Albert Y. Kim
Adriana Escobedo-Land
Reed College
OkCupid Profile Data for Introductory Statistics and Data Science Courses

Keywords: OkCupid, online dating, data science, big data.

Abstract

We present a dataset consisting of user profile data for 59,946 San Francisco OkCupid users (an online dating site) from June 2012. The data set includes typical user information, lifestyle variables, and text responses to 10 essays questions. We present four example analyses suitable for use in undergraduate introductory probability & statistics and data science classes that use `R`. The statistical and data science concepts covered include basic data visualizations, exploratory data analyses, multivariate relationships, text analysis, and logistic regression for prediction.

1 Introduction

In light of the field of data science gaining more prominence in academia and industry, many statisticians are arguing that statistics needs to stake a bigger claim in data science in order to avoid marginalization by other disciplines such as computer science and computer engineering [1, 2]. While precise definition of the difference between statistics and data science and its implications for statistics education can be debated [3], one consensus among many in statistics education circles is that at the very least statistics needs to incorporate a heavier computing component and the use of technology for both developing conceptual understanding and analyzing data [4, 5]. Relatedly, in the hopes of making introductory undergraduate statistics courses more relevant, many statistics educators are placing a higher emphasis on the use of real data in the classroom, a practice the American Statistical Association's Guidelines for Assessment and Instruction in Statistics Education (GAISE) project's reports strongly encourages [4]. Of particular importance in the success of such ambitions are the data sets considered, as they provide the context of the analyses and thus will ultimately drive student interest [6].

It is in light of these discussions that we present this paper centering on data from the online dating website OkCupid, specifically a snapshot of San Francisco California users taken on June 2012. We describe the dataset and present a series of example analyses along with corresponding pedagogical discussions. The example analyses presented in this paper were used in a variety of settings at Reed College: a 90 minute introductory tutorial on R, an introductory probability and statistics class, and a follow-up two-hundred level data science class titled "Case Studies in Statistical Analysis." The statistical and data science concepts covered include basic data visualizations, exploratory data analyses, multivariate relationships, text analysis, and logistic regression for prediction. All examples are presented using the R statistical software program and make use of the `mosaic`, `dplyr`, `stringr`, and `ggplot2` packages [7, 8, 9, 10].

2 Data

The data consists of the public profiles of 59,946 OkCupid users who were living within 25 miles of San Francisco, had active profiles on June 26, 2012, were online in the previous year and had at least one picture in their profile. Using a Python script, data was scraped from users' public profiles on June 30, 2012; any non-publicly facing information such as messaging was not accessible.

Variables include typical user information (such as sex, sexual orientation, age, and ethnicity) and lifestyle variables (such as diet, drinking habits, smoking habits). Furthermore, text responses to the 10 essay questions posed to all OkCupid users are included as well, such as "My Self Summary", "The first thing people usually notice about me...", and "On a typical Friday night I am..." For a complete list of variables and more details, see the accompanying codebook `okcupid_codebook.txt`. We load the data as follows:

```
profiles <- read.csv(file="profiles.csv", header=TRUE, stringsAsFactors=FALSE)
n <- nrow(profiles)
```

Analysis of similar data has received much press of late, including Amy Webb's TED talk "How I Hacked Online Dating" [11] and Wired magazine's "How a Math Genius Hacked OkCupid to Find True Love" [12]. OkCupid co-founder Christian Rudder pens periodical analysis of their data on the OkTrends blog [13] and has recently published a book "Dataclysm: Who We Are When We Think No One's Looking" describing similar data-driven insights [14]. Such publicity surrounding data-driven online dating and the salience of dating matters among students makes this dataset one with much potential to be of interest to students, hence facilitating the instruction of statistical and data science concepts.

Before we continue we note that even though this data consists of publicly facing material, one should proceed with caution before scraping and using data in fashion similar to ours, as the Computer Fraud and Abuse Act (CFAA) makes it a federal crime to access a computer without authorization from the owner [15]. In our case, permission was given by the owners of the data (See Section 5).

3 Example Analyses

The example analyses we present address the following questions:

1. How do the heights of male and female OkCupid users compare?
2. What does the San Francisco online dating landscape look like? Or more specifically, what is the relationship between users' sex and sexual orientation?
3. Are there differences between the sexes in what words are used in the responses to the 10 essay questions?
4. How accurately can we predict a user's sex using their listed height?

For each question, we present an exercise as would be given to students in a lab setting followed by a pedagogical discussion.

3.1 Male and Female Heights

3.1.1 Exercise

We compare the distribution of male and female OkCupid users' heights. Height is one of 3 numerical variables in this dataset (the others being age and income). This provides us an opportunity to investigate numerical summaries using the `favstats()` command from the `mosaic` package:

```
favstats(height, data=profiles)

##  min Q1 median Q3 max mean sd     n missing
##    1 66     68 71  95   68  4 59943       3
```

We observe that some of the heights are nonsensical, including heights of 1 inch and 95 inches (equaling 7'11"). We deem heights between 55 and 80 inches to be reasonable and remove the rest. While there is potential bias in discarding users with what we deem non-reasonable heights, since out of the 59946 users there are only 117 who would be discarded, the effect would not be large hence. We keep only those users with heights between 55 and 80 inches either using the `subset()` command or using the `filter()` command from `dplyr` package. Note that the following commands are equivalent:

```
profiles.subset <- subset(profiles, height>=55 & height <=80)
profiles.subset <- filter(profiles, height>=55 & height <=80)
```

We compare the distributions of male and female heights using histograms. While we could plot two separate histograms without regard to the scale of the x-axis, in Figure 1 we instead

1. Plot them with bin widths matching the granularity of the observations (inches)
2. Plot them simultaneously in a panel consisting of two rows and one column of plots using the command `par(mfrow=c(2,1))`
3. Plot them with the same scale on the x-axis (by matching the histograms' `breaks=55:80`) and the y-axis (by selecting a density histogram and not a frequency histogram using `prob=TRUE`) to facilitate comparisons between the two distributions

```
profiles.male <- filter(profiles.subset, sex=='m')
profiles.female <- filter(profiles.subset, sex=='f')
par(mfrow=c(2,1))
hist(profiles.female$height, breaks=55:80, main="Females", prob=TRUE, xlab="Height (in.)")
hist(profiles.male$height, breaks=55:80, main="Males", prob=TRUE, xlab="Height (in.)")
```

3.1.2 Pedagogical Discussion

This first exercise stresses many important considerations students should keep in mind when working with real data. Firstly, it emphasizes the importance of performing an exploratory data analysis to identify anomalous observations and confronts students with the question of what to do with them. For example, while a height of 1 inch is clearly an outlier that needs to be removed, at what point does a height no longer become reasonable and what impact does their removal have on the conclusions? In our case, since only a small number of observations were removed, the impact is minimal.

Secondly, this exercise demonstrates the power of data visualizations as simple as histograms to convey insight and hence emphasizes the importance of putting careful thought into their construction. In our case, while having students plot two histograms simultaneously on the same scale in order to demonstrate that males have on average greater height may seem to be a rather pedantic goal at first, we encouraged students to take a closer look at the histograms and steered their focus towards the unusual peaks at 72 inches (6 feet) for males and 64 inches (5'4") for females. Many of the students could explain the phenomena of the peak at 72 inches for men: sociological perceptions of the rounded height of 6 feet. On the other hand, consensus was not as strong about perceptions of the height of 5'4" for women. Instructors can then refer students to the entry on OkCupid's blog OkTrends "The Biggest Lies in Online Data"[16] to show they've replicated (on a smaller scale) a previous analysis and then show other analyses conducted by OkCupid.

Further questions that can be pursued from this exercise include "How can we question if those peaks are significant or due to chance?", "Are we only observing men who are just under 6 feet rounding up, or are men just over 6 feet rounding down as well?", or "How can we compare the distribution of listed heights on OkCupid to the actual San Francisco population's height distribution?"

3.2 Relationship Between Sex and Sexual Orientation

3.2.1 Exercise

Since among the most important considerations in assessing a potential mate are their sex and sexual orientation, in this exercise we investigate the relationship between these two variables. At the time, OkCupid allowed for two possible sex choices (male or female) and three possible sexual orientation (gay, bisexual, or straight)[1]. First, we perform a basic exploratory data analysis on these variables using barcharts in Figure

[1] OkCupid has since relaxed these categorizations to allow for a broader range of choices for both sex and sexual orientation.

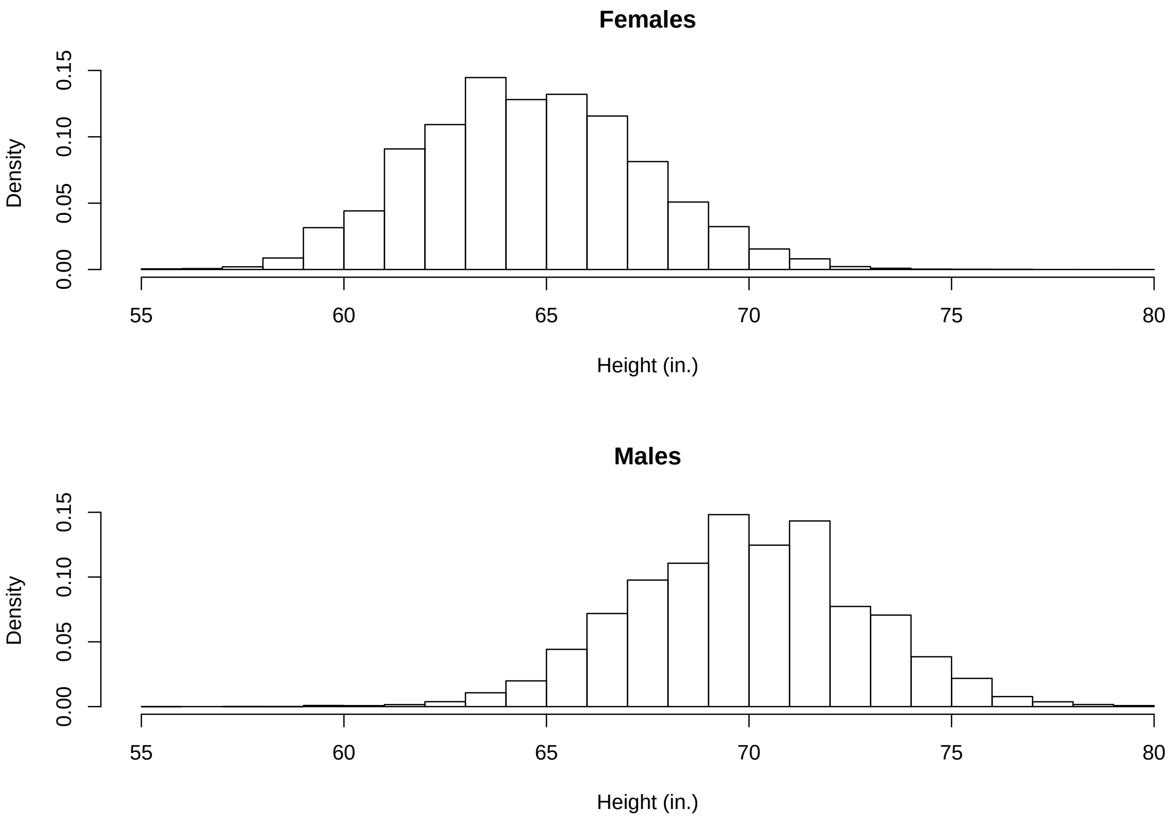

Figure 1: Histograms of user heights split by sex.

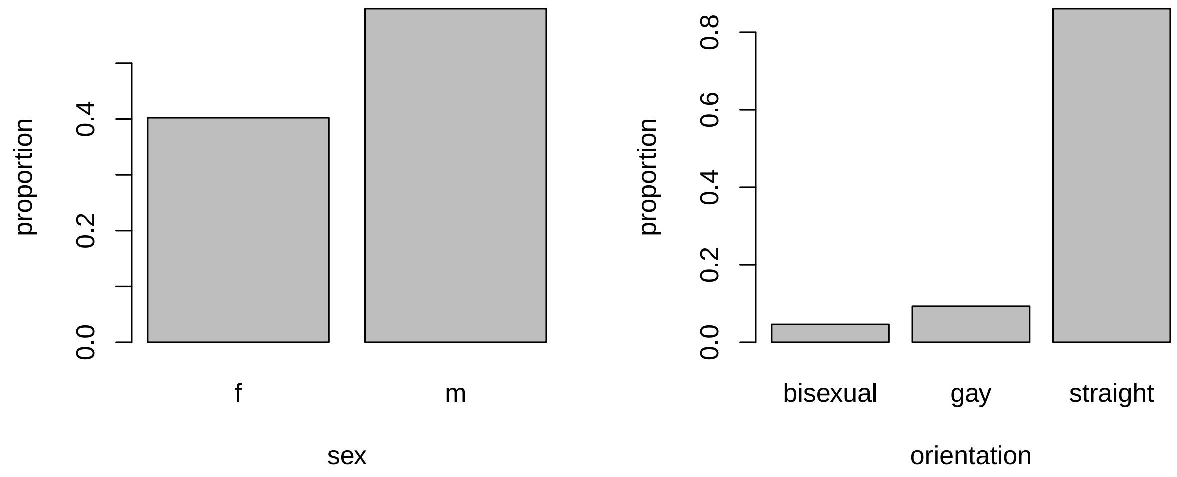

Figure 2: Distributions of sex and sexual orientation.

2:

```
par(mfrow=c(1, 2))
barplot(table(profiles$sex)/n, xlab="sex", ylab="proportion")
barplot(table(profiles$orientation)/n, xlab="orientation", ylab="proportion")
```

However, in order to accurately portray the dating landscape we can't just consider the **marginal distributions** of these variables, we must consider their **joint** and **conditional distributions** i.e. their cross-classifications. We describe the distribution of sexual orientation conditional on sex. For example, of all females (the condition) what proportion are bisexual? We do this using the `tally()` command from the `mosaic` package and note both columns sum to 1. Furthermore, we visualize their joint distribution, as represented by a contingency table, via the mosaicplot shown in Figure 3.

```
tally(orientation ~ sex, profiles, format='proportion')

##              sex
## orientation     f     m
##    bisexual 0.083 0.022
##    gay      0.066 0.111
##    straight 0.851 0.867

sex.by.orientation <- with(profiles, table(sex, orientation))
sex.by.orientation

##    orientation
## sex bisexual   gay straight
##   f     1996  1588    20533
##   m      771  3985    31073

mosaicplot(sex.by.orientation/n, main="Sex vs Orientation")
```

Do these results generalize to the entire San Francisco population?

3.2.2 Pedagogical Discussion

This exercise was an opportunity to concretize the statistical notions such as marginal/joint/conditional distributions and sampling bias. The data indicate that the San Francisco OkCupid dating population skews male and while the proportions of males and females who list themselves as straight are similar, a higher proportion of males list themselves as gay while a higher proportion of females list themselves as bisexual. Many students were not surprised by this fact as they were well aware of the gender imbalance issues in the large technology sector in the San Francisco Bay Area and San Francisco's history of being a bastion for the gay community.

The question of generalizability was presented in an introductory probability and statistics class homework. Almost all students were able to recognize the selection biases of who signs up for this site and hence the non-generalizability of the results. For example some recognized that OkCupid's demographic is most likely different than other dating website demographics such as `match.com` (which is not free) or `christiansingles.com` (which is targeted towards Christians). So while 59946 users may initially seem like a fairly large sample, we emphasized that "bigger" isn't always "better" when it comes to obtaining generalizable results due to the aforementioned selection biases. This proved an excellent segue to Kate Crawford of Microsoft Research's YouTube talk "Algorithmic Illusions: Hidden Biases of Big Data" [17].

Further questions one can pose to students include "Which dating demographic would you say has it the best and worst in terms of our simplified categorization?", "What variable do you think should be incorporated next in order to faithfully represent the OkCupid dating pool?", and "Even though not perfectly generalizable, to what degree can we apply these results to, for example, New York OkCupid users?".

3.3 Text Analysis

3.3.1 Exercise

The next exercise focuses on the responses to the essay questions, providing an opportunity to perform text analysis. Words are also called "strings" in the context of computer programming. Manipulating text data in R is often a complicated affair, so we present some code that is at an intermediate level to preprocess the essay responses for analysis. In order to output a single vector `essays` that contains all 10 essay responses for each user concatenated together, the following code:

- Uses the `select()` command from the `dplyr` package to select the 10 essay columns as identified by the fact they `starts_with("essay")`.
- For each user concatenates the 10 columns to form a single character string. The code applies the function `paste(x, collapse=" ")` to every essay, where `x` is a user's set of 10 essay responses and the `paste()` function collapses `x` across columns while separating the elements by a space. We do this for each set of essays (i.e. each row of `essays`) via the `apply()` command and setting the second argument to `1`.
- Replace all HTML line breaks (`\n`) and paragraph breaks (`<br />`) with spaces as well to make the outputs more readable.

```r
essays <- select(profiles, starts_with("essay"))
essays <- apply(essays, 1, paste, collapse=" ")
essays <- str_replace_all(essays, "\n", " ")
essays <- str_replace_all(essays, "<br />", " ")
```

We ask: Do male and female OkCupid users use words at different rates in their essay responses? We search for the presence of a word in a user's essays using the `str_detect()` command in the `stringr`

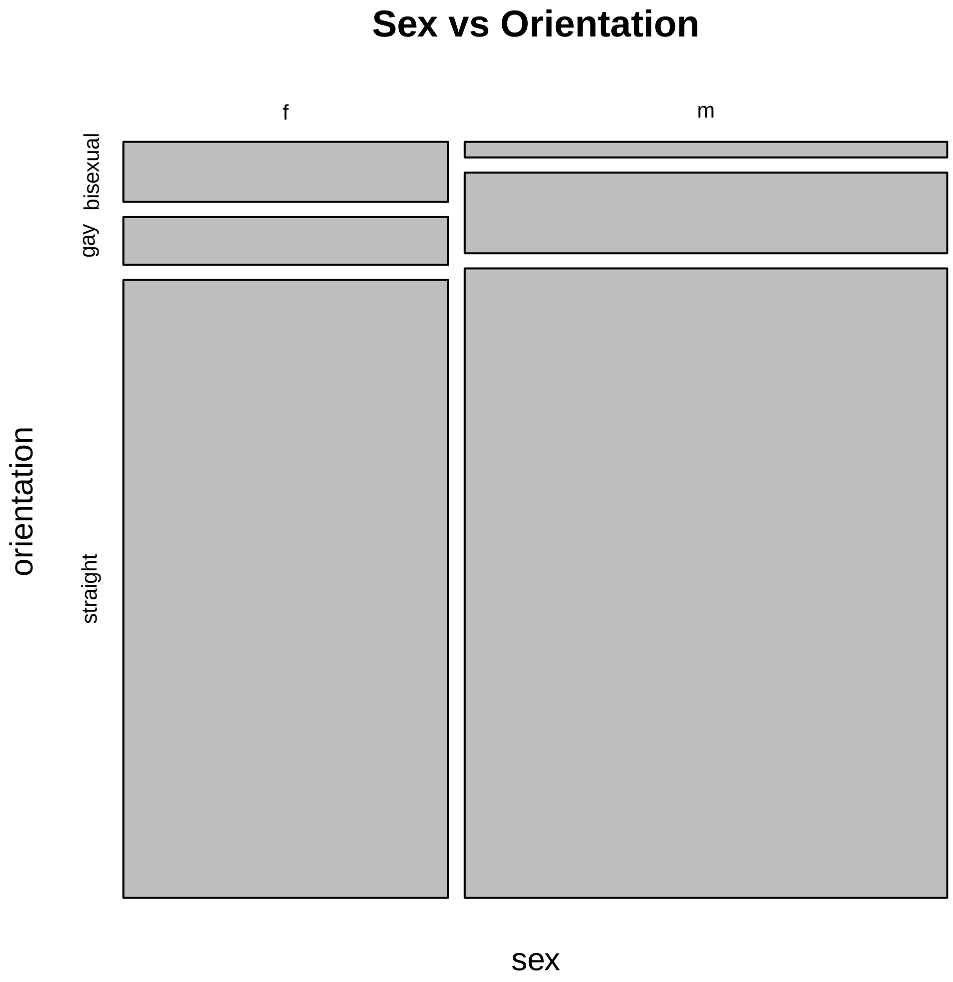

Figure 3: Cross-classification of sex and sexual orientation.

package. We then use the `tally()` illustrated in Section 3.2 to compute the distribution conditional on sex. For example, the word "book" is used by 62% of female profiles and 55% of male profiles. In Table 1 we make similar comparisons for the use of the words "travel", "food", "wine", and "beer."

```
profiles$has.book <- str_detect(essays, "book")
tally(has.book ~ sex, profiles, format='proportion')

##          sex
## has.book    f    m
##    TRUE  0.62 0.55
##    FALSE 0.38 0.45
```

word	female	male
travel	0.386	0.299
food	0.652	0.601
wine	0.201	0.117
beer	0.087	0.109

Table 1: Proportions of each sex using word in essays.

We also verify the co-occurrence of words, such as "wine" and "travel", visualizing their relationship in a mosaicplot in Figure 4.

```
profiles$has.wine <- str_detect(essays, "wine")
profiles$has.travel <- str_detect(essays, "travel")
travel.vs.wine <- tally(~has.travel + has.wine, data=profiles)
mosaicplot(travel.vs.wine, main="", xlab="travel", ylab="wine")
```

We evaluate the statistical significance of the difference in the use of the word "football" via a two-sample proportions test using the `prop.test()` function where you specify vectors `x` of the successes of each group and `n` of the numbers in each group. While the difference of around 0.5% is statistically significant for almost any α-level, it can be argued that this difference is of little practical significance.

```
profiles$has.football <- str_detect(essays, "football")
results <- table(profiles$has.football, profiles$sex)
prop.test(x=results[2, ], n=colSums(results), alternative="two.sided")

##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  x and n
## X-squared = 13, df = 1, p-value = 0.0002929
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0085 -0.0026
## sample estimates:
## prop 1 prop 2
##  0.031  0.036
```

And finally, consider the following fun exercise: we generate the top 500 words used by males and females respectively. The following code uses the "pipe" `%>%` operator from the `dplyr` package to send the output of

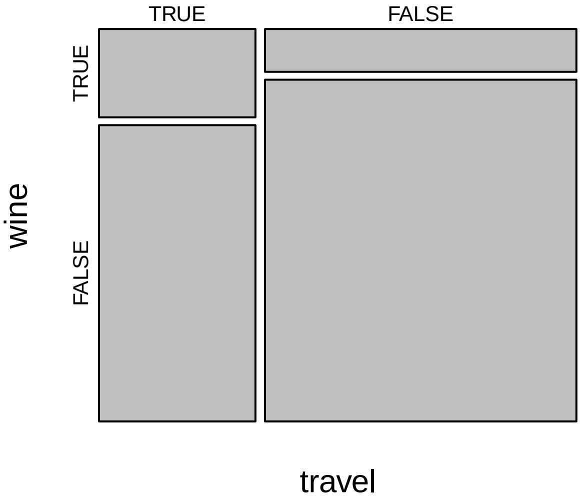

Figure 4: Co-occurrence of ‘travel’ and ‘wine.’

one function into the next function wherever the period is located. For example, the following two lines of code perform the identical task:

```
c(1, 2, 3, 4) %>% sum(.)
sum(c(1, 2, 3, 4))
```

This allows us to avoid having multiple R functions nested in a large number of parentheses and highlights the sequence of commands performed. In our case, the code below:

- Pulls the `subset` of `essays` corresponding to males and females.
- Splits up the each user's essay text at each space, i.e. cuts it up into words, using the `str_split()` command from the `stringr` package.
- Converts the list of words into a vector of words.
- Computes the frequency table using the `table()` command.
- Sorts them in decreasing order.
- Extracts the words (and not the frequency counts), which are the `names` of each element of the vector.

```
male.words <- subset(essays, profiles$sex == "m") %>% str_split(., " ") %>%
  unlist(.) %>% table(.) %>% sort(., decreasing=TRUE) %>% names(.)
female.words <- subset(essays, profiles$sex == "f") %>% str_split(., " ") %>%
  unlist(.) %>% table(.) %>% sort(., decreasing=TRUE) %>% names(.)
```

However, for both males and females, the top words are not interesting (as reflected by the top 25 below), as they include many particles such as "I", "and", and "the." Therefore, we consider the top 500 words used by males *that are not in the top 500 words used by females*, and vice-versa, by taking the difference in sets using the `setdiff()` command. Note that we didn't correct for punctuation.

```
# Top 25 male words:
male.words[1:25]

##  [1] ""                "i"               "and"
##  [4] "the"             "to"              "a"
##  [7] "of"              "my"              "in"
## [10] "i'm"             "you"             "for"
## [13] "with"            "that"            "is"
## [16] "have"            "like"            "on"
## [19] "but"             "or"              "<a"
## [22] "at"              "class=\"ilink\"" "it"
## [25] "am"

# Top 25 female words
female.words[1:25]

##  [1] ""     "i"    "and"  "the"  "to"   "a"    "my"   "of"   "in"   "i'm"
## [11] "with" "for"  "you"  "that" "have" "is"   "love" "am"   "but"  "like"
## [21] "or"   "on"   "at"   "it"   "be"

# Words in the males top 500 that weren't in the females' top 500:
setdiff(male.words[1:500], female.words[1:500])
```

```
##  [1] ","          "video"      "company"    "sports"     "/"
##  [6] "internet"   "future"     "computer"   "star"       "well,"
## [11] "well."      "away"       "john"       "until"      "business"
## [16] "us"         "type"       "couple"     "generally"  "2"
## [21] "more."      "went"       "bar"        "science"    "woman"
## [26] "work."      "started"    "does"       "here."      "found"
## [31] "three"      "lost"       "means"      "do."        "become"
## [36] "run"        "that,"
```

```
# Words in the male top 500 that weren't in the females' top 500:
setdiff(female.words[1:500], male.words[1:500])
```

```
##  [1] "loving"       "dancing,"     "love,"        "appreciate"   "dog"
##  [6] "hair"         "beautiful"    "laughing"     "passionate"   "red"
## [11] "cooking,"     ";)"           "laugh."       "please"       "kids"
## [16] "local"        "drinking"     "kinds"        "family."      "healthy"
## [21] "adventure"    "explore"      "laugh,"       "men"          "smile."
## [26] "comfortable"  "crazy"        "nature"       "hiking,"      "day."
## [31] "chocolate"    "huge"         "change"       "dating"       "sex"
## [36] "met"          "movies."
```

3.3.2 Teaching Goals and Discussions

This exercise provides students with experience performing basic text processing, mining, and analysis. Given the more advanced tools used in the last component of the exercise (the top words used by males and females), we suggest this be reserved for students with more familiarity with R. We deliberately did not pre-process the data for students to remove punctuation and HTML tags both to keep the code simple and to demonstrate the reality to students that “real” data is often very messy requiring work to clean up.

Statistical concepts include the difference between practical and statistical significance as demonstrated by the difference in proportion of males and females that used the word “football”. This can lead to discussions of what it means to conduct hypothesis tests when the sample size is as large as 59946. Furthermore, we demonstrate that simple comparisons via basic set operations can be very powerful tools yielding valuable insight. For example, the difference in lengths of words and number of adjective used by males and females in our surface-level analysis is striking.

The richness of the essay data allows for students to verify and challenge prior sociological beliefs and preconceptions using empirical data. Furthermore, such analyses can expose and confront students with biases, prejudices, and behaviors that they didn’t know they held or exhibited.

Another interesting investigation for students to pursue is to what degree do the above results hold when the groups we are comparing are further refined (grouping by sex *and* sexual orientation for example). Even bolder goals include introducing text analysis concepts such as regular expressions, inverse document frequency, natural language processing, and Latent Dirichlet analysis[18].

3.4 Predictors of Sex

The final exercise provides an opportunity to fit a predictive model for sex using logistic regression. In order to reinforce the concepts of logistic regression, we keep things simple and consider only one predictor variable in the logistic model: height. We restrict to only those users whose heights are reasonable as defined previously in Section 3.1 and take a random sample of 5995 users (10% of the data) to speed up computation using the `sample_n()` command from the `dplyr` package.

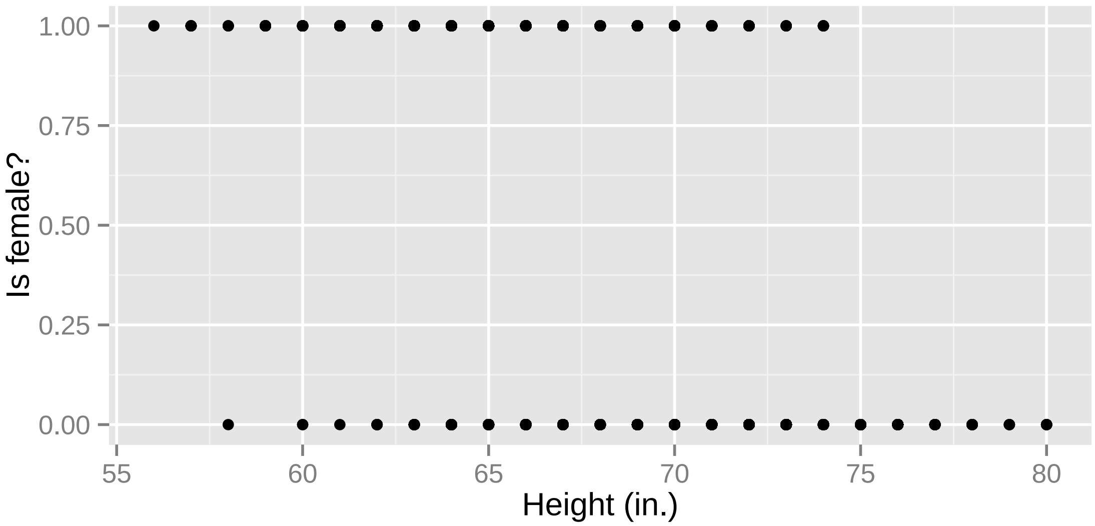

Figure 5: Female indicator vs height.

```
profiles <- filter(profiles, height>=55 & height <=80)
profiles <- sample_n(profiles, 5995)
```

3.4.1 Exercise

We convert the sex variable to a binary is.female variable, where 1 if female and 0 if male, using the ifelse() function. We plot the points as in Figure 5, making use of the ggplot2 package.

```
profiles$is.female = ifelse(profiles$sex=="f", 1, 0)
ggplot(data=profiles, aes(x=height, y=is.female)) + geom_point() +
  xlab("Height (in.)") + ylab("Is female?")
```

This plot is not very useful, as the overlap of the points makes it difficult for determine how many points are involved. We use the jitter() function to add a little random noise to each clump of points as shown in Figure 6.

```
ggplot(data=profiles, aes(x=jitter(height), y=jitter(is.female))) +
  geom_point() + xlab("Height (in.)") + ylab("Is female?")
```

We fit both linear and logistic regression models using height as the sole predictor. In order to summarize the results, we use the msummary() from the mosaic package as its output is much more digestible than the output of the standard summary() command. Furthermore, we extract the coefficients of the linear model using the coef() command.

```
linear.model <- lm(is.female ~ height, data=profiles)
msummary(linear.model)

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.42887    0.08373    76.8   <2e-16 ***
```

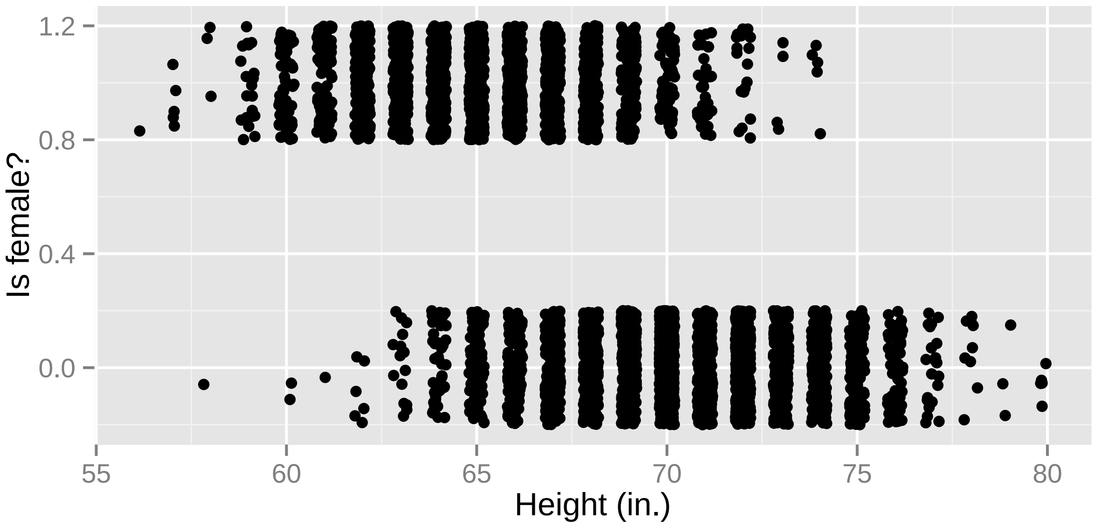

Figure 6: Female indicator vs height (jittered).

```
## height      -0.08831    0.00122   -72.1   <2e-16 ***
##
## Residual standard error: 0.36 on 5993 degrees of freedom
## Multiple R-squared:  0.465,Adjusted R-squared:  0.464
## F-statistic: 5.2e+03 on 1 and 5993 DF,  p-value: <2e-16

b1 <- coef(linear.model)
b1

## (Intercept)      height
##       6.429      -0.088
```

```
logistic.model <- glm(is.female ~ height, family=binomial, data=profiles)
msummary(logistic.model)

## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  44.9999     1.1374    39.6   <2e-16 ***
## height       -0.6705     0.0169   -39.8   <2e-16 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8075.1  on 5994  degrees of freedom
## Residual deviance: 4460.2  on 5993  degrees of freedom
## AIC: 4464
##
## Number of Fisher Scoring iterations: 6

b2 <- coefficients(logistic.model)
b2
```

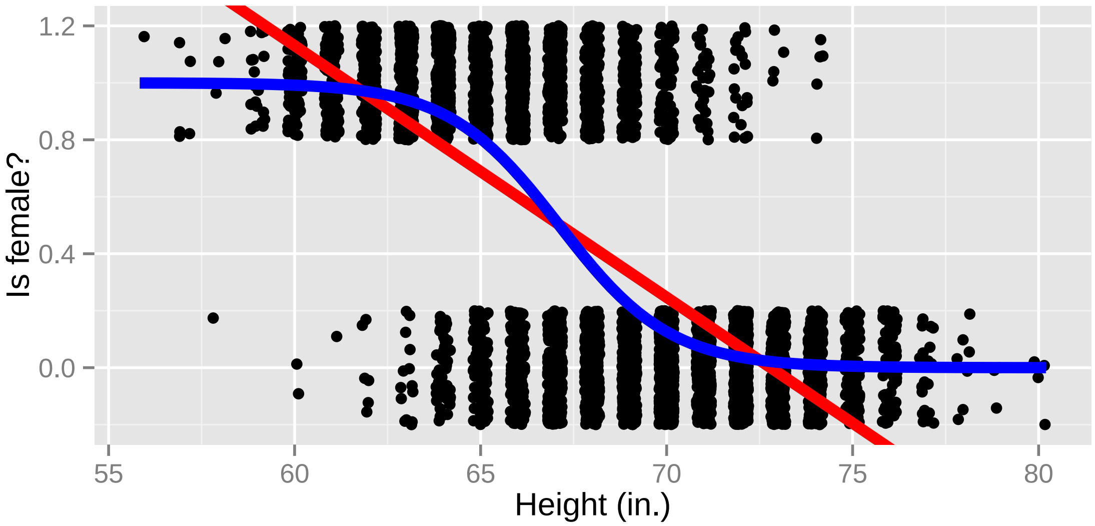

Figure 7: Linear and logistic regression curves.

```
## (Intercept)      height
##       45.00       -0.67
```

In both cases, we observe that the coefficient associated with height is negative, in other words, as height increases, the fitted probability of being female decreases. We plot both regression lines in Figure 7, with the linear regression in red and the logistic regression in blue. The latter necessitates the function `inverse.logit()` in order to compute the inverse logit of the linear equation to obtain the fitted probabilities $\widehat{p}_i$:

$$\widehat{p}_i = \frac{1}{1 + \exp\left(-(\widehat{\beta}_0 + \widehat{\beta}_1 \times \text{height}_i)\right)}$$

```
inverse.logit <- function(x, b){
  linear.equation <- b[1] + b[2]*x
  1/(1+exp(-linear.equation))
}
ggplot(data=profiles, aes(x=jitter(height), y=jitter(is.female))) +
  geom_point() + xlab("Height (in.)") + ylab("Is female?") +
  geom_abline(intercept=b1[1], slope=b1[2], col="red", size=2) +
  stat_function(fun = inverse.logit, args=list(b=b2), color="blue", size=2)
```

We observe that linear regression (red curve) yields fitted probabilities less than 0 for heights less than 61 inches for and fitted probabilities greater than 1 for heights over 73 inches, which do not make sense. This is not a problem with logistic regression as the shape of the logistic curve ensures fitted probabilities are between 0 and 1 for all heights. We therefore deem logistic regression to be a more appropriate technique for this data than simple linear regression.

However, when predicting a user's gender, just using the fitted probabilities $\widehat{p}_i$ is insufficient; a decision threshold is necessary. In other words, a point at which if the probability of being female is exceeded, we *predict* that user to be female. Looking at the histogram of fitted probabilities, we pick an appropriate

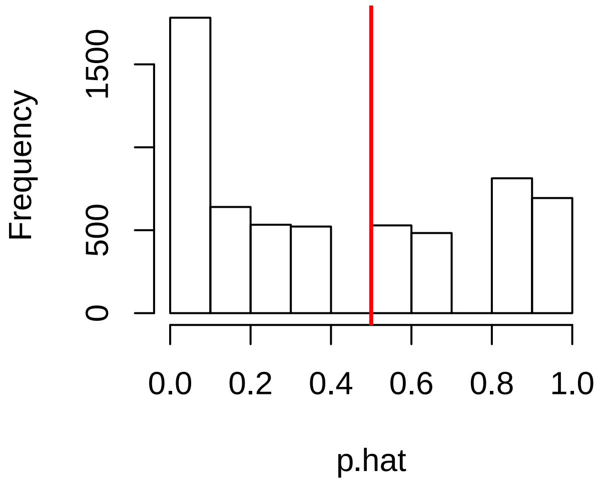

Figure 8: Fitted probabilities of being female and decision threshold (in red).

decision threshold p^* such that for all users whose $\widehat{p}_i > p^*$, we predict the user to be female. We opt for $p^* = 0.5$ and highlight this in red in Figure 8. In order to evaluate the performance of our model and our decision threshold, we produce the contingency table comparing the true and predicted values.

```
profiles$p.hat <- fitted(logistic.model)
hist(profiles$p.hat, xlab="p.hat", main="")
abline(v=0.5, col="red", lwd=2)
```

```
profiles$predicted.female <- profiles$p.hat >= 0.5
table(truth=profiles$is.female, prediction=profiles$predicted.female)

##      prediction
## truth FALSE TRUE
##     0  3024  566
##     1   452 1953
```

How did our predictions fare?

3.4.2 Pedagogical Discussion

We find that the jump from linear to logistic regression is hard for many students to grasp at first. For example, students often ask "Why the log and exp functions?" and "So we are not modelling the outcome variable Y_i, we're modeling the probability p_i that Y_i equals 1?" This exercise allows students to build up to the notion of logistic regression from the ground up using data visualizations. We also argue that on top of fitting the model and interpreting any results, students should also use the results to make explicit predictions and evaluate the logistic model's predictive power. We ask the students "For what proportion of people did you guess wrong" referring to the misclassification error rate, in this case 16.98%. Also solving for height using $p^* = 0.5$ yields a height of 67.11 inches, corresponding to 5 foot 7 inches, which is the smallest height in Figure 1 at which the proportion of males exceeds the proportion of females. This point can be highlighted to students, tying together the exercise in Section 3.1 to this one.

Further questions to ask of students include building a model with more than one predictor, incorporating essay information from Section 3.3, evaluating the *false positive rate* (the proportion of users who were predicted to be female who were actually male), evaluating the *false negative rate* (the proportion of user's who were predicted to be male who were actually female), varying the decision threshold p^*, and asking questions about out-of-sample predictions (using different data to fit and evaluate the model).

4 Conclusions

We feel that this dataset is an ideal one for use in introductory statistics and data science classes as the salience of the dataset provides students with an interesting vehicle for learning important concepts. By presenting questions to students that allow for the use of their background knowledge, whether it be from the news, stereotypes, sociological knowledge, students are much better primed to absorb statistical lessons. Furthermore,

1. The data consists of a rich array of categorical, ordinal, numerical, and text variables.

2. This is an instance of real data that is messy, requires much data manipulation, has many suspicious values, and includes categorical variables of a complicated nature (for instance, there are 218 unique responses to the ethnicity variable). This reinforces to students that time and energy must be often invested into preparing the data for analysis.

3. The dataset is of modest size. While $n = 59946$ is not an overwhelmingly large number of observations, it is still much larger than typical datasets used in many introductory probability and statistics classes.

All the files, including the original data and the R Sweave `.Rmd` file used to create this document, can be found at `https://github.com/rudeboybert/JSE_OkCupid`. Note that the file `profiles.csv.zip` must be unzipped first. All R code used in this document can be outputted into an R script file by using the `purl()` command in the `knitr` package on the `JSE.Rnw` R Sweave document.

```
purl(input="JSE.Rnw", output="JSE.R", quiet=TRUE)
```

5 Acknowledgements

First, we thank OkCupid president and co-founder Christian Rudder for agreeing to our use of this dataset (under the condition that the dataset remains public). Second, we thank Everett Wetchler of Bayes Impact `everett.wetchler@gmail.com` for providing the data; the python script used to scrape the data can be found at `https://github.com/evee746/okcupid`. Finally, we thank the reviewers for their helpful comments.

Albert Y. Kim
Mathematics Department
Reed College
3203 SE Woodstock Blvd
Portland, OR 97202
albert.kim@reed.edu

Adriana Escobedo-Land
Reed College
3203 SE Woodstock Blvd

Portland, OR 97202
escobad@reed.edu

References

[1] Bin Yu. IMS presidential address: let us own data science. *IMS Bulletin*, 43(7):1, 2014.

[2] Marie Davidson. Aren't we data science? *AMSTATNEWS*, July 2013.

[3] Hadley Wickham. How are data science and statistics different? *IMS Bulletin*, 43(6):7, 2014.

[4] GAISE College Group. Guidelines for assessment and instruction in statistics education. Technical report, American Statistical Association, Alexandria, VA, 2005.

[5] Deborah Nolan and Duncan Temple Lang. Computing in the statistics curricula. *The American Statistician*, 64(2):97–107, 2010.

[6] Robert Gould. Statistics and the modern student. *International Statistics Review*, 78(2):297–315, 2010.

[7] Randall Pruim, Daniel Kaplan, and Nicholas Horton. *mosaic: Project MOSAIC (mosaic-web.org) statistics and mathematics teaching utilities*, 2014. R package version 0.9.1-3.

[8] Hadley Wickham and Romain Francois. *dplyr: dplyr: a grammar of data manipulation*, 2014. R package version 0.2.

[9] Hadley Wickham. *stringr: Make it easier to work with strings.*, 2012. R package version 0.6.2.

[10] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[11] How I hacked online dating. http://www.ted.com/talks/amy_webb_how_i_hacked_online_dating. Accessed: 2015-01-15.

[12] How a math genius hacked okcupid to find true love. http://www.wired.com/wiredscience/2014/01/how-to-hack-okcupid/. Accessed: 2015-01-15.

[13] OkTrends: Dating Research from OkCupid. http://blog.okcupid.com/.

[14] Christian Rudder. *Dataclysm: Who We Are When We Think No One's Looking*. Crown, 2014.

[15] Did the mathematician who hacked OKCupid violate federal computer laws? http://pando.com/2014/01/22/did-the-mathematician-who-hacked-okcupid-violate-federal-computer-laws/. Accessed: 2015-03-27.

[16] The biggest lies in online data. http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating. Accessed: 2015-03-26.

[17] Algorithmic illusions: Hidden biases of big data. https://www.youtube.com/watch?v=irP5RCdpilc. Accessed: 2015-01-15.

[18] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.