Quiz: Exploring the Data

Eddo W. Hintoso

October 21, 2015

1. After untaring the the dataset, how many files are there (including the documentation pdfs)?

7

---

2. The data files are in what format?

`.json`

---

3. How many lines of text are there in the reviews file (in orders of magnitude)?

```
yelpReviewRDS <- readRDS('../yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_review.rds')
dim(yelpReviewRDS)
```

```
## [1] 1569264      10
```

One million

---

4. Consider line 100 of the reviews file. "I've been going to the Grab n Eat for almost XXX years"

```
yelpReviewRDS$text[100]
```

```
## [1] "I have been coming to Gab n Eat for almost 20 years and They have never let me down. I get a ty
```

20

---

5. What percentage of the reviews are five star reviews (rounded to the nearest percentage point)?

```
100 * length(yelpReviewRDS$stars[yelpReviewRDS$stars == 5]) / length(yelpReviewRDS$stars)
```

```
## [1] 36.92986
```

37%

---

6. How many lines are there in the businesses file?

```
yelpBusinessRDS <- readRDS('../yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_business.r
dim(yelpBusinessRDS)
```

```
## [1] 61184   105
```

About 60 thousand

---

7. Conditional on having an response for the attribute "Wi-Fi", how many businesses are reported for having free wi-fi (rounded to the nearest percentage point)?

```
100 * length(na.omit(yelpBusinessRDS$`attributes.Wi-Fi`[yelpBusinessRDS$`attributes.Wi-Fi` == "free"]))
```

```
## [1] 40.91519
```

40%

---

8. How many lines are in the tip file?

```
yelpTipRDS <- readRDS('../yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_tip.rds')
dim(yelpTipRDS)
```

```
## [1] 495107      6
```

About 500 thousand

---

9. In the tips file on the 1,000th line, fill in the blank: "Consistently terrible _____"

```
yelpTipRDS$text[1000]
```

```
## [1] "Consistently terrible service. What's with the attitudes?"
```

service

---

10. What is the name of the user with over 10,000 compliment votes of type "funny"?

```
yelpUserRDS <- readRDS('../yelp_dataset_challenge_academic_dataset/yelp_academic_dataset_user.rds')
subset(yelpUserRDS, compliments.funny > 10000)$name
```

```
## [1] "Brian"
```

Brian

---

11. Create a 2 by 2 cross tabulation table of when a user has more than 1 fans to if the user has more than 1 compliment of type "funny". Treat missing values as 0 (fans or votes of that type). Pass the 2 by 2 table to fisher.test in R. What is the P-value for the test of independence?

```
yelpUserRDS$compliments.funny[is.na(yelpUserRDS$compliments.funny)] <- 0
yelpUserRDS$fans[is.na(yelpUserRDS$fans)] <- 0
cross.table <- matrix(c(sum(yelpUserRDS$compliments.funny > 1),
                        sum(yelpUserRDS$compliments.funny <= 1),
                        sum(yelpUserRDS$fans > 1),
                        sum(yelpUserRDS$fans <= 1)),
                      nrow = 2)
fisher.test(cross.table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cross.table
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5185328 0.5344792
## sample estimates:
## odds ratio
##  0.5264532
```

less than .001