Overview:

This data set consists of diagnostic criteria from 768 patients with similar characteristics (females 21 years or older and of Pima Indian heritage) and whether they have diabetes. It has the following features:

Outcome: This feature is 1 if the patient has diabetes and 0 if they do not.

Age: The patient's age in years.

DiabetesPedigreeFunction (DPF): A score which predicts the likelihood of diabetes based on family history.

BMI: Body mass index in kg/m^2

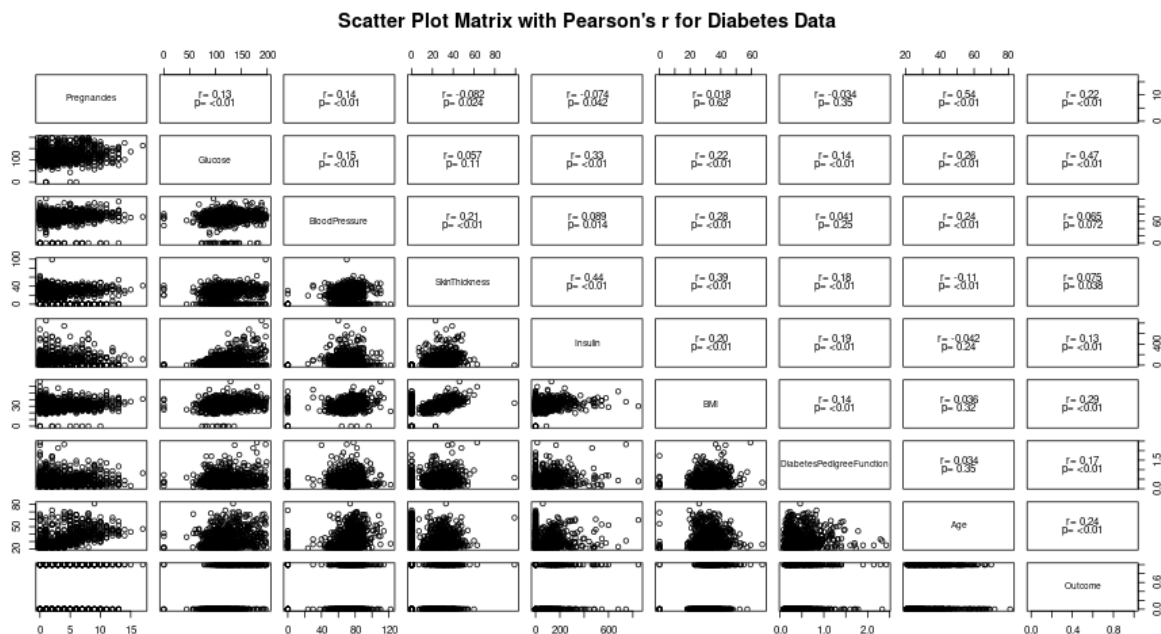Insulin: 2-hour serum insulin in muU/ml.

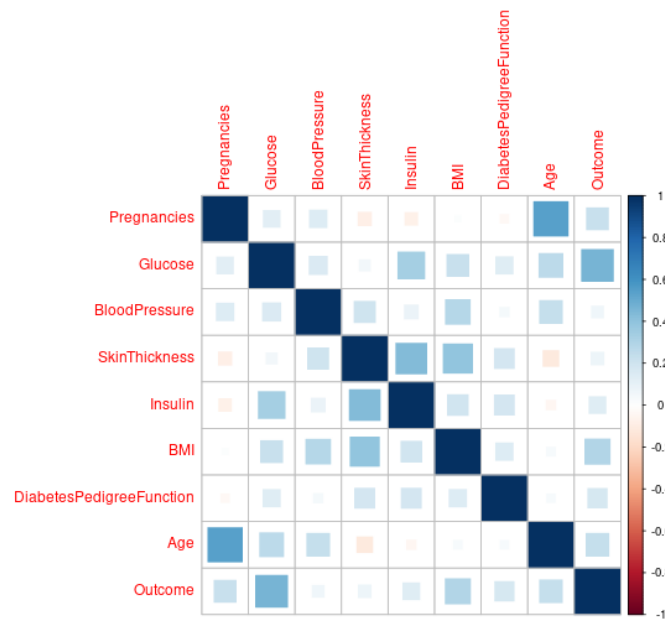SkinThickness: Tricep skin fold thickness in mm.

BloodPressure: Diastolic blood pressure in mm Hg

Glucose: Plasma glucose concentration as determined by an oral glucose tolerance test.

Pregnancies: The number of times the patient has been pregnant.

Summary:

Scatter Plot Matrix with Pearson's r for Diabetes Data

For the most part, the quality of this data seems to be good. The one exception is that certain tuples have measurements which don't make sense, such as those with a skin thickness or BMI of 0 mm. It is possible that some of these were rounded down, given that the skin thickness measurements are integers, but others might be substitutes for missing data.

Although this data set contains medical data, it lacks identifying information and most likely does not pose any privacy concerns.

The features which describe diagnostic criteria all seem to be useful without any cleanup or alteration. Since outcome describes a characteristic that patients either have or do not have, it may need to be replaced with a nominal feature that contains the same information.
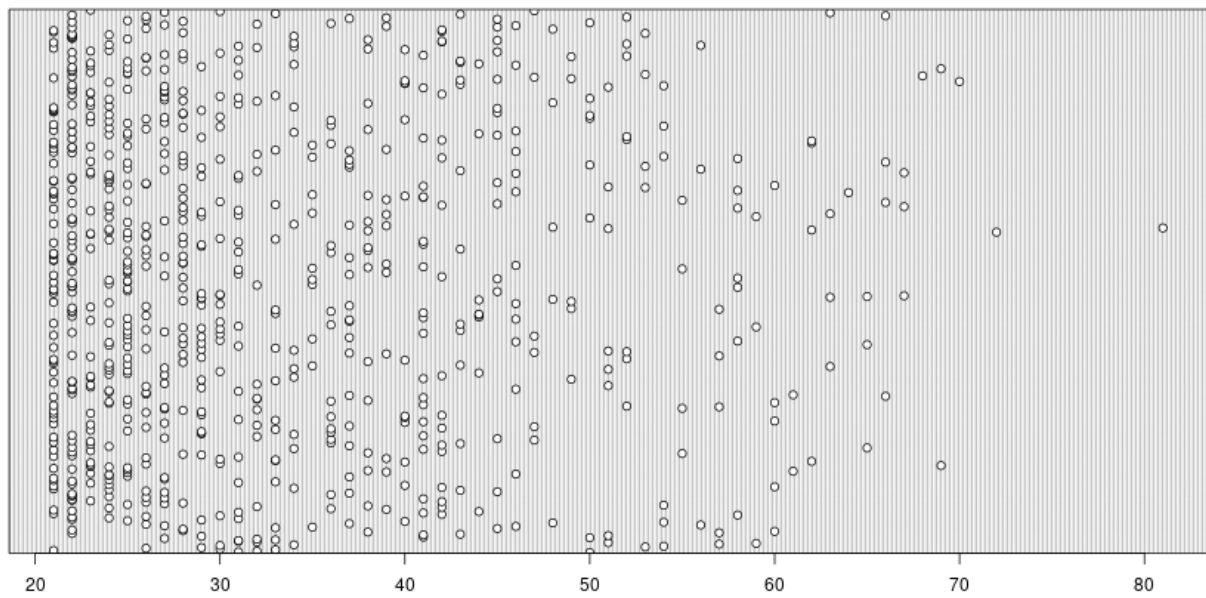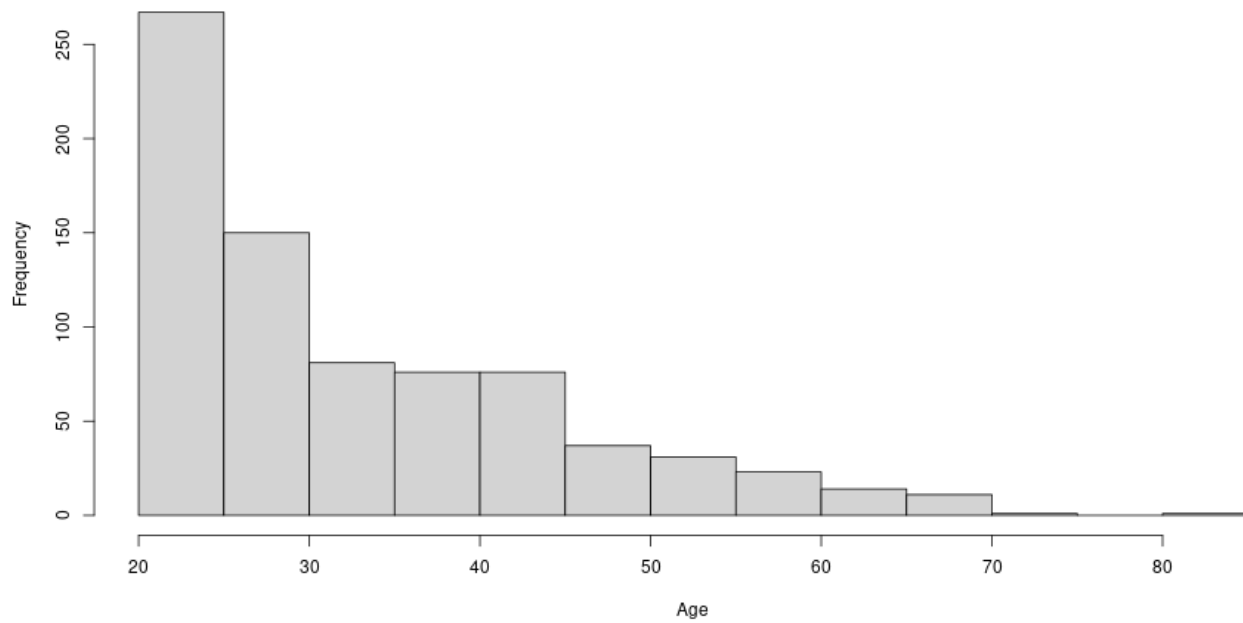
Details:

Outcome

Outcome Frequencies
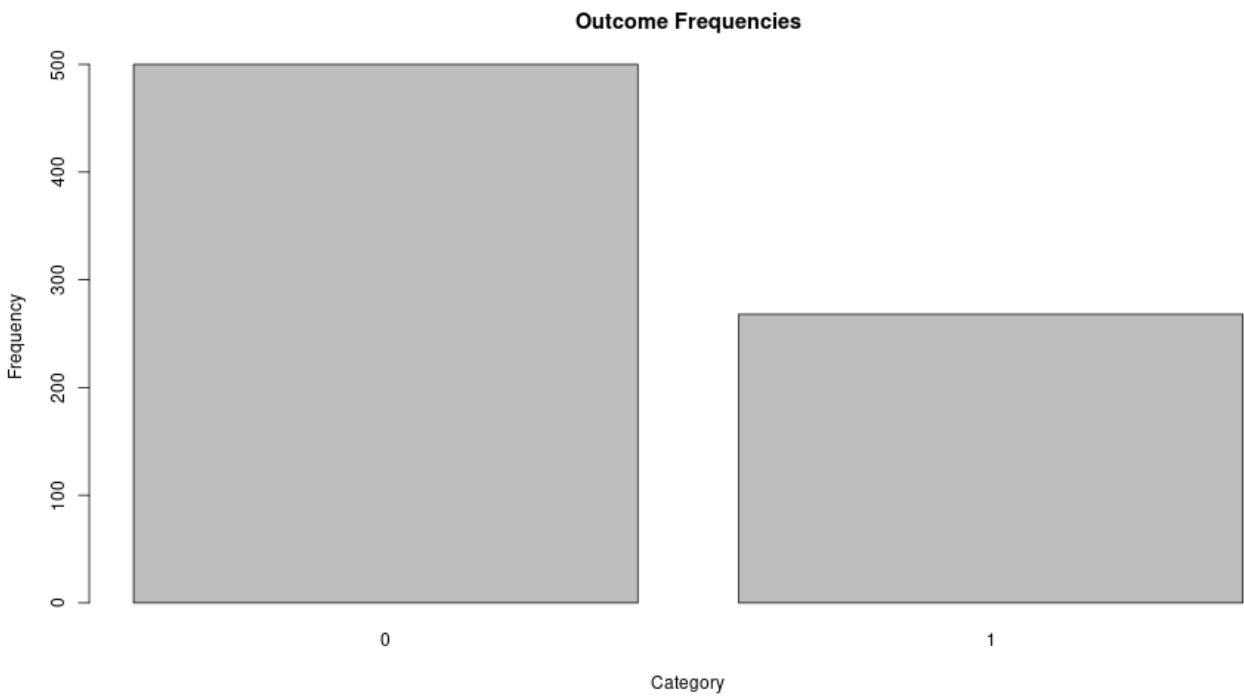
Number of patients with diabetes: 268
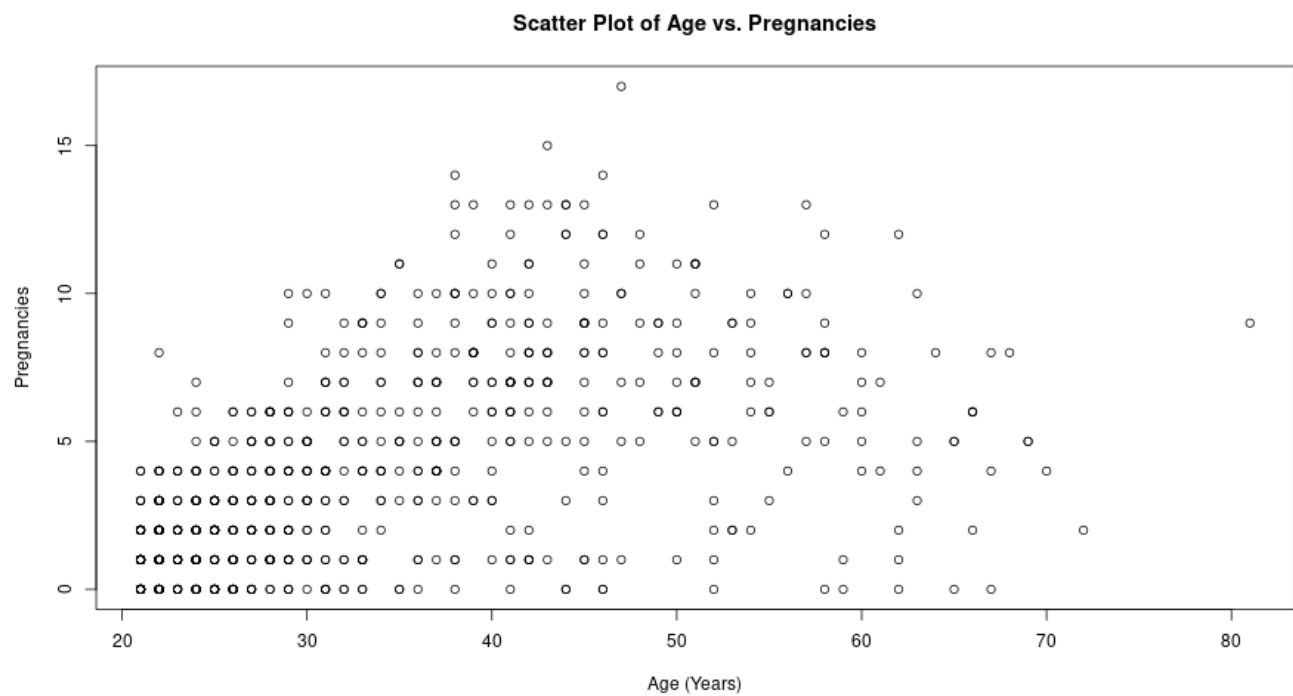Number of patients without diabetes: 500

Age

Scatter Plot of Age vs. Pregnancies

Mean: 33.24

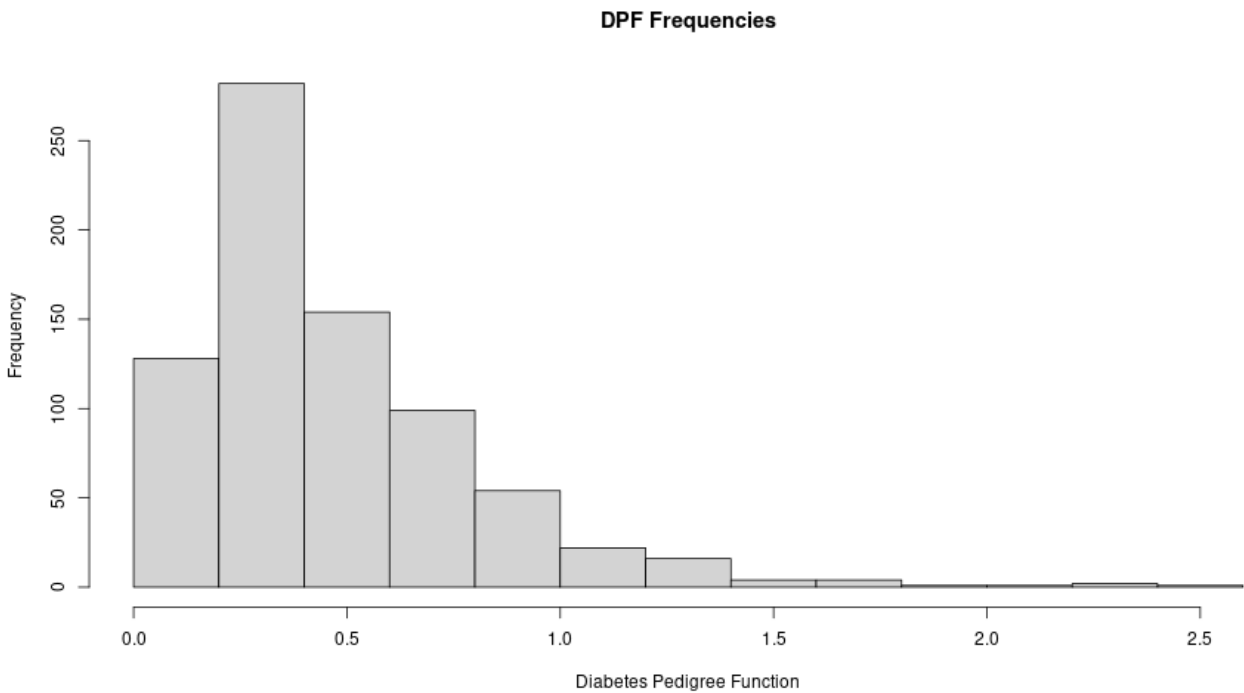Five-number summary:
Minimum: 21
First Quartile: 24
Median: 29
Third Quartile: 41
Maximum: 81

DiabetesPedigreeFunction

DPF Frequencies

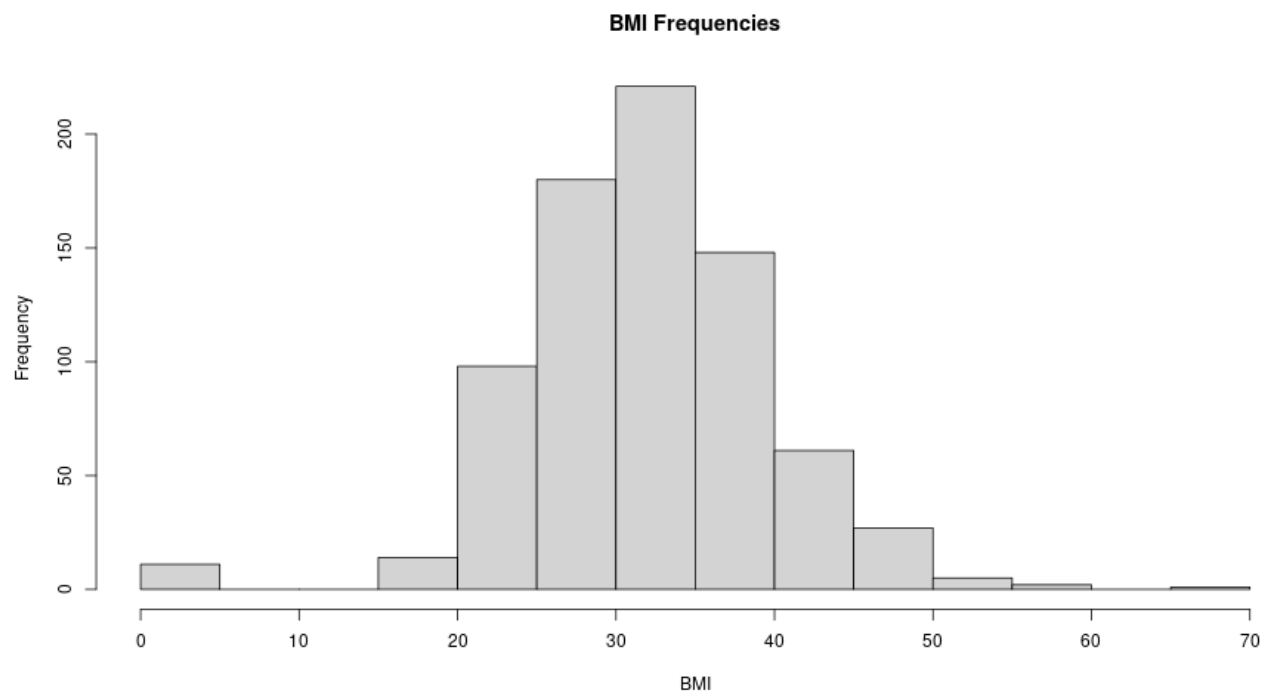
Mean: 0.4719

Five-number summary:
Minimum: 0.078
First Quartile: 0.2437
Median: 0.3725
Third Quartile: 0.6262
Maximum: 2.42

BMI

BMI Frequencies
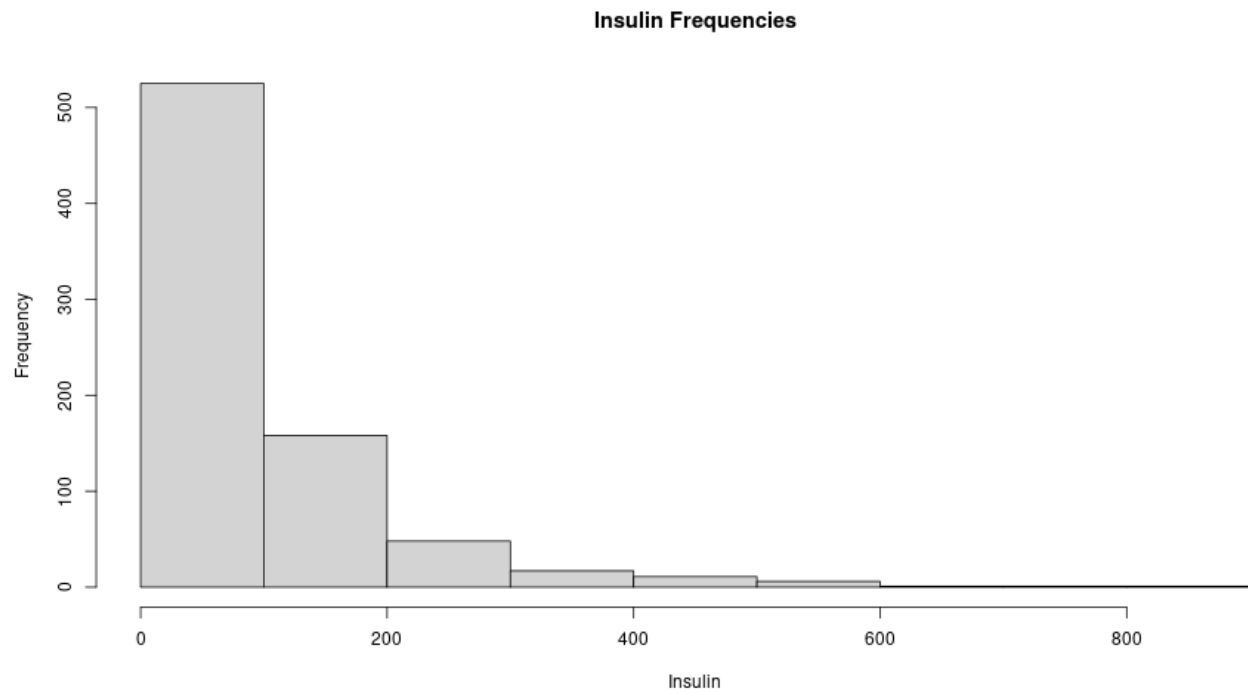

Mean: 31.99

Five-number summary:
Minimum: 0
First Quartile: 27.3
Median: 32
Third Quartile: 36.6
Maximum: 67.1

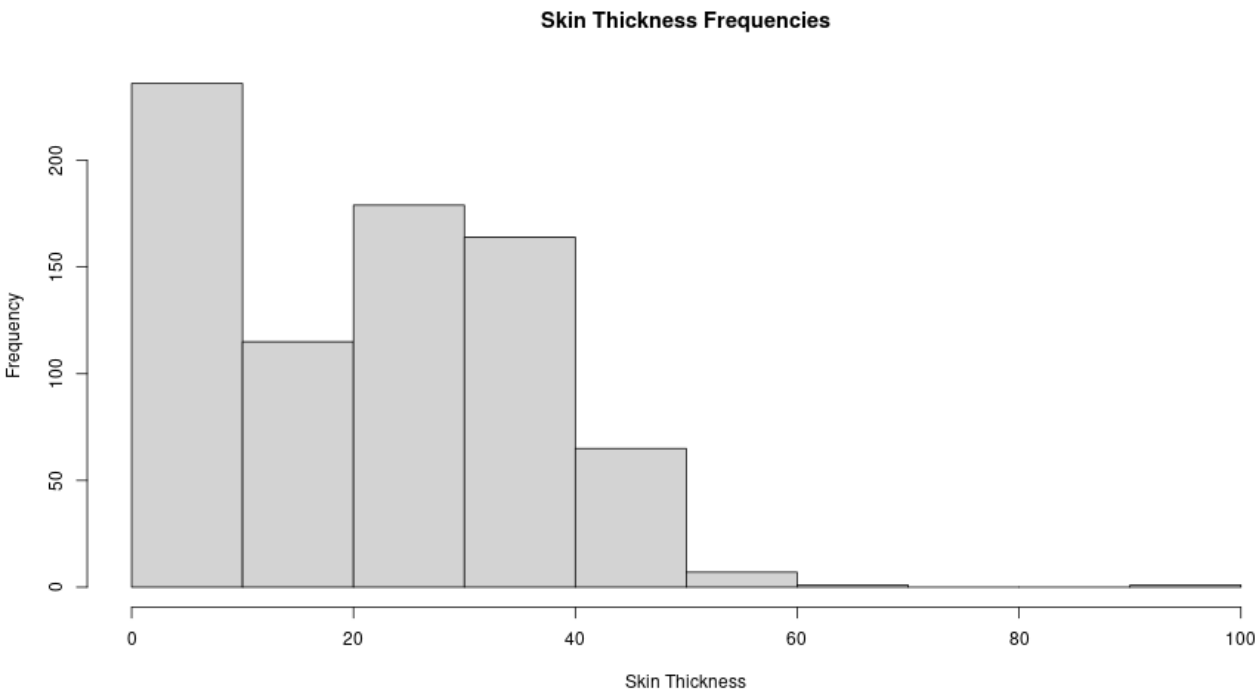**Insulin**

Mean: 79.8

Five-number summary:
Minimum: 0
First Quartile: 0
Median: 30.5
Third Quartile: 127.2
Maximum: 846

SkinThickness

Skin Thickness Frequencies

Mean: 20.54

Five-number summary:
Minimum: 0
First Quartile: 0
Median: 23
Third Quartile: 32
Maximum: 99

BloodPressure

Strip Chart with Y Limit 1-20 for Blood Pressure

Blood Pressure

Blood Pressure Frequencies

Blood Pressure
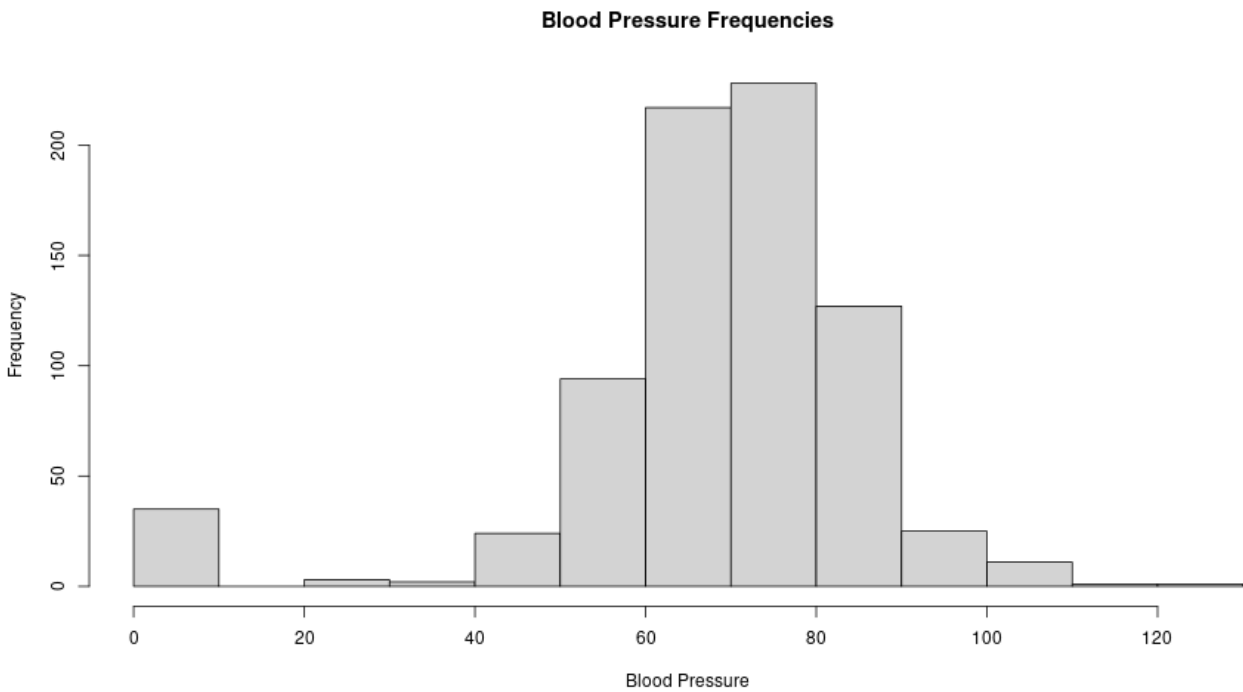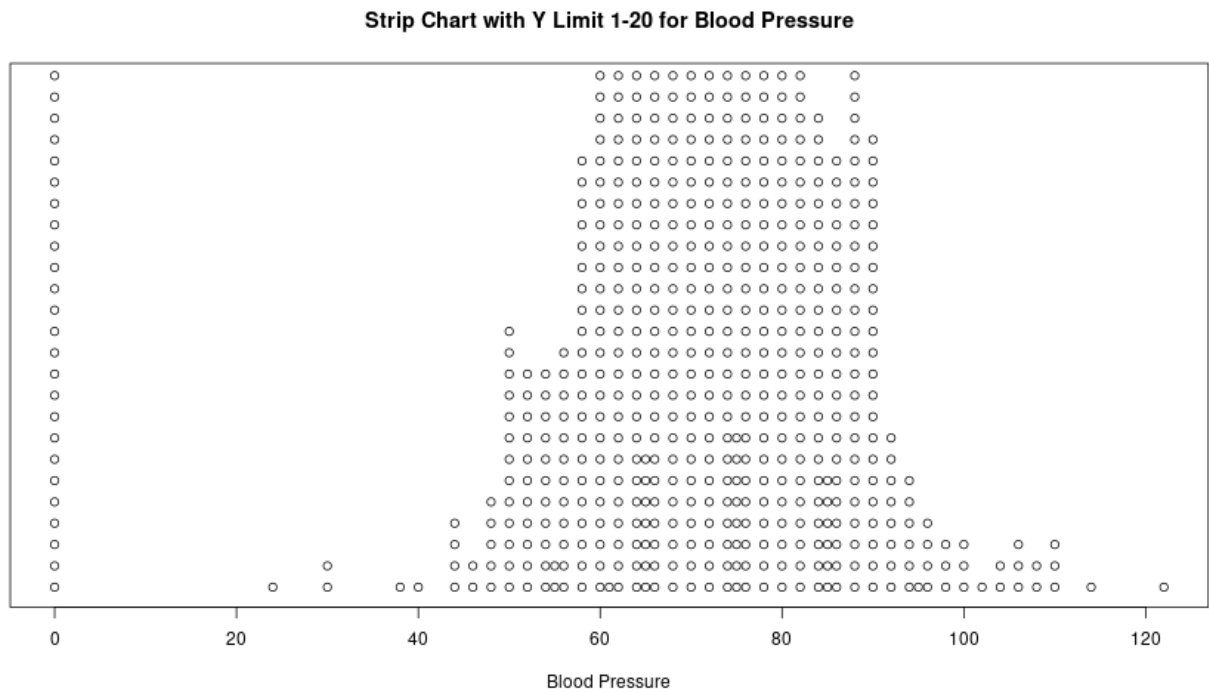
Mean: 69.11

Five-number summary:
Minimum: 0
First Quartile: 62
Median: 72

Third Quartile: 80
Maximum: 122

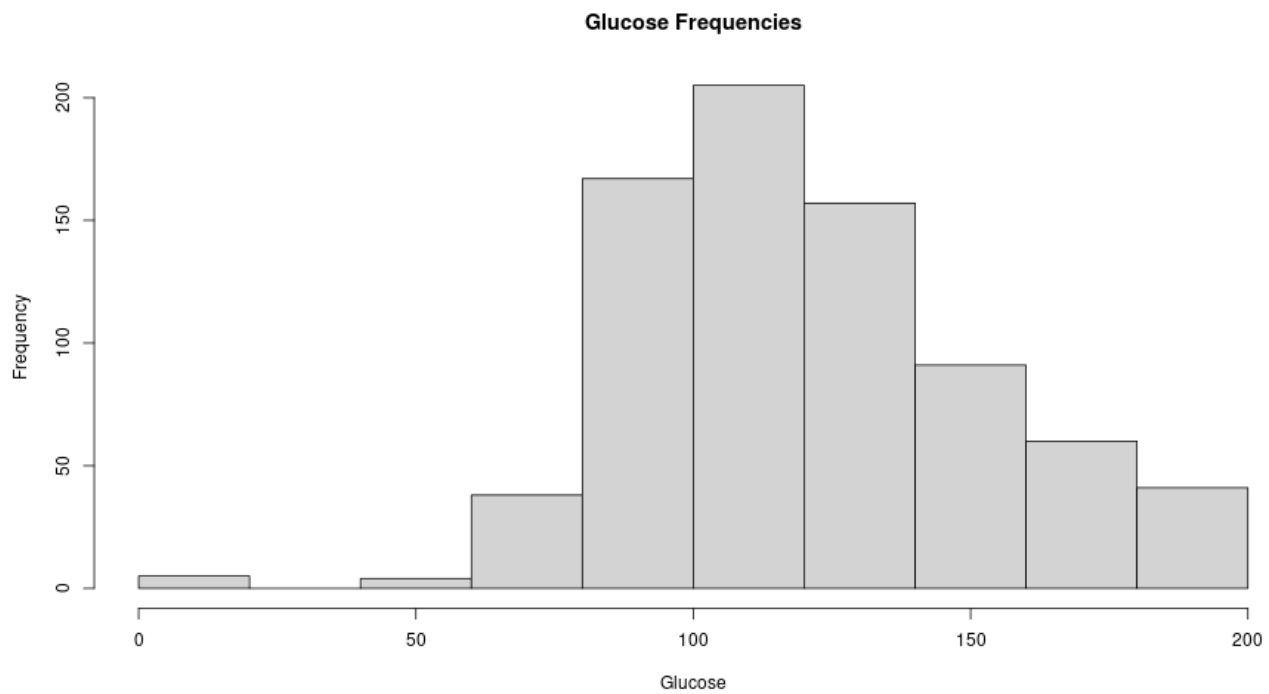**Glucose**

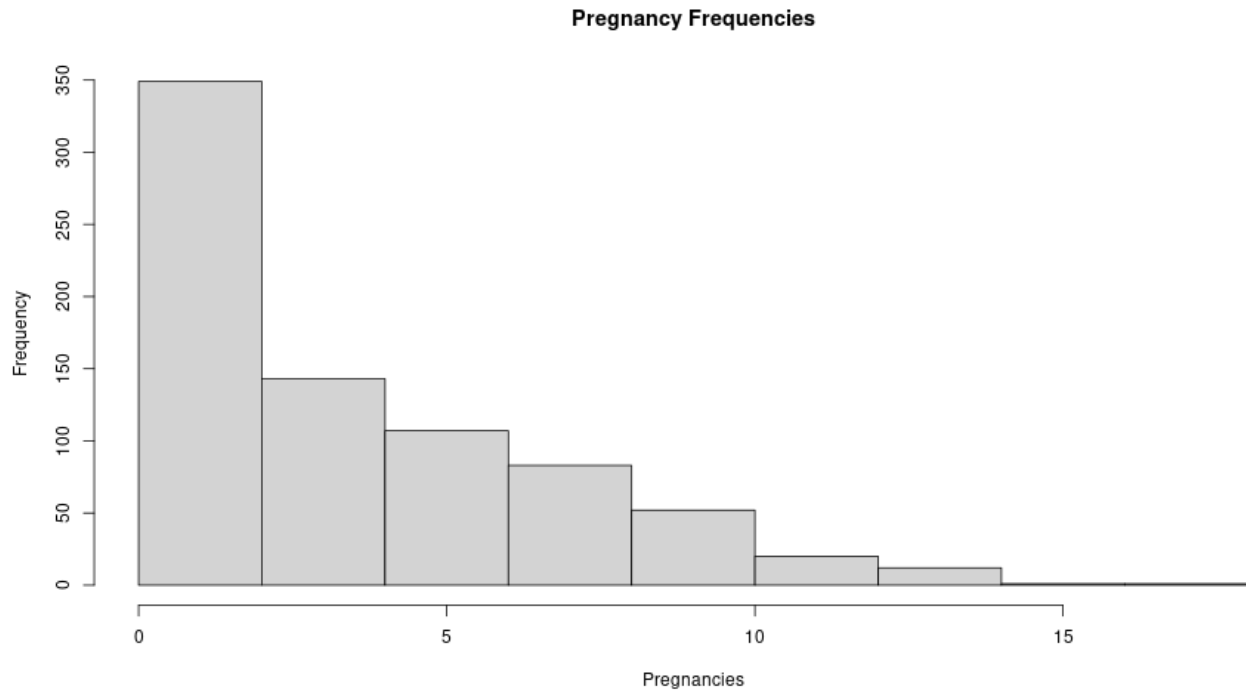Mean: 120.9

Five-number summary:
Minimum: 0
First Quartile: 99.0
Median: 117
Third Quartile: 140.2
Maximum: 199.0

Pregnancies

Pregnancy Frequencies

See also: The Age vs. Pregnancies scatterplot under Age.

Mean: 3.84

Five-number summary:
Minimum: 0
First Quartile: 1
Median: 3
Third Quartile: 3
Maximum: 17