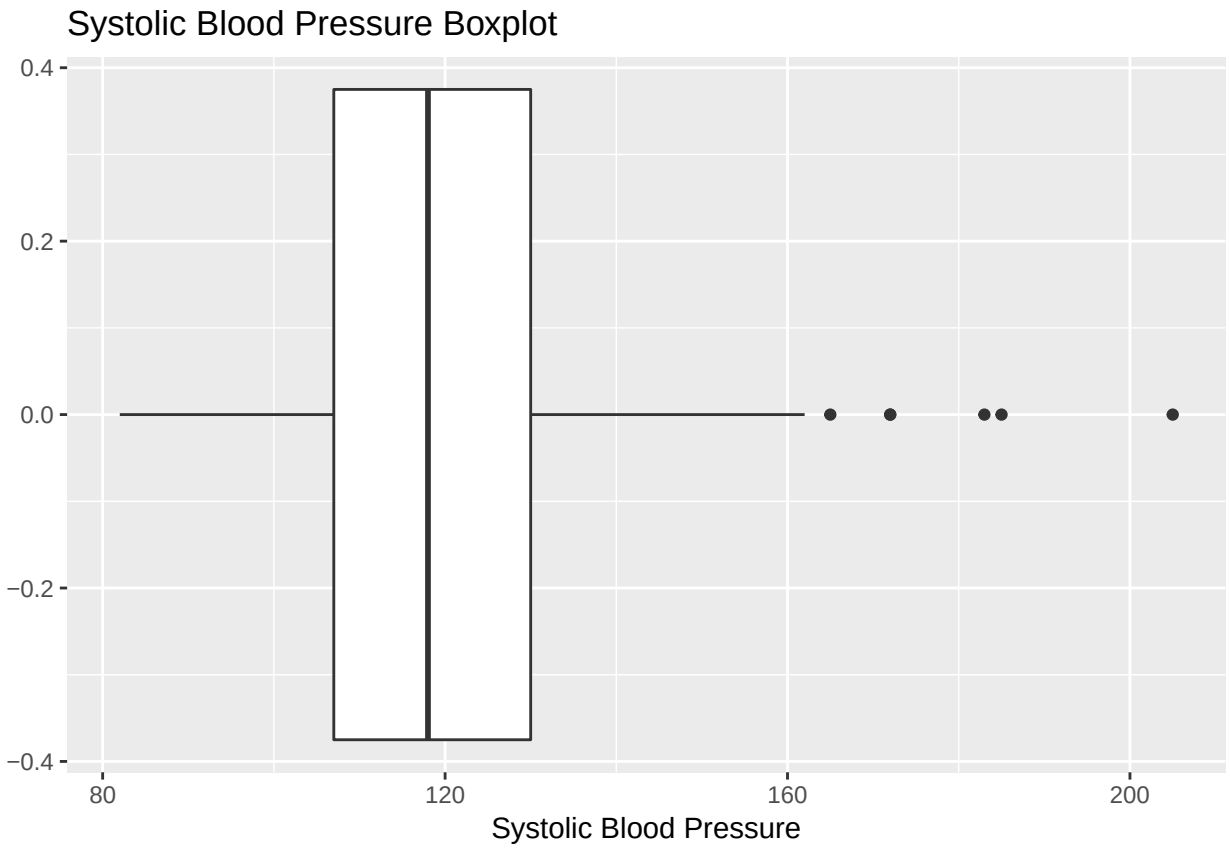
# 662 Hw2

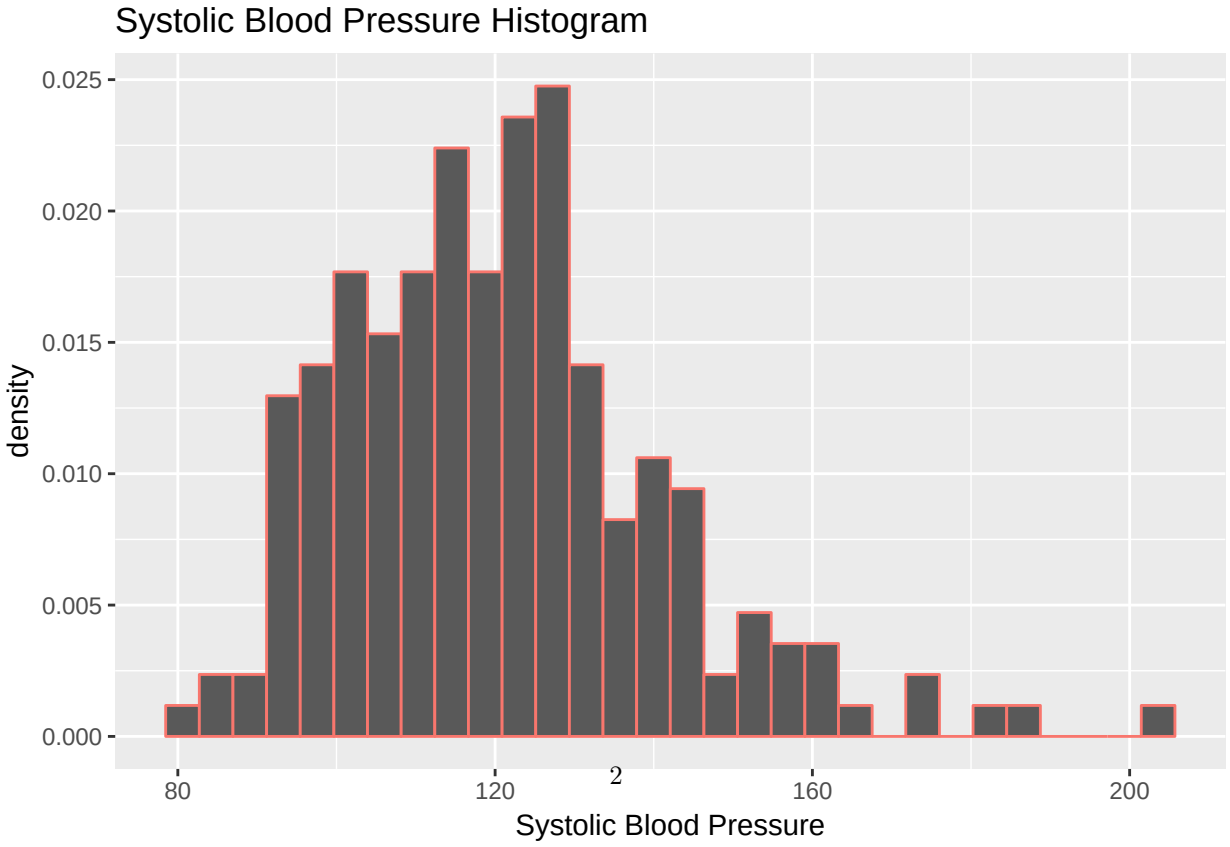
*Ty Darnell*

*September 6, 2018*

## Problem 1

a)


Systolic Blood Pressure Boxplot
0.4
0.2
0.0
−0.2
−0.4
80
120
160
200
Systolic Blood Pressure


Systolic Blood Pressure Histogram
0.025
0.020
0.015
0.010
0.005
0.000
density
80
120
160
200
Systolic Blood Pressure

## b)

25th, 50th, 75th percentiles

```
25% 50% 75%
107 118 130
```

## c)

IQR

```
[1] 23
```

## d)

25th - 1.5 IQR, 75th + 1.5 IQR

```
[1]  72.5 164.5
```

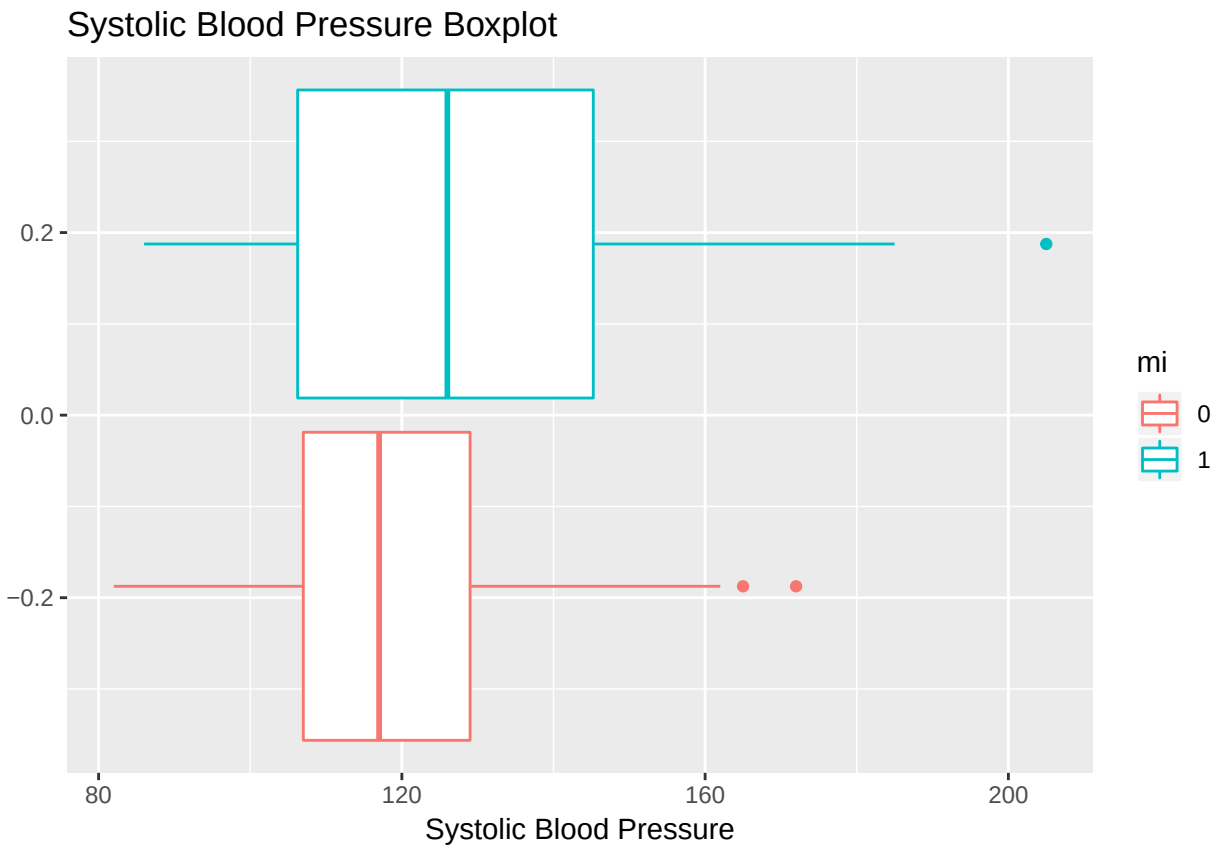
Smallest, Largest Nonoutliers

```
[1]  82 162
```

From the boxplot there appears to be 5 outliers. However in the dataset there are six ouliers (as shown below). However two of the values for `sbp` are the same so they are overlapping and thus not visually apparent. Hence the boxplot does agree with the defintion from our notes.

```
# A tibble: 6 x 2
  mi      sbp
  <fct> <int>
1 0       165
2 0       172
3 1       185
4 1       205
5 1       172
6 1       183
```

e)



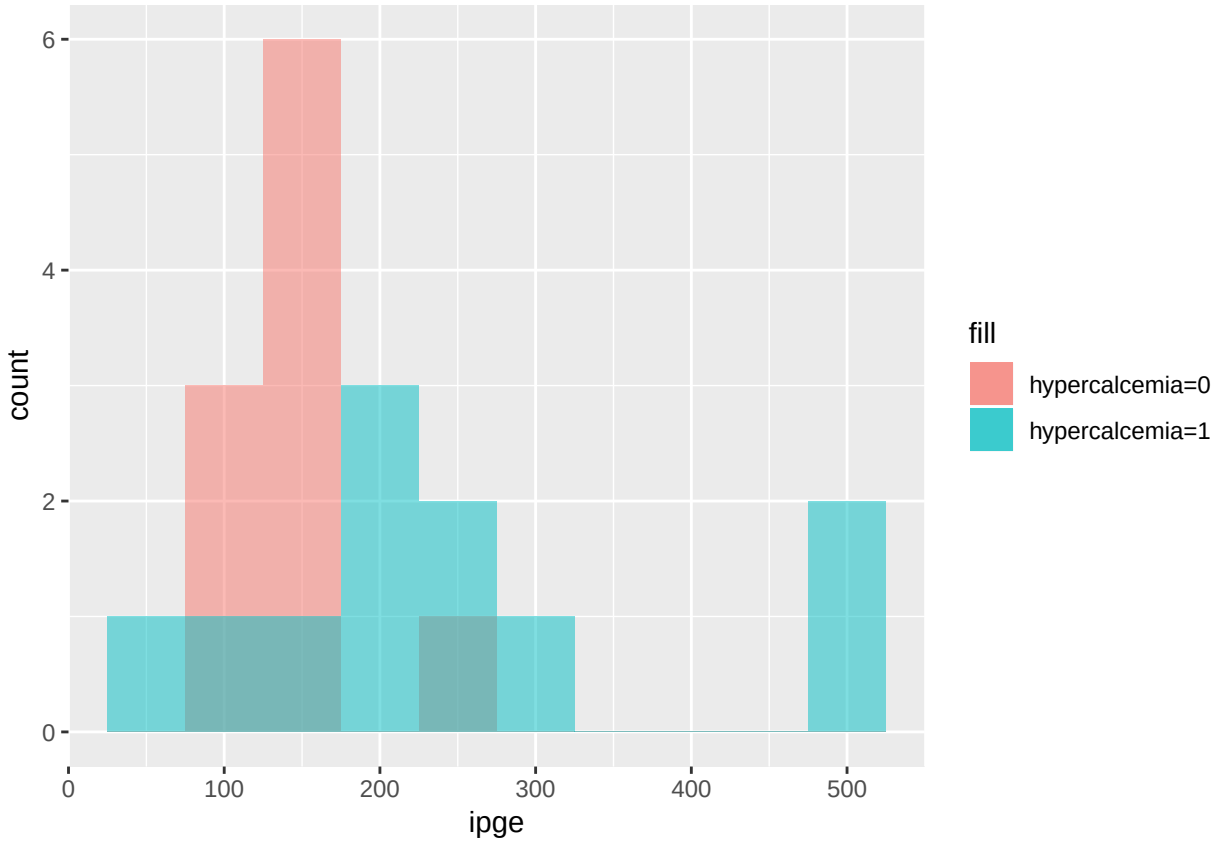

From the plot, blood pressures of the group who had an MI appears to be higher. The median of the MI group is almost equal to the 75th percentile of the group without MI.

# Problem 2

## a)

Looking at the means and standard deviations, they appear to be very different. However looking an overlaid histogram of `iPGE` separated by `hypercalcemia` we see that there are two extreme outliers that are impacting the mean of the `hypercalcemia`= 1 group. Based on this, I do not think that there is enough evidence to conclude that the means of the two groups differ significantly.

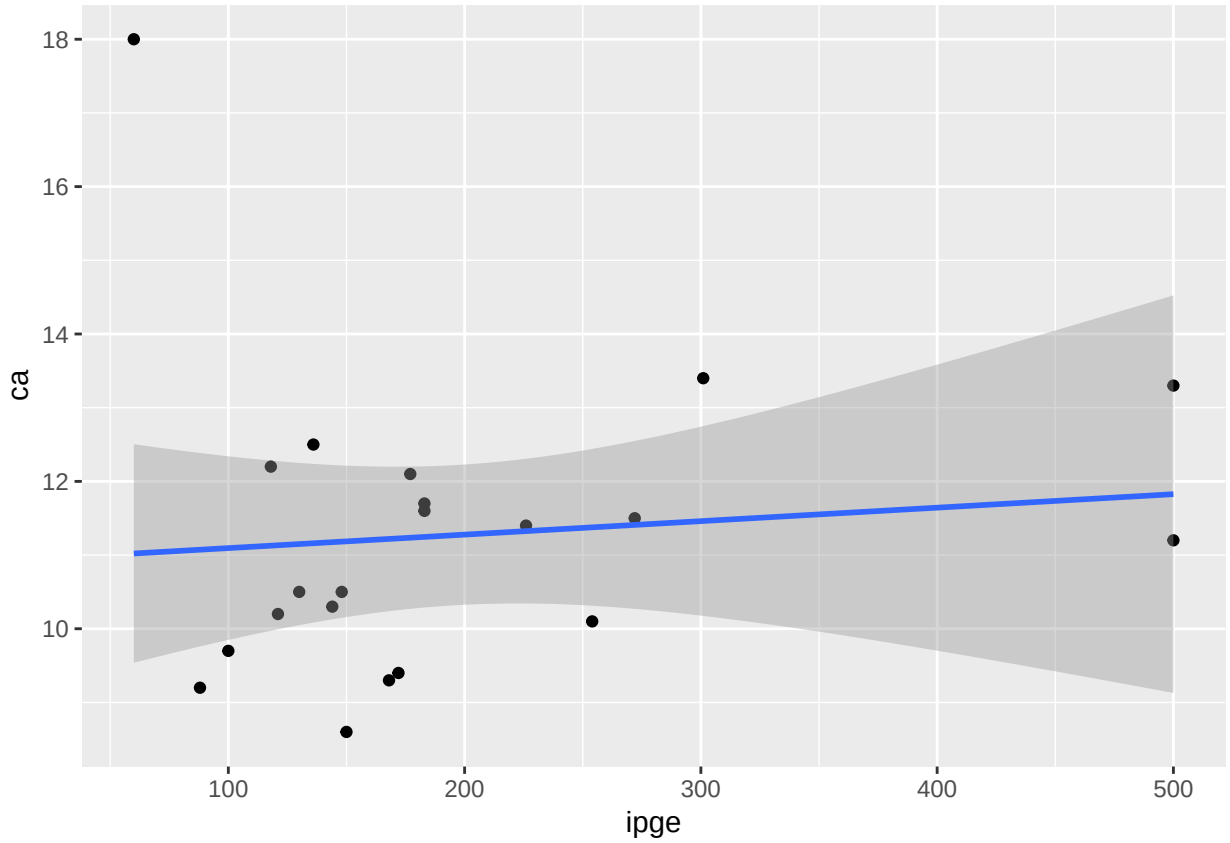
```
# A tibble: 2 x 3
  hypercalcemia `mean(ipge)` `sd(ipge)`
  <fct>                <dbl>      <dbl>
1 0                     148.       46.2
2 1                     241.      144.
```




## b)

Since the correlation between `iPGe` and `ca` is .105 there does not appear to be a strong association between the two variables.

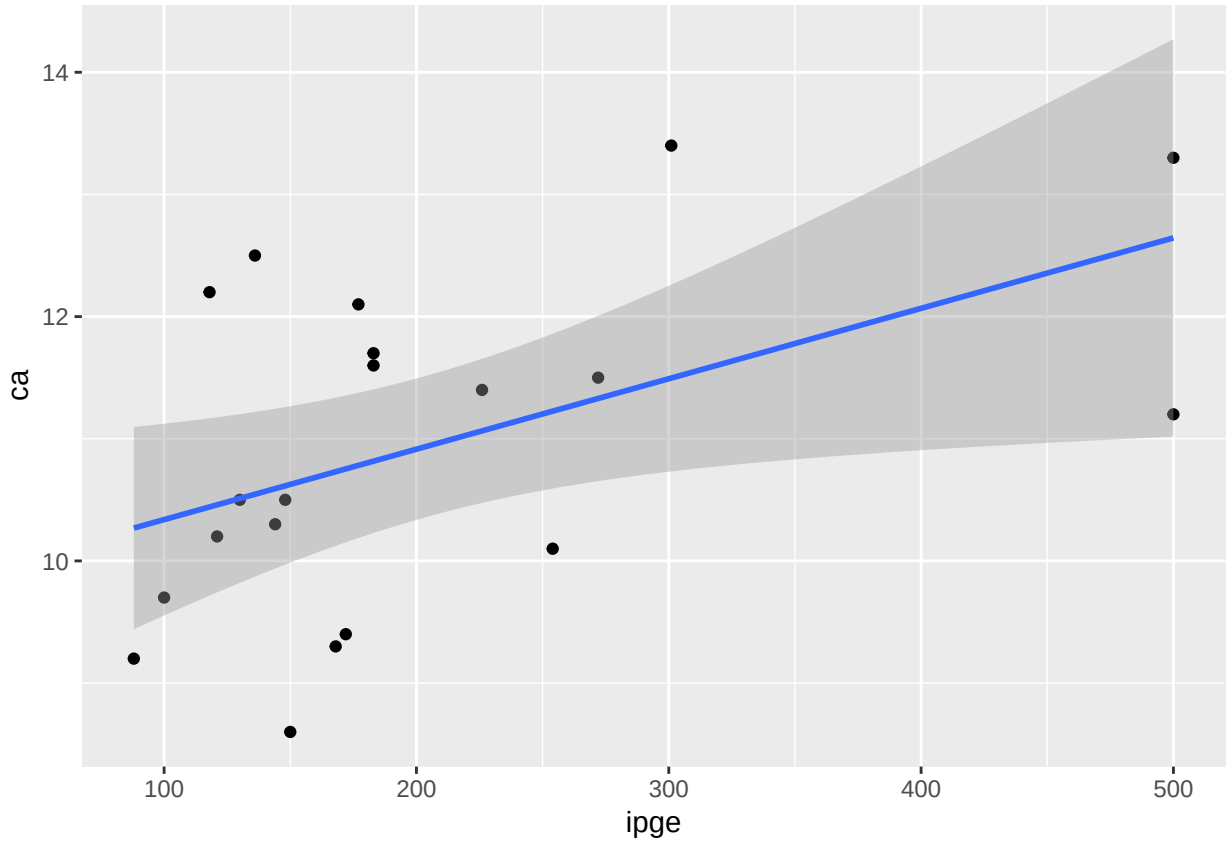
```
# A tibble: 1 x 1
  correlation
        <dbl>
1       0.105
```




c)

`patient` 11 has a `ca` value of 18 and an `iPGE` value of 60. This is the highest `ca` value by far and the lowest `iPGE` value by far. This appears to be an anomaly. After removing this point and replotting the data, there appears to be a stronger association between the two variables. The correlation between `iPGE` and `ca` become .486 which is much a stronger correlation than before but still does not suggest a strong association between the variables.

```
# A tibble: 1 x 4
  patient  ipge    ca hypercalcemia
    <int> <int> <dbl> <fct>
1      11    60    18 1
```




```
# A tibble: 1 x 1
  correlation
        <dbl>
1       0.486
```

```
Call:
lm(formula = ca ~ ipge, data = remove)

Coefficients:
(Intercept)         ipge
   9.761447     0.005765
```

Creating a linear model of the new data set and using the lm (shown below) to find a better `ca` value for `patient` 11, we obtain 9.761447+.005765*(60)=10.10735. Thus the new `ca` = 10.1

```
Call:
lm(formula = ca ~ ipge, data = remove)

Coefficients:
(Intercept)         ipge
   9.761447     0.005765
```

## d)

The new value of `ca` would cause `patient` 11 to be part of the `hypercalcemia` $= 0$ data subset instead of the `hypercalcemia` $= 1$ subset since the new value for `ca` $< 10.5$ which is the cutoff point for `hypercalcemia`. This would increase the mean of `iPGE` and decrease the standard deviation `iPGE` of the `hypercalcemia` $= 1$ subset. Also for the `hypercalcemia` $= 0$ subset the mean of `iPGE` would increase and the standard deviation would decrease. However the effect on the `hypercalcemia` $= 1$ subset would be greater since `iPGE` $= 60$ is a much more extreme outlier in that subset.