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

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

Abstract

1 Introduction

In light of the field of data science generating more interest in academia, industry, and government, many prominent statisticians are arguing that statistics needs to stake a bigger claim in data science or risk marginalization in favor of other disciplines such as computer science and computer engineering [1, 2]. While the 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 technogloy for both developing conceptual understanding and analyzing data [4, 5]. Relatedly, in the hopes of making more relevant introductory undergraduate statistics courses, 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 an approach are the data sets considered, as they provide the scientific context of the analyses and thus will ultimately drive student interest [6].

It is in light of such discussions that we present this paper centering on user profile data scraped from the online dating website OkCupid. Specifically, we describe the data set and present example analyses along with corresponding discussions of their pedagogical value given the aforementioned needs in statistics education. The analyses presented in this paper were used in a variety of settings: as examples and exercises in introductory statistics classes and in statistical software tutorials. However, we present the material in a manner best suited for use in a short introductory course to data science course using the R statistical software package [7]. This course provides students a glimpse of what data science is like, in particular the difficulties associated with using real and unprocessed data and the use computational tools and concepts such as data manipulation, data visualization and text mining.

This paper does assume that students have.

2 Data

The data consists of the public profiles of 59,946 OkCupid users. Each of these users had active profiles on June 26th 2012, were living within 25 miles of San Francisco, were online in the previous year and had at least one picture in their profile. Data was scraped using a Python script from users' public profiles on June 30th 2012; any non-publicly facing information such as messaging and login behavior 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). Futhermore, text strings of the 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`.

Analysis of similar data has received much press of late, including Amy Webb's TED talk "How I Hacked Online Dating" [8] and Wired magazine's "How a Math Genius Hacked OkCupid to Find True Love" [9]. OkCupid co-founder Christian Rudder pens periodical analysis of their data via the OkTrends blog [10] and has recently published a book "Dataclysm: Who We Are When We Think No One's Looking" describing similar data driven insights [11]. Such publicity surrounding data-driven online dating and the salience of dating matters among students makes this dataset one with great pedagogical usefulness. The possible questions one could answer with this dataset can increase the captivity of the audience.

Before we continue we note that even though this data only consists of what once was publicly facing material, one should proceed with caution before using data in fashion similar to ours. Even though the profiles are public, the Computer Fraud and Abuse Act (CFAA) makes it a federal crime to access a computer without authorization from the owner [12]. In our case, permission was given by the owners of the data (See Section 5).

3 Example Analyses

We describe five example analyses presented to students in a variety of settings, including: an 90 minute introductory tutorial on R, an introductory probability and statistics class (consisting of chiefly math, biology, and economics majors), and a follow-up two-hundred level data science class title ”Case Studies in Statistical Analysis”.

The example analyses we used are described as 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 sex and sexual orientation differences in what terms users use in the responses to the 10 essay questions?

4. Can we predict a users' sex using their profile information?

3.1 Male and Female Heights

3.1.1 Exercise

We compare the distribution of OkCupid user's heights using histograms. Height is one of 3 numerical variables in this dataset (the other being height and income). This provides us an opportunity to investigate numerical summaries using the `favstats()` command from the `mosaic` package [13]:

```
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 deems heights between 55 and 80 inches to be reasonable, and while there is potential bias in discarding these cases, of the 59946 users, since there are 39 and only 78 users who would be discarded, we proceed safely. We remove these users from the dataset either using the `subset()` command in base R or using the `filter()` command from `dplyr` package [14]. Note: the following commands are equivalent:

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

We proceed by comparing 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, we rather

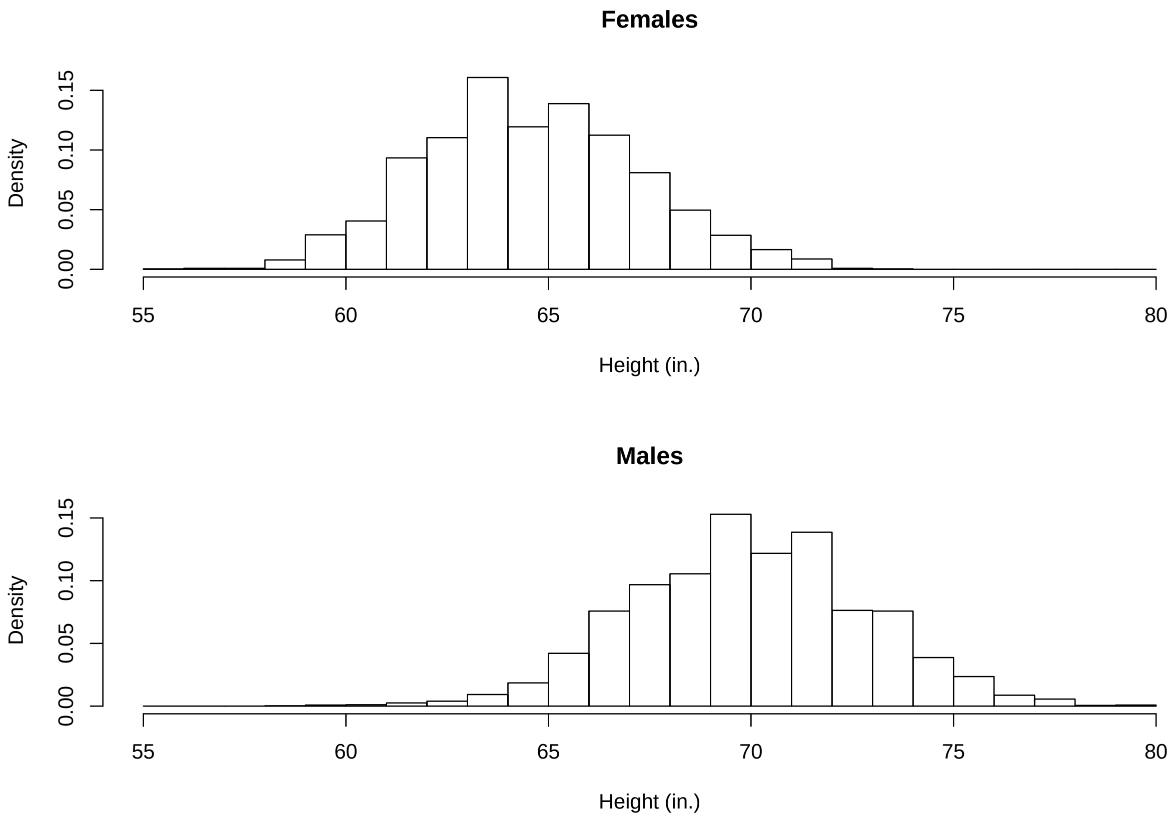

Figure 1: Histograms of User Heights Split by Sex

1. Plot them with binwidths 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 the density and not frequency using `prob=TRUE`) to fascilitate comparisons between the two distributions

We demonstrate this in Figure 1.

```
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 an instructor can ask at what point does a height no longer become reasonable and what impact does their removal have on their removal?

Secondly, this exercise demonstrates the power of a simple data visualization and hence emphasizes the importance of putting careful thought into their construction. The better the construction, the better the conveying of the insight. In our case, while having students plot two histograms simultaneously on the same scale to demonstrate that males have on average greater height may seem to be a rather pedantic goal, 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 easily 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"[15] to show they've replicated (on a smaller scale) a previous analysis done by OkCupid and then show the other examples of analysis of the blog.

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's population 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. First, we perform a basic exploratory data analysis on these variables using barcharts in Figure 2 using the code below:

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

However, when it comes to dating we can't just consider the **marginal distributions** (i.e. the margins of a table) of these variables, we must consider their cross-classification i.e. their **joint distribution**, and **conditional**. Only then will we have a more accurate picture of the dating landscape. 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 visualize this breakdown via a mosaicplot (as shown in Figure 3).

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

##             sex
## orientation     f     m
##    bisexual 0.078 0.018
##    gay      0.065 0.110
##    straight 0.857 0.871

sex.by.orientation <- with(profiles, table(sex, orientation))
mosaicplot(sex.by.orientation/n, main="Sex vs Orientation")
```

We now proceed with the *chi-square tests for independence* to perform one using the `chisq.test()` command.

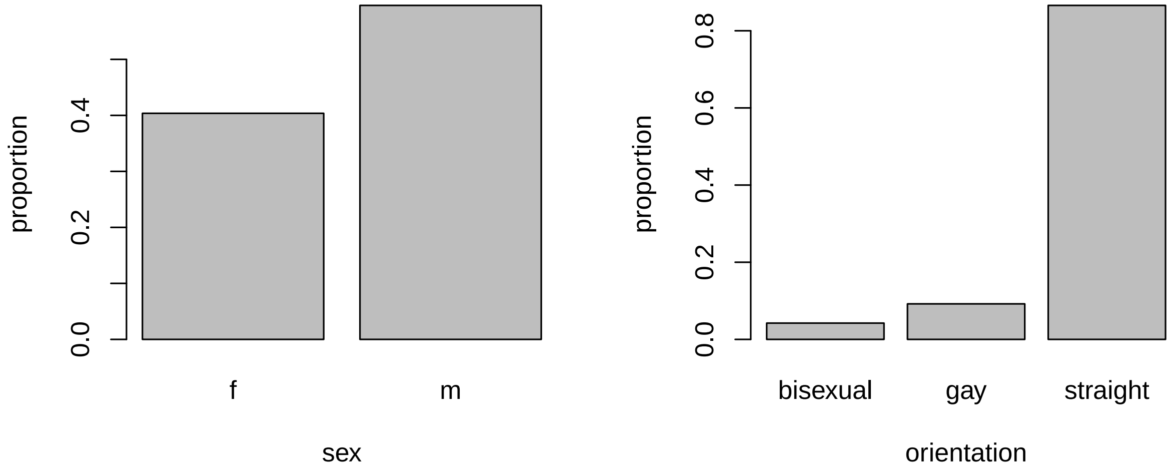

Figure 2: Distributions of Sex and Sexual Orientation.

```
chisq.test(sex.by.orientation)

##
##  Pearson's Chi-squared test
##
## data:  sex.by.orientation
## X-squared = 154, df = 2, p-value < 2.2e-16
```

3.2.2 Pedagogical Discussion

This exercise was an opportunity to concretize the statistical notions such as marginal / joint / conditional distribtions, sampling bias, and “big data.”

he 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 are well aware of the gender imbalance issues in the Technology sector, the rise of the San Francisco Bay Area as a formidabe tech capitol, and San Francisco’s history of being a bastion community for gay men in the recent past. An interesting discussion arose on the sociological factors in why more women listed themselves as bisexual then men; this topic has previously been covered on the blog OkTrends[15].

Students could make statement that are conditional probability statements such as “Of the female population, $X\%$ were Y’s and some could even recognize the etiomogical origins of the term ”marginal distribution” as their counts are presented in the margins of the table.

The question of generalizability was presented in an introductory probability and statistics class homework. Almost all students were able to recognize the non-randomness of the sampling and hence the results are non-generalizable. For example they recognized that OkCupid’s demographic is most likely different than other demographics, such as match.com (which is not free) or christiansingles.com (which is targeted towards Christians), or less technology-savvy individuals.

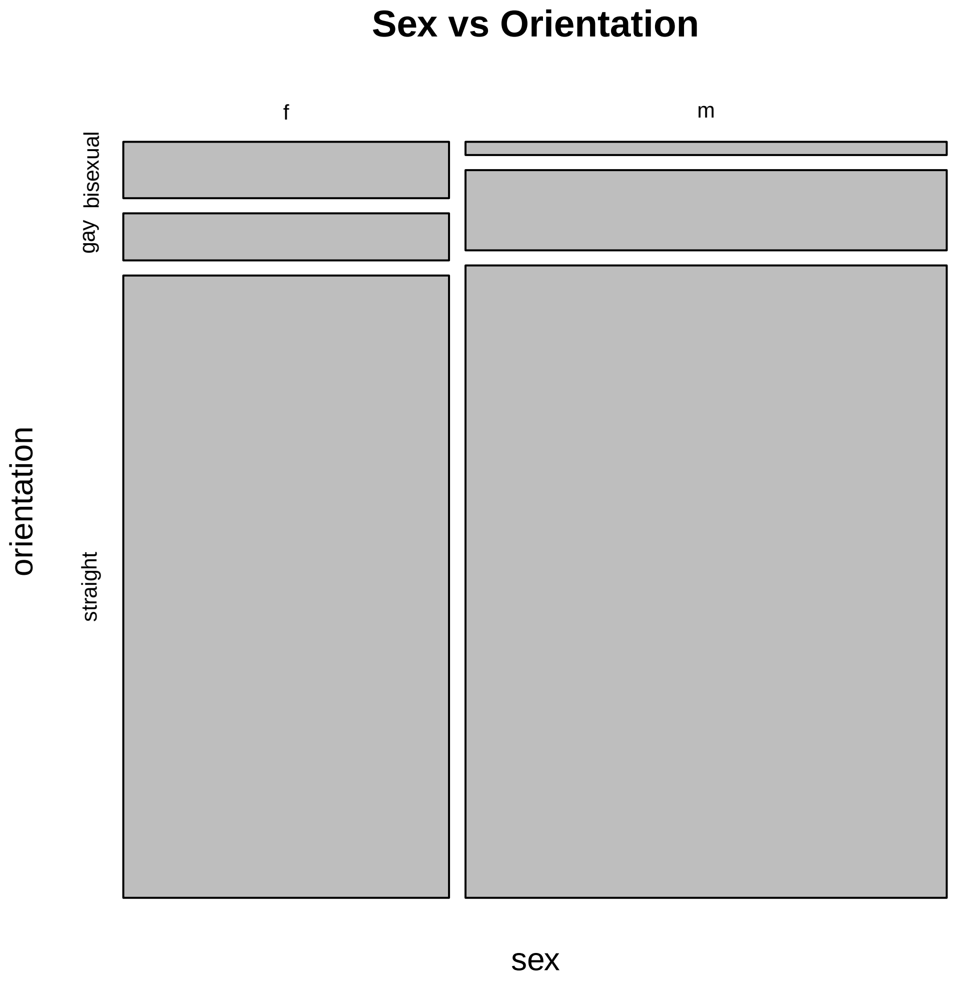

Figure 3: Cross-classification of sex and sexual orientation.

So while 59946 users may seem like a fairly large sample to many students, we emphasized that one must be wary of self-selection biases that can affect the generalizability of the results. This proved an excellent segue to discussions on John Tukey's famous quote to the sexologist Alfred Kinsey "A random selection of three people would have been better than a group of 300 chosen by Mr. Kinsey." and Kate Crawford of Microsoft Research's YouTube talk "Algorithmic Illusions: Hidden Biases of Big Data" [16].

Further questions one can pose to students include "Which dating demographic would you say has it the best and worst in terms of our simplied categorization?", "What variable do you think should be incorporated next after sex and sexual orientation in faithfully representing the OkCupid landscape?", "Are we sure we can't *pratically* assume generalizability of the results to Brooklyn OkCupid?".

3.3 Text Analysis

3.3.1 Exercise

The next exercise focuses on the responses to the essay questions, providing you with an opportunity to perform text analysis. Manipulating text data in R is often a complicated affair, so we present some code you should not worry if it doesn't make sense at first, as it is for more skilled coders. The following code

- Uses the `select()` command from the `dplyr` package to select the 10 essay columms as identified by the fact they `starts_with("essay")`.
- Converts the essays from list to matrix format.
- For each user concatenates the 10 columns to form a single string. The code applies the function `paste(x, collapse=" ")` to every essay where `x` is the set of 10 essay responses, and the `paste()` function collapses `x` across columns while separate 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 line breaks (`\n`) and paragraph breaks (`<br />`) with spaces as well to make the outputs more readable.

The output is a single vector `essays` that contains all 10 essay responses for each user concatenated together.

```
essays <- select(profiles, starts_with("essay"))
essays <- matrix(unlist(essays), ncol = ncol(essays), byrow = FALSE)
essays <- apply(essays, 1, paste, collapse=" ")
essays <- str_replace_all(essays, "\n", " ")
essays <- str_replace_all(essays, "<br />", " ")
```

We ask: Do men and women OkCupid users use words at different rates in their essay responses? We can do this using the `str_detect()` command in the `stringr` package to search essays for the word in quotation marks and the `tally()` command from the `mosaic` package to compute the conditional (on sex) distributions. For example, the word "book":

```
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
```

We test out various preconceptions of word use between any of the categorical variables: sex, sexual orientation, etc. For example, in Table **??** we compare the proportion of use of words in essays split by sex.

We also verify the co-occurence of words:

query	female	male
travel	0.387	0.293
food	0.655	0.602
wine	0.195	0.116
beer	0.088	0.105

Table 1: Proportion of use of words in essays split by sex.

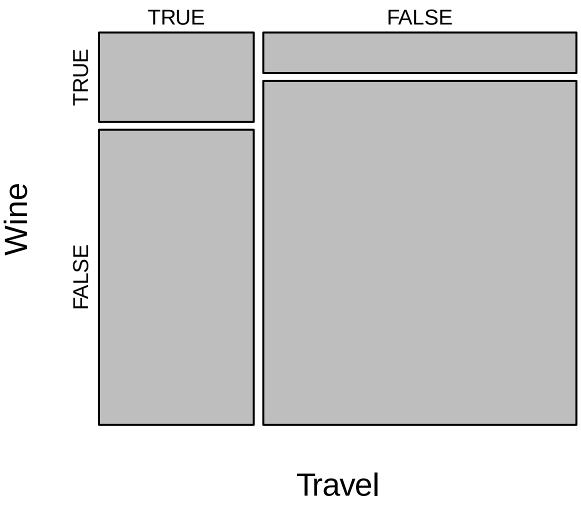

Figure 4: Co-occurrence of ’Travel’ and ’Wine’

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

##           has.wine
## has.travel TRUE FALSE
##      TRUE   461  1522
##      FALSE  423  3589

mosaicplot(tally(~has.travel + has.wine, data=profiles), main="", xlab="Travel",
           ylab="Wine")
```

And finally 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 `x` the successes of each group and `n` the numbers in each group. While the difference of around 0.5% is statistically significant for almost any α-level, it could 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 = 2, df = 1, p-value = 0.1546
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0166  0.0024
## sample estimates:
## prop 1 prop 2
##  0.030  0.037
```

And finally, consider the following fun exercise: the words in the male's top 500 that weren't in the women's top 500 and vice-versa.

```
male.words <- str_split(essays[profiles$sex =="m"], " ") %>% unlist() %>% table() %>%
  sort(decreasing=TRUE) %>% names()
female.words <- str_split(essays[profiles$sex =="f"], " ") %>% unlist() %>% table() %>%
  sort(decreasing=TRUE) %>% names()
# Words in the males top 500 that weren't in the women's top 500
setdiff(male.words[1:500], female.words[1:500])

##  [1] "words"    "his"      "company"  ","        "video"    "sports"  
##  [7] "else"     "computer" "/"        "john"     "future"   "world."  
## [13] "internet" "well,"    "run"      "three"    "bar"      "work."   
## [19] "short"    "away"     "business" "until"    "side"     "us"      
## [25] "found"    "went"     "science"  "type"     "here."    "isn't"   
## [31] "started"  "stuff."   "career"   "well."    "more."    "show,"   
## [37] "serious"  "star"     "said"     "to."      "women"    "couple"  

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

##  [1] "loving"     "hair"       "dog"        "seeing"     "dancing,"  
##  [6] ";)"         "appreciate" "kinds"      "laughing"   "red"       
## [11] "passionate" "hate"       "laugh."     "love,"      "beautiful" 
## [16] "kids"       "david"      "she"        "please"     "crazy"     
## [21] "change"     "2"          "laugh,"     "hiking,"    "smile."    
## [26] "men"        "adventure"  "except"     "sex"        "met"       
## [31] "passion"    "willing"    "does"       "become"     "local"     
## [36] "nature"     "healthy"    "myself."    "you'll"     "beach"     
## [41] "dating"     "human"     
```

3.3.2 Teaching Goals and Discussions

This exercise provides students with experience performing basic text processing, mining, and analysis. The messiness of the data is once against exemplified by the precense of HTML tags in the text data. This messiness is both liability, as only students who are somewhat familiar with R will feel comfortable, as it is asset, as it confronts students with the fact that data-cleaning is a large part of the work.

The nature of the data allows for students to judge prior sociological beliefs and preconceptions using data. Many students were confronted, even with these simple analyses, with the fact that there is still evident gendered language being used in profile essays, and further investigation is warranted.

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. Why it was important to show which words were being used by men *but not by women* as individually both groups will have as most frequently used words that are mostly particles like ”the”, ”and”, and ”or”.

Further avenues of inquest include asking “What are the implications of performing a large number of these proportions test on the essays?”, “Can we predict demographic attributes of users based on word use?” “For what words do we see the biggest differences in use between the three categories of sexual orientation?”. An even bolder goal include introducing concepts, such as constructing wordclouds, inverse document frequency, natural language processing, and Latent dirichlet analysis[17].

3.4 Predictors of Sex

The final exercise provides an opportunity to fit a predicitive model for sex using logistic regression. In order to reinforce the concepts of logistic regression, we only consider using one predictor variable in the logistic model: height. Given our observations in Section 3.1, we restrict to only those users whose heights are “reasonable”.

```
profiles <- filter(profiles, height>=55 & height <=80)
profiles <- sample_n(profiles, n*0.1)

## Error:  Sample size (5994.6) greater than population size (5983).  Do you want replace =
TRUE?
```

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 **??**, 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?")
```

We question the usefulness of this plot, 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 **??**.

```
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. Furthermore, we use the `msummary()` from the `mosaic` command to obtain the regression summary as its output is much more digestible than the output of the `summary()` command. Furthermore, we extract the coefficients of the linear model using the `coef()` command.

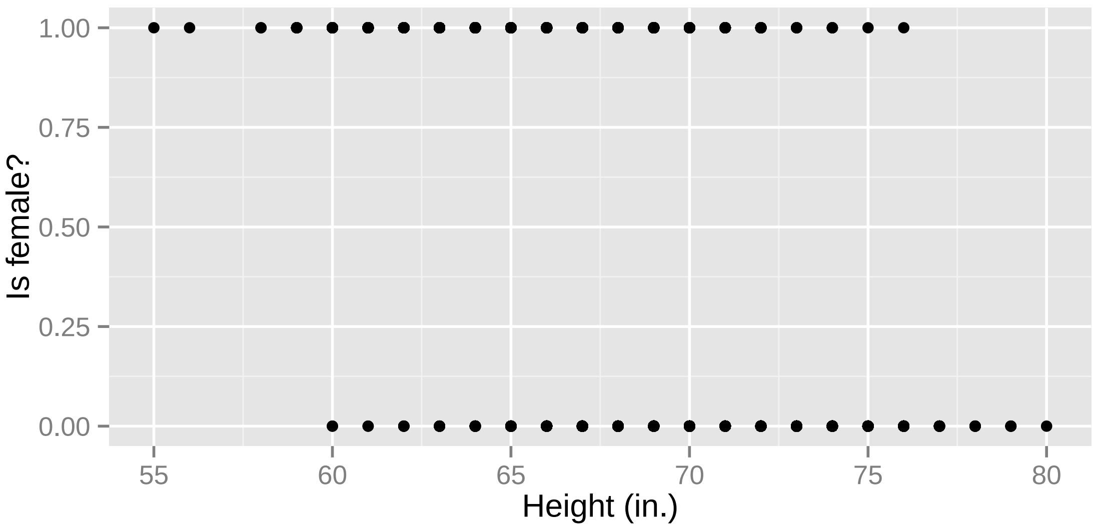

Figure 5: Female Indicator vs Height

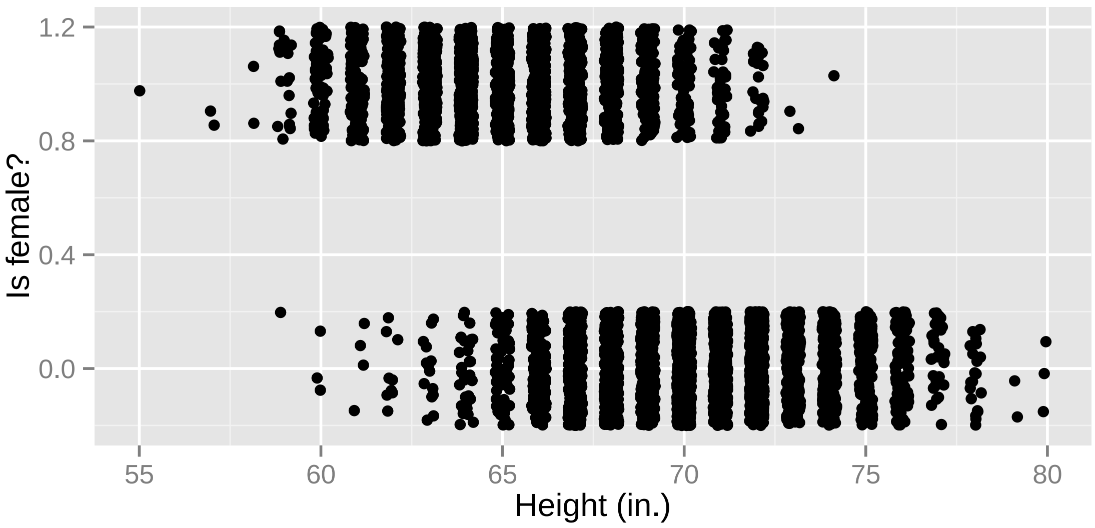

Figure 6: Female Indicator vs Height

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

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.3836     0.0819    77.9   <2e-16 ***
## height       -0.0876     0.0012   -73.1   <2e-16 ***
##
## Residual standard error: 0.36 on 5982 degrees of freedom
## Multiple R-squared:  0.472,Adjusted R-squared:  0.472
## F-statistic: 5.34e+03 on 1 and 5982 DF,  p-value: <2e-16

b1 <- coef(linear.model)
```

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

## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  45.5213     1.1566    39.4   <2e-16 ***
## height       -0.6778     0.0171   -39.6   <2e-16 ***
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8075.5  on 5983  degrees of freedom
## Residual deviance: 4381.5  on 5982  degrees of freedom
## AIC: 4386
##
## Number of Fisher Scoring iterations: 6

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

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. However the interpretability of the logistic model is difficult. We thus plot both regressions lines, with the linear regression in red and the logistic regression in blue. The latter necessitates a function for the inverse logit of the linear equation to obtain:

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

```
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

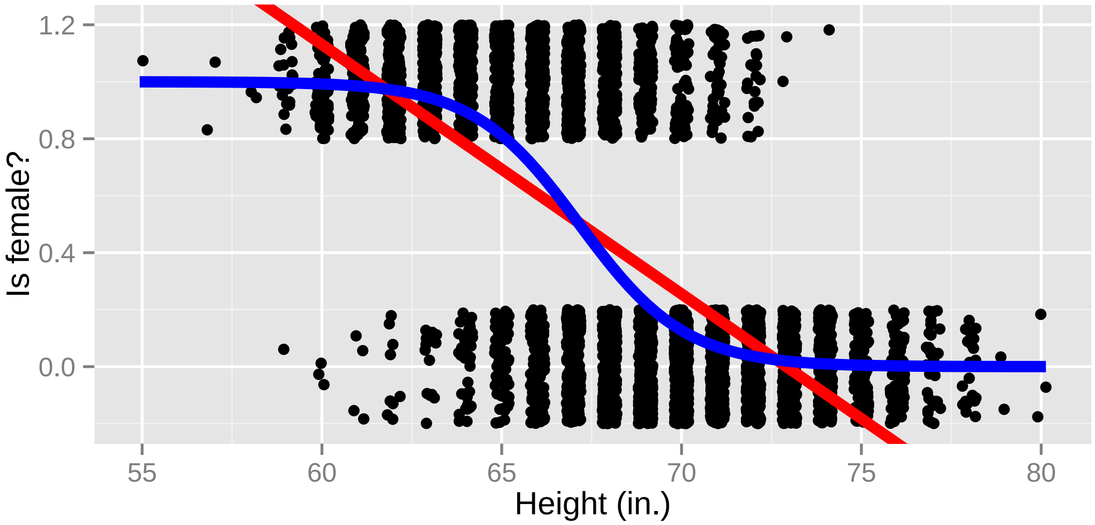

Figure 7: Female Indicator vs Height

not a problem with logistic regression as the inverse-S shaped of the logistic curve ensures fitted probabilities are between 0 and 1 for all heights. We therefore deem logistic regression to be the appropriate technique for this data.

However, when predicting a user's gender, just using the fitted probabilities $\widehat{p}_i$ are insufficient; a decisions threshold is necessary. In other words, a point at which if the probability of being a woman is exceeded, we declare that user to be female.

Looking at the histogram of fitted probabilities, we pick an appropriate decision threshold p^* such that for all users who's $\widehat{p}_i > p^*$ we predict the user to be female. We opt for $p^* = 0.5$ and highlight this in red in Figure **??**. In order to evaluate the performance of our model and our decision threshold, we produce the contingency table of the truth and predicted values to evaluate the model performance.

```
profiles$p.hat <- fitted(logistic.model)
hist(profiles$p.hat, xlab="p.hat", main="Fitted Probabilities of Being Female")
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  3014  550
##     1   449 1971
```

3.4.2 Pedagogical Discussion

We find that the jump from linear to logistic regression is really hard for students to grasp. For example, students often ask "Why the log and exp functions?", "So we're not modelling Y_i the outcome variable, we're modeling the probability p_i that it equals 1?", "Why does the linear equation model the odds and not p_i or Y_i". This exercise provides students to build up to the notion of logistic regression from the ground up and

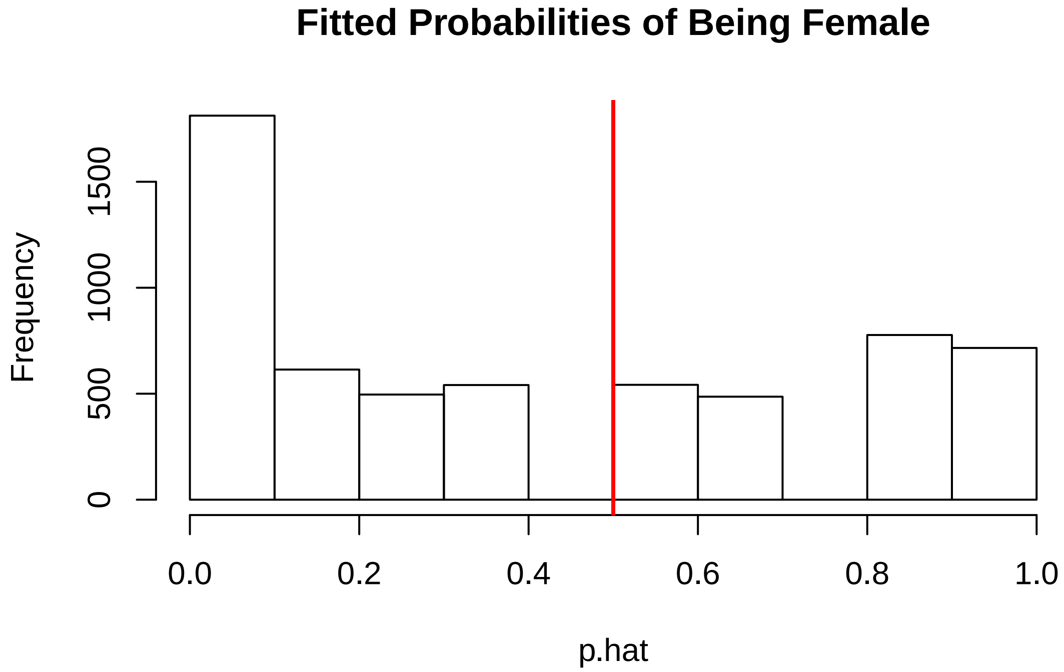

Figure 8: Female Indicator vs Height

using data visualizations. This example does however require familiarity with the `ggplot2` package. We feel this is a tool any data scientist primarily using `R` should use, and have incorporated it heavily into our data science class.

We also argue that it is insufficient to merely fit the model and interpret any regressions, we feel that students benefit by also having predictions explicitly made and evaluating the logistic model's predictive power. Wee ask the students "For what proportion of people did you guess wrong" referring to the misclassification error rate, in this case 16.69%.

Further investigations including more than one predictor? Perhaps using essay based information like in Section 3.3? What is our *false positive rate* i.e. the proportion of user's who were predicted to be female who were actually male? What is our *false negative rate* i.e. the proportion of user's who were predicted to be male who were actually female? If you wanted to be sure you predicted all women correctly, what decision threshold p^* should you use?

4 Conclusions

We feel that this dataset is an ideal one for introducing students to data science as it is

1. The data consists of an array of categorial, ordinal, numerical, and text variables.
2. This is an instance of real data that is messy and thus requiring data-wrangling, 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.

We argue that given the salience of the dataset to students, this dataset provides a very fruitful channel to introduce statistical and data sciences to students of vary undergraduate levels. By presenting questions to students that allow for the use of background knowledge of the problem, whether it be the news, stereotypes, sociological knowledge, students are much better primed to absorb statistical lessons. Hence, we took a ”lead by the hand” approach to the tone of the exercises. Other statistics education papers taking a similar approach to ours but using different data include BLAH

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 outputed into an R script file by using the `purl()` command in the `knitr` package on the `JSE.Rnw` R Sweave document found at .

```
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] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

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

[9] 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.

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

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

[12] 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.

[13] 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.

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

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

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

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