

STOR 320.1
Course Overview and
Review of Basic Concepts

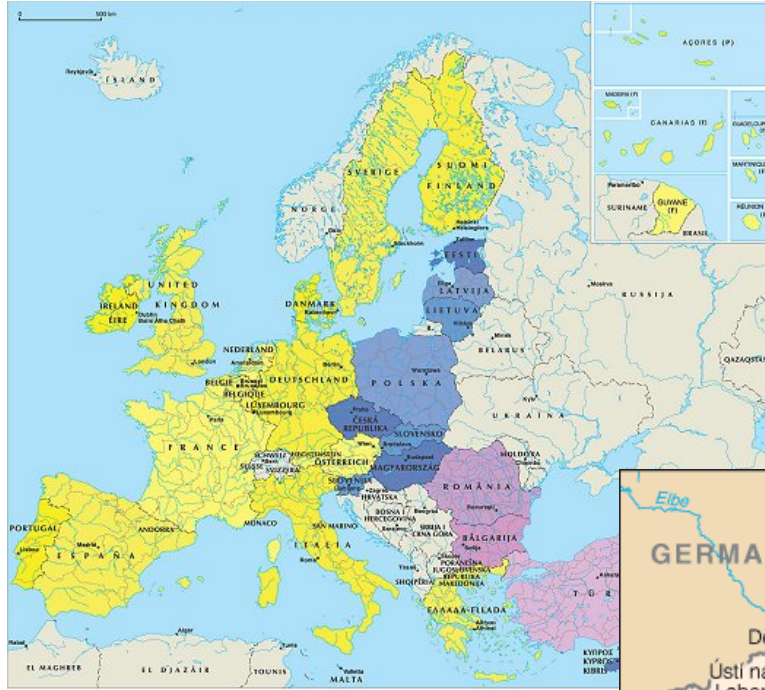
Jan Hannig

- 330 Hanes Building
- jan.hannig@unc.edu,
- (919) 962-7511
- <https://hannig.cloudapps.unc.edu/>
- Twitter @janhannig
- Office hours: MW 2-3pm



Where am I from?

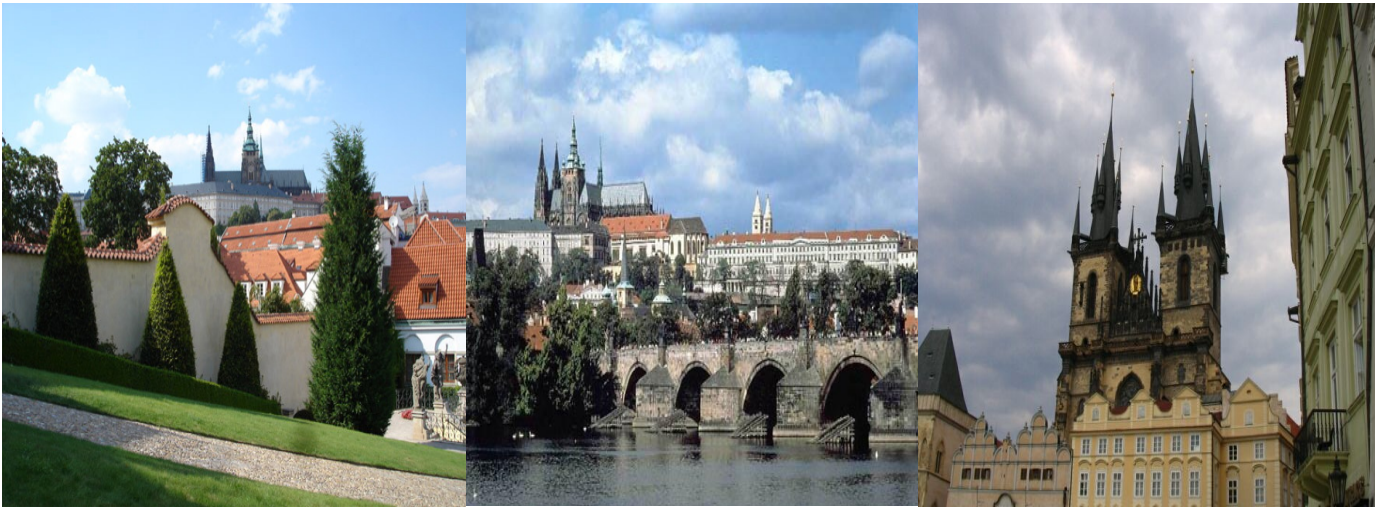
- Can you venture a guess?
- A) Netherlands
- B) Ukraine
- C) Czechia
- D) South Africa



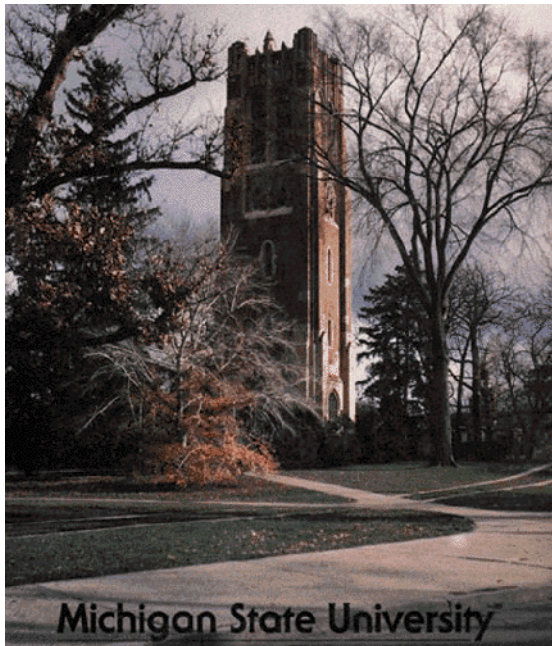
Czech Republic



Prague



Michigan State



Married to Shevaun Neupert



Klára and Declan



Interests

- Mountain biking
- Cello
- My church (Lifepointe in Cary)
- Of course

–Research

–Teaching

Tell me about you

- Go to Sakai –forums and leave me some fun fact about you!

Lectures and Labs

- Lectures TuTh 2:00 PM - 3:15 PM
- Labs
 - 320.400 by Taylor: W 5:45 PM to 6:35 PM
 - 320.401 by Pavlos: W 4:40 PM to 5:30 PM
 - 320.402 by Sam: F 4:40 PM to 5:30 PM
- Email Christine (crikeat@email.unc.edu)

Instructional Assistant

- Taylor Patty (400)
 - Email: tmpetty@live.unc.edu
 - Office Hours: M 2:30 – 3:30pm & W 1:30 – 2:30pm
- Pavlos Zoubouloglou (401)
 - Email: pavlos@live.unc.edu
 - Office Hours: M 10:00 – 11:00am. & F 1:00 – 2:00pm
- Sam Booth (402)
 - Email: slbooth@live.unc.edu
 - Office Hours: TuTh 4:00pm – 5:00pm

Outline

- Administrative details
- What's the course about?
 - Introduction to R

Ask Questions in Class

- By default, your microphone will be muted.
- If you have a question, feel free to unmute yourself and ask questions. (Preferred)
- Also, you can type your question in the in-meeting chat window.

Remote Instruction

- This will be a remote course:
 - a) lectures will be held live online during the scheduled time and recorded so that you can watch them later on Panopto
 - b) labs will be online and **not** recorded (attendance will be taken!)
 - c) office hours will be held online but not recorded
 - d) all assignments will be done remotely.

Questions

- Three ways to ask questions:
 - post questions on Sakai forum;
 - come to the virtual office hours on Zoom;
 - send an email to the instructor or the IAs.

Grading

Lab Attendance	10%
Lab Assignments	15%
Homework	45%
Final Project	30%

A	94 to 100	B	83 to 86.99	C	73 to 76.99	D	60 to 66.99
A-	90 to 93.99	B-	80 to 82.99	C-	70 to 72.99	F	0 to 59.99
B+	87 to 89.99	C+	77 to 79.99	D+	67 to 69.99		

Homework and Labs

- Around 7 homework assignments and 4 data analysis assignments. They will be posted on Gradescope and there will be about one week to complete the homework and about two weeks to complete data analysis assignments.
- Lab assignment:
 - Due 30 minutes after the lab ends.
 - No late submission will be accepted.
 - will be based on the topics discussed in lecture or related to your final project.

Project

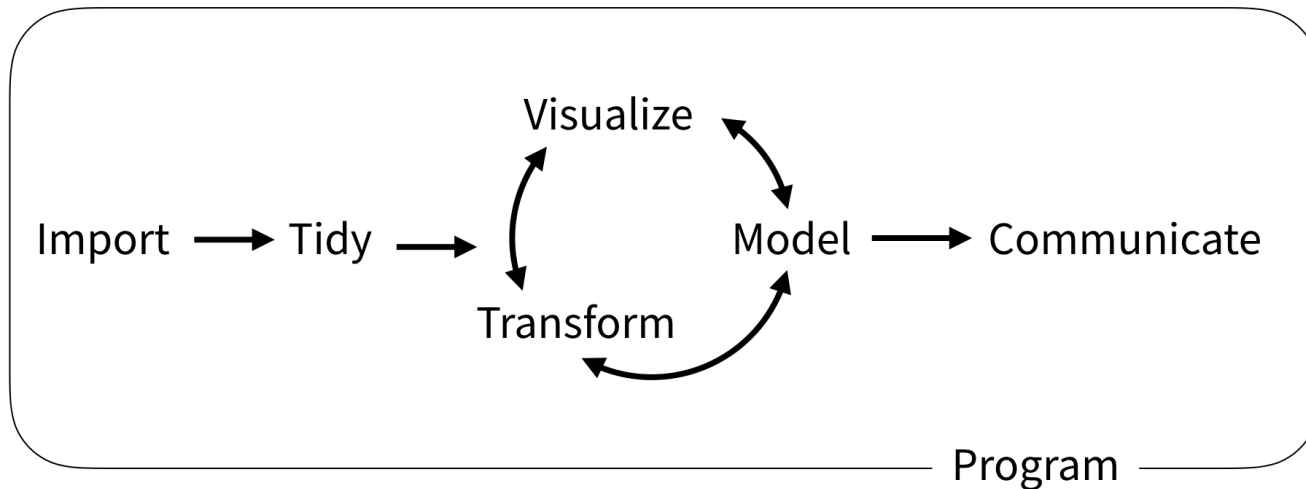
- For the final project, each lab section of STOR 320 will be divided into groups of 5. I will assign you to groups at random (using R program) after drop deadline.
- More information on the project is on the website.

Big Picture



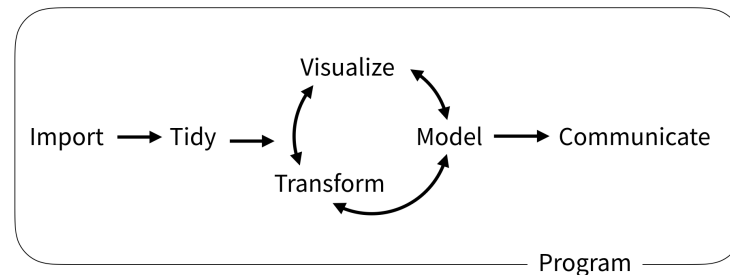
shutterstock.com • 1247255884

What is data science?



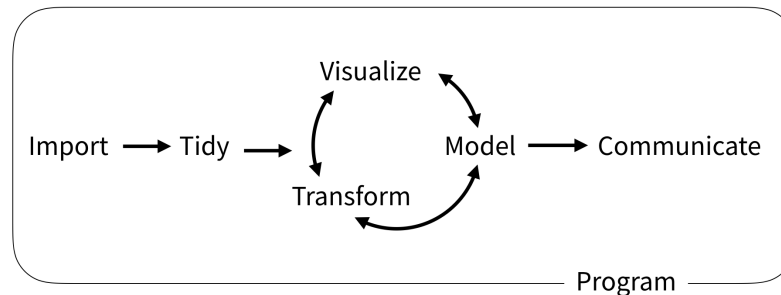
Wickham and Grolemund
(2017)

The model of data science



- First we must *import* our data.
- *Tidy* data → consistent structure
- Transformation:
 - narrowing in on observations of interest
 - creating new variables
 - calculating a set of summary statistics

The model of data science



- *Visualization*: show you things that you did not expect or raise new questions about the data.
- Use a *model* to answer your questions
- *Communication*: an absolutely critical part of any data analysis project.
- Surrounding all these tools is programming.

R and RStudio



The screenshot shows the RStudio interface with a file named "5-parameters.Rmd" open. The editor contains the following R Markdown code:

```
1 ---
2 title: "Visualizing the ocean floor"
3 output: html_document
4 params:
5   data: "hawaii"
6 ---
7
8 ```{r include = FALSE}
9 library(marmap)
10 library(ggplot2)
11 ```
12
13
14 The [marmap](https://cran.r-project.org/web/packages/marmap/index.html) package provides tools and data for visualizing the ocean floor. Here is an example contour plot of marmap's ``r
15 params$data`` dataset.
16
17 ```{r echo = FALSE}
18 data(list = params$data)
19 autoplot(get(params$data))
20 ```
21:1 (Top Level) >
```

The console at the bottom shows the command: `> render("5-parameters.Rmd", params = list(data = "aleutians"))`

The right-hand pane displays the rendered HTML output. It features a title "Visualizing the ocean floor" and a paragraph: "The `marmap` package provides tools and data for visualizing the ocean floor. Here is an example contour plot of marmap's `aleutians` dataset."

The contour plot shows the ocean floor topography around the Aleutian Islands. The x-axis is labeled 'X' and ranges from 170 to 210. The y-axis is labeled 'Y' and ranges from 50 to 65. The plot displays depth contours, with the deepest parts of the ocean floor (trenches) shown as dark lines, and shallower areas shown as lighter lines. The Aleutian Islands are visible as a series of peaks and valleys along the coast.

Why R?

- Easy to learn and easy to use.
- Very popular and one of the standard languages for statistics, data science, computational biology, finance, industry, etc.
- Free and open-source.
- A lot of high-quality packages.
- New technology and ideas often appear first in R.
- Supported by a vast community that maintains and updates R.
- Runs on basically any platform.

Learning Programming

- Transfer the concepts to other languages
- How you approach a computational task and reason about the computations is similar
- Learning another programming language will be much easier in the future

Statistical Learning

- Linear regression.
- Classification (logistic regression, LDA, K-nearest neighbors).
- Cross-validation and bootstrap.
- Principal component analysis.
- Clustering methods (K-means clustering and hierarchical clustering).
- Recommender systems.
- Neural networks.

Textbooks

- *Required: R for Data Science.* Hadley Wickham. Legally free online, but can be purchased for less than \$40 on Amazon.
- *Good supplement: The elements of statistical learning: data mining, inference, and prediction.* Hastie, Trevor, Robert Tibshirani, and Jerome Friedman.

Data Visualization



What are we learning?

- To Install Some Key R Packages
 - Tidyverse
 - Rmarkdown
- To Practice Coding via R Scripts
- To Learn Elements of ggplot2
- Practice Making Visually Stunning Pictures

Initial Steps in RStudio

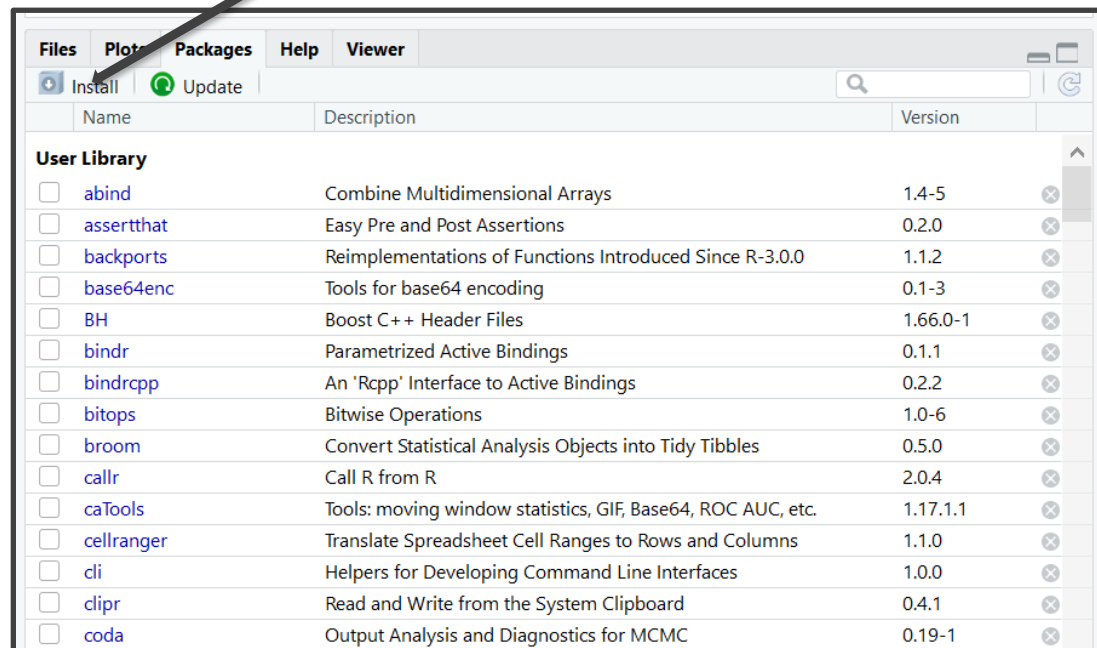
- Install Tidyverse Package

```
> install.packages("tidyverse")
```

- Other Packages To Be Installed

- RColorBrewer
- Rmarkdown

Select Install and Search on CRAN

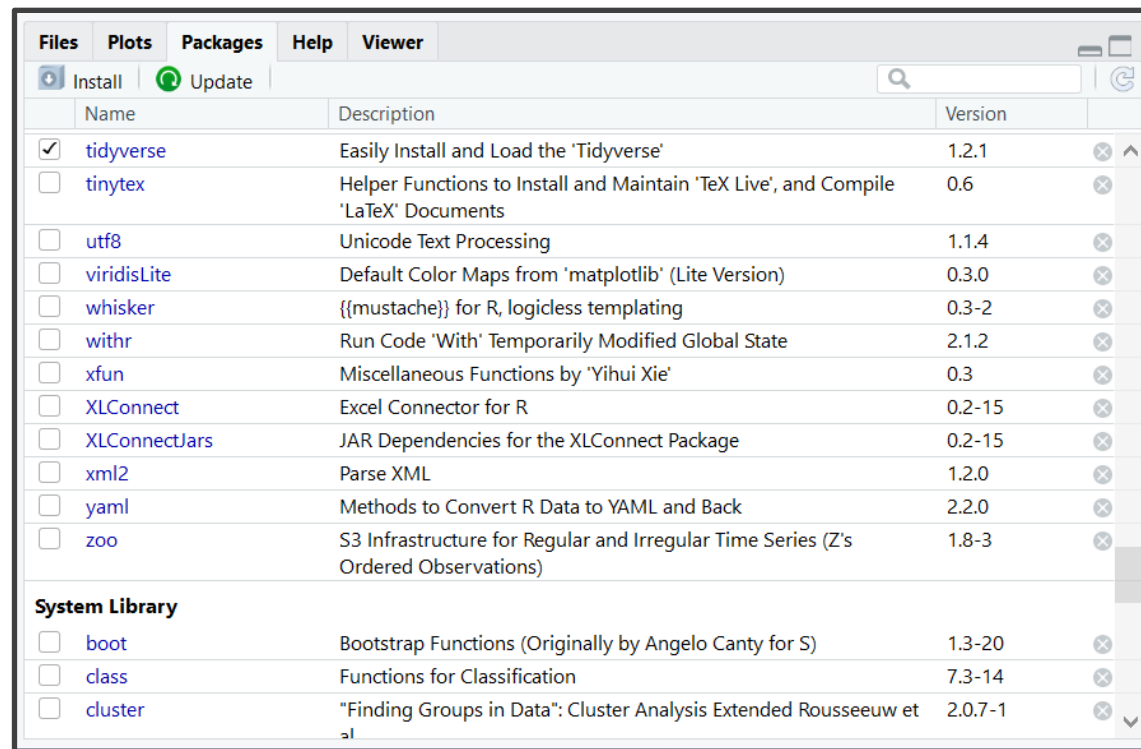


Initial Steps in RStudio

- To Use the Package
 - Code

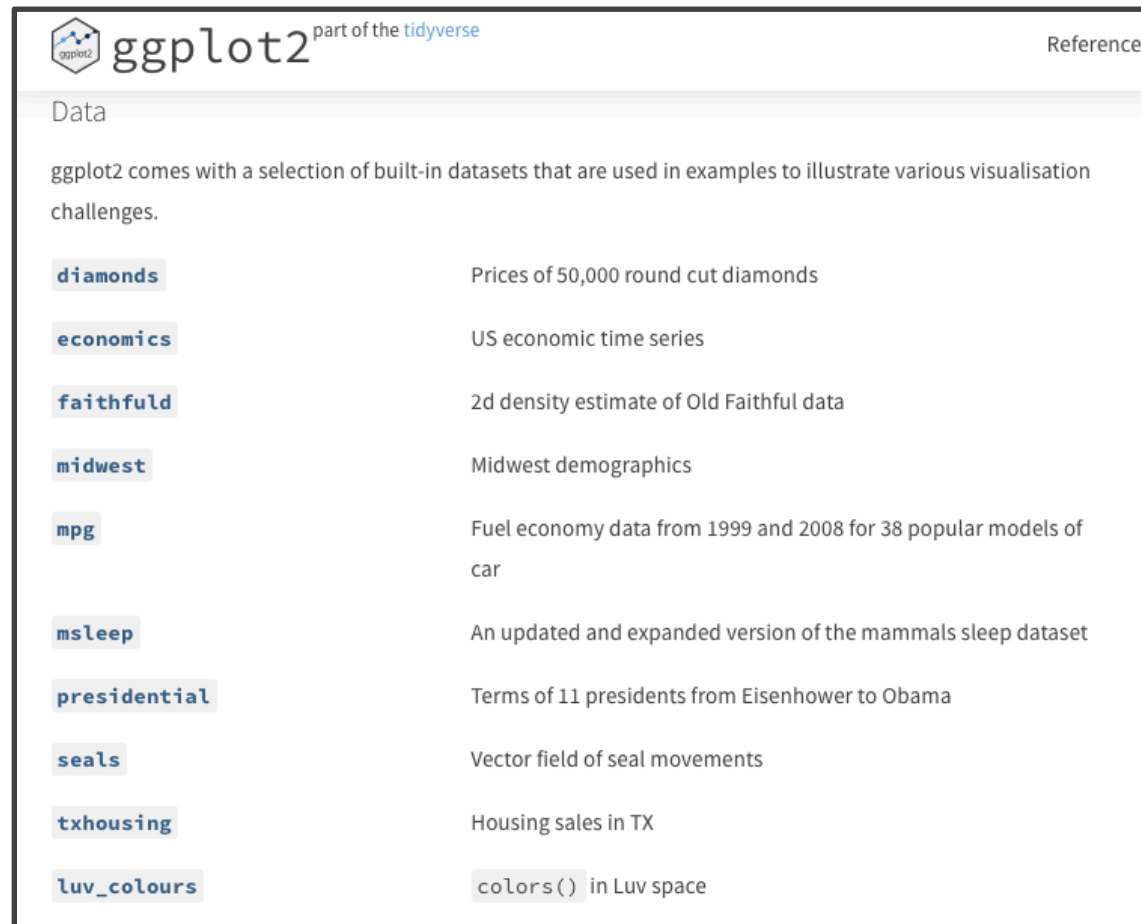
```
> library("tidyverse")
```

- Check Box for Tidyverse



ggplot2

- Help Page: [Link](#)
- Comes with Preloaded Datasets



The screenshot shows the ggplot2 help page, which is part of the tidyverse. The page title is "Data" and it includes a "Reference" link. The main text states: "ggplot2 comes with a selection of built-in datasets that are used in examples to illustrate various visualisation challenges." Below this, there is a list of datasets with their descriptions:

Dataset Name	Description
<code>diamonds</code>	Prices of 50,000 round cut diamonds
<code>economics</code>	US economic time series
<code>faithfuld</code>	2d density estimate of Old Faithful data
<code>midwest</code>	Midwest demographics
<code>mpg</code>	Fuel economy data from 1999 and 2008 for 38 popular models of car
<code>msleep</code>	An updated and expanded version of the mammals sleep dataset
<code>presidential</code>	Terms of 11 presidents from Eisenhower to Obama
<code>seals</code>	Vector field of seal movements
<code>txhousing</code>	Housing sales in TX
<code>luv_colours</code>	<code>colors()</code> in Luv space

ggplot2

- Many Useful Plots and Charts Provided
 - See Cheat Sheet: [Link](#)
(Also on Course Website)
 - Called Geoms (Geometric Objects)
 - The Geom you choose Must Comply with the Type of Variables You are Analyzing
- Organized by Type of Data
 - Univariate
 - Bivariate
 - Mixtures of Categorical and Numeric

ggplot2

- ggplot2 General Form

Fill in Blank With
Name of Data in R

```
> ggplot(data=_____) +  
  geom_TYPE(mapping=aes(x=____,y=____, etc.))
```

Fill in Blanks
from Variables
in Data

Tutorial 1

- Locate Tutorial 1 on Course Website
- Download Rmd File
- Open Rmd File on Computer
- Knit the Rmd File to PDF format
- View Graphs with Me