# Caffeine Project Write-Up

Alexandra Blitch
Fabi Estrada
Cagan Abney

Our group selected the “Caffeine Content of Drinks” dataset from Kaggle for this project. Our inspiration for this dataset was based on our familiarity with caffeinated drinks. We also believe it is important to know the caffeine and calorie amounts in today’s popular drinks. This dataset is 29.21 kB which consists of 5 columns and 611 rows. The data includes volume of drinks (mL), calorie content, caffeine content (mg), drink name, and type.

After selecting our dataset, we began thinking of research questions. Our groups decided on the following research questions:

- Does coffee/tea have a lower volume-to-caffeine ratio than other drink categories?
- Does soda have a higher calorie-to-caffeine ratio than other drink categories?
- Which drink type has the most caffeine per 100 mL?

After listing our research questions, we began writing our code. We started by importing modules into the notebook. The modules imported are as follows:

- import pandas as pd
- import matplotlib.pyplot as plt
- from pathlib import Path
- import numpy as np
- from scipy.stats import linregress
- import seaborn as sns
- import scipy.stats as st

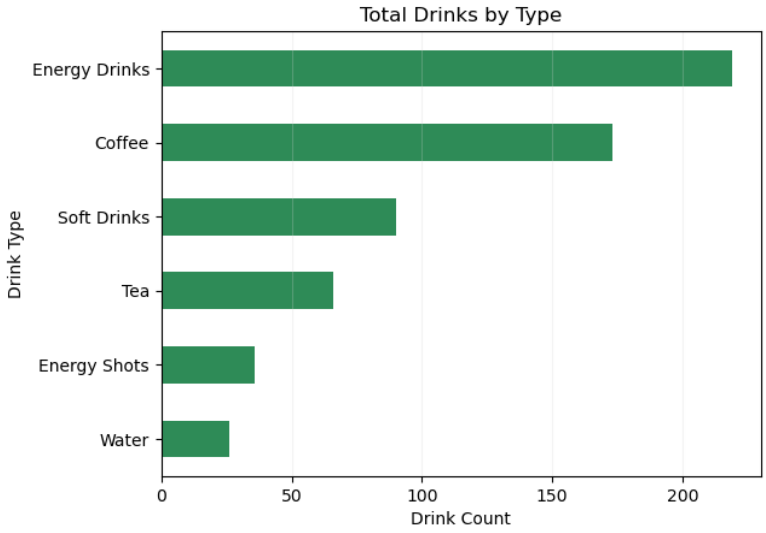
Below the modules, we added the colors for our visualizations and assigned them as variables as we wanted to use specific colors and easily call on them. In the next cell, we added and read the CSV from our dataset. We then ran a df.head() to confirm this worked. We then wanted to change the column names since some were lowercase and some were not. We changed “Volume (ml)” to “volume” and “Caffeine (mg)” to “caffeine”. Once the columns were renamed, we then ran a df.info() to check for null values. Our dataset came back with zero null values.

In the next cell, we ran a df.describe() to get a grasp on the averages, min, max, and quartiles of our data. The results are shown below:

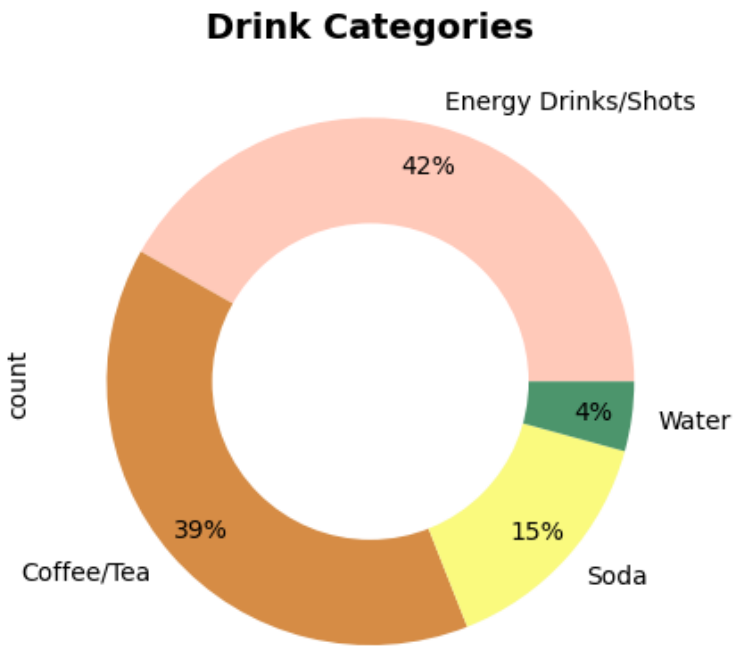
| | volume | calories | caffeine |
|---|---|---|---|
| **count** | 610.000000 | 610.00000 | 610.000000 |
| **mean** | 346.543630 | 75.527869 | 134.693443 |
| **std** | 143.747738 | 94.799919 | 155.362861 |
| **min** | 7.393375 | 0.000000 | 0.000000 |
| **25%** | 236.588000 | 0.000000 | 50.000000 |
| **50%** | 354.882000 | 25.000000 | 100.000000 |
| **75%** | 473.176000 | 140.00000 | 160.000000 |
| **max** | 1419.52800 | 830.00000 | 1555.00000 |

After viewing the table above, we noticed that the max volume, max calories, and max caffeine were oddly high. In the next three cells in our notebook, we sorted by values of each category to show us these outliers. This strategy was very insightful as it showed us that there may be some drinks that should not be in this dataset, for example, the 'Starbucks Bottled Iced Coffee', which has a volume of over 1419 mL. This is likely because it is a package or bundle of drinks rather than one single drink. We also saw that the drink with max calories was the 'Arby's Jamocha Shake', which had a relatively low level of caffeine and normal volume, so we decided to keep it in our dataset as it is representative of coffee-type drinks on the market.

The last step of our initial data engineering was to check for duplicate values. We found that there were no duplicate values in our dataset and we were ready to complete our first visualization. Below you can see a bar graph of the total drinks by type:

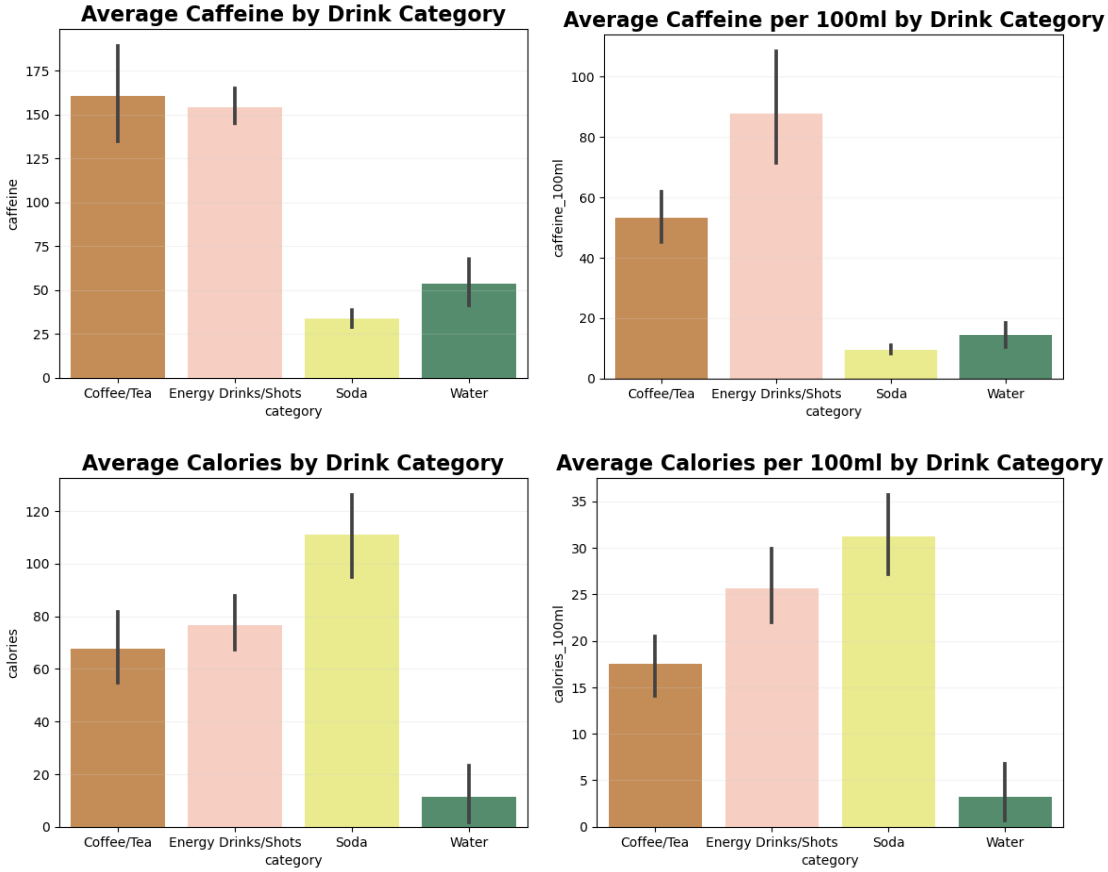



After creating this bar graph, we decided that the data would be better displayed by combining the three smaller drink types with the three larger drink types. In our notebook, we used the df.loc() to combine Coffee and Tea, Energy Drinks and Energy Shots, and Water and Soda. Once we combined these categories we realized that Water and Soda ran bimodally and would need to be their own parent categories. After settling for these four parent categories we wanted to show these categories by percentages. The new drink categories can be seen below:




We then created some bar graphs comparing the average amount of Caffeine and Calories to Caffeine and Calories per 100 mL. Coffee and Tea's average caffeine dropped significantly

when we compared it to the average Caffeine per 100mL. Soda had significantly higher Calories in both the average and the 100mL. The bar graphs are shown below:




This led us to more questions: how strong is the correlation between calories and volume compared to our original dataset? What about caffeine and volume? Lastly, calories and caffeine? Below is the original dataset:

| | volume | calories | caffeine |
|---|---|---|---|
| **volume** | 1.000000 | 0.341998 | 0.110770 |
| **calories** | 0.341998 | 1.000000 | -0.126021 |
| **caffeine** | 0.110770 | -0.126021 | 1.000000 |

Below, we can compare the original dataset to Coffee and Tea, Energy Drinks and Shots, Soda, and Water:

```
Coffee/Tea
```

| | volume | calories | caffeine | caffeine_100ml | calories_100ml |
|---|---|---|---|---|---|
| **volume** | 1.000000 | 0.409093 | 0.162447 | -0.197330 | 0.181752 |
| **calories** | 0.409093 | 1.000000 | -0.102982 | -0.213935 | 0.917602 |
| **caffeine** | 0.162447 | -0.102982 | 1.000000 | 0.862237 | -0.139445 |
| **caffeine_100ml** | -0.197330 | -0.213935 | 0.862237 | 1.000000 | -0.145618 |
| **calories_100ml** | 0.181752 | 0.917602 | -0.139445 | -0.145618 | 1.000000 |

According to the chart, calories and volume had a stronger correlation than the original dataset, as did caffeine and volume. However, calories and caffeine had a slightly weaker correlation.

```
Energy Drinks/Shots
```

| | volume | calories | caffeine | caffeine_100ml | calories_100ml |
|---|---|---|---|---|---|
| **volume** | 1.000000 | 0.326900 | 0.110534 | -0.646098 | -0.218562 |
| **calories** | 0.326900 | 1.000000 | -0.158162 | -0.276000 | 0.518352 |
| **caffeine** | 0.110534 | -0.158162 | 1.000000 | 0.287678 | -0.213583 |
| **caffeine_100ml** | -0.646098 | -0.276000 | 0.287678 | 1.000000 | 0.259719 |
| **calories_100ml** | -0.218562 | 0.518352 | -0.213583 | 0.259719 | 1.000000 |

Energy Drinks and Shots seemed to have a stronger correlation between calories and caffeine, caffeine and volume seemed to be pretty even, and calories and volume had a slightly weaker correlation.

```
Soda
```

| | volume | calories | caffeine | caffeine_100ml | calories_100ml |
|---|---|---|---|---|---|
| **volume** | 1.000000 | 0.277084 | 0.073379 | -0.033439 | 0.031704 |
| **calories** | 0.277084 | 1.000000 | -0.156792 | -0.182037 | 0.960212 |
| **caffeine** | 0.073379 | -0.156792 | 1.000000 | 0.984448 | -0.168080 |
| **caffeine_100ml** | -0.033439 | -0.182037 | 0.984448 | 1.000000 | -0.178579 |
| **calories_100ml** | 0.031704 | 0.960212 | -0.168080 | -0.178579 | 1.000000 |

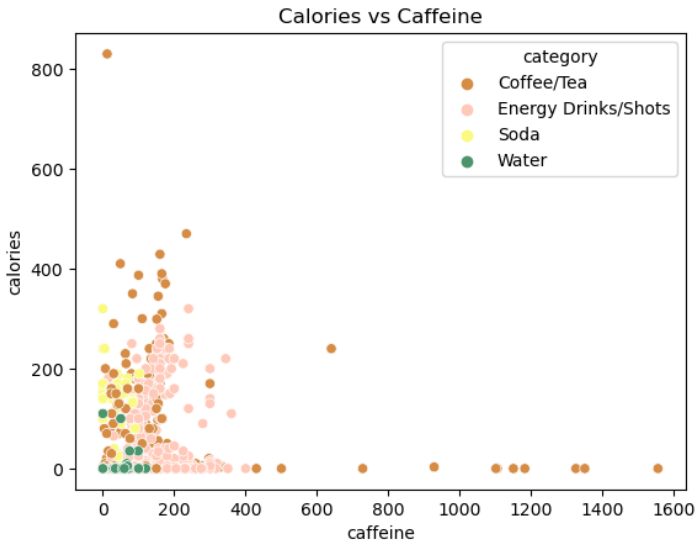
Soda's chart appeared to have a weaker correlation in both calories and caffeine when compared to volume, though calories and caffeine seemed to show a slightly stronger correlation when compared to each other.

Water

| | volume | calories | caffeine | caffeine_100ml | calories_100ml |
|---|---|---|---|---|---|
| **volume** | 1.000000 | 0.043064 | 0.038770 | -0.299305 | -0.130185 |
| **calories** | 0.043064 | 1.000000 | -0.147638 | -0.128281 | 0.946101 |
| **caffeine** | 0.038770 | -0.147638 | 1.000000 | 0.925890 | -0.162300 |
| **caffeine_100ml** | -0.299305 | -0.128281 | 0.925890 | 1.000000 | -0.089144 |
| **calories_100ml** | -0.130185 | 0.946101 | -0.162300 | -0.089144 | 1.000000 |

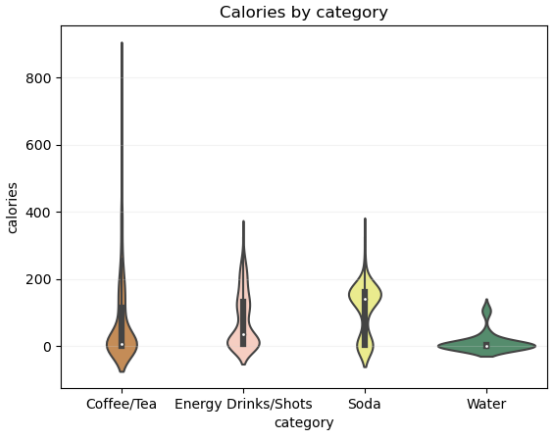
Both calories and caffeine compared to volume had a much weaker correlation than the original dataset, though, like Soda, calories and caffeine had a slightly stronger correlation.

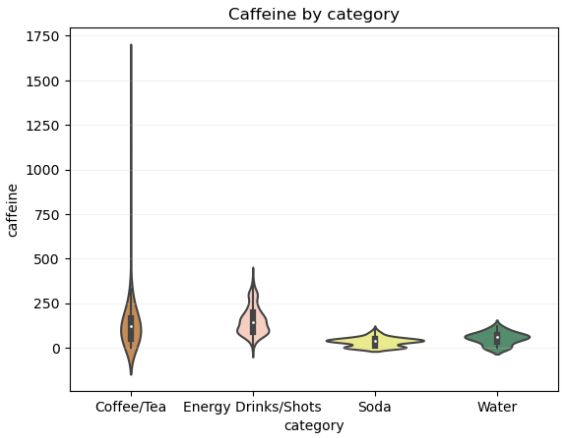
After completing the correlations, we made a scatter plot to show Calories vs. Caffeine as shown below:

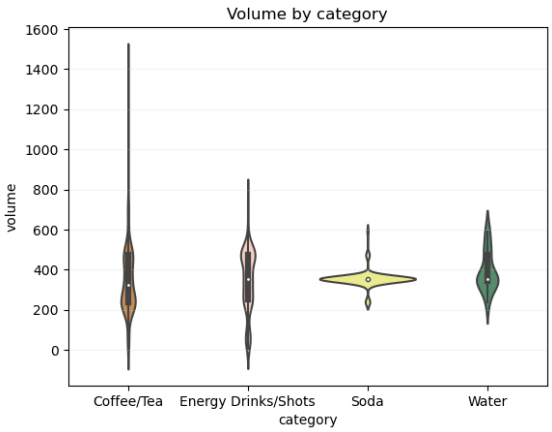



This graph shows us that Coffee/Tea was lower in calories and higher in caffeine on average, Energy Drinks and Shots were about equal in calories and higher in caffeine, Soda was higher in calories and lower in caffeine, and Water was lower in calories and caffeine on average.

Next, we made violin graphs that compare calories, caffeine, and volume. As stated above, Water and Soda ran bimodally; you can see below that the datasets were just too different to combine. Also, Coffee and Tea seemed to be way higher across the board, and Soda has more calories in comparison to volume and caffeine.




We also did some T-Tests to see how different the caffeine levels were statistically. You can see in the first T-Test, that Coffee and Tea, and Energy Drinks and Shots were pretty similar, though Energy Drinks and Shots were significantly different than Water. Soda and Water also seem to be statistically insignificant.

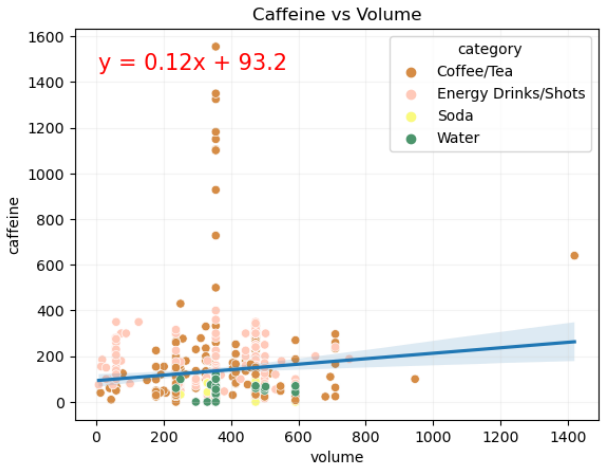
```
4]:  1 grp0 = df.loc[df.category == "Coffee/Tea","caffeine"]
     2 grp1 = df.loc[df.category == "Energy Drinks/Shots","caffeine"]
     3 grp2 = df.loc[df.category == "Soda","caffeine"]
     4 grp3 = df.loc[df.category == "Water","caffeine"]
```

```
5]:  1 st.ttest_ind(grp0,grp1)
5]: TtestResult(statistic=0.4278307582724591, pvalue=0.6689615437595988, df=492.0)
```

```
]:  1 st.ttest_ind(grp1,grp3)
]: TtestResult(statistic=6.4543005696776605, pvalue=4.79774554259269e-10, df=279.0)
```

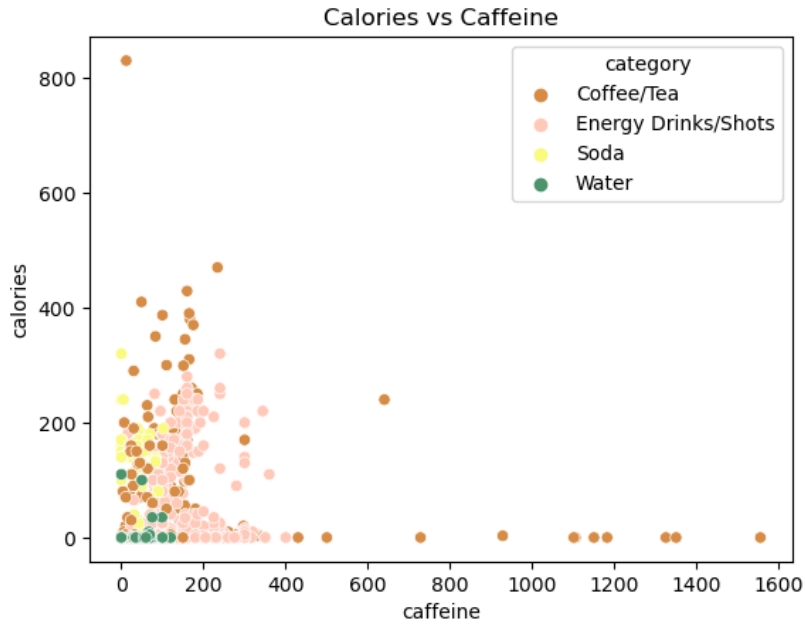
```
1 st.ttest_ind(grp2,grp3)
TtestResult(statistic=-3.3129377635121506, pvalue=0.0012376061281805858, df=114.0)
```

Next, we looked back at our research questions to see what answers we could gather. For our first research question, "Does coffee/tea have a lower volume-to-caffeine ratio than other drink categories?", we discovered that as expected coffee/tea has a lower volume-to-caffeine ratio than other categories. This is because many of the coffee/tea drinks were extremely high in caffeine.

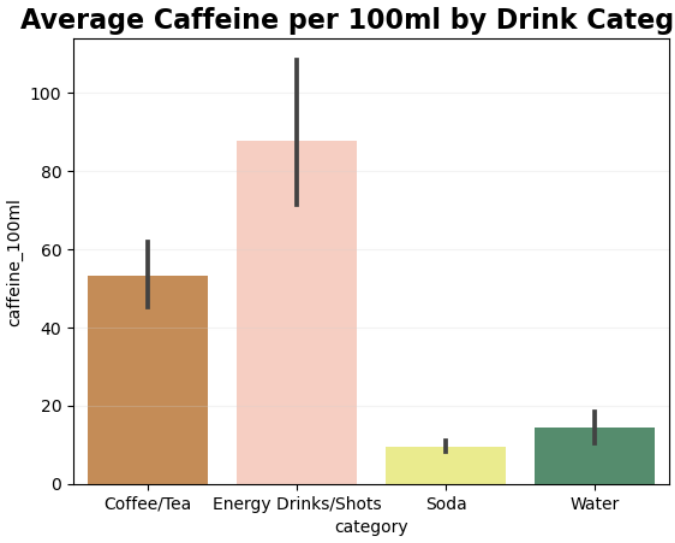



Similarly in the second research question, "Does soda have a higher calorie-to-caffeine ratio than other drink categories?", our results showed the expected, that yes soda does have a higher calorie-to-caffeine ratio than other drink categories. Soda's caffeine levels remained low

while some of the sodas were very high in calories.

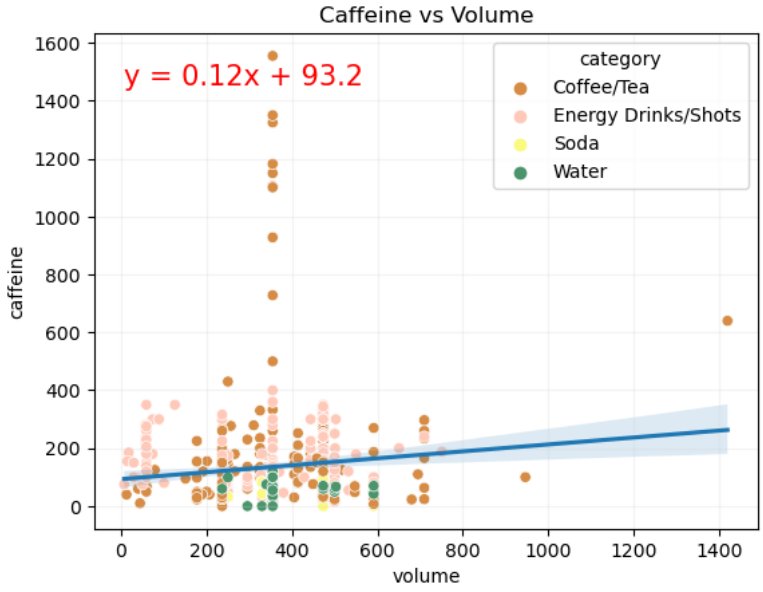



For the third research question, “Which drink type has the most caffeine per 100 mL?”, we discovered that energy drinks/shots had the most caffeine per 100 mL. This is likely caused by the fact that most of the energy shots are less than 100 mL, so comparing 100 mL quantities shows even higher caffeine levels than would be in a typical serving of an energy shot.

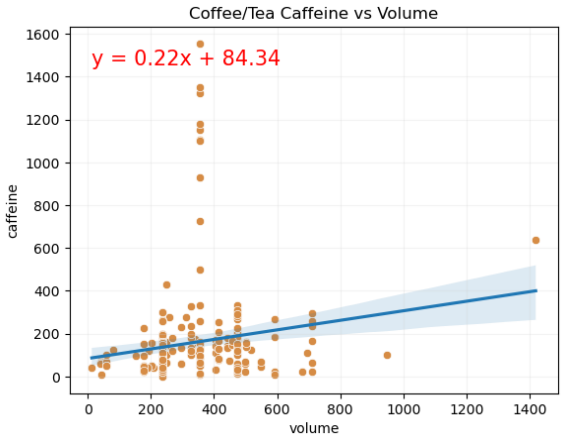


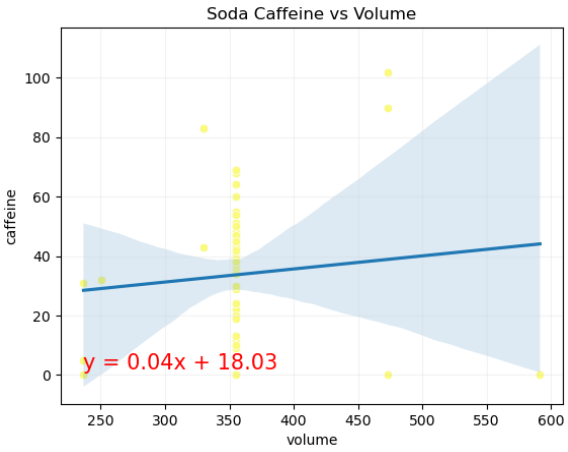

To further evaluate our dataset, we ran a regression comparing caffeine and volume. Our regression showed that the correlation between caffeine and volume is not very statistically significant. All drink types except coffee/tea had a weak correlation, so this result is expected.

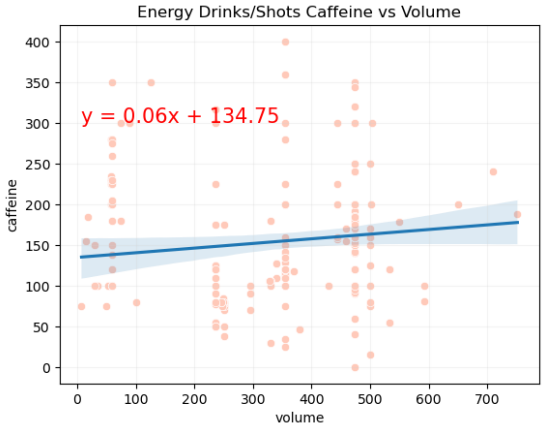
For example, energy drinks/shots all had relatively similar levels of caffeine despite the energy shots having very low volume levels and the energy drinks having higher volume levels.

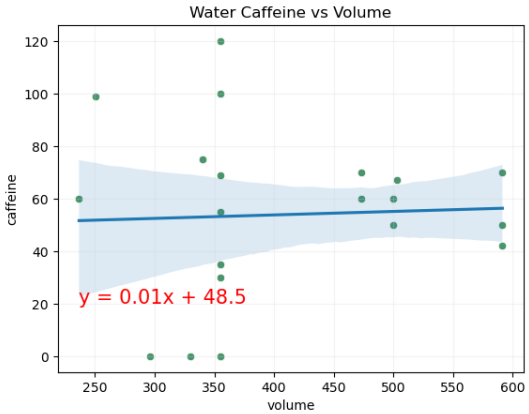



To confirm this, we ran regressions for each of the categories individually:



In conclusion, coffee/tea drinks are typically the most highly caffeinated, but energy drinks/shots have the most caffeine per 100 mL. Soda has significantly lower caffeine vs calorie levels on average. Water is low in both calories and caffeine on average.

In further studies, it may be more effective only to compare drinks from restaurants and coffee shops, or to compare personal-sized beverages from grocery stores. Further work could involve grouping brands to compare within a certain brand (e.g. which drink at Starbucks has the most caffeine?) or to compare between brands (e.g. is there more caffeine in a Starbucks iced coffee or a Dunkin iced coffee?). Furthermore, including prices would make this information even more valuable to consumers, as you could compare which drink gets you the best value if you are looking for more caffeine or more volume, for example.