




COVID-19:

Statistical Analysis of cases in Mexico and risk factors for mortality of patients

- Berenice Cedillo
- Luis Cavazos
- Mario Ibarra
- René Mondragón

Really? COVID-19?



Though it may be the last thing we all want to think about, after months of a worldwide struggle with COVID-19, we believe that there are still a lot of key points to be found in data.

Our Hypothesis

There are risk factors that can help to identify Mexican patients with poor prognosis at an early stage, either personal, demographic or medical preconditions.

Our Questions

- Are there medical preconditions that make it more feasible for a Mexican COVID-19 patient to die?
- Are there demographic characteristics that make it more feasible for a Mexican COVID-19 patient to die?
- Is there a significant relation between the date when COVID-19 symptoms were first presented, the COVID-19 confirmation and the defunction date?

Our Findings

- Although there are factor of statistical significance, the available factors are not enough to explain the behavior of death rates in mexican patients due to COVID-19.

Questions & Data

- ***Are there medical preconditions that make it more feasible for a Mexican COVID-19 patient to die?***

Due to what have been learnt about COVID-19, certain medical conditions could cause an infection to be more dangerous (particularly respiratory and cardiovascular conditions). Are there particular conditions that could be identified as more critical for a COVID-19 patient?

- **Are there demographic characteristics that make it more feasible for a Mexican COVID-19 patient to die?**

Additional to medical preconditions, certain characteristics of COVID-19 patients have been deemed as dangerous, particularly, the age of the patient. Does data verify this affirmation and is there any other characteristic of the patient that could be relevant in his/her prognosis.

- **Is there a significant relation between the date when COVID-19 symptoms were first presented, the COVID-19 confirmation and the defunction date?**

There are important period of times in a patient process, could the length of these periods be an important variable in the development of a patient, either as an independent or a dependent variable?

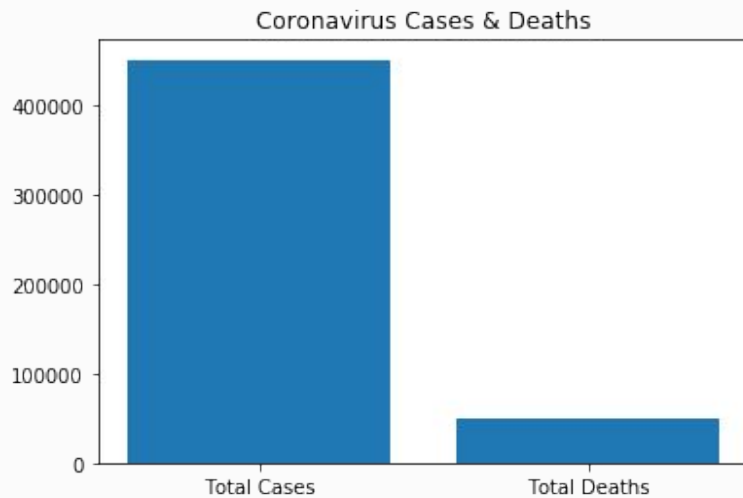
Our Data

- Daily updated Databases of all COVID-19 cases nationwide provided by the Health Department, includes:
 - Details of the case (Registered Date, Institution, Results, etc.)
 - Patient demographics (Gender, Residency, Age, Nationality, etc.)
 - Patient medical preconditions (Pneumonia, Diabetes, Obesity, Smoking, etc.)

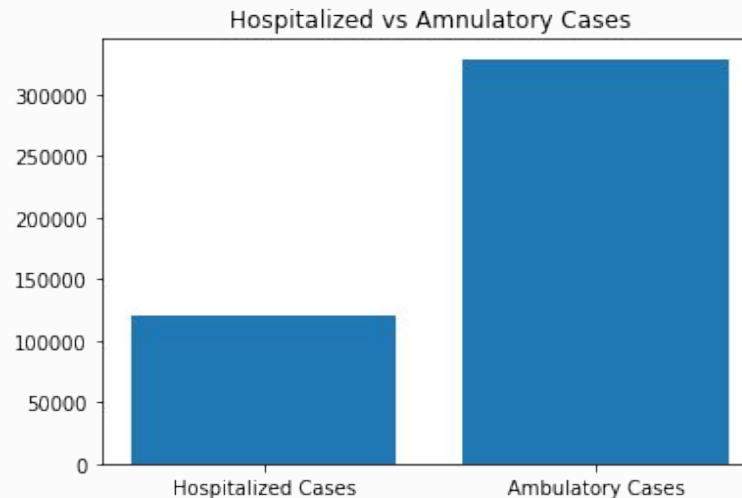
Data Cleanup & Exploration

- Process to extract csv from zip, import it to dataframe.
- Merging with catalogs provided by Health Department.

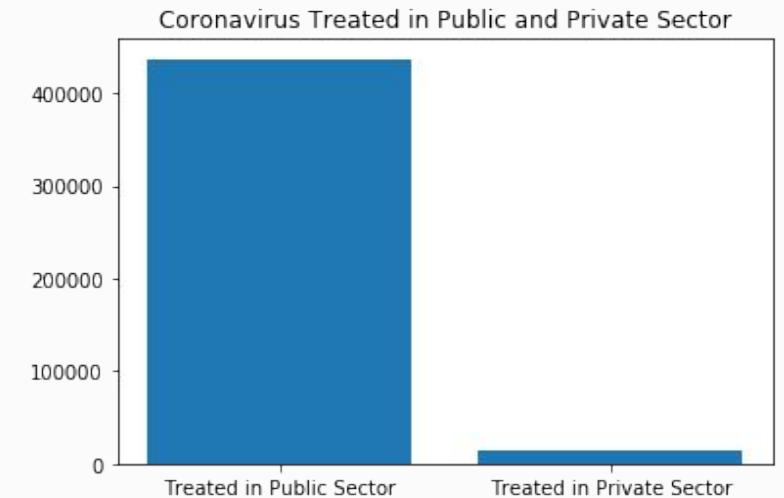
Key Statistics Summary



Coronavirus in Mexico has a 10.86% Mortality Rate



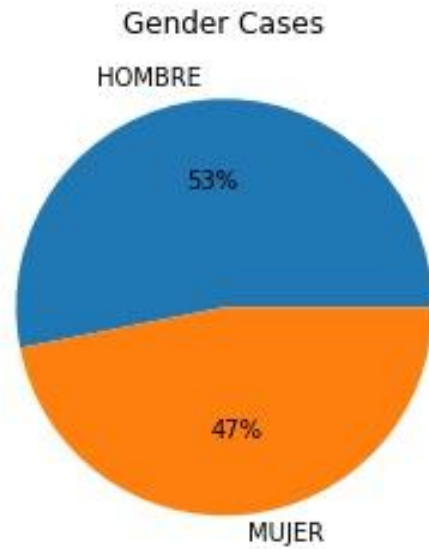
Most of Coronavirus cases where ambulatory (73%).



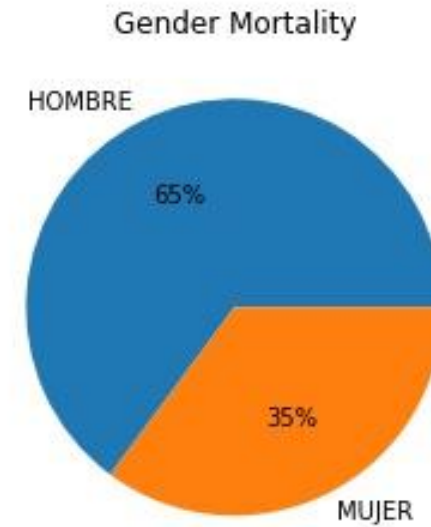
Coronavirus was treated mostly treated by Public Health service (97%).

Data Cleanup & Exploration

Cases



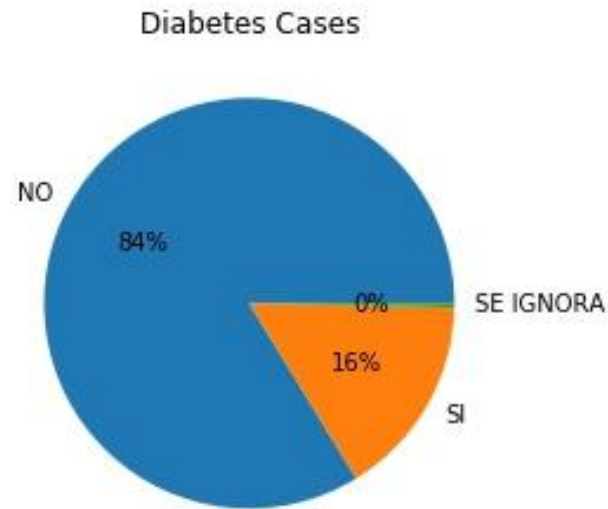
Deaths



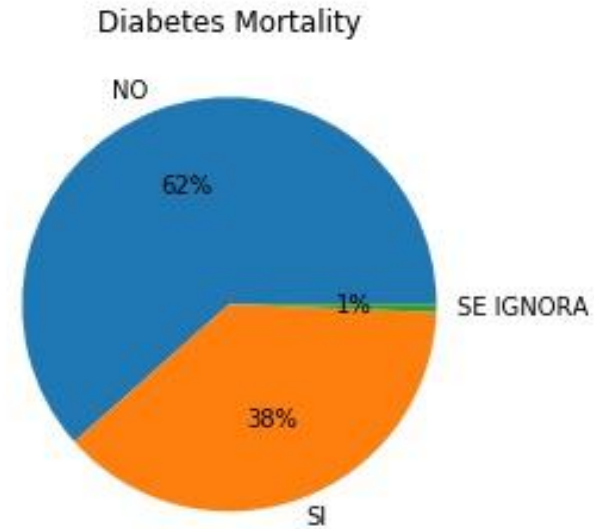
Men are more vulnerable than women against covid. Despite men just have 3% more cases than women, they got a 65% death rate.

Data Cleanup & Exploration

Cases



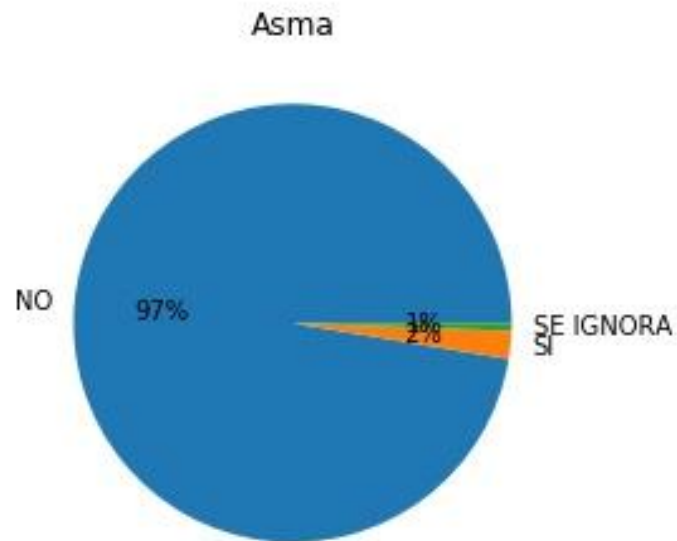
Deaths



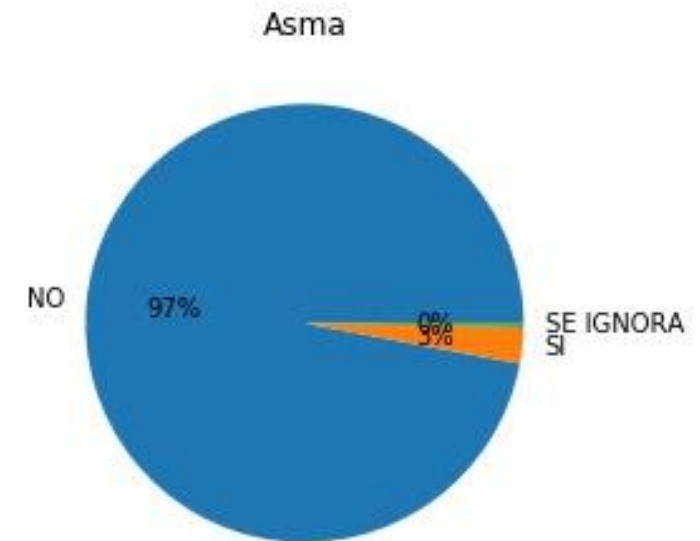
The most relevant medical condition on coronavirus deaths is diabetes. 38% of people who die from coronavirus has this disease.

Data Cleanup & Exploration

Cases



Deaths

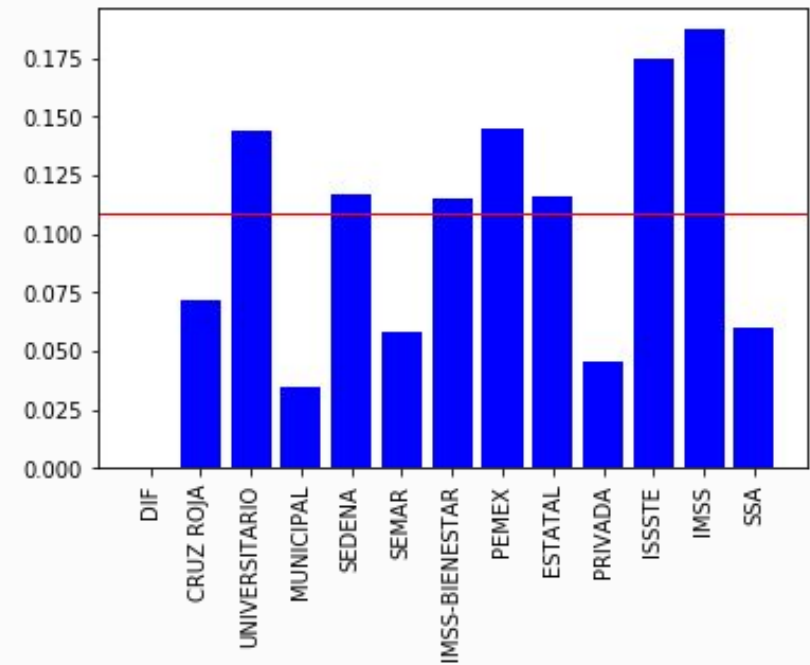


The less relevant medical condition on coronavirus deaths is asma. This pre existing medical condition doesn't have any impact on death rate.

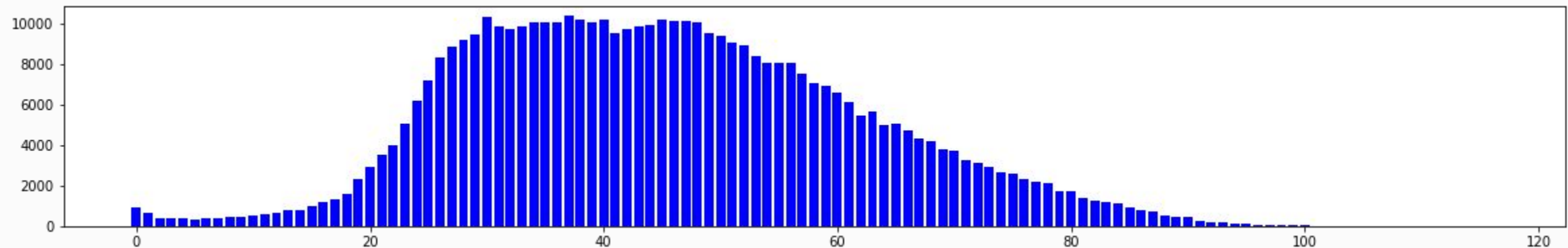
Data Cleanup & Exploration



Death Rate by Medical Institution



Number of cases by Age



Date Analysis Data Cleanup

Data was analyzed and cleared, in the CSV there was date information that didn't fit any format (9999-99-99), this information was removed from this study as it didn't helped answer the question. 3 values were added in which we could detect the days between the following: Symptom detection until Hospitalization, Hospitalization until Demise and Symptom detection until demise.

```
df_fall["Sintomas a hospitalizacion"] = (df_fall['FECHA_INGRESO'] - df_fall['FECHA_SINTOMAS']).dt.days
df_fall["Sintomas a Defuncion"] = (df_fall['FECHA_DEF'] - df_fall['FECHA_SINTOMAS']).dt.days
df_fall["Hospitalizacion a Defuncion"] = (df_fall['FECHA_DEF'] - df_fall['FECHA_INGRESO']).dt.days
```

Following this information we proceeded to clean the data by removing outliers that fell outside of the quantiles for all 3 added values as well as the age value:

	EDAD	Sintomas a hospitalizacion	Sintomas a Defuncion	Hospitalizacion a Defuncion
count	46364.000000	46364.000000	46364.000000	46364.000000
mean	61.859740	4.344319	11.116621	6.772302
std	13.334847	3.326185	6.108281	5.549489
min	25.000000	0.000000	0.000000	0.000000
25%	53.000000	2.000000	6.000000	2.000000
50%	62.000000	4.000000	10.000000	5.000000
75%	72.000000	7.000000	15.000000	10.000000
max	90.000000	14.000000	29.000000	24.000000

We also proceeded to separate in bins the Age of all patients in 5 groups divided by 15 years

```
bins = [0, 15, 30, 45, 60, 100]
group_names = ["Menos de 15", "16 a 30", "31 a 45", "46 a 60", "mayores de 60"]
data_df = df_out
df_out["Grupo de Edad"] = pd.cut(df_out["EDAD"], bins, labels=group_names, include_lowest=True)
```

After finishing the cleanup, we proceeded to Analyse the data to find any correlation between the days between Symptom detection until Hospitalization, Hospitalization until Demise and Symptom detection until demise.

Date Analysis Data Exploration

The correlation between the patient's Sex and the time from symptom detection till demise is:

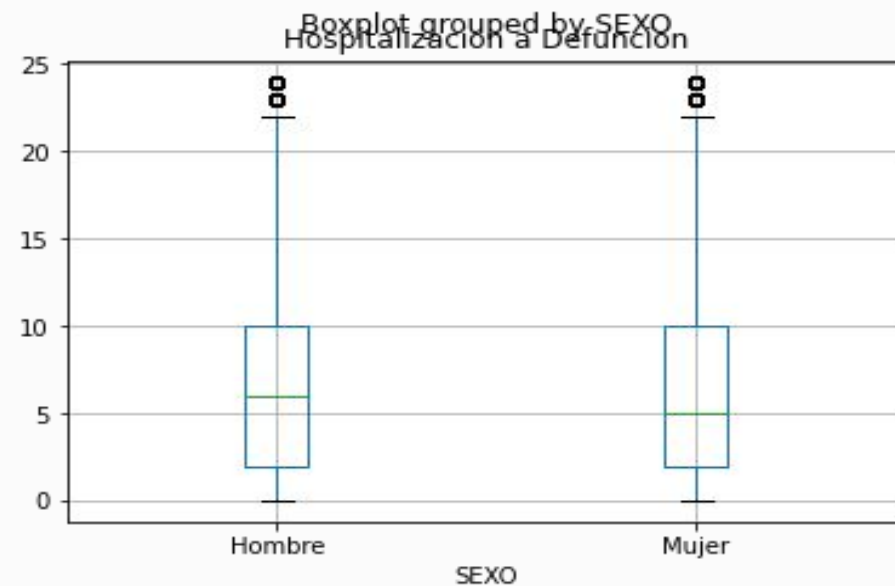
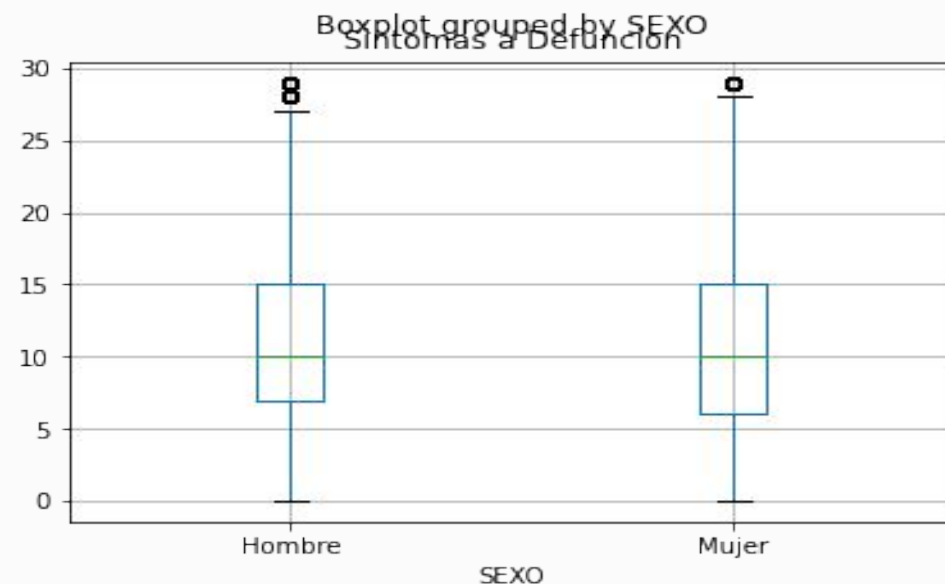
Hombre 0.031429

Mujer -0.031429

The correlation between the patient's Sex and the time from hospitalization till demise is:

Hombre 0.013096

Mujer -0.013096



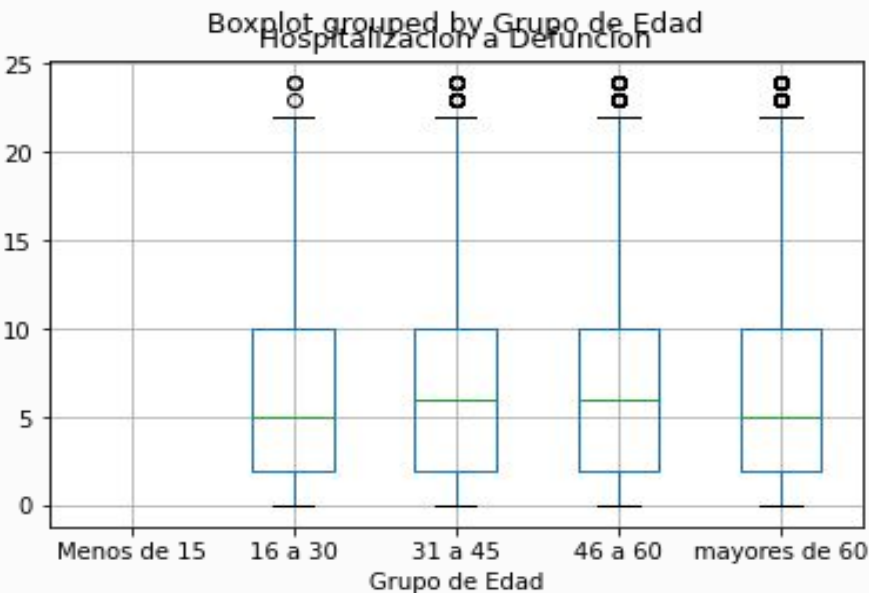
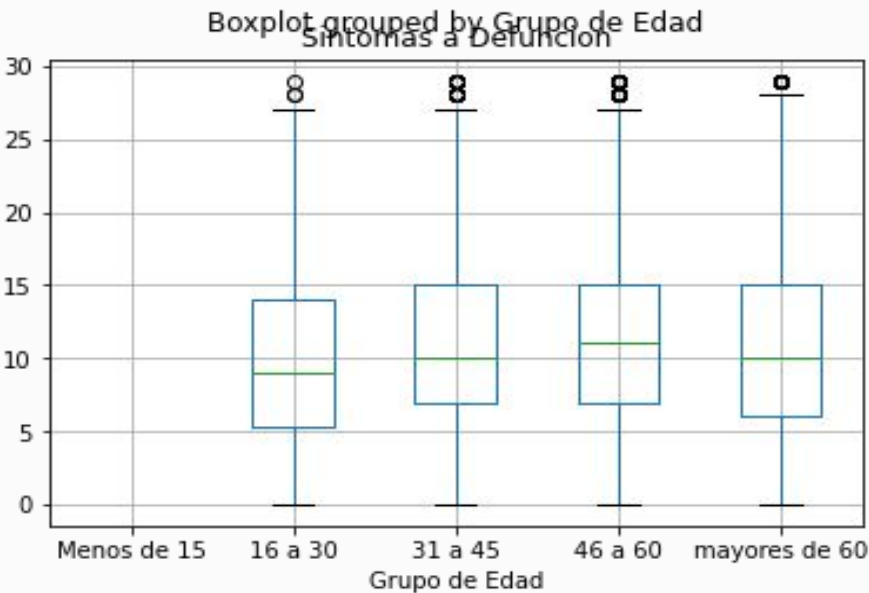


The correlation between the patient's Age Group and the time from symptom detection till demise is:

16 a 30	-0.012589
31 a 45	0.000354
46 a 60	0.032448
mayores de 60	-0.028125

The correlation between the patient's Age Group and the time from hospitalization till demise is:

16 a 30	-0.009423
31 a 45	-0.001215
46 a 60	0.023571
mayores de 60	-0.019452



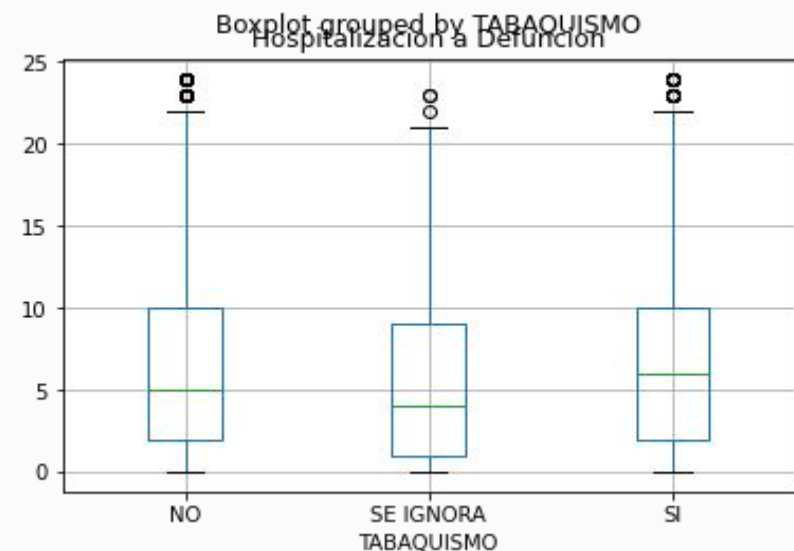
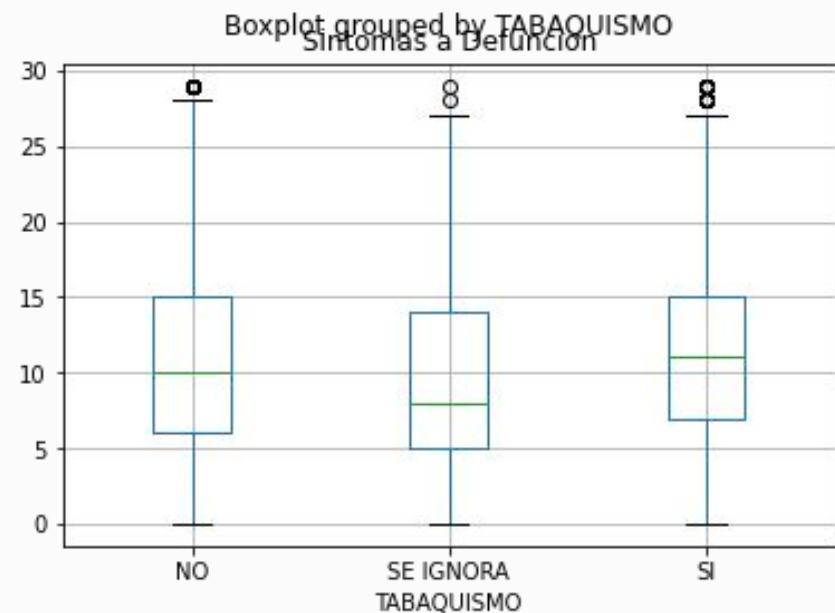


The correlation between the patient's symptom detection and it's tobacco usage is:

NO	-0.008151
SE IGNORA	-0.016140
SI	0.013002

The correlation between the patient's hospitalization until demise and it's tobacco usage is:

NO	-0.002907
SE IGNORA	-0.015574
SI	0.007424

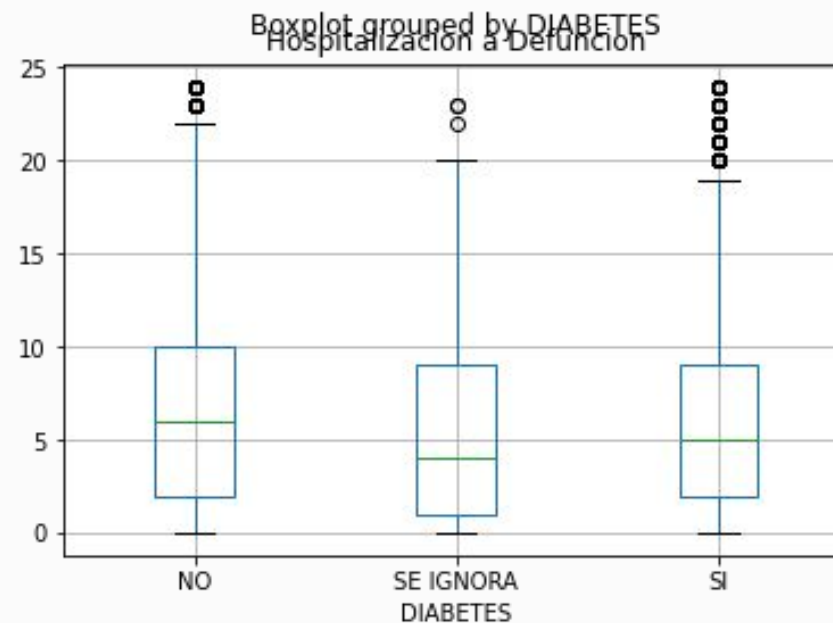
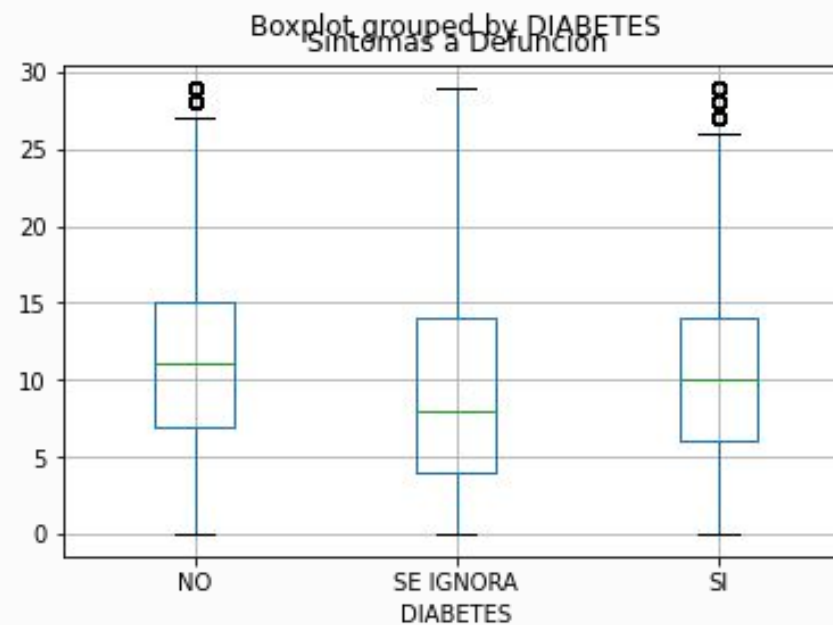



The correlation between the patient's symptom detection until demise and Diabetes is:

NO	0.065779
SE IGNORA	-0.018351
SI	-0.063067

The correlation between the patient's hospitalization until demise and Diabetes is:

NO	0.066067
SE IGNORA	-0.017381
SI	-0.063509





We continued to try to find correlation between Symptom until demise and Hospitalization until demise involving the following factors:

The correlation between the patient's symptom detection until demise and Asthma is:

NO 0.004441
SE IGNORA -0.017672
SI 0.004309

The correlation between the patient's symptom detection until demise and Obesity is:

NO 0.007924
SE IGNORA -0.016231
SI -0.005076

The correlation between the patient's Immunosuppressed disease and the time from symptom detection till demise is:

NO 0.022445
SE IGNORA -0.018424
SI -0.015890

The correlation between the patient's Hospitalization until demise and Asthma is:

NO 0.004820
SE IGNORA -0.016821
SI 0.003435

The correlation between the patient's Hospitalization until demise and Obesity is:

NO 0.038360
SE IGNORA -0.020382
SI -0.035010

The correlation between the patient's Immunosuppressed disease and the time from hospitalization till demise is:

NO 0.012682
SE IGNORA -0.016959
SI -0.005807

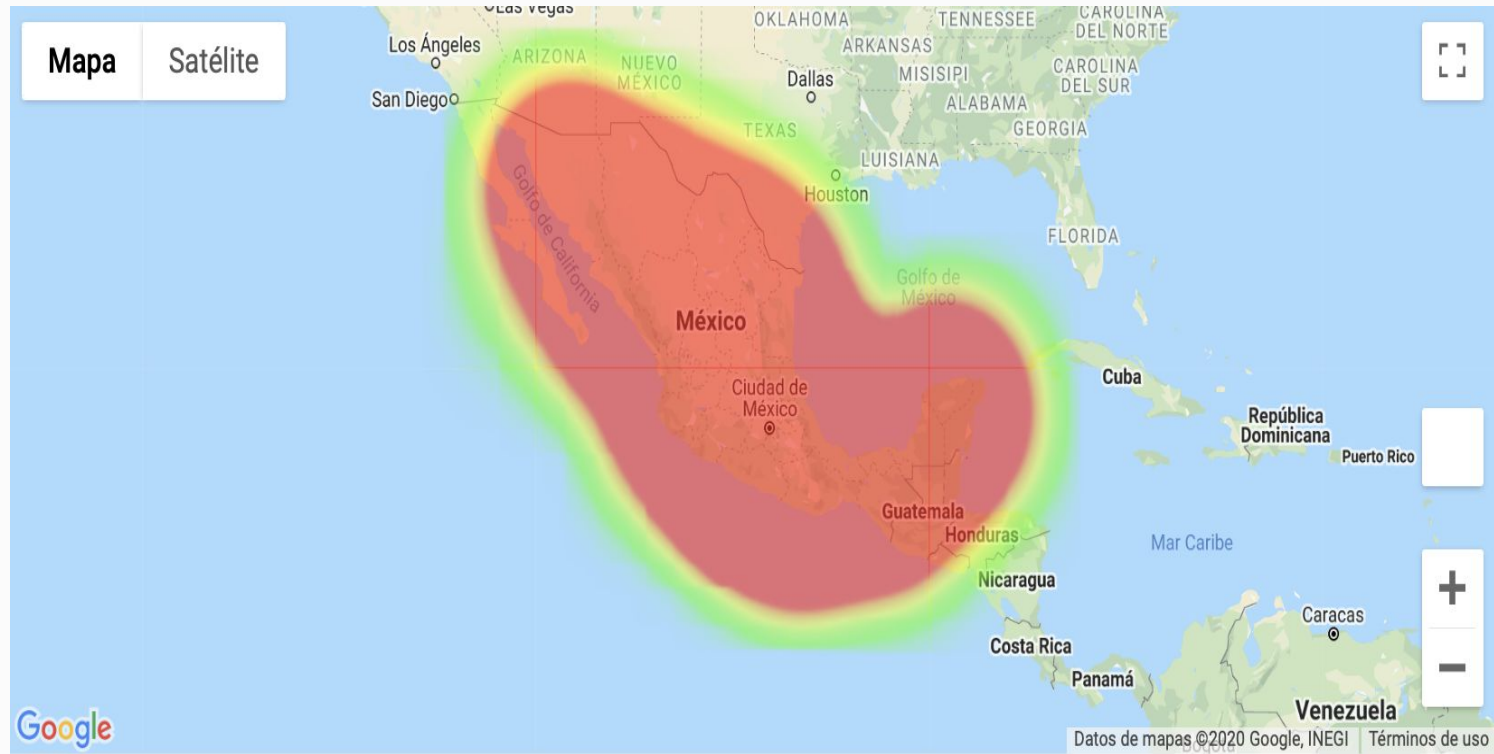


The correlation between the patient's symptom detection until demise and State of residency is:

AGUASCALIENTES	-0.007068
BAJA CALIFORNIA	0.031151
BAJA CALIFORNIA SUR	-0.001654
CAMPECHE	0.010961
CHIAPAS	0.004246
CHIHUAHUA	-0.010514
CIUDAD DE MÉXICO	0.057223
COAHUILA DE ZARAGOZA	-0.055868
COLIMA	0.004497
DURANGO	-0.013042
GUANAJUATO	-0.021990
GUERRERO	-0.016783
HIDALGO	-0.004010
JALISCO	0.000569
MEXICO	-0.018255
MICHOACAN DE OCAMPO	-0.001767
MORELOS	-0.044030
NAYARIT	-0.024695
NUEVO LEON	-0.002966
OAXACA	-0.029358
PUEBLA	0.015726
QUERETARO	-0.001269
QUINTANA ROO	0.000213
SAN LUIS POTOSI	-0.017841
SINALOA	0.037156
SONORA	0.022745
TABASCO	0.017917
TAMAULIPAS	-0.016146
TLAXCALA	-0.002955
VERACRUZ	-0.028536
YUCATAN	0.004944
ZACATECAS	-0.019326

The correlation between the patient's hospitalization until demise and State of residency is:

AGUASCALIENTES	0.009362
BAJA CALIFORNIA	0.032127
BAJA CALIFORNIA SUR	0.010194
CAMPECHE	-0.007450
CHIAPAS	0.005086
CHIHUAHUA	0.002315
CIUDAD DE MÉXICO	0.056013
COAHUILA DE ZARAGOZA	-0.021281
COLIMA	0.008440
DURANGO	-0.001837
GUANAJUATO	-0.014850
GUERRERO	-0.029736
HIDALGO	-0.012902
JALISCO	0.012034
MEXICO	0.013341
MICHOACAN DE OCAMPO	-0.007995
MORELOS	-0.027446
NAYARIT	-0.016271
NUEVO LEON	0.014573
OAXACA	-0.023247
PUEBLA	-0.019923
QUERETARO	0.026066
QUINTANA ROO	-0.002113
SAN LUIS POTOSI	-0.015684
SINALOA	0.011519
SONORA	0.011141
TABASCO	-0.029687
TAMAULIPAS	-0.016045
TLAXCALA	-0.018515
VERACRUZ	-0.029593
YUCATAN	0.000592
ZACATECAS	-0.006585

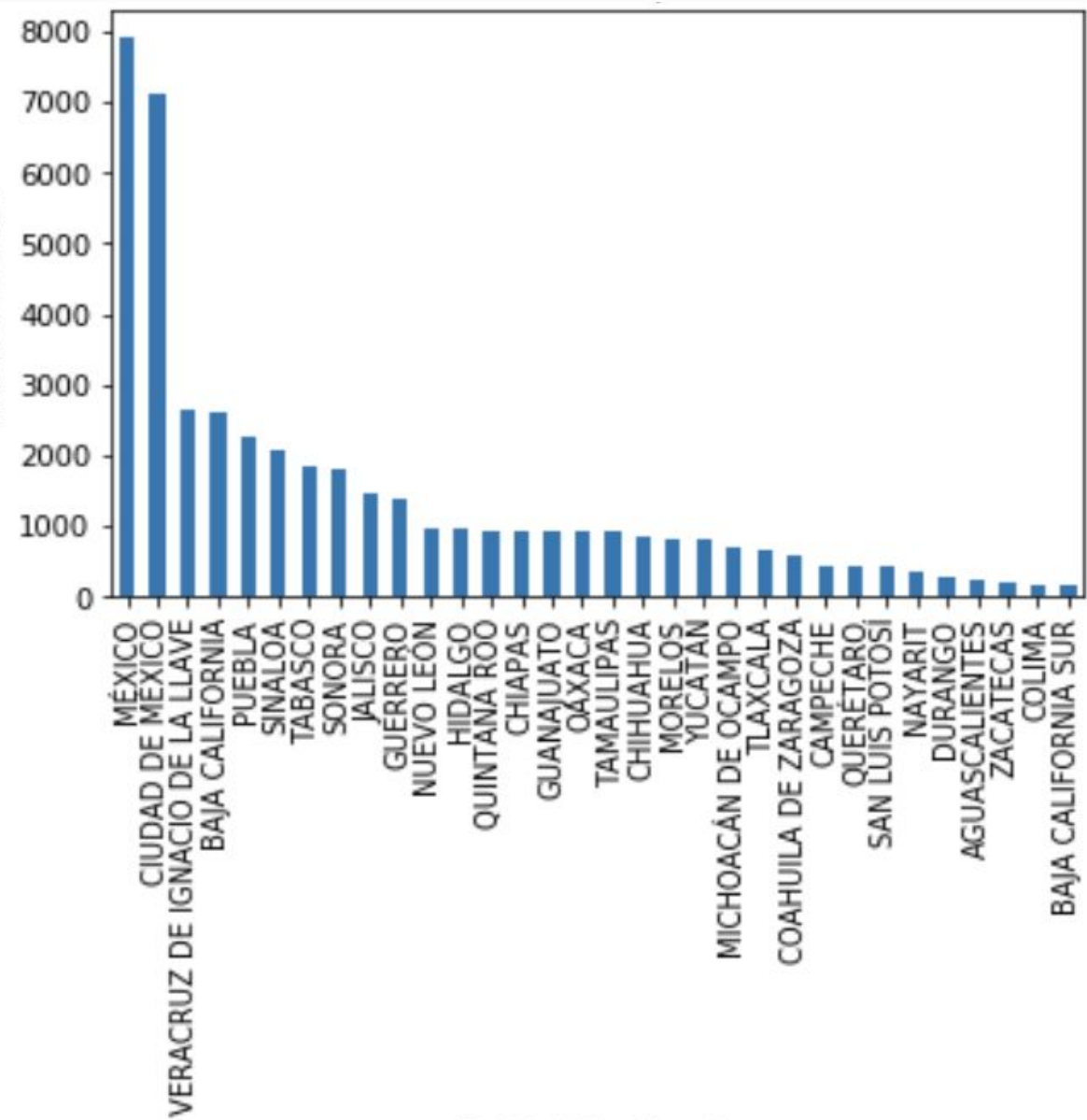



The map of Mexico shows a high rate of infection and mortality, which is alarming and suggests taking drastic measures to manage and solve this pandemic.



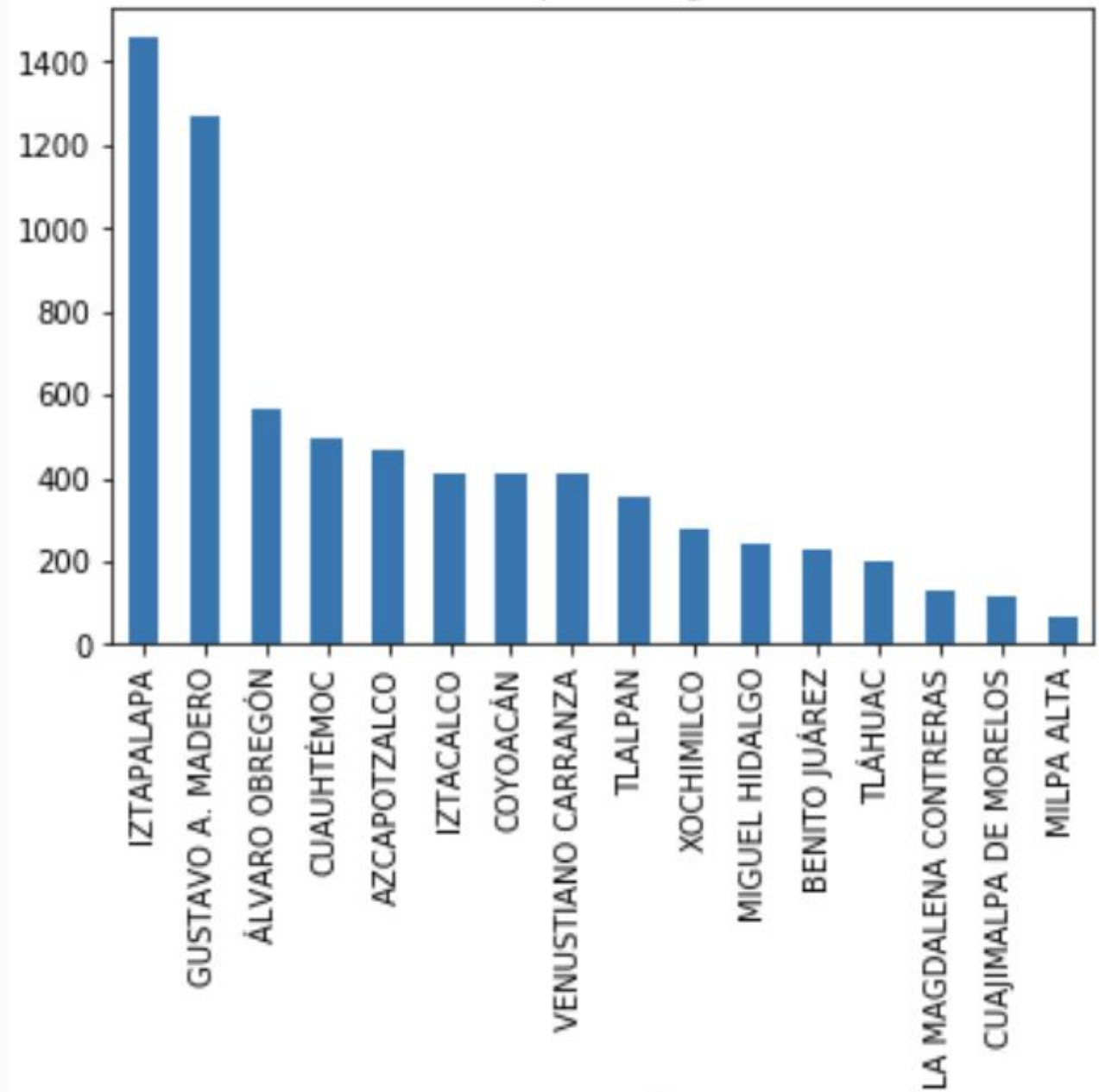
The bar graph of cases by states shows that State of Mexico and Mexico City are states with the highest number of covid cases, this is due to the number of people and their mobility, which is why confinement and protection measures were taken.

The states with less mobility present fewer covid cases, such as Aguascalientes, Zacatecas, Colima and Baja California Sur.



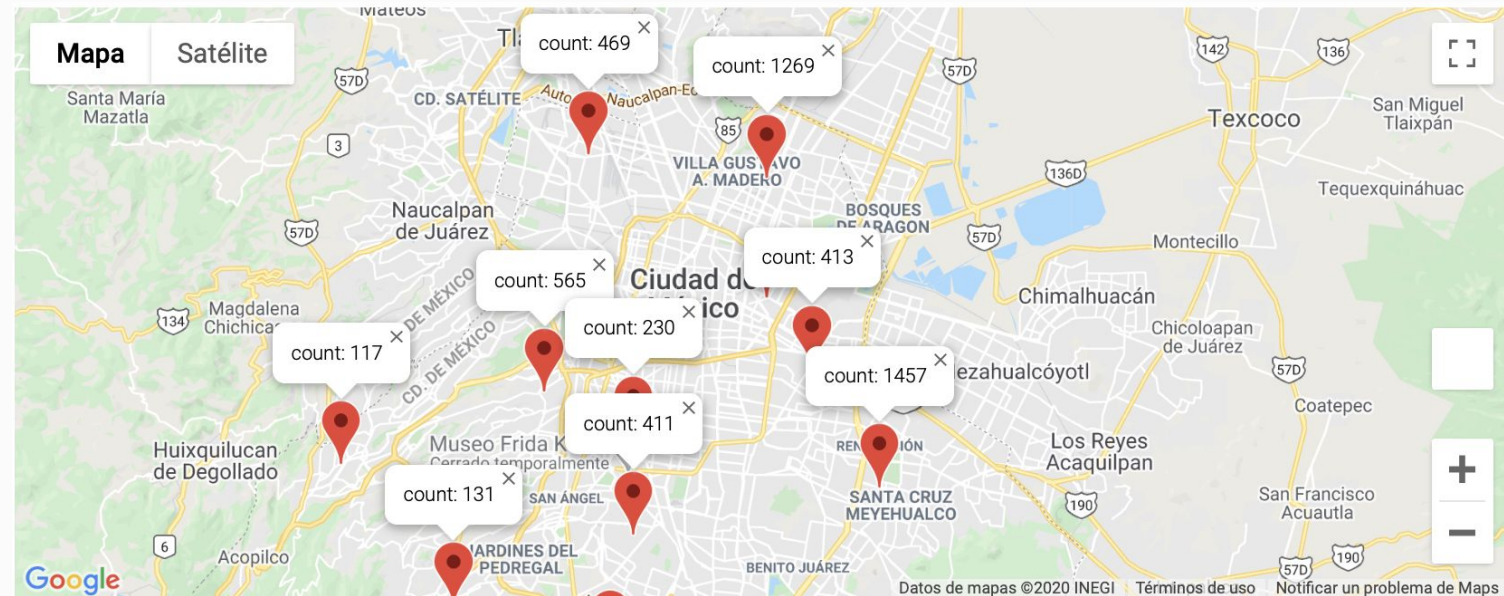


In the graph of cases per delegation in Mexico City shows that Iztapalapa and Gustavo A. Madero have a greater number of cases, due to their proximity to State of Mexico and the mobility of the people, which is why care and confinement measures were applied in the State and Mexico City.



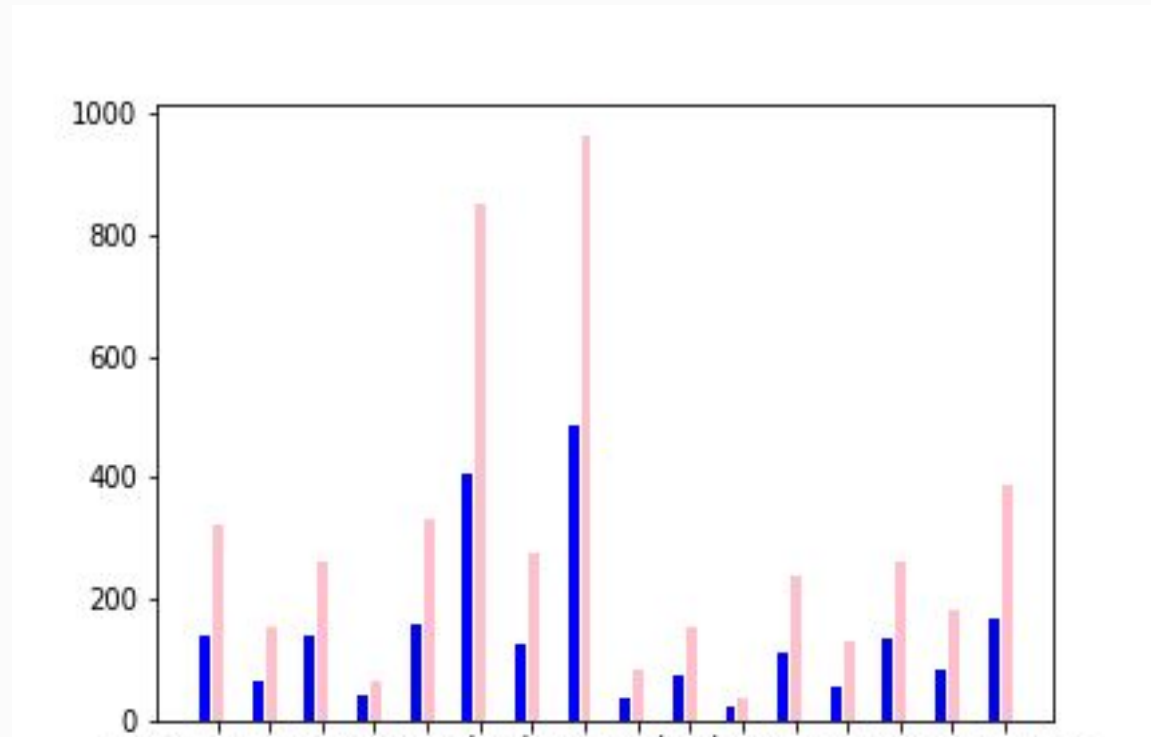


The map shows the delegations with cases of covid, Santa Cruz Meyehualco being the one with the highest number of infections (1457) and Gustavo A Madero (1269).



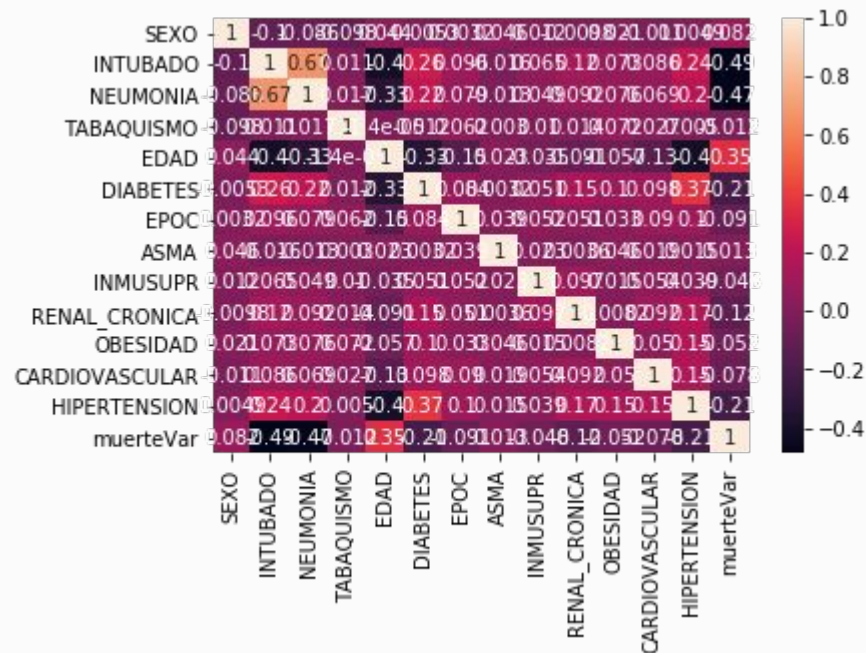
The delegations with the lowest number of infections are Santa Fe and Ajusco, and it can be seen that, due to their proximity or adjacency to State of Mexico, these municipalities have more cases of covid.





The bar graph of the Mexico City shows delegations by gender.

Women are the most vulnerable to overexploitation, being the same delegations from. Gustavo A. Madero and Iztapalapa are the ones that present the highest number of cases, according to the data registered, more care should be taken or attention should be focused on this sector of the population.



With the available variables, the best model isn't good enough to explain the behavior of death rates in Mexican patients due to COVID-19.

Though some of the variables seem to have a relevant effect, for example, Renal diseases and Pneumonia.

Dep. Variable:	muerteVar	R-squared:	0.275
Model:	OLS	Adj. R-squared:	0.275
Method:	Least Squares	F-statistic:	1.882e+04
Date:	Wed, 05 Aug 2020	Prob (F-statistic):	0.00
Time:	20:29:27	Log-Likelihood:	-39560.
No. Observations:	447481	AIC:	7.914e+04
Df Residuals:	447471	BIC:	7.925e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	1.0107	0.012	84.769	0.000	0.987	1.034
SEXO	0.0249	0.001	31.352	0.000	0.023	0.026
EDAD	0.0035	2.79e-05	124.992	0.000	0.003	0.004
NEUMONIA	-0.2960	0.001	-280.683	0.000	-0.298	-0.294
DIABETES	-0.0350	0.001	-29.121	0.000	-0.037	-0.033
EPOC	-0.0538	0.003	-16.630	0.000	-0.060	-0.047
INMUSUPR	-0.0355	0.004	-9.646	0.000	-0.043	-0.028
RENAL_CRONICA	-0.1193	0.003	-40.889	0.000	-0.125	-0.114
CARDIOVASCULAR	-0.0210	0.003	-7.476	0.000	-0.026	-0.015
HIPERTENSION	-0.0261	0.001	-22.904	0.000	-0.028	-0.024

Omnibus:	116219.886	Durbin-Watson:	1.971
Prob(Omnibus):	0.000	Jarque-Bera (JB):	303995.060
Skew:	1.409	Prob(JB):	0.00
Kurtosis:	5.892	Cond. No.	1.50e+03



Our Findings


- There are demographic variables that have a clear impact in the prognosis of a mexican COVID-19 patients: Gender and Age.
- Particular medical preconditions have a statistical relevant effect, such as, Renal diseases and Pneumonia.
- Although there are statistical significant variables, with the available factor, the best model isn't good enough to explain the behavior of death rates in Mexican patients due to COVID-19.
- Mexico City and the State of Mexico have a greater number of people dead and infected by COVID-19 than other states due to higher population density than other states.



Our Findings

- We detected that on average it takes 4.3 days from Symptom detection until patient hospitalization and 7.7 days from Hospitalization until patient demise.
- We couldn't detect any visible correlation between Age, Sex, State of Residency and several medical preconditions such as: Diabetes, Asthma, Obesity, Tobacco condition and the time it takes until the patient's demise.
- It was also noticed that patients who referred to ignore that they had medical preconditions had a lower life expectancy compared to patients who provided the complete information regarding its medical conditions.

Now what...?

- 
- There were some factors that didn't seem to be complete accurate, such as the geographical information.
 - Further analysis could be done by including information particular to the geographical location of each case.
 - A low socioeconomic level was identified as a risk factor that makes the death of a COVID patient more likely, because this level is associated with poor access to medicines, health resources and the health system but we don't delved deeper into this issue.
 - General Analysis of COVID-19 have pointed out that the treatment quality of hospital seems to be a good predictor for death rates. If we had information available for each hospital, this could be an interesting analysis topic.
 - Although date information was cleaned of outliers, several of them still had outliers present.
 - Review evolution of information, testing of model.
 - Any question?