Project 3.0

Gavin Gunawardena

Background

Over the summer Dr. Christopher Saunders was supporting a chemist who was developing a method using LC-MS/MS (see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3052391/ for background on liquid chromatography-tandem mass spectrometry (LC-MS/MS).) to analyze the chemical make up of various brands of aspirin.

At the current stage of development we are simply interested in whether or not there is a difference between the three different brands of aspirin that we have analyzed to date.

We have the following data set- for each of seven peaks.

- Three pills from three different brands of aspirin. (Bayer, PV. and Wall)
- Each Pill has been analyzed in triplicate- this means that we measured the same pill three times for one peak.

We want to know if there is a difference between the the three brands of pills for each peak.

Please see the accompanying video where I will describe the data set and what we want you to look at.

Project

Analyze the corresponding data sets and give an appropriate data analysis

Objective

The objective of this project is to take the 63 tab separated text files of aspirin test data, analyze them in R, and test the null hypothesis that the 3 brands of aspirin are the same with a 95% confidence interval. I plan on doing this by utilizing R to pull the data into a dataframe with the columns: Brand, Peak, Replicant, Aliquot, Time, and Intensity. Once this is done, I can more easily navigate and group the data, allowing for easier analysis of it via visualizations and summary functions. Then, I plan on utilizing summary functions to measure the Area Under a Peak (AUP) for each aliquot of each replicant. Finally, I plan on utilizing Analysis of Variance (ANOVA) functions in R to compare the AUP between brands for each peak in order to confirm or deny the null hypothesis.

For some further context, here is what some of the less obvious variables I mentioned in the last paragraph mean, and their significance to the dataset. Peak is when the intensity of each pill is at its highest and usually indicates when a chemical or combination of chemicals is released from the pill. This usually happens in artificially timed intervals. Replicants

indicate replicas of the same pill, as 3 replicas of 3 different brands of aspirin are tested in this dataset. Aliquot indicates a series of measurements within a timeframe of less than a second of a replicant of a pill. There are 3 aliquots for each replicant, 3 replicants for each peak, and 7 peaks for each brand in this dataset. Finally, intensity is the measured intensity of the chemical being released and time is the time in seconds since the pill activated.

Assumptions made for this project include that the scientists who supplied the data utilized a standardized process for attaining the measurements. They also include that there's the possibility for error within the process utilized to attain the data, and thus I'll be attempting to correct for it by first removing missing data and also by running the comparison on a version of the dataset with outliers and on a version without outliers. Also, I'll be skipping over the "blank time" and "blank intensity" columns due to lack of information on what they entail.

1. Format the data contained in the txt files for use in R. (Make sure to document all of the steps that you have done to prepare the data sets. You should turn in a new data set if you manually edited the data before loading it in R.)

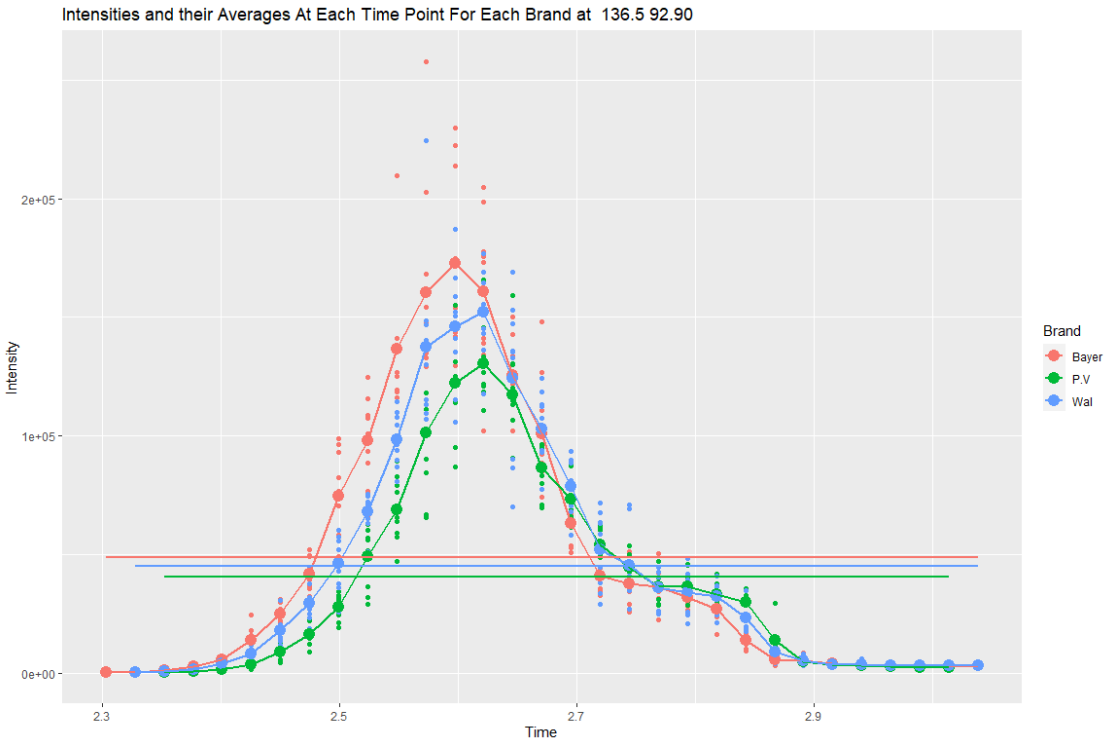
```
##           [,1]
## Brand     0
## Peak      0
## Replicant 0
## Aliquot   0
## Time      0
## Intensity 0

## [1] 0

##     Brand               Peak             Replicant       Aliquot
##  Length:4798        Length:4798        Min.   :1.00   Min.   :1.000
##  Class :character   Class :character   1st Qu.:1.00   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median :2.00   Median :2.000
##                                        Mean   :1.99   Mean   :2.003
##                                        3rd Qu.:3.00   3rd Qu.:3.000
##                                        Max.   :3.00   Max.   :3.000
##       Time          Intensity
##  Min.   :2.303   Min.   :    10
##  1st Qu.:2.548   1st Qu.:   270
##  Median :2.720   Median :   785
##  Mean   :2.717   Mean   : 10071
##  3rd Qu.:2.867   3rd Qu.:  2590
##  Max.   :3.161   Max.   :257870
```

Here I loaded all of the data into a dataframe, Data.All.df, in order to make it easier to process and analyze since I would then by able to group and aggregate the data similarly to a relational database table. This was done by first making sure each file follows the same naming conventions and adjusting the file names if they did not, recursively looping through each file, and pulling info from the file name as well as the file contents to populate the dataframe. Afterwards, I updated the datatypes of any numerical values of the dataframe so that they were numerical in order to allow for calculations and visualizations.

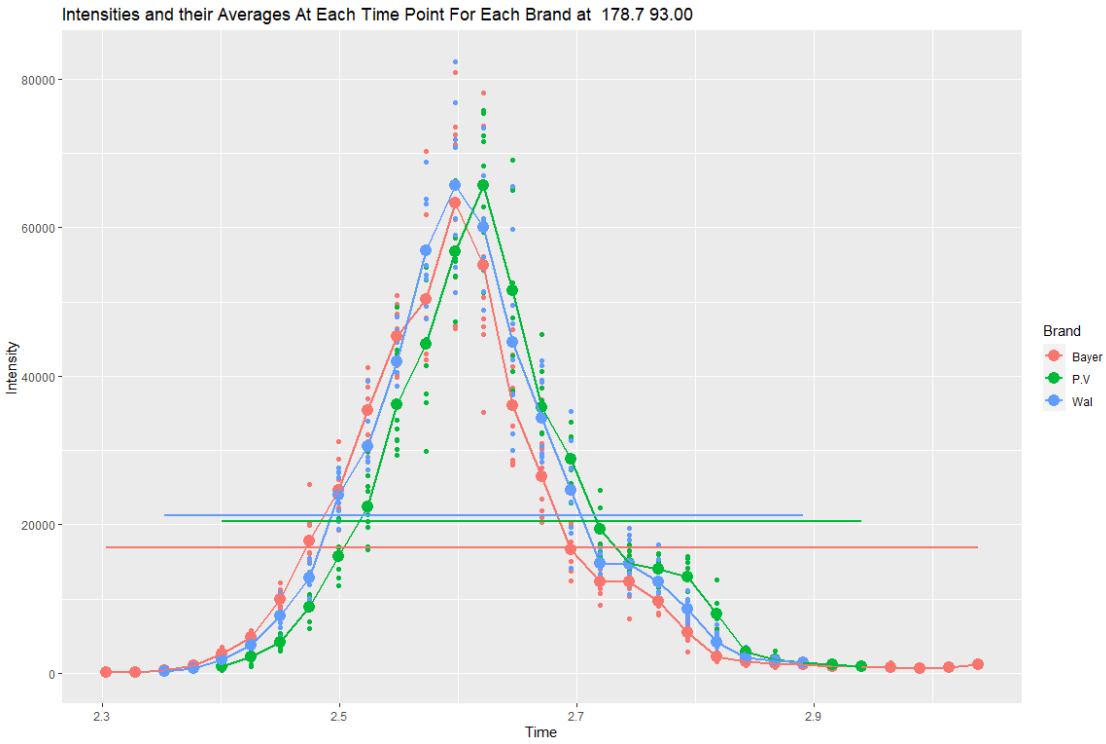
Finally, I checked for any null values or duplicate values that could indicate issues in the dataset or my data processing. Luckily, neither of these issues were found. Above is a summary of the dataframe.

2. Perform an exploratory data analysis.

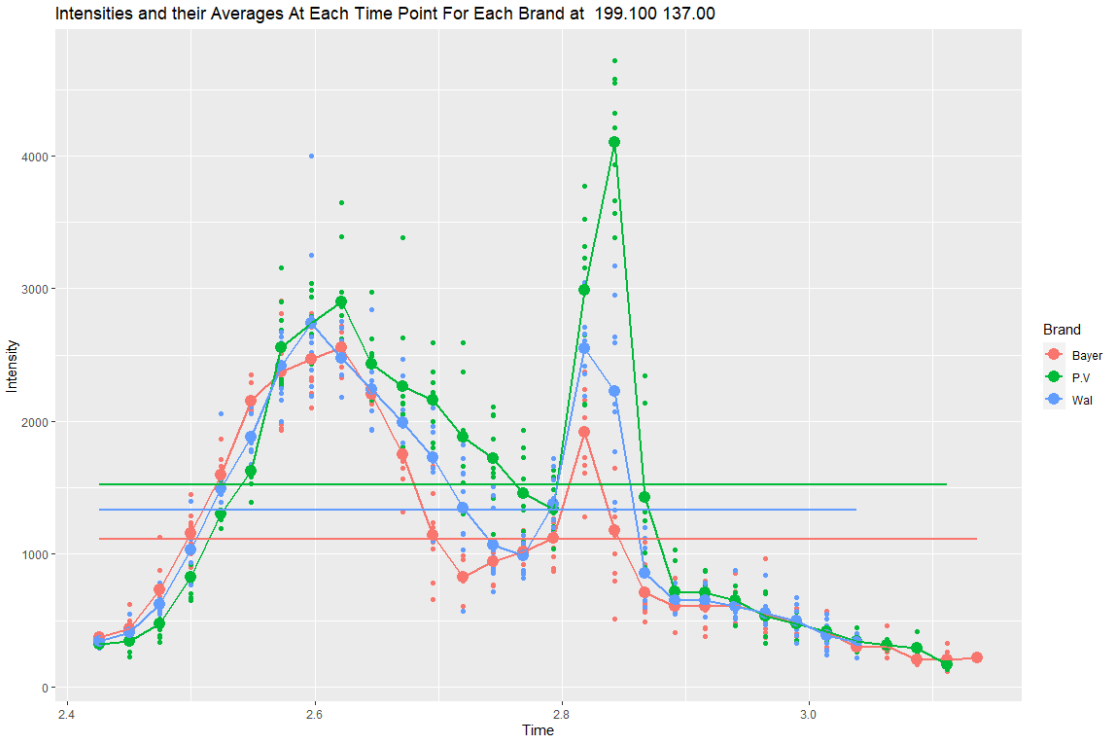
```
## [[1]]
```



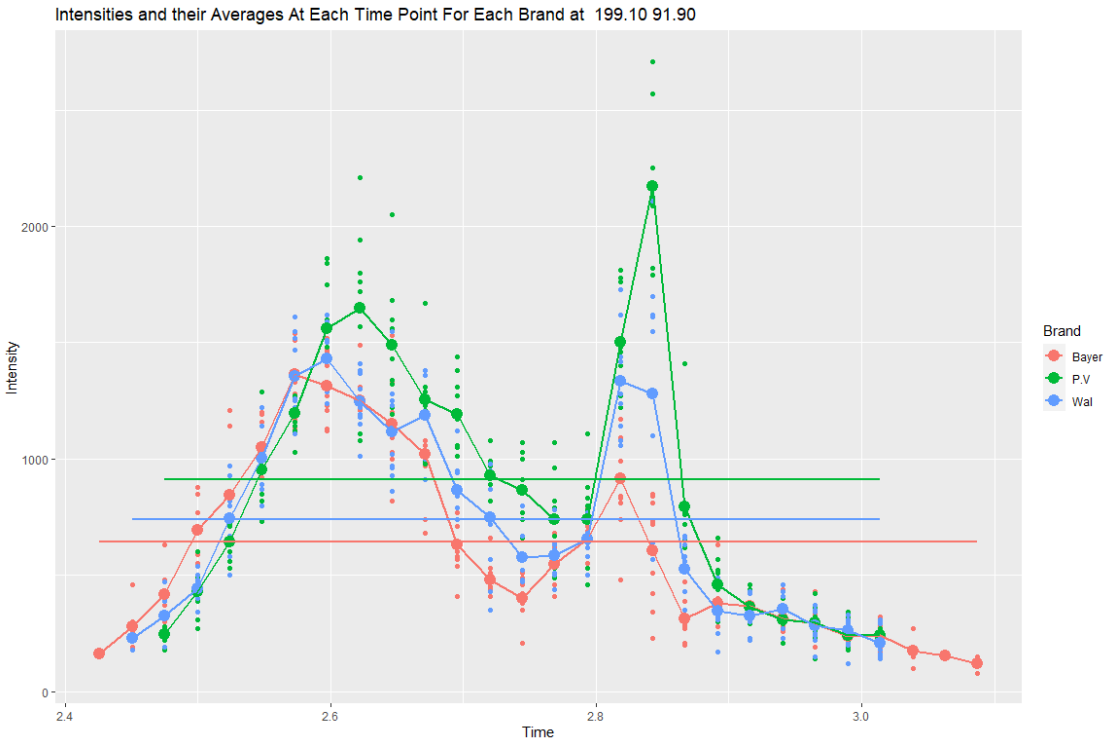

```
##
## [[2]]
```


Intensities and their Averages At Each Time Point For Each Brand at 178.7 93.00
Intensity
Time
Brand
Bayer
P.V
Wal

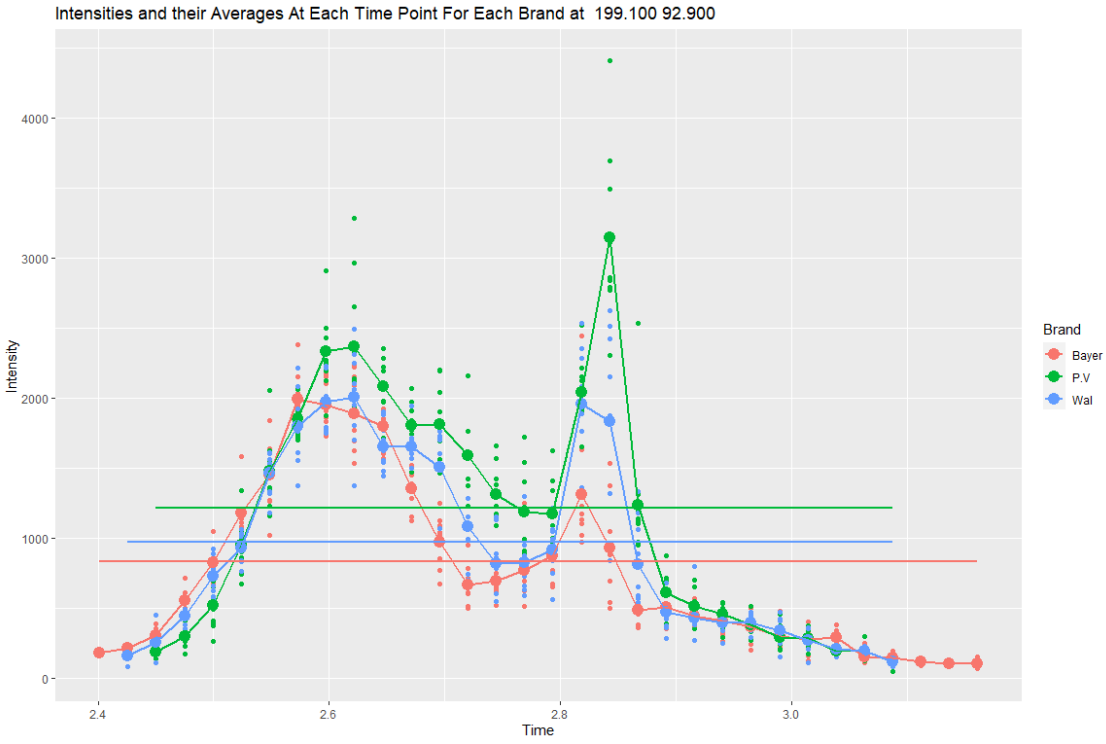
```
##
## [[3]]
```


Intensities and their Averages At Each Time Point For Each Brand at 199.100 137.00
Intensity
Time
Brand
Bayer
P.V
Wal

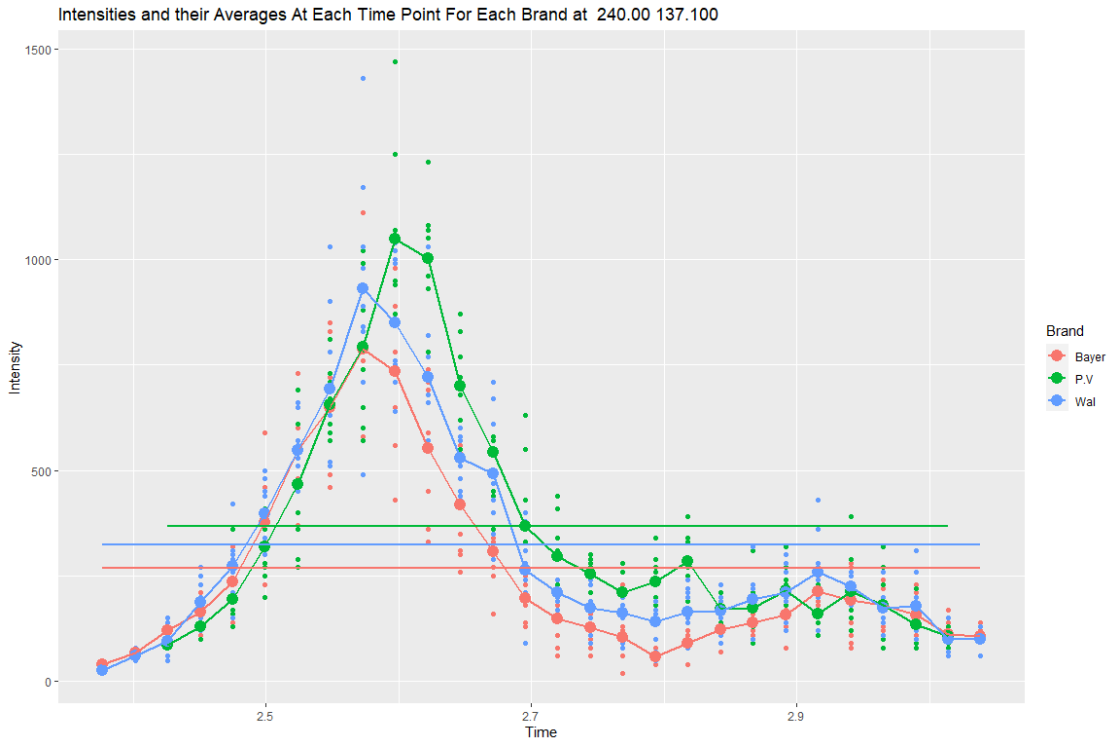
```
##
## [[4]]
```


Intensities and their Averages At Each Time Point For Each Brand at 199.10 91.90
Intensity
2000
1000
0
2.4
2.6
2.8
3.0
Time
Brand
Bayer
P.V
Wal

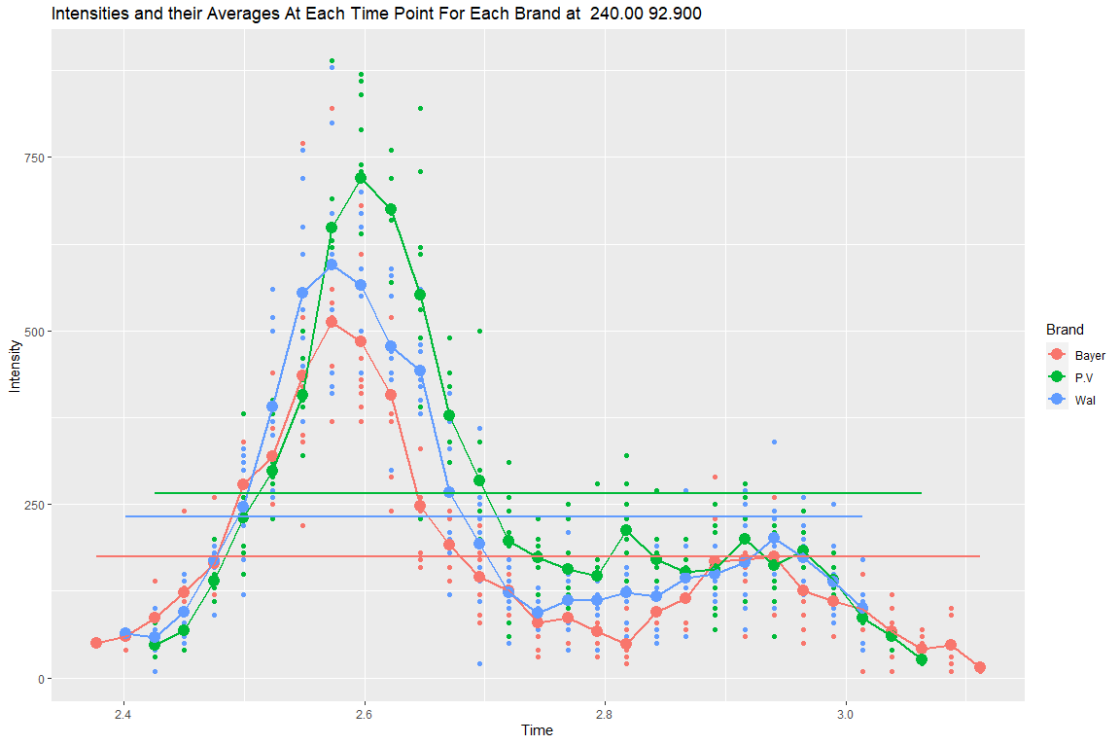
```
##
## [[5]]
```


Intensities and their Averages At Each Time Point For Each Brand at 199.100 92.900
Intensity
4000
3000
2000
1000
0
2.4
2.6
2.8
3.0
Time
Brand
Bayer
P.V
Wal

```
##
## [[6]]
```



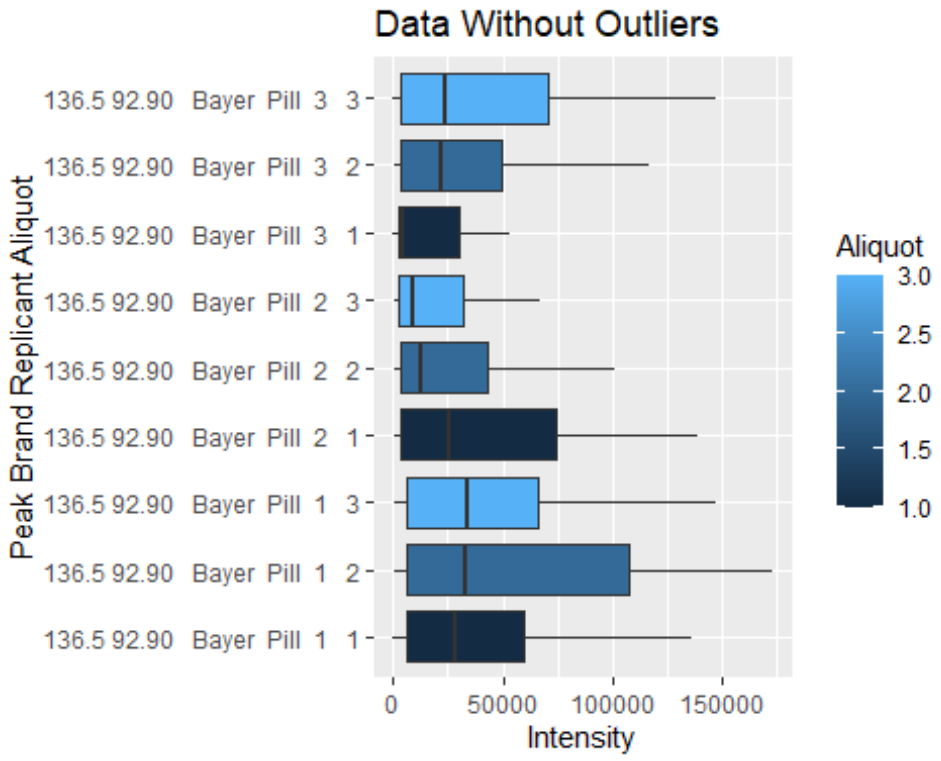

```
##
## [[7]]
```

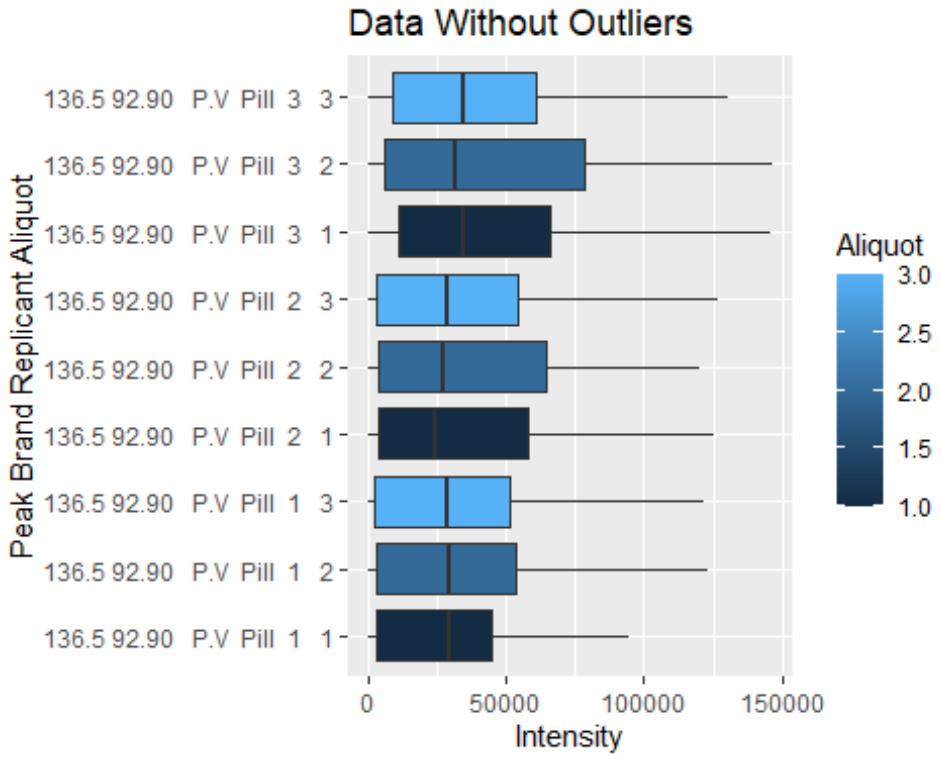


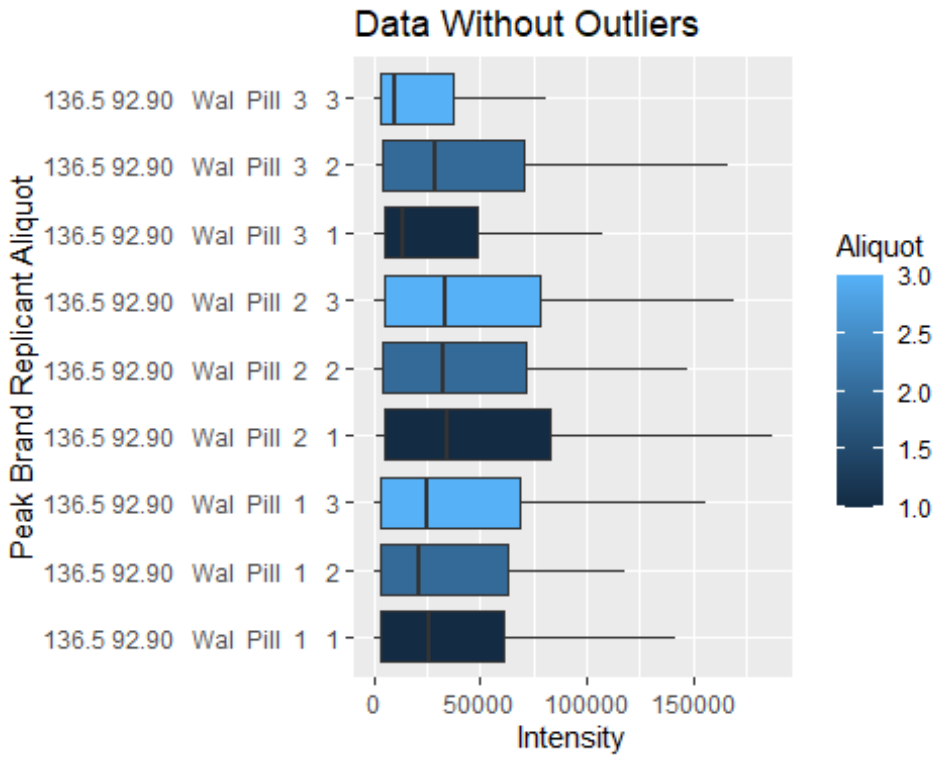
My exploratory analysis utilized scatter/line charts for each peak, showing the intensity measurements, the average intensity at each time point, and the overall average intensity

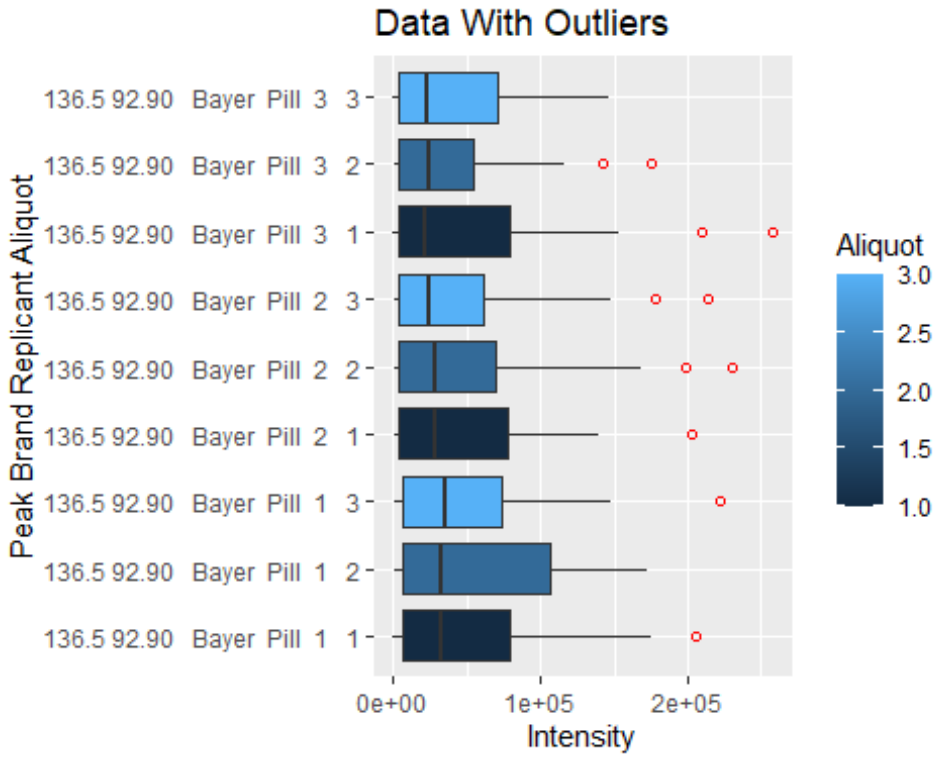
for the peak of all of the replicants, aliquots, and brands. I used this method since the same amount of replicants and aliquots were measured for each brand and peak, and this would allow for a simple visual comparison of the brands at each peak. Small circles represent intensities within each aliquot at each time point while larger circles and the lines connecting them represent the mean intensity and change in mean intensity between the time points for each brand. The three horizontal lines represent the overall average intensity for each brand at each peak. From these charts, I noticed that P.V. brand aspirin had the highest average intensity in all but two peaks, Bayer brand aspirin had the lowest average intensity at all but one peak, and Walgreens brand aspirin had average intensities between Bayer and P.V. at all but one peak.

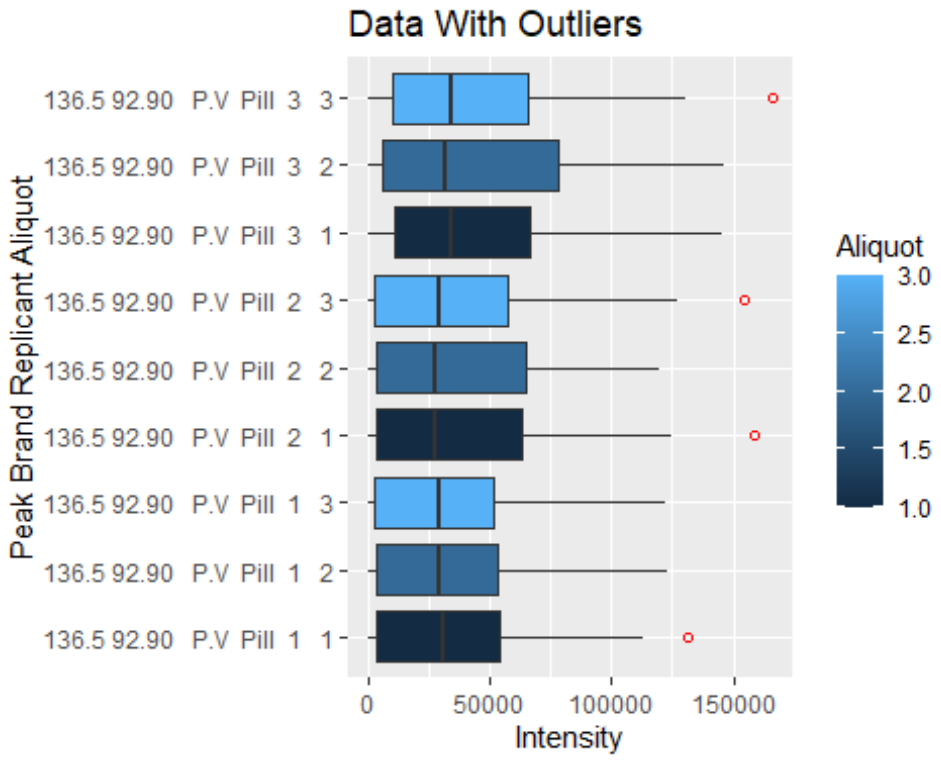
The above chart also shows many outliers in the dataset due to inconsistencies between replicants and aliquots within the same brand and time point. Due to the data being taken into account for this project being the area under the curve of each of the aliquots and due to myself not having access to the researchers in order to ask further questions, I've decided to test the null hypothesis with and without outliers being removed. In order to remove the outliers, I'll be removing intensities in each aliquot that are either above 1.5 times the interquartile range plus the third quartile or below 1.5 times the interquartile range minus the first quartile of each aliquot.

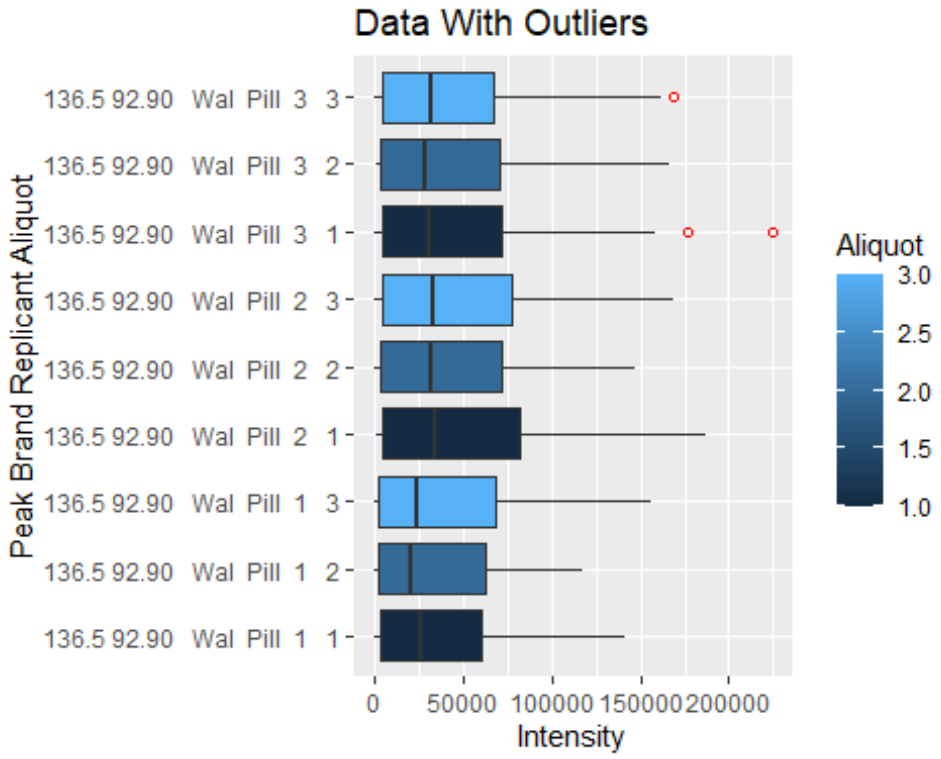

Data Without Outliers
136.5 92.90 Bayer Pill 3 3
136.5 92.90 Bayer Pill 3 2
136.5 92.90 Bayer Pill 3 1
136.5 92.90 Bayer Pill 2 3
136.5 92.90 Bayer Pill 2 2
136.5 92.90 Bayer Pill 2 1
136.5 92.90 Bayer Pill 1 3
136.5 92.90 Bayer Pill 1 2
136.5 92.90 Bayer Pill 1 1
Peak Brand Replicant Aliquot
0
50000
100000
150000
Intensity
Aliquot
3.0
2.5
2.0
1.5
1.0


Data Without Outliers
136.5 92.90 P.V Pill 3 3
136.5 92.90 P.V Pill 3 2
136.5 92.90 P.V Pill 3 1
136.5 92.90 P.V Pill 2 3
136.5 92.90 P.V Pill 2 2
136.5 92.90 P.V Pill 2 1
136.5 92.90 P.V Pill 1 3
136.5 92.90 P.V Pill 1 2
136.5 92.90 P.V Pill 1 1
Peak Brand Replicant Aliquot
0
50000
100000
150000
Intensity
Aliquot
3.0
2.5
2.0
1.5
1.0


Data Without Outliers
136.5 92.90 Wal Pill 3 3
136.5 92.90 Wal Pill 3 2
136.5 92.90 Wal Pill 3 1
136.5 92.90 Wal Pill 2 3
136.5 92.90 Wal Pill 2 2
136.5 92.90 Wal Pill 2 1
136.5 92.90 Wal Pill 1 3
136.5 92.90 Wal Pill 1 2
136.5 92.90 Wal Pill 1 1
Peak Brand Replicant Aliquot
0
50000
100000
150000
Intensity
Aliquot
3.0
2.5
2.0
1.5
1.0


Data With Outliers
136.5 92.90 Bayer Pill 3 3
136.5 92.90 Bayer Pill 3 2
136.5 92.90 Bayer Pill 3 1
136.5 92.90 Bayer Pill 2 3
136.5 92.90 Bayer Pill 2 2
136.5 92.90 Bayer Pill 2 1
136.5 92.90 Bayer Pill 1 3
136.5 92.90 Bayer Pill 1 2
136.5 92.90 Bayer Pill 1 1
Peak Brand Replicant Aliquot
0e+00
1e+05
2e+05
Intensity
Aliquot
3.0
2.5
2.0
1.5
1.0





Here I confirmed that my method for removing the outliers worked by plotting box plots for one of the peaks in order to check if the outliers are still present. Now, none of the outliers are present in the Data.All.df.No.Outliers dataframe while the Data.All.df dataframe is keeping all outliers.

3. Perform an Analysis of Variance or related analysis if you deem it appropriate.

Variance and Standard Deviation of the Aliquots, Grouped by Peak, Brand, and Replicant

Method	Standard_Deviation	Variance
With Outliers	486.9008	1267944
Without Outliers	1198.3651	9062003

Here I obtained some summary statistics in order to compare the 3 brands of aspirin in the 2 new datasets (with outliers and without outliers). I then grouped the data by peak, brand, replicant, and aliquot. Then I obtained the count, standard deviation, and mean of each group. The mean would also be considered the area under the peak for this dataset, so I labeled it as AUP. Afterwards I printed the first 6 rows of each of these datasets.

Furthermore, I checked for standard deviation and variance between the AUPs, grouped by peak, brand, and replicant. I found that the aliquots are very spread out between each other, especially between those that had the outliers removed. This may be indicative of an issue with the measurements.

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Brand         2 1.446e+07   7229092   0.028  0.973
## Residuals   186 4.827e+10 259516456

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Aliquots_AUP$AUP and Aliquots_AUP$Brand
##
##     Bayer P.V
## P.V 0.85  -
## Wal 0.98  0.83
##
## P value adjustment method: none

##              Df    Sum Sq   Mean Sq F value Pr(>F)
## Brand         2 1.270e+08  63487548   0.317  0.729
## Residuals   186 3.724e+10 200222989

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Aliquots_AUP.No.Outliers$AUP and Aliquots_AUP.No.Outliers$Brand
##
##     Bayer P.V
## P.V 0.53  -
## Wal 0.46  0.91
##
## P value adjustment method: none
```

ANOVA and T-Test Results Summary

Brand	AoV_P_Value	Pairwise_T_Test_P_Value
Brand - With Outliers	0.973	
Brand - w/o Outliers	0.729	
P.V. vs Bayer - With Outliers		0.85
Wal vs Bayer - With Outliers		0.98
Wal vs P.V. - With Outliers		0.83
P.V. vs Bayer - w/o Outliers		0.53
Wal vs Bayer - w/o Outliers		0.46
Wal vs P.V - w/o Oultiers		0.91

Here I conducted a one way analysis of variance to check if brand has an effect on the area under the peak as well as a pairwise t-test with no adjustment for error to check if there was a significant difference that can be found between the individual brands. One-way ANOVA is generally used to find out if there's a statistically significant difference between the means of 3 or more independent groups. It allows me to generate an F value that can be in turned used to obtain a P statistic which allows for measures of significance. The T-test is a common hypothesis test based on the Student's t-distribution that's often used to make pairwise comparisons.

These tests were ran on the datasets with and without outliers. At a 95% confidence interval, there was no significant difference found in either of the brands. The pairwise t-tests with and without outliers also did not find a significant difference between the brands.

```
ANOVA Formula:
```

$$F = \frac{MST}{MSE} = \frac{\sum_{j=1}^{k}\sum_{j=1}^{l}(\bar{x}_j - x_j)^2}{df_w}$$

```
T-Test Formula:
```

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Sources: https://www.vedantu.com/formula/anova-formula https://www.educba.com/t-test-formula/

```
##                diff       lwr      upr     p adj
## P.V-Bayer 1584.6195 -4371.963 7541.202 0.8046378
```

```
## Wal-Bayer 1859.9952 -4096.587 7816.578 0.7413751
## Wal-P.V    275.3757 -5681.207 6231.958 0.9934445

## # A tibble: 3 x 10
##   .y.   group1 group2    n1    n2 statistic    df     p p.adj p.adj.signif
## * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl> <dbl> <chr>
## 1 AUP   Bayer  P.V       63    63     0.831    62 0.409 0.818 ns
## 2 AUP   Bayer  Wal       63    63    -0.143    62 0.887 0.887 ns
## 3 AUP   P.V    Wal       63    63    -2.10     62 0.039 0.118 ns

##                diff       lwr      upr     p adj
## P.V-Bayer 1584.6195 -4371.963 7541.202 0.8046378
## Wal-Bayer 1859.9952 -4096.587 7816.578 0.7413751
## Wal-P.V    275.3757 -5681.207 6231.958 0.9934445

## # A tibble: 3 x 10
##   .y.   group1 group2    n1    n2 statistic    df     p p.adj p.adj.signif
## * <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl> <dbl> <chr>
## 1 AUP   Bayer  P.V       63    63    -1.85     62 0.07  0.14  ns
## 2 AUP   Bayer  Wal       63    63    -2.05     62 0.044 0.132 ns
## 3 AUP   P.V    Wal       63    63    -0.493    62 0.624 0.624 ns
```

Ad-Hoc Test Results Summary

Brand	Tukey_Kramer	Bonferroni
P.V. vs Bayer - With Outliers	0.489	0.818
Wal vs Bayer - With Outliers	0.984	0.887
Wal vs P.V. - With Outliers	0.389	0.118
P.V. vs Bayer - w/o Outliers	0.078	0.140
Wal vs Bayer - w/o Outliers	0.030	0.132
Wal vs P.V - w/o Oultiers	0.924	0.624

Here, for further analysis I ran a Tukey-Kramer post-hoc test which has an error correction that assists when comparing pairs within a group with sample sizes that aren't exactly the same. This isn't the case in this study as although the amount of measurements within each aliquot vary between 20 and 29, the analysis is being done on the area under the peak for each aliquot, or in other words, the mean of each aliquot. The Tukey-Kramer method is optimally used when all possible pairs of a group are being compared.

Source: https://www.statology.org/tukey-vs-bonferroni-vs-scheffe/

```
Tukey-Kramer post-hoc formula:
```

$$y_i - y_j \pm q\alpha, k, N - k \sqrt{(\frac{MST}{2})(\frac{1}{n_i} + \frac{1}{n_j})}$$

Source: https://aaronschlegel.me/tukeys-test-post-hoc-analysis.html

Here, I also ran a Bonferroni post-hoc test which has an error correction that, similarly to the Tukey-Kramer test, also assists when comparing pairs within a group with sample sizes that aren't exactly the same. This test is optimally used when a subset of pairs from a group is being compared.

Source: https://www.statology.org/tukey-vs-bonferroni-vs-scheffe/

Bonferroni post-hoc formula Confidence Intervals:

$$\hat{C}_1 = \frac{\bar{Y}_1 + \bar{Y}_2}{2} - \frac{\bar{Y}_3 + \bar{Y}_4}{2}$$

and

$$\hat{C}_2 = \frac{\bar{Y}_1 + \bar{Y}_3}{2} - \frac{\bar{Y}_2 + \bar{Y}_4}{2}$$

Bonferroni post-hoc formula Point Estimate and Variance of the Confidence Intervals:

$$\sum_{\iota=1}^{4} \frac{c_\iota^2}{\eta_\iota}$$

and

$$\sigma_\epsilon^2 \sum_{\iota=1}^{4} \frac{c_\iota^2}{4}$$

Source: https://www.itl.nist.gov/div898/handbook/prc/section4/prc473.htm

For this project, I am prioritizing the results of the pure ANOVA and pairwise T-tests without error correction, although stakeholders are free to use the results of the Tukey-Kramer and/or Bonferroni tests if they so wish as they all have their merits and depending on factors mentioned in my descriptions of them above, can lead to more accurate results.

- Explain the conclusions of your analysis.

Conclusion

The goal of this project was to fail to reject or to reject the null hypothesis that the aspirin pills from the three companies are the same. My conclusion for this project is to fail to reject the null hypothesis that the aspirin pills are the same. The initial ANOVA test revealed that at a 95% confidence rate, there is no difference between the brands with or without the outliers being removed. This result was further confirmed when running pairwise t-tests between the three brands.

When outliers are not removed and when comparing the individual brands between each other using either Tukey-Kramer or Bonferroni ad-hoc analyses, the null hypothesis is not rejected and the aspirins from the 3 companies are the same at a 95% confidence rate. When the outliers are removed and the Tukey-Kramer ad-hoc analysis but not the

Bonferroni ad-hoc analysis is used, the Walgreens and Bayers brand aspirin pills are found to be different at a 95% rate of confidence while the other pairs of brands are found to be the same.

For this project, as mentioned in the previous section, I'm prioritizing the results of the ANOVA and T-Test without error corrections. Thus, I failed to reject the null hypothesis. If a stakeholder wishes to prioritize the results of the Tukey-Kramer ad-hoc analysis when outliers are removed, they would reject the null hypothesis. In all other cases of this study including when using Bonferroni ad-hoc analysis, the null hypothesis fails to be rejected

These results could be useful for confirming whether or not the aspirin brands are the same depending on whether conditions are adjusted(outliers are removed) or different types of error correction are used. Something that may be necessary to be looked into though before accepting these results would be the high variance found between the aliquots as shown in part 3 as they may be indicating an issue with the data collection or may just be due to the nature of the process being measured.

4. Provide a one-page write-up (excluding graphs, tables and figures) explaining your analysis of the dataset and your recommendations on the usefulness of your predictions.

5. As a secondary component provide annotated code that replicates your analysis.

Datasets

The datasets are organized into a set of folders.

1. Each folders refers to the peak location.
2. Within each folder there is a single txt file for each pill.
3. Each text file will have 3 columns that correspond to the intensities of three aliqouts and other information as well.
4. The other columns correspond to blank time and retention time for which the intensity was measured.
5. All of the non-relevant intensities have been deleted.
6. To calculate the area under a peak (AUP)- simply take the mean of the intensity values for a given aliqout.

Notes

1. Make sure to check for consistency between the three AUP's associated with the same pill- these are all measurements of the same object.

2. Make sure to document everything that you have done in your rmd file… this is your lab notebook for this type of project.

3. I hoping to receive an ANOVA with a set of post hoc tests between the three brands of pills, but if you are concerned about the assumptions for an ANOVA please do not hesitate to do something else or to discuss why we can not perform a formal analysis.

4. Do watch the video where I introduce the problem before starting the task.
5. You are expected to work by yourself on the project but we will have a message board for questions about the project open.