Data Pre processing

- Deleting duplicate records
- Removing bad data(dates exceeding now)
- Removing or unimportant columns(improvement_surcharge)

taxi-sample

VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount
2	11/04/2084 12:32:24 PM	11/04/2084 12:47:41 PM	1	1.34	1	N	238	236	2	10	0	0.5	0	0	0.3	10.8
2	11/04/2084 12:32:24 PM	11/04/2084 12:47:41 PM	1	1.34	1	N	238	236	2	10	0	0.5	0	0	0.3	10.8
2	11/04/2084 12:25:53 PM	11/04/2084 12:29:00 PM	1	0.32	1	N	238	238	2	4	0	0.5	0	0	0.3	4.8
2	11/04/2084 12:25:53 PM	11/04/2084 12:29:00 PM	1	0.32	1	N	238	238	2	4	0	0.5	0	0	0.3	4.8
2	11/04/2084 12:08:33 PM	11/04/2084 12:22:24 PM	1	1.85	1	N	236	238	2	10	0	0.5	0	0	0.3	10.8
2	11/04/2084 12:08:33 PM	11/04/2084 12:22:24 PM	1	1.85	1	N	236	238	2	10	0	0.5	0	0	0.3	10.8
2	11/04/2084 11:41:35 AM	11/04/2084 11:59:41 AM	1	1.65	1	N	68	237	2	12.5	0	0.5	0	0	0.3	13.3
2	11/04/2084 11:41:35 AM	11/04/2084 11:59:41 AM	1	1.65	1	N	68	237	2	12.5	0	0.5	0	0	0.3	13.3
2	11/04/2084 11:27:28 AM	11/04/2084 11:39:52 AM	1	1.07	1	N	170	68	2	9	0	0.5	0	0	0.3	9.8
2	11/04/2084 11:27:28 AM	11/04/2084 11:39:52 AM	1	1.07	1	N	170	68	2	9	0	0.5	0	0	0.3	9.8
2	11/04/2084 11:19:06 AM	11/04/2084 11:26:44 AM	1	1.3	1	N	107	170	2	7.5	0	0.5	0	0	0.3	8.3
2	11/04/2084 11:19:06 AM	11/04/2084 11:26:44 AM	1	1.3	1	N	107	170	2	7.5	0	0.5	0	0	0.3	8.3
2	11/04/2084 11:02:59 AM	11/04/2084 11:15:51 AM	1	1.85	1	N	113	137	2	10	0	0.5	0	0	0.3	10.8
2	11/04/2084 11:02:59 AM	11/04/2084 11:15:51 AM	1	1.85	1	N	113	137	2	10	0	0.5	0	0	0.3	10.8
2	11/04/2084 10:46:05 AM	11/04/2084 10:50:09 AM	1	0.62	1	N	231	231	2	4.5	0	0.5	0	0	0.3	5.3
2	11/04/2084 10:46:05 AM	11/04/2084 10:50:09 AM	1	0.62	1	N	231	231	2	4.5	0	0.5	0	0	0.3	5.3
2	07/11/2053 01:25:33 PM	07/11/2053 01:25:33 PM	1	0	1	N	264	264	2	0	0	0	0	0	0	0
2	12/04/2042 08:51:43 AM	12/04/2042 08:54:47 AM	1	0.29	1	N	162	162	2	4	0	0.5	0	0	0.3	4.8
2	12/04/2042 08:51:43 AM	12/04/2042 08:54:47 AM	1	0.29	1	N	162	162	2	4	0	0.5	0	0	0.3	4.8
2	06/25/2041 08:46:37 PM	06/25/2041 08:52:37 PM	1	1.34	1	N	239	151	2	7	0.5	0.5	0	0	0.3	8.3
2	11/17/2037 09:24:28 PM	11/17/2037 09:46:03 PM	1	2.99	1	N	170	143	1	15	0.5	0.5	1.7	0	0.3	18
2	11/17/2037 09:24:28 PM	11/17/2037 09:46:03 PM	1	2.99	1	N	170	143	1	15	0.5	0.5	1.7	0	0.3	18
2	02/02/2032 12:39:23 AM	02/02/2032 01:11:39 AM	4	23.21	1	N	132	228	2	62	0.5	0.5	0	0	0.3	63.3
2	02/13/2031 05:36:35 PM	02/13/2031 05:45:36 PM	1	1.44	1	N	236	237	2	8	1	0.5	0	0	0.3	9.8
2	02/13/2031 05:21:28 PM	02/13/2031 05:35:36 PM	1	1.69	1	N	141	236	2	8	1	0.5	0	0	0.3	9.8
2	05/06/2029 08:43:14 PM	05/06/2029 09:03:14 PM	4	4.47	1	N	162	80	1	17.5	0.5	0.5	4.91	5.76	0.3	29.47
2	05/05/2029 11:22:18 PM	05/06/2029 02:02:00 AM	1	11.51	1	N	148	244	1	34.5	0.5	0.5	0	0	0.3	35.8
2	02/13/2026 11:53:54 AM	02/13/2026 11:58:02 AM	2	0.85	1	N	161	43	2	5	1	0.5	0	0	0.3	6.8
2	02/13/2026 11:06:18 AM	02/13/2026 06:26:09 PM	2	3.14	1	N	163	246	2	20	1	0.5	0	0	0.3	21.8
2	09/13/2021 12:19:52 PM	09/13/2021 12:22:07 PM	1	0	1	N	193	193	2	0	0	0	0	0	0	0
2	12/10/2020 08:34:26 PM	12/10/2020 08:54:46 PM	1	4.62	1	N	50	231	2	17.5	0.5	0.5	0	0	0.3	18.8
2	12/10/2020 08:23:43 PM	12/10/2020 08:32:35 PM	1	2.44	1	N	90	50	1	9	0.5	0.5	2.06	0	0.3	12.36
2	08/01/2020 12:20:58 AM	08/01/2020 12:47:09 AM	1	16.71	1	N	143	138	1	45.5	0	0.5	10.41	5.76	0.3	62.47
2	08/01/2020 12:07:04 AM	08/01/2020 12:20:28 AM	1	2.3	1	N	238	143	2	11	0	0.5	0	0	0.3	11.8
2	03/05/2020 06:44:16 PM	03/06/2020 03:14:32 PM	1	2.39	1	N	125	161	2	11	0	0.5	0	0	0.3	11.8
2	03/05/2020 06:44:16 PM	03/06/2020 03:14:32 PM	1	2.39	1	N	125	161	2	11	0	0.5	0	0	0.3	11.8
2	03/05/2020 06:33:57 PM	03/05/2020 06:40:39 PM	1	1.04	1	N	231	125	1	6.5	0	0.5	1.46	0	0.3	8.76
2	03/05/2020 06:33:57 PM	03/05/2020 06:40:39 PM	1	1.04	1	N	231	125	1	6.5	0	0.5	1.46	0	0.3	8.76

Importing Data

We have imported data into three tables:

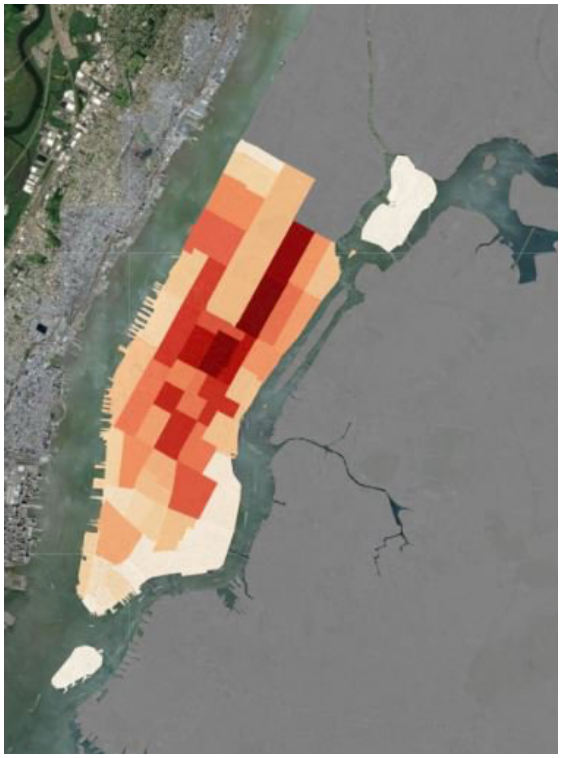
1. Trips (vendor id, total amount, …)
2. Zones (location id, borough text, zone text, …)
3. Payments (credit card, cash, no charge, …)
4. Ratecodes (standard rate, JFK, Newark,…)

```
import sqlite3, csv
from datetime import datetime
from zipfile import ZipFile

DATASET = {
    'taxi' : 'taxi-sample',
    'zones' : 'taxi+_zone_lookup.csv'
}

def setup_database():
    """
    Description: dataset pre processing
    """
    taxi_trips_file = DATASET.get('taxi')
    taxi_zf = ZipFile(taxi_trips_file+'.zip')

    # connect db
    conn = sqlite3.connect('assignment2.db')
    c = conn.cursor()

    # create tables
    # table trips
    c.execute(''' DROP TABLE IF EXISTS trips; ''')
    c.execute('''
    CREATE TABLE trips (
        VendorID int,
        tpep_pickup_datetime timestamp,
        tpep_dropoff_datetime timestamp,
        passenger_count int,
        trip_distance real,
        RatecodeID int,
        store_and_fwd_flag int,
        PULocationID int,
        DOLocationID int,
        payment_type int,
        fare_amount real,
        extra real,
        mta_tax real,
        tip_amount real,
        tolls_amount real,
        improvement_surcharge real,
        total_amount real
    );
    ''')

    # table zones : zones look up
    c.execute(''' DROP TABLE IF EXISTS zones; ''')
    c.execute('''
    CREATE TABLE zones (
        LocationID int PRIMARY KEY,
        Borough text,
        Zone text,
        service zone text
```

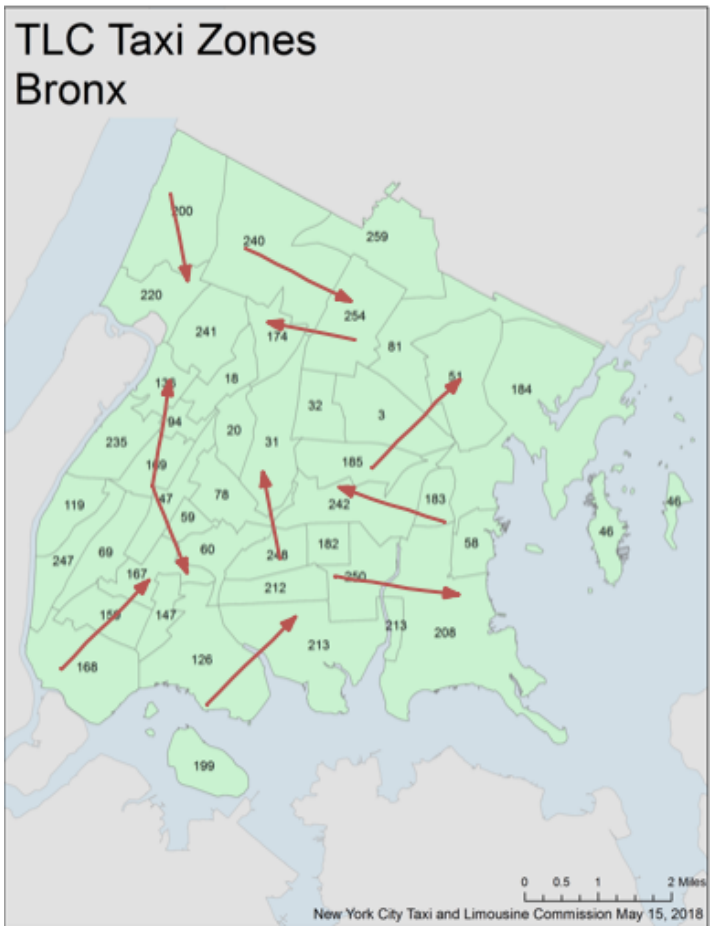
Future plans for visualizations

Three different types of heat map:

1. Start point
2. End point
3. Inter district travel



Outer district travels



Other visualizations

- Comparing average fares from each zone to another zone
- Comparing different trips according to their payment method
- Filtering out trips with high rates

Map shaper

Link: https://mapshaper.org

CSV file: Taxi Zone Shapefile(CSV)