# **Linear regression on California housing dataset**

***Abstract***

Linear regression is a commonly used technique for predictive analysis, forecasting data and error reduction. The following report focuses on the methods used for pre-processing the data and analysing attributes of the California housing dataset.

1. **Introduction**

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables. It can be used to fit a predictive model to an observed [data set](https://en.wikipedia.org/wiki/Data_set) of values of the response and explanatory variables. A fitted linear regression model can be used to identify the relationship between a single predictor variable and the response variable when all the other predictor variables in the model are "held fixed".

1. **Dataset**

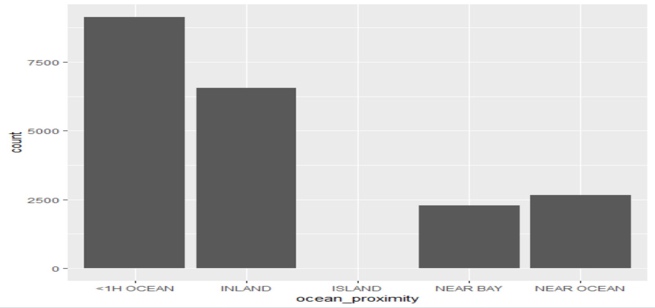
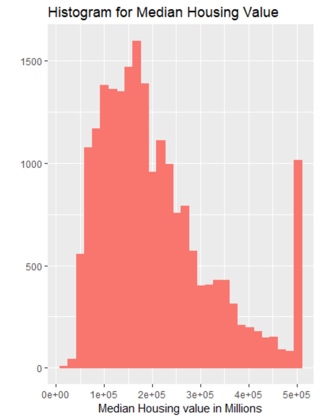
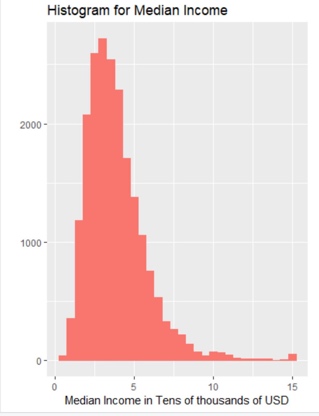
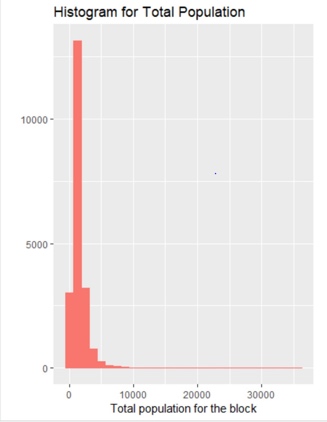
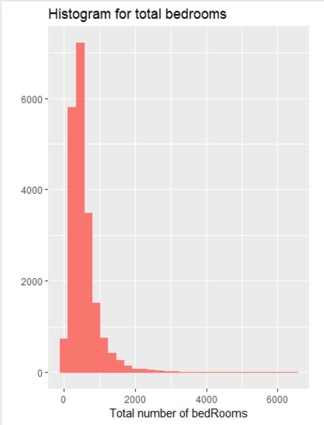
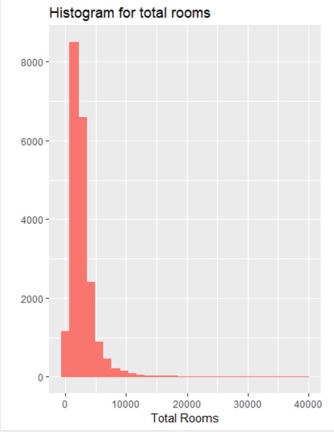
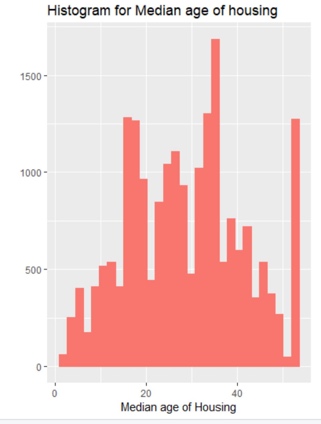
California housing dataset pertains to the houses found in a given California district and some summary stats about them based on the 1990 census data.

The column names and their data types are as follows, their names are self-explanatory:

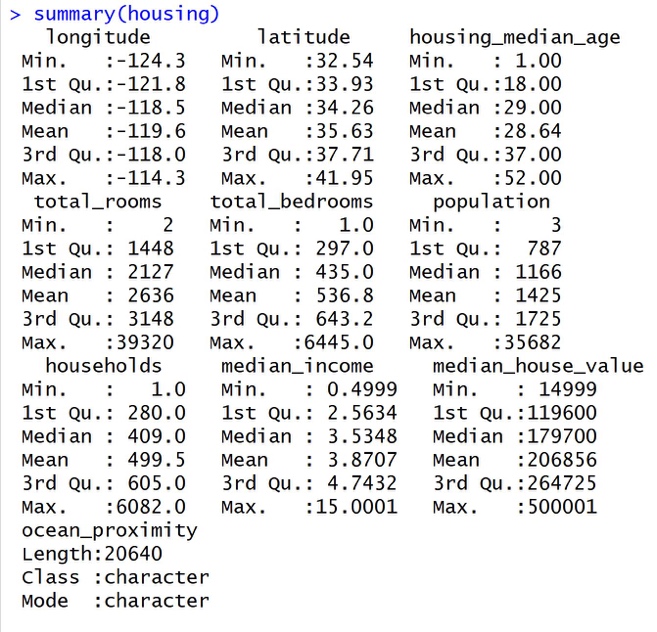
|  |  |
| --- | --- |
| *Column Name* | *Data Type* |
| Longitude | Numerical |
| Latitude | Numerical |
| Housing\_median\_age | Numerical |
| Total\_rooms | Numerical |
| Total\_bedrooms | Numerical |
| Population | Numerical |
| Households | Numerical |
| Median\_income | Numerical |
| Median\_house\_value | Numerical |
| Ocean\_proximity | Categorical |

1. **Attribute analysis**

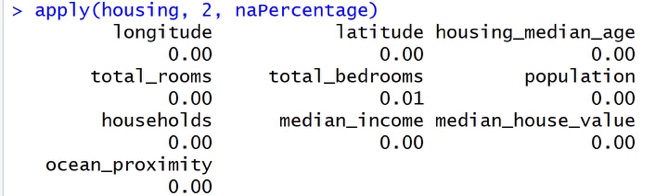
The California housing data distribution:



Summary of the dataset is as follows,



We have observed that the initial dataset contained null values. The distribution of null values is as follows:



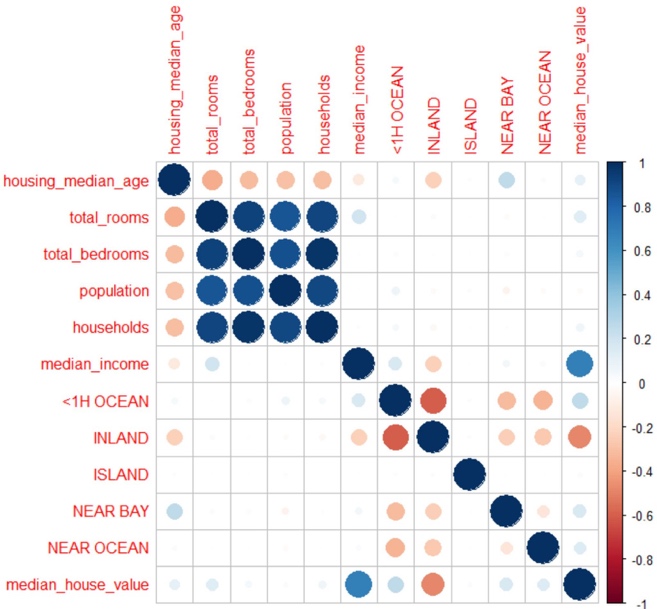
The null values in *total\_bedrooms* where replaced by the mean value of that attribute.

Ocean proximity was categorical data, So the attribute values were spread and converted to numerical data. Further, the data was normalized using standard normalization.

Normalized Data:



Correlation Plot - Circle:



Correlation Plot - Values:

