Data Driven Decision Making

EXAM03: Data Science Group Assignment

Group:	The 3rd Bulgarian empire

Student names	Student numbers
1. Silvia Popova	1. 95653
2. Georgi Chitarliev	2. 95567
3. Chavdar Tsvetkov	3. 98008
4. Simeon Atanasov	4. 95616
5. Dimitar Petrov	5. 93425

HZ UNIVERSITY OF APPLIED SCIENCES

WWW.HZ.NL

Content

Iteration 1

BUSINESS UNDERSTANDING

Situation Understanding:

The real estate agency Enchanted Estates works in a magical world where property values vary a lot across different realms and enclaves. The agency has collected historical property data, but they need our help with making sense of the market trends. They believe that a deeper understanding of their data will allow them to improve their client services and build more effective marketing approaches in the market.

Business Objectives:

To provide buyers and sellers with an accurate market value, Enchanted Estates wants to improve its ability to properly evaluate and price properties. The agency can boost client satisfaction, accelerate the selling process, and improve its marketing strategy by developing a greater understanding of market trends and the factors affecting property values. Additionally, by using data-driven insights to provide pricing recommendations, the company aims to grow into a respected leader in the magical real estate market.

Data Mining Goals:

Creating a predictive model that accurately calculates property values from past data is the main objective of data mining. This involves identifying the main factors influencing TradeValue, evaluating price trends across different realms and enclaves, and looking for errors or inconsistencies in trade values. Given the importance of accurate evaluation, we will measure the model's accuracy using Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Success Criteria:

Both technical and business criteria will be used to measure the project's success rate. On the technical side, we want to develop a predictive model with MAE ≤ 60,000. From a business perspective, success will be reflected in the improved level of client satisfaction, accurate property valuations, and reduced time-to-sale for listed properties. All these metrics will provide a clear indication of whether our data analysis is supporting Enchanted Estates' business goal.

DATA UNDERSTANDING

Data Exploration:

Our data consists of 2363 rows and 7 columns, 6 of which represent actual data. Here is the summary:

DI: Data Index. Does not represent data.

#	Column	Non-null	Count	Dtype
0	DI	2,363	Non-null	Int64
1	TradeValue	2,363	Non-null	Float64
2	RealmType	2,363	Non-null	Object
3	Enclave	2,363	Non-null	Object
4	LivingQuarters	2,244	Non-null	Float64
5	ParcelSize	2,363	Non-null	Float64
6	ParcelSizeUnit	2,363	Non-null	Object

Figure 1: Column names inspected using df.info()

TradeValue: Represents the value of the property in gold coins (the cost).

RealmType: The general realm where the property is located.

Enclave: Represents the region where the property is (the neighborhood).

LivingQuarters: Represents the living space of the property.

ParcelSize: The land area of the property.

ParcelSizeUnit: The unit of measurement.

Summary Statistics:

We have 3 columns of numerical value and 3 categorical. 2 of them contain errors/missing information. The summary below was made without omitting the errors:

	DI	TradeValue	LivingQuarters	ParcelSize
count	2,363	2,363	2,244	2,363
mean	1,170	181,355.98	1,157.59	9,825.32
std	676	82,615.06	395.43	7,309.12
min	0	63	256	-13,159
25%	586	129,500	866	7200
50%	1,169	160,000	1,110	9337
75%	1,755	213,000	1,347	11,475
max	2,339	755,000	4,339	164,660

Figure 2: Mean, std and quartile summary

	RealmType	Enclave	ParcelSizeUnit
count	2,363	2,363	2,363
unique	5	47	2
top	Meadowlands	Dragonspire	sqft
freq	1,836	3,50	2,340

Figure 3: Detailed information for the categorical data

Data Insights:

Trade Value varies significantly between **Realm Types.** For example, properties in Aether Isles boast a much higher value than those in Arcane Market. (See *Figure 4*). Also, some realms have many extreme outliers, while others are more stable.

The **Trade Value** seems to increase as the **Living Quarters** increase. (*Figure 6*)

There seems to be no connection between the **Parcel Size** and the **Trade Value**. (See *Figure 5*)

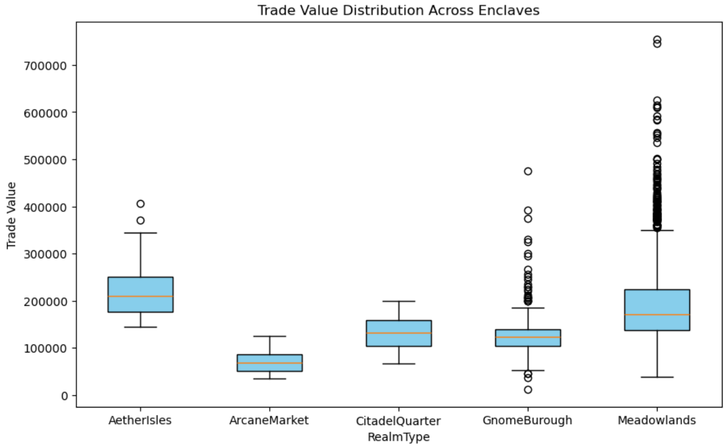


Figure 4: Trade Value distribution boxplot

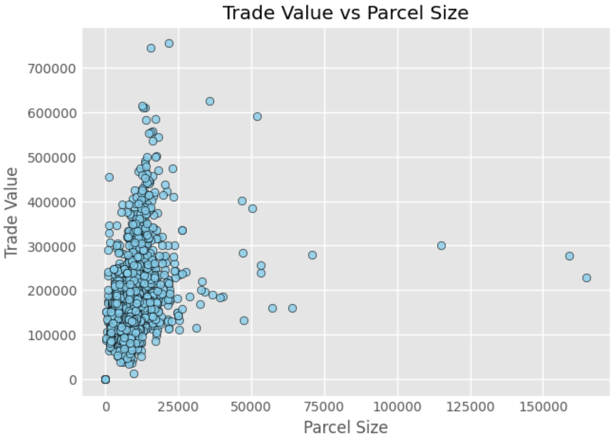


Figure 5: Trade value vs Parcel Size scatterplot

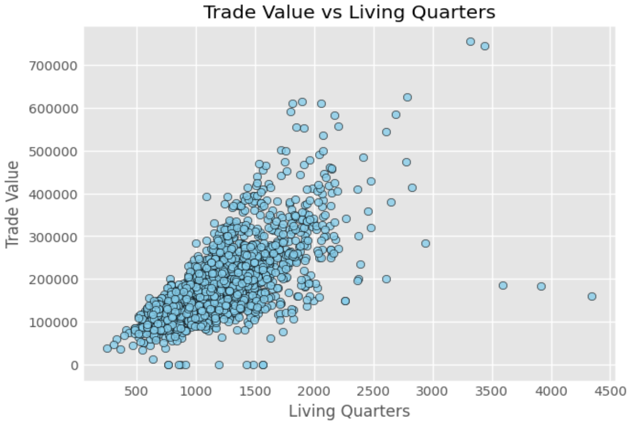


Figure 6: Trade Value vs Living Quarters scatterplot

Data Quality Issues:

We cannot gather insights about the connection between **Trade Value** and **Enclave**, since **Enclave** is full of typos and duplicate records.

In the dataset, there are 2 major errors, which can heavily influence our insights. The ParcelSize should never hold negative values and the LivingQuarters has some missing data (119 empty rows). There are 23 duplicate rows, and the column Enclave has many typos (capital letters, plural).

Furthermore, we can observe some other issues. There is an inconsistency in the ParcelSizeUnit column. It holds 2 values: sqft(predominant) and sqm. Another potential error might be the minimal value of TradeValue - 63. Although the ParcelSize for this data endpoint is also incredibly small, it still seems like an error/outlier.

And lastly, by using the 1.5*IQR method we search for potential outliers. We can deduce that **TradeValue** has 133 potential outliers, **LivingQuarters** has 57 potential outliers, **ParcelSize** has 129 potential outliers.

DATA PREPARATION

The dataset underwent a structured cleaning process to ensure data accuracy and consistency. Based on the errors defined in the previous chapter, here are the steps we undertook to make the dataset prepared to be evaluated (for code snippets of the steps *See Team 8: the 3rd Bulgarian empire - Iteration 1 Notebook*):

- Any records with negative values in the ***ParcelSize*** column were removed, because it cannot be negative.
- Instead of dropping missing values, which would reduce our dataset and potentially discard valuable information, we impute them with the median as it better represents the average property and is less sensitive to outliers, and it prevents biasing predictions toward larger homes.
- Entries where the ***ParcelSize*** was less than 1 were removed since an extremely small parcel size is unrealistic and likely represents errors or misreported/entered size.
- Duplicate rows were identified and eliminated to prevent redundant data points from influencing predictions.
- The ***Enclave*** column was standardized by stripping unnecessary spaces, fixing inconsistent casing (lower- and uppercase)**.** Additionally, variations of enclave names were mapped to their most frequently used correct version to prevent inconsistencies in grouping.
- ***ParcelSize*** was converted from sqm to sqft to ensure all measurements were in a consistent unit.
- We have decided not to remove outliers because doing so would eliminate actual high-value properties that are essential for understanding the full range of trade values. While removing outliers might slightly improve the model's ability to predict general property values, it would come at the cost of failing to predict real extreme cases.

DATA UNDERSTANDING – trend analysis

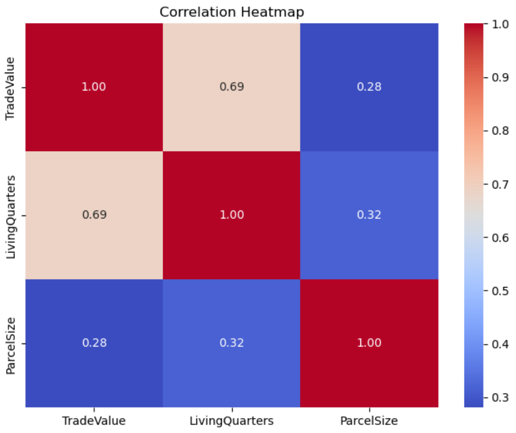


Figure 7: Heatmap for the correlation of all columns

After we have cleaned the data, we can easily answer what are the key factors that could influence trade values and how do they vary across different enclaves. We can look into the correlation between all columns and can vizualize it in a heatmap for better understanding.

LivingQuarters and ***ParcelSize*** appear to have the strongest positive relationships with ***TradeValue***. The correlation between ***LivingQuarters*** and ***TradeValue*** suggests that larger the dwelling is, the higher the trade value is. This makes sense, as larger living areas are generally associated with higher property worth. Similarly, ***ParcelSize*** has a smaller correlation but can suggest that larger parcels may have higher trade values.

Furthermore, we can investigate ***Enclave*** - specific enclaves may have distinct underlying dynamics that affect trade differently. Unleashing our "imagination" for this point, some enclaves might be more commercial (supporting higher trade values), while others might be more restricted. We have used a bar chart to show that trade values vary significantly across enclaves, showing that enclave location plays a role in determining trade value. This suggests that enclave-specific factors can drive trade value to change more than the ***RealmType*** classifications.

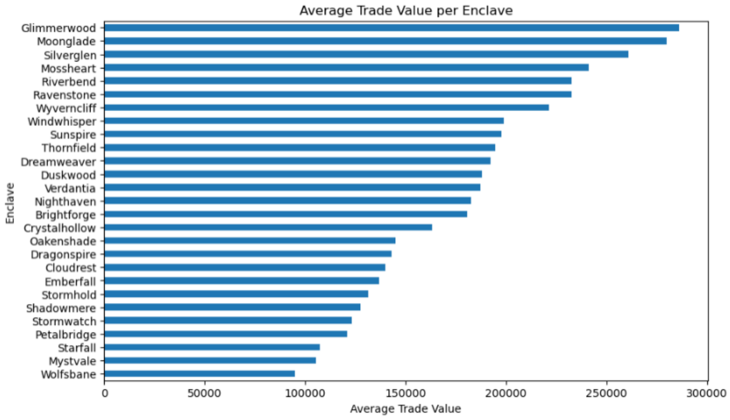


Figure 8: Barchart for the trade values average in each enclave

Now that we have finalized the data, cleaned and polished it, it is ready to be handled by the modeling phase where we will develop a predictive benchmark regression model to analyze the trade value and predict it properly. With a better-structured dataset, we can now train, validate, and optimize our model to improve prediction accuracy and extract useful insights that we will use to help the Enchanted Estates.

MODELING

We have chosen to use the median for our model because it provides a more stable central value than the mean, which is sensitive to extreme values. Since there may be outliers or errors in the dataset that we cannot fully control, using the median makes sure that one or a few extremely high or low values do not skew predictions. While the mean can be shifted by large trade values, the median remains consistent, making it a more reliable benchmark for our predictive model in a dataset with varying property values.

Then we separate our data into two categories: categorical data and data we want to predict. They are separated in X, y respectively: ***X = df.drop(["ParcelSizeUnit", "DI", "TradeValue", "Enclave", "RealmType"], axis=1) & y = df['TradeValue']***. Then we assign the data where we are using 70% of the data for training:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

The final step is to create the basic predictive model. We are using the median and the length of the y_train, and to create the model. *(See Team 8: the 3rd Bulgarian empire - Iteration 1 Notebook for full function)*.

After the predictions we are using metrics, such as Mean Absolute Error and Mean squared error to see what the efficiency of our model is with the MAE = 54,498.995689655174 and the MSE = 6,533,157,540.837644.

EVALUATION

Our model's predictions are, on average, 54,499 magical coins currency off from the actual trade values (MAE – mean absolute error). The Mean Squared Error (MSE = 6,533,157,540.84) further reveals that some predictions are significantly off, especially in high-value enclaves where price estimation becomes more complex. These metrics suggest that while the model captures general trends in property prices, it struggles with properties at the higher end of the market - possibly due to unaccounted variables or uneven distribution of training data across different realms. This highlights that relying solely on basic central tendency methods (like the median in our case) does not sufficiently model the true complexity of trade values across different enclaves and realms.

Trade values are mainly influenced by ***RealmType***, ***LivingQuarters*** size, and ***Enclave***. Larger LivingQuarters generally lead to higher prices. Some enclaves, like '***Shadowmere***'and '***Ravenstone***', consistently have higher property values, while others, like '***Emberfall***' and '***Dragonspire***', fall in the mid-range. This suggests that location within a realm has a significant impact on pricing, and different enclaves may have unique economic or magical influences affecting property worth.

One major problem with our model is that while it performs reasonably well on training data, its errors remain large when tested on new validation data. This occurs because the model does not fully capture the relationships between ***RealmType***, ***LivingQuarters***, and ***Enclave***. To better understand this issue, we need to analyse specific validation set errors - check whether mispredictions occur in specific enclaves, dwelling sizes, or extreme trade values. To improve accuracy, besides needing more data, we should test alternative more complex models and adapt our approach to recognize these patterns better. Additionally, using a cross-validation can help ensure that the model generalizes well to new data and does not just memorize patterns from the training set.

PERSONAL CONTRIBUTION

	Student name	Contribution
1	Georgi Chitarliev	Data Understanding • Include your data exploration (e.g., summary statistics, visualizations). • Summarize key trends in the data - e.g., average trade values by enclave • Check the data quality - check missing values, detect outliers, detect duplicates, detect typos and other errors
2	Silvia Popova	Evaluation • Assess the benchmark model and discuss limitations of using a centrality-based benchmark. • Provide initial insights into trade value patterns. Data Preparation • Define trends in the cleaned data and answer Q1 and Q2 from the iteration.
3	Chavdar Tsvetkov	Data Preparation • Clean the dataset - fix the data quality issues you found in Data Understanding and prepare it for analysis. • Describe and justify the steps taken to clean and preprocess data.
4	Simeon Atanasov	Modeling • Create a benchmark regression model using a measure of centrality • Describe and justify how you have set up/created the model • Describe and justify how you have tested the model • Describe and justify the performance of your model using appropriate metrics
5	Dimitar Petrov	Business Understanding • Describe the situation as you understand it. • Define and justify the business objective(s) • Define and justify the data mining goal(s) • Determine success criteria for this iteration.

Iteration 2

BUSINESS UNDERSTANDING

Situation Understanding:

The HomeFirst Enchantment program was started by the High Council of Realms to help low-income families in locating affordable dwellings across several enchanted enclaves. Affordable homes, according to the council, are ones whose yearly housing expenses don't surpass 30% of a household's income. They require assistance in creating a classification model that will help them assess if the property is cheap to support this project.

Business Objectives:

By utilising a classification model, the High Council of Realms can effectively determine how affordable housing is. That will guarantee that families are guided towards appropriate dwellings that fulfil the affordability standards. Among the business goals are the following: maximising housing policy efficiency through data-driven recommendations; transparency in the housing market by defining clear rules to determine affordability; and improving decision-making for council officials and real estate consultants by classifying dwellings based on affordability.

Data Mining Goals:

Creating a basic rule-based classification model for determining a home's affordability is the main objective of data mining. This involves figuring out the main elements that affect affordability. Analyzing affordability trends across different enclaves to identify areas with a shortage of affordable housing is another objective. Through an optimization of the decision rules, the model can be continually improved to boost classification accuracy.

Success Criteria:

To ensure that most of the predictions match real affordability, a good model should be able to categorize the most affordable homes with an accuracy of at least 80%. For the model to efficiently identify inexpensive homes, precision and recall should both exceed 85%. To prevent classification mistakes from having a major influence on decision-making, the misclassification rate should stay below 20%. Furthermore, to ensure that the model continues to perform well overall in finding inexpensive properties while minimizing classification mistakes, the F1 Score should be at least 0.85.

DATA UNDERSTANDING – DATA EXPLORATION

The add-on dataset introduces 2 additional columns (***StructuralIntegrity, ExteriorCharm***) to our existing dataset. They provide insight into the condition and appeal of the dwellings. While these factors are not directly related to finances, well-maintained homes may have higher trade values, while homes that are visually appealing but structurally weaker might be less affordable.

DATA PREPARATION – MERGING, VARIABLES, TARGET

After cleaning the Iteration 1 dataset, it should be exported and re-imported as ***set8_it1_cleaned.csv*** (exported in *Team 8: the 3rd Bulgarian empire - Iteration 1 Notebook*). The add-on dataset (***df_it2***) should be checked for duplicates, which can be easily removed. Then both datasets are ready for a left merge using the DI column as the key: ***df = df_it1.merge(df_it2, on='DI', how='left')***. We merge on the left so any outliers we have dropped will not be visible in the new dataset and then check if there are any missing values because of the merge. Next, we define the key financial variables from the assignment (See *Team 8: the 3rd Bulgarian empire - Iteration 2 Notebook*). The variables are ***annual_income, monthly_income, interest_rate_year, interest_rate_month, loan_term,*** and ***total_payments.*** We also created separate columns for the ***Downpayment, LoanAmount, MonthlyPayment,*** and ***AffordableDwelling*** to be able to easily track those values.

DATA UNDERSTANDING

The new dataset consists of 2 318 properties, with an average trade value of 182 213 and an average monthly loan payment of around 696. That shows that a lot of the dwellings close to 3/4 are classified as affordable. However, there is a big variety in property values and loan amounts, with some dwellings reaching up to 755 000 in trade value and loan monthly payments going over 2 884 - that suggests an affordability difference in different enclaves.

	Trade-Value	Living-Quarters	Parcel-Size	Structural-Integrity	Down-payment	Loan-Amount	Monthly-Payment	Affordable-Dwelling
count	2,318	2,318	2,318	2,318	2,318	2,318	2,318	2,318
mean	182,213.3	1,155.3	10,060.29	5.57	36,442.66	145,770.6	695.9	0.69
std	82,064.9	386.4	7,158.77	1.11	16,412.99	65,651.96	313.4	0.46
min	12,789	256	1,470	1	2,557.8	10,231.2	48.8	0
25%	129,900	879	7,410.25	5	25,980	103,920	496.1	0
50%	160,500	1,110.5	9,378.5	5	32,100	128,400	613	1
75%	213,220.75	1,329	11,546.25	6	42,644.15	170,576.6	814.4	1

Table 1: Mean, std and quartile summary

Data Insights:

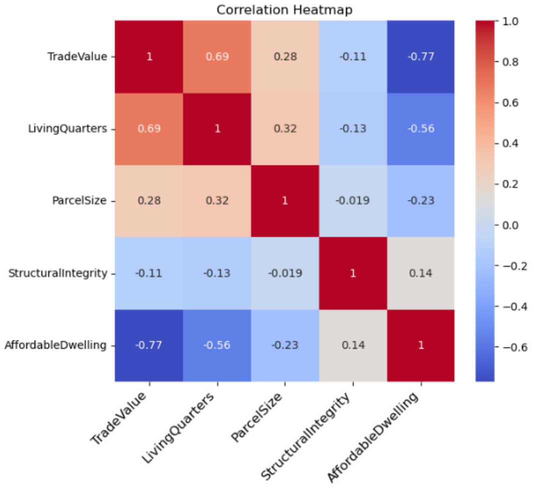
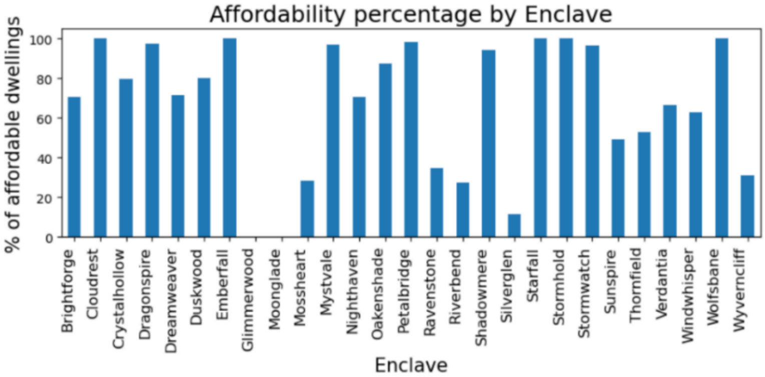


Figure 1: Correlation Heatmap

To highlight key relationships between variables we can use the correlation heatmap.

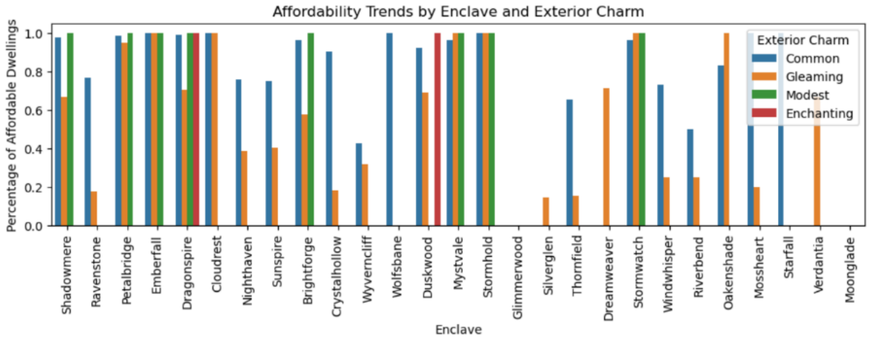
The primary metric for affordability is based on whether monthly loan payments exceeded 30% of a household's monthly income. That is the reason for the strongest negative correlation is between ***AffordableDwelling*** and ***TradeValue (-0.81)*** - meaning that as TradeValue increase, affordability significantly decreases.

Interestingly, ***StructuralIntegrity*** has a weak positive correlation (0.14) with affordability - better-maintained homes may slightly contribute to affordability but are not a major factor. These insights show that ***Enclave***, ***LivingQuarters***, and ***ExteriorCharm*** are the strongest predictors of affordability.



On the connection between affordable dwellings and enclave, another key visualization is a bar chart showing the percentage connection. It appears that there is a significant variation of the number of affordable dwellings between different enclaves. The affordability is strongly influenced by ***TradeValue***, with enclaves that have lower median dwellings' prices having higher affordability rates.

Figure 2: Affordability percentage by enclave boxplot



Adding the ***ExteriorCharm*** as a feature we can see how affordability trends vary across enclaves and different levels of exterior charm. In many enclaves, homes with "**Common**" or "**Modest**" exterior charm tend to have the highest affordability rates, while homes with "**Gleaming**" or "**Enchanting**" charm are generally less affordable. Some enclaves, such as **'Ravenstone'** and **'Crystalhollow'**, show an evident affordability gap, where homes with higher charm levels are much less accessible, indicating that aesthetic appeal may drive up trade values and reduce affordability.

Figure 3: Affordability trends by enclave and exterior charm boxplot

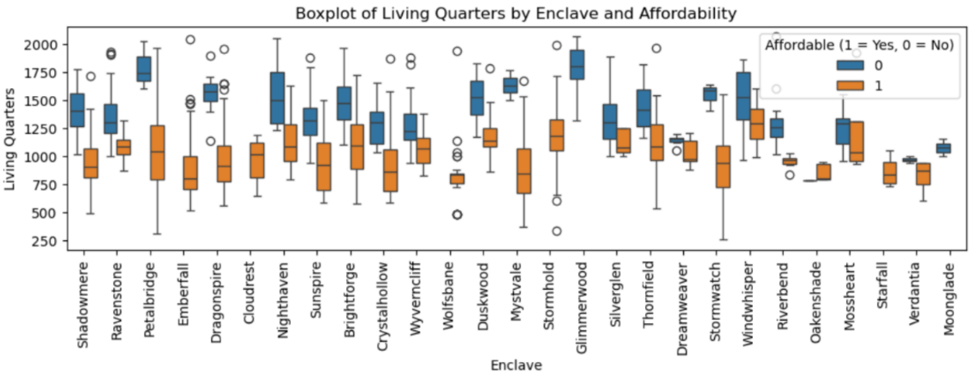


Figure 4: Affordability trends by enclave and living quarters boxplot

That is why we need to investigate ***LivingQuarters***. In many enclaves, non-affordable homes tend to have significantly larger living spaces, while affordable homes generally have smaller median sizes. Some enclaves, such as **'Stormwatch'** and **'Shadowmere'**, show a clear separation between affordable and non-affordable homes, whereas others, like **'Mossheart'** and **'Moonglade'**, have a more balanced distribution, indicating that affordability in those areas may be influenced by factors beyond home size.

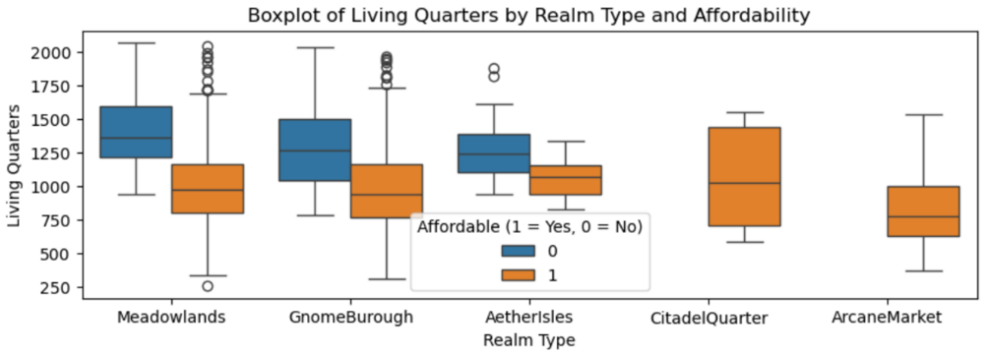


Figure 5: Affordability trends by real type and living quarters boxplot

Expanding on our findings, certain realms tend to have larger homes that are mostly unaffordable, while others show a greater mix of home sizes. In '**Meadowlands'** and **'GnomeBorough'**, larger homes are typically less affordable, while smaller homes in these realms have a higher likelihood of being classified as affordable. '**CitadelQuarter**'and '**ArcaneMarket**', on the other hand, consist entirely of affordable homes, suggesting that either these realms have smaller average home sizes or different economic conditions that make housing more accessible.

MODELLING

First, since around 70% of dwellings are affordable, we create a simple benchmark model which would be one that always predicts True (1) – it will reflect the class imbalance in our dataset and will help us determine if our rule-based models provide added value beyond guessing the most frequent outcome. It has an accuracy of 69.25% on the test split.

After that, we will build a rule-based classification model because affordability follows clear, structured patterns based on enclave location, realm type, and property characteristics. First, we must split the set into 70% training and 30% testing sets, separated into x and y respectively. For this, we define the target and features first: ***target = df3.pop('AffordableDwelling') & features = df3.*** And then we split: ***feature_train, feature_test, target_train, target_test = train_test_split(features, target, train_size=0.7, test_size=0.3, random_state=42).*** To ensure accurate predictions, we must define classification rules based on the insights gained from the Data Understanding phase. We derived the following rules:

- Always affordable:
  - Realm types: 'CitadelQuarter', 'ArcaneMarket'
  - Exterior Charm: 'Modest'
  - Enclaves: 'Starfall', 'Stormhold', 'Wolfsbane', 'Oakenshade', 'Cloudrest', 'Emberfall'
- Never affordable:
  - LivingQuarters > 1500
  - Enclaves: 'Glimmerwood', 'Moonglade'
- Conditional affordability:
  - 'Enchanting' + enclave in ['Duskwood', 'Dragonspire'] → affordable

- 'Common' or 'Gleaming' charm has partial affordability based on enclave. For some enclaves, we introduce randomness to simulate borderline cases that cannot be captured by strict rules (e.g., 76% chance for affordability in 'Ravenstone' and 'Sunspire').

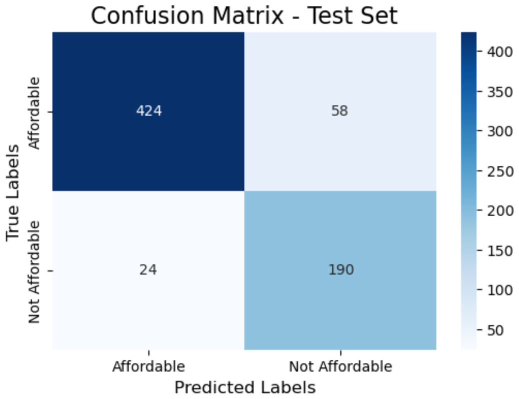
The final rule-based model relies on multiple if-else statements, structured conditions, and, in one version, randomized elements. We introduce randomness intentionally to reflect real-world variability in affordability where no clear deterministic rule applies. For comparison, we also test the model without randomness to evaluate whether it significantly affects accuracy. While randomness introduces minor fluctuations in results, it might help simulate the nuances of housing affordability across magical enclaves without using complex machine learning algorithms. The model used is detailed in the *Team 8: The 3rd Bulgarian Empire – Iteration 2 Notebook*.

After generating predictions, we calculate the accuracy using ***accuracy_score*** from ***sklearn.metrics***. The model with randomness achieves 83% test accuracy, while the model without randomness performs slightly better at 88%, making it the preferred version in terms of raw accuracy. However, the random-enhanced model still offers value in capturing uncertainty and may be more suitable when interpretability and realism are desired over exactness (more explained in the next chapter: Evaluation)**.**

EVALUATION

After the model is being created and trained it comes time to how efficient it is. We create a confusion matrix, which looks like the following:

Figure 6: Confusion matrix of the model



The metrics we are using to determine if our model is efficient are:

- **Accuracy**: Measures the proportion of correctly classified dwellings out of all predictions.
- **Precision**: Indicates how many of the predicted affordable dwellings are affordable. This is important to minimize false positives and avoid incorrectly labeling unaffordable homes as affordable, which could mislead low-income families into considering properties beyond their financial capabilities, causing financial issues.
- **Recall**: Measures of how many of the truly affordable dwellings are correctly identified by the model. We use this to ensure that we do not overlook affordable homes, which could prevent families from accessing suitable housing options.
- **F1 score**: A single metric to assess both false positives and false negatives. This is useful for evaluating the model when both types of errors carry significant consequences, as is the case with our project
- **Misclassification rate**: Shows the percentage of incorrect predictions made by the model.

Based on these metrics, our model successfully meets the defined success criteria, demonstrating a very good performance in affordability classification. Here are our results:

- Accuracy of 0.83 - good overall performance that the model correctly classifies the majority of dwellings
- Precision of 0.88 - the model predicts dwelling as affordable 88% of the time
- Recall of 0.88 - the model is identifying 88% of the truly affordable dwellings
- F1 Score of 0.88 - there is a good balance between precision and recall
- Misclassification rate of 0.12 - the model makes an error 12% of the time, meaning it still maintains a high level of accuracy

To improve the model in the future we may need to dive deeper into some aspects of our model. The first aspect is better understanding of the misclassification of our features so we can reduce the misclassification rate to a minimum. Secondly, we may need to focus better on the data preparation stage, where we can change the values of the outliers with the median values. The affordability classification model will be further improved, and performance will be maximized with these improvements.

PERSONAL CONTRIBUTION

	Student name	Contribution
1	Georgi Chitarliev	Modeling • Create a benchmark classification model using simple rules. • Improve the benchmark model by adding more rules, based on the findings of the data understanding.
2	Silvia Popova	Data Understanding • Include your data exploration (e.g., summary statistics, visualizations). • Summarize key trends in affordability patterns (e.g., percentage of affordable dwellings by enclave). Overall document structure check
3	Chavdar Tsvetkov	Business Understanding • Describe the current situation and define the problem and business objectives. • Define and justify the data mining goal(s) to be able to determine the success criteria for this iteration.
4	Simeon Atanasov	Evaluation • Evaluate the model using a confusion matrix and appropriate classification metrics – accuracy, precision, recall, F1 score, misclassification. • Discuss how well the benchmark model performs and identify areas for improvement.
5	Dimitar Petrov	Data Preparation • Merge, clean and prepare the dataset • Create calculation variables based on assignment and a binary target variable (AffordableDwelling) correctly following the formula stated in the case.

Iteration 3

BUSINESS UNDERSTANDING

Situation Understanding:

RenovateNow is a property management company specializing in acquiring enchanted dwellings that require restoration. A dwelling needs restoration, if it was built before 1300 and if the StructuralIntegrity is 6 or less. They want to refine their understanding of trade values by creating a more accurate regression model and an additional decision tree model to classify whether a dwelling needs restoration. Overall they are looking for more advanced models to enhance their restoration strategy.

Business Objectives:

By utilising a decision tree model, the company wants to effectively determine whether a dwelling needs restoration or not. That will guarantee that all dwellings are in prime condition. The business goals are maximising restoration choice efficiency through data-driven insights; transparency in the choice by defining clear rules to determine restoration; and improving decision-making for the company by classifying dwellings based on restoration necessity and lastly, refining their understanding of TradeValue through data-driven insights.

Data Mining Goals:

Creating a basic decision tree model for determining a home's restoration needs is the main data mining objective, which involves determining the main elements that affect restoration necessity. A secondary goal is to gather insights around the dwelling's TradeValue. Through an optimization of the decision rules, the model can be continually improved to boost classification accuracy, as well as the regression model – can be improved based on the data gathered from the classification model.

Success Criteria:

To ensure that the predictions match real restoration needs, a good classification model should be able to categorize the most damaged dwellings with an accuracy of at least 80%. For the model to efficiently identify ran-down dwellings, precision and recall should both exceed 85%. To prevent classification mistakes from having a major influence on decision-making, the misclassification rate should stay below 15%. Furthermore, to minimize mistakes, the F1 Score should be at least 0.75. For the linear regression model predicting the TradeValue, we want to improve our initial model and results, so the success criteria is MAE ≤ 25,000 and MSE ≤ 1,110,000,000.

DATA UNDERSTANDING

The new dataset provides us with two new columns (‘***Craftsmanship’***, ‘***EraConstructed’***) to our already existing dataset. They provide information about the how good a dwelling was built and when it was built. There may be correlation we should investigate between craftsmanship and era constructed.

DATA PREPARATION

After trying to clear the dataset in a sense to remove the duplicates, to improve the inconsistency in the type of variables, we have reached the conclusion that the dataset does not have lacking sides and can be merged without any problem. We have then merged on the left with the cleaned dataset from Iteration 2, because we need the previous data so we can understand the new one better, and check if there are no missing values because of the merge. According to the task from Iteration 3 we should create a new variable called ***NeedsRestoration***, which will be dependable on ***EraConstructed*** and ***StructuralIntegrity***.

DATA UNDERSTANDING

The newly merged dataset consists of 2,318 rows, with an average trade value of approximately 182,213. This indicates a diverse range of property values, with some reaching up to 755,000. The dataset reveals significant variability in key features. The Living Quarters with mean size 1,155, with a range from 386 to 4,339, highlighting differences in living space across properties. Parcel Size ranges from 1,470 to 164,660, showing a wide variety in land sizes. Structural Integrity with an average score around 5.6, suggesting most properties are moderately well-maintained.

We begin with a correlation heatmap. ***LivingQuarters*** shows a strong positive correlation with ***TradeValue*** (0.69), meaning larger dwellings tend to be more valuable. Even stronger is the relationship between ***Craftsmanship*** and ***TradeValue*** (0.80), suggesting high-quality builds significantly boost their value. While not included in this numeric heatmap, categorical features like RealmTyp, Enclave and ExteriorCharm should also be investigated further, as from the previous Iterations we know they could be important drivers of value.
To gain deeper insights, both Pearson (linear) and Spearman (rank-based) correlations were calculated (See *Team 8: The 3rd Bulgarian Empire – Iteration 3 Notebook*). While the overall trends were consistent, Spearman revealed slightly stronger associations with non-linear patterns, such as ***ParcelSize*** and ***TradeValue***.

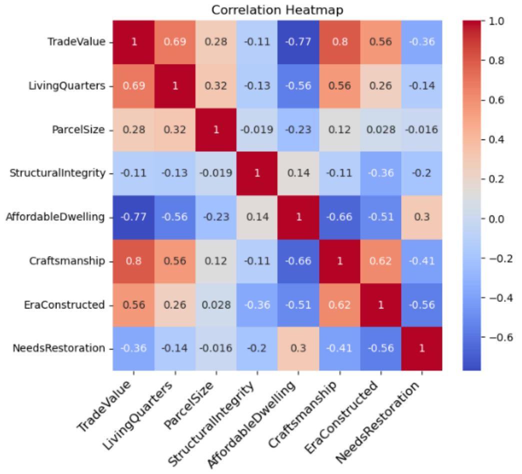


Figure 1: Heatmap

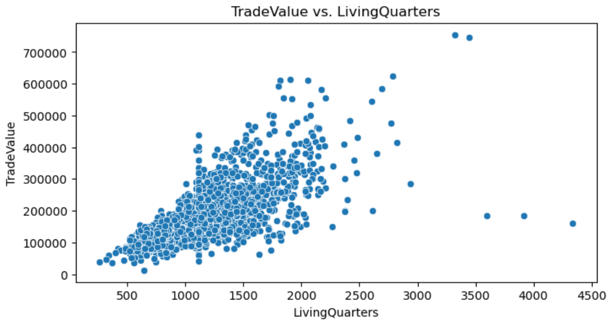


Figure 2: Scatterplot for LivingQuarters and TradeValue

Starting with the relationship between ***TradeValue*** and ***LivingQuarters*** showcased with this scatterplot. We see a positive trend, indicating that as the size of the living quarters increases, the trade value of the property tends to rise. This suggests that larger living spaces are generally associated with higher property values (as mentioned several times). While the overall trend is positive, there is noticeable variability. Properties with similar living quarters can have different trade values, likely due to other influencing factors such as location, condition, etc.

Coming to how ***TradeValue*** varies across different levels of ***ExteriorCharm*** with this boxplot. For the Common charm we see lower median trade value compared to other categories and a wide range of trade values. For the Gleaming charm, there is a higher median trade value than Common, suggesting a positive impact on property value and wider range, meaning there are diverse property values within this category. For the Modest exterior charm, we see a similar structure to Common but with less variability. There are more consistent trade values with fewer outliers. The Enchanting charm has the highest median trade value - strong positive impact on property value, but there is less variability. All of this suggests that aesthetic appeal significantly influences property value.

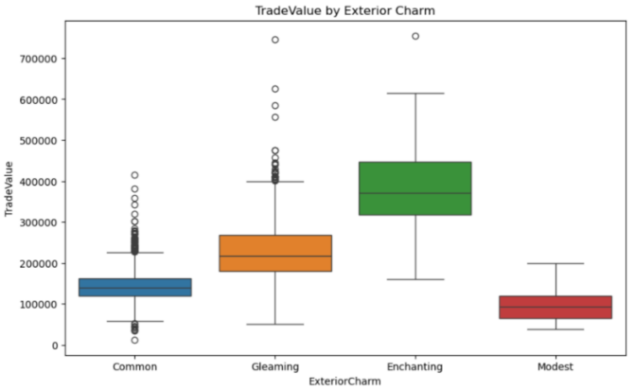


Figure 3: Boxplot showing TradeValue vs ExteriorCharm

Lastly on ***TradeValue*** trends, we will see a strong positive connection with ***Craftsmanship***. As craftsmanship ratings increase, so does the median trade value of the properties, with the spread of values also becoming wider at higher levels. Lower craftsmanship levels (1–4) have relatively low and more tightly clustered trade values, while higher levels (especially 8–10) not only have significantly greater value but also show more variability. This suggests that craftsmanship is a key influence of dwellings trade value.

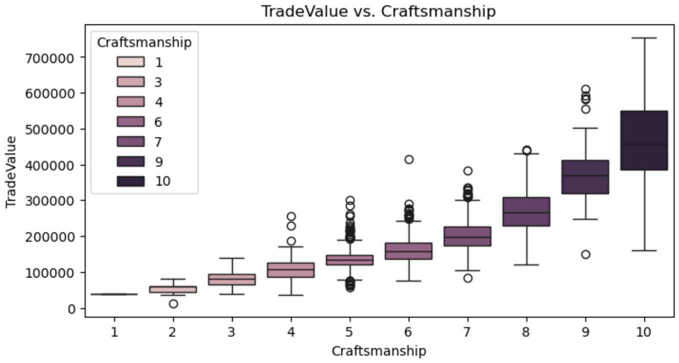
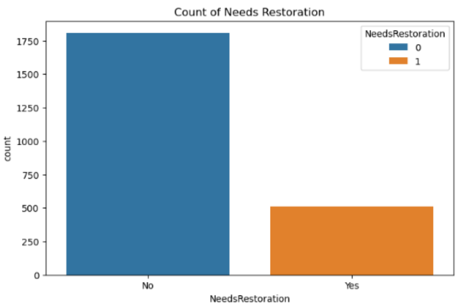


Figure 4: Countplot for *TradeValue vs. Craftsmanship (ordinal variable)*



Moving forward to restoration and how many properties need one. A significant majority of properties are marked as False for ***NeedsRestoration***, indicating they do not require significant restoration. A much smaller portion of properties are marked as True. This suggests that most properties are in a condition that does not require immediate restoration, which could imply better overall structural integrity or more recent construction dates.

Figure 5: Countplot showing the number of dwellings (not)/in need of restoration

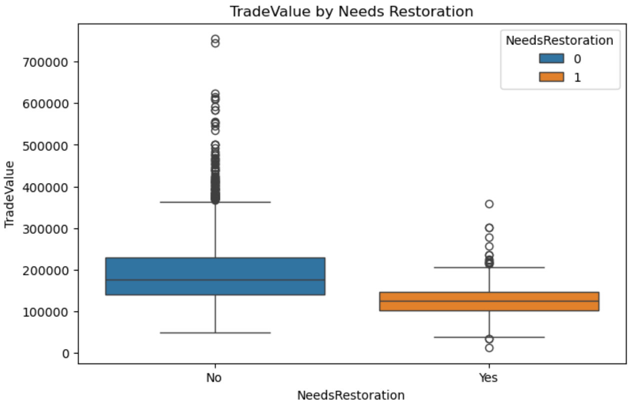


Figure 6: Boxplot for TradeValue and NeedsRestoration

And we can also combine ***TradeValue*** and ***NeedsRestoration***, to see how the value varies based on whether a property needs to have a restoration done on it. The properties that don't need restoration have a higher median trade value compared to those needing restoration. The range of trade values is also broader, which indicates more variability in property values. Accordingly, the properties needing restoration have a lower median trade value, suggesting that restoration needs may decrease property value and the range is narrower, showing more consistency in trade values among properties needing restoration.

To further analyse what influences the decision to renovate a dwelling or not, we can also combine ***Craftsmanship*** and ***NeedsRestoration***. The countplot shows that dwellings with higher craftsmanship ratings are far less likely to require restoration, while those with mid to low craftsmanship levels have a much higher need of restoration. In particular, craftsmanship levels around 5 and below show an increase in properties marked for restoration. This suggests that lower craftsmanship is a strong indicator of restoration necessity.

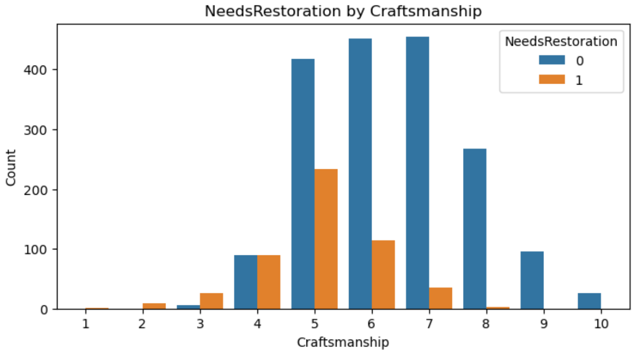
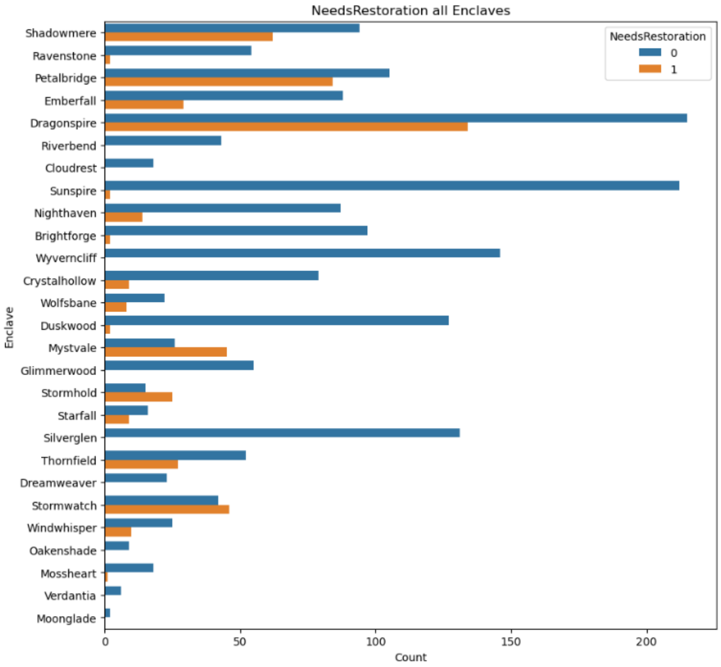


Figure 7: Countplot showing the number of dwellings needing restoration based on their craftsmanship



Finally, we will see the distribution of properties needing restoration across different enclaves. While most enclaves have a larger number of properties that do not require restoration, certain enclaves like Dragonspire, Petalbridge, and Emberfall show noticeably higher counts of dwellings in need of restoration compared to others. In contrast, enclaves such as Sunspire, Stormhold, and Starfall have very few or no restoration needs. This variation suggests that location plays a significant role in the condition of properties, making ***Enclave*** a potentially valuable feature in predicting restoration needs.

Figure 8: Countplot showing the number of dwellings needing restoration per Enclave

MODEL - linear regression model

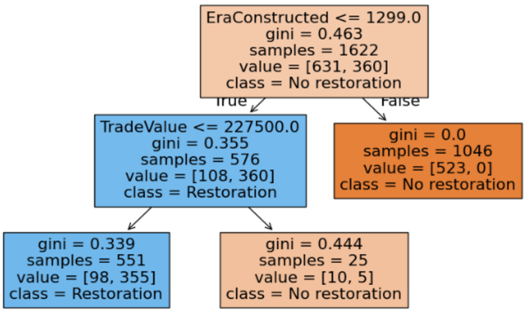
In the linear regression model, we aim to predict the TradeValue. We first preprocess the data by removing irrelevant or target-leaking columns, such as ***Downpayment*** and ***LoanAmount***, and transform nominal variables like ***ExteriorCharm***, ***Enclave*** and ***RealmType*** using one-hot encoding (converting categorical variables into binary columns) to make them suitable for modelling. We decided to use this approach because dropping them would mean losing potentially valuable information for predicting ***TradeValue***. After splitting the data into training and test sets, we trained a linear regression model. Afterwards we explore the regression coefficient, which confirm that location and charm have the strongest impact on TradeValue, with properties in **Moonglade** and those with **Enchanting** exteriors contributing the most to higher values. Features like **AffordableDwelling** have a strong negative coefficient, confirming that affordable homes are generally cheaper. Features such as ***LivingQuarters/ParcelSize*** have smaller effects.

EVALUATION - linear regression

Set	MAE	MSE
test	20 470.13	1 104 503 677.47

The linear regression model performs significantly better than our initial baseline model, which simply predicted the median ***TradeValue*** and resulted in a high MAE of 54,499 and MSE of over 6.5 billion. In contrast, our improved model achieved a much lower MAE of 20,470 and MSE of approximately 1.1 billion on the test set, indicating a more accurate and reliable prediction of trade values and meaning the model meets the success criteria. This model demonstrates consistent performance across training and test data, with no clear signs of overfitting. We can conclude that key factors influencing ***TradeValue*** are the size of the ***LivingQuarters***, the ***Enclave*** and ***RealmType*** the dwelling belongs to, and the ***ExteriorCharm***. All of these features were found out as important during data understanding and confirmed through the model's coefficients. These features capture the essence of what makes a property more or less valuable in this housing market.

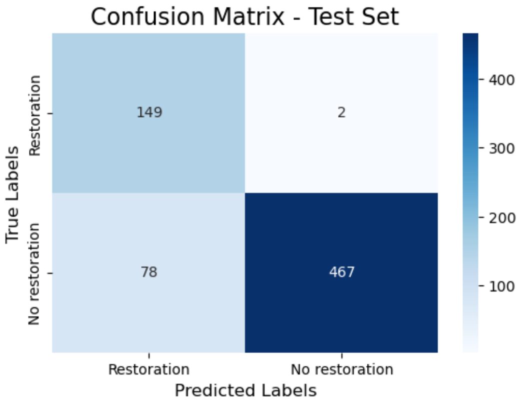
Figure 9: Decision tree plottree



MODEL - classification model (tree)

Since around 3/4 dwellings don't need restoration, we use a benchmark model that always predicts False (no restoration). This will reflect the class imbalance in our dataset and helps evaluate whether our model performs better than a naïve guess. This baseline achieves an accuracy of approximately 78% on the test set but 0% recall. Then we do our actual decision tree classifier model with the goal to predict whether a dwelling needs restoration. The model was trained using a shallow decision tree with a limited depth to maintain interpretability and avoid overfitting. The key predictors were ***EraConstructed***, ***Craftsmanship***, and ***TradeValue***.

Figure 10: Confusion matrix classification model



EVALUATION - classification model

Metric	Value	What it Means
Accuracy	0.89	89% overall correct predictions
Precision	0.66	Of all predicted restorations, only 66% were correct
Recall	0.99	Almost all actual restorations were correctly found
F1 Score	0.79	Harmonic mean of precision & recall – solid
Misclassification	0.11	11% of predictions were wrong

Based on the metrics, our model achieved an accuracy of 89%, a recall of 99%, and an F1 Score of 0.79, all of which meet or exceed the success criteria. However, the precision of 66% falls below the 85% target, indicating room for improvement in reducing false positives. This trade-off is acceptable as catching all at-risk properties is more important than avoiding some false alarms. The misclassification rate is 11%, which stays within the acceptable range (≤ 15%). The key indicators influencing restoration needs are primarily EraConstructed and TradeValue, with older and lower-valued dwellings being far more likely to require restoration. Compared to our benchmark model that always predicts "No restoration" (84.1% accuracy but 0% recall), our model clearly performs better by identifying nearly all true cases.

PERSONAL CONTRIBUTION

	Student name	Contribution
1	Georgi Chitarliev	Business understanding • Define both regression and classification objectives. • State the business goals and the data mining goals
2	Silvia Popova	Model • Build a linear regression model to predict TradeValue - Select relevant predictors based on exploratory analysis • Build a decision tree classifier to predict NeedsRestoration - Visualize the decision tree structure and explain each split Overall document structure check
3	Chavdar Tsvetkov	Understanding • For regression: Use MAE/MSE to evaluate performance. • For classification: Use confusion matrix metrics to evaluate performance • What are the most important factors influencing trade values? • What are the key indicator of restoration needs?
4	Simeon Atanasov	Data preparation • Clean and preprocess the dataset. • Engineer new features if necessary • Ensure all predictors are in the correct format for modeling.
5	Dimitar Petrov	Data Understanding • Explore relationships between features and both target variables (TradeValue and NeedsRestoration). Include relevant visualizations (e.g., scatterplots, boxplots) to highlight trends.

As a team, we learned how each iteration built upon the previous, allowing us to progressively refine our approach - from basic data cleaning and benchmark models in Iteration 1, to structured rule-based classification in Iteration 2, and more advanced regression and classification modeling in Iteration 3. The data preparation techniques we developed early on, such as filling missing values and identifying outliers, directly improved the accuracy and reliability of our later models. We improved our collaboration by dividing tasks based on individual strengths, supporting one another with technical challenges, and integrating feedback throughout the project. While our individual reflections in the following chapter highlight personal growth, this shared journey helped us balance interpretability with complexity, and taught us how to work more effectively as a team in data science.

Personal lessons learned

Each of the team members will reflect on their personal learning outcomes:

1. What went well in the iterations?
2. What challenges did you face, and how did you overcome them?
3. What would you do differently in future projects?

Student name: Georgi Chitarliev

1. Cooperation within the team was solid, therefore efficiency was high. Everybody knew their tasks and did them on time. Help was allocated wherever and whenever it was needed. Overall, I would confidently say that everything went well.
2. The main problem came in the face of the 4-page limit per iteration. The work we had to do itself was not the biggest challenge, but structuring the text was what we found to be the hardest.
3. As mentioned, since everything went smoothly in my view within the team, I see this as a strong indication that no alterations to my behavior/work need to be undergone.

Student name: Simeon Atanasov

1. Cooperation, Communication and Work ethic were aligned with the expectation of each one of the team members, therefore everybody knew their task, the deadline of the tasks he/she was assigned to and most importantly we kept the professional environment.
2. The one and only challenge which we faced during the three iterations was fitting the information on maximum 4 pages for each one of the iterations. This was quite challenging, because we wanted to add a lot of information.
3. Everything went smoothly so there is nothing I would do differently in future projects.

Student name: Chavdar Tsvetkov

1. Team alignment regarding cooperation, communication, and work ethic exceeded my expectations - every member was clear about their allocated duties, deadlines, and responsibilities. This clarity helped to preserve a professional and effective atmosphere all during the project.
2. Following the four-page limit for every iteration presented the main difficulty we faced during the three iterations. We had significant material we wanted to incorporate, therefore compressing it while maintaining important points required extra work. This proved challenging.
3. The project went generally without major hicc-ups. Based on this experience, there is nothing I would change in the next projects considering the good result and effective procedure.

Student name: Dimitar Petrov

1. Throughout the project each team member was clear about their roles and responsibilities, and we maintained a professional environment while working. Communication was good, meetings were arranged in advance and regularly and everyone contributed to the report.
2. The main challenge that we faced was the four-page limit per iteration. We had to decrease the amount of information in the report which made it difficult to condense our work while retaining all critical points. To overcome this, we focused on prioritizing the most essential information but without losing the core.
3. In my opinion the project went well, we had a structured plan, regular feedback and set up a goal that we fulfilled. Given these positive outcomes, I wouldn’t change anything in our approach.

Student name: Silvia Popova

1. Throughout the iterations, our team gradually improved both the model performance and our data understanding. Because of our collaboration and the adaptation of our approach based on feedback, we managed to develop more accurate models and understand the CRISP-DM cycle better.
2. One challenge we encountered was balancing model complexity with explainability, especially in the beginning of the course when we weren’t sure what complexity is required from us for each Iteration. We overcame this by testing multiple versions for the model, comparing their performance, and choosing the model that offered both accuracy and clarity, but also following more strictly the rubric.
3. I wouldn’t change much, as I’m happy with how the project went overall. However, I would make sure to read and fully understand the rubric from the very beginning instead of jumping into experimentation before being completely clear on what’s expected.