Extraction, Transformation, and Load Technical Report

Dr Lamba’s Minute Clinic

ETL for Data on Diabetes and Cardiovascular Conditions

Contents

[**1.0 Introduction** 3](#_Toc19209232)

[1.1 Summary 3](#_Toc19209233)

[1.2 Scope 3](#_Toc19209234)

[1.3 Technologies and Resource Contributions 3](#_Toc19209235)

[1.4 Definitions, Acronyms and Abbreviations 4](#_Toc19209236)

[**2.0 ETL Details** 4](#_Toc19209237)

[2.1 Data Import/Extract Sources and Method 4](#_Toc19209238)

[2.2 Data Acquisition 5](#_Toc19209239)

[2.3 Data Transform 5](#_Toc19209240)

[2.4 Data Integrity 6](#_Toc19209241)

[2.5 Data Refresh Frequency 7](#_Toc19209242)

[2.6 Data Security 7](#_Toc19209243)

[2.7 Data Loading and Availability 7](#_Toc19209244)

[**3.0 Data Quality** 7](#_Toc19209245)

# **1.0 Introduction**

## Summary

Dr. Lamba’s Minute Clinic is looking to provide additional information to diabetic patients that come in for any screening and monitoring that relate to their diabetic conditions. Dr. Lamba’s clinic wants to find out if there is any correlation between body vitals such as Body Mass Indices (BMI), Age and medical conditions like diabetes and heart attack. They would also like to investigate if there exists any interdependency between diabetes and heart attack.

**Hypothesis:**

1. There is a correlation between body vitals such as age, BMI and diabetes
2. There is a correlation between body vitals such as age, BMI and heart attack
3. There is correlation between patients with diabetes and propensity for heart attack

**Null:**

1. There is no correlation between body vitals such as age, BMI and diabetes
2. There is no correlation between body vitals such as age, BMI and heart attack
3. There is no correlation between patients with diabetes and propensity for heart attack

## Scope

Our primary source of data was CSV files from Kaggle.com. All of the data collected did not need to be HIPPA compliant. Our goal was not to find the correlation, but simply to clean and transform the relevant data. Additional analysis will be conducted by the team at Dr. Lamba’s Minute Clinic. We ensured that our data sources had common fields that could be used to join data and provide opportunity for comparison.

Because this data is static, and we are not collecting information that is updated regularly using data scraping techniques or API connections, this data should be considered a one-time snap shot. Additional factors that could be included but were not are things like vegan, vegetarian, or red meat diet, patient medications, etc.

## Technologies and Resource Contributions

**Team:**

1. Dinesh Lamba
2. Vin Dixit
3. Wesley Stone
4. Kendall Gouldthorpe

**Technologies Used:**

1. Python
2. NumPy
3. Pandas
4. PostgreSQL
5. Excel
6. Seaborn
7. Matplotlibs
8. Quickdatabasediagrams.com

## Definitions, Acronyms and Abbreviations

Abbreviations/Acronyms found in the data:

* Ap\_hi – Systolic Blood Pressure
* Ap\_lo – Diastolic Blood Pressure
* Gluc – Glucose
* Alco – Alcohol Use
* Smoke – Tobacco Use
* Active – Level of Exercise
* BMI – Body Mass Index
* ETL – Extract, Transform, Load

# **2.0 ETL Details**

## 2.1 Data Import/Extract Sources and Method

**Data Set #1:**

Cardiovascular Disease Dataset

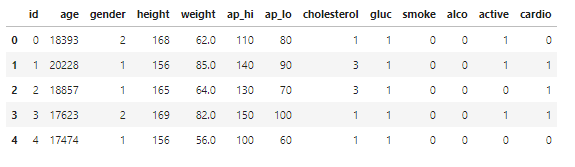
<https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>

File Type: .csv

Rows: 70,000

Data includes:

1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm)
3. Weight | Objective Feature | weight | float (kg)
4. Gender | Objective Feature | gender | categorical code
5. Systolic blood pressure | Examination Feature | ap\_hi | int
6. Diastolic blood pressure | Examination Feature | ap\_lo | int
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal
9. Smoking | Subjective Feature | smoke | binary
10. Alcohol intake | Subjective Feature | alco | binary
11. Physical activity | Subjective Feature | active | binary
12. Presence or absence of cardiovascular disease | binary



**Data Set #2:**

Diabetes

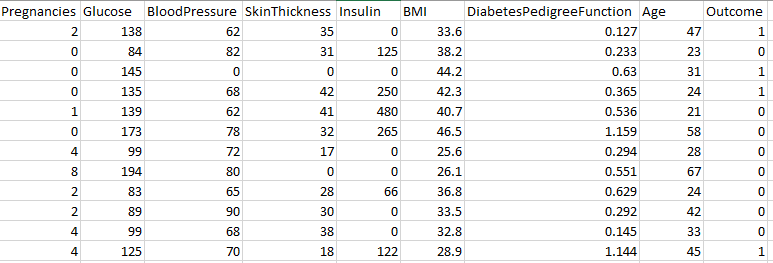
<https://www.kaggle.com/johndasilva/diabetes>

File Type: .csv

Rows: 2,000

Data includes:

1. Pregnancies | Pregnancies | int
2. Glucose | Glucose | int
3. Blood Pressure | BloodPressure | int
4. Skin Thickness | SkinThickness | int
5. Insulin | Insulin | int
6. BMI | BMI | float
7. Diabetes Pedigree Function | DiabetesPedigreeFunction | float
8. Age | Age | int
9. Outcome | Outcome | binary



## 2.2 Data Acquisition

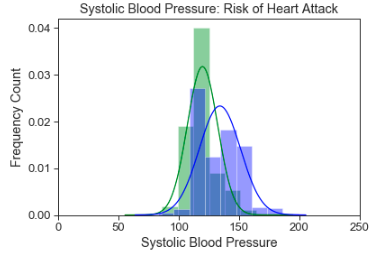
Our datasets are static; this information all came from one-time studies. All information can be found at Kaggle.com with a free login sign up.

For future updates, additional research into data sources will need to be conducted. Potential future data could come from sources like data.gov API or other medical institutions that will require permissions and subscriptions. Further research can be negotiated for data scraping and API use depending on the client’s needs.

## 2.3 Data Transform

We inspected the data using data visualization techniques to identify medically improbable outliers. Once they were identified, we were able to remove them and create cleaned versions that could be passed into PostgreSQL for further transformation.

Below is a sample of the python code used to visualize and remove outliers:



**Bell Curve was showing as skewed while Systolic Blood Pressure included numbers over 200. (Medically Improbable)**

**Here we are removing the outliers to ensure that the included data is accurate.**

Once the data was deemed cleaned, we were able to use PostgresSQL to create a database and join the datasets to start the comparison process for the client.

Below are suggested views:

**Age View – Joining tables on age to see if diabetes or cardiovascular condition is present**

**Age-BMI Combination View – Joining tables on age while including BMI to see if diabetes or cardiovascular condition is resent**

## 2.4 Data Integrity

The datasets we collected were relatively clean. Using data visualization techniques, we were able to spot the medical improbabilities and remove them. In addition, for a more robust comparison, we used the given information of weight and height to calculate BMI.

Our datasets were from one-time studies, and therefore will not ever be updated.

## 2.5 Data Refresh Frequency

There will be no refreshing of these data sources. The datasets are static.

## 2.6 Data Security

The datasets include personal medical information, but all personal information has been previously scrubbed. There is no concern for HIPPA violations.

Our recommendation is to always back up any data or project work. We recommend sites like GitHub to store projects and data for retrieval as well as collaboration.

## 2.7 Data Loading and Availability

In order to further manipulate the datasets, we recommend using PostgresSQL. This is a free database software that will allow the client to further manipulate and join the datasets to discover correlations between age, BMI, diabetes and cardiovascular conditions. We are also including the cleaned CSV files which can be imported into any program the client chooses to use to manipulate that file type. [IE – Jupyter Lab, local terminal, Python, MongoDB, etc.]

# **3.0 Data Quality**

Once combined, the dataset is large enough (row count 1,183,216) for the client to preform predictive analysis using right train/test data split (80/20) without losing statistical significance of the data volume.

Based on our original proposal to Dr. Lamba’s Minute Clinic, we promised to provide the following deliverables:

1. Report describing methodology of data collection and data cleaning
2. Two cleaned datasets
3. Various views joining the datasets that would help client to analyze for given hypothesis

The views were created to assist the client to make statistical predictions without further cleaning or transformations.