ALY 6000: Introduction to Analytics

FINAL REPORT

Exploratory Data Analysis of females of Pima Indian Heritage

Submitted by: Sakshi Mehta

Submitted To: Prof. Zhi He

Abstract

Diabetes is a chronic and usually a long-term disease that affects how your body turns what you eat into energy. It occurs when there is too much sugar in the blood. It is regulated by the hormone called Insulin which is released when the sugar level in the blood rises. There are multiple factors that can be responsible for the development of diabetes in a person like their genes, family history, race or ethnicity, weight, insulin levels, blood pressure, body mass index, skin thickness, lifestyle, and much more. Diabetes can lead to the failure of multiple organs especially the eyes, nerves, veins, kidneys. Therefore, the objective of this analysis report is to find out the trends and patterns relative to the diagnosis of diabetes.

The dataset used in this analysis is from the National Institute of Diabetes and Digestive and Kidney Diseases. This record is particularly of females who are at least 21 years old and are from Pima Indian Heritage.

The dataset has multiple variables (8 columns and 768 rows) and a final column of outcomes that shows if the person is diabetic or not.

It consists of 9 variables:

-Pregnancies (number of pregnancies a woman had)

-Glucose (Plasma glucose concentration)

-BloodPressure (Diastolic blood pressure in mm Hg)

-SkinThickness (Tricep skin thickness in mm)

-Insulin (serum insulin in mu U/ml)

-BMI (Body Mass Index (weight in kg/(height in m)^2))

-DiabetesPedigreeFunction (likelihood of diabetes based on family history)

-Age (in years)

-Outcome: (0 – diabetic, 1-non-diabetic)

(Source of dataset: https://www.kaggle.com/mathchi/diabetes-data-set)

Structure of the dataset – Diabetes.csv (uncleaned and without normalization)

```
> structure(Diabetes)
   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome
1            6     148            72            35       0 33.6                    0.627  50       1
2            1      85            66            29       0 26.6                    0.351  31       0
3            8     183            64             0       0 23.3                    0.672  32       1
4            1      89            66            23      94 28.1                    0.167  21       0
5            0     137            40            35     168 43.1                    2.288  33       1
6            5     116            74             0       0 25.6                    0.201  30       0
7            3      78            50            32      88 31.0                    0.248  26       1
8           10     115             0             0       0 35.3                    0.134  29       0
9            2     197            70            45     543 30.5                    0.158  53       1
10           8     125            96             0       0  0.0                    0.232  54       1
```

Maximum, minimum and average (mean and median) age of women whose details have been recorded:

```
> round(mean(Diabetes$Age))
[1] 33
> median(Diabetes$Age)
[1] 29
> min(Diabetes$Age)
[1] 21
> max(Diabetes$Age)
[1] 81
```

The survey has been done on women with an average age of 29 and the details clearly indicate that the range of the age is (81 - 21 = 61).

Dataset after cleaning

Console | Terminal | Jobs

R 3.6.3 · ~/

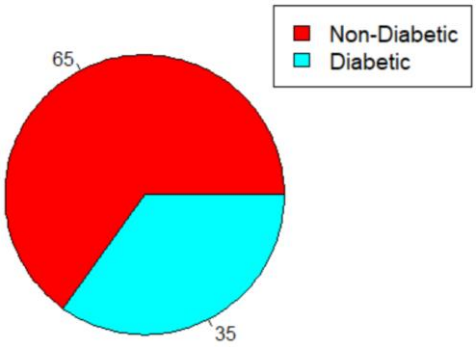
```
> Diabetes
    Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome
1             6     148            72            35       0 33.6                    0.627  50       1
2             1      85            66            29       0 26.6                    0.351  31       0
3             8     183            64            21       0 23.3                    0.672  32       1
4             1      89            66            23      94 28.1                    0.167  21       0
5             0     137            40            35     168 43.1                    2.288  33       1
6             5     116            74            21       0 25.6                    0.201  30       0
7             3      78            50            32      88 31.0                    0.248  26       1
8            10     115             0            21       0 35.3                    0.134  29       0
9             2     197            70            45     543 30.5                    0.158  53       1
10            8     125            96            21       0  0.0                    0.232  54       1
```

The most unusual thing that the original dataset had was that the SkinThickness is printed as zero which is not possible as in a literal sense that means there is no skin layer at the triceps. Therefore, to make the wrong values or uncleaned dataset clear and meaningful the missing values were replaced with the mean value of the entire column of skin thickness. The cleaned dataset above has replaced all the values that are equal to 0 to the mean value of all the records by using the if-else condition check first and then if the condition of missing values is satisfied, it is then replaced with the average value.

Overall count of females

```
>   ifDiabetes
  x freq
1 0  500
2 1  268
>   tot<- sum(ifDiabetes$freq)
>   tot
[1] 768
```

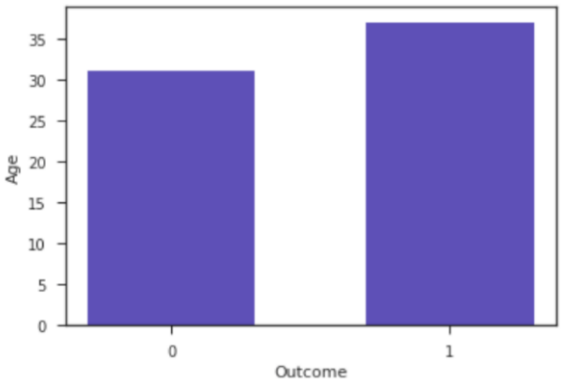
Out of the total of 768 females, 500 do not have diabetes and 268 of them have diabetes. This is the descriptive methodology to define it.



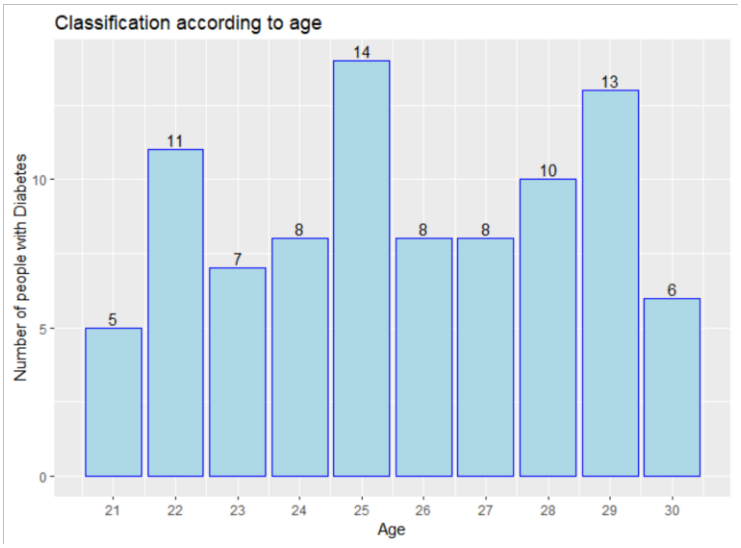
Pie chart for the distribution of people with diabetes and no diabetes.

The pie chart on the right shows that out of 100%, 65% that is shown with red color are non-diabetic and the rest 35% with blue color have been diagnosed with diabetes.

The bar plot on the left shows the average age of non-diabetic and diabetic people in the "Diabetes.csv" dataset. It is evident from the plot that the mean age of people having diabetes is more than the mean age of people with no diabetes.

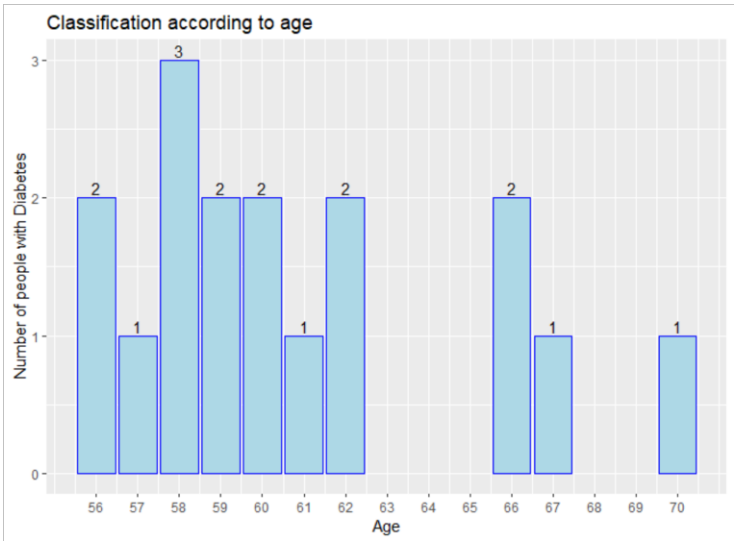


Number of females diagnosed with diabetes from the age group (21-30)



The above bar plot shows the number of people diagnosed with diabetes from the age group of 21 to 30. We can see that 25 is the age where the maximum number of people have diabetes and that is 14. Also, it is noticeable that the minimum count of diabetes is seen at the age of 21 and that is 5.

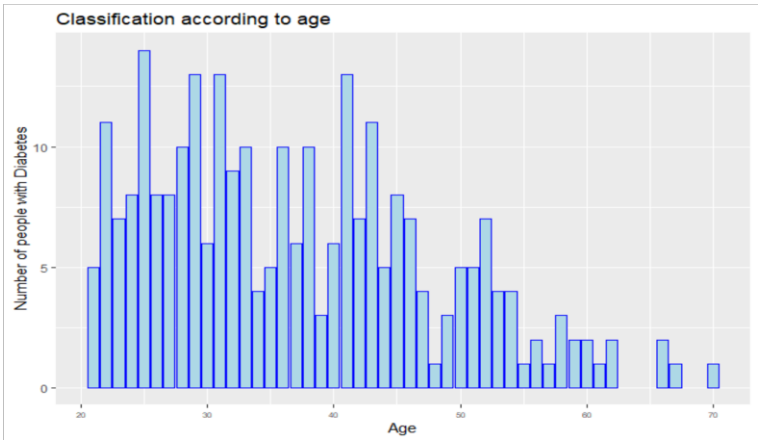
Number of females diagnosed with diabetes from the age group (56-70)



The above bar plot shows the number of people diagnosed with diabetes from the age group of 56 to 70. We can see that 58 is the age where the maximum number of people have diabetes and that is 3. Also, it is noticeable that the minimum count of diabetes is 1. Also, the sample dataset does not have anyone with the age 63 to 65 and 68-69. Therefore, there is a possibility that the data might be a little biased as the sample is not complete.

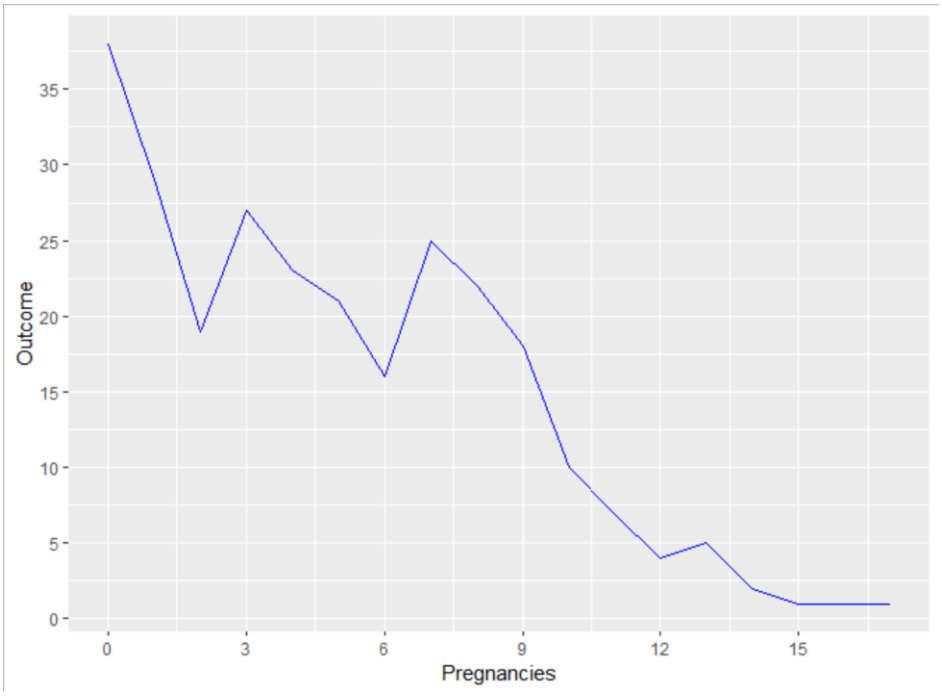
From the above 2 graphs, one thing that is comprehensible is that after the age of 58, the occurrence of diabetes has reduced to a great extent. People of the age between (58-70) are barely facing abrupt sugar levels.

Overall plotting of number of people diagnosed with diabetes from the age group (21-70)



This is the overall representation of all the ages that the sample set has with the mapping of the frequency of diabetic people in that age. The plotting seems quite irregular however the overall trend shows that the diagnosis reduces after the age of 45 and the age of 25 remains to be critical. Therefore 25 is the age of maximum concern in relation to this.

pregnancy VS outcome line graph



The line graph above shows that the women with the least that is in our case zero pregnancies have had the highest diabetes record. The line graph above slowly moves towards the bottom right which means that the higher the pregnancy countless are the ladies prone to diabetes. This helps us to get towards a conclusion that otherwise is difficult to predict. This is a strange relation and opens a topic of research. There has been no such conclusion made until now regarding the matter that states the number of pregnancies and its relationship with the occurrence of diabetes in women.

Addition of a column – new_glucose

	Diabetes.Glucose	Diabetes.new_glucose
1	148	74.37186
2	85	42.71357
3	183	91.95980
4	89	44.72362
5	137	68.84422
6	116	58.29146
7	78	39.19598
8	115	57.78894
9	197	98.99497
10	125	62.81407

All the glucose values are made relative to one maximum value and that is 199.

Conclusion

To sum up the analysis that is done above, it can be concluded that huge datasets can be plotted, and certain trends and correlations can be made from them which otherwise becomes impossible to read. There are multiple factors and patterns that can be seen in the women of Pima Indian heritage. This also opens up a question of further research.

Scope of improvement

- Larger samples can be used to get a more precise and realistic conclusion.
- Better cleaning and manipulating can be done by using normalization techniques best suitable for any particular variable.
- More details about the person like race, ethnicity, eating practice, stress level could be used.
- Multiple plotting techniques and better software could be used for better visualization.

References

- https://www.datasciencemadesimple.com/scaling-or-normalizing-the-column-in-r-2/
- https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16#:~:text=DiabetesPedigreeFunction%3A%20Diabetes%20pedigree%20function%20(a,%2Ddiabetic%2C%201%20if%20diabetic)
- https://cmdlinetips.com/2021/04/categorize-multiple-numerical-columns-in-r/
- https://www.statmethods.net/graphs/line.html
- https://community.rstudio.com/t/plot-error-and-size-too-large-very-large/57865/3
- https://www.kaggle.com/mathchi/diabetes-data-set
- https://statisticsglobe.com/replace-missing-values-by-column-mean-in-r
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0175-6
- https://www.geeksforgeeks.org/count-the-frequency-of-a-variable-per-column-in-r-dataframe/