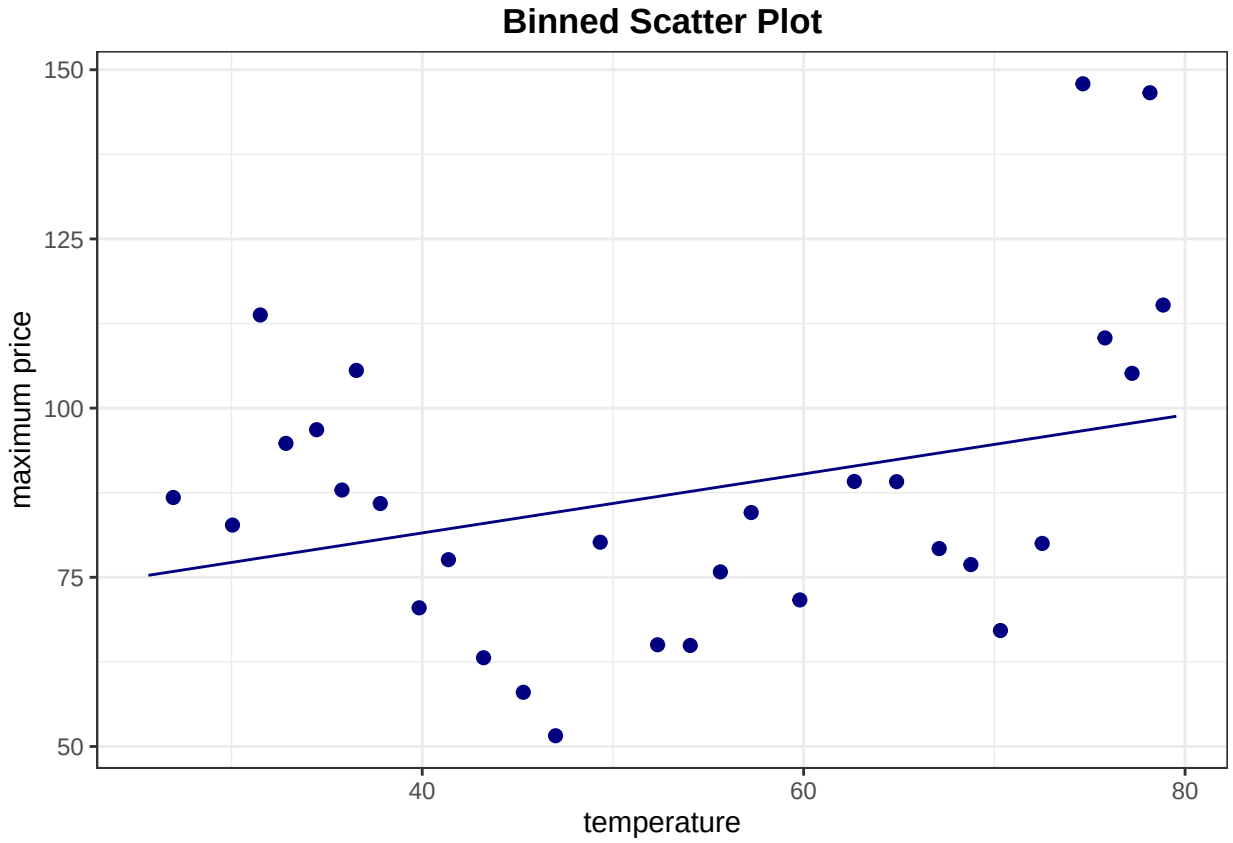
Empirical Methods for Applied Micro

Problem Set 2

Alberto Cappello

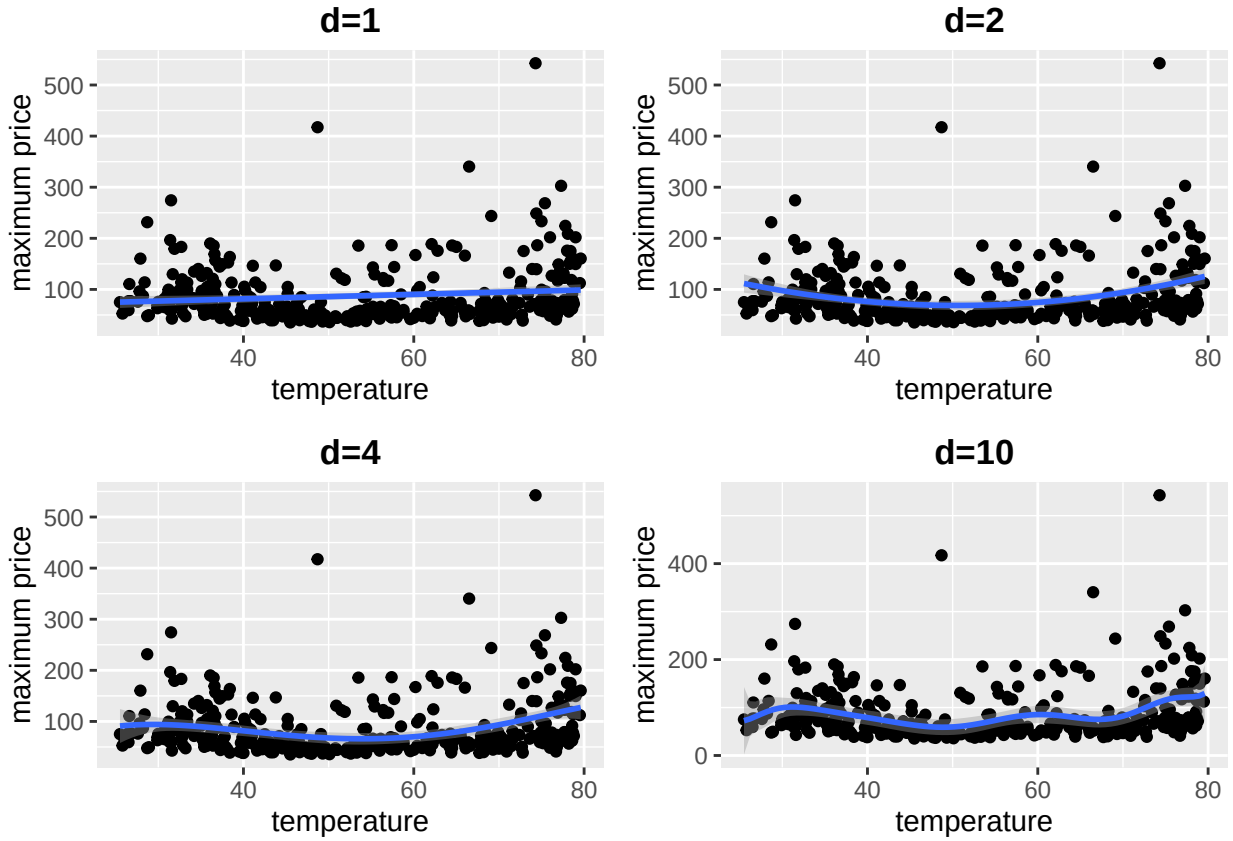
2/9/2021

Binned Scatterplot



	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	360	1151831.14				
2	359	1074482.49	1	77348.65	25.92	0.0000
3	358	1069812.47	1	4670.02	1.57	0.2118
4	357	1067825.95	1	1986.51	0.67	0.4151
5	356	1058421.54	1	9404.41	3.15	0.0767
6	355	1055415.63	1	3005.91	1.01	0.3162
7	354	1050341.67	1	5073.96	1.70	0.1931
8	353	1049634.46	1	707.21	0.24	0.6267
9	352	1048883.61	1	750.85	0.25	0.6162
10	351	1047367.22	1	1516.39	0.51	0.4764

Polynomial Regression



Cross Validation

```
df.shuffled <- df[sample(nrow(df)),]
K=10; order=10
folds <- cut(seq(1,nrow(df.shuffled)),breaks=K,labels=FALSE)
#Creating empty object to hold fit information
rmse = matrix(NA,nrow=K,ncol=order)

#Perform K-fold cross validation
for(i in 1:K){
```

```r
    #Segement data by fold using the which() function
    testIndexes <- which(folds==i,arr.ind=TRUE)
    testData <- df.shuffled[testIndexes, ]
    trainData <- df.shuffled[-testIndexes, ]
    #Use the test and train data partitions
    #Model fitting and evaluation
    for (j in 1:order){
      #training regression model on training folds
        fit.train = lm(lmp ~ poly(temp,j), data = trainData)
        #evaluating fit on the test fold
            fit.test = predict(fit.train, newdata=testData)
        rmse[i,j] = sqrt(mean(fit.test - testData$lmp)^2)
    }
}

#Averaging fit at each order
cvmse <- colMeans(rmse)
#plotting cross-validated prediction accuracy
plot(cvmse, type='l',xlab = "degree",ylab="CV MSE")
```