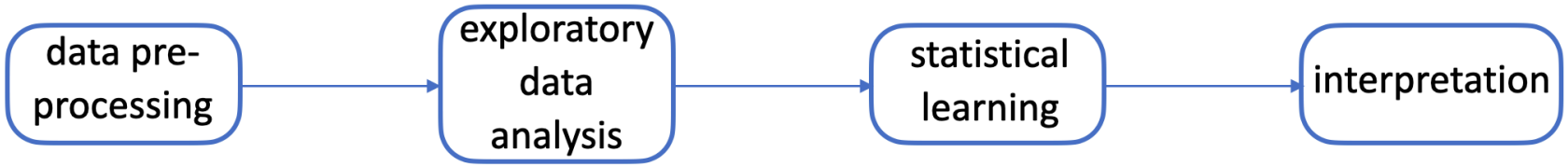
Exploratory Data Analysis

36-600

The Canonical Analysis Workflow



- *Data Pre-Processing*: the act of extracting analyzable data (e.g., a structured data table) from unstructured sources (e.g., images, audio files, text, etc.), as well as the act of editing the data table to mitigate missing data
- *Exploratory Data Analysis*: the act of visualizing the observed data--via, e.g., histograms, scatter plots, box plots, etc., etc.--so as to build intuition about them
- *Statistical Learning*: the attempt to find meaningful structures in the data or to uncover relationships between elements of the dataset
- *Interpretation*: what did you discover through your analysis?

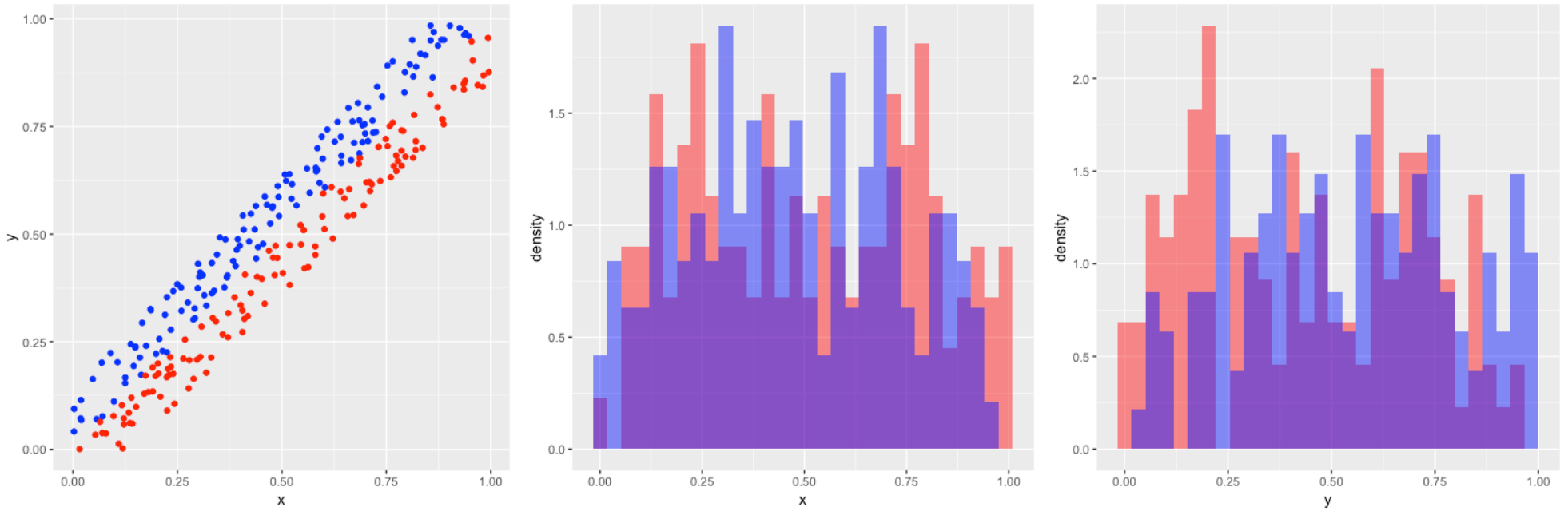
What is Exploratory Data Analysis?

- As stated above: the act of visualizing the observed data--via, e.g., histograms, scatter plots, box plots, etc., etc.--so as to build intuition about them
  - what is the sample size and what is the number of variables?
  - are there uninformative variables (e.g., a column of row numbers)?
  - are the informative variables quantitative or categorical?
  - how are the variable values empirically distributed (e.g., unimodally or bimodally? symmetric or skewed?)
  - are there missing data?
  - are there anomalous data?
  - are there apparent linear relationships between predictor variables?
  - are there visually apparent associations between the predictor variables and the response variable?
  - etc.

Exploratory Data Analysis is Not a Replacement for Statistical Learning!

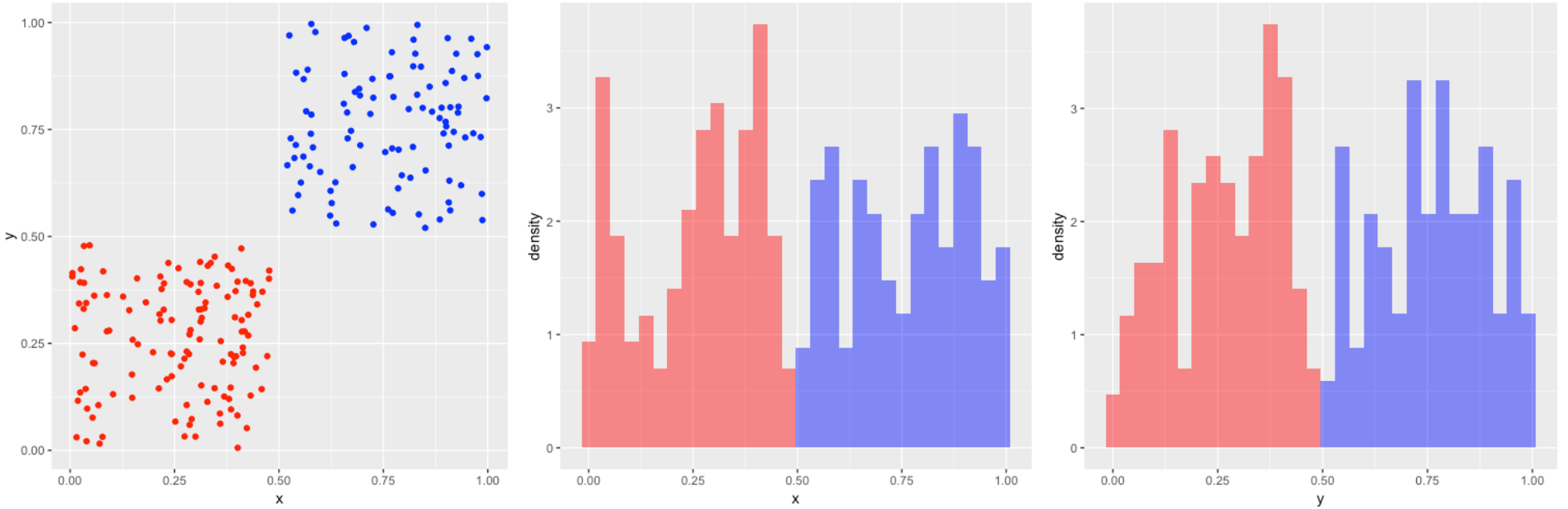
- If the number of predictor variables is larger than three, one cannot directly visualize the native space of the data; one can only visualize *projections* of that space
    - if those projections yield useful information, great!
    - if they do not: one should not give up, *because information may have been lost in projection*

Exploratory Data Analysis is Not a Replacement for Statistical Learning!



- If we can only "see" in one dimension, then we would conclude when we use EDA that there is no association between the (x, y) coordinates and the object class...however, if we could visualize the native space, we'd conclude there is definitely an association between coordinates and class
    - EDA: visualization of *projected* data
    - statistical learning: modeling of data in their *native space*

The Limits of Exploratory Data Analysis



- Here, we would conclude when we use EDA that there is an association between the (x, y) coordinates and the object class
- If by eye we see clear associations between, e.g., at least some predictor variables and the response variable, then we know that we when apply statistical learning techniques we will uncover meaningful information about associations between variables

EDA: A Starting Point

- We start by calling `summary()`, as it will identify amounts of missing data (assuming they have been identified and marked as `NA`) and give a sense of how the data are distributed

```
dim(df)           # 10 measurements for each of 3,456 galaxies
```

```
## [1] 3456   10
```

```
summary(df[,1:4]) # showing a subset only so as to fit on the slide
```

```
##     field          Gini              M20                C
##  COSMOS:905   Min.   :0.03829   Min.   :-2.2154   Min.   :1.093
##  EGS   :750   1st Qu.:0.41397   1st Qu.:-1.7181   1st Qu.:2.665
##  GOODSN:464   Median :0.45774   Median :-1.5802   Median :3.008
##  GOODSS:588   Mean   :0.44359   Mean   :-1.5135   Mean   :3.023
##  UDS   :749   3rd Qu.:0.48947   3rd Qu.:-1.3431   3rd Qu.:3.402
##               Max.   :0.85754   Max.   :-0.4342   Max.   :4.922
```

- `field` is a categorial variable...all the others are quantitative
- There are no missing data

EDA: But What Happens of Data Are Missing?

Statistical learning models will not run if data are missing. They are not expected to know where to place missing data along the plotted lines!

- If data are missing...
  - you will have to remove data from your sample, or
  - perform data imputation (which is beyond the scope of this class)

EDA: But What Happens of Data Are Missing?

- There are no simple heuristics in how to go about removing data
- For instance, if you have six columns of data and a sample size of 100, and 30 of the data in column 6 are missing, do you…
  - remove the 30 rows with missing data, or
  - remove the sixth column?

EDA with ggplot

- `ggplot2` is a package for creating graphics, utilizing principles discussed in the book *The Grammar of Graphics*
    - `ggplot2` is part of the `tidyverse`

```
suppressMessages(library(tidyverse))
```

- A *very* basic call to `ggplot()` has the following structure:

```
ggplot(data=<data frame>,mapping=aes(x=<x axis variable>,...)) +
  geom_<plot type>(<arguments>) + ...
```

EDA: Single Categorical Variable

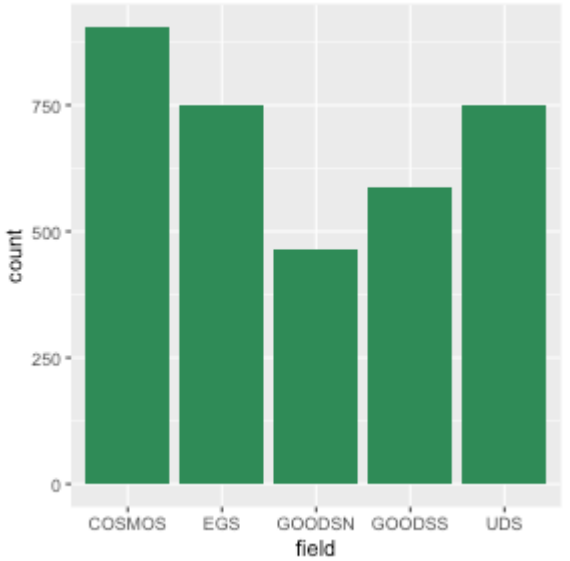
- To visualize a single categorical variable, we could use a `table()`...

```
df %>% select(.,field) %>% table(.)
```

```
## field
## COSMOS     EGS GOODSN GOODSS     UDS
##    905     750     464     588     749
```

- ...or a bar chart...

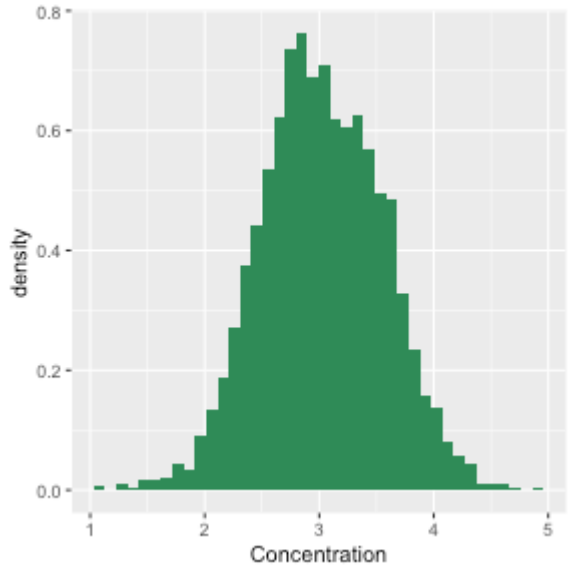
```
ggplot(data=df,mapping=aes(x=field)) +
  geom_bar(fill="seagreen")
```



EDA: Single Quantitative Variable

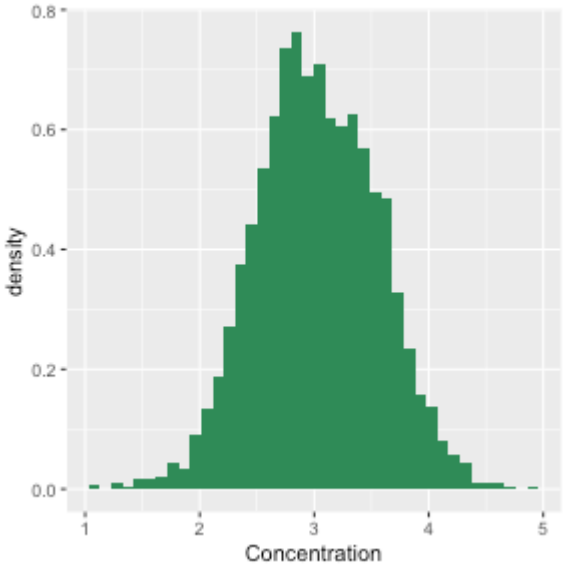
- To visualize a single quantitative variable, we would most often use a histogram

```
ggplot(data=df,mapping=aes(x=C,y=after_stat(density))) +
  geom_histogram(bins=40,fill="seagreen") +
  xlab("Concentration")
```



EDA: Single Quantitative Variable

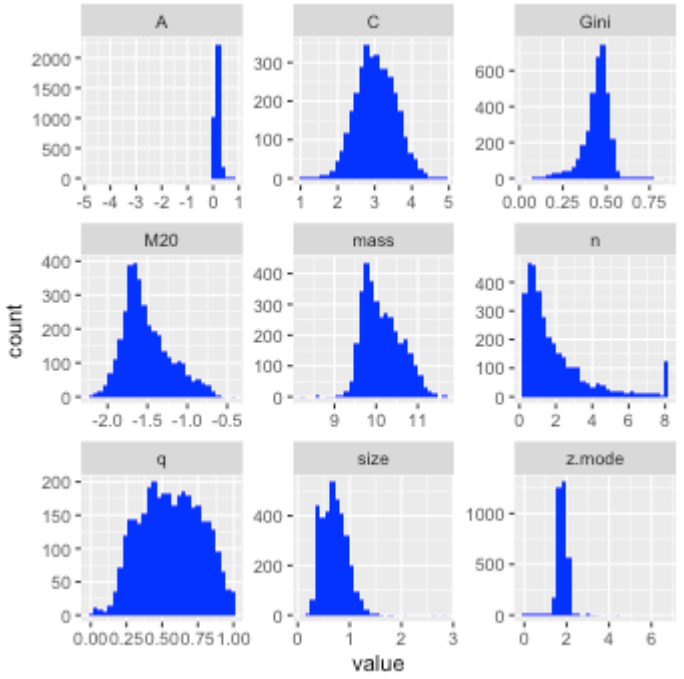
- Features of histograms to keep in mind:
  - they exhibit a nonparametric estimate of the shapes of underlying data distributions
  - they are sensitive to bin width: too few bins, and noise and localized features are smoothed out; too many bins, and noise obscures the underlying distribution



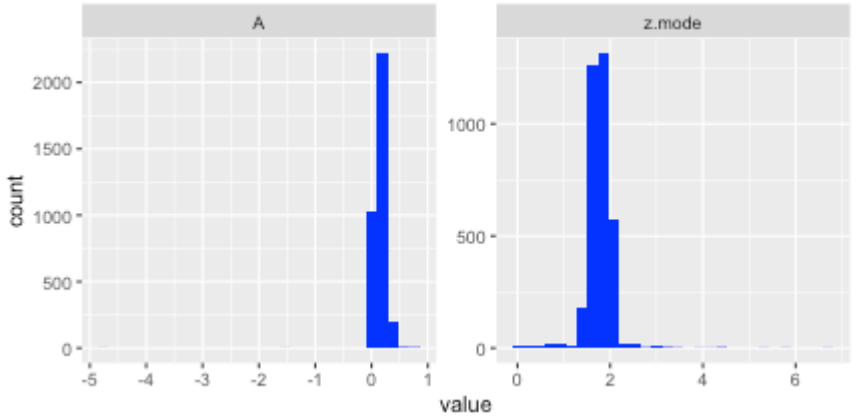
EDA: Single Quantitative Variable

- What do we observe here?
  - `C` and `q` have unimodal and symmetric distributions
  - `M20`, `n`, `mass`, and `size` are somewhat positively skewed
  - `Gini` is somewhat negatively skewed
  - we cannot make firm conclusions about `A` and `z.mode` other than there are apparent outlier values for these variables

```
df.new  <- df %>% select(.,-field) %>% gather(.)
ggplot(data=df.new,mapping=aes(x=value)) +
  geom_histogram(fill="blue",bins=30) +
  facet_wrap(~key,scales='free')
```



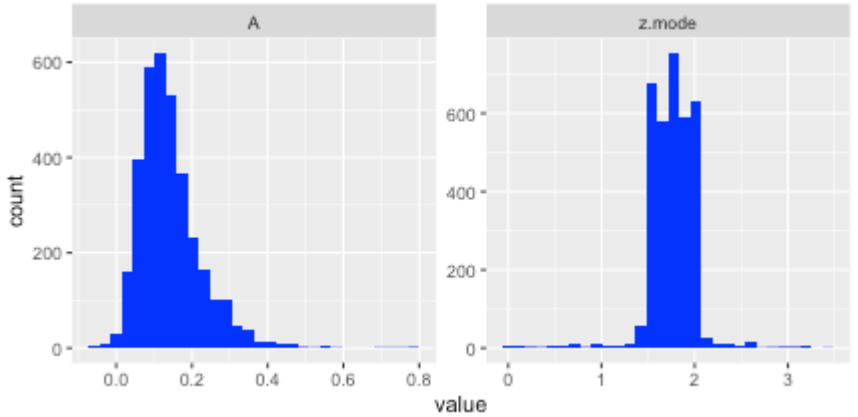
EDA: Anomaly Detection



- One goal of EDA is to perform *anomaly detection*, i.e., to find possible outlying values in the data
  - we will assume that we will detect anomalous values "by eye," using the qualitative criterion that data that lie "far from the core" of any histogram cannot be plausibly generated by the same mechanisms that generated the rest of the data
  - **be sure to record any removal of apparent anomalous values so that others can reproduce your analyses later!**
- Here, I will remove rows of data with values of `A` less than -0.5 and values of `z.mode` greater than 3.5. I can do that using the `dplyr` function `filter()`:

```
df.filtered <- df %>% filter(.,A>-0.5,z.mode<3.5)
```

EDA: Anomaly Removal



- Do not use the "1.5 IQR rule" for determining whether a datum is anomalously valued
- This rule is *far* too conservative to use when the sample size is (moderately) large
  - if we have $n = 100$ normally distributed data, the rule will on average identify 0.7 *valid* data as being anomalous
  - if we have $n = 10000$ data, 70 *valid* data, on average, will be flagged as anomalous
  - etc.

EDA: A Categorical Response versus a Categorial Predictor

- To visually ascertain whether there might be an association between a quantitative response variable and a quantitative predictor variable, the simplest approach is to use a table

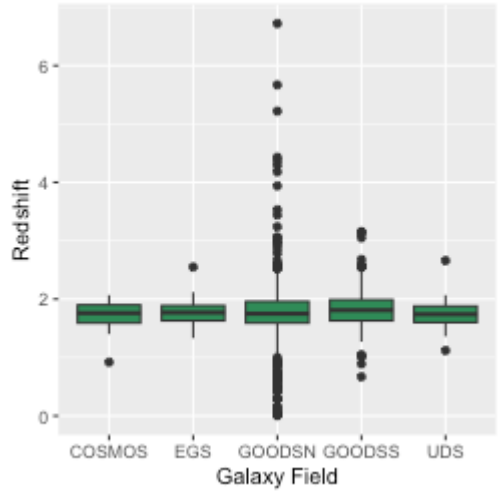
```
fake.var <- factor(sample(2,nrow(df),replace=TRUE))
table(fake.var,df$field)
```

```
##
## fake.var COSMOS EGS GOODSN GOODSS UDS
##        1    451 398    227    283 345
##        2    454 352    237    305 404
```

EDA: A Quantitative Response versus a Categorical Predictor

- To visually ascertain whether there might be an association between a quantitative response variable and a categorical predictor variable, we would most often use grouped boxplots

```
ggplot(data=df,mapping=aes(x=field,y=z.mode)) +
  geom_boxplot(fill="seagreen") +
  xlab("Galaxy Field") + ylab("Redshift")
```

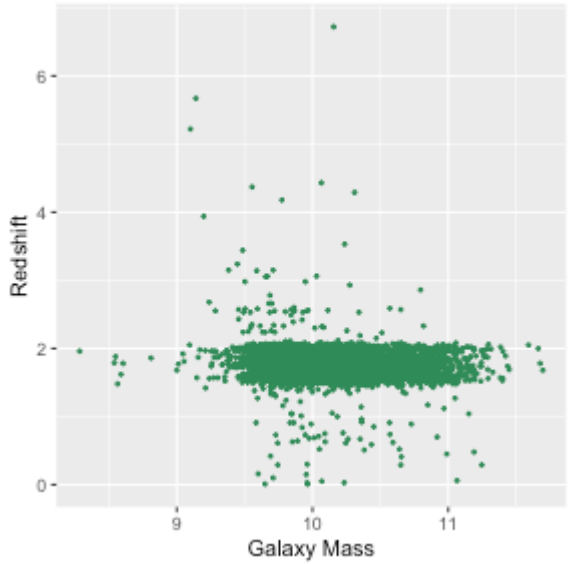


- If we draw a line through the boxes, it is flat: there is no apparent association between galaxy distance and sky location
- If we want to visualize a categorical response versus a quantitative predictor, we would simply "turn the plot 90 degrees" (reverse `x` and `y` in the `ggplot()` function call)

EDA: A Quantitative Response versus a Quantitative Predictor

- To visually ascertain whether there might be an association between a quantitative response variable and a quantitative predictor variable, we would most often use scatter plots

```
ggplot(data=df,mapping=aes(x=mass,y=z.mode)) +
  geom_point(col="seagreen",cex=0.5) +
  xlab("Galaxy Mass") + ylab("Redshift")
```



EDA: Scatter Plots

- **IMPORTANT!** Here are some tips of the trade to keep in mind when constructing scatter plots:
  - if your sample size is $\gtrsim 10^4$, randomly sample ~ 1000 points for plotting
  - to improve interpretability, mitigate the effect of point overlap by both reducing the size of points and altering point transparency
  - change the plot limits manually if necessary to zoom in on the bulk of the points

EDA: A Quantitative Response versus a Quantitative Predictor

```
set.seed(101)
s <- sample(nrow(df),1000)
ggplot(data=df[s,],mapping=aes(x=mass,y=z.mode)) +
  geom_point(col="seagreen",cex=0.5,alpha=0.5) +
  xlab("Galaxy Mass") + ylab("Redshift") +
  xlim(9,11.5) + ylim(1,2.5)
```

```
## Warning: Removed 28 rows containing missing values (`geom_point()`).
```