510 Final Project Document

Kuan-Hui Lin

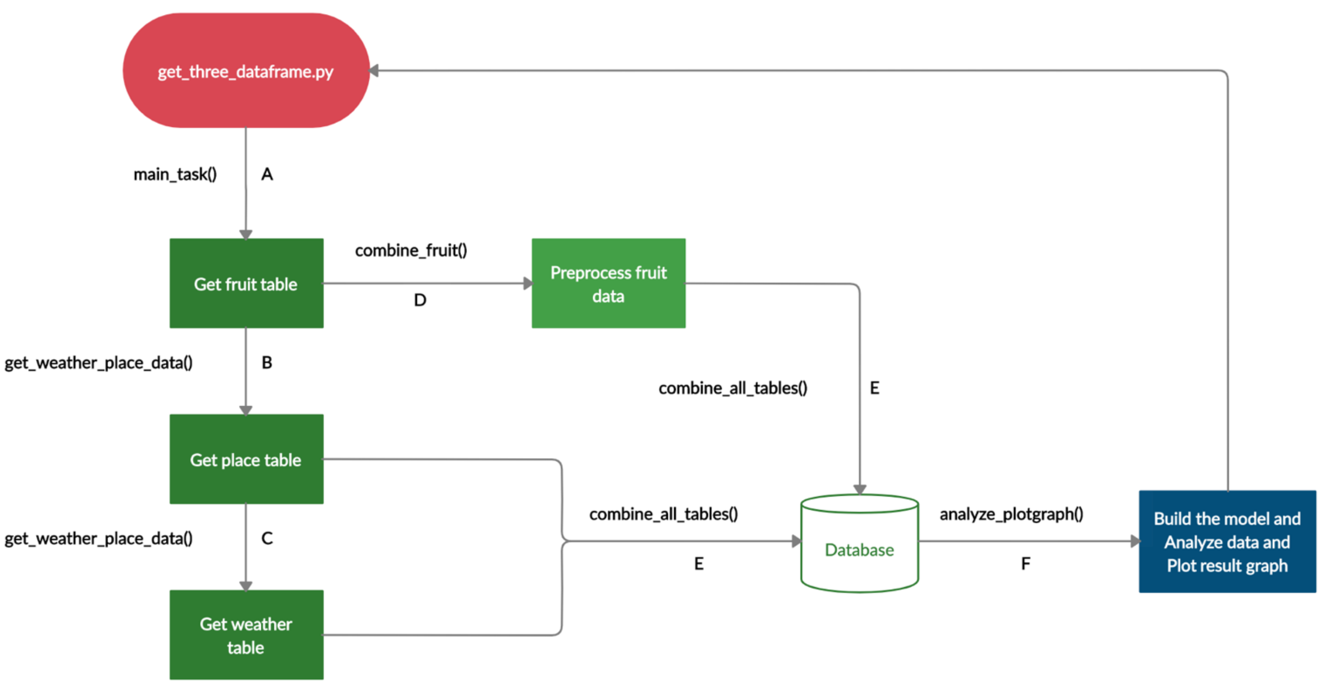
Section 1: Goals of the project.

Fruit is an indispensable food in our lives, and the market prices of fruits in different cities will affect the purchase intention of local consumers, so what factors will cause fluctuations in fruit market prices become an issue that I want to explore. Based on this problem, I came up with many hypothetical factors, temperature, location, and type of fruits, and also established a hypothesis that the temperature of different cities will affect the market price of different kinds of fruit. The purpose of the project is not only to find significant factors but also to verify whether or not the hypothesis is true which is whether or not there is a correlation with these hypothetical factors.

Section 2 / Section 3: Description of what your code does. Try to include high-level flow diagrams in order to understand the code flow easily. Description of your code itself.

- **Overview of my code:**

  In section 2 and section 3, first, I will give the overview of my code and then describing what my code does in detail. **get_three_dataframe.py** is my entry file, and in this file, it includes many multiple functions which are in each step to complete its job in the whole project.



- **Present content of each table:**
    - fruit table:

fruit_price

name of features

Commodity_Name	City	Date	Low_Price	High_Price	Origin	Item_Size
ORANGES	ATLANTA	01/19/2019	20.00	23.00	FLORIDA	56s
ORANGES	ATLANTA	01/19/2019	20.00	21.00	FLORIDA	80s
ORANGES	ATLANTA	01/19/2019	21.25	21.25	FLORIDA	100s
ORANGES	ATLANTA	01/19/2019	18.50	21.50	FLORIDA	125s
ORANGES	ATLANTA	01/19/2019	25.00	28.50	CALIFORNIA	113s
ORANGES	ATLANTA	01/27/2019	20.00	23.00	FLORIDA	56s
ORANGES	ATLANTA	01/27/2019	20.00	21.00	FLORIDA	80s
ORANGES	ATLANTA	01/27/2019	21.25	21.25	FLORIDA	100s
ORANGES	ATLANTA	01/27/2019	18.50	21.50	FLORIDA	125s
ORANGES	ATLANTA	01/27/2019	25.00	28.50	CALIFORNIA	113s

- place table:

place

name of features

City	Marketname	Latitude	Longitude	Address
Atlanta	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
Baltimore	Sprouts Market Corner Coffee	34.0875457	-118.3442673	915 N La Brea Ave, West Hollywood, CA 90038, United States
Boston	Boston Public Market	42.3619639	-71.05703610000000	100 Hanover St, Boston, MA 02108, United States
Chicago	Lincoln Square Farmers Market	41.9664711	-87.68788710000000	N Lincoln Ave & W Leland Ave, Chicago, IL 60625, United States
Columbia	Columbia Farmers Market	38.9581722	-92.3645985	1769 W Ash St, Columbia, MO 65203, United States
Detroit	Northwest Detroit Farmers' Market	42.4068148	-83.2236095	18445 Scarsdale St, Detroit, MI 48223, United States
Los Angeles	Central Avenue Farmers' Market	34.0053028	-118.2567959	4301 S Central Ave, Los Angeles, CA 90011, United States
MiaMi	Legion Park Farmers Market	25.836543	-80.183916	6601 Biscayne Blvd, Miami, FL 33138, United States
New York	Hak Farmer Market	40.596646	-73.978116	167 Avenue U, Brooklyn, NY 11223, United States
Philadelphia	Headhouse Farmers' Market	39.9421239	-75.1453135	2nd & Lombard Sts, Philadelphia, PA 19147, United States

- weather table:

weather

name of features

City	Latitude	Longitude	WeatherCode	Date	LowTemp	HighTemp	AvgTemp
Atlanta	33.749	-84.388	113	01/01/2019	12	16	14
Atlanta	33.749	-84.388	113	01/02/2019	9	14	11
Atlanta	33.749	-84.388	113	01/03/2019	10	11	10
Atlanta	33.749	-84.388	113	01/04/2019	10	15	12
Atlanta	33.749	-84.388	113	01/05/2019	4	12	9
Atlanta	33.749	-84.388	113	01/06/2019	4	15	11
Atlanta	33.749	-84.388	113	01/07/2019	5	15	10
Atlanta	33.749	-84.388	113	01/08/2019	8	17	12
Atlanta	33.749	-84.388	113	01/09/2019	2	9	7
Atlanta	33.749	-84.388	113	01/10/2019	-4	4	2

- After combine all three tables to be a database:

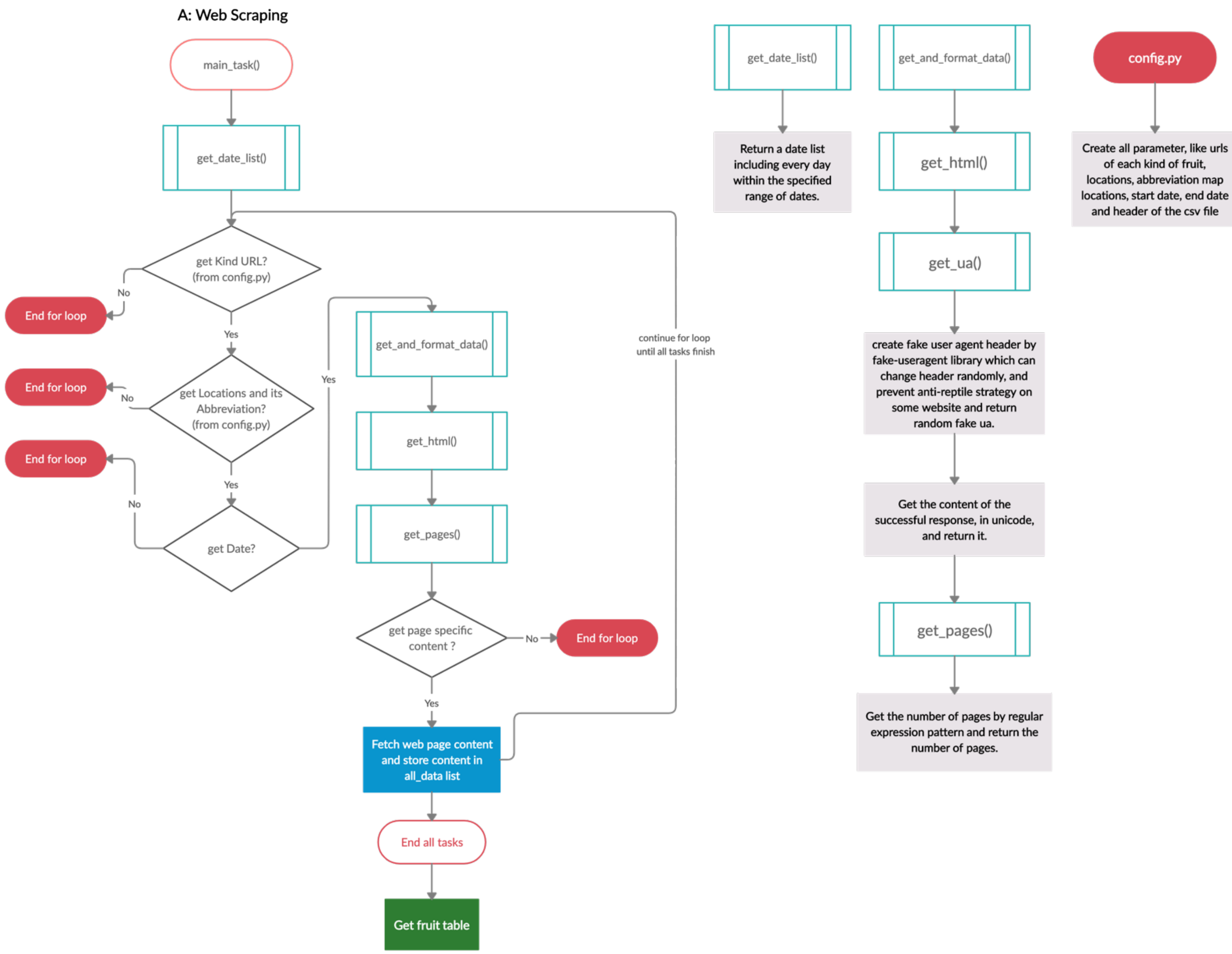
name of features

all_merge_result

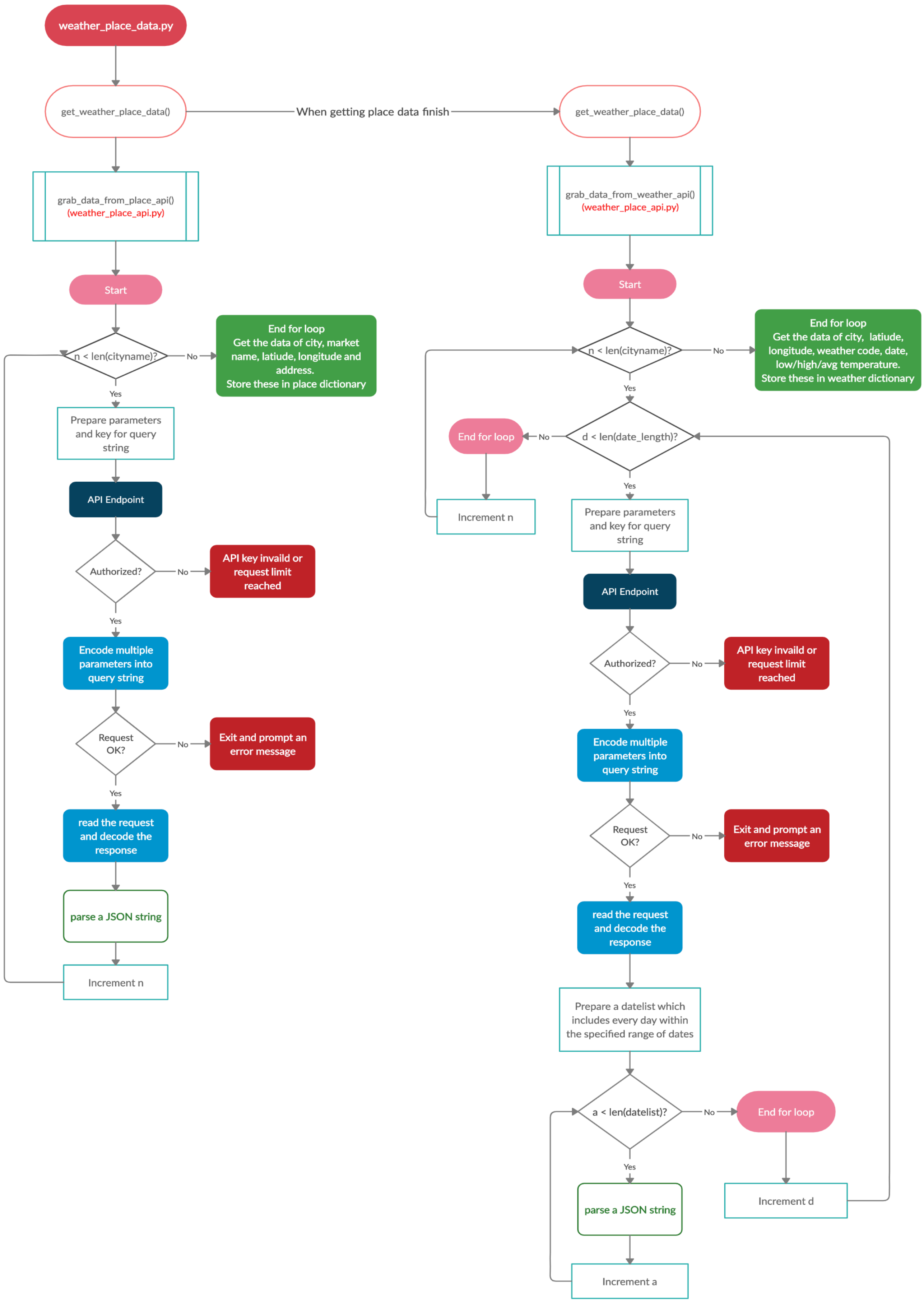
City	Latitude_x	Longitude_x	WeatherCode	Date	LowTemp	HighTemp	AvgTemp	Commodity	Mean_Price	Marketname	Latitude_y	Longitude_y	Address
ATLANTA	33.749	-84.388	113	01/02/2019	9	14	11	LIMES	12.875	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/02/2019	9	14	11	STRAWBERRIES	39.25	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/02/2019	9	14	11	GUAVA	23.0	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/02/2019	9	14	11	BLUEBERRIES	26.625	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/02/2019	9	14	11	ORANGES	34.5	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/03/2019	10	11	10	GUAVA	23.0	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/03/2019	10	11	10	STRAWBERRIES	37.25	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/03/2019	10	11	10	BLUEBERRIES	25.0	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/03/2019	10	11	10	LIMES	12.875	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States
ATLANTA	33.749	-84.388	113	01/03/2019	10	11	10	ORANGES	34.5	Peachtree Road Farmers Market	33.8312225	-84.38624590000000	2744 Peachtree Rd NW, Atlanta, GA 30305, United States

- **Description of my code itself and what each function does.**

  Each flow diagram will describe what each function does based on the overview of my code mentioned above.

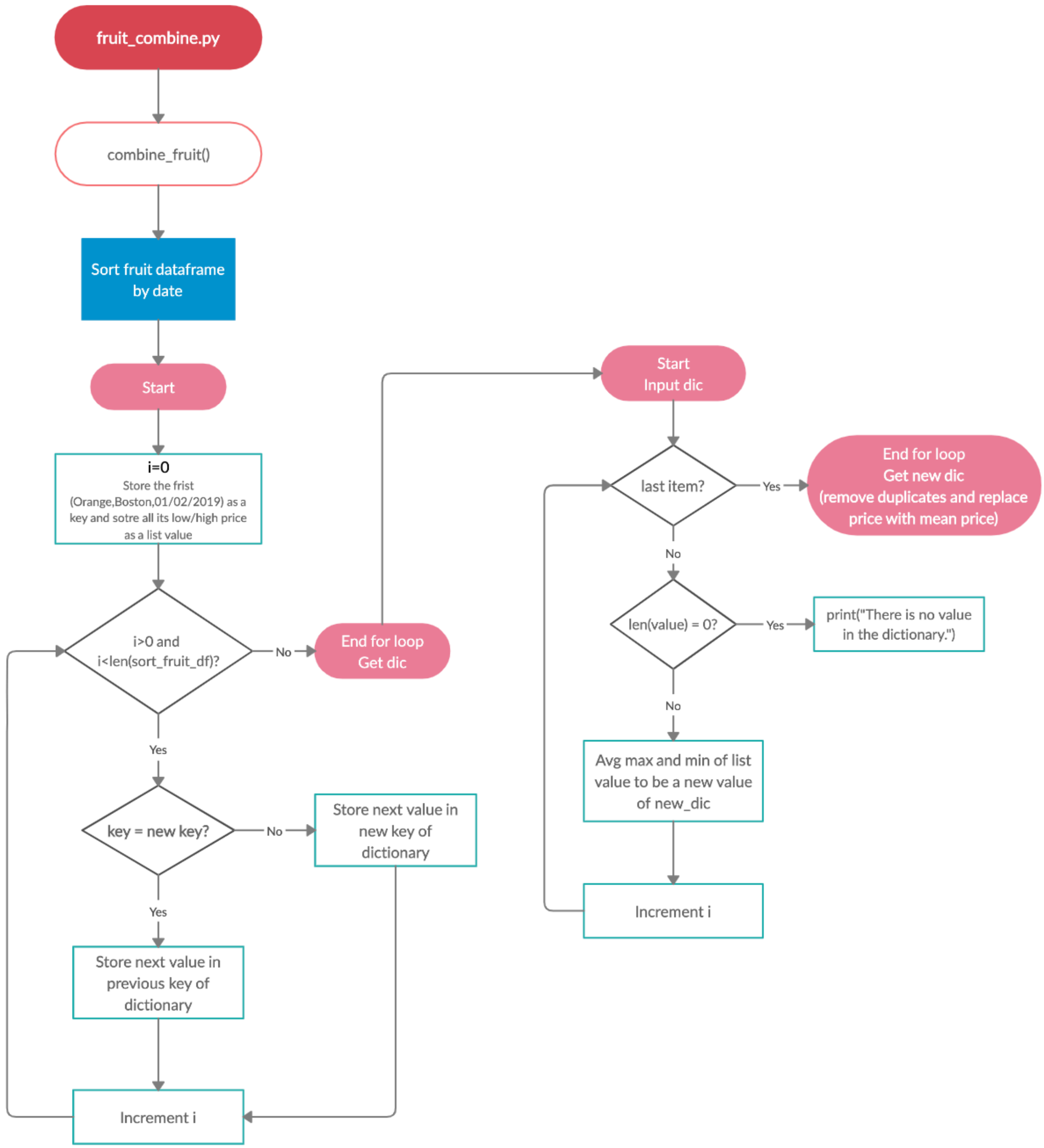


B/C: API

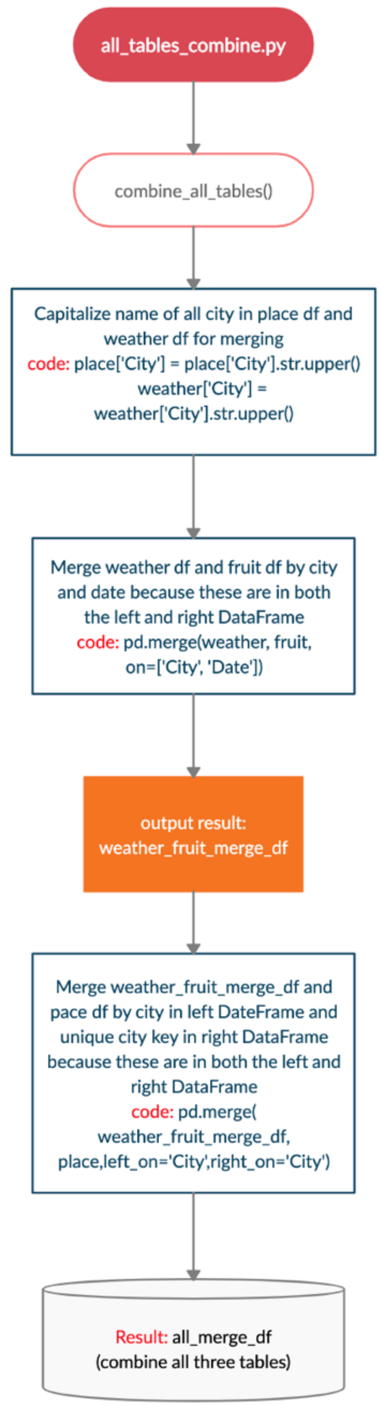


weather_place_data.py
get_weather_place_data()
When getting place data finish
get_weather_place_data()
grab_data_from_place_api()
(weather_place_api.py)
grab_data_from_weather_api()
(weather_place_api.py)
Start
n < len(cityname)?
No
End for loop
Get the data of city, market name, latiude, longitude and address.
Store these in place dictionary
Yes
Prepare parameters and key for query string
API Endpoint
Authorized?
API key invaild or request limit reached
Encode multiple parameters into query string
Request OK?
Exit and prompt an error message
read the request and decode the response
parse a JSON string
Increment n
Start
n < len(cityname)?
End for loop
Get the data of city, latiude, longitude, weather code, date, low/high/avg temperature.
Store these in weather dictionary
d < len(date_length)?
End for loop
Increment n
Prepare parameters and key for query string
API Endpoint
Authorized?
API key invaild or request limit reached
Encode multiple parameters into query string
Request OK?
Exit and prompt an error message
read the request and decode the response
Prepare a datelist which includes every day within the specified range of dates
a < len(datelist)?
End for loop
parse a JSON string
Increment d
Increment a

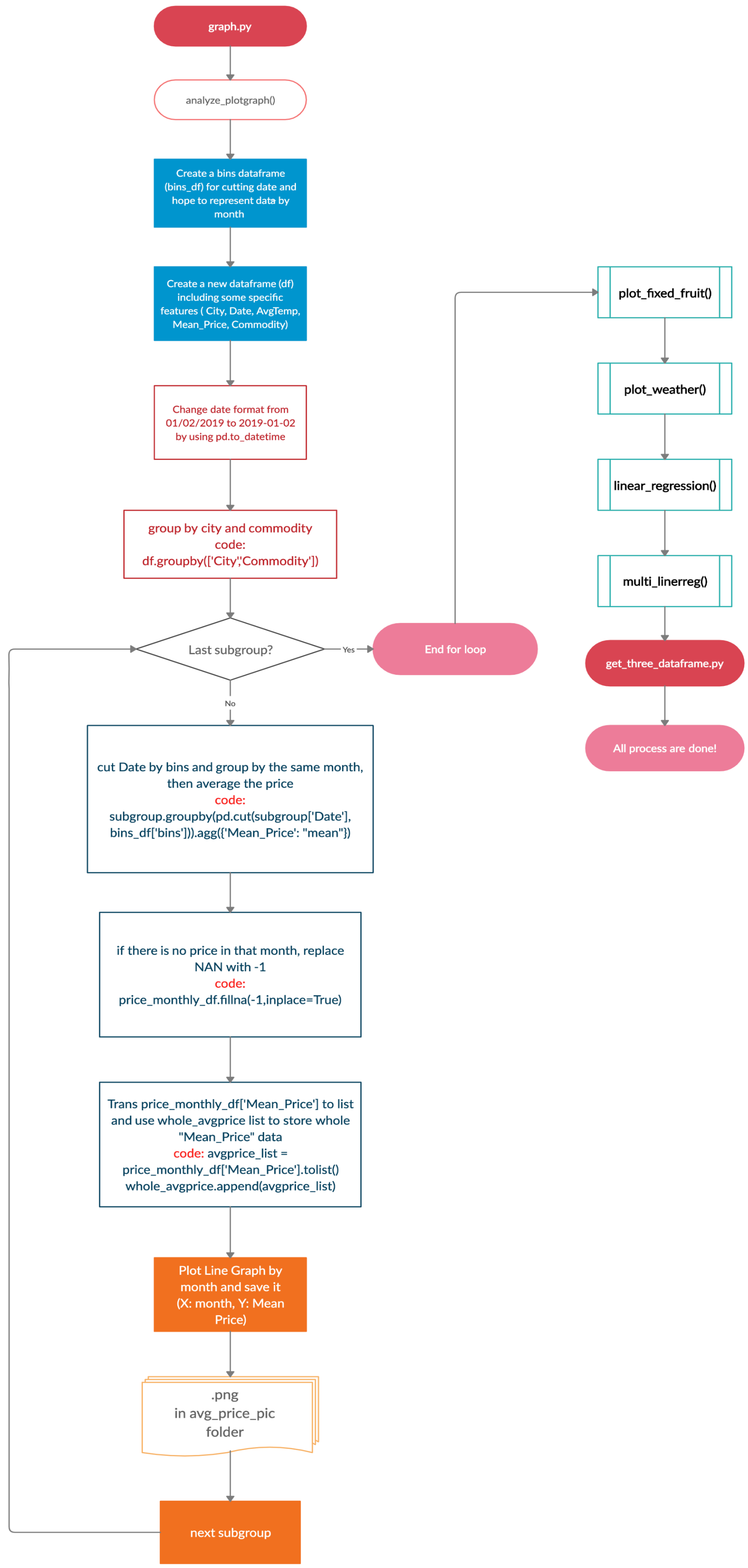
D: Preprocess Fruit Data

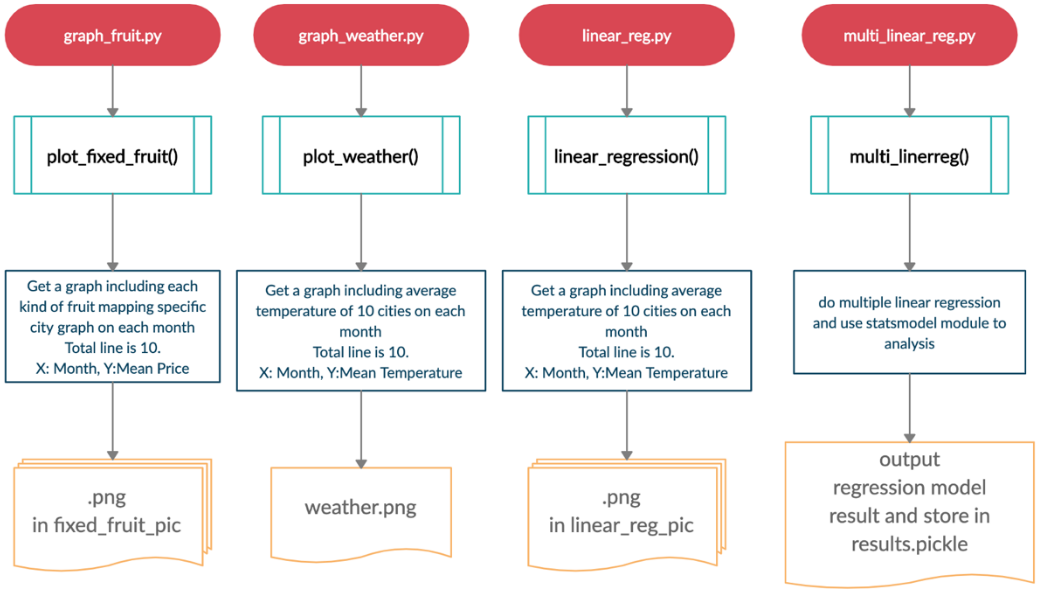


E: combine all tables



F: Analyze Data and Plot Data





- **Show what my code does and describe them in detail on each .py files about plotting the graph.**
  - graph_fruit.py

```python
def plot_fixed_fruit(path, month, price):

    city = ['ATLANTA','BALTIMORE','BOSTON','CHICAGO','COLUMBIA',
            'DETROIT','LOS ANGELES','MIAMI','NEW YORK','PHILADELPHIA']

    colours=['tab:blue', 'tab:orange', 'tab:green', 'tab:red', 'tab:purple',
            'tab:brown', 'tab:pink', 'tab:gray', 'tab:olive', 'tab:cyan']

    create_directory(path)

    # fruit fix
    fruit_dic = defaultdict(list)
    for j in range(len(price)):
        if (j+1)%5 == 1:
            fruit_dic['BLUEBERRIES'].append(price[j])
        elif (j+1)%5 == 2:
            fruit_dic['GUAVA'].append(price[j])
        elif (j+1)%5 == 3:
            fruit_dic['LIMES'].append(price[j])
        elif (j+1)%5 == 4:
            fruit_dic['ORANGES'].append(price[j])
        elif (j+1)%5 == 0:
            fruit_dic['STRAWBERRIES'].append(price[j])

    # plot Line Graph
    leg_list = list()
    plt.figure(figsize = (11,10))
    for key in fruit_dic.keys():    #key is fruit name
        leg_list = list()
        for f in range(len(fruit_dic[key])):
            leg_list.append((str(key),city[f]))
            plt.plot(month, fruit_dic[key][f], c=colours[f%10], label='High',  alpha=0.5, linewidth = 2.0, linestyle = '-', marker='o')
            if (f+1)%10 == 0:
                plt.legend(leg_list, loc='best')
                plt.title("Fruit Fixed and City Changed")
                plt.xlabel("Month")
                plt.ylabel("Mean Price for each type")
                plt.savefig(path +'/'+ key +'.png')
                plt.close()
                plt.figure(figsize = (11,10))
```

set two variables: city and colours

Get each average market price of each fruit from price. Because there are five fruit kinds, use the remainder to heap data.

I want to draw a graph based on each fruit, so go through each key name which is fruit name and get price data by key and then plot them. Save all pictures in the folder, fixed_fruit_pic

- graph_weather.py

```python
def plot_weather(path, bins, weather, month):

    weather['Date'] = pd.to_datetime(weather['Date'])
    bins['bins'] = pd.to_datetime(bins['bins'])
    weather['AvgTemp'] = weather['AvgTemp'].astype(float)
    city_grouped = weather.groupby(weather['City'])

    colours=['tab:blue', 'tab:orange', 'tab:green', 'tab:red', 'tab:purple', 'tab:brown', 'tab:pink', 'tab:gray', 'tab:olive', 'tab:cyan']
    plt.figure(figsize = (12,8))

    i=0
    for name,subgroup in city_grouped:
        weather_monthly_df = subgroup.groupby(pd.cut(subgroup['Date'], bins['bins'],labels=bins.iloc[:-1,0])).agg({'AvgTemp': 'mean'})
        meantemp = weather_monthly_df['AvgTemp'].tolist()

        # plot Line Graph
        plt.plot(month, meantemp, c=colours[i], label=name,  alpha=0.5, linewidth = 2.0, linestyle = '-', marker='v')
        plt.legend()
        # give text on each point
        # for a,b in zip(month,weather_monthly_df['AvgTemp']):
        #     plt.text(a, b, '%.2f' %b, ha='right', va= 'baseline',fontsize=10)
        i+=1

    plt.title("Weather")
    plt.xlabel("Month")
    plt.ylabel("Mean Temperature")
    create_directory(path)
    plt.savefig(path + '/' + 'weather.png')
    plt.close()
```

Because I want to present data by month, I need to use the same date format for both df and handle each subgroup with the same city name.

Because I want to present my data by month, still go through each city(subgroup) and groupby Date in subgroup by bins['bins'] which already is created by previous function and average "AvgTemp". Then store all of them and plot all of them in the same graph.

- linear_reg.py

```python
def linear_regression(path,_df):

    create_directory(path)

    for fruit in _df['Commodity'].unique():

        fix_fruit_df = _df[_df['Commodity'] == fruit]
        slope, intercept, r_value, p_value, std_err = stats.linregress(fix_fruit_df['AvgTemp'],fix_fruit_df['Mean_Price'])

        sns.set(rc={'figure.figsize':(8, 7)})
        ax = sns.regplot(x="AvgTemp", y="Mean_Price", data=fix_fruit_df, color='b',
                         line_kws={'color':'red','label':"y={0:.5f}x+{1:.5f}".format(slope,intercept)})
        # plot legend
        ax.legend()
        size_label = fruit
        plt.title("Commodity = {}".format(size_label))
        plt.savefig(path +'/'+'linear_reg_result_'+ fruit +'.png')
        plt.close()
```

Because I want the data including two features, average temperature and market fruit price, to do linear regression based on each fruit kind, go through each fruit kind and get fix_fruit_df with the same fruit and then do linear regression (stats.linregress). Then it will output the value of regression result. And use sns.regplot to plot resut and save it in linear_reg_pic folder.

- multi_linear_reg.py

```python
def multi_linerreg(path, _df):

    cleanup_nums = {"Commodity":{"STRAWBERRIES": 1, "BLUEBERRIES": 2,'LIMES':3,'GUAVA':4,'ORANGES':5}}

    df_dum = pd.get_dummies(_df['City'])
    df2 = _df.join(df_dum)
    df2_final = df2.iloc[:,2:]        # df2_final not include original city and date
    df2_final.replace(cleanup_nums, inplace=True)
    y = df2_final['Mean_Price']
    X = df2_final[['AvgTemp', 'Commodity','ATLANTA','BALTIMORE','BOSTON','CHICAGO','DETROIT','LOS ANGELES','MIAMI','NEW YORK','PHILADELPHIA']]
    X_expan = sm.add_constant(X)      #increase dimension (i.e. add constant 1 in first column) y=kx+b
    regmodel = sm.OLS(y, X_expan)     #OLS : ordinary least square model
    results = regmodel.fit()
    print(results.summary())
    results.save(path + '/' + "results.pickle")
```

First, before doing multiple linear regression, I need to trans the categorical data to numerical number. So, in terms of City feature, I trans it by using one-hot encoding (pd.get_dummies[_df['City']]) and in terms of Commodity, I replace it (df2_final.replace) by using assigning the number(cleanup_nums).

And get the value of Mean_Price feature to be y and get the value of 11 festures to be variables X, then do multiple linear regression and output the result.

Section 4: Description of any additional packages that you may have used apart from the ones that are discussed in the class.

(1). **from __future__ import unicode_literals:** If your system is Python2, this can help developers let the Python 2 interpreter behave as close to Python 3 as possible and all string literals will become Unicode literals.

(2). **from fake_useragent import UserAgent:** When using Python as a crawler, we need to disguise the header information to deceive the website's anti-crawling strategy. The third-party module fake_useragent can solve this problem, and it will return us a randomly encapsulated good header.

(3). **import os:** It provides a way of using operating system dependent functionality, and it provides allows you to interface with the underlying operating system that Python is running on Mac, Windows or Linux.

(4). **import datetime:** This module can help developers handle dates and times such as date and time arithmetic and output formatting.

(5). **import statsmodels.api as sm**: This module that provide functions for different statistical models and it has many statistical models which can be used in data analysis.

Section 5: Additional procedures to be followed by the grader in order to run the code.

According to this part, I already created a **requirements.txt** which includes all needed python environment, so before running my code, the grader should make sure its computer has these requirements, if not, please install them first.

$ pip install -r requirements.txt

There is the **only one** entry file, **get_three_dataframe.py**, and it will call all functions and then complete all takes in this project.

Section 6: Summary/Presentation of results

1. **Abstract:**

Fruit is an indispensable food in our lives, and the market prices of fruits in different cities will affect the purchase intention of local consumers, so what factors will cause fluctuations in fruit market prices become an issue. Many previous studies focused on demand of consumers, quantity of supply from origin and natural disaster. Therefore, my project tends to find the significant factors which are different from previous research that could influence on the market price of fruit and observed whether or not there is a correlation with these factors. Hence, there are multiple impact factors which are proposed in this project and also discuss their relationship. This study consists of three stages: (a) collected data (b) preprocessed data related to fruit market price and combined three tables (c) analyze data by building model, multiple linear regression, simple linear regression and multiple line graphs. The results show that indeed, the average temperature of all cities, fruit kinds, some specific cities have the relationship with the market price of fruits. Among them, the average temperature of cities is a significant factor. Furthermore, in terms of GUAVA, Los Angeles with relatively high average temperature, it also presents the lowest market price of GUAVA, and contrary to ORANGES, it presents the highest market price of ORANGES from June to September.

2. **Data:**

In this project, I adopted weather, place and market price of different kinds of fruit to be my three datasets. First, I collected the place data from Place API of Google Maps API based on farmers market of city and name of city. Second, I obtained all historical daily weather data in 2019 from Weatherstack API which is not static data source and update everyday based on name of city and the specified range of dates. Third, all market fruit price used in this project was sourced from the website, freshfruitportal.com, which includes different kinds of fruit and its daily market price, and I collected the whole year historical market price of five kinds of fruits in 2019. Hence, the total number of collecting place data was 10 and the total number of collecting weather data was 3,650 and total market price fruit data was more than 90,000. After preprocessing and removing redundant fruit data and combining all three datasets, I obtained the total data was about 18,190 which including 14 features for further studies. In this project, I only extracted 5 highly related features, name of city(City), fruit kinds(Commodity), fruit market price(Mean_Price), average temperature(AvgTemp) and date(Date), to do research, build model and analyze.

3. **Methods:**

The goal of my project is to find key factors and to observe their relationship between these factors that is whether or not the average temperature factor in some specific cities influences the market price of different kinds of fruits in some specific cities. There are three stages in the method. First, in order to find key factors and explain the relationship between one continuous dependent variable and two or more independent variables which can be numerical or categorical data, I adopted the multiple linear regression model fitting a linear equation to observed data. In order to make name of city variable meaningful, I used one-hot encoding. Because there is no order in the city variable, they cannot be converted into numerical values with differences in size. The model was used in this project, I set variable y is mean price of each kinds of fruit and explanatory variables(x) are remaining variables except for the date, and this relationship can be expressed by:

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \beta_0$$

The equation describes how the explanatory variables influence y to change. In the multiple linear regression model, β_0 is called regression constant and β_p is called regression coefficient.

In the study, it is assumed that the market price of fruit is determined by the twelve variables, name of cities, commodity, average temperature. Through multiple linear regression, it can help me filter out some useful features. Second, I fitted a simple linear regression model to predict the response, and in this model, I only analyzed the relationship between average fruit market price and average temperature. Finally, through line graphs with each month, I analyzed three features's relationship--market prices of fruit, name of the city and fruit kinds, and at the same time referenced the weather graph including name of the city and average temperature of the city in each month. Hence, I can obtain the relationship between four variables, average market price, name of the city, average temperature of the city and fruit kinds and observe the line change of line graph and compare them together.

4. Analysis and Results

4.1 Effect Factors of Market Price of Fruit

The primary drawback in using simple regression analysis for the project is that it is very difficult to draw the conclusions about how x affects y, thus, multiple regression analysis is more amenable to this part analysis because it allows us to explicitly control for many other factors that simultaneously affect the dependent variable. Dependent fruit price variables are presented daily and were collected from freshfruitportal website which including the fruit price of U.S. Department of Agriculture (USDA). The average temperature also is obtained daily. Estimation is from the January of 2019 to the December of 2019. And all of the estimations are under the 95% confidence.

First, the average market price of fruit is estimated as the dependent variable. The regression model can be got as:

$$Y_{avg\_market\_price} = (-0.0316)x_{temp} + (1.5161)x_{kind} + (1.2007)x_{ATLANTA} + \cdots + (-0.1198)x_{PHILADELPHIA} + 18.3767$$

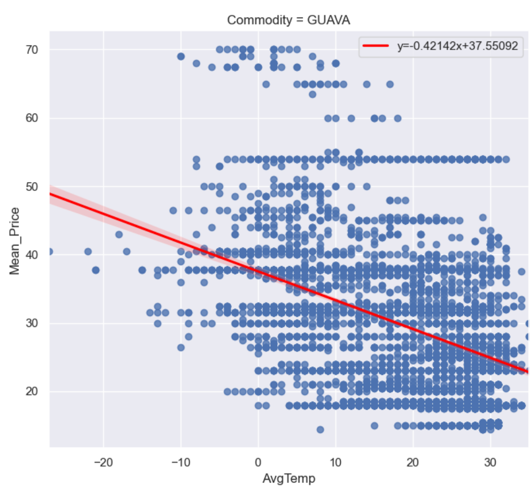
According to the regression estimation [1], I found that each of predictors except for PHILADELPHIA, DETROIT and MIAMI could be used to reject the null hypothesis H0: βj =0 because the p-values of these predictors were less than the confidence level, often 0.05. It indicates that ten variables except for PHILADELPHIA, DETROIT and MIAMI, show statistical significance which means all of these variables can influence the average market price of fruits. When the average daily temperature increase, it will bring 0.0316 dollar decrease to fruit average market price. In addition, the variety of fruits will affect the fruit average market price. When increased demand for a certain type of fruit, it will also make fruit average market prices rise. Either positive or negative influence is working that is name of city, and in this estimation, when fruit in ATLANTA, it will increase by 1.2007 units. There are some insignificant influence factors are PHILADELPHA, DETROIT and MIAMI because its p-value is 0.707, 0.406 and 0.096 which is not less than 0.05. That is to say, when increasing the value of the variables, PHILADELPHIA, DETROIT and MIAMI, it will not cause significant average market price of fruits fluctuations. Hence, in the regression model, indeed, a significant effect exists in these factors except for PHILADELPHA, DETROIT and MIAMI such as average temperature of cities, kinds of fruit, and the place location.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Mean_Price   R-squared:                       0.071
Model:                            OLS   Adj. R-squared:                  0.071
Method:                 Least Squares   F-statistic:                     127.0
Date:                Mon, 11 May 2020   Prob (F-statistic):          3.92e-282
Time:                        22:01:17   Log-Likelihood:                -66807.
No. Observations:               18189   AIC:                         1.336e+05
Df Residuals:                   18177   BIC:                         1.337e+05
Df Model:                          11
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           18.3767      0.311     59.064      0.000      17.767      18.987
AvgTemp         -0.0316      0.007     -4.272      0.000      -0.046      -0.017
Commodity        1.5161      0.050     30.344      0.000       1.418       1.614
ATLANTA          1.2007      0.316      3.799      0.000       0.581       1.820
BALTIMORE        2.4612      0.318      7.737      0.000       1.838       3.085
BOSTON           4.4286      0.322     13.748      0.000       3.797       5.060
CHICAGO          1.1235      0.326      3.444      0.001       0.484       1.763
DETROIT          0.2699      0.325      0.831      0.406      -0.366       0.906
LOS ANGELES      2.3505      0.316      7.441      0.000       1.731       2.970
MIAMI           -0.5301      0.318     -1.665      0.096      -1.154       0.094
NEW YORK         2.1057      0.319      6.595      0.000       1.480       2.732
PHILADELPHIA    -0.1198      0.318     -0.376      0.707      -0.743       0.504
==============================================================================
Omnibus:                    15883.367   Durbin-Watson:                   2.119
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1146202.780
Skew:                           3.856   Prob(JB):                         0.00
Kurtosis:                      41.117   Cond. No.                         217.
==============================================================================
```

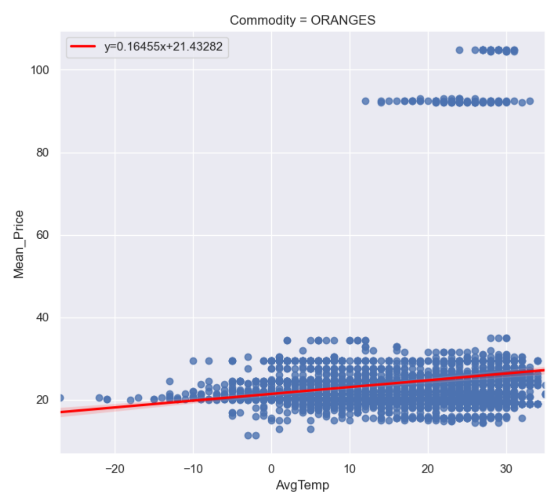
[1]

4.2 Temperature factor effect the market price of fruit base on different fruit kinds.

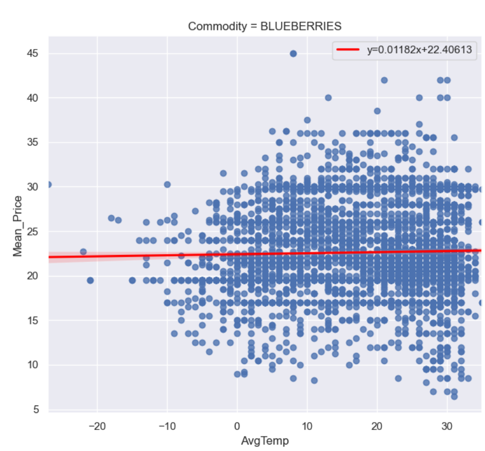
After knowing which factors are higher or lower influence on average market price of fruits, in order to back up my assertion, because there is a key factor, average temperature, in my assumption, I concentrated on the analysis of relationship between average temperature of cities and average market fruit price fluctuations. I extracted the average temperature of all cities feature and average market price of fruits feature based on different kinds of fruit for further discussion. According to five simple linear regression result graphs [2,3,4,5,6], I found that these average market price of all five kinds of fruits are related with change in average temperature of cities. The regression coefficient told me the influence of one variable on another. Hence, among them, the average market price of GUAVA is relatively sensitive to average temperature changes, followed by ORANGES. In terms of GUAVA, when increasing the average temperature one unit, it will lower the average market price of the fruit 0.42142 dollars. Contrary to ORANGES, when increasing the mean temperature one unit, it will also higher the average market price of the fruit 0.16455.



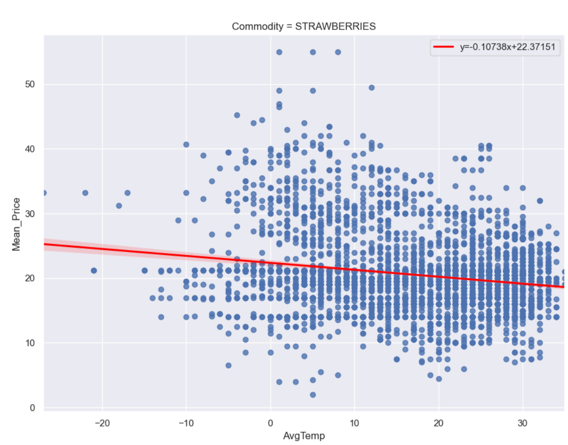
[2]



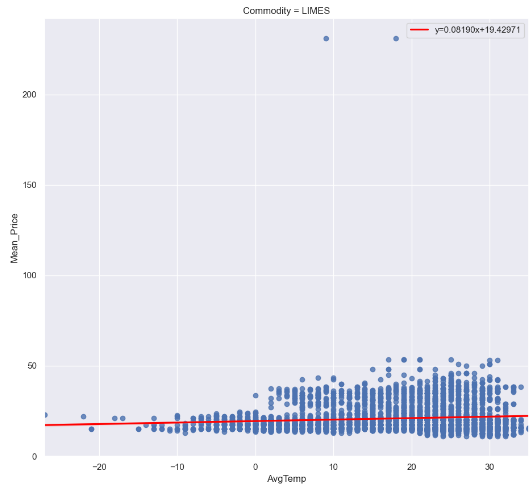
[3]



[4]



[5]

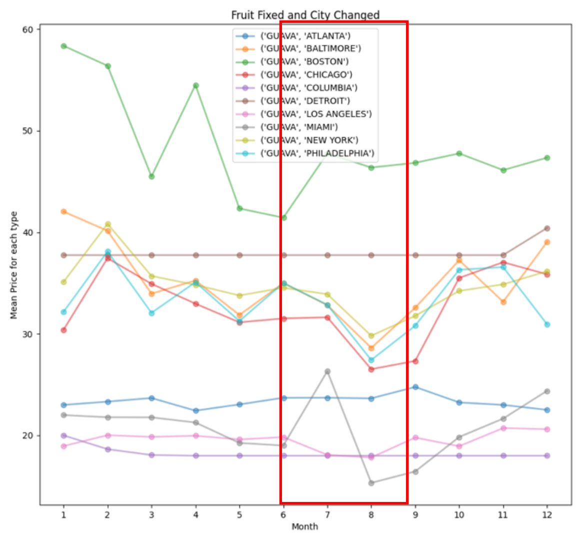


[6]

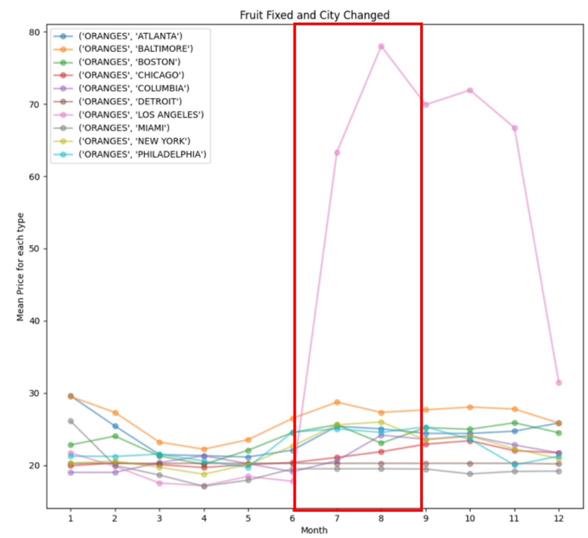
4.3 Comprehensive comparison

According to previous analysis [4.1], I only know that indeed, average temperature, fruit kinds, some cities except for PHILADELPHIA, DETROIT and MIAMI will affect the average market price of fruits, but how to affect will be analyzed in this section. Due to the mention in 4.2, in terms of GUAVA, the higher average temperature will lead to the average market price of GUAVA decrease. Mapping the line graph of weather of each city [12], in general, the average temperature from June to September is relatively high compared to other months, and in the GUAVA graph [7], I focused on the average market price of GUAVA from June to September and found that majority of cities, like, Los Angeles, Chicago, Baltimore and New York, had lower average market price of GUAVA. Among them, the lowest average market price of GUAVA is in Los Angeles, and through weather graph, I can know the 2019 average temperature in Los Angeles is higher than in Chicago, Baltimore and New York.

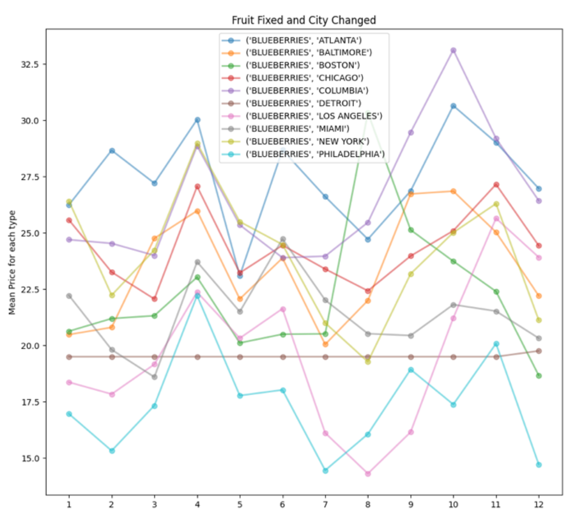
Another example is ORANGES [8]. According to previous analysis in 4.2, I know that ORANGES also have the second highest correlation to the average temperature. It indicates that the average temperature and average market price of ORANGES have a positive correlation. Similarly, I observe the average market price changes from June to September in different cities. Remove insignificant variables such as PHILADELPHIA, DETROIT and MIAMI, mentioned in section 4.1 and still observe the change of average market price of ORANGES from June to September which are months with relatively high average temperatures. According to weather graph [12], it shows that the highest average temperature in 2019 is Los Angeles, and I also found that the average market price of ORANGES in Los Angeles is the highest. In addition, three cities with relatively low average market prices of ORANGES from June to September is MIAMI, DETROIT and CHICAGO. But I do not consider MIAMI, DETROIT because these two variables are insignificant factors in section 4.1. Hence, the city, CHICAGO, represents the positive correlation between average market price of ORANGES and average temperature of CHICAGO. That is to say, the lowest average temperature is CHICAGO and the lowest average market price of ORANGES is still CHICAGO from June to September. In this section, I analyzed the top two most weather-related fruits, GUAVA and ORANGES, which are showed in section 4.2 and also discussed the relationship between the three variables, city, fruit kinds, and average temperature based on month.



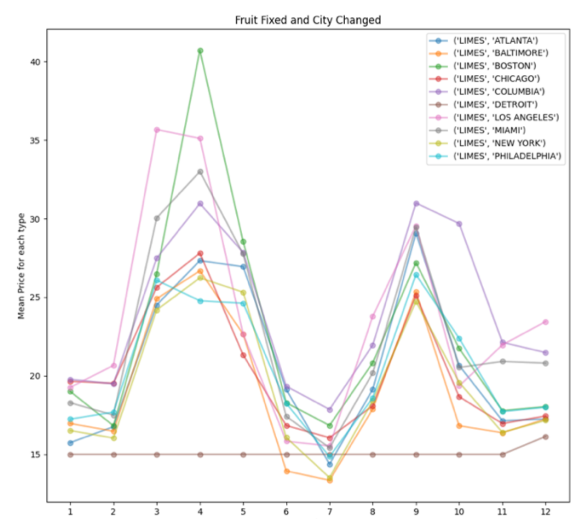
[7]



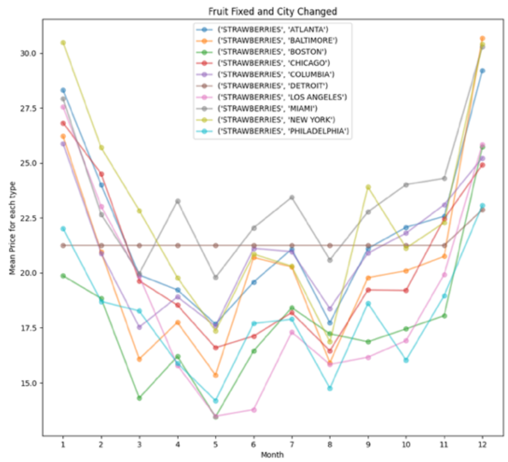
[8]



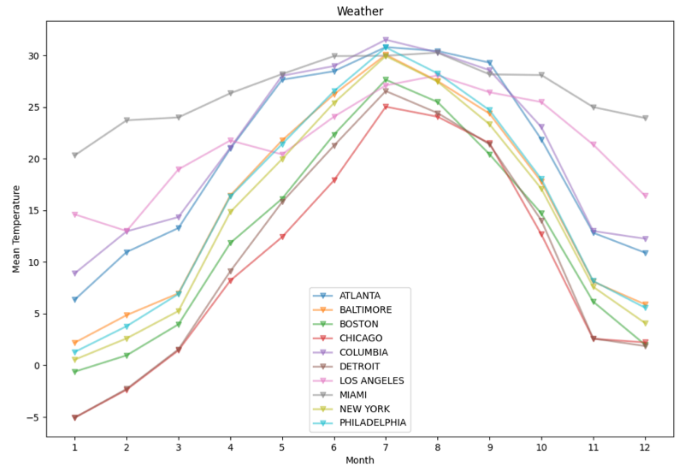
[9]



[10]



[11]



[12]

5. Conclusion/Discussion

Through this study, it was found that average temperature has an impact on fruit average market prices, with ORANGES and GUAVA being the most significant. My experiment is divided into three stages. First, through the multiple linear regression analysis of section 4.1, I know that some variables do not make any contribution to average market fruit price fluctuations, and then through the simple linear regression analysis of section 4.2, I know the average market price of what kinds of fruit has positively correlated or negatively correlated with average temperature. Finally, using the results from 4.1 and 4.2, I analyze the average market fruit prices of seven cities and two kinds of fruits with line graph of weather.

From the average temperature graph [12], I can know that the average temperature in all cities in 2019 increased from June to September. Therefore, I observed the changes in average market prices of same kinds of fruit from June to September in different cities. Two analysis results can be drawn. First, when the average temperature is higher, the average market price of GUAVA is lower. And Los Angeles has the highest average temperature of all cities in 2019, so the average market price of GUAVA in Los Angeles is the lowest compared to Chicago, Baltimore and New York. Second, when the average temperature is higher, the average market price of ORANGES is higher, and Los Angeles has the highest average temperature of all cities in 2019, so the average market price of ORANGE in Los Angeles is also the highest. Not only that, in the same month interval, I found that except for MIAMI and DETROIT, the average temperature in Chicago is the lowest, and the average market price of ORANGES is also the lowest. Thus, the assumption in the project can be supported and the temperature is one of key factors that can affect fruit market price fluctuation.

In this project, I only adopted four features, name of city, average temperature, average fruit market price and date to do my research, but there are still many factors, like typhoon, fruit appearance…etc., which could affect the average market price of fruits. In the future, the researchers who are interested in this topic, adding more related features can let this research more complete.