Challenge B

Philip Petersen (github.com/philipjadach), Bastian Lippmann (github.com/bastilip) Carles Mano (github.com/carlesmano)

December 8, 2017

Task 1B - Predicting house prices in Ames, Iowa

Step 1

We choose a random forest method. Random Forests considers two types of "trees"; classification and regression trees. Thereby Random Forest is an ensemble learning method for classification and regression trees. In a regression tree the dependent variable is continuoues, and in a classification tree the dependent variable is discrete. Random Forests correct for the habbit of the decision trees to overfit to their training set.

Step 2

Before testing the random forest method on the training we want to get rid of variables with a lot of missing data. We also want to get rid of the variabel Id.

```
train1 <- names(train) %in% c("Id")
train_without_id <- train[!train1]
```

Now the variable Id is no longer a part of the train dataset. Hereby we can use the random forest method on the new dataset. Not only do we have to remove the variable Id, but we also have to get rid of any missing values. If the variables have a lot of missing values, then we will get rid of these variables in the dataset.

Here we want to remove all variables, which have more than 100 missing observations. After that we summarise our dataset to see is there is still some missing observations left.

```
##          feature missing.observations
## 1    MasVnrType                    8
## 2    MasVnrArea                    8
## 3      BsmtQual                   37
## 4      BsmtCond                   37
## 5  BsmtExposure                   38
## 6  BsmtFinType1                   37
## 7  BsmtFinType2                   38
## 8    Electrical                    1
## 9    GarageType                   81
## 10  GarageYrBlt                   81
## 11 GarageFinish                   81
## 12   GarageQual                   81
## 13   GarageCond                   81
```

The answer here is yes, and we therefore remove the missing observations. Finally, we check to see if the dataset is all clean, and the result indicates that there are no rows left with missing observations.

```
## [1] feature              missing.observations
## <0 rows> (or 0-length row.names)
```

After taking the missing values out of the dataset, we now return to our random forest method. Before testing the random forest method we first set a seed. Now we can test the random forest method on our data set, and this gives us the following:

```
##
## Call:
##  randomForest(formula = SalePrice ~ ., data = train_without_id,      importance = FALSE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 24
##
##           Mean of squared residuals: 791626791
##                     % Var explained: 87.28
```

We can see that after testing the random forest method on our data, we can explain 87.28% of the variables.

Step 3

Firstly, we run a linear regression with $SalePrice$ as our dependent variable. Next, we want to keep the variables in the linear regression model that have coefficients significant at the 1% level.

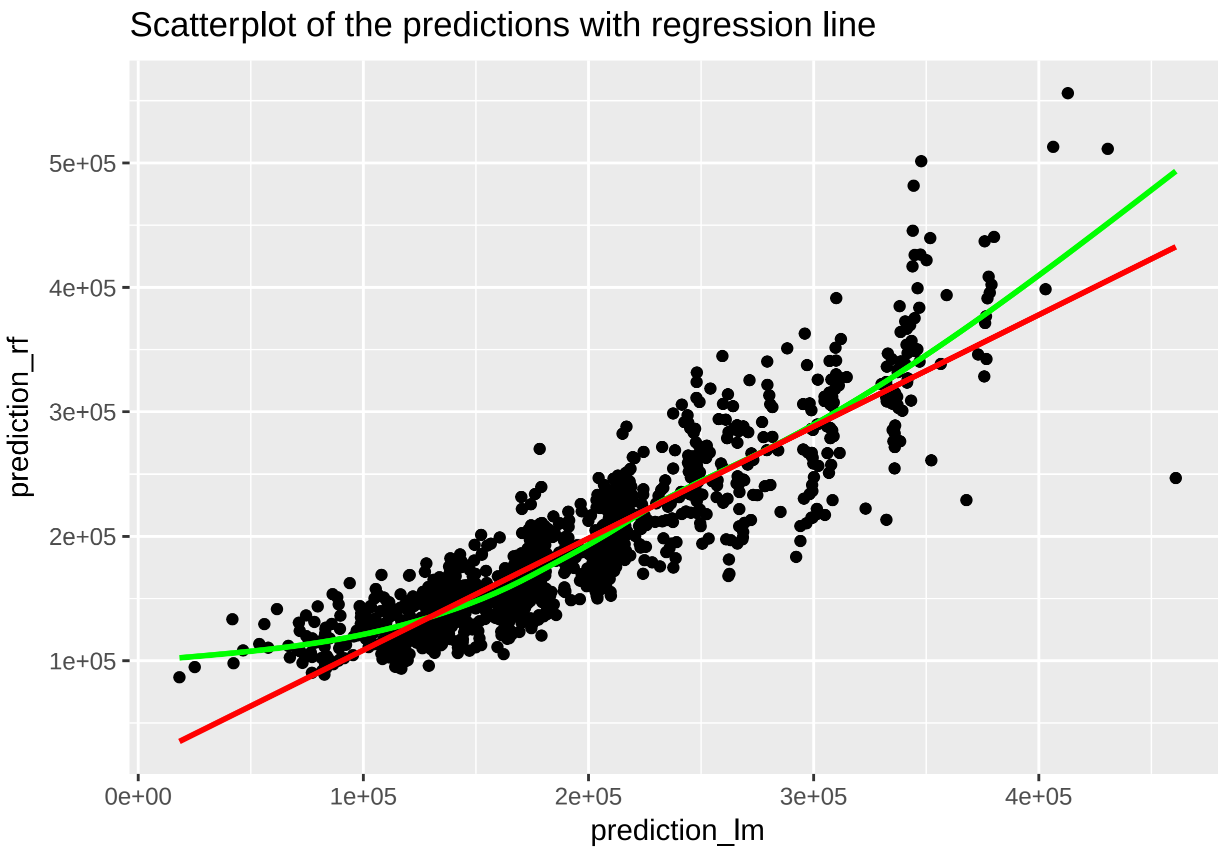
This gives us a smaller regression model where the dependent variable is still $SalePrice$ and we choose the variables in our model as $MSZoning, LotArea, Neighborhood, YearBuilt, OverallQual$. This linear regression model gives an R^2 above 70%. This can be seen from result below:

```
##
## Call:
## lm(formula = SalePrice ~ MSZoning + LotArea + Neighborhood +
##     YearBuilt + OverallQual, data = train_without_id)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217205  -22277   -1857   16999  342090
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -7.800e+05  1.820e+05  -4.286 1.95e-05 ***
## MSZoningFV           2.296e+03  2.059e+04   0.111 0.911242
## MSZoningRH          -3.057e+03  2.193e+04  -0.139 0.889180
## MSZoningRL           2.271e+04  1.769e+04   1.284 0.199468
## MSZoningRM           3.500e+03  1.673e+04   0.209 0.834332
## LotArea              1.221e+00  1.239e-01   9.859  < 2e-16 ***
## NeighborhoodBlueste  1.107e+04  3.195e+04   0.346 0.729084
```

```
## NeighborhoodBrDale  -7.151e+03  1.637e+04  -0.437 0.662347    
## NeighborhoodBrkSide  3.049e+04  1.420e+04   2.147 0.032005 *  
## NeighborhoodClearCr  3.779e+04  1.384e+04   2.731 0.006405 ** 
## NeighborhoodCollgCr  1.636e+04  1.078e+04   1.518 0.129350    
## NeighborhoodCrawfor  5.981e+04  1.301e+04   4.596 4.73e-06 ***
## NeighborhoodEdwards  1.212e+04  1.208e+04   1.003 0.316055    
## NeighborhoodGilbert  9.123e+03  1.129e+04   0.808 0.419164    
## NeighborhoodIDOTRR   3.084e+04  1.640e+04   1.881 0.060220 .  
## NeighborhoodMeadowV  2.584e+04  1.737e+04   1.488 0.137089    
## NeighborhoodMitchel  1.815e+04  1.235e+04   1.470 0.141705    
## NeighborhoodNAmes    1.809e+04  1.132e+04   1.598 0.110363    
## NeighborhoodNPkVill -3.695e+03  1.745e+04  -0.212 0.832348    
## NeighborhoodNWAmes   2.151e+04  1.159e+04   1.857 0.063553 .  
## NeighborhoodNoRidge  1.045e+05  1.221e+04   8.558  < 2e-16 ***
## NeighborhoodNridgHt  7.421e+04  1.135e+04   6.536 9.05e-11 ***
## NeighborhoodOldTown  3.344e+04  1.441e+04   2.321 0.020423 *  
## NeighborhoodSWISU    2.592e+04  1.538e+04   1.685 0.092206 .  
## NeighborhoodSawyer   2.034e+04  1.202e+04   1.692 0.090816 .  
## NeighborhoodSawyerW  2.118e+04  1.182e+04   1.792 0.073427 .  
## NeighborhoodSomerst  3.461e+04  1.371e+04   2.525 0.011704 *  
## NeighborhoodStoneBr  7.518e+04  1.326e+04   5.668 1.78e-08 ***
## NeighborhoodTimber   2.624e+04  1.254e+04   2.092 0.036620 *  
## NeighborhoodVeenker  5.095e+04  1.636e+04   3.114 0.001886 ** 
## YearBuilt            3.539e+02  9.076e+01   3.899 0.000101 ***
## OverallQual          3.337e+04  1.295e+03  25.763  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 41850 on 1306 degrees of freedom
## Multiple R-squared:  0.7253, Adjusted R-squared:  0.7188 
## F-statistic: 111.2 on 31 and 1306 DF,  p-value: < 2.2e-16
```

After defining a linear regression model, we now want make predictions of the random forest method on the test data. We therefore download the test data, and make the predictions. Next, we make a plot to compare the predictions of random forest and the linear regression.

```
## `geom_smooth()` using method = 'gam'
```



Task 2B - Overfitting in Machine Learning

Step 1

We estimate the low-flexibility local linear model on the training data with a bandwith on 0.5. Here we also calculate the fitted values for this model, which we are going to use later in step 3 to make the plot. When estimating this model we get and R^2 of approximately 85%.

```
##
## Regression Data: 122 training points, in 1 variable(s)
##                x
## Bandwidth(s): 0.5
##
## Kernel Regression Estimator: Local-Linear
## Bandwidth Type: Fixed
## Residual standard error: 1.085438
## R-squared: 0.8540956
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 1
```

Step 2

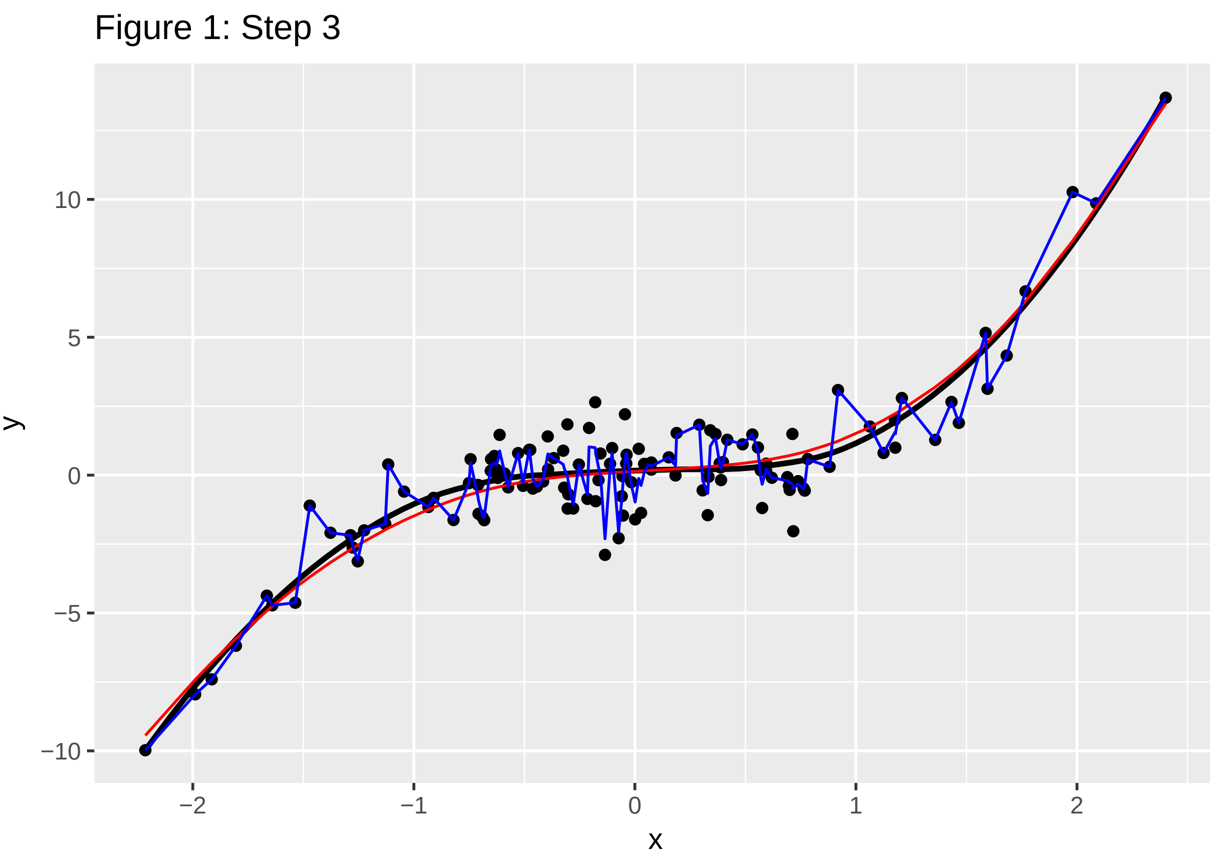
We estimate the high-flexibility local linear model on the training data with a bandwith on 0.01. When estimating this model we get an R^2 of approximately 97%.

```
##
## Regression Data: 122 training points, in 1 variable(s)
##                  x
## Bandwidth(s): 0.01
##
## Kernel Regression Estimator: Local-Linear
## Bandwidth Type: Fixed
## Residual standard error: 0.5070779
## R-squared: 0.9680171
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 1
```

Step 3

After estimating both the high- and low-flexibility local linear model, we try to make the scatterplot of x-y along with the calculated fitted values of the two models.

```
## `geom_smooth()` using method = 'loess'
```



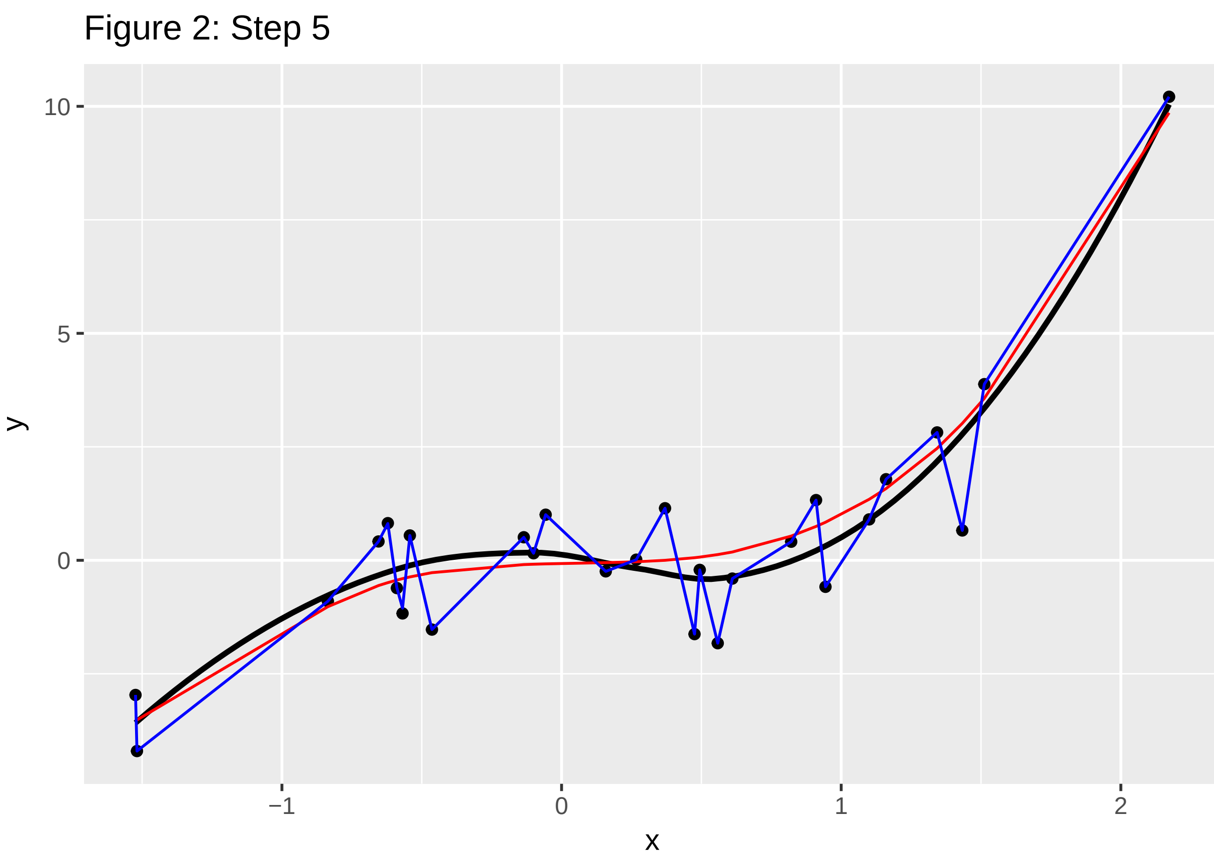
Step 4

You can see from figure 1 above, that the high-flexibility local linear model (blue line) has the least bias. This model has a lot of variance. It is typically this bias-variance trade off one has to face when working on non-parametric estimation. Here it is the choice of the bandwidth which makes the blue line more gittery - it is over-fitting the data. So one has to keep in mind to balance this bias-variance trade-off. Bias occurs if the bandwidth is high, and variance if the bandwidth is low. A method of choosing the right number of bandwidth can be The Jackknife Cross-Validation Method. Here you have a cross-validation function, and the right number of bandwidths to use, is the number that minimizes this cross-validation function.

Step 5

Now we turn to the same problem as in step 3, but instead we consider the test data. This means that we first estimate the high- and low-flexibility local linear model on the test data before making the plot. Again we also calculate the fitted values of the two models on the test data. After doing this, we can now make the scatterplot.

```
## `geom_smooth()` using method = 'loess'
```



Step 6

We create a vector going from 0.01 to 0.5 and every time add a step of 0.001.

```
v <- seq(from = 0.01, to = 0.5, by = 0.001)
```

Step 7

Here we estimate a local linear model on the training data with each bandwidth. We start by creating an empty list. In this list we want to store the output from our estimation. The estimation is such that we estimate the model for each bandwidth, and each estimation is being added into the list. So for everytime it runs the estimation for a new bandwidth the output is being added to the list. We do this by using a loop. The code is included here to illustrate this.

```
list_train <- list()
v <- seq(from = 0.01, to = 0.5, by = 0.001)
for (i in v) {
    output_ll_train <- npreg(y ~ x, data = training_set, bws = i, regtype = "ll")
    list_train[[length(list_train) + 1]] <- output_ll_train
}
```

Step 8

After estimating our model in step 7 we now want to focus on the MSE. From the estimation the MSE is included into the list we created before. Therefore, we want to create a vector from the list of each bandwidth, which only consists of the MSE. The code can be seen below.

```
vector_train <- c()
for (i in 1:length(v)) {
    vector_train <- append(vector_train, list_train[[i]]$MSE)
}
```

Step 9

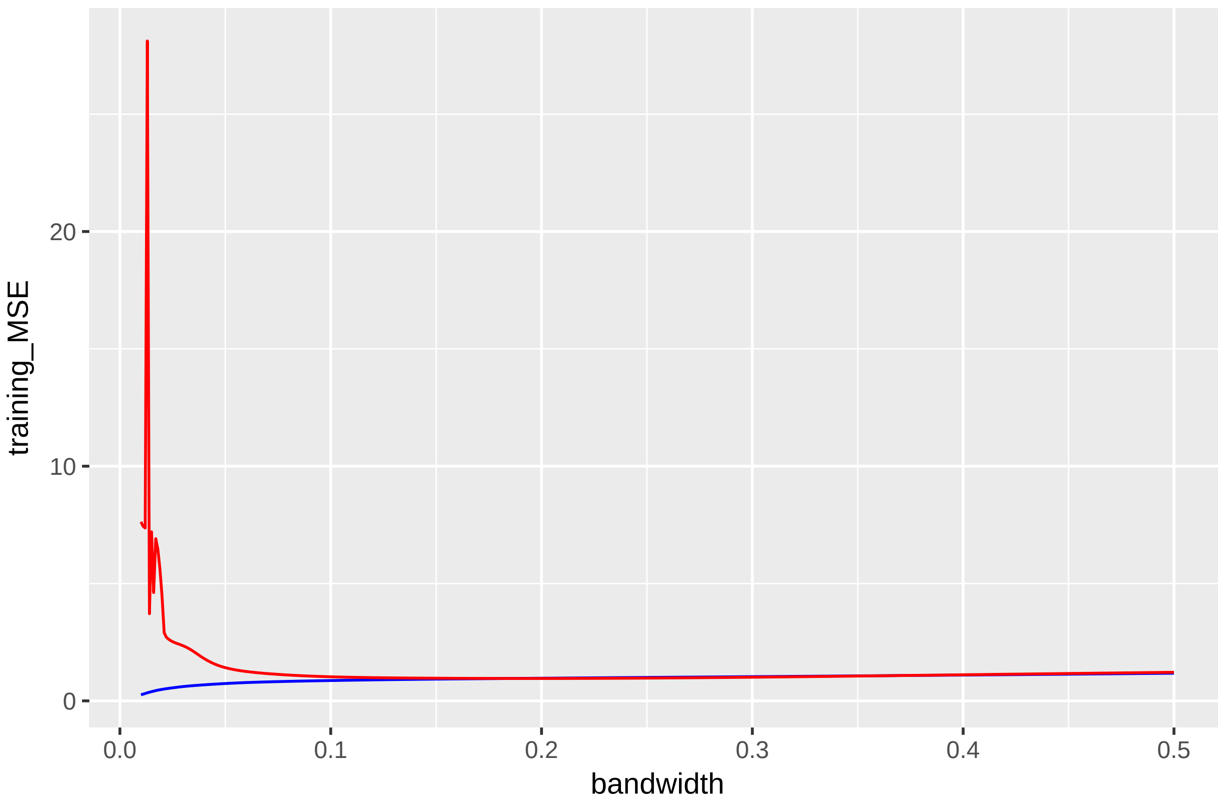
We still have to have the model from step 7 in mind. This is still the model we are going to use to compute the MSE on the test data. Firstly, we have to make predictions from the model in step 7, but on the new data set (test data). This means we have to find $\hat{y}$ for each bandwidth. Finally, we create a vector of the MSE. But now we cannot simply find the MSE in a list compared to earlier. Therefore, we have to write the equation for the MSE into our code in order to create the vector. This can be seen from the code below:

`r   test_MSE <- c()   for (i in 1:length(v)) {       y_hat <- predict(object = list_train[[i]], newdata = test_set)       test_MSE <- append(test_MSE, mean((y_hat - test_set$y)^2))   }`

Step 10

After finding the MSE in both the training data and the test data, we can now draw both MSE in the same plot.

Figure 3: Step 10



Task 3B - Privacy regulation compliance in France

Step 1

We looked on the website for the document, downloaded it and opened the file with the read() command.

```
data_cnil <- read.csv(file = "CNIL.csv", header = TRUE, quote = "", sep = ";")
head(data_cnil)
```

```
##   X...Siren        Responsable                       Adresse Code_Postal
## 1 788349926 """LA RIVE BLEUE"""              3/5 RUE BOILEAU       49100
## 2 421715731          01 DIRECT      58 AVENUE DE RIVESALTES       66240
## 3 409869708       01DB-METRAVIB       200 CHEMIN DES ORMEAUX       69760
## 4 444600464          1.2.3. SAS 57-59 -61 RUE HENRI BARBUSSE       92110
## 5 922002968       100 % ASNIERES       70 AVENUE D'ARGENTEUIL       92600
## 6 429621311          1000MERCIS         28 RUE DE CHATEAUDUN       75009
##          Ville
## 1       ANGERS
## 2 SAINT ESTEVE
## 3     LIMONEST
## 4       CLICHY
## 5     ASNIERES
```

```
## 6        PARIS
##                                                                              NAF
## 1                  8790A Autres activit\303\251s d'h\303\251bergement social
## 2                                526B Commerce de d\303\251tail hors magasin
## 3           7120B Activit\303\251s de contr\303\264le et analyses techniques
## 4 524C Autres commerces de d\303\251tail en magasin sp\303\251cialis\303\251
## 5                                     913C Autres organisations associatives
## 6      6201Z Programmation, conseil et autres activit\303\251s informatiques
##         TypeCIL                    Portee
## 1       INTERNE                   Etendue
## 2       EXTERNE G\303\251n\303\251rale
## 3 PROFESSIONNEL                   Etendue
## 4       EXTERNE                   Etendue
## 5       INTERNE                   Etendue
## 6       INTERNE                   Etendue
```

Step 2

We created a new variable by splitting up the first two digits of the postal code and called it "department".

```r
twodigits <- t(sapply(data_cnil$Code_Postal, function(x) substring(x, first = c(1,
    2), last = c(2, 4))))
data_cnil1 <- cbind(data_cnil, twodigits[, 1])
colnames(data_cnil1) <- c("SIREN", "Responsable", "Adresse", "Code_Postal", "Ville",
    "NAF", "TypeCIL", "Portee", "Department")
```

We remove any duplicates because we need just unique combinations

```r
data_cnil_unique <- unique(data_cnil1[c("SIREN", "Department")])
head(data_cnil_unique)
```

```
##       SIREN Department
## 1 788349926         49
## 2 421715731         66
## 3 409869708         69
## 4 444600464         92
## 5 922002968         92
## 6 429621311         75
```

We want determine the amount of designated an CIL-responsible for each department by firm. Doing so using the unique function.

```r
table_uniq <- table(unlist(duplicated(data_cnil_unique$SIREN)))
table_uniq
```

```
##
## FALSE  TRUE
## 17756   240
```

So there are 17756 firms with one unique responisble, and 240 firms have one for two or more departemens.

Step 3

First we have to import the SIREN dataset. Doing so by the read-command. By including some arguments in read-command we can reduce the working process. We plug in following arguments.

```
data_siren <- read.table(file = "siren.csv", header = TRUE, fill = TRUE, sep = ";",
    na.strings = "EMPTY", strip.white = TRUE, comment.char = "", stringsAsFactors = FALSE,
    nrows = 1048576)
```

We merge the list of CIL representatives and the SIREN data set by the variable "SIREN" since this variable is the same in the two data sets.

```
x <- data_cnil1
y <- data_siren
data_merge <- merge(x, y, by = "SIREN", all = FALSE)
```

Data_merge only contains the firms that have a cil representative in the data set SIREN.

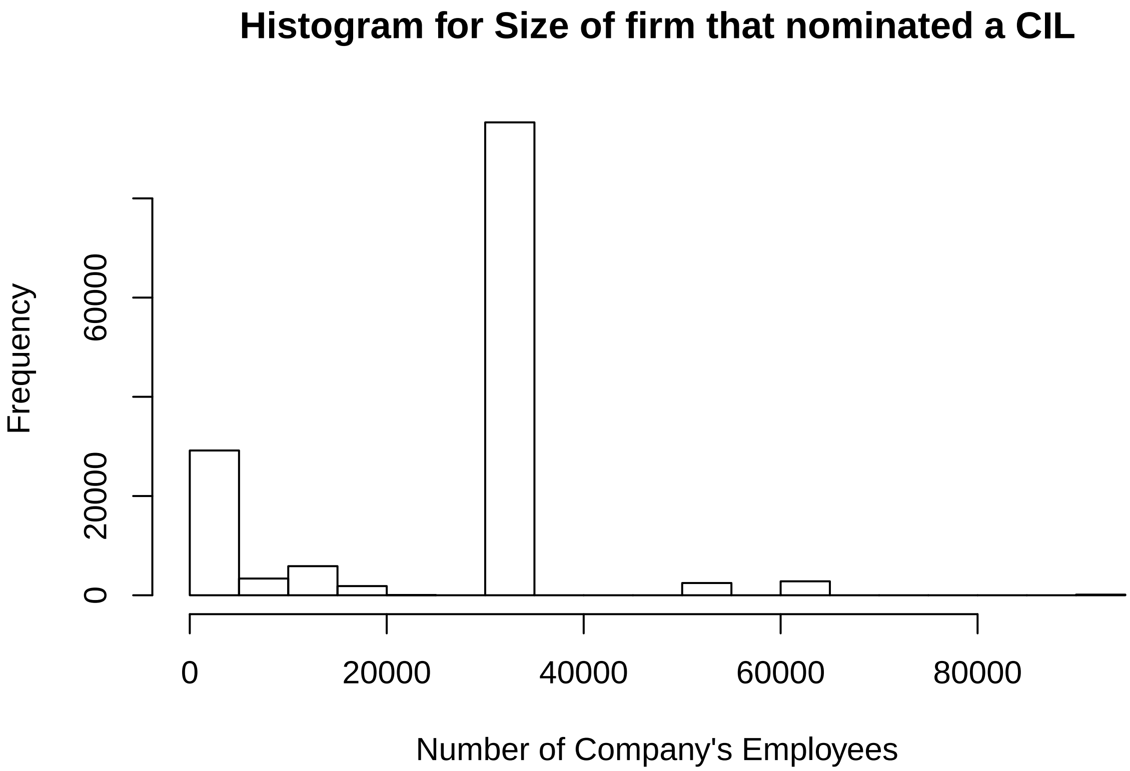
Step 4

We simply plot a histogram by using the variable EFENCENT. This works out since we only selected firms with CIL. See Appendix.

```
histo <- transform(data_merge, EFENCENT = as.numeric(EFENCENT))
```

```
## Warning: NAs introduced by coercion
```

```
hist(histo$EFENCENT, main = "Histogram for Size of firm that nominated a CIL", xlab = "Number o
```



Appendx

```r
head(histo)
```

```
##       SIREN                                                   Responsable
## 1 100000017                             PRESIDENCE DE LA REPUBLIQUE
## 2 110000023                                                     SENAT
## 3 110000122 COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTES
## 4 110000239                          AUTORITE DES MARCHES FINANCIERS
## 5 110000262                                       COUR DE CASSATION
## 6 110000296              CONSEIL SUP\303\211RIEUR DE L'AUDIOVISUEL
##                               Adresse Code_Postal          Ville
## 1        55 RUE DU FAUBOURG SAINT HONORE       75008          PARIS
## 2                   15 RUE DE VAUGIRARD      750291          PARIS
## 3          3 PLACE DE FONTENOY TSA 80715       75334 PARIS CEDEX 07
## 4                 17 PLACE DE LA BOURSE       75002          PARIS
## 5                    5, QUAI DE L'HORLOGE       75001          PARIS
## 6 39/43 QUAI ANDR\303\211 CITRO\303\213N       75015          PARIS
##                                                                     NAF
## 1 8411Z Administration g\303\251n\303\251rale, \303\251conomique et sociale
## 2 8411Z Administration g\303\251n\303\251rale, \303\251conomique et sociale
## 3
## 4 8411Z Administration g\303\251n\303\251rale, \303\251conomique et sociale
## 5                            8423Z Services de pr\303\251rogative publique
## 6 8411Z Administration g\303\251n\303\251rale, \303\251conomique et sociale
```

```
##   TypeCIL  Portee Department NIC                        L1_NORMALISEE
## 1 INTERNE Etendue         75  10      REPUBLIQUE FRANCAISE PRESIDENCE
## 2 INTERNE Etendue         75  17                                SENAT
## 3 INTERNE Etendue         75  33 COMMISSION NAT INFORMATIQUE LIBERTES
## 4 INTERNE Etendue         75  19      AUTORITE DES MARCHES FINANCIERS
## 5 INTERNE Etendue         75  11                    COUR DE CASSATION
## 6 INTERNE Etendue         75 241   CONSEIL SUPERIEUR DE L AUDIOVISUEL
##                            L2_NORMALISEE L3_NORMALISEE
## 1                                  PARIS 8
## 2                                  PARIS 6
## 3                         TSA80715 PARIS 7
## 4                                  PARIS 2
## 5                                  PARIS 1
## 6 COMITE TERRITORIAL DE L AUDIOVISUEL DE
##                      L4_NORMALISEE L5_NORMALISEE          L6_NORMALISEE
## 1 55 RUE DU FAUBOURG SAINT HONORE                 75800 PARIS CEDEX 08
## 2             15 RUE DE VAUGIRARD                 75291 PARIS CEDEX 06
## 3      3 PLACE DE FONTENOY UNESCO                 75334 PARIS CEDEX 07
## 4           17 PLACE DE LA BOURSE                 75082 PARIS CEDEX 02
## 5             5 QUAI DE L HORLOGE                 75055 PARIS CEDEX 01
## 6           69 RUE ANATOLE FRANCE               63000 CLERMONT FERRAND
##   L7_NORMALISEE                          L1_DECLAREE
## 1        FRANCE      REPUBLIQUE FRANCAISE PRESIDENCE
## 2        FRANCE                                SENAT
## 3        FRANCE COMMISSION NAT INFORMATIQUE LIBERTES
## 4        FRANCE      AUTORITE DES MARCHES FINANCIERS
## 5        FRANCE                    COUR DE CASSATION
## 6        FRANCE   CONSEIL SUPERIEUR DE L'AUDIOVISUEL
##                              L2_DECLAREE L3_DECLAREE
## 1
## 2
## 3
## 4
## 5
## 6 COMITE TERRITORIAL DE L AUDIOVISUEL DE
##                        L4_DECLAREE      L5_DECLAREE            L6_DECLAREE
## 1 55 RUE DU FAUBOURG SAINT HONORE          PARIS 8   75800 PARIS CEDEX 08
## 2             15 RUE DE VAUGIRARD          PARIS 6   75291 PARIS CEDEX 06
## 3       3 PL DE FONTENOY - UNESCO TSA80715 PARIS 7   75334 PARIS CEDEX 07
## 4                17 PL DE LA BOURSE          PARIS 2   75082 PARIS CEDEX 02
## 5             5 QUAI DE L HORLOGE          PARIS 1   75055 PARIS CEDEX 01
## 6           69 RUE ANATOLE FRANCE                 63000 CLERMONT FERRAND
##   L7_DECLAREE NUMVOIE INDREP TYPVOIE                  LIBVOIE CODPOS CEDEX
## 1                  55            RUE DU FAUBOURG SAINT HONORE  75008 75800
## 2                  15            RUE             DE VAUGIRARD  75006 75291
## 3                   3             PL    DE FONTENOY - UNESCO  75007 75334
## 4                  17             PL             DE LA BOURSE  75002 75082
## 5                   5           QUAI             DE L HORLOGE  75001 75055
```

```
## 6                  69            RUE          ANATOLE FRANCE  63000    NA
##   RPET                   LIBREG DEPET ARRONET CTONET COMET           LIBCOM
## 1   11        \316le-de-France    75       1     NA   108           PARIS 8
## 2   11        \316le-de-France    75       1     NA   106           PARIS 6
## 3   11        \316le-de-France    75       1     NA   107           PARIS 7
## 4   11        \316le-de-France    75       1     NA   102           PARIS 2
## 5   11        \316le-de-France    75       1     NA   101           PARIS 1
## 6   84 Auvergne-Rh\364ne-Alpes    63       2     99   113 CLERMONT FERRAND
##   DU TU UU      EPCI TCD ZEMET SIEGE
## 1 00  8 51 200054781  80  1101     1
## 2 00  8 51 200054781  80  1101     1
## 3 00  8 51 200054781  80  1101     1
## 4 00  8 51 200054781  80  1101     1
## 5 00  8 51 200054781  80  1101     1
## 6 63  7  1 246300701  61  8310     0
##                                          ENSEIGNE IND_PUBLIPO DIFFCOM
## 1                                                           0       0
## 2                                                           1       0
## 3                                                           1       0
## 4                                                           0       0
## 5                                                           1       0
## 6 COMITE TERRITORIAL DE L AUDIOVISUEL DE CLERMONT-FD           0       0
##   AMINTRET NATETAB LIBNATETAB APET700
## 1   201209      NA              8411Z
## 2   201209      NA              8411Z
## 3   201611      NA              8411Z
## 4   201209      NA              8411Z
## 5   201209      NA              8423Z
## 6   201704      NA              8411Z
##                               LIBAPET DAPET TEFET
## 1 Administration publique g\351n\351rale  2008    NN
## 2 Administration publique g\351n\351rale  2008    51
## 3 Administration publique g\351n\351rale  2016    22
## 4 Administration publique g\351n\351rale  2008    00
## 5                               Justice  2008    41
## 6 Administration publique g\351n\351rale  2017    NN
##                     LIBTEFET EFETCENT DEFET ORIGINE    DCRET  DDEBACT
## 1    Unit\351s non employeuses       NN    NA       1 19830301 19830301
## 2 2 000 \340 4 999 salari\351s     3300  2015       1 19830301 19830301
## 3     100 \340 199 salari\351s       NN  2016         20161101 20161101
## 4                0 salari\351        0  2015       1 20031101 20031101
## 5     500 \340 999 salari\351s      500  2015       1 19830301 19830301
## 6    Unit\351s non employeuses       NN    NA       1 20170501 20170501
##   ACTIVNAT LIEUACT ACTISURF SAISONAT MODET PRODET PRODPART AUXILT
## 1       NR      99       NA        P     S      N       NA      0
## 2       NR      99       NA        P     S      N       NA      0
## 3                        NA        P     S      N       NA      0
## 4       NR      99       NA        P     S      N       NA      0
```

```
## 5       NR      99       NA        P     S      N       NA      0
## 6                        NA        P     S      N       NA      0
##                                                    NOMEN_LONG SIGLE NOM PRENOM
## 1                       REPUBLIQUE FRANCAISE PRESIDENCE
## 2                                                       SENAT
## 3 COMMISSION NATIONALE DE L'INFORMATIQUE ET DES LIBERTES  CNIL
## 4                       AUTORITE DES MARCHES FINANCIERS   AMF
## 5                                    COUR DE CASSATION
## 6                    CONSEIL SUPERIEUR DE L'AUDIOVISUEL   CSA
##   CIVILITE RNA NICSIEGE RPEN DEPCOMEN ADR_MAIL   NJ
## 1       NA           10   11    75108       NA 7111
## 2       NA           17   11    75106       NA 7111
## 3       NA           33   11    75107       NA 7112
## 4       NA           19   11    75102       NA 7112
## 5       NA           11   11    75101       NA 7112
## 6       NA           27   11    75115       NA 7112
##                                                         LIBNJ APEN700
## 1                        Autorit\351 constitutionnelle    8411Z
## 2                        Autorit\351 constitutionnelle    8411Z
## 3           Autorit\351 administrative ind\351pendante    8411Z
## 4 Autorit\351 administrative ou publique ind\351pendante   8411Z
## 5           Autorit\351 administrative ind\351pendante    8423Z
## 6 Autorit\351 administrative ou publique ind\351pendante   8411Z
##                               LIBAPEN DAPEN APRM ESS DATEESS TEFEN
## 1 Administration publique g\351n\351rale  2008              NA    NN
## 2 Administration publique g\351n\351rale  2008              NA    51
## 3 Administration publique g\351n\351rale  2008              NA    31
## 4 Administration publique g\351n\351rale  2008              NA    00
## 5                               Justice  2008              NA    41
## 6 Administration publique g\351n\351rale  2008              NA    32
##                    LIBTEFEN EFENCENT DEFEN CATEGORIE    DCREN AMINTREN
## 1    Unit\351s non employeuses       NA    NA       PME 19470116   201209
## 2 2 000 \340 4 999 salari\351s     3300  2015       ETI 19590108   201209
## 3     200 \340 249 salari\351s      200  2015       PME 19780106   201209
## 4                0 salari\351        0  2015       PME 20031101   201209
## 5     500 \340 999 salari\351s      500  2015       ETI 19590108   201209
## 6     250 \340 499 salari\351s      300  2015       ETI 19861001   201209
##   MONOACT MODEN PRODEN ESAANN TCA ESAAPEN ESASEC1N ESASEC2N ESASEC3N
## 1       1     S      N     NA  NA
## 2       1     S      N     NA  NA
## 3       1     S      N     NA  NA
## 4       1     S      N     NA  NA
## 5       1     S      N     NA  NA
## 6       1     S      N     NA  NA
##   ESASEC4N VMAJ VMAJ1 VMAJ2 VMAJ3             DATEMAJ
## 1            NA    NA    NA    NA 1983-03-01T00:00:00
## 2            NA    NA    NA    NA 2013-06-18T00:00:00
## 3            NA    NA    NA    NA 2016-11-25T00:00:00
```

```
## 4              NA    NA    NA    NA 2016-09-15T00:00:00
## 5              NA    NA    NA    NA 1999-07-29T00:00:00
## 6              NA    NA    NA    NA 2017-04-04T00:00:00
```