Analyzing Metabolic and Behavioral Risk Factors for Noncommunicable Diseases Within Canada

Wardah Ali, Navyasri Chinthapatla, Danae McCulloch, Safeen Mridha, Ayda Takehei

DATA 604: Working with Data at Scale

FALL 2024

Table of Contents

List of Figures

Introduction

Noncommunicable diseases (NCDs) pose a significant challenge to public health systems worldwide as they play a major role in global mortality every year. It was stated that 86% of all deaths in Canada were accounted by a NCD in the year of 2021 (World Health Organization, n.d.)**.** NCDs are defined as illnesses that cannot be transmitted from one person to another, which include conditions such as Cardiovascular diseases, Diabetes, Respiratory diseases and Cancer (World Health Organization, n.d.)**.** Since NCDs are not transferable, factors such as lifestyle behaviors play an integral role in increasing metabolic risk factors that heighten the susceptibility of these diseases. Given the ubiquitous nature of the stress that NCDs place on public health systems worldwide, understanding the impact of lifestyle behaviors is critical for the development of effective prevention and intervention strategies.

The scope of this project will be looking at "Analyzing Metabolic and Behavioral Risk Factors for Noncommunicable Diseases Within Canada". The analysis will be divided into two categories: Metabolic and Behavioral. Metabolic factors are physiological and are largely influenced by genetics, making them mostly uncontrollable. Metabolic factors will include Cholesterol levels, Hypertension, and Obesity. Behavioral factors are controllable and relate to actions such as physical activity, diet, and other day-to-day habits. For this project, the behavioral factors will be specific to Physical Inactivity, Tobacco use, and Alcohol consumption. The analysis will explore the relationship between all factors and the impact they have in Canada. Data regarding these factors are organized in separate datasets which have been derived from the World Health Organization (WHO). Please note that this analysis does not consider other uncontrollable factors that may impact on our results such as environmental and socio-economic factors which limit the findings. Investigating these factors within the population is crucial for assessing overall population health and alerting public health professionals to potential surges in NCDs. This project also compares Canada's results to global data, providing insight into how the country ranks in terms of health outcomes. By analyzing this data, we can identify areas where Canada excels and areas needing improvement.

Individual Dataset

Each team member was responsible for a specific risk factor throughout the project. In other words, the individual was in charge of knowing and preparing the dataset that relates to their assigned factor. The responsibilities and assigned factors were the following: Safeen for Alcohol consumption, Danae for both Cholesterol and Hypertension, Ayda for Obesity, Wardah for Tobacco use, and Navya for Physical Inactivity. Each member was also responsible for incorporating the mortality dataset and exploring any trends or findings with their assigned factor(s).

Alcohol Dataset

The Alcohol dataset used throughout this study was obtained through studies conducted by the WHO. It aimed to look at the total yearly per capita consumption of pure Alcohol by males and females aged 15+ across 189 countries worldwide from the years 2000 to 2019. The official unit of measurement was liters of pure Alcohol per person per year. Alcohol consumption values were calculated by considering the regional production, import, export, and sales data often via taxation of Alcoholic beverages within a country. There were four columns that included measurements of pure Alcohol consumption in this dataset: a low value, a high value, an overall average and a range. The range consisted of the low, high and average values. For the use of this project, we elected to use the averaged values. Furthermore, the dataset used regression models to account for the unrecorded consumption of Alcohol as well, which were incorporated into the averaged values and helped provide a more accurate estimate. In addition to measuring Alcohol consumption among males and females, this dataset also considered Alcohol consumption among both sexes as its own column; it is important to note that the "Both sexes" column is the average between Male total pure Alcohol consumption per year and Female total pure Alcohol consumption per year. Notably, this dataset was clean and did not possess blank value- no cleaning on our end was required. Considering that Alcohol consumption is a major lifestyle risk factor that can be linked to the formation of NCDs such as cancer and Cardiovascular diseases, we elected to use this dataset. From this dataset, I learned that there is a great disparity in Alcohol consumption between men and women. It was common for men to record much higher per capita levels of Alcohol consumption each year, which can shed light into the fact that men are more likely to partake in risky lifestyle behaviors that make them more prone to the onset of NCDs. I also learned that some countries do not have any Alcohol consumption per capita at all (they had values of 0 each year), markedly where Alcohol is forbidden due to religious reasons.

Cholesterol Dataset

Cholesterol values were measured and collected by the WHO for 191 countries worldwide. The dataset was clean and contained no missing values; therefore, no additional data cleaning was required besides the renaming of columns. This dataset was used for analysis as Cholesterol is known to be a metabolic risk factor for NCDs. Cholesterol levels were measured in millimoles per liter (mmol/L). Mean Cholesterol values were organized with years and countries as rows, while the columns included three subcategories under "Mean total Cholesterol, age-standardized" stating the gender populations: Female, Male and Both sexes. Estimates of mean Cholesterol levels from 1980 to 2018 were derived from 1,127 population-based studies involving 102.6 million individuals aged 18 years and older. Results were only selected to 2018 as the dataset did not go beyond this year. The dataset included age-standardized values, along with the low, high, average, and the range for Cholesterol levels. Therefore, the main columns of interest include Indicator, Parent Location, Country, Year, Gender, Cholesterol Value, Cholesterol High Value, Cholesterol Low Value and Range. From visually looking at this dataset, I learned that Cholesterol levels differ

from each country and gender, and that Canada has been recording healthier results since the 2000s which is most likely due to increased awareness and education.

Hypertension Dataset

Hypertension data were measured and collected by the WHO for 195 countries worldwide. Hypertension was defined as having a systolic blood pressure greater than or equal to 140 millimeters of mercury (mmHg), a diastolic blood pressure greater than or equal to 90 mmHg or requiring Hypertension medication. This dataset was used for analysis as Hypertension is a metabolic risk factor for NCDs. The dataset was clean and contained no missing values. Only one step was required to prepare the dataset for analysis: one of the two indicators included in the dataset was removed. Specifically, the indicator representing the "Prevalence of Hypertension among adults aged 30–79 years, crude" was dropped, as the analysis focused solely on age-standardized results. Additionally, renaming the columns was also performed. After performing these steps, the final dataset was organized with years and countries as rows, while the columns included three subcategories under "Prevalence of Hypertension among adults ages 30-79 years, age-standardized" stating the gender populations: Female, Male and Both sexes. This dataset included age-standardized values, along with the low, high, average, and the range in a percentage of population. Data covered the years from 1990 to 2019, incorporating 1,201 population-based studies and data from 104 million individuals aged 30–79 years. Therefore, the main columns of are Indicator, Parent Location, Country, Year, Gender, Hypertension Value, Hypertension High Value, Hypertension Low Value and Range.

From exploring the initial dataset, I noticed a trend where Hypertension tends to affect at least 30% of the male and female population in most countries worldwide, with prevalence reaching as high as 50% in some countries. Fortunately, Canada has been showing a decreasing trend in Hypertension prevalence for both genders.

Obesity Dataset

The prevalence of Obesity among adults aged 18 years and older was recorded by WHO for 198 countries worldwide. Obesity was defined as a body mass index (BMI) of 30kg/m^2 or higher. This dataset was used for analysis as Obesity is a metabolic risk factor for NCDs. Given its strong association with physiological factors and genetics, understanding the prevalence of Obesity can help identify valuable health trends. The dataset included the average age standardized estimates for the percentage of the population with Obesity, along with the low and high estimates, disaggregated by sex and based on measured height and weight. The Obesity estimates from 1990 to 2022 were derived from 3,663 population-based surveys involving 222 million individuals, including children, adolescents, and adults. Results were limited to 2022; however, to maintain consistency with the other datasets included in this study, data up to 2018 will be used. Importantly, the dataset was clean and free of missing values, so no additional data cleaning was necessary, aside from renaming the columns. From this dataset, I learned that the prevalence of Obesity varies

across countries, genders and regions. Within Canada, the trend since 2000 has been positive with males having higher estimates then females.

Physical Inactivity Dataset

The Prevalence of insufficient Physical activity among adults aged 18 years and older was recorded by the WHO for 195 countries worldwide. WHO defined Physical Inactivity as the activity of less than 150 minutes per week of moderate-intensity aerobic activity, or less than 75 minutes per week of vigorous-intensity activity. It is recorded in percentages and represents the proportion of adults who engage in insufficient physical inactivity which can lead to NCDs, including Cancer, Cardiovascular diseases, Respiratory diseases, and Diabetes. The dataset runs from 2000 to 2022 with no missing years in between. It reflects physical inactivity by gender, hence showing the impact of this behavioral risk factor on non-communicable diseases. This behavioral risk factor dataset is used to test against the metabolic risk factor and then examine its effects on each NCD. EDA and analysis are for all these years up to 2022. However, data until 2018 is used to maintain consistency with all other datasets used in this study. The main columns of interest include Indicator, Parent Location, Country, Year, Gender, Physical Inactivity Value, Physical Inactivity High, Physical Inactivity Low, and Range.

Tobacco Dataset

Prevalence of Tobacco use among people aged 15 years and older was recorded by the WHO for 165 countries worldwide. The data was recorded in percentages meaning; percentage of population aged 15 years and over who currently use any Tobacco products on a daily or non-daily basis. Since Tobacco use is a major contributor to illness and death from NCDs and there is no safe level for Tobacco consumption, this dataset will be used under behavioral risk factors to analyze its impact on metabolic ricks factors and its significance in different NCDs. The data has been recorded from 2000 to 2022 with few years missing in between, it also contains predictions for 2025 and 2030 for all the countries for both genders. Initially the dataset contained 34 columns but since our primary area of interest is to look at Tobacco values over the past eighteen year from 2000 to 2018 in Canada for both genders and see the correlation with NCDs we reduced the dataset down to 8 columns i.e. Region, Country, Year, Gender, Tobacco value, Tobacco value high, Tobacco value low and range. From this dataset, I learned that Tobacco use varies across countries, genders and regions. Within Canada, the trend since 2000 has been negative with males consuming more Tobacco than females.

Mortality Dataset

The mortality values of the four NCDs were measured and collected by the WHO for 183 countries worldwide. This dataset was derived from the WHO Global Health Estimates (GHE) and includes the number of deaths by cause, disaggregated by age and sex for Cardiovascular diseases, Diabetes mellitus, Respiratory diseases, and Malignant neoplasms (Cancer). This data has been recorded from 2000 to 2019 with no missing values for any NCDs, only renaming of the column

was performed to make them understandable. Initially, this dataset obtained 34 columns, but we majorly used five columns for the analysis namely; Country, Year, Gender, Mortality Value, and Disease as the rest of the columns were redundant. From this dataset, I learned that the mortality rate varies across countries and regions for each of the NCDs. Within Canada, Respiratory diseases and Malignant neoplasms (cancer) have shown an increase in the number of deaths while Cardiovascular diseases and Diabetes showed no trend.

Exploratory Data Analysis

In this exploratory data analysis, each member explored their assigned factor(s). Please note that the factor values are compared to the mortality data to explore trends or other findings. Our project will be analyzing the years from 2000 to 2018 as these years are present in all seven datasets. For visual purposes, a facet grid was performed on top of the exploration process to purposely show the Mortality trend for each NCD for the respective factor. This was done by using SQL and calling the 'JOIN ON' built-in function to create one table that shows both the results from the factor and the mortality values. Additionally, the query selected the appropriate years (2000 to 2018) and the Country (Canada). The SQL-Alchemy and Panda's libraries were used together to run SQL queries and convert them to data frames for further visualizations of trend lines and scatter plots. Note that this technique was also used to answer our guiding questions regarding the trend in Metabolic and Behavioral risk factors along with their correlation.

Alcohol

I was able to learn about the trends of Alcohol consumption between 2000 to 2018 through the exploration of Alcohol dataset using SQL. A query was used to extract the values for Canada's total pure Alcohol consumption per capita from the years 2000 to 2018 for both men and women. Across both genders, it was revealed that the yearly Alcohol consumption per capita remained mostly constant over the time frame, with men hovering around 15L of pure Alcohol consumed per capita, and women hanging around 4.2L of pure Alcohol consumed per capita (Appendix 1.1.B)**;** there was a slight increase in about 2009 in Alcohol consumption for both men and women (Appendix 1.1.B). From this, I learned that men consumed almost 4 times more pure Alcohol per year compared to women throughout the 19 years, suggesting that men in Canada indulge more in lifestyle risk factors, like drinking (Appendix 1.1.B). Considering these trends, I was curious to see where Canada ranked overall in terms of Alcohol consumption per capita, and if this rank increased or decreased during our time frame; so, two more queries were made to see Canada's world rank- one for the year 2000, and another for the year 2018. It was revealed that Canada's rank in 2000 was 40^{th} in the world, whereas it was 33^{rd} in the world by the year 2018 (Appendix 1.1.A). This shows that Canada's rank slightly increased over time, suggesting that compared to the 189 other countries in this study, Canada's Alcohol consumption is comparatively rising at a global scale.

Finally, considering that Alcohol consumption is considered as a risk factor for the onset of NCDs, I created queries to investigate the trends between Alcohol consumption and NCD-caused mortality. To do so, I joined the Alcohol dataset with the Mortality dataset, which quantified the number of NCD-caused mortalities that accumulated yearly across the world. Within this dataset, four NCDs were considered: Diabetes Mellitus, Cardiovascular Diseases, Malignant Neoplasms (Cancer) and Respiratory Disease. I merged the two datasets using the key "Year" and "Gender" and created four tables, one with each NCD being related to Alcohol consumption, to see if the consumption of Alcohol is correlated with NCD-caused mortalities, and if there were any sex-related trends (Appendix 1.1.C, Appendix 1.1.D, Appendix 1.1.E, Appendix 1.1.F). It was revealed that Alcohol consumption does not seem to be correlated with NCD mortalities, as the number of mortalities across all 4 NCDs remained constant, regardless of the level of Alcohol consumed (Appendix 1.1.C, Appendix 1.1.D, Appendix 1.1.E, Appendix 1.1.F). Upon the creation of these tables, I utilized them to create a facet grid that visualizes the trends between Alcohol use and NCD mortality prevalence (Figure 1).

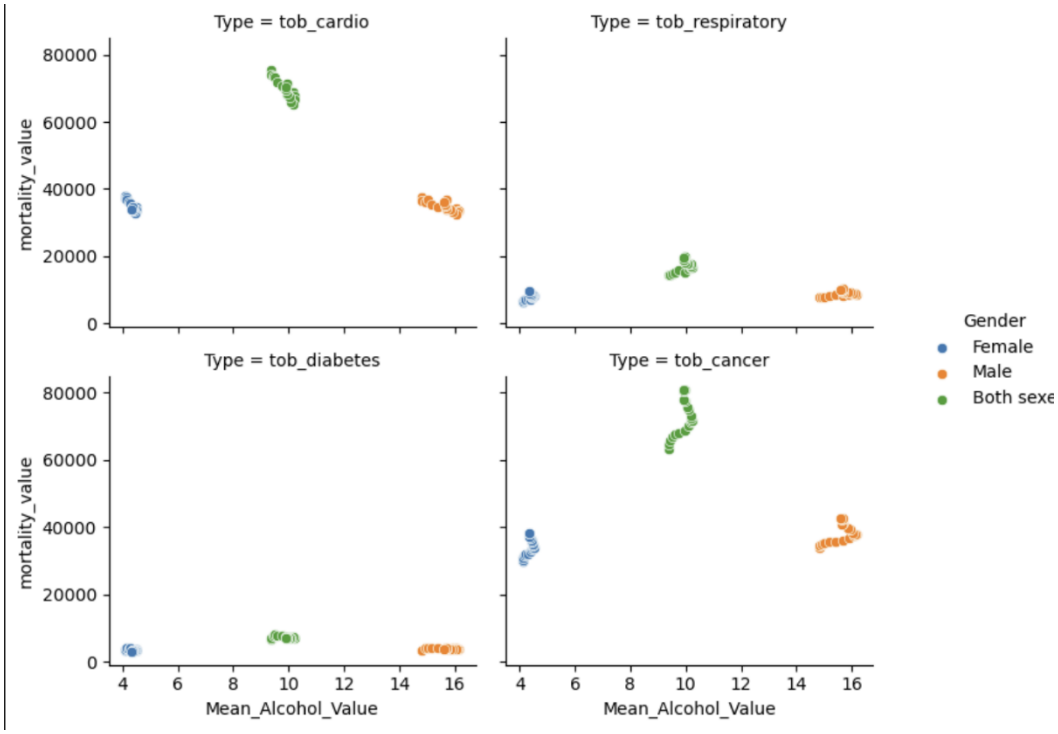

Figure 1: Facet Grid showing correlation of Alcohol with NCDs

Cholesterol

From the Cholesterol dataset, I learned more about the Cholesterol trend in Canada and its ranking compared to the rest of the world. This was done by forming a query that selected the Cholesterol values for Canada between the years of 2000 and 2018 as shown in Appendix 1.2.A.

The query showed that Cholesterol levels have been steadily decreasing for both the Male and Female population however the Canadian Male population has been experiencing high Cholesterol levels compared to females. I compared this trend to mortality for each NCD: Cardiovascular disease, Respiratory disease, Diabetes and cancer (Appendix 1.2.C/D/E/F). There was no distinct trend between the Mortality values and Cholesterol; however Cardiovascular disease mortality results also had a slight decrease since 2000 which aligns with the Cholesterol trend. The other three NCD had an increase in their mortality which contrasts with the Cholesterol trend in Canada. A rank for the highest Cholesterol level was done to look at the data on a larger scale by comparing the countries. I found that Canada was ranked 75th out of 191 countries in 2018 with an average Cholesterol level of 4.6 mmol/L for both gender populations combined. This is a significant improvement from 2000 when Canada was ranked 31st out of 191 countries with an average Cholesterol level of 5.2 mmol/L (Appendix 1.2.B). This is interesting to note as the WHO states that a high level of Cholesterol is diagnosed when it exceeds 5.0 mmol/L.

A facet grid was also derived to look at the correlation between each NCD and the Cholesterol level in Figure 2. This graph was used for visually purposes, and it clearly supports the earlier statements indication that there is a positive association between Cholesterol mortality caused by Cardiovascular disease whereas the rest of Respiratory disease and Malignant neoplasms (Cancer) show a negative association. Diabetes shows no pattern therefore this may indicate that Cholesterol does not impact mortality caused by Diabetes mellitus. This was possible by adjusting for the 'Disease' column in the mortality dataset (Appendix 1.2.G). Overall, by exploring our data and the facet grid, the results can suggest that Cholesterol can impact an individual's Cardiovascular health.

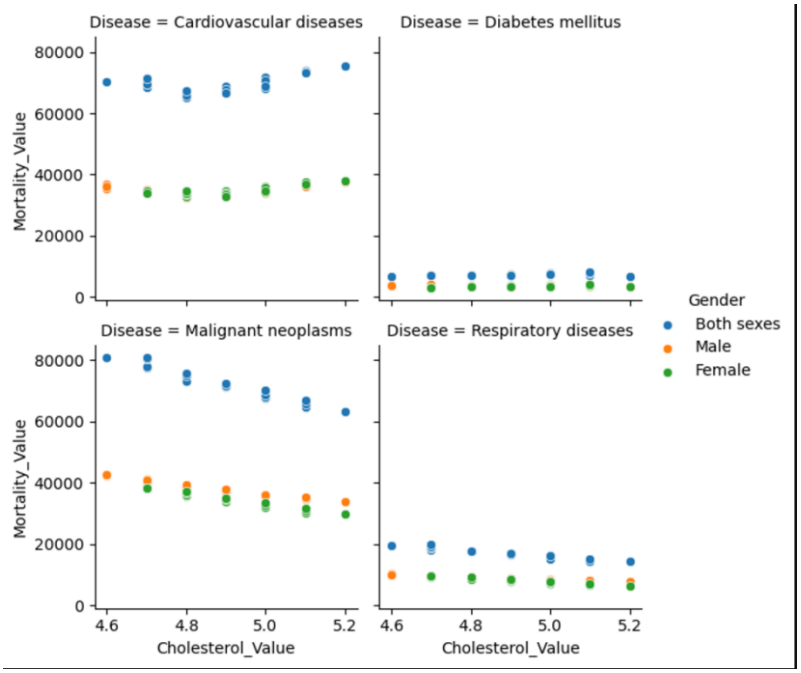

Figure 2: Facet Grid showing correlation of Cholesterol with NCDs

Hypertension

The Hypertension dataset was able to show that the Canadian population has been experiencing a consistent decrease from 2000 to 2018. This was done by forming a query that selected the Hypertension values for Canada between the years of 2000 and 2018 (Appendix 1.3.A). Similarly to Cholesterol, the Male population is more prevalent to Hypertension compared to Female population. The Male population have decreased from 30.8% to 24.5% whereas the Females have decreased from 25.6% to 20.2% over the span of 18 years. The Hypertension trend in Canada was also compared to each the mortality dataset for each NCD (Appendix 1.3.C/D/E/F). Similarly to Cholesterol, no distinct relationship was found, however Cardiovascular mortality rates have also been declining (Appendix 1.3.C). A rank for the highest Hypertension percentage was done and I found that Canada was ranked a surprisingly 192nd out of 195 countries in 2018 with an average Hypertension level of 22.4% for both genders. This indicates a recovery for Canada as the year 2000 resulted an average Hypertension value of 28.2% which ranked Canada at 180th worldwide (Appendix 1.3.B).

The facet grid was able to visually show the data trends in Hypertension compared to each NCD show in Figure 3. As mentioned in the introductory paragraph, this was done by joining the Hypertension data with Mortality data to get one table output by adjusting for the 'Disease' column in the mortality dataset (Appendix 1.3.G). Similarly to the Cholesterol dataset, it clearly supports the earlier statements indication that there is a positive association between Hypertension mortality caused by Cardiovascular disease. There is a drastic decrease in Malignant neoplasms (Cancer) which is strongly stating a negative association. Respiratory disease shows a steady decrease, therefore also indicating a negative association. Diabetes shows no trend as the datapoints sit at 0

indicating that Hypertension does not relate to mortality caused by Diabetes mellitus. In summary, these results can suggest that Hypertension is a significant risk factor for Cardiovascular health.

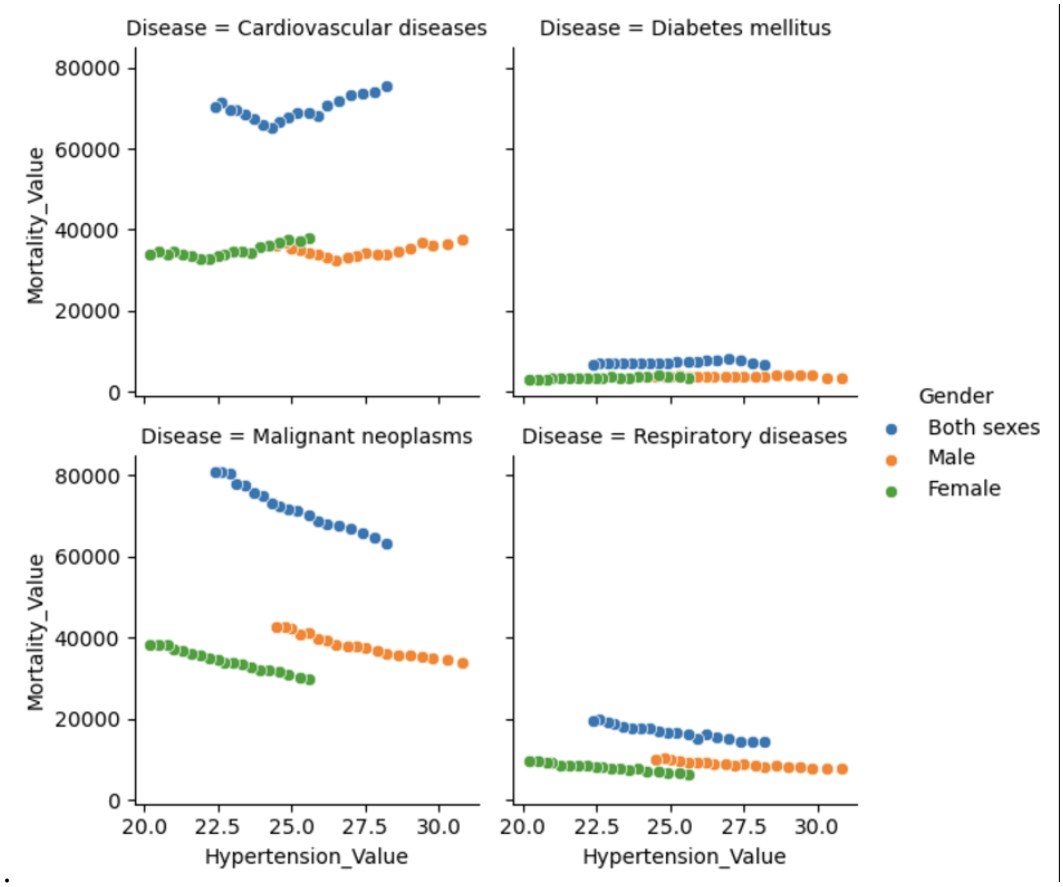

Figure 3: Facet Grid showing correlation of Hypertension with NCDs

Obesity

As mentioned briefly before, the Obesity dataset has shown a positive trend between 2000 to 2018 in Canada. Both genders have seen a consistent increase in Obesity values with males taking the lead in comparison to females. This relationship was seen and further analyzed using SQL queries as shown in Appendix 1.4.A. This was done by extracting the prevalence of Obesity values, the average and the low and high values, for the years 2000 to 2018 for both males and females in Canda. In 2000, Canada ranked 39th out of 198 countries with an Obesity value of 20.19% for both genders (Appendix 1.4.B). By 2018, Canadas rank had shifted to 70th with an increased Obesity value of 25.57% (Appendix 1.4.B). These rankings were organized in decreasing order. Despite the increase in Obesity values, the worldwide rank decreased, which might be indicative of an overall increasing trend on a global scale. On average, Canada's rank for Obesity for the 18-year trend is 52nd with an average value of 23.56% for both genders (Appendix 1.4.B). This analysis was done using three SQL queries to extract the required data. To further explore the effects and relationship between the prevalence of Obesity and the development

of NCDs defined as - Diabetes, Respiratory disease, Cardiovascular disease and Cancer - the dataset was merged with the NCD mortality dataset using four queries, one query for each disease (Appendix 1.4.C-F). A facet grid of four scatter plots was constructed using a SQL query and code (Appendix 1.4.G), each illustrating the relationship between Obesity prevalence and mortality values across four NCDs, with each plot corresponding to a specific type of disease. These relationships were categorized by gender. The facet grid is shown in Figure 4.

There appears to be a weak positive relationship between Obesity prevalence and mortality values for Cardiovascular diseases, where both sexes show higher mortality rates. Males and females show a similar consistent trend at lower mortality values. The relationship between Respiratory disease is subtle with lower mortality values for both males and females and slightly higher values for both sexes. Mortality values for Diabetes was low for males and females as well as both sexes, showing no strong correlation. Lastly, the trend for cancer was a stronger positive relationship in comparison to the other diseases, with both sexes observing a much higher trend in comparison to males and females individually.

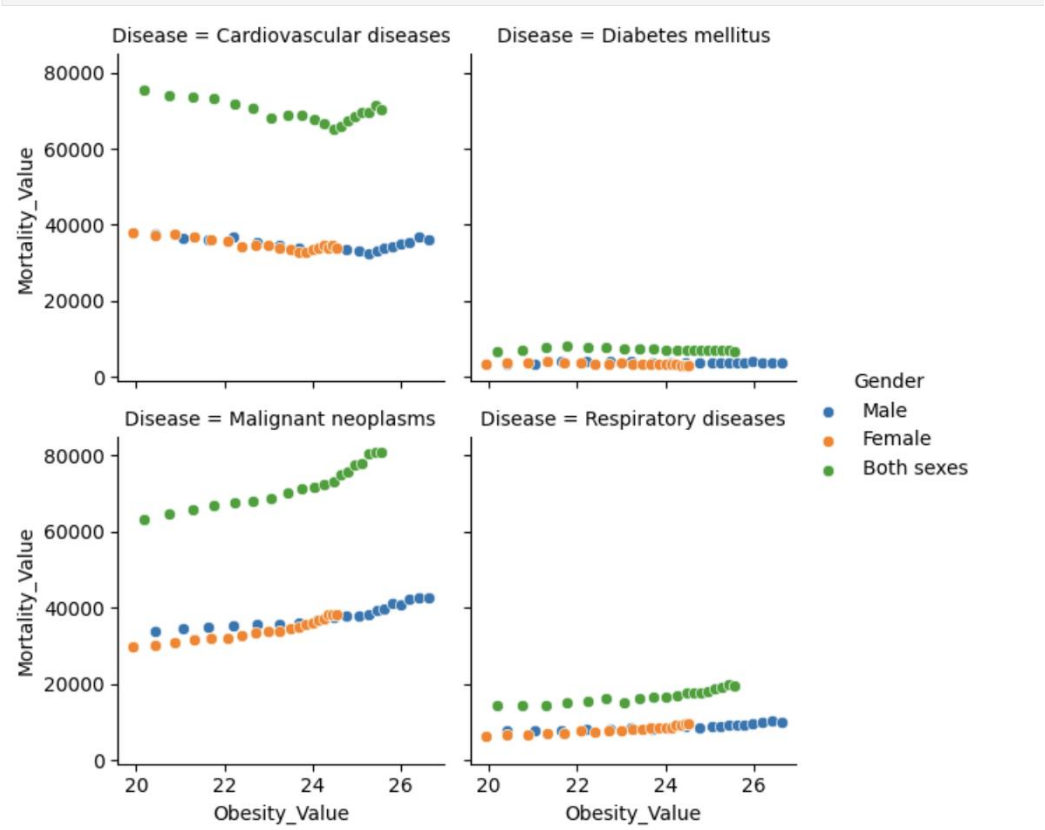

Figure 4: Facet Grid showing correlation of Obesity with NCDs

Physical Inactivity

From the Physical Inactivity dataset, I was able to learn and examine trends between 2000 to 2018. Firstly, all the irrelevant columns were dropped and only relevant columns as mentioned above were included. A query is performed on this dataset to extract specific columns like year, region, Physical Inactivity value, etc... of both males and females within the timeframe. From the table, I analyzed that there is a gradual increase of Physical Inactivity over these years specifically females having higher rates than males. The lowest rates were observed in the year 2000 when the female Physical Inactivity rate was 28% and that of males was 25%. In the next 18 years, those numbers increased to 37% for females and 33% for males reflecting a consistent inactivity. Physical Inactivity has been recognized as one of the leading risk factors for NCDs, which comprise Cardiovascular diseases, Diabetes, and Cancer.

To look in-depth and analyze more, we compared Canada among all 195 countries in physical inactivity aspect. Canada stood in 88th position out of all 195 countries in the year 2000. As physical inactivity increased, gradually the rank decreased to 45 in the year 2018. When we look at overall physical inactivity for both males and females in the timestamp of 2000 to 2018, it stands at 66th position across the world. The queries and results are given below. The Rank Queries are shown in Appendix 1.5.B

We compared this trend to mortality rates for these NCDs and found mixed results: Mortality due to Cardiovascular diseases and Diabetes has a constant trend with both increasing and decreasing slightly. Whereas Mortality due to Respiratory diseases and Cancer has increased since 2000. The queries for merging Physical Inactivity and mortality for each specific disease are mentioned in Appendix 1.5.C, 1.5.D, 1.5.E and 1.5.F. Finally, we plotted a facet grid that shows four plots for four different diseases that is affected by Physical Inactivity. It is shown in Figure 5.

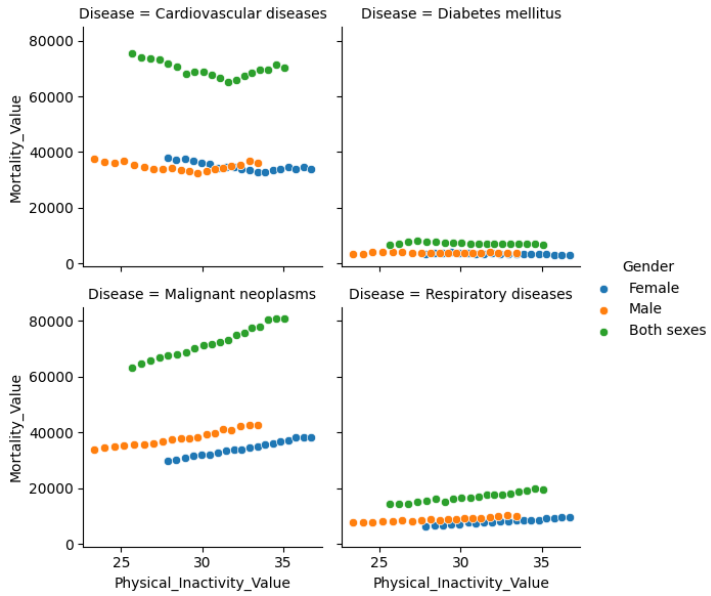

Figure 5: Facet Grid showing correlation of Physical Inactivity with NCDs

Tobacco

Firstly, data wrangling was performed on this dataset, irrelevant columns were dropped, and relevant columns were renamed as discussed earlier in the data section. Since this dataset had missing values for years 2001, 2002, 2003, 2004, 2006, 2008, 2009, 2011, 2012, 2013, 2014, 2016, 2017, and 2018 we decided to fill them out with the average of the adjacent years as the adjacent years would have similar populations and awareness on harmfulness of Tobacco. Looking at the Appendix 1.6.B, we learned that Tobacco use has been declining in the past eighteen years for both male and female, with the all-time lowest use of 10.15% for female population and 14.40% for male population in 2018. While 2000 having the highest values, with 31.00% for female and 36.10% in male population. We also wanted to learn that where Canada stands in terms of Tobacco use compared to the rest of the world. As seen in the Appendix 1.6.C, Canada Ranks 104th on average Tobacco use in the past eighteen years from 2000-2018 among 165 countries. For the year 2000 it ranked 92nd and for the year 2018 the ranked dropped to 116^{th,} indicating that fortunately Tobacco use has been declining since 2000. We then joined this dataset with mortality dataset to see the correlation of Tobacco with each of the NCDs. From Appendix 1.6.D, we learned that mortality rate for Cardiovascular disease has been decreasing as Tobacco use decreased but for 2017 and 2018 there is a small spike in mortality. While the number of deaths attributed to Respiratory diseases and Malignant neoplasm (Cancer) has been increasing over the years, Tobacco use has been reduced (Appendix 1.6.G). We could not see any pattern with Diabetes as the mortality rate has been fluctuating up and down, as seen in Appendix 1.6.F. The Figure 6 is the visualization of the tables seen in Appendix 1.6.D, 1.6.E, 1.6.F, 1.6.G. The code for this Figure 6 has been given in Appendix 1.6.H.

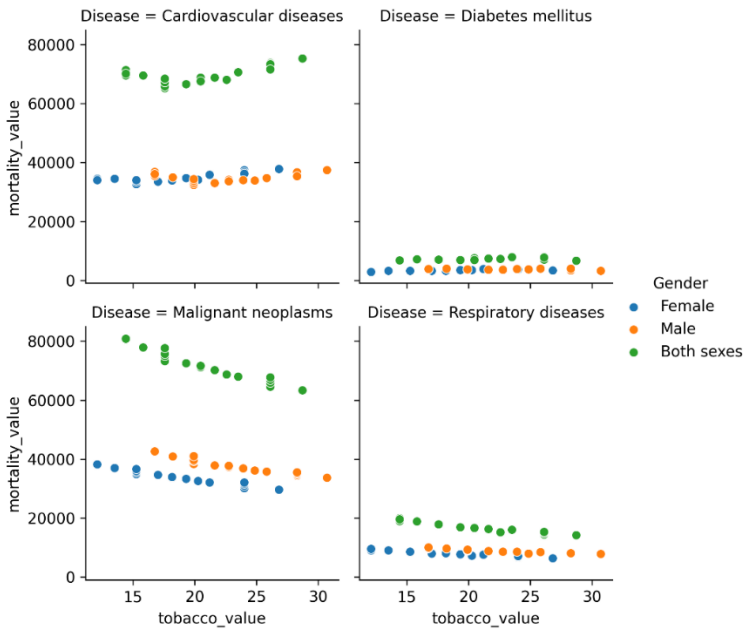

Figure 6: Facet Grid showing correlation of Tobacco with NCDs

Guiding Question 1: Behavioral Factors

Behavioral risk factors are modifiable behaviors and lifestyle choices that increase the likelihood of developing NCDs. This guiding question focuses on analyzing the trends in behavioral risk factors-Physical Inactivity, Obesity and Alcohol consumption from 2000 to 2018. To observe these trends, scatter plots will be utilized using SQL queries where both females and males as well as both genders will be selected for data values in Canada from 2000 to 2018. The trend for Physical Inactivity is positive with females having higher Physical Inactivity rates than males (Figure 7). The Tobacco usage scatter plot shown in Figure 8 illustrates a negative relationship over the 18 years with females having lower consumption rates than males. Finally, the Alcohol consumption plot shown in Figure 9 displays a steady trend over the 18 years, indicating a consistent pattern in Alcohol intake with no significant increases or decreases during this period. Additionally, the data reveals that males consistently had higher Alcohol consumption levels compared to females throughout the observed years. This is also shown in the outputs from our queries in Appendix 1.1.A, 1.5.A, and 1.6.A.

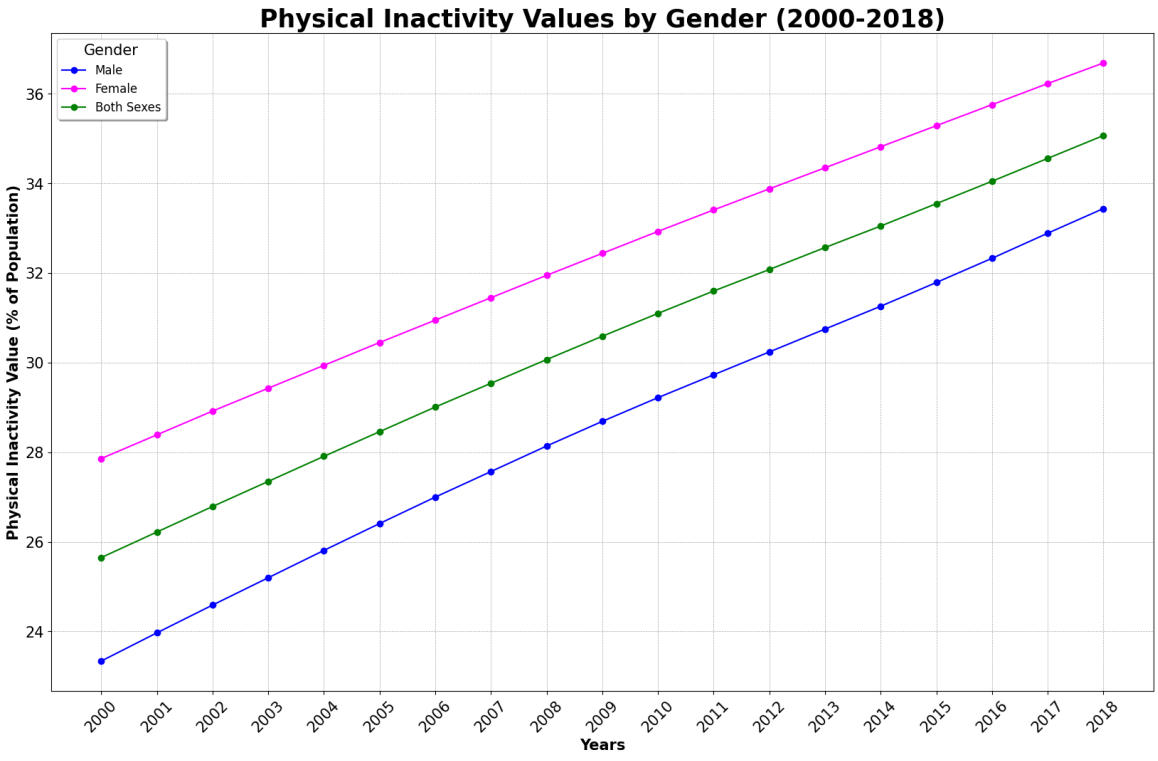

Figure 7: Physical Inactivity Trend overtime (2000 – 2018)

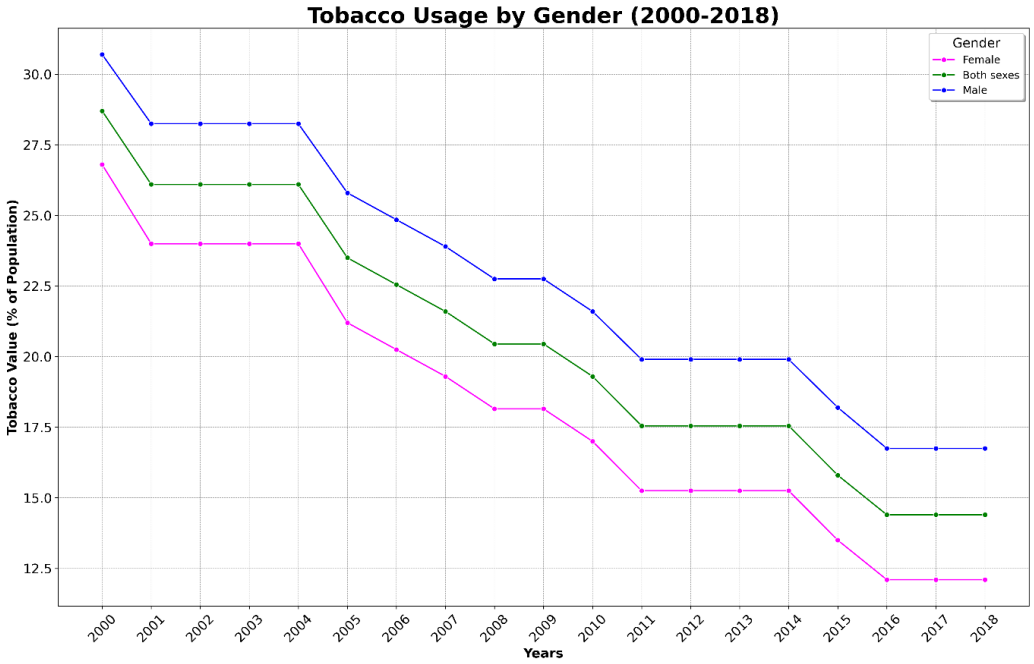

Figure 8: Tobacco Trend overtime (2000 – 2018)

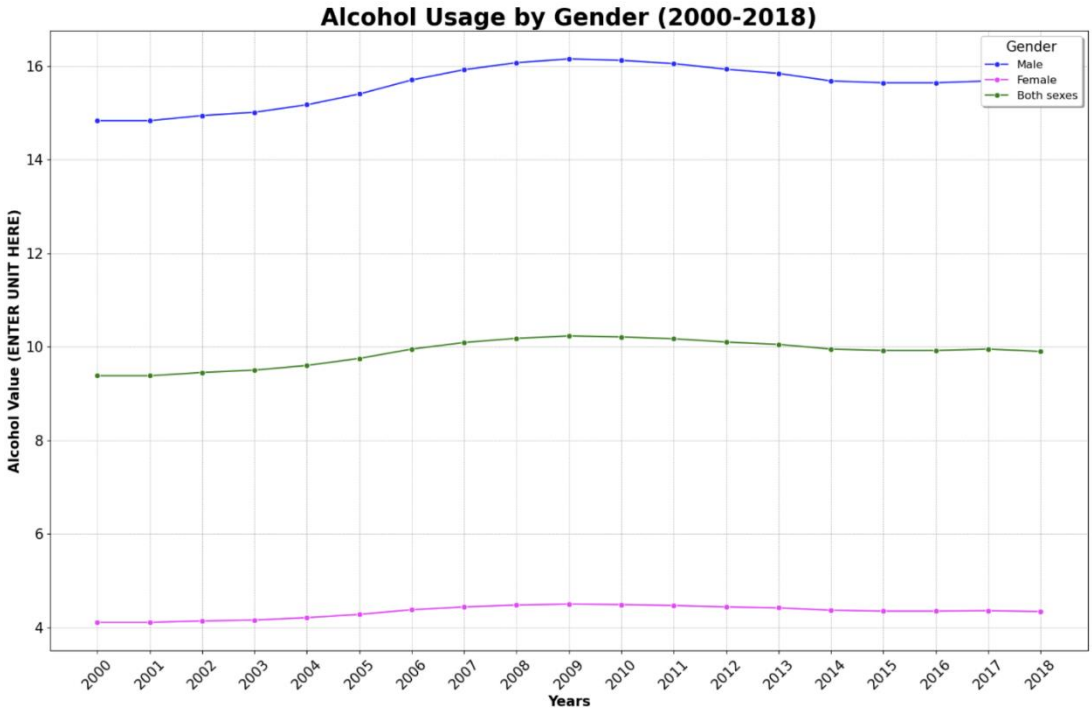

Figure 9: Alcohol Trend overtime (2000 – 2018)

Guiding Question 2: Metabolic Factors

In relation to the research topic "Examining NCD's in Canada", one of the focuses was to analyze the metabolic risk factor trends from 2000 to 2018 for both genders. As a reminder, metabolic factors are physiological factors that are generally genetic and are therefore uncontrollable. The required queries selected data from Gender, Year, Average Factor Value, and Country from the following datasets: Hypertension, Cholesterol and Obesity. The output was then converted to pandas for graphing to showcase the trend visually. It is concluded that Hypertension and Cholesterol have a decreased trend as seen in Figures 10 and 11, whereas Obesity has an increased trend as shown in Figure 12. Interestingly enough, the Male population is more prevalent to Hypertension, Cholesterol and Obesity compared to females in all metabolic trends. This is also shown in the outputs from our queries in Appendix 1.2.A, 1.3.A, and 1.4.A.

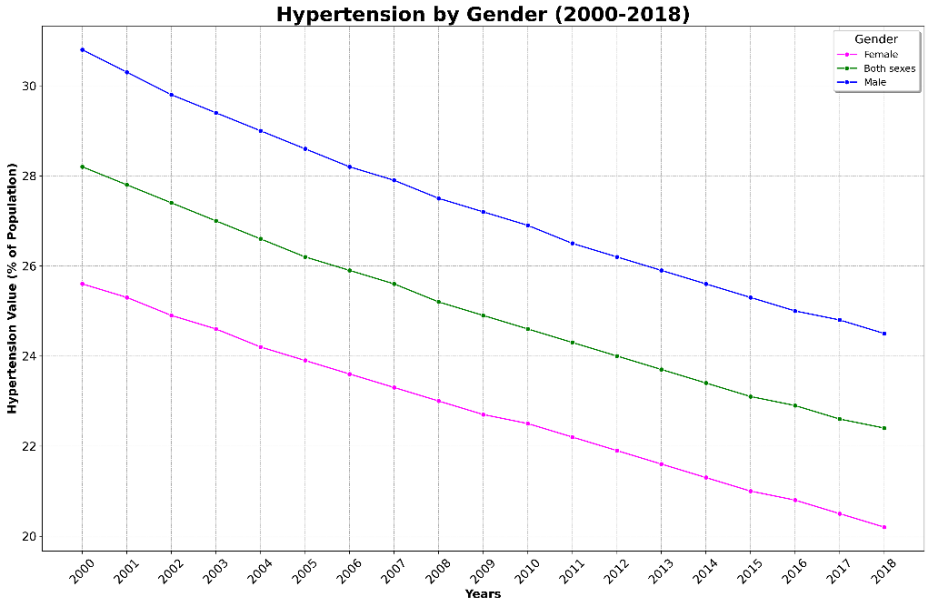

Figure 10: Hypertension Trend overtime (2000 – 2018)

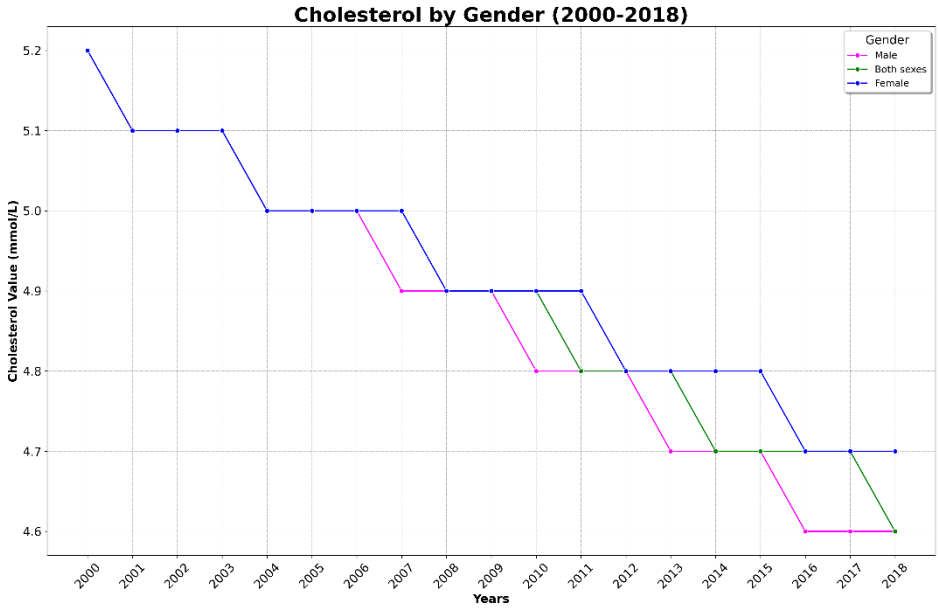

Figure 11: Cholesterol Trend overtime (2000 – 2018)

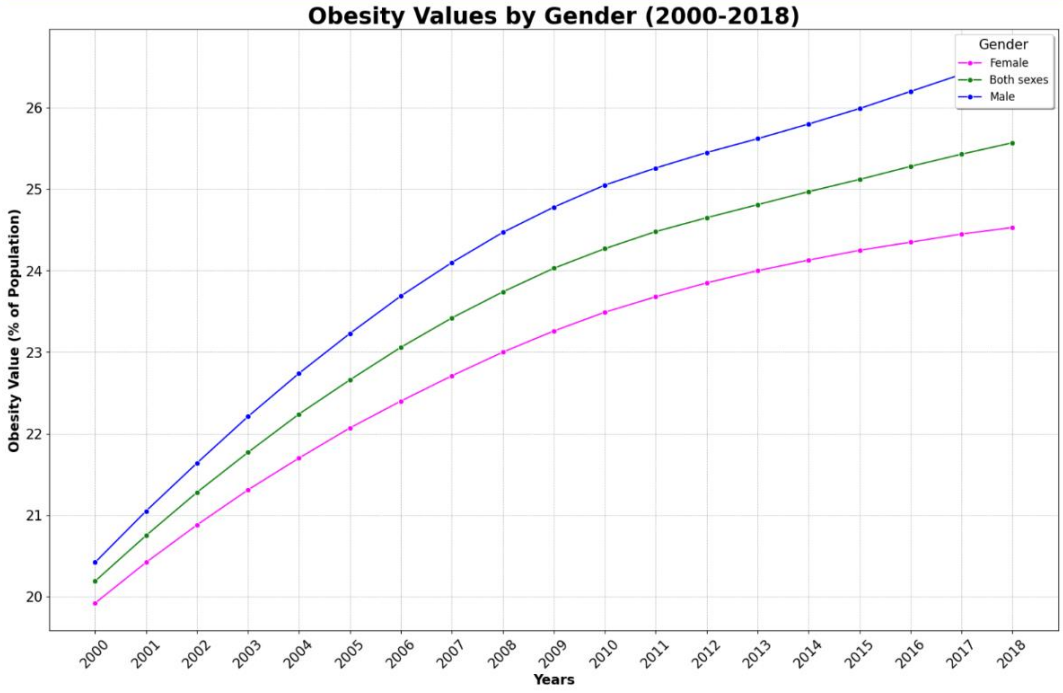

Figure 12: Obesity Trend overtime (2000 – 2018)

Guiding Question 3. Which Factors are Strongly Correlated?

A correlation matrix has been plotted to understand what factors are strongly correlated in order to focus on the most impactful areas where Behavior and Metabolic conditions are interconnected. Query has been performed to join all the 6 tables that included both Metabolic and Behavioral risk factors. The query for joining all factors is shown in Appendix 4.1. Using that combined table, correlation matrix is plotted for which the code is shown in the Appendix 4.2 and the results are as follows:

There is a moderate correlation of 0.53 between Physical Inactivity and Obesity, suggesting that people who are less Physically active are likely to be more obese. Next, Tobacco has a strong correlation of 0.79 with Cholesterol and 0.89 with Hypertension that indicates a clear relationship between smoking and these health conditions. Similarly, Alcohol shows significant correlation of 0.77 with Hypertension, suggesting that excessive Alcohol use leads to Hypertension. From the correlation matrix shown in Figure 13, the strongest correlation pairs occur between both Behavioral and Metabolic risk factors.

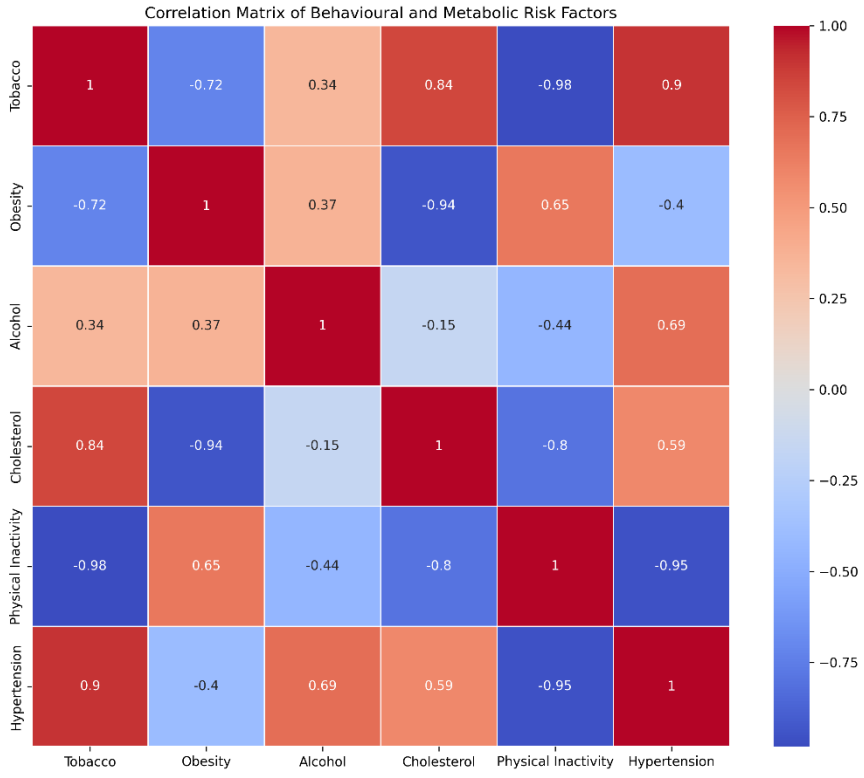

Figure 13: Correlation Matrix of all Risk factors

This analysis reviews the relationship between the Metabolic and Behavioral risk factors. For that join query is used to merge the datasets and see the pattern. First, Alcohol dataset is merged with three metabolic risk factors individually. The query we used for joining all these risk factors individually is shown in the Appendix 4.3

In this first plot of relationship between Alcohol and Cholesterol level, the graph shows how Cholesterol level changes at different levels of Alcohol consumption across the three different genders: Female, Male, and Both sexes. Male subjects show maximum levels of Cholesterol at every Alcohol consumption level while clustering at the top of the line. Females remain consistently lower than males and both sexes. Amongst men, the peak Cholesterol levels were observed at higher levels of Alcohol consumption closer to 5.0, and females at the lower levels recorded lower Cholesterol levels at around 4.6 to 4.7.

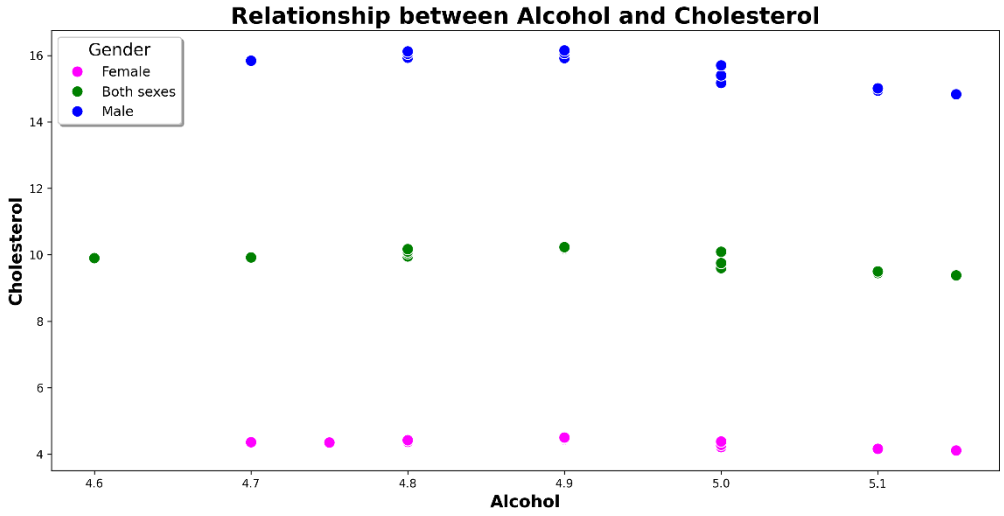

Figure 14: Correlation between Alcohol and Cholesterol

Next plot shows the relationship between Pure Alcohol usage and Hypertension, with males consuming more Alcohol have higher blood pressure compared to females having lower blood pressure as they consume less Alcohol.

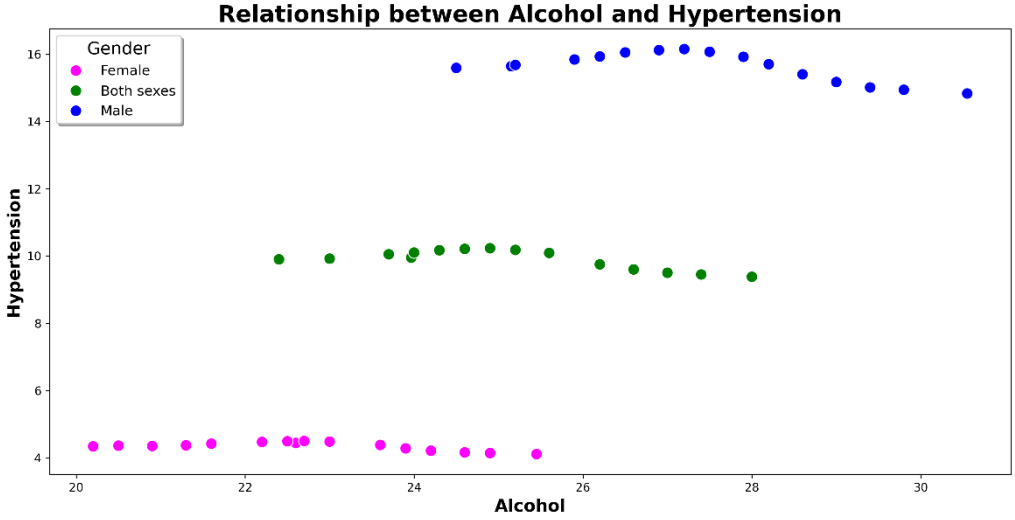

Figure 15: Correlation between Alcohol and Hypertension

The last plot representing the relationship between Alcohol and Obesity shows that the level of Obesity goes up in males with excessive consumption of Alcohol. Males, again, show high values compared to females for all Alcohol levels, with Obesity peaking up as Alcohol consumption increases. Both sexes show moderate values of Obesity, following a similar pattern with low overall values. The female population have always shown low values of Obesity through all levels of Alcohol consumption. The findings imply that the level of Alcohol consumption, along with Obesity itself, somehow has a stronger effect on males than on females.

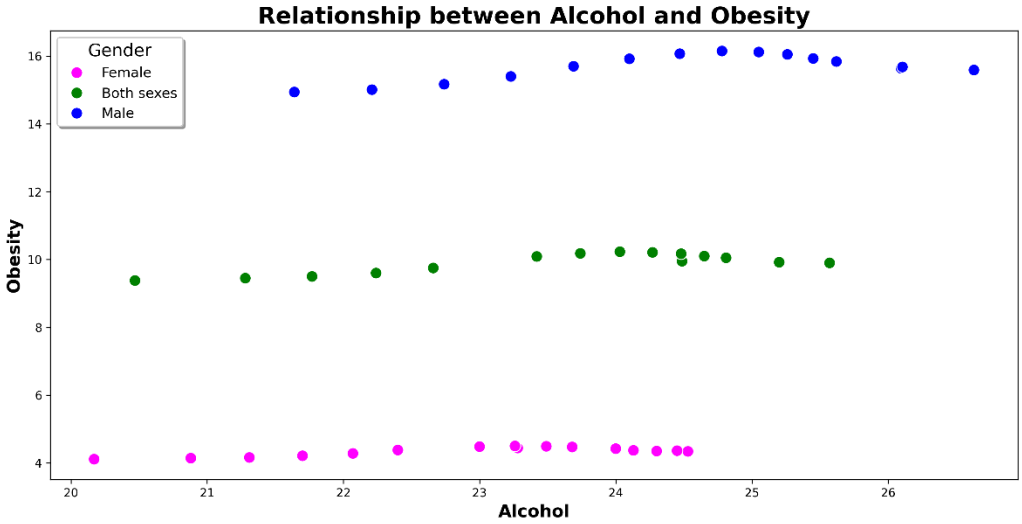

Figure 16: Correlation between Alcohol and Obesity

From the analysis between Physical Inactivity and Cholesterol, Cholesterol levels tend to increase with increased inactivity. Males recorded high Cholesterol levels in comparison to females despite the level of inactivity and in both genders combined. Male Cholesterol values are higher than Female still show a definite gender difference in this phenomenon. This might mean that even though inactivity is strongly associated with Cholesterol levels, the effect on those levels is mediated by gender.

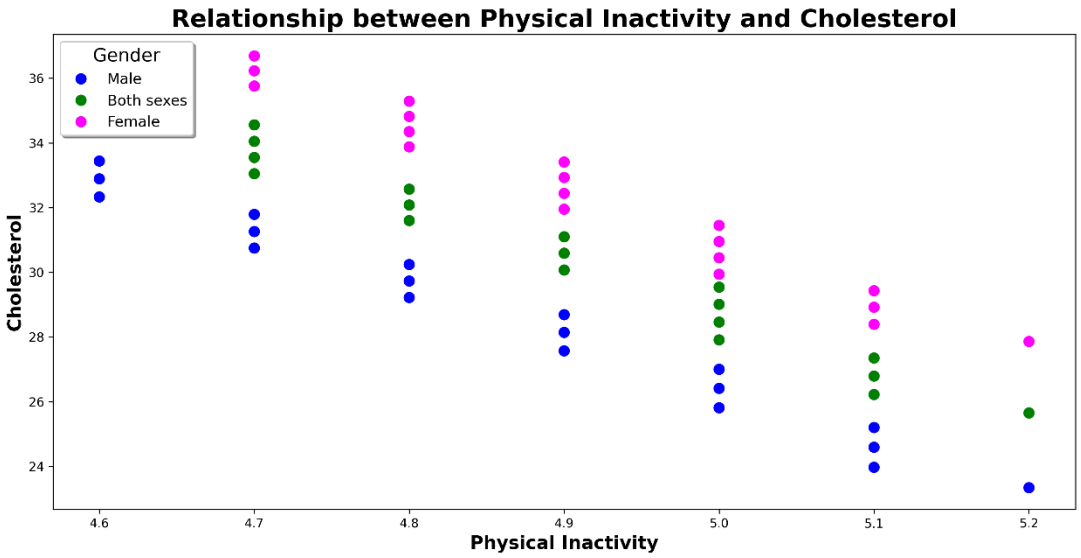

Figure 17: Correlation between Physical Inactivity and Cholesterol

The plot shows a reverse correlation between Physical Inactivity and Hypertension; Thus, it has an inverse relation with Physical activity. On all the different levels of inactivity, the Male score of Hypertension was always higher than that of Females and even a combination of Both sexes. The lowest scores of Hypertension were recorded for Females. Therefore, Physical Inactivity affects the Hypertension considerably negatively.

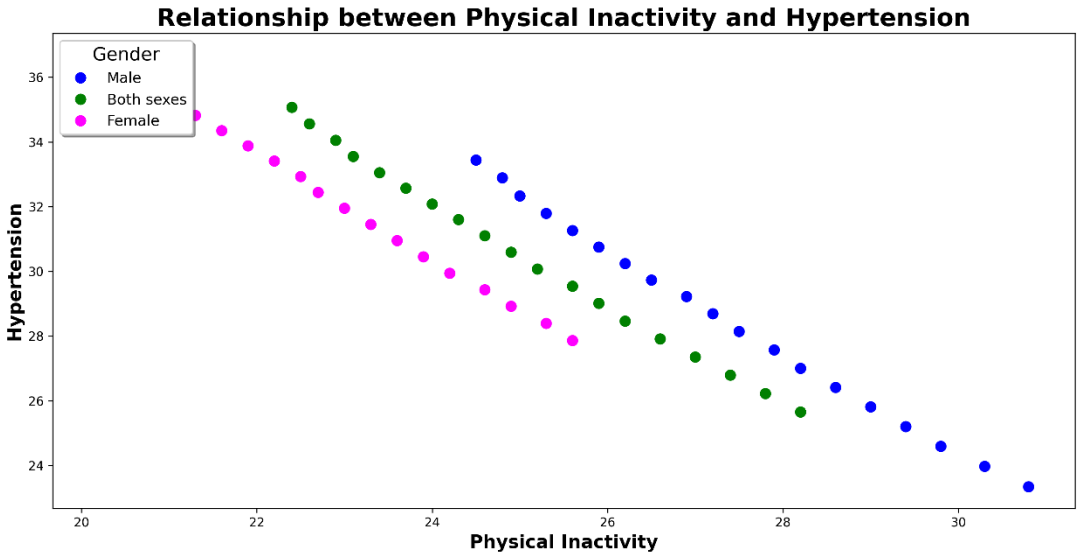

Figure 18: Correlation between Physical Inactivity and Hypertension

The last plot with Obesity shows that Obesity has a positive relationship with Physical Inactivity. With Obesity, Physical Inactivity rises for each gender. Males showed the highest levels of Physical Inactivity, followed closely by both sexes combined; Females demonstrated the lowest

values at the same Obesity levels. This trend attended by gender differences demonstrates that as the levels of Obesity increase, levels of Physical Inactivity increase accordingly.

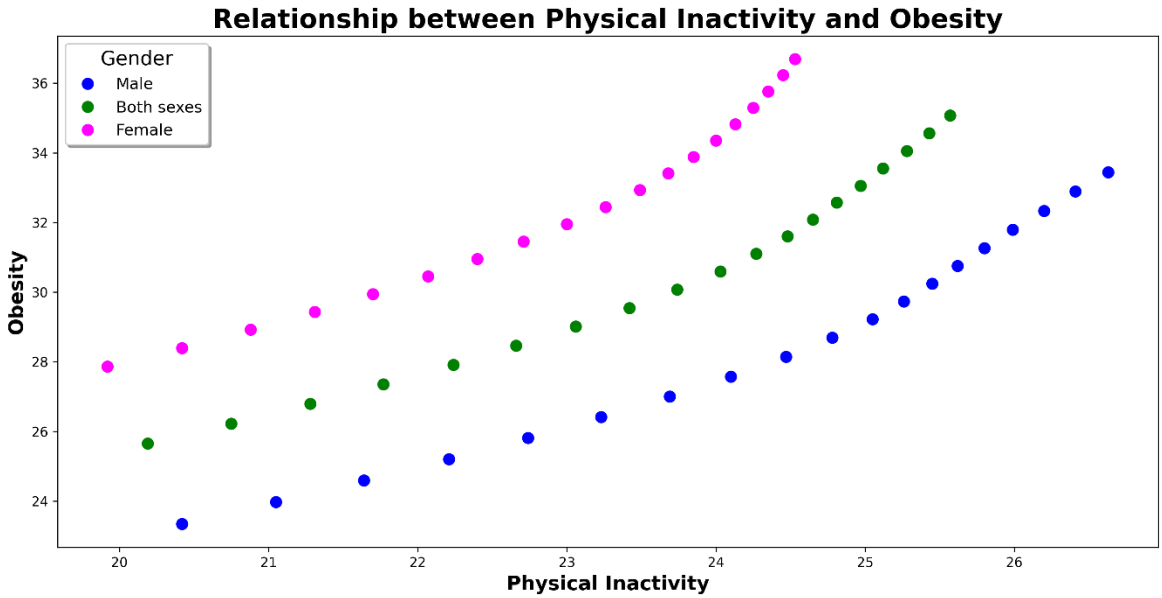

Figure 19: Correlation between Physical Inactivity and Obesity

This scatter plot represents a positive correlation between the use of Tobacco and Cholesterol levels. This indicates that the increase in Tobacco use has induced rising Cholesterol levels uniformly across each of the sexes. The highest Cholesterol levels were obtained for males, followed by males and females combined, while the lowest values were noted in females for the same Tobacco consumption levels. This kind of trend indicates a direct association whereas increased Tobacco consumption will lead to high Cholesterol levels.

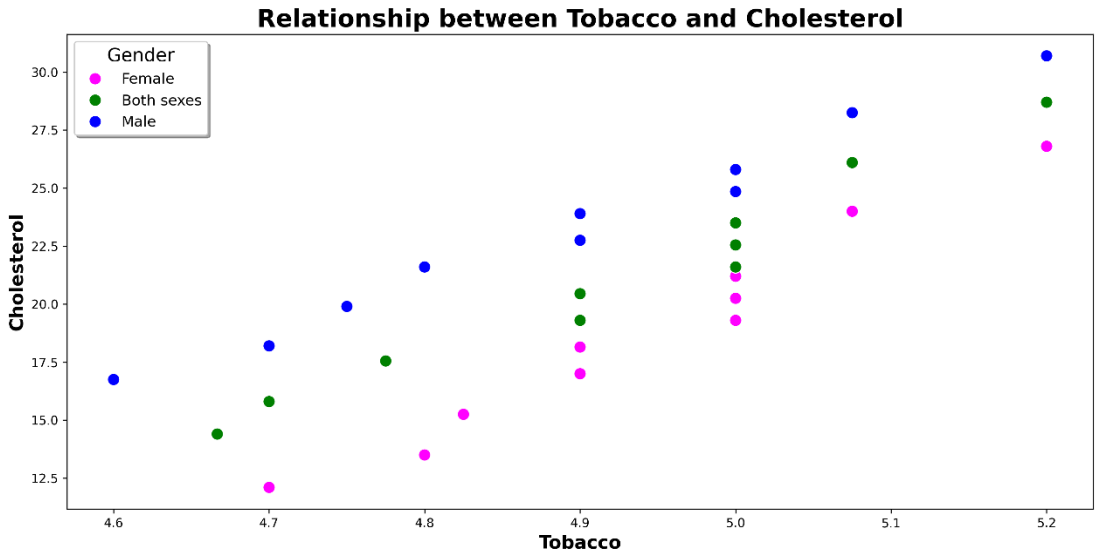

Figure 20: Correlation between Tobacco and Cholesterol

The second scatter plot with Tobacco indicates a positive relationship between Tobacco consumption and Hypertension. With the increase in Tobacco consumption, Hypertension levels also increase for every gender group. For Tobacco amounts that can be considered similar, males showed the highest values of Hypertension, followed by both sexes combined and finally females. These findings clearly show that a rise in Tobacco consumption leads to a rise in high blood pressure with specific differences arising due to gender.

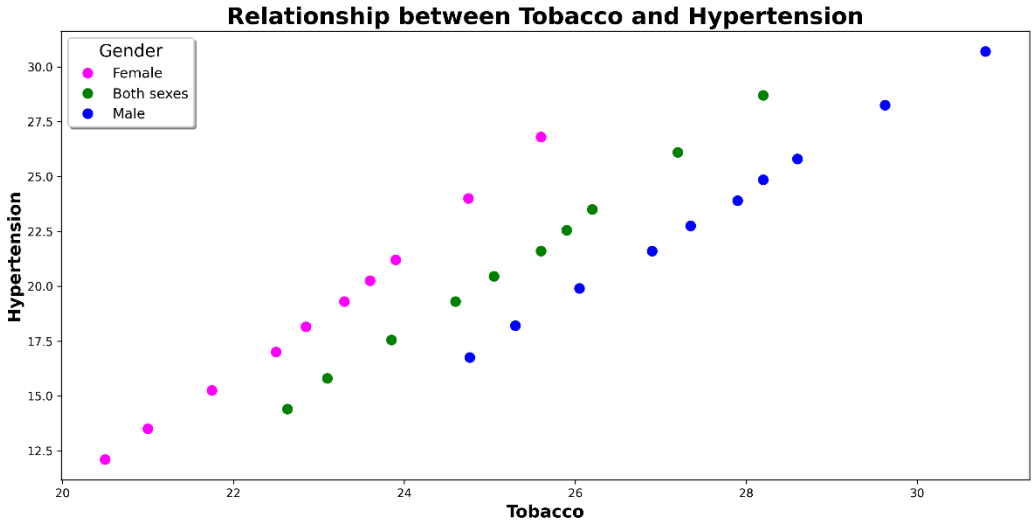

Figure 21: Correlation between Tobacco and Hypertension

The last scatter plot states that there is a negative relationship between Tobacco consumption and Obesity levels. As Tobacco consumption increases, Obesity levels decline for all genders. Furthermore, males have the highest levels of Obesity, with both sexes combined in the middle,

while females at that Tobacco level have the lowest levels of Obesity. Thus, it indicates that higher Tobacco consumption levels result in lower Obesity levels.

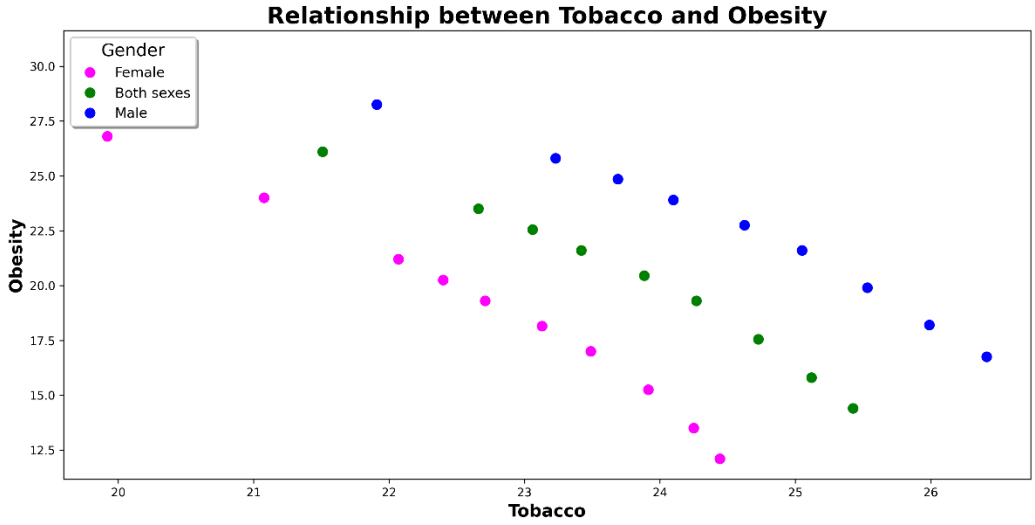

Figure 22: Correlation between Tobacco and Obesity

Guiding Question 4: Analysis of Metabolic and Behavioral Risk Factors for Each NCD: Multiple Regression

To investigate whether any of the metabolic and behavioral risk factors have an effect on NCD-caused mortalities, four linear regressions were performed, with each respective NCD as the response variable. The alpha value that we chose to use was 0.05. The outputs are shown below.

Respiratory Diseases:

	coef	std err	t	P>|t|	[0.025	0.975]
Tobacco	-107.1985	41.345	-2.593	0.014	-191.415	-22.982
Obesity	-456.2854	126.453	-3.608	0.001	-713.862	-198.708
Alcohol	261.7262	82.396	3.176	0.003	93.891	429.561
Cholesterol	-73.7842	1041.266	-0.071	0.944	-2194.774	2047.206
Physical Inactivity	478.5890	74.565	6.418	0.000	326.704	630.474
Hypertension	181.2237	172.772	1.049	0.302	-170.702	533.149

Figure 23: Multiple regression Analysis of all Risk factors vs Respiratory Diseases

Based on the model above, Tobacco use, Obesity, Alcohol consumption and Physical inactivity had p-values of 0.014, 0.001, 0.003 and 0.000, respectively, all of which are less than 0.05; therefore, we can conclude that these four risk factors play a significant role in causing mortalities when having Respiratory disease.

Cardiovascular Diseases:

	coef	std err	t	P>|t|	[0.025	0.975]
Tobacco	657.5965	156.897	4.191	0.000	338.007	977.186
Obesity	-1889.4363	479.871	-3.937	0.000	-2866.902	-911.971
Alcohol	69.2330	312.680	0.221	0.826	-567.676	706.142
Cholesterol	-8658.6348	3951.447	-2.191	0.036	-1.67e+04	-609.801
Physical Inactivity	2245.0469	282.964	7.934	0.000	1668.667	2821.427
Hypertension	1563.1026	655.644	2.384	0.023	227.599	2898.606

Figure 24: Multiple regression Analysis of all Risk factors vs Cardiovascular Diseases

Based on the model above, Tobacco use, Obesity, Cholesterol, Physical inactivity and Hypertension had p-values of 0.000, 0.000, 0.036, 0.000, and 0.023 respectively, all of which are less than 0.05; therefore, we can conclude that these four risk factors play a significant role in causing mortalities when having Cardiovascular disease.

Diabetes Mellitus:

	coef	std err	t	P>|t|	[0.025	0.975]
Tobacco	118.4934	36.220	3.271	0.003	44.715	192.272
Obesity	444.4644	110.780	4.012	0.000	218.814	670.115
Alcohol	-180.3123	72.183	-2.498	0.018	-327.345	-33.280
Cholesterol	-1909.1488	912.203	-2.093	0.044	-3767.246	-51.052
Physical Inactivity	-85.8011	65.323	-1.313	0.198	-218.860	47.258
Hypertension	178.8669	151.357	1.182	0.246	-129.438	487.172

Figure 25: Multiple regression Analysis of all Risk factors vs Diabetes Mellitus

Based on the model above, Tobacco use, Obesity, Alcohol and Cholesterol had p-values of 0.003, 0.000, 0.018, 0.044, respectively, all of which are less than 0.05; therefore, we can conclude that these four risk factors play a significant role in causing mortalities when having Diabetes mellitus.

Malignant Neoplasms:

	coef	std err	t	P>|t|	[0.025	0.975]
Tobacco	-22.4278	76.916	-0.292	0.772	-179.100	134.245
Obesity	-847.3202	235.248	-3.602	0.001	-1326.504	-368.136
Alcohol	558.1394	153.286	3.641	0.001	245.907	870.372
Cholesterol	-5804.3303	1937.122	-2.996	0.005	-9750.119	-1858.541
Physical Inactivity	1695.4028	138.718	12.222	0.000	1412.844	1977.962
Hypertension	1107.4487	321.417	3.446	0.002	452.743	1762.154

Figure 26: Multiple regression Analysis of all Risk factors vs Malignant Neoplasms (Cancer)

Based on the model above, Obesity, Alcohol use, Cholesterol, Physical inactivity and Hypertension had p-values of 0.001, 0.001, 0.005, 0.000 and 0.002, respectively, all of which are less than 0.05; therefore, we can conclude that these five risk factors play a significant role in causing mortalities when having Malignant neoplasms.

Discussion

Please note that this discussion includes a paragraph from each individual team member in order to showcase our individual learnings. How we would extend our analysis is explained after the individual paragraphs.

Ayda:

From this project, I gained a deeper understanding of the relationship between the prevalence of Obesity and its effect on an NCD. Throughout this process, I found out that mortality values for Malignant Neoplasm (Cancer) were positively correlated with Obesity while Diabetes Mellitus and Respiratory disease had a steady relationship with Obesity. Through analyzing the data, I improved my skills in creating queries in SQL, using the join function to join multiple tables and combining SQL queries with python for data analysis and visualization. Given that I do not have a computer science background, this project gave me the opportunity to deepen my understanding of SQL, write appropriate queries and develop a better understanding of database management.

Wardah:

From this project, I learned that even though Tobacco use has been reduced in the past years in Canada, it still plays a major role in the mortality rates of all four NCDs, especially Respiratory diseases. I also learnt that Tobacco being behavioral risk factor has strong impact on all three metabolic risk factors, specifically Cholesterol and Hypertension indicating consuming Tobacco increasing the blood pressure and Cholesterol levels. It was a little challenging handling

seven data sets together, but it was a new experience. I learnt a lot about the efficiency of SQL and how convenient data wrangling and manipulation can be compared to when done in python.

Safeen:

From this project, I was able to investigate the relationship between Alcoholism and the effect it has on contributing to NCD-caused deaths. Given the fact that Alcohol is commonly regarded as a lifestyle risk factor causing a wide variety of NCDs, it was surprising to see that there wasn't really a correlation between excess Alcoholism and mortality levels. Despite the lack of correlation, it was still determined that Alcohol seems to influence NCD-related deaths, as it was found to be a significant variable in the regression models for 3 of the 4 types of NCDs that were investigated in our project. This solidifies to me the notion that correlation and causation are two different avenues when it comes to the world of data and analytics. In terms of technical skills, this project helped refine my skills of using SQL. Through the merging of datasets and utilization of queries, I was able to effectively investigate the data and learn about trends regarding different factors. My skills in using different SQL functions, as well as implementing proper syntax, were all refined throughout the duration of this project. As someone who does not have a computer science background, I can confidently say that I am much more proficient in SQL than I was before. My python knowledge also improved, as I used it to create visualizations throughout the project.

Danae:

In this project, I was responsible for investigating the Cholesterol and Hypertension dataset. Exploring the datasets allowed me to become more proficient in both SQL and Python as I have no coding background. I believe SQL was the best approach for extracting the data and it was ideal that the two datasets that I analyzed had no missing values. The only necessary approach was to rename the columns for the queries and dropping certain columns. I noticed more improvement in exploring the datasets by calculating average for the different gender populations and ranking Canada compared to other countries worldwide. I became more proficient in SQL as I was outputting less errors towards the end of this project because I was more familiar with the syntax. Overall, analyzing this topic allowed me to get a deeper understanding and awareness of NCDs and what actions are necessary to minimize the risk of an NCD diagnosis.

Navya:

From this project, I have worked and analyzed physical inactivity dataset. As I have no prior knowledge regarding all these risk factors and non-communicable diseases, I got to learn how physical inactivity is measured. I have seen physical inactivity trend increased from 2000 to 2018. I became comfortable with joining tables using SQL. I have noticed physical inactivity which is a behavioral risk factor has significant effect on Respiratory diseases, Cancer and Cardiovascular diseases. Diabetes is consistent among all years with physical inactivity. For the multiple regression analysis, all the 7 datasets are merged. It is a difficult task for us all. But, I learnt how to efficiently merge all these datasets.

As a group we discussed a few ways of extending this project in the future to provide more valuable insights. To further improve this project, a Glucose dataset with up-to-date data could be incorporated to examine the overall trend over the years and to examine its effect on the NCDs, more importantly on Diabetes Mellitus. Another way to extend this project is to incorporate more factors to receive more accurate data for NCDs and to collect data from recent years as the group was forced to stop the analysis beyond 2018 due to missing data. A dietary pattern and Physical activity levels dataset could be incorporated with Obesity to get a more holistic understanding of Obesity trends. The Malignant neoplasm mortality dataset had certain limitations that impacted our analysis. The dataset was not specific to a certain type of cancer but rather aggregated mortality data for all types of cancers combined. This lack of precision made it challenging to investigate the relationship between cancer and other risk factors

Conclusion

In this analysis, we examined "Metabolic and Behavioral Risk Factors for Non-Communicable Diseases (NCDs) in Canada." Data extraction was performed using SQL, while Python was used to create visualizations. Our analysis revealed a significant correlation between metabolic and behavioral factors, indicating that lifestyle behaviors have a substantial impact on metabolic health.

Key findings from the regression analysis include the following: Tobacco use, Obesity, Alcohol Consumption and Physical Inactivity play a significant role in predicting Respiratory Disease mortalities; Tobacco use, Obesity, Cholesterol, Physical Inactivity and Hypertension can significantly predict Cardiovascular disease mortalities; Tobacco use, Obesity, Alcohol and Cholesterol have a significant role in projecting Diabetes Mellitus-caused mortalities; and finally, Obesity, Alcohol use, Cholesterol, Physical Inactivity and Hypertension are significant variables in projecting Malignant Neoplasm-caused deaths. By examining trends in Canada, we observed healthier values for Hypertension and Cholesterol compared to the early 2000s, likely due to increased awareness and education. Alcohol consumption has shown a linear trend indicating that the Canadian population has neither increased nor decreased in overall consumption. However, Obesity and Physical Inactivity showed an upward trend, signaling areas that require attention. When compared globally, Canada ranks positively in most factors but demonstrates a clear need for improvement in addressing Obesity and Physical inactivity. These findings highlight opportunities for targeted public health interventions and policies to further improve Canada's health outcomes.

References:

World Health Organization. (n.d.). Noncommunicable diseases: Risk factors. WHO. https://www.who.int/data/gho/data/themes/topics/topic-details/GHO/ncd-risk-factors.

Appendix:

Alcohol:

Appendix 1.1.A: SQL queries for determining Canada's average Alcohol consumption rank in 2000, 2018, and the 19-year average for ranking alcohol consumption

```sql
# creating query to investigate the world rank of Canada in terms of alcohol consumption for the year 2000
query_alc_trend_2000 = """
WITH RankedAlcohol AS (
    SELECT
        Country,
        Mean_Alcohol_Value,
        RANK() OVER (ORDER BY Mean_Alcohol_Value DESC) AS rank_value
    FROM
        alcohol_levels
    WHERE
        Gender = 'Both sexes' AND Year = 2000
)
SELECT
    Country,
    Mean_Alcohol_Value,
    rank_value
FROM
    RankedAlcohol
WHERE
    Country = 'Canada';
    """
alc_trend_2000 = pd.read_sql_query(query_alc_trend_2000, engine)
print("Canada's Average Alcohol Rank during 2000 out of 189 countries")
alc_trend_2000
```

```
Canada's Average Alcohol Rank during 2000 out of 189 countries
```

	Country	Mean_Alcohol_Value	rank_value
0	Canada	9.38	40

```sql
# creating query to investigate the world rank of Canada in terms of alcohol consumption for the year 2018
query_alc_rank_2018 = """
WITH Country_Averages AS (
    SELECT
        Country,
        AVG(Mean_Alcohol_Value) AS Average_Alcohol_Value
    FROM
        alcohol_levels
    WHERE
        Gender = 'Both sexes' and Year = 2018
    GROUP BY
        Country
),
Ranked_Countries AS (
    SELECT
        Country,
        Average_Alcohol_Value,
        RANK() OVER (ORDER BY Average_Alcohol_Value DESC) AS Country_Rank
    FROM
        Country_Averages
)

SELECT
    Country,
    Average_Alcohol_Value,
    Country_Rank
FROM
    Ranked_Countries
WHERE
    Country = 'Canada'
ORDER BY
    Country_Rank DESC;

"""

alc_rank_2018 = pd.read_sql_query(query_alc_rank_2018, engine)
print("Canada's Average Alcohol Rank for 2018 out of 189 countries")
alc_rank_2018
```

Canada's Average Alcohol Rank for 2018 out of 189 countries

	Country	Average_Alcohol_Value	Country_Rank
0	Canada	9.9	33

```sql
# creating query to investigate the world rank of Canada in terms of alcohol consumption for the year 2018
query_alc_rank_trend = """
WITH Country_Averages AS (
    SELECT
        Country,
        AVG(Mean_Alcohol_Value) AS Average_Alcohol_Value
    FROM
        alcohol_levels
    WHERE
        Gender = 'Both sexes'
    GROUP BY
        Country
),
Ranked_Countries AS (
    SELECT
        Country,
        Average_Alcohol_Value,
        RANK() OVER (ORDER BY Average_Alcohol_Value DESC) AS Country_Rank
    FROM
        Country_Averages
)

SELECT
    Country,
    Average_Alcohol_Value,
    Country_Rank
FROM
    Ranked_Countries
WHERE
    Country = 'Canada'
ORDER BY
    Country_Rank DESC;

"""

alc_rank_trend = pd.read_sql_query(query_alc_rank_trend, engine)
print("Canada's Average Alcohol Rank out of 189 countries between 2000-2018")
alc_rank_trend
```

Canada's Average Alcohol Rank out of 189 countries between 2000-2018

	Country	Average_Alcohol_Value	Country_Rank
0	Canada	9.879524	37

Appendix 1.1.B: SQL queries for determining Canada's average Alcohol consumption rank across Male and Females between 2000 to 2018

```
# creating query to investigate the overall world rank of Canada for alcohol consumption for the years 2000-2018
query_alc_rank_trend = """
SELECT
    Country,
    Year,
    Gender,
    Mean_Alcohol_Value,
    Alcohol_Value_Low,
    Alcohol_Value_High
FROM
    alcohol_levels
WHERE
    Country = 'Canada' and
    Gender != "Both sexes"
ORDER BY
    Year DESC;

"""

alc_rank_trend = pd.read_sql_query(query_alc_rank_trend, engine)
print("Canada's Average Alcohol Rank between 2000 to 2018 out of 189 countries")
alc_rank_trend
```

Canada's Average Alcohol Rank between 2000 to 2018 out of 189 countries

	Country	Year	Gender	Mean_Alcohol_Value	Alcohol_Value_Low	Alcohol_Value_High
0	Canada	2020	Male	15.59	10.64	20.70
1	Canada	2020	Female	4.31	2.87	5.67
2	Canada	2019	Male	15.59	10.63	20.50
3	Canada	2019	Female	4.34	2.93	5.78
4	Canada	2018	Male	15.59	10.63	20.50
5	Canada	2018	Female	4.34	2.93	5.78
6	Canada	2017	Male	15.68	10.75	20.68
7	Canada	2017	Female	4.36	2.95	5.82
8	Canada	2016	Male	15.64	10.85	20.75
9	Canada	2016	Female	4.35	2.96	5.84
10	Canada	2015	Male	15.64	10.84	20.64
11	Canada	2015	Female	4.35	2.97	5.82
12	Canada	2014	Male	15.68	10.86	20.59
13	Canada	2014	Female	4.37	2.98	5.80
14	Canada	2013	Male	15.84	10.80	20.77
15	Canada	2013	Female	4.42	3.00	5.85
16	Canada	2012	Male	15.93	10.89	20.93
17	Canada	2012	Female	4.44	3.03	5.92
18	Canada	2011	Male	16.05	11.11	21.05
19	Canada	2011	Female	4.47	3.10	5.94
20	Canada	2010	Male	16.12	11.35	21.21
21	Canada	2010	Female	4.49	3.17	5.98
22	Canada	2009	Male	16.15	11.35	21.13
23	Canada	2009	Female	4.50	3.16	5.93
24	Canada	2008	Male	16.07	11.02	21.08
25	Canada	2008	Female	4.48	3.08	5.90
26	Canada	2007	Male	15.92	10.77	20.86
27	Canada	2007	Female	4.44	2.98	5.85
28	Canada	2006	Male	15.70	10.62	20.75
29	Canada	2006	Female	4.38	2.93	5.80
30	Canada	2005	Male	15.40	10.39	20.45
31	Canada	2005	Female	4.28	2.86	5.73
32	Canada	2004	Male	15.17	10.35	20.11
33	Canada	2004	Female	4.21	2.87	5.63
34	Canada	2003	Male	15.01	10.31	19.88
35	Canada	2003	Female	4.16	2.83	5.58
36	Canada	2002	Male	14.94	10.40	19.89
37	Canada	2002	Female	4.14	2.85	5.57
38	Canada	2001	Male	14.83	10.14	19.79
39	Canada	2001	Female	4.11	2.77	5.55
40	Canada	2000	Male	14.83	10.14	19.79
41	Canada	2000	Female	4.11	2.77	5.55

Appendix 1.3.C: Trend between Alcohol Consumption and Cardiovascular Disease Mortality Values using SQL.

```
# creating query to investigate trend between Alcohol Usage and Cardiovascular Disease Mortalities between 2000 to 2018
query_alc_join_cardio = """
SELECT
    alcohol_levels.Country,
    alcohol_levels.Year,
    alcohol_levels.Gender,
    alcohol_levels.Mean_Alcohol_Value AS Alcohol_Value,
    mortality_levels.Mortality_Value AS mortality_value,
    mortality_levels.Disease
FROM
    alcohol_levels
JOIN
    mortality_levels
ON
    alcohol_levels.Year = mortality_levels.Year AND
    alcohol_levels.Gender = mortality_levels.Gender AND
    alcohol_levels.Country = mortality_levels.Country
WHERE
    alcohol_levels.Country = 'Canada' AND
    alcohol_levels.Year BETWEEN 2000 AND 2018 AND
    alcohol_levels.Gender = 'Both sexes' AND
    mortality_levels.Disease = 'Cardiovascular diseases';
"""

alc_join_cardio = pd.read_sql_query(query_alc_join_cardio, engine)
alc_join_cardio
```

	Country	Year	Gender	alcohol_value	mortality_value	Disease
0	Canada	2018	Both sexes	9.90	70147.0	Cardiovascular diseases
1	Canada	2017	Both sexes	9.95	71372.0	Cardiovascular diseases
2	Canada	2016	Both sexes	9.92	69553.0	Cardiovascular diseases
3	Canada	2015	Both sexes	9.92	69548.0	Cardiovascular diseases
4	Canada	2014	Both sexes	9.95	68507.0	Cardiovascular diseases
5	Canada	2013	Both sexes	10.05	67296.0	Cardiovascular diseases
6	Canada	2012	Both sexes	10.10	65860.0	Cardiovascular diseases
7	Canada	2011	Both sexes	10.17	65294.0	Cardiovascular diseases
8	Canada	2010	Both sexes	10.21	66670.0	Cardiovascular diseases
9	Canada	2009	Both sexes	10.23	67611.0	Cardiovascular diseases
10	Canada	2008	Both sexes	10.18	68827.0	Cardiovascular diseases
11	Canada	2007	Both sexes	10.09	68833.0	Cardiovascular diseases
12	Canada	2006	Both sexes	9.95	68097.0	Cardiovascular diseases
13	Canada	2005	Both sexes	9.75	70622.0	Cardiovascular diseases
14	Canada	2004	Both sexes	9.60	71670.0	Cardiovascular diseases
15	Canada	2003	Both sexes	9.50	73397.0	Cardiovascular diseases
16	Canada	2002	Both sexes	9.45	73741.0	Cardiovascular diseases
17	Canada	2001	Both sexes	9.38	73945.0	Cardiovascular diseases
18	Canada	2000	Both sexes	9.38	75398.0	Cardiovascular diseases

Appendix 1.3.D: Trend between Alcohol Consumption and Respiratory Disease Mortality Values using SQL.

```
# creating query to investigate trend between Alcohol Usage and Respiratory Disease Mortalities between 2000 to 2018
query_alc_join_resp = """
SELECT
    alcohol_levels.Country,
    alcohol_levels.Year,
    alcohol_levels.Gender,
    alcohol_levels.Mean_Alcohol_Value AS Alcohol_Value,
    mortality_levels.Mortality_Value AS Mortality_Value,
    mortality_levels.Disease
FROM
    alcohol_levels
JOIN
    mortality_levels
ON
    alcohol_levels.Year = mortality_levels.Year AND
    alcohol_levels.Gender = mortality_levels.Gender AND
    alcohol_levels.Country = mortality_levels.Country
WHERE
    alcohol_levels.Country = 'Canada' AND
    alcohol_levels.Year BETWEEN 2000 AND 2018 AND
    alcohol_levels.Gender = 'Both sexes' AND
    mortality_levels.Disease = 'Respiratory diseases';
"""

alc_join_resp = pd.read_sql_query(query_alc_join_resp, engine)
alc_join_resp
```

	Country	Year	Gender	Alcohol_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	9.90	19700.0	Respiratory diseases
1	Canada	2017	Both sexes	9.95	19973.0	Respiratory diseases
2	Canada	2016	Both sexes	9.92	18990.0	Respiratory diseases
3	Canada	2015	Both sexes	9.92	18873.0	Respiratory diseases
4	Canada	2014	Both sexes	9.95	17989.0	Respiratory diseases
5	Canada	2013	Both sexes	10.05	17848.0	Respiratory diseases
6	Canada	2012	Both sexes	10.10	17670.0	Respiratory diseases
7	Canada	2011	Both sexes	10.17	17578.0	Respiratory diseases
8	Canada	2010	Both sexes	10.21	16903.0	Respiratory diseases
9	Canada	2009	Both sexes	10.23	16665.0	Respiratory diseases
10	Canada	2008	Both sexes	10.18	16621.0	Respiratory diseases
11	Canada	2007	Both sexes	10.09	16335.0	Respiratory diseases
12	Canada	2006	Both sexes	9.95	15287.0	Respiratory diseases
13	Canada	2005	Both sexes	9.75	16040.0	Respiratory diseases
14	Canada	2004	Both sexes	9.60	15307.0	Respiratory diseases
15	Canada	2003	Both sexes	9.50	14938.0	Respiratory diseases
16	Canada	2002	Both sexes	9.45	14502.0	Respiratory diseases
17	Canada	2001	Both sexes	9.38	14278.0	Respiratory diseases
18	Canada	2000	Both sexes	9.38	14291.0	Respiratory diseases

Appendix 1.3.E: Trend between Alcohol Consumption and Diabetes Mellitus Mortality Values using SQL.

```
# creating query to investigate trend between Alcohol Usage and Diabetes Mellitus Mortalities between 2000 to 2018
query_alc_join_diab = """
SELECT
    alcohol_levels.Country,
    alcohol_levels.Year,
    alcohol_levels.Gender,
    alcohol_levels.Mean_Alcohol_Value AS Alcohol_Value,
    mortality_levels.Mortality_Value AS Mortality_Value,
    mortality_levels.Disease
FROM
    alcohol_levels
JOIN
    mortality_levels
ON
    alcohol_levels.Year = mortality_levels.Year AND
    alcohol_levels.Gender = mortality_levels.Gender AND
    alcohol_levels.Country = mortality_levels.Country
WHERE
    alcohol_levels.Country = 'Canada' AND
    alcohol_levels.Year BETWEEN 2000 AND 2018 AND
    alcohol_levels.Gender = 'Both sexes' AND
    mortality_levels.Disease = 'Diabetes Mellitus';
"""

alc_join_diab = pd.read_sql_query(query_alc_join_diab, engine)
alc_join_diab
```

	Country	Year	Gender	Alcohol_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	9.90	6883.0	Diabetes mellitus
1	Canada	2017	Both sexes	9.95	7002.0	Diabetes mellitus
2	Canada	2016	Both sexes	9.92	6962.0	Diabetes mellitus
3	Canada	2015	Both sexes	9.92	7230.0	Diabetes mellitus
4	Canada	2014	Both sexes	9.95	7091.0	Diabetes mellitus
5	Canada	2013	Both sexes	10.05	7056.0	Diabetes mellitus
6	Canada	2012	Both sexes	10.10	7003.0	Diabetes mellitus
7	Canada	2011	Both sexes	10.17	7232.0	Diabetes mellitus
8	Canada	2010	Both sexes	10.21	6964.0	Diabetes mellitus
9	Canada	2009	Both sexes	10.23	6938.0	Diabetes mellitus
10	Canada	2008	Both sexes	10.18	7542.0	Diabetes mellitus
11	Canada	2007	Both sexes	10.09	7423.0	Diabetes mellitus
12	Canada	2006	Both sexes	9.95	7311.0	Diabetes mellitus
13	Canada	2005	Both sexes	9.75	7916.0	Diabetes mellitus
14	Canada	2004	Both sexes	9.60	7864.0	Diabetes mellitus
15	Canada	2003	Both sexes	9.50	8028.0	Diabetes mellitus
16	Canada	2002	Both sexes	9.45	7929.0	Diabetes mellitus
17	Canada	2001	Both sexes	9.38	7154.0	Diabetes mellitus
18	Canada	2000	Both sexes	9.38	6759.0	Diabetes mellitus

Appendix 1.3.F: Trend between Alcohol Consumption and Malignant Neoplasms Mortality Values using SQL.

```
# creating query to investigate trend between Alcohol Usage and Malignant Neoplasm (cancer) Mortalities between 2000 to 2018
query_alc_join_cancer = """
SELECT
    alcohol_levels.Country,
    alcohol_levels.Year,
    alcohol_levels.Gender,
    alcohol_levels.Mean_Alcohol_Value AS Alcohol_Value,
    mortality_levels.Mortality_Value AS Mortality_Value,
    mortality_levels.Disease
FROM
    alcohol_levels
JOIN
    mortality_levels
ON
    alcohol_levels.Year = mortality_levels.Year AND
    alcohol_levels.Gender = mortality_levels.Gender AND
    alcohol_levels.Country = mortality_levels.Country
WHERE
    alcohol_levels.Country = 'Canada' AND
    alcohol_levels.Year BETWEEN 2000 AND 2018 AND
    alcohol_levels.Gender = 'Both sexes' AND
    mortality_levels.Disease = 'Malignant neoplasms';
"""

alc_join_cancer = pd.read_sql_query(query_alc_join_cancer, engine)
alc_join_cancer
```

	Country	Year	Gender	Alcohol_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	9.90	80881.0	Malignant neoplasms
1	Canada	2017	Both sexes	9.95	80994.0	Malignant neoplasms
2	Canada	2016	Both sexes	9.92	80696.0	Malignant neoplasms
3	Canada	2015	Both sexes	9.92	78020.0	Malignant neoplasms
4	Canada	2014	Both sexes	9.95	77763.0	Malignant neoplasms
5	Canada	2013	Both sexes	10.05	75755.0	Malignant neoplasms
6	Canada	2012	Both sexes	10.10	74983.0	Malignant neoplasms
7	Canada	2011	Both sexes	10.17	73328.0	Malignant neoplasms
8	Canada	2010	Both sexes	10.21	72535.0	Malignant neoplasms
9	Canada	2009	Both sexes	10.23	71726.0	Malignant neoplasms
10	Canada	2008	Both sexes	10.18	71197.0	Malignant neoplasms
11	Canada	2007	Both sexes	10.09	70257.0	Malignant neoplasms
12	Canada	2006	Both sexes	9.95	68729.0	Malignant neoplasms
13	Canada	2005	Both sexes	9.75	68005.0	Malignant neoplasms
14	Canada	2004	Both sexes	9.60	67746.0	Malignant neoplasms
15	Canada	2003	Both sexes	9.50	66883.0	Malignant neoplasms
16	Canada	2002	Both sexes	9.45	65980.0	Malignant neoplasms
17	Canada	2001	Both sexes	9.38	64668.0	Malignant neoplasms
18	Canada	2000	Both sexes	9.38	63429.0	Malignant neoplasms

Cholesterol

Appendix 1.2.A: SQL query for determining the trend of Cholesterol in Canada from 2000 to 2018.

```
#Cholesterol trend in Canada from 2000 to 2018
query = """ SELECT Country, Year, Gender, Cholesterol_Value, Cholesterol_Low_Value, Cholesterol_High_Value
FROM cholesterol_table
WHERE Year >= '2000'
  AND Year <= '2018'
  AND Country = 'Canada'
  AND Gender != 'Both sexes';"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Cholesterol_Value	Cholesterol_Low_Value	Cholesterol_High_Value
0	Canada	2018	Male	4.6	4.3	4.8
1	Canada	2018	Female	4.7	4.4	5.0
2	Canada	2017	Male	4.6	4.3	4.8
3	Canada	2017	Female	4.7	4.5	5.0
4	Canada	2016	Male	4.6	4.4	4.8
5	Canada	2016	Female	4.7	4.5	4.9
6	Canada	2015	Male	4.7	4.5	4.9
7	Canada	2015	Female	4.8	4.6	5.0
8	Canada	2014	Male	4.7	4.5	4.9
9	Canada	2014	Female	4.8	4.6	5.0
10	Canada	2013	Male	4.7	4.6	4.9
11	Canada	2013	Female	4.8	4.6	5.0
12	Canada	2012	Male	4.8	4.6	4.9
13	Canada	2012	Female	4.8	4.7	5.0
14	Canada	2011	Male	4.8	4.6	5.0
15	Canada	2011	Female	4.9	4.7	5.0
16	Canada	2010	Male	4.8	4.7	5.0
17	Canada	2010	Female	4.9	4.7	5.0
18	Canada	2009	Male	4.9	4.7	5.0
19	Canada	2009	Female	4.9	4.7	5.1
20	Canada	2008	Male	4.9	4.7	5.1
21	Canada	2008	Female	4.9	4.8	5.1
22	Canada	2007	Male	4.9	4.8	5.1
23	Canada	2007	Female	5.0	4.8	5.1
24	Canada	2006	Male	5.0	4.8	5.1
25	Canada	2006	Female	5.0	4.8	5.2
26	Canada	2005	Female	5.0	4.8	5.2
27	Canada	2005	Male	5.0	4.8	5.2
28	Canada	2004	Female	5.0	4.8	5.2
29	Canada	2004	Male	5.0	4.9	5.2
30	Canada	2003	Female	5.1	4.9	5.3
31	Canada	2003	Male	5.1	4.9	5.3
32	Canada	2002	Female	5.1	4.9	5.3
33	Canada	2002	Male	5.1	4.9	5.3
34	Canada	2001	Female	5.1	4.9	5.3
35	Canada	2001	Male	5.1	4.9	5.3
36	Canada	2000	Female	5.2	4.9	5.4
37	Canada	2000	Male	5.2	5.0	5.4

Appendix 1.2.B: SQL queries for ranking data in the year 2000, 2018 and the 18-year average ranking for Cholesterol.

[263]:
```
#Canada cholesterol rank compared to the rest of the world in 2000

query = """ WITH Ranked_Cholesterol AS (SELECT Country, Cholesterol_Value, RANK() OVER (ORDER BY Cholesterol_Value DESC) AS Rank_Cholesterol
    FROM cholesterol_table WHERE Gender = 'Both sexes' AND Year = 2000) SELECT Country, Cholesterol_Value, Rank_Cholesterol
    FROM Ranked_Cholesterol WHERE Country = 'Canada';"""

ranked_data = pd.read_sql_query(query, engine)
print("Canada rank for 2000 out of 191 countries")
ranked_data.head()
```

```
Canada rank for 2000 out of 191 countries
```

[263]:

	Country	Cholesterol_Value	Rank_Cholesterol
0	Canada	5.2	31

[243]:
```
#Canada cholesterol rank compared to the rest of the world in 2018

query = """ WITH Ranked_Cholesterol AS (SELECT Country, Cholesterol_Value, RANK() OVER (ORDER BY Cholesterol_Value DESC) AS Rank_Cholesterol
    FROM cholesterol_table WHERE Gender = 'Both sexes' AND Year = 2018) SELECT Country, Cholesterol_Value, Rank_Cholesterol
    FROM Ranked_Cholesterol WHERE Country = 'Canada';"""
ranked_data = pd.read_sql_query(query, engine)
print("Canada rank for 2018 out of 191 countries")
ranked_data.head()
```

```
Canada rank for 2018 out of 191 countries
```

[243]:

	Country	Cholesterol_Value	Rank_Cholesterol
0	Canada	4.6	75

[347]:
```
#Ranking countries based on cholesterol (from highest to lowest).

query = """WITH Ranked_Cholesterol AS (SELECT Country, AVG(Cholesterol_Value) AS Average_Cholesterol,
        RANK() OVER (ORDER BY AVG(Cholesterol_Value) DESC) AS Ranked_Cholesterol
    FROM cholesterol_table
    WHERE Gender = 'Both sexes' AND Year >= 2000 AND Year <= 2018
    GROUP BY Country)
SELECT Country, Average_Cholesterol, Ranked_Cholesterol
FROM Ranked_Cholesterol
WHERE Country = 'Canada';"""
print("AVG Canada rank for Cholesterol is out of 191 countries")
ranked_data = pd.read_sql_query(query, engine)
ranked_data.head()
```

```
AVG Canada rank for Cholesterol is out of 191 countries
```

[347]:

	Country	Average_Cholesterol	Ranked_Cholesterol
0	Canada	4.894737	54

Appendix 1.2.C: Trend between Cholesterol and Cardiovascular Disease Mortality Values using SQL.

```
#Looking at trend between cardiovascular disease mortality and cholesterol

query = """
SELECT c.Country, c.Year, c.Gender, c.Cholesterol_Value, m.Mortality_Value, m.Disease FROM cholesterol_table c
    JOIN mortality_table m ON c.Year = m.Year AND c.Gender = m.Gender and c.Country=m.Country
    WHERE c.Country = 'Canada' AND c.Year BETWEEN 2000 AND 2018
    AND c.Gender= 'Both sexes' AND m.Disease = 'Cardiovascular diseases';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Cholesterol_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	4.6	70147.0	Cardiovascular diseases
1	Canada	2017	Both sexes	4.7	71372.0	Cardiovascular diseases
2	Canada	2016	Both sexes	4.7	69553.0	Cardiovascular diseases
3	Canada	2015	Both sexes	4.7	69548.0	Cardiovascular diseases
4	Canada	2014	Both sexes	4.7	68507.0	Cardiovascular diseases
5	Canada	2013	Both sexes	4.8	67296.0	Cardiovascular diseases
6	Canada	2012	Both sexes	4.8	65860.0	Cardiovascular diseases
7	Canada	2011	Both sexes	4.8	65294.0	Cardiovascular diseases
8	Canada	2010	Both sexes	4.9	66670.0	Cardiovascular diseases
9	Canada	2009	Both sexes	4.9	67611.0	Cardiovascular diseases
10	Canada	2008	Both sexes	4.9	68827.0	Cardiovascular diseases
11	Canada	2007	Both sexes	5.0	68833.0	Cardiovascular diseases
12	Canada	2006	Both sexes	5.0	68097.0	Cardiovascular diseases
13	Canada	2005	Both sexes	5.0	70622.0	Cardiovascular diseases
14	Canada	2004	Both sexes	5.0	71670.0	Cardiovascular diseases
15	Canada	2003	Both sexes	5.1	73397.0	Cardiovascular diseases
16	Canada	2002	Both sexes	5.1	73741.0	Cardiovascular diseases
17	Canada	2001	Both sexes	5.1	73945.0	Cardiovascular diseases
18	Canada	2000	Both sexes	5.2	75398.0	Cardiovascular diseases

Appendix 1.2.D: Trend between Cholesterol and Respiratory Disease Mortality Values using SQL.

```
#Looking at trend between respiratory disease mortality and cholesterol

query = """
SELECT c.Country, c.Year, c.Gender, c.Cholesterol_Value, m.Mortality_Value, m.Disease FROM cholesterol_table c
    JOIN mortality_table m ON c.Year = m.Year AND c.Gender = m.Gender and c.Country=m.Country
    WHERE c.Country = 'Canada' AND c.Year BETWEEN 2000 AND 2018
    AND c.Gender= 'Both sexes' AND m.Disease = 'Respiratory diseases';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Cholesterol_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	4.6	19700.0	Respiratory diseases
1	Canada	2017	Both sexes	4.7	19973.0	Respiratory diseases
2	Canada	2016	Both sexes	4.7	18990.0	Respiratory diseases
3	Canada	2015	Both sexes	4.7	18873.0	Respiratory diseases
4	Canada	2014	Both sexes	4.7	17989.0	Respiratory diseases
5	Canada	2013	Both sexes	4.8	17848.0	Respiratory diseases
6	Canada	2012	Both sexes	4.8	17670.0	Respiratory diseases
7	Canada	2011	Both sexes	4.8	17578.0	Respiratory diseases
8	Canada	2010	Both sexes	4.9	16903.0	Respiratory diseases
9	Canada	2009	Both sexes	4.9	16665.0	Respiratory diseases
10	Canada	2008	Both sexes	4.9	16621.0	Respiratory diseases
11	Canada	2007	Both sexes	5.0	16335.0	Respiratory diseases
12	Canada	2006	Both sexes	5.0	15287.0	Respiratory diseases
13	Canada	2005	Both sexes	5.0	16040.0	Respiratory diseases
14	Canada	2004	Both sexes	5.0	15307.0	Respiratory diseases
15	Canada	2003	Both sexes	5.1	14938.0	Respiratory diseases
16	Canada	2002	Both sexes	5.1	14502.0	Respiratory diseases
17	Canada	2001	Both sexes	5.1	14278.0	Respiratory diseases
18	Canada	2000	Both sexes	5.2	14291.0	Respiratory diseases

Appendix 1.2.E: Trend between Cholesterol and Diabetes mellitus Mortality Values using SQL.

```
#Looking at trend between diabetes mortality and cholesterol

query = """
SELECT c.Country, c.Year, c.Gender, c.Cholesterol_Value, m.Mortality_Value, m.Disease FROM cholesterol_table c
    JOIN mortality_table m ON c.Year = m.Year AND c.Gender = m.Gender and c.Country=m.Country
    WHERE c.Country = 'Canada' AND c.Year BETWEEN 2000 AND 2018
    AND c.Gender= 'Both sexes' AND m.Disease = 'Diabetes mellitus';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Cholesterol_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	4.6	6883.0	Diabetes mellitus
1	Canada	2017	Both sexes	4.7	7002.0	Diabetes mellitus
2	Canada	2016	Both sexes	4.7	6962.0	Diabetes mellitus
3	Canada	2015	Both sexes	4.7	7230.0	Diabetes mellitus
4	Canada	2014	Both sexes	4.7	7091.0	Diabetes mellitus
5	Canada	2013	Both sexes	4.8	7056.0	Diabetes mellitus
6	Canada	2012	Both sexes	4.8	7003.0	Diabetes mellitus
7	Canada	2011	Both sexes	4.8	7232.0	Diabetes mellitus
8	Canada	2010	Both sexes	4.9	6964.0	Diabetes mellitus
9	Canada	2009	Both sexes	4.9	6938.0	Diabetes mellitus
10	Canada	2008	Both sexes	4.9	7542.0	Diabetes mellitus
11	Canada	2007	Both sexes	5.0	7423.0	Diabetes mellitus
12	Canada	2006	Both sexes	5.0	7311.0	Diabetes mellitus
13	Canada	2005	Both sexes	5.0	7916.0	Diabetes mellitus
14	Canada	2004	Both sexes	5.0	7864.0	Diabetes mellitus
15	Canada	2003	Both sexes	5.1	8028.0	Diabetes mellitus
16	Canada	2002	Both sexes	5.1	7929.0	Diabetes mellitus
17	Canada	2001	Both sexes	5.1	7154.0	Diabetes mellitus
18	Canada	2000	Both sexes	5.2	6759.0	Diabetes mellitus

Appendix 1.2.F: Trend between Cholesterol and Cancer Mortality Values using SQL.

```
#Looking at trend between cancer mortality and cholesterol

query = """
SELECT c.Country, c.Year, c.Gender, c.Cholesterol_Value, m.Mortality_Value, m.Disease FROM cholesterol_table c
    JOIN mortality_table m ON c.Year = m.Year AND c.Gender = m.Gender and c.Country=m.Country
    WHERE c.Country = 'Canada' AND c.Year BETWEEN 2000 AND 2018
    AND c.Gender= 'Both sexes' AND m.Disease = 'Malignant neoplasms';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Cholesterol_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	4.6	80881.0	Malignant neoplasms
1	Canada	2017	Both sexes	4.7	80994.0	Malignant neoplasms
2	Canada	2016	Both sexes	4.7	80696.0	Malignant neoplasms
3	Canada	2015	Both sexes	4.7	78020.0	Malignant neoplasms
4	Canada	2014	Both sexes	4.7	77763.0	Malignant neoplasms
5	Canada	2013	Both sexes	4.8	75755.0	Malignant neoplasms
6	Canada	2012	Both sexes	4.8	74983.0	Malignant neoplasms
7	Canada	2011	Both sexes	4.8	73328.0	Malignant neoplasms
8	Canada	2010	Both sexes	4.9	72535.0	Malignant neoplasms
9	Canada	2009	Both sexes	4.9	71726.0	Malignant neoplasms
10	Canada	2008	Both sexes	4.9	71197.0	Malignant neoplasms
11	Canada	2007	Both sexes	5.0	70257.0	Malignant neoplasms
12	Canada	2006	Both sexes	5.0	68729.0	Malignant neoplasms
13	Canada	2005	Both sexes	5.0	68005.0	Malignant neoplasms
14	Canada	2004	Both sexes	5.0	67746.0	Malignant neoplasms
15	Canada	2003	Both sexes	5.1	66883.0	Malignant neoplasms
16	Canada	2002	Both sexes	5.1	65980.0	Malignant neoplasms
17	Canada	2001	Both sexes	5.1	64668.0	Malignant neoplasms
18	Canada	2000	Both sexes	5.2	63429.0	Malignant neoplasms

Appendix 1.2.G: SQL query and Python code for Cholesterol Facet Grid Plot.

```
#Correlation between cholesterol and mortality diseases
query_corr_chol = """SELECT c.Country, c.Year, c.Gender,  c.Cholesterol_Value, m.Mortality_Value, m.Disease FROM cholesterol_table c
    JOIN mortality_table m ON c.Year = m.Year AND c.Gender = m.Gender and c.Country=m.Country
    WHERE c.Country = 'Canada' AND c.Year BETWEEN 2000 AND 2018 order by m.Disease, m.Year;"""

result=pd.read_sql_query(query_corr_chol, engine)
pivot_df = result.pivot_table(index='Year', columns='Disease', values='Mortality_Value', aggfunc='first')
pivot_df = pivot_df.reset_index()
result = pd.merge(result, pivot_df, on='Year', how='left')

g = sns.FacetGrid(result, col="Disease", hue="Gender", col_wrap=2)
g.map_dataframe(sns.scatterplot, x="Cholesterol_Value", y="Mortality_Value")
g.add_legend()
plt.savefig("604 Project/cholesterol_mortality_pairplot.png", format='png', dpi=300)
plt.show()
```

Hypertension

Appendix 1.3.A: SQL query for determining the trend of Cholesterol in Canada from 2000 to 2018.

```
#Hypertension trend in Canada from 2000 to 2018

query = """ SELECT Country, Year, Gender, Hypertension_Value, Hypertension_Low_Value, Hypertension_High_Value
FROM hypertension_table
  WHERE Year >= '2000'
  AND Year <= '2018'
  AND Country = 'Canada'
  AND Gender != 'Both sexes';"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Hypertension_Value	Hypertension_Low_Value	Hypertension_High_Value
0	Canada	2018	Female	20.2	16.8	24.1
1	Canada	2018	Male	24.5	20.6	28.6
2	Canada	2017	Female	20.5	17.4	24.0
3	Canada	2017	Male	24.8	21.2	28.4
4	Canada	2016	Female	20.8	17.9	24.0
5	Canada	2016	Male	25.0	21.6	28.4
6	Canada	2015	Female	21.0	18.3	24.0
7	Canada	2015	Male	25.3	22.0	28.6
8	Canada	2014	Female	21.3	18.6	24.2
9	Canada	2014	Male	25.6	22.4	28.9
10	Canada	2013	Female	21.6	18.9	24.5
11	Canada	2013	Male	25.9	22.8	29.2
12	Canada	2012	Female	21.9	19.0	24.9
13	Canada	2012	Male	26.2	23.0	29.6
14	Canada	2011	Female	22.2	19.2	25.3
15	Canada	2011	Male	26.5	23.1	30.1
16	Canada	2010	Female	22.5	19.3	25.7
17	Canada	2010	Male	26.9	23.3	30.6
18	Canada	2009	Female	22.7	19.4	26.2
19	Canada	2009	Male	27.2	23.4	31.1
20	Canada	2008	Female	23.0	19.6	26.6
21	Canada	2008	Male	27.5	23.5	31.7
22	Canada	2007	Female	23.3	19.6	27.2
23	Canada	2007	Male	27.9	23.7	32.2
24	Canada	2006	Female	23.6	19.7	27.6
25	Canada	2006	Male	28.2	23.8	32.8
26	Canada	2005	Female	23.9	19.9	28.1
27	Canada	2005	Male	28.6	23.9	33.5
28	Canada	2004	Female	24.2	20.0	28.6
29	Canada	2004	Male	29.0	24.1	34.1
30	Canada	2003	Female	24.6	20.2	29.1
31	Canada	2003	Male	29.4	24.4	34.8
32	Canada	2002	Female	24.9	20.4	29.7
33	Canada	2002	Male	29.8	24.6	35.4
34	Canada	2001	Female	25.3	20.8	30.1
35	Canada	2001	Male	30.3	24.9	36.0
36	Canada	2000	Female	25.6	21.1	30.5
37	Canada	2000	Male	30.8	25.4	36.7

Appendix 1.3.B: SQL queries for ranking data in the year 2000, 2018 and the 19-year average ranking for Hypertension.

[261]:
```
#Canada hypertension rank compared to the rest of the world in 2000
query = """ WITH Ranked_Hypertension AS (SELECT Country, Hypertension_Value, RANK() OVER (ORDER BY Hypertension_Value DESC) AS Rank_Hypertensi
    FROM hypertension_table WHERE Gender = 'Both sexes' AND Year = 2000) SELECT Country, Hypertension_Value, Rank_Hypertension
    FROM Ranked_Hypertension WHERE Country = 'Canada';"""

ranked_data = pd.read_sql_query(query, engine)
print("Canada rank for 2000 out of 195 countries")
ranked_data.head()
```

```
Canada rank for 2000 out of 195 countries
```

[261]:

	Country	Hypertension_Value	Rank_Hypertension
0	Canada	28.2	180

•[233…
```
#Canada hypertension rank compared to the rest of the world in 2018

query = """ WITH Ranked_Hypertension AS (SELECT Country, Hypertension_Value, RANK() OVER (ORDER BY Hypertension_Value DESC) AS Rank_Hypertensi
    FROM hypertension_table WHERE Gender = 'Both sexes' AND Year = 2018) SELECT Country, Hypertension_Value, Rank_Hypertension
    FROM Ranked_Hypertension WHERE Country = 'Canada';"""
ranked_data = pd.read_sql_query(query, engine)
print("Canada rank for 2018 out of 195 countries")
ranked_data.head()
```

```
Canada rank for 2018 out of 195 countries
```

[233]:

	Country	Hypertension_Value	Rank_Hypertension
0	Canada	22.4	192

[283]:
```
#Ranking countries based on hypertension (from highest percentage to lowest percentage).

query = """WITH Ranked_Hypertension AS (SELECT Country, AVG(Hypertension_Value) AS Average_Hypertension,
        RANK() OVER (ORDER BY AVG(Hypertension_Value) DESC) AS Ranked_Hypertension
    FROM hypertension_table
    WHERE Gender = 'Both sexes' AND Year >= 2000 AND Year <= 2018
    GROUP BY Country)
SELECT Country, Average_Hypertension, Ranked_Hypertension
FROM Ranked_Hypertension
WHERE Country = 'Canada';"""
print("AVG Canada rank for Hypertension is out of 195 countries")
ranked_data = pd.read_sql_query(query, engine)
ranked_data.head()
```

```
AVG Canada rank for Hypertension is out of 195 countries
```

[283]:

	Country	Average_Hypertension	Ranked_Hypertension
0	Canada	25.042105	190

Appendix 1.3.C: Trend between Hypertension and Cardiovascular Disease Mortality Values using SQL.

```
#Joining the combined Hypertension data with Mortality regarding Cardiovascular disease to find trend
query = """
SELECT h.Country, h.Year, h.Gender,  h.Hypertension_Value, m.Mortality_Value, m.Disease FROM hypertension_table h
    JOIN mortality_table m ON h.Year = m.Year AND h.Gender = m.Gender and h.Country=m.Country
    WHERE h.Country = 'Canada' AND h.Year BETWEEN 2000 AND 2018
    AND h.Gender= 'Both sexes' AND m.Disease = 'Cardiovascular diseases';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Hypertension_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	22.4	70147.0	Cardiovascular diseases
1	Canada	2017	Both sexes	22.6	71372.0	Cardiovascular diseases
2	Canada	2016	Both sexes	22.9	69553.0	Cardiovascular diseases
3	Canada	2015	Both sexes	23.1	69548.0	Cardiovascular diseases
4	Canada	2014	Both sexes	23.4	68507.0	Cardiovascular diseases
5	Canada	2013	Both sexes	23.7	67296.0	Cardiovascular diseases
6	Canada	2012	Both sexes	24.0	65860.0	Cardiovascular diseases
7	Canada	2011	Both sexes	24.3	65294.0	Cardiovascular diseases
8	Canada	2010	Both sexes	24.6	66670.0	Cardiovascular diseases
9	Canada	2009	Both sexes	24.9	67611.0	Cardiovascular diseases
10	Canada	2008	Both sexes	25.2	68827.0	Cardiovascular diseases
11	Canada	2007	Both sexes	25.6	68833.0	Cardiovascular diseases
12	Canada	2006	Both sexes	25.9	68097.0	Cardiovascular diseases
13	Canada	2005	Both sexes	26.2	70622.0	Cardiovascular diseases
14	Canada	2004	Both sexes	26.6	71670.0	Cardiovascular diseases
15	Canada	2003	Both sexes	27.0	73397.0	Cardiovascular diseases
16	Canada	2002	Both sexes	27.4	73741.0	Cardiovascular diseases
17	Canada	2001	Both sexes	27.8	73945.0	Cardiovascular diseases
18	Canada	2000	Both sexes	28.2	75398.0	Cardiovascular diseases

Appendix 1.3.D: Trend between Hypertension and Respiratory Disease Mortality Values using SQL.

```
#Looking at trend between respiratory disease mortality and hypertension

query = """
SELECT h.Country, h.Year, h.Gender,  h.Hypertension_Value, m.Mortality_Value, m.Disease FROM hypertension_table h
    JOIN mortality_table m ON h.Year = m.Year AND h.Gender = m.Gender and h.Country=m.Country
    WHERE h.Country = 'Canada' AND h.Year BETWEEN 2000 AND 2018
    AND h.Gender= 'Both sexes' AND m.Disease = 'Respiratory diseases';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Hypertension_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	22.4	19700.0	Respiratory diseases
1	Canada	2017	Both sexes	22.6	19973.0	Respiratory diseases
2	Canada	2016	Both sexes	22.9	18990.0	Respiratory diseases
3	Canada	2015	Both sexes	23.1	18873.0	Respiratory diseases
4	Canada	2014	Both sexes	23.4	17989.0	Respiratory diseases
5	Canada	2013	Both sexes	23.7	17848.0	Respiratory diseases
6	Canada	2012	Both sexes	24.0	17670.0	Respiratory diseases
7	Canada	2011	Both sexes	24.3	17578.0	Respiratory diseases
8	Canada	2010	Both sexes	24.6	16903.0	Respiratory diseases
9	Canada	2009	Both sexes	24.9	16665.0	Respiratory diseases
10	Canada	2008	Both sexes	25.2	16621.0	Respiratory diseases
11	Canada	2007	Both sexes	25.6	16335.0	Respiratory diseases
12	Canada	2006	Both sexes	25.9	15287.0	Respiratory diseases
13	Canada	2005	Both sexes	26.2	16040.0	Respiratory diseases
14	Canada	2004	Both sexes	26.6	15307.0	Respiratory diseases
15	Canada	2003	Both sexes	27.0	14938.0	Respiratory diseases
16	Canada	2002	Both sexes	27.4	14502.0	Respiratory diseases
17	Canada	2001	Both sexes	27.8	14278.0	Respiratory diseases
18	Canada	2000	Both sexes	28.2	14291.0	Respiratory diseases

Appendix 1.3.E: Trend between Hypertension and Diabetes mellitus Mortality Values using SQL.

```
#Looking at trend between diabetes mortality and hypertension

query = """
SELECT h.Country, h.Year, h.Gender,  h.Hypertension_Value, m.Mortality_Value, m.Disease FROM hypertension_table h
    JOIN mortality_table m ON h.Year = m.Year AND h.Gender = m.Gender and h.Country=m.Country
    WHERE h.Country = 'Canada' AND h.Year BETWEEN 2000 AND 2018
    AND h.Gender= 'Both sexes' AND m.Disease = 'Diabetes mellitus';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Hypertension_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	22.4	6883.0	Diabetes mellitus
1	Canada	2017	Both sexes	22.6	7002.0	Diabetes mellitus
2	Canada	2016	Both sexes	22.9	6962.0	Diabetes mellitus
3	Canada	2015	Both sexes	23.1	7230.0	Diabetes mellitus
4	Canada	2014	Both sexes	23.4	7091.0	Diabetes mellitus
5	Canada	2013	Both sexes	23.7	7056.0	Diabetes mellitus
6	Canada	2012	Both sexes	24.0	7003.0	Diabetes mellitus
7	Canada	2011	Both sexes	24.3	7232.0	Diabetes mellitus
8	Canada	2010	Both sexes	24.6	6964.0	Diabetes mellitus
9	Canada	2009	Both sexes	24.9	6938.0	Diabetes mellitus
10	Canada	2008	Both sexes	25.2	7542.0	Diabetes mellitus
11	Canada	2007	Both sexes	25.6	7423.0	Diabetes mellitus
12	Canada	2006	Both sexes	25.9	7311.0	Diabetes mellitus
13	Canada	2005	Both sexes	26.2	7916.0	Diabetes mellitus
14	Canada	2004	Both sexes	26.6	7864.0	Diabetes mellitus
15	Canada	2003	Both sexes	27.0	8028.0	Diabetes mellitus
16	Canada	2002	Both sexes	27.4	7929.0	Diabetes mellitus
17	Canada	2001	Both sexes	27.8	7154.0	Diabetes mellitus
18	Canada	2000	Both sexes	28.2	6759.0	Diabetes mellitus

Appendix 1.3.F: Trend between Hypertension and Cancer Mortality Values using SQL.

```
#Looking at trend between cancer mortality and hypertension

query = """
SELECT h.Country, h.Year, h.Gender,  h.Hypertension_Value, m.Mortality_Value, m.Disease FROM hypertension_table h
    JOIN mortality_table m ON h.Year = m.Year AND h.Gender = m.Gender and h.Country=m.Country
    WHERE h.Country = 'Canada' AND h.Year BETWEEN 2000 AND 2018
    AND h.Gender= 'Both sexes' AND m.Disease = 'Malignant neoplasms';
"""
pd.read_sql_query(query, engine)
```

	Country	Year	Gender	Hypertension_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	22.4	80881.0	Malignant neoplasms
1	Canada	2017	Both sexes	22.6	80994.0	Malignant neoplasms
2	Canada	2016	Both sexes	22.9	80696.0	Malignant neoplasms
3	Canada	2015	Both sexes	23.1	78020.0	Malignant neoplasms
4	Canada	2014	Both sexes	23.4	77763.0	Malignant neoplasms
5	Canada	2013	Both sexes	23.7	75755.0	Malignant neoplasms
6	Canada	2012	Both sexes	24.0	74983.0	Malignant neoplasms
7	Canada	2011	Both sexes	24.3	73328.0	Malignant neoplasms
8	Canada	2010	Both sexes	24.6	72535.0	Malignant neoplasms
9	Canada	2009	Both sexes	24.9	71726.0	Malignant neoplasms
10	Canada	2008	Both sexes	25.2	71197.0	Malignant neoplasms
11	Canada	2007	Both sexes	25.6	70257.0	Malignant neoplasms
12	Canada	2006	Both sexes	25.9	68729.0	Malignant neoplasms
13	Canada	2005	Both sexes	26.2	68005.0	Malignant neoplasms
14	Canada	2004	Both sexes	26.6	67746.0	Malignant neoplasms
15	Canada	2003	Both sexes	27.0	66883.0	Malignant neoplasms
16	Canada	2002	Both sexes	27.4	65980.0	Malignant neoplasms
17	Canada	2001	Both sexes	27.8	64668.0	Malignant neoplasms
18	Canada	2000	Both sexes	28.2	63429.0	Malignant neoplasms

Appendix 1.3.G: SQL query and Python code for Hypertension Facet Grid Plot.

```
# Correlation between hypertension and mortality diseases
query_corr_hyp = """SELECT h.Country, h.Year, h.Gender,  h.Hypertension_Value, m.Mortality_Value, m.Disease FROM hypertension_table h
    JOIN mortality_table m ON h.Year = m.Year AND h.Gender = m.Gender and h.Country=m.Country
    WHERE h.Country = 'Canada' AND h.Year BETWEEN 2000 AND 2018 order by m.Disease, m.Year;"""

result=pd.read_sql_query(query_corr_hyp, engine)
pivot_df = result.pivot_table(index='Year', columns='Disease', values='Mortality_Value', aggfunc='first')
pivot_df = pivot_df.reset_index()
result = pd.merge(result, pivot_df, on='Year', how='left')

g = sns.FacetGrid(result, col="Disease", hue="Gender", col_wrap=2)
g.map_dataframe(sns.scatterplot, x="Hypertension_Value", y="Mortality_Value")
g.add_legend()
plt.savefig("604 Project/hypertension_mortality_pairplot.png", format='png', dpi=300)
plt.show()
```

Obesity

Appendix 1.4.A: SQL query for determining the trend of Obesity in Canada from 2000 to 2018.

```
#Create a new table for obesity values in Canada from 2000-2018 for both sexes
query1 = """SELECT Country, Year, Gender, Obesity_Value, Obesity_Value_Low, Obesity_Value_High FROM obesity_table
WHERE Country = 'Canada'
  AND Year >= '2000'
  AND Year <= '2018'
  AND Gender!='Both sexes';"""

pd.read_sql_query(query1, engine)

canada_obesity=pd.read_sql_query(query1, engine)

canada_obesity.to_sql("canada_obesity", engine, index=False, if_exists='replace')
canada_obesity=pd.read_sql_table("canada_obesity", engine)
canada_obesity
```

	Country	Year	Gender	Obesity_Value	Obesity_Value_Low	Obesity_Value_High
0	Canada	2018	Female	24.53	22.38	26.79
1	Canada	2018	Male	26.63	24.26	29.13
2	Canada	2017	Female	24.45	22.52	26.47
3	Canada	2017	Male	26.41	24.30	28.67
4	Canada	2016	Female	24.35	22.58	26.14
5	Canada	2016	Male	26.20	24.28	28.24
6	Canada	2015	Female	24.25	22.57	25.92
7	Canada	2015	Male	25.99	24.26	27.82
8	Canada	2014	Female	24.13	22.55	25.69
9	Canada	2014	Male	25.80	24.16	27.49
10	Canada	2013	Female	24.00	22.49	25.49
11	Canada	2013	Male	25.62	24.08	27.23
12	Canada	2012	Female	23.85	22.39	25.32
13	Canada	2012	Male	25.45	23.90	27.04
14	Canada	2011	Female	23.68	22.20	25.18
15	Canada	2011	Male	25.26	23.72	26.84
16	Canada	2010	Female	23.49	22.03	24.94
17	Canada	2010	Male	25.05	23.53	26.60
18	Canada	2009	Female	23.26	21.82	24.74
19	Canada	2009	Male	24.78	23.29	26.34
20	Canada	2008	Female	23.00	21.52	24.49
21	Canada	2008	Male	24.47	22.98	26.02
22	Canada	2007	Female	22.71	21.24	24.22
23	Canada	2007	Male	24.10	22.63	25.66
24	Canada	2006	Female	22.40	20.89	23.96
25	Canada	2006	Male	23.69	22.17	25.29
26	Canada	2005	Female	22.07	20.53	23.65
27	Canada	2005	Male	23.23	21.69	24.87
28	Canada	2004	Female	21.70	20.15	23.35
29	Canada	2004	Male	22.74	21.17	24.43
30	Canada	2003	Female	21.31	19.72	22.99
31	Canada	2003	Male	22.21	20.60	23.95
32	Canada	2002	Female	20.88	19.25	22.58
33	Canada	2002	Male	21.64	19.99	23.42
34	Canada	2001	Female	20.42	18.78	22.12
35	Canada	2001	Male	21.05	19.35	22.86
36	Canada	2000	Female	19.92	18.28	21.63
37	Canada	2000	Male	20.42	18.71	22.23

Appendix 1.24.B: SQL queries for ranking data in the year 2000, 2018 and the 19 years average ranking.

```
query = """ WITH Ranked_Obesity AS ( SELECT Country, Obesity_Value, RANK() OVER (ORDER BY Obesity_Value DESC) AS Rank_Obesity
    FROM obesity_table WHERE Gender = 'Both sexes' AND Year = 2000) SELECT Country,  Obesity_Value, Rank_Obesity
    FROM Ranked_Obesity WHERE Country = 'Canada';"""
ranked_data = pd.read_sql_query(query, engine)
print("Canada rank for 2000 out of 198 countries")
ranked_data.head()
```

```
Canada rank for 2000 out of 198 countries
```

	Country	Obesity_Value	Rank_Obesity
0	Canada	20.19	39

```
query = """ WITH Ranked_Obesity AS ( SELECT Country, Obesity_Value, RANK() OVER (ORDER BY Obesity_Value DESC) AS Rank_Obesity
    FROM obesity_table WHERE Gender = 'Both sexes' AND Year = 2018) SELECT Country, Obesity_Value, Rank_Obesity
    FROM Ranked_Obesity WHERE Country = 'Canada';"""
ranked_data = pd.read_sql_query(query, engine)
print("Canada rank for 2018 out of 198 countries")
ranked_data.head()
```

```
Canada rank for 2018 out of 198 countries
```

	Country	Obesity_Value	Rank_Obesity
0	Canada	25.57	70

```
query = """WITH Ranked_Obesity AS (SELECT Country, AVG(Obesity_Value) AS avg_obesity,
        RANK() OVER (ORDER BY AVG(Obesity_Value) DESC) AS Rank_Obesity
    FROM obesity_table
    WHERE Gender = 'Both sexes' AND Year >= 2000 AND Year <= 2018
    GROUP BY Country)
SELECT Country, avg_obesity, Rank_Obesity
FROM Ranked_Obesity
WHERE Country = 'Canada';"""
print("AVG Canada rank for Obesity us out of 198 countries")
ranked_data = pd.read_sql_query(query, engine)
ranked_data.head()
```

```
AVG Canada rank for Obesity us out of 198 countries
```

	Country	avg_obesity	Rank_Obesity
0	Canada	23.564211	52

Appendix 1.4.C: Trend between Obesity and Cardiovascular Disease Mortality Values using SQL.

```
#Query: Join Obesity values with Cardiovascular Disease Mortality Values
query = """SELECT
    o.Region,
    o.Country,
    o.Year,
    o.Gender,
    o.Obesity_Value,
    m.Mortality_Value,
    m.Disease
FROM obesity_table o
JOIN mortality_table m
ON o.Year = m.Year
   AND o.Gender = m.Gender
   AND o.Country = m.Country
WHERE o.Country = 'Canada'
  AND o.Year BETWEEN 2000 AND 2018
  AND o.Gender = 'Both sexes'
  AND m.Disease = 'Cardiovascular diseases';"""

merged_data = pd.read_sql_query(query, engine)
merged_data
```

	Region	Country	Year	Gender	Obesity_Value	Mortality_Value	Disease
0	Americas	Canada	2018	Both sexes	25.57	70147.0	Cardiovascular diseases
1	Americas	Canada	2017	Both sexes	25.43	71372.0	Cardiovascular diseases
2	Americas	Canada	2016	Both sexes	25.28	69553.0	Cardiovascular diseases
3	Americas	Canada	2015	Both sexes	25.12	69548.0	Cardiovascular diseases
4	Americas	Canada	2014	Both sexes	24.97	68507.0	Cardiovascular diseases
5	Americas	Canada	2013	Both sexes	24.81	67296.0	Cardiovascular diseases
6	Americas	Canada	2012	Both sexes	24.65	65860.0	Cardiovascular diseases
7	Americas	Canada	2011	Both sexes	24.48	65294.0	Cardiovascular diseases
8	Americas	Canada	2010	Both sexes	24.27	66670.0	Cardiovascular diseases
9	Americas	Canada	2009	Both sexes	24.03	67611.0	Cardiovascular diseases
10	Americas	Canada	2008	Both sexes	23.74	68827.0	Cardiovascular diseases
11	Americas	Canada	2007	Both sexes	23.42	68833.0	Cardiovascular diseases
12	Americas	Canada	2006	Both sexes	23.06	68097.0	Cardiovascular diseases
13	Americas	Canada	2005	Both sexes	22.66	70622.0	Cardiovascular diseases
14	Americas	Canada	2004	Both sexes	22.24	71670.0	Cardiovascular diseases
15	Americas	Canada	2003	Both sexes	21.77	73397.0	Cardiovascular diseases
16	Americas	Canada	2002	Both sexes	21.28	73741.0	Cardiovascular diseases
17	Americas	Canada	2001	Both sexes	20.75	73945.0	Cardiovascular diseases
18	Americas	Canada	2000	Both sexes	20.19	75398.0	Cardiovascular diseases

Appendix 1.4.D: Trend between Obesity and Malignant Neoplasms Mortality Values using SQL.

```
#Query: Join Obesity values with Malignant Neoplasms Mortaliity Values
query= """SELECT
    o.Region,
    o.Country,
    o.Year,
    o.Gender,
    o.Obesity_Value,
    m.Mortality_Value,
    m.Disease
FROM obesity_table o
JOIN mortality_table m
ON o.Year = m.Year
   AND o.Gender = m.Gender
   AND o.Country = m.Country
WHERE o.Country = 'Canada'
  AND o.Year BETWEEN 2000 AND 2018
  AND o.Gender = 'Both sexes'
  AND m.Disease = 'Malignant Neoplasms';"""

merged_data = pd.read_sql_query(query, engine)
merged_data
```

	Region	Country	Year	Gender	Obesity_Value	Mortality_Value	Disease
0	Americas	Canada	2018	Both sexes	25.57	80881.0	Malignant neoplasms
1	Americas	Canada	2017	Both sexes	25.43	80994.0	Malignant neoplasms
2	Americas	Canada	2016	Both sexes	25.28	80696.0	Malignant neoplasms
3	Americas	Canada	2015	Both sexes	25.12	78020.0	Malignant neoplasms
4	Americas	Canada	2014	Both sexes	24.97	77763.0	Malignant neoplasms
5	Americas	Canada	2013	Both sexes	24.81	75755.0	Malignant neoplasms
6	Americas	Canada	2012	Both sexes	24.65	74983.0	Malignant neoplasms
7	Americas	Canada	2011	Both sexes	24.48	73328.0	Malignant neoplasms
8	Americas	Canada	2010	Both sexes	24.27	72535.0	Malignant neoplasms
9	Americas	Canada	2009	Both sexes	24.03	71726.0	Malignant neoplasms
10	Americas	Canada	2008	Both sexes	23.74	71197.0	Malignant neoplasms
11	Americas	Canada	2007	Both sexes	23.42	70257.0	Malignant neoplasms
12	Americas	Canada	2006	Both sexes	23.06	68729.0	Malignant neoplasms
13	Americas	Canada	2005	Both sexes	22.66	68005.0	Malignant neoplasms
14	Americas	Canada	2004	Both sexes	22.24	67746.0	Malignant neoplasms
15	Americas	Canada	2003	Both sexes	21.77	66883.0	Malignant neoplasms
16	Americas	Canada	2002	Both sexes	21.28	65980.0	Malignant neoplasms
17	Americas	Canada	2001	Both sexes	20.75	64668.0	Malignant neoplasms
18	Americas	Canada	2000	Both sexes	20.19	63429.0	Malignant neoplasms

Appendix 1.4.E: Trend between Obesity and Diabetes Mellitus Mortality Values using SQL.

```
#Query: Join Obesity values with Diabetes Mellitus Mortality Values
query= """SELECT
    o.Region,
    o.Country,
    o.Year,
    o.Gender,
    o.Obesity_Value,
    m.Mortality_Value,
    m.Disease
FROM obesity_table o
JOIN mortality_table m
ON o.Year = m.Year
   AND o.Gender = m.Gender
   AND o.Country = m.Country
WHERE o.Country = 'Canada'
  AND o.Year BETWEEN 2000 AND 2018
  AND o.Gender = 'Both sexes'
  AND m.Disease = 'Diabetes mellitus';"""

merged_data = pd.read_sql_query(query, engine)
merged_data
```

	Region	Country	Year	Gender	Obesity_Value	Mortality_Value	Disease
0	Americas	Canada	2018	Both sexes	25.57	6883.0	Diabetes mellitus
1	Americas	Canada	2017	Both sexes	25.43	7002.0	Diabetes mellitus
2	Americas	Canada	2016	Both sexes	25.28	6962.0	Diabetes mellitus
3	Americas	Canada	2015	Both sexes	25.12	7230.0	Diabetes mellitus
4	Americas	Canada	2014	Both sexes	24.97	7091.0	Diabetes mellitus
5	Americas	Canada	2013	Both sexes	24.81	7056.0	Diabetes mellitus
6	Americas	Canada	2012	Both sexes	24.65	7003.0	Diabetes mellitus
7	Americas	Canada	2011	Both sexes	24.48	7232.0	Diabetes mellitus
8	Americas	Canada	2010	Both sexes	24.27	6964.0	Diabetes mellitus
9	Americas	Canada	2009	Both sexes	24.03	6938.0	Diabetes mellitus
10	Americas	Canada	2008	Both sexes	23.74	7542.0	Diabetes mellitus
11	Americas	Canada	2007	Both sexes	23.42	7423.0	Diabetes mellitus
12	Americas	Canada	2006	Both sexes	23.06	7311.0	Diabetes mellitus
13	Americas	Canada	2005	Both sexes	22.66	7916.0	Diabetes mellitus
14	Americas	Canada	2004	Both sexes	22.24	7864.0	Diabetes mellitus
15	Americas	Canada	2003	Both sexes	21.77	8028.0	Diabetes mellitus
16	Americas	Canada	2002	Both sexes	21.28	7929.0	Diabetes mellitus
17	Americas	Canada	2001	Both sexes	20.75	7154.0	Diabetes mellitus
18	Americas	Canada	2000	Both sexes	20.19	6759.0	Diabetes mellitus

Appendix 1.4.F: Trend between Obesity and Respiratory Disease Mortality Values using SQL.

```
#Query: Join Obesity values with Respiratory Disease Mortality Values
query= """SELECT
    o.Country,
    o.Year,
    o.Gender,
    o.Obesity_Value,
    m.Mortality_Value,
    m.Disease
FROM obesity_table o
JOIN mortality_table m
ON o.Year = m.Year
   AND o.Gender = m.Gender
   AND o.Country = m.Country
WHERE o.Country = 'Canada'
  AND o.Year BETWEEN 2000 AND 2018
  AND o.Gender = 'Both sexes'
  AND m.Disease = 'Respiratory Diseases';"""

merged_data = pd.read_sql_query(query, engine)
merged_data
```

	Country	Year	Gender	Obesity_Value	Mortality_Value	Disease
0	Canada	2018	Both sexes	25.57	19700.0	Respiratory diseases
1	Canada	2017	Both sexes	25.43	19973.0	Respiratory diseases
2	Canada	2016	Both sexes	25.28	18990.0	Respiratory diseases
3	Canada	2015	Both sexes	25.12	18873.0	Respiratory diseases
4	Canada	2014	Both sexes	24.97	17989.0	Respiratory diseases
5	Canada	2013	Both sexes	24.81	17848.0	Respiratory diseases
6	Canada	2012	Both sexes	24.65	17670.0	Respiratory diseases
7	Canada	2011	Both sexes	24.48	17578.0	Respiratory diseases
8	Canada	2010	Both sexes	24.27	16903.0	Respiratory diseases
9	Canada	2009	Both sexes	24.03	16665.0	Respiratory diseases
10	Canada	2008	Both sexes	23.74	16621.0	Respiratory diseases
11	Canada	2007	Both sexes	23.42	16335.0	Respiratory diseases
12	Canada	2006	Both sexes	23.06	15287.0	Respiratory diseases
13	Canada	2005	Both sexes	22.66	16040.0	Respiratory diseases
14	Canada	2004	Both sexes	22.24	15307.0	Respiratory diseases
15	Canada	2003	Both sexes	21.77	14938.0	Respiratory diseases
16	Canada	2002	Both sexes	21.28	14502.0	Respiratory diseases
17	Canada	2001	Both sexes	20.75	14278.0	Respiratory diseases
18	Canada	2000	Both sexes	20.19	14291.0	Respiratory diseases

Appendix 1.4.G: SQL query and code for the Facet grid to observe the relationship between Obesity and each NCD Mortality Value.

```
# Facet Grid Query to observe the relationship between obesity and each NCD Mortality Value
query = """SELECT o.Country, o.Year, o.Gender,  o.Obesity_Value, m.Mortality_Value, m.Disease FROM obesity_table o
    JOIN mortality_table m ON o.Year = m.Year AND o.Gender = m.Gender and o.country=m.country
    WHERE o.Country = 'Canada' AND o.Year BETWEEN 2000 AND 2018 order by m.Disease, m.Year;"""

pivot_df = result.pivot_table(index='Year', columns='Disease', values='Mortality_Value', aggfunc='first')
pivot_df = pivot_df.reset_index()
result = pd.merge(result, pivot_df, on='Year', how='left')

g = sns.FacetGrid(result, col="Disease", hue="Gender", col_wrap=2)
g.map_dataframe(sns.scatterplot, x="Obesity_Value", y="Mortality_Value")
g.add_legend()
plt.show()
```

Physical Inactivity:

Appendix 1.5.A: SQL query for determining the trend of Physical inactivity in Canada from 2000 to 2018.

```
# Over all trend of physical inactivity for males and females in recenty years in Canada
query4 = """SELECT Country,Year,Gender,Physical_Inactivity_Value, Physical_Inactivity_Value_Low, Physical_Inactivity_Value_High
            FROM physicalinactivity
            where year<=2018 and Country='Canada' and gender!='Both sexes' order by year DESC;
        """
result=pd.read_sql_query(query4, engine)
result
```

✓ 0.1s

	Country	Year	Gender	Physical_Inactivity_Value	Physical_Inactivity_Value_Low	Physical_Inactivity_Value_High
0	Canada	2018	Male	33.44	25.78	41.39
1	Canada	2018	Female	36.69	27.98	45.56
2	Canada	2017	Male	32.89	25.59	40.42
3	Canada	2017	Female	36.23	27.90	44.71
4	Canada	2016	Male	32.33	25.22	39.73
5	Canada	2016	Female	35.76	27.81	43.94
6	Canada	2015	Male	31.79	24.95	39.09
7	Canada	2015	Female	35.29	27.52	43.30
8	Canada	2014	Male	31.26	24.48	38.46
9	Canada	2014	Female	34.82	27.18	42.85
10	Canada	2013	Male	30.75	23.93	37.97
11	Canada	2013	Female	34.35	26.65	42.37
12	Canada	2012	Male	30.24	23.29	37.57
13	Canada	2012	Female	33.88	26.00	42.04
14	Canada	2011	Male	29.73	22.62	37.39
15	Canada	2011	Female	33.41	25.42	41.89
16	Canada	2010	Male	29.22	21.89	37.28
17	Canada	2010	Female	32.93	24.74	41.71
18	Canada	2009	Male	28.69	21.04	37.15
19	Canada	2009	Female	32.44	24.00	41.65
20	Canada	2008	Male	28.14	20.16	37.16

Appendix 1.5.B: SQL queries for ranking data in the year 2000, 2018 and the 19 years average ranking.

```
# Rank of canada among all countries for 2000

query1 = """ WITH RankedPhysical AS ( SELECT country, Physical_Inactivity_Value, RANK() OVER (ORDER BY Physical_Inactivity_Value DESC) AS rank_value
    FROM physicalinactivity WHERE Gender = 'Both sexes' AND Year = 2000) SELECT country, Physical_Inactivity_Value, rank_value
    FROM RankedPhysical WHERE country = 'Canada';"""

ranked_data = pd.read_sql_query(query1, engine)

print("Canada rank for 2000 out of 195 countries")

ranked_data.head()
```

✓ 0.1s

```
Canada rank for 2000 out of 195 countries
```

	country	Physical_Inactivity_Value	rank_value
0	Canada	25.65	88

```
# Rank of canada among all countries for 2018

query2 = """ WITH RankedPhysical AS ( SELECT Country, Physical_Inactivity_Value, RANK() OVER (ORDER BY Physical_Inactivity_Value DESC) AS rank_value
    FROM physicalinactivity WHERE Gender = 'Both sexes' AND Year = 2018) SELECT Country, Physical_Inactivity_Value, rank_value
    FROM RankedPhysical WHERE Country = 'Canada';"""

ranked_data = pd.read_sql_query(query2, engine)

print("Canada rank for 2018 out of 195 countries:")

ranked_data.head()
```

✓ 0.1s

```
Canada rank for 2018 out of 195 countries:
```

	Country	Physical_Inactivity_Value	rank_value
0	Canada	35.07	45

```
# Average Physical Inactivity in Canada Rank among all countries from 2000 to 2018

query3 = '''WITH RankedTobacco AS (
    SELECT
        Country,
        AVG(Physical_Inactivity_Value) AS avg_physical_value,
        RANK() OVER (ORDER BY AVG(Physical_Inactivity_Value) DESC) AS rank_value
    FROM physicalinactivity
    WHERE Gender = 'Both sexes' AND Year >= 2000 AND Year <= 2019
    GROUP BY Country)
SELECT Country, avg_physical_value, rank_value
FROM RankedTobacco
WHERE Country = 'Canada';'''

print("AVG Canada rank for Physical Inactivity out of 195 countries:")

ranked_data = pd.read_sql_query(query3, engine)

ranked_data.head()
```

✓ 0.1s

```
AVG Canada rank for Physical Inactivity out of 195 countries:
```

	Country	avg_physical_value	rank_value
0	Canada	30.74	66

Appendix 1.5.C: Trend between Physical Inactivity and Cardiovascular Disease Mortality Values using SQL.

```
#Cardivascular vs Physical Inactivity
query5 = '''SELECT p.Country, p.Year, p.Gender, m.Disease,p.Physical_Inactivity_Value, m.Mortality_Value FROM physicalinactivity p JOIN mortality m ON
        p.Year = m.Year AND p.Gender = m.Gender AND p.Country = m.Location WHERE p.Country = 'Canada' AND m.Disease = 'Cardiovascular Diseases'; '''
cardio_physical = pd.read_sql_query(query5, engine)
cardio_physical
```

✓ 0.1s

	Country	Year	Gender	Disease	Physical_Inactivity_Value	Mortality_Value
0	Canada	2019	Male	Cardiovascular diseases	34.00	36230.0
1	Canada	2019	Both sexes	Cardiovascular diseases	35.58	70202.0
2	Canada	2019	Female	Cardiovascular diseases	37.16	33972.0
3	Canada	2018	Male	Cardiovascular diseases	33.44	36074.0
4	Canada	2018	Both sexes	Cardiovascular diseases	35.07	70147.0
5	Canada	2018	Female	Cardiovascular diseases	36.69	34074.0
6	Canada	2017	Male	Cardiovascular diseases	32.89	36871.0
7	Canada	2017	Both sexes	Cardiovascular diseases	34.56	71372.0
8	Canada	2017	Female	Cardiovascular diseases	36.23	34501.0
9	Canada	2016	Male	Cardiovascular diseases	32.33	35508.0
10	Canada	2016	Both sexes	Cardiovascular diseases	34.05	69553.0
11	Canada	2016	Female	Cardiovascular diseases	35.76	34045.0
12	Canada	2015	Male	Cardiovascular diseases	31.79	34971.0
13	Canada	2015	Both sexes	Cardiovascular diseases	33.55	69548.0
14	Canada	2015	Female	Cardiovascular diseases	35.29	34577.0

Appendix 1.5.D: Trend between Physical Inactivity and Diabetes Mellitus Mortality Values using SQL.

```
query6 = '''SELECT p.Country, p.Year, p.Gender, m.Disease,p.Physical_Inactivity_Value, m.Mortality_Value FROM physicalinactivity p JOIN mortality m ON
        p.Year = m.Year AND p.Gender = m.Gender AND p.Country = m.Location WHERE p.Country = 'Canada' AND m.Disease = 'Diabetes Mellitus';'''

diabetes_physical = pd.read_sql_query(query6, engine)
diabetes_physical
```

✓ 0.1s

	Country	Year	Gender	Disease	Physical_Inactivity_Value	Mortality_Value
0	Canada	2019	Male	Diabetes mellitus	34.00	3891.0
1	Canada	2019	Both sexes	Diabetes mellitus	35.58	6821.0
2	Canada	2019	Female	Diabetes mellitus	37.16	2930.0
3	Canada	2018	Male	Diabetes mellitus	33.44	3889.0
4	Canada	2018	Both sexes	Diabetes mellitus	35.07	6883.0
5	Canada	2018	Female	Diabetes mellitus	36.69	2994.0
6	Canada	2017	Male	Diabetes mellitus	32.89	3970.0
7	Canada	2017	Both sexes	Diabetes mellitus	34.56	7002.0
8	Canada	2017	Female	Diabetes mellitus	36.23	3032.0
9	Canada	2016	Male	Diabetes mellitus	32.33	3769.0
10	Canada	2016	Both sexes	Diabetes mellitus	34.05	6962.0
11	Canada	2016	Female	Diabetes mellitus	35.76	3193.0
12	Canada	2015	Male	Diabetes mellitus	31.79	3980.0
13	Canada	2015	Both sexes	Diabetes mellitus	33.55	7230.0

Appendix 1.5.E: Trend between Physical Inactivity and Malignant Neoplasms (Cancer) Mortality Values using SQL.

```
query7 = '''SELECT p.Country, p.Year, p.Gender, m.Disease,p.Physical_Inactivity_Value, m.Mortality_Value FROM physicalinactivity p JOIN mortality m ON
        p.Year = m.Year AND p.Gender = m.Gender AND p.Country = m.Location WHERE p.Country = 'Canada' AND m.Disease = 'Malignant neoplasms';'''

cancer_physical = pd.read_sql_query(query7, engine)
cancer_physical
```

✓ 0.1s

	Country	Year	Gender	Disease	Physical_Inactivity_Value	Mortality_Value
0	Canada	2019	Male	Malignant neoplasms	34.00	42952.0
1	Canada	2019	Both sexes	Malignant neoplasms	35.58	81565.0
2	Canada	2019	Female	Malignant neoplasms	37.16	38614.0
3	Canada	2018	Male	Malignant neoplasms	33.44	42620.0
4	Canada	2018	Both sexes	Malignant neoplasms	35.07	80881.0
5	Canada	2018	Female	Malignant neoplasms	36.69	38261.0
6	Canada	2017	Male	Malignant neoplasms	32.89	42817.0
7	Canada	2017	Both sexes	Malignant neoplasms	34.56	80994.0
8	Canada	2017	Female	Malignant neoplasms	36.23	38177.0
9	Canada	2016	Male	Malignant neoplasms	32.33	42380.0
10	Canada	2016	Both sexes	Malignant neoplasms	34.05	80696.0
11	Canada	2016	Female	Malignant neoplasms	35.76	38316.0
12	Canada	2015	Male	Malignant neoplasms	31.79	40944.0
13	Canada	2015	Both sexes	Malignant neoplasms	33.55	78020.0
14	Canada	2015	Female	Malignant neoplasms	35.29	37077.0

Appendix 1.5.F: Trend between Physical Inactivity and Respiratory disease Mortality Values using SQL.

```
query8 = '''SELECT p.Country, p.Year, p.Gender, m.Disease,p.Physical_Inactivity_Value, m.Mortality_Value FROM physicalinactivity p JOIN mortality m ON
        p.Year = m.Year AND p.Gender = m.Gender AND p.Country = m.Location WHERE p.Country = 'Canada' AND m.Disease = 'Respiratory Diseases';'''
resp_physical = pd.read_sql_query(query8, engine)
resp_physical
```

✓ 0.1s

	Country	Year	Gender	Disease	Physical_Inactivity_Value	Mortality_Value
0	Canada	2019	Male	Respiratory diseases	34.00	10194.0
1	Canada	2019	Both sexes	Respiratory diseases	35.58	19987.0
2	Canada	2019	Female	Respiratory diseases	37.16	9793.0
3	Canada	2018	Male	Respiratory diseases	33.44	10097.0
4	Canada	2018	Both sexes	Respiratory diseases	35.07	19700.0
5	Canada	2018	Female	Respiratory diseases	36.69	9603.0
6	Canada	2017	Male	Respiratory diseases	32.89	10257.0
7	Canada	2017	Both sexes	Respiratory diseases	34.56	19973.0
8	Canada	2017	Female	Respiratory diseases	36.23	9716.0
9	Canada	2016	Male	Respiratory diseases	32.33	9893.0
10	Canada	2016	Both sexes	Respiratory diseases	34.05	18990.0
11	Canada	2016	Female	Respiratory diseases	35.76	9097.0
12	Canada	2015	Male	Respiratory diseases	31.79	9740.0
13	Canada	2015	Both sexes	Respiratory diseases	33.55	18873.0
14	Canada	2015	Female	Respiratory diseases	35.29	9132.0

Appendix 1.5.G: SQL query and code for the Facet grid to observe the relationship between Physical Inactivity and each NCD Mortality Value.

```
# Correlation between physical inactivity and mortality diseases
query = """SELECT p.Country, p.Year, p.Gender,  p.Physical_Inactivity_Value, m.Mortality_Value, m.Disease FROM physicalinactivity p
    JOIN mortality m ON p.Year = m.Year AND p.Gender = m.Gender and p.Country=m.Location
    WHERE p.Country = 'Canada' AND p.Year BETWEEN 2000 AND 2018 order by m.Disease, m.Year;"""
result=pd.read_sql_query(query, engine)

pivot_df = result.pivot_table(index='Year', columns='Disease', values='Mortality_Value', aggfunc='first')
pivot_df = pivot_df.reset_index()
result = pd.merge(result, pivot_df, on='Year', how='left')

g = sns.FacetGrid(result, col="Disease", hue="Gender", col_wrap=2)
g.map_dataframe(sns.scatterplot, x="Physical_Inactivity_Value", y="Mortality_Value")
g.add_legend()
plt.show()
✓ 5.3s
```

Tobacco:

Appendix 1.6.A: SQL query for Filling out missing years Tobacco dataset.

```
query1 = """
WITH AvailableYears AS (
    SELECT DISTINCT Year, Country, Gender, tobacco_value, tobacco_value_low, tobacco_value_high
    FROM tobacco
    WHERE Year IN (2000, 2005, 2007, 2010, 2015, 2020)
),
YearRanges AS (
    SELECT 2001 AS MissingYear, 2000 AS PrevYear, 2005 AS NextYear UNION ALL
    SELECT 2002, 2000, 2005 UNION ALL
    SELECT 2003, 2000, 2005 UNION ALL
    SELECT 2004, 2000, 2005 UNION ALL
    SELECT 2006, 2005, 2007 UNION ALL
    SELECT 2008, 2007, 2010 UNION ALL
    SELECT 2009, 2007, 2010 UNION ALL
    SELECT 2011, 2010, 2015 UNION ALL
    SELECT 2012, 2010, 2015 UNION ALL
    SELECT 2013, 2010, 2015 UNION ALL
    SELECT 2014, 2010, 2015 UNION ALL
    SELECT 2016, 2015, 2020 UNION ALL
    SELECT 2017, 2015, 2020 UNION ALL
    SELECT 2018, 2015, 2020 UNION ALL
    SELECT 2019, 2015, 2020
),
CalculatedAverages AS (
    SELECT
        r.MissingYear,
        a1.Country,
        a1.Gender,
        AVG(a1.tobacco_value) AS avg_tobacco_value,
        AVG(a1.tobacco_value_low) AS avg_tobacco_value_low,
        AVG(a1.tobacco_value_high) AS avg_tobacco_value_high
    FROM YearRanges r
    JOIN AvailableYears a1 ON a1.Year IN (r.PrevYear, r.NextYear) AND a1.Country = a1.Country
    GROUP BY r.MissingYear, a1.Country, a1.Gender
)
SELECT
    'Estimate of current tobacco use prevalence' AS Indicator,
    'Americas' AS Region,
    c.Country,
    c.MissingYear AS Year,
    c.Gender,
    c.avg_tobacco_value AS Tobacco_value,
    c.avg_tobacco_value_low AS Tobacco_value_low,
    c.avg_tobacco_value_high AS Tobacco_value_high,
    NULL AS tobacco_range
FROM CalculatedAverages c;
"""

# Assuming you have a connection object named engine
result = pd.read_sql_query(query1, engine)
result.to_sql('tobacco', con=engine, if_exists='append', index=False)
tobacco = pd.read_sql_table("tobacco", engine)
tobacco.tail(5)
```

✓ 0.9s

	Indicator	Region	Country	Year	Gender	tobacco_value	tobacco_value_low	tobacco_value_high	tobacco_range
27220	Estimate of current tobacco use prevalence	Americas	Albania	2001	Female	10.35	7.90	12.85	None
27221	Estimate of current tobacco use prevalence	Americas	Chad	2004	Both sexes	10.40	7.65	13.20	None
27222	Estimate of current tobacco use prevalence	Americas	Chad	2003	Both sexes	10.40	7.65	13.20	None
27223	Estimate of current tobacco use prevalence	Americas	Chad	2002	Both sexes	10.40	7.65	13.20	None
27224	Estimate of current tobacco use prevalence	Americas	Chad	2001	Both sexes	10.40	7.65	13.20	None

Appendix 1.6.B: SQL query for determining the trend of Tobacco in Canada from 2000 to 2018.

```
# Over all trend of tobacco use for males and females in recenty years in Canada
query = """SELECT Country,Year,Gender,tobacco_value, tobacco_value_low, tobacco_value_high FROM tobacco
 where year<=2018 and Country='Canada' and gender!='Both sexes' order by year DESC;
 """
result=pd.read_sql_query(query, engine)
result
```

✓ 0.0s

	Country	Year	Gender	tobacco_value	tobacco_value_low	tobacco_value_high
0	Canada	2018	Female	12.10	10.15	14.00
1	Canada	2018	Male	16.75	14.40	19.15
2	Canada	2017	Male	16.75	14.40	19.15
3	Canada	2017	Female	12.10	10.15	14.00
4	Canada	2016	Male	16.75	14.40	19.15
5	Canada	2016	Female	12.10	10.15	14.00
6	Canada	2015	Female	13.50	11.30	15.60
7	Canada	2015	Male	18.20	15.80	20.60
8	Canada	2014	Female	15.25	12.90	17.55
9	Canada	2014	Male	19.90	17.00	22.85
10	Canada	2013	Female	15.25	12.90	17.55
11	Canada	2013	Male	19.90	17.00	22.85
12	Canada	2012	Female	15.25	12.90	17.55
13	Canada	2012	Male	19.90	17.00	22.85
14	Canada	2011	Male	19.90	17.00	22.85
15	Canada	2011	Female	15.25	12.90	17.55
16	Canada	2010	Female	17.00	14.50	19.50
17	Canada	2010	Male	21.60	18.20	25.10
18	Canada	2009	Female	18.15	15.45	20.90
19	Canada	2009	Male	22.75	19.25	26.30
20	Canada	2008	Female	18.15	15.45	20.90
21	Canada	2008	Male	22.75	19.25	26.30
22	Canada	2007	Female	19.30	16.40	22.30
23	Canada	2007	Male	23.90	20.30	27.50
24	Canada	2006	Female	20.25	17.25	23.30
25	Canada	2006	Male	24.85	21.05	28.65
26	Canada	2005	Female	21.20	18.10	24.30
27	Canada	2005	Male	25.80	21.80	29.80
28	Canada	2004	Female	24.00	20.25	27.70
29	Canada	2004	Male	28.25	23.55	32.95
30	Canada	2003	Female	24.00	20.25	27.70
31	Canada	2003	Male	28.25	23.55	32.95
32	Canada	2002	Female	24.00	20.25	27.70
33	Canada	2002	Male	28.25	23.55	32.95
34	Canada	2001	Female	24.00	20.25	27.70
35	Canada	2001	Male	28.25	23.55	32.95
36	Canada	2000	Female	26.80	22.40	31.10
37	Canada	2000	Male	30.70	25.30	36.10

Appendix 1.6.C: SQL queries for ranking Tobacco data in the year 2000, 2018 and the 18-year average ranking.

```
# Rank of Canada for average tobacco use in past 18 years among 165 countries
query = """WITH RankedTobacco AS (SELECT country, AVG(Tobacco_value) AS avg_tobacco, RANK() OVER (ORDER BY AVG(Tobacco_value) DESC) AS rank_value
    FROM tobacco WHERE Gender = 'Both sexes' AND Year >= 2000 AND Year <= 2018 GROUP BY country)
    SELECT country, avg_tobacco, rank_value FROM RankedTobacco WHERE country = 'Canada';"""
print("Canada rank for average TObacco use out of 165 countries")
ranked_data = pd.read_sql_query(query, engine)
ranked_data
```

✓ 0.0s

```
Canada rank for average TObacco use out of 165 countries
```

	country	avg_tobacco	rank_value
0	Canada	20.534211	104

```
# Rank of Canada for tobacco use in 2018 among 165 countries
query = """ WITH RankedTobacco AS ( SELECT country, Tobacco_value, RANK() OVER (ORDER BY Tobacco_value DESC) AS rank_value
    FROM tobacco WHERE Gender = 'Both sexes' AND Year = 2018) SELECT country, Tobacco_value, rank_value
    FROM RankedTobacco WHERE country = 'Canada';"""
ranked_data = pd.read_sql_query(query, engine)
print("Rank of Canada out of 165 countries in 2018")
ranked_data.head()
```

✓ 0.0s

```
Rank of Canada out of 165 countries in 2018
```

	country	Tobacco_value	rank_value
0	Canada	14.4	116

```
# Rank of Canada for tobacco use in 2000 among 165 countries
query = """ WITH RankedTobacco AS ( SELECT country, Tobacco_value, RANK() OVER (ORDER BY Tobacco_value DESC) AS rank_value
    FROM tobacco WHERE Gender = 'Both sexes' AND Year = 2000) SELECT country, Tobacco_value, rank_value
    FROM RankedTobacco WHERE country = 'Canada';"""
ranked_data = pd.read_sql_query(query, engine)
print("Rank of Canada out of 165 countries in 2000")
ranked_data.head()
```

✓ 0.0s

```
Rank of Canada out of 165 countries in 2000
```

	country	Tobacco_value	rank_value
0	Canada	28.7	92

Appendix 1.6.D: Trend between Tobacco and Cardiovascular Disease Mortality Values using SQL.

```
# Cardiovascular vs tobacco values to see if they have been increasing or decreasing over the years
query = """SELECT t.Country, t.Year, t.Gender,  t.tobacco_value, m.mortality_value, m.Disease FROM tobacco t
           JOIN mortality m ON t.Year = m.Year AND t.Gender = m.Gender and t.country=m.country
           WHERE t.Country = 'Canada' AND t.Year BETWEEN 2000 AND 2018
           AND t.Gender= 'Both sexes' AND m.Disease = 'Cardiovascular diseases';"""
result=pd.read_sql_query(query, engine)
result
```

✓ 0.1s

	Country	Year	Gender	tobacco_value	mortality_value	Disease
0	Canada	2018	Both sexes	14.40	70147.0	Cardiovascular diseases
1	Canada	2017	Both sexes	14.40	71372.0	Cardiovascular diseases
2	Canada	2016	Both sexes	14.40	69553.0	Cardiovascular diseases
3	Canada	2015	Both sexes	15.80	69548.0	Cardiovascular diseases
4	Canada	2014	Both sexes	17.55	68507.0	Cardiovascular diseases
5	Canada	2013	Both sexes	17.55	67296.0	Cardiovascular diseases
6	Canada	2012	Both sexes	17.55	65860.0	Cardiovascular diseases
7	Canada	2011	Both sexes	17.55	65294.0	Cardiovascular diseases
8	Canada	2010	Both sexes	19.30	66670.0	Cardiovascular diseases
9	Canada	2009	Both sexes	20.45	67611.0	Cardiovascular diseases
10	Canada	2008	Both sexes	20.45	68827.0	Cardiovascular diseases
11	Canada	2007	Both sexes	21.60	68833.0	Cardiovascular diseases
12	Canada	2006	Both sexes	22.55	68097.0	Cardiovascular diseases
13	Canada	2005	Both sexes	23.50	70622.0	Cardiovascular diseases
14	Canada	2004	Both sexes	26.10	71670.0	Cardiovascular diseases
15	Canada	2003	Both sexes	26.10	73397.0	Cardiovascular diseases
16	Canada	2002	Both sexes	26.10	73741.0	Cardiovascular diseases
17	Canada	2001	Both sexes	26.10	73945.0	Cardiovascular diseases
18	Canada	2000	Both sexes	28.70	75398.0	Cardiovascular diseases

Appendix 1.6.E: Trend between Tobacco and Respiratory Disease Mortality Values using SQL.

```
# Respiratory diseases vs tobacco values to see if they have been increasing or decreasing over the years
query = """SELECT t.Country, t.Year, t.Gender,  t.tobacco_value, m.mortality_value, m.Disease FROM tobacco t
    JOIN mortality m ON t.Year = m.Year AND t.Gender = m.Gender and t.country=m.country
    WHERE t.Country = 'Canada' AND t.Year BETWEEN 2000 AND 2018
    AND t.Gender= 'Both sexes' AND m.Disease = 'Respiratory diseases';"""
result=pd.read_sql_query(query, engine)
result
```

✓ 0.0s

	Country	Year	Gender	tobacco_value	mortality_value	Disease
0	Canada	2018	Both sexes	14.40	19700.0	Respiratory diseases
1	Canada	2017	Both sexes	14.40	19973.0	Respiratory diseases
2	Canada	2016	Both sexes	14.40	18990.0	Respiratory diseases
3	Canada	2015	Both sexes	15.80	18873.0	Respiratory diseases
4	Canada	2014	Both sexes	17.55	17989.0	Respiratory diseases
5	Canada	2013	Both sexes	17.55	17848.0	Respiratory diseases
6	Canada	2012	Both sexes	17.55	17670.0	Respiratory diseases
7	Canada	2011	Both sexes	17.55	17578.0	Respiratory diseases
8	Canada	2010	Both sexes	19.30	16903.0	Respiratory diseases
9	Canada	2009	Both sexes	20.45	16665.0	Respiratory diseases
10	Canada	2008	Both sexes	20.45	16621.0	Respiratory diseases
11	Canada	2007	Both sexes	21.60	16335.0	Respiratory diseases
12	Canada	2006	Both sexes	22.55	15287.0	Respiratory diseases
13	Canada	2005	Both sexes	23.50	16040.0	Respiratory diseases
14	Canada	2004	Both sexes	26.10	15307.0	Respiratory diseases
15	Canada	2003	Both sexes	26.10	14938.0	Respiratory diseases
16	Canada	2002	Both sexes	26.10	14502.0	Respiratory diseases
17	Canada	2001	Both sexes	26.10	14278.0	Respiratory diseases
18	Canada	2000	Both sexes	28.70	14291.0	Respiratory diseases

Appendix 1.6.F: Trend between Tobacco and Diabetes Mellitus Mortality Values using SQL.

```
# Diabetes mellitus vs tobacco values to see if they have been increasing or decreasing over the years
query = """SELECT t.Country, t.Year, t.Gender,  t.tobacco_value, m.mortality_value, m.Disease FROM tobacco t
    JOIN mortality m ON t.Year = m.Year AND t.Gender = m.Gender and t.country=m.country
    WHERE t.Country = 'Canada' AND t.Year BETWEEN 2000 AND 2018
    AND t.Gender= 'Both sexes' AND m.Disease='Diabetes mellitus';"""
result=pd.read_sql_query(query, engine)
result
```

✓ 0.0s

	Country	Year	Gender	tobacco_value	mortality_value	Disease
0	Canada	2018	Both sexes	14.40	6883.0	Diabetes mellitus
1	Canada	2017	Both sexes	14.40	7002.0	Diabetes mellitus
2	Canada	2016	Both sexes	14.40	6962.0	Diabetes mellitus
3	Canada	2015	Both sexes	15.80	7230.0	Diabetes mellitus
4	Canada	2014	Both sexes	17.55	7091.0	Diabetes mellitus
5	Canada	2013	Both sexes	17.55	7056.0	Diabetes mellitus
6	Canada	2012	Both sexes	17.55	7003.0	Diabetes mellitus
7	Canada	2011	Both sexes	17.55	7232.0	Diabetes mellitus
8	Canada	2010	Both sexes	19.30	6964.0	Diabetes mellitus
9	Canada	2009	Both sexes	20.45	6938.0	Diabetes mellitus
10	Canada	2008	Both sexes	20.45	7542.0	Diabetes mellitus
11	Canada	2007	Both sexes	21.60	7423.0	Diabetes mellitus
12	Canada	2006	Both sexes	22.55	7311.0	Diabetes mellitus
13	Canada	2005	Both sexes	23.50	7916.0	Diabetes mellitus
14	Canada	2004	Both sexes	26.10	7864.0	Diabetes mellitus
15	Canada	2003	Both sexes	26.10	8028.0	Diabetes mellitus
16	Canada	2002	Both sexes	26.10	7929.0	Diabetes mellitus
17	Canada	2001	Both sexes	26.10	7154.0	Diabetes mellitus
18	Canada	2000	Both sexes	28.70	6759.0	Diabetes mellitus

Appendix 1.6.G: Trend between Tobacco and Malignant Neoplasm Mortality Values using SQL.

```
# Malignant neoplasms(cancer) vs tobacco values to see if they have been increasing or decreasing over the years
query = """SELECT t.Country, t.Year, t.Gender,  t.tobacco_value, m.mortality_value, m.Disease FROM tobacco t
    JOIN mortality m ON t.Year = m.Year AND t.Gender = m.Gender and t.country=m.country
    WHERE t.Country = 'Canada' AND t.Year BETWEEN 2000 AND 2018
    AND t.Gender= 'Both sexes' AND m.Disease='Malignant neoplasms';"""
result=pd.read_sql_query(query, engine)
result
```

✓ 0.0s

	Country	Year	Gender	tobacco_value	mortality_value	Disease
0	Canada	2018	Both sexes	14.40	80881.0	Malignant neoplasms
1	Canada	2017	Both sexes	14.40	80994.0	Malignant neoplasms
2	Canada	2016	Both sexes	14.40	80696.0	Malignant neoplasms
3	Canada	2015	Both sexes	15.80	78020.0	Malignant neoplasms
4	Canada	2014	Both sexes	17.55	77763.0	Malignant neoplasms
5	Canada	2013	Both sexes	17.55	75755.0	Malignant neoplasms
6	Canada	2012	Both sexes	17.55	74983.0	Malignant neoplasms
7	Canada	2011	Both sexes	17.55	73328.0	Malignant neoplasms
8	Canada	2010	Both sexes	19.30	72535.0	Malignant neoplasms
9	Canada	2009	Both sexes	20.45	71726.0	Malignant neoplasms
10	Canada	2008	Both sexes	20.45	71197.0	Malignant neoplasms
11	Canada	2007	Both sexes	21.60	70257.0	Malignant neoplasms
12	Canada	2006	Both sexes	22.55	68729.0	Malignant neoplasms
13	Canada	2005	Both sexes	23.50	68005.0	Malignant neoplasms
14	Canada	2004	Both sexes	26.10	67746.0	Malignant neoplasms
15	Canada	2003	Both sexes	26.10	66883.0	Malignant neoplasms
16	Canada	2002	Both sexes	26.10	65980.0	Malignant neoplasms
17	Canada	2001	Both sexes	26.10	64668.0	Malignant neoplasms
18	Canada	2000	Both sexes	28.70	63429.0	Malignant neoplasms

Appendix 1.6.H: Facet Grid code for Tobacco and NCDs

```
# Correlation between Tobacco and mortality diseases
query = """SELECT t.Country, t.Year, t.Gender,  t.tobacco_value, m.mortality_value, m.Disease FROM tobacco t
    JOIN mortality m ON t.Year = m.Year AND t.Gender = m.Gender and t.country=m.country
    WHERE t.Country = 'Canada' AND t.Year BETWEEN 2000 AND 2018 order by m.disease, m.year;"""
result=pd.read_sql_query(query, engine)

pivot_df = result.pivot_table(index='Year', columns='Disease', values='mortality_value', aggfunc='first')
pivot_df = pivot_df.reset_index()
result = pd.merge(result, pivot_df, on='Year', how='left')

g = sns.FacetGrid(result, col="Disease", hue="Gender", col_wrap=2)
g.map_dataframe(sns.scatterplot, x="tobacco_value", y="mortality_value")
g.add_legend()
plt.savefig("Visuals/tobacco_mortality pairplot.png", format='png', dpi=300)
plt.show()
```

Guiding Question 1:

Appendix 2.1: Alcohol SQL query and code

```
# plotting alcohol trends in Canada from 2000-2018
import seaborn as sns
import matplotlib.pyplot as plt

query_alc_trend_vis = """SELECT Year, Mean_Alcohol_Value, Gender FROM alcohol_levels where Year<=2018 and Year>=2000 and Country='Canada';"""
trend_alc=pd.read_sql_query(query_alc_trend_vis, engine)
trend_alc['Year'] = trend_alc['Year'].astype(int)
plt.figure(figsize=(20, 12))

sns.lineplot(data=trend_alc, x=trend_alc['Year'], y=trend_alc['Mean_Alcohol_Value'], hue='Gender', marker='o', markersize=6, palette=['Blue','Magenta','Green'])
plt.title("Alcohol Usage by Gender (2000-2018)", fontsize=25, weight='bold')
plt.xlabel("Years", fontsize=15,weight='bold')

# ENTER UNIT HERE
plt.ylabel("Alcohol Value (ENTER UNIT HERE)", fontsize=15, weight='bold')

plt.legend(frameon=True, shadow=True, loc='upper right',fontsize=12,title="Gender", title_fontsize=15)
years = trend_alc['Year'].unique()
plt.grid(color='gray', linestyle='--', linewidth=0.5, alpha=0.7)
plt.xticks(years, rotation=45)
plt.xticks(ticks=years, labels=years, rotation=45,fontsize=15)
plt.yticks(fontsize=15)
```

Appendix 2.2: Physical Inactivity Trend Graph SQL query and code

```
# Physical Inactivity trend in Canada from 2000 to 2018
query = """SELECT Year, Physical_Inactivity_Value, Gender FROM physicalinactivity where year<=2018 and country='Canada';"""
canada_physical_inactivity=pd.read_sql_query(query, engine)
canada_physical_inactivity['Year'] = canada_physical_inactivity['Year'].astype(int)

plt.figure(figsize=(20, 12))

sns.lineplot(data=canada_physical_inactivity, x=canada_physical_inactivity['Year'], y=canada_physical_inactivity['Physical_Inactivity_Value'],
             hue='Gender', marker='o', markersize=6, palette=['Magenta','Green','Blue'])
plt.title("Physical Inactivity by Gender (2000-2018)", fontsize=25, weight='bold')
plt.xlabel("Years", fontsize=15,weight='bold')
plt.ylabel("Physical Inactivity Value (% of Population)", fontsize=15, weight='bold')
plt.legend(frameon=True, shadow=True, loc='upper right',fontsize=12,title="Gender", title_fontsize=15)
years = canada_physical_inactivity['Year'].unique()
plt.grid(color='gray', linestyle='--', linewidth=0.5, alpha=0.7)
plt.xticks(years, rotation=45)
plt.xticks(ticks=years, labels=years, rotation=45,fontsize=15)
plt.yticks(fontsize=15)

# Show the plot
plt.show()
```

Appendix 2.3: Tobacco Trend Graph SQL query and code

```
#Trend line for tobacco use in canada from 2000-2018
query = """SELECT Year, tobacco_value, Gender FROM tobacco where year<=2018 and country='Canada';"""
canada_tobacco=pd.read_sql_query(query, engine)
canada_tobacco['Year'] = canada_tobacco['Year'].astype(int)
plt.figure(figsize=(20, 12))

sns.lineplot(data=canada_tobacco, x=canada_tobacco['Year'], y=canada_tobacco['tobacco_value'], hue='Gender', marker='o', markersize=6, palette=['Magenta','Green','Blue'])
plt.title("Tobacco Usage by Gender (2000-2018)", fontsize=25, weight='bold')
plt.xlabel("Years", fontsize=15,weight='bold')
plt.ylabel("Tobacco Value (% of Population)", fontsize=15, weight='bold')
plt.legend(frameon=True, shadow=True, loc='upper right',fontsize=12,title="Gender", title_fontsize=15)
years = canada_tobacco['Year'].unique()
plt.grid(color='gray', linestyle='--', linewidth=0.5, alpha=0.7)
plt.xticks(years, rotation=45)
plt.xticks(ticks=years, labels=years, rotation=45,fontsize=15)
plt.yticks(fontsize=15)

# Show the plot
plt.savefig("Visuals/tobacco_trends_by_gender.png", format='png', dpi=300)
plt.show()
```

Guiding Question 2:

Appendix 3.1: Cholesterol Trend Graph code

```
query_chol_trend = """ SELECT Country, Year, Gender, Cholesterol_Value
FROM cholesterol_table
  WHERE Year >= '2000'
  AND Year <= '2018'
  AND Country = 'Canada';"""
```

Appendix 3.2: Hypertension Trend Graph Code

```
query_hyper_trend = """ SELECT Country, Year, Gender, Hypertension_Value
FROM hypertension_table
  WHERE Year >= '2000'
  AND Year <= '2018'
  AND Country = 'Canada';"""
```

Appendix 3.3: Obesity Trend Graph Code

```
query = """
SELECT Year, Obesity_Value, Gender, Country
FROM obesity_table
WHERE Year >= '2000' AND Year <= '2018'
AND Country = 'Canada';
"""
```

Correlation of Factors:

Appendix 4.1: SQL query for joining the six Risk Factors

```
query="""SELECT t.Country, t.Year, t.Gender, t.tobacco_value, o.obesity_value, a.alchohol_value, c.cholesterol_value,
ph.phy_inactivity_value, h.hypertension_value FROM tobacco t
JOIN obesity o ON t.Year = o.Year AND t.Gender = o.Gender and t.country=o.country
JOIN alchohol a ON t.Year = a.Year AND t.Gender = a.Gender and t.country=a.country
JOIN cholesterol c ON t.Year = c.Year AND t.Gender = c.Gender and t.country=c.country
JOIN phy_inactivity ph ON t.Year = ph.Year AND t.Gender = ph.Gender and t.country=ph.country
JOIN hypertension h ON t.Year = h.Year AND t.Gender = h.Gender and t.country=h.country
WHERE t.Country = 'Canada' AND t.Year BETWEEN 2000 AND 2018;"""
result=pd.read_sql_query(query, engine)
result.rename(columns={'tobacco_value':'Tobacco', 'obesity_value':'Obesity' ,'alchohol_value' : 'Alcohol','cholesterol_value' : 'Cholesterol',
                       'phy_inactivity_value': 'Physical Inactivity', 'hypertension_value' : 'Hypertension'}, inplace=True)
result.head()
```

	Country	Year	Gender	Tobacco	Obesity	Alcohol	Cholesterol	Physical Inactivity	Hypertension
0	Canada	2018	Male	16.75	26.63	15.59	4.6	33.44	24.5
1	Canada	2018	Both sexes	14.40	25.57	9.90	4.6	35.07	22.4
2	Canada	2018	Female	12.10	24.53	4.34	4.7	36.69	20.2
3	Canada	2017	Male	16.75	26.41	15.68	4.6	32.89	24.8
4	Canada	2017	Both sexes	14.40	25.43	9.95	4.7	34.56	22.6

Appendix 4.2: Correlation of Factors code

```
#select only numerical columns
factors = result.iloc[:, 3:]
correlation_matrix = factors.corr()
#plotting correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Matrix of Behavioural and Metabolic Risk Factors')
plt.savefig(f"Visuals/Correlation btw factors.png", format='png', dpi=300)
plt.show()
```

Appendix 4.3: Scatter plot code for Behavioral vs Metabolic Risk Factors

```
#plotting scatter plot for for each behavioral factor against metabolic factors
metabolic_factors=['Cholesterol', 'Hypertension', 'Obesity']
behavioural_factors=['Tobacco', 'Physical Inactivity', 'Alcohol']
for bfactor in behavioural_factors:
    for factor in metabolic_factors:
        aggregated = result.groupby([bfactor, 'Gender'])[factor].mean().reset_index()
        plt.figure(figsize=(15, 7))
        sns.scatterplot(data=aggregated, x=factor, y=bfactor, hue='Gender', palette={'Female':'Magenta','Both sexes':'Green', 'Male':'Blue'}, s=100)
        plt.title(f"Relationship between {bfactor} and {factor}", fontsize=20, weight='bold')
        plt.xlabel(f"{bfactor}",  fontsize=15,weight='bold')
        plt.ylabel(f"{factor}",  fontsize=15, weight='bold')
        plt.legend(frameon=True, shadow=True, loc='upper left',fontsize=12,title="Gender", title_fontsize=15)
        plt.savefig(f"Visuals/{bfactor} vs {factor}.png", format='png', dpi=300)
        plt.show()
```

Regression Analysis

Appendix 5.1: SQL query and table for joining the six Risk Factors with Mortality data

```
# Joining all seven datasets
query = """SELECT t.Country, t.Year, t.Gender, t.tobacco_value, o.obesity_value, a.alchohol_value, c.cholesterol_value,
ph.phy_inactivity_value, h.hypertension_value, m.mortality_value, m.Disease FROM tobacco t
JOIN obesity o ON t.Year = o.Year AND t.Gender = o.Gender and t.country=o.country
JOIN alchohol a ON t.Year = a.Year AND t.Gender = a.Gender and t.country=a.country
JOIN cholesterol c ON t.Year = c.Year AND t.Gender = c.Gender and t.country=c.country
JOIN phy_inactivity ph ON t.Year = ph.Year AND t.Gender = ph.Gender and t.country=ph.country
JOIN hypertension h ON t.Year = h.Year AND t.Gender = h.Gender and t.country=h.country
JOIN mortality m ON t.Year = m.Year AND t.Gender = m.Gender and t.country=m.country
WHERE t.Country = 'Canada' AND t.Year BETWEEN 2000 AND 2018 AND t.Gender != 'Both sexes';"""
result=pd.read_sql_query(query, engine)
result.rename(columns={'tobacco_value':'Tobacco', 'obesity_value':'Obesity' ,'alchohol_value' : 'Alcohol','cholesterol_value' : 'Cholesterol',
                       'phy_inactivity_value': 'Physical Inactivity', 'hypertension_value' : 'Hypertension', 'mortality_value':'Mortality'}, inplace=True)
result.head()
```

	Country	Year	Gender	Tobacco	Obesity	Alcohol	Cholesterol	Physical Inactivity	Hypertension	Mortality	Disease
0	Canada	2018	Male	16.75	26.63	15.59	4.6	33.44	24.5	10097.0	Respiratory diseases
1	Canada	2018	Female	12.10	24.53	4.34	4.7	36.69	20.2	2994.0	Diabetes mellitus
2	Canada	2018	Female	12.10	24.53	4.34	4.7	36.69	20.2	34074.0	Cardiovascular diseases
3	Canada	2018	Male	16.75	26.63	15.59	4.6	33.44	24.5	36074.0	Cardiovascular diseases
4	Canada	2018	Female	12.10	24.53	4.34	4.7	36.69	20.2	38261.0	Malignant neoplasms

Appendix 5.2: Python code for Regression analysis of Respiratory Diseases against the six factors

```
# extract respiratory disease data from above query
X = result[result['Disease'] == 'Respiratory diseases'][['Tobacco', 'Obesity', 'Alcohol', 'Cholesterol', 'Physical Inactivity', 'Hypertension']]
y = result[result['Disease'] == 'Respiratory diseases'][[ 'Mortality']]

# regression
model = sm.OLS(y, X).fit()
print(model.summary())
```

Appendix 5.3: Python code for Regression analysis of Cardiovascular Diseases against the six factors

```
# extract Cardio disease data from above query
X = result[result['Disease'] == 'Cardiovascular diseases'][['Tobacco', 'Obesity', 'Alcohol', 'Cholesterol', 'Physical Inactivity', 'Hypertension']]
y = result[result['Disease'] == 'Cardiovascular diseases'][[ 'Mortality']]

# regression
model = sm.OLS(y, X).fit()
print(model.summary())
```

Appendix 5.4: Python code for Regression analysis of Diabetes mellitus against the six factors

```
# extract diabetes data from above query
X = result[result['Disease'] == 'Diabetes mellitus'][['Tobacco', 'Obesity', 'Alcohol', 'Cholesterol', 'Physical Inactivity', 'Hypertension']]
y = result[result['Disease'] == 'Diabetes mellitus'][[ 'Mortality']]

# regression
model = sm.OLS(y, X).fit()
print(model.summary())
```

Appendix 5.5: Python code for Regression analysis of Malignant neoplasms against the six factors.

```
# extract Malignant neoplasms data from above query
X = result[result['Disease'] == 'Malignant neoplasms'][['Tobacco', 'Obesity', 'Alcohol', 'Cholesterol', 'Physical Inactivity', 'Hypertension']]
y = result[result['Disease'] == 'Malignant neoplasms'][[ 'Mortality']]

# regression
model = sm.OLS(y, X).fit()
print(model.summary())
```