# Analyzing the impact of Covid-19 Pandemic and the effect of quarantine in Argentina

## Applied Data Science Capstone

**Martín Alejandro Heredia**

# Introduction

## The Problem

From the beginning of Covid-19 Pandemic, people’s life has changed drastically due to the fast and exponential growth of cases, which leads governments to take decisions on how to prevent the expansion rate of the disease, health system saturation and economic impact.

The common denominator is quarantine. This helps to reduce infection rate, giving more time to governments to prepare health care system. In contrast to this positive consequence of quarantine, is the economic, social and industrial impact that quarantine has on a country. This positive and negative effects of quarantine makes it a “trade-off” solution.

The analysis of growth of cases, distribution of disease on the territory, impact of quarantine, among others; is fundamental for governments to take decisions. Also, viewing the evolution of the disease and its influence on indicators such like economics or industrial ones, gives governments tools to evaluate and reorganize its own decisions.

This study is focused on Argentina. First, statistical data of Covid-19 in Argentina will be analyzed: how numbers of cases are distributed in its geography and where are the most affected provinces. Then, the effect of quarantine in other areas (such like transport, industry, economic) will be studied and compared with the same period of the previous year.

# Data

To analyze the effect of Covid-19 on the territory, data from the following website of Argentinian Government will be used:

https://sisa.msal.gov.ar/datos/descargas/covid-19/files/Covid19Casos.csv

```
In [3]: url='https://sisa.msal.gov.ar/datos/descargas/covid-19/files/Covid19Casos.csv'
        cases=pd.read_csv(url,encoding='utf-16')
        cases.head()
```

Out[3]:

| | id_evento_caso | sexo | edad | edad_años_meses | residencia_pais_nombre | residencia_provincia_nombre | residencia_departa |
|---|---|---|---|---|---|---|---|
| 0 | 672064 | M | 52.0 | Años | Argentina | Buenos Aires | |
| 1 | 717629 | F | 46.0 | Años | Argentina | Buenos Aires | |
| 2 | 717926 | F | 41.0 | Años | Argentina | CABA | |
| 3 | 718029 | F | 52.0 | Años | Argentina | Buenos Aires | |
| 4 | 718055 | F | 34.0 | Años | Argentina | CABA | SIN |

5 rows × 25 columns

All the categories (columns) of this dataset are listed below:

- **CLASIFICACION** : Manual classification of the case
- **asistencia_respiratoria_mecanica** : Indication if mechanical respiratory assistance was required
- **carga_provincia_id** : ID of province were the case was registered
- **carga_provincia_nombre** : Name of province were the case was registered
- **clasificacion_resumen** : General classification of the case
- **cuidado_intensivo** : Indication if the case was in intensive care
- **edad** : Age
- **edad_años_meses** : Indication if “edad” is in months or years
- **fallecido** : Indication of deceased
- f**echa_apertura** : Date when case was open
- **fecha_cui_intensivo** : Date of admission to intensive care, if applicable

- **fecha_diagnostico** : Diagnosis date
- **fecha_fallecimiento** : Date of death, if applicable
- **fecha_inicio_sintomas** : Symptom onset date
- **fecha_internacion** : Hospitalization date
- i**d_evento_caso** : Number of case
- **origen_financiamiento** : Funding source (Public or Private)
- **residencia_departamento_id** : Residence Department code
- **residencia_departamento_nombre** : Residence Department name
- **residencia_pais_nombre** : Residence Country name
- **residencia_provincia_id** : Residence Province code
- **residencia_provincia_nombre** : Residence Province name
- **sepi_apertura** : Opening date of Epidemiological Week
- **sexo** : gender
- **ultima_actualizacion** : Last update

This columns provide information such like:

- Age and gender of the patients
- Dates (diagnosis date, hospitalization date, date of death, among others)
- Geographical information (province name, department name)
- Classification (active, suspect, cured, among others)
- ID of each case

With this dataset, we can visualize the impact of this disease with indicators such like:

- Total of cases
- Total of deaths
- Total of recovered patients
- Total of active cases
- Total number of hospitalized patients
- Total of cases on intensive care

In addition to this data, the use of **folium** will help us to visualize the distribution of cases on the country.

After this preliminary study, the work will be focused on analyzing and visualizing the impact of quarantine on the following areas:

- Economic activity:
  https://www.indec.gob.ar/ftp/cuadros/economia/sh_emae_mensual_base2004.xls
  https://www.indec.gob.ar/ftp/cuadros/economia/sh_emae_actividad_base2004.xls

- Industrial activity:
  https://www.indec.gob.ar/ftp/cuadros/economia/sh_ipi_manufacturero_2020.xls

- Transport (railway): https://servicios.transporte.gob.ar/gobierno_abierto/descargar.php?t=trenes&d=pasajeros

Next, each of the previous datasets will be introduced.

## Economics

There are two datasets in this category. Both of them shows the evolution of EMAE estimator ("Economic Activity Monthly Estimation" in English). The first dataset, shows the evolution of this estimator, related to economic activity of year 2004:

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 205 | Abril | 153,3 | 0,1 | 147,2 | -3,2 | 147,8 | -0,6 |
| 206 | Mayo | 159,6 | -4,7 | 143,7 | -2,4 | 146,8 | -0,6 |
| 207 | Junio | 149,0 | -6,5 | 142,2 | -1,0 | 145,9 | -0,6 |
| 208 | Julio | 145,6 | -2,9 | 142,9 | 0,5 | 145,0 | -0,6 |
| 209 | Agosto | 146,2 | -1,8 | 145,9 | 2,1 | 144,2 | -0,5 |
| 210 | Septiembre | 137,8 | -6,3 | 143,3 | -1,8 | 143,6 | -0,4 |
| 211 | Octubre | 143,0 | -4,0 | 144,2 | 0,7 | 143,1 | -0,3 |
| 212 | Noviembre | 140,9 | -7,3 | 141,9 | -1,6 | 142,7 | -0,2 |
| 213 | Diciembre | 136,8 | -7,0 | 142,1 | 0,1 | 142,5 | -0,2 |
| 214 | | | | | | | |
| 215 | **2019** | | | | | | |
| 216 | Enero | 134,9 | -5,7 | 143,4 | 0,9 | 142,4 | -0,1 |
| 217 | Febrero | 132,4 | -4,7 | 143,4 | 0,0 | 142,4 | 0,0 |
| 218 | Marzo | 144,8 | -7,0 | 142,0 | -1,0 | 142,4 | 0,0 |
| 219 | Abril | 150,7 | -1,7 | 142,3 | 0,2 | 142,5 | 0,0 |
| 220 | Mayo | 162,8 | 2,0 | 142,3 | 0,0 | 142,5 | 0,0 |
| 221 | Junio | 148,4 | -0,4 | 141,7 | -0,4 | 142,5 | 0,0 |
| 222 | Julio | 146,2 | 0,4 | 144,7 | 2,1 | 142,4 | -0,1 |
| 223 | Agosto | 140,9 | -3,6 | 144,0 | -0,4 | 142,2 | -0,1 |
| 224 | Septiembre | 134,9 | -2,1 | 141,6 | -1,7 | 142,0 | -0,2 |
| 225 | Octubre | 141,6 | -0,9 | 143,9 | 1,6 | 141,7 | -0,2 |
| 226 | Noviembre | 137,9 | -2,2 | 141,3 | -1,8 | 141,3 | -0,3 |
| 227 | Diciembre | 136,4 | -0,3 | 141,1 | -0,1 | 140,8 | -0,3 |
| 228 | | | | | | | |
| 229 | **2020** | | | | | | |
| 230 | Enero | 132,3 | -1,9 | 141,1 | 0,0 | 140,3 | -0,3 |
| 231 | Febrero | 129,2 | -2,4 | 139,2 | -1,3 | 139,8 | -0,4 |
| 232 | Marzo | 128,2 | -11,5 | 125,6 | -9,8 | 139,3 | -0,3 |
| 233 | | | | | | | |
| 234 | **Fuente:** INDEC. | | | | | | |

The second one, shows the same estimator for different activities:

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Estimador Mensual de Actividad Económica. Números índice, base 2004=100 y variaciones porcentuales | | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | | | |
| 3 | Período | A - Agricultura, ganadería, caza y silvicultura | B - Pesca | C - Explotación de minas y canteras | D - Industria manufacturera | E - Electricidad, gas y agua | F - Construcción | G - Comercio mayorista, minorista y reparaciones | H - Hoteles y restaurantes | I - Transporte y comunicaciones | J - Intermediación financiera | K - Actividades inmobiliarias, empresariales y de alquiler | L - Administración pública y defensa; planes de seguridad social de afiliación obligatoria | M - Enseñanza | N - Servicios sociales y de salud | O - Otras actividades de servicios comunitarios, sociales y personales |
| 4 | | | | | | | | | | | | | | | | |
| 206 | Mayo | 224,3 | 130,9 | 91,5 | 135,4 | 142,0 | 155,3 | 160,4 | 153,3 | 201,5 | 197,5 | 146,9 | 153,8 | 159,9 | 193,4 | 150,3 |
| 207 | Junio | 155,6 | 197,5 | 87,2 | 124,2 | 156,5 | 149,0 | 148,5 | 149,5 | 192,6 | 189,9 | 146,8 | 154,3 | 161,2 | 198,4 | 151,4 |
| 208 | Julio | 104,9 | 265,8 | 89,0 | 127,3 | 155,8 | 149,5 | 144,5 | 168,0 | 195,1 | 193,8 | 144,2 | 153,5 | 162,0 | 182,4 | 170,2 |
| 209 | Agosto | 72,9 | 292,8 | 88,0 | 132,9 | 149,1 | 161,7 | 154,8 | 165,4 | 192,4 | 200,2 | 144,5 | 152,7 | 162,2 | 181,4 | 150,6 |
| 210 | Septiembre | 64,5 | 291,2 | 86,5 | 123,2 | 132,8 | 149,2 | 135,2 | 160,5 | 185,7 | 189,8 | 139,8 | 152,8 | 162,6 | 177,8 | 144,6 |
| 211 | Octubre | 75,6 | 304,9 | 88,6 | 131,2 | 133,6 | 146,3 | 149,5 | 171,8 | 189,6 | 185,6 | 143,4 | 153,4 | 163,1 | 173,4 | 149,8 |
| 212 | Noviembre | 91,1 | 121,7 | 85,0 | 124,8 | 131,2 | 141,1 | 150,1 | 169,1 | 186,4 | 177,2 | 142,1 | 153,9 | 163,5 | 173,2 | 145,0 |
| 213 | Diciembre | 103,4 | 84,5 | 87,8 | 111,8 | 139,9 | 129,0 | 129,0 | 174,1 | 184,7 | 188,3 | 144,4 | 153,9 | 163,0 | 165,6 | 142,8 |
| 214 | | | | | | | | | | | | | | | | |
| 215 | **2019** | | | | | | | | | | | | | | | |
| 216 | Enero | 73,6 | 150,2 | 87,5 | 104,9 | 146,5 | 145,8 | 125,4 | 173,6 | 188,0 | 179,0 | 136,1 | 153,6 | 156,3 | 181,1 | 155,8 |
| 217 | Febrero | 77,2 | 164,5 | 81,7 | 104,1 | 138,7 | 150,3 | 131,5 | 164,3 | 180,5 | 169,5 | 137,5 | 153,9 | 156,9 | 171,7 | 147,5 |
| 218 | Marzo | 156,5 | 153,4 | 88,4 | 114,6 | 135,7 | 149,5 | 149,0 | 160,9 | 193,1 | 172,4 | 138,8 | 153,8 | 159,6 | 186,1 | 144,4 |
| 219 | Abril | 223,5 | 152,7 | 87,4 | 119,7 | 130,1 | 142,7 | 138,2 | 156,5 | 194,7 | 171,7 | 141,4 | 153,8 | 159,9 | 190,1 | 144,2 |
| 220 | Mayo | 329,0 | 91,5 | 93,0 | 126,5 | 139,4 | 150,7 | 144,7 | 153,4 | 205,5 | 168,6 | 143,9 | 154,5 | 161,4 | 193,7 | 147,0 |
| 221 | Junio | 220,4 | 180,9 | 88,4 | 115,9 | 138,4 | 138,6 | 137,1 | 152,4 | 196,6 | 161,7 | 143,5 | 155,0 | 162,0 | 198,7 | 150,9 |
| 222 | Julio | 126,2 | 252,5 | 91,6 | 124,5 | 151,8 | 149,1 | 141,4 | 174,5 | 197,5 | 168,2 | 145,5 | 153,8 | 163,8 | 183,0 | 168,6 |
| 223 | Agosto | 78,3 | 290,5 | 91,4 | 123,9 | 143,3 | 157,1 | 140,4 | 167,1 | 191,3 | 170,1 | 144,0 | 153,7 | 163,5 | 182,0 | 146,1 |
| 224 | Septiembre | 71,3 | 163,3 | 88,3 | 116,7 | 132,2 | 142,3 | 127,9 | 162,5 | 187,5 | 162,3 | 142,4 | 154,0 | 164,1 | 178,3 | 140,9 |
| 225 | Octubre | 77,6 | 220,7 | 90,9 | 128,5 | 137,0 | 134,1 | 146,6 | 173,5 | 189,2 | 168,9 | 144,5 | 154,6 | 164,5 | 173,8 | 146,2 |
| 226 | Noviembre | 92,6 | 94,8 | 86,6 | 118,7 | 136,6 | 130,0 | 140,4 | 170,3 | 186,2 | 162,9 | 142,8 | 154,8 | 164,7 | 173,8 | 144,2 |
| 227 | Diciembre | 103,0 | 95,8 | 88,1 | 112,8 | 145,5 | 117,9 | 130,5 | 176,4 | 185,4 | 172,0 | 146,3 | 155,4 | 164,4 | 167,0 | 142,8 |
| 228 | | | | | | | | | | | | | | | | |
| 229 | **2020** | | | | | | | | | | | | | | | |
| 230 | Enero | 68,0 | 89,7 | 87,7 | 103,2 | 152,3 | 125,4 | 126,1 | 174,1 | 186,8 | 165,2 | 137,2 | 154,2 | 158,1 | 178,5 | 155,7 |
| 231 | Febrero | 75,0 | 168,1 | 85,0 | 102,5 | 140,2 | 119,7 | 129,7 | 167,9 | 179,1 | 157,2 | 136,5 | 154,5 | 158,8 | 169,2 | 147,7 |
| 232 | Marzo | 143,8 | 78,9 | 85,7 | 96,8 | 144,8 | 80,1 | 132,2 | 111,4 | 164,4 | 165,3 | 128,7 | 152,5 | 157,7 | 168,8 | 126,4 |
| 233 | | | | | | | | | | | | | | | | |
| 234 | **Fuente:** INDEC. | | | | | | | | | | | | | | | |
| 235 | | | | | | | | | | | | | | | | |

This two datasets will help us to compare the estimator at the first month of quarantine against the estimator at the same month of previous years, viewing which activities were more affected by quarantine.

## Industrial

This datasets shows the IPI (Industrial Production Index) of Argentina referred to the last 4 years, and it's also divided by activities:

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Cuadro 3. IPI manufacturero nivel general, divisiones y subclases. Serie original, base 2004=100, en variación porcentual interanual. Años 2016-2020 | | | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | | | | | | |
| 3 | Código (CLaNAE 2004) | | | **Nivel general** | **15** | 15111 | 15112 | 15113 | 15130 | 15140 | 15200 | 15311/12/13/20 | 15411/12/41/42 | 15420/30 | 15491/2/3 | 15510/29/30/41/42/49 | 15521 | 15120/ |
| 4 | Período | | | **IPI Manufacturero** | **Alimentos y bebidas** | Carne vacuna | Carne aviar | Fiambres y embutidos | Preparación de frutas, hortalizas y legumbres | Molienda de oleaginosas | Productos lácteos | Molienda de cereales | Galletitas, productos de panadería y pastas | Azúcar, productos de confitería y chocolate | Yerba mate, té y café | Gaseosas, aguas, sodas, cervezas, jugos para diluir, sidras y bebidas espirituosas | Vino | Ot prod alime |
| 5 | | | | % | | | | | | | | % | | | | | | |
| 6 | 2016 | Enero | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 7 | | Febrero | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 8 | | Marzo | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 9 | | Abril | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 10 | | Mayo | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 11 | | Junio | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 12 | | Julio | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 13 | | Agosto | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 14 | | Septiembre | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 15 | | Octubre | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 16 | | Noviembre | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 17 | | Diciembre | | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// | /// |
| 18 | 2017* | Enero | | **-1,1** | **0,6** | 8,9 | 13,1 | 1,9 | 6,9 | -20,8 | 8,0 | 8,6 | 4,8 | -2,5 | 2,4 | -3,2 | -4,1 | |
| 19 | | Febrero | | **-8,0** | **-8,1** | -2,7 | 7,8 | -6,9 | 2,2 | -18,7 | -4,9 | -9,7 | -5,6 | 2,3 | 0,4 | -8,5 | -23,2 | |
| 20 | | Marzo | | **-0,8** | **1,7** | 9,7 | 11,6 | 6,0 | 3,2 | 3,2 | 3,5 | -1,8 | -0,9 | 12,2 | 5,8 | 2,6 | -11,6 | |
| 21 | | Abril | | **-3,8** | **-3,0** | 1,9 | 2,4 | -4,5 | 0,1 | 7,2 | 9,3 | -12,3 | -7,0 | 0,2 | -13,3 | -5,5 | -18,1 | |
| 22 | | Mayo | | **3,9** | **4,1** | 7,9 | 6,3 | 7,2 | -10,0 | 5,7 | 28,8 | -8,1 | 1,8 | 7,9 | 9,9 | 2,3 | -6,4 | |
| 23 | | Junio | | **7,6** | **3,9** | 12,0 | 3,3 | 5,8 | -14,3 | -6,7 | 11,9 | -1,3 | 0,2 | 27,6 | 4,0 | 8,6 | 9,3 | |
| 24 | | Julio | | **6,9** | **4,8** | 11,9 | 5,8 | 5,0 | 2,2 | 9,9 | -2,7 | -5,7 | 4,7 | 19,1 | 2,7 | 0,3 | 1,7 | |
| 25 | | Agosto | | **6,9** | **1,1** | 6,1 | -1,6 | 2,3 | 26,0 | 2,6 | -5,5 | -10,9 | -3,4 | 3,2 | -11,4 | 8,2 | -13,4 | |
| 26 | | Septiembre | | **3,7** | **-2,0** | 7,8 | -6,8 | 2,2 | 4,3 | -6,2 | -7,0 | -9,6 | -5,9 | -6,9 | -8,8 | 4,2 | -14,1 | |

It will help us to evaluate how much industries were affected on the first month of quarantine.

### Transport

This dataset shows the total amount of tickets by month and by train station for every Train Line on the region of Buenos Aires. This information will be used to compare the first month of quarantine with same month of the previous year and see how much were railway mobility reduced:

In [4]:
```
url = 'https://servicios.transporte.gob.ar/gobierno_abierto/descargar.php?t=trenes&d=pasajero
trenes_pasajeros_df = pd.read_csv(url)
trenes_pasajeros_df.head(10)
```

Out[4]:

| | mes;linea;estacion;cantidad |
|---|---|
| 0 | 12/2019;belgranonorte;"Boulogne Sur";189118 |
| 1 | 12/2019;belgranonorte;Carapachay;44049 |
| 2 | 12/2019;belgranonorte;"Del Viso";76691 |
| 3 | 12/2019;belgranonorte;"Don Torcuato";128986 |
| 4 | 12/2019;belgranonorte;Florida;37226 |
| 5 | 12/2019;belgranonorte;"Grand Bourg";220363 |
| 6 | 12/2019;belgranonorte;"Ing. P. Nogués";103460 |
| 7 | 12/2019;belgranonorte;"Los Polvorines";137590 |
| 8 | 12/2019;belgranonorte;Munro;98282 |
| 9 | 12/2019;belgranonorte;"R. S. Ortíz";29953 |

# Methodology and Results

In this section, exploratory data analysis will be explained.

## 1. Covid-19 Cases on Argentina

First, analysis of Covid-19 cases in Argentina will be discussed.

### Data Exploration and Cleaning

In previous section, Covid-19 dataset was introduced, and every column in it was briefly explained. "**CLASIFICACION**" column refers to the manual classification of every case in the dataset. The first step to clean dataset is to select only cases which match the classification of interest for this study, those are:

* Active cases

* Non-active cases

* Death cases

Having this in mind, DataFrame can be cleaned in order to drop every case who's classification doesn't match the required ones (suspect cases, invalid cases, etc.):

```
#List of categories out of interest for this study:
removeList = ['Caso Descartado',\
                  'Caso Invalidado Epidemiologicamente',\
                  'Otro diagnostico',\
                  'Caso sospechoso - Con muestra sin resultado',\
                  'Caso Sospechoso - Sin muestra',\
                  'Caso Sospechoso - Muestra no apta',\
                  'Sin clasificar']
#Indexes of rows to remove:
ind = cases[cases.CLASIFICACION.isin(removeList)].index
#Drop rows:
casesClean = cases.drop(ind)
#Verify:
casesClean['CLASIFICACION'].value_counts()
```

```
Caso confirmado - Activo                                29773
Caso confirmado - No activo (por tiempo de evolución)   26462
Caso confirmado - Activo Internado                      16718
Caso confirmado - No Activo por criterio de laboratorio 12370
Caso confirmado - Fallecido                              1694
Name: CLASIFICACION, dtype: int64
```

**Total of confirmed cases is equal to:**

Caso confirmado - Activo +

Caso confirmado - Activo Internado +

Caso confirmado - No Activo por criterio de laboratorio +

Caso confirmado - No activo (por tiempo de evolución) +

Caso confirmado - Fallecido

After that, DataFrame is clean and ready to calculate:

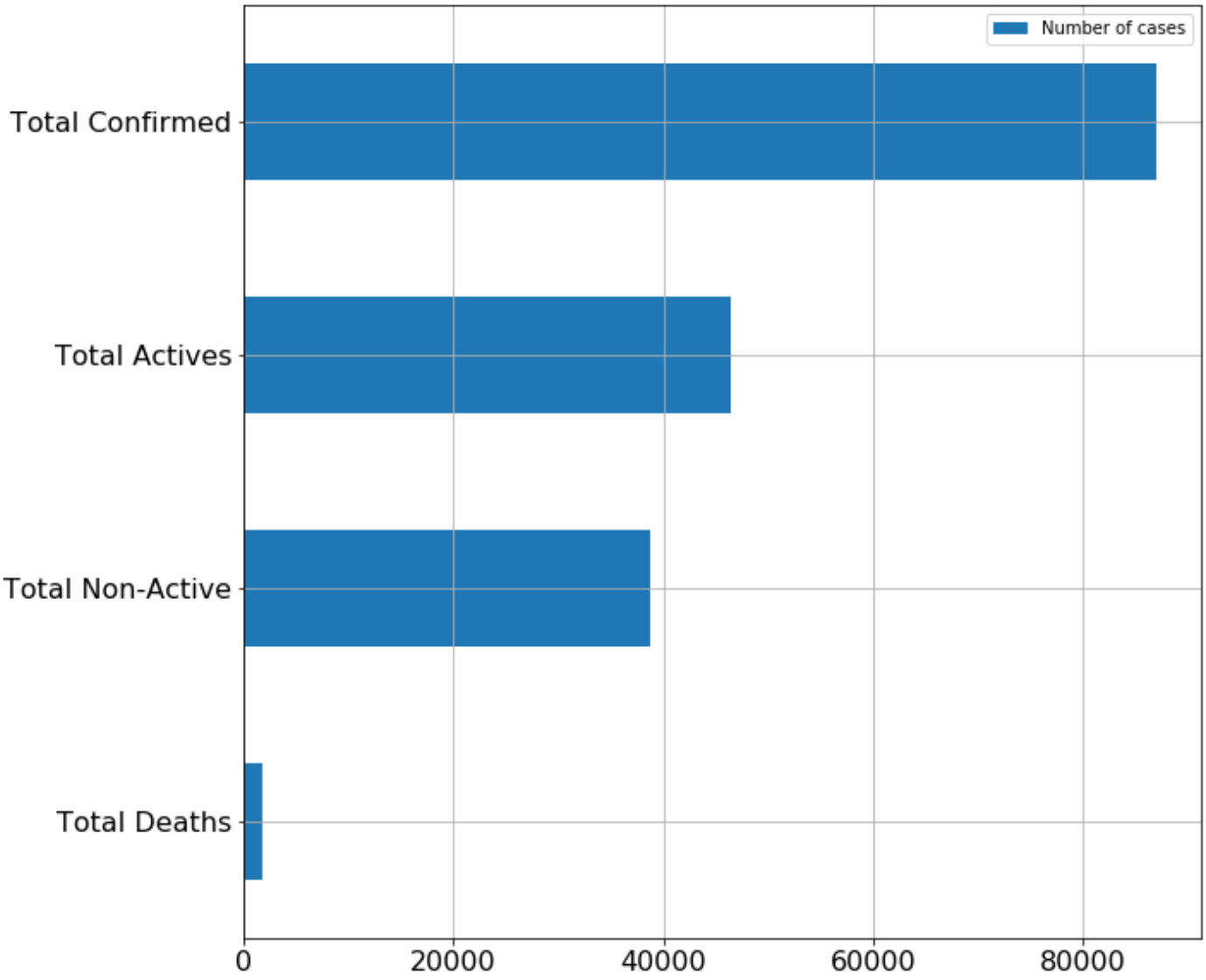
* Total number of confirmed cases

* Total active cases

* Total number of non-active

* Total number of deaths

And, with this values, a Bar Plot can be made to visualize number of cases for each category:




## Distribution of confirmed cases by age and gender

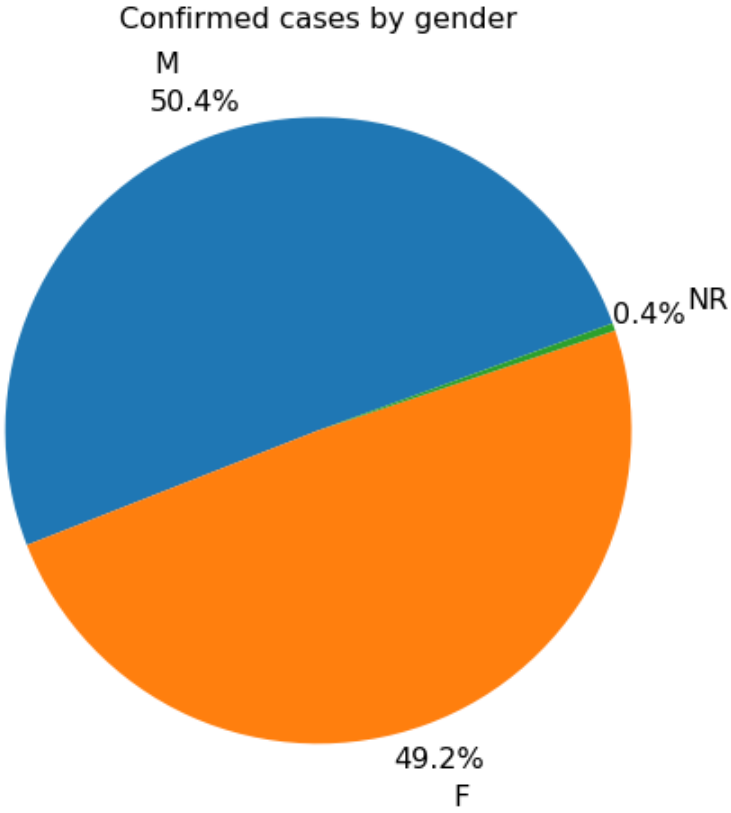
To make this analysis, it’s necessary first to remove rows with NaN values on 'CLASIFICACION' column and sort by diagnosis date:

**Remove rows with NaN values on 'CLASIFICACION' column and sort by diagnosis date:**

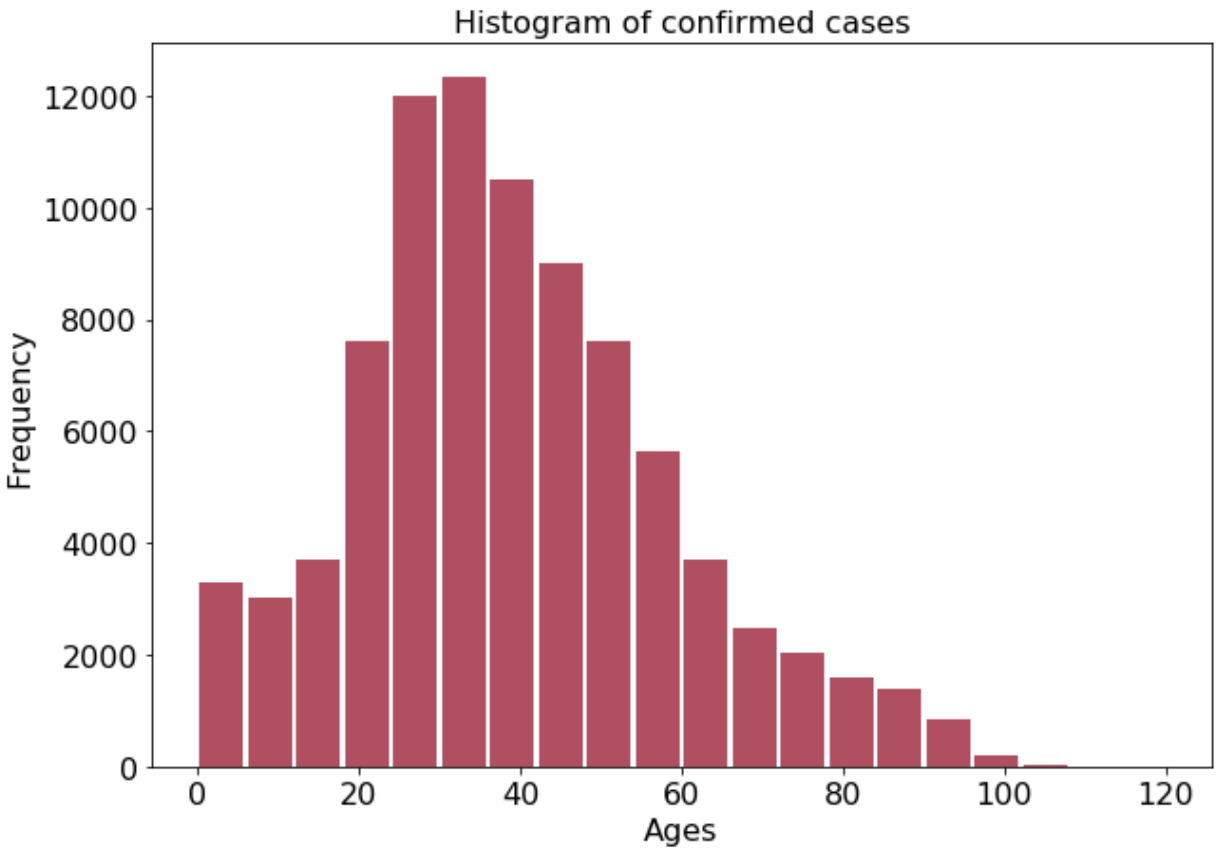
```
ind = []
for i in casesClean.index:
    if str(casesClean['CLASIFICACION'][i]) == 'nan':
        ind.append(i)
    #end if
#end for

confirmed_df = casesClean.drop(ind)
confirmed_df.sort_values('fecha_diagnostico', ascending=True,inplace=True)
print(confirmed_df.shape)
confirmed_df['fecha_diagnostico'].head()
```

```
(87017, 25)

42    2020-03-03
51    2020-03-06
59    2020-03-06
74    2020-03-06
87    2020-03-06
Name: fecha_diagnostico, dtype: object
```

After that, Pie chart and histogram of confirmed cases can be made:

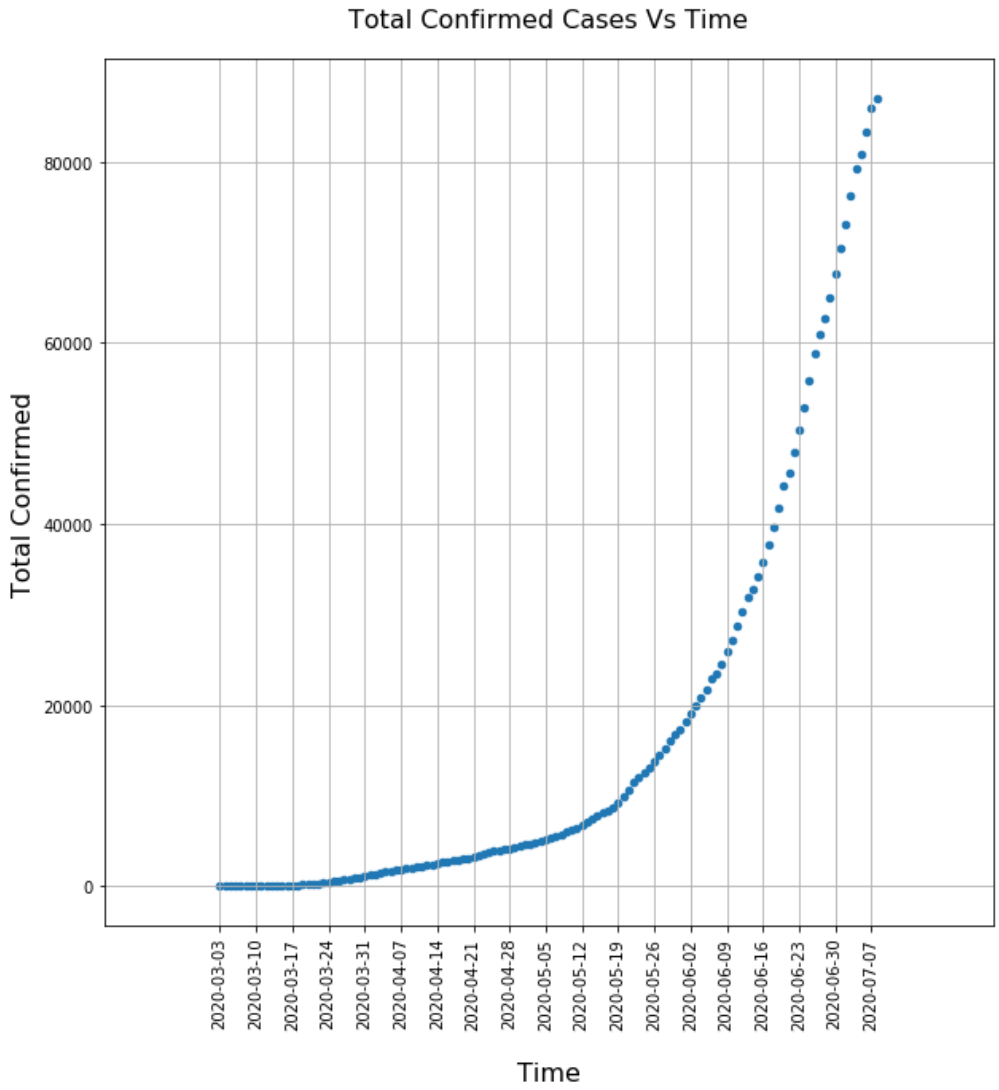



To make the histogram, all ages must be converted to “years”. For example, an age of 6 months must be converted to 0.5 years. Then, histogram can be made:



From the histogram above, it can be seen that people between 20 and 60 are more likely to get infected, regardless of its gender. Moreover, this range of ages corresponds to people's "working age".

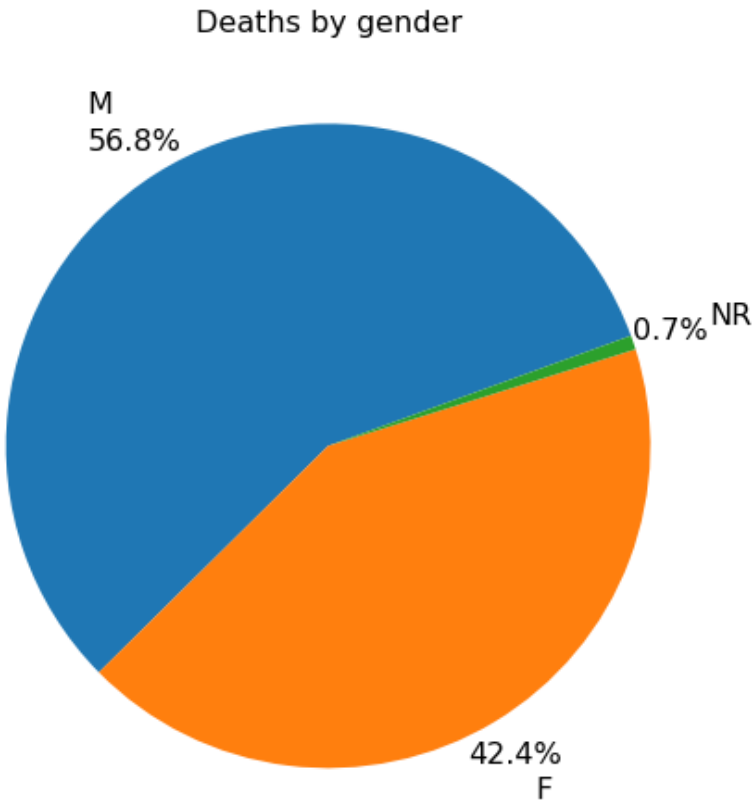
Following, a graph shows evolution of confirmed cases (i.e. active + non-active + deaths):



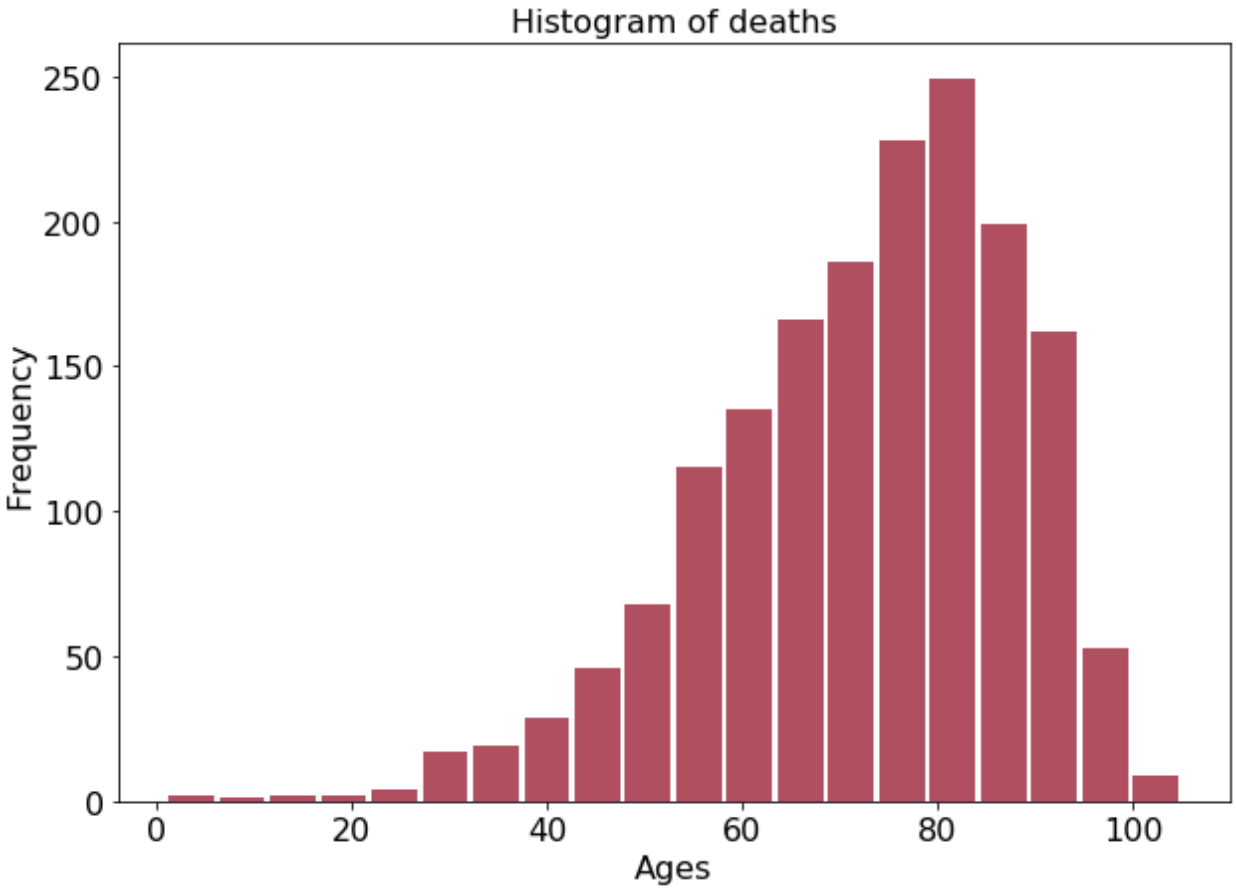

The figure above shows the known exponential curve, which represents growth of cases over time.

## Distribution of deaths by age and gender

Similar to previous analysis, here rows with 'NaN' values in column '**fecha_fallecimiento**' are dropped from the DataFrame. After that, Pie chart and histogram can be made:

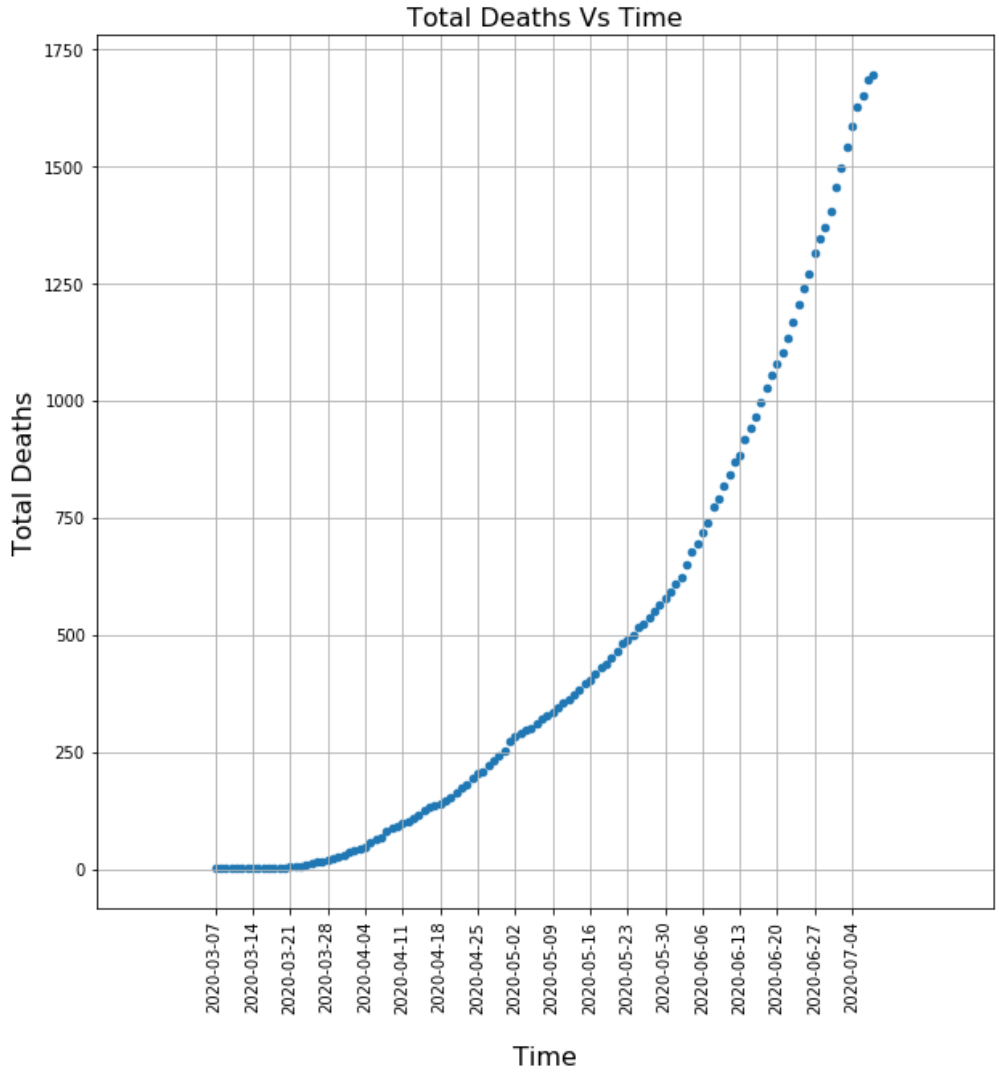



To make the
histogram, all ages must be converted to “years”. For example, an age of 6 months must be converted to 0.5 years. Then, histogram can be made:



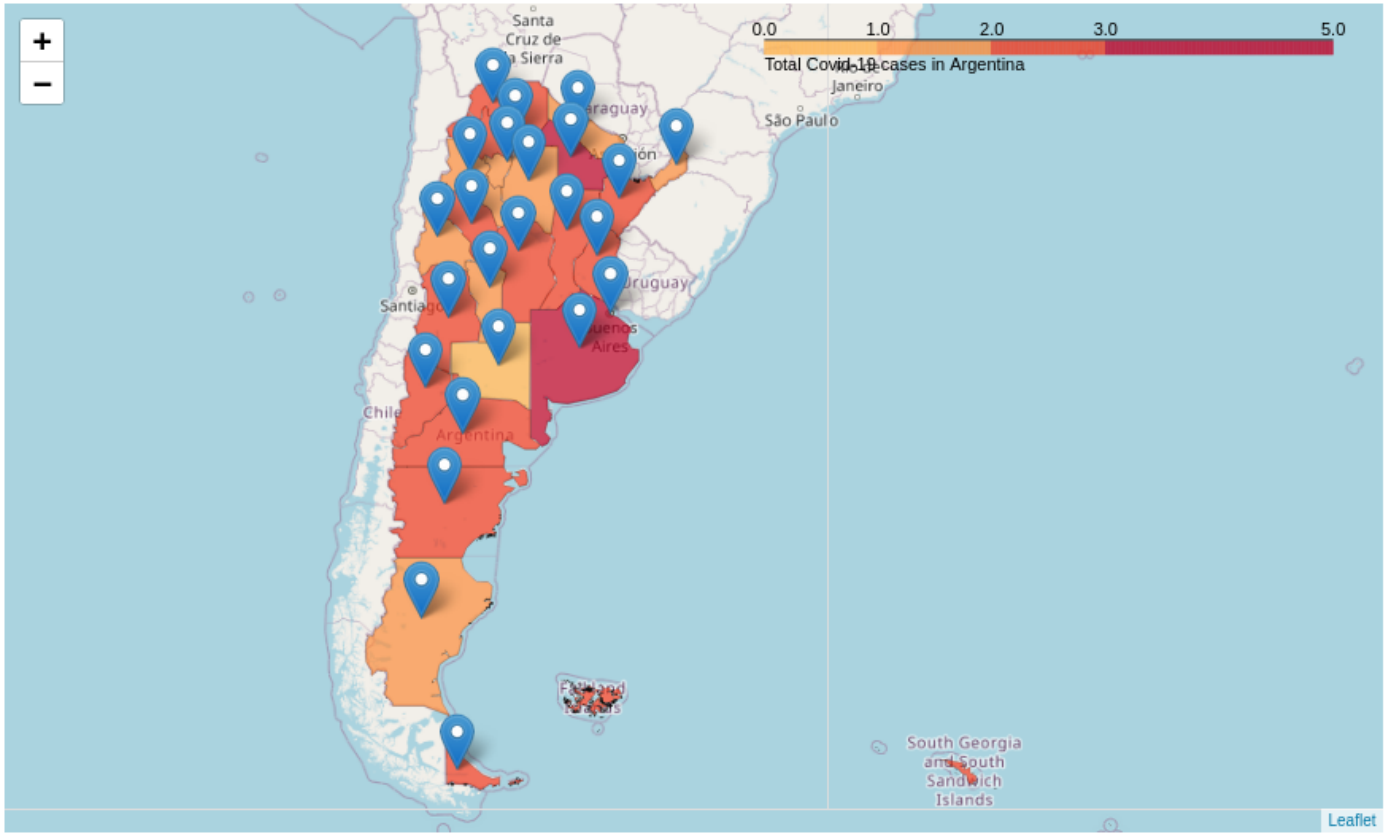
From the above Pie chart, we can see that infected men are more likely to die than other genders. Also, ages between 60 and 90 has higher death rates.

The following curve shows evolution of deaths over time:




## Distribution of cases in the country

In order to analyze distribution of cases on the territory, DataFrame must be grouped by Provinces, more specifically, by "**carga_provincia_nombre**" column. After that, "**TotalCases**", "**Latitude**" and "**Longitude**" columns are added to DataFrame (last two columns are included using **Nominatim** from **geopy.geocoders**). With this done, an interactive Choropleth Map of Argentina can be made using **folium**:




It can be seen from the map above, that Buenos Aires, CABA and Chaco are the critical zones in number of cases

## Conclusion

Here, Covid-19 Dataset was explored in terms of: number of cases and deaths, and its distribution in terms of age, gender and territory.

The next steps in this study is to see how much the economy, industry and mobility were affected by Covid-19 Pandemic and Quarantine.

# 2. Analyzing Economic Activity

In this section, the Economic Activity Monthly Estimation (EMAE in spanish) will be analyzed:

1. First, we'll plot the evolution of EMAE over time, from 2004 to 2020. This will show that Argentina is currently living an 'outlier' period in its Economy. To complete the analysis, Dataframe called “mensual_df” will be used. This is a dataset with general level EMAE (i.e. not discriminated by activity).

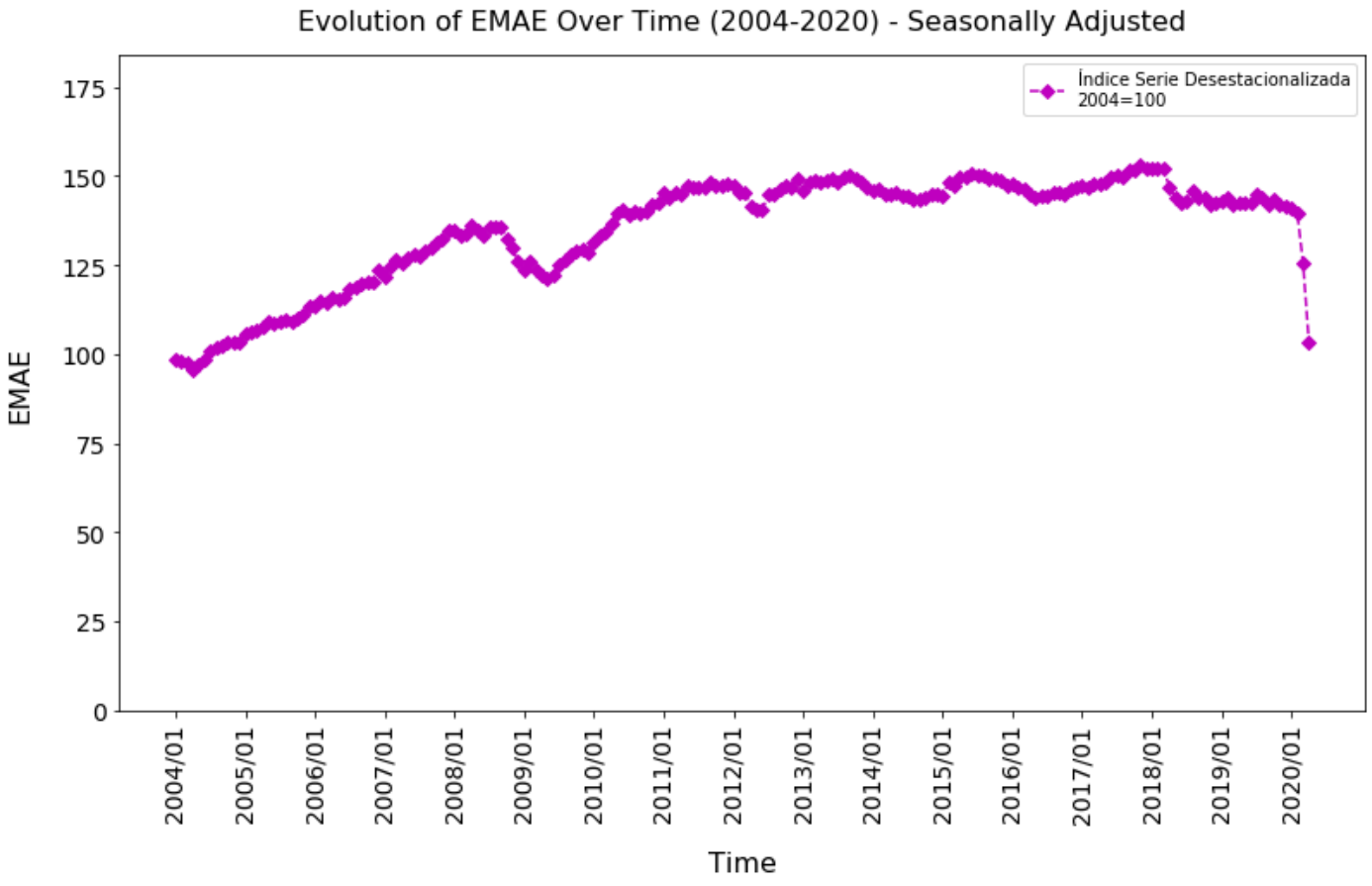
2. Next, EMAE variation relative to same month of previous year will be analyzed for every activity in the dataset. This will allow us to determine which economic activities were most affected by Pandemic and Quarantine. In this case, DataFrame called “actividad_var_df” (a dataframe with EMAE discriminated by activity) will be helpful to compare economic variation among activities.

## Graph evolution of EMAE over time

First, column "**PERÍODO**" will be converted to format 'YYYY/MM'. Then, "**Índice Serie Desestacionalizada\n2004=100**" will be used to represent evolution of EMAE over time. The reason to use this column is because it represents the seasonally adjusted index, which gives a "softy curve" of EMAE over time:

| | PERÍODO | Índice Serie Original\n2004=100 | Var % respecto a igual período del año anterior | Índice Serie Desestacionalizada\n2004=100 | Var % respecto al mes anterior | Índice Serie Tendencia-Ciclo\n2004=100 | Var % respecto al mes anterior.1 |
|---|---|---|---|---|---|---|---|
| 0 | 2004/01 | 92.627506 | NaN | 98.317580 | NaN | 96.003297 | NaN |
| 1 | 2004/02 | 90.186179 | NaN | 98.222593 | -0.096613 | 96.612123 | 0.634171 |
| 2 | 2004/03 | 101.883298 | NaN | 97.557758 | -0.676865 | 97.286335 | 0.697855 |
| 3 | 2004/04 | 102.567430 | NaN | 95.391911 | -2.220067 | 97.989131 | 0.722400 |
| 4 | 2004/05 | 109.877504 | NaN | 96.847940 | 1.526365 | 98.719394 | 0.745249 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 191 | 2019/12 | 135.443184 | -0.185625 | 141.758949 | -0.067824 | 141.139168 | -0.283129 |
| 192 | 2020/01 | 131.930264 | -1.928503 | 140.882115 | -0.618539 | 140.705711 | -0.307113 |
| 193 | 2020/02 | 129.376635 | -2.414515 | 139.802665 | -0.766208 | 140.263195 | -0.314498 |
| 194 | 2020/03 | 127.919209 | -11.432783 | 125.597351 | -10.160975 | 139.839055 | -0.302389 |
| 195 | 2020/04 | 110.302177 | -26.385864 | 103.571995 | -17.536482 | 139.427735 | -0.294138 |

196 rows × 7 columns



The figure above shows a steep drop in the EMAE. This could be the worst drop in Argentina's Economy in the last 15 years. To complete this exploratory analysis, a bar graph of economic decline in several activities will be made in the following section.

## List most affected activities

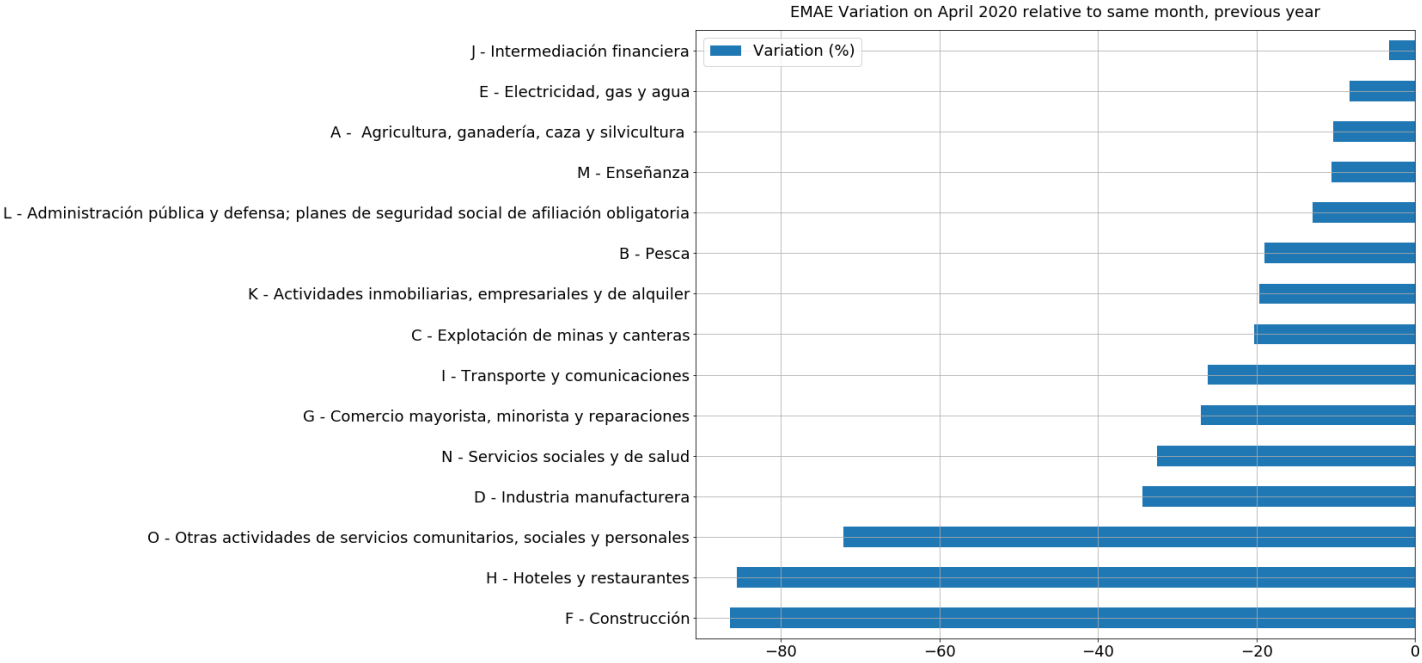
To make a list with most affected activities, “actividad_var_df” is used:

| | PERÍODO | A - Agricultura, ganadería, caza y silvicultura | B - Pesca | C - Explotación de minas y canteras | D - Industria manufacturera | E - Electricidad, gas y agua | F - Construcción | G - Comercio mayorista, minorista y reparaciones | H - Hoteles y restaurantes | I - Transporte y comunicaciones | J - Intermediación financiera | i e |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2005 | - | - | - | - | - | - | - | - | - | - | |
| 1 | Enero | -0.454394 | -15.7663 | 0.489363 | 6.38979 | 9.15374 | 9.7761 | 8.23983 | 9.28139 | 12.3031 | 1.14996 | |
| 2 | Febrero | 6.59691 | 42.4717 | -2.29607 | 5.90775 | 6.78087 | 10.2792 | 9.04294 | 8.94971 | 12.4247 | 5.82669 | |
| 3 | Marzo | 25.2082 | -15.7381 | 1.33685 | 6.05691 | 3.9884 | 10.77 | 8.42798 | 9.60376 | 14.4171 | 7.99795 | |
| 4 | Abril | 39.2054 | -20.3986 | 2.5106 | 9.25904 | 6.39246 | 13.5384 | 13.1805 | 10.1362 | 15.5631 | 11.8796 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 195 | 2020 | - | - | - | - | - | - | - | - | - | - | |
| 196 | Enero | -7.7782 | -41.6715 | 0.205644 | -1.68481 | 3.97768 | -8.4772 | -1.31315 | 0.601747 | -0.620573 | -7.29364 | |
| 197 | Febrero | -2.43034 | -2.22622 | 4.07081 | -1.48839 | 1.30627 | -14.2851 | -3.33069 | 2.71121 | -0.995624 | -6.71821 | |
| 198 | Marzo | -7.38471 | -49.6344 | -2.95643 | -15.4137 | 6.14244 | -39.224 | -13.4698 | -35.1168 | -14.2538 | -3.47938 | |
| 199 | Abril | -10.2773 | -18.9771 | -20.3506 | -34.3732 | -8.27272 | -86.3952 | -27.0362 | -85.551 | -26.1395 | -3.24854 | |

200 rows × 17 columns

To make the list, row 199 is selected (April 2020):

| | Activity | Variation (%) |
|---|---|---|
| 0 | F - Construcción | -86.3952 |
| 1 | H - Hoteles y restaurantes | -85.551 |
| 2 | O - Otras actividades de servicios comunitario... | -72.1068 |
| 3 | D - Industria manufacturera | -34.3732 |
| 4 | N - Servicios sociales y de salud | -32.581 |
| 5 | G - Comercio mayorista, minorista y reparaciones | -27.0362 |
| 6 | I - Transporte y comunicaciones | -26.1395 |
| 7 | C - Explotación de minas y canteras | -20.3506 |
| 8 | K - Actividades inmobiliarias, empresariales y... | -19.6902 |
| 9 | B - Pesca | -18.9771 |
| 10 | L - Administración pública y defensa; planes d... | -12.9385 |
| 11 | M - Enseñanza | -10.5836 |
| 12 | A - Agricultura, ganadería, caza y silvicultura | -10.2773 |
| 13 | E - Electricidad, gas y agua | -8.27272 |
| 14 | J - Intermediación financiera | -3.24854 |




The figure above shows how much activities were affected by Pandemic. Construction, Hotels, Restaurants and other activities of community and social services were the most affected ones. Financial, Energy, Agriculture and Livestock were the less affected ones.

In the following section, activities related to Industry (which has a variation lower than -30% as can be seen in the figure above) will be analyzed, compared to each other and clustered to identify similarities between them.

## 3. Analyzing Industrial Activities

The Industrial Production Index (IPI) includes an exhaustive survey of all the economic activities that make up the manufacturing industry sector, with coverage for the entire country.

This indicator measures the evolution of the industrial activity on a monthly basis, and is calculated by production variables such like physical units, sales, utilization of resources, worked hours, among others.

As a whole, all the selected variables provide monthly data on more than 5,000 industrial containers.

### Data Analysis

In this section, Industrial Activity on Argentina will be analyzed based on a dataset called “sh_ipi_manufacturero_2020.xls”. This file is composed of 5 sheets. First sheet is called

“Índice”, and it’s only an index with a brief description of the other 4 sheets. The remaining sheets are called “Cuadro 1”, “Cuadro 2”, “Cuadro 3” and “Cuadro 4”.

Here, sheets 'Cuadro 2' and 'Cuadro 3' will be under analysis. 'Cuadro 2' represents Argentinian Industrial Production Index (IPI) in absolute values for each industrial activity and in general level. 'Cuadro 3' represents Argentinian IPI's variation percentage relative to same month, previous year; again, specified by activity and in general level.

After cleaning, the analysis to make with this datasets can be summarized in the following steps :

1. Visualize the most affected activities

2. Show evolution of IPI for the most and less affected activities, from January 2019 to April 2020

3. Cluster activities by its IPI number and variation percentage relative to previous year

The following sections covers the steps defined here.

## 1. Find most affected activities

This analysis will be focused on DataFrame N°3 (Cuadro 3). Here, variation percentage on April 2020 will be the key value to sort DataFrame and look for the most affected activities:

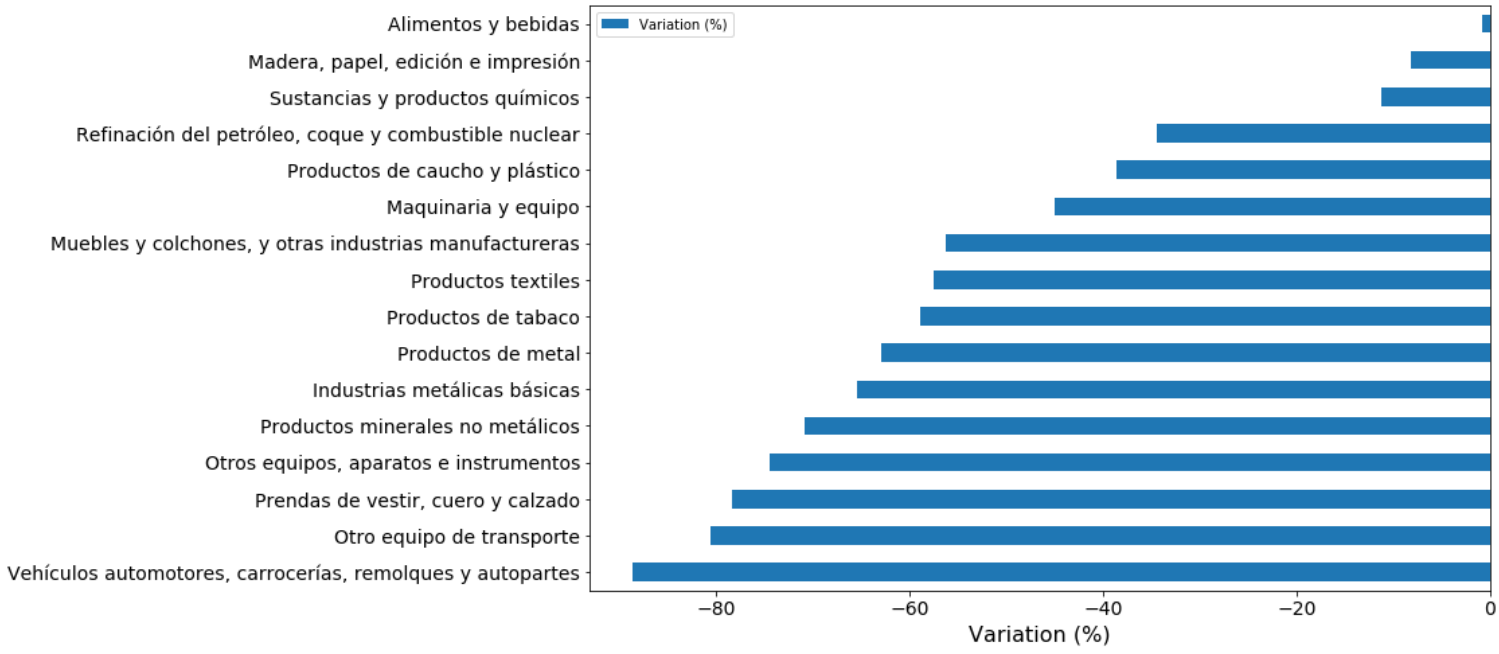
| | Período | Month | IPI Manufacturero | Alimentos y bebidas | Carne vacuna | Carne aviar | Fiambres y embutidos | Preparación de frutas, hortalizas y legumbres | Molienda de oleaginosas | Productos lácteos | ... | Vehículos automotores | Carrocerías, remolques y semirremolques | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 35 | 2018* | Diciembre | -14.8362 | -2.45715 | 3.90395 | 4.6494 | -8.57018 | 9.77583 | 11.2168 | 4.19174 | ... | -30.9664 | -41.681 | |
| 36 | 2019* | Enero | -11.1712 | -2.46003 | 0.901916 | 13.6158 | -8.55314 | 12.0778 | -5.85337 | -8.4447 | ... | -37.4431 | -34.8858 | |
| 37 | 2019* | Febrero | -8.41078 | -0.354001 | -1.25503 | 10.8443 | -3.14636 | 18.4297 | 1.49169 | -2.14087 | ... | -13.6595 | -31.1552 | |
| 38 | 2019* | Marzo | -14.111 | -8.53911 | -10.65 | -3.73095 | -17.5837 | 15.0799 | -3.54599 | -8.19182 | ... | -39.4647 | -49.3064 | |
| 39 | 2019* | Abril | -8.86641 | -1.45586 | -2.12008 | 3.18865 | -7.32355 | -2.00761 | 14.1855 | -7.07268 | ... | -31.8544 | -37.0844 | |
| 40 | 2019* | Mayo | -6.91018 | -1.1757 | 3.36037 | 6.82774 | -11.3249 | 10.6707 | 7.5049 | -8.37796 | ... | -31.1644 | -32.3959 | |
| 41 | 2019* | Junio | -7.16707 | 1.11942 | -2.4574 | 1.64866 | -17.6097 | 5.19717 | 27.0399 | -1.89268 | ... | -32.9751 | -27.4862 | |
| 42 | 2019* | Julio | -1.71254 | 4.4986 | 10.0052 | 10.4665 | -3.09628 | -4.03455 | 28.5631 | -1.82427 | ... | -40.7748 | -10.3312 | |
| 43 | 2019* | Agosto | -6.43121 | 0.463002 | 0.544581 | 4.03887 | -13.476 | 3.95639 | 36.5828 | 4.22243 | ... | -32.2463 | -11.7287 | |
| 44 | 2019* | Septiembre | -5.00693 | 0.906734 | 17.6279 | 12.704 | -9.38215 | -7.48254 | 1.8288 | -2.46637 | ... | -24.6368 | 15.9292 | |
| 45 | 2019* | Octubre | -1.88457 | 0.25664 | 8.58953 | 6.91917 | -6.26499 | -13.6941 | 16.2278 | -8.65371 | ... | -13.3428 | 1.87596 | |
| 46 | 2019* | Noviembre | -4.33699 | -1.61468 | 7.3273 | 2.64104 | -12.3624 | -23.3608 | -9.43199 | -8.02131 | ... | -20.966 | -8.54143 | |
| 47 | 2019* | Diciembre | 1.41533 | 7.33395 | 13.138 | 10.7799 | -2.11196 | -14.3872 | 9.88314 | -4.00827 | ... | -23.1662 | -1.84925 | |
| 48 | 2020* | Enero | -0.267145 | 4.08932 | 1.66989 | 4.88014 | -7.09479 | -18.5006 | -10.2365 | -1.58361 | ... | 38.5527 | -24.7372 | |
| 49 | 2020* | Febrero | -0.900185 | 5.44362 | 0.89587 | 4.94499 | -10.1125 | -7.21741 | 7.04061 | 0.901998 | ... | -21.8092 | 5.05799 | |
| 50 | 2020* | Marzo | -16.5347 | -1.8081 | 6.24759 | 7.55893 | -5.81431 | -20.4099 | -6.97478 | -2.10978 | ... | -35.2953 | -21.0922 | |
| 51 | 2020* | Abril | -33.4046 | -0.793792 | 9.36978 | 12.6669 | -7.46169 | -1.70751 | -6.51124 | 6.43876 | ... | -100 | -37.5458 | |

As can be seen in the picture above, Industrial Activity falls 33% respect to the same month of the previous year (column “**IPI Manufacturero**”, row 51), which is right according to bar graph introduced in previous economic analysis.

The list with activities sorted by its IPI variation is showed below:

| | Activity | Variation (%) |
|---|---|---|
| 0 | Vehículos automotores, carrocerías, remolques ... | -88.6001 |
| 1 | Otro equipo de transporte | -80.5692 |
| 2 | Prendas de vestir, cuero y calzado | -78.3671 |
| 3 | Otros equipos, aparatos e instrumentos | -74.3855 |
| 4 | Productos minerales no metálicos | -70.7479 |
| 5 | Industrias metálicas básicas | -65.3359 |
| 6 | Productos de metal | -62.8877 |
| 7 | Productos de tabaco | -58.8574 |
| 8 | Productos textiles | -57.4428 |
| 9 | Muebles y colchones, y otras industrias manufa... | -56.2169 |
| 10 | Maquinaria y equipo | -44.9894 |
| 11 | Productos de caucho y plástico | -38.5085 |
| 12 | Refinación del petróleo, coque y combustible n... | -34.4272 |
| 13 | Sustancias y productos químicos | -11.2086 |
| 14 | Madera, papel, edición e impresión | -8.20136 |
| 15 | Alimentos y bebidas | -0.793792 |

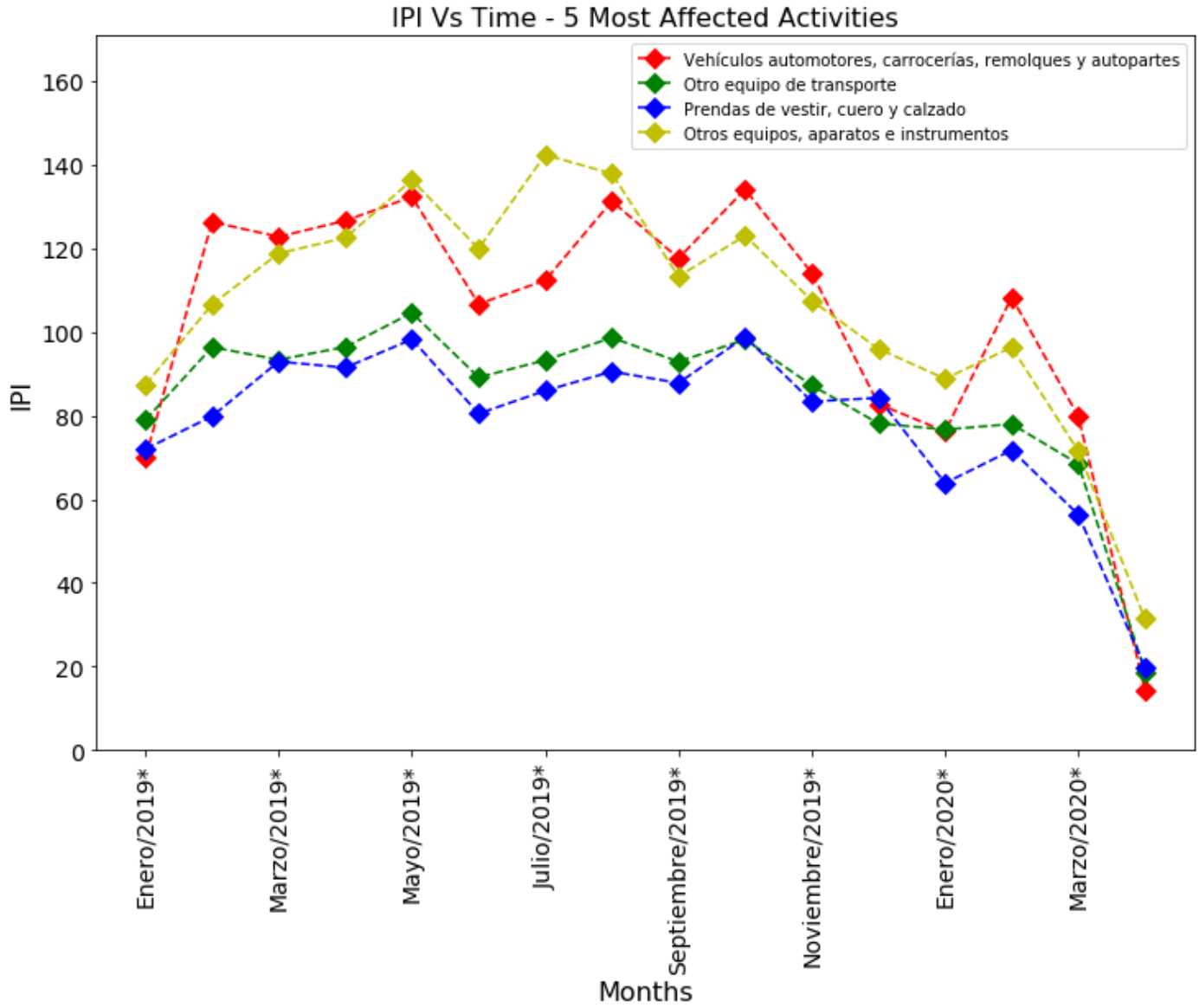
From the previous list, it can be seen that activities related to automotive and textile industries were more affected by pandemic and quarantine. The following bar graph will help to visualize this variations:



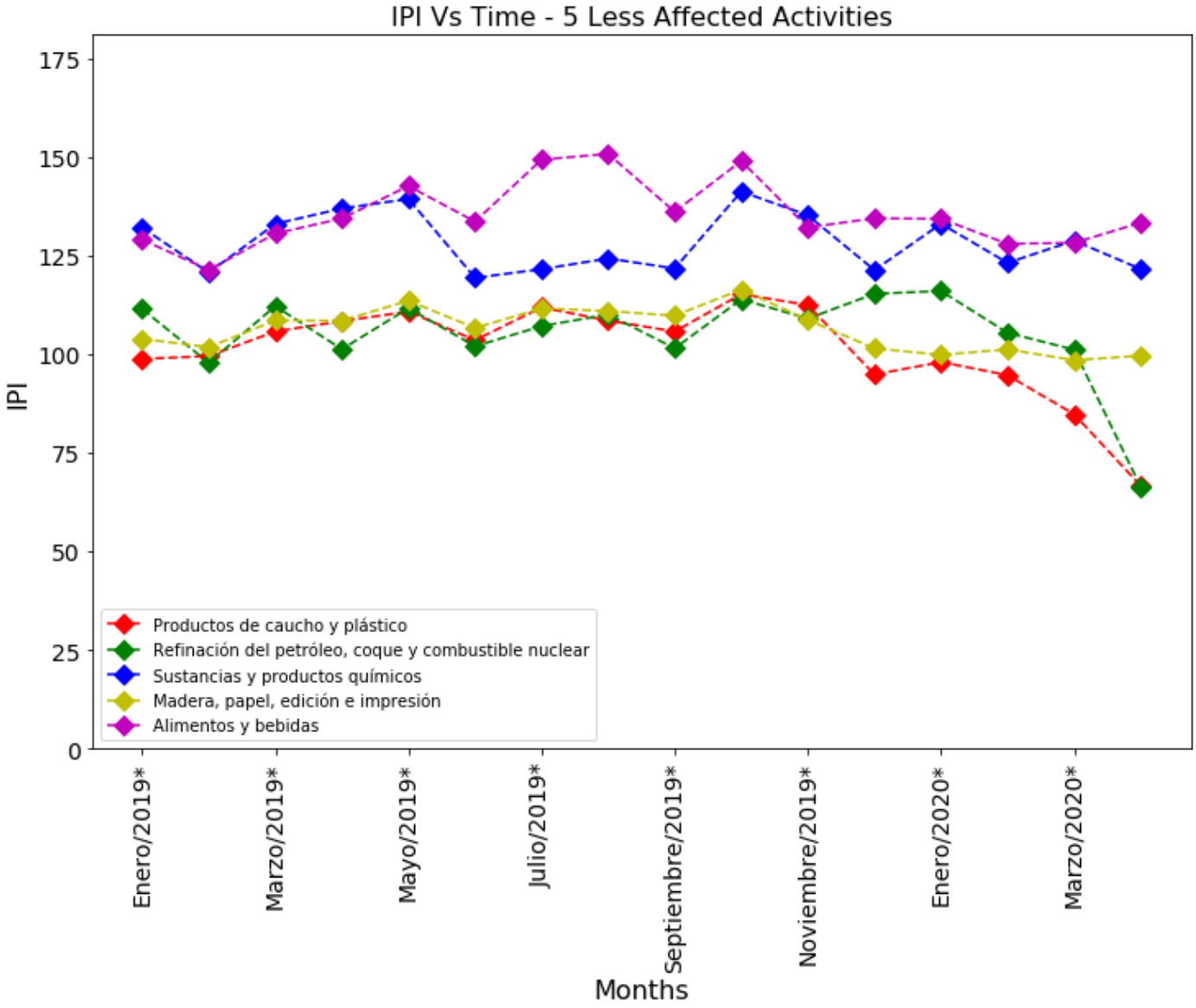
## 2. Show evolution of IPI for the most and less affected activities, from January 2019 to April 2020

With sorted DataFrame from previous analysis, here, DataFrame N°2 will be used to represent evolution of IPI over time.

First, 5 most affected activities in April 2020:




Next, 5 less affected activities in April 2020:




## 3. Cluster activities by its IPI number and variation percentage relative to previous year

To apply K Means algorithm, the following steps will be made:

1. Create a Numpy array X with IPI and Variation % values
2. Normalize the array
3. Plot samples previous to clustering to see its distribution
4. Initialize k means with number of clusters and iterations and fit
5. Plot samples and cluster's centers.

### 1. Creating a Numpy array

From DataFrames previously used, we will get IPI and Variation % values:

| | Activity | IPI |
|---|---|---|
| 0 | Alimentos y bebidas | 133.389 |
| 1 | Industrias metálicas básicas | 37.4343 |
| 2 | Madera, papel, edición e impresión | 99.6503 |
| 3 | Maquinaria y equipo | 64.9525 |
| 4 | Muebles y colchones, y otras industrias manufa... | 46.229 |
| 5 | Otro equipo de transporte | 18.7361 |
| 6 | Otros equipos, aparatos e instrumentos | 31.384 |
| 7 | Prendas de vestir, cuero y calzado | 19.7891 |
| 8 | Productos de caucho y plástico | 66.6778 |
| 9 | Productos de metal | 43.1464 |
| 10 | Productos de tabaco | 34.952 |
| 11 | Productos minerales no metálicos | 48.9175 |
| 12 | Productos textiles | 38.3518 |
| 13 | Refinación del petróleo, coque y combustible n... | 66.4255 |
| 14 | Sustancias y productos químicos | 121.645 |
| 15 | Vehículos automotores, carrocerías, remolques ... | 14.4367 |

| | Activity | Variation (%) |
|---|---|---|
| 0 | Alimentos y bebidas | -0.793792 |
| 1 | Industrias metálicas básicas | -65.3359 |
| 2 | Madera, papel, edición e impresión | -8.20136 |
| 3 | Maquinaria y equipo | -44.9894 |
| 4 | Muebles y colchones, y otras industrias manufa... | -56.2169 |
| 5 | Otro equipo de transporte | -80.5692 |
| 6 | Otros equipos, aparatos e instrumentos | -74.3855 |
| 7 | Prendas de vestir, cuero y calzado | -78.3671 |
| 8 | Productos de caucho y plástico | -38.5085 |
| 9 | Productos de metal | -62.8877 |
| 10 | Productos de tabaco | -58.8574 |
| 11 | Productos minerales no metálicos | -70.7479 |
| 12 | Productos textiles | -57.4428 |
| 13 | Refinación del petróleo, coque y combustible n... | -34.4272 |
| 14 | Sustancias y productos químicos | -11.2086 |
| 15 | Vehículos automotores, carrocerías, remolques ... | -88.6001 |

Above DataFrames columns can be merged to create a 2D numpy array:

In [103]:
```
X=np.array([April_IPI_df['IPI'].astype(float).tolist(),\
            April_var_df['Variation (%)'].astype(float).tolist()])
X=np.transpose(X)
print(X.shape)
X
```

```
(16, 2)
```

Out[103]:
```
array([[133.38888386,  -0.79379236],
       [ 37.4343231 , -65.33587512],
       [ 99.65026005,  -8.20136214],
       [ 64.95254447, -44.98937425],
       [ 46.2289712 , -56.21686492],
       [ 18.73606616, -80.56920796],
       [ 31.38404931, -74.38552421],
       [ 19.78911891, -78.36713656],
       [ 66.67783696, -38.50845446],
       [ 43.14644779, -62.88767746],
       [ 34.95204763, -58.85737786],
       [ 48.91747635, -70.74794848],
       [ 38.35175021, -57.44275975],
       [ 66.42553176, -34.42722244],
       [121.64514046, -11.20857398],
       [ 14.43669716, -88.60010522]])
```

2. Normalizing array:

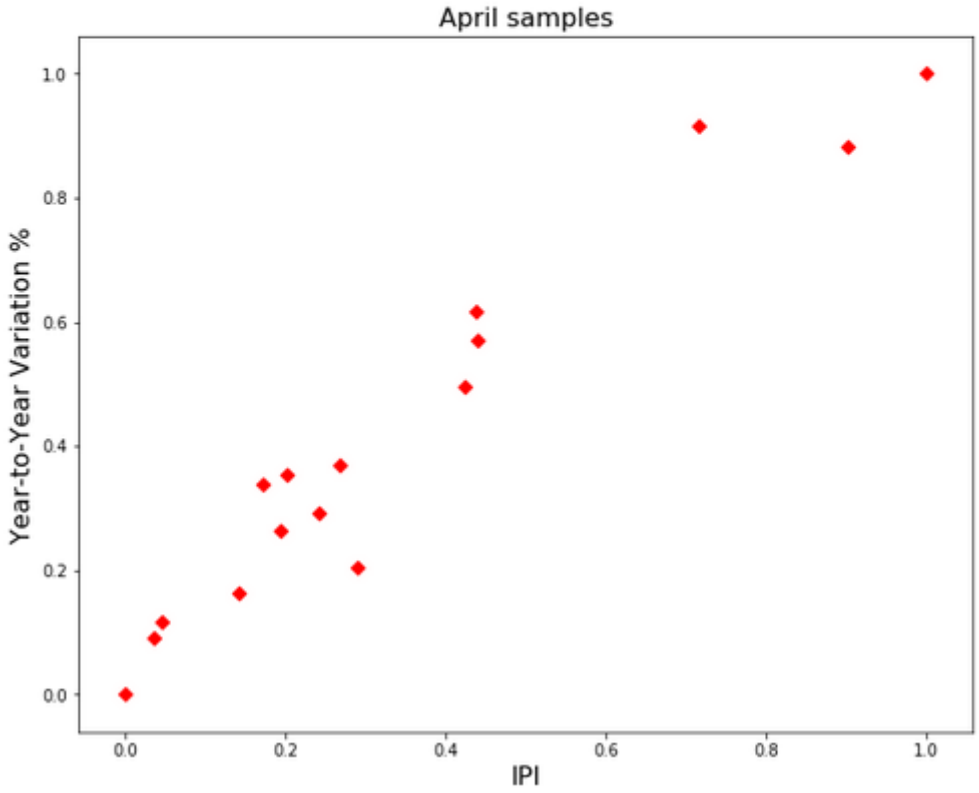
**Normalize array:**

In [106]:
```
from sklearn import preprocessing

min_max_scaler = preprocessing.MinMaxScaler()
X = min_max_scaler.fit_transform(X)
X
```

Out[106]:
```
array([[1.        , 1.        ],
       [0.19333504, 0.2649494 ],
       [0.71636819, 0.91563739],
       [0.42467355, 0.49666965],
       [0.26726935, 0.3688031 ],
       [0.03614367, 0.0914615 ],
       [0.14247197, 0.16188564],
       [0.04499641, 0.11654024],
       [0.43917763, 0.57047892],
       [0.24135538, 0.2928312 ],
       [0.1724672 , 0.33873108],
       [0.28987091, 0.20331291],
       [0.20104761, 0.35484175],
       [0.43705657, 0.61695886],
       [0.90127341, 0.88138915],
       [0.        , 0.        ]])
```

3. Plotting samples previous to clustering to see its distribution:

```
plt.figure(figsize=(10, 8))
plt.scatter(X[:,0], X[:,1], marker='D',c='r')
plt.title("April samples",fontsize=16)
plt.xlabel("IPI",fontsize=16)
plt.ylabel('Year-to-Year Variation %',fontsize=16)

plt.savefig("images/AprilSamples.png",format='png',bbox_inches='tight')
plt.show()
```




4. Initializing k means with number of clusters and iterations and fit:

```
from sklearn.cluster import KMeans
```

```
n_clusters = 3
k_means = KMeans(init="k-means++", n_clusters=n_clusters, n_init=10)
```

```
k_means.fit(X)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=3, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=None, tol=0.0001, verbose=0)
```

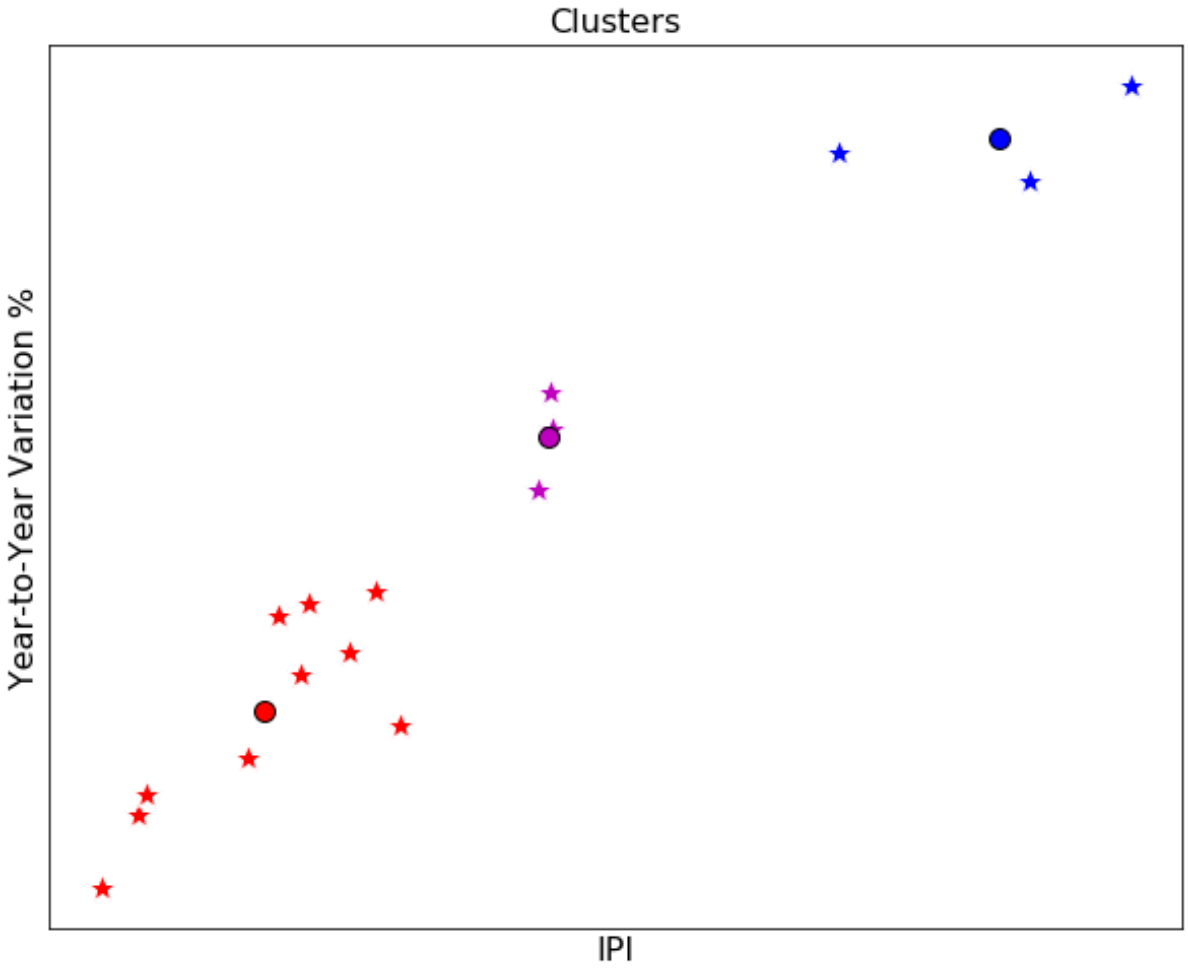
```
k_means_labels = k_means.labels_
k_means_labels
```

```
array([1, 0, 1, 2, 0, 0, 0, 0, 2, 0, 0, 0, 0, 2, 1, 0], dtype=int32)
```

```
k_means_cluster_centers = k_means.cluster_centers_
k_means_cluster_centers
```

```
array([[0.15889576, 0.21933568],
       [0.8725472 , 0.93234218],
       [0.43363592, 0.56136915]])
```

5. Plotting samples and cluster's centers:




The figure above shows 3 clusters, which in general term has the following characteristics:

a. Low IPI, High negative variation (%)

b. Medium IPI, Medium negative variation (%)

c. High IPI, Low negative variation (%)

Following image shows a list with activities and its cluster N°:

| | Activity | IPI | Variation (%) | Cluster N° |
|---|---|---|---|---|
| 1 | Industrias metálicas básicas | 37.4343 | -65.3359 | 0 |
| 4 | Muebles y colchones, y otras industrias manufa... | 46.229 | -56.2169 | 0 |
| 5 | Otro equipo de transporte | 18.7361 | -80.5692 | 0 |
| 6 | Otros equipos, aparatos e instrumentos | 31.384 | -74.3855 | 0 |
| 7 | Prendas de vestir, cuero y calzado | 19.7891 | -78.3671 | 0 |
| 9 | Productos de metal | 43.1464 | -62.8877 | 0 |
| 10 | Productos de tabaco | 34.952 | -58.8574 | 0 |
| 11 | Productos minerales no metálicos | 48.9175 | -70.7479 | 0 |
| 12 | Productos textiles | 38.3518 | -57.4428 | 0 |
| 15 | Vehículos automotores, carrocerías, remolques ... | 14.4367 | -88.6001 | 0 |
| 0 | Alimentos y bebidas | 133.389 | -0.793792 | 1 |
| 2 | Madera, papel, edición e impresión | 99.6503 | -8.20136 | 1 |
| 14 | Sustancias y productos químicos | 121.645 | -11.2086 | 1 |
| 3 | Maquinaria y equipo | 64.9525 | -44.9894 | 2 |
| 8 | Productos de caucho y plástico | 66.6778 | -38.5085 | 2 |
| 13 | Refinación del petróleo, coque y combustible n... | 66.4255 | -34.4272 | 2 |

### Conclusions

In this section, industrial dataset were explored and analyzed. A comparison between activities were made, showing which kind of industrial activities were less affected by pandemic and quarantine, and a clustering algorithm were applied to group activities with similar IPI and Variation % relative to last year. This segmentation could help governments to take decisions based on industries characteristics and its response capacity to pandemic and quarantine.

## 4. Analyzing mobility (trains) in Buenos Aires and CABA

In this section, train mobility in number of purchased tickets will be analyzed. Specifically, trains belonging to AMBA region (Buenos Aires Metropolitan Area in English) are of interest.

First, the amount of tickets sold by station in March 2020 (first month of Argentinian quarantine) will be compared with March 2019 for each train line.

After that, a comparison between train lines will be made in order to see mobility reduction per line.

## Exploring and Cleaning

The dataset used for this study is called “trenes_pasajeros.csv”:

**Read dataset:**

```
In [118]: url = 'https://servicios.transporte.gob.ar/gobierno_abierto/descargar.php?t=trenes&d=pasajeros'
          trenes_file = 'Datasets/trenes_pasajeros.csv'

          #Download dataset from url:
          trenes_pasajeros_df = pd.read_csv(url)

          #The following commented line of code can be used instead of url download:
          #trenes_pasajeros_df = pd.read_csv(trenes_file)

          trenes_pasajeros_df.head()
```

Out[118]:

| | mes;linea;estacion;cantidad |
|---|---|
| 0 | 12/2019;belgranonorte;"Boulogne Sur";189118 |
| 1 | 12/2019;belgranonorte;Carapachay;44049 |
| 2 | 12/2019;belgranonorte;"Del Viso";76691 |
| 3 | 12/2019;belgranonorte;"Don Torcuato";128986 |
| 4 | 12/2019;belgranonorte;Florida;37226 |

As seen in the figure above, all data is merged in a single column, so first, it’ll be necessary to split:

```
trains_df = trenes_pasajeros_df['mes;linea;estacion;cantidad'].str.split(pat=';',expand=True)
trains_df.columns = ['Month','Line','Station','Total']
trains_df.head()
```

| | Month | Line | Station | Total |
|---|---|---|---|---|
| 0 | 12/2019 | belgranonorte | "Boulogne Sur" | 189118 |
| 1 | 12/2019 | belgranonorte | Carapachay | 44049 |
| 2 | 12/2019 | belgranonorte | "Del Viso" | 76691 |
| 3 | 12/2019 | belgranonorte | "Don Torcuato" | 128986 |
| 4 | 12/2019 | belgranonorte | Florida | 37226 |

The result is a DataFrame with the following columns:

* **Month**: Month and year

* **Line**: Train Line

* **Station**: Train station

* **Total**: Total amount of tickets sold in the specified station and month

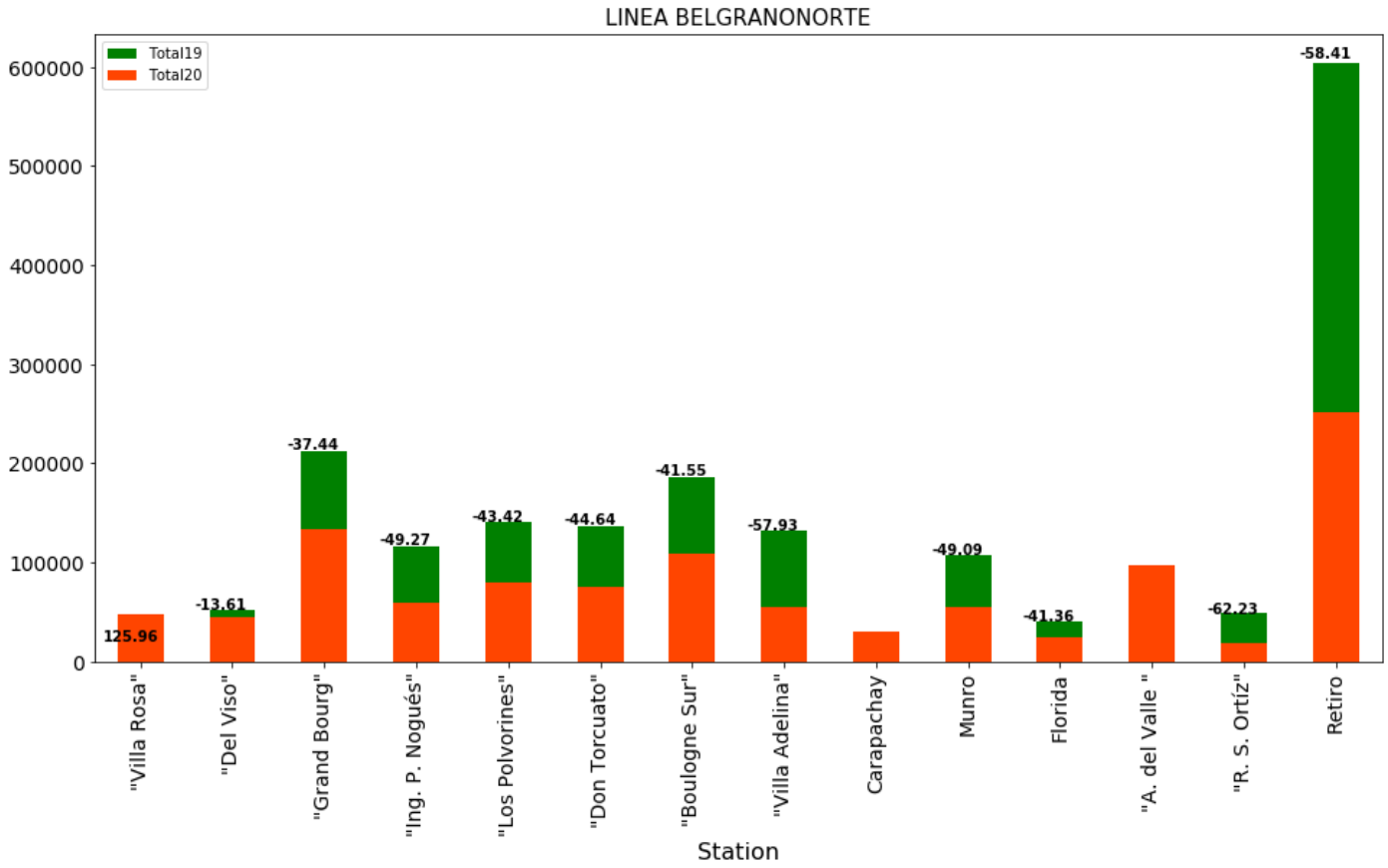
The next step is to create a Set of DataFrames, where each DataFrame of the set represents a particular Train Line. This will help to make all the plots for the analysis:

```
dataframes = {}
linesList = trainsClean['Line'].unique().tolist()
for i,lineName in zip(range(len(linesList)),linesList):
    line_df = trainsClean[trainsClean['Line'] == lineName]
    #Group by Station's total amount of passengers of 2019
    g19 = line_df[line_df['Month'] == '2019/03'].groupby('Station').agg(
                Total19=('Total',lambda x:int(x)))
    #Convert group to DataFrame
    g19.transform(lambda x:x)

    #Group by Station's total amount of passengers of 2020
    g20 = line_df[line_df['Month'] == '2020/03'].groupby('Station').agg(
                Total20=('Total',lambda x:int(x)))

    #Convert group to DataFrame
    g20.transform(lambda x:x)

    #Concatenate DataFrames:
    dataframes[i]=pd.concat([g19,g20],axis=1)

    #Add Variation column
    dataframes[i]['Variation %'] = ((dataframes[i]['Total20']/dataframes[i]['Total19'])-1)*100
    dataframes[i]['Variation %'] = [float('{:.2f}'.format(i)) for i in dataframes[i]['Variation %']]

    #Add line name and change order of columns:
    dataframes[i]['LineName'] = [lineName]*dataframes[i].shape[0]
    col = dataframes[i].columns.tolist()
    col = col[-1:] + col[:-1]
    dataframes[i] = dataframes[i][col]
#end for
```

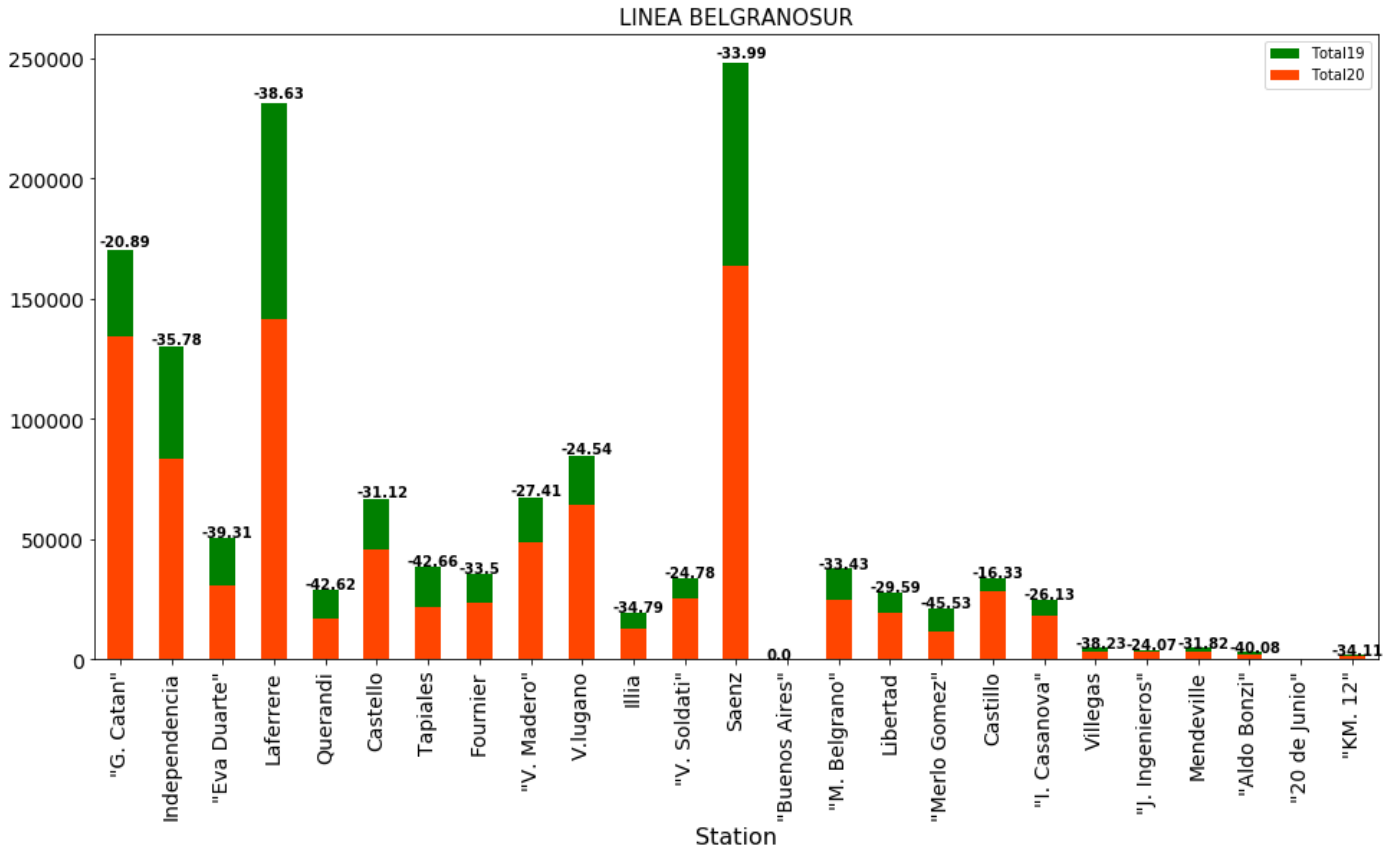
## 1. Subplots to show variation per Station in number of tickets and percentages:

The following plots represent sold tickets per Train Line. Each bar represents a train station. Green Bars represent total amount of tickets sold in March 2019. Orange Bars represent total amount of tickets sold in March 2020. There are stations with only one kind of bar (green or orange, not both), this is due to maintenance activities that had taken place in those stations when data acquisition was made.

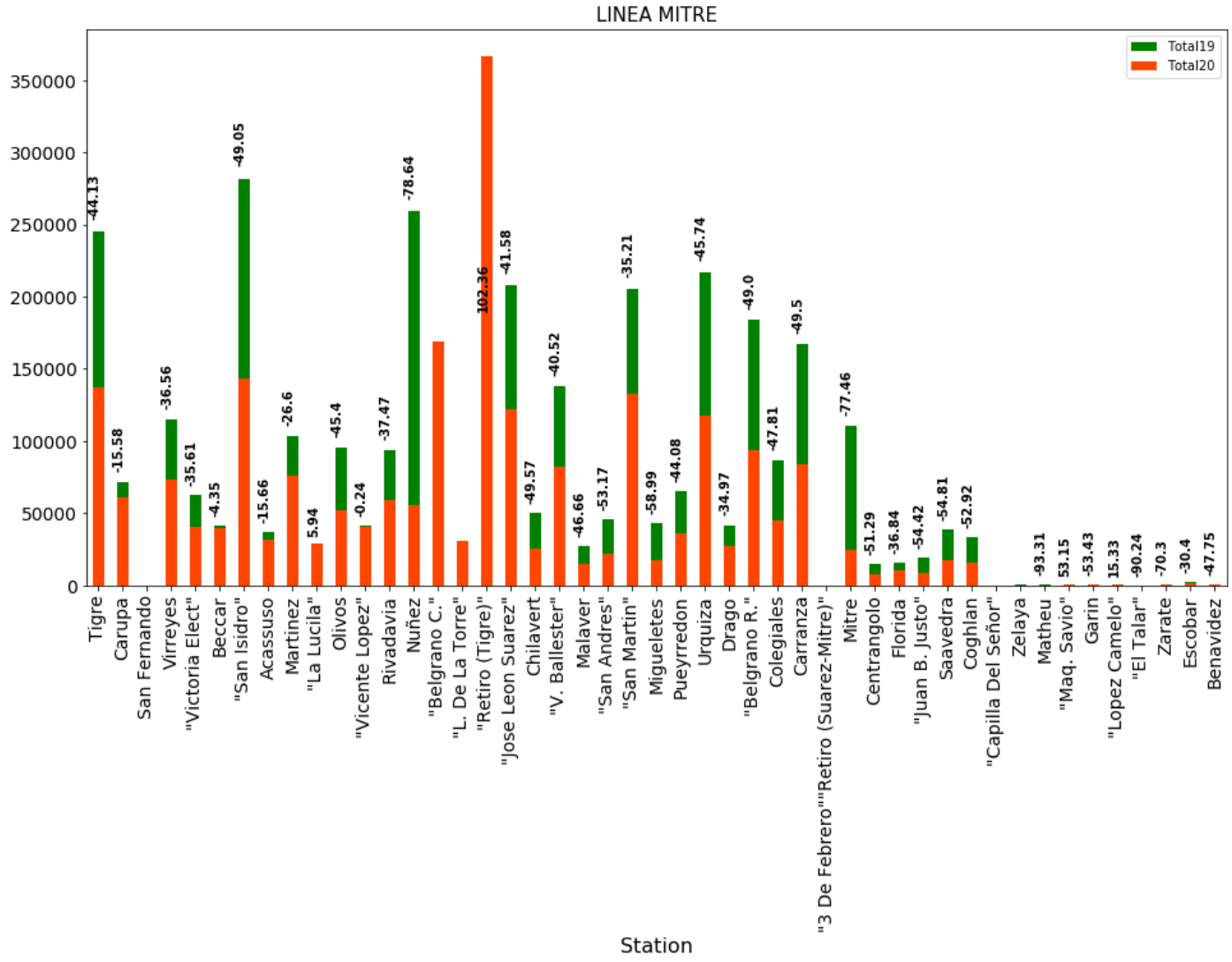
## Belgrano Norte Line:



## Belgrano Sur Line:

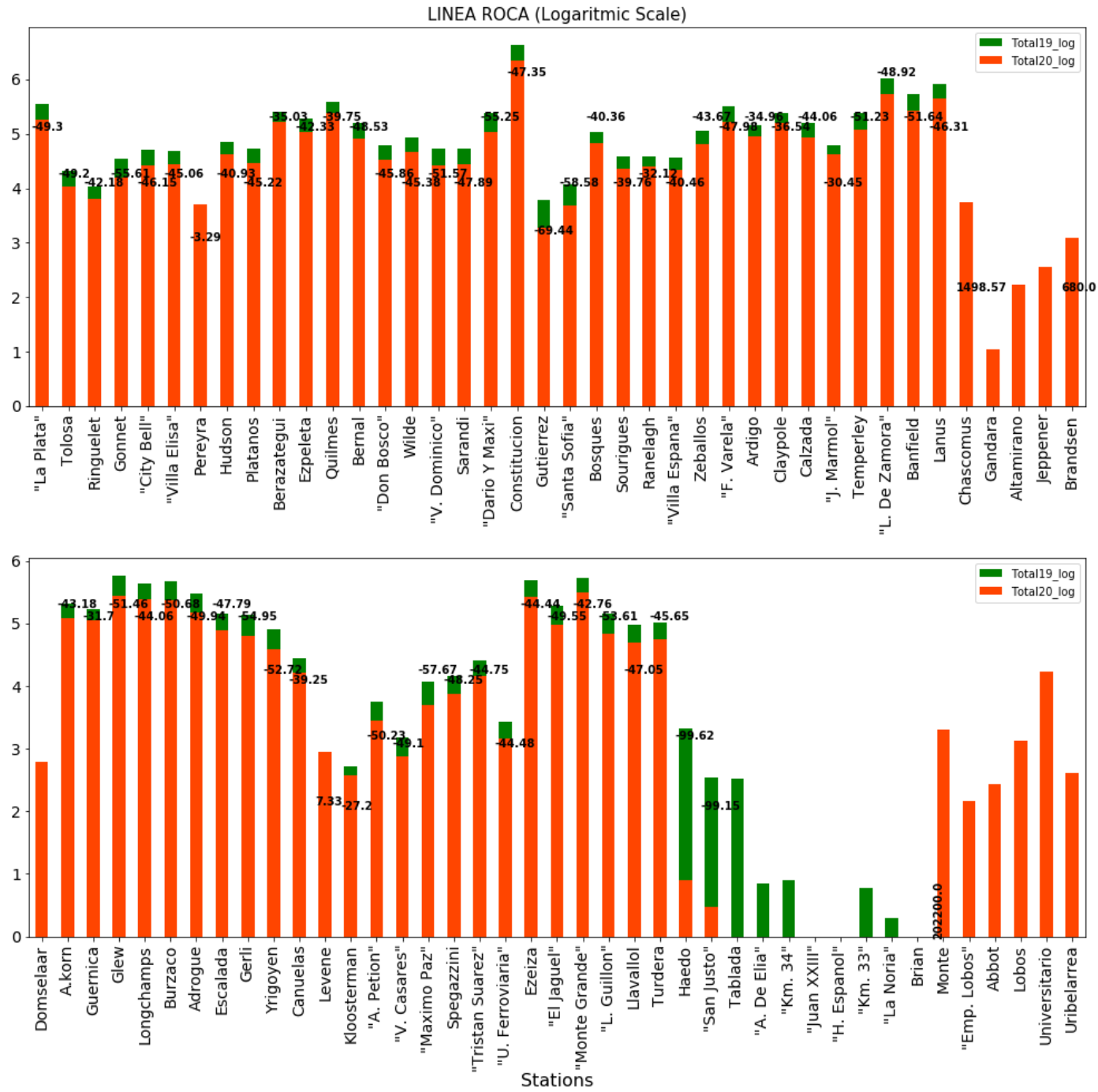



## Mitre Line:




## Roca Line (Logarithmic Scale)

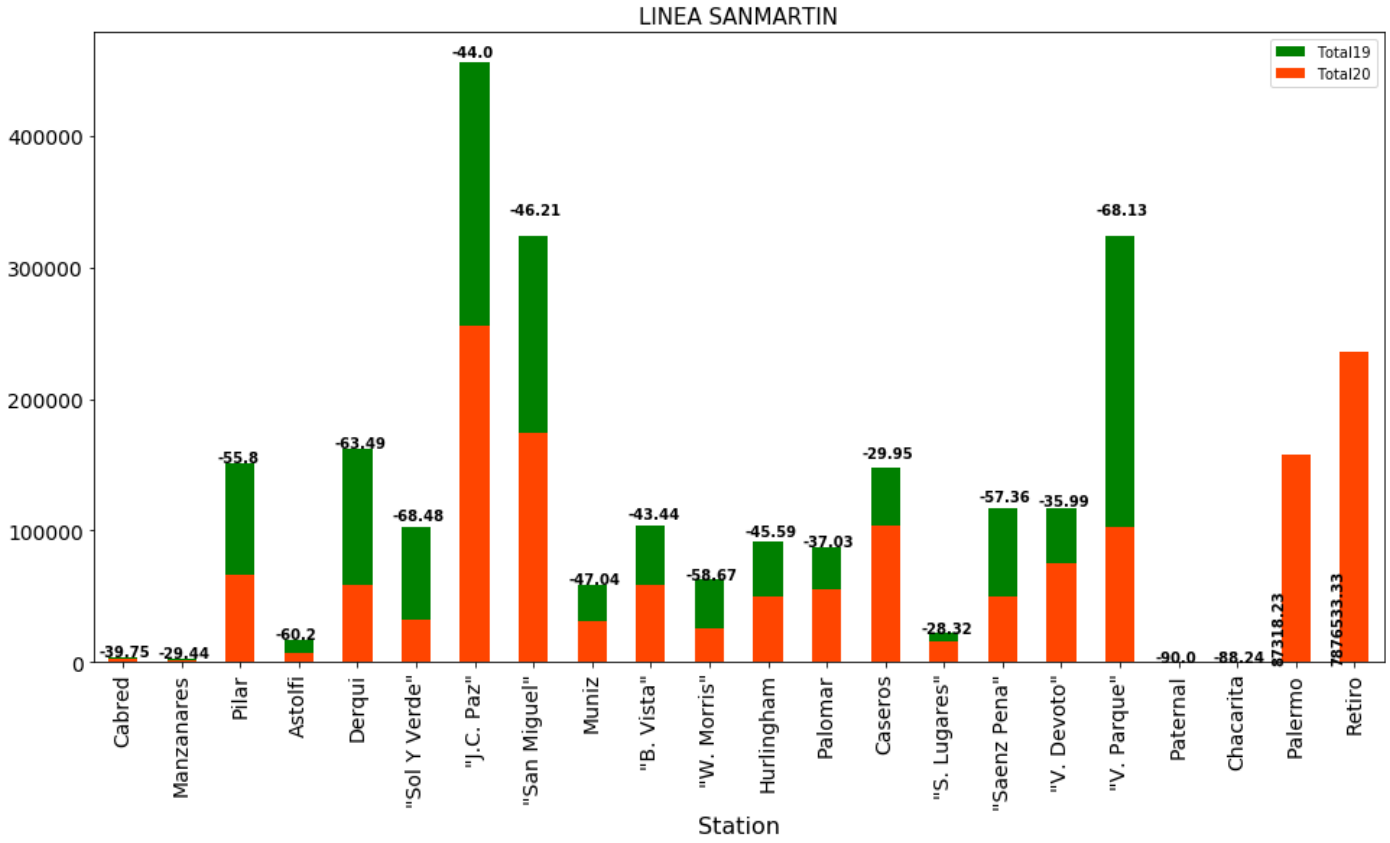
This line shows a huge range of number of sold tickets between stations. To make data visualization more comfortable to readers, vertical axis will be represented in logarithmic scale. Also, due to the large number of stations of Roca Line (81), bar graph will be splitted in two:



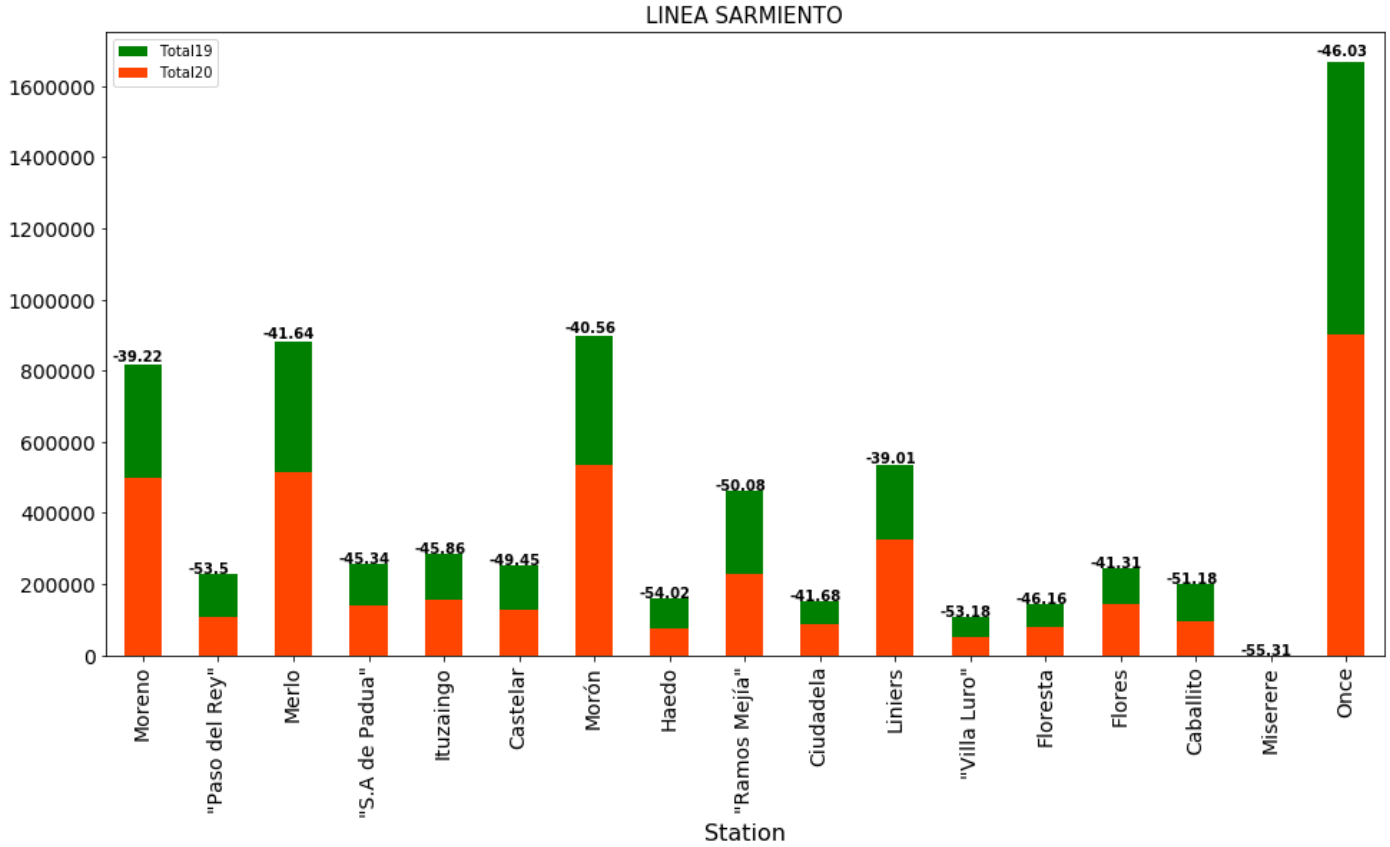




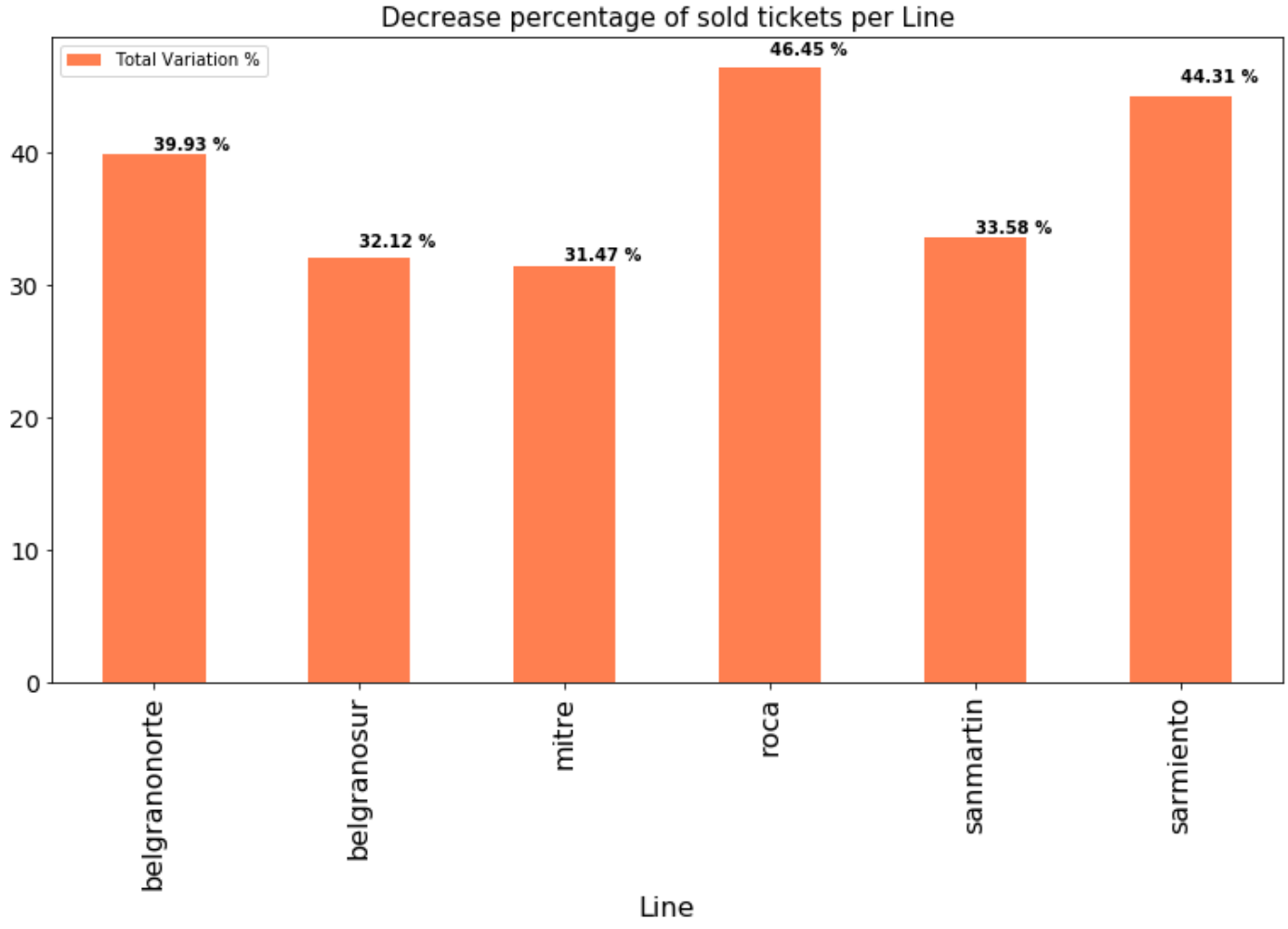
## San Martín Line:




## Sarmiento line:



## 2. Variation percentage for each line:




## Conclusions

From this analysis, we can see that the number of sold tickets in train lines of Metropolitan Area were reduced between 30 % and 47 % in the first month of quarantine. Some train stations shows variations lower than -60%. In some cases, March 2020 registered higher amount of sold tickets, which can be caused by maintenance activities done there in March 2019.

# Discussion

The analysis done here throws several points to discuss:

* Distribution of Covid-19 cases in Argentina is not homogeneous. A further analysis can be done here to understand where are the most dense zones, even inside Provinces (i.e., departments, neighborhoods, etc.)

* Cluster techniques can be applied to find similarities between cases. Features like age, gender, province, department, funding source, diagnosis date, symptoms onset date, among others, can be used to complete this study.

* A model could be inferred to **predict** deaths and cases in territory based on previous analysis.

* Industrial and Economic activities in general were affected in different ways. A deep analysis could be made here to take decisions and politics in order to help those sectors where Pandemic and Quarantine had been more damaging.

* Mobility study should be made to know how much people (and virus) are moving, specially in critical zones like Buenos Aires and CABA.

# Conclusion

To conclude this study, I want to mention the importance of data collection and analysis in critical scenarios like the one we are experiencing right now. Having accurate and reliable data is as important as medical care, financial and economic aid. Today, governments are taking decisions funded on data, which gives enormous importance and responsibility to Data Science and Data Scientists.