Methods of Advanced Data Engineering

Prepared by Arefa Binti Sumaya (23187010)

Title: Analysis of Traffic Volume and Population Trends in New York City

The goal of this project is to combine population and traffic data for New York State to find trends, understand relationships, and offer useful insights for urban planning and transportation.

Question: How can we combine and evaluate population demographics and traffic data to guide urban development decisions?

Data Sources

Dataset-1 (Population Data)

This dataset provides population estimates by age and sex, helping to understand trends and their impact on traffic. It includes states and counties, making it useful for local urban planning.

- **Source:** The data comes from the U.S. Census Bureau and is available on StatsAmerica, a platform that organizes demographic and economic information.
- **Metadata URL:** `https://www.statsamerica.org/downloads/default.aspx#population`
- **Data URL:** `https://www.statsamerica.org/downloads/Population-by-Age-and-Sex.zip`

Population by Age and Sex - US, States, Counties

_Geo_ID	Statefips	Countyfips	Description	Year	Total Population	Population 0-4	Population 5-17	Population 18-24	Population 25-44	Population 45-64	Population 65+	Population Under 18	Population 18-54	Population 55+	Male Population	Female Po
0	0	0	U.S.	2000	282162411.0	19178293.0	53197896.0	27315274.0	84973340.0	62428040.0	35069568.0	72376189.0	150287588.0	59498634.0	138443407.0	143719004
0	0	0	U.S.	2001	284968955.0	19298217.0	53372958.0	27992652.0	84523274.0	64491563.0	35290291.0	72671175.0	151902194.0	60395586.0	139891492.0	145077463
0	0	0	U.S.	2002	287625193.0	19429192.0	53507265.0	28480708.0	83990295.0	66695526.0	35522207.0	72936457.0	152463197.0	62225539.0	141230559.0	146394634
0	0	0	U.S.	2003	290107933.0	19592446.0	53508312.0	28916746.0	83398001.0	68828899.0	35863529.0	73100758.0	153134701.0	63872474.0	142428897.0	147679036
0	0	0	U.S.	2004	292805298.0	19785885.0	53511850.0	29302179.0	83066831.0	70935234.0	36203319.0	73297735.0	153998940.0	65508623.0	143828012.0	148977286
0	0	0	U.S.	2005	295516599.0	19917400.0	53606269.0	29441546.0	82764185.0	73137401.0	36649798.0	73523669.0	154701635.0	67291295.0	145197078.0	150319521
0	0	0	U.S.	2006	298379912.0	19938883.0	53818831.0	29602839.0	82638980.0	75216272.0	37164107.0	73757714.0	155527978.0	69094220.0	146647265.0	151732647
0	0	0	U.S.	2007	301231207.0	20125962.0	53893443.0	29808025.0	82509693.0	77068373.0	37825711.0	74019405.0	156257657.0	70954145.0	148064854.0	153166353
0	0	0	U.S.	2008	304093966.0	20271127.0	53833475.0	30194274.0	82399959.0	78617510.0	38777621.0	74104602.0	157054680.0	72934684.0	149489951.0	154604015
0	0	0	U.S.	2009	306771529.0	20244518.0	53889649.0	30530346.0	82211153.0	80272688.0	39623175.0	74134167.0	157608587.0	75028775.0	150807454.0	155964075
0	0	0	U.S.	2010	309321666.0	20188815.0	53931851.0	30762380.0	82191286.0	81769110.0	40478224.0	74120666.0	157940058.0	77260942.0	152077478.0	157249665

Figure 1: Raw Population Dataset

Dataset-2 (Traffic Volume Data)

The traffic dataset shows detailed vehicle counts on New York City roads, helping to understand traffic patterns and their link to demographics for better planning.

- **Source:** The dataset is provided by the New York City Department of Transportation (NYC DOT) and is available on the NYC Open Data platform.
- **Metadata URL:** `https://catalog.data.gov/dataset/traffic-volume-counts`
- **Data URL:** `https://data.cityofnewyork.us/api/views/btm5-ppia/rows.csv`

ID	SegmentID	Roadway Name	From	To	Direction	Date	12:00-1:00 AM	1:00-2:00AM	2:00-3:00AM	3:00-4:00AM	4:00-5:00AM	5:00-6:00AM	6:00-7:00AM	7:00-8:00AM	8:00-9:00AM
1	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	NB	01/09/2012	20	10	11	14	13	20	34	66	100
2	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	NB	01/10/2012	21	16	8	6	13	13	31	70	67
3	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	NB	01/11/2012	27	14	6	5	12	16	34	75	69
4	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	NB	01/12/2012	22	7	7	8	11	12	33	75	89
5	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	NB	01/13/2012	31	17	7	5	13	28	29	68	84
6	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	NB	01/14/2012	42	27	21	18	21	13	17	18	46
7	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	SB	01/09/2012	27	12	12	4	22	27	66	154	155
8	15540	BEACH STREET	UNION PLACE	VAN DUZER STREET	SB	01/10/2012	26	16	11	13	16	27	59	156	177

Figure 2: Raw Traffic Volume Dataset

Structure

Both datasets are provided in CSV format, with population data including geographic identifiers, age groups, and sex distributions, and traffic volume data featuring roadway names, dates, time intervals, traffic volumes, and metadata on collection methods and context.

Quality Criteria for the Two Datasets

- **Accuracy:** The population data from the U.S. Census Bureau and the traffic data from NYC DOT are reliable and reflect real-world trends.
- **Completeness:** The population data includes detailed demographic fields but only up to 2019, while the traffic data has gaps in year-round sampling for some locations.
- **Consistency:** Both datasets are consistent after processing, with standardized formats for demographic fields and traffic intervals.
- **Timeliness:** The population data is outdated (up to 2019), and the traffic data, though periodically updated, may not reflect the latest trends.
- **Relevance:** Both datasets are relevant for analyzing population and traffic patterns but may require additional real-time data for current insights.

Licenses and Permissions

The population data is free to use from StatsAmerica, sourced from the U.S. Census Bureau under their open data policy, requiring no extra permissions. The traffic data, provided by the NYC DOT on the NYC Open Data platform, is also free to use under the NYC Open Data policy, provided the source is credited.

Data Pipeline

Technology Stack

Programming Language: Python; **Libraries:** Pandas, Requests, SQLite, Zipfile; **Storage:** SQLite database

Pipeline Steps

- **Data Collection:** Fetch population and traffic data from online sources.
- **Data Cleaning:** Remove inconsistencies and fill missing values in the data.
- **Data Transformation:** Group traffic data into broader time intervals for easier analysis.
- **Data Integration:** Combine population and traffic data for deeper insights.
- **Storage:** Save the cleaned and transformed data in an SQLite database for easy access and analysis.

Transformation and Cleaning Steps

The population data was processed by removing unnecessary metadata fields, such as `IBRC_Geo_ID`, and filtering the dataset to include only records for New York. For the traffic data, unused columns like `SegmentID` and `Direction` were dropped, hourly traffic data was grouped into broader intervals (e.g., 12:00 AM–4:00 AM), and column names were standardized to ensure consistency during processing.

Challenges and Solutions

- **Handling Missing Data:** Filled missing traffic intervals with zeros to make the data complete.
- **Data Format Inconsistencies:** Made column names consistent using Python string operations.
- **Large File Sizes:** Processed large files in chunks to use memory efficiently.

Error and Meta-Quality Measures

- **Error Handling:** Used try-except blocks to catch errors during file downloads and ZIP file extraction, allowing the pipeline to manage network issues or broken files without stopping.
- **Change Detection:** Used logging to track changes in the source data structure, ensuring the pipeline remains compatible with future updates.

Results and Limitations

Output Data of the Pipeline

- A cleaned dataset of New York population data, showing age and sex details.
- An aggregated traffic dataset with vehicle counts grouped into broader time intervals.

Both datasets are stored in an SQLite database for easy access and analysis. SQLite is efficient for managing structured data, supports queries, and is scalable for future analysis.

	Description *TEXT*	Year *INTEGER*	Total Population *REAL*	Population 0–4 *REAL*	Population 5–17 *REAL*	Population 18–24 *REAL*	Population 25–44 *REAL*
1	New York	2000	19001780.0	1237078.0	3450296.0	1770901.0	5823214.0
2	New York	2001	19082838.0	1222011.0	3450414.0	1802422.0	5775563.0
3	New York	2002	19137800.0	1212221.0	3440011.0	1817402.0	5719243.0
4	New York	2003	19175939.0	1204454.0	3415052.0	1837040.0	5659284.0
5	New York	2004	19171567.0	1197698.0	3376367.0	1847979.0	5583555.0
6	New York	2005	19132610.0	1180745.0	3333711.0	1855760.0	5492496.0
7	New York	2006	19104631.0	1160997.0	3296780.0	1869014.0	5417603.0
8	New York	2007	19132335.0	1157021.0	3253928.0	1898728.0	5357444.0
9	New York	2008	19212436.0	1159239.0	3212931.0	1936825.0	5324739.0
10	New York	2009	19307066.0	1157343.0	3185583.0	1968102.0	5290037.0
11	New York	2010	19399878.0	1159747.0	3160055.0	1983446.0	5269491.0

Figure 3: Modified Population Dataset

	roadway_name *TEXT*	date *TEXT*	400_800_am *REAL*	400_800_pm *REAL*	1200_400_am *REAL*	800_1200_pm *REAL*
1	BEACH STREET	01/09/2012	133.0	441.0	55.0	305.0
2	BEACH STREET	01/10/2012	127.0	432.0	51.0	236.0
3	BEACH STREET	01/11/2012	137.0	468.0	52.0	277.0
4	BEACH STREET	01/12/2012	131.0	464.0	44.0	285.0
5	BEACH STREET	01/13/2012	138.0	476.0	60.0	320.0
6	BEACH STREET	01/09/2012	269.0	549.0	55.0	522.0
7	BEACH STREET	01/10/2012	258.0	548.0	66.0	523.0
8	BEACH STREET	01/11/2012	273.0	585.0	65.0	596.0
9	BEACH STREET	01/12/2012	241.0	584.0	61.0	533.0

Figure 4: Modified Trafiic Dataset

Data Structure and Quality

- **Structure:** The output data is stored in SQLite tables with structured columns for easy access.
- **Quality:** The population data is complete and consistent, while the traffic data is cleaned, aggregated, and free of inconsistencies. Missing intervals were filled with zeros, and standardized formats ensure compatibility.

Critical Reflection and Potential Issues

The population data only covers up to 2019 and may not reflect current trends. Traffic data has gaps since not all locations are sampled year-round, and grouping time intervals may hide details. Combining the data could cause errors due to differences in location and time.