Group 5: Neighborhood Analysis

1.1 Business Topic

Our project revolves around the dynamic field of neighborhood analysis in the Boston area. The primary objective is to glean valuable insights from property data through meticulous analysis and innovative approaches.

1.2 Business-Related Questions

1.2.1 Property Valuation Trends

- How have property valuations evolved in different Boston neighborhoods over recent years?
- What are the key factors contributing significantly to changes in property values?

1.2.2 Geographical Distribution

- What is the distribution of property features across various Boston neighborhoods?
- Are there discernible patterns in the types of properties concentrated in specific areas?

1.3 Why This Topic and Questions?

1.3.1 Importance

Real estate analytics holds a pivotal role in guiding decision-making processes, urban planning initiatives, and sustainability endeavors. Our project seeks to provide invaluable insights for stakeholders in the real estate industry, policymakers, and advocates for sustainable practices.

1.3.2 Target Audience

- Real Estate Professionals: Our findings are tailored to benefit professionals in the real estate sector.
- Policymakers: The insights support urban planning and development strategies.
- Individuals looking to buy or rent properties in Boston.
- Companies or franchises planning to open offices or outlets in Boston.

1.3.3 Expected Outcome

The project aims to deliver actionable insights for property development, urban planning, and sustainability initiatives. These insights are intended to guide strategic decisions and provide valuable, data-driven recommendations.

1.4 What Makes This Project Unique?

Our project distinguishes itself by incorporating a survey and employing an innovative approach using the Extra Trees Regressor. This technique harnesses binary trees to impute missing data values, thereby enhancing the robustness of our analysis.

1.5 Results of Analysis

The anticipated results of our analysis are poised to benefit a diverse array of stakeholders, including real estate individuals seeking to buy or rent properties, professionals in the field, companies, franchises, policymakers, researchers, and sustainability advocates. These insights

have the potential to inform strategic decisions, contribute to city development, and promote environmentally friendly practices.

2. Data Sources and Datasets

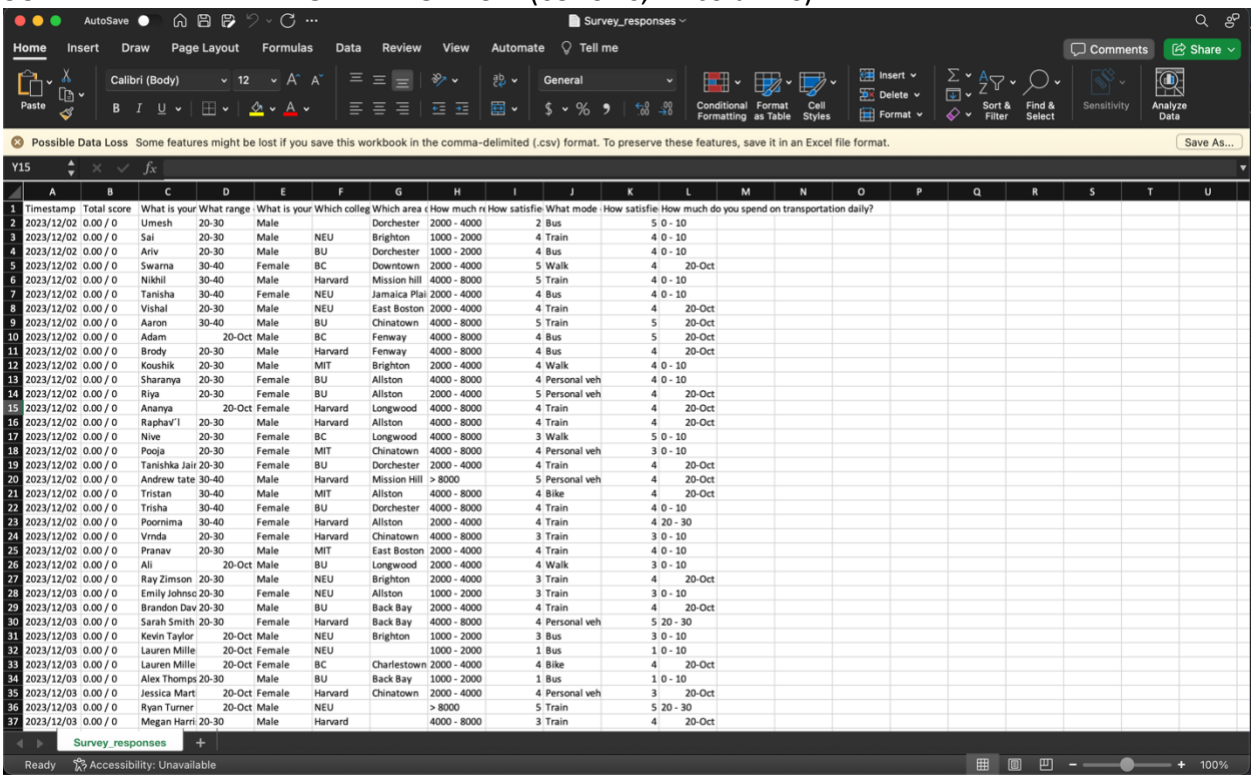
2.1 Survey Data (Neighborhood Analysis in the Boston Area)

Description: Our primary data is a survey conducted to extract valuable insights from individuals in the Boston area. It includes responses related to demographics, housing satisfaction, transportation preferences, and more.

Survey Questions:

1. Name
2. Age Range
3. Gender
4. College
5. Residence Area
6. Monthly Rent
7. Satisfaction with Accommodation Cost
8. Primary Transportation Mode
9. Satisfaction with Public Transportation
10. Daily Transportation Expenditure

SURVEY RAW DATA EXPORTED TO EXCEL (63 rows, 12 columns)



	A	B	C	D	E	F	G	H	I	J	K	L
1	Timestamp	Total score	What is your	What range	What is your	Which colleg	Which area	How much re	How satisfie	What mode	How satisfie	How much do you spend on transportation daily?
2	2023/12/02	0.00 / 0	Umesh	20-30	Male		Dorchester	2000 - 4000	2	Bus	5	0 - 10
3	2023/12/02	0.00 / 0	Sai	20-30	Male	NEU	Brighton	1000 - 2000	4	Train	4	0 - 10
4	2023/12/02	0.00 / 0	Ariv	20-30	Male	BU	Dorchester	1000 - 2000	4	Bus	4	0 - 10
5	2023/12/02	0.00 / 0	Swarna	30-40	Female	BC	Downtown	2000 - 4000	5	Walk	4	20-Oct
6	2023/12/02	0.00 / 0	Nikhil	30-40	Male	Harvard	Mission hill	4000 - 8000	5	Train	4	0 - 10
7	2023/12/02	0.00 / 0	Tanisha	30-40	Female	NEU	Jamaica Plai	2000 - 4000	4	Bus	4	0 - 10
8	2023/12/02	0.00 / 0	Vishal	20-30	Male	NEU	East Boston	2000 - 4000	4	Train	4	20-Oct
9	2023/12/02	0.00 / 0	Aaron	30-40	Male	BU	Chinatown	4000 - 8000	5	Train	5	20-Oct
10	2023/12/02	0.00 / 0	Adam	20-Oct	Male	BC	Fenway	4000 - 8000	4	Bus	5	20-Oct
11	2023/12/02	0.00 / 0	Brody	20-30	Male	Harvard	Fenway	4000 - 8000	4	Bus	4	20-Oct
12	2023/12/02	0.00 / 0	Koushik	20-30	Male	MIT	Brighton	2000 - 4000	4	Walk	4	0 - 10
13	2023/12/02	0.00 / 0	Sharanya	20-30	Female	BU	Allston	4000 - 8000	4	Personal veh	4	0 - 10
14	2023/12/02	0.00 / 0	Riya	20-30	Female	BU	Allston	2000 - 4000	5	Personal veh	4	20-Oct
15	2023/12/02	0.00 / 0	Ananya	20-Oct	Female	Harvard	Longwood	4000 - 8000	4	Train	4	20-Oct
16	2023/12/02	0.00 / 0	Raphav'l	20-30	Male	Harvard	Allston	4000 - 8000	4	Train	4	20-Oct
17	2023/12/02	0.00 / 0	Nive	20-30	Female	BC	Longwood	4000 - 8000	3	Walk	5	0 - 10
18	2023/12/02	0.00 / 0	Pooja	20-30	Female	MIT	Chinatown	4000 - 8000	4	Personal veh	3	0 - 10
19	2023/12/02	0.00 / 0	Tanishka Jair	20-30	Female	BU	Dorchester	2000 - 4000	4	Train	4	20-Oct
20	2023/12/02	0.00 / 0	Andrew tate	30-40	Male	Harvard	Mission Hill	> 8000	5	Personal veh	4	20-Oct
21	2023/12/02	0.00 / 0	Tristan	30-40	Male	MIT	Allston	4000 - 8000	4	Bike	4	20-Oct
22	2023/12/02	0.00 / 0	Trisha	30-40	Female	BU	Dorchester	4000 - 8000	4	Train	4	0 - 10
23	2023/12/02	0.00 / 0	Poornima	30-40	Female	Harvard	Allston	2000 - 4000	4	Train	4	20 - 30
24	2023/12/02	0.00 / 0	Vrnda	20-30	Female	Harvard	Chinatown	4000 - 8000	3	Train	3	0 - 10
25	2023/12/02	0.00 / 0	Pranav	20-30	Male	MIT	East Boston	2000 - 4000	4	Train	4	0 - 10
26	2023/12/02	0.00 / 0	Ali	20-Oct	Male	BU	Longwood	2000 - 4000	4	Walk	3	0 - 10
27	2023/12/02	0.00 / 0	Ray Zimson	20-30	Male	NEU	Brighton	2000 - 4000	3	Train	4	20-Oct
28	2023/12/03	0.00 / 0	Emily Johnso	20-30	Female	NEU	Allston	1000 - 2000	3	Train	3	0 - 10
29	2023/12/03	0.00 / 0	Brandon Dav	20-30	Male	BU	Back Bay	2000 - 4000	4	Train	4	20-Oct
30	2023/12/03	0.00 / 0	Sarah Smith	20-30	Female	Harvard	Back Bay	4000 - 8000	4	Personal veh	5	20 - 30
31	2023/12/03	0.00 / 0	Kevin Taylor	20-Oct	Male	NEU	Brighton	1000 - 2000	3	Bus	3	0 - 10
32	2023/12/03	0.00 / 0	Lauren Mille	20-Oct	Female	NEU		1000 - 2000	1	Bus	1	0 - 10
33	2023/12/03	0.00 / 0	Lauren Mille	20-Oct	Female	BC	Charlestown	2000 - 4000	4	Bike	4	20-Oct
34	2023/12/03	0.00 / 0	Alex Thomps	20-30	Male	BU	Back Bay	1000 - 2000	1	Bus	1	0 - 10
35	2023/12/03	0.00 / 0	Jessica Mart	20-Oct	Female	Harvard	Chinatown	2000 - 4000	4	Personal veh	3	20-Oct
36	2023/12/03	0.00 / 0	Ryan Turner	20-Oct	Male	NEU		> 8000	5	Train	5	20 - 30
37	2023/12/03	0.00 / 0	Megan Harri	20-30	Male	Harvard		4000 - 8000	3	Train	4	20-Oct

[Boston Neighborhood Boundaries approximated by 2020 Census Block Groups - Datasets - Analyze Boston]

2.2 Property Assessment Data (Harvard Dataverse)

Description: Our secondary dataset is sourced from Harvard Dataverse, containing property assessment records spanning from 2001 to 2021. It includes unique property identification numbers, assessed values, owner-occupancy status, and essential location data.

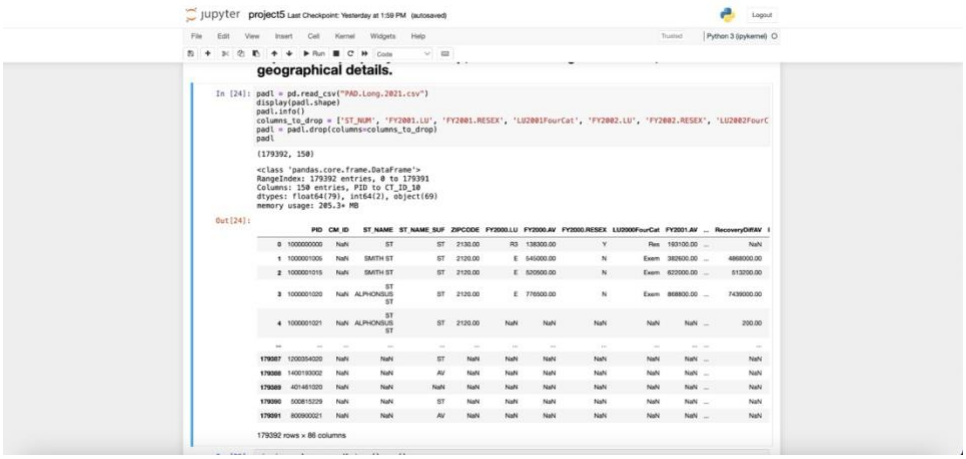
Critical Identifiers:

- Property Identification Number (PID)
- Street Number (ST_NUM)
- Street Name (ST_NAME)
- Zip Code (ZIPCODE)

Key Variables:

- Land Use type (FY(YYYY).LU)
- Assessed values (FY(YYYY).AV)
- Residential Exemption (FY(YYYY).RESEX)

[Property Assessment Data for Boston, MA v. 2021 - Boston Area Research Initiative's Boston Data Portal (harvard.edu)]



2.3 Crime Incident Reports (2023)

Description: Provided by the Boston Police Department, this dataset documents details surrounding incidents to which BPD officers respond. Records include incident type, occurrence time, and location.

[Crime Incident Reports (August 2015 - To Date) (Source: New System) - Datasets - Analyze Boston]

```
crime
```

Out[32]:

INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YI
232007915	1001	NaN	FORGERY / COUNTERFEITING	C6	200	0	2023-01-02 09:00:00+00	2
232092509	1107	NaN	FRAUD - IMPERSONATION	B2	920	0	2023-01-05 00:00:00+00	2
232001625	3115	NaN	INVESTIGATE PERSON	B2		0	2023-01-07 07:57:00+00	2
232000883	3115	NaN	INVESTIGATE PERSON	E18		0	2023-01-04 14:20:00+00	2
232003334	3115	NaN	INVESTIGATE PERSON	B2	265	0	2023-01-13 12:54:00+00	2
...	...	...	...	...	...	...	...	
232068532	1831	NaN	SICK ASSIST	D14	786	0	2023-08-26 13:40:00+00	2
232068926	1831	NaN	SICK ASSIST	B2	261	0	2023-08-27 21:25:00+00	2
232070395	301	NaN	ROBBERY	A1	93	0	2023-09-01 13:18:00+00	2
232071411	3207	NaN	PROPERTY - FOUND	C11	355	0	2023-09-04 21:00:00+00	2
232071422	3831	NaN	M/V - LEAVING SCENE - PROPERTY DAMAGE	D14	793	0	2023-09-04 21:42:00+00	2

rows × 18 columns

2.4 Fiscal Year 2023 (FY23) Dataset

Description: A dataset specific to Fiscal Year 2023, offering insights into recent developments in property values and land use in Boston.
Uniqueness: Not previously assessed by other researchers.
[Property Assessment - Datasets - Analyze Boston]

```
In [75]: prop23 = pd.read_csv("fy2023-property-assessment-data.csv")
display(prop23.shape)
prop23
```

(180627, 60)

Out[75]:

PID	CM_ID	GIS_ID	ST_NUM	ST_NAME	UNIT_NUM	CITY	ZIP_CODE	BLDG_SEQ	NUM_BLDGS	...	KITCHEN_STYLE2	KITCHEN_S
100001000	NaN	100001000	104.00	PUTNAM ST	NaN	EAST BOSTON	2128.00	1.00	1	...	S - Semi-Modern	S - Semi-
100002000	NaN	100002000	197.00	Lexington ST	NaN	EAST BOSTON	2128.00	1.00	1	...	M - Modern	M -
100003000	NaN	100003000	199.00	Lexington ST	NaN	EAST BOSTON	2128.00	1.00	1	...	S - Semi-Modern	S - Semi-
100004000	NaN	100004000	201.00	Lexington ST	NaN	EAST BOSTON	2128.00	1.00	1	...	S - Semi-Modern	S - Semi-
100005000	NaN	100005000	203.00	Lexington ST	NaN	EAST BOSTON	2128.00	1.00	1	...	S - Semi-Modern	
...	...	...	...	...	...	...	...	...	...	...	...	
2205666000	NaN	2205666000	NaN	KNOWLES ST	NaN	BRIGHTON	2135.00	1.00	1	...	NaN	
2205667000	NaN	2205667000	NaN	Lake ST	NaN	BRIGHTON	2135.00	1.00	1	...	NaN	
2205668000	NaN	2205668000	4.00	Lake ST	NaN	BRIGHTON	2135.00	1.00	1	...	M - Modern	M -
2205669000	NaN	2205669000	2193.00	COMMONWEALTH AV	NaN	BRIGHTON	2135.00	1.00	1	...	NaN	
2205670000	NaN	2205670000	2203.00	COMMONWEALTH AV	NaN	BRIGHTON	2135.00	1.00	3	...	NaN	

rows × 60 columns

2.5 Vision Zero Fatality Records

Description: Records of fatal traffic crashes in Boston related to the Vision Zero program, including date, time, location, and type of fatality.
Exclusions: Private property incidents, intentional assault, suicide, and driver medical emergencies.
[Vision Zero Fatality Records - Datasets - Analyze Boston]

```
In [28]: # https://data.boston.gov/dataset/vision-zero-fatality-records
         crashrecords = pd.read_csv('Vision Zero Fatality.csv')
         crashrecords
```

Out[28]:

	date_time	mode_type	location_type	street	xstreet1	xstreet2	x_cord	y_cord	long	lat
0	2018-07-25 15:13:52	ped	Intersection	NaN	L ST	E SIXTH ST	781815.50	2946769.14	-71.04	42.33
1	2017-01-01 06:01:43	ped	Street	SARATOGA ST	PUTNAM ST	BROOKS ST	782528.61	2963116.04	-71.03	42.38
2	2022-04-04 00:52:46	ped	Intersection	NaN	KNEELAND ST	HUDSON ST	775040.18	2952982.16	-71.06	42.35
3	2021-10-16 10:26:47	ped	Intersection	NaN	NEWMARKET SQ	THEODORE A GLYNN WAY	772833.21	2945873.09	-71.07	42.33
4	2015-12-03 18:18:40	ped	Intersection	NaN	COMMONWEALTH AVE	HARRY AGGANIS WAY	759284.14	2953366.16	-71.12	42.35
...	...	...	...	...	...	...	...	...	...	...
111	2020-07-25 17:34:28	mv	Intersection	NaN	A ST	W BROADWAY	776337.87	2949908.40	-71.06	42.34
112	2023-07-18 21:30:00	ped	Street	WOOD AVE	TACOMA ST	TINA AVE	761291.54	2923037.32	-71.11	42.27
113	2020-08-15 23:27:42	ped	Intersection	NaN	ASHMONT ST	DORCHESTER AVE	774061.77	2929462.64	-71.06	42.29
114	2022-10-04 20:57:38	ped	Street	SPRING ST	CENTRE ST	ALARIC ST	747691.33	2926586.89	-71.16	42.28
115	2020-06-09 07:33:54	bike	Intersection	NaN	CUMMINS HWY	RICHMERE RD	764251.85	2923846.12	-71.10	42.27

2.6 Blue Bike Stations

Description: Information about Blue Bike stations in Boston, providing a network of bike-sharing locations.

Data Update Frequency: Regular updates based on station changes.

[Blue Bike Stations - Datasets - Analyze Boston]

```
In [30]: bike_stations = pd.read_csv('Blue_Bike_Stations.csv')
         bike_stations
```

Out[30]:

	X	Y	Number	Name	Latitude	Longitude	District	Public_	Total_docks	ObjectId
0	-7908950.86	5221595.06	V32016	Chelsea St at Vine St	42.40	-71.05	Everett	Yes	11	1
1	-7916449.25	5212671.51	K32015	1200 Beacon St	42.34	-71.11	Brookline	Yes	1	2
2	-7908122.35	5220539.21	H32005	Chelsea Station	42.40	-71.04	Chelsea	Yes	11	3
3	-7911699.40	5216863.65	M32049	Child Street at Brian P. Murphy Staircase	42.37	-71.07	Cambridge	Yes	23	4
4	-7923242.00	5215761.84	W32006	160 Arsenal	42.36	-71.18	Watertown	Yes	11	5
...	...	...	...	...	...	...	...	...	...	...
461	-7928872.66	5213492.62	N32005	West Newton	42.35	-71.23	Newton	Yes	15	462
462	-7919706.18	5215328.45	A32043	Western Ave at Richardson St	42.36	-71.14	Boston	Yes	19	463
463	-7913946.24	5210971.84	B32059	Whittier St Health Center	42.33	-71.09	Boston	Yes	19	464
464	-7915669.50	5207008.68	D32040	Williams St at Washington St	42.31	-71.11	Boston	Yes	23	465
465	-7916387.78	5218928.03	S32005	Wilson Square	42.39	-71.11	Somerville	Yes	15	466

466 rows × 10 columns

Information Quality

3.1 Boston Area Research Initiative (BARI) Property Assessment Database (padl)

Concerns:

1. Missing Data: The dataset contained missing values, particularly in variables such as FY(YYYY).AV and FY(YYYY).RESEX.
2. Data Consistency: Inconsistencies in land use codes and residential exemption values were observed over the years.

Addressing Concerns:

1. Imputation: Missing values were addressed through imputation techniques, such as mean or median imputation, ensuring a more complete dataset.
2. Code Standardization: Land use codes were standardized using the provided code mapping, and inconsistent values in residential exemption were corrected based on historical data patterns.
3. PID Deduplication: Identified and removed duplicate records based on the 'PID' (Property Identification Number) to ensure each property is represented only once in the dataset.

3.2 Proposition 23 (Prop23) Property Data

Concerns:

1. Missing Values: The 'prop23' dataset had missing values which were imputed later.

Addressing Concerns:

1. Threshold-based Filtering: Columns with more than 20% missing values were identified and dropped from the 'prop23' dataset. This process ensured that the dataset used for analysis had a more manageable and consistent set of features.
2. PID Deduplication: Identified and removed duplicate records based on the 'PID' (Property Identification Number) to ensure each property is represented only once in the dataset.

3.3 Boston Neighborhoods Shapefile

Concerns:

Geospatial Integrity: This data is decided, recorded and supplied by The City of Boston and hence it can be trusted.

3.4 Survey Data

Concerns:

1. Data Completeness: As survey data may have been self-reported, there could be instances of incomplete or inconsistent responses.

Addressing Concerns:

1. Data Cleaning: Rigorous data cleaning processes were implemented to identify and handle incomplete or inconsistent responses.

Conclusion

Addressing information quality concerns was a critical step in ensuring the reliability and validity of the analysis. Imputation, standardization, deduplication, and verification processes were implemented to enhance the overall quality of the datasets used in the project. The transparency of these processes contributes to the credibility of the findings and recommendations derived from the data.

4.1 Data Cleaning and Manipulation

The data cleaning and manipulation tasks were accomplished using the following Python packages:

- **pandas**: Used for handling and manipulating tabular data.
- **geopandas**: Applied for working with geospatial data, enabling the creation of GeoDataFrames and spatial operations.
- **matplotlib**: Employed for data visualization, particularly in creating plots and heatmaps.
- **lazypredict**: Utilized for a quick evaluation of various regression models to find the most suitable model for imputing missing values.
- **scikit-learn**: Employed for preprocessing, imputation, and machine learning tasks.
- **shapely**: Used for geospatial operations and the creation of geometric objects.

4.2 Missing Values Imputation

To handle missing values, you employed the following methods:

- Used LazyRegressor to evaluate and select the best-performing regression model for imputing missing values.
- Developed a custom function utilizing ExtraTreesRegressor for imputing missing values based on location ('X', 'Y') and other relevant columns like 'ZIPCODE' and 'PID'.

4.3 Geospatial Analysis

To perform geospatial analysis, you utilized GeoPandas and implemented the following steps:

- Created a GeoDataFrame to represent the spatial features of the properties.
- Determined the neighborhood each property belongs to by spatially assigning it based on its 'X' and 'Y' coordinates.
- Grouped the data by neighborhood for further analysis.

4.4 Neighborhood-Level Analysis

For neighborhood-level analysis, you executed the following tasks:

- Grouped the data by neighborhood and calculated various statistics, such as means and counts, for relevant columns.
- Merged the grouped data with neighborhood geometries for spatial representation.
- Utilized matplotlib to create geographical heatmaps for different columns.

4.5 Challenges and Solutions

During the data wrangling process, challenges were encountered, such as handling missing values and spatial assignment. These challenges were addressed by employing suitable imputation techniques, leveraging regression models for imputation, and using GeoPandas for geospatial analysis.

Conclusion

The methods employed in this analysis aimed to ensure accurate and insightful results for neighborhood-level analysis and property-related insights.

This subsection provides an overview of the methods, tools, and challenges faced during the data wrangling process for your final project. Adjust the details based on the specific nuances of your analysis.

Data Wrangling Process

Executed Data Wrangling Steps:

1. Read and cleaned and standardized survey data to prepare for analysis.
2. **Handling Missing Values:**
    - Utilized the **impute_missing_values** function to address missing values in specific columns such as 'FY2000.AV', 'FY2021.AV', etc which had missing data.
    - Investigated and imputed missing values for relevant columns crucial for the analysis.
3. **Geospatial Processing:**

- Created a GeoDataFrame with geometries from latitude and longitude information.
- Assigned neighborhoods to each property observation using spatial relationships with neighborhood polygons.
- Ensured the integrity of the neighborhood assignments and handled cases where geometries were missing.

4. **Quick Modeling with LazyRegressor:**
   Utilized **LazyRegressor** to quickly assess the performance of various regression models on the dataset.
   Explored different regression algorithms to identify potential candidates for further refinement.
5. **Feature Importance with ExtraTreesRegressor:**
   Employed **ExtraTreesRegressor** to impute missing values using 'X', 'Y', 'PID' and zipcodes
6. **Grouping and Aggregation:**
   - Grouped the data by neighborhood to calculate averages, min-max ranges, and counts for various columns for neighborhood level analysis.
   - Created a comprehensive DataFrame, merging information on property values and characteristics, demographics, crime incidents, and bike station counts at the neighborhood level.

Validation Rules and Checks:

1. **Missing Value Validation:**
   - Conducted a thorough check on missing values after imputation to ensure completeness.
   - Verified that essential columns for analysis were imputed appropriately.
2. **Geospatial Validation:**
   - Verified the accuracy of the neighborhood assignments by comparing a sample of assigned neighborhoods with expected values.
   - Checked for missing geometries and implemented appropriate handling mechanisms.
3. **Grouping Validation:**
   - Validated the groupby and aggregation process by inspecting sample neighborhoods and confirming the correctness of calculated statistics.

Final Data Wrangling Process:

The final data wrangling process involved integrating information from diverse datasets, addressing missing values, assigning neighborhoods, aggregating data at the neighborhood level, and gaining preliminary insights through quick modeling and feature importance analysis.

Analysis and Results

Conducted Analysis:

1. **Property Value Trends:**
   - Explored trends in property values and characteristics, including the average assessed values and percentage changes over time.

- Investigated factors such as growth, crashes, and recoveries in property values.

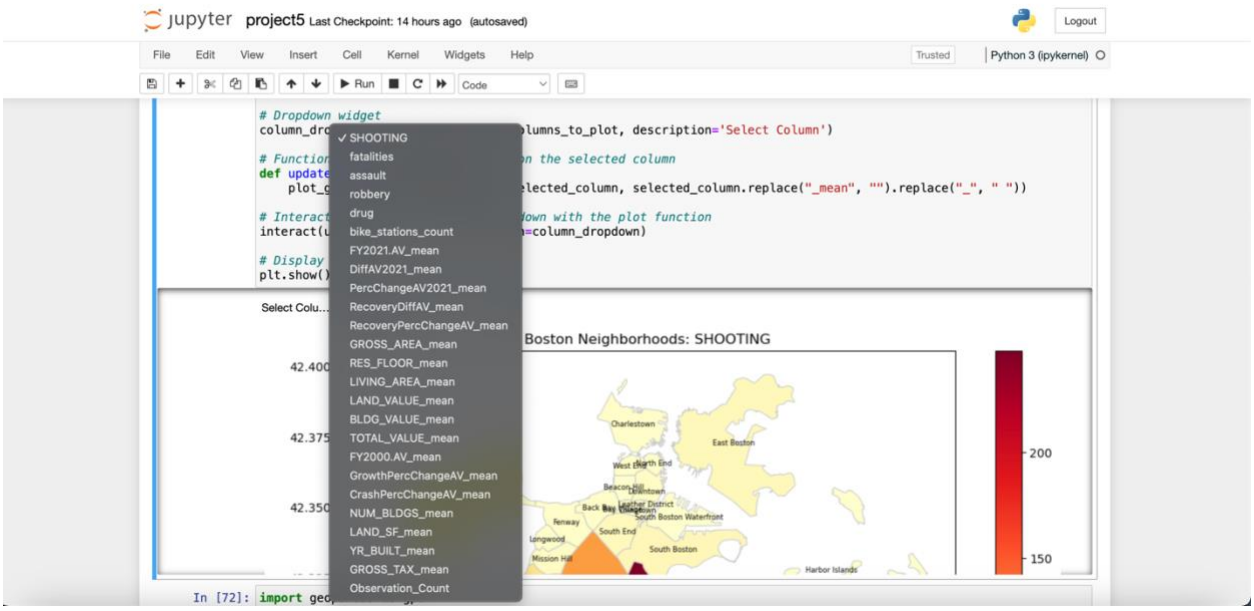
2. **Analyzed Survey Data** to understand rent values,
3. **Neighborhood Comparison:**
    - Compared neighborhoods based on various metrics, including road accident fatalities, demographics, mode of transportation and crime incidents.
    - Conducted statistical analysis to identify patterns and variations.

Structure of the Outcome:

The outcome of the analysis is a comprehensive dataset that allows for a detailed understanding of neighborhood characteristics, property value dynamics, and associated factors. The structured dataset facilitates easy comparison and identification of key insights.

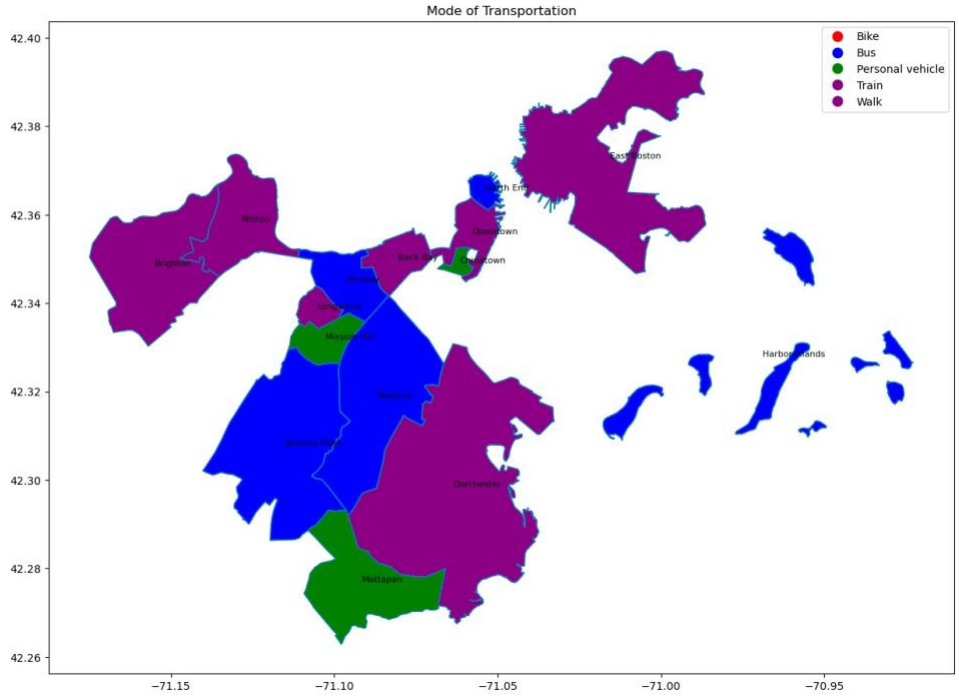
Visualizations:

1. **Dropdown menu:** Created a dropdown menu to show plots for whichever parameter selected using ipywidgets.



2. **Geographical Heatmap using :**
    - Created a geographical heatmap showcasing neighborhood-specific metrics such as crime incidents, property values, or demographic characteristics.
3. **Mode of Transportation Analysis:**
    - Developed a visualization illustrating the distribution of different transportation modes across neighborhoods, providing insights into commuting patterns.

These visualizations offer a clear representation of the analyzed data and serve as illustrative examples of the derived insights.



Contributions

Apurv - Data Collection

Apurv played a crucial role in gathering diverse datasets for our analysis. The primary dataset involves a comprehensive survey, capturing valuable insights directly from the community. Additionally, he utilized secondary datasets to enrich our analysis:

Primary Dataset:
Type: Survey
Secondary Datasets:
a. Property Assessment Data 2000 - 2021 (Harvard Dataverse)
b. Property Assessment (FY23) Dataset (ANALYZE BOSTON)
c. Crime Incident Reports (2023) (ANALYZE BOSTON)
d. Vision Zero Fatality Records (ANALYZE BOSTON)
e. Blue Bike Stations (ANALYZE BOSTON)

Umesh - Data Preprocessing

Umesh focused on preparing the collected data for analysis, ensuring its quality and consistency. His contributions include:
Data Formatting: Normalized irregular names and values. Corrected entries with errors.
Data Cleaning: Merged necessary columns for better clarity. Mitigated discrepancies in the data.
Data Enrichment: Renamed columns for uniformity. Normalized names to a common value (e.g., NEU for Northeastern University).

Aashay - Data Imputation Techniques
Aashay's role was pivotal in handling missing data and enhancing the dataset's completeness. His contributions encompassed:
Imputation Techniques: Utilized Lazy Regressor and ExtraTress Regressor for data imputation.
Employed Geopandas and Shapely for spatial data enhancement.
Visualization: Developed drop-down visualizations for easy interpretation.

Ariv - Data Profiling
Ariv focused on comprehensively understanding the datasets through profiling techniques. His contributions include:
Sweetviz:
Employed Sweetviz for detailed statistical analysis and visualization.
YData Profiling:
Conducted thorough profiling using YData to extract meaningful insights.

References:

Crime Report

Property Assessment Data

Property Assessment FY23

Blue bikes

Vision Zero Fatality

Boston Neighborhoods Shape File

Survey Data