מבוא למערכות לומדות

מספר קורס: 67577

תרגיל 2

שם: מיטב עובדיה

ת"ז: 211860465

Practical part:

Part 3.1 - Fitting A Linear Regression Model:

Q2: Describing the analysis process that lead me to the decisions of preprocess_data:

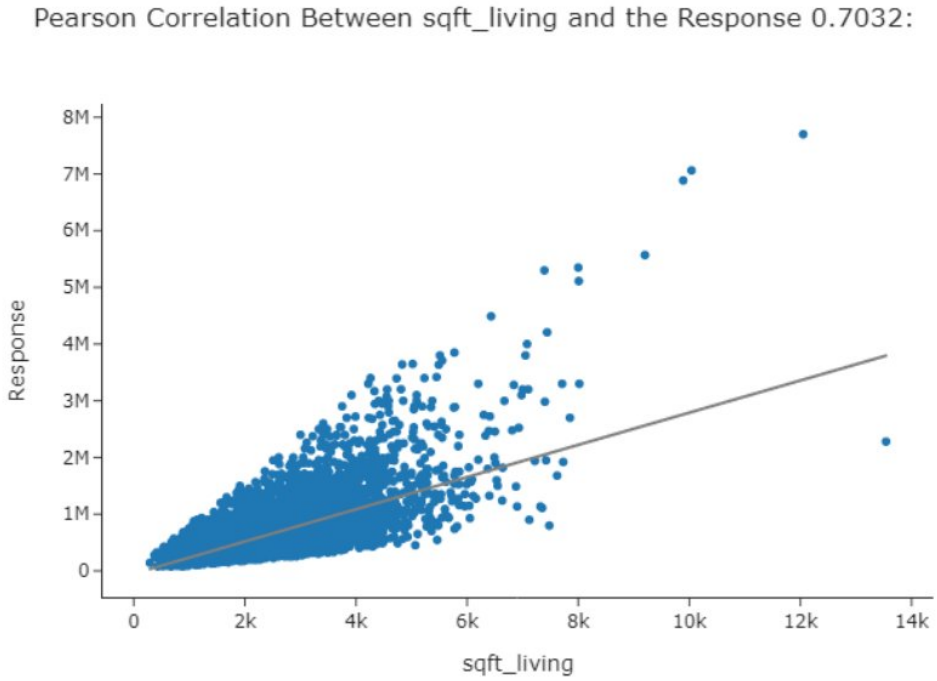
I took time to have good understanding of the dataset using House Sales in King County, USA | Kaggle. I understood that feature like the 'id' are not relevant to the model because the don't have affect on the price. Contrary to that, there were features that affected the price of the house such as the sqftbliving of the house. I choose to not include irrelevant features during computing the model, so I drop them during the process function. I wrote a function' helper to preprocess_data that make this dropping.

The categorical features are features like "bedrooms", "bathrooms", "sqft_living", "sqft_lot", "floors, "waterfront", "view", "condition", "grade," "sqft_above", "sqft_basement" . These are features whose content affects the house price significantly. I filtered their values by duing tests on these features such as is the feature is between the minimum value and the maximum value of this feature (I played with the data on the website I mentioned above) this logic in the helper function called: deal_with_invalid_vals. I also differentiated between the training set and the test set during dealing with invalid features. At first I thought that the zip code is not a feature that affects the price, then I realized that the zip code is determined by the area and the quality of the area affects the price of the house.

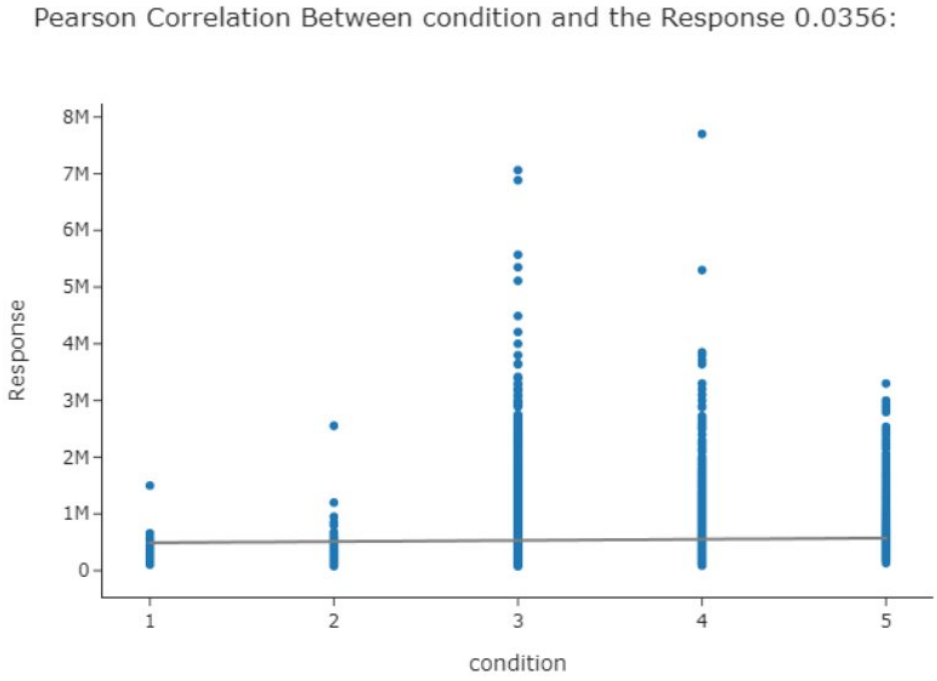
While in the training set I removed samples where the values were invalid, in the test set I replaced the value of the sample with the average value of the values of this particular feature. I saved the averages in a dictionary that maps between the name of the feature and the average of its values. See functions init_mean_dict, deal_with_invalid_vals in the code. Similarly, when there were nan values in the training set, I drop the sample, while in the test set, I replaced it with the average. This can be seen in the function: replace_nan_with_mean. Before processing function I deleted samples where the price was Nan. In the preprocess_data function, I distinguished between the case of train data and the case of test data according to the value of y (y was None if it was the test set).

Q3: Choosing two features, one that seems to be beneficial for the model and one that does not and explain:

- Beneficial for the model:
  From this figure we can see that more the living area , more the price though data is concentrated towards a particular price zone. From the figure I see that the data points seem to be in linear direction. I could see some irregularities that the house with the highest square feet was sold for very less , maybe there is another factor or probably the data must be wrong. In addition I printed the Pearson Correlation and the value close to 1, so there is almost linear relation between living area and price.

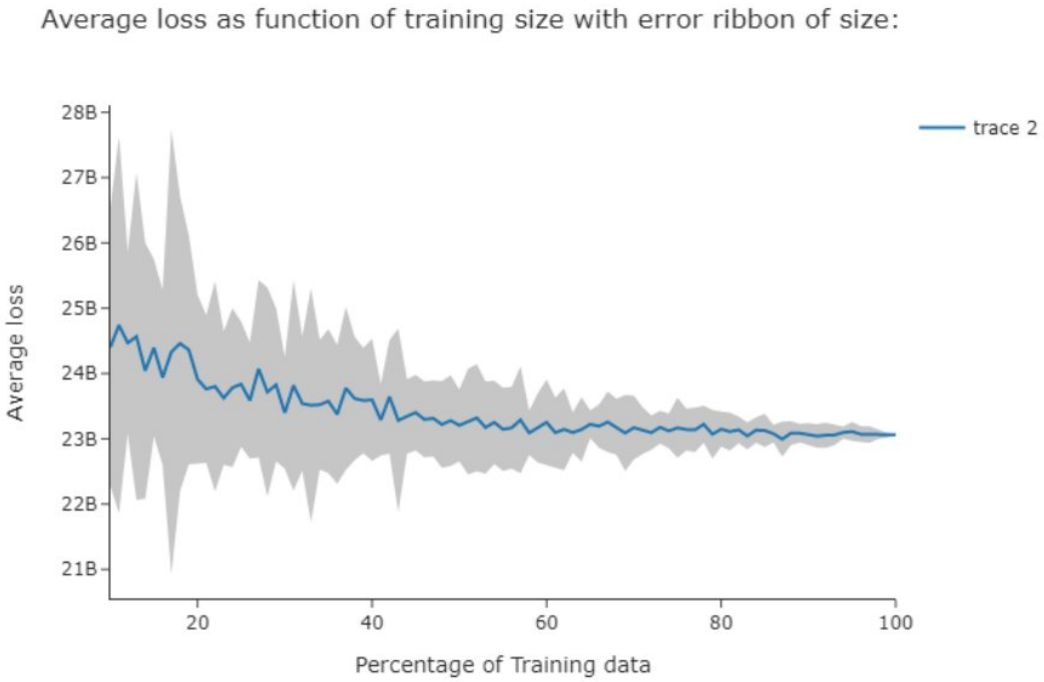


- Doesn't Beneficial for the model:



The condition value does not beneficial for the model, I realized that from the figure because the Pearson Correlation and the condition value very close to 0, so there is almost no linear relation between condition and price.

Q4: Explanation what is seen:

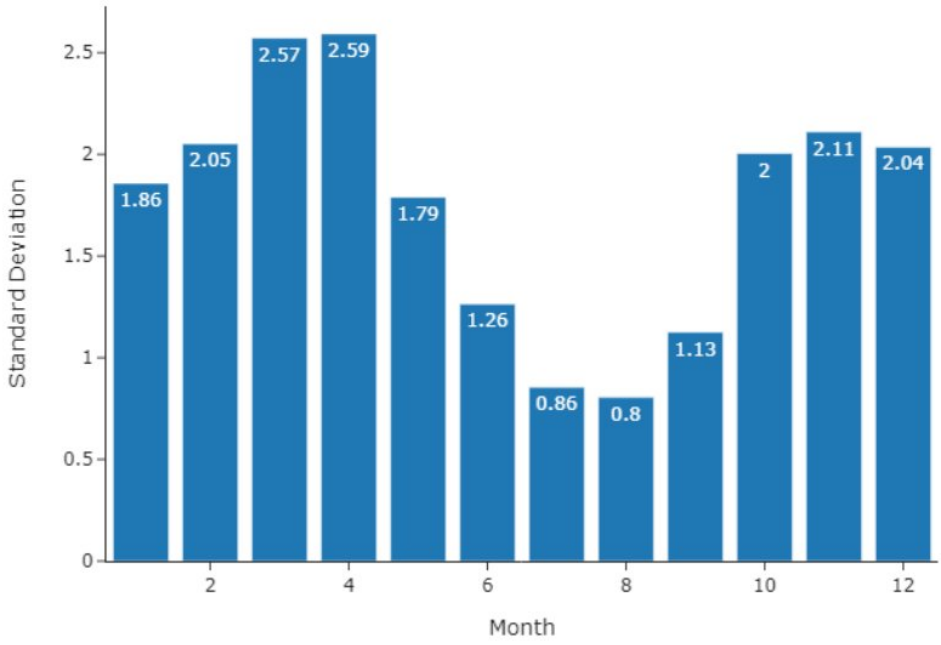


In the graph above we can see the average error as a function of the precentage of the training set. The more we increased the percentage of the training set, the smaller the confidence interval. So, we can understand that when fitting linear model as we increase that, we will see that the test error decreases and the variance of the prediction decreases too.

Part 3.2 - Polynomial Fitting:

Q2: Polynomial degree might be suitable for this data:

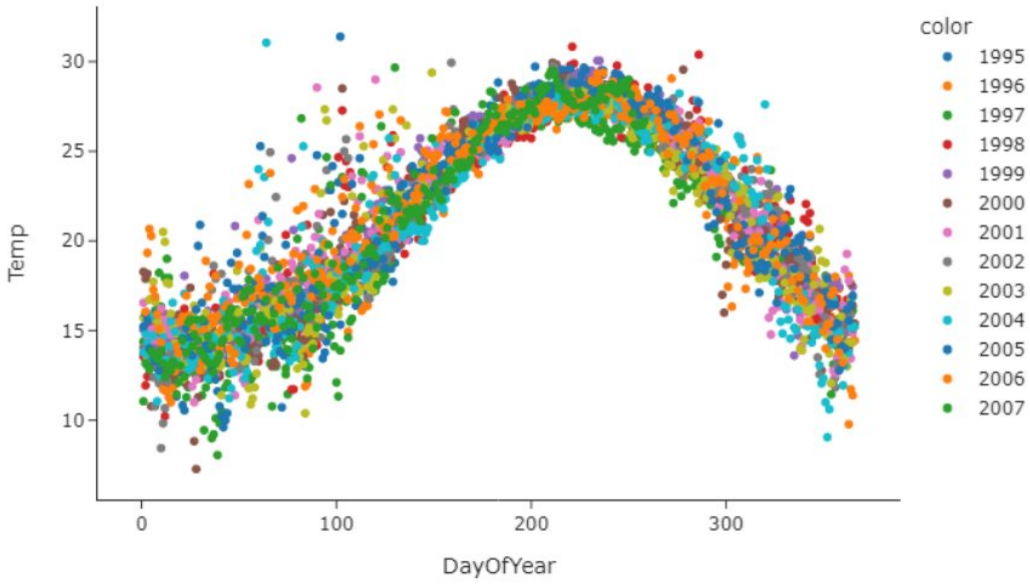
The standard deviation of the daily temperatures:



I think that the Polynomial degree might be suitable for this data is 3 or 4. We will need a degree greater than 2 because the temperatures do not continue to drop from both sides of the peak but remain stable.

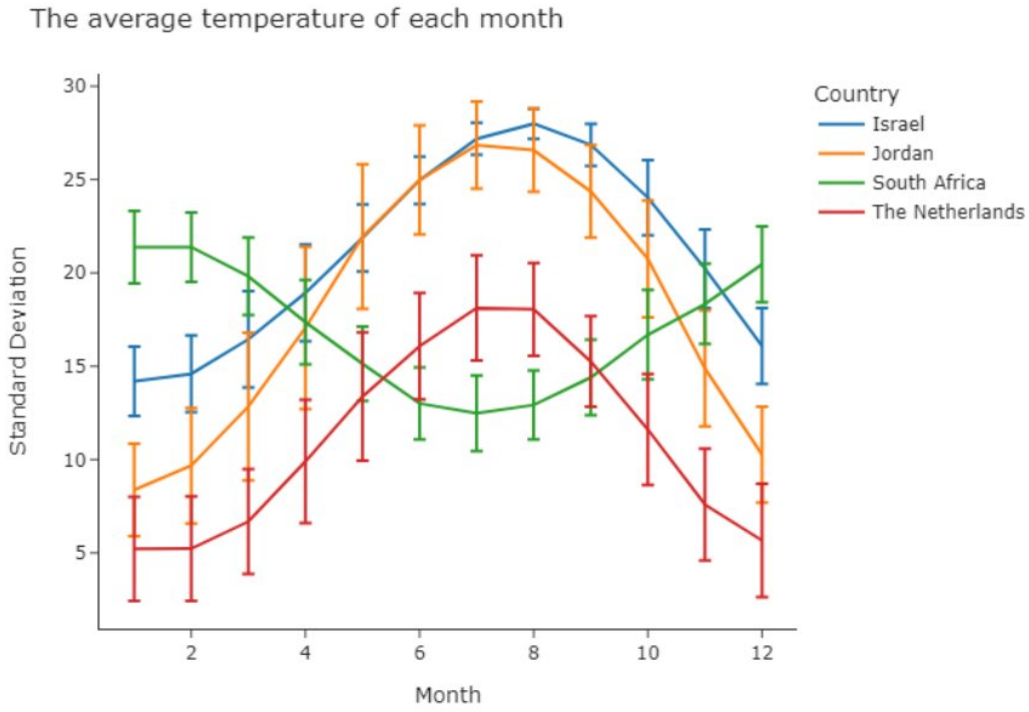
- Expection of the model success equally over all months or are there times of the year:

The relation between average daily temperature to DayOfYear in Israel:



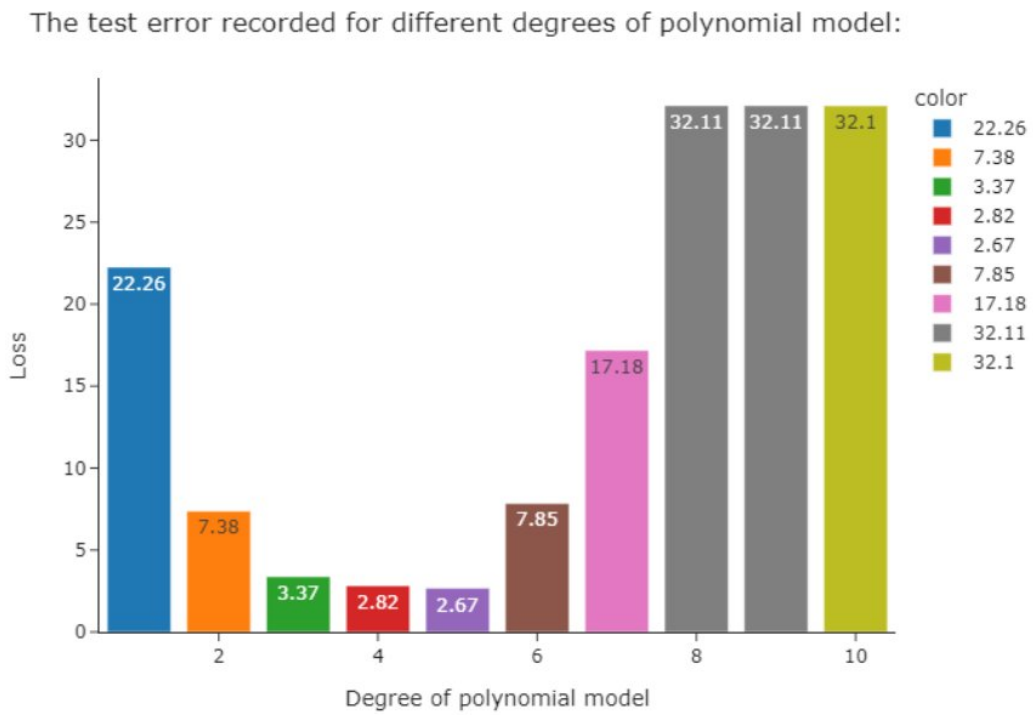
In my opinion, based on the graph the model will not equally be as successful in predicting across the year. The reason for this is the differences in variance between the months. There are months with high variability where the forecasting performance was less. In contrast, in the months of low variability the prediction will be better and more accurate (if the distribution from which the test is taken is similar to the test set).

Q3: Do all countries share a similar pattern? For which other countries is the model fitted for Israel likely to work well and for which not:



Countries don't share a similar pattern. We can see from the graph that different countries have a different distribution of average daily temperatures to month during the year. We can see that fitting the model on Jordan will likely to work well. However, fitting the model to countries like The Netherlands or South Africa likely to don't work well.

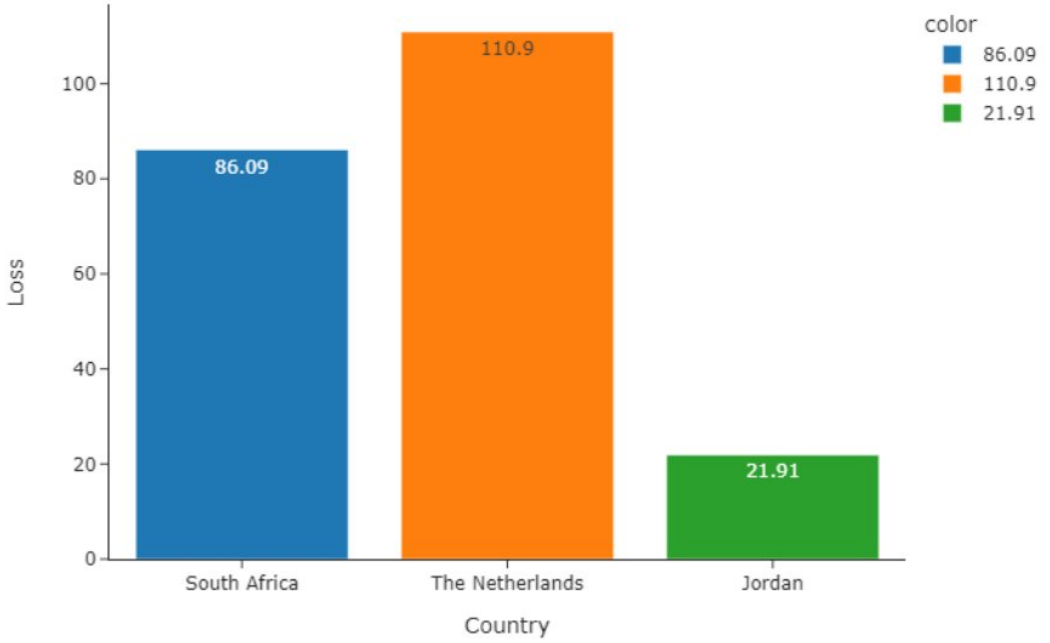
Q4: Which value of k best fits the data? In the case of multiple values of k achieving the same loss select the simplest model of them. Are there any other values that could be considered?



The best value of k is 5 because model fitted with k = 5 got the lowest test error value : 2.67. Other values that could be considered are 4 or 3, because model fitted with k = 4 got the a test error value : 2.82 and model fitted with k = 3 got a test error: 3.37.

Q5: Explain my results:

The fitted over Israel model's error over each of the countries:



The model fitted with k=5 over the observations of Israel performed less good over observations from other countries, which makes a lot of sense. In the third question the distribution of temperatures in Jordan has been similar to the distribution of temperatures in Israel. Thus, the model performed on Jordan better then it preformed on South Africa and The Netherlands. The fitted model performed poorly on South Africa and The Netherlands because their distributions were further very different of distributions of Israel.