Final Presentation

Presented by: Mehreen Akmal, Jen Bushey, Eric Diep, Hao Liu, Corey Yang Smith

Where we left off

- Processed our datasets separately

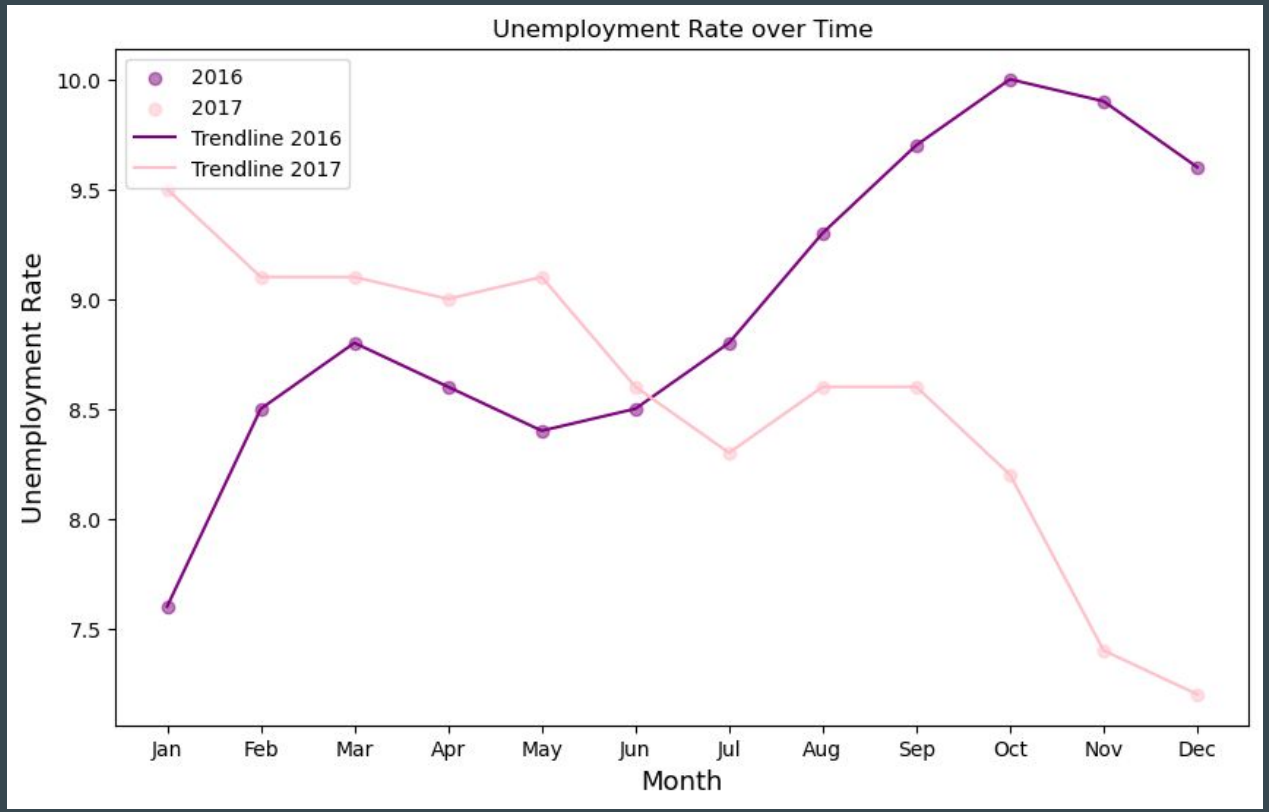
Table 1: Summary table of data collection and feature selection.

#	Target Feature	Category	Dataset
0	Assessed Value	Target Variable	Calgary Assessed Property Values
1	Assessed Year	Economic Indicator	Any Dataset
2	Average Home Price	Economic Indicator	Prosperity Indicator
3	Unemployment Rate	Economic Indicator	Prosperity Indicator
4	Housing Starts	Economic Indicator	Prosperity Indicator
5	City Population	Economic Indicator	Census
6	Dwelling Type	Individual Price Indicator	Census
7	Density by Community	Group Price Indicator	Historical Community Populations
8	Community Name/Code	Group Price Indicator	Calgary Assessed Property Values
9	Crime Count (?)	Group Price Indicator	Calgary Crime Data
10	Resident Count (?)	Group Price Indicator	Historical Community Populations
11	Land Use Designation (?)	Individual Price Indicator	Calgary Assessed Property Values
12	Property Type (?)	Individual Price Indicator	Calgary Assessed Property Values
13	Land Size	Individual Price Indicator	Calgary Assessed Property Values
14	Interprovincial Immigration	Economic Indicator	Interprovincial Migration Data
15	New Development by Community	Group Price Indicator	Building Permits/Developer Permits
16	Inflation Rate	Economic Indicator	Historical CPI

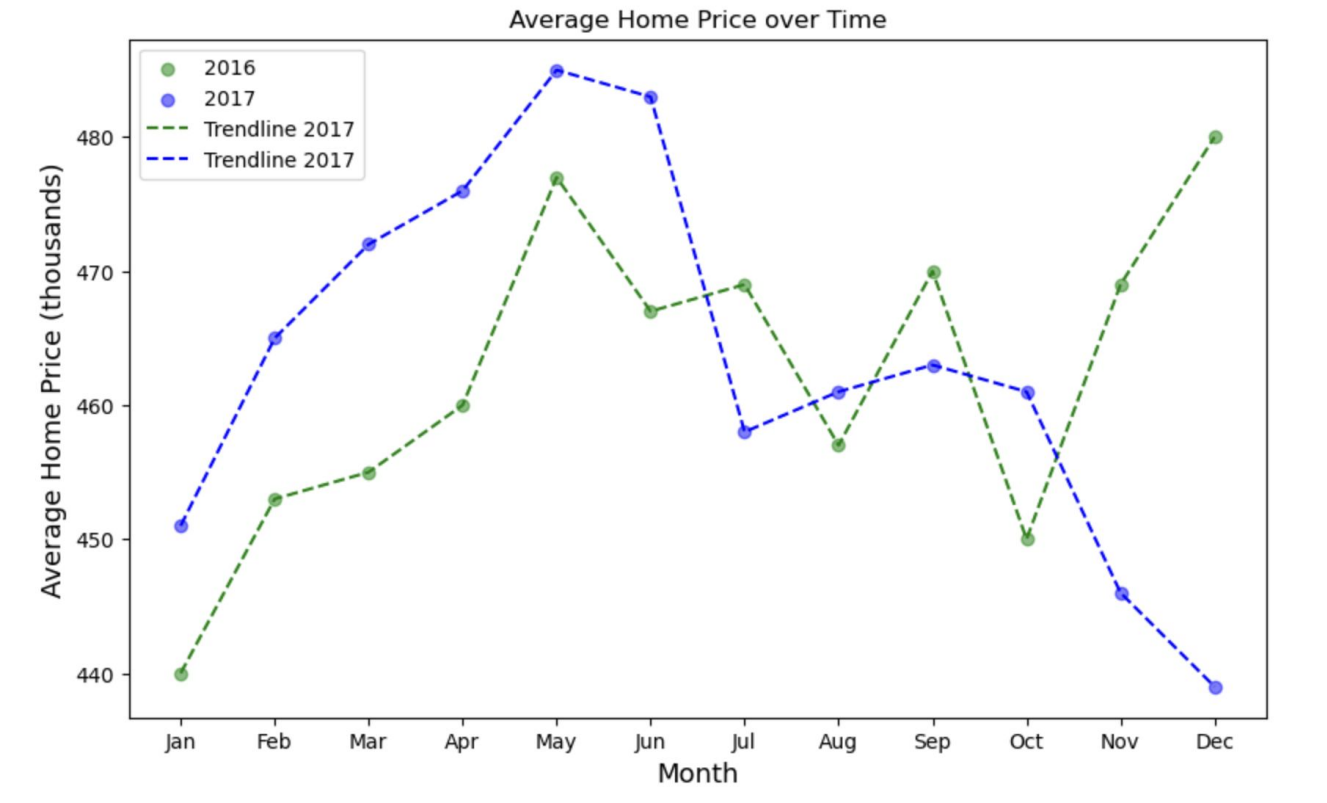
What did we do?

- Converted our data processing from pandas → pyspark
- Combined all datasets together into a single RDD
- Performed EDA
- One Hot Encoding
- Tested preliminary regression models

Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis

Combining Data

```
Macro-Scale Dataset
[(2017, 50396, 1.2283333333333337, 463.3666666666666, 961.1666666666666, 2.6624062087645566, 8.5583333

Community-Scale Dataset
[(2017, 'GRV', 'GREENVIEW', '2036', '1032', 'GREENVIEW', 240, 2036), (2017, 'CHN', 'CHINATOWN', '2124'

Individual-Scale Dataset
[(2017, 'SAD', '41 SADDLEBACK RD NE', 'RE', '411000.0', 'R-1N', 'LI', '307.9'), (2017, 'SAD', '45 SADD
```

```
print("Final RDD")
print(final_rdd.take(5))
print('')
```

```
Intermediate RDD (Macro + Community)
[(2017, 'GRV', 'GREENVIEW', '2036', '1032', 'GREENVIEW', 240, 2036, 50396, 1.2283333333333337, 463.3666666666666, 961.16666666

Individual RDD
[(2017, 'SAD', '41 SADDLEBACK RD NE', 'RE', '411000.0', 'R-1N', 'LI', '307.9'), (2017, 'SAD', '45 SADDLEBACK RD NE', 'RE', '36

Processing...

Final RDD
[(2017, 'ERL', 'ERLTON', '1307', '694', 'ERLTON', 186, 1307, 50396, 1.2283333333333337, 463.3666666666666, 961.1666666666666,
```

Preliminary Models - Linear Regression

```
Mean Squared Error (MSE): 52288364994.18242
Mean Absolute Error (MAE): 130479.20747614212
R-squared: -11.430414436682128
```

```
Labels (Actual):
385000.0
274500.0
439000.0
494500.0
254000.0

Predictions:
472110.380917034
400607.8803676837
400607.8803676837
400607.8803676837
400607.8803676837
```

Preliminary Models - Linear Regression

```
Mean Squared Error (MSE): 52288364994.18242
Mean Absolute Error (MAE): 130479.20747614212
R-squared: -11.430414436682128
```

```
Labels (Actual):
385000.0
274500.0
439000.0
494500.0
254000.0

Predictions:
472110.380917034
400607.8803676837
400607.8803676837
400607.8803676837
400607.8803676837
```

What do we need to do?

- Further process some of our more difficult data and incorporate into dataset
- Implement scaling to numerical data
- Explore non-linear models (random forest)
- Further refine our selected model
- Determine where **data loss** is occurring
  - Currently some error with combining our RDD
  - ~412,000 Rows → ~139,000 Rows
- Further **process data**
  - Normalize: Convert Community Population & Occupied Dwellings → Community Density
  - Drop “Year” - Cannot process time series data
  - Drop redundant data → include resident count twice

Data Loss

```
print("Calgary Assessed Property Values RDD Information")
print(assessed_vals_rdd.count())
print('')
```

```
Calgary Assessed Property Values RDD Information
412209
```

```
Training Data Count:  69201
Validation Data Count:  34723
Test Data Count:  34859
```

Sum:
138,783

Thank You

...

=)