# DAT405 Assignment 2 – Group 111

Jihad Almahal- (4 hrs)
Oscar Karbin- (4 hrs)

February 15, 2023

## 1 First Question

### 1.a

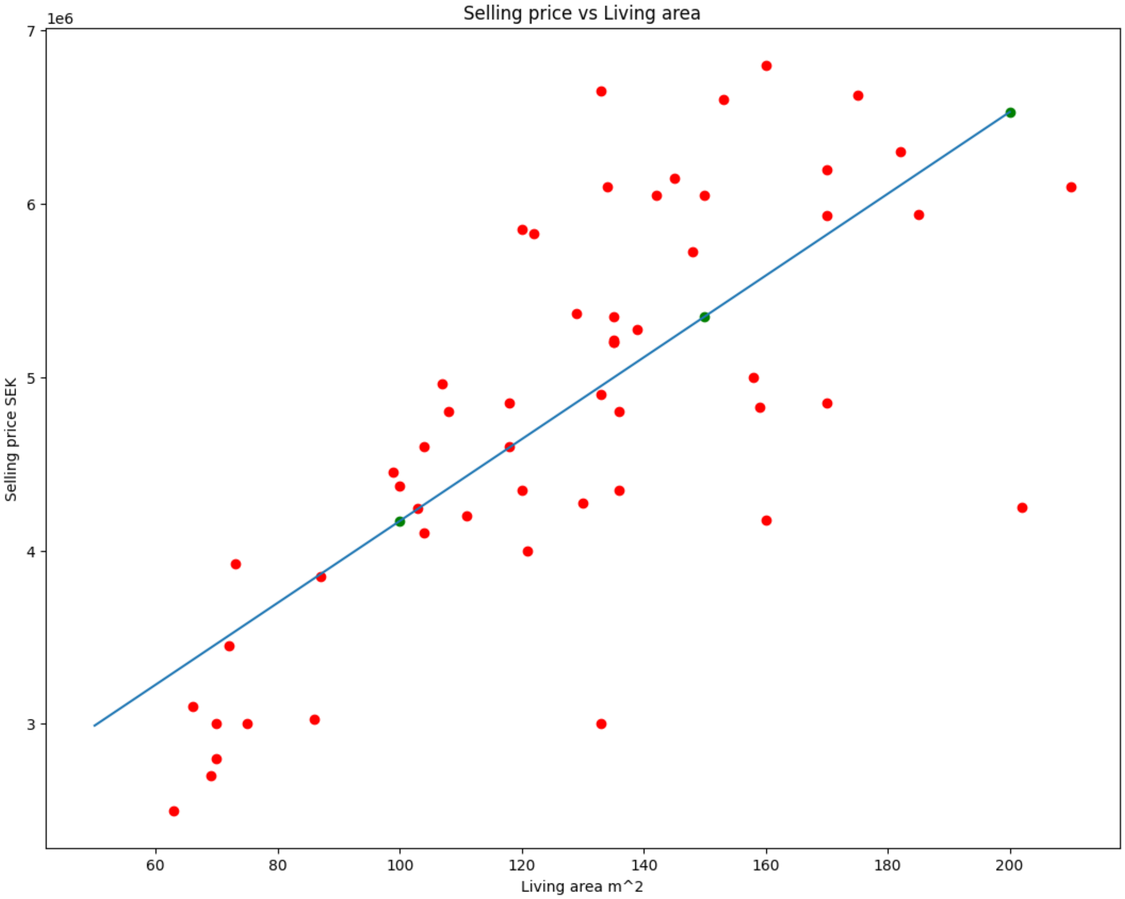


Figure 1: Living area vs selling price where the green points are predictions

From the given data we removed two data points. These data points were not relevant because of them likely not being villas.

## 1.b

Slope of the line: [23 597] Intercept of the line: [1 809 821]

## 1.c

living area 100 m2 is 4 169 600 SEK, 150 m2 is 5 349 490 SEK and 200 m2 is 6 529 380 SEK.

## 1.d

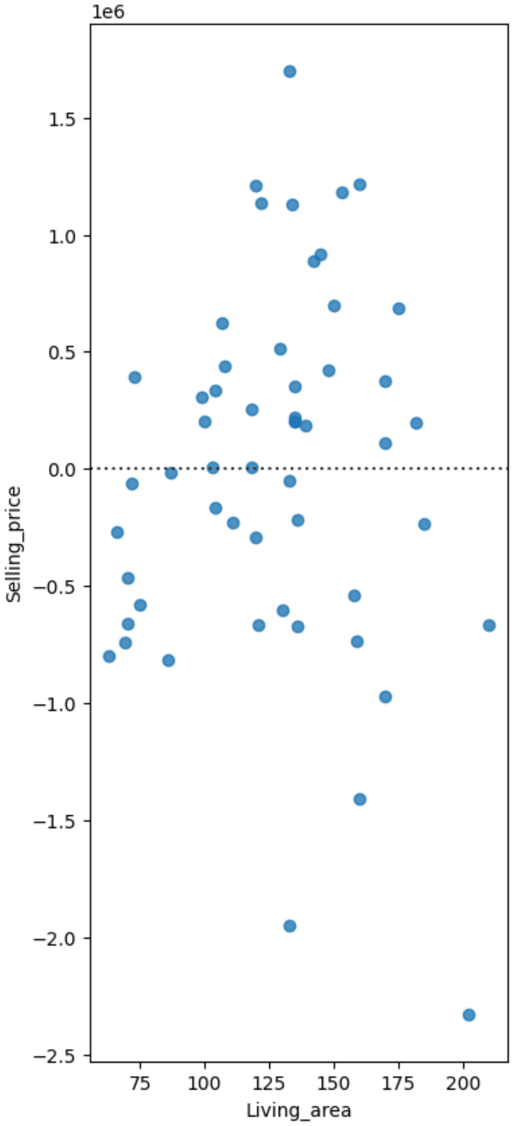
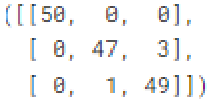


Figure 2: Living area vs selling price where the green points are predictions

# 2 Second Question

## 2.a

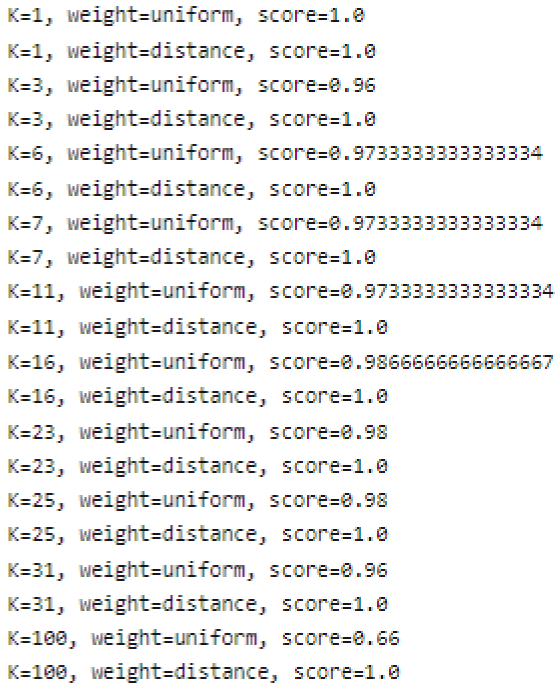
The confusion matrix we get with the help of logistic regression is as follows The matrix shows that all setosa's are classified correctly, while 3 versicicolor are detected as virgincia and 1 virgincia detected as a versiciolor.



```
([[50,  0,  0],
  [ 0, 47,  3],
  [ 0,  1, 49]])
```

Figure 3: confusion matrix

## 2.b

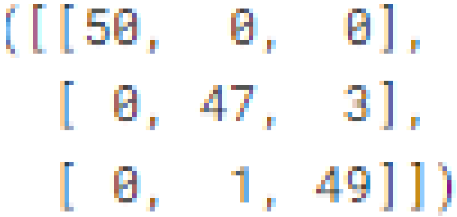


```
K=1, weight=uniform, score=1.0
K=1, weight=distance, score=1.0
K=3, weight=uniform, score=0.96
K=3, weight=distance, score=1.0
K=6, weight=uniform, score=0.9733333333333334
K=6, weight=distance, score=1.0
K=7, weight=uniform, score=0.9733333333333334
K=7, weight=distance, score=1.0
K=11, weight=uniform, score=0.9733333333333334
K=11, weight=distance, score=1.0
K=16, weight=uniform, score=0.9866666666666667
K=16, weight=distance, score=1.0
K=23, weight=uniform, score=0.98
K=23, weight=distance, score=1.0
K=25, weight=uniform, score=0.98
K=25, weight=distance, score=1.0
K=31, weight=uniform, score=0.96
K=31, weight=distance, score=1.0
K=100, weight=uniform, score=0.66
K=100, weight=distance, score=1.0
```

Figure 4: Iris data set with some different values for k, and with uniform and distance-based weights

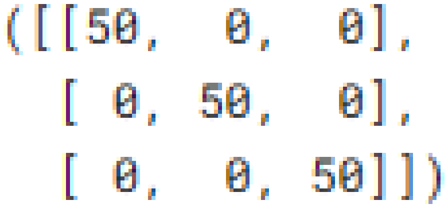
When searching for a suitable value for k we test several different values of k on the given dataset. We also know that a greater k-value generates a better generalization for classification of the iris. Additionally we have 3 different classes of iris flowers. Meaning we will not receive any ties.

When testing different values of k. We clearly see that the distance weight generates a better score, especially for larger k-values, and seems to be the most relevant option for our dataset. Therefore choosing a large K-value combined with distance 'weight' is the best option when striving for noise reduction



```
([[50,  0,  0],
  [ 0, 47,  3],
  [ 0,  1, 49]])
```

Figure 5: logistic regression



```
([[50,  0,  0],
  [ 0, 50,  0],
  [ 0,  0, 50]])
```

Figure 6: k-nearest neighbours

### 2.c

From the confusion matrix's in Figures 5 and 6 we can see that the k-nearest neighbours preforms better than the logistic regression on this dataset. Overall The performance of the KNN model will depend on the value of k and the type of weighting used. As discussed in the previous answer, larger values of k lead to smoother decision boundaries and more robust predictions. The performance of the logistic regression model will depend on the choice of regularization and the choice of solver. Logistic regression is a relatively simple and fast model, but may not fit complex decision boundaries as well as KNN.