# Relationship between Preference on Ice Cream and Tempearture

**Description:** The dataset used in analyzing for this project is named “result.csv”. The result dataset is created by usinig data from three different datasets. One of the three data sourse is a file named “data.json”. The other two dataset are extracted by using api. All datasourse and descriptions are provided in the subfolder “data”. By analyzing the mentioned datasets, I have analyzed the relationship between the willingness of people in different cities to have ice cream and the tempearture of those cities.

**Motivation:** It’s an obvious fact that people like having ice cream in hot summer days to help them stay cool. However, whether people will hold their love towards ice cream still in cold days. On the one hand, it seems normal that people are less willing to have something cold like ice cream in winter because the ice cream will make them feel colder. On the other hand, ice cream is one of the most popular dessert in the United States. It is also normal for people to enjoy their beloved dessert regardless the temperarture. It happens to be winter in the U.S. right now. While Los Angeles, the city that I live still has pleasant weather, there are many other cities that have already started snow. I was motivated to find out the relationship between the preference on ice cream and the tempearture by using recent data in different cities. To obtain the relationship answer, I have performed statistical analysis on the dataset that I collected.

**Running the Code:** The requirement.txt is uploaded on github repository. There is only one script in the form of Jupyter notebook. The result can be re-produced by running the Jupyter notebook directly. No other action required. The link to the github repository is https://github.com/Aster-Yu/510-Project.

**Data Sourse:**

- The first data sourse is a json file named “data.json” which is downloaded from the website https://worldpopulationreview.com/us-cities. The nature of it is population. It contains the city information of the 200 cities that have most population in the U.S.. This dataset is used to obtain the list of cities that have top 100 large population in order to obtain more accurate result (avoid situations such as certain small cities may share common food preference such as dislike in ice cream). Only the name column is used in this dataset. Following is the sample data.


```
[
  {
    "rank": 1,
    "name": "New York City",
    "usps": "NY",
    "pop2022": 8930002,
    "pop2020": 8804190,
    "growth": "0.00709",
    "density": 29729,
    "aland_sqmi": "300.3810"
  },
  {
    "rank": 2,
    "name": "Los Angeles",
    "usps": "CA",
```

- The second data sourse is from the link https://rapidapi.com/weatherapi/api/weatherapi-com/. This is an API link. The nature of this data sourse is weather. This data sourse is used to get weather information in each city. By iterating the city name list obtained from “data.json”, the recent tempearture data can be extracted from this API link using the search key of city name. I have averaged the recent tempearture in each city and store them in “result.csv”. I have draw a historam to show the temperature spread of large population cities in the U.S.. The following is the sample data of New York weather in a certain day in json format by using json fomatter.


```
{
  "location":{
    "name":"New York",
    "region":"New York",
    "country":"United States of America",
    "lat":40.71,
    "lon":-74.01,
    "tz_id":"America/New_York",
    "localtime_epoch":1671002995,
    "localtime":"2022-12-14 2:29"
  },
  "forecast":{
    "forecastday":[
      {
        "date":"2022-12-07",
        "date_epoch":1670371200,
        "day":{
          "maxtemp_c":12.5,
          "maxtemp_f":54.5,
          "mintemp_c":11.5,
          "mintemp_f":52.7,
          "avgtemp_c":12.0,
          "avgtemp_f":53.7,
          "maxwind_mph":11.6,
          "maxwind_kph":18.7,
          "totalprecip_mm":6.6,
          "totalprecip_in":0.26,
          "avgvis_km":3.9,
          "avgvis_miles":2.0,
          "avghumidity":98.0,
          "condition":{
            "text":"Light rain",
            "icon":"//cdn.weatherapi.com/weather/64x64/day/296.png",
            "code":1183
          },
          "uv":0.0
```

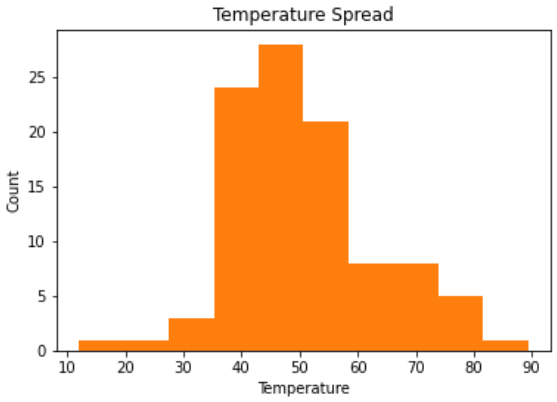

- The third data sourse is from the link https://www.yelp.com/developers/documentation/v3/business_search. This is an yelp API link. The nature of this data sourse is restrunant information. This data sourse is used to get information about ice cream shops in each city. By iterating the city name list obtained from “data.json”, the most popular restrunants information can be extracted from this API link using the search key of city name. I have sorted the results by using the top 20 most reviewed restrunants (small shops included). Not everyone is willing to write a review for a restrunants. In this case, restrunants with most reviews can be viewed as popular in their city. I have counted the number of restrunants that have “icecream” in their tags. The following is the sample data of New York most reviewed restrunant information in json format by using json fomatter.

```
{
  "businesses":[
    {
      "id":"H4jJ7XB3CetIr1pg56CczQ",
      "alias":"levain-bakery-new-york",
      "name":"Levain Bakery",
      "image_url":"https://s3-media3.fl.yelpcdn.com/bphoto/DH29qeTmPotJbCSzkjYJwg/o.jpg",
      "is_closed":false,
      "url":"https://www.yelp.com/biz/levain-bakery-new-york?adjust_creative=fAdd2NO5XS4zjg
      "review_count":9343,
      "categories":[
        {
          "alias":"bakeries",
          "title":"Bakeries"
        }
      ],
      "rating":4.5,
      "coordinates":{
        "latitude":40.779961,
        "longitude":-73.980299
      },
      "transactions":[
```


**Changes and Challenges on Data Collecting:** In my orginal plan, I was planning to analyze the relationship between temperature and food preference in each city. However, this has two problems. Firstly, requires far more data to analyze relationships among temperature and such amount of different food genres. I have to run correlation test on all of them. Secondly, temperature is definitely not the only reason that have influence on food preference. There is high possibility that two cities that have similar tempearture have completely food preference due to different belives and so on. I don't believe my original plan is practical and sound. So I changed the target to Ice cream only because ice cream is a very common and popular food in the U.S.. And major religions does not have strict restriction on having ice cream while different meat could be prohibitted by some believes. The biggest challenge for me is to extract the tempearture data. I cannot find one data source that provide history average weather data directly so I have to calculate the average temperature by myself.

**Data Analysis and Visualization:**

- The following two figures shows the temperature spread of the top 100 cities that has most population and the temperature stats. The first figure is drawn by using matplotlib.pyplot. The x-axis shows the temperature and the y-axis shows the count of cities. The 100 cities are divided into 10 bins representing different temperature zones. The interesting part is that the tempearture data is clearly in normal distribution. My original thought is that the first figure will show a very right-skewed distrobution (i.e. most of cities will have extreme cold weather).




```
Mean of Temperature: 50.83700000000001
Median of Temperature: 49.165
Minimum of Temperature: 12.0
Maximum of Temperature: 89.33
Standard Deviation of Temperature: 13.068371294190733
```

- The following figure is a scatter plot with x-axis shows the temperature of cities and y-axis shows the number of ice cream restrunants of cities. It turns out that there is no any linear relationship between the number of ice cream restrunants and temperature. With this figure, I can tell that the popularity of ice cream is not highly related to the temperature even without running the statistical analysis. This is also very interesting because the figure almost shows no linear relationship. In my original thought, there should be at least some influence on the preference on ice cream regarding the temperature.

- The following figure shows the result of the pearson correlation between the temperature and the number of popular ice cream restrunants and the OLS Regression result. With the correlation coefficient of -0.034 which is close to 0, we can tell that there is almost no linear dependency between the two variables. Besides, with the p-value of 0.735 which is more than the general threshold 0.05, the conclusion can be drawn from this result that a statistically significant difference exist. That means, we cannot tell that the temperature has impact on people's preference on ice cream.

```
Pearson Correlation: -0.034247065222225655
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Count   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.009
Method:                 Least Squares   F-statistic:                    0.1151
Date:                Wed, 14 Dec 2022   Prob (F-statistic):              0.735
Time:                        02:04:53   Log-Likelihood:                -173.66
No. Observations:                 100   AIC:                             351.3
Df Residuals:                      98   BIC:                             356.5
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.8841      0.560      3.364      0.001       0.773       2.996
Temperature   -0.0036      0.011     -0.339      0.735      -0.025       0.018
==============================================================================
Omnibus:                       22.055   Durbin-Watson:                   2.166
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               28.313
Skew:                           1.148   Prob(JB):                     7.11e-07
Kurtosis:                       4.234   Cond. No.                         212.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

**Changes and Challenges on Analyzing and Visualizing:** The only change I made is regarding the scatter plot. After collecting all the necessary data, I sorted the "result.csv" file by ascending temperature. Originally, I tried to use plt.plot(col1, col2) to draw a line chart. This is because I thought a positive correlation will exist (i.e. with higher temperature, people's willingness on having ice cream will also be higher). However, the fact is that I was wrong. The line chart is not practical in showing the result. The challenge is how to show the result in a clear and intuitive way. Finally I chose to change the line chart to scatter chart in order to better show the spread of dots. By looking at the scatter chart, it can be clearly tell that there does not exist a linear relationship.

**Future Work:** Since yelp does not allow user to search the data in a specific time duration, what I can do now to test the correlation is only by comparing data in different cities. However, if I can wait until next winter and get the restrunant popularity in the same cities in every month, I can done a comparision on same city with different tempearture to get a more accurate result. While the relationship can still be not very strong, there is possibility that I can dig out more deep relationship between the ice cream preference and temperature.