# A Data Science Approach to Galactic Compactness and Environment

*Kaelee Parker, Case Western Reserve University*

*04 May, 2019*

## Contents





```
# Loading in packages
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------------------------

## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------------- tidy
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(magrittr)
```

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

## The following object is masked from 'package:purrr':
##
##     set_names

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

```r
library(ggplot2)
library(imager)
```

```
## Warning: package 'imager' was built under R version 3.5.2

##
## Attaching package: 'imager'

## The following object is masked from 'package:magrittr':
##
##     add

## The following object is masked from 'package:stringr':
##
##     boundary

## The following object is masked from 'package:tidyr':
##
##     fill

## The following objects are masked from 'package:stats':
##
##     convolve, spectrum

## The following object is masked from 'package:graphics':
##
##     frame

## The following object is masked from 'package:base':
##
##     save.image
```

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(GGally)
```

```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```r
library(leaps)

# Setting up dataframes
```

```r
data <- as.data.frame(read.table('ScaleLengths.tsv', sep = '\t'))
colnames(data) <- c('Right ascension (J2000) [deg]', 'Declination (J2000) [deg]',
                    'Right ascension 2 (J2000)', 'Declination 2 (J2000)', 'z',
                    'z_err', 'Morphological type', 'type_err', 'rMAG', 'Ar',
                    'r_sc [pc]', 'mu0', 'Sloan', 'DR7')

data %<>%
  select('Right ascension (J2000) [deg]', 'Declination (J2000) [deg]', 'z',
         'z_err', 'Morphological type', 'type_err', 'rMAG', 'Ar',
         'r_sc [pc]')

data <- data[-c(1, 2, 3), ]
data <- map_df(data, function(k) {
  return(as.numeric(as.character(k)))
})

# Adding in the galaxies from SDSS
SDSS <- read.csv('SkyServer_5x5.csv')
SDSS %<>% distinct(SDSS$ra, .keep_all = TRUE)
SDSS <- filter(SDSS, SDSS$dec < 55)
# Filtering Vizier data to cover same region of space as SDSS data
Vizier <- filter(data, data$z < 0.085
                 & data$`Right ascension (J2000) [deg]` > 120
                 & data$`Right ascension (J2000) [deg]` < 215
                 & data$`Declination (J2000) [deg]` > 30
                 & data$`Declination (J2000) [deg]` < 55
                 & data$z_err < 0.00075)

# Combining the datasets
tol = 0.0001
`%~%` <- function(x,y) {
out <- logical(length(x))
for(i in 1:length(x)) out[i] <- any(abs(x[i] - y) <= tol)
out
}

# Narrowing datasets down to desired quantities
SDSS %<>%
  select('ra', 'dec', 'z')
SDSS$sc <- NaN

Vizier %<>%
  select('Right ascension (J2000) [deg]', 'Declination (J2000) [deg]', 'z', 'r_sc [pc]')
colnames(Vizier) <- c('ra', 'dec', 'z', 'sc')

# Finding overlapping galaxies
matches <- SDSS[, 1] %~% Vizier[, 1]
# Removing overlapping galaxies from SDSS
SDSS <- SDSS[!matches, ]

# Denoting data and combining into one
SDSS$source <- 'SDSS'
Vizier$source <- 'Vizier/SDSS'
```

```r
All <- rbind(SDSS, Vizier) # All the galaxies: Vizier + SDSS
```

# 1 Abstract

The question that I am approaching through this project is whether spiral galaxies in high density environments are more compact than their low density environment counterparts. Mechanisms like tidal stripping would likely have this effect so a component of this project is to determine whether they have a noticeable contibution to the dynamics of cluster galaxies. Through a number of data science methods, I have determined that, at least for my sample of 58,924 galaxies, there is actually an opposite effect. Spiral galaxies in higher density regions tend to be less compact than those in lower density regions. I do not have an explanation as of yet on what could cause this trend.

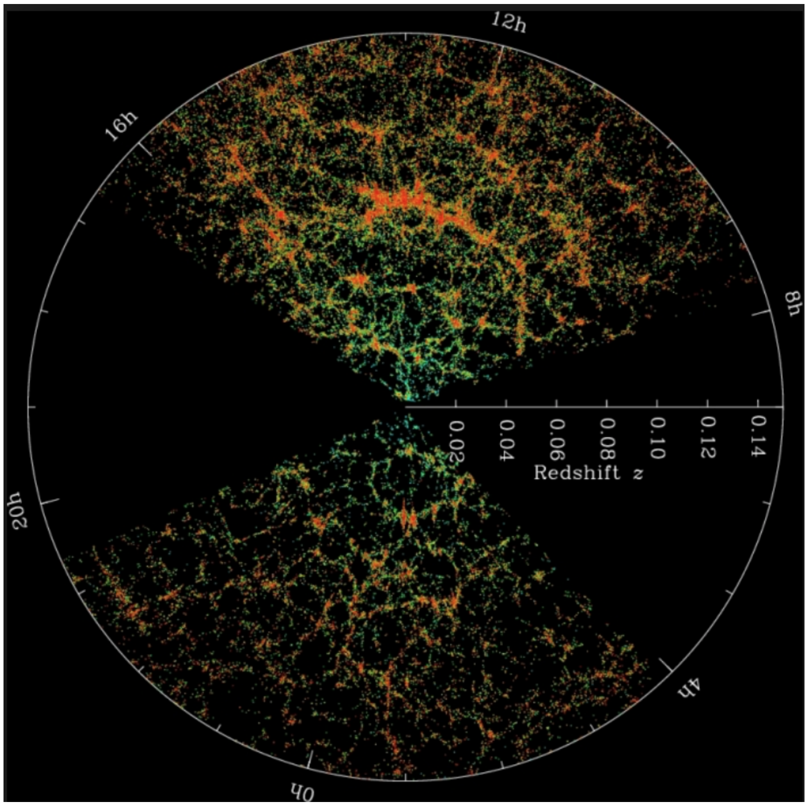
# 2 Introduction

## 2.1 Large Scale Structure of the Universe

A common expression in astronomy is that "the universe is homogenous on the largest scales." This, of course, is an approximation when looking at some of the largest scales of the known universe because if it were true on a smaller scale, galaxies, planets, and other lumps would not exist. Even on the scales that astronomers often observe though, the universe is still full of lumps and structure. We see galaxies and clusters of galaxies, filaments of high density and regions void of nearly everything. This structure to the universe originated from extremely small fluctuations in density at the time of the big bang and has since been exaggerated by the effects of gravity. More dense regions have a stronger gravitational potential and thus attract more matter then less dense regions.

The figure below shows a set of observed galaxies from the Sloan Digital Sky Survey. This image shows areas of higher and lower density, looking at a single slice of the sky. The distance from the origin is approximated by the redshifts of the galaxies.

```r
knitr::include_graphics("structure.png")
```




Since there are different environments in the universe (high density versus low density regions), it is possible to study how the environment of a galaxy effects its properties. This is an idea I will be investigating in my project.

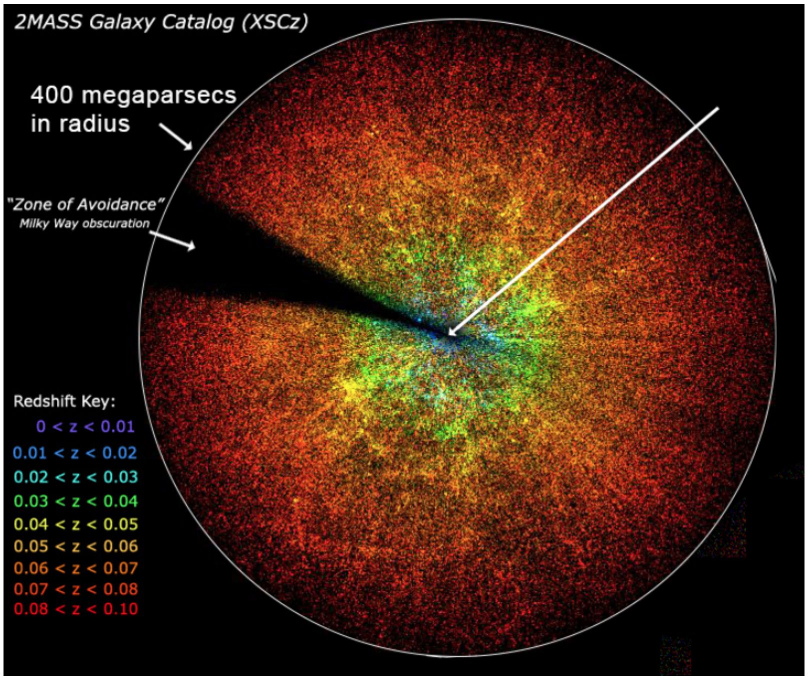
## 2.2 Using Redshift as a Distance Proxy

Redshift is a value that quantifies how quickly an object is moving away from us. As an object moves away, the light from it is stretched slightly, resulting in a longer wavelength/redder light reaching Earth. Since the universe is expanding and accelerating, galaxies that are farther away are moving away from us faster than the closer galaxies are moving away. This means that, in general, farther galaxies will have a larger redshift than nearer galaxies. However, not all motion is due to the expansion of the universe, some motion is from the proper motion of the galaxies themselves. For example, the Andromeda galaxies is blueshifted rather than redshifted because it is moving towards us. This kind of motion- motion that is not due to the expansion of the universe- is called proper motion and is what makes redshifts less reliable as a distance indicator.

For the most part the redshifts should be an adequate proxy for distance, especially since my project is focused more on environment than on exact distance values. It would be nearly impossible to get such a complete dataset using a difference method for distance measurements which is why redshifts are the best we can do with this kind of project.

The figure below shows the distribution of galaxies from the 2MASS Galaxy Catalog along with their redshifts versus estimated distances. For the most part, more distant galaxies have larger redshifts but there are certainly some regions where redshift is not a perfect distance indicator.

```
knitr::include_graphics("redshift.png")
```




## 2.3 Galaxy Scale Length

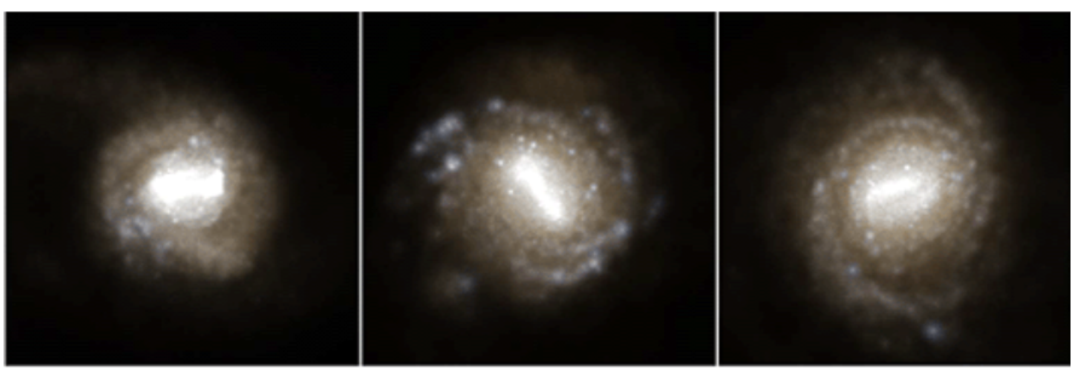
The main property that I am investigating in this project is compactness of spiral galaxies. The compactness of a spiral galaxy can be described by the galaxy's scale length. A scale length is a measure for how quickly the brightness profile for the galaxy decreases. This is shown in the equation below:

$$b = b_0 e^{-r/r_0}$$

where $b$ is the brightness at radius $r$, $b_0$ is the brightness at the center of the galaxy, and $r_0$ is the scale length.

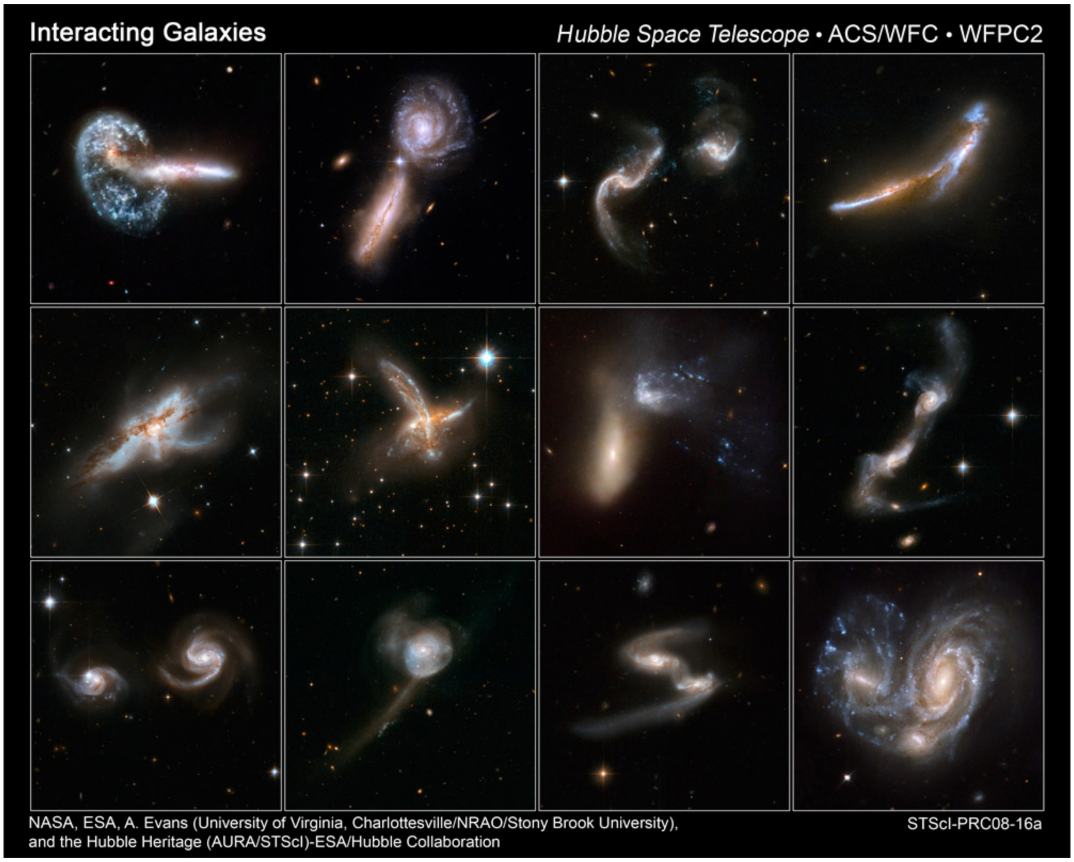
The figure below shows an example of galaxies with different scale lengths. Scale length is not a measure of a galaxy's size, as can be seen below.

```r
knitr::include_graphics("scalelength.png")
```



The compactness of a galaxy can change over the course of its lifetime. Galaxies often interact with each other, and especially in higher density regions, it can be common for smaller galaxies to have gas and/or stars stolen from their outskirts by larger galaxies. This phenomenon is known as tidal stripping. The figure below shows an example of this kind of interaction. If tidal stripping is rather prevalent in high density regions of the universe, it might be expected that the galaxies in these regions could be noticeably more compact than those in lower density environments where tidal stripping is less common.

```
knitr::include_graphics("interactions.jpg")
```




## 2.4 My Dataset - Aquisition and Cleaning

The dataset I am using to answer this question is from two separate sources. The first set is from the database Vizier and consists of 30,000 spiral galaxies, each with a v (essentially longitude and latitude on the sky), a redshift and a scale length. This should allow me to analyze the three dimensional structure of the galaxies in the dataset and determine whether scale length is dependent on a galaxy's environment.

However this dataset is not a complete census of the galaxies in each region of the sky. In fact, when loading in data for the galaxies in the regions from Vizier from the Sloan Digital Sky Survey (SDSS), there are approximately 10 times as many galaxies than are contained within Vizier. Unfortunately SDSS does not have the scale lengths for those galaxies but I will use their positions in space (right ascension, declination, and redshift) to include them in my calculations for environment.

In order to lead to the most accurate analysis of the data, I needed to ensure my data was volume-limited, accurate, and in complete regions of the sky. To ensure this, I only used galaxies that have a redshift that is less than 0.085 with an error of 0.00075 or less. I chose this value because a histogram of the number of galaxies per redshift had a sharp cutoff around a redshift of 0.085 which means beyond this point, many of the galaxies are not being included. There is also an effect due to the spherical geometry of the region, but using a cutoff of 0.085 is conservative enough that other effects would be neglidgable. It is necessary to make this cutoff in order to ensure that even the faintest galaxies are represented in the dataset. I also limited the data to regions that appear to be sampled more thoroughly. I focused on a specific region that where $120 < RA < 215$ and $30 < DEC < 55$. In this region, there are 6,072 galaxies that are in both the Vizier and SDSS datasets and 52,852 additional galaxies added by the SDSS dataset, leading to a total of 58,924 galaxies to work with for this project.

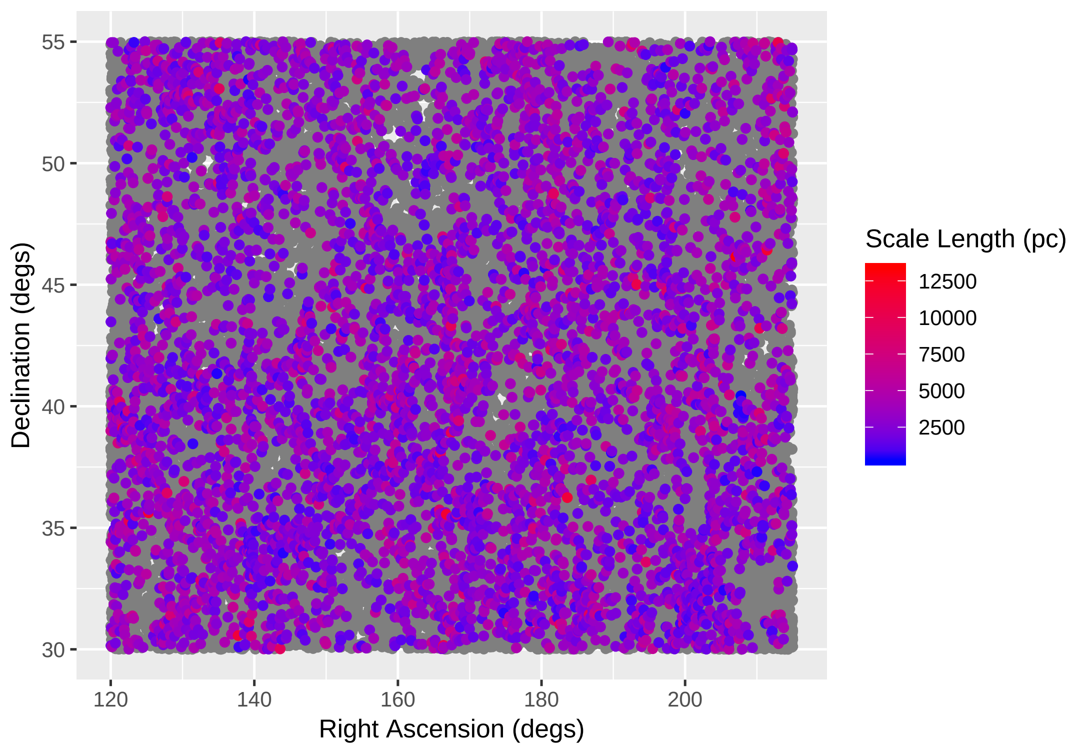
Below is the header of the complete dataset (Vizier + SDSS) for this project:

```
head(All)
```

```
##         ra      dec          z  sc source
## 1 151.9795 39.82400 0.06327377 NaN   SDSS
## 2 151.7876 39.88637 0.08028392 NaN   SDSS
## 3 151.8330 39.89106 0.08051044 NaN   SDSS
## 4 153.9673 39.02075 0.06320294 NaN   SDSS
## 5 152.1329 39.93545 0.05747681 NaN   SDSS
## 6 154.0141 39.44171 0.05500858 NaN   SDSS
```

A 2D image of the full data set:

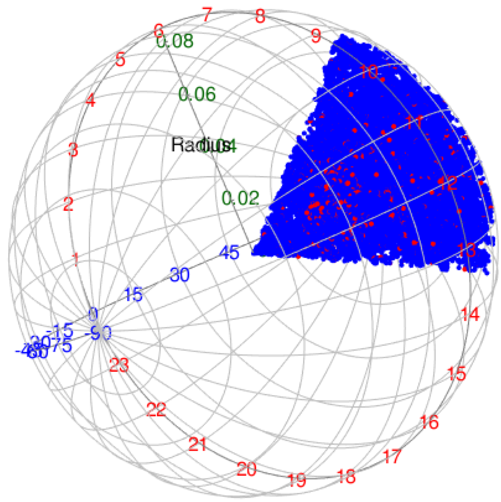
```
# 3D data projected onto 2D space
ggplot(data = All, aes(x = All$ra, y = All$dec, color = All$sc)) + geom_point() +
  scale_color_gradient(low = "blue", high = "red") + labs(x = 'Right Ascension (degs)',
                                                         y = 'Declination (degs)', color = 'Scale Leng
```



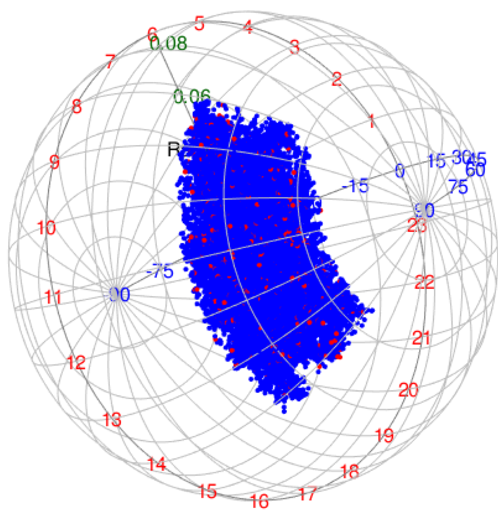

The colored points are the galaxies in both datasets, as they are the ones with known scale lengths. The gray points are the ones that were only in the SDSS dataset and therefore do not have known scale lengths.

A 3D image of the full data set:

```
# Created using sphereplot package
knitr::include_graphics("both1.png")
```




```r
knitr::include_graphics("both2.png")
```




The blue points are for galaxies that are only in SDSS while the red points are the ones that are in both datasets.

# 3 Data Science Methods

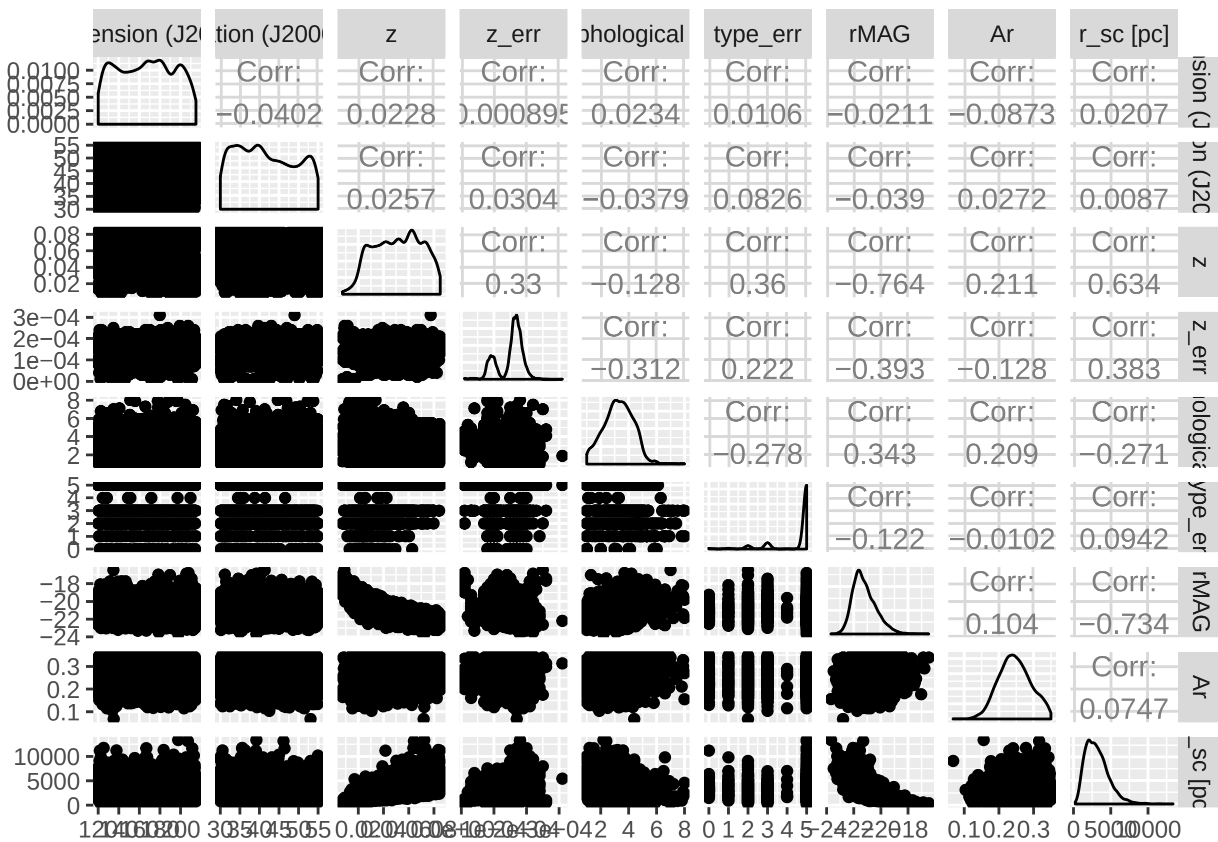
This project used a variety of data science methods. The main ones include k-means clustering, principal component analysis, subset selection, and multiple linear regression. These will be expored further with respect to my data in Section 5.

# 4 Exploratory Data Analysis

## 4.1 Investigating Vizier Variables

The Vizier dataset includes multiple variables that I have not used in this project. Below is a pairwise plot of all the variables from Vizier.

```
# Bringing in all the Vizier variables
vars_Vizier <- filter(data, data$z < 0.085
                & data$`Right ascension (J2000) [deg]` > 120
                & data$`Right ascension (J2000) [deg]` < 215
                & data$`Declination (J2000) [deg]` > 30
                & data$`Declination (J2000) [deg]` < 55
                & data$z_err < 0.00075)

ggpairs(vars_Vizier)
```

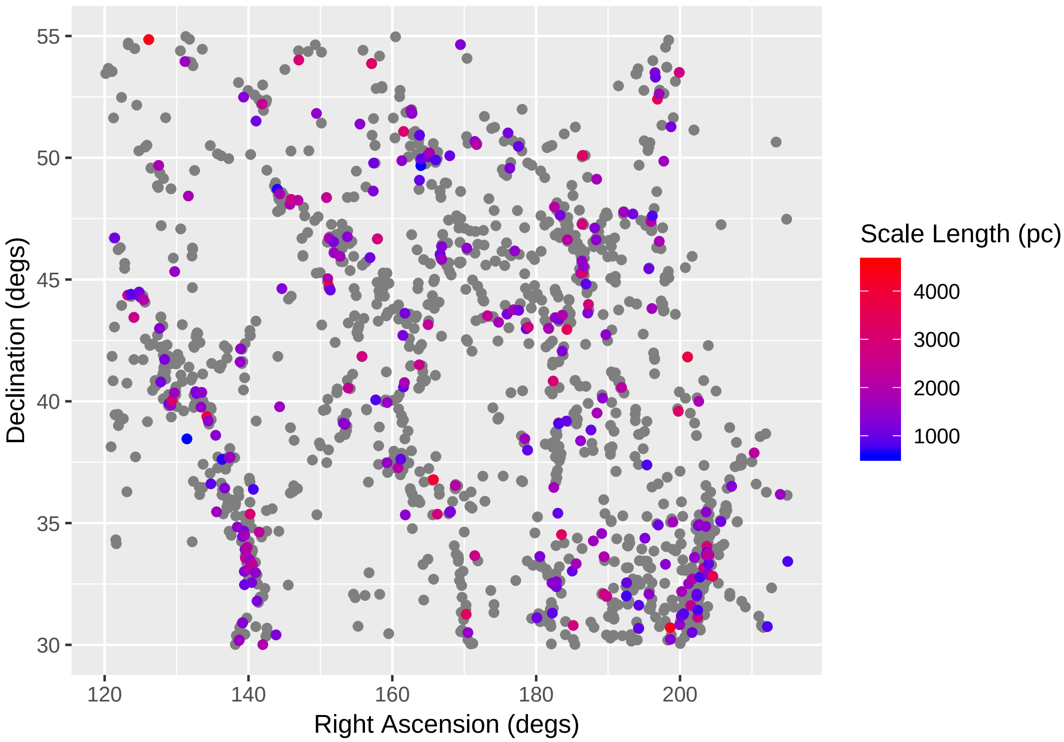


The strongest correlation is between the apparent magnitude (rMAG) and the redshift (z). This makes sense because the apparent brightness of a galaxy is extremely dependent on its distance from Earth.

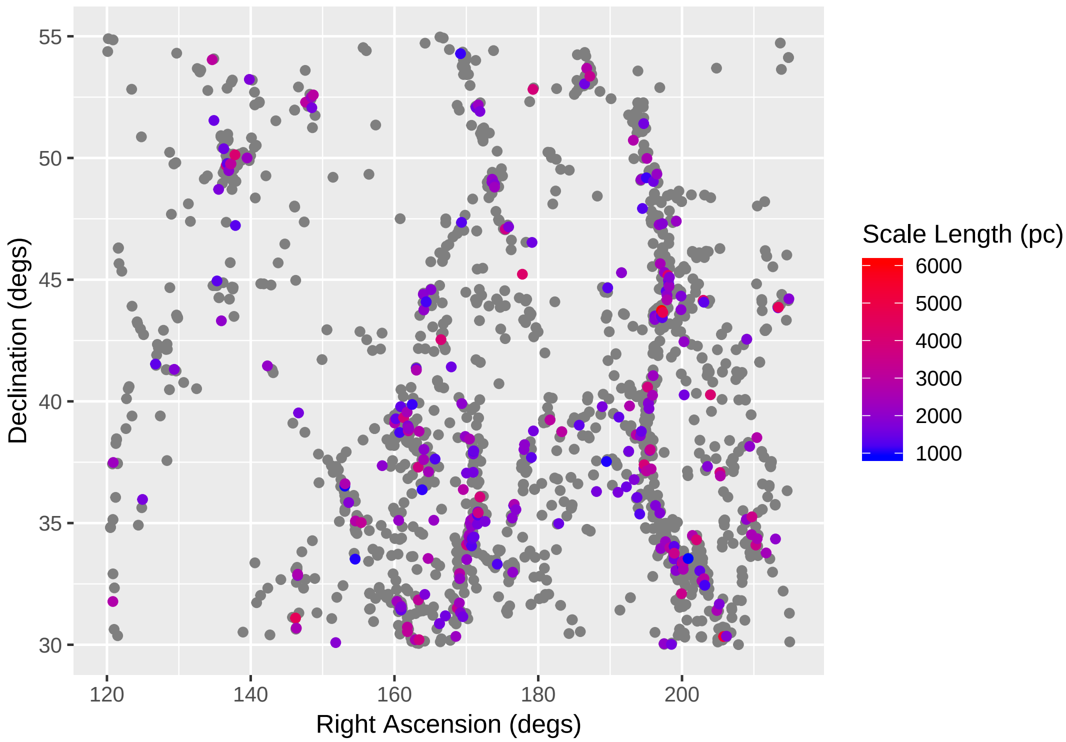
## 4.2 Thin Slices in Redshift

To get an initial idea for what to expect, I looked at thin slices in redshift to see if there were obvious trends for the scale lengths of galaxies based on environment. Below are two examples of this. Again, the colored points represent galaxies with scale lengths whereas the gray points are galaxies that just contribute to the density of the environment.

```r
# Thin slice 1
range1 <- filter(All, All$z > 0.023 & All$z < 0.025)
ggplot(data = range1, aes(x = range1$ra, y = range1$dec, color = range1$sc)) +
  geom_point() + scale_color_gradient(low = "blue", high = "red") +
  labs(x = 'Right Ascension (degs)', y = 'Declination (degs)', color = 'Scale Length (pc)')
```




```r
# Thin slice 2
range2 <- filter(All, All$z > 0.034 & All$z < 0.037)
ggplot(data = range2, aes(x = range2$ra, y = range2$dec, color = range2$sc)) +
  geom_point() + scale_color_gradient(low = "blue", high = "red") +
  labs(x = 'Right Ascension (degs)', y = 'Declination (degs)', color = 'Scale Length (pc)')
```




It is difficult to tell by eye whether or not these plots suggest a relationship between scale length of spiral galaxies and density of the environment. It is necessary to turn to more quantitative methods to confirm or deny at relationship.

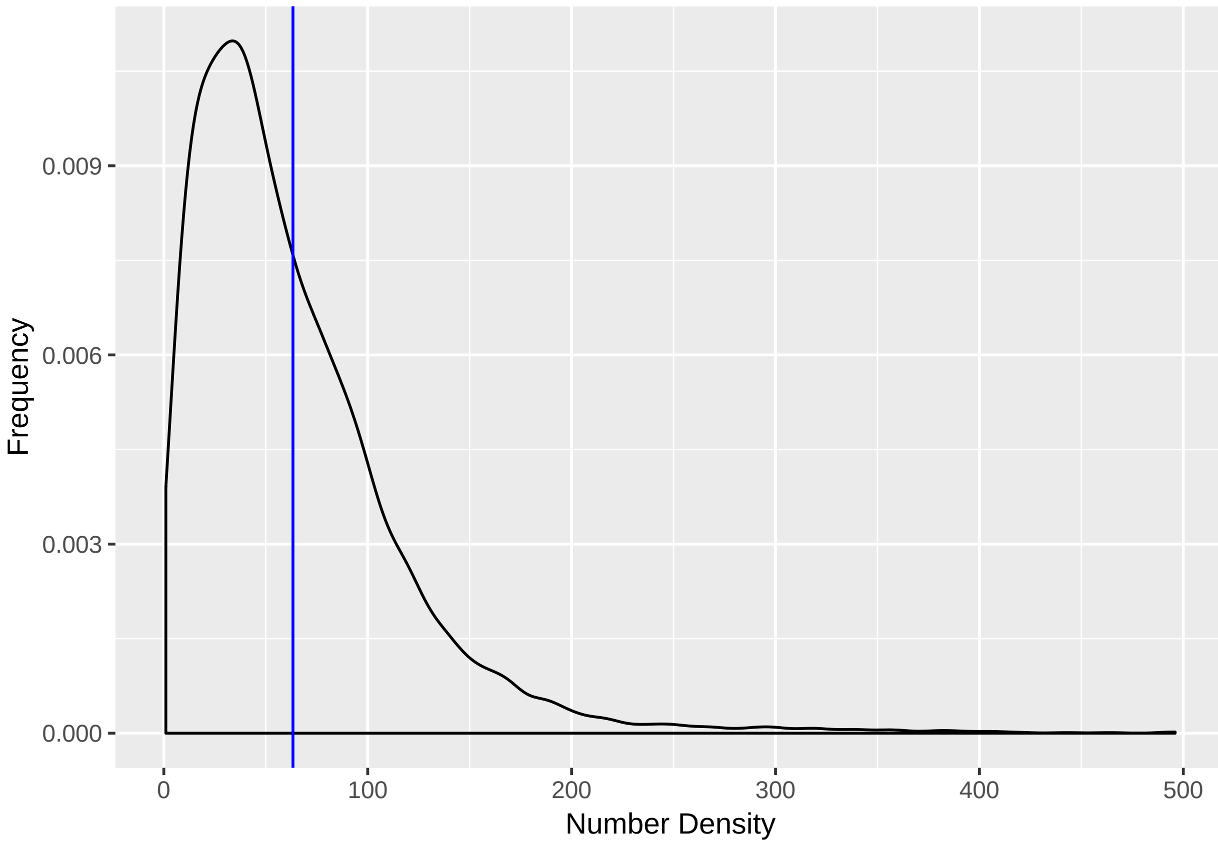
## 4.3 Fixed Volume Method

What I call the 'Fixed Volume Method' is an algorithm for determining the number of galaxies within a fixed volume of space. It centers this volume on each galaxy that has a scale length (the 4658 from Vizier) and counts the number of neighboring galaxies around it (including the SDSS ones). This provides a numeric value of density that can be used to quantify the environment. Here is the code for the algorithm:

```
# Counting the number of galaxies per fixed volume for each of the Vizier galaxies
count <- list()
counter <- 0
for (i in 1:length(All$z)) {
  if (All$source[i] == "Vizier/SDSS") {
    counter <- counter + 1
    RA <- All$ra[i]
    DEC <- All$dec[i]
    z <- All$z[i]
     # the fixed volume 1x1x0.01
    ind <- length(which(All$ra > RA - 0.5 & All$ra < RA + 0.5 & All$dec > DEC - 0.5
                        & All$dec + 0.5 & All$z > z - 0.005 & All$z < z + 0.005))
    count[counter] = ind
  }
}

Vizier$number <- as.numeric(as.character(count))
```

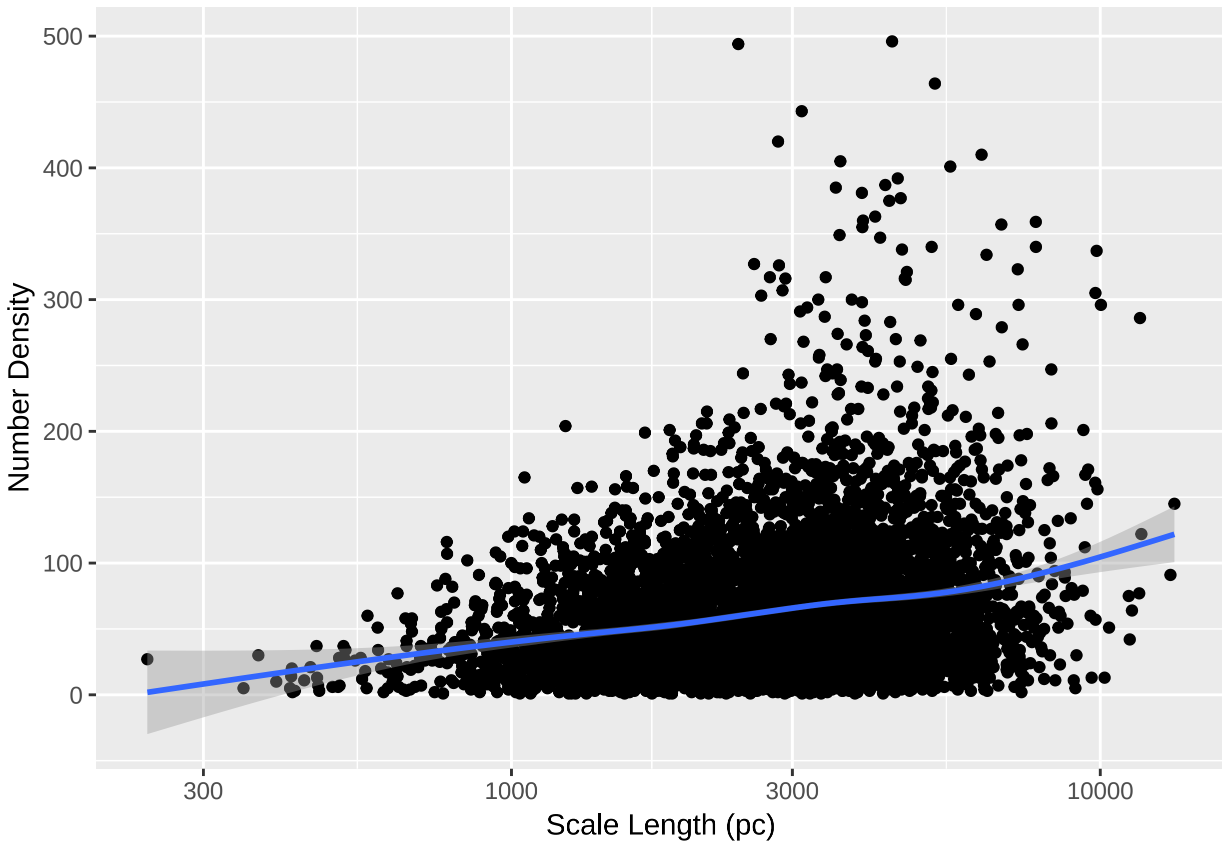
Below is a histogram of the number densities calculated for the Vizier galaxies. The average number of galaxies per 1x1x0.01 volume is about 63 (denoted with the blue line below).

```
# Distribution of number densities per fixed volume
ggplot(data = Vizier, aes(Vizier$number)) + geom_density() +
  geom_vline(xintercept = mean(Vizier$number), color = "blue") +
  labs(x = 'Number Density', y = 'Frequency')
```




The following is a plot of scale length vesus number density, where the number density is what was calculated through the Fixed Volume Method.

```
ggplot(data = Vizier, aes(x = Vizier$sc, y = Vizier$number)) + geom_point() +
  geom_smooth() + scale_x_continuous(trans = 'log10') +
  labs(x = 'Scale Length (pc)', y = 'Number Density')
```




# 5 Statistical Learning: Modeling and Prediction

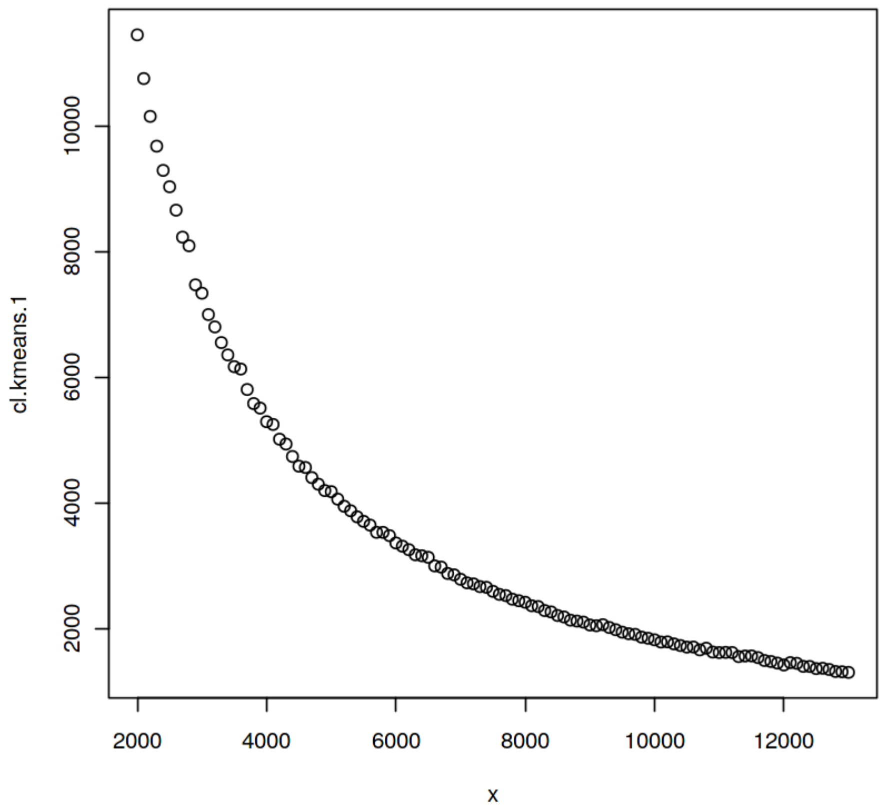
## 5.1 K-means calculations

I used Kaggle to test k-means algorithms going from 700 clusters to 13000 clusters in order to determine the ideal number of clusters for the dataset. The following code is what I ran on Kaggle:

```
# This code was run on Kaggle
df <- scale(SDSS)
distance <- dist(df, method = "euclidean")
df.big <- as.big.matrix(SDSS)

getSS <- function(k) {
    return(sum(bigkmeans(df.big, centers = k, nstart = 1, iter.max = 50,
                         dist = "euclid")$withinss))
}

x <- seq(2000, 13000, by = 100)
cl.kmeans.1 <- mclapply(x, function(k) getSS(k))
plot(x, cl.kmeans.1)
```

This code created this elbow plot:

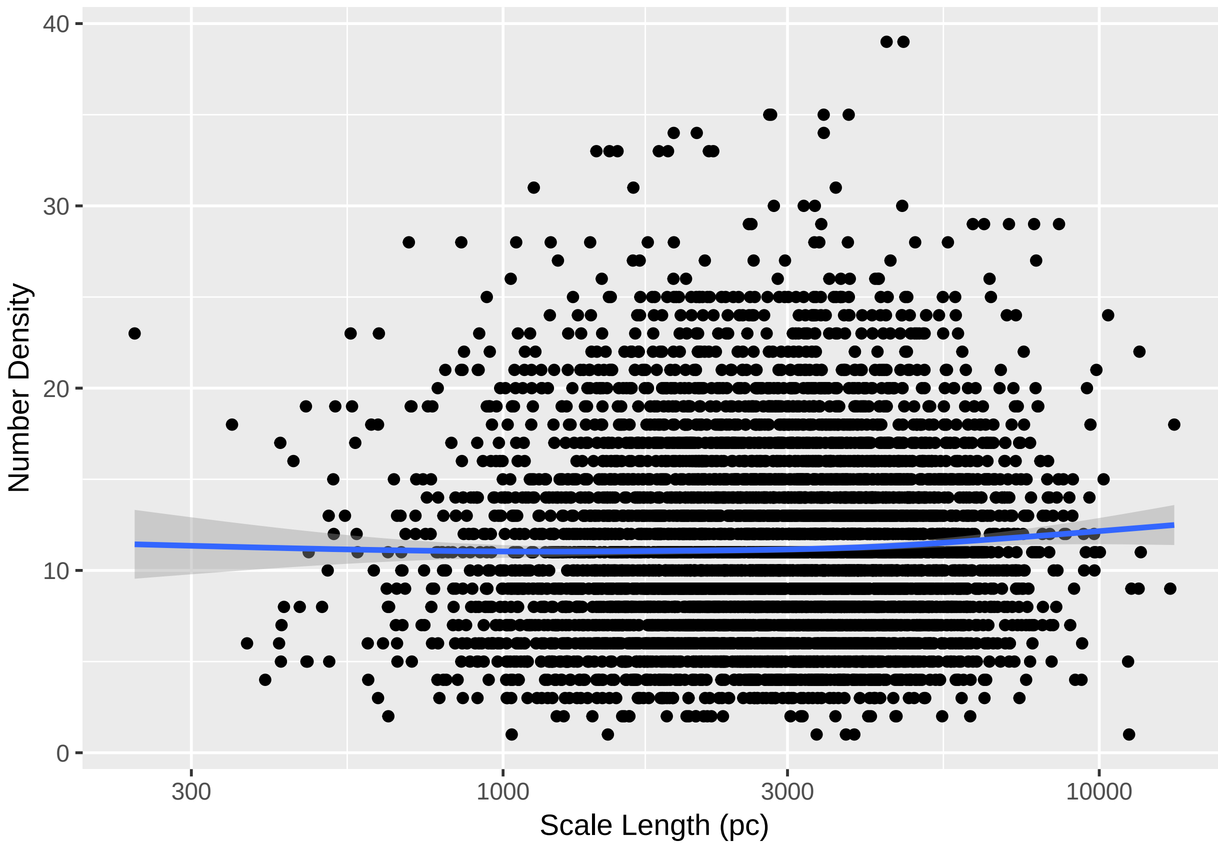
```
knitr::include_graphics("kmean_elbow.png")
```




This shows that the ideal number of clusters is probably between 6000 and 8000 clusters. For the rest of the k-means analysis, I used 7000 clusters. Below is a comparison between the results from the k-means method with 7000 clusters and the Fixed Volume Method. For the k-means method, I found the number of galaxies per each cluster.

```
# Determining cluster information through k-means
All_kmeans <- map_df(All, function(k) {
  return(as.numeric(as.character(k)))
})

All_kmeans <- All_kmeans[,-c(4, 5)]
All_kmeans <- as.data.frame(scale(All_kmeans))
obj <- kmeans(All_kmeans, 7000, iter.max = 10, nstart = 1, trace = FALSE)

# Counting the number of galaxies per cluster
kmeans_count <- list()
for (i in 1:length(All_kmeans$z)) {
  kmeans_count[i] = length(which(obj$cluster == obj$cluster[i]))
}

All$count <- as.numeric(as.character(kmeans_count))

# Plotting in the same manner as the Fixed Volume Method
ggplot(data = All, aes(x = All$sc, y = All$count)) + geom_point() +
```

```r
geom_smooth() + scale_x_continuous(trans = 'log10') +
labs(x = 'Scale Length (pc)', y = 'Number Density')
```




This method did not seem to lead to as useful of a result as the Fixed Volume Method. It did help determine an optimal clustering size though.

## 5.2 Multiple Linear Regression

Using the Vizier data and the number density from the Fixed Volume Method, I worked on creating a multiple linear regression to help determine whether or not there is a trend between scale length and density of environment. One of the better models is shown below, dependent on the log of scale length and the interactions between RA, DEC, and redshift. This model led to an adjusted $R^2$ of 0.4005. Although this is not a 'good' fit by any means, the question is whether it is enough to confirm or deny whether there is a relationship.

```r
# Create a linear model using the scale length and interactions between the coordinates
y.lm <- lm(number ~ log10(sc) + (ra + dec + z)^2, data = Vizier)
summary(y.lm)
```

```
##
## Call:
## lm(formula = number ~ log10(sc) + (ra + dec + z)^2, data = Vizier)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -133.33  -22.21   -5.88   14.34  359.43
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.004e+01  2.433e+01   0.413  0.67995
## log10(sc)    1.719e+01  3.203e+00   5.368 8.25e-08 ***
## ra           8.935e-03  1.200e-01   0.074  0.94066
## dec          3.366e-01  4.739e-01   0.710  0.47758
## z            1.427e+03  2.421e+02   5.894 3.97e-09 ***
## ra:dec      -8.094e-03  2.570e-03  -3.149  0.00165 **
## ra:z         9.064e+00  1.037e+00   8.737  < 2e-16 ***
## dec:z       -5.404e+01  3.870e+00 -13.965  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.45 on 6064 degrees of freedom
## Multiple R-squared:  0.4012, Adjusted R-squared:  0.4005
## F-statistic: 580.4 on 7 and 6064 DF,  p-value: < 2.2e-16
```

### 5.3 Principle Component Analysis and Subset Selection

To try and improve this fit, or at least investigate it further, I applied principal component analysis to the Vizier dataset. This lead to an adjest $R^2$ of 0.2857.

```r
vars_Vizier %<>%
  select('Right ascension (J2000) [deg]', 'Declination (J2000) [deg]', 'z',
         'r_sc [pc]', 'Morphological type', 'Ar')

vars_Vizier$`r_sc [pc]` <- log10(vars_Vizier$`r_sc [pc]`)
vars_Vizier <- map_df(vars_Vizier, function(k) as.numeric(k))

# Determine principal components
pr.mat <- prcomp(as.matrix(vars_Vizier), tol = 0.01)$x
pr.mat <- cbind.data.frame(data.frame(pr.mat), number = unlist(count))

# Create linear model for principal components
y2.lm <- lm(number ~ ., data = pr.mat)
summary(y2.lm)
```

```
##
## Call:
## lm(formula = number ~ ., data = pr.mat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -95.90 -25.32  -6.67  14.74 390.81
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 63.33333    0.56666 111.767  < 2e-16 ***
## PC1          0.16930    0.02085   8.119 5.63e-16 ***
## PC2          3.68768    0.07678  48.028  < 2e-16 ***
## PC3          4.13726    0.54003   7.661 2.13e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 44.16 on 6068 degrees of freedom
## Multiple R-squared:  0.2861, Adjusted R-squared:  0.2857
## F-statistic: 810.4 on 3 and 6068 DF,  p-value: < 2.2e-16
```

The above model shows three principal components so I used subset selection to determine the best component for this data.

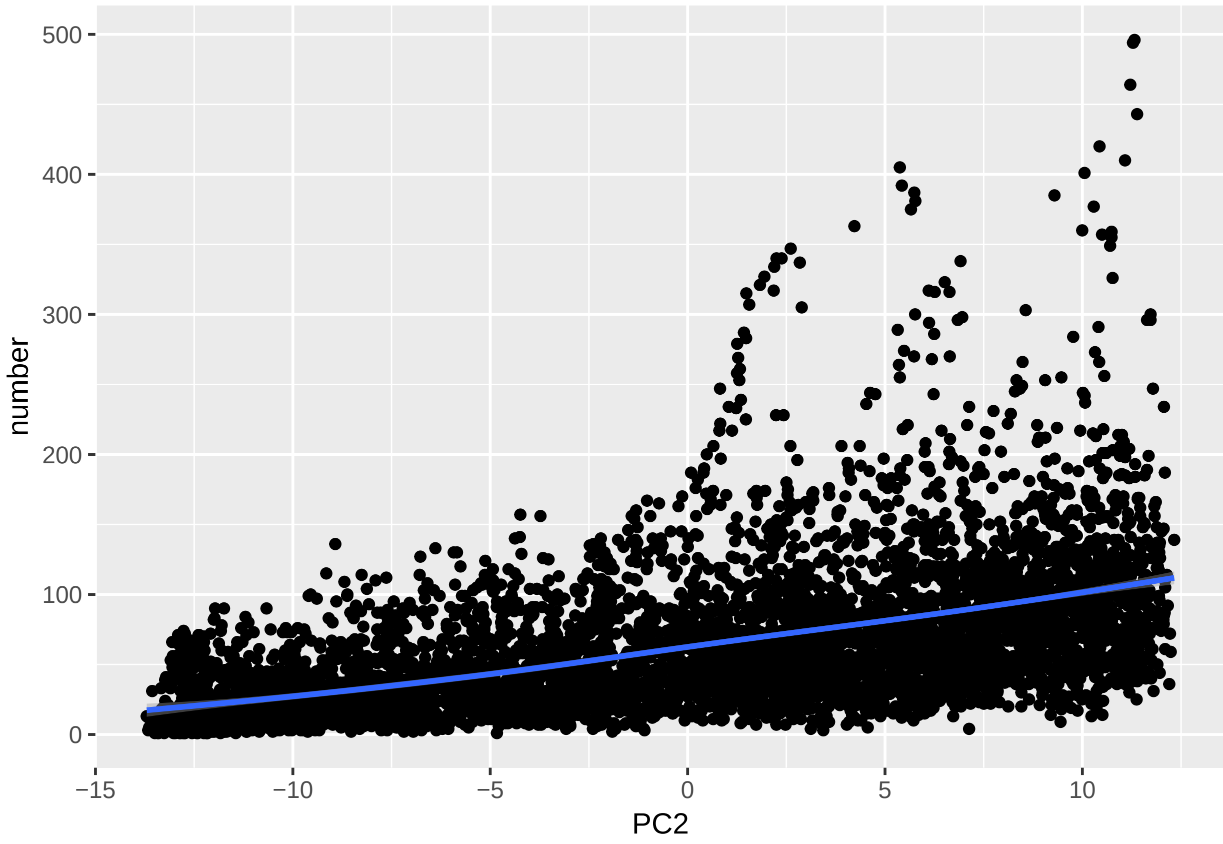
```r
pcaSub <- regsubsets(x = select(pr.mat, -number), y = pr.mat$number)
summary(pcaSub)
```

```
## Subset selection object
## 3 Variables  (and intercept)
##     Forced in Forced out
## PC1     FALSE      FALSE
## PC2     FALSE      FALSE
## PC3     FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          PC1 PC2 PC3
## 1  ( 1 ) " " "*" " "
## 2  ( 1 ) "*" "*" " "
## 3  ( 1 ) "*" "*" "*"
```

The second principal component is the most significant. Here is a plot of this principal component.

```r
ggplot(data = pr.mat, aes(x = PC2, y = number)) + geom_point() + geom_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



### 5.4 Confidence Intervals

Using the multiple linear regression model from Section 5.2, I found estimates for the fits at two different scale lengths (1000 pc and 4000 pc) at the same location in RA and DEC. These predictive models made estimates for the confidence intervals of the fit.

```
new1.dat <- data.frame(sc = 1000, ra = 160, dec = 45) # Lower scale length example
predict(y.lm, newdata = new1.dat, interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 30.86222 28.39218 33.33225
```

```
new2.dat <- data.frame(sc = 4000, ra = 160, dec = 45) # Higher scale length example
predict(y.lm, newdata = new2.dat, interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 41.21295 38.06083 44.36506
```

As is shown above, the confidence intervals lie outside of each others bounds which signifies that there is a real trend in the data. For a scale length of 1000 pc, this fit estimates approximately 30.86 +/- 2.47 galaxies within the fixed volume. For a scale length of 4000 pc (a less compact galaxy than one with a scale length of 1000 pc), the fit predicts 41.21 +/- 3.15 galaxies within the same volume.

## 6 Discussion

The multiple linear regressions and the confidence interval determinations show that there is a trend between the scale length of a spiral galaxy and the environment it is located in. It shows that more compact galaxies tend to be located in lower density environments than their more sprawling counterparts. This is opposite to the expectation I had going into this project. It shows that some other element of physics must be at play, beyond tidal stripping, in order to account for the difference in compactness.

## 7 Conclusions

Although there is not a strong relationship between the density of a spiral galaxy's environment and it's scale length, there does appear to be a statistically significant trend based on the 58,924 galaxies I used in this analysis. The physics behind this trend still needs to be determined in order to explain why galaxies in more highly populated regions of space tend to be slightly less compact than galaxies in more remote locations. This would be an interesting field for further study within the scientific community.

## 8 Acknowledgements

I would like to acknowledge Dr. French and Menghong for their help through this course. They inspired me to keep trying new methods throughout the semester especially when my project got discouraging. I would also like to acknowledge Dr. Mark Allen for his help obtaining the data for this project. Lastly, I'd like to acknowledge Dante Velasquez for his brillant ideas on how to approach aspects of my project.

## 9 References

Vizier - http://vizier.u-strasbg.fr/viz-bin/VizieR

Sloan Digital Sky Survey - http://skyserver.sdss.org/dr14/en/tools/search/sqs.aspx