REPORT DATA ANALYSIS CAPSTONE FINAL

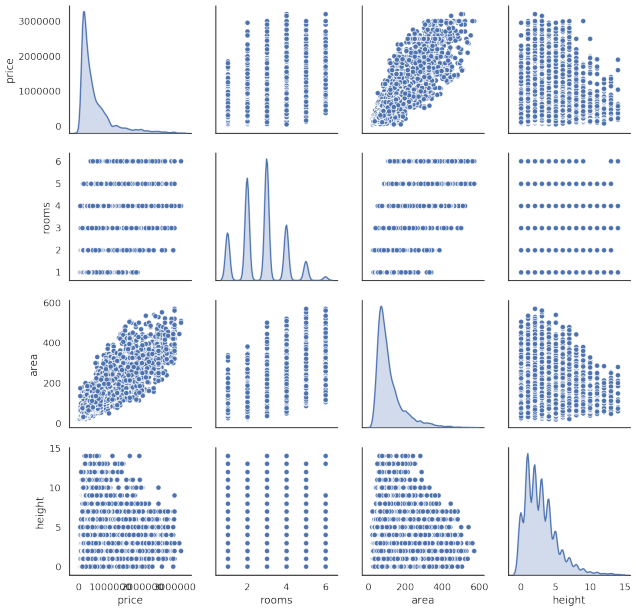
REPORT Xavier Martínez

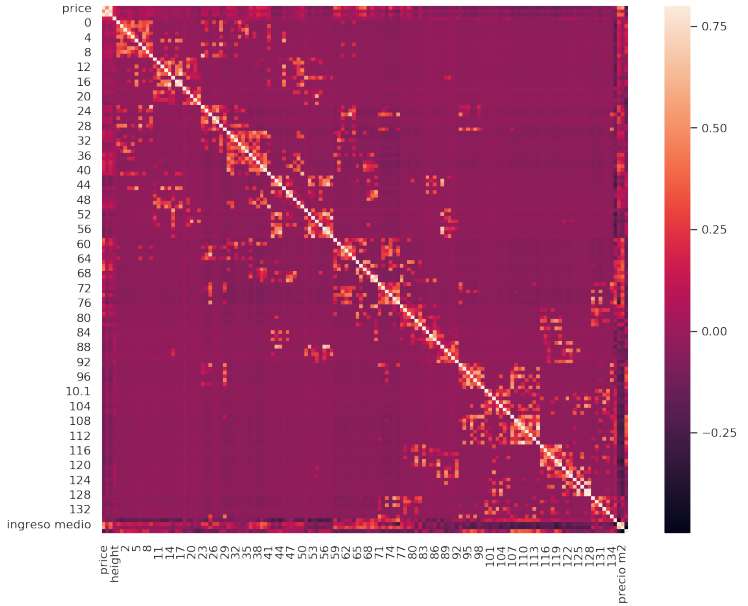
1. DESCRIPTION OF THE PROBLEM

- Now we are going to predict and analyze the price of buildings in Madrid. After doing a web scraping, we will show the data and basic statistics of it. We will use the tools studied during the course, such as seaborn, matplotlib or pandas.

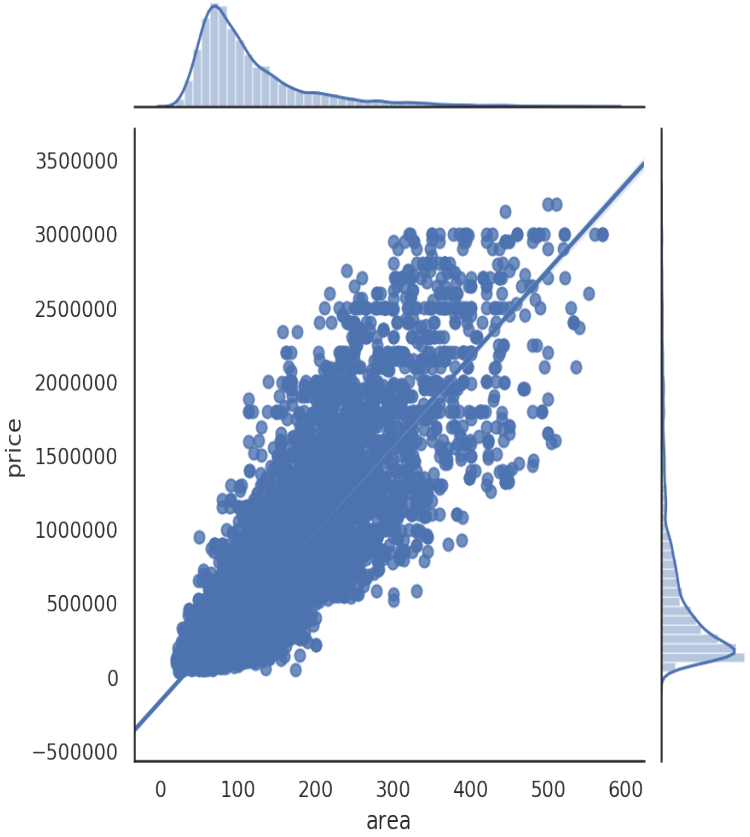
2. DESCRIPTION OF THE DATA

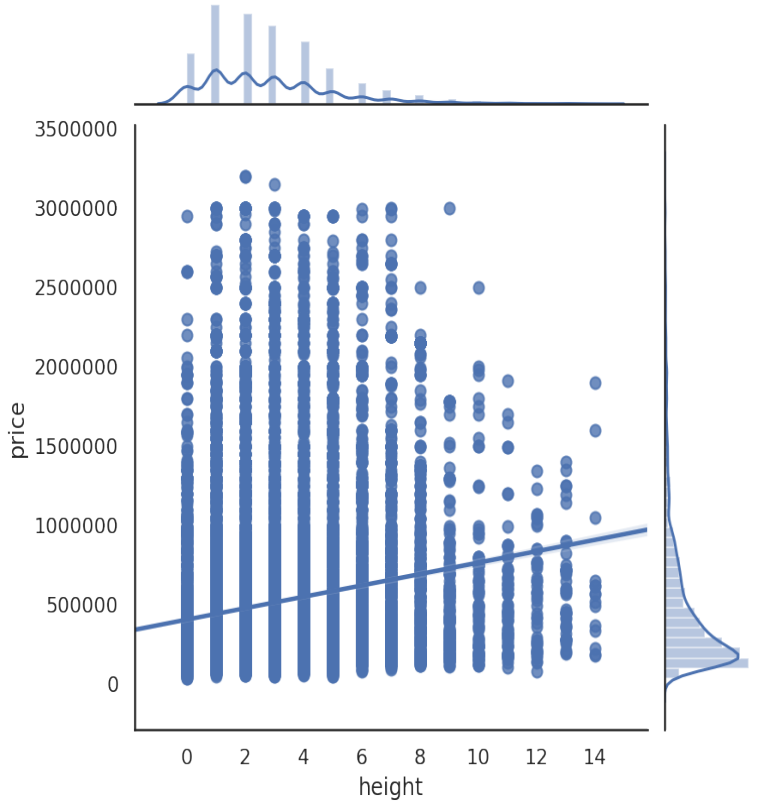
- We have approx. 20.000 different buildings and we are going to predict the label, which is the price. We observe some basic characteristics and we observe that some of the features employed are the number of rooms, the crime rate per area, the height, the area and the numbers from 0 to 134 are the different neighborhoods of Madrid

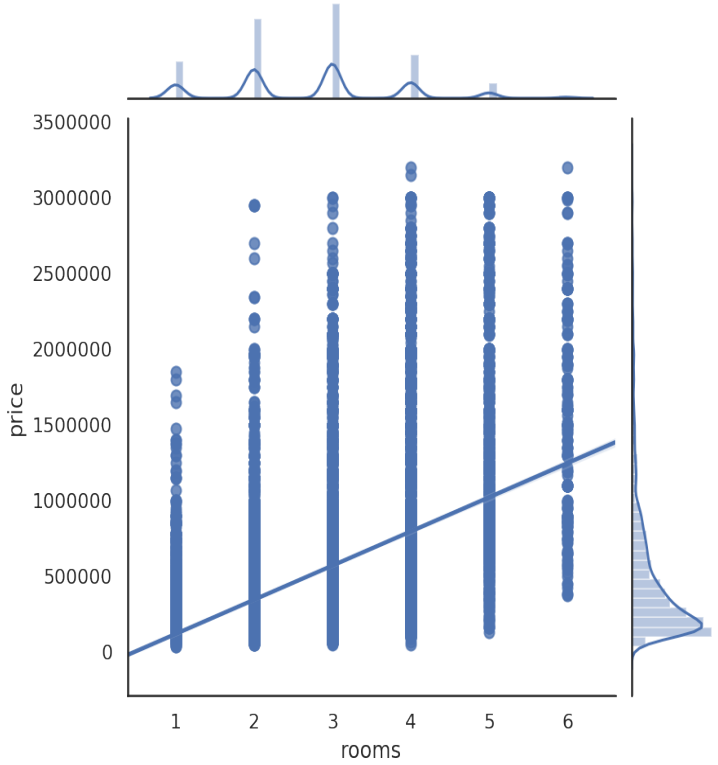


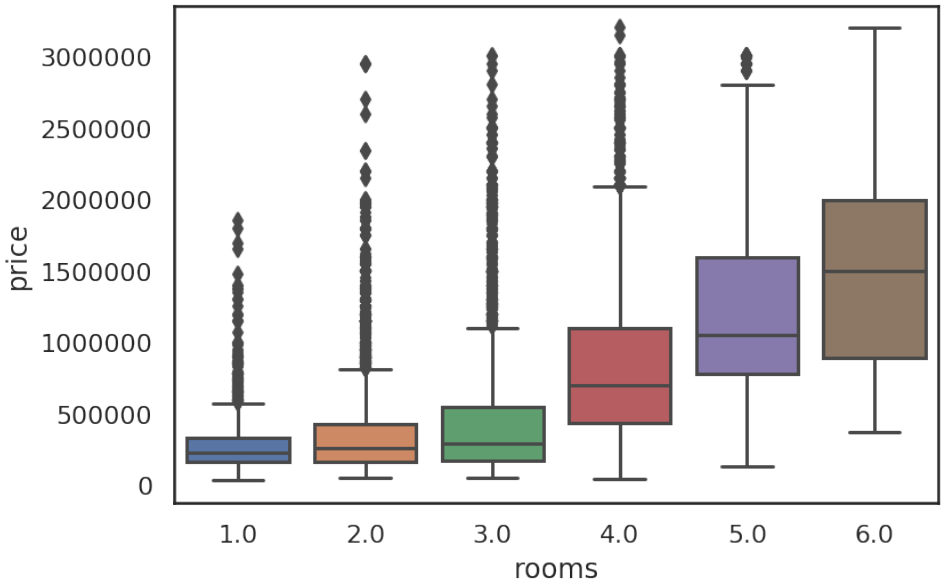


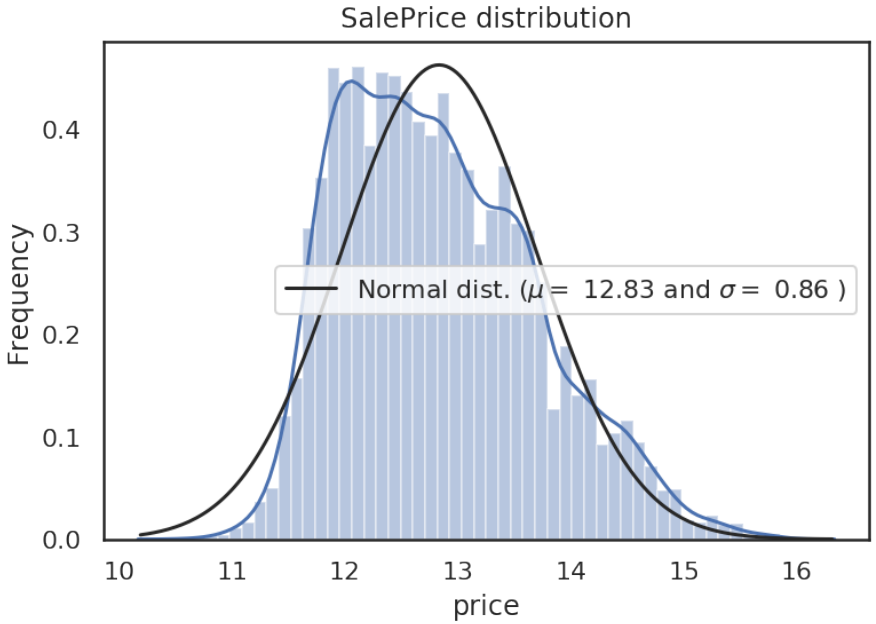
DISTRIBUTIONS











BOXPLOT AND NORMALIZED DISTRIBUTION

3. INTERESTED AGENTS

- It would be interesting for real estate investors, it will help to predict the price of the buildings

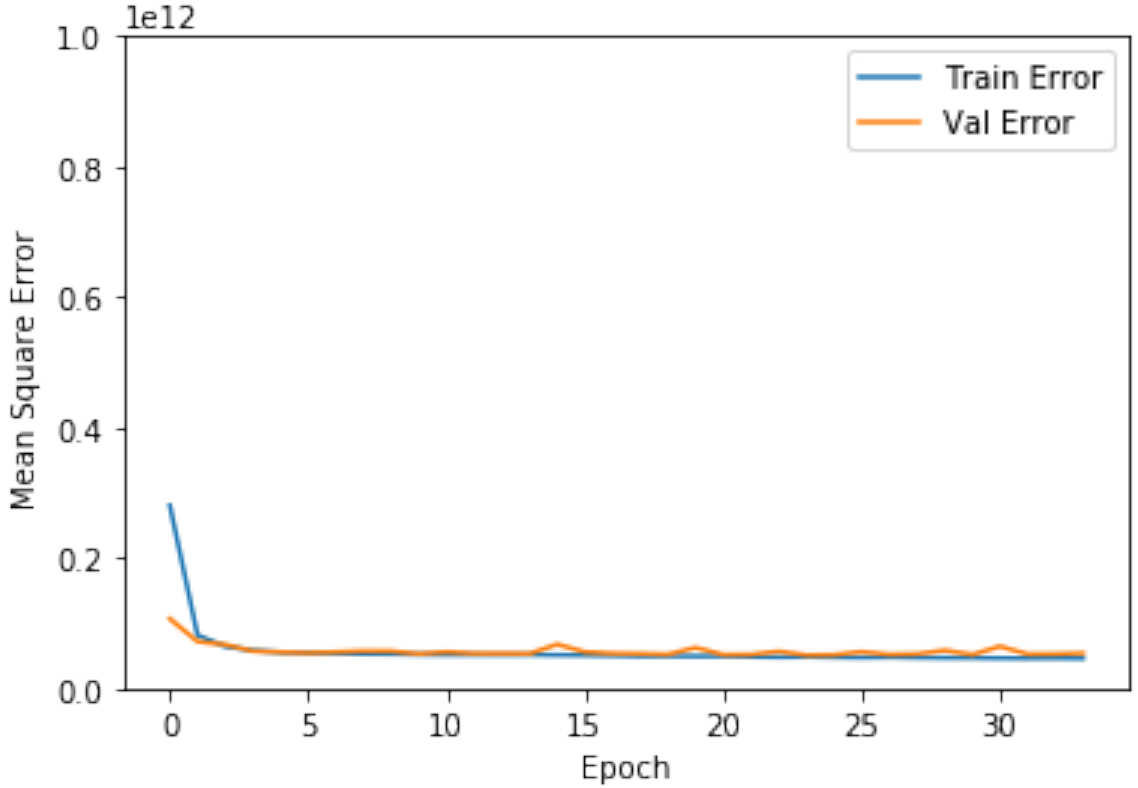
4. METHODOLOGY

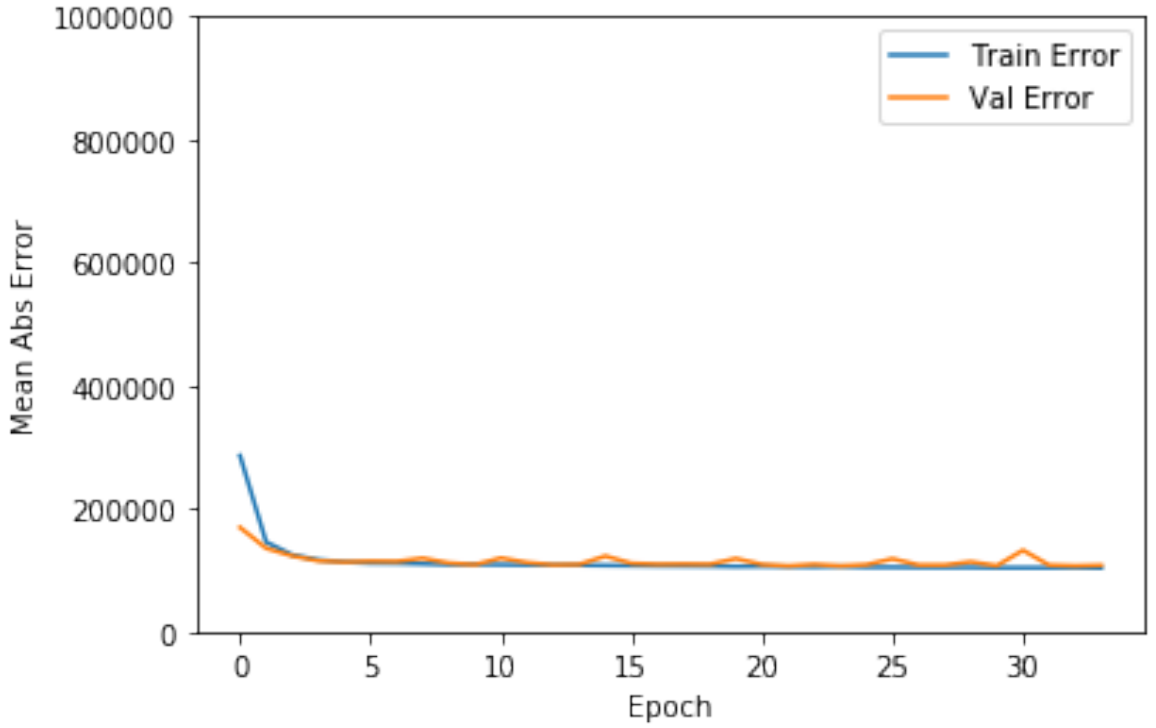
- The description of the data is shown by using describe(). Also explanatory data analysis is observable and it is important to delete anomalies that could negatively affect the result of the prediction. The method used is neural networks
- We split data into train and test, this data is normalized and the structure of the neural network could be observable in the following image.

```python
def build_model():
  model = keras.Sequential([
    layers.Dense(256, activation=tf.nn.relu, input_shape=[len(train_dataset.keys())]),
    layers.Dense(128, activation=tf.nn.relu),
    layers.Dense(128, activation=tf.nn.relu),
    layers.Dense(64, activation=tf.nn.relu),
    layers.Dense(64, activation=tf.nn.relu),
    layers.Dense(1)
  ])

  optimizer = tf.keras.optimizers.RMSprop(0.001)

  model.compile(loss='mean_squared_error',
                optimizer=optimizer,
                metrics=['mean_absolute_error', 'mean_squared_error'])
  return model
model = build_model()
model.summary()
```

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_12 (Dense)             (None, 256)               36608
_________________________________________________________________
dense_13 (Dense)             (None, 128)               32896
_________________________________________________________________
dense_14 (Dense)             (None, 128)               16512
_________________________________________________________________
dense_15 (Dense)             (None, 64)                8256
_________________________________________________________________
dense_16 (Dense)             (None, 64)                4160
_________________________________________________________________
dense_17 (Dense)             (None, 1)                 65
=================================================================
Total params: 98,497
Trainable params: 98,497
Non-trainable params: 0
_________________________________________________________________
```

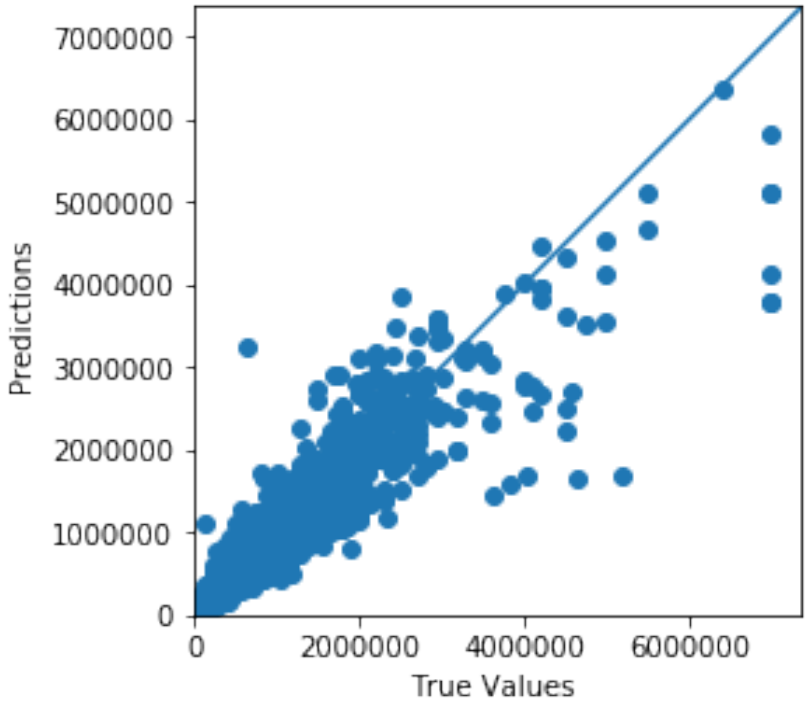
NEURAL NETWORK TRAINING

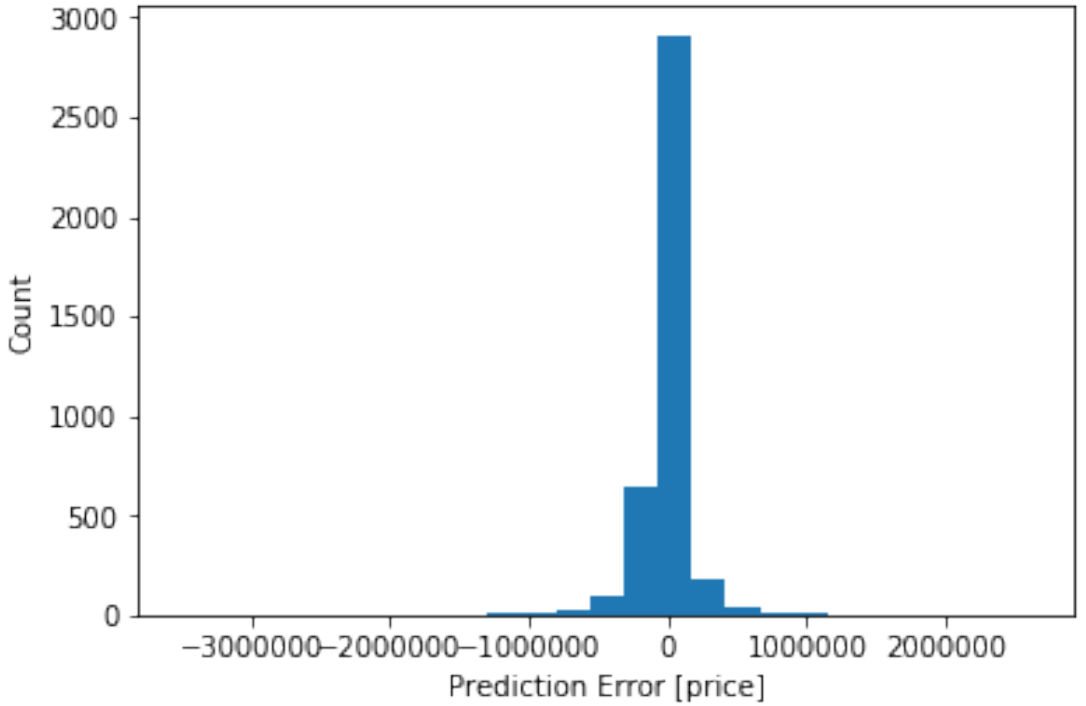




5. RESULTS

- It is observable the prediction in comparison with the true values, the error is quite Gaussian and we have a good model that is able to predict the price of the houses.





6. CONCLUSION

- The conclusion is that using web scraping and this kind of analysis we can predict the price per area in Madrid.