Homework 3

1. **Parallel Coordinates**

(a) Draw a parallel coordinates plot of the data in "ManhattanCDResults.csv" in the data folder on CourseWorks. (Original data source and additional information about the data can be found here: https://cbcny.org/research/nyc-resident-feedback-survey-community-district-results). Your plot should have one line for each of the twelve Manhattan community districts in the dataset.

(b) Do there appear to be greater differences across *indicators* or across *community districts*? (In other words, are Manhattan community districts more alike or more different in how their citizens express their satisfaction with city life?

There appear to be greater differences across indicators. For most situations, communities have a relatively same trend. But for some indicators like V1, V17 and V20, they are so different from each other.

(c) Which indicators have wide distributions (great variety) in responses?

By observation, V1, V3, V17, V18, V19, V20, V21 and V39 have relatively wide distributions.

To get a more precise result, we can also calculate all variance of the indicators.

(d) Does there appear to be a correlation between districts and overall satisfaction? In order words, do some districts report high satisfaction on many indicators and some report low satisfaction on many indicators or are the results more mixed? (Hint: a different color for each community district helps identify these trends).

Yes. As shown in the parallel coordinates plot, line cd7 has high satisfaction nearly on all indicators, while lines cd3, cd11 and cd12 have most low satisfaction on most indicators.

2. **Mosaic Plots**

Using the "Death2015.txt" data from the previous assignment, create a mosaic plot to identify whether `Age` is associated with `Place of Death`. Include only the top four `Age` categories. Treat `Age` as the independent variable and `Place of Death` as the dependent variable. (Hint: the dependent variable should be the last cut and it should be horizontal.) The labeling should be clear enough to identify what's what, that is, "good enough," not perfect. Do the variables appear to be associated? Describe briefly.

According to the dataset, the top 4 groups which have more samples are 65-74 years, 75-84 years, 85+ years, and 55-64 years.

3. **Time Series**

(a) Use the `tidyquant` package to collect stock information on four stocks of your choosing. Create a line chart showing the closing price of the four stocks on the same graph, employing a different color for each stock.

(b) Transform the data so each stock begins at 100 and replot. Do you learn anything new that wasn't visible in part (a)?

4. **Missing Data**

For this question, explore the New York State Feb 2017 snow accumulation dataset available in the data folder on CourseWorks: "NY-snowfall-201702.csv". The original data source is here: https://www.ncdc.noaa.gov/snow-and-ice/daily-snow/

(a) Show missing patterns graphically.

(b) Is the percent of missing values consistent across days of the month, or is there variety?

(c) Is the percent of missing values consistent across collection stations, or is there variety?

According to the plot, we can see that, percent of missing values are not consistent across collection stations. Some stations have almost 1 missing values while some stations only have nearly zero percent of missing values.

(*d*) Is the daily average snowfall correlated with the daily missing values percent? On the basis of these results, what is your assessment of the reliability of the data to capture true snowfall patterns? In other words, based on what you've discovered, do you think that the missing data is highly problematic, or not?

From the plot, we can see that higher percent of missing data is more likely to appear on low daily average snowfall.

The missing data is not problematice since existing data already could capture the true snowfall patterns.