Lab Assignment #1

ENSF 619.25 Machine Learning— Winter 2019

Instructor: Zahra Shakeri, University of Calgary

Objectives

Objectives for our first lab assignment:

- Understanding different data types
- Applying different analysis techniques
- Understanding exploratory data analysis
- Applying various visualization techniques to better explore and understand a dataset*

*Some sample visualization techniques: bar charts, Histogram, box plot, scatter plot, radar chart

About the Dataset

Main Features

This data set is based on the nutrition analysis of menu items on a coffee shop including coffee drink and foods. This dataset contains information about calories, fat, carb, fiber, protein, and sodium and etc.

A Sample Visualization

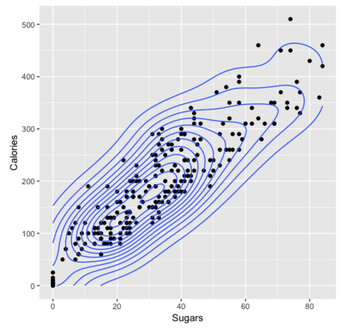

Figure 1: 2D density plot

Instruction

Please follow the below instruction:

- Download the dataset from D2L (/content/Lab Assignments/Assignment 1)
- Create a new project in RStudio
- Install required packages
- Call Libraries and load the data
- Start by plotting the distribution of sugar in each drink...

Exploratory Analysis

Sample techniques:

1. Correlation analysis
2. Frequency distribution
3. Means
4. Centrality
5. Variance and standard deviation

A Successful Submission ...

A successful submission will include a thoughtful data analysis on the provided dataset and will detail specific and understandable descriptions of each finding.
We expect to see at least three interesting findings of your exploratory data analysis. Each finding must include a *visualization*, *a description of the finding's corresponding analysis technique*, and *your justification* for choosing each of the analysis methods.

Example Code Snippet (R)

```r
#Correlation
library(corrplot)
# Load data
data("mtcars")
data <- mtcars[, c(1,3,4,5,6,7)]
# print the first 6 rows
head(data, 6)
correlation <- cor(data)
round(correlation, 2)
corrplot(correlation,
method="number")
#try the chart with "pie", "color",
and "number"
```

Example Code Snippet (Python)

```python
import pandas as pd
import matplotlib.pyplot as plt
#read .csv file of nutrition data
data = pd.read_csv("../input/TimiTomi_expanded.csv")
#isolate the intended column of data
fat = nutrition_data["fat"]
#plot a histogram of calories data
plt.hist(calories, bins=9, edgecolor
= "black")
plt.title("fat in TimiTomi menue")
# add a title
plt.xlabel("fat") # label the x axes
plt.ylabel("Count") # label the y
axes
```

How to Submit?

Please submit your assignments via D2L as a single *.zip* file which contains all the needed deliverables by 1:50 PM on January 31^{th}. Make sure to put the names of all group members on the first page of your assignment. Only submit **one** copy per group! All teams will also demo their assignments during the Lab on the same day.

Note: This assignment is pretty free-form! This is intentional; projects you work on in industry will rarely be very specific. Please feel free to show early results to Zahra as well as your TAs to get a feedback you can use to ensure a successful submission/Demo!

Possible Scenarios

- Exploring the shape of different variables' distribution, their central value, and their variability.
- Exploring which variables have similar values or if there are any outliers amongst each variable.
- Exploring how much one variable is affected by other variables.

R Cheat Sheet

Purpose	Operation
`?function`	Get help of a particular function.
`install.packages()`	Download and install a package from CRAN.
`library()`	Load the package into the session
`setwd('C://path')`	Change the current working directory.
`df <- read.csv('file.csv')`	Read a CSV file
`rm(list=ls())`	Clear the environment

Table 1: R Cheat Sheet #1

Some Necessary and Useful APIs

- `ggplot2`
- `RColorBrewer`
- `corrplot`
- `dplyr`
- `Hmisc`
- `devtools`