Nibu Jacob
FA14-BL-INFO-I590-19491
School of Informatics and Computing
Indiana University

Mining Electronic Medical Record Data

The purpose of this assignment is to use R together with the packages ggplot2 and corrplot to analyze a given anonymized patient data and answer nine questions regarding the data. The questions cover data science aspects such as descriptive statistics, data visualization, and finding correlations between variables.

Data preparation

The given data is a random sample from Medical Quality Improvement Consortium (MQIC) database of GE Healthcare [(1)]. A flat file in CSV format was loaded into R workspace and examined.

```
data <- read.csv('MQIC_Detailed_slice.csv', stringsAsFactors=FALSE) str(data)
```

The data consists of 2029 records of 25 variables. The factor variables loaded as integer or character datatype were fixed as per the variable description given with the data [(1)].

```
data["STATE"] <- as.factor(data["STATE"])
data["AGE_CATEGORY"] <- as.factor(data["AGE_CATEGORY"])
levels(data$AGE_CATEGORY) <- c("18 to 44 years","45 to 64 years","65 to 79 years","80+ years")
data["GENDER"] <- as.factor(data["GENDER"])
data["DISEASE_CATEGORY"] <- as.factor(data["DISEASE_CATEGORY"])
levels(data$DISEASE_CATEGORY) <- c("diabetes", "hypertension")
```

Identifying the 5 U.S. states in the dataset which have the largest number of patients

The number of patients was aggregated per US state using the aggregate function and the resulting records were ordered in decreasing order according to patient number. The top five records were then selected. *df1 <- aggregate(data$PATIENTS ~ data$STATE, data=data, FUN=sum)*

df1 <- df1[order(df1[,2], decreasing=TRUE),] head(df1, 5)

STATE	PATIENTS
TX	202861
NY	161792
PA	136463
WA	121108
MO	109381

Identifying the 5 U.S. states with the highest number of diabetes patients

A subset of the data consisting of only diabetes patients was taken and the same procedure as above (aggregation followed by ordering) was followed to get the five US states with the highest number of diabetes patients. *df2 <- subset(data, data$DISEASE_CATEGORY == 1)*

df2 <- aggregate(df2$PATIENTS ~ df2$STATE, data=df2, FUN=sum)

df2 <- df2[order(df2[,2], decreasing=TRUE),] head(df2, 5)

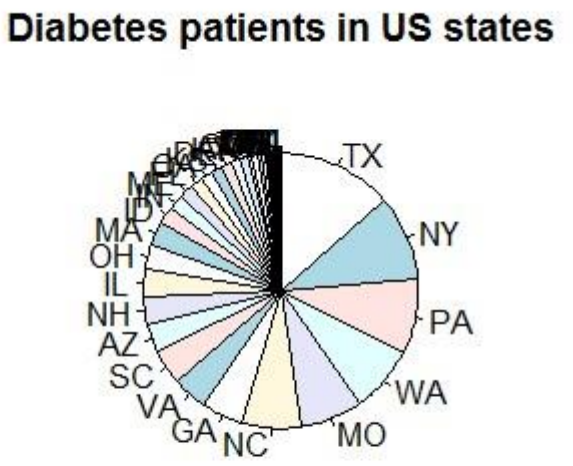
STATE	PATIENTS
TX	102126
NY	75512
PA	66941
WA	58645
MO	54971

Number of diabetes patients in each state

The same subset of data used in the above question was used to create a pie chart. R has limited features for pie chart since pie charts are generally not recommended for various reasons[2, 3]. The pie() function was used to draw a pie chart.

pie(df2[,2], labels=df2[,1], clockwise=TRUE) title(main="Diabetes

patients in US states", font.main=2)



States having the highest and lowest mean BMIs

The records were ordered in ascending order according to mean BMI value and the states corresponding to the highest and lowest values were selected.

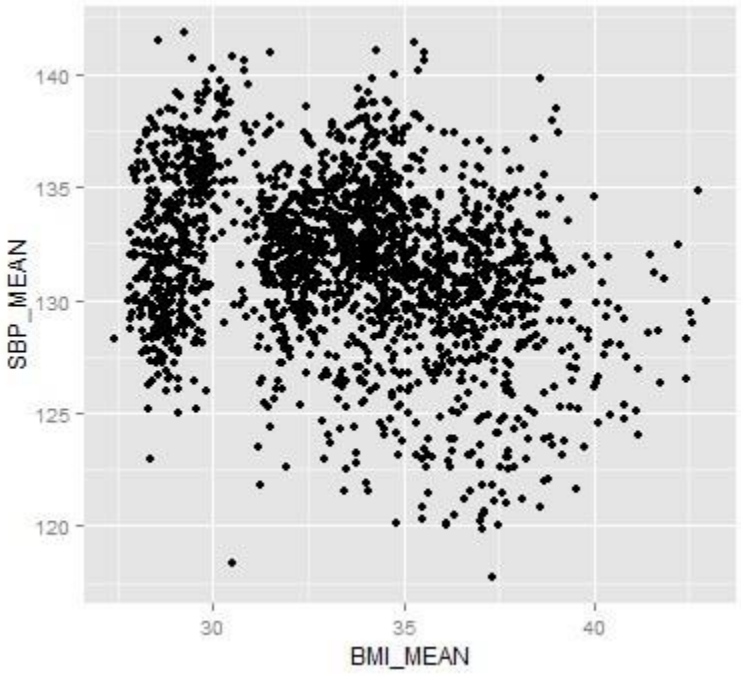
```
df4 <- subset(data, select=c(STATE, BMI_MEAN)) df4
<- df4[order(df4$BMI_MEAN),]
head(df4, 1) tail(df4,
1)
```

The state MO has the highest mean BMI value for any patient group (BMI = 42.98) while the state AR has the lowest mean BMI value for any patient group (BMI = 27.38).

Plotting mean BMI against mean systolic blood pressure to find correlation

The package ggplot2 was used for plotting a scatter chart of the two variables [(4)].

```
library(ggplot2)
df5 <- subset(data, select=c(BMI_MEAN, SBP_MEAN))
ggplot(data=df5, aes(x=BMI_MEAN, y=SBP_MEAN))+
geom_point()
```

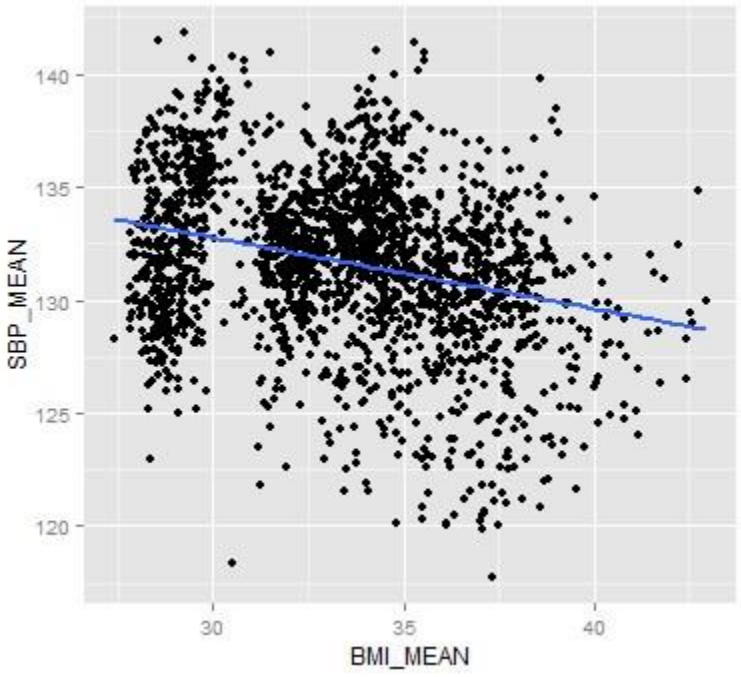


The plot shows the data points scattered without an apparent linear structure. A Pearson productmoment correlation coefficient was computed to assess the relationship between mean BMI and mean SBP.

```
> cor.test(df5$BMI_MEAN, df5$SBP_MEAN)

        Pearson's product-moment correlation

data:  df5$BMI_MEAN and df5$SBP_MEAN t = -
12.9266, df = 2027, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.3156916 -0.2352752 sample
estimates:
       cor
-0.2759662
```

The results suggest a statistically significant weak negative correlation between the two variables, r = -0.28, df = 2027, p < 0.05. Since the absolute value of r is less than 0.3, the correlation can be considered as weak [5,6]. Having established a significant linear relationship, a line of best fit was drawn using a linear model of the data [7].

```
ggplot(data=df5, aes(x=BMI_MEAN, y=SBP_MEAN))+
geom_point()+
 geom_smooth(method="lm",se=FALSE,size=1)
```

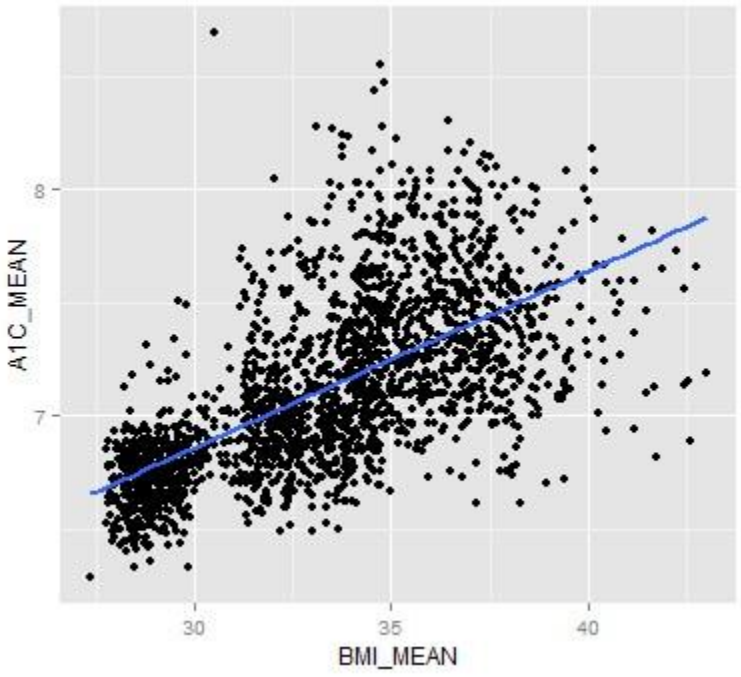


Plotting mean BMI against mean A1C to find correlation

The package ggplot2 was used for plotting a scatter plot of the two variables.

```
df6 <- subset(data, select=c(BMI_MEAN, A1C_MEAN))
ggplot(data=df6, aes(x=BMI_MEAN, y=A1C_MEAN))+
geom_point()+
 geom_smooth(method="lm",se=FALSE, size=1)
```

Since the scatter plot showed a rough linear structure, a line of best fit was also drawn using a linear model of the two variables [7].



A Pearson product-moment correlation coefficient was computed to assess the relationship between mean BMI and mean A1C.

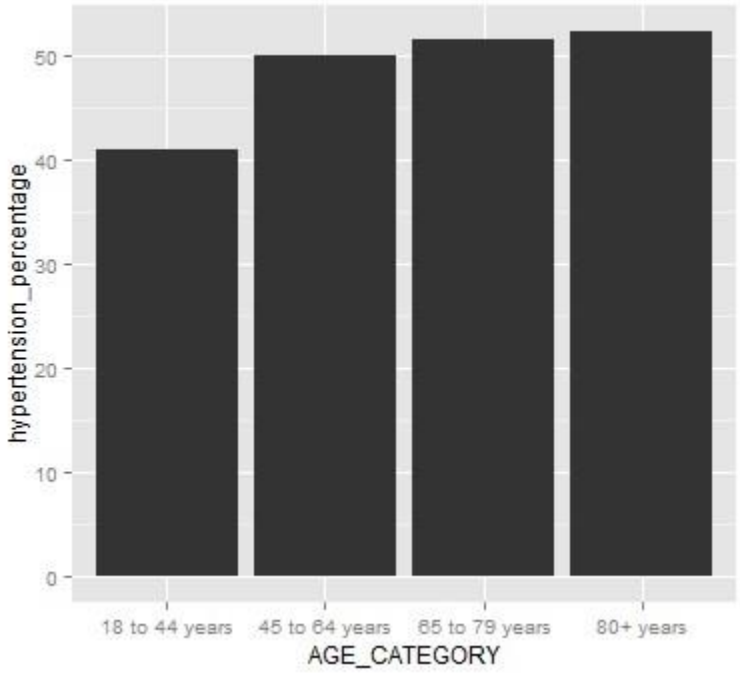
```
> cor.test(df6$BMI_MEAN, df6$A1C_MEAN)

Pearson's product-moment correlation

data:  df6$BMI_MEAN and df6$A1C_MEAN t =
37.2637, df = 2027, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6110453 0.6627354 sample
estimates:
      cor  0.6376074
```

The result suggests a statistically significant strong positive correlation between the two variables, r = 0.64, df = 2027, p < 0.05. As the mean BMI increases there is a corresponding increase in mean A1C. The r value of 0.64 is closer to 1 and that indicates a linear relationship.

Age bracket having the highest frequency of hypertension

A subset of the data containing records of only patients with hypertension was taken and the number of patients was aggregated per age bracket. A barplot of the percentage of patients with hypertension (out of the total number of patients in a patient bracket) against the age brackets was created using ggplot2.

```
df7 <- subset(data, DISEASE_CATEGORY=="hypertension", select=c(AGE_CATEGORY, PATIENTS))
df7 <- aggregate(PATIENTS ~ AGE_CATEGORY, data=df7, FUN=sum)
totalPatients <- aggregate(data$PATIENTS ~ data$AGE_CATEGORY, data=data, FUN=sum) df7
<- cbind(df7,totalPatients[,2])
df7$hypertension_percentage <- (df7[,2]/df7[,3])*100
ggplot(data=df7, aes(x=AGE_CATEGORY, y=hypertension_percentage))+
geom_bar(stat="identity")
```



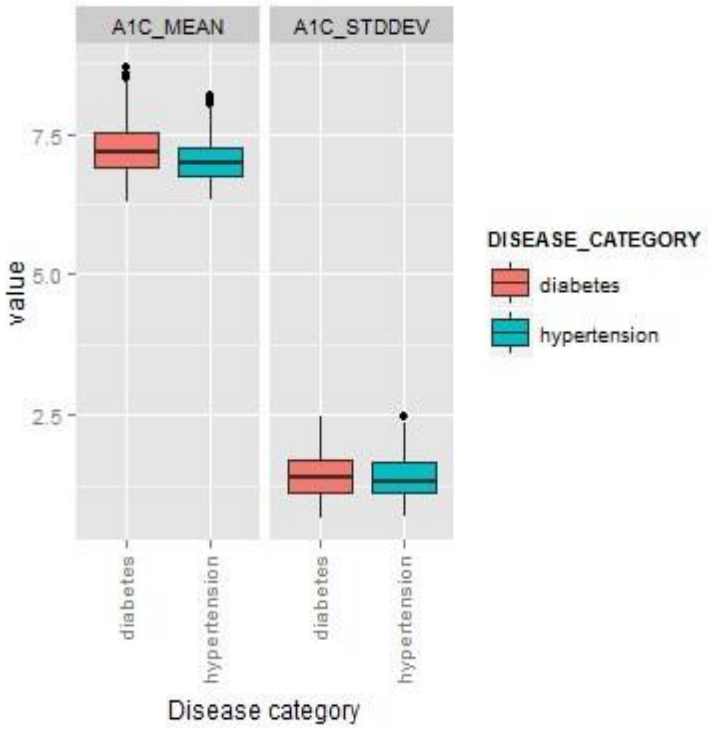
From the barplot it is clear that the age group "80+ years" has the highest frequency of hypertension.

Box plot (showing means, deviations) of all mean descriptors (A1C, weight, BMI, FBG, SBP, DBP) for both patient groups

Box plot is a non-parametric (no assumptions about the distribution of underlying data) descriptive statistical display tool used to compare the distribution of values between different levels of a factor variable [(8)]. A box plot gives information about the median, range, interquartile range (IQR), and skewness of the data [(9)]. The following custom R function was used to construct the box plots for side-by-side comparison of similar variables.

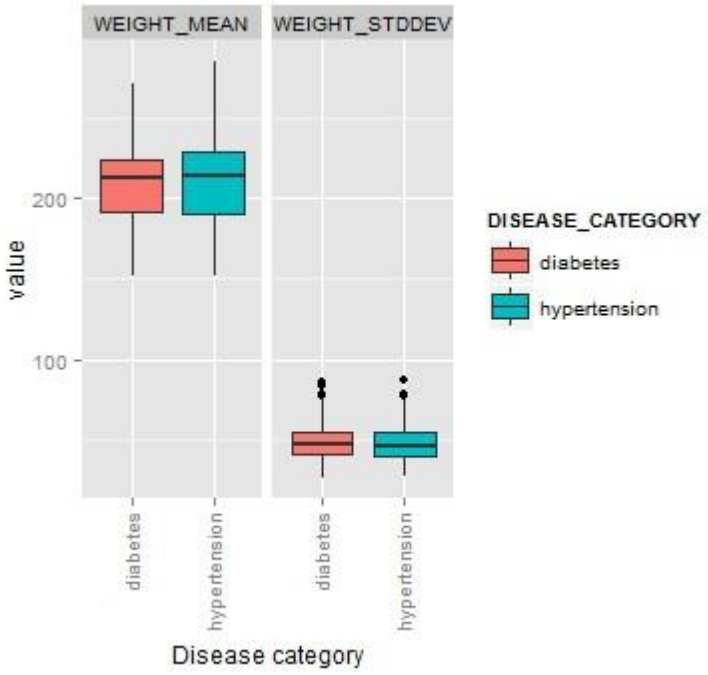
```
library(reshape2) df8 <- subset(data,
select=c(DISEASE_CATEGORY,A1C_MEAN,A1C_STDDEV,WEIGHT_MEAN,WEIGHT_STDDEV,
BMI_MEAN,
```

```
BMI_STDDEV,FBG_MEAN,FBG_STDDEV,SBP_MEAN,SBP_STDDEV,DBP_MEAN,DBP_STDDEV))
  plot_boxplot <- function(df8,y1,y2){    df8a <- melt(df8,
measure.vars = y1:y2)    ggplot(df8a, aes(x=DISEASE_CATEGORY,
y=value,fill=DISEASE_CATEGORY))+      geom_boxplot()+
facet_grid(.~variable)+      labs(x="Disease category")+
    theme(axis.text.x=element_text(angle=90,
vjust=0.4,hjust=1))  }

plot_boxplot(df8,2,3)
```



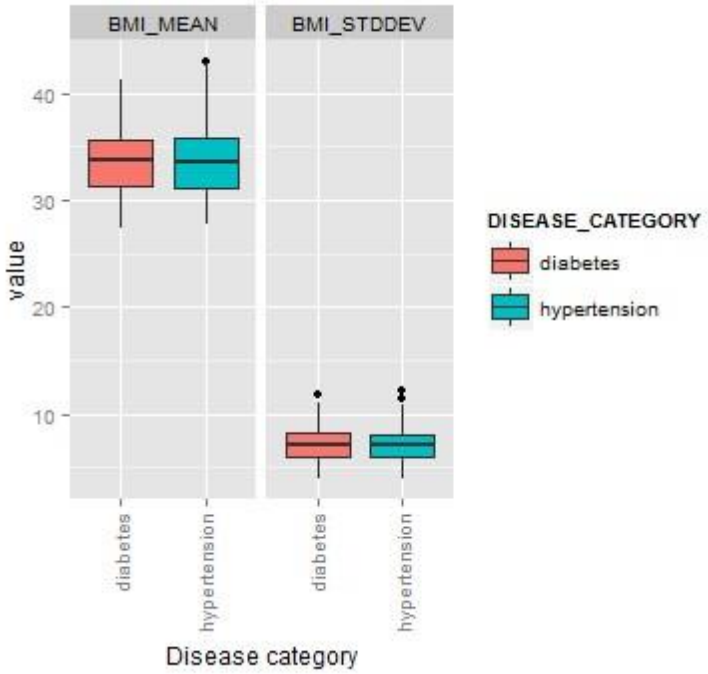
The first two variables chosen were the mean and standard deviation of A1C. The distribution of these two variables were compared against the diabetes and hypertension patient groups. The diabetes group clearly has higher median value and range for the mean of A1C. This is as expected since high A1C is a diagnostic criteria for diabetes [(10)]. Compared to the mean values, the standard deviation values are not very small, indicating there is some variance in A1C values between the patients [(11)].

```
plot_boxplot(df8,4,5)
```



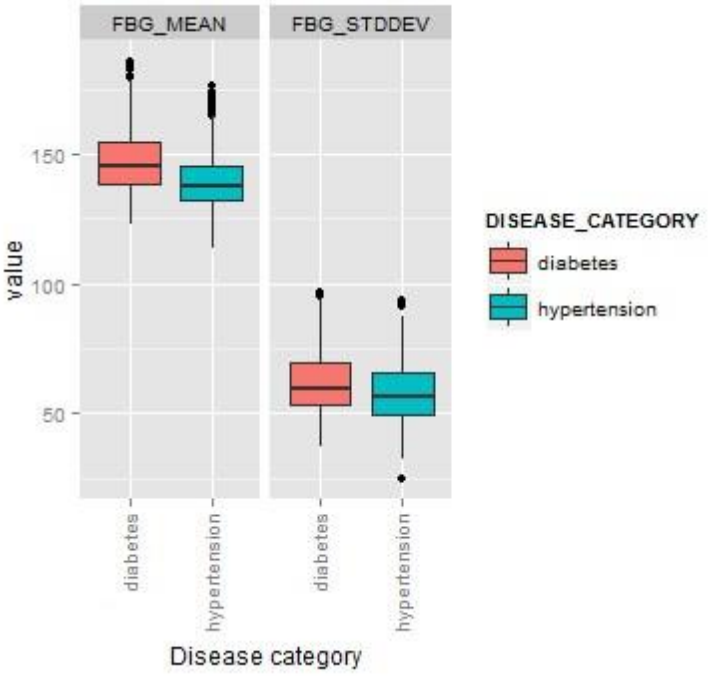
When it comes to body weight, even though both diabetes and hypertension groups show similar median values, the hypertension group has a higher range. Both groups show a skewness indicating there are more of higher body weight values in the distribution.

```
plot_boxplot(df8,6,7)
```



The median value of mean BMI for diabetes patient group is slightly higher than that of hypertension group. However, hypertension group shows a higher range of BMI values. The boxplots of mean weight and mean BMI show remarkably similar distribution. This goes per expectation since BMI is calculated using body weight.

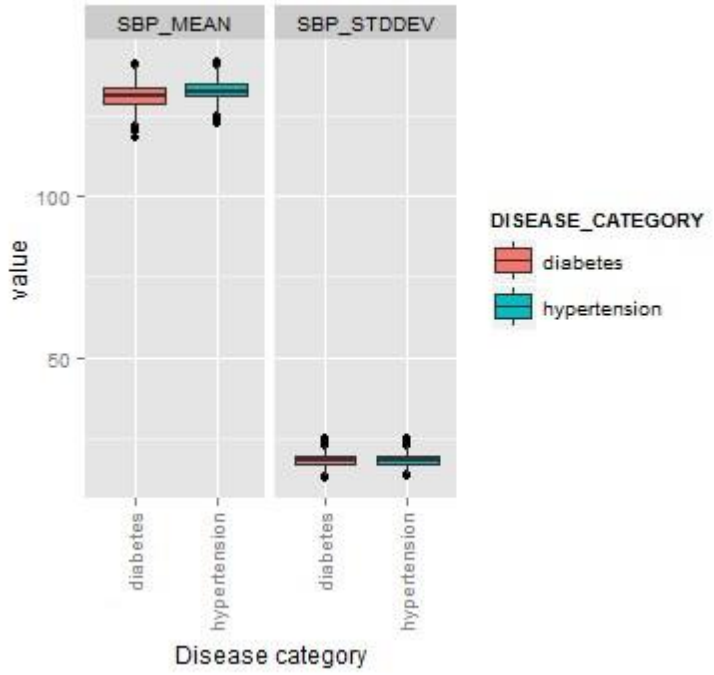
```
plot_boxplot(df8,8,9)
```



Fasting Blood Glucose (FBG) is another diagnostic criteria of diabetes. The box plots show a marked difference in mean FBG values between diabetes and hypertension groups. Both the range and the median values are significantly higher in diabetes group. The standard deviation for both patient groups is quite large compared to the mean values, indicating significant variation among individual patients. The plots also show many outliers at the upper end of the mean values.

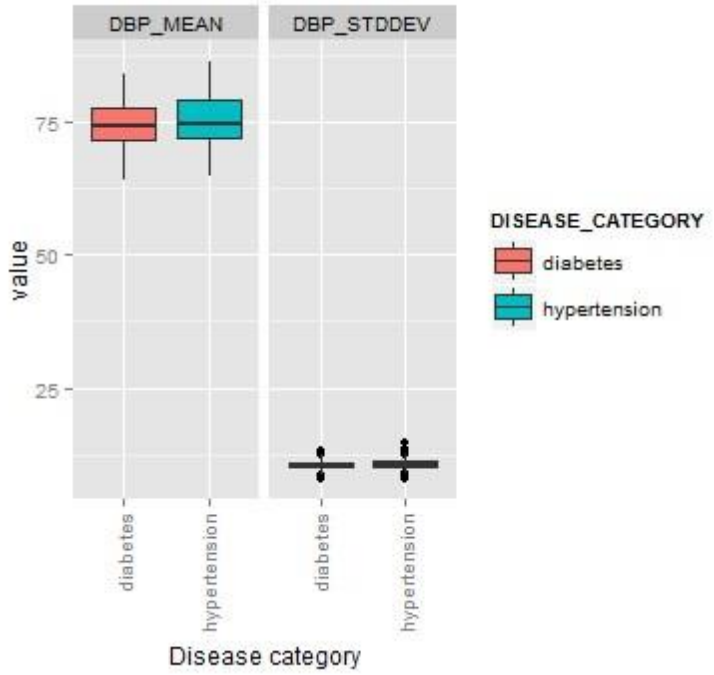
Compared to the boxplots of A1C, FBG shows a remarkable difference between the two patient group and thus supports the current view that FBG is the gold standard of diagnosis for diabetes[12].

```
plot_boxplot(df8,10,11)
```



As expected, the mean systolic blood pressure (SBP) values for hypertension group is significantly higher in both median and range. The relatively small interquartile range and standard deviation shows the values are fairly uniform among the patients.

```
plot_boxplot(df8,12,13)
```

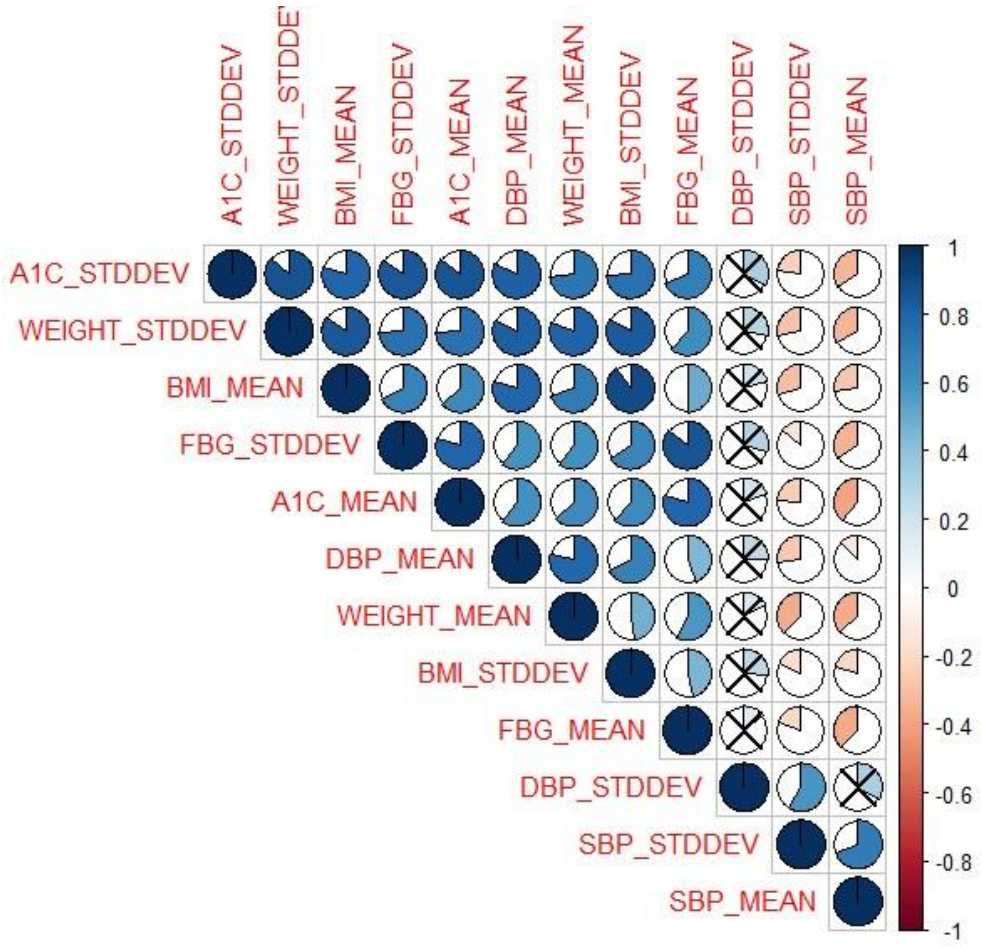


Mean diastolic blood pressure (DBP) shows a higher IQR in hypertension group. However, compared to the SBP, DBP is less characteristic of the hypertension group. The extremely small standard deviation shows the DBP values of individual patients were fairly close to the mean value.

Identifying all highly correlated variables in the dataset

The package corrplot was used to create a correlation matrix of the 12 variables. The package provides a corrplot() function that outputs a plot that visually shows the strength of correlation using colors [13]. A significance test was performed using a custom function cor.mtest() that outputs a list with the first element being a matrix of p-values. This matrix was given as parameter p.mat in the corrplot() function to visually cross out statistically insignificant correlations ($p > 0.05$). The method pie was used to clearly show the strength of correlation (in addition to the intensity of color). The correlation matrix was ordered using First Principal Component (FPC) order that orders the variables according to variance (variable with the largest variance being first) [14].

```
library(corrplot)
df9 <- subset(data, select=c(A1C_MEAN,A1C_STDDEV,WEIGHT_MEAN,WEIGHT_STDDEV,
BMI_MEAN,BMI_STDDEV,FBG_MEAN,FBG_STDDEV,SBP_MEAN,SBP_STDDEV,DBP_MEAN,DBP_STDDEV))
M <- cor(df9)
  cor.mtest <- function(mat, conf.level = 0.95)
{   mat <- as.matrix(mat)   n <- ncol(mat)
  p.mat <- lowCI.mat <- uppCI.mat <- matrix(NA, n, n)
diag(p.mat) <- 0   diag(lowCI.mat) <- diag(uppCI.mat) <- 1   for (i
in 1:(n - 1)) {     for (j in (i + 1):n) {       tmp <-
cor.test(mat[, i], mat[, j], conf.level = conf.level)
p.mat[i, j] <- p.mat[j, i] <- tmp$p.value       lowCI.mat[i, j] <-
lowCI.mat[j, i] <- tmp$conf.int[1]       uppCI.mat[i, j] <-
uppCI.mat[j, i] <- tmp$conf.int[2]
    }
}
  return(list(p.mat, lowCI.mat, uppCI.mat))
}
res1 <- cor.mtest(M, 0.95)
corrplot(M, type="upper", method="pie", p.mat = res1[[1]], insig="pch", order="FPC")
```



Among the statistically significant correlations, the highly correlated (absolute value of $r > 0.7$) variables were identified from the plot:

1) A1C_STDDEV and WEIGHT_STDDEV
2) A1C_STDDEV and BMI_MEAN
3) WEIGHT_STDDEV and BMI_MEAN
4) A1C_STDDEV and FBG_STDDEV
5) WEIGHT_STDDEV and FBG_STDDEV
6) A1C_STDDEV and A1C_MEAN
7) WEIGHT_STDDEV and A1C_MEAN
8) FBG_STDDEV and A1C_MEAN
9) A1C_STDDEV and DBP_MEAN
10) WEIGHT_STDDEV and DBP_MEAN
11) BMI_MEAN and DBP_MEAN

12) A1C_STDDEV and WEIGHT_MEAN
13) WEIGHT_STDDEV and WEIGHT_MEAN
14) DBP_MEAN and WEIGHT_MEAN
15) A1C_STDDEV and BMI_STDDEV
16) WEIGHT_STDDEV and BMI_STDDEV
17) BMI_MEAN and BMI_STDDEV
18) FBG_STDDEV and FBG_MEAN
19) A1C_MEAN and FBG_MEAN

All of the 19 highly correlated variables are showing positive correlation. All negative correlations among the variables are weak. 10 correlations are statistically insignificant ($p > 0.05$). The rest of the correlations are moderate or weak in strength.

Conclusion

A given patient data was prepared for analysis and analysed using R. The packages ggplot2 and corrplot were used to create plots that visually represent the results. Various descriptive statistics like order, distribution, correlation, etc. were explored for answering a given set of questions about the data.

References

1. http://www.visualizing.org/datasets/mqic-patient-data-detailed-sample
2. http://www.statmethods.net/graphs/pie.html
3. https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/pie.html
4. http://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
5. http://onlinestatbook.com/2/describing_bivariate_data/pearson.html
6. http://www.strath.ac.uk/aer/materials/4dataanalysisineducationalresearch/...
7. http://docs.ggplot2.org/current/geom_abline.html
8. http://en.wikipedia.org/wiki/Box_plot
9. http://stattrek.com/statistics/charts/boxplot.aspx

10. http://diabetes.niddk.nih.gov/dm/pubs/A1CTest/
11. https://www.gastro.org/practice/quality-initiatives/performancemeasures/Calculating_Mean_and_Standard_Deviation.pdf
12. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024379/
13. http://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
14. http://en.wikipedia.org/wiki/Principal_component_analysis