Lecture 5 Notes

2024-07-08

Introducing new data

```
require(tidyverse)
```

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

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── 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
```

```
library(labelled)

nba <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/nba_pla
yers_2018.Rds")

nba %>%
  select(namePlayer)
```

```
## # A tibble: 530 × 1
##    namePlayer
##    <chr>
##  1 LaMarcus Aldridge
##  2 Quincy Acy
##  3 Steven Adams
##  4 Alex Abrines
##  5 Bam Adebayo
##  6 Rawle Alkins
##  7 Grayson Allen
##  8 Deng Adel
##  9 Jaylen Adams
## 10 DeVaughn Akoon-Purcell
## # ℹ 520 more rows
```

Summarizing single variables

```
summary(nba$pts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   115.0   419.0   516.2   759.5  2818.0
```

```
summary(nba %>% select(pts))
```

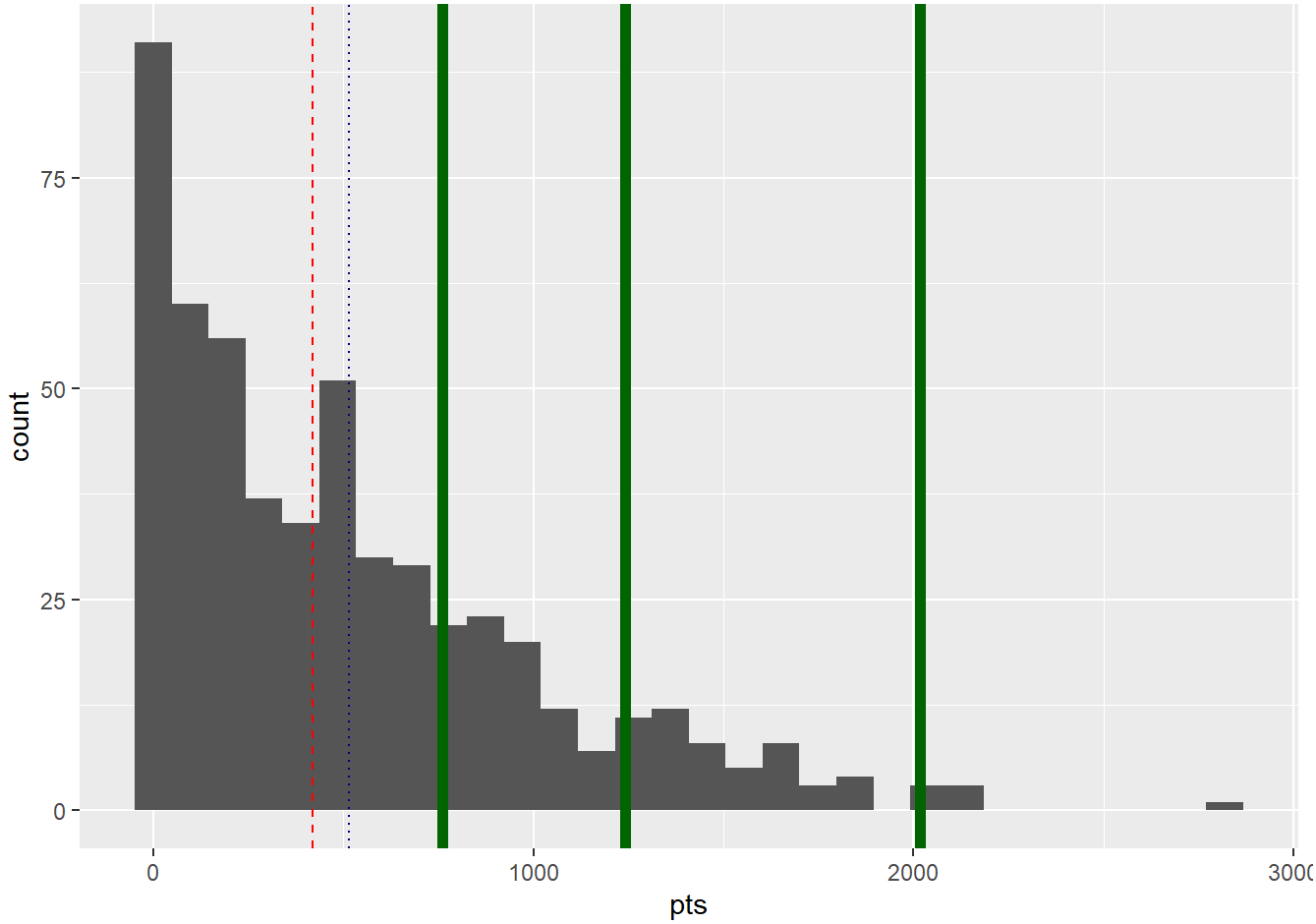
```
##       pts
##  Min.   :   0.0
##  1st Qu.: 115.0
##  Median : 419.0
##  Mean   : 516.2
##  3rd Qu.: 759.5
##  Max.   :2818.0
```

Visualizing with summary statistics

```
nba %>%
  ggplot(aes(x = pts)) +
  geom_histogram() + # geom_density
  geom_vline(xintercept = mean(nba$pts),
             color = 'darkblue',linetype = 'dotted') +
  geom_vline(xintercept = median(nba$pts),
             color = 'red',linetype = 'dashed') +
  geom_vline(xintercept = quantile(nba$pts,c(.75,.90,.99)),
             color = 'darkgreen',size = 2)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

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



Looking at every variable

```
glimpse(nba)
```

```
## Rows: 530
## Columns: 37
## $ namePlayer         <chr> "LaMarcus Aldridge", "Quincy Acy", "Steven Adams", …
## $ idPlayer           <dbl> 200746, 203112, 203500, 203518, 1628389, 1628959, 1…
## $ slugSeason         <chr> "2018-19", "2018-19", "2018-19", "2018-19", "2018-1…
## $ numberPlayerSeason <dbl> 12, 6, 5, 2, 1, 0, 0, 0, 0, 0, 8, 5, 4, 3, 1, 1, 1,…
## $ isRookie           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE…
## $ slugTeam           <chr> "SAS", "PHX", "OKC", "OKC", "MIA", "CHI", "UTA", "C…
## $ idTeam             <dbl> 1610612759, 1610612756, 1610612760, 1610612760, 161…
## $ gp                 <dbl> 81, 10, 80, 31, 82, 10, 38, 19, 34, 7, 81, 72, 43, …
## $ gs                 <dbl> 81, 0, 80, 2, 28, 1, 2, 3, 1, 0, 81, 72, 40, 4, 80,…
## $ fgm                <dbl> 684, 4, 481, 56, 280, 13, 67, 11, 38, 3, 257, 721, …
## $ fga                <dbl> 1319, 18, 809, 157, 486, 39, 178, 36, 110, 10, 593,…
## $ pctFG              <dbl> 0.519, 0.222, 0.595, 0.357, 0.576, 0.333, 0.376, 0.…
## $ fg3m               <dbl> 10, 2, 0, 41, 3, 3, 32, 6, 25, 0, 96, 52, 9, 24, 6,…
## $ fg3a               <dbl> 42, 15, 2, 127, 15, 12, 99, 23, 74, 4, 280, 203, 34…
## $ pctFG3             <dbl> 0.2380952, 0.1333333, 0.0000000, 0.3228346, 0.20000…
## $ pctFT              <dbl> 0.847, 0.700, 0.500, 0.923, 0.735, 0.667, 0.750, 1.…
## $ fg2m               <dbl> 674, 2, 481, 15, 277, 10, 35, 5, 13, 3, 161, 669, 1…
## $ fg2a               <dbl> 1277, 3, 807, 30, 471, 27, 79, 13, 36, 6, 313, 1044…
## $ pctFG2             <dbl> 0.5277995, 0.6666667, 0.5960347, 0.5000000, 0.58811…
## $ agePlayer          <dbl> 33, 28, 25, 25, 21, 21, 23, 22, 23, 26, 28, 24, 25,…
## $ minutes            <dbl> 2687, 123, 2669, 588, 1913, 120, 416, 194, 428, 22,…
## $ ftm                <dbl> 349, 7, 146, 12, 166, 8, 45, 4, 7, 1, 150, 500, 37,…
## $ fta                <dbl> 412, 10, 292, 13, 226, 12, 60, 4, 9, 2, 173, 686, 6…
## $ oreb               <dbl> 251, 3, 391, 5, 165, 11, 3, 3, 11, 1, 112, 159, 48,…
## $ dreb               <dbl> 493, 22, 369, 43, 432, 15, 20, 16, 49, 3, 498, 739,…
## $ treb               <dbl> 744, 25, 760, 48, 597, 26, 23, 19, 60, 4, 610, 898,…
## $ ast                <dbl> 194, 8, 124, 20, 184, 13, 25, 5, 65, 6, 104, 424, 1…
## $ stl                <dbl> 43, 1, 117, 17, 71, 1, 6, 1, 14, 2, 68, 92, 54, 22,…
## $ blk                <dbl> 107, 4, 76, 6, 65, 0, 6, 4, 5, 0, 33, 110, 37, 13, …
## $ tov                <dbl> 144, 4, 135, 14, 121, 8, 33, 6, 28, 2, 72, 268, 58,…
## $ pf                 <dbl> 179, 24, 204, 53, 203, 7, 47, 13, 45, 4, 143, 232, …
## $ pts                <dbl> 1727, 17, 1108, 165, 729, 37, 211, 32, 108, 7, 760,…
## $ urlNBAAPI          <chr> "https://stats.nba.com/stats/playercareerstats?Leag…
## $ n                  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
## $ org                <fct> Texas, NA, Other, FC Barcelona Basquet, Kentucky, N…
## $ country            <chr> NA, NA, NA, "Spain", NA, NA, NA, NA, NA, NA, NA, "S…
## $ idConference       <int> 2, 2, 2, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, …
```

Categorical

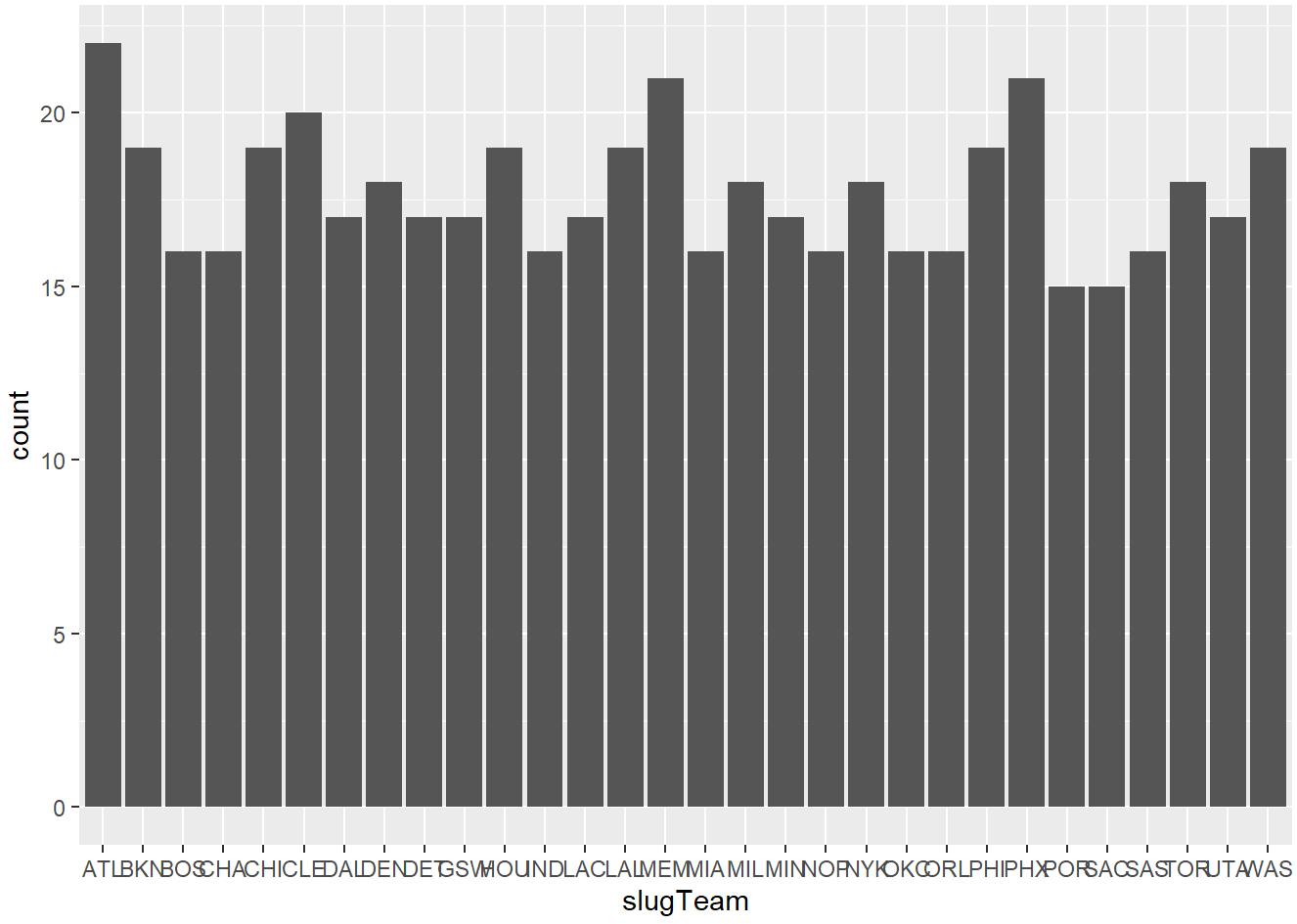
```
# Step 1: Look
summary(nba$slugTeam)
```

```
##    Length     Class      Mode
##       530 character character
```

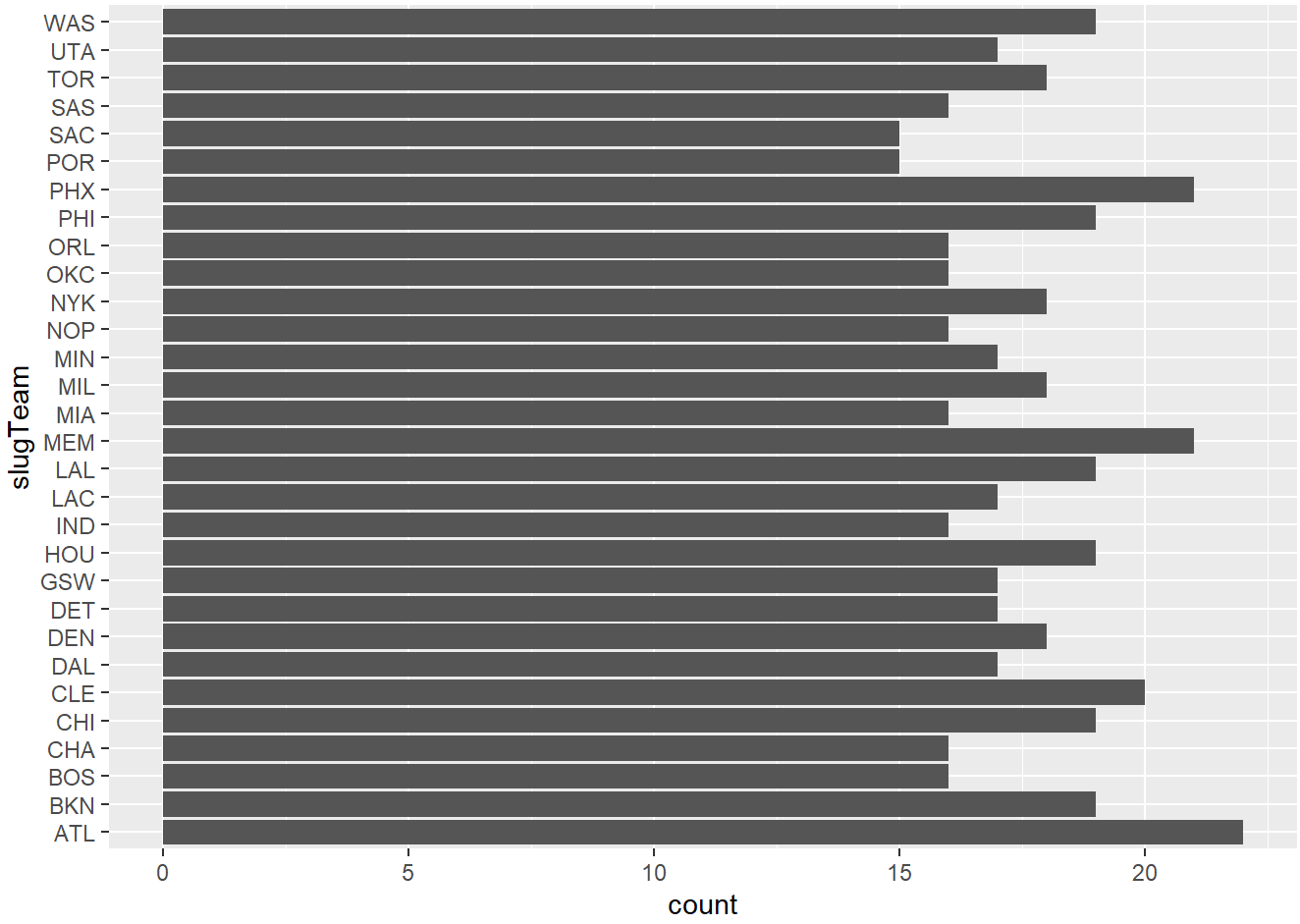
```
nba %>%
  select(slugTeam)
```

```
## # A tibble: 530 × 1
##    slugTeam
##    <chr>
##  1 SAS
##  2 PHX
##  3 OKC
##  4 OKC
##  5 MIA
##  6 CHI
##  7 UTA
##  8 CLE
##  9 ATL
## 10 DEN
## # ℹ 520 more rows
```

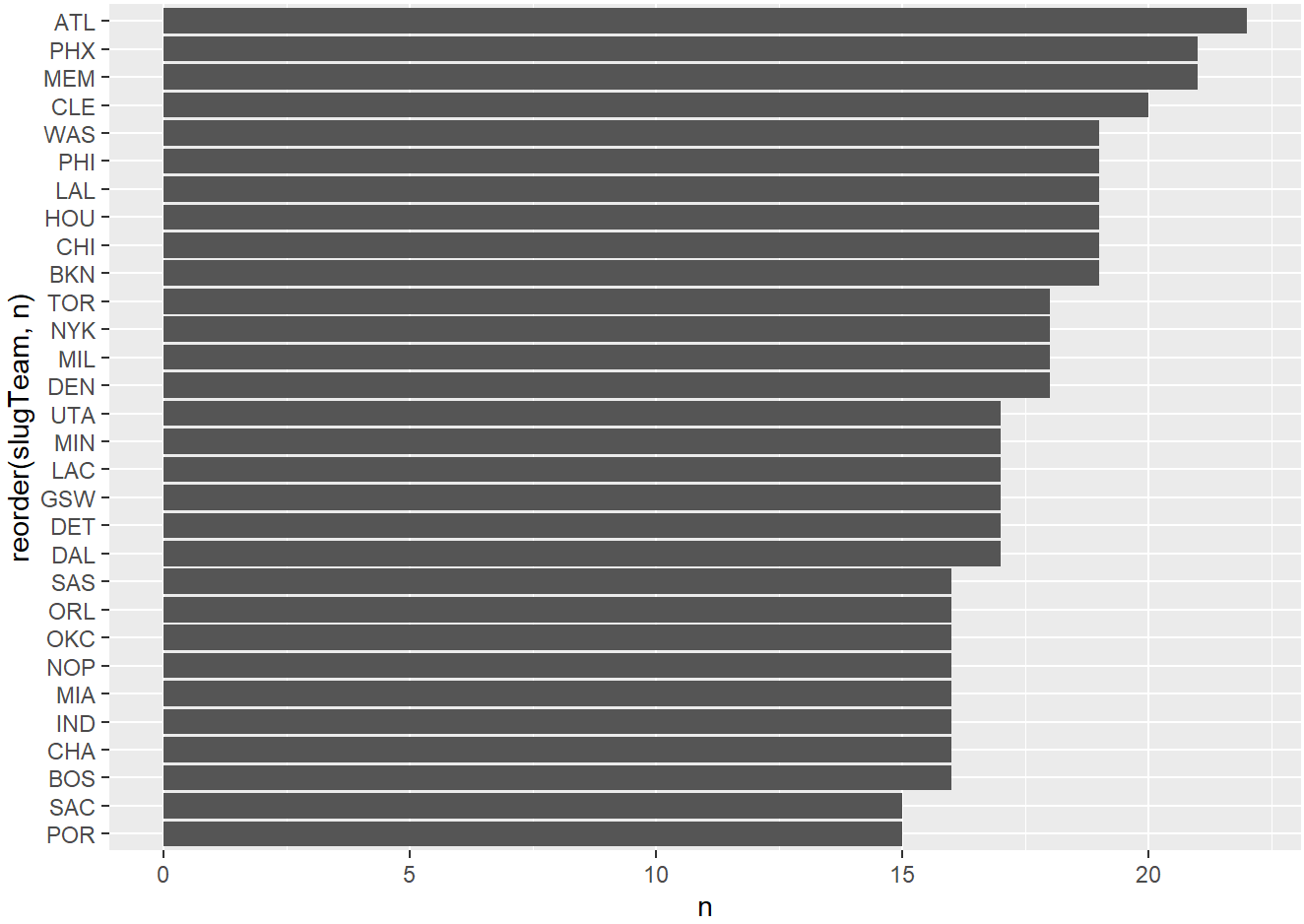
```
# Step 2: Visualize
nba %>%
  ggplot(aes(x = slugTeam)) +
  geom_bar()
```



```
nba %>%
  ggplot(aes(y = slugTeam)) +
  geom_bar()
```




```
# To order by values, need one extra step using count()
nba %>%
  count(slugTeam) %>%
  ggplot(aes(x = n,
             y = reorder(slugTeam,n))) +
  geom_bar(stat = 'identity')
```



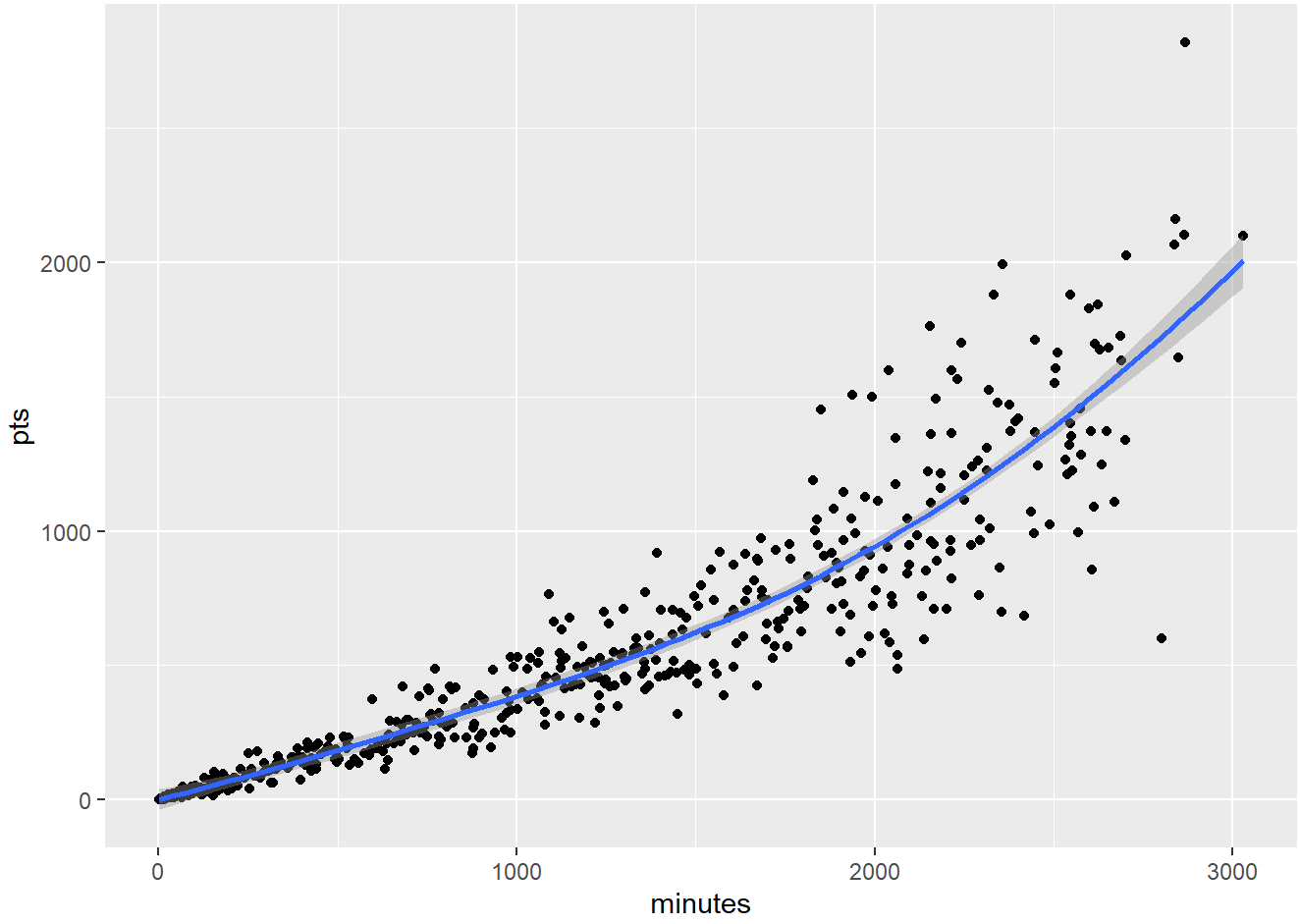

```
# Step 3: Summarize
nba %>%
  count(slugTeam) %>%
  arrange(desc(n))
```

```
## # A tibble: 30 × 2
##    slugTeam     n
##    <chr>    <int>
##  1 ATL         22
##  2 MEM         21
##  3 PHX         21
##  4 CLE         20
##  5 BKN         19
##  6 CHI         19
##  7 HOU         19
##  8 LAL         19
##  9 PHI         19
## 10 WAS         19
## # ℹ 20 more rows
```

Multivariate Visualization

```
nba %>%
  ggplot(aes(x = minutes,
             y = pts)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



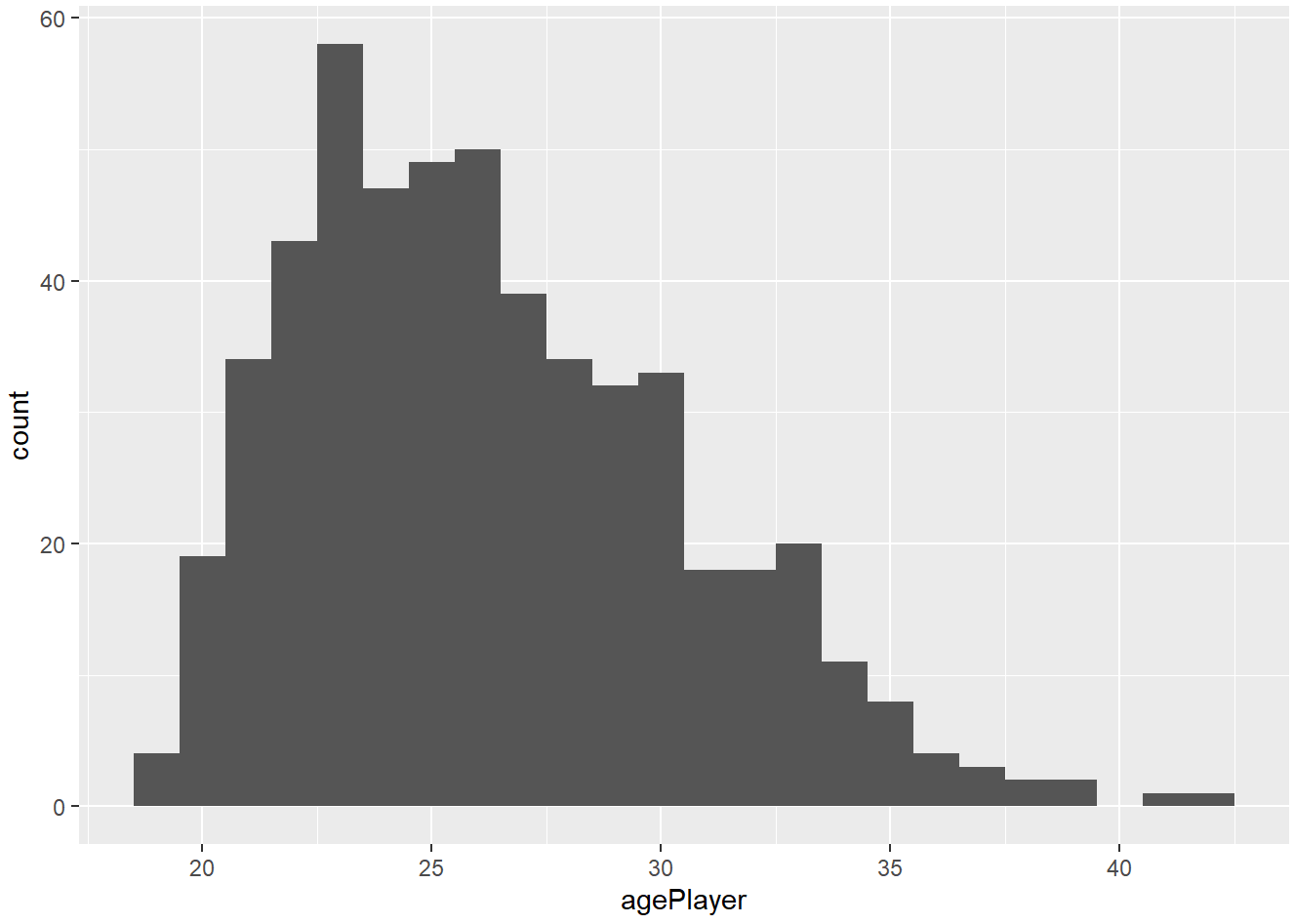
BREAK

Quick detour on player age

```
# Step 1: Look
nba %>%
  select(agePlayer)
```

```
## # A tibble: 530 × 1
##    agePlayer
##        <dbl>
##  1        33
##  2        28
##  3        25
##  4        25
##  5        21
##  6        21
##  7        23
##  8        22
##  9        23
## 10        26
## # ℹ 520 more rows
```

```
# Step 2: Visualize
nba %>%
  ggplot(aes(x = agePlayer)) +
  geom_histogram(binwidth = 1)
```




```
  # geom_density()
  geom_bar()
```

```
## geom_bar: just = 0.5, width = NULL, na.rm = FALSE, orientation = NA
## stat_count: width = NULL, na.rm = FALSE, orientation = NA
## position_stack
```

```
# Step 3: Summarize
```

New data

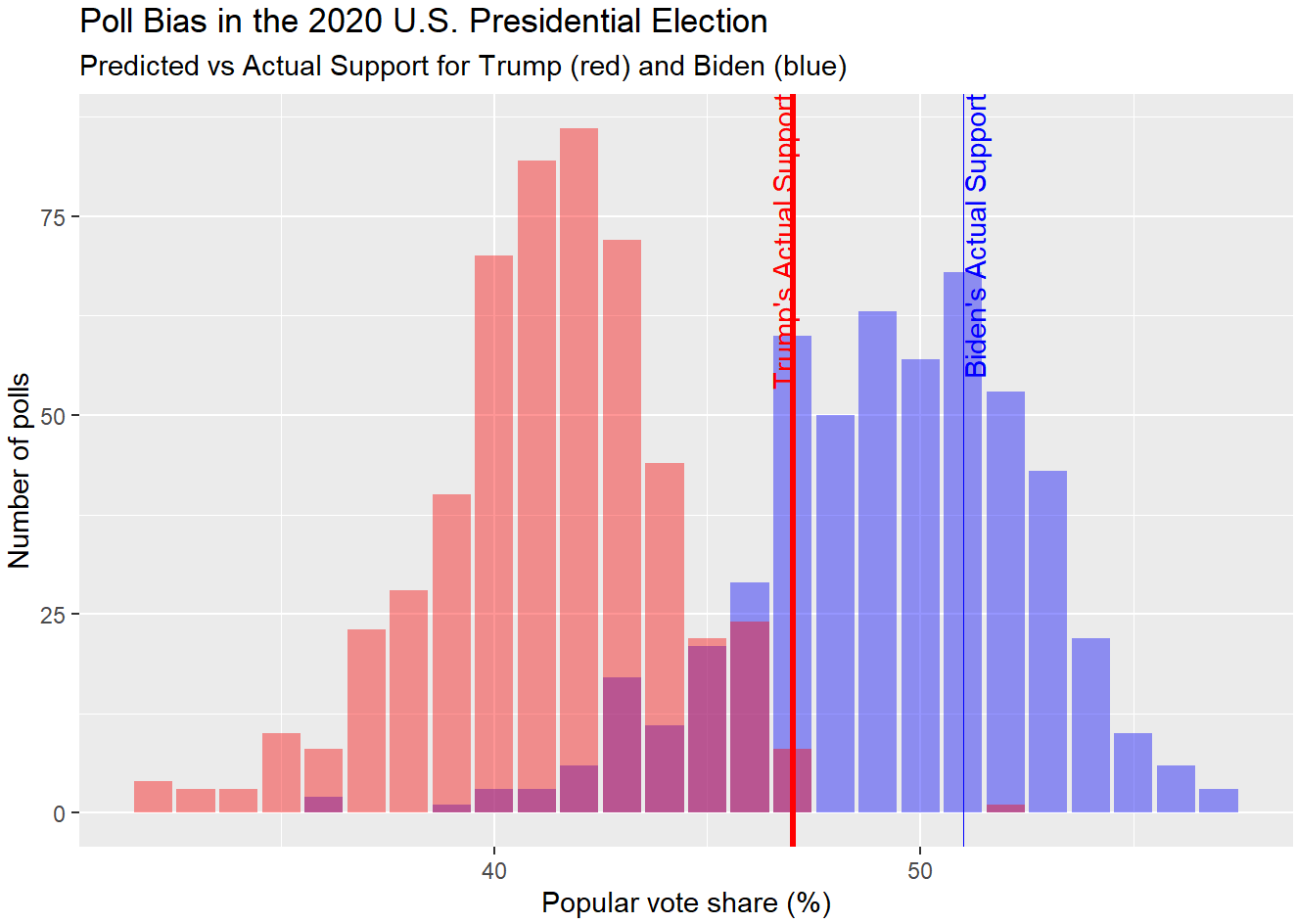
```
pres <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/Pres2020_PV.Rds")

pres
```

```
## # A tibble: 528 × 16
##    poll.id Geography Poll   StartDate EndDate DaysinField   MoE Mode  SampleSize
##      <dbl> <chr>     <chr>  <chr>     <chr>         <dbl> <dbl> <chr>      <dbl>
##  1    1942 NAT       Econo… 10/31/20… 11/2/2…           3 NA    Onli…       1363
##  2    1941 NAT       Resea… 10/31/20… 11/2/2…           3  3.1  Onli…        974
##  3    1940 NAT       Ipsos  10/29/20… 11/2/2…           5  3.7  Onli…        914
##  4    1939 NAT       Swaya… 11/1/2020 11/1/2…           1  1.7  Onli…       5174
##  5    1938 NAT       John … 11/1/2020 11/1/2…           1  3.2  <NA>        1008
##  6    1937 NAT       Yahoo… 10/30/20… 11/1/2…           3 NA    Onli…       1360
##  7    1936 NAT       Surve… 10/31/20… 11/2/2…           3  1    Onli…     799401
##  8    1935 NAT       Redfi… 10/30/20… 11/1/2…           3 NA    Onli…       8765
##  9    1934 NAT       Qriou… 10/29/20… 11/1/2…           4  2.2  Onli…       3505
## 10    1933 NAT       CNBC/… 10/29/20… 11/1/2…           4  2.26 Onli…       1880
## # ℹ 518 more rows
## # ℹ 7 more variables: Biden <dbl>, Trump <dbl>, DemCertVote <dbl>,
## #   RepCertVote <dbl>, Winner <chr>, Funded <chr>, Conducted <chr>
```

```
glimpse(pres)
```

```
## Rows: 528
## Columns: 16
## $ poll.id     <dbl> 1942, 1941, 1940, 1939, 1938, 1937, 1936, 1935, 1934, 1933…
## $ Geography   <chr> "NAT", "NAT", "NAT", "NAT", "NAT", "NAT", "NAT", "NAT", "N…
## $ Poll        <chr> "Economist/YouGov", "Research Co.", "Ipsos", "Swayable", "…
## $ StartDate   <chr> "10/31/2020", "10/31/2020", "10/29/2020", "11/1/2020", "11…
## $ EndDate     <chr> "11/2/2020", "11/2/2020", "11/2/2020", "11/1/2020", "11/1/…
## $ DaysinField <dbl> 3, 3, 5, 1, 1, 3, 3, 3, 4, 4, 2, 5, 5, 14, 2, 3, 3, 3, 3, …
## $ MoE         <dbl> NA, 3.10, 3.70, 1.70, 3.20, NA, 1.00, NA, 2.20, 2.26, 2.50…
## $ Mode        <chr> "Online", "Online", "Online", "Online", NA, "Online", "Onl…
## $ SampleSize  <dbl> 1363, 974, 914, 5174, 1008, 1360, 799401, 8765, 3505, 1880…
## $ Biden       <dbl> 53, 53, 52, 52, 48, 53, 52, 53, 52, 52, 48, 50, 49, 54, 48…
## $ Trump       <dbl> 43, 44, 45, 46, 42, 43, 46, 41, 41, 42, 47, 39, 46, 43, 39…
## $ DemCertVote <dbl> 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51…
## $ RepCertVote <dbl> 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47, 47…
## $ Winner      <chr> "Dem", "Dem", "Dem", "Dem", "Dem", "Dem", "Dem", "Dem", "D…
## $ Funded      <chr> "Economist", "Research Co.", "Reuters", "Swayable", "John …
## $ Conducted   <chr> "YouGov", "Research Co.", "Ipsos", "Swayable", "John Zogby…
```

```
pres %>%
  ggplot(aes(x = Biden)) +
  geom_bar(fill = 'blue',alpha = .4) +
  geom_bar(aes(x = Trump),fill = 'red',alpha = .4) +
  geom_vline(xintercept = mean(pres$DemCertVote),
             color = 'blue',size = 0.2) +
  geom_vline(xintercept = mean(pres$RepCertVote),
             color = 'red',size = 1.2) +
  annotate(geom = "text",x = mean(pres$DemCertVote),
           y = Inf,label = "Biden's Actual Support",
           color = 'blue',angle = 90,hjust = 1,vjust = 1) +   annotate(geom = "text",x =
mean(pres$RepCertVote),
           y = Inf,label = "Trump's Actual Support",
           color = 'red',angle = 90,hjust = 1,vjust = 0) +
  labs(x = "Popular vote share (%)",
       y = "Number of polls",
       title = "Poll Bias in the 2020 U.S. Presidential Election",
       subtitle = "Predicted vs Actual Support for Trump (red) and Biden (blue)")
```



Theory 1: Mode of survey

```
pres %>%
  count(Mode)
```

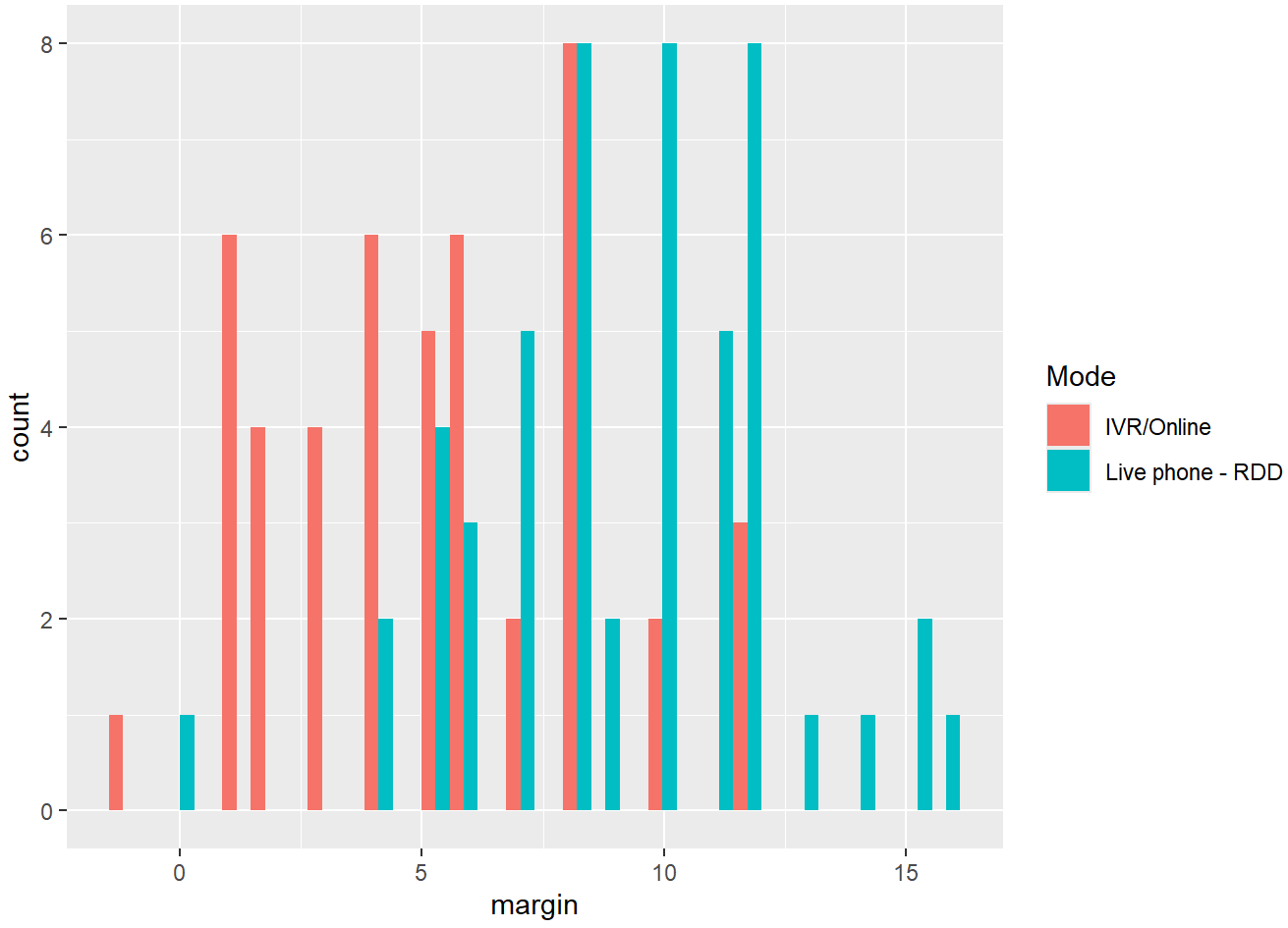
```
## # A tibble: 9 × 2
##   Mode                n
##   <chr>           <int>
## 1 IVR                 1
## 2 IVR/Online         47
## 3 Live phone - RBS    13
## 4 Live phone - RDD    51
## 5 Online             366
## 6 Online/Text         1
## 7 Phone - unknown     1
## 8 Phone/Online       19
## 9 <NA>               29
```

```
pres %>%
  filter(Mode == "IVR/Online" | Mode == "Live phone - RDD") %>%
  count(Mode)
```

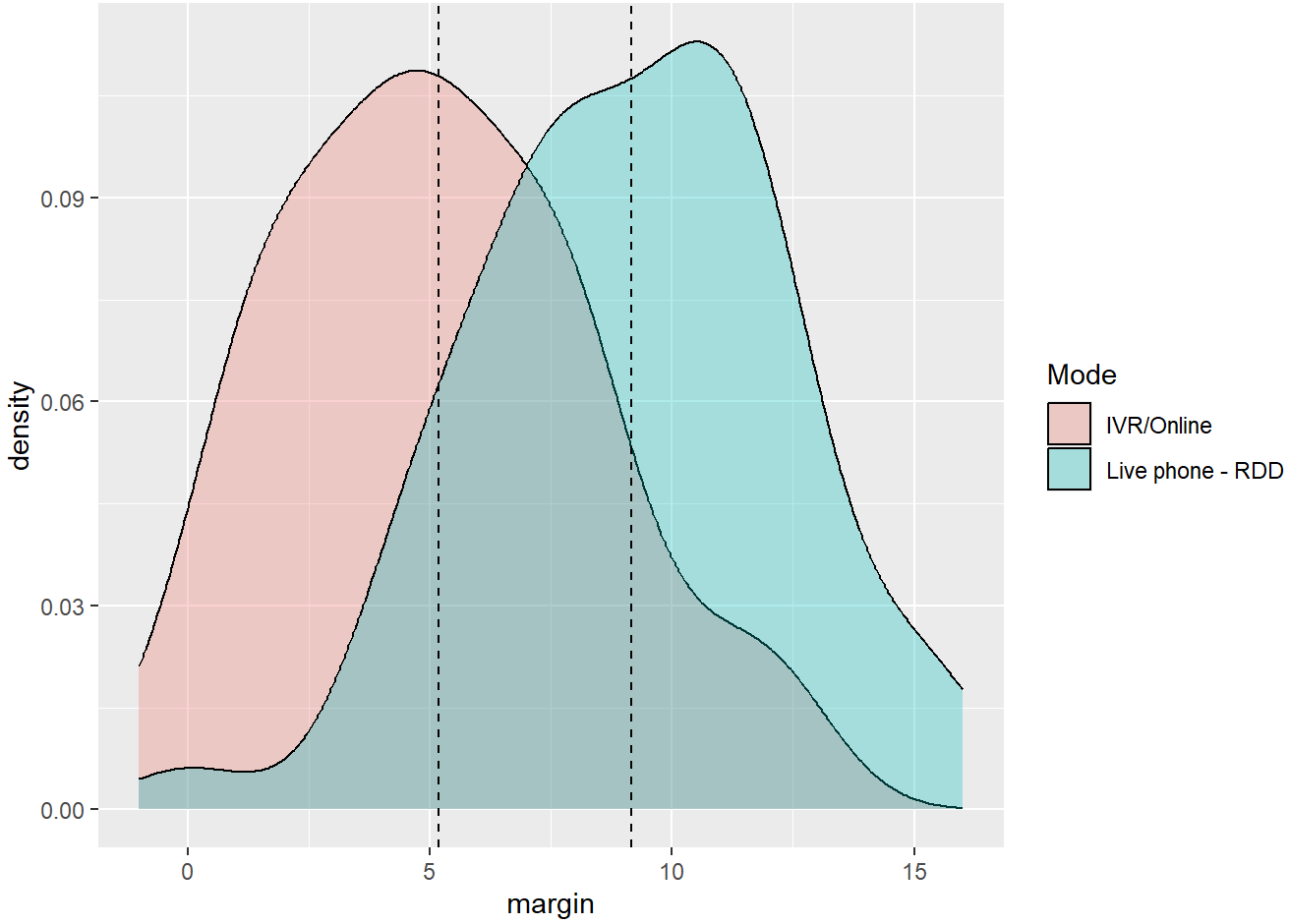
```
## # A tibble: 2 × 2
##   Mode                 n
##   <chr>            <int>
## 1 IVR/Online          47
## 2 Live phone - RDD    51
```

```
pres <- pres %>%
  mutate(margin = Biden - Trump)

pres %>%
  filter(Mode == "IVR/Online" | Mode == "Live phone - RDD") %>%
  ggplot(aes(x = margin,fill = Mode)) +
  geom_histogram(position = 'dodge')
```

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

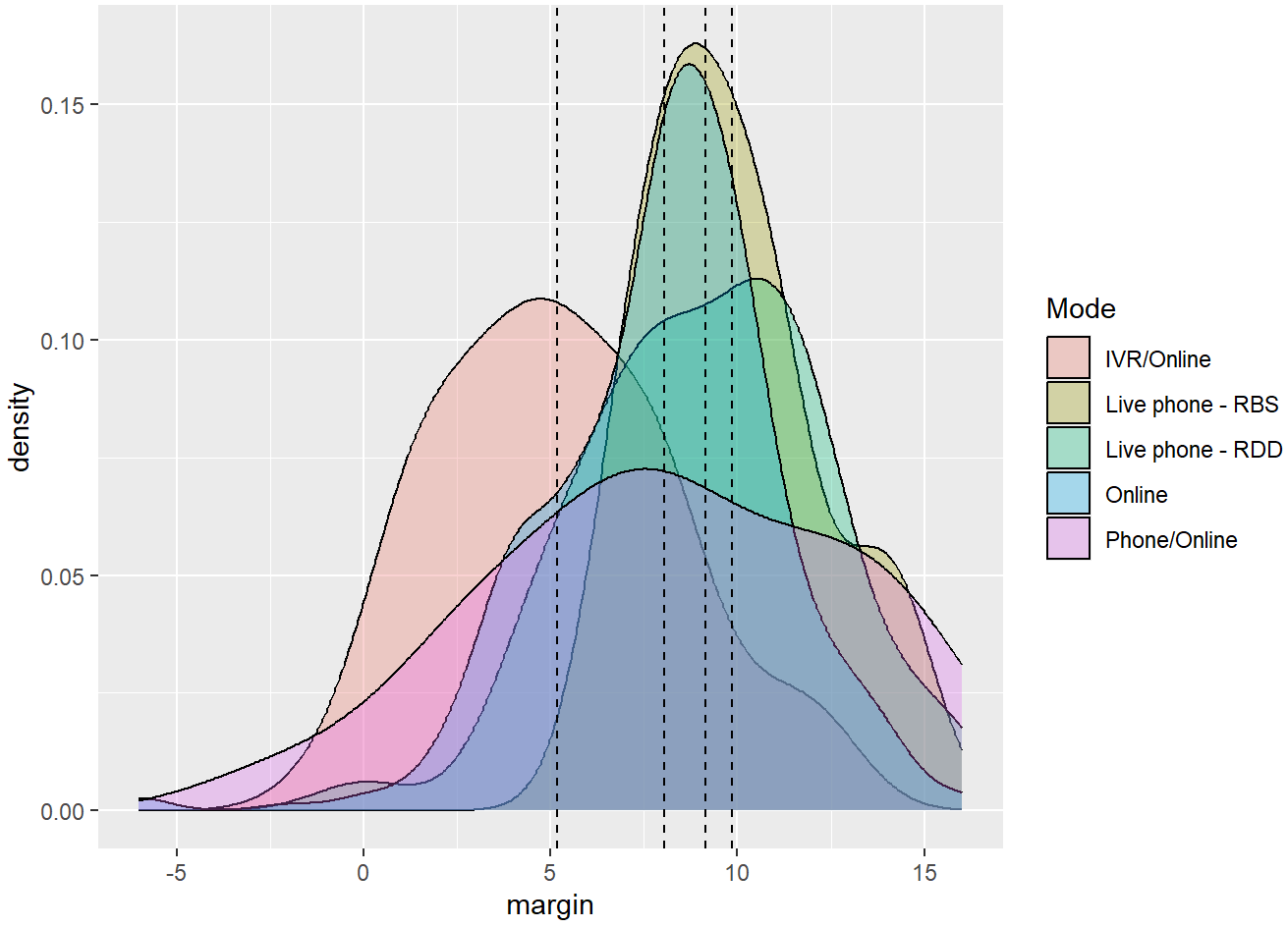


```
# Ugly! Try density instead
toplot <- pres %>%
  filter(Mode == "IVR/Online" | Mode == "Live phone - RDD")
toplot %>%
  ggplot(aes(x = margin,fill = Mode)) +
  geom_density(alpha = .3) +
  geom_vline(data = toplot %>%
               group_by(Mode) %>%
               summarise(margin = mean(margin)),
             aes(xintercept = margin),
             linetype = 'dashed')
```



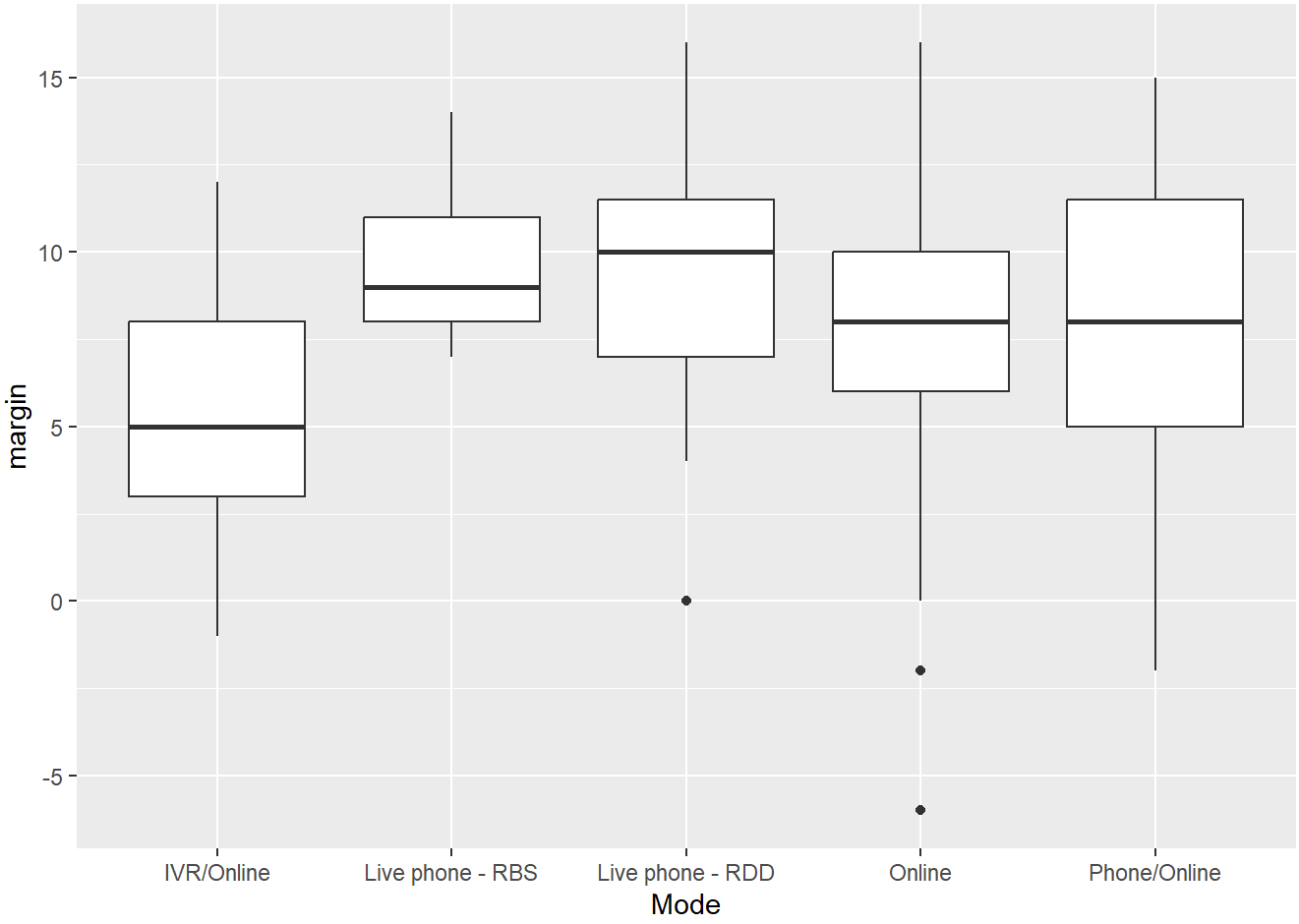
More modes

```
toplot <- pres %>%
  filter(Mode %in% c('IVR/Online','Live phone - RBS',
                     'Live phone - RDD',
                     'Online','Phone/Online'))

toplot %>%
  ggplot(aes(x = margin,fill = Mode)) +
  geom_density(alpha = .3) +
  geom_vline(data = toplot %>%
               group_by(Mode) %>%
               summarise(margin = mean(margin)),
             aes(xintercept = margin),
             linetype = 'dashed')
```



Introducing new geom: geom_boxplot()

```
toplot %>%
  ggplot(aes(x = Mode,y = margin)) +
  geom_boxplot()
```


15
10
5
0
-5
margin
IVR/Online
Live phone - RBS
Live phone - RDD
Online
Phone/Online
Mode