2: Data Wrangling with dplyr

Suppose we have the following toy data set, named df. The first two columns are numeric while the third column is categorical.

x1	x2	cat1	y1	New var
1	3	Yes	0.333	small x1
7	NA	Yes	NA	large x1
4	2	No	2	large x1

Create New Variables with mutate() (Perhaps with case_when() or if_else())

name of new variable

```
df %>% mutate(y1 = x1 / x2)
```

operations on old variables

```
df %>% mutate(newvar = if_else(x1 > 3,
                        true = "large x1",
                        false = "small x1"))
```

Choose Rows to Keep with filter()

condition

```
df %>% filter(x1 > 6)

=>  x1   x2   cat1
    7    NA   Yes

df %>% filter(cat1 == "No")
df %>% filter(x1 == max(x1))
=>  x1   x2   cat1
    7    NA   Yes
```

Choose Columns to Keep with select()

```
df %>% select(x2, cat1)

=>  x2   cat1
    3    Yes
    NA   Yes
    2    No
```

x1	x2	cat1
1	3	Yes
7	NA	Yes
4	2	No

Order/Sort Your Data Set with arrange()

```
df %>% arrange(x1)
```

=>

x1	x2	cat1
1	3	Yes
4	2	No
7	NA	Yes

Obtain Numerical Summaries with summarise()

```
df %>% summarise(x1m = median(x1))
```

=>

x1m
4

Obtain Numerical Summaries by Group with group_by() and summarise()

```
df %>% group_by(cat1) %>%
    summarise(nummins = min(x1))
```

mean(), median(), max(), min(), sd(), var(), quantile(), etc.

=>

cat1	nummns
Yes	1
No	4