# STA130 Fall 2019 - T0107

## Week 6 – Joining Dataframes

(Materials used in this presentation are provided by the U of T Statistical Sciences Department.

This presentation was prepared by Vivian Ngo.)

**Github.com/vivianngo97/STA130-Fall-2019**

**viv.ngo@mail.utoronto.ca**

# Agenda

- Term Test Reminder
- Practice problem and vocabulary review
- Group discussion
- Oral presentations
- Time to create poster groups – poster project details to be provided next week!

# Term Test Reminder

- Reminders:
- The term test is two weeks from today during your usual tutorial time (November 1st)
  - You MUST attend the correct section's midterm.
  - AM section:
    - HS106: Last names A – F
    - MS3153: Last names G – V
    - NL6: Last names W – Z
  - PM Section:
    - MS2170: Last names A – Kalinins
    - MS2172: Last names Kandiah – Nyakong
    - MS3154: Last names Pan – Z
  - Includes all material up to & including Oct 28th (mostly a review class)
  - Format: Multiple choice, fill in the blanks, written answers (make sure to write complete sentences)
- Example midterms have been posted to Quercus

# Practice problem and vocabulary review

- Vocabulary:
  - Join
  - Key
  - Data frame
- What is a data frame?
- How to join data frames? (e.g. need a common variable, often called a key)
- Any questions on other materials to date, including R code and content
  - Ticket out the door from last week

# Group discussion

- Question 1. Based on Practice Problem #2:
- State which data summary (or summaries) would be appropriate to address each of the following questions based on data in the heroes data frame **and why**.
- (a) Do superhero weights tend to vary by publisher?
- (b) Are good superheroes more likely to be agile than bad superheroes?

# Group discussion

- Question 2. Based on Practice Problem #3: Describe what each of the following sets of code is doing in 1-2 sentences.

(a)
```
data <- full_join(hero_info, hero_power, key="name")
result <- data %>%
select(name,Publisher) %>%
filter(Publisher=="Marvel Comics"|Publisher=="DC Comics") %>%
group_by(Publisher) %>%
summarise(n=n())
result
```

(b)
```
data <- left_join(hero_info, hero_power, key="name")
results <- data %>%
mutate(weightclass=ifelse(Weight>100,"heavyweight","lightweight")) %>%
filter(!is.na(Flight)) %>%
group_by(Flight) %>%
summarise(prop=sum(weightclass=="heavyweight")/n()) %>%
arrange(desc(prop))
results
```

# Group discussion

- Question 3. You are interested in exploring which characteristics of neighborhoods in Toronto are associated with higher real-estate values – however, you can only find single pieces of information related to each characteristic in separate databases! For example, you've found one source of data with information on housing sale prices and another with condo prices. You found other databases with information on the number of parks and schools in the community, and yet another with the number of restaurants and stores in the community. You also found two other databases with information on community crime rates and walkability scores.

  What type of information would you need to combine all of these databases into one detailed data set that you can use to explore factors associated with real-estate values? How might you define neighbourhoods?

# Oral presentations

- Provide a summary of your results for practice problem 1, based on the hypothesis that there are twice as many non-flying superheroes as superheroes who can fly and a significance level of 5%.
- Provide a summary of your results for practice problem 4, based on the hypothesis that there are an equal number of non-flying superheroes as superheroes who can fly and a significance level of 5%.
- Based on practice problem 4, how would your conclusion change based on your chosen significance level (i.e., 2.5% versus 5%)? What type of errors (type 1/2) could you have made?

# Create poster groups

- **Start creating groups for poster presentation.**
- Groups should include 3-4 students (no 2-person groups or groups larger than 4 are allowed).
- You will be given the details of the poster project on Monday. We will be doing an activity regarding the poster project (in your poster groups) next Friday.