NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Week 15: Lecture 1
Classification

How can we predict qualitative variables instead of quantitative variables?

DS-UA 111
Data Science for Everyone

Week 15: Lecture 1
Classification

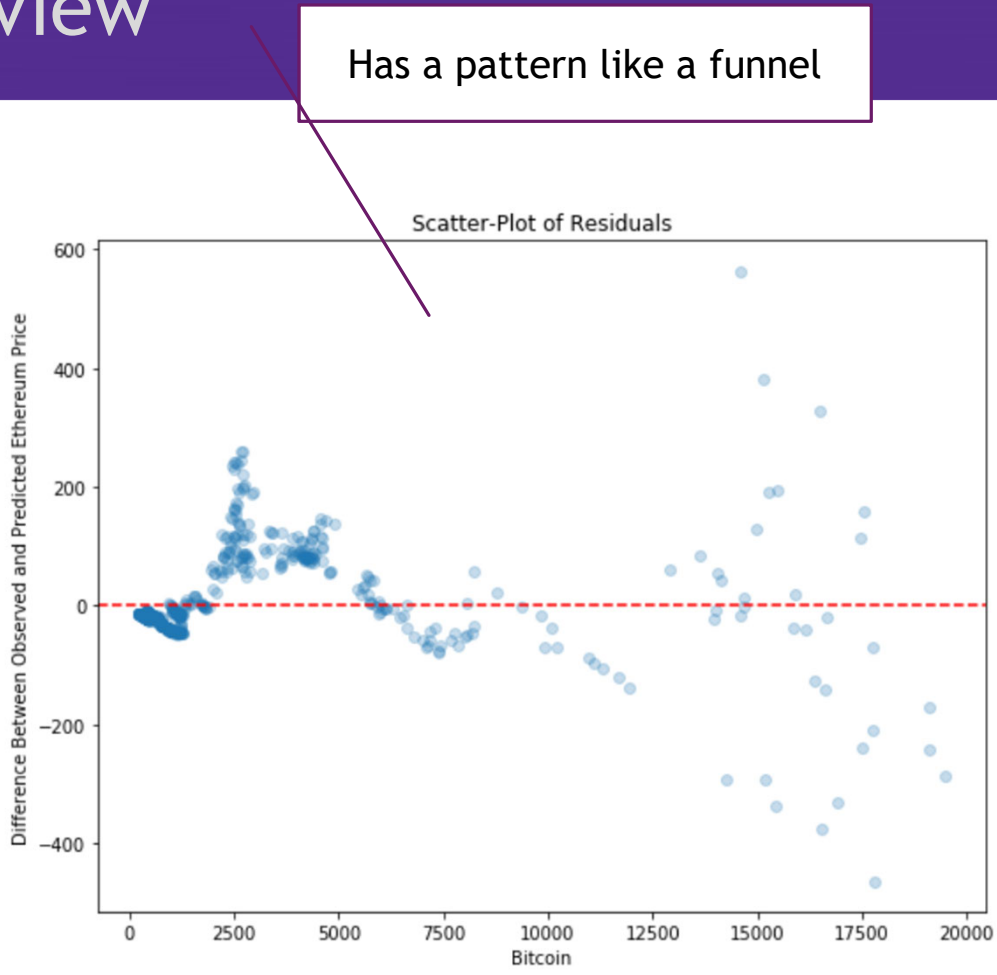
Adapted from Adhikari, DeNero, Wagner, Milner

Announcements

- Please check Week 15 agenda on NYU Classes
  - Project
  - Lab 10
- Refer to the Calendar linked to NYU Classes

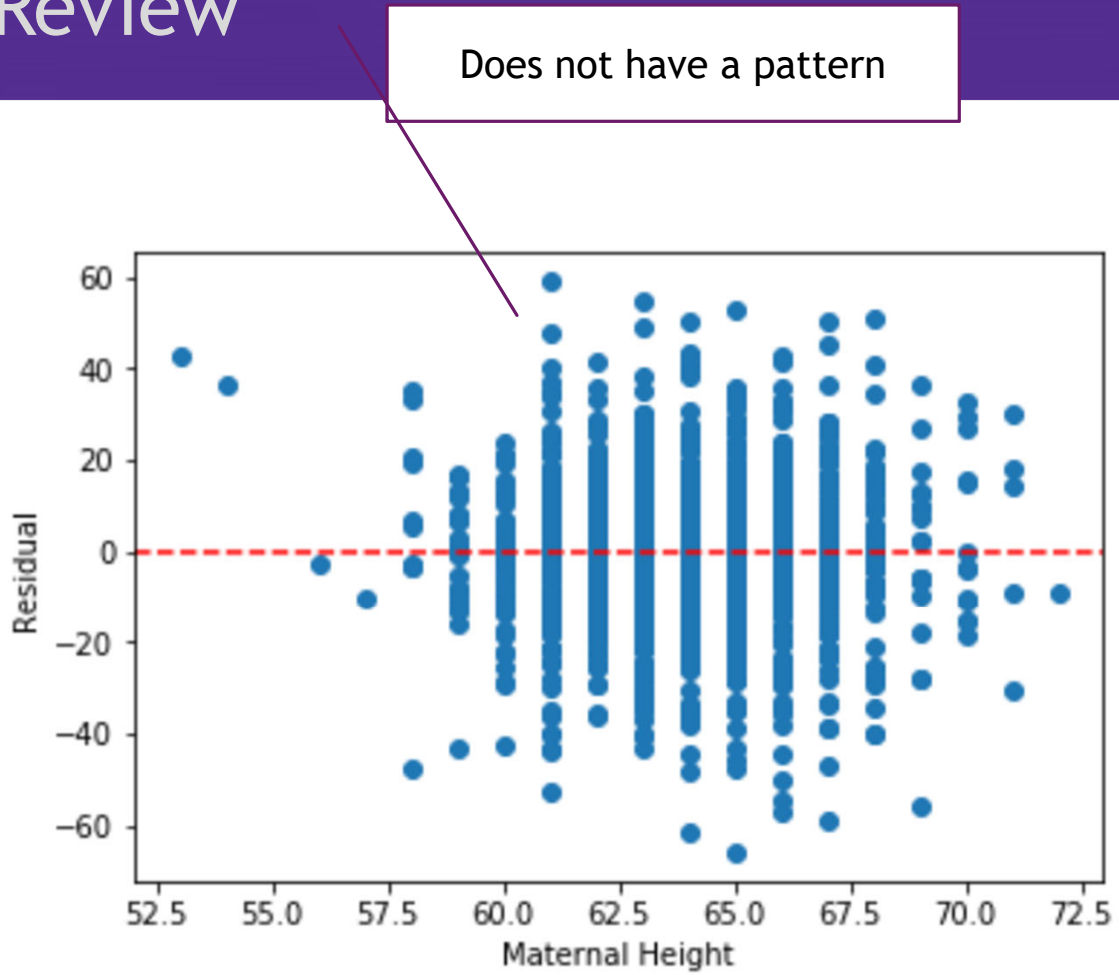
Review

- We can generate a scatter-plot to visualize the residuals. We want
  - About half the points above 0 and about half the points below 0
  - Comparable differences from 0 throughout the points
  - No discernible trend or pattern
- Otherwise we should explore other explanatory variables



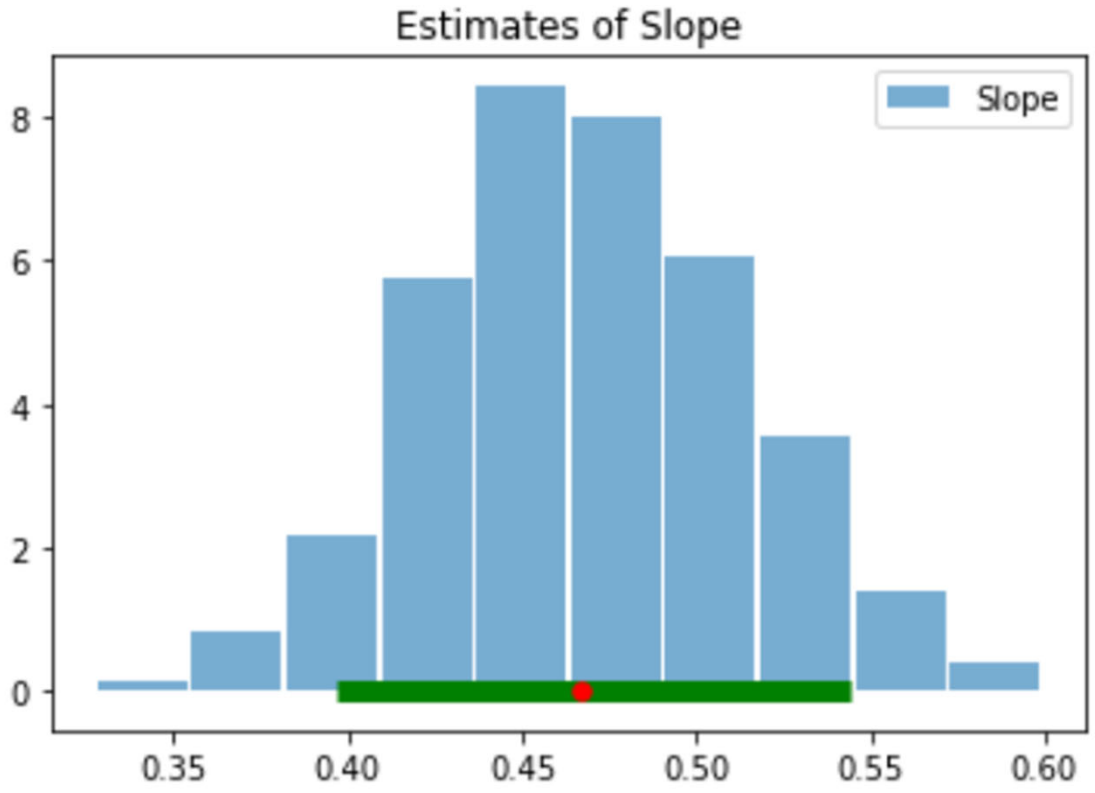
Review

- We can generate a scatter-plot to visualize the residuals. We want
  - About half the points above 0 and about half the points below 0
  - Comparable differences from 0 throughout the points
  - No discernible trend or pattern
- Otherwise we should explore other explanatory variables



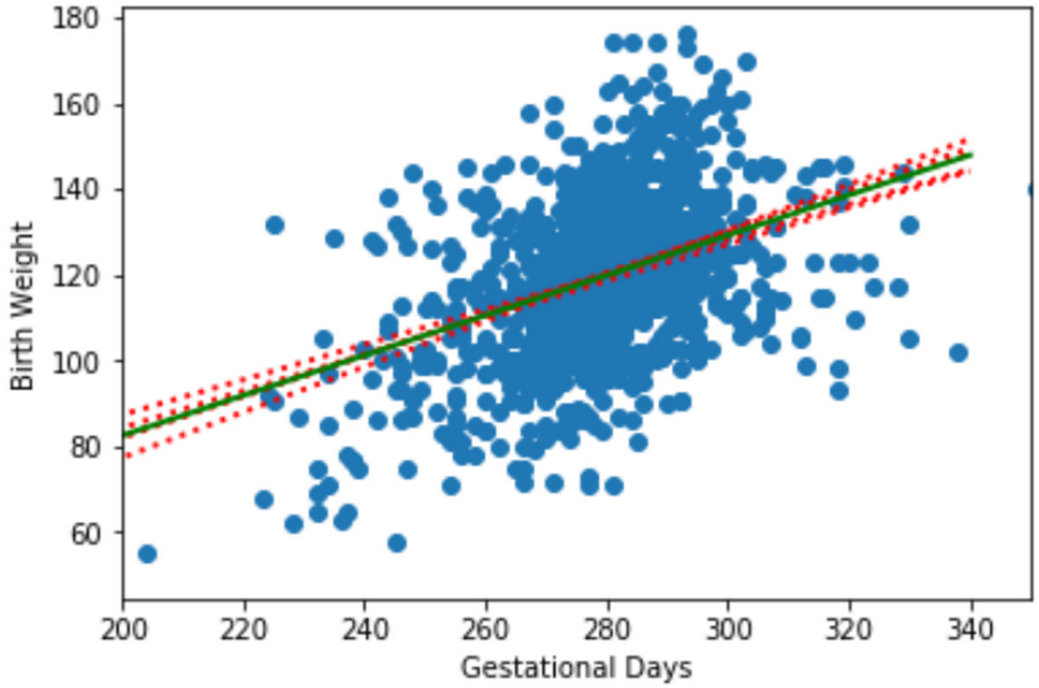
Review

- We determine the slope and intercept through fitting the line to the data. The data is a sample from the population.
- We can quantify the variation across samples in the slope and interval through resampling.
- Bootstrap resampling allows us to generate many slopes and intercepts across replications



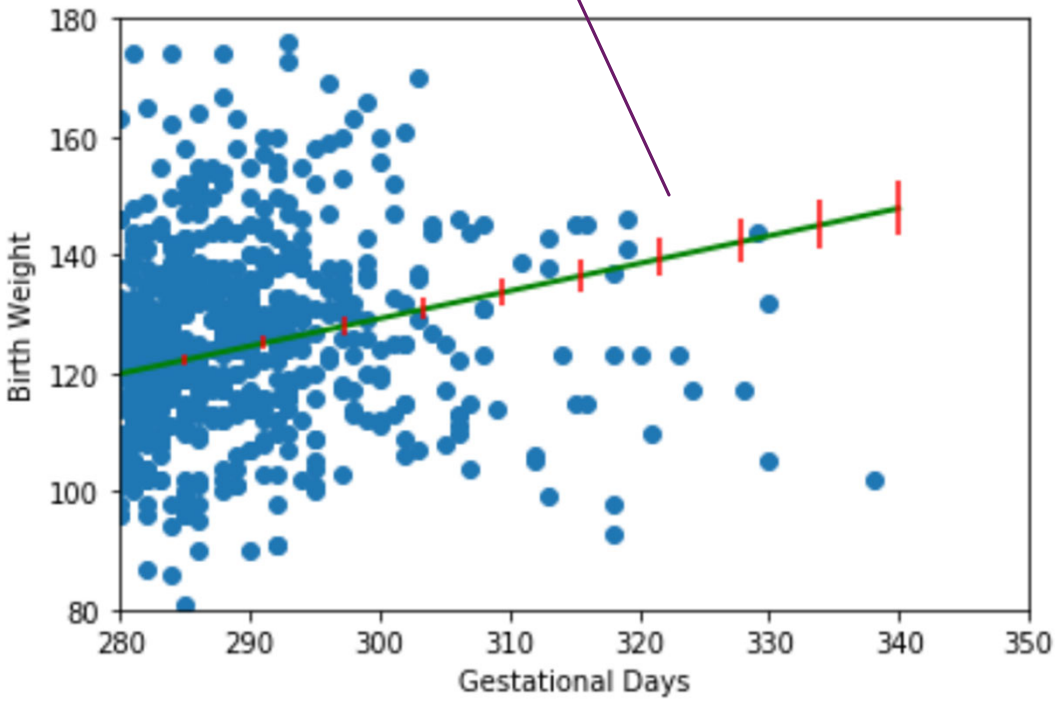
Review

- For each replication, we have a resample. We can fit a line to the data in the resample to determine the slope and intercept.
- We can calculate confidence intervals from these numbers by determining percentiles like 5^{th} and 95^{th}
- Here we have bootstrap confidence intervals for slope and intercept



Review

- If we fix a value for the explanatory variable, then for each replication we have a slope and intercept to make a prediction.
- We can calculate confidence intervals from these numbers by determining percentiles like 5^{th} and 95^{th}
- Here we have bootstrap confidence intervals predictions



Exercise

- Suppose we determine a 90% confidence interval of predicted wait for duration of 2 minutes. Label the following statements as True or False:

1. The confidence interval covers 90% of waits in the sample that had duration of 2 minutes.
2. This confidence interval tells us the differences between observed waits and predicted waits.
3. The confidence interval quantifies our uncertainty about different possible predictions for waits.

Agenda

- Nearest Neighbors

References

- Classification
  - Chapter 17.1-17.2

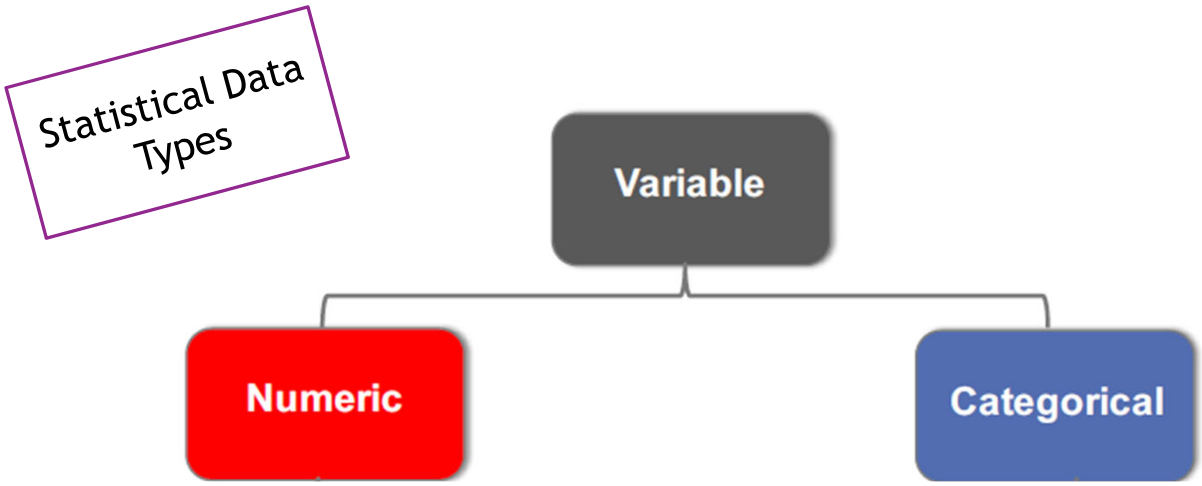
Data Types

Computational Data Types

- We store data as variables with different types.
- The types allow for different operations on the variable like adding or appending

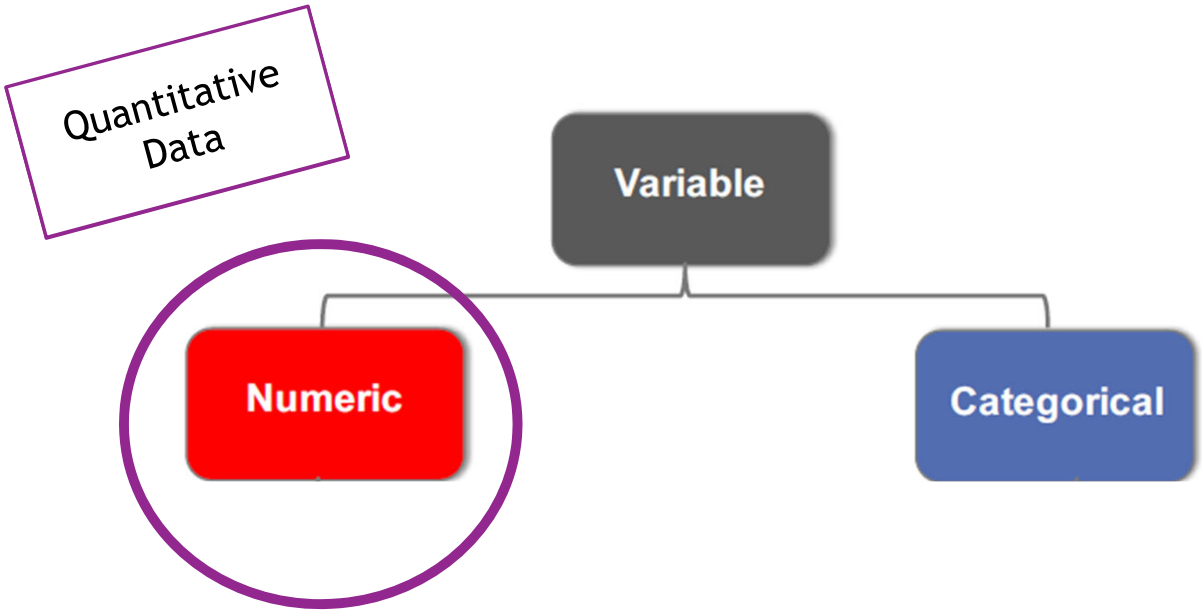
Example	Data Type
`x = "Hello World"`	str
`x = 20`	int
`x = 20.5`	float
`x = {"name" : "John", "age" : 36}`	dict
`x = {"apple", "banana", "cherry"}`	set
`x = ["apple", "banana", "cherry"]`	list
`x = ("apple", "banana", "cherry")`	tuple
`x = True`	bool

Data Types



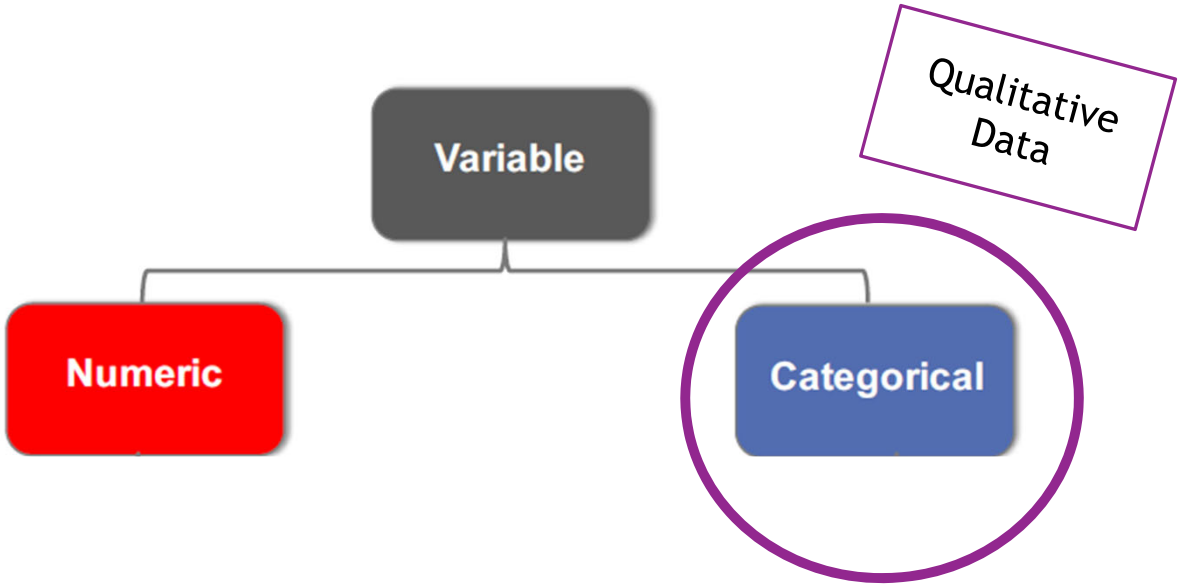
- We study data with different properties. We divide these properties across two data types
  - Numbers
  - Categories

Data Types



- Sometimes we can count values like the numbers 1,2,3,…
- Other times we have too many values. We cannot count numbers like 1.54, 2.43, 3.14,…

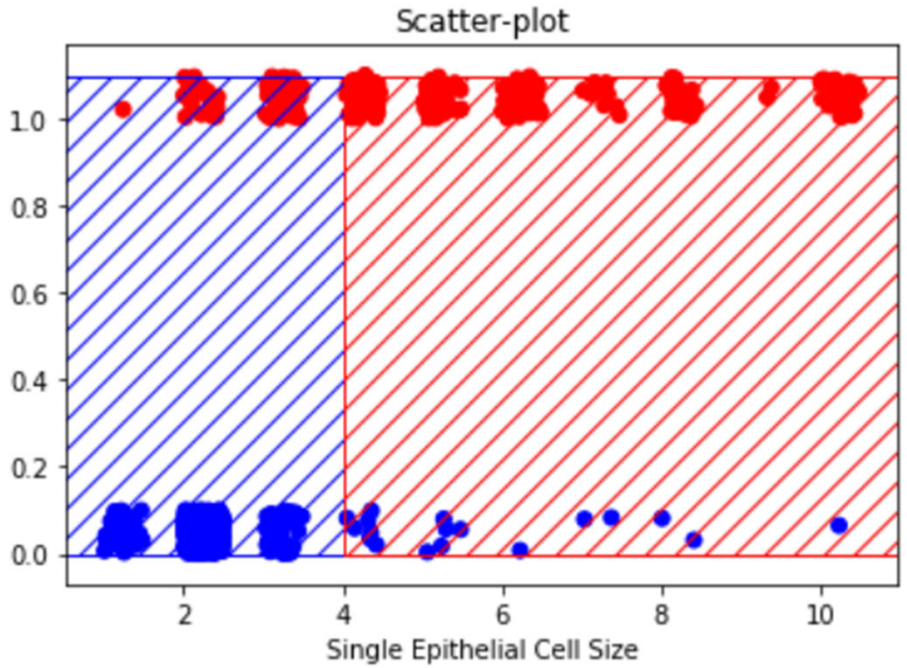
Data Types



- We could think of categories as labels like blue, green and red.
- While we cannot perform calculations with categories, we can sometimes put the categories in order like high, medium, low

Classification

- With regression we predict a quantitative response variable from explanatory variables
- With classification we predict a qualitative response variable from explanatory variables



Summary

- Nearest Neighbors

Goals

- Use hypothesis testing with confidence intervals to study estimated slope and intercept.
- Understand the nearest neighbors approach to classification