Team Sherlock Homes

W210 Capstone Final Presentation

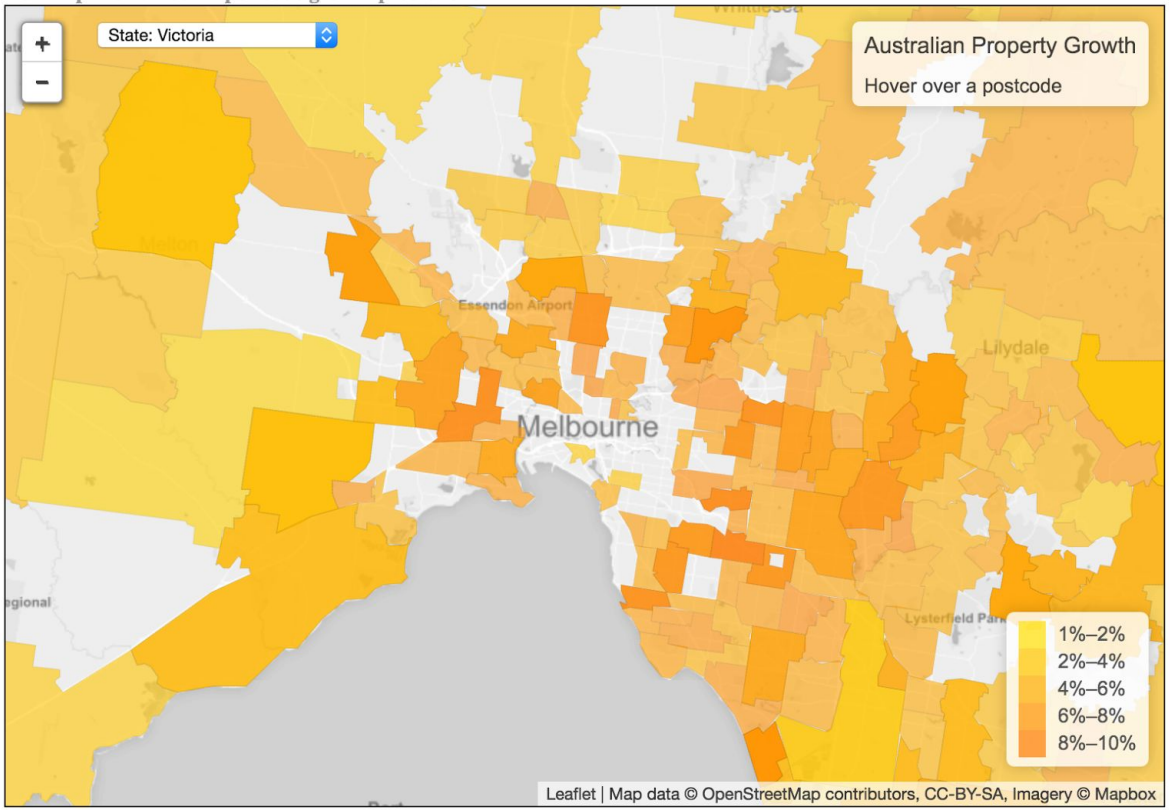
August 11, 2016

Thomas Atkins, Megan Jasek, Charles Kekeh, Eric Whyne

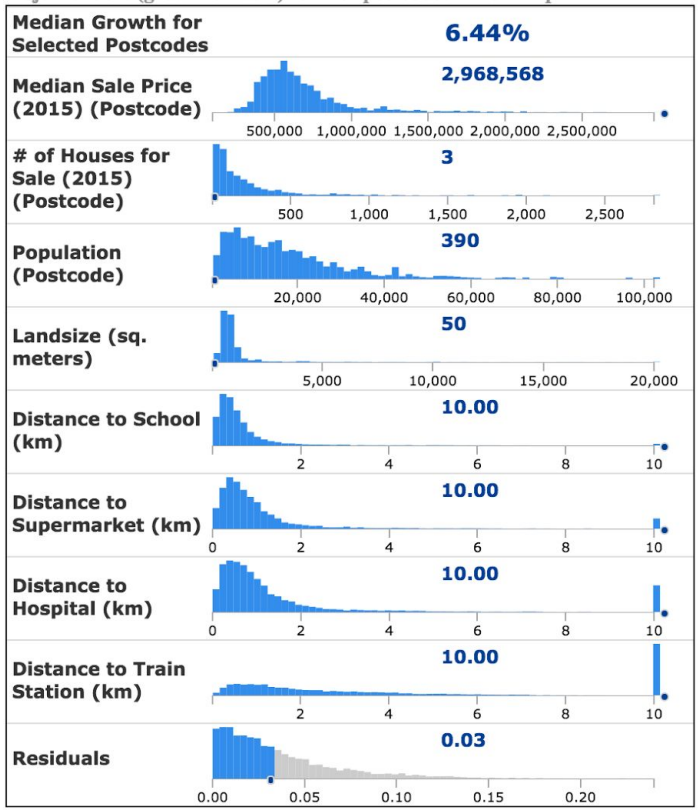
Investigating Home Prices in Australia

About the science

Heatmap shows median postcode growth predictions



Adjust sliders (growth drivers) to filter postcodes in heatmap



Click on a row to display a property on the map. Click on a column header to sort properties by that column.

Street	Postcode	Suburb	State	Beds	Baths	Cars	Landsize	Growth %	Current Value $
				>= 1 x	>= 1 x	>= '1' x	>=		

Scope, Impact, and Rationale

Problem Statement: How can we make home buying in Australia (in the states of NSW, VIC and QLD) less emotional and more profitable?

Proposed Solution: Create a tool that

- Predicts 1-year annualized growth of a property in Australia (in NSW, VIC and QLD)
- Displays confidence interval for the prediction
- Displays factors that are most important to property growth

Impact: Home buyers will be more confident in their home purchase decision which will reduce stress levels and improve quality of life. Home buyers might also experience increased profits because they know some of the factors that go into growth.

Customer Segment: Home buyers in the NSW, VIC and QLD states of Australia

Challenges

The data is:

- Messy
- Proprietary
- Contradictory
- Inaccurate
  - (i.e. time deltas)
- Delayed
- Missing

Our users are:

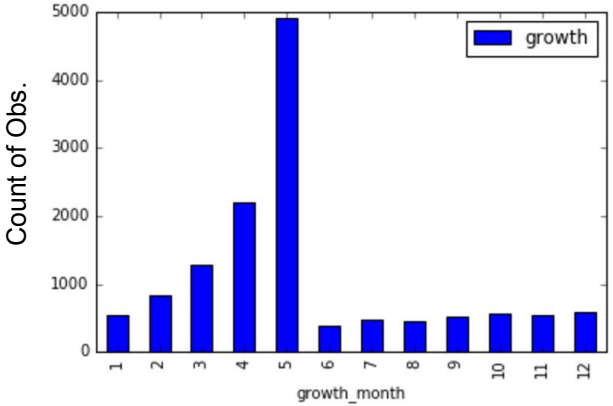
- Varied
- Unknown
- Stressed out
- Missing variables

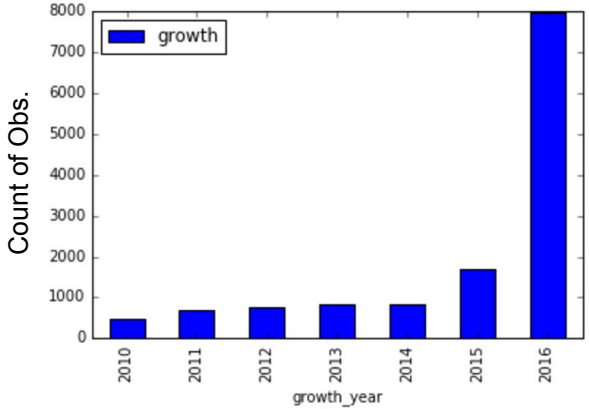
The math is:

- Not intuitive
- Difficult to explain
- Prone to wild errors
- Missing variables

Seasonality Issues:

- Uneven sampling over months and years





Home Ownership in Australia

70% of homes in Australia are owner-occupied, one of the largest proportions of any country.

Increases in owner-occupied property values are not subject to capital gains tax. This helps to make home ownership a smart investment.

A third of Australian homes are owned by their owners outright, about another third are still paying mortgages. (The last 30% are renting.)

Home loan interest rates are at all time lows of ~4% encouraging further investment into the property market.

Impact: Example Calculation

Australian National Average home

6% growth

An area found by our model (postcode 3129) is showing

10% growth

Amount better than average

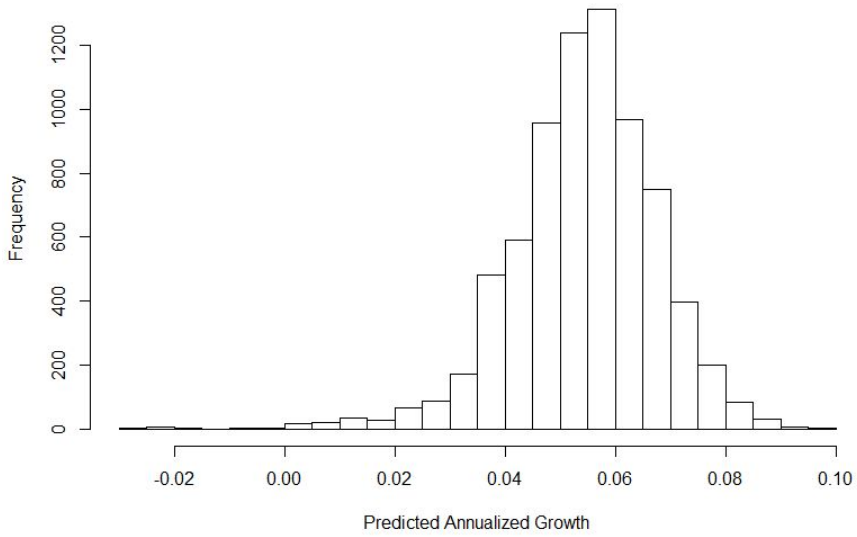
10% - 6% = 4% growth

The median home value in postcode 3129

$1.0 million

Extra 4% growth on a home in that area would mean

extra $40K a year



As a comparison, Term Deposits with local banks are currently only returning ~3%

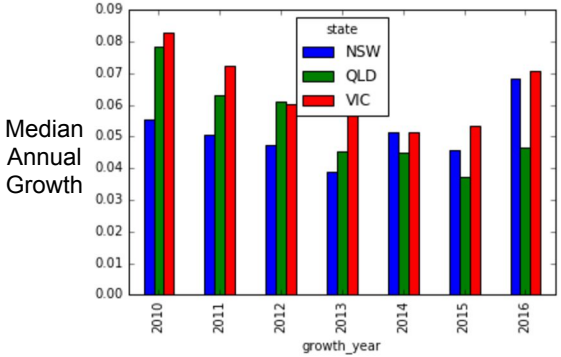
Data Description

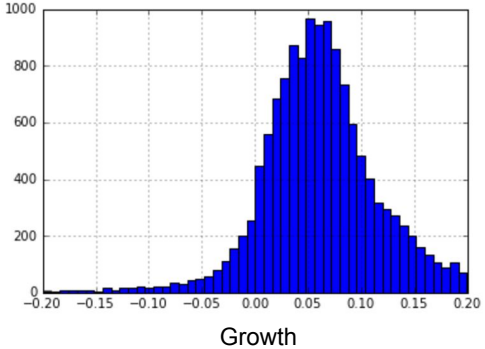
166 variables initially collected and analyzed for each property at the postcode and property level

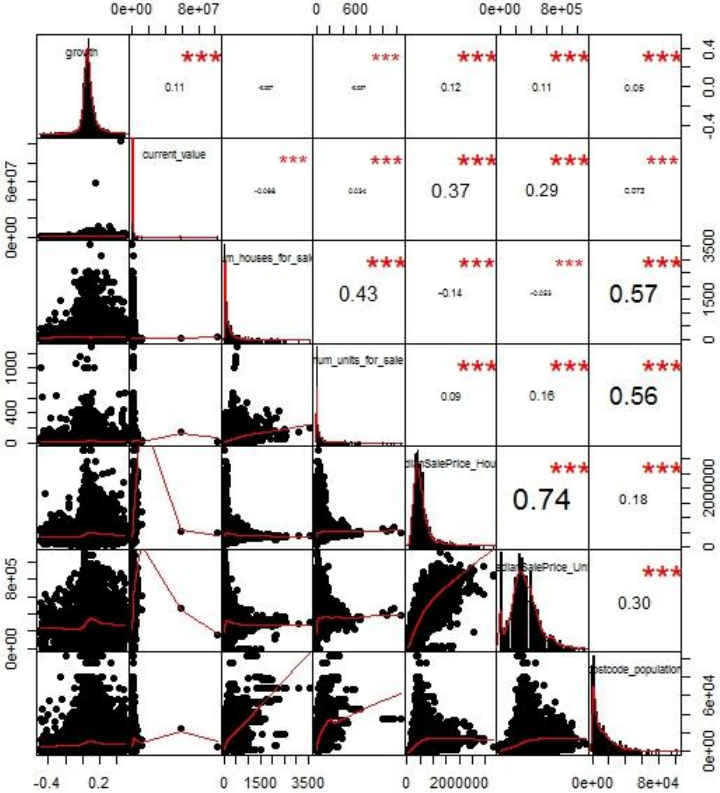
Property - # Beds, Baths, Cars, Distance to Features
Postcode - Historical Trends, Business Features

Data variance in both temporal and spacial dimensions

Sales History used to calculate Compound Annual Growth Rate (CAGR) which is our target variable







Modelling

Model Type: Gradient Boosted Regression Trees (sklearn.ensemble.GradientBoostingRegressor)

Feature Selection:

- Initially include all individual property level features
- Check importance scores from GBRT (gbm.feature_importances_)
- Drop features with close to zero importance
- Tune model parameters and repeat

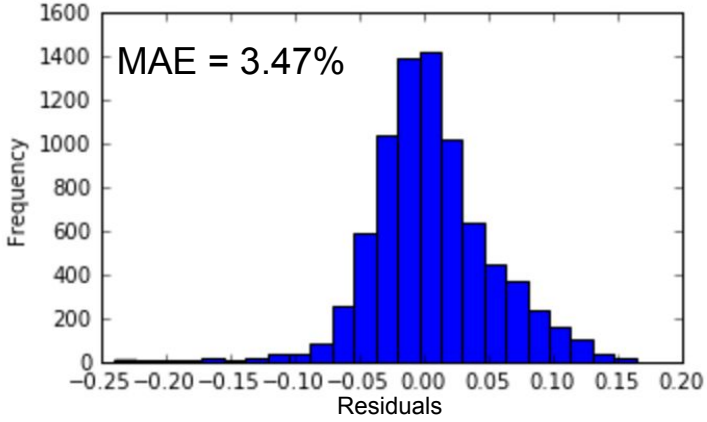
Model Tuning:

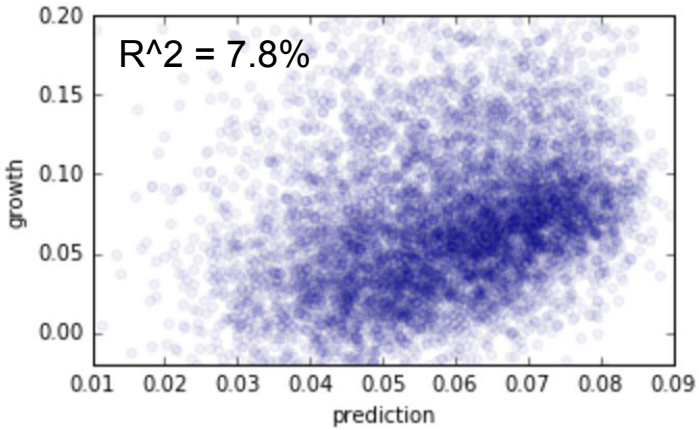
- Tune: loss function, learning rate, n_estimators, subsample, min_weight_fraction, max_depth
- Multiple Grid searches over a range of values for each tuned parameter, update with optimal then continue to optimise other variables
- Optimised: R^2, MAE

Performance & Evaluation:

- Train model on growth from 2009 to 2015
- Generate predictions of growth for 2016 (Prediction Window up to 6 months)
- Ensure that future data is not used to train model - information leakage!

Residuals & Predicted v Actuals





- R^2 is fairly low
  - Predictions not precise
  - A property sale is often an irrational human event

- LR model found a number of interesting variables to be statistically significant

70% of predicted growth >6% had actual growth >6%

Model	R^2	MAE
GBRT - 2 Month Prediction Window	12%	3.3%
GBRT - 6 Month Prediction Window	7.8%	3.47%
RF - 6 Month Prediction Window	5.6%	3.5%
LR - 6 Month Prediction Window	3.6%	4.2%

Drivers of Growth

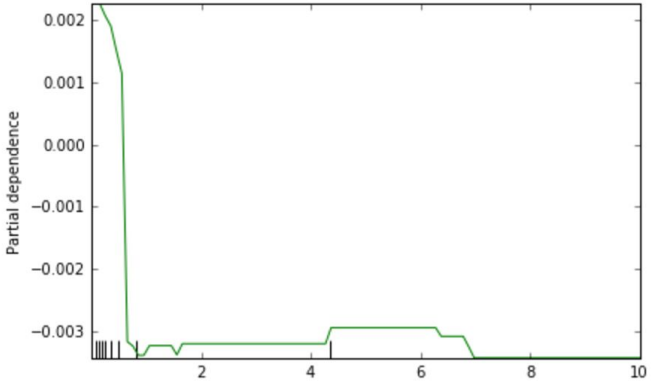
gbm.feature_importances_

feature	imp
growth_year	0.088807
num_houses_for_saleprev	0.078039
landsize	0.061173
MedianSalePrice_Unitsprev	0.052566
hospital_distance	0.048761
MedianSalePrice_Housesprev	0.048374
train_station_distance	0.043676
bus_stop_distance	0.043253
supermarket_distance	0.039473
school_distance	0.037151
flat_count	0.036586
building_count	0.036072
MedianSalePrice_3bedHousesprev	0.033418
address_count	0.031818

Supply of properties for sale is an important feature. **Less availability = More growth**

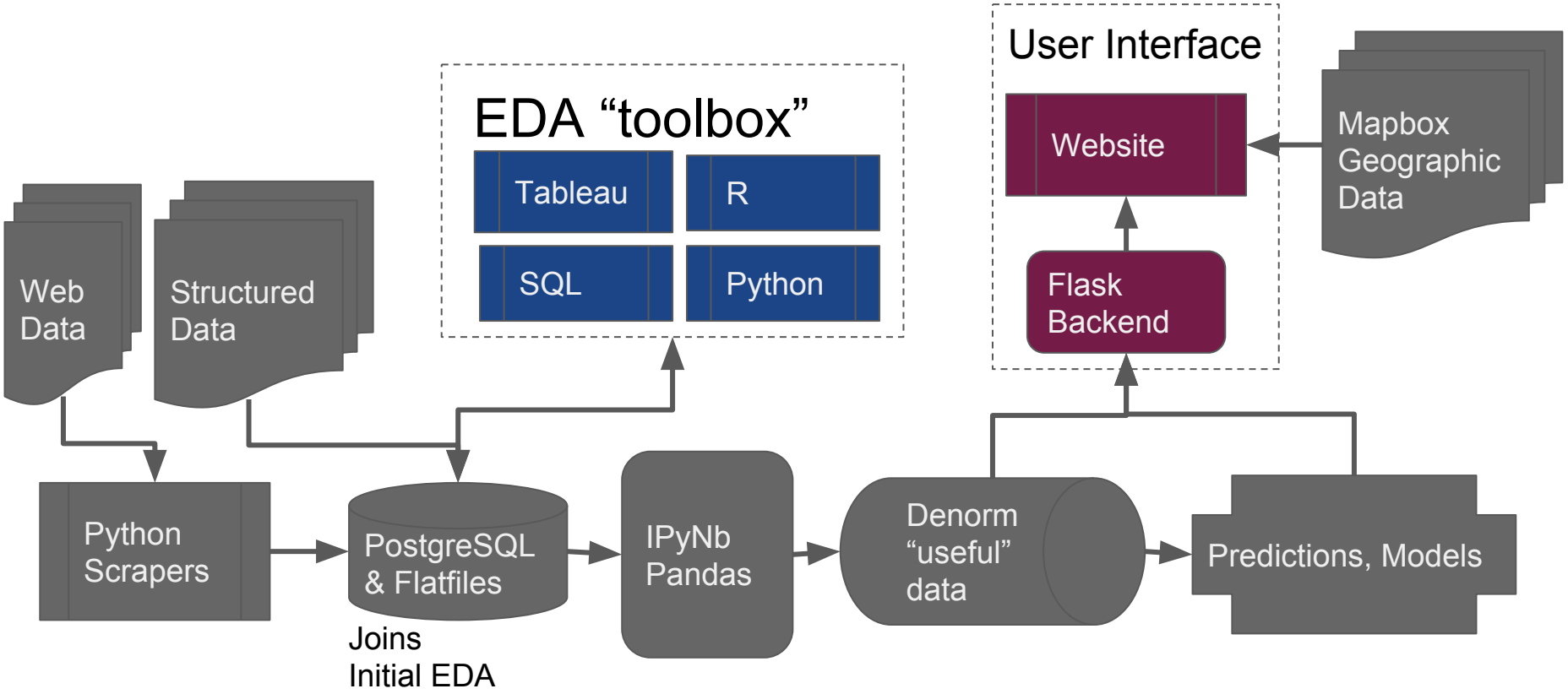
Proximity to transport is highly important

Partial Dependence Plot



Distance to Bus Stop (km)

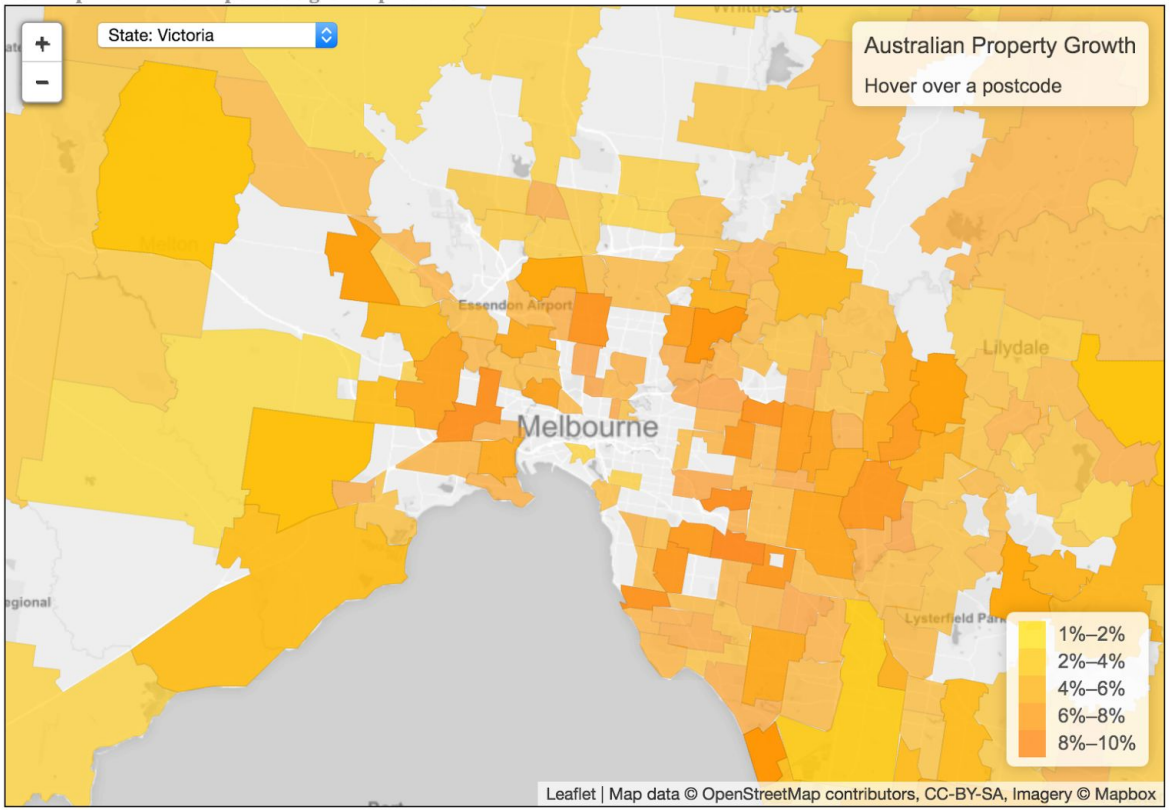
Architecture and Tools: Where we are now



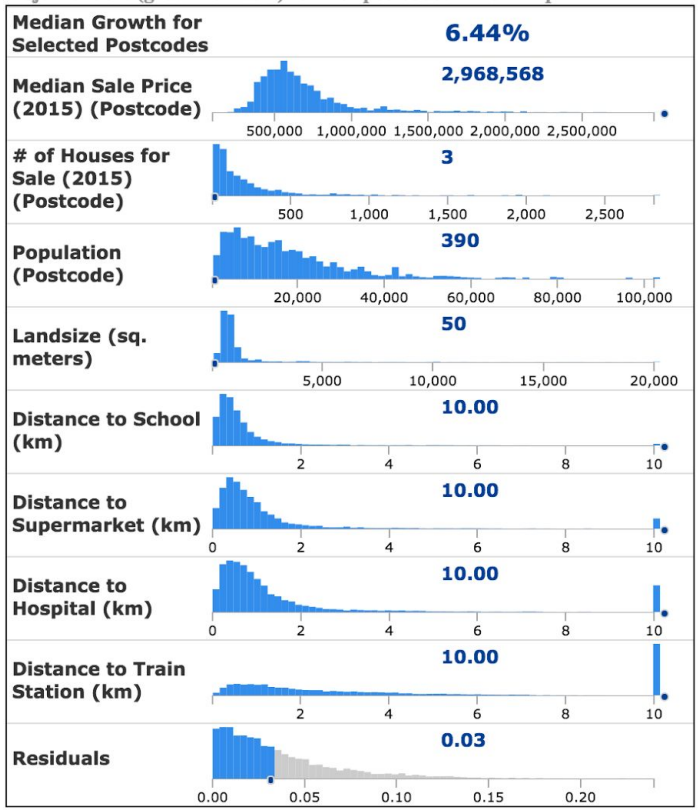
Investigating Home Prices in Australia

About the science

Heatmap shows median postcode growth predictions



Adjust sliders (growth drivers) to filter postcodes in heatmap



Click on a row to display a property on the map. Click on a column header to sort properties by that column.

Street	Postcode	Suburb	State	Beds	Baths	Cars	Landsize	Growth %	Current Value $
				>= 1 x	>= 1 x	>= '1' x	>=		

Usability Testing

Informal testing

- Users were shown UX and asked to give feedback

Who participated

- Andy Reagan (instructor for MIDS data visualization course)
- Tom’s boss, Tom’s wife, Megan’s friend

Results - Highest Priority Items

- Move map up (make it prominent), move sliders to right side
- Use small histograms for the slider variables
    - When sliders move across histogram, gray values out
- Display model uncertainty
- Add instructions over map and sliders (not clear what sliders do)
- Create logo around Sherlock Holmes

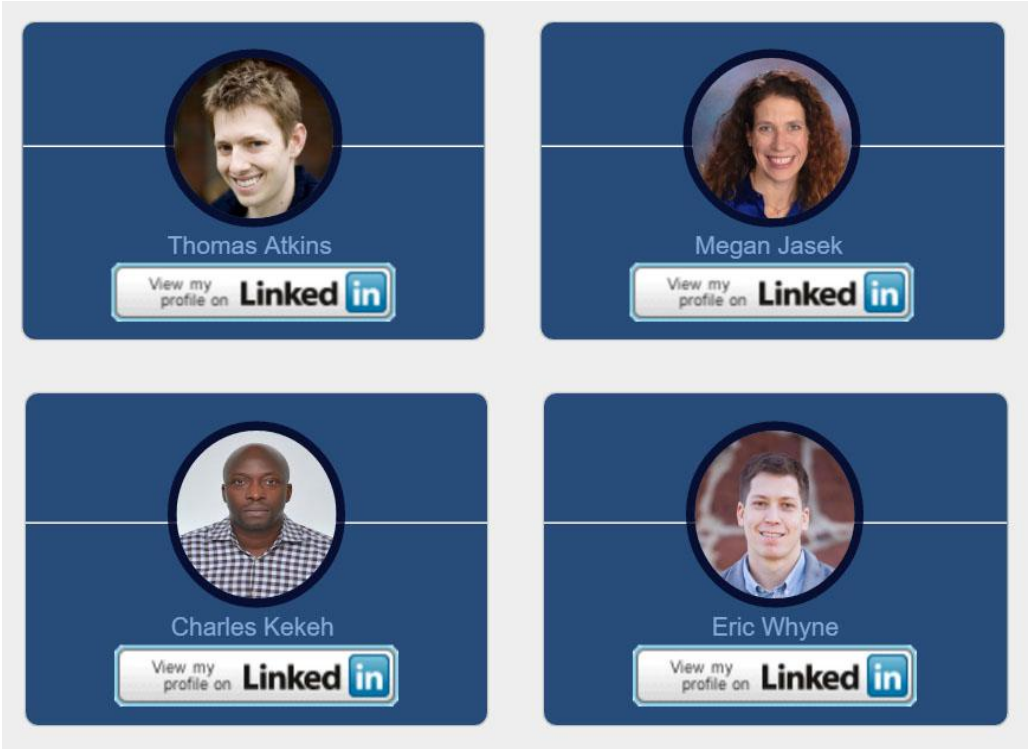
Response

- Full list of prioritized items in the Appendix
- Almost all items fixed

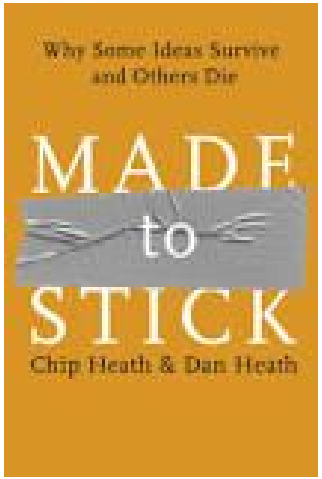
Next Steps

- **Model**
  - Use of more advanced modeling techniques - neural networks, ensembles
  - Create model at postcode level
  - Make future predictions instead of historical predictions
  - Collect more property level data - distance to workplace
  - Consider converting target variable to categorical with ranges of growth
  - Create sensitivity analysis - deeper analysis of what is causing uncertainty in the model
- **User Interface**
  - Home search - enter hypothetical values into sliders → create real time prediction → search for properties close to those values
  - Comparison - compare one or more postcodes side by side
  - Drill down into a postcode - create finer levels of granularity in UX

Contact Us


Thomas Atkins
View my profile on LinkedIn
Megan Jasek
View my profile on LinkedIn
Charles Kekeh
View my profile on LinkedIn
Eric Whyne
View my profile on LinkedIn

Fun Fact



Sherlock Holmes never actually said "Elementary, my dear Watson" in any of the books.

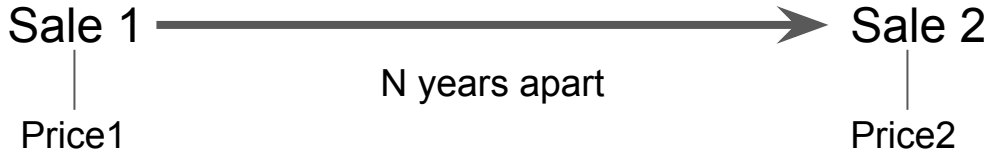
Appendix: EDA: Feature List

166 variables initially collected and analyzed for each property at the postcode and property level.

Feature Class	Feature List
Property-level features	# Bedrooms, # Bathrooms, # Car Spaces, Land size, School distance , Supermarket distance, Hospital distance, Train station distance, Bus stop distance, Number of parking spots, growth year, Distance to airport, Growth (CAGR), Growth Year, Growth Month, and Current value.
Postcode-level features	Medium sales price (previous, current, and YoY), Number of units for sale (previous, current, and YoY), Number of houses for sale (previous, current, and YoY) Building count, Postcode population, Number of flats, Streets count, Restaurants, Train stations, Parks, Number of businesses in the postcode across 60+ business categories.

Appendix: Modelling: Deriving a Target Variable

Sales history divided into adjacent pairs of sales



Compound Annual Growth Rate (CAGR)

If this were an investment that grew at a constant % growth rate compounding annually, the CAGR is the rate that achieves the price change from Price1 to Price2.

$$CAGR = \left(\frac{Price2}{Price1}\right)^{\frac{1}{N}} - 1$$

Appendix: UX: Final Iteration

- Use predictions over 2016 held-out dataset to power the UX
- Use geocoding to display the retained test set
- Data served to the UX by a Flask web service
- Allow filtering of properties based on drivers of growth
  - Median Growth for Selected Postcodes
  - Median Sale Price (2015) (Postcode)
  - # of Houses for Sale (2015 (Postcode)
  - Population (Postcode)
  - Landsize (sq. meters)
  - Distance to School (km)
  - Distance to Supermarket (km)

Appendix: Usability Testing - MoSCoW Prioritization

MUST Fix

- Move map up (make it prominent) and put sliders on right side
- Use small histograms for the slider variables
- Display model uncertainty somehow
- When move sliders create graying
- NOT CLEAR what the sliders do (Added additional labels)
- Instructions over map and sliders. How model works.

SHOULD Fix

- State and postcode inputs could go on top of map
- Update growth Label: 'Projected property growth for selected postcode'
- Create logo - Sherlock Holmes icon. Magnifying glass over house. Hat. glass.
- Could do a walkthrough of the visualization using http://introjs.com/
- Use gray background for map and sliders
- Update heatmap colors to be more distinct
- Add scale for heatmap
- Format variables names so they display nicely

COULD Fix

- Use Suburb Names rather than Postcodes
- Histogram color: use a complimentary color for yellow/red. Only use one color (same color for starting at min or max) and use text to explain it.
- App is slow to respond (User located in Australia, Webserver in USA)
- Do usability testing with actual people

WON'T Fix

- When you hover over a zip code, give historical info and/or graph of growth
- Feature: could add distance from where you work
- Use population density rather than overall population (not addressed - next steps)
- Feature: It would be great to COMPARE the neighboring postcodes to each other
- Add 2nd slider to create minimum and maximum values