02450 Project 1 - Group 46

Team members:

Jordan Shopp, s226743

Natalie Valett, s226441

Ka Chun So (Andy), s226545

#### A description of the data set

The California Housing Prices dataset from Kaggle (Found at [https://www.kaggle.com/datasets/camnugent/california-housing-prices?resource=downl](https://www.kaggle.com/datasets/camnugent/california-housing-prices?resource=download)oad) contains data from the California housing census in 1990. It includes California housing data, and some demographic/regional data, aggregated by “housing block groups”, which are the smallest geographical unit for which the US Census Bureau publishes sample data (containing around 600 to 3,000 people).

This dataset has been used in numerous publications before, including a 1997 paper “Sparse spatial autoregressions” which utilized the geographic nature of this data to research optimization of regression on data with sparse spatial representations. It was also used in the textbook ‘Hands-On Machine learning with Scikit-Learn and TensorFlow’, which focuses on predicting the median housing price of a district using a few different machine learning methods. As these tasks are left to the reader, the results and analysis of the predictions aren’t specified in the textbook.

Our problem of interest with this data is to see if we can predict the median income in an area based on attributes about the housing market there, such as housing prices, proximity to the ocean, and population. To achieve this goal, we first need to focus on understanding and representing the data. Before anything can be done with the data, we need to account for the 207 entries that contain null values. Null values for numeric or categorical attributes can introduce problems in the computations required for principal component analysis (PCA), regression, and classification, so we’ve chosen to discard any rows/entries containing null values. These entries only account for 0.01% of our data, so their elimination shouldn’t have a profound effect on our results.

Before diving into machine learning applications, we will aim to reduce redundancy and glean insight into the correlation between attributes in our data through feature extraction. This will allow us to represent our data more concisely by reprojecting datapoints onto a more concise coordinate system, consisting of components which account for most of the variation in the data. After this dimensionality reduction, we aim to use classification to predict the “ocean proximity” attribute of housing blocks based on their median housing value, population, total bedrooms, households, and median income. Lastly, we will use regression to predict the median income in a housing block given the median house value, ocean proximity, housing median age, total rooms, total bedrooms, population, and households.

#### A detailed explanation of the attributes of the data

The dataset consists of 10 attributes. The first attribute is "longitude", which measures how far west a house is; a higher value indicates that it is farther west. The second attribute is "latitude", which measures how far north a house is; a higher value indicates that it is farther north. Both of these attributes are continuous and can be considered interval data, as they are measured on a continuous scale and the zero point is not the true zero. For example, the zero point of Longitude represents Prime Meridian, which is an arbitrary reference point rather than an absence of the attribute being measured.

The third attribute is "housing\_median\_age", which represents the median age of houses in the area, in years. This attribute is discrete since the raw data has been already rounded to the nearest integer. It can also be considered ratio data, as the zero age means no age.

The next four attributes are "total\_rooms", "total\_bedrooms", "population" and "households”, which represent the total number of rooms, bedrooms, population and total number of households in the respective blocks. Notice that the data use longitudes and latitudes to divide blocks. These attributes are discrete and can be considered ratio data. as the number of these attributes are recorded in a whole number and have a true zero point.

The seventh attribute is "median\_income", which represents the median income of households in the area and is measured in tens of thousands of US Dollars. This attribute is continuous and can be considered ratio data. The eighth attribute is "median\_house\_value", which represents the median value of owner-occupied homes in the area and is measured in US Dollars. This attribute is continuous and can be considered ratio data. The last attribute is "ocean\_proximity", which represents the proximity of the housing location to the ocean. This attribute is nominal and can be considered categorical data, as there is no inherent order to the categories, which are “<1 H OCEAN”, “INLAND”, “ISLAND”, “NEAR BAY”, and “NEAR OCEAN”.

It's worth noting that there is 0.01% missing data in the "total\_bedrooms" attribute, which would need to be addressed if this dataset were to be used in further analysis. To tackle the problem, we choose to eliminate it as mentioned above. Overall, this dataset provides a variety of information about housing in California that could be used for further analysis and modeling.

Statistics for each numerical attribute:

| **Attribute** | **Mean** | **Standard Deviation** | **Median** | **Range** |
| --- | --- | --- | --- | --- |
| longitude | -119.5697 | 2.0035 | -118.49 | 10.04 |
| latitude | 35.6319 | 2.1359 | 34.26 | 9.41 |
| housing\_median\_age | 28.6395 | 12.5856 | 29 | 51 |
| total\_rooms | 2635.7631 | 2181.6153 | 2127 | 39318 |
| total\_bedrooms | 537.8706 | 421.3851 | 435 | 6444 |
| population | 1425.4767 | 1132.4621 | 1166 | 35679 |
| households | 499.5397 | 382.3298 | 409 | 6081 |
| median\_income | 3.8707 | 1.8998 | 3.5348 | 14.5 |
| median\_house\_value | 206855.8169 | 115395.6159 | 179700 | 485002 |

Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA)

(i) Data visualization(s)

To begin the analysis, box plots and histograms were utilized to detect outliers and examine the distribution of attributes. Box plots were used to visualize the distribution of the data and to identify any values that fall outside of the interquartile range (IQR), which are commonly known as outliers. Outliers are typically defined as data points that are more than 1.5 times the IQR above the third quartile or below the first quartile. Any outliers present in the data will be displayed as points outside of the whiskers in the box plot. Additionally, histograms were used to group data into intervals or bins and display the frequency of observations within each bin as a bar. Histograms can reveal the shape of the distribution, such as whether it is symmetrical or skewed, and provide insight into the range and concentration of values within the data.

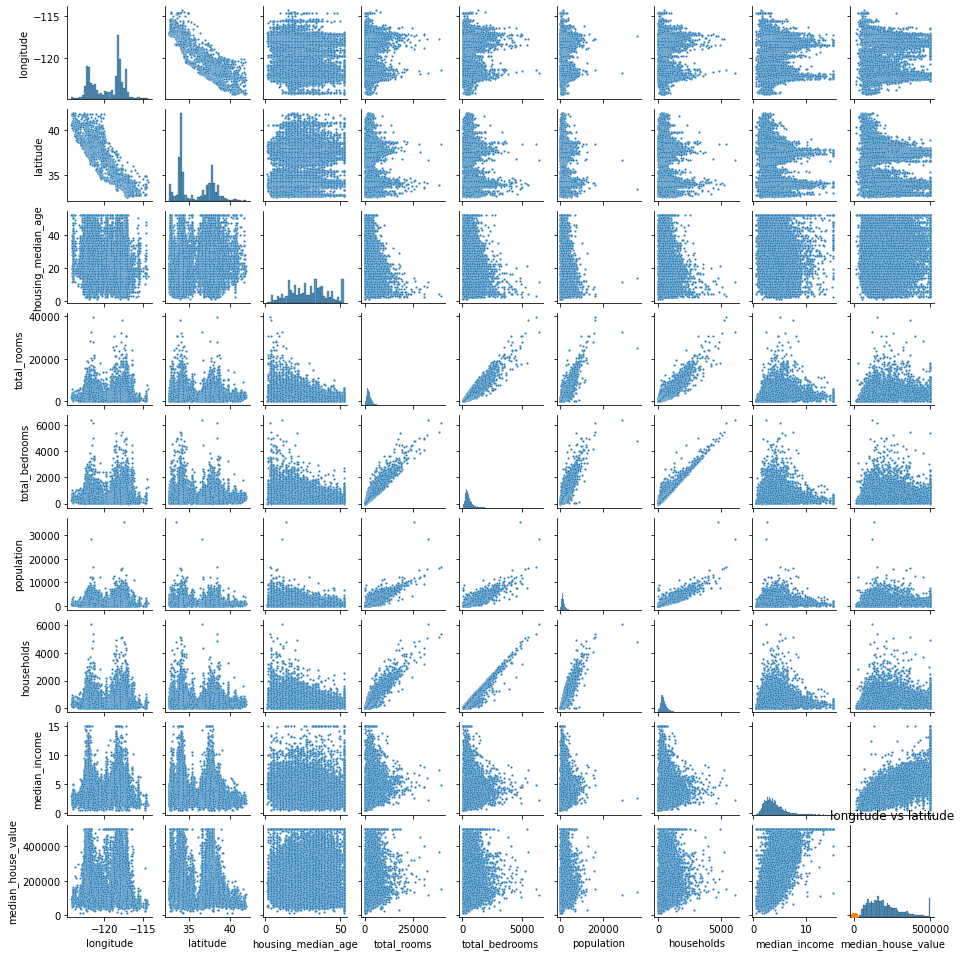
|  | **Box Plots (Identifying Outliers)** | **Histograms (Checking the distribution of attributes)** |
| --- | --- | --- |
| **longitude** |  |  |
| **latitude** |  |  |
| **housing\_median\_age** |  |  |
| **total\_rooms** |  |  |
| **total\_bedrooms** |  |  |
| **population** |  |  |
| **households** |  |  |
| **median\_income** |  |  |
| **median\_house\_value** |  |  |

We can conclude that there are no outliers for the attributes "longitude", "latitude" and "housing\_median\_age". For "median\_house\_value", there are a few outliers that lie above the maximum whisker of the boxplot. However, these outliers are not significantly distant from the rest of the data and are still within a reasonable range for the housing prices in California. On the other hand, there are certain obvious outliers for the attributes "total\_rooms", "total\_bedrooms", "population", "households" and "median\_income". These outliers are significantly distant from the rest of the data and might need to be treated or removed before further analysis. For example, the "total\_rooms" attribute has an outlier value of 39320, which is much higher than the rest of the values and might affect the analysis if not treated properly.

For distributions, the distribution of “median\_income” is approximately normally distributed, with the majority of the data centered around the mean and symmetrical tails on either side. On the other hand, the distributions of “total\_rooms”, “total\_bedrooms”, “population”, and “households” are all skewed right, with a few large values far from the majority of the data. Finally, the distribution of “median\_house\_value” is bimodal, with one peak around $170,000 and another around $500,000. For other attributes, we consider them irregularly distributed.

To evaluate the feasibility of the primary machine learning modeling aim, we created scatter plots of each pair of numeric attributes. The scatter plots help visualize the relationship between each pair of attributes, allowing us to gain insights into further analysis.

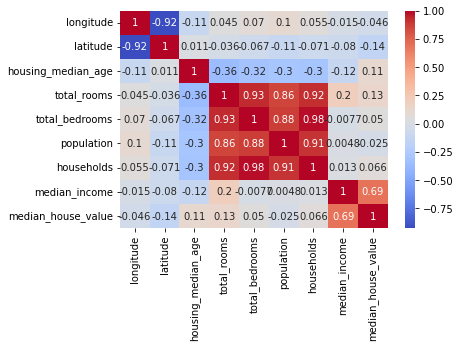
**Scatter Plot from scatter\_plot.py**



Looking at the scatter plot, we can see that several pairs of attributes are correlated. For example, the scatter plot of "median\_income" vs. "median\_house\_value" shows a clear positive correlation, indicating that higher median incomes are associated with higher median house values. Similarly, the scatter plot of "total\_rooms" vs. "median\_house\_value" shows a weaker positive correlation, suggesting that larger homes tend to be more expensive, but that there are other factors at play as well. It is feasible for us to continue further machine learning studies like house value predictions.

We have created a correlation matrix heatmap also. It is a graphical representation that shows the correlation between pairs of attributes in the dataset. The heatmap uses color coding to represent the strength of the correlation, with darker colors indicating stronger positive or negative correlations and lighter colors indicating weaker correlations or no correlation.

**Correlation Matrix Heatmap of Each Pair of Attributes from basic\_stat.py**



The heatmap that we have created for our dataset shows that there are several attributes that are strongly correlated with each other. For example, “total\_rooms” and “total\_bedrooms” have a strong positive correlation, which suggests that they may provide redundant information when used together in a machine learning model.

Similarly, “population” and “households” also have a strong positive correlation, which suggests that these two attributes may be closely related. On the other hand, “median\_income” and “median\_house\_value” have a moderately strong positive correlation, which suggests that there may be a relationship between the two variables.

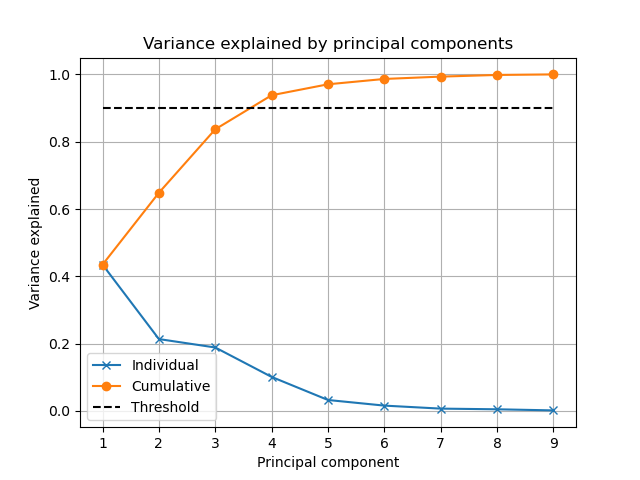
(ii) **PCA:**

Before carrying out principal component analysis on our data, we need to standardize the attributes in order to work with data of different scales (for instance, comparing dollar amounts to population). We standardize by subtracting each value by the mean of that column, then dividing by the standard deviation of the column.

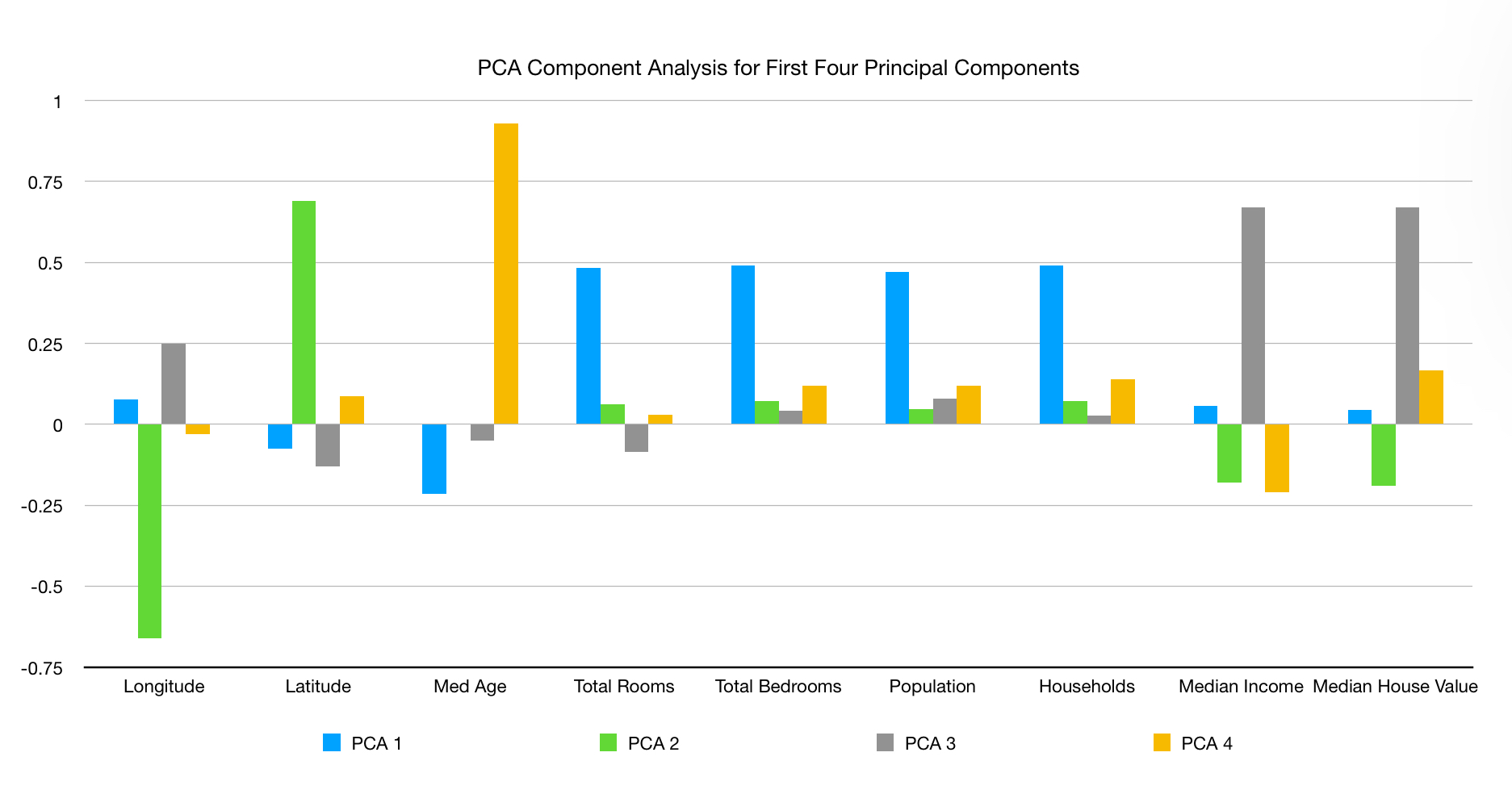
Now we can carry out PCA on our data, and determine how much variation we can convey in as few dimensions as possible. We find that 93.79% of the variation in our data can be expressed using just 4 principal components. That means that we’re able to reduce our data’s dimensionality from 9 to 4, while still maintaining almost 94% of the accuracy. If we want to increase the variance we’re able to represent in our data at the expense of increasing dimensionality, we can use the table or graph below to evaluate how much variation we can retain based on the number of PCA components we consider.

|  | **PCA 1** | **PCA 2** | **PCA 3** | **PCA 4** | **PCA 5** | **PCA 6** | **PCA 7** | **PCA 8** | **PCA 9** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variance %** | 0.4347 | 0.2136 | 0.1885 | 0.1011 | 0.03255 | 0.01586 | 0.00693 | 0.00494 | 0.001643 |

**Variance conveyed by each principal component**

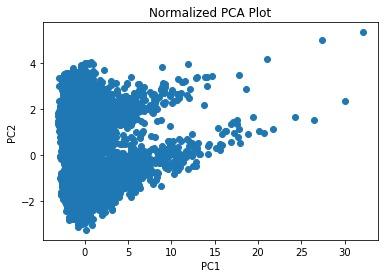


**Principal directions of the considered PCA components**



From the visual above, we are able to determine the magnitude and direction of the primary component vectors’ terms. For instance we are able to infer that the first principal component is largely determined by Total Rooms, Total Bedrooms, Population, and Households. Additionally, the similarities between those for attributes’ bars in the graph indicate that they largely influence the principal components in similar ways, which indicates that they probably measure, to some extent, the same cause of variability.

**Projection of data set onto first two principal components**

****

By projecting our higher dimensional dataset on to the two dimensional plane of perpendicular vectors PC1 and PC2, we can observe that the distribution of the data is skewed right with respect to the first principal component, but bimodal with respect to the second principal component.

#### A discussion explaining what you have learned about the data.

In this project, we focused on understanding our dataset and variables, identifying principal components to simplify a nine-dimensional dataset, and visualizing the distributions of each variable and the relationships between variables.

During the analysis of our data, we have found that there are some strong relationships between some variables. Latitude and longitude have a strong negative correlation, which we can rationalize by considering California’s landmass itself is at a diagonal with a negative slope. Additionally, total rooms, total bedrooms, population, and households are all strongly, positively related. The cause of this can be inferred that more people would need more houses and bedrooms and rooms. There is also a weaker, but yet still strong correlation between median income and median house value. This reasoning behind this can also be inferred because typically people with more money buy more expensive homes.

Because of the strong correlation between some variables, only four principal components are needed to explain over 90% of variation in the dataset, which in future projects, will allow us to represent and manipulate our data with fewer variables. It will also enable us to do regression using the principal components with high accuracy.

In the future, we plan on classifying observations regarding ocean proximity, our only categorical variable. While we have not yet run calculations, it will be interesting to see if we can classify our observations using the other attributes.

#### Mandatory Section

**Member contributions:**

| **Sections** | Description of dataset | Detailed explanation of the data attributes | Data visualization | PCA Computation | Principal component analysis (PCA) | What you learned about the data | Exam problems |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Jordan |  |  |  | 50% | 50% | 100% | 33% |
| Natalie | 100% |  |  | 50% | 50% |  | 33% |
| Ka Chun (Andy) |  | 100% | 100% |  |  |  | 33% |

**Exam Problems**

1.

Option A: False: (*Time of Day*) is not nominal

Option B: False: (*Time of Day*) is also not ratio

Option C: False: (*Time of Day*) is not ordinal

Option D: True: (*Time of Day*) is interval because its relative magnitude has a constant physical meaning. (*Traffic Lights*) and (*Running over*) are ratio because for them to be 0 implies the lack thereof. And (*Congestion level*) is ordinal, as a rating system

2.

Option A: True: dp=inf(x14, x18) =

Option B: False: dp=3(x14, x18) = != 3.688

Option C: False: dp=1(x14, x18) = != 1.286

Option D: False: dp=4(x14, x18) = != 4.311

3. Variance explained by the first n principal components

Option A: True:

Option B: False:

Option C: False:

Option D: False:

4. To solve this problem, one simply does a dot product multiplication of the correct component vector (focusing on the direction) and the signs of the observation vector to find if the resulting projection vector is positive or negative.

Option A: False: Negative

Option B: False: Negative

Option C: False: Positive

Option D: True: Positive

5. The answer is Option A: 0.153846.

For s1, the set of words is {the, bag, of, words, representation, becomes, less, parsimonious}. For s2, the set of words is {if, we, do, not, stem, the, words}. The intersection of these sets is {the, words}, and the union is {the, bag, of, words, representation, becomes, less, parsimonious, if, we, do, not, stem}. Therefore, the Jaccard similarity between s1 and s2 is 2/13 or approximately 0.153846.

6. The answer is Option B: 0.84

According to the Bayes’ Theorem,