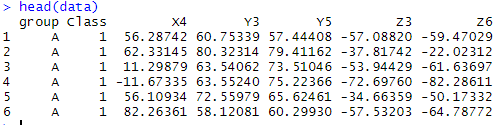
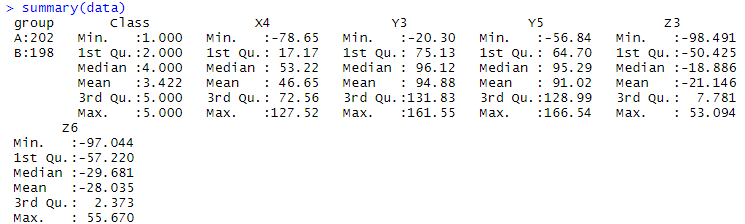
**Midterm Project**  
Benedikt Willecke  
IES19372

**Data Overview**  
The data describes the position of motion trackers while performing certain hand gestures(class1-5) by group A&B.

Load the data:  

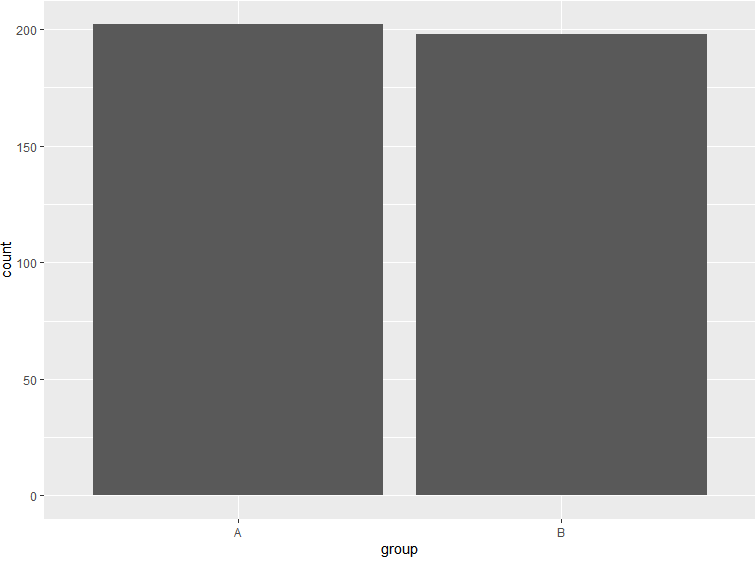


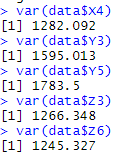

Here is a first look:  
  
We can see all the variables, the group and the class.

Summarize the data:  


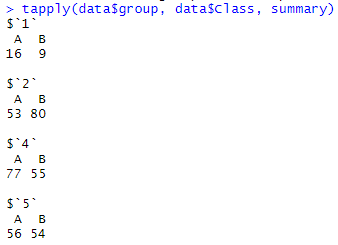
We can see many important metrics like mean, median, max and min value. We can also see that we have 202 data entries for group A and 198 for group B. So 400 in total.

We can also visualize this using a bar chart.

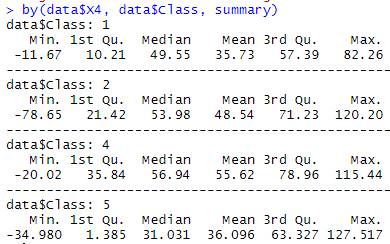
  


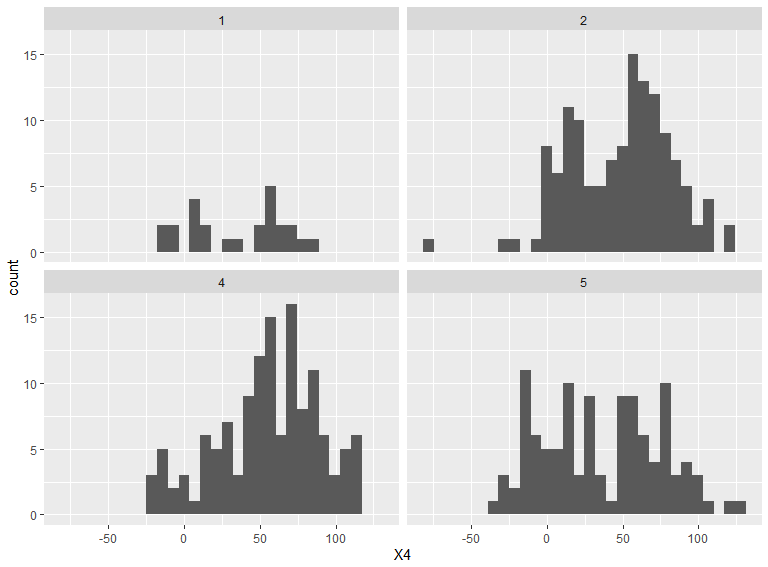


We can also observe that each variable (X4, Y3 Y5, Z3, Z6) has a lot of variance. This is possible because they have different values depending on the class.

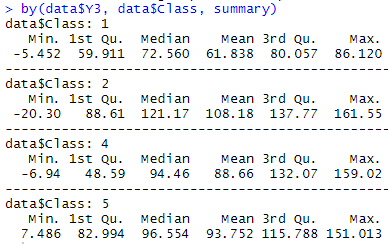


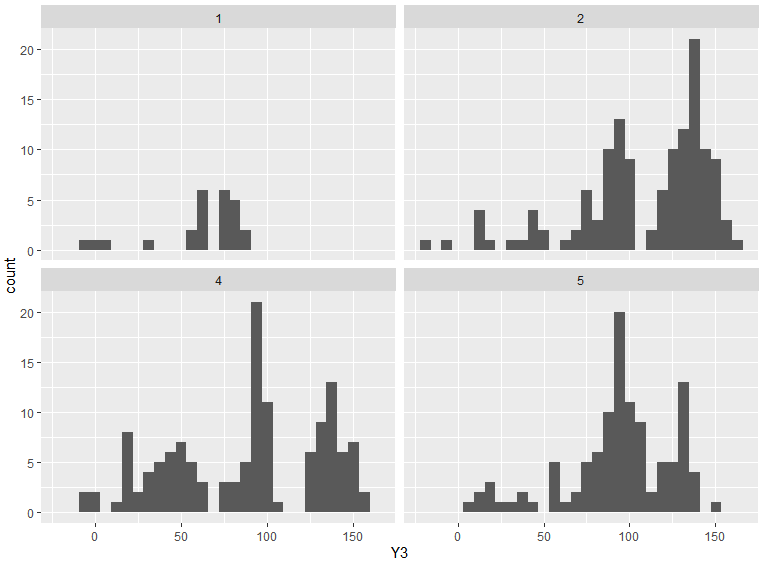
There also seems to be no data available for class 3. Also, it should be mentioned that for class 1 there is less data available than for the other classes. Furthermore, we can confirm that in general the amount of data available is not equally distributed among classes and groups. It goes from as little as 9 datapoints in class one group B to as high as 80 datapoints in class 2 group B.

Next we can get a summary for each variable by class.  


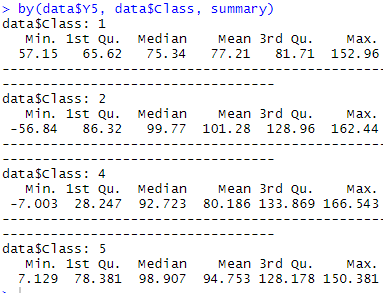
Here is the Histogram:  


By doing so we can get a feel for how high/low the values should be. Class 1, 2 and 4 seem to have very similar values for X4 with the values slightly lower in class 1 compared to the others. Class 2 has the overall lowest minimum value for X4 and Class 5 has the highest maximum value while having the lowest median. On should also note that Class two has the highest difference between the lowest and the highest value. One also should mention that in class 5 seems to be fairly equally spread out.



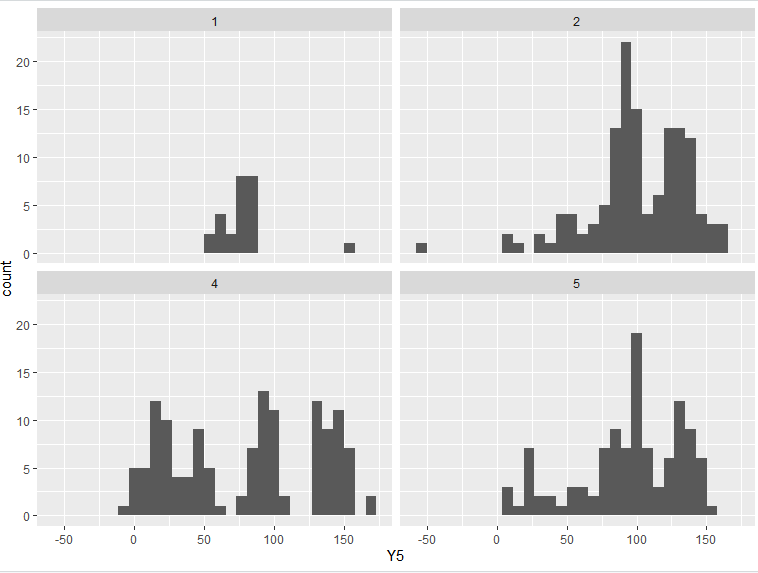
Here is the Histogram:  
  


Overall the values for Y3 seem to be much higher throughout all classes compared to variable X4. Class 2 has again the lowest minimum value and class 5 has again the highest maximum value. One should also note that Class two has the highest difference between the lowest and the highest value. Also the count of the values seem to spike at particular points while being fairly flat for other values.

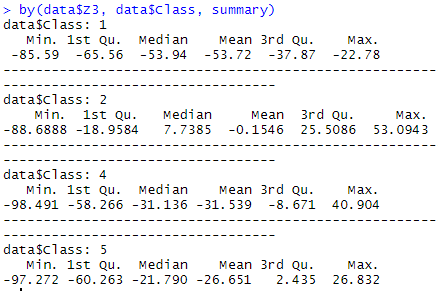


Histogram:



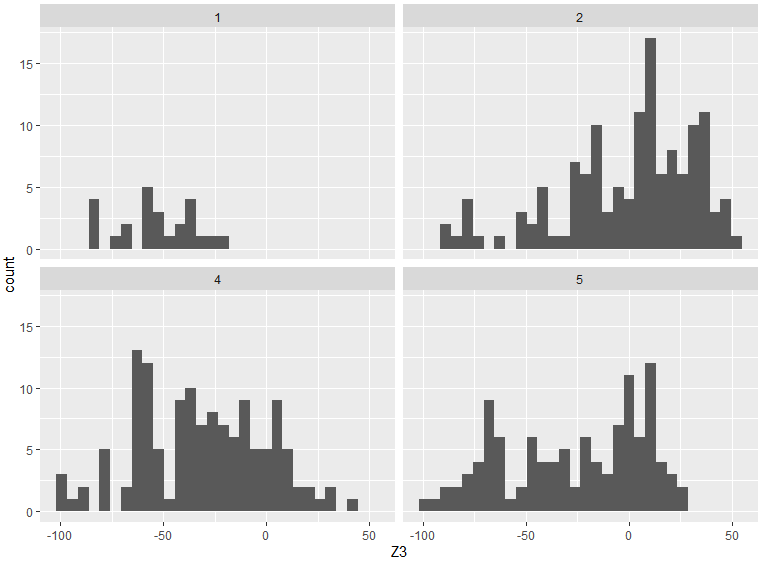


Overall the values for Y5 seem to be much higher throughout all classes compared to variable X4 as well. Class 2 has again the lowest minimum value and class 4 has again the highest maximum value. One should also note that Class two has the highest difference between the lowest and the highest value. Again we can see many spikes in frequency at certain points. This is especially visible in class 4. We can also see that class 1 is basically on spike between values of 50-100 with one outlier to the right. One might argue that class 2 and 5 are skewed to the right but this is not so clear as they have many local minima and maxima.

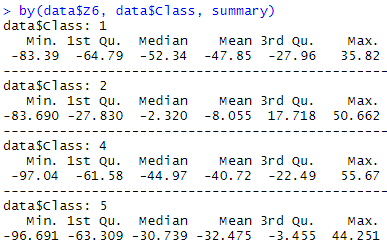


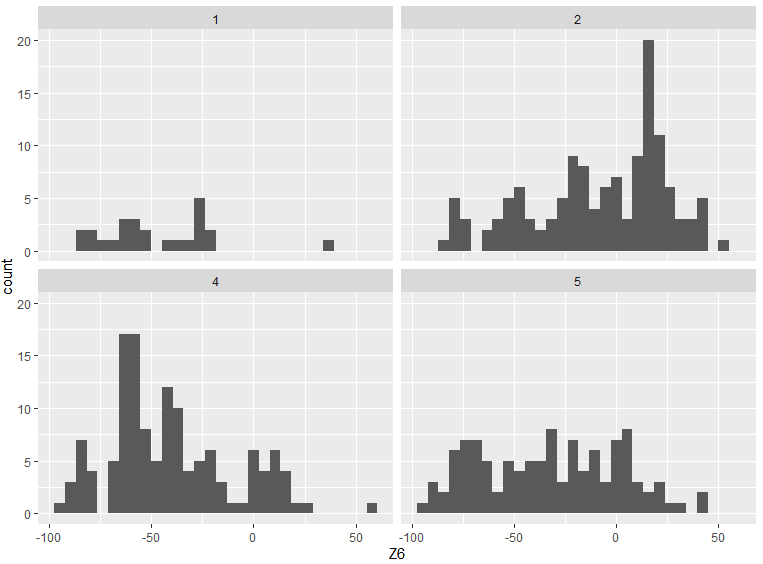
Histogram:





Here we can see that the values overall seem to be much lower than the previous two variables we looked at. In fact, they are negative to a large degree. The highest maximum value has Class two and the lowest minimum has class four. Class 4 seems to be skewed to the right and class 2 to the left while class 1 and five seem to be somewhat uniform with regards to the distribution. We can see again here that class 1 does not seem to have many datapoints and it very dense at one point.



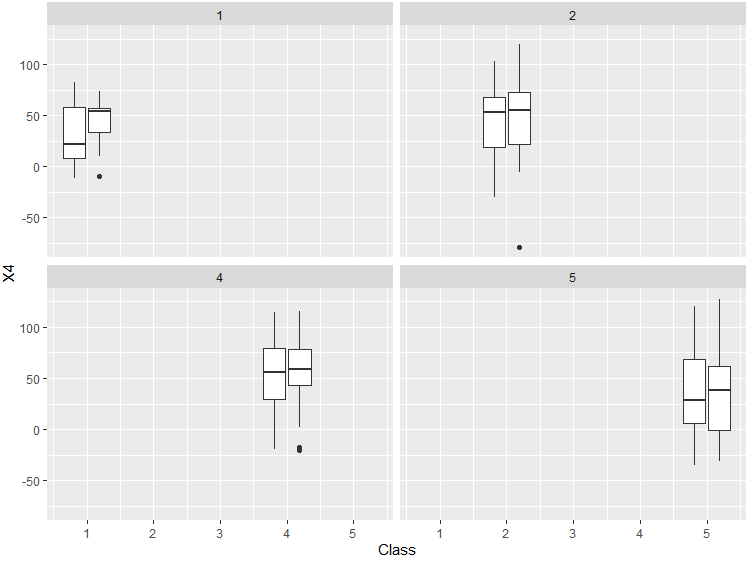
Histogram:  
  


Again, we can see that the values overall seem to be much lower than the previous Y variables we looked at. In fact, they are negative to a large degree like Z3. The highest maximum value has Class four and the lowest minimum has class four as well and thus the largest difference between minimum and maximum. All classes are very stretched out with class 5 being very uniform in its distribution. Class two seems to be skewed to the left and class 4 to the right. For class 1 it seems to be uniform but it is impossible to tell as there are so few datapoints.

**Boxplot**

Now we can compare the distributions per class per group graphically using boxplot:

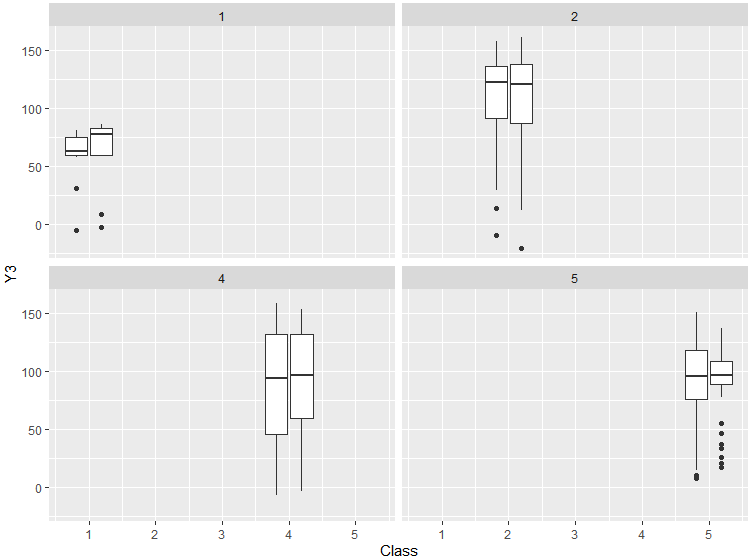
For variable X4:  

One can see for variable X4 in class 2 and 4 the median is almost the same and in class 5 there is just a slight difference. In class one however the is a very big difference. This might be due to the low amount of datapoints available as well as the amount of data available in group A compared to group B. In class 1 group B is heavily skewed to the left while group A is quiet strongly skewed to the right. Also, in group B the distribution is much more condensed. In class 2 in both groups the distribution of X4 is skewed to the left. In class 4 both have a fairly equal distribution while group B is slightly more condensed. In class 4 group A seems to be skewed a bit to the right and group B slightly to the left. There are a few outliers in class 1,2 and 4 and they are all in class B.

For variable Y3:

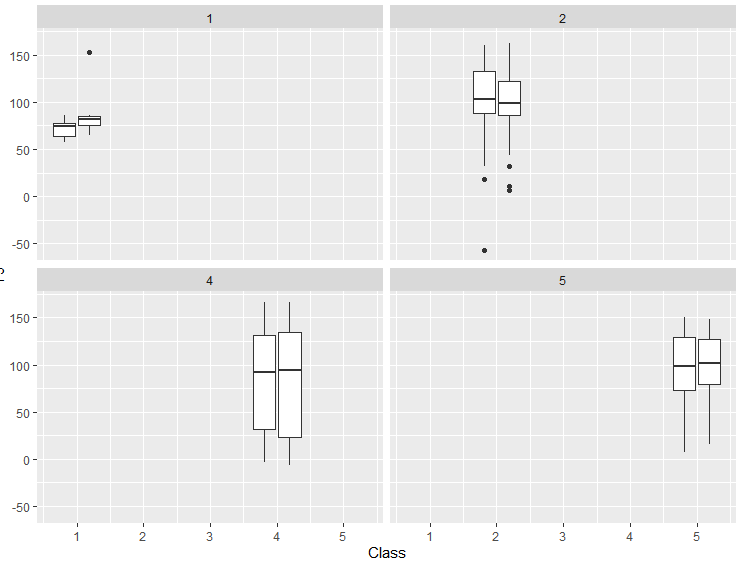




At first glance one can notice that there are a lot of outliers to the left side in both groups except in class 4 and especially in class 5. All distributions seem fairly similar when comparing both groups with each other. In class one the distributions of both groups are very condensed. Here group A is strongly skewed to the right while group B is strongly skewed to the left. In class 2 and 4 there is just a slight difference between the groups. Class two is skewed to the left and class 4 seems equally distributed. In class 5 both distributions are pretty condensed while group B is very condensed here in terms of distribution while also having a lot of outliers. The distributions are both pretty equal however. The distribution of class 4 seems to be very stretched out.

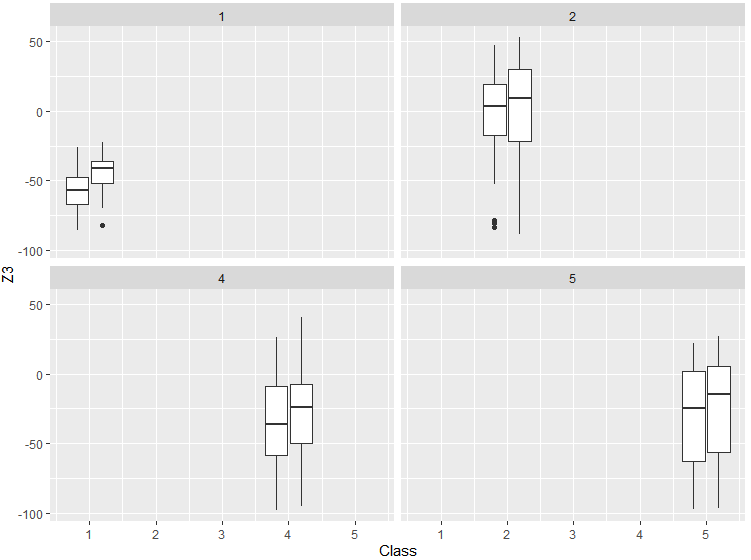
For variable Y5:



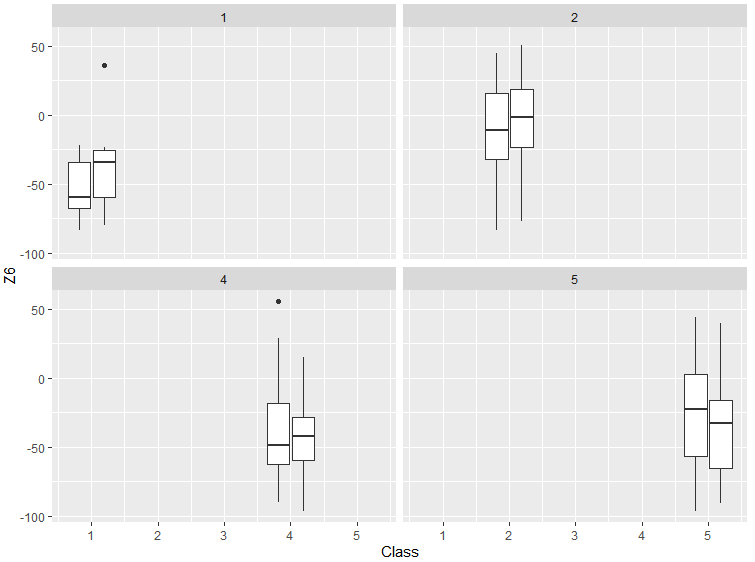


Like variable Y3 in class one has a very condensed distribution. Group B even more so while having a positive outlier as well. Class 1 seems to be skewed to the left and class 2 skewed to the right while class 4 and 5 have almost an equal distribution. The distribution of class 4 seems to be very stretched out. Furthermore, there are many outliers to the left in class 2.

For variable Z3:  

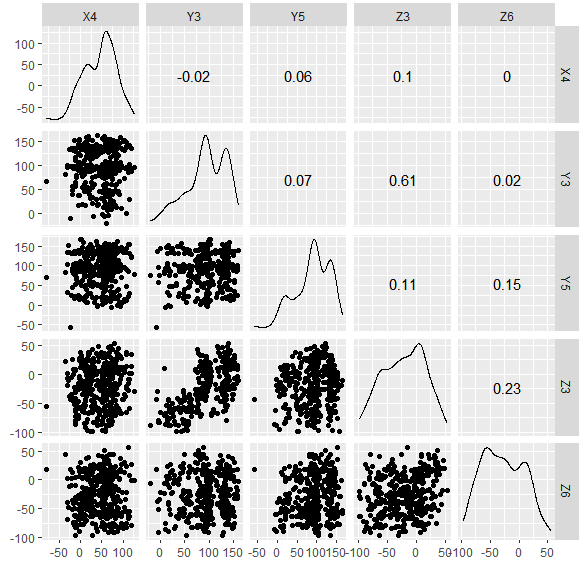
As previously observed the Z variables’ distribution is mostly in the negative area. Class 1 again has a very condensed distribution while class 2, 4 and 5 are fairly stretched out. Here the median always differs a little with group B having a higher median throughout all classes. All classes seem to have a equal distribution with a slight skew to the left in class 1, 4 and 5 of group B and group A only having a slight skew to the right in class 4. Group B has one negative outlier in class 1 and group A has a few in class 2.  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
For variable Z6:  

Again, all distributions are mostly in the negative area while all distributions are fairly stretched out. In class 1 we can observe a big difference in skew. Here group A is strongly skewed to the right and group B strongly to the left. And group B has an outlier to the right as well. In class 2 there is almost no difference between the groups. Both groups are pretty much equally distributed.

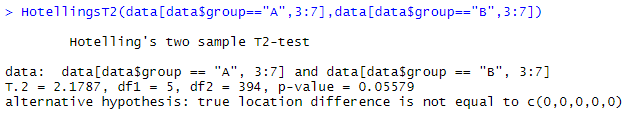
**Scatter Matrix**

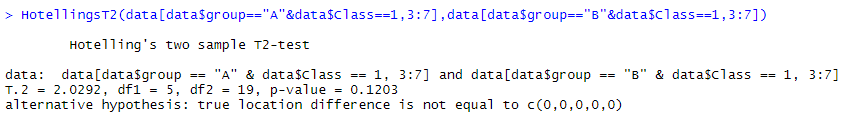
Now that we analyzed the relationship between the class, group and the variables we can further analyzed the data by investigating if the variables have any correlation with each other or not. This can be done with a scatter matrix.

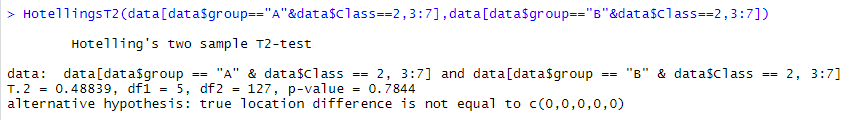
First one can see the skewedness to the left of the two Y variables. Also one can see that the Z variables are very much in the negative numbers while also being pretty equally spread out. The X variable on the other hand seems to be somewhat equally distributed while slightly condensed. There seems to be no correlation between the variables except a strong positive correlation between Y3 and Z3 of 0.61 and a slight positive correlation between Z3 and Z6 of 0.23, Y5 and Z6 of 0.15 and Y3 and Z5 of 0.11.

**Hotellings T2**

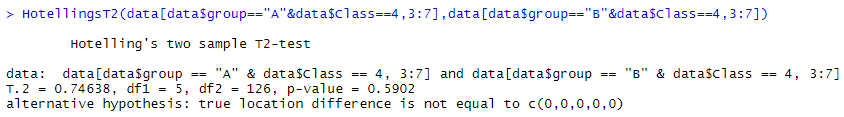
Now I would like to analyses if there are any differences between the groups and classes.  


First with this HotellingsT2 test we can see if overall the mean of group A and B is the same. With 95% confidence we can say that there seems to be no difference in the overall mean. We cannot reject H0 stating that there is no difference. At the 0.1 significance level we can reject H0 however.  


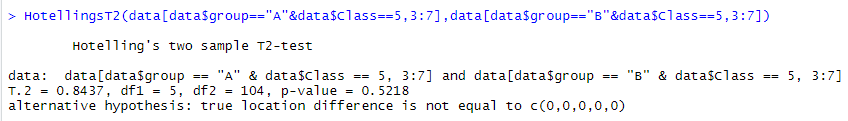
Next we can see if there is a difference between class 1 of group A and B. This seems not to be the case. At least we cannot reject H0 stating that the means are same with 95% confidence.



Here we can see if there is a difference in the class 2 between the two groups. This does not seem to be the case. We cannot reject the Null hypothesis.

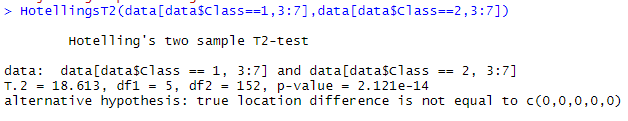


The same case is with class 4. Here we can say with high confidence that there seems to be no difference in the two means.

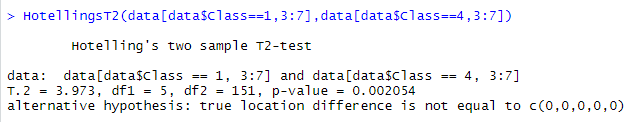


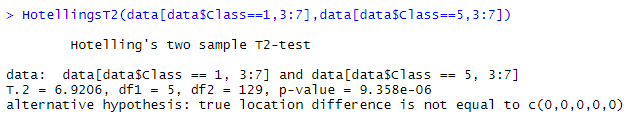
Lastly the same case is with class 5.  
  
Thus we can conclude that there is no significant difference between group A and B.

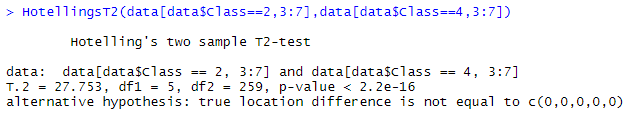
This means that we don’t have to distinguish between group A and B when we do a comparison between the classes. As the hand positions in each class is different one would expect the means of the classes to be different. We can test this using the following commands:  
Class 1:

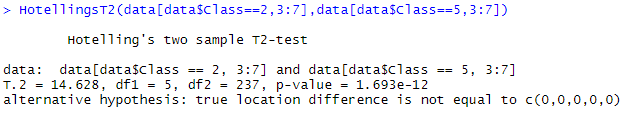


With this comparison between class 1 and 2 we can reject H0, there is a difference.

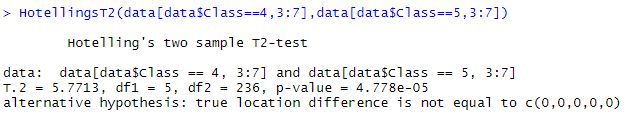




With a 0.05 significance level this is also the case when comparing class 1 with class 4 and 5.  
  
Class 2:  




This is also the case for class 2 in comparison with class 4 and 5 (and class 1 from the previous batch of comparisons).

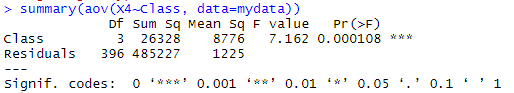
Class 4 & Class 5:  


Class 4 and 5 also seem to have a different mean.

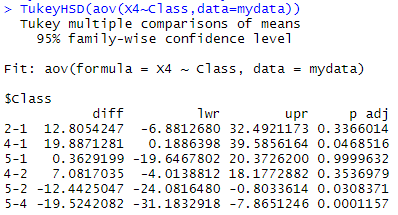
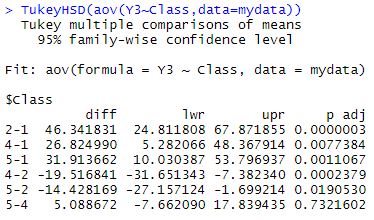
This we can draw the conclusion that all the classes are quite different. Probably due to the different hand position.

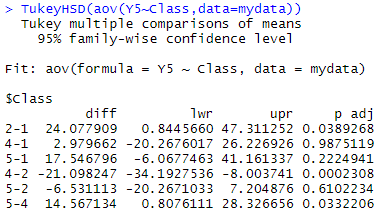
**TuckeyHSD**

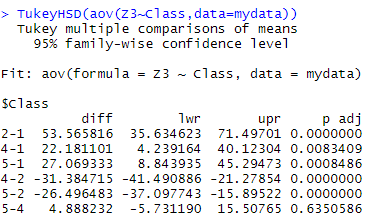
Next I would like to compare the variables per class for difference. For this I have to first convert the Class variable from numeric to categorical.  

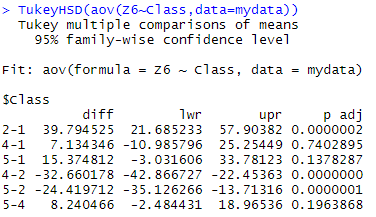

With ANOVA we can only test if there is a difference between the classes, but we don’t know which class.  


Here we can reject H0, thus there seems to be a difference but we cannot know between which class by using ANOVA.  
Then I can test for difference on variable after another throughout all classes using the TukeyHSD.  
TuckeyHSD gives more detailed result regarding to which classes are different.

X4:  
  
One can see that the value is very similar in class 1 and 5. And in class 2 and 4.  
Y3:  


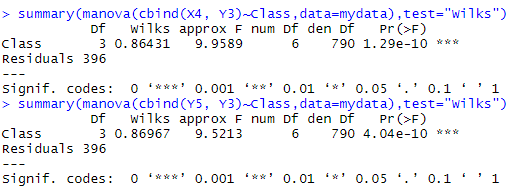
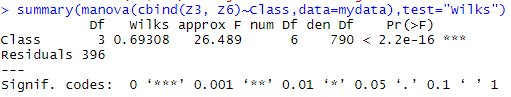
Here class 5 and 4 are very similar. And class 1 and 2 have different values respectively.  
Y5:  


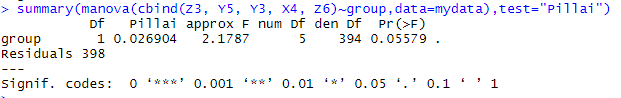
Here class 4 and 1 seem to be similar. And class 5 and 2.  
Z3:  


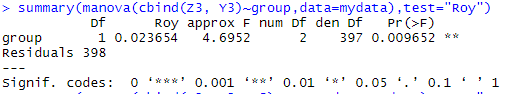
Similar like variable Y3 here 5 and 4 seem to have similar values but 1 and 2 are both different and different from each other.  
Z6:  


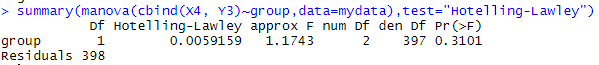
Here class 1, 4 and 5 seem to be pretty similar while class 2 having a different value.  
  
In conclusion one can notice that for variable Y3, Z3 and Z6 at least class 4 and 5 seem to be similar.

**MANOVA**

With MANOVA we can also check and confirm the difference of variables between classes.  
Here are a few examples:  
  
  
We can see that at all three MANOVA we can reject H0 as there seems to be a difference between the classes.

Instead of Wilk’s Lambda we can also use other methods:  


Here we can see that when looking at all variables we cannot reject H0(H0 = no difference in mean). Here I used Pillai’s Trace.  


When looking at only Z3 and Y3 we can reject H0 however. Here I used Roy`s greatest root. This difference in variable mean per group might be due to the fact that the different groups did different hand gestures to a different degree. We saw in the beginning that group A did certain hand gestures more often than B did. This could lead to differences in mean when looking at a certain variable.   


For other variables though we are not able to reject H0. Here I used Hotelling Lawley Trace.  
All the methods while being different usually give very similar results.

**Summary and Conclusion**In summary one can say that there is not enough evidence to support the claim that there is a significant difference between group A and B. Due to different hand postures in the different classes there is however a significant difference in the position between the classes. Some classes as mentioned before have a few similar traits, like having the same variables be similar to each other or having certain variables strongly in the negative or positive. We can also detect a difference in the variables per group. But this might be due to the not equally distributed amount of data points.