DTU

DTU Compute

02450 Introduction to Machine Learning and Data Mining (Spring 2025)

Danish Residential Housing Prices

Freature Extraction and Visualization

s236115 - Dimitar Ilev

Table of Contents

1 Introduction **1**

1.1 Literature review 1

1.2 Methodology 1

2 Data Attributes **2**

2.1 Irregularities 3

2.2 Summary statistics 3

3 Data Analysis **4**

3.1 Extreme Values and Outliers 4

3.2 Distribution 6

3.3 Correlation 7

3.4 PCA Analysis 8

4 Discussion **10**

5 Collaboration **11**

Bibliography **12**

A Appendix **13**

1 Introduction

Housing is a crucial component of any human's life. It is a significant economic driver, especially in Denmark, where access to affordable housing is essential for societal well-being. Despite the country's economic stability, the housing market experiences regular fluctuations and price increases.

This steady price increase, particularly in major cities like Copenhagen, Odense, and Aarhus, poses challenges for young people and first-time buyers looking to enter the housing market. As students in Denmark contemplating future homeownership, analyzing housing price dynamics and their long-term implications is of strong personal interest. The "Danish Residential Housing Prices 1992-2024" [1] contains approximately 1.5 million residential property sales in Denmark from 1992 to 2024. This makes it ideal for detailed analysis of market development over the past three decades and a discussion of future outlook.

This project aims to analyze housing market trends, the impact of house features on sales prices, and overall market dynamics in recent history (past 30 years).

1.1 Literature review

Comparing the average purchase and square meter price between properties in the capital region, larger cities, smaller towns and rural areas [2] it turns out that it decreases with decreasing city size and it is the lowest in the rural Danish areas. In 2021 the price per square meter of a single-family home was DKK 41,800 in the capital area and DKK 9.400 in the rural areas. The apartments of 80-90 sqm in the capital region match the price of three single family houses of 140-150 sqm on the country side. The research remarks that newer build homes on the country side after the year 2000 do have a much higher pricing on the market. Overall, the paper concludes that since 2013 to 2021 the housing market has increased in pricing both in the cities as well as on the country side.

In mid-2023, residential property prices declined sharply, with the House Price Index recording a year-on-year drop of 6.16% (9.38% inflation-adjusted) [3]. However, recent data suggests that the market is recovering, with a 1.3% annual increase (0.3% inflation-adjusted) recorded in Q1 2024. Experts attribute this rebound to stable household finances, rising salaries, and recent interest rate cuts. The Danish Ministry of Economic Affairs forecasts further price increases of 3.2% in 2024 and 3.0% in 2025, reinforcing the need for a deeper understanding of market trends.

1.2 Methodology

Regression analysis will be performed to **predict housing prices** based on attributes such as *square meters*, *number of rooms*, *year*, *zip code*, and *region*. **Linear regression** will be used as a baseline model to capture linear correlations between variables.

The goal is to predict the total purchase price of a house using these characteristics and understanding how these factors contribute to a positive or negative price development.

The other objective is **to predict the price per square meter**, which can provide information on regional pricing trends and property valuation. **Random Forest Regressor** will be implemented to better capture nonlinearity. Using this information, the model will generate price estimates to aid in market analysis and comparative analysis between properties. **XGBoost** is a gradient boosting method that works great with large data sets and will be a key factor in model performance.

Classification analysis will be based on housing attributes such as *square meters*, *number of rooms*, *build year*, and *region*. Converting the *percentage change between the offer and the purchase price* into a binary class, where a price increase or no change is marked as 1, and a price decrease is marked as 0. Using **logistic regression** as an initial starting point, the model can predict whether a property will sell above or below the listed market value. **Random Forrest Classifier** is a great approach to handle any imbalances between the different classes of attributes. Also here using **XGBoost** or **CatBoost** will be beneficial for a performance increase of the classifier.

Data transformation is necessary to increase precision and reduce classification error. One is transforming the transaction *date* into separate features such as year or quarter. **Normalizing** numeric features such as *purchase price* and *price per sqm*. Also using **log transformation** to highly skewed variables to stabilize the variance. Convert nominal variables such as *area*, *house type*, *region* and *sales type* into numerical form using **one hot encoding** or label encoding. Depending on frequency, removerties with extreme values in *year built*, *purchase price*, or *sqm price* to ensure betterbetteruality.

The objective of this project is, trough the mentioned methods, to analyze housing market trends, the impact of house features on sales prices, and overall market dynamics in recent history.

2 Data Attributes

Below is the detailed description and explanation of all the dataset[1] attributes. These being subcategorized and in the order: attribute name, type, followed by a short description.

Transactional Details:

Date: (Discrete/Interval) yyyy-mm-dd. **Quarter:** (Discrete/Ordinal) Q1-Q4.
House ID: (Discrete/Nominal) Unique identifier. **Sales Type:** (Nominal/Nominal) Sale type (*regular_sale, auction*). **Purchase Price:** (Discrete/Interval) Price in DKK.

Property Characteristics:

House Type: (Nominal/Nominal) Specifies the type of property (*Villa, Farm, Apartment, etc.*). **Year Built:** (Discrete/Interval) Construction year of the property year 1000 to 2024. **Number of Rooms:** (Discrete/Interval) Total number of rooms in the property. **Square Meters (Sqm):** (Discrete/Interval) The total area of the house in square meters. **Price per Sqm:** (Discrete/Ratio) The price per square meter of the property.

Location Details:

City: (Nominal/Nominal) The city where the property is located. **Area:** (Nominal/Nominal) Geographic classification of the location. **Region:** (Nominal/Nominal) Broader regional classification (*Zealand, Jutland, etc.*).

Economic Indicators:

% Interest Rate: (Continuous/Ratio) Danish nominal interest rate. **% Annual Inflation Rate:** (Continuous/Ratio) Danish annual inflation rate. **Yield on Mortgage Credit Bonds (%):** (Continuous/Ratio) Mortgage bond yield (30-year).

2.1 Irregularities

The *date* attribute is converted into a pandas date time object so it can be used for time series plotting. The attribute *sales_type* contains 19 values marked as '-', therefore the type is recategorized as *other_sale*. There are 4 attributes containing missing/ NaN values.

The two features *sqm* and *sqm_price* both miss two values in relation to the same two properties. This irregularity is not crusial as the data could have been forgotten or is simply not available to the public at the time of creation of the dataset. To keep the integrity of the data, the two rows were removed from the dataset which does not impose any risk to disturb the overall analysis due to its information richness.

The two features *dk_ann_infl_rate%* and *yield_on_mortgage_credit_bonds%* both are missing 1193 values in the period of 01.10.2024 until 26.10.2024, this being the last month before the extraction of the dataset. This can be explained as the information might not have been available to the public at the time of the table creation since it contains such recent data. The missing values can be found available on the internet. The *anual inflation rate* being at *4.1%* [**tradingeconomics_dk_inflation**] and the *fixed yield on mortage credit bonds* laying at *1.6%* [**finansdanmark_mortgage_rates**] for the month of Oktober 2024.

2.2 Summary statistics

Table 1 provides a comprehensive overview of the numerical characteristics of the Danish housing data. After clearing missing values and imputing the data there is a total

of 1.507.906 property purchases. The average property was build in roughly 1955, holds a price of 1.915.469 DKK, which is 2% below the initial asking price, has 4 rooms, 129sqm with 16.345,25 DKK/sqm, gains 1,68% in interest, looses 1.93% due to inflation and yields 4,11% on mortage credit bonds. The 75% quartile of construction years indicates that a significant portion of the houses were built before 1980.

	Year Built	Purchase Price	% Change	No. of Rooms	SQM	SQM Price	ZIP Code	Interest%	Inflation%	Yield%
Count	1,507,906	1,507,906	1,507,906	1,507,906	1,507,906	1,507,906	1,507,906	1,507,906	1,507,906	1,507,906
Mean	1954.94	1,915,469	-2.08	4.37	129.25	16,345.25	5959.60	1.68	1.93	4.11
Std	45.84	1,765,655	4.81	1.65	57.20	13,626.27	2,369.32	2.04	1.64	2.19
Min	1000	250,010	-49	1	26	269.86	1050	0.00	0.25	1.10
25%	1931	800,000	-3	3	89	6,741.57	4000	0.00	0.79	2.12
50%	1965	1,400,000	0	4	123	12,006.58	6000	0.75	1.85	4.34
75%	1980	2,450,000	0	5	160	21,317.83	8260	3.25	2.34	5.50
Max	2024	46,800,000	49	15	997	75,000	9990	9.50	7.70	10.14

Table 1: Updated Statistical Summary of all numerical housing data. *Shortened column names: % Change (%_change_between_offer_and_purchase), Interest% (nom_interest_rate%), Inflation% (dk_ann_infl_rate%), Yield% (yield_on_mortgage_credit_bonds%).*

3 Data Analysis

3.1 Extreme Values and Outliers

This segment will focus on the most relevant findings from this analysis. It is important to mention that not all extreme values are equivalent to an outlier. As there is a good reason why a property could be much more expensive or why it has been purchased for a significantly less money than the original offer.

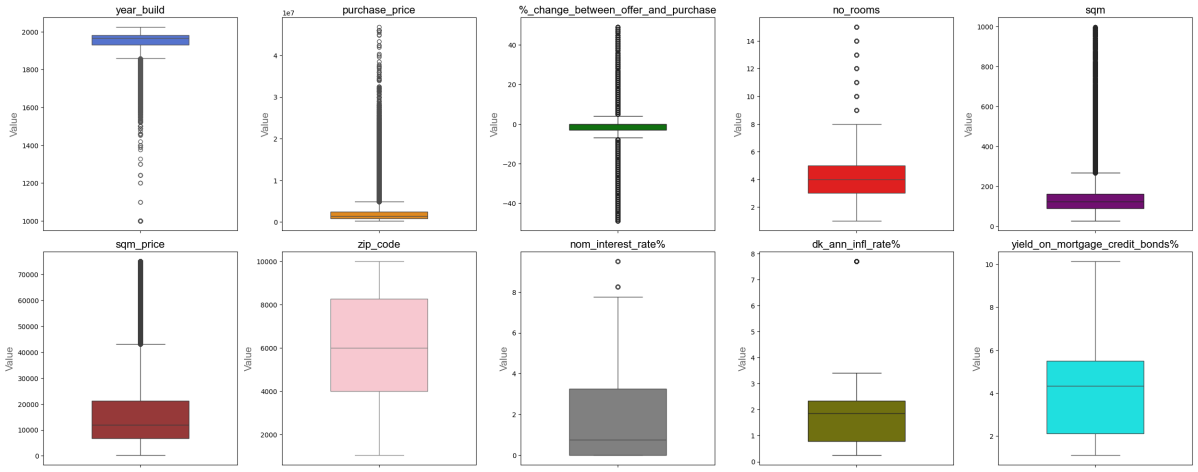


Figure 1: Boxplots

The data distribution is observed using histograms 2a and extreme values are reviewed using box plots 1. A complete table of all numeric attributes can be found in the **??** Appendix. The **Interquatile Range** (IQR) method and **Z-Score** method are both used for the targeted observation and analysis of extreme values and outliers.

Figure 1 provides an overview of the attributes with the most outliers. It shows that *%_Change* has a lot of extreme values outside the upper and lower ends. *Year_build*

has many outliers at the lower end, suggesting that there are a few very old properties in the data. *Purchase_price*, *sqm* and *sqm_price* all have many outliers at the upper end, suggesting that there are also a few rather expensive places in the dataset, these being farms, villas and luxury apartments.

Attributes	IQR Method		Z-Score Method	
	Outliers (%)	Bounds	Outliers (%)	Outliers (Count)
year_build	2.61%	[1857.50, 2053.50]	1.16%	17455
purchase_price	5.49%	[-1675000.00, 4925000.00]	1.75%	26390
%_change_between_offer_and_purchase	10.56%	[-7.50, 4.50]	2.11%	31889
no_rooms	1.79%	[0.00, 8.00]	0.85%	12890
sqm	2.05%	[-17.50, 266.50]	1.04%	15639
sqm_price	6.04%	[-15122.81, 43182.22]	2.23%	33602
zip_code	0.00%	[-2390.00, 14650.00]	0.00%	0
nom_interest_rate%	2.02%	[-4.88, 8.12]	2.02%	30528
dk_ann_infl_rate%	5.43%	[-1.53, 4.66]	5.43%	81814
yield_on_mortgage_credit_bonds%	0.00%	[-2.95, 10.57]	0.00%	0

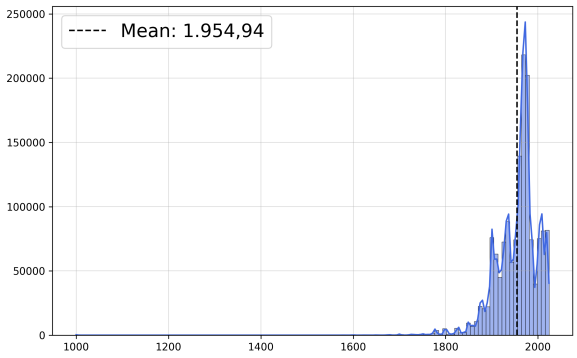
Table 2: Outlier Analysis Summary

To further analyse these values Table 2 showcases that according to IQR the *%_Change* has the highest number of extreme values at 10.56%, followed by *sqm_price* at 6.04%, *purchase_price* at 5.49%, and *dk_ann_infl_rate%* at 5.43%. The Z-score detects way less outliers due to the non-normal distribution of the data. The attributes zip code and %yield have zero values as they have a uniform distribution.

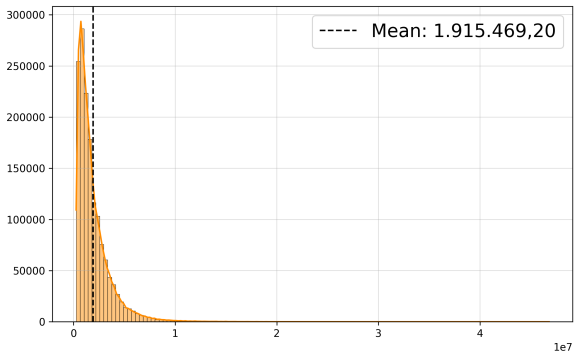
Purchase_prise, *sqm*, and *sqm_price* have negative lower bounds, this can be explained by the high skewness of the data, since they have higher percentages of outliers explained by luxury properties or special market conditions. Even though, *% change* has the highest IQR percentage of outliers, the annual inflation rate % has the most with 81.814 outliers according its Z-score. This again can be explained by the distribution of the data.

**Supplementary visualizations in a tabular overview of distribution, outliers, and regions can be found in the appendix A.*

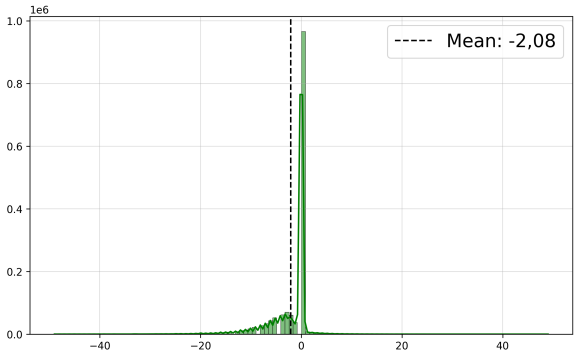
3.2 Distribution



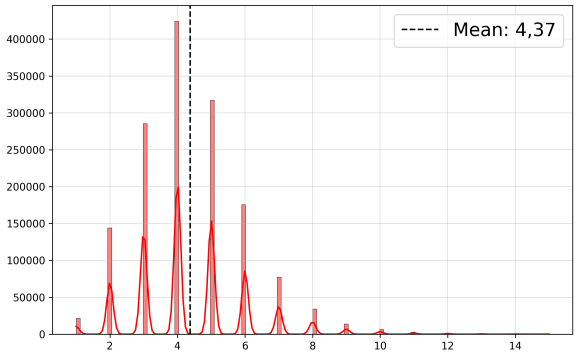
Year Build: Non-normal distribution with strong increase after year 1800 and a peak around 1965 with a long tail to ancient properties, registered until the year 1000



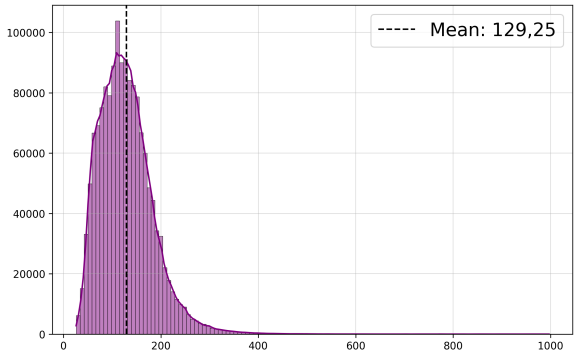
Purchase Price: Non-normal distribution, positively skewed to the right due to some expensive properties up to DKK 46 Million, the majority priced below DKK 2 Million.



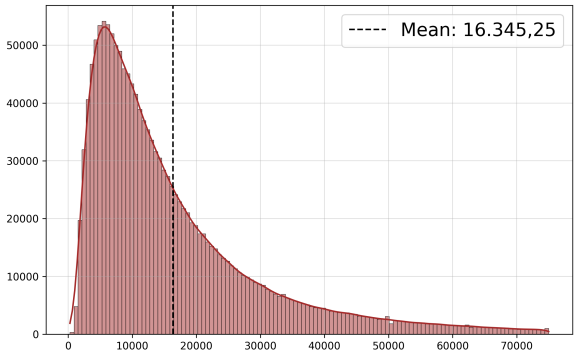
% Change: Non-normal distribution and has a peak around zero, suggesting that most properties sold close to the offer price.



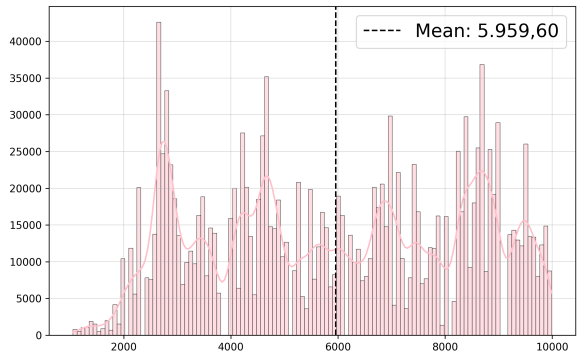
Number of Rooms: Discrete distribution with peaks around 3 to 5 rooms reflecting normal properties.



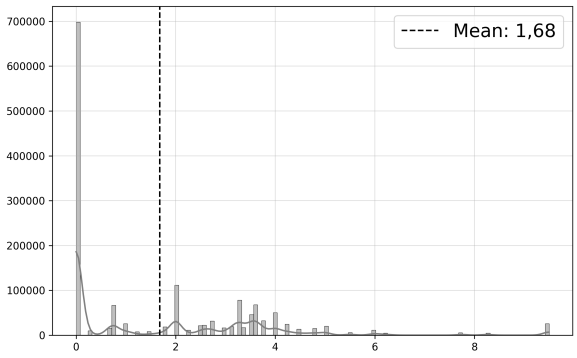
Square Meters: Non-normal distribution and skewed to the right, most properties being at around 130sqm.



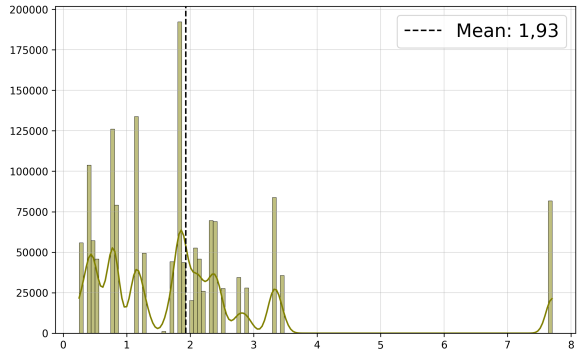
Price/sqm: Non-normal distribution and skewed to the right with a mean around DKK 16.345.



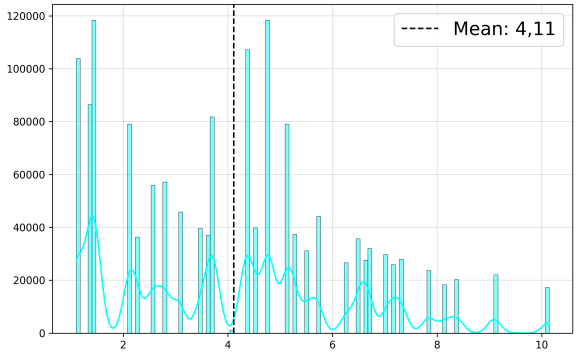
Zip Code: Non-normal distribution with peaks and valleys suggesting popular areas, such as Kopenahgen with 2000.



Nominal Interest Rate: Non-normal distribution and heavily skewed to the right, with most being around zero.



Annual Inflation Rate: Non-normal distribution with some rates being more common than others.



Mortgage Yield: Non-normal distribution with peaks around 1% and 4,5% and reaching up to 10%.

3.3 Correlation

According to the correlation matrix in Figure 7 there is a strong positive correlation between purchase price and square meter price of 0.85. This means that that the property becomes more expensive with higher per meter price. The number of rooms and the total square meter area also show a high correlation score of 0.81. Suggesting that larger homes generally have more rooms. Additionally, the nominal interest rate and yield percentage are strongly correlated with a score of 0.89, meaning higher nominal interest rates are associated with higher mortgage yields. A moderate positive correlation is observed between the annual inflation rate and the nominal interest rate at 0.51, implying that inflation impacts interest rates to some extent. There is a positive correlation between the annual inflation rate and the yield on credit mortage bonds of 0.62, suggesting that inflation also influences the yield of the properties.

The purchase price and zip code correlation of -0.27 may suggest that lower property prices are concentrated in specific areas. Zip code and sqm price has the lowest correlation at -0.33. Square meter area and square meter price are negatively correlated with a score of -0.32, meaning that larger properties tend to overall have a lower price

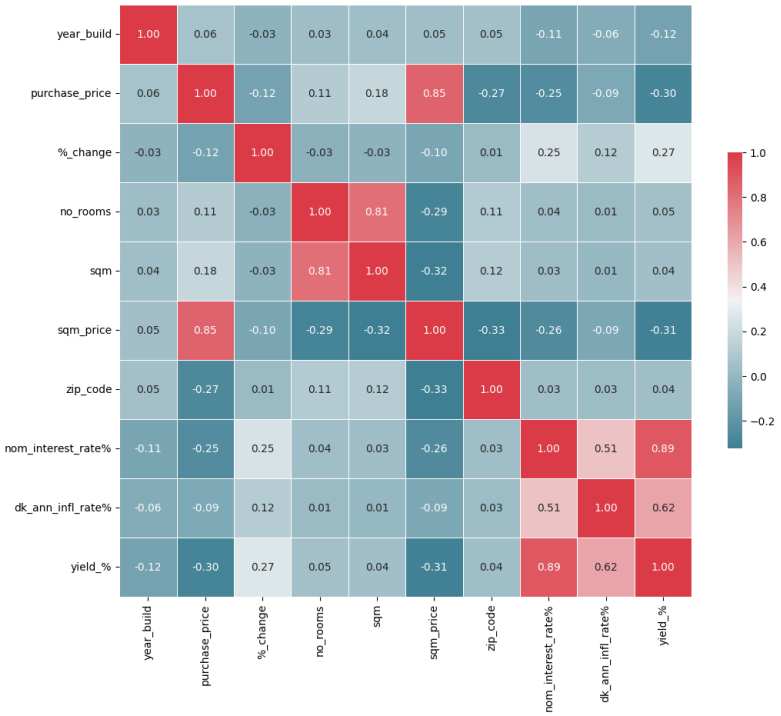


Figure 7: Spearman Correlation Matrix

per square meter. Sqm price is also negatively correlated with the number of rooms at -0.29. There is a negative correlation between yield percentage and purchase price of -0.30. Iandicating that properties with lower purchase prices tend to have higher mortgage yields. Some attributes show very weak correlations, such as year build, which has little to no relationship with most other features. The percentage change between offer and purchase price also does not exhibit strong correlations with other variables.

3.4 PCA Analysis

Figure 8 represents a heatmap of the correlation between the original features and the principal components. Strong positive correlations are indicated with darker red and strong negative correlations are indicated in darker blue.

Year_build at 0.73 and *%_Change* at 0.64 are strongly correlated with *PC4*, meaning this principal component is heavily influenced by these features. *Purchase_price* and *sqm* have high loadings on PC2 and PC3, indicating that these PCs capture variations in property pricing and size. PC8 has a large influence of 0.74 on the *number of rooms* while it has the lowest influence of -0.63 on the *sqm*. The inflation rate has the highest score of 0.78 on PC5. Nominal interest also has a relatively low value of -0.56 on PC9.

Figure 9 presents a plot showing the variance explained by each of the principal components. The cumulative variance curve (orange line) shows how much of the total variance is captured by compbining the components. The threshold lines for 90% and 95% are there to distinguish the amount of explained variance. Six principal components are necessary to explain close to 90% of the variance in the dataset and seven

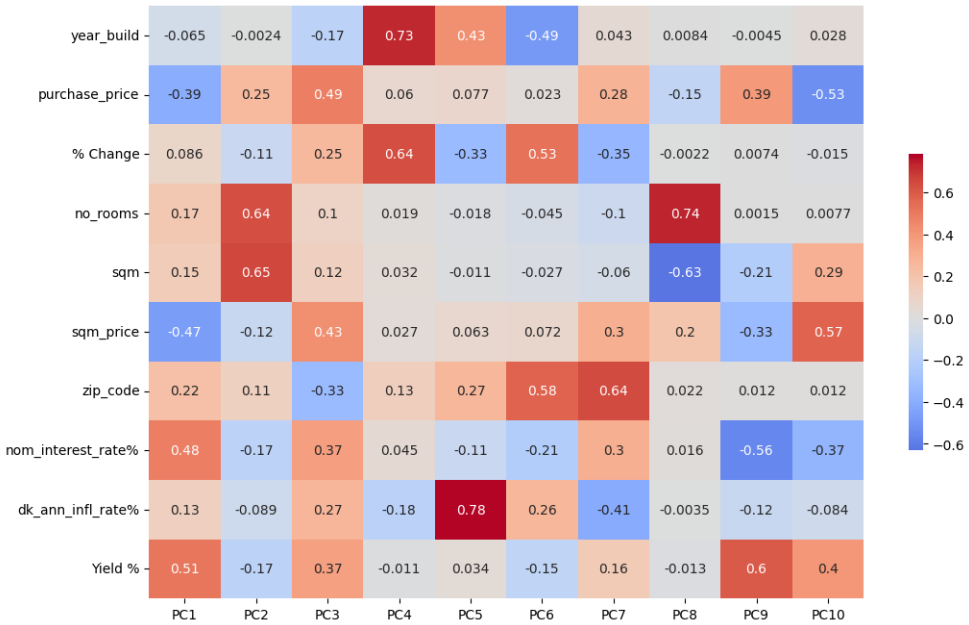


Figure 8: Principal Directions and Attributes

are necessary to explain more than 95% of the total variance. Dimensionality reduction to these components will retain most of the original information while reducing complexity.

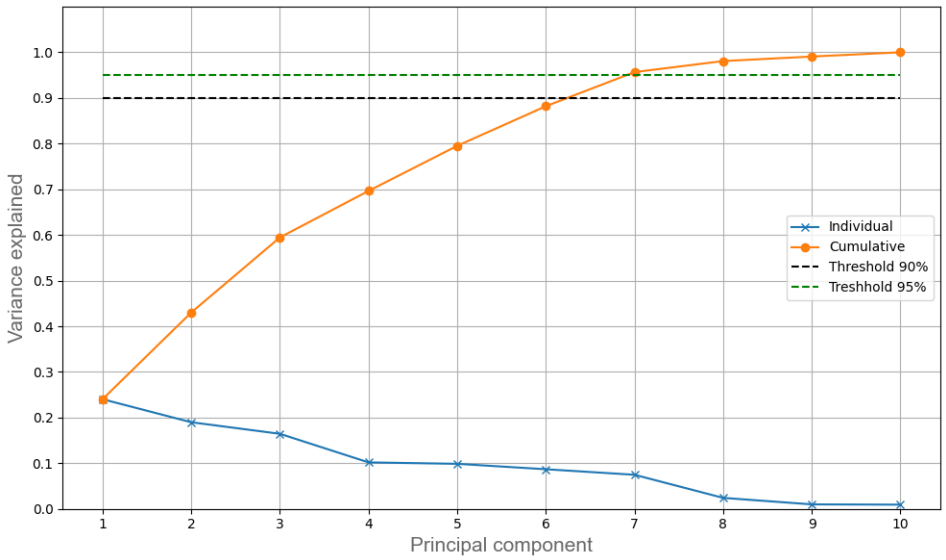


Figure 9: Explained Variance

Figure 10 presents a 2D scatter plot of the first seven principal components responsible for more than 95% of the total variance of the dataset. To better comprehend the distributions the coloring is centered around the predictor variable *purchase_price*. From the grapfik it becomes clear that PC1 has a negative and PC2 has a positive relation and PC3 has a very strong positive influence from that attribute. This aligns the previous results on the principal directions.

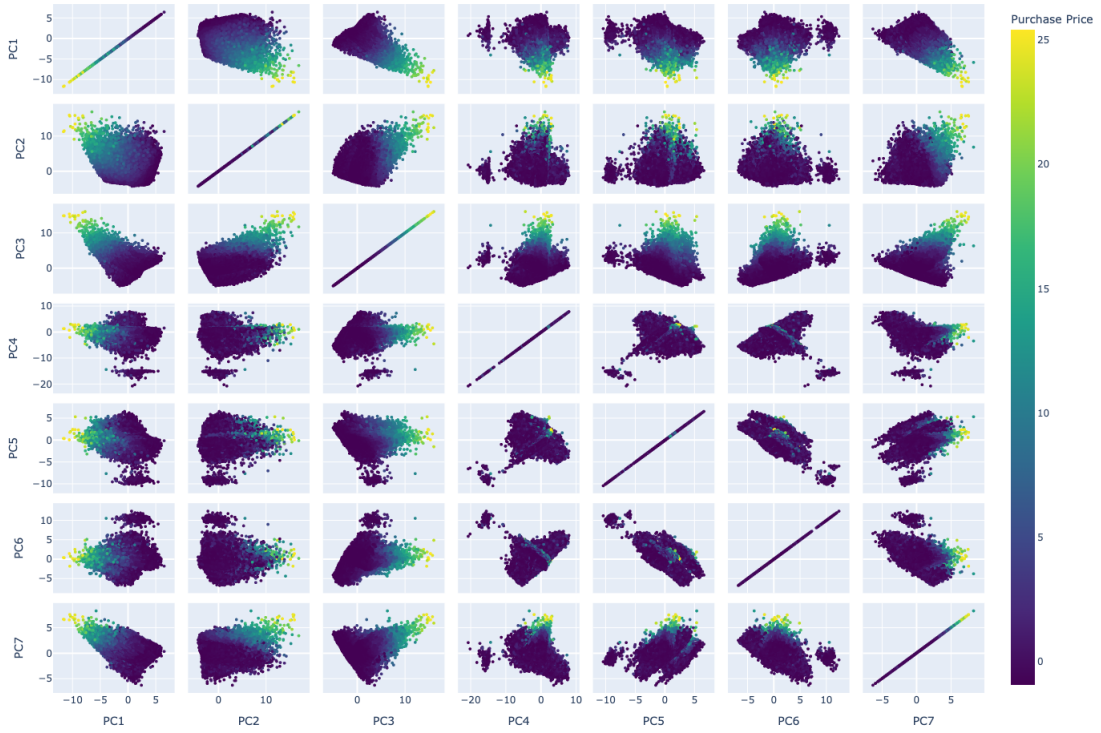


Figure 10: Scatter of first 7 PCs color coded based on purchase price attribute

4 Discussion

This project has provided a comprehensive analysis of the Danish housing market between 1992 and 2024, analyzing key features, outliers, distribution, correlation and dimensionality reduction.

The property size (sqm) and location (zip code) are strong determinants of the purchase price, suggesting that these factors should be central to any predictive model.While economic indicators such as interest rates and inflation exhibit correlations with property prices, these relationships appear to be more complex and may require nonlinear modeling.

Based on the visualization, the first machine learning aim is feasible. The correlation shows a robust 0.85 between square meter and purchase price, indicating that a linear regression model might be a good starting point for predicting the purchase price of a property. The model may require robust regression techniques or data transformations to prevent those outliers from unduly influencing the model, however, it should be considered feasible.

The classification model may not be as feasible. Visualizations and and further statistics show there aren not any robust correlations between nominal rates, or ann infl rates. These features seem to be weak indicators and may not provide the desired predictive ability.

5 Collaboration

Since this project was conducted alone, *Open AI* was used as a sparing partner for code optimization and suggestion of possible changes. Especially for the improved look of the individual plots. Additionally, *Grammarly* has been used to correct grammatical errors made in the writing process of this document. The main sources for troubleshooting have been Google such as *StackOverflow*, *Scikit-learn* and similar.

Bibliography

[1] Martin Frederiksen. *Danish Residential Housing Prices 1992-2024*. Kaggle. Accessed: February 22, 2025. 2024. URL: https://www.kaggle.com/datasets/martinfrederiksen/danish-residential-housing-prices-1992-2024/data.

[2] Statistics Denmark. *Land og by: Hvor forskellig er boligprisen?* Accessed: February 23, 2025. Feb. 2025. URL: https://www.dst.dk/da/Statistik/nyheder-analyser-publ/Analyser/48868-land-og-by-hvor-forskellig-er-boligprisen.

[3] Global Property Guide. *Denmark Price History*. Accessed: February 23, 2025. Dec. 2024. URL: https://www.globalpropertyguide.com/europe/denmark/price-history.

A Appendix

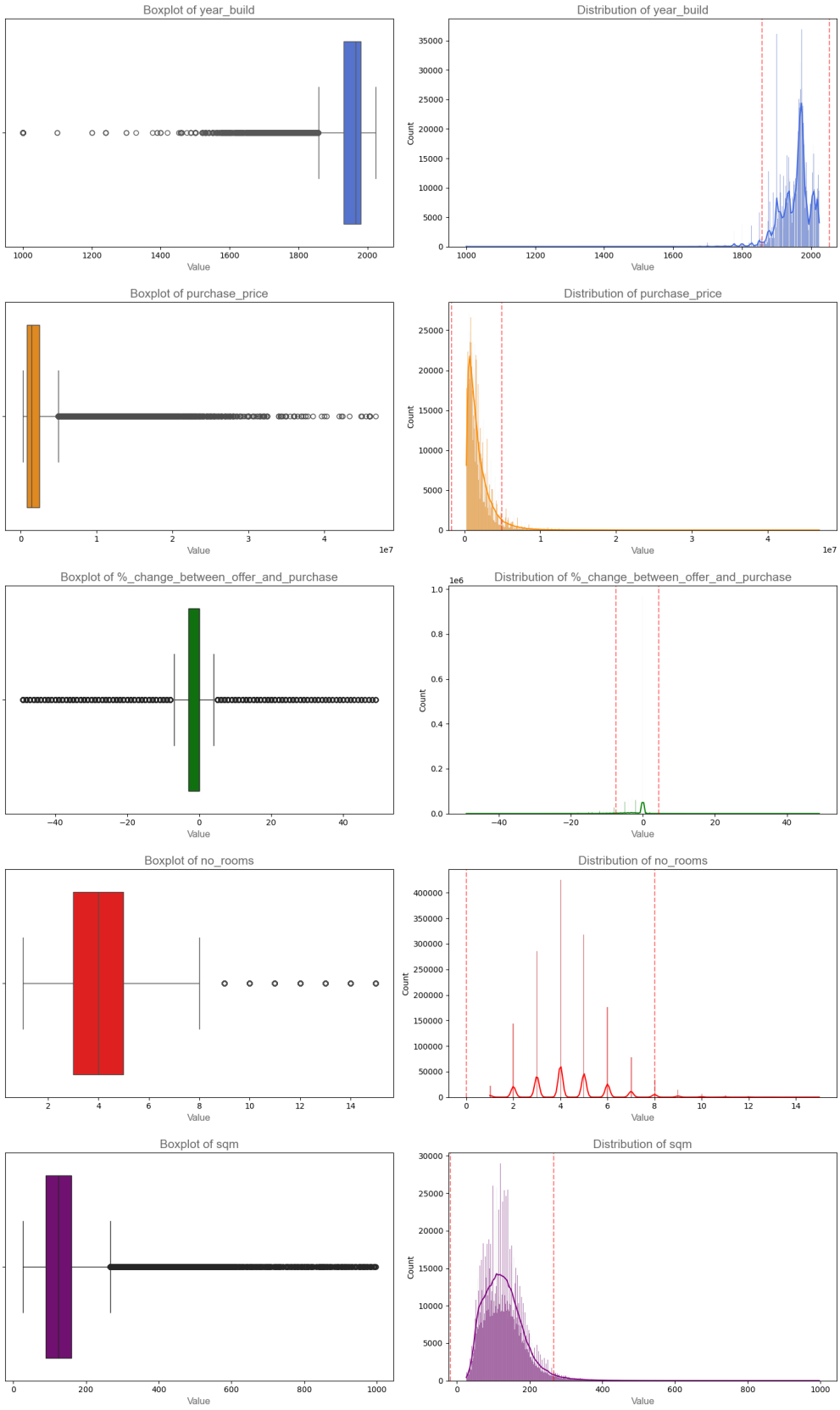


Figure 11: Outliers Analysis Part 1

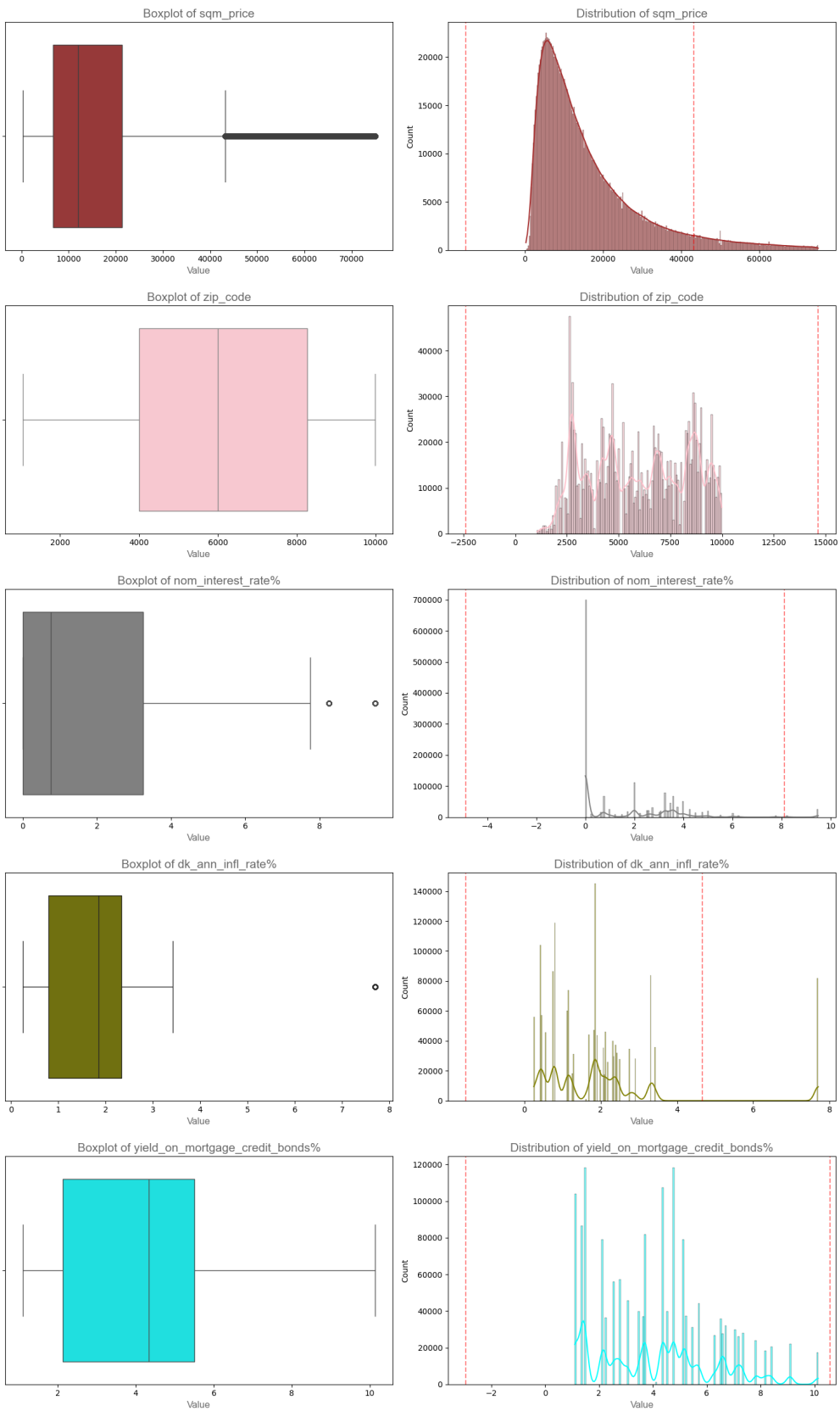


Figure 12: Outliers Analysis Part 2

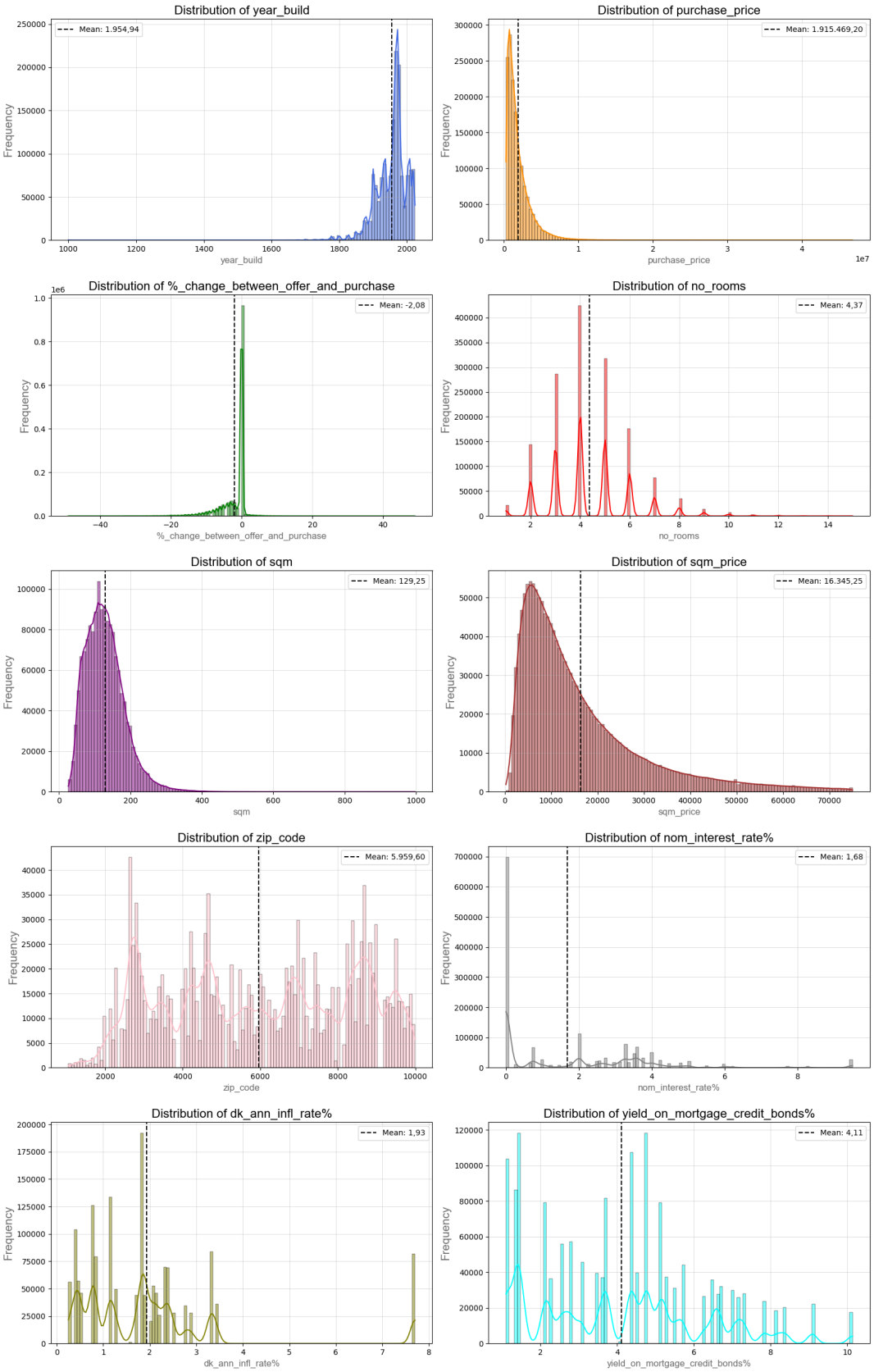


Figure 13: Distribution Analysis

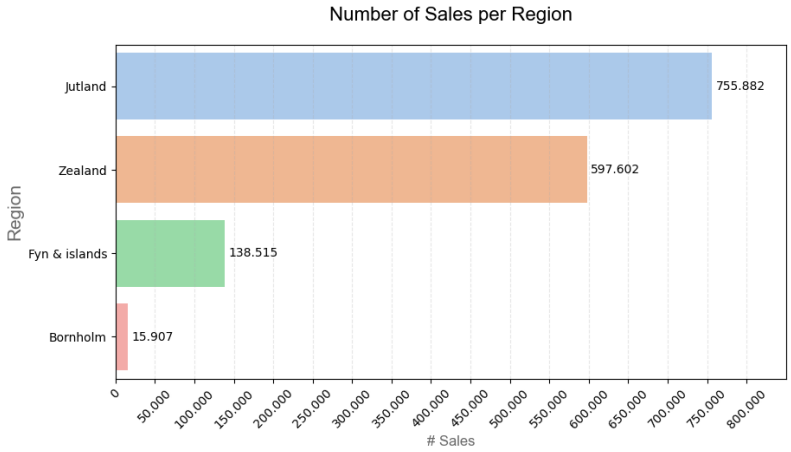


Figure 14: Sales per Region

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
year_build	-0.065	-0.002	-0.174	0.731	0.431	-0.492	0.043	0.008	-0.005	0.028
purchase_price	-0.395	0.247	0.489	0.060	0.077	0.023	0.284	-0.146	0.385	-0.532
%_change	0.086	-0.107	0.252	0.638	-0.329	0.528	-0.350	-0.002	0.007	-0.015
no_rooms	0.169	0.638	0.102	0.019	-0.018	-0.045	-0.101	0.735	0.002	0.008
sqm	0.148	0.653	0.124	0.032	-0.011	-0.027	-0.060	-0.632	-0.215	0.293
sqm_price	-0.474	-0.121	0.431	0.027	0.063	0.072	0.298	0.195	-0.330	0.574
zip_code	0.216	0.115	-0.332	0.133	0.266	0.576	0.640	0.022	0.012	0.012
interest_rate%	0.481	-0.170	0.369	0.045	-0.111	-0.212	0.305	0.016	-0.564	-0.366
infl_rate%	0.128	-0.089	0.272	-0.180	0.782	0.265	-0.407	-0.003	-0.119	-0.084
yield%	0.513	-0.171	0.368	-0.011	0.034	-0.153	0.156	-0.013	0.604	0.399

Table 3: PCA Loadings and Attributes Correlation

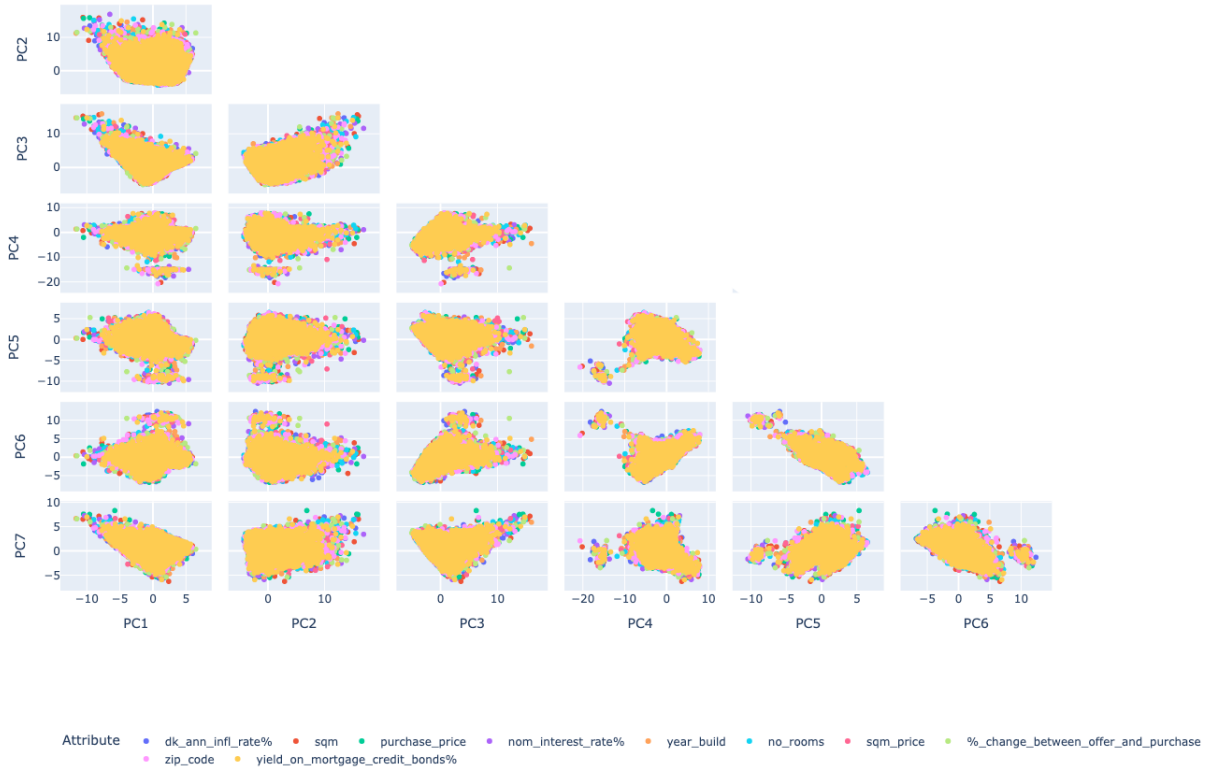


Figure 15: Matrix of principal conponents and attributes

Technical
University of
Denmark

Brovej, Building 118
2800 Kgs. Lyngby
Tlf. 4525 1700

www.byg.dtu.dk