STAT 302: Data Wranglin’

Sheridan Grant

University of Washington Statistics Department

slgstats@uw.edu

April 15, 2020

Sample Data Frame

	Country	Region	Date	Type	Count
1	USA	WA	4/1	Conf.	500
2	China	Hubei	4/1	Death	100
3	China	Sichuan	4/8	Conf.	1000
4	USA	WA	4/8	Death	15
5	USA	WA	4/15	Conf.	3000
6	China	Sichuan	4/8	Death	20

Table: `covid`

- `dim(covid)` returns [6,5]
- `colnames(covid)` returns ['Country', 'Region', 'Date', 'Type', 'Count']

Filtering

	Country	Region	Date	Type	Count
1	USA	WA	4/1	Conf.	500
2	China	Hubei	4/1	Death	100
3	China	Sichuan	4/8	Conf.	1000
4	USA	WA	4/8	Death	15
5	USA	WA	4/15	Conf.	3000
6	China	Sichuan	4/8	Death	20

Filter with `covid <- covid[covid$Type == ‘Conf.’,]` or `covid <- covid %>% filter(Type == ‘Conf.’)`:

	Country	Region	Date	Type	Count
1	USA	WA	4/1	Conf.	500
3	China	Sichuan	4/8	Conf.	1000
5	USA	WA	4/15	Conf.	3000

Selecting

	Country	Region	Date	Type	Count
1	USA	WA	4/1	Conf.	500
3	China	Sichuan	4/8	Conf.	1000
5	USA	WA	4/15	Conf.	3000

Select with `covid <- covid[,c(2,3,5)]` or `covid <- covid %>% select`:

	Region	Date	Count
1	WA	4/1	500
3	Sichuan	4/8	1000
5	WA	4/15	3000

Computing New Variables

	Region	Date	Count
1	WA	4/1	500
3	Sichuan	4/8	1000
5	WA	4/15	3000

```
WApop <- 10^7
Spop <- 10^8
```

Add a new variable `Pop` with

`covid$Pop <- ifelse(covid$Region == ‘WA’, WApop, Spop)` or

`covid <- covid %>% mutate(Pop = ifelse(Region == ‘WA’, WApop,`

	Region	Date	Count	Pop
1	WA	4/1	500	1e7
3	Sichuan	4/8	1000	1e8
5	WA	4/15	3000	1e7

Computing New Variables Cont'd

	Region	Date	Count	Pop
1	WA	4/1	500	1e7
3	Sichuan	4/8	1000	1e8
5	WA	4/15	3000	1e7

Add a new variable `FracInfected` with `covid$FracInfected <- covid$Count/covid$Pop` or `covid <- covid %>% mutate(FracInfected = Count/Pop)`:

	Region	Date	Count	Pop	FracInfected
1	WA	4/1	500	1e7	5e-5
3	Sichuan	4/8	1000	1e8	1e-5
5	WA	4/15	3000	1e7	3e-4

Compute Summary Statistics

	Region	Date	Count	Pop	FracInfected
1	WA	4/1	500	1e7	5e-5
3	Sichuan	4/8	1000	1e8	1e-5
5	WA	4/15	3000	1e7	3e-4

Find the maximum case count with `max(covid$Count)` or
`covid %>% summarise(maxCases = max(Count))`

Compute Summary Statistics **by Group**

	Region	Date	Count	Pop	FracInfected
1	WA	4/1	500	1e7	5e-5
3	Sichuan	4/8	1000	1e8	1e-5
5	WA	4/15	3000	1e7	3e-4

Find the maximum case count in each region with

```
covid %>% group_by(Region) %>% summarise(maxCases = max(Count)
```

Not so simple without `dplyr`!

	Region	maxCount
1	Sichuan	1000
2	WA	3000