BMI Score and Tricep Skinfold Thickness in Pima Dataset

By Jaclyn Andrews, Farhana Faruqe and Clayton Hirsh

Overview

- Background of Dataset
- Geomap
- Data Exploration
- Background Research
- SMART Question Development
- Epicycles of Analysis (before & after research)
- Methodology
- Correlation
- Tools
- Conclusion

Background of Dataset - Pima.te

Population: Pima Indian women over 21 years old near Phoenix, Arizona

Sample: 332 randomly selected women from population

Source: US National Institute of Diabetes and Digestive and Kidney Diseases

R Package: MASS

The Pima Indian population near Phoenix, AZ has been under study by the NIDDKD since 1965 due to the population's high rate of diabetes.

Geomap

Showing the location, gender and age group of “Pima” dataset using google chart

Initial Variable Exploration (Variable Meanings)

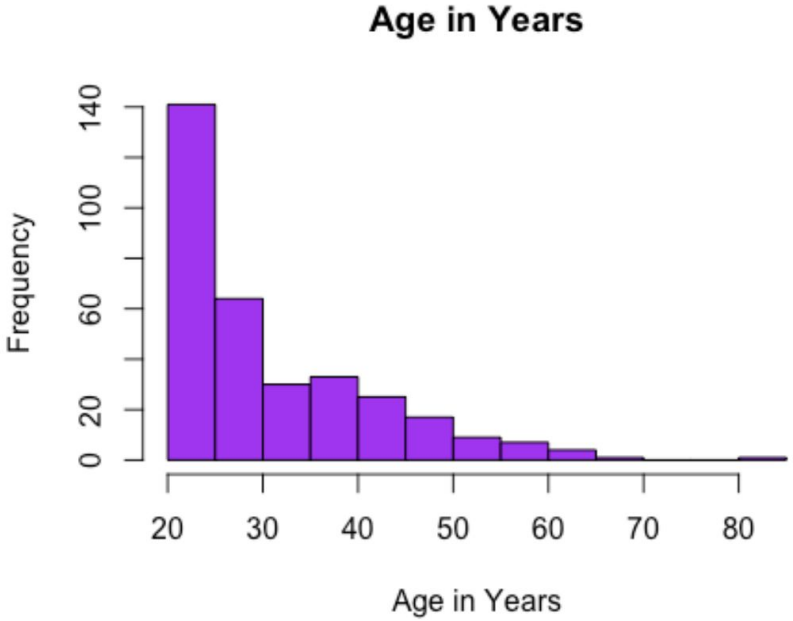
- Npreg: number of pregnancies
- Glu: plasma glucose concentration (blood sugar)
- Bp: diastolic blood pressure (mm Hg)
- Skin: triceps skinfold thickness (mm)
- Bmi: body mass index score
- Ped: diabetes pedigree function (based on genetic history of diabetes)
- Age: age in years
- Type (Yes or No): diabetic by WHO criteria (if the 2 hour post-load glucose was at least 200 mg/dl at any survey exam)

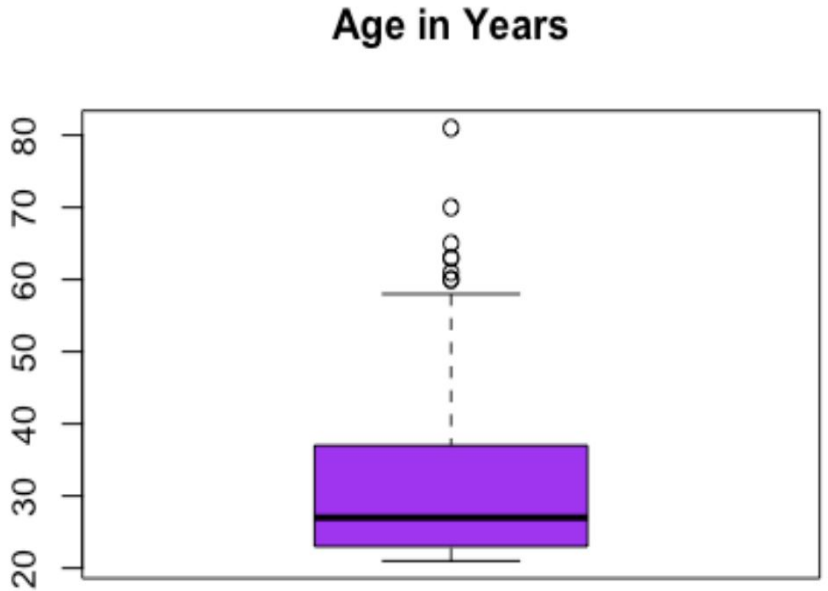
```
## 'data.frame':    332 obs. of  8 variables:
##  $ npreg: int  6 1 1 3 2 5 0 1 3 9 ...
##  $ glu  : int  148 85 89 78 197 166 118 103 126 119 ...
##  $ bp   : int  72 66 66 50 70 72 84 30 88 80 ...
##  $ skin : int  35 29 23 32 45 19 47 38 41 35 ...
##  $ bmi  : num  33.6 26.6 28.1 31 30.5 25.8 45.8 43.3 39.3 29 ...
##  $ ped  : num  0.627 0.351 0.167 0.248 0.158 0.587 0.551 0.183 0.704 0.263
...
##  $ age  : int  50 31 21 26 53 51 31 33 27 29 ...
##  $ type : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 2 1 1 2 ...
```

Initial Variable Exploration (Descriptive Statistics)

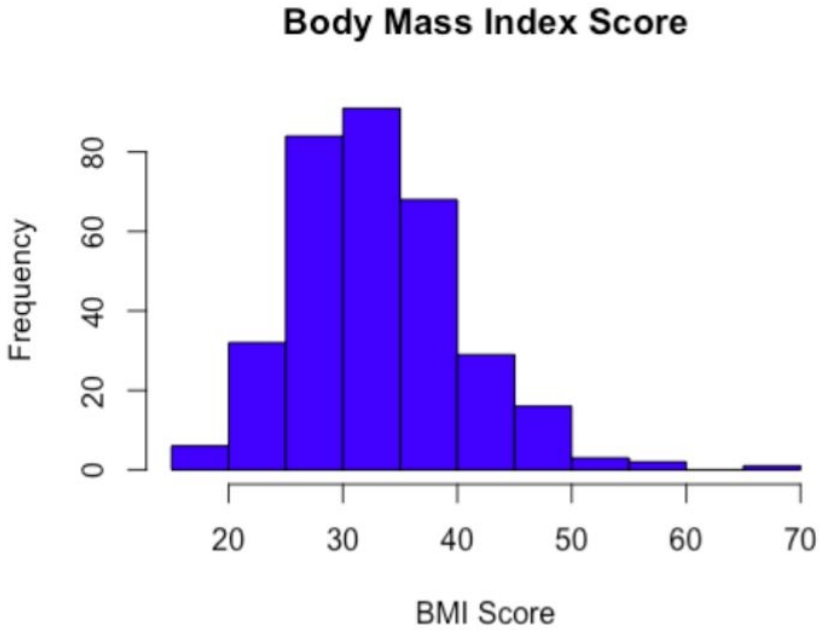
	count	mean	std	min	25%	50%	75%	max
npreg	332.0	3.484940	3.283634	0.000	1.000	2.00	5.00000	17.00
glu	332.0	119.259036	30.501138	65.000	96.000	112.00	136.25000	197.00
bp	332.0	71.653614	12.799307	24.000	64.000	72.00	80.00000	110.00
skin	332.0	29.162651	9.748068	7.000	22.000	29.00	36.00000	63.00
bmi	332.0	33.239759	7.282901	19.400	28.175	32.90	37.20000	67.10
ped	332.0	0.528389	0.363278	0.085	0.266	0.44	0.67925	2.42
age	332.0	31.316265	10.636225	21.000	23.000	27.00	37.00000	81.00
diab	332.0	0.328313	0.470308	0.000	0.000	0.00	1.00000	1.00

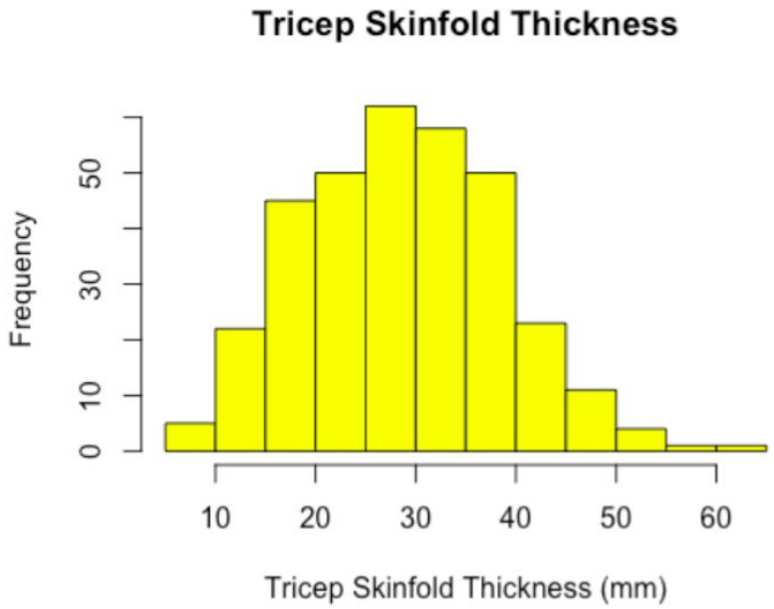
Initial Variable Exploration (Histograms and Boxplots)



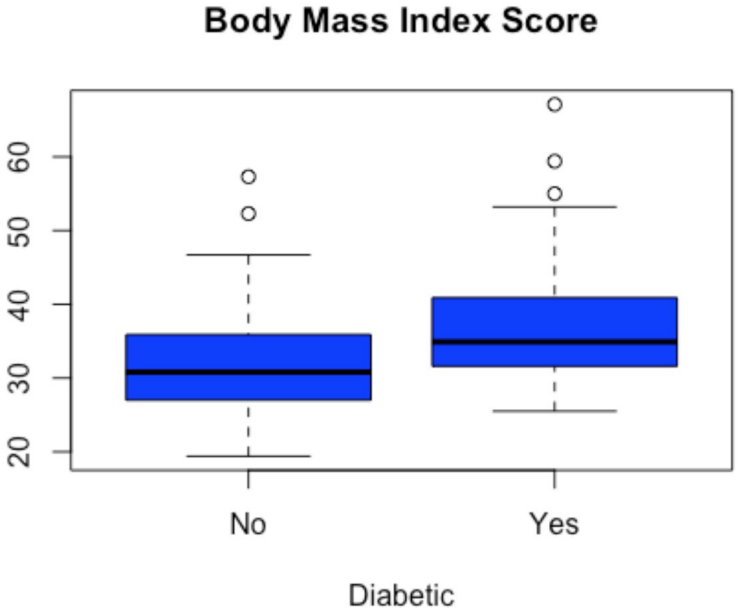


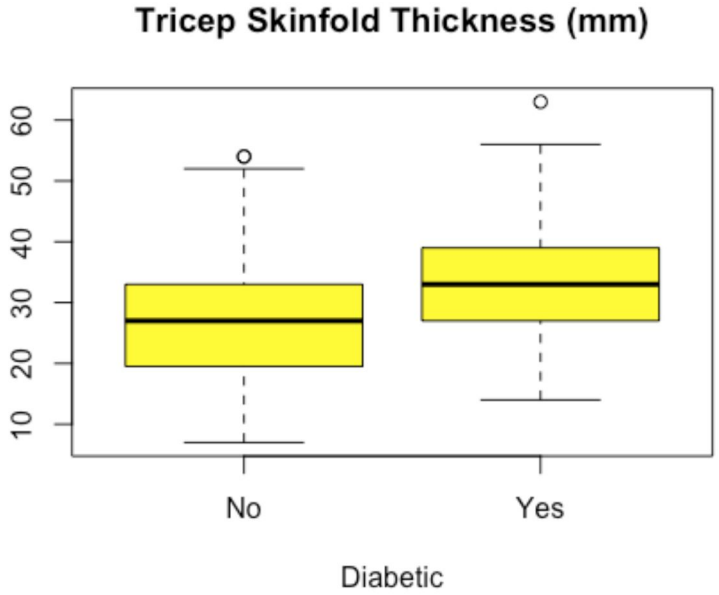
Histograms of BMI and Triceps Skinfold Thickness





BMI and Triceps Skinfold Thickness Exploration: Diabetic vs. Non-diabetic





Background Research

- Body fat a known indicator of diabetes
- Diabetics are more likely to hold excess fat in abdomen

Flaws in Two Variables

BMI Score:

- Doesn't take into account lean tissue
- Overexaggerates thinness in short people and fatness in tall people

Triceps Skinfold Thickness:

- Not meant to be a stand alone measurement
- Should be measured along with other skinfold areas

SMART Question Development

Are BMI score and triceps skinfold thickness strongly correlated (r > 0.7) in the Pima dataset sample population of women?

Specific: about a specific population of women, focused on two variables

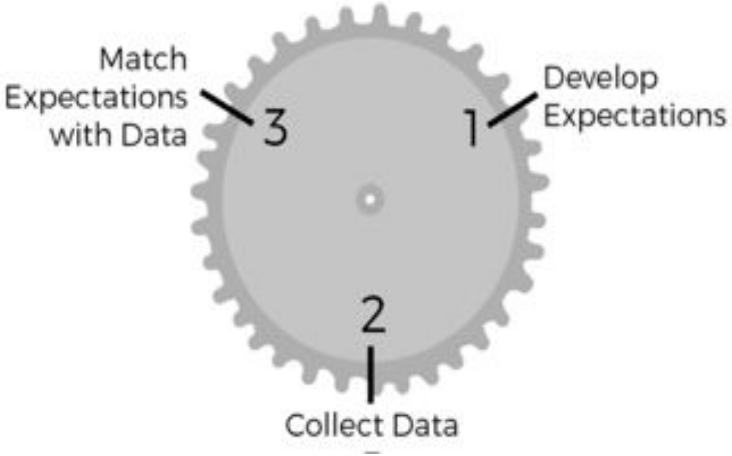
Measurable: can be measured with a correlation coefficient

Answerable: yes if (r > 0.7), no if (r < 0.7)

Relevant: useful for researchers looking at relationship between BMI and triceps skinfold thickness

Time bound: data is already collected, so question can be answered in a timely manner

Epicycles of Analysis



1. **Expectations:** BMI score and triceps skinfold thickness are positively correlated
2. **Data collection:** investigate pima dataset
3. **Match expectations with data:** Yes

Goals of Programming (Easy to update)

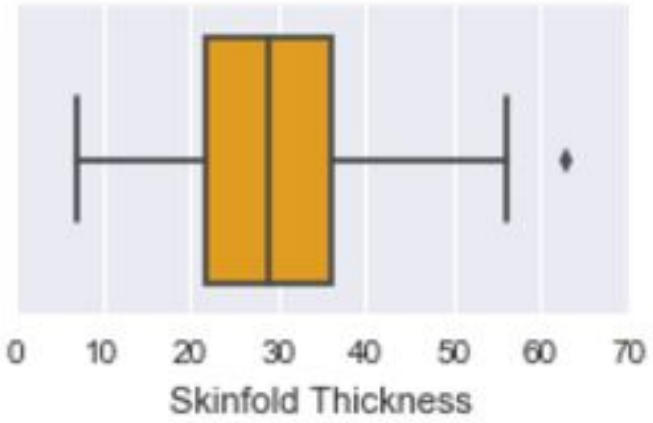
Defining a Function to Draw Histogram & Boxplot

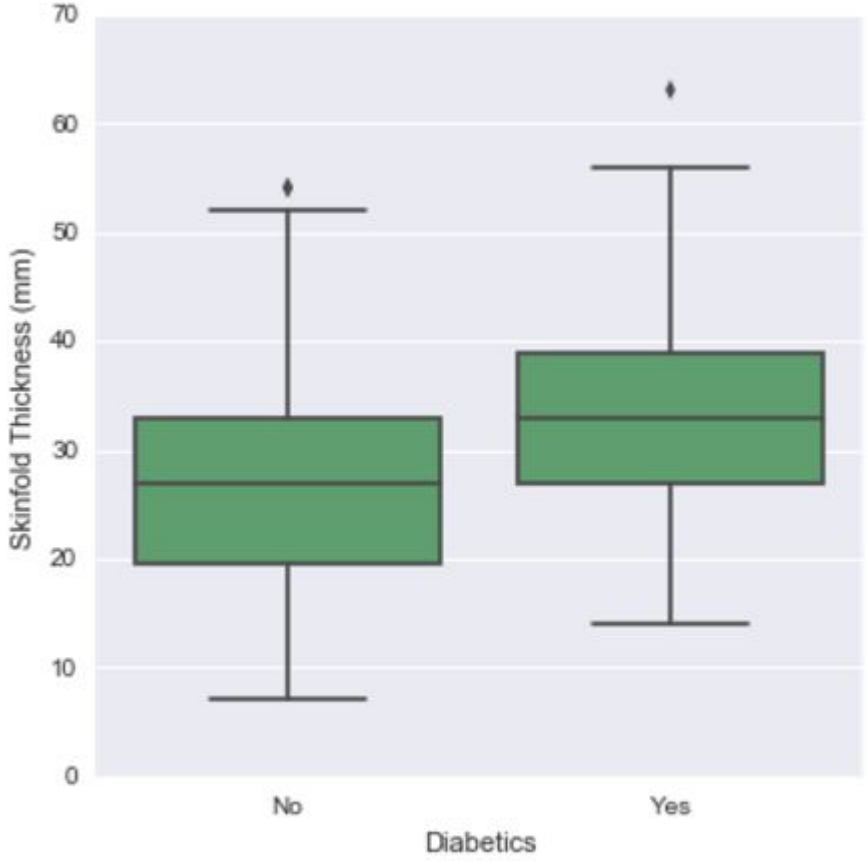
```
def drawhistbox(val,xaxis,col): ### passing data, title and color to draw histogram and boxplot
    hist = sns.FacetGrid(df, col="type")
    hist = hist.map(plt.hist, val, color=col)
    hist.set_titles("Diabetics: {col_name}  ")
    hist.set_axis_labels(xaxis,"")

    box = sns.FacetGrid(df, col="type",col_order=["Yes", "No"],size=2, aspect=1.5)
    box = box.map(sns.boxplot, val, color=col)
    box.set_titles("")
    box.set_axis_labels(xaxis,"")
```

Calling "drawhistbox" funtion to draw histogram & boxplot for five numeric variables

```
drawhistbox("bmi","Body Mass Index","m")
drawhistbox("age","Age","y")
drawhistbox("bp", "Blood Pressure","g")
drawhistbox("skin","Skinfold Thickness","orange")
drawhistbox("npreg", "Number of Pregnancies","b")
```

Example


0
10
20
30
40
50
60
70
Skinfold Thickness


70
60
50
40
30
20
10
0
Skinfold Thickness (mm)
No
Yes
Diabetics

Interpret the Result

```
print("***** Correalation *****")

print("\nBody Mass Index & Triceps Skinfold Thickness")
typeofcorrelation(correlatoion(triceps_skinfold_thickness,bmi))
```

```
***** Correalation *****

Body Mass Index & Triceps Skinfold Thicknes
Moderate Correlation:  0.658542792738
```

Correlation

Defining correlation function & interpret the type Of correlation

```
def correlatoion(var1, var2):
    return(np.corrcoef(var1, var2)[1,0])

def typeofcorrelation(val):
    if val<.3:
        print("Weak Correlation: ", val)
    elif val>=.3 and val<.7:
        print("Moderate Correlation: ", val)
    elif val>=.7:
        print("Strong Correlation: ", val)
```

Correlation

```
***** Correlation *****

Body Mass Index & Triceps Skinfold Thickness
Moderate Correlation:  0.658542792738
```

Goals of Programming (Make code readable)

Drawing Scatterplot Triceps Skinfold Thickness vs BMI with "colorCategorizedBMI" & Plot Function

Getting scatterplot of tricep skinfold thickness by BMI score for all observations, color coding based on BMI weight categories:

#if BMI is less than 25.0 then

    #Normal

#elseif BMI is between 25.0-29.9 then

    #Obese

#elseif BMI is between 30.0-39.9 then

    #Overweight

#elseif BMI is greater than or equal to 40.0 then

    #Extremely obese

Defining a function which will return color based on logic to draw a plot

```
color=[]
def colorCategorizedBMI(): ## assisgn color based on bmi category
    for i in range(len(bmi)):
        if bmi[i]<=25:
            color.append("green")
        elif bmi[i]>25 and bmi[i]<=29:
            color.append("yellow")
        elif bmi[i]>29 and bmi[i]<=39.9:
            color.append("orange")
        elif bmi[i]>=40:
            color.append("red")
    return color
```

Scatterplot of BMI vs Triceps Skinfold Thickness (color coded BMI)

Tools, Packages & Charts

- R
- Python
- Google chart
- jsreportstudio
- Packages:
  - Numpy, pandas, statistics, matplotlib, seaborn, collections,
- Charts:
  - Histogram
  - Boxplot
  - Scatter plot
  - Geomap

Conclusion

SMART Question: Are BMI score and triceps skinfold thickness strongly correlated ($r > 0.7$) in the Pima dataset sample population of women?

BMI and tricep skinfold thickness are not strongly correlated ($r = 0.66$). We can come to the conclusion that one or both methods (BMI and Tricep Skinfold Thickness) are not strong indicators for body fat.

Questions

References

(2012, January 7). Setting the color for an individual data point. Retrieved October 9, 2016, from http://stackoverflow.com/questions/8774002/setting-the-color-for-an-individual-data-point

BMI Formula. (n.d.). Retrieved October 10, 2016, from http://www.bmi-calculator.net/bmi-formula.php

Brownlee, J. (2014, March 29). Case Study: Predicting the Onset of Diabetes Within Five Years (part 1 of 3) - Machine Learning Mastery. Retrieved October 9, 2016, from http://machinelearningmastery.com/case-study-predicting-the-onset-of-diabetes-within-five-years-part-1-of-3/

Finding the mode of a list. (2012, May 29). Retrieved October 9, 2016, from http://stackoverflow.com/questions/10797819/finding-the-mode-of-a-list

Nordqvist, C. (2013, August 25). Obesity / Weight Loss / Fitness Public Health Why BMI is inaccurate and misleading. Retrieved October 10, 2016, from http://www.medicalnewstoday.com/articles/265215.php

R Mean, Median and Mode. Retrieved October 9, 2016, from https://www.tutorialspoint.com/r/r_mean_median_mode.htm

Rogers, J. Z., & Still, C. D. (n.d.). Obesity and Type 2 Diabetes. Retrieved October 9, 2016, from http://www.obesityaction.org/educational-resources/resource-articles-2/obesity-related-diseases/obesity-and-type-2-diabetes

Waskom, M. (n.d.). Seaborn.lmplot. Retrieved October 9, 2016, from https://stanford.edu/~mwaskom/software/seaborn/generated/seaborn.lmplot.ht

Wood, R. (2008). Skinfold Measurement. Retrieved October 9, 2016, from http://www.topendsports.com/testing/tests/skinfolds.htm