**Glossary:**

Neighborhood,
Physical locations within Ames city limits:

| | |
|---|---|
| Blmngtn | Bloomington Heights |
| Blueste | Bluestem |
| BrDale | Briardale |
| BrkSide | Brookside |
| ClearCr | Clear Creek |
| CollgCr | College Creek |
| Crawfor | Crawford |
| Edwards | Edwards |
| Gilbert | Gilbert |
| IDOTRR | Iowa DOT and Rail Road |
| MeadowV | Meadow Village |
| Mitchel | Mitchell |
| Names | North Ames |
| NoRidge | Northridge |
| NPkVill | Northpark Villa |
| NridgHt | Northridge Heights |
| NWAmes | Northwest Ames |
| OldTown | Old Town |
| SWISU | South & West of Iowa State University |
| Sawyer | Sawyer |
| SawyerW | Sawyer West |
| Somerst | Somerset |
| StoneBr | Stone Brook |
| Timber | Timberland |
| Veenker | Veenker |

- totalBath: sum of total bathrooms in the house.
- BedroomAbvGr: Bedrooms above grade (does NOT include basement bedrooms)
- KitchenAbvGr: Kitchens above grade.
- Fireplaces: Number of fireplaces.
- GarageCars: Size of garage in car capacity
- totalSpcBlt_standardized = Total space built: houses['totalSpcBlt'] = houses['2ndFlrSF'] + houses['1stFlrSF'] + houses['LowQualFinSF'] + houses['TotalBsmtSF']
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- TotalBsmtSF: Total square feet of basement area
- LotArea: Lot size in square feet
- GrLivArea: Above grade (ground) living area square feet
- BsmtFinType1: Rating of basement finished area
    - GLQ Good Living Quarters
    - ALQ Average Living Quarters
    - BLQ Below Average Living Quarters
    - Rec Average Rec Room
    - LwQ Low Quality
    - Unf Unfinshed
    - NA No Basement
- SalePrice: Sale price of the hose.
- OverallQual: Quality rating for the house.

**Intermediate Results:**

From the EDA "played" around with many features and built different models. Down are the main conclusions:
- The younger is the house the highest the price.
- Total rooms above grade also increase the price, as the minimum prices go up.
- Full bathroom, from 2 to 3 there is a significant increase in the price, whereas higher numbers funnily has a negative correlation.
- First floor squared feet, the more room built the higher the price.
- The largest the basement the highest the price, makes sense, the more sqfeet built the more expensive.
- Garage doesn't seem to be affected significantly to the price, I concluded there is a set of amenities that together have a meaningful impact on the price but by themself doesn't seem to affect too much.
- Overall quality, has a high direct impact too, although at the beginning of the process was considered as a feature in the end focused the model on tangible features.
- While checking multicollinearity faced some bug, I had to split the data and finally found the issue. Then rerun the code on the whole data set getting the final results.

Also considered the implementation of other features such as Quality for both, the overall rating for the house as well experimented targeting the quality of certain features like the related to Kitchen:

```
houses['KitExQual'] = 'Exteranl Quality' + 'Kitchen Quality':
```

**Some Implementation Details:**

1. I apply Standardization for totalSpcBlt, LotArea: This technique will rescale the feature to have a mean of 0 and a standard deviation of 1. Given the bell-shaped form of the distribution, standardization would work well, especially for linear regression, because it can handle moderate skewness and the algorithm benefits from features that are scaled and centred. Standardization would not drastically alter the shape of the distribution, just reposition it, which is often desirable for linear models.
2. Normalization: Given the left-skewed distribution and use case, standardization still seems to be the more appropriate preprocessing step, especially for linear regression. It will adjust for the scale of the data without distorting the skewness, which could be informative for the model.
3. Checked for Multicollinearity before saving the final dataset. Based on the results I deleted columns with a VIF higher than 7 - 8 like "Exterior1st" (this process of deleting the columns are also made above). On the other side, even some neigbourhood have higher multicollinearity than 7 I kept them to maintain the integrity of the whole group.
4. As a mentioned in the poster, Lasso regression was used for feature selection. Used this method for simplicity.

**Extra resources used:**

The final code for my machine learning, although I made it my own, I watched/checked out the following:

Linear Regression;

- https://www.youtube.com/watch?v=D2vZmz-JsLw
- https://www.youtube.com/watch?v=V0ktXKaqnTs
- https://www.youtube.com/watch?v=luyNcTVICF8
- https://www.youtube.com/watch?v=oig4Z34vM5s
- https://www.youtube.com/watch?v=VQsPCtU7Uik
- https://www.youtube.com/watch?v=BYGBr_o_KZ4
- https://www.youtube.com/watch?v=AX_ZDX6aTT0
- https://www.youtube.com/watch?v=nCFvKOMUWV4
- https://www.youtube.com/watch?v=Wf2N2GIc2ls
- https://www.youtube.com/watch?v=fjGO3mrjskc
- Linear Regression in MATLAB (youtube.com)
- AI Experts 27s (youtube.com)
- Linear Regression in Matlab (youtube.com)
- https://www.youtube.com/watch?v=ySBF4lD4LZc
- https://www.youtube.com/watch?v=V_C6luIhvjg&t=152s

Lasso Regression:

- https://www.youtube.com/watch?v=1_mLYfHX3Ro
- https://blog.devgenius.io/understanding-ridge-regression-in-machine-learning-with-matlab-code-8f9ba3ae6e48

Decision Tree Regression:

- https://www.youtube.com/watch?v=Ga-qUawiY2U
- https://www.youtube.com/watch?v=cyRpVwVP_CU
- https://www.youtube.com/watch?v=Ga-qUawiY2U&t=9s

Others:

- https://stackoverflow.com/questions/1960430/decision-tree-in-matlab
- Documentation - MATLAB & Simulink - MathWorks United Kingdom
- https://uk.mathworks.com/help/stats/understanding-linear-regression-outputs.html
- https://uk.mathworks.com/help/stats/improving-classification-trees-and-regression-trees.html
- https://uk.mathworks.com/help/stats/lasso.html