In [1]:
```
NAME = "Dustin Seltz"
```

Purpose:

Combine the data we already have from earlier questions with the data from https://scriptin.github.io/kanji-frequency/ (https://scriptin.github.io/kanji-frequency/)

Input:

combined.csv

Output:

combined_with_frequency.csv

In [2]:
```
import pandas as pd
```

In [4]:
```
filename = "combined_genki.csv"
df = pd.read_csv(filename)
df
```

Out[4]:

	kanji	strokes	frequency	grade	jlpt	parts	radicals	on_readings	kun_readings	on_read
0	亜	7	1509.0	junior high	N1	{'一': 'http://jisho.org/search/%E4%B8%80%20%...	{'二': 'two'}	{'ア': 'http://jisho.org/search/%E4%BA%9C%20%...	{'つ.ぐ': 'http://jisho.org/search/%E4%BA%9C%2...	['亜 (indica
1	哀	9	1715.0	junior high	N1	{'亠': 'http://jisho.org/search/%E4%BA%A0%20%...	{'口': 'mouth, opening'}	{'アイ': 'http://jisho.org/search/%E5%93%80%20...	{'あわ.れ': 'http://jisho.org/search/%E5%93%80%...	['condole
2	挨	10	2258.0	junior high	NaN	{'ム': 'http://jisho.org/search/%E5%8E%B6%20%...	{'手 (扌)': 'hand'}	{'アイ': 'http://jisho.org/search/%E6%8C%A8%20...	{'ひら.く': 'http://jisho.org/search/%E6%8C%A8%...	['g
3	愛	13	640.0	grade 4	N3	{'冖': 'http://jisho.org/search/%E5%86%96%20%...	{'心 (忄, ⺗)': 'heart'}	{'アイ': 'http://jisho.org/search/%E6%84%9B%20...	{'いと.しい': 'http://jisho.org/search/%E6%84%9B...	affection
4	曖	17	NaN	junior high	NaN	{'冖': 'http://jisho.org/search/%E5%86%96%20%...	{'日': 'sun, day'}	{'アイ': 'http://jisho.org/search/%E6%9B%96%20...	{'くら.い': 'http://jisho.org/search/%E6%9B%96%...	['曖昧 【 ar
...	...	...	...	...	...	...	...	...	...	
2131	脇	10	1806.0	junior high	NaN	{'力': 'http://jisho.org/search/%E5%8A%9B%20%...	{'肉 (月)': 'meat'}	{'キョウ': 'http://jisho.org/search/%E8%84%87%2...	{'わき': 'http://jisho.org/search/%E8%84%87%20...	['脇侍 image (e
2132	惑	12	777.0	junior high	N1	{'口': 'http://jisho.org/search/%E5%8F%A3%20%...	{'心 (忄, ⺗)': 'heart'}	{'ワク': 'http://jisho.org/search/%E6%83%91%20...	{'まど.う': 'http://jisho.org/search/%E6%83%91%...	['planet', '
2133	枠	8	922.0	junior high	N1	{'十': 'http://jisho.org/search/%E5%8D%81%20%...	{'木': 'tree'}	NaN	{'わく': 'http://jisho.org/search/%E6%9E%A0%20...	['frame
2134	湾	12	545.0	junior high	N2	{'亠': 'http://jisho.org/search/%E4%BA%A0%20%...	{'水 (氵, 氺)': 'water'}	{'ワン': 'http://jisho.org/search/%E6%B9%BE%20...	{'いりえ': 'http://jisho.org/search/%E6%B9%BE%2...	['湾 inlet', '
2135	腕	12	1163.0	junior high	N2	{'⺼': 'http://jisho.org/search/%E5%8D%A9%20%...	{'肉 (月)': 'meat'}	{'ワン': 'http://jisho.org/search/%E8%85%95%20...	{'うで': 'http://jisho.org/search/%E8%85%95%20...	['腕 phys

2136 rows × 13 columns

In [ ]:
```
base_path = "./"
twitter_filename = base_path + "KanjiFrequencyOnTwitter"
twitter_df = pd.read_csv(twitter_filename)
wikipedia_filename = base_path + "KanjiFrequencyOnWikipedia"
wikipedia_df = pd.read_csv(wikipedia_filename)
news_filename = base_path + "KanjiFrequencyOnNews"
news_df = pd.read_csv(news_filename)
aozora_filename = base_path + "KanjiFrequencyOnAozora"
aozora_df = pd.read_csv(aozora_filename)
twitter_df.head()
```

In [ ]:
```
freq_datasets = [(twitter_df, "Twitter"), (wikipedia_df, "Wikipedia"), (news_df, "N
ews"), (aozora_df, "Aozora")]
```

In [ ]:
```
#Actually it's easier to just use pd df library functions.
#combine_on = "kanji"
#def combine(result_df, incoming_df, label):
#    label = label+"_"
#    for kanji in result_df[combine_on]:
#        match = incoming_df.loc[incoming_df[combine_on] == kanji]
#        for column in match:
#            result_df[combine_on][column] = match[column]
```

In [ ]:
```
len(df)
```

In [ ]:
```
#I want to add that frequency information to all the characters we have in the comb
ined.csv dataset.
#I will add: Rank of frequency, number of appearances, %. For each dataset.
for (dataset, label) in freq_datasets:
    dataset['temporary'] = dataset.index + 1
    dataset.columns = ["kanji", "Number of Appearances on "+label, "Percentage of A
ppearances on "+label, "Rank of Appearances on "+label]
    #combine(df, dataset, label)
    #https://stackoverflow.com/q/28174752
    df = df.merge(dataset, on="kanji", how="outer")
```

In [ ]:
```
df.head()
```

In [ ]:
```
#Verify that the numbers are matched up correctly.
for (dataset, _) in freq_datasets:
    print(dataset.loc[dataset["kanji"] == "亜"])
```

In [ ]:
```
#Verify that kanji that didn't match were dropped.
len(df)
#Whoops. Well, we can fix that.
```

In [ ]:
```
#Any one of the original columns should do.
column_it_should_have = "strokes"
#https://stackoverflow.com/a/13413845
df = df[pd.notnull(df[column_it_should_have])]
```

In [ ]:
```
print(len(df))
#Fixed! Looks good.
df.head()
```

In [ ]:
```
#Store the combined data
file_name = "combined_with_frequency.csv"
df.to_csv(file_name, index=False)
```

In [ ]: