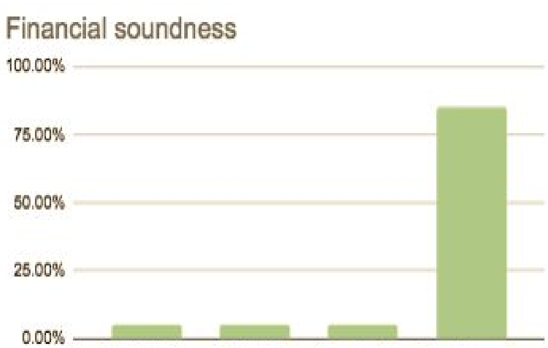
This entails that each value contained in a master variable is the sum of the weighted values of the variables contained in that category for each row. The next graphs demonstrated the weighted value of variables for each master variable.

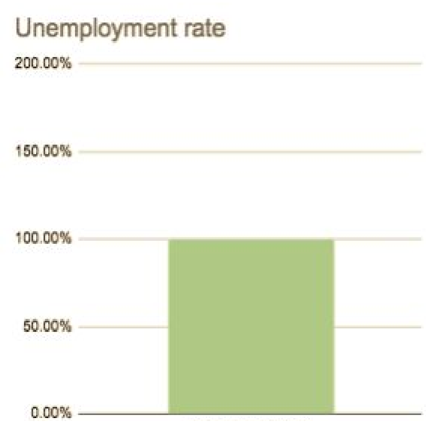












Data Preprocessing

Data standardization was performed for master variables with preprocessing module from sklearn Python package.

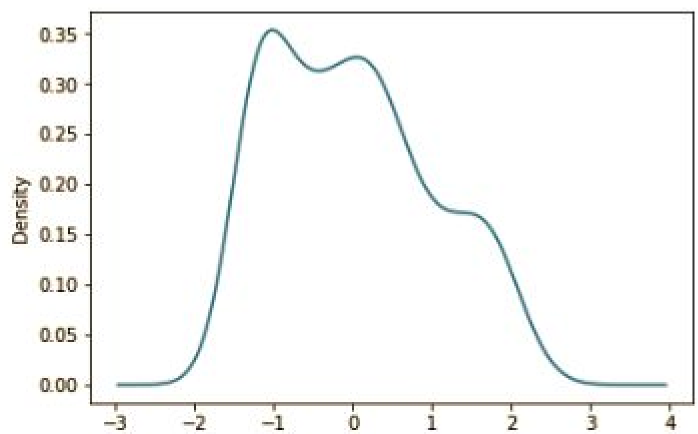
The command was the following:

- *df_processed = preprocessing.StandardScaler().fit_transform(df_weights)*[4]

Below are the summary statistics of the center and scaled dataset where mean of values is 0 with a marginal difference and the standard deviation is 1 with a miniscule deviation.

	Master_Target	Master_one	Master_two	Master_three	Master_four	Master_five	Master_six
count	1.700000e+02	1.700000e+02	1.700000e+02	1.700000e+02	1.700000e+02	1.700000e+02	1.700000e+02
mean	1.345329e-16	2.533921e-16	1.253899e-16	5.498869e-16	3.608225e-16	4.649875e-16	9.992007e-17
std	1.002954e+00	1.002954e+00	1.002954e+00	1.002954e+00	1.002954e+00	1.002954e+00	1.002954e+00
min	-1.239331e+00	-2.006937e+00	-1.921928e+00	-2.969766e+00	-2.730663e+00	-2.021888e+00	-1.068088e+00
25%	-9.420275e-01	-7.363685e-01	-6.683470e-01	-6.644969e-01	-5.917480e-01	-6.015174e-01	-6.728093e-01
50%	-9.102850e-02	-1.492681e-01	-1.856960e-01	1.960577e-01	-4.653714e-02	-1.801731e-01	-1.612632e-01
75%	7.129175e-01	6.913757e-01	3.184062e-01	6.281572e-01	7.117465e-01	6.796194e-01	2.574732e-01
max	2.219003e+00	3.270443e+00	2.311652e+00	1.700850e+00	2.795885e+00	2.388033e+00	3.799977e+00

However, as seen in the following visual. The variables not necessarily have a normal distribution. Below is the distribution of the target master variables which is center and scaled:



To assess lack of normality in the distribution, the *Normalizer* command from the preprocessing module of sklearn was used as below:

- *normalizer = preprocessing.Normalizer().fit(df_processed)*

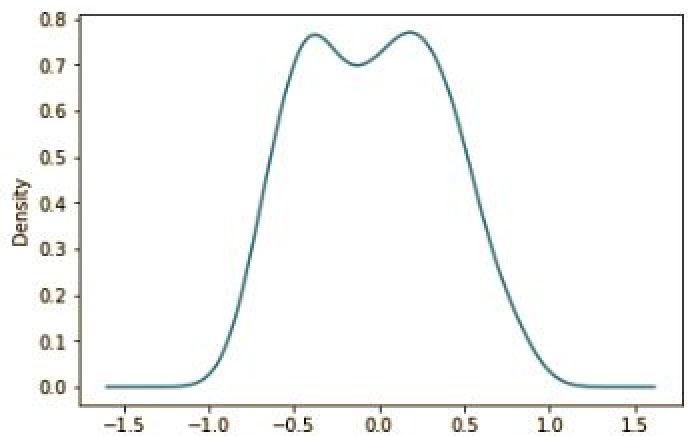
Now the remaining dataset has the following summary statistics, where the mean, the range of values, and deviation, differ from the previous table.

---

[4] *df_weights* contains the unscaled values of a master variable dataset.

	Master_Target	Master_one	Master_two	Master_three	Master_four	Master_five	Master_six
count	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000
mean	-0.035363	-0.039938	-0.027425	0.018928	-0.020067	-0.024902	-0.023457
std	0.405696	0.373298	0.353210	0.390303	0.385002	0.383123	0.352580
min	-0.796123	-0.784726	-0.653000	-0.854506	-0.814434	-0.855128	-0.725385
25%	-0.394860	-0.349482	-0.302073	-0.281068	-0.323934	-0.277682	-0.295066
50%	-0.045881	-0.083153	-0.098409	0.097145	-0.017768	-0.077127	-0.070114
75%	0.275142	0.311096	0.178392	0.353097	0.307789	0.288469	0.109393
max	0.811511	0.846386	0.724413	0.730769	0.850500	0.693979	0.924303

Note that the distribution of the target variable now appears similar to a normal distribution.



Hypotheses

The following are considered important hypothesis to determine their results in order to gain valuable insights of socioeconomic factors the emergence of new Fintechs in our sample.

- **H1:** How capitalized an economy is, positively affects the emergence of new Fintech companies.
- **H2:** Available technology in countries foster the emergence of new Fintech companies.
- **H3:** Sound financial systems enable Fintech company emergence.
- **H4:** The mobile subscription total number is a determinant for the emergence of new Fintech companies.
- **H5:** A large market size of an economy relates positively to the emergence of new Fintech companies.
- **H6:** Higher levels of unemployment fosters emergence of new Fintech companies.

The method applied for all hypotheses is one tailed test, with Pearson's Correlation Coefficient to test whether two variables have a linear relationship. In this sense, the master target variable will be tested against the 6 master variables to determine the result of each hypothesis.

As seen explored in previous section the assumptions to perform this statistical test are fulfilled, which are the following:

- Observations in each variable are independent.
- Observations in each variable are normally distributed.
- Observations in each variable have the same variance.

Below is the resulting correlation between master variables.

	Master_Target	Master_one	Master_two	Master_three	Master_four	Master_five	Master_six
Master_Target	1.000000	0.222047	0.369964	0.220902	0.243976	0.345109	-0.069661
Master_one	0.222047	1.000000	0.459184	0.385297	-0.084341	0.269466	-0.121491
Master_two	0.369964	0.459184	1.000000	0.422635	0.082841	0.465691	-0.115146
Master_three	0.220902	0.385297	0.422635	1.000000	-0.150941	0.134943	0.163930
Master_four	0.243976	-0.084341	0.082841	-0.150941	1.000000	-0.224944	0.127224
Master_five	0.345109	0.269466	0.465691	0.134943	-0.224944	1.000000	-0.073717
Master_six	-0.069661	-0.121491	-0.115146	0.163930	0.127224	-0.073717	1.000000

The hypothesis interpretation is as follows:

- **H0 or null hypothesis:** the two variables are independent.
- **H1 or alternative hypothesis:** there exists a linear dependency between the variables.

Results

H1: How **capitalized an economy** is, positively affects the emergence of new Fintech companies.
p value**:** `0.003611959034829`
ACCEPT ALTERNATIVE HYPOTHESIS

H2: **Available technology** in countries foster the emergence of new Fintech companies.
p value**:** `6.853359914525474e-07`
ACCEPT ALTERNATIVE HYPOTHESIS

H3: **Sound financial systems** enable Fintech company emergence.
p value**:** `0.003793977886372`

ACCEPT ALTERNATIVE HYPOTHESIS

H4: Mobile subscriptions is a determinant for the emergence of new Fintech companies.
p value: `0.001345110752896`
ACCEPT ALTERNATIVE HYPOTHESIS

H5: A large **market size** relates positively to the emergence of new Fintech companies.
p value: `4.048732405146512e-06`
ACCEPT ALTERNATIVE HYPOTHESIS

H6: Higher levels of **unemployment** fosters emergence of new Fintech companies.
p value: `0.366701975270771`
ACCEPT NULL HYPOTHESIS

Accepted alternative hypothesis ordered by level of significance:

technology (Hypothesis 2) > market (Hypothesis 5) > mobile phone (Hypothesis 4) > capital (Hypothesis 1) > financial sector (Hypothesis 3)

Part II.
Latin American Case

Dataset
"Mastersheet_part2.csv": shape 21 x 131

	Countries_fintechs	CB_Ranking	Country	Year	oneA	oneB	oneC	twoA
0	4	229	Argentina	2018	2.447560	2.805120	4.061309	3.430802
1	32	300	Brazil	2018	2.459301	3.302393	4.566283	3.238514
2	1	424	Chile	2018	3.504282	4.101409	3.883189	3.885442
3	7	175	Colombia	2018	2.886229	3.680098	3.975811	3.878223
4	6	176	Mexico	2018	3.152481	3.589537	4.556806	3.626325
5	1	316	Bolivia	2018	2.150270	3.869697	4.496598	3.345480
6	2	153	Ecuador	2018	2.353610	3.398379	4.410720	3.717029

Target variables

For this part, the target variable will be a combination of two variables: Countries Fintechs and CB_Ranking.

Countries Fintechs : the number of founded Fintechs in a particular year for a specific country.
CB_Ranking: company ranking assessed by CrunchBase.

Independent variables

Independent variables are the same as in part I.

Time period
2009 - 2019

Countries
13

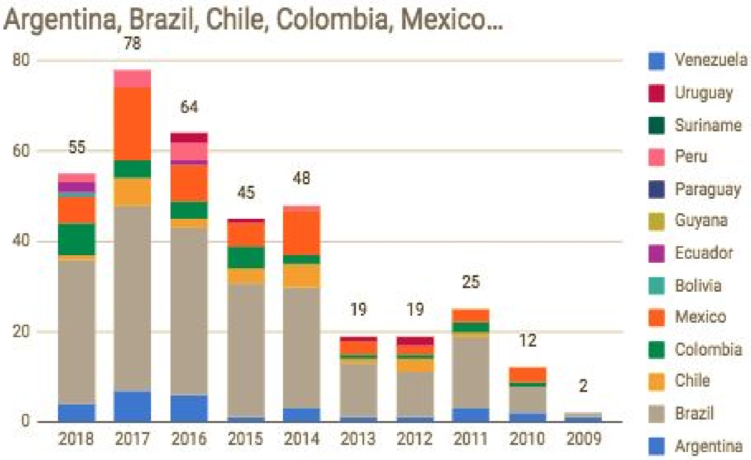
The criteria is the geographic region where the company in question wants to expand which is South America and Mexico. The countries are:

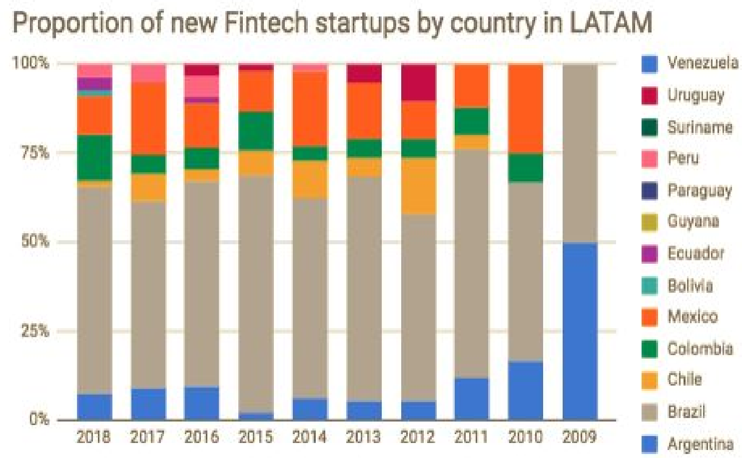
- Mexico, Brazil, Argentina, Colombia, Chile, Peru, Paraguay, Uruguay, Bolivia, Venezuela, Ecuador, Suriname and Guyana.

Completeness
The following columns contained missing values: : ["1B", "1C", "2B", "2D", "2E", "3A","3B","3C","4A","5B","5C"] this were not filled in the extracted information. **Iterative imputation by fancy impute* package from Python** was used to estimate the missing values.

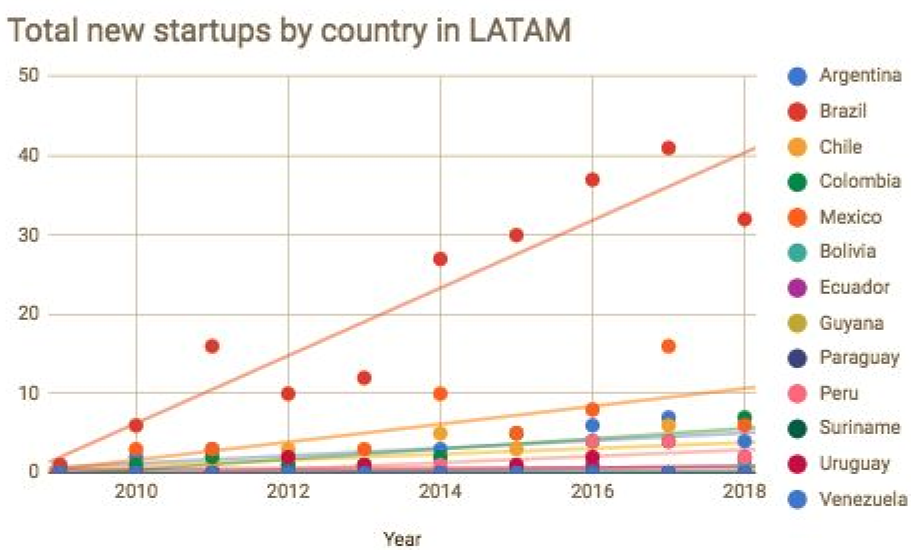
Exploratory Analysis part II

In this sample, Brazil dominates the sector and as seen in the left picture the first 5 years of the sample (2009-2013) average less number of new Fintechs than the second half (2014-2018) and 2017 is the year with more new Fintech companies in the region. The second visual in order complements the fact that Brazil has the largest share of the market, however Mexico and Argentina appear to be in the second and third position. Other countries have a represent a minor part of the Fintech companies founded in the region in that time period.

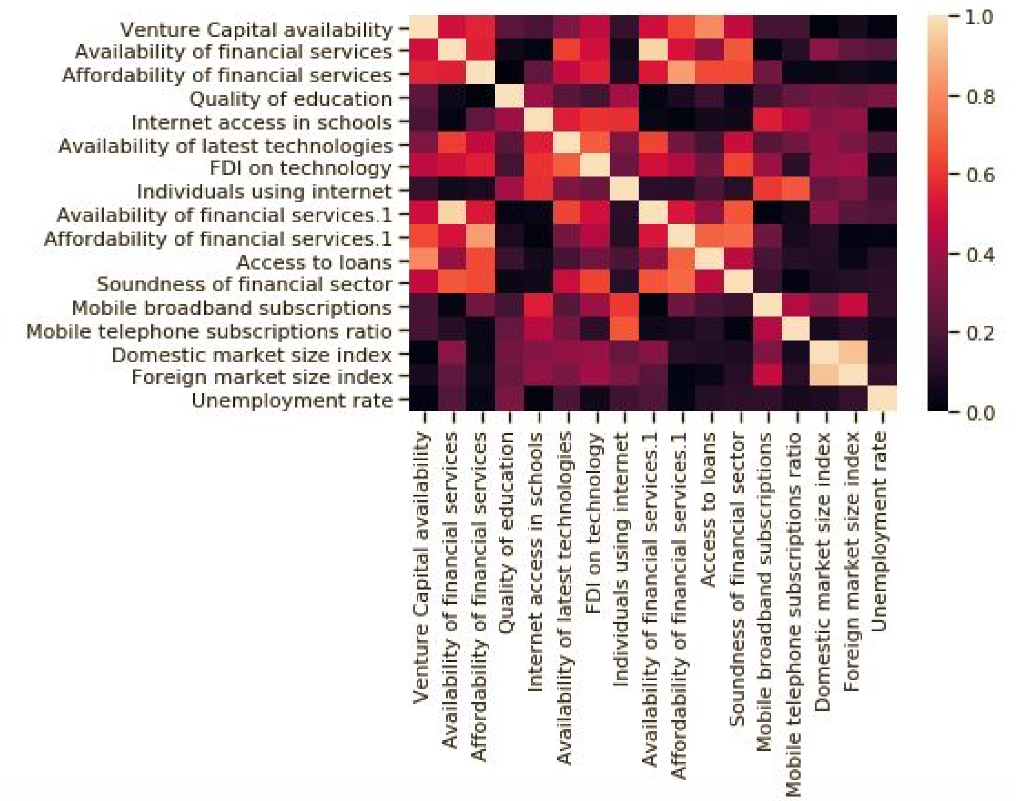




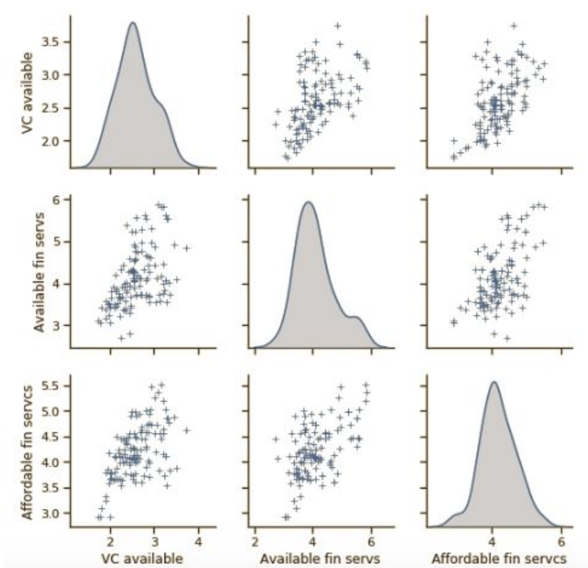
As seen below there is a growing tendency of new Fintech companies which peak in recent years for Latin American countries.

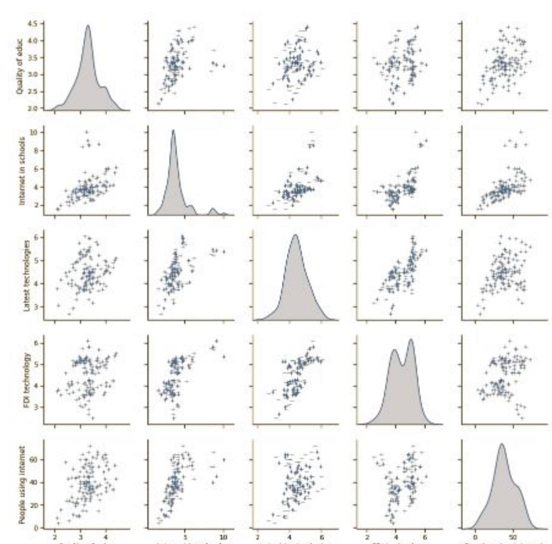


The correlation map shows interesting relations that will not be explored in depth.

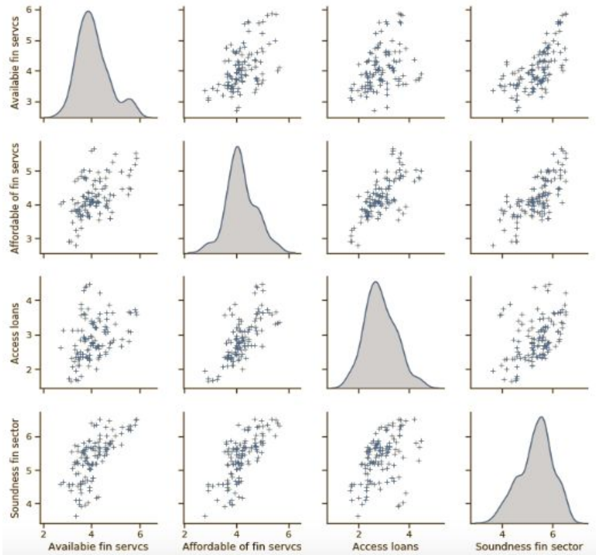


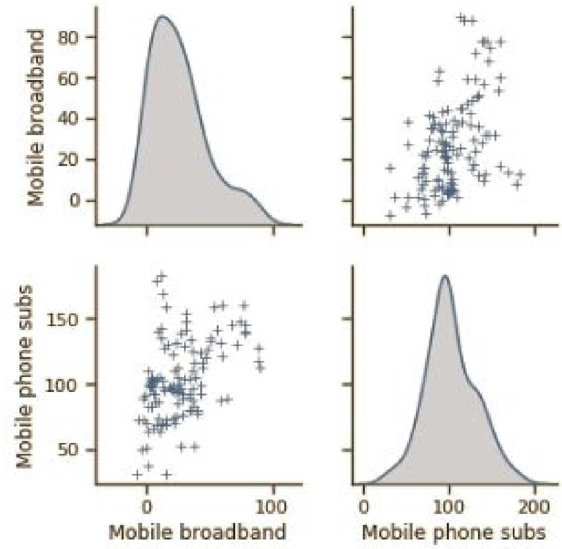
The graph on the left shows contain distribution and pairplots for the following capital development related variables in the following order: Venture capital availability, Availability of financial services, Affordability of financial services. The second one contains the same information for technology available variables: Quality of education, Internet access in schools, Availability of latest technologies, FDI on technology, and Individuals using internet.



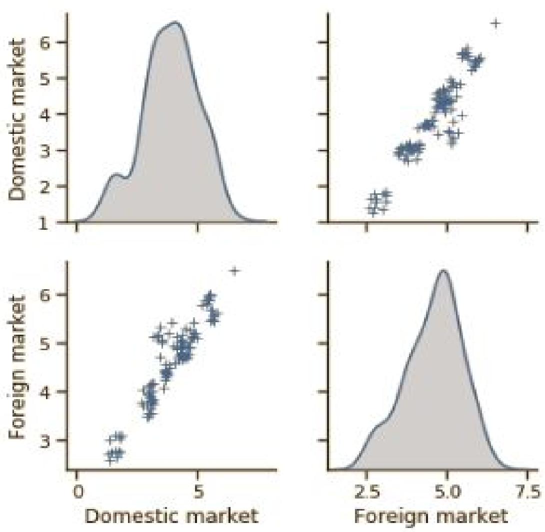


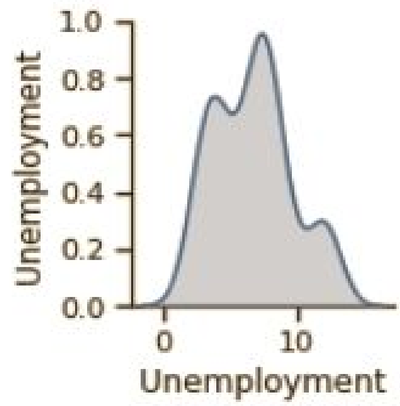
Below is the normality and pairplot relation among financial soundness variables (5) and mobile phone use variables (2).



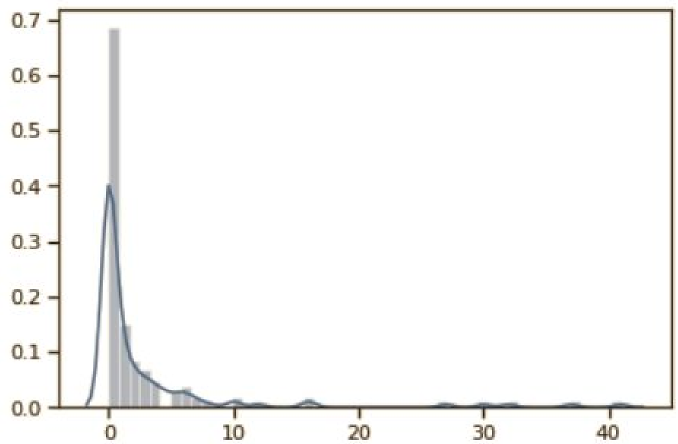


The next visuals is the normality and pairplot relation among financial market size variables (3) and the unemployment variable, which was extracted from the World Bank data.





The number of Fintechs per country for each year variable behaves similar to the international case studied in part I. It demonstrates that the data does have a normal distribution, and this time the mayor outlier is Brazil which consistently has a notably above average number of new Fintech companies each year.



Weighted variables

Weighted variables follow the same structure as in part I.

Weighted Target variable

For this part, not only the number of Fintechs will be contemplated as a target variable, but also the average Crunchbase ranking in that year will be considered. For this it there must be introduced two auxiliary variables that enable both concepts to have the same range of values.

Fintechs_decile:
Takes values from a range (0,1). It is an ordinary variable that takes the highest number of Fintech companies in a row as a 1 and the lowest number as a 0 and the remaining values take values from 0 to 1 in accordance to their ranking.

CB_decile:
Takes values from a range (0,1). It is an ordinary variable that takes 1 for the best Crunchbase ranking average of Fintech companies in a row, while 0 accounts for the lowest ranking and the remaining values take values from 0 to 1 in accordance to their ranking.

Master_Target:
This variable takes the weighted values of “Fintechs_decile” and “CB_decile”. In this case “Fintechs_decile” takes 80% of the whole variable and “CB_decile” takes the remaining 20% as follows:

Master_Target = (Fintechs_decile x 4) + CB_decile

Note direct relation between “Countries_fintech” and “Fintech_decile” variable and the indirect relation among “CB_ranking” and “CB_decile” (a high ranking means low quality and vice versa), whereas “Master Target” is the result of 4x “Fintechs_decile” plus “CB_decile”.

Countries_fintechs	CB_Ranking	Fintechs_decile	CB_decile	Master_Target	Country	Latam	Year
4	229	0.797	0.642	2.046696	Argentina	1	2018
32	300	0.976	0.593	2.315072	Brazil	1	2018
1	424	0.537	0.561	1.205028	Chile	1	2018
7	175	0.894	0.715	2.556840	Colombia	1	2018
6	176	0.862	0.699	2.410152	Mexico	1	2018

Preprocessing

Data preprocessing follows the exact same process as in part one where:

- Data standardization was performed for master variables with preprocessing module from sklearn Python package.
- Normality in the sample was assured, the *Normalizer* command from the preprocessing module of Python’s sklearn.

The remaining dataset containing only master variables which are centered, scaled and normalized has the following first rows:

	Master_Target	Master_one	Master_two	Master_three	Master_four	Master_five	Master_six
count	130.000000	130.000000	130.000000	130.000000	130.000000	130.000000	130.000000
mean	-0.024455	-0.012779	-0.015403	-0.006703	-0.013165	-0.006263	-0.014799
std	0.443610	0.355034	0.356775	0.365946	0.356899	0.363020	0.403808
min	-0.833976	-0.707899	-0.708827	-0.772367	-0.659865	-0.727285	-0.763423
25%	-0.381084	-0.263427	-0.277221	-0.261839	-0.269088	-0.242225	-0.370242
50%	-0.242583	-0.032275	-0.050160	0.050909	-0.117549	0.014519	0.025031
75%	0.414759	0.271353	0.236364	0.262070	0.247888	0.265398	0.243196
max	0.890645	0.699971	0.755580	0.581655	0.770594	0.645470	0.931286

While the master target variable has a normal distribution.

The next table shows the correlation between master variables for the Latin American sample.

	Master_Target	Master_one	Master_two	Master_three	Master_four	Master_five	Master_six
Master_Target	1.000000	0.215843	0.431304	0.359351	0.383174	0.509423	0.106377
Master_one	0.215843	1.000000	0.386681	0.662622	-0.048118	0.124221	-0.014651
Master_two	0.431304	0.386681	1.000000	0.440742	0.467604	0.313826	0.153095
Master_three	0.359351	0.662622	0.440742	1.000000	0.067570	0.116716	0.051079
Master_four	0.383174	-0.048118	0.467604	0.067570	1.000000	0.051741	0.088565
Master_five	0.509423	0.124221	0.313826	0.116716	0.051741	1.000000	-0.006851
Master_six	0.106377	-0.014651	0.153095	0.051079	0.088565	-0.006851	1.000000

The hypothesis interpretation is as follows:

- **H0 or null hypothesis:** the two variables are independent.
- **H1 or alternative hypothesis:** there exists a linear dependency between the variables.

* Note that the following hypothesis are the same as in part one. However the geographic region studied is different. This is done so, to identify any unique characteristics of the socioeconomic factors that enable Fintech company emergence in Latin America in comparison to the international realm.

- **H1:** How capitalized an economy is, positively affects the emergence of new Fintech companies in the sample studied.
- **H2:** Available technology in countries foster the emergence of new Fintech companies in the sample studied.
- **H3:** Sound financial systems enable Fintech company emergence in the sample studied.

- **H4:** The mobile subscription total number is a determinant for the emergence of new Fintech companies in the sample studied.
- **H5:** A large market size of an economy relates positively to the emergence of new Fintech companies in the sample studied.
- **H6:** Higher levels of unemployment fosters emergence of new Fintech companies in the sample studied.

Method and assumptions are the same as in part I.[5]

Results

H1: How **capitalized an economy** is, positively affects the emergence of new Fintech companies.
p value: `0.013649375512200086`
ACCEPT ALTERNATIVE HYPOTHESIS

H2: Available technology in countries foster the emergence of new Fintech companies.
p value: `3.0082661727792637e-07`
ACCEPT ALTERNATIVE HYPOTHESIS

H3: Sound financial systems enable Fintech company emergence.
p value: `2.6833653524601116e-05`
ACCEPT ALTERNATIVE HYPOTHESIS

H4: Mobile subscriptions is a determinant for the emergence of new Fintech companies.
p value: `6.807286556366098e-06`
ACCEPT ALTERNATIVE HYPOTHESIS

H5: A large **market size** relates positively to the emergence of new Fintech companies.
p value: `6.032682940493941e-10`
ACCEPT ALTERNATIVE HYPOTHESIS

H6: Higher levels of **unemployment** fosters emergence of new Fintech companies.
p value: `0.22835972790643308`
ACCEPT NULL HYPOTHESIS

Accepted alternative hypothesis ordered by level of significance:

Market size (Hypothesis 5) > technology (Hypothesis 2) > mobile phones (Hypothesis 4) > financial soundness (Hypothesis 3) > capital (Hypothesis 1)

---

[5] Method: one tailed Peason correlation test. Assumptions: normality of variables, independence of variables, and the same variance.

Observations

- The same alternative hypothesis are accepted in the International case (part I) and the Latin American case (part II).
- Level of significance vary from both regions.
- Order of significance between regions differ.

Hypothesis 7 Determinants for Fintech expansion are increasingly favorable in recent years.

Methodology

All master variables that prove to have significance in order to explain linear dependencies with Fintech company emergence in the Latin American sample will be considered.[6]

A variable called score is built, where each of the variables have the same weight regardless of different levels of significance. Each of the five variables account for 20% of the weighted value of the score as follows[7]:

Score = (master1*0.2)+(master2*0.2)+(master3*0.2)+(master3*0.2)+(master4*0.2)+(master5*0.2).

Each datapoint is the result of the division of each value by the maximum value of that variable of column multiplication by two (in order to have 0 as the minimum value addressable and 10 the maximum value possible) as follows:

Datapoint = (x/(max value in column)) x 2

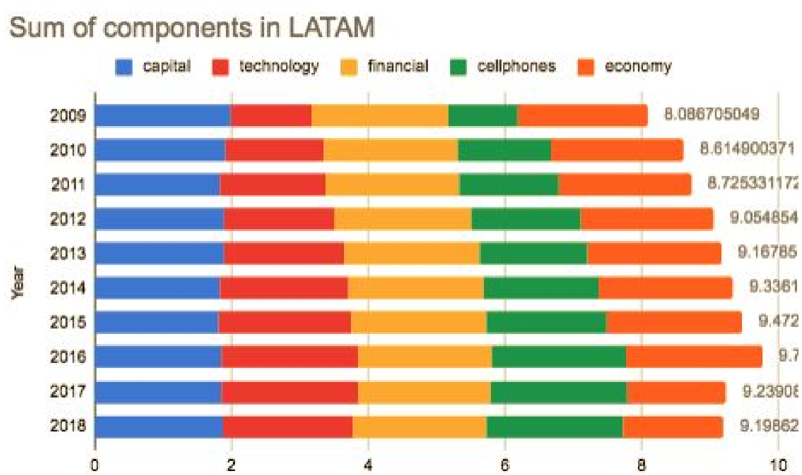
The resulting table is and graphs are as follows, where the column “General” is the horizontal sum of each datapoint:

[6] master1 (capital), master2 (technology) , master3 (financial soundness), master4 (mobile phones), master5 (market size).

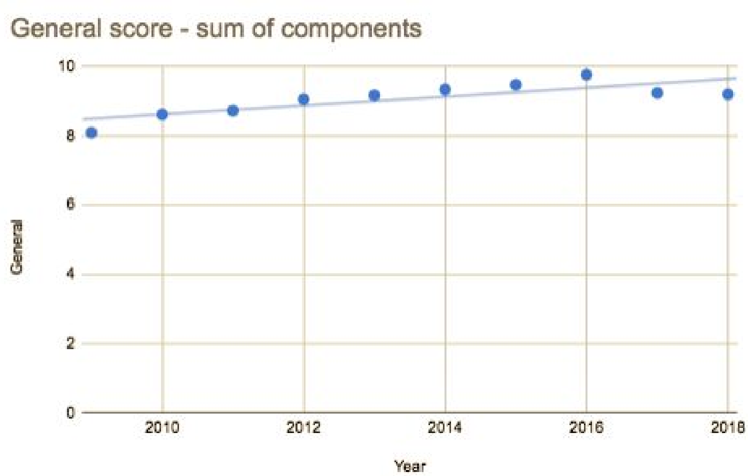
[7] All values from master variables already centered, scaled and normalized as in previous hypotheses in part II.

Year	capital	technology	financial	cellphones	economy	General
2009	1.997309014	1.180095543	1.988182027	1.024323552	1.896794914	8.086705049
2010	1.913399907	1.430607921	1.968264065	1.36356059	1.939067889	8.614900371
2011	1.827337523	1.537778032	1.965718548	1.462135484	1.932361585	8.725331172
2012	1.889808149	1.619409856	1.999935444	1.607199029	1.938501804	9.054854282
2013	1.890781782	1.752921982	1.997658587	1.567353172	1.959135803	9.167851326
2014	1.837148197	1.870730597	1.980537775	1.679019377	1.968711651	9.336147597
2015	1.815008255	1.936164279	1.980537775	1.767919581	1.973084706	9.472714596
2016	1.861121136	1.993197924	1.949052446	1.960397757	1.999962296	9.763731558
2017	1.859338193	1.999448648	1.940640009	1.985947506	1.45370818	9.239082536
2018	1.875302663	1.901469225	1.948991893	1.999755614	1.473103169	9.198622564

The sum of each component is represented in the following graph.



The tendency shows a clear improvement in the general component score.



To verify statistically if there is statistical evidence of improvement in the score of Fintech socioeconomic enablers, there will be compared the first 5 years general score (2009-2013) against the last five years considered in the sample (2014-2018).

This time **Student's t-test** will be performed to test whether the means the means two independent samples are significantly different. To perform this test, the same assumptions of part I and II have to be covered, and are fulfilled.

The hypothesis interpretation is as follows:

- **H0 or null hypothesis**: the means of the samples are equal.
- **H1 or alternative hypothesis**: the means of the samples are unequal.

Result

Hypothesis 7: Determinants for Fintech expansion are increasingly favorable in recent years.
p value**:** `2.772827354775973e-05`
ACCEPT ALTERNATIVE HYPOTHESIS

Part III.
Fintech B2B-B2C market trend analysis

Dataset
"Mastersheet_part3.csv": shape 15 x 77

	Date	Quarter	B2B	B2C	Mean CB rank	CB rank percentile	Mean CB rank_B2B	Mean CB rank_B2C	Sum of both	First Half	Second half	Difference
0	2000-03-31	2000 Q1	3	6	390,421	0.333	158	239	9	1	0	-3
1	2000-06-30	2000 Q2	1	1	567675	0.250	567	567	2	1	0	0
2	2000-09-30	2000 Q3	0	0	316495.9863	0.833	312	289	0	1	0	0
3	2000-12-31	2000 Q4	0	0	316495.9863	0.000	312	289	0	1	0	0
4	2001-03-31	2001 Q1	1	2	552,223	0.286	440	422	3	1	0	-1

Source
Crunchbase

Variables
Date
Quarter
B2C
B2C
Mean CB rank
Mean CB rank_B2B
Mean CB rank_B2C
* other columns displayed in the table are operations from the variables listed above.

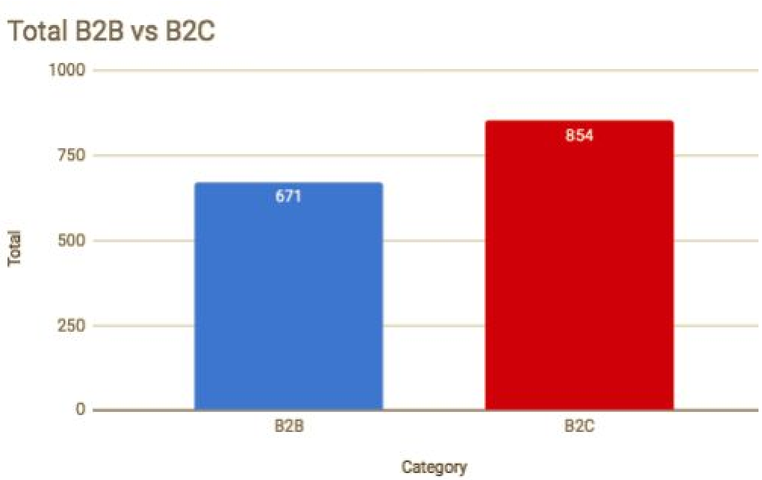
Time period
2000 - 2018

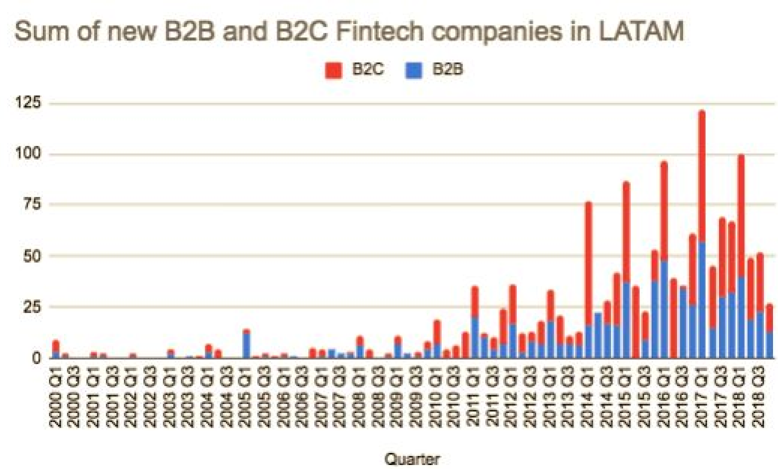
* Countries evaluated are the same as in part II (all South American countries and Mexico).

Completeness
Information retrieved was complete

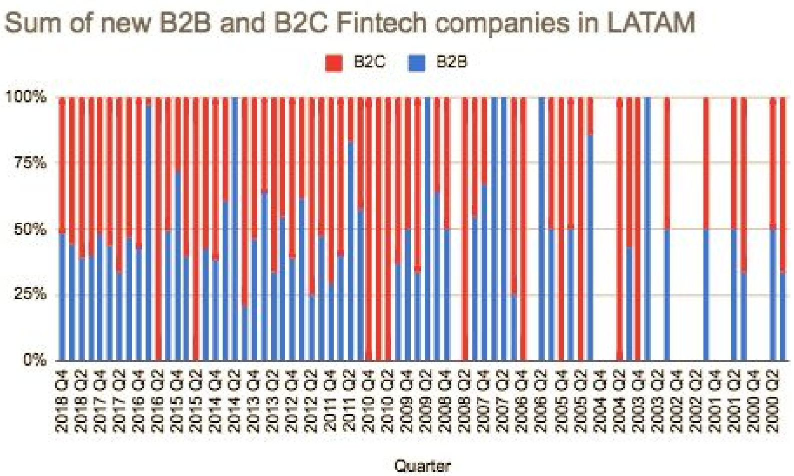
Exploratory Analysis part III

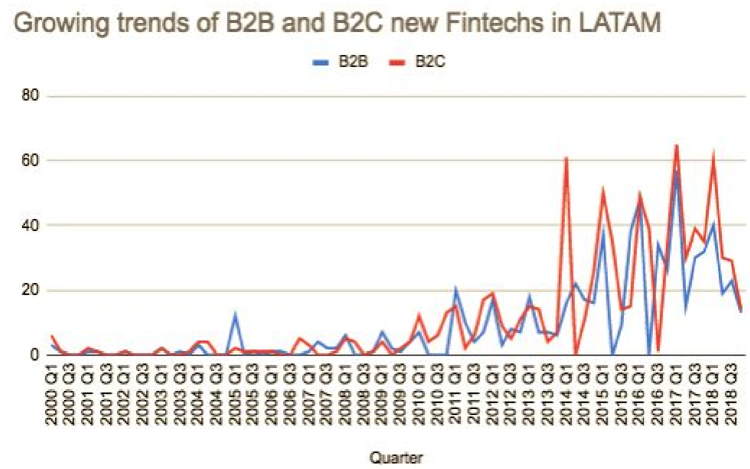
This section will explore valuable market insights and identify any trends regarding B2B and B2C Fintech companies in Latin America. The left graph demonstrates that there are more B2C new established companies in the period studied (2000-2018) in Latin America and Mexico. The right sided visual illustrates a growing tendency of Fintech founded companies reaching more than 100 new companies in the first quarter of 2017.



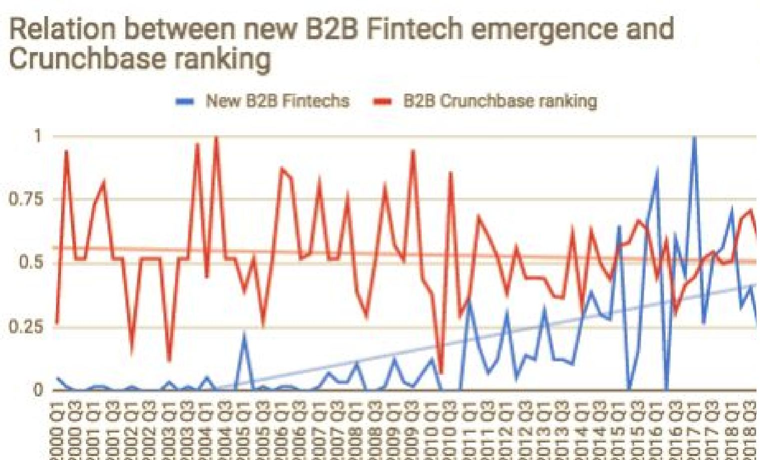


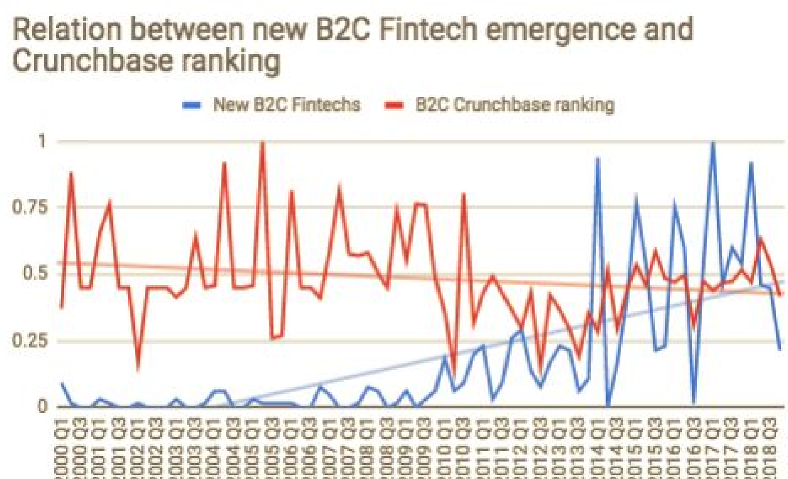
The first graph from left to right illustrates the share of B2B and B2C new companies in Latin America. In recent years the percentage has become similar for both categories, while in previous quarters the relation among both seem unstable because the sample was not numerous. The second graph illustrates a linear tendency between the number of B2B new companies and B2B new companies in the region studied.



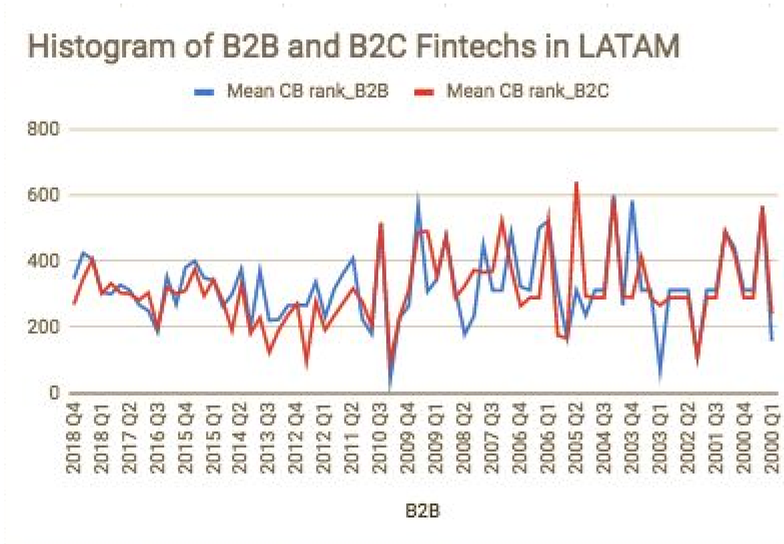


Left and right graphs show the relation between Fintech new companies and Crunchbase ranking of B2B and B2C respectively. For both cases, there is a notable and sustained increase in the number Fintechs. However, the quality of these companies, which is covered in the Crunchbase ranking, demonstrates that on average these companies don't have a better ranking.





The next visual demonstrates an arbitrary relation between B2B and B2C Fintech company emergence and time. Nevertheless, there seems to be a linear dependence of B2B and B2C ranking, this can be explained partially due to the fact that some companies are considered to be both B2B and B2C and thus have the same ranking.



Time Series Analysis for new Fintech companies in Latin America

The objective of this section is to execute a forecast of the development of the Fintech market in Latin American. The following aspect of the Fintech market for the sample studied will the predicted:

- How will the total number of new Fintech companies evolve in the next 3 years.

For this section Univariate Time Series Forecasting will be applied. The ARIMA algorithm stands for 'AutoRegressive Integrated Moving Average', will be used[8].

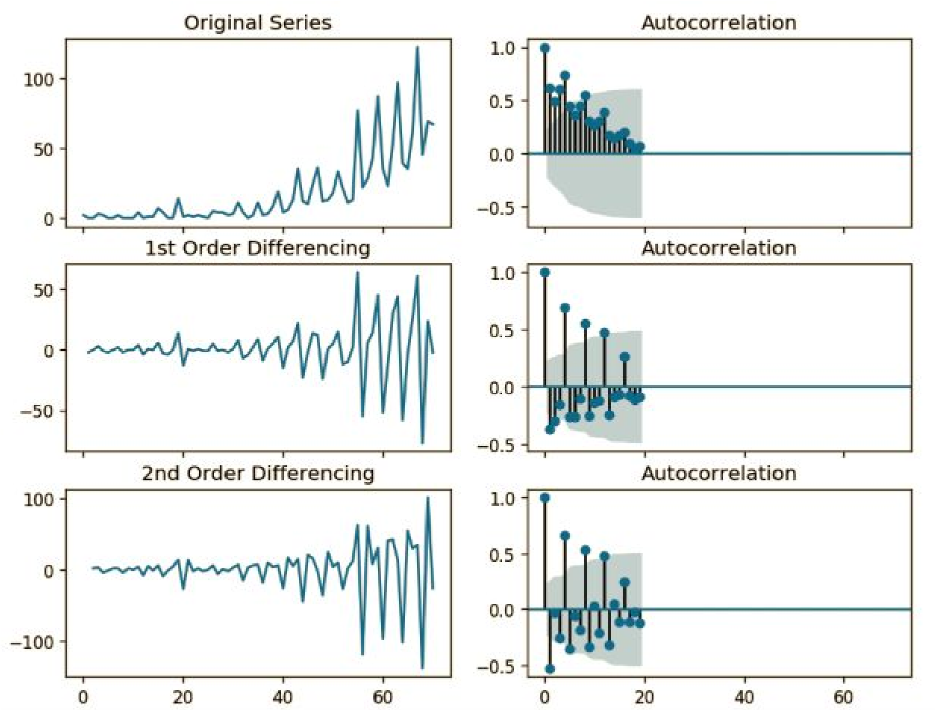
An ARIMA model is has 3 main parameters: p represents the order of the AutoRegressive term, q refers to the order of the Moving Average parameter, while d is the number of differencing to convert the time series into stationary.

Testing is series is stationary.

[8] This algorithm is based on the idea that the past values of the time series can solely be used to predict the future values.

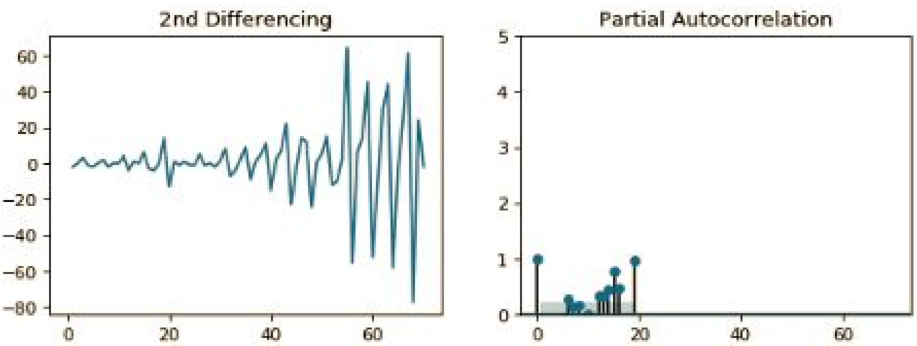
```
ADF Statistic: 4.311433
p-value: 1.000000
```

Since the p-value of the ADF test does not fulfill the significance level (0.05) then the time series is not considered stationary. In this case, there must be found the order of differencing.



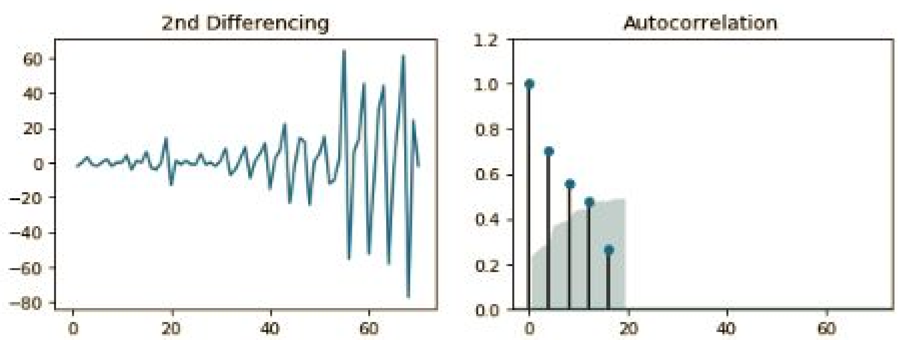
The time series has stationarity with both orders of differencing. This time the second order differencing (d) will be selected.

The following graph shows the *second order* differencing order series and the partial autocorrelation.



In this case the P term selected will be *1* for the AR (autoregressive term).

Next is illustrated the second differencing and autocorrelation of the second differencing.



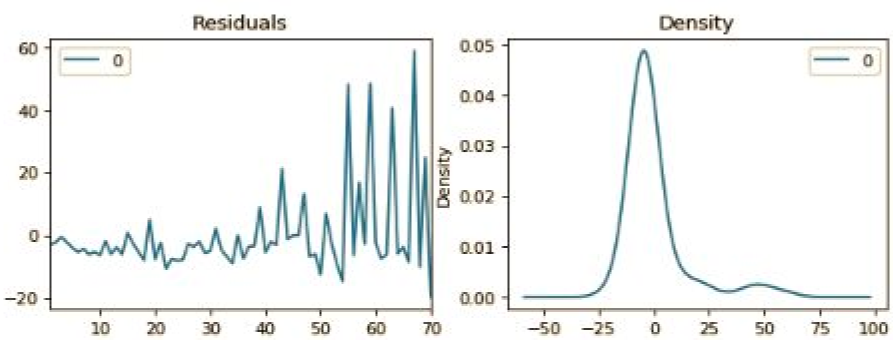
The q term selected for this model is 1.

The Arima selected model takes d value as 2,p as 1 and q as 1. That is portrayed in the model architecture below.

```
                              ARIMA Model Results
==============================================================================
Dep. Variable:                D.value   No. Observations:                   70
Model:                 ARIMA(2, 1, 1)   Log Likelihood                -285.184
Method:                       css-mle   S.D. of innovations             14.040
Date:                Sat, 07 Sep 2019   AIC                            580.369
Time:                        14:11:18   BIC                            591.611
Sample:                             1   HQIC                           584.834

=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             1.0086      0.337      2.989      0.004       0.347       1.670
ar.L1.D.value    -0.2803      0.111     -2.536      0.014      -0.497      -0.064
ar.L2.D.value    -0.5035      0.105     -4.805      0.000      -0.709      -0.298
ma.L1.D.value    -0.6550      0.075     -8.749      0.000      -0.802      -0.508
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1           -0.2783           -1.3815j            1.4093           -0.2816
AR.2           -0.2783           +1.3815j            1.4093            0.2816
MA.1            1.5266           +0.0000j            1.5266            0.0000
-----------------------------------------------------------------------------
```

To ensure there are no particular trends for residuals to ensure a normality by showing constant mean and variance plot will be executed as follows:



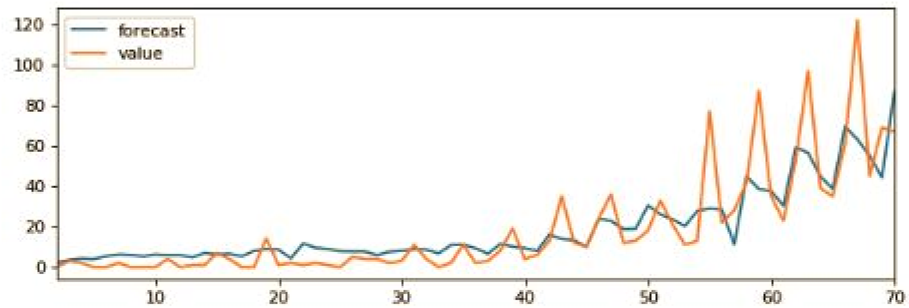
Residual errors seem behave with zero mean and uniform variance therefore, no transformation need to be done.

Model training and test split

The fist 85% data points will be used to dest the model, and the rest 15% will be used to test the model as follows:

train = df.value[:60]
test = df.value[60:70]

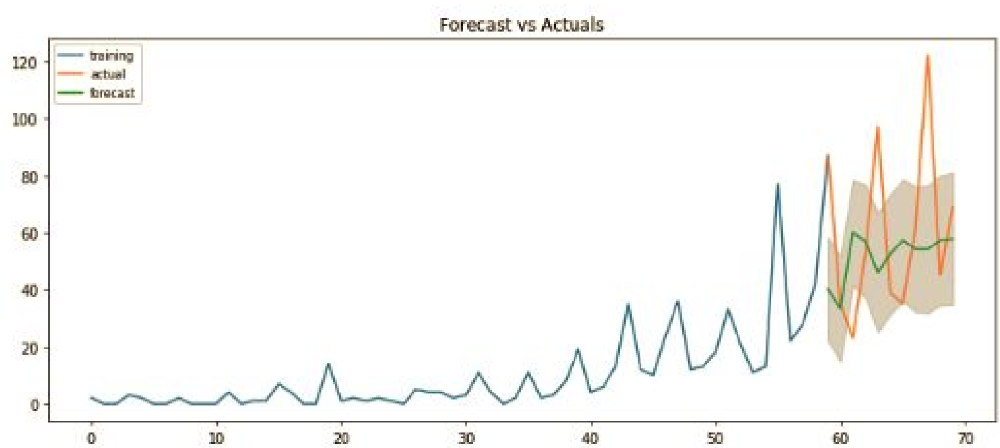
The following graph shows in the x axis uniform periods of time starting with the first quarter of the year 2000 and finishing with last quarter of 2018. With the selected model architecture the model forecasts behave in the following fashion:



As the graph above shows, the forecast follows a clear increasing tendency of new Fintech startups as the years pass. However sudden increases in recent years where notable local peaks appear do not necessarily are followed accurately by the model forecasts.

Validation model

This is a graphic representation of the validation model, which forecasts the results portrayed by the green line. As seen below, the sudden increase in the last years of Fintech was not accurately predicted by our model. That give evidence of a market boom of Fintech in Latin America in recent years.



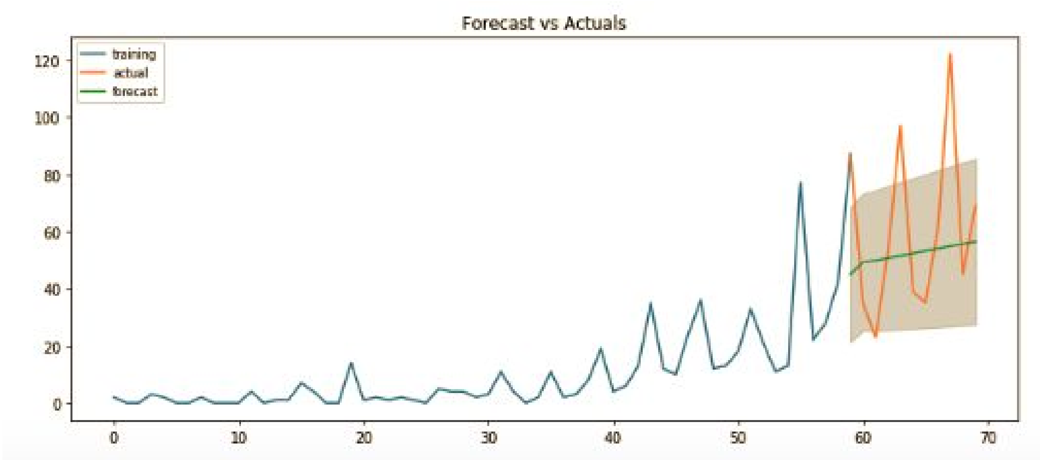
Modifying parameters

By testing in the programming framework and modifying parameter d=1 gave better results and reduced the AIC curve as seen in the next visual:

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                D.value   No. Observations:                   59
Model:                 ARIMA(1, 1, 1)   Log Likelihood                -231.169
Method:                       css-mle   S.D. of innovations             12.074
Date:                Fri, 06 Sep 2019   AIC                            470.339
Time:                        22:32:26   BIC                            478.649
Sample:                             1   HQIC                           473.583

=================================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
---------------------------------------------------------------------------------
const             0.8280      0.387      2.138      0.037       0.069       1.587
ar.L1.D.value    -0.0783      0.172     -0.456      0.650      -0.414       0.258
ma.L1.D.value    -0.7541      0.097     -7.739      0.000      -0.945      -0.563
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1          -12.7775           +0.0000j           12.7775            0.5000
MA.1            1.3260           +0.0000j            1.3260            0.0000
-----------------------------------------------------------------------------
```

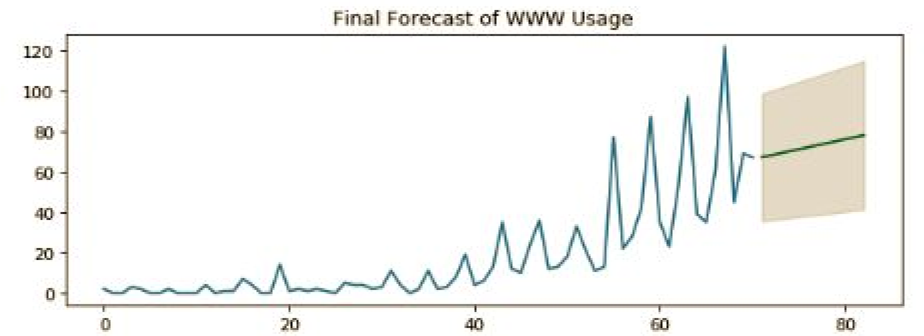
Validation model



This model performs better with first order differencing. The parameter selection of models is as follows: d = 1, p = 1, d = 1

Forecast

As presented in the beginning of this section, the aim of this segment is to predict how the total number of new Fintech companies evolve in the next 3 years. The model created visually presents the following trend, where the green line represents the predicted number of new Fintech companies for the next three years with a 95% interval of confidence covered by the gray area.



The array displayed is the following:

([67.13875436, 68.13077966, 69.12280496, 70.11483026, 71.10685556,
72.09888086, 73.09090617, 74.08293147, 75.07495677, 76.06698207,
77.05900737, 78.05103267])

This translates to the following time periods:

2019 Q1: 67.13875436
2019 Q2: 68.13077966
2019 Q3: 69.12280496
2019 Q4: 70.11483026
2020 Q1: 71.10685556
2020 Q2: 72.09888086
2020 Q3: 73.09090617
2020 Q4: 74.08293147
2021 Q1: 75.07495677
2021 Q2: 76.06698207
2021 Q3: 77.05900737
2021 Q4: 78.05103267

Mean of new Fintech companies forecasted in next three years: 72.59
Mean of new Fintech companies contained in the sample (2000-2018): 18.14

There is a **544.5% increase** in the mean of companies expected for the next three years 2019 - 2021 compared to the year period 2000-2018.

Discussion of results

The accuracy in terms of 'RMSE' is 30.290729784714063 which equated to 25.25% of the maximum value of the sample. The 'RMSE' of the model selected is 65% larger than the mean of the sample which is 18.14. This is evidence that the sector is impressively volatile and is experiencing a boom in recent years which represents high levels of complexity in order to model and forecast.

Conclusions and Future Work

Conclusions

When tested statistically, the same alternative hypothesis were accepted in the International case (part I) and the Latin American case (part II). However, the level of significance varied for both regions, this may have many academic explanations that exceed the scope of this work. The components of Fintech emergence proved to be statistically significant displayed increasingly favorable results when compared to previous time periods, this translates to an increasingly favorable environment for Fintech company establishments in recent years.

When analyzing market trends for Fintech in Latin America, B2B and B2C companies, have consistently increased in number and the tendency shows a consistent increase of new companies in the industry. Time series analysis results forecasting give evidence that the sector is impressively volatile and is experiencing a major boom today.

Bibliography[9]

1. “15 Statistical Hypothesis Tests in Python (Cheat Sheet).” *Machine Learning Mastery*, 8 Aug. 2019, machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/.

2. Alois, JD. “3 Biggest Fintech Trends Shaping Latin America.” Crowdfund Insider, 9 June 2019, www.crowdfundinsider.com/2019/06/148254-3-biggest-fintech-trends-shaping-latin-america/.

3. “ARIMA Model - Complete Guide to Time Series Forecasting in Python: ML .” *Machine Learning Plus*, 28 Apr. 2019, www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/.

4. “The Berkus Method – Valuing the Early Stage Investment.” *BERKONOMICS*, berkonomics.com/?p=131.

5. Fontinelle, Amy. “What Exactly Is a Startup?” *Investopedia*, Investopedia, 14 July 2019, www.investopedia.com/ask/answers/12/what-is-a-startup.asp.

6. Haddad, et al. “The Emergence of the Global Fintech Market: Economic and Technological Determinants.” *SSRN*, 27 Aug. 2016, poseidon01.ssrn.com/delivery.php?ID=654125087110031107002006089026029076025033067063006028024025076018103083121105108006025011001016026124055019092071100105117091117023037077029103070025102118112079060075054120121069003000121016027095075086125110015026079029089096077064065093126123021&EXT=pdf.

7. “IPOs.” *Investopedia*, Investopedia, www.investopedia.com/ipos-4689824.

8. Jain, Nirvikar. “Fintech Valuation Methods for Money Transfer Disruptors.” *Toptal Finance Blog*, Toptal, 19 Mar. 2019, www.toptal.com/finance/valuation/how-to-value-a-company.

9. Jaksic, et al. “The Future of Banking: The Role of Information Technology.” *SSRN*, 16 Oct. 2015, poseidon01.ssrn.com/delivery.php?ID=616064002106075071121016115023112121037016025093044007014073016026002103074070123120028062030124045033010026076087080099124114046083078061083102071031066084124001000030016120077066066114106115070116101092122099094098121101110122125067075089096007125

[9] Style: MLA 8th Edition (assisted with EasyBib application from Google)

&EXT=pdf.

10. Mearian, Lucas. “What Is FinTech (and How Has It Evolved)?” *Computerworld*, Computerworld, 18 Sept. 2017, www.computerworld.com/article/3225515/what-is-fintech-and-how-has-it-evolved.html.

11. Nordqvist, Christian. “What Is a Unicorn Company? Definition and Examples.” *Market Business News*, Market Business News, 19 Feb. 2019, marketbusinessnews.com/financial-glossary/unicorn/.

12. “Perspectives.” *Greylock Perspectives*, news.greylock.com/perspectives/home.

13. Reiff, Nathan. “Series A, B, C Funding: How It Works.” *Investopedia*, Investopedia, 25 June 2019, www.investopedia.com/articles/personal-finance/102015/series-b-c-funding-what-it-all-means-and-how-it-works.asp.

14. Robehmed, Natalie. “What Is A Startup?” *Forbes*, Forbes Magazine, 15 May 2015, www.forbes.com/sites/natalierobehmed/2013/12/16/what-is-a-startup/#2215478f4044.

15. “Unicorns.” *Investopedia*, Investopedia, www.investopedia.com/unicorns-4689822.