Lecture 10 Notes

2024-02-20

Loading the data

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

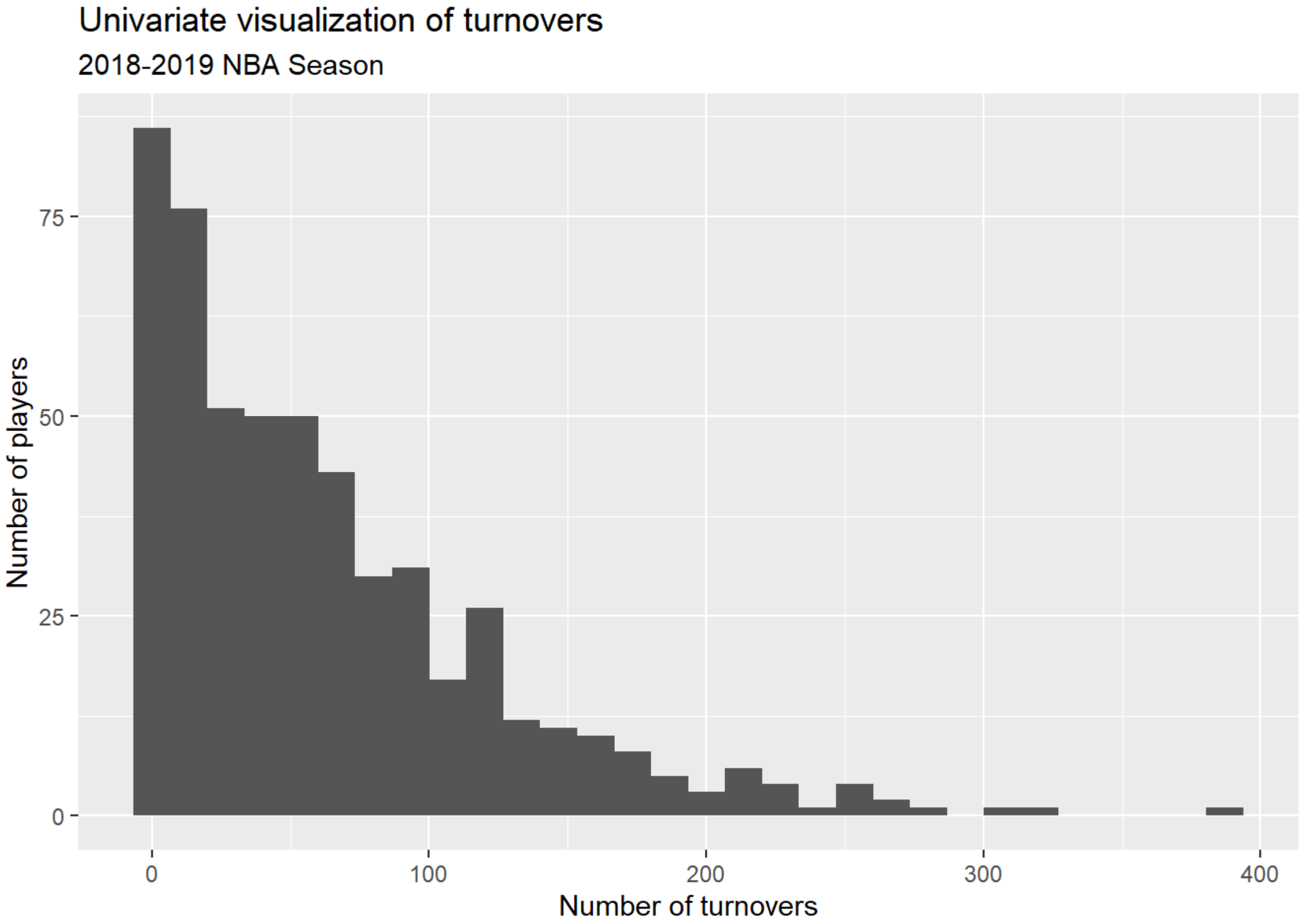
```
nba <- read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/nba_players_201
8.Rds')
glimpse(nba %>% select(tov,isRookie))
```

```
## Rows: 530
## Columns: 2
## $ tov      <dbl> 144, 4, 135, 14, 121, 8, 33, 6, 28, 2, 72, 268, 58, 23, 103, …
## $ isRookie <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TR…
```

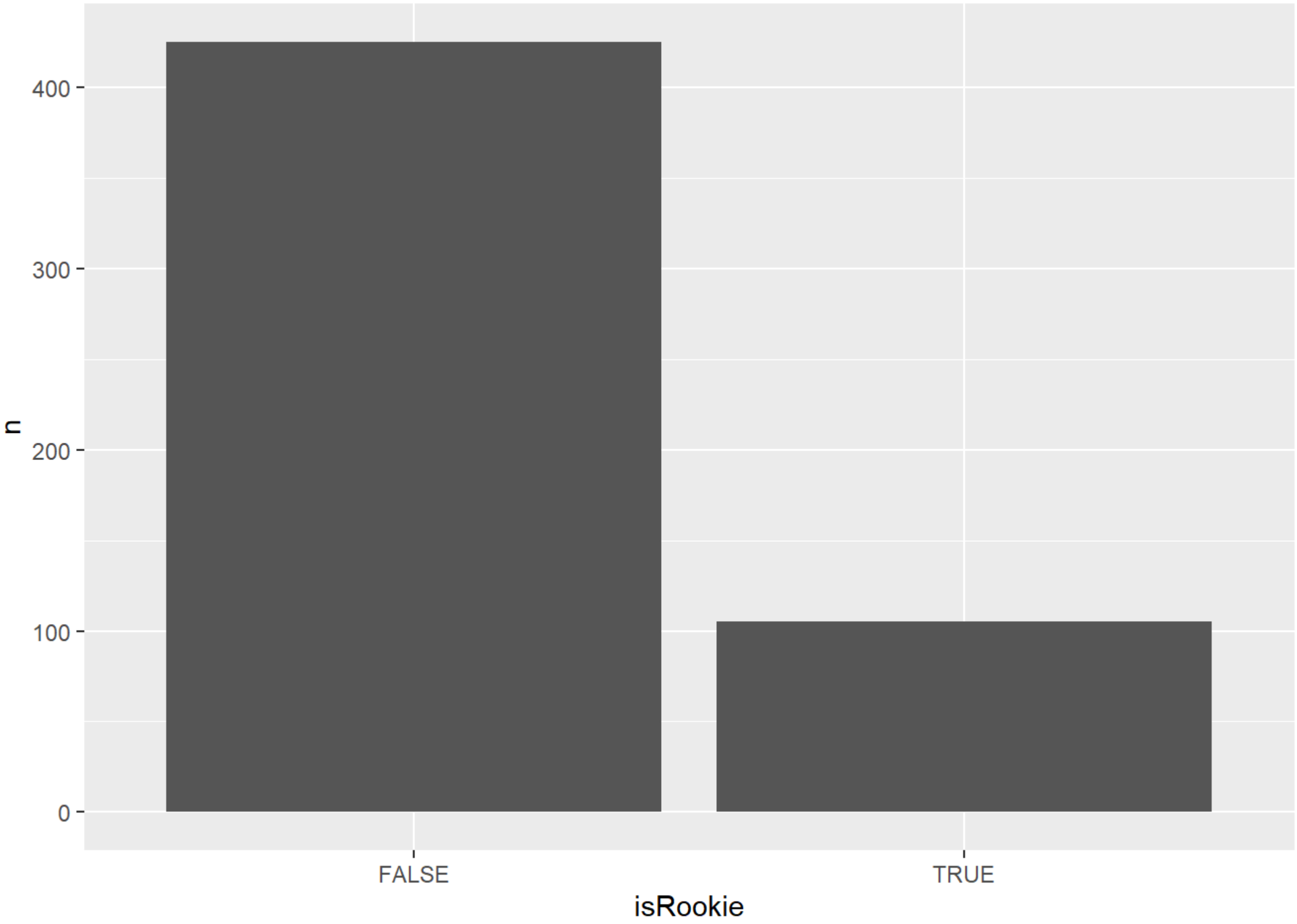
Univariate visualizations of \(X\) and \(Y\)

```
# Y
nba %>%
  ggplot(aes(x = tov)) +
  geom_histogram() +
  labs(x = 'Number of turnovers',
       y = 'Number of players',
       title = 'Univariate visualization of turnovers',
       subtitle = '2018-2019 NBA Season')
```

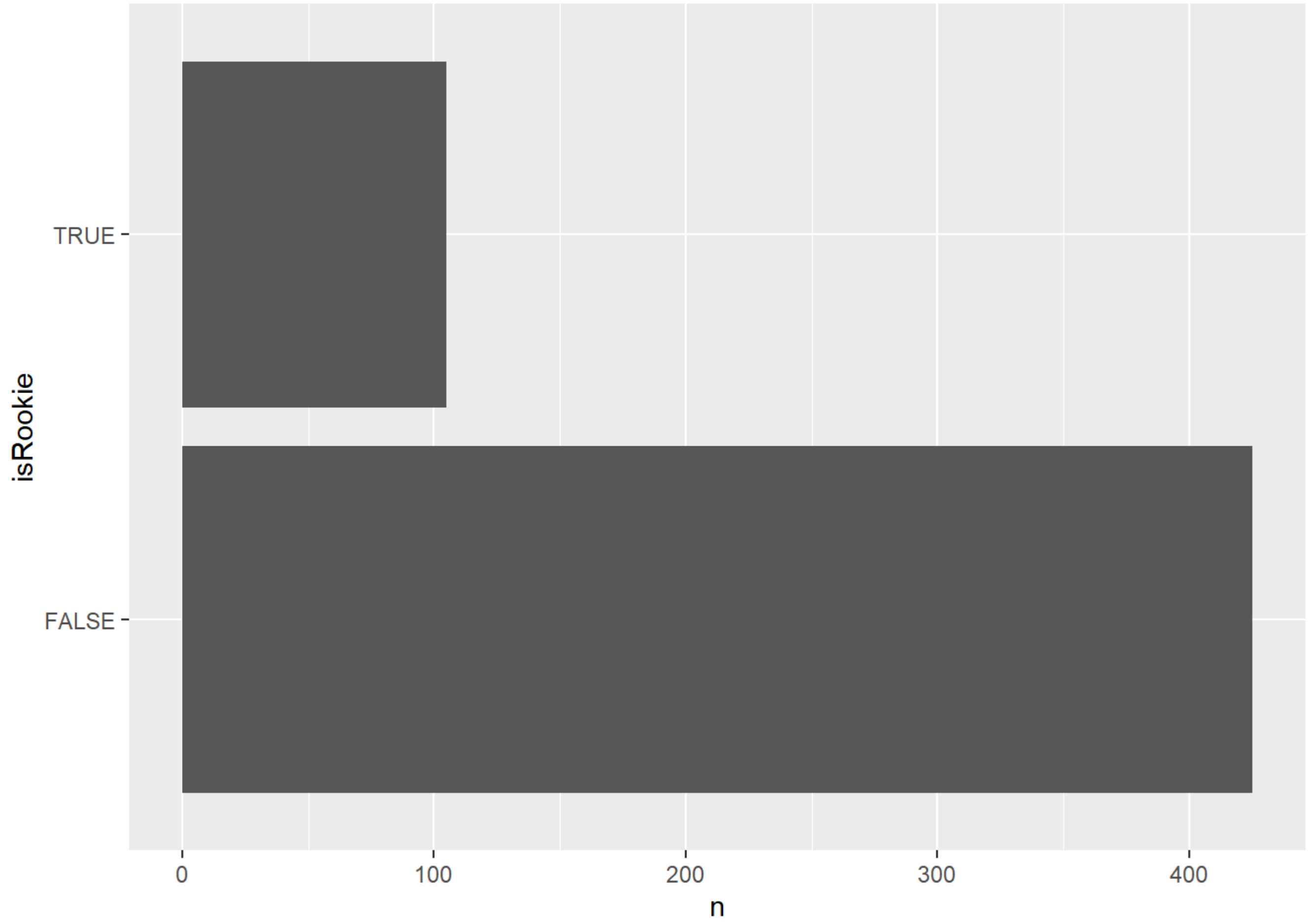
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



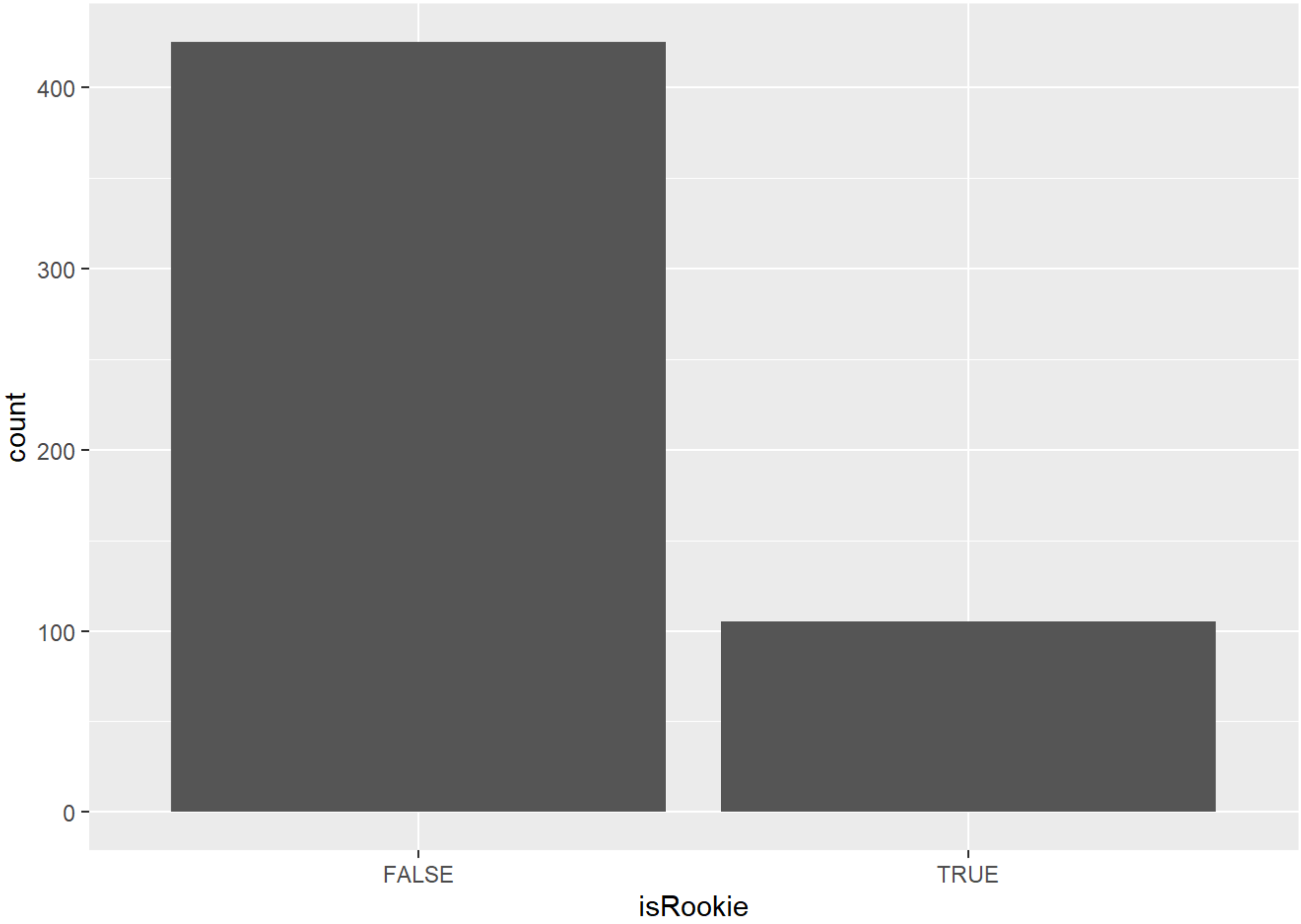
```
# X
nba %>%
  count(isRookie) %>%
  ggplot(aes(x = isRookie,y = n)) +
  geom_bar(stat = 'identity')
```



```
nba %>%
  count(isRookie) %>%
  ggplot(aes(y = isRookie,x = n)) +
  geom_bar(stat = 'identity')
```

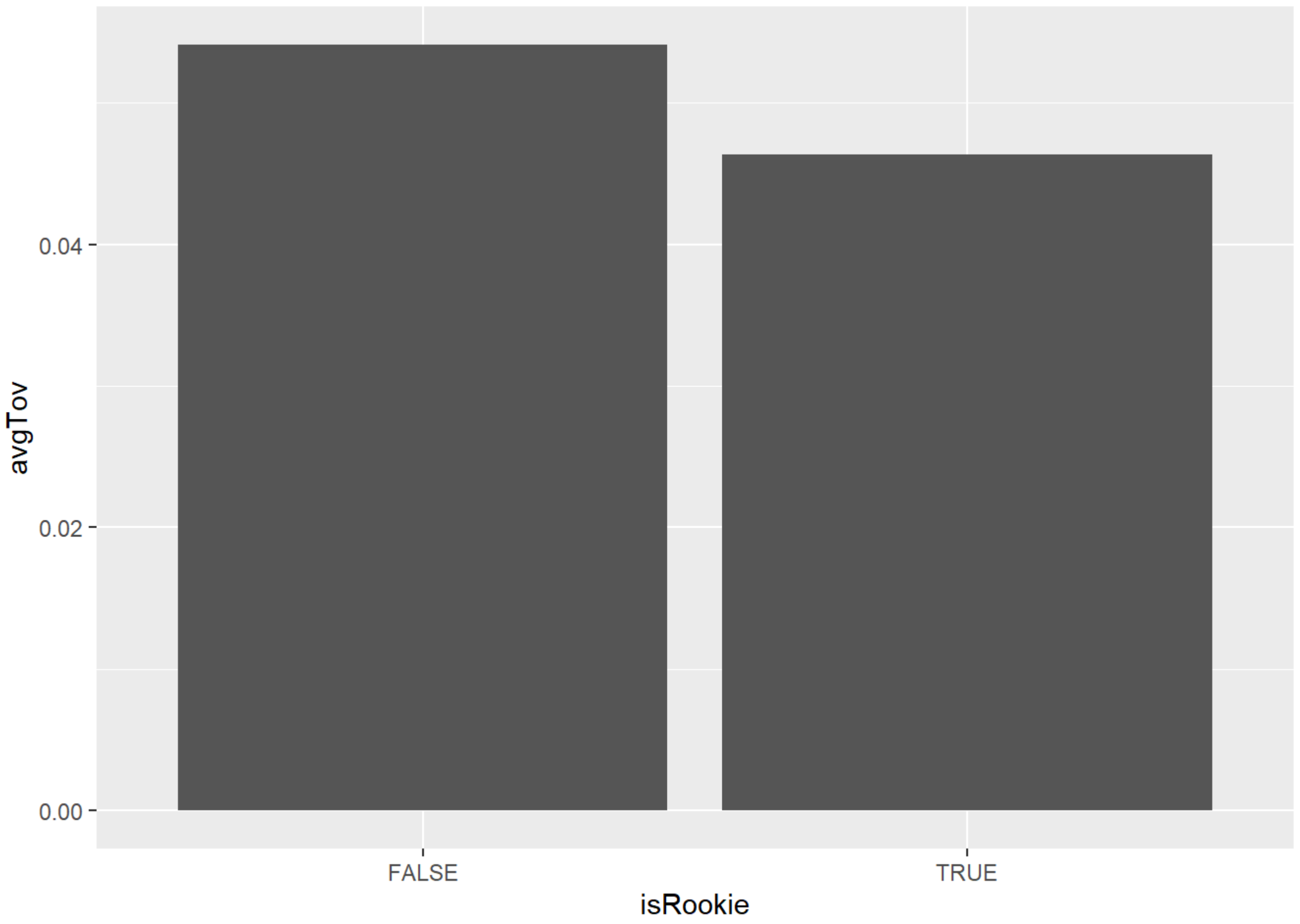


```
nba %>%
  ggplot(aes(x = isRookie)) +
  geom_bar()
```



Multivariate Visualization

```
nba %>%
  group_by(isRookie) %>%
  summarise(avgTov = mean(tov/minutes,na.rm=T)) %>%
  ggplot(aes(x = isRookie,
             y = avgTov)) +
  geom_bar(stat = 'identity')
```



sample_n()

```
set.seed(123)

nba %>%
  sample_n(size = 530,replace = T) %>%
  select(namePlayer,isRookie,tov,minutes) %>%
  group_by(isRookie) %>%
  summarise(avgTov = mean(tov/minutes))
```

```
## # A tibble: 2 × 2
##   isRookie avgTov
##   <lgl>     <dbl>
## 1 FALSE    0.0526
## 2 TRUE     0.0461
```

for()

```
results <- NULL
for(i in 1:100) {
  results <- results %>%
    bind_rows(nba %>%
                sample_n(size = nrow(nba),replace = T) %>%
                select(namePlayer,isRookie,tov,minutes) %>%
                group_by(isRookie) %>%
                summarise(avgTov = mean(tov/minutes)) %>%
                mutate(simNumber = i))
}

results %>%
  pivot_wider(names_from = 'isRookie',
              values_from = 'avgTov')
```

```
## # A tibble: 100 × 3
##    simNumber `FALSE` `TRUE`
##        <int>   <dbl>  <dbl>
##  1         1  0.0530 0.0444
##  2         2  0.0509 0.0504
##  3         3  0.0537 0.0441
##  4         4  0.0552 0.0523
##  5         5  0.0548 0.0444
##  6         6  0.0560 0.0438
##  7         7  0.0552 0.0461
##  8         8  0.0537 0.0485
##  9         9  0.0553 0.0521
## 10        10  0.0547 0.0465
## # ℹ 90 more rows
```

```
results %>%
  spread(isRookie,avgTov) %>%
  # summarise(conf = mean(`FALSE` > `TRUE`))
  rename(Veteran = `FALSE`,
         Rookie = `TRUE`) %>%
  mutate(rookieBetter = ifelse(Rookie < Veteran,
                               'Rookie is better',
                               'Veteran is better')) %>%
  summarise(conf = mean(rookieBetter == 'Rookie is better'))
```

```
## # A tibble: 1 × 1
##    conf
##   <dbl>
## 1  0.99
```