Group Project Stage 2

Group:
DATA1002-CC-L1-G3

Members:
Ethan Yong Vi Hsiong | SID: 510067773 | Unikey: eyon5512

Alexandre Sauze Contreras | SID: 520604830 | Unikey: asau0735

Sebastian Klemens | SID : 510594198 | Unikey: skle4070

Thomas Drageset | SID: 520562286 | Unikey: tdra0337

Sebastian: Suicide mortality rate depending on the gender in each reigon between 2005-2015 (suicide mort, gender, reigon, year)

Introduction

Topic: What variables impact different mortality rates?

We used the “SDG_goal3_clean.csv” dataset that was provided by the DATA1002 teaching staff. The dataset has already been cleaned and contains 28 columns and 163 rows consisting of information on mortality rates and healthcare quality with each row representing data of a specific country on a specified year. Conclusions made on this topic will prove to be crucial to the healthcare industry and to governments as by investigating the variables that impact mortality rates the most, the industry and the government will be able to gain insights and make adjustments to decrease overall mortality rates. An example would be to focus more funding towards the education of healthcare workers if healthcare worker density is proven to decrease mortality rates.

Country	Year	Region	Maternal mortality ratio	Proportion of births attended by skilled health personnel (%)	Infant mortality rate (deaths per 1,000 live births):::BOTHSEX	Infant mortality rate (deaths per 1,000 live births):::MALE	Infant mortality rate (deaths per 1,000 live births):::FEMALE
Albania	2000	Europe	23	99.1	24.1	27.4	20.6
Armenia	2000	Asia	43	96.8	27	29.8	24.1
Armenia	2005	Asia	35	97.8	21.3	23.5	18.8
Armenia	2010	Asia	32	99.5	16.5	18.3	14.6
Australia	2000	Oceania	7	99.3	5.1	5.6	4.6
Australia	2005	Oceania	5	99.4	4.8	5.2	4.3
Australia	2010	Oceania	5	99.1	4	4.4	3.6
Australia	2015	Oceania	6	98.8	3.3	3.5	3
Austria	2000	Europe	6	98.3	4.6	5	4.1

Table 1: Data schema of original dataset (subset)

Part A:

Unikey tdra0337, SID: 520562286

Group-Aggregate Summaries:

The dataset that was used to create the grouped-aggregate summaries was “SDG_goal3_clean.csv”. The clean dataset is a 163x28 csv that is read into a pandas dataframe. The variables contain information about country, region, year, healthcare and mortality ratios. All 28 variables have 163 Non-Null entries, and out of the 28 variables 23 are of type float 64. Country and Region have data type object. While Year, Universal health coverage (UHC) service coverage index and Maternal mortality ratio have data type int64. Data schema:

```
RangeIndex: 163 entries, 0 to 162
Data columns (total 23 columns):
 #   Column                                                                  Non-Null Count  Dtype
---  ------                                                                  --------------  -----
 0   Country                                                                 163 non-null    object
 1   Year                                                                    163 non-null    int64
 2   Region                                                                  163 non-null    object
 3   Maternal mortality ratio                                                163 non-null    int64
 4   Proportion of births attended by skilled health personnel (%)           163 non-null    float64
 5   Infant mortality rate (deaths per 1,000 live births):::BOTHSEX          163 non-null    float64
 6   Infant mortality rate (deaths per 1,000 live births):::MALE             163 non-null    float64
```

Figure 1: Figure is an example structure for the data schema, for the sake of readability some of the columns were removed from the dataset. For the full data schema you can run the code from the “project 2 data1002 thomas.py python file”.

The first grouped-aggregate summary:

In the first grouped-aggregate summary the data is grouped by the nominal qualitative variable region. The groups are Africa, Americas, Asia, Europe and Oceania. Then the agg() function takes the stat list, which contains the statistical attributes of interest, median, mean, max and min, and calculates them for the grouped dataframe and outputs a grouped-aggregate summary. From the grouped-aggregate summary the column containing the quantitative variable "Proportion of births attended by skilled health personnel (%)" was chosen. The grouped-aggregate summary for this data can be viewed in the following table. The code for how the grouped-aggregate summary and the table was made can be found in the “project 2 data1002 thomas.py” python file.

	median	mean	max	min
Region				
Africa	98.75	81.425000	99.8	39.7
Americas	97.10	94.966667	100.0	70.9
Asia	98.60	95.985417	100.0	26.5
Europe	99.70	99.331765	100.0	95.4
Oceania	98.55	98.090000	99.4	96.3
	median	mean	max	min

,

Figure 2: Figure 2 is the table containing the values "Proportion of births attended by skilled health personnel (%)" of the first grouped-aggregate summary.

the second grouped-aggregate summary:

In the second grouped-aggregate summary the data is grouped a binned version of the quantitative variable "Universal health coverage (UHC) service coverage index". This index goes from 0-100 and for the purpose of grouping this data it was binned into 5 bins, (0,20], (20,40], (40,60], (60,80] and (80,100], using the cut() function. This function takes the chosen quantitative column and places each element into one of the 5 bins. Then the agg() function takes the stat list, which contains the statistical attributes of interest, median, mean, max and min, and calculates them for the grouped dataframe and outputs a grouped-aggregate summary. From the grouped-aggregate summary the column containing the quantitative variable "Infant mortality rate (deaths per 1,000 live births):::BOTHSEX " was chosen. The grouped-aggregate summary for this data can be viewed in the following table. the “project 2 data1002 thomas.py” python file.

UHC_bins	median	mean	max	min
(0, 20]	NaN	NaN	NaN	NaN
(20, 40]	50.6	56.816667	107.2	8.7
(40, 60]	21.0	22.960000	67.6	4.2
(60, 80]	5.1	7.142105	30.3	2.0
(80, 100]	3.3	3.752941	8.8	1.8

Figur 3:Figure 3 is the table containing the values "Infant mortality rate (deaths per 1,000 live births):::BOTHSEX" of the second grouped-aggregate summary. The Nan values in the first row show that no countries had a UHC score of less or equal to 20.

The Charts:

Chart 1:

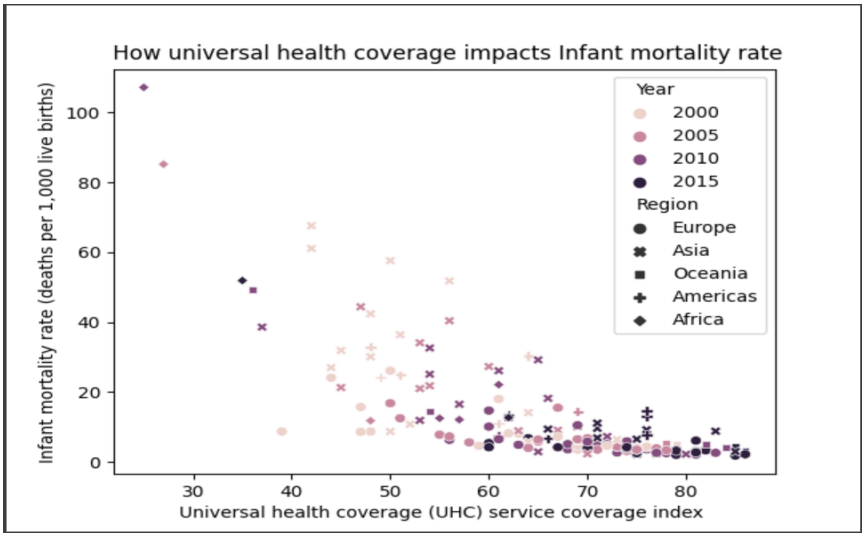

Chart 1: chart 1 shows "Universal health coverage (UHC) service coverage index " on the x-axis and "Infant mortality rate (deaths per 1,000 live births):::BOTHSEX " on the y-axis. The colour of the points show what year the data was collected and the shape of the point shows what region(Europe, Asia, Africa, Americas) the country is a part of.

The first plot shows "Universal health coverage (UHC) service coverage index " on the x-axis and "Infant mortality rate (deaths per 1,000 live births):::BOTHSEX " on the y-axis. The colour of the points show what year the data was collected and the shape of the point shows what region(Europe, Asia, Africa, Americas) the country is a part of.
The variable "Universal health coverage (UHC) service coverage index " is described by the WHO as an index to measure how well a country “Achieve universal health coverage, including financial risk protection, access to quality essential health-care services and access to safe, effective, quality and affordable essential medicines and vaccines for all”.(source url:https://www.who.int/data/gho/indicator-metadata-registry/imr-details/4834)
The scatter plot shows a strong negative correlation(corr = -0.74) between a country's score on the Universal health coverage (UHC) service coverage index and the reduction of infant mortality rate. Most of the countries show an improvement in the UHC service coverage index from year 2000 to year 2015, since the grouping goes from primarily red to primarily white as value of the x-axis increases. This indicates an improvement of the UHC service coverage index with the increase of time. Overall the plot shows that increasing the UHC score will in most cases lead to a reduction in Infant mortality. The code that produced chart 1, and the correlation between the variables can be found in the “project 2 data1002 thomas.py” python file.

Chart 2:

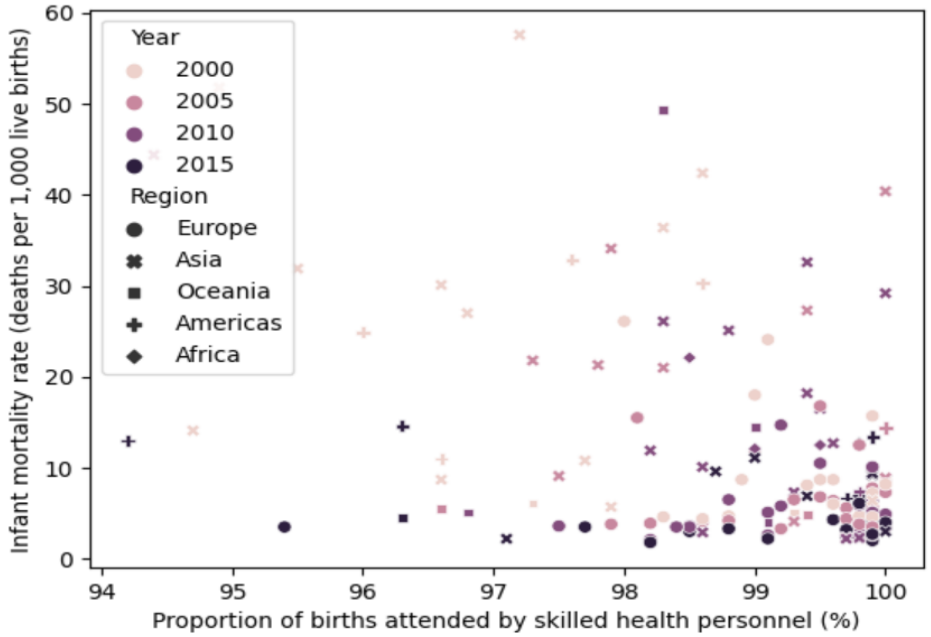

Chart 2: Chart 2 shows "Proportion of births attended by skilled health personnel (%)" on the x-axis and "Infant mortality rate (deaths per 1,000 live births):::BOTHSEX " on the y-axis. The plot also starts from a Proportion of births attended by skilled health personnel at 93% to have a readable graph while removing the least amount of data points. The colour of the points show what year the data was collected and the shape of the point shows what region the country is a part of.

The second plot shows "Proportion of births attended by skilled health personnel (%)" on the x-axis and "Infant mortality rate (deaths per 1,000 live births):::BOTHSEX " on the y-axis. The plot also starts from a Proportion of births attended by skilled health personnel at 93% to

have a readable graph while removing the least amount of data points. The colour of the points show what year the data was collected and the shape of the point shows what region the country is a part of. This graph gives a high negative correlation(corr = -0.608). The plot also indicates a negative correlation between year and infant mortality, as the overall infant mortality reduces from 2000 to 2015 which you can see from the increasing degree of whiteness on the bottom right of the plot and an increasing degree of redness as we move up the y-axis. The code that produced chart 2, and the correlation between the variables can be found in the “project 2 data1002 thomas.py” python file.

Chart 3:

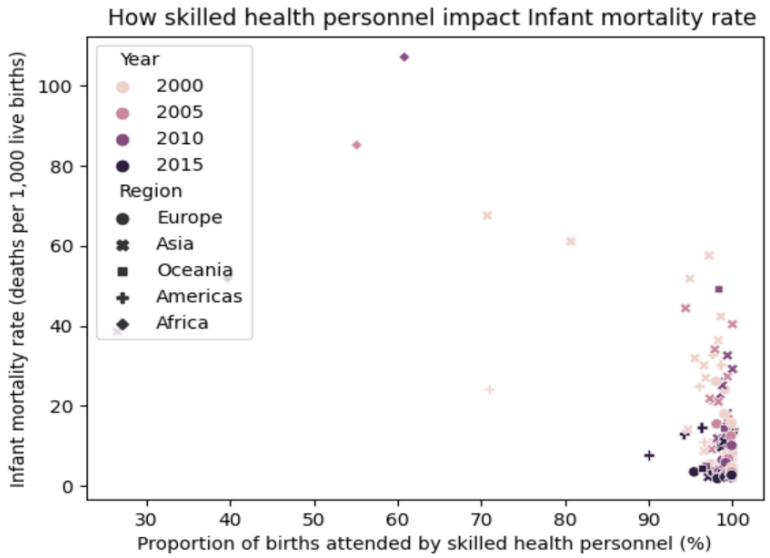

Chart 3: Chart 2 shows "Proportion of births attended by skilled health personnel (%)" on the x-axis and "Infant mortality rate (deaths per 1,000 live births):::BOTHSEX " on the y-axis. The colour of the points show what year the data was collected and the shape of the point shows what region the country is a part of.

The third plot is the same data as in the second plot, but it starts the x-axis at 0%. This plot shows the extent of the negative correlation that could be lost if you view the second plot without thinking about where the x-axis begins. The reason for separating these plots is that the readability of the data points in the second plot is better, but the extent of the negative correlation between the x- and y-axis in the third plot is more evident. There could be inaccuracies in the analysis because of the clustering of the data on such a small area. The clustering could give a higher negative correlation because of confounding variables that are consistent within two or more of the variables included in this plot. To discover and resolve potential confounding variables further analysis of the dataset and the sources of the dataset is required. The code that produced chart 3, and the correlation between the variables can be found in the “project 2 data1002 thomas.py” python file.

Chart 4:

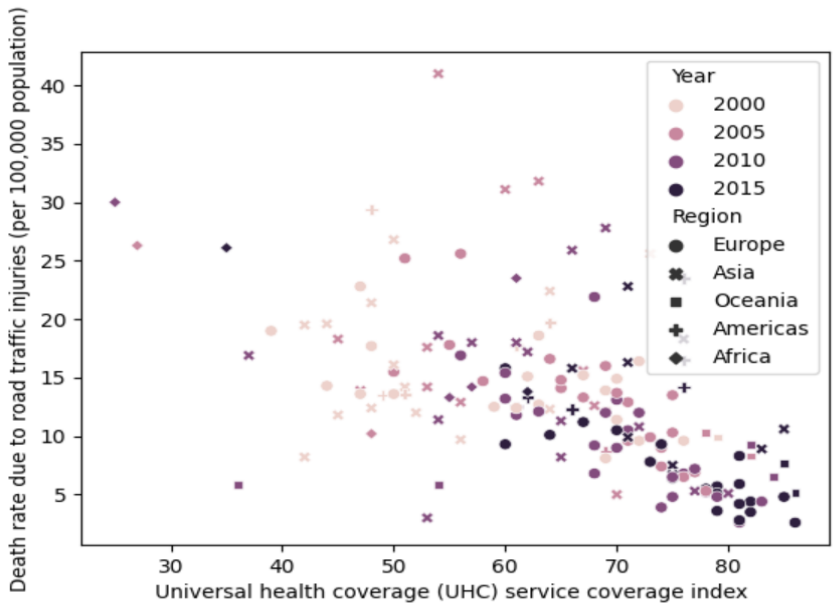

Chart 4: Chat 4 shows "Universal health coverage (UHC) service coverage index " on the x-axis and "Death rate due to road traffic injuries, by sex (per 100,000 population):::BOTHSEX " on the y-axis. The colour of the points show what year the data was collected and the shape of the point shows what region the country is a part of.

The fourth plot shows "Universal health coverage (UHC) service coverage index " on the x-axis and "Death rate due to road traffic injuries, by sex (per 100,000 population):::BOTHSEX " on the y-axis. The colour of the points show what year the data was collected and the shape of the point shows what region the country is a part of. This plot shows a weak negative correlation(corr = -0.55) between the x- and y-axis, this is visually indicated by the spread of the points. By visual inspection of chart 4 it seems like the correlation improves for the data points collected in 2010 and 2015, And the correlation of European countries seems to be stronger. But to draw any conclusions further data analysis is needed. The code that produced chart 4, and the correlation between the variables can be found in the “project 2 data1002 thomas.py” python file.

Grouped Aggregate Summaries

Unikey: eyon5512
SID: 510067773

A subset of the original dataset was used to generate group aggregates and charts, the data subset consisted of 8 columns and 163 rows containing information on death rates due to road traffic injuries and health worker density with each row representing the data of a country in a specific year. The data subset was generated in python using the pandas module and was exported into a csv file.

Data Subset Schema:

Country	Year	Death rate due to road traffic injuries, by sex (per 100,000 population):::MALE	Death rate due to road traffic injuries, by sex (per 100,000 population):::FEMALE	Universal health coverage (UHC) service coverage index	Health worker density, by type of occupation (per 10,000 population)::PHYSICIAN	Health worker density, by type of occupation (per 10,000 population)::NURSEMIDWIFE	Health worker density, by type of occupation (per 10,000 population)::PHARMACIST
Albania	2000	22.4	6.1	44	13.821	40.17	3.432
Armenia	2000	32.2	8.6	44	27.007	59.089	0.345
Armenia	2005	28.6	9.2	45	25.643	49.884	0.319
Armenia	2010	27.7	9.5	57	28.419	52.396	0.427
Australia	2000	14.0	5.7	79	24.944	100.871	8.056
Australia	2005	12.2	4.4	82	27.794	97.628	7.384
Australia	2010	9.6	3.5	84	33.429	104.017	8.683
Australia	2015	7.5	2.9	86	34.886	122.016	8.474
Austria	2000	19.3	6.1	59	38.492	56.942	5.616

Table 2: Data schema of a subset of the original dataset

Average health worker density and average death rates due to traffic injuries across the years for each country

Group aggregation was performed in python. The data file was first imported into python and the values were defined into variables (e.g. nurse_density, death_rate, country):

Secondly, dictionaries were used to group the quantitative variables based on the country variable (nominal attribute):

```
        # Grouping nurse density in different years by country
        if country not in dic:
            dic[country] = [nurse_density]
        else:
            dic[country].append(nurse_density)
```

This set of code was performed for each quantitative variable resulting in 4 dictionaries, further details can be found in Table1.py.

Thirdly, the values for each key in each dictionary was averaged with the following code:

```
for key in sorted(dic):
    if key not in dic2:
        dic2[key] = round((sum(dic[key])/len(dic[key])),3)
```

This set of code was performed for each of the four dictionaries, further details can be found in Table2.py.

Fourthly, the four dictionaries were merged based on their keys as they all had the same keys (Country) with the following code utilising the collections module:

```
# Merging the dictionaries
dd = defaultdict(list)
for d in (dic2, dic3, dic4, dic5):
    for key, value in d.items():
        dd[key].append(value)
table_data = pd.DataFrame(dd).T
```

The final merged dictionary was converted into a dataframe using the pandas module.

Lastly, the data frame was converted to a table format using the tabulate module. The output was written into a csv (‘output.csv'), an excel file (table1.xlsx) of the output was also generated for chart generation:

```
# Export table output into csv file
head = ["Country", "Average Nurse Density per 100,000 Population",
        "Average Death Rate due to Road Injuries per 100,000 Population",
        "Average Pharmacist Density per 100,000 Population",
        "Average Physician Density per 100,000 Population"]
table = tabulate(table_data, headers=head, tablefmt='tsv')
table_data.to_excel('table1.xlsx', header=["Average Nurse Density per 100,000 Population",
        "Average Death Rate due to Road Injuries per 100,000 Population",
        "Average Pharmacist Density per 100,000 Population",
        "Average Physician Density per 100,000 Population"])
output_file = open("output.csv", "w")
output_file.write(table)
output_file.close()
```

Output (subset of output, further details can be found in ‘output.csv’):

Country	Average Nurse Density per 100,000 Population	Average Death Rate due to Road Injuries per 100,000 Population	Average Pharmacist Density per 100,000 Population	Average Physician Density per 100,000 Population
Albania	40.17	14.3	3.432	13.821
Armenia	53.79	18.633	0.364	27.023
Australia	106.133	7.475	8.149	30.263
Austria	64.02	8.65	6.358	45.224
Azerbaijan	80.849	11.167	2.329	36.064
Bahrain	26.553	11.475	1.738	9.847
Bangladesh	1.83	16.9	0.825	3.635
Barbados	47.448	8.7	9.084	17.698
Belarus	106.145	21.9	2.899	32.498
Bosnia and Herzegovina	52.51	17.375	1.046	17.37

Table 3: Average health worker density and death rates due to road accidents of each country (subset of full data)

Average death rates due to traffic injuries grouped by max UHC bins across the years

The max UHC bins of each country were first aggregated in python together with the average death rates of each country using the pandas module and exported into a csv file, 'output2.csv', further details can be found in 'Table2.py'. This csv file was then used to obtain the average death rates due to traffic injuries for each UHC bin through the following code:

```
# Sort death rates into UHC bins
dic = {}
first_line = True
for lines in open('output2.csv'):
    if first_line:
        first_line = False
    else:
        values = lines.rstrip("\n").split("\t")
        country = values[1]
        uhc = int(values[2])
        av_death_rate = float(values[3])
        bins = uhc//20
        if bins not in dic:
            dic[bins] = [av_death_rate]
        else:
            dic[bins].append(av_death_rate)

# Obtain average death rate in each UHC bin
dic2 = {}
dic2["UHC Bins"] = "Average Death Rate due to Road Traffic Injuries per
100,000 Population"
for key in sorted(dic):
    if key not in dic2:
        dic2[key] = round((sum(dic[key])/len(dic[key])),3)

# Export table output to csv file
table_data = pd.DataFrame.from_dict(dic2, orient="index")
head = ('UHC Bins', '20-39', '40-59', '60-79', '80-100')
table3 = tabulate(table_data.T, headers = head, tablefmt= 'tsv')
print(table3)
output_file = open("output3.csv", "w")
output_file.write(table3)
output_file.close()
```

This code generated the following output ('output3.csv'):

UHC Bins	20-39	40-59	60-79	80-100
Average Death Rate due to Road Traffic Injuries per 100,000 Population	21.02	16.942	14.476	7.36

Table 4: Average death rates due to road traffic injuries in each UHC bin

Charts

Unikey: eyon5512
SID: 510067773

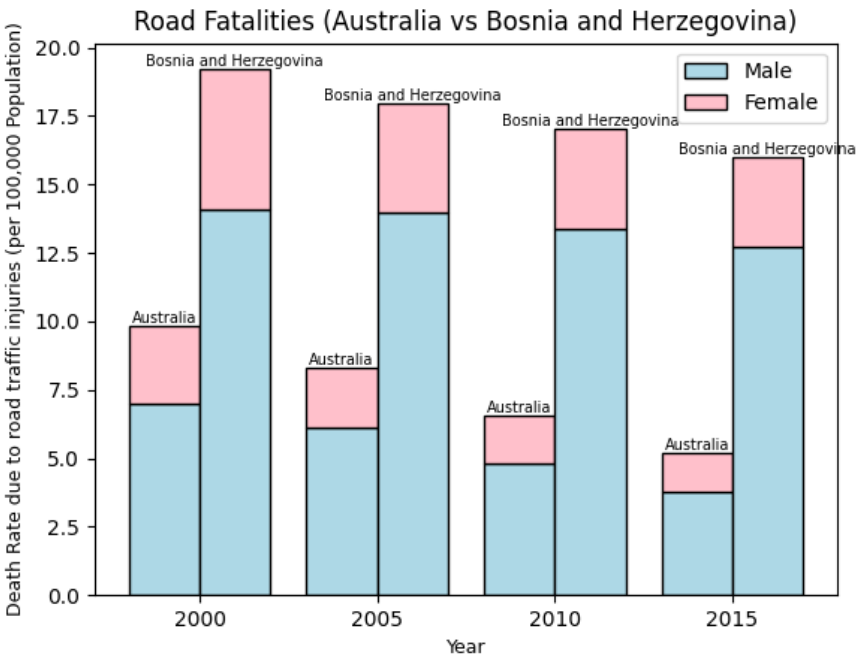

Chart 5: Bar graph of road fatalities in Australia vs Bosnia and Herzegovina split by gender from 2000-2015

There are 4 data attributes encoded in Figure 1, one quantitative attribute (Death rate due to road traffic injuries) and 3 nominal attributes (gender, country and year). Axis labels, bar labels, scales and legends are clearly labelled. The bar labels represent which country is being represented by the bar, the height of the bar represents the death rates due to road traffic injuries, the y-axis represents which year the data is from and the colour of the bar represents if the death is male or female. The countries Australia and Bosnia and Herzegovina were selected as Australia had the highest UHC among all the countries and Bosnia and Herzegovina had the lowest UHC rating among the countries which had data for all 4 years (2000-2015). Only 2 countries were selected as adding more countries to the graph would have adverse effects on the readability of the bar graph as there would be way too many bars to interpret. Therefore, if more data was obtained, it would not benefit the quality of this chart design. However, the gender attribute is only represented by colours which would prove to be a limitation as colour blind individuals will have trouble trying to view the graph. The bar graph (Figure 1) was generated in python with the matplotlib.pyplot module using the following code:

```
# Setting plot axis
years = australia["Year"].to_list()
bar_width = 0.4
xticks_index = range(0, len(years))
aus_index = [x - bar_width/2 for x in xticks_index]
bos_index = [x + bar_width/2 for x in xticks_index]

# Plotting
```

```
plt.xticks(xticks_index, years)
plt.xlabel("Year", fontsize = 9)
plt.ylabel("Death Rate due to road traffic injuries (per 100,000 Population)",
fontsize = 9)
plt.title('Road Fatalities (Australia vs Bosnia and Herzegovina)')
aus_bar_m = plt.bar(aus_index, australia_death_rate_m, bar_width, label =
"Male", color = "lightblue", edgecolor = 'black')
aus_bar_f = plt.bar(aus_index, australia_death_rate_f, bar_width, bottom =
australia_death_rate_m, label = "Female", color = "Pink", edgecolor = 'black')
bos_bar_m = plt.bar(bos_index, bosnia_death_rate_m, bar_width, color =
'lightblue', edgecolor = 'black')
bos_bar_f = plt.bar(bos_index, bosnia_death_rate_f, bar_width, bottom =
bosnia_death_rate_m, color = 'Pink', edgecolor = 'black')
plt.bar_label(aus_bar_f, labels= ["Australia", "Australia", "Australia",
"Australia"], fontsize = 7)
plt.bar_label(bos_bar_f, labels = ['Bosnia and Herzegovina', 'Bosnia and
Herzegovina', 'Bosnia and Herzegovina', 'Bosnia and Herzegovina'], fontsize = 7)
plt.legend()
plt.show()
```

Further details on the variables can be found in 'Charts.py'.

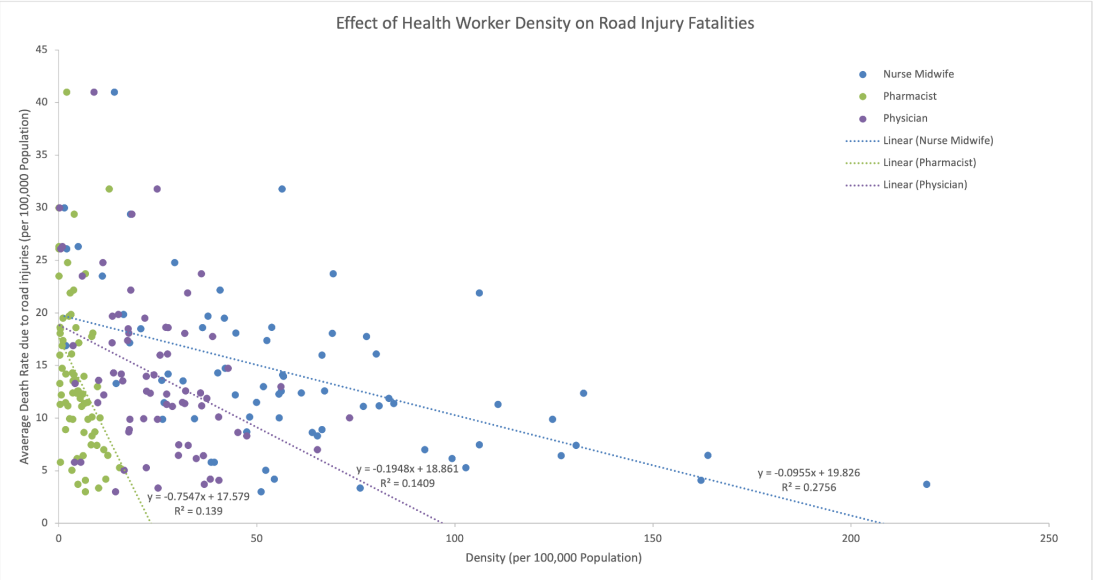

Chart 6: Scatterplot of health worker density plotted against average death rate due to road injuries. Health worker density was grouped by occupation.

The scatter plot (Figure 2) was generated in excel in the 'table1.xlsx' excel spreadsheet which reflects the group aggregated data (Table 3) in excel format. Plot generation was performed by inserting a scatterplot in excel and creating a series for each of the occupation types. For each series, the y values were set as the average death rates and the x values were set as the health worker densities of that occupation. The axis labels, axis ticks, legend, trendline, trendline equation and R^2 values were then added to the plot.

There are 3 attributes encoded in the plot (Figure 2), 2 quantitative attributes (health worker density and average death rate due to road injuries) and one nominal attribute (health worker occupation). The x-position of the point represents the average death rate due to road injuries, the y-position of the point represents the health worker density and the colour

of the point represents the type of occupation. Each point represents data from a different country. A scatterplot was used as it is the most effective plot to use when looking for relationships between variables. Adding trendlines to the plot allowed for clear displays of the relationship between the x and y variables. Separating the health worker densities by occupation allows further insight on the effect of each occupation density on road fatalities. Obtaining more data would have a positive effect on the quality of the scatterplot and the change would also be reflected on the R^2 values of each trendline which represents the fit quality of each trendline.

Grouped Aggregate Summaries and Charts

Unikey: asau0735
SID: 520604830

In the original dataset we found five different columns containing information about deaths by cardiovascular diseases, cancer, diabetes and chronic respiratory diseases. Columns 13, 14 and 15 represented the mortality rate or probability of dying by said diseases divided by both sexes, males and females respectively. Column 16 contains the information about the fraction of diabetes deaths between said diseases while column 17 compares the fraction deaths caused by cancer between the same diseases. The 13th column containing information for the set of diseases for both sexes was selected for this analysis along with the diabetes, cancer, country, year and region columns were saved in a separate file using excel to facilitate the process. The following data summary was generated using python.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 163 entries, 0 to 162
Data columns (total 6 columns):
 #   Column                                                                                                                                  Non-Null Count  Dtype
---  ------                                                                                                                                  --------------  -----
 0   Country                                                                                                                                 163 non-null    object
 1   Year                                                                                                                                    163 non-null    int64
 2   Region                                                                                                                                  163 non-null    object
 3   Mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease (probability):::BOTHSEX           163 non-null    float64
 4   Fraction of deaths due to diabetes, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease        163 non-null    float64
 5   Fraction of deaths due to cancer, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease          163 non-null    float64
dtypes: float64(3), int64(1), object(2)
```

Figure 1. Summary of Cleaned Dataset

As we can see the new dataset contains just five columns with 163 values where the country and region's values are strings; the year's values are integers; and the mortality rate of the diseases, fraction of death caused by diabetes and cancer columns' values are floats.

To obtain the grouped aggregate summaries a procedure was made in python to obtain the median, mean, standard deviation, maximum value and minimum value of the mortality rate attributed to cardiovascular disease, cancer, diabetes, and chronic respiratory diseases for each region as seen in figure 2. In figure 3 and 4 a representation of the same summary but for the information of the fraction of deaths by diabetes and cancer in said disease group for each region respectively. A detailed explanation of the coding process can be seen in the "asau0735_DATA1002_Part2.py file.

Region	median mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease	mean mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease	sd mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease	max mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease	min mortality rate attributed to cardiovascular disease, cancer, diabetes or chronic respiratory disease
Europe	18.40	18.637647	6.671202	35.3	9.2
Asia	25.25	24.404167	8.832727	41.6	8.4
Oceania	12.35	17.470000	12.717315	53.0	9.3
Americas	12.90	15.091667	5.120296	28.7	9.8
Africa	23.60	22.662500	3.426346	26.4	14.8

Figure 2.

median fraction of deaths due to diabetes, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	mean fraction of deaths due to diabetes, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	sd fraction of deaths due to diabetes, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	max fraction of deaths due to diabetes, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	min fraction of deaths due to diabetes, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease
0.0230	0.026624	0.016690	0.081	0.005
0.0485	0.069062	0.058661	0.270	0.015
0.0385	0.072800	0.070428	0.216	0.031
0.1015	0.112417	0.046852	0.198	0.056
0.0940	0.149625	0.099668	0.327	0.082

Figure 3.

median fraction of deaths due to cancer, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	mean fraction of deaths due to cancer, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	sd fraction of deaths due to cancer, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	max fraction of deaths due to cancer, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease	min fraction of deaths due to cancer, among deaths due to cardiovascular disease, cancer, diabetes or chronic respiratory disease
0.3220	0.315765	0.072969	0.469	0.148
0.2165	0.227396	0.096633	0.474	0.102
0.3845	0.361700	0.075403	0.426	0.167
0.3325	0.328500	0.043026	0.396	0.253
0.1910	0.229625	0.069835	0.333	0.153

Figure 4.

To produce the charts the previous tables were exported as csv files. After opening the original dataset and the grouped aggregate summaries in the same file the creation of bar charts and a boxplot was simple enough. For Charts 7, 8 and 9 the summary was used to show the mean per region. Chart relates the average probability result per region.

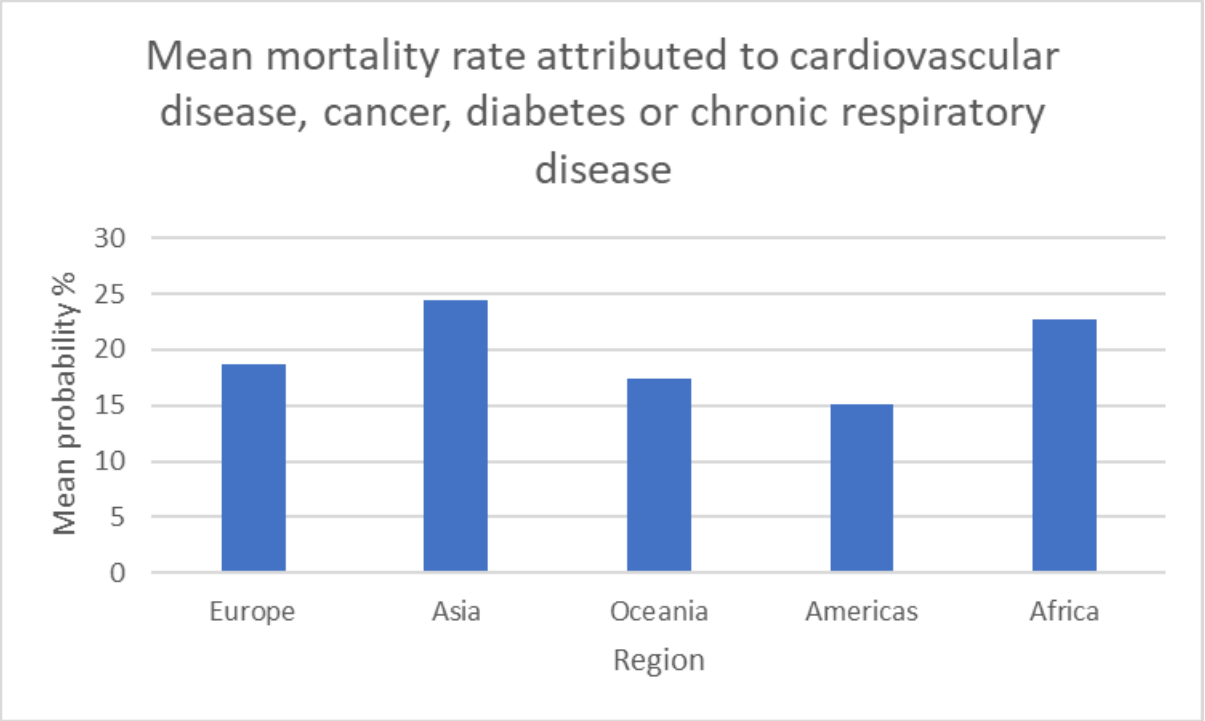

Chart 7. Average mortality rate of cancer, diabetes, cardiovascular and chronic respiratory diseases by region.

Charts 8 and 9 were made by selecting the column containing the region and the mean of fraction of contribution of diabetes and cancer respectively. The program automatically creates a new chart.

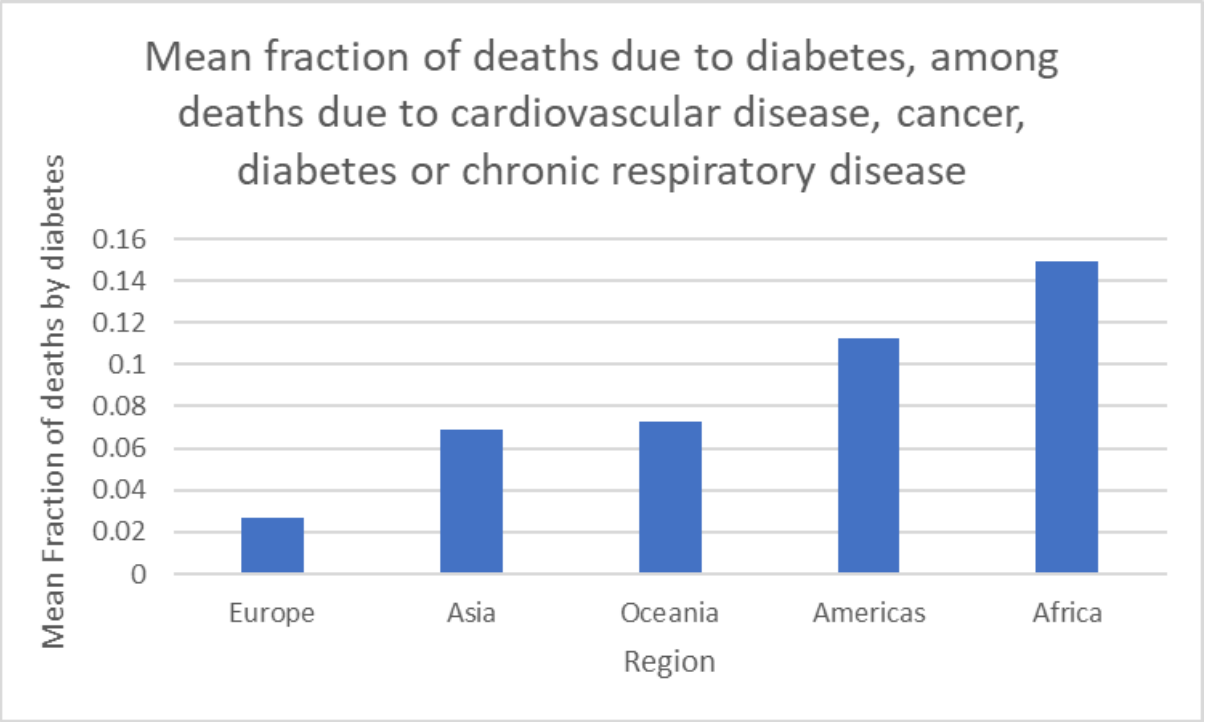

Chart 8. Average fraction of deaths due to diabetes per region

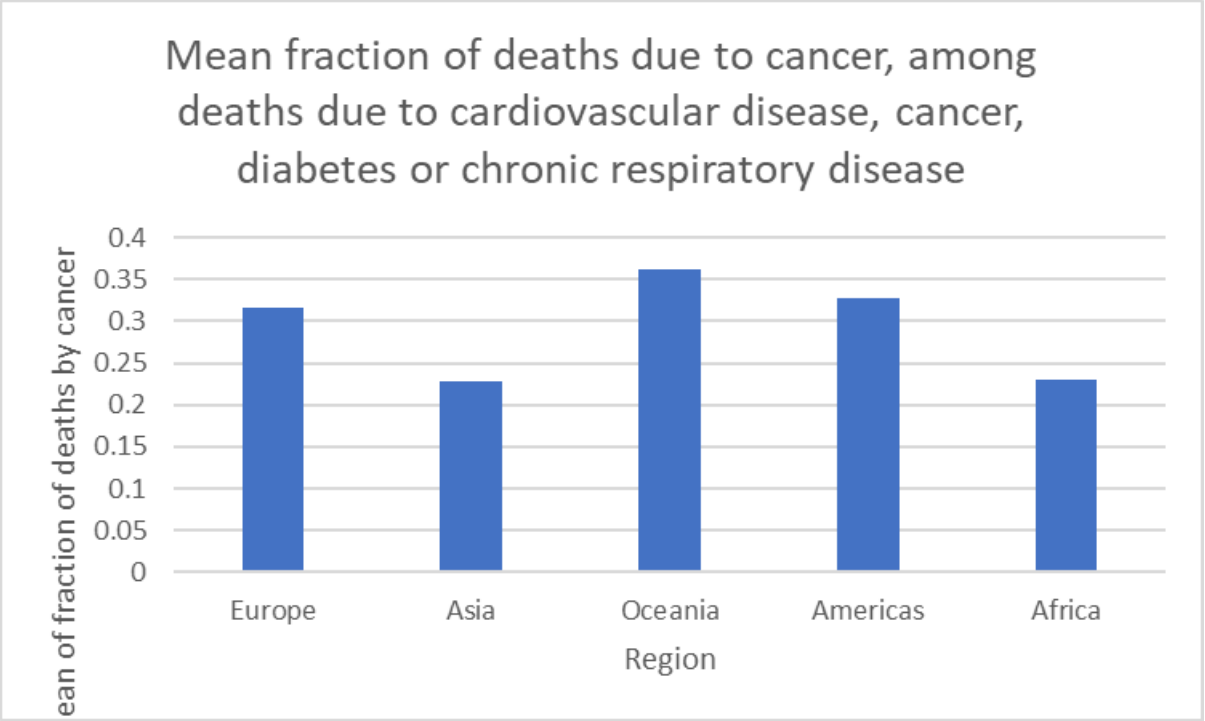

Chart 9. Average fraction of deaths due to cancer per region

The boxplot was created in a different fashion. This time the first version of the dataset was used to graph the whole values of the dataset and create five boxplots, each showing the distribution of the data.

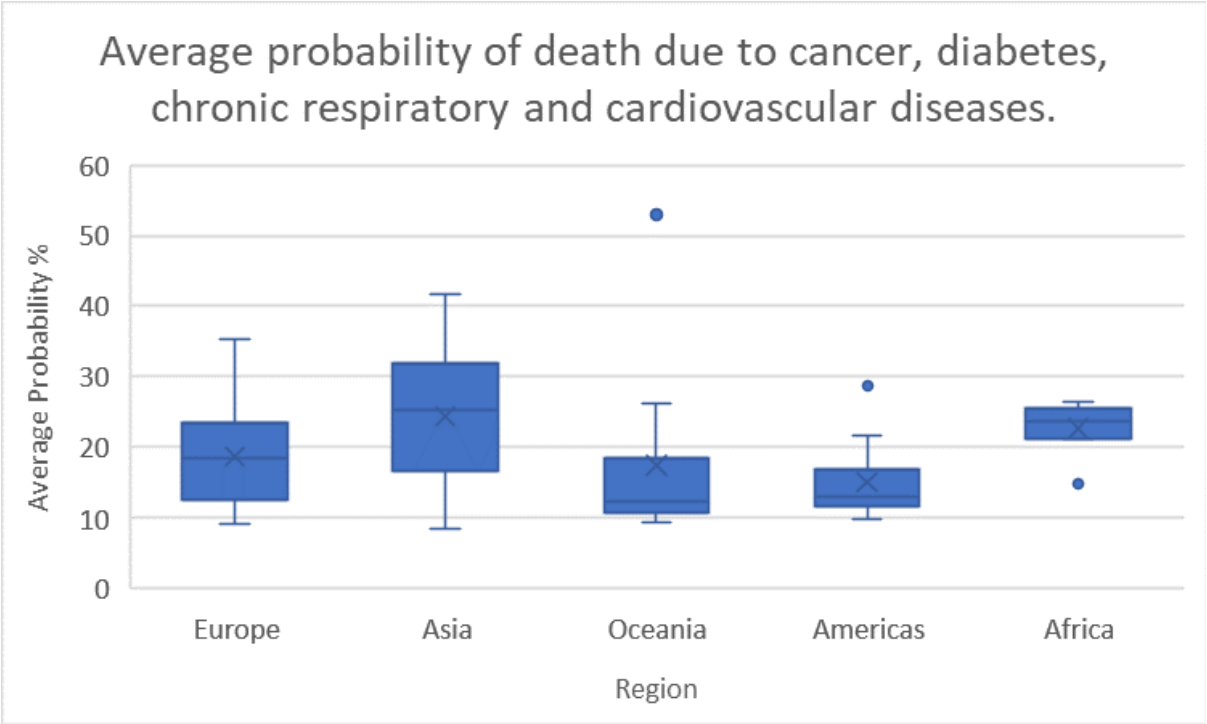
Average probability of death due to cancer, diabetes, chronic respiratory and cardiovascular diseases.
Average Probability %
60
50
40
30
20
10
0
Europe
Asia
Oceania
Americas
Africa
Region

Chart 10.

Grouped Aggregate Summaries and Charts

Unikey: skle4070
SID: 510594198

Part B: Conclusion

Intended audience:

The intended audience of this report is governments, healthcare professionals and policymakers inside the healthcare profession. The result of the data analysis regarding the different kinds of mortality and what variables can impact them on a national or regional scale, can be invaluable regarding the decisions around healthcare policies around the world and help countries decide on what direction they want to develop their national healthcare programs.

eyon5512:

Issue:

Road traffic injuries are inevitable all around the world and when these accidents happen, health workers are theoretically the individual's best chance of survival as they would be the ones responsible for the recovery of the individual. Which begs the question, when it comes to road injuries, does having a higher health worker density actually increase the chances of survival?

Insights:

We generated a table which shows the average death rates due to road injuries, and the average health worker density for each occupation (Nurse, Pharmacist and Physician) across 71 countries (Table 3). However, this information would be difficult to comprehend as there is way too much data in one table and so this information was visualised into a scatter plot (Chart 6).
The scatterplot (Chart 6) explains the relationship between the average death rates due to road injuries and health worker density (split by occupation). The trend lines indicate that there is a negative relationship between death rates due to road injuries and health worker density in general. This also relates to Table 4 where it shows increasing death rates the lower you go in the UHC bins due to health worker density being one of the factors in calculating the UHC rating of a country.
Furthermore, the trend lines also provide further insights of how each occupation affects the death rate. The chart indicates that nurse density has the lowest impact on the death rates due to road traffic injuries as it has the lowest slope (-0.0955) in the trendline equation and

that surprisingly, pharmacist density has the largest impact on death rates due to road traffic injuries with the largest slope (-0.7547).
The scatter plot (Chart 6) therefore suggests that by training more pharmacists, a country can effectively reduce their death rates due to road traffic injuries. However, training any health worker occupation would still mitigate death rates due to road traffic injuries.
The fit of the trend lines are however poor with very low R^2 values, with the highest R^2 value being 0.2756.

Tdra 0337:

Issue:

The issues that are focused are, how does the variables "Universal health coverage (UHC) service coverage index " and "Proportion of births attended by skilled health personnel (%)" impakt the “Infant mortality rate (deaths per 1,000 live births)”, also how the variabel "Universal health coverage (UHC) service coverage index " impact the “Death rate due to road traffic injuries (per 100,000 population)” were investigated.

Insights:

Cart 1 and the data used to create chart 1 indicates that there is a strong negative correlation (corr = -0.74) between “UHC service coverage index” and infant mortality. Showing that increasing the “UHC service coverage index” will in most cases lead to a reduction of infant mortality. Also cart 2 and the data used to create chart 2 indicates a high negative correlation (corr = -0.6) between the "Proportion of births attended by skilled health personnel (%)" and infant mortality. This again shows that increasing the "Proportion of births attended by skilled health personnel (%)” will in most cases lead to a reduction of infant mortality. Further reach of the UHC service coverage index shows that Proportion of births attended by skilled health personnel (%) could be a factor in determining the UHC service coverage index of a country. This indicates that there are a lot of factors in the reduction of infant mortality, but that perhaps Proportion of births attended by skilled health personnel (%) is one of the more important.

Chart 4 and the data used to create chart 4 indicates a weak negative correlation (corr = -0.55) between the “UHC service coverage index” and Death rate due to road traffic injuries (per 100,000 population). The spread of the data in the scatterplot (chart 4) of these to variables makes it hard to draw any further conclusions. A possible way of investigating the possibility of a correlation between these variables might be to subdivide the data by year and/or by region and do a new correlation analysis. But this again can lead to a situation where you find random correlations because you reduce the sampleise to a population that can be too small to do a meaningful statistical analysis.

asau0735:

Issue:

The reduced dataset used focused on deaths by some diseases and conditions such as cancer, diabetes, cardiovascular and chronic respiratory diseases. This information shows the deadliness of each disease by giving out information about the death ratio of said disease group by country and divided into four possible years in each country (2000, 2005, 2010, adn 2015). By obtaining a mean of said deaths per region we can seemingly understand how deadly these diseases are in each region, hence hinting about how advanced a continent is in its healthcare. Also by analysing the contribution of diabetes and cancer in the deaths per the comorbidities we can further understand which diseases have a higher mortality rate in each region.

Insight:

Chart 7 shows the average mortality rate of cancer, diabetes, cardiovascular and chronic respiratory diseases by region. The goal of this chart is to understand how likely it is to die due to contracting the listed diseases. As we can see, continents like Oceania and the Americas are less likely to die when suffering from one of these conditions. Asia and Africa have the highest probability of not surviving one of these diseases.
A similar analysis would be to understand the effect of diabetes inside these diseases. Chart 8 shows how there is a difference in each region as to the contribution of diabetes in deaths between said conditions. Europe and Asia have a small contribution which may indicate a better diet. In the Americas and Africa diets may be less nutritious, diabetes may increase. This may also be linked to genetic diseases. The contribution of deaths by cancer is quite remarkable in each region with almost no variation seen between continents.
To further analyse the probability of death with the comorbidities we can also graph a boxplot to show attributes from the dataset obtained in the grouped aggregate summary. The results show some outliers in Oceania, Americas and Africa which could represent some countries with very different results than their counterparts in the same continent.