Project Notebook

Keith Engwall

1/23/2018

Project Objectives

Create a predictive model for identifying patients with Diabetes

See Capstone Proposal for details.

Project Questions

Does there tend to be a difference in age between diabetic patients and non-diabetic patients?

Is either gender more prone to diabetes?

Are there allergy types that correlate with diabetes?

Are there diagnosis categories that correlate with diabetes?

Are there medications that correlate with diabetes?

Are there characteristics (weight, blood pressure, lab results) that are significantly different in the diabetic population than in non-diabetics?

Is there any correlation between diabetes and smoking?

Project Dataset

Practice Fusion De-Identified Data Set containing EHR data for approximately 10,000 de-identified patients, including data points for diagnoses, medication, transcript data, and lab observations. See the Data Dictionary for details.

Project Notes

Load R Packages

The following packages are needed to work on the project.

Note that dbplyr is required to make database connections using dplyr functions.

```
library(tidyr)
library(dplyr)
library(dbplyr)
library(readr)
library(ggplot2)
library(Hmisc)
```

Load Data from csv files

Read the data files into R. Use select statements to pre-filter the desired columns.

```
# read patient table into data frame.
patient <- read_csv("db/training_patient.csv") %>%
  select(-PracticeGuid)

# read diagnosis table into data frame.
diagnosis <- read_csv("db/training_diagnosis.csv") %>%
  select(DiagnosisGuid, DiagnosisDescription, PatientGuid, ICD9Code, StartYear, StopYear, Acute)

# read allergy table into data frame. Change column name of MedicationNdcCode to AllergyMedicationNdcCo
allergy <- read_csv("db/training_allergy.csv") %>%
  select(AllergyGuid, PatientGuid, AllergyType, AllergyStartYear = StartYear, ReactionName, SeverityNam

# read prescription table into data frame.
prescription <- read_csv("db/training_prescription.csv") %>%
  select(PrescriptionGuid, PatientGuid, MedicationGuid, PrescriptionYear, Quantity, GenericAllowed)

# read medication table into data frame.
medication <- read_csv("db/training_medication.csv") %>%
  select(MedicationGuid, PatientGuid, MedicationNdcCode, MedicationName, MedicationStrength, Schedule, I

# read immunization table into data frame.
immunization <- read_csv("db/training_immunization.csv") %>%
  select(ImmunizationGuid, PatientGuid, VaccineName, AdministeredYear, CvxCode)

# read lab tables into data frames.
labResult <- read_csv("db/training_labResult.csv") %>%
  select(LabResultGuid,PatientGuid, TranscriptGuid)

labPanel <- read_csv("db/training_labPanel.csv") %>%
  select(LabResultGuid, LabPanelGuid, PanelName)

labObservation <- read_csv("db/training_labObservation.csv")


# read transcript table into data frame.
transcript <- read_csv("db/training_transcript.csv") %>%
  select(TranscriptGuid, PatientGuid, VisitYear, Height, Weight, BMI, SystolicBP, DiastolicBP, Respirat

# read smoke table into data frame.
smoke <- read_csv("db/training_smoke.csv") %>%
  select(PatientGuid, SmokeEffectiveYear, SmokingStatus_Description, SmokingStatus_NISTCode)

#read join tables into data frames.
transDiag <- read_csv("db/training_transcriptDiagnosis.csv")
transMed <- read_csv("db/training_transcriptMedication.csv")
transAllergy <- read_csv("db/training_transcriptAllergy.csv")
```

Data Cleaning

To prepare the data for analysis, add columns and make other modifications to the data.

```r
# add age column to patient data frame: derived from subtracting 2010, the median year of the patient d
patient <- patient %>% mutate(age = 2010 - YearOfBirth)

# add daibetesStatus column to patient data frame for display in graphs.
patient$diabetesStatus <- ifelse(patient$dmIndicator, "Diabetic", "NonDiabetic")

# change type of AllergyMedicationNdcCode column to character.
allergy <- transform(allergy, AllergyMedicationNdcCode = as.character(AllergyMedicationNdcCode))

# change transcript Height & Weight to numeric types
transcript <- transform(transcript, Height = as.numeric(Height), Weight = as.numeric(Weight), Temperatu

# add pulsePressure column to transcript (SystolicBP - DiastolicBP)
transcript <- transcript %>%
  filter(!is.na(SystolicBP)) %>% filter(!is.na(DiastolicBP)) %>% filter(SystolicBP > 0 & DiastolicBP > 0
  mutate(pulsePressure = SystolicBP - DiastolicBP)
```

Prepare data frames

```r
# prescription data frame
presMed <- left_join(prescription, medication)
presJoin <- left_join(presMed, transMed) %>% select(-TranscriptMedicationGuid)
presTran <- left_join(presJoin, transcript)

# diagnosis data frame
diagJoin <- left_join(diagnosis, transDiag) %>% select(-TranscriptDiagnosisGuid)
diagTran <- left_join(diagJoin, transcript)

# allergy data frame
allerJoin <- left_join(allergy, transAllergy) %>% select(-TranscriptAllergyGuid)
allerTran <- left_join(allerJoin, transcript)

# labs data frame
labs <- left_join(
  left_join(labResult, labPanel, by="LabResultGuid"),
  labObservation,
  by="LabPanelGuid"
)

# smoking data frame
smoke <- smoke %>% mutate(TranscriptGuid = "SMOKE")

# master data frame (comprised of all data frames)
dataJoin <- full_join(presJoin, diagJoin)
dataJoin <- full_join(dataJoin, allerJoin)
dataJoin <- full_join(dataJoin, labs)
dataJoin <- full_join(dataJoin, smoke)
dataMaster <- left_join(dataJoin, transcript)

# patient data frames (join patient table to each data frame)
patientPrescription <- left_join(patient, presTran)
patientDiagnosis <- left_join(patient, diagTran)
```

```r
# patientAllergy <- inner_join(patient, allerTran)
patientAllergy <- inner_join(patient, allerTran)
# patientLabs <- inner_join(patient, labs)
patientLabs <- left_join(patient, labs)
patientSmoke <- left_join(patient, smoke)
patientMaster <- left_join(patient, dataMaster)
# patientTranscript <- inner_join(patient, transcript)
patientTranscript <- left_join(patient, transcript)

# diabetic data frames filter patient data frames based on dmIndicator == 1
diabeticBase <- patient %>% filter(dmIndicator == 1)
diabeticPrescription <- patientPrescription %>% filter(dmIndicator == 1)
diabeticDiagnosis <- patientDiagnosis %>% filter(dmIndicator == 1)
diabeticAllergy <- patientAllergy %>% filter(dmIndicator == 1)
diabeticLabs <- patientLabs %>% filter(dmIndicator == 1)
diabeticSmoke <- patientSmoke %>% filter(dmIndicator == 1)
diabeticMaster <- patientMaster %>% filter(dmIndicator == 1)
```

Additional Data Cleaning

Due to the large amount of medication allergy data, we must filter the data down to medications for which at least one diabetic patient has an allergy and for which at least 20 patients have allergies

```r
# create medicationMap data frame linking medication names to their NDC Codes
medicationMap <- medication %>% select(MedicationNdcCode, MedicationName) %>% group_by(MedicationNdcCod

# use inner join to filter patients to those
# with medication allergies, and pull in the names for the medications
allergyMeds <- inner_join(patientAllergy,medicationMap, by = c("AllergyMedicationNdcCode" = "MedicationN

# identify the medications by name for which diabetic patients have allergies
diabeticMedNames <- allergyMeds %>% filter(dmIndicator == "1") %>% select(MedicationName) %>% distinct()

# use inner join to filter patients to those using the medications for which
# diabetic patients also have allergies (filter out all medication allergies
# for which diabetic patients do not have allergies)
diabeticAllergyMeds <- inner_join(allergyMeds,diabeticMedNames)

# identify the medications for which at least 20 patients have allergies
topAllergyNdcCodes <- diabeticAllergyMeds %>%
  group_by(AllergyMedicationNdcCode) %>%
  summarise(n = n()) %>%
  ungroup() %>%
  filter(n >= 20) %>%
  select(AllergyMedicationNdcCode)

# use inner join to filter data to those medications for which at least 20
# patients have allergies
diabeticAllergyMeds <- inner_join(allergyMeds,topAllergyNdcCodes)

# identify which medications are most used by diabetic patients in comparison to non-diabetic patients

# get a count of prescriptions for medications used by diabetic patients
```

```
diabeticMedicationList <- patientPrescription %>% filter(dmIndicator == 1) %>% group_by(MedicationName)

# join with a count of prescriptions for medications used by all patients
diabeticMedicationList <- inner_join(diabeticMedicationList, patientPrescription %>% group_by(Medicatio

# get the ratio between diabetic prescriptions and total prescriptions
diabeticMedicationList <- diabeticMedicationList %>%
  mutate(useRatio = n.x/n.y)

# had to tweak the filter to get a reasonably small set of the medications with the highest ratio of di
topDiabeticMedicationList <- diabeticMedicationList %>% filter(n.y > 300 & useRatio > .6) %>% arrange(d

# create a data frame limited to the top diabetic prescription list
topDiabeticPrescriptions <- inner_join(patientPrescription, topDiabeticMedicationList, by="MedicationNa

# sort the abnormal statuses in the labs data
patientLabs$AbnormalFlagsSorted = factor(patientLabs$AbnormalFlags, levels = rev(c("Panic Low", "Alert
```

For diagnoses, rather than identify each individual diagnosis type, we can create categories of diagnoses based on the ICD9 codes

```
# create diagCat column in patientDiagnosis containing diagnosis categories corresponding to ranges of
patientDiagnosis$diagCat <-
  ifelse((as.integer(patientDiagnosis$ICD9Code) < 140),
    "Infectious/Parasitic",
    ifelse((as.integer(patientDiagnosis$ICD9Code) >= 140 &
      as.integer(patientDiagnosis$ICD9Code) < 240),
      "Neoplasms",
      ifelse((as.integer(patientDiagnosis$ICD9Code) >=240 &
        as.integer(patientDiagnosis$ICD9Code) < 280),
        "Endocrine/Nutritional/Metabolic",
        ifelse((as.integer(patientDiagnosis$ICD9Code) >= 280 &
          as.integer(patientDiagnosis$ICD9Code) < 290),
          "Blood",
          ifelse((as.integer(patientDiagnosis$ICD9Code) >= 290 &
            as.integer(patientDiagnosis$ICD9Code) < 320),
            "Mental",
            ifelse((as.integer(patientDiagnosis$ICD9Code) >= 320 &
              as.integer(patientDiagnosis$ICD9Code) < 390),
              "Nervous",
              ifelse((as.integer(patientDiagnosis$ICD9Code) >= 390 &
                as.integer(patientDiagnosis$ICD9Code) < 460),
                "Circulatory",
                ifelse((as.integer(patientDiagnosis$ICD9Code) >= 460 &
                  as.integer(patientDiagnosis$ICD9Code) < 520),
                  "Respiratory",
                  ifelse((as.integer(patientDiagnosis$ICD9Code) >= 520 &
                    as.integer(patientDiagnosis$ICD9Code) < 580),
                    "Digestive",
  ifelse((as.integer(patientDiagnosis$ICD9Code) >= 580 &
    as.integer(patientDiagnosis$ICD9Code) < 630),
    "Genitourinary",
    ifelse((as.integer(patientDiagnosis$ICD9Code) >= 630 &
      as.integer(patientDiagnosis$ICD9Code) < 680),
```

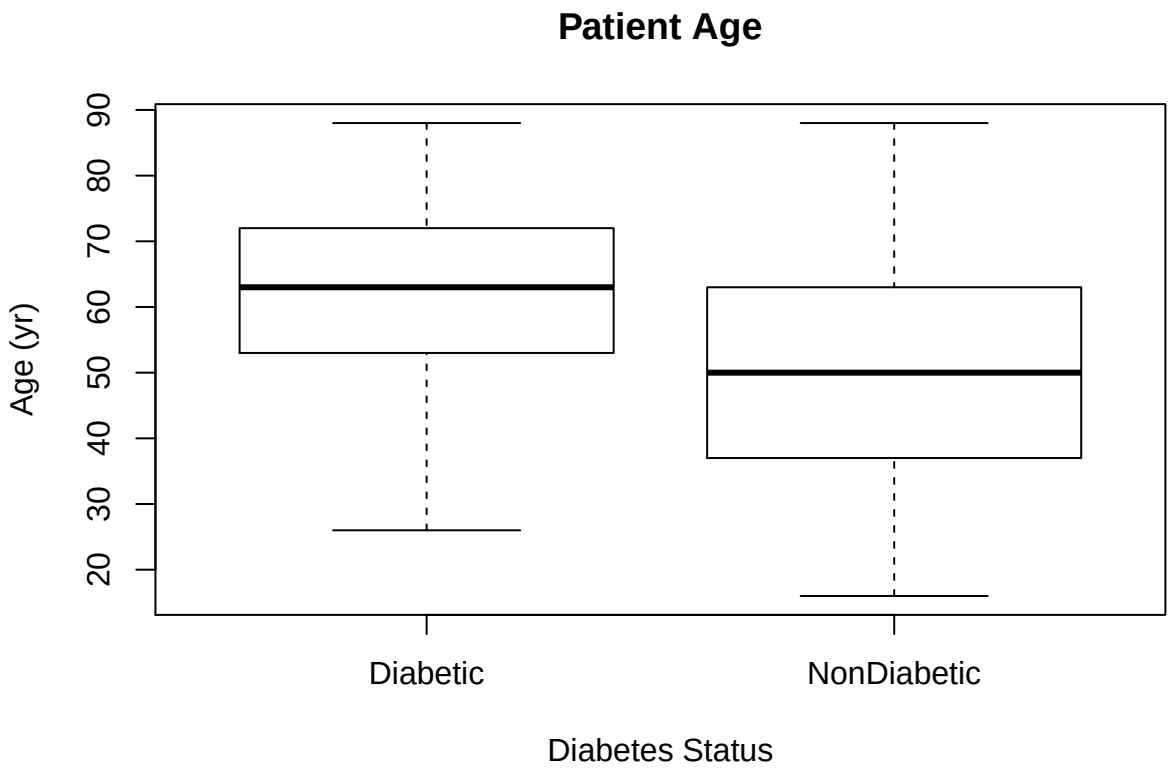
```
      "Pregnancy",
      ifelse((as.integer(patientDiagnosis$ICD9Code) >=680 &
        as.integer(patientDiagnosis$ICD9Code) < 710),
          "Skin",
          ifelse((as.integer(patientDiagnosis$ICD9Code) >=710 &
            as.integer(patientDiagnosis$ICD9Code) < 740),
            "Musculoskeletal",
            ifelse((as.integer(patientDiagnosis$ICD9Code) >= 740 &
              as.integer(patientDiagnosis$ICD9Code) < 760),
              "Congenital",
              ifelse((as.integer(patientDiagnosis$ICD9Code) >= 760 &
                as.integer(patientDiagnosis$ICD9Code) < 780),
                "Perinatal",
                ifelse((as.integer(patientDiagnosis$ICD9Code) >= 780 &
                  as.integer(patientDiagnosis$ICD9Code) < 800),
                  "Ill Defined",
                  ifelse((as.integer(patientDiagnosis$ICD9Code) >= 800),
                    "Injury/Poisoning",
                    "NULL"
                    )
                  )
                )
            )
          )
      )
    )
))))))))))
```

Data analysis

Age

It appears that the central tendency for Age is higher in the diabetic population than in the non-diabetic population

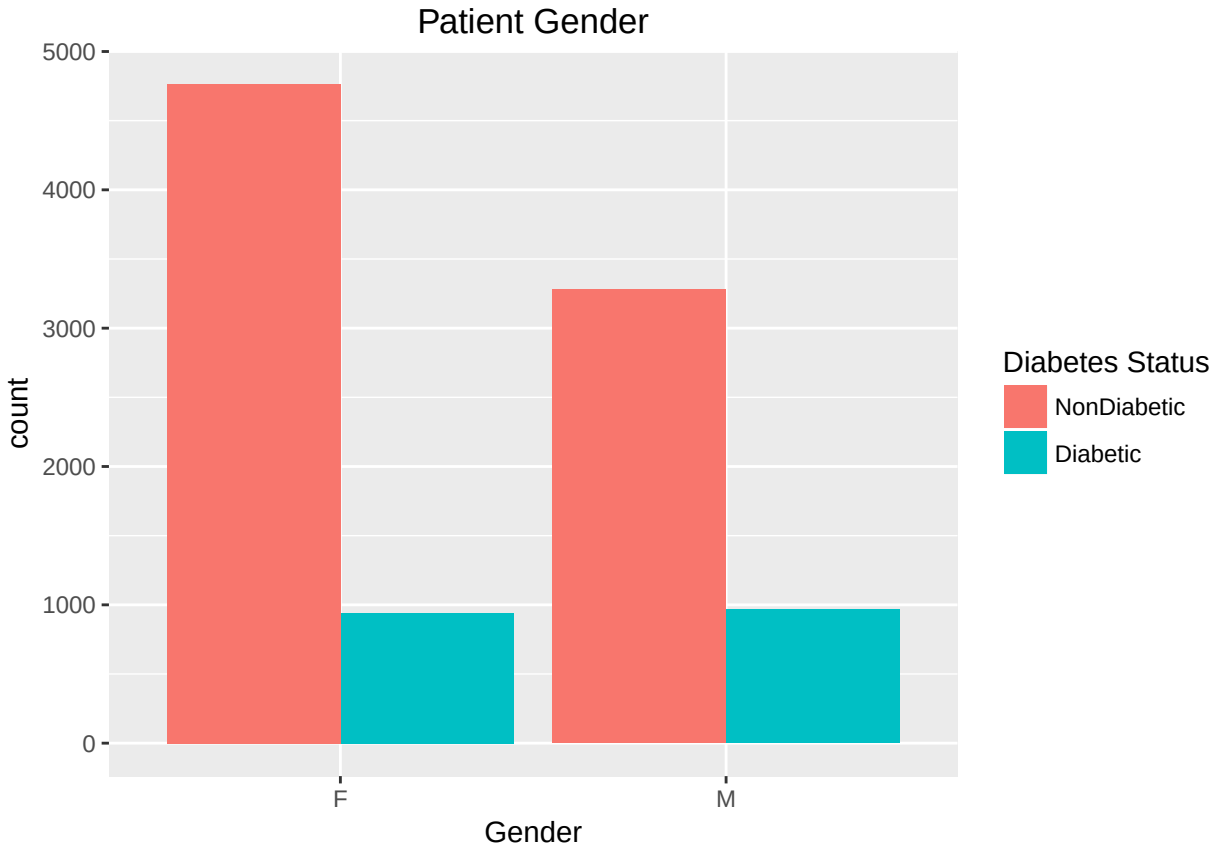
```
#ggplot(patient, aes(x = diabetesStatus, y = age)) +
#  geom_boxplot()

boxplot(age~diabetesStatus,data=patient, outline = FALSE, main = "Patient Age", ylab = "Age (yr)", xlab
```



Gender

The number of patients with diabetes is similar between males and females, but the ratio of males with diabetes is higher.

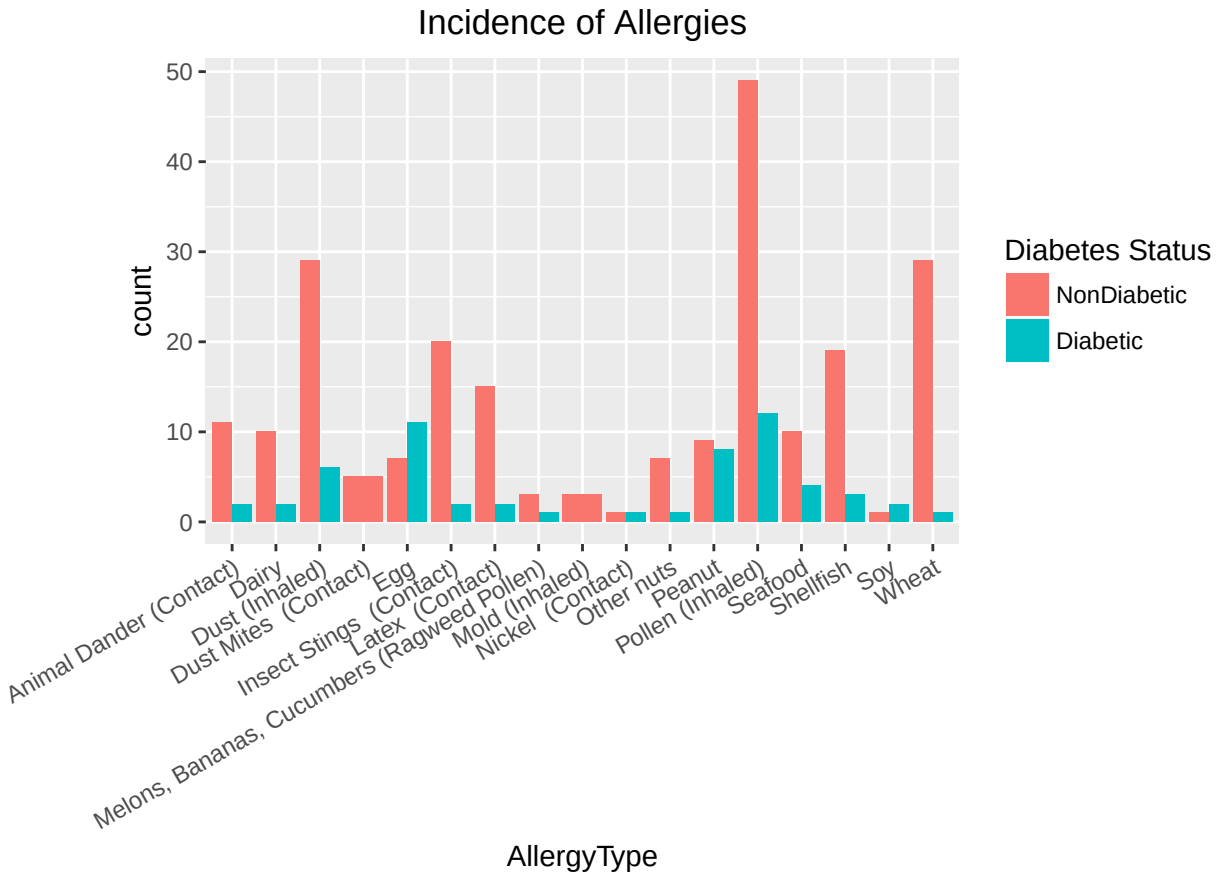
```
ggplot(arrange(patient, rev(dmIndicator)), aes(x=Gender,fill=factor(diabetesStatus, levels = c("NonDiab
  geom_bar(position = "dodge") +
  labs(title = "Patient Gender", fill="Diabetes Status") +
  theme(plot.title = element_text(hjust = "0.5"))
```



Allergies

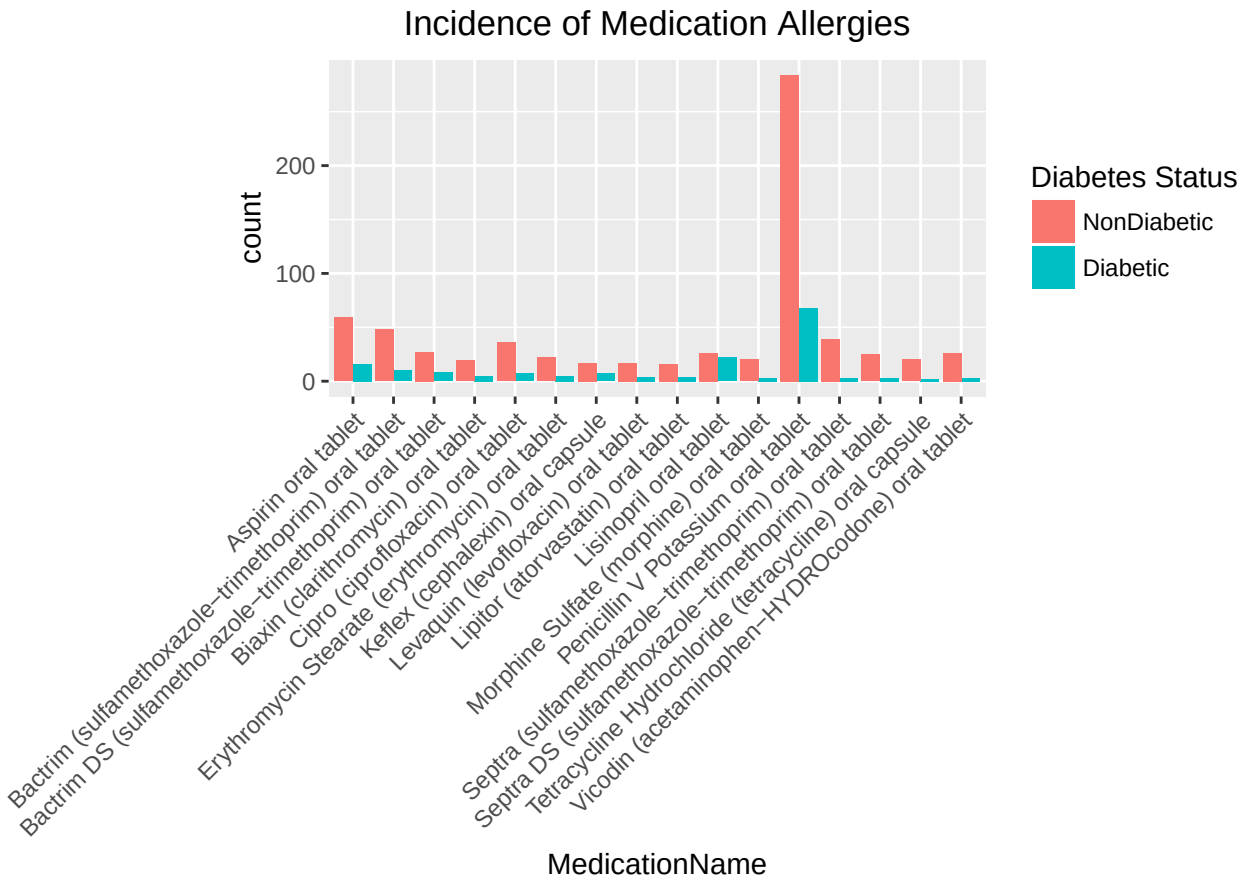
Among non-medical allergy types, there appears to be a large proportion of diabetic patients with egg and peanut allergies in comparison to the general population. The number of diabetic patients with egg allergies actually outnumbers non-diabetic patients.

```
patientAllergy %>%
  filter(AllergyType != "Medication") %>%
  ggplot(aes(x=AllergyType, fill=factor(diabetesStatus, levels = c("NonDiabetic","Diabetic"))))+
  geom_bar(position="dodge") +
  labs(title = "Incidence of Allergies", fill = "Diabetes Status") +
  theme(plot.margin = margin(0,0,0,2,"cm"), axis.text.x = element_text(angle = 30, hjust = 1), plot.titl
```



Among the allergies to medicine, there appears to be an large proportion of diabetic patients with an allergy to Lisinopril compared to the general population.

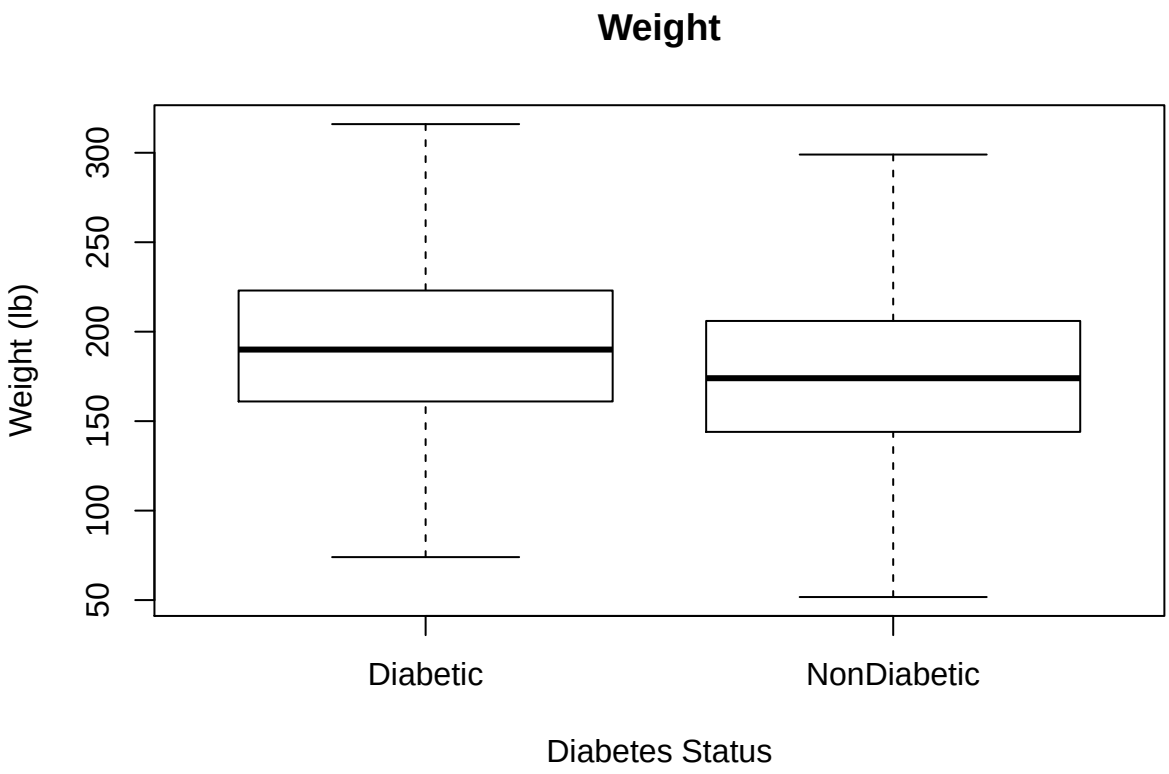
```
diabeticAllergyMeds %>%
  ggplot(aes(x=MedicationName, fill=factor(diabetesStatus, levels = c("NonDiabetic","Diabetic"))))+
  geom_bar(position="dodge") +
  labs(title="Incidence of Medication Allergies", fill = "Diabetes Status") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.margin = margin(0,0,0,3.1,"cm"),
        plot.title = element_text(hjust = 0.5))
```



Weight, Height & BMI

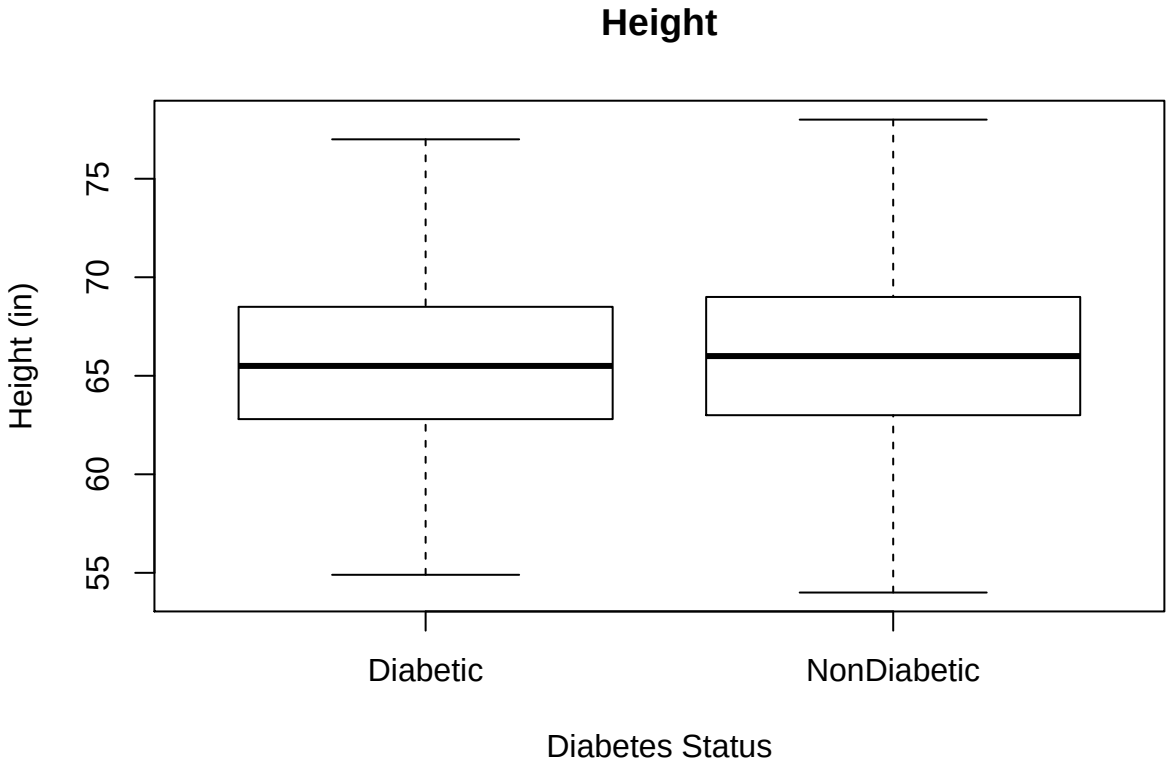
Diabetic patients have a slightly higher weight than non-diabetic patients.

```
#patientTranscript %>%
#  filter(Weight < 500 & Weight > 0) %>%
#  ggplot(aes(x=Weight,fill=diabetesStatus)) +
#  geom_histogram(binwidth=50, position="dodge")

boxplot(Weight~diabetesStatus, data=patientTranscript %>% filter(!is.na(Weight)) %>%filter(Weight < 600
```



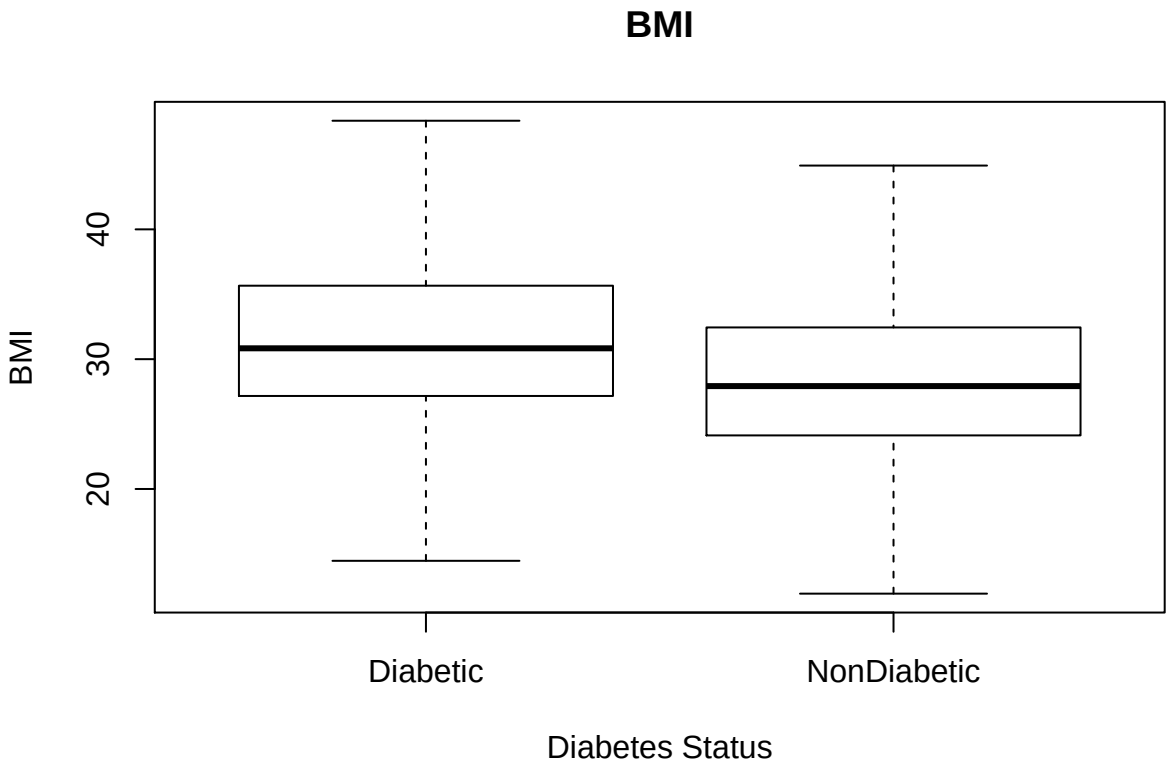
Diabetic patients have a slightly shorter height than non-diabetic patients.

```
#heightData = patientTranscript %>% distinct(PatientGuid,.keep_all = TRUE)
boxplot(Height~diabetesStatus, data = patientTranscript %>% filter(!is.na(Height)) %>% filter(Height <
```



Diabetic patients have a higher BMI than non-diabetic patients

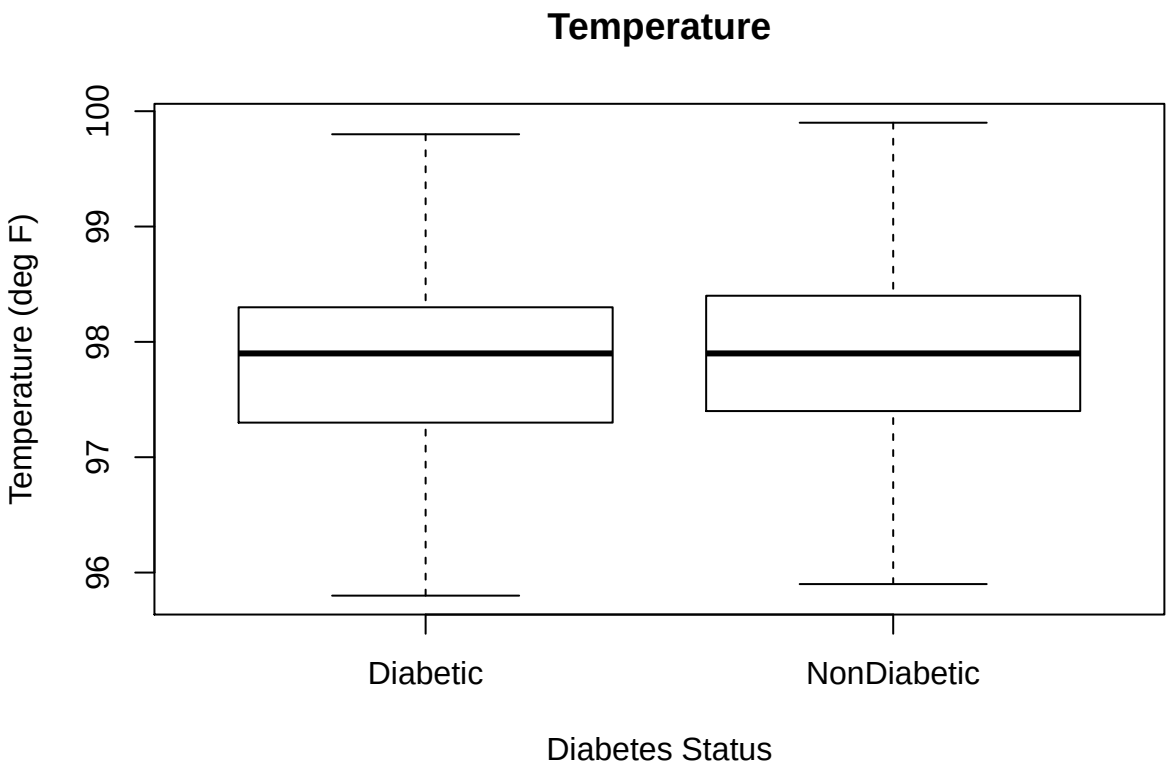
```
boxplot(BMI~diabetesStatus, data=patientTranscript %>% filter(!is.na(BMI)) %>% filter(BMI >0), outline=
```



Temperature

There is no significant difference in temperature between diabetic and non-diabetic patients.

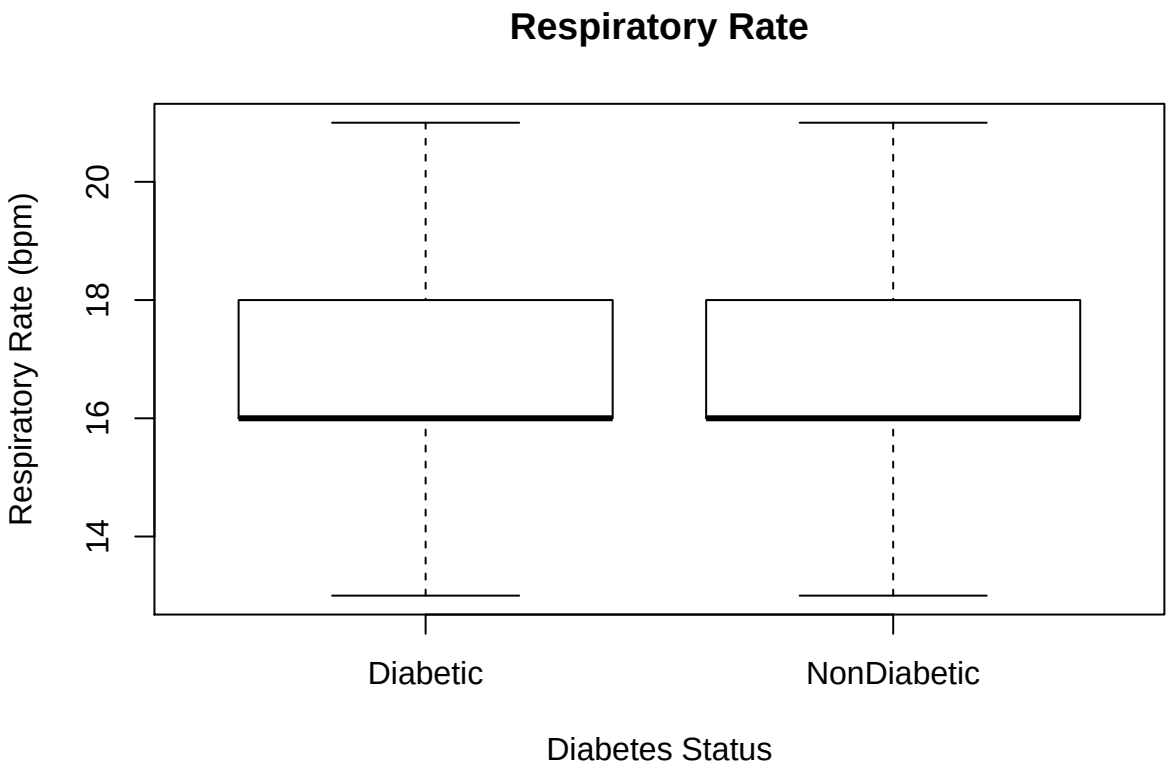
```
boxplot(Temperature~diabetesStatus, data=patientTranscript %>% filter(!is.na(Temperature)), outline = F
```



Respiratory Rate

There is no significant difference in respiratory rate between diabetic and non-diabetic patients.

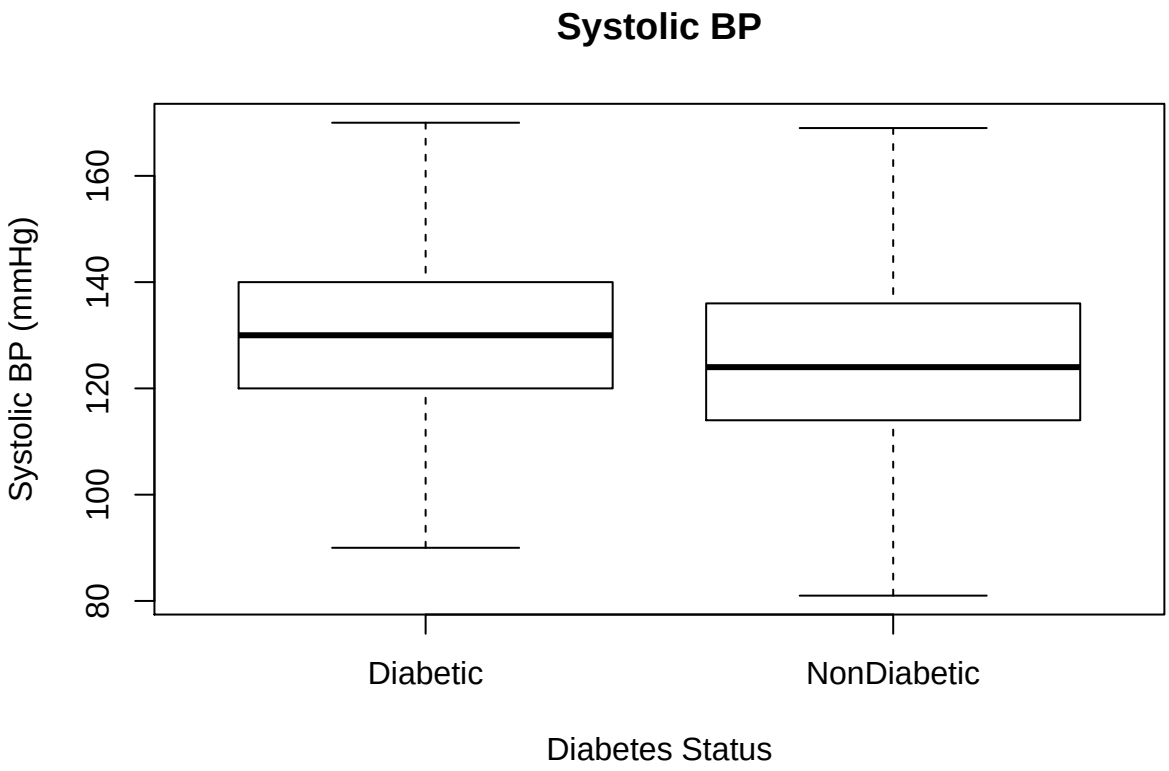
```
boxplot(RespiratoryRate~diabetesStatus, data=patientTranscript %>% filter(!is.na(RespiratoryRate)), outl
```



Systolic BP, Diastolic BP, and Pulse Pressure

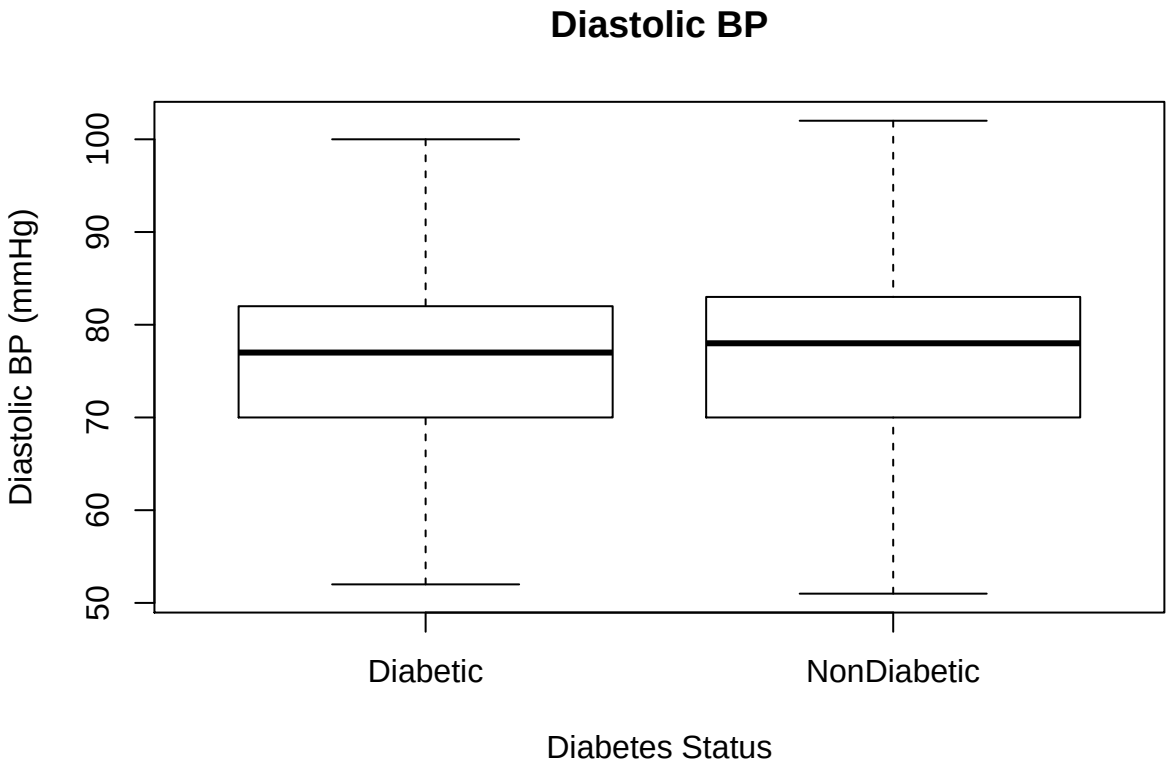
Systolic BP is higher for diabetic patients than non-diabetic patients.

```
boxplot(as.integer(SystolicBP)~diabetesStatus, data=patientTranscript %>% filter(!is.na(SystolicBP)) %>%
```



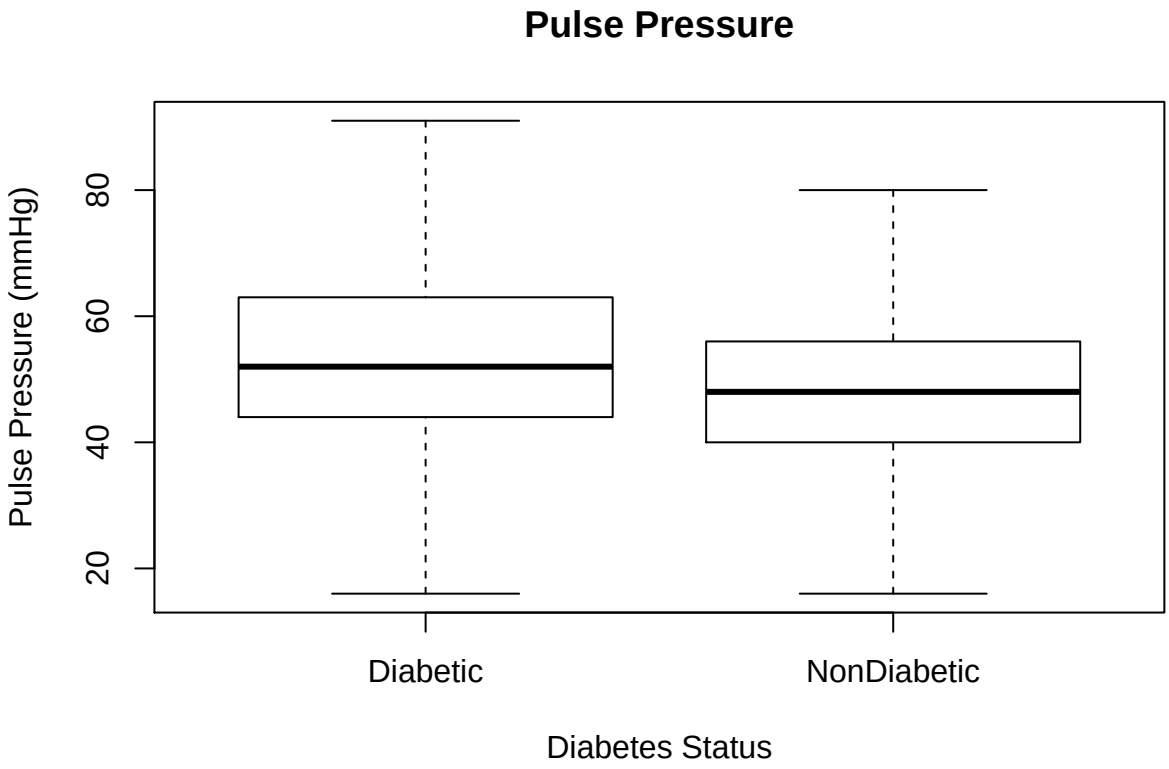
Diastolic BP is slightly lower for diabetic patients than non-diabetic patients.

```
boxplot(as.integer(DiastolicBP)~diabetesStatus, data=patientTranscript %>% filter(!is.na(DiastolicBP)) %
```



The pulse pressure of diabetic patients is slightly higher than non-diabetic patients.

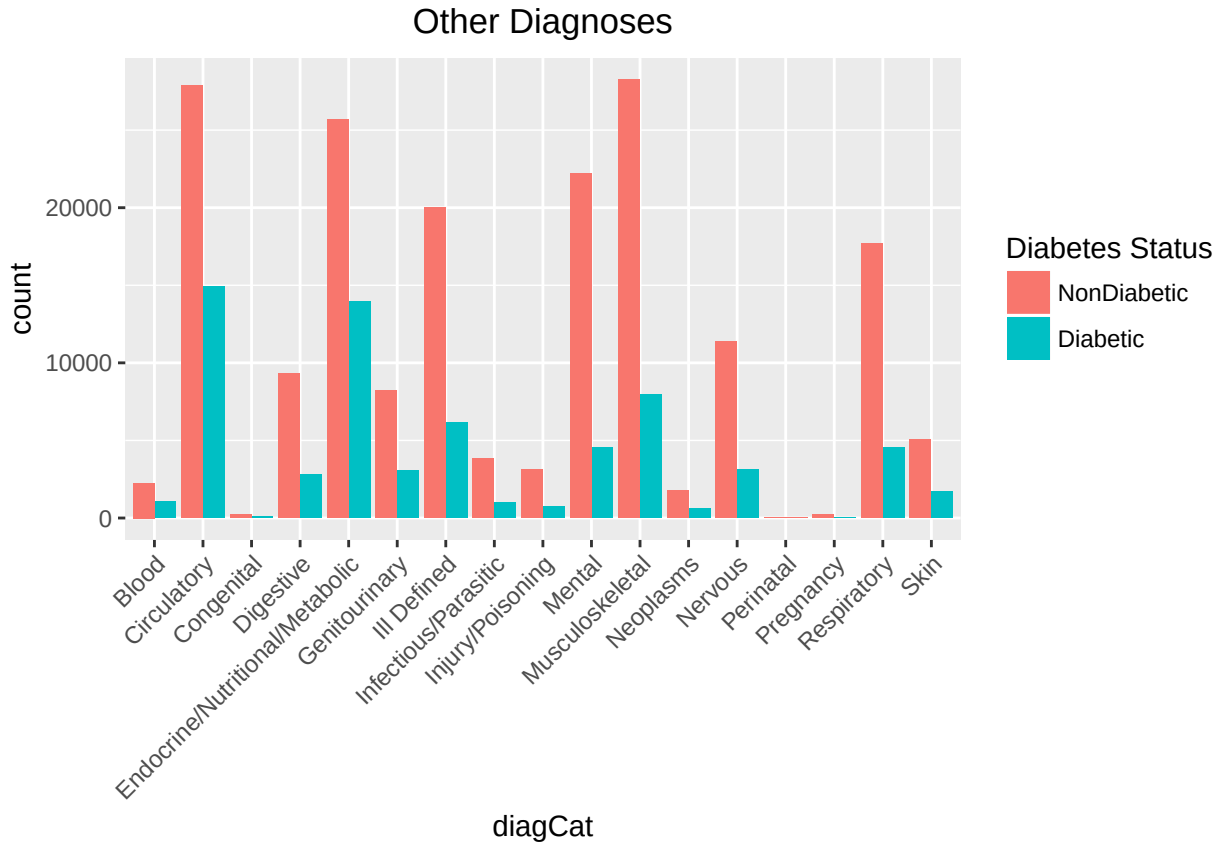
```
boxplot(as.integer(pulsePressure)~diabetesStatus, data=patientTranscript %>% filter(pulsePressure != "NU
```



Diagnosis Categories

The diagnosis categories with the highest ratio between diabetic and non-diabetic patients are Circulatory and Endocrine/Nutritional/Metabolic.

```
patientDiagnosis %>%
  filter(diagCat != "NULL") %>%
  ggplot(aes(x=diagCat, fill=factor(diabetesStatus, levels = c("NonDiabetic","Diabetic"))))+
  geom_bar(position="dodge") +
  labs(title="Other Diagnoses", fill = "Diabetes Status") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
#  labs(fill = "Diabetes Status") +
#  theme(axis.text.x = element_text(angle = 30, hjust = 1))
```

Abnormal Labs

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.

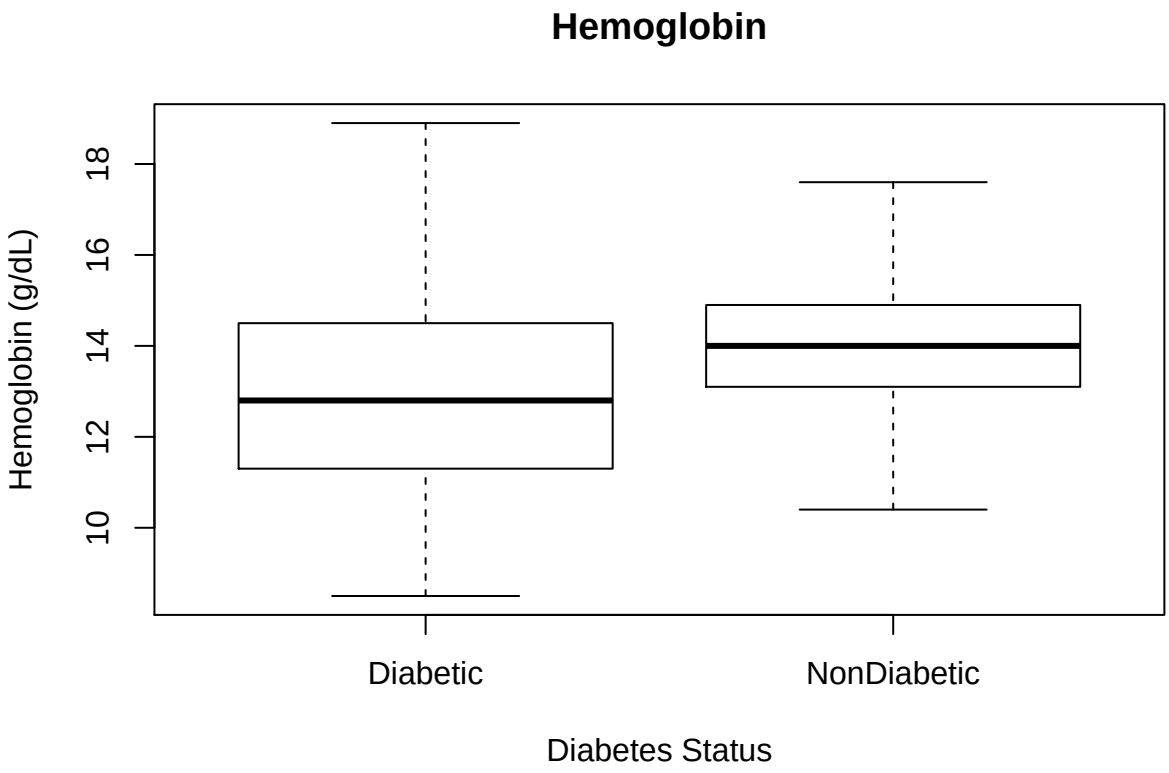
```
patientLabs %>% filter(dmIndicator == 1) %>% filter(IsAbnormalValue == 1) %>% group_by(HL7Text) %>% summ
```

```
## # A tibble: 58 x 2
##    HL7Text                    n
##    <chr>                  <int>
##  1 Hemoglobin                39
##  2 Hematocrit                31
##  3 Chloride                  14
##  4 Triglyceride              14
##  5 Platelets                 13
##  6 RBC DISTRIBUTION WIDTH     9
##  7 ABS SEGS                   8
##  8 DIFFERENTIAL: SEGS         8
##  9 EOSINOPHILS                8
## 10 LYMPHOCYTES                8
## # ... with 48 more rows
```

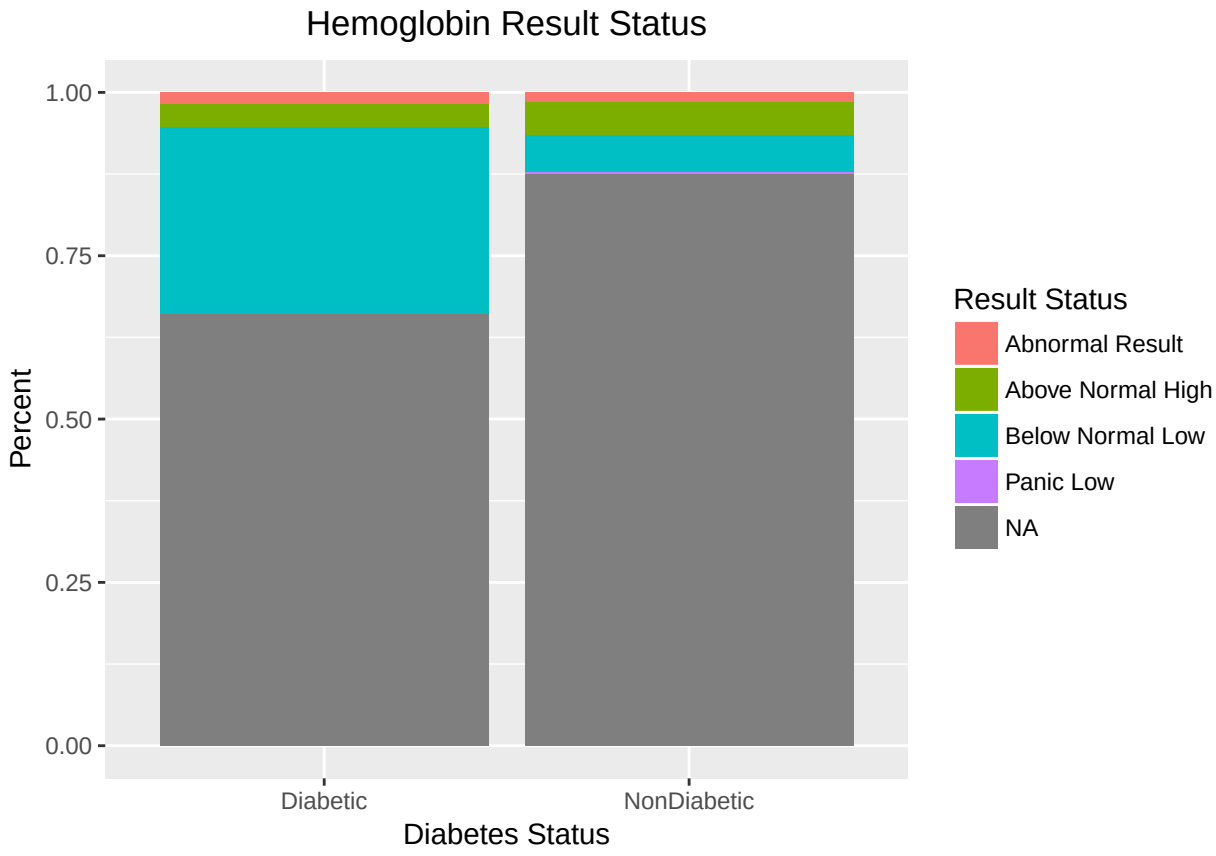
Hemoglobin Lab Results

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.

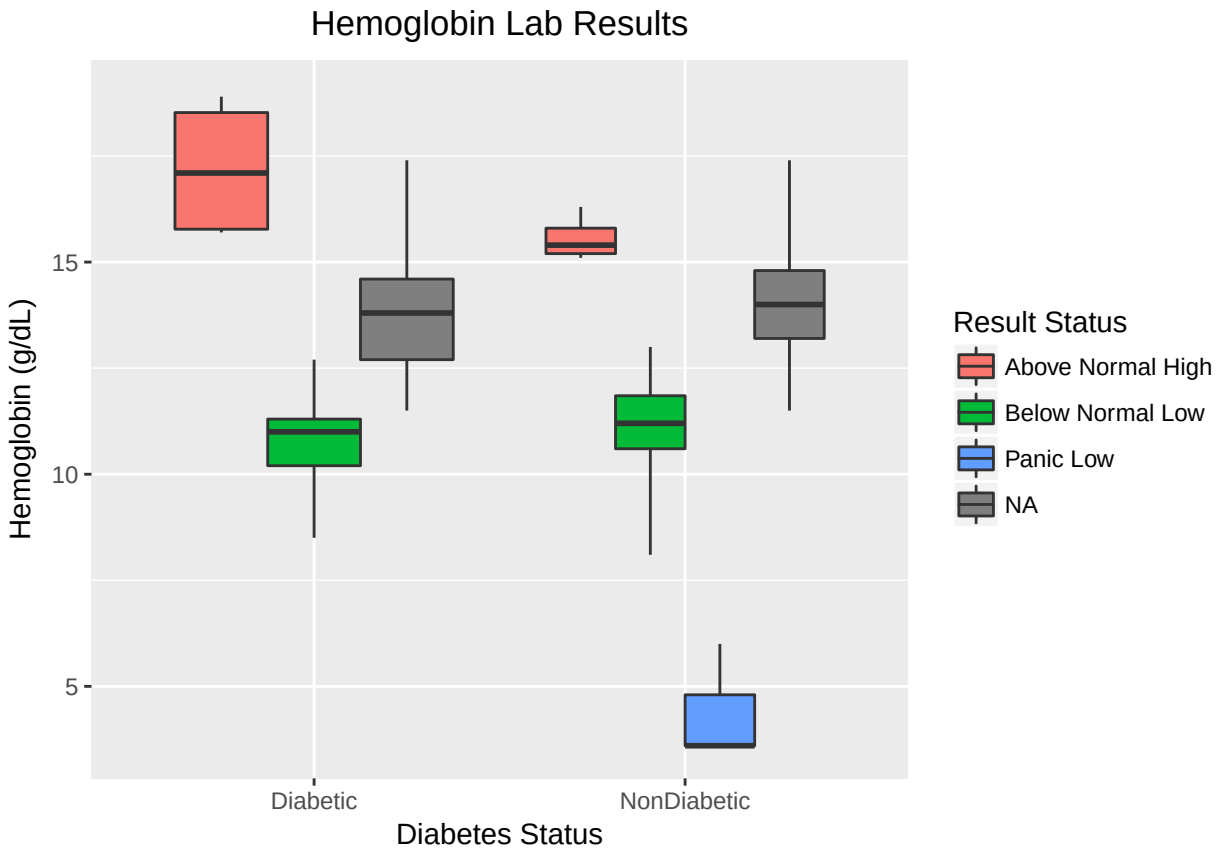
```
boxplot(ObservationValue~diabetesStatus, data=patientLabs %>% filter(HL7Text == "Hemoglobin"), outline=
```




```
patientLabs %>%
  filter(HL7Text == "Hemoglobin") %>%
  ggplot(aes(x=diabetesStatus, fill=AbnormalFlagsSorted))+
  geom_bar(position="fill") +
  labs(title = "Hemoglobin Result Status", x = "Diabetes Status", fill = "Result Status", y = "Percent"
  theme(plot.title = element_text(hjust = 0.5))
```



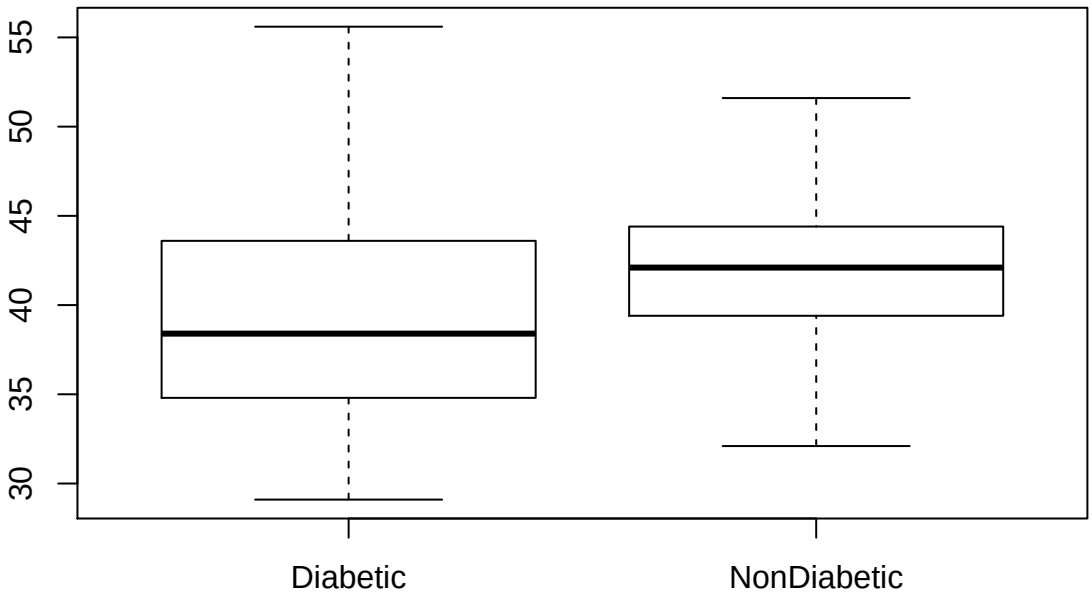

```
patientLabs %>%
  filter(HL7Text == "Hemoglobin") %>%
  ggplot(aes(x=diabetesStatus,y=ObservationValue, fill=AbnormalFlags)) +
  stat_boxplot(geom="boxplot", position="dodge", coef=2, outlier.shape=NA, na.rm=TRUE, show.legend = TRU
  labs(title = "Hemoglobin Lab Results", x = "Diabetes Status", fill = "Result Status", y = "Hemoglobin
  theme(plot.title = element_text(hjust = 0.5))
```



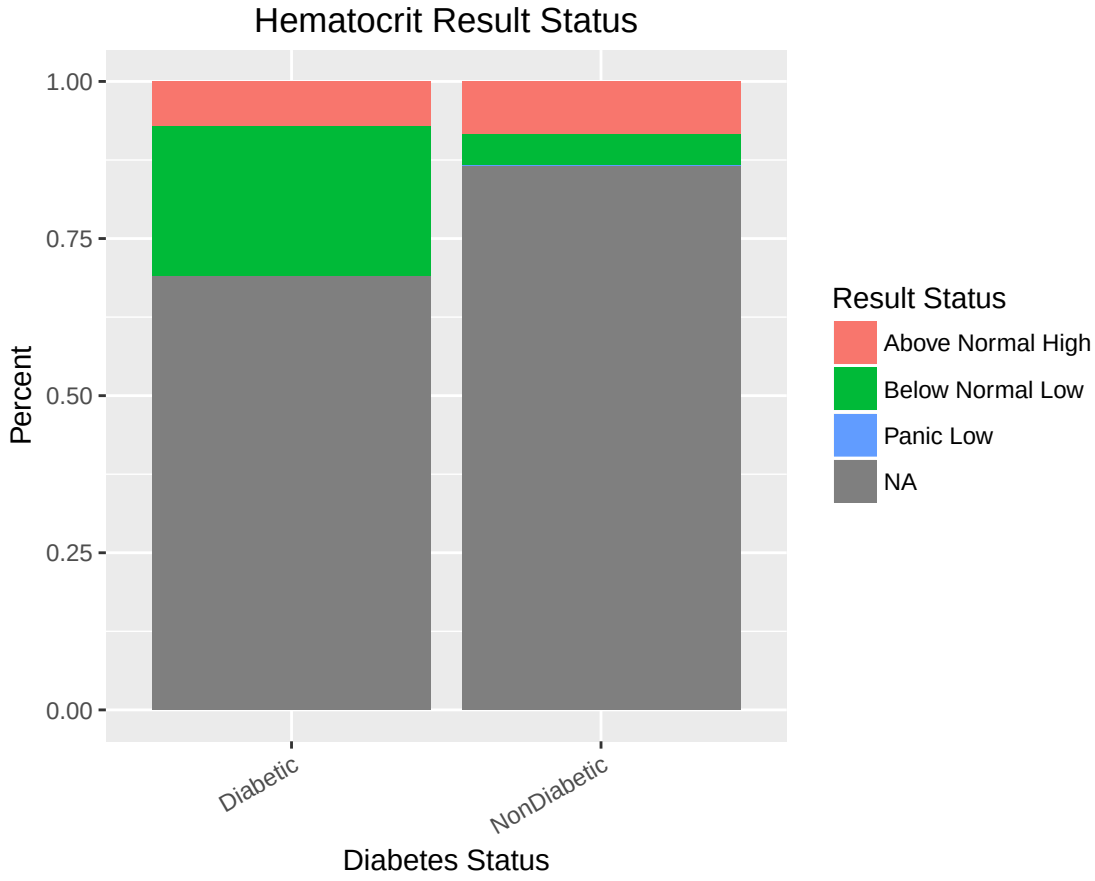
Hematocrit Lab Results

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.

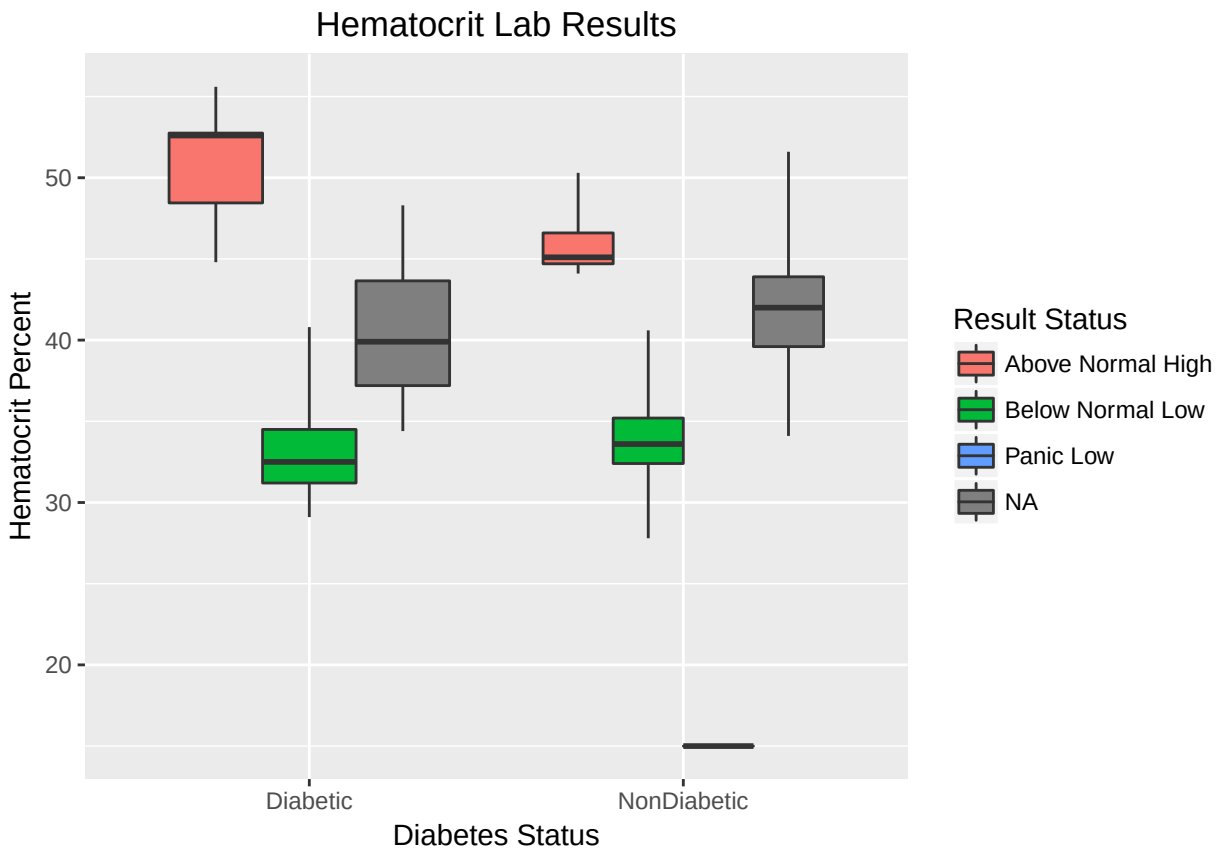
```
boxplot(ObservationValue~diabetesStatus, data=patientLabs %>% filter(HL7Text == "Hematocrit"), outline=
```




```
patientLabs %>%
  filter(HL7Text == "Hematocrit") %>%
  ggplot(aes(x=diabetesStatus, fill=AbnormalFlagsSorted))+
  geom_bar(position="fill") +
  labs(title = "Hematocrit Result Status", x = "Diabetes Status", fill = "Result Status", y = "Percent")
  theme(plot.margin = margin(0,0,0,2,"cm"), axis.text.x = element_text(angle = 30, hjust = 1), plot.titl
```



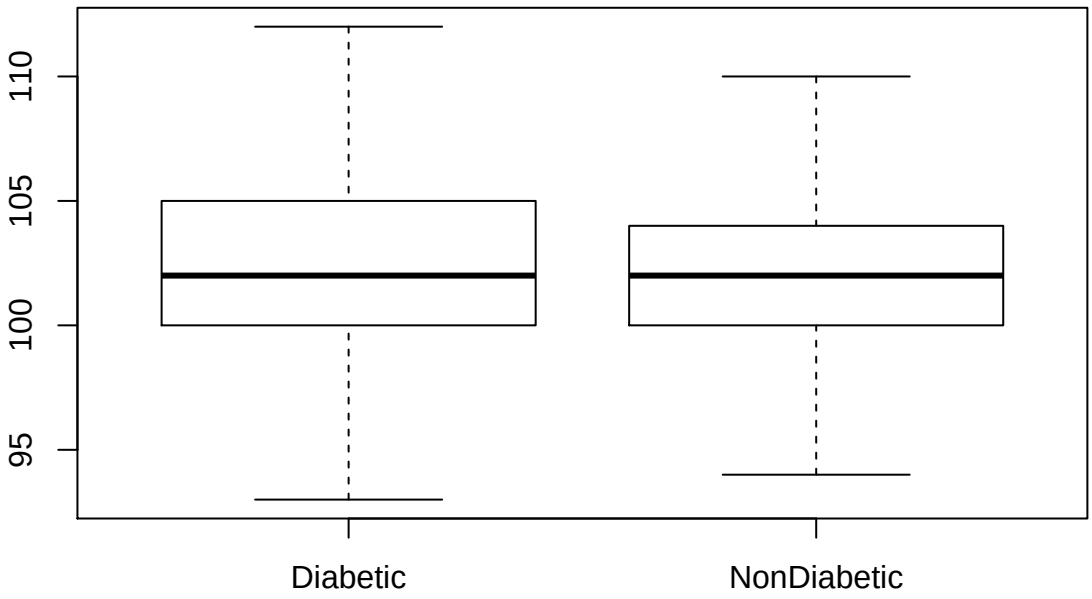
```
patientLabs %>%
  filter(HL7Text == "Hematocrit") %>%
  ggplot(aes(x=diabetesStatus,y=ObservationValue, fill=AbnormalFlags)) +
  stat_boxplot(geom="boxplot", position="dodge", coef=2, outlier.shape=NA, na.rm=TRUE, show.legend = TRU
  labs(title = "Hematocrit Lab Results", x = "Diabetes Status", fill = "Result Status", y = "Hematocrit
  theme(plot.title = element_text(hjust = 0.5))
```



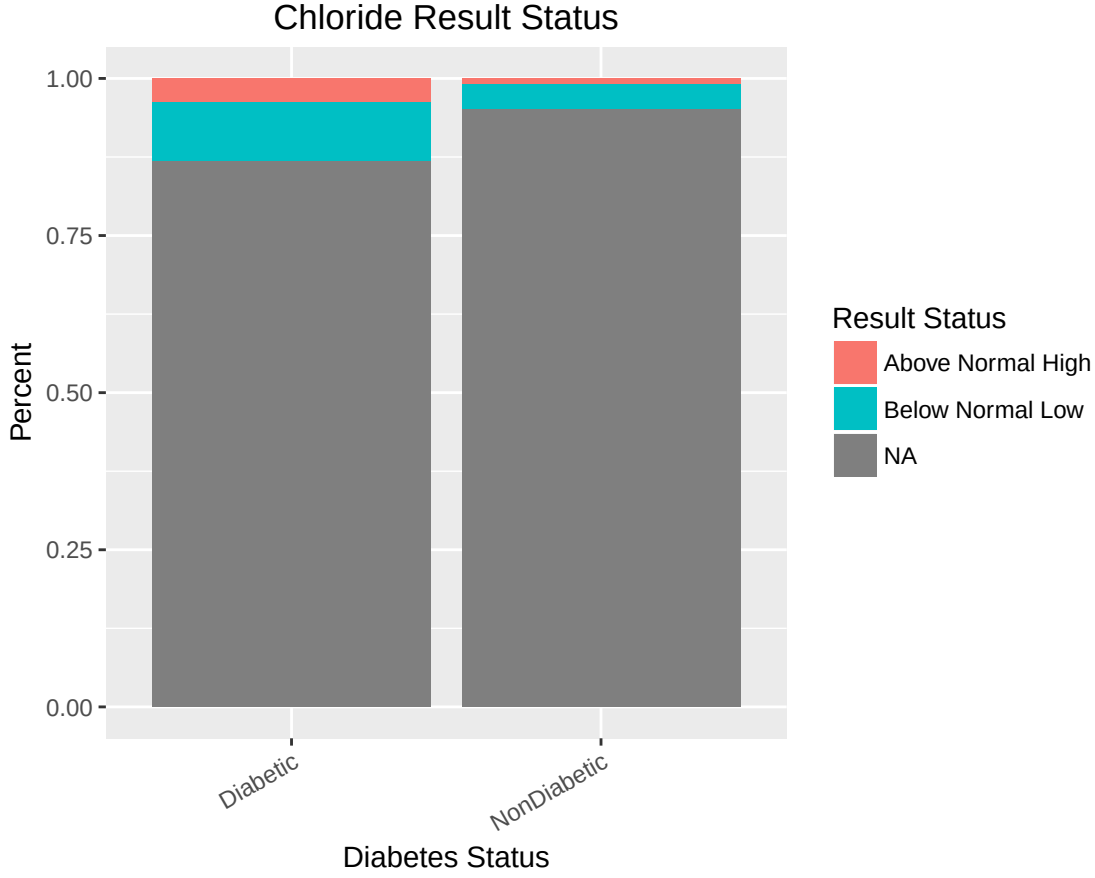
Chloride Lab Results

The Chloride lab results are more spread out for the diabetic population than for the non-diabetic population. There's a larger ratio of both above and below normal readings in the diabetic population, but the abnormal levels aren't as severe.

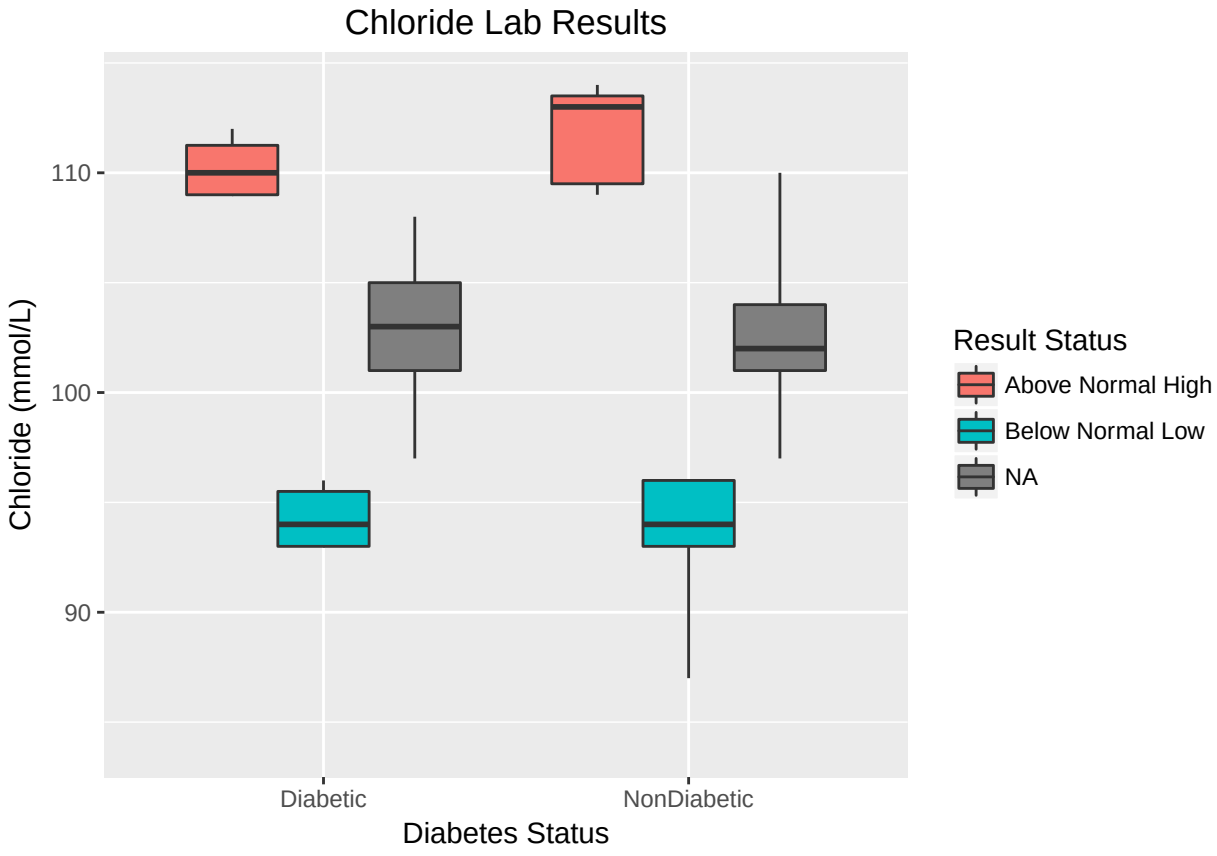
```
boxplot(ObservationValue~diabetesStatus, data=patientLabs %>% filter(HL7Text == "Chloride"), outline=FA
```




```
patientLabs %>%
  filter(HL7Text == "Chloride") %>%
  ggplot(aes(x=diabetesStatus, fill=AbnormalFlagsSorted))+
  geom_bar(position="fill") +
  labs(title = "Chloride Result Status", x = "Diabetes Status", fill = "Result Status", y = "Percent") +
  theme(plot.margin = margin(0,0,0,2,"cm"), axis.text.x = element_text(angle = 30, hjust = 1), plot.titl
```



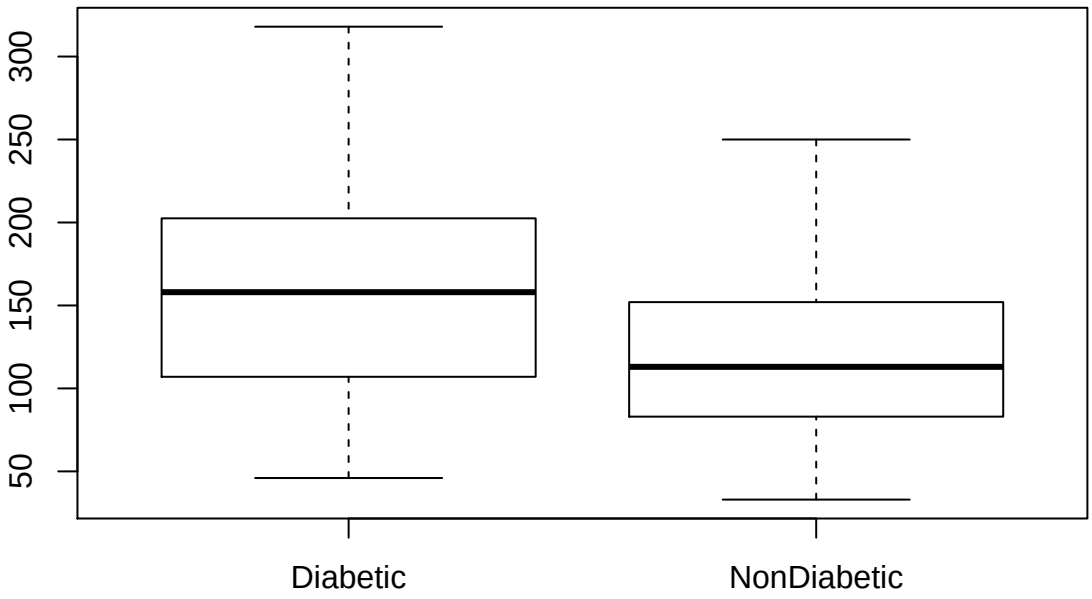
```
patientLabs %>%
  filter(HL7Text == "Chloride") %>%
  ggplot(aes(x=diabetesStatus,y=ObservationValue, fill=AbnormalFlags)) +
  stat_boxplot(geom="boxplot", position="dodge", coef=2, outlier.shape=NA, na.rm=TRUE, show.legend = TRU
  labs(title = "Chloride Lab Results", x = "Diabetes Status", fill = "Result Status", y = "Chloride (mm
  theme(plot.title = element_text(hjust = 0.5))
```



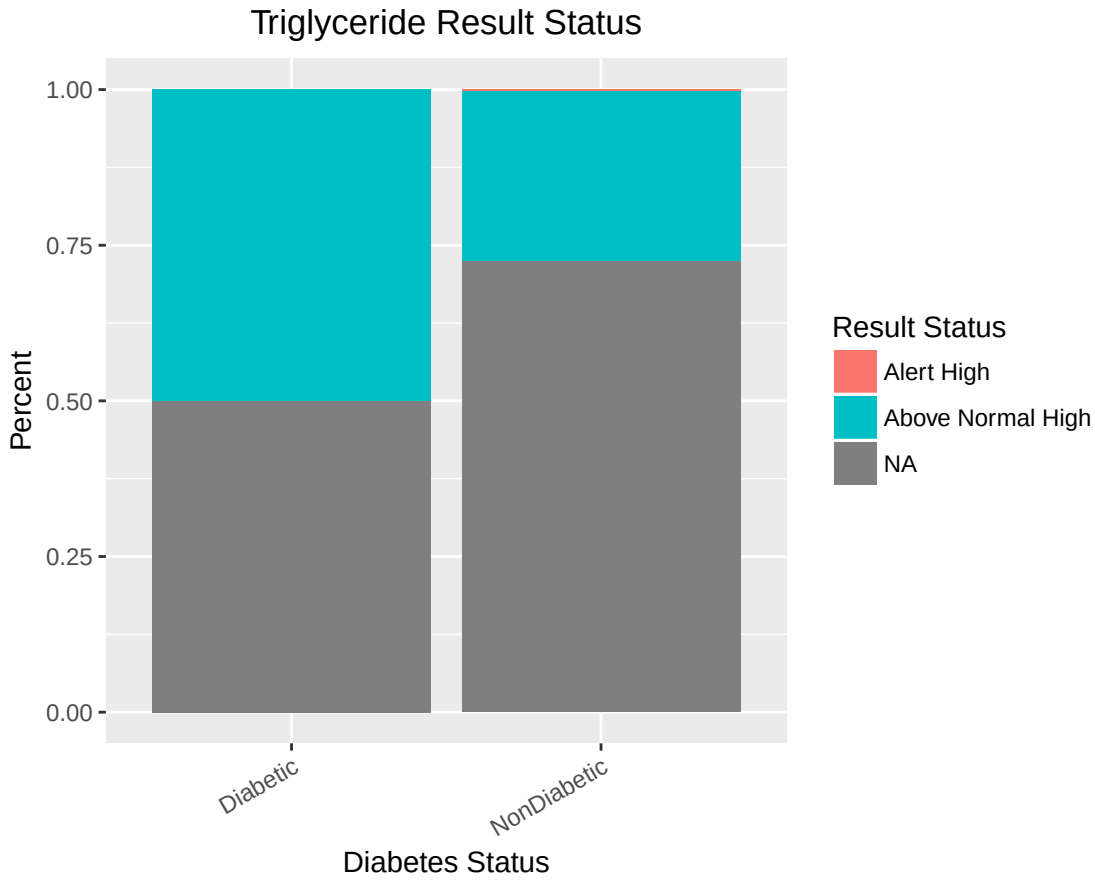
Triglyceride Lab Results

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.

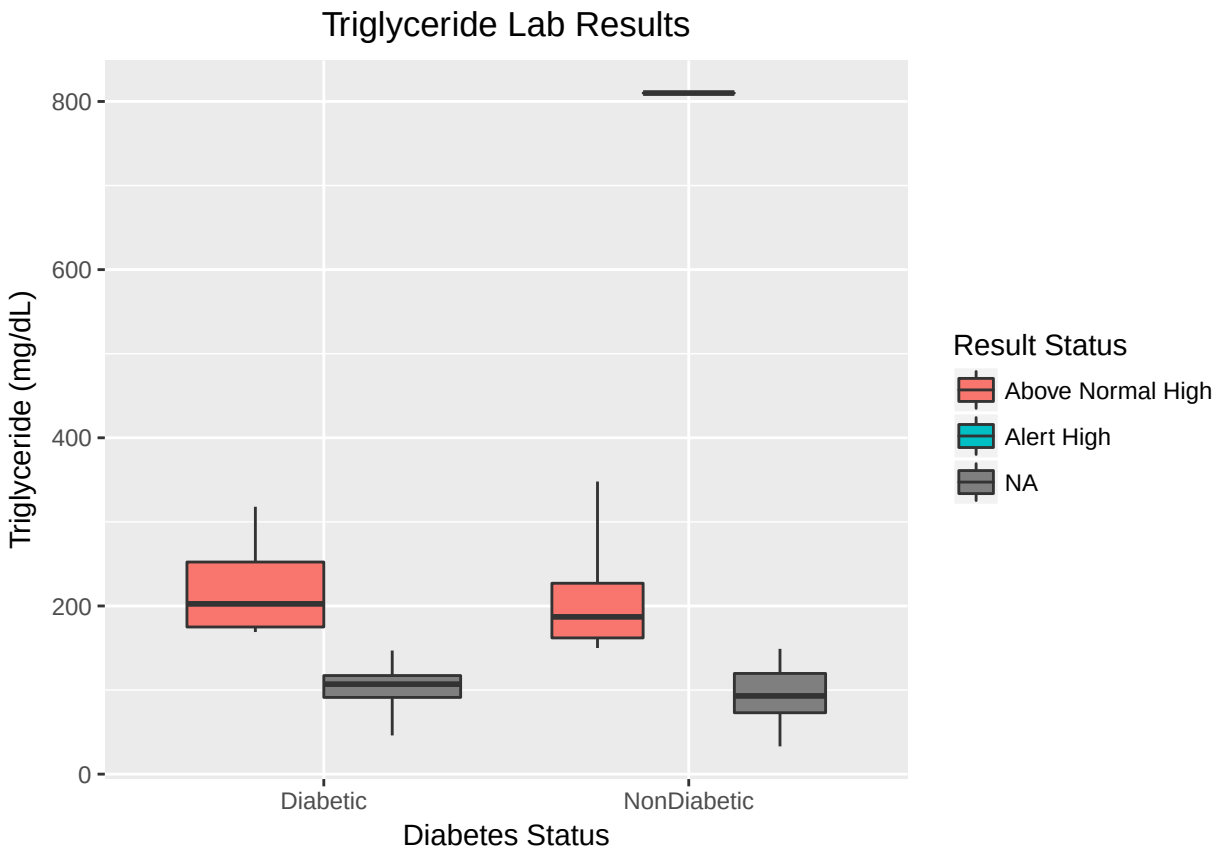
```
boxplot(ObservationValue~diabetesStatus, data=patientLabs %>% filter(HL7Text == "Triglyceride"), outlin
```




```
patientLabs %>%
  filter(HL7Text == "Triglyceride") %>%
  ggplot(aes(x=diabetesStatus, fill=AbnormalFlagsSorted))+
  geom_bar(position="fill") +
  labs(title = "Triglyceride Result Status", x = "Diabetes Status", fill = "Result Status", y = "Percen
  theme(plot.margin = margin(0,0,0,2,"cm"), axis.text.x = element_text(angle = 30, hjust = 1), plot.titl
```



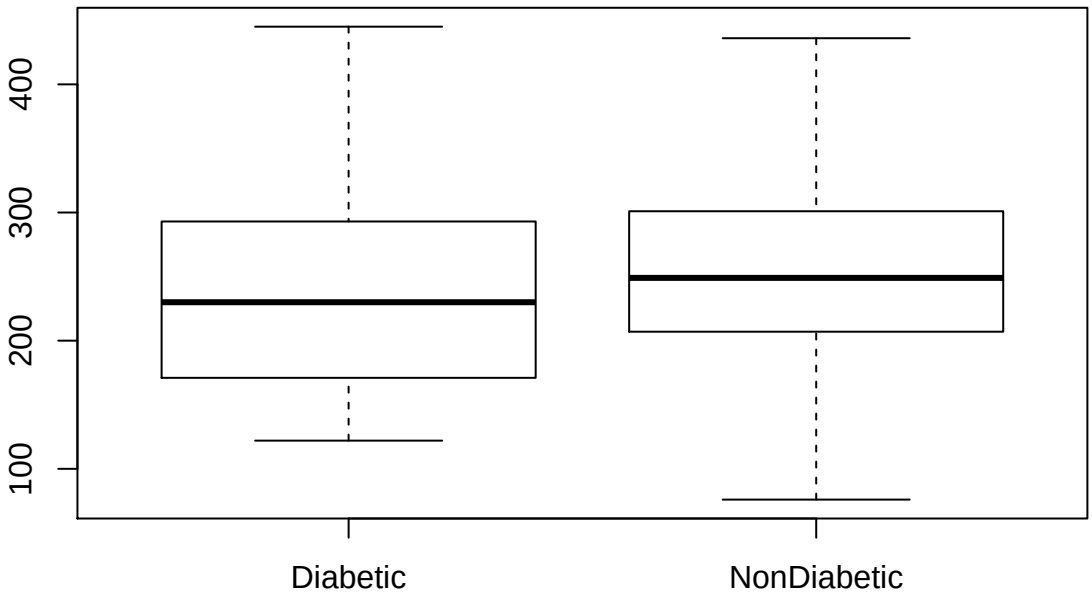
```r
patientLabs %>%
  filter(HL7Text == "Triglyceride") %>%
  ggplot(aes(x=diabetesStatus,y=ObservationValue, fill=AbnormalFlags)) +
  stat_boxplot(geom="boxplot", position="dodge", coef=2, outlier.shape=NA, na.rm=TRUE, show.legend = TRU
  labs(title = "Triglyceride Lab Results", x = "Diabetes Status", fill = "Result Status", y = "Triglyce
  theme(plot.title = element_text(hjust = 0.5))
```



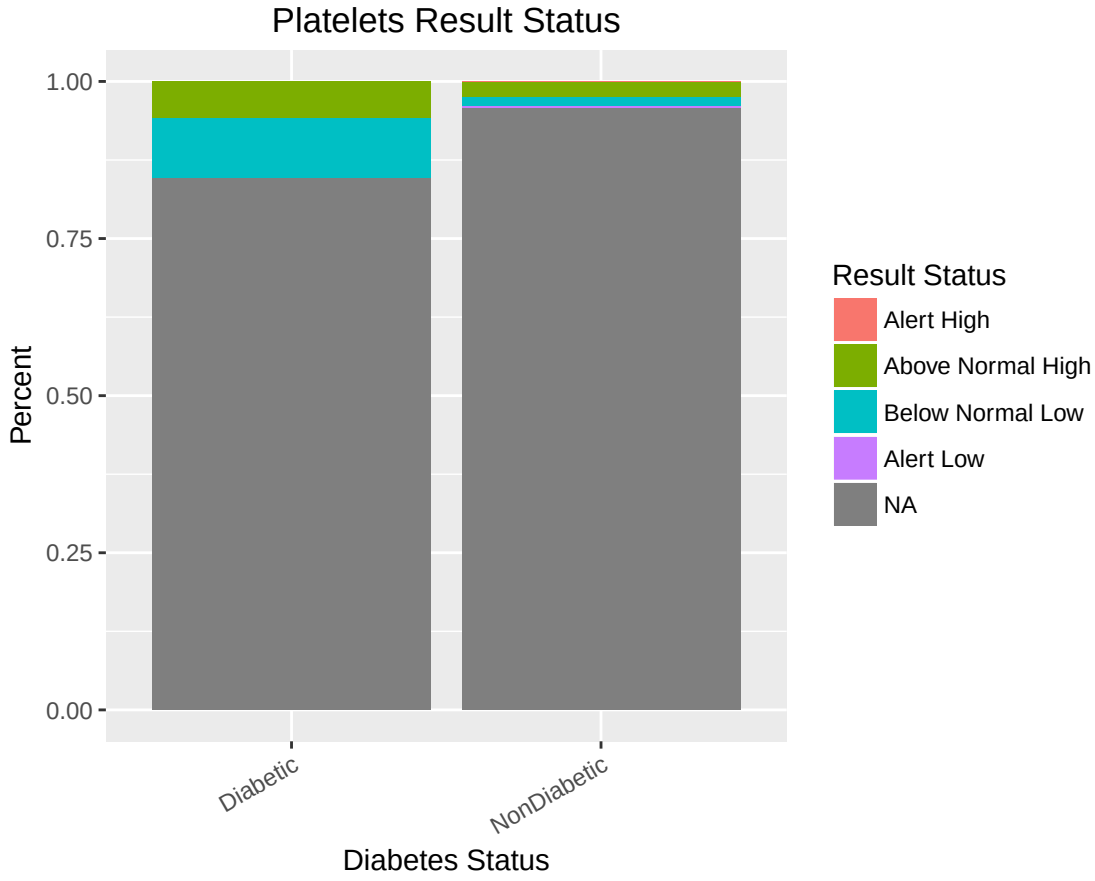
Platelets Lab Results

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.

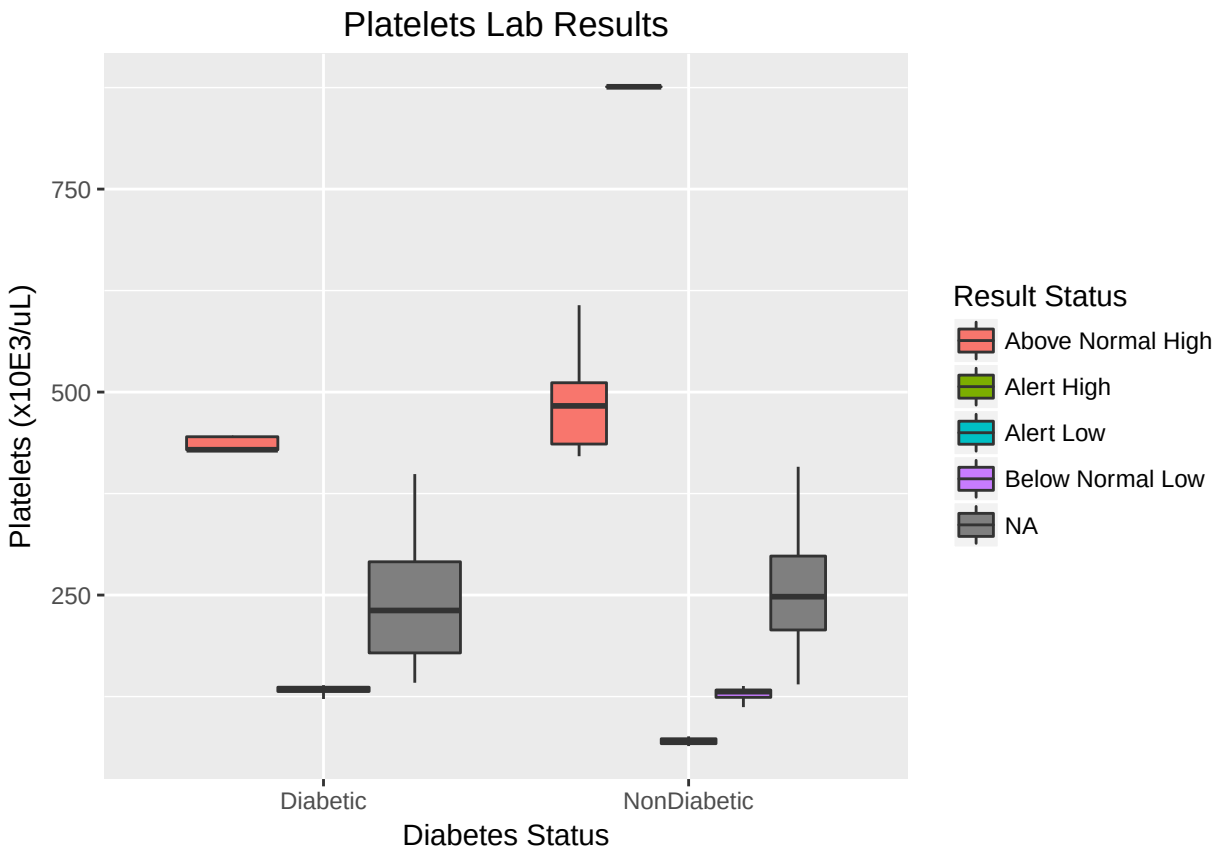
```r
boxplot(ObservationValue~diabetesStatus, data=patientLabs %>% filter(HL7Text == "Platelets"), outline=F
```



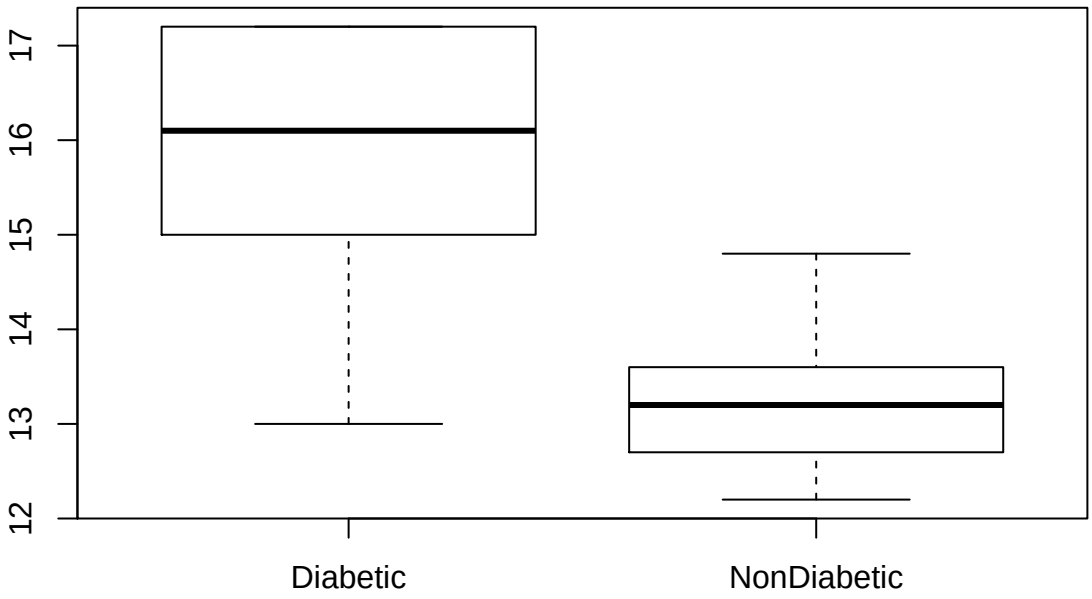

```
patientLabs %>%
  filter(HL7Text == "Platelets") %>%
  ggplot(aes(x=diabetesStatus, fill=AbnormalFlagsSorted))+
  geom_bar(position="fill") +
  labs(title = "Platelets Result Status", x = "Diabetes Status", fill = "Result Status", y = "Percent")
  theme(plot.margin = margin(0,0,0,2,"cm"), axis.text.x = element_text(angle = 30, hjust = 1), plot.titl
```



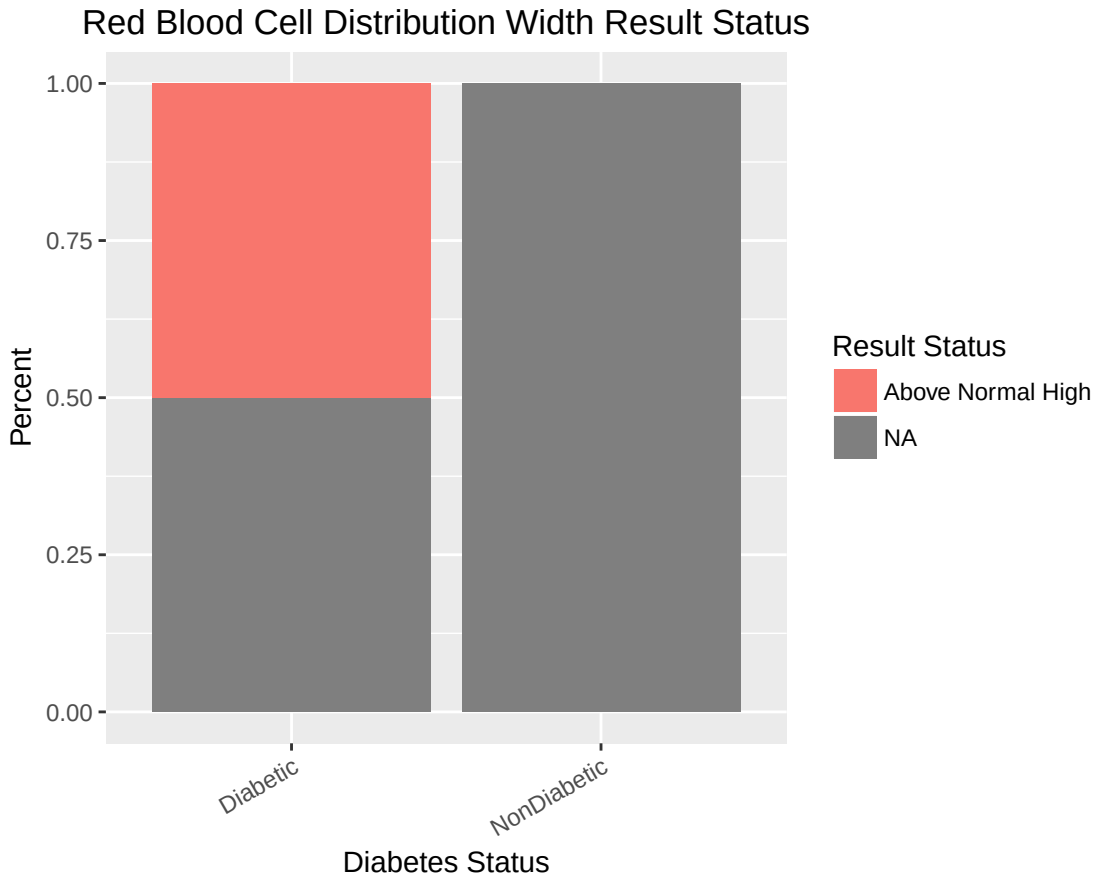
```r
patientLabs %>%
  filter(HL7Text == "Platelets") %>%
  ggplot(aes(x=diabetesStatus,y=ObservationValue, fill=AbnormalFlags)) +
  stat_boxplot(geom="boxplot", position="dodge", coef=2, outlier.shape=NA, na.rm=TRUE, show.legend = TRU
  labs(title = "Platelets Lab Results", x = "Diabetes Status", fill = "Result Status", y = "Platelets (
  theme(plot.title = element_text(hjust = 0.5))
```



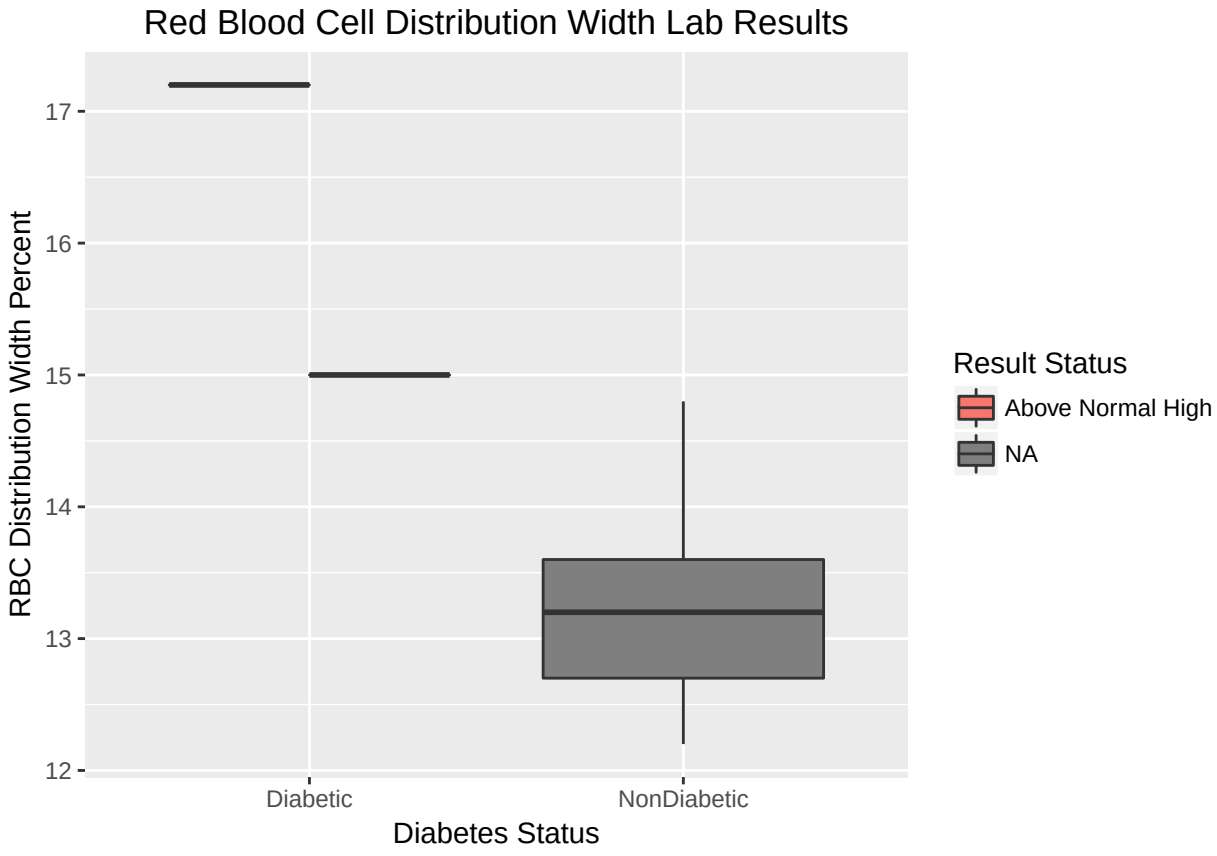

```r
boxplot(ObservationValue~diabetesStatus, data=patientLabs %>% filter(HL7Text == "RBC DISTRIBUTION WIDTH
```




```
patientLabs %>%
  filter(HL7Text == "RBC DISTRIBUTION WIDTH") %>%
  ggplot(aes(x=diabetesStatus, fill=AbnormalFlagsSorted))+
  geom_bar(position="fill") +
  labs(title = "Red Blood Cell Distribution Width Result Status", x = "Diabetes Status", fill = "Result
  theme(plot.margin = margin(0,0,0,2,"cm"), axis.text.x = element_text(angle = 30, hjust = 1), plot.titl
```



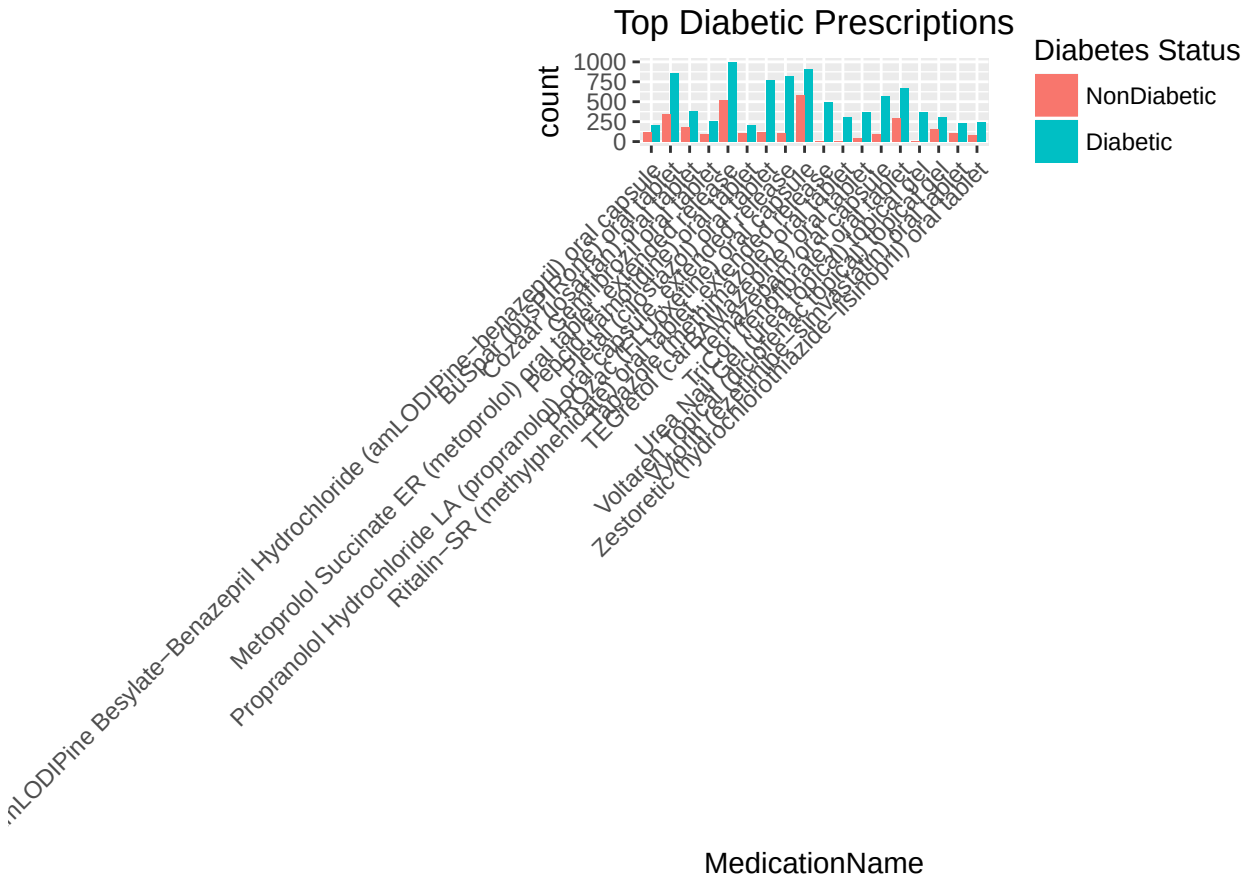
```
patientLabs %>%
  filter(HL7Text == "RBC DISTRIBUTION WIDTH") %>%
  ggplot(aes(x=diabetesStatus,y=ObservationValue, fill=AbnormalFlags)) +
  stat_boxplot(geom="boxplot", position="dodge", coef=2, outlier.shape=NA, na.rm=TRUE, show.legend = TRU
  labs(title = "Red Blood Cell Distribution Width Lab Results", x = "Diabetes Status", fill = "Result St
  theme(plot.title = element_text(hjust = 0.5))
```



Prescription Data

There are 18 medications with at least 300 prescriptions and at least 60% use by diabetic patients.

```
topDiabeticPrescriptions %>%
  arrange(desc(useRatio)) %>%
  ggplot(aes(x=MedicationName, fill=factor(diabetesStatus, levels = c("NonDiabetic","Diabetic"))))+
  geom_bar(position="dodge") +
  labs(title = "Top Diabetic Prescriptions", fill = "Diabetes Status") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.margin = margin(0,0,0,7,"cm"),
        plot.title = element_text(hjust = 0.5))
```


Top Diabetic Prescriptions
count
1000
750
500
250
0
Diabetes Status
NonDiabetic
Diabetic
nLODIPine Besylate−Benazepril Hydrochloride (amLODIPine−benazepril) oral capsule
Metoprolol Succinate ER (metoprolol) oral tablet
Propranolol Hydrochloride LA (propranolol) oral capsule
Ritalin−SR (methylphenidate) oral tablet
MedicationName