STA130 Winter 2022, Tutorial 2

Ente Kang

University of Toronto

2022-01-24

What is tidy data?

- Each variable is in one column
- Each row only has one observation
- Each cell contains only one value

```
# A tibble: 6 x 4
  country      year  cases population
  <chr>       <int>  <int>      <int>
1 Afghanistan  1999    745   19987071
2 Afghanistan  2000   2666   20595360
3 Brazil       1999  37737  172006362
4 Brazil       2000  80488  174504898
5 China        1999 212258 1272915272
6 China        2000 213766 1280428583
```

```
  country      year type            count
  <chr>       <int> <chr>           <int>
1 Afghanistan  1999 cases             745
2 Afghanistan  1999 population   19987071
3 Afghanistan  2000 cases            2666
4 Afghanistan  2000 population   20595360
5 Brazil       1999 cases           37737
6 Brazil       1999 population  172006362
# … with 6 more rows
```

Which one is tidy?

Source

Our setup

- We have a tibble called **df** (Ah, yes, very creative :P)
- Our variables are **gender, age, height, weight**
- Further assume that it is **tidy**

Removing a column

After eating a KFC 30 piece family bucket alone, we don't want to talk about or see our **weight**

- Let's get rid of that column!

```
df %>% select(-c(weight))
```

- OR

```
subset(df, select = -c(weight))
```

Try to see if you can find other methods

Filtering based on a condition

- Are there people that are 200cm tall?

```
df %>% filter(height == 200)
```

- Is anyone 20 years old?

```
df %>% filter(age == 20)
```

- Individuals that are 24 years of age and 185cm

```
df %>% filter(age == 24 & height == 185)
```

Again, other solutions are possible

Subsetting into a column(s)

- Accessing just the gender and age column

```
df %>% select(gender, age)
```

Renaming variables

```
rename(df, new_col1 = old_col1, new_col2 = old_col2, ...)
```

Handling missing values

- Drop all missing

```
df %>% drop_na()
```
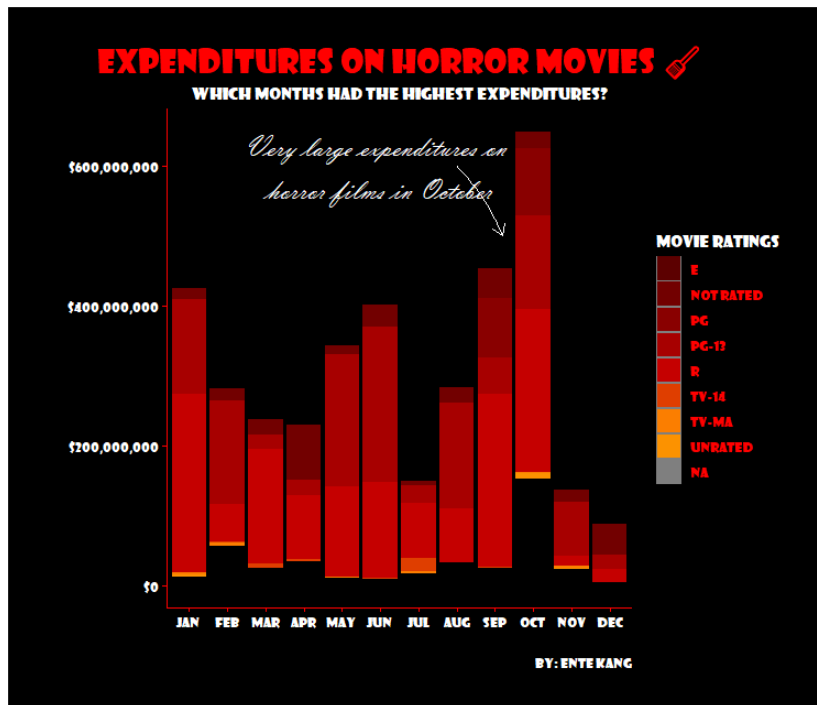
- Drop all missing for a variable

```
df %>% drop_na(your_variable)
```

- Impute the missing values

```
df[is.na(df), ] <- what you wish to fill in
```

Some cool things you can make with R

R IS GREAT!

Thank you! See you next week!