Group number: 18

Group name: Boring Mushroom

First and Last Name of each student in the group: M. Chai, Y. Lyu

1. What is the general domain/subject area of this project?

   This project focuses on diabetes prediction through diagnostic measurements provided within it.

2. What data will you use, and what is the source?

   The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. It was posted publicly on Kaggle by UCI Machine Learning. The dataset contains several medical predictor variables and one response variable.

   Dataset: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

3. What primary questions will you seek to answer?

   - How do pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, and age impact diabetes risks?
   - Which model can predict diabetes using the most effective factors for diabetes prediction?

4. What secondary questions will you seek to answer?

   Since the variables are diagnostic measurements, it is likely that they are correlated with each other. Thus, we need to answer:
   - What's the association in each medical predictor?
   - Are there confounding and effect modifications among the variables? (e.g. Is blood pressure a confounder of BMI and diabetes? Is there effect modification between age and pregnancies?)

5. What outcome(s)/endpoint(s) will you use? (could be continuous, binary, polytomous, Poisson, survival,...and you may be considering more than one--and this may be updated/added to, as the semester progresses)

We plan to use the “outcome” column as one of the outcomes. The “outcome” column contains the category variables (1 or 0), 1 means the person has diabetes, and 0 means not.

6. What is your *draft* Statistical Analysis Plan? (this should be a very thorough, detailed, bullet point outline, demonstrating that you have broadly thought this through and included details - this may be updated of course as the semester proceeds, and with feedback) ** Note that we will be discussing all forms of outcome/endpoint data in this course, and at present have not yet covered each of these...so this plan may be updated/added to as the semester progresses, but you still should be able to plan out the structure and significant details of your plan. If your outcome data is other than continuous for instance (and we encourage you to consider any/all outcome data forms!), you can still include for instance 'Regression modeling involving 'Y' outcome data of interest, involving these variables (list them)...' and any other concerns or methods of interest (listing potential confounders, effect modifiers, the potential use of splines or additive modeling, potential missing data considerations, data reduction methods, regularization methods, etc,...or none of these--you will want to consider what is most appropriate for your data and questions at hand). Recall the BST 210 Regression Models Overview Table from which most extensions arise.

   (1) Data Cleaning: if there are any NAs, we need to use the method to extract or fill out the missing value. Here we notice that a lot of values in BMI and insulin are filled with “0”.
   (2) Residual analysis: Check if there are outliers (possible ways include standardized residual, leverage, and Cook’s distance). If so, figure out possible reasons for outlying observations and decide if we should fix, discard, or keep the outliers. In this step, we need factual knowledge such as the range of BMI and average pregnancies in India to evaluate each outlier.
   (3) Exploratory Analysis of predictor variables: Plot histogram and boxplot of each variable. Plot scatter plot with a smooth line of predictor variable and response variable.
   (4) Model selection: We plan to use a linear model for our prediction. The backward Elimination method and covariate’s p-value AIC will be used to determine the covariates used in the model.
   (5) Confounding and effect modification: based on factual knowledge, we will test the confounding and effect modification of the association between predictor and response to improve our fitted model.
   (6) (Possible) Machine learning: try to predict whether the person has diabetes or not based on the current medical predictor.

7. What are the biggest challenges you foresee in answering your proposed questions and completing this project? (logistical, statistical, etc, if there are any)

   Data cleaning can be challenging because there are many “0”s in the column “insulin” and “skin sickness”.
   We need more background knowledge to understand diagnostic measurements and their relationships.
   The machine learning part is also challenging since we never learned this before.

8. Will you seek domain expertise? Why or why not? If so, from whom? *(Seeking domain expertise is strongly encouraged!)*

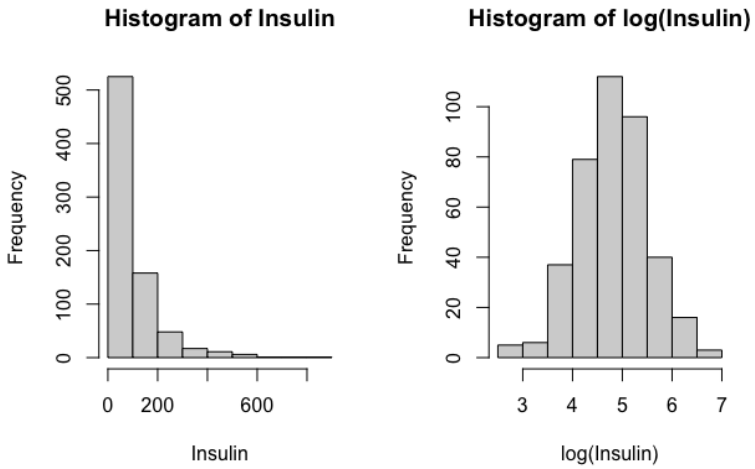
   We will seek domain expertise. The first thing we will do is to read the class notes and make sure we understand all the concepts. Then, we will see some insights on Kaggle since a lot of people post questions and knowledge notes on this website, and we may learn something from them.

9. What software package(s) will you use to complete this project? (It is absolutely fine for different group members to use different packages; in fact, some tasks are easier in some packages over others and vice versa.)

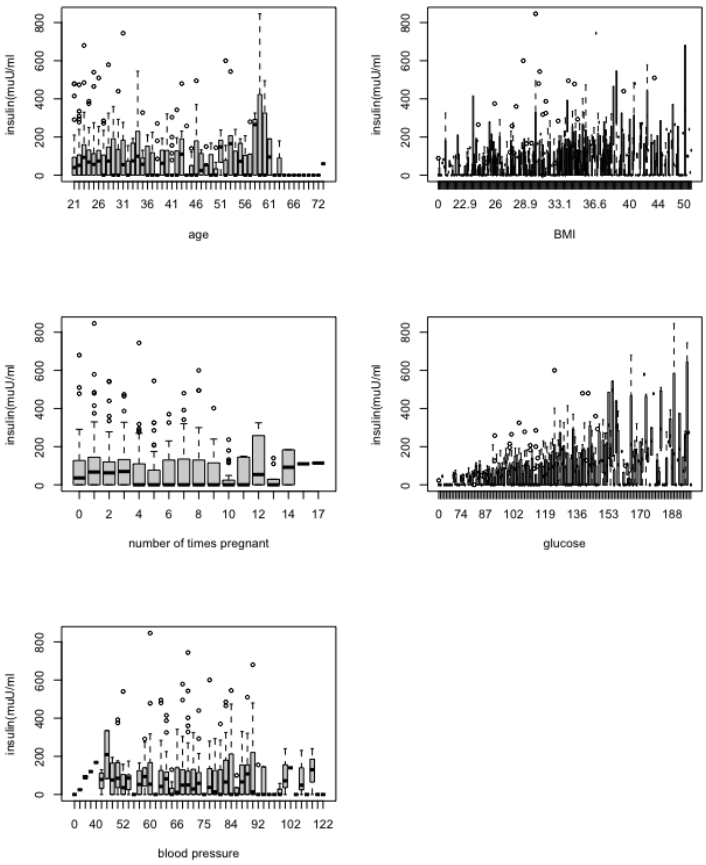
   We will mainly use R (tidyverse, ggplot2, glm, etc.) and maybe a little bit of Python (seaborn, heatmap, etc.). But we will add more when the project continues.

10. Complete an initial round of exploratory analyses on your data that would be relevant to your plan and responses above, and include/report **several MAIN** (not all) critical plots or summaries. Include ALL code after that in the usual appendix. Please carry out exploratory analysis for outcome(s) of continuous form however/wherever possible even if your ultimate goals/questions involve a different form of outcome data such as binary, polytomous, etc. (You may consider this initial analysis as a potential sub-analysis later on.)

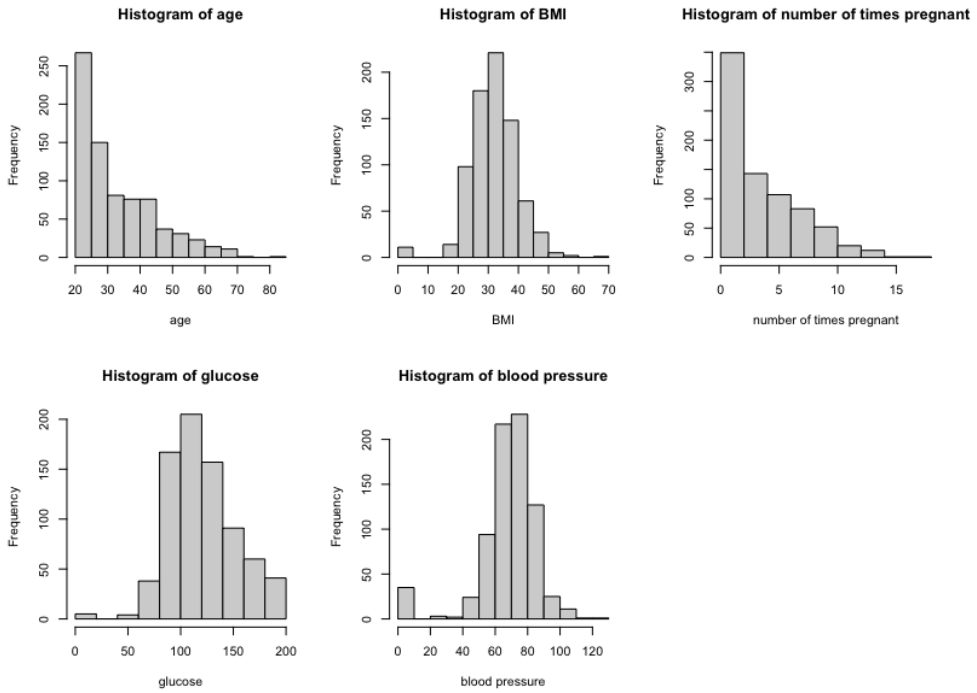
    Select Insulin as a continuous outcome:



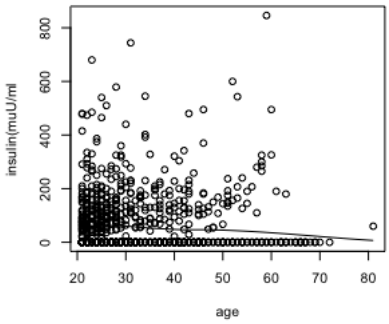
This is the histogram of insulin and log(insulin). From the histogram, we can see that the range of insulin is pretty wide from 0 muU/ml to 600 muU/ml. But for most insulin concentrations are within 0 - 200 muU/ml.

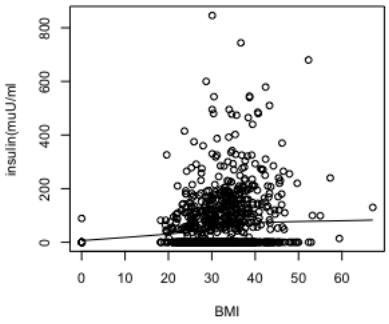


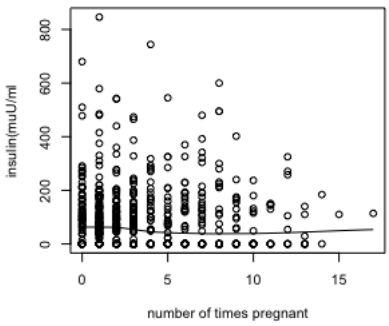
These are the boxplots between each predictor (age, BMI, number of times pregnant, glucose concentration, and blood pressure) and outcome (2-hour serum insulin).

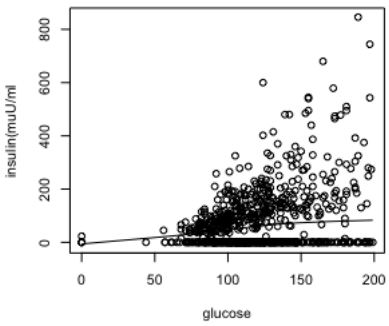


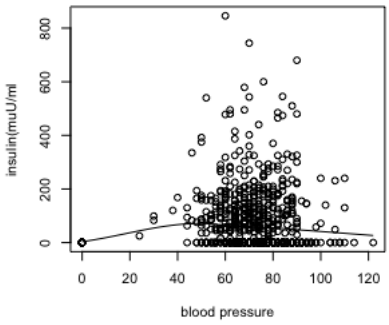
These are the histograms of each predictor (age, BMI, number of times pregnant, glucose concentration, and blood pressure). We can see that all participants are above 20, and most of them are between 20 - 30. For the BMI, most people's BMI is between 25 to 40. Then most people's pregnancy times are less than 5. For the glucose, most people fall into around 70 to 150. For the blood pressure, most people are between 50 to 90.



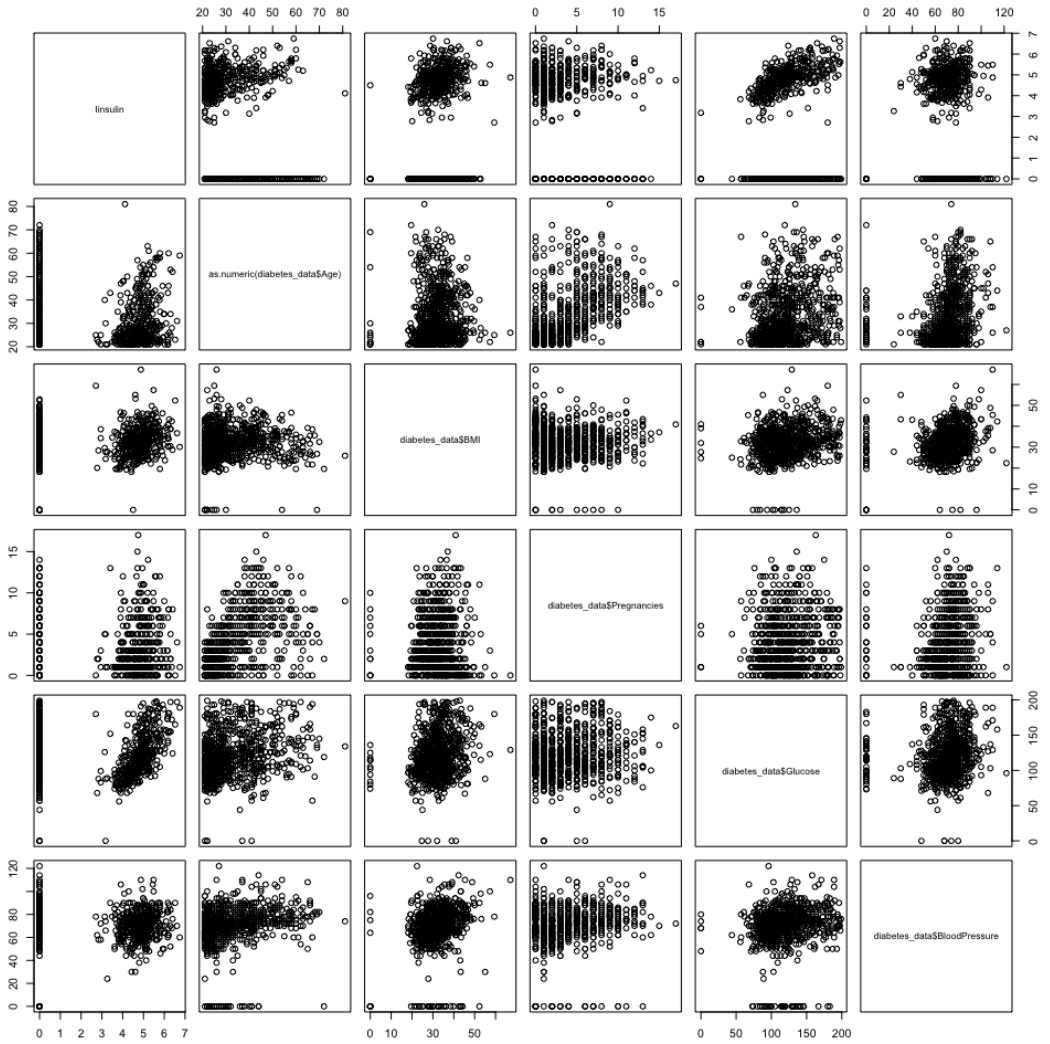








These are the scatter smooth plots of each predictor (age, BMI, number of times pregnant, glucose concentration, and blood pressure). We can see that, except for the number of times pregnant, most values are within 0 - 200. Two plots show the “skeptical” trend: the number of times pregnant plot shows the increasing number of times pregnant with the decrease of insulin; the glucose plot shows the increasing of glucose with the increasing of insulin.



11. Project Attestation: No member of this group is using these data or the same/similar questions in any other course or course project, at HSPH. *By listing your name as a group member on your project, and submitting this assignment, you are attesting to this statement above. Groups must include this attestation here under Question 11 for credit!*

    By listing our name as group members on our project, and submitting this assignment, we are attesting to this statement above.

Code Appendix

```
library(dplyr)
library(tidyr)
library(stringr)
library(tidyverse)

# load data
diabetes_data <- read.csv("diabetes.csv")

# check the data
head(diabetes_data)
summary(diabetes_data)

# check the missing value
colSums(is.na(diabetes_data))

# no missing value

# first select insulin as outcome
summary(diabetes_data$Insulin)

# Some EDA (exploratory data analysis of outcome insulin, and
predictors)

# create histogram for outcome
par(mfrow=c(1,2))
hist(diabetes_data$Insulin, main="Histogram of Insulin",
xlab="Insulin")
hist(log(diabetes_data$Insulin), main="Histogram of log(Insulin)",
xlab="log(Insulin)")

#draw the boxplots of outcome insulin, and predictors
par(mfrow=c(3,2))
boxplot(diabetes_data$Insulin ~ diabetes_data$Age, xlab="age",
ylab="insulin(muU/ml")
boxplot(diabetes_data$Insulin ~ diabetes_data$BMI, xlab="BMI",
ylab="insulin(muU/ml")
boxplot(diabetes_data$Insulin ~ diabetes_data$Pregnancies,
```

```
xlab="number of times pregnant", ylab="insulin(muU/ml")
boxplot(diabetes_data$Insulin ~ diabetes_data$Glucose,
xlab="glucose", ylab="insulin(muU/ml")
```

```
boxplot(diabetes_data$Insulin ~ diabetes_data$BloodPressure,
xlab="blood pressure", ylab="insulin(muU/ml")

#draw the histograms of outcome insulin, and predictors
par(mfrow=c(2,3))
hist(diabetes_data$Age, main="Histogram of age", xlab="age")
hist(diabetes_data$BMI, main="Histogram of BMI", xlab="BMI")
hist(diabetes_data$Pregnancies, main="Histogram of number of times
pregnant", xlab="number of times pregnant")
hist(diabetes_data$Glucose, main="Histogram of glucose",
xlab="glucose")
hist(diabetes_data$BloodPressure, main="Histogram of blood pressure",
xlab="blood pressure")


#draw the scatter plots of outcome insulin, and predictors
par(mfrow=c(3,2))
scatter.smooth(as.numeric(diabetes_data$Age), diabetes_data$Insulin,
xlab="age", ylab="insulin(muU/ml")
scatter.smooth(diabetes_data$BMI, diabetes_data$Insulin, xlab="BMI",
ylab="insulin(muU/ml")
scatter.smooth(diabetes_data$Pregnancies, diabetes_data$Insulin,
xlab="number of times pregnant", ylab="insulin(muU/ml")
scatter.smooth(diabetes_data$Glucose, diabetes_data$Insulin,
xlab="glucose", ylab="insulin(muU/ml")
scatter.smooth(diabetes_data$BloodPressure, diabetes_data$Insulin,
xlab="blood pressure", ylab="insulin(muU/ml")

Iinsulin <- log(diabetes_data$Insulin+1)
pairs(Iinsulin ~
as.numeric(diabetes_data$Age)+diabetes_data$BMI+diabetes_data$Pregnan
cies+diabetes_data$Glucose+diabetes_data$BloodPressure)
```