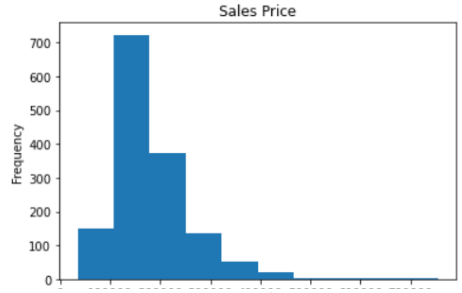
Part 1

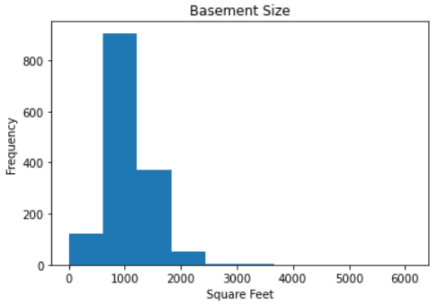
1.a) There are 81 features of which 43 are categorical and 38 are numerical. However there are only 79 explanatory variables (excluding ID and Sales Price) of which 43 are categorical and 36 numerical. There are 1460 instances in the house price data.

b)

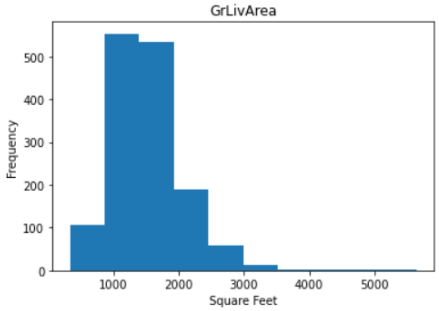
1.Overall Quality: 0.79
2.GrLiveArea: 0.71
3.GarageCars: 0.64
4.GarageArea: 0.62
5.Total BsmtSF: 0.61



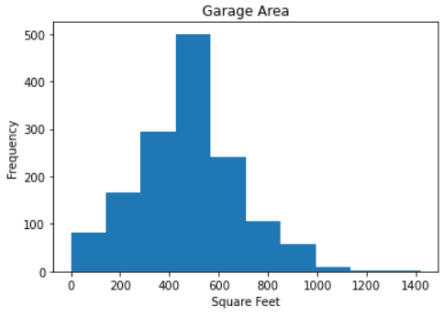
c) With the target variable being Sales Price there is a skewness of 1.88 which means there is a large skewness to the right. The value of kurtosis is 6.51 means the data has a very high peak and heavy edges.



With the Total Square Feet of Basement feature there is a skewness of 1.52 which means there is a moderately large skewness to the right again for the data. The value of kurtosis is 13.20 which means the data has a high peak also with heavy edges likely due to the outlier at 6110 square feet.

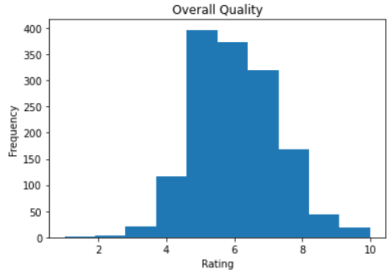


With the GrLivArea (above ground living area square feet) feature there is a skewness of 1.37 which means there is a moderately large skewness to the right again for the data. The value of kurtosis is 4.87 means the data has a very high peak and heavy edges.



With the Garage Area feature there is a skewness of 0.18 this means the data is very symmetric. The value of kurtosis is 0.91 which means the data is very close to being normally distributed but will have a slightly more prominent peak and slightly heavier tail.

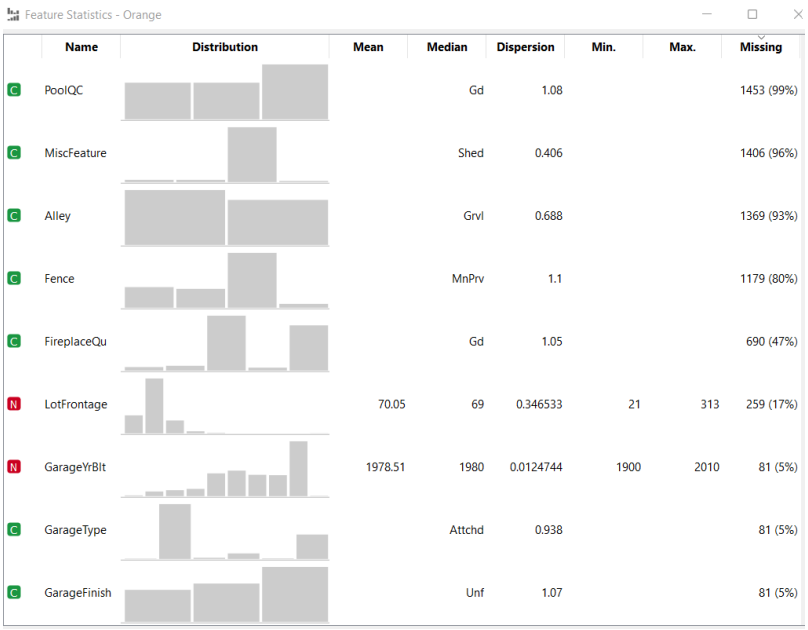
With the number of Garage Cars features there is a skewness of -0.34 which means the data is very symmetric with a slight skew to the left. The value of kurtosis is 0.22 which means the data is essentially normally distributed.



The Overall Quality feature has a skewness of 0.22 which means the data is very symmetric with a slight skewness to

the right. The value of kurtosis is 0.09 which means the data is normally distributed.

d) There are 19 features with at least 1 missing value according to Orange(found using feature statistics and looking at the rightmost column). However this is incorrect as some values have incorrectly been considered to be missing an example of this is Pool Quality is considered to have the largest amount of missing values with 99% missing however this is because Orange counted all values which are “NA” to be missing when these are actually valid figures categorising the pool quality as the person doesn’t have a pool. The same goes for Misc Features, Alley and other features where “NA” is a category seeing as not everyone has a pool or basement etc. Features which do contain missing values(found by analysing the data raw) include LotFrontage with 259(17%) missing values seeing as people who don't have area connected to the driveway have put NA instead of 0 square feet. Garage Year Built also contains 81(5%) missing values most likely caused by owners not knowing the year their garage was built or owners not having a garage and not having an option to select which makes sense considering Orange has 81 missing values or Garage Quality and Type however this is people not owning a garage and correctly inputting “NA” as no garage. There were 8(0%) missing values for Masonry Veneer Area and Type features as some people input NA as the area and type instead of 0 and None. Finally there was 1(0%) missing value for the Electrical feature. Overall there were 5 features with missing values.



	Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
C	PoolQC			Gd	1.08			1453 (99%)
C	MiscFeature			Shed	0.406			1406 (96%)
C	Alley			Grvl	0.688			1369 (93%)
C	Fence			MnPrv	1.1			1179 (80%)
C	FireplaceQu			Gd	1.05			690 (47%)
N	LotFrontage		70.05	69	0.346533	21	313	259 (17%)
N	GarageYrBlt		1978.51	1980	0.0124744	1900	2010	81 (5%)
C	GarageType			Attchd	0.938			81 (5%)
C	GarageFinish			Unf	1.07			81 (5%)

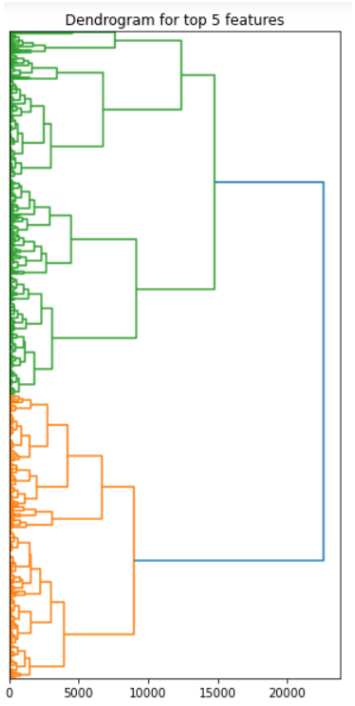
2.a.

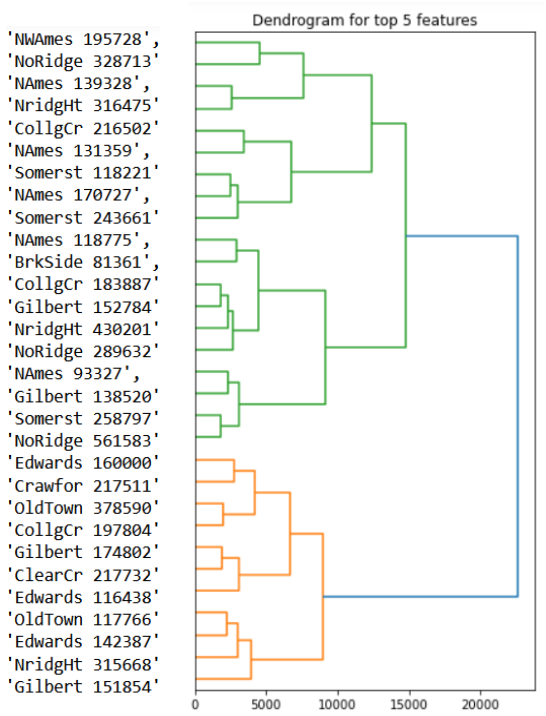
- Find the features with the highest correlation to the target variable, sales price.
- Model these key variables to predict the sales price based on these key variables to see how it affects sales price.

b. Dimensionality Reduction and Regression.
Dimensionality reduction will show which features are irrelevant/redundant and which are the useful features to predict house price.

Regression will show the trend with sales price based on the most useful features. This will be useful for analysis to see the relations in features that affect sales prices and how they affect the sales prices.

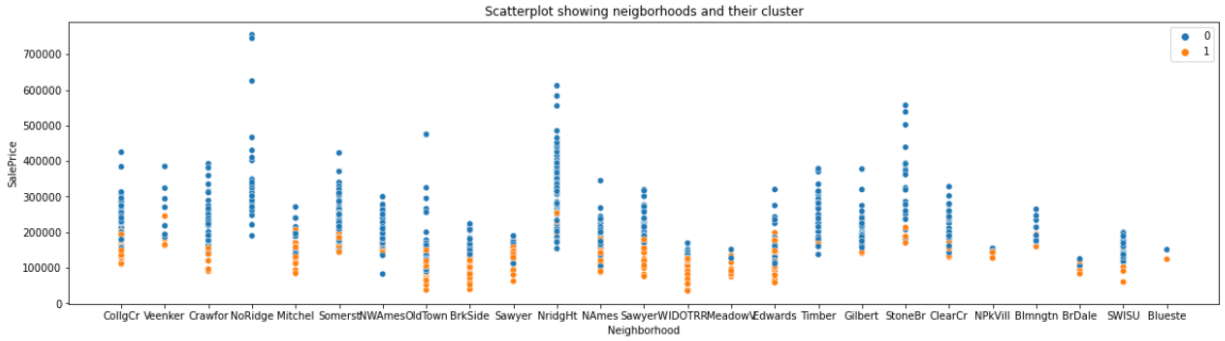
3.To help answer the question “does house prices vary by neighbourhood?” I started the analysis with a Dendrogram plotting the top 5 numerical features with the highest correlation to house price hence the relation to house price. The first dendrogram shows the output of the with all 1460 instances plotted on the y axis. This dendrogram shows there are 2 clear clusters: the top one and bottom one joined together by the blue line. This means that each cluster is closely linked by a walking distance calculated based on this rough “sales price” (top 5 features). To more clearly visualise what the dendrogram shows I truncated the model as shown in the second dendrogram. This makes the model easier to read as instead of 1460 instances there are 30 on the y axis. I then gathered the mean sales price and mode neighbourhood of each of these leaves to do further analysis as the labels on y axis using Agglomerative clustering as they use the same process of hierarchical clustering. From what I can see the mode neighbourhoods don’t repeat too often meaning that each leaf (truncated cluster) does highly correlate to a neighbourhood. This is useful information as it shows that each neighbourhood can be clustered together by the sales price as these clusterings are set based on the top 5 features correlated to sales price.



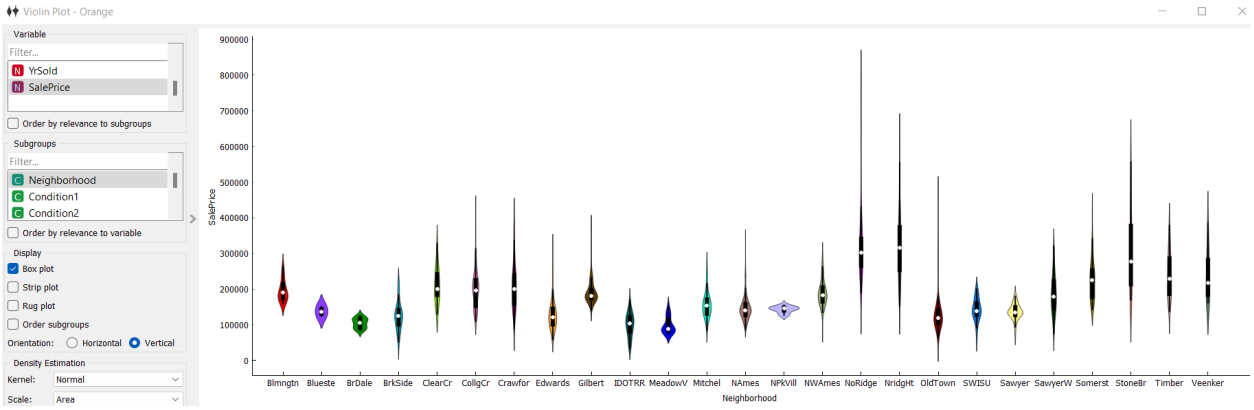


Next I took these 2 clear clusters and plotted them using a scatter plot. On the x axis are all the neighbourhoods and the y axis is the sales price of each. The 2 colours (blue and orange) represent the 2 clusters each of the data has been grouped into. From the analysis I can see that the majority of the lower sales prices are the orange cluster 1 and the higher sales price come from the blue cluster 0. This means that the dendrogram has clustered the data based on high sales price and low sales price. This further shows the correlation with the top 5 features and their relation to sales price. This scatterplot also shows that house prices do vary between neighbourhoods differently. Some neighbourhoods such as NoRidge, NridgHt and Timber are predominantly in the blue cluster 0 and have high sales prices on average meaning it is probably a richer neighbourhood and houses in neighbourhoods such as Sawyer and IDOTRR

are predominantly in the orange cluster 1 have lower house prices on average so it's probably a poorer neighbourhood. Other neighbourhoods tend to have a mix between clusters and have no clear indication of whether their house price depends on the neighbourhood based on hierarchical clustering so it's fair to say they have similar house prices and they dont vary based on neighbourhood.



Then to further see the relation between sales price and neighbourhood I plotted it on a violin plot as seen below. It clearly shows that house price does depend on neighbourhood seeing as there is variance between the neighbourhoods in means a ranges etc. The fact all the sales prices aren't identical between neighbourhoods shows that the neighbourhood the house is in depends on the sales price.



Overall it is clear to see that house prices do vary between neighbourhoods in some neighbourhoods but the majority of the neighbourhoods have very similar house prices seeing as they have mixes of clusters so it is unfair to conclude that house prices do vary between neighbourhoods 100% as only some neighbourhoods do.

Part 2

1.To preprocess the data I first started by splitting the data into training and test splits with a 70/30 split. This was done to avoid data leakage. After that I imputed the valid missing values found from Part A with their mean value for the numerical features and mode for categorical. This is because these were the real missing values and the other missing values were fake because NA was a valid response to questions such as Basement Quality where some people don't own a basement. After this I divided the data into the categorical and numerical types. This

was done to fill in the valid missing values with “None” for the categorical and “0” for the numerical. Then I encoded the categorical data using Label Encoder, this assigned a numerical value to the categorical data preparing it for scaling/normalising. To normalise the data I used MinMaxScaler where it sets the range of all values between 0 and 1. This keeps a consistent scale which will be beneficial for modelling. Finally I dropped duplicate rows to tidy up the data and keep it unique. I repeated the same process for the test set however I used the fitted models from the training set to transform the test set.

2. The dimensionality reduction techniques I used were SelectKBest and Recursive Feature Elimination(RFE). Both techniques are feature selection techniques which are suitable for finding irrelevant features/finding the best features.

RFE is a feature selection technique based on feature importance where each feature is assigned a score based on relative importance and is ranked recursively. These scores are assigned based on a supervised learning estimator which I used Random Forest Regressor which is a good estimator which uses variance reduction. Random Forest Regressor makes a collection of decision trees which it averages to provide a score. RFE splits the features in half by relevance. This resulted in the 40 features below for the training set being selected as the most relevant and the other 40 being considered irrelevant and being rejected for the next regression modelling.

```
'Id', 'LotFrontage', 'LotArea', 'OverallQual', 'OverallCond',
'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1', 'BsmtUnfSF',
'TotalBsmtSF', '1stFlrSF', '2ndFlrSF', 'GrLivArea', 'FullBath',
'BedroomAbvGr', 'TotRmsAbvGrd', 'Fireplaces', 'GarageYrBlt',
'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF', 'PoolArea',
'MoSold', 'LotShape', 'LandContour', 'Neighborhood', 'HouseStyle',
'Exterior1st', 'Exterior2nd', 'BsmtQual', 'BsmtExposure',
'BsmtFinType1', 'KitchenQual', 'FireplaceQu', 'GarageType',
'GarageFinish', 'PoolQC', 'SaleCondition'],
```

As seen with the features being selected above all the top 5 highest correlation features from part 1 are present in the top 40 features which is a good indication to the technique working. There was a slight difference in the testing set with the selected features, namely “Pool Area” and “Pool Quality” aren’t selected and instead “Land Contour” and “Year Sold” were selected. These slight differences make sense seeing as these features could have been ranked very closely between being rejected or selected and the differences caused by the test set using the training fitted model can cause some variance.

SelectKBest is another feature selection model similar to RFE where it selects the key features based on the highest k score. Instead of an estimator this model uses a score function which is a supervised method once again and I used regression since it was appropriate for the data type. The regression method used for SelectKBest was f_regression which is a univariate linear regression which returns the scores. SelectKBest then picks the top specified amount of features as the most relevant ones. To find the optimal number of features, k, I tested a variety

of values and fitted them to a regression model (next question) to find the value which receives the lowest mean squared error. The lower the mean squared error the more accurate the model and hence the more relevant and useful the features. I found that k=30 received the lowest mean squared error and any below meant that I was missing out useful features and any higher meant there are too many irrelevant features. The top 30 features using this model resulted in the features below being selected.

```
'LotFrontage', 'LotArea', 'OverallQual', 'YearBuilt', 'YearRemodAdd',
'MasVnrArea', 'BsmtFinSF1', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF',
'GrLivArea', 'FullBath', 'HalfBath', 'TotRmsAbvGrd', 'Fireplaces',
'GarageYrBlt', 'GarageCars', 'GarageArea', 'WoodDeckSF', 'OpenPorchSF'
'LotShape', 'ExterQual', 'Foundation', 'BsmtQual', 'BsmtExposure',
'HeatingQC', 'CentralAir', 'KitchenQual', 'GarageType', 'GarageFinish'
```

As seen with the features being selected above once again all the top 5 highest correlation features from part 1 are present in the selected features. The difference between the selection on the training and testing sets is very small with the only difference being “Lot Shape” being replaced with “Sales Condition” in the test set.

3. To collect the results below, I first fitted each of the 3 models (Linear Regression, Ridge Regression and Random Forest Regression) 2 times for each dimensionality reduction technique, one for the training set and one for the testing. I split the data from the question above and fitted it using the new selected x training data which excludes the irrelevant features and the sales price. I then used this fitted model to predict the sales price using the test set and calculated the mean squared error to see how accurate the prediction was based on the selected features. I also calculated the root mean squared error to make the numbers easier to compare and tabulated it. The tables below represent the results from this on each model for the training and test sets for the dimensionality reduction techniques.

SKB Training Set K=10

Model	MSE	RMSE
Linear Regression	1.48315e+09	38511.6
Ridge Regression	1.47831e+09	38448.8
Random Forest Regression	1.1958e+09	34580.4

SKB Test Set K=10

Model	MSE	RMSE
Linear Regression	1.00686e+09	31731
Ridge Regression	9.4905e+08	30806.7
Random Forest Regression	1.2907e+09	35926.3

RFE Training Set

Model	MSE	RMSE
Linear Regression	2.01302e+09	44866.7
Ridge Regression	1.54959e+09	39364.8
Random Forest Regression	1.19387e+09	34552.4

RFE Test Set

Model	MSE	RMSE
Linear Regression	9.61593e+08	31009.6
Ridge Regression	9.05136e+08	30085.5
Random Forest Regression	1.35675e+09	36834.1

From the results above it is clear to see that the 2 dimensionality reduction techniques, SelectKBest and RFE, performed very similarly with the only difference being that SelectKBest performed better than RFE for Linear Regression on the training set 38511.6 RMSE compared to 44866.7 for RFE. This could be caused due to Linear Regression performing better on fitted models where there are less features. Them having similar results makes sense because both are very similar feature selection techniques where they perform based on a ranking of scores assigned hence have similar features selected. Both techniques resulted in the training set Random Forest Regression performing much better in the training set than the test set. This is because Random Forest Regression is a more powerful ensemble regression method and so the method works very well on fitted data but has the tendency to overfit the testing set. The method creates multiple models trained over the data and averaging the results of each model to find a more powerful predictive result and so it works well with familiar data but can overcomplicate and hence overfit the testing data hence why it performs worse on the testing sets. Another discovery to be made from the tables above is that Linear Regression and Ridge Regression seems to have very similar results in terms of mean squared error. This is because both models act similarly in that they are both linear and simple. This means that these models generalise very successfully hence why the methods perform better on the testing sets than training sets as they are much better at predicting off less data, 483 instances compared to 1022 for the training set and also don't tend to overfit unlike Random Forest Regression. The better model of the 2 and best for the test sets is Ridge Regression as it has the lowest mean squared error. This is because Ridge Regression is simpler and more fancy than Linear Regression as it uses regularisation by adding a penalty to the cost function using "Alpha" which can make a more accurate prediction.