**CAPSTONE PROJECT # 2 BY LISA HAHN-WOERNLE, PHD**

**Problem Statement**

**Are diseases of affluence making societies more perceptible to COVID-19? And can health and age demographics indicate the susceptibility of a country to COVID-19?**

**Context**

[Diseases of affluence](https://en.wikipedia.org/wiki/Diseases_of_affluence) correspond to diseases and other health conditions commonly associated with an increasing wealth in society. Examples are: type 2 diabetes, cardiovascular diseases and obesity. The statistics of the COVID-19 cases show great discrepancies among countries globally and also throughout the western world. Many factors are assumed to have an impact on the spread of the virus, such as the health system and the political measures taken to prevent the spread. In this project, the spread of the COVID-19 virus is analyzed in the context of diseases of affluence as well as the age and gender structure within a society.

**Criteria of Success**

The vast differences in COVID-19 cases and fatalities, not only globally but also throughout the western world indicate that there must be differences in societal structures affecting the spread and impact of the disease. In an attempt to identify these differences in the health and age demographics of a country, different analytical methods, such as correlations and predictive modeling, will be tested. A model or indicator will be classified as predictive or correlated if the predictability lies above 80 %.

**Scope of Solution Space**

The scope of this study is to identify societal factors that affect the susceptibility of a country to COVID-19. One important application of the gained insights could be targeted preventive actions for the current and potentially future control of the spread. For example, if a strong correlation between cardiovascular diseases and the cases of COVID-19 infections is revealed, these groups should be especially protected and be the first to be vaccinated once a vaccine is available. Another valuable application could be a susceptibility prediction for countries, if not regions, that have not yet been tested or are at the beginning of their testing. This could serve as an early warning system and help target preventive measures to potential “hot spots”.

**Constraints**

As with many other data sets, the differences in the data collection, such as the availability and intensity of tests, could introduce a high uncertainty and thereby error. Another limitation is the kind of available data: while COVID-19 statistics are available as a daily time-series, health and age demographics are at most annually available and not for the year 2020. The exploratory data analysis should therefore also address potential indicators for the temporal evolution and current stage of the spread, for example the time passed since x% of the population was infected or since the peak of the daily infections.

**Stakeholders**

Controlling the spread of COVID-19 is of global interest and each one of us is a stakeholders of such a data study. Especially countries that cannot afford wide spread testing or preventive measures due to time-limitations or economical reasons, will benefit from the outcome of this study. Targeting hot spots and high-risk groups can save precious time and means in the mission of protecting as many lives as possible.

**Data Sources**

**Daily COVID-19** statistics are sourced from the following GitHub repository: <https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports/05-02-2020.csv>

**Population** statistics by age and sex (5-year groups) are based on the Population Dynamics data of the UN Department of Economic and Social Affairs: <https://population.un.org/wpp/Download/Standard/CSV/>

Alternatively the **population** data set by the Population Reference Bureau can be used: https://datacenter.prb.org/download/international/indicator/population/csv

**Diseases of affluence:**

**Obesity** statistics are sourced from the “COVID-19 Healthy Diet Dataset“ available on Kaggle: <https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset>

Data on **cardiovascular disease** and **diabetes** is retrieved from “Our World in Data”: https://ourworldindata.org/grapher/cardiovascular-death-rate-vs-gdp-per-capita and https://ourworldindata.org/grapher/diabetes-prevalence

**Strategy Outline**

The study will be structured as follows:

1. Data wrangling will align data from different sources and clean it from potential outliers or false/missing data. Two datasets will be generated: one only containing countries for which COVID-19 data is available, and one containing demographic data of countries without COVID-19 statistics.
2. Exploratory data analysis will investigate correlations and features with tools like histograms, box plots, biplots and correlation heatmaps. Beside gaining a rough sense for the data set, central goals of this step are to sort out collinear features and identify the potential of measures for the temporal evolution of the spread. This step will also tell whether some features are more correlated to the spread than others, e.g. individual diseases of affluence.
3. In the pre-processing and data training development phase, different models and sets of data are tested for their predictability and performance compared to testing data. Models to be tested are for example logistic regression, random forest and k-nearest neighbors. The purpose of the model will be to predict the spread of COVID-19 in countries that are not yet tested.
4. In the Modeling phase, the “winner” of step 3 will be used to predict COVID-19 statistics for the 2nd data set, the one missing the COVID-19 features. Previous error analysis will allow us to estimate the possibility and error of the resulting analysis.

**Deliverables**

Key components of the deliverables will be, next to this document: a set of four Jupyter Notebooks addressing the four analysis steps above, a read-me file describing the data set, a final project report and a slide deck present the main findings.