**Assignment Part 1**

Please capture all your work in an Orange file (with extension .ows) a Word file and an Excel file (you will need to have all three). Please copy the images and reports that you generate in Orange into the Word file to illustrate your answers as needed.

The dataset for this assignment is called “PortlandHousing.xls” and is available on GitHub at the following URL:

<https://github.com/jsub10/Machine-Learning-Course-2018/blob/master/DataSets/PortlandHousingPrices.xlsx>

For the assignment, carry out the following steps:

1. Using Orange, build a pipeline for visualizing the data and visualize it using a scatter plot and a distribution plot. Briefly (1-2 sentences) explain your observations for each plot. Make sure you capture the images and paste them into your Word file.

Bedrooms:

|  |  |
| --- | --- |
|  | It appears that the houses recorded in our dataset, have at least one bedroom and maximum of 5. Houses with 3 bedrooms have the highest frequency. The plot looks like a normal distribution. |
| A close up of a logo  Description generated with very high confidence | This plot is the same as the one above except it is shown as bins rather than a smooth graph. You can play with the toggle in Orange to change the number of the bins. |
| A screenshot of a computer  Description generated with high confidence | This plot shows the price of houses vs their number of bedrooms. A clear trend shown is that as the number of bedrooms in the houses increase, the price increases as well. |

Area:

|  |  |
| --- | --- |
| A close up of an animal  Description generated with very high confidence | The recorded area of houses in our dataset range from to 852 to 4,478 (ft2). The highest frequency of area seems to be between 1,000 to 2,000 ft2. |
|  | This plot is the same as the one above except it is shown as bins rather than a smooth graph. You can play with the toggle in Orange to change the number of the bins. |
| A picture containing sky, text  Description generated with very high confidence | Similar to the bedrooms, it appears, as expected, that in general, the increase in the area of the house translates to an increase in the price. The two values are highly positively correlated. |

Price:

|  |  |
| --- | --- |
| A close up of a logo  Description generated with very high confidence | Recorded prices in our data set range between approximately $170k to $700k. The highest frequency of prices seems to be $200k to $300k |
|  | This plot is the same as the one above except it is shown as bins rather than a smooth graph. You can play with the toggle in Orange to change the number of the bins. |

1. Create a model for predicting the house price based on the number of bedrooms and the area of houses. Specifically, write down the expression for this model (using addition, multiplication, etc.). Write this down in your Word file.

(w0 \* x0) + (w1 \* x1) + (w2 \* x2) =

1. In the expression you wrote for step 2, clearly identify (a) the target, (b) the features, and (c) the parameters. Do this in your Word file.

x1 = number of bedrooms (the first feature)

x2 = size in square feet (the second feature)

w0, w1, w2 = parameters

= predicted house price (target)

1. Based on the expression for the model, expand the datasheet by adding the new columns. Clearly label the columns. Do this in your Excel file.

See Excel

1. Pick two penalty functions. Is one of your penalty functions better than the other? Explain why or why not. Answer this question in your Word file.
2. Linear Penalty
3. Squared Penalty

Squared Penalty would be better since all penalty values will be positive and won’t cancel each other. Linear penalty would penalize the predicted values that are higher but incentivize the predicted values that are lower than actual values.

1. Create columns for the parameters and calculate the cost for the following values of w0, w1, and w2: w0 = -10, w1 = 400, w2 = 600. Calculate the cost for each penalty function you chose in question (5). Do this in your Excel file.

See Excel