Predicting Ireland's residential property prices

✨ Workshop 4 - Data Cleaning ✨

Hannah O'Connor & Catherine Mooney (group 22)

Before

```
Dublin mean: €390448.66
Dublin median: €290000.00
Dublin max: €139165000.00
Dublin min: €5250.00
Dublin standard deviation: €1135687.72
Cork mean: €219793.65
Cork median: €183500.00
Cork max: €69873482.00
Cork min: €5030.53
Cork standard deviation: €624223.93
Galway mean: €196268.59
Galway median: €165000.00
Galway max: €34781000.00
Galway min: €5864.00
Galway standard deviation: €339919.34
Kildare mean: €253729.18
Kildare median: €235000.00
Kildare max: €26500000.00
Kildare min: €6500.00
Kildare standard deviation: €288071.70
Meath mean: €229081.93
Meath median: €216000.00
Meath max: €5536500.00
Meath min: €6000.00
Meath standard deviation: €148263.66
```

← 5# summary of top 5

Entries per county →

↓ Head of our dataset

	Date of Sale (dd/mm/yyyy)	Address	Postal Code	County	Price (€)	Not Full Market Price	VAT Exclusive	Description of Property	Property Size Description
0	01/01/2010	5 Braemor Drive, Churchtown, Co.Dublin	NaN	Dublin	€343,000.00	No	No	Second-Hand Dwelling house /Apartment	NaN
1	03/01/2010	134 Ashewood Walk, Summerhill Lane, Portlaoise	NaN	Laois	€185,000.00	No	Yes	New Dwelling house /Apartment	greater than or equal to 38 sq metres and less...
2	04/01/2010	1 Meadow Avenue, Dundrum, Dublin 14	NaN	Dublin	€438,500.00	No	No	Second-Hand Dwelling house /Apartment	NaN
3	04/01/2010	1 The Haven, Mornington	NaN	Meath	€400,000.00	No	No	Second-Hand Dwelling house /Apartment	NaN
4	04/01/2010	11 Melville Heights, Kilkenny	NaN	Kilkenny	€160,000.00	No	No	Second-Hand Dwelling house /Apartment	NaN

```
Dublin        128093
Cork           43002
Galway         19952
Kildare        19123
Meath          15255
Limerick       14684
Wexford        13289
Wicklow        12067
Kerry          11465
Donegal        10759
Tipperary      10428
Louth          10421
Waterford      10323
Mayo            9277
Clare           9039
Westmeath       7628
Kilkenny        6209
Cavan           6163
Laois           5954
Sligo           5875
Roscommon       5580
Offaly          4659
Carlow          4282
Leitrim         3558
Longford        3383
Monaghan        3076
Name: county, dtype: int64
```

Data Preparation

Removed unnecessary attributes/rows...

- Had Irish descriptions, property size, etc.
- Weren't full market price + removed column

Also...

- Converted prices to floats
- Converted upper case addresses
- Transformed header names
- Split up csv files to individual counties

Went from 383,116 rows to 374,770 rows.

Missing Values..

date_of_sale	0
address	0
postal_code	305009
county	0
price	0
vat_exclusive	0
property_description	0
property_size	323669

```
df['property_size'].unique()
```

```
array([nan,
       'greater than or equal to 38 sq metres and less than 125 sq metres',
       'greater than 125 sq metres', 'less than 38 sq metres',
       'greater than or equal to 125 sq metres'], dtype=object)
```

```
df['postal_code'].unique()
```

```
array([nan, 'Dublin 14', 'Dublin 2', 'Dublin 13', 'Dublin 12', 'Dublin 4',
       'Dublin 9', 'Dublin 24', 'Dublin 15', 'Dublin 22', 'Dublin 5',
       'Dublin 18', 'Dublin 6', 'Dublin 6w', 'Dublin 7', 'Dublin 16',
       'Dublin 11', 'Dublin 8', 'Dublin 3', 'Dublin 1', 'Dublin 17',
       'Dublin 20', 'Dublin 10'], dtype=object)
```

- → Removed postal_code, property_size & created area_code column.
- → For Dublin, no town column & a lot of addresses did not include as part.
- → Went from 374,770 rows to 357,795.

Missing Values...

- ★ Ignore the tuple
- ★ Fill in the missing value
- ★ Global constant for missing value
- ★ Attribute mean for missing value
- ★ Attribute mean for samples in same class as the tuple
- ★ Probable value to fill in the missing value

Noisy Data...

Three ways to deal with Noisy Data...

- ★ **Clustering:** organising values into groups of similarity to highlight outliers.
- ★ **Binning:** smooth sorted data values by consulting the values around it.
- ★ **Regression:** fitting the data through a function.

We chose **Binning!**

Binning

Before (Creating bins per town):

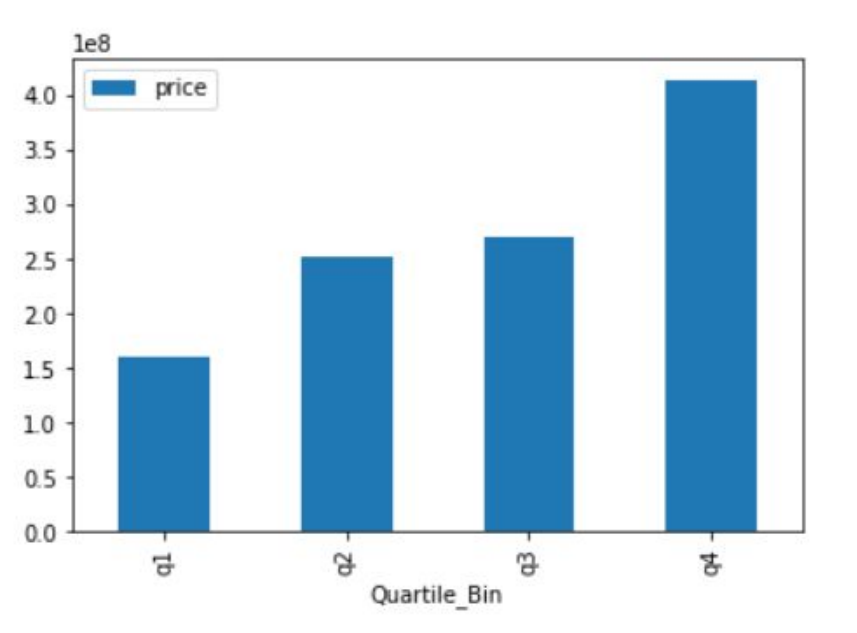
	date_of_sale	address	county	price	vat_exclusive	town	area_code	quartile_bins	decile_bins
0	06/01/2010	18 Earlsfort Court, Lucan, Co. Dublin	Dublin	280000.0	No	Lucan	Co. Dublin	(275000.0, 325990.858]	(275000.0, 295000.0]
1	06/01/2010	9 Colthurst Green, Huntington Glen, Lucan	Dublin	228000.0	No	Lucan	Co. Dublin	(218375.0, 275000.0]	(200000.0, 230000.0]
2	21/01/2010	4 Caislean Riada Avenue, Lucan	Dublin	272500.0	No	Lucan	Co. Dublin	(218375.0, 275000.0]	(253000.0, 275000.0]
3	22/01/2010	13 Beech Park, Lucan	Dublin	478000.0	Yes	Lucan	Co. Dublin	(325990.858, 3432000.0]	(404500.0, 3432000.0]
4	22/01/2010	4 Larkfield Grove, Lucan	Dublin	230000.0	No	Lucan	Co. Dublin	(218375.0, 275000.0]	(200000.0, 230000.0]

After (Adding labels for bins):

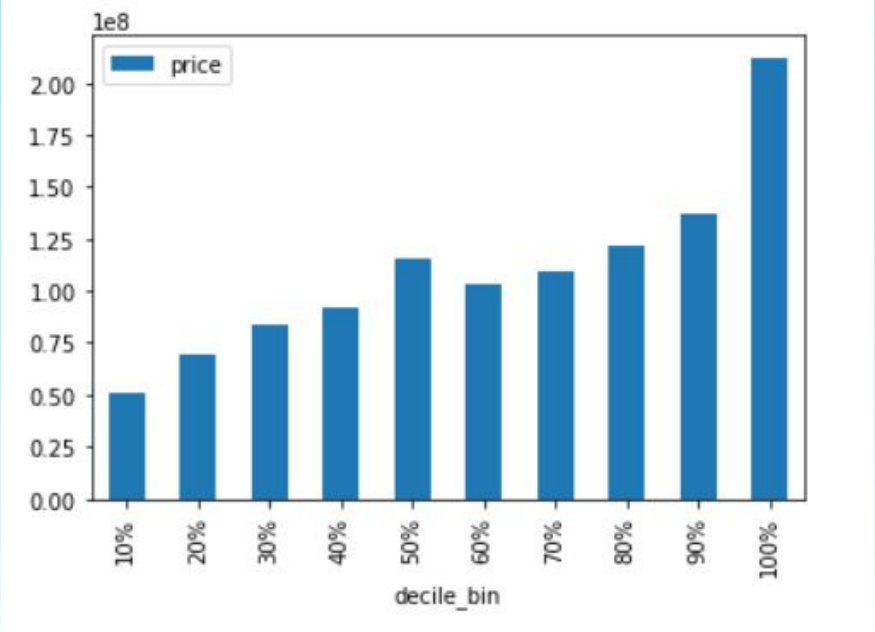
	date_of_sale	address	county	price	vat_exclusive	town	area_code	quartile_bins	decile_bins
0	06/01/2010	18 Earlsfort Court, Lucan, Co. Dublin	Dublin	280000.0	No	Lucan	Co. Dublin	q3	60%
1	06/01/2010	9 Colthurst Green, Huntington Glen, Lucan	Dublin	228000.0	No	Lucan	Co. Dublin	q2	30%
2	21/01/2010	4 Caislean Riada Avenue, Lucan	Dublin	272500.0	No	Lucan	Co. Dublin	q2	50%
3	22/01/2010	13 Beech Park, Lucan	Dublin	478000.0	Yes	Lucan	Co. Dublin	q4	100%
4	22/01/2010	4 Larkfield Grove, Lucan	Dublin	230000.0	No	Lucan	Co. Dublin	q2	30%

Binning (2)

Quartile Graph



Decile Graph



Data Cleaning as a Process!

Discrepancy Detection

Discrepancies

- They're caused by several factors, from poor designed data entry forms to human error in data entry + deliberate errors Discrepancies may also arise from inconsistent data representations and the inconsistent use of codes.

Discrepancy Detection(2)

1) Use Metadata

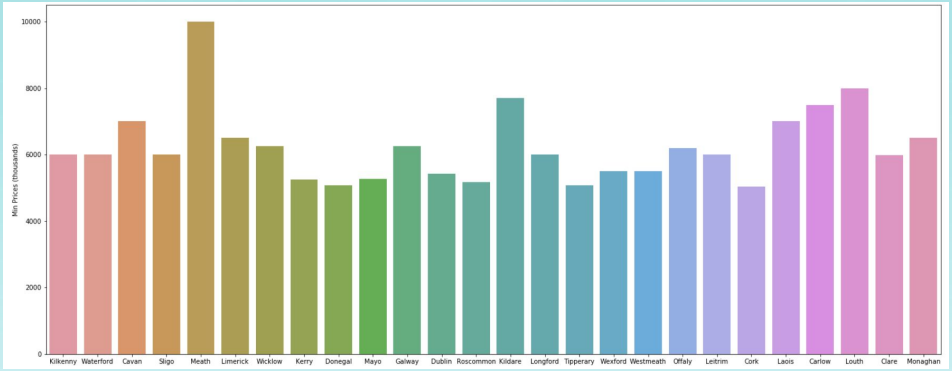
- Didn't originally contain useful metadata
- Added Latitude & Longitude coordinates & its accuracy.
- Also retrieved eircode based on given address. (all through google geocoding API)

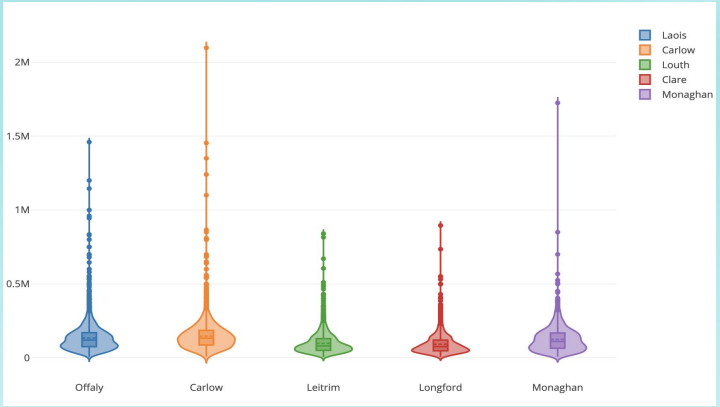
address	postal_code	county	price	vat_exclusive	property_description	property_size	town	area_code	eircode	longitude	latitude	accuracy
5 Braemor Drive, Churchtown, Co.Dublin	NaN	Dublin	343000.0	No	Second-Hand Dwelling house /Apartment	NaN	Churchtown	Dublin 14	D14 NX40	-6.263783	53.302391	ROOFTOP

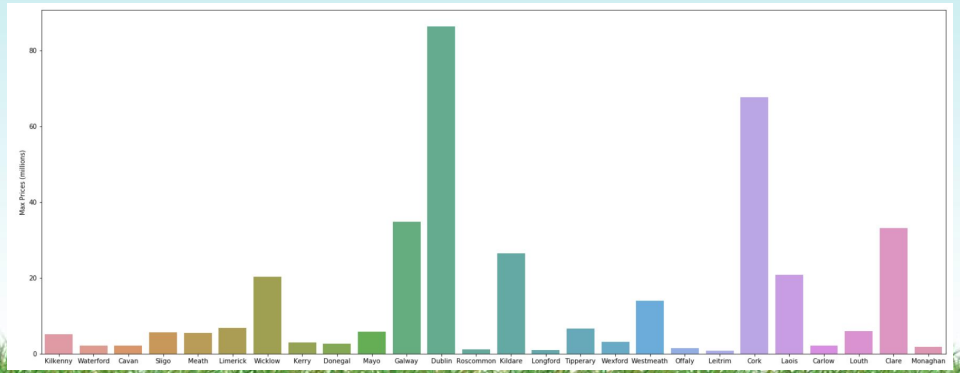
- This also helped to resolve incomplete addresses & misspellings like: "3 Cois Chnoic, Tirkeenan, Momaghan"
- To: "3 Cois Chnoic, Connolly Park, Monaghan, H18 F201"

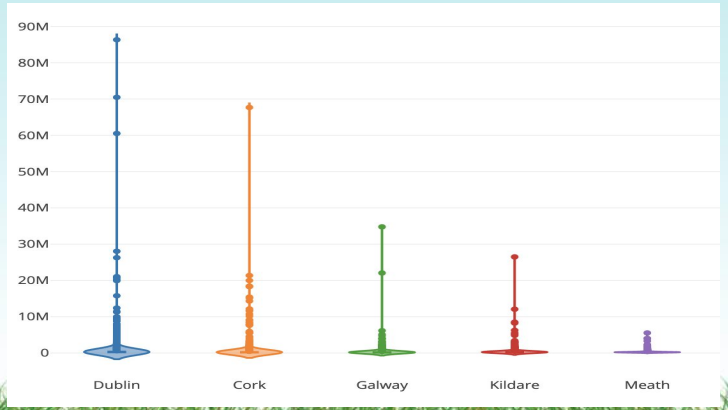
Discrepancy Detection(3)

2) Identify outliers / anomalies









Discrepancy Detection(3)

2) Identify outliers / anomalies

- Example: we can see that the max property price in Dublin was 76 million. Yet looking at the address:

```
df[df['price'] == df[df['county'] == 'Dublin']['price'].max()]['address'].values
```

We get: 'Block F, K And L, Central Park, Leopardstown' - which are commercial office spaces.

- Other massive outliers in the data turnt out to be complete apartment blocks too.

Cork: '1 The Elysian, Eglinton Street, Cork' -> Entire Apartment Block

Galway: 'Student Accomodation Known As, 'Cuirt Na Coiribe Headford Road, Cuirt Na Coiribe, Headford Road'

Kildare: 'Castlemartin House, Kilcullen' -> Georgian Mansion

Meath: '1-26 Blackcastle Manor, Slane Road, Navan' -> Residential Development

Monaghan: 'As caill Rois, Carrickmacross' -> Housing Estate

Discrepancy Detection(4)

3) Identify Inconsistent Data

- checked the date format for each entry:

```
df['date_of_sale'] =
pd.to_datetime(df['date_of_sale'],
dayfirst=True, format='%d/%m/%Y')
```

- Irish vs English descriptions

```
In [8]: df['property_description'].unique()

Out[8]: array(['Second-Hand Dwelling house /Apartment',
               'New Dwelling house /Apartment', 'Teach/Árasán Cónaithe Atháimhe',
               'Teach/Árasán Cónaithe Nua', 'Teach/?ras?n C?naithe Nua'],
              dtype=object)
```

- Not all dublin addresses had postal codes.

4) Identify Field Overloading

- When two or more separate concepts are being used in a single data field
- Area Code, Town & Address

Before:

- Address = "22 Laverna Way, Castleknock, Dublin 15"

After:

22 Laverna Way, Castleknock, Dublin 15	Dublin	355000.0	No	Second-Hand Dwelling house /Apartment	NaN	Castleknock	Dublin 15	D15 KC8C	53.373652	-6.383476

Conclusions

- **Unexpected findings?**
  - Discovered addresses in Dublin files outside of the county, like:

```
“30 Cnoc Tiarnach, Grange End, Dunshaughlin”,,Dublin

"11 Wingfield, Corke Abbey, Bray", Dublin 18, Dublin
```

  - A lot of human errors, misspellings & null values for particular columns
  - Originally may not have been suitable for machine learning, lack of numerical attributes.
- **Future Plans?**
  - Utilize lat & lon coordinates to extract other data (distance to public transport stops, city centre)

Thanks for Listening!