PANAMA CITY RESIDENTIAL: DATA ENRICHMENT,CLEANING AND INSIGHTS

OUR TEAM

Rishika Lall

Vaishali Patidhar

Mirudula Muralidhar

Govardhan Gattu

Manish Bhatia

INTRODUCTION

01 Our Team

02 Steps taken

03 Data Enrichment

04 Data Cleaning

05 Challenges

06 Next Steps

STEPS TAKEN

- Visualization
- Data Cleaning
- Data Enrichment
- Identifying features

DATA ENRICHMENT

01

We had a total of 88,000 records in the address fabrics dataset.

02

This data extracts insights from geodemographics and property datasets using Python.

03

Data was divided into segments of 20,000 rows and assigned to each person to extract the data.

04

We extracted elevation and living square footage from the property dataset, and property type, household income, and property tenure from the demographics dataset.

DATA ENRICHMENT QUERIES

```python
def generate_psyteGeodemographics_query(self, precisely_id):
    # GraphQL query for psyteGeodemographics (same as before)
    return f"""
    query addressByPreciselyID {{
      getById(id: "{precisely_id}", queryType: PRECISELY_ID) {{
        addresses {{
          data {{
            psyteGeodemographics {{
              data {{
                PSYTECategoryCode
                PSYTEGroupCode
                PSYTESegmentCode {{
                  description
                }}
                censusBlock
                censusBlockGroup
                censusBlockPopulation
                censusBlockHouseholds
                householdIncomeVariable {{
                  value
                  description
                }}
                propertyValueVariable {{
                  value
                  description
                }}
                propertyTenureVariable {{
                  value
                  description
                }}
                propertyTypeVariable {{
                  value
                  description
                }}
                urbanRuralVariable {{
                  value
                  description
                }}
              }}
            }}
          }}
        }}
      }}
    }}
    """
```

```python
def generate_coastalRisk_query(self, precisely_id):
    # GraphQL query for coastalRisk (same as before)
    return f"""
    query coastalRisk {{
        getById(id: "{precisely_id}", queryType: PRECISELY_ID) {{
            addresses {{
                data {{
                    coastalRisk {{
                        data {{
                            preciselyID
                            waterbodyName
                            nearestWaterbodyCounty
                            nearestWaterbodyState
                            nearestWaterbodyType {{
                                value
                                description
                            }}
                            nearestWaterbodyAdjacentName
                            nearestWaterbodyAdjacentType
                            distanceToNearestCoastFeet
                            windpoolDescription
                        }}
                    }}
                }}
            }}
        }}
    }}
    """
```

```python
def generate_floodRisk_query(self, precisely_id):
    # GraphQL query for floodRisk (same as before)
    return f"""
    query floodRisk {{
      getById(id: "{precisely_id}", queryType: PRECISELY_ID) {{
        addresses {{
          data {{
            floodRisk {{
              data {{
                preciselyID
                floodID
                femaMapPanelIdentifier
                floodZoneMapType
                stateFIPS
                floodZoneBaseFloodElevationFeet
                floodZone
                additionalInformation
                baseFloodElevationFeet
                communityNumber
                communityStatus
                mapEffectiveDate
                letterOfMapRevisionDate
                letterOfMapRevisionCaseNumber
                floodHazardBoundaryMapInitialDate
                floodInsuranceRateMapInitialDate
                addressLocationElevationFeet
                year100FloodZoneDistanceFeet
                year500FloodZoneDistanceFeet
                elevationProfileToClosestWaterbodyFeet
                distanceToNearestWaterbodyFeet
                nameOfNearestWaterbody
              }}
            }}
          }}
        }}
      }}
    }}
    """
```

PROPERTY DATASET

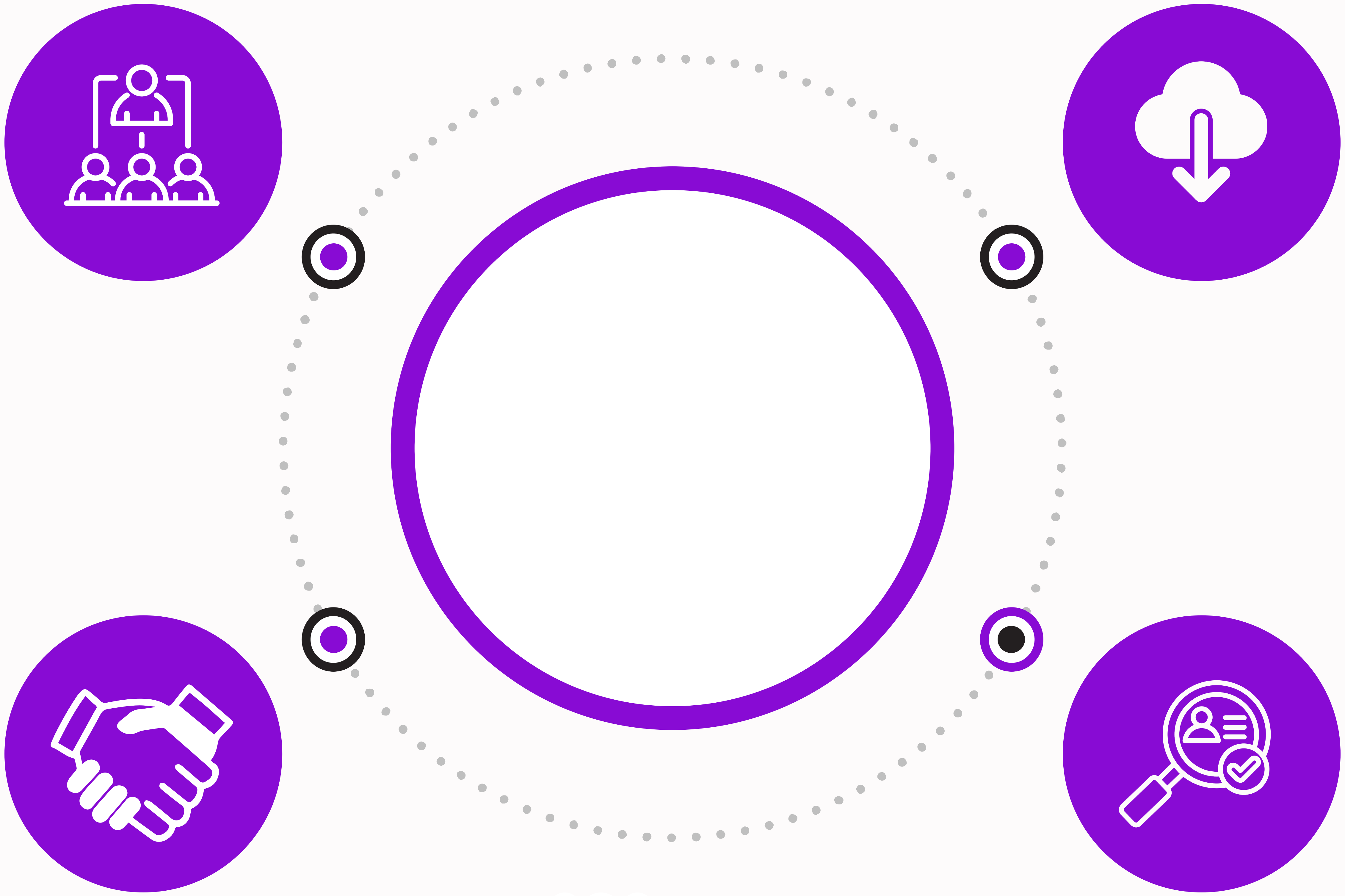
PBKEY	ADD_N	STREETNAME	CITY	STATE	ZIPCOD	PLUS4	LOC_CC	GEOID	LAT	LON	PROP_T	FIPS	LivingS	Bedroo	Bathro	SaleAmour	ParcelI	ParcelA	Elevati	Geome	BuildingID	MaxEle	MinEle	Buildin
P00005K4F	7726	BETTY LOUISE DR	PANAMA CITY	FL	32404	8536	P05	120050000000	30.14071	-85.56158	R	12005	1636	3	2	75000	C000CU68\	8950	18	{"type":"Mu	B000CTQ7M9CP	18	17	2596
P00005L2A	5376	MILLIE WAY	PANAMA CITY	FL	32404	3838	P07	120050000000	30.23639	-85.57021	R	12005	2588	5	3	107500000	C000CU6A\	5771	48	{"type":"Mu	B000CTSOYF4K	48	48	1725
P00005K4F	5422	NICOLE BLVD	PANAMA CITY	FL	32404	3051	P05	120050000000	30.23799	-85.56475	R	12005	1537	3	2	150500	C000CU5R	9221	55	{"type":"Mu	B000CTQ7L531	55	53	2623
P00005K4F	617	N 9TH ST	PANAMA CITY	FL	32404	6893	P05	120050000000	30.13572	-85.60192	R	12005	2508	5	3	82000	C000CU5N	10841	24	{"type":"Mu	B000CTQ7KZBA	24	23	2815
P00005K4S	7604	BAYOU GEORGE DR	PANAMA CITY	FL	32404	4053	P05	120050000000	30.26058	-85.5213	R	12005	1536	3	2	25000	C000CU72I	217433	41	{"type":"Mu	B000CTQ7L7J3	43	42	1730
P00005K4F	7124	CHIPEWA ST	PANAMA CITY	FL	32404	8112	P05	120050000000	30.144	-85.56905	R	12005	1613	3	2	160000	C000CU5PI	9257	30	{"type":"Mu	B000CTQ7ME2A	31	30	1704
P00005KH(	119	REDFISH WAY	PANAMA CITY	FL	32404	8962	P05	120050000000	30.24094	-85.59843	R	12005	1768	4	2	234900	C000CU5S	6739	5	{"type":"Mu	B000CTSOYAZW	6	5	2830
P00005K4F	627	HIGHWAY 2297	PANAMA CITY	FL	32404	2713	P05	120050000000	30.13535	-85.50581	R	12005	1586	3	2	290000	C000CU7G.	40707	11	{"type":"Mu	B000CTQ7JV19	11	8	1820
P00005K4F	4141	E 15TH ST	PANAMA CITY	FL	32404	5895	P05	120050000000	30.17458	-85.59333	R	12005	226068	2	2	34300	C000CU5J5	434058	33	{"type":"Mu	B000CTQ7NRQZ	33	31	12330
P00005K4F	714	FLIGHT AVE	PANAMA CITY	FL	32404	5907	P05	120050000000	30.16272	-85.60122	R	12005	1258	3	2	76000	C000CU6U	8717	23	{"type":"Mu	B000CTQ7NJG5	24	23	1529
P00005K4F	5824	CHERRY ST	PANAMA CITY	FL	32404	6446	P05	120050000000	30.1448	-85.58784	R	12005	7884	3	1	399000	C000CU7BI	32579	30	{"type":"Mu	B000CTQ7NQY4	31	30	747
P00005K4F	1026	S COMET AVE	PANAMA CITY	FL	32404	9621	P05	120050000000	30.12801	-85.57899	R	12005	1504	3	2	42500	C000CU5V	8758	12	{"type":"Mu	B000CTQ7JWHP	12	12	2420
P00005KLV	105	SEA FOX DR	PANAMA CITY	FL	32404	9813	P05	120050000000	30.12643	-85.57025	R	12005	1768	4	2	229900	C000CU77I	6051	6	{"type":"Mu	B000CTSOYEGJ	7	6	2554
P00005KAH	7408	SWEETBRIAR RD	PANAMA CITY	FL	32404	4109	P05	120050000000	30.25875	-85.5244	R	12005	1512	3	2	189000	C000CU6E	28933	36	{"type":"Mu	B000CTSOYG83	45	40	1745
P00005K4F	5010	HICKORY ST	PANAMA CITY	FL	32404	6861	P05	120050000000	30.13814	-85.59903	R	12005	1226	2	2	65500	C000CU6K	2642	31	{"type":"Mu	B000CTQ7JGO0	31	30	1243
P00005K4F	4128	CHERRY LN	PANAMA CITY	FL	32404	6223	P05	120050000000	30.14808	-85.60048	R	12005	1352	4	2	54500	C000CU5F5	13972	31	{"type":"Mu	B000CTQ7JJ0Q	31	30	2222

DEMOGRAPHICS DATASET

PSYTE	psyte_PSY	psyte_census	psyte_censu	psyte_c	psyte_c	psyte_PSYTESegr	psyte_h	psyte_householdIncomeVariable.	psyte_p	psyte_propertyV	psyte_p	psyte_propertyTenureVar	psyte_propertyTypeVarial	psyte_propertyTypeVar	psyte_u	psyte_
1	7	120050000000000	1.2005E+11	142	53	Renter Fringe	BAV	Bottom 30-49.99% of households by h	B20	Bottom 10-19.99%	MIX	Mixed tenure	SFR	Majority of properties sing	URB	Block
1	9	120050000000000	1.2005E+11	33	8	Youthful Heartlands	AAV	Top 30-50% of households by househ	AAV	Top 30-50% of prop	MOR	Majority of properties owned	SFR	Majority of properties sing	URB	Block
2	7	120050000000000	1.2005E+11	111	36	Farmland Families	BAV	Bottom 30-49.99% of households by h	BAV	Bottom 30-49.99%	MOR	Majority of properties owned	SFR	Majority of properties sing	URB	Block
4	9	120050000000000	1.2005E+11	87	44	Strapped Rural Com	BAV	Bottom 30-49.99% of households by h	B20	Bottom 10-19.99%	RNT	Majority of properties rented	SFR	Majority of properties sing	URB	Block
2	7	120050000000000	1.2005E+11	66	16	Farmland Families	BAV	Bottom 30-49.99% of households by h	B20	Bottom 10-19.99%	MOR	Majority of properties owned	MOB	Majority of properties mob	URB	Block
2	10	120050000000000	1.2005E+11	28	14	Working to Live	AAV	Top 30-50% of households by househ	B20	Bottom 10-19.99%	OWN	Majority of properties owned	SFR	Majority of properties sing	URB	Block
3	4	120050000000000	1.2005E+11	88	21	Flourishing Modern	T20	Top 10-19.99% of households by hous	AAV	Top 30-50% of prop	MOR	Majority of properties owned	SFR	Majority of properties sing	URB	Block
5	6	120050000000000	1.2005E+11	87	42	Comfortable Comm	BAV	Bottom 30-49.99% of households by h	BAV	Bottom 30-49.99%	OWN	Majority of properties owned	SFR	Majority of properties sing	RUR	Block
2	8	120050000000000	1.2005E+11	132	72	Hometown Singles	AAV	Top 30-50% of households by househ	BAV	Bottom 30-49.99%	RNT	Majority of properties rented	SFR	Majority of properties sing	URB	Block
3	11	120050000000000	1.2005E+11	42	13	Hard Times Country	B20	Bottom 10-19.99% of households by h	B30	Bottom 20-29.99%	OWN	Majority of properties owned	SFR	Majority of properties sing	URB	Block
7	8	120050000000000	1.2005E+11	73	16	Rural Family Values	AAV	Top 30-50% of households by househ	BAV	Bottom 30-49.99%	MOR	Majority of properties owned	SFR	Majority of properties sing	URB	Block
4	12	120050000000000	1.2005E+11	216	82	Urban Family Strugg	BAV	Bottom 30-49.99% of households by h	B30	Bottom 20-29.99%	RNT	Majority of properties rented	SFR	Majority of properties sing	URB	Block
4	7	120050000000000	1.2005E+11	247	93	Comfortably Conne	BAV	Bottom 30-49.99% of households by h	AAV	Top 30-50% of prop	MOR	Majority of properties owned	SFR	Majority of properties sing	URB	Block
5	7	120050000000000	1.2005E+11	103	22	Wide Open Country	BAV	Bottom 30-49.99% of households by h	B30	Bottom 20-29.99%	MOR	Majority of properties owned	SFR	Majority of properties sing	URB	Block
5	9	120050000000000	1.2005E+11	61	33	Smalltown Tenants	AAV	Top 30-50% of households by househ	B10	Bottom 9.99% of pr	RNT	Majority of properties rented	TNH	Majority of properties towi	URB	Block
3	12	120050000000000	1.2005E+11	64	17	Young Working Clas	B30	Bottom 20-29.99% of households by h	BAV	Bottom 30-49.99%	RNT	Majority of properties rented	SFR	Majority of properties sing	URB	Block

DATA CLEANING OVERVIEW

- MISSING VALUES
- REMOVING DUPLICATES
- HANDLING IRRELEVANT DATA
- DATA STANDARDIZATION



DATA CLEANING CODE

```
# Step 1: Define numeric and categorical columns based on the dataset
numeric_cols = ['LivingSquareFootage', 'BedroomCount', 'BathroomCount', 'SaleAmount',
                'ParcelArea', 'Elevation', 'MaxElevation', 'MinElevation', 'BuildingArea']

categorical_cols = ['PBKEY','ADD_NUMBER','STREETNAME','CITY','STATE','ZIPCODE','LOC_CODE','Geometry','GEOID','

# Print the names of numeric and categorical columns
print(f"Numeric columns: {numeric_cols}")
print(f"Categorical columns: {categorical_cols}")

# Step 2: Handle missing values
# Drop columns with more than 50% missing values
threshold = len(data) * 0.5
data.dropna(axis=1, thresh=threshold, inplace=True)

# Print the shape after dropping columns with more than 50% missing values
print(f"After dropping columns with >50% missing values: {data.shape}")

# Ensure that missing values for numeric columns are treated as np.nan
data[numeric_cols] = data[numeric_cols].replace('N/A', np.nan)

# Fill missing values for numerical columns with median
data[numeric_cols] = data[numeric_cols].fillna(data[numeric_cols].median())

# Fill missing values for categorical columns with 'N/A'
data[categorical_cols] = data[categorical_cols].fillna('N/A')
```

```
# Step 3: Standardize data types
# Convert date columns to datetime format if applicable
date_cols = [col for col in data.columns if 'Date' in col]
for col in date_cols:
    data[col] = pd.to_datetime(data[col], errors='coerce')

# Print the shape after standardizing data types
print(f"After standardizing data types: {data.shape}")

# Step 4: Remove duplicates
data.drop_duplicates(inplace=True)

# Print the shape after removing duplicates
print(f"After removing duplicates: {data.shape}")

# Step 5: Strip extra whitespace from string columns
string_cols = [col for col in categorical_cols if data[col].dtype == 'object'
data[string_cols] = data[string_cols].apply(lambda x: x.str.strip())

# Print the shape after stripping extra whitespace
print(f"After stripping extra whitespace: {data.shape}")


# Save the final cleaned data to a new CSV file with the given name
final_cleaned_file_path = 'Final_Cleaned_Residential_Data(12).csv'
```

CHALLENGES

DATA ENRICHMENT

- Large Data Set
- Missing Fields
- Time Constraints

DATA CLEANING

- Key Mismatches
- Outliers
- Missing values

NEXT STEPS

1

CONTINUE DATA CLEANING

2

VISUALIZATION

3

BUSINESS INSIGHTS

THANK YOU