Jonathan Tybirk (s216136), Gabriel S. Hornstrup (s234833)
Lucas R.G. Pedersen (s234842)

# Explorative analysis of the Boston House Value dataset

## Description

In this report, we will analyze the features of the Boston Housing dataset. We mainly want to determine how the features correlate with each other in preparation for a later report which will analyze how these features can be used together to predict the median housing value of a 1978 Boston town. The dataset describes the median value of owner-occupied housing in 507 towns of Boston as well as 13 features about the respective towns, ranging widely from population statistics like pupil-teacher ratios to age of buildings to air quality. A full list of features can be found in the next section.

The Boston Housing dataset was originally collected by David Harrison Jr. and Daniel L. Rubinfeld for their 1978 paper examining the effect of air quality (measured by concentration of nitric oxides) on housing prices. The features of the data were collected from sources including the 1970 U.S. Census and data published by various organizations. The authors transformed the data, fit it in a multiple linear regression model, and concluded that low air quality had a significant negative impact on the value of housing in an area.

By analyzing the data, we are preparing to build a model for predicting the median house value of a town in 1970's Boston based on these attributes in a later report. This is a task most suited for regression, which would give a continuous estimate of the median. Classification could also be done but requires bracketing housing prices into different categories such as cheap, middle, and expensive. The cut-off between these categories could then e.g. be the 33% and 67% percentiles of the median house values in the dataset. We will initially try the methods after standardizing and normalizing the data. Looking at the scatter plots in the visualization section, it is not immediately clear how a feature could be transformed to better fit a linear regression.

Either method would allow us to predict the housing prices and gauge the importance of each feature in our model.

## Attributes:

Attribute descriptions are based on the descriptions from the 1978 paper[1]. All the variables are on a per-town basis.

| Feature | Description | Data type |
|---|---|---|
| CRIM | Crime rate per capita. | Continuous, ratio |
| ZN | Proportion of residential land zoned for lots greater than 25,000 square feet. | Continuous, ratio |
| INDUS | Proportion of non-retail business acres. | Continuous, ratio |
| Chas | Charles River dummy: =1 if tract bounds the Charles River; =0 otherwise. | Discrete, nominal |
| NOX | Nitrogen oxide concentrations (parts per 10 million). | Continuous, ratio |
| RM | Average number of rooms in owner units. | Continuous, ratio |
| AGE | Proportion of owner units built prior to 1940. | Continuous, ratio |
| DIS | Weighted distance to five employment centers in the Boston region. | Continuous, interval |
| RAD | Index of accessibility to radial highways. | Discrete, ordinal |
| TAX | Average full value property tax rate per $10,000. | Continuous, ratio |
| PTRATIO | Average pupil-teacher ratio by town school district. | Continuous, ratio |
| B | $1000(Bk - 0.63)^2$ where Bk is the proportion of black people by town. | Continuous, interval |
| LSTAT | Proportion of population of lower status. Average of two values: proportion of adults without some high school education and proportion of male workers classified as laborers. | Continuous, ratio |
| MEDV | Median value of owner-occupied homes per $1,000. | Continuous, ratio |

Note: We classified B and DIS as interval, as one can extract some meaning from the differences between two datapoints (a large difference in B implies a large difference in proportion of black people between two towns, and likewise a high difference in DIS corresponds to generally being further or closer to employment centers). However, as is visible from the attribute description, B measures the difference between the proportion of black people and 0.63 *squared*. This means a

[1] Hedonic Housing Prices and the Demand Clean Air, Harrison and Rubinfeld, published in Journal of Environmental Economics and Management 5, 81-102 (1978)

large B value can arise from both a very small and a very large proportion of black people. A low difference in B value therefore does not necessarily imply a low difference in proportion of black people. Originally, this transformation was done as the authors believed that both a low proportion and a high proportion of black people led to higher housing prices, but it prevents us from knowing the true proportion as squaring is non-invertible. We have decided to use the attributes despite these uncertainties, to maintain the maximum amount of information. We will later discard them if they turn out to insignificant in our linear regression.

Furthermore, 132 of the datapoints had a RAD of 24 and comparatively high TAX and very high crime rates (CRIM). We decided to keep these datapoints, despite the RAD value being way out of range of the rest (which were integers from 1-8) and it seeming suspicious that they all had the same high TAX, INDUS and PTRATIO and all had very high crime rates compared to the rest. The best explanation we could find was that there was only general data about these towns for certain attributes, so the same value was assigned to all of them. The high RAD value could be explained by supremely good access to highways, which most likely corresponds to dense cities which may also explain the high crime and tax rates. We have not been able to find any critique of the data in other analyses.

The following table contains summary statistics for all the features.

| | Mean | Standard Deviation | Min | First Quartile | Median | Third quartile | Max |
|---|---|---|---|---|---|---|---|
| **CRIM** | 3.61 | 8.59 | 0.01 | 0.08 | 0.26 | 3.68 | 88.98 |
| **ZN** | 11.36 | 23.30 | 0.00 | 0.00 | 0.00 | 12.50 | 100.00 |
| **INDUS** | 11.14 | 6.85 | 0.46 | 5.19 | 9.69 | 18.10 | 27.74 |
| **Chas** | 0.07 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| **Nox** | 0.55 | 0.12 | 0.39 | 0.45 | 0.54 | 0.62 | 0.87 |
| **RM** | 6.28 | 0.70 | 3.56 | 5.89 | 6.21 | 6.62 | 8.78 |
| **AGE** | 68.57 | 28.12 | 2.90 | 45.02 | 77.50 | 94.07 | 100.00 |
| **DIS** | 3.80 | 2.10 | 1.13 | 2.10 | 3.21 | 5.19 | 12.13 |
| **RAD** | 9.55 | 8.70 | 1.00 | 4.00 | 5.00 | 24.00 | 24.00 |
| **TAX** | 408.24 | 168.37 | 187.00 | 279.00 | 330.00 | 666.00 | 711.00 |
| **PTRATIO** | 18.46 | 2.16 | 12.60 | 17.40 | 19.05 | 20.20 | 22.00 |
| **B** | 356.67 | 91.20 | 0.32 | 375.38 | 391.44 | 396.23 | 396.90 |
| **LSTAT** | 12.65 | 7.13 | 1.73 | 6.95 | 11.36 | 16.96 | 37.97 |
| **MEDV** | 22.53 | 9.19 | 5.00 | 17.02 | 21.20 | 25.00 | 50.00 |

From the summary statistics we see that the data looks as expected. The proportions are all in the interval [0; 100], the tax range seems plausible (187k to 711k), so does the PTRATIO (12.6 to 22) and so on. The range of B may seem weird, but as we explained earlier it has already been transformed. The only weird thing one might notice is that MEDV is bounded by very clean values. We do not know the reason for this, and in other analyses it is theorized that the values were capped at 50. Only 16 datapoints have a value of 50, and based on the spread of the data it doesn't seem too unreasonable that this number of towns have a value close to 50. Keeping in mind that it only affects a fraction of the dataset, we judge that it is not detrimental to our analysis and decide to keep it.

For the rest of our analysis, we have standardized all the data to a mean of 0 and a standard deviation of 1. We did this to avoid putting any importance on one attribute over another. This also makes it possible to easily compare the cross-correlations of the features.

To better understand the distribution of our features, we checked normality:

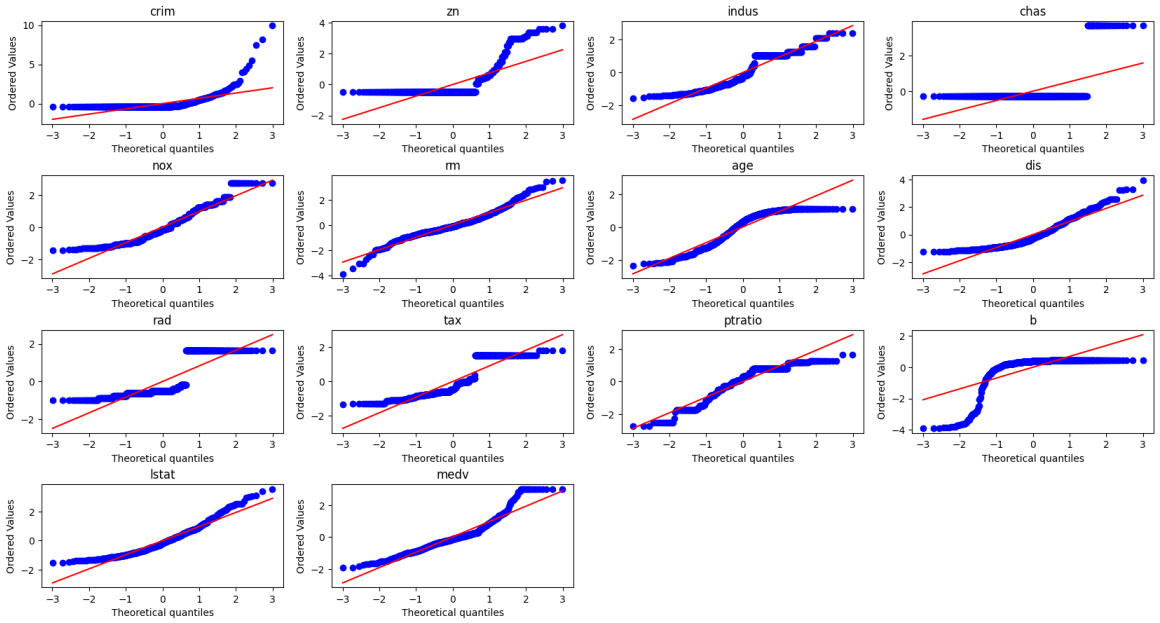


*Figure 1: QQ-plots of the attributes.*

As seen in figure 1 of the QQ-plots of all attributes, only RM seems reasonably close to normally distributed, however with what seems like systematic deviation at the tails. However, as the deviation is small, and it is normal for normally distributed variables to have deviant tails, we conclude that it can reasonably be estimated by a normal distribution.

Jonathan Tybirk (s216136), Gabriel S. Hornstrup (s234833)
Lucas R.G. Pedersen (s234842)

## Visualization:

To learn more about the data, we constructed the covariance matrix (see figure 2). We see that LSTAT and RM, corresponding to proportion of people of lower status (LSTAT) and average number of rooms per dwelling (RM) respectively, are highly correlated with the median house value (MEDV) - RM positively and LSTAT negatively. Many of the other variables also show slight correlation with the MEDV. Specifically, PTRATIO, TAX, INDUS are negatively correlated, though to a lesser degree. Examining the covariance between features, the most noticeable pairs are (RAD TAX), (NOX INDUS), (INDUS AGE), (NOX AGE), (DIS AGE) and (LSTAT RM), where the first two pairs have positive correlation and the rest have negative. Notice the pair (LSTAT RM) are correlated; this might explain why both are correlated with MEDV and makes it harder to determine which one has a causal relationship with MEDV. (RAD TAX) are almost perfectly correlated, but as we discussed in the last section, many datapoints share the same TAX and RAD values, which may cause the correlation statistic to be higher than anticipated. This is not the case for the other correlation pairs. Either way, high correlation may cause problems for linear regression, so it is something to note for later.

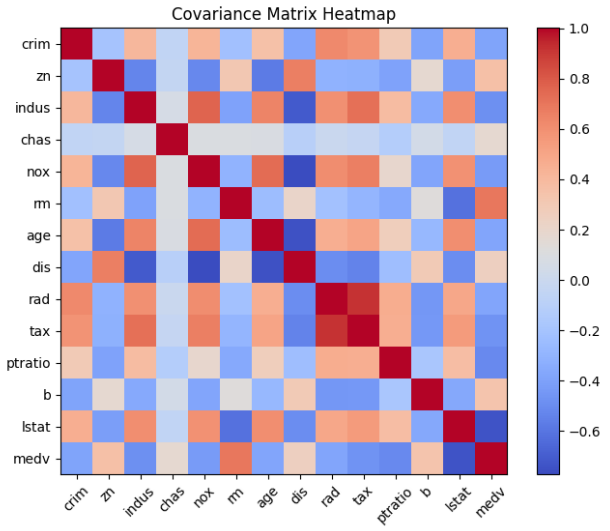


*Figure 2: Covariance matrix of all features against each other. Notice that due to standardization this is the correlation matrix as well.*

To further analyze the relationships in the data and spot any nonlinear relationships, we plotted all the features against the median house value in 13 scatterplots (figure 3).

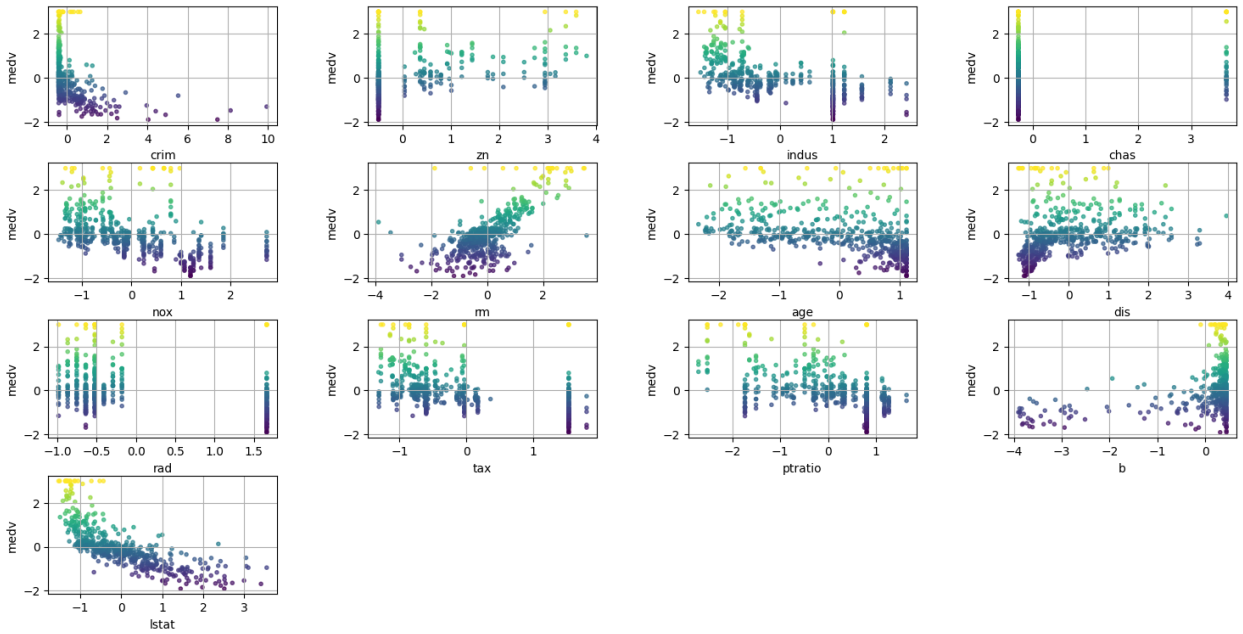


*Figure 3: scatterplots of all features against the median house value.*

Jonathan Tybirk (s216136), Gabriel S. Hornstrup (s234833)
Lucas R.G. Pedersen (s234842)

From the scatterplots in figure 3, it seems that most variables do hold some information regarding MEDV. However, the exact relationships are hard to confirm. It seems that a few of the other variables are correlated with the lower bound for MEDV, but within a big interval where the upper bound does not change. One should also note for this illustration that a large amount of datapoints naturally give a wider range as there are more possibilities for "rare" values. As an example, for B it is hard to judge whether only a high value of B can correspond to a high MEDV, as this might simply be a spurious relationship caused by the larger amounts of datapoints with a high B value compared to a low B value. To get an overall picture of the relationship between all the features and MEDV, we perform a dimensionality reduction with a PCA. To start off the following analysis, we examine the explained variance based on the number of principal components, plotted in figure 4.

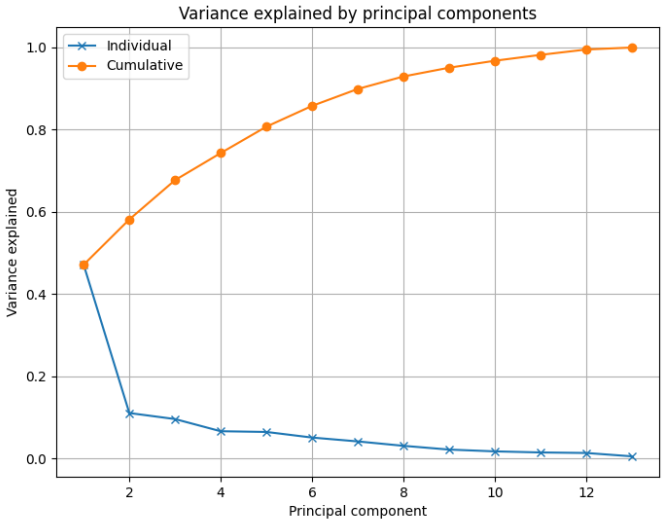


*Figure 4: Variance explained by number of principal components.*

To achieve an explained variance of at least 80%, which would be a low requirement, we need at least 5 principal components. However, as we want to visualize the data, we are only able to use at most 3 principal components. This achieves an explained variance of about 68%. Large aspects of the data will thus not be accounted for in the visualizations of figure 5 and figure 6.

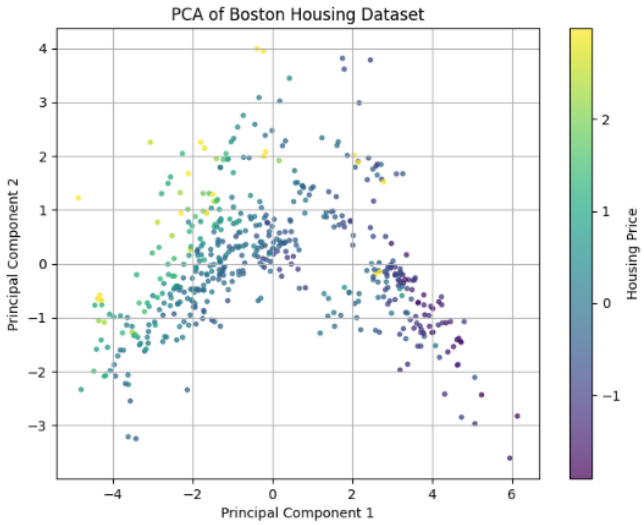


*Figure 5: Principal component analysis of dataset with 2 principal components.*

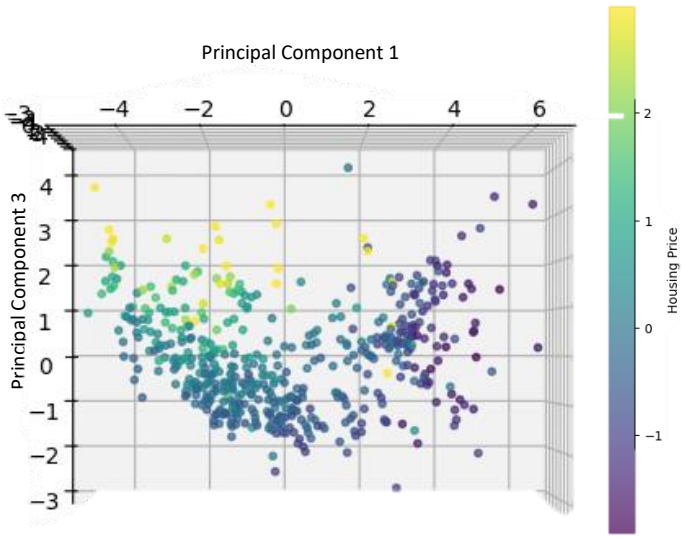


*Figure 6: PCA of dataset with 3 principal components, pitched 90 degrees upwards compared the perspective in figure 4 (if it had been 3D). This rotation is done to better see the effect of principal component 3*

Looking at figure 5, housing prices seem to be lower at a high principal component 1 (PC1) and low PC2, though the highest (yellow) housing prices are hard to separate from the rest. However, using PC3 in figure 6, we can see that many of these are located at a high PC3 value, and that at low PC1 values, PC3 is positively correlated with MEDV. To gain insight

into what the principal components stand for, we have created bar plots for every principal component in figure 7, showing how much each feature contributes to its value.

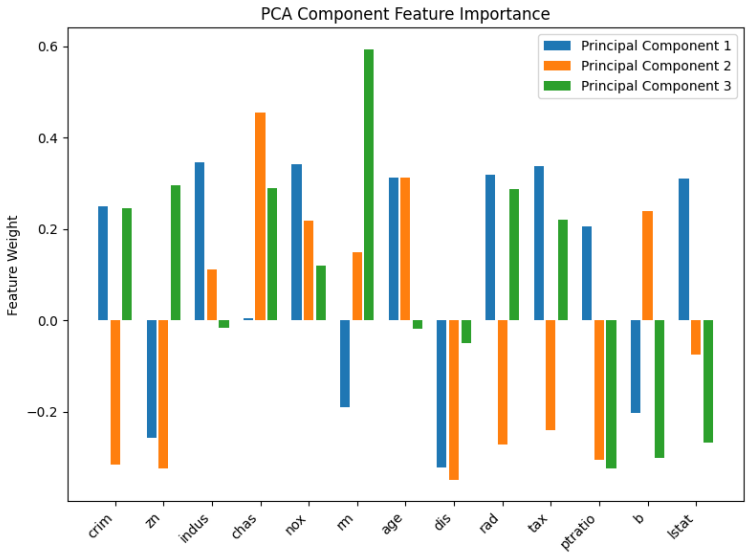


*Figure 7: Contribution of each feature to the 3 PCs plotted in figure 5 and 6.*

If we simplify our findings from figures 5 and 6, PC1 is negatively correlated with MEDV while PC2 and PC3 are positively correlated. Looking at the bar plots, we see that LSTAT contributes positively to PC1, and negatively to PC2 and PC3, while the exact opposite is true for RM. As we already know, RM and LSTAT are respectively positively and negatively correlated with MEDV, meaning at least some of the correlation we see in our principal components to MEDV come from the simple RM and LSTAT correlations. This is especially true for PC1 and PC3 which the two features contribute a lot to, whereas to PC2 the contributions aren't as large. But, as PC2 still is useful for predicting MEDV, it is plausible that the other features also contain useful information for predicting house values, which was something we expected from inspecting the scatter plots of figure 3. The exact relation is hard to pinpoint as there are many large contributions to PC2, namely from CRIM, ZN, CHAS, NOX, AGE, DIS, RAD, TAX, PTRATIO and B.

Jonathan Tybirk (s216136), Gabriel S. Hornstrup (s234833)
Lucas R.G. Pedersen (s234842)

## Discussion

First off, examining our attributes, we noticed three things that may impede our understanding of the data. The feature B is transformed in advance in a non-invertible fashion, many datapoints had a very high RAD value of 24 compared to 1-8 for the rest and MEDV seems to be capped at 50. While this does not necessarily cause any problems, it may explain future problems we encounter.

From our data analysis, we learned that two features, particularly LSTAT and RM are correlated with MEDV, however, these two features are also correlated with each other, which could pose problems for our linear regression and might mean we need to drop one. A few of the other features are also highly correlated, particularly TAX and RAD. We also learned that many of the other features seem to have some information about MEDV, but less obviously so, possibly with non-linear relationships.

Based on all this, it seems highly feasible to predict the median house value of a Boston town from the features. Plotting just 2 or 3 principal components with a color gradient representing housing price, a clear pattern already emerges. This is despite the 3D PCA only explaining 68% of the variance of the data. When we do linear regression in the next report, we will look for ways to transform the data to better capture potential non-linear effects as well, which could increase the feasibility of our goal even further.

## Responsibility distribution

The entire report was developed in collaboration.

| | Description | Attributes | Visualization | Discussion | Exam |
|---|---|---|---|---|---|
| Gabriel | 30% | 30% | 30% | 40% | 40% |
| Jonathan | 40% | 40% | 30% | 30% | 30% |
| Lucas | 30% | 30% | 40% | 30% | 30% |

## Exam questions

1. Option D: You can meaningfully add and subtract time of day, but not scale it, thus $x1$ is interval. $x2$ and $x3$ describe amounts for which a value of $0$ corresponds to 'nothing', and so scaling is meaningful. $x2$ and $x3$ are thus ratios. Finally, $y$ is ordinal, as you can rank the possible values (low, high) but not add or subtract them.

2. Option A: From the beneath calculations, it is apparent that $d_{p=\infty}$ gives the only result matching one of the options.

$$d_{p=\infty}(x_{14}, x_{18}) = \max(|26-19|, |2-0|) = |26-19| = 7$$

$$d_{p=3} = \sqrt[3]{|26-19|^3 + |2-0|^3} \approx 7.054004$$

$$d_{p=1}(x_{14}, x_{18}) = \sqrt[1]{|26-19|^1 + |2-0|^1} = 9$$

$$d_{p=4}(x_{14}, x_{18}) = \sqrt[4]{|26-19|^4 + |2-0|^4} \approx 7.011633$$

3. Option A: To find the percentage of variance explained by each of the principal component's S is squared and divided by the sum of its diagonal. Starting with option A we can already conclude that this is the correct answer since the first 4 eigenvalues sum up to be $0{,}8667932$ which is greater than 0.8.

Jonathan Tybirk (s216136), Gabriel S. Hornstrup (s234833)
Lucas R.G. Pedersen (s234842)

$$S \coloneqq \begin{bmatrix} 13.9 & 0 & 0 & 0 & 0 \\ 0 & 12.47 & 0 & 0 & 0 \\ 0 & 0 & 11.48 & 0 & 0 \\ 0 & 0 & 0 & 10.03 & 0 \\ 0 & 0 & 0 & 0 & 9.45 \end{bmatrix}$$

$$S \cdot S \approx \begin{bmatrix} 193{,}21 & 0 & 0 & 0 & 0 \\ 0 & 155{,}5009 & 0 & 0 & 0 \\ 0 & 0 & 131{,}7904 & 0 & 0 \\ 0 & 0 & 0 & 100{,}6009 & 0 \\ 0 & 0 & 0 & 0 & 89{,}3025 \end{bmatrix}$$

$$trace(S \cdot S) = 193{,}21 + 155{,}5009 + 131{,}7904 + 100{,}6009 + 89{,}3025 = 670{,}4047$$

$$S \cdot S \cdot \frac{1}{670{,}4047} \approx \begin{bmatrix} 0{,}2881991 & 0 & 0 & 0 & 0 \\ 0 & 0{,}2319508 & 0 & 0 & 0 \\ 0 & 0 & 0{,}1965833 & 0 & 0 \\ 0 & 0 & 0 & 0{,}15006 & 0 \\ 0 & 0 & 0 & 0 & 0{,}1332069 \end{bmatrix}$$

$$0{,}2881991 + 0{,}2319508 + 0{,}1965833 + 0{,}15006 = 0{,}8667932$$

4. Option D: When looking at the principal component 2 column in matrix V, a negative value corresponds with the $x_1$ attribute and positive values for $x_2, x_3$ and $x_5$. Furthermore, option D asks about an observation with a low value for $x_1$ and positive values for $x_2, x_3$ and $x_5$. For these reasons we will typically have observations which have a positive value of the projection onto principal component 2.

5. Option A: For a Jacard similarity the total vocabulary size is not needed for the calculation and only the intersection and union of the two sets are needed for the calculation. It is calculated by the intersection over union and the result gives approximately $0{,}1538462$ which corresponds to option A.

$$s_1 \cap s_2 = \{the, words\} = 2$$

$$s_1 \cup s_2 = \begin{Bmatrix} the, bag, of, words, representation, becomes, \\ less, parsimoneous, if, we, do, not, stem \end{Bmatrix} = 13$$

$$\frac{2}{13} \approx 0{,}1538462$$

6. Option B: To find the probability of observing particular values of $\hat{x}_2$ and $\hat{x}_7$ we take the corresponding percentage chance of $\hat{x}_2$ being equal 0 in the column $y = 2$ and dividing it by the sum of all probabilities of $y = 2$ happening.

$$\frac{0.81 + 0.03}{0.81 + 0.03 + 0.10 + 0.06} = 0.84$$