



DSI Project 2

Ames Housing Project

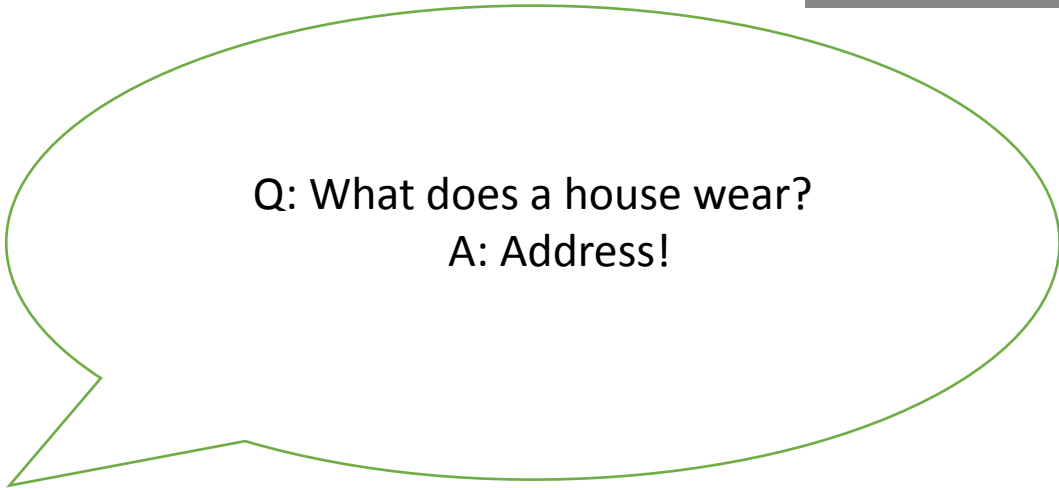
Group 2: Cindy, Shin, Chun Shan and Jason



Problem statement

- We are from a data-driven real estate company that wants to determine the most important characteristics that affect housing prices in Ames, Iowa.
- Target audience: Investors in our company

Sound effect:
Ba-dum-tss

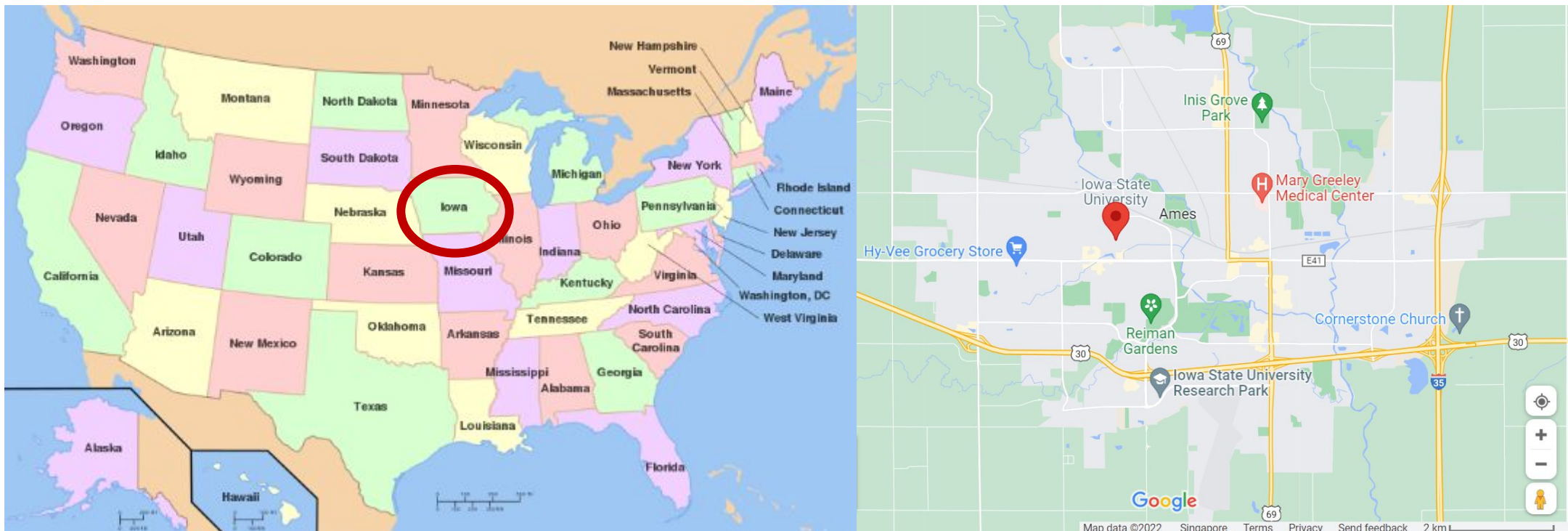


Q: What does a house wear?
A: Address!

Problem statement	EDA	Regression Model	Variables	Discussion
-------------------	-----	------------------	-----------	------------

Where is Ames located geographically?

- USA, Iowa
- Home to Iowa State University



Problem statement	EDA	Regression Model	Variables	Discussion
-------------------	-----	------------------	-----------	------------

- Method:
 - To fit a regression model to determine the important determinants of Ames housing sale prices, and identify the most important variables.
- Questions to answer:
 - Which features appear to add the most value to a home?
 - Which features hurt the value of a home the most?
 - What are things that homeowners could improve in their homes to increase the value?
 - What neighborhoods seem like they might be a good investment?

Problem statement	EDA	Regression Model	Variables	Discussion
-------------------	-----	------------------	-----------	------------

2 additional observation identifiers: *PID* and *Id*

Y variable = *SalePrice*

Ordinal Variables (Numerical) - 23

E.g. Overall condition of the house rating, overall material and finish rating, Type of utilities available

Nominal Variables (Categorical) - 23

E.g. Type of dwelling, Exterior covering on house, Garage Location

Continuous Variables (Numerical) - 20

E.g. Size of garage, Lot Area, Pool Area

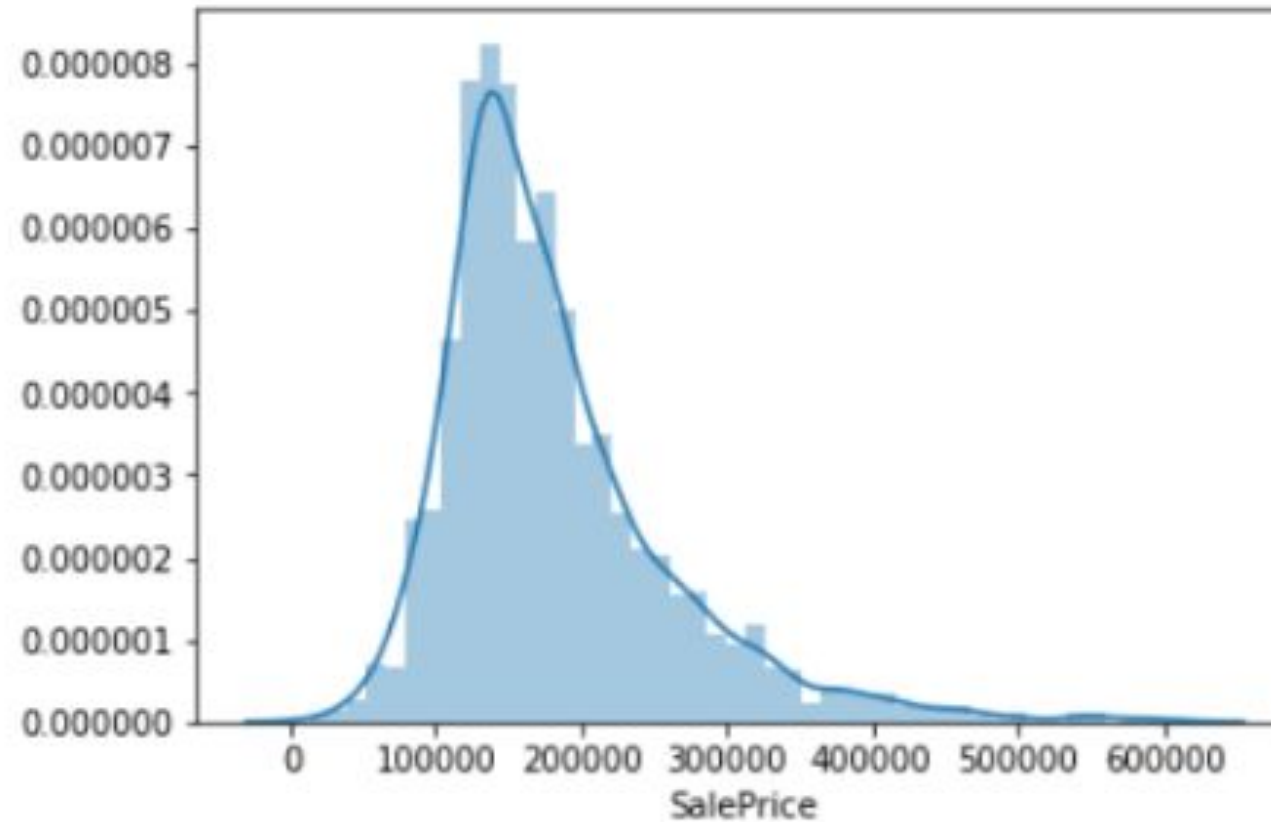
Discrete Variables (Numerical) - 14

E.g. Month Sold, Year Sold, Bedrooms Above Grade, Number of Fireplaces



Optional features for houses:
Garage, Pool, Fireplace, Basement, Driveway

Frequency of Sale Prices in Ames Housing



Problem statement	EDA	Regression Model	Variables	Discussion
-------------------	-----	------------------	-----------	------------

Are any variables strongly correlated?

For numerical data, we need to do something about the multicollinearity:

- **'Gr Liv Area' and 'TotRms AbvGrd' (corr = 0.81)**
 - TotRms AbvGrd can be removed as it has a lower correlation to SalePrice
- **'Garage Cars' and 'Garage Area' (corr = 0.90)**
 - Garage Area can be removed as Garage Cars is easier to visualize a discrete variable
- **'Garage Yr Blt' and 'Year Built' (corr = 0.78)**
 - The Garage and house are usually built in the same year, GarageYrBlt can be removed
- **'1st Flr SF' and 'Total Bsmt SF' (corr = 0.79)**
 - The 1st Flr SF and Total Bsmt SF can be combined (with the area of other floors) to form a new variable through feature engineering

Final model characteristics

R^2 : 0.913

Variables with the largest positive coefficients

Neighborhood_GrnHill	Exter Qual
Neighborhood_StoneBr	Overall Qual
MS SubClass_30	Kitchen Qual
MS SubClass_45	Garage Cars
MS SubClass_75	
Garage Type_NA	

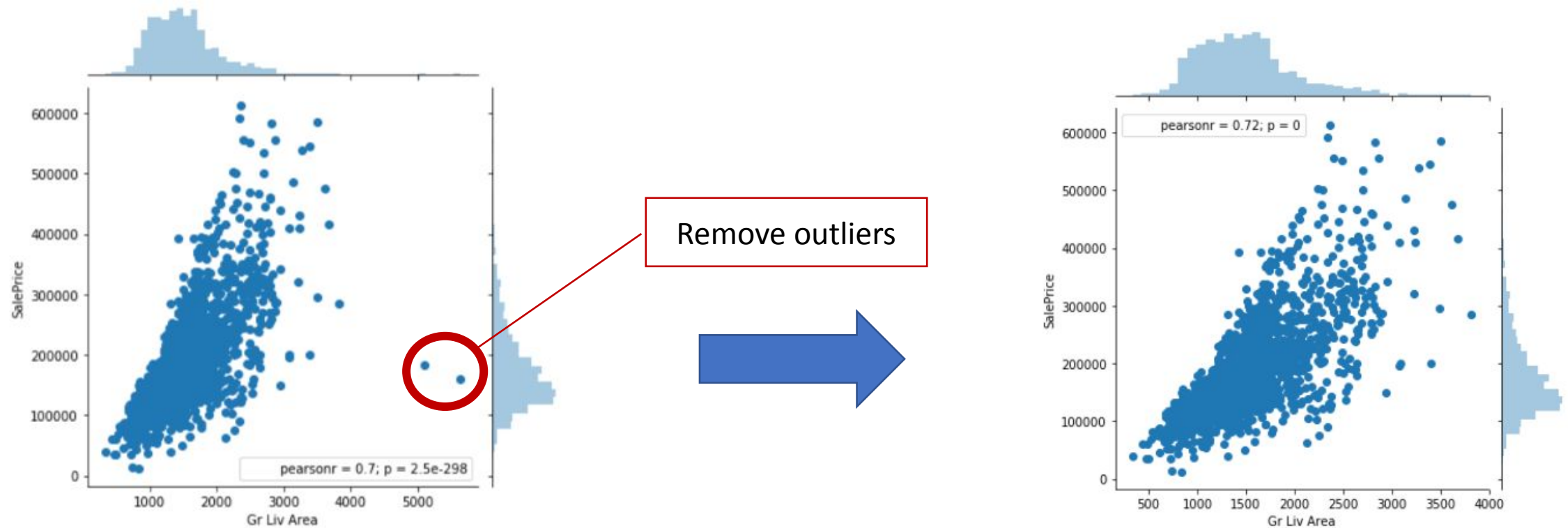
Variables with the largest negative coefficients

Roof Style_Mansard	Yr Sold
Neighborhood_NWAmes	Neighborhood_OldTown
Neighborhood_Gilbert	Neighborhood_NAmes
Neighborhood_SawyerW	Bedroom AbvGr

Problem statement	EDA	Regression Model	Variables	Discussion
-------------------	-----	------------------	-----------	------------

Continuous Variables (Numerical)

- 2 houses are outliers as they have the largest areas but low sale prices.

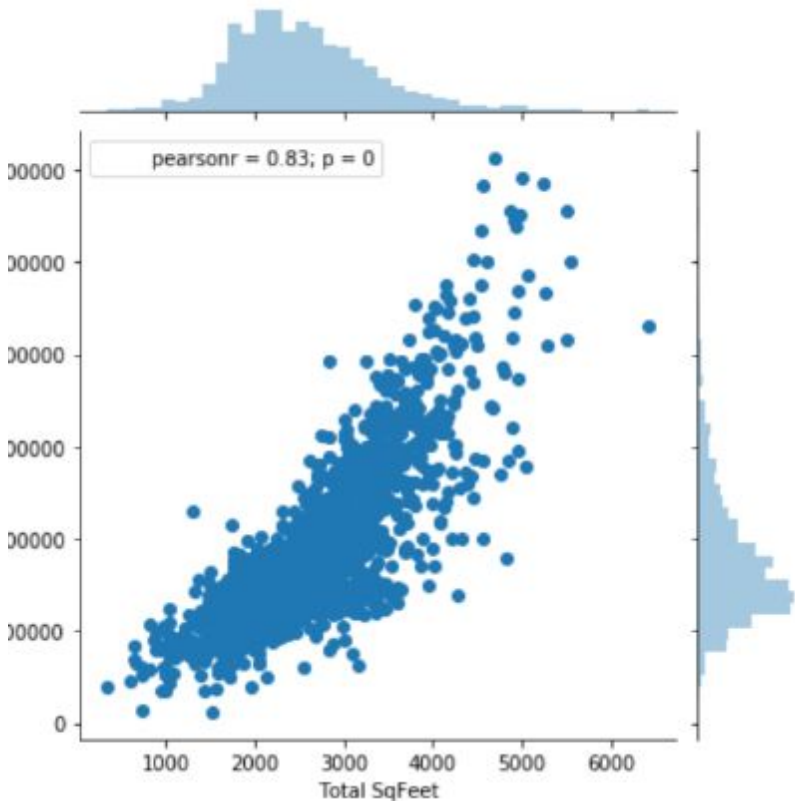
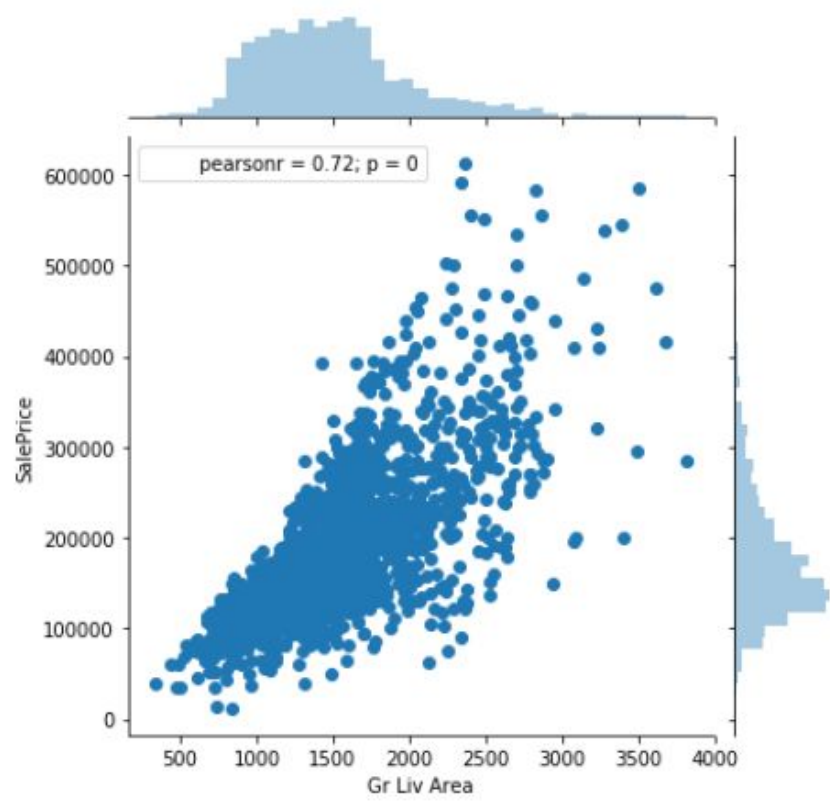


Which variable is a better predictor of Sale Price?

Total SqFeet = Total Bsmt SF + 1st Flr SF + 2nd Flr SF

Gr Liv Area (Continuous): Above grade (ground) living area square feet

Total Sq Feet (obtained through Feature Engineering)



Correlation between Gr Liv Area and Total Sq Feet = 0.856

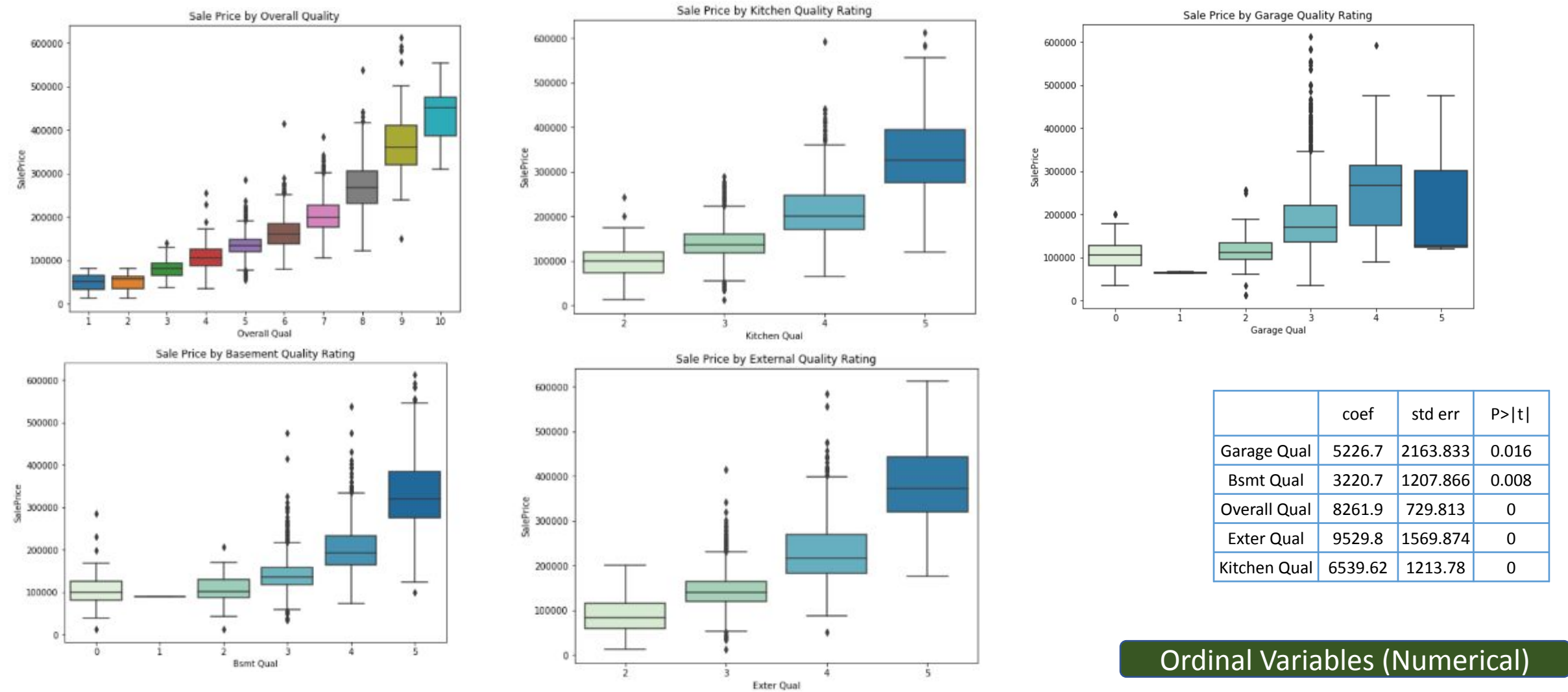
> Drop Gr Liv Area

'Total Sq Feet' coef	std err
56.8759	2.365

Continuous Variables (Numerical)

Better Quality ratings significantly increase Sale Price

Ex	Excellent	5
Gd	Good	4
TA	Average/Typical	3
Fa	Fair	2
Po	Poor	1



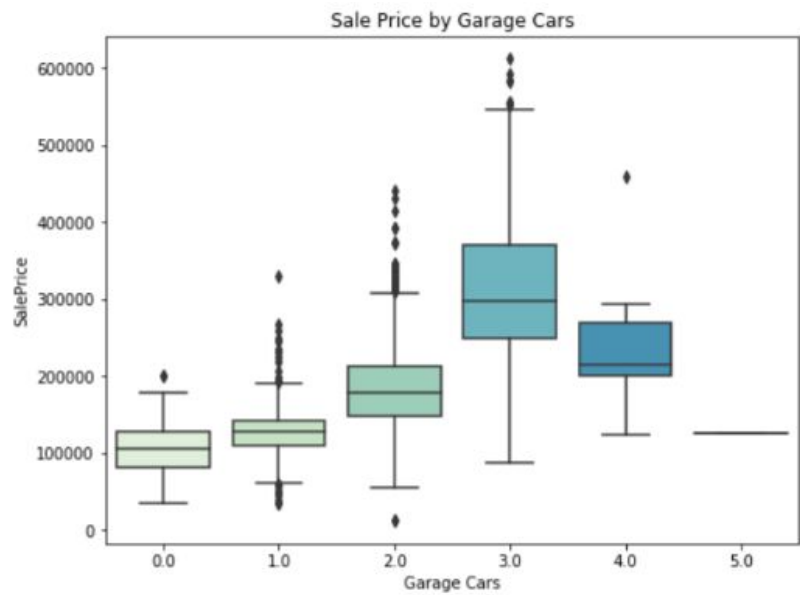
	coef	std err	P> t
Garage Qual	5226.7	2163.833	0.016
Bsmt Qual	3220.7	1207.866	0.008
Overall Qual	8261.9	729.813	0
Exter Qual	9529.8	1569.874	0
Kitchen Qual	6539.62	1213.78	0

Ordinal Variables (Numerical)

Note: '0' means there is no basement

How Garage affects Sale Price

The mean sale price of the house peaks at car capacity of 3 and decreases for values of *Garage Cars* > 3.



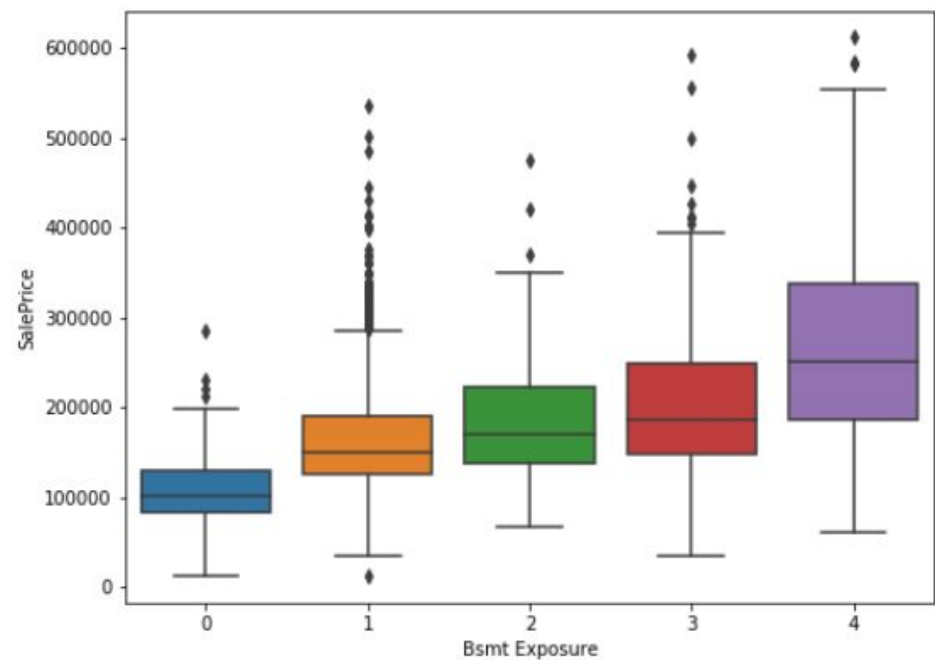
- Among houses with a garage, the ones that can afford to have the largest garages tend to be further away from the city where land is cheaper per sq foot, hence overall sale prices are lower.

Variable	coef	std err	t
Garage Type_NA	2.63E+04	6835.559	3.846
Garage Cars	6206.3205	1134.808	5.469



How Basement affects Sale Price

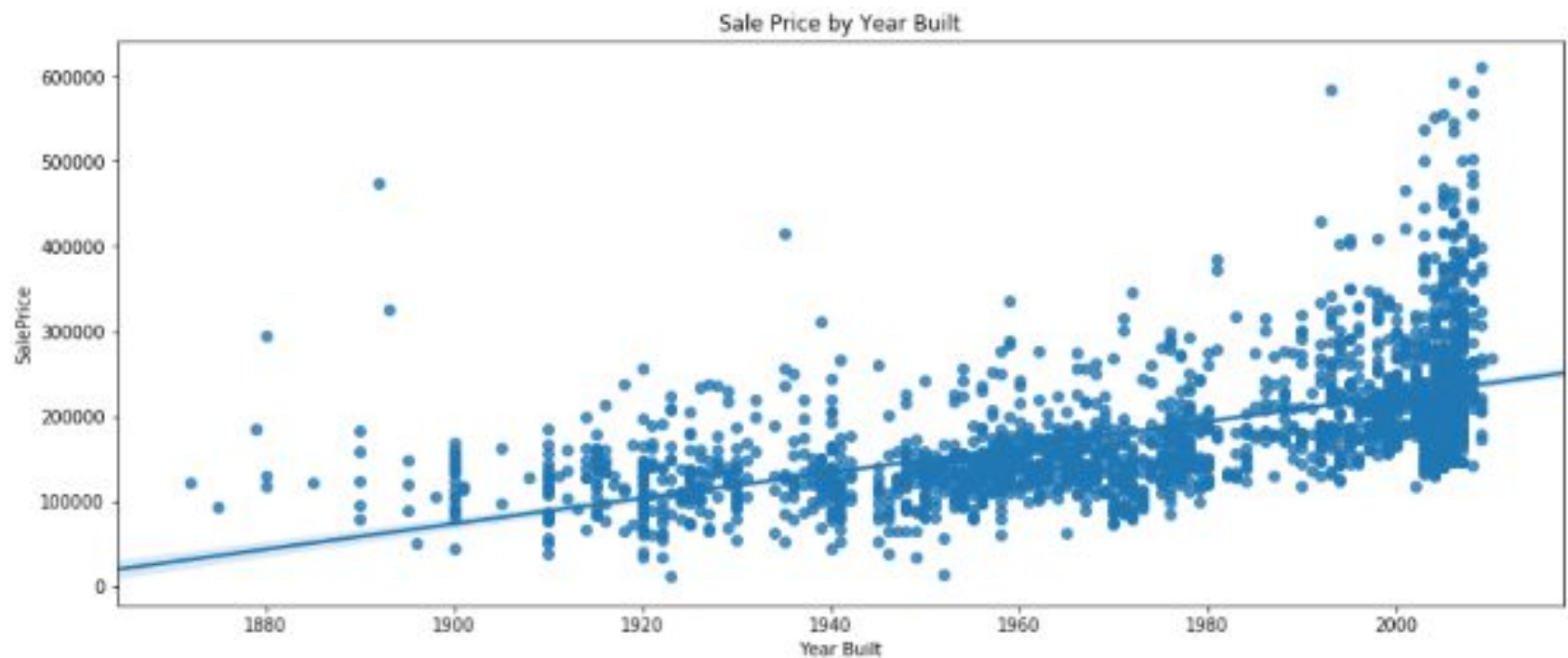
Higher Basement exposure rating increases the sale price of a house significantly.
Basement sq feet are inversely related to Sale Price (could be because of maintenance costs).



Variable	Explanation	coef
Bsmt Exposure	Refers to walkout or garden level walls	3543.9872
BsmtFin SF 1	Type 1 finished square feet	-17.0737
BsmtFin SF 2	Type 2 finished square feet	-30.4843
Bsmt Unf SF	Unfinished square feet of basement area	-40.7742

Sale price increase across the years houses are built

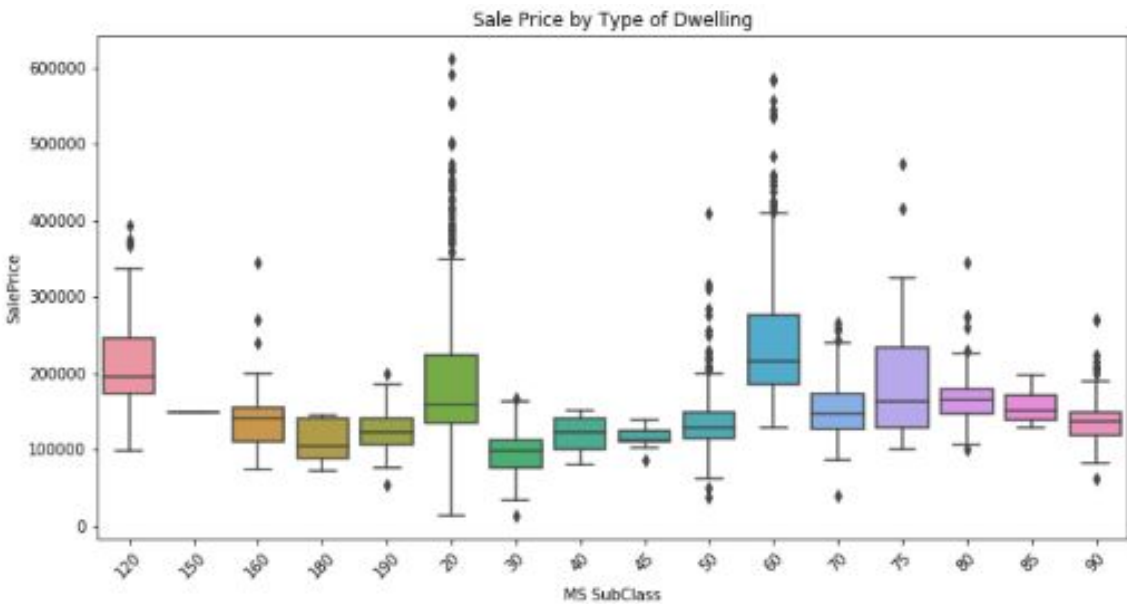
- New houses command a higher Sale Price
- Possible inflation of housing prices over years



	coef
Year Built	361.9918

Discrete Variables (Numerical)

Which type of dwelling tend to have high or low sale prices?

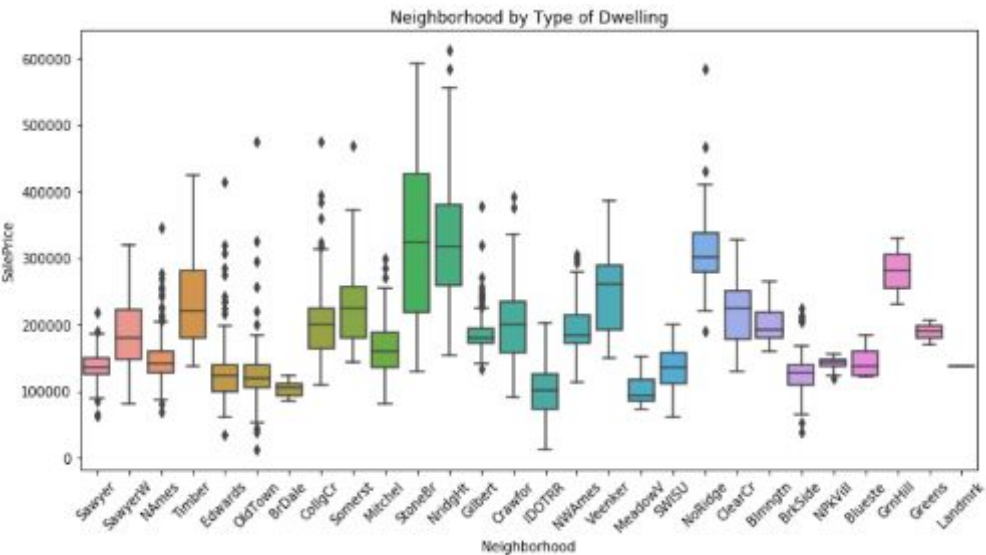


Variable	Subclass	coef
MS SubClass_30	1-STORY 1945 & OLDER	3.55E+04
MS SubClass_45	1-1/2 STORY - UNFINISHED ALL AGES	3.42E+04
MS SubClass_75	2-1/2 STORY ALL AGES	3.35E+04
MS SubClass_40	1-STORY W/FINISHED ATTIC ALL AGES	3.25E+04
MS SubClass_70	2-STORY 1945 & OLDER	3.02E+04
MS SubClass_50	1-1/2 STORY FINISHED ALL AGES	2.74E+04
MS SubClass_20	1-STORY 1946 & NEWER ALL STYLES	2.64E+04
MS SubClass_60	2-STORY 1946 & NEWER	2.27E+04
MS SubClass_80	SPLIT OR MULTI-LEVEL	2.09E+04
MS SubClass_85	SPLIT FOYER	2.02E+04
MS SubClass_90	DUPLEX - ALL STYLES AND AGES	4338.3159
MS SubClass_150	1-1/2 STORY PUD - ALL AGES	-4.57E+04

Nominal Variables (Categorical)

Which neighbourhood in Ames tend to have high or low sale prices?

	coef
Neighborhood_Edwards	-4747.9275
Neighborhood_CollgCr	-6993.9083
Neighborhood_Gilbert	-1.09E+04
Neighborhood_GrnHill	1.12E+05
Neighborhood_NAmes	-7457.669
Neighborhood_NWAmes	-1.56E+04
Neighborhood_NoRidge	1.35E+04
Neighborhood_NridgHt	2.39E+04
Neighborhood_OldTown	-8228.4859
Neighborhood_SawyerW	-1.07E+04
Neighborhood_StoneBr	3.61E+04



Nominal Variables (Categorical)

Problem statement	EDA	Regression Model	Variables	Discussion
-------------------	-----	------------------	-----------	------------

Applications as a real estate company

- Things homeowners can do to increase the value of the house:
 - Downsize (or remove) the basement and convert it into general living area
 - Upgrade the mansard roof to more modern options

Problem statement	EDA	Regression Model	Variables	Discussion
-------------------	-----	------------------	-----------	------------

Limitations of the model

- Dataset was provided only for a limited time period. It is difficult to predict housing prices on the future because it is heavily dependent on the economy in general.
- Model probably cannot be generalised across USA cities because each city has different population size, income levels, etc. We would need to have to adjust the Sale Prices to the income levels and population of each state after finding out the 2 variables' relationship with housing Sale Price.

Discussion: Possible further improvements

- Can explore outliers to see the characteristics of houses are undervalued
- Can explore interaction variables (e.g. Basement Quality x Basement Size, and Subclass x Neighbourhood)
- Monthly analysis can be done to find the best month to sell a house for the real estate company. Month should be treated as a categorical variable.



The End (Ant)

I was in my room and I saw a group of 10 ants just running frantically. I felt bad, so I made a small house out of a cardboard box.



This technically makes me their landlord and they are my... TenANTS.



- Adapted from:
<https://darrylspeaks.com/bad-dad-jokes-for-real-estate-agents/>