

1.2-exploratory_analysis_socioeconomic

October 21, 2020

1 Imports

```
[1]: import sys
import os

import matplotlib as mpl

[2]: project_path = os.path.abspath(os.path.join('.'))

if project_path not in sys.path:
    sys.path.append(f'{project_path}/src/visualizations/')

from covid_data_viz import CovidDataViz
```

2 Setup

```
[3]: mpl.rcParams['figure.figsize'] = (9, 5)
```

3 Goal

My goal is to visualize various aspect of the COVID-19 pandemic.

4 Data sources

In this notebook I use data from the following sources: - <https://github.com/CSSEGISandData/COVID-19> - JHU CSSE COVID-19 Data. - [GDP per capita PPP](#) - The World Bank. - [Population](#) - The World Bank. - [Rural population](#) - The World Bank. - [Life expectancy at birth](#) - The World Bank. - [Current healthcare expenditure](#) - The World Bank. - <https://datahub.io/JohnSnowLabs/country-and-continent-codes-list> - country codes and continents.

5 Data loading

```
[4]: cdv = CovidDataViz()
```

Yemen is an outlier and is excluded from the analysis.

6 Socioeconomic data.

To enhance the analysis I used data available freely at from the **World Bank**. In this part of the analysis I use the last available value for each country. This is a reasonable thing to do given that these specific do not undergo wild fluctuations from year to year.

7 Correlation matrix

Note that **Rural population %** and **Cases per mln** have a correlation of -0.46% . Possible reasons could be the virus has a harder time spreading in scarcely populated countries. Note also that **GDP Healthcare** and **Dead per mln** have a positive correlation of 0.38 . This could be considered an expected result given that excluding **Asia** most of the countries around the world do not have experience in dealing with such matters.

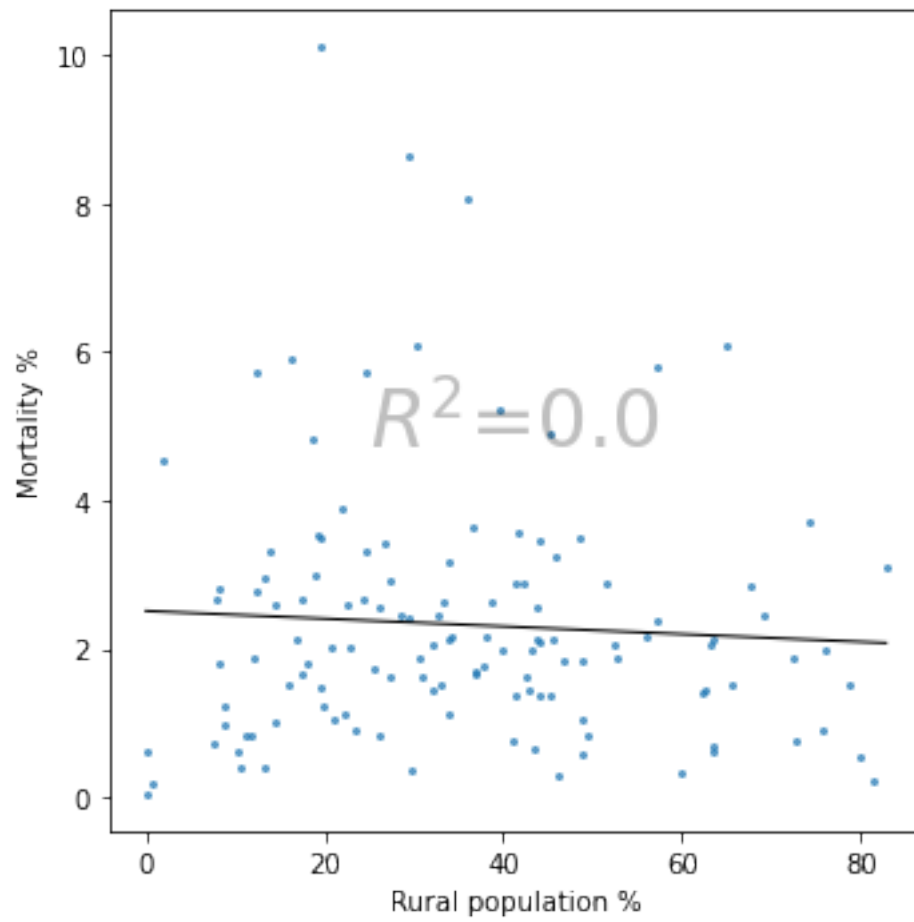
```
[5]: cdv.show_corr_mat()
```

```
<pandas.io.formats.style.Styler at 0x7f2468577280>
```

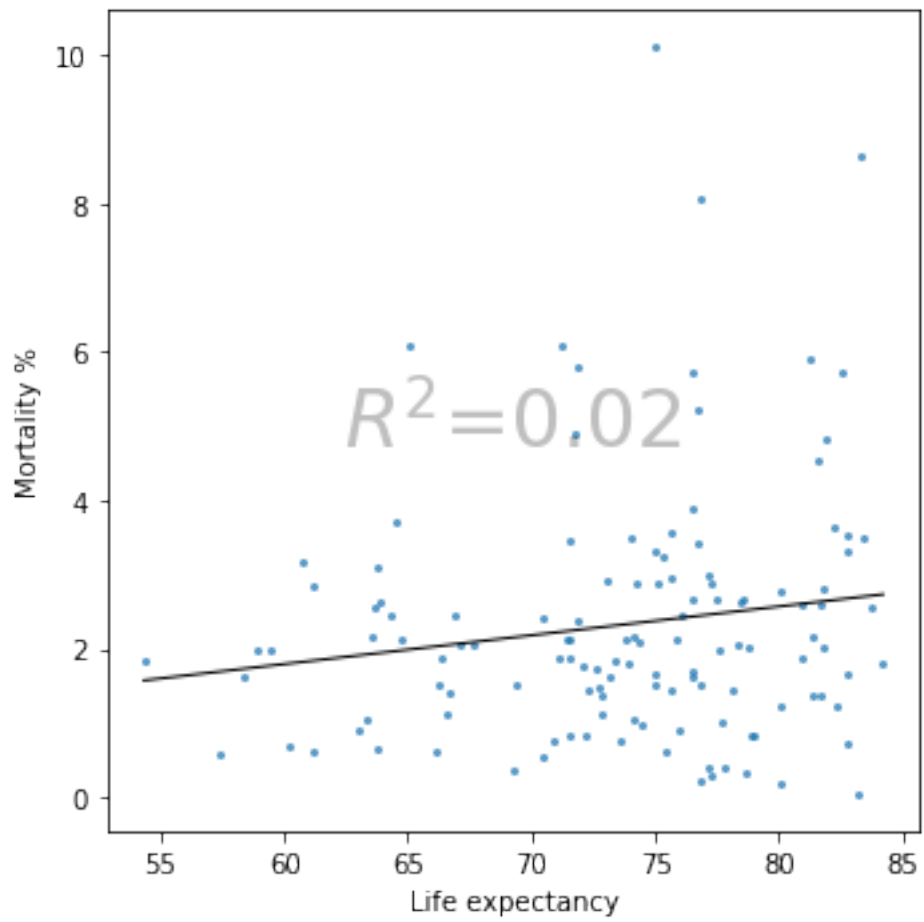
8 Mortality

As a reminder note that $\text{mortality} = \text{dead} / \text{confirmed}$. Observe that the **Life expectancy** vs. **Mortality %** scatter plot has a few nasty outliers to right of the plot. One possible explanation is that people from the risk group are more prevalent in countries with higher **life expectancy**. In the scatter plots we do not find any nice linear relationships. Perhaps examining the data on a per continent level would yield more fruitful results.

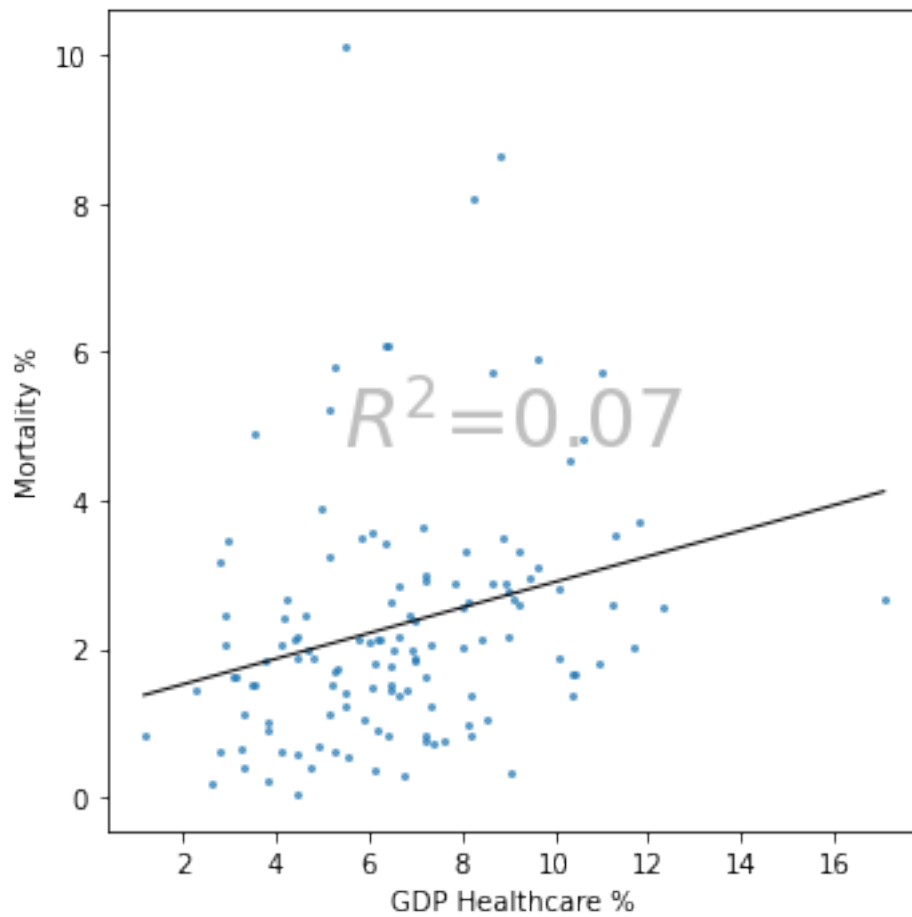
```
[6]: cdv.plot_with_slope('Rural population %', 'Mortality %')
```



```
[7]: cdv.plot_with_slope('Life expectancy', 'Mortality %')
```



```
[8]: cdv.plot_with_slope('GDP Healthcare %', 'Mortality %')
```



```
[9]: cdv.plot_with_slope('GDP per capita', 'Mortality %')
```

