Capstone project milestone report

Ashwini Bhatte

02/08/2017

Introduction

Insulin dependent diabetes mellitus (IDDM) also known as **Type 1 diabetes** is a chronic illness characterised by the body's inability to produce insulin due to the autoimmune destruction of the beta cells in the pancreas.

Patients with type 1 diabetes mellitus (DM) require lifelong insulin therapy. Most require 2 or more injections of insulin daily, with doses adjusted on the basis of self-monitoring of blood glucose levels. Long-term management requires a multidisciplinary approach that includes physicians, nurses, dieticians, and selected specialists.

Outpatient management of IDDM relies principally on three interventions: diet, exercise and exogenous insulin. Proper treatment requires careful consideration of all three interventions.

Patients with diabetes face a lifelong challenge to achieve and maintain blood glucose levels as close to the normal range as possible.

Hyperglycemia: BG $>$ 200 mg/dl

Average BG: 150 mg/dl

Hypoglycemia: BG $<$ 80 mg/dl

But it is desirable to keep 90% of all BG measurements $<$ 200 mg/dl.

With appropriate glycemic control, the risk of both microvascular and neuropathic complications is decreased markedly. In addition to being a lifelong disease, the treatment of DM is multifaceted making it necessary to take many variables into account.

The goal of this project is to extract understandable patterns and associations from data through data visualisation.

Dataset

The dataset consists of 70 .tsv files and a data code and a domain description file. Each of which represent the data for a single DM patient. Each file contains 4 attributes: Date, Time, Code and Value.

Date

The dataset covers several weeks' to months' worth of outpatient care on 70 patients.

The date is in MM-DD-YYYY format.

Time

Diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records. The automatic device had an internal clock to timestamp events, whereas the paper records only provided "logical time" slots (breakfast, lunch, dinner, bedtime). For paper records, fixed times were assigned to breakfast (08:00), lunch (12:00), dinner (18:00), and bedtime (22:00).

The time is in HH:MM format.

Code

The Code field is deciphered as follows:

Code	explanation
33	Regular insulin dose
34	NPH insulin dose
35	UltraLente insulin dose
48	Unspecified blood glucose measurement
57	Unspecified blood glucose measurement
58	Pre-breakfast blood glucose measurement
59	Post-breakfast blood glucose measurement
60	Pre-lunch blood glucose measurement
61	Post-lunch blood glucose measurement
62	Pre-supper blood glucose measurement
63	Post-supper blood glucose measurement
64	Pre-snack blood glucose measurement
65	Hypoglycemic symptoms
66	Typical meal ingestion
67	More-than-usual meal ingestion
68	Less-than-usual meal ingestion
69	Typical exercise activity
70	More-than-usual exercise activity
71	Less-than-usual exercise activity
72	Unspecified special event

Value

This attribute represents a blood glucose (BG) concentration at a particular time after a particular activity (see Code).

BG concentration vary even in individuals with normal pancreatic hormonal function.

Normal person BG concentration

pre-meal: 80-120 mg/dl

post-meal: 80-140 mg/dl

DM patient BG concentration

The target range for an individual with diabetes mellitus is very controversial.

So to cut the Gordian knot,

- it would be very desirable to keep 90% of all BG measurements < 200 mg/dl
- average BG: 150 mg/dl or less

Hyperglycemia (BG > 200 mg/dl, over several years) is associated with a poor long-term outcome with the risk of vascular complications.

And **Hypoglycemia** (BG < 80) may have symptoms such as headache, abdominal pain, sweating. Additionally BG < 40 mg/dl may result in neuroglycopenic symptoms such as lethargy, weakness, disorientation, seizures and passing out.

patient_num

This column has been added, to represents the patient number (1-70)

This is how the data frame looks like

```
##
## 1 function (x, df1, df2, ncp, log = FALSE)
## 2 {
## 3     if (missing(ncp))
## 4         .Call(C_df, x, df1, df2, log)
## 5     else .Call(C_dnf, x, df1, df2, ncp, log)
## 6 }
```

Data limitations

The main limitation of this dataset is the occurrence of unspecified blood glucose measurement twice in the code feature. These two codes are treated separately as this is an important information, which can not be rejected.

Also, it would have been useful to include demographics of a patient such as Age and Sex, to predict how blood glucose concentration varies.

Data wrangling and cleaning

The diabetes data set from UCI machine learning repository is relatively comprehensive and well put together and did not require any major transformations.

However, since there are 70 different .tsv files for individual patients, they need to be merged in a single data frame.

Here are the data wrangling steps for diabetes data set

Packages required

```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(caTools)
library(ROCR)
library(nnet)
```

Load the data sets

```
# create an empty data frame
df <- data.frame()
# for loop for data files 1:70
for(i in 1:70) {
  num = paste("", i, sep = "")
# if i < 10, paste 0 followed by the file name, e.g 01, 02, etc.
  if (i < 10) num = paste("0", i, sep = "")
  fname = paste("~/R Projects/diabetes-dataset/Diabetes-Data/data-", num, sep = "")
# creat vector to read files and add column names
  temp <- read_delim(fname, col_names = c("date", "time", "code", "bg_conc"),
   col_types = cols(
    date = col_date(format = "%m-%d-%Y"),
```

```
# %R = Equivalent to %H:%M.
    time = col_time(format = "%R"),
# original string converted to integer
    code = col_integer(),
    bg_conc = col_character()
  ), "\t",
    escape_double = FALSE, trim_ws = TRUE);
# add patient_num column to represent each patients records (1:70)
  temp <- mutate(temp, patient_num = i)
# bind data files (1:70) & create df file
  df <- rbind(df, temp);
}
```

- Create clean_df to clean the data
- Remove "000" & "3A" from bg_conc as they are illogical values
- Clean_df contains code "4", "36, &"56", but since there is no entry of such codes in the data description, these values are removed as well.
- 29 NA's in date and 8 NA's in bg_conc are removed

```
clean_df <- subset(df, bg_conc != "000" & bg_conc != "3A")
clean_df <- subset(clean_df, code != "4" & code != "36" & code != "56")
clean_df$bg_conc <- as.numeric(clean_df$bg_conc)
clean_df <- na.omit(clean_df)
```

- Match the pattern with regex from time column and create 2 hours time bins

```
clean_df <- clean_df %>%
  mutate(time_bin = gsub("^00:[0-9][0-9]:00|^01:[0-9][0-9]:00", "00-02", x = time))%>%
  mutate(time_bin = gsub("^02:[0-9][0-9]:00|^03:[0-9][0-9]:00", "02-04", x = time_bin))%>%
  mutate(time_bin = gsub("^04:[0-9][0-9]:00|^05:[0-9][0-9]:00", "04-06", x = time_bin))%>%
  mutate(time_bin = gsub("^06:[0-9][0-9]:00|^07:[0-9][0-9]:00", "06-08", x = time_bin))%>%
  mutate(time_bin = gsub("^08:[0-9][0-9]:00|^09:[0-9][0-9]:00", "08-10", x = time_bin))%>%
  mutate(time_bin = gsub("^10:[0-9][0-9]:00|^11:[0-9][0-9]:00", "10-12", x = time_bin))%>%
  mutate(time_bin = gsub("^12:[0-9][0-9]:00|^13:[0-9][0-9]:00", "12-14", x = time_bin))%>%
  mutate(time_bin = gsub("^14:[0-9][0-9]:00|^15:[0-9][0-9]:00", "14-16", x = time_bin))%>%
  mutate(time_bin = gsub("^16:[0-9][0-9]:00|^17:[0-9][0-9]:00", "16-18", x = time_bin))%>%
  mutate(time_bin = gsub("^18:[0-9][0-9]:00|^19:[0-9][0-9]:00", "18-20", x = time_bin))%>%
  mutate(time_bin = gsub("^20:[0-9][0-9]:00|^21:[0-9][0-9]:00", "20-22", x = time_bin))%>%
  mutate(time_bin = gsub("^22:[0-9][0-9]:00|^23:[0-9][0-9]:00", "22-24", x = time_bin))
```

- From BG concentration measurements, we can derive if the patient is Hypoglycemic, Hyperglycemic or has a blood glucose in Normal range.

To simplify and study the effects of time and activity against the BG symptoms; 3 bins are created

```
clean_df <- clean_df %>%
  mutate(symptom = gsub("^([0-9]|[0-7][0-9])$", "Hypoglycemia", x = bg_conc))%>%
  mutate(symptom = gsub("^(8[0-9]|9[0-9]|1[0-9][0-9])$", "Normal", x = symptom))%>%
  mutate(symptom = gsub("^(2[0-9][0-9]|3[0-9][0-9]|4[0-9][0-9]|50[0-1])$",
                        "Hyperglycemia", x = symptom))%>%
  mutate(symptom = gsub("^(1.5|2.5|3.5|4.5|6.5|7.5)$", "Hypoglycemia", x = symptom))
```

- Let's add a new column "activity" which represent a type of activity patient has done just before or after measuring the BG

code	type of activity
33 - 35	insulin dose
58 - 64	meal (before - after)
69 - 71	exercise

```
clean_df <- clean_df %>%
  mutate(activity = gsub("^(3[3-5])$", "insulin", x = code)) %>%
  mutate(activity = gsub("^(5[8-9]|6[0-4]|66|67|68)$", "meal", x = activity)) %>%
  mutate(activity = gsub("^(6[9]|7[0-1])$", "exercise", x = activity)) %>%
  mutate(activity = gsub("^(48|57|72)$", "unspecified", x = activity)) %>%
  mutate(activity = gsub("^(65)$", "hypo-measurement", x = activity))
```

- For predictive analysis add one more column similar to the time_bin.

This represents time groups in a numeric format e.g. 00-02 time_bin = 0 bin_num, etc.

```
clean_df <- clean_df %>%
  mutate(bin_num = gsub("^00:[0-9][0-9]:00|^01:[0-9][0-9]:00", "0", x = time))%>%
  mutate(bin_num = gsub("^02:[0-9][0-9]:00|^03:[0-9][0-9]:00", "2", x = bin_num))%>%
  mutate(bin_num = gsub("^04:[0-9][0-9]:00|^05:[0-9][0-9]:00", "4", x = bin_num))%>%
  mutate(bin_num = gsub("^06:[0-9][0-9]:00|^07:[0-9][0-9]:00", "6", x = bin_num))%>%
  mutate(bin_num = gsub("^08:[0-9][0-9]:00|^09:[0-9][0-9]:00", "8", x = bin_num))%>%
  mutate(bin_num = gsub("^10:[0-9][0-9]:00|^11:[0-9][0-9]:00", "10", x = bin_num))%>%
  mutate(bin_num = gsub("^12:[0-9][0-9]:00|^13:[0-9][0-9]:00", "12", x = bin_num))%>%
  mutate(bin_num = gsub("^14:[0-9][0-9]:00|^15:[0-9][0-9]:00", "14", x = bin_num))%>%
  mutate(bin_num = gsub("^16:[0-9][0-9]:00|^17:[0-9][0-9]:00", "16", x = bin_num))%>%
  mutate(bin_num = gsub("^18:[0-9][0-9]:00|^19:[0-9][0-9]:00", "18", x = bin_num))%>%
  mutate(bin_num = gsub("^20:[0-9][0-9]:00|^21:[0-9][0-9]:00", "20", x = bin_num))%>%
  mutate(bin_num = gsub("^22:[0-9][0-9]:00|^23:[0-9][0-9]:00", "22", x = bin_num))
```

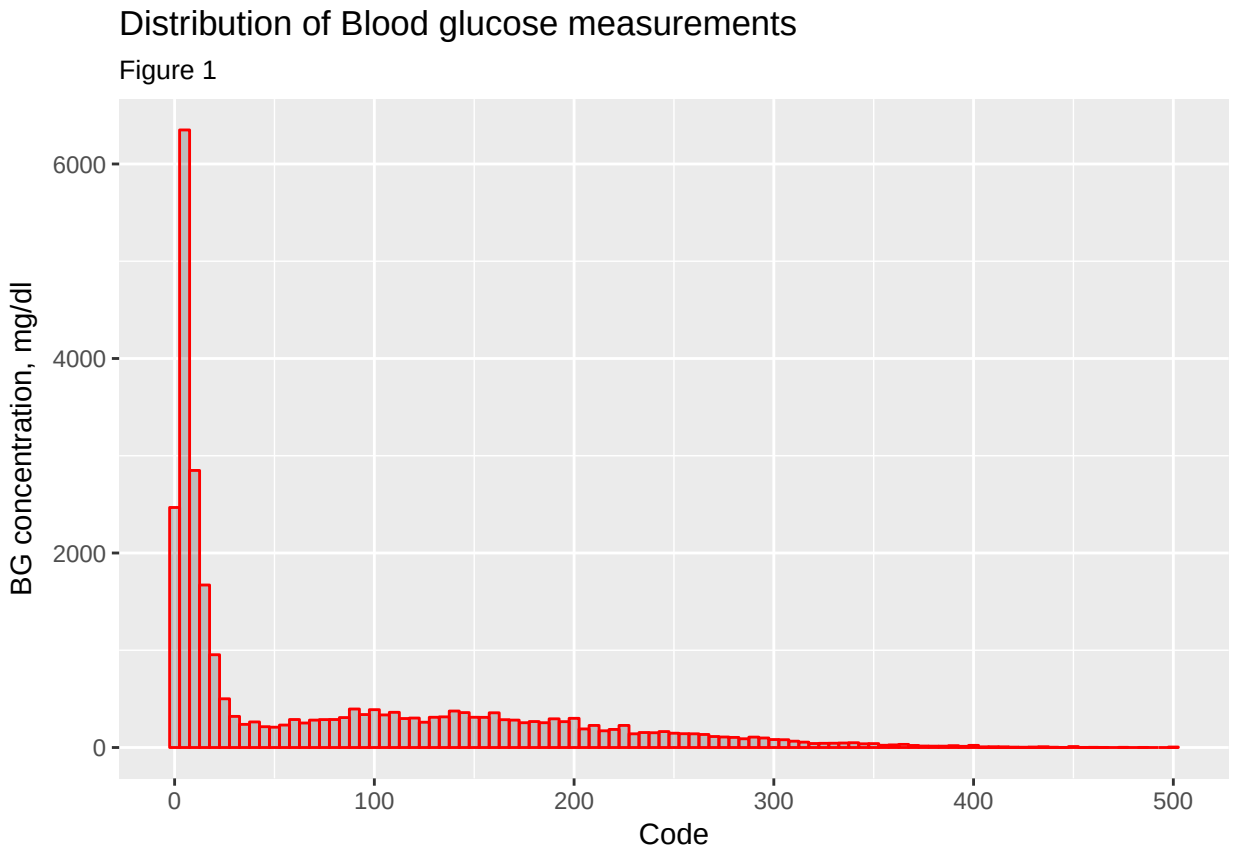
Exploratory Data analysis

The following variables will be investigated in this section:

- Code vs Blood Glucose concentration
- Time vs Blood Glucose concentration

Since bg_conc is the only numeric variable in the data set; it's distribution is observed by plotting the histogram.

```
ggplot(clean_df, aes(bg_conc)) +
  geom_histogram(binwidth = 5, fill = "black", col = "red", alpha = 0.2) +
  labs(x = "Code", y = "BG concentration, mg/dl") +
  ggtitle("Distribution of Blood glucose measurements", subtitle = "Figure 1")
```



We can see the high peak at 10 - 20 mg/dl BG

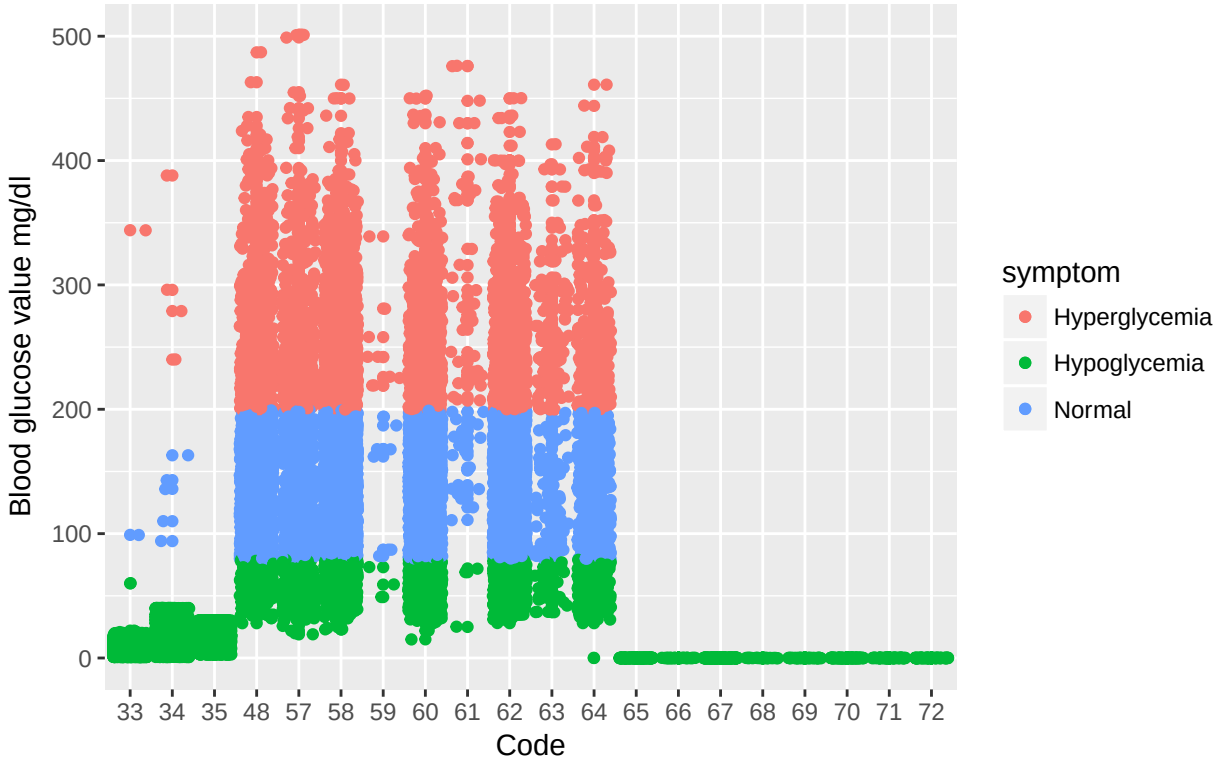
Now that we have determined the distribution of BG concentration, the focus of the further data visualisation and exploration will be to deep dive into other variables and determine how they interact with one another.

Code vs Blood Glucose concentration

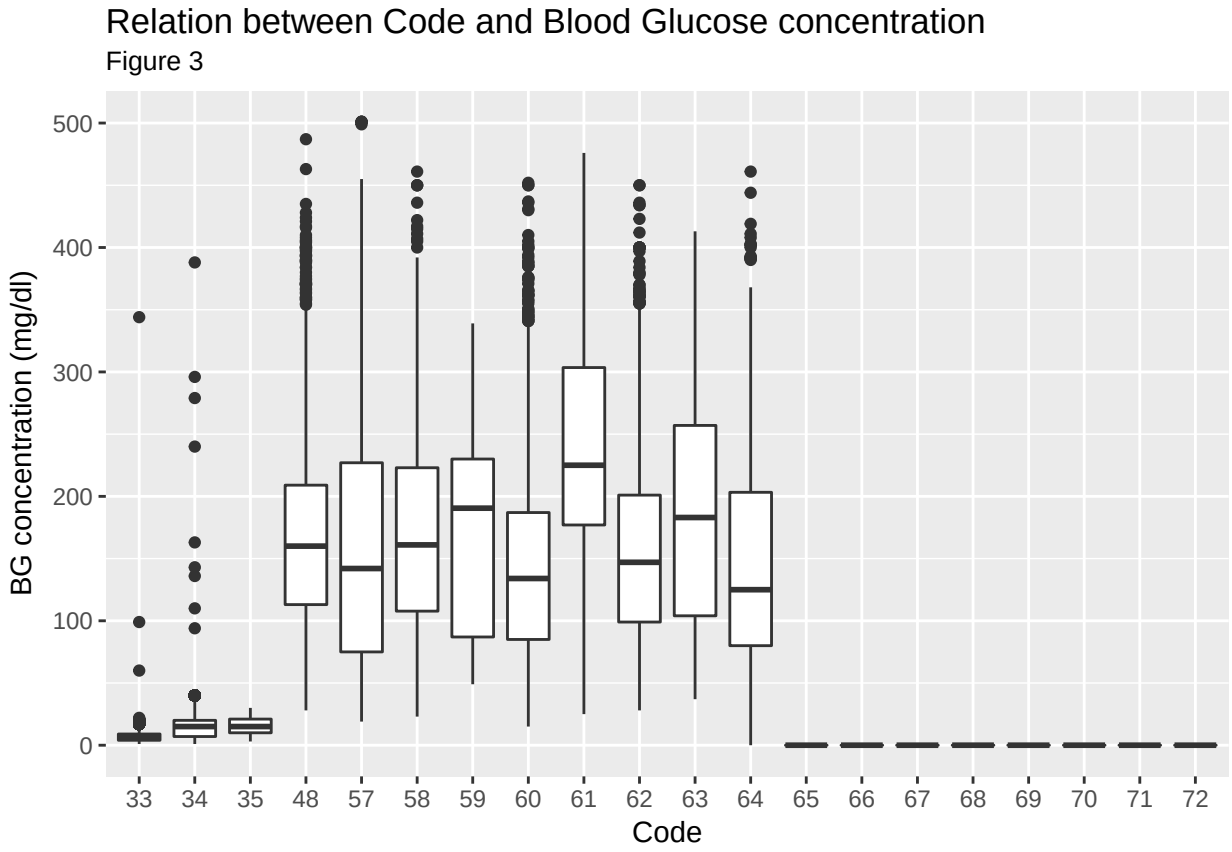
```
ggplot(clean_df, aes(factor(code), bg_conc, col = symptom)) +
  geom_point() +
  geom_jitter() +
  labs(x = "Code", y = "Blood glucose value mg/dl") +
  ggtitle("Distribution of BG measurements based on Code and BG concentration",
          subtitle = "Figure 2")
```

Distribution of BG measurements based on Code and BG concentration

Figure 2




```
ggplot(clean_df, aes(factor(code), bg_conc)) +
  geom_boxplot() +
  labs(x = "Code", y = "BG concentration (mg/dl)") +
  ggtitle("Relation between Code and Blood Glucose concentration", subtitle = "Figure 3")
```

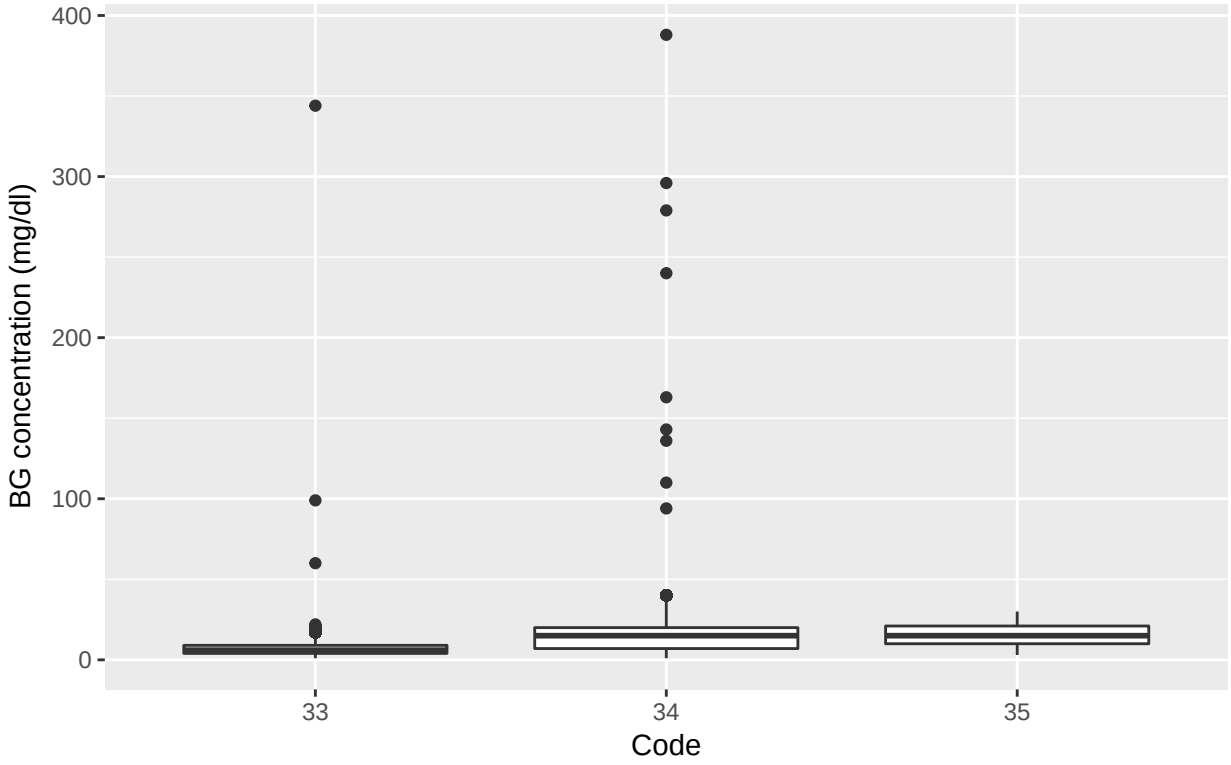


To gain a better perspective at this, let's plot the graph of BG concentration vs code by grouping the codes into 3 sub categories.

```
c1 <- c(33:35)
code_df1 <- clean_df[clean_df$code %in% c1,]
ggplot(code_df1, aes(factor(code), bg_conc)) +
  geom_boxplot() +
  labs(x = "Code", y = "BG concentration (mg/dl)") +
  ggtitle("Relation between Insulin dose and Blood Glucose concentration", subtitle =
            "Figure 4")
```

Relation between Insulin dose and Blood Glucose concentration

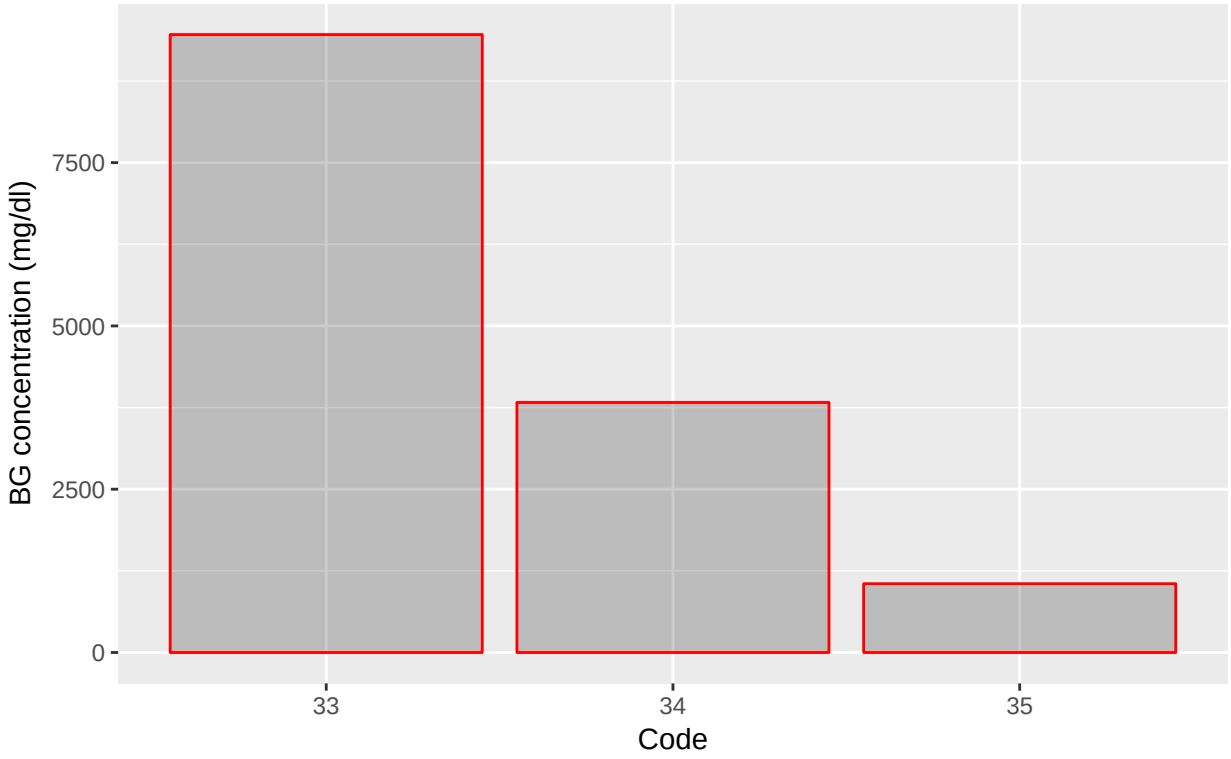
Figure 4




```
ggplot(code_df1, aes(factor(code))) +
  geom_histogram(stat = "count", fill= "black", col= "red", alpha= 0.2) +
  labs(x = "Code", y = "BG concentration (mg/dl)") +
  ggtitle("Distribution of code (Insulin)", subtitle = "Figure 5")
```

Distribution of code (Insulin)

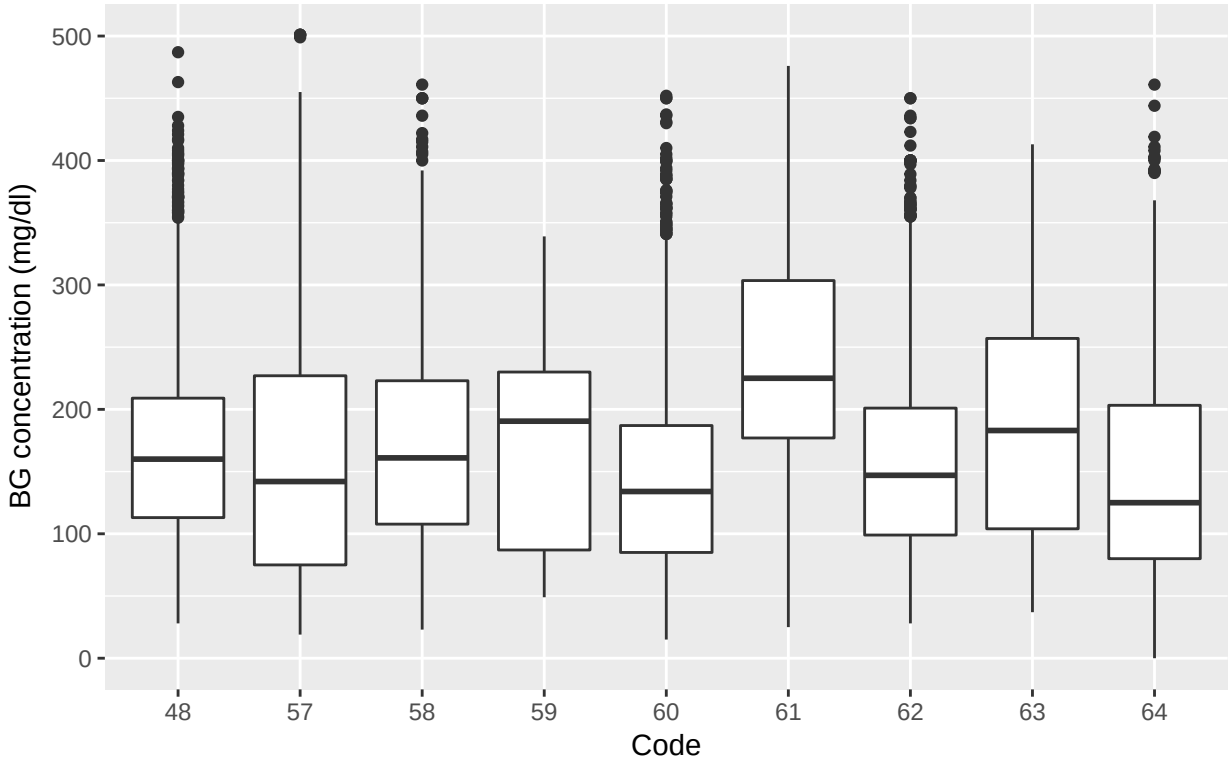
Figure 5



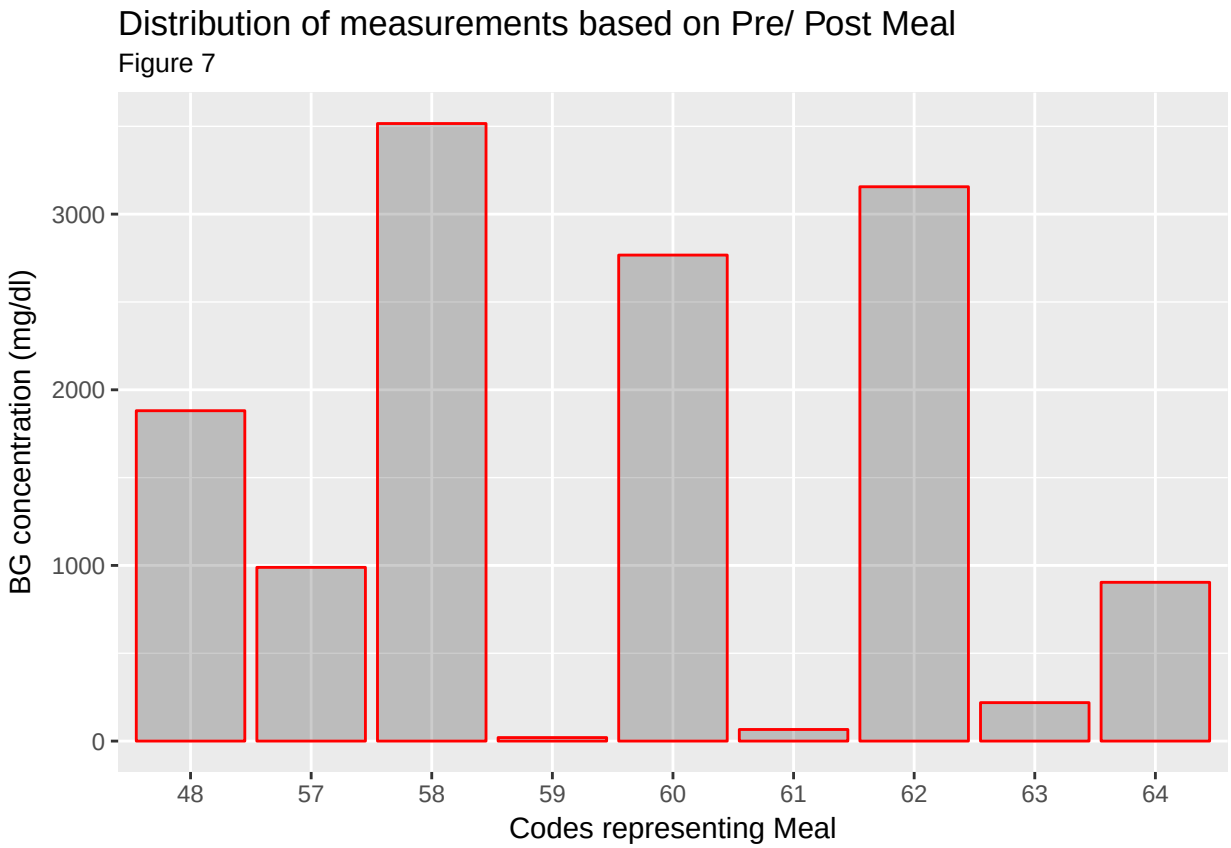

```
c2 <- c(48, 57:64)
code_df2 <- clean_df[clean_df$code %in% c2,]
ggplot(code_df2, aes(factor(code), bg_conc)) +
  geom_boxplot() +
  labs(x = "Code", y = "BG concentration (mg/dl)") +
  ggtitle("Relation between Meal and Blood Glucose concentration", subtitle = "Figure 6")
```

Relation between Meal and Blood Glucose concentration

Figure 6

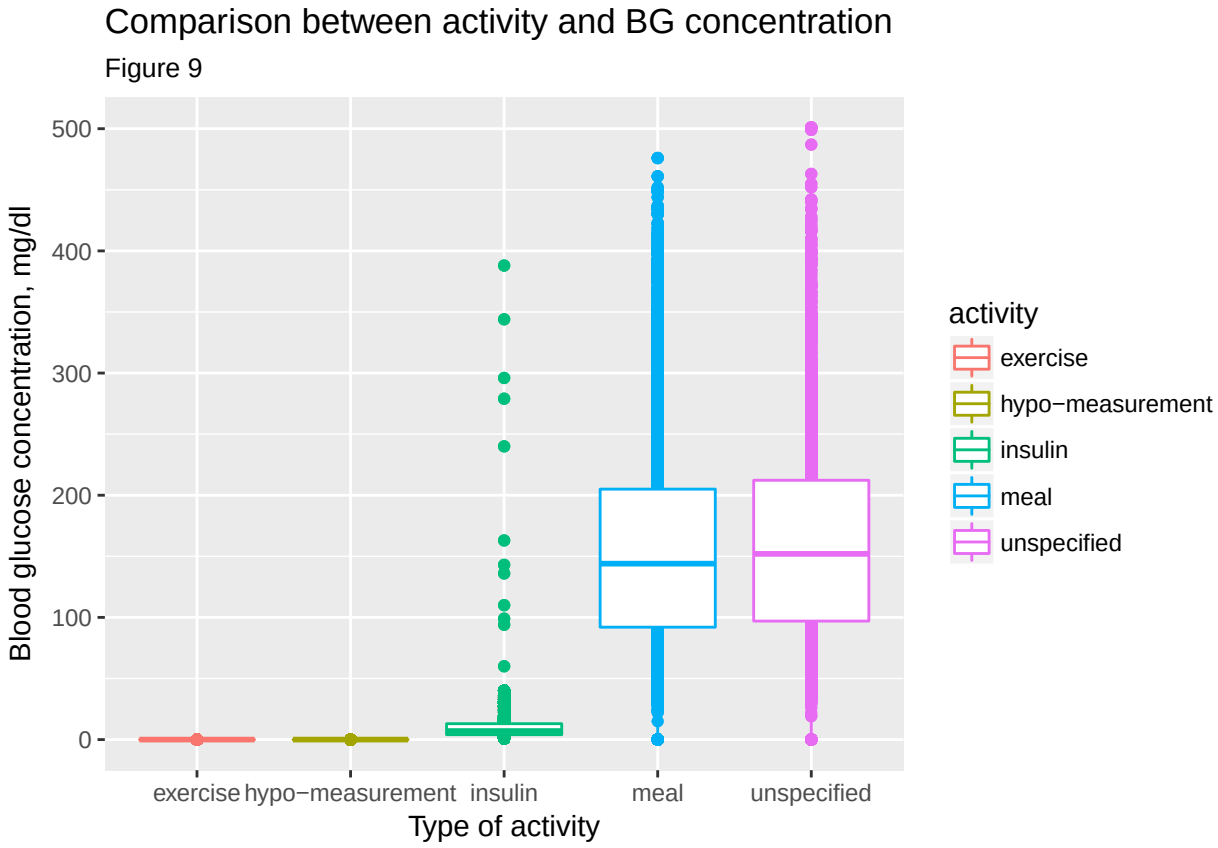



```
ggplot(code_df2, aes(factor(code))) +
  geom_histogram(fill= "black", col= "red", alpha= 0.2, stat = "count") +
  labs(x = "Codes representing Meal", y = "BG concentration (mg/dl)") +
  ggtitle("Distribution of measurements based on Pre/ Post Meal", subtitle = "Figure 7")
```




```
c3 <- c(65:72)
code_df3 <- clean_df[clean_df$code %in% c3,]
ggplot(code_df3, aes(factor(code), bg_conc)) +
  geom_boxplot() +
  labs(x = "Code", y = "BG concentration (mg/dl)") +
  ggtitle("Relation between exercise and Blood Glucose concentration", subtitle = "Figure 8")
```


```
ggplot(clean_df, aes(factor(activity), bg_conc, group=activity, col=activity)) +
  geom_point() +
  geom_boxplot() +
  labs(x = "Type of activity", y = "Blood glucose concentration, mg/dl") +
  ggtitle("Comparison between activity and BG concentration", subtitle = "Figure 9")
```



Conclusion 1: Exploratory Data Analysis

Code vs Blood Glucose concentration

- From Figure 1, we can clearly see a pattern between BG concentration and the type of activity. The patients suffer from Hyperglycemia before or after a meal.
- Also there are three groups which show the similar patterns. To explore more, the three individual plots have been created. Statistical summary of which is as follows
- Figure 4, Relation between Insulin dose and Blood Glucose concentration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   7.000   9.643  13.000 388.000
```

- Figure 6, Relation between Meal and Blood Glucose concentration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    97.0   149.0   160.2   210.0   501.0
```

- Figure 8, Relation between exercise and Blood Glucose concentration

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

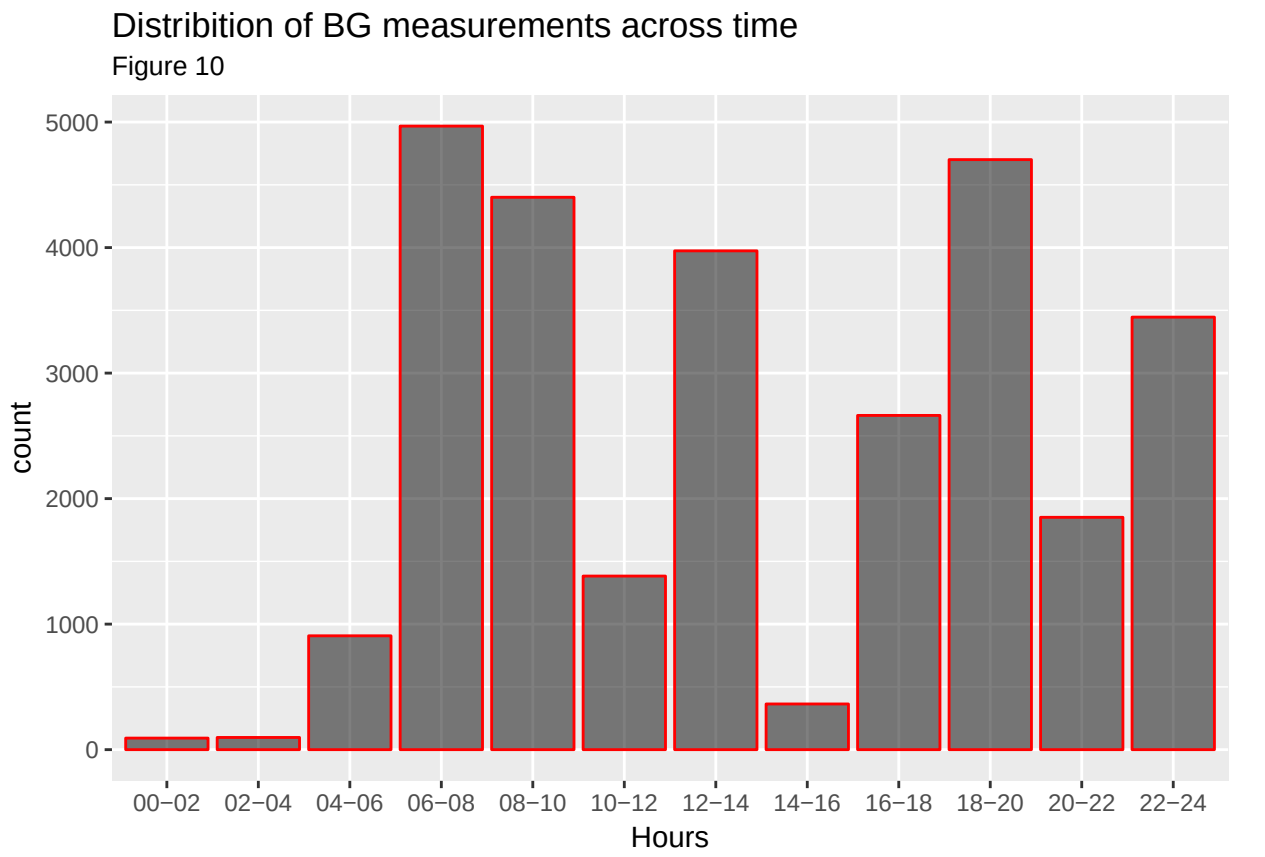
- Note that, patients take Regular insulin dose more often than NPH and Ultralente insulin dose and rarely measure BG after a meal (Figure 5 & 7)
- Figure 9 nicely compares between the types of activities and BG concentration. Blood glucose level decreases just after taking the isulin dose. It even drops down further down after the exercise. However, the BG notably increases after a meal with a median of around 150 mg/dl

Time vs Blood Glucose concentration

Since time is a continuous variable, it will be difficult to explore the data when time is plotted on x- axis. Which is why the time intervals of 2 hours (time_bin) were created

We want to explore the distribution of Hypoglycemia, Normal BG concentration and Hyperglycemia based on time intervals.

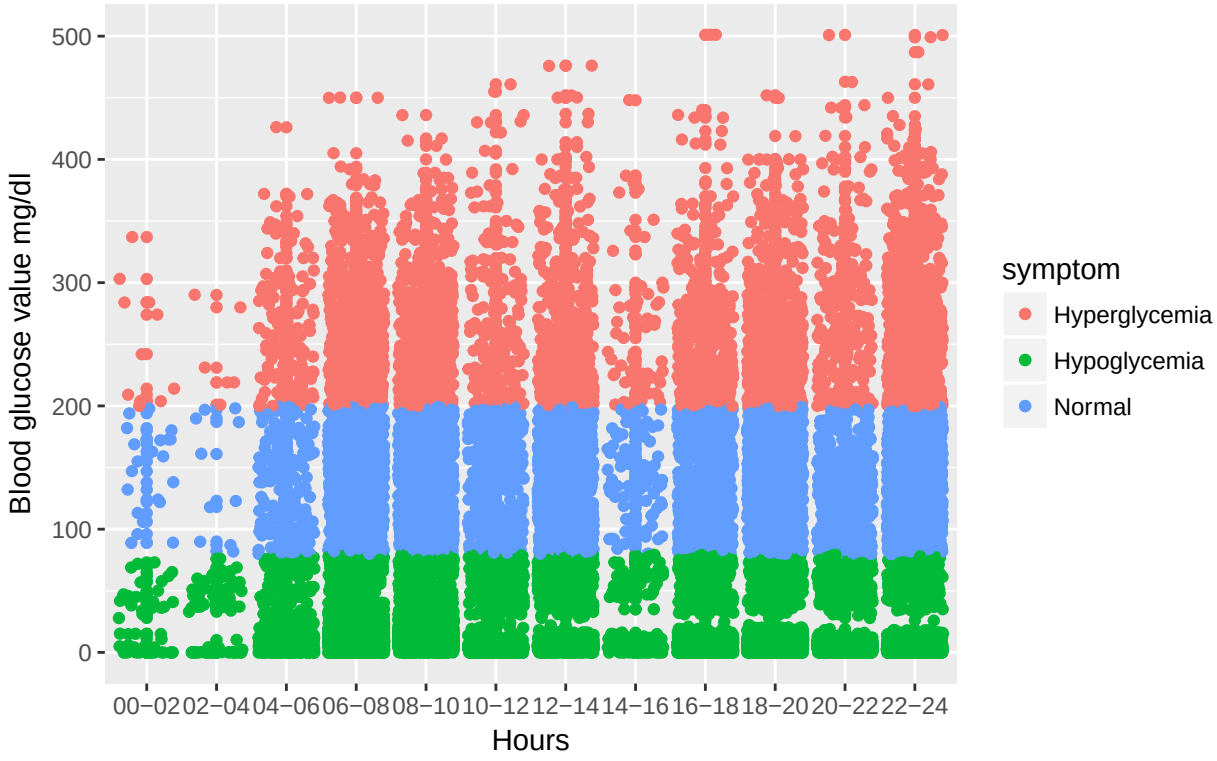
```
ggplot(clean_df, aes(factor(time_bin))) +
 geom_histogram(stat = "count", fill= "black", col= "red", alpha= 0.5) +
 labs(x = "Hours") +
 ggtitle("Distribition of BG measurements across time", subtitle = "Figure 10")
```




```
ggplot(clean_df, aes(factor(time_bin), bg_conc, col = symptom)) +
  geom_point() +
  geom_jitter() +
  labs(x = "Hours", y = "Blood glucose value mg/dl") +
  ggtitle("Distribution of BG symptoms over 24 hours", subtitle = "Figure 11")
```

Distribution of BG symptoms over 24 hours

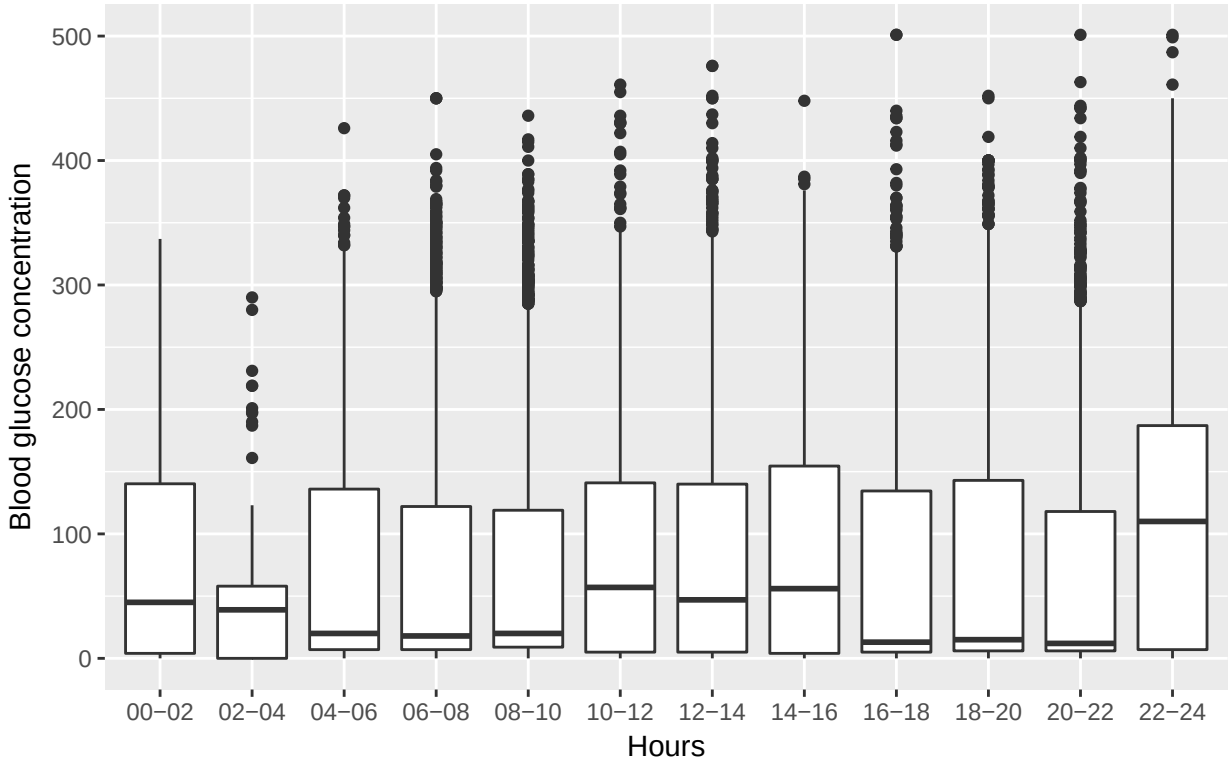
Figure 11




```
ggplot(clean_df, aes(factor(time_bin), bg_conc)) +
  geom_boxplot() +
  labs(x = "Hours", y = "Blood glucose concentration") +
  ggtitle("Association between time and BG concentration", subtitle = "Figure 12")
```

Association between time and BG concentration

Figure 12



Conclusion 2: Exploratory Data Analysis

Time vs Blood Glucose concentration

Each point here represents number of Blood glucose measurements by the patients in 24 hours for several weeks or months.

We do not see any precise time at which the patient was particulary showing Hypoglycemic or Hyperglycemic symptoms, as the symptoms are distributed across 24 hours. However we can say that there are comparatively less measurements showing these symptoms from 00 - 04 in the morning.

This could be due to the fact that there were less number of measurements taken at these time intervals.

Howerver, there is a notable drop in median of BG concentration from 4 - 10 and from 16 - 22.
