# Statistical analysis - Capstone project

*Ashwini Bhatte*

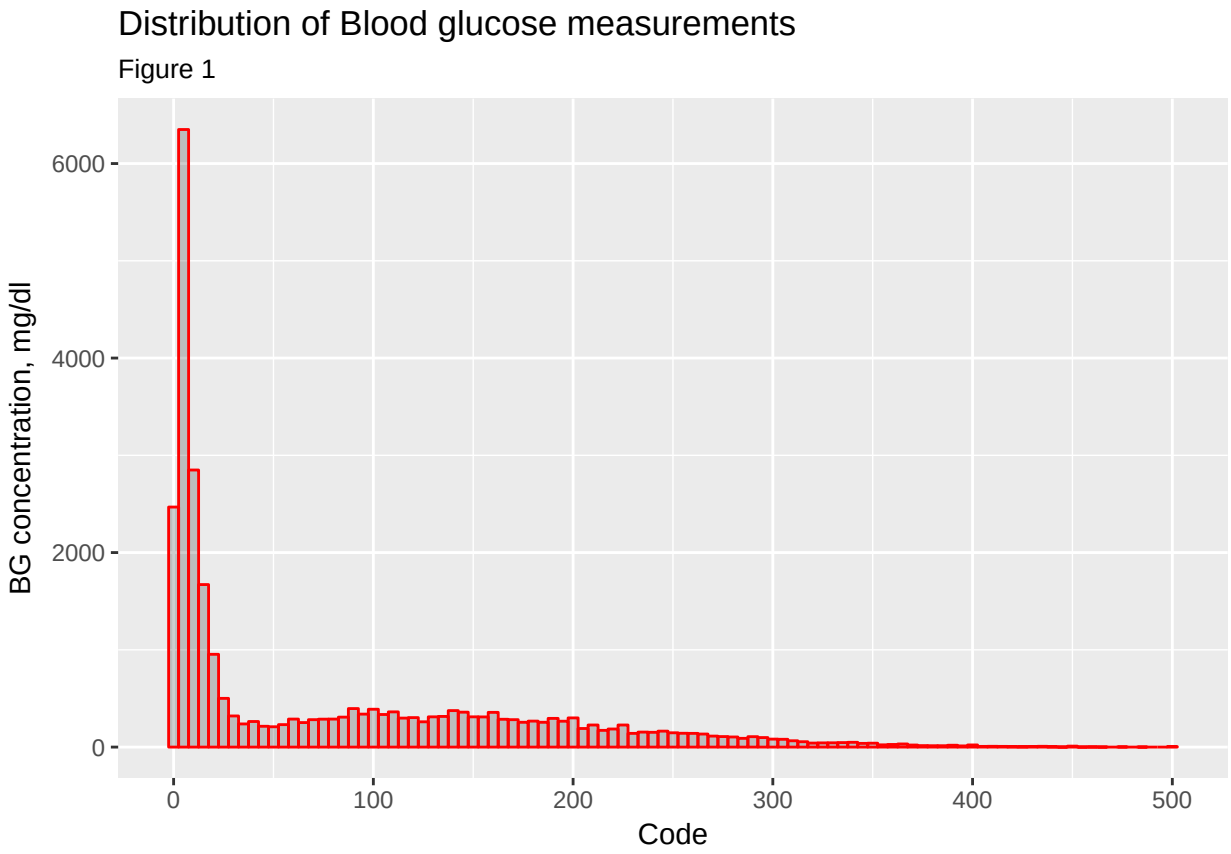
*05/08/2017*

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")
```



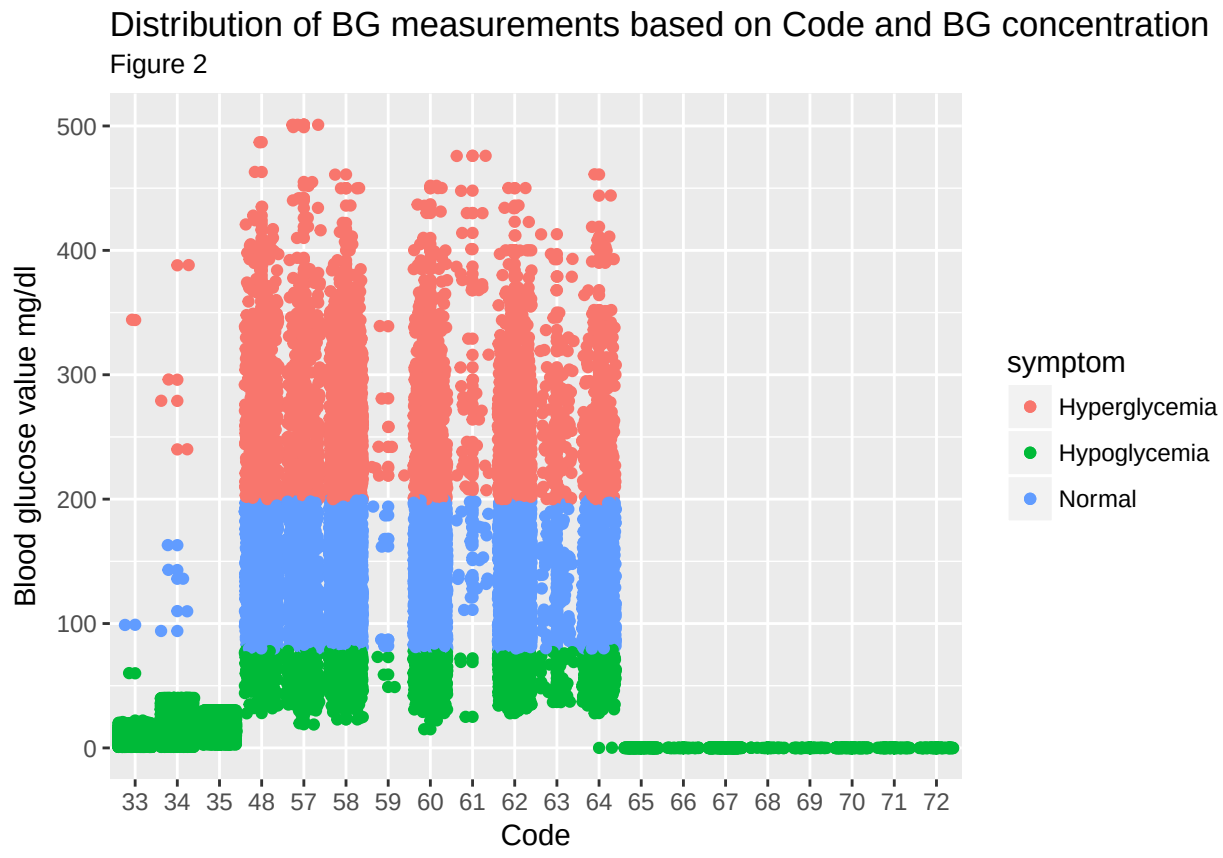

We can see the high peak at 10 - 20 mg/dl BG

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.

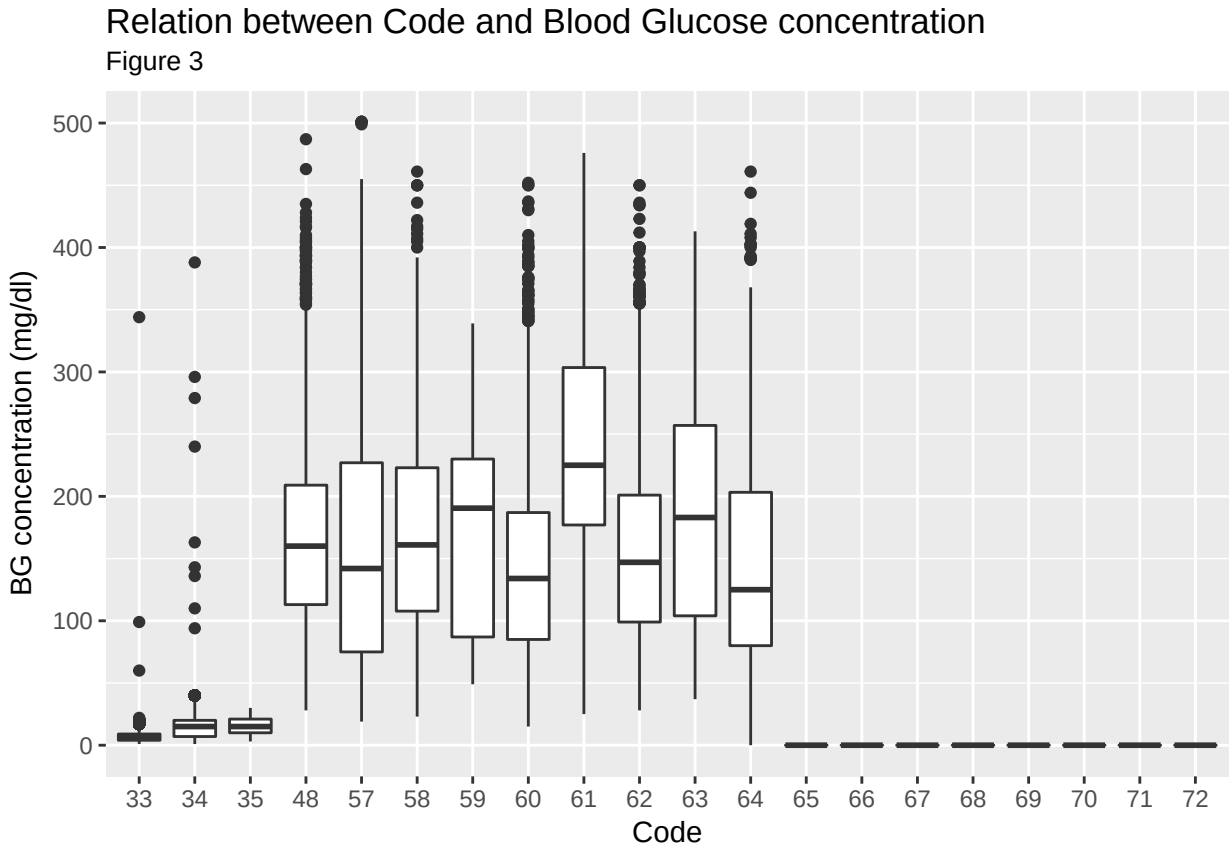
---

**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")
```




```
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")
```

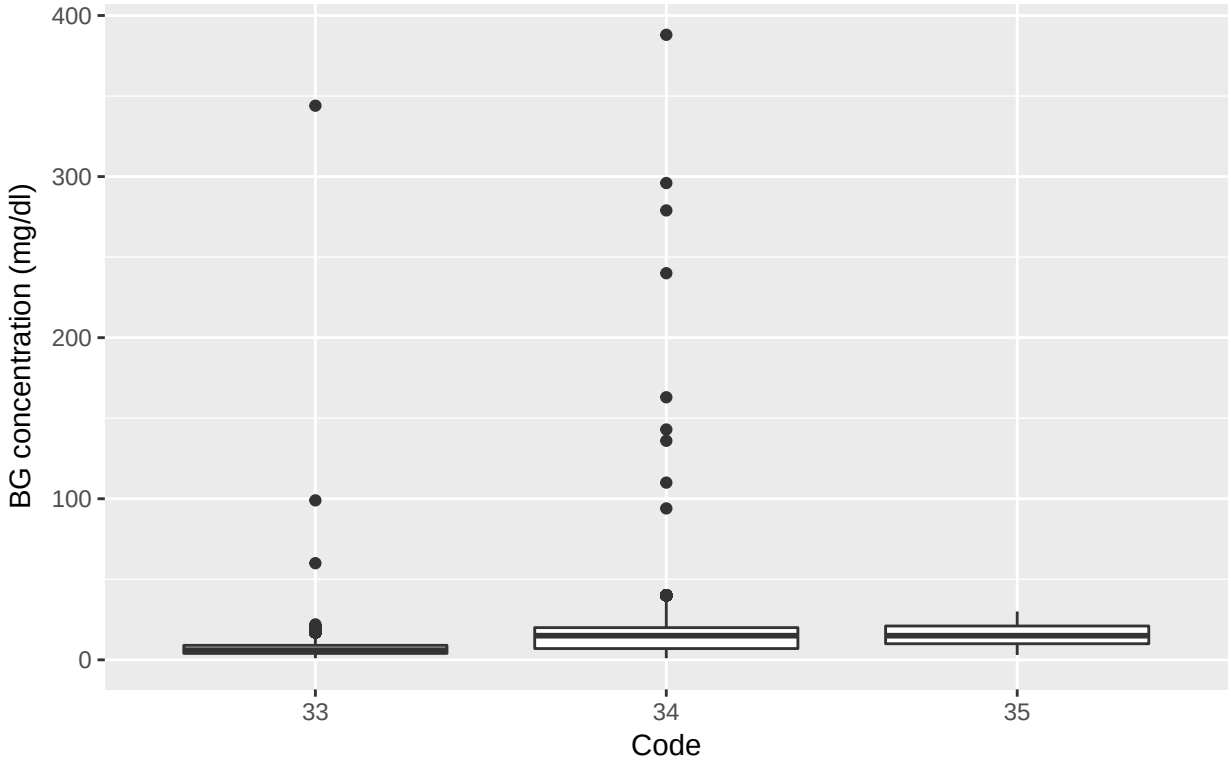



To gain a better perspective at this, let's plot the graph of BG concentration vs code by grouping the codes into 3 sub categories.

```
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

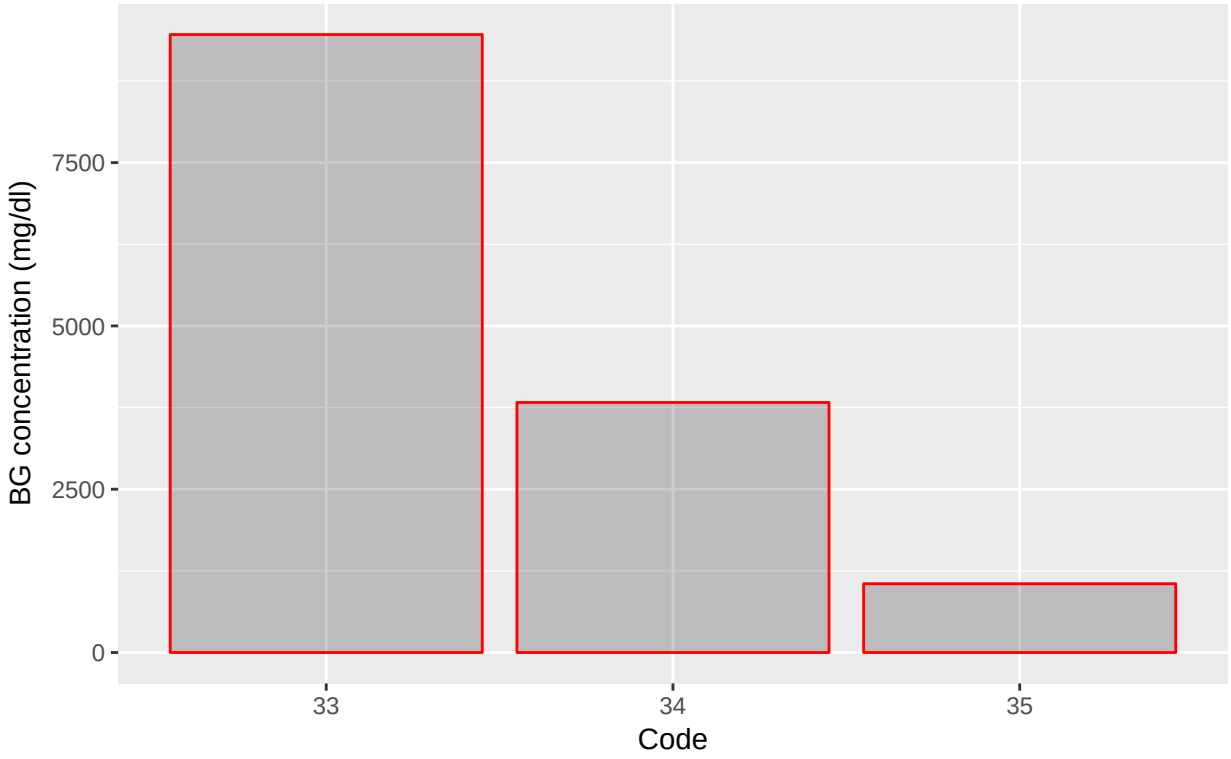
Figure 4




```
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)

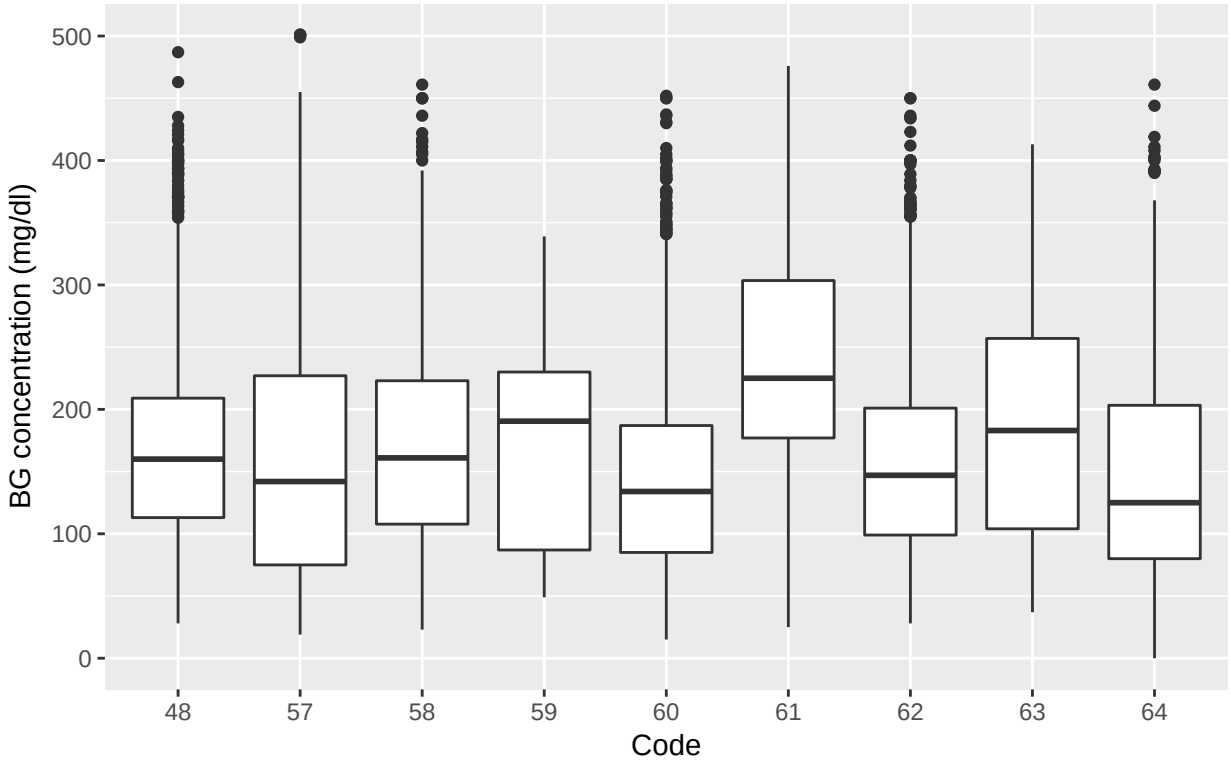
Figure 5



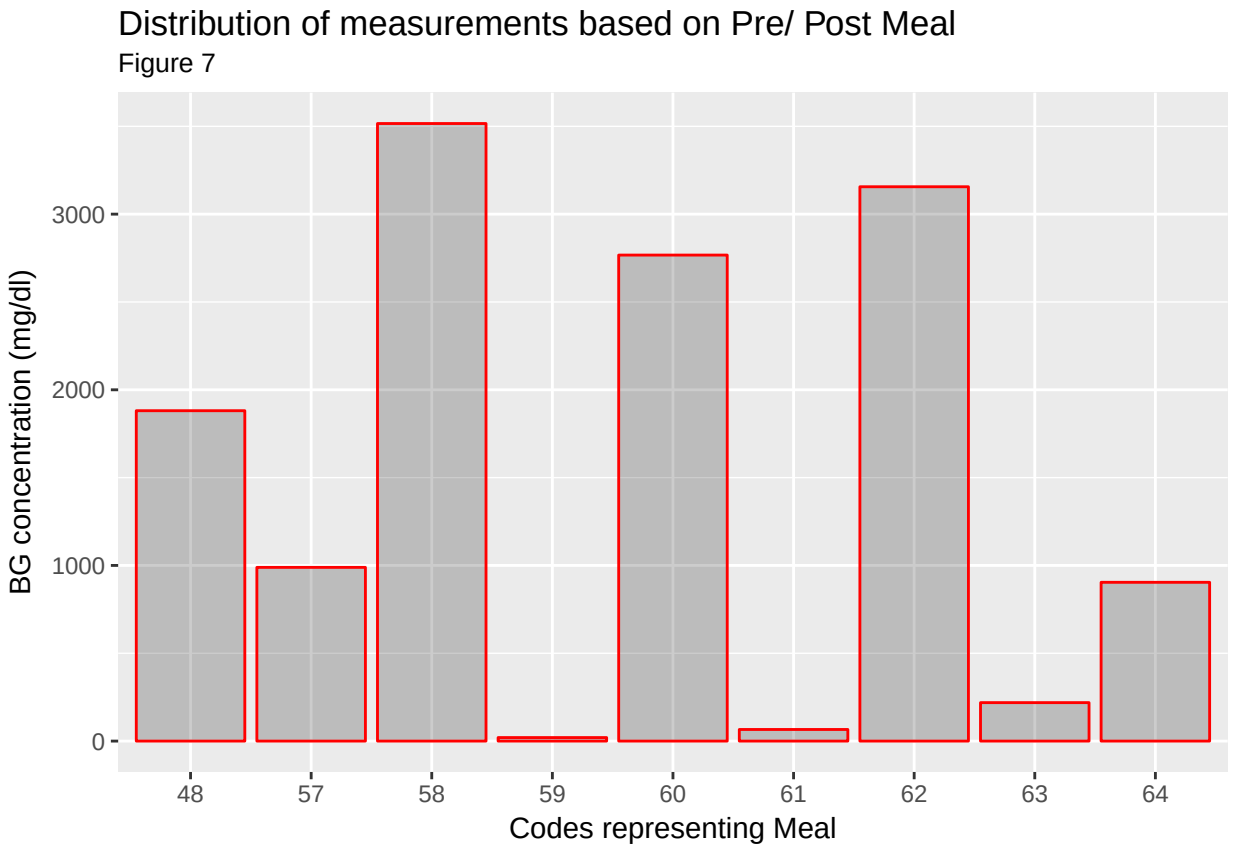

```
c2 <- c(48, 57:64)
code_df2 <- clean_df[clean_df$code %in% c2,]
ggplot(code_df2, aes(factor(code), bg_conc)) +
  geom_boxplot() +
  labs(x = "Code", y = "BG concentration (mg/dl)") +
  ggtitle("Relation between Meal and Blood Glucose concentration", subtitle = "Figure 6")
```

Relation between Meal and Blood Glucose concentration

Figure 6



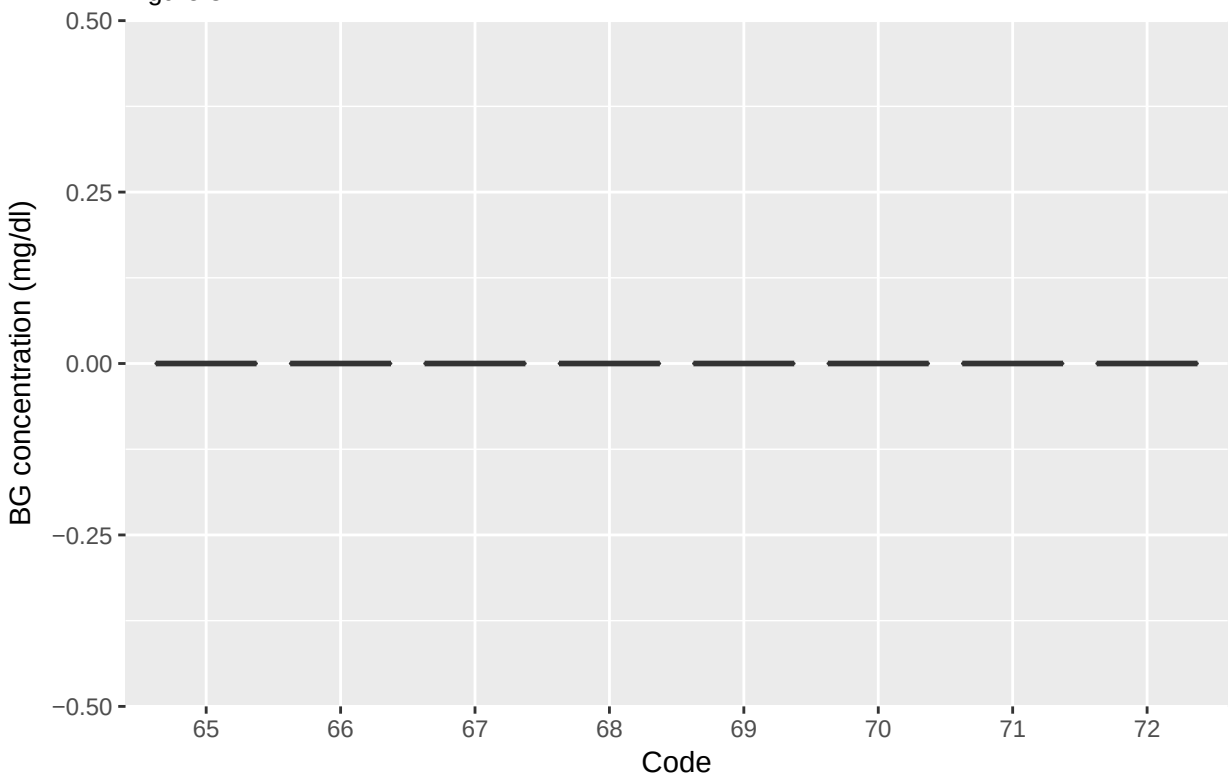

```
ggplot(code_df2, aes(factor(code))) +
  geom_histogram(fill= "black", col= "red", alpha= 0.2, stat = "count") +
  labs(x = "Codes representing Meal", y = "BG concentration (mg/dl)") +
  ggtitle("Distribution of measurements based on Pre/ Post Meal", subtitle = "Figure 7")
```



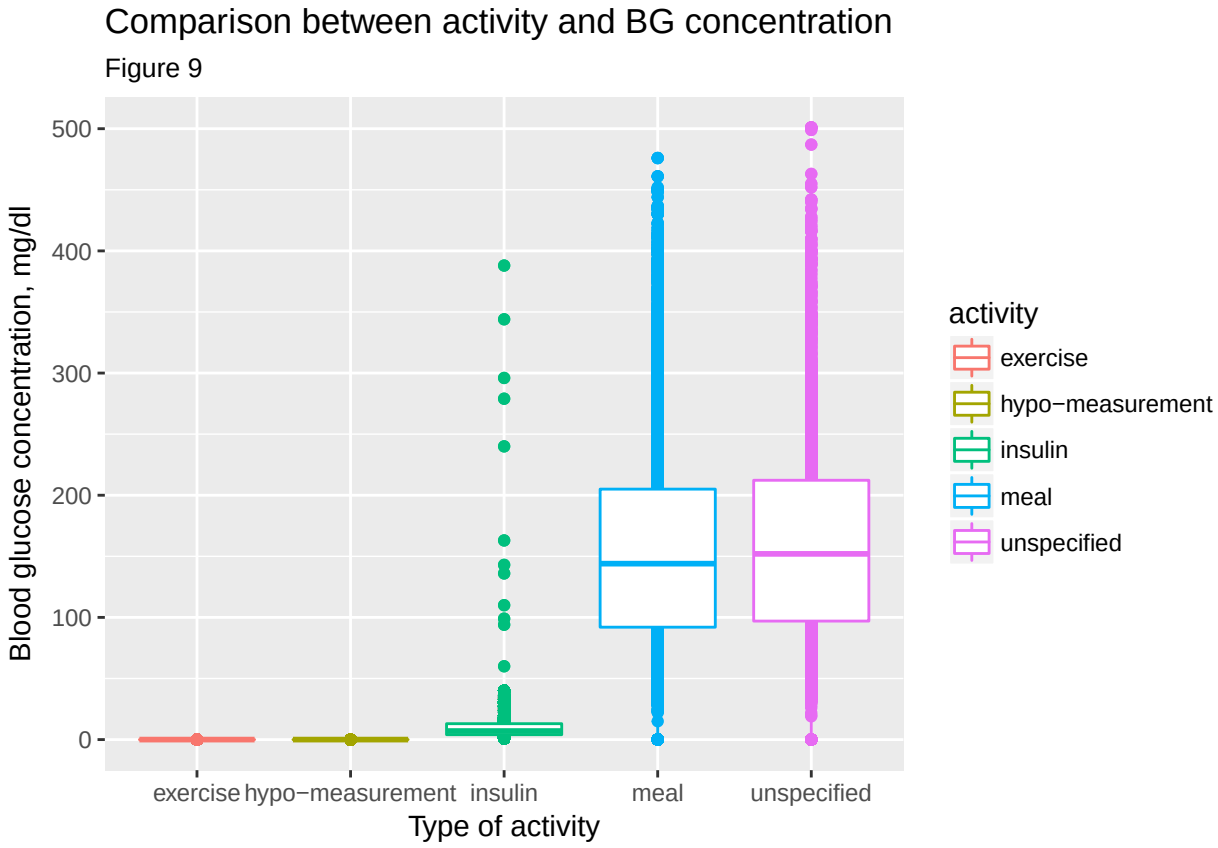

```
c3 <- c(65:72)
code_df3 <- clean_df[clean_df$code %in% c3,]
ggplot(code_df3, aes(factor(code), bg_conc)) +
  geom_boxplot() +
  labs(x = "Code", y = "BG concentration (mg/dl)") +
  ggtitle("Relation between exercise and Blood Glucose concentration", subtitle = "Figure 8")
```

Relation between exercise and Blood Glucose concentration

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

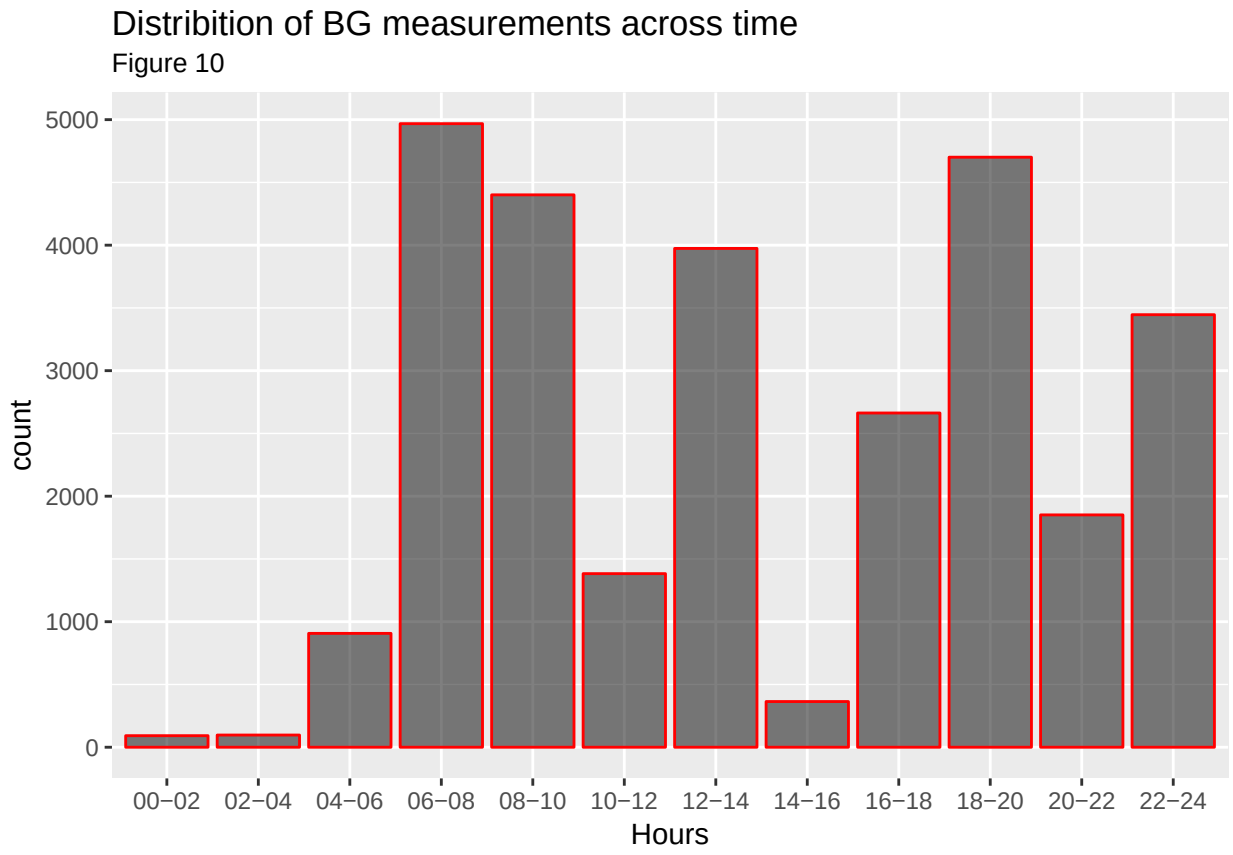
---

**Time vs Blood Glucose concentration**

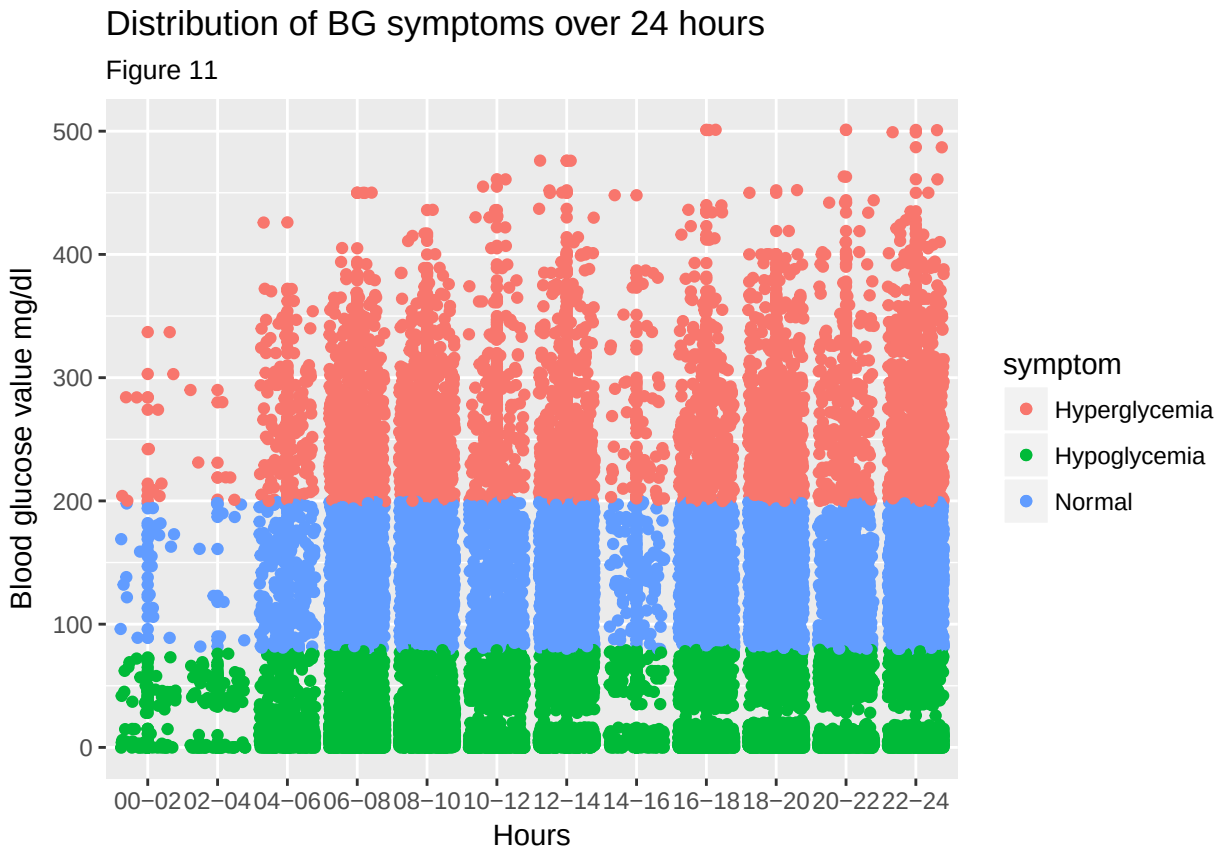
Since time is a continuous variable, it will be difficult to explore the data when time is plotted on x- axis. Which is why the time intervals of 2 hours (time_bin) were created

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")
```



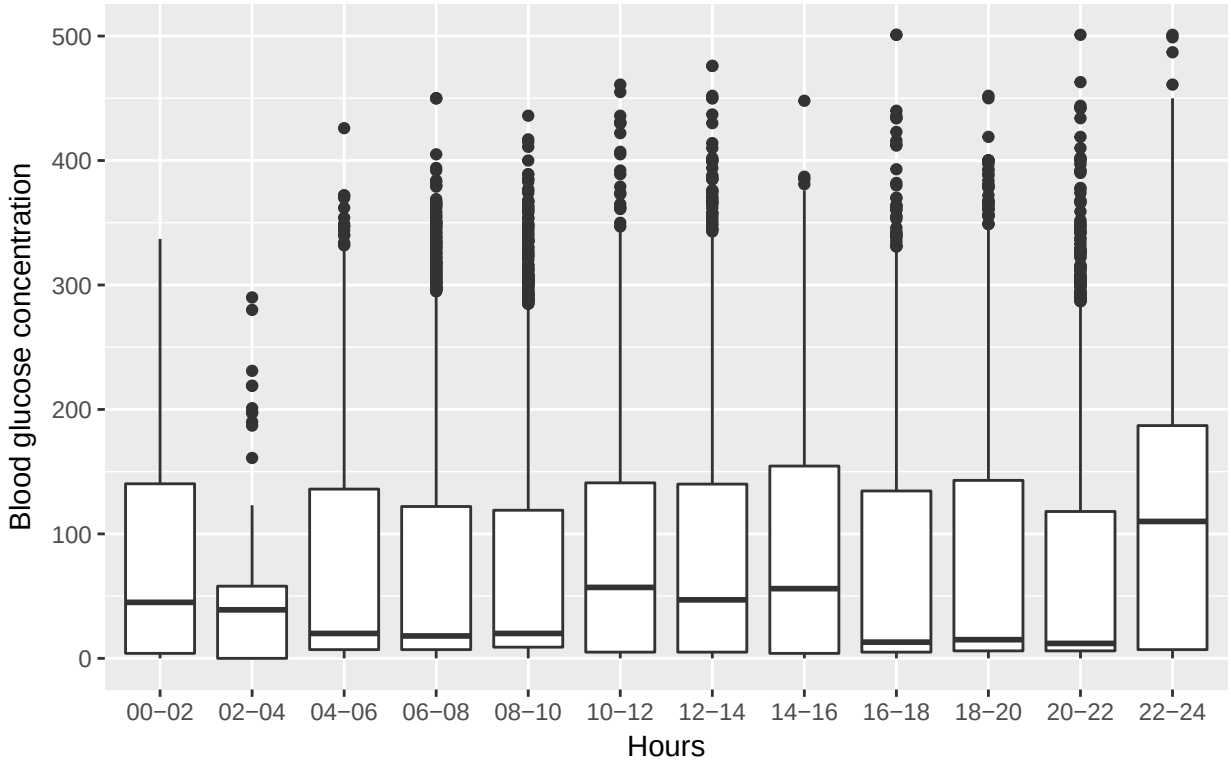

```
ggplot(clean_df, aes(factor(time_bin), bg_conc, col = symptom)) +
  geom_point() +
  geom_jitter() +
  labs(x = "Hours", y = "Blood glucose value mg/dl") +
  ggtitle("Distribution of BG symptoms over 24 hours", subtitle = "Figure 11")
```




```
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.

---