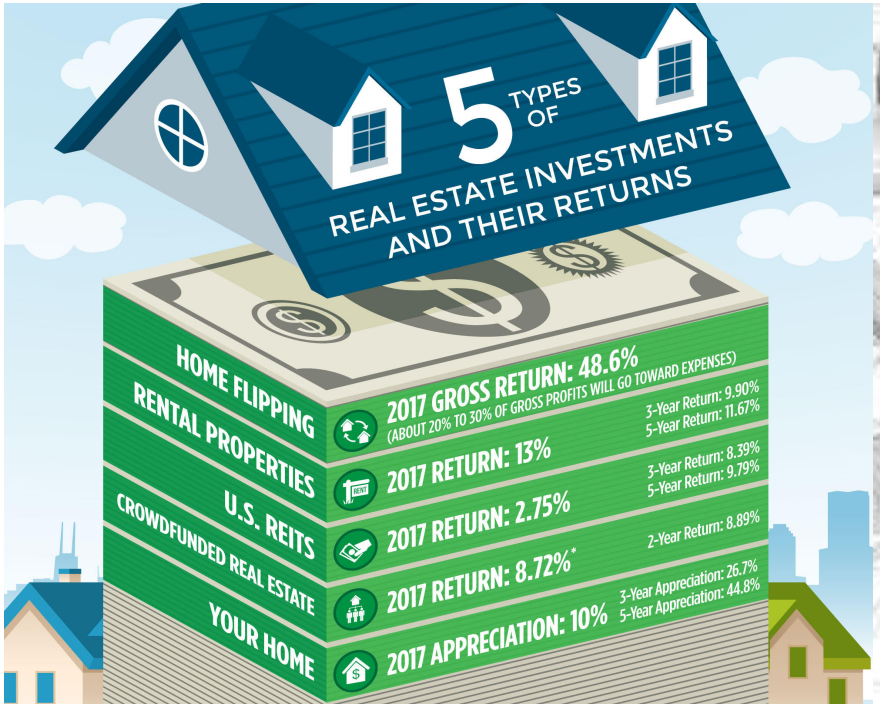
House Price Prediction

Shunling Guo, PhD

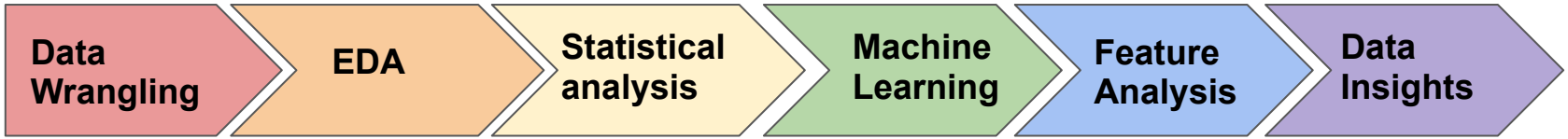
Overview

- Problem to solve and Audience
- Solution:
  - Data Cleaning and wrangling
  - Exploratory data analysis
  - Statistical analysis
  - Machine learning modeling
  - Case Story
- Data Insights

Predict house price is essential for Real Estate Investment and Development



Solutions



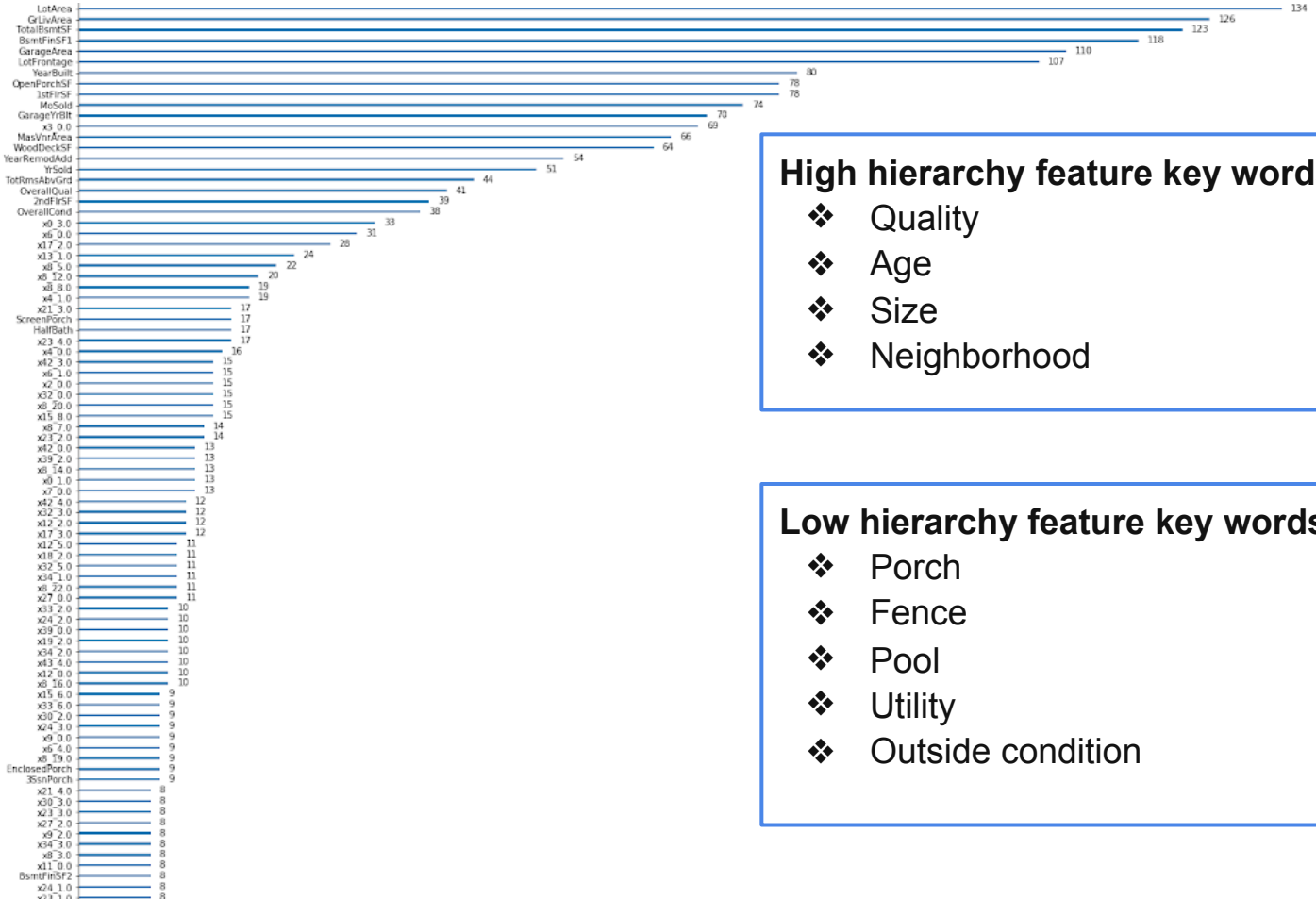
Data Wrangling	EDA	Statistical analysis	Machine Learning	Feature Analysis	Data Insights
•Fill NA •Parse date •Change year to age •preprocess categorical data to dummy table •normalize data	•Feature distribution •Feature correlation •Check Imbalance •Time series plot	•Infer statistics of house price for specific features or categories •Test hypothesis	•Polynomial regression with feature selection tuning •Random forest with parameter tuning •Xgboost with fine tuning	•Feature importance hierarchy •Case study of feature importance with causal/ correlation validation to mine driven forces.	•recommend feature for marketing •draw conclusion for investment •insight for real estate development

Results

- **Comparison of different regression models:**

regression	Best Cross validation r2 score	Best Testing score
Polynomial	0.84	0.76
Random forest	0.87	0.83
Xgboost	**>0.99**	**0.87**

- **Feature importance hierarchy:**



High hierarchy feature key words:

- ❖ Quality
- ❖ Age
- ❖ Size
- ❖ Neighborhood

Low hierarchy feature key words:

- ❖ Porch
- ❖ Fence
- ❖ Pool
- ❖ Utility
- ❖ Outside condition

Case study for mining driven forces:

- **Choose one feature from important feature category**:
  - Quality: OverallQual
  - Age: YearBuilt
  - Size: GrLivArea

- **Choose one feature from instantly thought important:**
  - Bedroom number: BedroomAbvGr

Time series view:

- Age of house and Sale Price correlation trends over time.
  Y axis is average house sale price (red, SalePrice) or average home age (black, YearBuilt) of sold houses within a month. X axis is Sold Time with precision of a month.
- We see a negative correlation along time between the two.

Time series view:

- House quality and Sale Price correlation trends over time.
  Y axis is average house sale price (red, SalePrice) or average home quality (black, OverallQual) of sold houses within a month. X axis is Sold Time with precision of a month.

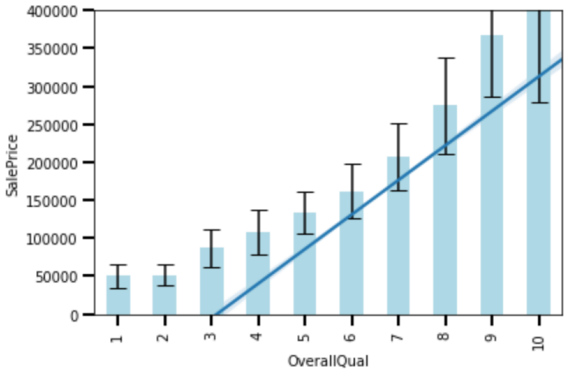
- We see a negative correlation along time between the two.

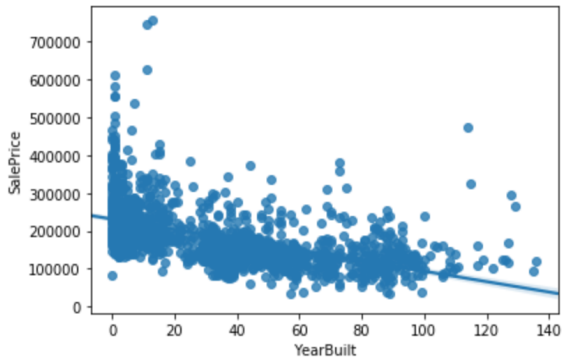
Time series view:

- Living area and Sale Price correlation trends over time.

  Y axis is average house sale price (red, SalePrice) or average living area (black, GrLivArea) of sold houses within a month. X axis is Sold Time with precision of a month.
- The correlation is harder to predict, most times are positively correlated, but sometimes see negative effect.

Statistical inference:



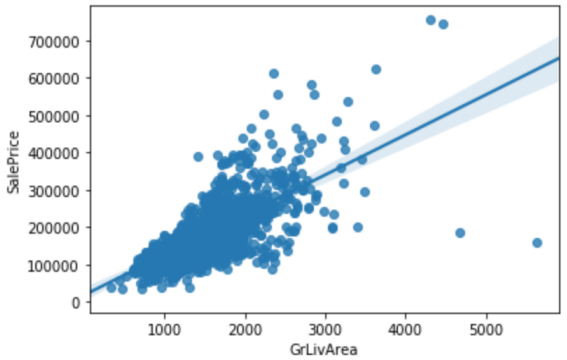


OLS Regression Results

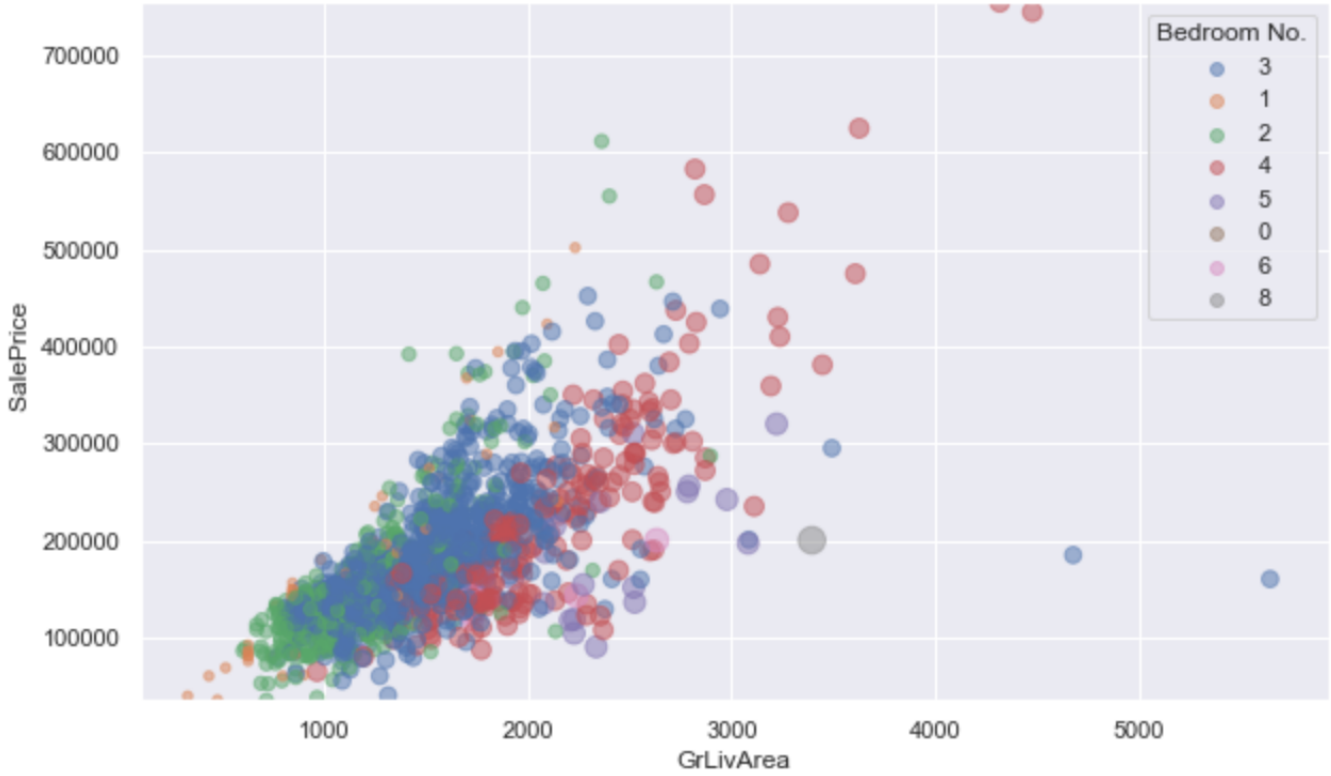
Dep. Variable:	OverallQual	R-squared (uncentered):	0.928
Model:	OLS	Adj. R-squared (uncentered):	0.928
Method:	Least Squares	F-statistic:	1.879e+04
Date:	Wed, 04 Sep 2019	Prob (F-statistic):	0.00
Time:	10:30:10	Log-Likelihood:	-2828.1
No. Observations:	1460	AIC:	5658.
Df Residuals:	1459	BIC:	5663.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
SalePrice	3.049e-05	2.22e-07	137.068	0.000	3.01e-05	3.09e-05

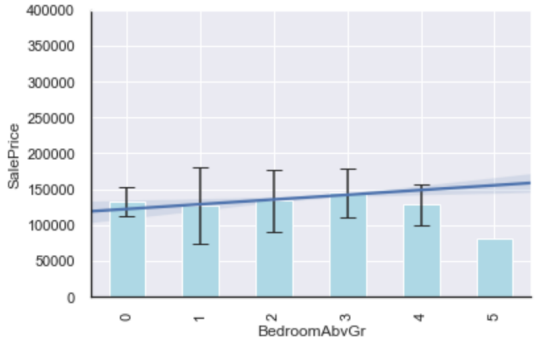
Omnibus:	812.455	Durbin-Watson:	1.794
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10480.098
Skew:	-2.319	Prob(JB):	0.00
Kurtosis:	15.278	Cond. No.	1.00

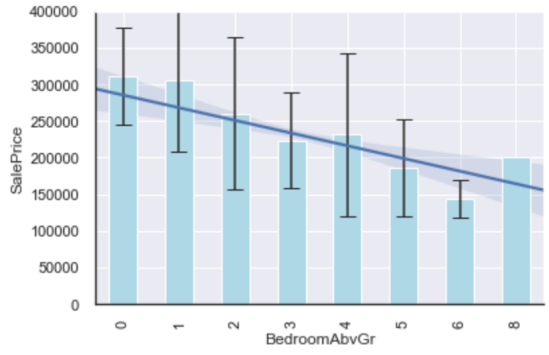


Bedroom number influence:

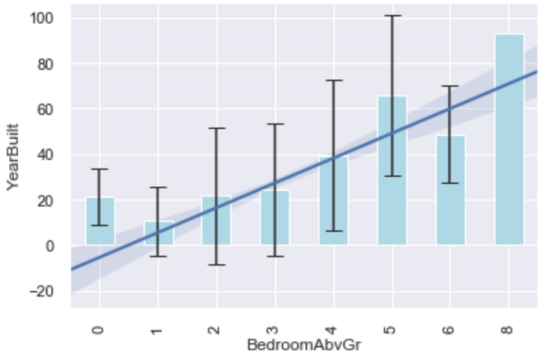


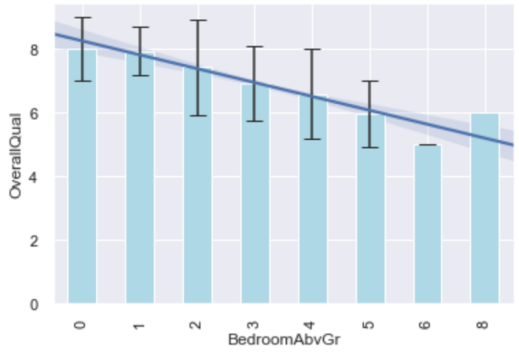
Bedroom number negatively correlated with house price in big houses (>1500 sqft).



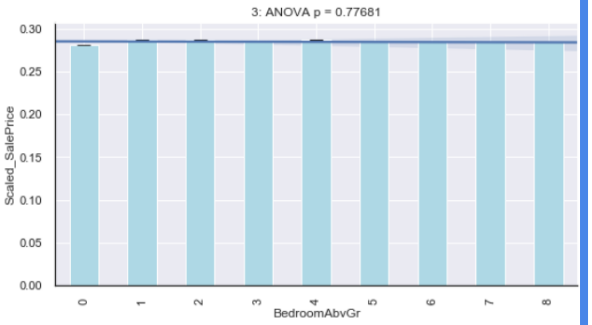


Bedroom number correlated with house age and quality.

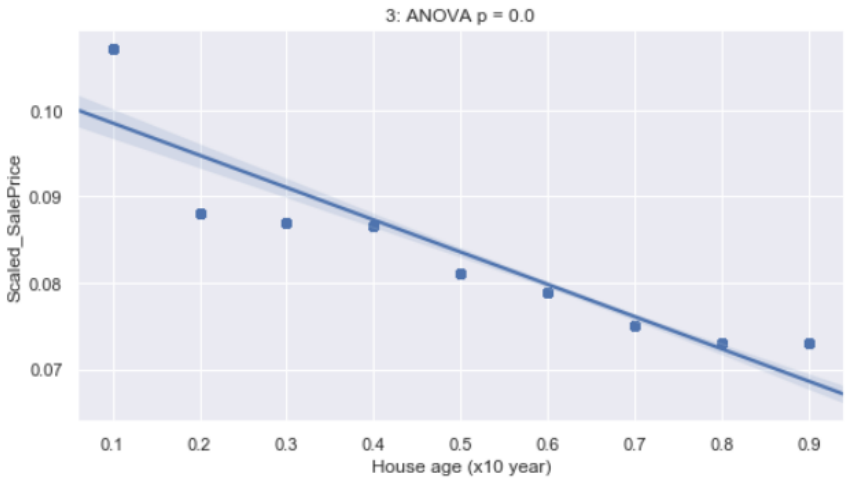


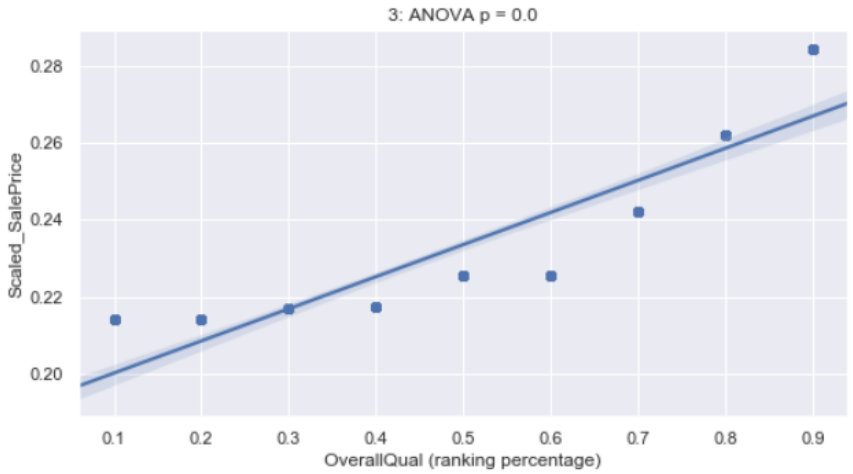


Bedroom number is not a driven force to affect house price according to model based simulation.



House age and quality are two driven forces:





Business insights:

- **Marketing:**
  - In this area, bedroom number is not a popular feature, therefore no need to highlight.
  - Top 10 important features:
    - Area of: Basement, living, garage, lot, porch and miscellaneous
    - Age of house and remodel
    - Month sold, wood deck
- **Investment:**
  - Price during time is quite stable, no appreciation value, but with low risk.
- **Real Estate Development:**
  - Using model to predict house prices and estimate revenue.