# K-Nearest Neighbors Algorithm on Kidney Data

Lily Samuel

12/10/2022

```
library("FNN")
library("readr")
kidney<-read_delim(paste0("https://raw.githubusercontent.com/",
"mcgillstat/regression/main/data/kidney.txt"),
" ", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 157 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: " "
## dbl (2): age, tot
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(kidney)
```

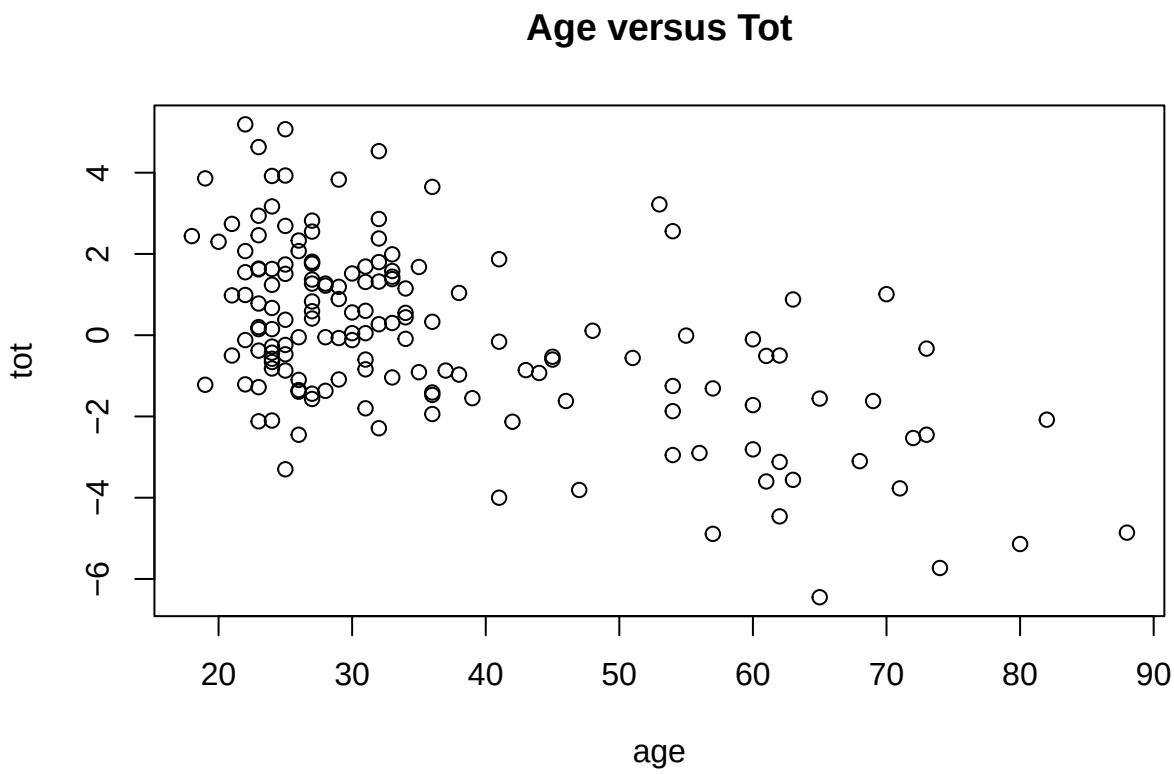
```
## # A tibble: 6 x 2
##     age   tot
##   <dbl> <dbl>
## 1    18  2.44
## 2    19  3.86
## 3    19 -1.22
## 4    20  2.3
## 5    21  0.98
## 6    21 -0.5
```

Task 1:

Plot the observations for the response against the observations for the predictor. Is the relationship between tot and age linear?

Solution:

```
x_kidney<-kidney$age
y_kidney<-kidney$tot

plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "Age versus Tot",
     col="black")
```




We can see that there is a negative linear relationship between age and tot. As age increases, the overall kidney function decreases.

Task 2:

Create a "test" set, that is a grid of age values at which we will predict tot.

Solution:

The range of the grid is [mini(agei), maxi(agei)].

```
data.b<-seq(min(y_kidney), max(x_kidney), by=0.01)
grid.of.age<-data.frame(age=data.b)
head(grid.of.age)
```

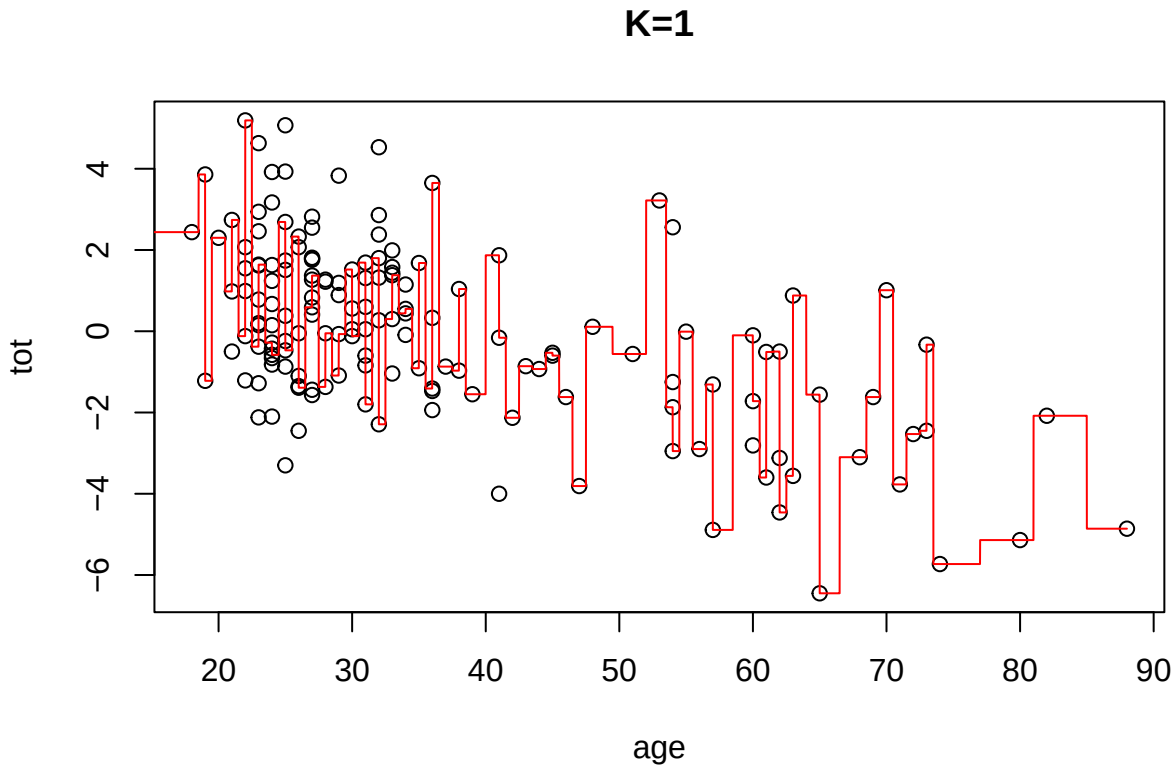
```
##     age
## 1 -6.45
## 2 -6.44
## 3 -6.43
## 4 -6.42
## 5 -6.41
## 6 -6.40
```

Task 3:

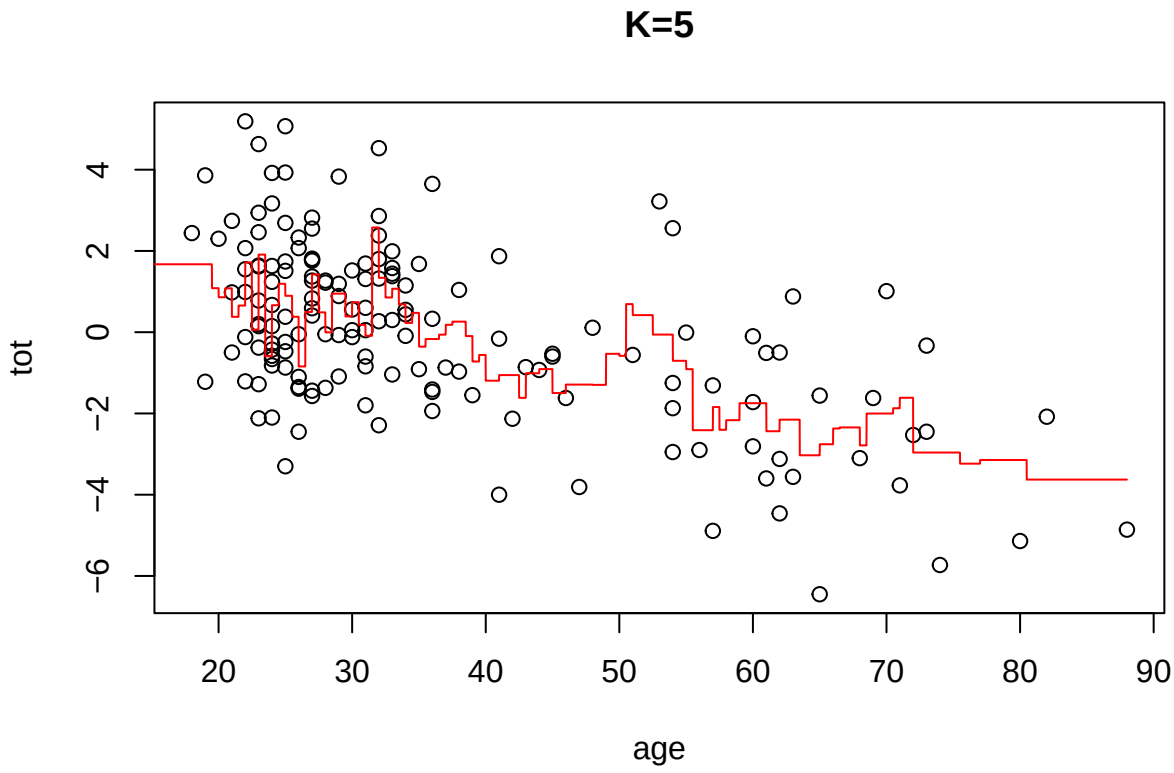
Plot the KNN predictions using the "test" set defined in task 2 for various k = {1, 5, 10, 25, 50, 157}. Which value of k gives the best prediction?

Solution:

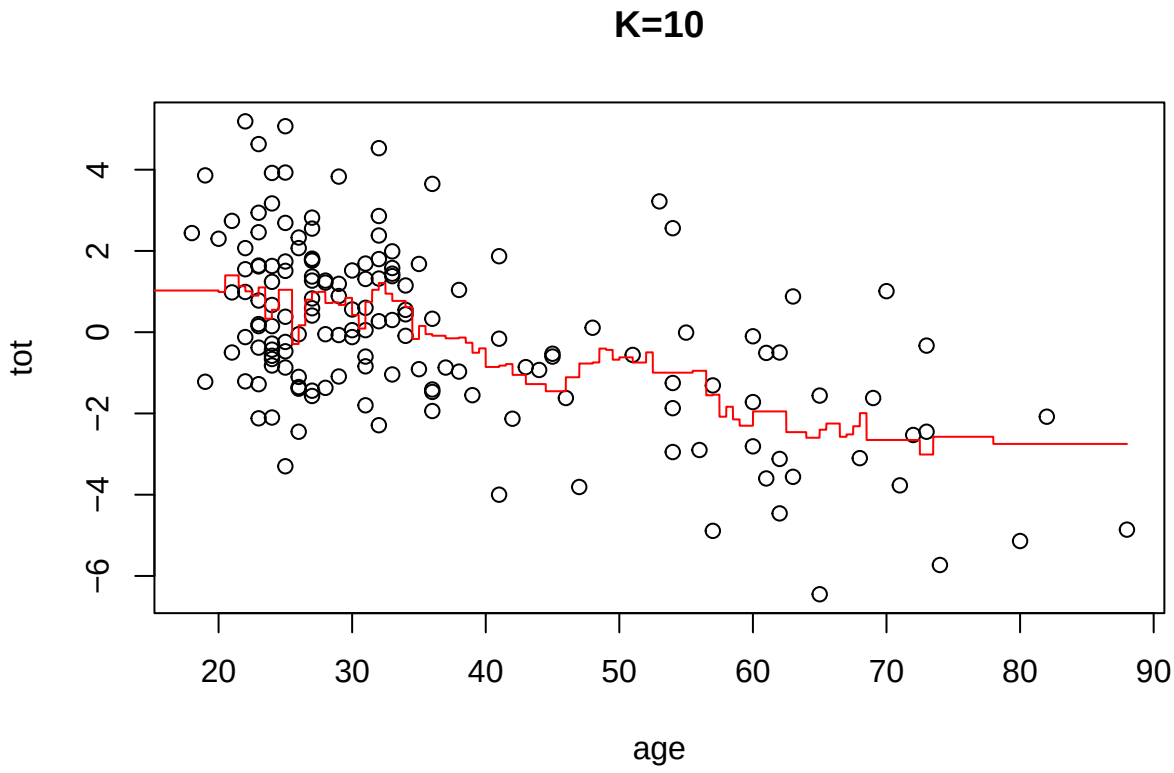
```r
KNN.Pred1 <- knn.reg(x_kidney, grid.of.age, y_kidney, k = 1)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=1",
     col="black")
lines(grid.of.age$age, KNN.Pred1$pred, col = "red", lwd = 1)
```



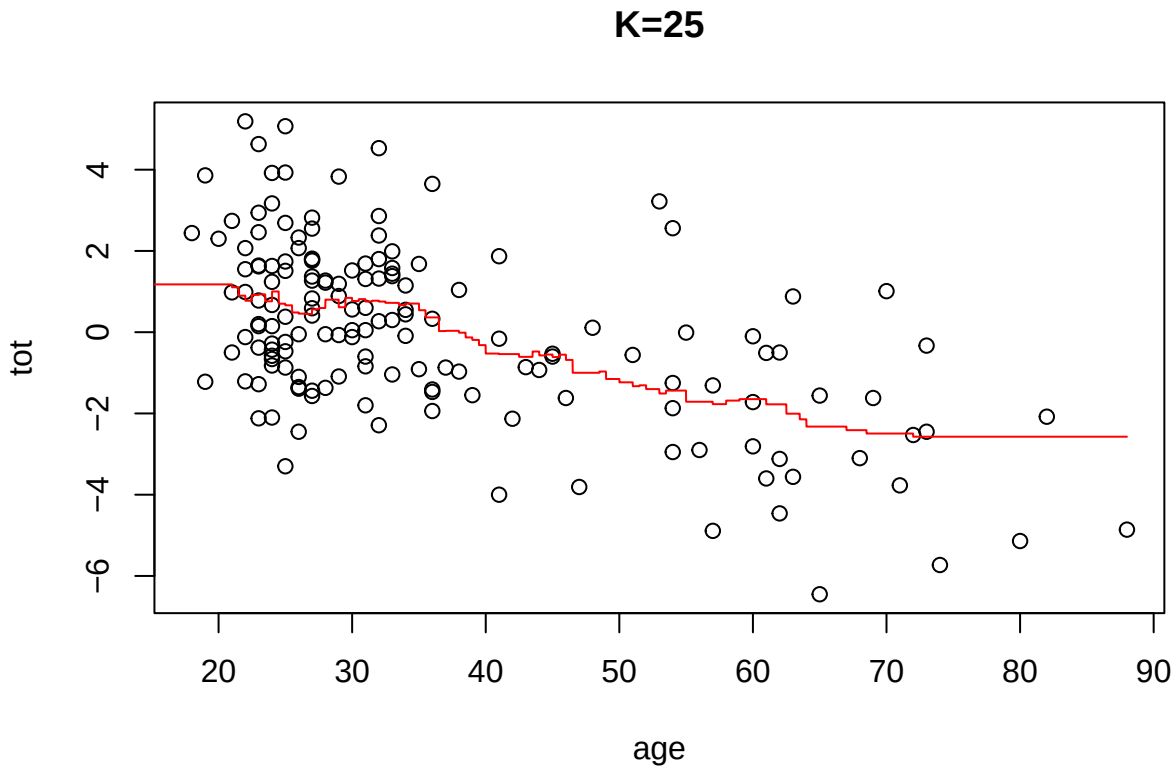

```r
KNN.Pred5 <- knn.reg(x_kidney, grid.of.age, y_kidney, k = 5)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=5",
     col="black")
lines(grid.of.age$age, KNN.Pred5$pred, col = "red", lwd = 1)
```



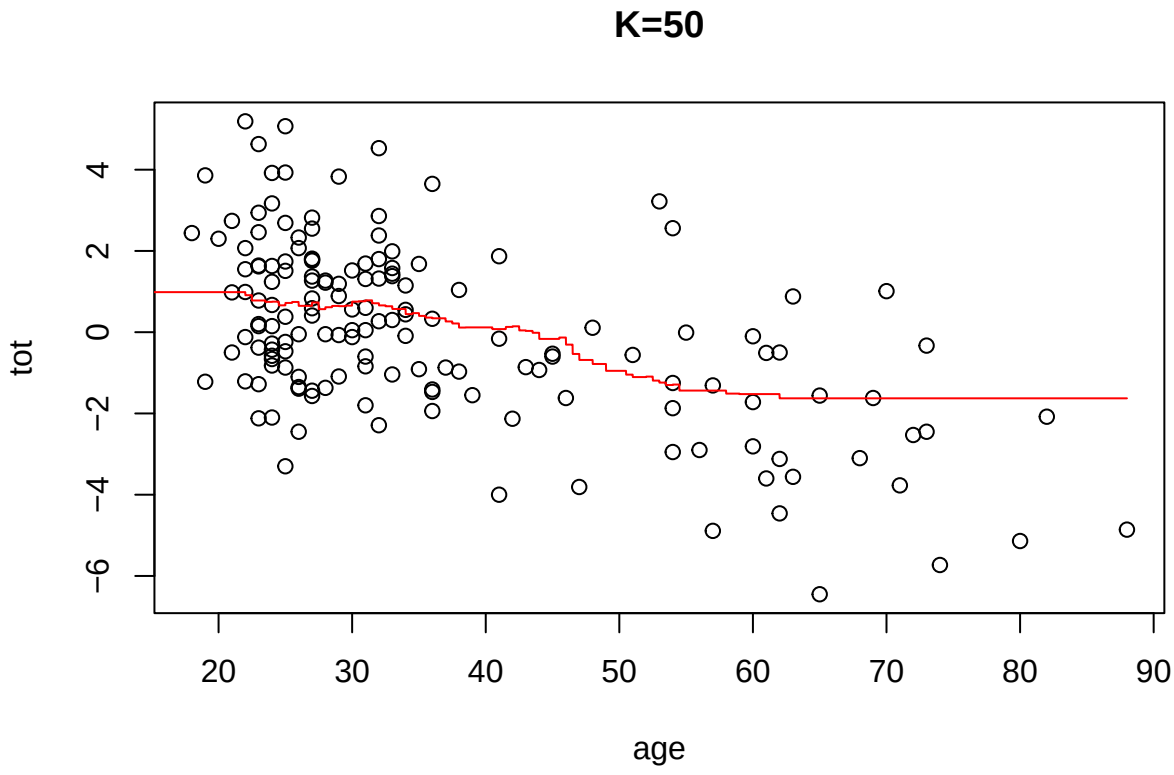

```
KNN.Pred10 <- knn.reg(x_kidney, grid.of.age, y_kidney, k = 10)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=10",
     col="black")
lines(grid.of.age$age, KNN.Pred10$pred, col = "red", lwd = 1)
```



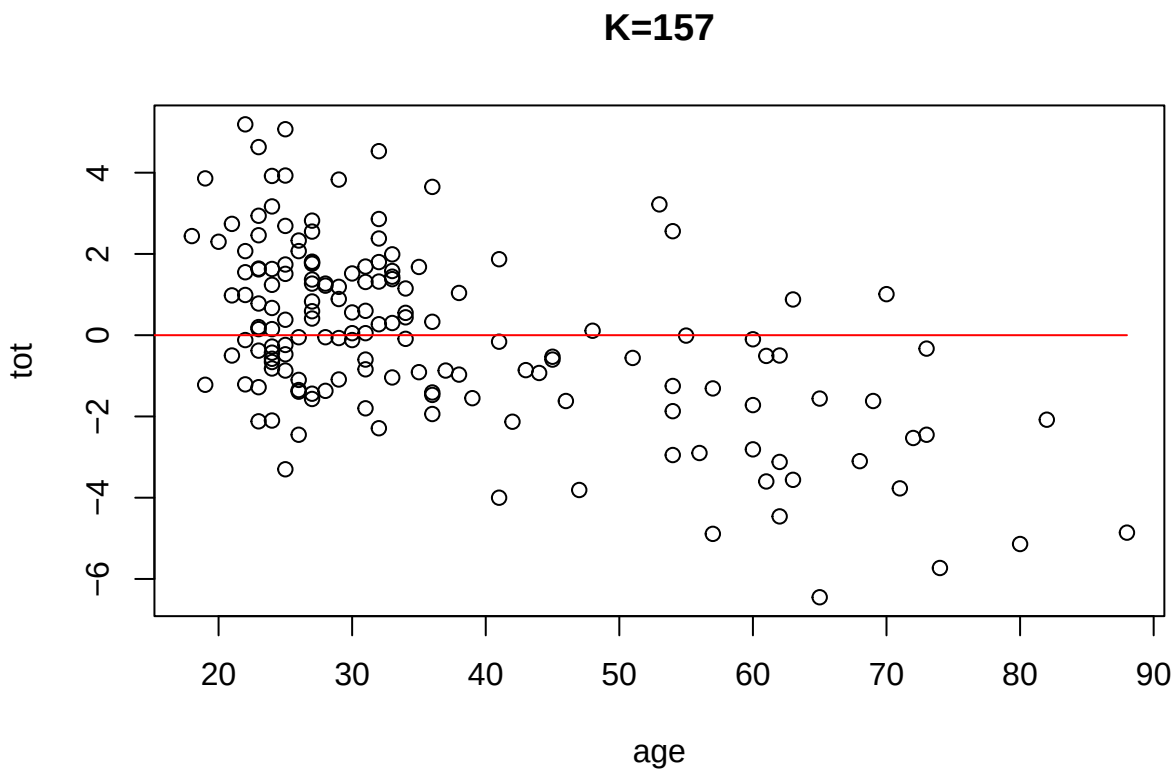

```
KNN.Pred25 <- knn.reg(x_kidney, grid.of.age, y_kidney, k = 25)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=25",
     col="black")
lines(grid.of.age$age, KNN.Pred25$pred, col = "red", lwd = 1)
```




```
KNN.Pred50 <- knn.reg(x_kidney, grid.of.age, y_kidney, k = 50)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=50",
     col="black")
lines(grid.of.age$age, KNN.Pred50$pred, col = "red", lwd = 1)
```




```
KNN.Pred157 <- knn.reg(x_kidney, grid.of.age, y_kidney, k = 157)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=157",
     col="black")
lines(grid.of.age$age, KNN.Pred157$pred, col = "red", lwd = 1)
```



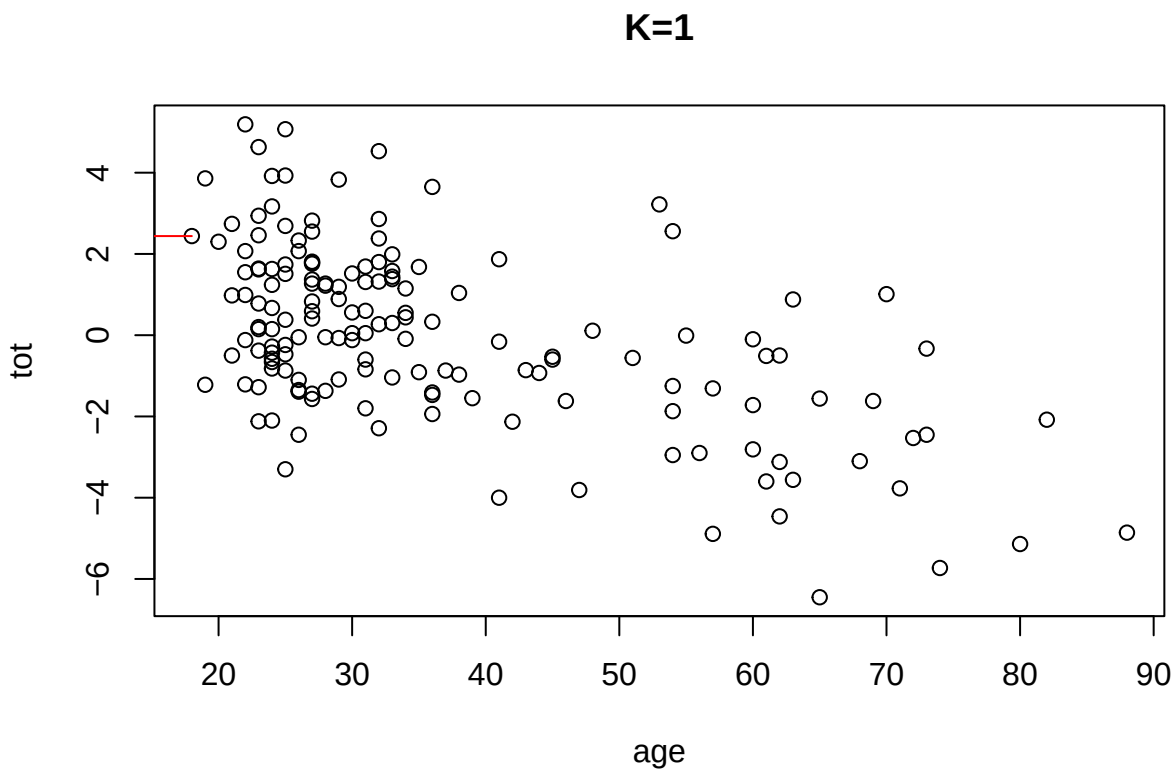

Looking at these graphs, we can see that the best prediction comes from k = 50. At k = 1, K=5, K=10, and k=50, the curve is over fitted. The curve is under fitted at k = 157. Therefore, at k=50, the graph is most appropriately fitted.

Task 4:

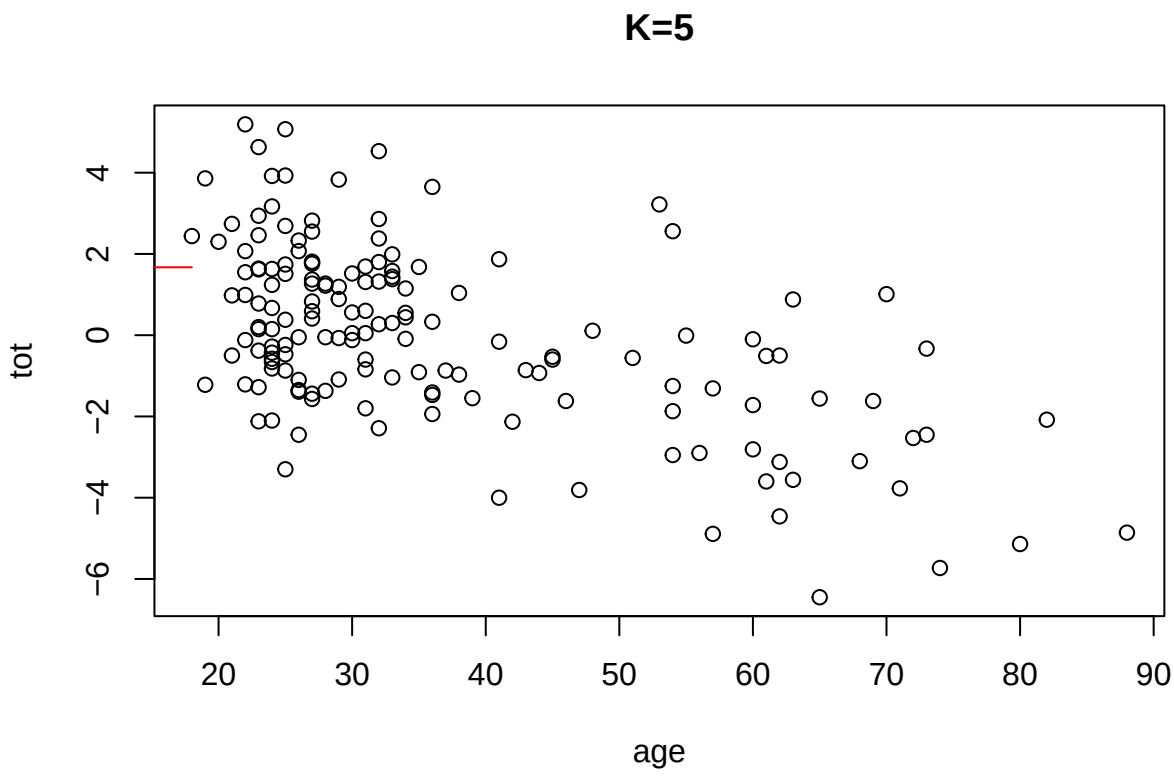
Change the range of the grid for the test set to [1,18]. Repeat task 3 for this test set. How do the predictions change? What's the problem of the predictions?

Solution

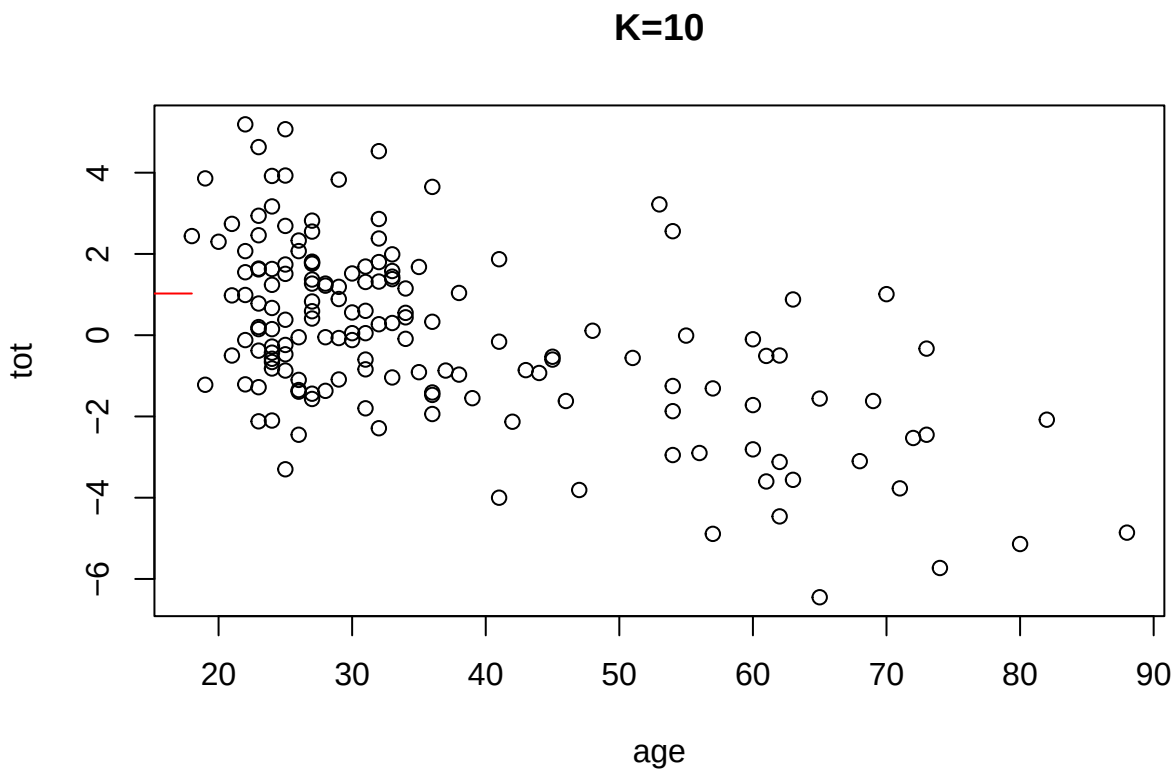
```
data.2a<-seq(min(1), max(18), by=0.01)
grid.of.age2<-data.frame(age=data.2a)

KNN_Pred1 <- knn.reg(x_kidney, grid.of.age2, y_kidney, k = 1)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=1",
     col="black")
lines(grid.of.age2$age, KNN_Pred1$pred, col = "red", lwd = 1)
```



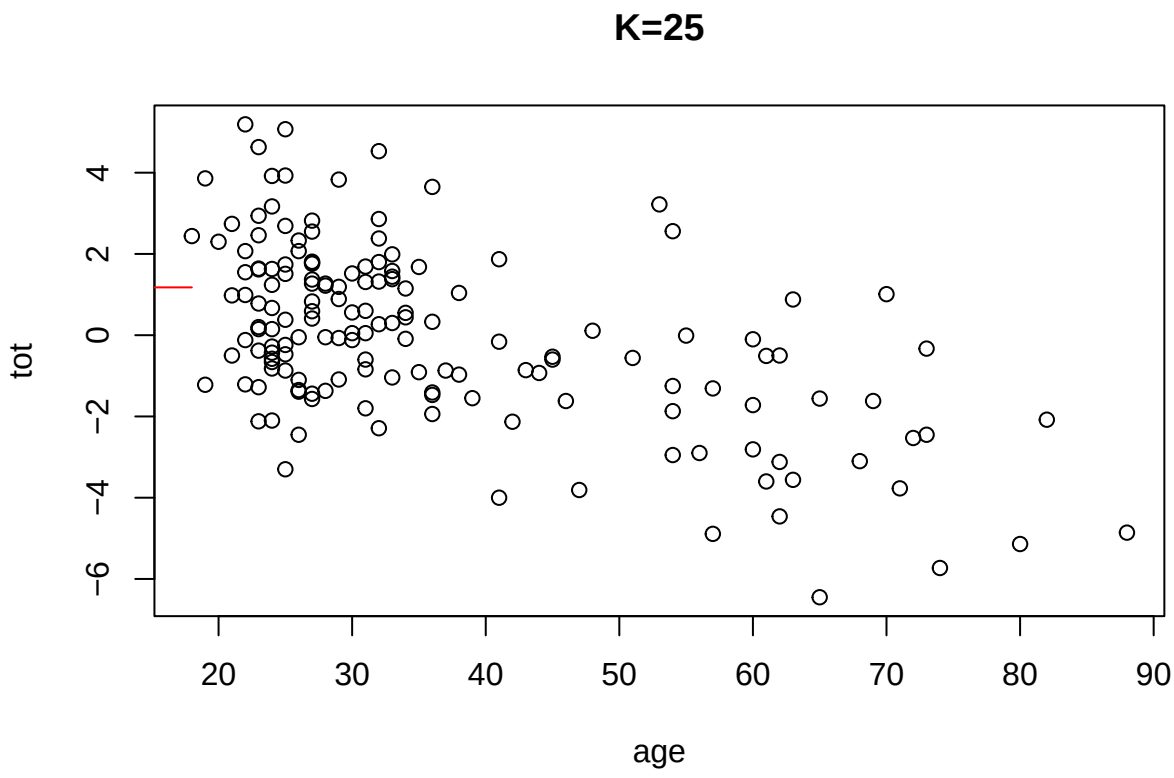

```
KNN_Pred5 <- knn.reg(x_kidney, grid.of.age2, y_kidney, k = 5)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=5",
     col="black")
lines(grid.of.age2$age, KNN_Pred5$pred, col = "red", lwd = 1)
```



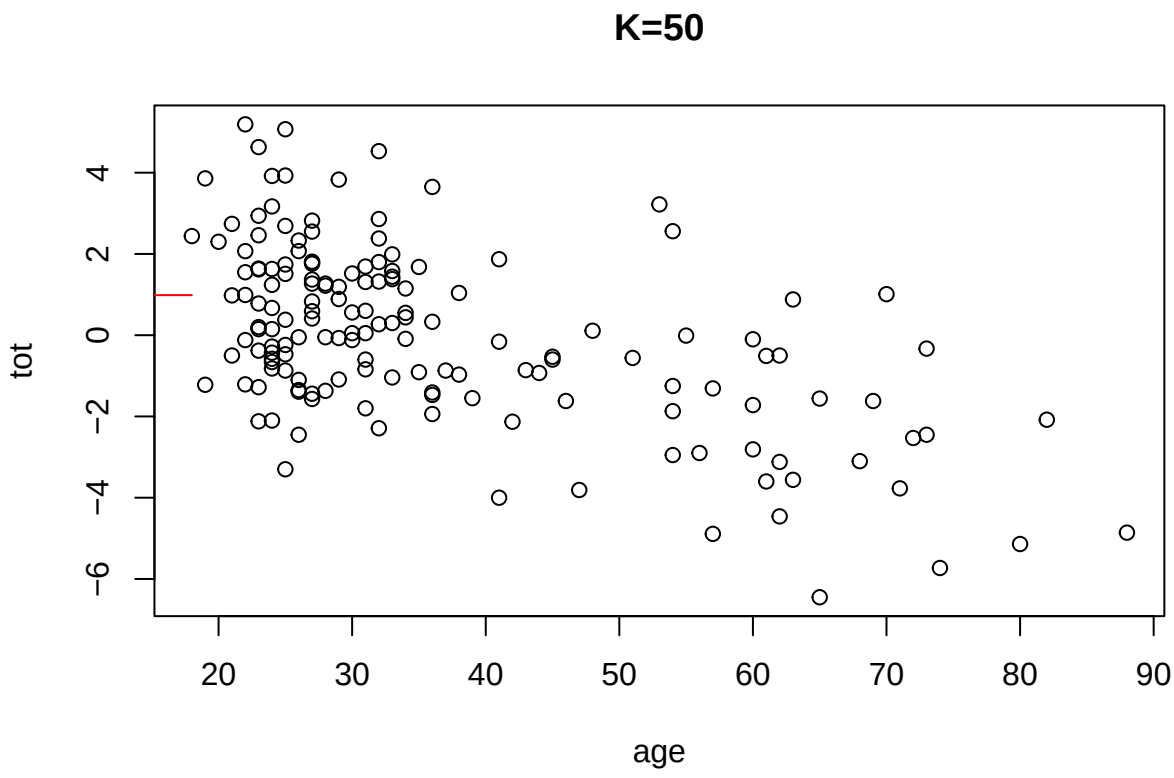

```
KNN_Pred10 <- knn.reg(x_kidney, grid.of.age2, y_kidney, k = 10)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=10",
     col="black")
lines(grid.of.age2$age, KNN_Pred10$pred, col = "red", lwd = 1)
```



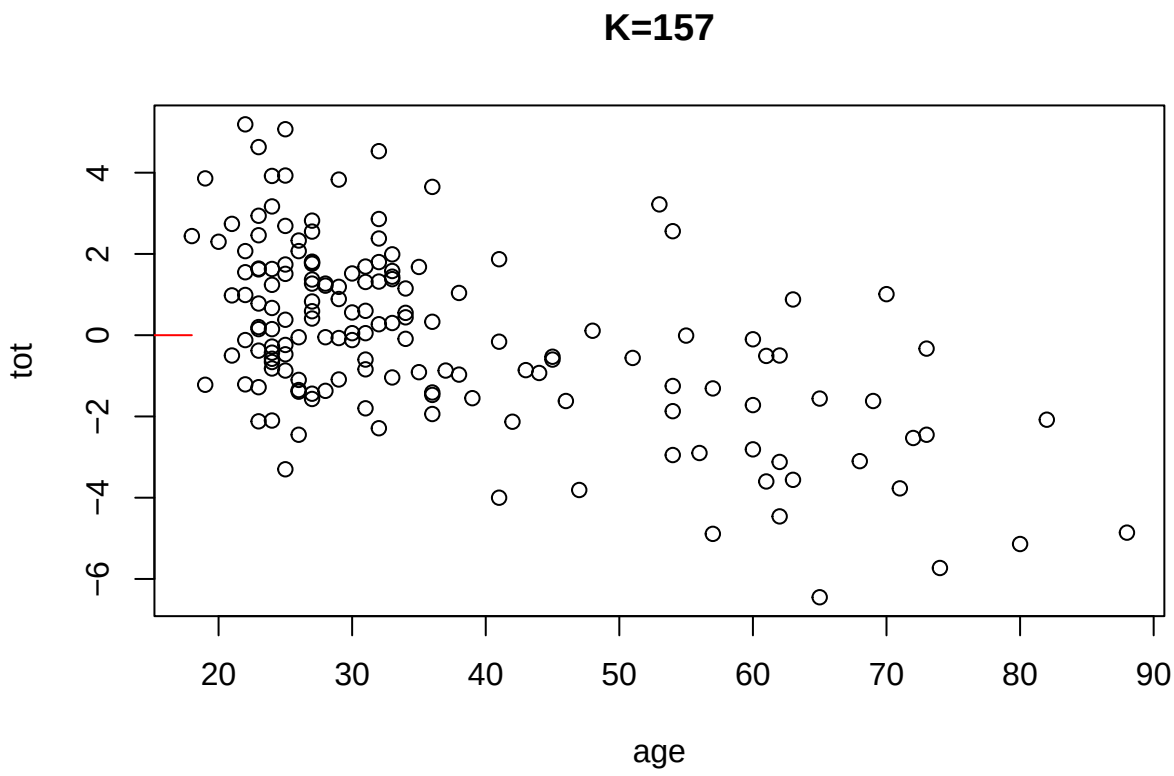

```
KNN_Pred25 <- knn.reg(x_kidney, grid.of.age2, y_kidney, k = 25)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=25",
     col="black")
lines(grid.of.age2$age, KNN_Pred25$pred, col = "red", lwd = 1)
```




```
KNN_Pred50 <- knn.reg(x_kidney, grid.of.age2, y_kidney, k = 50)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=50",
     col="black")
lines(grid.of.age2$age, KNN_Pred50$pred, col = "red", lwd = 1)
```




```
KNN_Pred157 <- knn.reg(x_kidney, grid.of.age2, y_kidney, k = 157)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=157",
     col="black")
lines(grid.of.age2$age, KNN_Pred157$pred, col = "red", lwd = 1)
```



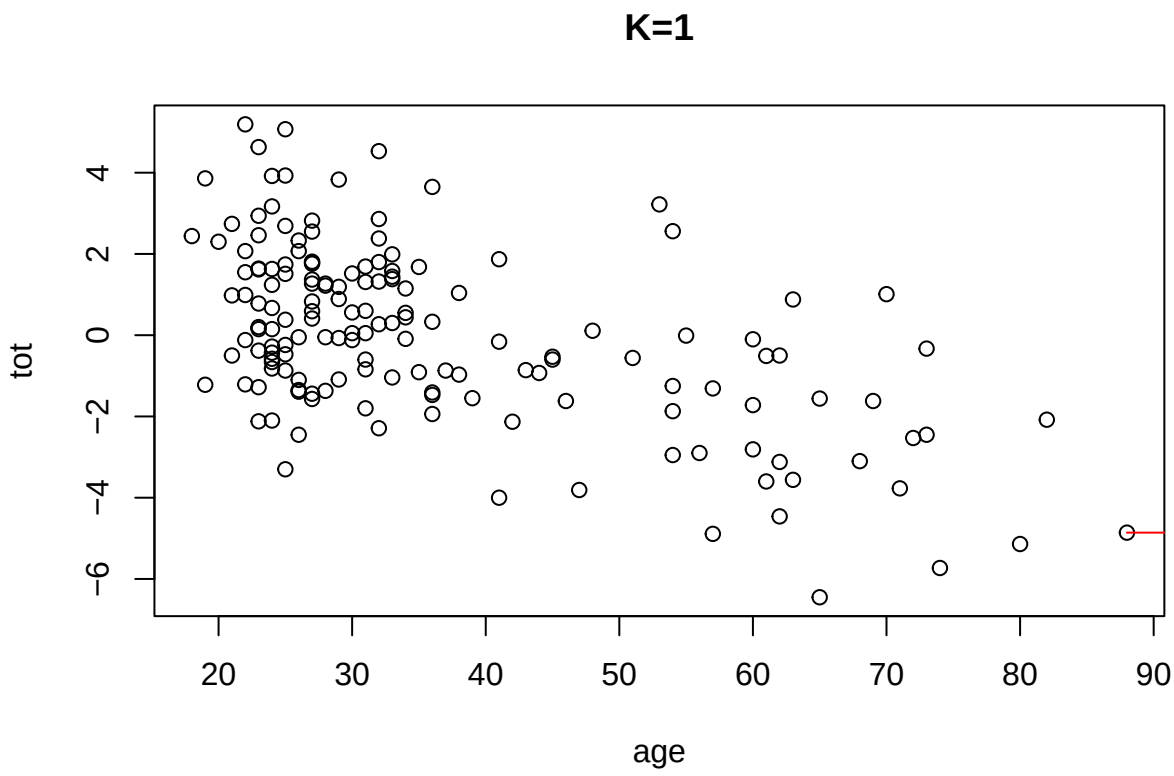

The age range [1,18] is outside the range [18,88]. The prediction being outside the training set makes the data meaningless and not able to be used for our purposes. As k increases, the line of prediction on the y axis decreases, not necessarily meaning that the prediction is more accurate.

Task 5:

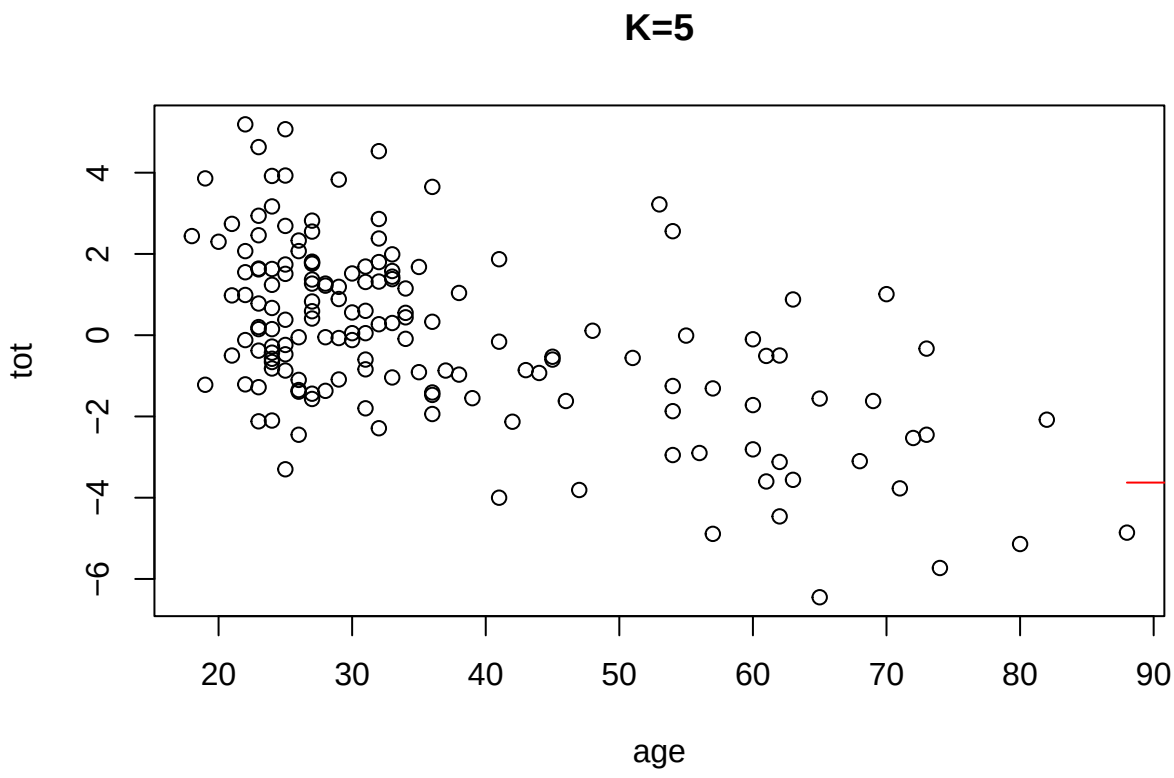
Change the range of the grid for the test set to [88,120]. Repeat task 3 for this test set. How do the predictions change? What's the problem of the predictions?

Solution:

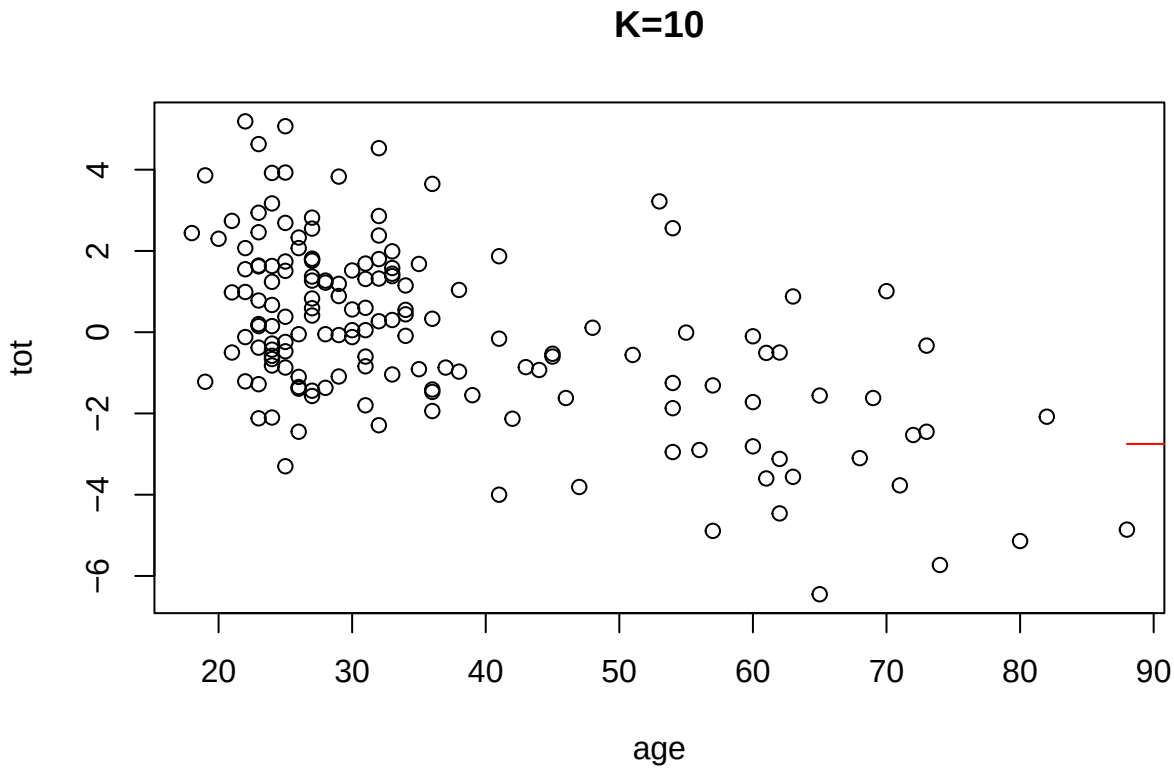
```
data.2b<-seq(min(88), max(120), by=0.01)
grid.of.age3<-data.frame(age=data.2b)

KNN_Pred1 <- knn.reg(x_kidney, grid.of.age3, y_kidney, k = 1)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=1",
     col="black")
lines(grid.of.age3$age, KNN_Pred1$pred, col = "red", lwd = 1)
```



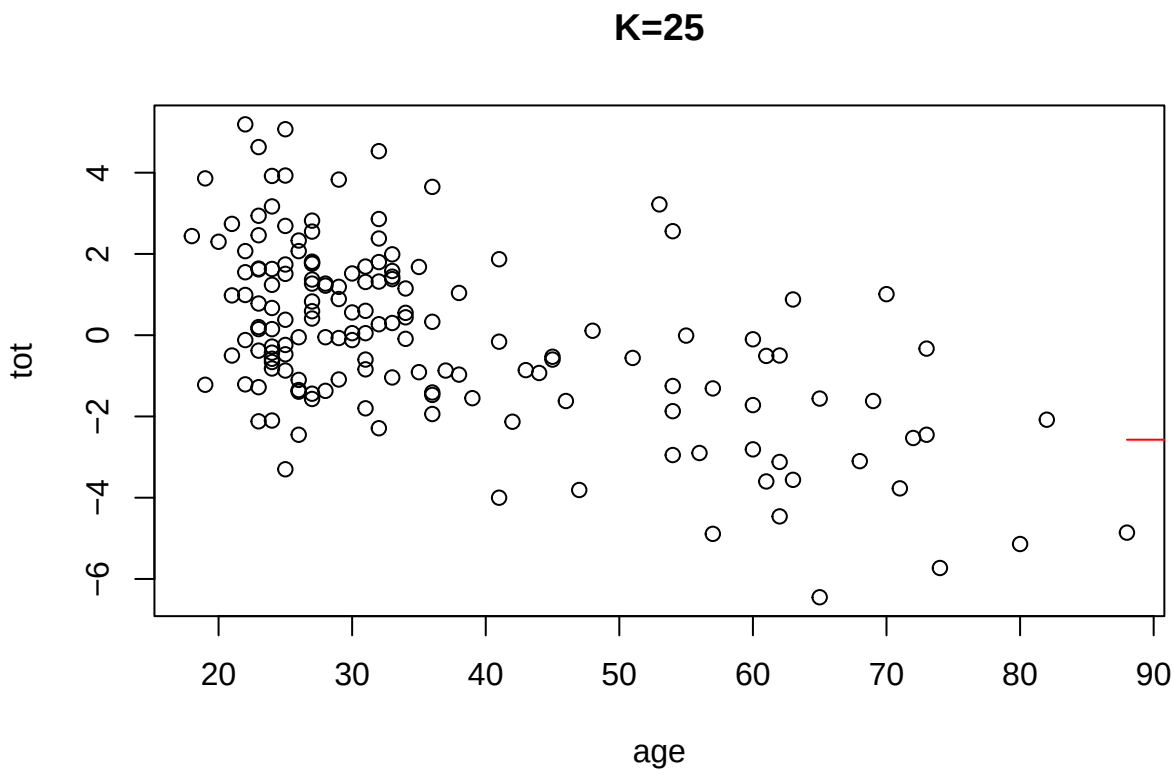

```
KNN_Pred5 <- knn.reg(x_kidney, grid.of.age3, y_kidney, k = 5)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=5",
     col="black")
lines(grid.of.age3$age, KNN_Pred5$pred, col = "red", lwd = 1)
```



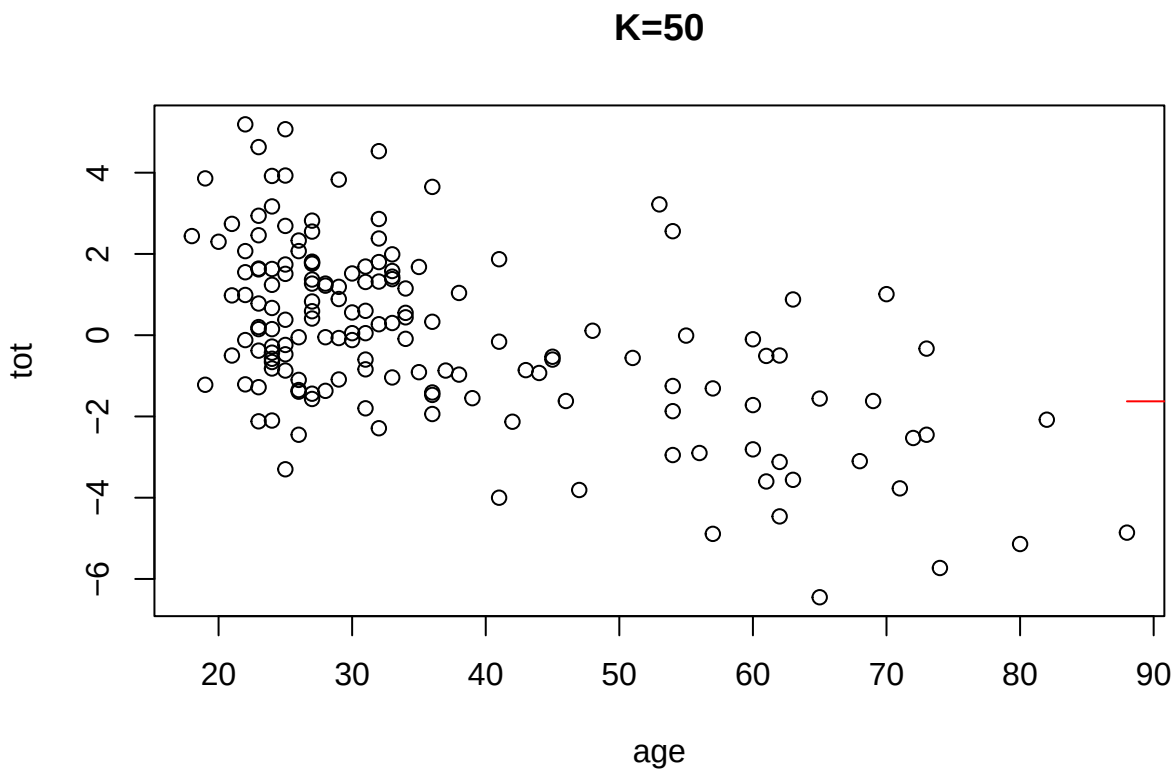

```
KNN_Pred10 <- knn.reg(x_kidney, grid.of.age3, y_kidney, k = 10)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=10",
     col="black")
lines(grid.of.age3$age, KNN_Pred10$pred, col = "red", lwd = 1)
```



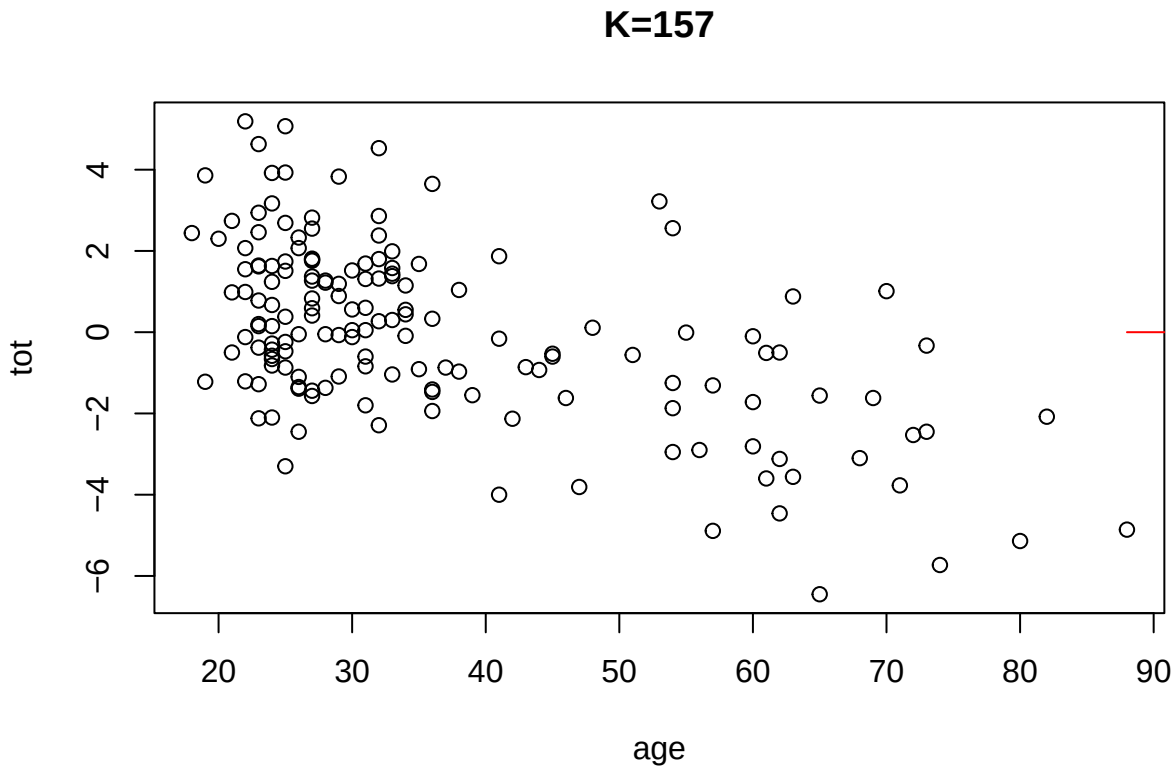

```
KNN_Pred25 <- knn.reg(x_kidney, grid.of.age3, y_kidney, k = 25)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=25",
     col="black")
lines(grid.of.age3$age, KNN_Pred25$pred, col = "red", lwd = 1)
```




```
KNN_Pred50 <- knn.reg(x_kidney, grid.of.age3, y_kidney, k = 50)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=50",
     col="black")
lines(grid.of.age3$age, KNN_Pred50$pred, col = "red", lwd = 1)
```




```
KNN_Pred157 <- knn.reg(x_kidney, grid.of.age3, y_kidney, k = 157)
plot(kidney$age, kidney$tot,
     xlab = "age",
     ylab= "tot",
     main= "K=157",
     col="black")
lines(grid.of.age3$age, KNN_Pred157$pred, col = "red", lwd = 1)
```




The line of prediction is now on the right side of the training set, as oppose to the left side in the previous question. The age range [1, 18] is outside the training set the same way it was in the previous question. As k increases, the line of prediction increases, which is different from what was observed in the previous question. However, the increase of the line of prediction does not make the prediction more useful of have meaning.