Project 2: Ames Housing Data and Kaggle Challenge

Group 3
Linus, Kelvyn, Samuel, Fionna

Problem Statement

Develop a regression model to predict final selling price of homes based on housing features in the Ames Housing dataset.

Scenario - Real estate agents and investors want to know key characteristics that affect housing prices

Overview

The data set we are using to build our model on contains housing features from the city of Ames:

- Train data: sample size of 2051, 41 categorical and 39 numerical features and the target variable of 'SalePrice'.

- Test data: sample size of 879, 41 categorical and 39 numerical features, excluding the target variable of 'SalePrice'.

Overview

Data cleaning & EDA → Data Transformation → Feature Selection → Model Selection & Finetuning → Recommendations → Limitations

Data cleaning

1. For missing object:

   - Fill ‘no facility’ :

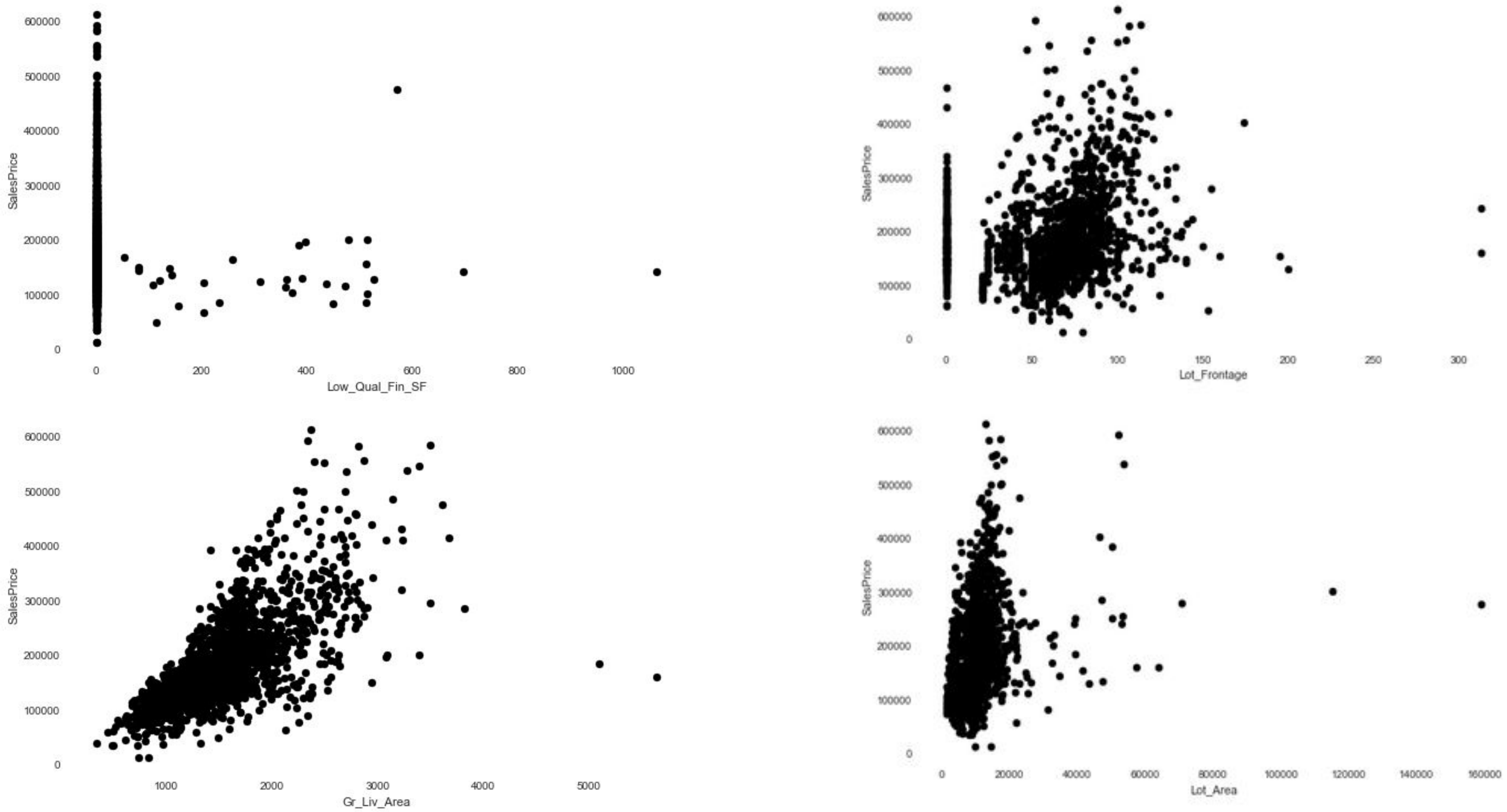
   df['FireplaceQu'] = df['FireplaceQu'].replace({np.nan: 'No fireplace'})

2. For missing values:

   - Fill ‘0’ (for big missing quantities) :

   df[zero_col] = df[zero_col].fillna(0)

   - Fill mean (for one missing value)

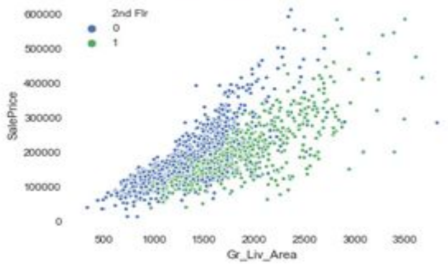


SalesPrice
Low_Qual_Fin_SF
SalesPrice
Lot_Frontage
SalesPrice
Gr_Liv_Area
SalesPrice
Lot_Area

Exploratory Data Analysis & Data Visualization

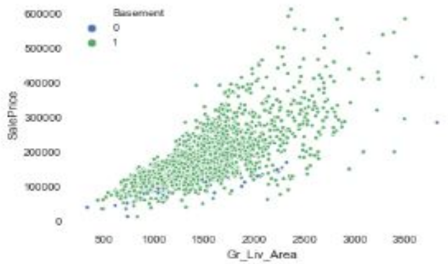
In [33]:
```
#From the plot below, having 2nd Flr features does not meant higher sale price

sns.scatterplot('Gr_Liv_Area','SalePrice',hue='2nd Flr', s=15, data=ames_train);
```



In [34]:
```
#From the plot below, there's a huge amount of house with basement

sns.scatterplot('Gr_Liv_Area','SalePrice',hue='Basement', s=15, data=ames_train);
```

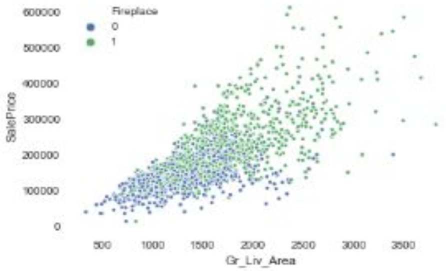


- Scatterplots to depict if additional features will affect the sale price of the house.

Exploratory Data Analysis & Data Visualization

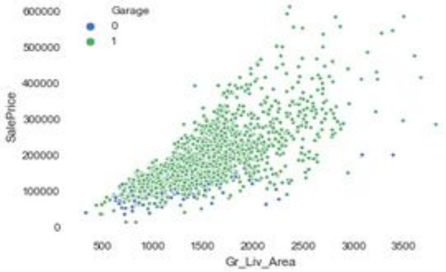
```
In [35]: #From the plot below, house with fireplaces fetch a higher sale price

         sns.scatterplot('Gr_Liv_Area','SalePrice',hue='Fireplace', s=15, data=ames_train);
```




```
In [36]: #From the plot below, most of the houses have garage

         sns.scatterplot('Gr_Liv_Area','SalePrice',hue='Garage', s=15, data=ames_train);
```

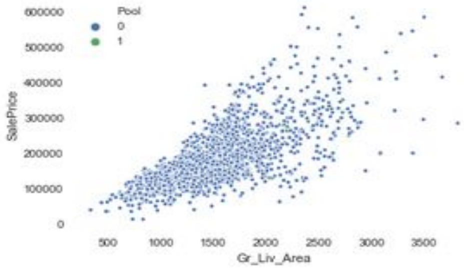


- Scatterplots to depict if additional features will affect the sale price of the house.

Exploratory Data Analysis & Data Visualization

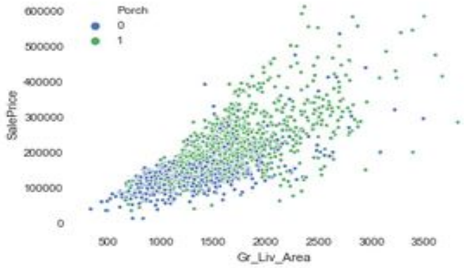
In [37]:
```
#From the plot below, there's a small amount of house with pool

sns.scatterplot('Gr_Liv_Area','SalePrice',hue='Pool', s=15, data=ames_train);
```



In [38]:
```
#Most of the house with porch fetch a higher sale price

sns.scatterplot('Gr_Liv_Area','SalePrice',hue='Porch', s=15, data=ames_train);
```

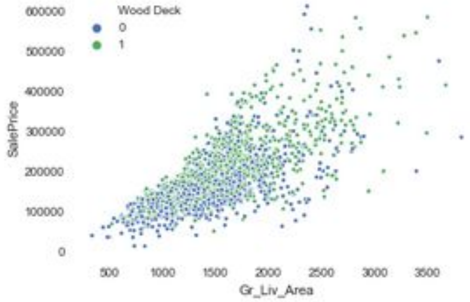


- Scatterplots to depict if additional features will affect the sale price of the house.

Exploratory Data Analysis & Data Visualization

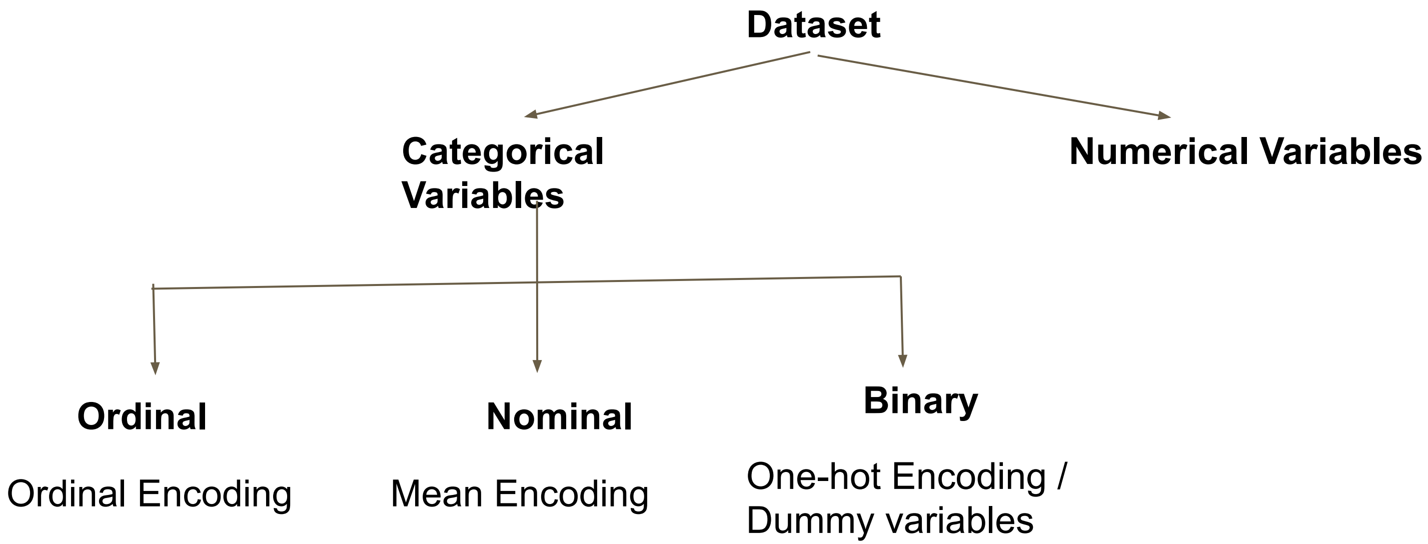
```
In [39]: #Does not describe much as scatterplot are distributed

         sns.scatterplot('Gr_Liv_Area','SalePrice',hue='Wood Deck', s=15, data=ames_train);
```



- Scatterplots to depict if additional features will affect the sale price of the house.

Data Transformation



Data Transformation (Ordinal vars)

	lot_shape	utilities	exter_qual	exter_cond	bsmt_qual	bs
0	IR1	AllPub	Gd	TA	TA	
1	IR1	AllPub	Gd	TA	Gd	
2	Reg	AllPub	TA	Gd	TA	
3	Reg	AllPub	TA	TA	Gd	
4	IR1	AllPub	TA	TA	Fa	

Individual data dictionaries

Ordinal Scales (scale 1 being the worst) h

Order rank 3

- garagefinish 1 unfinished 3 finished

Order rank 4 (custom dict)

- lotshape 1 irregular - 4 regular
- utilities 1 elo 4 allpub
- poolqc 1 fair 4 excellent
- fence 1 min wood/wire 4 good privacy

Order rank 5 (standard dict)

- exterqual 1 poor 5 excellent
- extercond 1 poor 5 excellent
- bsmtqual 1poor 5 excellent
- bsmtcond 1poor 5 excellent
- heatingqc 1 poor 5 excellent
- kitchenqual 1 poor 5 excellent
- fireplacequ 1 poor 5 excellent
- garagequal 1 poor 5 excellent
- garagecond 1 poor 5 excellent

Mapped over to ordinal vars

lot_shape_ord	utilities_ord	fe
3	4	
3	4	
4	4	
4	4	
3	4	

Data Transformation (Nominal vars)

```
MS Zoning (Nominal): Identifies the general zoning classification of the sale.

       A        Agriculture
       C        Commercial
       FV       Floating Village Residential
       I        Industrial
       RH       Residential High Density
       RL       Residential Low Density
       RP       Residential Low Density Park
       RM       Residential Medium Density
```

	ms_zoning	street	alley	land_contour
0	RL	Pave	NaN	Lvl
1	RL	Pave	NaN	Lvl
2	RL	Pave	NaN	Lvl
3	RL	Pave	NaN	Lvl
4	RL	Pave	NaN	Lvl

	ms_zoning_mean_enc	street_mean_enc	alley_mean_enc	land_contour_mean_enc	lo
0	191235.164581	181793.565558	NaN	178998.56484	
1	191235.164581	181793.565558	NaN	178998.56484	
2	191235.164581	181793.565558	NaN	178998.56484	
3	191235.164581	181793.565558	NaN	178998.56484	
4	191235.164581	181793.565558	NaN	178998.56484	

Data Transformation (Binary vars)

	central_air	paved_drive	misc_feature
0	Y	Y	NaN
1	Y	Y	NaN
2	Y	Y	NaN
3	Y	Y	NaN
4	Y	N	NaN

	central_air_dum	paved_drive_dum	misc_feature_dum
0	1	1	0
1	1	1	0
2	1	1	0
3	1	1	0
4	1	0	0

Data Transformation

1) Train-Test-Split

```
X train shape, Y train shape : (1640, 19) (1640,)
X test shape, Y test shape : (411, 19) (411,)
```

Train: 80%

Test: 20%

2) Scaling of all variables

Feature Selection I

```
ames_nomord_reg['target'].sort_values()
```

```
lot_shape_ord              -0.294542
bsmtfin_type_2_ord         -0.021038
utilities_ord               0.026404
exter_cond_ord              0.036418
land_slope_mean_enc         0.063163
street_mean_enc             0.069841
roof_matl_mean_enc          0.110623
heating_mean_enc            0.111181
functional_ord              0.125682
garage_cond_ord             0.152981
condition_2_mean_enc        0.161266
lot_config_mean_enc         0.164137
bsmt_cond_ord               0.176309
bldg_type_mean_enc          0.201220
garage_qual_ord             0.209884
fence_ord                   0.217405
condition_1_mean_enc        0.222024
land_contour_mean_enc       0.233183
electrical_mean_enc         0.257219
roof_style_mean_enc         0.268874
house_style_mean_enc        0.274636
fireplace_qu_ord            0.321086
bsmtfin_type_1_ord          0.324551
ms_zoning_mean_enc          0.332927
sale_type_mean_enc          0.377781
exterior_2nd_mean_enc       0.420871
bsmt_exposure_mean_enc      0.421535
pool_qc_ord                 0.422219
exterior_1st_mean_enc       0.438345
heating_qc_ord              0.458354
mas_vnr_type_mean_enc       0.460091
garage_type_mean_enc        0.467798
garage_finish_ord           0.525776
foundation_mean_enc         0.537040
alley_mean_enc              0.549612
bsmt_qual_ord               0.678307
kitchen_qual_ord            0.692336
exter_qual_ord              0.712146
neighborhood_mean_enc       0.760650
target                      1.000000
```

shortlist

Consolidation of all numerical and categorical variables

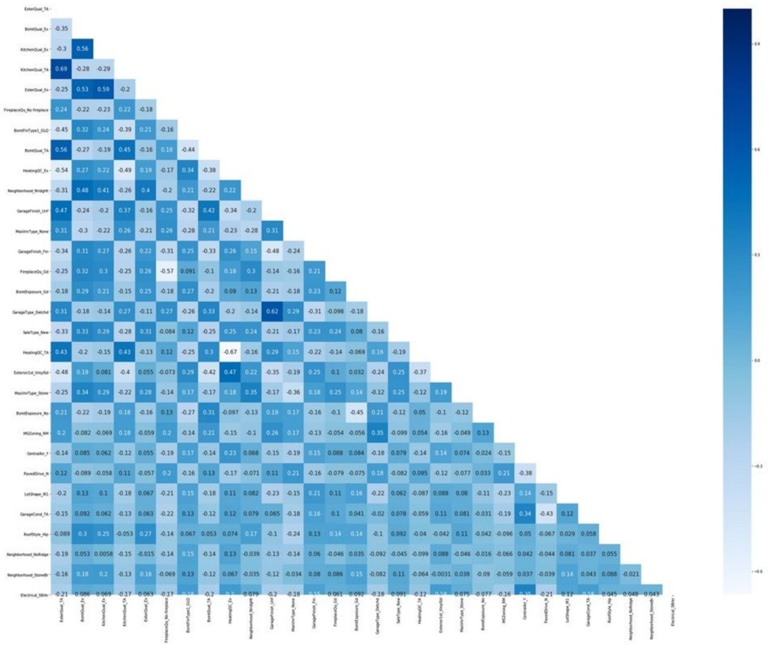
	overall_qual	gr_liv_area	garage_area	garage_cars	total_bsmt_sf	1st_flr_sf	year_built	year_remod/add
0	6	1479	475.0	2.0	725.0	725	1976	2005
1	7	2122	559.0	2.0	913.0	913	1996	1997
2	5	1057	246.0	1.0	1057.0	1057	1953	2007
3	5	1444	400.0	2.0	384.0	744	2006	2007
4	6	1445	484.0	2.0	676.0	831	1900	1993

```
overall_qual
gr_liv_area
garage_area
garage_cars
total_bsmt_sf
1st_flr_sf
year_built
year_remod/add
neighbourhood_mean_enc
exter_qual_ord
kitchen_qual_ord
bsmt_qual_ord
alley_mean_enc
foundation_mean_enc
garage_finish_ord
lot_shape_ord
central_air_dum
paved_drive_dum
misc_feature_dum
```

80 -> 19 variables

Feature selection II

1. Category variables(dummied):
   - Select variables with high corr(Xi,y)
   - Drop variables with high corr(Xi,Xj)
   - Check P value to drop variables further
2. Numerical variables(scaled):
   - Use feature selections RFE
   - Drop some variables with high p values
3. Combine the remained features as fina
   feature list for model



Model Selection

K-Fold Cross Validation

```
lr_scores = cross_val_score(lr, X_train, y_train, cv=10)
print(lr_scores)
print(lr_scores.mean())
```

```
[0.86485124 0.83015069 0.79762124 0.81684913 0.81939452 0.82564035
 0.84166456 0.71798323 0.49109148 0.82608055]
0.7831326986160876
```

Baseline Model Evaluation

```
train_score = lr.score(X_train_ss, y_train)
print("Baseline Model train_score:", train_score)
```

```
Baseline Model train_score: 0.8020482376099151
```

```
test_score = lr.score(X_test_ss, y_test)
print("Baseline Model test_score:", test_score)
```

```
Baseline Model test_score: 0.8555231576919783
```

OLS Regression Results

Dep. Variable:	saleprice	**R-squared:**	0.813
Model:	OLS	**Adj. R-squared:**	0.811
Method:	Least Squares	**F-statistic:**	489.5
Date:	Thu, 21 Nov 2019	**Prob (F-statistic):**	0.00

Model Finetuning

1) Hyper-parameter tuning: Grid search with Ridge and Lasso regressions

2) (Manual) recursive feature elimination based on T-test p-values

foundation_mean_enc	-0.0135	0.029	-0.470	0.638
garage_finish_ord	2408.8745	1181.089	2.040	0.042
lot_shape_ord	-3223.2406	1433.652	-2.248	0.025
central_air_dum	-1312.4906	3417.409	-0.384	0.701
paved_drive_dum	197.1003	3112.480	0.063	0.950
misc_feature_dum	-5302.3653	4373.543	-1.212	0.226

Final Model Selection

Criteria: Best R^2 test-score

Option 1: Best fit Ridge regression test scores (0.853)

Option 2: Best fit Lasso regression score (0.855)

Option 3: Lin reg baseline model score (0.855)

Winner: **<u>Option 3</u>**

Final Model

Name	Type	Description
overall_qual	*int64*	oridnal scale from 1-10, 1 being the poorest. rates the overall material of the house
gr_liv_area	*int64*	Above grade(ground) living area square feet
garage_area	*float64*	ordinal scale of 5. Size of garage in square feet
1st_flr_sf	*int64*	First floor square feet
exter_qual_ord	*int64*	ordinal scale of 4. Evaluates present condition of the material on the exterior
kitchen_qual_ord	*int64*	ordinal scale of 4. Kitchen quality
bsmt_qual_ord	*float64*	ordinal scale of 5. Evaluates the height of the basement
garage_finish_ord	*float64*	ordinal scale of 3. Interior finish of the garage
ClearCr	*uint8*	dummy variable of neighborhoods. Clear creek
CollgCr	*uint8*	dummy variable of neighborhoods. College creek
Crawfor	*uint8*	dummy variable of neighborhoods. Crawford
Mitchel	*uint8*	dummy variable of neighborhoods. Mitchell
NAmes	*uint8*	dummy variable of neighborhoods. Northwest Ames
NoRidge	*uint8*	dummy variable of neighborhoods. Northridge
NridgHt	*uint8*	dummy variable of neighborhoods. Northridge Heights
Sawyer	*uint8*	dummy variable of neighborhoods. Sawyer
Somerst	*uint8*	dummy variable of neighborhoods. Somerset
StoneBr	*uint8*	dummy variable of neighborhoods. Stonebrook
Timber	*uint8*	dummy variable of neighborhoods. Timberland
Veenker	*uint8*	dummy variable of neighborhoods. Veenker

8 variables

12 dummies for Neighborhood

Recommendations

Common features selected amongst our group members included:

- Overall material and finish quality
- Size of garage in square feet
- Above grade (ground) living area square feet
- Height of the basement
- Kitchen quality

Limitations

Limitations of model and areas for future improvement :

- Linear regression may not be the suitable model especially with so many features.
- Capture ordinal variable data on a numerical ordinal scale instead of word descriptors.
- The significance of features like ‘fireplaces’ and ‘pool_area’ on housing prices are state/climate dependant, valued more in colder cities like Ames[1]. And may not be suitable to train a generalised model.

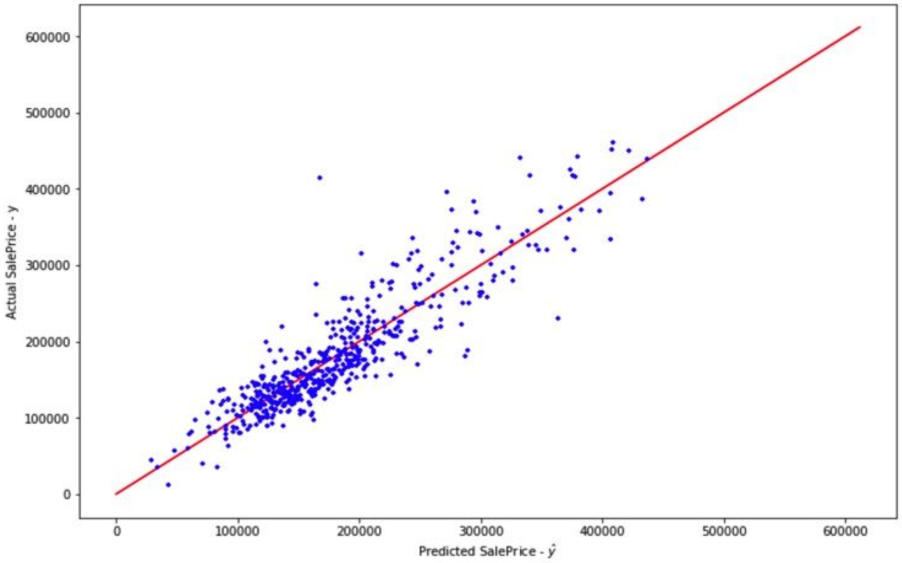
[1] Weather Spark https://weatherspark.com/y/10339/Average-Weather-in-Ames-Iowa-United-States-Year-Round

Thank you for your patience!

Model Fitting and Evaluation

Use linear regression model

- Data split after scaling
- Cross validation
- Linear regression model fit
- Model score and plot



Remove later**

Is the problem statement clearly presented?

Does a strong narrative run through the presentation building toward a fina

Are the conclusions/recommendations clearly stated?

Is the level of technicality appropriate for the intended audience?

Is the student substantially over or under time?

Does the student appropriately pace their presentation?

Does the student deliver their message with clarity and volume?

Are appropriate visualizations generated for the intended audience?

Are visualizations necessary and useful for supporting conclusions/explaining findings?

```
The optimal Lasso alpha value is: 288.3318675523569

The Lasso model has a score of: 0.900256487278 7225

               col_names          coef
2            gr_liv_area  24760.713194
85      roof_matl_CompShg  22378.731419
87      roof_matl_Tar&Grv  15979.873272
171       kitchen_qual_TA  12778.909284
0            overall_qual  12384.137330
..                   ...           ...
111  exterior_2nd_MetalSd      0.000000
105  exterior_2nd_Brk Cmn      0.000000
103  exterior_1st_WdShing      0.000000
101  exterior_1st_VinylSd      0.000000
220         sale_type_WD      0.000000
```