Data Science 1st Assignment

Dataset: House Prices

Mahan Madani – 99222092

1. Overview

This dataset is comprised of 1460 different houses and details about their sale, providing comprehensive descriptions on each one. This includes detailed information about the house's living spaces such as the number of rooms, the size of each area and condition of each section.

2. Dataset Exploration & Analysis

The dataset stores 81 different attributes for each house that is available for sale. Some of the columns are irrelevant and therefore will be omitted from this report.

Below you can find a list of all attributes:

- **SalePrice** - the property's sale price in dollars
- **MSSubClass**: The building class
- **MSZoning**: The general zoning classification
- **LotFrontage**: Linear feet of street connected to property
- **LotArea**: Lot size in square feet
- **Street**: Type of road access
- **Alley**: Type of alley access
- **LotShape**: General shape of property
- **LandContour**: Flatness of the property
- **Utilities**: Type of utilities available
- **LotConfig**: Lot configuration
- **LandSlope**: Slope of property
- **Neighborhood**: Physical locations within Ames city limits
- **Condition1**: Proximity to main road or railroad
- **Condition2**: Proximity to main road or railroad (if a second is present)

- **BldgType**: Type of dwelling
- **HouseStyle**: Style of dwelling
- **OverallQual**: Overall material and finish quality
- **OverallCond**: Overall condition rating
- **YearBuilt**: Original construction date
- **YearRemodAdd**: Remodel date
- **RoofStyle**: Type of roof
- **RoofMatl**: Roof material
- **Exterior1st**: Exterior covering on house
- **Exterior2nd**: Exterior covering on house (if more than one material)
- **MasVnrType**: Masonry veneer type
- **MasVnrArea**: Masonry veneer area in square feet
- **ExterQual**: Exterior material quality
- **ExterCond**: Present condition of the material on the exterior
- **Foundation**: Type of foundation
- **BsmtQual**: Height of the basement
- **BsmtCond**: General condition of the basement
- **BsmtExposure**: Walkout or garden level basement walls
- **BsmtFinType1**: Quality of basement finished area
- **BsmtFinSF1**: Type 1 finished square feet
- **BsmtFinType2**: Quality of second finished area (if present)
- **BsmtFinSF2**: Type 2 finished square feet
- **BsmtUnfSF**: Unfinished square feet of basement area
- **TotalBsmtSF**: Total square feet of basement area
- **Heating**: Type of heating
- **HeatingQC**: Heating quality and condition
- **CentralAir**: Central air conditioning
- **Electrical**: Electrical system
- **1stFlrSF**: First Floor square feet
- **2ndFlrSF**: Second floor square feet
- **LowQualFinSF**: Low quality finished square feet (all floors)
- **GrLivArea**: Above grade (ground) living area square feet
- **BsmtFullBath**: Basement full bathrooms
- **BsmtHalfBath**: Basement half bathrooms
- **FullBath**: Full bathrooms above grade
- **HalfBath**: Half baths above grade
- **Bedroom**: Number of bedrooms above basement level

- **Kitchen**: Number of kitchens
- **KitchenQual**: Kitchen quality
- **TotRmsAbvGrd**: Total rooms above grade (does not include bathrooms)
- **Functional**: Home functionality rating
- **Fireplaces**: Number of fireplaces
- **FireplaceQu**: Fireplace quality
- **GarageType**: Garage location
- **GarageYrBlt**: Year garage was built
- **GarageFinish**: Interior finish of the garage
- **GarageCars**: Size of garage in car capacity
- **GarageArea**: Size of garage in square feet
- **GarageQual**: Garage quality
- **GarageCond**: Garage condition
- **PavedDrive**: Paved driveway
- **WoodDeckSF**: Wood deck area in square feet
- **OpenPorchSF**: Open porch area in square feet
- **EnclosedPorch**: Enclosed porch area in square feet
- **3SsnPorch**: Three season porch area in square feet
- **ScreenPorch**: Screen porch area in square feet
- **PoolArea**: Pool area in square feet
- **PoolQC**: Pool quality
- **Fence**: Fence quality
- **MiscFeature**: Miscellaneous feature not covered in other categories
- **MiscVal**: $Value of miscellaneous feature
- **MoSold**: Month Sold
- **YrSold**: Year Sold
- **SaleType**: Type of sale
- **SaleCondition**: Condition of sale

3. Data Preprocessing

3.1 Drop Unnecessary Features:

Due to the high number of features included in this dataset, only a selection of them was chosen for this research. The following attributes are of little use, as such they should be dropped:

'MSSubClass' - 'LotFrontage' - 'BsmtExposure' - 'BsmtFinType1' - 'BsmtFinSF1' - 'BsmtFinType2' - 'BsmtFinSF2' - 'BsmtUnfSF' - 'TotalBsmtSF' - 'EnclosedPorch' - '3SsnPorch' - 'ScreenPorch' - 'PoolArea' - 'PoolQC' - 'MiscVal' - "MiscFeature" - 'Alley' - 'MasVnrType' - 'FireplaceQu' - 'Fence' - 'KitchenAbvGr' - 'LowQualFinSF' - 'BsmtFullBath' - 'BsmtHalfBath' - 'WoodDeckSF', 'OpenPorchSF'

3.2 Handle Null Values:

Out of all the selected features, several contain null values that must be dealt with.

About 5 percent of all records are missing the attributes related to the state of the house's garage (e.g., "GarageType", GarageQual). This number is too high to manually handle or replace with randomly generated values. Ultimately, the best course of action is to not utilize these records for samples that aim to analyze the House's garage.

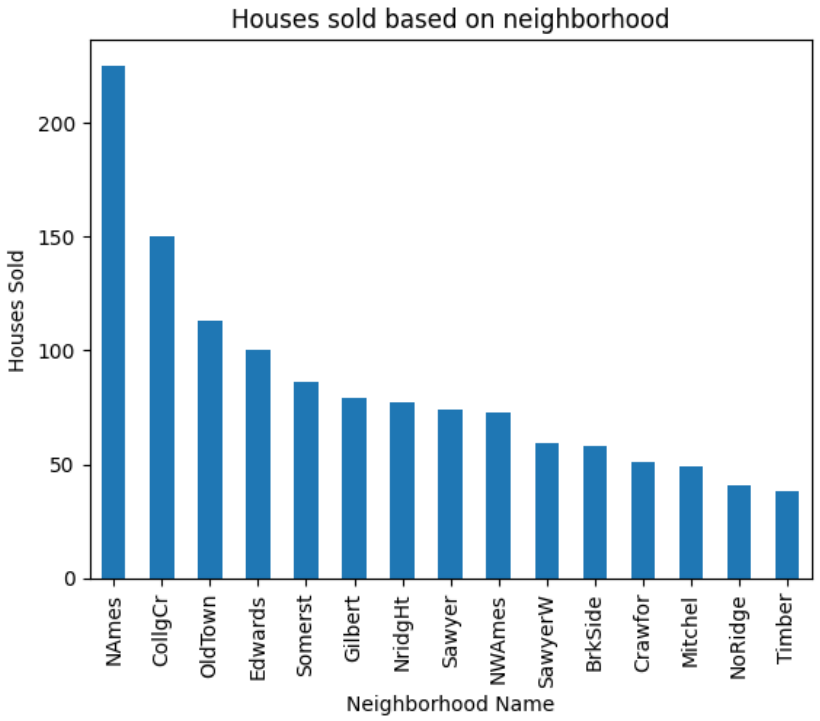
3.3 Convert Data Types:

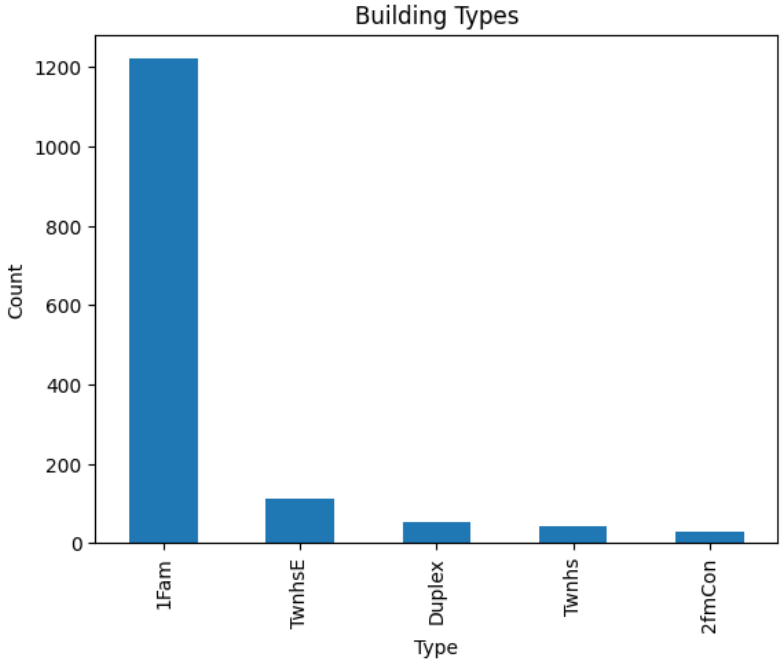
- "MoSold" is stored as an integer based on the month's numeric order. By converting it to the corresponding month title, it can be utilized as a categorical data for other purposes.

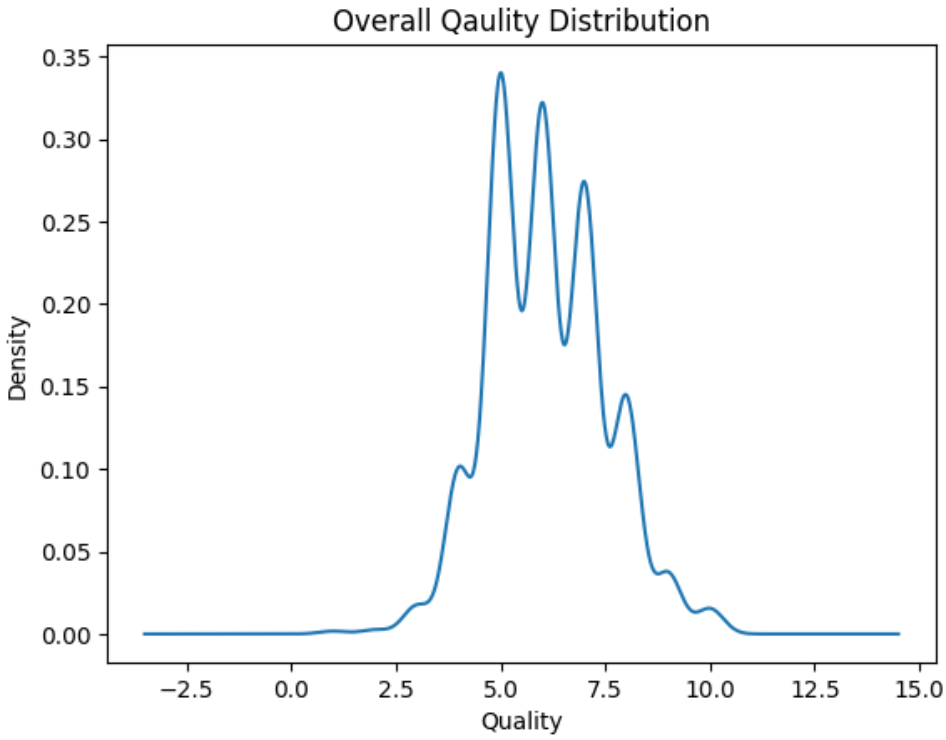
All of the other attributes in the dataset are using the correct datatype so no additional modification is required.

4. Data Visualization:

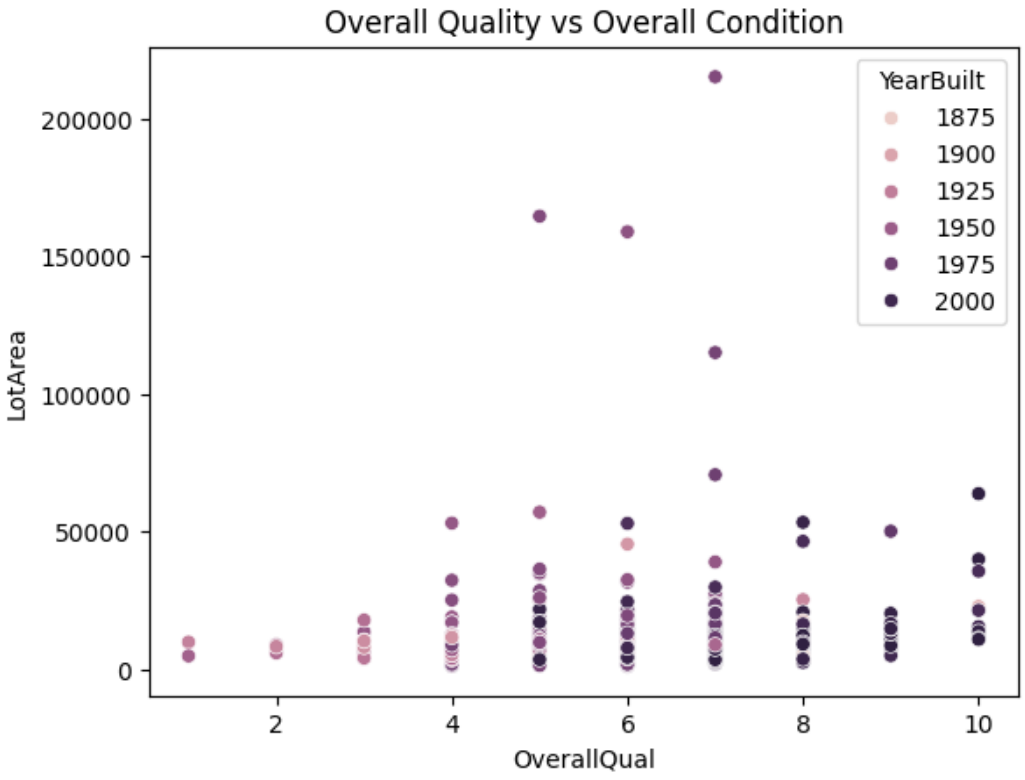
Below you can find a number of plots visualizing the data so that it can be more easily understood. These plots will be used in the next step to conduct a set of statistical tests.

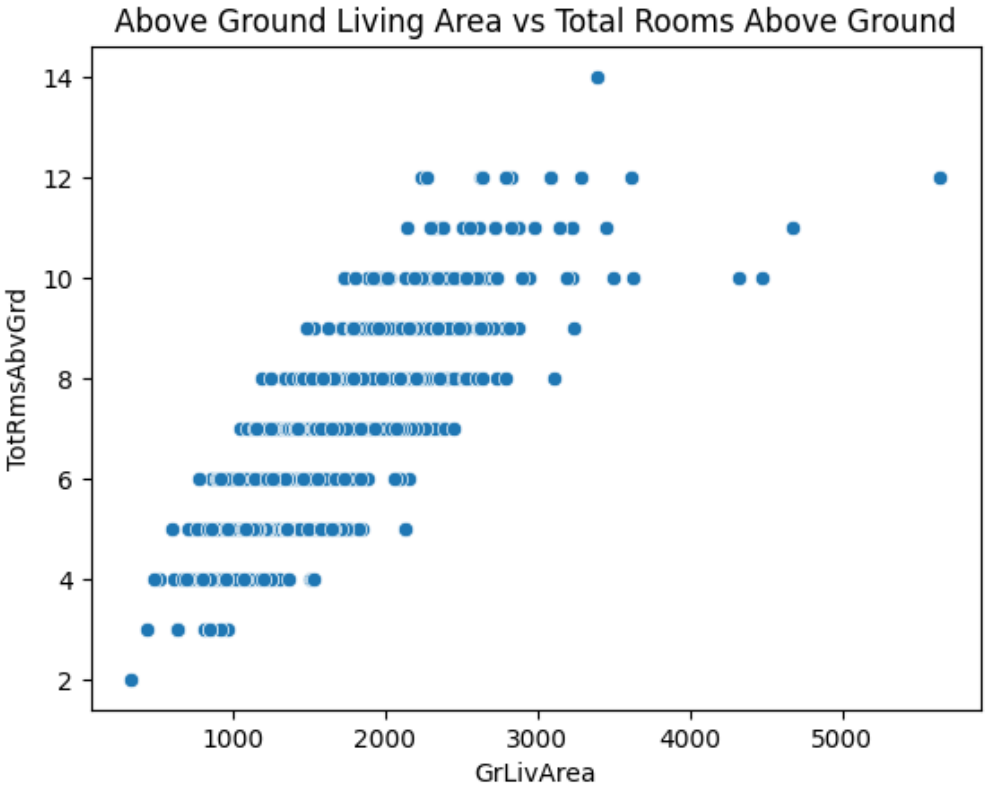




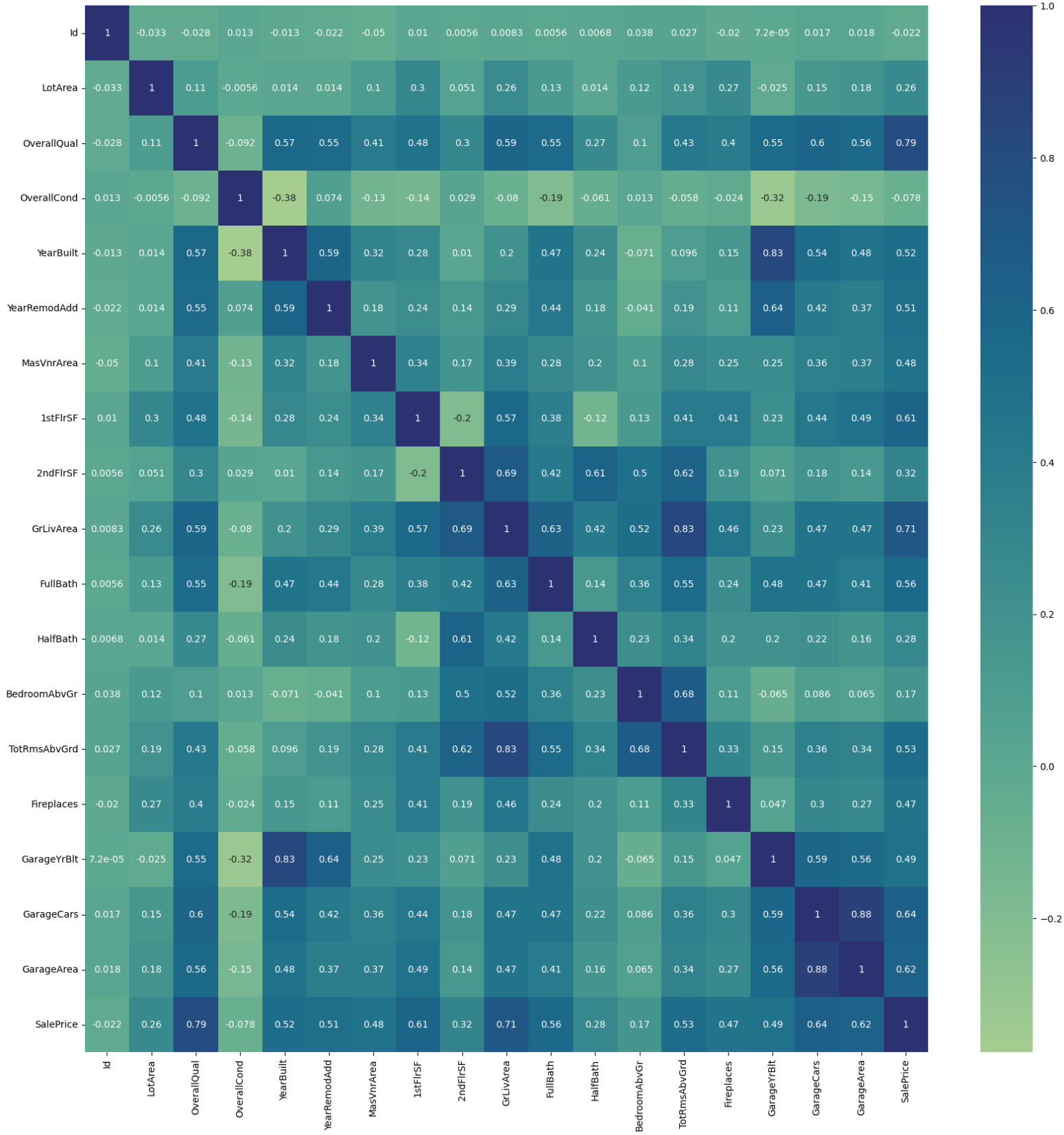

Overall Qaulity Distribution
Density
Quality
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
15.0


Sale Price vs Overall Quality
OverallQual
SalePrice
YearBuilt
1875
1900
1925
1950
1975
2000
10
8
6
4
2
0
100000
200000
300000
400000
500000
600000
700000


Overall Quality vs Overall Condition
YearBuilt
1875
1900
1925
1950
1975
2000
LotArea
200000
150000
100000
50000
0
2
4
6
8
10
OverallQual


Above Ground Living Area vs Total Rooms Above Ground
TotRmsAbvGrd
14
12
10
8
6
4
2
1000
2000
3000
4000
5000
GrLivArea

Correlation Matrix Heatmap:



5. Statistical Tests:

For hypothesis testing, a number of statements will be presented and then by utilizing a statistical test, their credibility is evaluated.

Significance Level (**α**) = 0.01

If the calculated p-value is smaller than **α**, the null hypothesis is rejected and the alternative hypothesis is accepted.

5.1 Null Hypothesis 1:

There is no relationship between Sale Price and Overall Quality.

Both Sale Price and Quality are quantitative, so an appropriate test would be the Pearson Correlation test.

Test type: **Spearman Correlation**, sample size = **50**

Statistic	P-Value	Result
0.764749	1.0155e-10 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Sale Price and Overall Quality.

5.2 Null Hypothesis 2:

There is no relationship between Above Ground Living Area and Total Rooms Above Ground.

Test type: **Spearman**, sample size = **50**

Statistic	P-Value	Result
0.862382	8.5160e-16 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Above Ground Living Area and Total Rooms Above Ground.

5.3 Null Hypothesis 3:

There is no relationship between YearBuilt and Overall Condition.

Test type: **Pearson Correlation**, sample size = **50**

Statistic	P-Value	Result
0.466630	0.0006 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between YearBuilt and Overall Condition.

5.4 Null Hypothesis 4:

There is no relationship between Sale Price and Above Ground Living Area.

Test type: **Spearman Correlation**, sample size = **50**

Statistic	P-Value	Result
0.759359	1.6360e-10 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Sale Price and Above Ground Living Area.

5.5 Null Hypothesis 5:

There is no relationship between Overall Quality and Garage Cars.

Test type: **Pearson Correlation**, sample size = **50**

Statistic	P-Value	Result
0.568592	1.6487e-05 < α	Reject Null Hypothesis

Alternative Hypothesis: There is a significant relationship between Overall Quality and Garage Cars.

6. Conclusion:

Based on the analyzed data, the visualization and the statistical tests, we can conclude that many attributes of the dataset are correlated to each other. Some dependencies are easier to see but others may need deeper analysis.