Capstone 1 : Investing with Univision TV

Dævyd Hjelmstad --- Thinkful Data Science Intensive Cohort #8

Introduction

A sales representative named Lisa from Univision TV has requested an analysis of a potential client for advertising with Univision. The client is Chaz's, a new restaurant chain in Univision City that serves Mexican and Italian entrees (we will refer to these types of restaurants as "Mexitalian" restaurants later on in our study)

Lisa has requested an analysis & presentation which illustrates the strength of the Hispanic population in Univision City as consumers at sit-down restaurants. She would also like this presentation to include an analysis of how profitable an investment in Univision TV could be for Chaz's.

Hypotheses

1

- H_{1_0} : There is no difference between the Hispanic vs. Non-Hispanic populations as consumers at sit-down restaurants in Univision City.
- H_{1_a} : The Hispanic population is significantly stronger group of consumers at sit-down restaurants in Univision City.

2

- H_{2_0} : An investment in advertising on Univision TV produces no net profit for the client.
- H_{2_a} : An investment in advertising on Univision TV will produce a net profit for the client.

Data

The data set was provided by Lisa via email, so it is safe to assume this should be sufficient for the scope of this project. The excel dataset was cleaned up to make it easier to read into python & is available at

https://drive.google.com/file/d/1WG0Ug8SUzSjjSOYemZLrI0_9-SmifUOf/view?usp=sharing (https://drive.google.com/file/d/1WG0Ug8SUzSjjSOYemZLrI0_9-SmifUOf/view?usp=sharing)

In [1]:
```
# setup
%reload_ext nb_black

import math
import pandas as pd
import numpy as np
from scipy import stats
import seaborn as sns
import matplotlib.pyplot as plt


data_file = "Univision City Data for Capstone I.xlsx"
xlsx = pd.ExcelFile(data_file)
xlsx_sheets = xlsx.sheet_names

df_restaurant_info = pd.read_excel(xlsx, xlsx_sheets[0])
df_restaurant_advertising_spending = pd.read_excel(xlsx, xlsx_sheets[8])
```

Method

We will be using the data frame called 'df_restaurant_info' to explore a significant variation in the means of the '%' & 'Index to Total Market' columns of Hispanic vs. Non-Hispanic populations. Lisa gave no info about how the 'Index to Total Market' was calculated, but it is safe to assume it is a measurement of popularity within the demographic. After some analysis, it appears the 'Index to Total Market' in this case is in fact defined as follows:

$$p_1 = \frac{\text{\# of Hispanics who eat at sit down restaurants}}{\text{total people who eat at sit down restaurants}}$$

$$p_2 = \frac{\text{\# of Hispanics who eat at Restaurant X}}{\text{total people who eat at Restaurant X}}$$

$$\text{Index to Total Market of Restaurant X} = \frac{p_1}{p_2}$$

Hypothesis 1:

In [2]:
```
df_restaurant_info.head()
```

Out[2]:

	Sit Down Restaurants used in the Past 6 Months (Among Adults Age 18+)	# of Hispanics	% of Hispanics	Hispanic Index to Total Market	# of Non-Hispanics	% of Non-Hispanics	Non-Hispanic Index to Total Market
0	Any sit-down restaurant	828786	0.787	98	2635867	0.801	101
1	Applebee's	117084	0.113	130	252718	0.077	90
2	Buca di Beppo	21706	0.023	141	41627	0.013	87
3	California Pizza Kitchen	13581	0.015	62	76284	0.023	112
4	Carrabba's Italian Grill	34180	0.034	135	70243	0.021	89

It looks like the first element contains information for 'any sit-down restaurant', so it is an aggregated number. This will be useful later in our analysis, but for now, we drop this data element & procede with the current tests.

In [3]:
```
# defining the columns of df_restaurant_info as a list,
#since some of the names are long

restaurant_columns = df_restaurant_info.columns
for i, item in enumerate(restaurant_columns, start=0):
    print(i,"-",item)
```

```
0 - Sit Down Restaurants used in the Past 6 Months (Among Adults Age 18
+)
1 - # of Hispanics
2 - % of Hispanics
3 - Hispanic Index to Total Market
4 - # of Non-Hispanics
5 - % of Non-Hispanics
6 - Non-Hispanic Index to Total Market
```

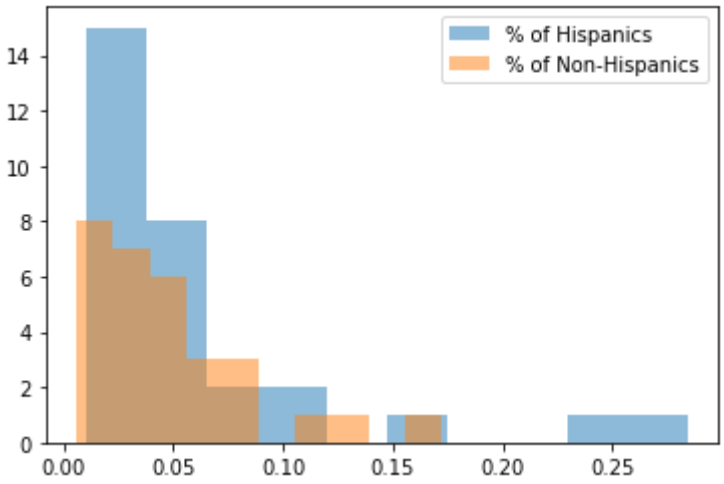
In [4]:
```
# drop the first element which contains aggregated data
df_restaurant_info = df_restaurant_info[
    df_restaurant_info[restaurant_columns[0]] != "Any sit-down restauran
t"
]
```

First, let's plot histograms of the distributions of '%' and 'Index' columns

In [5]:
```
# plot histogram & print stats for '% of Hispanics' & '% of Non-Hispanics' columns

plt.hist(df_restaurant_info[restaurant_columns[2]], alpha=0.5, label="% of Hispanics")
plt.hist(
    df_restaurant_info[restaurant_columns[5]], alpha=0.5, label="% of Non-Hispanics"
)
print(stats.describe(df_restaurant_info[restaurant_columns[2]]))
print(stats.describe(df_restaurant_info[restaurant_columns[5]]))
plt.legend()
plt.show()
```

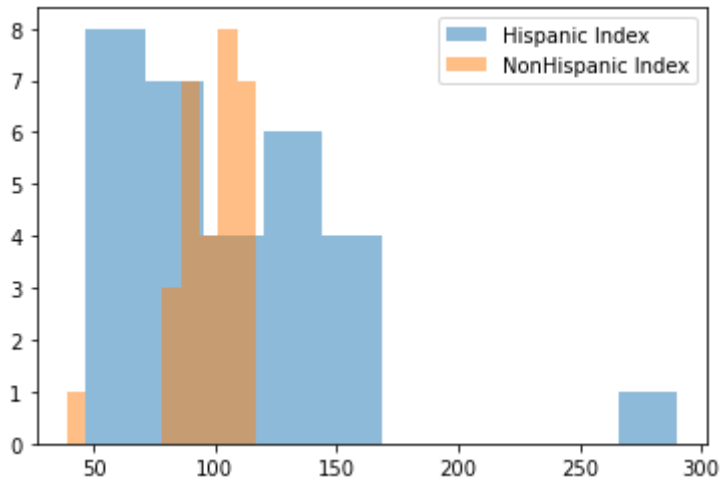
```
DescribeResult(nobs=30, minmax=(0.01, 0.284), mean=0.05763333333333332, variance=0.004387826436781607 5, skewness=2.2315847013744423, kurtosis=4.484604992004228)
DescribeResult(nobs=30, minmax=(0.006, 0.172), mean=0.04746666666666666 4, variance=0.0014879816091954024, skewness=1.5083238312376936, kurtosis=2.2029355085746767)
```



In [6]:
```
# plot histogram & print stats for
# 'Hispanic index to total market' &
#'Non-Hispanic index to total market' columns

plt.hist(df_restaurant_info[restaurant_columns[3]], alpha=0.5, label="Hi
spanic Index")
plt.hist(
    df_restaurant_info[restaurant_columns[6]], alpha=0.5, label="NonHisp
anic Index"
)
print(stats.describe(df_restaurant_info[restaurant_columns[3]]))
print(stats.describe(df_restaurant_info[restaurant_columns[6]]))
plt.legend()
plt.show()
```

```
DescribeResult(nobs=30, minmax=(47, 290), mean=107.03333333333333, vari
ance=2390.2402298850566, skewness=1.669826517913023, kurtosis=4.4459552
062788505)
DescribeResult(nobs=30, minmax=(39, 117), mean=97.73333333333333, varia
nce=245.3747126436782, skewness=-1.6756496629909805, kurtosis=4.4990987
48229652)
```



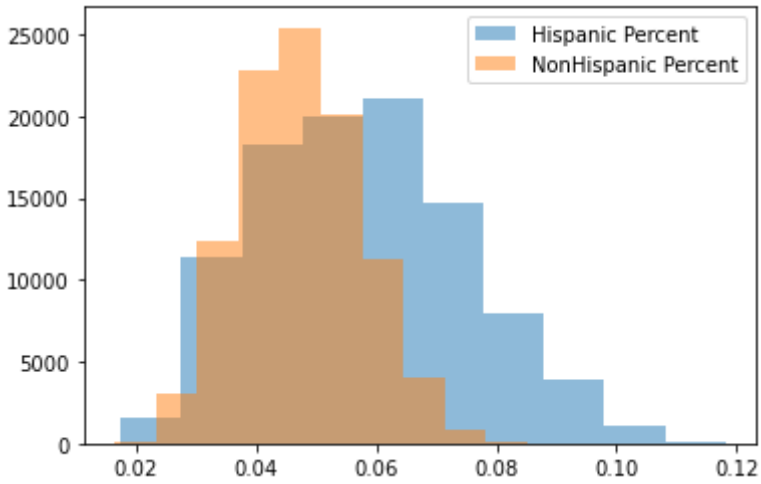
So, it looks like the Hispanic population has both a higher mean % and a higher mean Index to total market across all restaurants that we have data for! This is a good indication, but we would like to run a t-test to ensure that this is a statistically significant difference.

However, the data is not very normal so we will not be able to use a t-test directly to determine if this is a significant difference or not. I would like to use a bootstrapping re-sampling technique to take the means of 10 restaurants at a time & get a normal data distribution to use to find confidence intervals.

In [7]:
```
sample_means_Hispanic_Percent = []
sample_means_Hispanic_Index = []
sample_means_NonHispanic_Percent = []
sample_means_NonHispanic_Index = []
for i in range(100000):
    sample_means_Hispanic_Percent.append(
        df_restaurant_info[restaurant_columns[2]].sample(10).mean()
    )
    sample_means_Hispanic_Index.append(
        df_restaurant_info[restaurant_columns[3]].sample(10).mean()
    )
    sample_means_NonHispanic_Percent.append(
        df_restaurant_info[restaurant_columns[5]].sample(10).mean()
    )
    sample_means_NonHispanic_Index.append(
        df_restaurant_info[restaurant_columns[6]].sample(10).mean()
    )
```

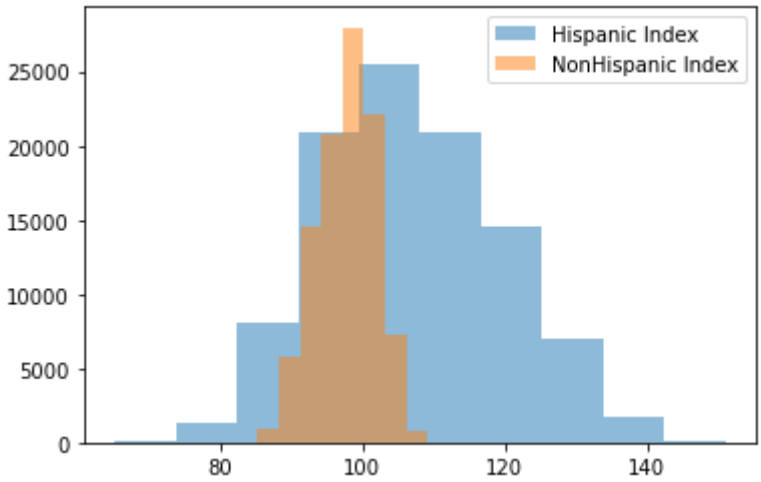
In [8]:
```
plt.hist(sample_means_Hispanic_Percent, alpha=0.5, label="Hispanic Percent")
plt.hist(sample_means_NonHispanic_Percent, alpha=0.5, label="NonHispanic Percent")
print(stats.describe(sample_means_Hispanic_Percent))
print(stats.describe(sample_means_NonHispanic_Percent))
plt.legend()
plt.show()
```

```
DescribeResult(nobs=100000, minmax=(0.017500000000000005, 0.11810000000000001), mean=0.057731248000000006, variance=0.0002936461972244683, skewness=0.30659909793496215, kurtosis=-0.42836970414579945)
DescribeResult(nobs=100000, minmax=(0.0165, 0.0851), mean=0.047433231, variance=9.914740327467176e-05, skewness=0.20704492945065475, kurtosis=-0.31851058661067855)
```



In [9]:
```
plt.hist(sample_means_Hispanic_Index, alpha=0.5, label="Hispanic Index")
plt.hist(sample_means_NonHispanic_Index, alpha=0.5, label="NonHispanic I
ndex")
print(stats.describe(sample_means_Hispanic_Index))
print(stats.describe(sample_means_NonHispanic_Index))
plt.legend()
plt.show()
```

```
DescribeResult(nobs=100000, minmax=(65.3, 150.9), mean=107.000147, vari
ance=158.07731785156952, skewness=0.2305926488035491, kurtosis=-0.43486
058184694265)
DescribeResult(nobs=100000, minmax=(82.3, 112.1), mean=97.7293199999999
9, variance=16.344091178511786, skewness=-0.23844932681553593, kurtosis
=-0.43325992933787427)
```



these data sets are sufficiently normal for us to calculate the t-test and confidence intervals:

In [10]:
```
print(stats.ttest_ind(sample_means_Hispanic_Percent, sample_means_NonHis
panic_Percent))
print(stats.ttest_ind(sample_means_Hispanic_Index, sample_means_NonHispa
nic_Index))
```

```
Ttest_indResult(statistic=164.31280237700307, pvalue=0.0)
Ttest_indResult(statistic=221.98241999541503, pvalue=0.0)
```

from this we can certainly reject the null in both cases ! Let's see what confidence intervals we have:

In [11]:
```
def get_95_ci(array_1, array_2):
    sample_1_n = len(array_1)
    sample_2_n = len(array_2)
    sample_1_mean = np.mean(array_1)
    sample_2_mean = np.mean(array_2)
    sample_1_var = np.var(array_1)
    sample_2_var = np.var(array_2)
    mean_difference = sample_2_mean - sample_1_mean
    std_err_difference = math.sqrt(
        (sample_1_var / sample_1_n) + (sample_2_var / sample_2_n)
    )
    margin_of_error = 1.96 * std_err_difference
    ci_lower = mean_difference - margin_of_error
    ci_upper = mean_difference + margin_of_error
    return (
        "The difference in means at the 95% confidence interval (two-tai
l) is between "
        + str(ci_lower.round(4))
        + " and "
        + str(ci_upper.round(4))
        + "."
    )


print("Percentage confidence interval: ")
print(get_95_ci(sample_means_Hispanic_Percent, sample_means_NonHispanic_
Percent))
print("Index to Total Market confidence interval: ")
print(get_95_ci(sample_means_Hispanic_Index, sample_means_NonHispanic_In
dex))
```

```
Percentage confidence interval:
The difference in means at the 95% confidence interval (two-tail) is be
tween -0.0104 and -0.0102.
Index to Total Market confidence interval:
The difference in means at the 95% confidence interval (two-tail) is be
tween -9.3527 and -9.189.
```

Result

The data was not normally distributed, so a re-sampling technique was utilized to move forward with the t-test. Given the resulting p-value of zero, the null hypothesis may be rejected and thus there is some difference in the turn-out of Hispanics vs. Non-Hispanics at sit-down restaurants in Univision City.

The 95% confidence interval suggests that the difference in percentages of people who are Hispanic vs. Non-Hispanic will be between 1.02% and 1.04%, and the Total Index to Market is higher in the Hispanic Population over Non-Hispanic population by 9.189 to 9.3527 Index to Total Market points.

These statistics imply that restaurants in general in Univision City are more likely to be visited by customers from the Hispanic population, as opposed to the Non-Hispanic population.

Hypothesis 2:

Method

For this hypothesis, we will use the df_restaurant_spending dataframe in conjunction with the df_restaurant_info dataframe to draw conclusions about the effectiveness of advertising with Univision City. We will filter out restaurants which do not have Mexican or Italian entrees on their menu, so that our analysis will more accurately predict the expectations for Chaz's:

In [12]:
```
mexitalian_restaurants = [
    "Buca di Beppo",
    "California Pizza Kitchen",
    "Mellow Mushroom",
    "Carrabba's Italian Grill",
    "Olive Garden",
    "Oregano's Pizza Bistro",
    "Postino",
    "Sweet Tomatoes",
    "Macayo's",
]

df_mexitalian_restaurant_info = df_restaurant_info[
    df_restaurant_info[restaurant_columns[0]].isin(mexitalian_restaurant
s)
]

df_mexitalian_restaurant_advertising_spending = df_restaurant_advertisin
g_spending[
    df_restaurant_advertising_spending['Restaurant'].isin(mexitalian_res
taurants)
]
```

In [13]:
```
df_mexitalian_restaurant_info.head(10)
```

Out[13]:

	Sit Down Restaurants used in the Past 6 Months (Among Adults Age 18+)	# of Hispanics	% of Hispanics	Hispanic Index to Total Market	# of Non-Hispanics	% of Non-Hispanics	Non-Hispanic Index to Total Market
2	Buca di Beppo	21706	0.023	141	41627	0.013	87
3	California Pizza Kitchen	13581	0.015	62	76284	0.023	112
4	Carrabba's Italian Grill	34180	0.034	135	70243	0.021	89
16	Macayo's	25509	0.024	95	85221	0.026	102
17	Mellow Mushroom	11394	0.011	146	20769	0.006	85
19	Olive Garden	297583	0.284	142	566018	0.172	87
20	Oregano's Pizza Bistro	65465	0.062	112	175503	0.053	96
24	Postino	10075	0.010	47	78871	0.024	117
28	Sweet Tomatoes	27421	0.026	59	162795	0.049	113

In [14]:
```
df_mexitalian_restaurant_advertising_spending.head(10)
```

Out[14]:

	Restaurant	Amount Spent on Advertising
0	Buca di Beppo	312723
1	Carrabba's Italian Grill	129783
5	Olive Garden	103679
6	Oregano's Pizza Bistro	94048

calculate the mean percentage of Hispanics who go to a restaurant that serves Mexican and/or Italian entrees:

In [26]:
```
df_mexitalian_restaurant_info[restaurant_columns[2]].mean()
```

Out[26]: 0.05433333333333333

We will make an estimate of the cost of an average household in Univision City, based on the given data, and an estimate for the average cost of a taco dish at Mexican restaurants in the US, according to modernrestaurantmanagement.com, which is $11.28. From the data given, 48\% of Hispanic households have one or more children, and we will estimate that a child eats half as much as an adult, so the average entree cost for any client will be calculated as mean(11.28,5.64,5.64), which is \$7.52, a reasonable estimate.

Household Information	# of Hispanic HHLDs	% of Hispanic HHLDs	Hispanic Index to Total Market	# of Non-Hispanic HHLDs	% of Non-Hispanic HHLDs	Non-Hispanic Index to Total Market
One or more Children under 18 living in the HHLD	217,770	0.48	165	398,433	0.24	82

From the above analysis, we can remember that the mean % of Hispanics who ate at any given Mexitalian restaurant in Univision city was 5.43%. Therefore, we can estimate the total number of entrees purchased by the Hispanic population, and thus the total revenue from the Hispanic population by multiplying this with the average entree cost above.

Lisa has provided data about the average household size in the Hispanic population, as well as an indication of how many times Hispanics go out to eat in general:

frequency of dining in at restaurants

Number of Times used a Sit-Down Restaurant in Past 6 Months (Among Adults Age 18+)	# of Hispanics	% of Hispanics	Index to Total Market	# of Non-Hispanics	% of Non-Hispanics	Index to Total Market
5	192,503	0.18	129	421,654	0.13	91
10	164,868	0.16	108	464,376	0.14	97

average household size

Hispanic HHLDs	Non-Hispanic HHLDs
3.33	2.37

clientele at dine-in restaurants population size

Sit Down Restaurants used in the Past 6 Months (Among Adults Age 18+)	# of Hispanics	% of Hispanics	Hispanic Index to Total Market	# of Non-Hispanics	% of Non-Hispanics	Non-Hispanic Index to Total Market
Any sit-down restaurant	828,786	0.79	98	2,635,867	0.80	101

So, we can make an estimate of the number of entrees sold at a dine-in restaurant in Univision City in 2019 as follows: (192,503*(5-1))+(164,868*(10-6))+ 828,786 = 2,258,270 entrees

In [16]:
```
192503 * 4 + 164868 * 4 + 828786
```

Out[16]:
```
2258270
```

We then get the average entrees ordered at any given restaurant by taking the total entrees & multiplying by the mean percentage of Hispanics who ate at a given single restaurant in Univision City:

In [17]:
```
2258270 * 0.0543
```

Out[17]:
```
122624.061
```

So, the average restaurant in Univision City in 2019 generated 122,624 * $7.52 = \$922,132.48 from the Hispanic population

In [18]:
```
122624 * 7.52
```

Out[18]:
```
922132.48
```

Now, we compare this with the average amount spent on advertising by a restaurant in Univision City using the df_univision_city_restaurant_spending dataframe

In [19]:
```
df_mexitalian_restaurant_advertising_spending.head(10)
```

Out[19]:

	Restaurant	Amount Spent on Advertising
0	Buca di Beppo	312723
1	Carrabba's Italian Grill	129783
5	Olive Garden	103679
6	Oregano's Pizza Bistro	94048

In [20]:
```
df_mexitalian_restaurant_advertising_spending["Amount Spent on Advertising "].mean()
```

Out[20]:
```
160058.25
```

In [21]:
```
1258735.2 / 160058.25
```

Out[21]:
```
7.864231928063689
```

So, our analysis reveals that the average Mexitalian restaurant in Univision City could expect to bring in $1,258,735.20 in revenue from the Hispanic population, while only spending on average \$160,058.25 on advertising with Univision TV, apparently reaching plenty of potential hungry customers. This represents an 786% average return on investment for a Mexitalian Restaurant, in this sense.

Lastly, we present one specific case example of a specific restaurant who advertises with Univision TV and experiences a huge turnout from the Hispanic population of restaurant-goers: Olive Garden

In [22]:
```
df_restaurant_info[df_restaurant_info[restaurant_columns[0]] == "Olive G
arden"]
```

Out[22]:

	Sit Down Restaurants used in the Past 6 Months (Among Adults Age 18+)	# of Hispanics	% of Hispanics	Hispanic Index to Total Market	# of Non-Hispanics	% of Non-Hispanics	Non-Hispanic Index to Total Market
19	Olive Garden	297583	0.284	142	566018	0.172	87

In [23]:
```
round(297583 * 7.52, 4)
```

Out[23]:
```
2237824.16
```

In [24]:
```
df_restaurant_advertising_spending[
    df_restaurant_advertising_spending["Restaurant"] == "Olive Garden"
]
```

Out[24]:

	Restaurant	Amount Spent on Advertising
5	Olive Garden	103679

As we can see, the Olive Garden in Univision City spent $103,679 on advertising with Univision TV, and had 297,583 Hispanic people dine in at their restaurant, resulting in an estimated $2,243,775 in revenue, a 2,164% return on investment, in this sense.

In [25]:
```
2243775.82 / 103679
```

Out[25]:
```
21.641565022810788
```

These favorable figures clearly reject the null hypothesis that investing with Univision TV will not lead to an increased revenue in the establishment, via clear counterexamples from similar restaurants in Univision City in 2019.

Discussion and recommendation

The data has shown that both the percentage of, as well as the Index to Total Market of, the Hispanic population who dines in at sit-down restaurants in Univision city is significanty higher than the population of their Non-Hispanic counterparts in the city. It follows that advertisements with Univision TV, whose audience is overwhelmingly people of the Hispanic population, are reaching the densest pocket of clientelle at sit-down restaurants in the city, and thus will be well-viewed.

It has also shown that the average return on investment for a 'Mexitalian' restaurant who advertised with Unvision City in 2019 was 786%, and that one outstanding establishment, Olive Garden, experienced a return on investment of 2,164% from an investment in the Hispanic population via Univision City.

It is our recommendation that an investment in advertising with Univision TV will greatly boost the revenue for the new Chaz's establishments that open up in the city this year.