BUSINESS INTELLIGENCE PROJECT

ISYS 850

COVID-19 DATA ANALYSIS - *BELGIUM*

Team members
Chisako Tani
Hema Vivekanandan
Kavita Kathaith
Siddhant Navaratna
Sushma Srinivas

For this project we will be analyzing Covid data for Belgium. We will try to analyze the effects of various factors like age, location and other such factors on mortality and number of confirmed cases.

Data source:

For the purpose of this study we have selected data from Sciensano which is a research and national public health institute of Belgium - https://epistat.wiv-isp.be/Covid/ . The dataset includes data by age, sex, province, number of tests performed. The time period of the dataset is 1st March 2020 to 29th April 2020.

The weather data is collected from https://weather.com.

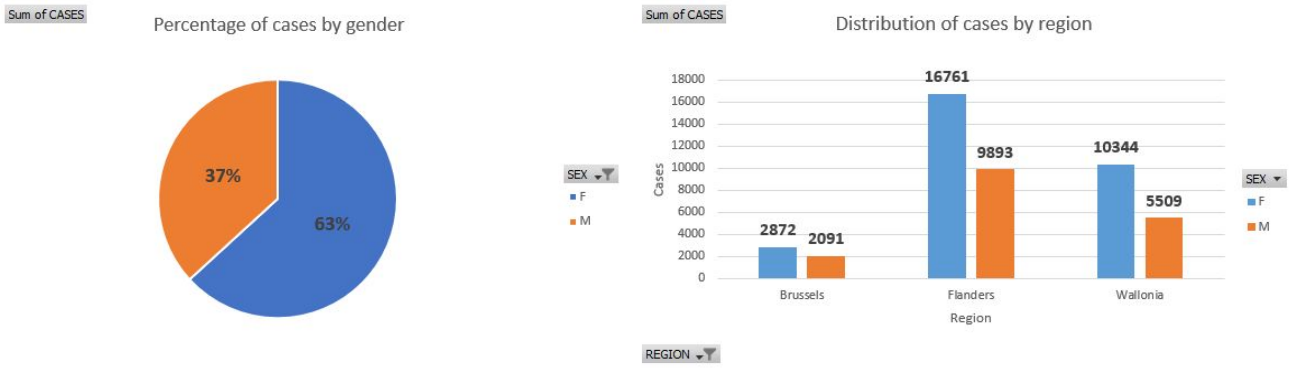
Data cleaning:

- The missing values were deleted from all the datasets.
- Unnecessary columns were deleted for each analysis.
- Duplicate records were checked and deleted for each dataset.
- Entry errors in the AGEGROUP column of the 'COVID19BE_CASES_AGESEX' dataset were corrected.
- Added temperature data for each entry corresponding to their respective dates and regions.

Hypothesis 1

Coronavirus affects one gender more than the other.

The graph below shows the distribution of cases by gender. We can see that Female cases are at a high compared to male cases.
This contradicts the general trend of male cases being higher than female cases in other countries. This makes us question our hypothesis more.



Next, we check if there is a trend in cases based on the region. From the graph below, we see that the pattern is the same for all the regions. Number of female cases is more than male cases for all the three regions.
We then see the correlation among the three variables – Gender, Region and Cases. From the correlation matrix we clearly see that the number of cases has weak correlation with Gender.

Hence, we reject our hypothesis that the number of cases is gender biased.

	CASES	gender_M	region_Flanders	region_Wallonia
CASES	1.000000	-0.201233	0.102704	-0.106079
gender_M	-0.201233	1.000000	0.011645	-0.017906
region_Flanders	0.102704	0.011645	1.000000	-0.809902
region_Wallonia	-0.106079	-0.017906	-0.809902	1.000000

In conclusion, the gender bias in number of cases maybe due to the following factors:

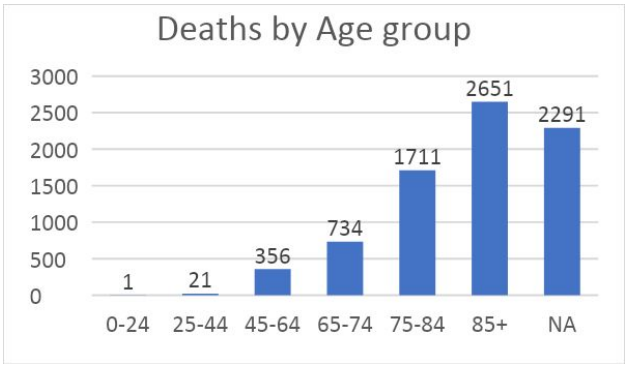
- Underlying medical condition
- Overall male and female population of the country

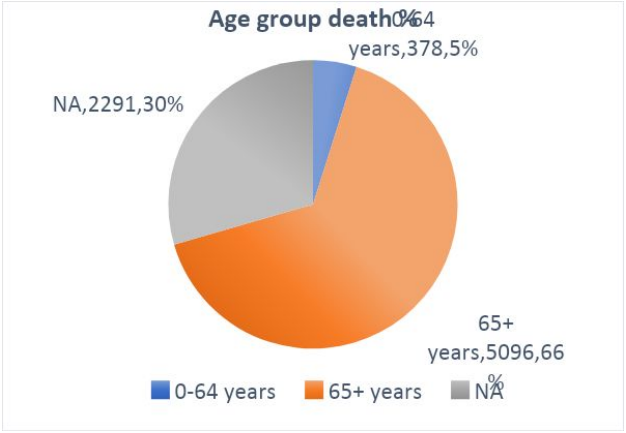
- Belgium might have a higher population of female healthcare workers
- Population of men and women being tested

Hypothesis 2

Coronavirus affects a certain age group more than it does others.

As of 1st May Belgium has about 7,765 deaths. We can see the breakup of deaths in the following graphs according to the age groups





We can clearly see in the pie chart that of total deaths people with 65 years and above are at most risk. More than 65% of infected people above the age of 65 have died. We can also see that as the age increases the number of deaths has also gone up.
In conclusion, there might a few reasons as to why the death rates is high amongst certain age groups like other chronic illnesses, mark of biological aging and declining immunity

Hypothesis 3

The number of hospitals in each region have an effect on mortality

	REGION	CASES	NR_REPORTING	DEATHS
0	Brussels	5055	15	1219
1	Flanders	27331	52	3904
2	Wallonia	16132	37	2721

	CASES	NR_REPORTING	DEATHS
CASES	1.00	0.99	1.00
NR_REPORTING	0.99	1.00	1.00
DEATHS	1.00	1.00	1.00

Dep. Variable:	DEATHS	R-squared:	0.998
Model:	OLS	Adj. R-squared:	0.997
Method:	Least Squares	F-statistic:	614.7
Date:	Sun, 03 May 2020	Prob (F-statistic):	0.0257
Time:	14:56:34	Log-Likelihood:	-15.628
No. Observations:	3	AIC:	35.26
Df Residuals:	1	BIC:	33.45
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	110.0250	110.296	0.998	0.501	-1291.416	1511.466
NR_REPORTING	72.2493	2.914	24.794	0.026	35.224	109.275

Omnibus:	nan	Durbin-Watson:	1.699
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.506
Skew:	-0.670	Prob(JB):	0.777
Kurtosis:	1.500	Cond. No.	94.3

There seem to be correlations between those three variables. Most of them are 1.00 but the colors are different. It means that although they are shown as 1.00, the actual numbers would be 0.999.
Then, we moved on to the regression analysis. The dependent variable is Deaths and the independent variable is the number of hospitals in each region.

The p-value of NR_REPORTING is 0.026, which means this variable is significant because it is less than 0.05. However, we had only 3 rows so it is difficult to conclude there is a robust relationship between the number of deaths and hospitals.
We then expanded the scope of our analysis to see more closely how the number of hospitals affects mortality using

those data by countries in the EU. Instead of the number of hospitals, we used the number of beds in each country, and added the data of the number of doctors, nurses, the amount of health spendings. To analyze with different countries, the number of those variables were converted to the number per 1000 inhabitants.
The picture below is the result of the regression analysis between the deaths per 1000 inhabitants and the variables mentioned above.

Dep. Variable:	Deaths_1000	**R-squared:**	0.361
Model:	OLS	**Adj. R-squared:**	0.147
Method:	Least Squares	**F-statistic:**	1.691
Date:	Sun, 03 May 2020	**Prob (F-statistic):**	0.216
Time:	14:56:40	**Log-Likelihood:**	6.4048
No. Observations:	17	**AIC:**	-2.810
Df Residuals:	12	**BIC:**	1.356
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5949	0.336	1.769	0.102	-0.138	1.328
HEALTHEXP	0.0002	0.000	1.684	0.118	-5.15e-05	0.000
HOSPITALBED	-0.0368	0.027	-1.359	0.199	-0.096	0.022
MEDICALDOC	-0.0914	0.085	-1.077	0.303	-0.276	0.093
NURSE	-0.0367	0.034	-1.082	0.301	-0.111	0.037

Omnibus:	4.131	**Durbin-Watson:**	2.393
Prob(Omnibus):	0.127	**Jarque-Bera (JB):**	2.265
Skew:	0.876	**Prob(JB):**	0.322
Kurtosis:	3.358	**Cond. No.**	1.88e+04

Dep. Variable:	Recovered_1000	**R-squared:**	0.246
Model:	OLS	**Adj. R-squared:**	-0.005
Method:	Least Squares	**F-statistic:**	0.9795
Date:	Sun, 03 May 2020	**Prob (F-statistic):**	0.455
Time:	14:56:40	**Log-Likelihood:**	-27.065
No. Observations:	17	**AIC:**	64.13
Df Residuals:	12	**BIC:**	68.30
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4530	2.409	0.188	0.854	-4.796	5.702
HEALTHEXP	0.0007	0.001	0.943	0.364	-0.001	0.002
HOSPITALBED	-0.0343	0.194	-0.177	0.862	-0.457	0.388
MEDICALDOC	-0.3227	0.608	-0.531	0.605	-1.647	1.002
NURSE	0.0293	0.243	0.120	0.906	-0.501	0.559

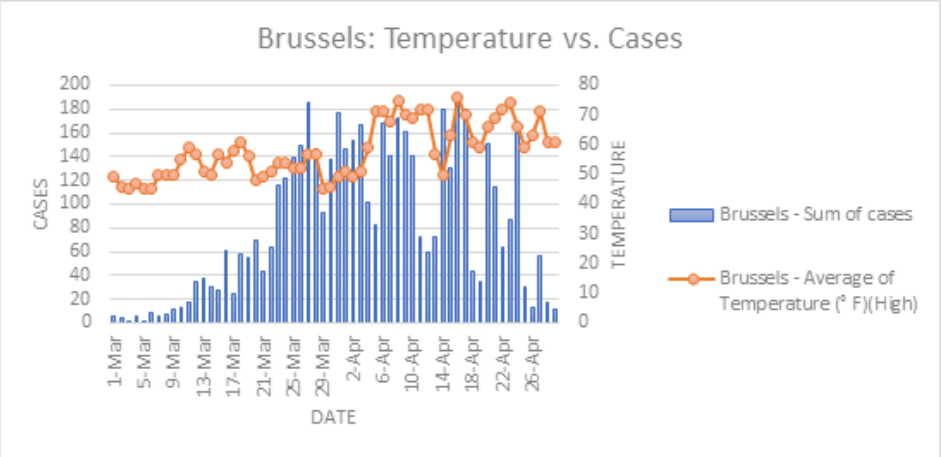
Omnibus:	11.033	**Durbin-Watson:**	1.548
Prob(Omnibus):	0.004	**Jarque-Bera (JB):**	7.858
Skew:	1.329	**Prob(JB):**	0.0197
Kurtosis:	5.006	**Cond. No.**	1.88e+04

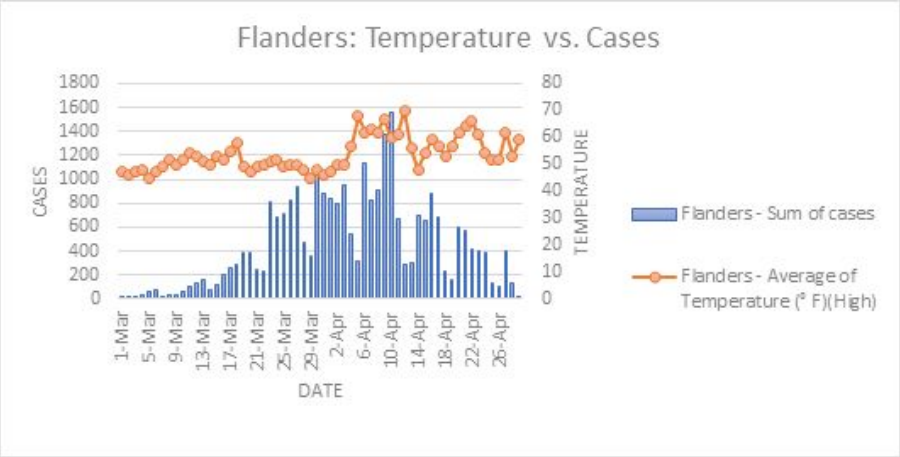
Although we saw the correlation in the analysis of data in Belgium, all of the variables in this analysis are not significant. Also, we changed the dependent variable from the number of deaths to the number of recovered people, but the result was almost the same.
We can conclude that there is no clear relationship between health facilities and resources and the number of deaths. In order to reveal what factors have an effect on the number of deaths, we have to consider many other things such as environment, habits, demographic characteristics, etc.

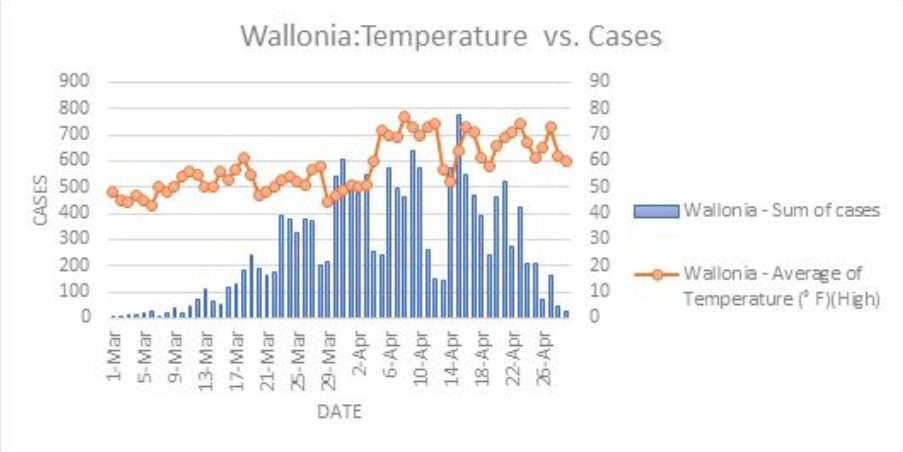
Hypothesis 4

The potency of the coronavirus decreases in higher temperatures.

Analysis:





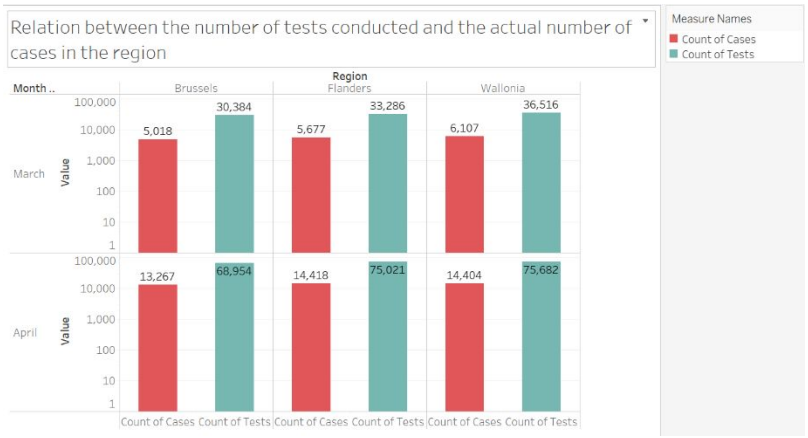


We can see that the temperature of the region does not affect the rate of infection from the coronavirus.
In conclusion, the infection through coronavirus might be affected due to other factors like:

- Age of the person.
- Medical condition.
- Recent travel history.

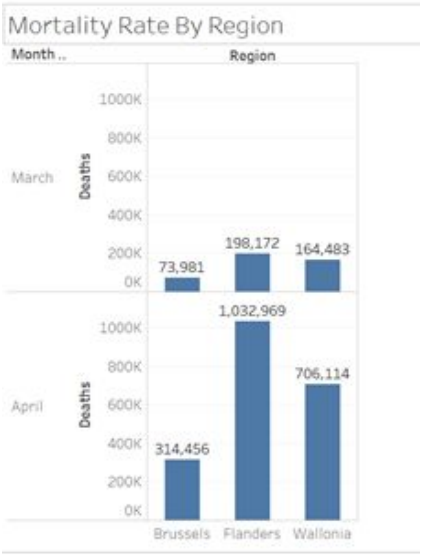
Hypothesis 5

Effects of increase in number of tests



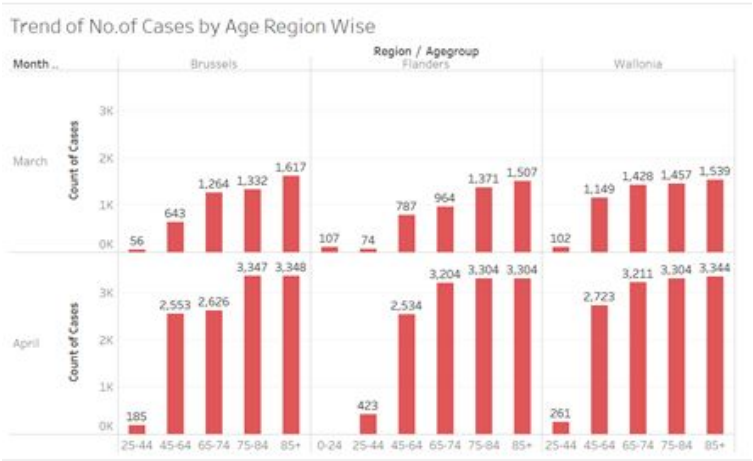
Though the number of tests conducted are significantly greater than the number of cases reported, it does not seem to have a negative correlation with the number of cases reported.

Checking for the most affected Region



As we can see from Flanders seems to be most affected by the virus having a very high increase in the mortality rate from March to April.
All the regions have an alarming increase; however, Flanders tops the list during both the months.
Considering the number of deaths alone we can say that Flanders has a very high infection rate.

Verifying if the same age groups are affected similarly across the regions.



The trend seems to be uniform. The senior population is the most affected in all the regions.
In April it can be observed that in Flanders, there are no cases reported who are younger than 25.
This is a significant decrease from the previous month.