Diabetes Data

Peter Sullivan

2/8/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
setwd("C:/Users/pjsul/OneDrive/Desktop/R HWS/")
diabetes_data <- read.csv("diabetes.csv", sep = ',')
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
head(diabetes_data)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
## 4           1      89            66            23      94 28.1
## 5           0     137            40            35     168 43.1
## 6           5     116            74             0       0 25.6
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
## 4                    0.167  21       0
## 5                    2.288  33       1
## 6                    0.201  30       0
```

```
colnames(diabetes_data)
```

```
## [1] "Pregnancies"              "Glucose"
## [3] "BloodPressure"            "SkinThickness"
## [5] "Insulin"                  "BMI"
## [7] "DiabetesPedigreeFunction" "Age"
## [9] "Outcome"
```

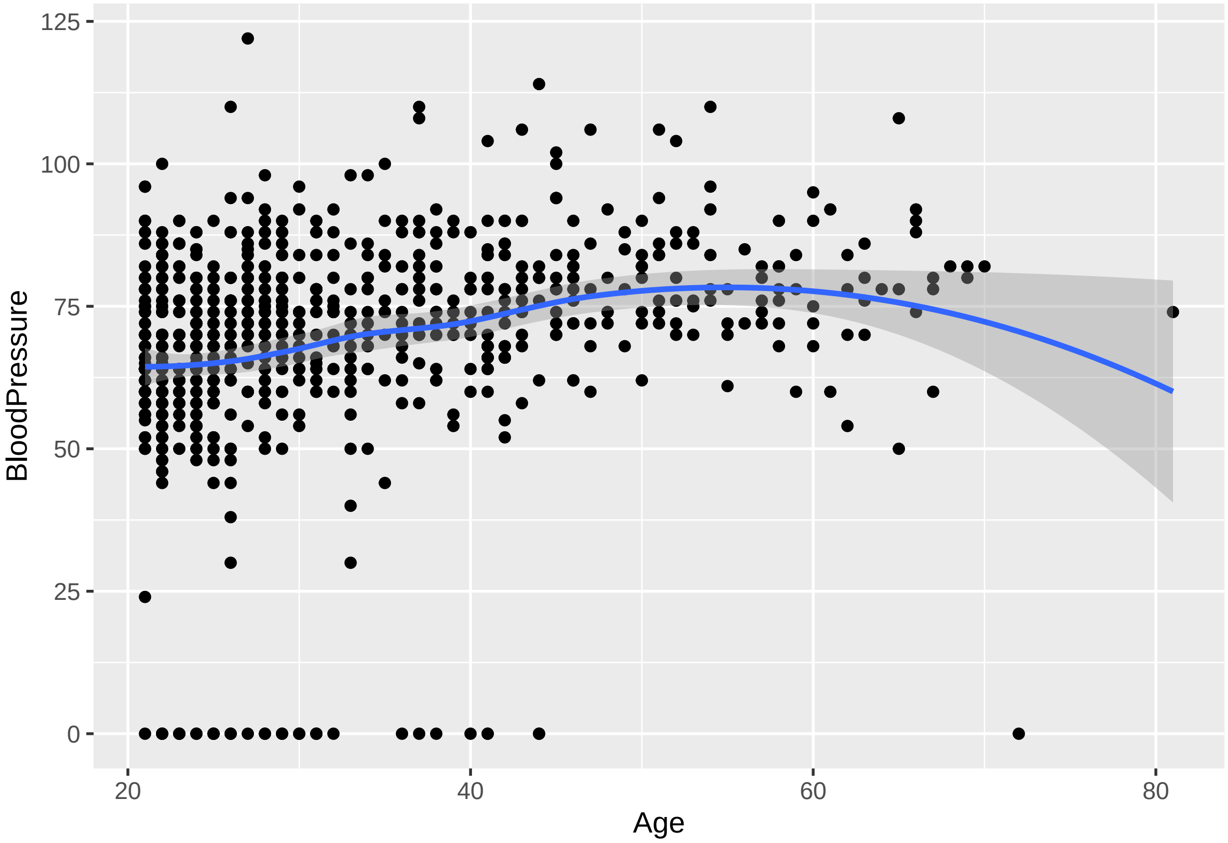
```
dim(diabetes_data)
```

```
## [1] 768   9
```

When looking at the Diabetes data set, there are 768 observations, and 9 variables. This data set was posted on kagle.com as a machine learning exercise, to see if someone can predict whether the patient has diabetes based on certain features. The target column, “Outcome”, states whether someone has diabetes; 1 for diabetes and 0 for not having diabetes.

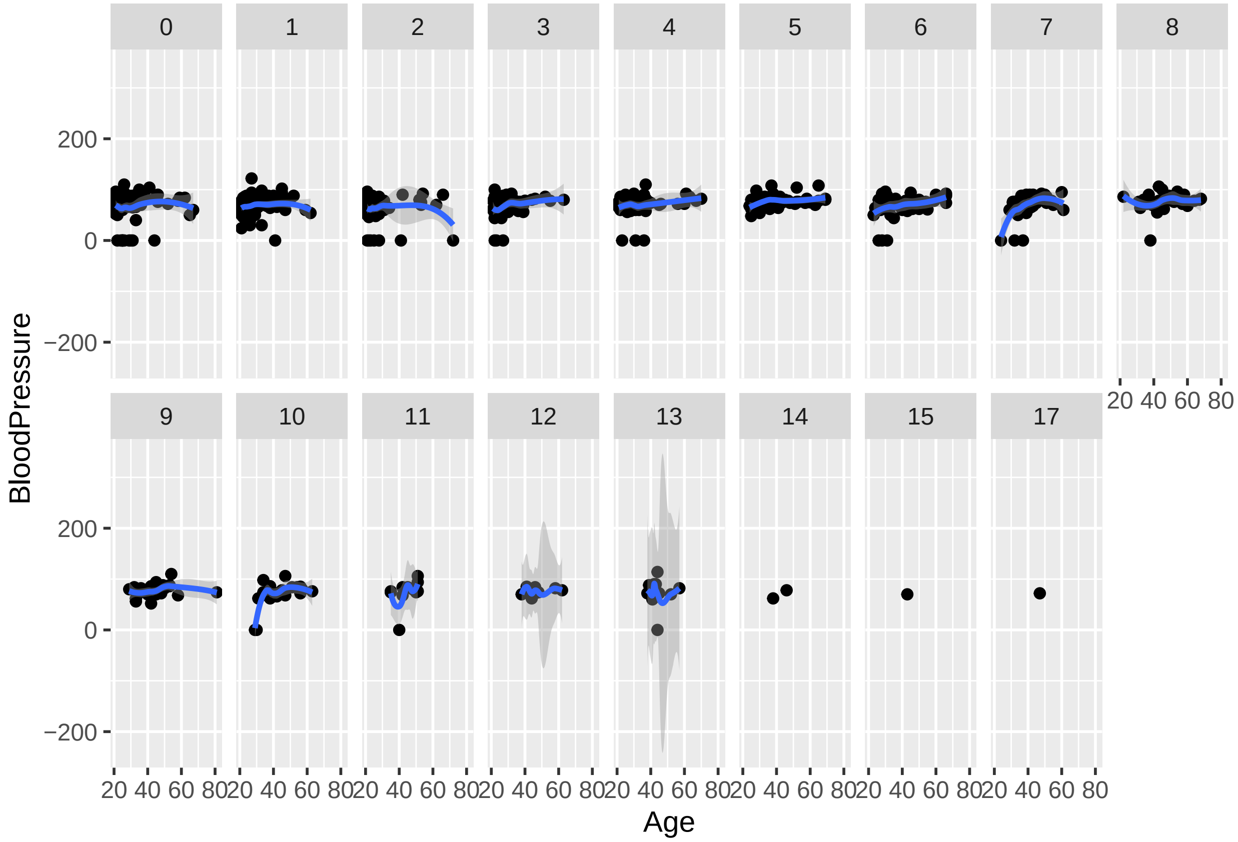
Here are the column names: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome

```
diabetes_data %>%
ggplot() +
  geom_point(mapping = aes(x = Age, y = BloodPressure))+
  geom_smooth(mapping = aes(x = Age, y = BloodPressure))
```



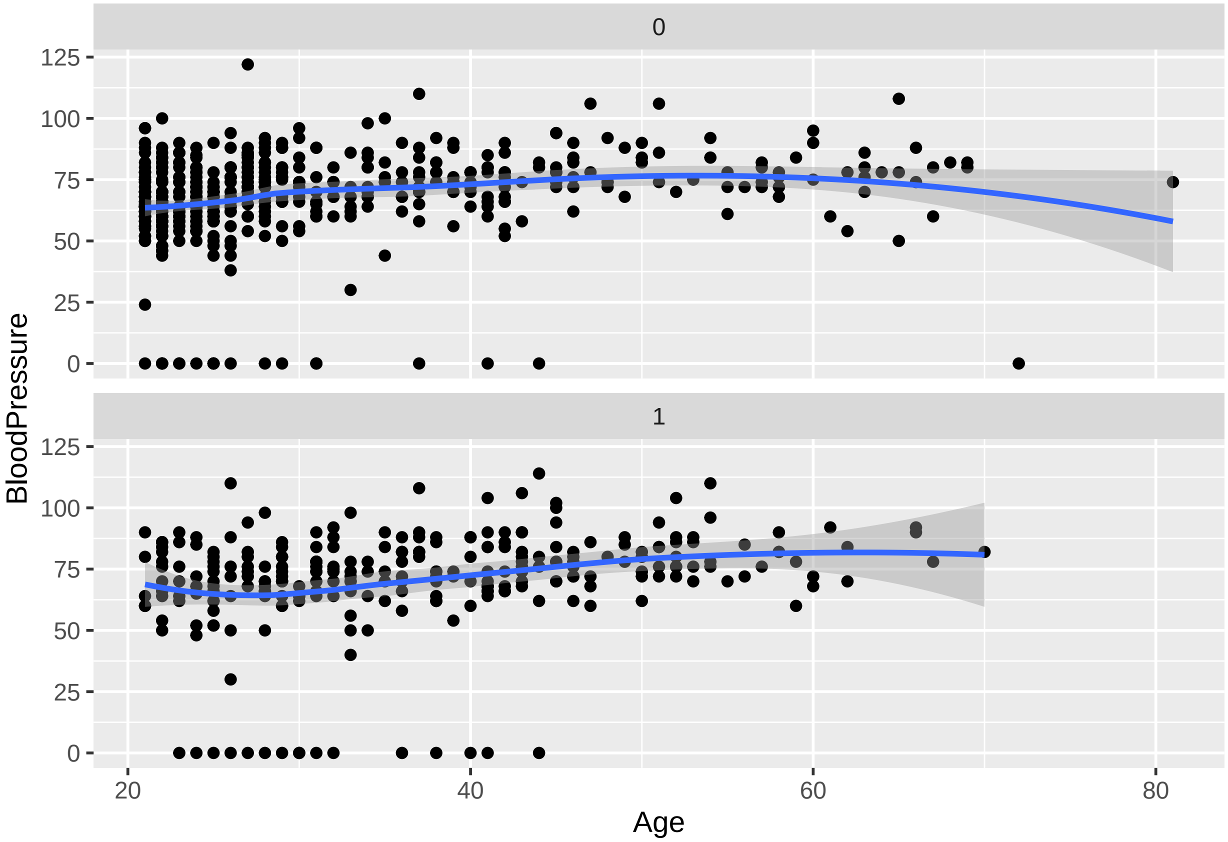
Here is an attempt at using a Facet wrap, specifically looking at the number of pregnancies per patient:

```
diabetes_data %>%
ggplot() +
  geom_point(mapping = aes(x = Age, y = BloodPressure)) +
  geom_smooth(mapping = aes(x = Age, BloodPressure)) +
  facet_wrap( ~ Pregnancies , nrow = 2)
```



This is another example of using a Facet wrap but instead of looking at pregnancies, we are looking at the Diabetes Outcome.

```
ggplot(data = diabetes_data) +
  geom_point(mapping = aes(x = Age, y = BloodPressure)) +
  geom_smooth(mapping = aes(x = Age, BloodPressure)) +
  facet_wrap( ~ Outcome , nrow = 2)
```


0
1
BloodPressure
125
100
75
50
25
0
Age
20
40
60
80