SG house price analysis

by Zhu Ai Ling

Outline

- **Data exploration and preprocessing**
- **Modeling**
- **Insights**

Data exploration

- **Data source:**
- https://data.gov.sg/dataset/resale-flat-prices
- Mar 2012-July 2017
- 100331 records , 9 attributes and resale_price

Data Exploration and preprocessing

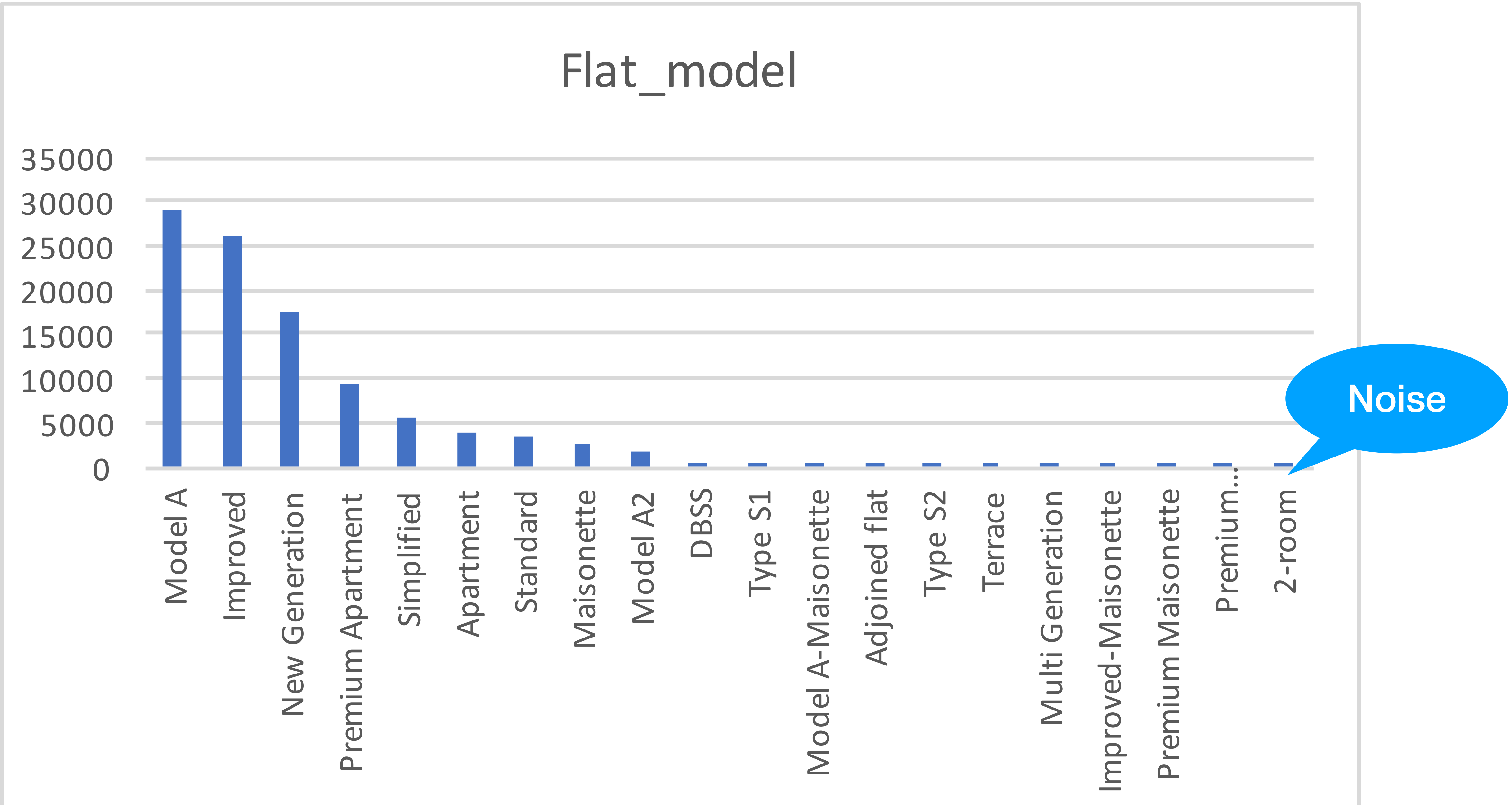
- **Data exploration:** Check and remove missing value, noise
- **Feature Engineering:** Create **new** features and drop **unwanted** features
- Further **Data exploration:** univariate analysis and bivariate analysis
- **Further Feature Engineering**:
  - Normalize numeric features
  - Create Dummy variables for the categorical features
- **Data splitting**: Split data into training and test

Data sample

Target

month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	resale_price
2012-03	ANG MO KIO	2 ROOM	172	ANG MO KIO AVE 4	06 TO 10	45	Improved	1986	250000
2012-03	ANG MO KIO	2 ROOM	510	ANG MO KIO AVE 8	01 TO 05	44	Improved	1980	265000
2012-03	ANG MO KIO	3 ROOM	610	ANG MO KIO AVE 4	06 TO 10	68	New Generation	1980	315000
2012-03	ANG MO KIO	3 ROOM	474	ANG MO KIO AVE 10	01 TO 05	67	New Generation	1984	320000
2012-03	ANG MO KIO	3 ROOM	604	ANG MO KIO AVE 5	06 TO 10	67	New Generation	1980	321000
2012-03	ANG MO KIO	3 ROOM	154	ANG MO KIO AVE 5	01 TO 05	68	New Generation	1981	321000

Data exploration: flat_model: noise



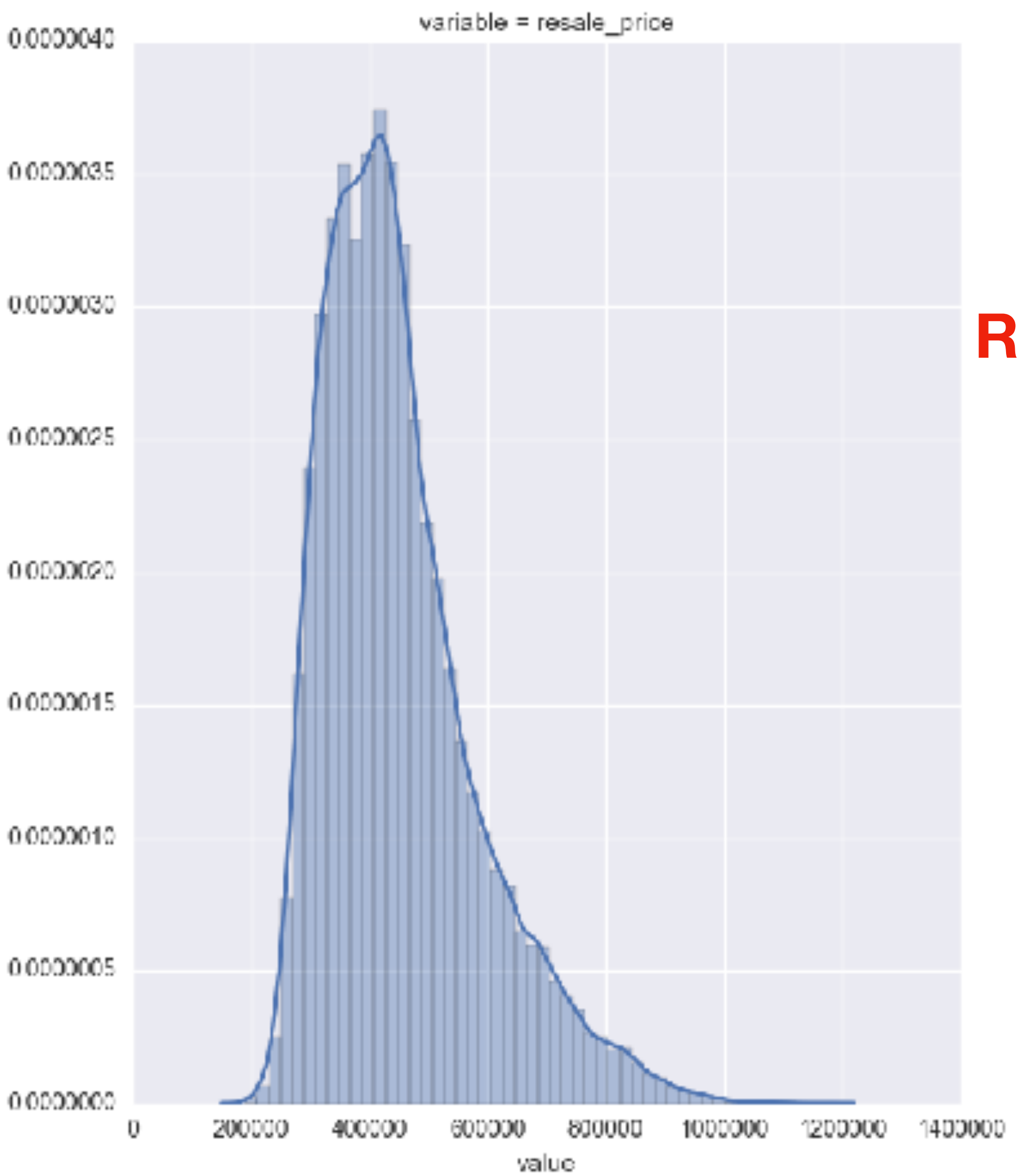
Data exploration and preprocessing: Feature engineering

- Create **New features** :
  - **age_at_sale** =‘year’-'lease_commence_date'
  - extract **year** and **month** from ‘year-month'
  - **price_per_sqm**=‘resale_price']/'floor_area_sqm'
- **Drop** unwanted features: unwanted = {'block', 'street_name'}

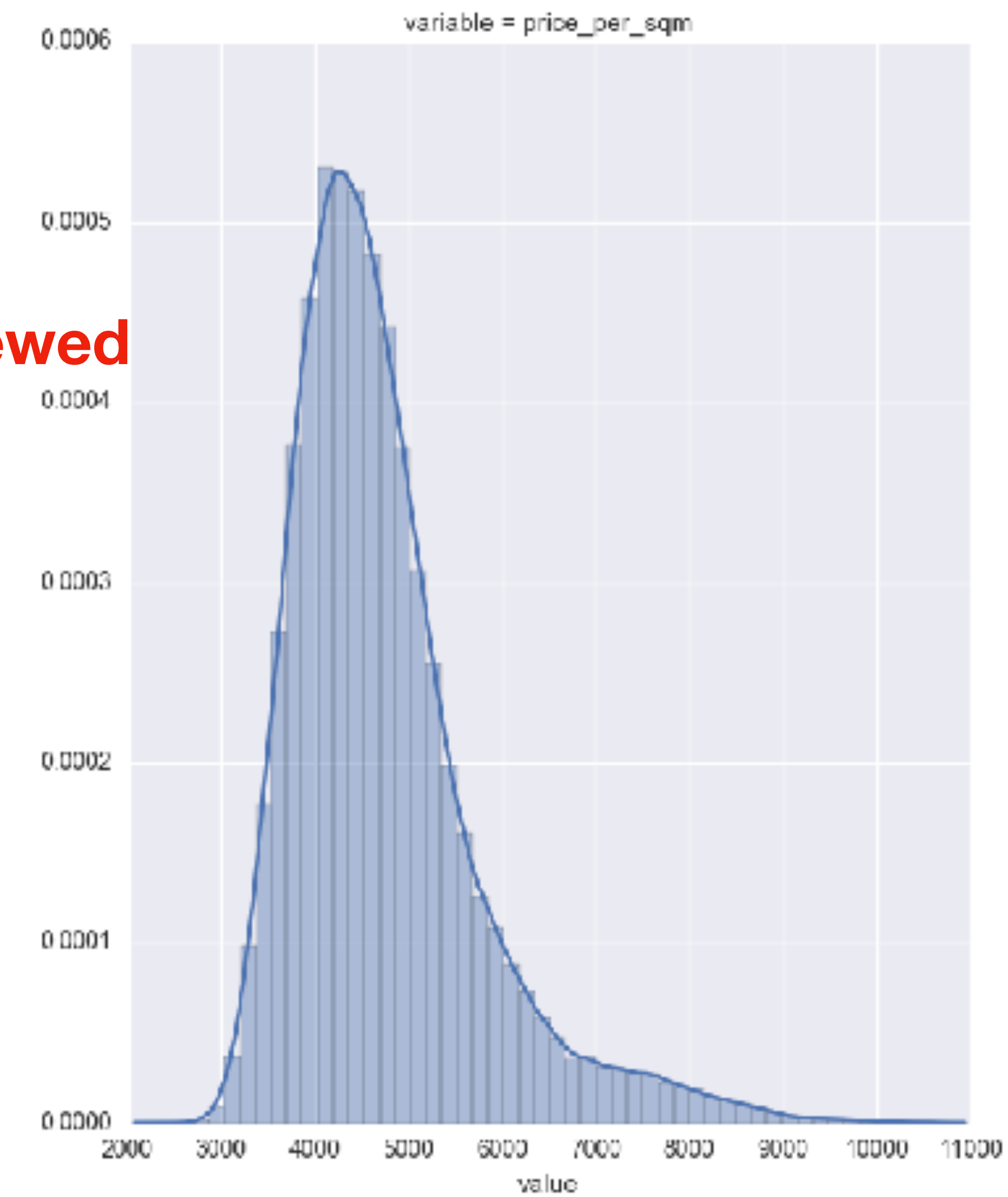
Data exploration: Data summary

	floor_area_sqm	resale_price	price_per_sqm	age_at_sale
count	**100331**	**100331**	**100331**	**100331**
mean	96.611177	450036.5626	4728.254694	23.99367095
std	24.60016607	130669.9166	1003.445685	10.60193973
min	31	190000	2375	1
25%	74	355000	4054.054054	15
50%	95	425000	4530.201342	26
75%	111	515000	5144.821492	32
max	280	1180000	10645.16129	51

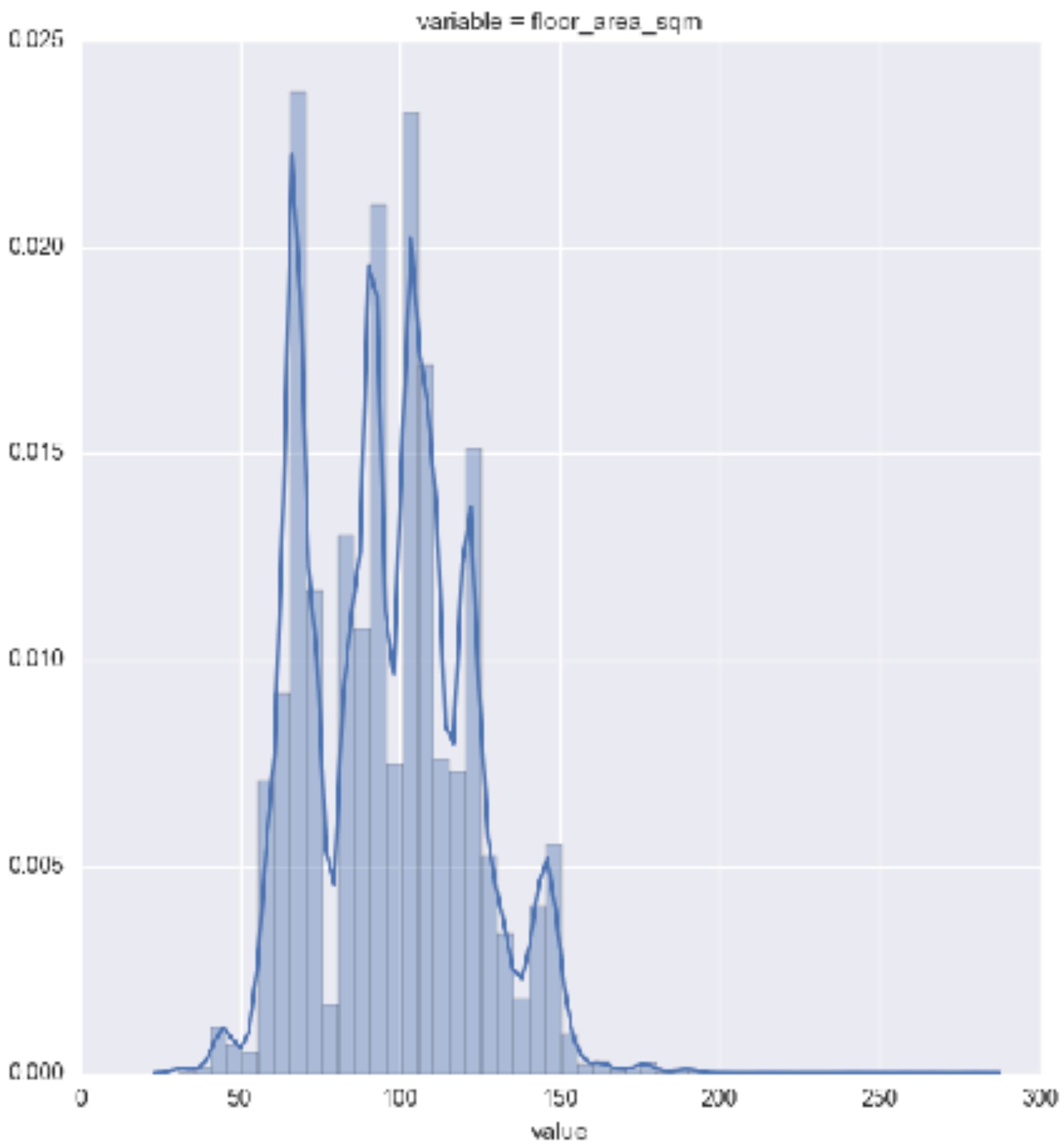
Data exploration:
Univariate analysis: distribution of resale_price

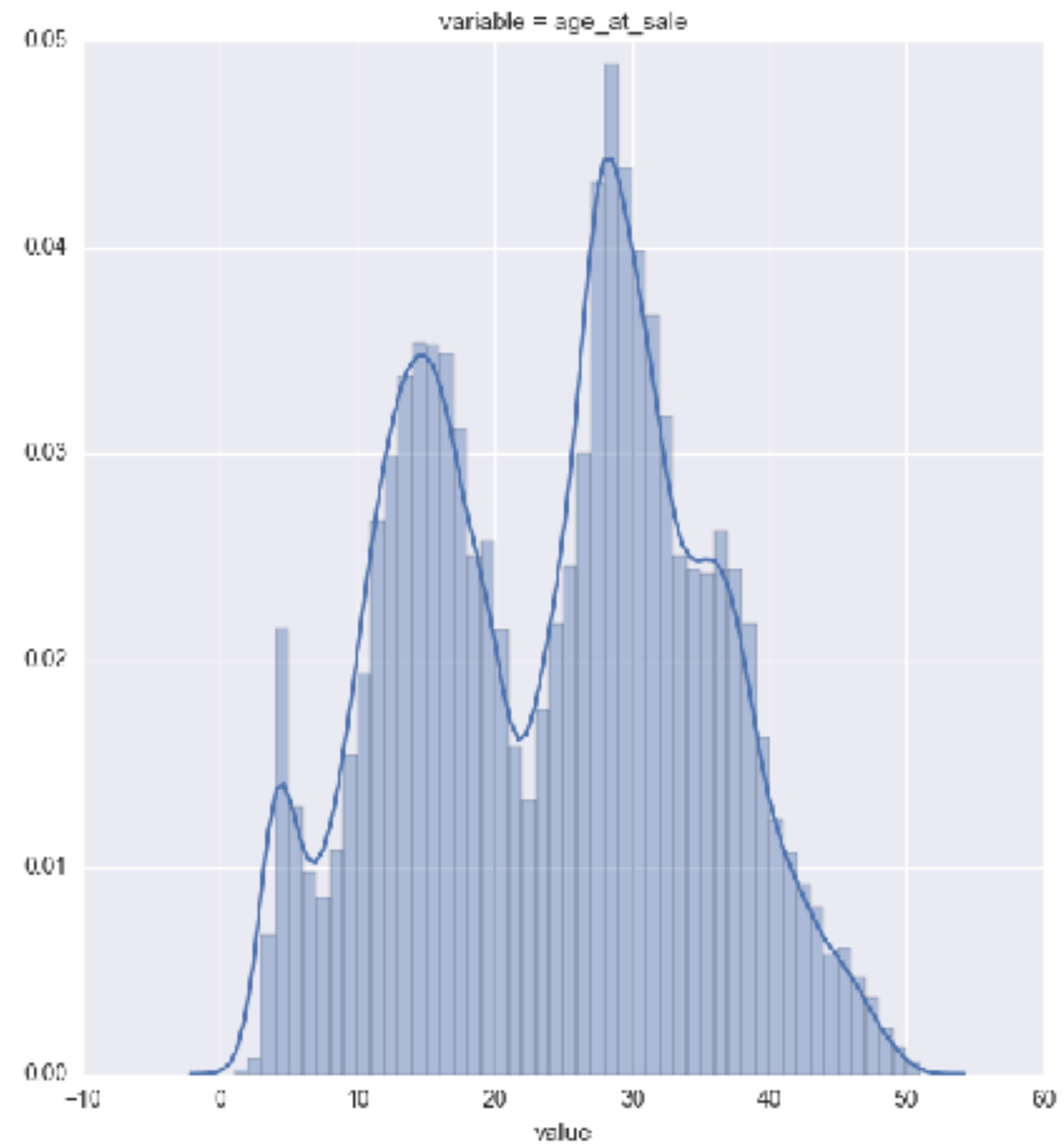


Right Skewed

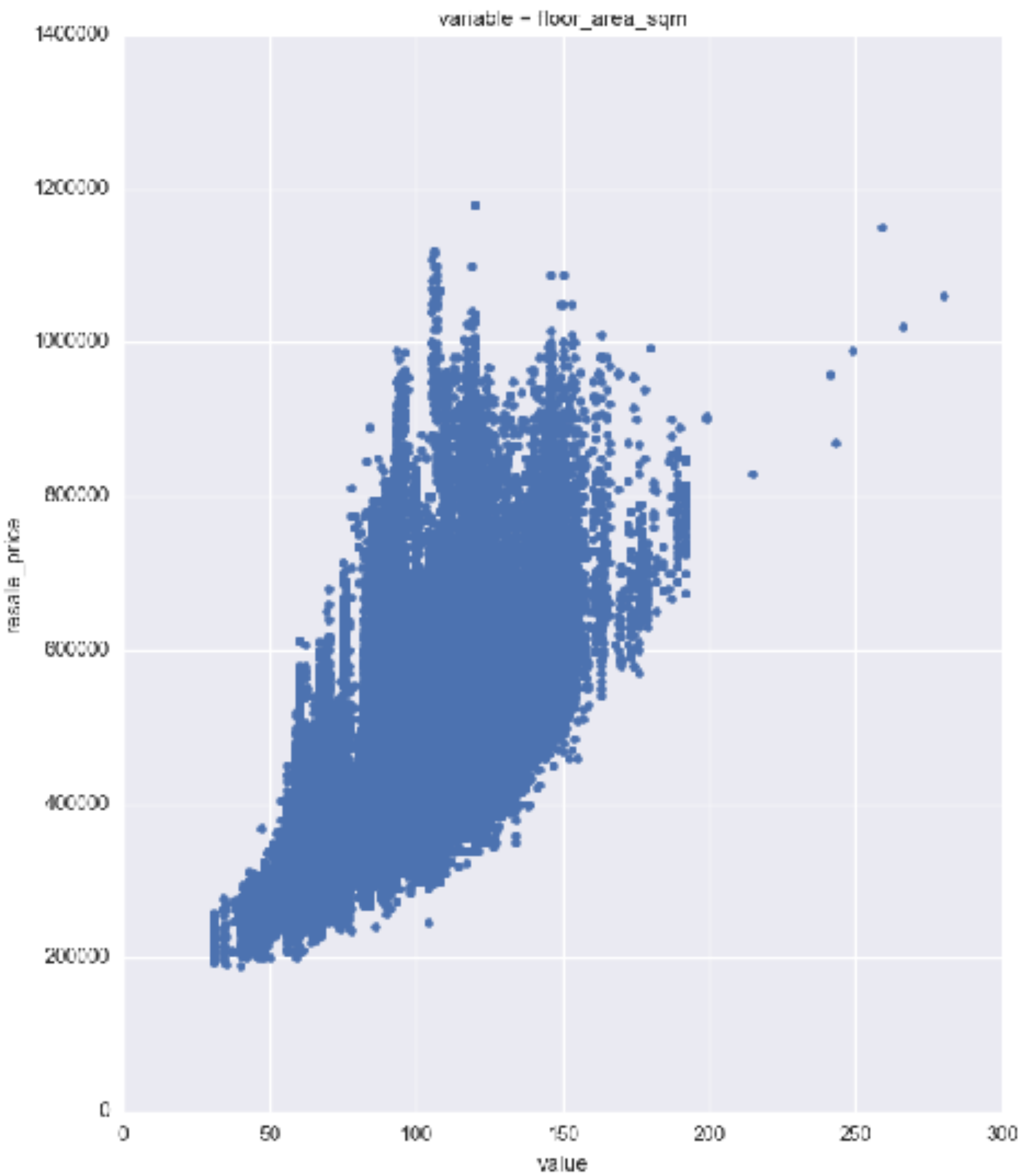


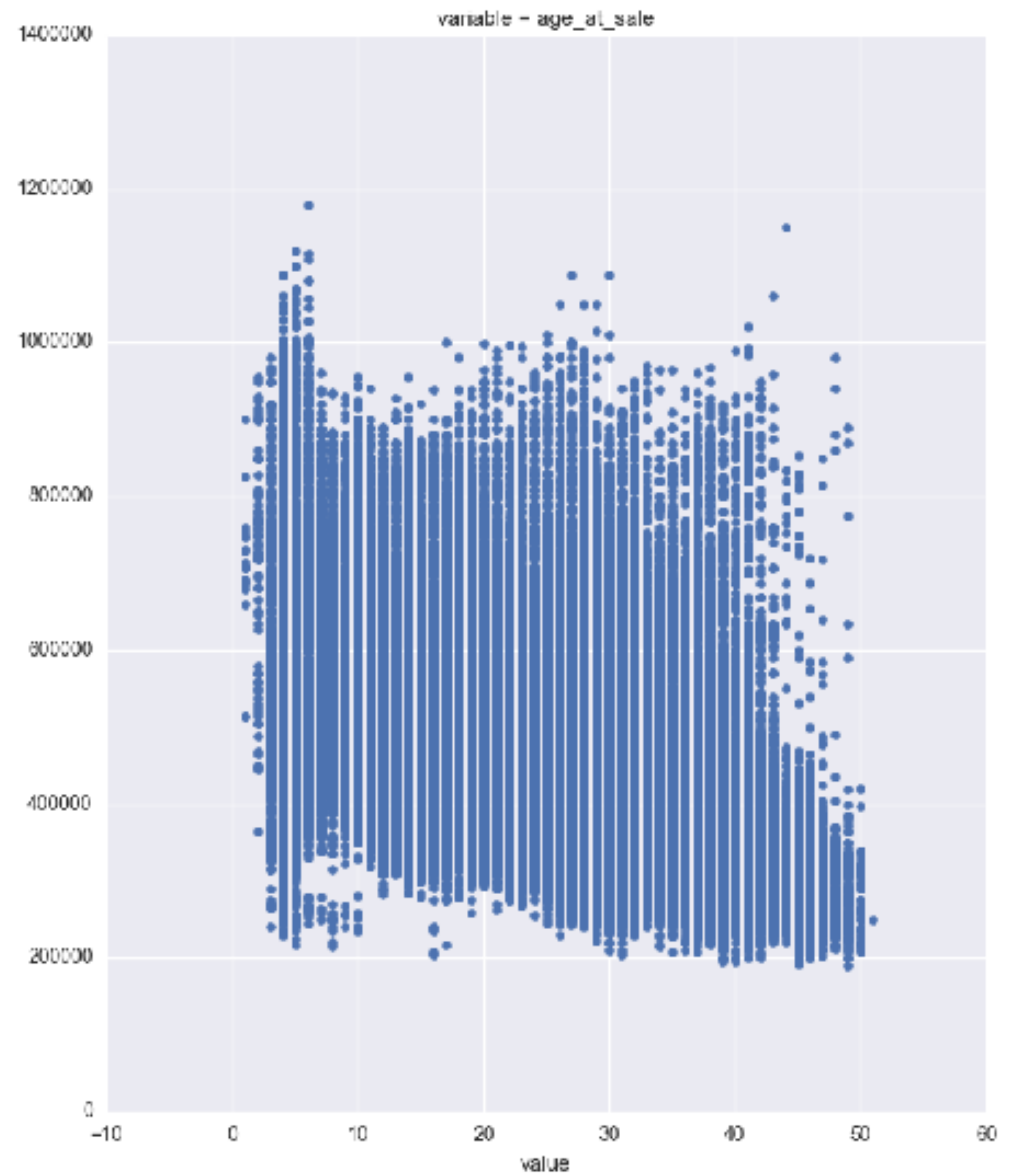
Univariate analysis: distribution of floor_area and age_at_sale



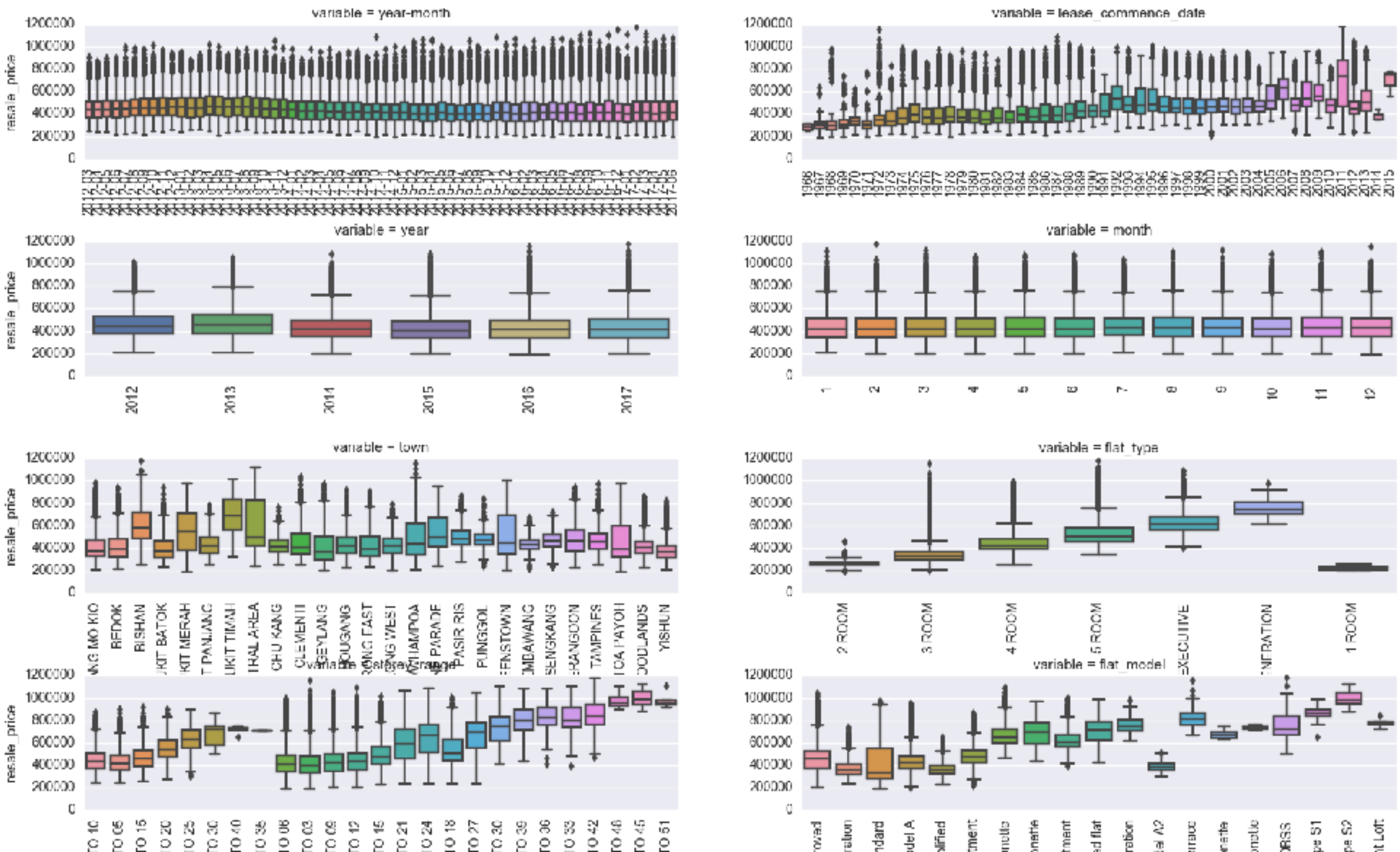


Bivariate analysis: resale_price vs. numerical variables





Bivariate analysis: resale_price vs. Categorical variables



Data Engineering

- **Normalize numeric features:**
  - transform the skewed numeric features by taking log(feature + 1)
- Create **Dummy variables** for the categorical features
- 100330 records, 211 features

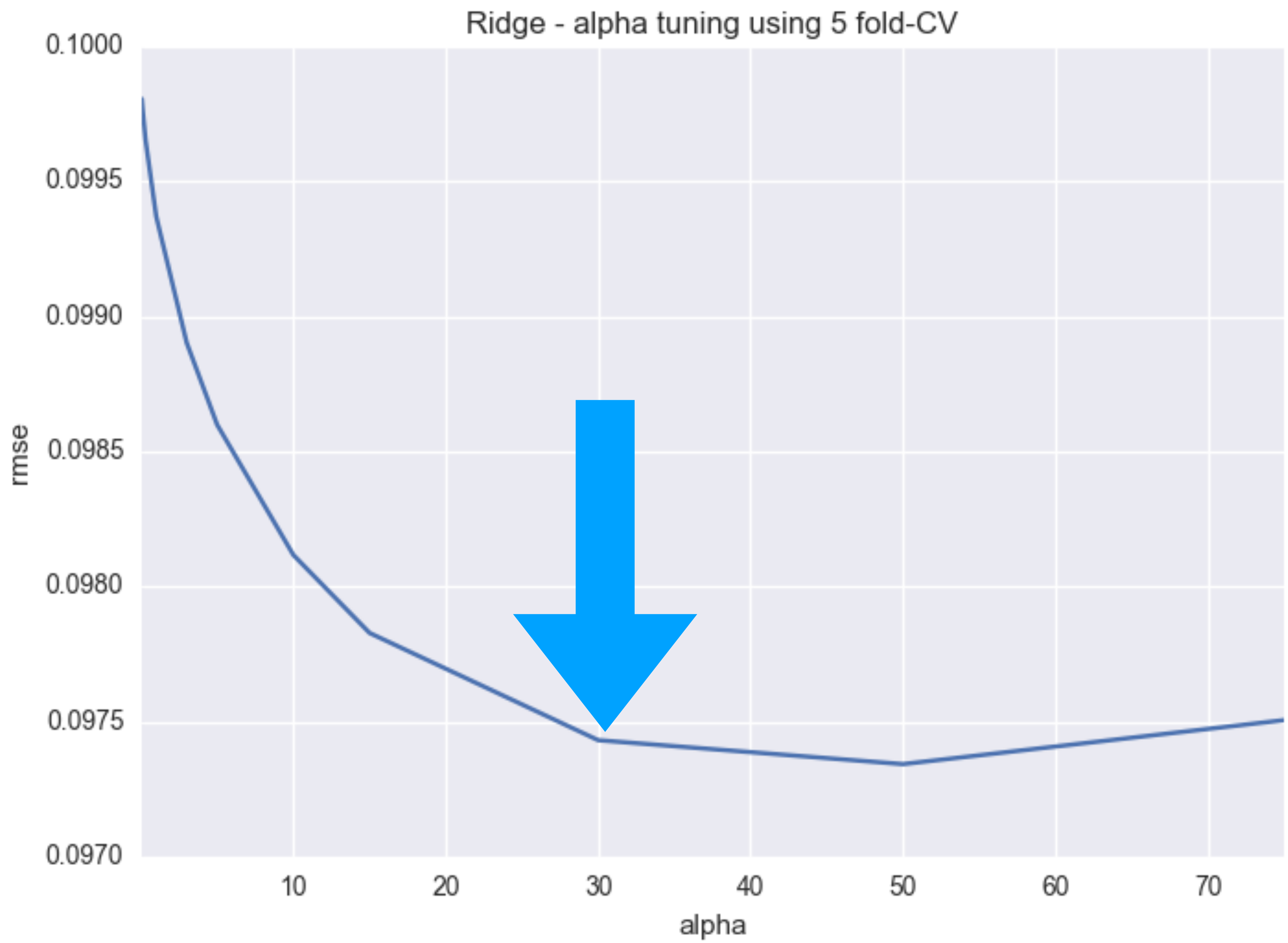
Data splitting

- Time series data:
- split data into training(2012-2015) and test (2016-2017)
- test: 2016-19379, 2017-9715 total (29094)
- train: 2012-1015: total (71236)

Regression modeling

- **Setting**:
  - Tune parameter using CV
  - Evaluation criteria: RMSE
- **Models considered:**
  - Simple: Ridge, Lasso
  - Ensemble: Random Forest, XGboost

Ridge-alpha tuning

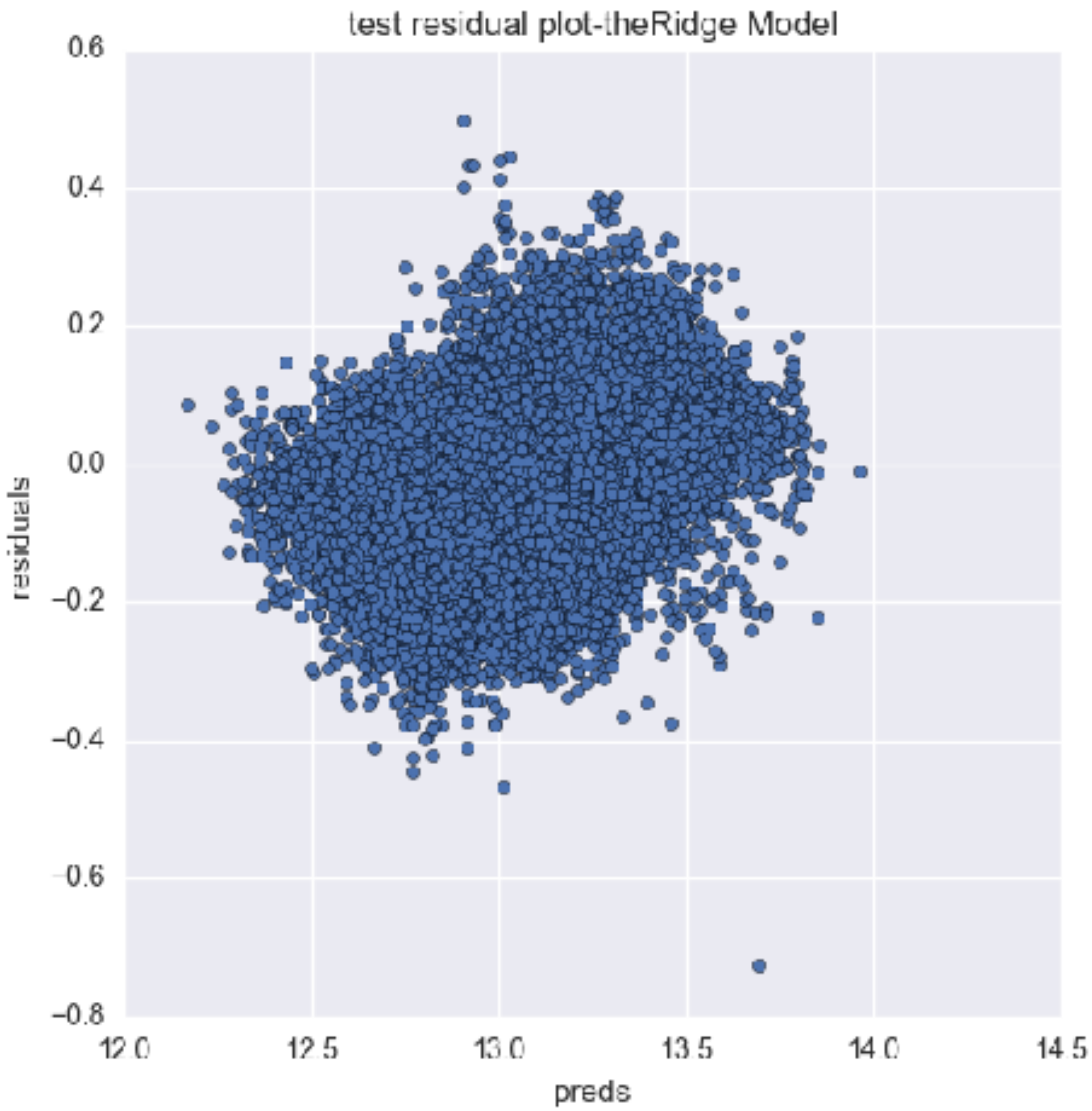


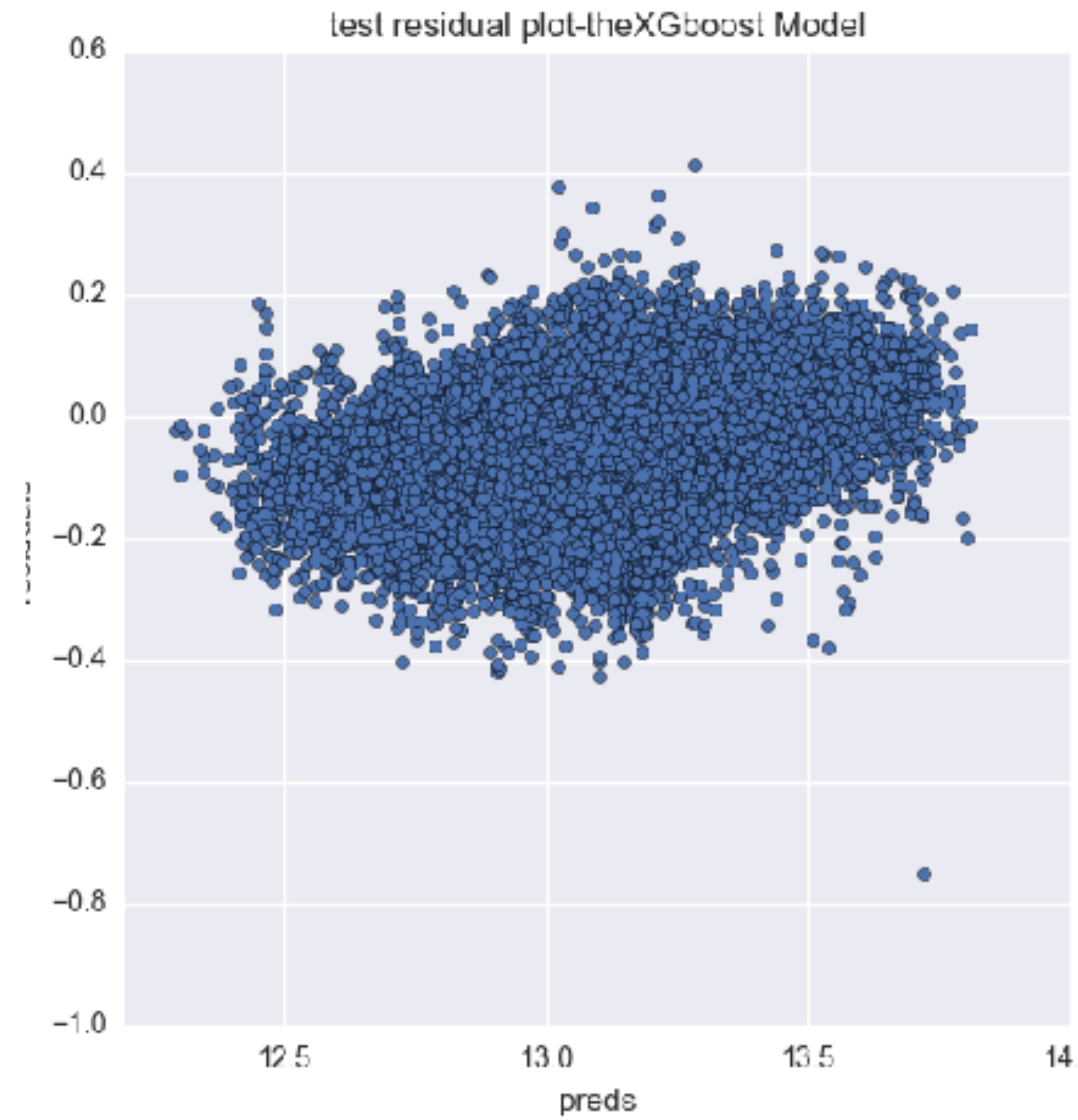
Regression modelling: Preliminary Results

Model	RMSE
Ridge	0.118403911
Lasso	0.126580114
RF	0.121871804
XGboost	0.115110421

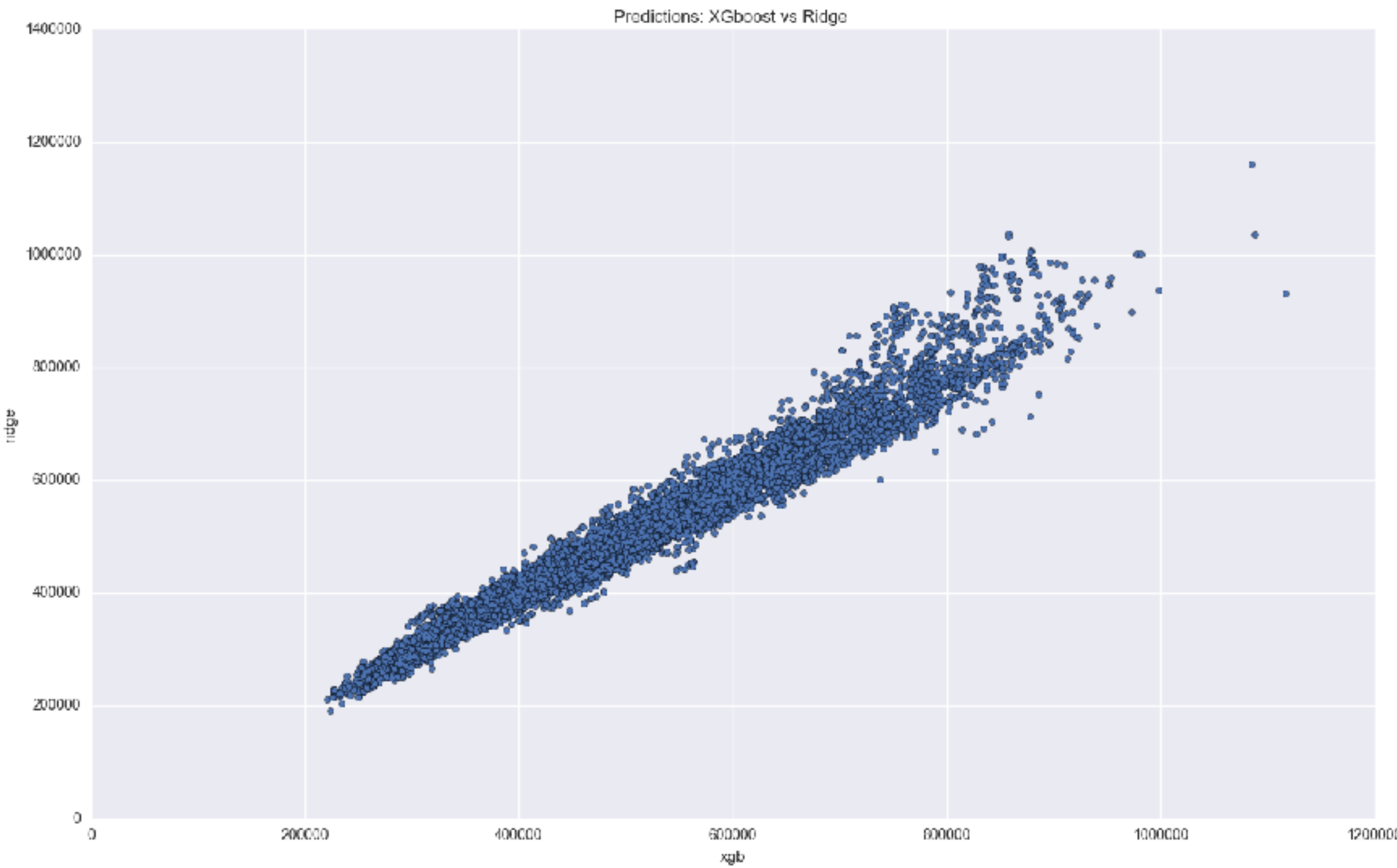
Ridge model does a good job

Regression modeling : Ridge vs XGboost

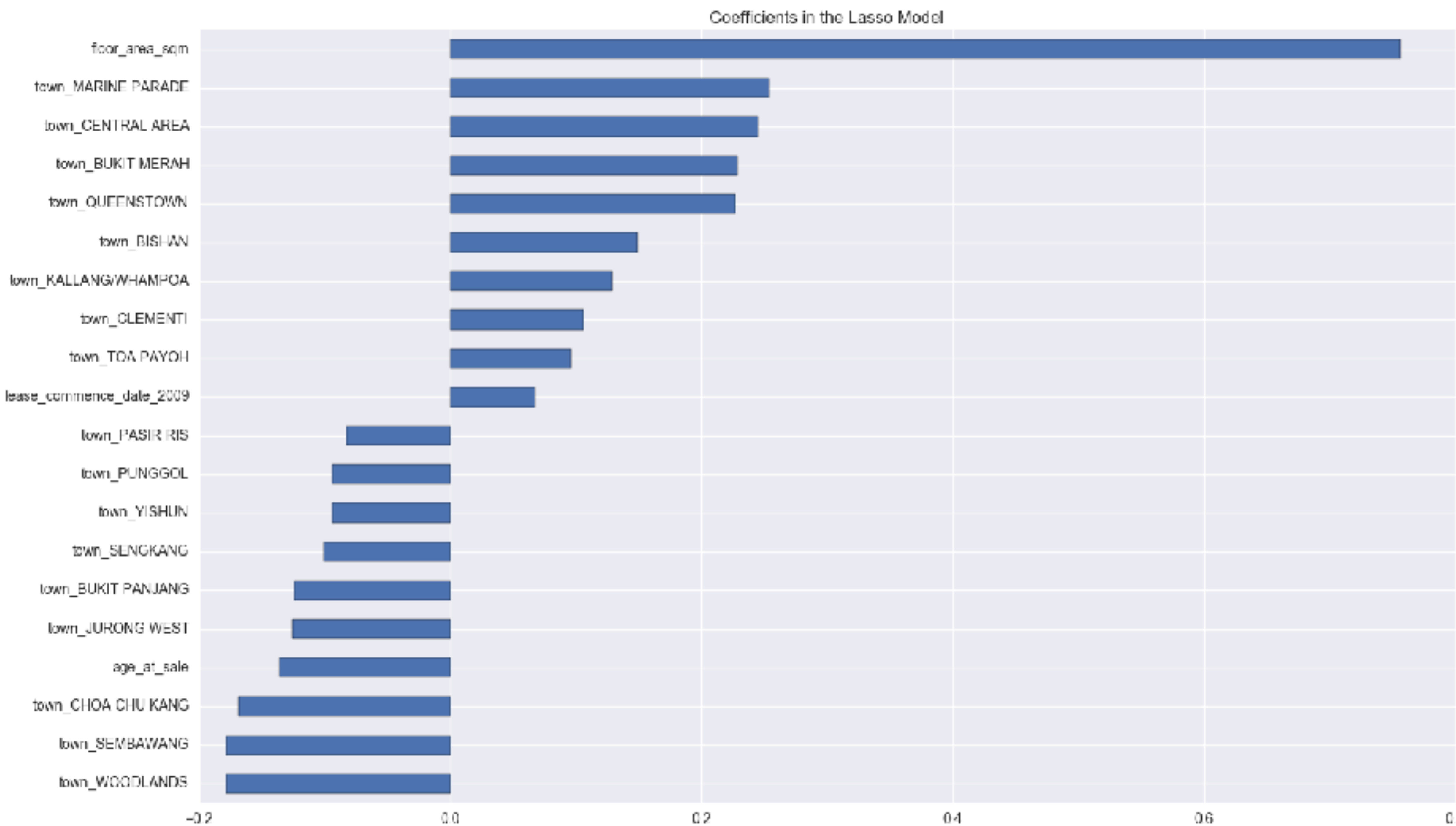




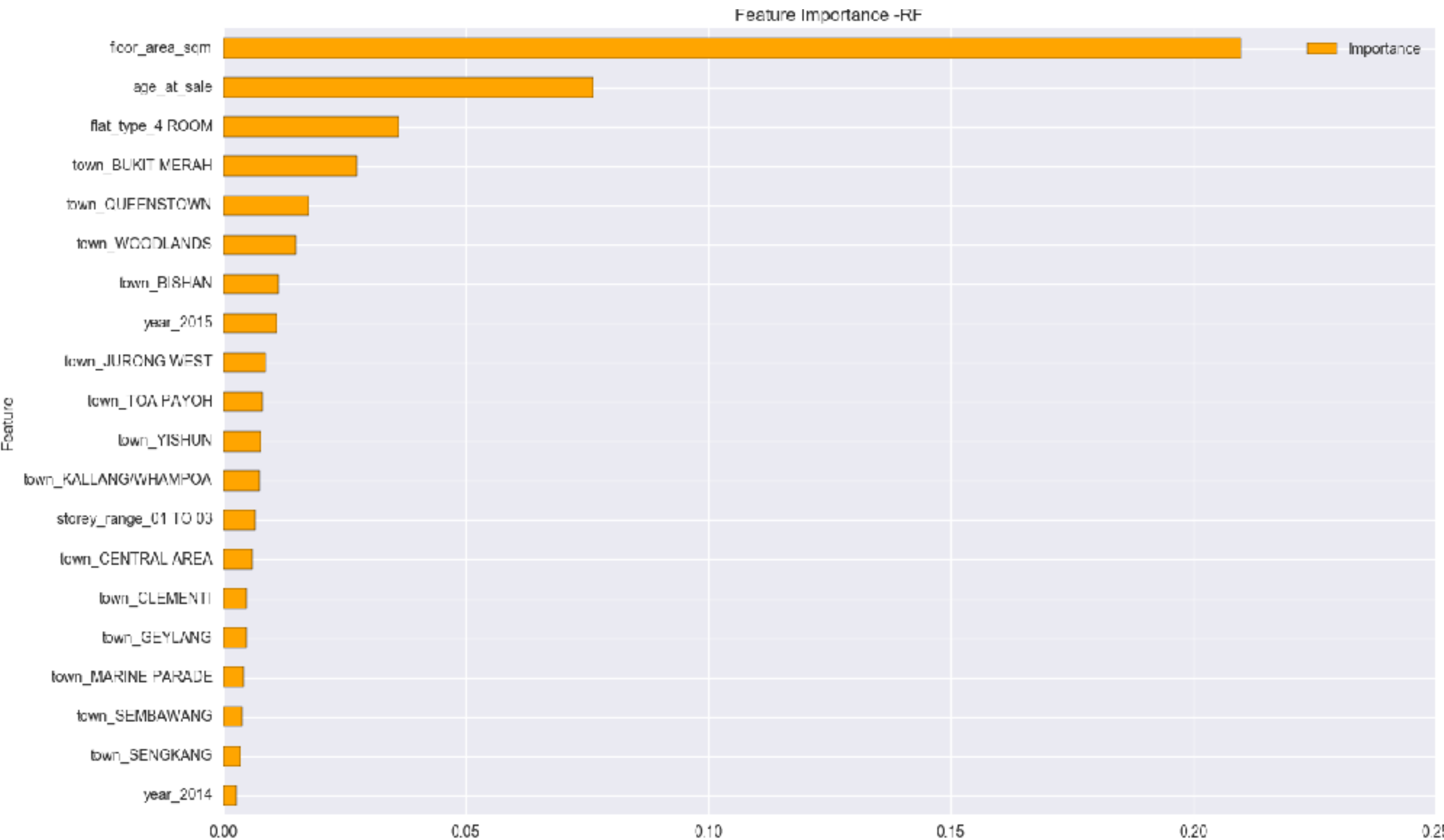
Regression modeling : Ridge vs XGboost



Feature Importance: Lasso coefficients



Feature Importance: RF



Insights

- Based on existing data and simple experiment setting:
  - **Simple model does a good job**
- More convincing conclusion can be drawn from:
  - **Nested Cross-validation and careful data splitting**
  - F**ine tuning parameters for XGboost**
- Performance can be improved by further **Feature Engineering**:
- **Augment data with other features** like **Neighbourhood** features: distance to CBD, MRT, Schools, also some **Economic** factors like consumer price index etc