Report 1.- California Housing Prices

02450: Introduction to Machine Learning and Data Mining

Lucas Besnard	s201778
Jan Cedric Issel	s201232
Ayoub El Outati	s201602

Lyngby, October 2020

Table of Contents

1. Description of the Dataset

The observations of our dataset were collected in 1990 by California census and are about attributes found in different housing districts in California. The data is composed of different metrics like median house value, median income, population as well as locations. Thus, it provides information about the location of the respective district, the houses in that district, and the people living there. The dataset contains a total of 20,649 observations. Each observation corresponds to a district of houses. The main problem of interest concerning this dataset is to predict the median housing value of the housing districts using the other attributes. We obtained the dataset as a .csv-file on Kaggle at https://www.kaggle.com/camnugent/california-housing-prices.

This dataset has been studied in a number of competitions on Kaggle and was subject to several data analysis projects. However, since the information in the dataset is not very up to date, most of the publications are of a didactic or tutorial character. Several visualizations of the attributes and their repartitions are available online. We can find different machine learning methods being used on this dataset to predict the median house value. At https://www.kaggle.com/subashdump/california-housing-price-prediction, one can find an example of the decision tree method and the random forest method which both give very accurate results. At https://jmyao17.github.io/Kaggle/California_Housing_Prices.html, a support vector machine implementation is used on the dataset. Also, at https://www.kaggle.com/prashant111/explain-your-model-predictions-with-shapley-values one can find a notebook that uses the Shapley values from game theory to estimate the housing median value.

The regression task in this project will be to predict the median house value based on other fitting attributes of the dataset. In general, all the attributes can be pertinent to explain the price of the housing. However, one of the more suitable attributes to explain it could be the location since the price of the housing is higher near the coast or in metropolitan areas like San Francisco and Los Angeles. Another relevant attribute could be the median income because this is a major factor if one has to consider which house he will buy. For the classification task a discrete attribute is needed. The only attribute in the dataset that fulfills this criterium is the ocean proximity. The ocean proximity is composed of five classes called “Inland”, “<1h from Ocean”, “Near Ocean”, “Near Bay”, and “Island”. Therefore, in the classification task we aim to predict the respective class of ocean proximity based on the other attributes. Using this machine learning techniques, we hope to gain more insights about the factors that influence the value of houses in California. Although the collection of this data is quite some time ago, the general relationships between the attributes of the dataset can still be considered valid today. Especially a solution for the regression task can provide a guidance for potential house buyers who wish to evaluate the price level of a house they are interested in. Regarding the clustering a task for this dataset could be to find different clusters of populations that share similar characteristics. A reason to use association mining for this dataset could be to find higher dimensional relations between several attributes like for example the influence of ocean proximity and population on the median house value. In case that the regression model doesn’t perform as expected, an anomaly detection like k-means could be an option to find outliers that distort the data distribution. However, for this dataset the main machine learning goal will be the prediction of median house value with a high accuracy.

2. Explanation of the Attributes

The dataset consists of ten attributes that describe different aspects related to the housing in California. Below these attributes are described in detail with regard to continuity and scale.

Longitude	The longitude is a measure of how far west a housing block is. The values range from -180 to +180 degrees. This attribute is **continuous**. It is **interval** scaled because there is no absence of what is measured. Whatever the location is, there will always be a longitude value.
Latitude	The latitude is a measure of how far north a housing block is. The values range from -90 to +90 degrees. Analogous to the longitude this attribute is **continuous**, and **interval** scaled as well.
Housing Median Age	The housing median age means the median age of a house within a block where new buildings have lower numbers and older buildings have higher numbers. This attribute is **continuous**. It is **ratio** scaled because if a building has just been built, the housing age is not existent.
Total Rooms	With the total rooms the total number of rooms in a building aggregated for the whole block is meant. This attribute is **discrete** because it describes a finite number of values that are integer, i.e. there are no half rooms. It is **ratio** scaled because it is a count and because it is theoretically possible to measure a housing block that doesn't have any rooms. Thus, there exists a true zero.
Total Bedrooms	The attribute of total bedrooms is analogous to the attribute of total rooms.
Population	The population describes the total number of people living in buildings within a block. This attribute is **discrete** because the number of residents in a block is finite and integer, i.e. there are no half persons. It is **ratio** scaled because it is a count and because it is theoretically possible to measure a housing block without residents. Thus, there exists a true zero.
Households	Households means the total number of households within a block since one building can contain more than one house. This attribute is **discrete** because the number of households in a block is finite and integer, i.e. there are no half households. It is **ratio** scaled because it is a count and because it is theoretically possible to measure a housing block without households. Thus, there exists a true zero.
Median Income	The median income for households within a block is measured in tens of thousands of US Dollars. In general, money can be considered discrete or continuous. However, in this dataset the attribute is **continuous** because it consists statistically derived median values with floating point numbers. It is **ratio** scaled because it is possible to measure the absence of income, i.e. the median income per household in a block is zero.
Median House Value	The median house value for households within a block is measured in US Dollars. Analogous to the median income it can be argued that this attribute is **continuous**, and **ratio** scaled
Ocean Proximity	The ocean proximity describes the location of the houses with respect to the ocean in words. This means that the data are descriptive instead of numerical. The attribute consists of five different proximity descriptions called "near bay", "<1h ocean", "inland" "near ocean" or "island". Thus, it is a **discrete** attribute because a continuous value between two of these descriptions is not defined. The attribute is **nominal** scaled because the five proximity descriptions can't be ordered without further interpretation.

In the following the dataset is examined regarding any data issues like missing values or corrupted data. The reason for this approach is to prevent using a biased sample for further analysis and to enable the correct representation of any phenomena hidden in the data.

First the dataset is filtered for missing values. In total there are 207 measurement instances that contain a missing value for an attribute. All the missing values belong to the "Total Bedrooms" attribute. In general, there are different options to handle missing values in a dataset. It is possible to eliminate the measurement instances that contain the missing values from the dataset, i.e. to delete the respective rows. Another option is to estimate the missing values by using e.g. an average. On the other hand, it is possible to just ignore the missing values during the analysis. In this case a look at the correlation between the attributes "Total Rooms" and "Total Bedrooms" shows that the attributes are with a correlation of 0.93 strongly positive related which means we can estimate the number of bedrooms from the number of rooms in a housing block. Thus, the ratios of total bedrooms to total rooms is computed for each instance in the dataset without missing values. The average of the result is then used to estimate the missing values based on the number of rooms in the respective measurement.

Corrupted data can be data with e.g. a deviating datatype or a deviating sign. In the current dataset a checkup for datatype of each attribute shows no evidence for deviating datatypes. Subsequently every attribute that describes a number (e.g. number of household) is examined with regard to the sign of its values. However, these are all positive. Therefore, it is concluded that the dataset is free of corrupted datapoints. In Table 1 we supply some summary statistics for all numeric attributes.

Table 1: Summary statistics

	housing median age	total rooms	total bedrooms	population	house-holds	median income	median house value
unit	years	number	number	number	number	$10000	$
count	20640	20640	20640	20640	20640	20640	20640
mean	28.64	2635.76	537.95	1425.48	499.54	3.87	206855.82
std	12.59	2181.62	420.99	1132.46	382.33	1.90	115395.62
min	1.00	2.00	1.00	3.00	1.00	0.50	14999.00
25%	18.00	1447.75	296.00	787.00	280.00	2.56	119600.00
50%	29.00	2127.00	435.00	1166.00	409.00	3.53	179700.00
75%	37.00	3148.00	648.00	1725.00	605.00	4.74	264725.00
max	52.00	39320.00	6445.00	35682.00	6082.00	15.00	500001.00

3. Data Visualization and PCA

Concerning the outliers in our data set, we use a histogram for each numeric attribute in order to detect potential distortion and to have a clear visualization of the distribution of the data. The matrix of these histograms can be seen in Figure 1, we have outliers in the housing median age and the median house value. The frequency is well distributed but on the right-hand side both attributes show a strange increase of frequency for extreme values. In the description of the dataset this is explained by a preprocessing of the collected data. For instance, values superior to $500.000 in the median house value attribute have been substituted by $500.000. Similar operations have been done for the housing median age and the median income. Therefore, there are a lot of observations gathered in one bin for the respective attributes. Consequently, we need to filter our data set and take these extreme values out of the dataset to avoid distortions and other problems in our machine learning approach. In addition, we noticed that the median income attribute is scaled in an inconsistent way. The values are scaled and capped to stay between 0 and 15 in thousand US$ while other attributes with currency unit are not scaled like this. This anomaly can disturb our machine learning predictions for example if we want to use the median income to predict the median house value.

The histograms in Figure 1, can also be used to check the distribution of the attributes. By visually interpreting these histograms, we can get an insight if an attribute follows a normal distribution or not. We do this by evaluating if a bell-shaped curve can be fitted into the distribution. Therefore, we can conclude that none of the attributes follow a normal distribution in our data set. Especially the attributes of total rooms, total bedrooms and population are more matching with the curve of a Poisson distribution. Indeed, many histograms are tail-heavy to the right of the median which highlights the asymmetry.

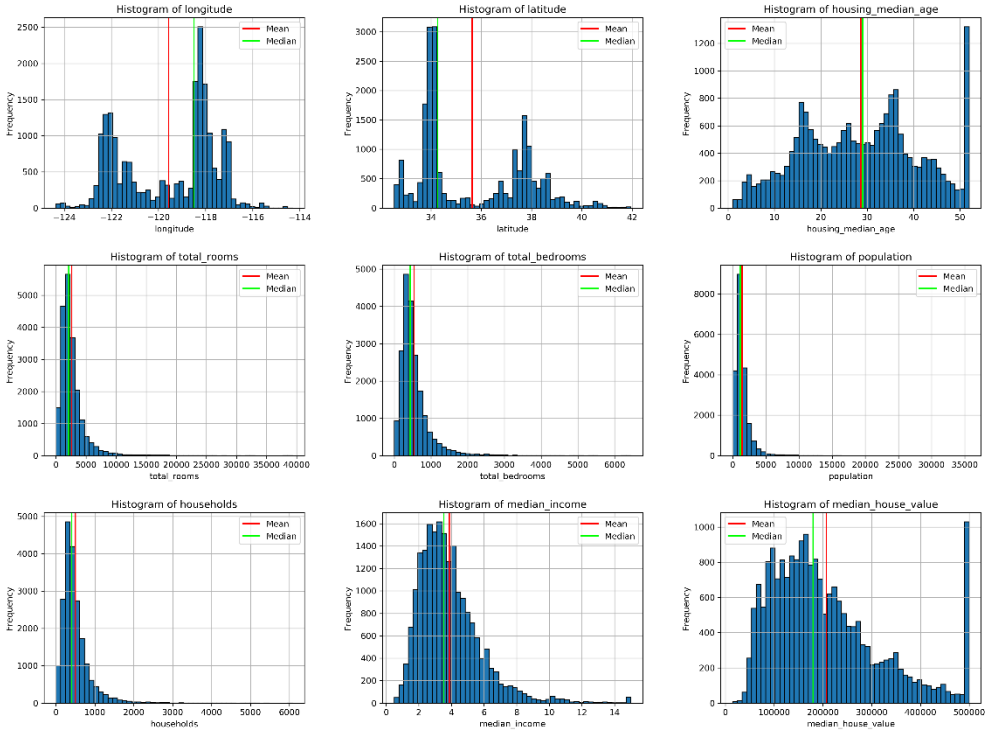


Figure 1: Histograms of all numeric attributes with mean and median

From the previous observations we conclude that some preprocessing operations have to be done to receive a proper correlation between the data and to make our data ready for the machine learning task. First, we create three new attributes that are ratios of existing attributes which they will

substitute. With this feature engineering step, we aim to get attributes that are more normal distributed and that show good correlation with other attributes. The respective attributes and their substitutes can be seen in Table 2. The distributions of the new attributes are visualized in Figure 2. One can see that the attributes have a bell shape which indicates a normal distribution. Next, we define a number of thresholds to filter the mentioned outliers from the dataset. Last, we rescale the median income attribute so that it has the same scale as other attributes with a currency unit.

Table 2: Original attributes and their substitutes

Original Attribute	**Substitute**
Total rooms	Rooms per household
Total bedrooms	Bedrooms per household
Population	Population per household

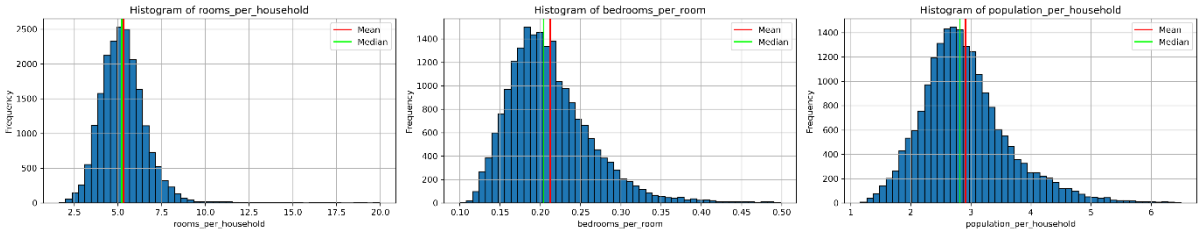
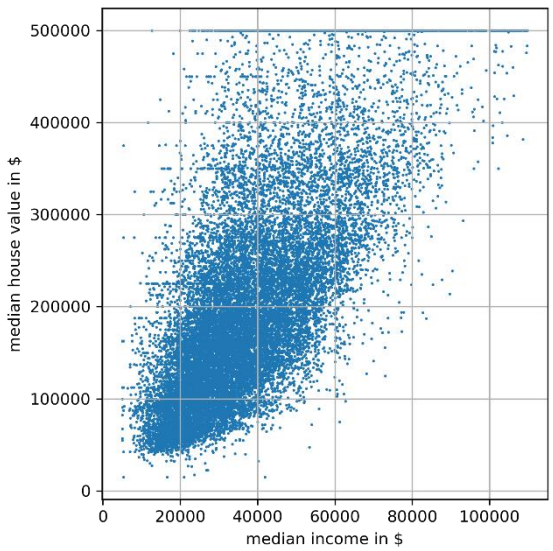


Figure 2: Distributions of new attributes

After preprocessing the dataset, we can now evaluate the correlation between the attributes. Since we aim to predict the median house value with regression, we focus on the Pearson correlation coefficients of the median house value as well as the median income with the other attributes. In Figure 3, we list the calculated correlations on the right-hand side. We notice that the correlation of the median house value with the median income is relatively high. The scatter plot on the left-hand side of Figure 3 confirms this observation by visualizing that the points show an upwards trend and have a relatively low dispersion. Thus, it indicates a positive relation of the attributes. The location attributes stand out because they show almost no correlation to the evaluated attributes. However, an independence of the house value from the location seems unlikely. It is more likely that the correlation is just not visible due to the geographical coding in longitude and latitude. The new created attributes all show some considerable correlation to the examined attributes. Thus, the reason for their creation is given and we keep them for the upcoming investigations.



	Median Income	Median House Value
Longitude	-0.02	-0.05
Latitude	-0.08	-0.15
Housing Median Age	-0.15	0.11
Households	0.03	0.08
Median Income	1	0.69
Median House Value	0.69	1
Rooms Per Household	0.57	0.25
Bedrooms Per Room	-0.66	-0.27
Population Per Household	-0.08	-0.28

Figure 3: Correlation of the median house value with all other attributes

Based on the insights from the visualization and the resulting conclusions, we can now evaluate if our machine learning goal seems feasible. We were able to estimate the missing values in the dataset and we filtered outliers that could disturb the training process of our machine learning approach. We substituted some of the attributes by ratios which resulted in more normally distributed attributes. Last but not least we found several attributes that are correlated to the median house value as well as the median income. Thus, our primary machine learning goal, to predict the median house value with regression, seems to be feasible with the current dataset. Another factor contributing to this conclusion is that several machine learning algorithms have already successfully been used for this dataset.

After analyzing and preprocessing our data, we can now apply a principal component analysis (PCA). We do this on the one hand to evaluate how much of the variance in the data can be explained by a single principal component (PC) and on the other hand to see which PC explains which attribute. Another purpose of this method is to consider a dimensionality reduction without losing a lot of information in order to have an easier perception of our data. To get reliable results from the PCA, there are some steps to be followed. First, we omit the ocean proximity attribute for this analysis because it consists of discrete classes. Since the remaining attributes have different units (currency, numbers, coordinates), we need to standardize the data first by subtracting the mean and dividing the result by the standard deviation. Once we standardized the data, we compute the PCA on our data and obtain the principal components. Each principal component explains a certain amount of variation in the data. In Figure 4 we display this variation as a function of the number of attributes. One can see that it roughly takes the first five out of nine PCs to explain 90% of the variance. This means that we could project our data onto five PCs instead of nine attributes and could still cover 90% of the original variance.

Next, we take a look at the principal directions given by the PCA by plotting the principal components for each attribute. Thereby, we can evaluate how much of the variation of each attribute is explained by the respective PC. Since, we found out that the first five PCs explain 90% of the variance, we do the evaluation only for this more important PCs. Based on the corresponding graph in Figure 5, we notice that the attributes described by the feature vectors of PC1 are mainly the median income, the median house value, the rooms per household and the bedrooms per room. The feature vector of PC2 mostly

describes the geo attributes like latitude and longitude. PC3 is made of the population and the population per household attribute.

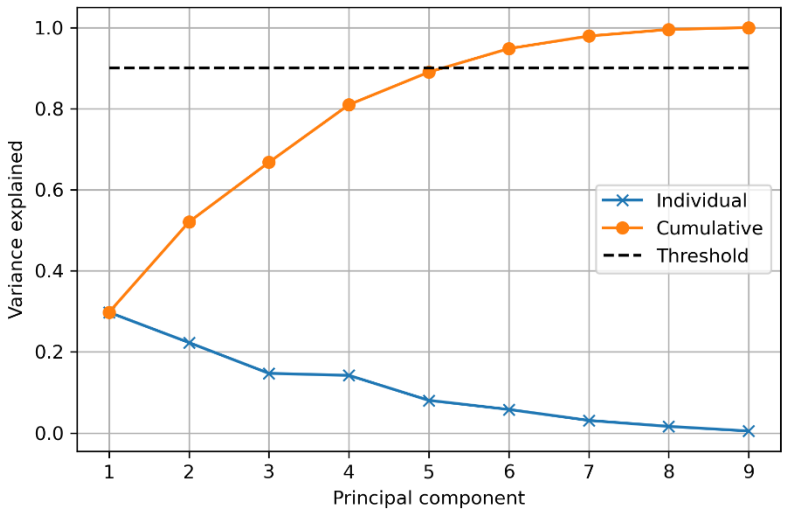


Figure 4: Variance explained by the principal components

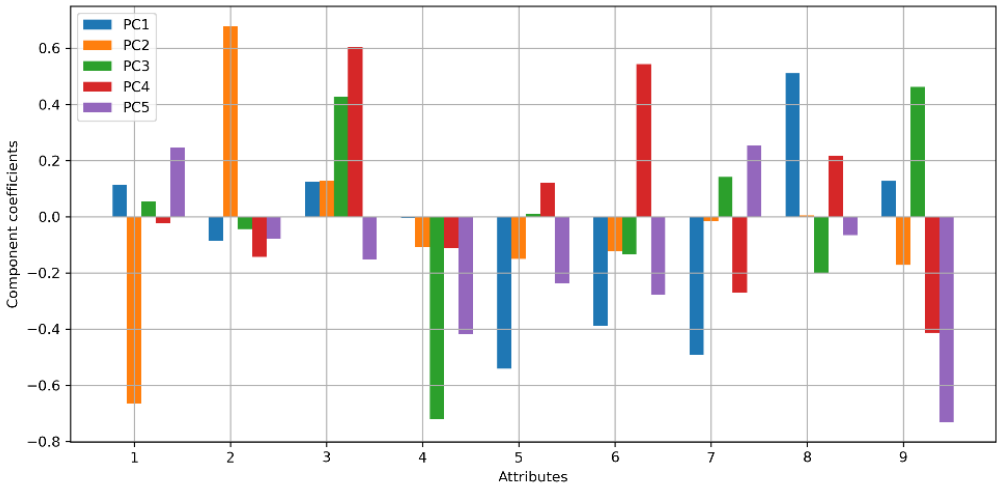


Figure 5: Principal directions obtained using the PCA

So far, we have gained considerable insights by computing the variance explained by the PCs and by looking at specific PCs that explain the variance of specific attributes particularly good. However, we have not gained any knowledge about the interference of the non-numeric ocean proximity attribute with the other attributes. Since we plan to perform a classification with this attribute it seems advisable to figure out if the calculated PCs show a relation to the ocean proximity although it has not been considered in the PCA. Therefore, we use the standardized data and project them onto the principal component space build by the vectors of the first two calculated PCs which gives us a scatter plot of the data. Next, we use the five classes of the ocean proximity attribute to create a colormap for the scatter. The result can be seen in Figure 6. It shows that the different classes of ocean proximity are connected to a correlation between the two PCs. For instance, the data points labeled as a location less than one hour to the ocean show a slow linear increase of PC2 for an increase of PC1. The data points labeled inland on the other hand show a strong variation for PC2. In general, it becomes apparent that the classes are somewhat distinguishable from each other but also still overlap considerably. An interpretation of this observation is that the variance contained in the attributes

apart from the ocean proximity which is explained by the two PCs can classify the ocean proximity but only to a certain extent. Therefore, a conclusion for our classification goal is that it may be necessary to do some further feature engineering to help the classifier learning to distinguish between the different ocean proximities.

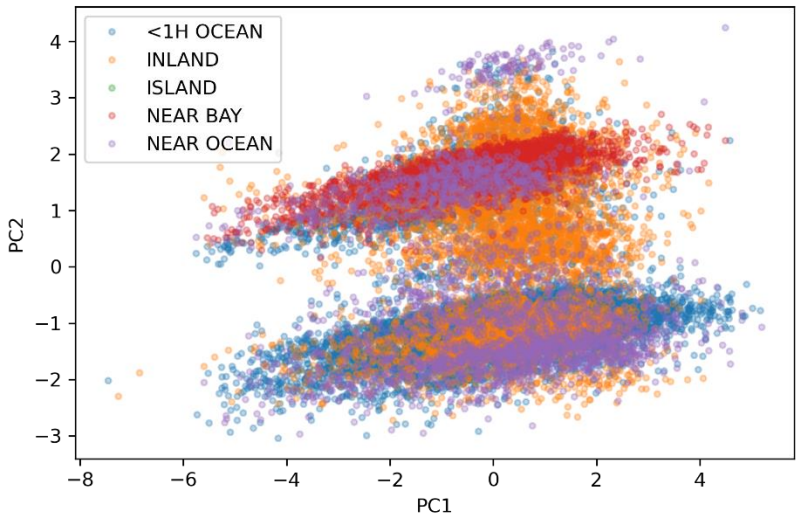


Figure 6: Projection of the PCs on the data

4. Discussion

In this report we started by gaining insights about the data by computing some summary statistics as well as plotting a histogram for each attribute. Thereby, we learned that there are distortions in the distribution of several attributes among others caused by some preprocessing of the data. In connection with this we also defined a set rules to filter outliers from the dataset. We decided to generate some new attributes as ratios of existing attributes to receive attributes that are normally distributed and show a good correlation towards our target attribute, the median house value. From the following analysis of the correlation matrix we learned that the target attribute is strongly correlated to the median income and that the combination of newly generated attributes could be a good predictor as well. Next, we applied a PCA to get a better visualization of our high dimensional dataset. From these visualizations we learned that it takes five PCs to explain 90% of the variation in all numeric attributes. We found out which PC explains the variation in which attribute particularly good. To obtain a visualization of the variance explained by the PCs with regard to the discrete classes of the ocean proximity attribute we projected the data onto the principal component space and used a colormap. Thereby, we learned that classification for this attribute seems possible but may need some further feature engineering.

Concerning our primary machine learning aim, the prediction of the median house value with regression, we have created a good base with the performed data processing and feature engineering. Especially considering the correlations of the attributes and the histograms of the generated attributes income it seems possible to do an accurate prediction of the target attribute. Thus, we conclude that the primary machine learning aim is feasible.

Table of Responsibilities

Sections	**Responsible for more than 40%**	**Contributor**
Description of the Dataset	Jan Cedric Issel	Lucas Besnard
Explanation of the Attributes	Jan Cedric Issel	
Data Visualization and PCA	Ayoub El Outati	Lucas Besnard, Jan Cedric Issel
Discussion	Jan Cedric Issel	