W03 Problem Set

Sami Cemek

Introduction

This week we will use R to explore Hollywood Movies. Our data set contains information on 1295 movies released in Hollywood between 2012 and 2018.

I've modified the data set slightly using skills we will practice next week in order to create additional options when visualizing the data.

```
str(HollywoodMovies)
```

```
## 'data.frame':    1295 obs. of  16 variables:
##  $ Movie           : chr  "2016: Obama's America" "21 Jump Street" "A Late Quartet" "A Royal
Affair" ...
##  $ LeadStudio      : chr  "Rocky Mountain Pictures " "Sony Pictures Releasing " "Entertainmen
t One " "Magnolia Pictures " ...
##  $ RottenTomatoes  : int  26 85 76 90 35 27 91 56 11 44 ...
##  $ AudienceScore   : int  73 82 71 82 51 72 62 47 47 63 ...
##  $ Genre           : chr  "Documentary" "Comedy" "Drama" "Drama" ...
##  $ TheatersOpenWeek: int  1 3121 9 7 3108 3039 132 245 2539 3192 ...
##  $ OpeningWeekend  : num  0.03 36.3 0.08 0.04 16.31 ...
##  $ BOAvgOpenWeekend: int  30000 11631 8889 5714 5248 8055 8636 2857 4490 6739 ...
##  $ Budget          : num  3 42 NA NA 68 12 NA 7.5 35 50 ...
##  $ DomesticGross   : num  33.35 138.45 1.56 1.55 37.52 ...
##  $ WorldGross      : num  33.4 202.8 6.3 7.6 137.5 ...
##  $ ForeignGross    : num  0 64.36 4.74 6.05 99.97 ...
##  $ Profitability   : num  1334 483 NA NA 202 ...
##  $ OpenProfit      : num  1.2 86.4 NA NA 24 ...
##  $ Year            : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
##  $ BudgetSize      : Factor w/ 3 levels "Small","Medium",..: 1 2 NA NA 2 1 NA 1 2 2 ...
```

PS1 Budget

Use an appropriate graph or graphs to show the distribution of budgets for Hollywood movies. Modify the title, labels, colors, bins, and themes on your graph(s) to make them "publication ready" (ie something you'd be willing to share in a presentation or news article).

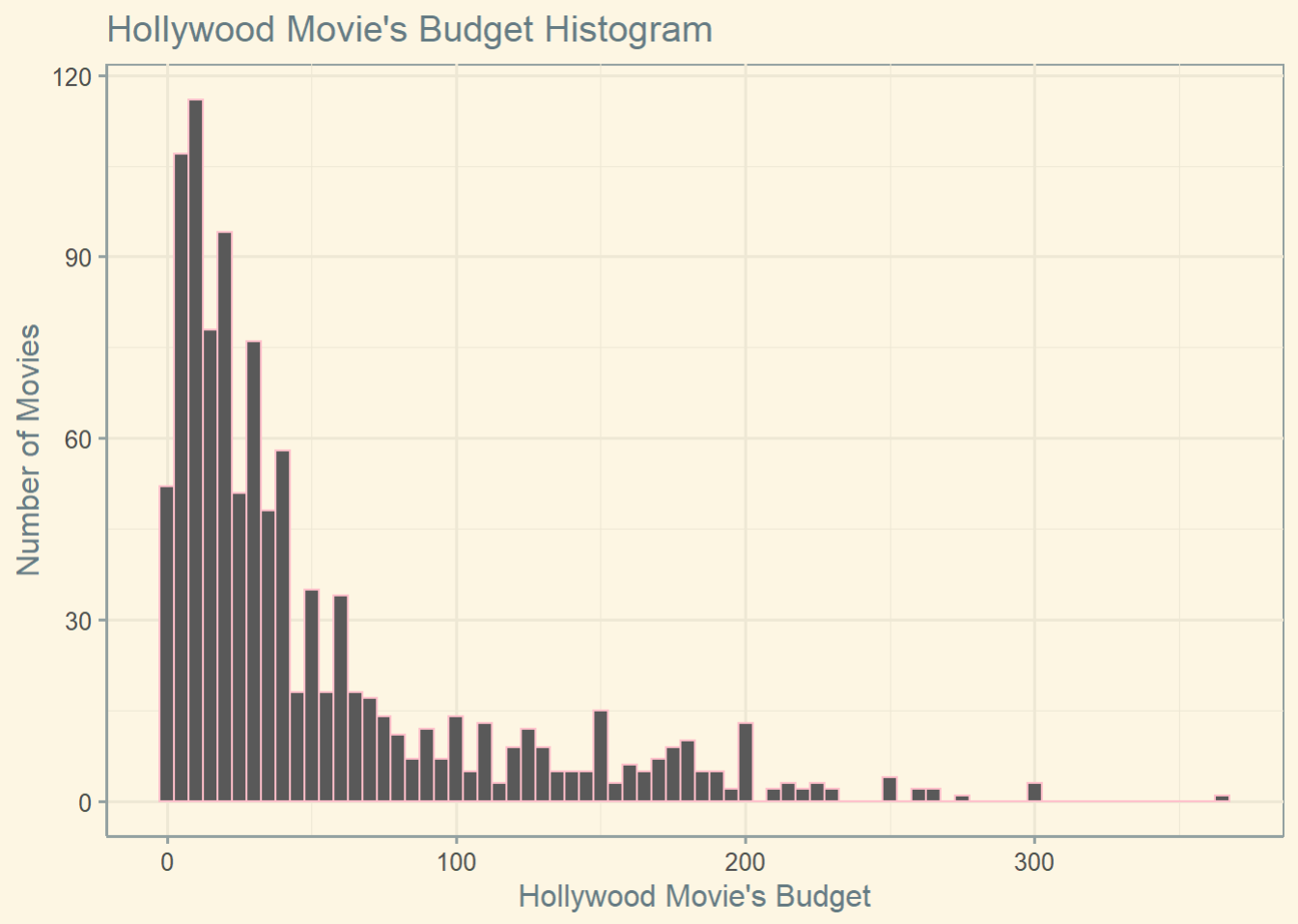
Intermediate/Advanced: annotate the graph to add information about any outliers or unusual features.

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.1.1
```

```
#x axis suppose to be the quantitative variable
ggplot(HollywoodMovies, aes(x=Budget)) + geom_histogram(binwidth = 5, color="pink") + labs(x =
"Hollywood Movie's Budget", y = "Number of Movies", title = "Hollywood Movie's Budget Histogram
 ") + theme_solarized()
```

```
## Warning: Removed 239 rows containing non-finite values (stat_bin).
```



PS2 Budget by Genre

We'd like to explore the distribution of budget by genre, but there are a lot of different genres here, and some genres only have a few movies. Modify the data set to only include the 5 most frequently occuring genres; save this as a new data set called PopularGenres. Then use boxplots to show the distribution of budgets in each genre. Write a couple sentences about what you've found by visualizing the data.

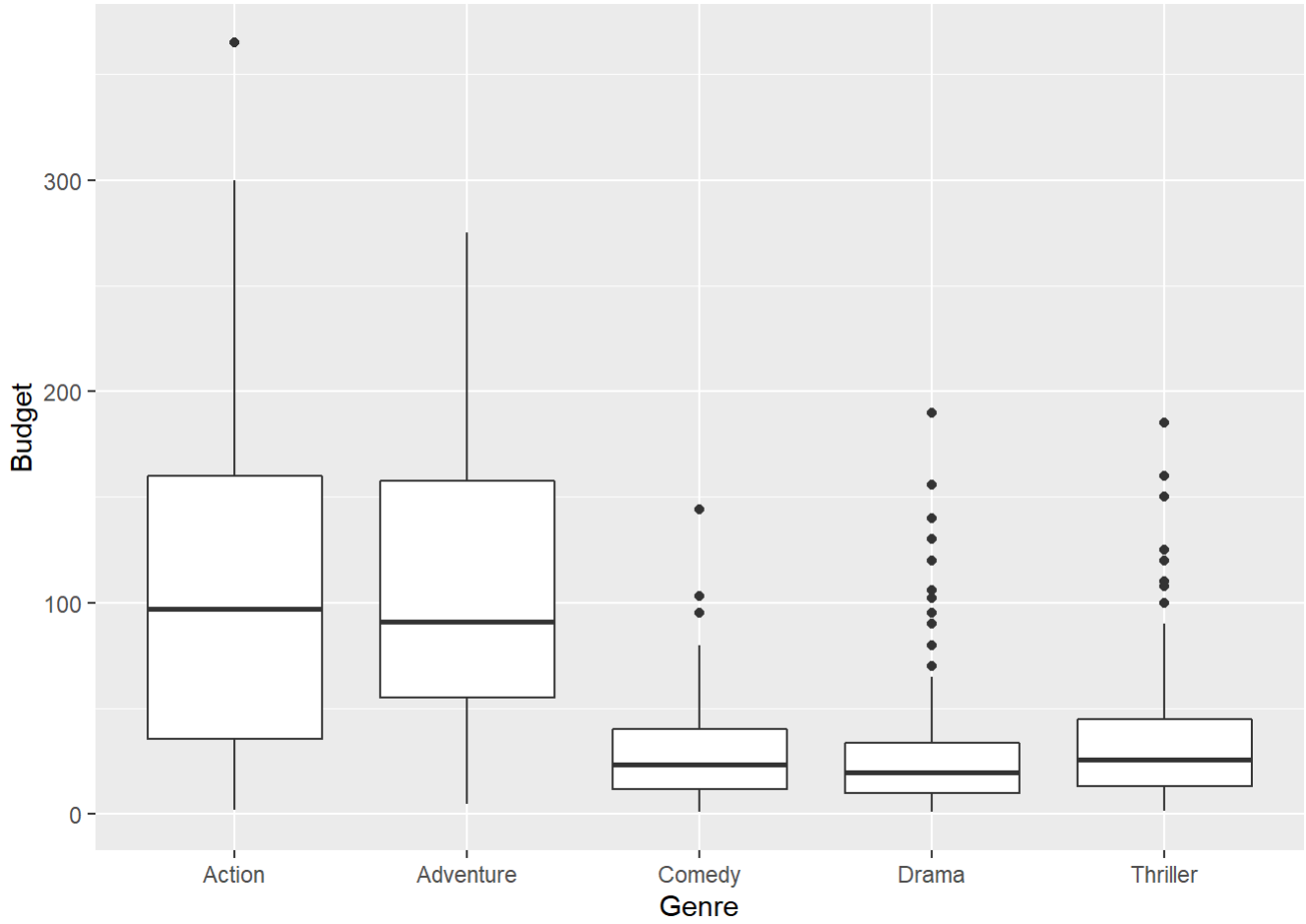
Intermediate/Advanced: Customize the graph theme, colors, labels, etc. and add a red line showing the overall median budget for the PopularGenres data set.

```
#part 1
PopularGenres <- names(sort(table(HollywoodMovies$Genre), decreasing=TRUE)[1:5])
PopularGenres
```

```
## [1] "Drama"     "Comedy"    "Action"    "Adventure" "Thriller"
```

```
#part2
top5genre <- HollywoodMovies %>% filter(Genre %in% PopularGenres)
#part3
ggplot(top5genre, aes(x=Genre, y=Budget)) + geom_boxplot()
```

```
## Warning: Removed 173 rows containing non-finite values (stat_boxplot).
```



Your interpretation: Write a couple sentences about what you've found by visualizing the data.

We found the 5 most frequently occuring genres are: Action,Adventure,Comedy,Drama and Thriller. - The action genre has one outlier. The median is around 100 million. The lower quartile is around 47 and the higher quartile is around 155 million. - The adventure genre has no outlier. The median is around 100 million. The lower quartile is around 50 and the higher quartile is around 155 million. - The comedy genre has three outliers. The median is around 25 million. The lower quartile is around 20 and the higher quartile is around 30 million. - The drama genre has multiple outliers. The median is around 23 million. The lower quartile is around 20 and the higher quartile is around 28 million. - The thriller genre has multiple outliers. The median is around 25 million. The lower quartile is around 20 and the higher quartile is around 45 million. We can see that Action and Adventure genres have more budget most of the times than other three genres. Comedy, Drama, and Thriller have more outliers than other two genres.

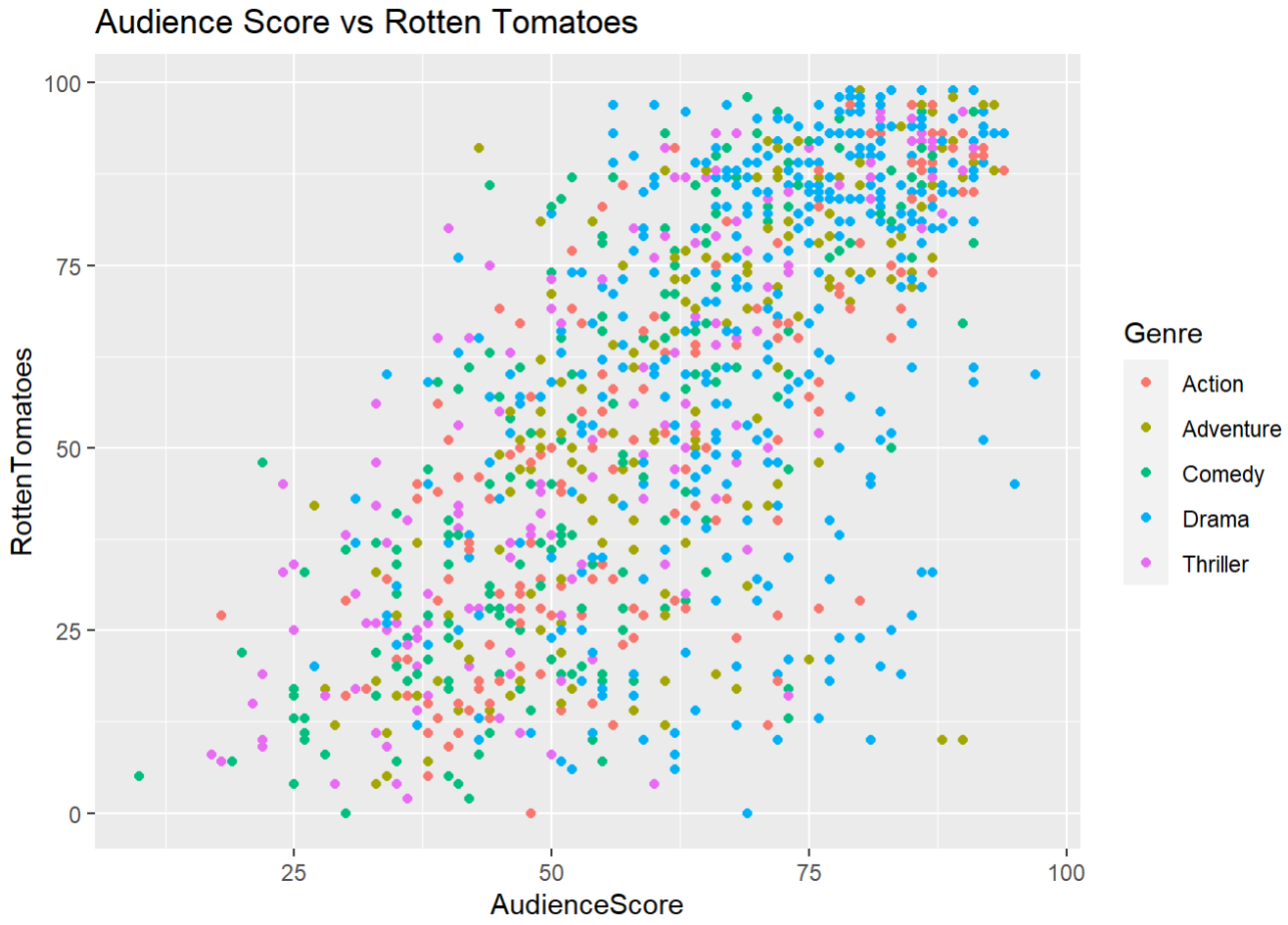
PS3 Movie Popularity

Do movies with higher critic reviews also get higher ratings from audiences? Create a scatterplot using the PopularGenres data to show the association between Rotten Tomatoes critic rating and the Audience rating, paying attention to how genre may (or may not) affect the results.

Intermediate: Explore additional options and graphs to see how *both* genre and budget may be linked to movie profitability, as well as other variables like worldwide gross. You may need to create multiple visualizations to see if there are any patterns, but focus on 2 visualizations to highlight what you found or didn't find.

```
ggplot(top5genre, aes(x=AudienceScore, y=RottenTomatoes, color = Genre)) + labs(title = "Audienc
e Score vs Rotten Tomatoes") + geom_point()
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```




```
#Since we are using PopularGenres data, we will see the distribution for the top 5 genres.
```

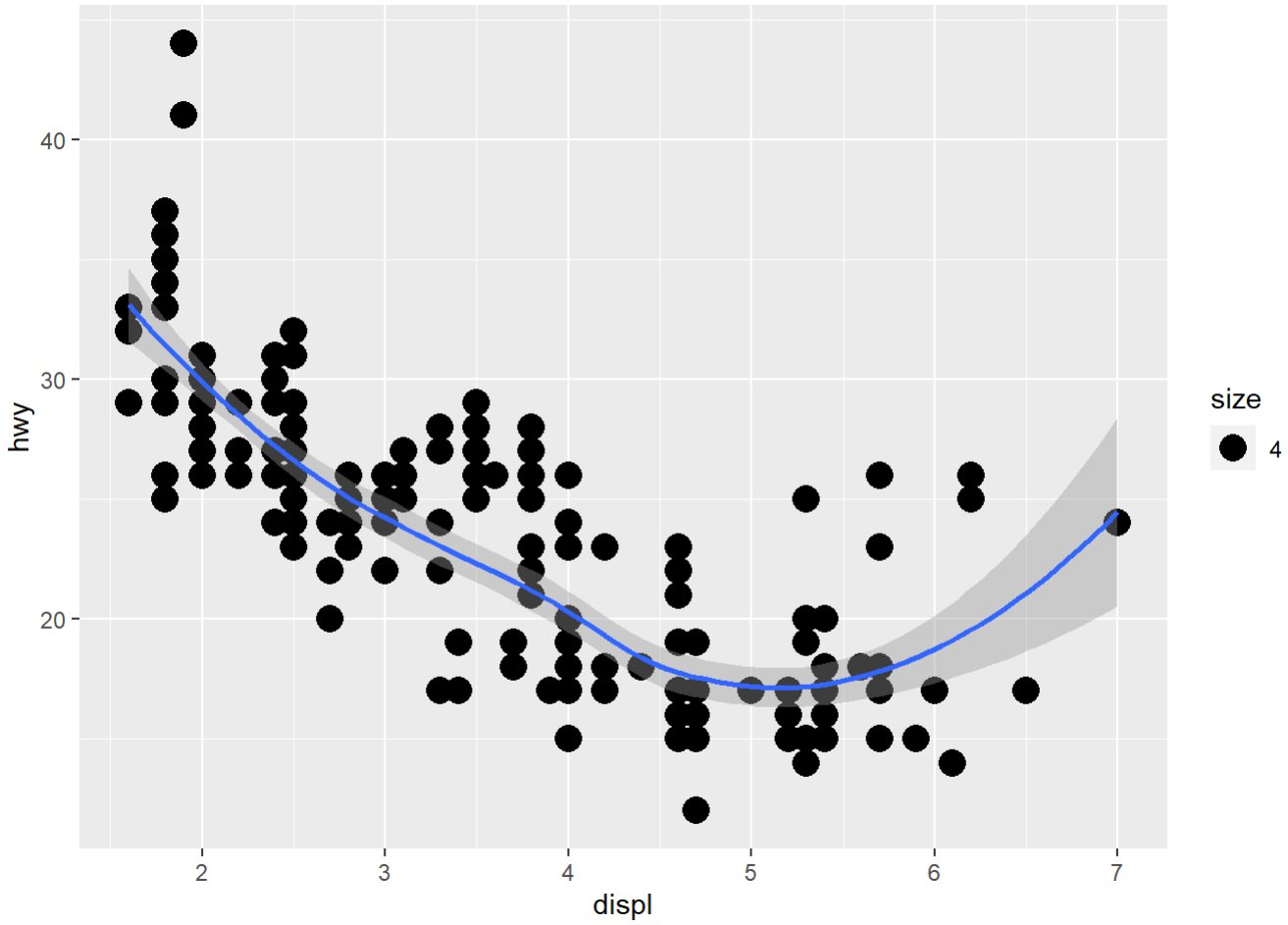
PS4 R4DS textbook exercise

Complete Exercise 6 in R4DS Section 3.6 (https://r4ds.had.co.nz/data-visualisation.html#exercises-3).

Show your code to create each of the 6 graphs here:

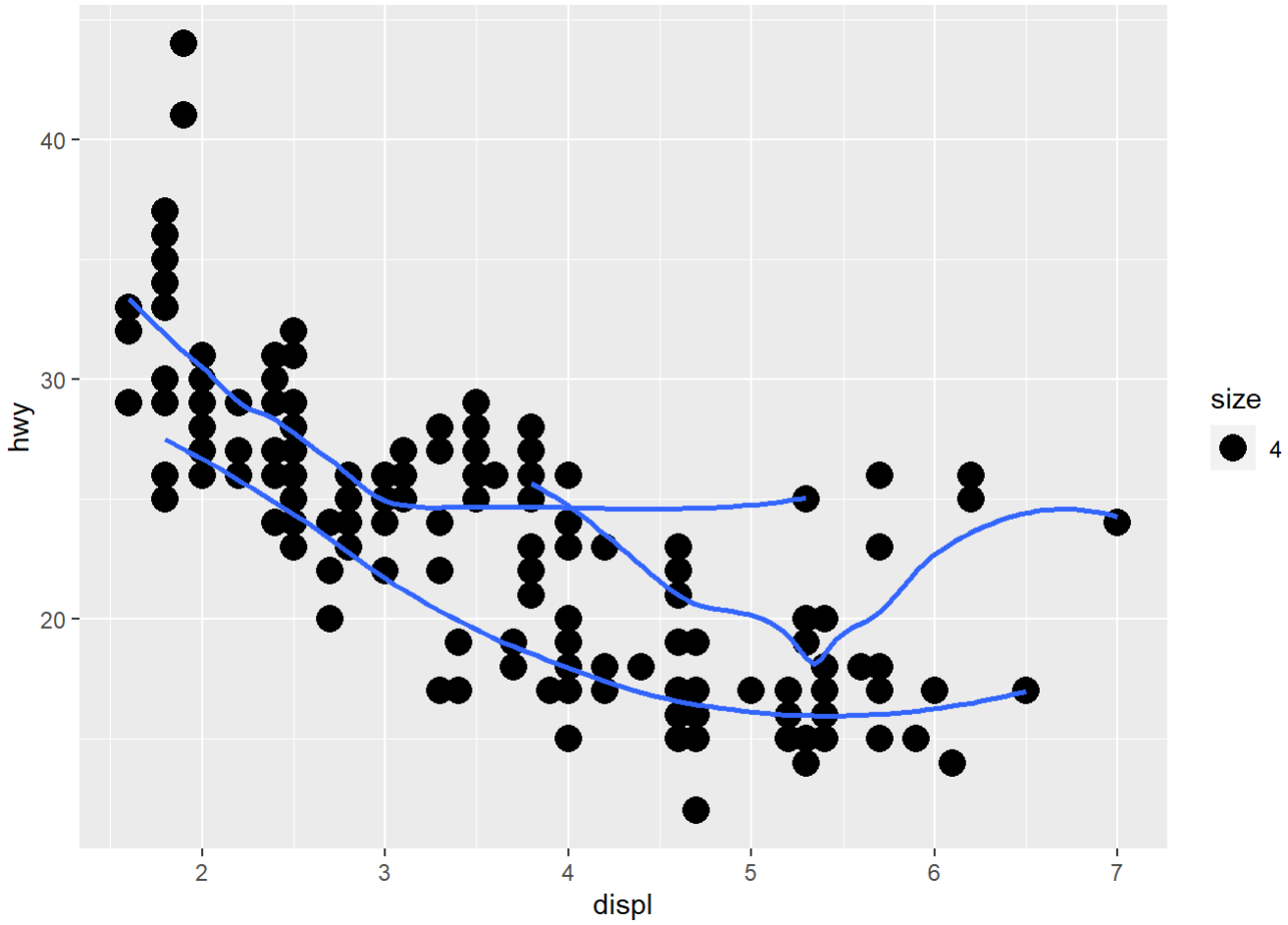
```
# Top Left
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = 4)) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

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



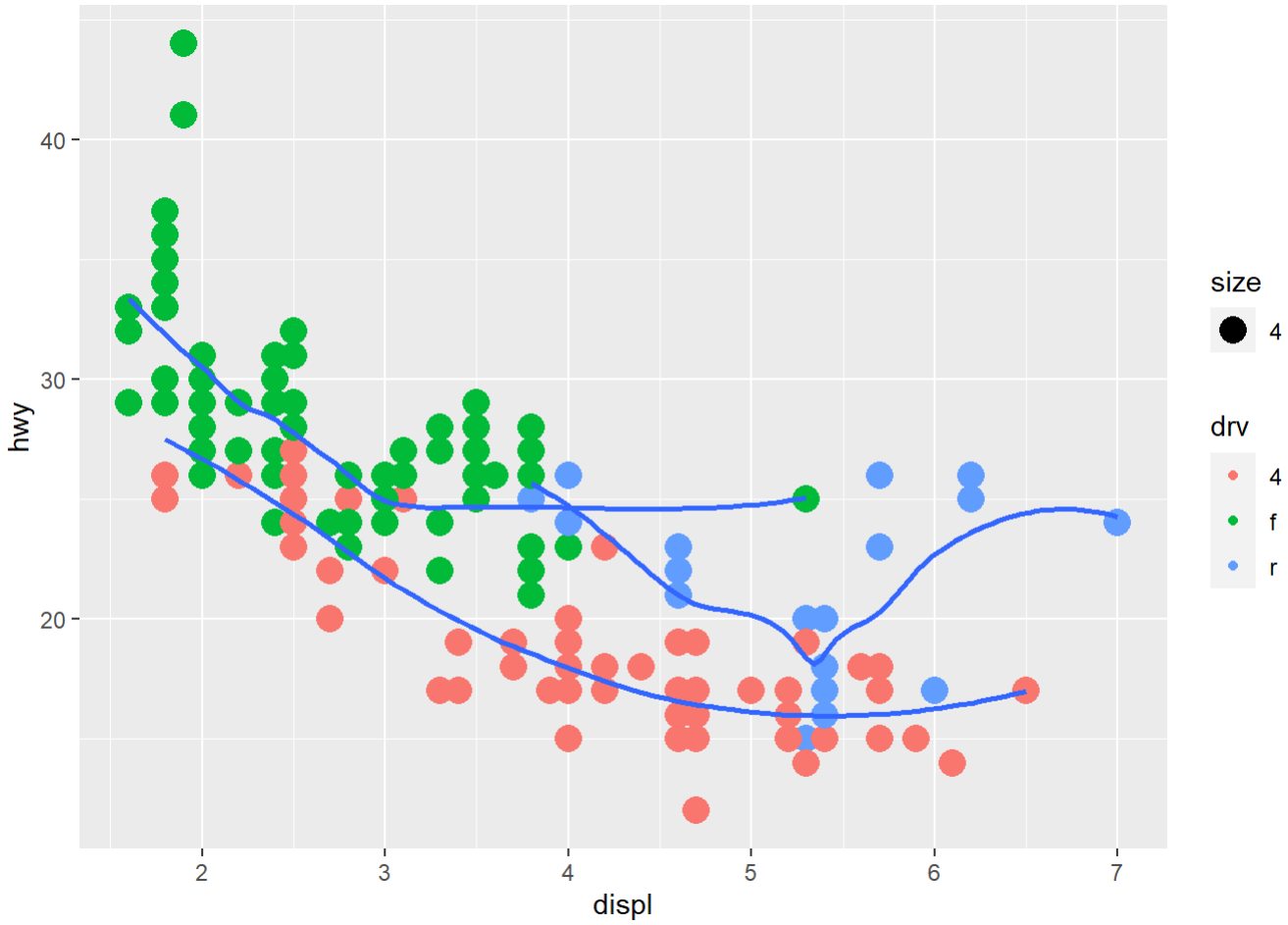

```
# Top right
ggplot() +
  geom_point(data = mpg, mapping = aes(x = displ, y = hwy, size = 4)) +
  geom_smooth(data = mpg, mapping = aes(x = displ, y = hwy, fill = drv),show.legend = FALSE , se
= FALSE)
```

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



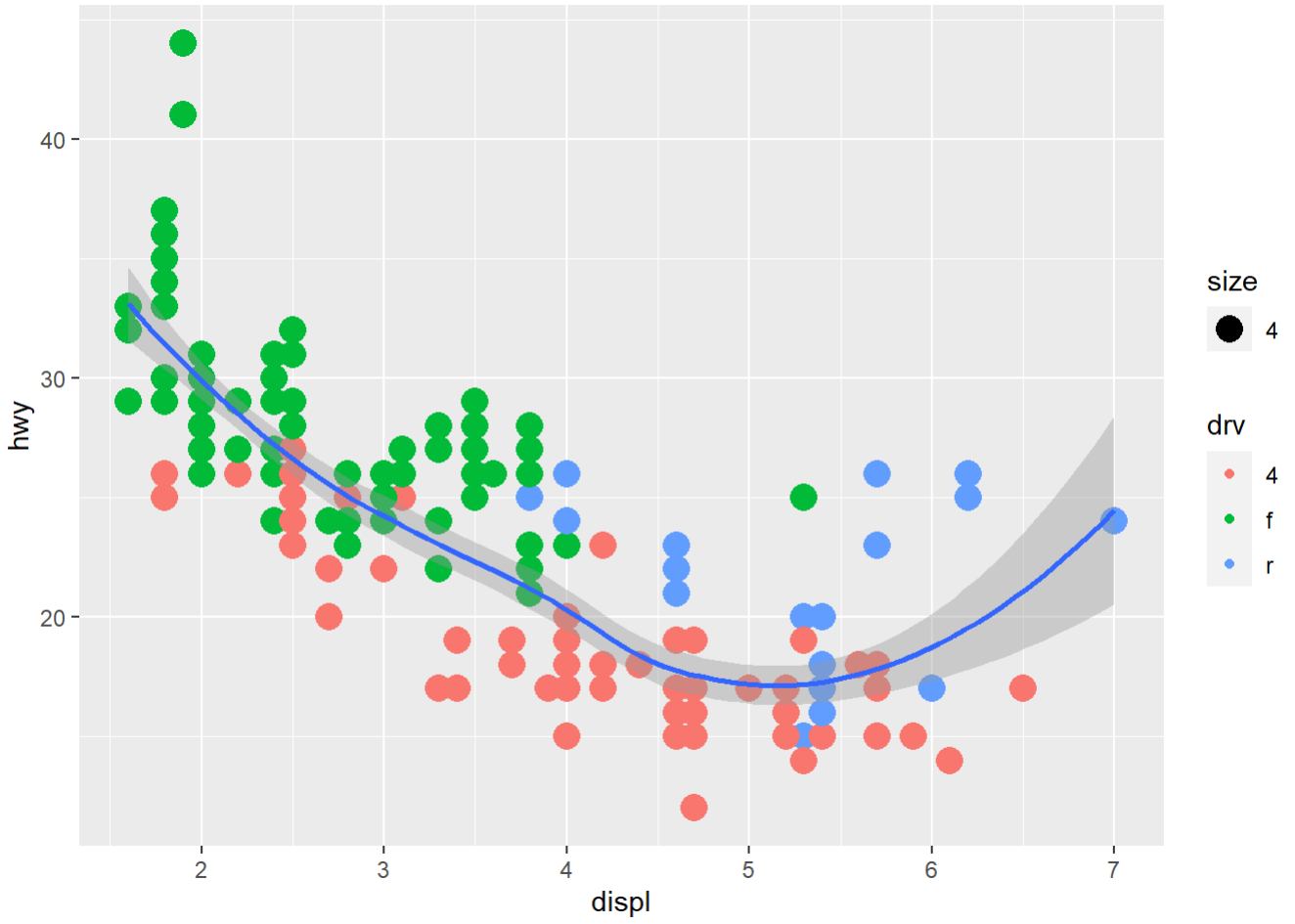

```
# Middle Left'
ggplot() +
  geom_point(data = mpg, mapping = aes(x = displ, y = hwy, size = 4, color = drv)) +
  geom_smooth(data = mpg, mapping = aes(x = displ, y = hwy, fill = drv),show.legend = FALSE , se
= FALSE)
```

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



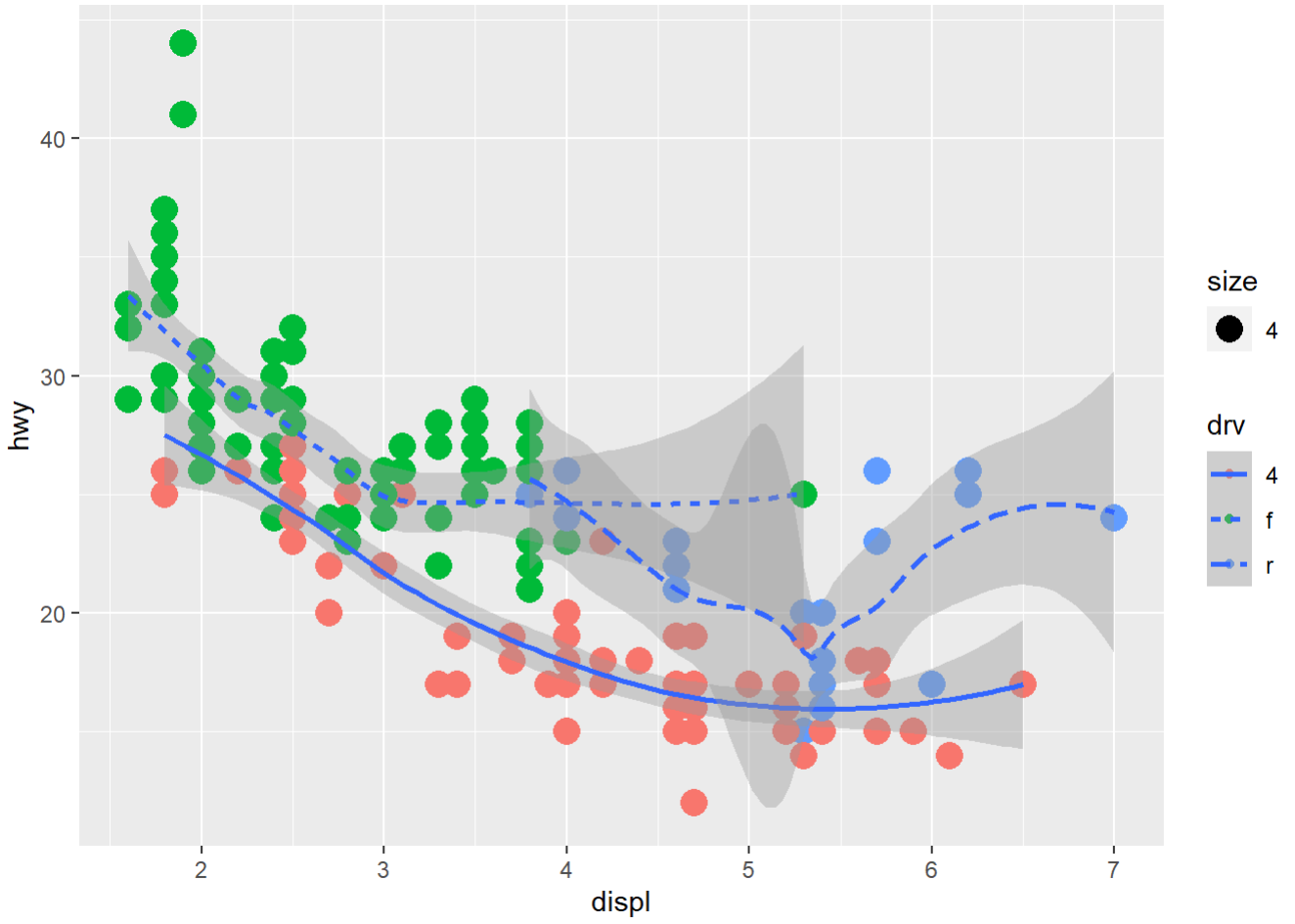

```
# Middle right
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = 4, color = drv)) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

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



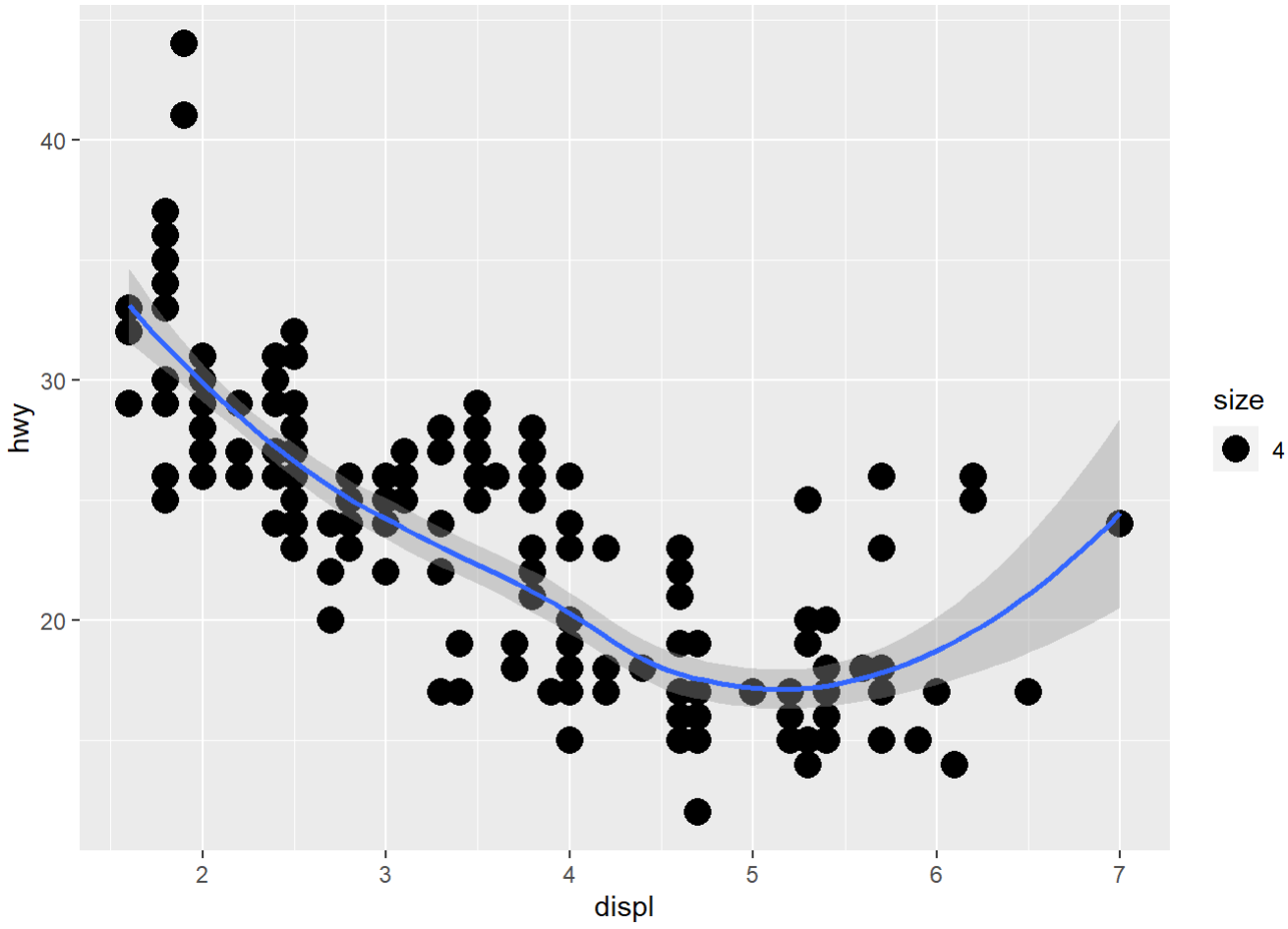

```
# Bottom left
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = 4, color = drv)) +
  geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv))
```

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




```
# Bottom right
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = 4)) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

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


40
30
20
hwy
2
3
4
5
6
7
displ
size
4