Final Project for MPA 635

Christopher Law

October 21, 2017

Executive Summary

This is an analysis of over 500,000 reviews of Eurpean hotels. I'm presenting this to hotel owners that want to know how to focus their effors. I found some interesting data showing that there are large differences in nationalities and how they perceive hotel quality. That was probably my biggest finding. I also found some trends in the positive and negative reviews. My final Illustrator graphic is completely miserable because I really underestimated how much time it takes to make that part look nice. Go easy on me, please!

Data Background

This comes from Bookings.com. I didn't have to do much tidying - it was already pretty clean. According to kaggle, the data was scraped directly from the website. I wish it had other information such as a country of the hotel. Also, the data covers only 8 months of a single year. I was hoping to have a longer span of time so I could look at temporal trends.

Data cleaning

As you can see from the rmarkdown code, I used a healthy amount of group_by, select, filter, summarize, and other parts of the tidyverse. I also used tidytext to analyze the reviews.

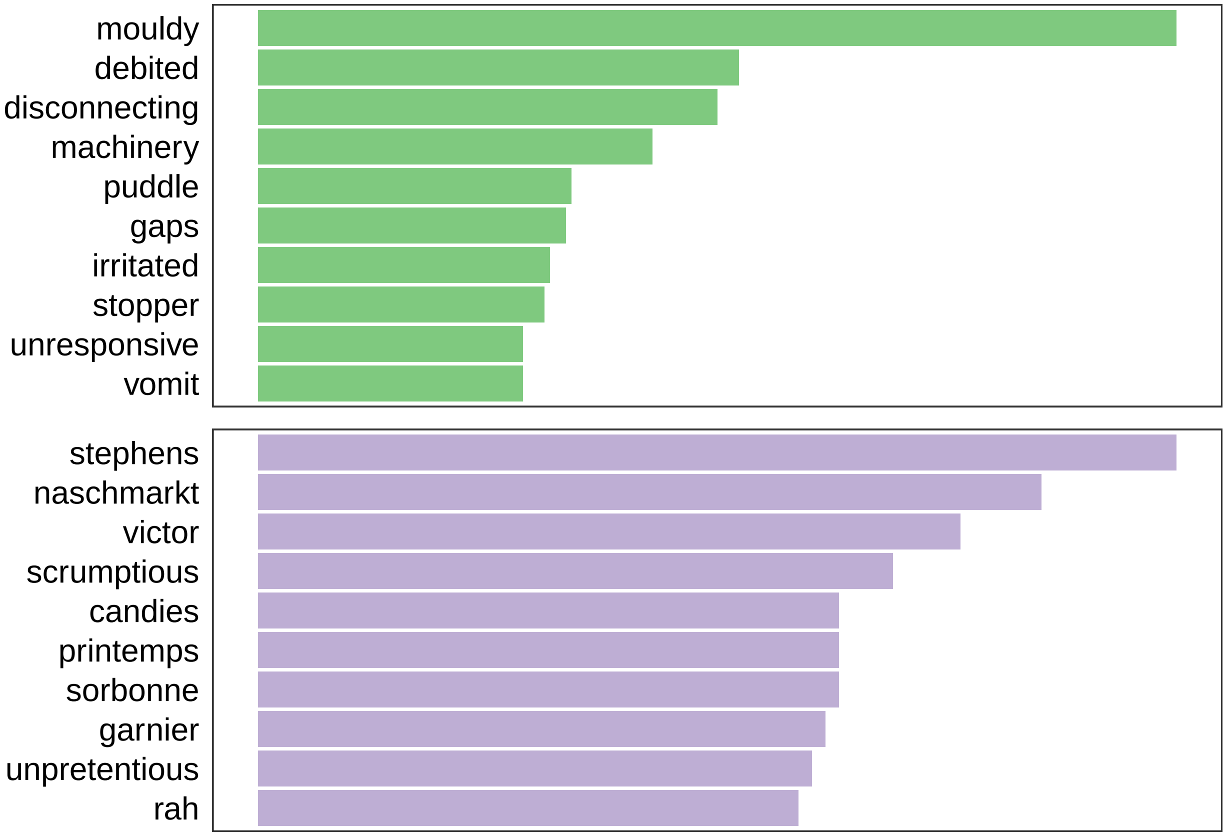
```r
library(tidyverse)
library(tidytext)
library(forcats)
library(extrafont)
library(RColorBrewer)

hotels <- read_csv("data/Hotel_Reviews.csv")
```

Individual figures

Figure 1

```r
gathered <- hotels %>%
  select(Positive_Review, Negative_Review) %>%
  gather(review_type, text) %>%
  unnest_tokens(word, text) %>%
  count(review_type ,word, sort = TRUE) %>%
  bind_tf_idf(word, review_type, n) %>%
  group_by(review_type) %>%
  top_n(10, tf_idf) %>%
  mutate(word = fct_rev(fct_inorder(word, ordered = TRUE)))
```

```r
pos_neg_words <- ggplot(gathered, aes(x = word, y = tf_idf, fill = review_type))+
  geom_col(show.legend = FALSE) +
  scale_fill_brewer(palette = "Accent") +
  facet_wrap(~ review_type, scales = "free", nrow = 2) +
  labs(x = NULL, y = NULL, title = NULL) +
  theme_bw() +
  theme(strip.text.x = element_blank(),
        axis.text.x = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text.y=element_text(size = 12, colour = "black"),
        axis.ticks = element_blank(),
        panel.background = element_rect(fill = "transparent", colour = NA),
        plot.background = element_rect(fill = "transparent", colour = NA)) +
  coord_flip()
pos_neg_words
```

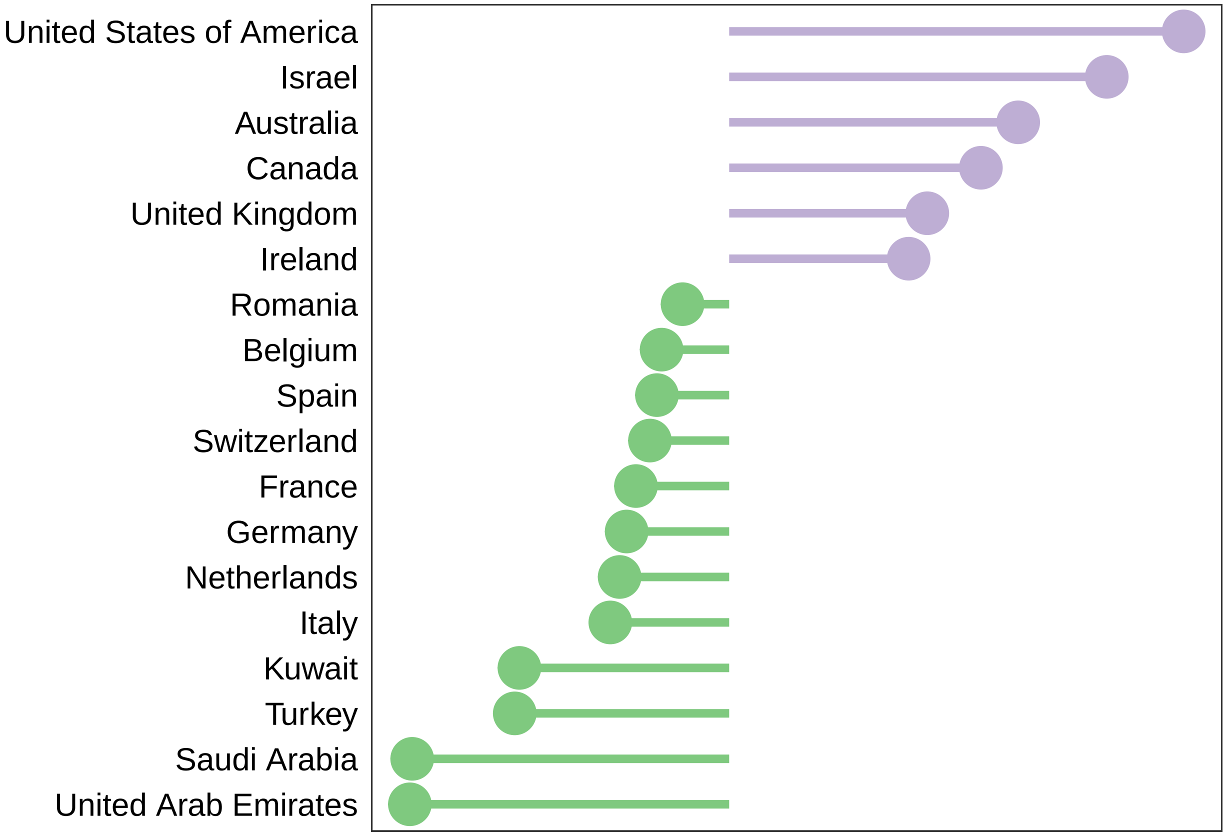


```r
ggsave(pos_neg_words, filename = "images/pos_neg_words.pdf", width = 7, height = 6)
```

Thanks for helping me on this one. I had a hard time making the right data frame so I could analyze the words like I wanted to. One frustration that I have with this chart is that I didn't like how the bars are bigger toward the top. I would rather that the bars were bigger in the middle and that the outside had the smaller values. I spent too much time trying to figure that out. There are no values for the x-axis becuase of the tf_idf measurement - it's basically meaningless.

Figure 2

```
lots_reviews <- hotels %>%
  group_by(Reviewer_Nationality) %>%
  summarize(m = mean(Reviewer_Score), n = n()) %>%
  ungroup() %>%
  filter(n>4000) %>%
  arrange(m)

sd_lots_reviews <- sd(lots_reviews$m)
mean_lots_reviews <- mean(lots_reviews$m)

reviews4000 <- lots_reviews %>%
  mutate(z_score = round((m - mean_lots_reviews)/sd_lots_reviews,2)) %>%
  mutate(cust_type = ifelse(z_score < 0 , "Tough Customers", "Golden")) %>%
  mutate(cust_type = fct_inorder(cust_type, ordered = TRUE)) %>%
  mutate(Reviewer_Nationality = fct_inorder(Reviewer_Nationality, ordered = TRUE))


country_revs <- ggplot(reviews4000, aes(x = Reviewer_Nationality, y = z_score,
                                        color = cust_type))+
  geom_point(stat = 'identity', size = 7) +
  geom_segment(aes(y = 0,
                   x = Reviewer_Nationality,
                   yend = z_score,
                   xend = Reviewer_Nationality,
               color = cust_type), size = 1.5) +
  scale_color_brewer(palette = "Accent") +
  coord_flip() +
  theme_bw() +
  labs(x = NULL, Y = NULL) +
  theme(strip.text.x = element_blank(),
        axis.title = element_blank(),
        axis.text.x=element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.text.y=element_text(size = 12, colour = "black"),
        axis.ticks = element_blank(),
        panel.background = element_rect(fill = "transparent", colour = NA),
        plot.background = element_rect(fill = "transparent", colour = NA),
        legend.position="none")
country_revs
```

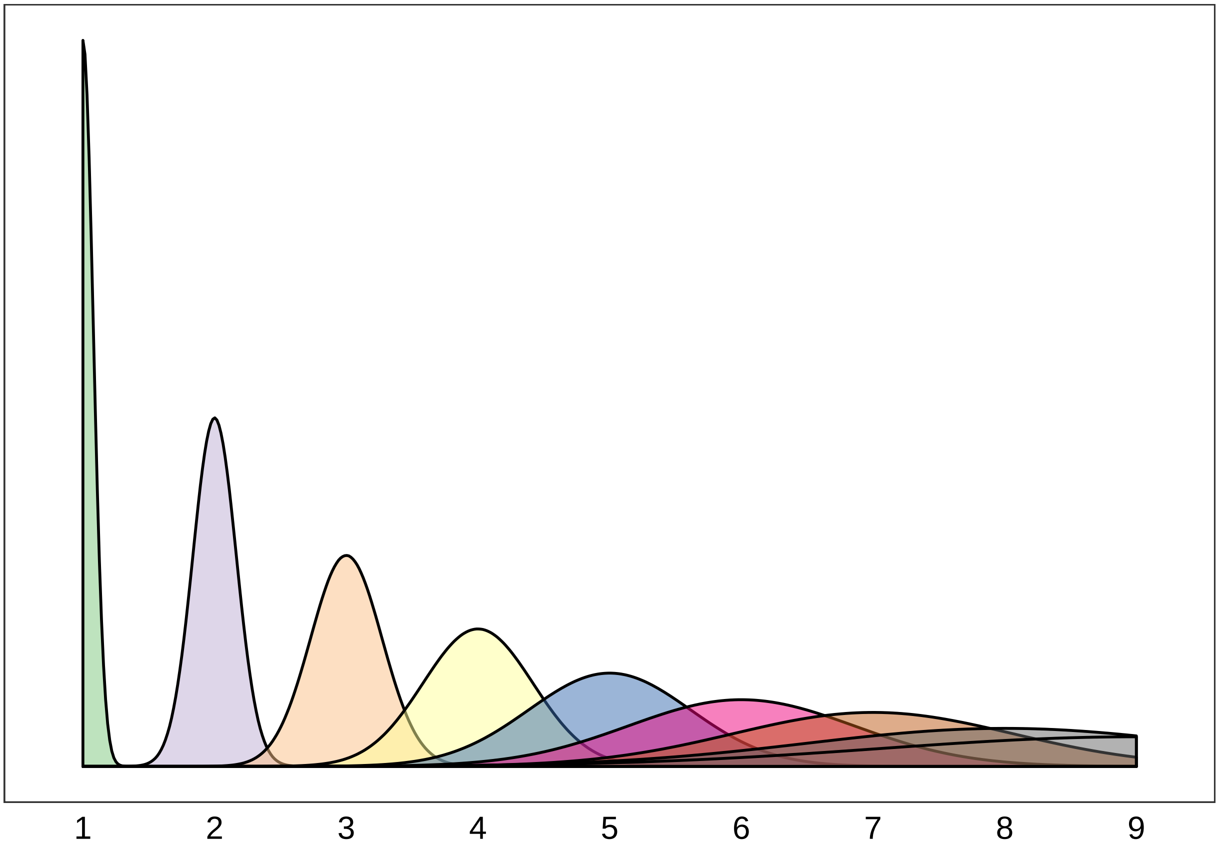


```
ggsave(country_revs, filename = "images/country_reviews.pdf", width = 5, height = 7)
```

This was my favorite finding. There really is a large difference between the nationalities of the reviewers. I guess people just are raised with different expecations. I also thought that it was interesting that the most harsh critics are generally from the Middle East. I first had this chart as a normal bar chart but it was hard to see the difference in the raw values (on a scale of 1-10, then roughly ranged from 7.5-8.9). That's when I thought about displaying this as z-scores and having the bars diverge from a z-score of 0. Z-scores might be beyond the understanding of the lay reader but I think the graphic conveys the idea without a pure understanding of the math.

Figure 3

```
nights <- hotels %>%
  select(Tags, Reviewer_Score) %>%
  unnest_tokens(word, Tags) %>%
  filter(word %in% c("1", "2", "3", "4", "5", "6", "7", "8", "9")) %>%
  group_by(word)

night_density <- ggplot(nights, aes(x = word, fill = word)) +
  geom_density(alpha = .5) +
  theme_bw()+
  scale_fill_brewer(palette = "Accent") +
  theme(strip.text.x = element_blank(),
        strip.text.y = element_blank(),
        axis.text.y=element_blank(),
        axis.title = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "none",
        axis.ticks = element_blank(),
        panel.background = element_rect(fill = "transparent", colour = NA),
        plot.background = element_rect(fill = "transparent", colour = NA),
        axis.text.x=element_text(size = 12, colour = "black"))
night_density
```

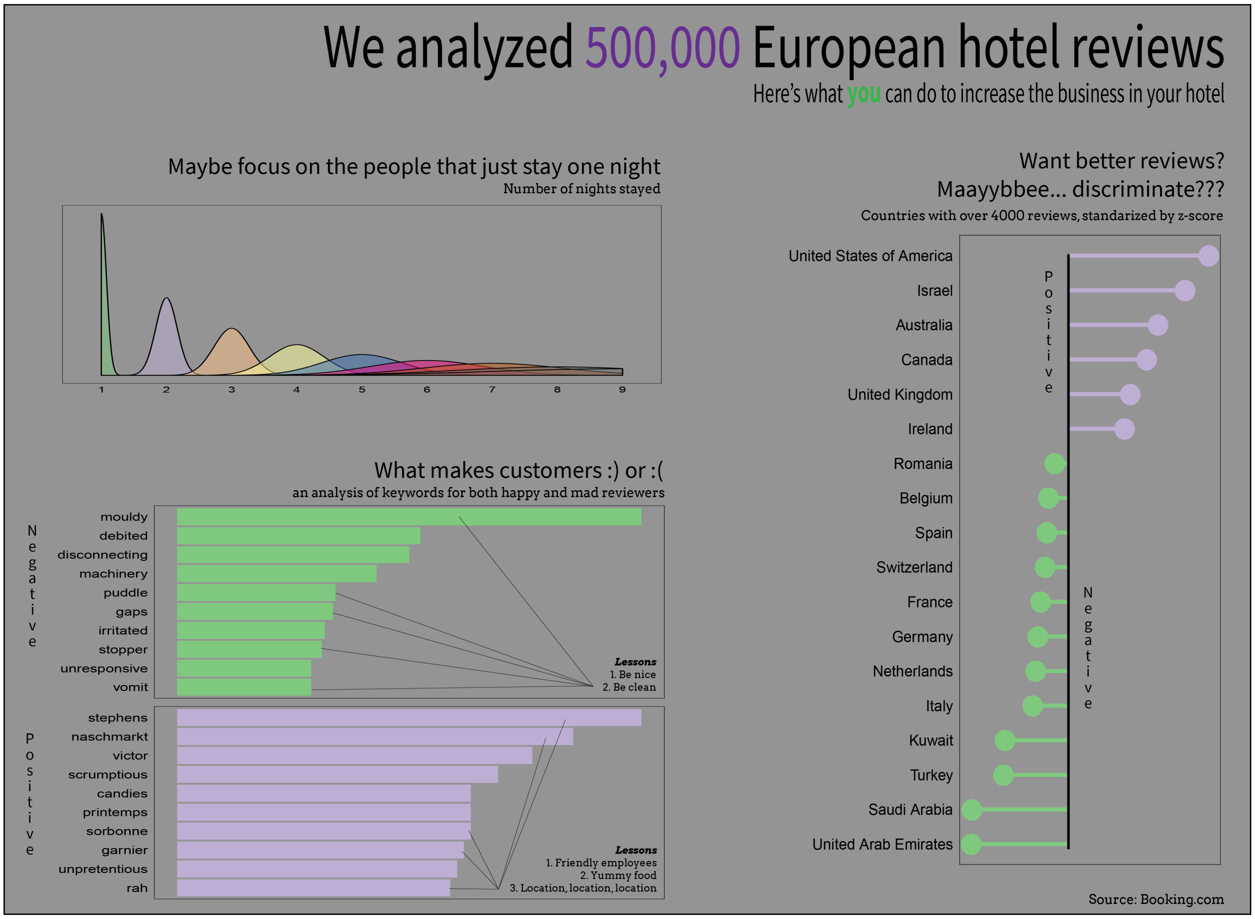


```
ggsave(night_density, filename = "images/nights.pdf", width = 8, height = 4)
```

Admitedly, this graph is pretty stupid. It was probably the 5th graph I built looking at different aspects of the data. I was really having a hard time coming up with anything insightful. I tried looking at the relationship between the number of reviews a person has given and the scores they give but the results were pretty insignificant. I played with other parts of the actual reviewer text and that's when I realized that a different column had the nights each person stayed. Again, this isn't really insightful - obviously more people just stay for one night. The business implication is unclear because you risk requent vacancy if all your customers are only staying for one night. Alas, I ran out of time and couldn't find anything more insightful.

Final figure

First first thought when seeing this is a scene from the first Home Alone movie. "Buzz's girlfriend. WOOF!" I know that this is super ugly but I really ran out of time. I don't like the grey background but I was having a hard time coming up with a consistent them for each of the three R charts that rendered well on a white background. Grey allowed for the charts to stand out. I tried to give everything a right alignment. There is some truth is this graph - people from different nationalities have diverse expectations when it comes to hospitality and there is consistency among positive and negative reviews. I have two different fonts - one for the titles and one for the subtitles.



We analyzed 500,000 European hotel reviews

Here's what you can do to increase the business in your hotel

Maybe focus on the people that just stay one night

Number of nights stayed

What makes customers :) or :(

an analysis of keywords for both happy and mad reviewers

Lessons
1. Be nice
2. Be clean

Lessons
1. Friendly employees
2. Yummy food
3. Location, location, location

Want better reviews?
Maaayybbee... discriminate???

Countries with over 4000 reviews, standarized by z-score

Source: Booking.com

Figure 1: Whoop, there it is!