IbovRange Project

HarvardX: PH125.9x Data Science: Capstone

Rebello, Carlos E. Piffer

December, 2020

Contents

1 Executive Summary

This project deals with a behavioral assessment of Ibovespa, that is the São Paulo Stock Exchange Index. It is a theoretical stock portfolio that contains the assets that drive the highest trading volumes, something around 80% of daily total and is considered the benchmark for variable income.

To this end, several internal and global indicators were used, such as interest, foreign exchange, commodities, other markets and more.

The objective here is to train a machine learning algorithm that makes predictions of the Ibovespa daily high and low. The goal is an accuracy greater than 80% of the values with a 95% reliability and, for future, to develop strategies to operate in this market.

1.1 Data Set

The ibov dataset was compiled by me according to the needs of the project. The information used is available on the website of B3, the company that manages the stock exchange, on the investing.com platform and on BCB website (Central Bank of Brazil) using IBGE (Brazilian Institute of Geography and Statistics), FGV (Getúlio Vargars Foundation), CNI (Nacional Confederation of Industry) and Fecomercio (São Paulo State Commerce Federation) as a source. At the end of report, links will be available.

The first step was to observe the index 's companies and understand their weight. Below we can see that, with a total of 77 companies, the top 20 represent 68.72% of total.

code	company	part. 100k	accum
VALE3	VALE	12460	0.12460
ITUB4	ITAUUNIBANCO	6799	0.19259
PETR4	PETROBRAS	5746	0.25005
BBDC4	BRADESCO	5182	0.30187
B3SA3	B3	4776	0.34963
PETR3	PETROBRAS	4486	0.39449
MGLU3	MAGAZ LUIZA	3193	0.42642
ABEV3	AMBEV S/A	3078	0.45720
WEGE3	WEG	2564	0.48284
ITSA4	ITAUSA	2422	0.50706
BBAS3	BRASIL	2192	0.52898
NTCO3	GRUPO NATURA	2134	0.55032
GNDI3	INTERMEDICA	1993	0.57025
SUZB3	SUZANO S.A.	1987	0.59012
RENT3	LOCALIZA	1950	0.60962
LREN3	LOJAS RENNER	1797	0.62759
JBSS3	JBS	1786	0.64545
VVAR3	VIAVAREJO	1455	0.66000
BBDC3	BRADESCO	1362	0.67362
RADL3	RAIADROGASIL	1358	0.68720

Analyzing these 20 companies closely, 4 are banks, a holding company with a large participation in the banking sector, in addition to B3 itself. We have 8 exporting companies, or with activities in other countries and, 6 retail companies or dependent on the domestic market.

Based on these characteristics, the following were chosen as predictors:

- DI - interbank deposit (interest rate resulting from negotiations between banks).

- SLEIC - basic interest rate determined by the BCB.
- Exchange rate between USD / BRL and EUR / BRL
- Economic performance indicators:
  - IECI - Industrial Entrepreneur Confidence Index,
  - CCI - Consumer Confidence Index,
  - IPCA & IGPM - price inflation indices. Columns were created with accumulated indexes.
  - Industry Activity Index and
  - Unemployment rate.

As Ibovespa's largest company, with a 12.46% stake, is a big iron ore exporter, TIOc1 values (future iron market) is in dataset. In addition, 2 of the top 6 shares represent Petrobras, an oil company, so WTI values (future oil market) was also collected to try to explain the São Paulo stock index behavior.

Investors are globalized and B3 receives investments from all over the world. I understood that it is important to compare the Ibovespa movement with the main stock exchanges in the world. those used in the study were:

- Nasdaq - US
- S&P 500 - US
- Dow-Jones - US
- DAX - Germany
- Euronext - Netherlands / France
- LSE - UK
- Nikkei - Japan
- SSE - China

Investors control their risk and always look for a safe place for money. Thus, FED bonds are coveted and any change in their rate affects the world market. Is therefore in the dataset, besides, of course, Ibovespa historical series with open, close, high, low, volume traded and variation in percentage.

We also have the operations dates that are spread over:

- Year,
- Month,
- Month day,
- Week day and
- Quarter

Let's see how the data is arranged.

date	close	open	high	low	volM	var	year	month
2019-04-02	95386.76	96062.1	96690.17	94824.93	3.83	-0.7	2019	4

month_day	week_day	quarter	di	selic	fed_rate	dax	euronext	dj
2	3	2	6.4	6.4	2.41	11703.24	1054.02	26235

lse	nasdaq	sp500	nikkei	sse	usd	eur	wti	iron
4793	7473.51	2868.5	21505.31	3176.82	3.8523	4.3189	61.81	89.5

ieci	cci	ipca	industry_activity	unemployment	igpm	igpm_acc	ipca_acc
62.6	136.66	0.57	83.4	12.5	0.92	151.2221	63.25597

It contains 35 variables, high and low, that are the target of our predictions and others 33 predictors. Each line represents a market day. We have 5,979 lines in total, starting on 2000-12-27 and ending on 2020-12-01.

The dataset contains several NAs due to the dates when one market is open and the other is not, for more recent data that have not yet been made available for some predictors and indicators or markets that haven't existed for so long.

Looking at a summary of the data, note that there are actually many NAs.

	date	close	open	high	low	volM
	Min. :2000-12-27	Min. : 8371	Min. : 8397	Min. : 8513	Min. : 8225	Min. : 1.00
	1st Qu.:2007-01-04	1st Qu.: 38817	1st Qu.: 38805	1st Qu.: 39384	1st Qu.: 38303	1st Qu.: 3.13
	Median :2013-01-22	Median : 57716	Median : 57704	Median : 58293	Median : 57258	Median : 4.59
	Mean :2012-05-20	Mean : 59220	Mean : 59188	Mean : 59774	Mean : 58613	Mean : 63.29
	3rd Qu.:2018-12-26	3rd Qu.: 85685	3rd Qu.: 85673	3rd Qu.: 86397	3rd Qu.: 84924	3rd Qu.: 87.45
	Max. :2020-12-01	Max. :119528	Max. :119528	Max. :119593	Max. :118108	Max. :994.96
	NA	NA	NA	NA	NA	NA's :98

	var	year	month	month_day	week_day	quarter
	Min. :-14.78000	Min. :2000	Min. : 1.000	Min. : 1.00	Min. :2.000	Min. :1.000
	1st Qu.: -0.84000	1st Qu.:2007	1st Qu.: 3.000	1st Qu.: 8.00	1st Qu.:3.000	1st Qu.:1.000
	Median : 0.08000	Median :2013	Median : 6.000	Median :16.00	Median :4.000	Median :2.000
	Mean : 0.06677	Mean :2012	Mean : 6.074	Mean :15.82	Mean :3.997	Mean :2.371
	3rd Qu.: 1.02000	3rd Qu.:2018	3rd Qu.: 9.000	3rd Qu.:23.00	3rd Qu.:5.000	3rd Qu.:3.000
	Max. : 14.66000	Max. :2020	Max. :12.000	Max. :31.00	Max. :6.000	Max. :4.000
	NA	NA	NA	NA	NA	NA

	di	selic	fed_rate	dax	euronext	dj
	Min. : 1.90	Min. : 1.90	Min. :0.040	Min. : 2204	Min. : 427.6	Min. : 6523
	1st Qu.: 6.40	1st Qu.: 6.40	1st Qu.:0.160	1st Qu.: 5664	1st Qu.: 681.9	1st Qu.:10882
	Median :10.82	Median :10.90	Median :1.420	Median : 7843	Median : 860.5	Median :15402
	Mean :11.32	Mean :11.37	Mean :1.641	Mean : 8361	Mean : 841.4	Mean :17047
	3rd Qu.:14.13	3rd Qu.:14.15	3rd Qu.:2.400	3rd Qu.:11578	3rd Qu.:1008.1	3rd Qu.:24917
	Max. :26.32	Max. :26.35	Max. :6.670	Max. :13774	Max. :1180.5	Max. :30029
	NA's :28	NA's :28	NA's :50	NA's :87	NA's :65	NA's :381

	lse	nasdaq	sp500	nikkei	sse	usd
	Min. : 7.8	Min. : 800.9	Min. : 674.8	Min. : 7055	Min. :1012	Min. :1.539
	1st Qu.:1293.3	1st Qu.: 1686.9	1st Qu.:1195.6	1st Qu.:10696	1st Qu.:2029	1st Qu.:2.117
	Median :3251.0	Median : 2771.9	Median :1550.8	Median :15619	Median :2729	Median :2.857
	Mean :3430.8	Mean : 4017.7	Mean :1863.2	Mean :15597	Mean :2583	Mean :2.900
	3rd Qu.:4840.0	3rd Qu.: 6840.9	3rd Qu.:2702.0	3rd Qu.:20776	3rd Qu.:3057	3rd Qu.:3.748
	Max. :9230.0	Max. :12417.5	Max. :3643.8	Max. :26788	Max. :6092	Max. :5.992
	NA's :2238	NA's :181	NA's :100	NA's :384	NA's :410	NA's :23

	eur	wti	iron	ieci	cci	ipca
	Min. :1.788	Min. : 12.96	Min. : 38.54	Min. :34.70	Min. : 86.86	Min. :-0.3800
	1st Qu.:2.623	1st Qu.: 46.33	1st Qu.: 73.98	1st Qu.:56.40	1st Qu.:122.89	1st Qu.: 0.2000
	Median :3.324	Median : 58.30	Median : 92.53	Median :62.00	Median :133.69	Median : 0.4000
	Mean :3.416	Mean : 61.55	Mean : 97.53	Mean :59.62	Mean :134.41	Mean : 0.4077
	3rd Qu.:4.232	3rd Qu.: 75.12	3rd Qu.:120.25	3rd Qu.:64.10	3rd Qu.:147.63	3rd Qu.: 0.5700
	Max. :6.750	Max. :145.19	Max. :188.90	Max. :70.70	Max. :171.70	Max. : 1.3200
	NA's :23	NA's :112	NA's :2562	NA's :1308	NA's :84	NA's :1347

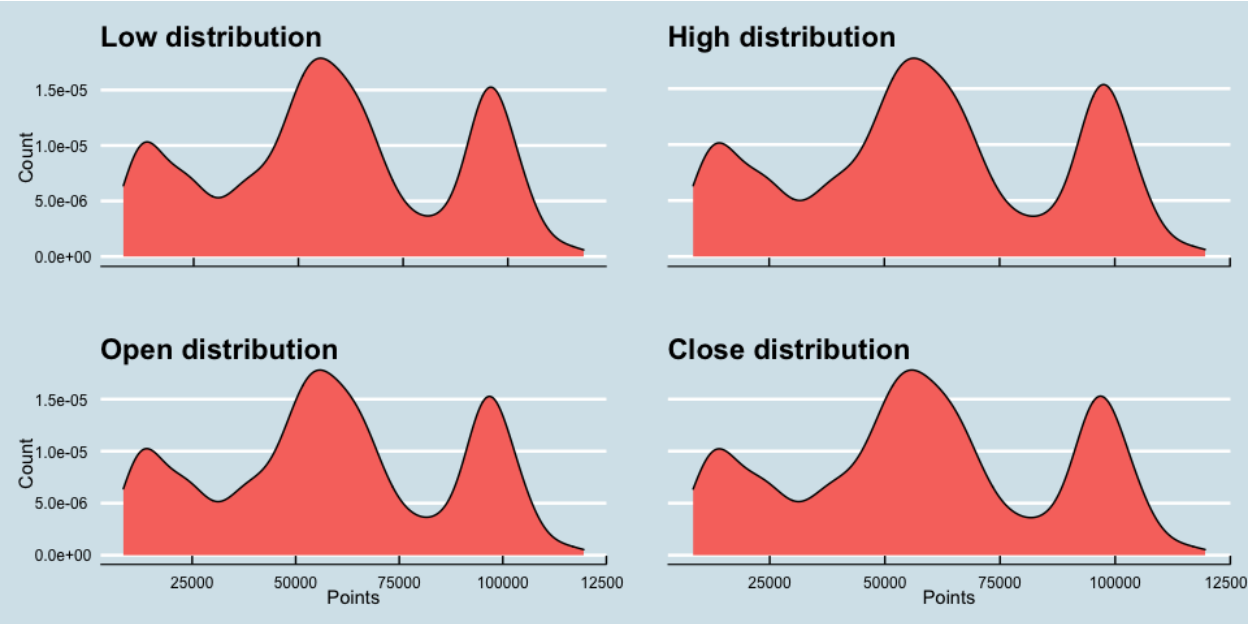
	industry_activity	unemployment	igpm	igpm_acc	ipca_acc
	Min. : 60.40	Min. : 4.30	Min. :-1.1000	Min. : 0.02094	Min. :-0.0738
	1st Qu.: 83.70	1st Qu.: 7.40	1st Qu.: 0.2000	1st Qu.: 48.97326	1st Qu.:15.8557
	Median : 92.20	Median :10.90	Median : 0.5600	Median : 93.55384	Median :43.0337
	Mean : 92.22	Mean : 9.85	Mean : 0.6393	Mean : 93.04126	Mean :38.7606
	3rd Qu.: 99.20	3rd Qu.:12.20	3rd Qu.: 0.9200	3rd Qu.:144.07138	3rd Qu.:61.5569
	Max. :112.60	Max. :14.60	Max. : 5.1900	Max. :190.48191	Max. :69.8670
	NA's :1410	NA's :698	NA's :42	NA's :42	NA's :1347

To develop the algorithm and define the parameters that maximize our results, cross validation was used, so ibov was divided into train_set, with 90%, and test_set, with 10% of observations and, train set was divided into train_set2, with 90%, and test_set2, with 10% of observations.

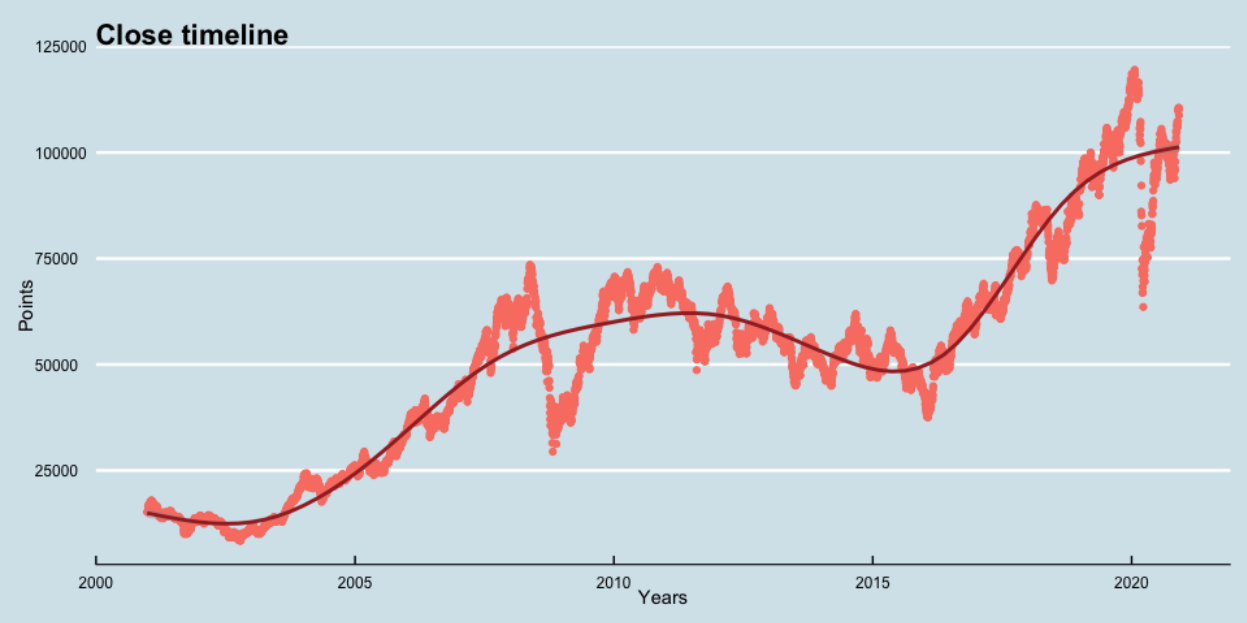
The data were analyzed from seasonality and between Ibovespa and predictors relationship point of view. Some of them were selected for predictions where logistic and local weighted regressions were used, in addition to k-nearest neighbors and random forest. The last two had a better result compared to the first.

2 Methods and Analysis

The first step was to observe the distribution of Ibovespa points and, as expected, open, close, high and low have the same density curve.



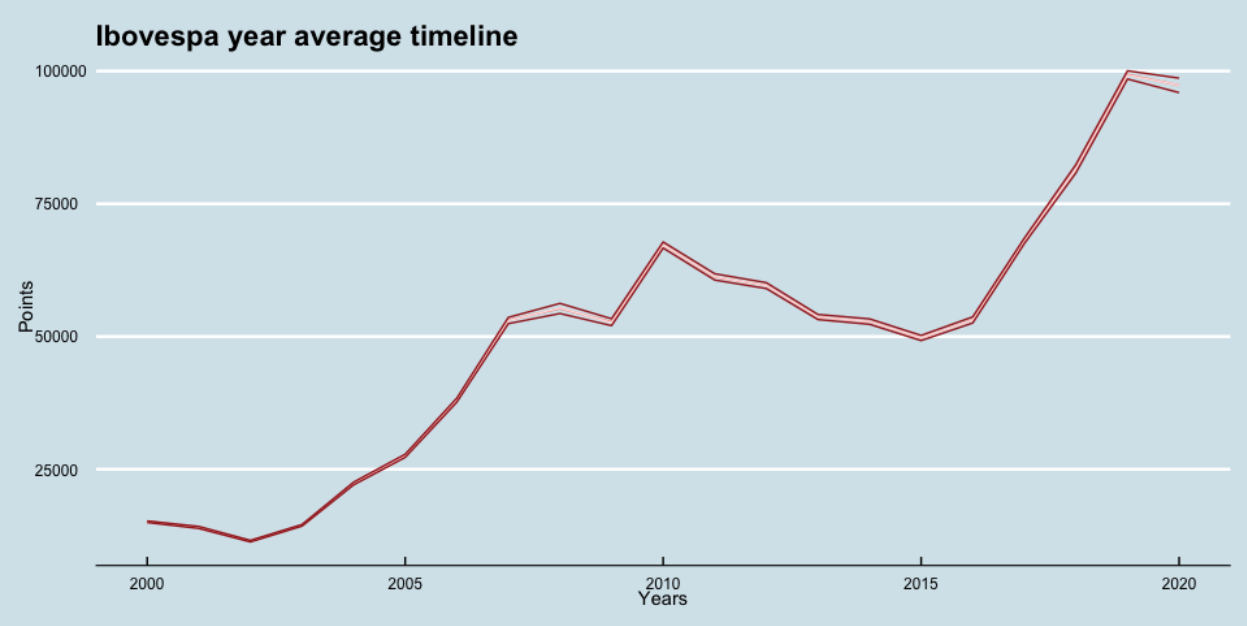
The first thing catches our attention is we have 3 peaks, as 3 distinct approximately normal distributions. To try to understand this phenomenon we will observe Ibovespa development overtime using closing numbers.



We can clearly see that the index remained at 3 levels: Below 25 thousand points in the 2000s, close to 70 thousand points in the 2010s and in 2020 it reached the level of 100 thousand points. This explains the 3

humps in the data distribution. We can see a major depression in 2008, due to the financial crisis, another around 2016/17, when President Dilma Rousseff's had impeachmeant and, after a good recovery, another depression in 2020 with COVID-19 pandemic.

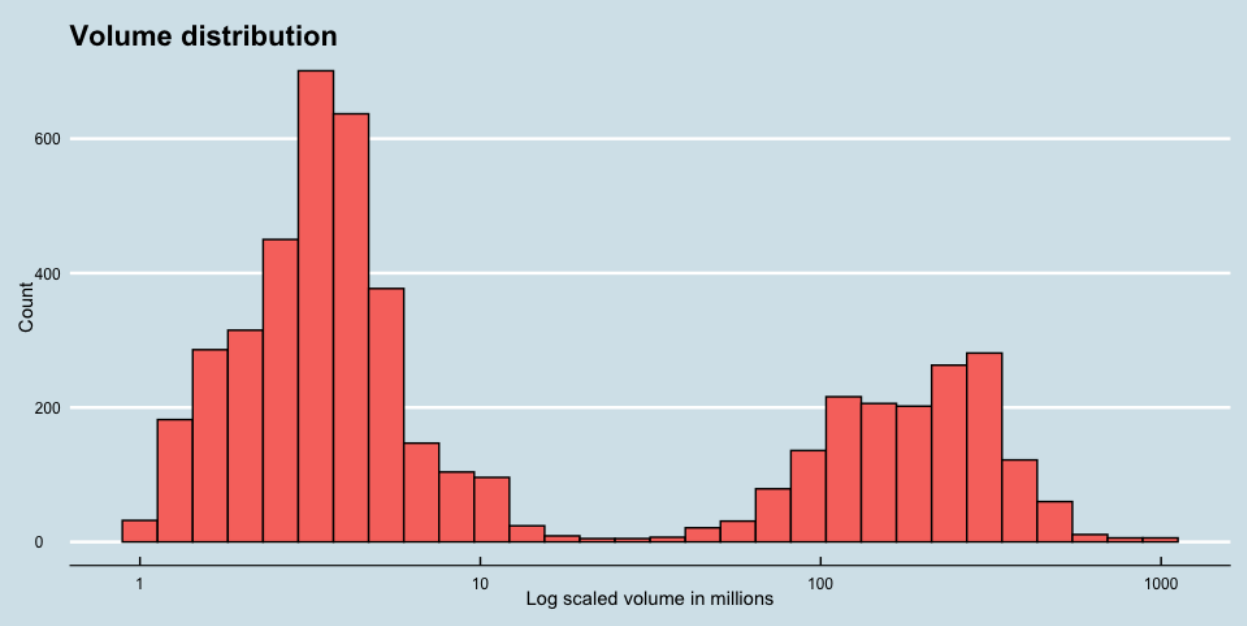
Observing the timeline of the 4 Ibovespa points columns:



We can see the difference between high and low is small in the 2000s and the market suffered greater volatility in years around 2008. It falls and, between 2010 and 2016, rises again. Starts to decrease and now, between 2019 and 2020 it has increased again. We can see when markets are rising, volatility drops and, when they are falling, with crises, the variance increases. Opening and closing is always near and obviously between high and low.

2.1 Volume Effect

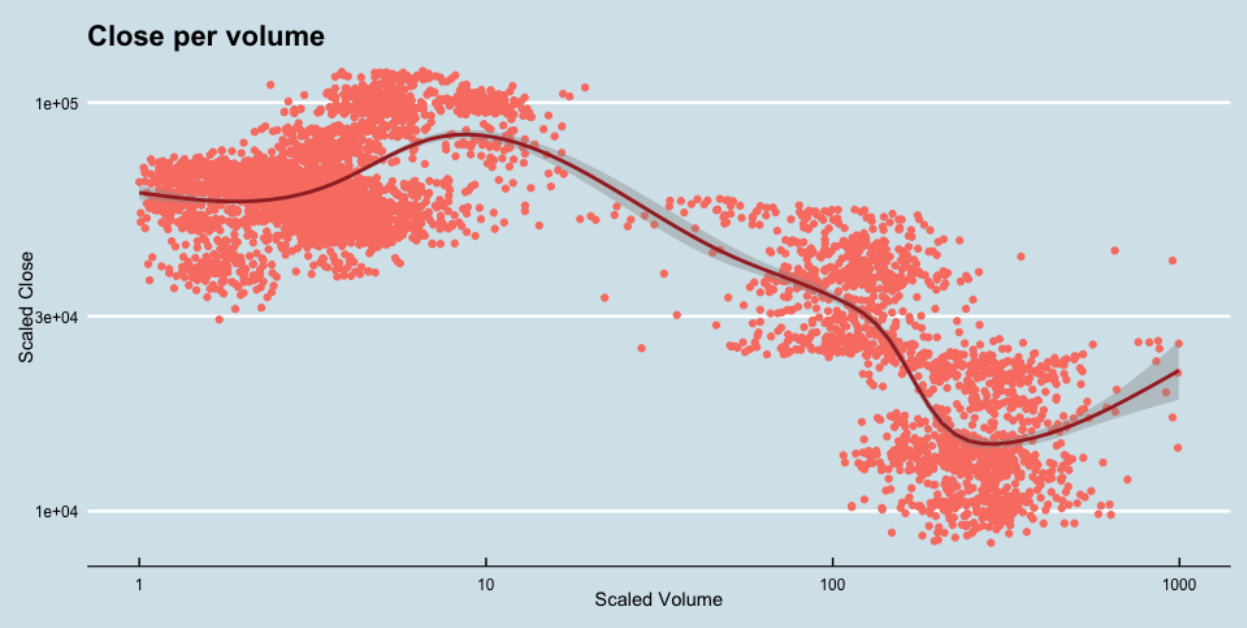
The volume of money can be expected to directly affect the value of assets. But in what way? First, let's see how the data is distributed.



We now have two humps, different from what we saw with the stock index. Let's look at timeline.

We realized that the financial volume of operations was very high and, in mid-2008 it plummeted. On the logarithmic scale chart we notice that the volume shows a reaction, but with normal scale we can see that compared to the pre 2008 phase, the trading volume remains practically dead.

Looking at plot, the relation between volume and index score seems obvious an inverse relation.



The lower the volume, the higher the price index. But does it make any sense? Shouldn't the larger financial volume show a more euphoric market with rising prices? The market may be euphoric to sell as well, so in 2 moments this correlation is reversed. The correlation between index and volume is -0.69, but between 2000 and 2010 is -0.76 and, between 2010 and 2020, 0.39. However, in this case the historical maxims without volume draw attention.

2.2 Seasonality Effect

The financial market has a recurring schedule. There are dividend distribution, contracts that expire every month, often on the same day, others with expirations linked to week day, others are quarterly and this movement can influence prices. Another relevant factor is the routines within the institutions, with goals and targets often weekly. In addition, we have the annual seasonality of the market with its cycles of commodities, taxes and others.

We have seen that over the years the index has varied between a minimum of 8,224.61 and a maximum of 119,593.1 points. 14.5 times the minimum. This makes it very difficult to perceive the effect of seasonality. As this variation is well explained by time, loess was used to minimize the variation in value over the years.

2.2.1 Local weighted regression (loess)

To smooth the line over the years, local weighted regression (loess) was used, which allows us to consider different sizes of data windows using Taylor's theorem, which tells us that if you look closely at any smooth function $f(x)$, it will look like a line. We can consider larger window sizes with the linear assumption than with a constant. For each point $x0$ loess defines a window and fits a line within that window (h).

$$E[Y_i \mid X_i = x_i] = \beta_0 + \beta_1(x_i - x_0) \quad if \quad \mid x_i - x_0 \mid \leq h$$

The value adjusted at $x0$ becomes our estimate $\hat{f}(x0)$. The final result is a smoother fit, as we use larger sample sizes to estimate our local parameters. Loess maintains the same number of points used in the local

adjustment, which is controlled by the span argument that expects the proportion of the total data. Instead of using least squares, we minimize a weighted version:

$$\sum_{i=1}^{N} w_0(x_i)[Y_i - \{\beta_0 + \beta_1(x_i - x_0)\}]^2$$

Instead of the Gaussian kernel, loess uses a function called Tukey's triple weight that reaches the values closer to the maximum.

$$W(u) = \left(1 - \mid u \mid^3 \right)^3 \quad if \quad \mid u \mid \leq 1 \quad and \quad W(u) = 0 \quad if \quad \mid u \mid > 1$$

$$w_0(x_i) = W\left(\frac{x_i - x_0}{h}\right)$$

Taylor's theorem tells us if we don't look so closely at the mathematical function, it looks like a parable. This is the standard function procedure and can be adjusted through the degree parameter, being 1 to adjust polynomials of degree 1 (lines) and 2 for parabolas. To smooth the Ibovespa close scores over the years, degree 1 and span 0.1 were used.

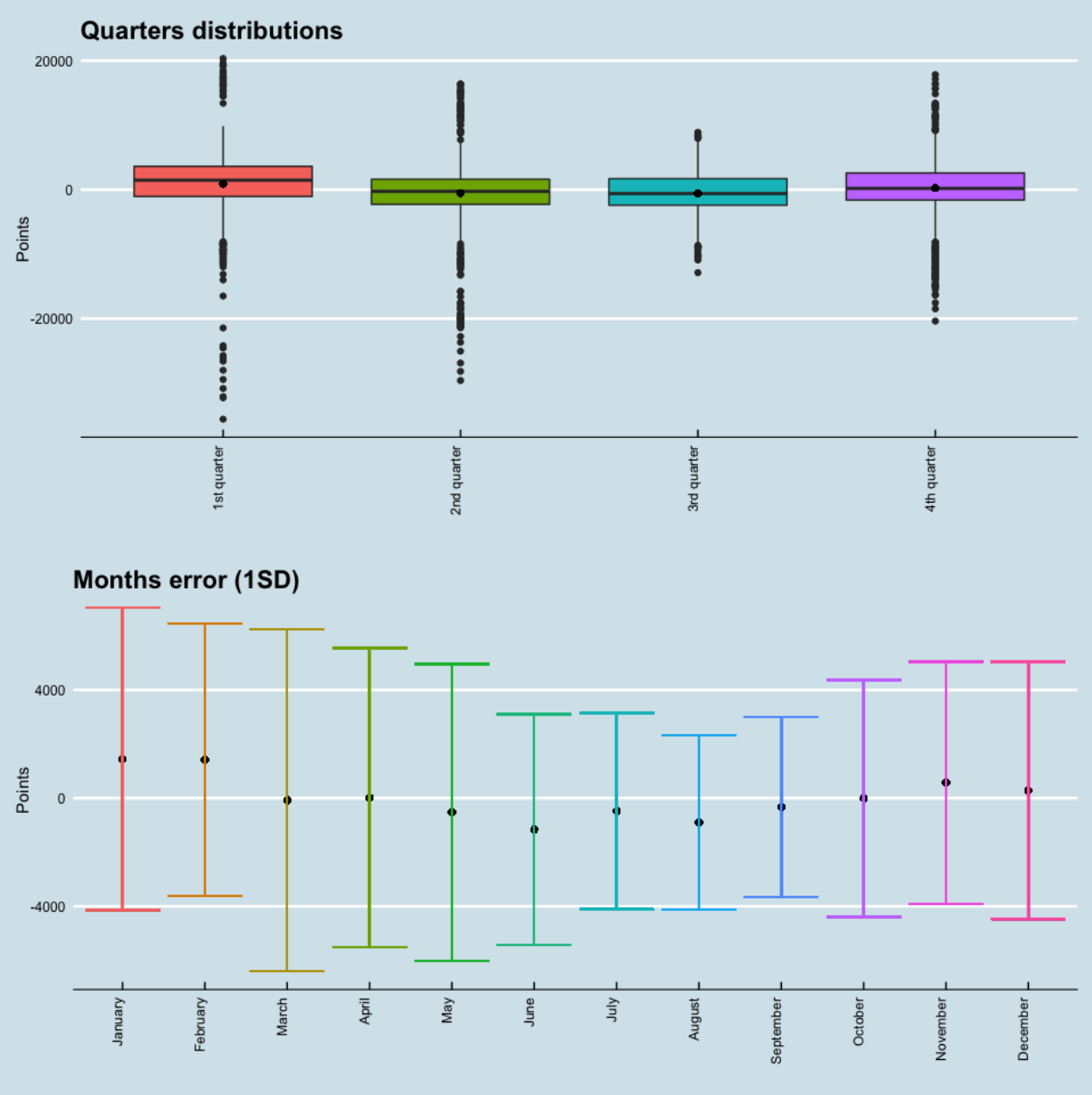
The line generated by loess l was subtracted from the closing value c on the same date d forming Center Close CC with average 0.

$$CC_d = c_d - l_d$$

```
## Generalized Additive Model using LOESS
##
## 5979 samples
##    1 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 5979, 5979, 5979, 5979, 5979, 5979, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4613.342  0.9742634  3143.748
##
## Tuning parameter 'span' was held constant at a value of 0.1
## Tuning
##  parameter 'degree' was held constant at a value of 1
```

2.2.2 Over the year

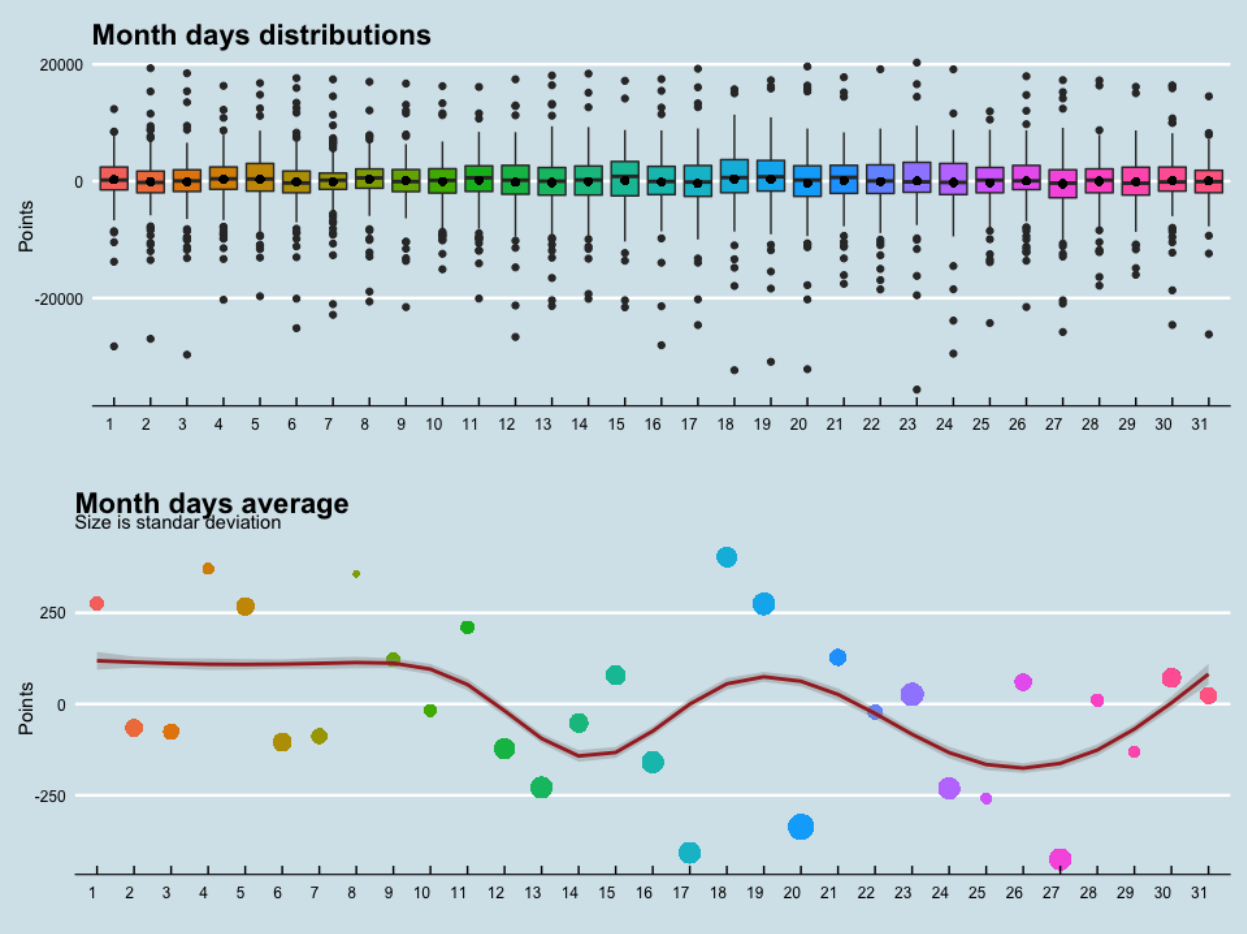
To understand the index behavior, we'll analyze the distribution over the quarters and months.



Looking at the plots, we can see the average and median are very close, in addition to the first and third quartiles, which are not very far apart. Usually between May and August we have a low moment, with the lowest average in June. January and February are the peak months. We can also see a cycle of variance, the smallest in August and continues to increase until its peak in March, when it returns to retract.

2.2.3 Over the month

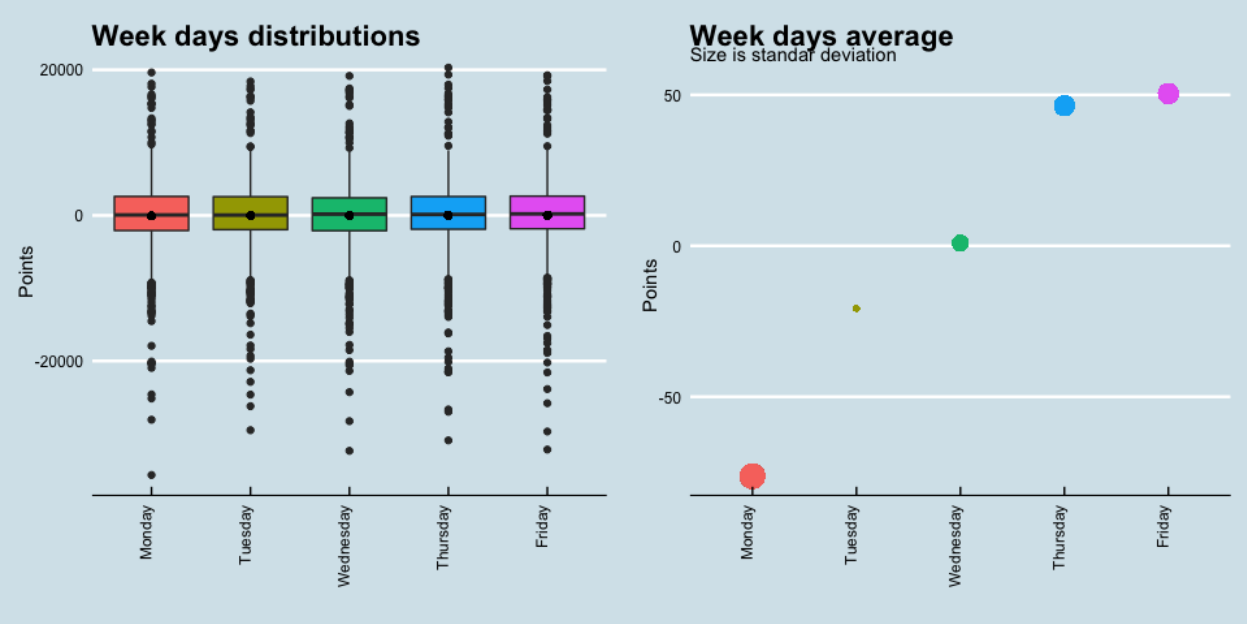
With the intention of finding days that every month behave similarly, let's look at the data.



We can see every day is well distributed, with 50% of the data are relatively close to the average and the median. Looking at the means closely, we notice the first month days tend to be more valued than the last ones, with major depressions around the 14th and 17th. 18 and 4 are the top ones above the average, 403 and 371 points respectively. The worst scores are those of days 27 and 17 with 425 and 407 points below average. The difference between the most positive and most negative days average is 828 points.

2.2.4 Week days effect

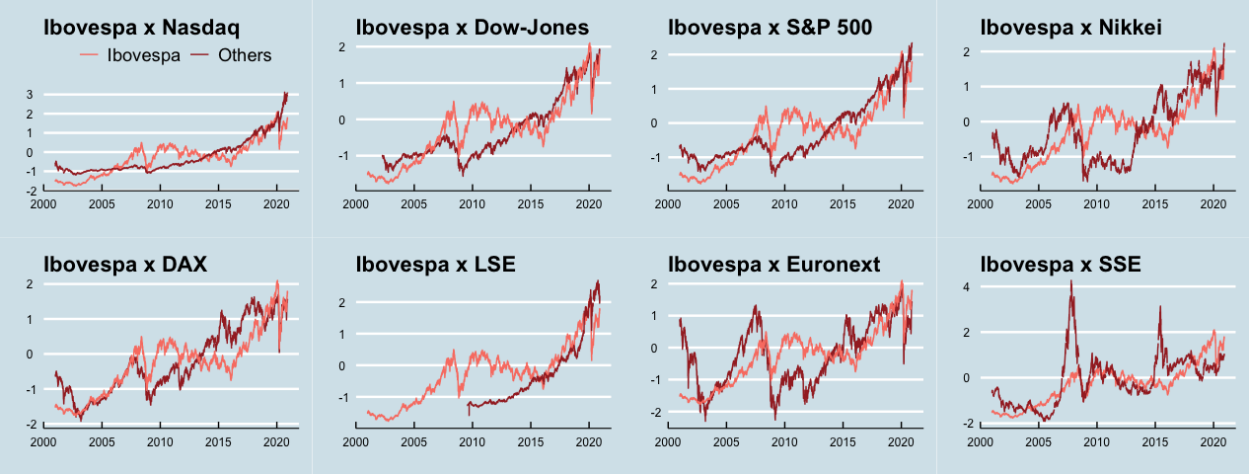
As previously mentioned, financial institutions that operate on the stock exchange have weekly routines, in addition to contracts that have their maturities linked to week day. So let's look for patterns that can help us with predictions.



We can see the week days are well distributed and, looking at the means closely, we can see a linear rise from Monday to Friday. Between Monday and Friday the average difference is 127 points. Tuesday has the smallest standard deviation, 4,530 while Monday the biggest, 4,770, only 5.29% bigger.

2.3 Markets Around the World

Large investment funds move billions of dollars daily across the world according to events. We will analyze the movements of the main world exchanges in comparison to São Paulo. As each index has its score, the values have been scaled.



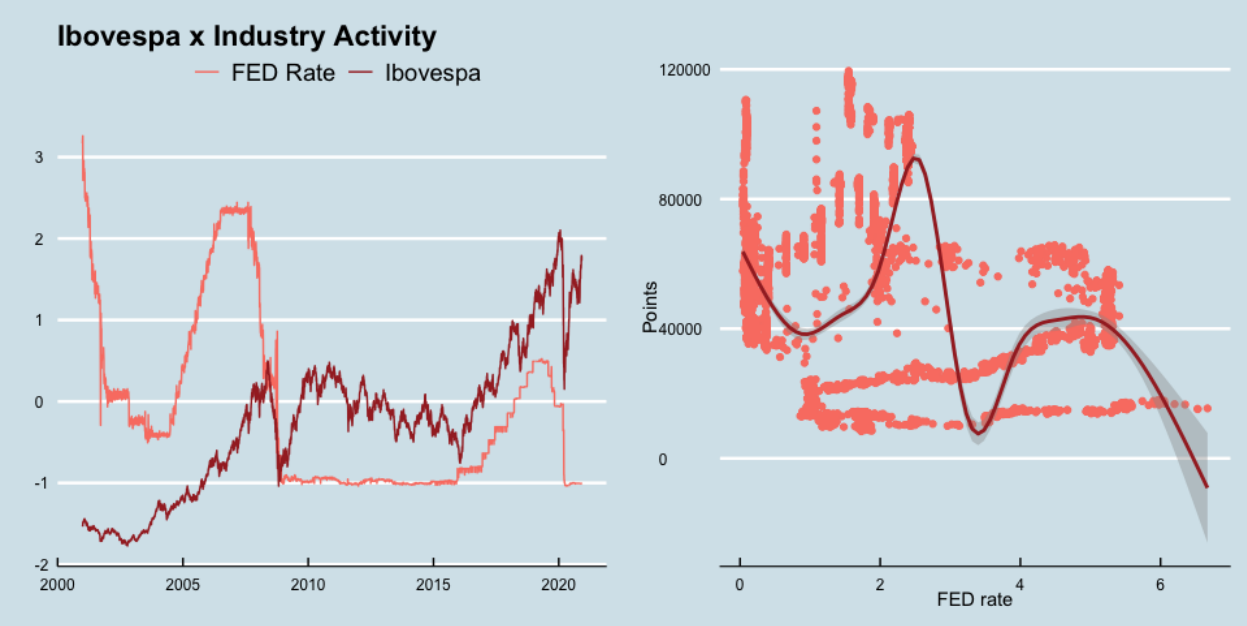
We note that all exchanges come from a depression after the fall of the twin towers in September 2001. The European ones had more severe crashes and have a good recovery, when in 2008, all markets plummeted, especially Euronext. While DAX had a recovery more like the Dow-Jones and S&P 500, Nasdaq and LSE had an almost exponential recovery.

Exchange	Correlation
Nasdaq	0.86
Dow Jones	0.87
S&P 500	0.85
DAX	0.84
LSE	0.82
Euronext	0.69
Nikkei	0.74
SSE	0.59

We can see al exchanges around the world are highly correlated and, the greatest correlation is with the Dow Jones index, but all others scores almost the same, with the exception of SSE and Euronext slightly below.

2.3.1 FED Bonds

As previously stated, US Treasury bills are considered to be the safest investment and attracts investors, so we are going to observe the Ibovespa variation compared to changes in FED rates.



The correlation between Ibovespa and the FED rate is -0.07, it really seems difficult to see many relationships between the movement of the two lines, if not for the fact that the FED seems to anticipate the big falls. Despite the zigzag, the correlation is clearly negative.

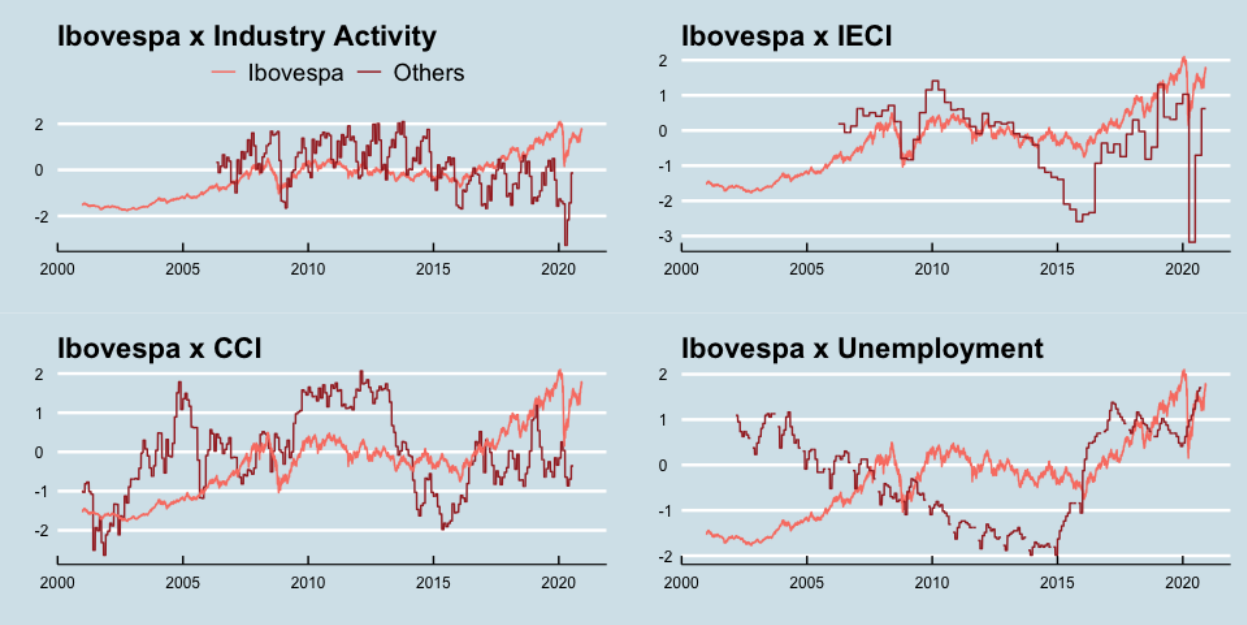
2.4 Local Economy

As, theoretically, the stock exchange is a portrait of main companies in the country, it should be affected by the local economic activity. Let's start with industrial activity, as among the 20 largest we have 8 industries. This data came from IBGE.

As the market lives on expectations, we will use the IECI, the industrial entrepreneur confidence index, to help us understand the relationship between the expectations of entrepreneurs and the movement of the stock exchange. It came from CNI opinion polls.

Of the 20 main companies, 11 depend significantly on local retail, which obviously also influences industrial production, so let's assess the correlation between Ibovespa and the CCI, consumer confidence index. That information came from Fecomercio opinion polls.

As the basis of all this is people's income, let's look at the performance of stock exchange compared to unemployment rate. All values were scaled to facilitate comparison.



Industrial activity seems to follow the Ibovespa but with a delay and more volatility. After 2015, it was stagnant below average while the stock index soared upward, acting homogeneously in the pandemic. The correlation between them is -0.42, negative and low.

The businessmen mood was glued to the stock market's performance when, with a rapid recovery after 2008, confidence surpassed the market, but when show itself weak, the discouragement was much greater than Ibovespa fall. Even with the index rising sharply in recent years, the mood of businessmen has not kept up with the same intensity, in oposit, with the pandemic it fell to historic low, but recovering at the same speed. The correlation between them is 0.46, positive, but not strong.

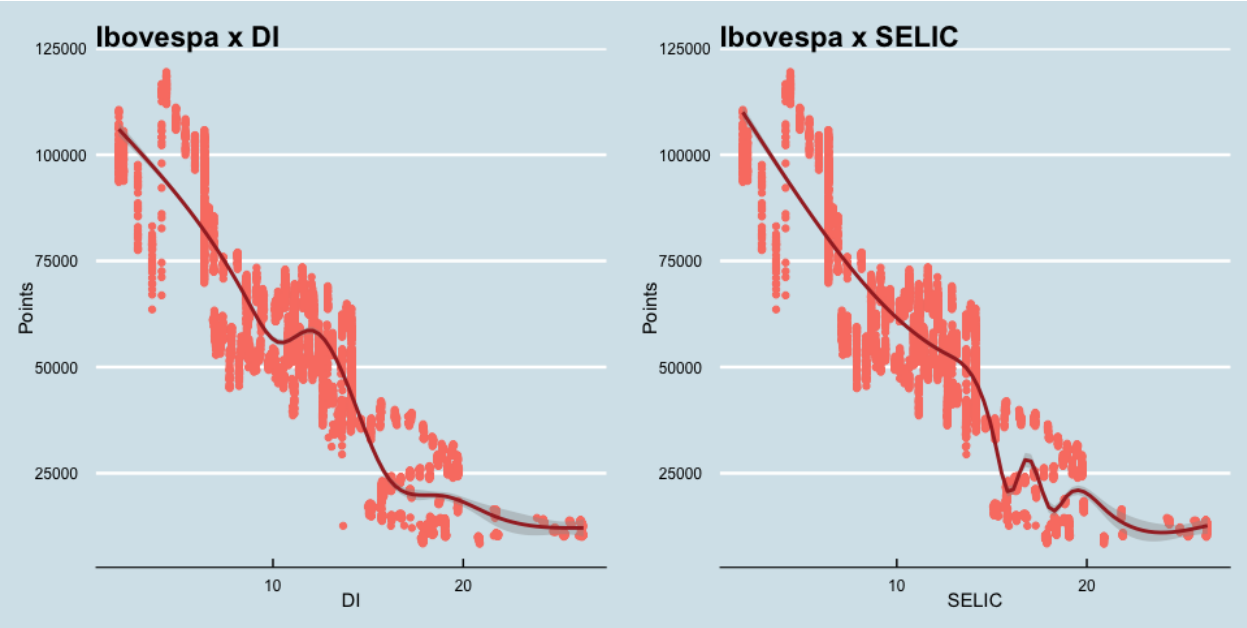
With a correlation of 0.27, consumer confidence seemed overestimated before 2013, when the mood changed and seems underestimated until today compared to Ibovespa.

Between 2000 and 2015, the unemployment rate was negatively correlated with market performance, as expected. However, from 2015 to today, inexplicably, unemployment and the value of shares are constantly growing together. Thus the total correlation in the period was 0.26.

2.4.1 Interest rates

The interest rate directly impacts the economy, as if too high it tends to discourage investments in production and consumption, if it is too low it can cause an exaggerated increase in credit, generating bubbles and inflation. In addition, 6 of the 20 largest companies in the index are banks or related to the sector.

We will analyze 2 rates, DI and SELIC. DI, interbank deposit, is a market rate and works as parameter for investments in fixed income. The SELIC, in turn, is the basic interest rate determined by BCB, it is its main monetary policy instrument to control inflation.



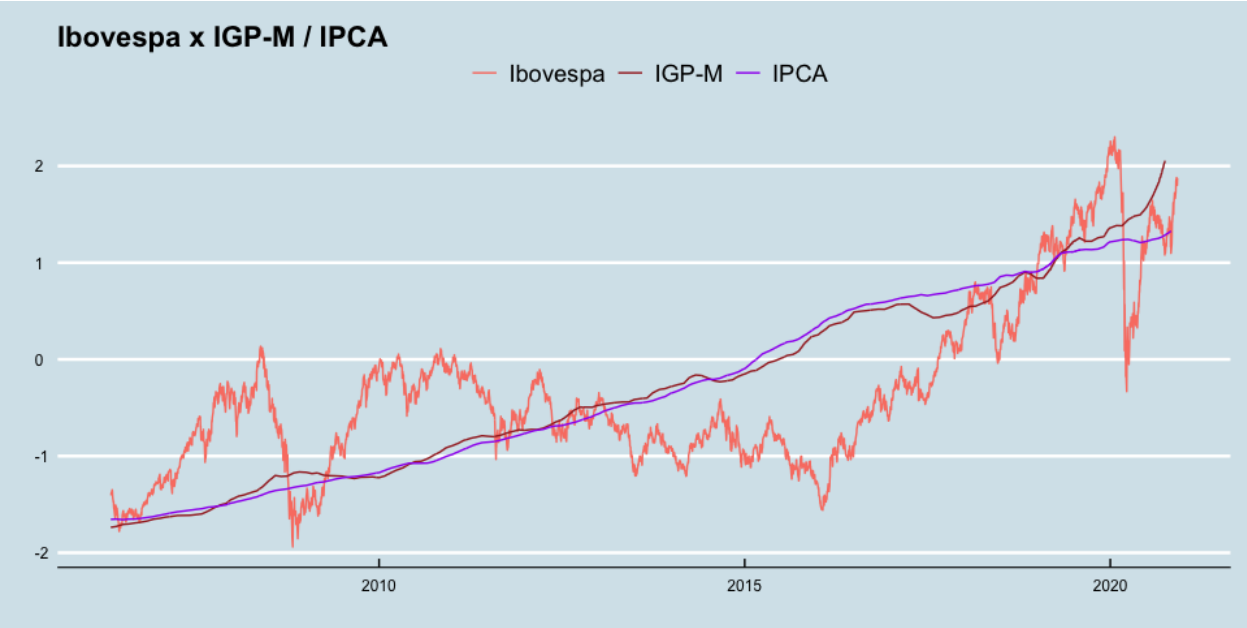
We can see the two rates are totally correlated each other and, in relation to Ibovespa, both with negative correlation equal to 0.9.

2.4.2 Price inflation indices

Inflation has many causes and consequences, it can be a sign of consumption or scarcity of goods, it can be related to an abundance of credit, an increase in amount of money in market, or to exchange rate, since Brazil imports a wide range of essential products.

Two different indices were selected for comparison.

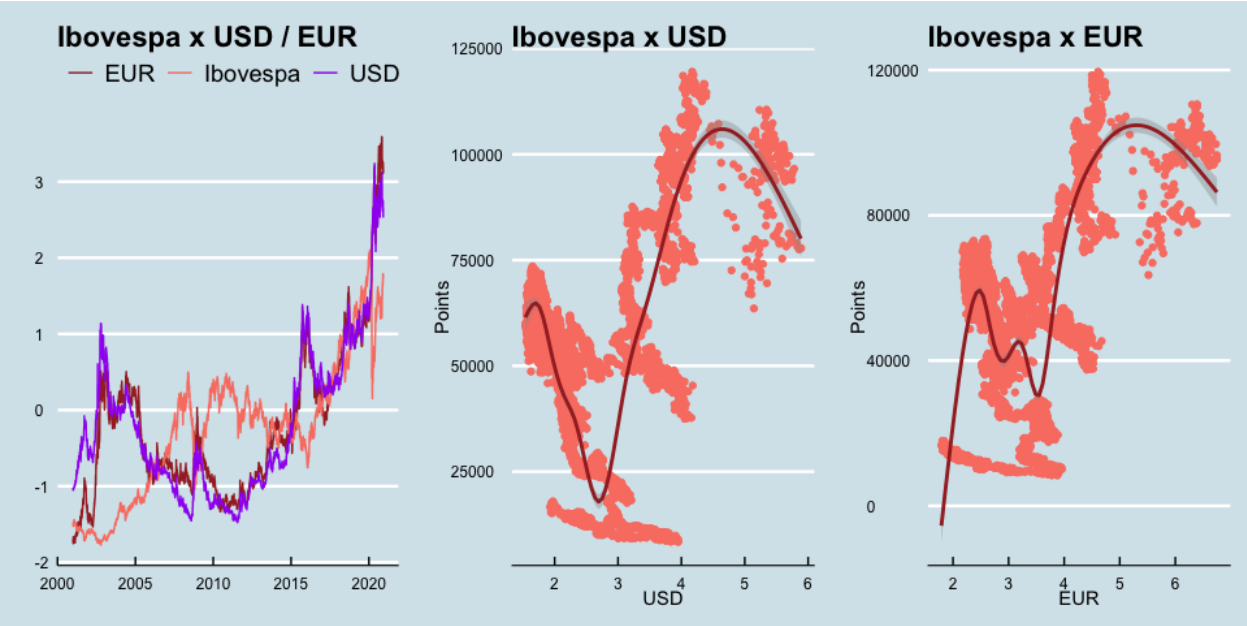
- IPCA - Acronym in Portuguese for Extended National Consumer Price Index, is calculated by IBGE and is considered the official inflation index of country. The indicator serves as a reference for Central Bank to decide Selic, measures the prices of products and services charged to families living in metropolitan regions with monthly incomes of 1 to 40 minimum wages and, the following groups are researched: food and beverages, housing, residence, clothing, transport, health and personal care, personal expenses, education and communication.
- IGP-M - Acronym in Portuguese for General Price Index - Market, is calculated by FGV and checks wholesale prices, for the producer (60%), retail, for the consumer (30%) and in the construction sector (10%). The IGP-M is used to correct rent contracts and electricity supply prices.



It's easy to see that both accumulated inflation indexes are extremely correlated with each other, 0.99, and very correlated with Ibovespa, 0.78 for IPCA and 0.9 for IGP-M, the highest among all indicators. We can say the Ibovespa is only paying for price increase.

2.5 Exchange Rate

At least half of the 20 largest companies on Ibovespa export, import or have representation in other countries. In addition, Brazil is dependent on imports of petroleum fuel, chemical fertilizers, products from manufacturing industry and telecommunications equipment. Until very recently, Brazil had attractive interest rates, which increased the inflow of foreign capital keeping exchange rate controlled. This has been changing lately, let's see the impact of Euro and Dollar on Ibovespa.

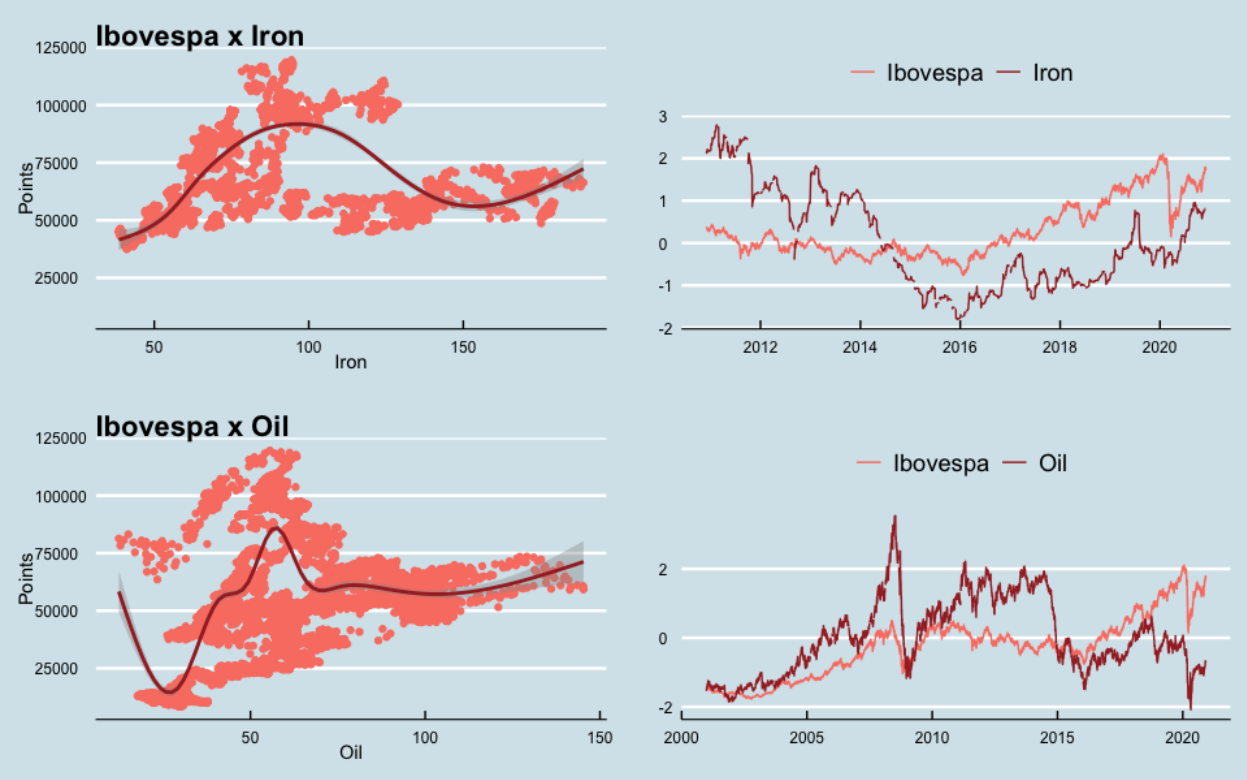


It is easy to see the correlation between the two currencies, with Ibovespa the correlation with dollar is 0.51 and with euro is 0.6. We can see between 2000 and 2017 movements suggest a negative correlation, while between 2017 and 2020 they have been going together, separating again with pandemic.

We noticed correlation between Ibovesta and USD, with dollar between R$ 1.50 and R$ 2.70 is very negative, as well as between R$ 4.50 and R$ 6.00. In opposit, between R$ 2.70 and R$ 4.50 we have a very positive correlation. Euro had a similar behavior, but with a very positive correlation with Ibovespa between R$ 1.50 and R$ 2.50 and, between R$ 3.60 and R$ 5.00, in opposit, a negative correlation between R$ 2.50 and R$ 3.60 and R$ 5.00 to more.

2.6 Iron & Oil

Ibovespa's largest company is a steelmaker with a weight of 12.46% of the index, so we will analyze the impact of world future price of iron ore on São Paulo stock exchange. The second largest company among 20 largest is Petrbras, with more than 10% of importance in the index, so we will also assess the future price of oil in relation to Ibovespa. These are purchase and sale contracts for iron and oil with a future maturity. They are derivatives, but can help us to understand expectations about these commodities and how they affect the index.



We can see the correlation between Ibovespa and iron ore was positive while iron was between 50 and 100 dollars. Between 100 and 150 dollars the correlation became negative, returning to be positive above 150 dollars.

Looking at iron ore and index through the timeline, we notice that between 2010 and 2016 the price of iron ore fell a lot while the Ibovespa remained sideways . From 2016 to 2020, both followed an upward trend and fell sharply with the COVID-19 pandemic.

The correlation, throughout the series, between Ibovespa and iron ore is -0.03 and between index and oil is 0.26, both with a small correlation with the Ibovespa.

2.7 Predictions

To build our forecasting model we have to choose which indicators will be used and assess whether they need any changes. We will evaluate used techniques and their results.

2.7.1 Dates

For our prediction model, we cannot use variables that we will not have access to on a daily basis, such as the closing of the Ibovespa on the same day, for example. With that in mind we will use the Ibovespa opening. Closing, high, low and volume will be from the previous day. On international exchanges we will also use the opening values, with the exception of Asian ones where the closing values could be used due to fuser. The used values for USD, EUR, DI, WTI and iron ore will also be the opening.

The other indicators that are released monthly or quarterly will be used in their normal dates and, when doing future projections, for these data we can use Focus Report estimates, which considers the market expectations collected until the Friday before its release every Monday.

The report presents the weekly behavior of the projections for price indices, economic activity, foreign exchange, SELIC rate, among other indicators. It is released by the BCB, but the projections are for the market. During the week the values will be repeated and so on until the official data are released.

2.7.2 Predictors

In addition to the Ibovespa historical data with appropriate dates, month, month day and week day were used. Quarter was not in preference to month. As we saw, DI and SELIC rates are totally correlated each other and, in relation to Ibovespa, both with negative correlation equal to 0.9, so only DI was used once it is traded daily, which will allow us to work with more updated data on day by day.

FED rate was used and from world stock exchanges, Dow-Jones was chosen representing the American ones for the highest correlation. Among Europeans, Euronext was used for the lowest correlation with Dow-Jones and, as for Asians, the two were excluded, Japanese for having a very high correlation with American (0.92) and European (0.94) and, Chinese for low correlation with Ibovespa.

Between currencies the choice was made in the same way. As EUR is more correlated with Ibovespa than USD and both are very correlated with each other, 0.96, only euro was used.

As both accumulated inflation indexes are extremely correlated each other, 0.99, and very correlated with Ibovespa, 0.78 for IPCA and 0.9 for IGP-M, the last one was chosen. Other economic indicators, CCI, industry activity and unemployment rate, were also used despite the low correlation with the index, as well as iron and oil.

2.7.3 Logistic regression

The logistical transformation for a proportion or rate p is defined as: $g(p) = log(p/(1-p))$ when p is a ratio or probability, the quantity being logged, $p/(1-p)$, is called probabilities. The transformation of the log makes this symmetrical. If the rates are the same, then the probability of logging is 0. Increases or decreases of times become positive and negative increments, respectively. Using the log, these fold changes become constant increases. Logistic regression is a specific case of a set of generalized linear models. The function $\beta_0 + \beta_1 x$ can take on any value, including negatives and values greater than 1. But we are estimating a probability: $Pr(Y = 1 \mid X = x)$ which is restricted between 0 and 1. The idea of generalized linear models (GLM) is:

1) to define a Y distribution that is consistent with your possible results and,

2) find a function g so that $g(Pr(Y = 1 \mid X = x))$ can be modeled as a linear combination of predictors.

Logistic regression is the most commonly used GLM. It is an extension of linear regression that ensures that the estimate of $Pr(Y = 1 \mid X = x)$ is between 0 and 1. This logistic transformation converts the probability into a log of chances. A good feature of this transformation is that it converts the probabilities to symmetric around 0.

$$g\{p(x_1, x_2, ..., x_{20})\} = g\{Pr(Y = 1 \mid X_1 = x_1, X_2 = x_2, ..., X_{20} = x_{20})\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{20} x_{20}.$$

As we have 20 predictors, our estimate is:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_{20} x_{i,20} + \varepsilon_i.$$

Two predictions were made, one for high (h) and other for low (l), being:

$$Y_{h,i} = \beta_{h,0} + \beta_{h,1} x_{i,1} + \beta_{h,2} x_{i,2} + ... + \beta_{h,20} x_{i,20} + \varepsilon_{h,i}.$$
$$Y_{l,i} = \beta_{l,0} + \beta_{l,1} x_{i,1} + \beta_{l,2} x_{i,2} + ... + \beta_{l,20} x_{i,20} + \varepsilon_{l,i}.$$

For $Y_{h,i}$ using *glm* function we had:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1598.1005103	1456.3652174	1.0973213	0.2725280	-1256.3228641	4452.5238847
date	-0.0651220	0.1001160	-0.6504659	0.5154066	-0.2613458	0.1311017
open	0.3065785	0.0063217	48.4960101	0.0000000	0.2941881	0.3189688
month	4.5904757	2.6920401	1.7052033	0.0881883	-0.6858259	9.8667773
month_day	1.1881867	0.4711640	2.5218110	0.0116910	0.2647221	2.1116512
week_day	-10.2428244	2.8861488	-3.5489592	0.0003886	-15.8995720	-4.5860767
di	-19.0831634	6.0519671	-3.1532166	0.0016197	-30.9448010	-7.2215258
fed_rate	68.0609174	18.6398669	3.6513628	0.0002622	31.5274496	104.5943852
euronext	-0.1939732	0.2116196	-0.9166125	0.3593683	-0.6087400	0.2207936
dj	0.0189058	0.0127214	1.4861413	0.1372740	-0.0060277	0.0438394
eur	154.1501332	30.3234329	5.0835317	0.0000004	94.7172969	213.5829695
wti	-2.5174162	0.8231046	-3.0584403	0.0022309	-4.1306716	-0.9041609
iron	2.6997106	0.5511472	4.8983478	0.0000010	1.6194820	3.7799392
cci	0.0970448	0.7954610	0.1219981	0.9029030	-1.4620302	1.6561197
industry_activity	-2.5281255	1.2005187	-2.1058610	0.0352419	-4.8810990	-0.1751521
unemployment	17.5686266	6.4007351	2.7447826	0.0060662	5.0234162	30.1138369
igpm_acc	-2.4020353	2.5070314	-0.9581194	0.3380263	-7.3157267	2.5116560
close	0.3249140	0.0068359	47.5307419	0.0000000	0.3115160	0.3383121
‘high_-1‘	0.5221416	0.0087074	59.9654192	0.0000000	0.5050754	0.5392077
‘low_-1‘	-0.1630266	0.0082893	-19.6671046	0.0000000	-0.1792733	-0.1467799
volM	-6.4833267	4.6905451	-1.3822118	0.1669383	-15.6766260	2.7099727

And for $Y_{l,i}$:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	801.2748467	1859.0528140	0.4310124	0.6664689	-2842.4017141	4444.9514075
date	0.0886806	0.1277983	0.6939104	0.4877549	-0.1617995	0.3391606
open	0.2643285	0.0080697	32.7557083	0.0000000	0.2485122	0.2801448
month	10.7504108	3.4363940	3.1283988	0.0017628	4.0152023	17.4856192
month_day	2.0160133	0.6014417	3.3519677	0.0008054	0.8372091	3.1948174
week_day	-7.8871251	3.6841741	-2.1408123	0.0323139	-15.1079737	-0.6662766
di	-11.0253631	7.7253469	-1.4271674	0.1535636	-26.1667647	4.1160385
fed_rate	134.4645823	23.7938235	5.6512390	0.0000000	87.8295451	181.0996195
euronext	-0.5474072	0.2701328	-2.0264372	0.0427471	-1.0768577	-0.0179566
dj	0.0263192	0.0162389	1.6207459	0.1051044	-0.0055085	0.0581469
eur	7.1901483	38.7079165	0.1857539	0.8526416	-68.6759739	83.0562706
wti	-2.4726596	1.0506945	-2.3533573	0.0186245	-4.5319829	-0.4133362
iron	3.0528384	0.7035404	4.3392513	0.0000144	1.6739247	4.4317522
cci	-4.6638434	1.0154074	-4.5930758	0.0000044	-6.6540054	-2.6736814
industry_activity	-4.3330697	1.5324643	-2.8275176	0.0047006	-7.3366446	-1.3294949
unemployment	31.1890948	8.1705499	3.8172577	0.0001358	15.1751113	47.2030783
igpm_acc	-10.2470927	3.2002301	-3.2019862	0.0013692	-16.5194285	-3.9747569
close	0.2992637	0.0087260	34.2956319	0.0000000	0.2821611	0.3163664
‘high_-1‘	-0.2718450	0.0111150	-24.4575306	0.0000000	-0.2936300	-0.2500601
‘low_-1‘	0.6996295	0.0105813	66.1193557	0.0000000	0.6788905	0.7203684
volM	1.0969302	5.9874892	0.1832037	0.8546420	-10.6383329	12.8321934

As a confidence interval around the projection of high and low, RMSE was used.

RMSE

It is a measure of accuracy for comparing forecasting errors from different models for a given dataset. In this case, it is the typical mistake we make when predicting high and low.

We define $y_{h,i}$ as the Ibovespa high score for day i and denote our prediction with $\hat{y}_{h,i}$. The high RMSE is then defined as:

$$high\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{h,i} - y_{h,i})^2}$$

The same for low RMSE:

$$low\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{l,i} - y_{l,i})^2}$$

We obtained 443.55 as high RMSE and 647.72 as low RMSE. Analyzes were performed with 1 and 2 RMSE as a confidence interval obtaining the following results:

- 2 RMSE
  - Low accuracy 96.7%
  - High accuracy 94.95%
  - Total accuracy 93.3%

At first glance we had great results as we reached 95% reliability in max and min, in addition to a much better accuracy than the 80% desired for the total hit, but the results were not so good. Our trading margin, the difference between the upper limit of low and the lower limit of high, was negative in average at 643.13, which means that the confidence margins crossed and the prediction was not useful.

- 1 RMSE
  - Low accuracy 90%
  - High accuracy 86.24%
  - Total accuracy 82.29%
  - Partial accuracy 93.94%
  - Trading margin 448.13.

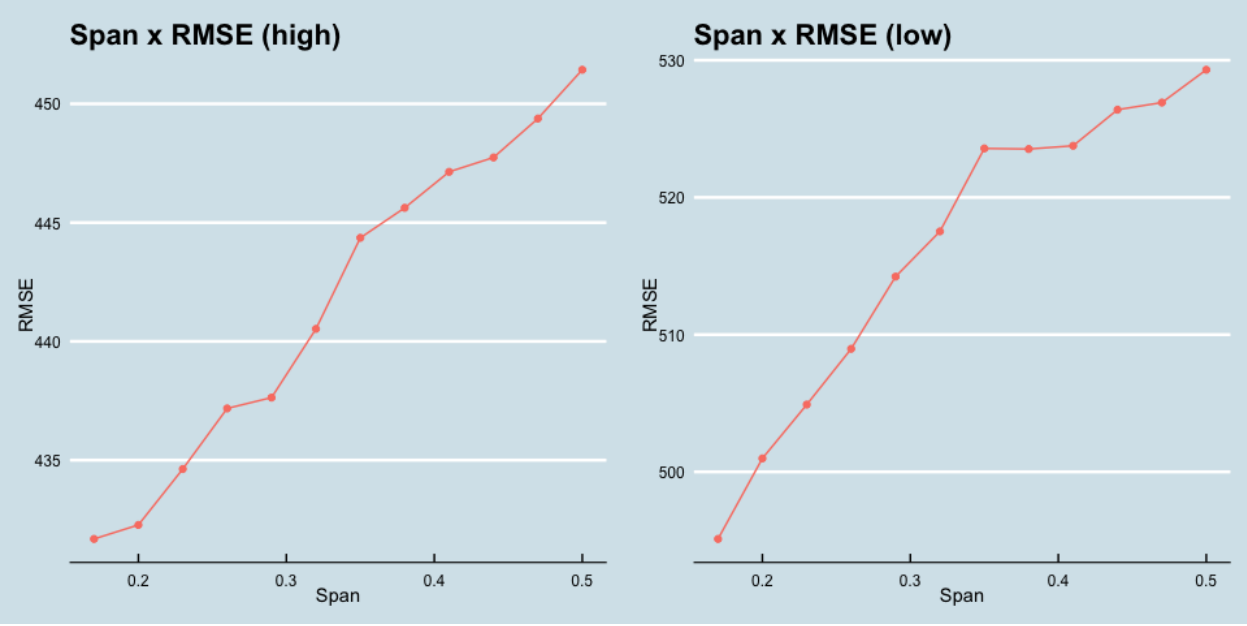
With 1 RMSE as confidence interval, we maintained total accuracy above the initial target, but the low and high reliability were below 95%. This time the trading margin is positive and we have a partial accuracy of 93.94%, compared to 98.35% with 2 RMSE. This information is very important because when we hit only low or high we still have chances to make good deals. If we hit low and buy at the maximum low limit, for example, if high is above our forecast, or close within the trading margin, the deal is profitable, causing loss only with closing within the low confidence interval.

2.7.4 Loess

As explained before, loess was used, as logistic regression, to estimate $Y_{l,i}$ and $Y_{h,i}$. It was made with the same predictors, except for high and low with lagged dates, 18 predictors at all.

$$Y_{h,i} = \beta_{h,0} + \beta_{h,1}x_{i,1} + \beta_{h,2}x_{i,2} + ... + \beta_{h,18}x_{i,18} + \varepsilon_{h,i} \quad if \quad \mid x_i - x_0 \mid \leq h.$$
$$Y_{l,i} = \beta_{l,0} + \beta_{l,1}x_{i,1} + \beta_{l,2}x_{i,2} + ... + \beta_{l,18}x_{i,18} + \varepsilon_{l,i} \quad if \quad \mid x_i - x_0 \mid \leq h.$$

Train set was divided into train set 2 (90%) and test set 2 (10%) to find the best span for low and high by cross validation. We can see that bigger the span bigger the error, so span = 0.17 was used, as small as possible.



For $Y_{h,i}$ using *gamLoess* method of *train* function we had:

```
## Generalized Additive Model using LOESS
##
## 9791 samples
##   21 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 9791, 9791, 9791, 9791, 9791, 9791, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   416.7224  0.9992093  290.3493
##
## Tuning parameter 'span' was held constant at a value of 0.17
## Tuning
##  parameter 'degree' was held constant at a value of 1
```

And for $Y_{l,i}$:

```
## Generalized Additive Model using LOESS
##
## 9791 samples
##   21 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 9791, 9791, 9791, 9791, 9791, 9791, ...
## Resampling results:
##
##   RMSE     Rsquared   MAE
##   543.177  0.9986276  362.6612
##
## Tuning parameter 'span' was held constant at a value of 0.17
## Tuning
##  parameter 'degree' was held constant at a value of 1
```

Now was obtained 494.49 as high RMSE and 648.74 as low RMSE, worse than logistic, and the following results:

- 2 RMSE
  - Low accuracy 97.71%
  - High accuracy 97.71%
  - Total accuracy 95.87%
  - Partial accuracy 99.54%
  - Trading margin -757.57.

We achieved excellent accuracy, better than with logistic regression, but the margin of negotiation was worse.

- 1 RMSE
  - Low accuracy 90.09%
  - High accuracy 81.47%
  - Total accuracy 76.42%
  - Partial accuracy 95.14%
  - Trading margin 385.66.

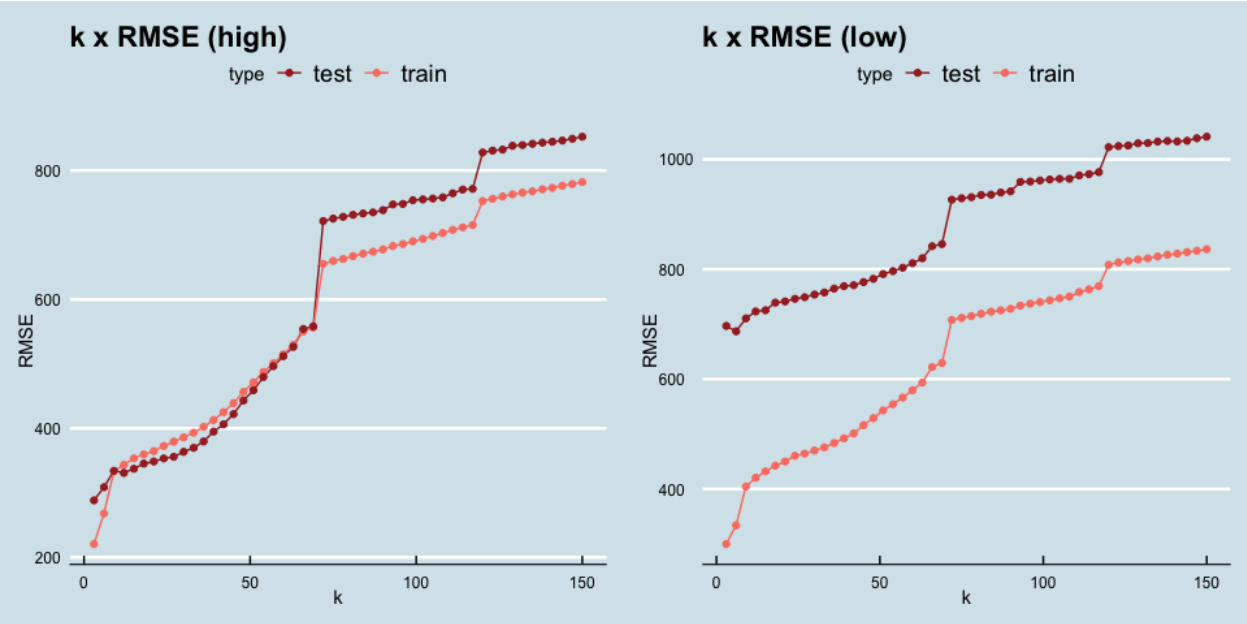
With 1 RMSE as a confidence interval, we failed in low, high and total accuracy. This method was better in partial precision and allowed a positive trading margin.

2.7.4 K-nearest neighbors

This technique is easy to adapt to multiple dimensions and consists of defining the distance between all observations based on the predictors and then, for any point we want an estimate, we look for the k closest points and average associated values with those points.

We refer to the set of points used to calculate the mean as a neighborhood and the number of neighbors can be used to control the flexibility of our estimate using the parameter k. Larger ks result in smoother and lower ks result in more flexible and fluctuating estimates.

To define the size of neighborhood, cross validation between train set 2 and test set 2 was used and the values were estimated on the two databases.



Estimates were generated using between 3 and 150 neighbors and, we can see in both cases the error between train and test set is always greater and increasing, so k was chosen based on lowest RMSE and was 3 for high and 6 for low.

For $Y_{h,i}$ using *knn* method of *train* function we had:

```
## k-Nearest Neighbors
##
## 9791 samples
##   21 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 9791, 9791, 9791, 9791, 9791, 9791, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   339.1579  0.9994696  105.373
##
## Tuning parameter 'k' was held constant at a value of 3
```

And for $Y_{l,i}$:

```
## k-Nearest Neighbors
##
## 9791 samples
##   21 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 9791, 9791, 9791, 9791, 9791, 9791, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   469.5632  0.9989692  135.1429
##
## Tuning parameter 'k' was held constant at a value of 6
```

Now was obtained 304.45 as high RMSE and 536.19 as low RMSE, considerably better than the first two and, the following results:

- 2 RMSE
  - Low accuracy 98.72%
  - High accuracy 96.42%
  - Total accuracy 95.87%
  - Partial accuracy 99.27%
  - Trading margin -149.41.

We achieved best accuracy until now, but the margin for negotiation still negative.

- 1 RMSE
  - Low accuracy 94.5%
  - High accuracy 91.1%
  - Total accuracy 89.54%
  - Partial accuracy 96.06%
  - Trading margin 691.23.

With 1 RMSE as a confidence interval we practically hit low, stayed close to high and had excellent total and partial accuracy. In addition to a much higher trading margin.

2.7.5 Random forest

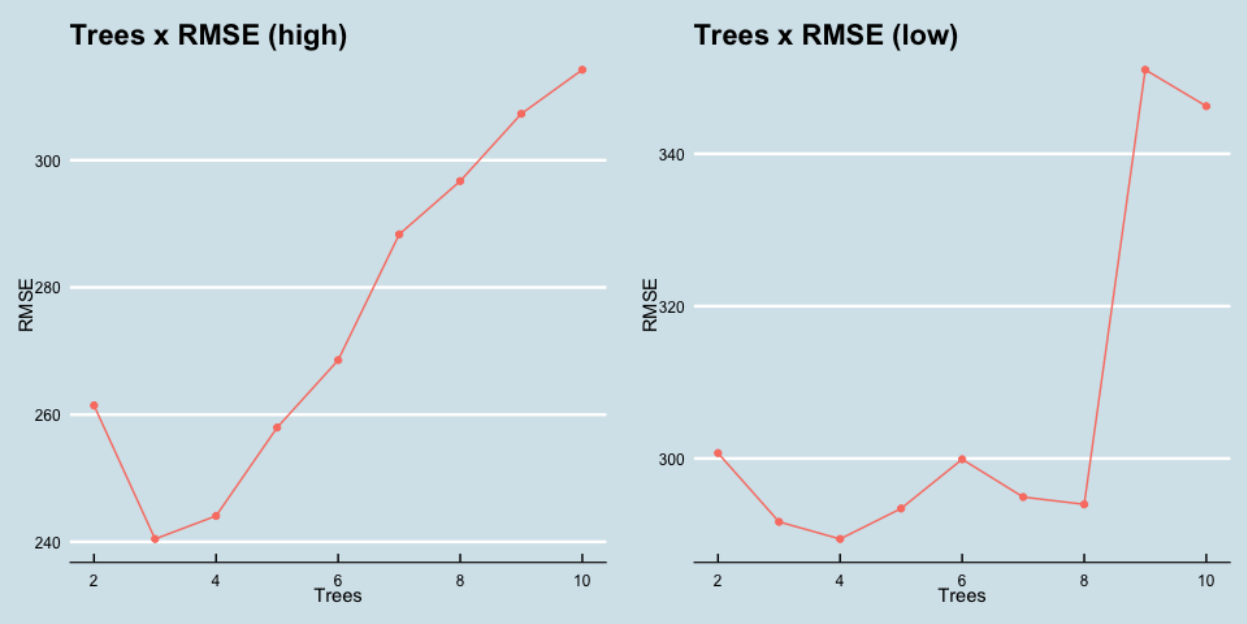
Decision trees are used in forecasting problems where the result is categorical and we form predictions by calculating which class is the most common among observations.

Random forests addresses the deficiencies of decision trees where the objective is to improve forecast performance and reduce instability by averaging several decision trees (a randomly constructed tree forest).

The first step is packaging of bootstrap. The general idea is to generate many predictors, each using regression trees, then form a final prediction based on the average prediction for all those trees. To ensure that the individual trees are not the same, we use the bootstrap, making the individual trees randomly different. The combination of trees is the forest.

We built B decision trees using the training set we refer to the adjusted models as $T_1, T_2, ..., T_B$. For each observation in the test set we made a prediction $\hat{y}_j$ using the T_j tree. As we are using continuous results, the final forecast was formed with the average $\hat{y} = \frac{1}{B}\sum_{j=1}^{B}\hat{y}_j$.

In the train function, with "rf" method, we have mtry parameter to define the number of trees used and, to define the best, cross validation between train and test set 2 was used.



The number of trees that minimizes the RMSE was used, 3 for high and 4 for low.

For $Y_{h,i}$ using *rf* method of *train* function we had:

```
## Random Forest
##
## 9791 samples
##   21 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 9791, 9791, 9791, 9791, 9791, 9791, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   284.7276  0.9996272  90.24329
##
## Tuning parameter 'mtry' was held constant at a value of 3
```

And for $Y_{l,i}$:

```
## Random Forest
##
## 9791 samples
##   21 predictor
```

```
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 9791, 9791, 9791, 9791, 9791, 9791, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   397.5048  0.9992626  102.3881
##
## Tuning parameter 'mtry' was held constant at a value of 4
```

Now was obtained 249.02 as high RMSE and 255.55 as low RMSE, the best at all, with the following results:

- 2 RMSE
  - Low accuracy 97.98%
  - High accuracy 97.89%
  - Total accuracy 96.79%
  - Partial accuracy 99.08%
  - Trading margin 547.02.

Random forest gave us the best overall result among all techniques. Low accuracy was practically tied with knn, high and total accuracy were the best, while partial accuracy was slightly behind loess and knn. The big difference is that with random forest we get an positive trading margin average of more than 500 points and, with that, we managed to easily reach the project initial objective.

- 1 RMSE
  - Low accuracy 93.94%
  - High accuracy 94.13%
  - Total accuracy 91.19%
  - Partial accuracy 96.88%
  - Trading margin 1,051.58.

With 1 RMSE as a confidence interval we practically hit high, missing 1% to hit low, total accuracy was achieved with a lot of surplus, we got the best partial accuracy and, the trading margin is practically double the others average.

3 Results

Using 2 RMSE as confidence margin, all techniques achieved the proposed objectives and, with ease when it comes to total accuracy. The main problem was trading margin, which was negative in all cases, with exception of random forest.

names	Low.Accuracy	High.Accuracy	Total.Accuracy	Partial.Accuracy	Trading.Margin
Logistic	96.70	94.95	93.30	98.35	-643.13
Loess	97.71	97.71	95.87	99.54	-757.57
Knn	98.72	96.42	95.87	99.27	-149.41
RF	97.98	97.89	96.79	99.08	547.02

The best low accuracy was k-nearst neighbors, 98.72%, the best high and total accuracy was random forest, 97.89% and 96.79%. Local weighted regression was the best partial accuracy hiting 99.54% of time. Because of trading margin, only random forest if shown to be viable.

Using 1 RMSE, all techniques reached the expected total accuracy, with the exception of loess with 76.42%. The best performance was with random forest that reached 91.19%.

names	Low.Accuracy	High.Accuracy	Total.Accuracy	Partial.Accuracy	Trading.Margin
Logistic	90.00	86.24	82.29	93.94	448.13
Loess	90.09	81.47	76.42	95.14	385.66
Knn	94.50	91.10	89.54	96.06	691.23
RF	93.94	94.13	91.19	96.88	1051.58

About high and low accuracy, no prediction reached the goal, but knn low and random forest high almost succeeded, 94.5% and 94.13% respectively. Random forest also had the best partial results with 96.88% accuracy and an incredible trading margin of 1,051.58 points. Using 1 RMSE as a confidence interval, all techniques had a positive trading margin.

4 Conclusion

We can say this project was successful in developing an algorithm to predict Ibovespa's high and low daily scores using its timeline, seasonality, currencies, exchanges, interest rates and economic indicators. Among used techniques, with 2 RMSE as confidence margin, random forest obtained the best results, being the only one that achieved more than 80% total accuracy and superior 95% reliability around high and low, maintaining a positive trading margin, being:

	names	Low.Accuracy	High.Accuracy	Total.Accuracy	Partial.Accuracy	Trading.Margin
4	RF	97.98	97.89	96.79	99.08	547.02

With 1 RMSE as confidence margin, random forest still the best results, but could not achieved more than 95% reliability around high and low, getting close. K-nn came in second and narrowly missed the goal. More distant, less efficient, were local weighted and logistic regressions.

	names	Low.Accuracy	High.Accuracy	Total.Accuracy	Partial.Accuracy	Trading.Margin
3	Knn	94.50	91.10	89.54	96.06	691.23
4	RF	93.94	94.13	91.19	96.88	1051.58

For future model development, early news can be used as a daily mood index, but for that it is necessary a history of news, which was an impilience at that first moment. As we achieve results that allow us to do business, another step will be to collect intraday data and calculate the probabilities on days when we do not reach high and / or low. Even if we lose high and low, we can make a profit if low is below the low minimum and high is above the high maximum, for example. With the probabilities in hand, we will choose the most profitable strategy using Monte Carlo simulation.