An Analysis of Galaxies and Their Properties

By Aarya, Zaid and Rafaye

Project Description

- The purpose of this project is to apply statistics in astronomical research work to better understand the universe!
- The data we collected (from the Galaxy Zoo database) describes the properties that Galaxies possess and how that correlates with other features we might be interested in.
- We examined three research questions and completed visualizations, hypothesis tests and descriptions, detailing the process.

Our Plan

- We will first complete two appropriate visualizations for each research question through R studio. This will give us an approximate idea about making out hypothesis and setting up the test.
- Then, we will complete the hypothesis tests using the methods mentioned previously and obtain definite answer on whether to reject the null hypothesis we determined through the previous step.

Project Objectives

- We would like to obtain definite results for our research questions through the processes mentioned. Hence, we will explain all the steps we undertook and the rationale behind each step.
- We need to use accurate and deductive tests for each research question. Therefore, we must be careful with the method in which we graph our variables, the choice of variables themselves and finally the reasoning behind our choice for the null and alternate hypothesis. Finally, we can undertake the test (which also needs careful review) and come to a conclusion.

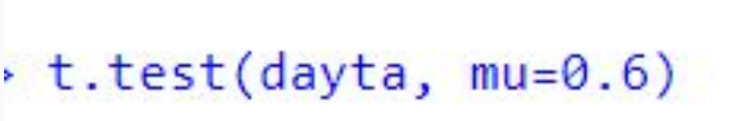
Question 1 “How does the Mean Spectra for each Galaxy differ?”

- For this question, we would need to gather data that contained different galaxies and their mean spectra located in one column. Dataset chosen for this task was **GalaxyZoo**, table 2 which contains the classifications for galaxies as well as their spectra.

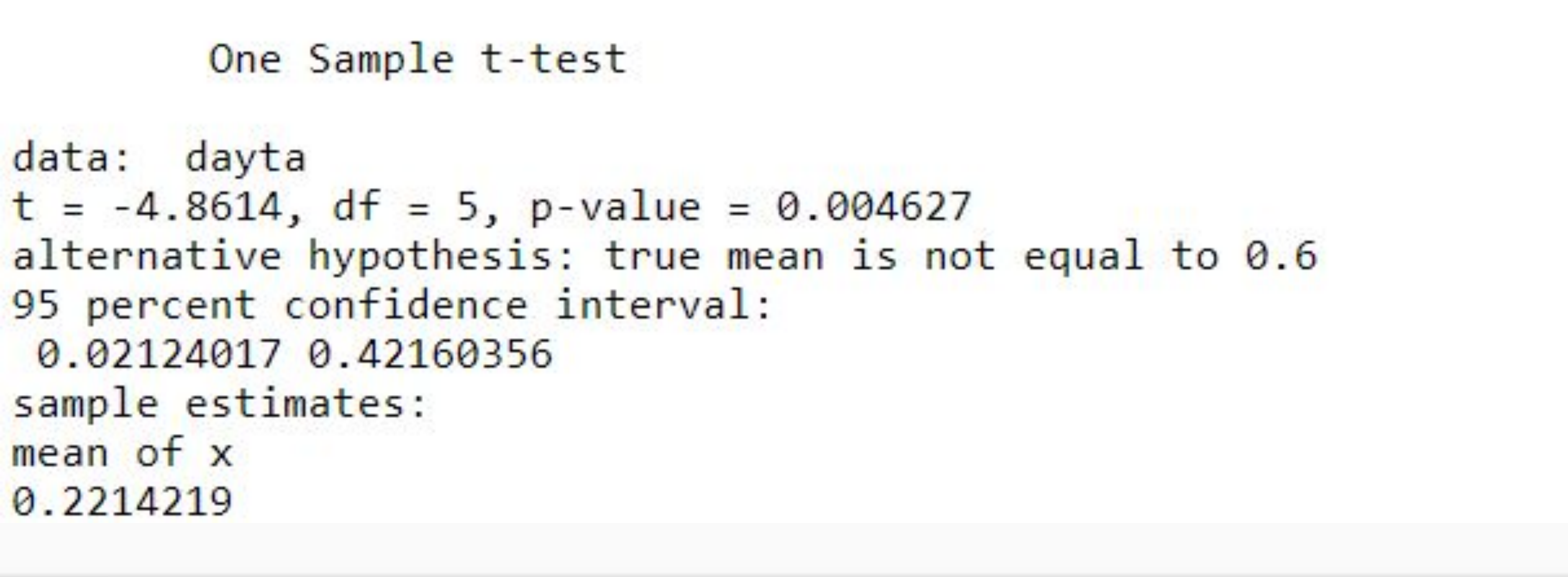
- By using R, the mean value of the spectra for each galaxy was extracted and then could be used for the data visualization.

- **For our First Visualization** We decided to use a graphical data visualization with a bar chart as we felt this was the most effective way to communicate our results to an audience as our results contained both quantitative and qualitative data.

- For the **Second Visualization**, in order to properly illustrate the distribution of the average spectra across the different galaxies, we felt a density plot would effectively present the distribution of the spectral characteristics across multiple galaxies in the Galaxy Zoo Dataset. Although the first visualization shows the mean spectra for each galaxy this visualization focuses more heavily on the actual distribution between galaxies.



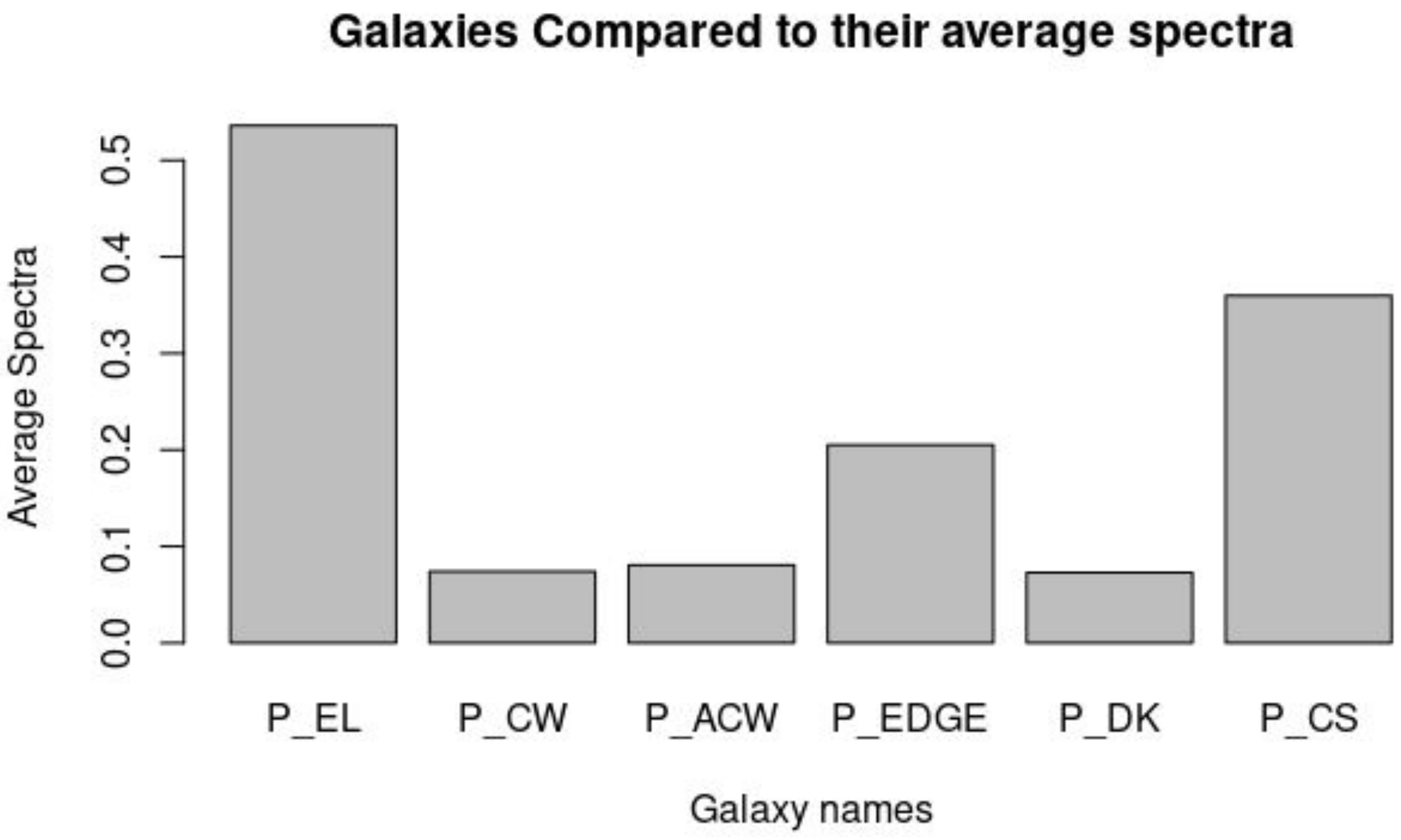
```
› t.test(dayta, mu=0.6)
```

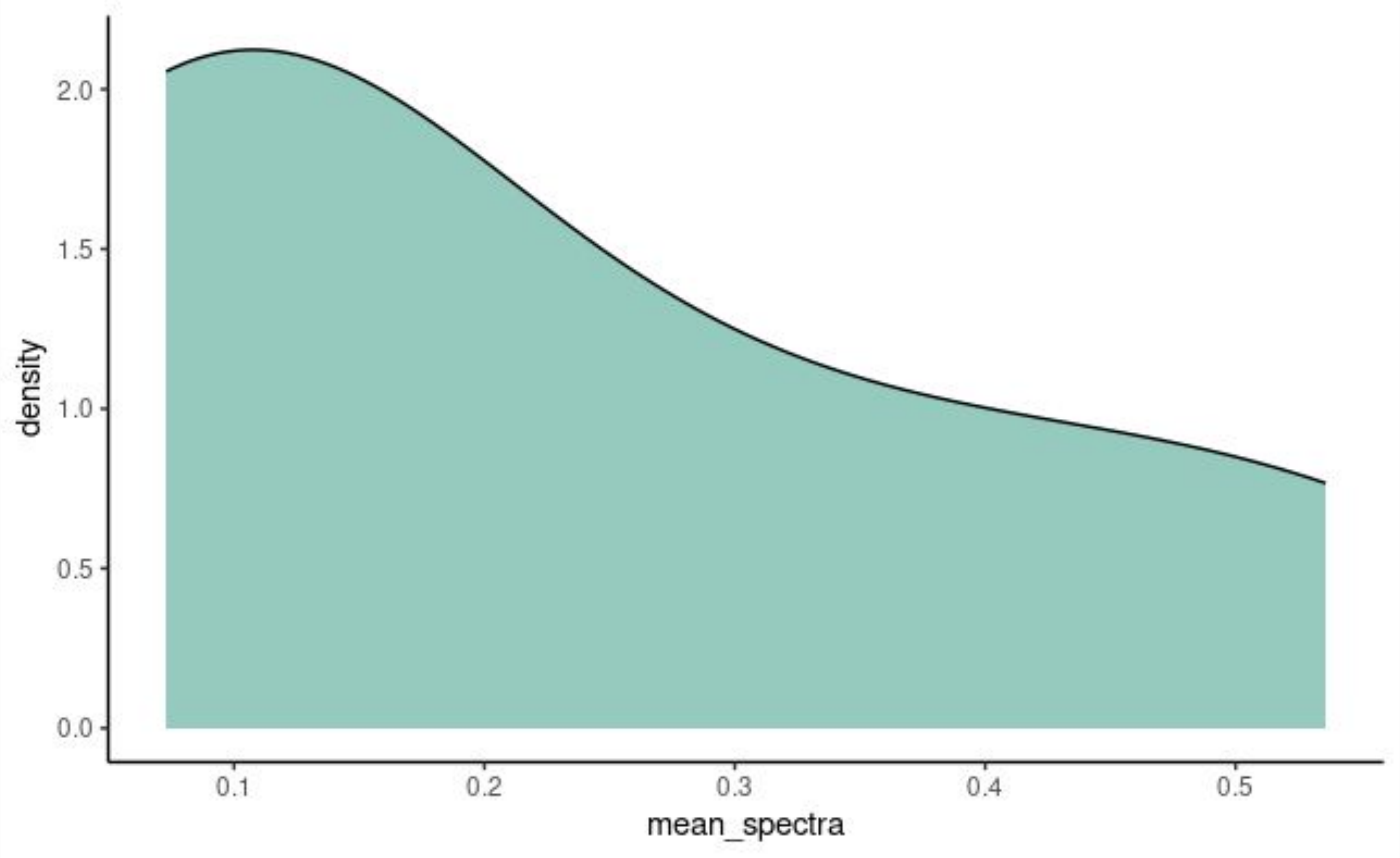


```
        One Sample t-test

data:  dayta
t = -4.8614, df = 5, p-value = 0.004627
alternative hypothesis: true mean is not equal to 0.6
95 percent confidence interval:
 0.02124017 0.42160356
sample estimates:
mean of x
0.2214219
```

Hypothesis Test Q1

Aim of the hypothesis test A is to perform a one sample T-Test is performed on Data in order to determine whether the mean of the population (Average spectra/brightness for each galaxy) from which the sample was drawn is significantly different from the value specified by ‘mu’. The output of the test shows that the calculated t value is -4.86 with a p value of 0.000462. The p value indicates the probability of obtaining a sample mean as extreme as the observed one, **assuming the null hypothesis** (that the true mean is equal to 0.6) is true. A small p value suggests that the null hypothesis can be rejected in favor of the alternative hypothesis. In this case the p-value is less than 0.05 which suggests that the null hypothesis should be rejected and that the true mean of the average spectra of all the galaxies is not likely equal to 0.6. Furthermore, the sample mean of ‘0.0221’ is given, along with a 95% confidence interval of ‘(0.021,0.4216’). This interval suggests that we can be 95% confidence that the average spectra for each galaxy falls within this range.


Galaxies Compared to their average spectra
Average Spectra
0.0
0.1
0.2
0.3
0.4
0.5
P_EL
P_CW
P_ACW
P_EDGE
P_DK
P_CS
Galaxy names


density
2.0
1.5
1.0
0.5
0.0
0.1
0.2
0.3
0.4
0.5
mean_spectra

Question 2 “Are there equal fractions of mwebv across all supernova classes in the training and testing data?”

Transient

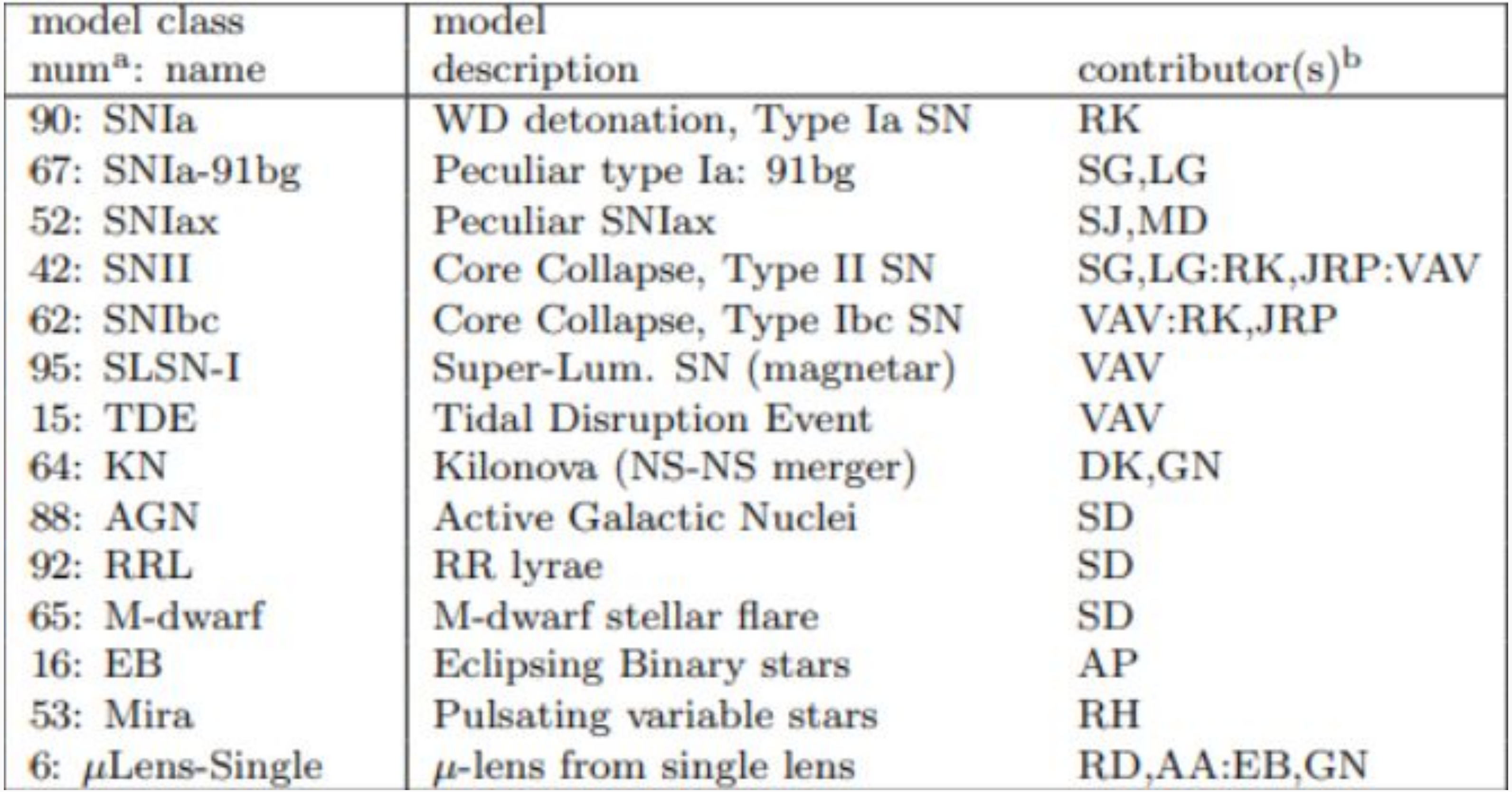
- Transients refer to astronomical phenomena with durations of fractions of a second to weeks or years. Typically they are extreme, short-lived events associated with the total or partial destruction of an astrophysical object.
- Most massive stars end their lives by exploding spectacularly, known as a supernova, a major type of transients.

MWEBV

- mwebv stands for "Milky Way E(B-V)", where E(B-V) is the color excess, a measure of how much the observed colors of an object are shifted by dust extinction.
- is essentially a measure of the amount of dust along the line of sight to the object being observed, with higher values indicating greater dust extinction.
- The mwebv value is typically derived from multi-band photometry, which involves observing an object in multiple different wavelength bands and comparing the observed colors to those of a reference star or model.

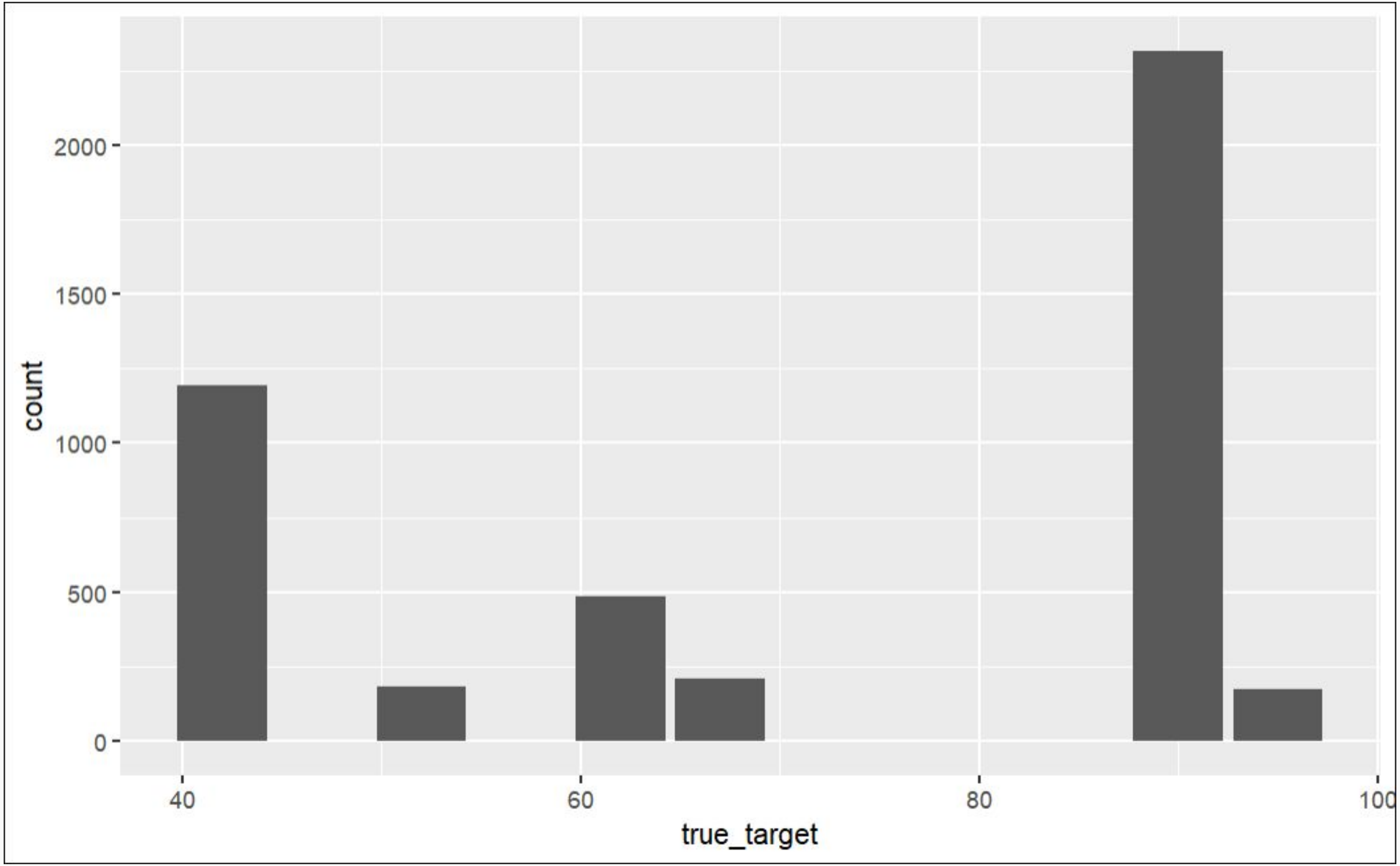
Dataset

- For this particular question, we will be using the PLAsTiCC dataset as it holds all the relevant information such as the mwebv values for each respective transient class.
- There are a total of 14 transient classes within this dataset.
- However we will only be looking at the supernova transients for this question. Thus we will focus on classes 90, 67, 52, 42, 62, and 95.

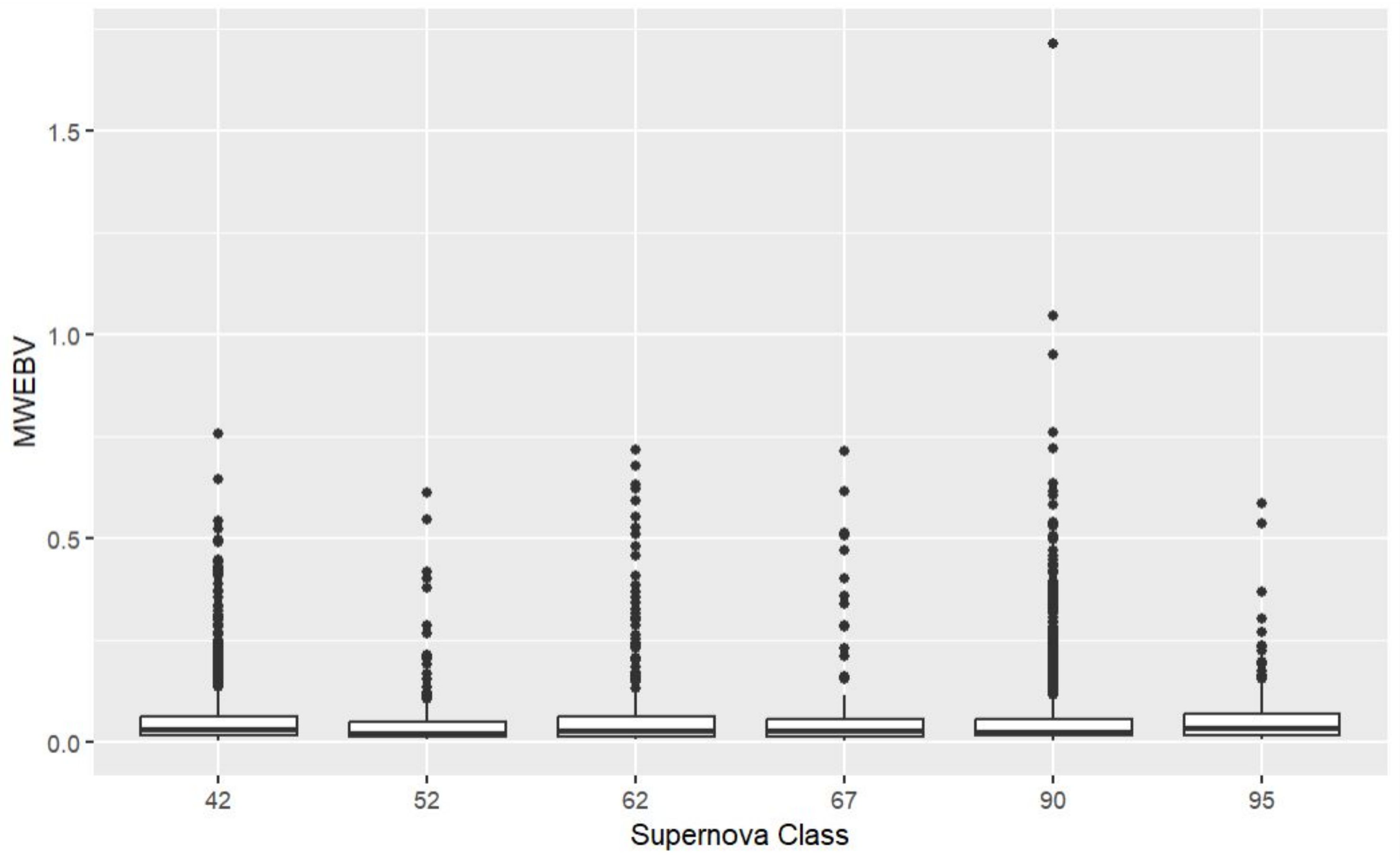


model class	model	
num[a]: name	description	contributor(s)[b]
90: SNIa	WD detonation, Type Ia SN	RK
67: SNIa-91bg	Peculiar type Ia: 91bg	SG,LG
52: SNIax	Peculiar SNIax	SJ,MD
42: SNII	Core Collapse, Type II SN	SG,LG:RK,JRP:VAV
62: SNIbc	Core Collapse, Type Ibc SN	VAV:RK,JRP
95: SLSN-I	Super-Lum. SN (magnetar)	VAV
15: TDE	Tidal Disruption Event	VAV
64: KN	Kilonova (NS-NS merger)	DK,GN
88: AGN	Active Galactic Nuclei	SD
92: RRL	RR lyrae	SD
65: M-dwarf	M-dwarf stellar flare	SD
16: EB	Eclipsing Binary stars	AP
53: Mira	Pulsating variable stars	RH
6: μLens-Single	μ-lens from single lens	RD,AA:EB,GN

Classes of transients in the dataset



This bar plot shows how the data looks when we filter out to only include the transient classes we are interested in.



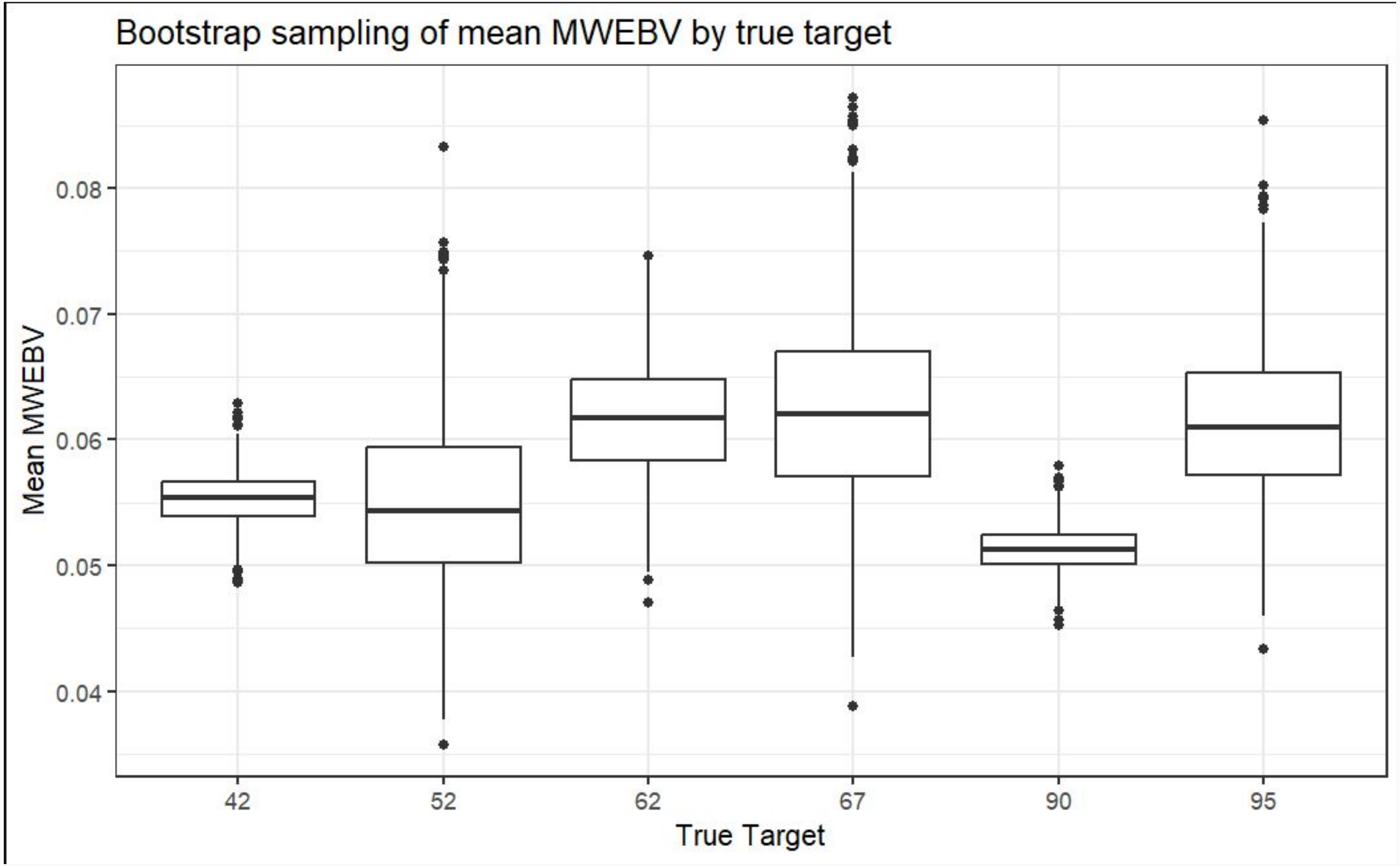
This boxplot shows the mwebv values for each transient supernova class

Bootstrapping & Confidence Intervals

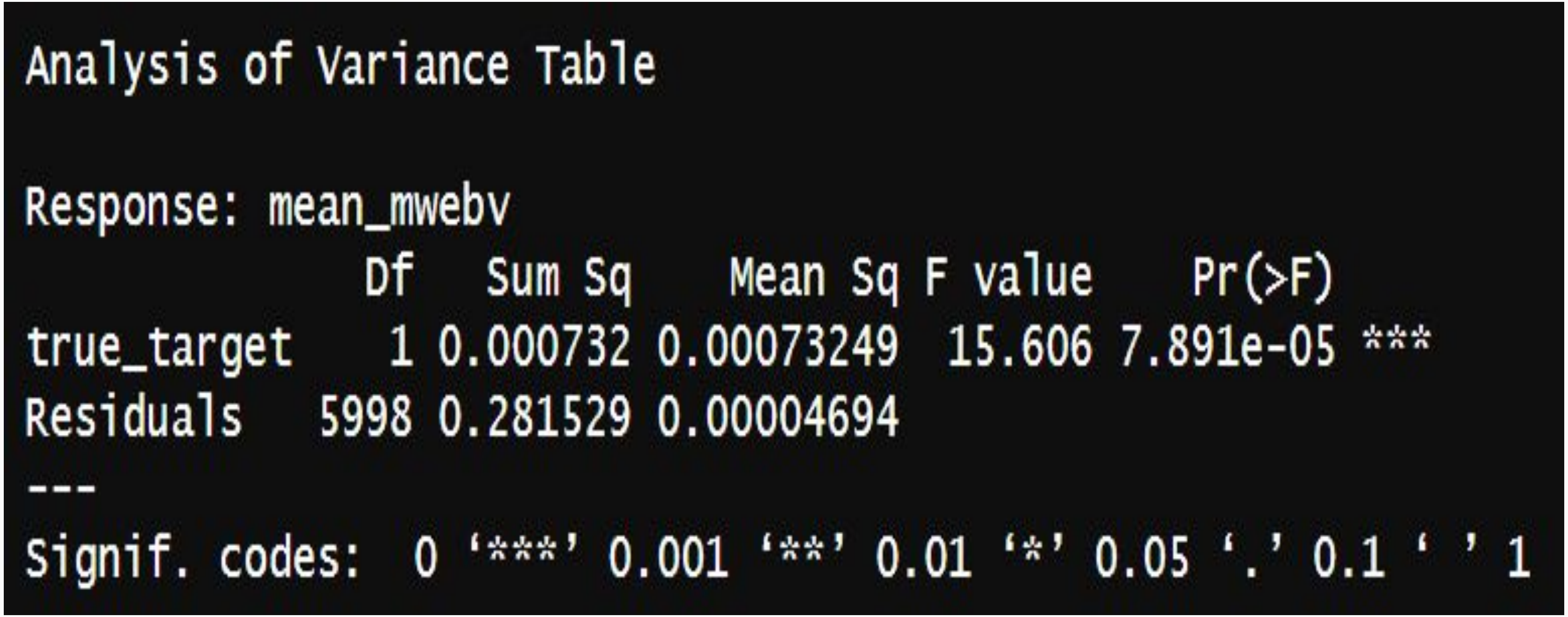
- We first resample the training data (with replacement) and generate multiple resampled datasets. Each resampled dataset is the same size as the original.
- Next, for each resampled dataset we compute the average mwebv value for each respective supernova class and also compute any relevant statistics using ANOVA to compare the average mwebv values between the different supernova classes.
- We then perform bootstrap distribution and collect the test statistic values for each resampled dataset. This bootstrap distribution represents the distribution of the test statistic under the null hypothesis, which in this case is assuming that there is no difference in mwebv values between the supernova classes.
- We then finally compute the confidence intervals for the test statistic values based on the bootstrap distribution.

Conclusion

- Based on the analysis, we can conclude that there is a significant difference in the mean mwebv values across the different supernova classes in the PLAsTiCC dataset. The ANOVA test indicates that the observed differences in mean mwebv values between the supernova classes are statistically significant ($p < 0.001$). Additionally, the confidence intervals computed from the bootstrap distribution suggest that the difference in mean mwebv values between at least two of the supernova classes is likely to be meaningful and not due to chance.
- The confidence interval obtained ranges from 0.0003043015 to 2.246597. Note that this is a 95% confidence interval so if we were to repeat this resampling procedure many times, approximately 95% of the intervals generated would contain the true population parameter. This confidence interval suggests that true difference in mean mwebv values between the different supernova classes is likely to be between 0.0003 and 2.2466. With this we can safely reject the null as we know that the difference of mean within the mwebv values are evident.



Boxplots showing the mean mwebv values for each transient



```
Analysis of Variance Table

Response: mean_mwebv
              Df   Sum Sq    Mean Sq F value    Pr(>F)
true_target    1 0.000732 0.00073249  15.606 7.891e-05 ***
Residuals   5998 0.281529 0.00004694
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabular summary of ANOVA results

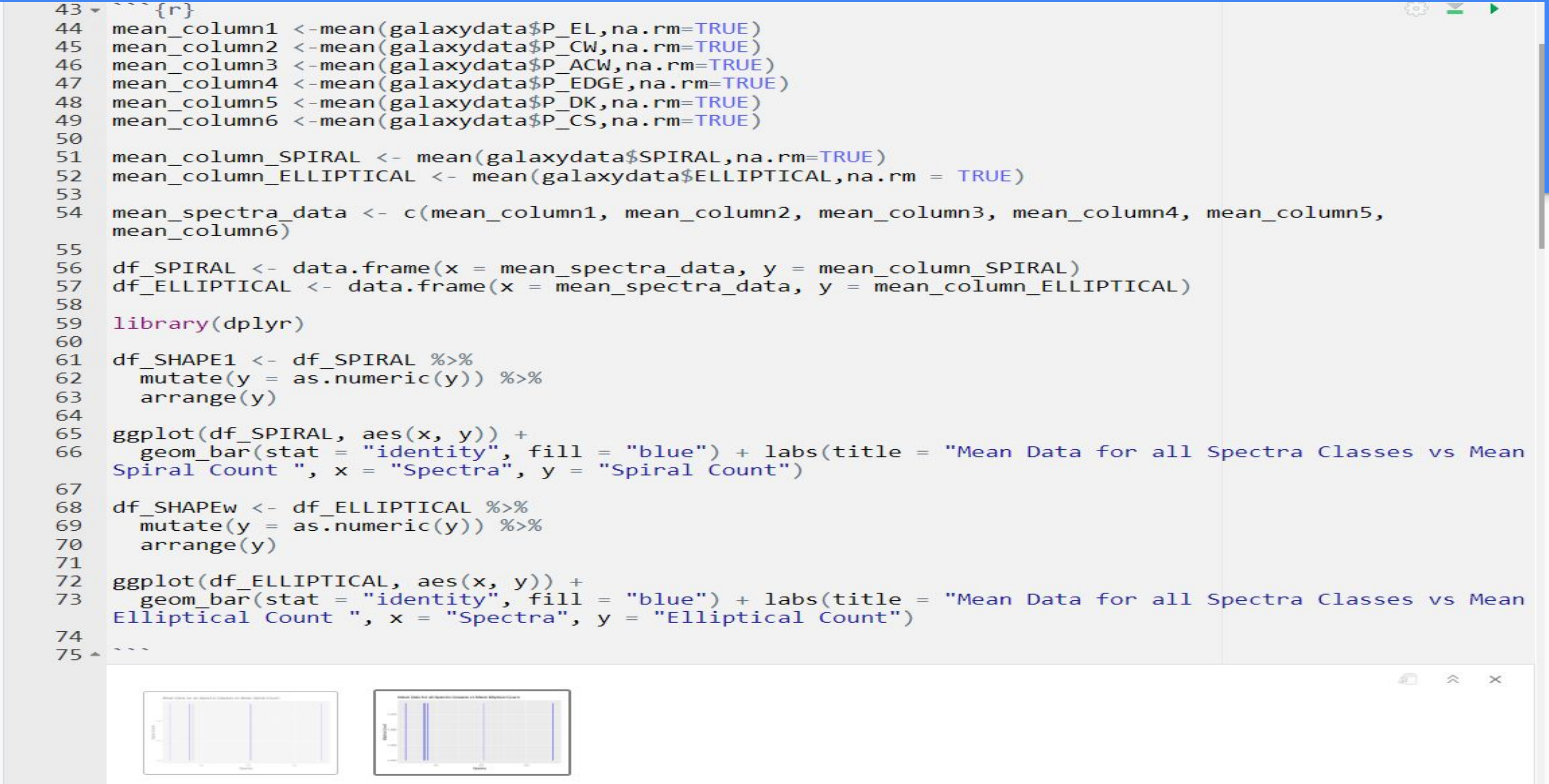
Question 3: “Does the shape of the galaxy (elliptical or spiral) correlate with its spectra (colour)?”

- For this question, we collected data from the Galaxy Zoo set (table 2) and matched the spectra listed for 6 different classes of galaxies and matched them to whether they were spiral or elliptical (shape.
- We first wanted a mean value for the spectra and realized that for all classes along with the mean spiral and elliptical data (1 being yes and 0 being no on the table) where we could calculate the likelihood of each galaxy being either spiral or elliptical.
- We found that for all classes, the likelihood of a galaxy being spiral was much higher than that for it being elliptical.

- **1st Visualization:**

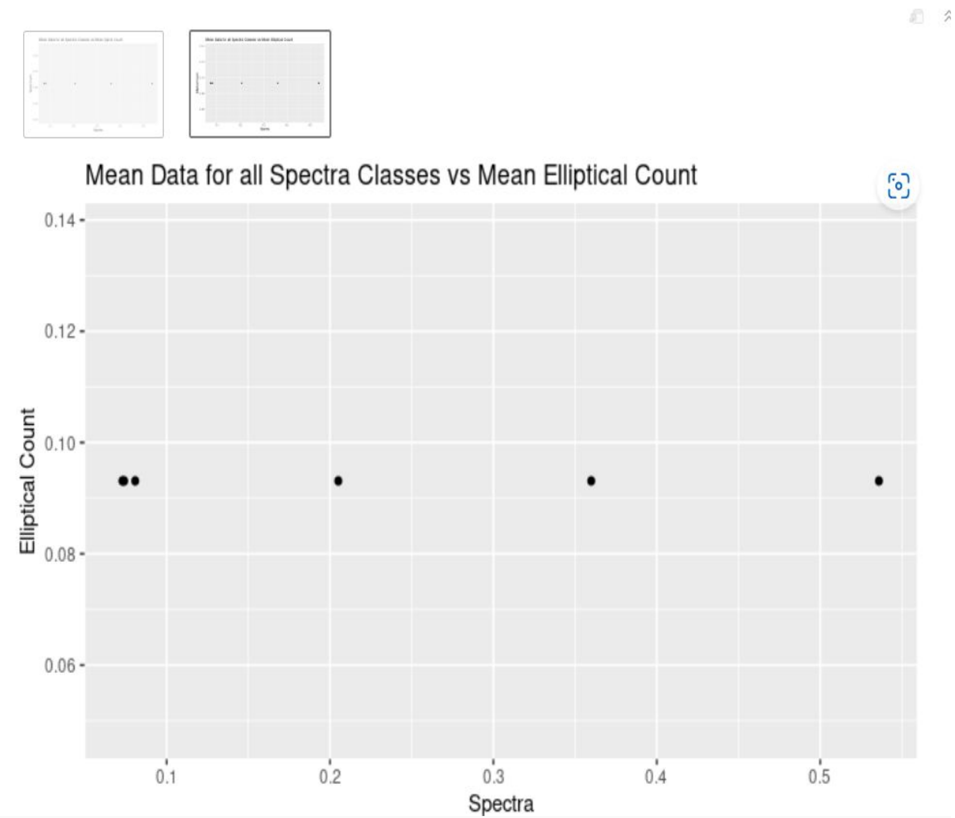
  2 scatter plots showing the ranging spectra for each class of galaxy, showing that there are definitely more spiral shaped galaxies versus the elliptical counterpart. We graphed the mean of the spectra for watch class and found that all the values correlated with the same mean for both spiral and elliptical galaxies.

Code for Q3 Visualization

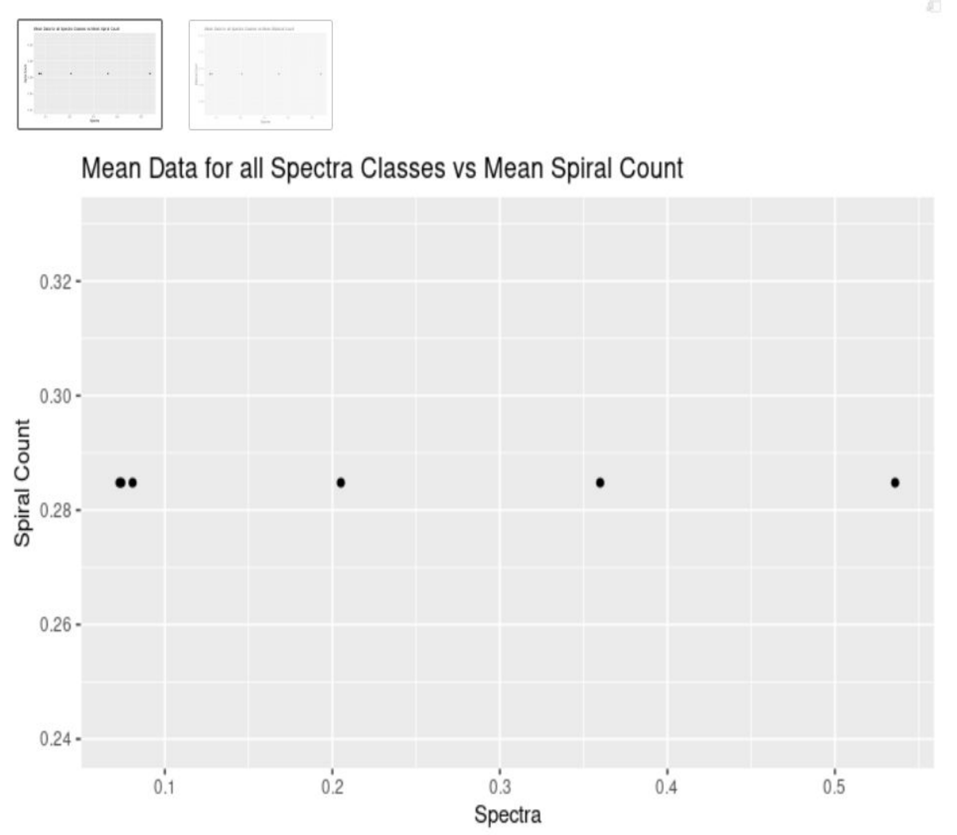


~~~
```{r}
mean_column1 <-mean(galaxydata$P_EL,na.rm=TRUE)
mean_column2 <-mean(galaxydata$P_CW,na.rm=TRUE)
mean_column3 <-mean(galaxydata$P_ACW,na.rm=TRUE)
mean_column4 <-mean(galaxydata$P_EDGE,na.rm=TRUE)
mean_column5 <-mean(galaxydata$P_DK,na.rm=TRUE)
mean_column6 <-mean(galaxydata$P_CS,na.rm=TRUE)

mean_column_SPIRAL <- mean(galaxydata$SPIRAL,na.rm=TRUE)
mean_column_ELLIPTICAL <- mean(galaxydata$ELLIPTICAL,na.rm = TRUE)

mean_spectra_data <- c(mean_column1, mean_column2, mean_column3, mean_column4, mean_column5, mean_column6)

df_SPIRAL <- data.frame(x = mean_spectra_data, y = mean_column_SPIRAL)
df_ELLIPTICAL <- data.frame(x = mean_spectra_data, y = mean_column_ELLIPTICAL)

library(dplyr)

df_SHAPE1 <- df_SPIRAL %>%
  mutate(y = as.numeric(y)) %>%
  arrange(y)

ggplot(df_SPIRAL, aes(x, y)) +
  geom_bar(stat = "identity", fill = "blue") + labs(title = "Mean Data for all Spectra Classes vs Mean Spiral Count ", x = "Spectra", y = "Spiral Count")

df_SHAPEw <- df_ELLIPTICAL %>%
  mutate(y = as.numeric(y)) %>%
  arrange(y)

ggplot(df_ELLIPTICAL, aes(x, y)) +
  geom_bar(stat = "identity", fill = "blue") + labs(title = "Mean Data for all Spectra Classes vs Mean Elliptical Count ", x = "Spectra", y = "Elliptical Count")

```
~~~

Visualizations



Mean Data for all Spectra Classes vs Mean Elliptical Count

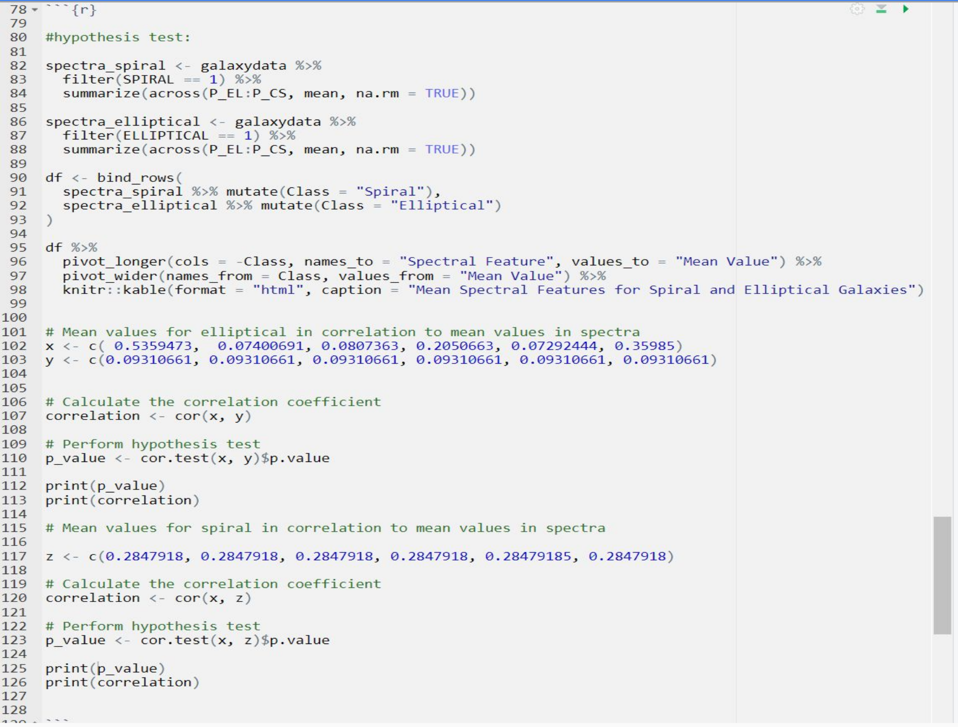


Mean Data for all Spectra Classes vs Mean Spiral Count

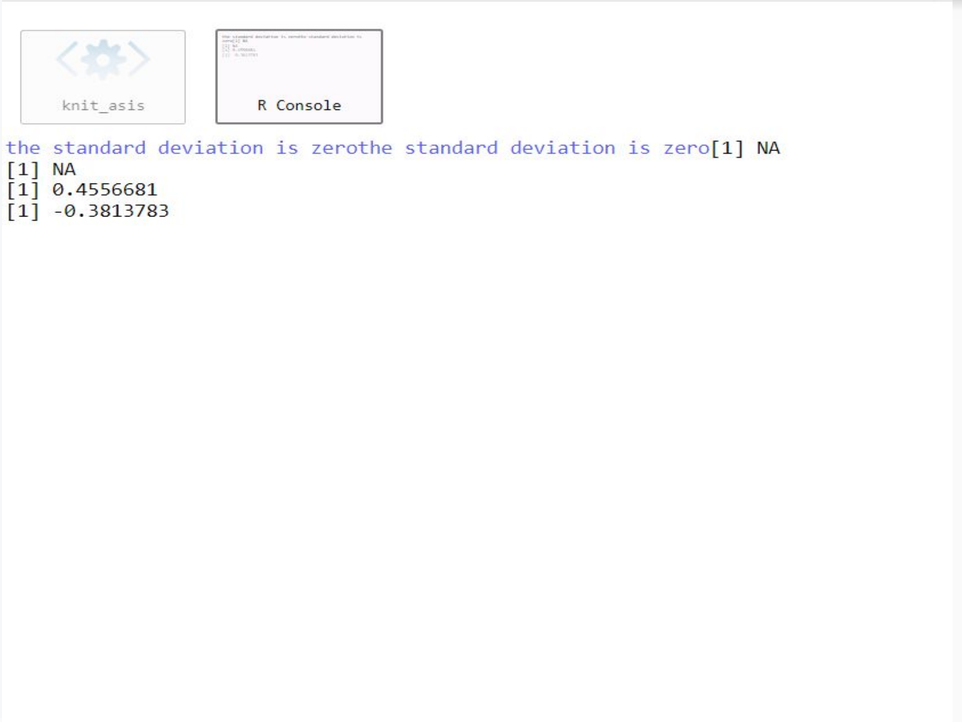
Hypothesis Test for Q3:

- For the Hypothesis test, we set the null hypothesis to be that there is no correlation between spectra and the shape of a galaxy. Hence the obtained p-values would be lower than 0.05 (our significance level).
- The alternate hypothesis was that there is a correlation.
- We set up the test with the two vectors containing the mean spectra value (x) for each class of galaxy and then the related variable (y), the likelihood of a galaxy being both spiral and elliptical.
- We found that for all classes, the likelihood was differed between a galaxy being either spiral or elliptical, however for all classes, the likelihood for each was the same.
- Hence, we attained no correlation between the shape and spectra (colour) of a galaxy. We observed the same y value for each observation and therefore could fail to reject the null hypothesis. We believe that even though the results of the code show that the null hypothesis is undetermined for one of the shapes (spiral), we could not possibly have a correlation if the likelihood did not range at all.

Code and Results for Q3



~~~
```{r}

#hypothesis test:

spectra_spiral <- galaxydata %>%
  filter(SPIRAL == 1) %>%
  summarize(across(P_EL:P_CS, mean, na.rm = TRUE))

spectra_elliptical <- galaxydata %>%
  filter(ELLIPTICAL == 1) %>%
  summarize(across(P_EL:P_CS, mean, na.rm = TRUE))

df <- bind_rows(
  spectra_spiral %>% mutate(Class = "Spiral"),
  spectra_elliptical %>% mutate(Class = "Elliptical")
)

df %>%
  pivot_longer(cols = -Class, names_to = "Spectral Feature", values_to = "Mean Value") %>%
  pivot_wider(names_from = Class, values_from = "Mean Value") %>%
  knitr::kable(format = "html", caption = "Mean Spectral Features for Spiral and Elliptical Galaxies")


# Mean values for elliptical in correlation to mean values in spectra
x <- c( 0.5359473,  0.07400691, 0.0807363, 0.2050663, 0.07292444, 0.35985)
y <- c(0.09310661, 0.09310661, 0.09310661, 0.09310661, 0.09310661, 0.09310661)


# Calculate the correlation coefficient
correlation <- cor(x, y)

# Perform hypothesis test
p_value <- cor.test(x, y)$p.value

print(p_value)
print(correlation)

# Mean values for spiral in correlation to mean values in spectra

z <- c(0.2847918, 0.2847918, 0.2847918, 0.2847918, 0.28479185, 0.2847918)

# Calculate the correlation coefficient
correlation <- cor(x, z)

# Perform hypothesis test
p_value <- cor.test(x, z)$p.value

print(p_value)
print(correlation)


```
~~~



```
the standard deviation is zerothe standard deviation is zero[1] NA
[1] NA
[1] 0.4556681
[1] -0.3813783
```

Contributions

Project Description -> Aarya

Our Plan -> Rafaye

Project Objectives -> Zaid

Question 1 -> Rafaye

Question 2 -> Aarya

Question 3 -> Zaid