House Price Investigation for Washington Cities

The Outliers:
Sagar Bansal, Yiqing Bu, Parth Padia, Ziting Yuan

AGENDA

- Introduction
- Data Characteristics
- Initial Multiple Linear Regression Model
- Transformed Model
- Analysis Summary

Introduction

Introduction

Motivation of analysis:

- In the Real Estate Industry, understanding the trend of property prices is an invaluable tool as they are generally good indicators of economy.

Project Goals:

- For this purpose , we have chosen a sample of 10 cities from the dataset of house prices to conduct our analysis.


Data Characteristics

Data Characteristics

Original Data Description:

- Observational and Cross-sectional in nature
- 4,600 number of observations & 18 variables

Cleaned Data:

- 3,412 number of observations
- Predictors:-
  - Quantitative - Sqft_living, Sqft_lot, Bathrooms, Bedrooms, Age of the house and Floors.
  - Qualitative - City.
- Response - Price

Data Characteristics

Table showing summary statistics of quantitative data used

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
price	3412	87500.0000	7062500.000	559877.4125	353808.5124
bathrooms	3412	.75	8.00	2.1491	.77291
sqft_living	3412	370	13540	2100.01	936.518
sqft_lot	3412	638	1074218	11753.26	28512.876
floors	3412	1.0	3.5	1.525	.5531
bedrooms	3412	1	9	3.38	.916
Age	3267	0	114	45.47	31.910
Valid N (listwise)	3267				

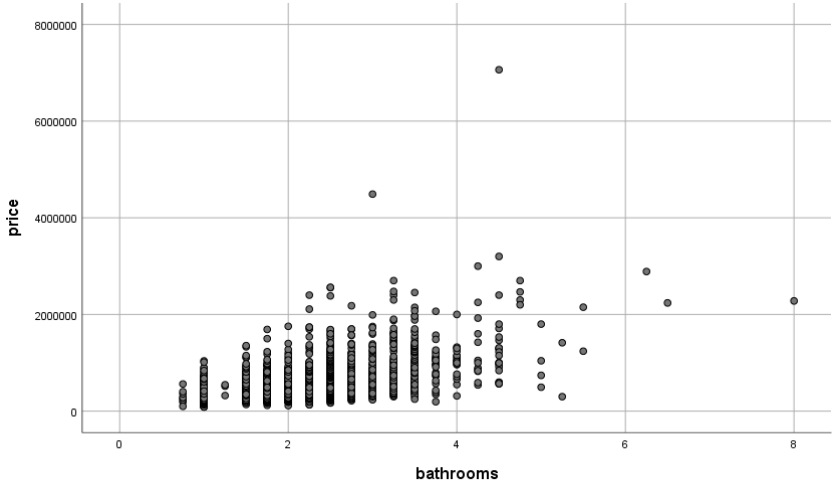
Data Characteristics

Frequency of the ten cities studied

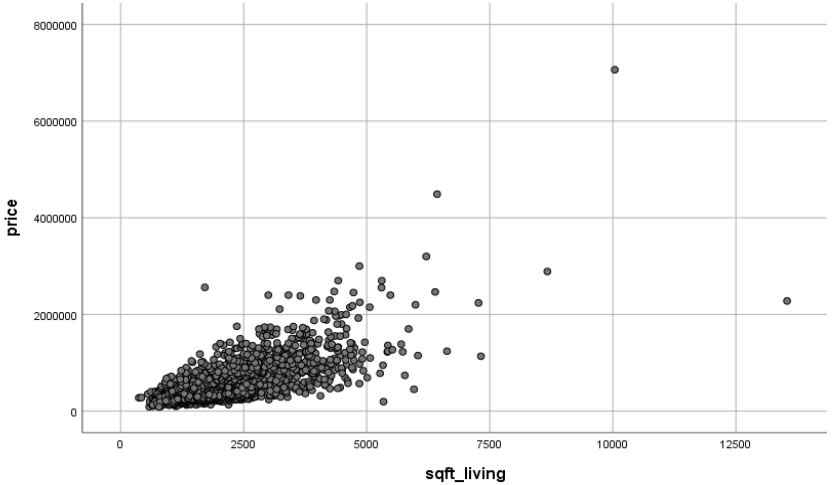
		Frequency	Percent
Valid	Auburn	175	5.1
	Bellevue	281	8.2
	Federal	145	4.2
	Issaquah	186	5.5
	Kent	183	5.4
	Kirkland	187	5.5
	Redmond	234	6.9
	Renton	291	8.5
	Sammamish	171	5.0
	Seattle	1559	45.7
	Total	3412	100.0

Data Characteristics

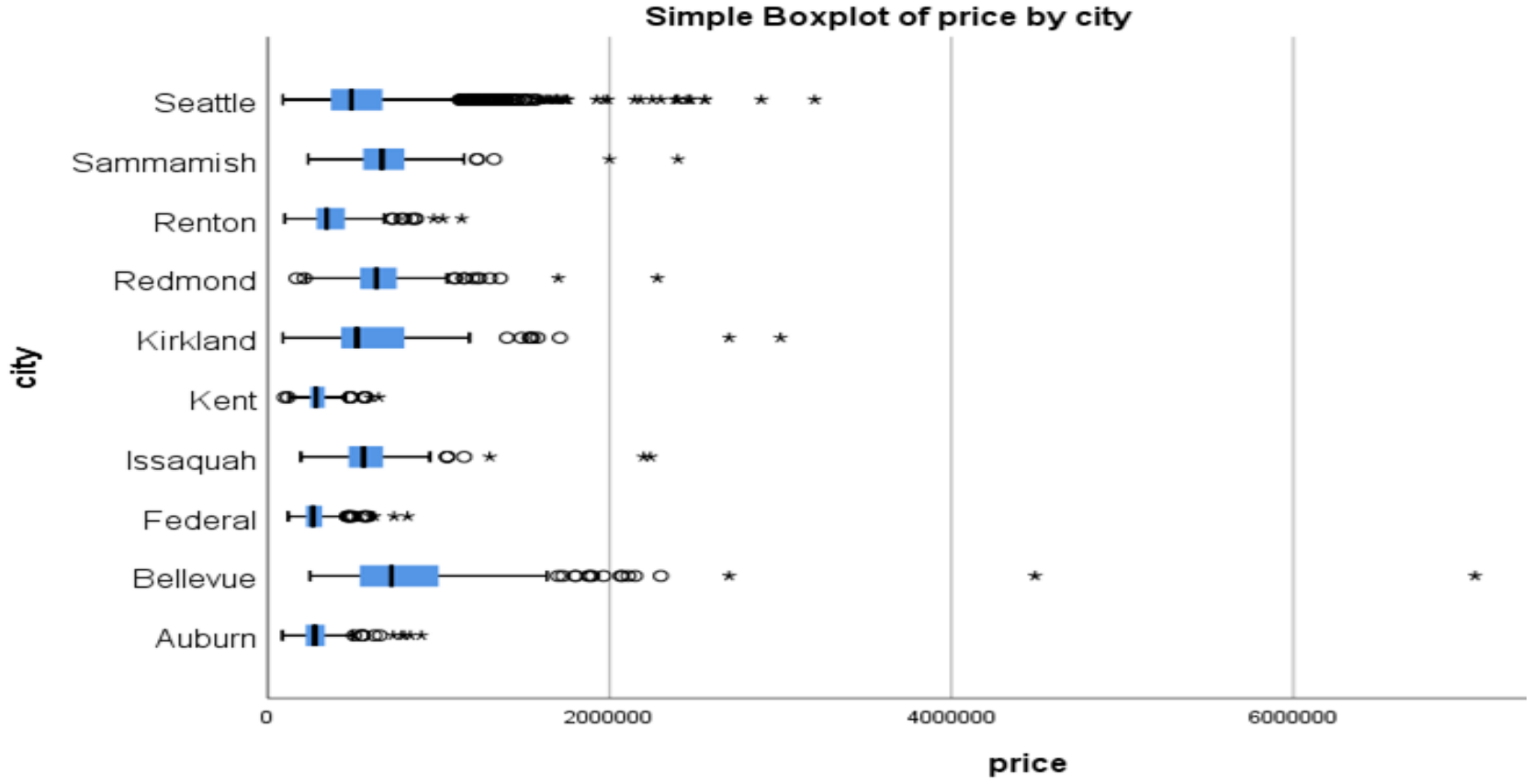
Scatter plot of Price vs no. of Bathrooms



Scatter Plot of Price vs Sqft_Living



Data Characteristics



Initial Multiple Linear Regression Model

Hypothesized Model

Price = $b_0 + b_1$Bedroom + b_2Bathroom + b_3Sqft_living + b_4Floors + b_5Sqft_lot + b_6AgeoftheHouse + b_{c1}CitySeattle + b_{c2}CityRenton + b_{c3}CityBellevue + b_{c4}CityRedmond + b_{c5}CityKirkland + b_{c6}CityIssaquah + b_{c7}Kent + b_{c8}Auburn + b_{c9}Sammamish + e

Assumptions:

Fe ted as the reference level!

$\epsilon_i \overset{Indp}{\sim} Normal(0, \sigma)$

Check Utility

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.796[a]	.633	.631	214783.0340

a. Predictors: (Constant), CItySammamish, sqft_lot, cityKirkland, CityKent, floors, CityRenton, CityAuburn, cityBellevue, CityRedmond, bedrooms, CityIssaquah, Age of the house, sqft_living, bathrooms, CitySeattle

b. Dependent Variable: price

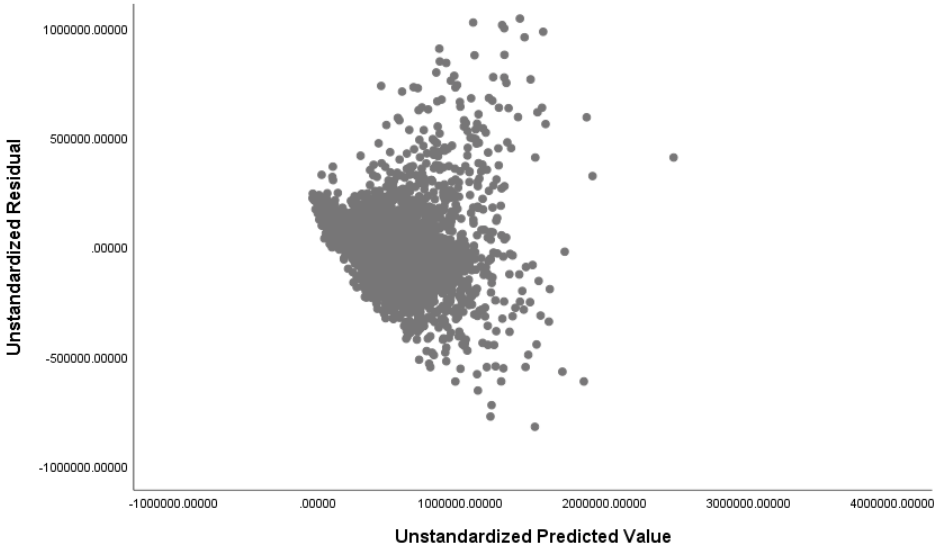
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.703E+14	15	1.802E+13	390.660	.000[b]
	Residual	1.567E+14	3396	4.613E+10		
	Total	4.270E+14	3411			

a. Dependent Variable: price

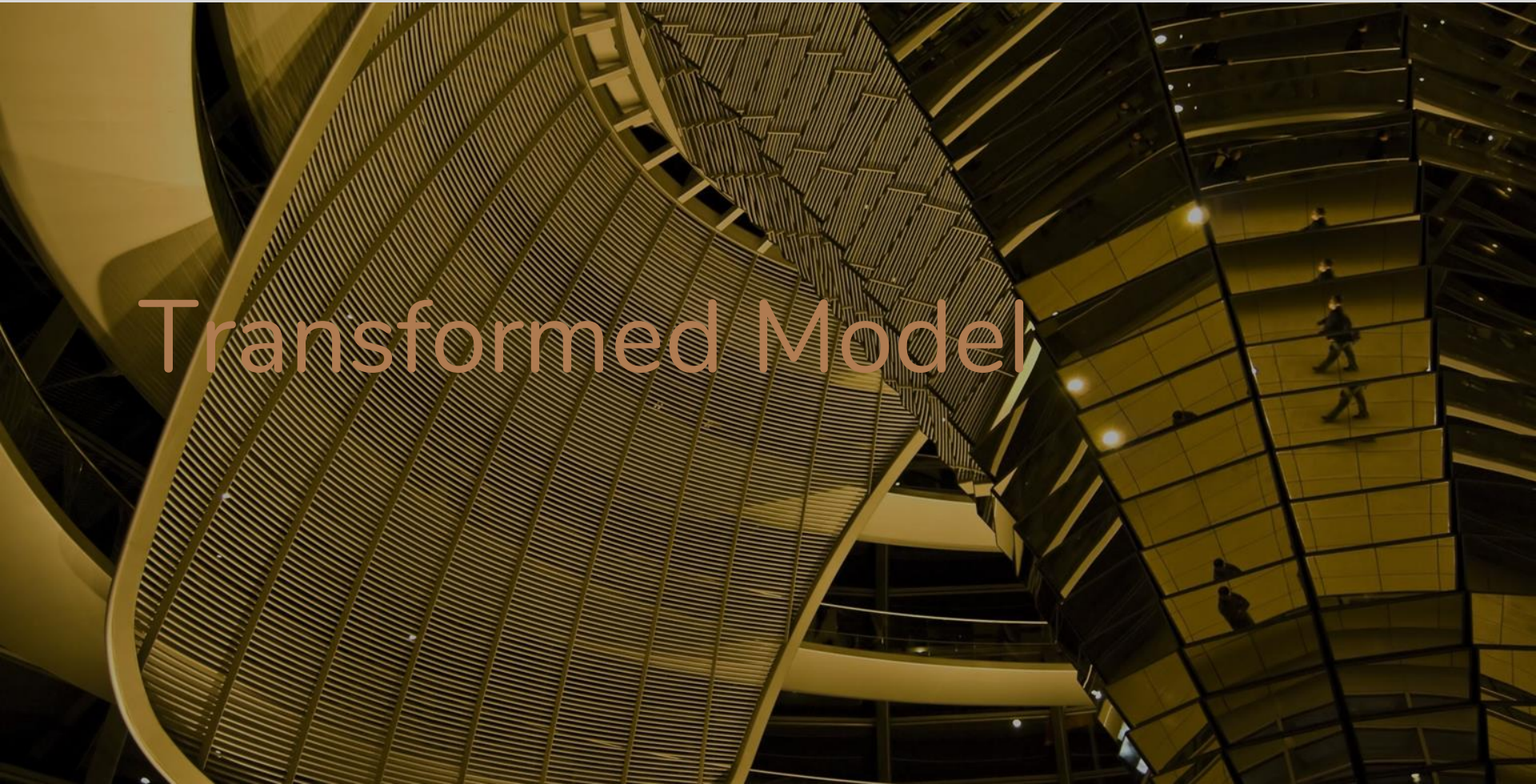
b. Predictors: (Constant), CItySammamish, sqft_lot, cityKirkland, CityKent, floors, CityRenton, CityAuburn, cityBellevue, CityRedmond, bedrooms, CityIssaquah, Age of the house, sqft_living, bathrooms, CitySeattle

Check Lack of Fit & Homogeneity



Residual vs. Fitted Values Plot

- Multiplicative pattern
- Lack of Fit
- Have heterogeneity issue
- Need to modify the model


Transformed Model

Transformed Model

ln(Price) = $b_0 + b_1$Bedroom + b_2Bathroom + b_3Sqft_living + b_4Floors + b_5Sqft_lot + b_6AgeoftheHouse + b_{c1}CitySeattle + b_{c2}CityRenton + b_{c3}CityBellevue + b_{c4}CityRedmond + b_{c5}CityKirkland + b_{c6}CityIssaquah + b_{c7}Kent + b_{c8}Auburn + b_{c9}Sammamish + e

Assumptions:

$\epsilon_i \overset{Indp}{\sim} Normal(0, \sigma)$

Federal Way is treated as the reference level!

Check Utility

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.833[a]	.694	.693	.28757

a. Predictors: (Constant), citySammamish, sqft_lot, cityKirkland, cityKent, floors, cityRenton, cityAuburn, cityBellevue, cityRedmond, bedrooms, cityIssaquah, Age of the house, sqft_living, bathrooms, citySeattle

b. Dependent Variable: logprice

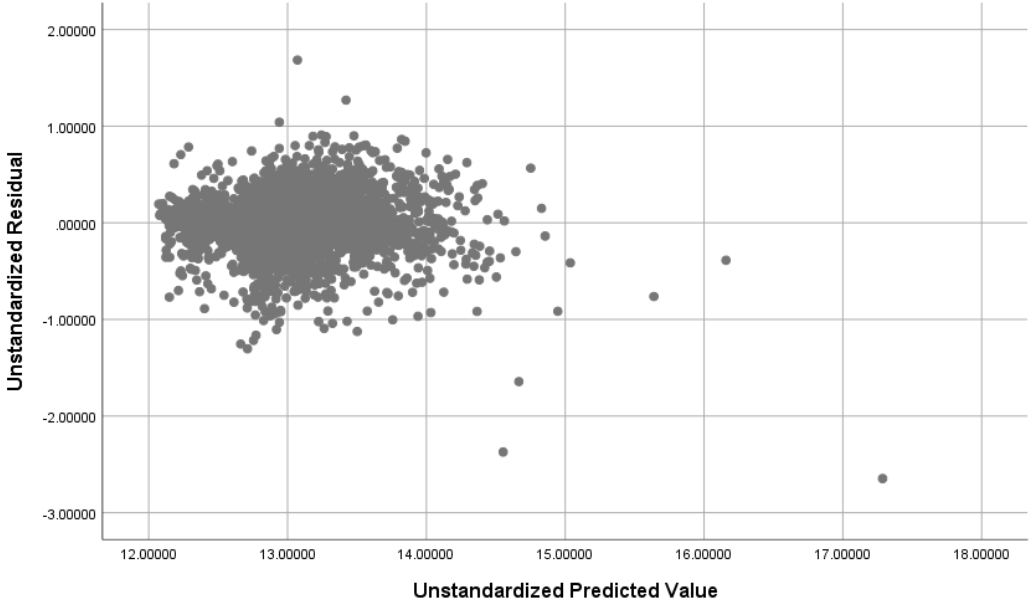
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	638.013	15	42.534	514.325	.000[b]
	Residual	280.846	3396	.083		
	Total	918.859	3411			

a. Dependent Variable: logprice

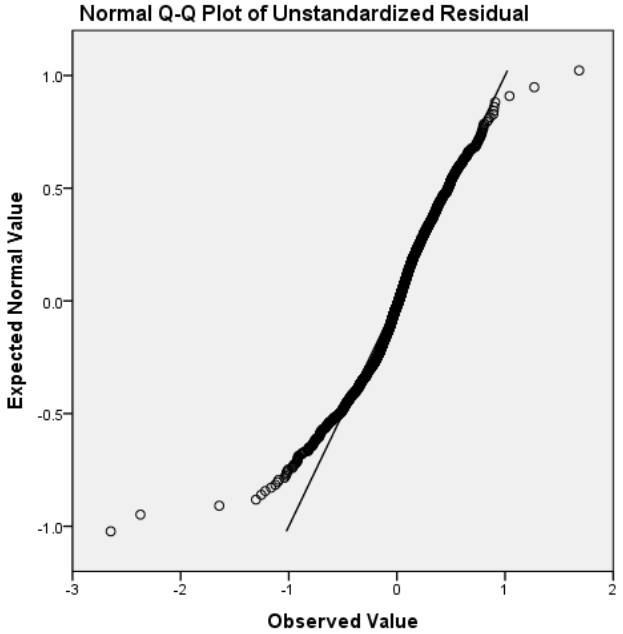
b. Predictors: (Constant), citySammamish, sqft_lot, cityKirkland, cityKent, floors, cityRenton, cityAuburn, cityBellevue, cityRedmond, bedrooms, cityIssaquah, Age of the house, sqft_living, bathrooms, citySeattle

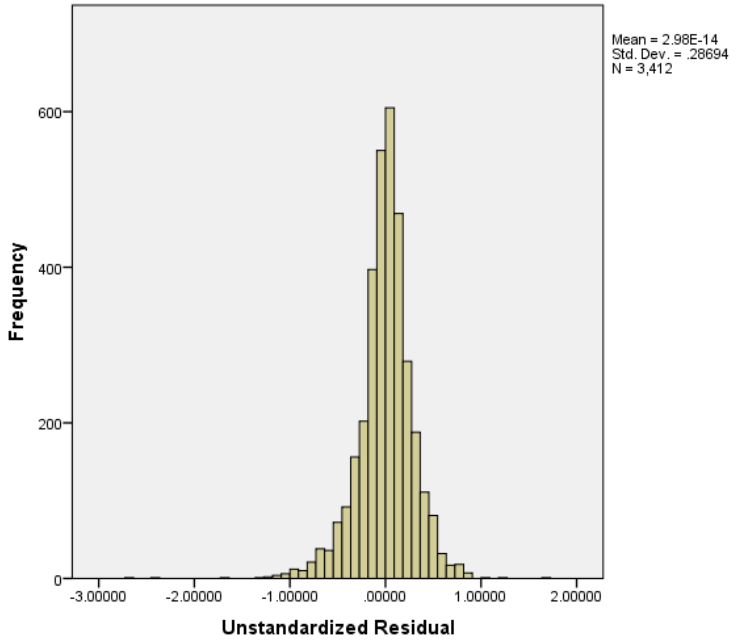
Check Lack of Fit & Homogeneity



- Data evenly spread out
- No obvious pattern

Checking Normality





- Residuals appear closer to normal line
- Almost normally distributed
- Future work will need further investigation

Checking Independence

- No time-series structure
- No need to perform Durbin-Watson test
- Therefore, we assume that error terms are independent of each other

Model Interpretation

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	11.585	.036		324.792	.000
	bedrooms	-.039	.007	-.069	-5.577	.000
	bathrooms	.101	.011	.151	8.942	.000
	sqft_living	.000	.000	.566	34.836	.000
	sqft_lot	-3.012E-7	.000	-.017	-1.641	.101
	floors	.103	.011	.109	8.971	.000
	Age of the house	.003	.000	.164	12.047	.000
	citySeattle	.579	.026	.556	22.219	.000
	cityRenton	.215	.029	.116	7.343	.000
	cityBellevue	.771	.030	.408	26.002	.000
	cityRedmond	.664	.031	.323	21.657	.000
	cityKirkland	.642	.032	.282	20.148	.000
	cityIssaquah	.532	.032	.233	16.498	.000
	cityKent	.047	.032	.020	1.454	.146
	cityAuburn	.016	.032	.007	.504	.614
	citySammamish	.599	.033	.252	18.214	.000

a. Dependent Variable: logprice

- Only **Sqft_lot, CityKent, CityAuburn** are not statistically significant
- Coefficients of **bedrooms** and **Sqft_lot** are negative


Summary

Analysis Summary

- All cities except Kent and Auburn are statistically different from Federal Way
- The house prices of Federal Way are least expensive among the ten cities which we have been studied.
- Since bedroom has negative correlation with house price, in the future study, we could conduct further investigation to figure out why bedroom does not have positive relationship with house price.
- The model cannot be used for forecasting future house prices.

Thank You!