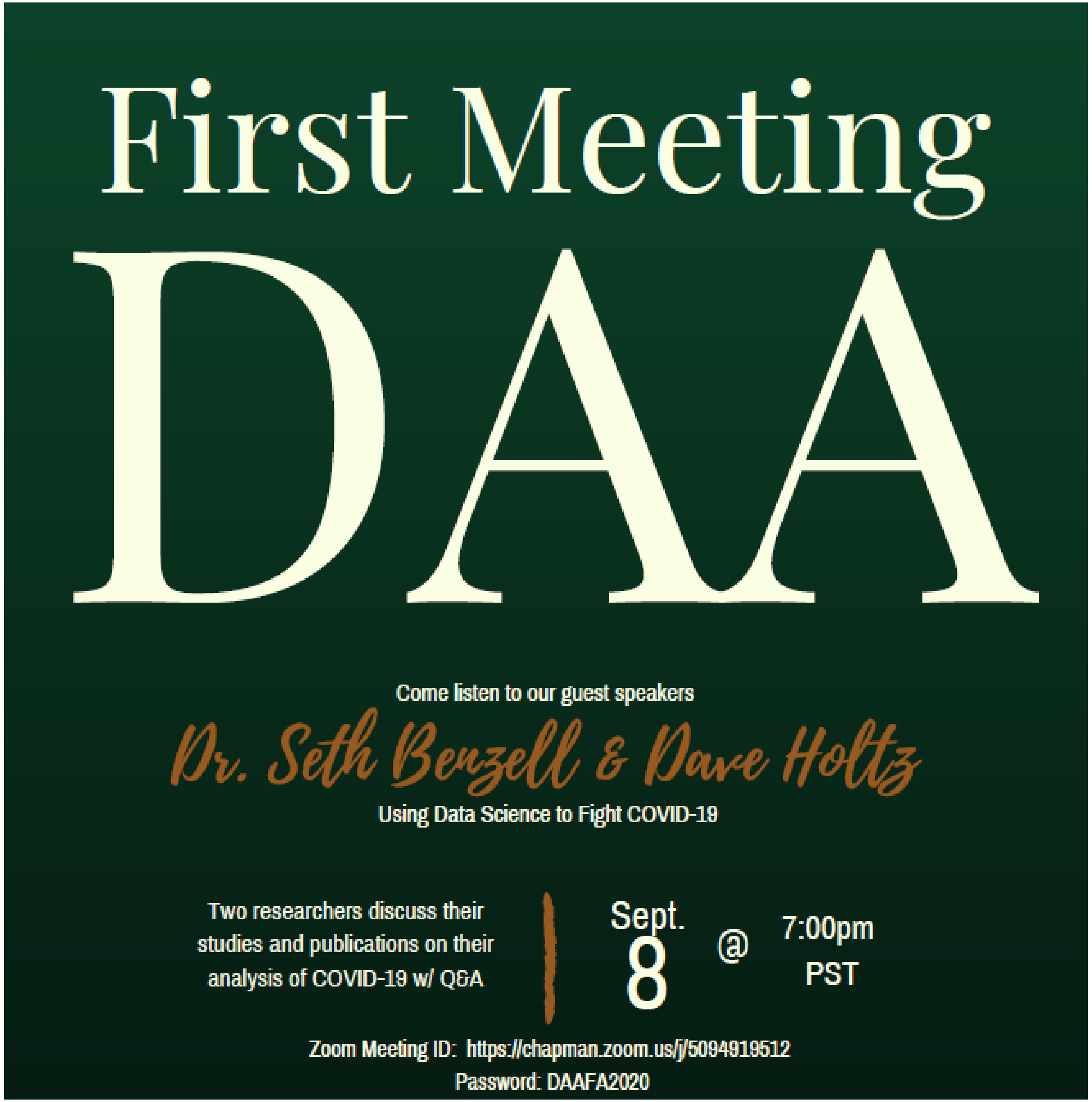
Class 2: Rmarkdown and Data Analysis in Dplyr

BUS 696

Prof. Jonathan Hersh

Class 2: Announcements

1. Data Analytics Club Meeting Yesterday
2. Data Analytics Week! October 5-9
3. Problem Set 1 posted tomorrow – Due Sept 23
   - Must submit compiled HTML file using RMarkdown



First Meeting

DAA

Come listen to our guest speakers

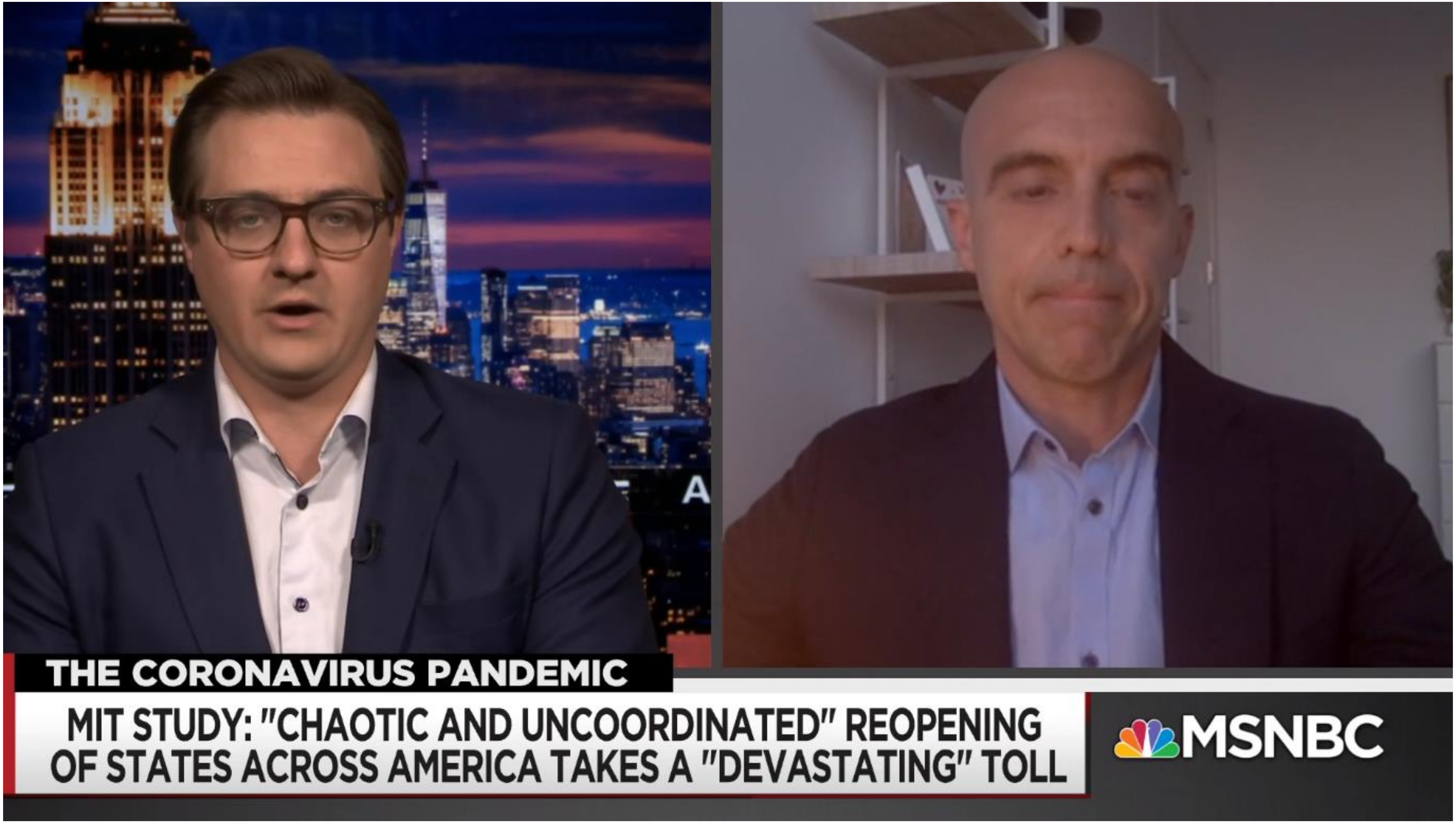
Dr. Seth Benzell & Dave Holtz

Using Data Science to Fight COVID-19

Two researchers discuss their studies and publications on their analysis of COVID-19 w/ Q&A

Sept. 8 @ 7:00pm PST

Zoom Meeting ID: https://chapman.zoom.us/j/5094919512
Password: DAAFA2020



THE CORONAVIRUS PANDEMIC

MIT STUDY: "CHAOTIC AND UNCOORDINATED" REOPENING OF STATES ACROSS AMERICA TAKES A "DEVASTATING" TOLL

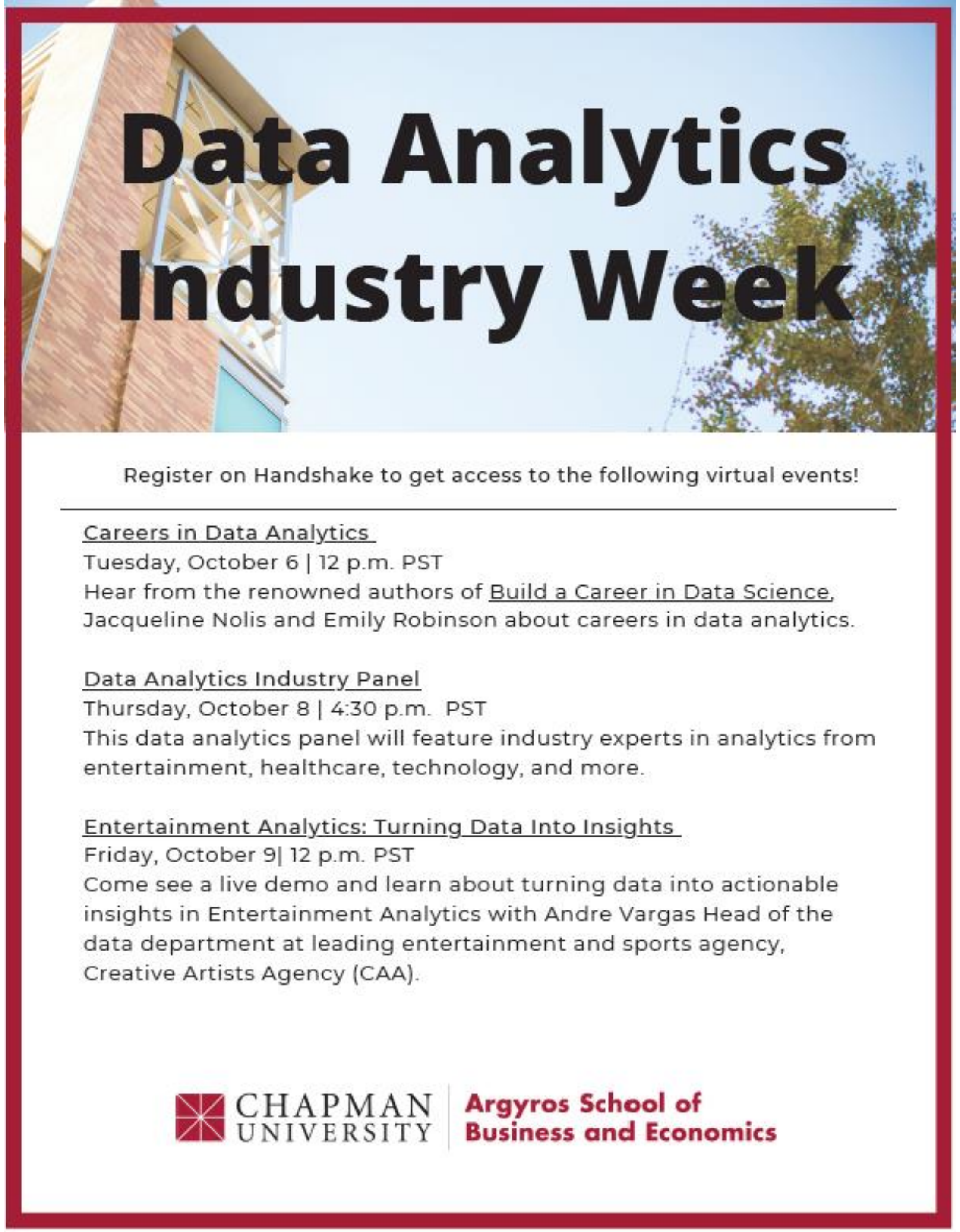
Hello!

Thank you to everyone who came to our first meeting of the semester! We were able to listen to Dr. Seth Benzell and Dave Holtz talk about their work on COVID-19 papers as well as a little about their backgrounds! As requested, We have provided a **link to the meeting recording** for those who were unable to make it. The recording starts right when the speakers were introduced:
[goog_1863276205] https://drive.google.com/file/d/1-2XdN4j-jJS3uHJH8F_sXYcfbc0VDSAE/view?usp=sharing

Lastly, there is **contact form** that I would ask you to give to your friends if they are interested, or if you are *interested in joining the executive team*! Here is the link https://forms.gle/FaoiZzqB5MaGvn1E7

Reminder **our next meeting will be Tuesday, October 13th @ 7pm PDT.** Be on the lookout for emails on internship opportunities or updates till then!

Best,
DAA Executive Team



Data Analytics Industry Week

Register on Handshake to get access to the following virtual events!

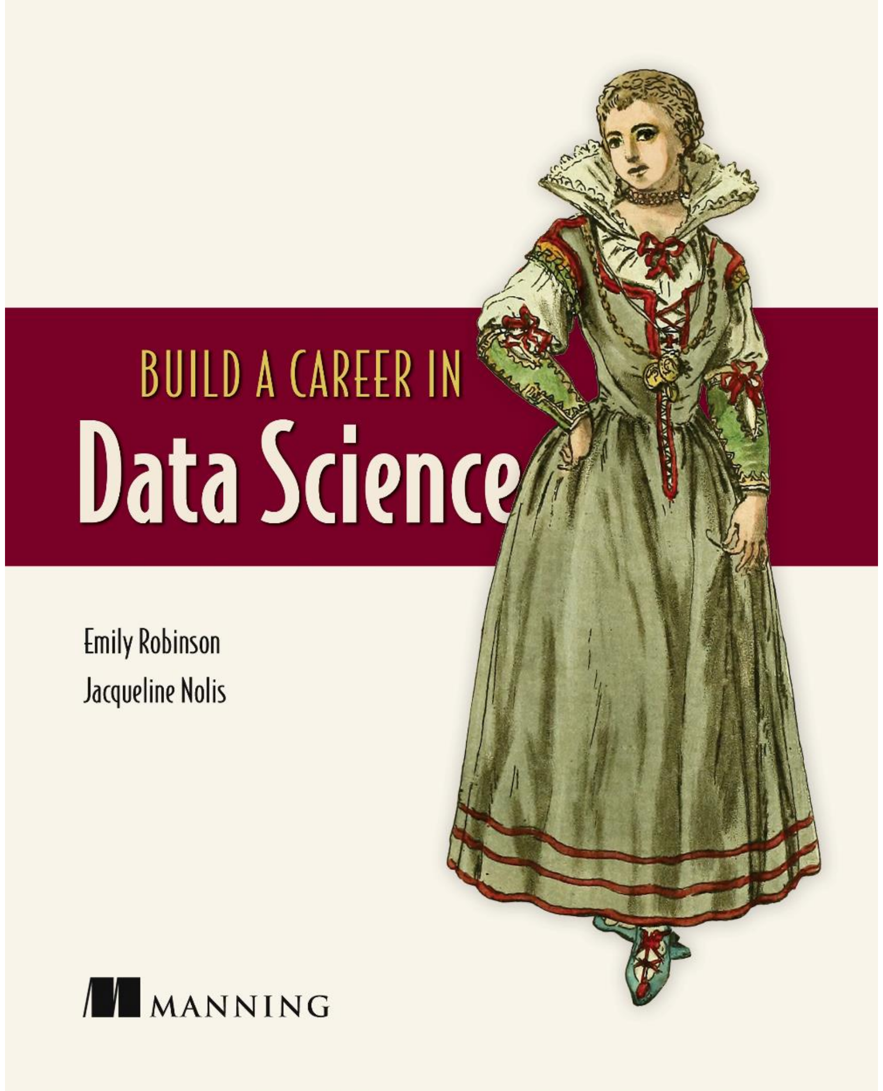
Careers in Data Analytics
Tuesday, October 6 | 12 p.m. PST
Hear from the renowned authors of Build a Career in Data Science, Jacqueline Nolis and Emily Robinson about careers in data analytics.

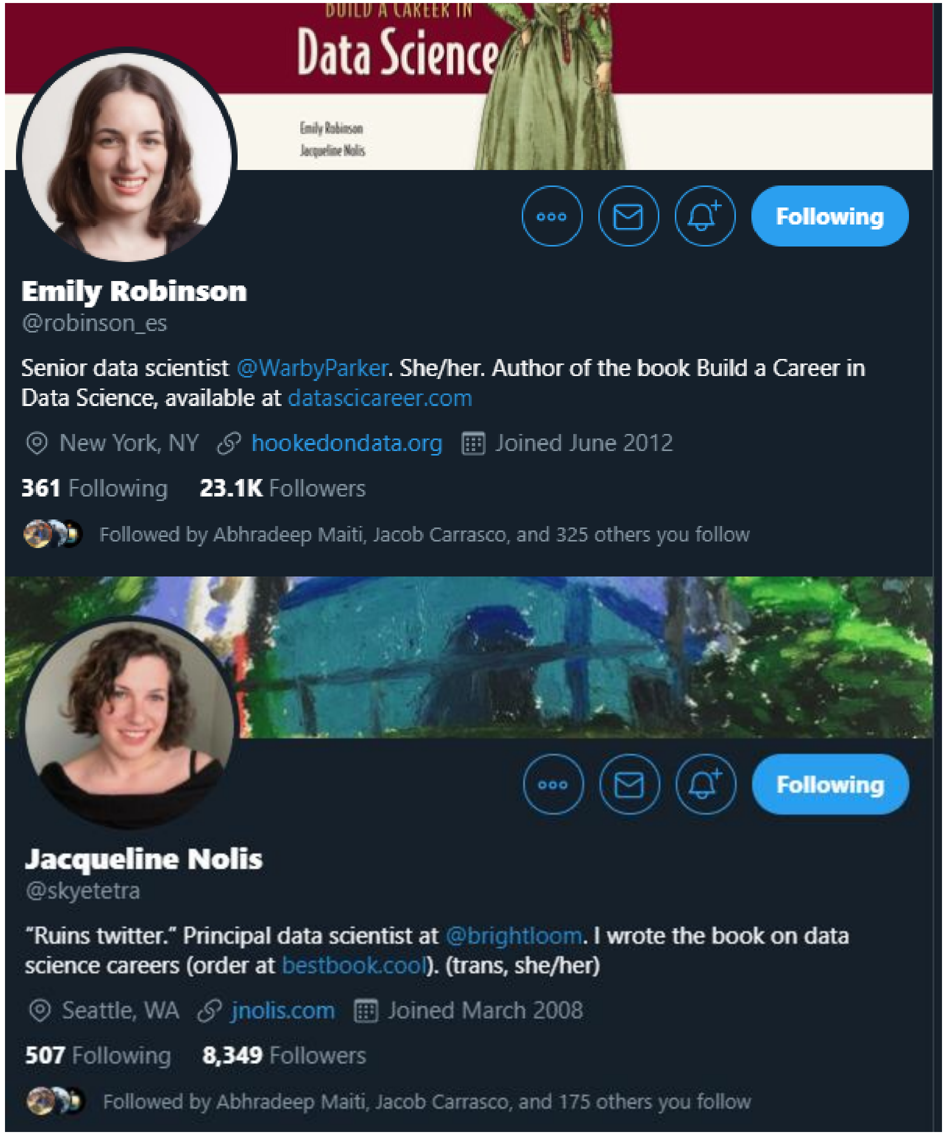
Data Analytics Industry Panel
Thursday, October 8 | 4:30 p.m. PST
This data analytics panel will feature industry experts in analytics from entertainment, healthcare, technology, and more.

Entertainment Analytics: Turning Data Into Insights
Friday, October 9| 12 p.m. PST
Come see a live demo and learn about turning data into actionable insights in Entertainment Analytics with Andre Vargas Head of the data department at leading entertainment and sports agency, Creative Artists Agency (CAA).

CHAPMAN UNIVERSITY | Argyros School of Business and Economics

Emily Robinson and Jacqueline Nolis are Awesome





Emily Robinson
@robinson_es

Senior data scientist @WarbyParker. She/her. Author of the book Build a Career in Data Science, available at datascicareer.com

New York, NY hookedondata.org Joined June 2012

361 Following **23.1K** Followers

Followed by Abhradeep Maiti, Jacob Carrasco, and 325 others you follow

Jacqueline Nolis
@skyetetra

"Ruins twitter." Principal data scientist at @brightloom. I wrote the book on data science careers (order at bestbook.cool). (trans, she/her)

Seattle, WA jnolis.com Joined March 2008

507 Following **8,349** Followers

Followed by Abhradeep Maiti, Jacob Carrasco, and 175 others you follow

Data Analytics Accelerator Program Info Sesh October 5 @ Noon

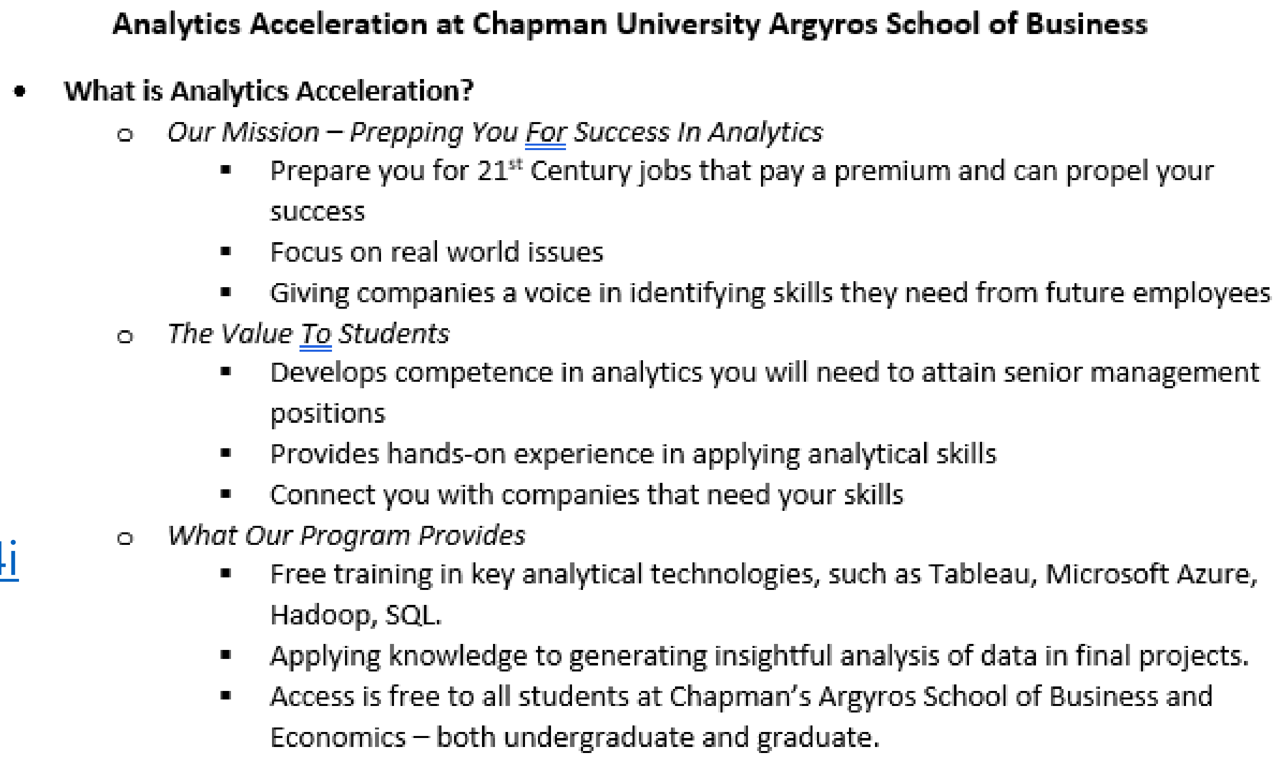
Join Zoom Meeting
https://chapman.zoom.us/j/94331484366

Meeting ID: 943 3148 4366
One tap mobile
+16699006833,,94331484366# US (San Jose)
12532158782,,94331484366# US
+(Tacoma)

Meeting ID: 943 3148 4366
Find your local
number: https://chapman.zoom.us/u/aMMe54i7h

Meeting ID: 943 3148 4366

Join by Skype for Business
https://chapman.zoom.us/skype/94331484366



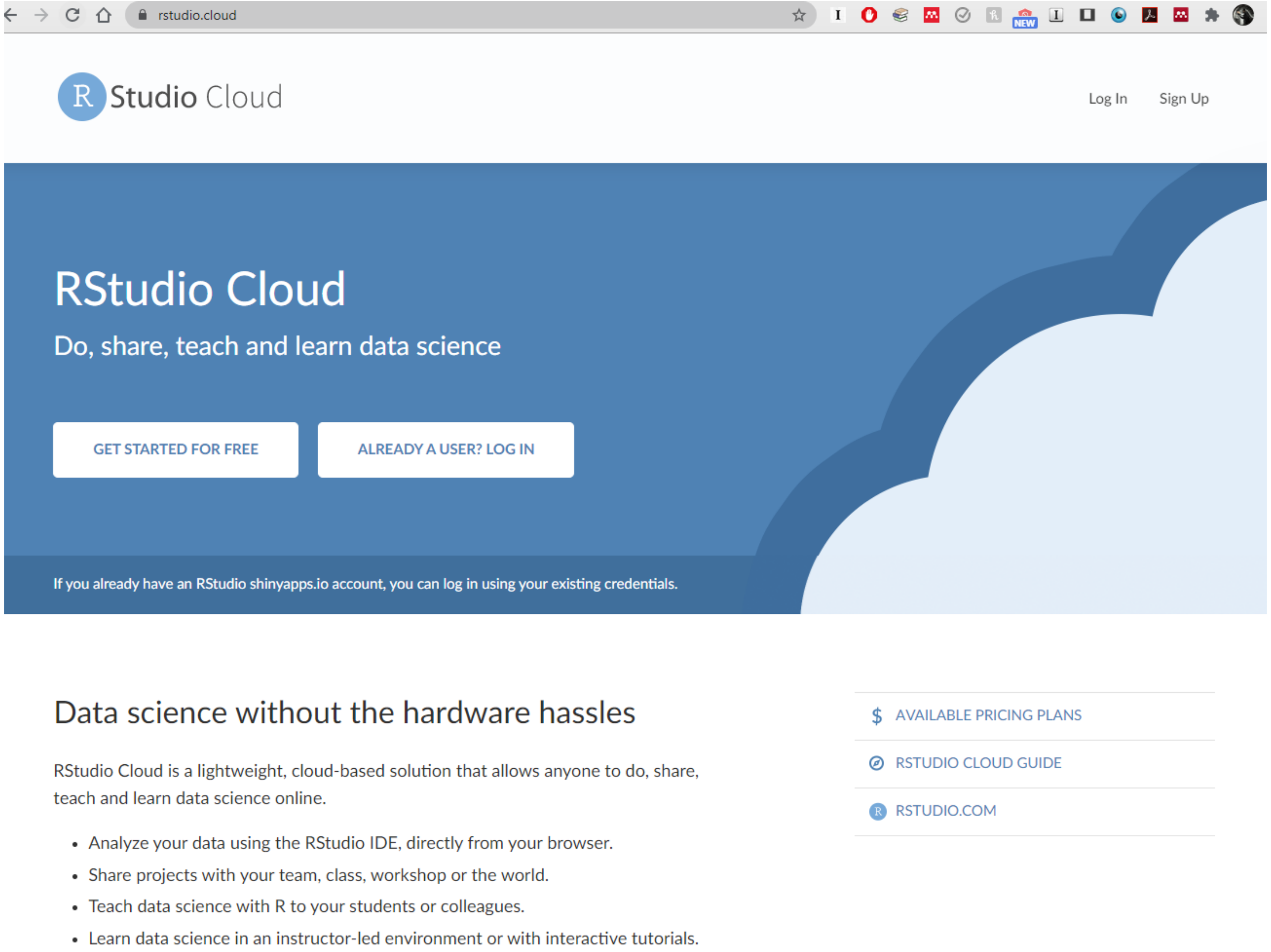
Analytics Acceleration at Chapman University Argyros School of Business

- **What is Analytics Acceleration?**
  - *Our Mission – Prepping You For Success In Analytics*
    - Prepare you for 21st Century jobs that pay a premium and can propel your success
    - Focus on real world issues
    - Giving companies a voice in identifying skills they need from future employees
  - *The Value To Students*
    - Develops competence in analytics you will need to attain senior management positions
    - Provides hands-on experience in applying analytical skills
    - Connect you with companies that need your skills
  - *What Our Program Provides*
    - Free training in key analytical technologies, such as Tableau, Microsoft Azure, Hadoop, SQL.
    - Applying knowledge to generating insightful analysis of data in final projects.
    - Access is free to all students at Chapman's Argyros School of Business and Economics – both undergraduate and graduate.

Class 2: Outline

1. Qs from last week?
2. Rmarkdown Lab
3. **Data Analysis**
   - Loading data
   - Glimpse to view
   - Pipe operator
   - slice() to select rows
   - arrange() to order data frame
   - select() to choose variables
   - rename() to rename variables
   - filter() to select rows matching characteristics
   - Missing values
   - Loops
   - mutate to transform variables
   - Remove duplicates with distinct
   - Outputting “clean” data file”
4. Data Analysis Lab 1
5. **Data Analysis by Groups**
   - group_by() function
   - summarize() to create group variables
6. Data Analysis Lab 2

R Studio Cloud



RStudio Cloud

Do, share, teach and learn data science

If you already have an RStudio shinyapps.io account, you can log in using your existing credentials.

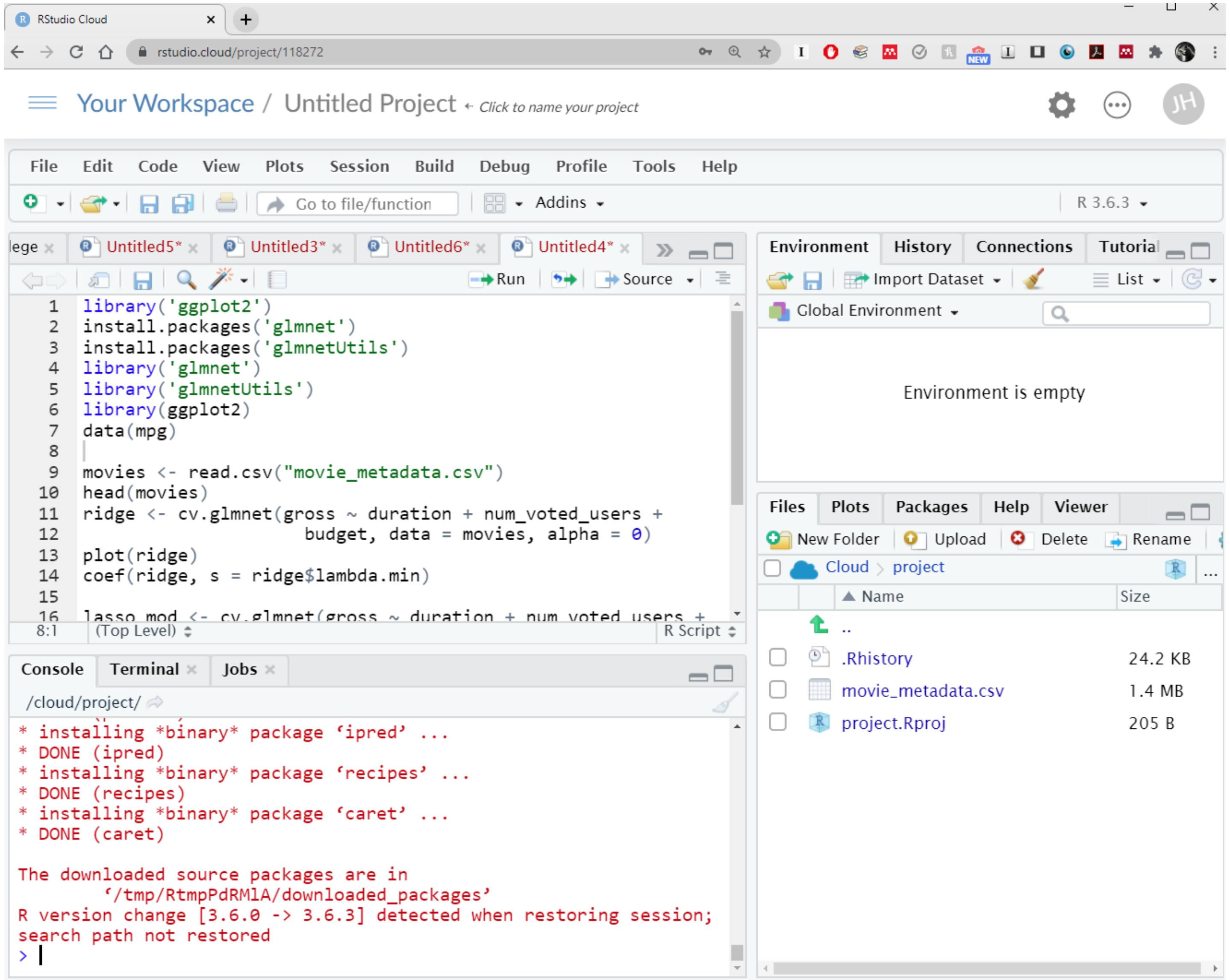
Data science without the hardware hassles

RStudio Cloud is a lightweight, cloud-based solution that allows anyone to do, share, teach and learn data science online.

- Analyze your data using the RStudio IDE, directly from your browser.
- Share projects with your team, class, workshop or the world.
- Teach data science with R to your students or colleagues.
- Learn data science in an instructor-led environment or with interactive tutorials.

- Go to rstudio.cloud if your version of R is ever not working

R Studio Cloud



```
library('ggplot2')
install.packages('glmnet')
install.packages('glmnetUtils')
library('glmnet')
library('glmnetUtils')
library(ggplot2)
data(mpg)

movies <- read.csv("movie_metadata.csv")
head(movies)
ridge <- cv.glmnet(gross ~ duration + num_voted_users +
                   budget, data = movies, alpha = 0)
plot(ridge)
coef(ridge, s = ridge$lambda.min)

lasso_mod <- cv.glmnet(gross ~ duration + num_voted_users +
```

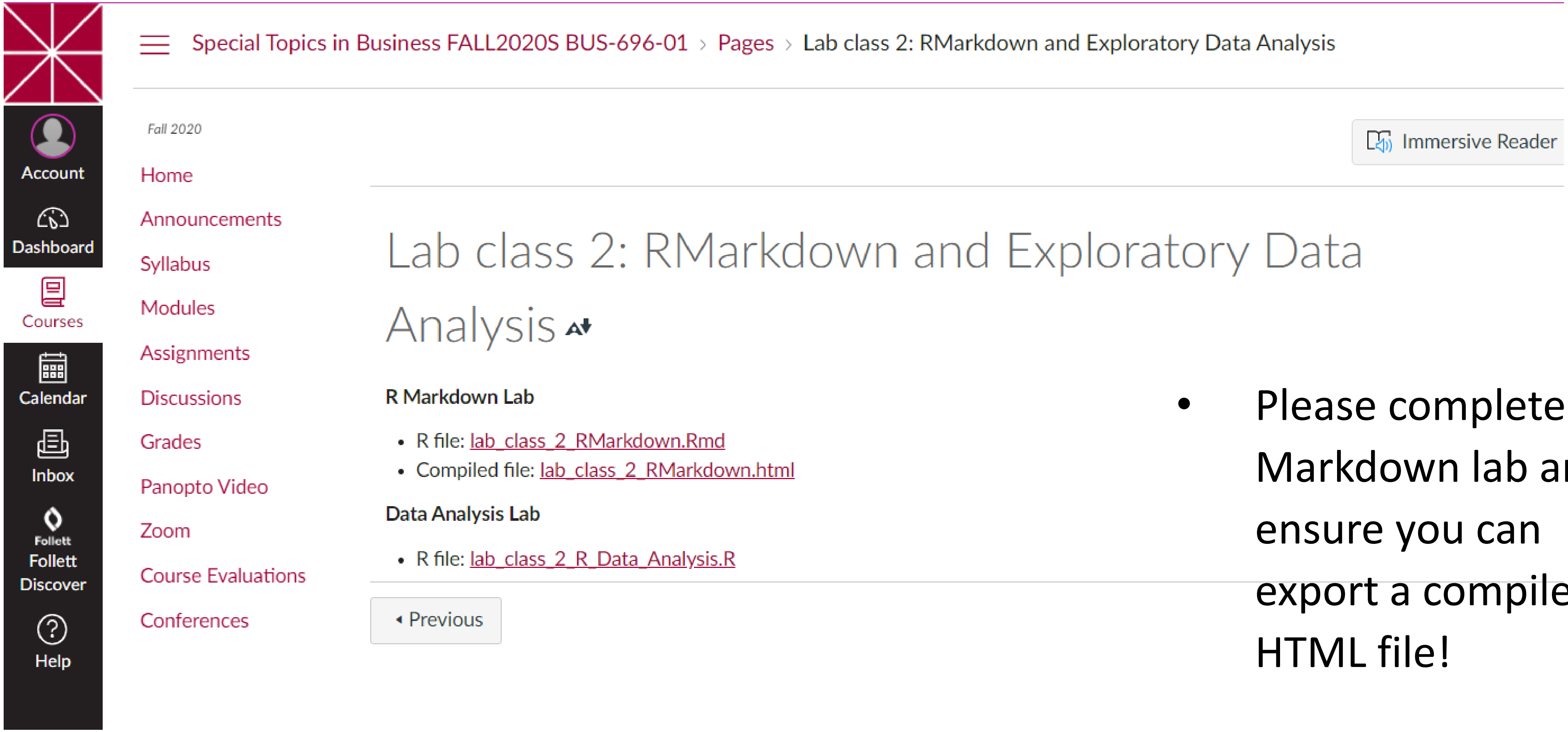
Environment is empty

Name	Size
..	
.Rhistory	24.2 KB
movie_metadata.csv	1.4 MB
project.Rproj	205 B

```
* installing *binary* package ‘ipred’ ...
* DONE (ipred)
* installing *binary* package ‘recipes’ ...
* DONE (recipes)
* installing *binary* package ‘caret’ ...
* DONE (caret)

The downloaded source packages are in
        ‘/tmp/RtmpPdRMlA/downloaded_packages’
R version change [3.6.0 -> 3.6.3] detected when restoring session;
search path not restored
>
```

- R Studio Cloud is a full featured version of R in your browser!

R Markdown & Data Analysis Lab



Special Topics in Business FALL2020S BUS-696-01 › Pages › Lab class 2: RMarkdown and Exploratory Data Analysis

Fall 2020

- Home
- Announcements
- Syllabus
- Modules
- Assignments
- Discussions
- Grades
- Panopto Video
- Zoom
- Course Evaluations
- Conferences

Lab class 2: RMarkdown and Exploratory Data Analysis

R Markdown Lab

- R file: lab_class_2_RMarkdown.Rmd
- Compiled file: lab_class_2_RMarkdown.html

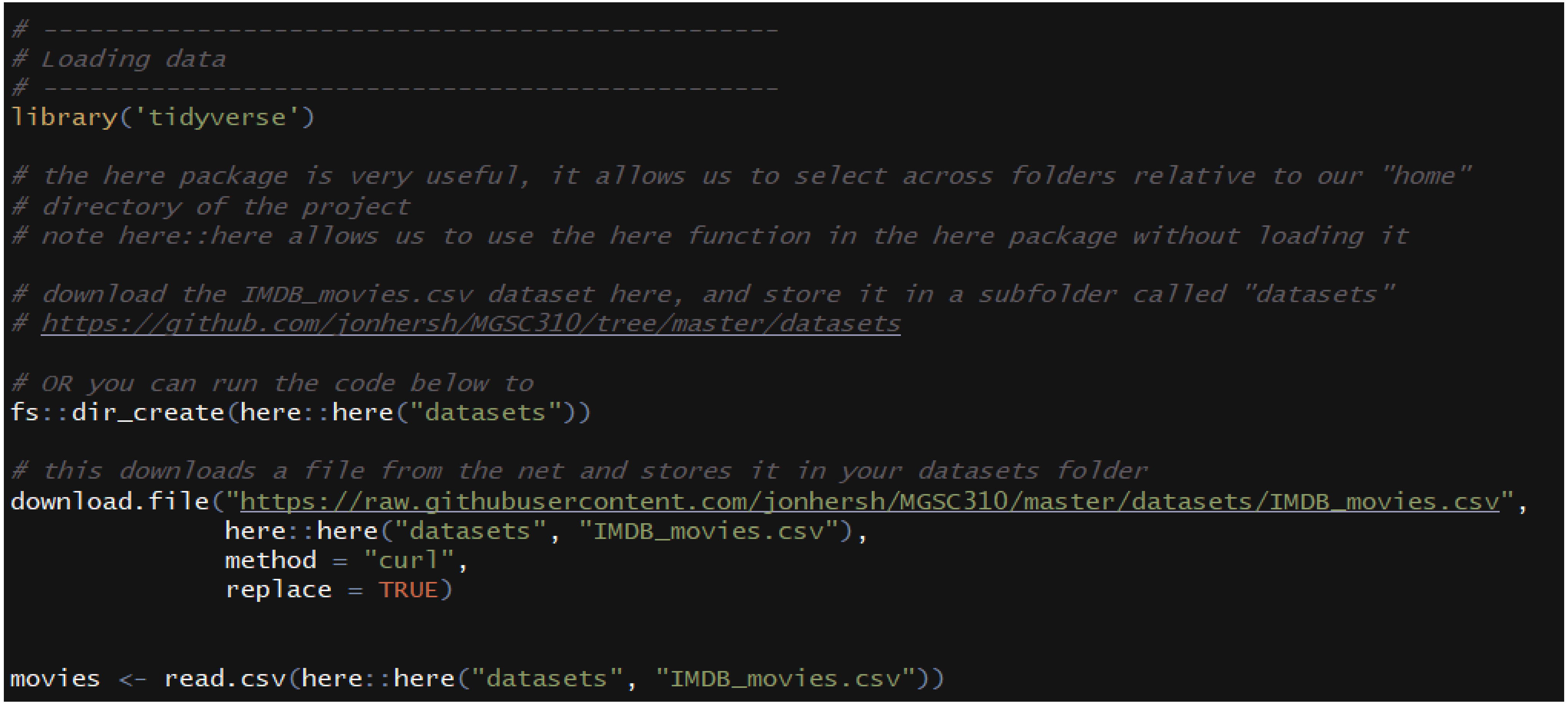
Data Analysis Lab

- R file: lab_class_2_R_Data_Analysis.R

◂ Previous

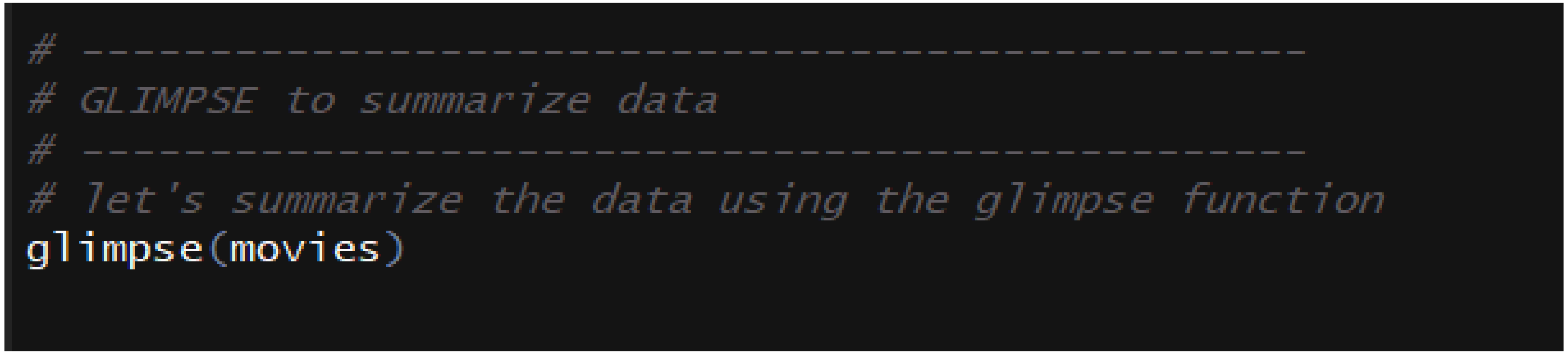
- Please complete R Markdown lab and ensure you can export a compiled HTML file!

Loading Data



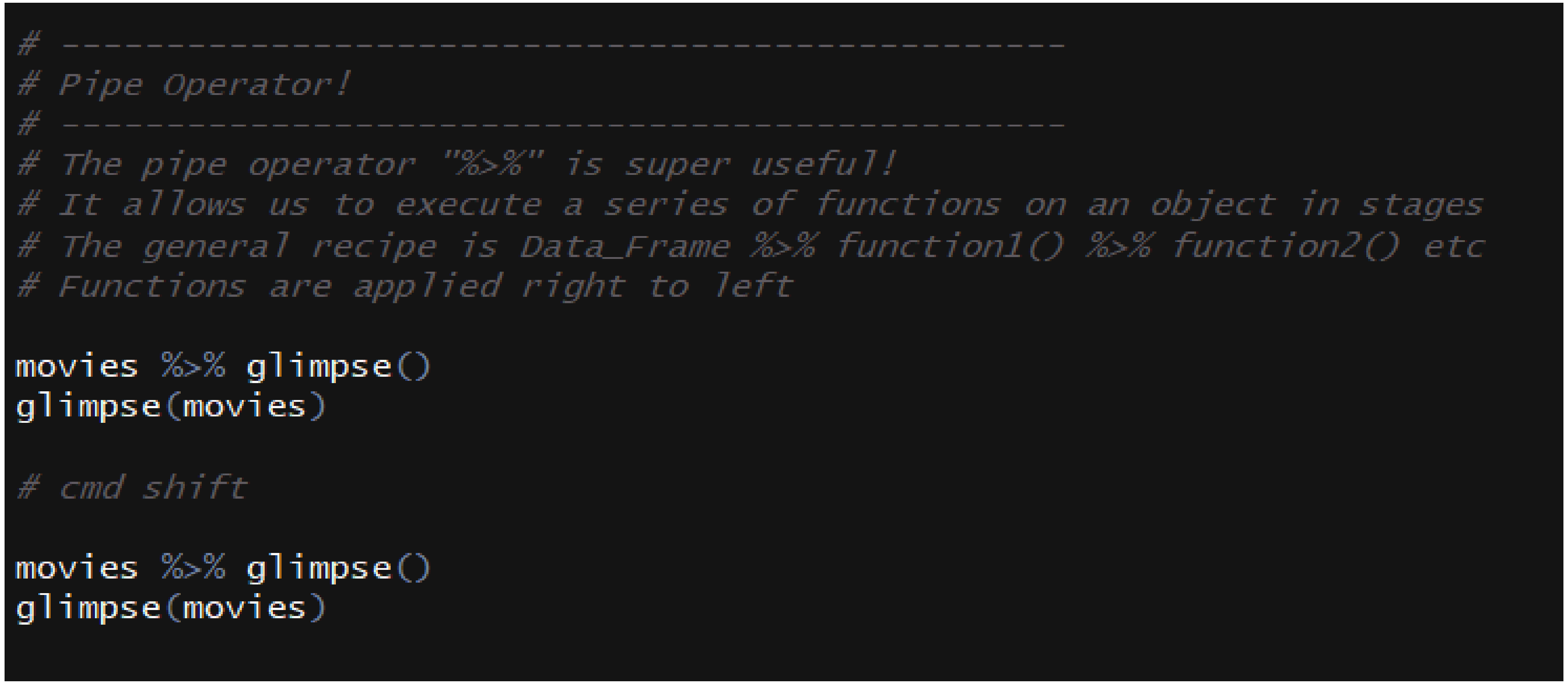
```
# ------------------------------------------------
# Loading data
# ------------------------------------------------
library('tidyverse')

# the here package is very useful, it allows us to select across folders relative to our "home"
# directory of the project
# note here::here allows us to use the here function in the here package without loading it

# download the IMDB_movies.csv dataset here, and store it in a subfolder called "datasets"
# https://github.com/jonhersh/MGSC310/tree/master/datasets

# OR you can run the code below to
fs::dir_create(here::here("datasets"))

# this downloads a file from the net and stores it in your datasets folder
download.file("https://raw.githubusercontent.com/jonhersh/MGSC310/master/datasets/IMDB_movies.csv",
              here::here("datasets", "IMDB_movies.csv"),
              method = "curl",
              replace = TRUE)


movies <- read.csv(here::here("datasets", "IMDB_movies.csv"))
```

Glimpse to Summarize Data



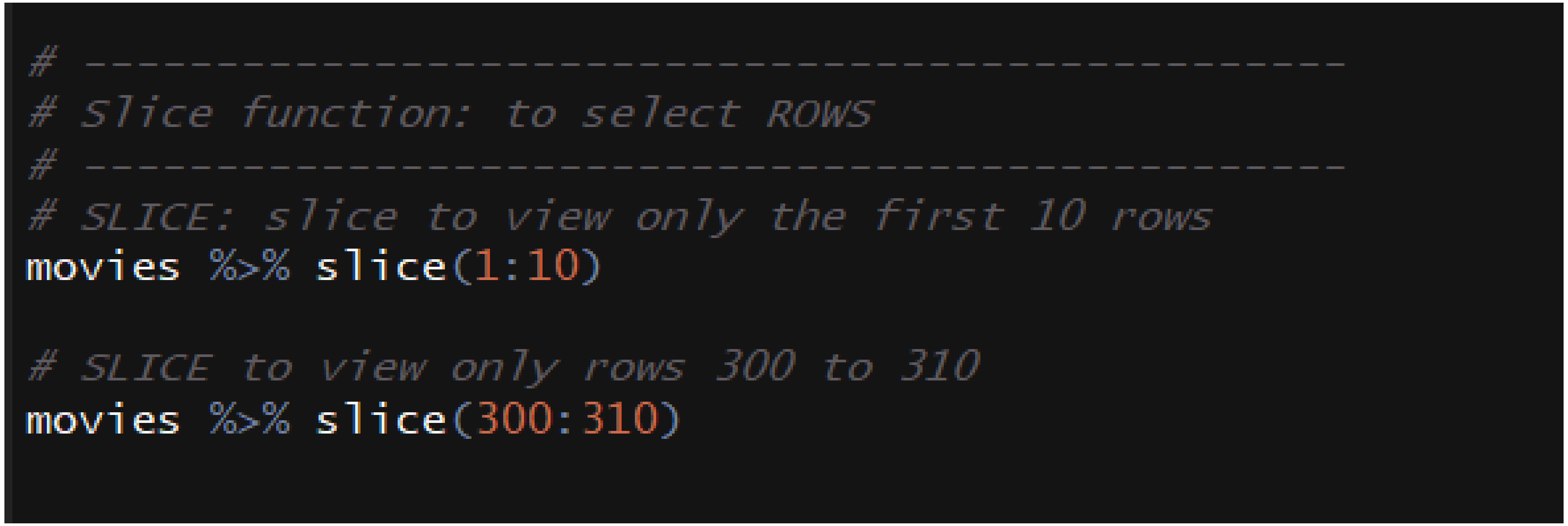
```
# ------------------------------------------------
# GLIMPSE to summarize data
# ------------------------------------------------
# let's summarize the data using the glimpse function
glimpse(movies)
```

Pipe Operator



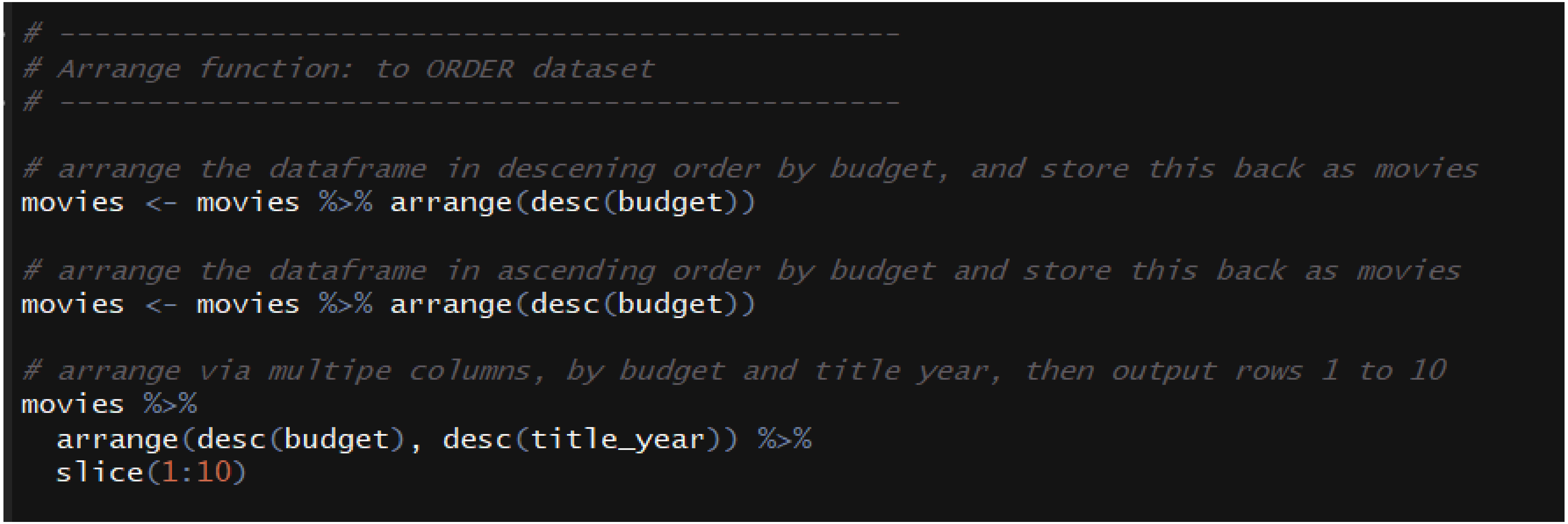
```
# ------------------------------------------------
# Pipe Operator!
# ------------------------------------------------
# The pipe operator "%>%" is super useful!
# It allows us to execute a series of functions on an object in stages
# The general recipe is Data_Frame %>% function1() %>% function2() etc
# Functions are applied right to left

movies %>% glimpse()
glimpse(movies)

# cmd shift

movies %>% glimpse()
glimpse(movies)
```

Slice to View Rows



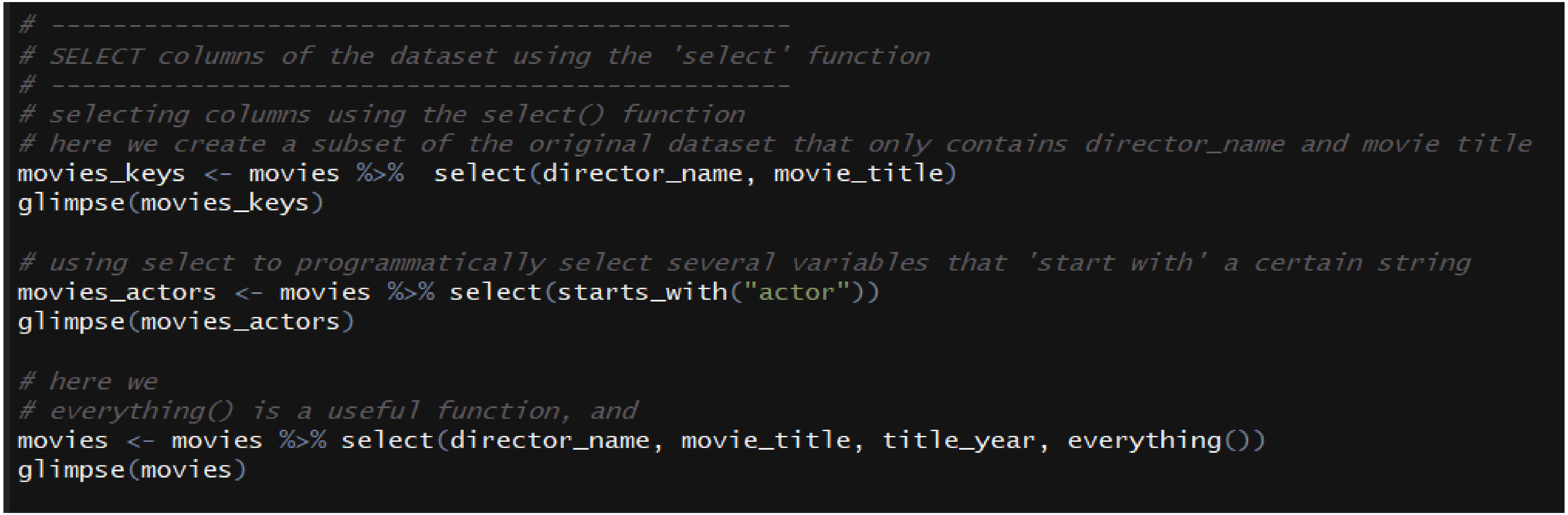
```
# ------------------------------------------------
# Slice function: to select ROWS
# ------------------------------------------------
# SLICE: slice to view only the first 10 rows
movies %>% slice(1:10)

# SLICE to view only rows 300 to 310
movies %>% slice(300:310)
```

Arrange function: to ORDER dataset



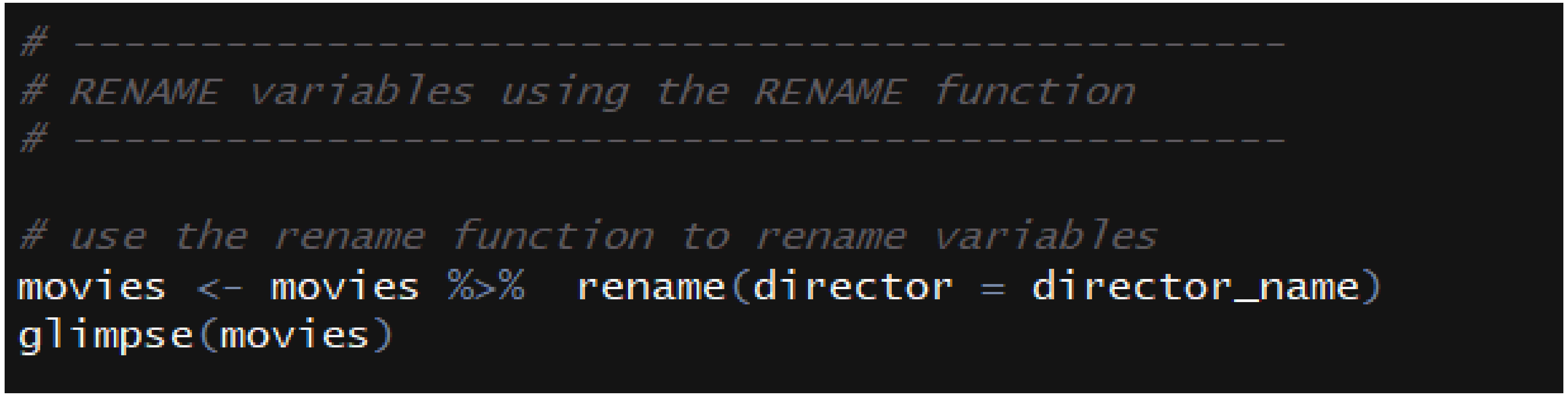
```
# ----------------------------------------------
# Arrange function: to ORDER dataset
# ----------------------------------------------

# arrange the dataframe in descening order by budget, and store this back as movies
movies <- movies %>% arrange(desc(budget))

# arrange the dataframe in ascending order by budget and store this back as movies
movies <- movies %>% arrange(desc(budget))

# arrange via multipe columns, by budget and title year, then output rows 1 to 10
movies %>%
  arrange(desc(budget), desc(title_year)) %>%
  slice(1:10)
```

SELECT columns of the dataset using the 'select' function



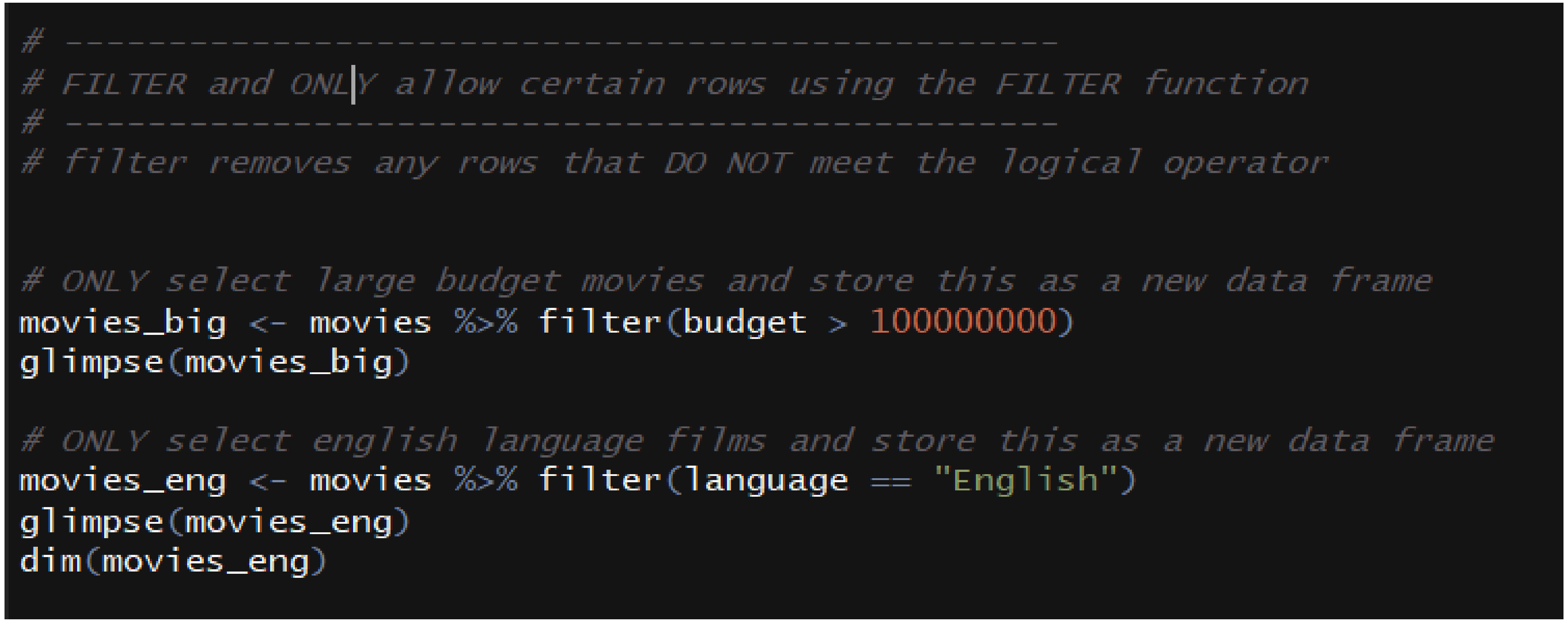
```
# ------------------------------------------------
# SELECT columns of the dataset using the 'select' function
# ------------------------------------------------
# selecting columns using the select() function
# here we create a subset of the original dataset that only contains director_name and movie title
movies_keys <- movies %>%  select(director_name, movie_title)
glimpse(movies_keys)

# using select to programmatically select several variables that 'start with' a certain string
movies_actors <- movies %>% select(starts_with("actor"))
glimpse(movies_actors)

# here we
# everything() is a useful function, and
movies <- movies %>% select(director_name, movie_title, title_year, everything())
glimpse(movies)
```

RENAME variables using the RENAME function



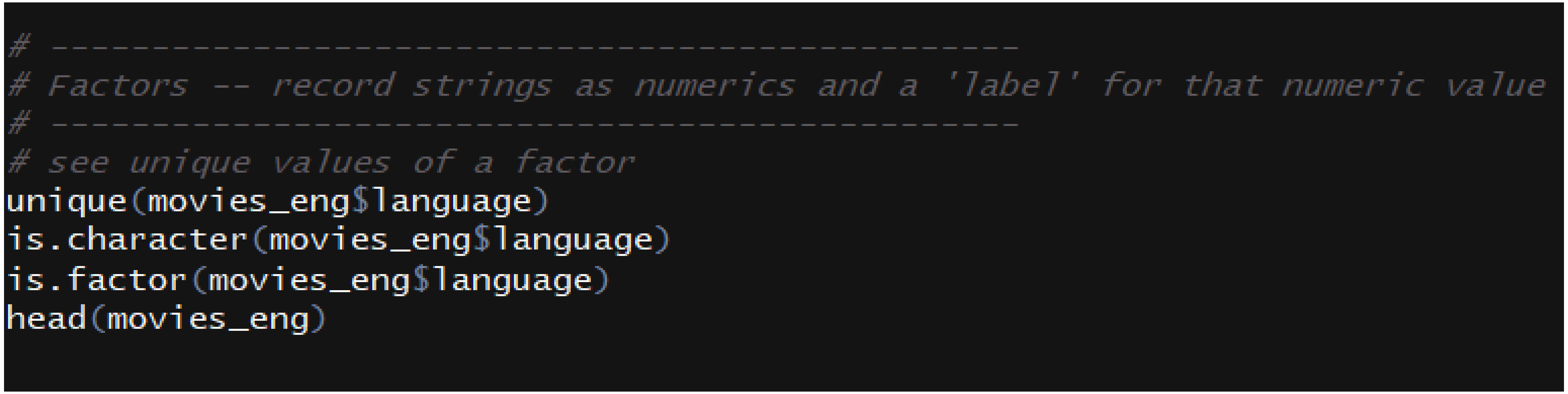
```
# ------------------------------------------------
# RENAME variables using the RENAME function
# ------------------------------------------------

# use the rename function to rename variables
movies <- movies %>%  rename(director = director_name)
glimpse(movies)
```

FILTER and ONLY allow certain rows using the FILTER function



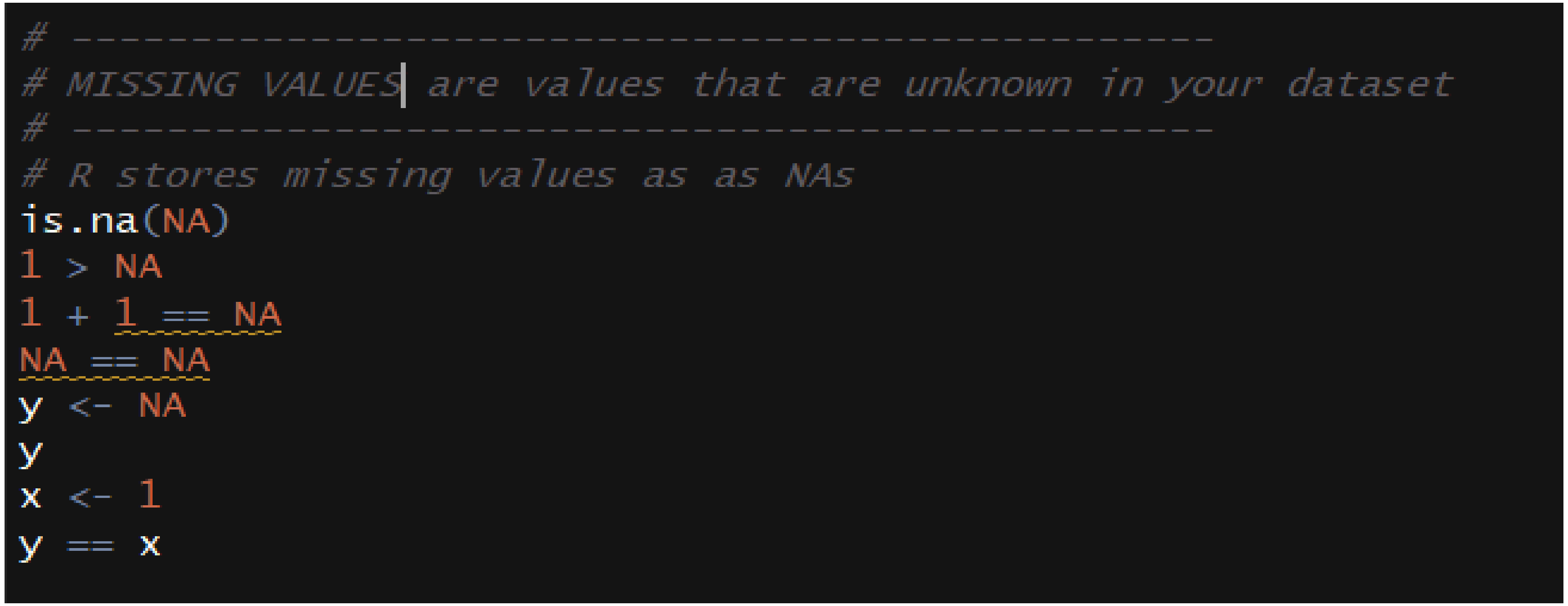
```
# ------------------------------------------------
# FILTER and ONLY allow certain rows using the FILTER function
# ------------------------------------------------
# filter removes any rows that DO NOT meet the logical operator


# ONLY select large budget movies and store this as a new data frame
movies_big <- movies %>% filter(budget > 100000000)
glimpse(movies_big)

# ONLY select english language films and store this as a new data frame
movies_eng <- movies %>% filter(language == "English")
glimpse(movies_eng)
dim(movies_eng)
```

Factors -- record strings as numerics and a 'label' for that numeric value



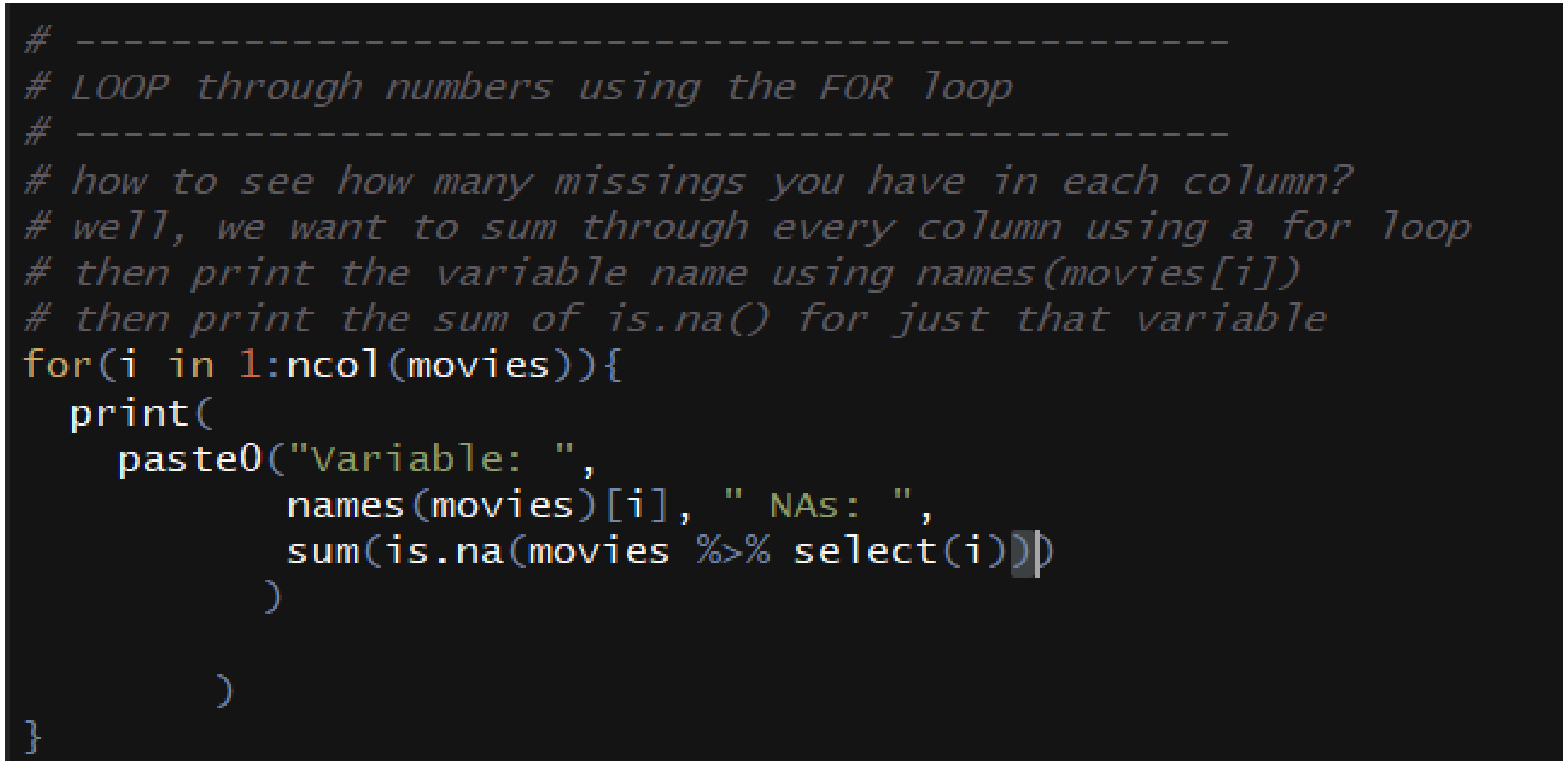
```
# ------------------------------------------------
# Factors -- record strings as numerics and a 'label' for that numeric value
# ------------------------------------------------
# see unique values of a factor
unique(movies_eng$language)
is.character(movies_eng$language)
is.factor(movies_eng$language)
head(movies_eng)
```

MISSING VALUES are values that are unknown in your dataset



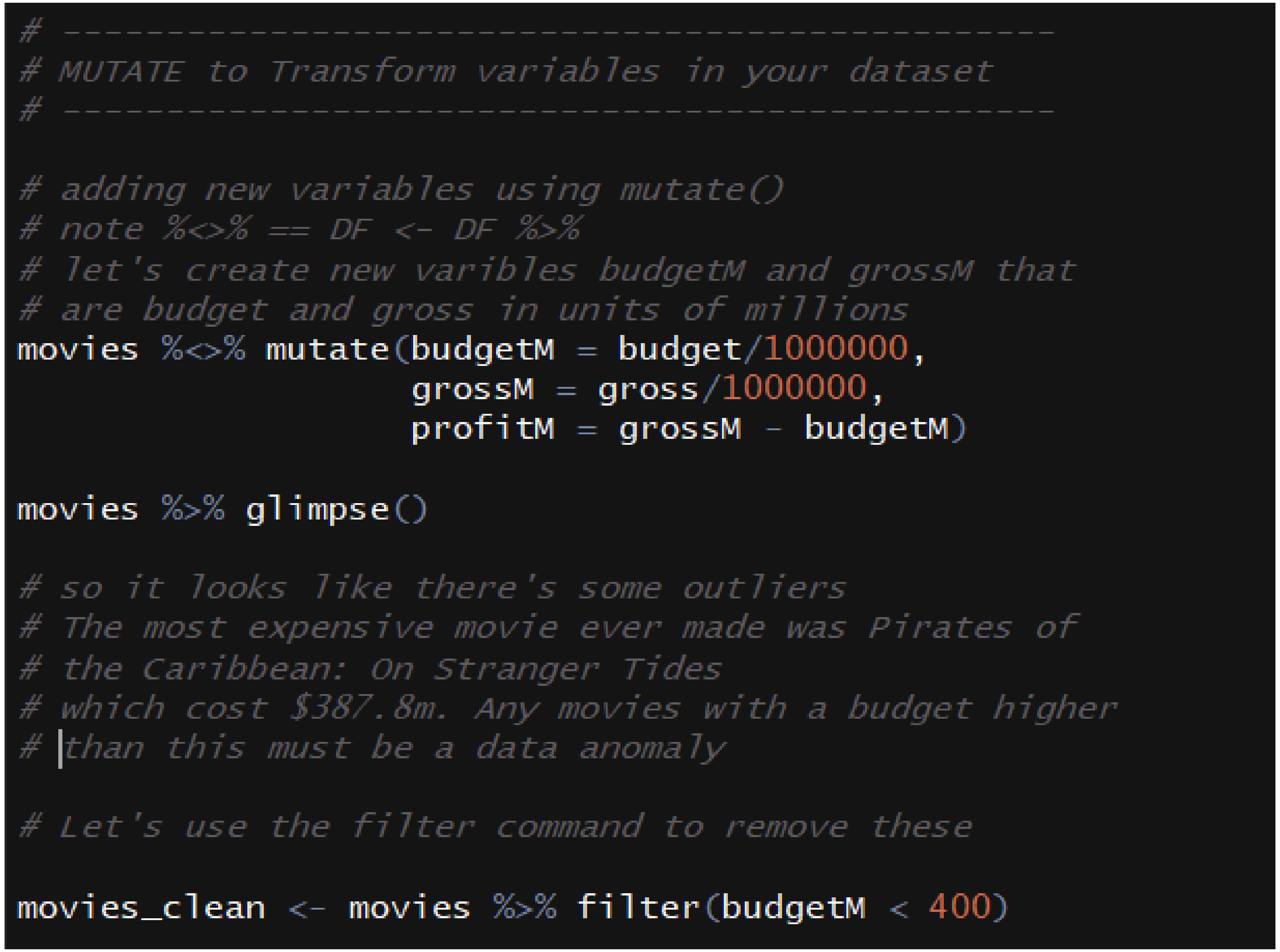
```
# ------------------------------------------------
# MISSING VALUES are values that are unknown in your dataset
# ------------------------------------------------
# R stores missing values as as NAs
is.na(NA)
1 > NA
1 + 1 == NA
NA == NA
y <- NA
y
x <- 1
y == x
```

LOOP through numbers using the FOR loop



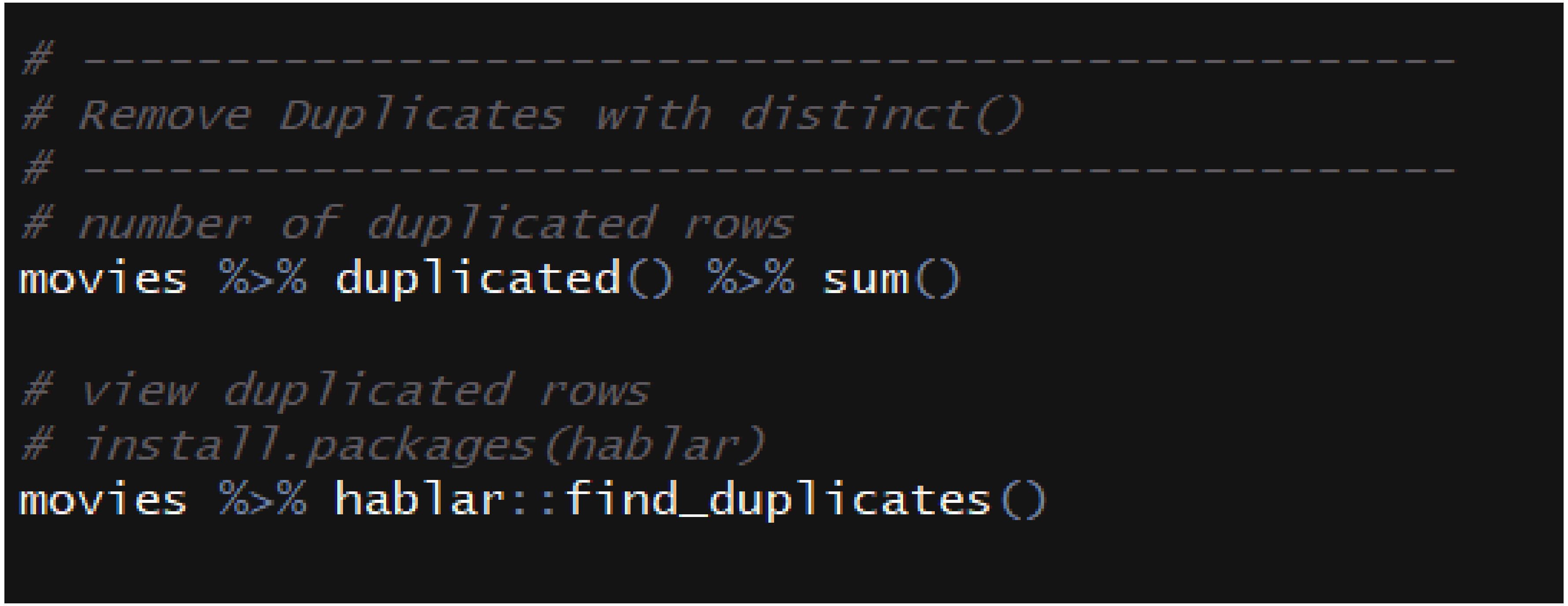
```
# -------------------------------------------------
# LOOP through numbers using the FOR loop
# -------------------------------------------------
# how to see how many missings you have in each column?
# well, we want to sum through every column using a for loop
# then print the variable name using names(movies[i])
# then print the sum of is.na() for just that variable
for(i in 1:ncol(movies)){
  print(
    paste0("Variable: ",
           names(movies)[i], " NAs: ",
           sum(is.na(movies %>% select(i))))
          )

        )
}
```

MUTATE to Transform variables in your dataset



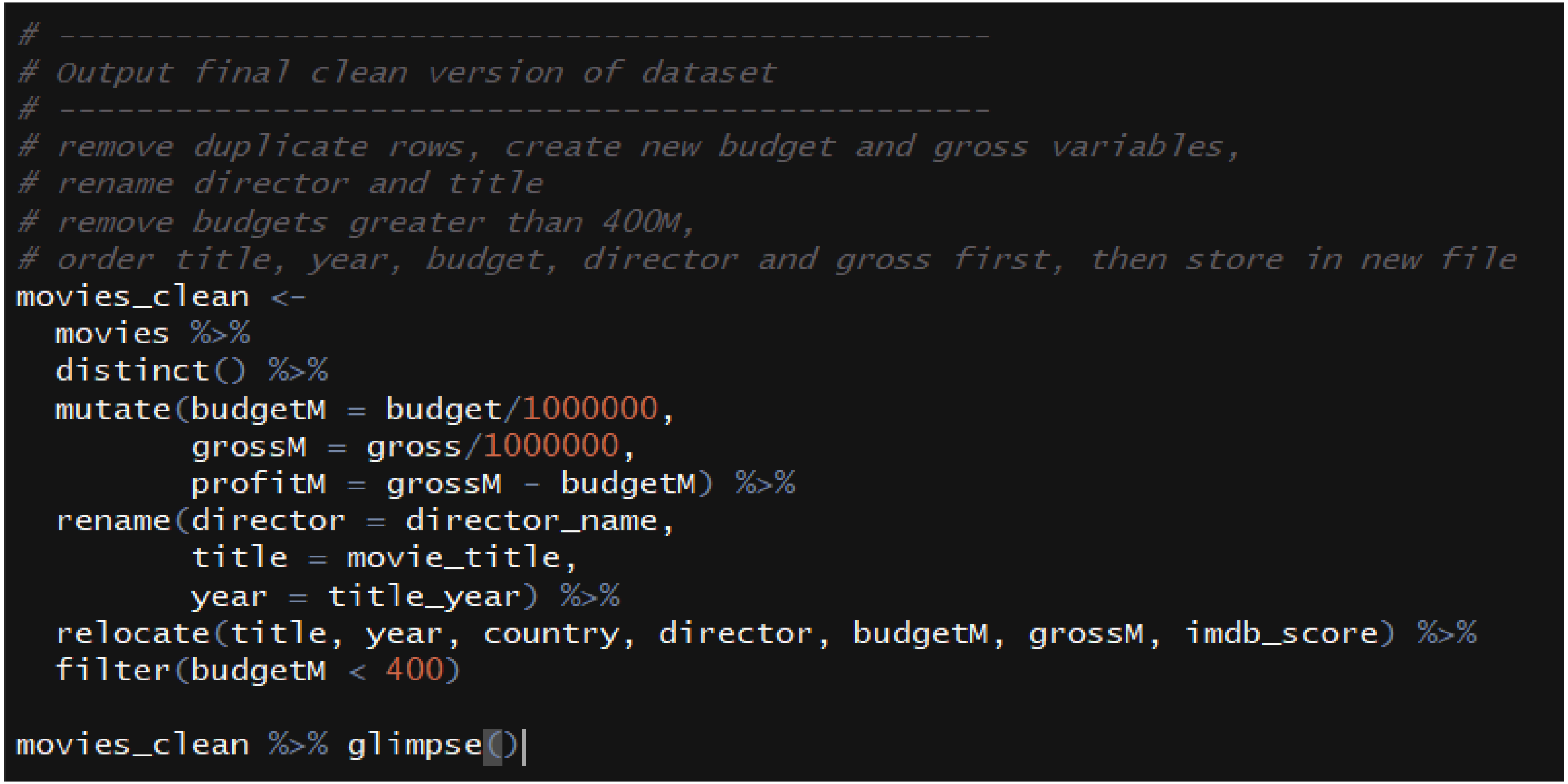
```
# ------------------------------------------------
# MUTATE to Transform variables in your dataset
# ------------------------------------------------

# adding new variables using mutate()
# note %<>% == DF <- DF %>%
# let's create new varibles budgetM and grossM that
# are budget and gross in units of millions
movies %<>% mutate(budgetM = budget/1000000,
                   grossM = gross/1000000,
                   profitM = grossM - budgetM)

movies %>% glimpse()

# so it looks like there's some outliers
# The most expensive movie ever made was Pirates of
# the Caribbean: On Stranger Tides
# which cost $387.8m. Any movies with a budget higher
# than this must be a data anomaly

# Let's use the filter command to remove these

movies_clean <- movies %>% filter(budgetM < 400)
```

Remove Duplicates with distinct()



```
# ------------------------------------------------
# Remove Duplicates with distinct()
# ------------------------------------------------
# number of duplicated rows
movies %>% duplicated() %>% sum()

# view duplicated rows
# install.packages(hablar)
movies %>% hablar::find_duplicates()
```

Output final clean version of dataset

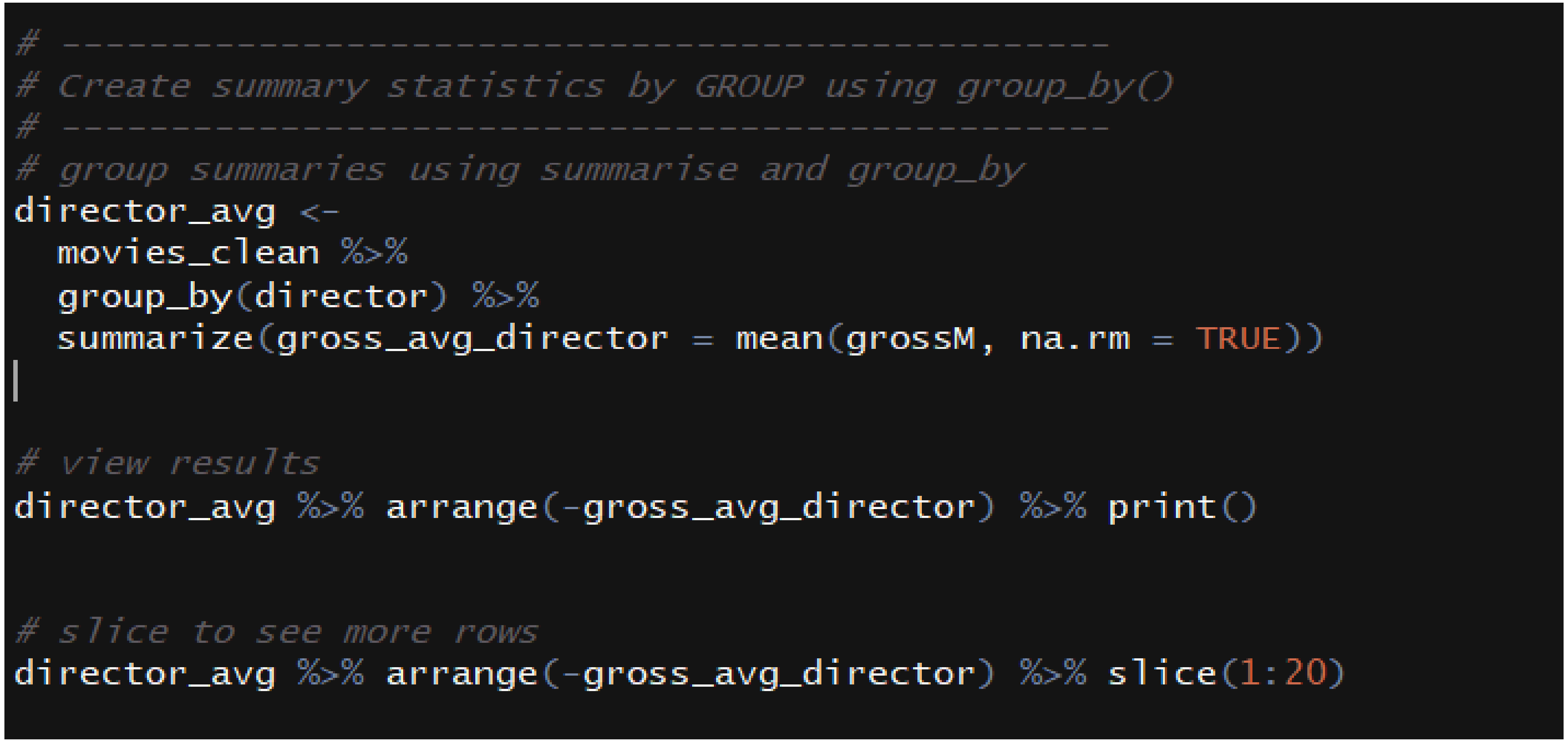


```
# ------------------------------------------------
# Output final clean version of dataset
# ------------------------------------------------
# remove duplicate rows, create new budget and gross variables,
# rename director and title
# remove budgets greater than 400M,
# order title, year, budget, director and gross first, then store in new file
movies_clean <-
  movies %>%
  distinct() %>%
  mutate(budgetM = budget/1000000,
         grossM = gross/1000000,
         profitM = grossM - budgetM) %>%
  rename(director = director_name,
         title = movie_title,
         year = title_year) %>%
  relocate(title, year, country, director, budgetM, grossM, imdb_score) %>%
  filter(budgetM < 400)

movies_clean %>% glimpse()
```

Exercises - Lab

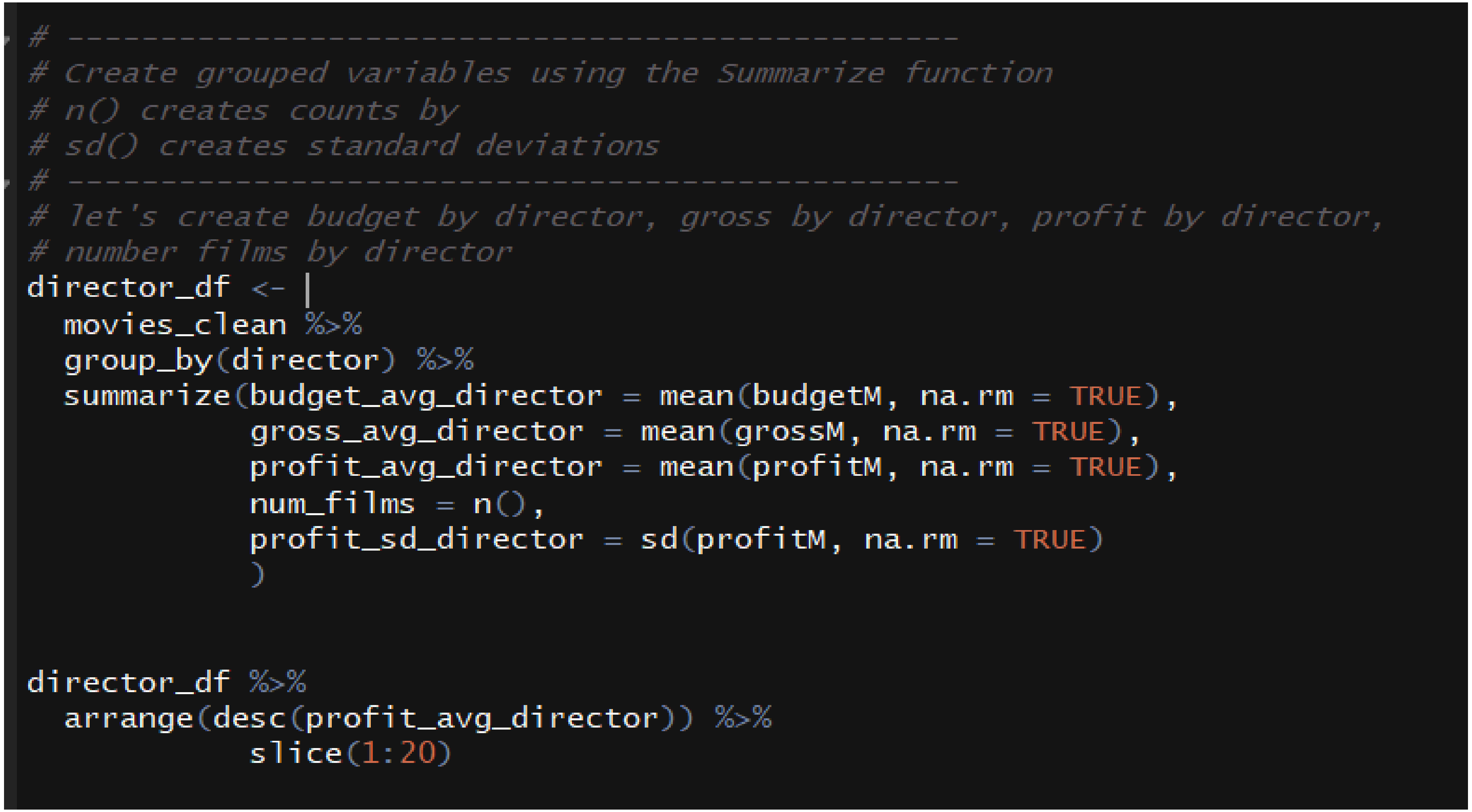
1. What are the highest grossing Steven Spielberg films?
2. What's the highest grossing film in the dataset?
3. Which film lost the most money?
4. Which film made the most money?
5. How many "PG-13" movies are there in the database?
6. Which movie has the most facebook likes?
7. Make 1-2 interesting ggplots using the movies_clean dataset

Create summary statistics by GROUP using group_by()



```
# ------------------------------------------------
# Create summary statistics by GROUP using group_by()
# ------------------------------------------------
# group summaries using summarise and group_by
director_avg <-
  movies_clean %>%
  group_by(director) %>%
  summarize(gross_avg_director = mean(grossM, na.rm = TRUE))


# view results
director_avg %>% arrange(-gross_avg_director) %>% print()


# slice to see more rows
director_avg %>% arrange(-gross_avg_director) %>% slice(1:20)
```

Create count and standard deviation by groups



```r
# ------------------------------------------------
# Create grouped variables using the Summarize function
# n() creates counts by
# sd() creates standard deviations
# ------------------------------------------------
# let's create budget by director, gross by director, profit by director,
# number films by director
director_df <- 
  movies_clean %>%
  group_by(director) %>%
  summarize(budget_avg_director = mean(budgetM, na.rm = TRUE),
            gross_avg_director = mean(grossM, na.rm = TRUE),
            profit_avg_director = mean(profitM, na.rm = TRUE),
            num_films = n(),
            profit_sd_director = sd(profitM, na.rm = TRUE)
            )


director_df %>%
  arrange(desc(profit_avg_director)) %>%
            slice(1:20)
```

Exercises – Data Analysis Lab 2

1. Which director made the most films in the IMDB 5000 database
2. Which director has the highest standard deviation of profit?
3. Which director has the highest profit?
4. Which director has the lowest profit?
5. How many movies has George Lucas Made?
6. Make 1-3 ggplots using the director_df showing revealing patterns.