California Air Quality Analysis and Investigation of Climate and Population Influences

Final Project Report

Manqi Li

1. Project description

Air pollution, coming from some harmful particulates or materials introduced into atmosphere, can degrade quality of life and damage economic system. The air quality is usually reflected by various pollutant concentrations. Inner relationships between various pollutant concentrations are an interesting aspect to be studied. Also, correlations between air quality and climate and human activity factors needs to be investigated. In this project, the real-time air quality index and various pollutant concentrations (CO, NO_2, O_3, SO_2, PM2.5, PM10) in different regions in California are achieved by latitude and longitude via API. The real-time local wind strength and temperature are requested by latitude and longitude via API. These scraped values can be considered as climate parameters. Population is a good parameter to assess the human activity in a given region, which can be accessed by census data in California (based on zip code). A conversion between the zip code and latitude and longitude is utilized for concatenating different data sources for further analysis. The air quality in different regions in California is visualized in a choropleth map plot. Air quality will also be plotted in a histogram and pie chart to further visualize the distribution. Air pollutant concentration boxplot and pairplot are plotted for distribution visualization and potential inner relationship analysis. The correlation of air pollutant concentrations is computed to analysis the relationship. Some regression models are established to explain the inner relationship. Furthermore, the correlation of air quality with both climate and population factors are studied.

2. How to run the code

Required libraries: pandas, pgeocode, requests, time, urllib, urlopen, json, plotly.express, urlopen, numpy, matplotlib, seaborn, statsmodels

Link to GitHub: https://github.com/manqili0127/DSCI510_Final.git

Below are detailed steps on how to run the code, and a flow chart can be found in Fig1.

1) Unzip the file. Under the "code" folder, there should be 3 code files: "get_the_data.py", "visualization_and_analysis.py", and "main.ipynb". Under the "data" folder, there are 4 CSV files: "pop-by-zip-code.csv", "population_zipcode_lat_lon.csv", "climate.csv", and "air_quality.csv".

2) The "get_the_data.py" is for data collection. Data collection needs the "pop-by-zip-code.csv" as an input. If run, please move the "pop-by-zip-code.csv" to be under the same directory of "get_the_data.py" (move to the "code" folder). After running, the output (web scrapped data) will be three csv fiels: "population_zipcode_lat_lon.csv"(population data), "climate.csv"(climate data), and "air_quality.csv"(air quality data). Sample output data files can be found under the "data" folder.

Note: The data collected is real-time data. If run, the data will be different from the sample data. Subsequent visualization and analysis are based on the collected data in the "data" folder.

3) The "visualization_and_analysis.py" contains the functions for data visualization and analysis. The "main.ipynb" is the main code of this project. Running this file will automatically call the functions in the "visualization_and_analysis.py". The main code needs 3 input CSV data files. So please put "population_zipcode_lat_lon.csv"(population data), "climate.csv"(climate data), and "air_quality.csv"(air quality data), "visualization_and_analysis.py", and "main.ipynb" in the same directory. The outputs are visualized figures and analysis results.

4) Visualized figures are saved under the "results" folder.

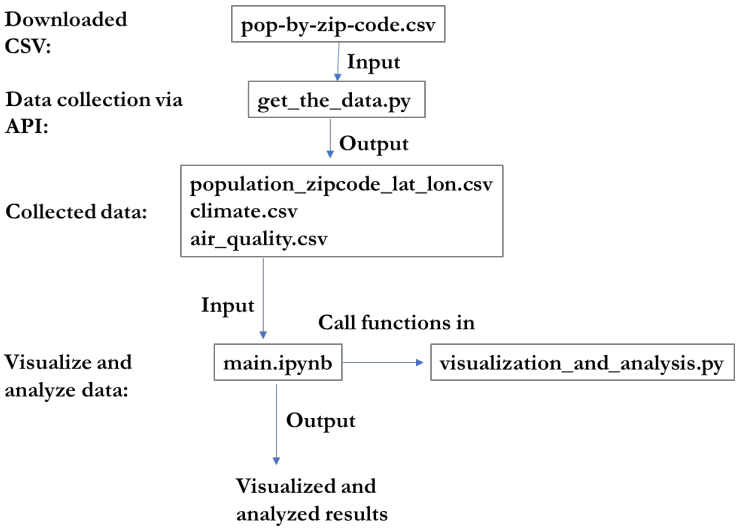


Fig.1 Flow chart on how to run the code

3. Data collection

3.1 Data sources and data collection

Three data sources are used in this project. The population dataset is available from census results and real-time air quality dataset and climate dataset are web-scrapped using APIs. Flow chart of data collection can be found in Fig.2. Detailed data collection procedures and data collection are listed below.

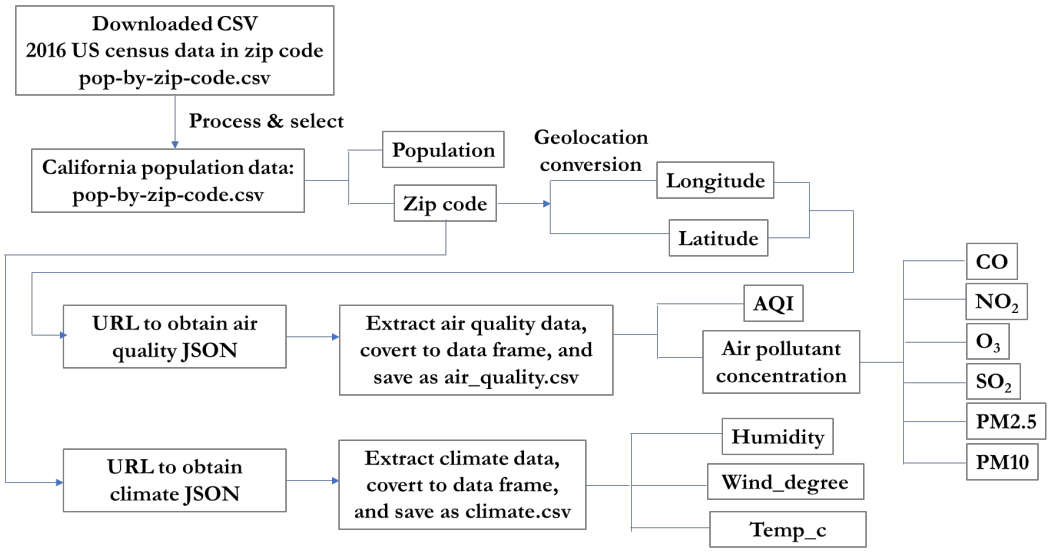


Fig.2 Flow chart of data collection

(1) Population by zip code in California

The population data is taken from the American Community Survey (https://data.world/lukewhyte/us-population-by-zip-code-2010-2016). This data is presented in a CSV file. Here, the latest data (year 2016) is used for future analysis. Since the dataset is in the national range, data with zip code ranging from 90001 to 96162 (California zip code) will be selected. The final data size is 1762. Data sample can be found in Fig.3.

	zip_code	y-2016
0	90001	57942
1	90002	51826
2	90003	70208
3	90004	63095
4	90005	39338

Fig.3 Population by zip code in California data

The dataset (1) is based on zip code, but other datasets scraped by API are based on latitude and longitude. Therefore, a geolocation conversion should be made to connect different data sources. A library for querying of GPS coordinates from postal codes (https://pypi.org/project/pgeocode/) can be sued here for geolocation conversion. Fig.4(a) shows a conversion example output from zip code (90015) to latitude and longitude (34.0434, -118.2716). The latitude and longitude information are extracted. All the converted geolocations are appended as latitude and longitude columns in the original data frame are shown in Fig.4(b).

(a)

postal_code	90015
country_code	US
place_name	Los Angeles
state_name	California
state_code	CA
county_name	Los Angeles
county_code	37.0
community_name	NaN
community_code	NaN
latitude	34.0434
longitude	-118.2716
accuracy	4.0

(b)

	zip_code	y-2016	latitude	longitude
0	90001	57942	33.9731	-118.2479
1	90002	51826	33.9497	-118.2462
2	90003	70208	33.9653	-118.2727
3	90004	63095	34.0762	-118.3029
4	90005	39338	34.0585	-118.3012

Fig.4 (a) Conversion example from zip code (90015) to latitude and longitude (34.0434, -118.2716); (b) Converted geolocation appended as latitude and longitude columns in the original data frame.

(2) Real-time air quality and various pollutant concentrations

An API (https://rapidapi.com/apininjas/api/air-quality-by-api-ninjas) is used here to collect current air quality data for any region (JSON file). The collected data includes overall air quality index and various pollutant concentrations (CO, NO2, O3, SO2, PM2.5, PM10). Since the zip code request does not work well in this API, the data will be requested by latitude and longitude converted by pgeocode in data source (1). Because the API is not stable when requesting large amount of data in a single run, the requested process is divided into 9 runs. For stability consideration , 200 data are in the first 8 runs and 176 data are in the last run, with 15s sleep time between each run. The final dataset is 1762 data points. Data sample can be found in Fig.5.

	CO	NO2	O3	SO2	PM2.5	PM10	aqi
0	867.84	122.01	3.67	13.83	27.26	42.34	123.0
1	867.84	122.01	3.67	13.83	27.26	42.34	123.0
2	867.84	122.01	3.67	13.83	27.26	42.34	123.0
3	307.08	43.87	60.80	3.10	4.31	12.11	55.0
4	307.08	43.87	60.80	3.10	4.31	12.11	55.0

Fig.5 Real-time air quality and various pollutant concentrations data

(3) Current wind strength, humidity and temperature

A weather and geolocation API is used to get current weather information in given region (https://rapidapi.com/weatherapi/api/weatherapi-com). The data (JSON file) is requested based on postal code in data source (1). Wind strength, humidity and temperature are collected in each region. The dataset size is the same as data source (1) with 1762 data points. Data sample can be found in Fig.6.

	wind_degree	humidity	temp_c
0	270.0	18.0	24.4
1	270.0	18.0	24.4
2	270.0	18.0	24.4
3	270.0	18.0	24.4
4	270.0	18.0	24.4

Fig.6 Real-time wind strength, humidity and temperature data

3.2 Data concatenating

Three datasets are concatenated into one single dataset as shown in Fig.7 for future analysis and visualization. The dataset includes population data, real-time data of air pollution, and current weather data in different regions in in California.

	zip_code	latitude	longitude	population	CO	NO2	O3	SO2	PM2.5	PM10	aqi	wind_degree	humidity	temp_c
0	90001	33.9731	-118.2479	57942	867.84	122.01	3.67	13.83	27.26	42.34	123.0	270.0	18.0	24.4
1	90002	33.9497	-118.2462	51826	867.84	122.01	3.67	13.83	27.26	42.34	123.0	270.0	18.0	24.4
2	90003	33.9653	-118.2727	70208	867.84	122.01	3.67	13.83	27.26	42.34	123.0	270.0	18.0	24.4
3	90004	34.0762	-118.3029	63095	307.08	43.87	60.80	3.10	4.31	12.11	55.0	270.0	18.0	24.4
4	90005	34.0585	-118.3012	39338	307.08	43.87	60.80	3.10	4.31	12.11	55.0	270.0	18.0	24.4

Fig.7 Concatenated dataset including three data sources

3.3 Collected dataset files

Three collected datasets are saved as CSV files under the folder “sample data”. The file names with corresponding descriptions can be found below in Table 1.

File name	Data description	Original format	Saved format
population_zipcode_lat_lon.csv	California population in each region with corresponding zip code, latitude and longitude	CSV from census result	CSV
air_quality.csv	Air quality data in each region including five major pollutant concentration and air quality index	Using API. The collected data (JSON file) is requested based on latitude and longitude in data source 1, Air quality index and various pollutant concentrations (CO, NO2, O3, SO2, PM2.5, PM10) are collected	CSV
climate.csv	Weather data in each region including temperature, humidity and wind degree	Using API. The data (JSON file) is requested based on postal code in data source 1. Wind strength, humidity and temperature are collected in each region	CSV

Table 1. File names with corresponding descriptions

4. Visualization and analysis

4.1 Air pollution data visualization

(1) Overall air quality index

To visualize the overall air quality index distribution in different regions in California, a choropleth map using plotly.express and carto base map is plotted as shown by Fig.8. Each region corresponds to a Federal Information Processing Standards (Fips) code, which is a five-digit code identified unique counties in the United States. The original collected air quality data is based on latitude and longitude, so an API is used here to convert latitude and longitude to Fips code. The color represents air pollution degree. The lighter the color, the better the air quality is in that area. It can be seen from the data that the air quality in northern areas is better than that in southern areas, while the air quality in coastal regions is better than that in inland regions.

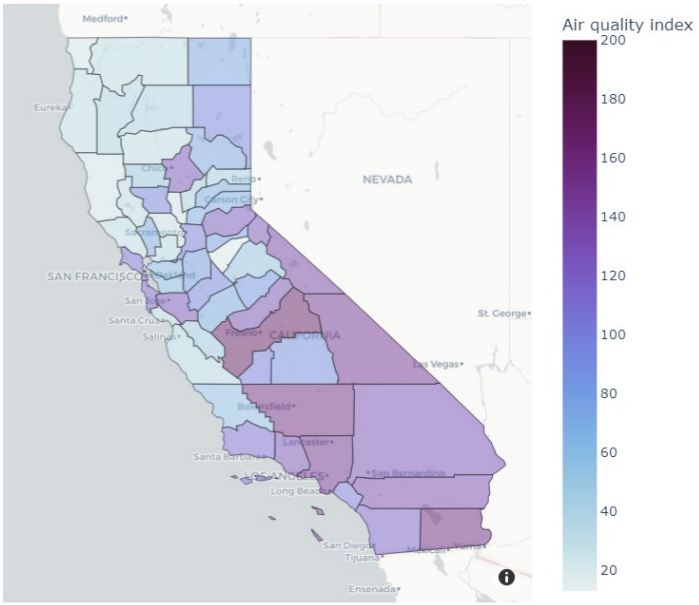
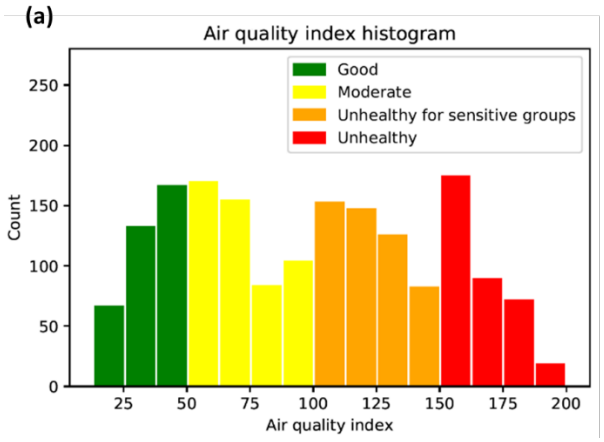


Fig.8 Choropleth map of overall air quality index distribution in California

To visualize the air quality index distribution, a histogram is plotted. Fig.9(a) shows the air quality index (AQI) histogram with 4 air quality levels: good (<50), moderate (50~100), unhealthy for sensitive groups (100~150), and unhealthy (>150). Statistical summary is shown in Fig.9(b). The mean AQI is 98.58 and median AQI is 100, both lying around the boundary between moderate level and unhealthy for sensitive groups level. A pie chart of AQI is plotted for further visualizing air quality level distribution in Fig.10. Most area lie in the moderate level and unhealthy for sensitive groups level. Levels with most counts to least counts are unhealthy for sensitive groups, moderate, good, and unhealthy.



(b)

	aqi
count	1761.000000
mean	98.581488
std	48.501198
min	13.000000
25%	55.000000
50%	100.000000
75%	139.000000
max	200.000000

Fig.9 (a) air quality index histogram, (b) air quality index statistical summary

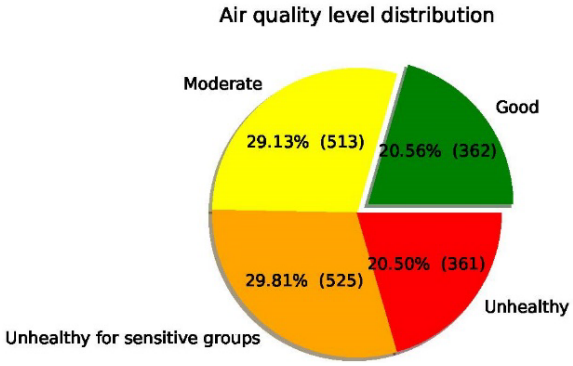


Fig.10 Air quality level distribution pie chart

(2) Various pollutant concentration

Five major pollutant concentrations (CO, NO_2, O_3, SO_2, PM2.5, PM10) are collected and visualized in a box plot as shown in Fig. 11. CO is the most dominant pollutant, while SO_2 is the least dominant pollutant. An air pollutant concentration pairplot is plotted in Fig. 12. From the scatter plots, it can be found some dependencies existing between pollutants, for example, CO vs NO_2, and PM2.5 vs PM10. When one pollutant increases, the other pollutant increase in a linear trend. Further analysis are implemented in the next session to investigate inner relationship between various pollutants.

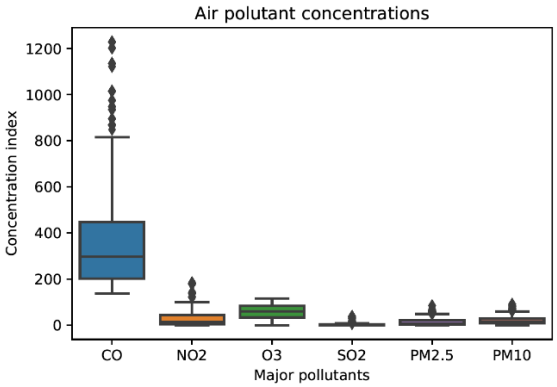


Fig.11 Air pollutant concentrations box plot

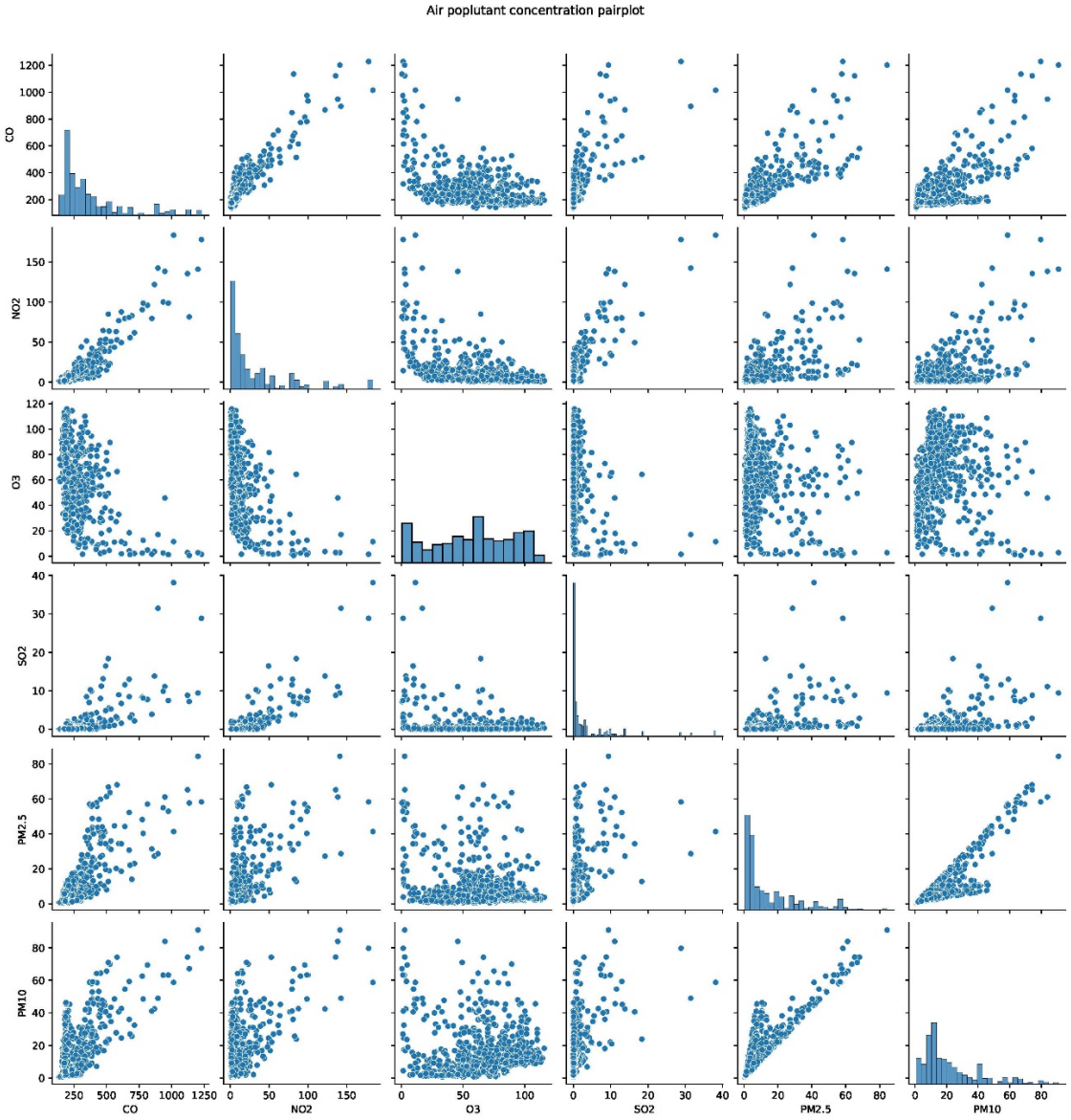


Fig.12 Air pollutant concentration pairplot

4.2 Data analysis

(1) Inner relationships between different pollutant concentrations

Correlations between different air pollutants are plotted in a heatmap shown in Fig.13 with color indicating how strong the relationship is. It can be seen that some strong relations existing between some pollutants. For instance, a positive correlation between CO and NO_2.

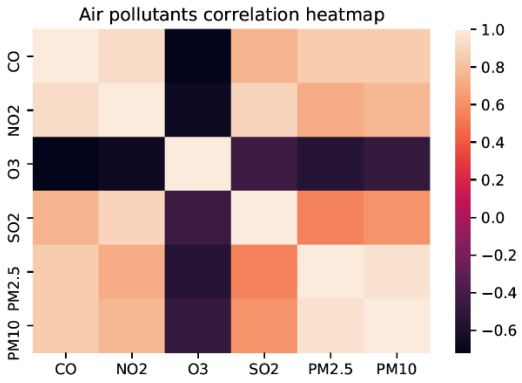


Fig.13 Air pollutant concentration correlation heatmap

To further analyze the inner relationship, a linear regression model is built here to describe the relationship between NO_2 and CO. The linear regression result is shown in Fig.14(a). The independent and dependent variables are CO and NO_2, respectively. The least squares method is used for regression. R-squared is calculated to be 0.863, indicating 86% of the variation in the output variable can be explained by the input variables. The linear regression model is NO_2 = -26.3 + 0.15*CO shown as the red line in Fig.14(b). A good agreement can be found between the model (red line) and data (blue dots).

(a)

OLS Regression Results

Dep. Variable:	NO2	R-squared:	0.863
Model:	OLS	Adj. R-squared:	0.863
Method:	Least Squares	F-statistic:	1.111e+04
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	0.00
Time:	20:13:51	Log-Likelihood:	-7320.3
No. Observations:	1761	AIC:	1.464e+04
Df Residuals:	1759	BIC:	1.466e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-26.3419	0.670	-39.292	0.000	-27.657	-25.027
CO	0.1549	0.001	105.402	0.000	0.152	0.158

Omnibus:	287.521	Durbin-Watson:	0.517
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2841.958
Skew:	-0.444	Prob(JB):	0.00
Kurtosis:	9.160	Cond. No.	830.

(b)

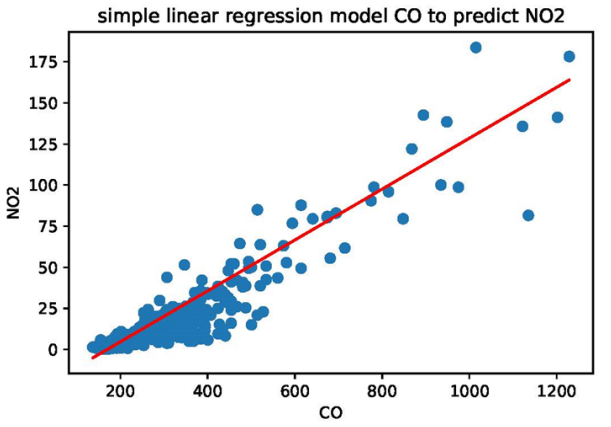


Fig.14 (a) OLS regression results with independent and dependent variable as CO and NO_2 respectively. (b) Simple linear regression model using CO to predict NO_2

Similar linear regression model is also built to predict PM10 with PM2.5. The linear regression result is shown in Fig.15(a). R-squared is calculated to be 0.911, indicating good agreement between model and data. The linear regression model is calculated to be PM10= 6.47 + 1.08*PM2.5, plotted as the red line in Fig.15(b). The model works better on input value larger than 20.

(a)

OLS Regression Results

Dep. Variable:	PM10	R-squared:	0.911
Model:	OLS	Adj. R-squared:	0.911
Method:	Least Squares	F-statistic:	1.802e+04
Date:	Sat, 26 Nov 2022	Prob (F-statistic):	0.00
Time:	20:14:02	Log-Likelihood:	-5549.1
No. Observations:	1761	AIC:	1.110e+04
Df Residuals:	1759	BIC:	1.111e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.4735	0.180	35.933	0.000	6.120	6.827
PM2.5	1.0791	0.008	134.226	0.000	1.063	1.095

Omnibus:	579.312	Durbin-Watson:	0.730
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2073.457
Skew:	1.602	Prob(JB):	0.00
Kurtosis:	7.242	Cond. No.	30.0

(b)

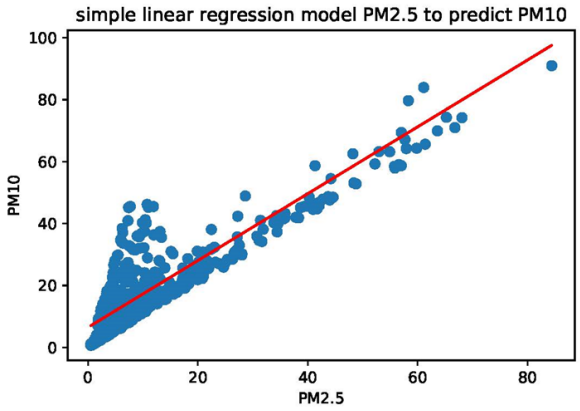


Fig.15 (a) OLS regression results with independent and dependent variable as PM2.5 and PM10 respectively. (b) Simple linear regression model using PM2.5 to predict PM10

(2) Correlation of air quality with both climate and population factors

The dependency of climate and population factors on air quality is investigated by calculating correlations. The correlation calculation result is shown in Fig.16. The absolute value of correlation ranges between 0.02 to 0.38. Humidity, temperature and population have a weak correlation with air quality (0.2<|correlation|<0.4). The wind degree has the correlation of 0.02, indicating very weak correlation (nearly no correlation) with air pollution.

	aqi	wind_degree	humidity	temp_c	population
aqi	1.000000	0.024362	-0.379538	0.218883	0.277160
wind_degree	0.024362	1.000000	0.068928	0.147862	0.038203
humidity	-0.379538	0.068928	1.000000	-0.701980	-0.237592
temp_c	0.218883	0.147862	-0.701980	1.000000	0.346221
population	0.277160	0.038203	-0.237592	0.346221	1.000000

Fig.16 Correlation of air quality with both climate and population factors

5. Conclusion and future work

In this project, air quality data, population data, and climate data in California are collected by API. To better understand the distribution, data is visualized in several methods including choropleth map plot, pair-plot, pie chart, etc. Inner relationships among air pollutant concentrations are analyzed and explained by regression models. Furthermore, the correlation of air quality with both climate and population factors are studied. The results shows that the influences of both climate and population factors on air quality are weak.

Some future works can be done to further improve this work:

1) Considering web scrapping time, this work only uses current day real-time climate and air quality data. In the future, data from a longer period (e.g., two weeks or one month) can be collected to better reflect the air quality and climate in one region.

2) More features can be included. In this work, to assess humanity activity, we use the population data. Other data, such as numbers of vehicle or factories can be included to better describe humane activities.

3) This work shows inner relationships exists between different pollutant concentrations. More data on pollutant sources (car exhaust, garbage burning, etc.) can be collected to further analyze the air pollution data.