Part II: Data Collection and Cleaning Report

Your Name

1 Data Sources and Collection Methods

For this study on the impact of remote work on urban housing prices, we have collected data from the following sources:

1. Remote Work Adoption: U.S. Census Bureau's American Community Survey (ACS)
    - Collection method: Annual survey data
    - Potential bias: Self-reporting bias, under-representation of certain demographics
2. Housing Prices: Zillow Home Value Index (ZHVI)
    - Collection method: Proprietary algorithm based on MLS data and public records
    - Potential bias: Overrepresentation of listed properties, potential regional biases
3. Urban Characteristics: U.S. Census Bureau's City and Town Population Totals
    - Collection method: Decennial census and intercensal estimates
    - Potential bias: Undercounting of certain populations, particularly in urban areas
4. Employment Data: Bureau of Labor Statistics
    - Collection method: Monthly surveys and administrative records
    - Potential bias: Potential undercounting of gig economy workers

2 Data Overview and Preprocessing Steps

Our dataset covers 100 major U.S. cities from 2019 to 2023. Key variables include:

- Remote work adoption rate (%)
- Median home value ($)
- Population
- Unemployment rate (%)
- Median household income ($)

Preprocessing steps included:

1. Merging datasets using city and year as keys
2. Handling missing values:
    - Imputed missing values for remote work adoption using k-nearest neighbors
    - Removed cities with >10% missing data across all variables
3. Normalizing variables:
    - Log-transformed home values and income
    - Standardized population to z-scores
4. Creating new variables:
    - Year-over-year change in home values
    - Remote work adoption change from 2019 baseline

3 Issues Encountered and Solutions

1. Inconsistent city definitions across datasets
    - Solution: Used Core-Based Statistical Areas (CBSAs) as consistent geographic units
2. Outliers in home value data
    - Solution: Winsorized top and bottom 1% of values
3. Temporal misalignment of ACS and monthly BLS data
    - Solution: Aggregated monthly BLS data to annual level, matching ACS timeframe

4 Preliminary Analysis

Descriptive statistics for key variables (2023 data):

Variable	Mean	Std Dev	Min	Max
Remote Work Adoption (%)	22.3	7.5	8.2	41.6
Median Home Value ($)	342,500	189,700	125,000	1,250,000
Population	652,000	1,245,000	100,000	8,500,000
Unemployment Rate (%)	5.2	1.8	2.1	11.3

Preliminary visualizations:

1. Scatter plot of remote work adoption vs. home value change (2019-2023)
2. Time series of average home values for high vs. low remote work cities
3. Choropleth map of remote work adoption rates across U.S. cities

5 Data Dictionary

A comprehensive data dictionary is available in our GitHub repository: https://github.com/yourusername/remote-work-housing-prices/data/data_dictionary.md

This data dictionary includes detailed descriptions of all variables, their units, sources, and any transformations applied during preprocessing.