David Luo

Homework 1: Cereal Data Analysys

(1) Insights

- Manufacturer N seems to have significantly lower average calories than all other companies. (additional note, Manufacturer R seems to have 2 distinct modes for its calories distribution). See figure 3
- It would seem that the calories distribution is relatively normal with a few outliers (notably near 0 calories). See figure 2
- Most Manufacturers make less than 10 cereals except G and K who make 22 and 23 types respectively. (note, Manufacturer A only makes 1 cereal). See tables
- Manufacturer N has many 0 sugars cereals. See figure 1
- Hot cereal has notably fewer sugars than Cold cereal (but there are only 3 hot cereals in this data set) See tables
- Most cereals fall into an area of <5 fiber, notable outliers include 4 cereals from K and 2 cereals from N. See figure 1

(2) Process

Cleaning:

The data was loaded into the pandas library in Python3. describe() was used to find summary statistics about the data. unique() was used to see if any values seemed off. It was found that some values were recorded as -1. replace() was used to replace these with NaN.

Explore:

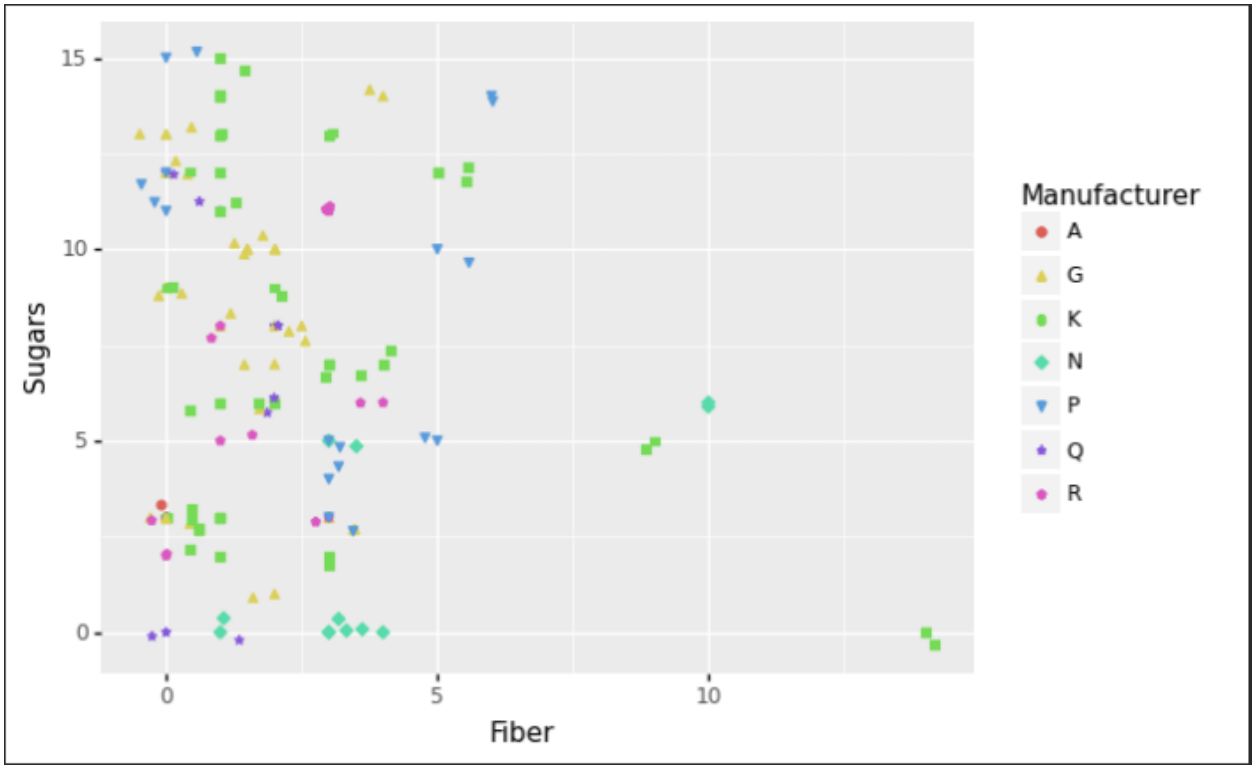
The groupby() function was used to see mean and counts by Manufacturer and Type.
From this, exploratory graphs were created using plotnine.

(3) Challenges/Problems

Plotnine does not allow for interactive visualizations. Tableau visualizations will be better for more interactive and in-depth analysis (I do not have access to Tableau on my primary device yet). It was very difficult to come up with insights to the data, as the data is very lopsided towards two manufacturers.

(4) Visualizations

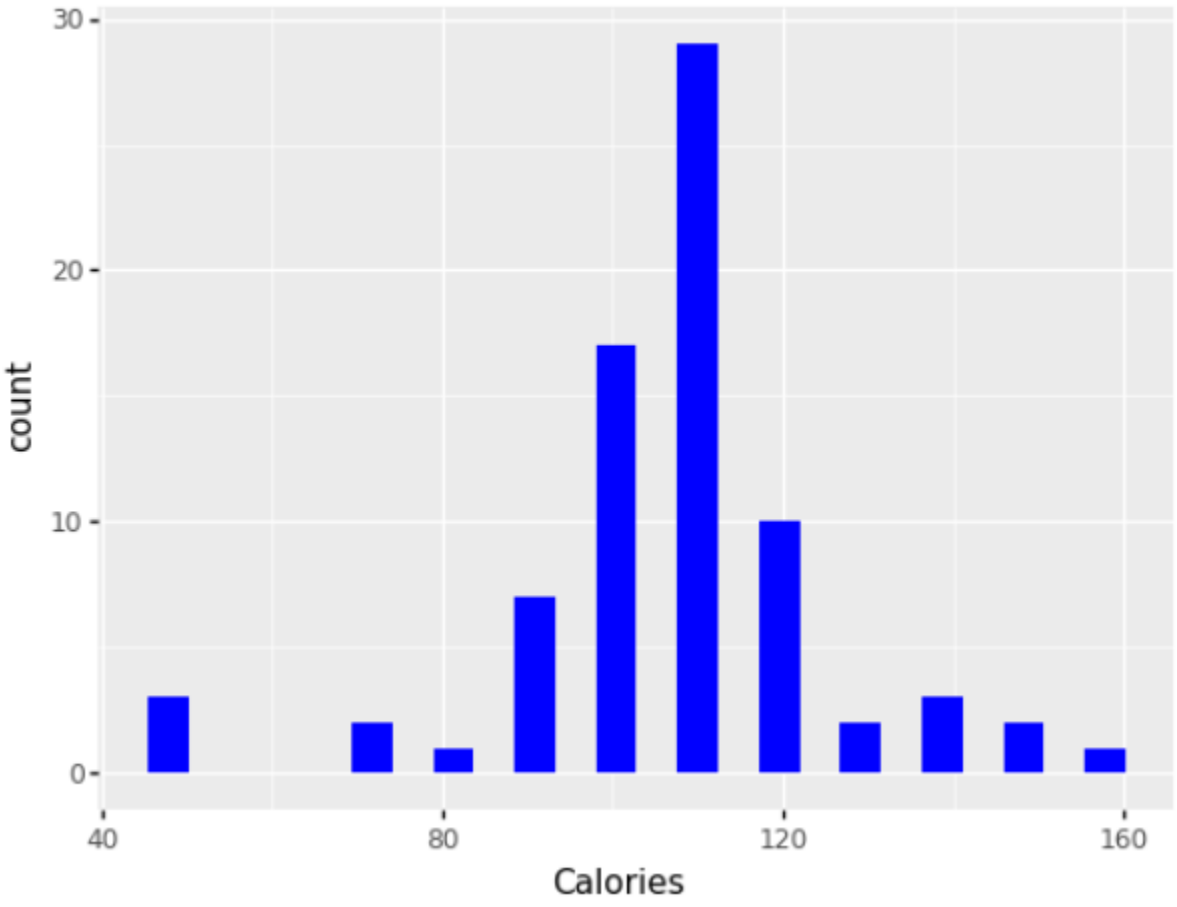
Figure 1



Fiber vs Sugars grouped by Manufacturer

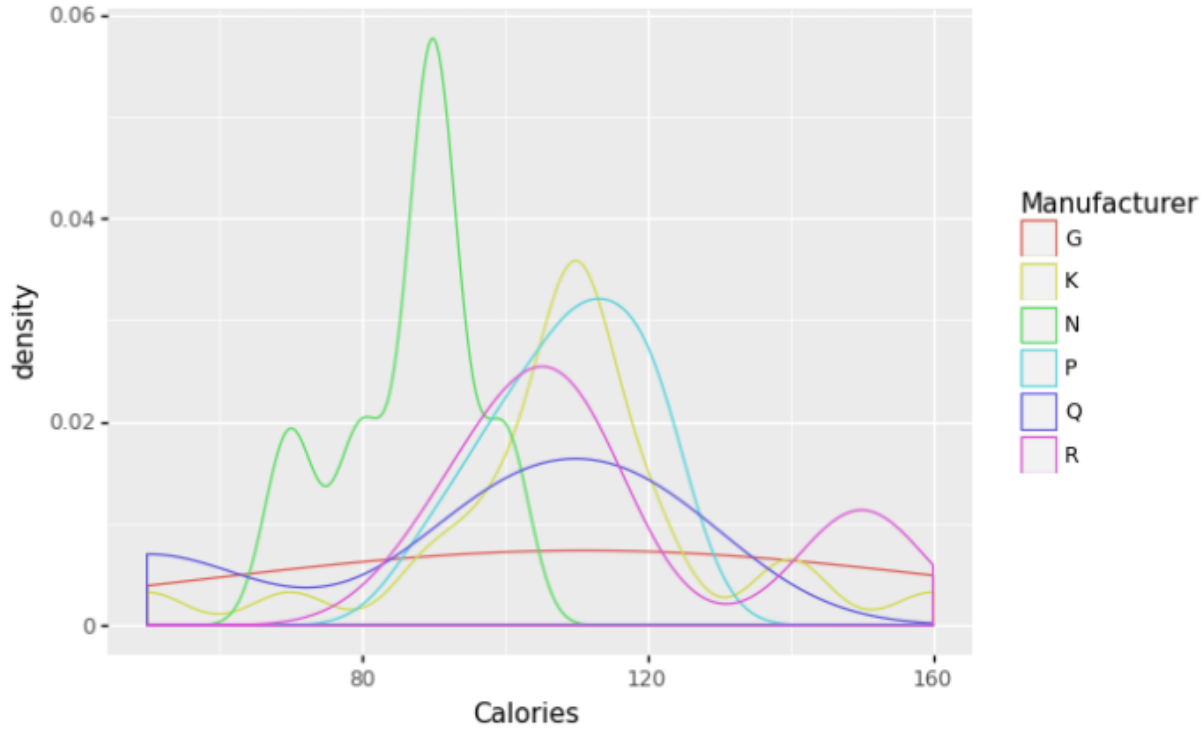
Note: The data was jittered to help prevent overlapping points

Figure 2



Histogram of Calories

Figure 3



Density of Calories by Manufacturer

Table 1 & 2

Manufacturer	Cereal
A	1
G	22
K	23
N	6
P	9
Q	8
R	8

Type	Cereal
C	74
H	3

Segmented sections from Counts table

Appendix

Code

```
#%%
from IPython.display import display
import numpy as np
import pandas as pd
import plotnine as p9
#%%
df = pd.read_csv("a1-cereals_csv.csv")
print(df)
df = df.replace(-1,np.NaN)
df.describe()
#%%
#Group By manufacture
print(df.groupby(["Manufacturer"]).mean())
print(df.groupby(["Type"]).mean())
print(df.groupby(["Manufacturer"]).count())
print(df.groupby(["Type"]).count())
#%%
p1 =
p9.ggplot(data=df,mapping=p9.aes(x="Fiber",y="Sugars",color="Manufacturer",shape="Manufa
cturer"))
p1+p9.geom_point()+p9.geom_jitter(width=.6)
#%%
p2 = p9.ggplot(data=df,mapping=p9.aes(x="Calories"))
p2+p9.geom_histogram(fill='blue')
#%%
p3 =
p9.ggplot(data=df,mapping=p9.aes(x="Calories",group="Manufacturer",color="Manufacturer"))
p3+p9.geom_density()

#%%
display(df.groupby(["Manufacturer"]).count())
display(df.groupby(["Type"]).count())
```