MW-Exploratory Data Project

Mason Wadsworth

2024-02-27

1. Introduction:

I decided to use the Physical Activity & BMI data set to analyze. The data set measures the BMI and physical activity where 1000 steps is 1 Physical Activity (PA). There are 100 subjects with both BMI and PA recorded. I wanted to answer the question: Is there a connection between BMI and physical activity. I want to analyze any visible correlation between these two variables.

This PABMI data set comes in the form of a downloadable text file. I will be labeling the data set “fit” for short.

```
"C:\\Users\\mason\\Downloads\\math130\\data\\PABMI.txt"
```

```
## [1] "C:\\Users\\mason\\Downloads\\math130\\data\\PABMI.txt"
```

```
fit <- read.table("C:/Users/mason/Downloads/math130/data/PABMI.txt", header=TRUE, sep="\t")
```

```
summary(fit)
```

```
##     SUBJECT             PA              BMI
##  Min.   :  1.00   Min.   : 3.186   Min.   :14.20
##  1st Qu.: 25.75   1st Qu.: 6.803   1st Qu.:21.10
##  Median : 50.50   Median : 8.409   Median :24.45
##  Mean   : 50.50   Mean   : 8.614   Mean   :23.94
##  3rd Qu.: 75.25   3rd Qu.:10.274   3rd Qu.:26.75
##  Max.   :100.00   Max.   :14.209   Max.   :35.10
```

I will use ggplot to describe my data with gridExtra for more detailed comparisons.

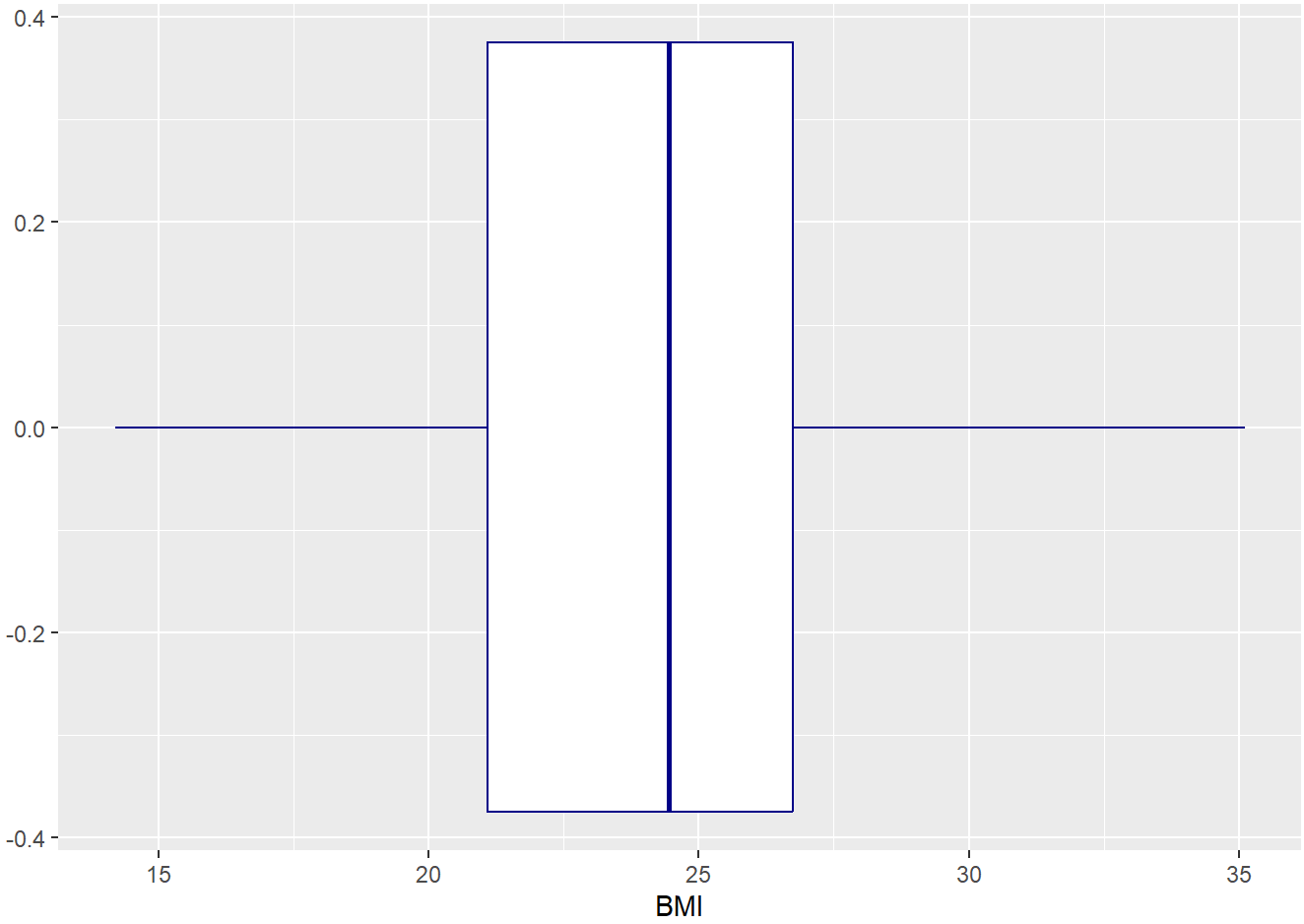
```
library(ggplot2)
library(gridExtra)
```

2. Analyzing BMI and PA Alone:

```
summary(fit$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.20   21.10   24.45   23.94   26.75   35.10
```

```
ggplot(fit, aes(x=BMI)) + geom_boxplot(color="darkblue")
```

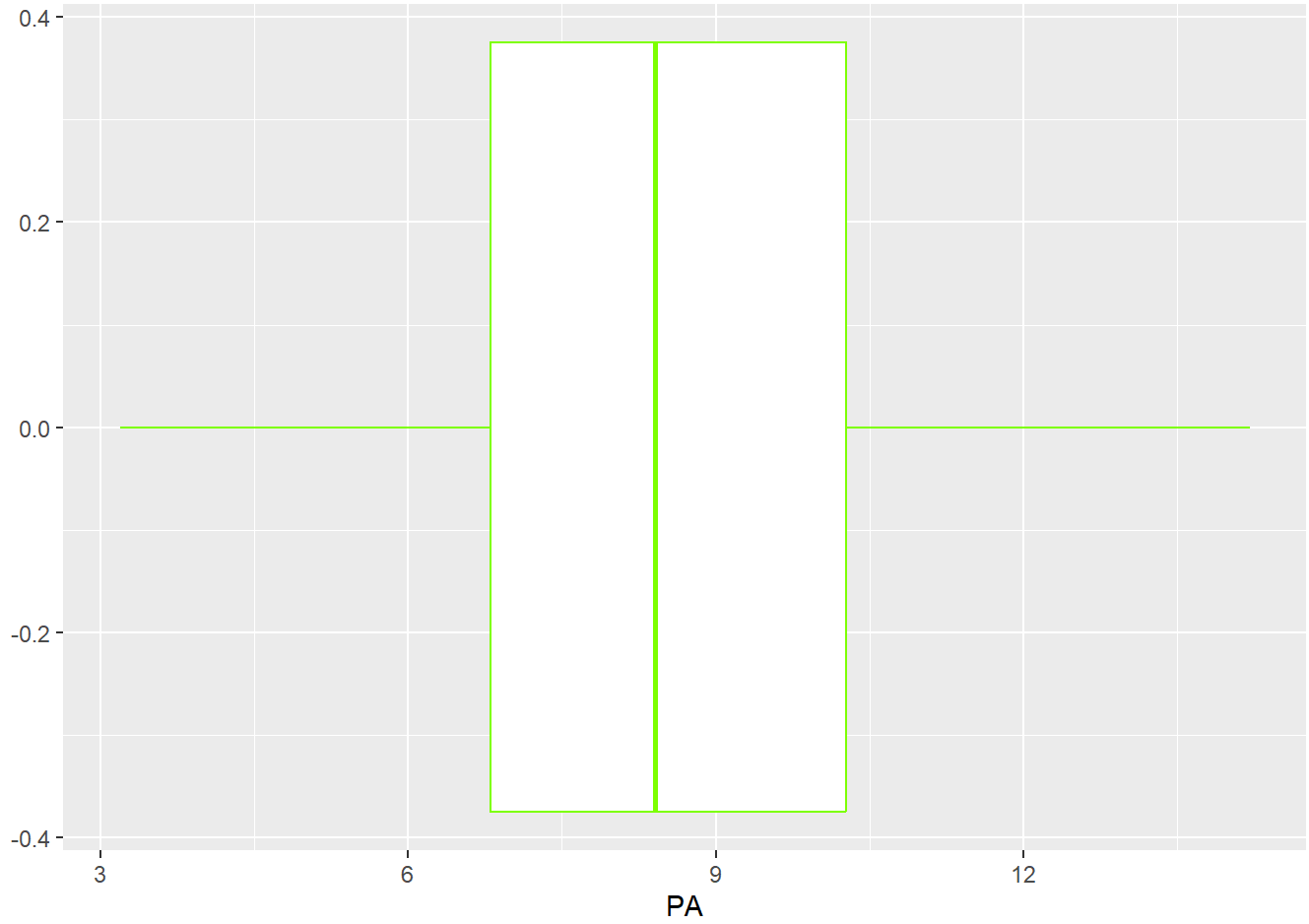



```
summary(fit$PA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.186   6.803   8.409   8.614  10.274  14.209
```

The BMI box plot and summary show that distribution is primarily centered around the median.

```
ggplot(fit, aes(x=PA)) + geom_boxplot(color="chartreuse")
```



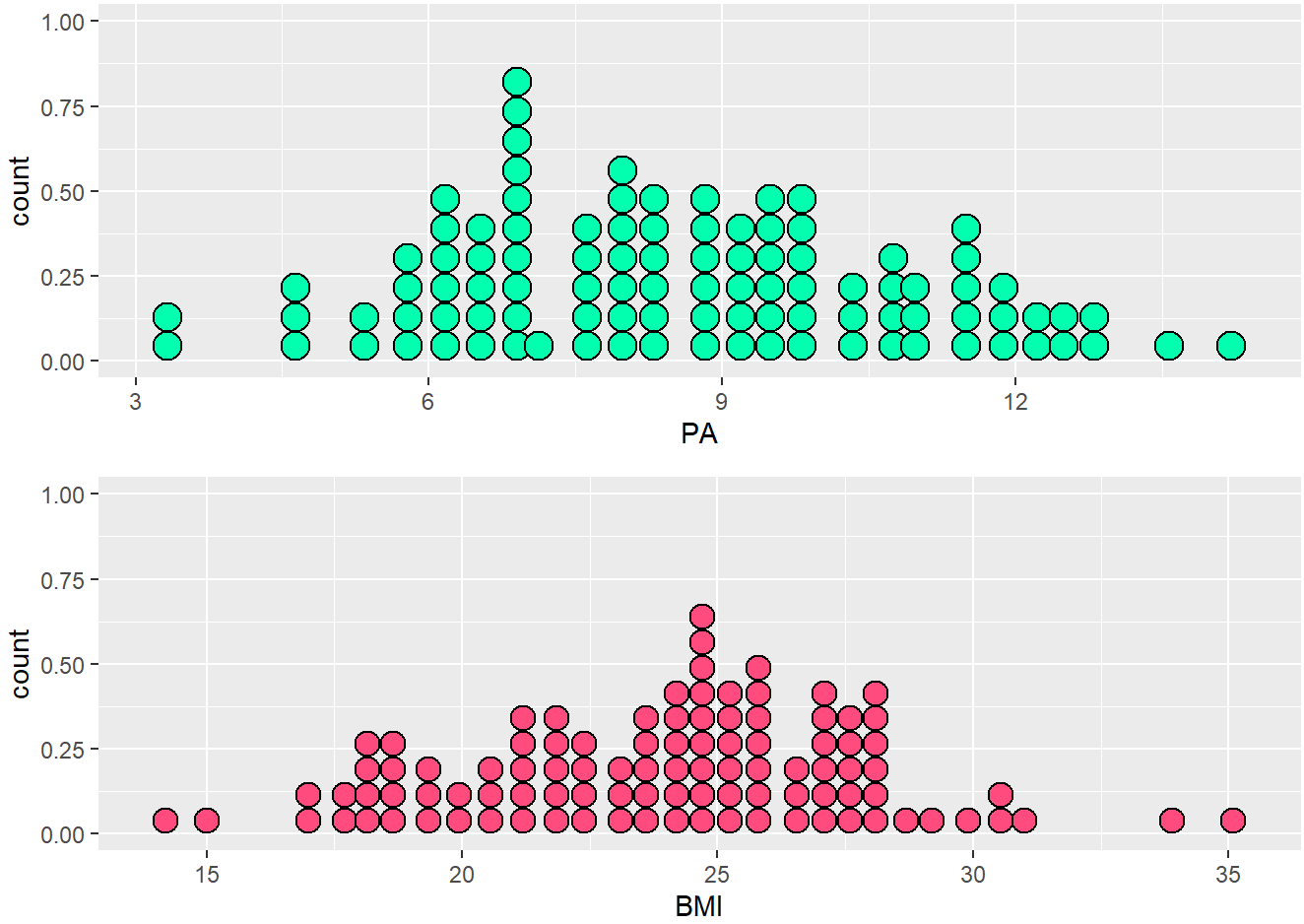

```
summary(fit$PA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.186   6.803   8.409   8.614  10.274  14.209
```

The PA box plot and summary show that there are few outliers and most data lies near the mean.

3. Comparison Between BMI and PA:

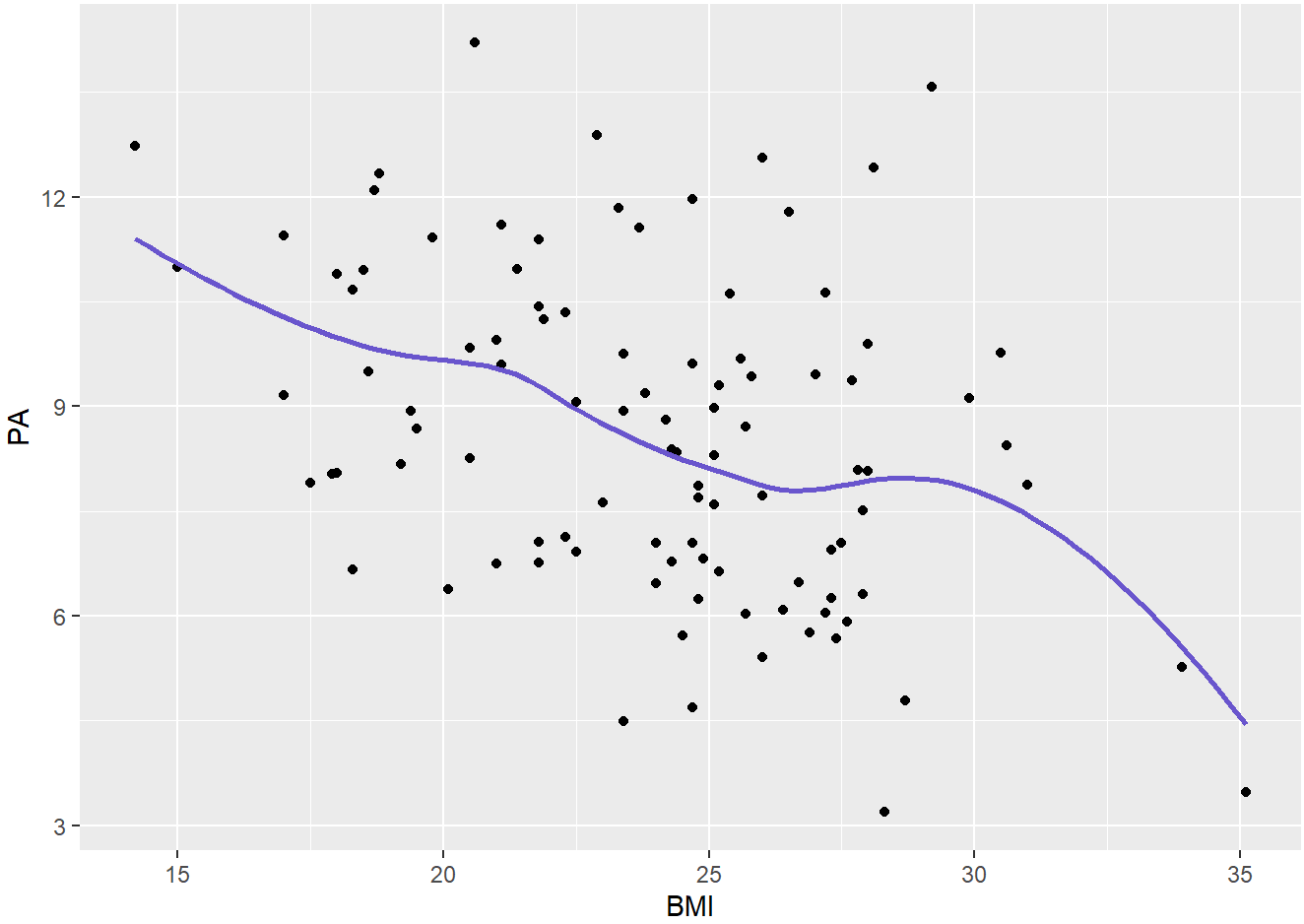
```
f1 <- ggplot(fit, aes(x=PA)) + geom_dotplot(fill=rgb(0,1,.7), binwidth=.3)
f2 <- ggplot(fit, aes(x=BMI)) + geom_dotplot(fill=rgb(1,.3,.5), binwidth=.5)
grid.arrange(f1, f2, nrow=2)
```



These dot plots show that distribution is not skewed one way or another and that the bulk of our population lies near the median. We can see both variables of interest and how they have similar distribution relative to each other.

```
ggplot(fit, aes(x=BMI, y=PA)) + geom_point() + geom_smooth(se=FALSE, color="slateblue3")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



We can see that nearly all the points are spread out near the middle whereas a few outliers are positioned around the edge with a negative trend. This scatter plot tells us how limited correlation is between most points as they form a near square shape in the middle of the plot.

4. Conclusion:

There appears to be some correlation between the two variables. While there are few extremes, the inverse relationship between BMI and PA is most visible in them. However, there is little correlation between the majority of recorded BMI and PA values. The data supports the hypothesis that there is correlation between increased physical activity and lowered body mass index but it does little to account for such a wide spread from the majority of measurements.