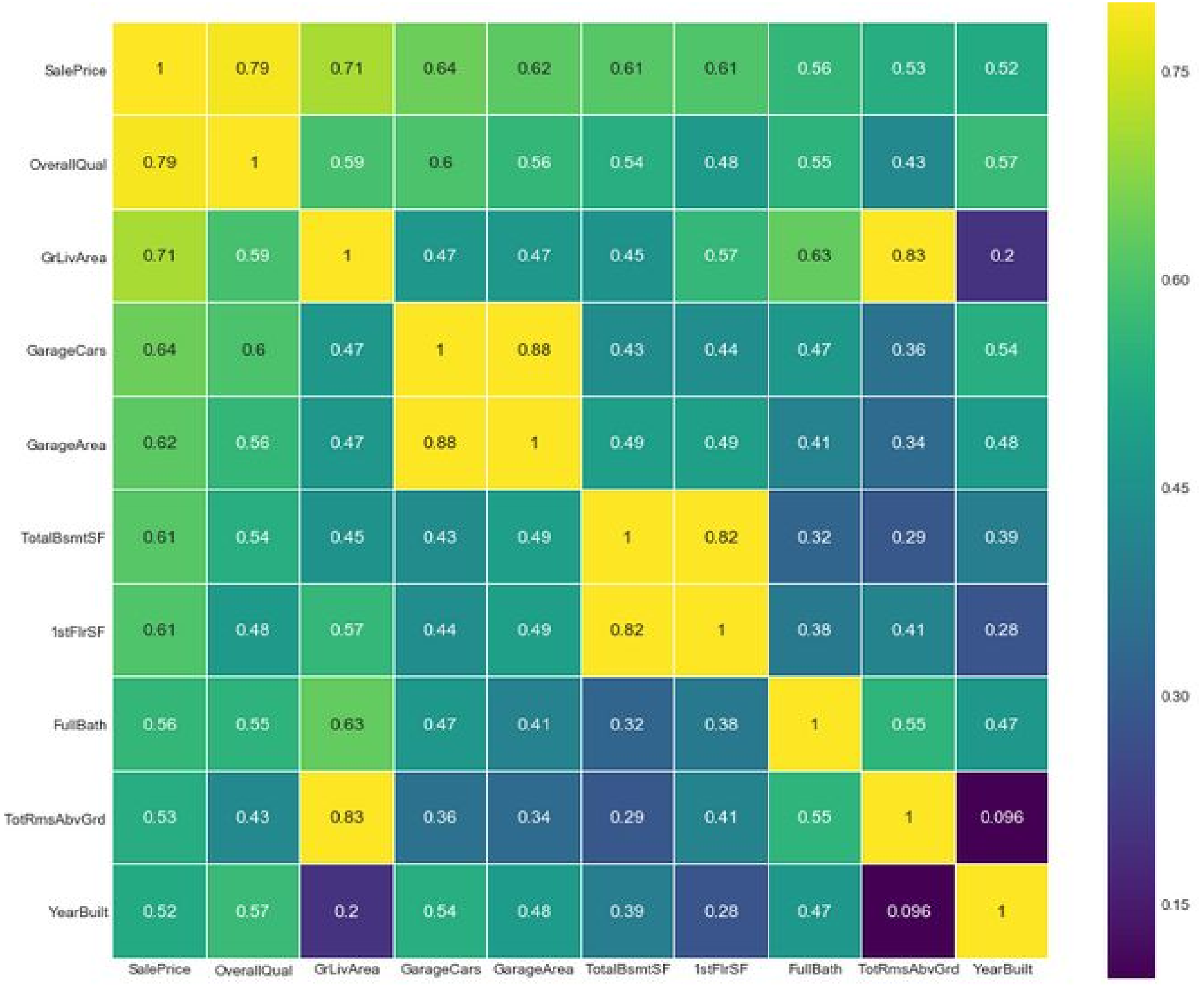
I personally think that how this eda and data processing is related to the models that are used to run. RMB that these models are Ridge, Lasso, Kernel Ridge, Elastic net, Gradient Boosting, XGBOost, and LightGBM
These are the questions i have for the script

Dropping out numerical features from looking at correlation

- I am not sure whether or not to drop the features apart from the top 10 correlated with SalesPrice
- So anything that has a correlation less than 0.52 (YearBuilt) can be neglected when modelling



	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
SalePrice	1	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1	0.59	0.6	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1	0.47	0.47	0.45	0.57	0.63	0.83	0.2
GarageCars	0.64	0.6	0.47	1	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1	0.49	0.49	0.41	0.34	0.48
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49	1	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1	0.096
YearBuilt	0.52	0.57	0.2	0.54	0.48	0.39	0.28	0.47	0.096	1

Applying log transformation to the skewed SalesPrice distribution

- Some of the kernels use log transformation, but some uses log1p which is log(1+x) but when I tried both of them it came out identical. Why is that?

Some of the categoricals are not applied with Label Encoder, why?

- I think its because you only apply the function to only ones that are ranks (not fully categorical)
- For example buildingType is not label encoded but extracted as a dummy variable with pd.get_dummies. On the other hand, bsmtCond is label encoded since it has to do with some rank

Filling out missing data

- I noticed that i used the median to fill out LotFrontage but its the median from all_data (train + test). Is it necessary to fill out median separately?

On the BOX COx transformation

- I still dont understand how to choose lambda, but I think its 0.15 is the best one interms of model prediction
- And is 0.75 skewness considered to be bad enuf for transformation?

Dummy Variable trap

- I see there the categoricals are prone to dummy variable trap and multi collinearity presence
- Do we have to deal with it? Or does the model get rid of them automatically