Assignment 2: Real World Data Handling, Modelling & Visualisation

COMP309: Charlene Leong 300322129

Introduction

The aim of this assignment is to investigate and analyse the reasons for the rising rent prices in Wellington and therefore validate or invalidate the claims in the following four Stuff articles. Relevant datasets were selected, preprocessed and merged in order to extract valuable insights to allow us to illustrate evidence that either supports or dismisses the Stuff article headlines.

News Article Analysis Summary

The following Stuff articles reporting on Wellington's rent prices were skimmed for insights to inform what datasets to choose and trends and predictions that would be useful to model. Points of interest have been *italicised* and corresponding possible attributes of interest **boldened**.

Wellington's skyrocketing rental prices are not an issue in some of the provinces
Stuff.co.nz 10:51, Jan 25 2018

This article revolves around Hannah and Kaleb Heberley's story of leaving Wellington with their three children for a significantly cheaper $535 four-bedroom seaside rental in central New Plymouth with relative market value to that of Oriental Bay in Wellington. Additional to housing, living costs including daycare and fuel are also cheaper. However, there are *limited number of jobs* in New Plymouth. However despite this, the Rental Division Manager at McDonald Real Estate, said she has definitely noticed the increase in people moving from central hubs to New Plymouth due to *inflating prices and increasing competition in Wellington.*

Therefore, it would be useful to include **employment statistics by region**, **population density by region** and **employment to population ratio by region** alongside relation to **average rent prices by region** in the ML model in order to validate the trends in this article.

Wellington rental prices match Auckland as property listings plummet
Stuff.co.nz 21:44, Feb 26 2018

This article states that *Wellington rents has caught up with Auckland*, with the median price reaching $550 per week in both cities in January. Trade Me figures suggested the parity was reached after rents hit record-breaking highs nationwide, with an average increase of 4.4 per cent to median rents, putting the average at $470 per week. Additionally, there was an 11 per cent *fall in the number of rental listings* – a drop that more than doubled the national decline of 5 per cent. Lower Hutt rents were up 13 per cent in January to $450 a week, and median rent in Upper Hutt increased 2 per cent to $400 a week hinting that more people are *moving further out due to a decrease in available rentals* in Wellington city. Student hotspots like the Wellington suburbs of Aro Valley and Te Aro received 16 and 12 inquiries on average, respectively, in the first two days, and there are early signs that students are increasingly ditching their usual hunts to *move further out.* Median rent prices in student suburbs were reported to be per week -

Kelburn	595
Mt Cook	615
Mt Victoria	550
Newtown	620
Te Aro	480
Wellington Central	460

Therefore, it would be useful to compare the rent prices of suburbs in Wellington with **proximity to tertiary institutions** and the **number of rental listings by suburb** in order to support this article's claims of students moving further out due to decrease of available rentals in Wellington city leading to increasing rent prices.

Wellington's rental prices cool, national rent stays at record high
Stuff.co.nz 08:02, Mar 26 2018

This article reports that Wellington's rental market has recovered from the *surge of students* looking for housing, Head of Trade Me Property Nigel Jeffries said. "Typically we see rental prices in the capital peak in December and January when students come into the city looking for a flat, before easing back in February and March." However, rents are still up 6.5 per cent compared to last year. Jeffries said demand was still very *high in the region without the supply* to meet it. The *number of rentals is slowly increasing* in Wellington, up 2 per cent since last year, Jeffries said.

Therefore, it is predicted that there will be a peak in Wellington's rental prices around December and January with a general upwards trend year after year. It will be useful to compare **student enrolment figures** and **student allowance and living costs** to see if they are contributing to price increases in the Wellington rental market.

Rents rise around the country as cost crunch goes on landlords
Stuff.co.nz 14:07, Apr 16 2018

This article aims to reports that extra costs for landlords are being blamed for rent rises around the country. Gareth Kiernan, of forecasting firm Infometrics informs that, "*Gradual rises in interest rates* will maintain upward pressure on rents during 2018 and 2019, and landlords are also likely to try and recoup increased costs associated with tougher standards for rental properties, including improved insulation requirements. The extension of the bright-line test for capital gains from two to five years will limit the willingness of new investors to enter the market, thereby leading to further increases in rent." However, economist Shaumbeel Eaqub said *how much demand there was for rental properties, the number of properties available for rent, and how much people could pay,* were much more important factors than landlord costs. Additionally, Andrew King of the NZ Property Investors Federation- "In Wellington rent went up a lot not because landlords' costs went up but because of the squeeze caused by the *earthquakes*, and an *increase in student numbers.*"

Therefore, additional to previously identified datasets for further investigation such as the **number of rental listings by suburb** (number of properties available for rent), **student allowance and living costs and employment statistics** (*how much people could pay),* **student enrolment figures** (*increase in student numbers)*, there is still room to investigate the correlation between **mortgage interest rates** and rising rent prices as well as history of **strong earthquakes in Wellington.**

Part 1: Evidence related to rental costs in Wellington

Selected Datasets

The goal is to build a pipeline to analyse attributes of the Wellington rental market compared to other regions. Therefore any time-series datasets with regional data not only informing rent prices but also average income of the demographic, supply and demand of the rental and housing market and trends in renting population are all relevant and useful in building a predictive model of Wellington's rental prices.

MBIE Rental Bond Datasets

All the following datasets show monthly data from Jan 1993 to July 2018.

Geometric mean rents by region	Monthly - Jan 1993 to July 2018	Geometric mean rents by region take into account the compounding that occurs from period to period and therefore is less affected by outliers in the dataset compared to the arithmetic mean.
Geometric mean rents by territorial authority	Monthly - Jan 1993 to July 2018	Geometric mean rents by territorial authority allows us to break down the regional data from the dataset above more granularly.
Synthetic lower quartile rents by region	Monthly - Jan 1993 to July 2018	Synthetic lower quartile of rents are relevant because students are more likely to rent in the lower quartile of rental prices.
Synthetic lower quartile rents by territorial authority	Monthly - Jan 1993 to July 2018	Synthetic lower quartile of rents by territorial authority allows us to break down the regional data from the dataset above more granularly.
Lodged bonds by region	Monthly - Jan 1993 to July 2018	Lodged bonds by region are relevant because they give an indication of the demand of rental properties on the market.
Lodged bonds by territorial authority	Monthly - Jan 1993 to July 2018	Lodged bonds by territorial authority allows us to break down the regional data from the dataset above more granularly.

Stats NZ Datasets

Regional GDP 2000-2016	Yearly - Mar 2000 to Mar 2016	It is hypothesised that regional GDP is highly correlated with rental price, therefore would be useful to see if this a contributor to rising rental prices.
National Population Estimates	Yearly - Jun 2008 to Jun 2018	One of the biggest stated reasons for rental price increase has been not having the supply to meet demand, therefore it will be useful to look at the population trends over time, especially the 20-29 age group which is the age group most likely to be renting.
Median Weekly Earnings from wages and salaries	Yearly - 1998-2018	It is hypothesised that median weekly earnings will not have increased at the same rate as that of rental prices to validate reason for students moving further out.
Building consents by region	Quarterly - Jun 1990 to Mar 2018	It is hypothesised that residential building consents in Wellington will not have increased at the same rate as the renting population to validate reason for supply not meeting demand.

Education Counts Datasets

Provider Summary Tables 2016	Yearly - 2008-2016	Statistics on enrolments, completions and EFTS by different characteristics for each provider for both international and domestics students will highlight the demand for student housing in each area. Additionally, It has been noted in the media that the international students choose to study in Wellington over other parts of the country [1] along with the increasing number of international students that choose to study in NZ. Therefore, it will be useful to see whether a rise in international students matches the rise in rent prices.

Immigration NZ Datasets

Student Visa Applications	Yearly - 2008-2018	Statistics on decided student visa applications is another useful attribute to consider that may contribute to the increased demand vs supply.
Work Visa Applications	Yearly - 2008-2018	Statistics on decided work visa applications is another useful attribute to consider that may contribute to the increased demand vs supply.

Ministry of Social Development Datasets

Amounts paid out for Student Allowance and Accommodation Benefit payments	Yearly - 1999 to 2017	It has been noted in the media that the recent increase of $50 in student allowance under the Labour government may have led to the rent spike in 2018 [2]. Unfortunately, there is no data for 2018 yet so it hard to confirm this correlation as Wellington rent prices traditionally spike around this time period regardless [2]. However it will be useful to see if this has historically happened before and validate whether this is a contributing factor to rising rent prices in 2018.

Reserve Bank NZ Datasets

Mortgage Interest Rates	Monthly - Jun 1998 to Jun 18	It is hypothesised that rising mortgage interest rates are putting pressure on landlords to increase their rent prices. Comparing this dataset with rental prices will be useful in evaluating if there is a correlation.
Inflation	Monthly - Jun 1991 to Jun 2018	Inflation is the term used to describe a rise of average prices through the economy indicating whether means that money is losing its value. It will be useful to see if the rental prices are rising at a rate higher than that of inflation. If so, that indicates rental prices are rising due to other factors.

Integration of Datasets and Feature Manipulation

Datasets were merged into a format suitable for building a prediction model with linear regression. From observing the availa ble scope and granularity of our datasets to be merged, it will be suitable to limit our model the last ten years of data Jan 200 8 - Jun 2018 and model our trends in a quarterly timestep. This will require downsampling monthly datasets and interpolating yearly datasets in order to impute missing values.

MBIE Datasets

Due to the all the selected MBIE datasets having the same format that shows the regions by month (Fig 1), it was simpler to implement a pipeline in Python to merge all the MBIE datasets into one dataset. Each dataset was transformed so that the Regi on and Mean rents by region are added as attributes (Fig 2). After this transformation, each dataset was appended together into one combined dataset describing rental data by region (Fig 3).

	Month	Auckland	Bay of Plenty	Canterbury	Gisborne	Hawke's Bay	Manawatu-Wanganui	Marlborough
0	1993-01-01	1909	411	849	60	230	504	71
1	1993-02-01	2291	473	1196	97	237	649	63
2	1993-03-01	2721	560	1325	74	257	824	49
3	1993-04-01	2425	435	982	82	281	530	55
4	1993-05-01	2421	533	1089	92	274	594	74

Fig 1. Geometric Mean Regions by Month RAW

	Month	Region	Mean rents by region
0	1993-01-01	Auckland	196
1	1993-02-01	Auckland	196
2	1993-03-01	Auckland	193
3	1993-04-01	Auckland	195
4	1993-05-01	Auckland	200
5	1993-06-01	Auckland	194
6	1993-07-01	Auckland	203
7	1993-08-01	Auckland	200

Fig 2. Geometric Mean Regions by Month REFORMATTED

	Month	Region	Territorial Authority	Mean rents by region	Mean rents by territorial authority	Geometric mean rents by region	Geometric mean rents by territorial authority	Synthetic lower quartile rents by region	Synthetic lower quartile rents by territorial authority	Lodged bonds by region	Lodged bonds by territorial authority
252	2014-01-01	Canterbury	Ashburton District	390.0	319.0	357.0	311.0	267.0	266.0	1635.0	59.0
253	2014-02-01	Canterbury	Ashburton District	400.0	315.0	365.0	308.0	271.0	264.0	1882.0	78.0
254	2014-03-01	Canterbury	Ashburton District	397.0	302.0	361.0	292.0	266.0	243.0	2011.0	76.0
255	2014-04-01	Canterbury	Ashburton District	401.0	307.0	364.0	297.0	270.0	247.0	1414.0	64.0
256	2014-05-01	Canterbury	Ashburton District	401.0	310.0	367.0	300.0	276.0	245.0	2098.0	90.0

Fig 3. Rental Data by Region and Territorial Authority 2008-2018

[1] International students choosing to study in Wellington over other parts of country - https://www.stuff.co.nz/national/education/89752366/international-students-choosing-to-study-in-wellington-over-other-parts-of-country

[2] Student allowance boost blamed for rent spikes - https://www.stuff.co.nz/national/politics/100485600/student-allowance-boost-blamed-for-rent-spikes

The same was done for rental data by territorial authority datasets. A lookup table linking region to territorial authority was created (see A.1) [1] in order to merge the regional and territorial authority datasets together to produce a combined dataset that has mean rents, geometric mean rents, synthetic lower quartile rents and lodged bonds into by region and territorial authority (Fig 3). The dataset was then downsampled from a monthly to quarterly time period. This rental data by region and territorial authority dataset is now the master combined rental dataset where features from new datasets will be transformed into a format compatible with this master dataset and appended as new attributes.

Due to the irregular format of the other datasets, it will be easier to transform features into attributes and append them manually using Excel to our MBIE combined rental dataset.

Regional GDP: Yearly - Mar 2000 to Mar 2016

From the raw dataset (Fig 4), the features of interest are the region and annual GDP figures per year from Mar 2000-2016. The raw table was cleaned and reformatted into a lookup table where quarterly GDP values were then interpolated from annual values by (see A2.1). GDP figures for 2017 and 2018 quarters were also interpolated. A Jupyter notebook *Lookup Table Transform* was then written to reformat the table to fit the combined rental dataset (see A2.2) using the INDEX MATCH formula.

	Series	Year ended March																
	ref:	2000	2001	2002	2003	2004	2005	2006	2007R	2008R	2009R	2010R	2011R	2012R	2013R	2014	2015P	2016P
Region	RNAA	$(million)																
Northland	SG01NAC34B0101ZZB	2,854	3,171	3,420	3,317	3,570	3,942	4,278	4,706	4,993	5,102	4,897	5,250	5,365	5,331	5,796	5,971	6,207
Auckland	SG01NAC34B0102ZZB	40,539	41,705	44,888	49,260	53,178	57,355	60,175	62,296	66,880	66,010	68,468	71,360	75,288	77,756	81,993	88,252	93,541
Waikato	SG01NAC34B0103ZZB	9,224	10,426	11,366	10,957	11,882	12,570	13,492	14,969	15,871	16,716	16,369	17,362	18,570	18,115	20,122	20,146	20,940
Bay of Plenty	SG01NAC34B0104ZZB	5,837	6,273	6,693	6,893	7,332	7,862	8,261	9,036	9,798	9,694	10,085	10,745	11,208	11,408	11,906	12,137	13,071
Gisborne	SG01NAC34B0105ZZB	968	1,014	1,058	1,085	1,098	1,156	1,172	1,287	1,331	1,409	1,449	1,533	1,583	1,587	1,639	1,715	1,759
Hawke's Bay	SG01NAC34B0106ZZB	3,640	3,910	4,165	4,401	4,702	5,055	5,190	5,376	5,272	5,486	5,577	5,929	6,082	6,254	6,591	6,648	6,817

Fig 4. Regional GDP RAW

National Population Estimates: Yearly - Jun 2008 to Jun 2018

The features of interest from the raw data (Fig 5) are the year and the number of people in the 15-39 age group which is the age group most likely to be renting. Quarterly figures were interpolated these annual figures in a similar fashion to the Regional GDP dataset (see A3.1). Piece-wise interpolation was performed due to the non-linear nature of the moving average trendline.

At 30 June		Age group (years)					Median age (years)
		All ages	Under 15	15–39	40–64	65+	
Total							
2008		4,259,800	895,500	1,460,400	1,368,800	535,000	36.4
2009		4,302,600	901,100	1,461,000	1,392,100	548,300	36.7
2010		4,350,700	908,100	1,464,600	1,414,400	563,500	36.9
2011		4,384,000	910,700	1,459,100	1,434,100	580,100	37.1
2012		4,408,100	909,800	1,450,500	1,444,700	603,000	37.4
2013		4,442,100	908,800	1,452,300	1,455,000	626,000	37.6
2014		4,509,700	911,100	1,481,100	1,467,100	650,400	37.5
2015		4,595,700	914,300	1,528,500	1,478,500	674,300	37.3
2016		4,693,200	921,600	1,583,400	1,489,800	698,400	37.1
2017		4,793,900	933,700	1,634,900	1,502,200	723,100	37.0
2018	P	4,885,300	944,600	1,680,800	1,513,100	746,900	36.9

Fig 5. National Population Estimates 15-39 age group RAW

Median weekly earnings: Yearly - 2008-2018

The raw dataset (Fig 6) was extracted from Stat NZ's Infoshare service showing median weekly earnings by region for the 20-24 and 25-29 age group. This dataset reformatted into a lookup table with median weekly earnings averaged between the two age groups. Unlike previous datasets whose missing values were imputed through linear interpolation, trends in median weekly earnings by regions fluctuate too rapidly for linear interpolation to give accurate quarterly figures. Therefore, imputation methods such as K Nearest Neighbours, multiple imputation (MCMC) algorithm and the NIPALS algorithm were explored in imputing quarterly figures where the NIPALS method proved to give the most promising results (A4.1). These figures were then added to the master combined rental dataset.

Region	Northland Region		Auckland Region		Waikato Region		Bay of Plenty Region		Gisborne/Hawkes		Taranaki Region		Manawatu-Wanganui		Wellington Region		Nelson/Tasman/Marl		Canterbury Region		Otago Re	
Sex	Total Both Sexes																					
Age Group	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	25 to 29	20 to 24	
Ethnic Group	Total Ethnic Groups																					
Measure	Median Weekly Income																					
Year																						
1998	322	372	364	480	392	430	318	416	357	418	380	363	250	354	342	537	360	418	310	439	143	
1999	306	320	360	518	305	450	329	467	315	400	340	400	262	300	360	518	338	422	340	450	156	
2000	320	383	360	499	331	460	380	455	302	440	339	355	160	431	398	560	400	432	236	451	147	
2001	400	442	372	528	376	475	384	400	335	459	400	414	258	389	479	575	420	479	306	500	158	

Fig 6. Median Weekly Earnings RAW

[1] Territorial authorities of New Zealand - https://en.wikipedia.org/wiki/Territorial_authorities_of_New_Zealand

Building Consents by region: Quarterly - Jun 1990 to Mar 2018

The raw dataset (Fig 7) extracted from Stats NZ shows all building consents by region. A subset of data consisting of residential building consents from Mar 2008 to Mar 2018 was separated using PivotTables, quarters from Jun - Dec 2018 were imputed using the NIPALS method and combined with the master combined rental dataset.

Series_reference	Period	Data_value	Suppressed	Status	Units	Magnitude	Subject	Group	Series_title_1	Series_title_2	Series_title_3	Series_title_4	Series_title_5
BLDQ.SF010001A1A	1990.06	284	n	F	Number	0	Building C	Building c	Northland Regio	All construction	New	Number	Actual
BLDQ.SF010001A1A	1990.09	315	n	F	Number	0	Building C	Building c	Northland Regio	All construction	New	Number	Actual
BLDQ.SF010001A1A	1990.12	371	n	F	Number	0	Building C	Building c	Northland Regio	All construction	New	Number	Actual
BLDQ.SF010001A1A	1991.03	258	n	F	Number	0	Building C	Building c	Northland Regio	All construction	New	Number	Actual
BLDQ.SF010001A1A	1991.06	272	n	F	Number	0	Building C	Building c	Northland Regio	All construction	New	Number	Actual
BLDQ.SF010001A1A	1991.09	278	n	F	Number	0	Building C	Building c	Northland Regio	All construction	New	Number	Actual

Fig 7. Building Consents by region RAW

Provider Summary Tables 2016: Yearly - 2008-2016

The raw dataset (Fig 8) shows student enrolment numbers, EFTS and student completion numbers for domestic and international students from 2008-2016 for tertiary providers. A separate lookup table was created to map tertiary provider to region in order to produce a summary table of student enrolments over time by region. Student enrolment numbers is the best indicator of student rental demand as EFTS only reports fulltime students and completions do not tell us much information about rental demand. This table was then interpolated using piecewise linear interpolation (A5.1) and combined with the master combined rental dataset.

Domestic and international student enrolments, EFTS and completions for providers 2008-2016

			Enrolments			EFTS			Completions		
Subsector	Provider	Year	International	Domestic	Total	International	Domestic	Total	International	Domestic	Total
		2014	3,965	27,260	31,225	3,015	15,655	18,670	1,220	5,360	6,575
		2015	4,210	26,995	31,205	3,205	15,460	18,665	1,315	5,440	6,755
		2016	4,550	26,620	31,170	3,430	15,495	18,925	1,310	4,780	6,090
	Victoria University of Wellington	2008	2,655	19,280	21,930	2,070	14,920	16,990	820	4,805	5,625
		2009	2,600	20,325	22,925	2,030	15,725	17,755	755	4,715	5,470
		2010	2,800	20,080	22,880	2,070	15,670	17,735	690	4,570	5,260
		2011	3,005	19,555	22,560	2,230	15,140	17,375	825	5,385	6,215
		2012	2,960	18,235	21,195	2,270	14,950	17,220	725	4,810	5,535
		2013	3,000	18,480	21,480	2,205	15,070	17,275	880	4,590	5,470
		2014	2,870	18,320	21,190	2,060	14,945	17,005	745	4,595	5,335
		2015	2,980	18,465	21,450	2,120	14,905	17,025	775	4,510	5,290
		2016	3,195	18,755	21,950	2,275	15,180	17,455	815	4,490	5,305
	University of Canterbury	2008	2,075	15,300	17,375	1,645	12,790	14,435	485	3,670	4,155
		2009	2,175	16,145	18,320	1,725	13,565	15,290	500	3,695	4,195
		2010	2,120	16,375	18,495	1,665	13,755	15,420	515	3,980	4,495

Fig 8. Provider Summary Tables 2016 RAW

Work and Student Visa Applications: Yearly 2008-2018

The raw datasets (Fig 9) show work and student visa applications by application criteria from 2008 -2018. This data was reformatted into a summary table displaying total approved student visa and approved work visa applications by year. Quarterly figures were then interpolated using piecewise linear interpolation (A6.1) and combined with the master combined rental dataset.

	Financial Year Decided																	
	2008/09			2009/10			2010/11			2011/12			2012/13			2013/14		
	Decision Type		Total	Decision Type		Total	Decision Type		Total	Decision Type		Total	Decision Type		Total	Decision Type		Total
	Approved	Declined		Approved	Declined		Approved	Declined		Approved	Declined		Approved	Declined		Approved	Declined	
Application Criteria																		
Child of NZ cit/res	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Dependant of Entrepreneur	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Dependant of a Diplomat	.	.	.	2	.	2	1	.	1	.	.	.	.	.	.	.	.	.
Dependant of a Worker - SMC	531	2	533	295	.	295	350	4	354	230	.	230	52	1	53	11	.	11
Dependent of a Worker	13,991	486	14,477	12,453	489	12,942	11,313	725	12,038	9,375	636	10,011	8,737	591	9,328	9,039	546	9,585

Fig 9. Work and Student Visa Applications RAW

Amounts paid out for Student Allowance and Accommodation Benefit payments: Yearly 1999 - 2017

The raw dataset (Fig 10) shows student allowance and accommodation benefit totals from the years 1999 to 2017. The data was reformatted into a lookup table showing average student allowance and accommodation benefit from 2008-2018. Quarterly figures were interpolated using piecewise linear interpolation (A7.1) and combined with the master combined rental dataset.

Allowance totals	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Student Allowance	64,292	68,084	70,231	68,869	64,036	60,826	56,806	59,431	62,479	65,702	82,633	95,945	99,271	96,908	85,094	79,670	75,050	71,028	65,352
Accommodation	50,287	53,257	54,370	53,516	50,446	48,143	43,644	45,068	45,976	46,109	57,976	67,819	69,672	67,512	58,977	54,923	51,161	48,193	44,438
Total number	64,316	68,086	70,235	68,877	64,053	60,958	56,811	59,459	62,505	65,705	82,638	95,948	99,277	96,910	85,099	79,672	75,051	71,029	65,354

Fig 10. Amounts paid out for Student Allowance and Accommodation Benefit payments RAW

Inflation and Mortgage Rates: Monthly - Jun 1998 to Jun 18

The raw datasets (Fig 11) shows monthly trends for the consumers price index and floating and fixed mortgage rates. These datasets were reformatted into lookup tables and added as attributes to the master combined rental dataset.

Mortgage Interest rates	Source: RBNZ
%pa	
Floating rate	2 year fixed rate
Jun-98	11.2	9.8
Jul-98	10.1	9.0
Aug-98	9.1	8.3
Sep-98	8.2	7.9

Fig 11. Inflation and Mortgage Rates RAW

Feature Engineering

The resulting raw combined master dataset has 21 features combined from 16 public datasets which allows for sufficient number of attributes depending on the desired regional granularity of the model (by territorial authority or by region). In order to build a model, categorical features such as region and territorial authority had to be one-hot encoded into numerical values using a label encoder. The date was converted to an ordinal value to be of numeric type. Possible target variables were identified and assumptions of associated attributes were made.

	Attribute	**Assumptions**
1.	Date	
2.	Region	
3.	Territorial Authority	
4.	Geometric mean rents by region	Target variable
5.	Geometric mean rents by territorial authority	Target variable
6.	Synthetic lower quartile rents by region	Target variable
7.	Synthetic lower quartile rents by territorial authority	Target variable
8.	Lodged bonds by region	Increases with geometric mean rents or synthetic lower quartile rents by region as indicates high demand to short supply
9.	Lodged bonds by territorial authority	Increases with geometric mean rents or synthetic lower quartile rents by territorial authority as an indicate of high demand to short supply
10.	Regional GDP	Increases with rent prices as regional GDP often dictates the average living cost including rent by region
11.	Population Estimate 15-39	Increases with rent prices as higher the demand to supply, the higher rent prices will be
12.	Median Weekly Earnings 20-29	Increases with rent prices as the more money people earn, the more rent they will be able to afford
13.	Number of Residential Building Consents	Increases with rent prices as the more residential building consents are lodged indicates high demand to short supply
14.	Student Enrolments by Region	Increases with rent prices as the more students enrolled, the more rentals in demand indicating high demand to short supply
15.	Approved Student Visa	Increases with rent prices as the more international students coming to NZ, the more rentals in demand, indicating high demand to short supply
16.	Approved Work Visa	Increases with rent prices as the more working internationals coming to NZ, the more rentals in demand indicating high demand to short supply
17.	Average Student Allowance	Increases with rent prices as an indicate of the baseline income of average domestic student on student allowance
18.	Average Accommodation Benefit	Increases with rent prices as an indicate of how much average domestic student has available for accommodation
19.	CPI	Increases with rent prices due to inflation
20.	Mortgage Floating Rate	Increases with rent prices as higher mortgage rates mean landlords will raise rents to cover costs
21.	Mortgage 2 year fixed rate	Increases with rent prices as higher mortgage rates mean landlords will raise rents to cover costs

Part 2: Feature Importance to rental costs in Wellington

Eliminating Collinearity

Collinearity is a problem in regression analysis that occurs when two independent variables are highly correlated. The relationship between the independent variables and the dependent variables is distorted by the very strong relationship between the independent variables, leading to the likelihood that our interpretation of relationships will be incorrect, decreased generalisation performance on the test set due to high variance and less model interpretability. In order to avoid collinearity, a correlation matrix was plotted to identify seven pairs of highly correlated features (Fig 12) with a correlation magnitude of over 90% .

	drop_feature	corr_feature	corr_value
0	Synthetic lower quartile rents by region	Geometric mean rents by region	0.984558
1	Synthetic lower quartile rents by territorial ...	Geometric mean rents by territorial authority	0.984013
2	Regional GDP	Lodged bonds by region	0.921240
3	Approved Work Visa	Population Estimate 15-39	0.941148
4	Average Student Allowance	Date	0.936628
5	Average Accommodation	Population Estimate 15-39	0.952357
6	Average Accommodation	Average Student Allowance	0.976617

Fig 12. Table of Collinear Features

The first two pairs between synthetic lower quartile rents and geometric mean rents by region and territorial authority are highly correlated because they are our nominal target variables dependent on the desired regional granularity of the model. The third pair between regional GDP and lodged bonds by region tells us that there is a very high correlation between the demand of rentals and the economic activity of a region indicating an insight that the more affluent a region is, the more people renting in that region. The fourth pair indicates that the population increase of 15-39 year olds have been due to an increase of working migrants. The seventh pair tells us that the average accommodation benefit and student allowance are very high correlated and that the fifth and sixth correlated pair indicates that they both have been steadily with the growing population of 15-39 year olds over time from quarter to quarter.

Model Selection

A gradient boosting regression model (GBM) form Sklearn was chosen to predict geometric mean rents by territorial authority to allow us to explore rent prices in Wellington region to a more granular degree broken down to its ten territorial authorities - Carterton, Kapiti Coast, Lower Hutt, Masterton, Porirua, South Wairarapa, Upper Hutt and Wellington. A GBM is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Therefore, according to our table of collinear features (Fig 13), we can remove the attributes synthetic lower quartile rents by region and territorial authority, geometric mean rents by region and lodged bonds by region in order to remove conflicting target variables and focus on values by territorial authority. Average student allowance and average accommodation have also been removed as they do not give us any extra information that the date and population estimate attributes do not account for. The result is a dataset of 13 attributes with no collinearity with "Geometric mean rents by territorial authority" as the target label. The dataset was split into a 60/40 training set.

Feature Selection

A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model. Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance. This importance is calculated explicitly for each attribute in the dataset, allowing attributes to be ranked and compared to each other. Importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. The performance measure may be the purity (Gini index) used to select the split points or another more specific error function. The feature importances are then averaged across all of the decision trees within the model. A feature selector [1] using Microsoft's LightGBM model was used to rank feature importance and eliminate features of zero and near-zero importance (Fig 13). The feature importances were averaged over 30 training runs of the GBM in order to reduce variance. Early stopping was implemented to prevent overfitting of the training data.

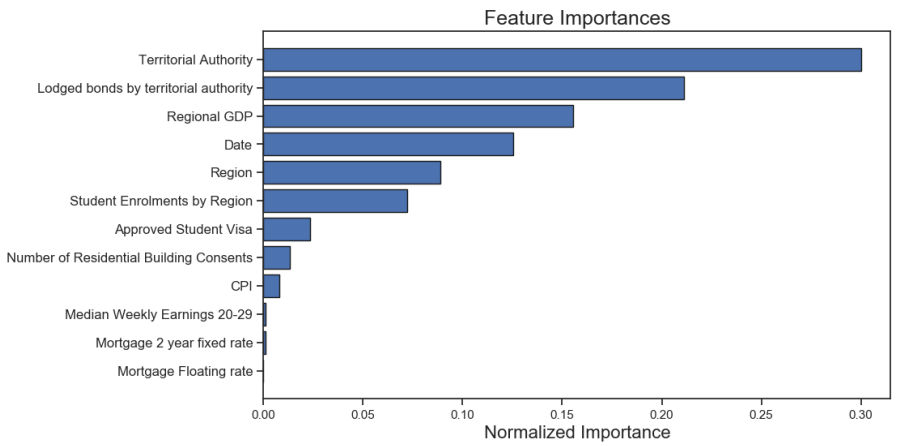

Fig 13. Feature Importances of Geometric Mean Rents by Territorial Authority Model

[1] Feature Selector: Simple Feature Selection in Python - https://github.com/WillKoehrsen/feature-selector

The cumulative feature importance shows that 8 features are required to achieve 99% of cumulative importance (Fig 14). Therefore, 5 of the least important features were removed as they have significant contribution to the total importance. However, we must remember training the gradient boosting machine is stochastic meaning the feature importances will change every time the model is run. This should not have a major impact (the most important features will not suddenly become the least) but it will change the ordering of some of the features. It also can affect the number of zero importance features identified.

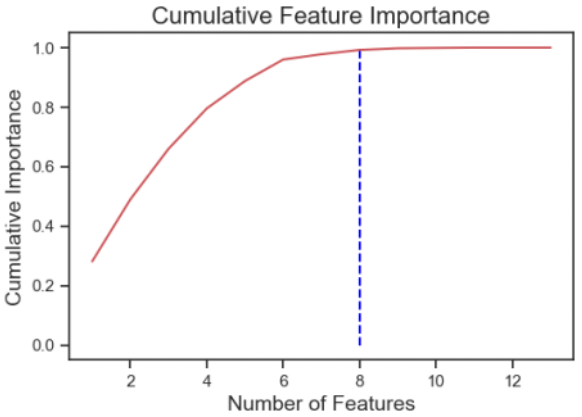

Fig 14. Cumulative Feature Importance of Geometric Mean Rents by Territorial Authority Model

Evaluating model with and without Dimensionality Reduction

Regression models are evaluated using R^2 (coefficient of determination) regression score function which provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0. Another important evaluation metric is root mean squared error (RMSE) which measures how much error there is between the predicted dataset and the test dataset. It is the standard deviation of the residuals (prediction errors) where residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are.

The model output was investigated for different stages of dimensionality reduction (see A8). It is observed training time decreases with the number of features as well as the risk of overfitting. Ten seems like a good number of features to proceed with the model as the R-squared value seems to improve and the RMSE value decreases after removing zero importance and collinear features. Removing the low importance features actually decreases the performance of the model.

	# Features	R-Squared	RMSE	Execution Time (s)
RAW Dataset	21	0.9858	9.4454	0.7458
After removing collinear features	13	0.9426	18.7723	0.6046
After removing zero importance features	10	0.9459	18.5136	0.5408
After removing low importance features	8	0.9351	19.7023	0.4789

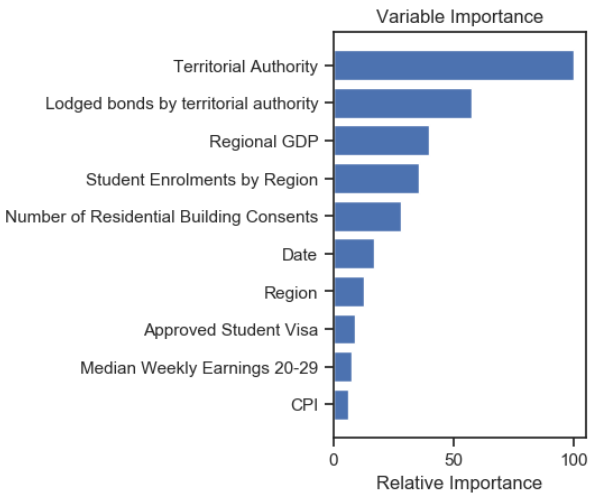

Fig 15. After Removing Zero Importance Features

Part 4: Consequences and ethics of reporting findings

Based on the model's relative feature importance rankings of Wellington region, the top 5 contributing features of significance are the number of lodged bonds indicating supply, the number of residential building consents indicating demand, the number of approved student visas indicating demand, the number of student enrolments indicating demand and the median weekly wage for 20-29 age group indicating how affordable the rent price is. Inflation (CPI) and the fixed 2 year mortgage rate are also influencing features but they are not within our means to control.

	NZ		Wellington		Auckland	
1	TA	100%	Lodged bonds by TA	100%	Lodged bonds by TA	100%
2	Lodged bonds by TA	67.37%	TA	64.36%	TA	62.63%
3	Student Enrolments	37.36%	No. of Residential Building Consents	33.91%	Date	27.03%
4	Regional GDP	36.70%	Approved Student Visa	28.72%	Approved Student Visa	26.39%
5	No. of Residential Building Consents	26.05%	Date	26.02%	No. of Residential Building Consents	25.52%
6	Date	19.98%	Student Enrolments	22.36%	Median Weekly Earnings 20-29	21.39%
7	Region	13.77%	Median Weekly Earnings 20-29	22.36%	CPI	20.80%
8	CPI	11.39%	CPI	22.26%	Mortgage floating rate	15.63%
9	Approved Student Visa	8.41%	Mortgage 2 year fixed rate	13.92%		

Fig 16. Relative Feature Importance for predicting geometric mean rents by TA

Fig 17. Geometric Mean rent vs Models Top 5 most important features for Wellington

Attribute	Hypothesis
Lodged bonds by territorial authority	Geometric mean rents *decreases with increase* of lodged bonds as an indication of gap between supply and demand closing - when num of lodged bonds has increased, mean rent has decreased Sep 16 and Mar 17
Number of Residential Building Consents	Geometric mean rents *decreases with increase* of residential building consents as indication of gap between supply and demand closing for the future
Approved Student Visa	Geometric mean rents not too largely affected by number of approved student visas - large decrease has not stopped rent increasing, this is because student visa numbers are not region-specific
Student Enrolments by Region	Geometric mean rents *decreases with decrease* in num of students enrolled as less students, less demand
Median Weekly Earnings 20-29	Geometric mean rents *decreases with decrease* in median weekly wage as the less renters earn, the lower the assumed living cost and the less one can afford in rent - ~50% of weekly wage spent on rent

Therefore, arbitrary number of increase and decreases of the respective features were applied for the year of 2017 to predict how rental prices were to change for that of 2018. The model was then trained on all data up to 2017 and the test set was set to predict figures for Mar and Jun of 2018. As predicted, the average geometric mean rent for Wellington does decrease with increase in the number of lodged bonds and residential building consents indicating that there are more houses and flats available to meet demand. When demand is low, tenants have more power of choice and landlords will be more inclined to lower rental prices to attract tenants.

A decrease in the number of student visas and therefore international students also leads to a decrease in the predicted rent prices but not as much as a decrease in student enrolment. Student enrolment numbers in Wellington have already dropped in the last year, which may be due to high rental prices deterring students from studying in Wellington.

If the cost of living were to be 20% higher than what was predicted to be affordable as indicated by a decrease in the median wage, this would most likely lead to a decrease in rent to match. However, if this is not the case and the cost of living were to be 20% lower than what was predicted to be affordable, it would ethical to report these findings and let landlords choose to raise rent prices if they wish.

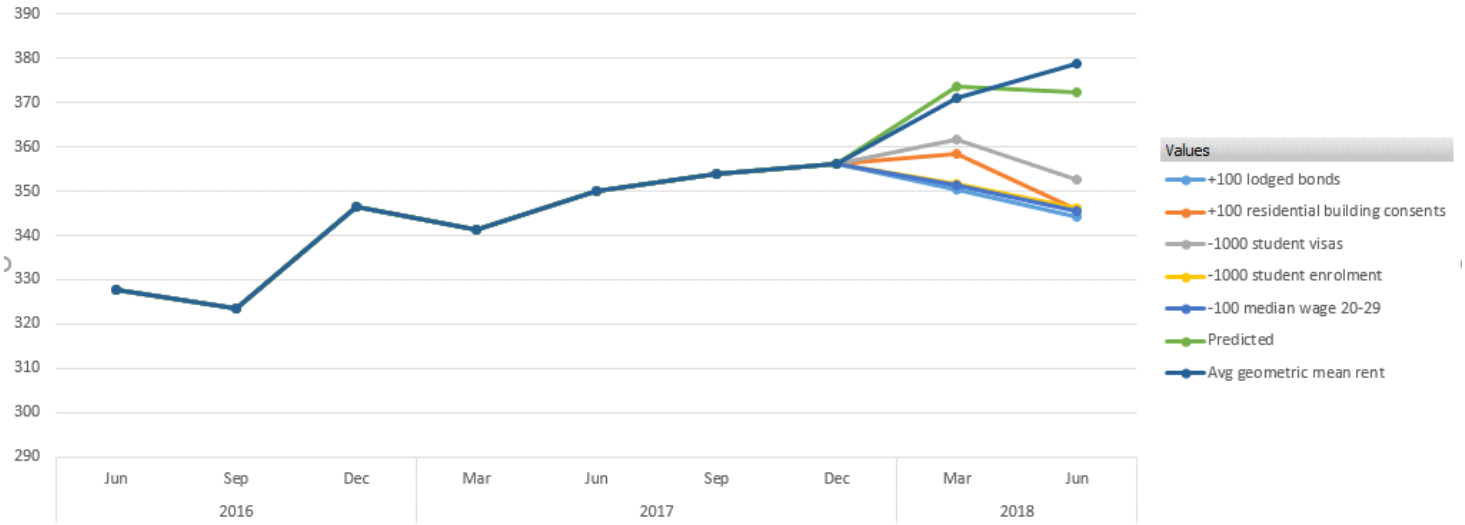

Fig 17. Avg Geometric Mean Rents by TA for Wellington w/ Predicted Outcomes

If the university decides to build a new 400 hall of residence, it would definitely decrease the average rental price as it will alleviate demand on the rental market in Wellington especially since the rent prices are greatly influenced by the student population.

Appendix

A Part 1: Evidence related to rental costs in Wellington - Look up Tables

A1 MBIE - Regional-TA Rental Datasets

Region	Territorial Authority
Canterbury	Ashburton District
Auckland	Auckland
West Coast	Buller District
Wellington	Carterton District
Hawke's Bay	Central Hawkes Bay Dist
Otago	Central Otago District
Canterbury	Christchurch
Otago	Clutha District
Otago	Dunedin
Northland	Far North District
Auckland	Franklin District
Gisborne	Gisborne District
Southland	Gore District

Fig A.1 - Regional-TA Lookup Table Snapshot

A2 Stats NZ - Regional GDP Dataset

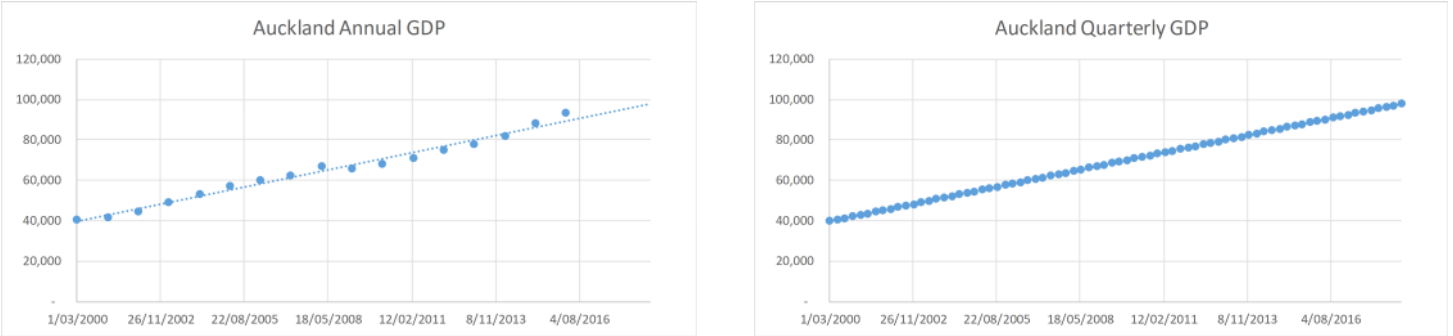

Fig A2.1 - Auckland GDP Annual-Quarterly Interpolation

Month	Auckland	Bay of Plenty	Canterbury	Gisborne	Hawke's Bay	Manawatu-Wanganui	Marlborough
2000-03-01	39,826	5,755	12,286	932	3,840	5,220	993
2000-06-01	40,601	5,866	12,598	945	3,888	5,290	1,018
2000-09-01	41,377	5,977	12,910	958	3,935	5,360	1,042
2000-12-01	42,153	6,088	13,222	971	3,983	5,431	1,067
2001-03-01	42,928	6,199	13,534	984	4,030	5,501	1,092
2001-06-01	43,704	6,310	13,846	997	4,078	5,572	1,116
2001-09-01	44,479	6,421	14,158	1,009	4,125	5,642	1,141

	Month	Region	Regional GDP
0	2000-03-01	Auckland	39,826
1	2000-06-01	Auckland	40,601
2	2000-09-01	Auckland	41,377
3	2000-12-01	Auckland	42,153
4	2001-03-01	Auckland	42,928
5	2001-06-01	Auckland	43,704
6	2001-09-01	Auckland	44,479
7	2001-12-01	Auckland	45,255

Fig A2.2 - Regional GDP Lookup Transformation

A3 Stats NZ - National Population Estimate Dataset

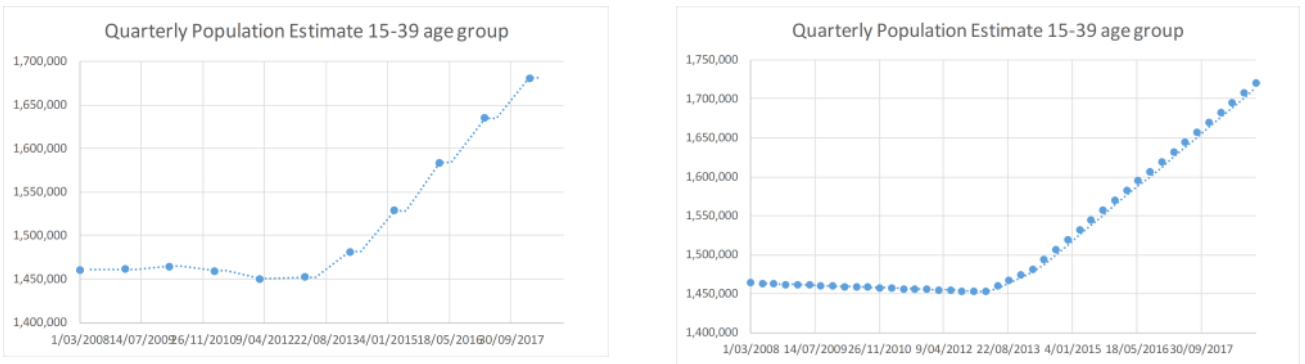

Fig A3.1 - Population Estimate 15-39 Age Group Annual-Quarterly Interpolation

A4 Stats NZ - Median weekly earnings Dataset

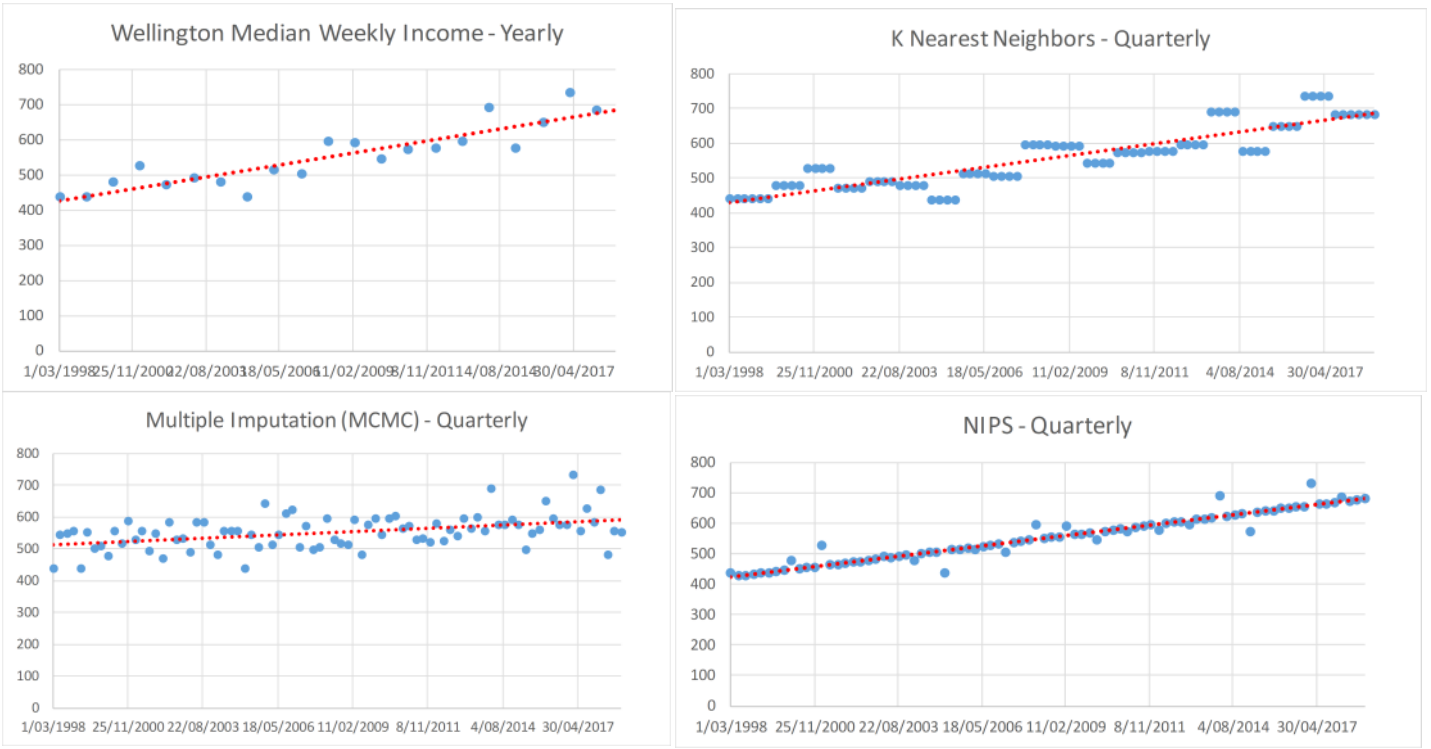

Fig A4.1 - Median weekly earnings Annual-Quarterly Interpolation

A5 Education Counts - Provider Summary Tables Dataset

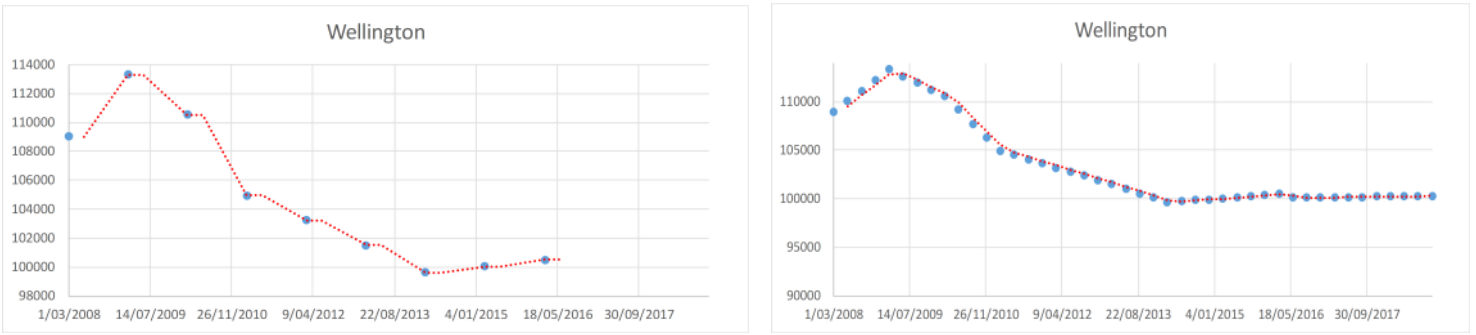

Fig A5.1 Wellington Student Enrolment Counts Annual-Quarterly Interpolation

A6 Immigration NZ - Approved Work and Student Visa

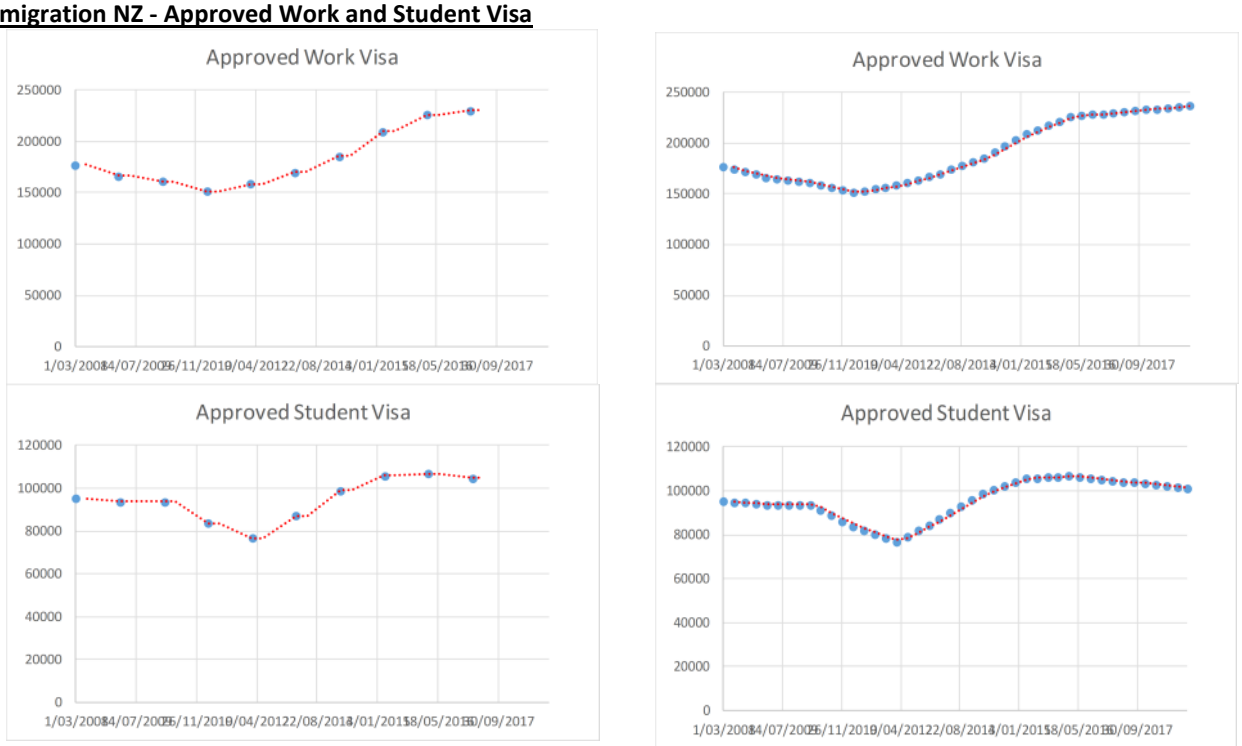

Fig A6.1 Approved Work and Student Visas Annual-Quarterly Interpolation

A7 MSD - Average Student Allowance and Accommodation Benefit

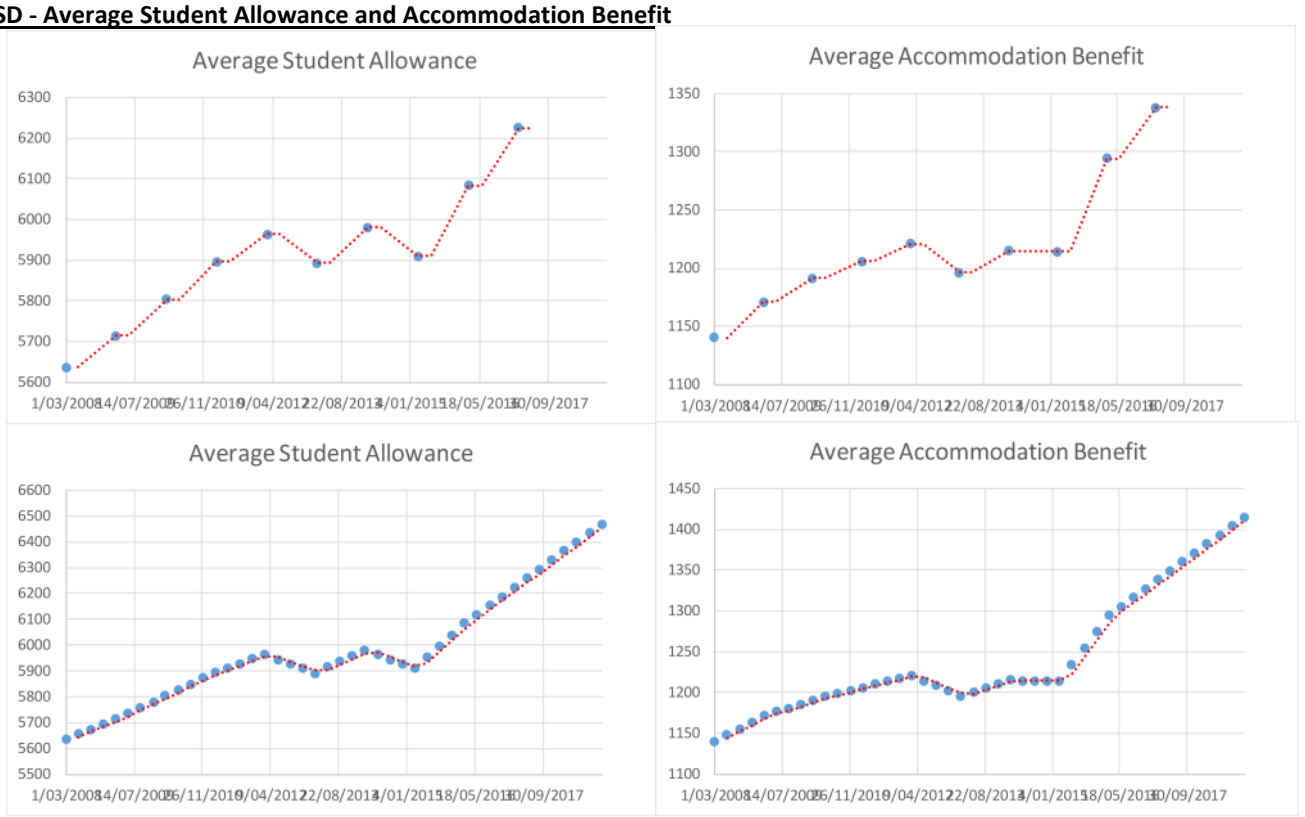

Fig A7.1 Average Student Allowance and Accommodation Benefit Annual-Quarterly Interpolation

A8 Evaluating model with and without Dimensionality Reduction

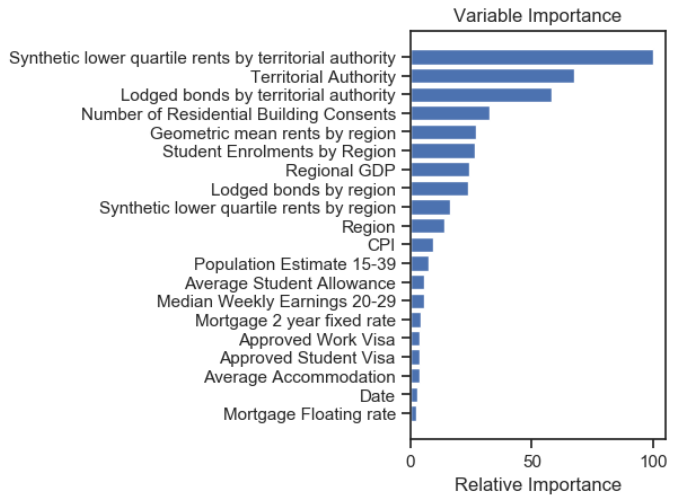

Fig A8.1 RAW Dataset

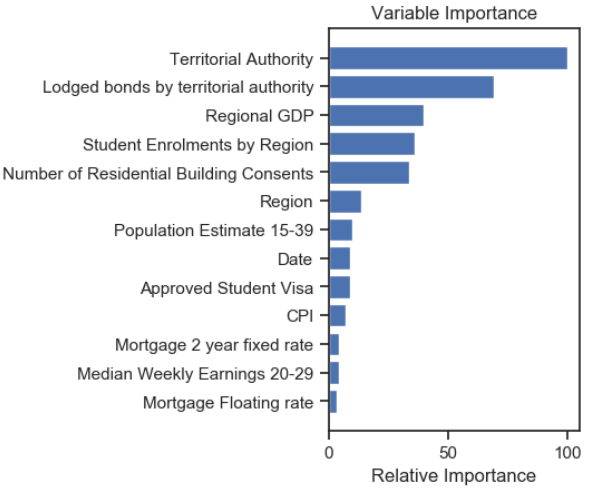

Fig A8.2 After Removing Collinear Features

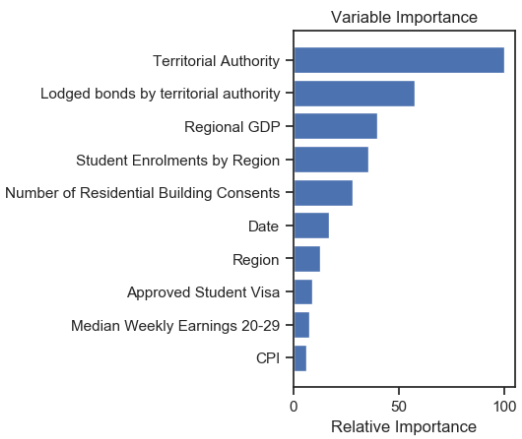

Fig A8.3 After Removing Zero Importance Features

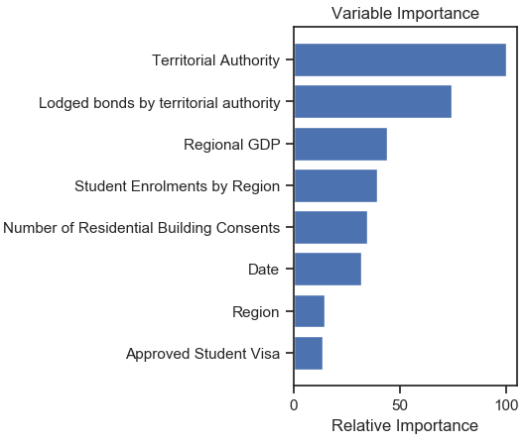

Fig A8.4 After Removing Low Importance Features