Data Assessment Document

Project Name: Housing Crisis: Fact or Fiction?

Data Source:

Data Sources	Data Collection	Carried out by,
House & Land-selling dataset	https://www.statcan.gc.ca/	Deep Sheta on 6th Oct 2023
Population Dataset	https://www.statcan.gc.ca/	Kaveri Patel on 4th Oct 2023
Migrants Dataset	https://www.statcan.gc.ca/	Kaveri Patel on 7th Oct 2023
Mortgage Dataset	https://www.statcan.gc.ca/	Deep Sheta on 8th Oct 2023
Interest Rate Dataset	https://www.statcan.gc.ca/	Jatin Kansra on 6th Oct 2023
Constructions Dataset	https://www.statcan.gc.ca/	Tirthak Bhingaradiya on 7th Oct 2023

Initial Data Overview:

- **House & Land-selling dataset**
  - This data assessment was performed by **Deep**.
  - This dataset contains information about housing and land sales, including property types, prices, locations, and transaction dates. It can be used to assess trends in the real estate market, property values, and demand for housing.
  - Data contains:
    - **REF_DATE:** Filter dataset to last 2 or 3 years.
    - **REF_DATE:** We will Group data quarterly like Q1 2021, q2 2021 and so on.
    - **GEO:** We will assign each province a unique number to keep it consistent in all the reference datasets. For example, Ontario: 1, Alberta: 2 and so on.
    - **Key Columns:**
      - **REF_DATE**
      - **GEO**
      - New housing price indexes
      - VALUE

- **Population Dataset**
  - This data assessment was performed by **Kaveri**.

- Population data can help in understanding the growth or decline of the local population. It's relevant for evaluating whether housing demand aligns with population changes, which could indicate a housing crisis or a surplus.
- Data contains:
  - **REF_DATE:** Filter dataset to last 2 or 3 years.
  - **REF_DATE:** We will Group data quarterly like Q1 2021, q2 2021 and so on.
  - **GEO:** We will assign each province a unique number to keep it consistent in all the reference datasets. For example, Ontario: 1, Alberta: 2 and so on.
  - **Key Columns:**
    - **REF_DATE**
    - **GEO**
    - VALUE

- **Migrants Dataset**
  - This data assessment was performed by **Kaveri**.
  - Information on migrants can be vital to the study of housing. It can provide insights into how immigration or emigration trends impact housing demand and availability.
  - Data contains:
    - **REF_DATE:** Filter dataset to last 2 or 3 years.
    - **REF_DATE:** We will Group data quarterly like Q1 2021, q2 2021 and so on.
    - **GEO:** We will assign each province a unique number to keep it consistent in all the reference datasets. For example, Ontario: 1, Alberta: 2 and so on.
    - **Key Columns:**
      - **REF_DATE**
      - **GEO**
      - VALUE

- **Mortgage Dataset**
  - This data assessment was performed by **Jatin**.
  - Mortgage data includes details on home loans, and mortgage types. This data can offer insights into the affordability of housing and the influence of interest rates on home purchases.
  - Data contains:
    - **REF_DATE:** Filter dataset to last 2 or 3 years.
    - **REF_DATE:** We will Group data quarterly like Q1 2021, q2 2021 and so on.
    - **GEO:** We will assign each province a unique number to keep it consistent in all the reference datasets. For example, Ontario: 1, Alberta: 2 and so on.
    - **Key Columns:**
      - **REF_DATE**
      - **GEO**
      - Type of unit
    - Type of market
    - VALUE

- **Interest Rate Dataset**
    - This data assessment was performed by **Deep**.
    - This dataset is crucial for understanding the financial aspect of housing. Changes in interest rates can significantly impact mortgage affordability and housing market dynamics.
    - Data contains:
        - **REF_DATE:** Filter dataset to last 2 or 3 years.
        - **REF_DATE:** We will Group data quarterly like Q1 2021, q2 2021 and so on.
        - **GEO:** We will assign each province a unique number to keep it consistent in all the reference datasets. For example, Ontario: 1, Alberta: 2 and so on.
        - **Key Columns:**
            - **REF_DATE**
            - **GEO**
            - VALUE

- **Constructions Dataset**
    - This data assessment was performed by **Tirthak**.
    - Data related to construction permits, housing starts, and building completions can help assess the supply side of the housing market. It can offer insights into whether housing shortages or surpluses exist and if new construction is meeting demand.
    - Data contains:
        - **REF_DATE:** Filter dataset to last 2 or 3 years.
        - **REF_DATE:** We will Group data quarterly like Q1 2021, q2 2021 and so on.
        - **GEO:** We will assign each province a unique number to keep it consistent in all the reference datasets. For example, Ontario: 1, Alberta: 2 and so on.
        - **Key Columns:**
            - **REF_DATE**
            - **GEO**
            - Building type
            - Construction type
            - VALUE

Dataset Feature:

Dependent and independent variables for each dataset:

1. **House & Land-selling Dataset:**

   Dependent Variable: Housing price indexes and VALUE

   Independent Variables: REF_DATE (time), GEO (geographical location), property types, transaction dates

2. **Population Dataset:**

   Dependent Variable: VALUE (population)

   Independent Variables: REF_DATE (time), GEO (geographical location)

3. **Migrants Dataset:**

   Dependent Variable: VALUE (number of migrants)

   Independent Variables: REF_DATE (time), GEO (geographical location)

4. **Mortgage Dataset:**

   Dependent Variable: VALUE (mortgage data, e.g., loan amounts)

   Independent Variables: REF_DATE (time), GEO (geographical location), Type of unit, Type of market

5. **Interest Rate Dataset:**

   Dependent Variable: VALUE (interest rates)

   Independent Variables: REF_DATE (time), GEO (geographical location)

6. **Constructions Dataset:**

   Dependent Variable: VALUE (construction-related data)

   Independent Variables: REF_DATE (time), GEO (geographical location), Building type, Construction type.

Ethical Assessment:

- We will make sure that all personal information is anonymized or removed from the datasets to protect individuals' privacy.
- We will be transparent about data sources, data preprocessing, and the methods used in the analysis.
- We will clearly define data ownership and permissible uses of the data. Ensure that data is used in accordance with legal and ethical standards.
- We will not share any data on open forums.
- We will try our best when presenting the findings, report them accurately and responsibly, to avoid sensationalism and misrepresentation.

Next Steps:

The next steps in the project "Housing Crisis: Fact or Fiction?" involve data preparation and preprocessing to ensure that the datasets are ready for analysis.

These crucial steps include data cleanup, filtering, and preprocessing.

Proper data cleanup, filtering, and preprocessing are essential for accurate and meaningful analysis of the housing crisis and its contributing factors.

1. Data Cleanup:

- Start by examining the datasets for any missing or inconsistent data. Address missing values by imputing them or removing incomplete records, depending on the impact of missing data on the analysis.
- Check for outliers, erroneous entries, and duplicates. Outliers should be identified and, if necessary, addressed appropriately.
- Ensure that data types are consistent and formatted correctly. For example, dates should be in a uniform date format, and numerical values should be in a consistent format.
- Correct any data errors or inconsistencies that could affect the quality and accuracy of the analysis.

2. Data Filtering:

- As indicated in your project description, filter the data to focus on the relevant time period, which may be the last 2 or 3 years.
- If applicable, group the data into quarters (e.g., Q1 2021, Q2 2021) as specified to facilitate time-based analysis.
- Verify that the data aligns with the unique numerical codes assigned to geographical locations (e.g., Ontario: 1, Alberta: 2) to ensure consistency across reference datasets.

3. Preprocessing Data:

- We will handle categorical data by encoding it into numerical format using techniques like one-hot encoding or label encoding, where necessary.
- We will try to normalize or standardize numerical features to bring them to a common scale.
- We will conduct exploratory data analysis (EDA) to gain insights into the data distribution, correlations, and potential trends.