1. **Introduction:**
2. **Literature Review:**

This topic is based on the analysis did by Aiello et al. on the Large-scale and high-resolution of analysis of food purchase and health outcomes. Compare to the traditional food classification and food-related health problem analysis, most of them have limited scale of the datasets and the process of the data takes a lot of time, however the data comes from online will make the research and analysis a lot more convenience. The analysis by Aiello et al comprises some 1.6 billion food items purchased in Tesco stores and 1.1 billion medical prescriptions over all areas of London (Aiello et al, 2019). The analysis contains detailed information derived from some 420m food items purchased from 411 Tesco stores in London in 2015. Information includes average nutrient profiles from baskets of goods purchased as well as categories of goods purchased. And it also contains the public health data such as diabetes and obesity of child and adult in the same year in London. By joining food consumption and medical prescription data together, further analysis can be done afterwards.

Since urbanization is leading the development of most countries, the society’s diet has become a big problem. Unlike the centuries before, people in the city nowadays do not need to worry about hunger and the energy intake is unconstrained. Overnutrition leads to disfunction of human body metabolism and causes many Chronic diseases. Similarly, Milburn (2004) found that how to use the Canadian indigenous people’s nutrition knowledge to solve contemporary health problems. Comparing the food pattern in Africa to industrial cities, medical doctor Denis Burkitt found chronic, degenerative diseases happens less in the Africa (Milburn, 2004). And it mainly caused by the significant dietary difference and he coined the term “Western disease” to describe the diet- and lifestyle-related disease happen to the people live in Western countries and urban area. The disease includes obesity, Type II diabetes, hypertension, coronary heart disease, peripheral vascular disease, varicose veins, diverticulosis, appendicitis, kidney stones, and some forms of cancer. These chronic degenerative diseases usually do not appear until later life. However, these diseases are induced in the early age on younger people by being overweight and obesity.

Being overweight and obesity is defined by the body mass index (BMI). BMI is the most commonly used tool to correlate risk of health problems with the weight and height of the human body. The BMI calculation divides an adult’s weight in kilograms by their height in metres squared. If the index is over 25, the person is defined as overweight (NHS, 2020). And when the BMI reaches 30, the person is defined as obesity. In 1991, the percentage of obesity in the US was 12. Then it nearly doubled in 10 years, 20 percent people in the US were obese in 2001 and it has risen to 36.2% in 2016 (Milburn, 2004). There are also 27.8% people are obese in the UK in the same year. Obesity has a really long-term cost, itself can lead to many health problems such as elevated cholesterol, asthma, arthritis cardiovascular risk. What is more important, in the same period Type II diabetes prevalence increase correspondingly from 4.9 percent in 1991 (Milburn, 2004) to 10.8 percent in 2019 in the US (indexmundi, 2019). Obesity is definitely playing a main role in the chronic degenerative disease. And to prevent the diseases, by studying indigenous nutrition and finding the way of how they structure the food intake. Indigenous nutrition is following their culture and bioregion. Although the indigenous scientist thinks and emerge knowledges in their own “ways of knowing”, there is still strong connection with western science in molecular components of food and hopefully it can be explained in the science way. The Vitamin-C and scurvy story in 1535 is well known in nutrition science, Jacques Cartier and his crew got scurvy during the sea voyages caused by lack of intake in vitamin-c. In early fifteenth century. Inuit and the Chinese avoid scurvy in another way, by intaking sprouted soy beans and various citrus fruits on the sea voyages (Milburn, 2004).

It has shown that various food intake can help human avoid from many diseases, even the lack of a kind of vitamin is harmful. In the indigenous nutrition, the food diet structure is “nutrition rich and calorie limited” (Milburn, 2004). Wild food contains more kinds of nutrition than similar cultivated crop such as calcium, magnesium, iron and vitamin-c. Even the fish and animals have less fat and less saturated fat in the wild. Compare to the traditional salt, the salt that has been processed contains less trace minerals like iron, manganese, zinc and cooper. Since the wild food also contains low energy, people have to eat lots of it to maintain the energy level in the body. It is true that various kinds of food intake will lead to an overconsumption, on the other hand a lot of physical activity leads to a very high energy output to keep the stored energy level balanced. However, the modern diets can be described as “calorie dense and nutrient poor”. For example, the grains that people now eat is only the endosperm part and the food process gets the rid of other parts to make it taste good and easy to absorb by the body. While the germ, the most nutrient valuable part has been discarded. The similar food processes have increased the calorie dense dramatically and only pointing on reaching the calories needs for a human body, but minimised the fibre intake at the same time. Less fibre intake causes the digestion disfunction and many intestinal diseases (Rahati, 2014).

This is similar to the result that found out by Monterio (2009), the nutrition and health issue is not food, nor nutrients, so much as processing. All the foods have been more or less processed in some way. They are classified into 3 groups, group 1 is of minimally processed foods, the processes here only include cleaning, remove the inedible parts, freezing, drying, etc. Fresh meat, milk, grains, fruit and vegetables are usually seen in this group. Group 2 is of products that extracted from whole foods like oils, flours and sugars. These products are always used as additive during cooking the food in group 1. Group 3 is based on the processing of food in group 2 and also called ultra-processed food, for instance confectionery, ice-cream, chocolate, cereals, chips and so on. The ultra-processed food is most designed to be ready-to-eat or at least ready-to-heat. With complex processing based on the food in group 2, products in group 3 are typically branded and followed by many advertisements and marketing strategy. According to the above, these foods in group 3 play the main role of modern diets and it change the habit of the way that human intake energy every day. There are two important factors has been mentioned by Monterio, one is nutrient density (nutrient per energy unit) and food energy density (energy per volume). The chronic degenerative diseases like obesity, dyslipidaemias and hypertension, are usually caused by consuming foods with low nutrient density on protein, trace mineral and various vitamins, or high nutrient density on saturated fat or sodium which comes from the processing, or both. Traditional diets mainly made up from the group 1 food with minimal processes and low level of energy and salt, so the massive amount brings the energy density down. Since the foods in group 2 are mainly additives and it is used to help cooking minimally processed foods or produce ultra-processed foods, it cannot be the only or main ingredients in any diets. This explains the problem with modern diets that include many ultra-processed foods in group 3 and these foods usually contain several of the unhealthy features of the group 2 ingredients which are low nutrient density: little dietary fibre, and excess simple carbohydrates, saturated fats, sodium, and trans fatty acids. And they also have a high level of energy density, it makes the modern diets intrinsically nutritionally unbalanced and cause chronic diseases to a human body. By solving this problem, more processing has been introduced. More ‘premium’ processed products appear in the market, which are usually the ultra-processed food with less fat or no trans-fat, less sugar, less salt, or with more micronutrients. The products with less fat may also reduce the flavour on the taste, so it may add more sugar to balance the taste then maintain the sales. The drinks may reduce sugar and add micronutrients in the ‘premium’ processing, while even it reduces the sugar or sodium, fat in the foods, the calories in the drink and foods still stay in a high level compare to the normal level of drinks and foods. It does not make these foods healthier but only let people think they are healthy. A part of ‘premium’ foods is truly positive by increasing the content of whole foods; however, they are higher in the prices and it leads to less affordable consumers to consume them. All of the changes above from the traditional diets finally induce the unhealth dietary patterns. And these disorder patterns induce the overweight and obesity. More and more researches show that energy-dense food may fool the biological mechanisms responsible for satiety responses and the body will over consume energy and store the fat (Mattes, 2006; Ludwig, 2008). So, concern from the point of view of human nowadays health, people should choose minimal processed foods and avoid the ultra-processed foods, at least minimise the consumption on them, rich the nutrients intake and reduce the total energy at the same time.

In the paragraphs above, the relationship between obesity and diabetes has been discussed. The same idea has been pointed out by Rahati et al (2014). which they said the obesity is the main inducement of Type 2 diabetes increasing rapidly worldwide. Besides, food pattern and lifestyle has also been mentioned in this article. They also emphasized that dietary habits in the developed and developing countries have affected the modern lifestyle and change towards an unhealthier direction. The There is need for encouragement at individual and population levels for getting back to a healthy lifestyle which includes dietary habits, to prevent the Type 2 diabetes reaching epidemic proportions in many countries. At this level, the paper is mainly focusing on explaining the metabolism of the human body, the macronutrients in dietary, and the energy consumption of the body in physical activity. Since body fat rate is highly correlated with type 2 diabetes, BMI, as mentioned in previous paragraph, is still a common and useful index to measure the risk of prevalence of diabetes. According to the study suggestions, BMI less than 21 has a lower risk of diabetes, and higher BMI have higher incidence rates of type 2 diabetes mellitus in earlier age groups. This unhealthy situation is happening more and more in the recent decades. Other studies suggest waist circumference or waist-to-hip ratio is better than BMI to do the predictions because this ratio is more likely to show the human body fat. Excess abdominal fat is related to the risk of cardiometabolic disease. Or with more advanced technology, measuring the risk through CAT scan of the intra-abdominal fat is the best predictor.

Poor lifestyle and diet can prevent modern people get back to a “healthier mode”. Physical inactivity is one of the problems; doing exercise correctly and efficiently is an important way for people who are overweight to lose fat and people with normal body fat to remain in an energetic lifestyle and keep a fine body metabolism. A series of planned and repeated activities of skeletal muscles can help with the energy consumption. The basis of efficiency exercise should have a duration about 30 minutes, starting with 5-10 minutes warm-up and ending with recovery exercises. Activities under this training volume can help people in regulation of blood glucose, metabolism of proteins and fats, improve the insulin action, so that to prevent the prevalence of diabetes. People should plan the exercises at least 3 times a week but take care when exceeding 5 times a week (Rahati et al, 2014). Even though the activities are needed, let the body rest, recovery, and intake correct foods are more important.

There are 3 macronutrients in all foods: carbohydrates, protein and fat. The first thing comes up is carbohydrates and it is the indispensable nutrient for human body. A study from Richard et al (2001). has suggested that the ratio of carbohydrates against fat has influence with the incidence of chronic diseases. To prevent it, he suggested taking more carbohydrates and reduce the fat intake. However, other studies reported increasing carbohydrate intake is not related to prevent chronic diseases (Parks, 2000). They make the point clear, the total carbs intake does not associate with diabetes risk, but the ratio of carbs to fat does have effects. The data in these studies showed increasing carbohydrate also increasing the secretion of insulin to maintain insulin homeostasis. On the other hand, the secretion of insulin output may be also affected by the ages, which means that output falls when the body gets old and diabetes become riskier. Also, fibre, which can create gel-like substance in human’s stomach by several metabolic hormones can help maintain the insulin secretion to a steady level then improve the control on glycaemic load which is directly linked with type 2 diabetes. So, the best suggestion from WHO/FAO is getting at least 55 percentage of energy intake on rich fibre carbohydrates, like all kinds of grains as the sources (Krauss, 2000; Joint FAO, 1998). Fat, saturated fat in this case, is playing the role of lowering insulin binding to its receptors, a high level of saturated fat diet may cause glucose intolerance and transport. Therefore, polyunsaturated fatty acids and monounsaturated fatty acids are recommended by replacing the daily saturated fatty acid intake, they can be easily found in vegetable oil, nuts, and fish. Unsaturated fats can improve glucose tolerance and reduce the risk on getting chronic disease. It is worth mentioning that 3 micronutrients have high correlation with diabetes. Vitamin E is an antioxidant vitamin and reduced this vitamin level in the body will increase the risk of chronic disease. Magnesium can be found as a component in the shell of cereals and it is strongly correlated with type 2 diabetes with a negative correlation. Chromium helped people with mild glucose intolerance as a supplementation to improve the tolerance and decrease the blood levels of insulin. The decrease on blood insulin levels means the tissue sensitivity gets enhanced. The Mediterranean Diet is recommended in this paper (Rahati et al, 2014) which has more details compare to the indigenous diet (Kastorini, 2010; Schroder, 2007). It encourages people eat fruit, vegetable, low fat dairy products and whole grains every day. And consume fish, poultry, tree nuts, legumes weekly, red meat monthly, as well as a moderate consumption of alcohol. It is clearly to see that Mediterranean diet is more like a guide to food intake which emphasizes the unsaturated fat as the food like fish and nuts are in a higher consumption frequency than the red meat.

What is the proportion of macronutrients other than fat should a person consume normally? Venn (2020) said high carbohydrate diet is used to cure type 2 diabetes, normal people are recommended to consume 50-55% carbohydrates, also mid-age people need to intake around 6% protein to prevent Alzheimer’s disease and this should increase to 17% when the age increases. The dietary intake for protein has been mentioned by Okreglicka et al. (2015) which should be 0.8g/kg body weight, unless athletes need to consume more protein, so does the elder people. However, in this paper rich protein diets (20-23% energy) has significantly improved the metabolic control in patients with type 2 diabetes. In obese women, low-calorie diets rich in proteins increase sensitivity to insulin and prevent muscle loss. A protein-rich diet which has protein as 25% of energy intake efficiently reduced blood pressure. There is insufficient scientific evidence to recommend an upper limit of protein consumption, but consider other two macronutrients, the range of protein is better to be 10-35%.

Overall, the papers have pointed out several diets which is healthy and trying to get people in modern life away from chronic diseases. There is analysis on the traditional and processed food, also the results on what nutrition that human body is needed every day. A proper diet for people nowadays should be not only limited in the total calories but also select various types of food. When selecting food, people should avoid the ultra-processed foods since these foods have high energy density which is against the diets and may cause the chronic disease in a long-term consumption. About the distribution of 3 macronutrients, low fat and rich protein will always be a good choice. Unsaturated fat is healthy to the body and should be included in the diet in preference to saturated fat which has probability on increasing the risk of chronic diseases.

Consider the environment of urbanization, this thesis is going to integrate the findings above to create a model for classifying the food in daily life. Since there are more and more processed foods and the data is now easier to acquire due to better labelling. The score of the food can be calculated by the consumer with the number of nutrients on the food label. Consumer can find out if the food is healthy to their body by the result and make decision on buying or check the daily dietary. This score may change through different ages because the metabolism for different age group changes. Besides, more factors can be added into this model which may have correlation with people maintaining health, for example weekly exercise, family income, education, and the present body situation, etc.

**3. Methodology:**

**3.1 Data Overview**

**3.1.1 Data source**

The basic dataset that has been used is the Tesco Grocery 1.0 dataset, uploaded by Aiello et al. in 2020 after his Large-scale and high-resolution of analysis (Aiello et al, 2019). This dataset is the aggregation data of the record of 420M food items purchased by 1.6 million membership card owners. These members usually shopped at the 411 Tesco stores in Greater London all over the year of 2015 during this course. For each level of area, the dataset provides the number of transactions and nutritional properties of the typical food item bought including the average caloric intake and the composition of nutrients. The data chose to process was the average all over the year instead of the record in each month in borough level. Here are the explanations of each column in the grocery dataset.

|  |  |
| --- | --- |
| **area\_id:** | identifier of the area |
| **weight:** | Weight of the average food product, in grams |
| **volume:** | Volume of the average drink product, in litres |
| **energy:** | Nutritional energy of the average product, in kcals |
| **energy\_density:** | Concentration of calories in the area's average product, in kcals/gram |
| **{nutrient}:** | Weight of {nutrient} in the average product, in grams. Possible nutrients are: carbs, sugar, fat, saturated fat, protein, fibre. The count of carbs includes sugars and the count of fats includes saturated fats |
| **energy\_{nutrient}:** | Amount of energy from {nutrient} in the average product, in kcals |
| **h\_nutrients\_weight:** | Diversity (entropy) of nutrients weight |
| **h\_nutrients\_weight\_norm:** | Diversity (entropy) of nutrients weight, normalized in [0,1] |
| **h\_nutrients\_calories:** | Diversity (entropy) of energy from nutrients |
| **h\_nutrients\_calories\_norm.** | Diversity (entropy) of energy from nutrients, normalized in [0,1] |
| **f\_{category}:** | Fraction of products of type {category} purchased. Possible categories are: beer, dairy, eggs, fats & oils, fish, fruit & veg, grains, red meat, poultry, readymade, sauces, soft drinks, spirits, sweets, tea & coffee, water, and wine. |
| **f\_{category}\_weight:** | Fraction of total product weight given by products of type {category} |
| **h\_category:** | Diversity (entropy) of food product categories |
| **h\_category\_norm:** | Diversity (entropy) of food product categories, normalized in [0,1] |
| **h\_category\_weight:** | Diversity (entropy) of weight of food product categories |
| **h\_category\_weight\_norm:** | Diversity (entropy) of weight of food product categories, normalized in [0,1]. |
| **representativeness\_norm:** | The ratio between the number of unique customers in the area and the number of residents as measured by the census; values are min-max normalized in [0,1] across all areas |
| **transaction\_days:** | Number of unique dates in which at least one purchase has been made by one of the residents in the area. |
| **num\_transactions:** | Total number of products purchased by Clubcard owners who are resident in the area. |
| **man\_day:** | Cumulative number of man-days of purchase (number of distinct days a customer has purchased something, summed all individual customers) |
| **population:** | Total population of residents in the area according to the 2015 census. |
| **male:** | Total male population in the area. |
| **female:** | Total female population in the area. |
| **age\_0\_17:** | Total number of residents between 0 and 17 years old |
| **age\_18\_64:** | Total number of residents between 18 and 64 years old. |
| **age\_65+:** | Total number of residents aged 65 years or more. |
| **avg\_age:** | Average age of residents according to the 2015 census |
| **area\_sq\_km:** | Surface of the area (km^2) |
| **people\_per\_sq\_km:** | Population density per km^2 |

In this data, where applicable, measures are accompanied with their standard deviation (fields with suffix **\_std**), the 95% confidence interval for the mean (suffix **\_ci95**), and the values of the 2.5th, 25th, 50th, 75th, and 97.5th percentiles (suffix **\_perc{value}**).

There are also validation data, which contains datasets of several health problems. At borough level it provides child obesity data from the English National Health Service (NHS) in 2015 with children in Reception class (aged 4 to 5) and year 6 (aged 10 to 11), adult obesity data from the Active People Survey (APS) in 2012, bariatric hospitalisation from NHS weight loss treatment records in 2015-2016, and at ward level it gives out child obesity in 2015 and diabetes estimate in 2016. There is one more area mapping data of greater London postcodes into larger geographical aggregations. Here are also the explanations of columns in the datasets.

|  |  |
| --- | --- |
| **number\_reception\_measured:** | number of children in reception year measured |
| **number\_y6\_measured:** | number of children in reception year measured |
| **prevalence\_overweight\_reception:** | the prevalence (percentage) of overweight children in reception year |
| **prevalence\_overweight\_y6:** | the prevalence (percentage) of overweight children in year 6 |
| **prevalence\_obese\_reception:** | the prevalence (percentage) of obese children in reception year |
| **prevalence\_obese\_y6:** | the prevalence (percentage) of obese children in year 6 |
| **number\_measured:** | number of people who participated in the survey |
| **prevalence\_healthy\_weight:** | the prevalence (percentage) of healthy-weight people |
| **prevalence\_overweight:** | the prevalence (percentage) of overweight people |
| **prevalence\_obese:** | the prevalence (percentage) of obese people |
| **total\_hospitalizations:** | total number of obesity-related hospitalizations |
| **total\_bariatric:** | total number of hospitalizations for bariatric surgery |
| **prevalence\_hospitalizations:** | prevalence (percentage) of obesity-related hospitalizations |
| **prevalence\_bariatric:** | prevalence (percentage) of bariatric surgery hospitalizations |
| **gp\_patients:** | total number of GP patients |
| **gp\_patients\_diabetes:** | total number of GP patients with a diabetes diagnosis |
| **estimated\_diabetes\_prevalence:** | prevalence (percentage) of diabetes |
| **pcd:** postcode |  |
| **lat:** | latitude |
| **long:** | longitude |
| **oa11:** | output area |
| **lsoa11:** | lower super output area |
| **msoa11:** | medium super output area |
| **osward:** | ward |
| **oslaua:** | borough |

There are further socio-economic data at borough level from London data store. Weekly income by place of residence in London, the education level which used the qualifications of working age population (16-64) in the same year and the sports participation data. Besides, for visualisation the shapefile data of London has also been used.

**3.1.2 Ethical Consideration**

Since the data from Aiello et al. has the licence CC BY 4.0, and all the other datasets are found in the London data store, the data is free to share and adapt. The websites are opened to public, as well as the datasets are approved for works. There will be no risk of using these datasets.

**3.2 Data preparation**

**3.2.1 Data Cleaning**

In this project, many datasets have been used which include shapefiles, XLS files with several work sheet, CSV files. Therefore, cleaning and data validation should be done in the first place for each dataset before processing. Firstly, change all XLS files to CSV files and save the needed work sheet, so that they are easy to read into data frames in python by using ‘Pandas’ package. Next step is read all csv and shape files individually, then drop the ‘NA’ rows in the data and remove any numerical data that contains characters, and rename any columns that may have conflict with python coding such as starting with numbers and spaces in the name. Finally, changing all numbers into type ‘float64’, this will remove the comma in numbers in some XLS files and the percentage sign in some prevalence and rate columns.

**3.2.2 Data selection**

According to the literature review and the research question, all of the nutrients in the food are selected to be the key component. Thus, they can clearly represent the structure of the food and the leading factors of the total energy in people dietary, and total energy has also been chosen since it directly linked with the intake calories. The columns of intake from food are: energy of fat, saturate fat, sugar, protein, carbohydrates, fibre, and total energy. And the last one column from grocery dataset is the diversity (entropy) of energy from nutrients which will help to prove the idea in Milburn’s paper (2004), “nutrition rich and calorie limited”. Then for validation data, number of year-6 child obesity, prevalence of two age groups child and adult health status. The status contains 3 types: health, overweight and obese which are separated by Body Mass Index. And the elements from socio-economic are weekly income in 2016, the qualifications of working age population and the sports participation of London citizens which includes the percentage of once and 3 times exercise per week, and no exercise per week. The times of exercises will be used to model and estimate the energy output for each person. Finally, combine all the data to one dataset, in this case, the original dataset is the Tesco Grocery 1.0 and the rest data combination has been done by using left join with either area id of each borough or the name of each area. Since the diabetes prevalence data is at ward level, the nutrients data will also change to the ward level before the combination. In the further data processing, the obesity and diabetes will be the main dependent variables and the nutrients data will be the energy intake independent variables and the exercises per week will be the energy output independent components.

By finding the relationship between each factor and the dependent variables, it will show the positive or negative influences by these key components. The correlation is the most important and efficient way of represent it. And by checking the correlation coefficient, it is easy to determine the relationship between variables. In this case, Pearson product-moment correlation coefficient is used as the method to calculate the result. The formula of Pearson correlation coefficient is:

= correlation coefficient,

, = values of x, y-variable in a sample,

, = mean of the values of the x, y-variable.

Pearson correlation coefficient demonstrates the correlation between 2 variables, and it has range from -1 to 1. When the coefficient is getting close to 1 then y has strong positive correlation with x, and increasing when x is increasing. Contrarily, when the coefficient gets close to -1, the correlation becomes strong negative and y decrease when x increase. Either positive or negative, when the absolute value of is small, x and y have week correlation, and when , there will be no influence on y when x changes.

Since the correlation is between 2 variables, a 2-Dimension figure is fit for visualisation. Correlation matrix can be the appropriate choice under this section. It can illustrate the calculated correlation coefficient values by plotting according to a colour bar. By observing the matrix, the information of relationship between each factor is convenience to get.

**3.3. Multiple linear regression**

Firstly, the model of obesity based from food nutrients can be created by linear regression method. Since there are several nutrient variables, the multiple linear regression should be used in this case. Before moving to Multiple linear regression model, understanding how does the Simple linear regression works is important.

Linear regression is a linear model which can estimate and approach to model the relationship between a dependent variable and one or more independent variables (Bhadra et al, 2020). If the model is based on only one independent component, it is recognised as simple linear regression. Otherwise, it becomes multiple linear regression if it contains two or more independent variables. In a simple linear regression, independent variable is shown as , and the dependent variable . The regression will generate out a line which is best fitted for all the input data, with the minimum sum of differences between data points and the model line. The function of this line is:

When the variables increase to 2 or more, the model also changes to multi linear regression. In this case, food nutrients and other key components like nutrition diversity and citizens exercises are independent variable , then the function of this model will become:

where in these two functions is the slope of the model (coefficient of each key components) and represents the intersection value of the model and the . There are two important values used to check the quality of regression models. An value and a value, they will be generated by the regression model. Where has the equation:

where,

: the sum of squared total errors (squared residuals)

,

: the sum of squared regression errors

,

value is the coefficient of determinant which is used to illustrate the correlation between the data points and the model line, if is close to 1, the data points are close to the fitting line, it shows the model demonstrate the data efficiently. If is close to 0 then the line may be considered as not fitted, so that model is not significant enough to describe the trend of the data. The -value is the measure of probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct and the lower -value represent a greater statistical significance of the observed difference. Normally, it needs to be small enough () to have significant evidence to reject the null hypothesis test: there is no relationship between the model and data.

Through multiple linear regression, it is convenience to get results by running code in python and improved the integrity compare with the simple linear regression that has been used in Aeillo’s paper (2019). In this way, it does not only analysis the whole dataset which factor has most impact on the health status at the same time, but also generate the model completely and efficiently. The model can also be manipulated easily afterwards.

**3.4 Advanced Regression**

**3.4.1 Stepwise Regression**

Since data analysis is based on statistics, there will be further process on the data which can improve the result of regression. The first value that can be considered is the -value. When selecting key components with smaller -value, it will abandon the independent variables that has less significantly influences on the dependent variables. Stepwise regression is the exact method working on filtering -value from the dataset with changeable thresholds. This method examines the statistical significance step by step. The first step is

**3.4.2 (VIF)**

**3.5 Model and classification**

**4. Result:**

**5. Discussion:**

**6. Conclusion:**

**7. Reference:**

Bhadra, A., Mukherjee, A., & Sarkar, K. (2020). Impact of population density on Covid-19 infected and mortality rate in India. *Modelling earth systems and environment*, 1–7. Advance online publication. https://doi.org/10.1007/s40808-020-00984-7