ANALYSIS OF CANCER LEVEL PATIENTS DATA SET

§ Afsar Nehan, * Alharthy Raiya, ¶ Dimos Konstadimos, † Hoyos Harrison, ‡ Sunilkumar Sidharth

§nehaafsar05@gmail.com, *raiyaalharthy1992@gmail.com, ¶konstadimos7@gmail.com,
†harrihoyos2680@gmail.com, ‡sidharthsunil74@gmail.com
Brainnest Data Analysis Trainee Program

Abstract

In 2019 the second most common cause of death worldwide was cancer just exceeded by cardiovascular diseases [1]. Cancer is an uncontrolled process in the division of the body's cells [2]. The national cancer institute defines how a tumor grade describes a level of cancer as low, medium, or high from a measure of 1 to 4. This report describes the procedure and conclusions to answer the following questions: What is the relationship between the features in the data set and the level of cancer? What are the most meaningful variables that define a level of cancer condition? What are the most significant insights into these variables? Could it be explained in numbers what the impact of the variables chosen in the cancer level is?
At first, it shows how the authors deal with missing values and outliers in the data set, then it is made a meticulous analysis to find the answers to the research proposed, and lastly defines a methodology to get predictions from new data to give a diagnosis in the level of cancer of patient having some features known with the objective to answer the last research question: Having these meaningful variables chosen could cancer levels be predicted for new patients from them?. In the meanwhile of the analysis, some relevant facts and clear misunderstandings have been found and presented in the conclusions.
keywords: cancer level, insights, predictions

I. Introduction

The goal of this report is purely educational, since attempting to analyze how different health symptoms and lifestyle choices or life conditions affect the risk of developing lung cancer could be quite a challenging task for amateur researchers. Nevertheless, the authors of this report have worked with a dataset of n = 1000 patients in their attempt to give a satisfactory answer to this question, with the age of the participants ranging from 14 years minimum to 73 years maximum and with the male patients amounting a 60.2% of the sample and the females amounting a 39.8% of the sample.
Within the dataset, there are 21 potentially decisive factors, split into 4 different categories:

1) Health Symptoms (i.e. Coughing of Blood, Wheezing, Dry Cough, Chest Pain, Fatigue, Swallowing Difficulty, Clubbing of finger Nails, Shortness of Breath, Frequent Cold, Weight Loss, Snoring)
2) Long-term Health Issues (i.e. Chronic Lung Disease, Obesity, Dust Allergy)
3) Lifestyle Factors (i.e. Alcohol Use, Smoking)
4) Environmental Factors (i.e. Air Pollution, Passive Smoking, Occupational Hazard)

Cancer patients have several factors in common that determine the result of the diagnosis of cancer level, these conditions have been tracked in different ways from diverse sources such as the data set (⬇) to be used in this analysis. This data source has one thousand registers of patients with cancer and their attributes that have been tracked (columns/features) such as Patient ID, Age from a range of 14 to 73 years old, Gender (1 to male, 2 to female), air pollution, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung diseases, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, swallowing difficulty, clubbing of fingernails, frequent cold, dry cold, snoring, all these 21 features are in a range from 1 to 9, the last attribute is the level of cancer of the patients that could be low, medium or high as is explained in detail by the national cancer institute cancer [3] have grades to describe the situation of a patient:

- Grade X: Grade cannot be assessed (undetermined grade)
- Grade 1: Well differentiated (low grade)
- Grade 2: Moderately differentiated (intermediate grade)
- Grade 3: Poorly differentiated (high grade)
- Grade 4: Undifferentiated (high grade)

The data as it is shown by itself is not completed, it is necessary at first to reduce as possible the factors that could impact the accuracy of the investigation results. Therefore cleaning the data is crucial and is explained how it been tackled in detail in the next section.
Some questions proposed by the authors give the direction of the research to find the most meaningful insights on the data given, the chosen ones are considered to be more qualified to make this purpose a reality:

1) What is the relationship between the features in the data set and the level of cancer?
   The feature target to analyze is the cancer level due to it shows the condition by itself that defines the diagnosis of the patients. Therefore comparing it with the other variables is meaningful to describe their relevant insights that could help understand the diagnosis

2) What are the most meaningful variables that define a level of cancer condition?
   Not all of the variables are important to describe the cancer-level condition and also some of them could bias the diagnosis. Then finding these meaningful variables that impact the results of the diagnosis could help to support treatment to the patients
3) What are the most significant insights in these variables?
   Find what these variables special to describe the diagnosis of the patients let's prove the hypothesis that as usual common people have such as, the age of the patient is something important to consider in the diagnosis because the older the patient more likely to have a high level of cancer, the people that smoke and consume alcohol also are going to have a result more likely to be high. The main idea is to see if these thoughts are right and find more insights than the said ones.
4) Having these meaningful variables chosen could cancer levels be predicted for new patients from them?
   chose groups divided just by the relevant features found and train a model from them is analyzed to see the accuracy of predictions that could be made for new patients without knowing the grade of their cancer

II. Cleaning Data

1) Missing values
   The data by itself shows different cells without information in the .xlsx file. Different approaches could be taken into account to deal with this problem. First, the rows (entities/patients) that have at least one column without information can be drooped, but this solution makes the data set lose important information. Another solution could be to fill the empty cells with a value and then when statistics came to the scene, the missing values could be the mean or the median of the feature analyzed, for example, if there is not a value of age for the patient with the ID P121 this cell could be filled by the mean or median of the patient's age; but how to decide to use mean or median? the answer is to analyze the frequency distribution of the feature, this could be proved in several ways and one analyzes the bar frequency chart, if it resembles a normal distribution then the best option is to replace it with the mean, but if isn't the case the median will be more optional. Following the example, the age histogram looks like the figure bellow:

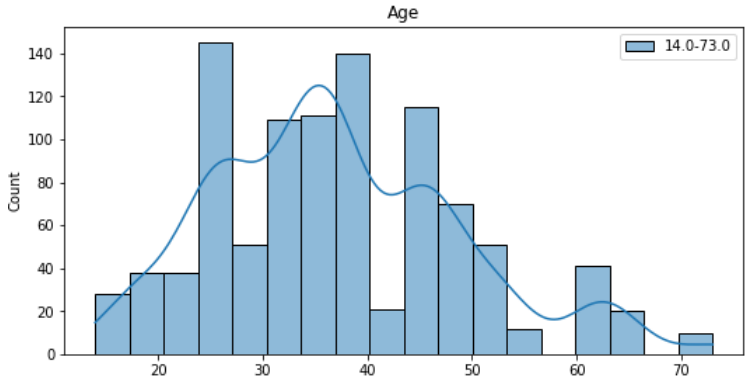


Fig. 1: Histogram of the age feature in the cancer data set

   The figure 1 shows that the distribution of the age is not normal so for the patient P121 is better to add a median age of the total patients. This process is automated using python, the code is shared here
2) Outliers:
   Another advantage to work with histograms is that let detect the outliers too, for example the figure **??** shows a detected outlier in dust allergy feature.

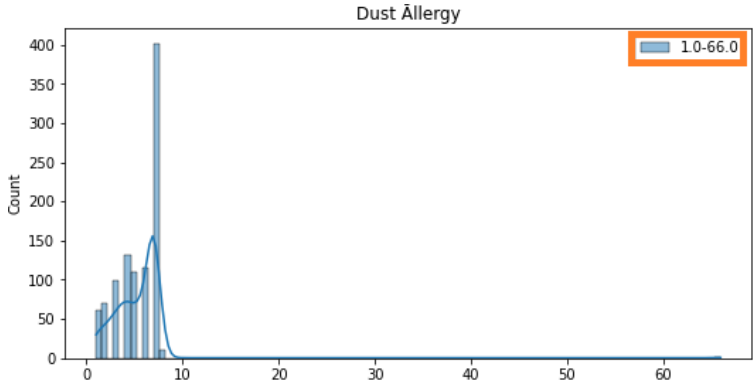


Fig. 2: Detected outlier in dust allergy feature on the cancer data set

The outliers were changed to the max value of the feature, another solution that also will help was change it to the median value or as has been discover, the outliers are numbers that were written wrong and that are expected with a point in the middle of two amounts, for example the outlier of the dust allergy is 66, the correct number that had to be written must be 6.6.

III. Data Analysis

The already cleaned data help to understand the relationship between features that could be meaningful in the diagnose of cancer level of a patient. To begin, the following question is answered:

1) **What is the relationship between the features in the data set and the level of cancer?**
This relationship could be shown in different ways, one that would be representative and meaningful in a visual way is a boxplot between the two variables analysed. It helps to visualize the mean, median and standard deviation at once.
The results of the boxplot for all the variables in the dataset in relationship with the cancer level of the patients are shown

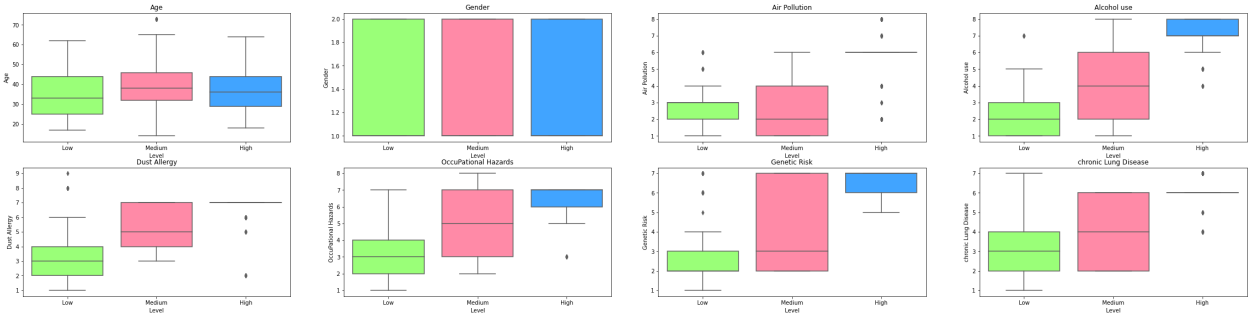


Fig. 3: Boxplot comparing the level of cancer of patients with the Age, Gender, Air pollution, Alcohol use, Dust Allergy, Occupational Hazards, Genetic risk and Chronic lung disease

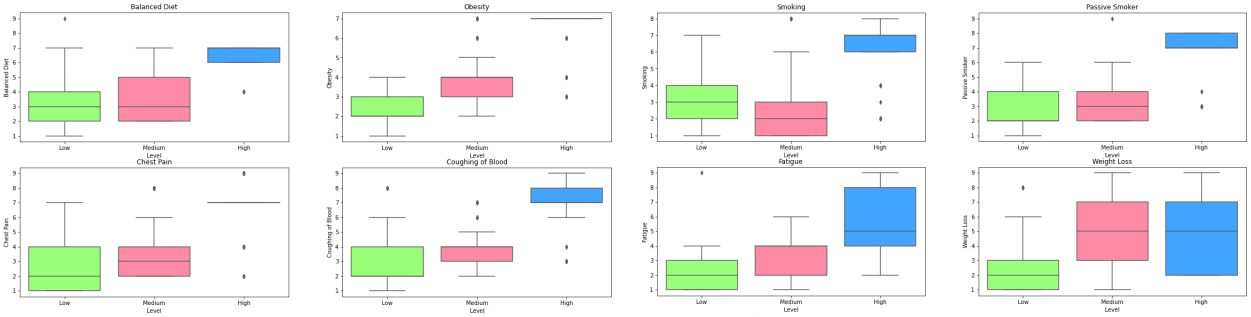


Fig. 4: Boxplot comparing the level of cancer of patients with the Balance diet, Obesity, Smoking, Passive smoker, Chest pain, Coughing blood, Fatigue and Weight loss

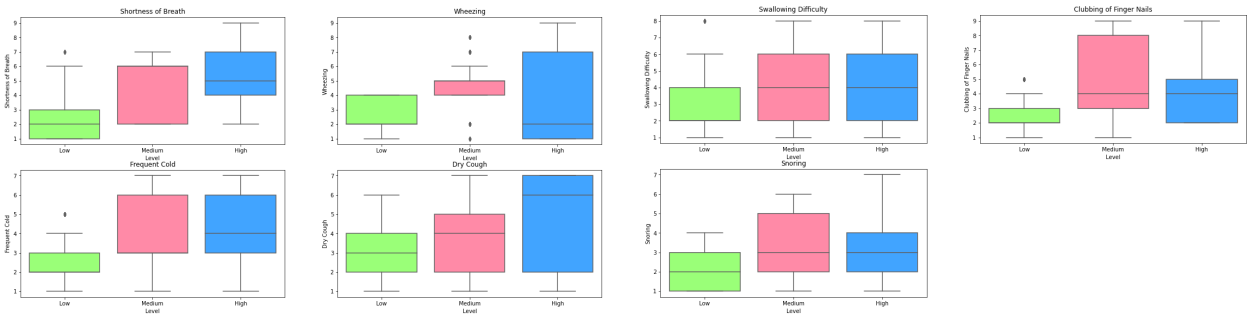


Fig. 5: Boxplot comparing the level of cancer of patients with the Shortness of breath, Whezzing, Swallowing difficulty, Clubbing of finger nails, Frequent cold and Dry cough

Having this relationships shown on figures 3, 4 and 5. The group chose four questions to be answered:

a) **What are the most meaningful variables that define a level of cancer condition?**
To analyze the information shown in the boxplots there are some factors that have to be into account to make a decision in how to chose these meaningful variables without still having more information about them. The correlation between how a variable grow and affects directly on the cancer level is the most relevant condition that the team chose as meaningful. The shape of the boxplot also is one condition that is considered due it show the distribution of data and let discard possible variables that are not so relevant to describe the diagnose of the patient. The variables considered are the following:

i) **Alcohol use**
ii) **Obesity**
iii) **Occupational hazards**
iv) **Gender**
v) **Age**

There is a strong relationship between the increment of these variables and the level of the diagnosis for the three first cases chosen, in terms of how the shape of the boxplot in these variables is distributed is in a way that have low variance in the majority of the cases.

b) **What are most significant insights in these variables?**
Alcohol use

Fig. 6: The figure shows a Power BI dashboard screenshot having into account the alcohol use level compare to the level of cancer. 1. boxplot of level of cancer compare to alcohol use 2. Number of patients, Average, STD, Variance and Median of alcohol use 3. Treemap of the count of alcohol use by level of cancer cases 4. Pie chart of the average alcohol use by level 5. Bar chart of Count of alcohol use cases organized by Gender 6. Histogram of alcohol use cases

The level of alcohol use is in the range from 1 to 8, the boxplot shown in the figure 6,1 demonstrate a correlation between the level of alcohol use and the cancer level of patients with a similar shape of normal distribution in the low and medium level of cancer cases, this is supported by the mean and median values of 4.58 and 5 respectively and that also show a high level of alcohol use by patients. An standard deviation of 2.6 show that also the distribution

is spread. for the sample of n = 1000 as figure 6,3 show there is a distribution of a 36.5%, 33.2%, 30.3% of high, medium and low level of cancer cases and in the case of alcohol use average shown in figure 6,4 there is a distribution of 6.82 (51.52%), 4.24 (31.81%) and 2.26 (16.99%) of high, medium and low level of cancer respectively and describe something meaningful that is **a high level of cancer patients has a huge consume of alcohol use near to 7 on average and represent more than a half of the patients**. The figure 6,5 describe that the 60.2% of the cases are male and the 39.8% are female.

What if the range of alcohol use change from the minimum value starting at the median?

Fig. 7: Power BI dashboard containing the same charts as figure 6 with a range of alcohol use level from 5 to 8, and with 533 patients being analyzed

Figure 7 show the case when the alcohol range change with the idea of making an analysis to demonstrate that as the level of alcohol use increase the impact on the level of cancer is direct. The mean of 6.82 and the median of 7 (7,2) are so close to each other that means the distribution is more likely to be normal, the standard deviation also show that in this range this distribution is less spread. There are now fewer patients (533) on this sample but the number of cases count by the level of cancer (7,3) have changed to be now distributed as 66.6%, 30.77%, 2.62% for high, medium and low cancer level cases respectively; that is an increment of 30.1% for the high level of the cases and a decrease of 2.43% and 27.68% for the medium and low level of cancer cases compare to the last dashboard (6,3) and means that from half of the alcohol use distribution (sample n=1000) is expected to have just high (more likely) and medium level of cancer cases. The figure 7,4 also has something to consider; as team the first though was that the expectancy for low cases on average would be the less value on the range of alcohol use, but instead it shows something more meaningful, the distribution of the pie chart is 6.9 (34.7%), 6.7 (33.69%) and 6.29 (31.61%) for high, medium and low cancer level respectively. That means first that this few amount of low cancer level patients (14) don't have the less alcohol use in the range but instead have a high one but it doesn't mean something important for the sample. **The average on alcohol use in this range change just 0.2% from high to medium and say that at least a low change on the habits to patients with a high consume of alcohol impact significantly on the cancer level**

Obesity

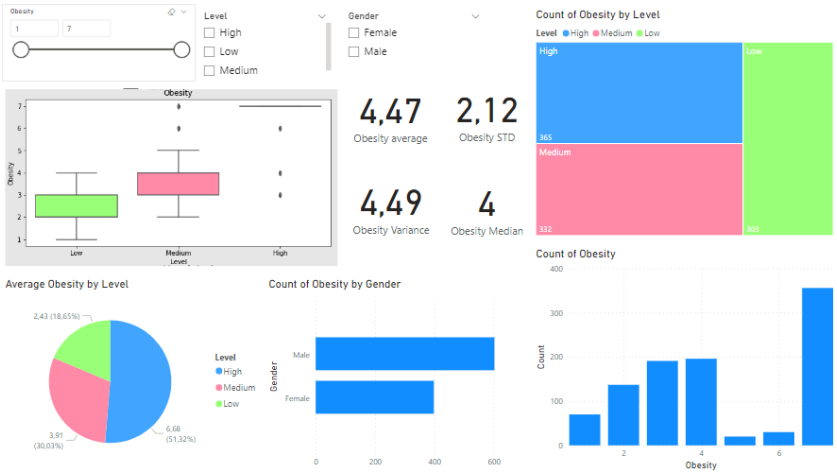


Fig. 8: The figure shows a Power BI dashboard screenshot having into account the obesity level compare to the level of cancer and the age of the patients in the data set. 1. boxplot of level of cancer compare to obesity 2. Average, STD, Variance and Median of obesity 3. Treemap of the count of obesity by level of cancer cases 4. Pie chart of the average obesity by level 5. Bar chart of Count of obesity cases organized by Gender 6. Histogram of Obesity cases

This human body condition show to be a common pattern that has to be directly with the cancer level of the patients. The most significant sign shown by the boxplot (8,1) is this direct relationship. The patients analyzed here have a range from 1 to 8 to measure the obesity level. The median of 4 and mean 4.47 (8,2) of this variable is quite similar one to other and let know why the distribution seams as shown in histogram (8,6). A standard deviation of 2.12 (8,2) is a high number compare to the range that is being analyzed.
The count of the cases of obesity by level of cancer (8,3) demonstrate that the mayority of the cases are high level follow by the medium and then the low. An average of obesity by level (8,4) demonstrate a 6.68 (51.32%), 3.91 (30.03%), 2,43 (18.65%) of high, medium and low respectively as the pie chart shows.

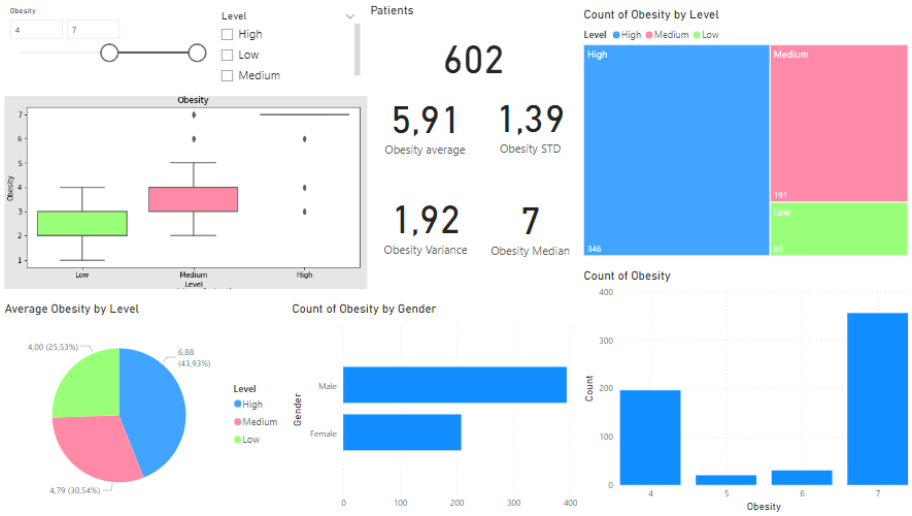


Fig. 9: Power BI dashboard containing the same charts as figure 8 with a range of obesity level from 4 to 8

In figure 9 the range of obesity change having the minimum value as 4 that is the median value of the complete dataset as figure 8 show. The main reason is to show how is the behavior of level of cancer in patients from middle to top of the frequency obesity distribution. There are several factors to have in mind such as there are now less patients to consider in this case, the standard deviation has decreased and means that in this range is expect to have less variance on data distribution, the median is the maximum value of the obesity range for this case and is concluded the hypotesis expected that had been declared with the boxplot of figure 8,1 **as higher values of obesity level more cases of high level of cancer in patients**, this is also supported by the Treemap (9,3) that shows also that the majority of cases is high with an increase of 20.97% in this kind of cases in compare to the last dashboard. There is still something else to consider as meaningful to define obesity as a variable to consider important to determine the level of cancer of patients. The average obesity in figure 9, 4 support the last observations

too, due it shows that for high, medium and low cancer level there is an average of 6.88 (43.93%), 4.79 (30.54%), 4 (25,53%) of obesity respectively and what does it means is that the low cancer level cases have the minimum value of obesity and on average as expected, and for high value on average obesity support what had been said so far.

Occupational hazards

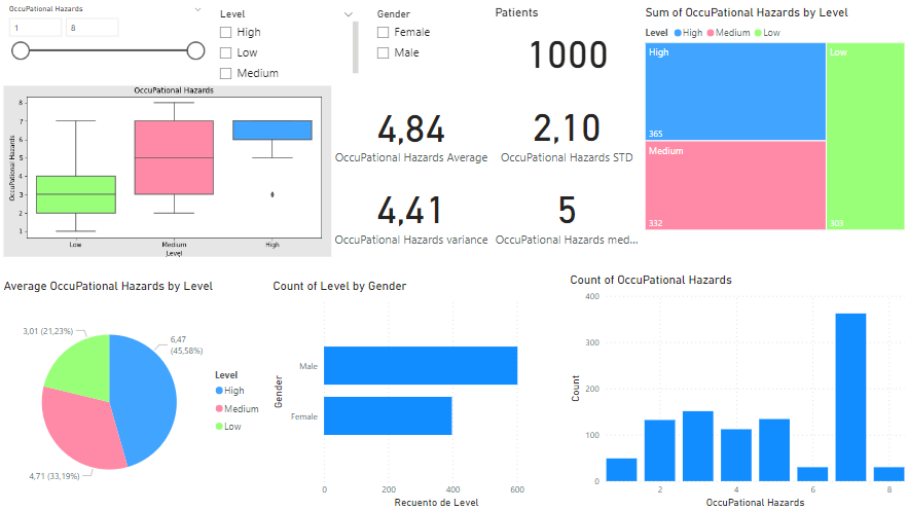


Fig. 10: The figure shows a Power BI dashboard screenshot having into account the occupational hazards level compare to the level of cancer. 1. boxplot of level of cancer compare to occupational hazards 2. Number of patients, Average, STD, Variance and Median of occupational hazards 3. Treemap of the count of occupational hazards by level of cancer cases 4. Pie chart of the average occupational hazards by level 5. Bar chart of Count of occupational hazards cases organized by Gender 6. Histogram of occupational hazards cases

The figure 10 show the whole data set analyzed in terms of occupational hazards and level of cancer of patients. This condition of occupational hazards is measured from a range of 1 to 8. The distribution of the boxplot (10,1) is spread for the medium level of cancer and is reflected in the standard deviation of 2.1 (10,2). The mean and median of 4.84 and 5 respectively shows that there is likely normal distribution that is shown on 10,6 biased by a huge amount of cases with a level in occupational hazards of 7. The same distribution of cases that were shown in figure 6,3 and 6,5 is the same for 10,3 and 10,5 respectively. Now the average occupational hazard by level (10,4) shows a distribution of 6.47 (45.58%), 4.71 (33,19%) and 3.01 (21.23%) for high, medium and low level of cancer actually showing that there is a strong relationship between how **the occupational hazards are a factor that impacts directly on cancer level**.

What happens when the range of occupational hazards change to be from the middle to the top of this level?

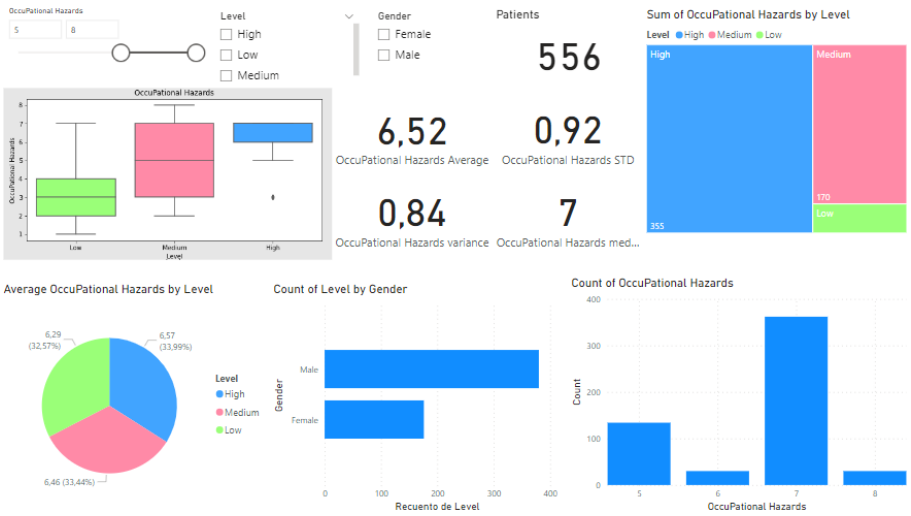


Fig. 11: Power BI dashboard containing the same charts as figure 10 with a range of occupational hazards level from 5 to 8 with 556 patients beign analized

Figure 11 show now that there are now fewer patients to analyze and indicates now that still there is a strong relationship in mean an median of 6,52 and 7 respectively. Also the standard deviation is pretty low therefore the data is less spread distributed. The count of cases by level of cancer also change to be 27.34% more for high level of cancer cases, 2.62% and 24.72% less for medium and low level of cancer cases respectively showing the same as alcohol use a strong relationship in how the occupational hazards affect the cancer level.

Age

Cancer is a disease that doesn't discriminate in terms of age, but, the anecdotal belief is that it affects older people more often than younger ones. So this notion would naturally move on in the case of the risk level of developing cancer things shouldn't be different.

In a sample of n = 1000 observations, the age of the patients in the dataset appears to approximate the normal distribution, with a mean of μ = 37.18 and a standard deviation σ = 11.98. The minimum age of patients is 14 and the maximum is 73 with a median value of 36. see figure 12

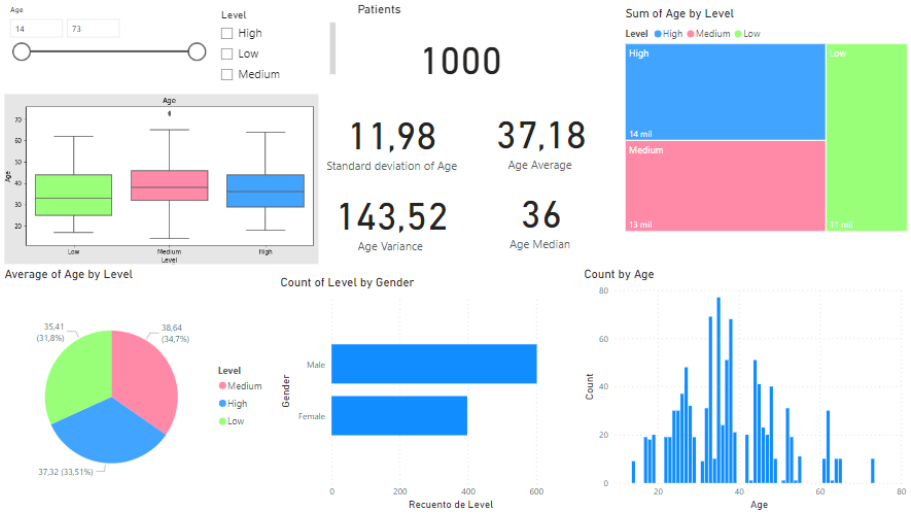


Fig. 12: The figure shows a Power BI dashboard screenshot having into account the age of patients compare to their level of cancer. 1. boxplot of level of cancer compare to age 2. Number of patients, Average, STD, Variance and Median of age 3. Treemap of the count of age by level of cancer cases 4. Pie chart of the average age by level 5. Bar chart of Count of age cases organized by Gender 6. Histogram of age cases

Having into account the gender of patients, the male patients of the analysis are distributed approximately normally with a mean μ = 39.18 and a standard deviation σ = 12.80 and they amount for 60.20% of the sample. The

maximum age of the male patients is 73 years and the minimum is 14 years, with a median age of 37. See figure 13

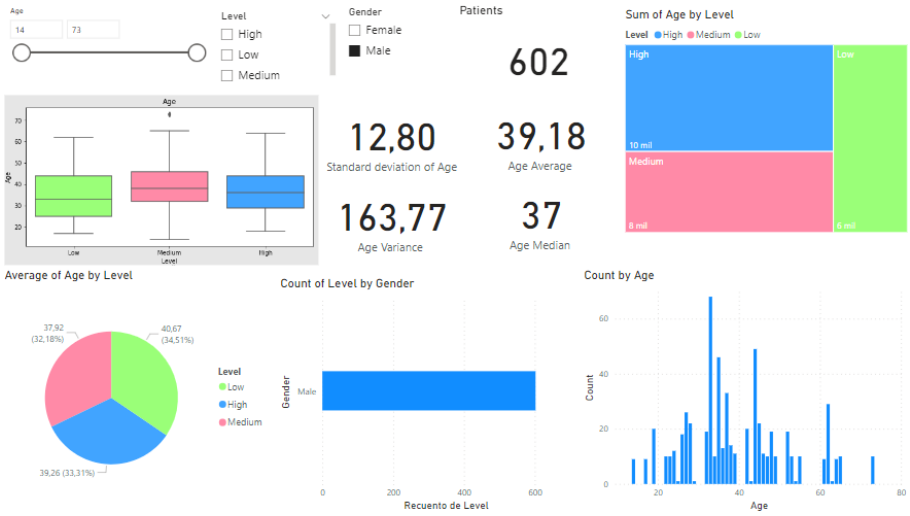


Fig. 13: Power BI dashboard containing the same charts as figure 12 having into account just the male gender with 602 patients being analyzed

The female part of the patients has a similar distribution with the male, with a mean age = 34.17 and a standard deviation = 9.89 and they amount for the 38.40% of the sample, with a maximum age of 64, minimum age of 17 and a median of 35. See figure 14

Fig. 14: Power BI dashboard containing the same charts as figure 12 having into account just the female gender with 398 patients being analyzed

Which ages and genders exhibit higher risk of developing cancer?
In figure 12 it is shown that the ages that have the greatest risk are not what have been initially expected. The way the sample is distributed again resembles the normal distribution with a mean value $\mu = 37.32$ years and a standard deviation $\sigma = 10.71$, so this sample refutes the notion that older ages posses a higher probability of developing tumors. The number of the patients that falls in this category is 365 (36,5 % of the sample), with 69.04% (i.e. 252) being male and the rest 30.96% (i.e. 113) being female.
How is the behavior of level of cancer in age range groups?:
There were 3 groups that have been selected:

- from the minimum value 14 to 30
- from 31 to 55
- from 56 to the maximum value 73

The first age group covers 300 patients (i.e.30%) of the sample, with 54%(i.e. 162) of them being male and 46%(i.e. 138) being female. The expectation for this age group is that the majority would exhibit a low risk of developing cancer. But still it can't be said that this expectation is met, as the 42.67% of the first age group(i.e. 128, 12.8% of the sample) exhibit low risk to develop cancer. The 34.67% of the first age group (i.e. 104, 10.4% of the sample) exhibit high risk to develop cancer and the remaining 22.67% of the first group(i.e. 68, 6.8% of the sample) has a medium risk to develop any form of cancer. see figure 15

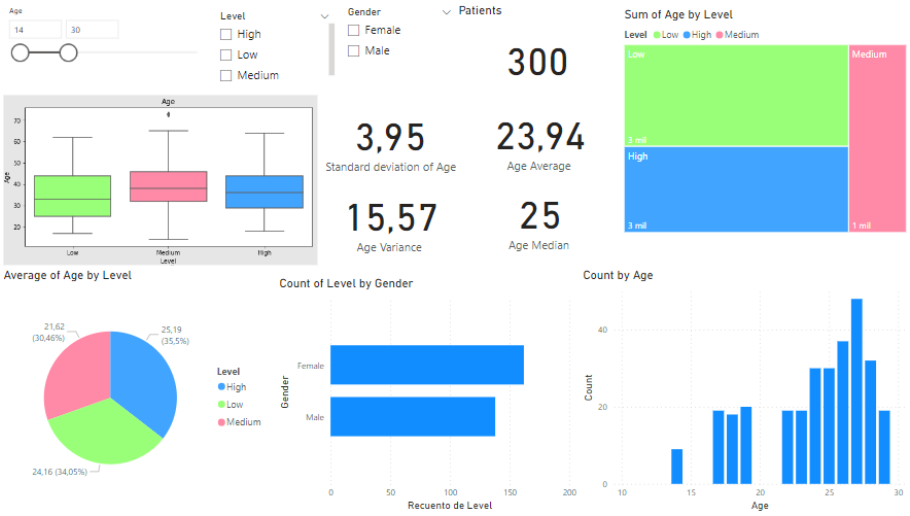


Fig. 15: Power BI dashboard containing the same charts as figure 12 changing the range of age from 14 to 30 with 300 patients being analyzed

The second age group (ages 31-55) is the most populous, as it is populated by 629 patients (i.e. 62.9% of the sample), with 62.96% of them (i.e. 396) being male and 37.04% (i.. 233) being female. The expectation for this age group is relatively balanced, or even shows a medium to lower inclination to exhibit cancer. The data say another story though, as 24.64% (i.e. 155, 15.5% of the sample) show low risk for developing cancer, the 37.2% of the second age group (i.e. 234, 23.4% of the sample) has a medium inclination to develop cancer and the biggest percentage, compared to the other two groups, with 38.16% (i.e. 240 patients and 24% of the sample) of them, showing higher tendency to develop cancer. See figure 16

Fig. 16: Power BI dashboard containing the same charts as figure 12 changing the range of age from 31 to 55 with 629 patients being analyzed

The third and last age group is the smallest counting 71 patients (7.1% of the sample). this age group is over represented by male as 95.77%(68) of them are male and merely 4.23%(3) are female. Given their age group, the team expected that the group would show higher risk levels of cancer. Instead, only 29.58%(21 patients, 2.1% of the sample) of the age group, are having a higher risk of developing cancer. Following that, the majority of the

age group, which reaches 42.25% (30 patients, 3% of the sample), is exhibiting medium risk to grow tumors. The smallest 28.17% (20 patients, 2% of the sample), is exhibiting low risk to develop cancer. See figure 17

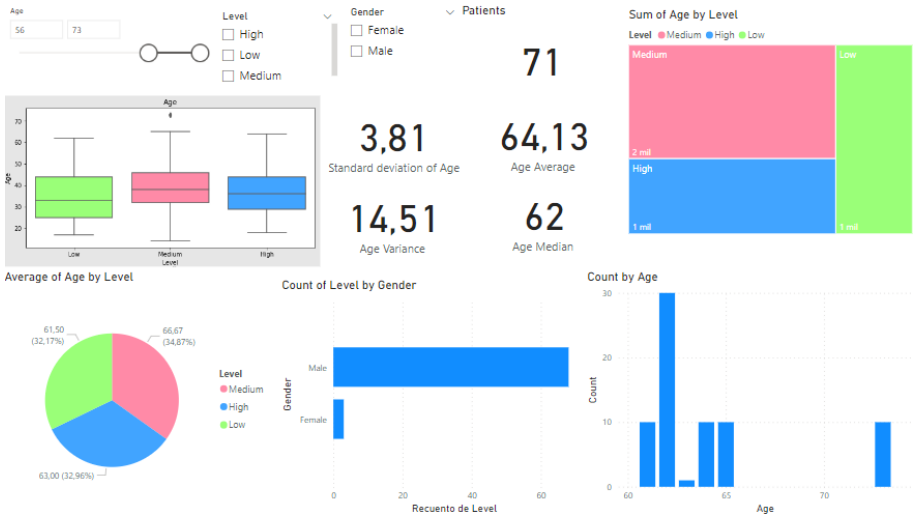


Fig. 17: Power BI dashboard containing the same charts as figure 12 changing the range of age from 56 to 73 with 71 patients being analyzed

There are other variables to consider as important?
The variables chosen describe some relationship in how is the behavior with the level of the condition (alcohol use, obesity and occupational hazards) and the cancer level. But there is still some meaningful variables that share this patterns too, such as **Dust Allergy and Fatigue**. Also despite the information shown by boxplots on figure 4 for the cases of the **patients that smoke or are being expose to smokers** as group is consider both as important by the explanation that those are factors that usually affect significantly the results of cancer level of patients.

c) **Could it be explain in numbers what is the impact of the variables chosen in the cancer level?**
To address this question it's needed to analyse the behaviour of how the variables affect separately to the level of cancer. To do so it is used the key elements analyser of Power BI, the following results determine the results of the analyser:

Condition	Increase in average	Impact in high level of cancer
Obesity	2.12	17.97
Passive smoker	2.31	9.72
Alcohol use	2.6	7.99
Dust Allergy	1.98	7.35
Occupational hazards	2.10	6.18
Smoking	2.49	5.06
Fatigue	2.25	4.76
Age	62 - 64	2.79
Age	17 - 37	1.5
Gender	Male	1.47

TABLE I: Likelihood of level of cancer being high compare to the increase of average in the conditions chosen as most relevant

There are many important factors to consider in table I. First at all the Obesity condition is the most important variable regarding to affect on high level of cancer patients. As group there was the hypothesis of having a high level of cancer would be determine by smoking and in less impact by the passive smoker conditions. What the results say is that the **passive smoker condition have more impact to determine a high level of cancer than the smoking itself by 47.94%**. The Alcohol use is an important factor to determine the level of cancer as expected. The Age as expected from the last analysis is not relevant to determine a high level condition.

2) **Risk Level**
Since the variables share common characteristics, the authors of this report decided to group the data to maintain both a cohesive analysis and uncluttered data analysis and graphs. More specifically, they decided to use the mean of the responses every given patient provided during the data collection process. This signifies the severity of the grouped factors that affect the grade of cancer. For instance, a patient with an average Symptoms score of 7 is expected to have on

average a relatively high level of adverse symptoms. The same is expected for the rest of the grouped factors: Lifestyle Choices, Long Term Health Issues and Environmental Choices. In the following graph we can see how the grouped factors behave given the risk level of cancer:

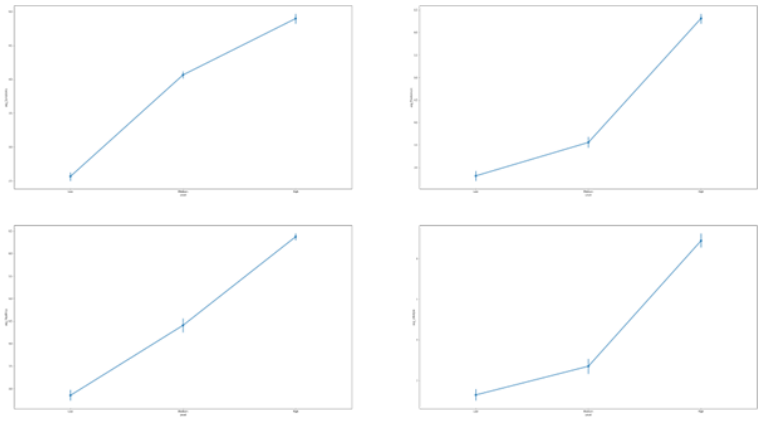


Fig. 18: Stripplots of the grouped factors scores given the risk level. From the upper left corner and clockwise: Average Score of Symptoms, Avg. Score of Environmental Factors, Avg. Score of Long Term Health Issues and Avg Score of Lifestyle Choices.

In these graphs, on the x-axis is depicted the risk level of cancer from Low to High and on the y-axis are the values for each variable from 1 to 9. The points in the graphs depict the average of the responses the participants provided during the data collection, for each risk level, while the lines that protrude from the dots depict the standard deviation of the values for each level.
The patients with:

- low-risk level have a group size $n_{low} = 303$ or 30.3% of the sample with 152 of them being male (15.2% of the sample) and 151 of them being female (15.1% of the sample)
- Medium risk level have a group size $n_{medium} = 332$ or 33.2% of the sample with 198 (19.8% of sample) being male and 134 (13.4%) being female
- and High risk level of cancer have a group size $n_{high} = 365$ or 36.5% of the sample, the bigger group, with 252 (25.2%) being male and 113 (11.3%) being female.

a) **Risk Level methodology analysis**
According to the graphs, it seems that the variables show some relationship with the risk level of cancer of the patients. In order to be certain, it has to be understood if the risk level groups are distinct from each other. Considering that the data are ordinal in nature, ANOVA tests should be executed for every variable given the group of the risk level. For the ANOVA test, though, to provide reliable results, the data should be distributed normally. To determine the normality of the data Skewness and Kurtosis z-score were used in conjunction with Probability-Probability Plots (P-P Plots). Kolmogorov-Smirnov Goodness of Fit tests were also executed for each variable to assess whether the normality presupposition is true. On every normality test, the data will be examined in their initial form and at a logarithmic transformed form by using their natural logarithm. Sometimes, by transforming non normal data with a logarithmic transformation, analysts can produce normally distributed data to work with.
The outcome of the normality tests will determine eventually, whether the parametric tests or its nonparametric equivalent will be more suitable for our analysis. Their usage helps to determine whether the variables under analysis are taken from the same population or not, or in other words, whether they share the same measures of central tendency. If the tests provide results that reject the hypothesis that the groups share the same measures of central tendency in either of these tests, it is safer to assume that the risk levels are correlated to the responses the patients provided during the data collection process of this dataset. This correlation can be further explored by the usage of Logistic Regression.
Multiple software packages were utilized to perform the analysis. For many of the visualizations Python was utilized and more specifically the statsmodels, Seaborn and matplotlib libraries alongside Pandas for tabular manipulation of the data. The normality testing and the ANOVA and Kruskal-Wallis analysis was performed through the IBM SPSS statistical software.

b) **Results Risk level analysis** Starting this section of the analysis, the hypothesis has to be set:

- H_0: The variables are normally distributed.
- H_a: The variables are not normally distributed.

Before more robust statistical approaches are implemented, heuristic approaches can be quite useful. Having in mind that normal distributions have the mean, the median and the mode equal, it could be helpful to check whether this applies in this case for both the initial and the log-transformed data

	Mean	Median	Mode
avg_Symptoms	3,91	3,91	5
avg_Enviroment	4,34	4,25	2,5
avg_HealthIss	4,65	5,25	6,75
avg_Lifestyle	4,26	4	4,5

Fig. 19: The Mean, Median and Mode of the variables at their initial state

It seems that most of the variables' measures of central tendency do not converge to a single value, rejecting the null, with a single exception. Lifestyle choices groups seem to have a really close measure to 4.25. Moving on to the log-transformed data, it seems that measures are closer to each other but this could be an effect of the logarithmic compression the data went through. See next figure:

	Mean	Median	Mode
avg_Symptoms	1,32	1,36	1,61
avg_Enviroment	1,38	1,45	0,92
avg_HealthIss	1,45	1,66	1,91
avg_Lifestyle	1,28	1,39	1,5

Fig. 20: The Mean Median and Mode of the log-transformed data

The second heuristic approach will look into the histograms of the variables in both their initial and log-transformed state. The goal is to detect whether the distributions depict the familiar bell shape of the Gaussian (normal) distribution. See Figure 21 and Figure 22:

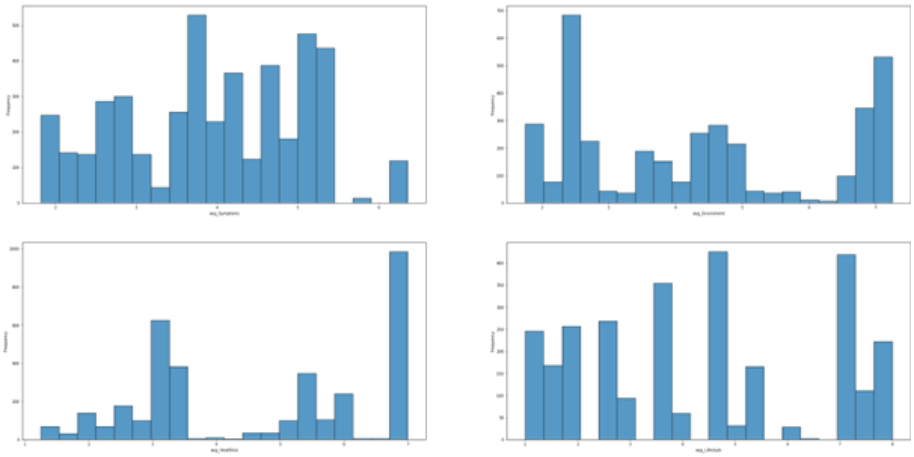


Fig. 21: Histogram of the grouped variables(initial values). From the upper left corner and clockwise: Average Score of Symptoms, Avg. Score of Environmental Factors, Avg. Score of Long Term Health Issues and Avg Score of Lifestyle Choices

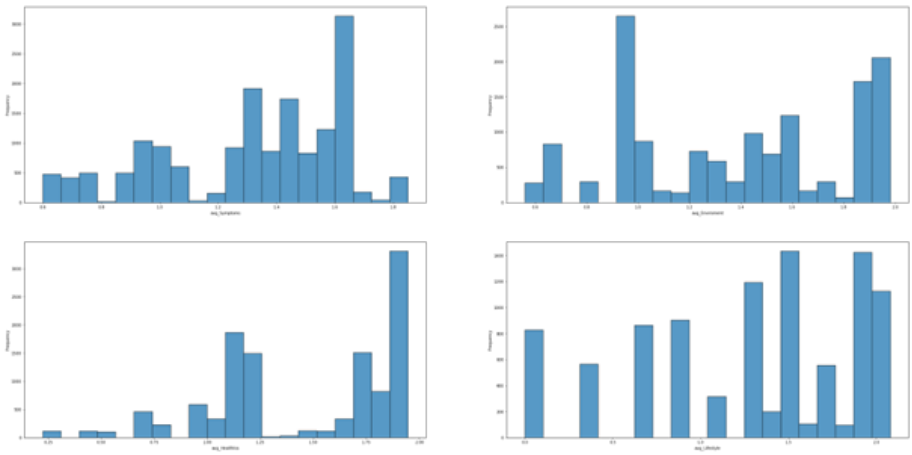


Fig. 22: Histogram of the grouped variables(log-transformed values). From the upper left corner and clockwise: Average Score of Symptoms, Avg. Score of Environmental Factors, Avg. Score of Long Term Health Issues and Avg Score of Lifestyle Choices

$z = \pm 1.96$ interval, then we can accept the null hypothesis. The following Figure 12 and Figure 13 depict the results of the tests:

	Skewness z-score	Kurtosis z-score
avg_Symptoms	-3,6320	-4,2562
avg_Enviroment	-2,2347	-3,1584
avg_HealthIss	-2,6637	-3,4735
avg_Lifestyle	-1,8011	-2,4481

Fig. 23: Skewness and Kurtosis z-scores (initial data)

	Skewness z-score	Kurtosis z-score
avg_Symptoms	-6,2253	-6,0199
avg_Enviroment	-3,3548	-6,2409
avg_HealthIss	-4,5384	-5,1473
avg_Lifestyle	-2,9525	-3,1191

Fig. 24: Skewness and Kurtosis z-scores (log-transformed data)

It is obvious that none of the values fall within the required values to accept the null hypothesis, with an exception of the skewness z-score of the Avg Score for Lifestyle choices. Probably still not enough to accept the null hypothesis. Additional testing is required
Following that, P-P plots will be used for each grouped variable of the dataset.

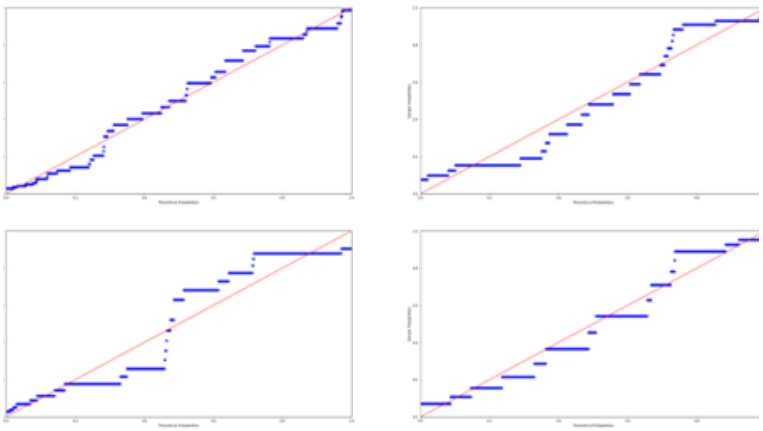


Fig. 25: These are the Probability-Probability Plots of the initial grouped variables. From the upper left corner and clockwise are: Avg Score of the Symptoms, Avg. Score of Environmental Factors, Avg. Score of Long Term Health Issues and Avg Score of Lifestyle Choices

In this graph, the initial values are utilized to create the P-P plot. In case the data were fit on the 45^o line, this would work as a strong indication that the values of each variable are normally distributed around. It's quite obvious from how the data are spread around the 45^o degree line that the variables are not normally distributed. The null hypothesis is rejected.
A similar pattern occurs in the log-transformed data in the following graph, as the log-transformation that was utilized to 'force' normality on the data, failed to produce appropriate values to perform the ANOVA:

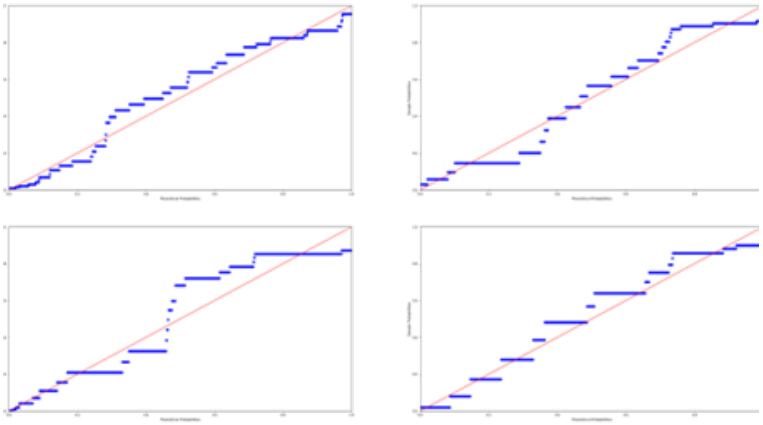


Fig. 26: These are the Probability-Probability Plots of the log-transformed grouped variables. From the upper left corner and clockwise are: Avg Score of the Symptoms, Avg. Score of Environmental Factors, Avg. Score of Long Term Health Issues and Avg Score of Lifestyle Choices

These graphs serve as an additional indication that the data do not fulfill the normality requirement to run the ANOVA test. In order to be certain that this requirement cannot be met, the Kolmogorov-Smirnov Tests.

	kstest Statistic	P-value
avg_Symptoms	0,966	0
avg_Enviroment	0,96	0
avg_HealthIss	0,948	0
avg_Lifestyle	0,847	0

Fig. 27: Kolmogorov-Smirnov Goodness of Fit test. (Normal Distribution CDF c - Initial Data)

The graph above provides in a compact way the results of the Kolmogorov-Smirnov Test for Goodness of Fit for every grouped variable in the dataset. The following graph shows how the Kolmogorov-Smirnov Test performed with the log-transformed data.

	kstest Statistic	P-value
avg_Symptoms	0,966	0
avg_Enviroment	0,96	0
avg_HealthIss	0,948	0
avg_Lifestyle	0,847	0

Fig. 28: Kolmogorov-Smirnov Goodness of Fit test. (Normal Distribution CDF used - Log-Transformed Data)

For this part of the analysis it is crucial to explain how this test works. Starting with the Kolmogorov-Smirnov Goodness of Fit Test, it constructs a CDF (Cumulative Distribution Function) which is derived from a random sample of our data, called EDF (Empirical Distribution Function), and then compares it with the theoretical CDF of a specified distribution, in this case the normal distribution. The hypotheses for the test are:

- H_0: The data come from the normal distribution.
- H_a: At least one of the values does not match the normal distribution.

Considering that and a level of statistical significance of 5% ($\alpha = 0.05$), the null hypothesis of normality in our data can be safely rejected , as every variable gives out a p-value of less than 0.001, far below the level of significance. Thanks to these tests it can be safely concluded that the ANOVA Test would not produce reliable results in the analysis, neither in the original data nor in the log-transformed data. The non parametric equivalent of ANOVA, Kruskal-Wallis H test and the Mann-Whitney U test, can be used instead.

c) **Testing the hypothesis**

Before presenting the results of the Kruskal-Wallis H test, it is important to explain how it works. As it was mentioned in previous parts of this report the Kruskal-Wallis test is the nonparametric equivalent of the ANOVA test and therefore sometimes it's called one-way ANOVA on ranks. The reason it's called that is that it utilizes a ranking system of the data in each group to determine if the groups of the analysis share the same median. The method goes by sorting the data in an ascending order and ranking them based on that to calculate the H statistic. The H statistic is then compared to the critical cutoff, determined by the critical ^{2} value. If the H statistic is less than the critical ^{2} for k-1 degrees of freedom, then the null hypothesis can be rejected. The formula utilized to produce the H statistic, is as follows:
where N is the sample size, R_i is the sum of ranks for group i, n_i is the size of the group i and k is the number of groups. Its hypothesis testing is structured as follows:

$$H = \frac{12N}{(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{1}$$

- H_0: med(low)=med(medium)=med(high), where the notation med() is utilized to designate the median of each group.
- H_a: At least one of the medians is not equal with the others.

After specifying the details on how the test works, it must be explained how the data were set in order to execute it. In the place of the response, the values of each variable was set and in place of the groups in question the risk level of cancer was set(i.e. Pr(variable—level of risk)).

	H-Statistic	p-value
avg_Symptoms	720,7455	0
avg_Enviroment	677,4203	0
avg_HealthIss	654,5126	0
avg_Lifestyle	514,798	0

Fig. 29: Kruskal-Wallis H test for the grouped variables

Considering a level of statistical significance of 5% ($\alpha = 0.05$), the null hypothesis can be safely rejected. From this, it can be understood that none of the groups of the analysis (i.e. Patients with Low, Medium, High Risk of developing cancer) have a common median and thus it is safer to assume that there is a relationship between the risk level of cancer and the severity of the factors that can cause lung cancer.
The Mann-Whitney test is going to build onto the Kruskal-Wallis test as it represents the nonparametric substitute of the independent sample t-test. Its goal is to determine if the measures of central tendency between two independent groups is the same. It utilizes a similar ranking philosophy with Kruskal-Wallis test by using the following formula to calculate the U statistic:

$$U_i = n_i n_j + \frac{(n_i + (n_i + 1))}{2} - T_i \tag{2}$$

With i, j being the ith and the jth group, n_i, n_j being the number of observations in group i and j respectively and T_i being the sum of ranks for group i. The same calculation is produced for both groups and the smaller U statistic is used to find the z-score based on the two groups. If the z-score is within the ±1.96, we accept the null hypothesis. The hypothesis structure goes as follows:

- H_0: There is no difference (in terms of central tendency) between the two groups in the population.
- H_a: There is a difference (with respect to the central tendency) between the two groups in the population.

For this test the results of the grouped variables are the following:

	U-Statistic	p-value
('avg_Symptoms', 'Low-Medium')	2870,5	0
('avg_Symptoms', 'Medium-High')	18924	0
('avg_Symptoms', 'Low-High')	615,5	0
('avg_Enviroment', 'Low-Medium')	25774,5	0
('avg_Enviroment', 'Medium-High')	5085,5	0
('avg_Enviroment', 'Low-High')	1480,5	0
('avg_HealthIss', 'Low-Medium')	14287,5	0
('avg_HealthIss', 'Medium-High')	15598,5	0
('avg_HealthIss', 'Low-High')	1753,5	0
('avg_Lifestyle', 'Low-Medium')	37201	0
('avg_Lifestyle', 'Medium-High')	13375	0
('avg_Lifestyle', 'Low-High')	5394,5	0

Fig. 30: Mann-Whitney U test for the grouped variables.

It is quite obvious from both the p-values and the statistics that the variable groups are quite distinct from one another in terms of measures of central tendency.
This is also quite evident from this graph, where each variable group shows differences for each grade of cancer.

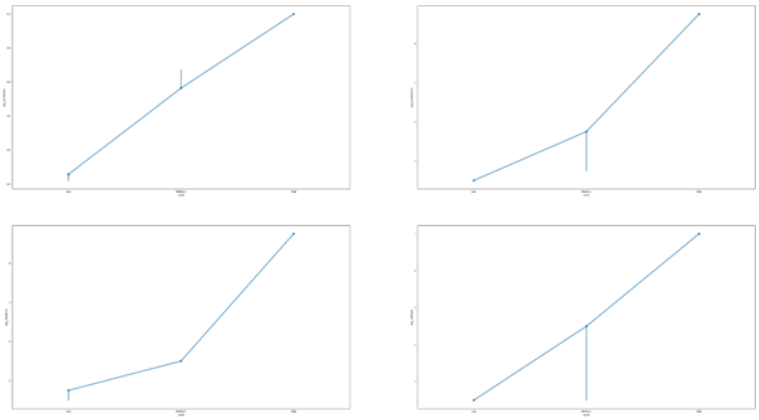


Fig. 31: Point Plots of the medians of the grouped variables given the grade of cancer. From the upper left corner and clockwise are: Avg Score of the Symptoms, Avg. Score of Environmental Factors, Avg. Score of Long Term Health Issues and Avg Score of Lifestyle Choices

d) **Correlations** Spearman rank correlation seems to confirm the alternative hypothesis

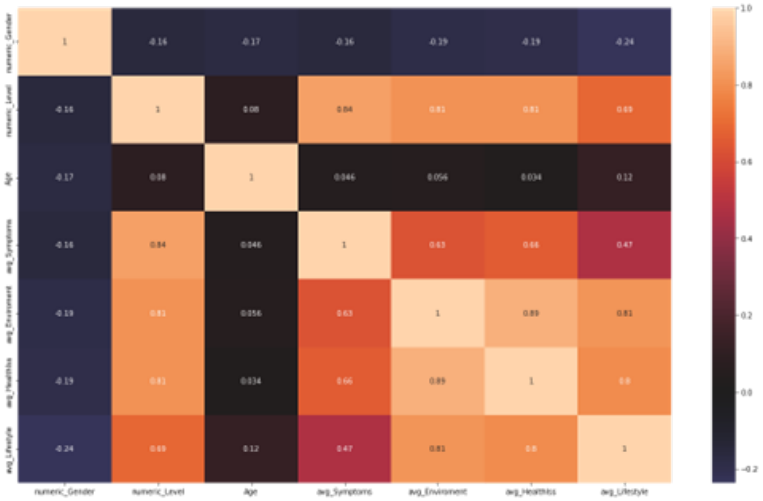


Fig. 32: Spearman Rank Correlation of the grouped variables, the Age of the patients, the Risk Level of Cancer and the Gender of the patients.

The usage of the Spearman's Rank Correlation, as the name suggests, is to find the level of interaction between the variables and it's quite suitable for ordinal data, like in this case. It should be noted that the variables of the Gender and the risk Level of cancer were coded numerically to perform the necessary computations to produce the correlation coefficients. It's quite obvious that the Level of cancer and the grouped variables show a very strong positive correlation with each other. The greatest interaction occurs between the Avg Score of the symptoms and the Risk Level of cancer with a correlation coefficient of 0.84. This means that for 1% increase of the Average Score of the Symptoms, the risk level increases by 0.84% also and vice versa. the Average Score of Environmental Causes share a coefficient of 0.81 with the Level of Cancer, the Average Score of the Long Term Health Issues share also a coefficient 0.81 with the Level of Cancer and the Average Score of the Lifestyle Choices share a coefficient of 0.69 with the Level of Cancer. Additionally, the variable of gender seem to have minimal effect on the grade of cancer, as its coefficient is within the range of ± 0.2. Lastly, the variable of Age seems to have no apparent effect on the grade of cancer or interaction with any other variable in any given circumstance. This analysis can be confirmed by the following p-values matrix.

Spearman's rank correaltion p-values	numeric_Gender	numeric_Level	Age	avg_Symptoms	avg_Enviroment	avg_Healthlss	avg_Lifestyle
numeric_Gender	0	0	0	0	0	0	0
numeric_Level	0	0	0,0117	0	0	0	0
Age	0	0,0117	0	0,1454	0,0774	0,2851	0,0001
avg_Symptoms	0	0	0,1454	0	0	0	0
avg_Enviroment	0	0	0,0774	0	0	0	0
avg_Healthlss	0	0	0,2851	0	0	0	0
avg_Lifestyle	0	0	0,0001	0	0	0	0

Fig. 33: Spearman's rank Correlation P-values for every corresponding coefficient

3) **Cancer level of a patient prediction**

Two approach for predictions are being selected and analysed

a) **Logistic Regression taking some variables into account**

This question was selected by thinking about how to get a diagnoses having some data in hand and what would be the accuracy of this result. To handle a solution to this question a logistic regression model in python (scikit-learn) is trained, the idea is by having separate data determine an accuracy in the prediction of cancer level, the groups are selected in the next way:

i) Gender, Obesity
ii) Gender, Passive smoker
iii) Gender, Alcohol use
iv) Gender, Dust Allergy
v) Gender, Obesity, Passive smoker, Alcohol use
vi) Gender, Obesity, Passive smoker, Alcohol use, Dust Allergy, Occupational hazards, Smoking, Fatigue
vii) All the variables

Having these groups assigned the models are trained by separate 80% of data to this purposed and the rest 20% of total data to test the models, the following results show the accuracy for the testing of the models, also in one specific example chose from the test data:

Example:

Patient Id	**Age**	**Gender**	**Air Pollution**	**Alcohol use**
P86	29	2	4	5
Dust Allergy	**OccuPational Hazards**	**Genetic Risk**	**chronic Lung Disease**	**Balanced Diet**
6	5	5	4	6
Obesity	**Smoking**	**Passive Smoker**	**Chest Pain**	**Coughing of Blood**
7	2	3	4	8
Fatigue	**Weight Loss**	**Shortness of Breath**	**Wheezing**	**Swallowing Difficulty**
8	7	9	2	1
Clubbing of Finger Nails	**Frequent Cold**	**Dry Cough**	**Snoring**	**Level**
4	6	7	2	High

TABLE II: Example for test the different models selected by the most meaningful variables on cancer level condition. For test g is ignore the Patient Id

Case	Variables chosen	Accuracy for 20% of total data set	Values in the example	Result of the example
a	Gender	76.4%	2 (female)	High (correct)
	Obesity		7	
b	Gender	59.29%	2 (female)	Low (incorrect)
	Passive Smoker		3	
c	Gender	64.32%	2 (female)	High (correct)
	Alcohol use		5	
d	Gender	65.32%	2 (female)	High (correct)
	Dust Allergy		6	
e	Gender	73.87%	2 (female)	High (correct)
	Obesity		7	
	Passive Smoker		3	
	Alcohol use		5	
f	Gender	85.93%	2 (female)	High (correct)
	Obesity		7	
	Passive Smoker		3	
	Alcohol use		5	
	Dust Allergy		6	
	Occupational hazards		5	
	Smoking		2	
	Fatigue		8	
g	All variables	97.98%	Table II	High (correct)

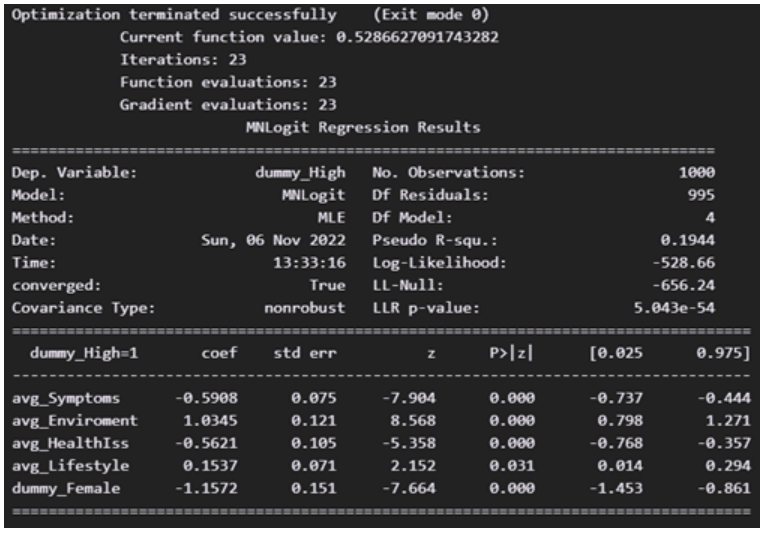
TABLE III: Results of testing the models for different scenarios proposed by the authors

The table III concluded something important, at first the most significant variable that could have more accuracy by itself is the obesity. The example e had less accuracy than the a one despite it has more variables in consideration, that means predicted from a combination of important features won't always have better results. With all the variables is possible to predict with so much certainty the level of cancer condition of a patient.

b) **Logistic Regression by significance**

For this estimation dummy variables of the high risk level variable were created and they were used as the dependent variable of the model. Each of the grouped variables were used to see how they affected the risk level. Alongside the grouped variables the age and the gender were used to estimate if there is an effect on the risk level of cancer. In case some of the variables seem to be statistically insignificant at the level of significance a=0.05 then the regression will be rerun excluding the variable. The hypothesis structure goes as follows:

- H_0: The variable coefficient estimation is not statistically significant(i.e. $\beta = 0$)
- H_a: The variable coefficient is statistically significant(i.e. $\beta 0$).



```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.5286627091743282
            Iterations: 23
            Function evaluations: 23
            Gradient evaluations: 23
                          MNLogit Regression Results
==============================================================================
Dep. Variable:             dummy_High   No. Observations:                 1000
Model:                        MNLogit   Df Residuals:                      995
Method:                           MLE   Df Model:                            4
Date:                Sun, 06 Nov 2022   Pseudo R-squ.:                  0.1944
Time:                        13:33:16   Log-Likelihood:                -528.66
converged:                       True   LL-Null:                       -656.24
Covariance Type:            nonrobust   LLR p-value:                5.043e-54
==============================================================================
  dummy_High=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
avg_Symptoms      -0.5908      0.075     -7.904      0.000      -0.737      -0.444
avg_Enviroment     1.0345      0.121      8.568      0.000       0.798       1.271
avg_HealthIss     -0.5621      0.105     -5.358      0.000      -0.768      -0.357
avg_Lifestyle      0.1537      0.071      2.152      0.031       0.014       0.294
dummy_Female      -1.1572      0.151     -7.664      0.000      -1.453      -0.861
==============================================================================
```

Fig. 34: Logistic Regression Result of High Risk Level of Cancer given Avg Score of Grouped Variables and Gender.

Thanks to these results, the hypothesis testing can begin. The variable for Age did not make it in the estimation as it seemed to be non statistically significant,in the previous test of rank correlation therefore in the test was used the rest of the variables. The most impactful variables, though, are the average score of environmental factors and the gender. The avg score of environmental factors has a coefficient of 1.29 and is statistically significant, rejecting the null hypothesis. The coefficient means that the probability of showing High risk to develop cancer increases by 1.03% with each increase in the score category. In the case of the gender a dummy variable was utilized to depict whether a patient is male or female, with 0 being male and 1 being female. The negative value of the coefficient agrees also with the Descriptive analysis chapter of this report. As the majority of the patients that had high risk to develop cancer were men. The coefficient seems to be statistically significant and it means that if the patient is a woman there is a lower probability for her to develop cancer, of about 1.16% less probable.
There is a problem though with this model. Variables, like the Grouped Symptoms and the grouped Health Issues show to have the opposite effect of what was expected and what was previously proven. This could be potentially a case of Type II error in our estimation. Therefore, the Multiple Logistic Regression approach to predict future cases will be scrapped due to ambiguity in its results.

IV. Conclusions

- The table I demonstrate that the three most relevant conditions to have a **high** level of cancer organized by importance are **obesity, passive smoker and alcohol use**. Those are circumstances that can be change by the patient habits and definitely a reduce on them will have a big impact in the cancer treatment.
- Thanks to the non parametric methods that were utilized in section 4 of the report, the medians of the grouped factors given the level of cancer are not equal, making it safer to accept the hypothesis that there is indeed a positive relation between the factors and the risk level of cancer.
- Despite the impact of the most meaningful conditions showed in table I on high level of cancer. There is not a direct condition to determine the accuracy on the test made by the models in table III. That means it was expected to have a correlation between table I and III, but what is shown is that the organization by meaningfulness of the variables in **general** that determine level of cancer is **obesity, dust allergy, alcohol use and passive smoker**.
- The gender of the Patients plays a role in the likelihood of developing cancer, with men being more prone to it.
- It's possible to have a good accuracy prediction of cancer level of a patient having into account all the variables that measure the data set, with a 97.98% of certainty in the diagnose of the patient by the model built by the authors. If it's only know the group of variables f in table III it's also possible to have good results with an accuracy of 85.93%. And lastly if it's just know the obesity of the patient in less of 23.6% of the cases the model a of table III would have an error.
- As figure 7 shows the consume of alcohol use determine a lot on the patient level of cancer condition but also that a minimum change on the consume of alcohol could result from having a high level condition to a medium one. This is shown by the average of alcohol use by level in the case of medium level having 6.7 of use and by high 6.9 of use.
- One of the most important insights of this analysis is that the patients that are exposed to smokers are more likely to have a high level of cancer than the smokers themselves. And a strong recommendation for cancer patients is to avoid smoke room scenarios
- A common belief is that age determines a huge impact on the level condition of cancer, when people get older have more tendency to have a likelihood of medium to high cancer levels. What is shown in table I is this affirmation is far from being true; it does not exist a huge impact on having a high level of cancer when the patient is old, there are factors more significant that determine the cancer condition.

References

[1] M. Roser and H. Ritchie, "Cancer," *Our World in Data*, 2015. https://ourworldindata.org/cancer.

[2] U. D. of Health and H. Services, "What is cancer?," *National cancer institute*, 2021. https://www.cancer.gov/about-cancer/understanding/what-is-cancerdefinition.

[3] U. D. of Health and H. Services, "Tumor grade," *National cancer institute*, 2022. https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-grade.

ANALYSIS OF CO2 EMISSIONS THROUGH HISTORY WORLDWIDE

§ Afsar Nehan, * Alharthy Raiya, ¶ Dimos Konstadimos, † Hoyos Harrison, ‡ Sunilkumar Sidharth

§nehaafsar05@gmail.com, *raiyaalharthy1992@gmail.com, ¶konstadimos7@gmail.com,
†harrihoyos2680@gmail.com, ‡sidharthsunil74@gmail.com
Brainnest Data Analysis Trainee Program

Abstract

The global environment had been ignored throughout humanity's history passing decades without still being in minds of countries. Nowadays the situation has changed drastically due to scientific research showing the possible and impactful changes that the destruction of the ozone layer could cause. And had been so taken into account that the France treatment of 2050 of zero CO2 emissions was signed in 2016. This research study is made with the idea of showing some impact in the results on the treatment and analysis of a dataset considering several factors that affect the global environment regarding different greenhouse gasses emissions most accordingly to CO2, to see the research code exploration the reader is invited to see the repository of this report here.

Firstly, a brief explanation of the data set is made with the idea that the reader has in mind the key concepts of the research, then how to handle the missing values in the data transformation part is introduced. An analysis of data is explained answering some important research questions made by the authors. The conclusion part is the last part of the report regarding the insights found.
keywords: global environment, ozone layer, greenhouse gasses, CO2

I. Introduction

Greenhouse gases are gases that capture heat in the atmosphere. This report contains data on the main greenhouse gas emissions and removals from the atmosphere. As human-caused greenhouse gas emissions rise, they accumulate in the atmosphere and warm the climate, causing a slew of other changes all over the world—in the atmosphere, on land, and in the oceans. The concentration, as well as abundance, is the amount of a specific gas in the air. Rising greenhouse gas emissions result in higher concentrations in the atmosphere. Greenhouse gas levels are measured in parts per million, billions, and trillions.

The report includes 19832 rows and 58 variables that are classified as follows:

1) CO2 includes CO2 growth in percent, CO2 growth, Trade CO2, CO2 per capita, consumption of CO2 per capita, share global CO2, cumulative CO2, share global CO2, CO2 per GDP, consumption CO2 per GDP, CO2 per unit energy.
2) CO2 is further classified as coal, cement, flaring, gas, oil, and other industry-related CO2.
3) CO2 per capita categories: coal per capita, cement per capita, flaring per capita, gas per capita, oil per capita, and other industry related per capita.
4) CO2 categories by share: coal, cement, flaring, gas, oil, and other industry-related CO2.
5) CO2 Categories in Cumulative Form: cumulative cement CO2, cumulative coal CO2, cumulative flaring CO2, cumulative Gas CO2, cumulative Oil CO2, cumulative
6) Other variables: Total Greenhouse Gases, Greenhouse gas per capita, methane, methane per capita, nitrous oxide, nitrous oxide per capita, population, GDP, primary energy consumption, energy per capita, and energy per GDP.

Carbon dioxide (CO2) is the atmospheric greenhouse gas released by human activity.). CO2 emissions are the primary cause of global climate change. It is widely acknowledged that to prevent the worst effects of climate change, the earth must urgently reduce carbon emissions. However, how this liability is distributed among regions, countries, and individuals has long been a source of disagreement in international discussions. As part of the Earth's carbon cycle, carbon dioxide is naturally present in the atmosphere (the natural circulation of carbon among the atmosphere, oceans, soil, plants, and animals. The primary source of CO2 emissions from humans is the use of fossil fuels like coal, natural gas, and oil for transportation and energy. CO2 is also released into the atmosphere with certain production plants and land-use changes.

Besides that, Methane is also emitted naturally from a variety of sources. The largest source is natural wetlands, which emit CH4 from bacteria that decompose organic materials in the absence of oxygen. Termites, oceans, sediments, volcanoes, and wildfires are illustrations of fairly small sources. Leaks from natural gas systems and livestock production are two examples of human activities that emit methane. Natural sources of methane, such as wetlands, also emit methane. Moreover, natural processes in soil and chemical reactions in the atmosphere assist in removing CH4 from the atmosphere. Methane has a much relatively short lifetime in the atmosphere than carbon dioxide (CO2), but CH4 is more efficient at trapping radiation than CO2.

N2O is about 300 times more powerful than carbon dioxide when it comes to heating the atmosphere. As well as, CO2, has a long lifetime in the atmosphere, lasting an estimated 114 years before collapsing. It also contributes to the depletion of the ozone layer. Overall, the carbon intensity of nitrous oxide is no surprise given. According to the Intergovernmental Panel on Climate Change (IPCC), nitrous oxide accounts for roughly 6% of greenhouse gas emissions, with agriculture accounting for

roughly three-quarters of N2O emissions.
However, despite their significant input to climate change, N2O emissions have been largely ignored in environmental legislation. And the gas keeps accumulating. Based on a 2020 overview of nitrous oxide sources and sinks, emissions have increased by 30% over the last 40 years and are now surpassing all previous levels.
Other gases are water vapor and ozone. These are two other greenhouse gases (O3). Water vapor is the most abundant greenhouse gas on the planet, but it is not monitored similarly to the other greenhouse gases since it is not directly released into the atmosphere by human activity and its repercussions are unknown. Similarly, troposphere ozone (not to be confused with the protective stratospheric ozone layer higher up) also isn't emitted directly into the air but emerges from complex reactions among pollutants in the air. Greenhouse gas emissions have had far environmental and health consequences. They contribute to respiratory disease due to smog and air pollution, and they end up causing global warming by trapping heat. Other effects of climate change caused by greenhouse gases include extreme weather, food supply disruptions, and increased wildfires. The weather patterns we've learned to expect will shift; some species will go extinct, while others will migrate or grow.
This report will address Carbon Dioxide and various other well-known gases that disrupt the climate due to their consumption and share in countries over 70 years, which will assist us in understanding which countries are strongly affected and producing harmful gases that impact the environment.

II. Cleaning Data

1) Missing Values Summary
The total amount of records filled or not in the dataset is 1.461.832. From this total value, there are 760.695 missing values, which represents 52% of the total dataset. To see the summary of total missing values and the percentage representation per column see table X_1, the column *percentage of missing values* i represented in the figure 1showing the columns that have more missing values in descending order

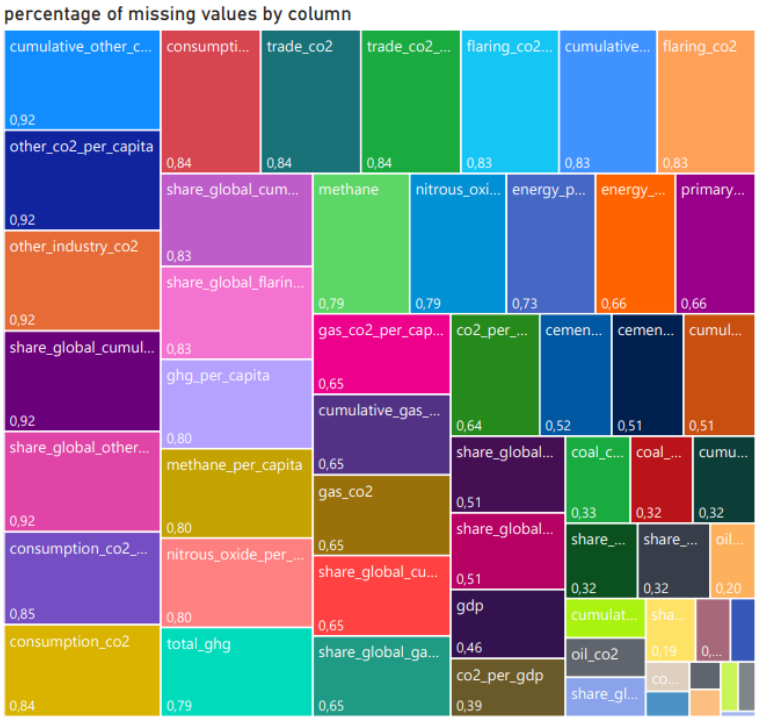


Fig. 1: Treemap of the percentage of missing values by column, showing for example that the three features with more missing values are cumulative other co2, other co2 per capita, other industry co2

To see a more deeply explanation of the count of missing values by column per country the table X_2 is introduced and to see the same explanation in percentage the table X_3 show these results. The figure 2 shows a map of the countries with more than 60% of total missing values where the diameter of the bubble is the percentage of missing values, the figure 3 shows the same data of figure 2 but in a treemap

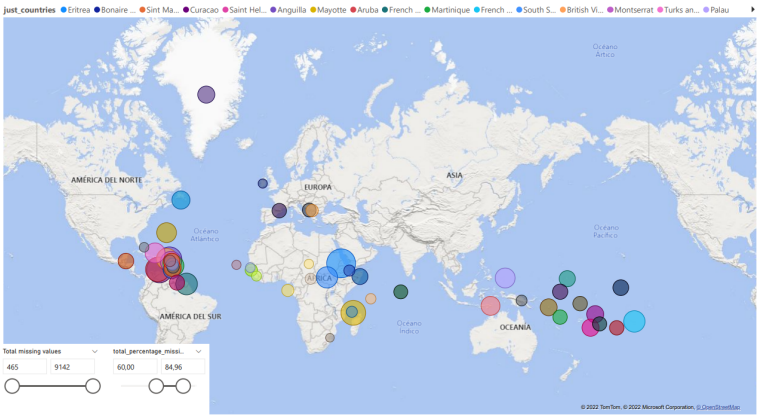


Fig. 2: Map of countries that have more than 60% of missing values

Fig. 3: Treemap containing the names of the countries and the percentage of their missing values with respect to the total

2) Remove data
Those countries showed in figure 2 are important for the total co2 emissions and other factors affecting the ozone layer globally?
These countries with the most quantity of missing values are in Central America, Africa, and islands near Oceania and Asia, according to the research of co2 emissions from Our World data [1] these regions are not quite relevant for the whole worldwide co2 emissions in total. The decision to delete those countries and their data from the data set was taken.
3) Filling data
The countries and the percentage of missing values after removing the data already explained is shown in the figure **??**; this Treemap shows that the country with fewer missing values in percentage of the total is Luxembourg with 28.97% and the highest after removing is Cote d'Ivoire with 59.10%

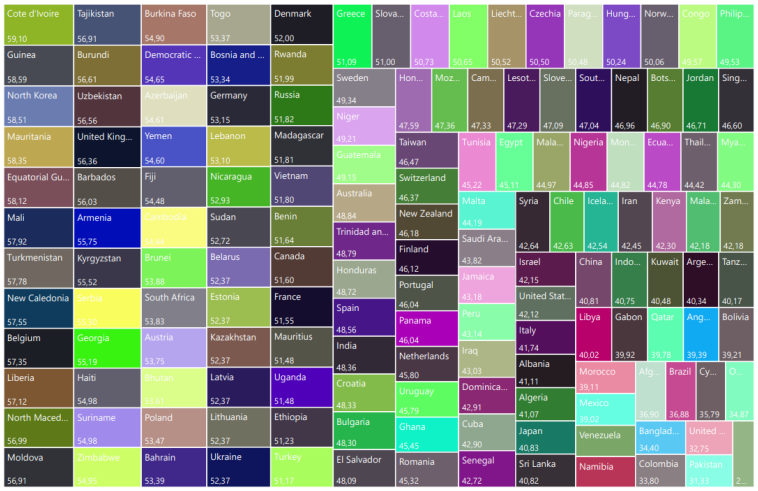


Fig. 4: Countries remaining after removing the ones with more than 60% of total missing values

There are still several missing values that need to be consider, also the analysis made so far is without having into account the world itself and the continents.
The decision to fill in the data missing is taken. An algorithm to replace these values is proposed. To explain the solution the figure **??**shows a part of comparison by columns over time for China data, the complete image is here, this is useful to understand some facts that have been found to replace the data missing:

3.1 The columns in comparison with the year follow different trends
3.2 The magnitude over time of this trend are always greater than zero

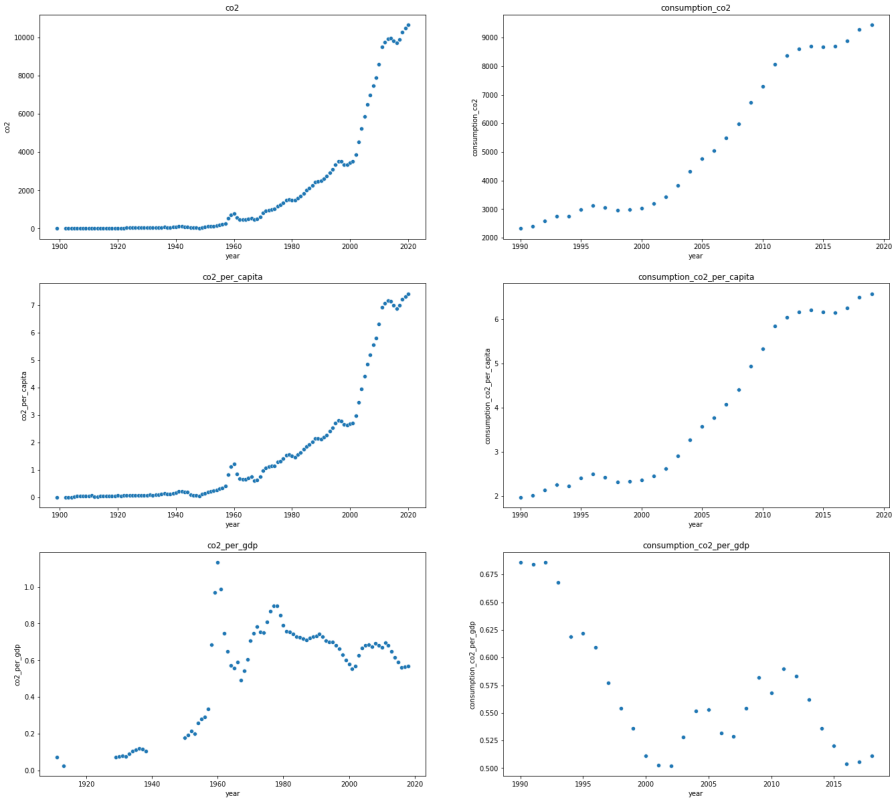


Fig. 5: Scatter plots showing the relationship between some columns in the dataset and the year for the China case with missing values

- Linear interpolation:
  To fill the data that is missing in-between values that are filled a linear interpolation is taken into account [2], the problem with this solution is that it doesn't help for years before the trend started

- Time series moving average:
  The data for China for example starts from 1900 and for the missing values that can't be filled by the linear interpolation the following approach is proposed:

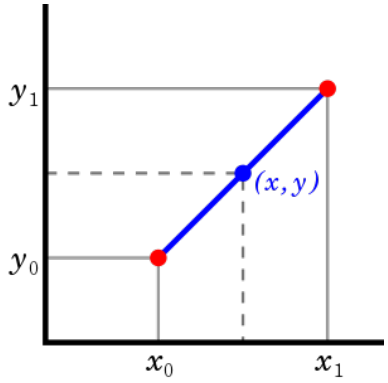
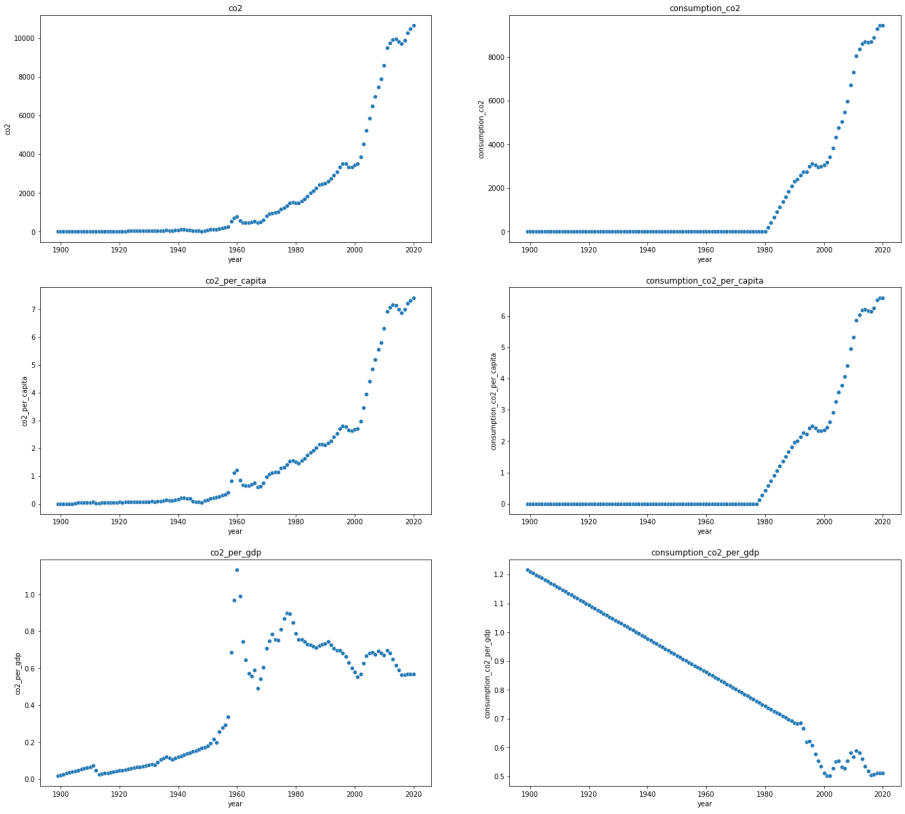


Fig. 6: Given two coordinates $(x_0, y_0), (x_1, y_1)$ the difference between them gives a factor that shows how the trend grows or shrinks

The figure 6shows that having two points in a plot the difference between them can be found and say what was the ratio of the changing from point a (x_0, y_0) to point b (x_1, y_1). For the case of analysis, the x's points will be the years and the difference between them is always 1. Having these differences for all the points in the trend an average of these differences is calculated. The average of the differences is a measure of the increase or decrease estimation of the trend. Starting from the tail point of the trend a reversed iteration is made to fill the data that is not in the trend due to missing values the value replaced is the previous value subtracted from the average of the calculated difference; if this calculated value is less than zero the replaced value is considered zero.
Figure **??**shows the same plots of figure 5 but with the missing values already filled (to see the complete figure with all the plots without missing values for China example see here). Note that in these plots all trends start since 1900



The s

Fig. 7: Same scatter plots in figure 5 but with the missing data already filled

III. Data Analysis

In this section, several questions are proposed to answer them based on the data in research. To address these questions some of the analysis presented by the authors are related to historical facts that have happened worldwide and complement what the data show.

1) **Are correlation between the data analyzed?**

1.1 **Correlational analysis of the Global Share cumulative of CO2 and Global share cumulative of CO2 coal**
Let's take a closer look at the results of the table I the strong correlation between co2 and coal where $r = .999$. It's based on the $N = 17188$ and its 2-tailed significance $p < .001$. this means there is $< .001$ probability of finding this sample correlation or a larger one.

		share global cumulative co2	share global cumulative coal co2
share global cumulative co2	Pearson Correlation	1	.999**
	Sig. (2-tailed)		<.001
	N	23949	17188
share global cumulative coal co2	Pearson Correlation	.999**	1
	Sig. (2-tailed)	<.001	
	N	17188	17188

TABLE I: Pearson correlation analysis between share global cumulative CO2 and share global cumulative coal CO2
*. Correlation is significant at the 0.01 level (2-tailed).

1.2 **Correlational analysis of the Global Share cumulative of CO2 and Global share cumulative of CO2 oil**
the results in the table IIshow the strong correlation between co2 and coal where $r = .823$. It's based on the $N = 20539$ and its 2-tailed significance $p < .001$. this means there is $< .001$ probability of finding this sample correlation or a larger one. So, these correlations indicate to which extent each pair of variables are linearly related. Finally, note that each correlation is computed on slightly different N, this is because SPSS (the software used for this analysis) uses pairwise deletion of missing values by default for correlation.

		share global cumulative co2	share global cumulative oil co2
share global cumulative co2	Pearson Correlation	1	.823**
	Sig. (2-tailed)		<.001
	N	23949	20539
share global cumulative oil co2	Pearson Correlation	.823**	1
	Sig. (2-tailed)	<.001	
	N	20539	20539

TABLE II: Pearson correlation analysis between shared global cumulative CO2 and share global cumulative oil CO2
*. Correlation is significant at the 0.01 level (2-tailed)

2) **How does economic activity affect CO2 emissions?**

2.1 **Which countries are the top CO2 cumulative emitters?**

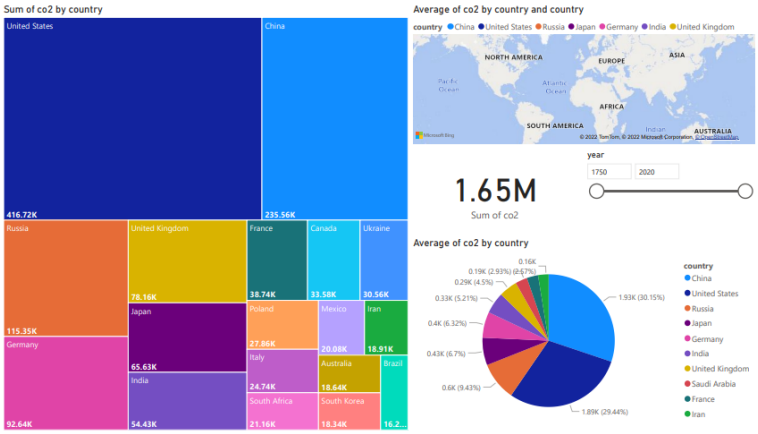


Fig. 8: Summary of measurements of Co2 emissions in history separated by countries
a. Treemap of co2 sum emissions by country b. Map of the top 7 places that emitted most co2 emissions on average c. Pie chart of the most meaningful average countries emitters of co2

The figreffig:co2_countriesshowsthetopcountriesthathaveemittedthemostco2bysumandbyaveragetroughhistory.Theseto
1001[3], $figure$**??**$showsthecorrelationbetweentheco2emissionsandGDPofthetopcountrylistinthefigure$8, $showingtha$

3) **What is the percentage representation of the top 10 countries with the highest cumulative co2 to the total GDP of the world?**

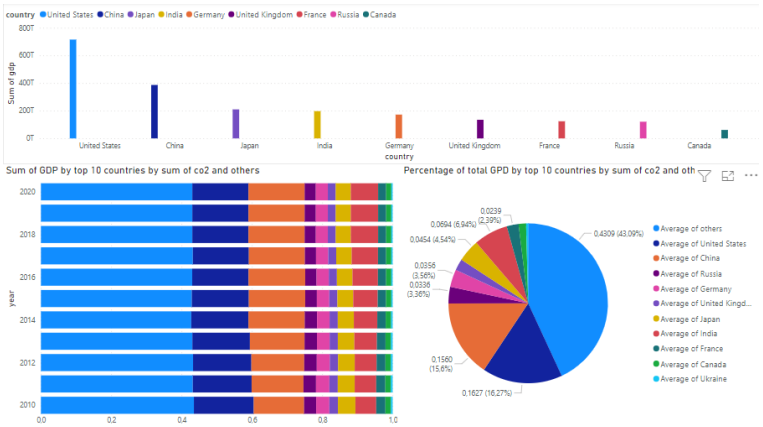


Fig. 10: Total representation of GDP by sum and on average from 2010 to 2020 by top 10 co2 emitters countries and others
a. Bar chart of the sum of GDP by top 10 co2 emitters countries
b. Stacked bar chart of the sum of GDP by top 10 co2 emitters countries and others
c. Pie chart of average GDP by top 10 co2 emitters countries and others
The color legend is the same for figures b and c

The figure 10,b shows the sum of total GDP by the top 10 co2 emitters countries and others, and 10,c shows that on average these top 10 countries represent nearly 57% of the total GDP between 2010 and 2020 saying that representation of the economy by these countries is more than a half of the total world.
The First Industrial Revolution was born in England in around 1760 and brought forth the steam engine as the driving force of production. The productivity of a single steam-powered machine could easily surpass the productivity of many professional workers in the same amount of time, increasing rapidly the total wealth a society could produce. But how did these magnificent contraptions work and make whole societies wealthier seemingly in one single night? The answer is of course coal. Coal was quite abundant in newly industrialized England and because of its high energy concentration per cm^3 and its capability to maintain higher temperatures compared to wood, it was preferred as an energy source. Therefore, the newly developed technology found a powerful energy source to drive. The side effect of this of course was the emission of huge amounts of CO2 into the atmosphere making the First Industrial Revolution also the dawn of Anthropogenic Climate change.

4) **CO2 and GDP Interaction**

To complement what the figure **??** show, the following graph depicts the CO2 emission per unit of GDP, which means the ratio at which the CO2 emissions grow compared to that of the GDP. If the ratio r produced by the calculation $r > 1$, it means that for 1 unit of GDP more than 1 unit of CO2 were produced, and in the case that the ratio is $0 < r < 1$, then for 1 unit of GDP between 1 and 0 units of CO2 was created. This ratio functions as a rough metric to assess the efficiency of production in terms of co2 emissions. See figure 11

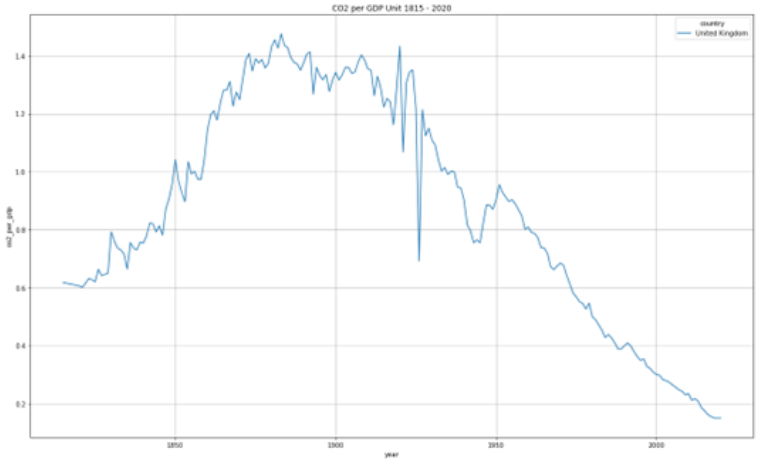


Fig. 11: The CO2 per GDP ratio of the United Kingdom over time from 1815 to 2020.

It would be easier to explain the point of this section by first analyzing one country and then moving on to multiple. In this graph, it seems that the CO2 emissions per unit of GDP reached a maximum ratio of 1.476 in the year 1883. From this point on the ratio shows a decreasing trend until it reaches 0.2 in 2020. There are many possible explanations for this behavior change:

- The creation of electrical power plants
- The usage of nuclear energy and other renewable energy sources (wind, hydroelectric, solar)
- the transition from coal-fueled to electricity-fueled means of mass transportation
- The mass automation of many productive procedures
- the development of the transistor and the creation of the digital computer
- the outsourcing of the major productive industries to developing countries and many others.

To test this notion though two linear regression models were executed. In the first model, the goal is to see if the slope of the model follows the rule behind the ratio r, if it is, basically, greater than 1 until the year where the apparent decoupling first occurs, in the case of the United Kingdom, 1883. The second model will attempt to prove if the slope of the regression will take values between 0 and 1. The model will use the GDP as an exogenous variable and the co2 emissions as an endogenous variable. It is assumed in this case that the relation of these variables is nonlinear in the form.

$$CO_2 = \beta_0 GDP^{\beta_1} \tag{1}$$

InToerform through the linear regression, both parts of the equation have to undergo a log transformation, to get a linear equation

$$logCO_2 = log\beta_0 + log\beta_1 GDP \tag{2}$$

The exponent β_1 in Equation 1, works as a coefficient of elasticity, or in other words, the percentage rate of change of the dependent variable in terms of the percent rate of change of the independent variable. If the value of the coefficient is greater than 1 then co2 grows exponentially; if it is between 0 and 1, it grows at a diminishing rate.

OLS Regression Results				
Dep.	Variable:	co2	R-squared:	0.710
Model:	OLS	Adj.	R-squared:	0.706
Method:	Least	Squares	F-statistic:	161.8
No.	Observations:	68	AIC:	53.65
Df	Residuals:	66	BIC:	58.09
Df	Model:	1		
Covariance	Type:	nonrobust		

coef	std	err	t	P>—t—	[0.025	0.975]
intercept	-0.3046	0.092	-3.308	0.002	-0.488	-0.121
GDP	1.0113	0.080	12.719	0.000	0.853	1.170

Omnibus:	28.344	Durbin-Watson:	0.157
Prob(Omnibus):	0.000	Jarque-Bera	(JB):67.130
Skew:	1.309	Prob(JB):	2.65e-15
Kurtosis:	7.104	Cond.	No.4.09

TABLE III: The co2 to GDP regression before the decoupling

OLS Regression Results			
Dep. Variable:	co2	R-squared:	0.317
Model:	OLS	Adj. R-squared:	0.312
Method:	Least Squares	F-statistic:	62.64
Date:	Tue, 08 Nov 2022	Prob (F-statistic):	
Time:		Log-Likelihood:	53.882
No. Observations:	137	AIC:	-103.8
Df Residuals:	135	BIC:	-97.92
Df Model:	1		
Covariance Type:	nonrobust		

coef	std	err	t	P>—t—	[0.025	0.975]
intercept	0.0208	0.021	1.004	0.317	-0.020	0.062
GDP	0.1497	0.019	7.915	0.000	0.112	0.187

Omnibus:	30.579	Durbin-Watson:	0.322
Prob(Omnibus):	0.000	Jarque-Bera	(JB):45.002
Skew:	-1.151	Prob(JB):	1.62e-10
Kurtosis:	4.611	Cond.	No.2.57

TABLE IV: The co2 to GDP regression after the decoupling

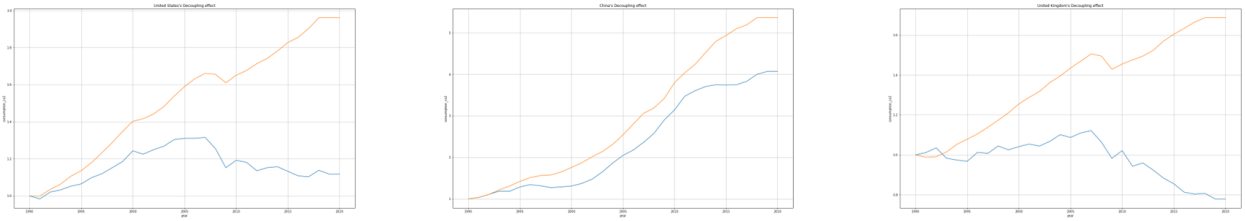
From both tests, it is apparent that the hypothesis is proven right. In the first model the slope is slightly greater than 1 and in the second is less than 1 and greater than 0. The coefficients seem to be statistically significant at a level of statistical significance a=0.05. The statistical significance of the models seems to be for both cases within the confidence interval.

5) **The Case of More than One countries**

How does the same notion work in other countries? In the following graph (see figure **??**) some of the major economic superpowers are present.

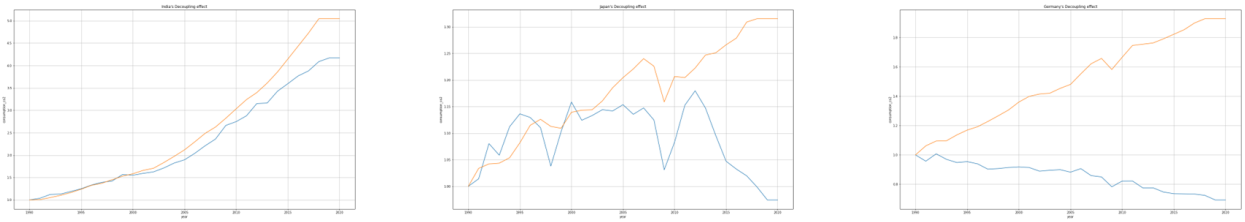
Fig. 12: Major economic superpowers in comparison co2 per GDP ratio

Some of the major superpowers that underwent heavy industrialization near the beginning of the industrial revolution are moving in a direction, where the efficiency of their production has caused less co2 emissions per unit of GDP. It can also be seen that some major historical tragedies have left their mark on the productive capabilities of these countries, such as the Great Depression of the 1930s, World War I, and World War II. There are, though, 2 superpowers that don't seem to fit the same pattern. India and China seem to have had a slight upward trend in the latest decades, especially after the 1970s. This can be further proved by the following graphs. :



(a) United States (b) China (c) United Kingdom

Fig. 13: United States, China and United Kingdom's Decoupling: the blue line depicts the consumption-based co2 emissions, and the orange line the gross domestic product



(a) India (b) Japan (c) Germany

Fig. 14: India's, Japan's, and Germany's Decoupling: the blue line depicts the consumption-based co2 emissions, and the orange line the gross domestic product

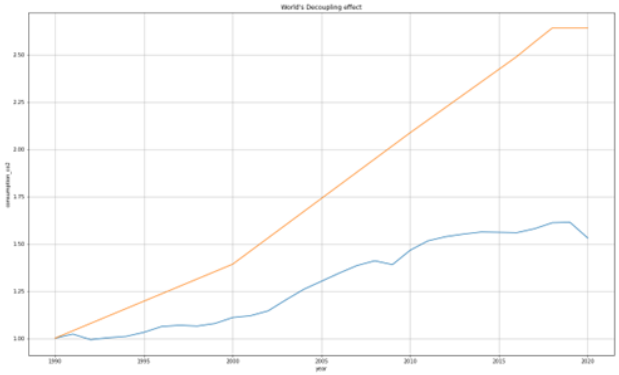


Fig. 15: World's decoupling: the blue line depicts the consumption-based co2 emissions and the orange line the gross domestic product

In figures 13, 14 and 15, 1990 was taken as a benchmark year. In the cases of China and India, there is still a strong correlation between how the GDP grows and the consumption-based co2 emissions. The rest of the superpowers seem to undergo a period of decoupling. In the case of Germany especially this is quite evident. The reason this happens could be that countries like China and India possess still growing industrial sectors, combined also with the fact that the industrial center of the world has moved from the west to the east. Many fabrication units have moved from Europe and North America to countries with lower operation costs like China and India.

6) **What is the industry that affects the co2 emissions to the world? (coal, flaring, gas, oil, cement, and other industries)**
The fig **??** shows the countries in history that have emitted the most greenhouse gasses concluding that Russia, United Kingdom, Germany, and Ukraine are countries that despite total co2 emissions were counted, here the impact on the

ozone layer is most notorious, also in the figure,b and c it's shown that the industries that emitted the most co2 in total are coal with a 46.3%, oil with 32.13% and gas with 14.35% of total emissions of co2 by industry respectively.

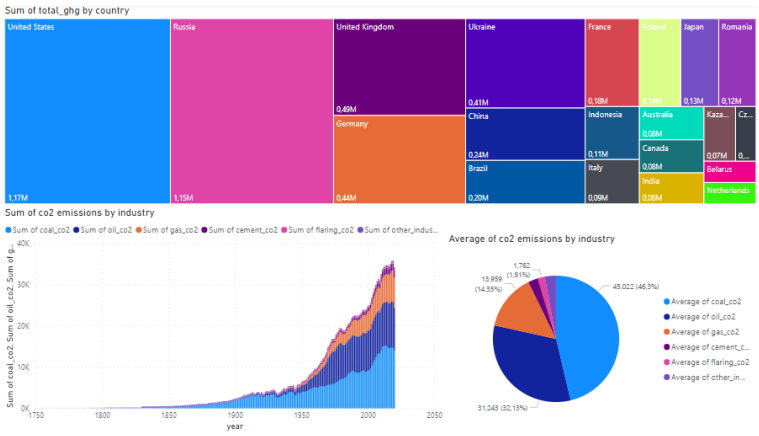


Fig. 16: a. Treemap of the total sum of greenhouse emissions by country
b. Stacked column chart of the sum of the CO2 emissions by industry through history
c. Pie chart of average co2 emissions by industry

The fig **??** shows the countries that produce more co2 by industry in terms of oil and gas United States has produced 234,7 thousand of Tones of co2 and in terms of coal and cement China has produced 192 thousand of Tones of co2 in total history; being both the most higher emitters of co2 by industry

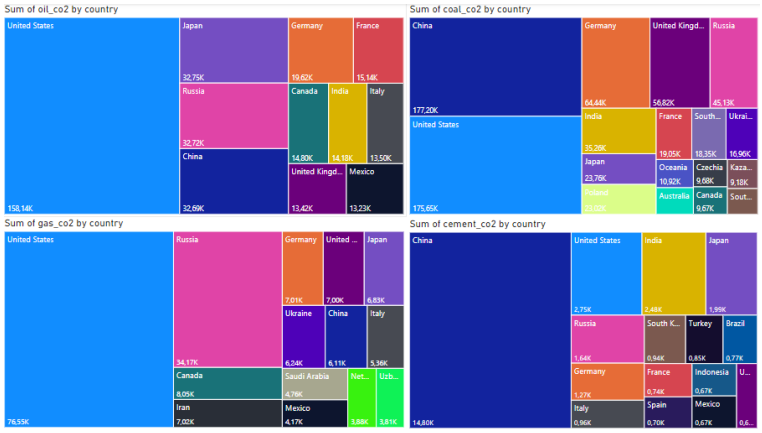


Fig. 17: Treemap of industrial production of CO2 emissions by country a. Oil, b. Coal, c. Gas, d. Cement

Global share of CO2

Since the industrial revolution in the early 1800, through the great wars that were fought around the world, it has reached a tipping point where the percentage of CO2 in our atmosphere can no longer be ignored. The major players are the USA, UK, Germany, Japan, China, and India, which account for the Global cumulative CO2. Most of the heavily industrialized countries have their major share in CO2 emission, regardless of the preventive measure that they have advocated reducing the pollution.

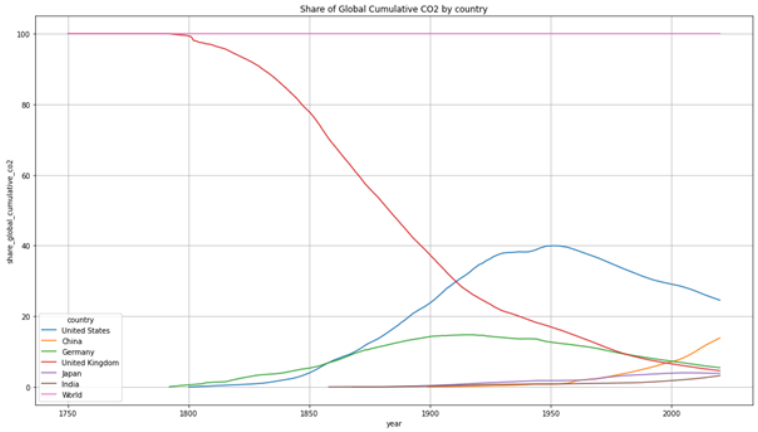


Fig. 18: Graph indicating the share of global co2 emissions from top economy countries

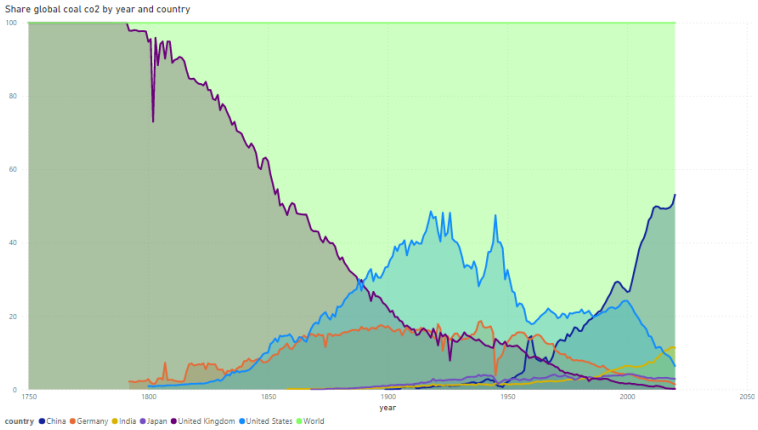


Fig. 19: Global share of coal-producing CO2 emissions by top economic power countries

Considering the global share of the CO2 from coal it's not other than the United Kingdom, takes the top place, while the dependency on coal in the UK is on a steady decline and their policies for a greener future are commendable, 2 centuries of heavy coal use did have its effects on the global CO2 footprint.
China is another key player since they produce to satisfy the global product demands, the heavily energy-dependent industries have no much choice, as the energy deficiency gap cannot be bridged with renewables alone.

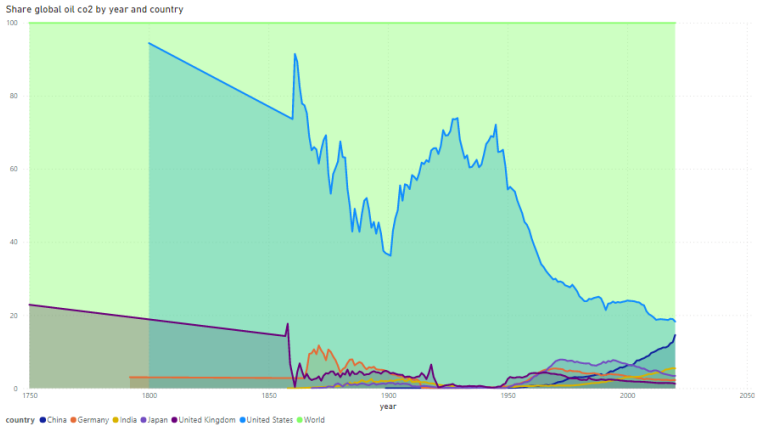


Fig. 20: Global share of oil-producing CO2 emissions by top economic power countries

Figures 19 and 20 shows also a complementing analysis of the co2 emissions provoked by the giant's economies that in addition to the emissions that already had in their place the pollution generated by the economic trade is also a reality and highly huge.

IV. Conclusions

- If the use by people in general of the top three industries that affects the most CO2 emissions coal, oil, and gas decreases the emissions in total CO2 would have a tremendous impact. Coal was tremendously used by countries to produce electricity and in the case of China, as is shown in the figure **??**and **??** is still a big part of its economy, the implementation of more strict polities around the use of fossil fuels and the incentives of renewable energy help in this situation. For the case of oil, the fastest and the nearest solution is avoiding the use of vehicles while transitioning to electric ones. Gas is mostly used for heating and a transition for warming with technology also implemented with the use of electric sources can help to decentralize the use of gas for heating and use different technologies that help the global environment.
- According to the results found in the figures **??**and **??** the possible solutions that really could help depends on the hands of just a few countries that have contributed the most to this problem.
  1) United States
  2) China
  3) Russia United Kingdom
  4) Germany
  5) Ukraine
  6) Japan
  7) Brazil
  8) India
  9) Canada
  10) France
  11) Poland
  12) Italy

  If the countries on the list implement more green transitions in politics according to their economy France's treatment of 2050 can be a reality. As the figure **??**shows it's possible to have an increase in economic growth without affecting the global environment in the process.

References

[1] H. Ritchie, M. Roser, and P. Rosado, "Co and greenhouse gas emissions," *Our World in Data*, 2020. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions.
[2] "Linear interpolation," *Wikipedia*, 2022, August 18. https://en.wikipedia.org/wiki/Linear_interpolation.
[3] "List of countries by gdp (nominal)," *Wikipedia*, 2022, November 1. https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal).