The data I selected was obtained from Kaggle here. It is a collection of observations of a random sample of many different people with varying ages, weights and other medical determinants. It was collected for the use of potentially linking positive diabetes diagnoses to any number of the collected variables. I found this interesting because several of my family members have diabetes.

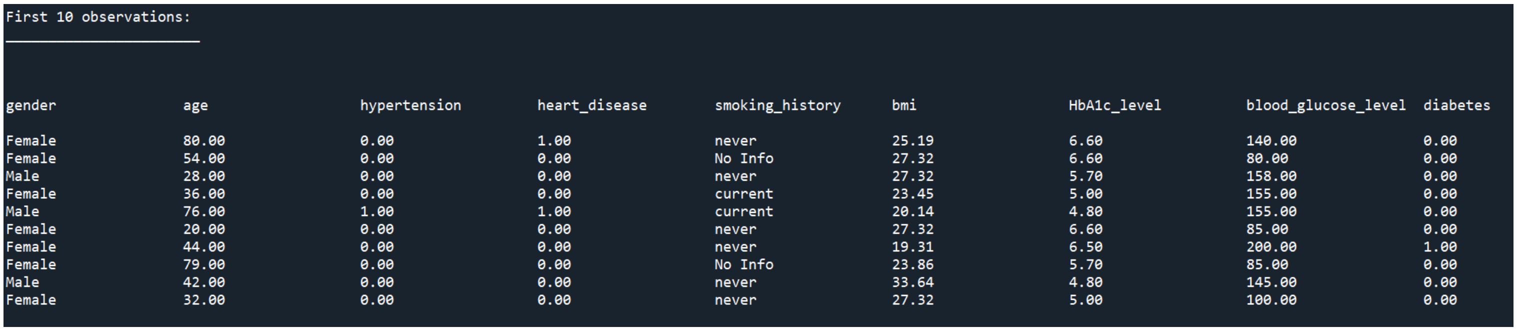
After observing the data, I proposed the question “Is bmi a predictor of diabetes?”.

The variables of the dataset are as follows:

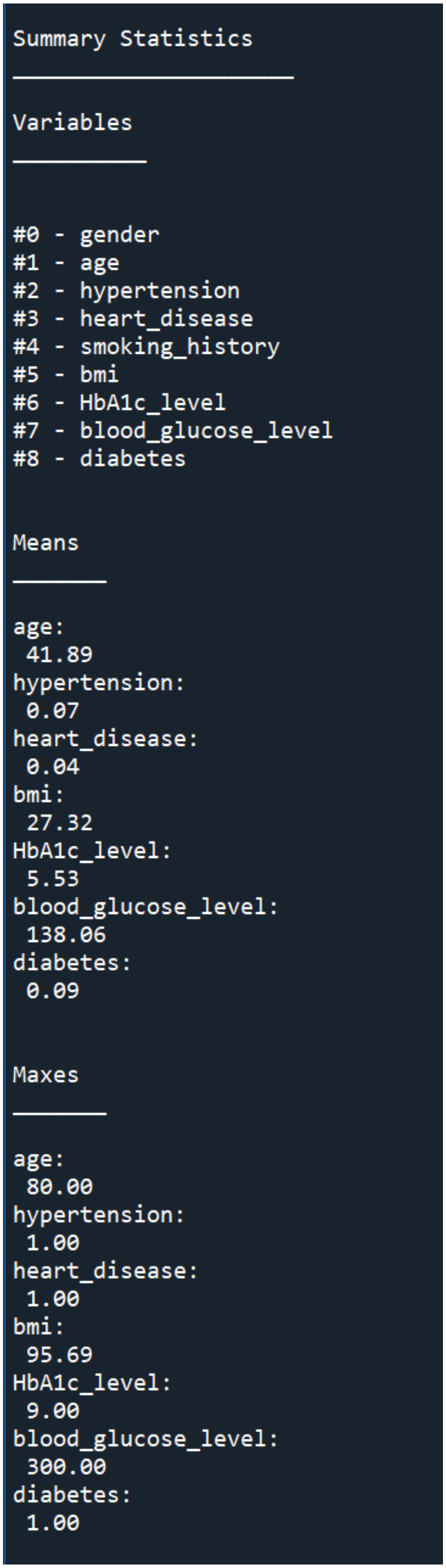
0 = gender
1 = age
2 = hypertension (y/n)
3 = heartdisease (y/n)
4 = smoker status
5 = bmi
6 = hbA1c level
7 = blood glucose levels
8 = diabetes (y/n)

Gender and smoker status are the only non numeric variables. Binary variables are 1 for yes and 0 for no and were considered to be real numbers for calculation purposes.

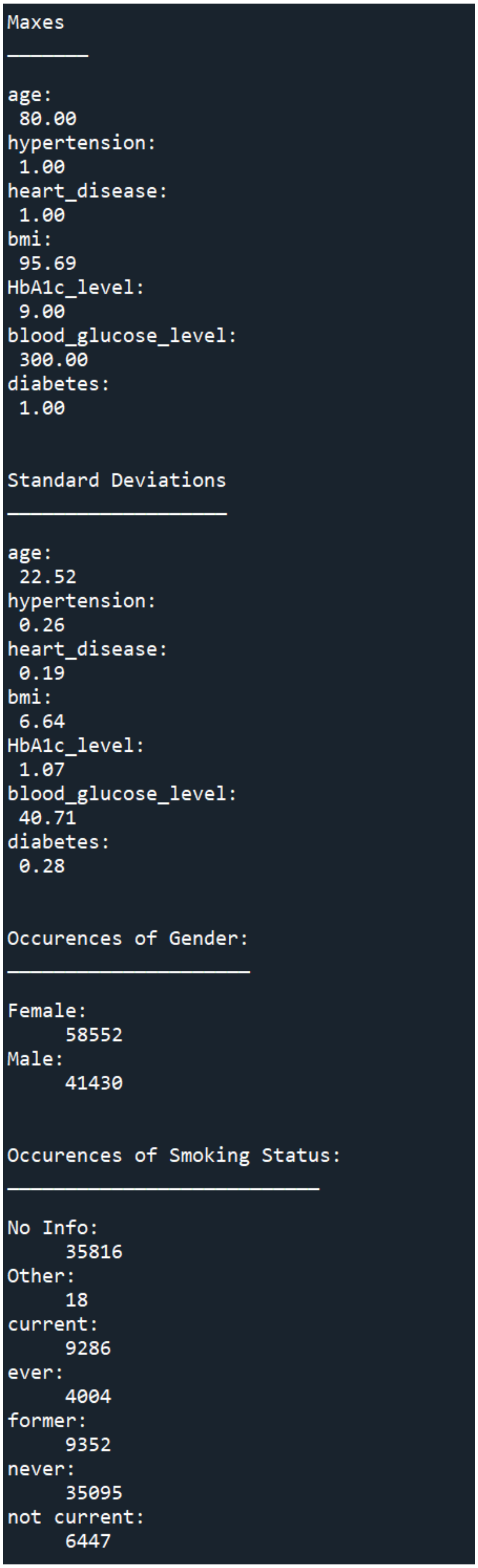
I produced output for the basic summary statistics for each variable. The first 10 observations are below and the basic summary statistics are on the next page. The console window size may need to be adjusted to see the output correctly to see the full formatting.



```
First 10 observations:
______________________


gender              age                 hypertension        heart_disease       smoking_history     bmi                 HbA1c_level         blood_glucose_level diabetes

Female              80.00               0.00                1.00                never               25.19               6.60                140.00              0.00
Female              54.00               0.00                0.00                No Info             27.32               6.60                80.00               0.00
Male                28.00               0.00                0.00                never               27.32               5.70                158.00              0.00
Female              36.00               0.00                0.00                current             23.45               5.00                155.00              0.00
Male                76.00               1.00                1.00                current             20.14               4.80                155.00              0.00
Female              20.00               0.00                0.00                never               27.32               6.60                85.00               0.00
Female              44.00               0.00                0.00                never               19.31               6.50                200.00              1.00
Female              79.00               0.00                0.00                No Info             23.86               5.70                85.00               0.00
Male                42.00               0.00                0.00                never               33.64               4.80                145.00              0.00
Female              32.00               0.00                0.00                never               27.32               5.00                100.00              0.00
```

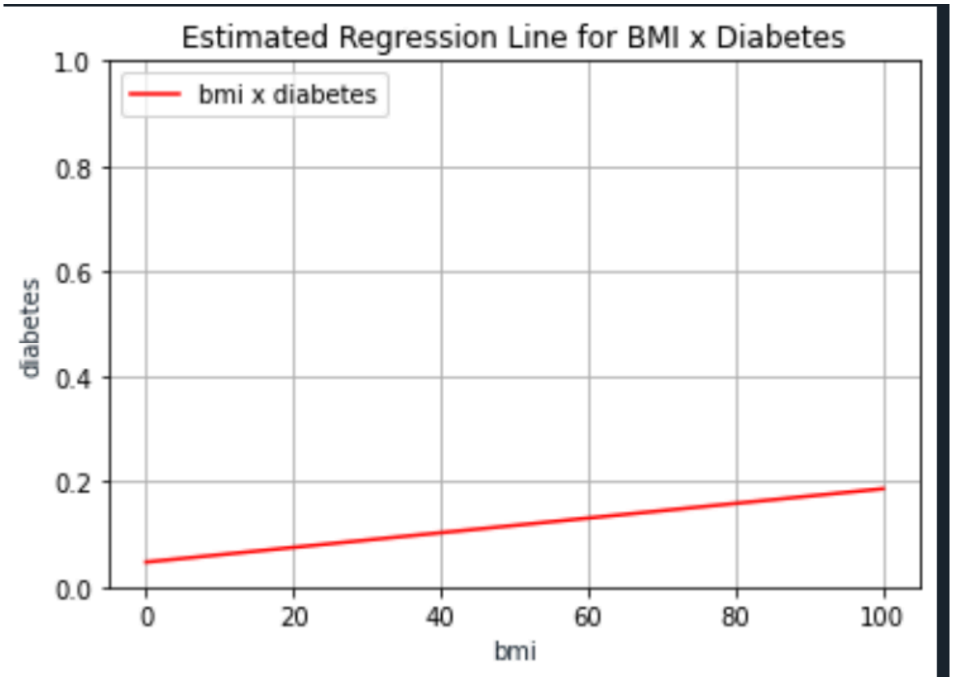


```
Summary Statistics
─────────────────────

Variables
──────────


#0 - gender
#1 - age
#2 - hypertension
#3 - heart_disease
#4 - smoking_history
#5 - bmi
#6 - HbA1c_level
#7 - blood_glucose_level
#8 - diabetes


Means
───────

age:
 41.89
hypertension:
 0.07
heart_disease:
 0.04
bmi:
 27.32
HbA1c_level:
 5.53
blood_glucose_level:
 138.06
diabetes:
 0.09


Maxes
───────

age:
 80.00
hypertension:
 1.00
heart_disease:
 1.00
bmi:
 95.69
HbA1c_level:
 9.00
blood_glucose_level:
 300.00
diabetes:
 1.00
```



```
Maxes
───────

age:
 80.00
hypertension:
 1.00
heart_disease:
 1.00
bmi:
 95.69
HbA1c_level:
 9.00
blood_glucose_level:
 300.00
diabetes:
 1.00


Standard Deviations
───────────────────

age:
 22.52
hypertension:
 0.26
heart_disease:
 0.19
bmi:
 6.64
HbA1c_level:
 1.07
blood_glucose_level:
 40.71
diabetes:
 0.28


Occurences of Gender:
─────────────────────

Female:
     58552
Male:
     41430


Occurences of Smoking Status:
───────────────────────────

No Info:
     35816
Other:
     18
current:
     9286
ever:
     4004
former:
     9352
never:
     35095
not current:
     6447
```

I created a simple linear model using only the variable bmi as a predictor. I calculated its estimated coefficients and produced the following graphic:



The rudimentary regression study I developed on bmi as a descriptor for diabetes seems to make a case that further investigation on this relationship might be worthwhile. I was surprised to see that the supposed possible correlation is as mild as it is. One thing to note is that the dataset did not specify between the prevalence of type 1 or type 2 diabetes. Without this distinction, statements made about the population refer to everyone with diabetes and this may cause issues down the line.