Italian vs. Mexican Food

The below script provides an analytic approach for assessing the American preference of Italian vs. Mexican food. Using data from the US Census and the Yelp API, the script randomly selects over 500 zip codes and aggregates the reviews of the 20 most popular Italian and Mexican restaurants in each area. Summary data is then reported using Python Pandas.

[548]
```
# Dependencies
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import requests
import time
import json
import seaborn
from scipy.stats import ttest_ind

# Yelp API Key
ykey_id = "1GwZyE0zIjSujpHtlMnodQ"
ykey_secret = "mcTmghB48JIH0xoNWLldvsX9uIiOLQfdi0gR8LWdFt02lboCAF9vxSSd1M
ykey_access_token = "gl6k6JmewUhzjMVBv0I2x4Bz_NRiEggSqjlGbTaejmbzvBJXgI36
```

Zip Code Sampling

[322]
```
# Import the census data into a Pandas DataFrame
census_pd = pd.read_csv("Census_Data.csv")

# Preview the data
census_pd.head()
```

	Zipcode	Address	Population	Median Age	Household Income	Per Capita Income

	Zipcode	Address	Population	Median Age	Household Income	Per Capita Income
0	15081	South Heights,	342	50.2	31500.0	22177

		PA 15081, USA				
1	20615	Broomes Island, MD 20615, USA	424	43.4	114375.0	43920
2	50201	Nevada, IA 50201, USA	8139	40.4	56619.0	28908
3	84020	Draper, UT 84020, USA	42751	30.4	89922.0	33164
4	39097	Louise, MS 39097, USA	495	58.0	26838.0	17399

[323]
```
# Sell all zip codes with a population over 1000 from a set of randomly s
selected_zips = census_pd.sample(n=700)
selected_zips = selected_zips[selected_zips["Population"].astype(int) > 1

# Visualize
selected_zips.head()
```

	Zipcode	Address	Population	Median Age	Household Income	Per Capita Income
621	29468	Pineville, SC 29468, USA	1768	43.8	19663.0	28526
226	25414	Charles Town, WV 25414, USA	17147	40.1	72833.0	32308
233	93545	Lone Pine, CA 93545, USA	2214	40.6	32473.0	18444
67	60565	Naperville, IL 60565, USA	40864	40.8	113581.0	45408
235	3064	Nashua, NH 03064, USA	14533	40.6	64026.0	34045

[324]
```
# Show the total number of zip codes that met our population cut-off
selected_zips.count()
```

```
Zipcode             521
Address             521
Population          521
Median Age          521
```

```
Household Income     520
Per Capita Income    521
dtype: int64
```

[325]
```
# Show the average population of our representive sample set
selected_zips["Population"].mean()
```

```
13860.940499040307
```

[326]
```
# Show the average population of our representive sample set
selected_zips["Household Income"].mean()
```

```
56293.278846153844
```

[327]
```
# Show the average population of our representive sample set
selected_zips["Median Age"].mean()
```

```
40.02053742802301
```

Yelp Data Retrieval

[328]
```
# Create Two DataFrames to store the Italian and the Mexican Data
italian_data = pd.DataFrame();
mexican_data = pd.DataFrame();

# Setup the DataFrames to have appropriate columns
italian_data["Zip Code"] = ""
italian_data["Italian Review Count"] = ""
italian_data["Italian Average Rating"] = ""
italian_data["Italian Weighted Rating"] = ""

mexican_data["Zip Code"] = ""
mexican_data["Mexican Review Count"] = ""
mexican_data["Mexican Average Rating"] = ""
mexican_data["Mexican Weighted Rating"] = ""

# Include Yelp Token
headers = {"Authorization": "Bearer gl6k6JmewUhzjMVBv0I2x4Bz_NRiEggSqjlGb
counter = 0

# Loop through every zip code
for index, row in selected_zips.iterrows():

    # Add to counter
```

```
        counter = counter + 1

        # Create two endpoint URLs:
        target_url_italian = "https://api.yelp.com/v3/businesses/search?term=
        target_url_mexican = "https://api.yelp.com/v3/businesses/search?term=

        # Print the URLs to ensure logging
        print(counter)
        print(target_url_italian)
        print(target_url_mexican)

        # Get the Yelp Reviews
        yelp_reviews_italian = requests.get(target_url_italian, headers=heade
        yelp_reviews_mexican = requests.get(target_url_mexican, headers=heade

        # Calculate the total reviews and weighted rankings
        italian_review_count = 0
        italian_weighted_review = 0

        mexican_review_count = 0
        mexican_weighted_review = 0

        try:

            # Loop through all records to calculate the review count and weig
            for business in yelp_reviews_italian["businesses"]:

                italian_review_count = italian_review_count + business["revie
                italian_weighted_review = italian_weighted_review + business[

            for business in yelp_reviews_mexican["businesses"]:
                mexican_review_count = mexican_review_count + business["revie
                mexican_weighted_review = mexican_weighted_review + business[

            # Append the data to the appropriate column of the data frames
            italian_data.set_value(index, "Zip Code", row["Zipcode"])
            italian_data.set_value(index, "Italian Review Count", italian_rev
            italian_data.set_value(index, "Italian Average Rating", italian_w
            italian_data.set_value(index, "Italian Weighted Rating", italian_

            mexican_data.set_value(index, "Zip Code", row["Zipcode"])
            mexican_data.set_value(index, "Mexican Review Count", mexican_rev
            mexican_data.set_value(index, "Mexican Average Rating", mexican_w
            mexican_data.set_value(index, "Mexican Weighted Rating", mexican_

        except:
            print("Uh oh")
```

```
1
https://api.yelp.com/v3/businesses/search?term=Italian&location=29468
```

```
https://api.yelp.com/v3/businesses/search?term=Mexican&location=29468
2
https://api.yelp.com/v3/businesses/search?term=Italian&location=25414
https://api.yelp.com/v3/businesses/search?term=Mexican&location=25414
3
https://api.yelp.com/v3/businesses/search?term=Italian&location=93545
https://api.yelp.com/v3/businesses/search?term=Mexican&location=93545
4
https://api.yelp.com/v3/businesses/search?term=Italian&location=60565
https://api.yelp.com/v3/businesses/search?term=Mexican&location=60565
5
https://api.yelp.com/v3/businesses/search?term=Italian&location=3064
https://api.yelp.com/v3/businesses/search?term=Mexican&location=3064
6
https://api.yelp.com/v3/businesses/search?term=Italian&location=38049
https://api.yelp.com/v3/businesses/search?term=Mexican&location=38049
7
https://api.yelp.com/v3/businesses/search?term=Italian&location=28529
https://api.yelp.com/v3/businesses/search?term=Mexican&location=28529
8
https://api.yelp.com/v3/businesses/search?term=Italian&location=93428
https://api.yelp.com/v3/businesses/search?term=Mexican&location=93428
9
https://api.yelp.com/v3/businesses/search?term=Italian&location=41564
https://api.yelp.com/v3/businesses/search?term=Mexican&location=41564
10
https://api.yelp.com/v3/businesses/search?term=Italian&location=98106
https://api.yelp.com/v3/businesses/search?term=Mexican&location=98106
11
https://api.yelp.com/v3/businesses/search?term=Italian&location=73014
https://api.yelp.com/v3/businesses/search?term=Mexican&location=73014
```

[329]
```
# Preview Italian Data
italian_data.head()
```

	Zip Code	Italian Review Count	Italian Average Rating	Italian Weighted Rating
621	29468	77	3.35065	258
226	25414	135	3.41111	460.5
233	93545	702	4.10684	2883
67	60565	2829	3.92807	11112.5
235	3064	1253	3.77813	4734

[330]
```
# Preview Mexican Data
mexican_data.head()
```

	Zip Code	Mexican Review Count	Mexican Average Rating	Mexican Weighted Rating
621	29468	199	3.8593	768

	Zip Code	Mexican Review Count	Mexican Average Rating	Mexican Weighted Rating
226	25414	206	4.15534	856
233	93545	213	3.64554	776.5
67	60565	2836	3.94059	11175.5
235	3064	544	3.72518	2026.5

Calculate Summaries

[331]
```
mexican_data["Mexican Review Count"].sum()
```

```
469100
```

[332]
```
italian_data["Italian Review Count"].sum()
```

```
561872
```

[333]
```
mexican_data["Mexican Weighted Rating"].sum() / mexican_data["Mexican Rev
```

```
3.905826049882754
```

[334]
```
italian_data["Italian Weighted Rating"].sum() / italian_data["Italian Rev
```

```
3.9436802332203778
```

[341]
```
# Combine Data Frames into a single Data Frame
combined_data = pd.merge(mexican_data, italian_data, on="Zip Code")
combined_data.head()
```

	Zip Code	Mexican Review Count	Mexican Average Rating	Mexican Weighted Rating	Italian Review Count	Italian Average Rating	Italian Weight Rating
0	29468	199	3.8593	768	77	3.35065	258
1	25414	206	4.15534	856	135	3.41111	460.5
2	93545	213	3.64554	776.5	702	4.10684	2883
3	60565	2836	3.94059	11175.5	2829	3.92807	11112.5

	Zip Code	Mexican Review Count	Mexican Average Rating	Mexican Weighted Rating	Italian Review Count	Italian Average Rating	Italian Weight Rating
4	3064	544	3.72518	2026.5	1253	3.77813	4734

[336]
```
# Total Rating and Popularity "Wins"
combined_data["Rating Wins"] = np.where(combined_data["Mexican Average Ra
combined_data["Review Count Wins"] = np.where(combined_data["Mexican Revi
```

[337]
```
# View Combined Data
combined_data.head()
```

	Zip Code	Mexican Review Count	Mexican Average Rating	Mexican Weighted Rating	Italian Review Count	Italian Average Rating	Italian Weight Rating
0	29468	199	3.8593	768	77	3.35065	258
1	25414	206	4.15534	856	135	3.41111	460.5
2	93545	213	3.64554	776.5	702	4.10684	2883
3	60565	2836	3.94059	11175.5	2829	3.92807	11112.5
4	3064	544	3.72518	2026.5	1253	3.77813	4734

[338]
```
# Tally number of cities where one type wins on ratings over the other
combined_data["Rating Wins"].value_counts()
```

```
Mexican    267
Italian    248
Name: Rating Wins, dtype: int64
```

[339]
```
# Tally number of cities where one type wins on review counts over the ot
combined_data["Review Count Wins"].value_counts()
```

```
Italian    298
Mexican    217
Name: Review Count Wins, dtype: int64
```

Display Summary of Results

[340]
```
# Model 1: Head-to-Head Review Counts
italian_summary = pd.DataFrame({"Review Counts": italian_data["Italian Re
                                "Rating Average": italian_data["Italian A
```

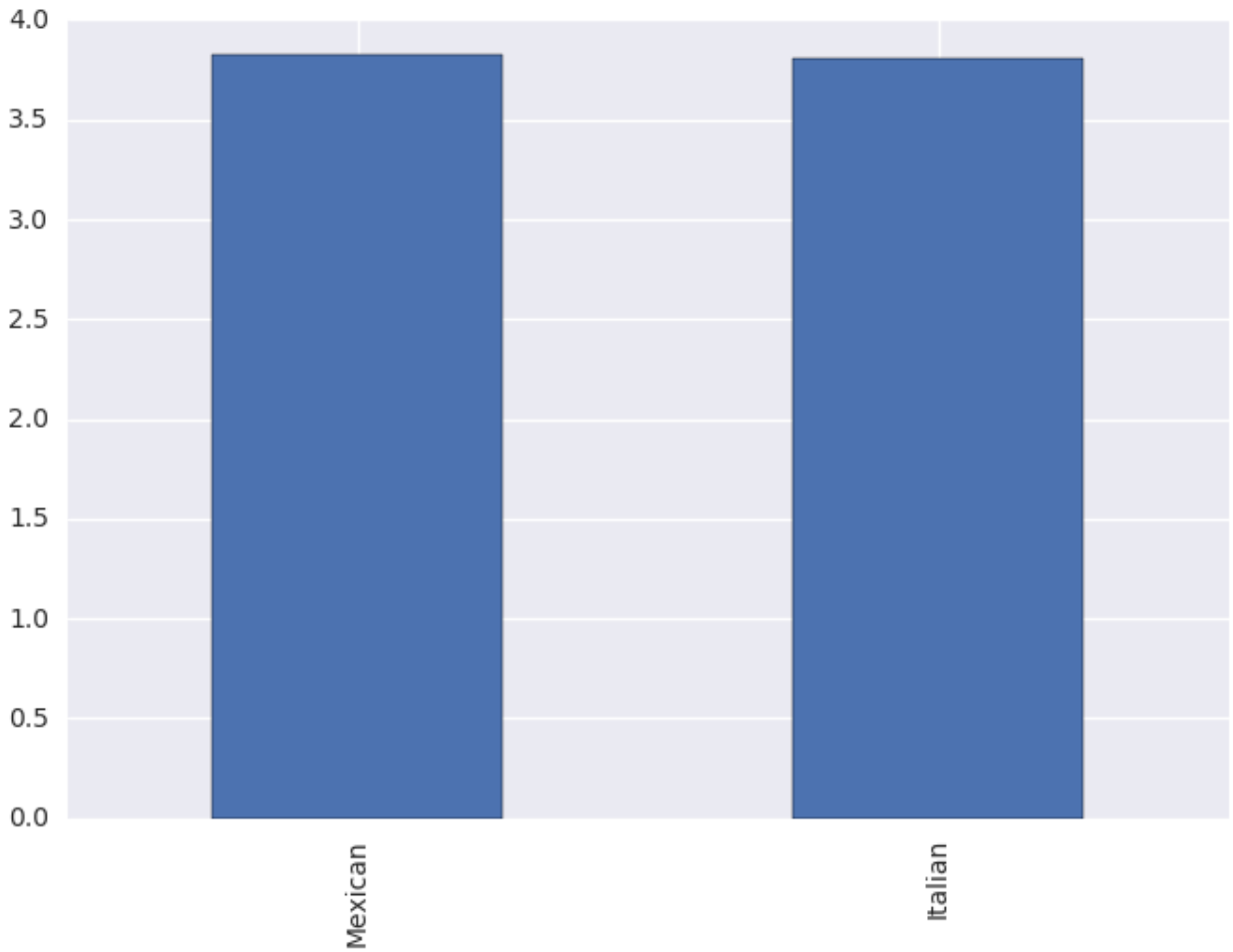
```
                                        "Review Count Wins": combined_data["Revie
                                        "Rating Wins": combined_data["Rating Wins

mexican_summary = pd.DataFrame({"Review Counts": mexican_data["Mexican Re
                                        "Rating Average": mexican_data["Mexican A
                                        "Review Count Wins": combined_data["Revie
                                        "Rating Wins": combined_data["Rating Wins

final_summary = pd.concat([mexican_summary, italian_summary])
final_summary
```

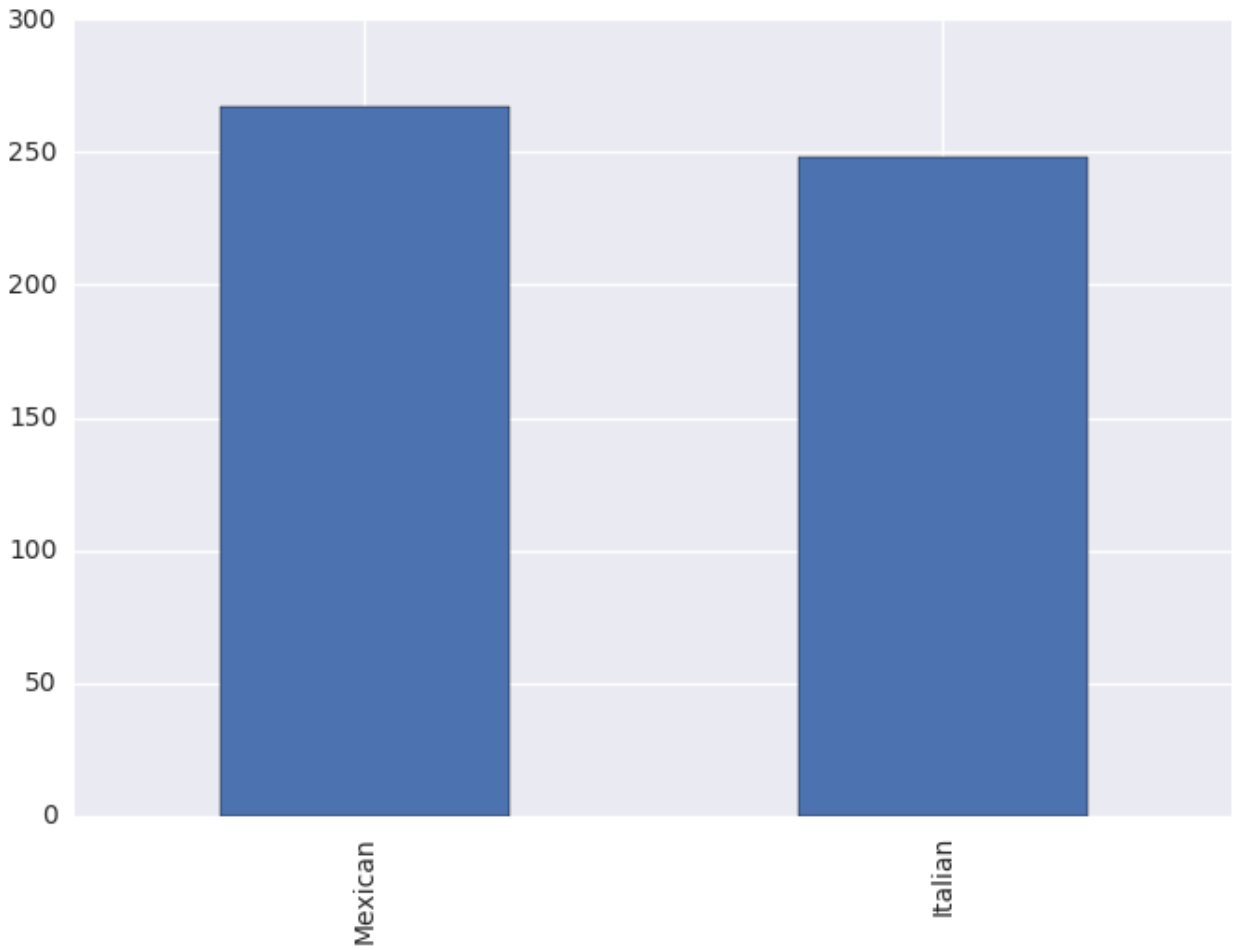
	Rating Average	Rating Wins	Review Count Wins	Review Counts
Mexican	3.825374	267	217	469100
Italian	3.807411	248	298	561872

[533]
```
# Plot Rating Average
plt.clf()
final_summary["Rating Average"].plot.bar()
plt.show()
```

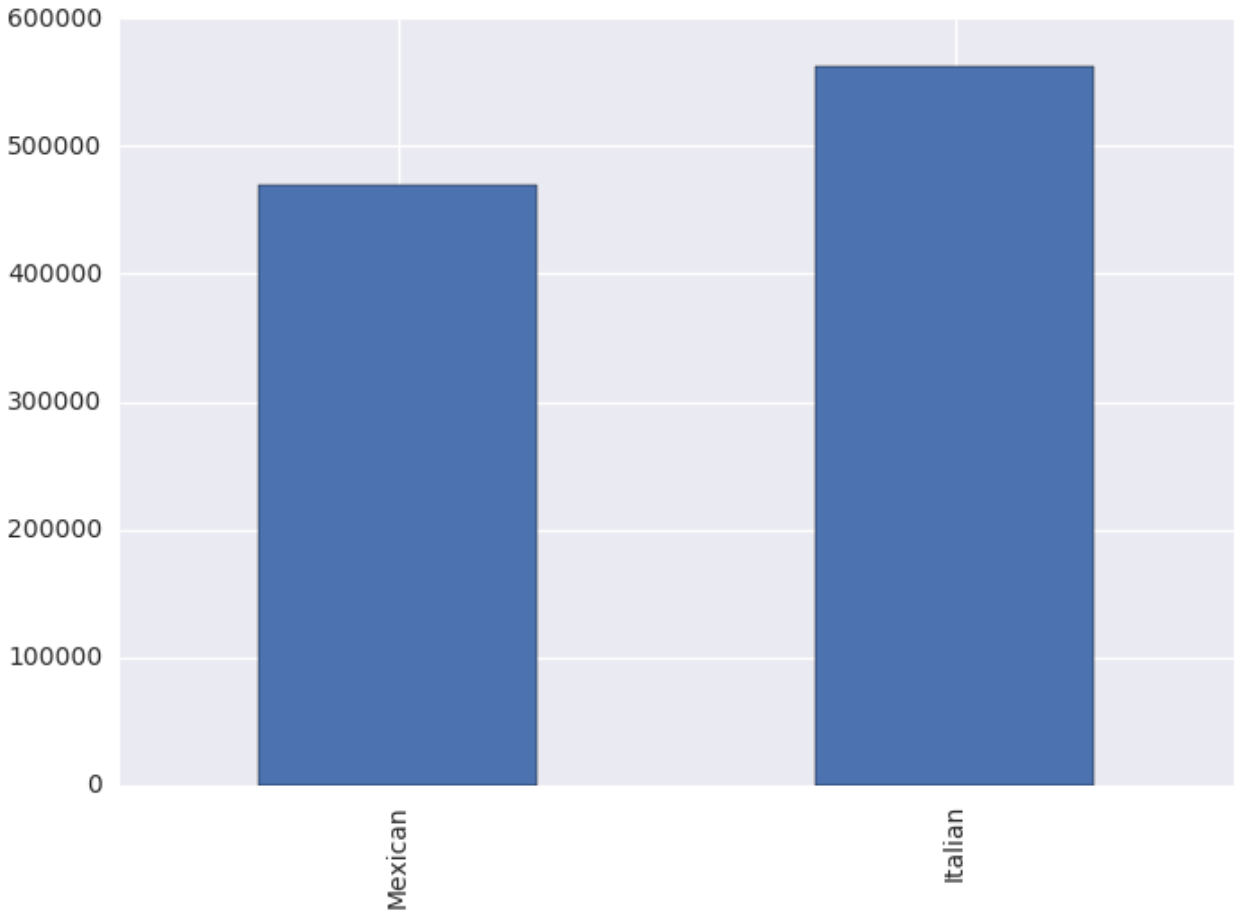


[534]
```
# Plot Rating Wins
plt.clf()
final_summary["Rating Wins"].plot.bar()
```

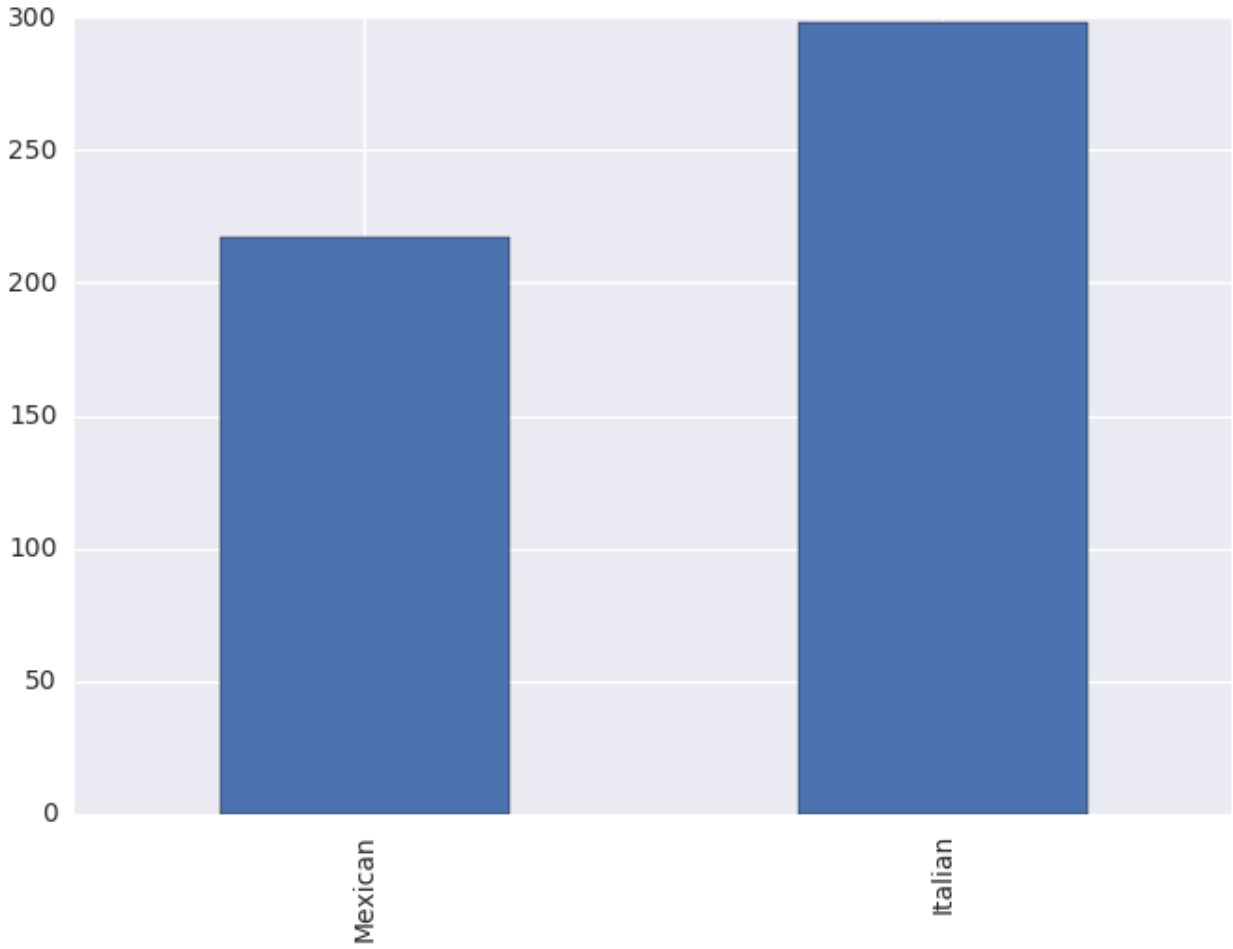
```
plt.show()
```



[535]
```
# Plot Review Count
plt.clf()
final_summary["Review Counts"].plot.bar()
plt.show()
```



[537]
```
# Plot Review Count
plt.clf()
final_summary["Review Count Wins"].plot.bar()
plt.show()
```

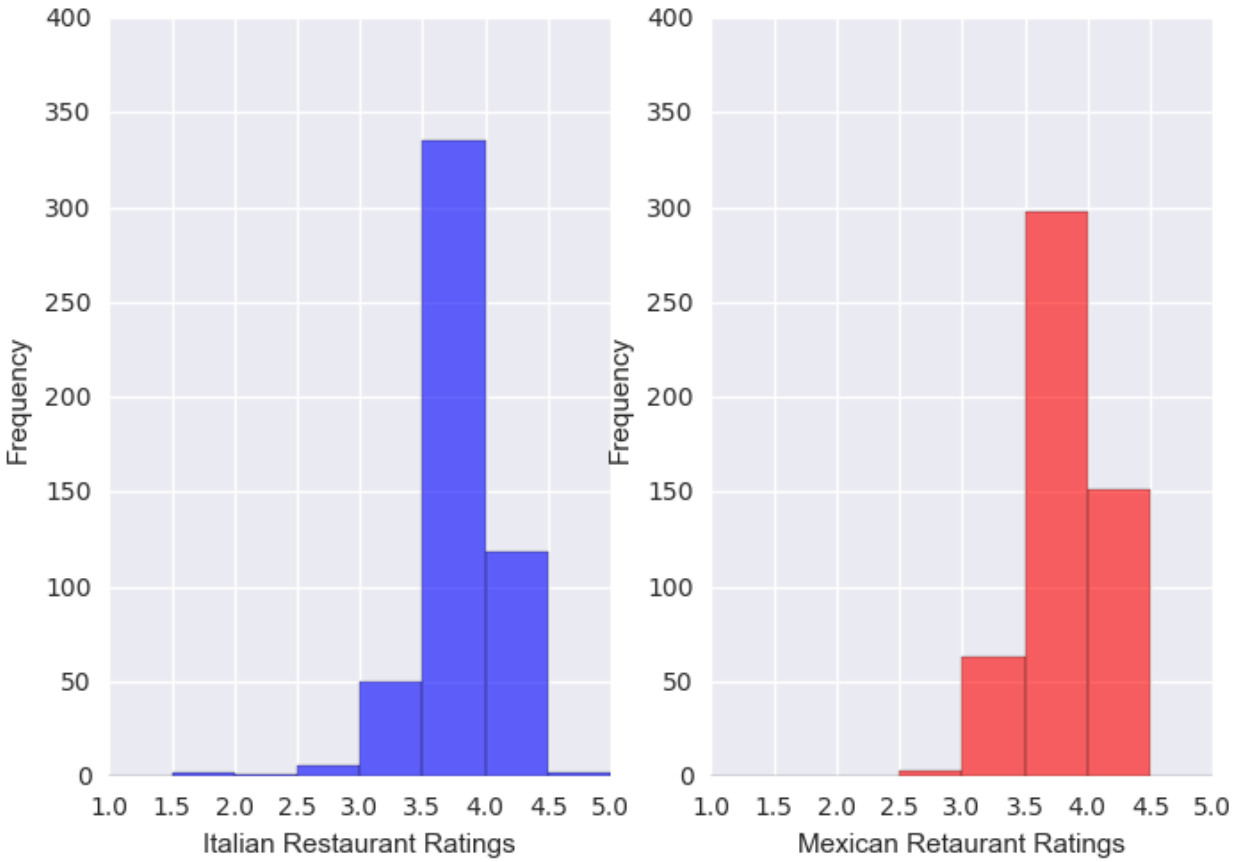


[542]
```
# Histogram Italian Food (Ratings)
plt.figure()

# Subplot 1 (Italian)
plt.subplot(121)
combined_data["Italian Average Rating"].plot.hist(bins=[0, 0.5, 1, 1.5, 2
plt.xlabel("Italian Restaurant Ratings")
plt.xlim([1, 5.0])
plt.ylim([0, 400])

# Subplot 2 (Mexican)
plt.subplot(122)
combined_data["Mexican Average Rating"].plot.hist(bins=[0, 0.5, 1, 1.5, 2
plt.xlabel("Mexican Retaurant Ratings")
plt.xlim([1, 5.0])
plt.ylim([0, 400])

# Show Plot
plt.show()
```



Statistical Analysis

[557]
```
# Run a t-test on average rating and number of reviewers
mexican_ratings = combined_data["Mexican Average Rating"]
italian_ratings = combined_data["Italian Average Rating"]

mexican_review_counts = combined_data["Mexican Review Count"]
italian_review_counts = combined_data["Italian Review Count"]
```

[558]
```
# Run T-Test on Ratings
ttest_ind(mexican_ratings.values, italian_ratings.values)
```

```
Ttest_indResult(statistic=0.97343025910497216, pvalue=0.33056849735046245)
```

[559]
```
# Run T-Test on Review Counts
ttest_ind(mexican_review_counts.values, italian_review_counts.values)
```

```
Ttest_indResult(statistic=-1.8606196373601354,
pvalue=0.063083275022233751)
```

Conclusions

Based on our analysis, it is clear that American preference for Italian and Mexican food are very similar in nature. As a whole, Americans rate Mexican and Italian restaurants at statistically similar scores. However, there does exist evidence that Americans do write more reviews on Italian restaurants. This may indicate that there is an increased interest in visiting Italian restaurants at an experiential level. (However, this data may also merely suggest that Yelp users happen to enjoy writing reviews on Italian restaurants more).

```
[ ]
```