Lecture 7: Data manipulation

Skidmore College FYE

Goals

- Data manipulation
- Using data manipulation to create better graphs
- Summary statistics

NFL kicker information

```
##   Team Year GameMinute Kicker Distance ScoreDiff Grass Success
## 1  PHI 2005          3  Akers       49         0 FALSE       0
## 2  PHI 2005         29  Akers       49        -7 FALSE       0
## 3  PHI 2005         51  Akers       44        -7 FALSE       1
## 4  PHI 2005         14  Akers       43        14  TRUE       0
## 5  PHI 2005         60  Akers       23         0  TRUE       1
## 6  PHI 2005         39  Akers       34        -3  TRUE       1
```

Possible questions of interest

Summary of R-commands for data manipulation

1. `summarise()`
2. `group_by()`
3. `arrange()`
4. `count()`
5. `slice()`
6. `filter()`
7. `top_n()`

Example 1

Among kickers with at least 50 attempts, what are the 5 best success rates?

Step 1:

```
nfl.kick %>%
  summarise(success.rate = mean(Success))
```

```
##   success.rate
## 1    0.8326629
```

Example 1

Among kickers with at least 50 attempts, what are the 5 best success rates?

Step 2:

```
nfl.kick %>%
  group_by(Kicker) %>%
  summarise(success.rate = mean(Success)) %>%
  top_n(5)
```

```
## # A tibble: 5 x 2
##   Kicker   success.rate
##   <fct>           <dbl>
## 1 Bailey          0.895
## 2 Boswell         0.923
## 3 Hopkins         0.897
## 4 Peterson        0.92
## 5 Scifres         1
```

Example 1

Among kickers with at least 50 attempts, what kickers have the 5 best success rates?

Step 3:

```
nfl.kick %>%
  group_by(Kicker) %>%
  summarise(success.rate = mean(Success),
            n.kicks = n()) %>%
  filter(n.kicks >= 50) %>%
  top_n(5, success.rate)
```

```
## # A tibble: 5 x 3
##   Kicker     success.rate n.kicks
##   <fct>             <dbl>   <int>
## 1 Andersen          0.882      51
## 2 Bailey            0.895     162
## 3 Catanzaro         0.894      66
## 4 Gostkowski        0.877     342
## 5 Tucker            0.885     156
```

Write your own code

Among kickers with at least 50 attempts, what kickers have the 5 longest average distances?

Example 2

Does the fraction of kicks on grass surface vary based on the kick distance?

Step 1:

```
nfl.kick %>%
  summarise(grass.rate = mean(Grass == "TRUE"))
```

```
##   grass.rate
## 1   0.548315
```

Example 2

Does the fraction of kicks on grass surface vary based on the kick distance?

Step 2:

```
nfl.kick %>%
  group_by(Distance) %>%
  summarise(grass.rate = mean(Grass == "TRUE"), n.kicks = n()) %>%
  filter(n.kicks >= 20) %>%
  slice(1:4)
```

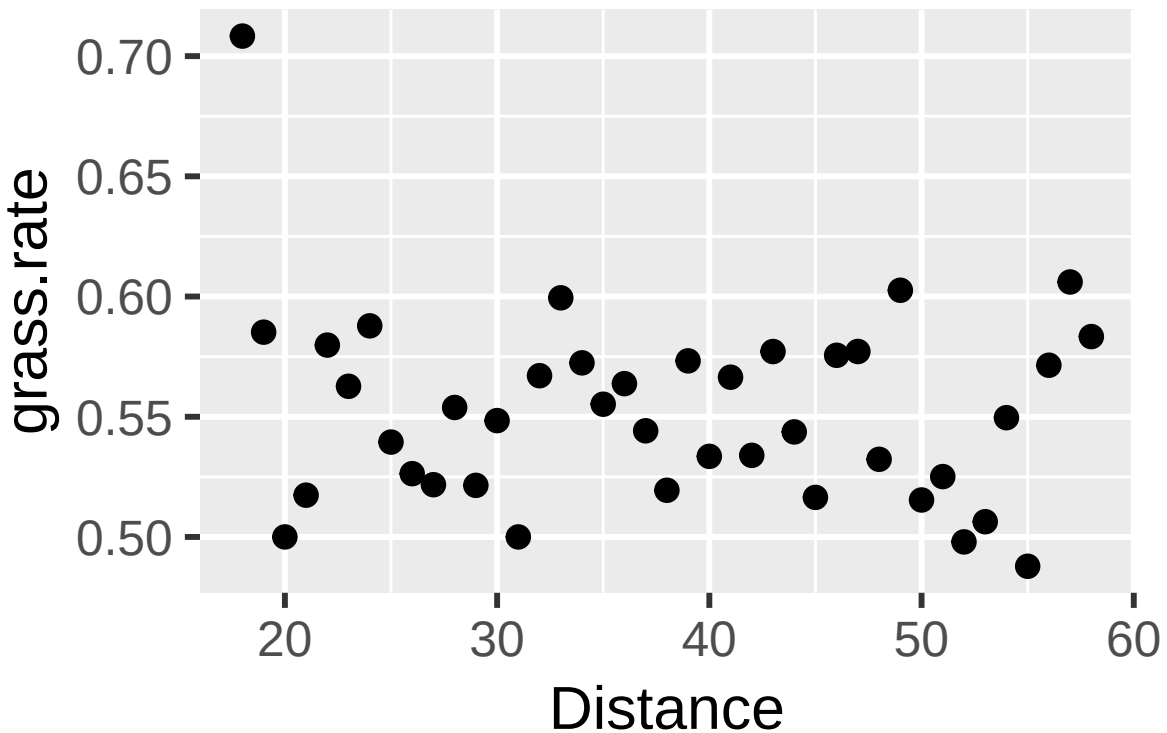
```
## # A tibble: 4 x 3
##   Distance grass.rate n.kicks
##      <int>      <dbl>   <int>
## 1       18      0.708      24
## 2       19      0.585     135
## 3       20      0.5       274
## 4       21      0.517     259
```

Example 2

Does the fraction of kicks on grass surface vary based on the kick distance?

Step 3:

```
df.kicks <- nfl.kick %>%
  group_by(Distance) %>%
  summarise(grass.rate = mean(Grass == "TRUE"), n.kicks = n()) %>%
  filter(n.kicks >= 20)
ggplot(data = df.kicks, aes(Distance, grass.rate)) +
  geom_point()
```



Write your own code

Does the fraction of made kicks vary based on game minute?

Example 3

Compare Matt Bryant and Steven Gostkowski based on kicks made on a similar surface and at similar distance. Which kicker was better?

Step 0:

```
nfl.kick %>%
  filter(Kicker == "Gostkowski" | Kicker == "Bryant") %>%
  group_by(Kicker) %>%
  summarise(success.rate = mean(Success), n.kicks = n())
```

```
## # A tibble: 2 x 3
##   Kicker     success.rate n.kicks
##   <fct>             <dbl>   <int>
## 1 Bryant            0.860     308
## 2 Gostkowski        0.877     342
```

Example 3

Compare Matt Bryant and Steven Gostkowski based on kicks made on a similar surface and at similar distance. Which kicker was better?

Step 1:

```
nfl.kick <- nfl.kick %>%
  mutate(Distance.cat = case_when(Distance <= 40 ~ "0 - 40",
                                  Distance <= 50 ~ "41 - 50",
                                  Distance > 50 ~ "50 or above"))
```

Example 3

Compare Matt Bryant and Steven Gostkowski based on kicks made on a similar surface and at similar distance. Which kicker was better?

Step 2:

```
kick.results <- nfl.kick %>%
  filter(Kicker == "Gostkowski" | Kicker == "Bryant") %>%
  group_by(Kicker, Grass, Distance.cat) %>%
  summarise(success.rate = mean(Success), n.kicks = n())
```

Example 3

Step 3:

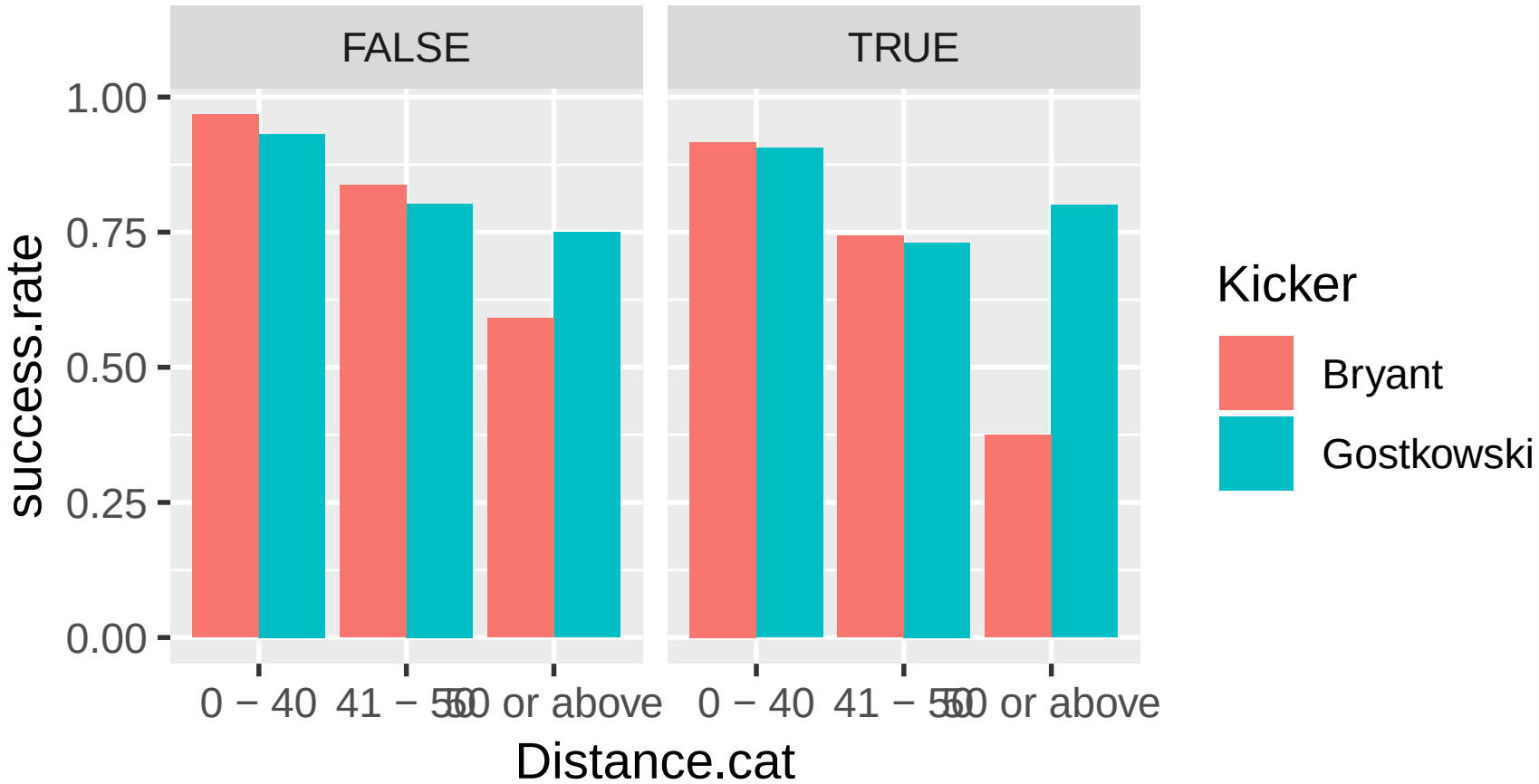
```
kick.results
```

```
## # A tibble: 12 x 5
## # Groups:   Kicker, Grass [?]
##    Kicker     Grass Distance.cat success.rate n.kicks
##    <fct>      <lgl> <chr>                <dbl>   <int>
##  1 Bryant     FALSE 0 - 40               0.968      94
##  2 Bryant     FALSE 41 - 50              0.837      49
##  3 Bryant     FALSE 50 or above          0.591      22
##  4 Bryant     TRUE  0 - 40               0.917      96
##  5 Bryant     TRUE  41 - 50              0.744      39
##  6 Bryant     TRUE  50 or above          0.375       8
##  7 Gostkowski FALSE 0 - 40               0.931     175
##  8 Gostkowski FALSE 41 - 50              0.803      71
##  9 Gostkowski FALSE 50 or above          0.75       12
## 10 Gostkowski TRUE  0 - 40               0.906      53
## 11 Gostkowski TRUE  41 - 50              0.731      26
## 12 Gostkowski TRUE  50 or above          0.8         5
```

Example 3

Step 4:

```
ggplot(data = kick.results, aes(x = Distance.cat, y = success.rate, fill = Kick
  geom_col(position='dodge') +
  facet_wrap(~Grass)
```



Write your own code