Adel Khorramrouz

ak8480@rit.edu (mailto:ak8480@rit.edu)

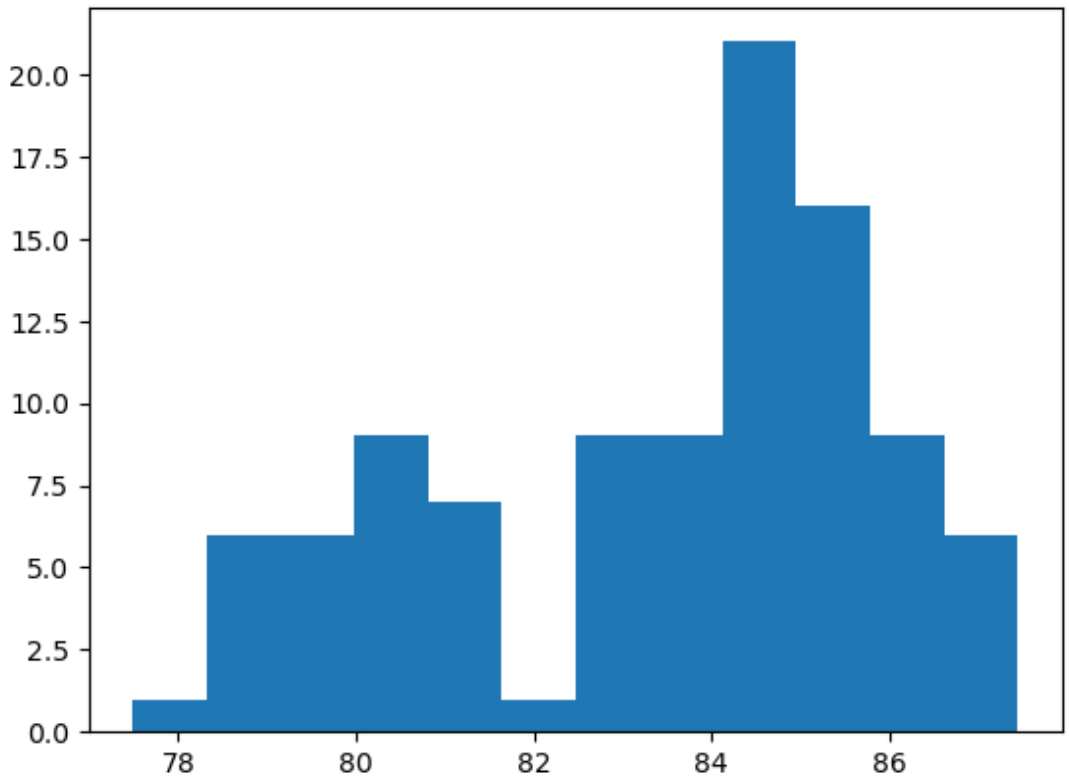
hw1 Stat 614

In [2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

Q1

Provide a univariate graphical summary of viscosity and comment on the shape of the distribution.

In [20]:
```
df = pd.read_csv("viscosity.txt", sep="\t",index_col=0)
plt.hist(df.Viscosity,bins=12)
plt.show()
```



As it is obvious in the histogram above (I used histogram because it is continues data) we have a bimodal distribution here

Provide numerical summaries of viscosity (measures of center and spread). Which measures would you use and why?

Since the data is not distributed symmetric (it is a bimodal data), it is not the best idea to use the mean, instead Mode will probably give a better understanding the data. to characterise how much data varries the IQR (85.3 - 81.08) explains data diversity better than the other metrics

In [23]:
```
1 df.describe().T
```

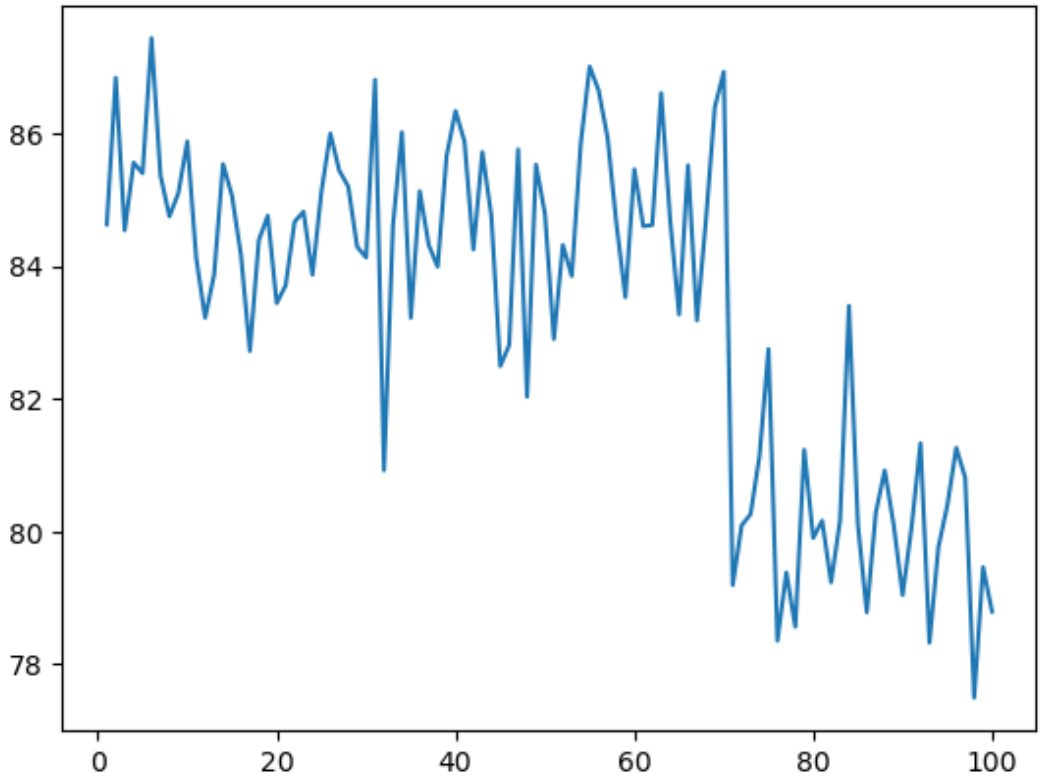
Out[23]:

	count	mean	std	min	25%	50%	75%	max
Viscosity	100.0	83.3643	2.535917	77.49	81.085	84.21	85.3775	87.44

Is this process stable over time? Explain. Provide a plot that helps explain your answer.

In [26]:
```
1 plt.plot(df.index, df.Viscosity)
```

Out[26]:
```
[<matplotlib.lines.Line2D at 0x1d5abab65b0>]
```



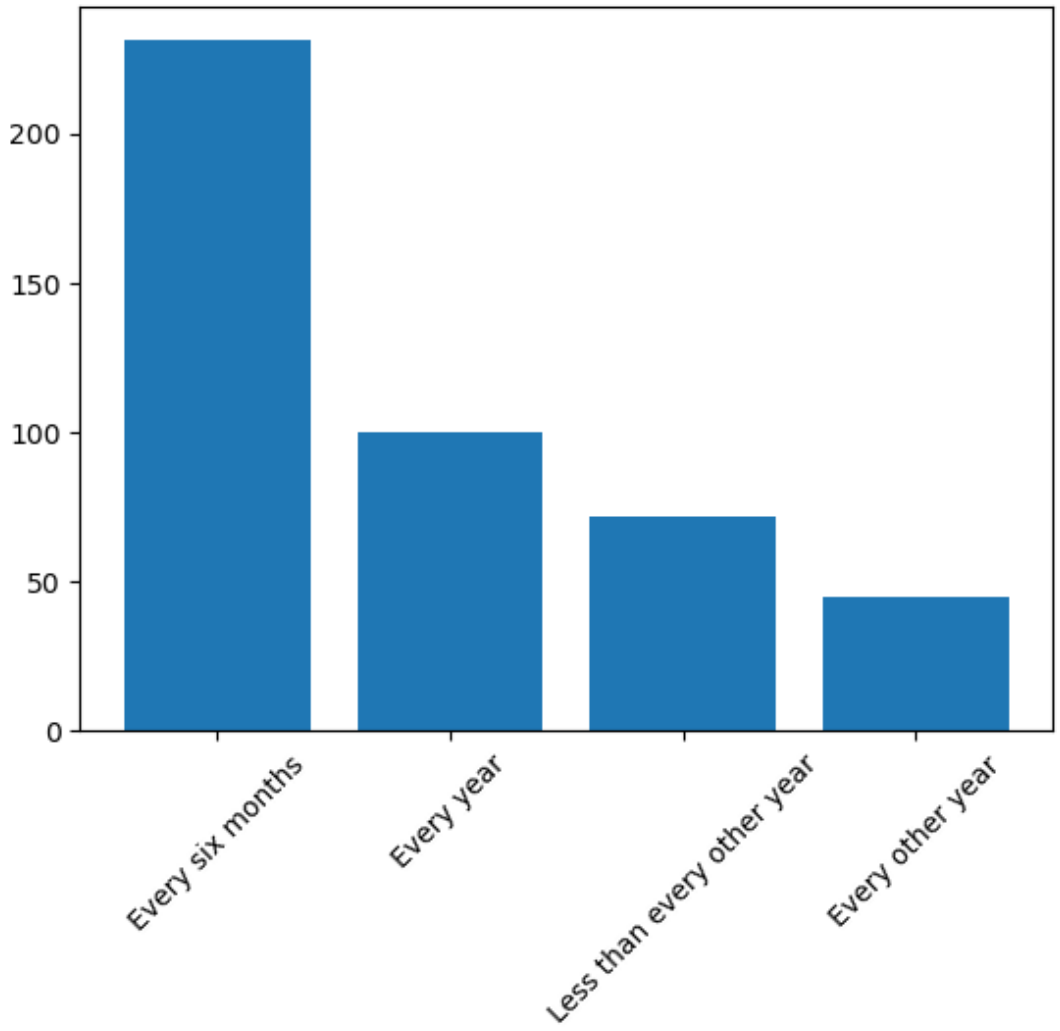
As it is obvious in the graph above, there is a big drop around day 60th. so however it is stable before and after the 60th days, yet there is a big drop happening there.

Q2

Provide a graphical summary of the frequency of teeth cleaning variable, include some comments on what you see.

In [3]:

```
df = pd.read_csv("tootthpaste.txt", sep="\t", index_col=0)

categories = pd.DataFrame(df['Frequency of Teeth Cleaning'].value_counts()
counts = pd.DataFrame(df['Frequency of Teeth Cleaning'].value_counts())['F

plt.bar(categories,counts)
plt.xticks(rotation=45)
plt.show()
```



This is a categorical data(it is ordinal) data so the best tool to visualise the data is a bar chart

B

For both variables individually, provide a table with the counts of each category and a table with the percentages or proportions of each category.

In [4]:
```
1 pd.DataFrame(df['Frequency of Teeth Cleaning'].value_counts())
```

Out[4]:

	Frequency of Teeth Cleaning
Every six months	231
Every year	100
Less than every other year	72
Every other year	45

In [5]:
```
1 pd.DataFrame(df['Frequency of Teeth Cleaning'].value_counts(normalize=True
```

Out[5]:

	Frequency of Teeth Cleaning
Every six months	0.515625
Every year	0.223214
Less than every other year	0.160714
Every other year	0.100446

In [6]:
```
1 pd.DataFrame(df['Single Status'].value_counts())
```

Out[6]:

	Single Status
Not Single	271
Single	177

In [7]:
```
1 pd.DataFrame(df['Single Status'].value_counts(normalize=True))
```

Out[7]:

	Single Status
Not Single	0.604911
Single	0.395089

C

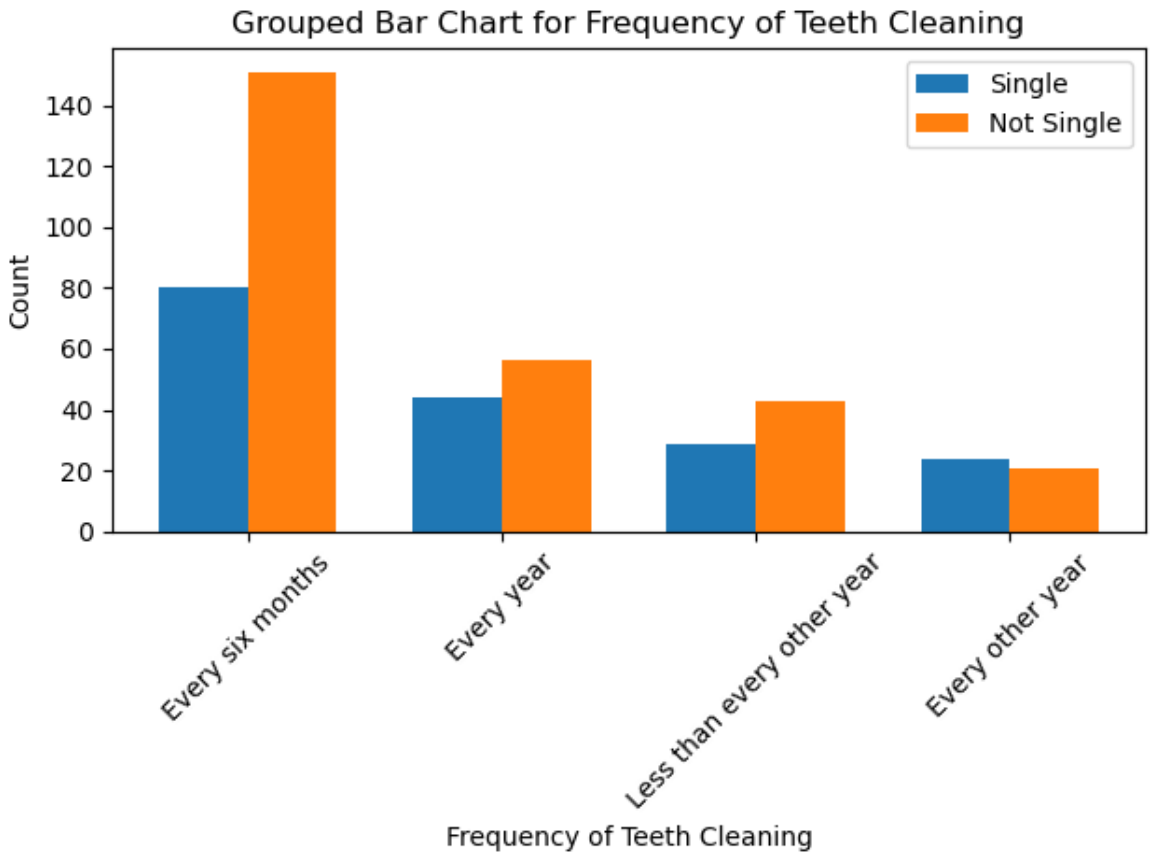
Do you think there is a relationship between relationship status and how frequently someone gets their teeth cleaned?

In [51]:

```
single = df.loc[df['Single Status'] == "Single"]
not_single = df.loc[~df.index.isin(single.index)]
single_counts = dict(single.drop(columns=['Single Status']).value_counts()
not_single_counts = dict(not_single.drop(columns=['Single Status']).value_
```

In [54]:

```
categories = [key[0] for key in single_counts.keys()]
values1 = list(single_counts.values())
values2 = list(not_single_counts.values())

# Create an index for the x-axis
x = np.arange(len(categories))
width = 0.35  # Width of the bars

# Create the figure and axis objects
fig, ax = plt.subplots()

# Plot the bars for the first dictionary
bar1 = ax.bar(x - width/2, values1, width, label='Single')
# Plot the bars for the second dictionary
bar2 = ax.bar(x + width/2, values2, width, label='Not Single')

# Set labels, title, and ticks
ax.set_xlabel('Frequency of Teeth Cleaning')
ax.set_ylabel('Count')
ax.set_title('Grouped Bar Chart for Frequency of Teeth Cleaning')
ax.set_xticks(x)
ax.set_xticklabels(categories, rotation=45)
ax.legend()

# Display the chart
plt.tight_layout()
plt.show()
```

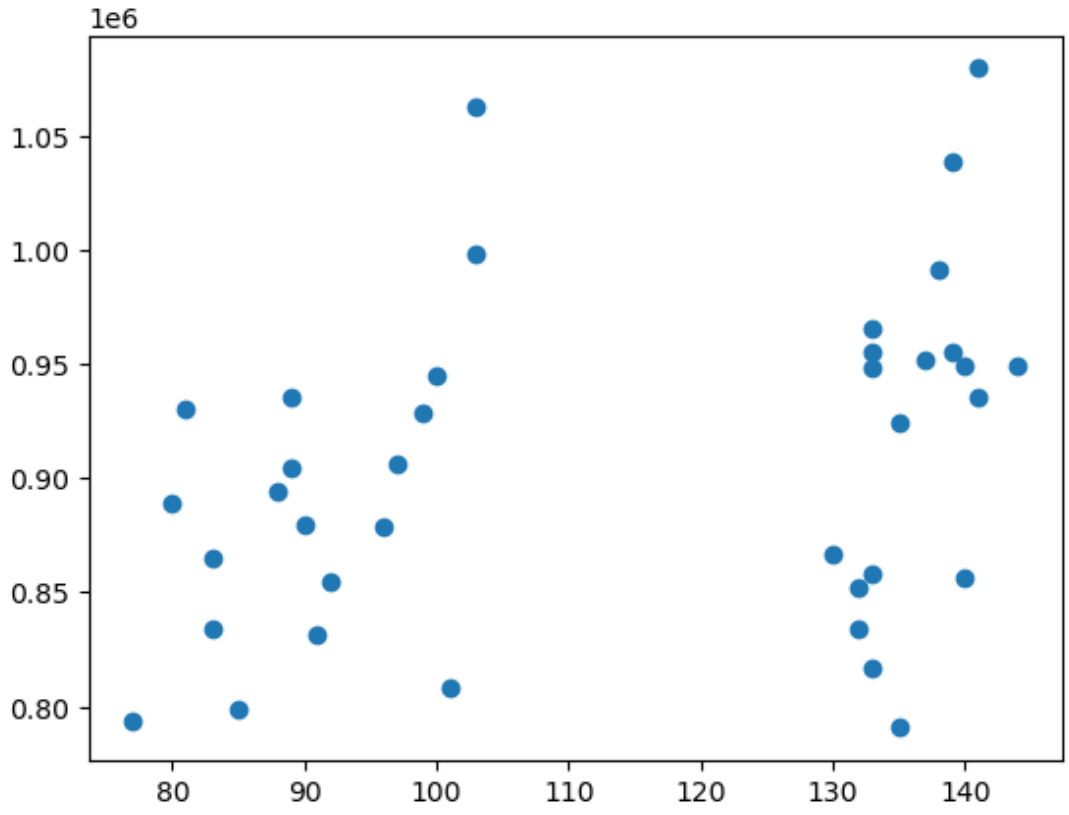


Based on the graph above those who clean their teeth more frequently are more likley to be in a relationship

Q3

In brainsize.txt, there are IQ scores and MRI data for 38 subjects. Is there a linear relationship between brain size (MRI count) and IQ scores (FSIQ)? Provide a plot that helps you answer this question.

In [5]:
```
df = pd.read_csv("brainsize.txt", sep='\t')
plt.scatter(df['FSIQ'],df['MRICount'])
plt.show()
```




```
<span style="font-size:20px; color:blue">
Based on the graph above there is no distinct linear relationship between theese two parameters
<span>
```

In [ ]:
```

```