Creating Evidence in a Real-World Health Data Science Scenario

Dogan Parlak

May 2022

1 Abstract

In order to learn from the Covid-19 to make better decisions next time, four exploratory data analysis are done and a machine learning model is trained. Firstly, before the first lock-down announced in Switzerland, the validated hypothesis is that increase in total number of cases in the most populous four cantons are exponential. Secondly, the validated hypothesis is that there is a direct proportion between number of cases in each canton and their population. Furthermore, validated hypothesis is that deaths due Covid-19 over time is surplus for older people compared to younger generations. Finally, rejected hypothesis is that proportions of death due to Covid-19 on people with varying number of vaccination doses is not dependent on gender. It is observed that males tend to die more compared to females. A second degree polynomial regression model is trained in order to predict total number of cases by looking at the population of a certain region. The Rˆ2 of the model is 0.996 so the model fits the data good enough.

2 Introduction

The aim of this work is to gain unknown knowledge and create evidence on about Swiss COVID-19 situation. The task is to find interesting patterns and validating them through hypothesis generation. Basically, the hypothesis are validated by visualizations and supported by numerical outcomes. In the first hypothesis, it is validated that there is an exponential increase in total number of cases in the most four populous cantons. For this purpose, Date of the data, number of cases in cantons and region names are utilized. What is more, another validated hypothesis suggests that there is a direct proportion between cantons population and number of Covid-19 cases in that region. This analysis is done by using number of cases for each canton, canton names and population of these cantons. Thirdly, the hypothesis that as people get older, they tend to die more compared to younger generations over time is validated. Date of the data, age intervals and total number of deaths in Switzerland for these age categories are utilized in analysis. Lastly, rejected hypothesis points out that there is no distinction between genders in terms of death proportions due to Covid-19 when the vaccination status is taken into account. This analysis requires vaccination statuses, genders and total number of deaths for each status. On the other side, a second degree polynomial regression model is trained. By looking at the population of a region, the model tries to estimate the total number of Covid-19 cases. Rˆ2 is used to estimate the quality of fit. For this analysis number of cases in cantons, canton names and population of regions are utilized.

3 Background and Related Work

3.1 Exponential increase in total number of cases

- People mistakenly perceive the coronavirus to grow in a linear manner, underestimating its actual potential for exponential growth [1]. Participants guessed the total number of coronavirus cases over the past 5 d, from Tuesday, March 17 to Saturday, March 21 [1]. As expected, participants displayed exponential growth bias. Although some participants accurately included exponential growth in their estimates, thus producing an overall significant quadratic trend (F = 18.78, P ¡ 0.0001). [1].

- This research uses exponential growth modelling studies to understand the spreading patterns of SARS-CoV-2 and identifies countries that showed early signs of containment until March 26, 2020 which is 6 days after the first lock down announced in Switzerland [7]. Epidemics are assumed to have an exponential growth at an early stage, and the number of infections reduces over time due to interventions like lock downs [7]. The study used a machine learning and exponential growth model [7]. The variables used as part of the predictive mode were: doctors per 1,000 population, beds per 1,000 population, average temperature, average humidity, days since official lockdown, percentage of lockdown days, total cases per million population, deaths per million population, days since the first contact, and percentage of serious cases of infection [7]. The countries that showed initial signs of containment were Austria, Chile, China, Czechia, Denmark, Greece, Iceland, Indonesia, Iran, Ireland, Israel, Japan, Netherlands, Norway, Peru, Poland, Portugal, Qatar, Republic of Korea, Saudi Arabia, Singapore, Slovenia, **Switzerland**[7].

3.2 Relation between Covid-19 cases and population of cantons

- Extreme public health interventions play a critical role in mitigating the local and global prevalence and pandemic potential [10]. Here, I use population size for pathogen transmission to measure the intensity of public health interventions, which is a key characteristic variable for nowcasting and forecasting of COVID-19 [10]. Recently, several epidemiological studies show that the population size is a crucial transmissibility factor [10]. Note that the evolution of an epidemic is a process of gradual spread [10]. The larger the population size N, the higher the epidemic peak [10].

3.3 Relation between deaths due to Covid-19 over time for certain age intervals

- Population fatality rates were analyzed and defined as the number of COVID-19 deaths divided by the total number of persons in the respective age groups [2]. Separate analyses were conducted per country[2]. In Model 1 COVID, the total numbers of COVID-19 deaths in 2020 were analyzed in a model with the metric variable age (centered at 65 years old and scaled in decades), the variable male sex (with female as the reference category), and the interaction age*×*sex with population size as an offset in each age by sex group [2]. The same model was used to analyze all-cause mortality in the individual countries (Model 1 ALL) [2]. However, the percentage of increase in COVID-19 mortality per 10 years of increasing age was generally larger than the percentage of increase of all-cause mortality over 10 years of age in the countries before the pandemic[2].

3.4 Proportions of death due to Covid-19 across genders with varying number of vaccination doses

- Following table represents data of deaths within 7 days after Covid-19 vaccination. It can be observed that male deaths are more than female deaths [8].

Deaths within 7 d after COVID-19 vaccination per 100 000 persons					
	United States				Taiwan
	50–59 y	60–64 y	65–79 y	80+ y	75+ y
Female	0.80	1.17	2.15	7.44	NA
Male	1.06	2.20	3.44	11.09	NA
Total	0.92	1.65	2.76	8.90	19.83
Background deaths per 7 d per 100 000 persons					
	United States				Taiwan
	50–59 y	60–64 y	65–79 y	80+ y	75+ y
Female	8.76	15.37	33.76	95.13	108.75
Male	14.52	25.53	50.23	128.17	152.10
Total	11.58	21.33	41.38	109.14	127.39

COVID = coronavirus disease; NA = not available.

Figure 1: Deaths due to Covid-19 after vaccination across genders

- As COVID-19 has swept across the globe, it has killed many more men than women[9]. Some have suggested that biological factors are driving the difference[9]. But researchers at Harvard's GenderSci Lab—including several students from Harvard T.H. Chan School of Public Health—think that social factors may be playing the largest role [9].

4 Methods

4.1 Exponential increase in total number of cases

First significant and general precaution against Covid-19 virus was lockdown in Switzerland. The hypothesis is as follows: Does total number of cases in the top four populous cantons display an exponential increase until the first lockdown in Switzerland announced?

First of all, the first lockdown date is determined and it is 20.03.2020. Only the data up to this date is used. Further, most populous cantons are determined and these are (Zurich) ZH, (Bern) BE, (Vaud) VD and (Aargau) AG. In addition to those, the behaviour of the increase is displayed for Switzerland in general by using the aggregate data.

The x-axis represents the date of the data recorded. The y-axis represents the number of cases. Finally colors are mapped to cantons in order to distinguish each four of them and observe their characteristic. A legend is used to indicate the matching between colors and cantons. In order to display the increase for Switzerland, only the data with abbreviation CH is taken into account. Hence, there is no need to use colors for mapping.

4.2 Relation between Covid-19 cases and population of cantons

It is known that Covid-19 virus in general spreads from person to person. Therefore, the following hypothesis is generated: Do population of cantons are directly proportional with total number of Covid-19 cases occurred in that particular canton?

First of all, names of cantons, their corresponding population and Covid-19 cases are extracted according to the latest day in the data. Since the Covid-19 cases are stored in a cumulative manner, data with the latest day contains the total cases for that canton. Afterwards, population of the cantons are sorted to ease the comparison with cases.

Colors are used to indicate the population size of cantons. The y-axis represents the cases whereas the x-axis represents the cantons. Bar plot is preferred to compare the cases utilizing the height of the bars. When the color is close to yellow, it means the population is low compared to the ones close to red.

4.3 Relation between deaths due to Covid-19 over time for certain age intervals

People with older ages tend to affect more compared to younger people. Therefore, the following hypothesis is generated: Are older people dying more from Covid-19 than young people?

A 2D array is used with size (age intervals, time) in order to store the number of deaths for each interval over time. The deaths are obtained from overall Switzerland data so not canton based.

The y-axis is used to represent the age intervals which are 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+. The x-axis is used to represent the time. Time 0 means the first day in the data and so on. Colors are used to indicate the number of deaths due to Covid-19. Brighter the color, more the deaths due Covid-19.

4.4 Proportions of death due to Covid-19 across genders with varying number of vaccination doses

The hypothesis is that Covid-19 does not cause the death of one gender more than the other regardless of number of Covid-19 vaccination status.

There are four main categories for vaccination status. These are fully vaccinated with one booster, fully vaccinated, partially vaccinated, not vaccinated. Data stores death of people due Covid-19 for overall time. For each of the vaccination status, sum of deaths due Covid-19 are stored separately for males and females. After that, the proportion of each of them are calculated such that sum of the proportions for a single vaccination status is equal to 1.

Horizontal barplot is utilized to represent the proportions. Blue and red colors are used to represent the males and females respectively. The y-axis is used to represent the vaccination status and in each row, the corresponding proportion is put inside the rectangles.

4.5 Polynomial regression to estimate the number of Covid-19 cases according to population of cantons

In order to observe the relation between population and number of cases, degree 2 polynomial regression is applied based on the visualization done for relation between Covid-19 cases and population of cantons. Two of the cantons Appenzell Ausserrhoden (AR) and Luzern (LU) are separated for testing purposes. Rest of the cantons are used for training with their population information as features. Afterwards using this fitted models coefficients, total number of Covid-19 cases in AR and LU are predicted.

Blue dots represents the train data whereas red dots are test data. Abbreviation of cantons are put on the dots to indicate the canton names. There are some overlaps since some cantons have similar number of Covid-19 cases. In order to reduce this situation, I have converted the x-axis to logarithmic level which represents the population. On the other hand, the y-axis represents the number of cases.

5 Data and Software

- **Tools:** For this task, python programming language is used with pandas, numpy and matplotlib libraries to analyze and visualize the data.

Following are the datasets used for analysis.

5.1 Daily record timelines by regions for cases

This dataset is used to extract columns which contains date, canton and number of cases information. It is further used to analyze the increase trend in Covid-19 cases over time in the most four populous cantons and in Switzerland [3].

- **Date:** This column contains strings indicating the date in which data is reported. In the analysis, I have utilized the data until the first lockdown announced in Switzerland.
- **Number of Cases:** This column contains integers indicating the number of Covid-19 cases reported.
- **Regions:** This column contains the regions as string in which the data is collected from. In the analysis, the most four populous cantons and Switzerland's overall data is utilized. These regions are stored as : ZH, BE, VD, AG and CH.

5.2 Daily record timelines for hospitalisations of fully vaccinated persons by vaccine

This dataset is used to extract columns which contains cantons, number of cases and populations of cantons. It is further used to analyze the relation between the population of cantons and total number of Covid-19 cases [4].

- **Number of Cases:** This column contains integers indicating the number of Covid-19 cases reported. I have aggregated this column for each canton by summing the cases for all dates.
- **Regions:** This column contains the regions as string in which the data is collected from. In the analysis, all 26 cantons are included.
- **Population of regions:** This column contains the population of each regions as integers. In the analysis, all 26 cantons population information are included.

5.3 Iso-Week record timelines by geoRegion and age brackets for tests

This dataset is used to extract the date, age-intervals and total number of deaths due to Covid-19. It is further used to analyze the relation between age-intervals and deaths during the pandemic [5].

- **Date:** This column contains strings indicating the date in which data is reported. In the analysis, I have utilized the whole date information.
- **Age Interval:** This column contains strings indicating the age-interval of people. In the analysis, I have utilized the following intervals: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+.
- **Total Number of Deaths:** This column contains integers indicating the number of deaths reported in the given data.

5.4 Iso-Week record timelines for deaths of fully vaccinated persons by vaccine and sex

This dataset is used to extract the vaccination status and, sex and total number of deaths due to Covid-19 [6].

- **Vaccination Status:** This column contains strings indicating the type of vaccination doses. In the analysis I have utilized fully vaccinated, partially vaccinated, fully vaccinated with one booster and not vaccinated types.
- **Sex:** This column contains strings indicating the sex of individuals for each data. In the analysis, I have utilized male and female as types.
- **Total Number of Deaths:** This column contains integers indicating the number of deaths reported in the given data. In the analysis, I have utilized these information to compare the proportion of deaths across males and females.

6 Results

Following are the exploratory data analysis results with the hypothesis formed accordingly.

6.1 Exponential increase in total number of cases

Following is the change in Covid-19 cases over time in Switzerland until first lockdown announced in Switzerland.

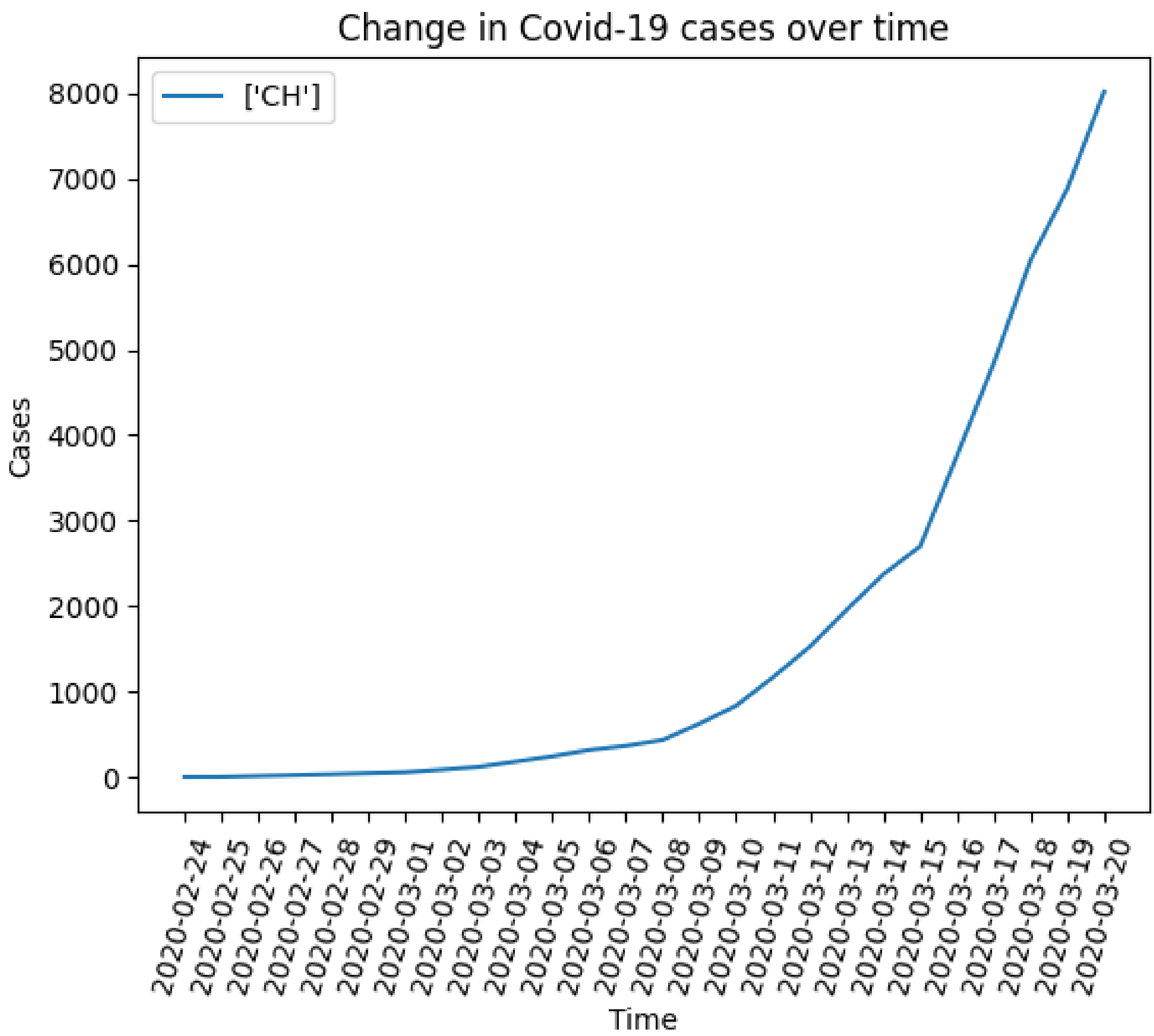

Figure 2: Covid-19 cases over time in Switzerland

Following is the change in Covid-19 cases over time in the most four populous cantons of Switzerland which are Zurich (ZH), Bern (BE), Vaud (VD), Aargau (AG).

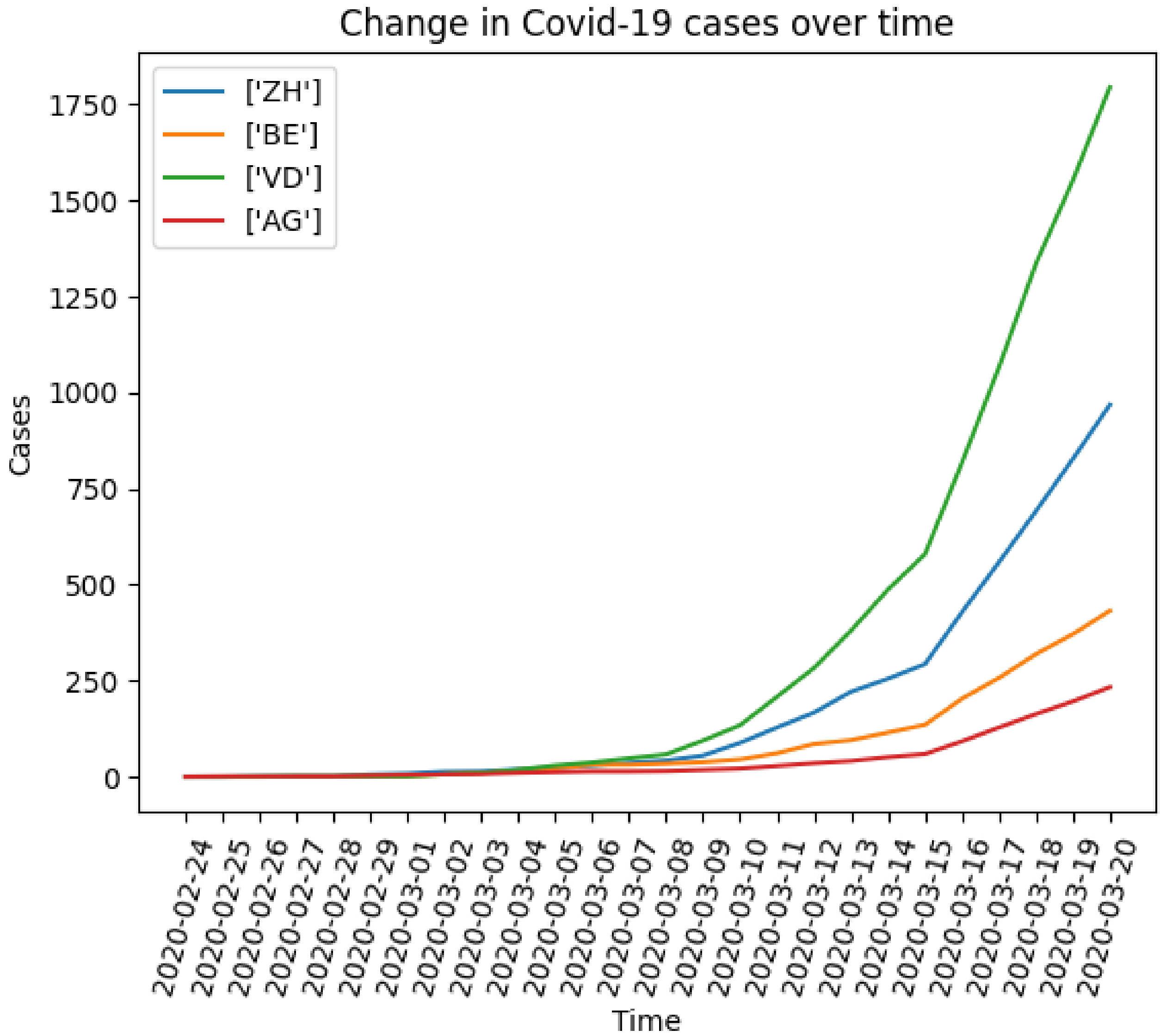

Figure 3: Covid-19 cases over time in Zurich, Bern, Vaud, Aargau

6.2 Relation between Covid-19 cases and population of cantons

Following is the total number of Covid-19 cases for each canton recorded between 24.02.2020 - 03.05.2022 together with the population of these cantons indicated by color map.

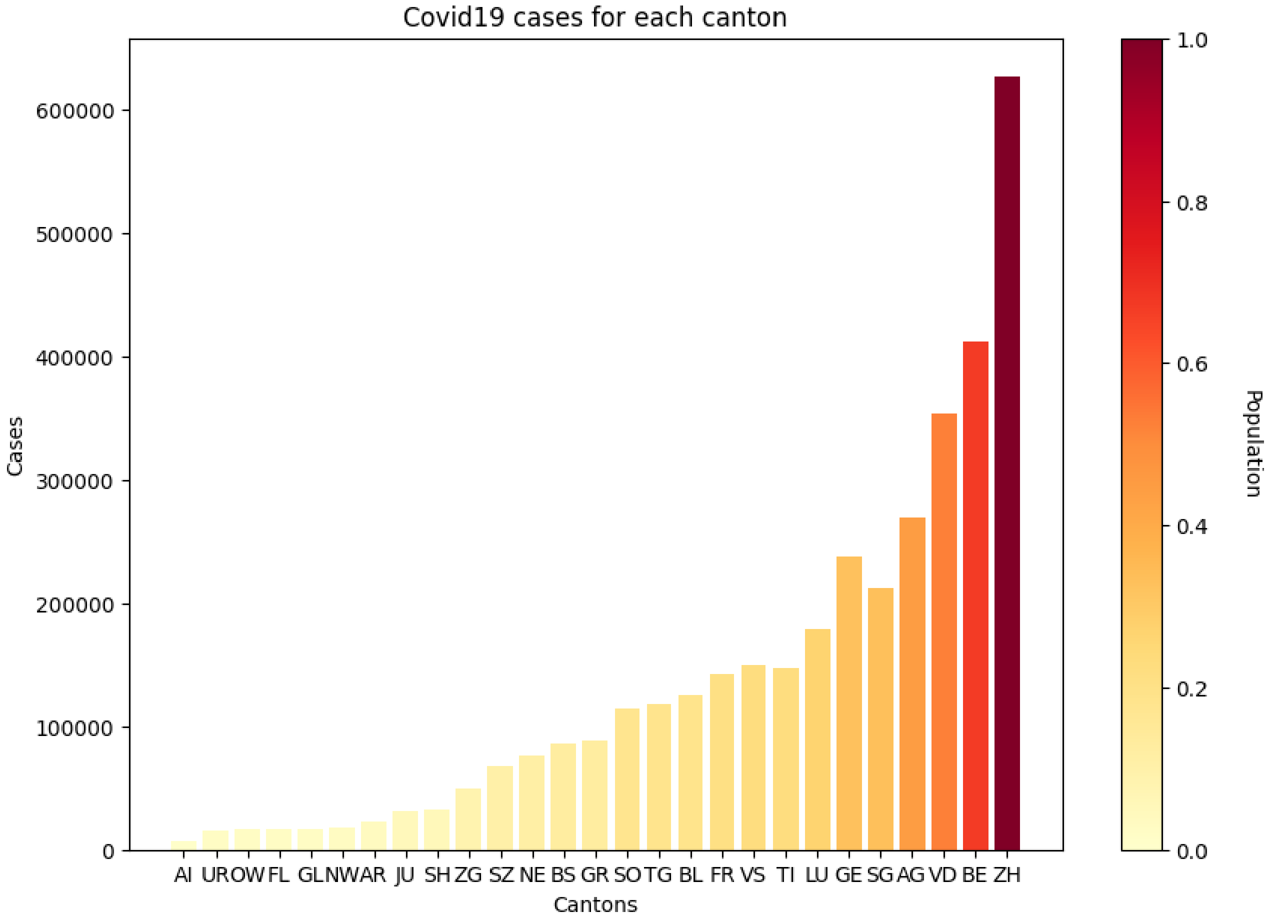

Figure 4: Covid-19 cases for each canton together with their population indicated

6.3 Relation between deaths due to Covid-19 over time for certain age intervals

Following is the change in number of deaths due to Covid-19 over time for certain age intervals. Deaths are indicated by color map. Data is collected for all deaths due to Covid-19 in Switzerland between 24.02.2020 - 20.04.2020.

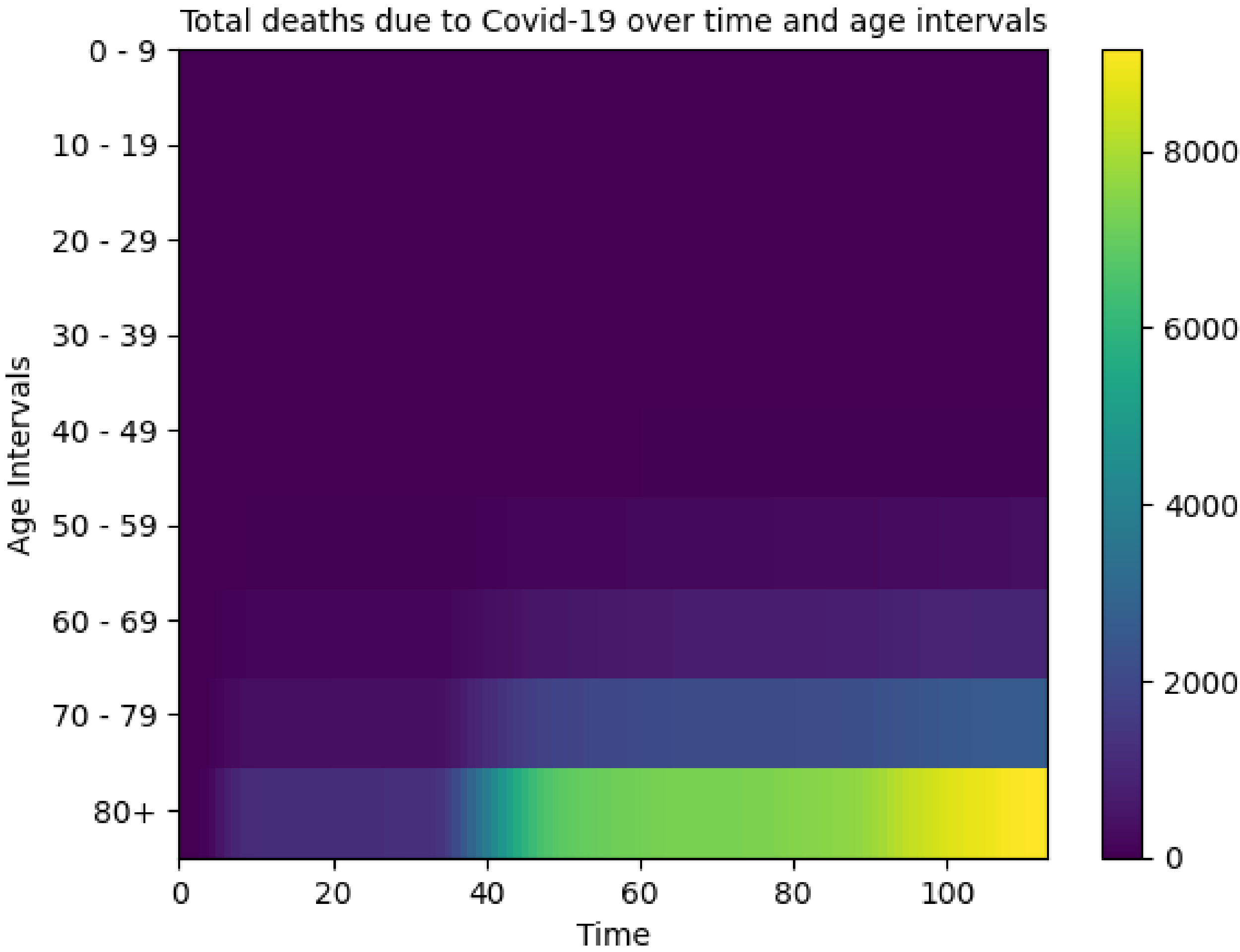

Figure 5: Covid-19 cases for each canton together with their population indicated

6.4 Proportions of death due to Covid-19 across genders with varying number of vaccination doses

Following is the proportions of death due to Covid-19 across males and females who have different number of Covid-19 vaccination doses. Data is collected between 25.01.2021 - 25.04.2022. Four different dose types are as follows: Not vaccinated, partially vaccinated, fully vaccinated, fully vaccinated with one booster.

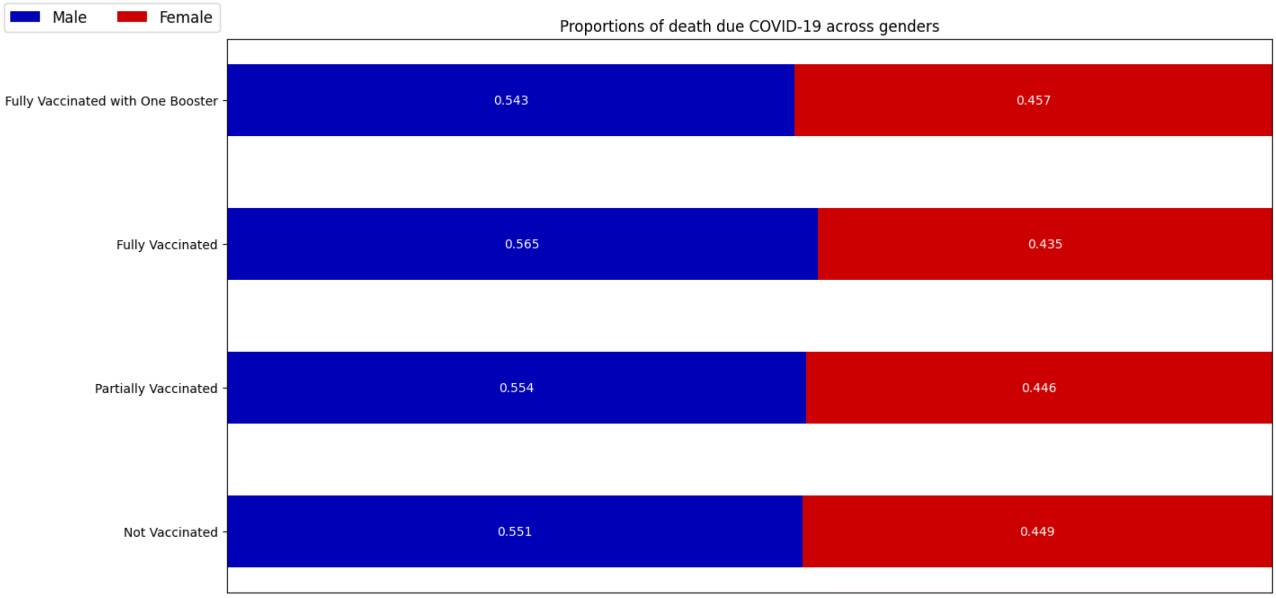

Figure 6: Proportions of death due to Covid-19 across genders with varying number of vaccination doses

Vaccination Status	Male	Female
Fully Vaccinated with One Booster	0.543	0.457
Fully Vaccinated	0.565	0.435
Partially Vaccinated	0.554	0.446
Not Vaccinated	0.551	0.449

6.5 Polynomial regression to estimate the number of Covid-19 cases according to population of cantons

Following is the visualization of the polynomial regression model which utilizes total number of Covid-19 cases for each canton recorded between 24.02.2020 - 03.05.2022 as label and population of each canton as feature space. Blue dots are used to train the model whereas red dots (2) used to test the model. Rˆ2 of the fit is obtained as 0.996. Second degree polynomial is utilized to fit.

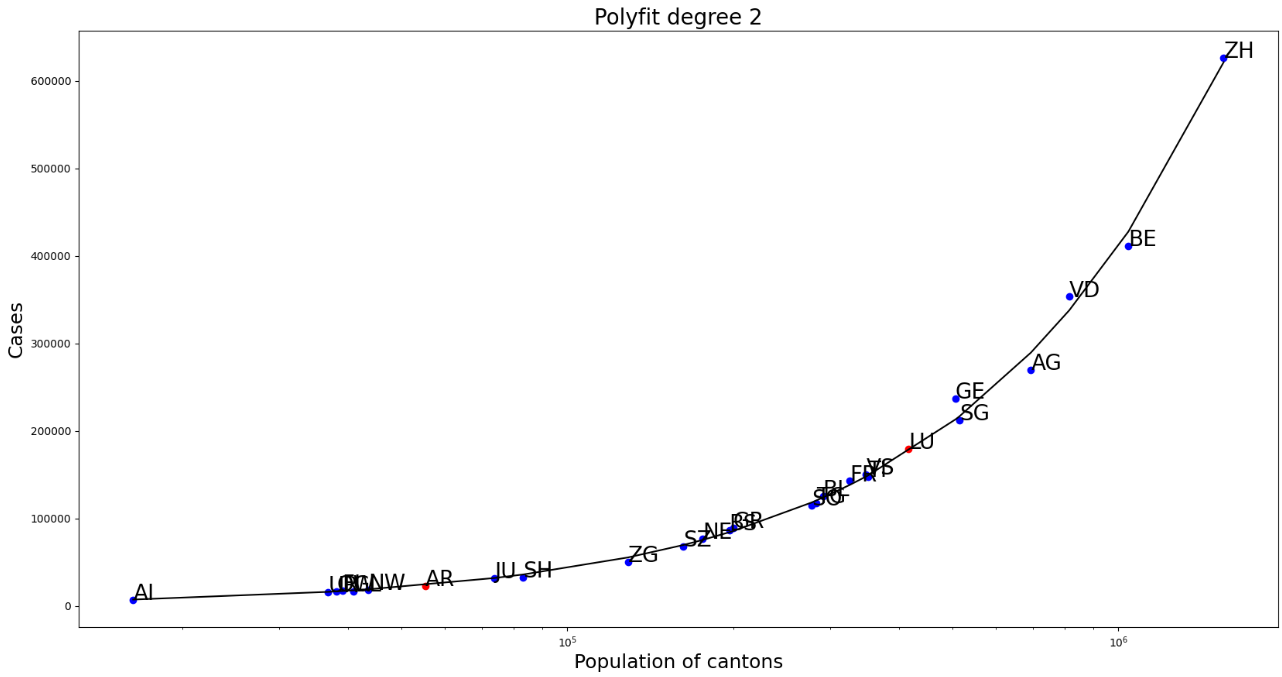

Figure 7: Polynomial regression result

7 Discussion

7.1 Exponential increase in total number of cases

My findings are consistent with the related works. There is a significant quadratic trend observed. Researchers uses exponential growth modelling in their studies until March 26, 2020 which verifies that the exponential behavior I have hypothesised earlier and discovered afterwards. As the population of region increases the exponential behavior becomes more apparent. Since there was no precaution against Covid-19 at its first times and people do not have enough information about this pandemic, exponential growth bias was inevitable. The hypothesis suggesting that increase in total number of cases in the most populous four cantons are exponential is validated.

7.2 Relation between Covid-19 cases and population of cantons

As several epidemiological studies display and my findings match with, population size is a crucial transmissibility factor. Considerin that evolution of an epidemic is a process of gradual spread, it is no suprise to obtain such a proportional increase in number of cases when the cantons are sorted according to their populations. There are a few exceptions such as Geneva and St Gallen. This can be acceptable since their population sizes are very close to each other (506.343 and 514.505 respectively). The hypothesis suggesting that there is a direct proportion between number of cases in each canton and their population is validated.

7.3 Relation between deaths due to Covid-19 over time for certain age intervals

The findings basically points out that the percentage of increase in COVID-19 mortality per 10 years of increasing age was generally larger than the percentage of increase of all-cause mortality over 10 years of age. Part of this outcome is partially consistent with my results. I divided the age intervals into 10 years categories. The distinction between total deaths begin from the interval 50-59 and becomes visible even more as ages get older. The hypothesis indicating that deaths due Covid-19 over time is surplus for older people compared to younger generations is validated.

7.4 Proportions of death due to Covid-19 across genders with varying number of vaccination doses

There is no specific research that is inline with my findings. There is a study suggesting that 7 days after Covid-19 vaccination, proportion of death due Covid-19 is more for males than females. I found that for all vaccination status categories, males tend to die more due Covid-19 than females. The hypothesis suggesting proportion of death due to Covid-19 on people with varying number of vaccination doses is not dependent on gender is rejected.

7.5 Polynomial regression to estimate the number of Covid-19 cases according to population of cantons

One can observe that second degree polynomial model fits the data good enough considering the R^2 of the fit which is 0.996. This analysis demonstrates that population of a canton is a useful feature that be accounted to forecast total number of Covid-19 cases in a certain region.

8 Limitations

8.1 Exponential increase in total number of cases

At the early stages of Covid-19, it was not easy to apply valid and enough number of cases to people with symptom. Therefore, the analysis assumes that this factor does not significantly effect the behavior of the increase.

8.2 Relation between Covid-19 cases and population of cantons

Since the data with the latest day is utilized, there can be minor changes in the population of cantons. Thus, the analysis assumes that there is no major change difference in cantons populations during the Covid-19 period. Similarly, applied Covid-19 test amount and validity of them is assumed to be satisfactory.

8.3 Relation between deaths due to Covid-19 over time for certain age intervals

This analysis assumes that reasons of deaths are recorded correctly. In other words, people who died from Covid-19 are not recorded with a different cause of death and vice versa. Such an impact is assumed not to affect the overall outcome of the study.

8.4 Proportions of death due to Covid-19 across genders with varying number of vaccination doses

This analysis assumes that vaccination status for each individual is correctly classified. There should not be any major difference that could affect the outcome of the analysis. Otherwise, the proportions may vary for certain vaccination statuses.

9 Conclusions

In the analysis I have done, different aspects of Covid-19 are covered. Trend in total number of cases, relation between cases and populations, relation between deaths over time for certain age intervals, proportions of death due Covid-19 across genders with varying number of vaccination status and a polynomial regression to estimate the cases according to population are five major topics that are analyzed. Further works should focus on the limitations and try to minimize these effects since it is impossible for them to be eliminated completely. Based on my findings, in a pandemic situatuon, an immediate lock-down would be benefical to prevent the spread of disease. Regions where population is high and more males live than females are critical and should be controlled primarily. There should be special regulations for older generations.

A APPENDIX

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.cm import ScalarMappable
from matplotlib.patches import ConnectionPatch
from sklearn.metrics import r2_score

path = "sources-csv/data/COVID19Cases_geoRegion.csv"
```

```
path2 = "sources-csv/data/COVID19Hosp_vaccpersons.csv"
path3 = "sources-csv/data/COVID19Death_geoRegion_AKL10_w.csv"
path4 = "sources-csv/data/COVID19Death_vaccpersons_sex_w.csv"

def csv_to_df(path):
   return pd.read_csv(path)

df1 = csv_to_df(path)
before_lock_df = df1[(df1["datum"] <= "2020-03-20")]
top_four_df = before_lock_df[(before_lock_df["geoRegion"] == "CH") |  (before_lock_df["geoRegion"] == "
                              | (before_lock_df["geoRegion"] == "AG")]
top_four_c = ["ZH", "BE", "VD", "AG"]

for c in top_four_c:
    df_temp = top_four_df[top_four_df["geoRegion"] == c]
    plt.plot(df_temp["datum"], df_temp["sumTotal"], label = df_temp["geoRegion"].unique())

plt.title("Change in Covid-19 cases over time")
plt.xticks(rotation = 75)
plt.xlabel("Time")
plt.ylabel("Cases")
plt.legend(loc="upper left")
plt.show()

df_temp = before_lock_df[before_lock_df["geoRegion"] == "CH"]
plt.plot(df_temp["datum"], df_temp["sumTotal"], label = df_temp["geoRegion"].unique())
plt.title("Change in Covid-19 cases over time")
plt.xticks(rotation = 75)
plt.xlabel("Time")
plt.ylabel("Cases")
plt.legend(loc="upper left")
plt.show()

df_case_pop = df1[(df1["datum"] == "2022-05-03") & (df1["geoRegion"] != "CH")& (df1["geoRegion"] != "CH
df_case_pop = df_case_pop.sort_values(by=['pop'])

cantons = df_case_pop["geoRegion"]
populations = df_case_pop["pop"]
total_cases = df_case_pop["sumTotal"]

populations_n = [x / max(populations) for x in populations]
fig, ax = plt.subplots(figsize=(10, 7))
my_cmap = plt.cm.get_cmap('YlOrRd')
colors = my_cmap(populations_n)
rects = ax.bar(cantons, total_cases, color=colors)
sm = ScalarMappable(cmap=my_cmap, norm=plt.Normalize(0,max(populations_n)))
sm.set_array([])
cbar = plt.colorbar(sm)
cbar.set_label('Population', rotation=270,labelpad=25)
plt.title("Covid19 cases for each canton")
plt.xticks(cantons)
plt.xlabel("Cantons")
plt.ylabel("Cases")
plt.show()

populations = list(populations)
cantons = list(cantons)
total_cases = list(total_cases)

#TEST FOR ---AR--LU---
population_test =  []
```

```
population_test.append(populations.pop(6))
population_test.append(populations.pop(19))

cantons_test = []
cantons_test.append(cantons.pop(6))
cantons_test.append(cantons.pop(19))

cases_test = []
cases_test.append(total_cases.pop(6))
cases_test.append(total_cases.pop(19))

print(cantons_test)
print(population_test)
print(cases_test)

coefs = np.polyfit(populations, total_cases, 2)
plt.figure()
pred_train = np.polyval(coefs, populations)
pred_test = np.polyval(coefs, population_test)
r2 = r2_score(total_cases, pred_train)
print(r2)
plt.plot(populations, pred_train, color="black")
plt.title("Polyfit degree "+str(2), fontsize = 20)
plt.scatter(populations,total_cases, color = "blue")
plt.scatter(population_test, cases_test, color ="red")
for i in range(len(populations)):
    plt.text(populations[i], total_cases[i], cantons[i], fontsize = 20)

for i in range(len(population_test)):
     plt.text(population_test[i], cases_test[i], cantons_test[i], fontsize = 20)
plt.xscale("log")
plt.gcf().set_size_inches(20, 10)
plt.xlabel("Population of cantons", fontsize = 18)
plt.ylabel("Cases", fontsize = 18)
plt.show()

df3 = csv_to_df(path3)

labels = df3["altersklasse_covid19"].unique()[:-1]
dates = df3["datum"].unique()

data_2d = np.zeros((len(labels), len(dates)))
for l in range(len(labels)):
    for d in range(len(dates)):
        data_2d[l,d] = df3[(df3["altersklasse_covid19"] == labels[l]) & (df3["datum"] == dates[d]) & (d

fig, ax = plt.subplots()
shw = ax.imshow(data_2d, extent=[0,113, 8.5, 0],  aspect='auto')
plt.xlabel("Time")
plt.ylabel("Age Intervals")
ax.set_yticklabels(labels)
plt.title("Total deaths due to Covid-19 over time and age intervals", fontsize = 10)
plt.colorbar(shw)
plt.show()

df4 = csv_to_df(path4)

vaccination_status = ["fully_vaccinated_first_booster", "fully_vaccinated", "partially_vaccinated", "no
death_by_vacc = []
for v in vaccination_status:
    male_death = df4[(df4["vaccination_status"] == v) & (df4["sex"] == "male")]["sumTotal"].sum()
```

```
    female_death = df4[(df4["vaccination_status"] == v) & (df4["sex"] == "female")]["sumTotal"].sum()
    total = male_death + female_death
    death_by_vacc.append([round(male_death/total,3), round(female_death/total,3)])

category_names = ['Male', 'Female']
results = {
    'Fully Vaccinated with One Booster': death_by_vacc[2],
    'Fully Vaccinated': death_by_vacc[0],
    'Partially Vaccinated': death_by_vacc[1],
    'Not Vaccinated': death_by_vacc[3],
}

def survey(results, category_names):
    """
    Parameters
    ----------
    results : dict
        A mapping from question labels to a list of answers per category.
        It is assumed all lists contain the same number of entries and that
        it matches the length of *category_names*.
    category_names : list of str
        The category labels.
    """
    labels = list(results.keys())
    data = np.array(list(results.values()))
    data_cum = data.cumsum(axis=1)
    category_colors = plt.colormaps['seismic'](
        np.linspace(0.15, 0.85, data.shape[1]))

    fig, ax = plt.subplots(figsize=(15, 8))
    ax.invert_yaxis()
    ax.xaxis.set_visible(False)
    ax.set_xlim(0, np.sum(data, axis=1).max())

    for i, (colname, color) in enumerate(zip(category_names, category_colors)):
        widths = data[:, i]
        starts = data_cum[:, i] - widths
        rects = ax.barh(labels, widths, left=starts, height=0.5,
                        label=colname, color=color)

        r, g, b, _ = color
        text_color = 'white' if r * g * b < 0.5 else 'darkgrey'
        ax.bar_label(rects, label_type='center', color=text_color)
    ax.legend(ncol=len(category_names), bbox_to_anchor=(0, 1),
              loc='lower right', fontsize= "large")

    return fig, ax


survey(results, category_names)
plt.title("Proportions of death due COVID-19 across genders")
plt.show()
```

References

[1] URL: https://www.pnas.org/doi/10.1073/pnas.2006048117.

[2] Peter Bauer et al. *An international comparison of age and sex dependency of covid-19 deaths in 2020: A descriptive analysis*. Sept. 2021. URL: https://www.nature.com/articles/s41598-021-97711-8.

[3] *Daily record timelines by regions for cases*. URL: https://www.covid19.admin.ch/api/data/documentation/models/sources-definitions-dailyincomingdata.md#sumtotal.

[4] *Daily record timelines for hospitalisations of fully vaccinated persons by vaccine.* URL: https://www.covid19.admin.ch/api/data/documentation/models/sources-definitions-dailyhospvaccpersonsincomingdat md.

[5] *Iso-Week record timelines by geoRegion and age brackets for tests.* URL: https://www.covid19.admin.ch/api/data/documentation/models/sources-definitions-weeklyincomingdata.md.

[6] *Iso-Week record timelines for deaths of fully vaccinated persons by vaccine and sex.* URL: https://www.covid19.admin.ch/api/data/documentation/models/sources-definitions-weeklydeathvaccpersonssexincomi md.

[7] Dharun Kasilingam et al. *Exploring the growth of covid-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning.* May 2021. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7436699/.

[8] aDepartment of Family Medicine. *Does covid-19 vaccination cause excess deaths? : Journal of the Chinese Medical Association.* URL: https://journals.lww.com/jcma/fulltext/2021/09000/does_covid_19_vaccination_cause_excess_deaths_.2.aspx.

[9] *More men than women are dying from covid-19. why?* Aug. 2020. URL: https://www.hsph.harvard.edu/news/features/more-men-than-women-are-dying-from-covid-19-why/.

[10] Xuqi Zhang et al. *The effect of population size for pathogen transmission on prediction of COVID-19 spread.* Sept. 2021. URL: https://www.nature.com/articles/s41598-021-97578-9.