Data Analysis Report – Draft 1: EDA 36-290 Fall 2021 due october 1

christina choi

9/26/2021

{r} setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)

```
file.path = "https://github.com/pefreeman/36-290/raw/master/PROJECT_DATASETS/ACTIVE_CLASS/active_class.Rdata"
load(url(file.path))
rm(file.path)
```

#Introduction

##Classification of Active Galaxies Observed by SDSS

The Sloan Digital Sky Survey has a catalog containing data on over 200 million of galaxies. Galaxy data typically involves images along with measures of brightness separated into five different bandpasses (u,g,r,i, and z) spanning the electromagnetic spectrum. A spectrum can reveal interesting information about galaxies. When a galaxy is *active*–meaning it forms stars at a relatively greater rate or has a supermassive black hole in its center that engulfs stars/gas/dust at an enhanced rate–its spectrum will reveal “spikes” called emission lines. These emission lines along with other features from a spectra can be used to make inferences about whether the galaxy is star-forming or whether it has active nucleus.

We will attempt to classify galaxies as either starform or having an active nucleus using features from their spectra.

There are a total of nine predictor variabels used in this investigation:

-**z** :galaxy redshift

redshift refers to the ratio of the observed wavelength of a photon from an object to its wavelength when it was emitted, minus 1

-**O3_Hb** :the relative strength of the [O III] emission line at 500.7 nanometers and the H beta line at 486.1 nm

O III refers to an emission line associated with oxygen atoms that are two electrons short of a full set of electrons

-**O2_Hb** :the relative strength of the [O II] emission line at 372.6 nanometers and the H beta line at 372.9 nm

O II refers to an emission line associated with oxygen atoms that are one electron short of a full set of electrons

-**sigma_star** the standard deviation of star velocities in the galaxy

-**sigma_o3** : the width of the [O III] line

-**u_g, g_r, r_i, i_z**: the four colors of the galaxy

#Data

```
## [1] 0
```

```
summary(df)
```

```
##        z              O3_Hb              O2_Hb            sigma_o3
##  Min.   :0.4000   Min.   :-1.02107   Min.   :-0.5577   Min.   :1.453
##  1st Qu.:0.5005   1st Qu.:-0.02219   1st Qu.: 0.1825   1st Qu.:1.957
##  Median :0.5784   Median : 0.40971   Median : 0.3375   Median :2.065
##  Mean   :0.5837   Mean   : 0.33569   Mean   : 0.3245   Mean   :2.110
##  3rd Qu.:0.6637   3rd Qu.: 0.66763   3rd Qu.: 0.4692   3rd Qu.:2.271
##  Max.   :0.8000   Max.   : 1.50399   Max.   : 1.3656   Max.   :2.749
##    sigma_star         u_g               g_r               r_i
##  Min.   :0.000   Min.   :-0.1599   Min.   :-0.1078   Min.   :-0.1589
##  1st Qu.:2.067   1st Qu.: 0.3470   1st Qu.: 0.6480   1st Qu.: 0.4559
##  Median :2.298   Median : 0.7572   Median : 1.0363   Median : 0.6590
##  Mean   :2.146   Mean   : 0.7323   Mean   : 0.9928   Mean   : 0.6234
##  3rd Qu.:2.594   3rd Qu.: 0.9882   3rd Qu.: 1.3640   3rd Qu.: 0.8004
##  Max.   :2.929   Max.   : 3.3347   Max.   : 2.5594   Max.   : 1.8681
##       i_z               label
##  Min.   :-0.3709   STARFORM:15521
##  1st Qu.: 0.1618   AGN     :13299
##  Median : 0.3047
##  Mean   : 0.2978
##  3rd Qu.: 0.4174
##  Max.   : 1.0742
```

```
names(df)
```

```
##  [1] "z"          "O3_Hb"      "O2_Hb"      "sigma_o3"   "sigma_star"
##  [6] "u_g"        "g_r"        "r_i"        "i_z"        "label"
```

```
Here is a summary of the dataset that we will be working with, and a list of the names of the nine predictor vari
ables that we will be working with. Upon further investigation, it seems that there is no missing data.

The following boxplot visualizes the summary above. It appears that g_r and z are fairly spread out and z appears
to be without much skew to either side. On the other hand, r_i and u_g appear fairly right skewed.
```

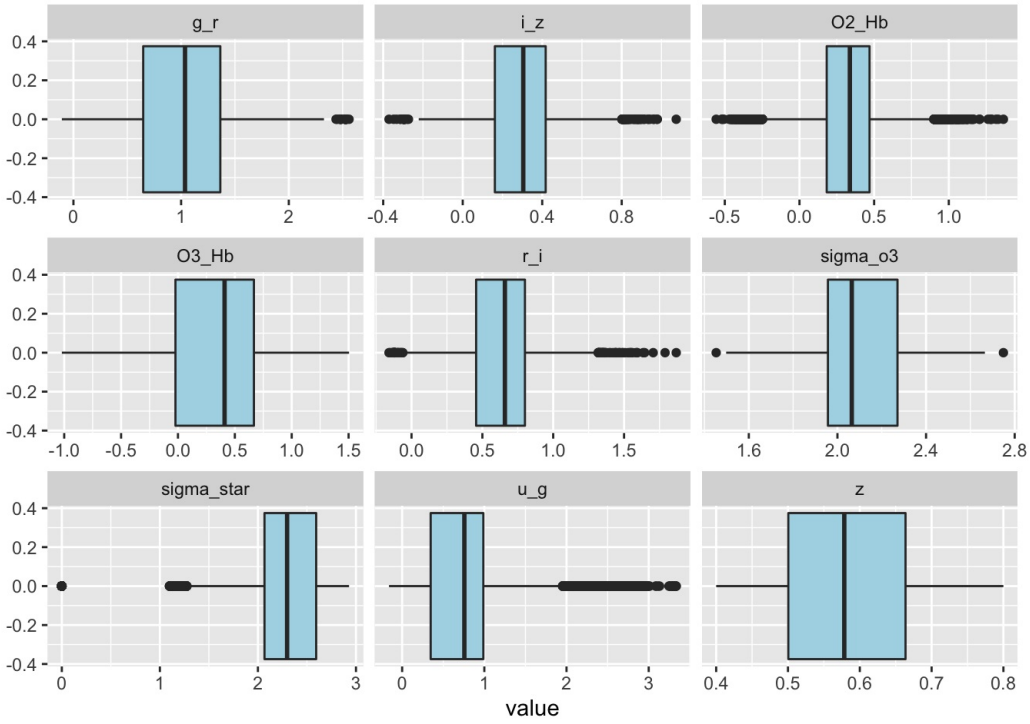
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
df.gathered = df %>% select(.,u_g,g_r,r_i,i_z,O3_Hb,O2_Hb,sigma_o3,sigma_star,z) %>% gather(.)
ggplot(data=df.gathered,mapping=aes(x=value)) +geom_boxplot(fill="lightblue") +
  facet_wrap(~key, scales='free_x')
```




```
library(GGally)
```

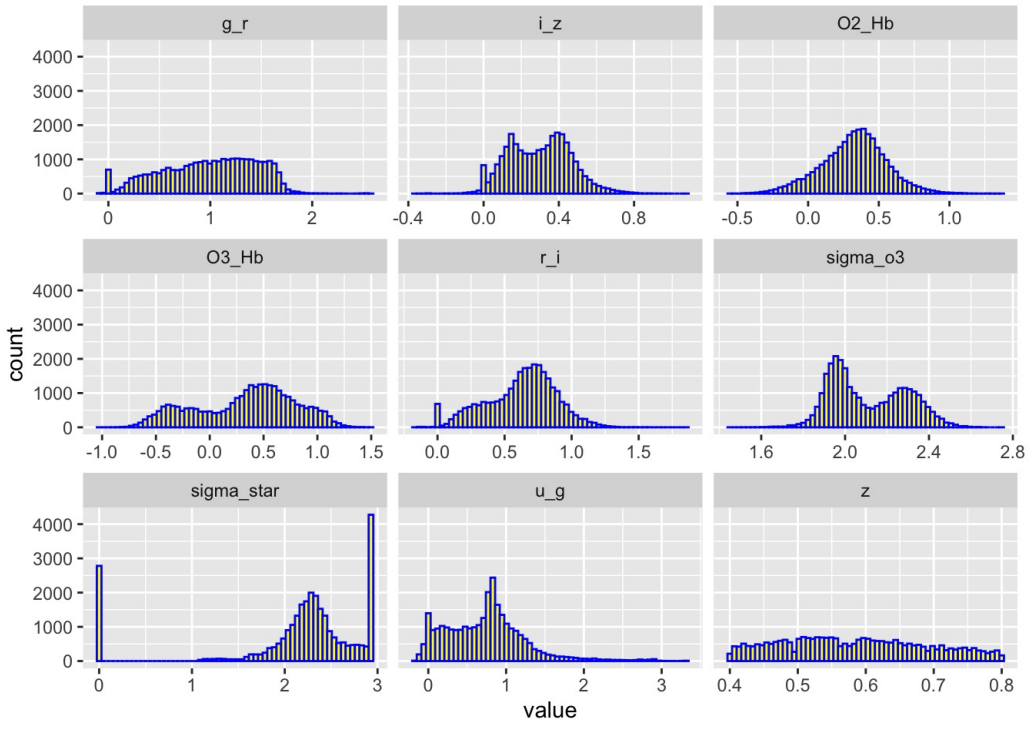
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(dplyr)
library(tidyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

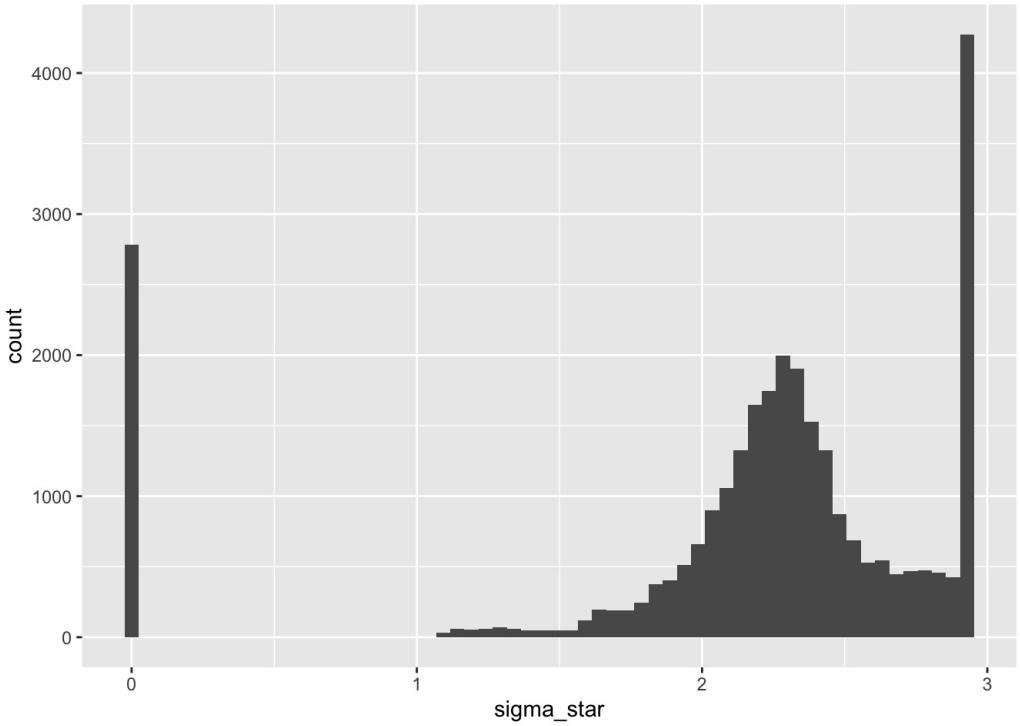
```
df.gathered = df %>% select(.,u_g,g_r,r_i,i_z,O3_Hb,O2_Hb,sigma_o3,sigma_star,z) %>% gather(.)
ggplot(data=df.gathered,mapping=aes(x=value)) +geom_histogram(color="blue",fill="yellow",bins=60) +
  facet_wrap(~key, scales='free_x')
```



Looking at the histograms, O3_Hb, r_i, sigma_o3, and i_z appear bimodal, with two distinct peaks in their distrib
ution. O2_Hb is the only graph that can be described as closest to a symetric, normal distribution. There are som
e outliers in u_g, g_r, r_i, i_z, O3_Hb, sigma_o3, sigma_star, and z that might need to be cut from the data.

```
library(ggplot2)

ggplot(data=df,mapping=aes(x=sigma_star)) +geom_histogram(bins=60)
```



Sigma_star in particular seems to have two distinct outliers that does not fit well with the overall graph, but g
iven the high frequency of these particular outliers it is uncertain if cutting them from the data will substanti
ally change the results. We will cut the outliers for now and perhaps check later on how the results will differ
with and without the outliers.

2
4
5
11
14
15
16
17
19
20

1-10 of 10,000 rows | 1-1 of 11 columns Previous **1** 2 Next

#References


```
overview on EDA:
https://r4ds.had.co.nz/exploratory-data-analysis.html

Documentation on subset()
https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/subset
```