Name: Kimanh Pham (38936)

Phuong Lieu (Fen) (38626)

Assignment 1
Data mining and Text mining Course

Q1.a.

There are 9 variables with total 768 observations.

The observed variables are pregrancies, plasma.glucose, blood.pressure, triceps.skin.thickness, insulin, BMI, diabetes.pedigree, age and diabetes.

Q1.b.

There are 5 missing values in the diabetes variable which are in the row 44, 211, 370,386 and 622.

```
diabetes.pedigree age diabetes
44              0.704  34       NA
211             0.150  29       NA
370             0.158  25       NA
389             0.311  35       NA
622             0.388  22       NA
```

There are 218 people with diabetes have BMI of 30 or higher in the dataset.

Q2.

Normal sugar level of healthy person is 100mg/dL. Sugar levels higher than normal mean either diabetes or pre-diabetes is present. In this case, the average plasma glucose level is about 120mg/dL (mean=120.8). Moreover, according to distribution chart, there are more people with plasma.glucose level above 100mg/dL than normal people. However, the ratio of person who has diabetes in the dataset is only about 35%. Therefore, the data is not so realistic for plasma.glucose variable.

Other than that, the dataset seems pretty reliable with normal distribution regarding to variables meaning.

Q3.a.

According to the observations, age of people is ranged from 21 to 81 years old.

Q3.b.

Diabetes		Diabetes (Ratio)	
0	1	0	1
497	266	0.6513761	0.3486239

There are 266 subjects have diabetes compared with 497 subjects who are healthy with respectively ration 35% and 65%.

Q3.c.

Mean, median, standard deviation and quantiles are parameters used to describe a standard normal distribution of blood pressure.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   62.00   72.00   69.11   80.00  122.00 
```

sd(blood.pressure)

```
[1] 19.35581
```

Q3.d.

```
0%     25%     50%     75%    100% 
 0.000 27.375 32.200 36.500 66.700 
```

There are 75% of the subjects have a BMI higher than 27.38.

Q3.e.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  21.00   24.00   29.00   33.24   41.00   81.00 
```

Mean is the average age of subject in the dataset which is 33.24. They are different because median is the central point of value of a frequency distribution of observed values in the dataset while mean is the average value.

Q4.a.

As be seen by the scatter plots, insulin is the most strongly correlate with plasma glucose. The scatter plot between plasma glucose variable and insulin shows a positive linear relationship (uphill pattern). It means that the level of insulin affects to the level of plasma glucose of subjects.

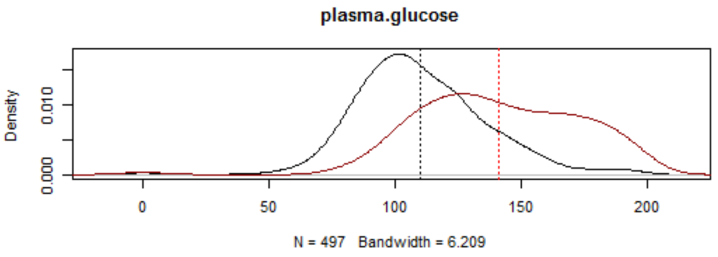
The scatterplot between age and blood pressure doesn't show much of anything happening. Therefore, its correlation is not very strong. They don't have much effect to each other.

Q4.b

As can be seen in the class-wise density plot, plasma glucose, blood pressure and BMI show remarkable pattern which indicates the distribution between non-diabetes and diabetes.

Density line represents the density of healthy people (normal color) or people with diabetes (darker line). Vertical line represents the average number of healthy people (black line) and with diabetes (red line). The black vertical line stands closely to the peak of density line of healthy people, is reasonable.

For example, in plasma glucose, the highest amount of diabetes people (the peak of darker density line) stands close to the average of people with diabetes (red vertical line) and respectively with two lines of healthy people.



Q5.a.

A strong relationship will have coefficient value from 0.5 to 1.

In the correlation analysis, the coefficient of correlation between plasma glucose and insulin is 0.33135711, which is the highest coefficient compared to others. Moreover, it is a positive value which explains positive relationship. When level of plasma glucose goes higher, the insulin level also goes higher relatively.

The coefficient of correlation between age and blood pressure is 0.23952795. Therefore, it doesn’t indicate any clear relationship. Age of observation does not affect to the blood pressure.

Q5.b.

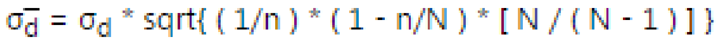
Diabetes pedigree variable is the one shows least association with age with coefficient 0.03362853. This result is reasonable because a person can have diabetes caused by pedigree regardless of age.

Q6.

The t-value is related to the size of the difference between the means of the two samples. The larger t is, the larger the difference.

Based on t- value, there are two variables exhibits the largest differences are plasma glucose (t-value -13.5653) and BMI (t-value -8.5966)

To calculate differences in terms of standard deviations, we use this



$$\sigma_{\bar{d}} = \sigma_d * \mathrm{sqrt}\{ (1/n) * (1 - n/N) * [N / (N - 1)] \}$$

with σ_d is the standard deviation of the population difference, *N* is the population size, and *n* is the size of observation with diabetes

Mean differences in plasma glucose with the presence of diabetes:

```
mean in group 0 mean in group 1 
       110.2334        141.0865 
```

In term of standard deviation, the difference is 1.080298
In term of the original unit, the difference is 30.8531

Mean differences in BMI with the presence of diabetes:

```
mean in group 0 mean in group 1 
       30.29276        35.14023 
```

In term of standard deviation, the difference is 0.169731
In term of the original unit, the difference is 4.84747