Group: Team Alpha
Project: NYC House Sale Price Analysis
Member: Peiying Yu, Jianxing Wan, Ziyuan Wang

Guide Sheet of Project Code

Assumption:

1. We assume that our code files (‘da_project’) are zipped into Desktop and use relative path like ('~/Desktop/da_project/csv_file/data_midtown.csv') to implement our code.
2. Except code (CrimeComplaintHeatmap), it must use absolute path like ('/Users/jianxingwan/Desktop/da_project/zipcode.geojson') instead of relative path ('~/Desktop/da_project/zipcode.geojson') !

Code File Name	**Code Description**	**Csv File Used**	**Csv File Generated**
Zillow Link Scraper	Catch all links of 12 neighborhoods of sale house		link data (stored in ‘link_data’ folder)
zillow_scrap	Use all links of sale house to scrap our target features data	all csv files in ‘link_data’ folder under ‘csv_file’ folder	all csv files in ‘Scraped_data’ folder under ‘csv_file’ folder
WordCloud_midtown +LDA	Use midtown sale house overview to make WorkCloud and LDA	‘data_midtown.csv’ from ‘dataset.csv’ that has been processed	WordCloud graph and LDA topic display in code
Complaint_data_scrap	Use crime complaint data to calculate number of complaints in each zip code area	‘NYPD_Complaint_Data_Current__Year_To_Date_.csv’	‘data_total_complaint.csv’
CrimeComplaintHeatmap	Use the crime complaint data to show crime heat map based on different zip code area	‘data_total_complaint.csv’	crime complaint heat map in code
Subway Data + KNN	1. Use NYC transit station data to calculate distance (in km) of each home to its nearest subway station with Haversine Formula 2. Generate price correlation plot and group data by neighborhood 3. Use KNN regressor and classifier predict home price	‘NYC_Transit_Subway_Entrance_And_Exit_Data.csv’ ‘data_with_subway_distance.csv' ‘data for visual.csv’	‘data_with_subway_distance.csv'
WordCloud_10neighborhoods	Generate word clouds for 10 neighborhoods	‘datause.csv’ from ‘dataset.csv’ that has been processed	WordCloud graph display in code
random+linearRegression	1. Linear regression for price prediction 2. Random forest regressor and classifier for price prediction	‘mldata_binary.csv’ ‘zonedata.csv’	
Haversine.py	Implementation of Haversine formula, can be imported as a package		
Data Clean + Visualization	1. Merge all scraped info 2. Generate Heatmap for home price	all csv files in ‘Scraped_data’ folder under ‘csv_file’ folder	‘dataset.csv’ Price Heatmap

data clean			data for visual.csv