Intro to Data Science - Lab 5

DATA 1501 — Dr. Mihail
Department of Computer Science
Valdosta State University

September 29, 2021

Introduction

Part 1 (20 points)

In this part, you will download a dataset containing information on 80 cereals, and make some basic visualizations. During this lab, have a Word of Google Docs document open and be prepared to take screenshots and answer questions as per the lab directions. Create a new Colab notebook and run the following code in a code cell:

```
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab6/cereal.csv
```

The data legend for this dataset is the following:

```
Fields in the dataset:

Name: Name of cereal
mfr: Manufacturer of cereal
    A = American Home Food Products;
    G = General Mills
    K = Kelloggs
    N = Nabisco
    P = Post
    Q = Quaker Oats
    R = Ralston Purina

type: cold or hot
calories: calories per serving
protein: grams of protein
fat: grams of fat
sodium: milligrams of sodium
fiber: grams of dietary fiber
carbo: grams of complex carbohydrates
```

```
sugars: grams of sugars
potass: milligrams of potassium
vitamins: vitamins and minerals - 0, 25, or 100, indicating the
    typical percentage of FDA recommended
shelf: display shelf (1, 2, or 3, counting from the floor)
weight: weight in ounces of one serving
cups: number of cups in one serving
rating: a rating of the cereals (Possibly from Consumer Reports?)
```

To create a Pandas dataframe, create a cell with the following code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv('cereal.csv')
df
```

Confirm you can see a few records in the dataset.

Part 2 (20 points)

Create a new cell and run the code below.

```
# create a histogram of brand frequencies in the dataset
frequency = pd.DataFrame(df['mfr'].value_counts())
frequency.columns = ['Frequency']
frequency.plot(kind='bar')
plt.ylabel("Number of Records")
```

Create a screenshot, add it to your word document, and answer the following questions:

- Describe, in your own words, what the plot represents.
- What is the most common cereal in the dataset?
- What is the least common cereal in the dataset?

Part 3 (20 points)

Create a new cell and run the code below:

```
kellogs_protein = df[df["mfr"]=="K"]["protein"].mean()
generalmills_protein = df[df["mfr"]=="G"]["protein"].mean()

l = [[kellogs_protein, generalmills_protein]]

protein_content = pd.DataFrame(l, columns = ["Kellogs", "General Mills"])
```

```
ax = protein_content.plot(kind='bar')
plt.setp( ax.get_xticklabels(), visible=False)
plt.ylabel("Average Protein content (grams)")
```

Create a screenshot of the resulting graph and answer the following questions:

- Describe, in your own words, what the plot represents.
- Which cereal brand has the most (on average) protein content?

Part 4 (20 points)

Create a new cell, and run the following code:

```
df.groupby(["mfr"]).mean().plot(kind='bar', y=["protein", "fat", "sugars","fiber"])
```

Add the plot to your Word document. Answer the following questoins about the plot you just created:

- Describe, in your own words, what the plot represents and how one might use it to differentiate between different brands of cereal.
- Which brand of cereal offers, on average, the most balance? Please explain your choice.

Part 5 (20 points)

Create a new cell, and run the following code:

```
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab6/starbucks.csv
```

Create a new cell, and run the following code:

```
import pandas as pd
import numpy as np
df = pd.read_csv('starbucks.csv')
df

df[" Calories"].hist(bins=25)
plt.xlabel("Calories")
plt.ylabel("Number of menu items")
plt.title("Distribution of menu item calories")
```

Add the plot to your Word document. Answer the following questoins about the plot you just created:

- Describe, in your own words, what the plot represents and how one might use it to tell a story about Starbucks food menu items.
- Based on the plot you just generated, what seems to be the most common caloric value of Starbucks food menu items.

Create a new cell, and run the following code:

```
sorted_values = df.sort_values(by=" Calories")
names = np.array(sorted_values["Item Name"])
cals = np.array(sorted_values[" Calories"])

fig, ax = plt.subplots(1, 2, figsize=(10, 10))
ax[0].bar(names[0:5], cals[0:5])
ax[0].set_ylabel("Calories")
ax[0].set_title("5 lowest caloric items")
ax[0].set_xticklabels(names[0:5], rotation=90)

ax[1].bar(names[-5:], cals[-5:])
ax[1].set_ylabel("Calories")
ax[1].set_title("5 highest caloric items")
ax[1].set_xticklabels(names[-5:], rotation=90)
```

Answer the following questions:

- What is the lowest calorie menu item at Starbucks?
- What is the highest calorie menu item at Starbucks?
- What is the difference (in calories) between the highest and lowest caloric content food menu items at Starbucks?

Create a new code cell and run the following code:

```
sorted_values = df.sort_values(by=" Calories")
names = np.array(sorted_values["Item Name"])
cals = np.array(sorted_values[" Calories"])

fig, ax = plt.subplots(1, 2, figsize=(10, 10))
ax[0].bar(names[0:5], cals[0:5])
ax[0].set_ylabel("Calories")
ax[0].set_title("5 lowest caloric items")
ax[0].set_xticklabels(names[0:5], rotation=90)
ax[0].set_ylim([cals[0] - 5, cals[4] + 5])

ax[1].bar(names[-5:], cals[-5:])
ax[1].set_ylabel("Calories")
ax[1].set_title("5 highest caloric items")
ax[1].set_xticklabels(names[-5:], rotation=90)
ax[1].set_ylim([cals[len(cals)-5]-5, cals[len(cals) - 1] + 5])
```

Add a screenshot of the resulting plot to your Word document, and answer the following questions:

- Describe, in your own words, the difference between the plots in the code cell before this current one.

- Given the current plot, what is the difference (in calories) between the highest and lowest caloric content food menu items at Starbucks?
- Does your estimate now agree with the previous estimate for the plot before this? Why/why not?.

Due Date: Before Midnight on Sunday, October 3rd.