**Group Project – California Houses**

**Course: Data Cleaning and Analysis**

**By Group 7 (Ata Avlar, Qiqi Bao, Gautier Hubert, King Man Siu, Romain Thomas)**

**Date: 15 Oct 2023**

1. **Data Exploration**

*Data Description*

The dataset, **california\_houses.csv**, represents districts in California. Each row corresponds to a distinct district, and the columns capture various attributes, including demographics, housing details, and geographic information.

*Observation*

* There are missing counts in the columns of 'Median\_Income', 'Median\_age', 'Tot\_Rooms' and 'Tot\_Bedrooms.'
* In 'Median\_Income' column, the minimal value is negative and the mean and std are both abnormally high.
* In 'Median\_Age' column, there is a minimal value of 1.000000.
* There are negative values in the 'Household' column.
* We suspected that the columns of 'Tot\_No\_Bedrooms' and 'Max\_Age' are either

irrelevant or columns with low significance.

表格

描述已自动生成

1. **Data Cleaning**

*Rationale*

* **Missing values**: missing data points can skew results and lead to inaccurate model predictions.
* **Redundant columns**: Certain columns, especially those indicating distances to various cities, were not necessary for our analysis. These were removed to reduce data dimensionality and improve efficiency.
* **Artificially generated columns**: our exploration revealed certain columns that appeared to be artificially generated and did not provide meaningful insights. These columns were removed from the dataset.
* **Handling outliers**: Outliers can significantly distort statistics and lead to misleading interpretations, adopt the interquartile range, we identify and address potential outliers.

*Before Cleaning*

* The dataset may contain visible outliers that deviate significantly from the general trend or cluster of data points.
* Missing values could lead to gaps or absences in visualizations.
* Data might show a wider range and variance.

*After cleaning*

* The dataset is cleaner and contains only relevant features.
* we can better understand the variables in the data frame.
* Missing data points have been addressed to ensure the dataset is complete.
* Potential outliers are managed to ensure more representative statistics.

*Feature1, representing a set of values that are randomly generated using a normal distribution, values primarily centered around 1000 with a spread determined by a standard deviation of 100.*

*Feature2, representing another set of values that are randomly generated using a normal distribution, values primarily centered around 500 with a spread determined, by a standard deviation of 50, and containing some missing data points.*

图表, 散点图

描述已自动生成

1. **Geo-data Handling**

*Create the Closest City Column*

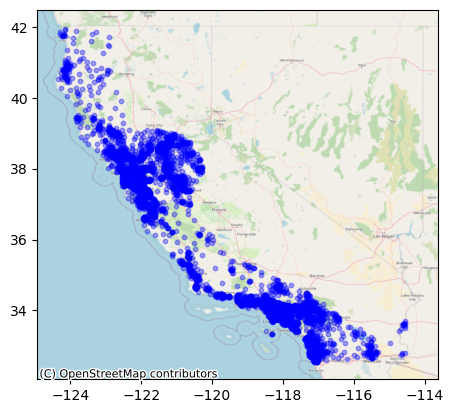
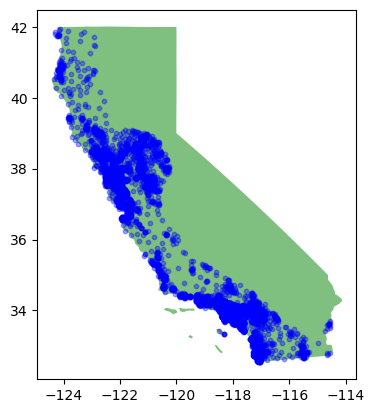
* + We determined which city of California (LA, SD, SJ, or SF) is the closest to each district in the dataset.
  + The data for each district's distance to these cities is present in columns named "Distance to\_<City Name>". The result adds a new column "Closest\_city" to the dataset that indicates which of the four cities is the nearest to that district.

图形用户界面, 表格

描述已自动生成

* The "Closest\_city" column has a total of 20,640 entries.
* There are 7, 592 districts close to Los Angeles, 3,753 districts close to San Francisco, 2,330 districts close to San Jose and 1,623 districts close to San Diego.

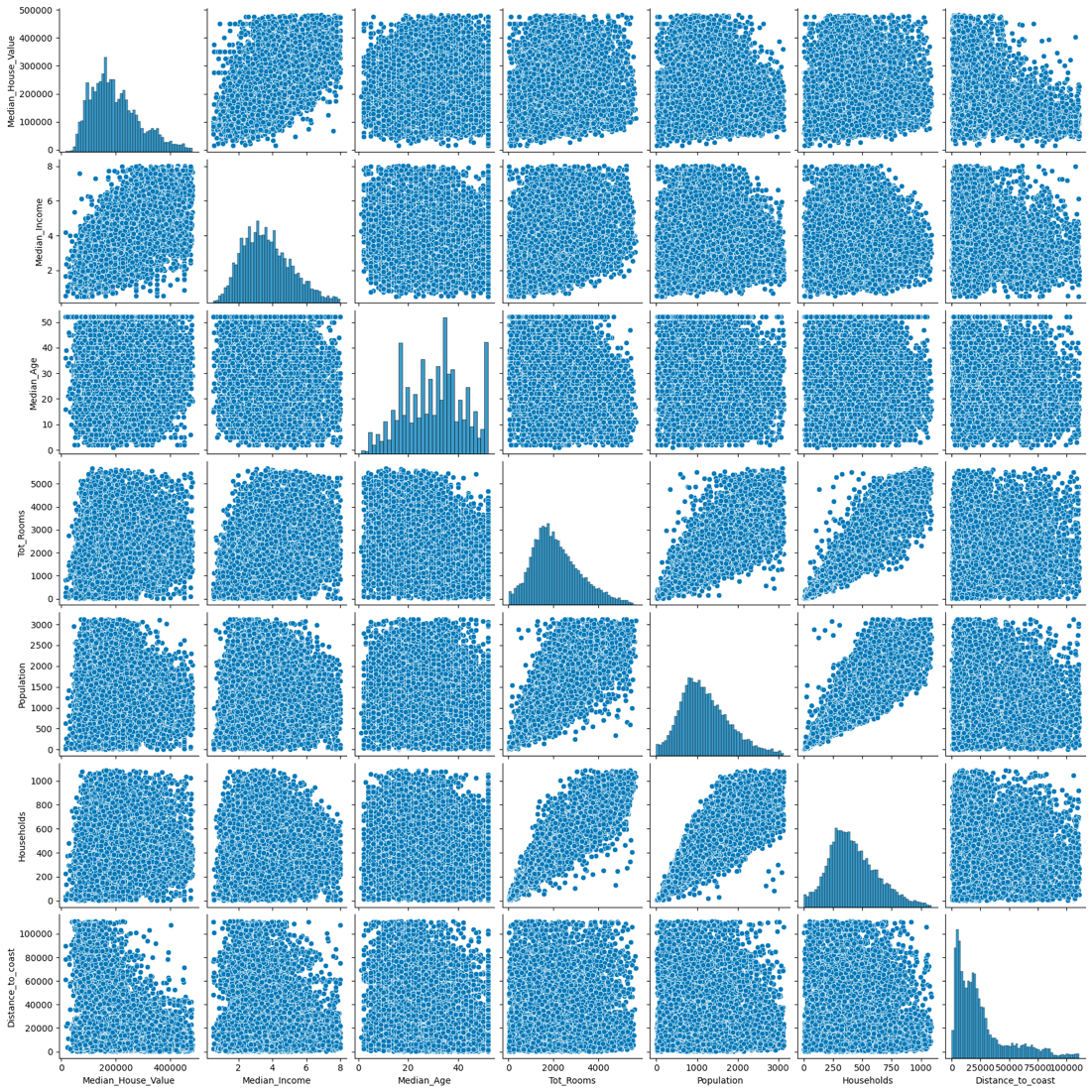
With the data above, we plotted the following graph with geopandas package to understand the district distribution on the map of California.

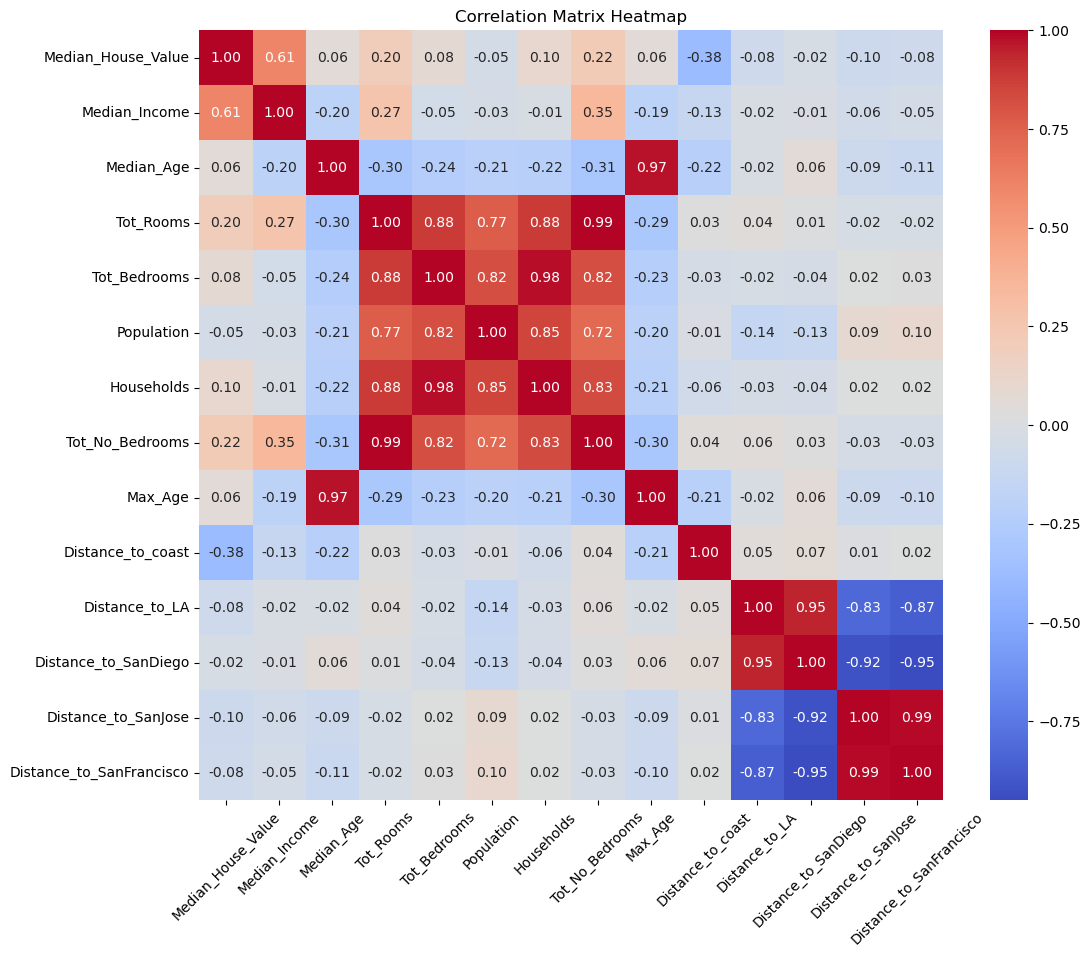


**4. Multivariate analysis**

*Scatterplot and Correlation matrix of all the quantitative variables*

We made a scatterplot matrix and a correlation matrix heatmap using seaborn to understand the correlation between variables.



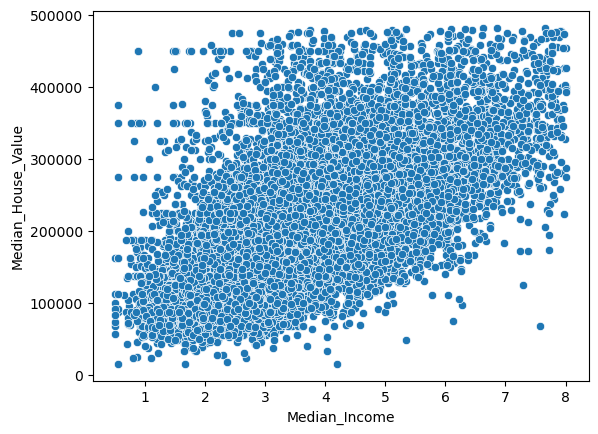


*Observations*

* We decided to drop columns "Tot\_No\_Bedrooms" since max age and median age are highly correlated that may be duplicated with a high correlation coefficient (0.97).
* With a similar observation, Tot\_Rooms and Tot\_No\_Bedrooms are highly correlated too with a correlation coefficient of (0.99).
* We found that the total number of rooms or bedrooms per district is not quite meaningful, as they have a very high correlation value of 0.93. We suggest that one of the columns can be dropped.
* We suggest using the meaning of the number of bedrooms to understand the average in correlation to the other variables, such as house value, income, population and household.
* From the plots, we also figured out that the median income and distance to coast have a more significant relationship with the median house value.

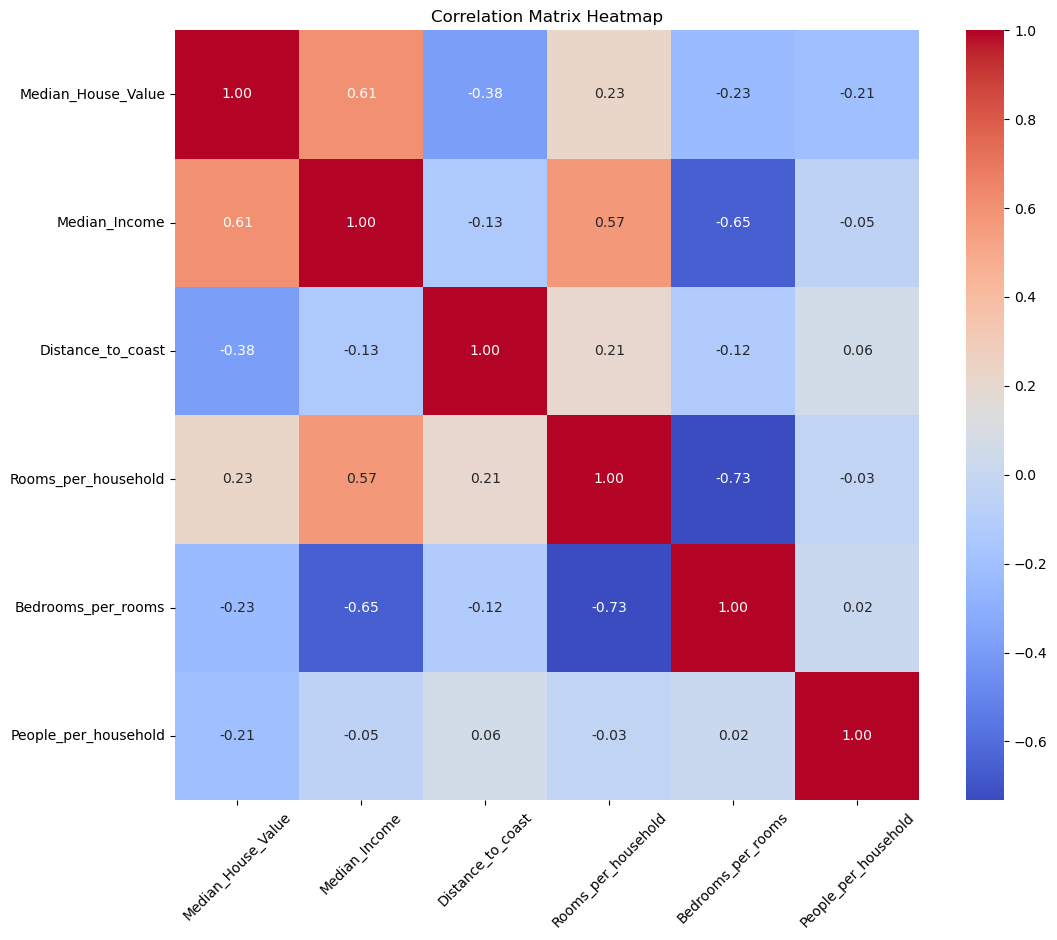
*Targets and Predictors*

* Median House Value is chosen as the target here, because the independent variables in the data set potentially affect the values of houses, such as income, age of district, number of rooms, population density, household size (single or family) and distance from the coast. In other words, Median House Value depends on these variables, and it has a higher business value to study.
* From the plots, we observed that the median house value has a positive correlation with median income (0.61). It suggests that as the median income in a district increases, the median house value also tends to increase. Higher income areas likely have a higher demand for housing, driving up prices. This predictor is the most significant among all, with a very strong positive correlation. Please refer to the scatterplot below.



## Explore new variables

We created three new variables and a new correlation matrix heatmap.



* The correlation between the median house value and rooms per household becomes slightly stronger with a weak positive correlation (0.23).
* The correlation between the median house value and bedrooms per rooms is weakly and negatively correlated (-0.23).
* The correlation between the median house value and people per household is weakly and negatively correlated (-0.21)

## Conclusion

**Data Exploration**: Our initial examination of the California housing dataset unveiled a plethora of attributes, each offering a unique facet of understanding. From median incomes to housing medians and geographic coordinates, the data painted a comprehensive picture of the housing landscape in the state. Visualizations further aided in grasping the distributions and relationships inherent within the data.

**Data Cleaning**: Rigorous data cleaning was fundamental in ensuring the dataset's integrity and reliability. By systematically addressing missing values, outliers, and potential inconsistencies, we bolstered the dataset's foundation, preparing it for more refined analyses and modeling tasks.

**Geo-data Handling**: Delving into the geographic nuances, we discerned the spatial distribution of districts relative to major cities and the coastline. The pronounced clustering around urban centers and coastal regions underscored the significance of location in the housing dynamics of California.

**Multivariate Analysis**: The interplay between multiple variables shed light on the intricate relationships shaping the dataset. In particular, the robust correlation between median incomes and house values accentuated the pivotal role of economic factors in housing valuations.

**New Variables**: The introduction and exploration of novel features derived from existing data enriched our dataset further. These new variables reveal insights into living conditions, housing demand, and district characteristics.

In summary, our holistic exploration and preparation of the California housing dataset have unveiled a multitude of insights, laying a robust groundwork for future modeling, analyses, and predictions. The multifaceted nature of the data, encompassing economic, geographic, and demographic dimensions, offers a rich tapestry of understanding, promising valuable revelations in subsequent analytical endeavors.