part2

February 12, 2020

1 Data of Interest

Our central focus will be on review.json. However, we will also make heavy use of the user, tip, and business data sets. Currently, we do not have plans to use the photo or check-in sets, but check-in may be integrated if we think there are possible insights to be gained from its use.

2 Data Preprocessing

Our project requires no specific data cleaning to be done because due to the nature of the questions we are analyzing, there is no specific set of mistakes that we would be able to check for and fix (e.g., if someone selects the wrong business, giving us the wrong business_id).

In terms of reorganization, this would simply take the form of loading the data into dataframes that include just the data we want to analyze for each problem we are considering. Because of the large quantity of reviews in the dataset, we will consider a random sample of 10% of businesses.

2.0.1 Getting the 10% random sample of businesses:

[1]:
```
# import pandas as pd
# import numpy as np

# path = "../yelp_dataset/business.json"

# df = pd.read_json(path, lines=True)

# df_sample = df.sample(19261)

# df_sample.to_json(r'business_sample.json',orient='records')
```

2.0.2 Getting all reviews for our random sample of businesses:

[20]:
```
# import pandas as pd
# import numpy as np
# import json

# df = pd.read_json("business_sample.json")
# all_ids = df['business_id'].to_numpy()
```

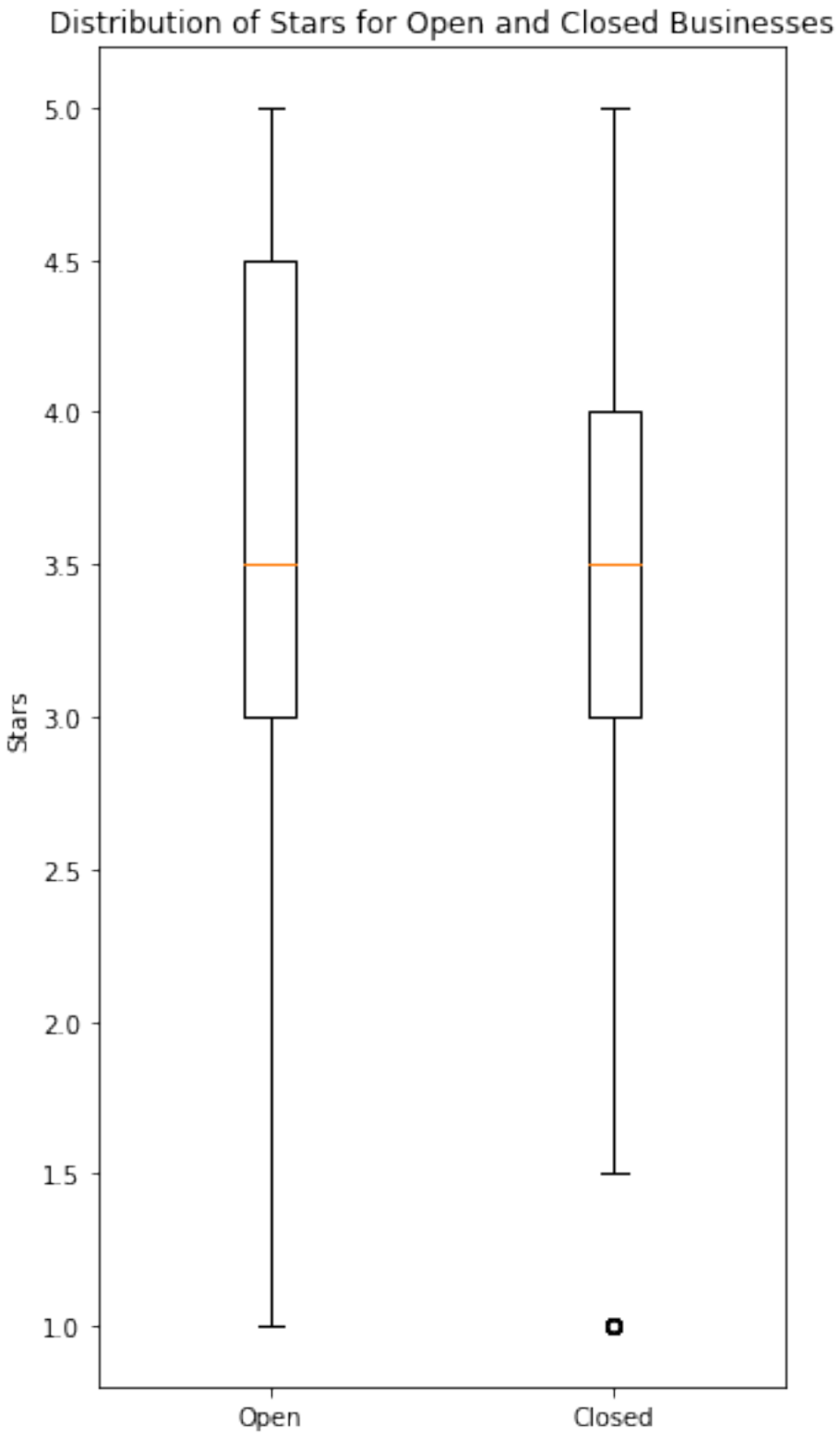
```
# business_reviews = []
# for l in open("../yelp_dataset/review.json").readlines():
#     data = json.loads(l)
#     if data["business_id"] in all_ids:
#         business_reviews.append(data)

# df_reviews = pd.DataFrame.from_records(business_reviews)

# df_reviews.to_json(r'reviews_sample.json',orient='records')
```

2.0.3 Are the distrbutions of stars for closed and open businesses the same?

We will use a two-sample t-test to analyze whether the distributions of stars for closed and open business are the same.

[6]:
```
import pandas as pd
import numpy as np
import json
import scipy.stats as stats
import matplotlib.pyplot as plt

df_business = pd.read_json("business_sample.json")

open_stars = df_business[df_business['is_open']==1]['stars'].to_numpy()
closed_stars = df_business[df_business['is_open']==0]['stars'].to_numpy()

t_stat, p_value = stats.ks_2samp(open_stars, closed_stars)
print(f't-stat: {t_stat} \t p-value: {p_value}')

data = [np.reshape(open_stars, (-1,1)), np.reshape(closed_stars, (-1,1))]
plt.figure(figsize=(5,10))
plt.boxplot(data)
plt.xticks([1,2],['Open','Closed'])
plt.ylabel('Stars')
plt.title('Distribution of Stars for Open and Closed Businesses')
plt.show()
```

```
t-stat: 0.11789647564234917      p-value: 8.70124840960327e-35
```


Distribution of Stars for Open and Closed Businesses
Stars
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Open
Closed

2.0.4 Insights

The p-value is << 0.05, so we reject the null hypothesis (that the distribution of stars is the same for both open and closed businesses) and have strong evidence that the distributions are different. The boxplot visually confirms this.

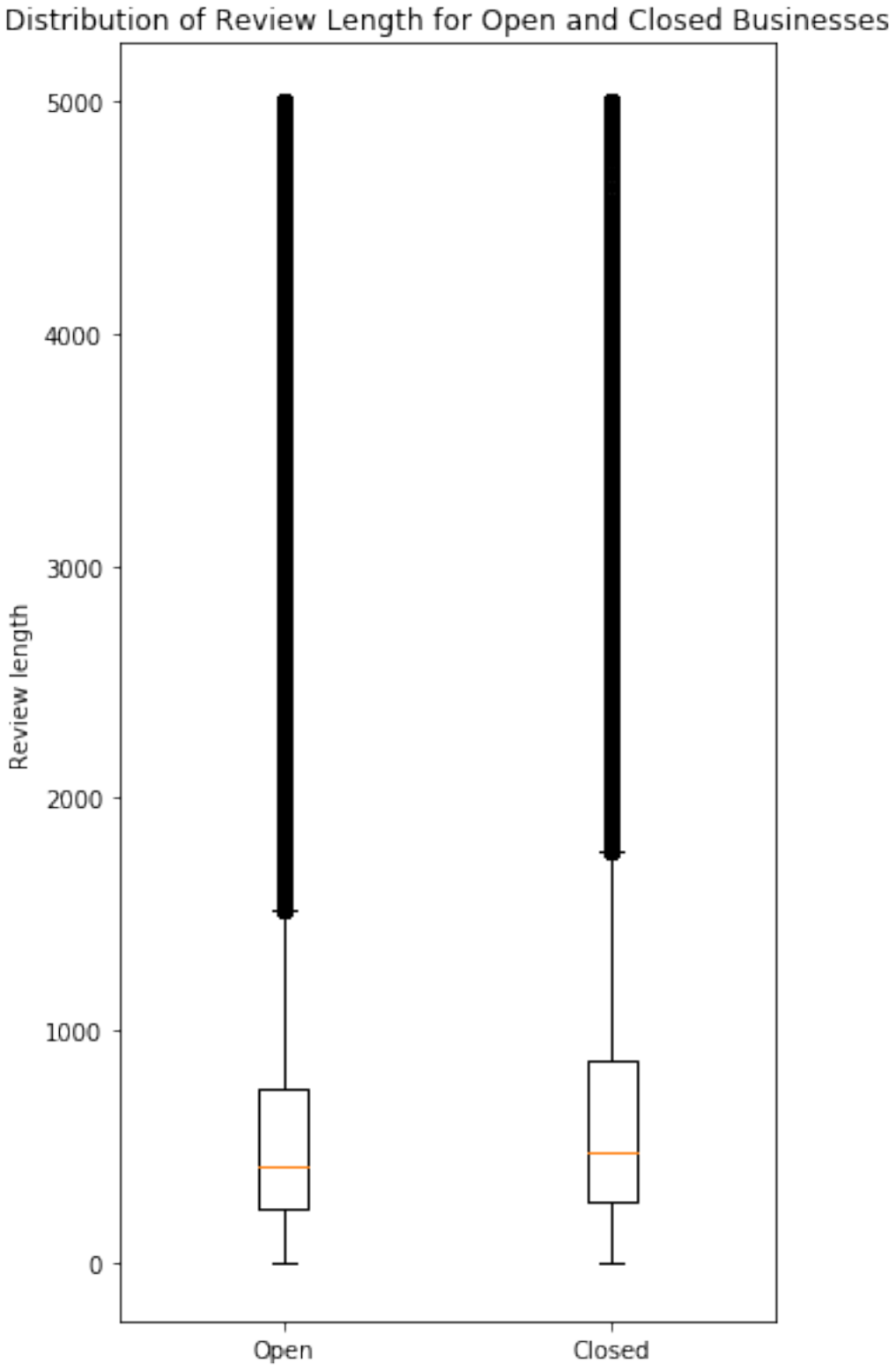
2.0.5 Is there a difference between review length for closed and open between?

We will use a two-sample t-test to analyze whether the distributions of review lengths for closed and open business are the same.

[8]:
```
import pandas as pd
import numpy as np
import json
import scipy.stats as stats
import matplotlib.pyplot as plt

df_review = pd.read_json('reviews_sample.json')
df_business = pd.read_json("business_sample.json")

df_join = df_review[['business_id', 'text']].join(df_business[['business_id',␣
↪'is_open']].set_index('business_id'), on='business_id')

open_text = df_join[df_join['is_open']==1]['text']
closed_text = df_join[df_join['is_open']==0]['text']

text_length = np.vectorize(len)
open_lengths = text_length(open_text.astype(str))
closed_lengths = text_length(closed_text.astype(str))

t_stat, p_value = stats.ks_2samp(open_lengths, closed_lengths)
print(f't-stat: {t_stat} \t p-value: {p_value}\n')
print(f'Open length mean: {np.mean(open_lengths)} \t Closed length mean: {np.
↪mean(closed_lengths)}')
print(f'Open length median: {np.median(open_lengths)} \t Closed length median:␣
↪{np.median(closed_lengths)}')

data = [np.reshape(open_lengths, (-1,1)), np.reshape(closed_lengths, (-1,1))]

plt.figure(figsize=(5,10))
plt.boxplot(data)
plt.xticks([1,2],['Open','Closed'])
plt.ylabel('Review length')
plt.title('Distribution of Review Length for Open and Closed Businesses')
plt.show()
```

```
t-stat: 0.06626003682356818      p-value: 2.776776516482881e-304
```

```
Open length mean: 591.964995085665        Closed length mean: 667.7090485637885
Open length median: 417.0         Closed length median: 482.0
```



2.0.6 Insights

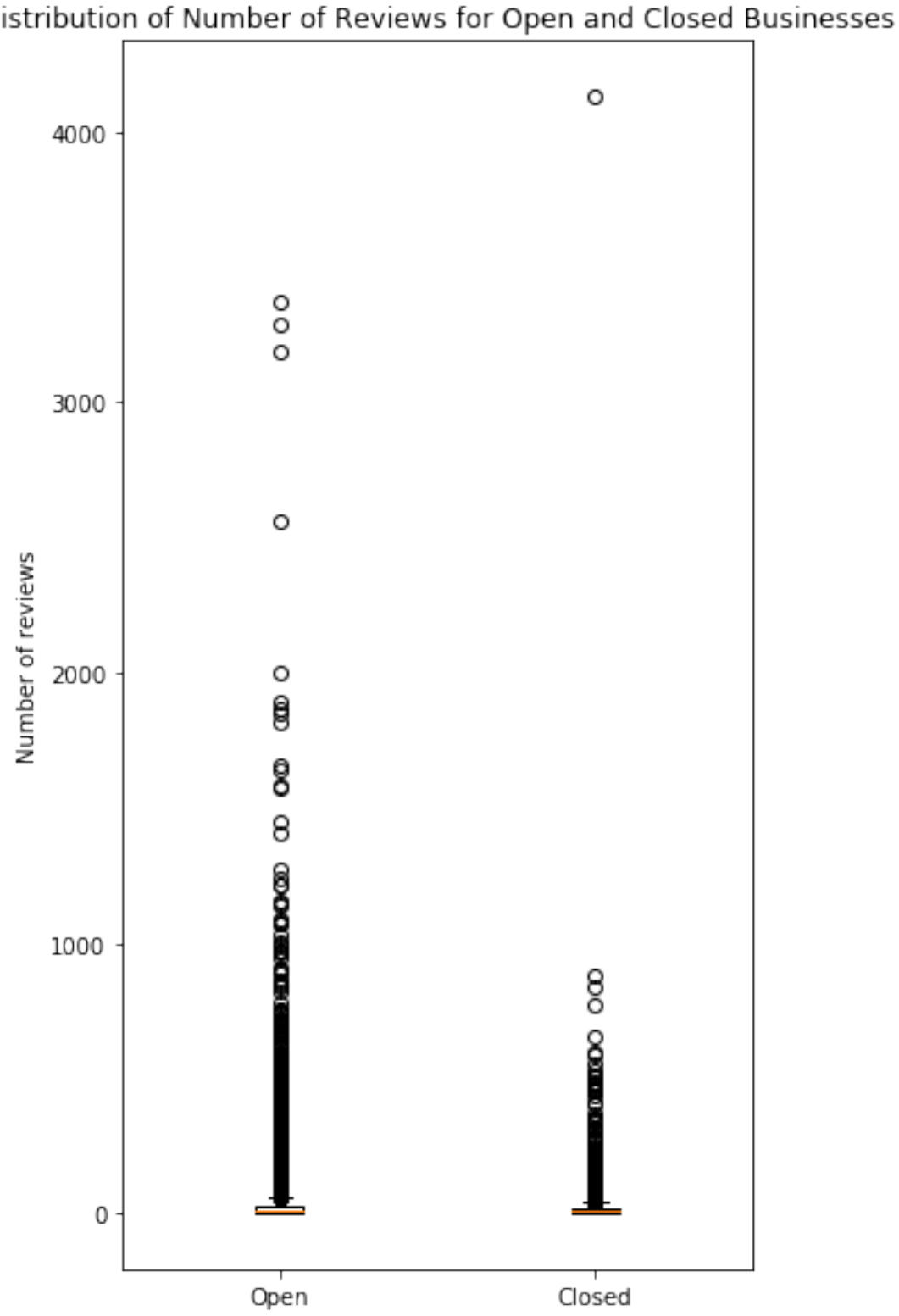
The p-value is << 0.05, so we reject the null hypothesis (that the distribution of review lengths is the same for both open and closed businesses) and have strong evidence that the distributions are different. The boxplot visually confirms this. From examining the mean length of each, it appears that closed reviews are typically longer.

2.0.7 Number of reviews for open versus closed businesses

We will use a two-sample t-test to analyze whether the distributions of the number of reviews for closed and open business are the same.

[16]:
```
import pandas as pd
import numpy as np
import json
import scipy.stats as stats
import matplotlib.pyplot as plt

df_review = pd.read_json('reviews_sample.json')
df_business = pd.read_json("business_sample.json")

df_join = df_review[['business_id', 'text']].join(df_business[['business_id',␣
↪'is_open']].set_index('business_id'), on='business_id')

open_reviews = df_join[df_join['is_open']==1].
↪groupby('business_id')['business_id'].count().to_numpy()
closed_reviews = df_join[df_join['is_open']==0].
↪groupby('business_id')['business_id'].count().to_numpy()

t_stat, p_value = stats.ks_2samp(open_reviews, closed_reviews)
print(f't-stat: {t_stat} \t p-value: {p_value}')

print(f'Open reviews mean: {np.mean(open_reviews)} \t Closed reviews mean: {np.
↪mean(closed_reviews)}')
print(f'Open reviews median: {np.median(open_reviews)} \t Closed reviews median:
↪ {np.median(closed_reviews)}')

data = [np.reshape(open_reviews, (-1,1)), np.reshape(closed_reviews, (-1,1))]

plt.figure(figsize=(5,10))
plt.boxplot(data)
plt.xticks([1,2],['Open','Closed'])
plt.ylabel('Number of reviews')
plt.title('Distribution of Number of Reviews for Open and Closed Businesses')
plt.show()
```

```
t-stat: 0.05053950965687193      p-value: 9.696907760450336e-07
Open reviews mean: 35.82823484704541     Closed reviews mean: 26.711981566820278
Open reviews median: 10.0        Closed reviews median: 8.0
```



2.0.8 Insights

Based on the above test, there is a difference in the distribution of the number of reviews for each open and closed business. However, this difference (closed have less reviews on average) is likely due to the fact that the business is closed, and therefore has had less time to accumulate reviews than still-open businesses.

2.0.9 Proprtion of each number of stars for open versus closed businesses

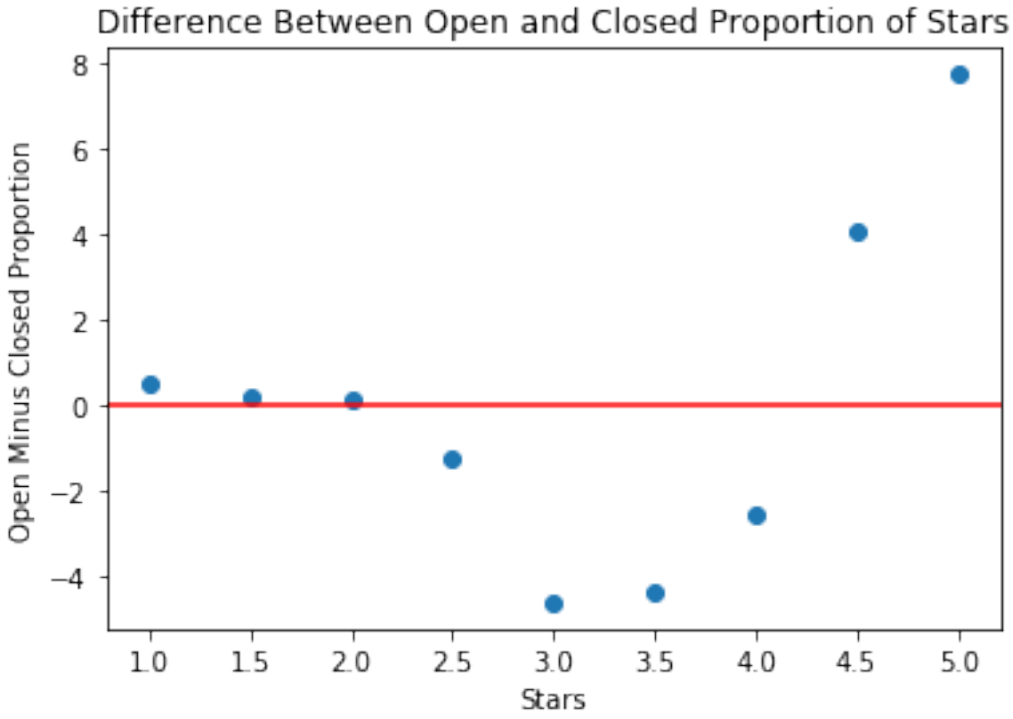
We will compare the proportion of each number of stars for open versus closed businsses and plot them.

[28]:
```
import pandas as pd
import numpy as np
import json
import scipy.stats as stats
import matplotlib.pyplot as plt

df_business = pd.read_json("business_sample.json")

open_stars = df_business[df_business['is_open']==1]['stars'].to_numpy()
closed_stars = df_business[df_business['is_open']==0]['stars'].to_numpy()

unique_open, counts_open = np.unique(open_stars, return_counts=True)
unique_closed, counts_closed = np.unique(closed_stars, return_counts=True)

total_open = np.sum(counts_open)
total_closed = np.sum(counts_closed)

print('Proportions:\n')
print(f'1 Star: \t Open: {counts_open[0]/total_open*100} \t Closed␣
 ↪{counts_closed[0]/total_closed*100} \t Diff: {counts_open[0]/
 ↪total_open*100-counts_closed[0]/total_closed*100} \n')
print(f'1.5 Stars: \t Open: {counts_open[1]/total_open*100} \t Closed␣
 ↪{counts_closed[1]/total_closed*100} \t Diff: {counts_open[1]/
 ↪total_open*100-counts_closed[1]/total_closed*100}\n')
print(f'2 Stars: \t Open: {counts_open[2]/total_open*100} \t Closed␣
 ↪{counts_closed[2]/total_closed*100} \t Diff: {counts_open[2]/
 ↪total_open*100-counts_closed[2]/total_closed*100}\n')
print(f'2.5 Stars: \t Open: {counts_open[3]/total_open*100} \t Closed␣
 ↪{counts_closed[3]/total_closed*100} \t Diff: {counts_open[3]/
 ↪total_open*100-counts_closed[3]/total_closed*100}\n')
print(f'3 Stars: \t Open: {counts_open[4]/total_open*100} \t Closed␣
 ↪{counts_closed[4]/total_closed*100} \t Diff: {counts_open[4]/
 ↪total_open*100-counts_closed[4]/total_closed*100}\n')
```

```
print(f'3.5 Stars: \t Open: {counts_open[5]/total_open*100} \t Closed␣
 ↪{counts_closed[5]/total_closed*100} \t Diff: {counts_open[5]/
 ↪total_open*100-counts_closed[5]/total_closed*100}\n')
print(f'4 Stars: \t Open: {counts_open[6]/total_open*100} \t Closed␣
 ↪{counts_closed[6]/total_closed*100} \t Diff: {counts_open[6]/
 ↪total_open*100-counts_closed[6]/total_closed*100}\n')
print(f'4.5 Star: \t Open: {counts_open[7]/total_open*100} \t Closed␣
 ↪{counts_closed[7]/total_closed*100} \t Diff: {counts_open[7]/
 ↪total_open*100-counts_closed[7]/total_closed*100}\n')
print(f'5 Stars: \t Open: {counts_open[8]/total_open*100} \t Closed␣
 ↪{counts_closed[8]/total_closed*100} \t Diff: {counts_open[8]/
 ↪total_open*100-counts_closed[8]/total_closed*100}\n')

diff = []
for i in range(0,len(counts_open)):
    diff.append(counts_open[i]/total_open*100-counts_closed[i]/total_closed*100)

plt.figure()
plt.scatter(unique_open, diff)
plt.axhline(y=0, color='r', linestyle='-')
plt.ylabel('Open Minus Closed Proportion')
plt.xlabel('Stars')
plt.title('Difference Between Open and Closed Proportion of Stars')
plt.show()
```

```
Proportions:

1 Star:          Open: 2.634745709037938         Closed 2.1025345622119813
Diff: 0.5322111468259565

1.5 Stars:       Open: 2.685413895750206         Closed 2.4769585253456223
Diff: 0.20845537040458373

2 Stars:         Open: 6.168851732218633         Closed 6.0195852534562215
Diff: 0.14926647876241184

2.5 Stars:       Open: 9.569953765279624         Closed 10.800691244239632
Diff: -1.2307374789600072

3 Stars:         Open: 12.673380201406042        Closed 17.252304147465438
Diff: -4.578923946059396

3.5 Stars:       Open: 17.328519855595665        Closed 21.687788018433178
Diff: -4.3592681628375125

4 Stars:         Open: 18.34188358984103         Closed 20.852534562211982
Diff: -2.5106509723709536
```

```
4.5 Star:        Open: 14.871112800050668        Closed 10.800691244239632
Diff: 4.070421555811036

5 Stars:         Open: 15.72613845082019         Closed 8.006912442396313
Diff: 7.7192260084238775
```



2.0.10 Insights

Both 4.5 and 5 star reviews are very significant with respect to whether a business is open. Both open and closed businsses have a similar proportion of 1 through 2 star reviews, and closed businsses have a slightly larger proportion of 2.5 through 4 star reviews, but open businesses have a much larger proportion of 4.5 to 5 star reviews than closed businesses. This is visually shown in the plot.