Milestone Report

Keith Engwall

4/8/2018

Project Introduction

This project will attempt to create a predictive model for identifying patients with diabetes (or at least at risk for diabetes) using patient characteristics such as age, gender, allergies, comorbidities, prescriptions, and a variety of measures such as BMI, blood pressure, hemoglobin levels, etc.

Data

The project makes use of an EMR data set from 2012 of approx. 10,000 patients. The data set includes tables for patients, allergies, diagnoses, and prescriptions, as well as tables for transcripts of visit data and labs. The key tables and fields are described below.

Patient Table

The patient table contains an indicator field to identify the diabetic population, as well as the patient's gender and year of birth

Allergy Table

The allergy table contains a field for allergy type to identify the category of allergy, as well as for reaction name for the type of reaction and severity name for the severity of the allergic reaction. The allergy table also contains a field for medication ndc codes to map medication allergies to a specific medication.

Diagnosis Table

The diagnosis table contains the ICD9 Code to specifically identify the diagnosis, as well as an Acute indicator to flag acute instances of a diagnosis. Although there is a diagnosis description field, its contents are not standardized and thus not suitable for analysis. Instead, the ICD9 codes can be mapped to names from a table available online: List of ICD-9 Codes.

Medication Table

The medication table contains the NdcCode field, which specifically identifies medications, as well as the Medication Name.

Transcript Table

The transcript table contains fields for height, weight, BMI, SystolicBP, DiastolicBP, respiratory rate, heart rate, and temperature.

Lab Observation Table

The lab observation table contains values from a variety of labs identified by an HL7 identifier. The text for these is more standardized and can be used to identify the type of lab the observations are for. The observation value and units fields provide the actual measurement values. There is an identifier for abnormal values as well as an abnormal flags field which indicates whether the abnormal value is high or low and to what degree.

Limitations

Although there is also an immunization table and a smoking status table, these do not provide sufficient data to analyze. Although the data structure allows for labs to link to visit transcripts, none do.

Data Cleaning

The tables contain several fields that are not relevant to analysis, such as the practice identifier, physician specialty, etc. The tables must be joined to link data to specific patients, as well as to each other (allergy & medication, prescription & medication, lab result & lab panel & lab observations, etc.).

Age was derived by subtracting the year of birth from 2010 (the median year for the data contained in the data set). The allergy field for the medication ndc code had to be renamed to disambiguate it from the prescription field of the same name. Numerous fields needed to be converted into numeric or integer types. In order to analyze systolic and diastolic blood pressure as a pair, a field for pulse pressure (systolic bp - diastolic bp) was added.

Some of the data had too much differentiation and needed to be chunked in order to be analyzed. For instance, the diagnoses were chunked into categories based on ICD-9 Code ranges. Some of the data needed to be filtered to remove insignificant data. For example, there were few if any records in the allergy table for the various types of medication allergies among the diabetic population. And for most of the remaining data, the ratio of diabetic patients with a particular medication allergy to non-diabetic patients was not of note. Therefore, the medication allergy data was filtered down to those for which there was at least one diabetic patient with an allergy and at least 20 patients overall with an allergy.

In order to identify the medication allergies, a map needed to be created between the ndc code and the medication name from the data in the medication table.

For medication usage, the vast amount of data needed to be filtered still more. Only data where diabetic patients accounted for greater than 60% were included. Also, only data where there were greater than 300 records were included.

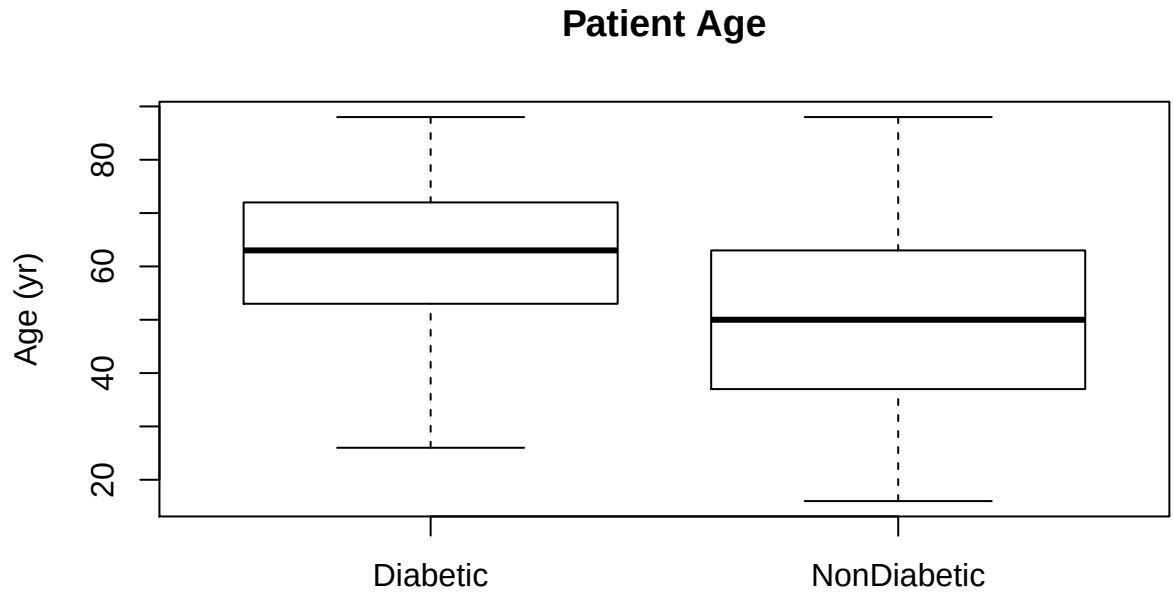
The abnormal flags needed to be reordered in order to display in a logical order (high to low) rather than in alphabetical order. Likewise, a field was added to provide a text equivalent for the diabetic patient indicator in order to be interpreted properly in the graphs.

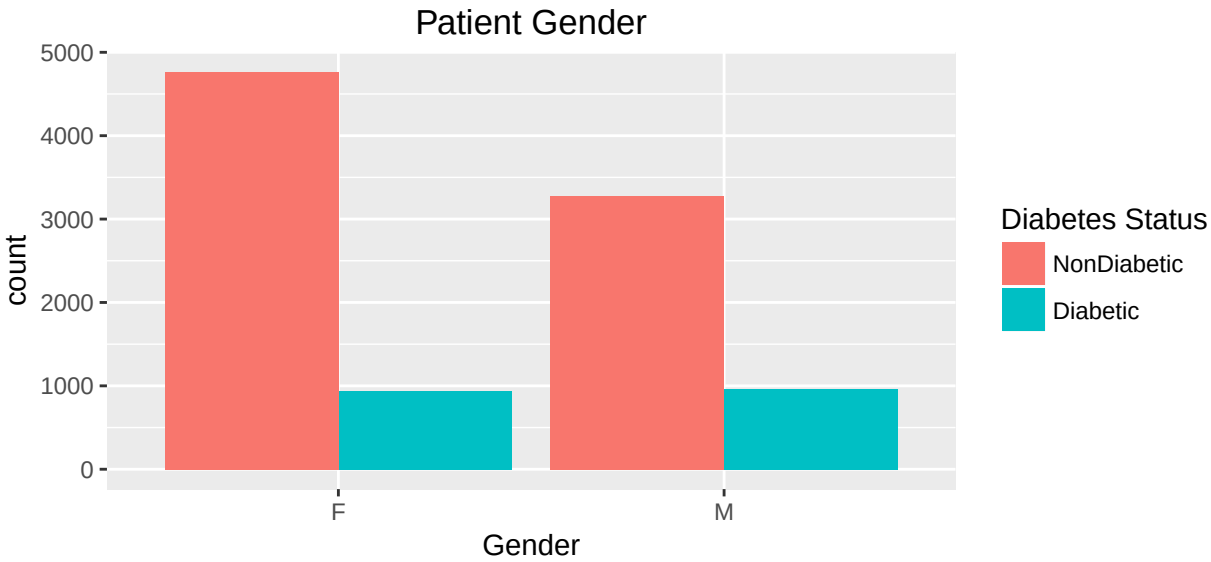
Initial Findings

Based on initial findings, there are some patient characteristics that may contribute to a predictive model.

Age & Gender

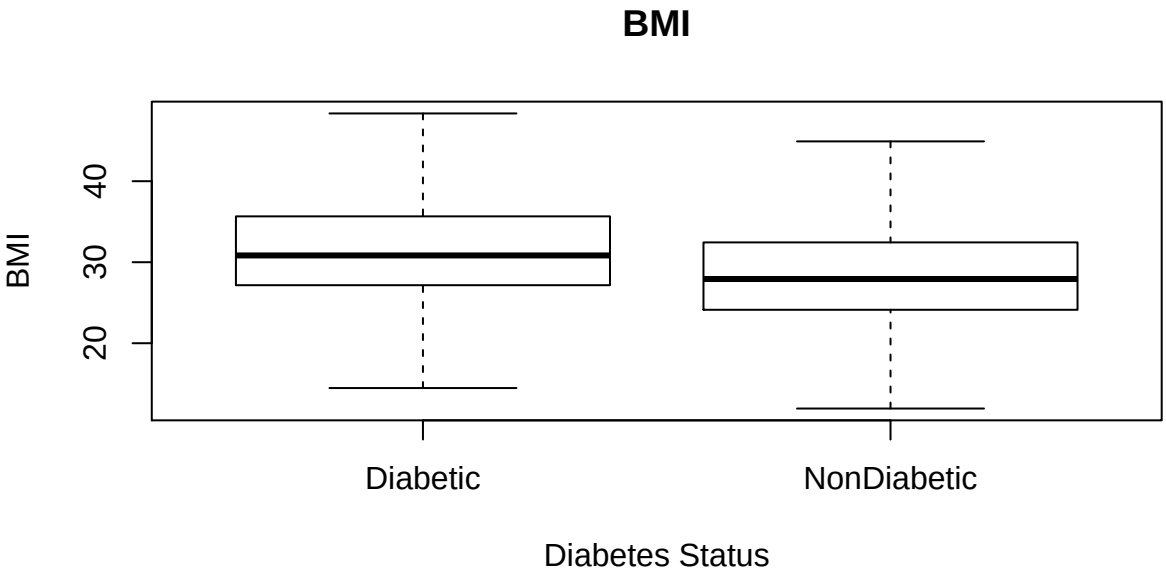
The diabetic population trends male and older than the non-diabetic population.





BMI

Diabetic patients have a slightly higher weight and shorter height than non-diabetic patients. Thus, their BMI trends higher.

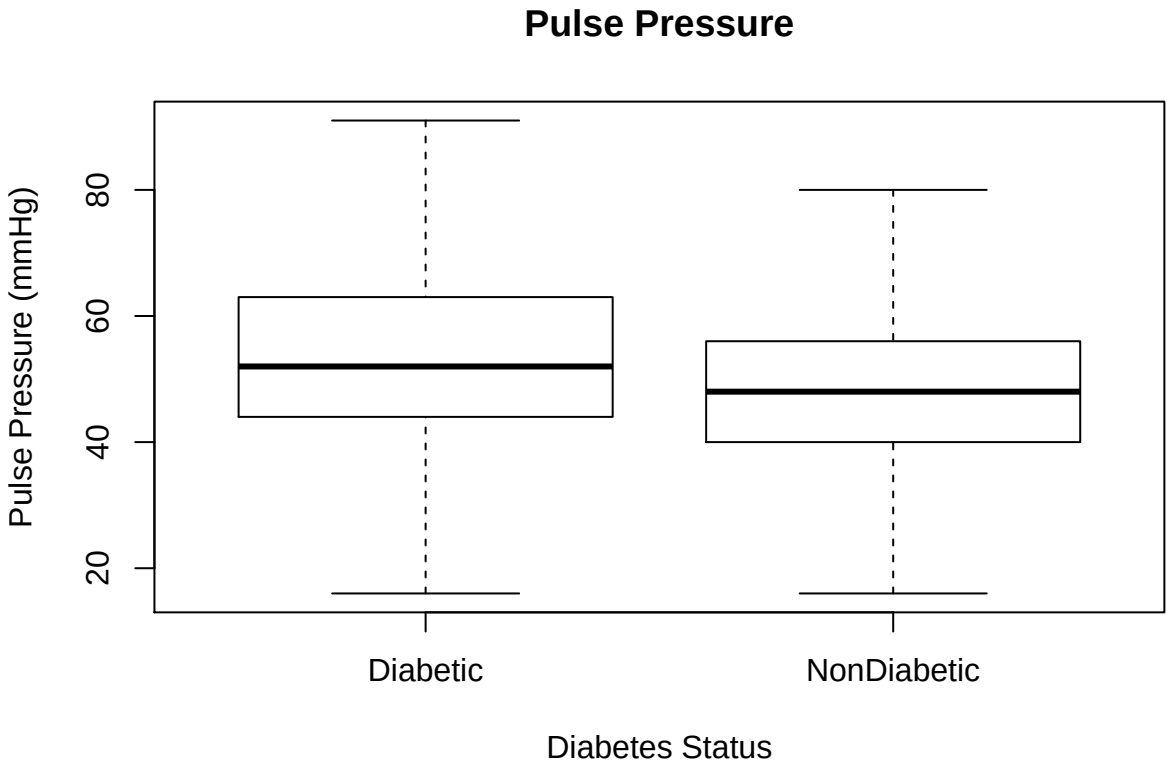


Allergies

Among non-medical allergy types, there appears to be a large proportion of diabetic patients with egg and peanut allergies in comparison to the general population. The number of diabetic patients with egg allergies actually outnumbers non-diabetic patients.

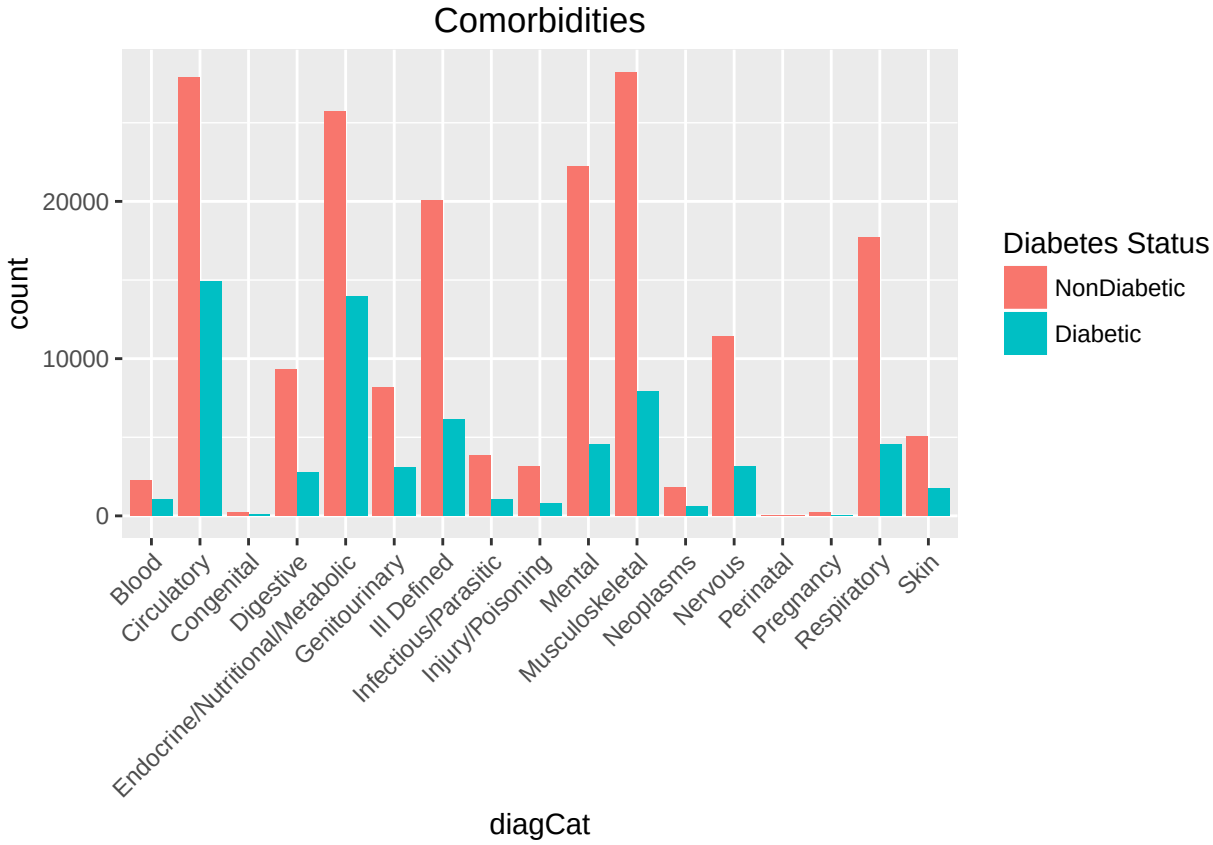
Blood Pressure

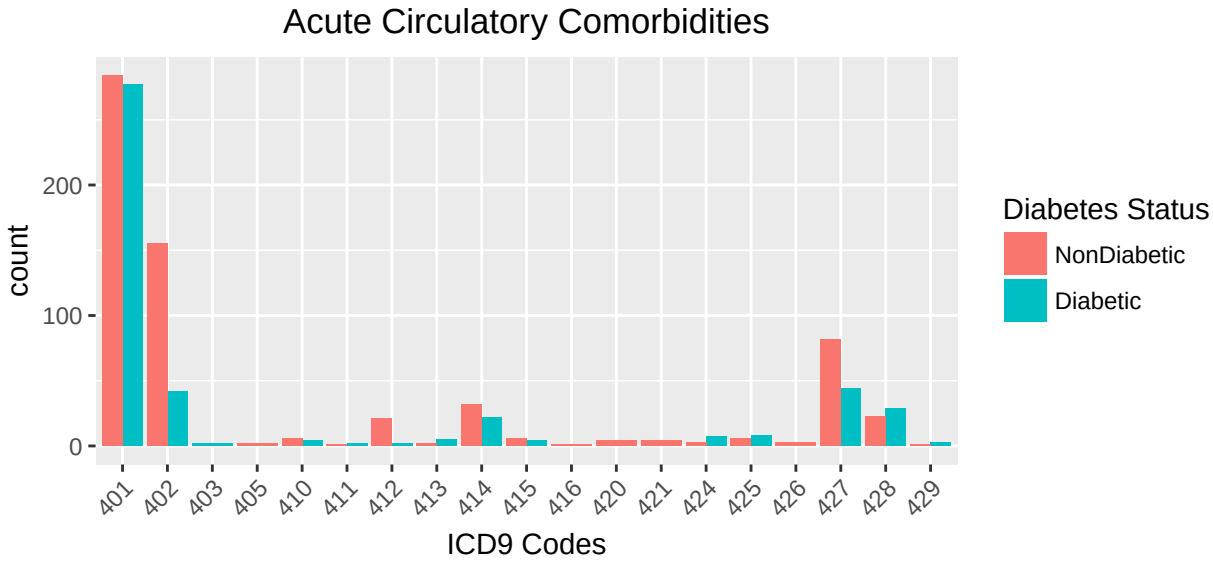
The pulse pressure (Systolic BP - Diastolic BP) of diabetic patients is slightly higher than non-diabetic patients.



Diagnosis Categories

To simplify analysis of diagnoses, the various diagnoses were divided into categories based on the ICD9 Codes. The diagnosis categories with the highest ratio between diabetic and non-diabetic patients is Circulatory. Within the Circulatory, there are almost as many diabetic patients with Essential Hypertension (ICD9 Code #401) as non-diabetic patients.



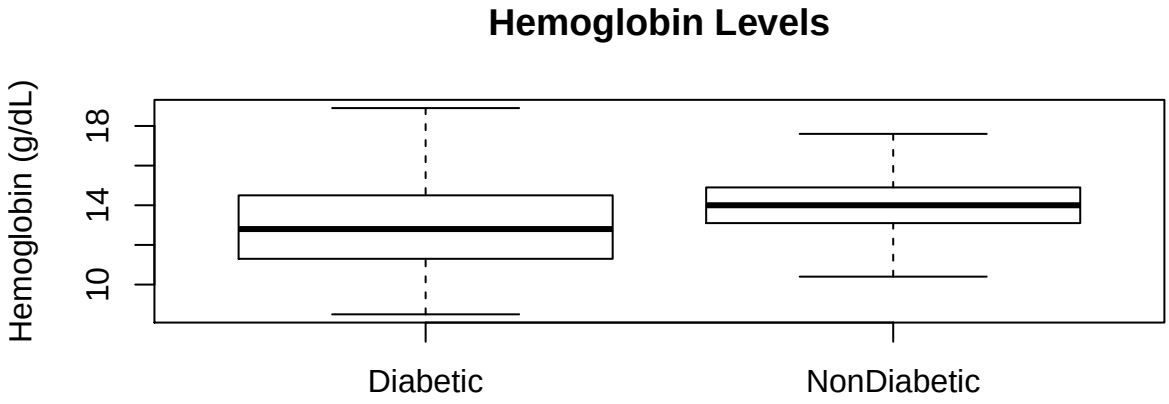


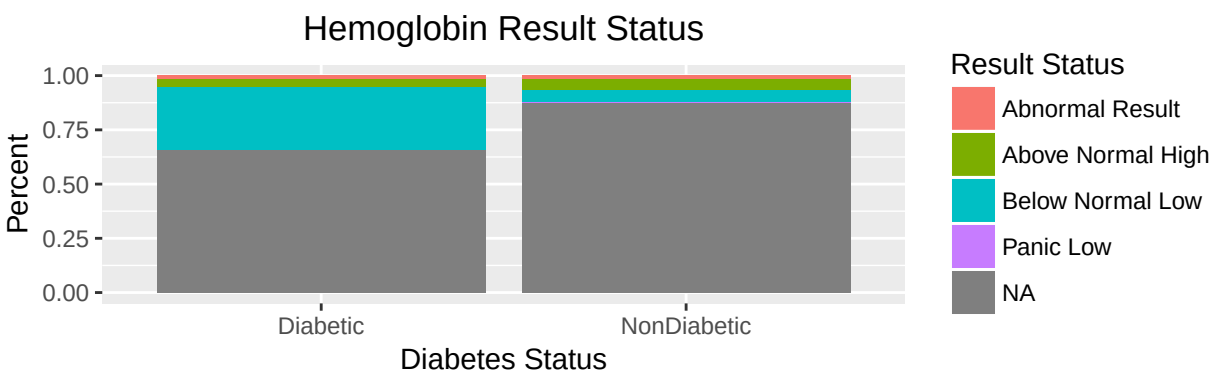
Lab Result Analysis

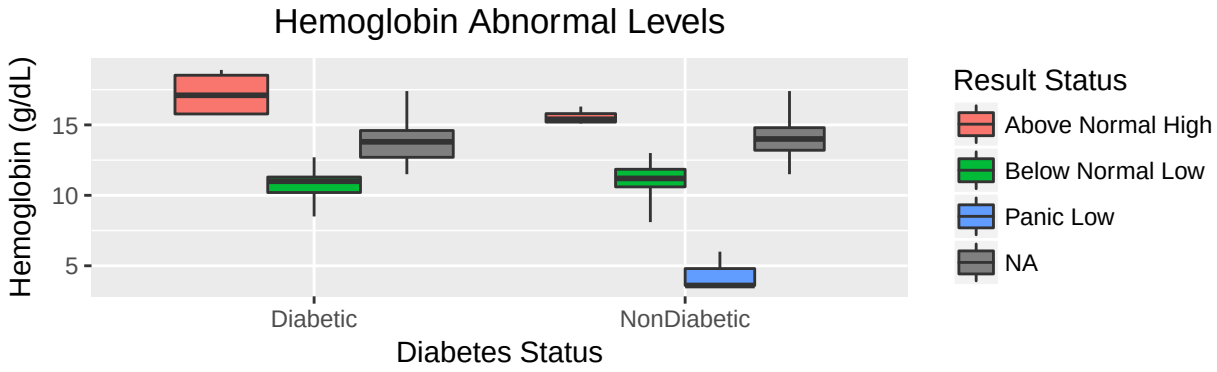
The lab results were limited to those for which there were sufficient abnormal readings data for the diabetic population. For each lab result, an overall measure of central tendency was analyzed, as well as an analysis of abnormal lab result status.

Hemoglobin Lab Results

Hemoglobin levels have a lower central tendency in diabetic patients than in non-diabetic patients. When results recorded as abnormal are separated out, the diabetic population has a larger percentage of below normal readings. The central tendency of above normal readings were higher and of below normal readings were lower among the diabetic population. The central tendency of normal readings was slightly lower in the diabetic population.

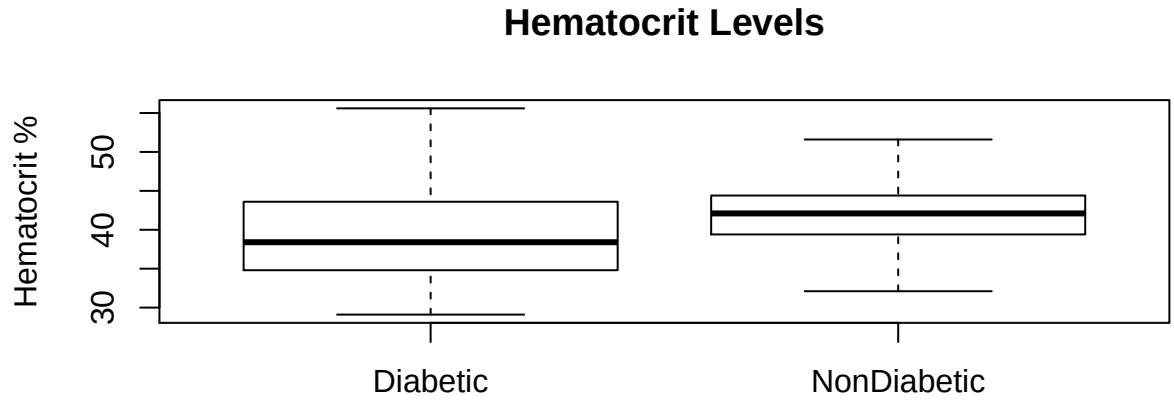


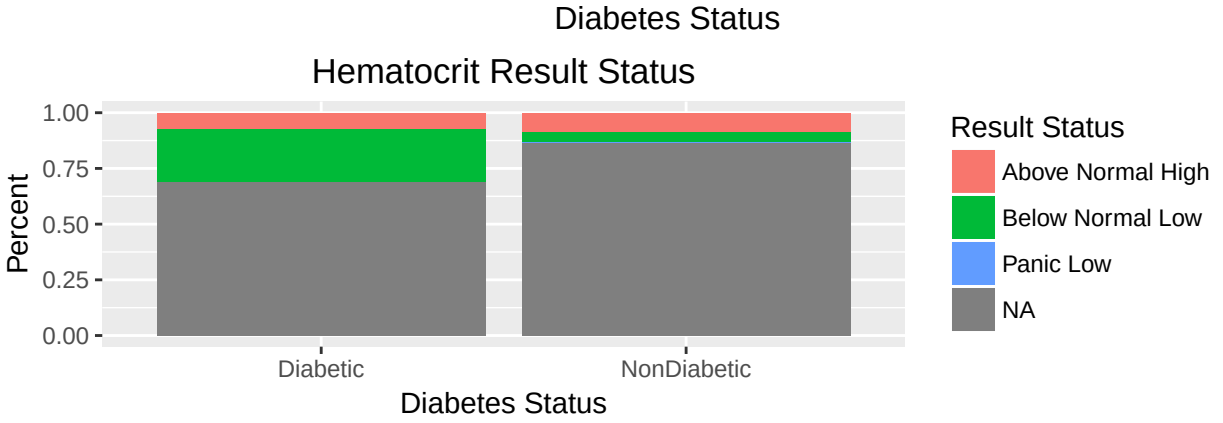


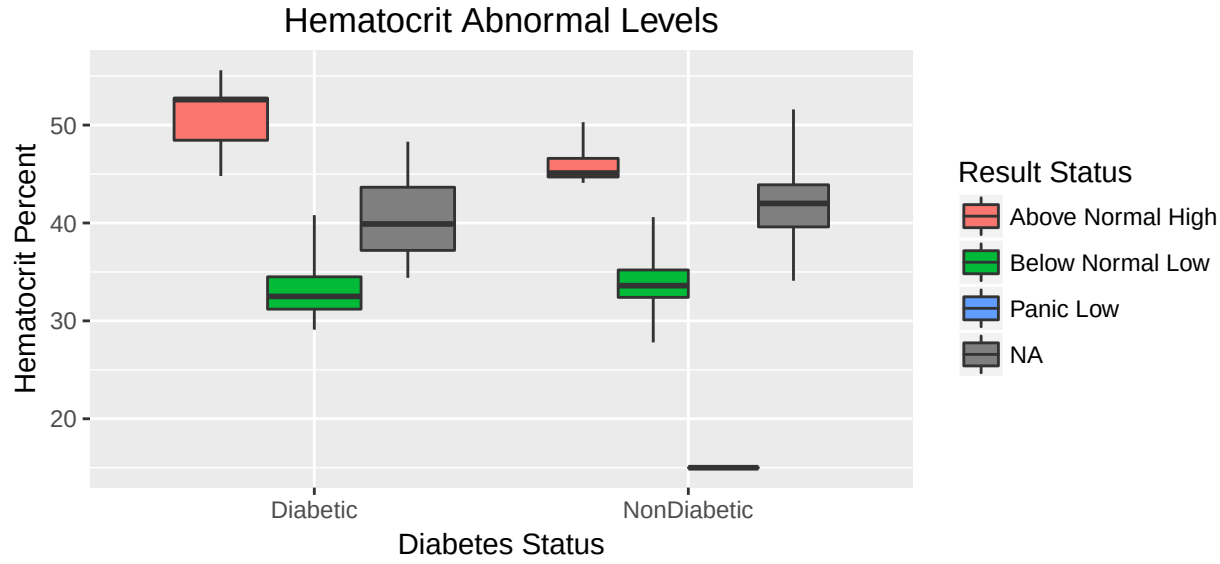


Hematocrit Lab Results

The central tendency of Hematocrit percentage was lower in the diabetic population than in the non-diabetic population. The ratio of above normal readings was lower and of below normal readings was higher among the diabetic population. The central tendency of above normal readings was higher and of below normal readings was lower in the diabetic population. The central tendency of normal readings was lower in the diabetic population.

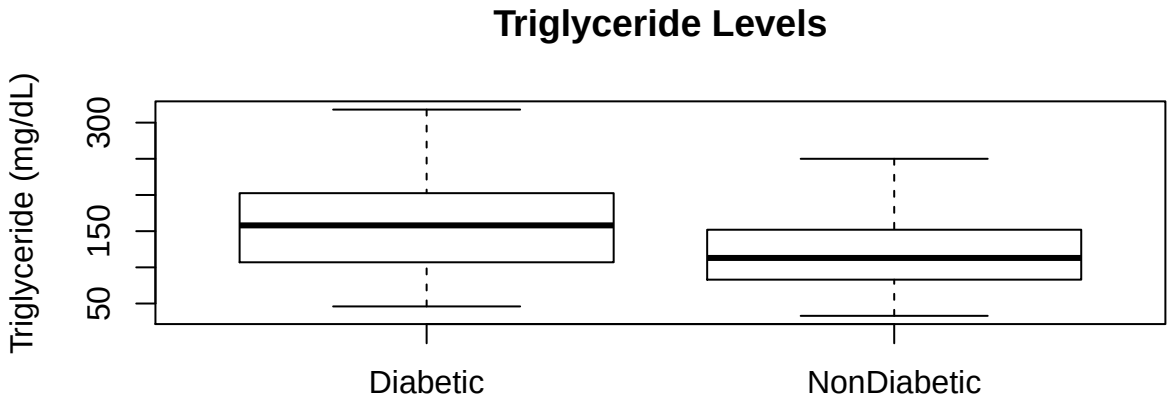


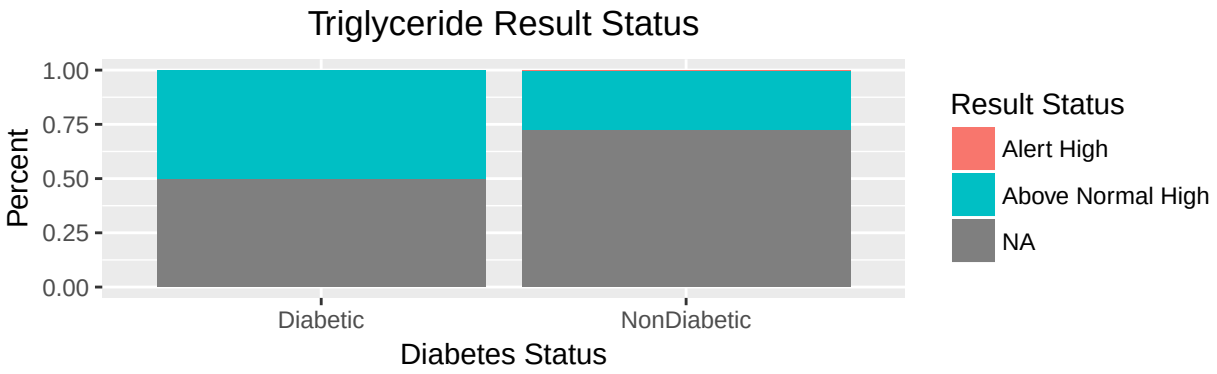


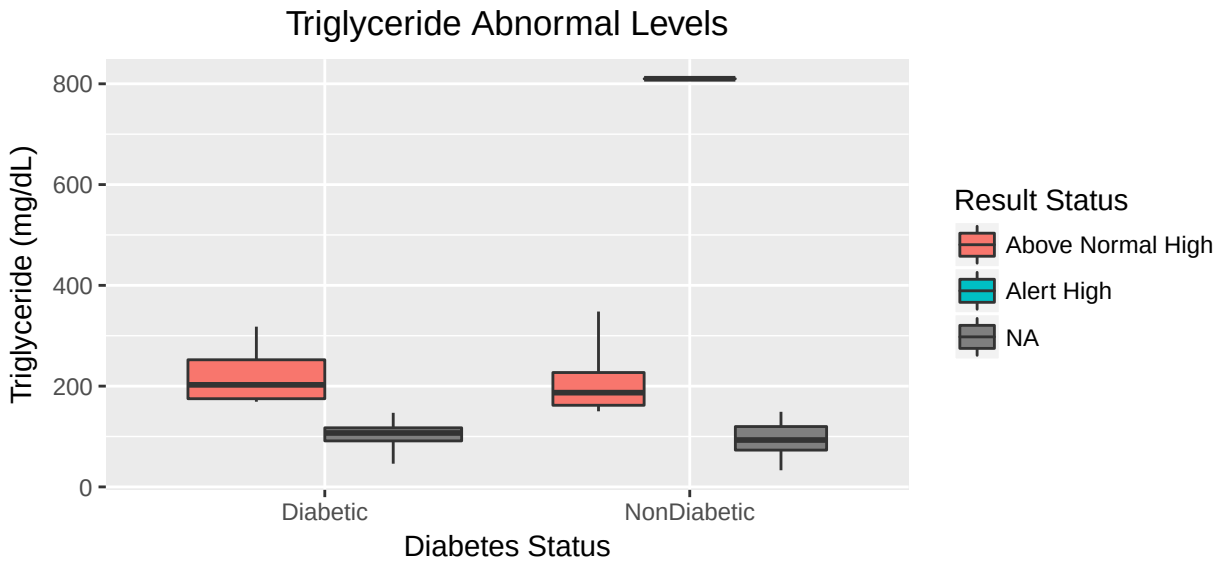


Triglyceride Lab Results

The central tendency of triglyceride levels is higher in the diabetic population than in the non-diabetic population. There is a higher ratio of above normal triglyceride readings in the diabetic population, and both the above normal and normal triglyceride levels are higher in the diabetic population.

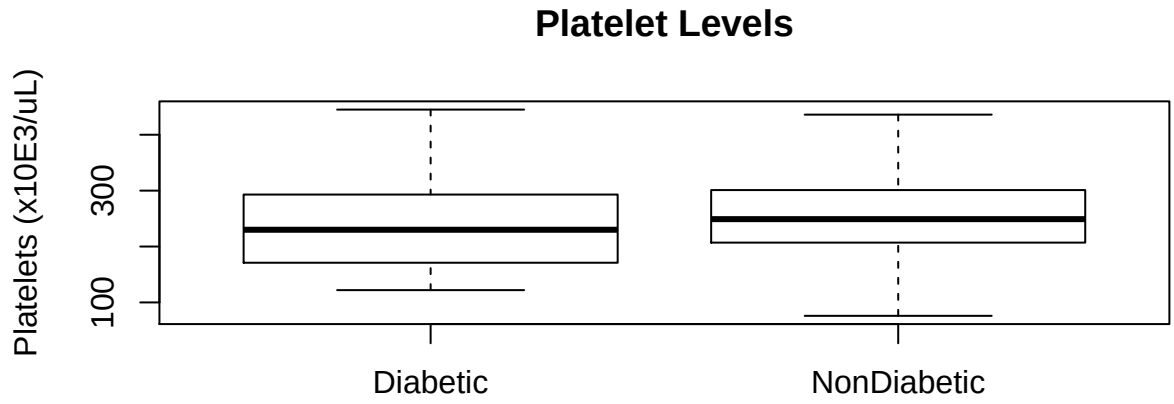


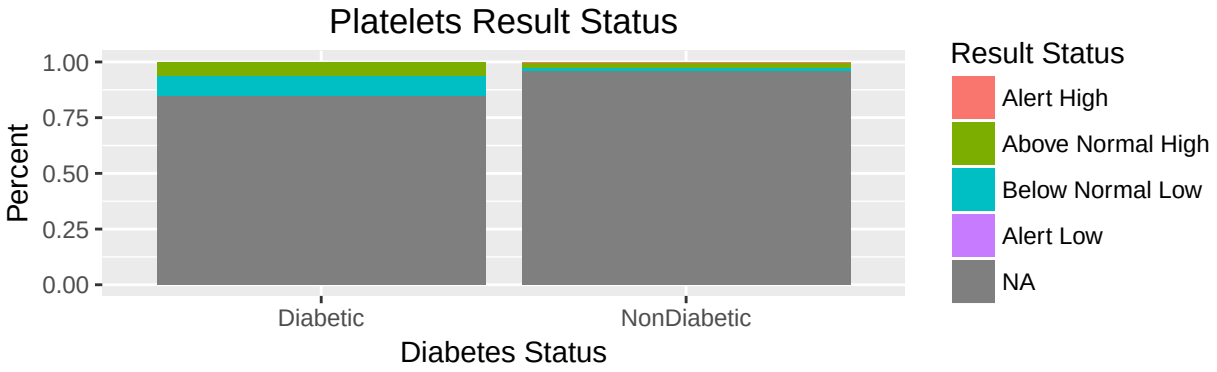


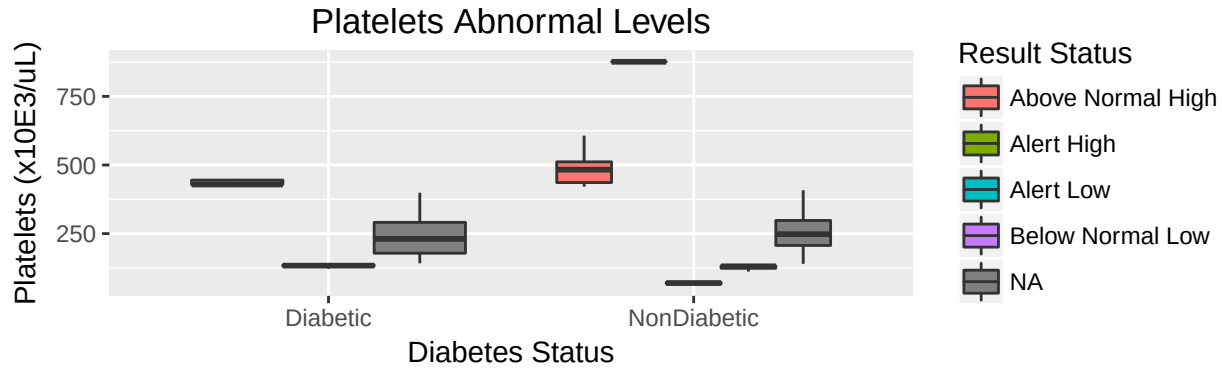


Platelets Lab Results

The central tendency of Platelet levels is lower in the diabetic population than in the non-diabetic population, particularly in the normal set. The ratio of both above and below normal readings are greater in the diabetic population, but the abnormal levels aren't as severe.







Approach

Before I can outline an approach, I need to review the materials on predictive analysis.