Question 1:

- Load an original dataset,
- comment on this dataset (what are the features, the dimensions of the problem,
- how does the correlation matrix looks like)

We choose to work with the California Housing Dataset from sklearn as this is a real-world, but still relatively clean, dataset and needs little preprocessing. The set has 20640 entries of 9 variables each

```
Shape: (20640, 9)
```

Number of Attributes: 8 numeric, predictive attributes and the target

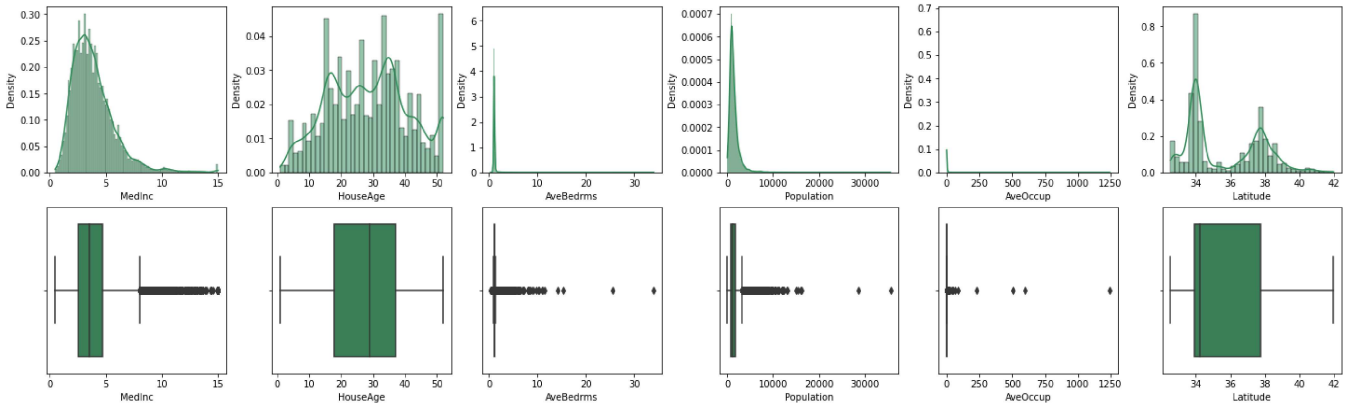
Attribute Information:

- MedInc median income in block group
- HouseAge median house age in block group
- AveRooms average number of rooms per household
- AveBedrms average number of bedrooms per household
- Population block group population
- AveOccup average number of household members
- Latitude block group latitude
- Longitude block group longitude

Consider some summary statistics for each feature:

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	L
count	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	20640.000000	206
mean	3.870671	28.639486	5.429000	1.096675	1425.476744	3.070655	35.631861	-1
std	1.899822	12.585558	2.474173	0.473911	1132.462122	10.386050	2.135952	
min	0.499900	1.000000	0.846154	0.333333	3.000000	0.692308	32.540000	-1
25%	2.563400	18.000000	4.440716	1.006079	787.000000	2.429741	33.930000	-1
50%	3.534800	29.000000	5.229129	1.048780	1166.000000	2.818116	34.260000	-1
75%	4.743250	37.000000	6.052381	1.099526	1725.000000	3.282261	37.710000	-1
max	15.000100	52.000000	141.909091	34.066667	35682.000000	1243.333333	41.950000	-1

The AveRooms column lists the maximum as 141.9. Even billionaires don't often have an average of 141 rooms in their mansions, so we investigate the existence of outliers.

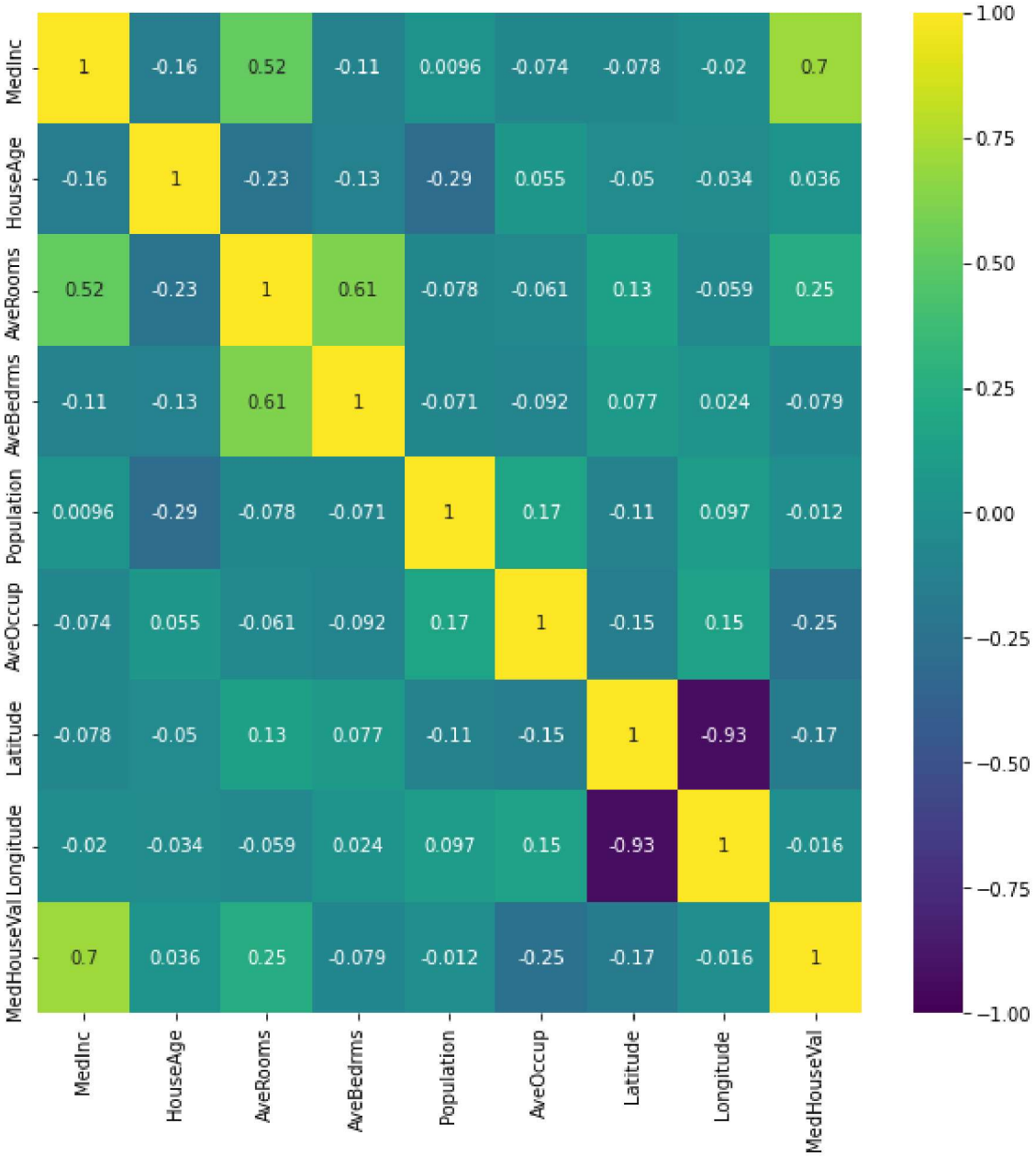


There are also definitely outliers present in the data. Except for these, the features seem to broadly be distibuted as Gaussians, with the exception of the geographical information which is bimodal and perhaps rather a mixture.

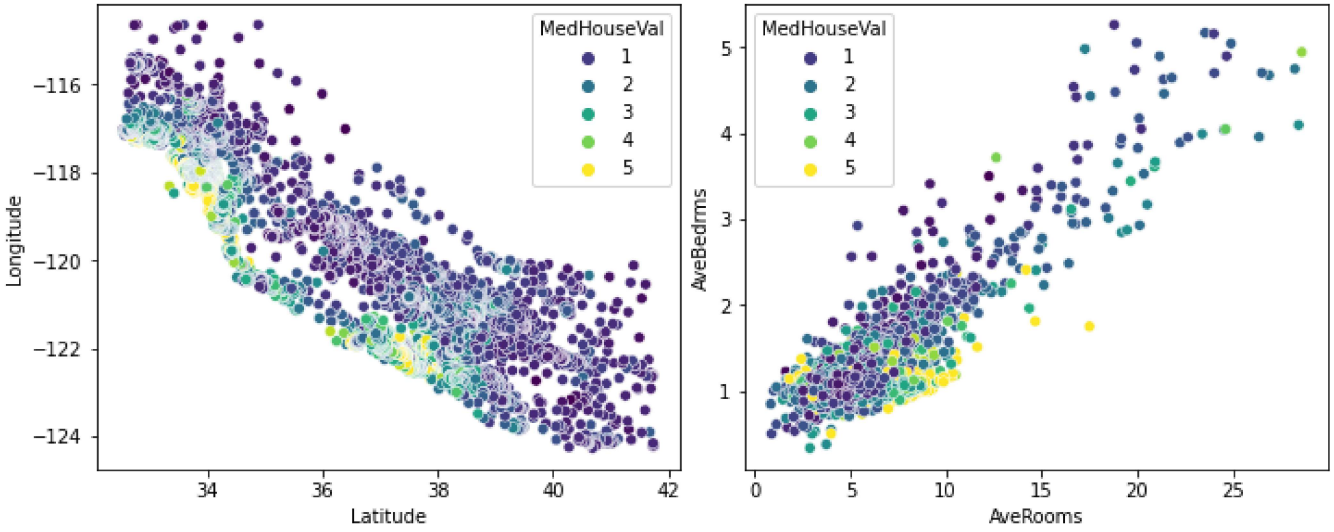
To this end, we remove the 99.9 percentile, which removes very few entries but gets rid of the most extreme outliers.

```
Removal of 0.069% of entries
```

We now consider the correlation matrix



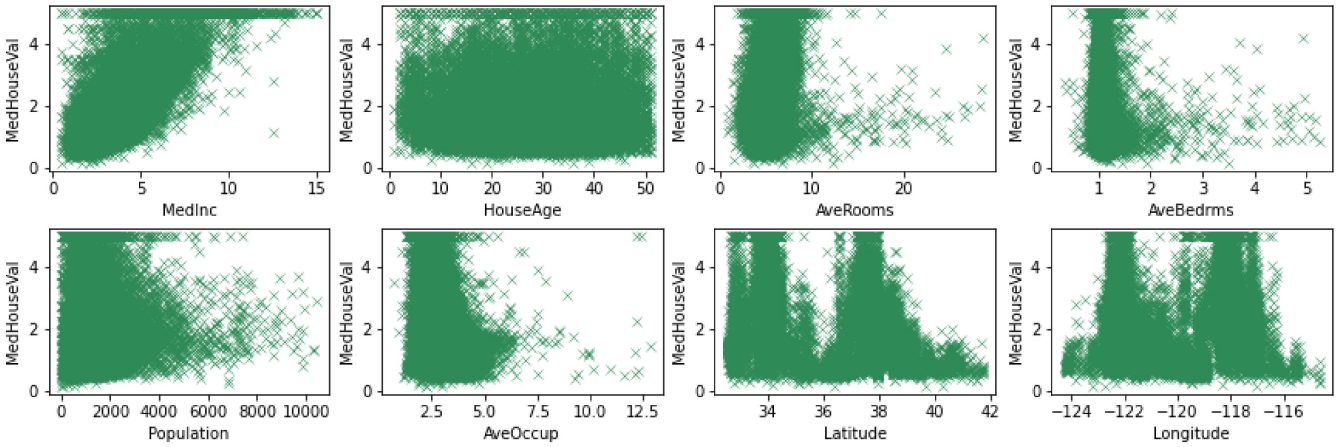
Except for two pairs of variables (Lat and Long, AveRooms and AveBdrms), there seems to be no strong correlation between the variables. Let us further explore the relationship between these two pairs.



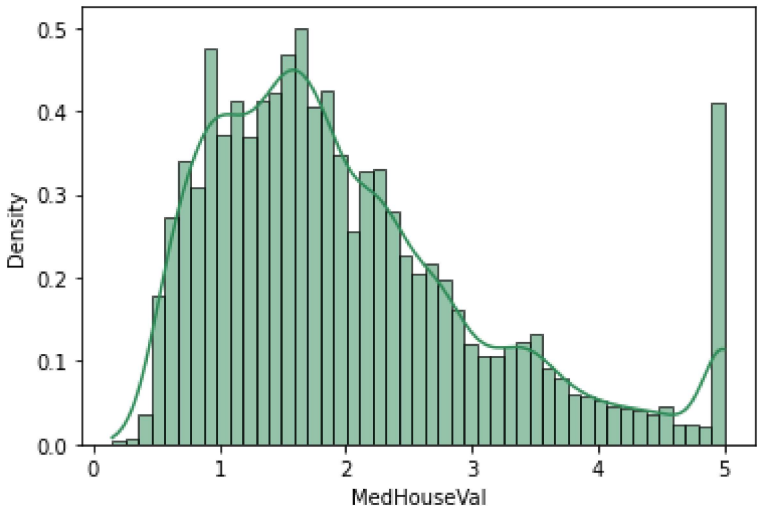
The data here represents housing information from California, which is a state on the West Coast of the US; the geography of the area considered explains the almost linear relationship between the 2 geographical features as it is immediately obvious that homes along the coast tend to be more expensive.

We also note that there is a strong positive correlation between the 2nd pair of variables. It seems that homes with a higher Rooms-to-Bedrooms proportion tend to be more expensive. This also makes sense as it would allow for alternatives such as sitting rooms and dining rooms to be included.

We now consider the relationship of the features w.r.t the y-label



As mentioned before, the geographical features seem to be bimodal. Finally, consider the distribution of the label, Median House Value which seems to be broadly follow a Poisson distribution, but with a single outlying mode to the right.



We now pass to the data-splitting. We first normalise the data and then, since this is a relatively large dataset, split it by the 80:20 ration into a train and test set. Finally, we add a column for the intercept so that the model $y = Ax + b$ may be succinctly rewritten as $y = Ax$

Question 2

Here we consider the minimisation problem

$$\min_{x \in \mathbb{R}^d} f(x)$$

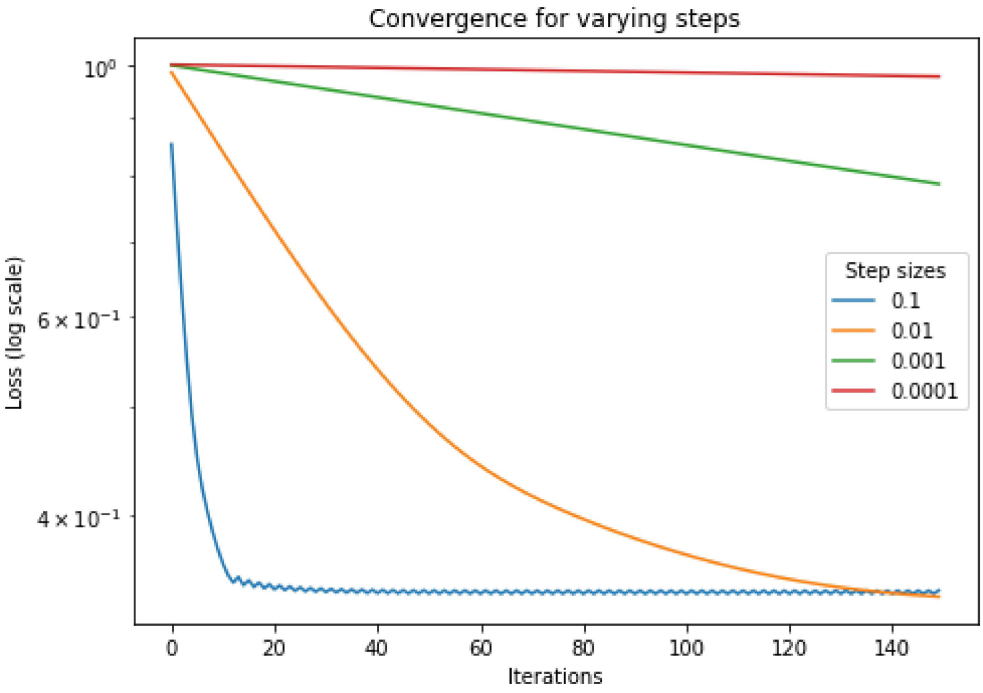
where

$$f(x) = \frac{1}{2n}\|Ax - y\|^2 + \frac{\lambda}{2}\|x\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(Ax_i - y_i)^2 + \lambda \sum x_i^2$$

and the gradient writes as $$ \nabla{f} = \frac{1}{n}A^{T}(Ax-y) + \lambda x

- _Implement gradient descent for regression (ℓ_2-loss) with a small ridge penalty._

- *Display the convergence rate on the training loss for several fixed step sizes.*



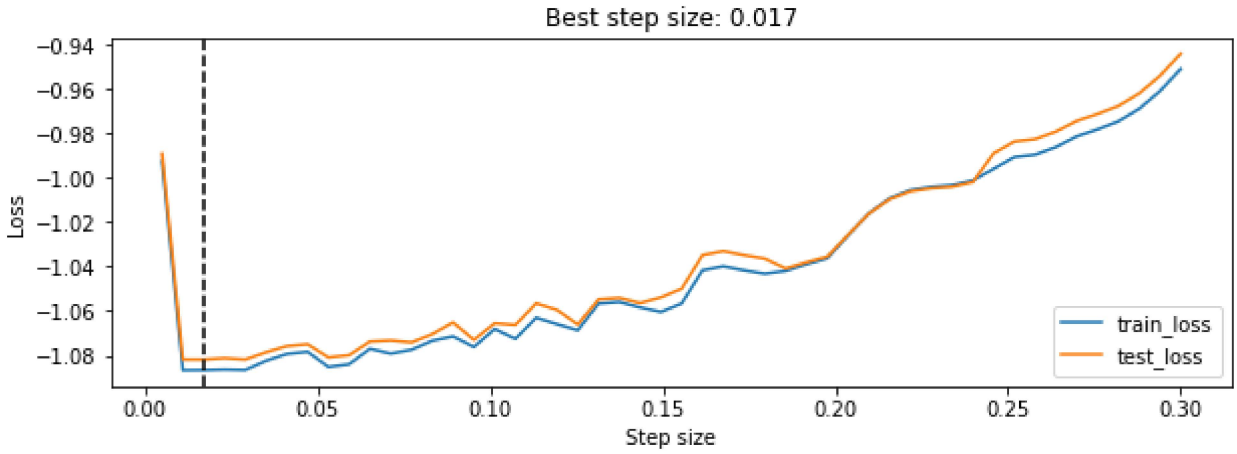
We consider 5 different step sizes α for a penalty of $\lambda = 10^{-5}$. When this optimisation problem has been regularised as above, we know that the system has a unique solution and is guaranteed to converge for a step size that is small enough.

We see above that a step size that is too small inhibits the speed of convergence. As the step sizes decrease, the method converges faster and for a step size of 0.1, the method is close to convergence after only approximately 30 iterations.

Question 3:

What is the optimal step choice ? How does this compare with the theory ?

We run GD for a number of different step sizes and plot the resulting train and test loss.



The best step size seems to occur at $\alpha = 0.017$

The theory states that the optimal step size is at $\alpha = \frac{2}{||X||}$ where $||\cdot||$ represents the matrix norm. For this problem:

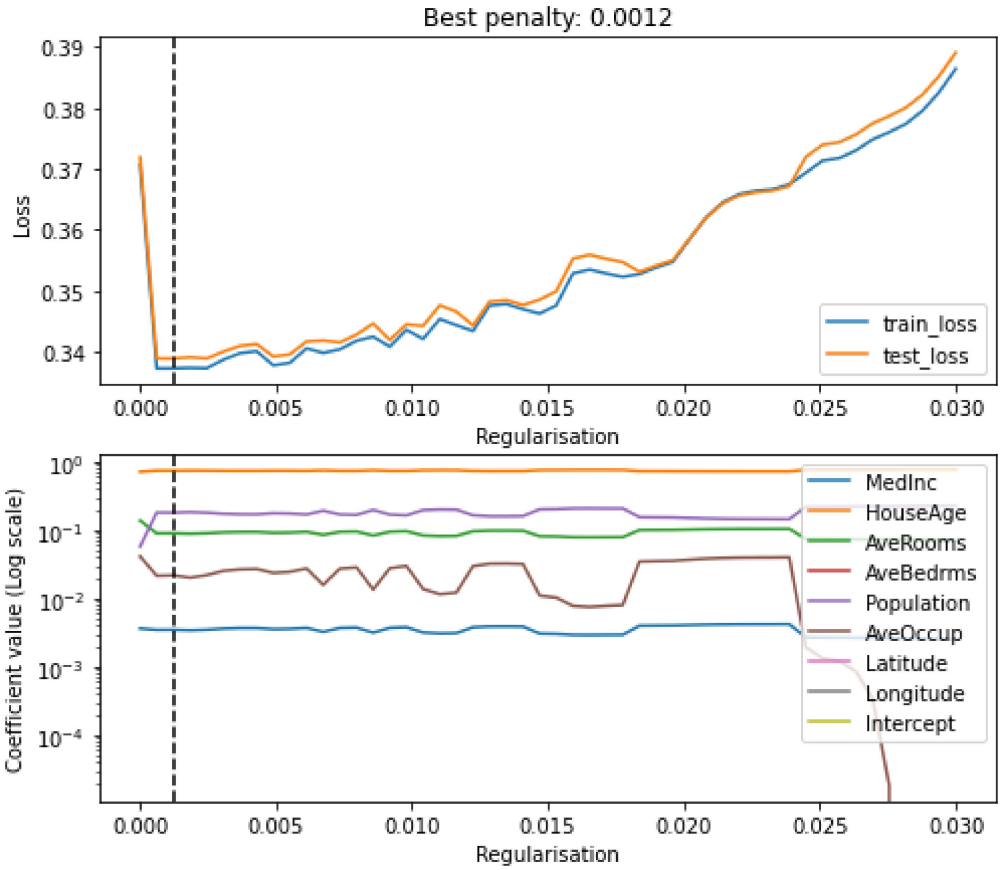
```
tau_opt=0.0112
```

The optimal step choice should be 0.012, which corresponds relatively well with the observed result of 0.017

Question 4

Show the regression performance on the test set as the ridge penalty changes.

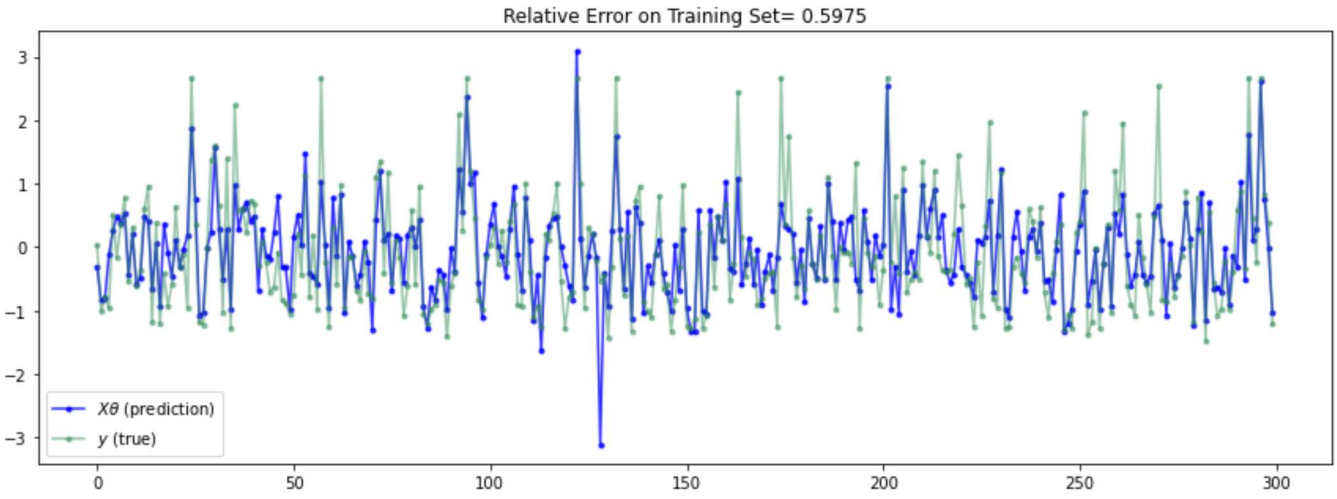
We set the step size to the value determined above and repeat the procedure as with Question 3

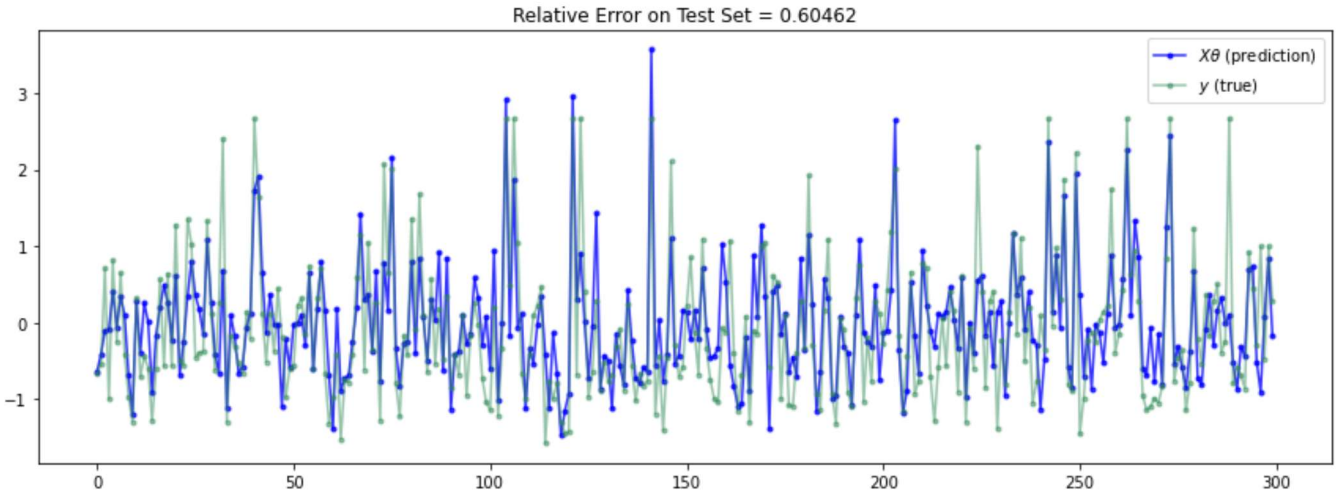


And the coefficients associated are

```
{'MedInc': 0.0035,
 'HouseAge': 0.7289,
 'AveRooms': 0.1248,
 'AveBedrms': -0.132,
 'Population': 0.1078,
 'AveOccup': 0.0352,
 'Latitude': -0.2112,
 'Longitude': -0.4033,
 'Intercept': -0.3272}
```

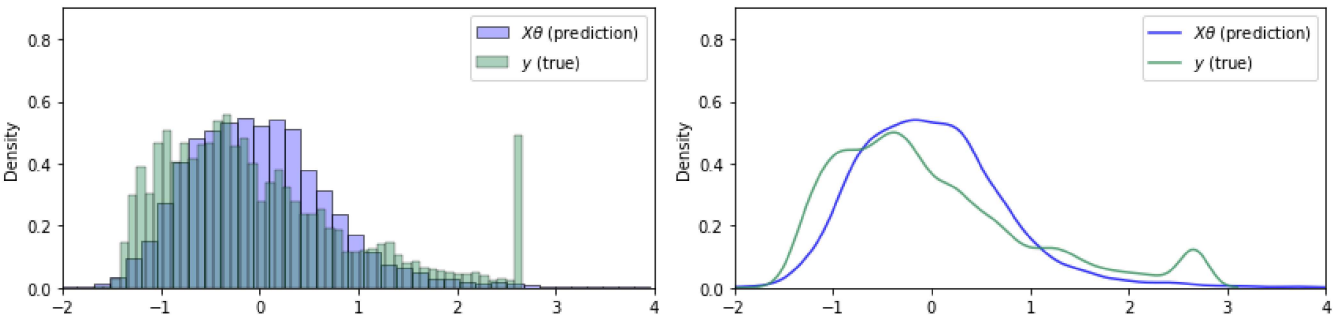
Let us now consider the predictive performance





The relative error on both the test and train set are comparable up to 2 significant digits and although not very close to 0, seems reasonable.

Consider also a histogram and kernel density plot of the observed and predicted lables y^* and $\hat{y}$ below.



We observe that these follow roughly the same distribution, indicating that long-term predictions should be statistically consistent.