Text_Analysis

Alex Wei

12/3/2021

```
listing <- read.csv('listings.csv')
calendar <- read.csv('calendar.csv')
# head(listing)
# head(calendar)
```

Price info in the calendar file, match the price and description, stratify the price and create word cloud for different levels.

```
calendar <- filter(calendar, price != '')

calendar <- calendar %>% mutate (price = as.numeric(substring(price,2)))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
# head(calendar)
```

```
price_tab <- calendar %>%
    group_by(listing_id) %>%
    summarise(mean_price = round(mean(price),2))

description_tab <- listing %>% select(listing_id = id,description)

# head(price_tab)
# head(description_tab)

dat <- left_join(price_tab,description_tab, by = 'listing_id')
```
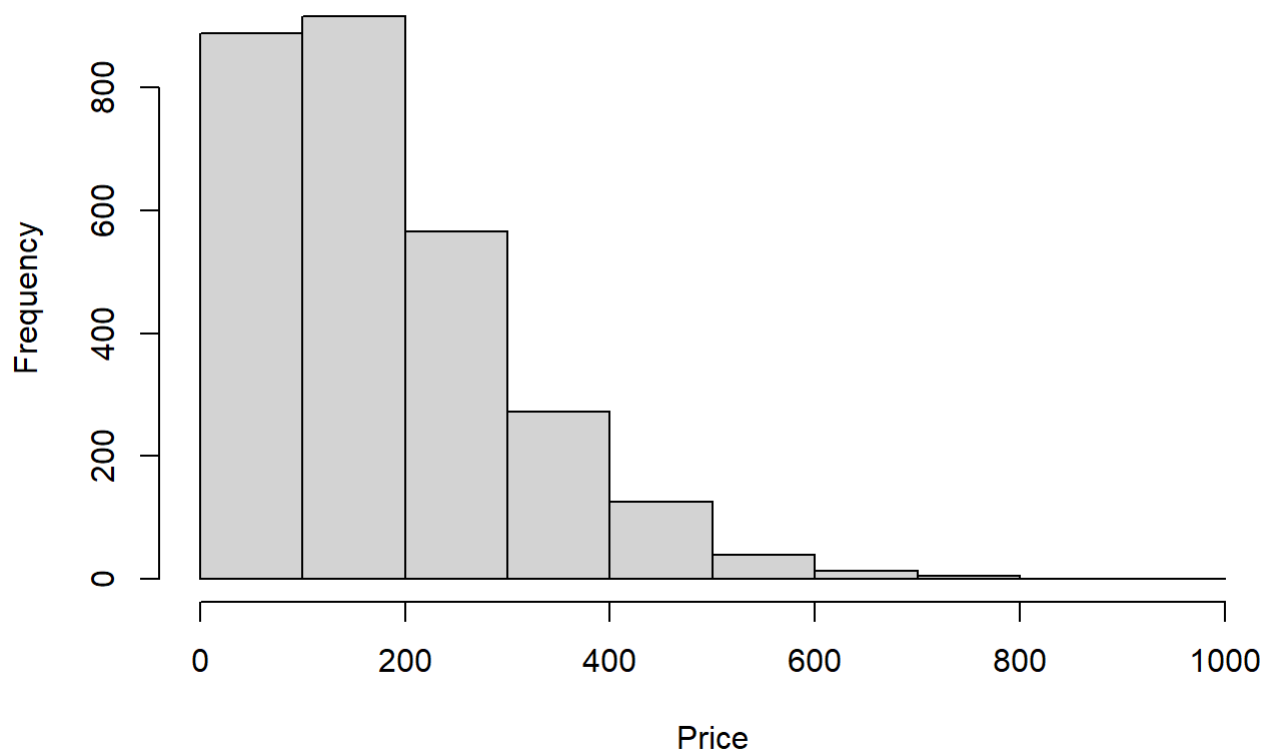
```
hist(dat$mean_price, main="Mean Price Distribution", xlab = "Price")
```

Mean Price Distribution


```
low_p <- dat %>% filter(mean_price < 100.0)

med_p <- dat %>% filter(mean_price >= 100.0 & mean_price < 400.0)

high_p <- dat %>% filter(mean_price >= 400.0)
```

```
# install.packages("tm")
library(tm)
```

```
## Loading required package: NLP
```

```
docs <- Corpus(VectorSource(low_p$description))

docs <- docs %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(., removeNumbers): transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(., removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(., stripWhitespace): transformation drops
## documents
```
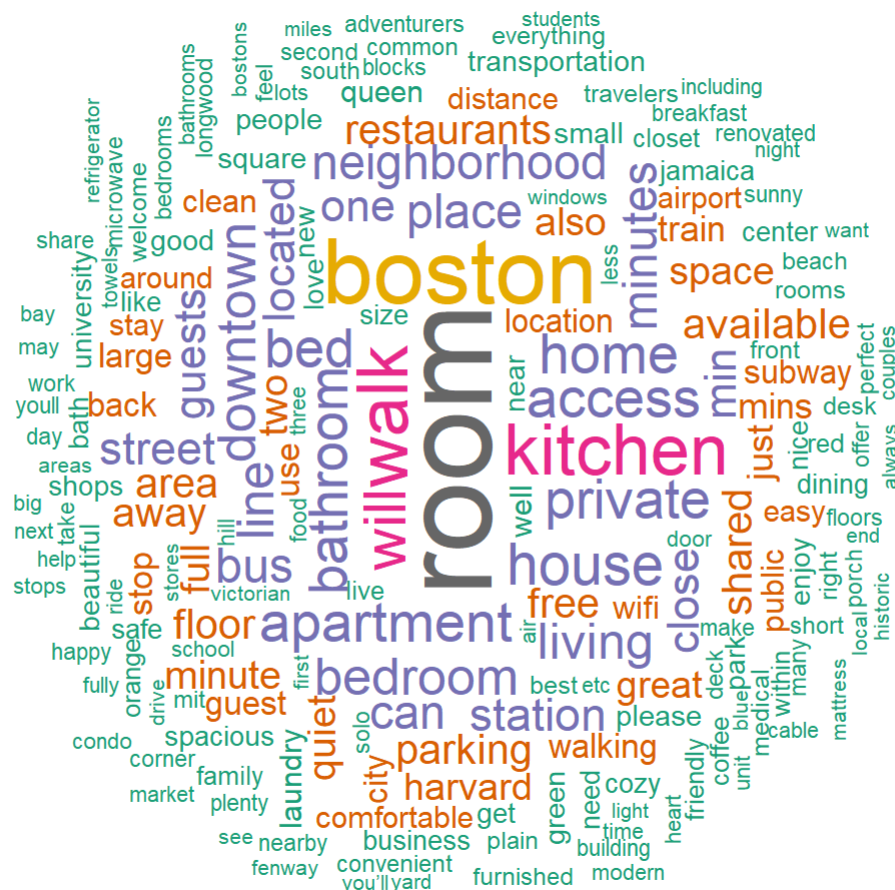
```
docs <- tm_map(docs, content_transformer(tolower))
```

```
## Warning in tm_map.SimpleCorpus(docs, content_transformer(tolower)):
## transformation drops documents
```

```
docs <- tm_map(docs, removeWords, stopwords("english"))
```

```
## Warning in tm_map.SimpleCorpus(docs, removeWords, stopwords("english")):
## transformation drops documents
```

```
dtm <- TermDocumentMatrix(docs)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
df <- data.frame(word = names(words),freq=words)

wordcloud(words = df$word,
          freq = df$freq,
          min.freq = 1,
          max.words=200,
          random.order=FALSE,
          rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

```
docs <- Corpus(VectorSource(med_p$description))

docs <- docs %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(., removeNumbers): transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(., removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(., stripWhitespace): transformation drops
## documents
```

```
docs <- tm_map(docs, content_transformer(tolower))
```

```
## Warning in tm_map.SimpleCorpus(docs, content_transformer(tolower)):
## transformation drops documents
```

```
docs <- tm_map(docs, removeWords, stopwords("english"))
```

```
## Warning in tm_map.SimpleCorpus(docs, removeWords, stopwords("english")):
## transformation drops documents
```

```
dtm <- TermDocumentMatrix(docs)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
df <- data.frame(word = names(words),freq=words)

wordcloud(words = df$word,
          freq = df$freq,
          min.freq = 1,
          max.words=200,
          random.order=FALSE,
          rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : central could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : garden could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : boston's could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : orange could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : convenient could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : microwave could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : closet could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : newly could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : entire could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : attractions could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : harvard could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : stainless could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : brownstone could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : you'll could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : includes could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : washerdryer could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : lounge could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : beds could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : ceilings could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : wireless could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : mattress could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : travelers could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : around could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : jamaica could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : prudential could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : block could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : washer could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : hall could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : make could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : across could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : second could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : nearby could not be fit on page. It will not be plotted.
```

```
docs <- Corpus(VectorSource(high_p$description))

docs <- docs %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(., removeNumbers): transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(., removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(., stripWhitespace): transformation drops
## documents
```

```
docs <- tm_map(docs, content_transformer(tolower))
```

```
## Warning in tm_map.SimpleCorpus(docs, content_transformer(tolower)):
## transformation drops documents
```

```
docs <- tm_map(docs, removeWords, stopwords("english"))
```

```
## Warning in tm_map.SimpleCorpus(docs, removeWords, stopwords("english")):
## transformation drops documents
```

```
dtm <- TermDocumentMatrix(docs)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
df <- data.frame(word = names(words),freq=words)

wordcloud(words = df$word,
          freq = df$freq,
          min.freq = 1,
          max.words=200,
          random.order=FALSE,
          rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : comfortable could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : financial could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : seaport could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : renovated could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : appliances could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : available could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : towels could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : trail could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : rooftop could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : service could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : proximity could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : bathrooms could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : gorgeous could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : three could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : attractions could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : freedom could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : unparalleled could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : hardwood could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : king could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : include could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : package could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : stainless could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : many could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : walking could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : event could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : basketball could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : courts could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : state could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : beyond could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : penthouse could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : gardens could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : welcome could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : ceilings could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : inside could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : waterfront could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : amenity could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : community could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : constructed could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : main could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : blocks could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : esplanade could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : condo could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : situated could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : perfectly could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : entire could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : experience could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(words = df$word, freq = df$freq, min.freq = 1, max.words =
## 200, : residents' could not be fit on page. It will not be plotted.
```