

“Exploratory Analysis for Smart Real Estate”

Course: Data Science for Business

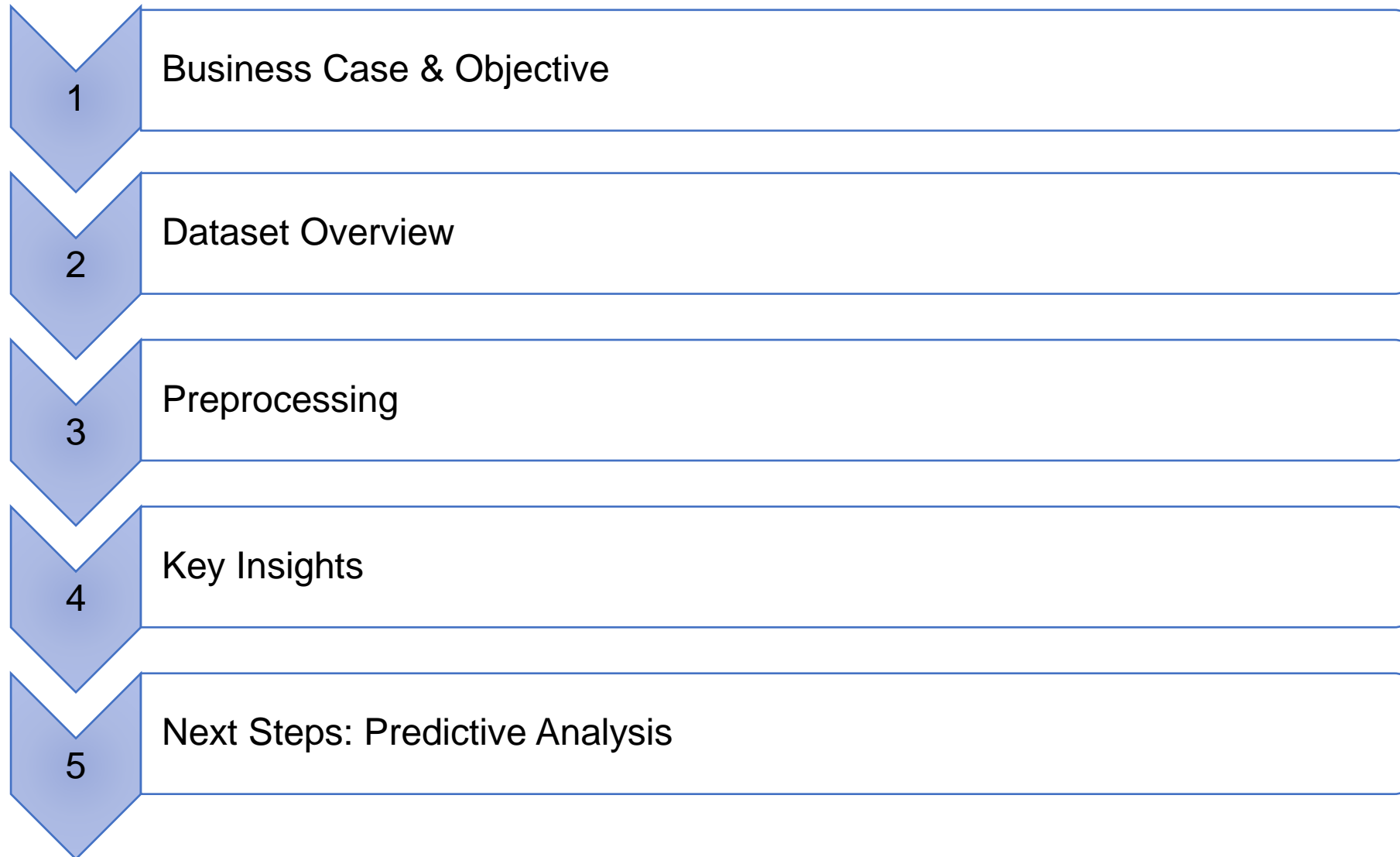
Study Program: Business Consulting Masters

WiSe 24-25

Subhashri Ravichandran

Prof. Dr. Holger Ziekow

Agenda



1. Business Case & Objective (1/2)

Business Case



1. BUSINESS CASE & OBJECTIVE

2. DATASET OVERVIEW

3. PREPROCESSING

4. KEY INSIGHTS

5. NEXT STEPS: PREDICTIVE ANALYSIS

1. Business Case & Objective (2/2)

Objectives



- 1. Analyze and assess the UK housing dataset**
- 2. Identify historical trends and patterns in house prices and sales**
- 3. Perform deeper predictive analysis for future insights (Phase 2)**
- 4. Give investment ideas or recommendations (Phase 2)**

2. Dataset Overview (1/2)

Dataset Source: <https://www.kaggle.com/hm-land-registry/uk-housing-prices-paid>

```
In [22]: df.head()
```

```
Out[22]:
```

	Transaction unique identifier	Price	Date of Transfer	\
0	{81B82214-7FBC-4129-9F6B-4956B4A663AD}	25000	1995-08-18 00:00	
1	{8046EC72-1466-42D6-A753-4956BF7CD8A2}	42500	1995-08-09 00:00	
2	{278D581A-5BF3-4FCE-AF62-4956D87691E6}	45000	1995-06-30 00:00	
3	{1D861C06-A416-4865-973C-4956DB12CD12}	43150	1995-11-24 00:00	
4	{DD8645FD-A815-43A6-A7BA-4956E58F1874}	18899	1995-06-23 00:00	

	Property Type	Old/New	Duration	Town/City	District	\
0	T	N	F	OLDHAM	OLDHAM	
1	S	N	F	GRAYS	THURROCK	
2	T	N	F	HIGHBRIDGE	SEDGEMOOR	
3	T	N	F	BEDFORD	NORTH BEDFORDSHIRE	
4	S	N	F	WAKEFIELD	LEEDS	

	County	PPDCategory	Type	Record Status - monthly file only
0	GREATER MANCHESTER		A	A
1	THURROCK		A	A
2	SOMERSET		A	A
3	BEDFORDSHIRE		A	A
4	WEST YORKSHIRE		A	A

```
In [23]: df.columns
```

```
Out[23]:
```

```
Index(['Transaction unique identifier', 'Price', 'Date of Transfer',  
      'Property Type', 'Old/New', 'Duration', 'Town/City', 'District',  
      'County', 'PPDCategory Type', 'Record Status - monthly file only'],  
      dtype='object')
```

```
In [27]: df.shape
```

```
Out[27]: (22489348, 11)
```

Data Overview

Columns and records available

2. Dataset Overview (2/2)

```
In [25]: df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22489348 entries, 0 to 22489347
Data columns (total 11 columns):
#   Column                                Dtype
---  -
0   Transaction unique identifier         object
1   Price                                int64
2   Date of Transfer                     object
3   Property Type                        object
4   Old/New                             object
5   Duration                             object
6   Town/City                           object
7   District                            object
8   County                              object
9   PPDCategory Type                    object
10  Record Status - monthly file only    object
dtypes: int64(1), object(10)
memory usage: 1.8+ GB
```

Datatype of Columns

```
In [26]: df.isnull().sum()
Out[26]:
Transaction unique identifier    0
Price                           0
Date of Transfer                 0
Property Type                    0
Old/New                         0
Duration                        0
Town/City                       0
District                        0
County                          0
PPDCategory Type                0
Record Status - monthly file only 0
dtype: int64
```

No. of null values

3. Preprocessing (1/2)

- *Extracting `Year` and `Month` from `Date of Transfer`*
- *Converting datatype of `Date of Transfer` to datetime*

```
...: pandarallel.initialize()
...: df['Date of Transfer'] = df['Date of Transfer'].parallel_apply(lambda x : x.split(' ')[0])
...: df['Year'] = df['Date of Transfer'].str[:4]
...: df['Date of Transfer'].iloc[1].split(" ")[0]
...: df['Date of Transfer'] = df['Date of Transfer'].parallel_apply(lambda x:
__import__('datetime').datetime.strptime(x, '%Y-%m-%d'))
...: df['Year'] = df['Date of Transfer'].dt.year
...: df['Month'] = df['Date of Transfer'].dt.month
INFO: Pandarallel will run on 8 workers.
INFO: Pandarallel will use standard multiprocessing data transfer (pipe) to transfer data between the
main process and workers.
```

WARNING: You are on Windows. If you detect any issue with pandarallel, be sure you checked out the
Troubleshooting page:
<https://nalepae.github.io/pandarallel/troubleshooting/>

```
In [4]: df.head()
```

```
Out[4]:
```

	Price	Date of Transfer	Property Type	Old/New	Town/City \
0	25000	1995-08-18	T	N	OLDHAM
1	42500	1995-08-09	S	N	GRAYS
2	45000	1995-06-30	T	N	HIGHBRIDGE
3	43150	1995-11-24	T	N	BEDFORD
4	18899	1995-06-23	S	N	WAKEFIELD

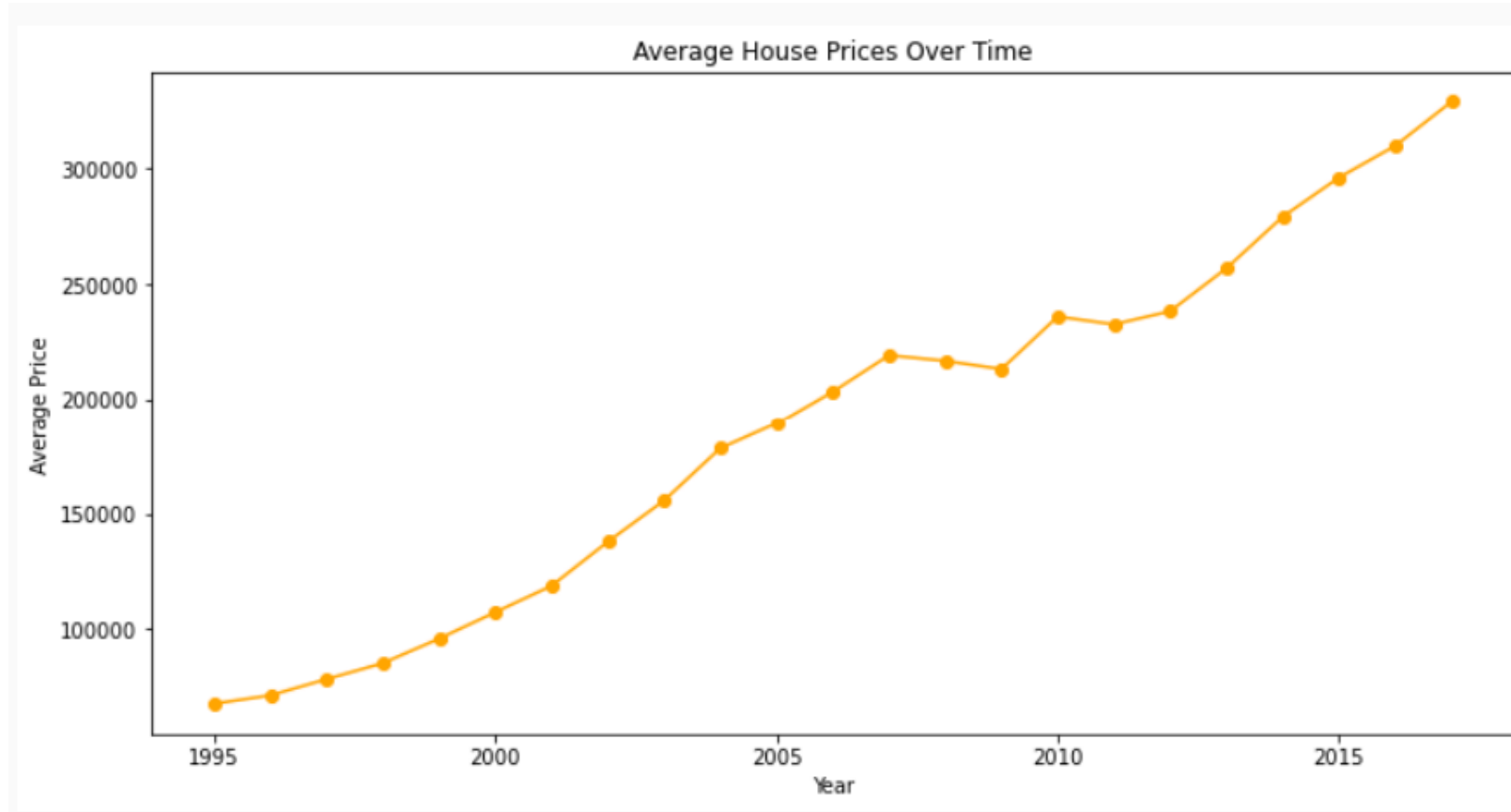
	District	County	Year	Month
0	OLDHAM	GREATER MANCHESTER	1995	8
1	THURROCK	THURROCK	1995	8
2	SEDGEMOOR	SOMERSET	1995	6
3	NORTH BEDFORDSHIRE	BEDFORDSHIRE	1995	11
4	LEEDS	WEST YORKSHIRE	1995	6

Drop unwanted columns

```
In [39]: df.drop(columns = 'Transaction unique identifier', axis = 1, inplace = True)
...: df.drop(columns = 'Duration', axis = 1, inplace = True)
...: df.drop(columns = 'PPDCategory Type', axis = 1, inplace = True)
...: df.drop(columns = 'Record Status - monthly file only', axis = 1, inplace = True)
...: df.shape
Out[39]: (22489348, 7)
```


4. Key Insights (1/8)

Overall Sales price trend analysis



- Avg Sales Price follows upward trend throughout the years (1995-2017)

Fig.1

4. Key Insights (2/8)

Overall Sales transactions trend analysis

Heatmap of Transactions Across Months and Years (2007-2017)

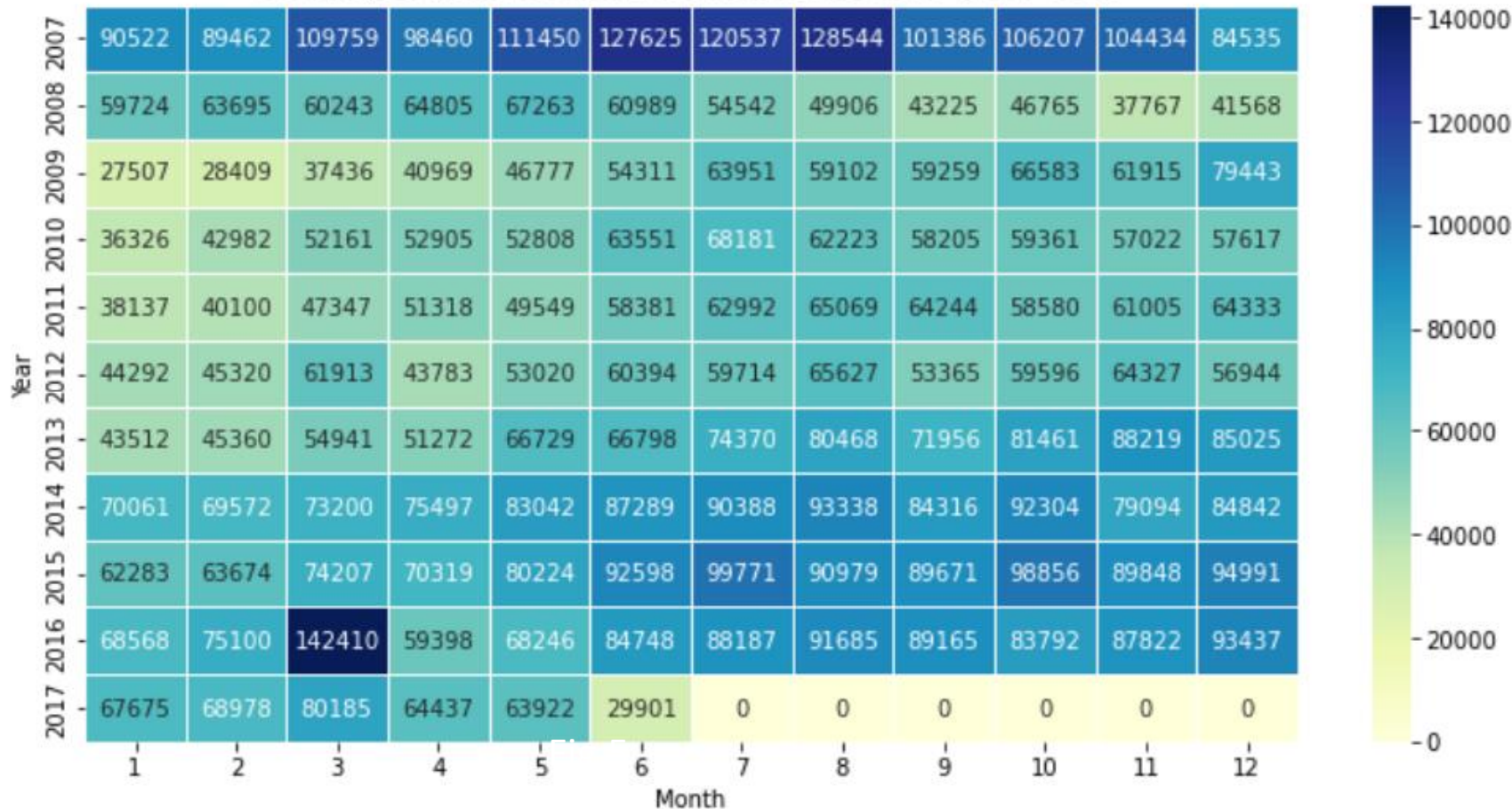


Fig.2

- High in 2007
- Drops from 2008
- Recovery starts from 2013
- Becomes stable in 2014, 2015, 2016
- Low in Jan and Feb
- Good from mid year

4. Key Insights (3/8)

Property analysis 1/2

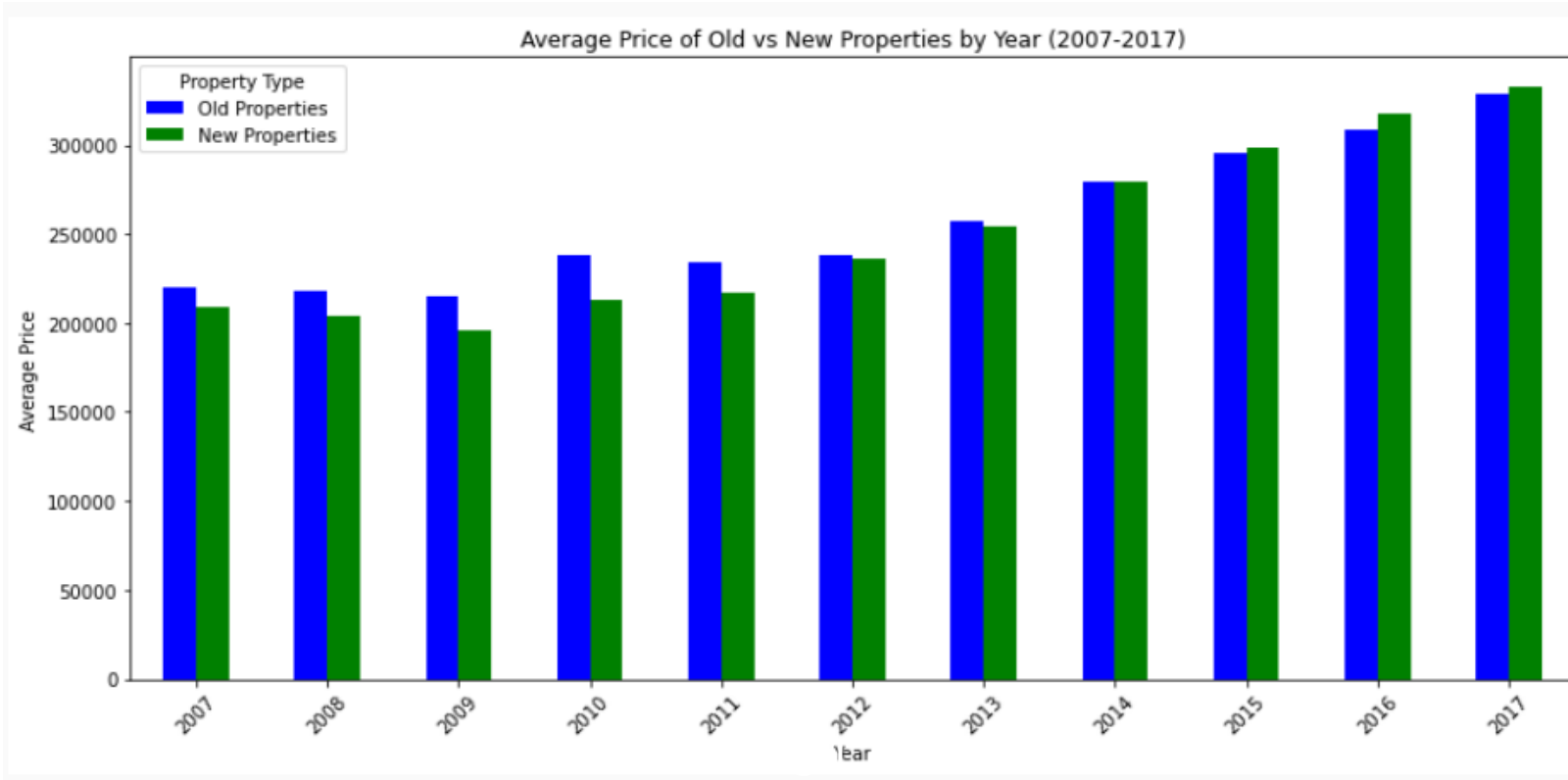


Fig.3

- 2007 – 2013 – Avg price of new property < Avg price of old property
- 2014 – Avg price of new property equals Avg price of old property
- 2015 – 2017 – Avg price of new property > Avg price of old property

4. Key Insights (4/8)

Property analysis 2/2

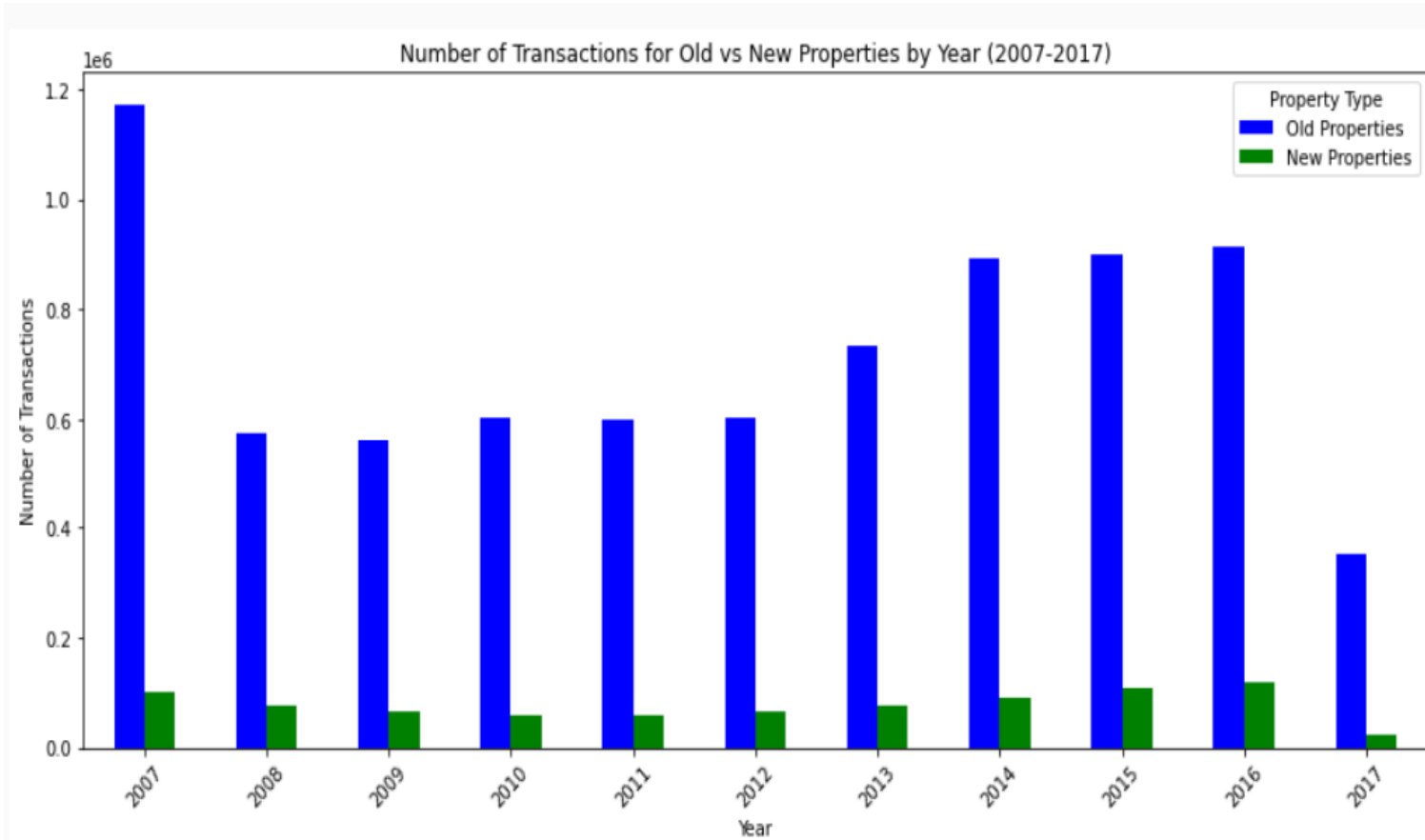


Fig.4

- The number of transactions of old properties beats the number of transactions of new properties

4. Key Insights (5/8)

Property type analysis 1/2

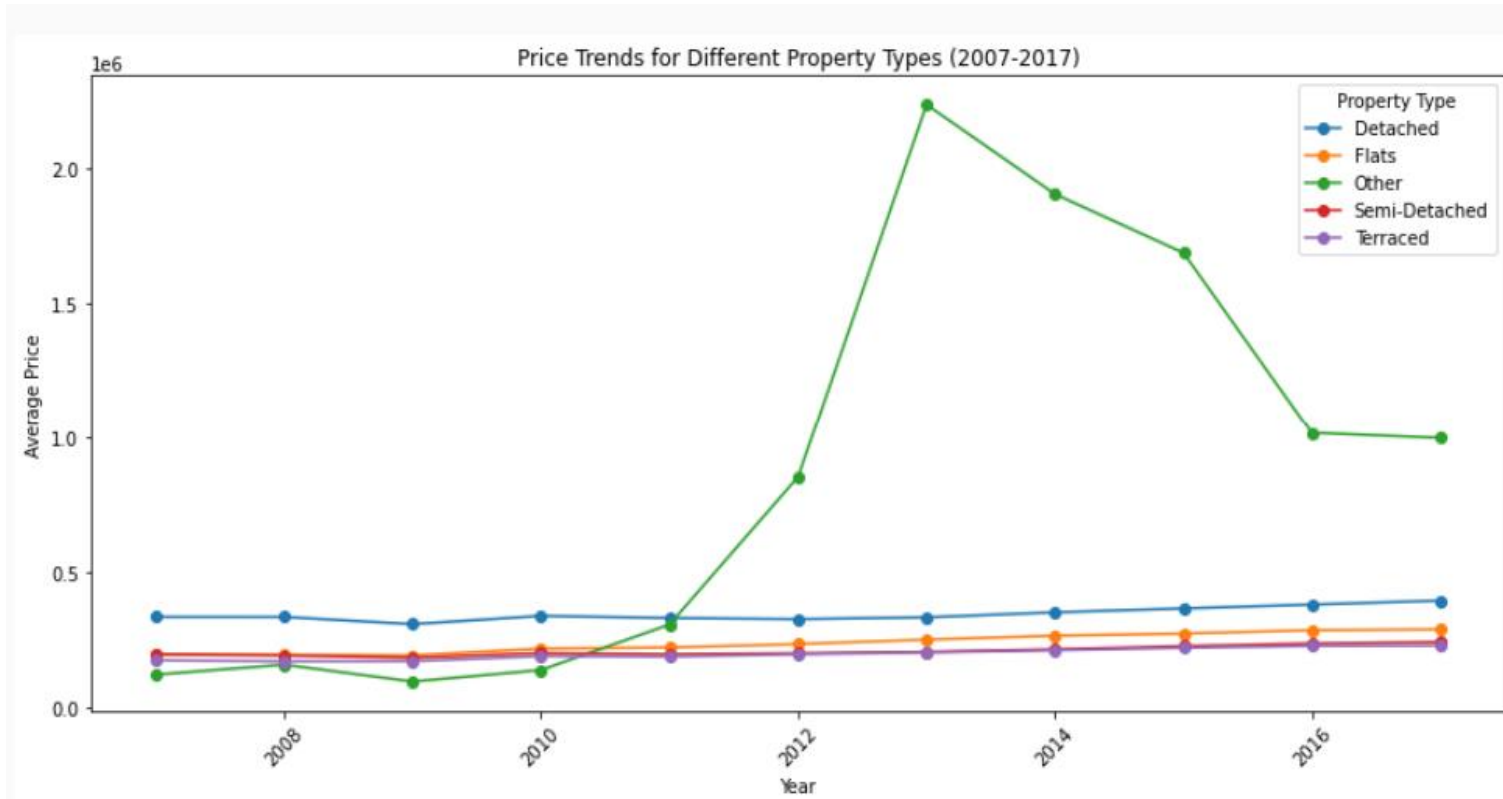


Fig.5

- The avg price of ` Other ` property type is high from 2011 onwards
- The avg price of ` Terraced ` property is the lowest through 2007-2017

4. Key Insights (6/8)

Property type analysis 2/2

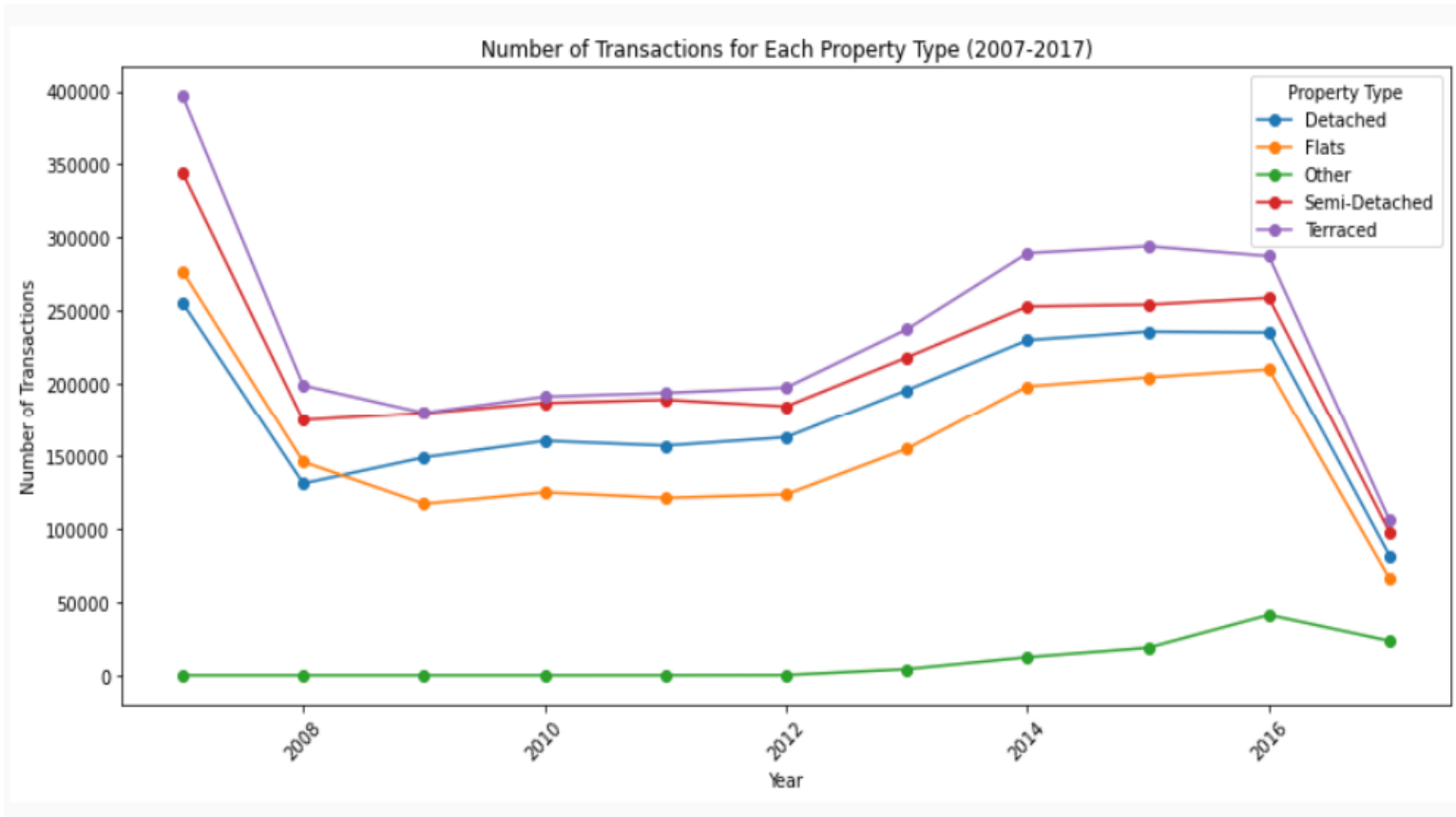


Fig.6

- Number of transactions is the lowest for `Other` property type.
- Number of transactions remains high for `Terraced` property type through 2007-2017

4. Key Insights (7/8)

Sales transactions trend analysis 1/2

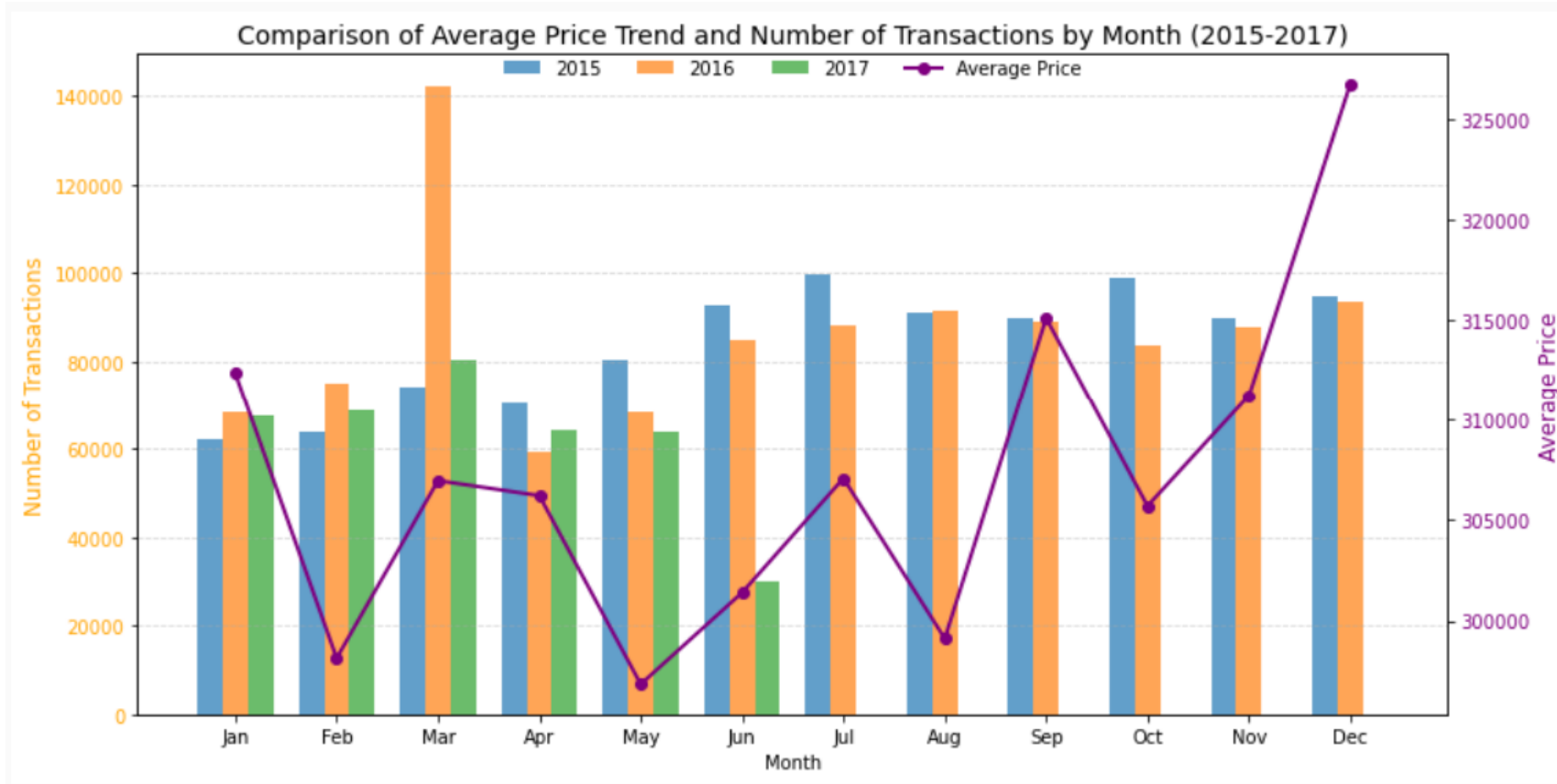


Fig.7

- Prices drop the most during February and May
- Prices attain it's peak during September and December

4. Key Insights (8/8)

Sales transactions trend analysis 2/2

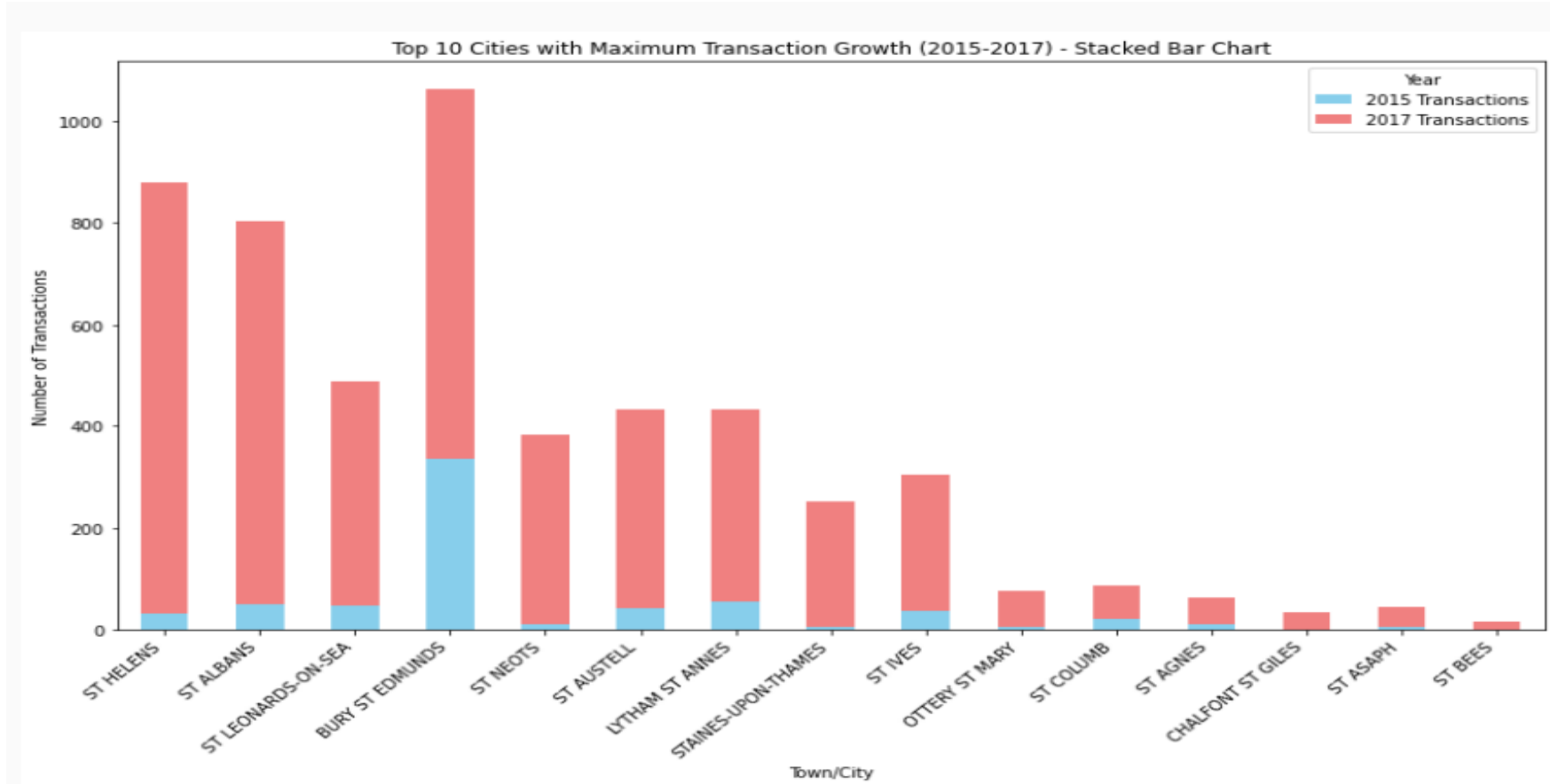


Fig.8

- 14 out of 15 cities belong to England
- 2017's Transaction: 6-month performance surpasses entire 2015

5. Next Steps: Predictive Analysis



In what should we invest ?
(Best property type)

When should we invest ?
(Optimal time/months)

Where should we invest ?
(Best cities)