Lab 09

Stat 131A

Welcome to lab 9! In this lab, you will:

- Explore some visualization methods for data sets with categorical variables.
- Learn about heat maps.

Multivariate Data visualization - Online news popularity

The data stored in `OnlineNewsPopularity.csv` includes several variables describing articles published by Mashable over a period of two years. This dataset contains 49 variables for each news post, including

- `weekday`: Days of week. Mon, Tue, Wed, etc.
- `channel`: Channel. Tech, Entertainment, Business, etc.
- `shares`: Number of shares.
- `num_imgs`: Number of images.
- `num_videos`: Number of videos.
- `n_tokens_title`: : Number of words in the title.
- `num_hrefs`: Number of links

Read in data.

```
popul <- read.csv("OnlineNewsPopularity.csv")
popul$weekday <- factor(popul$weekday, c("Sun", "Mon", "Tue", "Wed", "Thur", "Fri", "Sat"))
head(popul[,c("weekday", "channel", "shares", "num_imgs")])
```

```
##   weekday       channel shares num_imgs
## 1     Mon Entertainment    593        1
## 2     Mon      Business    711        1
## 3     Mon      Business   1500        1
## 4     Mon Entertainment   1200        1
## 5     Mon          Tech    505       20
## 6     Mon          Tech    855        0
```

Exercise 1

(a) Construct side-by-side bar plots for `weekdays` and `channels` the using `barplot()` and `table()` functions. Rotate the horizontal axis labels using argument `las = 2` in `barplot()`. *Hint*: use `par(mfrow...)`.

```
# Insert your code for plotting here
```

(b) Create a contingency table for `weekdays` and `channels` using `table()`. Use `barplot()` to visualize the relationship between two categorical variables.

```
# Insert your code for plotting here
```

(c) Use the contingency table you created in (b) to get separate bar plots for days of the week. *Hint*: use `beside = TRUE`.

```
# Insert your code for plotting here
```

Exercise 2

The following script takes a subset of 2000 rows from the news popularity dataset.

```
vars <- c("weekday", "channel", "shares", "num_imgs", "num_videos", "n_tokens_title", "num_hrefs")
sample.idx <- sample(nrow(popul), 2000)
subset <- popul[sample.idx, vars]
```

(a) Generate a matrix of scatter plots for all variable pairs in `popul.subset` that are of class `numeric` using the `pairs()` function.

```
# Insert your code for plotting here
```

(b) Generate a matrix of scatter plots for all variable pairs in `popul.subset` using the `gpairs()` function in the `gpairs` package.

```
library(gpairs)
# Insert your code for plotting here
```

For fun (not required)

(a) Plot the alluvial plot for `weekday` and `channel` using `alluvial()` in the `alluvial` package.

```
library(alluvial)
# Insert your code for plotting here
```

(b) Plot the mosaic plot for `weekday` and `channel` using `mosaicplot()`.

```
# Insert your code for plotting here
```

Places rated dataset

The data were taken from the *Places Rated Almanac* which rates cities according to nine criteria. For all but two of the criteria, the higher the score, the better. For Housing and Crime, the lower the score the better. The following are descriptions of the criteria:

- Climate & Terrain: very hot and very cold months, seasonal temperature variation, heating- and cooling-degree days, freezing days, zero-degree days, ninety-degree days.
- Housing: utility bills, property taxes, mortgage payments.
- Health Care & Environment: per capita physicians, teaching hospitals, medical schools, cardiac rehabilitation centers, comprehensive cancer treatment centers, hospices, insurance/hospitalization costs index, flouridation of drinking water, air pollution.
- Crime: violent crime rate, property crime rate.
- Transportation: daily commute, public transportation, Interstate highways, air service, passenger rail service.
- Education: pupil/teacher ratio in the public K-12 system, effort index in K-12, accademic options in higher education.
- The Arts: museums, fine arts and public radio stations, public television stations, universities offering a degree or degrees in the arts, symphony orchestras, theatres, opera companies, dance companies, public libraries.
- Recreation: good restaurants, public golf courses, certified lanes for tenpin bowling, movie theatres, zoos, aquariums, family theme parks, sanctioned automobile race tracks, pari-mutuel betting attractions, major- and minor- league professional sports teams, NCAA Division I football and basketball teams, miles of ocean or Great Lakes coastline, inland water, national forests, national parks, or national wildlife refuges, Consolidated Metropolitan Statistical Area access.
- Economics: average household income adjusted for taxes and living costs, income growth, job growth.

Read Data.

```
place_rated <- read.csv("place.csv", stringsAsFactors = FALSE)
place_rated[, 1:9] <- scale(place_rated[, 1:9])
CA_NY_cities <- which(place_rated$state %in% c("CA", "NY"))
example_cities <- CA_NY_cities[c(10, 13, 15, 22, 24, 25, 26)]
row.names(place_rated) <- paste0(place_rated$city, place_rated$state)
```

Heatmap

Exercise 3

Heatmaps map numeric data to colors and are usually used to visualize correlation matrices (and other matrices)

(a) Calculate the correlation matrix between nine rating criteria using function `cor`. What is the correlation between arts and education?

```
# insert your code here and save the
# correlation matrix as `place_rated.cor`

# place_rated.cor <-
```

(b) Plot the heatmap for the correlation matrix using the `pheatmap()` functiomn from `pheatmap` package. If you are going to divide the nine rating criteria into two categories based on similarity, how would you split the variables?

```
# insert your code here for heatmap
library(pheatmap)
```

(c) Plot the heatmap for data matrix.

```
# insert your code here for the heatmap
```