R Markdown Exercises

Dustin Pluta

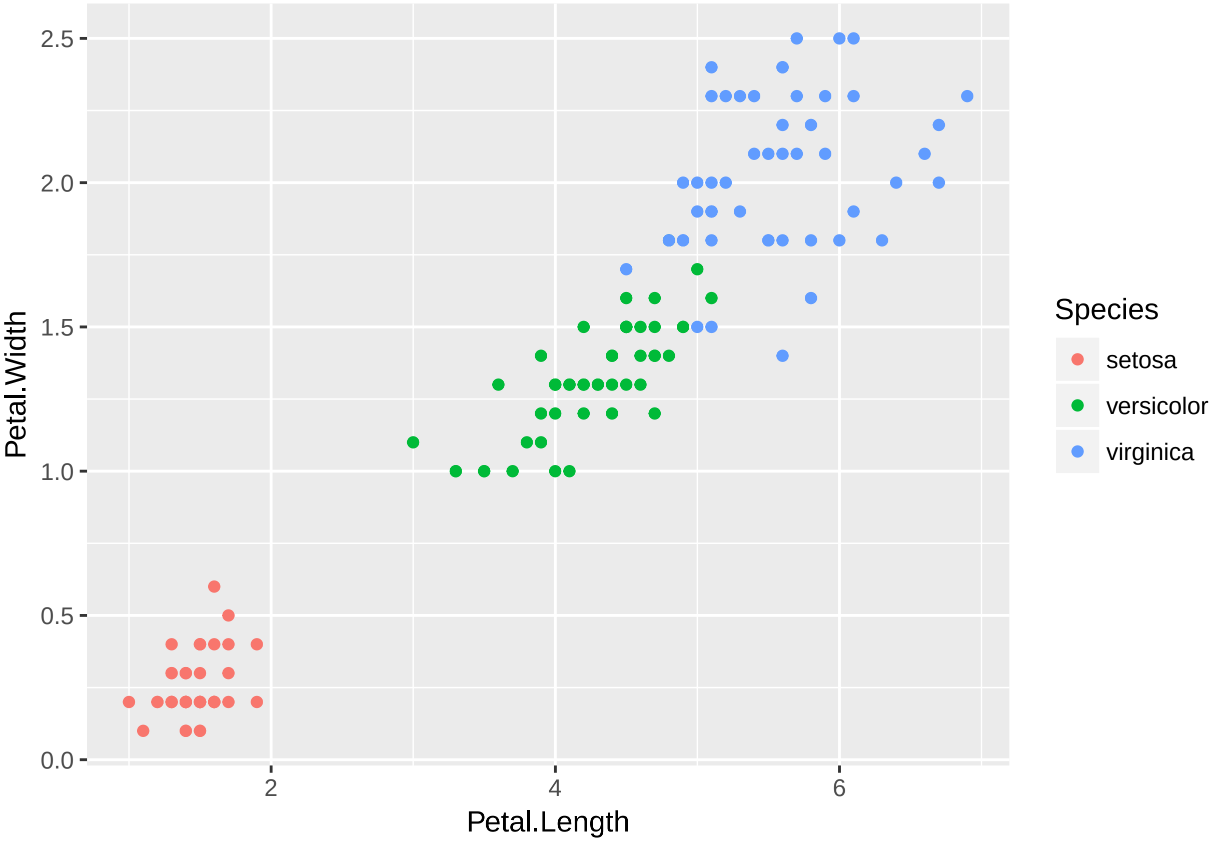
February 20, 2017

Exercises

Edit this R Markdown document to include your solutions. Experiment with formatting and the many available options.

Iris Data

1. Use `ggplot` to display the scatterplot the `iris` data with x-axis `Petal.Length` and y-axis `Petal.Width`. Color the points according to the species.



Solution

***

2.

- Add a *horizontal* **rule** between this problem and the previous problem using “***" on its own line, with newlines above and below.
- Add a link to the iris data set on the UCI Machine learning repository and a brief description of the data.

Iris Data on the UCI ML Repo

---

3. Use `dplyr` to select the petal length and petal width of only the Virginica species, and output this data as a table.

IMDB Data

4. Read in the IMDB Data and print the column names.

```
imdb <- read.csv("~/Data/movie_metadata.csv")
```

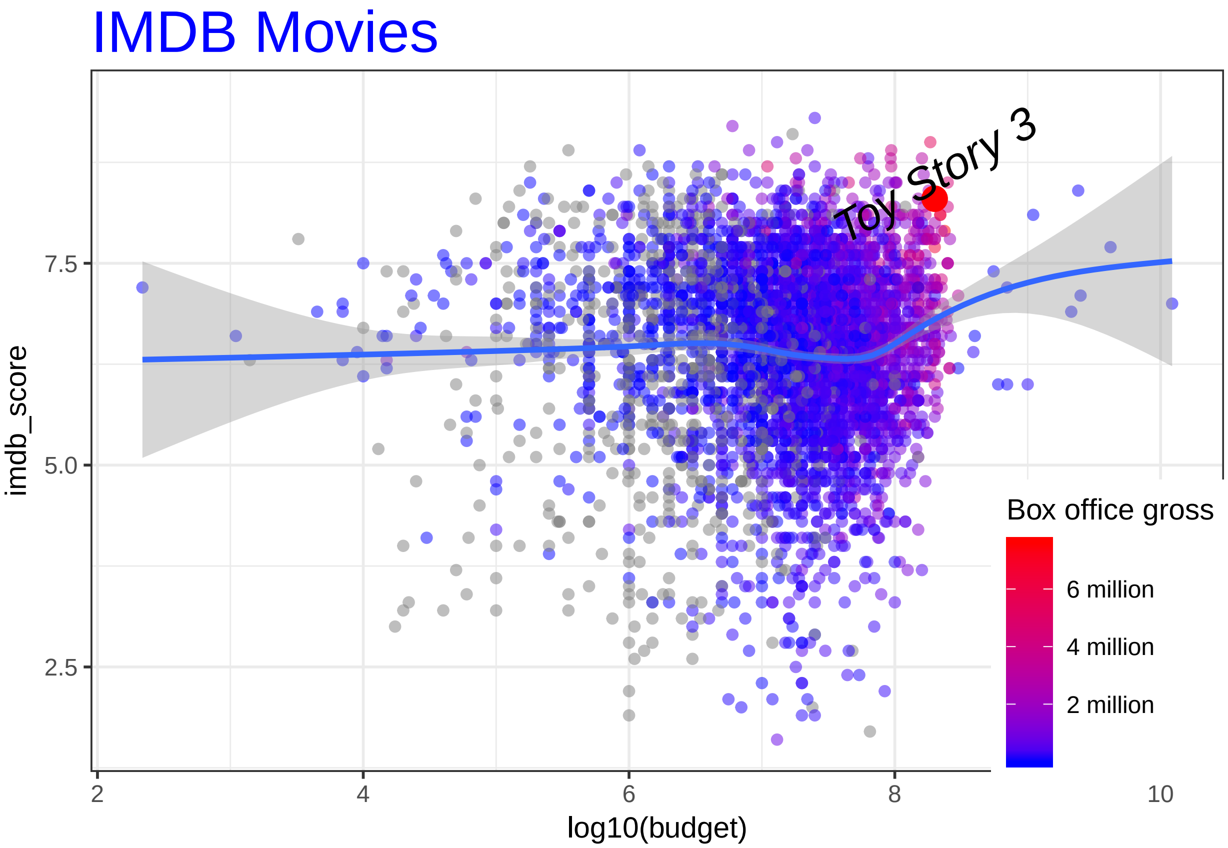
```r
imdb <- read.csv("Data/movie_metadata.csv")
colnames(imdb)
```

```
##  [1] "color"                     "director_name"
##  [3] "num_critic_for_reviews"    "duration"
##  [5] "director_facebook_likes"   "actor_3_facebook_likes"
##  [7] "actor_2_name"              "actor_1_facebook_likes"
##  [9] "gross"                     "genres"
## [11] "actor_1_name"              "movie_title"
## [13] "num_voted_users"           "cast_total_facebook_likes"
## [15] "actor_3_name"              "facenumber_in_poster"
## [17] "plot_keywords"             "movie_imdb_link"
## [19] "num_user_for_reviews"      "language"
## [21] "country"                   "content_rating"
## [23] "budget"                    "title_year"
## [25] "actor_2_facebook_likes"    "imdb_score"
## [27] "aspect_ratio"              "movie_facebook_likes"
```

5.

- Use `dplyr` to create a new data frame called `imdb_selected` with just movie score and $\log_{10}$(budget).
- Create a scatterplot of this data using `ggplot`

```r
library(ggplot2)
ggplot(data=imdb, aes(x=log10(budget), y=imdb_score)) +
  geom_point(aes(colour=gross), alpha=0.5) +
  geom_smooth() +
  scale_color_continuous(name='Box office gross', breaks = c(2e+8, 4e+8, 6e+8),
  labels = c('2 million', '4 million', '6 million'),
  low = 'blue', high = 'red') +
  annotate('point', x=8.3, y=8.3, colour='red', size=4) +
  annotate('text', x=8.3, y=8.6, label='Toy Story 3', fontface='italic', size=6, angle=30) +
  theme_bw() +
  labs(title='IMDB Movies') +
  theme(plot.title=element_text(size=rel(2), colour='blue')) +
  theme(legend.position=c(0.9, 0.2))
```



6. Use `dplyr` to find the average budgets of Nicolas Cage, Leonardo DiCaprio, and Bruce Willis movies and display these values in a table. Hint: `%in%` is useful for filtering on multiple strings.

```
library(dplyr)
avg_budgets <- filter(imdb, actor_1_name %in% c('Nicolas Cage', 'Bruce Willis')) %>%
  select(c(budget, actor_1_name)) %>%
  group_by(actor_1_name) %>%
  summarize(mean(budget))
knitr::kable(avg_budgets)
```

actor_1_name	mean(budget)
Bruce Willis	56066667
Nicolas Cage	51752121