BloodPressurePredict.Rmd

WONG Yuen Wah

16 July 2021

Blood Pressure Prediction without Medical Device

This project amis to predict Blood Pressure without medical device.

Why Blood Pressure is Important

Nearly half of adults in the United States (108 million, or 45%) have hypertension (also called high blood pressure (HBP)) (ie., systolic blood pressure ≥ 130 mm Hg).

Blood Pressure is an important indicator for predicting health.

Long-term high blood pressure is a major risk factor for stroke, heart disease, heart failure, vision loss, chronic kidney disease, and dementia.

Existing Mean to Measure Blood Pressure and the Limitation

To keep track a peron blood pressure, blood pressure guage can be used. However, it required professionals assistance, and most people may not have a blood pressure guage on hands.

Some people do not even know that they have high blood pressure issue.

AI Algorithm to Instantly Estimate Blood Pressure

Have a blood pressure AI algorithms to instantly estimated their blood pressure with simple parameters, such age, BMI, Hours of Sleep, etc. could provide an early alert and attention on their health. If there is double, they should go for a proper checkup and medical follow-up.

The Dataset

Dataset is from National Health and Nutrition Examination Survey of US (NHANES)

Size of Dataset

```
> [1] 6779   76

> The NHANES dataset contains 6779 unique people records.

> Each Records has 76 features.

> Number of Adult Female Aged 18-65 with Blood Pressure and BMI Records :  1886

> Number of Adult Male Aged 18-65 with Blood Pressure and BMI Records :  1908
```

More Idea about the Dataset

```
> Rows: 6,779
> Columns: 76


> $ ID                <dbl> 51624, 51625, 51630, 51638, 51646, 51647, 51654, 516…
> $ SurveyYr          <chr> "2009_10", "2009_10", "2009_10", "2009_10", "2009_10…
> $ Gender            <chr> "male", "male", "female", "male", "male", "female", …
> $ Age               <dbl> 34, 4, 49, 9, 8, 45, 66, 58, 54, 10, 58, 50, 9, 33, …
> $ AgeDecade         <chr> "30-39", "0-9", "40-49", "0-9", "0-9", "40-49", "60-…
> $ AgeMonths         <dbl> 409, 49, 596, 115, 101, 541, 795, 707, 654, 123, 700…
> $ Race1             <chr> "White", "Other", "White", "White", "White", "White"…
> $ Race3             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ Education         <chr> "High School", NA, "Some College", NA, NA, "College …
> $ MaritalStatus     <chr> "Married", NA, "LivePartner", NA, NA, "Married", "Ma…
> $ HHIncome          <chr> "25000-34999", "20000-24999", "35000-44999", "75000-…
> $ HHIncomeMid       <dbl> 30000, 22500, 40000, 87500, 60000, 87500, 30000, 100…
> $ Poverty           <dbl> 1.36, 1.07, 1.91, 1.84, 2.33, 5.00, 2.20, 5.00, 2.20…
> $ HomeRooms         <dbl> 6, 9, 5, 6, 7, 6, 5, 10, 6, 10, 10, 4, 3, 11, 5, 7, …
> $ HomeOwn           <chr> "Own", "Own", "Rent", "Rent", "Own", "Own", "Own", "…
> $ Work              <chr> "NotWorking", NA, "NotWorking", NA, NA, "Working", "…
> $ Weight            <dbl> 87.4, 17.0, 86.7, 29.8, 35.2, 75.7, 68.0, 78.4, 74.7…
> $ Length            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ HeadCirc          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ Height            <dbl> 164.7, 105.4, 168.4, 133.1, 130.6, 166.7, 169.5, 181…
> $ BMI               <dbl> 32.22, 15.30, 30.57, 16.82, 20.64, 27.24, 23.67, 23.…
> $ BMICatUnder20yrs  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ BMI_WHO           <chr> "30.0_plus", "12.0_18.5", "30.0_plus", "12.0_18.5", …
> $ Pulse             <dbl> 70, NA, 86, 82, 72, 62, 60, 62, 76, 80, 94, 74, 92, …
> $ BPSysAve          <dbl> 113, NA, 112, 86, 107, 118, 111, 104, 134, 104, 127,…
> $ BPDiaAve          <dbl> 85, NA, 75, 47, 37, 64, 63, 74, 85, 68, 83, 68, 63, …
> $ BPSys1            <dbl> 114, NA, 118, 84, 114, 106, 124, 108, 136, 102, NA, …
> $ BPDia1            <dbl> 88, NA, 82, 50, 46, 62, 64, 76, 86, 66, NA, 66, 56, …
> $ BPSys2            <dbl> 114, NA, 108, 84, 108, 118, 108, 104, 132, 102, 134,…
> $ BPDia2            <dbl> 88, NA, 74, 50, 36, 68, 62, 72, 88, 66, 82, 74, 64, …
> $ BPSys3            <dbl> 112, NA, 116, 88, 106, 118, 114, 104, 136, 106, 120,…
> $ BPDia3            <dbl> 82, NA, 76, 44, 38, 60, 64, 76, 82, 70, 84, 62, 62, …
> $ Testosterone      <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ DirectChol        <dbl> 1.29, NA, 1.16, 1.34, 1.55, 2.12, 0.67, 0.96, 1.16, …
> $ TotChol           <dbl> 3.49, NA, 6.70, 4.86, 4.09, 5.82, 4.99, 4.24, 6.41, …
> $ UrineVol1         <dbl> 352, NA, 77, 123, 238, 106, 113, 163, 215, 7, 29, 64…
> $ UrineFlow1        <dbl> NA, NA, 0.094, 1.538, 1.322, 1.116, 0.489, NA, 0.903…
> $ UrineVol2         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ UrineFlow2        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ Diabetes          <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No"…
> $ DiabetesAge       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ HealthGen         <chr> "Good", NA, "Good", NA, NA, "Vgood", "Vgood", "Vgood…
> $ DaysPhysHlthBad   <dbl> 0, NA, 0, NA, NA, 0, 10, 0, 4, NA, NA, 0, NA, 3, 7, …
> $ DaysMentHlthBad   <dbl> 15, NA, 10, NA, NA, 3, 0, 0, 0, NA, NA, 0, NA, 7, 0,…
> $ LittleInterest    <chr> "Most", NA, "Several", NA, NA, "None", "None", "None…
> $ Depressed         <chr> "Several", NA, "Several", NA, NA, "None", "None", "N…
> $ nPregnancies      <dbl> NA, NA, 2, NA, NA, 1, NA, NA, NA, NA, NA, NA, NA, NA…
> $ nBabies           <dbl> NA, NA, 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
> $ Age1stBaby        <dbl> NA, NA, 27, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ SleepHrsNight     <dbl> 4, NA, 8, NA, NA, 8, 7, 5, 4, NA, 5, 7, NA, 6, 6, 6,…
> $ SleepTrouble      <chr> "Yes", NA, "Yes", NA, NA, "No", "No", "No", "Yes", N…
> $ PhysActive        <chr> "No", NA, "No", NA, NA, "Yes", "Yes", "Yes", "Yes", …
> $ PhysActiveDays    <dbl> NA, NA, NA, NA, NA, 5, 7, 5, 1, NA, 2, 7, NA, NA, NA…
> $ TVHrsDay          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ CompHrsDay        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
> $ TVHrsDayChild     <dbl> NA, 4, NA, 5, 1, NA, NA, NA, NA, 4, NA, NA, 0, NA, N…
> $ CompHrsDayChild   <dbl> NA, 1, NA, 0, 6, NA, NA, NA, NA, 3, NA, NA, 1, NA, N…
> $ Alcohol12PlusYr   <chr> "Yes", NA, "Yes", NA, NA, "Yes", "Yes", "Yes", "Yes"…
> $ AlcoholDay        <dbl> NA, NA, 2, NA, NA, 3, 1, 2, 6, NA, NA, NA, NA, 3, 6,…
```

```
> $ AlcoholYear      <dbl> 0, NA, 20, NA, NA, 52, 100, 104, 364, NA, NA, 0, NA,…
> $ SmokeNow         <chr> "No", NA, "Yes", NA, NA, NA, "No", NA, NA, NA, "Yes"…
> $ Smoke100         <chr> "Yes", NA, "Yes", NA, NA, "No", "Yes", "No", "No", N…
> $ Smoke100n        <chr> "Smoker", NA, "Smoker", NA, NA, "Non-Smoker", "Smoke…
> $ SmokeAge         <dbl> 18, NA, 38, NA, NA, NA, 13, NA, NA, NA, 17, NA, NA, …
> $ Marijuana        <chr> "Yes", NA, "Yes", NA, NA, "Yes", NA, "Yes", "Yes", N…
> $ AgeFirstMarij    <dbl> 17, NA, 18, NA, NA, 13, NA, 19, 15, NA, NA, NA, NA, …
> $ RegularMarij     <chr> "No", NA, "No", NA, NA, "No", NA, "Yes", "Yes", NA, …
> $ AgeRegMarij      <dbl> NA, NA, NA, NA, NA, NA, NA, 20, 15, NA, NA, NA, NA, …
> $ HardDrugs        <chr> "Yes", NA, "Yes", NA, NA, "No", "No", "Yes", "Yes", …
> $ SexEver          <chr> "Yes", NA, "Yes", NA, NA, "Yes", "Yes", "Yes", "Yes"…
> $ SexAge           <dbl> 16, NA, 12, NA, NA, 13, 17, 22, 12, NA, NA, NA, NA, …
> $ SexNumPartnLife  <dbl> 8, NA, 10, NA, NA, 20, 15, 7, 100, NA, NA, 9, NA, 1,…
> $ SexNumPartYear   <dbl> 1, NA, 1, NA, NA, 0, NA, 1, 1, NA, NA, 1, NA, 1, NA,…
> $ SameSex          <chr> "No", NA, "Yes", NA, NA, "Yes", "No", "No", "No", NA…
> $ SexOrientation   <chr> "Heterosexual", NA, "Heterosexual", NA, NA, "Bisexua…
> $ PregnantNow      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …
```

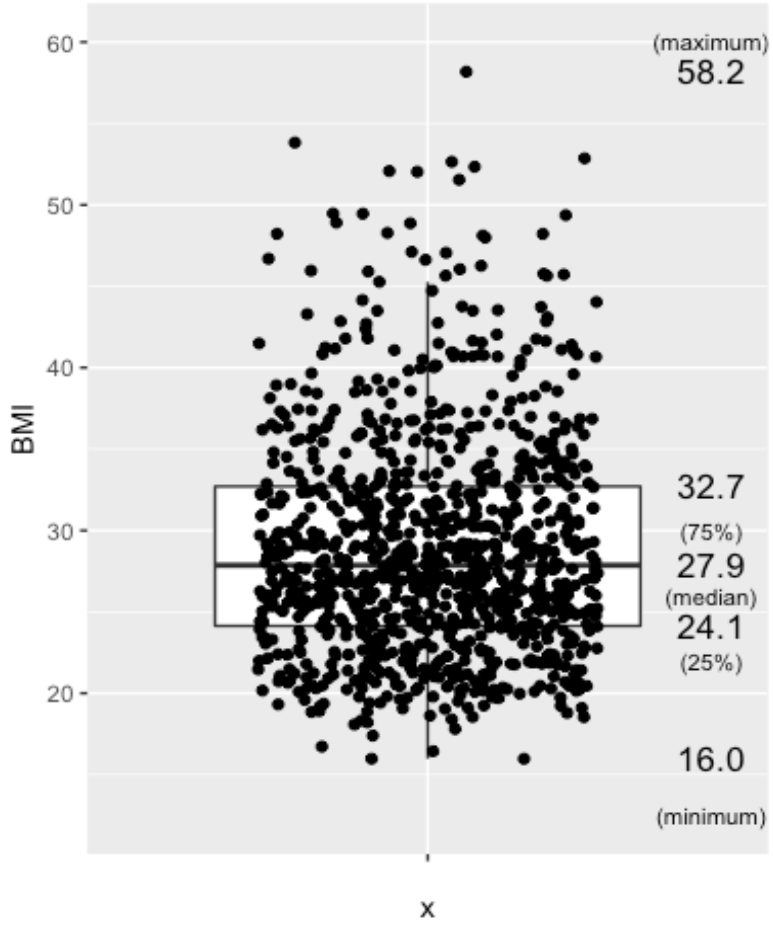
Data Visualization

Univaritate Statistics

Let's take a look on BMI to get a quick idea of the overall health status of population.

Boxplot

Boxplot – provide us with information on the mean/median values of the data.



Age over 65 year old are not considered in this analysis, due to more complicated medications this group of people may have.

The Mean BMI for adult (aged 18-65), with the 1000 samples (for a clearer view), the meam BMI is 27.9, which is considered as overweight, and is quite alarming regarding the health of the population.

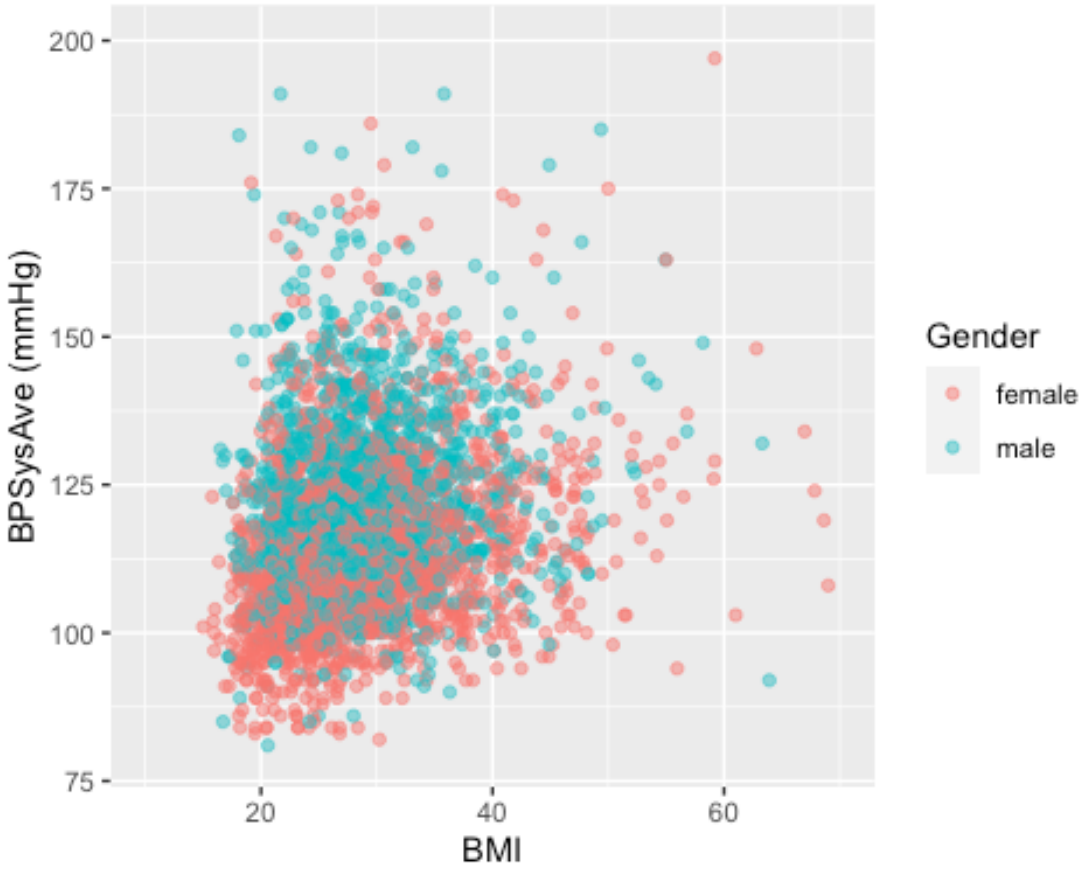
Normal BMI ranging between 18.5 and 24.9. Overweight: BMI between 25 and 29.9. Obese: BMI of 30 or higher.

The mean of BMI 27.9 implies the target population's health condition is at a concerning level. This analysis and Blood Pressure Prediction algorithm could acts as an useful trigger to people to take care of their Blood Pressure and Chronic Health.

Scatter Plot

Blood Pressure vs BMI

Under the same BMI, Male tends to have higher blood pressure then Female.
We would separate the analysis by Genders, and would choose Female in this project.



Confidence Interval

We want to see if the dataset contains the true mean of the population, by looking into the confidence interval (90%, 95%).

Let's look into healthy group people, compare their Blood Pressure with literatures.

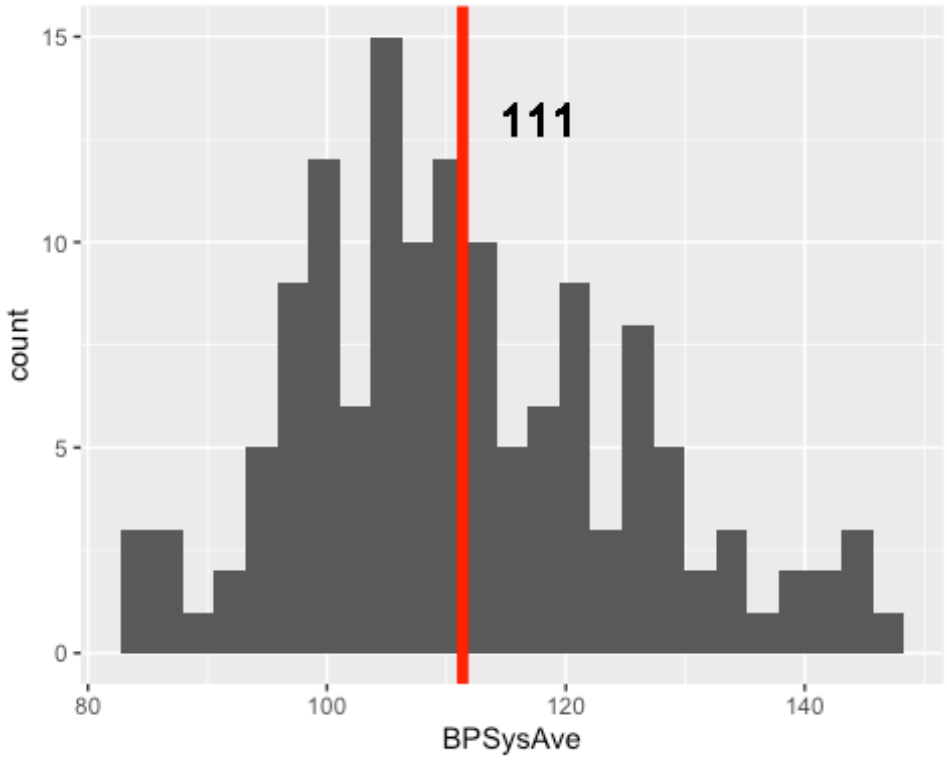
Here, we assume healthy individual as being :
* non-smoker,
* without a history of diabetes,
* no hard drugs,
* no sleeping trouble,
* with a general health that is not considered poor, and
* with a BMI between 18.5 and 25.

Besides,female group is chosen, as we have found female and male blood pressure seems behave a bit differently even under the same BMI.

To investigate the CI, we have to see if the data is normally distributed.

Histogram

Healthy Female Blood Pressure Distribution (mean = 111 mmHg)



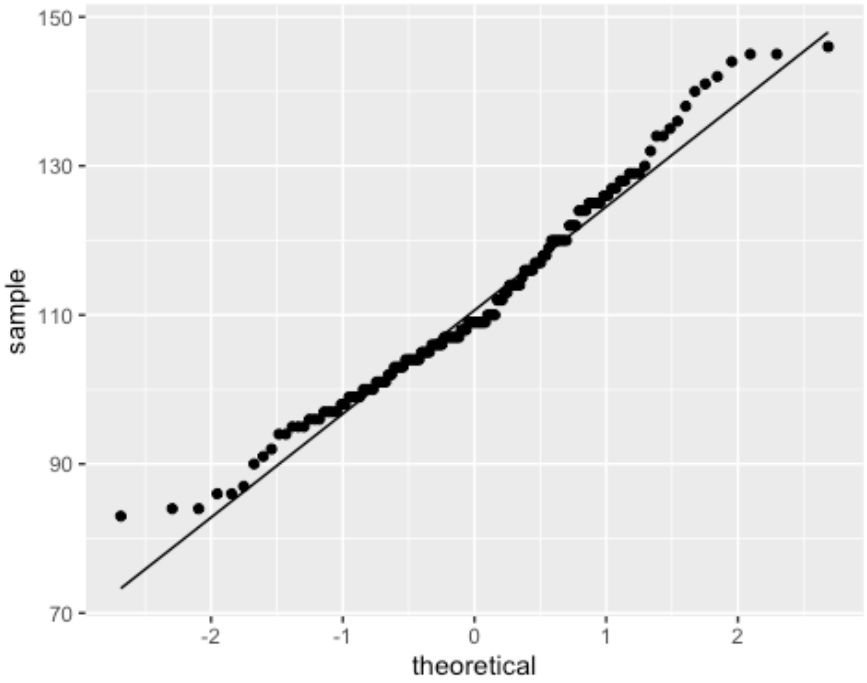

```
> The number of HEALTHY Female Aged 30 to 65 is only 138
```

Besides, it is quite obvious that it is not normally distributed. Long tail is quite common for health data.

Let's also see the QQ Plot.

QQplot

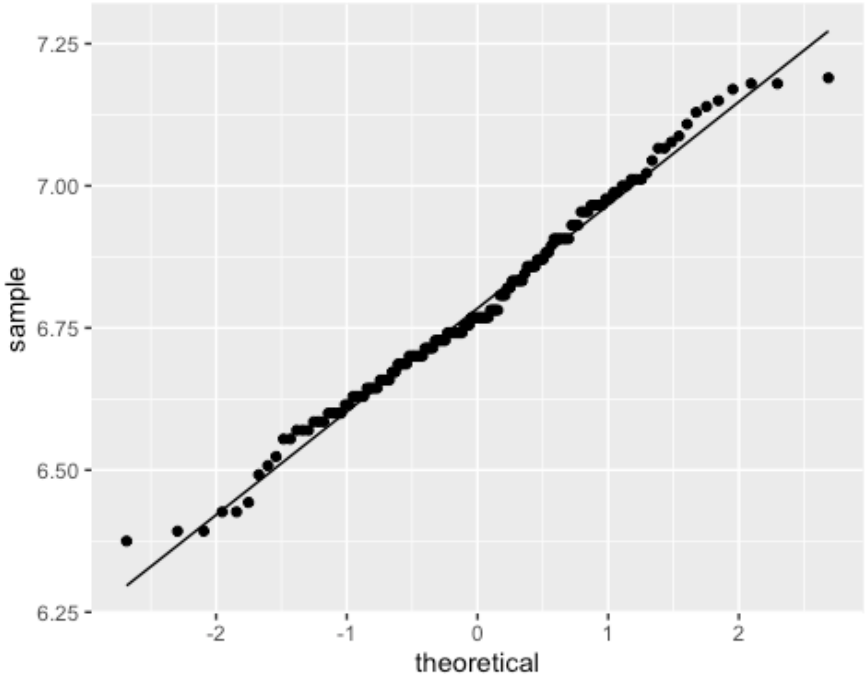
Distribution of Healhty Female (18-65) Blood Pressure



The two ends of the graph do not align well.

We use the log2 transformation, which QQ Plot looks closer to the normal distribution.

Distribution of Healthy Female (18-65) Blood Pressure (with Log2 scale)



The log2 systolic blood pressure for healthy subjects is approximately normally distributed.

90% or 95% confidence interval for the HEALTHY group

General Form of 95% Confidnece Interval = sample statistic mean +/- 2 * stardard error

```
> # A tibble: 1 x 4
>    mean    sd     n     se
>   <dbl> <dbl> <int>  <dbl>
> 1  6.79 0.181   138 0.0154

> Mean value (log2): 6.788

> Geometric mean value (log2): 110.50132885906

> Standard deviation (log2): 0.180701073811383
```

General Form of 95% Confidence Interval= sample statistic mean +/- 2 * stardard error

```
> 95% Confidence Interval (in log2 scale): [6.43;7.15]

> 95% Confidence Interval (mmHg in original scale): [86.02;142]
```

General Form of 90% Confidence Interval= sample statistic mean +/- stardard error

```
> 90% Confidence Interval (in log2 scale): [6.61;6.97]

> 90% Confidence Interval (mmHg in original scale): [97.49;125.2]
```

This allows us to set up the (90%) reference interval, for what we can consider to be normal blood pressure values.

Note that in the literature a value 125 mmHg for the systolic blood pressure is typically considered to be the upper limit of healthy group in aged of 30-65. (90% CI)

Note that in the literature a value 125 mmHg for the systolic blood pressure is typically considered to be the upper limit of normality in aged of 30-65. (95% CI)

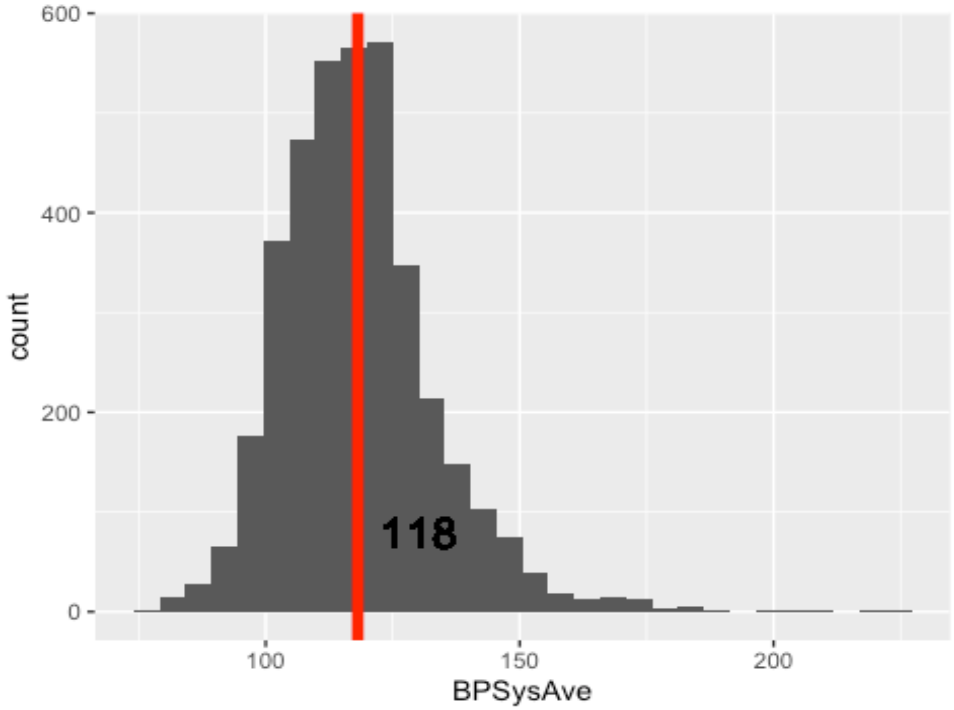
The dataset is reasonably representing, when considering 90% and 95% CI.

Let's continue of analysis

Histogram

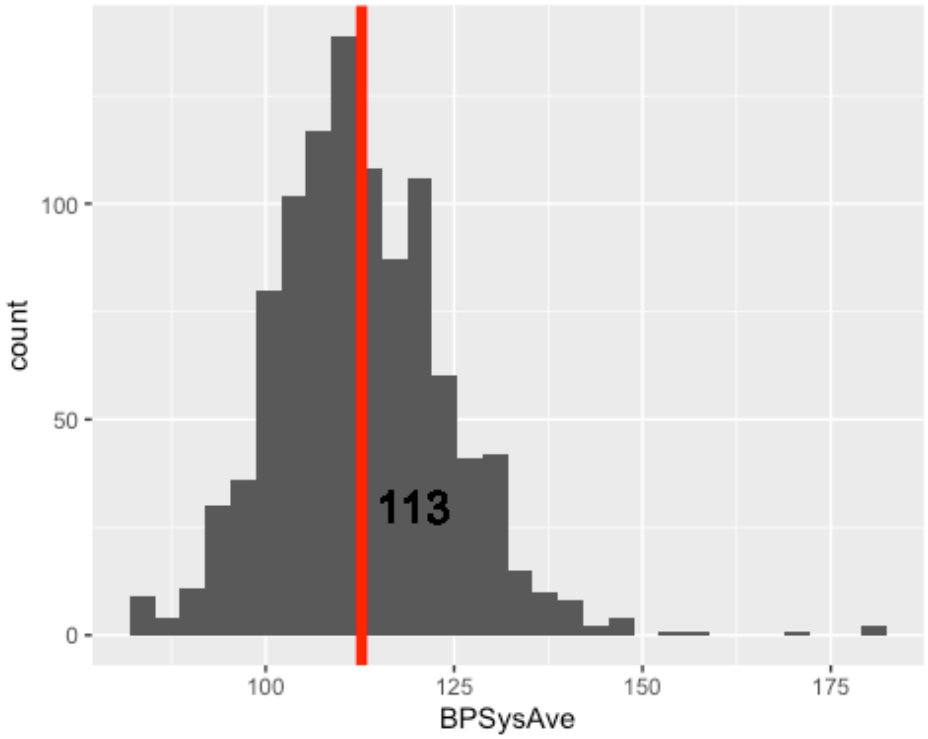
Distribution and Mean of Blood Pressure of All Gender Adult 18-65 (=118mmHg)

(Remarks: Including Healthy and Unhealthy)



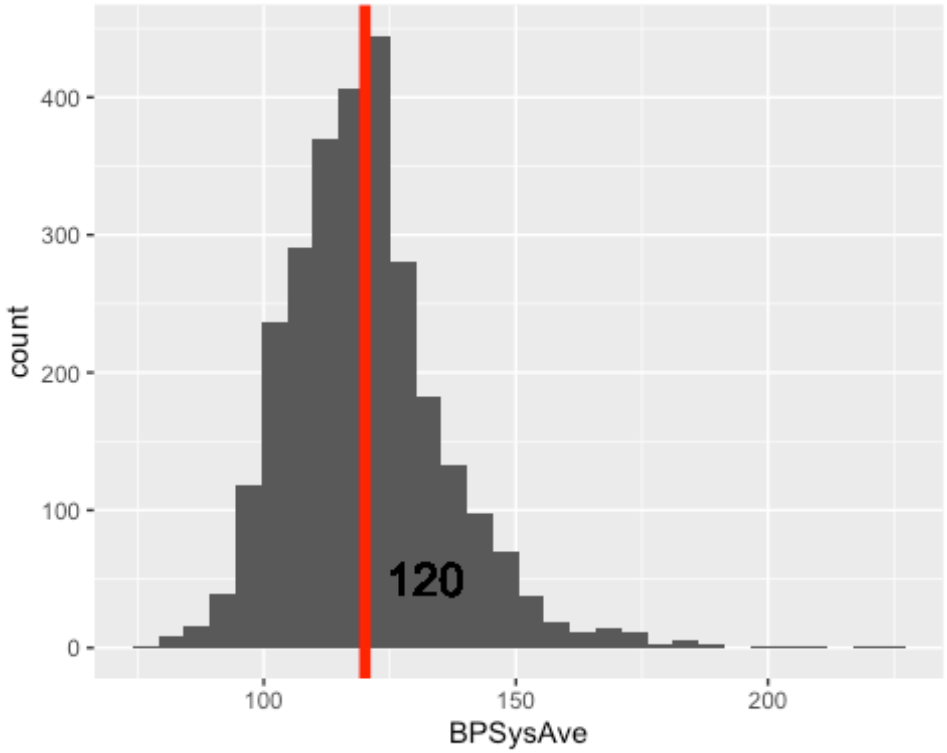
Histogram

Distribution and Mean of Blood Pressure of All Gender Adult 18-29 (=113mmHg)



Histogram

Distribution and Mean of Blood Pressure of All Gender Aged 30-65 (=120mmHg)

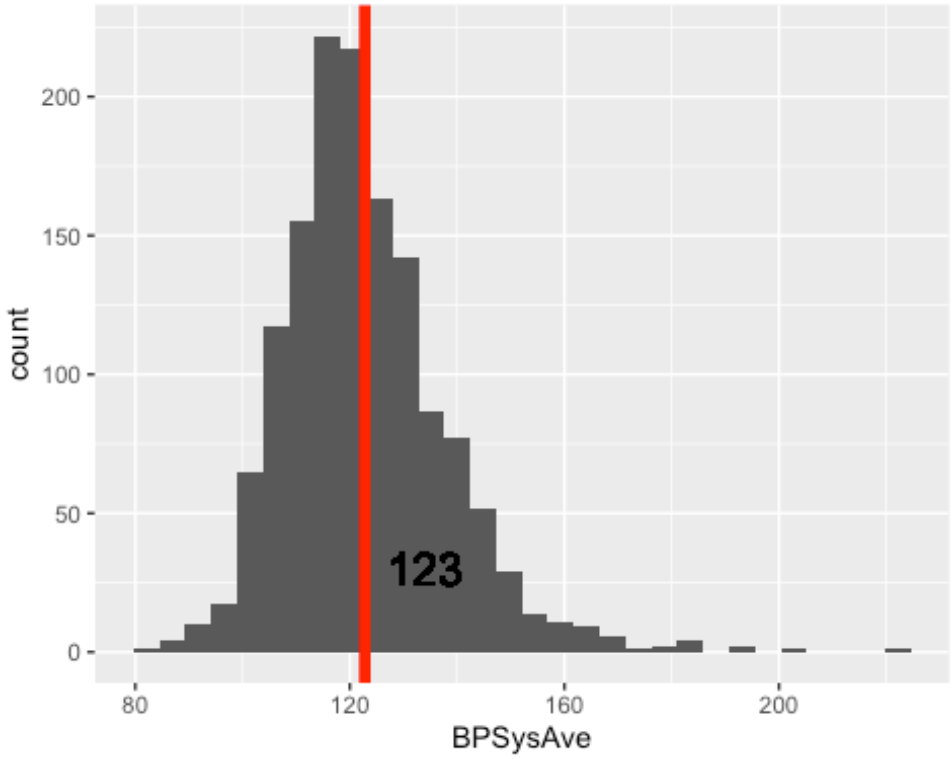


We can see Blood Pressure is higher for aged people.

Age	Blood Pressure
18-65	118mmHg
18-29	113mmHg
30-65	120mmHg

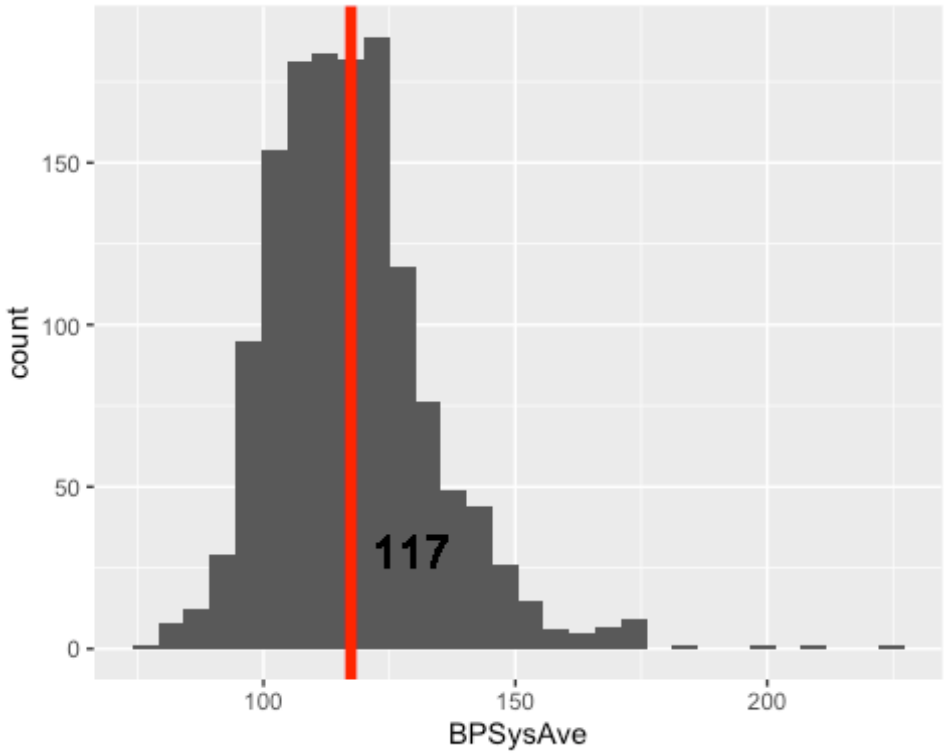
Histogram

Distribution and Mean of Blood Pressure of Male Aged 30-65 (=123mmHg)



Histogram

Distribution and Mean of Blood Pressure of Female Aged 30-65 (=117mmHg)



Summary

Age	Blood Pressure
18-65	118mmHg
18-29	113mmHg
30-65	120mmHg

Age (30-65)	Blood Pressure
Male	123mmHg
Female	117mmHg

Therefore, it makes sense to analyse the Blood Pressure by different Gender and Age Group.

Again, our aim is to predict Blood Pressure by using HANDY features of a person.

"Handy Features" means the parameter the user must know by themselves about their body easily. So that they can predict their blood pressure anytime and get a quick idea of their health conditions instantly.

We will focus on Female Aged 30 to 65. And see what Handy features can be used to predict blood pressure.

From the literature of human physologies, possible factor (featurers) that affect a person blood pressure would be: "Age", "Weight", "Height", "BMI", "SleepHrsNight", "PhysActiveDays", "AlcoholDay".

Let's see how they are related among one another and to the blood pressure.

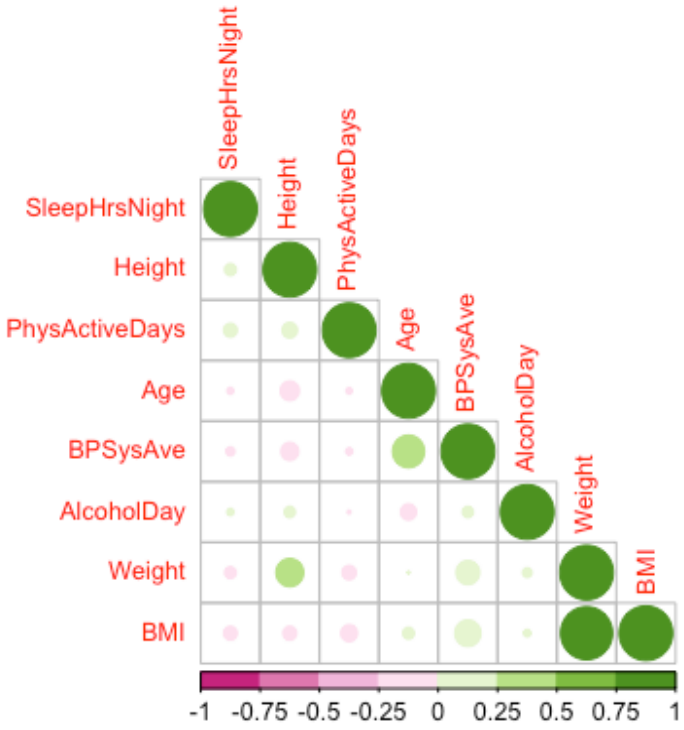
Female aged 30 to 65 — Female3065 (including Healthy and UnHealthy)

The Correlation Heatmap –

Explore the relationship (or the correlation) among all this parameter

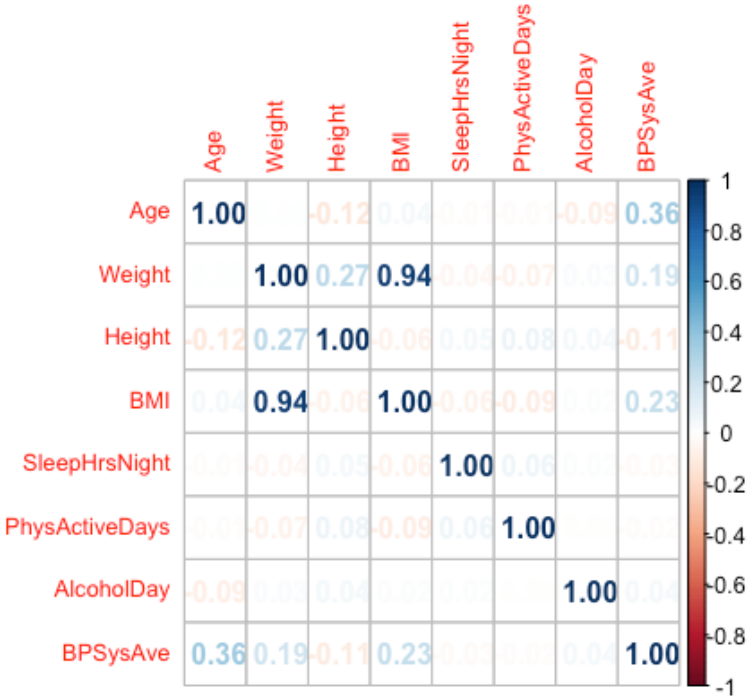
```
> The number of Female Aged 30 to 65 is 1446 (including all healthy and unhealthy).
```

Type 1 – color and sizes only



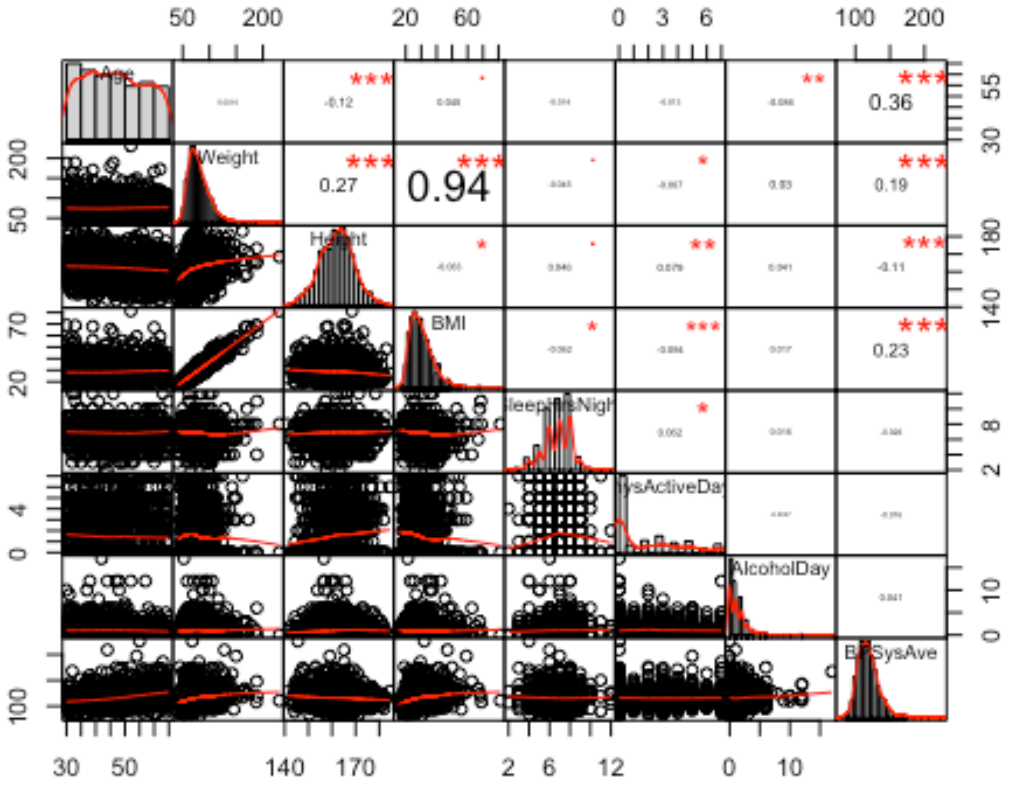
Type 2 – with parameters shown

The larger the parameters, the strong the relationship



Form the correlation table, Age is an obvious feature that affect Blood Pressure.
Besides, Weigth, Height and BMI are also being significantly correlated to Blood Pressure.
However, Weigth, Height and BMI are also correlated to one another by among themselves. In such case, we would only choose one of them.
As BMI = Weight/Height^2. And BMI is having the highest correlation parameter with Blood Pressure. Therefore, we choose BMI.

Type 3 – with graph plotted

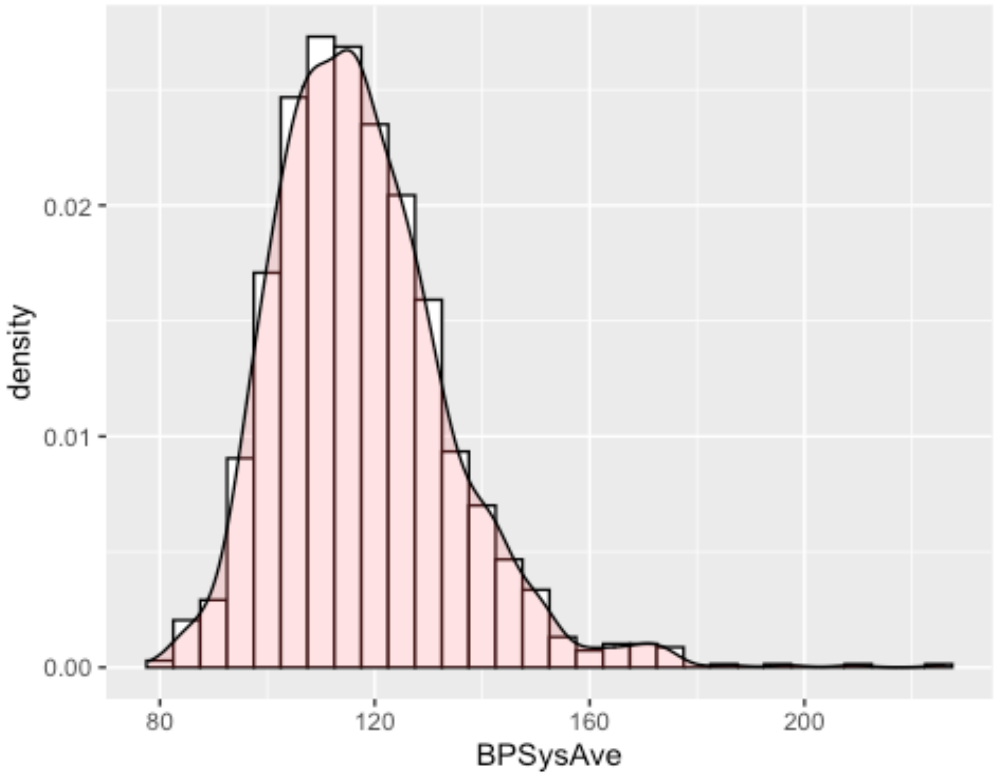


Deeper Look at EACH Handy Features

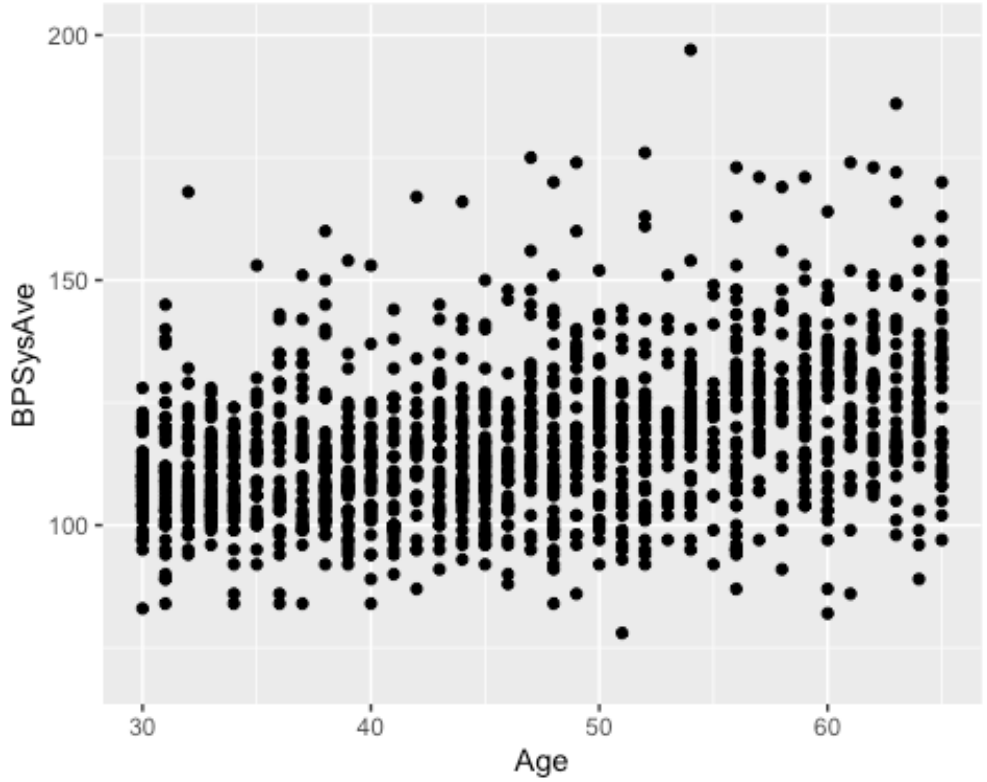
Target Feature --- Blood Pressure

The Distribution in the 1446 blood pressure sample

```
> Again, the number of Female Aged 30 to 65 is 1446 (including all healthy and unhealthy)
.
```

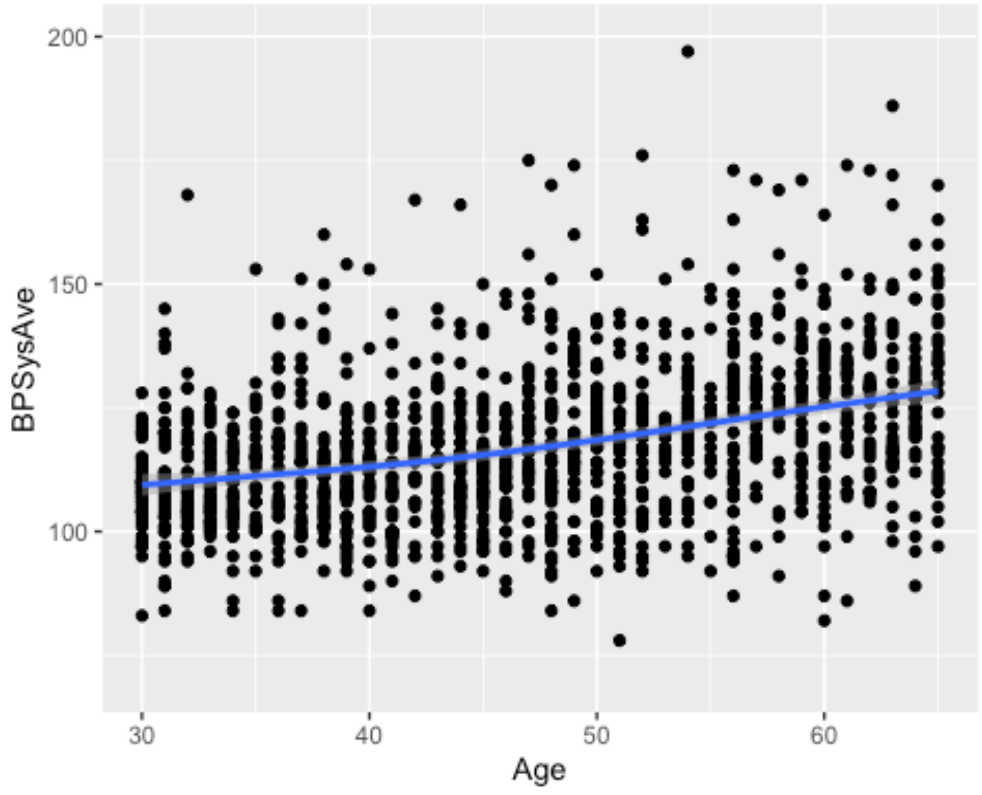


1. Age



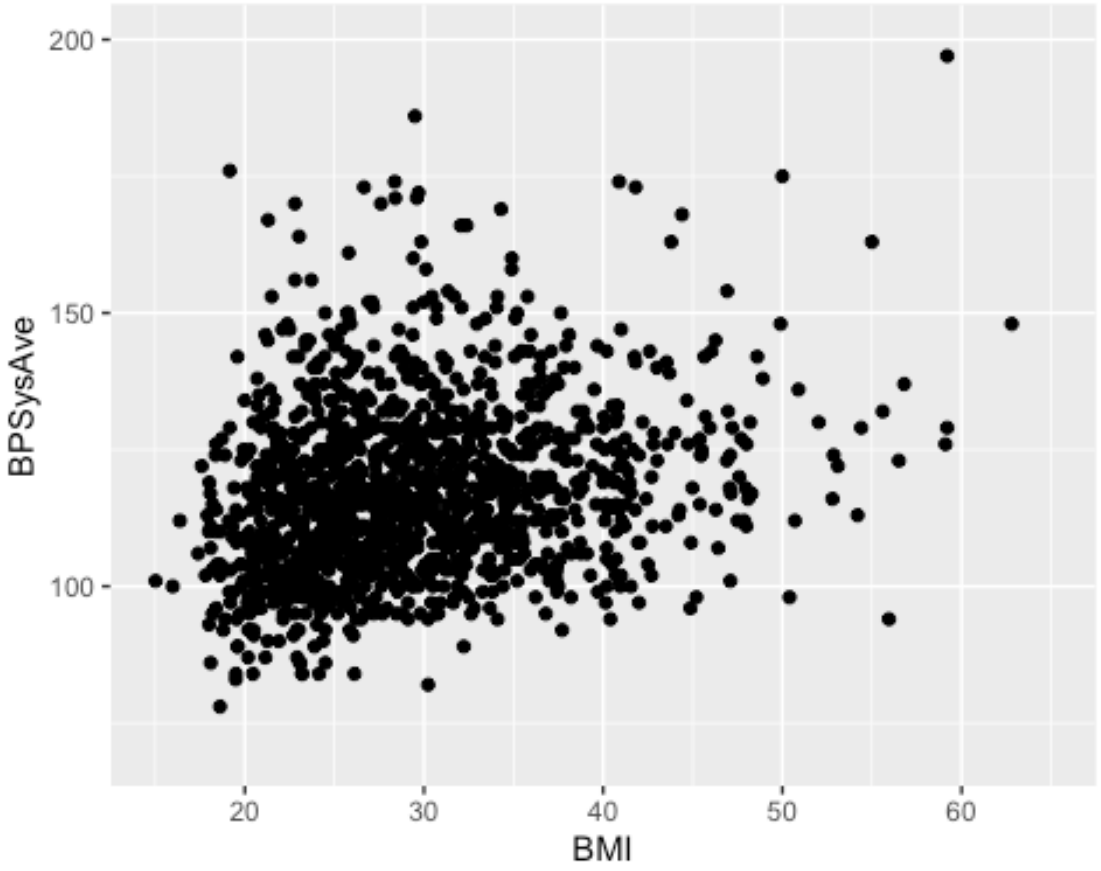
Let's plot for smoothing curve for a clear idea.

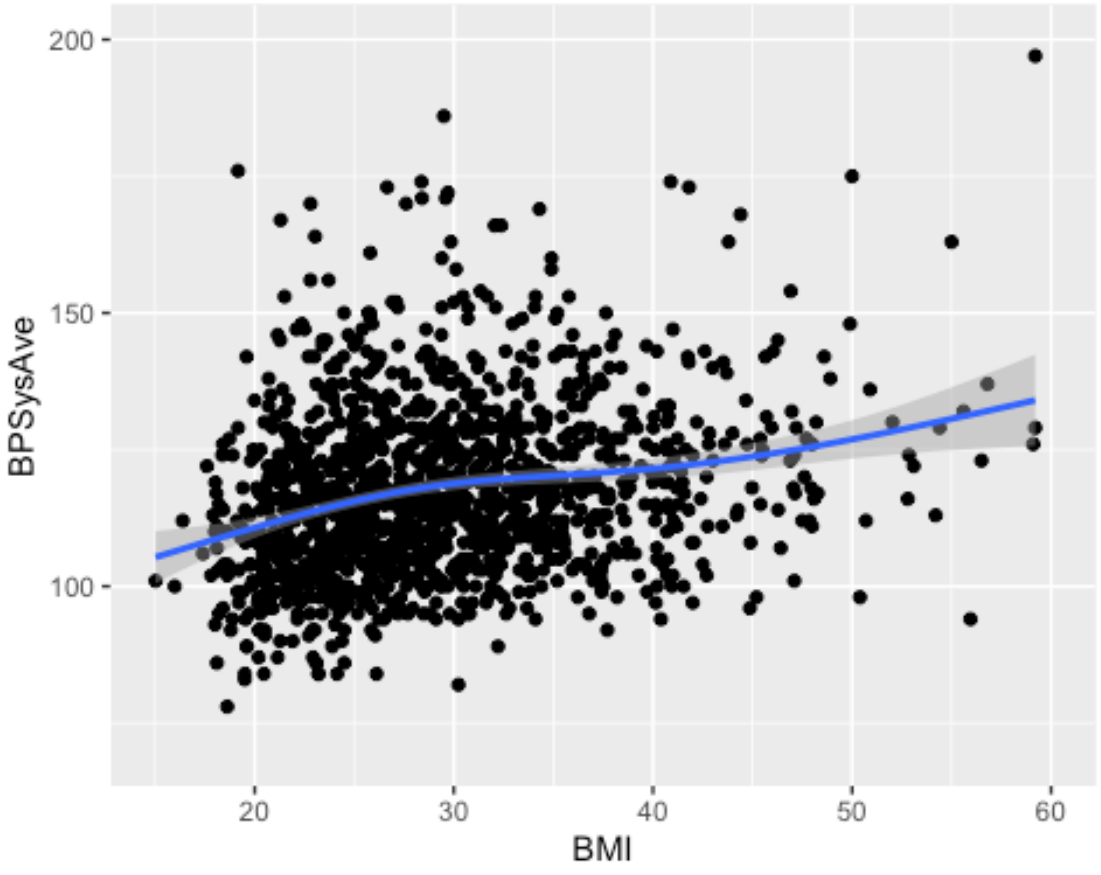
People at age of 30 is having a blood pressure of 113 mmHg. This evalates to almost 130 mmHg at the age of 65.



2. BMI

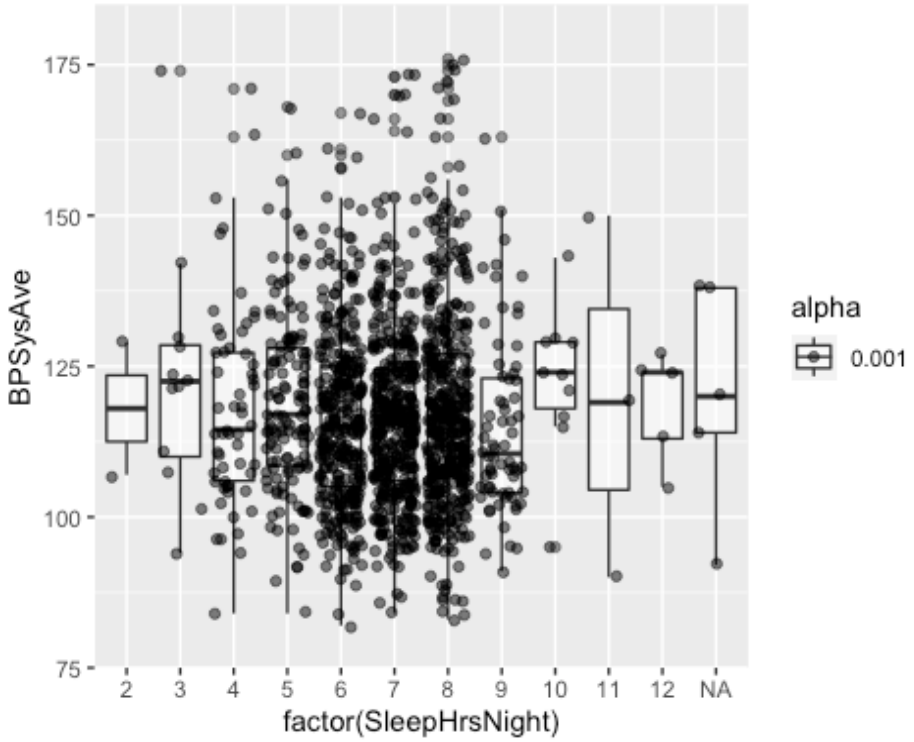
People of BMI = 25 or high is having Blood Pressure of 120 mmHg, the upper limit of healthy range.





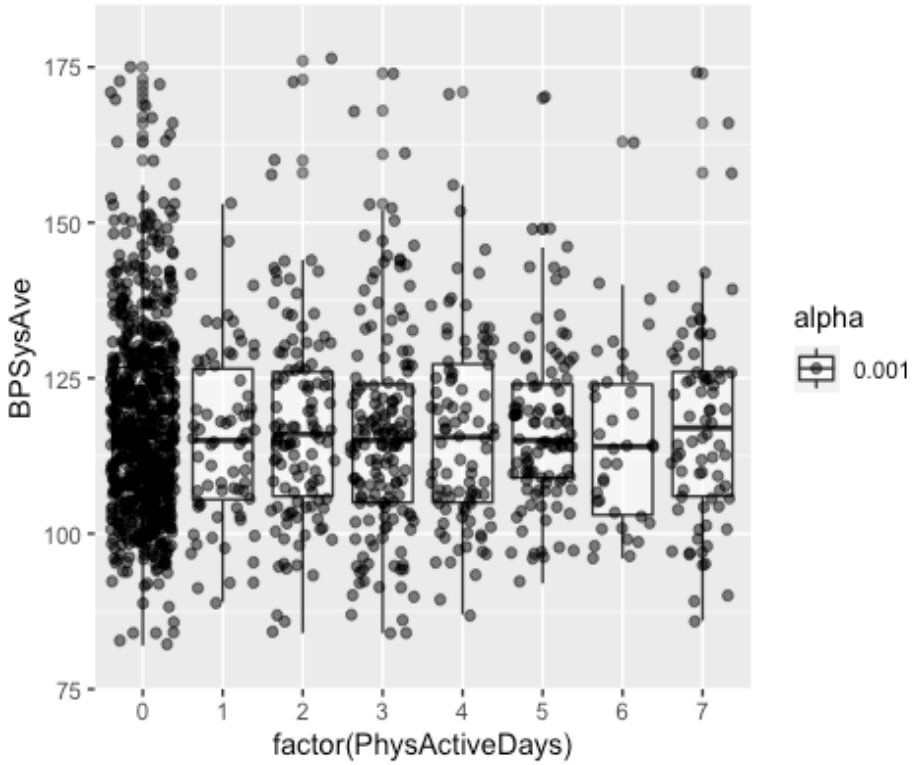
3. SleepHrsNight

People sleeps for 4 to 9 hours are having an average value of lower Blood Pressure closed to 110 mmHg or lower.
Sleeping less than 4 hours or more an 9 hours have an significant increase in the group blood pressure up to almost 125 mmHg.
This is something we can't observe just from the correlation chart (or the heat map).



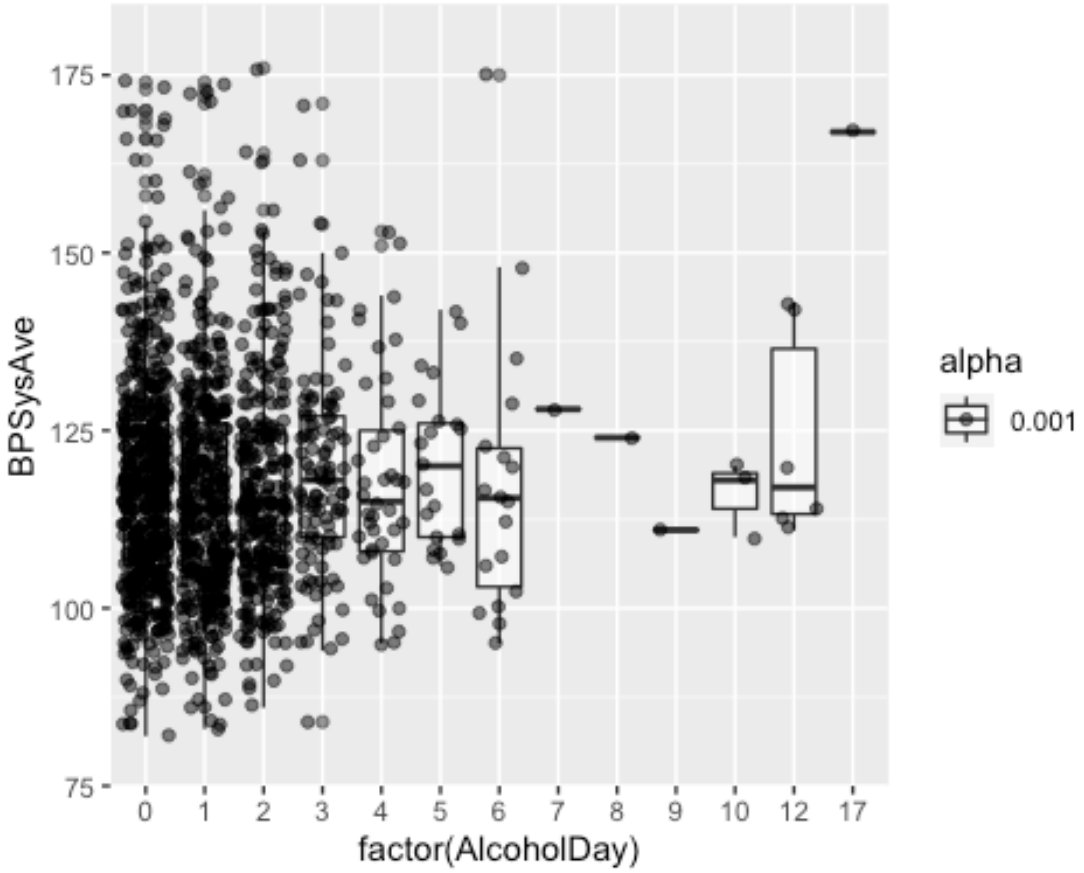
4. PhysActiveDays

Based on the available data, increase Physical Activities does not have much effect on Blood Presure.



5. AlcoholDay

In general, the lower the number of alcohol taken (1-2 portions per day), the lower the blood pressure.



The AI Algorithms

From the analysis, Age, BMI, SleepHrsNight, AlcoholDay are found to affect blood pressure.
The relationship between some of these features may not in a linear relationship with blood pressure.
But we would still try to predict blood pressure by Multiple Regression, as a baseline approach.
Besides the Multiple Regression, Keras Artificial Neural Network will also be used.

Next we try to predict by

(1) Multiple Regression (a baseline approach)
(2) Keras ANN

(1) Multiple Regression

Multiple regression generally explains the relationship between multiple independent or predictor variables and one dependent or criterion variable. … The multiple regression equation explained above takes the following form:
y = b1 x1 + b2 x2 + … + bn xn + c.
In our case:
y = b1 x Age + b2 x BMI + ... + b5 x AlcoholDay + c

Multiple Linear Regression

The purpose of a multiple linear regression is to:

1.Determine the size and nature of the coefficient for each feature in explaining the dependent variable.
2.Determine the signficance or insignificance of each feature.

Female Age 30-65 without NA

```
> Number of Female Aged 30-65 : 1355

> A Quick Look of the Database

> # A tibble: 6 x 6
>     Age   BMI SleepHrsNight PhysActiveDays AlcoholDay BPSysAve
>   <dbl> <dbl>         <dbl>          <dbl>      <dbl>    <dbl>
> 1    49  30.6             8              0          2      112
> 2    45  27.2             8              5          3      118
> 3    58  26.2             5              2          0      127
> 4    56  19.7             7              7          1       95
> 5    57  20.7             8              3          1      122
> 6    64  27.2             5              4          0      130
```

Split the Train_Set and Test_Set by 90:10

The split of train or test sets are usually 60:40, 70:30, 80:20, 90:10.
The choice are mostly depends on the size of the dataset, nature of application, etc.
We start with the most typical split of 70:30. But with several trials, we found that 90:10 provides the best performance, it is mostly because our relatively smaller dataset size (~1300), and NN is considered as a data intensive model.

```
> Female Aged 30-65
> -------
> Total Size of Dataset 1355

> Size of Train Set 1218

> Size of Test Set 137
```

Get the average value of Blood Pressure as a reference

```
avg <- mean(train_set$BPSysAve)

> The average Blood Pressure from train set (ie. our guessing) is 117.6 mmHg.
```

To compare performance of different models, the Root Mean Square Error (RMSE) is used, as it is more sensitive to large error.

When compared with Test Set, the RMSE of guessing with Average is computed:

```
RMSE_avg <- sqrt(mean((avg - test_set$BPSysAve)^2))

diff <- avg - test_set$BPSysAve

>
> RMSE is 15.3 if Guessing with average Blood Pressure.
```

Summary of RMSE

Algorithm	Root Mean Square Error (RMSE)
Guess with Average	15.30

Let's see if the Multiple Regression perform better than guessing with average:

Multiple Regression

```
>
> Call:
> lm(formula = BPSysAve ~ Age + BMI + SleepHrsNight + PhysActiveDays +
>     AlcoholDay, data = train_set)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -42.674  -9.120  -0.796   7.176 103.481
>
> Coefficients:
>                Estimate Std. Error t value Pr(>|t|)
> (Intercept)    75.90219    3.44306  22.045   <2e-16 ***
> Age             0.58020    0.04123  14.071   <2e-16 ***
> BMI             0.45657    0.05468   8.350   <2e-16 ***
> SleepHrsNight   0.02623    0.31142   0.084   0.9329
> PhysActiveDays -0.03110    0.19083  -0.163   0.8706
> AlcoholDay      0.64394    0.25681   2.507   0.0123 *
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 14.74 on 1212 degrees of freedom
> Multiple R-squared:  0.1882,  Adjusted R-squared:  0.1848
> F-statistic: 56.19 on 5 and 1212 DF,  p-value: < 2.2e-16
```

Intercept is 75.90219.
From the Estimate, the larger the number (Coefficient) of the more significants of the features.
In the above summary, 1 year older means 0.58 mmHg blood pressure higher.
1 Alcohol portion means 0.64 mmHg blood pressure higher.

```
>    (Intercept)            Age            BMI  SleepHrsNight PhysActiveDays
>    75.90218654     0.58020111     0.45657395     0.02623102    -0.03109749
>     AlcoholDay
>     0.64394086

>
> RMSE of Multiple Regression (with Age, BMI, SleepHrsNight, PhysActiveDays, AlcoholDay)
is:  14.78
```

RMSE has been improved a bit.

Summary of RMSE

Algorithms	Root Mean Square Error (RMSE)
Guess with Average	15.30
Multiple Regression (with Age, BMI, SleepHrsNight, PhysActiveDays, AlcoholDay)	14.78

Let's omit the most insignificant feature among those five : PhysActiveDays

Multiple Regression (2)

```
>
> Call:
> lm(formula = BPSysAve ~ Age + BMI + SleepHrsNight + AlcoholDay,
>     data = train_set)
>
> Residuals:
>     Min      1Q  Median      3Q     Max
> -42.622  -9.115  -0.803   7.173 103.525
>
> Coefficients:
>               Estimate Std. Error t value Pr(>|t|)
> (Intercept)   75.84859    3.42594  22.139   <2e-16 ***
> Age            0.58013    0.04121  14.076   <2e-16 ***
> BMI            0.45737    0.05444   8.401   <2e-16 ***
> SleepHrsNight  0.02316    0.31072   0.075   0.9406
> AlcoholDay     0.64399    0.25671   2.509   0.0123 *
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 14.73 on 1213 degrees of freedom
> Multiple R-squared:  0.1882,  Adjusted R-squared:  0.1855
> F-statistic: 70.28 on 4 and 1213 DF,  p-value: < 2.2e-16
```

Let's predict the Blood Pressure, and Work out the RMSE

```
>
> RMSE of Multiple Regression (with Age, BMI, SleepHrsNight, AlcoholDay) is:  14.76
```

RMSE after excluding the least significant feature for multiple regression is just very similar.

Summary of RMSE

Algorithms	Root Mean Square Error (RMSE)
Guess with Average	15.30
Multiple Regression (with Age, BMI, SleepHrsNight, PhysActiveDays, AlcoholDay)	14.78
Multiple Regression (with Age, BMI, SleepHrsNight AlcoholDay)	14.76

Next, we try Keras Artificial Neuron Network (Keras ANN)

Taking the findings of both the correlation plots and multiple linear regression into account, Age, BMI, SleepHrsNight, AlcoholDay are kept as the relevant features for the analysis.

(2) Keras ANN

Data Preparation

```
> [1] 1355    5
```

Neural network are very sensitive to non-normalized data, we need to normalize

Max-Min Normalization

The normalized train_set and test_set (90:10)

```
> train_set :

> X_train :

>            Age        BMI SleepHrsNight AlcoholDay
> [1,] 0.5428571 0.23478786           0.6 0.11764706
> [2,] 0.4285714 0.18450853           0.6 0.17647059
> [3,] 0.8000000 0.16910766           0.3 0.00000000
> [4,] 0.7428571 0.07111581           0.5 0.05882353
> [5,] 0.7714286 0.08515778           0.6 0.05882353
> [6,] 0.9714286 0.18360260           0.3 0.00000000

> [1] 1205    4

>
> y_train :

> [1] 0.2297297 0.2702703 0.3310811 0.1148649 0.2972973 0.3513514
```

```
> test_set :

> X_test :

>            Age        BMI SleepHrsNight AlcoholDay
> [1,] 0.1428571 0.08108108           0.6 0.05882353
> [2,] 0.7428571 0.21636721           0.5 0.00000000
> [3,] 0.8857143 0.38940057           0.2 0.05882353
> [4,] 0.2571429 0.06628416           0.4 0.05882353
> [5,] 0.2285714 0.13271931           0.5 0.05882353
> [6,] 0.1714286 0.09165031           0.4 0.00000000

> [1] 150   4

>
> y_test :

> [1] 0.1891892 0.4054054 0.3716216 0.1418919 0.1554054 0.1418919
```

We use a Sequential Model

Initialize a sequential model: The first step is to initialize a sequential model with keras_model_sequential(), which is the beginning of our Keras model. The sequential model is composed of a linear stack of layers.

Apply layers to the sequential model: Layers consist of the input layer, hidden layers and an output layer.

The input layer is for data to formated correctly. The hidden layers and output layers are what controls the ANN inner workings.

Hidden Layer(s): Hidden layer form the neural network nodes that enable non-linear activation using weights. We'll add one hidden layer. The hidden layer is created using layer_dense(). We would apply units = 32 (estimated as below), which is the number of nodes. We'll select kernel_initializer = "RandomNormal" and activation = "relu" for hidden layers.

The number of hidden layers, units, kernel initializers and activation functions are parameters can be optimized for the best results.

Input layer: Number of Features + 1

Hidden layer : Training Data Samples/(Factor * (Input Neurons + Output Neurons))

Output layer: 1

A factor of Scaling Factor is set in this case, the purpose of the factoris to prevent overfitting.

A factor can take a value between 5 and 10. With 5 neurons in the input layer, 1 neuron in the output layer and ~1000 entries in the training set:

The hidden layer is assigned 1200/((5 TO 10)*(5+1)) = (20-40) neurons.
We would choose between 20-40 nodes. With several trials, 32 is the most performing

Dropout in each layer is used to avoid overfitting.

```
model <- keras_model_sequential()

model %>%
    layer_dense(units = 5, activation = 'relu', kernel_initializer='RandomNormal', input_shape = c(4)) %>%
    layer_dropout(0.1) %>%
    layer_dense(units = 32, activation = 'relu', kernel_initializer='RandomNormal') %>%
    layer_dropout(0.1) %>%
    layer_dense(units = 1, activation = 'linear', kernel_initializer='RandomNormal')
```

```
> Model: "sequential"
> _________________________________________________________________________
> Layer (type)                 Output Shape              Param #
> =========================================================================
> dense_2 (Dense)              (None, 5)                 25
> _________________________________________________________________________
> dropout_1 (Dropout)          (None, 5)                 0
> _________________________________________________________________________
> dense_1 (Dense)              (None, 32)                192
> _________________________________________________________________________
> dropout (Dropout)            (None, 32)                0
> _________________________________________________________________________
> dense (Dense)                (None, 1)                 33
> =========================================================================
> Total params: 250
> Trainable params: 250
> Non-trainable params: 0
> _________________________________________________________________________
```

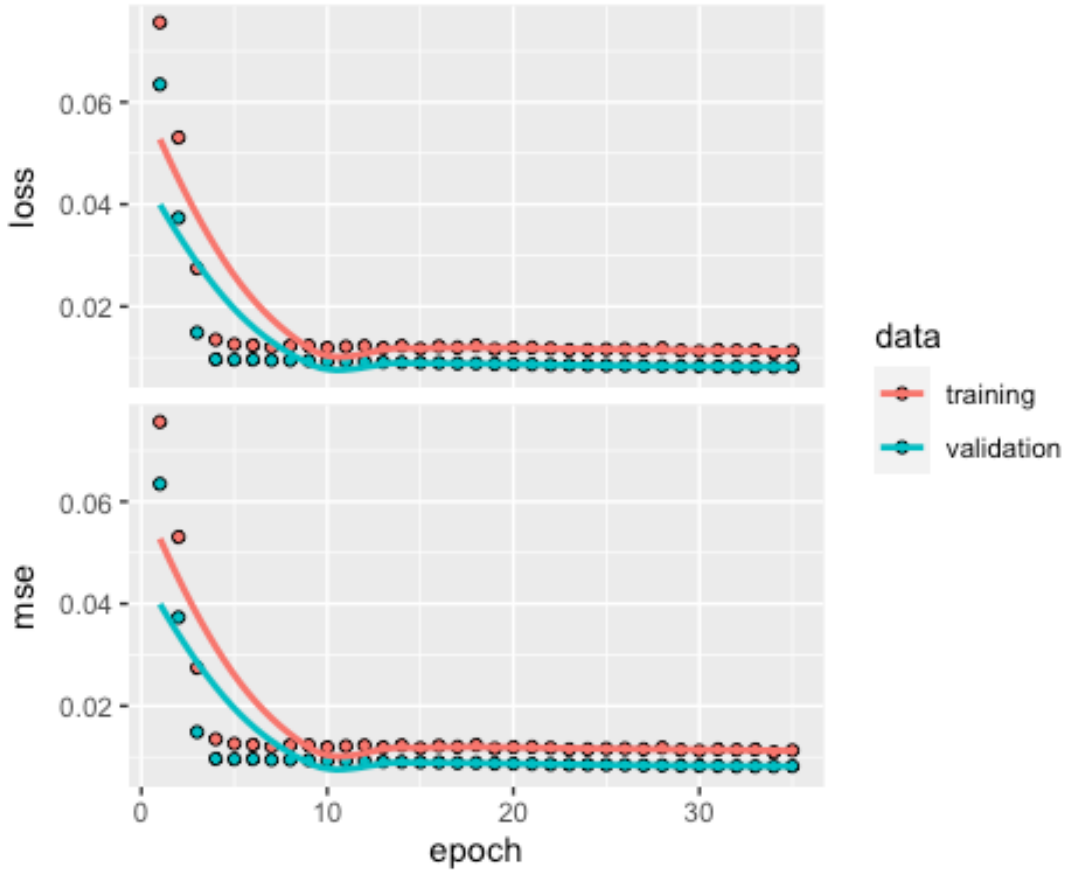
Optimizer: Adam realizes the benefits of both AdaGrad and RMSProp.
As from the following reference, Instead of adapting the parameter learning rates based on the average first moment (the mean) as in RMSProp, Adam also makes use of the average of the second moments of the gradients (the uncentered variance).
https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/
RMSE is used because it is suitable for target value having outliners. Optimizing MSE is same as optimizing RMSE.

```
model %>% compile(
    loss = 'mean_squared_error',
    optimizer = 'adam',
    metrics = c('mse')
)
```

We use the fit() function to run the ANN on our training data. The object is our model, and x and y are our training data in matrix and numeric vector forms, respectively. The batch_size = 50 sets the number samples per gradient update within each epoch. We set epochs = 35 to control the number training cycles. Typically we want to keep the batch size high since this decreases the error within each training cycle (epoch).

We can see that the model converged reasonably quickly and both train and test performance remained equivalent (in around episode 7). The performance and convergence behavior of the model suggest that mean squared error is a good match for a neural network used this problem.

We put validation_split = 0.2 to include 20% of the data for model validation, due to the small dataset size.



calculate the RMSE

```
model %>% evaluate(X_test, y_test)
pred <- data.frame(y = predict(model, as.matrix(X_test)))

predicted = pred$y * abs(diff(range(female3065_nona$BPSysAve))) + min(female3065_nona$BPSysAve)
actual = y_test * abs(diff(range(female3065_nona$BPSysAve))) + min(female3065_nona$BPSysAve)

> [1] 13.52028

> RMSE of Keras ANN (with Age, BMI, SleepHrsNight, AlcoholDay) is:  13.52
```

We have tried to add one more layers and change the number of node, batch size, the above parameters seems being optimized.

Summary of RMSE

Algorithms	Root Mean Square Error (RMSE)
Guess with Average	15.30
Multiple Regression (with Age, BMI, SleepHrsNight, PhysActiveDays, AlcoholDay)	14.78
Multiple Regression (with Age, BMI, SleepHrsNight AlcoholDay)	14.76
Keras Artificial Neuron Network (with Age, BMI, SleepHrsNight, AlcoholDay)	13.52

Conclusion

Different algorithms for Blood Pressure Prediction have been worked out and compared.

Keras ANN is found to be the most performing algorithm having the least RMSE 13.52.

Further Improvement Approach

- Excluding outliners – such as seriously underweight, obese person, or person on medication, etc. This is because their blood pressure might behave very abnormal. The above program is re-run by filtering seriously underweight and obese person:
  - Get Rid of Some Obvious overweight or underweight.
  - Add a filter : filter(between(BMI,18,40)) %>% to the data set
  - Decrease the number of node in hidden layers due to smaller dataset by excluding extreme BMI users

```
>
> RMSE of Keras ANN (with Age, BMI, SleepHrsNight, AlcoholDay) (excluding extreme BMI) is:
12.42
```

Algorithms	Root Mean Square Error (RMSE)
Guess with Average	15.30
Multiple Regression (with Age, BMI, SleepHrsNight, PhysActiveDays, AlcoholDay)	14.78
Multiple Regression (with Age, BMI, SleepHrsNight AlcoholDay)	14.76
Keras Artificial Neuron Network (with Age, BMI, SleepHrsNight, AlcoholDay)	13.52
Keras Artificial Neuron Network 2 (with Age, BMI, SleepHrsNight, AlcoholDay) (exclude extreme BMI)	12.42

- Keras ANN is now improved with the RMSE of 12.42. ~12mmHg different is somehow be reasonable for blood pressure prediction. A person may have ~10mmHg different from day time to night time.

There could also be other approach to improve the algorithms:

- Principle Component Analysis (PCA) – there are still other useful features on the dataset (> 70 features) that have not yet been taken in accounts. Including more useful features would improve the prediction, but also increase the complexity of the machine learning or deep learning. PCA is a common approaches to handle high dimension dataset, and should be a way to improve the RMSE performance.
- Collecting larger dataset – needless to say this help to generalize further the algorithms.
- Group user into finer Age groups – Right now we use 30 to 65 years old for a easy analysis. However, factors affecting blood pressure of ~30 years old female could be very differently with a ~50 years old female.

Final Words:

Again, our aim is to predict Blood Pressure with HANDY information, such as Age, BMI, SleepHrsNight, AlcoholDay, so a person get know their blood pressure instantly without using a device or without going to a clinic.

Blood Pressure is an useful indicator of a person health. People with comparatively high blood pressure (> 120 mmHg) should start to adopt a health lifestyle and pay attention to their health before getting into chronic illness.

Measuring Blood Pressure seems to be easy by using home-use gauge nowsadays, but algorithms can be developed similary way to predict Fasting Blood Glucose, Cholesterol, Uric Acide, Triglycerides, etc. They are all user indicators for predicting health of a user. It is a good way to reduce population's chronic illnesses.