Project Report

# GitHub URL

(https://github.com/CharlieJOD/UCDPA\_JohnODwyer)

# Abstract

This project combined three datasets, cleaned, investigated and analysed the data looking at interactions between diabetes and other risk factors such as age, BMI, general health results. The findings are explained using visualisations. Pandas, Numpy, Matplotlib and Seaborn were the packages used for data analysis.

# Introduction

# The following statistics are taken from the Centers for Disease Control and Prevention in the US.

* **Diabetes**
* **Total:** 37.3 million people have diabetes (11.3% of the US population)
* **Diagnosed:** 28.7 million people, including 28.5 million adults
* **Undiagnosed:** 8.5 million people (23.0% of adults are undiagnosed)
* **Pre-Diabetes**
* **Total:** 96 million people aged 18 years or older have prediabetes (38.0% of the adult US population)
* **65 years or older:** 26.4 million people aged 65 years or older (48.8%) have prediabetes. (ref: 1)

Analysing data on diabetes and it’s risk factors has the potential to improve people’s lives and help focus efforts to prevent and control diabetes, not only across the US but worldwide.

# Dataset

*“The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. Each year, the survey collects responses from over 400,000 Americans on health-related risk behaviors, chronic health conditions, and the use of preventative services. It has been conducted every year since 1984. For this project, a csv of the dataset available on Kaggle for the year 2015 was used. This original dataset contains responses from 441,455 individuals and has 330 features. These features are either questions directly asked of participants, or calculated variables based on individual participant responses.*

*This dataset contains 3 files:*

*diabetes \_ 012 \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_012 has 3 classes. 0 is for no diabetes or only during pregnancy, 1 is for prediabetes, and 2 is for diabetes. There is class imbalance in this dataset. This dataset has 21 feature variables*

*diabetes \_ binary \_ 5050split \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 70,692 survey responses to the CDC's BRFSS2015. It has an equal 50-50 split of respondents with no diabetes and with either prediabetes or diabetes. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is balanced.*

*diabetes \_ binary \_ health \_ indicators \_ BRFSS2015.csv is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. The target variable Diabetes\_binary has 2 classes. 0 is for no diabetes, and 1 is for prediabetes or diabetes. This dataset has 21 feature variables and is not balanced.”* (ref: 2)

# Implementation Process

Data Importing

I imported the relevant packages such as Numpy, Pandas, Matplotlib and Seaborn.

I downloaded three similar databases to my hard drive and then imported them as CSV files. I chose to include three datasets to get a larger volume of data to work with.

Data Exploration

I explored the data, looking at the shape for each of the three datasets and the columns. Upon seeing the columns I realized that for df1 the first column had a different name so I renamed it before then concatenating the three datasets to create df, the name of the dataset for this project. While the columns had different names the data in the columns was the same.

Data Cleaning

I then looked for duplicates, checked the shape was correct after merging and looked at some basic information on the dataset.

I was not happy with some of the names of the dataset so I renamed several.

I further examined the data finishing with df.corr() to allow a consideration of where there may be areas to consider more closely.

I then looked at BMI closely to get a feel for investigating the data and used several lines of code to provide a detailed understanding of this column.

I wrote a function that would allow me to categorise BMI into the relevant ranges (ref: 3) as shown here:

def categorize\_bmi(value):

if value < 18.5:

return 'Underweight'

elif value>=18.5 and value<=24.9:

return 'Healthy Weight'

elif value>=25.0 and value<=29.9:

return 'Overweight'

elif value>=30.0 and value<=39.9:

return 'Obese'

elif value>=40.0:

return 'Severely Obese'

I then analysed Age, manipulating the data to produce visualisations to provide insights.

Looking at these age distributions, it seems that the Diabetic population has a large concentration in the older age ranges. This would agree with the general prevalence of diabetes amongst the older members of a population.

I then analysed interactions with age and bmi and diabetes.

I then provided visualisations for High Cholesterol, Smoking, BMI and High Blood Pressure and how they interacted with the ‘Has Diabetes’ column.

In this analysis I used ratio for fig size as produces a clearer graph and saves time sizing headings. I also used sharey=True to compare on the same y axis

I completed some sorting and indexing exercises.

I used the pandas .groupby() function to split the data into groups based on certain variables. Pandas objects can be split on any of their axes. I did this to form groups based on more than one category.

Data Visualization

I used several types of visualisations in Seaborn, such as countplots and barplots which were very suitable for examining this dataset. Further on I used pointplots and finally I used some lineplots to examine the health variables and their interactions with diabetes.

For visualisations I showed several countplots side-by-side comparing Diabetic or Not Diabetic populations on variables such as High Cholesterol, Smoking, BMI and High BP. However, due to the large number of the population without diabetes I also showed countplots for just the population with diabetes.

Finally, I showed an example of subsetting the dataframe at the end and explored some tentative and simple data modelling procedures while not analyses.

Data Modelling

For making predictions on a categorical dependent variable we should use classification and in this instance it would be a multinomial classification

I have included some very simple code for machine learning at the end of the project. I have attempted to analyse BMI and General Health as predictors of reducing or increasing health.

# Results and Insights

# Insights

1: The first insight are some general comments on the dataset. This is a self-reported questionnaire and, as such, people inputs are subjective. Another thing to note is that there is a large portion of the population that do not have diabetes as can be seen from the graph produced below.

Chart, waterfall chart

Description automatically generated

2: Age Categories and BMI:

The below graph shows a very clear interaction, which is true for both males and females in this dataset, the population in middle-age have the highest incidences of having a higher BMI score. The female BMI score is higher in general which is expected. A healthy weight for BMI is below 25 and this graph shows that the vast majority of the population are overweight. Over 30 moves into the obese category but thankfully the population averages below that.

Chart, line chart

Description automatically generated

I find the following graph to show a very interesting story. It shows the interaction of BMI over age and also how the three categories of Has Diabetes, Has Pre-Diabetes and Does not have Diabetes plot. There is a nearly identical graph for the scores for BMI in the Has Diabetes or Has Pre-Diabetes categories. Both categories are also markedly above the BMI scores for the population without diabetes.

Chart, line chart

Description automatically generated

3: Interaction between Health scores and Diabetes:

The following graphs all show that a poorer score on self-reported health questions show a correlation with diabetes. Therefore, it can be assumed that that those people with diabetes record themselves as having poorer health.

This is most clearly seen in the graph below showing the Interaction of General Health Score and Diabetes. As the Gen Score on the x axis increases people are reporting themselves as less healthy, strongly correlating with the incidence of diabetes on the y axis.

Chart, line chart

Description automatically generated

Below, the interaction with physical health can also be seen although the pattern is not as defined.

Chart, line chart

Description automatically generated

Again, with the mental health scores plotted below there is a correlation although it would appear to be slightly weaker.

Chart

Description automatically generated

4: Age distribution and Diabetes

As can be seen below there is a marked and consistent increase in the instances of diabetes through every age category. This, again, fits with the general worldwide pattern of diabetes which is more prevalent as people age.

Chart, bar chart

Description automatically generated

Looking at these two age distributions below, it also seems that the Diabetic population has a large concentration in the older age ranges. This would agree with the general prevalence of diabetes amongst the older members of a population.

Chart, histogram

Description automatically generated

5: High Blood Pressure and Diabetes Interaction

Chart

Description automatically generated

This image shows the interaction of people with high blood pressure and diabetes. It shows that the people in this dataset who have diabetes have almost two and a half times the incidences of having high blood pressure.

*“High blood pressure is twice as likely to strike a person with diabetes than a person without diabetes. Left untreated, high blood pressure can lead to*[*heart disease*](http://www.hopkinsmedicine.org/healthlibrary/conditions/adult/diabetes/diabetes_and_heart_disease_85,p00341/)*and*[*stroke*](http://www.hopkinsmedicine.org/healthlibrary/conditions/adult/physical_medicine_and_rehabilitation/stroke_85,p01184/)*. In fact, a person with diabetes and high blood pressure is four times as likely to develop heart disease than someone who does not have either of the conditions. About two-thirds of adults with diabetes have blood pressure greater than 130/80 mm Hg or use prescription medications for hypertension.”*(Ref: 4)

This is taken from the John Hopkins website. Having this information will allow policy makers to engage funding research and practice to tackle the health issues that this population will have and to create heart healthy environments for this population, as well as tackling their current medical issues.

# References

Websites

1: <https://www.cdc.gov/diabetes/data/statistics-report/index.html>

2: <https://www.kaggle.com/code/analystarif/dadiabetes-data-analysis-project/input>

3: <https://www.cdc.gov/obesity/basics/adult-defining.html>

4: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure#:~:text=In%20fact%2C%20a%20person%20with,use%20prescription%20medications%20for%20hypertension>.

Other websites used for resources included:

[www.datacamp.com](http://www.datacamp.com)

[www.gitbub.com](http://www.gitbub.com)

[www.statology.org](http://www.statology.org)

[www.geeksforgeeks.org](http://www.geeksforgeeks.org)

Dataset references

This is the original source and code book for the data.

<https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf>

This is the original source and code book for the data.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Values |
|  |  |  |
| Diabetes\_012 | Diabetes Status |  0 - no diabetes   1 - prediabetes   2 - diabetes |
| HighBP | High Blood Pressure |  0 - no   1 - yes |
| HighChol | High Cholesterol |  0 - no   1 - yes |
| CholCheck | Cholesterol check in last 5 years |  0 - no   1 - yes |
| BMI | Body Mass Index | (numerical variables) |
| Smoker | smoked >=100 cigarettes(5 packs) in entire life |  0 - no   1 - yes |
| Stroke | Ever had a stroke? |  0 - no   1 - yes |
| HeartDiseaseorAttack | Ever had coronary heart disease (CHD) or myocardial infarction (MI)? |  0 - no   1 - yes |
| PhysActivity | Physical activity in past 30 days, not including job |  0 - no   1 - yes |
| Fruits | Consume fruit 1 or more times per day |  0 - no   1 - yes |
| Veggies | Consume vegetables 1 or more times per day |  0 - no   1 - yes |
| HvyAlcoholConsump | Heavy drinkers(adult men or women having more than 14 or 7 drinks/week respectively) |  0 - no   1 - yes |
| AnyHealthCare | Have any form of healthcare coverage |  0 - no   1 - yes |
| NoDocbcCost | Unable to see doctor due to cost any time in past 12 months |  0 - no   1 - yes |
| GenHlth | General Health rating |  1 - excellent   2 - very good   3 - good   4 - fair   5 - poor |
| MentHlth | No. of days of poor mental health in past 30 days | (number between 0 and 30) |
| PhysHlth | No. of days of poor physical health in past 30 days | (number between 0 and 30) |
| DiffWalk | Serious difficulty walking or climbing stairs? |  0 - no   1 - yes |
| Sex | Male or female? |  0 - female   1 - male |
| Age | 13-level age category |  1 - 18-24   2 - 25-29   3 - 30-34...   12 - 75-79   13 - >=80 |
| Education | Level of Education |  1 - Kindergarten/None   2 - Elementary (Grades 1-8)   3 - Some High School (Grades 9-11)   4 - High School Graduate(Grade 12 or GED)   5 - Some College (1-3 yrs) or Technical School   6 - College (1-3 yrs) Graduate |
| Income | Income($) scale |  1 - <10,000   5 - <35000   8 - >=75000 |