**Assignment 1:**

**Final Assignment**

**For** **exploration of a large global medical dataset**



**Prepared by:**

***Team Number:***

***Team Member 1*** ***Vu Kim Thanh (Stephen) 10648771***

***Team Member 2 Jie Wang N8663246***

***Team Member 3 Yuheng Zhang N10320920***

***<30/08/2020>***

**Abstract:**[**¶**](http://localhost:8888/notebooks/OneDrive%20-%20Queensland%20University%20of%20Technology/covid_explore.ipynb#Abstract:)

**Background:**

In March 2020, the World Health Organization (WHO) declared COVID-19 a pandemic, caused by the novel SARS-CoV-2 virus. Following the call from the WHO to immediately assess available data to learn what care approaches are most effective and evaluate the effects of therapies, this collection aims to report on original peer-reviewed research articles in methodological approaches to medical research related to COVID-19. As stated in NEXOID's official webpage, this data has been accumulated from their web-based self-reported survey since 25th of March 2020.

**Method:**

Extracted from the raw available data from NEXOID-UK of global respondents reporting their COVID-19 status along with other medical conditions. Such comorbidities as asthma, liver disease, lung disease, diabetes, kidney disease and hypertension will be discussed with other daily living aspects to provide the clear understanding of what elements cause the rash of the epidemic. This paper will introduce an informative outlook of the viral disease by applying Exploratory Data Analysis, in addition to optimizing logistic regression and association rules on given risk facts as independent univeriate and multivariate variables

**Result**

The result of our exploration will be demonstrated in the Section 4 (task 4) in this report. The final findings are essential to the future feature selection that might accelerates the data modeling for next assignment.

1. **Examine data types in assigned dataset for each variable and modify the ones needed to correspond with the given description**

survey\_date 5111 non-null object

region 1485 non-null object

country 5109 non-null object

ip\_latitude 5111 non-null float64

ip\_longitude 5111 non-null float64

ip\_accuracy 5111 non-null int64

sex 5111 non-null object

age 5111 non-null object

height 5111 non-null int64

weight 5111 non-null int64

bmi 5111 non-null float64

blood\_type 5111 non-null object

insurance 3943 non-null object

income 3943 non-null object

race 3943 non-null object

immigrant 3943 non-null object

smoking 5072 non-null object

contacts\_count 5065 non-null float64

house\_count 5111 non-null int64

public\_transport\_count 3943 non-null float64

working 5065 non-null object

worried 3943 non-null float64

covid19\_positive 5111 non-null int64

covid19\_symptoms 5111 non-null int64

covid19\_contact 5111 non-null int64

asthma 5111 non-null int64

kidney\_disease 5111 non-null int64

liver\_disease 5111 non-null int64

compromised\_immune 5111 non-null int64

heart\_disease 5111 non-null int64

lung\_disease 5111 non-null int64

diabetes 5111 non-null int64

hiv\_positive 5111 non-null int64

hypertension 5111 non-null int64

other\_chronic 5111 non-null int64

nursing\_home 5111 non-null int64

health\_worker 5111 non-null int64

risk\_infection 5111 non-null float64

risk\_mortality 5111 non-null float64

By comparing to the dataset description provided, such variables are considered for datatype correction: Survey\_date, Contacts\_count, Public\_transport\_count,Worried. To be more specific:

* 'survey\_date' as described in the assessment specification: it is a collection of dates recording data from the website, therefore, the most suitable datatype for this variable is Datetime.
* 'contacts\_count','public\_transport\_count','worried' should be converted to integer datatype due to these factors indicates the number of people and the level of individual concern about the pandemic. Hence, based on conventional logics, it should not be in 'float' type.
* Note that 'Int64' is a nullable datatype for integers.

1. **Using suitable statistical measures and functions:** 
   1. **Identify and report the skewness present in the variables**

One thing to look at two data sets and to decide that one is either symmetrical or asymmetrical. One measure of skewness, called Pearson’s first coefficient of skewness, is to subtract the mean from the mode, and then divide this difference by the standard deviation of the data, as displayed in below formula:

C:\Users\n10648771\Downloads\pearson-skewness.jpg

After imputing this formula on numerical univariate variables in the dataset, the results are shown as following:

|  |  |
| --- | --- |
| ip\_latitude -2.087471  ip\_longitude 1.652506  ip\_accuracy 3.157092  height -0.444816  weight 1.203467  bmi 2.205616  contacts\_count 1.001026  house\_count 1.777352  public\_transport\_count 6.091561  worried -0.301183  covid19\_positive 0.557063  covid19\_symptoms 2.354682  covid19\_contact 2.369179  asthma 2.284441  kidney\_disease 7.964242  liver\_disease 10.175924  compromised\_immune 3.605102  heart\_disease 4.768362  lung\_disease 6.159167  diabetes 2.988074  hiv\_positive 15.508909  hypertension 1.652573  other\_chronic 3.341092  nursing\_home 10.401023  health\_worker 3.653815  risk\_infection 0.406947  risk\_mortality 13.180574 |  |

“The values for asymmetry and kurtosis between -2 and +2 are considered acceptable in order to prove normal univariate distribution”.

*(George & Mallery, 2010). George, D., & Mallery, M. (2010). SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update (10a ed.) Boston: Pearson*.

As a general rule of thumb:

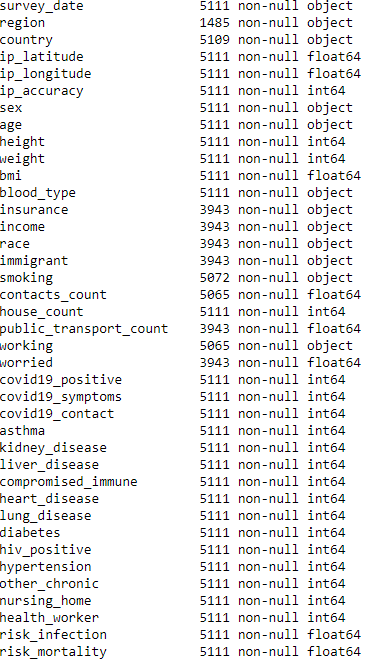
* If skewness is less than -2 or greater than 2, the distribution is highly skewed.
* If skewness is between -2 and -1 or between 1 and 2, the distribution is moderately skewed.
* If skewness is between -1 and 1, the distribution is approximately symmetric.

Our paper will only visualize the highly skewed distribution variables to have the general view of dataset. As only ip\_latitude satisfies the requirement of less than -2, however and ip\_latitude with ip\_longitude should not be appropriate for verifying distribution due to its special meaning, we will only include 17 variables that is considered as Positive Skewed Distribution**: 'ip\_accuracy', 'bmi', 'public\_transport\_count', 'covid19\_symptoms', 'covid19\_contact', 'asthma', 'kidney\_disease', 'liver\_disease', 'compromised\_immune', 'heart\_disease', 'lung\_disease', 'diabetes', 'hiv\_positive', 'other\_chronic', 'nursing\_home', 'health\_worker', 'risk\_mortality'.**

* 1. **Identify data errors and inconsistencies**

Data errors include **outliers, null values, discrepancies, noises**. For each data issues, we need to compute different techniques to correctly identify the problems in order to provide effective solutions for future model building

From the computed table, we can see that 'Region' has the least filled values with only 1485 while other missing-data columns share the common figure of 3943 replies. 'Country' also is incompletely surveyed with 2 entries left empty.

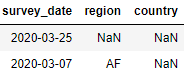


However, after examining the given CSV file, we detected that 'region' column is mis-intepreted by the computer's algorithm since the NA values (North American) is mistaken for Null values. To clarify our assumption, we use 'unique' function:



As you can see, the output for Region's unique values did not include North America, we need re-fill the values. But be careful, from the NEXOID's Covid19 Calculator, we can only input countries after we input our REGION values. Therefore, number of filled entries of Region column MUST be equal OR higher that of Country column (5109 to 5111 entries).

First we need to detect how many actual **missing values** of REGION based on COUNTRY column:



We can see that only 1 entry of REGION had not been filled in the original dataset (entry no 1529). The 1 missing point is too marginal to affect the general data points, therefore, we will fill it with string NA later on in Sub section 6.2.

We attempted to detect **data inconsistency** by doing 2 things:

* We try to use histogram and countplot (using df.hist() and sns.countplot()) for these variables to identify if there are any inappropriate values in the data columns. And we found that all entries are properly filled
* We initially suspect that the BMI values might not consistently be calculated from the Height and the Weight. But ultimately, it is correctly imputed.

Therefore our conclusion is that the dataset is perfectly consistent yet.

The final data issue that our team was aware of is "noise" or 'outlier'. As we can see in df.describe(). The max for height and accuracy is prominent to detect the outliers when comparing with mean and min. However, there should be a more appropriate approach to ensure we are not leave out any actual false values.

The approach to this process is simple, we will apply Z-Score method to identify outliers: For each variable, if the Z-Score of individual values is not included in the range from -3 to 3, we will consider that as outliers. With specific computation, the list of total outliers is displayed as below (‘ip\_latitude’, ‘ip\_longitude’ have been removed).

However, a reminder of the Z-Score is that it is unable differentiate binary variables (0 and 1 value) from normal continuous variables. Therefore, it also cannot be included in the list.

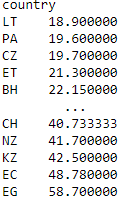
|  |
| --- |
| height : 45 outliers  weight : 76 outliers  bmi : 81 outliers  house\_count : 79 outliers  public\_transport\_count : 84 outliers  risk\_mortality : 53 outliers |

**What is the BMI patient from the US who is COVID-19 positive**

In order to get the value of US's patients, we need to group by the values of country and bmi where countries equals US and we also need to make sure that the value of covid19\_positive is 1 because we are looking for the infected people.

The average BMI of US is 30.9 which is considered to be overweighted. This might show a correlation between chronic diseases and the pandemic infection in the future exploitation.

However, on the global scale, it can be considered as average since the highest BMI is found in Egypt which nearly doubles the figure of US.



***Which country has the highest (average) risk of mortality:***

The risk of mortality rate in all countries are calculated by indexing individual nations and compute the mean of each subset. To identify the nation with highest risk, our team will rank the output in the descending order:

country

NZ 3.143556

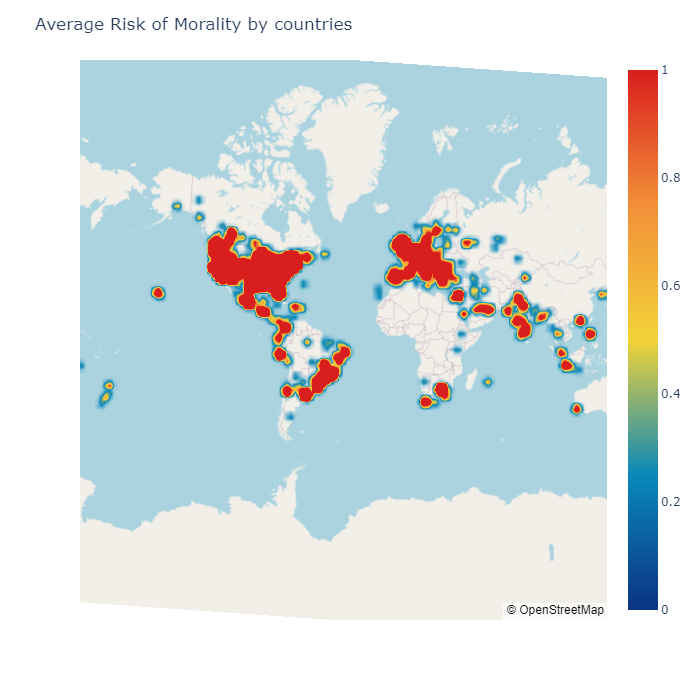
AM 3.291000

PT 3.603583

AD 3.994000

IT 10.031844

As we can see from the table, Italia has the most extreme risk of fatality with 10.03, followed by AD, PT, AM and New Zealand. Just for a clearer justification, we introduce a small map visualization along to demonstrate how tragic the situation is.



###### **How many countries with at least 10 positive cases?**

Categorizing the positive COVID19 cases by countries is more complex, we need to select the entries where column COVID19\_POSITIVE's values equal to 1 (which means active), then we need to group by nations and count the total number of it. Those countries are: Canada, USA, Belgium, Brazil, Chile, Spain, Italy, Mexico, United Kingdom, India, South Africa and Pakistan.

**Top 5 nations with highest number of positive cases:**

From the above dataset, we can easily now rank countries with highest value of positive cases:

Top 5 countries with highest number of cases:

US 1222

BR 154

GB 127

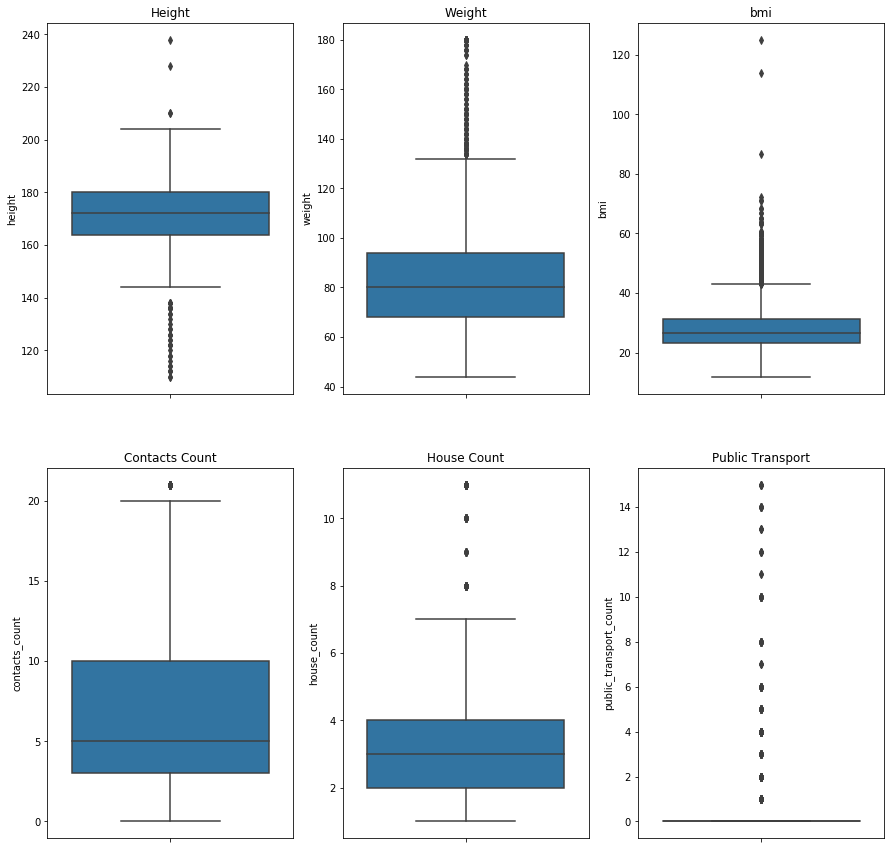
CA 51

IT 32

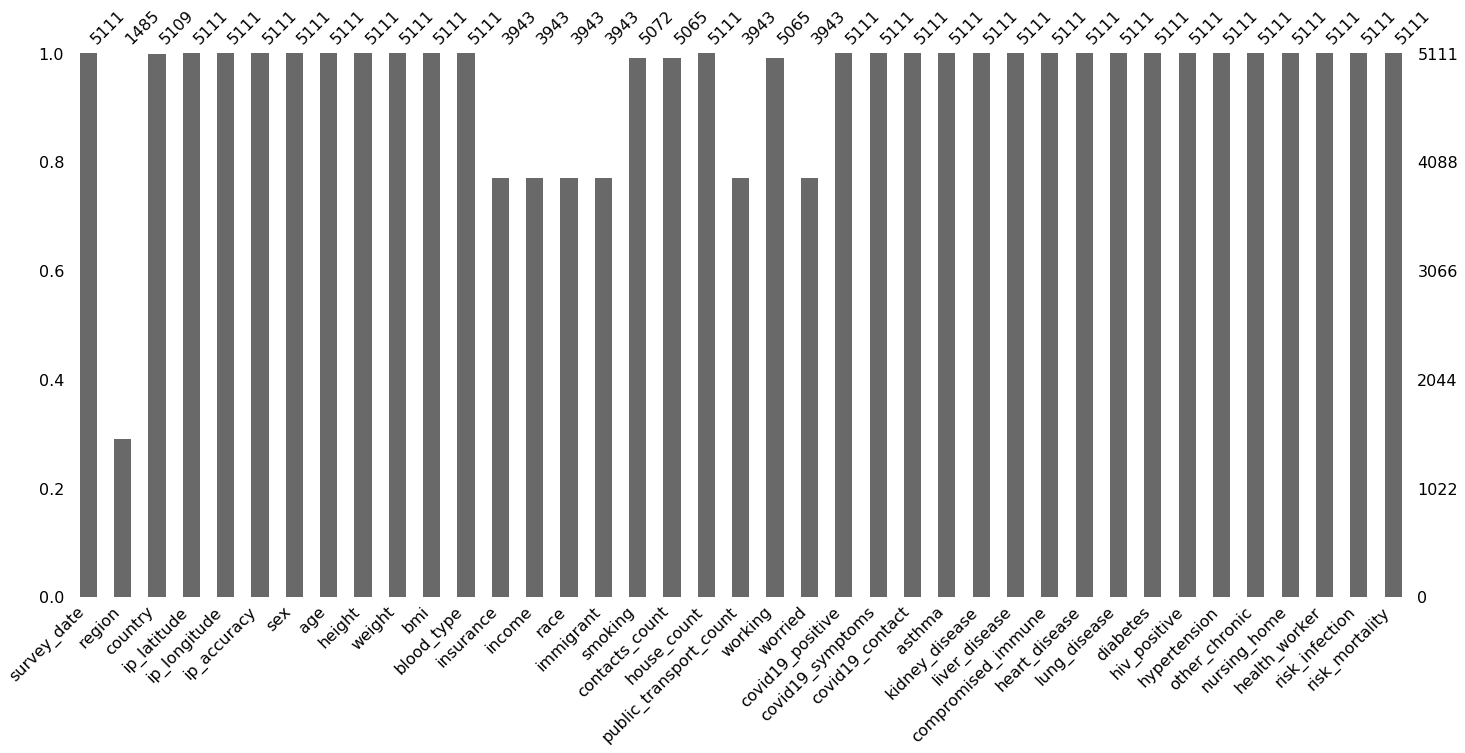
As can be inferred from the table, US is by far outripping following countries ( more than 9 times to the second place), proving an intensifying tension of the situation in the strongest power of the world. Moreover, the list above consists of countries only from Europe and North America, while in reality, it is believed that the pandemic is damaging harder with higher intensity in Asian countries namely Korea, China and Japan. Therefore, the credibility of this dataset should be re-contemplated.

1. **Performing suitable visualization for variables:** 
   1. **Understand the distribution of variables and identify data problems**

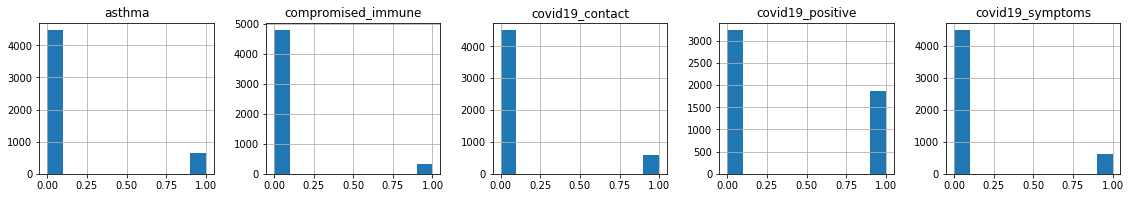
For continuous variables, it is advised to perform boxplot to fully view the distribution of values in each variables, we can get the overall insights as max, min, median and especially Outliers. In this case study, We will eliminate 2 continuous factors: longitude and latitude. Our team only wants to understand the dispersion of other 6 factors

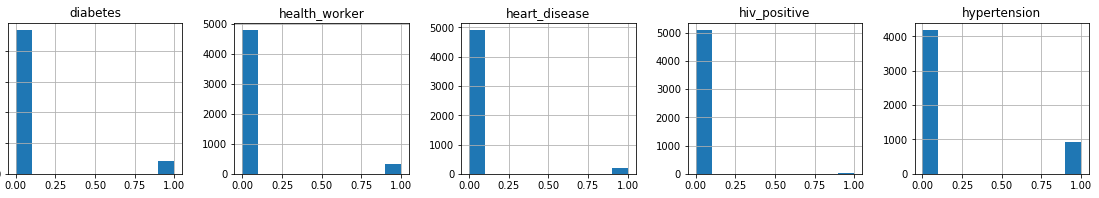


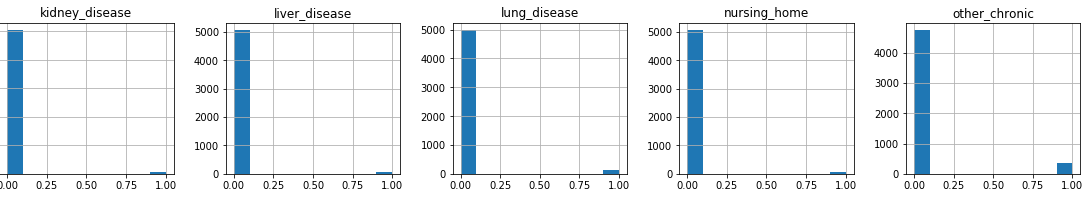
From the computed barchart, we can see that 'Region' has the least filled values with only 1485 while other missing-data columns share the common figure of 3943 replies. 'Country' also is incompletely surveyed with 2 entries left empty. But as we explained in the previous section, ‘Region’ is only the problem of mis-interpretation by the computer.



Histograms of below variables show that, the majority of these factors have the values of 0, meaning they show the negative results for all survey questions. The column has the highest number of '1' is unfortunately Covid19\_positive, this dictates that other medical conditions do not substantially impact the possibility of being infected by the pandemic.





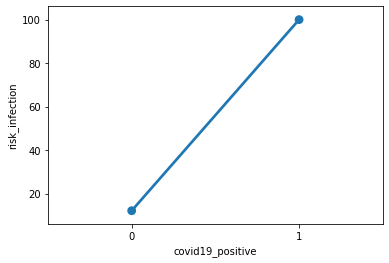


**There is no duplication as be seen in the dataset.**

###### **Correlation between 'covid19\_positive' and 'risk\_infection'**

Before examining the relationship of the two data, our team wants to state that: **correlation is a measure of the linear relationship between two variables**. To understand the existence of mutual dependence of these, we firstly need to categorize the variables: 'covid19\_positive' is a series of **binary outcomes** (0 or 1) stating the infection status of a person and 'risk\_infection' is a **continuous** variable having values ranged from 0-100 to measure the possiblity of being contracted with the virus. Our approach will clearly be demonstrated based on the calculation of point biserial algorithm and be illustrated with the help of scatter plot. Point biserial is a powerful tool, as stated in scipy documentation:

* "**The point biserial correlation** is used to measure the relationship between a binary variable, x, and a continuous variable, y. Like other correlation coefficients, this one varies between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply a determinative relationship." (Scipy, 2020). A point plot is ideal to illustrate this assumption.



We will dispute the coefficient of the 2 variables by using this formula:



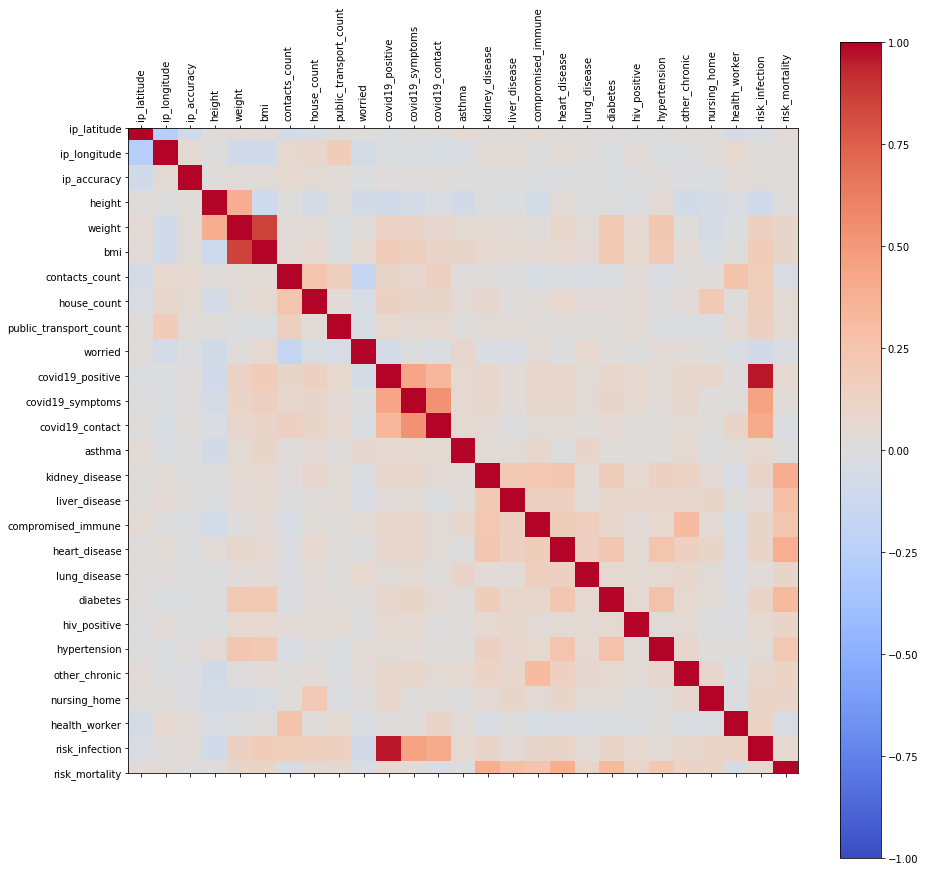
##### *The correlation coefficient outcome is 0.96, showing an extremely high correlation between these variables. As we see from the poiint plot above, the line represented for the relationship of 2 given variables is a straight line, therefore it is certainly a linear relationship. Nevertheless, the case seems distinct with the peculiar distribution of data points. As dictated from the chart, it shows an absolute correlation in which only produces 2 results: either negative with Covid19 or Positive with the virus. Also, it is an one-way relation: If a person is contracted with coronavirus, he/she is ensured to have 100% risk infection, if the rate is lower than that, he or she is believed to be negative with the pandemic test results, however, the vice versa cannot be true: unless the covid19 positive is one, the patient will always be negative to covid19 regardless of higher risk infection. Therefore, this is a linear dependent relationship.*

##### *In the data modelling process, the best algorithm to handle binary outcomes is logisitc regression, however, logistic models normally do not allow linear relationships. Therefore, in this case, we need to use mixed effect regression. This is a strong method to model the probability of binary events (whether a person is definitely infected or not infected). There are multiple ways to deal with the kind of data, but the simplest approach will just be aggregated. To explain our team's strategy of handling these variables, the goal of this report must be restated: Is to decide if a person is contracted with virus or not and within same determining factors, we can predict the future input of respondents of their covid19 test based on the given survey. Therefore, we are allowed to completely remove the column of RISK INFECTION: Because no matter how high of a person's risk rate, he/she will still get negative result of covid19 if the rate is not 100%. So it is no point of measuring their infection possibility.*

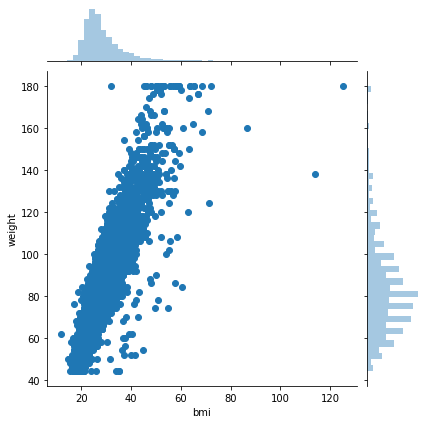
* 1. **Identify the highly correlated variable pair**

The most convenient way to identify a correlation between multivariate variables is to implement the correlation matrix. As indicated below, the darker red colors represents for higher positive correlation and vice versa with the blue color. Hence, we only determine the high correlated pairs if their co-efficient higher than 0.8. As the result, there are only 2 pairs of variables can be considered as "Highly Positive Correlated**":**

* **risk\_infection and covid\_positive**
* **bmi and weight**

****

Our approach to the first pair variables has been discussed above, we only need to view the bmi-weight relationship. Identifying highly correlated pairs is extremely important basic process in exploratory data mining tasks. The strong dependence between factors can speed up the search process. In the general aim of our report, we can use logistic regression or market basket to determine our derived variable (which is "Obesity" - will be discussed later in this report). The further handling approach will also be detailed in TASK 5.



1. **Summarize findings and elaborate on what kind of data preparation is needed.**

So far, we have explored that the dataset stores a great number of useful insights that can be worked for further understanding of the scenario. The findings can be summarized briefly as following:

* The correlations between variables in the dataset are relatively low, stating that most of these variables are independent in the analysis
* The number of people reported in the survey are mostly from US and NA region since more than 60% of data are reported from these places
* Though Italy has the highest number of dead cases, America should be wary of their huge number of the infected.
* The datatype is consistent but contains numerous errors including: more than 500 outliers, missing values in multiple columns and misconducting the datatype of several variables.
* Most of the medical values are 0, meaning that respondents are partly negative with any given conditions, meanwhile, the number of positive cases remains high. From that view, we can consider that the covid status is not strongly related to any comorbidities.

#### What data are needed to correct types?

So far in the exploratory process, everything is working fine. However, when it comes to the cleaning procedure, some features must be modified to facilitate better data preparation.

* contact\_count are currently stored in type of Float64 due to the fact that integer variable cannot have NaN values. This somehow can affect to the insight of these factors when we want to make the Mean computation, the number of objects and human cannot be in float type. Also, in case we decide to impute the missing values with the mean, this is also a problem. Therefore, our team decides that these variables should be converted to Int64 (not int64) - a nullable datatype that can store both null and integers. (We can also apply this on Worried and public\_transport\_count, however, the large amount of missing data makes it unnecessary to impute these variables)
* **Data Preparation methods for this dataset can be named as Drop Columns, Imputation, Aggregation, Drop Rows which will be detailed in section 6 to avoid repetition.**

1. **Selection of data mining task and feature selection:**

##### ***Explain the choice of data mining technique:***

The beauty of data world lies in the variety of techniques and the flexibility to apply either of them. The word "most suitable" might not be correct since, the most important objective in initiating a data mining project is its formulation statement guiding a group of data practicers to keep in one track, only within a clear vision and understanding, a data scientist can decide his/her approach toward a dataset.

Therefore, in order to conclude the choice of technique implementation, our team wants to restate the original goal of this data collection: **is to decide if a person is contracted with coronavirus based on their information inputs.** This says our practice revolves Predictive Analysis. We share the same intention with NEXOID (who formulated this dataset).

Though **Association** can be helpful in track patterns and correlations of variables, Cluster, however, based on our objective, we believe that **Classification** will be an efficient and appropriate tool assisting us to achieve desirable outputs. Within given various attributes in discernable categories, the classification will be facilitated and more ensured. **Regression** for predicting the likelihood of being infected or fatal is also a very good choice, however, calculating the exact rate requires extended numeric variables. And from our analysis, the accuracy of regression outputs is not reliable.

For current plan, univariate/multivariate logistic regression and analyses will be using medical risk factors as independent variables.

###### *2. Feature Selection:*

As team's focus is in Classification technique, we expect to apply Binary Classification to indicate whether a person is contracted with corona and whether he/she will be dead from the virus. The choice is inspired with a number of available binary variables collected in the dataset, all of them are fully filled, enhancing the precision of the mining process. Before listing the essential variables, the report will also introduce some derived variables that our team feel neccessary to be included in our future data mining task.

* First of all, we will formulate a column determining 'obesity' by filtering BMI over 30 and put it in binary value 0-1. Obesity is at epidemic proportions in the United States and in other developed and developing countries. The prevalence of obesity is increasing not only in adults, but especially among children and adolescents. Obesity is also strongly associated with an increased risk of all-cause mortality as well as cardiovascular and cancer mortality. Obesity must be identified to increase awareness of potential adverse outcomes. This will allow health care professionals to identify and implement appropriate interventions to reduce patient risk and mortality. The input for this variable will be from BMI.
* A new derived variable named: mortality\_level - a categorical factor displaying the intesity of covid19 fatality. Due to the fact that nominal variable will make the classification process easier, encoding the risk\_mortality into nominal bins will be a better option. We will divide the rate into different bins and rate them from 0-4.
* Since all ‘count’ columns display the number of interaction in community by people, it will make more sense if we can combine all [contact, public transportation and house count] in the same group. However, with a large amount of public transportation, it will make it hard to combine them with other 2 groups because it might make the model less efficiency. Therefore, we will only group: Contact and House counts. Then we will binarize them as 0 – 1 for contact or no contact lately and named it as RECENT\_CONTACT

*Our list of selected variables for this data mining task includes:*

* BMI: Because on the medical perspective, BMI can represent both Height and Weight to decide a person's fitness condition, moreover, BMI as discussed, is the input of Obese.
* Obese,asthma, kidney\_disease, liver\_disease, compromised\_immune, heart\_disease, lung\_disease, diabities, hiv\_positive, hypertension, other\_chronic, RECENT\_CONTACT, age, sex: A significant medical risk association with different age groups in term of genders needs to be defined to identify if pre-existing comorbidities in different demographics can affect the likelihood of developing the corona disease
* Covid\_positive, covid\_symptom, covid\_contact: risk\_infection will be redundant as explained in Task 3. covid\_symptom and covid\_contact are important to decide if a person is contracted with the virus

1. **Data preparation for mining:** 
   * 1. **Apply min-max and z score**

Normalization is generally required when we are dealing with attributes on a different scale, otherwise, it may lead to a dilution in effectiveness of an important equally important attribute(on lower scale) because of other attribute having values on larger scale. Many machine learning algorithms attempt to find trends in the data by comparing features of data points. However, there is an issue when the features are on drastically different scales.

In simple words, when multiple attributes are there but attributes have values on different scales, this may lead to poor data models while performing data mining operations. So they are normalized to bring all the attributes on the same scale. If the data normalization is not done, the machine learning algorithm in the future has the inclination to weigh greater values regardless of the unit of the values. The misconduct of normalizing data can lead to a lower precision and accuracy of future data modelling outputs.

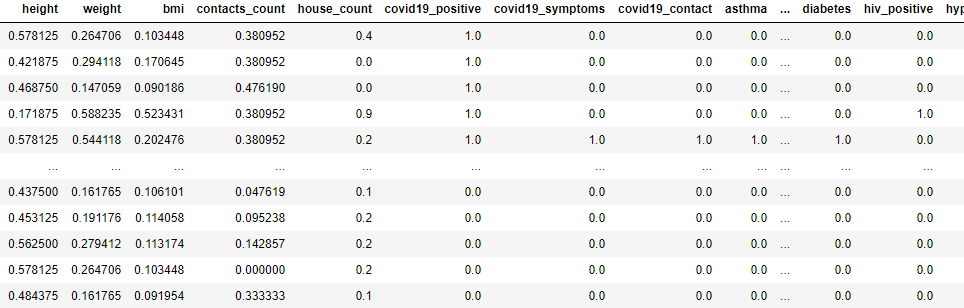
Therefore, when analysing the dataset for normalisation, we need to make sure 2 things:

* The variable has unidentified range with huge gap between variables
* The data should be on the same scale for future combination

All categorical and nominal variables will not be included in this pre-processing stage for the higher precision. Our team will extract a new DataFrame with numerical values first named it as dfr\_quant.

We will apply:

**Min Max Technique**:

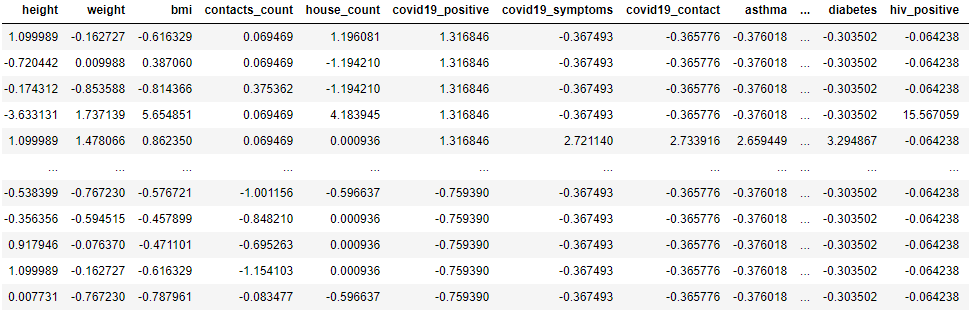


To compute the Min-Max technique, we implemented this function:



The aim of performing this method is to rescale data from multivariable values into the same measurement scale, hence, the computer can rightfully acknowledge the need of prioritizing correct variables when computing its algorithm. And the scale as indicated is from 0 to 1.

**Z-Score Technique:**



To compute the Z-Score technique, we implemented this function:



The range of all computed data seems vague as you see in the box.

**Which method is better?**

**There is no obvious answer of better method since each serves different purposes for different data manipulation, therefore, a problem formulation must be set to align our approach:**

*The aim of this process is to convert all unit values from columns to be in the same scale for more precision machine learning imputation.*

* **Min-Max:** Unable to deal with outliers but re-sample values in the same scale from 0-1
* **Z-Score:** Normalize the range not in the same scale and handle the data outliers,

To further decide the approach whether Z-Score or Min-Max, we also need to be aware of our task objectives: to facilitate our Classification approach in the future. When Z-Score were implemented, it counts the unnecessary outliers of compromised\_diseases, kidney\_diseases and some other binary values. This makes the outlier detection become less precise, thereby weakening the future prediction of our model. Hence, in a dataset like this, Min-Max should be prioritized for better outcome.

* + 1. **Data quality correction:**

Looking at the barchart of missing values, There are missing values in Region, Country, Insurance, Income, Race, Immigrant, Smoking, Contacts Count, Public Transportation Count, Working, Worried. Ratio of missing values in these columns are different, thus the technique of choice for fixing them will be different too.

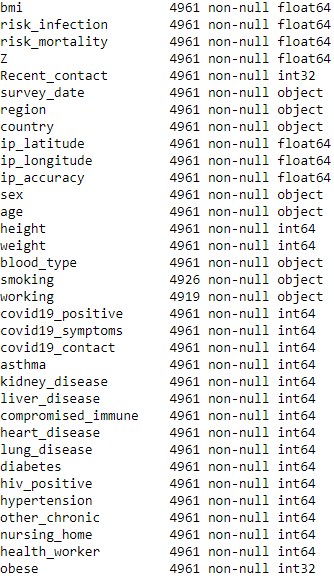
#### a. Handling missing data:

##### ***Before deciding the measure to handle data imputation, we need to fully understand the background of this research.*** *The data is collected from a serial responses from a web-based survey, by accessing to the main websites of NEXOID (The corporation created this dataset: (*[*https://www.covid19survivalcalculator.com/*](https://www.covid19survivalcalculator.com/)*), our team realizes that all of the above missing data columns (except Region, County and Worried) were compulsory to be filled in order to proceed the calculator of COVID19 risk rates. Though NEXOID guaranteed the confidentiality of the participants to be obligated to fullfil all information by giving them the option: "Blank". By acknowledging this procedure will help us to determine whether a variable should be removed or dropped for the future modelling.*

* **Re-import DataFrame:** Firstly, for Region column, 65% of the entry values are missing. However, we detect that it is the fault of Jupyter when importing this, it mistakes string value 'NA' for NaN values. However, if we replace the NaN value for string 'NA', we will automatically assume the value of actual missing data of this variable (this is fine since there is only actual value is missed in the dataframe, but the method will be ineffective for larger dataset with more data errors). We will fix this as we import the file and specify na\_values="",keep\_default\_na=False.
* **Drop rows with missing value**s: After importing the correct form of dataset, we can drop columns of **Smoking, Contact Count, Working, Region, Country** since its low ratio of missing data does not affect significantly the analysis.
* **Drop columns** of **Insurance, Income, Race, Immigrant, Worried, public\_transport\_count** due to their large missing data volumn that can mislead the modelling outcome.

|  |  |
| --- | --- |
| **Data Column Null**  survey\_date 0  region 0  country 0  ip\_latitude 0  ip\_longitude 0  ip\_accuracy 0  sex 0  age 0  height 0  weight 0  bmi 0  blood\_type 0  smoking 38  contacts\_count 45  house\_count 0  working 45 | covid19\_positive 0  covid19\_symptoms 0  covid19\_contact 0  asthma 0  kidney\_disease 0  liver\_disease 0  compromised\_immune 0  heart\_disease 0  lung\_disease 0  diabetes 0  hiv\_positive 0  hypertension 0  other\_chronic 0  nursing\_home 0  health\_worker 0  risk\_infection 0  risk\_mortality 0 |

* **Imputation handling missing data**: To the last 3 columns, we will fill missing values by using the majority of categorical data. However, each of variable has different solution.
* For **smoking**, since most of participants never took cigarretes before, we can assume the missing values have identical answers, we will compute by using mean
* For **working**, we will categorize the working status by **AGE** (because age is a good standard to decide of traveling condition). We will check that which age quantiles has the highest number of null data then we will fill them by sampling the majority working choice in that AGE quantiles.
* For **contacts\_count**, this is a continuous variable, most appropriate imputation will be MEAN function
* Aggragate Contact\_counts and House\_counts because they share the similar meaning
* Binarize the new aggregation column as 0,1.
* Remove outliers. All identified outliers take up a small amount of data rows, hence, we can either resample or remove it. We will remove it, but first we will CONVERT 3 column IP\_lat, ip\_lon, ip\_accuracy to advoid wrong imputation. Then we will remove outliers in all float datatype. Here is our final transformation:



* + 1. **Data Partrition:**

Data partrition is a technique to evaluate the accuracy and confidence of a dataset within the support of machine learning algorithm.

It can be used in both classification and regression or any other supervised learning algorithm.

We need to split our dataset into 2 piles: the bigger one called Train Set – Used for machine building model, and other is Test set to used for testing the validation of the model.

To determine how much of partition we want to compute, let’s rewind to the top of our report.

We see that all missing values (1168 values) are being missed at the same places, and if we re-direct to the official website of NEXOID: those variables are not supposed to leave blank. Therefore, we believe the missing type here is MNAR (Missing not at random). This might be in the intention of creating this dataset, although in our project we dropped those column, using this hint as a clue for splitting data is indeed useful. Hence we will Split our data set into 77.2% for Training set and 22.8% for Test set.