Project 1: Data Feature Extraction and Visualization

Group Members:
Dikaiopoulos Michail (s242816)
Papantzikos Dimitrios (s242798)
Tzimoulias Konstantinos (s242796)

September 2024

Contribution Table

Section	s242816	s242798	s242796
Data Description	40%	30%	30%
Attribute Analysis	30%	40%	30%
Data Visualization	30%	30%	40%
Discussion	33%	33%	33%
Exam Problems	33%	33%	33%

Table 1: Contribution of group members per section.

Abstract

This report explores **feature extraction and data visualization** techniques using methods learned in the first section of the course. The objective is to gain insights into an **apartment/room rental listings dataset**, preparing for further analysis in later reports.

Contents

1 Data Description

1.1 Overview of the Data

In the lifecycle of real estate assets, their value plays the most significant role. Various organisations in the real estate domain, utilise a big workforce in order to evaluate real estate prices. It has been of great interest to try and "correctly" evaluate these prices in an automated way, with the so called AVMs (Automated Valuation Models), which basically are machine/deep learning models trying to predict a real estate asset price based on its characteristics, but not with great success as it is a really complex task, depending on various external (and sometimes subjective) factors other than its characteristics. We will have our own approach in this problem, hopefully providing some insights into the rental world of apartments and rooms in Copenhagen.

1.2 Source of the Data

After firstly being approved by one of the Teaching Assistants, we scraped **this website**, while respecting web-scraping rules and ethics. Our inspiration was that we wanted to try and address a real world "live" problem and not one that has already been addressed. Since, our data originates from a live source, previous research is not relevant in this dataset. The scraping took place only once around the middle of September, so we only have a sample of one instance of the market.

1.3 Task Definition

Concisely, we hope to address the following in a successful way, defining success as being able to drive actual insights from the data.

- Identify which attributes play the most significant role (and at what volume) for the apartment and room rent prices.
- For regression, we will try to predict the rental price of apartments/rooms using the rest of its characteristics. For classification we will try to predict how long the asset will stay on the website. For the latter, it should be noted that we do not have the information of when an apartment/room was rented, so we will just try to classify the assets that are going to stay longer on the website compared to the ones that will not (threshold could be 6 months).
- Some data transformation has taken place, but on a high-level, we mostly applied (or at least tried to apply) the transformation of common sense.

2 Detailed Attribute Analysis

2.1 Types of Attributes

- **Numeric/Ordinal**:
  rooms (number of rooms), **floor** (floor of the apartment)

- **Continuous/Ratio**:
  monthly_rent (amount in DKK), **prepaid_rent** (upfront amount in DKK), **deposit** (upfront amount in DKK), **days_on_website** (number of days that the asset has stayed on the website as a result of a day difference between creation date of the listing and the scrape date), **total_monthly_rent** (monthly_rent plus monthly_aconto, because sometimes monthly_aconto was 0 as the aconto was included in the montlhy_rent price, so we added them together in order to

use the total rent for further analysis), **monthly_aconto** (monthly utilities in DKK), **size_sqm** (size of asset in square meters)

- **Discrete/Nominal**:
  roommate_friendly (yes/no/not applicable) **availability_in** (buckets of number of months e.g. 1-3 months, 3-6 months, etc. **area** (Copenhagen area e.g. København S), **energy_mark** (e.g. A15, A20, B), **housing_type** (Apartment or Room), **dryer** (yes/no/not applicable), **charging_station** (yes/no/not applicable), **washing_machine** (yes/no/not applicable), **dishwasher** (yes/no/not applicable), **parking** (yes/no/not applicable), **students_only** (yes/no/not applicable), **senior_friendly** (yes/no/not applicable), **elevator** (yes/no/not applicable), **pets_allowed** (yes/no/not applicable), **furnished** (yes/no/not applicable), **balcony_terrace** (yes/no/not applicable), **student_affordable** (Takes values True/False and we created it with the expectation to use it for our classification methods later on. In order to set the threshold for defining if an accommodation is affordable for a student, we took into consideration the average monthly student income, other expenses and general cost of living, as well as typical student expenses for rent and utilities. We gathered our information from Statistics Denmark and adjusted our threshold with findings from our fellow students)

2.2 Data Issues

- **prepaid_rent**: nulls were set to 0 by the assumption that no value in prepaid rent means that there is no **prepaid_rent** needed to be paid, therefore 0
- **energy_mark**: empty values were left as 'none' to form another group (which contains around 47.3% of the rows)

2.3 Summary Statistics

The dataset comprises apartment listings that provide insights into various attributes such as **rent**, **utilities**, **size**, **number of rooms**, **floor level**, **deposits**, **prepaid rent**, **days on the website**, and **total monthly rent**. Most apartments have monthly rents between **9,600** and **15,200**, with a median of **12,300**, although extreme outliers push the maximum rent to **1,758,802** and cause a high standard deviation. These high rents are likely correlated with other factors, such as apartment size, which ranges from **6** to **324** square meters, with larger apartments commanding significantly higher rents. This relationship is also reflected in the number of rooms, where larger apartments with more rooms (the median being **3**) tend to have higher rental prices.

Utilities (monthly **aconto**) remain relatively stable across different listings, suggesting that they may not be strongly tied to apartment size or rent. However, total monthly rent, which includes utilities, follows a similar pattern to rent itself, with a median of **13,200** and outliers that increase the maximum value to **1,760,302**, further demonstrating the impact of high-end properties on the overall dataset.

Deposits and prepaid rent are also notably connected to the rental price, with larger or more expensive apartments demanding higher upfront costs. While the median deposit is **35,700** and prepaid rent is **11,950**, extreme values (up to **4,826,250** for deposits and **3,732,018** for prepaid rent) suggest that more expensive listings have significantly higher initial payments.

Finally, the number of days a listing stays on the website could indicate **market demand** or **pricing strategy**. Listings tend to remain for about **35 days** on average, though more expensive or high-end properties might stay listed longer, as reflected by outliers staying up for over **1,400 days**. This suggests a potential connection between days on the website and pricing, where luxury apartments may have slower turnover due to their higher price points and target audience. Overall, the dataset reveals strong relationships between **rent**, **size**, **deposit**,

prepaid rent, and **total rent**, while **utilities** and the number of **days on the website** may be influenced by other factors like market conditions or pricing strategies.

Attribute	**Min**	**25%**	**50%**	**75%**	**Max**	**Mean**	**Standard Deviation**
monthly_rent	3127.00	9600.00	12300.00	15200.00	1758802.00	24088.24	124477.40
monthly_aconto	0.00	500.00	800.00	1000.00	3099.00	798.38	491.09
size_sqm	6.00	42.00	79.00	98.00	324.00	74.95	40.11
rooms	1.00	1.00	3.00	3.00	8.00	2.54	1.21
floor	0.00	1.00	2.00	3.00	27.00	2.32	2.37
deposit	0.00	23700.00	35700.00	44000.00	4826250.00	63113.57	338212.50
prepaid_rent	0.00	7200.00	11950.00	15000.00	3732018.00	23431.46	140350.80
days_on_website	10.00	19.00	35.00	71.00	1419.00	64.16	94.68
total_monthly_rent	3127.00	10040.00	13200.00	16200.00	1760302.00	24886.62	124515.20

Table 2: Summary statistics for attributes.

3 Data Visualization

3.1 Exploratory Data Visualization

Let's first examine our **total_monthly_rent** attribute, which will be the one we will use in our regression analysis later on. It is fairly obvious that we have some very big outliers that make our boxplot very thin on the left side of figure 1. By examining these outliers, we observe that their existence is completely justifiable, since it is not caused from corrupted values or some false computation but they are just very expensive due to a number of reasons. As a result, we have no reason to remove them from our study.

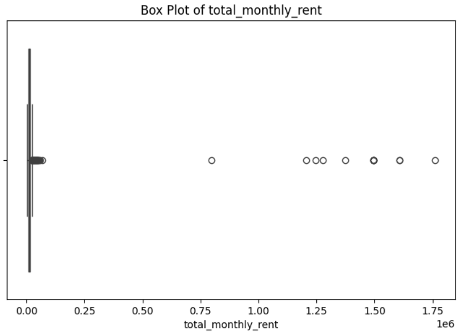

Figure 1: Total Monthly Rent boxplot

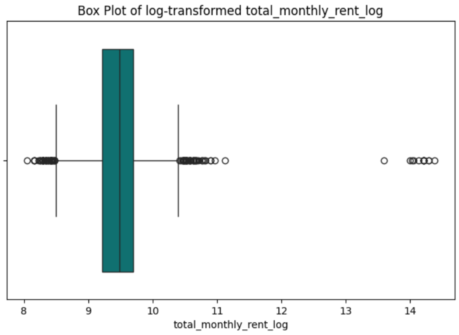

Figure 2: Log-transformed Total Monthly Rent boxplot

It is also important to have an idea about the distribution of our total monthly rent attribute. We already know that, because of the outliers that we saw in the boxplot diagram in figure 1, the distribution is going to appear very skewed if we don't apply a logarithmic transformation in our data, so that is what we are going to do. After applying the transformation, we get the distribution shown in figure 4. It seems that total monthly rent is normally distributed, with the outliers being obvious in the right side of the distribution.

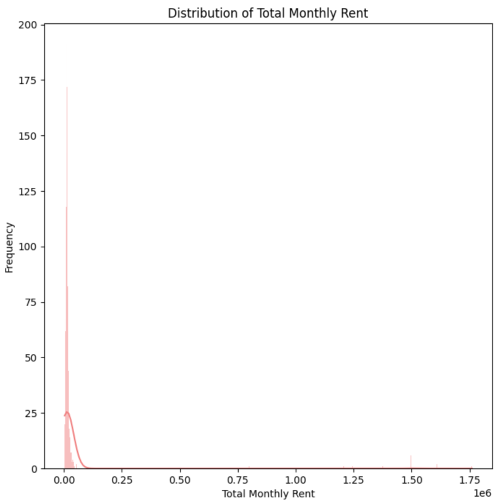

Figure 3: Total Monthly Rent distribution

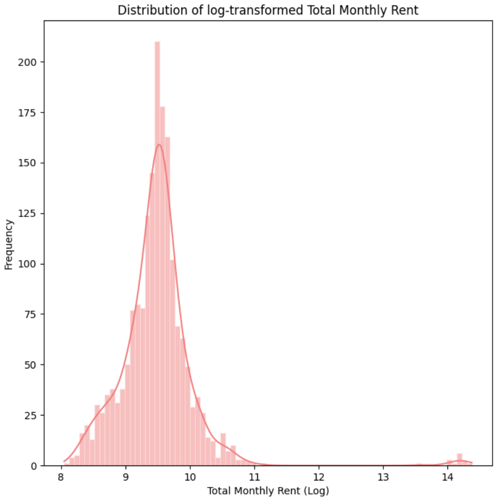

Figure 4: Log-transformed total monthly rent distribution

Let's now check the correlations between our attributes. We have created two correlation heatmaps below, one with the data linearly scaled and one with our data transformed in the logarithmic scale. The reasoning behind that comes again from the distribution of the data. The boxplot in figure 1 reveals a skewed distribution of the monthly rent, an observation we've made for other attributes as well, like deposit and prepaid rent which are also price metrics and strongly related to monthly rent. In these cases a log transformation can help us normalize the distribution and have more useful findings.

Our observations from Figure 6 show us that there is an expected strong correlation between total monthly rent and price metrics like deposit and prepaid rent, as well as a correlation with the size of the apartment and the number of the rooms. Intuitively, we would expect these correlations and we can see that they are much better represented in log-transformed heatmap.

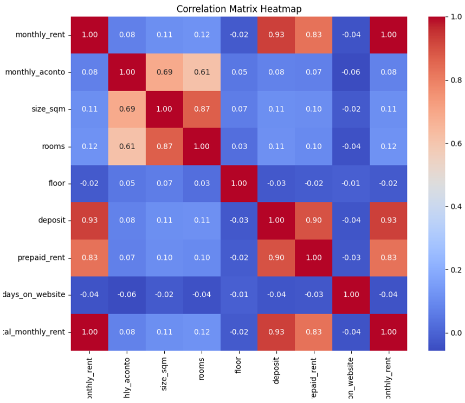

Figure 5: Original Variables

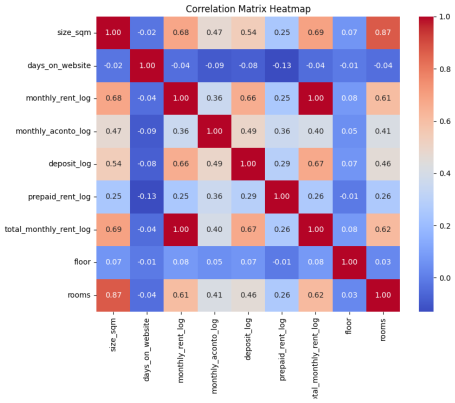

Figure 6: Log Transformed Variables

3.2 Principal Component Analysis (PCA)

Perform PCA on the dataset, and discuss:

- The amount of variation explained by the PCA components.

- Principal directions of the PCA components and their interpretation.
- Data projected onto the principal components.

Our next step is to perform a Principal Component Analysis on our dataset to examine our set in a lower dimensional space, trying to preserve however as much information as possible. In order to do that, we will need to exclude any categorical variables we have, to optimise the performance of our PCA implementation. We understand that there is some information that is lost by excluding these attributes, however there are more appropriate dimensionality reduction methods, like MCA for examining categorical variables as well. We are also aware that since we have attributes with different scales of values, a standardization is necessary to achieve the best possible results.

Our initial number of features is 7, which indicates our current dimensionality space. What we observe is that the number of components that we need to explain 90% of variance in our data is 3. In simple terms, this fact implies that the dataset lies primarily on a 3-dimensional subspace within the higher dimensional space. If we want to be even closer to our initial data variance, we can achieve a 95% similarity using 4 principal components.

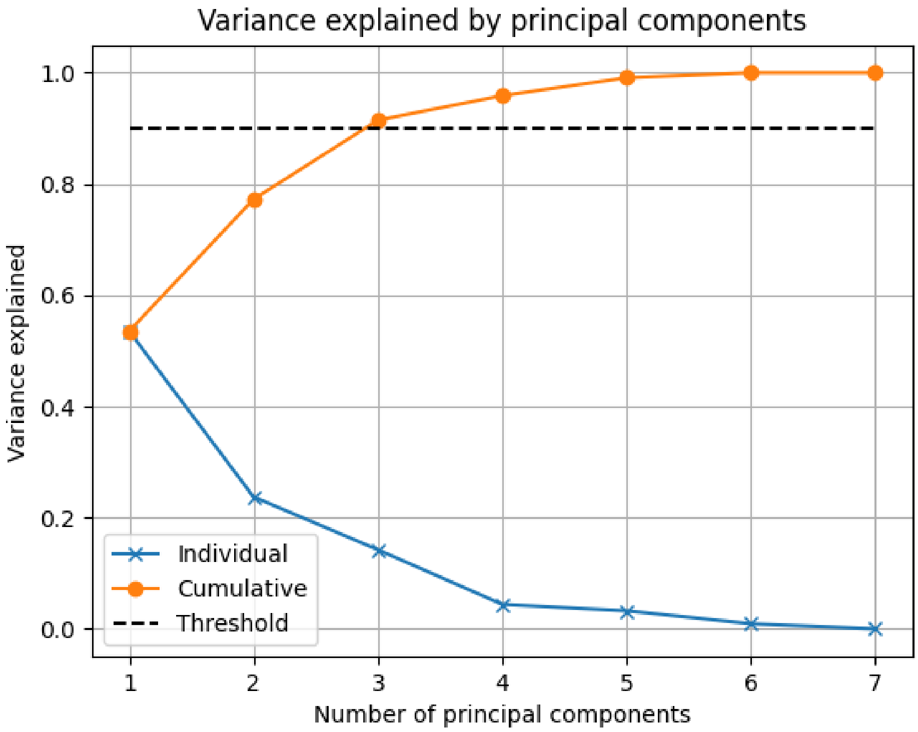

Figure 7: PCA variance explained

Let's now project our data in the 3-dimensional space that we extracted from our principal component analysis. We are labeling our data as affordable for students or not, using the student_affordable attribute on which we are planning to apply our classification techniques later on in the project.n Each of the 3 axes represent a principal component, with x-axis representing the most influential component. In figure 8 we can see there is some kind of clustering of the affordable and the non-affordable, indicating us that our principal component analysis could extract successfully information from our data in a lower dimensional space than the initial one.

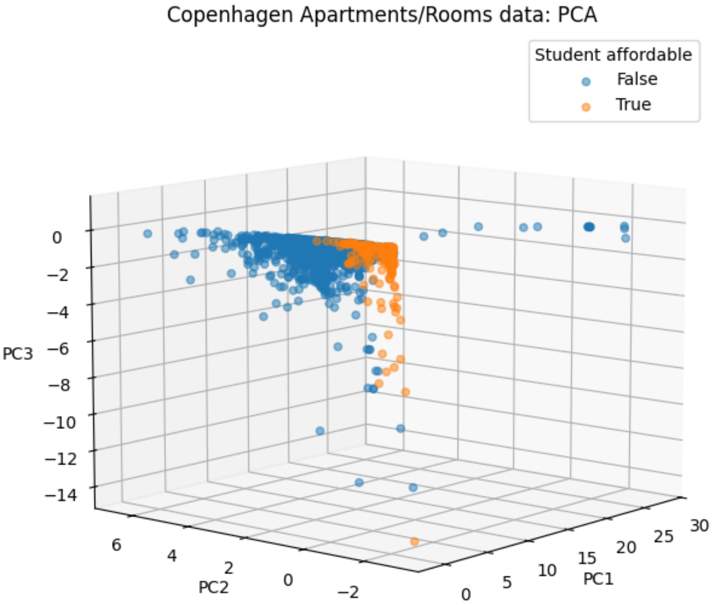

Figure 8: PCA projection of the dataset

4 Discussion

Summarize your key findings from the data exploration:

- The dataset consists of 1,821 entries with 36 columns, representing various features of rental properties. These include details like monthly rent, monthly aconto, move in price, Boligtype (property type), Størrelse (size), and Værelser (rooms). The data covers both numeric and categorical variables, indicating the potential for a mix of regression and classification tasks. Through preprocessing, we handled missing values, outliers, and transformed data types. For instance, (monthly rent), originally stored as an object, was converted to numeric data for analysis. Initial data exploration revealed key relationships between variables. For example, property size (Størrelse) and number of rooms (Værelser) are expected to correlate strongly with rental price (monthly rent), supporting their inclusion as predictors in a machine learning model.

- By visualizing our data, we were able to find out that we have some extreme outliers in attributes like total monthly rent, which however do not indicate any data errors or corrupted values. The skewed nature of our distribution however, forced us to transform logarithmically our data in order to draw better conclusions from their distribution.

- It was also quite obvious that there are certain strong correlations within our data, which makes us optimistic for the accuracy of our regression and classification techniques later on in the project. Since we have a big number of attributes, we focused on the continuous ones and examined whether we can lower the dimensionality space using the pca algorithm. By shrinking the dimensionality space, we can see that the variance of our data can be explained almost perfectly using only 3 components, indicating a simpler underlying data structure. By identifying the underlying relationships between our attributes, we can later extract some unnecessary features and optimise our learning algorithms.

- Overall, The data appears rich with predictive potential, especially for tasks related to rental price estimation or property type classification. The preprocessing has likely improved its quality by addressing missing values, outliers, and standardizing formats. The presence of both numerical and categorical variables supports the feasibility of using various machine learning algorithms, such as regression models for price prediction or clustering techniques for property segmentation or affordability classification.

5 Exam Problems

1. **Question 1**

   Correct: C
   x_1 (time of the day) is ordinal
   x_6 (Traffic lights) is ratio,
   x_7 (Running over) is ratio,
   y (Congestion level) is ordinal

2. **Question 2**

   Correct: A
   It's correct by the definition of p-norm while $p = \infty$ being the maximum of the absolute differences between corresponding elements.

3. **Question 3**

   Correct answer A
   The matrix V contains the principal components, and the diagonal matrix S contains singular values, which correspond to the variance explained by each principal component. The variance explained by each component is proportional to the square of its singular value. The total variance is the sum of the squares of all singular values. These are the first steps we need to do.
   We compute the total variance:

   $$S_{\text{total}} = 13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2 = 670.4$$

   Next, we'll compute the proportion of variance explained by different numbers of principal components. Doing the math on paper:

   $$\text{Variance (first 4 components)} = 13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 = 581.3$$

   Variance explained:

   $$\frac{581.13}{670.4} \approx 0.867$$

   We used (3.18) from the book.

4. **Question 4**

   Correct D
   Solution:

   $$V = \begin{bmatrix} 0.49 & -0.5 & 0.08 & -0.49 & 0.52 \\ 0.58 & 0.23 & -0.01 & 0.71 & 0.33 \\ 0.56 & 0.23 & 0.43 & -0.25 & -0.62 \\ 0.31 & 0.09 & -0.9 & -0.19 & -0.24 \\ -0.06 & 0.8 & 0.03 & -0.41 & 0.43 \end{bmatrix}$$

In PCA, the matrix V contains the eigenvectors (principal components) and tells us how much each original feature contributes to each principal component. Each row of this matrix corresponds to a feature (e.g., Time of day, Broken Truck, etc.), and each column corresponds to a principal component. So to make it more clear, The first row corresponds to **Time of day**, the second row to **Broken Truck** etc..
Each element in the matrix tells us how much the corresponding feature contributes to a given principal component. For example, the first element in the first column (0.49) tells us that Time of day has a weight of 0.49 for PC1.

We need to check how each feature contributes to PC2. The second column of V gives us the weights for PC2:
Time of day has a weight of -0.5. A low value of Time of day (since the weight is negative) will result in an increase in the projection onto PC2. **Broken Truck** has a weight of 0.23. A high value of Broken Truck (positive weight) will increase the projection onto PC2. **Accident victim** has a weight of 0.23. A high value of Accident victim (positive weight) will increase the projection onto PC2. **Defects** has a weight of 0.8. A high value of Defects (positive weight) will increase the projection onto PC2.

Since all these contributions either increase the projection or reduce it in a very minor way(this is because we have no information on the **Immobilized bus** feature), the overall projection onto PC2 will likely be positive, making Statement D true.

p.s although Immobilized bus has a positive weight, its small magnitude (0.09) means it doesn't significantly increase the projection compared to features with higher weights (like Defects, 0.8). This is why we say it "increases the projection in a minor way."

5. **Question 5**

   Correct A
   The Jaccard similarity between two text documents s_1 and s_2 measures the similarity between two sets by comparing the size of their intersection to the size of their union.
   The Jaccard similarity is defined as:

   $$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

   In our case, $A = s_1$ and $B = s_2$.
   $A \cup B = \{$the, bag, of, words, representation, becomes, less, parsimonious, if, we, do, not, stem$\}$
   $|A \cup B| = 13$
   $A \cap B = \{$the, words$\}$
   $|A \cap B| = 2$

   $$J(s_1, s_2) = \frac{|A \cap B|}{|A \cup B|} = \frac{2}{13} \approx 0.153846$$

6. **Question 6**

   We are interested in $p(\hat{x}_2 = 0|y = 2)$, which can be computed by summing over the possible values of $\hat{x}_7$:

   $$p(\hat{x}_2 = 0|y = 2) = p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) + p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2)$$

   We have $p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) = 0.81$ and
   $p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2) = 0.03$.
   So, $p(\hat{x}_2 = 0|y = 2) = 0.84$. In practice, we made use of the law of Total Probability, see: Law of Total Probability.