**1.1 Business Topic**

Our project revolves around the dynamic field of real estate analytics 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?

**1.2.3 Utility Bill Analysis**

* How does utility bill data correlate with property characteristics?
* What opportunities exist for improving energy efficiency based on utility consumption patterns?

**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.

**Top of Form**

**Bottom of Form**

**Top of Form**

**Bottom of Form**

**3. 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 columns with more than 20% missing values.

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. Outliers were scrutinized, and efforts were made to contact respondents for clarification.

**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.