**Property Price of California Data Analysis & Visualization**

Ravi Kumar Vedula

Mahesh Kumar Panuganty

Data Visualization

Professor: Dr. Anna Valeva

November 21, 2021

# Introduction

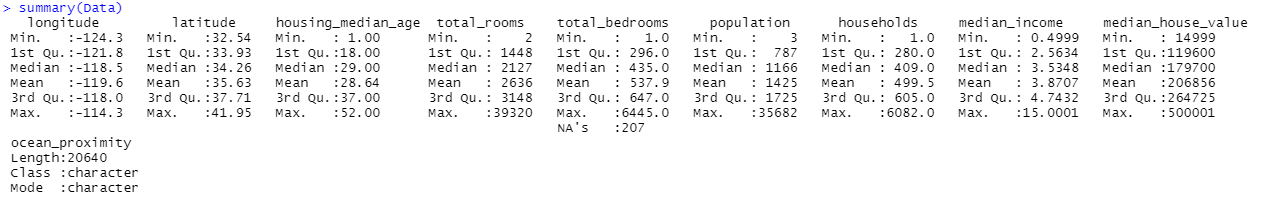
The information in the report comes from the California census taken in 1990. Consequently, although it may not be as useful for forecasting current house values as the Zillow Zestimate dataset, it does serve as an easy introduction to machine learning for anyone interested in understanding the fundamentals of machine learning. The information refers to the residences that can be located in a certain California district, as well as some basic statistics about them that are based on the 1990 census data. Please be aware that the data has not been cleansed and that some preparation procedures will be necessary! Here is a list of the columns, and their titles should be self-explanatory: longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, median house value, and ocean proximity.

# EDA

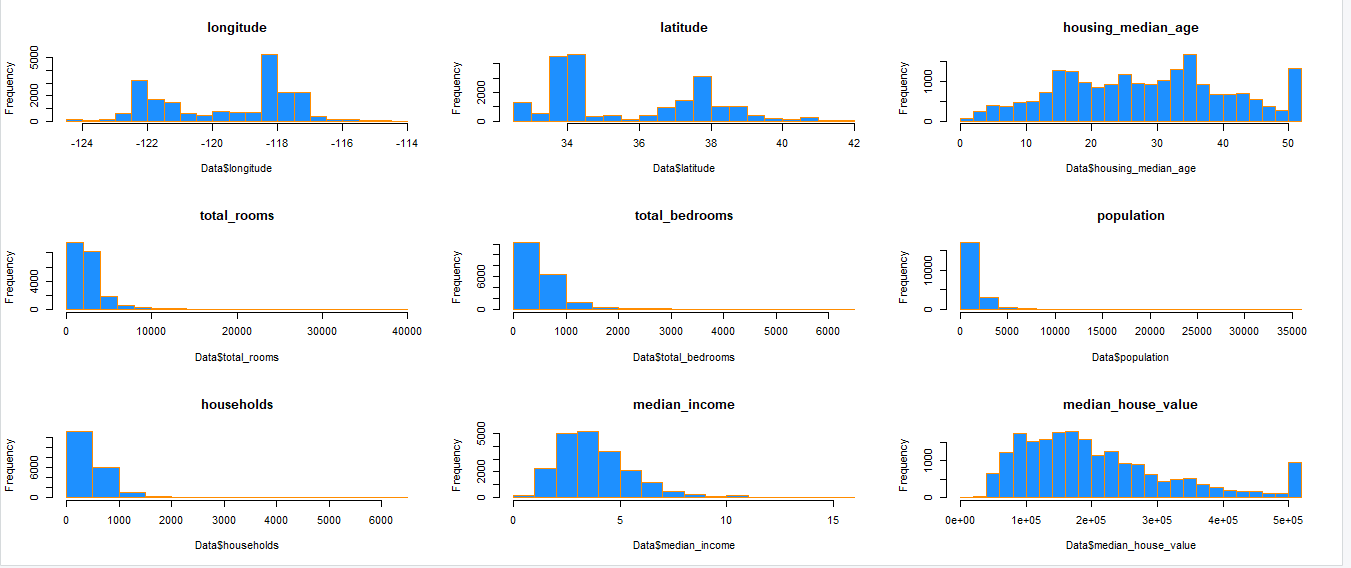
Performing exploratory data analysis, often known as EDA, is a critical stage in any Data Analysis or Data Science endeavor. When studying a dataset for patterns and anomalies (outliers), EDA is the process of developing hypotheses that are based on our knowledge of the dataset.

EDA entails the generation of summary statistics for numerical data in a dataset as well as the creation of different graphical representations to help the user better comprehend the data. With the use of an example dataset, we will be able to better comprehend EDA in this post. Here the R programming and Tableau software tools will be used for making visualizations.

## Summary of Dataset

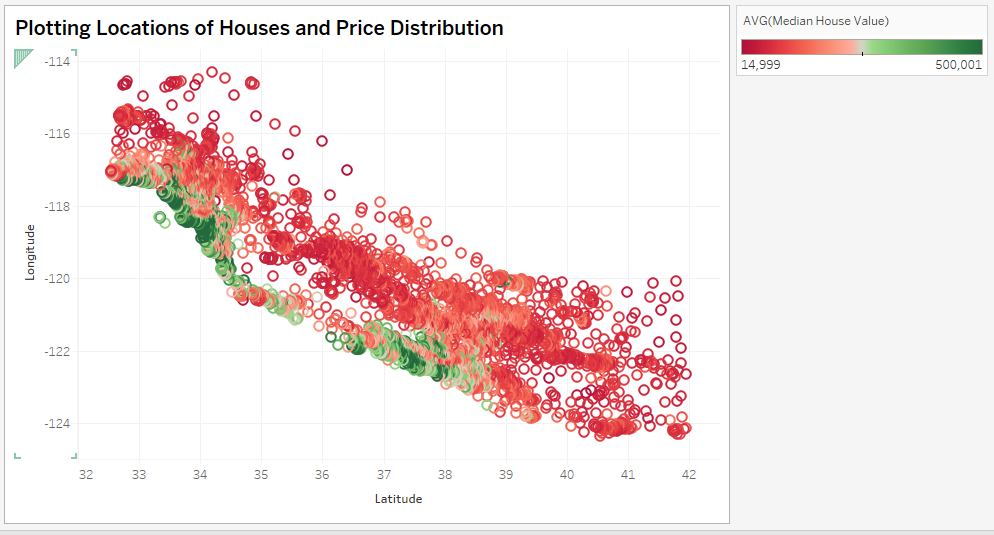


## Histogram of Variables



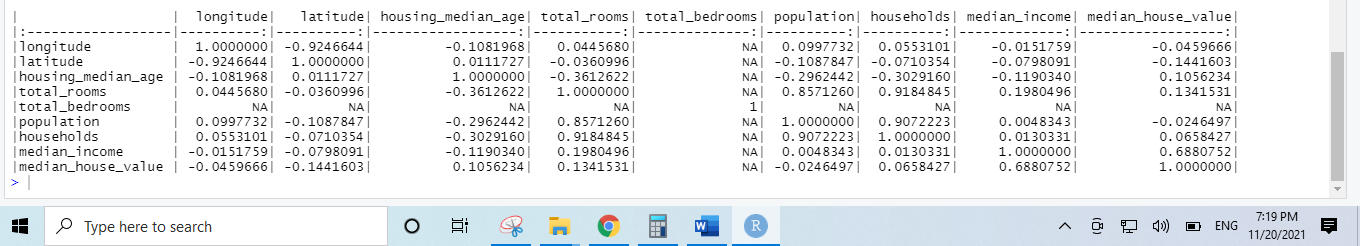
## Plotting Locations of Properties

The median home value of the houses closest to the beach is greater on average, but the median house value of the houses farther inland is lower. This is a significant difference, and it indicates that the variable ocean proximity will most likely play a significant role in forecasting median home value in the future. While it is true that some ocean-adjacent locations, such as the more isolated towns along California's northernmost coast, have lower median house prices than other coastal housing, some inland locations, such as the towns near popular tourist destinations such as Lake Tahoe or California's inland capital Sacramento, have higher median house prices than other inland housing. Furthermore, it is noteworthy that the homes nearest to major urban regions such as San Francisco and Los Angeles have among the highest median house prices, whilst the houses closest to the primarily agricultural Central Valley have among the lowest median property values.



## Correlation Matrix

It seems as if there are some more variables that may not be required owing to collinearity in the data. Correlation values are shown in the below given table. (Greenhill & Venkatesh, 2011) In this table the coefficient value 1 very high positive correlation and zero means very poor correlation.



# Conclusion

At the end of this research, the EDA on the selected dataset has been performed successfully. All the visualizations and explanations are documented here. From the performed EDA that is clear that the dataset contains zero null values. Also, the report contains information about the distribution of each variable.

# References

Greenhill, S., & Venkatesh, S. (2011). Data modelling and visualization using Noetica. *Data & Knowledge Engineering*, *33*(3), 241-276. https://doi.org/10.1016/s0169-023x(00)00003-3