Social Inequalities in Chronic Diseases

Huyen Tran

6 June 2023

Table of contents

1. Introduction
2. Data and Data Sources
3. EDA - Exploratory Data Analysis
    3.1. Data Cleaning (and columns definition)
    3.2. Data Distribution
    3.3. What is the difference between prevalence and incidence
    3.4. Weighted number of patients.
        3.4.1. Social class has an significant impact on the prevalence cases.
        3.4.2. Ranking of Weighted Number of Cases by Education Level.
        3.4.3. Ranking of Weighted Number of Cases by Profession
        3.4.4. Violin charts
    3.5. Standardised Rate with Direct and Indirect method
4. SQL
    4.1. Choose the Database type.
    4.2. Entities Relationship Diagram.
    4.3. Queries.
5. Conclusions
6. Recommendations

1. Introduction

In recent years, social inequalities in healthcare outcomes have become a subject of growing concern globally. In France, like many other countries, chronic diseases pose a significant burden on individuals and the healthcare system. However, there is a growing recognition that the impact of chronic diseases is not evenly distributed across the population, with certain social groups experiencing disproportionately higher rates of incidence, prevalence, and adverse outcomes.

For instance, studies have shown that people from lower socio-economic backgrounds are more likely to suffer from chronic diseases than those from higher socio-economic backgrounds. This is due to a range of factors, including differences in lifestyle factors, access to healthcare, and exposure to environmental risks. Furthermore, certain communities, such as immigrants may face additional challenges in accessing healthcare and managing chronic diseases.

This report seeks to shed light on the intersection of social inequalities and chronic diseases in France. By examining the social factors that contribute to health disparities, I aim to uncover insights that can drive informed decision-making and shape strategies to address these inequalities. By understanding the implications of social inequalities, organisations and policymakers can play a pivotal role in promoting better health outcomes for all.

Moreover, the implications of social inequalities in chronic diseases have important economic consequences. Chronic diseases place a significant burden on the healthcare system, with the costs of treatment and management being borne by individuals, healthcare providers, and society as a whole. By addressing social inequalities in healthcare outcomes, we can also promote greater economic prosperity by reducing the financial burden of chronic diseases.

With a better understanding of the implications of social inequalities, we can work towards a future where everyone has access to quality healthcare, regardless of their social background. By promoting health equity, we can create a more just and equitable society where everyone has the opportunity to live a healthy and fulfilling life.

2. Data and data sources

https://data.drees.solidarites-sante.gouv.fr/explore/dataset/er_inegalites_maladies_chroniques/information/?sort=poidstot&refine.catlib=Traitements+du+risque+vasculaire

Information about dataset

This data presents the incidence and prevalence rates of chronic diseases based on various socio-demographic variables such as Gender, Age Group, Region, Tenth of Standard of Living, Socio-professional group, and Education. The period covered is from 2016 to 2017 and the population studied includes those living in metropolitan France or in the overseas territories except for Mayotte, for which there is not sufficient data.

These statistics were compiled for the publication "Chronic diseases most often affect those on low-incomes and significantly reduce their life expectancy" (Studies and Results No. 1243)(1)

3. EDA - Exploratory data analysis

3.1. Data Cleaning

After reviewing the data, I have decided to rename columns to enhance readability and drop columns that are not essential to the data exploratory process. When completing the data modifications, I obtained the data's shape and information about any missing values in each column.

```
MORE DATA INFO :
-----------------------------------------------------
Data shape : (46176, 15)

Total rows in the dataset : 46,176
Total columns in the dataset : 15
Total duplicated values : 0
Total null values : 7,722

RATIO OF MISSING AND DUPLICATED VALUES IN OUR DATA :
-----------------------------------------------------

Percentage of null values in the data : 16.72%
Percentage of duplicates in the data : 0.0%
```

TOTAL MISSING VALUES FOR EACH COLUMN :

	MISSING VALUES
Type	0
Disease_Name	0
Disease_Group	0
Region	3016
Demographic	1768
Demographic_Indicator	1768
weights	4
weights_total	4
txNonStand	4
txStandDir	4
txStandDirModBB	4
txStandDirModBH	4
txStandIndir	382
txStandIndirModBB	382
txStandIndirModBH	382

After careful consideration of the high number of missing values in the "Region" column, I explored the option of filling them using K-Nearest Neighbours Imputer. However, it became apparent that this approach would introduce bias into the data. Therefore, I made the difficult decision to drop the column altogether and remove any remaining missing values in the other columns. This ensures a more robust and accurate dataset for analysis.

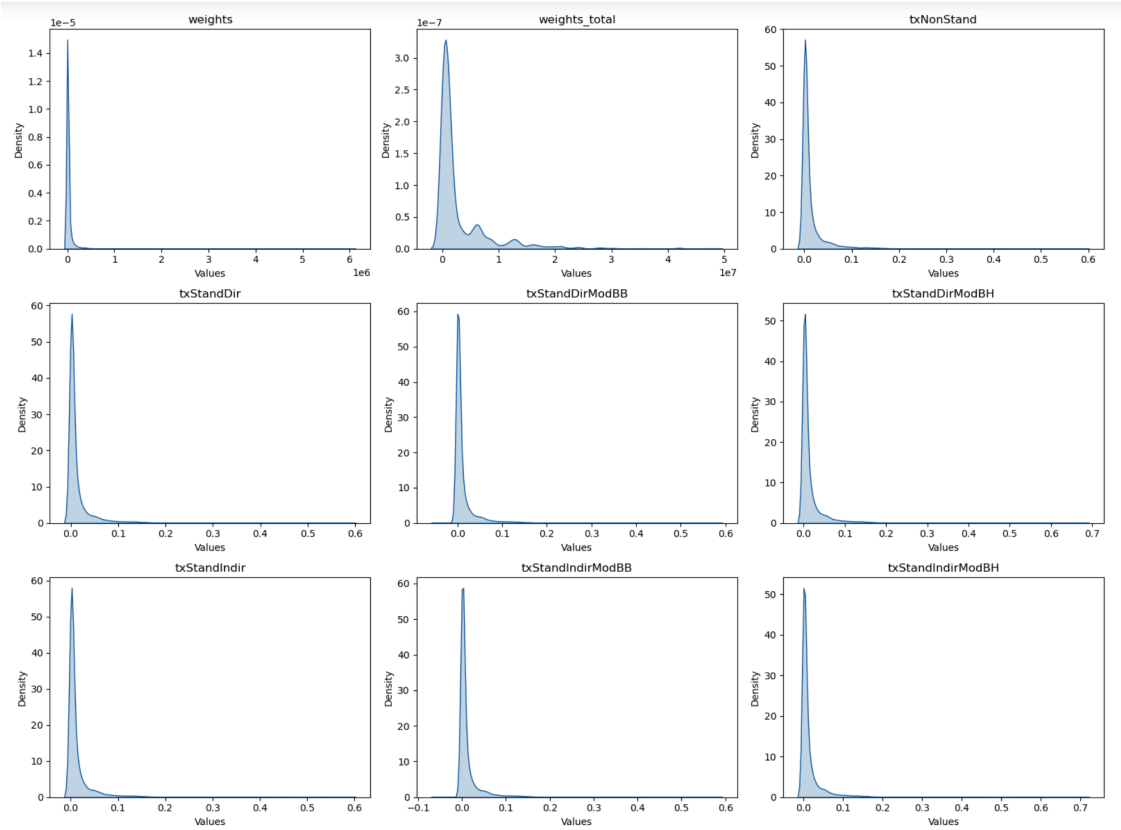
```
Type                    0
Disease_Name            0
Disease_Group           0
Demographic             0
Demographic_Indicator   0
weights                 0
weights_total           0
txNonStand              0
txStandDir              0
txStandDirModBB         0
txStandDirModBH         0
txStandIndir            0
txStandIndirModBB       0
txStandIndirModBH       0
dtype: int64
```

Columns definition:

- “Type”: Type of value ("incidence" or "prevalence")
- “Disease_Name”: Name of the Disease
- “Disease_Group”: Group of the Disease
- “Demographic”: Demographic Information such as Social Class, Genders, Age Group, Jobs and Degrees.
- “Demographic_Indicator”: Indicators that explain each Demographic Information
- “weights”: Weighted numbers of people who are ill (prevalence) or fall ill (incidence).
- “weights_total”: Total Weighted numbers of people who are ill (prevalence) or fall ill (incidence).
- “txNonStand”: Non-Standardised Rate: weights / weights_total
- “txStandDir”: Standardised Rate with the Direct Method
- “txStandDirModBB”: Standardised Rate with the Direct Method (lower limit of the 95% confidence interval)
- “txStandDirModBH”: Standardised Rate with the Direct Method (upper limit of the 95% confidence interval)
- “txStandIndir”: Standardised Rate with the Indirect Method
- “txStandIndirModBB”: Standardised Rate with the Indirect Method (lower limit of the 95% confidence interval)
- “txStandIndirModBH”: Standardised Rate with the Indirect Method (upper limit of the 95% confidence interval)

3.2. Data Distribution

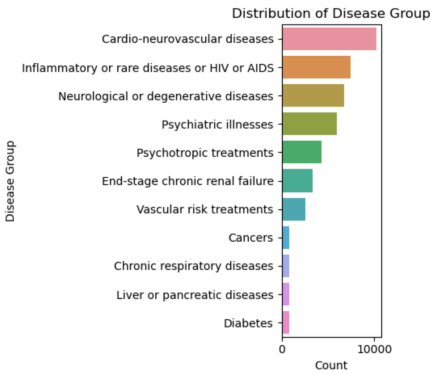
Graph a. Distribution of numeric columns:



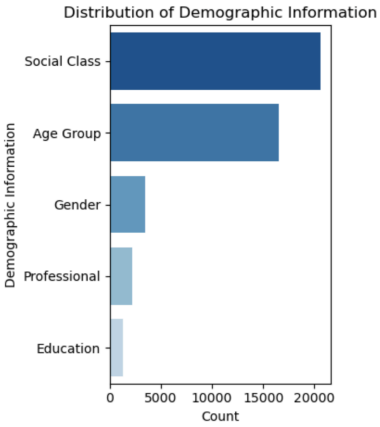
In the graph depicting the Distribution of numeric columns (graph a), it is evident that eight columns display a right skew. This observation can be attributed to the elongation of the distribution towards higher values, indicating that occurrences of higher values are less frequent but possess greater magnitude. The presence of a right skew suggests the existence of outliers or extreme values on the higher end of the distribution.

However, it is worth noting that in healthcare data, outliers may hold valuable information or represent significant events or conditions. Consequently, in consideration of the domain-specific nature of the data, it has been decided to retain these outliers for subsequent analysis and interpretation. It is important to acknowledge the potential influence of these outliers on the statistical properties of the data and to account for their presence in the analytical procedures.

The distribution of diseases group in the dataset reveals interesting patterns. The Cardio-neurovascular diseases group exhibits the highest prevalence, including a total of 12 different diseases. Following behind closely is the Inflammatory or rare diseases or HIV or AIDS group, encompassing 9 diseases. In contrast, the cancers group



does not provide a breakdown of specific cancer types, which explains the relatively lower representation of cancers in the overall distribution. This chart provides an overview of how the data categorises and presents each disease group.



The distribution of demographic information illustrates that the highest number of rows corresponds to the social class variable, indicating a primary focus on capturing social inequalities across different classes in France. On the other hand, the education variable has the lowest number of rows, suggesting relatively less emphasis on educational disparities within the dataset. It implies a specific interest in studying and addressing the inequities prevalent among different social strata.

3.3. What is the difference between prevalence and incidence.

By looking at the epidemiologist's bathtub (2), we can easily understand the definition of prevalence and incidence in healthcare data. Prevalence differs from incidence in that prevalence includes all cases, both new and preexisting, in the population at the specified time (in this case, the weighted number of patients in France from 2016 to 2017), whereas incidence is limited to new case only.

The number of rows for prevalence and incidence in our data:

```
prevalence    22124
incidence     21920
Name: Type, dtype: int64
```

3.4. Weighted number of patients.

The concept of weighted numbers (effectifs pondérés) is used to account for the varying degrees of impact on our patient population. By assigning weights based on the severity or frequency of the condition, the data provides a more accurate representation of the burden on our healthcare system. For example, out of the total patient population of 500, we found that 100 individuals were affected by the condition. However, using the concept of weighted numbers, we assigned higher weights to those with more severe symptoms or longer duration of illness. By doing so, we estimated that the weighted headcount of individuals affected by the condition was 120. This adjustment accounts for the fact that some patients may require more intensive treatment or have a greater impact on healthcare resources.

Similarly, when discussing the incidence of individuals falling ill, we considered the concept of weighted numbers to better capture the dynamic nature of the condition. By assigning weights based on factors such as age, pre-existing conditions, or susceptibility, we can assess the true burden of new cases on our patient population. For example, in the past year, we recorded 50 new cases of the condition. However, by applying the concept of weighted numbers, we found that the weighted incidence was 60. This adjustment reflects the fact that certain individuals may have a higher likelihood of contracting the condition due to their demographic or health-related factors.

3.4.1. Social class has an significant impact on the prevalence cases.

Table d. Total weighted number of patients by Social Class for each Disease Group:

Type	Disease_Group \ Demographic_Indicator	1	2	3	4	5	6	7	8	9	10
incidence	Cancers	212,172	254,828	296,205	302,589	308,269	295,988	299,974	316,566	348,876	384,580
	Cardio-neurovascular diseases	1,272,886	1,753,830	1,998,851	1,886,429	1,725,999	1,605,273	1,536,623	1,484,984	1,597,562	1,649,315
	Chronic respiratory diseases	640,698	653,756	633,626	606,400	592,856	575,793	554,705	531,173	530,143	491,877
	Diabetes	187,211	177,702	163,761	153,988	145,311	140,101	134,426	126,906	124,341	114,260
	End-stage chronic renal failure	20,535	20,694	19,592	17,657	19,284	15,012	20,212	15,625	13,379	13,050
	Inflammatory or rare diseases or HIV or AIDS	155,470	168,539	175,561	181,315	162,798	180,618	173,959	166,797	174,305	174,405
	Liver or pancreatic diseases	120,976	120,442	124,261	108,193	106,713	91,653	88,637	85,130	81,652	83,376
	Neurological or degenerative diseases	317,868	427,126	468,664	441,495	395,598	359,330	337,192	317,701	324,503	346,207
	Psychiatric illnesses	823,147	842,624	820,829	773,816	690,145	623,875	595,430	554,316	535,100	479,440
	Psychotropic treatments	1,715,879	1,933,990	2,054,173	2,098,965	2,038,025	2,051,986	2,010,759	2,040,744	2,110,078	2,065,059
	Vascular risk treatments	1,067,946	1,115,657	1,159,775	1,161,111	1,153,475	1,198,774	1,190,134	1,245,659	1,329,575	1,450,640
prevalence	Cancers	1,192,879	1,560,473	1,859,944	1,911,986	1,933,671	1,928,459	1,941,225	2,070,328	2,285,752	2,542,571
	Cardio-neurovascular diseases	5,373,920	7,353,257	8,400,943	8,052,555	7,502,942	6,980,628	6,639,148	6,499,843	6,872,951	6,971,750
	Chronic respiratory diseases	2,370,765	2,513,779	2,446,051	2,306,487	2,189,939	2,103,372	1,992,911	1,892,337	1,822,468	1,641,065
	Diabetes	2,406,421	2,591,320	2,670,841	2,484,489	2,346,200	2,146,647	2,009,323	1,957,522	1,878,108	1,675,478
	End-stage chronic renal failure	136,108	132,360	127,205	127,638	105,525	92,675	89,915	79,722	83,453	78,545
	Inflammatory or rare diseases or HIV or AIDS	1,302,858	1,486,927	1,504,716	1,555,082	1,479,614	1,436,906	1,429,144	1,457,773	1,473,509	1,347,288
	Liver or pancreatic diseases	461,146	428,069	392,064	350,616	335,090	295,137	284,619	265,656	259,814	240,067
	Neurological or degenerative diseases	1,942,533	2,409,381	2,556,265	2,387,721	2,097,733	1,957,818	1,775,650	1,728,172	1,744,962	1,787,834
	Psychiatric illnesses	5,368,761	5,474,137	4,737,868	4,078,321	3,530,604	3,113,311	2,801,501	2,571,054	2,399,883	2,077,016
	Psychotropic treatments	6,316,054	7,608,863	8,416,848	8,318,375	7,942,055	7,657,693	7,334,214	7,384,218	7,557,545	7,321,155
	Vascular risk treatments	7,613,148	9,829,293	11,534,845	11,857,824	12,122,995	12,346,350	12,309,596	13,061,666	13,986,157	14,313,414

Table e: Social Class definition:

1	Dixième le plus modeste de la population	10th decile (least affluent segment)
2	Deuxième dixième le plus modeste de la population	2nd decile from bottom (second least affluent segment)
3	Troisième dixième le plus modeste de la population	3rd decile from bottom (third least affluent segment)
4	Quatrième dixième le plus modeste de la population	4th decile from bottom (fourth least affluent segment)
5	Cinquième dixième le plus modeste de la population	5th decile from bottom (fifth least affluent segment)
6	Cinquième dixième le plus aisé de la population	5th decile from top (fifth most affluent segment)
7	Quatrième dixième le plus aisé de la population	4th decile from top (fourth most affluent segment)
8	Troisième dixième le plus aisé de la population	3rd decile from top (third most affluent segment)
9	Deuxième dixième le plus aisé de la population	2nd decile from top (second most affluent segment)
10	Dixième le plus aisé de la population	10th decile (most affluent segment)

In the graphic depicting the "Total weighted number of patients by Social Class for each Disease Group," the wealthiest population shows the highest weighted number of prevalence cases for Vascular risk treatment. This suggests that they likely have better access to healthcare facilities, specialised treatments, and preventive care, resulting in higher rates of diagnosis and treatment for vascular risk. Conversely, the poorest population faces obstacles such as lack of health insurance, limited healthcare infrastructure, and financial difficulties, leading to a significantly lower weighted number of treatment cases.

Regarding the incidence cases (or new cases in 2017) for Vascular risk treatment, the gap in weighted cases between the wealthiest and the poorest is smaller. While this is a positive trend, it is important to monitor the situation over time. If the poor population continues to face barriers in accessing healthcare facilities, this gap may widen progressively.

In France, the most common chronic diseases are Cardio-neurovascular (5.1 million people treated in 2019), diabetes (4.0 million), chronic respiratory diseases (3.7 million), cancer (3.3 million), psychiatric diseases (2.5 million), neurological or degenerative diseases (1.7 million), inflammatory or rare diseases or HIV/AIDS[3] (1.3 million) and liver or pancreatic diseases (0.6 million) [CNAM, 2021]. Let's examine the table d and the weights of cancer cases in order to understand why the prevalence is highest among the wealthiest population and lowest among the poorest. This observation can be attributed to the privileges that wealthier individuals have in terms of disease screening and access to treatment. They are more likely to undergo regular cancer screenings and receive timely treatment, resulting in higher detection rates. On the other hand, individuals from lower socioeconomic backgrounds may not have the same level of access to healthcare resources, leading to delayed or limited diagnosis of cancers. As a result, a higher proportion of cancer cases among the poorer population are detected at advanced stages, leading to increased mortality rates.

Furthermore, there is a significant disparity in the prevalence of psychiatric illnesses between the wealthiest and poorest populations. The prevalence of psychiatric disorders is notably higher among the poorest population compared to the wealthier population. Some psychiatric conditions, particularly those that manifest early in life, can have long-term impacts on education and employment opportunities. This, in turn, directly affects income levels and overall standards of living. Individuals in the average socioeconomic class who suffer from psychiatric illnesses may experience a downward socioeconomic shift, potentially transitioning to a lower socioeconomic class. This dynamic creates a substantial gap between the wealthiest and poorest populations, further exacerbating the disparities in psychiatric illness prevalence.

3.4.2. Ranking of Weighted Number of Cases by Education Level.

Type	Disease_Group	Demographic_Indicator 1	2	3	4
incidence	Cancers	7	6	6	6
	Cardio-neurovascular diseases	1	2	3	3
	Chronic respiratory diseases	6	5	5	5
	Diabetes	8	9	9	9
	End-stage chronic renal failure	11	11	11	11
	Inflammatory or rare diseases or HIV or AIDS	9	8	8	8
	Liver or pancreatic diseases	10	10	10	10
	Neurological or degenerative diseases	5	7	7	7
	Psychiatric illnesses	4	4	4	4
	Psychotropic treatments	2	1	1	1
	Vascular risk treatments	3	3	2	2

Type	Disease_Group	Demographic_Indicator 1	2	3	4
prevalence	Cancers	7	6	5	5
	Cardio-neurovascular diseases	2	3	3	3
	Chronic respiratory diseases	8	7	6	6
	Diabetes	6	5	9	8
	End-stage chronic renal failure	11	11	11	11
	Inflammatory or rare diseases or HIV or AIDS	9	9	8	7
	Liver or pancreatic diseases	10	10	10	10
	Neurological or degenerative diseases	5	8	7	9
	Psychiatric illnesses	4	4	4	4
	Psychotropic treatments	3	2	2	2
	Vascular risk treatments	1	1	1	1

1	Pas de diplôme	No degree
2	BEP/CAP	Professional degree
3	Baccalauréat	High school dgree
4	Enseignement supérieur	Higher education

The analysis reveals notable patterns in the distribution of weighted new cases (incidence) across different education levels in France. Among individuals with no degree, Cardio-neurovascular diseases exhibit the highest weighted number of new cases, indicating a significant health burden in this population segment. On the other hand, for the remaining education levels, Psychotropic treatments emerge as the category with the highest weighted number of new cases. This observation suggests a particular focus on addressing psychotropic health concerns in the country.

When considering the existing cases (prevalence), regardless of education level, vascular risk treatments consistently occupy the top position. This finding underscores the prevalence and importance of interventions targeting vascular risk factors in the overall healthcare landscape. These insights shed light on the distribution of chronic diseases in relation to education levels in France, providing valuable information for public health planning and resource allocation efforts.

3.4.3. Ranking of Weighted Number of Cases by Profession

Type	Disease_Group	1	2	3	4	5	6	8
incidence	Cancers	7	7	6	6	7	7	8
	Cardio-neurovascular diseases	1	1	2	2	2	1	5
	Chronic respiratory diseases	6	5	4	5	5	5	2
	Diabetes	9	8	9	9	9	8	9
	End-stage chronic renal failure	11	11	11	11	11	11	11
	Inflammatory or rare diseases or HIV or AIDS	8	9	8	8	8	9	7
	Liver or pancreatic diseases	10	10	10	10	10	10	10
	Neurological or degenerative diseases	4	6	7	7	6	6	6
	Psychiatric illnesses	5	4	5	4	4	4	3
	Psychotropic treatments	2	2	1	1	1	2	1
	Vascular risk treatments	3	3	3	3	3	3	4

Type	Disease_Group	1	2	3	4	5	6	8
prevalence	Cancers	7	6	4	5	7	8	9
	Cardio-neurovascular diseases	2	2	3	3	3	2	5
	Chronic respiratory diseases	8	8	7	8	8	7	4
	Diabetes	6	4	6	6	5	5	8
	End-stage chronic renal failure	11	11	11	11	11	11	11
	Inflammatory or rare diseases or HIV or AIDS	9	9	8	9	9	9	7
	Liver or pancreatic diseases	10	10	10	10	10	10	10
	Neurological or degenerative diseases	4	7	9	7	6	6	6
	Psychiatric illnesses	5	5	5	4	4	4	1
	Psychotropic treatments	3	3	2	2	2	3	3
	Vascular risk treatments	1	1	1	1	1	1	2

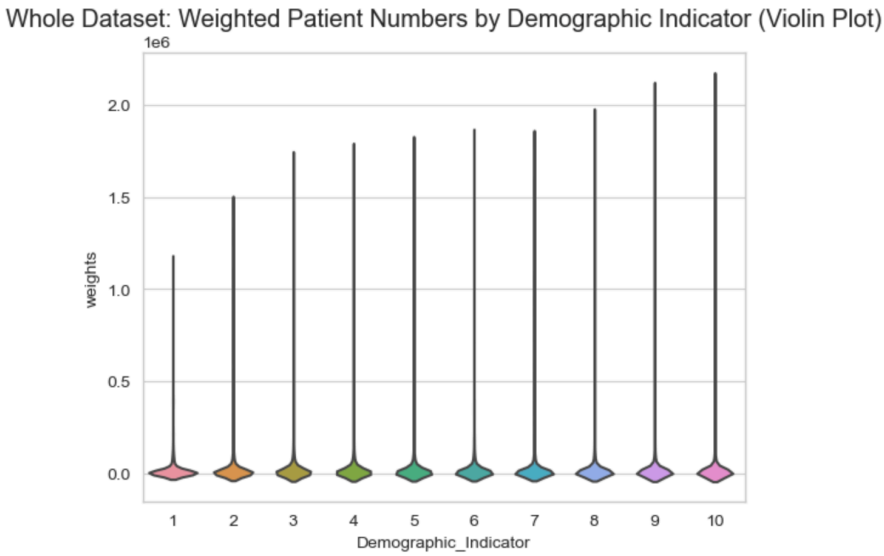
1	Agriculteurs exploitants	Farmers
2	Artisans, commerçants, chefs d'entreprise	Craftsmen, merchants, business owners
3	Cadres et professions intellectuelles supérieures	Executives and higher intellectual professions
4	Professions intermédiaires	Intermediate professions
5	Employés	Employees
6	Ouvriers	Workers
7	Retraités	Retired
8	Autres	Others

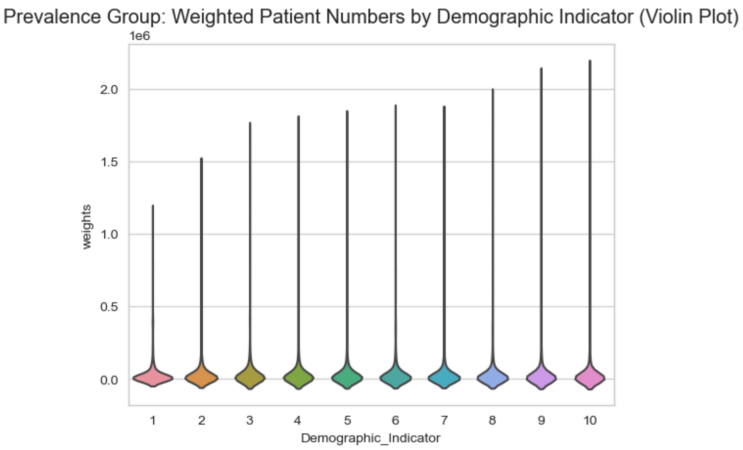
Farmers, craftsmen, merchants, business owners, and workers in France exhibit the highest weighted number of new cases for Cardio-neurovascular diseases. This finding sheds light on the prevalence of such diseases within these occupational groups. In terms of the overall population distribution, employees constitute 31%, intermediate professions make up 29%, executives and higher intellectual professions account for 18%, and farmers represent 2% of the population.

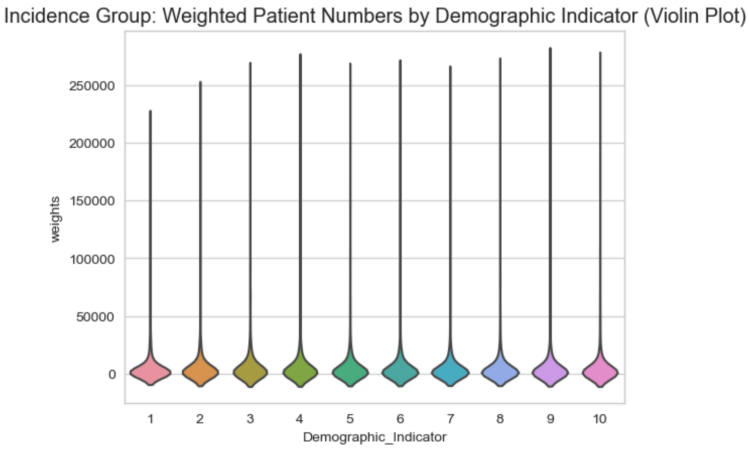
Interestingly, both employees and executives/higher intellectual professions share the highest number of new cases in the category of psychotropic treatments. However, when it comes to Neurological or degenerative diseases, their rankings differ. Executives and higher intellectual professions are placed at the 9th rank among the 11 diseases, while employees occupy the 6th rank. This discrepancy suggests variations in the prevalence and impact of Neurological or degenerative diseases within these professional categories.

These observations highlight the complex interplay between societal factors, specific occupations, and the occurrence of different diseases in France. Understanding these relationships can contribute to targeted interventions and policies aimed at addressing health disparities and promoting well-being within different segments of society.

3.4.4. Violin charts







These are the violin plots provides a visual representation of the distribution of the weighted number of patients across different categories of demographic information. Each category is represented by a "violin" shape, which displays the density or frequency of the data.

In the context of our analysis, the demographic information refers to variables such as age, gender, education level, or social class. By plotting the weighted number of patients on the y-axis and the demographic information on the x-axis, we can observe how the distribution of patients varies across different categories.

The width of the violin represents the density or frequency of patients within each category. A wider section indicates a higher concentration of patients, while a narrower section suggests a lower concentration. The shape of the violin provides insights into the distribution of the data, such as whether it is symmetric, skewed, or multimodal.

Analysing the violin plot allows us to identify patterns and trends in the distribution of patients based on demographic characteristics. We can observe which categories have higher or

lower concentrations of patients, as well as the variability within each category. Additionally, by comparing the violins of different demographic groups, we can assess if there are significant differences in the weighted number of patients across these groups.

Overall, the violin plot helps us visually understand the relationship between demographic information and the weighted number of patients, providing valuable insights into how different demographic factors may influence disease prevalence or healthcare utilisation.

3.5. Standardised Rate with the Direct and Indirect Method

In our data, there are columns named “txStandDir” (Standardised Rate with the Direct Method) and “txStandIndir” (Standardised Rate with the Indirect Method) which are the important techniques used to compare the observed and expected outcomes of a population or a group in the healthcare industry. These methods are often employed in epidemiological studies or healthcare research to account for differences in demographic or clinical characteristics between populations or groups.

Direct Standardisation is a method that adjusts for differences in the age structure of two or more populations or groups. It allows for a fair comparison of health outcomes by removing the effect of age, which can significantly influence disease rates or mortality rates. For example young people are relatively poorer than others. However, they are less often diabetic, so it could therefore be believed that poverty reduces the risk of diabetes. Here, the rates of incidence and prevalence have been standardised by age (10-year categories) and by sex, with a direct standardisation method.

Indirect Standardisation is a technique used to compare the observed and expected outcomes of a population or group by adjusting for differences in a specific variable, such as sex or socioeconomic status. It helps identify whether a population or group has higher or lower rates of a particular outcome compared to the expected rates based on a reference population. In this data, the rates of incidence and prevalence have been standardised by profession, education and social class with a indirect standardisation method.

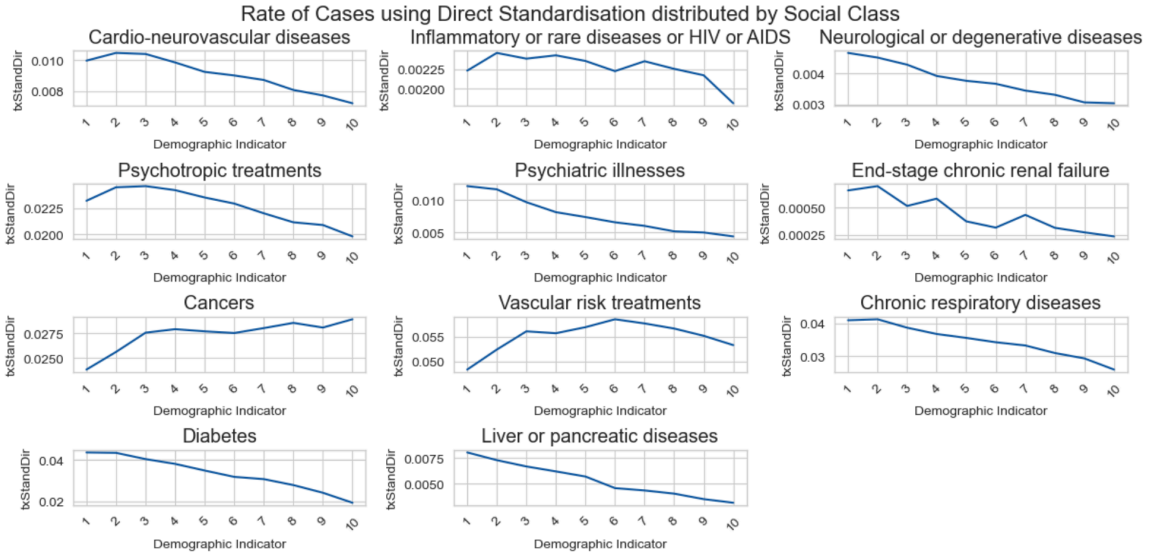
I’m making 2 charts to compare the Rate of Cases using Direct Standardisation and Rate of Cases using Indirect Standardisation distributed by Social Class to see if there is the significant difference of proportion of patients between these 2 methods.

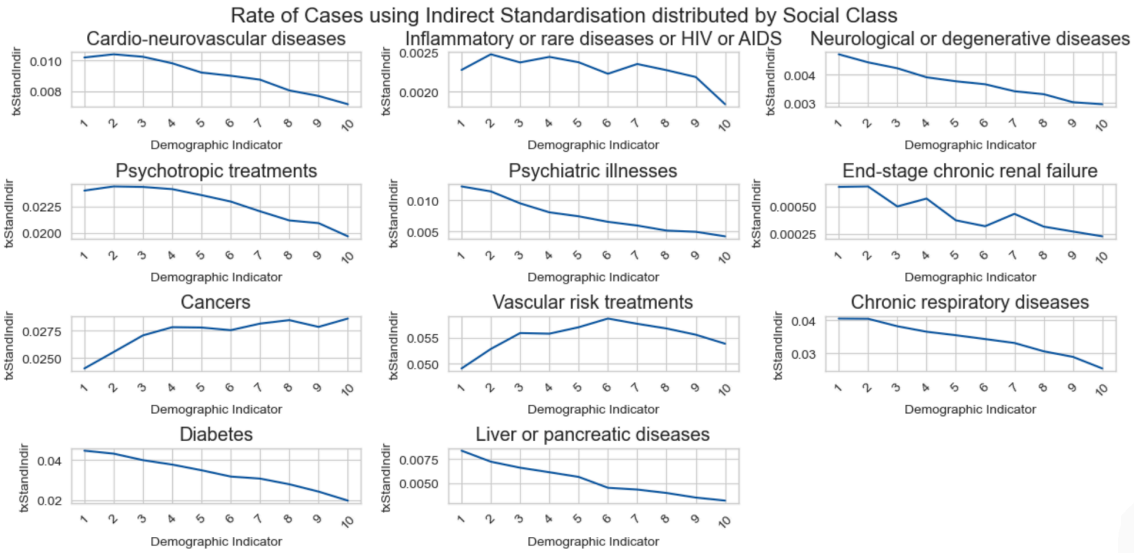
Among the 11 disease groups analysed, there are 4 groups that exhibit different rates when comparing the Direct and Indirect standardisation methods. These groups are Cardio-neurovascular diseases, psychotropic treatments, Inflammatory or rare diseases or HIV or AIDS, and End-stage chronic renal failure. When applying Direct Standardisation, which takes into account the effects of age and sex, the rate of patients in the aforementioned disease groups is observed to be lower compared to the Indirect Standardisation method. This suggests that after adjusting for age and sex, the prevalence or incidence of these diseases appears to be lower than when age and sex are not taken into consideration. Furthermore, among the four disease groups

mentioned, there is also a difference in rates specifically for the poorest population. This implies that within the poorest population segment, the Direct Standardisation method further highlights a lower rate of patients with Cardio-neurovascular diseases, psychotropic treatments, Inflammatory or rare diseases or HIV or AIDS compared to Indirect Standardisation.

These findings underscore the influence of age and gender on the occurrence of various diseases, as well as the significant impact of social class on disease prevalence. Age and gender emerge as influential factors, indicating that certain diseases may have a higher propensity to manifest in specific age groups or be more prevalent among males or females.

Furthermore, the role of social class becomes evident, as it demonstrates a significant association with the majority of diseases studied. Social class encompasses various socioeconomic factors, including education, income, and occupation, which can contribute to disparities in health outcomes. The data suggest that individuals from different social classes may face distinct risks and vulnerabilities to certain diseases.



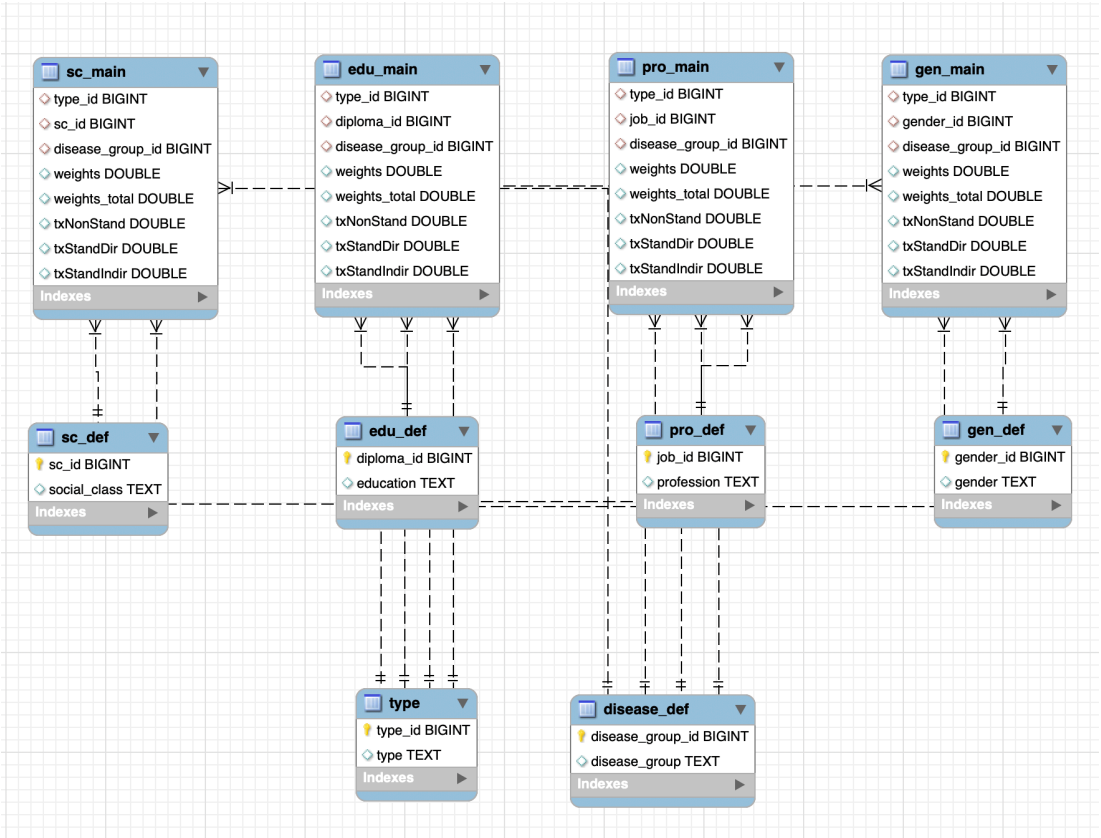


4. SQL

4.1. Choose the Database type.

I choose MySQL for its Versatility and Compatibility. It is highly versatile and compatible with different operating systems such as Windows, Linux, macOS, and various programming languages. This flexibility allows you to integrate MySQL seamlessly into your existing technology stack. MySQL has a rich ecosystem of tools, frameworks, and libraries that can enhance your development process. Other than that, MySQL offers efficient data storage and retrieval mechanisms. It can handle large amounts of data and high concurrent user connections. Additionally, MySQL supports advanced indexing and caching techniques, enabling faster query execution.

4.2. Entity-Relationship Diagram (ERD)



4.3. Queries.

Query 1. Top 10 highest weighted number of new cases (incidence) for the poorest population.

```
# Top 10 highest weighted number of new cases (incidence) for the poorest population
select sc_def.social_class, sc_main.type_id as disease_type, disease_def.disease_group as diease_name, sc_main.weights as highest_weights
from chronic_diseases.sc_main
left join chronic_diseases.disease_def on sc_main.disease_group_id=disease_def.disease_group_id
left join chronic_diseases.sc_def on sc_main.sc_id = sc_def.sc_id
where sc_main.type_id = 2 and sc_main.sc_id = 1
order by sc_main.weights desc limit 10;
```

social_class	disease_type	diease_name	highest_weights
10th decile (least affluent segment)	2	Psychotropic treatments	218766.553084868
10th decile (least affluent segment)	2	Chronic respiratory diseases	213565.92720769
10th decile (least affluent segment)	2	Psychotropic treatments	163763.34822732
10th decile (least affluent segment)	2	Vascular risk treatments	159227.810961478
10th decile (least affluent segment)	2	Vascular risk treatments	139291.372545082
10th decile (least affluent segment)	2	Psychotropic treatments	135866.886315331
10th decile (least affluent segment)	2	Cardio-neurovascular diseases	122229.751244613
10th decile (least affluent segment)	2	Chronic respiratory diseases	115861.708246178
10th decile (least affluent segment)	2	Psychiatric illnesses	109733.440878773
10th decile (least affluent segment)	2	Psychotropic treatments	105827.363807957

Query 2. Top 10 highest weighted number of new cases (incidence) for the wealthiest population.

```
# Top 10 highest weighted number of incidence cases for the wealthiest population
select sc_def.social_class, sc_main.type_id as disease_type, disease_def.disease_group as diease_name, sc_main.weights as highest_weights
from chronic_diseases.sc_main
left join chronic_diseases.disease_def on sc_main.disease_group_id=disease_def.disease_group_id
left join chronic_diseases.sc_def on sc_main.sc_id = sc_def.sc_id
where sc_main.type_id = 2 and sc_main.sc_id = 10
order by sc_main.weights desc limit 10;
```

social_class	disease_type	diease_name	highest_weights
10th decile (most affluent segment)	2	Psychotropic treatments	267585.066322509
10th decile (most affluent segment)	2	Vascular risk treatments	207681.435790414
10th decile (most affluent segment)	2	Vascular risk treatments	179934.80629602
10th decile (most affluent segment)	2	Psychotropic treatments	174320.920325922
10th decile (most affluent segment)	2	Cardio-neurovascular diseases	165503.285047724
10th decile (most affluent segment)	2	Chronic respiratory diseases	163958.877236791
10th decile (most affluent segment)	2	Psychotropic treatments	158532.086183803
10th decile (most affluent segment)	2	Psychotropic treatments	154027.913491984
10th decile (most affluent segment)	2	Cancers	128193.395339375
10th decile (most affluent segment)	2	Psychotropic treatments	109052.980138706

Query 3. How many weighted number of patients each disease group has at each profession

```
19  # How many weighted number of patients each disease group has at each profession.
20  select dd.disease_group_id as "disease_id", dd.disease_group as "disease_group", pd.profession as "profession",
21  sum(pm.weights) as "Number of weighted number"
22  from chronic_diseases.disease_def dd
23  inner join chronic_diseases.pro_main pm
24      on dd.disease_group_id = pm.disease_group_id
25  inner join chronic_diseases.pro_def pd
26      on pm.job_id = pd.job_id
27  group by dd.disease_group_id, pd.profession
28  order by dd.disease_group_id asc;
```

38:16

Result Grid Filter Rows: Search Export:

disease_id	disease_group	profession	Number of weighted num...
1	Cardio-neurovascular diseases	Craftsmen, merchants, business owners	5143335.269752712
1	Cardio-neurovascular diseases	Employees	13918568.967766207
1	Cardio-neurovascular diseases	Executives and higher intellectual professions	5096274.879106675
1	Cardio-neurovascular diseases	Farmers	3399447.060048797
1	Cardio-neurovascular diseases	Intermediate professions	9340338.109076591
1	Cardio-neurovascular diseases	Others	4185668.6685799346
1	Cardio-neurovascular diseases	Workers	15960150.029456185
2	Inflammatory or rare diseases or HIV or AIDS	Craftsmen, merchants, business owners	615095.824984554
2	Inflammatory or rare diseases or HIV or AIDS	Employees	2970139.000938609
2	Inflammatory or rare diseases or HIV or AIDS	Executives and higher intellectual professions	980610.0714420708
2	Inflammatory or rare diseases or HIV or AIDS	Farmers	282767.3331271705
2	Inflammatory or rare diseases or HIV or AIDS	Intermediate professions	1963431.7828026263
2	Inflammatory or rare diseases or HIV or AIDS	Others	1722062.6232495978
2	Inflammatory or rare diseases or HIV or AIDS	Workers	1949805.103500285
3	Neurological or degenerative diseases	Craftsmen, merchants, business owners	1050995.8857681814
3	Neurological or degenerative diseases	Employees	4325366.733803928
3	Neurological or degenerative diseases	Executives and higher intellectual professions	1033596.6659461475
3	Neurological or degenerative diseases	Farmers	900935.541349854
3	Neurological or degenerative diseases	Intermediate professions	2311327.475184918
3	Neurological or degenerative diseases	Others	2904271.314835079
3	Neurological or degenerative diseases	Workers	3491694.606244939
4	Psychotropic treatments	Craftsmen, merchants, business owners	3914701.828916698
4	Psychotropic treatments	Employees	20411848.95241939

Query 4. The average weighted number of patients of each disease at each education group.

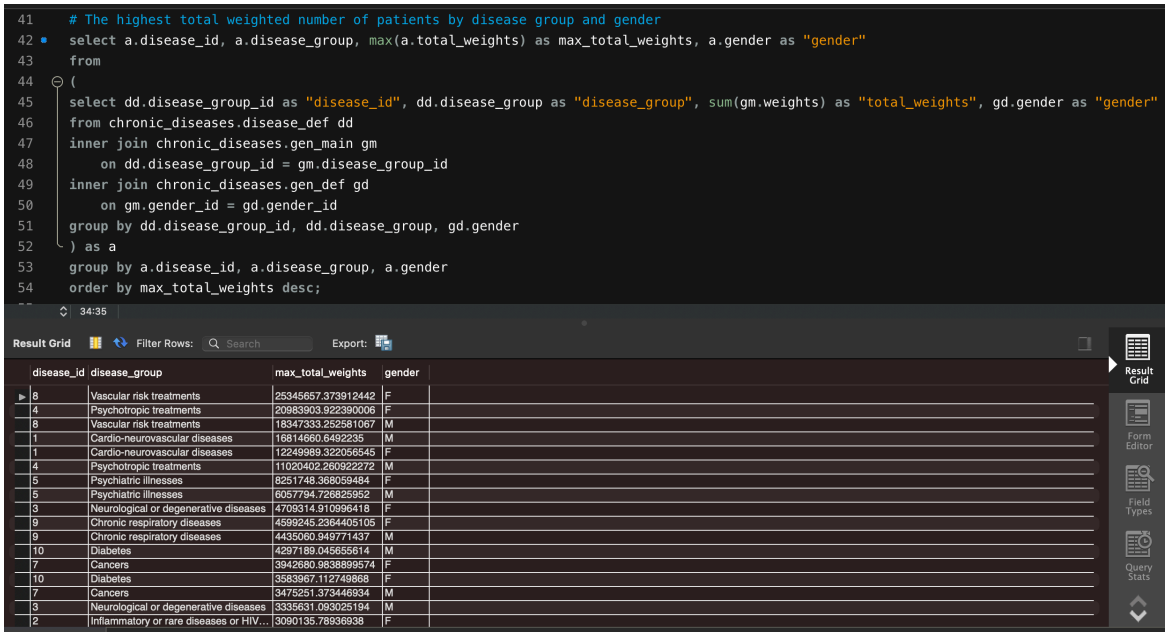
```
30  # The average weighted number of patients of each disease at each education group.
31  select dd.disease_group_id as "disease_id", dd.disease_group as "disease_group", ed.education as "education",
32  avg(em.weights) as "average_weights"
33  from chronic_diseases.disease_def dd
34  inner join chronic_diseases.edu_main em
35      on dd.disease_group_id = em.disease_group_id
36  inner join chronic_diseases.edu_def ed
37      on em.diploma_id = ed.diploma_id
38  group by dd.disease_group_id, dd.disease_group, ed.education
39  order by dd.disease_group asc;
```

76:42

Result Grid Filter Rows: Search Export:

disease_id	disease_group	education	average_weights
7	Cancers	Higher education	506755.8562264973
7	Cancers	High school dgree	309421.17751023255
7	Cancers	Professional degree	785693.8244984302
7	Cancers	No degree	808122.2953587114
1	Cardio-neurovascular diseases	No degree	359188.06427971163
1	Cardio-neurovascular diseases	Higher education	105869.62296305157
1	Cardio-neurovascular diseases	High school dgree	79681.08673818524
1	Cardio-neurovascular diseases	Professional degree	243080.516863615
9	Chronic respiratory diseases	High school dgree	335942.55222853326
9	Chronic respiratory diseases	Higher education	455634.5856764645
9	Chronic respiratory diseases	Professional degree	803351.2575363609
9	Chronic respiratory diseases	No degree	827820.2048055949
10	Diabetes	Higher education	340355.51058952766
10	Diabetes	No degree	1096592.0862263225
10	Diabetes	Professional degree	830403.382762995
10	Diabetes	High school dgree	253429.70705653736
6	End-stage chronic renal failure	High school dgree	3771.672723486887
6	End-stage chronic renal failure	Higher education	4991.950767167749
6	End-stage chronic renal failure	Professional degree	10472.557686542472
6	End-stage chronic renal failure	No degree	13978.666882120284
2	Inflammatory or rare diseases...	No degree	52554.5010033112
2	Inflammatory or rare diseases...	Professional degree	63468.05702918372
2	Inflammatory or rare diseases...	High school dgree	29176.64218795123

Query 5. The highest total weighted number of patients by disease group and gender (order by total weighted number descending).



```sql
# The highest total weighted number of patients by disease group and gender
select a.disease_id, a.disease_group, max(a.total_weights) as max_total_weights, a.gender as "gender"
from
(
select dd.disease_group_id as "disease_id", dd.disease_group as "disease_group", sum(gm.weights) as "total_weights", gd.gender as "gender"
from chronic_diseases.disease_def dd
inner join chronic_diseases.gen_main gm
    on dd.disease_group_id = gm.disease_group_id
inner join chronic_diseases.gen_def gd
    on gm.gender_id = gd.gender_id
group by dd.disease_group_id, dd.disease_group, gd.gender
) as a
group by a.disease_id, a.disease_group, a.gender
order by max_total_weights desc;
```

34:35

Result Grid | Filter Rows: Search | Export:

disease_id	disease_group	max_total_weights	gender
8	Vascular risk treatments	25345657.373912442	F
4	Psychotropic treatments	20983903.922390006	F
8	Vascular risk treatments	18347333.252581067	M
1	Cardio-neurovascular diseases	16814660.6492235	M
1	Cardio-neurovascular diseases	12249989.322056545	F
4	Psychotropic treatments	11020402.260922272	M
5	Psychiatric illnesses	8251748.368059484	F
5	Psychiatric illnesses	6057794.726825952	M
3	Neurological or degenerative diseases	4709314.910996418	F
9	Chronic respiratory diseases	4599245.2364405105	F
9	Chronic respiratory diseases	4435060.949771437	M
10	Diabetes	4297189.045655614	M
7	Cancers	3942680.9838899574	F
10	Diabetes	3583967.112749868	F
7	Cancers	3475251.373446934	M
3	Neurological or degenerative diseases	3335631.093025194	M
2	Inflammatory or rare diseases or HIV...	3090135.78936938	F

Result Grid | Form Editor | Field Types | Query Stats

5. Conclusion

In conclusion, our analysis of social inequalities in chronic diseases in France has revealed several important findings.

Firstly, we observed that certain disease groups have a higher prevalence among specific social classes. This suggests that social class plays a significant role in determining the risk and burden of chronic diseases. For example, Cardio-neurovascular diseases were more prevalent among individuals in lower social classes, while Psychotropic treatments had a higher prevalence among individuals in higher social classes.

Furthermore, our analysis highlighted the impact of demographic factors, such as age and gender, on the distribution of chronic diseases. Age was found to be a significant factor, with certain diseases exhibiting higher prevalence among specific age groups. Similarly, we observed differences in disease prevalence between genders, emphasising the influence of gender on disease patterns.

The findings also shed light on the importance of considering the weighted number of patients when analysing social inequalities in chronic diseases. By accounting for the weight of each patient, we gained a more comprehensive understanding of the disease burden across different social classes and demographic groups.

Overall, our analysis underscores the existence of social inequalities in chronic diseases in France. The findings highlight the need for targeted interventions and policies to address these inequalities and promote equitable access to healthcare services. By addressing social determinants of health and improving healthcare accessibility and affordability, we can strive towards reducing the burden of chronic diseases and promoting better health outcomes for all segments of society.

6. Recommendations

This part about Machine Leaning will be implemented with the presentation. In the context of my data analysis, I will be incorporating Unsupervised Machine Learning techniques. Unsupervised Machine Learning is a branch of Artificial Intelligence that enables the exploration and identification of patterns or structures within data without any predefined labels or target variables.

By employing Unsupervised Machine Learning algorithms, I aim to uncover hidden insights, relationships, or groupings within my data. These techniques are particularly useful when working with large and complex datasets, as they can assist in revealing meaningful patterns that may not be immediately apparent through manual analysis.