Exploratory Data Analysis for Machine Learning – Project

IBM Machine Learning Professional Certificate

By: Bernardita Štitić

Date: October 25, 2021

INDEX

TABLE OF CONTENTS

LIST OF TABLES....3

LIST OF FIGURES....3

LIST OF FIGURES IN APPENDICES....4

LIST OF APPENDICES....5

1. BRIEF DESCRIPTION OF THE DATA SET AND ITS ATTRIBUTES....6

2. INITIAL PLAN FOR DATA EXPLORATION....8

2.1. Pair plots (histograms and scatter plots): all and selected features....8

2.2. Box plot: selected features....10

2.3. Relevant statistics: selected features....11

3. DATA CLEANING AND FEATURE ENGINEERING: SELECTED FEATURES....11

3.1. Missing values....11

3.2. Skewness....12

3.3. Pair plot (after the logarithmic transformation)....14

3.4. Box plots (after the logarithmic transformation)....16

3.5. Relevant statistics (after the logarithmic transformation)....18

4. KEY FINDINGS AND INSIGHTS....19

5. HYPOTHESES ABOUT THE DATA 20

6. KOLMOGOROV-SMIRNOV TEST 21

7. NEXT STEPS 22

8. SUMMARY OF DATA QUALITY 25

REFERENCES 26

APPENDICES 27

LIST OF TABLES

Table 1.1. Column data types of a Pandas DataFrame with the Ames housing data set 6

Table 1.2. Pandas DataFrame with the Ames housing data set: variables with missing values 7

LIST OF FIGURES

Figure 2.1.1. Pair plot of the chosen feature subset and the target 9

Figure 2.2.1. Box plot of the chosen feature subset and the target.....10

Figure 2.3.1. Relevant statistics of the chosen feature subset and the target 11

Figure 3.2.1. Skewness of the chosen features and the target 12

Figure 3.2.2. Histograms of the target before and after applying the np.log1p transformation by NumPy.....13

Figure 3.2.3. Skewness of the transformed and original chosen variables 14

Figure 3.3.1. Pair plot of the chosen variable subset after applying a logarithmic transformation to 3 features and the target 15

Figure 3.4.1. Box plot of the chosen variable subset after the application of the logarithmic transformation.....16

Figure 3.4.2. Box plot of 5 of the chosen variables after the application of the logarithmic transformation.....17

Figure 3.4.3. Box plot of 4 of the chosen variables after the application of the logarithmic transformation.....18

Figure 3.5.1. Relevant statistics after applying the logarithmic transformation to 3 features and the target.....19

Figure 7.1. Scatter plot of the target ("SalePrice") and the feature "Gr Liv Area" after applying np.log1p.....23

Figure 7.2. Scatter plot of the target ("SalePrice") and the feature "Lot Area" after applying np.log1p.....23

Figure 7.3. Scatter plot of the target ("SalePrice") and the feature "Total Bsmt SF" 24

LIST OF FIGURES IN APPENDICES

Figure B-1. Pair plot of the first smaller Pandas DataFrame 29

Figure B-2. Pair plot of the second smaller Pandas DataFrame 30

Figure B-3. Pair plot of the third smaller Pandas DataFrame.....31

Figure B-4. Pair plot of the fourth smaller Pandas DataFrame.....32

Figure C-1. Histograms of the feature "Lot Area" before and after the transformation.....33

Figure C-2. Histograms of the feature "Total Bsmt SF" before and after the transformation.....33

Figure C-3. Histograms of the feature "Gr Liv Area" before and after the transformation.....34

Figure C-4. Bar plot of the feature "Garage Cars" before and after the transformation.....34

Figure C-5. Bar plot of the feature "Fireplaces" before and after the transformation.....35

LIST OF APPENDICES

Appendix A. List of column information of a Pandas DataFrame containing the Ames housing data set 27

Appendix B. Pair plots of the data set 29

Appendix C. Histograms and bar plots of the chosen features before and after applying a logarithmic transformation 33

1. BRIEF DESCRIPTION OF THE DATA SET AND ITS ATTRIBUTES

The chosen data set for this project corresponds to the Ames housing data set (De Cock, 2011), which describes the sale of residential property in Ames, Iowa, from 2006 to 2010. Table 1.1 summarizes the column data types of a Pandas DataFrame containing the Ames housing data set. Furthermore, the data set contains 2930 observations and 81 explanatory variables (features). Specifically, 23 features are nominal, 23 are ordinal, 14 are discrete and 20 are continuous (De Cock, 2011). A list with a brief description of each variable was developed by Kuhn et al. (2020), which is available in the references section of this report. Finally, the target or the 82nd variable corresponds to the sale price of the property.

Column data type	Number of instances
Float64	11
Int64	28
Object	43

Table 1.1. Column data types of a Pandas DataFrame with the Ames housing data set.

Table 1.1 shows that the proportion of categorical variables (52.44%) and numerical variables (47.56%) is roughly the same. Moreover, after exploring the data set with Pandas, it was found that 27 columns have missing values. These columns are illustrated in Table 1.2 with the respective number of missing values. Also, the complete list returned by Pandas is found in Appendix A.

Column name	Non-null values	Column name	Non-null values
Lot Frontage	2440	Bsmt Full Bath	2928
Alley	198	Bsmt Half Bath	2928
Mas Vnr Type	2907	Fireplace Qu	1508
Mas VnrArea	2907	Garage Type	2773
Bsmt Qual	2850	Garage YrBlt	2771
Bsmt Cond	2850	Garage Finish	2771
Bsmt Exposure	2847	Garage cars	2929
BsmtFin Type 1	2850	Garage Area	2929
BsmtFin SF 1	2929	Garage Qual	2771
BsmtFin Type 2	2849	Garage Cond	2771
BsmtFin SF 2	2929	Pool QC	13
BsmtUnf SF	2929	Fence	572
Total Bsmt SF	2929	Misc Feature	106
Electrical	2929	-	-

Table 1.2. Pandas DataFrame with the Ames housing data set: variables with missing values.

2. INITIAL PLAN FOR DATA EXPLORATION

A possible initial plan for data exploration may consist of studying the numerical features, as well as the target, to identify heavily skewed distributions, any multicollinearity issues and/or variable dependencies that might benefit from feature transformations. Finally, it is also relevant to study the presence of outliers and gather the most relevant statistics of the variables.

Therefore, the following steps can be taken to initially explore the data set:

1. Visualize the distributions of the numerical variables (histograms are a good option);
2. Visualize dependencies of variable pairs (scatter plots are a good option);
3. Visualize any outliers (box plots are good for quickly assessing this);
4. Gather relevant statistics of each feature and the target.

2.1. Pair plots (histograms and scatter plots): all and selected features

A good approach regarding the first and second steps consists of creating smaller Pandas DataFrames with the numerical variables to make pair plots with Seaborn. These include histograms or bar plots and scatter plots of the dependencies. In this project, the decision of splitting the numerical data into different Pandas DataFrames is for visualization purposes only. Moreover, in particular, 4 Pandas DataFrames were created with different features and the target, except for the features "Order" and "PID".

Furthermore, like in the course laboratory, as suggested by the data set author some outliers were removed. Specifically, rows whose value of the feature "Gr Liv Area" was larger than 4000 were dropped. In consequence, a data set with 2925 rows and 82 columns was obtained. After this step the pair plots, which are shown in Appendix B, were obtained. Subsequently, a subset of 5 features was chosen based on a selection made for a multiple regression model by the data set author (De Cock, 2011), namely:

- "Lot Area" (int64) or the lot area in square feet (Kuhn et al., 2020);
- "Total Bsmt SF" (float64) or square feet of the basement area (Kuhn et al., 2020);
- "Gr Liv Area" (int64) or above grade living area in square feet (Kuhn et al., 2020);
- "Garage cars" (float64) or capacity of the garage in terms of cars (Kuhn et al., 2020);
- "Fireplaces" (int64) or total number of fireplaces (Kuhn et al., 2020).

After selecting these features, a pair plot was made to study distributions and dependencies. Figure 2.1.1 illustrates the pair plot which shows the behavior of the previous subset of 5 features and the target.

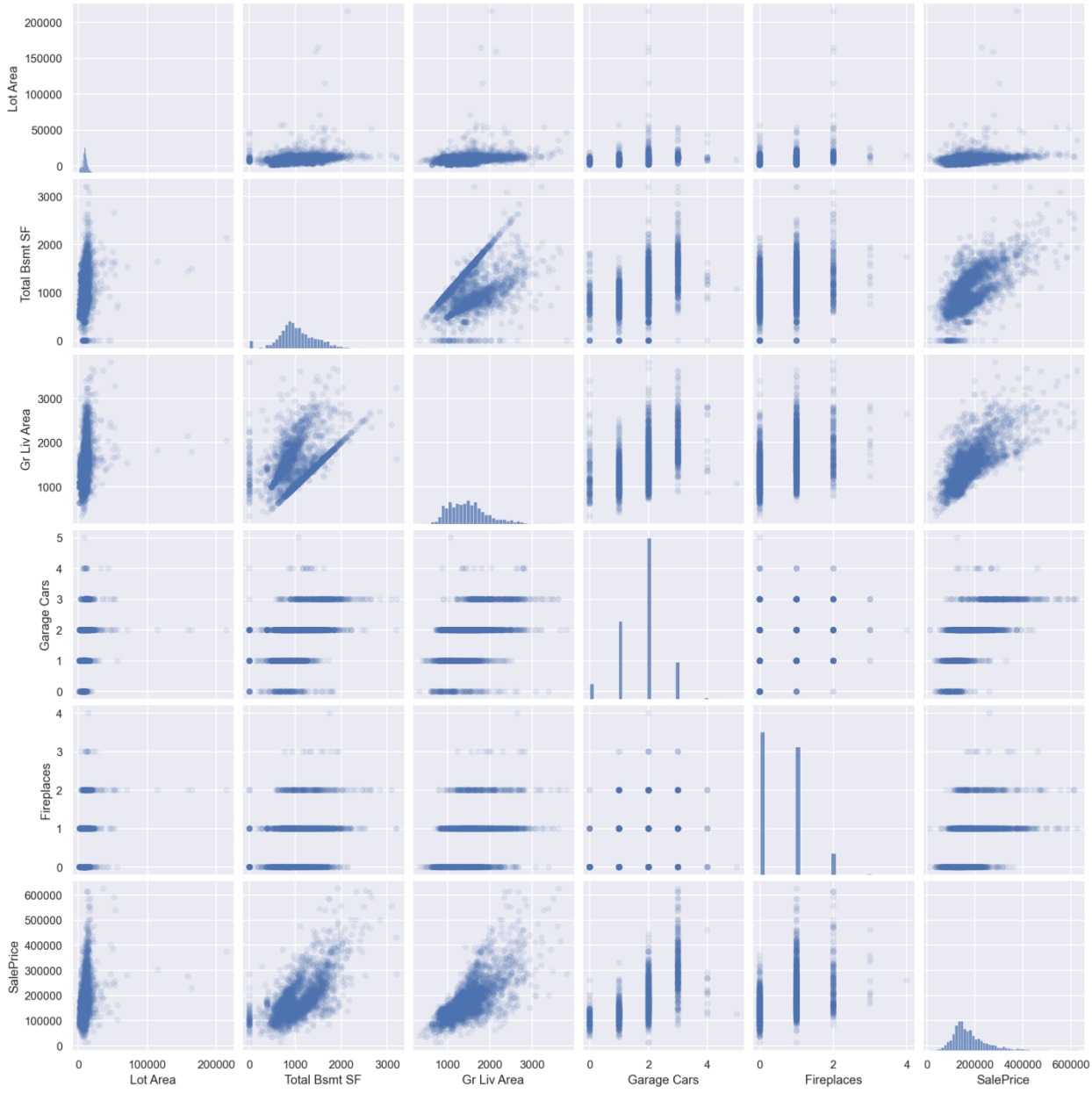

Figure 2.1.1. Pair plot of the chosen feature subset and the target.

2.2. Box plot: selected features

Similarly, to gain insight into the presence of outliers with respect to the selected features and the target, a box plot was made using the `.boxplot()` Pandas method. The plot is shown in Figure 2.2.1.

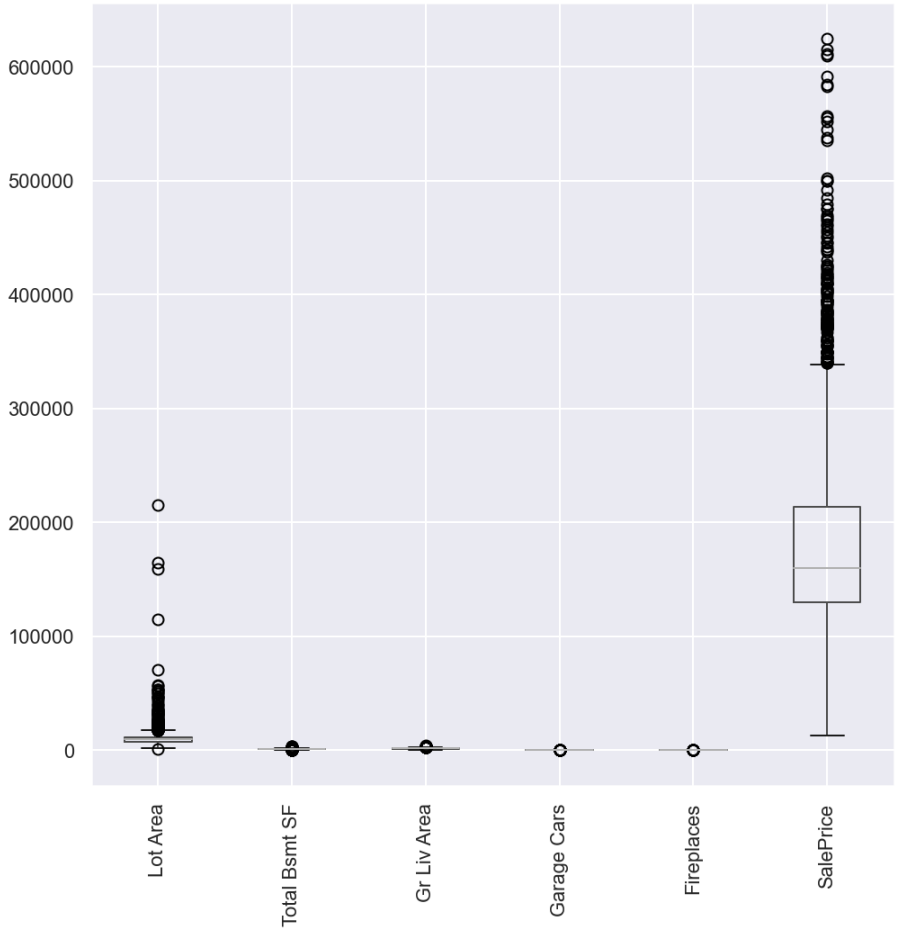

Figure 2.2.1. Box plot of the chosen feature subset and the target.

2.3. Relevant statistics: selected features

Using the Pandas .describe() method, relevant statistics of the chosen feature subset and the target were obtained. Specifically, the method calculates, for each variable, the number of non-null values, the sample mean, the sample standard deviation, the minimum value, relevant percentiles (25th, 50th and 75th) and the maximum value. Figure 2.3.1 illustrates the results.

	count	mean	std	min	25%	50%	75%	max
Lot Area	2925.0	10103.583590	7781.999124	1300.0	7438.00	9428.0	11515.00	215245.0
Total Bsmt SF	2924.0	1046.852257	421.109533	0.0	792.75	989.5	1299.25	3206.0
Gr Liv Area	2925.0	1493.978803	486.273646	334.0	1126.00	1441.0	1740.00	3820.0
Garage Cars	2924.0	1.765048	0.759834	0.0	1.00	2.0	2.00	5.0
Fireplaces	2925.0	0.596923	0.645349	0.0	0.00	1.0	1.00	4.0
SalePrice	2925.0	180411.574701	78554.857286	12789.0	129500.00	160000.0	213500.00	625000.0

Figure 2.3.1. Relevant statistics of the chosen feature subset and the target.

3. DATA CLEANING AND FEATURE ENGINEERING: SELECTED FEATURES

3.1. Missing values

Regarding the chosen subset of features, first the topic of missing values was addressed. As shown in Figure 2.3.1, only the features "Total Bsmt SF" and "Garage Cars" have missing values, specifically one each. Therefore, for each feature, the missing value was replaced with a 0, which is the minimum value for both "Total Bsmt SF" and "Garage Cars" according to Figure 2.3.1. Implicitly, by doing so it is being assumed that if the information is missing then most likely those properties do not have a basement or garage.

3.2. Skewness

After the missing values were replaced with a 0 for the chosen feature subset, the skewness of each distribution was explored. For this, like in the course laboratory, the .skew() Pandas method was employed. This method returns an unbiased skew value, which is normalized by N-1 (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.skew.html). Figure 3.2.1 shows the results.

	Skew
Lot Area	13.200004
SalePrice	1.591072
Gr Liv Area	0.878879
Fireplaces	0.732312
Total Bsmt SF	0.395191
Garage Cars	-0.221062

Figure 3.2.1. Skewness of the chosen features and the target.

Based on the histograms and bar plots in Figure 2.1.1, as well as on the results of Figure 3.2.1, a logarithmic transformation was applied to the variables. In particular, like in the course laboratory, the np.log1p transformation by NumPy was applied to each case (https://numpy.org/doc/stable/reference/generated/numpy.log1p.html). Also, similarly to the course laboratory, histograms and bar plots were created to visually appreciate how the distribution of the 5 features and the target were affected by the transformation. The case of the target ("SalePrice") is shown in Figure 3.2.2 whereas Appendix C shows the histograms and bar plots of the features.

Figure 3.2.2. Histograms of the target before and after applying the np.log1p transformation by NumPy.

Afterward, the skewness values were calculated again using the .skew() Pandas method. Regarding these new results, as well as the histograms and bar plots in Appendix C, it was observed that the np.log1p transformation was beneficial except for the features "Total Bsmt SF" and "Garage Cars". In fact, for these, the absolute values of the skewness increased from 0.395191 and 0.221062 as derived from Figure 3.2.1 to 4.956425 and 1.279444 respectively. Therefore, the transformation was applied only to the target and the remaining 3 features. The skewness results of the transformed and original variables are shown in Figure 3.2.3.

	Skew
Total Bsmt SF	0.395191
Fireplaces	0.235802
SalePrice	-0.041927
Gr Liv Area	-0.059228
Garage Cars	-0.221062
Lot Area	-0.534108

Figure 3.2.3. Skewness of the transformed and original chosen variables.

3.3. Pair plot (after the logarithmic transformation)

Figure 3.3.1 illustrates the obtained pair plot after applying the np.log1p transformation to the features "Lot Area", "Gr Liv Area", "Fireplaces" and the target "SalePrice", while leaving the features "Total Bsmt SF" and "Garage Cars" in their original form.

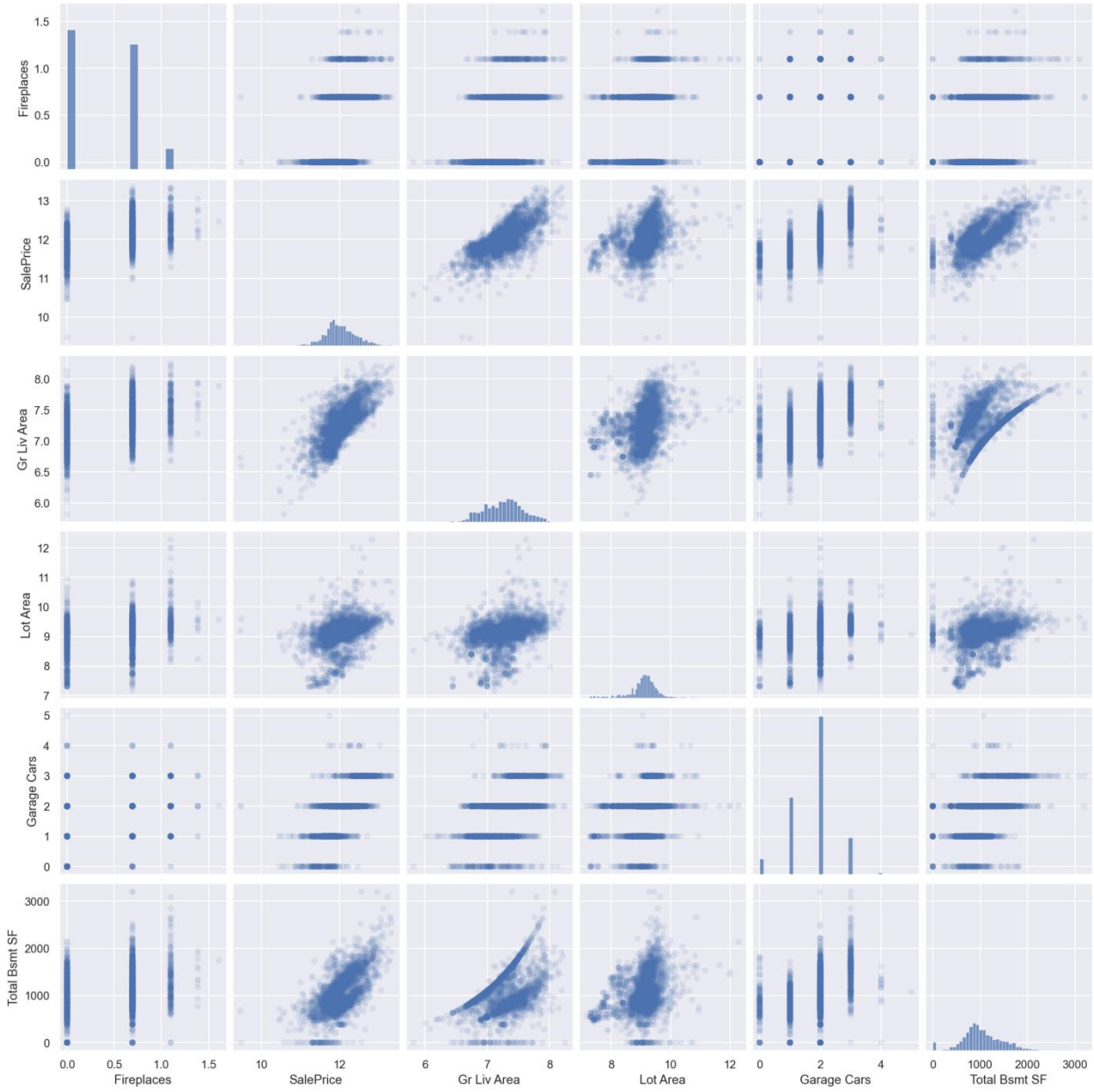

Figure 3.3.1. Pair plot of the chosen variable subset after applying a logarithmic transformation to 3 features and the target.

3.4. Box plots (after the logarithmic transformation)

Similarly, Figure 3.4.1, Figure 3.4.2 and Figure 3.4.3 show the new box plots. Due to a difference in the range of the variables, 3 box plots were produced to appreciate the new behavior.

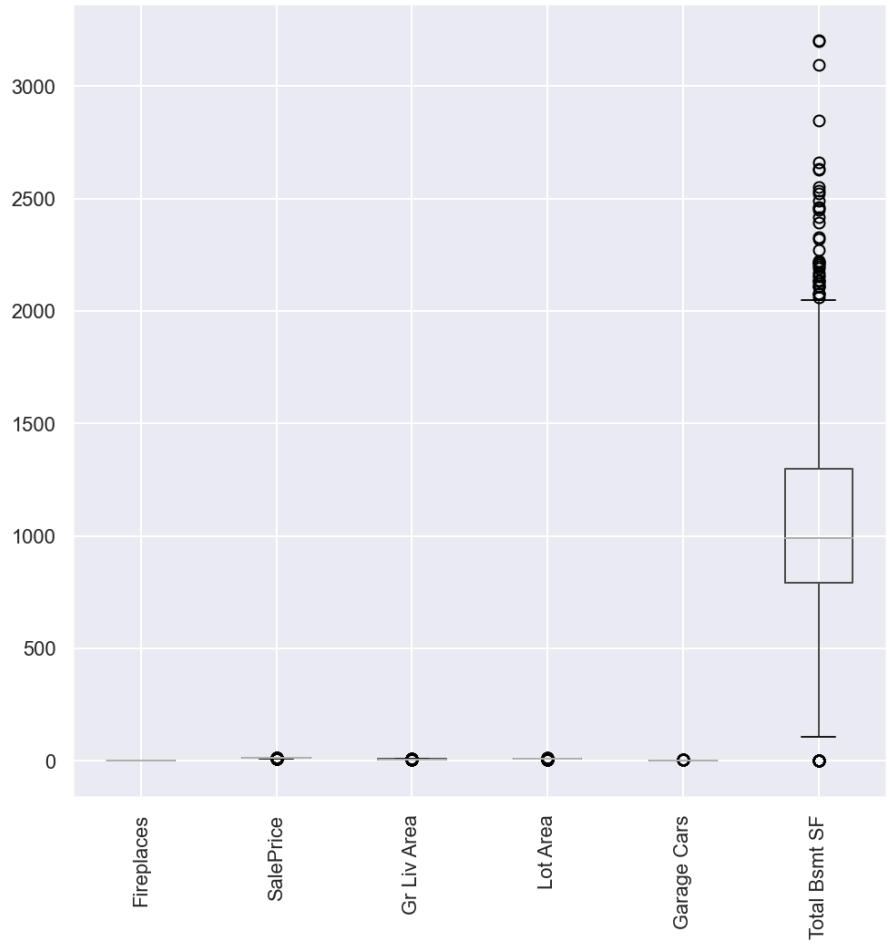

Figure 3.4.1. Box plot of the chosen variable subset after the application of the logarithmic transformation.

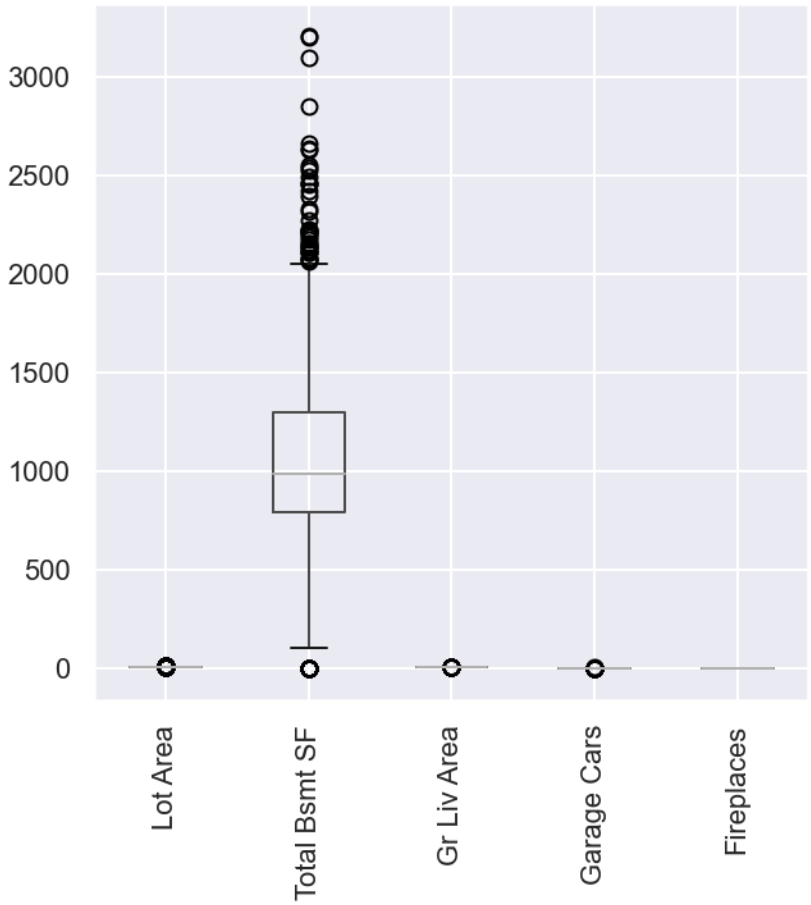

Figure 3.4.2. Box plot of 5 of the chosen variables after the application of the logarithmic transformation.

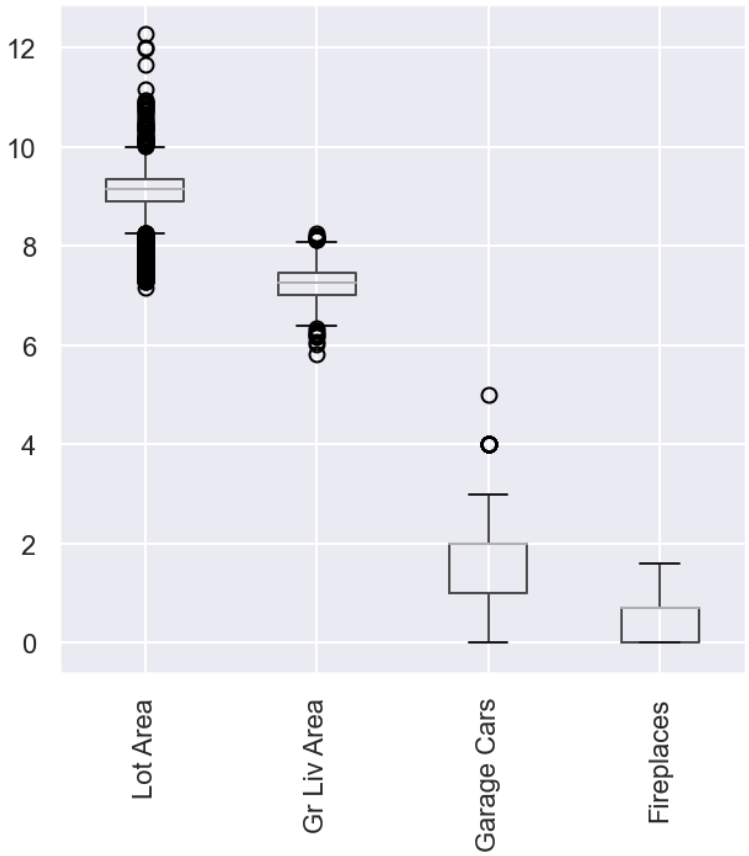

Figure 3.4.3. Box plot of 4 of the chosen variables after the application of the logarithmic transformation.

3.5. Relevant statistics (after the logarithmic transformation)

Likewise, Figure 3.5.1 shows the new statistics obtained with the .describe() Pandas method after applying the logarithmic transformation to 3 features and the target.

	Fireplaces	SalePrice	Gr Liv Area	Lot Area	Garage Cars	Total Bsmt SF
count	2925.000000	2925.000000	2925.000000	2925.000000	2925.000000	2925.000000
mean	0.389310	12.019887	7.258784	9.090148	1.764444	1046.494359
std	0.394534	0.406013	0.320753	0.508309	0.760405	421.482215
min	0.000000	9.456419	5.814131	7.170888	0.000000	0.000000
25%	0.000000	11.771444	7.027315	8.914492	1.000000	792.000000
50%	0.693147	11.982935	7.273786	9.151545	2.000000	989.000000
75%	0.693147	12.271397	7.462215	9.351493	2.000000	1299.000000
max	1.609438	13.345509	8.248267	12.279537	5.000000	3206.000000

Figure 3.5.1. Relevant statistics after applying the logarithmic transformation to 3 features and the target.

4. KEY FINDINGS AND INSIGHTS

Some of the key findings and insights about the Ames housing data set, with respect to the 5 chosen features and the target, include:

- After eliminating the outliers which the data set author suggested (De Cock, 2011), 5 outliers were eliminated.
    - This reduced the sample size to 2925 observations.
- Only 2 features (“Total Bsmt SF” and “Garage Cars”) had missing values, specifically one each.
    - These values were replaced with a 0, which is the minimum value of both features.
    - This can be interpreted as a property not having a basement or garage (this could be a reason why this information was not collected).
- For 3 features (“Lot Area”, “Gr Liv Area”, “Fireplaces”) and the target (“SalePrice”), the absolute values of the skewness decreased after applying the np.log1p transformation by NumPy.
    - Histograms and bar plots were consistent with this (sample distributions look less skewed).

    - The transformation was only applied to these variables; the remaining ones ("Total Bsmt SF" and "Garage Cars") were left unchanged.
- The pair plot obtained after the application of np.log1p shows dependencies between feature pairs that might be of concern. This should be addressed later.
    - Example: the scatter plot between "Lot Area" and "Gr Live Area" (4th row, 3rd column) shows that correlation in this case might be important.
- The pair plot also shows some dependencies that appear to be non-linear.
    - Example: the scatter plot between "SalePrice" and "Gr Live Area" (2nd row, 3rd column) exhibits polynomial behavior.
    - Example: the scatter plot between "Gr Liv Area" and "Total Bsmt SF" (3rd row, 6th column) exhibits behavior similar to a square root function.
- After applying the logarithmic transformation, the box plots show that there are still data points located past the whiskers.
    - Specifically, there are values beyond the 75th quartile + 1.5 IQR and below the 25th quartile – 1.5 IQR, where IQR is the Interquartile Range.
- Relevant statistics calculated after applying the transformation show that, considering the features, the variable "Total Bsmt SF" has a maximum order of magnitude of 10^3.
    - The other features reach a maximum order of magnitude of 10^2 or 10^1.

5. HYPOTHESES ABOUT THE DATA

Regarding hypothesis testing, and depending on what features are selected, some of the possible hypothesis tests to perform include:

- A One Sample T-Test to hypothesize about the coefficients of a multiple regression model (after studying the residuals);
- A Chi-Square Test of Independence to test the independence between a pair of categorical variables;
- A One Sample Kolmogorov-Smirnov Test to hypothesize about the distribution of a continuous variable.

For instance, it is possible to formulate the following:

- For a One Sample T-Test, we can consider the coefficient β_j of a given feature x_j in a multiple regression model.

- Null hypothesis: $\beta_j = 0$ (the feature is not significant for the analysis).
- Alternative hypothesis: $\beta_j \neq 0$ (the feature is significant for the analysis).

- For a Chi-Square Test of Independence, a pair of categorical variables can be chosen, such as "Neighborhood" and "House Style".
  - Null hypothesis: there is no relationship between the categorical variables.
  - Alternative hypothesis: there is a relationship between the categorical variables.
- For a One Sample Kolmogorov-Smirnov Test, a continuous variable such as "Total Bsmt SF" can be chosen.
  - Null hypothesis: the data fits a specific distribution (for example, a normal distribution).
  - Alternative hypothesis: the data does not fit the specific distribution (normal distribution in this case).

6. KOLMOGOROV-SMIRNOV TEST

The test that was performed corresponds to the One Sample Kolmogorov-Smirnov Test with respect to the feature "Total Bsmt SF". Specifically, the null hypothesis and the alternative hypothesis are similar to the ones that were formulated in the previous section. In particular:

- Null hypothesis: "Total Bsmt SF" fits a normal distribution with parameters $\mu = 1046.852257$ square feet and $\sigma = 421.109533$ square feet.
- Alternative hypothesis: "Total Bsmt SF" does not fit a normal distribution with parameters $\mu = 1046.852257$ square feet and $\sigma = 421.109533$ square feet.

The parameters above, which are reported in Figure 2.3.1, correspond to the sample mean and sample standard deviation (unbiased) that were calculated by Pandas with the .describe() method. Moreover, these results were calculated using the 2924 original observations before filling the missing value of the feature with a 0. Therefore, the hypothesis test was conducted using this data. Also, the values were sorted in ascending order and stored in a NumPy array.

Then, a significance level of 5% was set before conducting the test. Next, the kstest function by SciPy (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kstest.html) was called using the following arguments for the function parameters: cdf='norm' and args=(1046.852257,421.109533). The result, as reported by SciPy, was:

KstestResult(statistic=0.06386622788917229, pvalue=8.213032219133894e-11)

In other words, the test statistic was approximately 0.0639 and the p-value was lower than the significance level. Consequently, the null hypothesis was rejected. If the histogram shown in Figure C-2 in Appendix C is considered, this result is consistent with the fact that the histogram is not symmetric and that the tails of the sample distribution are not similar either. In particular, the left tail is noticeably larger than the right tail.

Moreover, since the sample size is large, it is likely that these results are correct. Also, precisely due to sample size, performing the test with the exact distribution or the asymptotic distribution of the test statistic leads to the previous results in both cases and thus to the rejection of the null hypothesis. Finally, it would be relevant to complement this analysis with knowledge from subject matter experts to help explain the results of the test further.

7. NEXT STEPS

Regarding the chosen variables, which were cleaned and transformed, some suggestions for next steps include the dependencies between the target and the features. Specifically:

- In Figure 3.3.1, the dependency between "SalePrice" and the logarithmic transformation of "Gr Liv Area" (2nd row, 3rd column) might benefit from a quadratic (polynomial) transformation. The scatter plot from the pair plot is illustrated in Figure 7.1.
- In Figure 3.3.1, the relationship between "SalePrice" and the logarithmic transformation of "Lot Area" (2nd row, 4th column) might benefit from a cubic transformation since the scatter plot shows an increase in prices, then a decrease in prices and finally another increase as the value of the independent variable grows. The scatter plot from the pair plot is highlighted in Figure 7.2.
- In Figure 3.3.1, and considering the relationship between "SalePrice" and "Total Bsmt SF" (2nd row, 4th column), due to the vertical behavior near the origin it might be relevant to consider adding a dummy variable to the feature set. The scatter plot from the pair plot is emphasized in Figure 7.3.

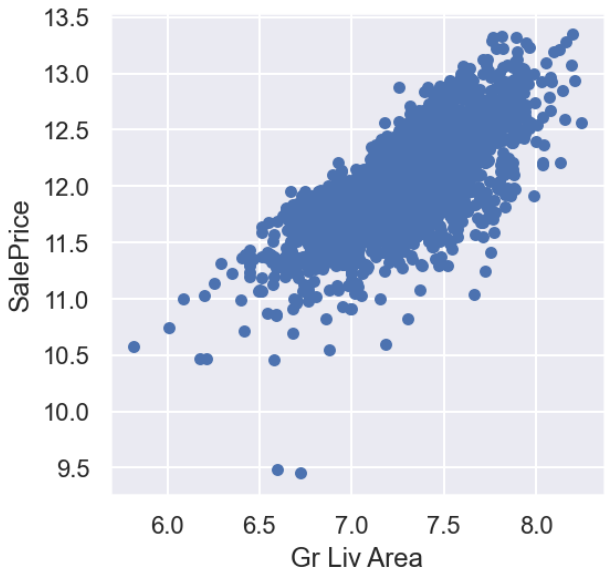

Figure 7.1. Scatter plot of the target (“SalePrice”) and the feature “Gr Liv Area” after applying np.log1p.

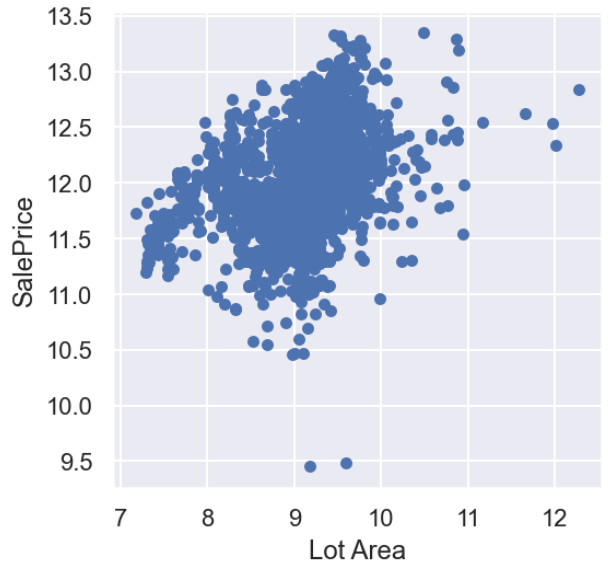

Figure 7.2. Scatter plot of the target (“SalePrice”) and the feature “Lot Area” after applying np.log1p.

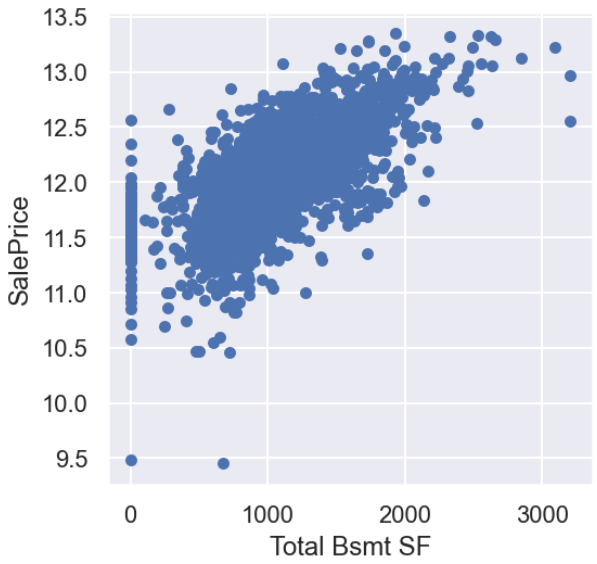

Figure 7.3. Scatter plot of the target (“SalePrice”) and the feature “Total Bsmt SF”.

Furthermore, other possible next steps with respect to the transformed feature subset discussed in this report include:

- Addressing the feature dependencies, which can be seen in Figure 3.3.1 as mentioned earlier, using an adequate approach. First, however, the course on regression, which is part of the Professional Certificate, will be completed.
- Addressing the topic of outliers further, which will require considering whether the data points observed in the box plots are true outliers. Moreover, it should then be decided whether it will be necessary to eliminate them.
- If a multiple linear regression model is created, it is advisable to study the residuals of the model and test for the statistical significance of the coefficients (for instance, with One Sample T-Student tests).
    - In this case, considering multicollinearity issues and relationships between each feature and the target will be crucial.
    - Calculating correlation coefficients is also advisable.
- Depending on what machine learning algorithm is later chosen, it might be necessary to scale the features due to the differences in orders of magnitude as seen in Figure 3.5.1.

8. SUMMARY OF DATA QUALITY

In general, the data quality of this data set is considered to be good. The data poses challenges which are common in EDA exercises but these, so far, appear manageable. Furthermore, some of these challenges are yet to be explored, such as the non-linear relationships between the target and certain features or the outliers. Therefore, any difficulties that might be discovered in the future will need to be addressed and managed properly.

One of the greatest challenges regarding the Ames housing data set is making a good feature selection so that an adequate model can be made. In fact, the amount of features is considerably high (81). Moreover, since the sample size is sufficiently large, at least considering the features that were chosen in this report hypothesis results are likely to be correct from a statistical point of view. However, it is good to complement these results with, for instance, histograms or knowledge from subject matter experts. For the moment, no more data to move forward with this project is necessary since there is enough information for more exploration.

REFERENCES

De Cock, D. (2011). Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education, 19*(3). http://jse.amstat.org/v19n3/decock.pdf

Kuhn, M., Perepolkin, D., & RStudio. (2020, June 23). Package 'AmesHousing'. CRAN. https://cran.r-project.org/web/packages/AmesHousing/AmesHousing.pdf

APPENDICES

Appendix A. List of column information of a Pandas DataFrame containing the Ames housing data set

Columns with missing values have been highlighted on the list below to quickly find those with missing values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2930 entries, 0 to 2929
Data columns (total 82 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
0   Order            2930 non-null   int64
 1   PID              2930 non-null   int64
 2   MS SubClass      2930 non-null   int64
 3   MS Zoning        2930 non-null   object
4   Lot Frontage     2440 non-null   float64
 5   Lot Area         2930 non-null   int64
 6   Street           2930 non-null   object
7   Alley            198 non-null    object
 8   Lot Shape        2930 non-null   object
 9   Land Contour     2930 non-null   object
10  Utilities        2930 non-null   object
11  Lot Config       2930 non-null   object
12  Land Slope       2930 non-null   object
13  Neighborhood     2930 non-null   object
14  Condition 1      2930 non-null   object
15  Condition 2      2930 non-null   object
16  Bldg Type        2930 non-null   object
17  House Style      2930 non-null   object
18  Overall Qual     2930 non-null   int64
19  Overall Cond     2930 non-null   int64
20  Year Built       2930 non-null   int64
21  YearRemod/Add    2930 non-null   int64
22  Roof Style       2930 non-null   object
23  RoofMatl         2930 non-null   object
24  Exterior 1st     2930 non-null   object
25  Exterior 2nd     2930 non-null   object
26  MasVnr Type      2907 non-null   object
27  MasVnr Area      2907 non-null   float64
28  Exter Qual       2930 non-null   object
29  Exter Cond       2930 non-null   object
30  Foundation       2930 non-null   object
31  Bsmt Qual        2850 non-null   object
32  Bsmt Cond        2850 non-null   object
33  Bsmt Exposure    2847 non-null   object
34  BsmtFin Type 1   2850 non-null   object
35  BsmtFin SF 1     2929 non-null   float64
```

```
36  BsmtFin Type 2   2849 non-null   object
37  BsmtFin SF 2     2929 non-null   float64
38  BsmtUnf SF       2929 non-null   float64
39  TotalBsmt SF     2929 non-null   float64
40  Heating          2930 non-null   object
41  Heating QC       2930 non-null   object
42  Central Air      2930 non-null   object
43  Electrical       2929 non-null   object
44  1st Flr SF       2930 non-null   int64
45  2nd Flr SF       2930 non-null   int64
46  Low Qual Fin SF  2930 non-null   int64
47  Gr Liv Area      2930 non-null   int64
48  Bsmt Full Bath   2928 non-null   float64
49  Bsmt Half Bath   2928 non-null   float64
50  Full Bath        2930 non-null   int64
51  Half Bath        2930 non-null   int64
52  BedroomAbvGr     2930 non-null   int64
53  KitchenAbvGr     2930 non-null   int64
54  Kitchen Qual     2930 non-null   object
55  TotRmsAbvGrd     2930 non-null   int64
56  Functional       2930 non-null   object
57  Fireplaces       2930 non-null   int64
58  Fireplace Qu     1508 non-null   object
59  Garage Type      2773 non-null   object
60  GarageYrBlt      2771 non-null   float64
61  Garage Finish    2771 non-null   object
62  Garage Cars      2929 non-null   float64
63  Garage Area      2929 non-null   float64
64  Garage Qual      2771 non-null   object
65  Garage Cond      2771 non-null   object
66  Paved Drive      2930 non-null   object
67  Wood Deck SF     2930 non-null   int64
68  Open Porch SF    2930 non-null   int64
69  Enclosed Porch   2930 non-null   int64
70  3Ssn Porch       2930 non-null   int64
71  Screen Porch     2930 non-null   int64
72  Pool Area        2930 non-null   int64
73  Pool QC          13 non-null     object
74  Fence            572 non-null    object
75  Misc Feature     106 non-null    object
76  Misc Val         2930 non-null   int64
77  Mo Sold          2930 non-null   int64
78  Yr Sold          2930 non-null   int64
79  Sale Type        2930 non-null   object
80  Sale Condition   2930 non-null   object
81  SalePrice        2930 non-null   int64
dtypes: float64(11), int64(28), object(43)
memory usage: 1.8+ MB
```

Appendix B. Pair plots of the data set

Figure B-1, Figure B-2, Figure B-3 and Figure B-4 show the pair plots obtained with Seaborn for the 4 Pandas DataFrames containing different features and the target.

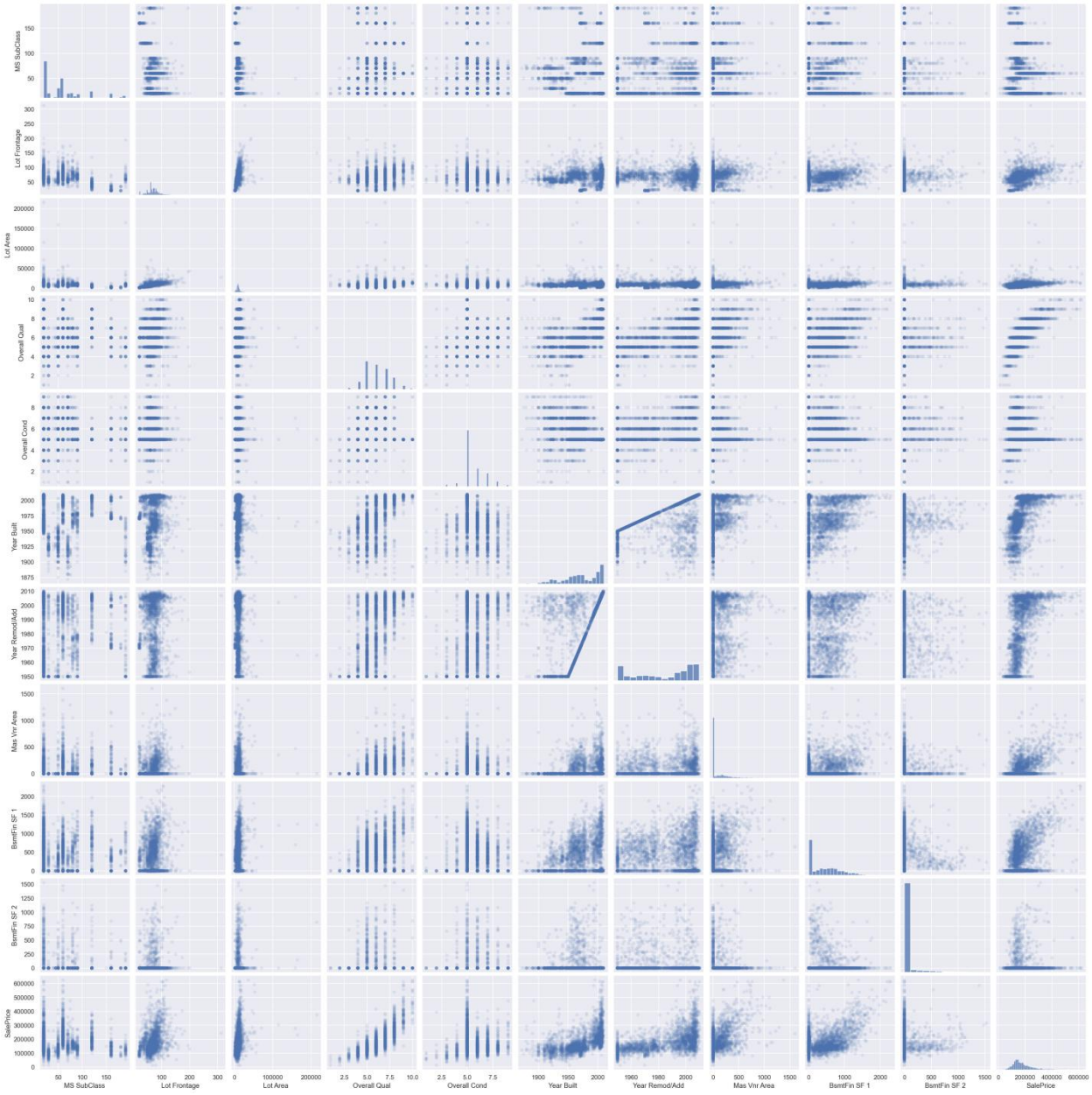

Figure B-1. Pair plot of the first smaller Pandas DataFrame.

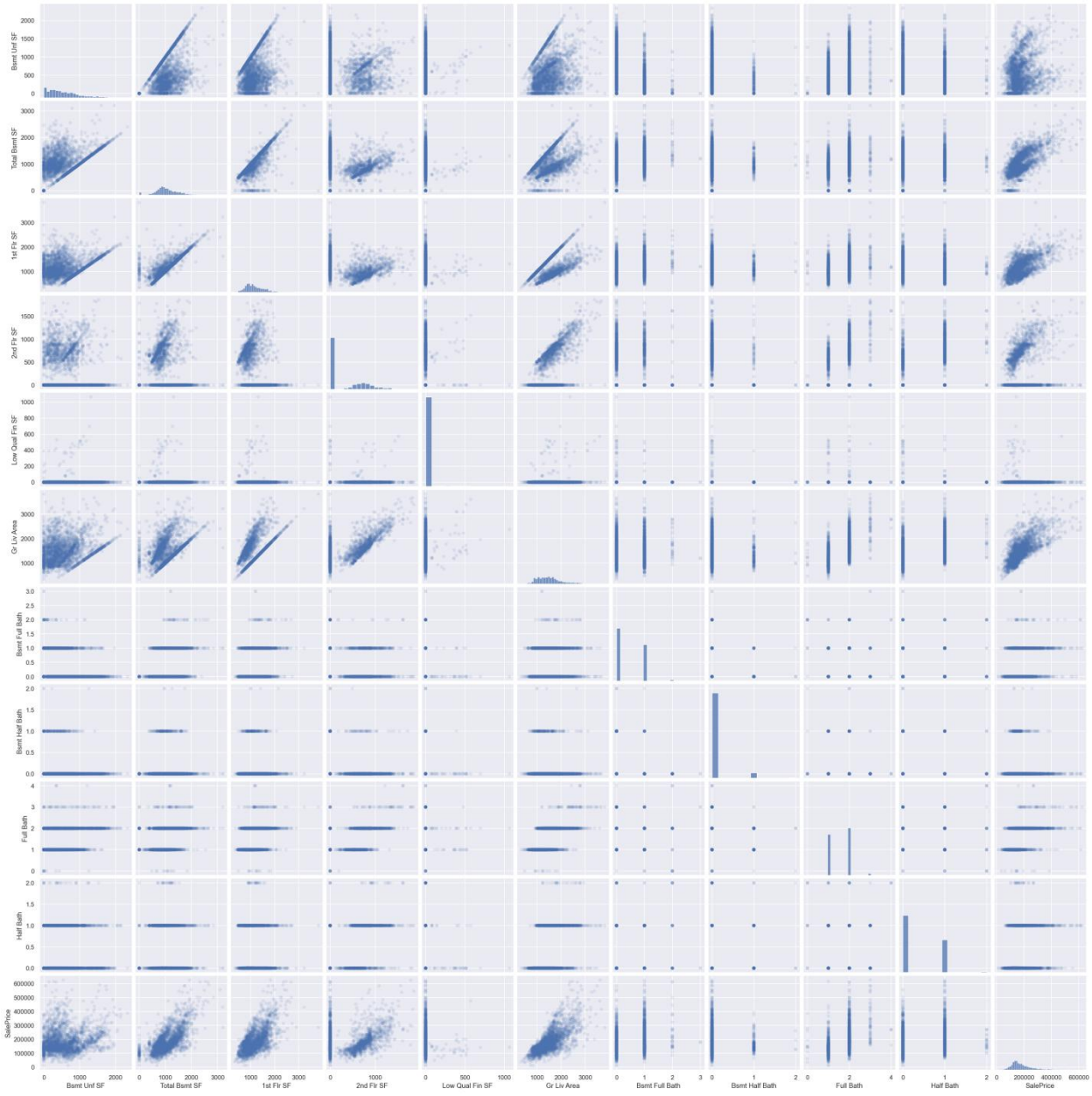

Figure B-2. Pair plot of the second smaller Pandas DataFrame.

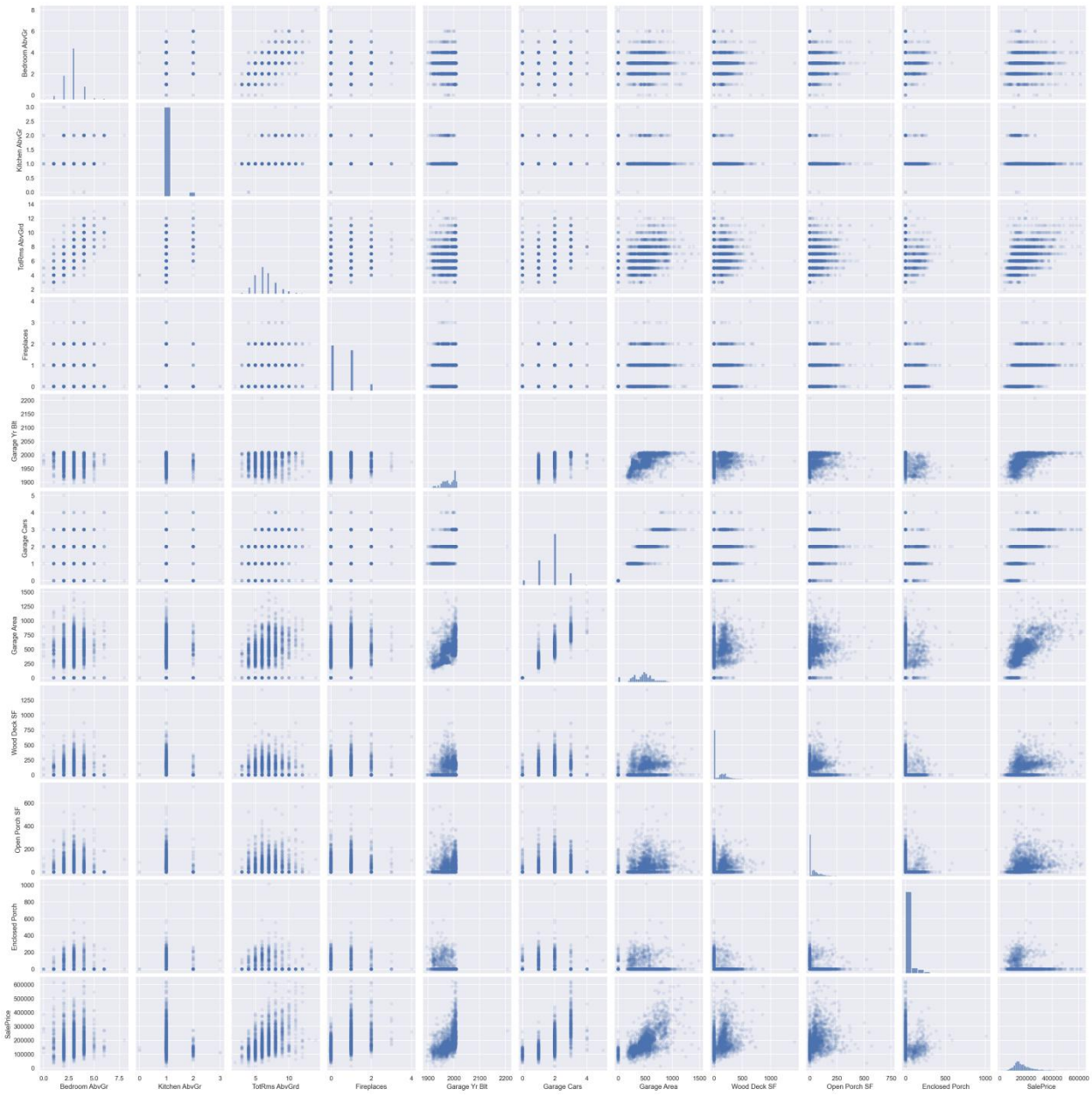

Figure B-3. Pair plot of the third smaller Pandas DataFrame.

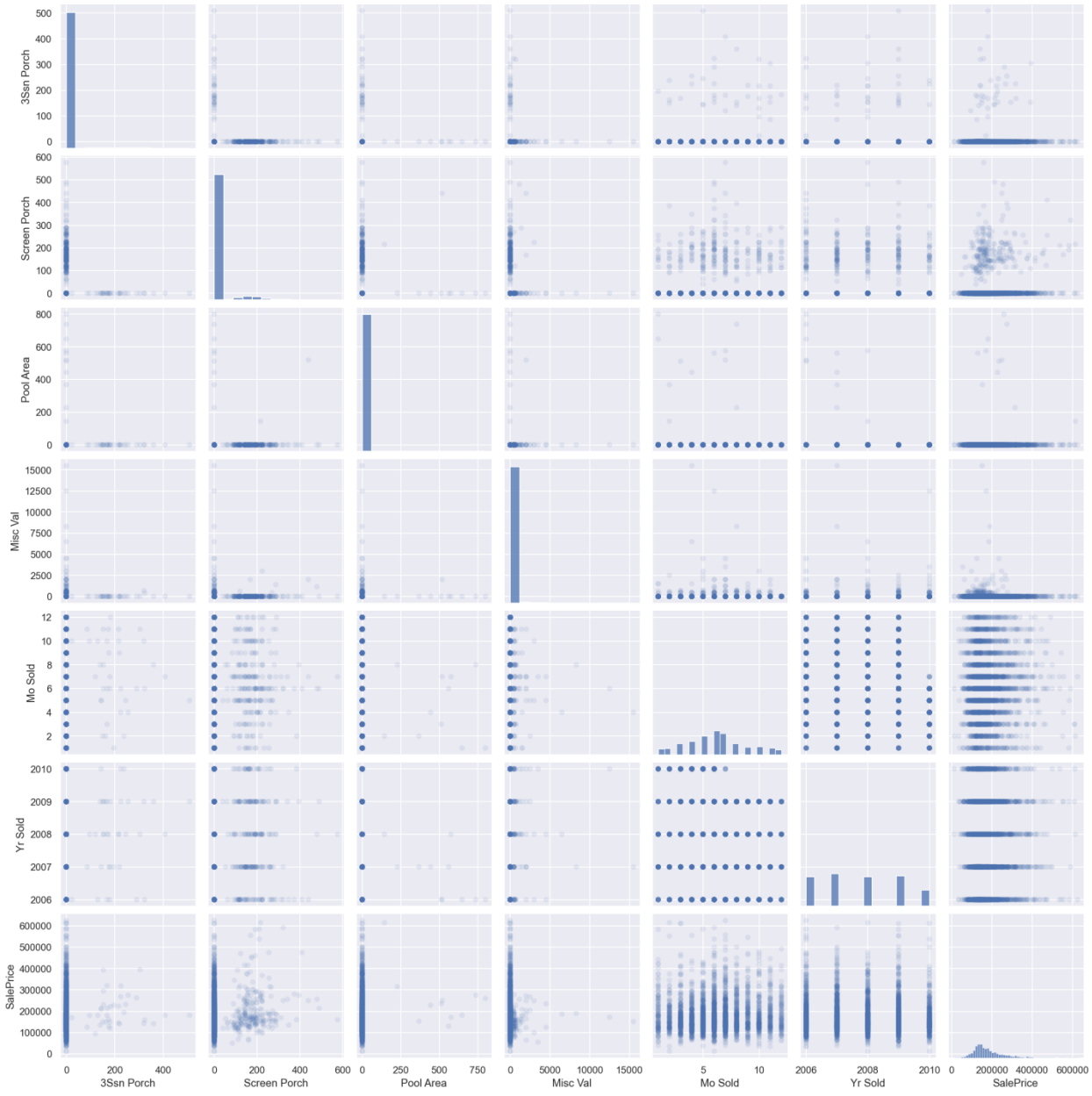

Figure B-4. Pair plot of the fourth smaller Pandas DataFrame.

Appendix C. Histograms and bar plots of the chosen features before and after applying a logarithmic transformation

Figure C-1, Figure C-2, Figure C-3, Figure C-4 and Figure C-5 illustrate the histograms and bar plots of the 5 chosen features before and after applying the np.log1p transformation.

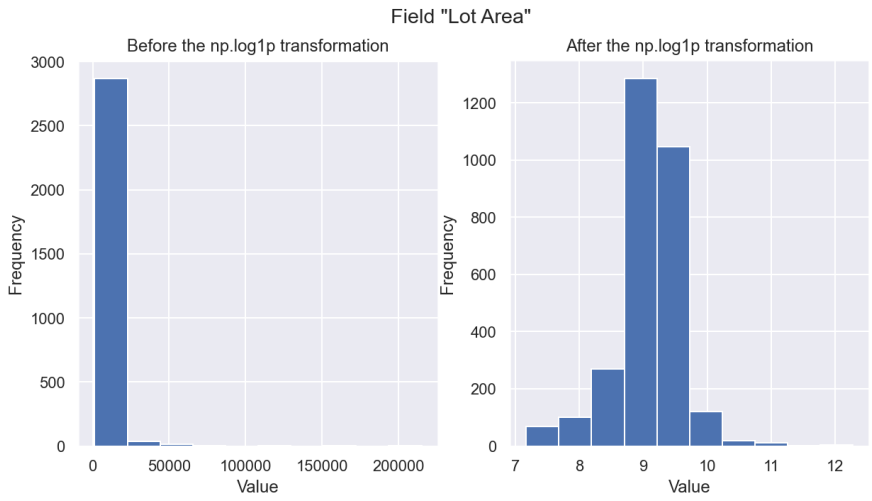

Figure C-1. Histograms of the feature "Lot Area" before and after the transformation.

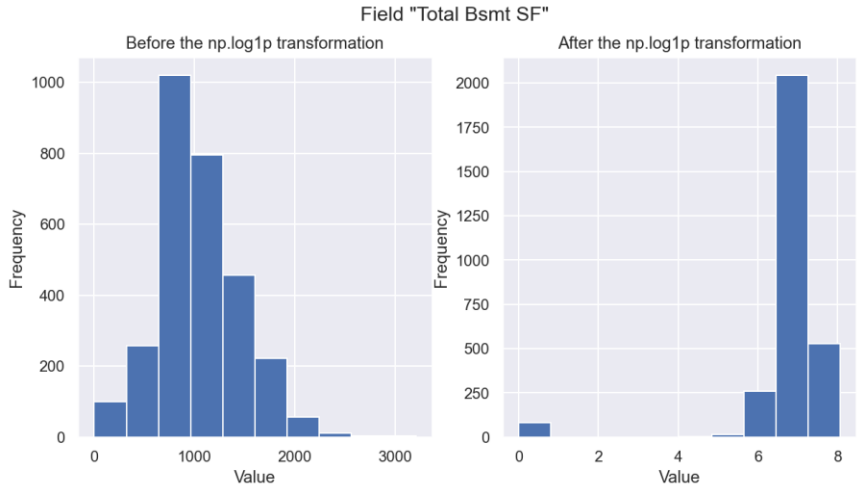

Figure C-2. Histograms of the feature "Total Bsmt SF" before and after the transformation.

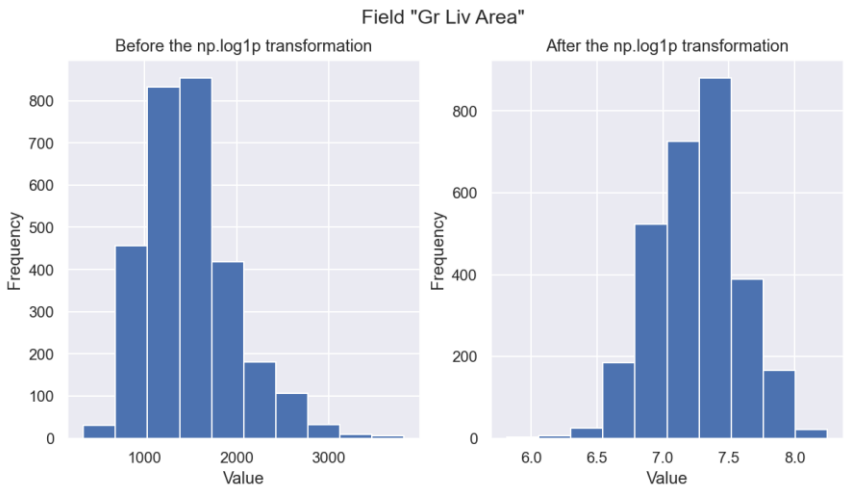

Figure C-3. Histograms of the feature "Gr Liv Area" before and after the transformation.

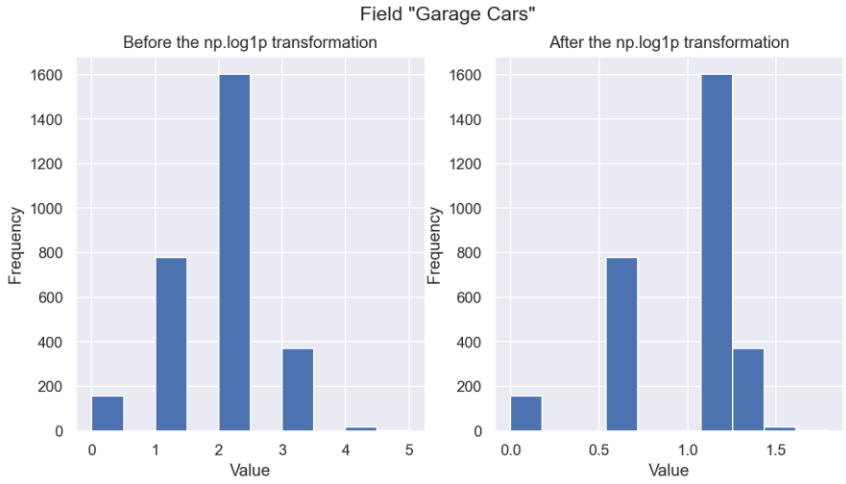

Figure C-4. Bar plot of the feature "Garage Cars" before and after the transformation.

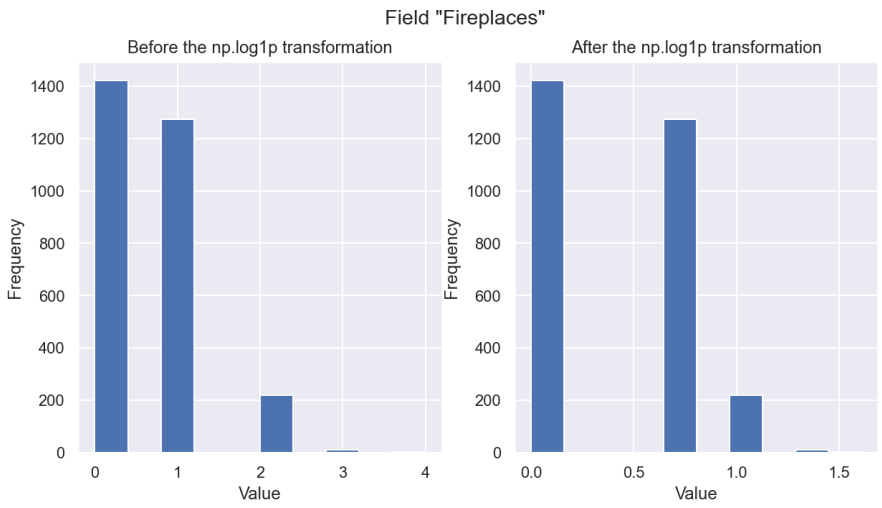

Figure C-5. Bar plot of the feature “Fireplaces” before and after the transformation.