KING COUNTY HOUSE PRICE ANALYSIS

Yukthi Papanna Suresh, Emily Lin, Caglar Kurtkaya

PROBLEM SELECTION

- House Price Prediction of King county the most populous county in Washington and 13th in USA
- Dataset from Kaggle
- **Solution**: Building a price prediction system using features
  - **Task 1:** Build regression model
  - **Task 2:** Build a clustering model using location

INITIAL DATA-FRAME

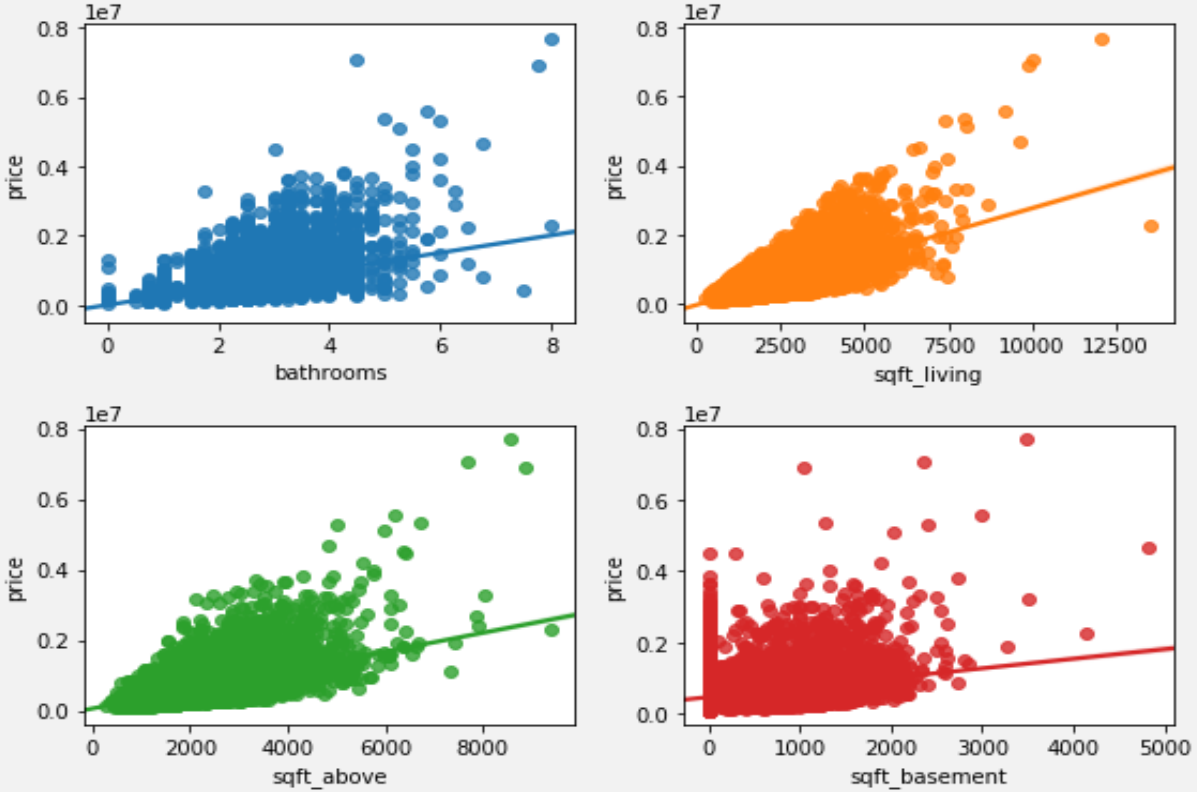
	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0
3	2487200875	20141209T000000	604000.0	4	3.00	1960	5000	1.0	0	0
4	1954400510	20150218T000000	510000.0	3	2.00	1680	8080	1.0	0	0

	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
1	3	7	2170	400	1951	1991	98125	47.7210	-122.319	1690	7639
2	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
3	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
4	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503

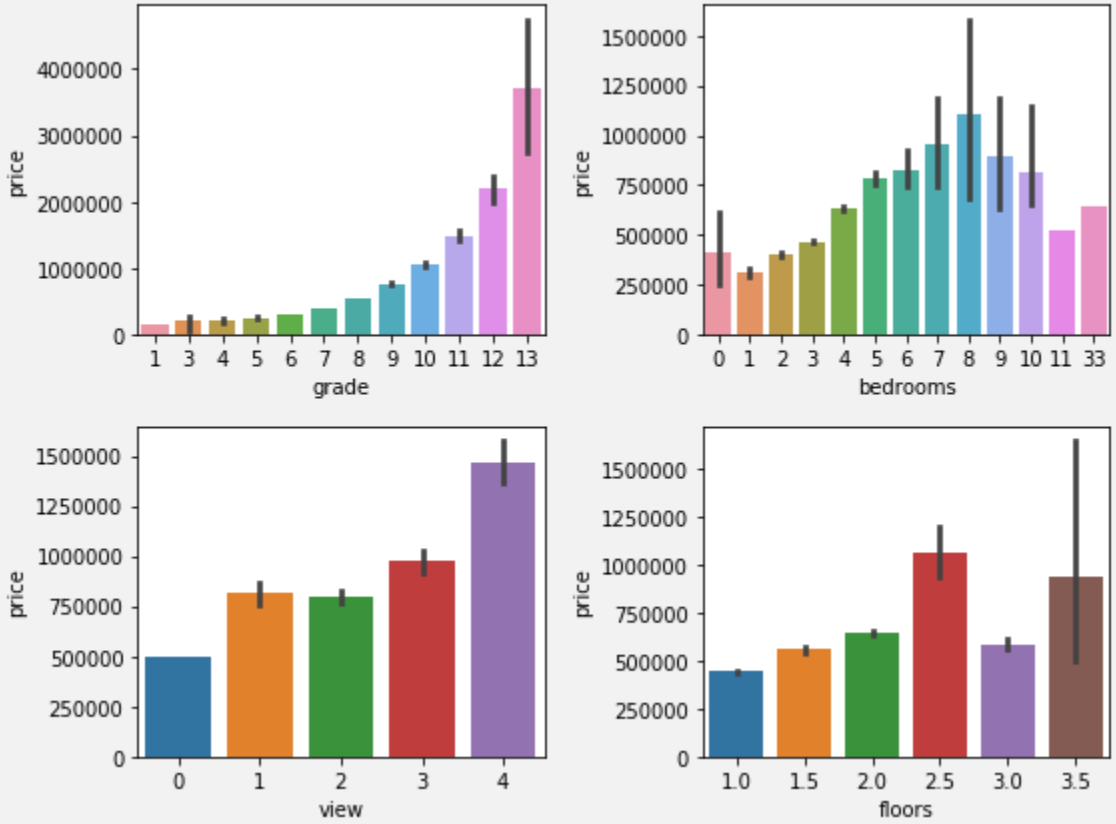
PRE-PROCESSING

- Trimmed date to year (Noise)
- Null values analysis (Missing Values)
- Create dummy variable for date (2014->0, 2015->1)
- Removed outliers (Outlier)

EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS CONT.



REGRESSION MODELS

- Top 12 attributes (using SelectKBest)
- Train, Test and Validation Set
- StandardScaler fit Training set
- Transformed Train, Test and Validation set (Mean of 0 and Standard deviation of 1)

REGRESSION CONT.

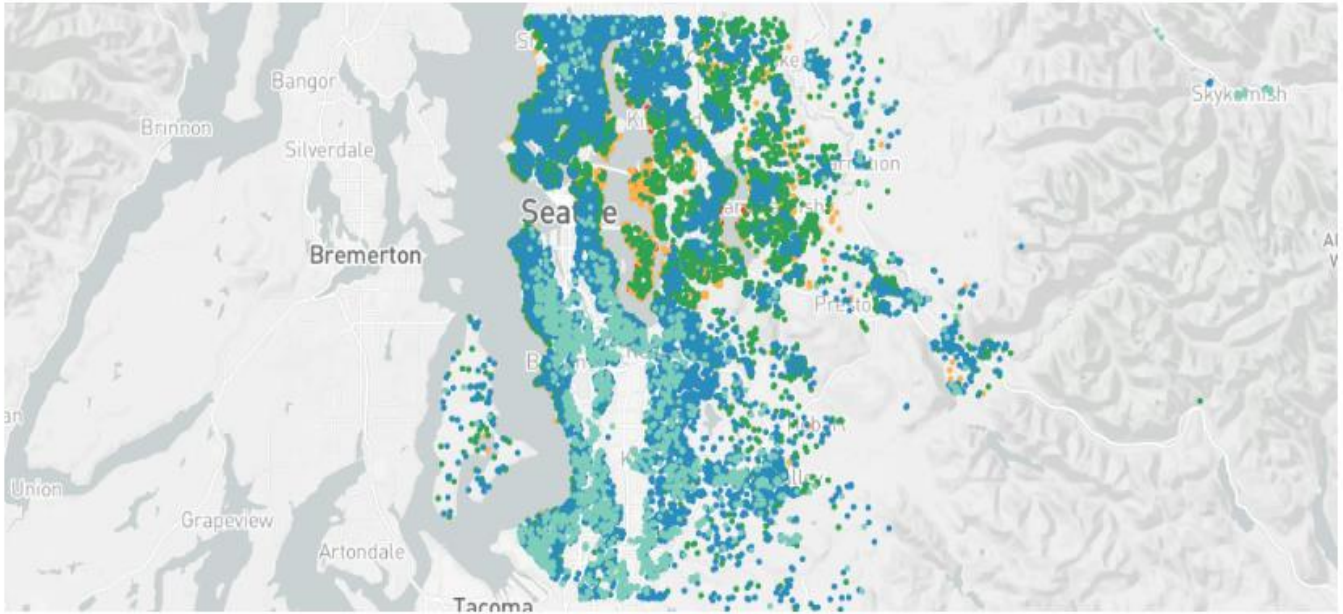
Summary of the R-squared values of the models using important features

Model	Train R-square	Validation R-square	Test R-square
Random Forest Regressor	**0.9571038167086312**	**0.804736200215252**	**0.7864689753166656**
KNN Regressor	0.811655569838343	0.7648410447043974	0.734203076978137
Lasso	0.6580630800394413	0.683640999182128	0.6536720919606233
Multiple Linear	0.6580630942955326	0.6836439383426598	0.6536695752128876
Ridge	0.6580630909744568	0.6836444387263839	0.653666585768011
Elastic Net	0.6320709061022907	0.6541556803876882	0.6206226725909685
Decision Tree Regressor	0.9993841300058871	0.6023396600684414	0.6195867930647317

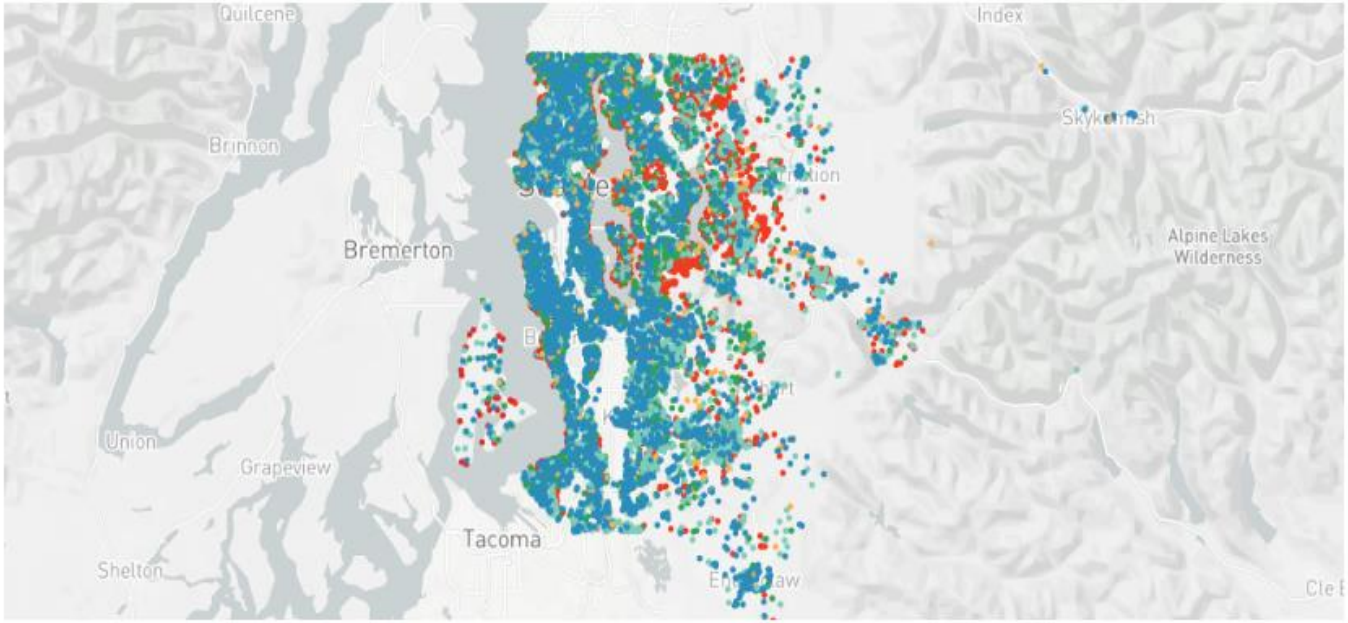
CLUSTERING MODELS

- **Custom Cluster**
  - Price < 250000 – Light Blue
  - 250000<price<500000 – Dark Blue
  - 500000<price<1000000 – Green
  - 1000000<price<3000000 – Yellow
  - 3000000<price<5000000 – Light Red
  - Price>5000000 – Dark Red
- **K-means**
  - **6 clusters** using important features
- **Ward's Linkage**
  - **6 clusters** using important features

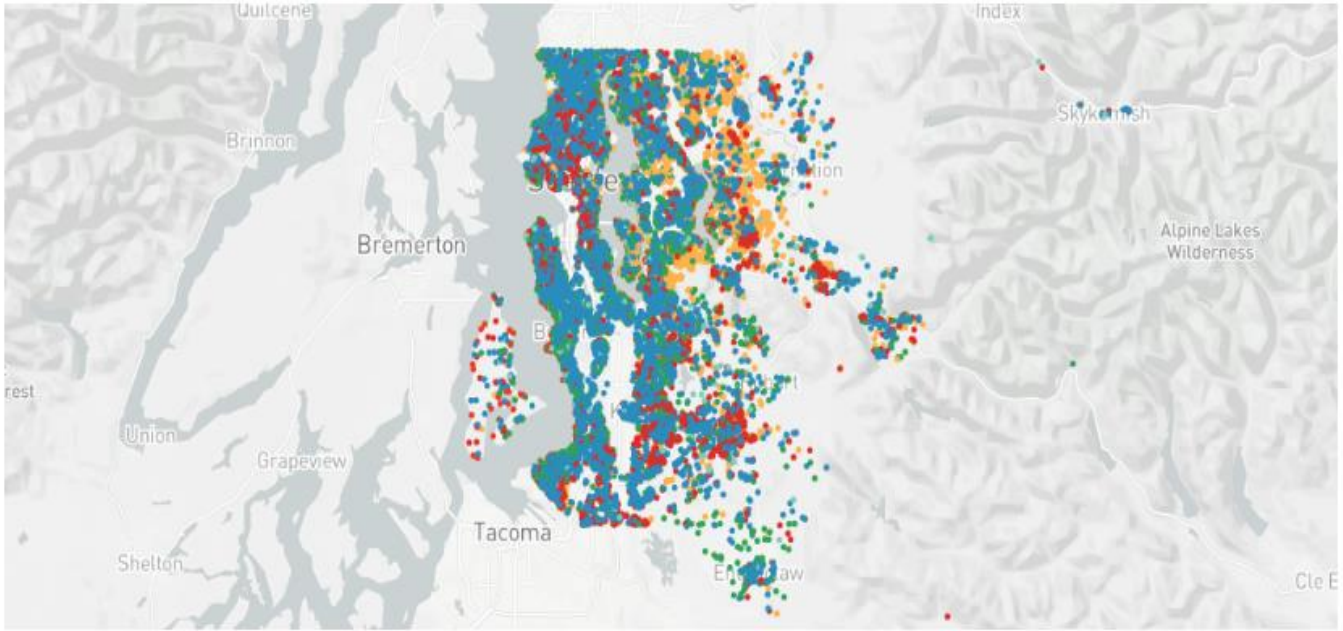
House prices in King County (Custom Price Clustering)



House prices in King County (K-Means) Clsutering



House prices in King County (Ward's Linkage) Clsutering



CONCLUSION

- Good Regression models
- Random Forest Regressor - 79% explanation is the variance of the dependent variable
- Bad Clustering Models due to the no clear trend in clusters of data points

THANK YOU