Dataquest Guided Project: Finding the Best Markets to Advertise In

Cindy Zhang

11/30/2020

Contents

Introduction **1**

Findings **1**

Step 1: Understanding the Data 1

Step 2: Checking for Sample Representativity 2

Step 3: New Coders - Locations and Densities 2

Step 4: Spending Money for Learning 2

Step 5: Dealing with Extreme Outliers 2

Step 6: Choosing the Two Best Markets 2

Conclusion **3**

Introduction

This is my solution to Dataquest's Guided Project from the second Probability and Statistics course, which practices analyzing distributions of data from a fictional e-learning company.

More details such as the RMD and csv files can be found in the repository in GitHub. More details about the survey response variables can be found here for 2014-15 and here (https://github.com/mircealex/Movie_ratings_2016_17/blob/master/README.md) for 2016-17.

Findings

Step 1: Understanding the Data

```
coders <- data.frame(read_csv("coders.csv"))
dim(coders)
colnames(coders)
for (i in colnames(coders)) {
  print(class(coders[[i]]))
}
```

Step 2: Checking for Sample Representativity

```
freq_dist_location <- coders %>%
  group_by(CountryLive) %>%
  summarize(Freq=n()) %>%
  arrange(desc(Freq))
```

```
head(freq_dist_location)
```

```
## # A tibble: 6 x 2
##   CountryLive               Freq
##   <chr>                    <int>
## 1 United States of America  5791
## 2 <NA>                      2839
## 3 India                     1400
## 4 United Kingdom             757
## 5 Canada                     616
## 6 Brazil                     364
```

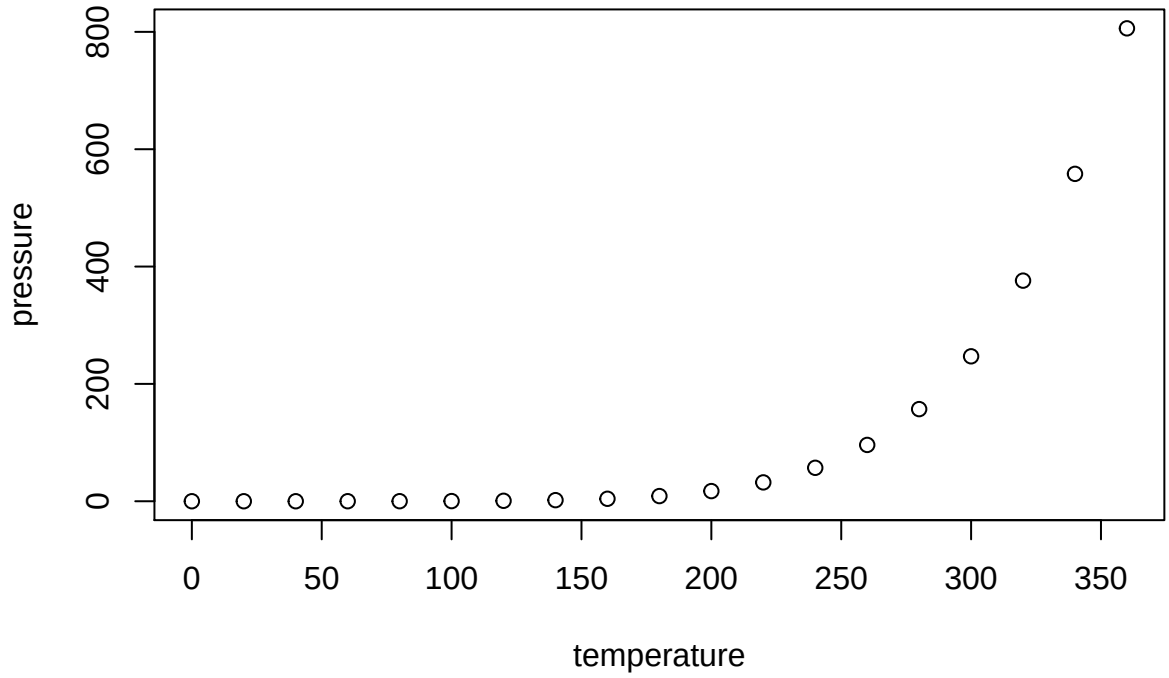
Step 3: New Coders - Locations and Densities

Step 4: Spending Money for Learning

Step 5: Dealing with Extreme Outliers

Step 6: Choosing the Two Best Markets

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Conclusion