Для початку завантажимо необхідні бібліотеки та датасет.

```
library(caret)
library(class)
library(dplyr)
library(psych)
library(e1071)
library(lattice)

train <- read.csv("D:/train.csv", header = TRUE)
train <- train[1:996,]
test3 <- train
head(test3)
```

```
##                 Dates       Category                     Descript DayOfWeek PdDistrict     Resolution
## 1 2015-05-13 23:53:00       WARRANTS                WARRANT ARREST Wednesday   NORTHERN ARREST, BOOKED
## 2 2015-05-13 23:53:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST Wednesday   NORTHERN ARREST, BOOKED
## 3 2015-05-13 23:33:00 OTHER OFFENSES       TRAFFIC VIOLATION ARREST Wednesday   NORTHERN ARREST, BOOKED
## 4 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO Wednesday   NORTHERN           NONE
## 5 2015-05-13 23:30:00  LARCENY/THEFT   GRAND THEFT FROM LOCKED AUTO Wednesday       PARK           NONE
## 6 2015-05-13 23:30:00  LARCENY/THEFT GRAND THEFT FROM UNLOCKED AUTO Wednesday  INGLESIDE           NONE
##                     Address         X        Y
## 1        OAK ST / LAGUNA ST -122.4259 37.77460
## 2        OAK ST / LAGUNA ST -122.4259 37.77460
## 3 VANNESS AV / GREENWICH ST -122.4244 37.80041
## 4  1500 Block of LOMBARD ST -122.4270 37.80087
## 5 100 Block of BRODERICK ST -122.4387 37.77154
## 6       0 Block of TEDDY AV -122.4033 37.71343
```

```
str(test3)
```

```
## 'data.frame':    996 obs. of  9 variables:
##  $ Dates     : chr  "2015-05-13 23:53:00" "2015-05-13 23:53:00" "2015-05-13 23:33:00" "2015-05-13 23:30:00"
...
##  $ Category  : chr  "WARRANTS" "OTHER OFFENSES" "OTHER OFFENSES" "LARCENY/THEFT" ...
```

```
##  $ Descript  : chr  "WARRANT ARREST" "TRAFFIC VIOLATION ARREST" "TRAFFIC VIOLATION ARREST" "GRAND THEFT FROM L
OCKED AUTO" ...
##  $ DayOfWeek : chr  "Wednesday" "Wednesday" "Wednesday" "Wednesday" ...
##  $ PdDistrict: chr  "NORTHERN" "NORTHERN" "NORTHERN" "NORTHERN" ...
##  $ Resolution: chr  "ARREST, BOOKED" "ARREST, BOOKED" "ARREST, BOOKED" "NONE" ...
##  $ Address   : chr  "OAK ST / LAGUNA ST" "OAK ST / LAGUNA ST" "VANNESS AV / GREENWICH ST" "1500 Block of LOMBA
RD ST" ...
##  $ X         : num  -122 -122 -122 -122 -122 ...
##  $ Y         : num  37.8 37.8 37.8 37.8 37.8 ...
```

Далі, проведемо **попередню обробку даних**:
де потрібно, змінимо типи атрибутів, видалимо незначущі змінні (котрі мають один рівень, і можуть бути конвертовані неправильно).

```
test3$Dates <- strptime(test3$Dates, "%Y-%m-%d %H:%M:%S")
test3$Dates <- as.POSIXct(test3$Dates, format="%Y-%m-%d %H:%M:%S")
test3$Month <- factor(format(test3$Dates, "%m"))
test3$Year <- factor(format(test3$Dates, "%Y"))
test3$Day <- factor(format(test3$Dates,"%d"))
test3$Hour <- factor(format(test3$Dates,"%H"))

test3 <- test3[,-c(1, 3, 6, 7, 11)]
test3$Month <- ifelse(test3$Month == "05", 1, 0)
test4 <- test3
test4 <- test4[,-c(6,8)]
```

Тепер ту ж процедуру, але вже із використанням інших функцій, проводимо із рештою змінних:

```
DayOfWeek <- as.data.frame(dummy.code(test4$DayOfWeek))
PdDistrict <- as.data.frame(dummy.code(test4$PdDistrict))
test4$Day <- as.numeric(as.factor(test4$Day))
test4 <- cbind(test4, DayOfWeek, PdDistrict)
test4 <- test4[,-c(2,3)]
```

Тепер переходимо до етапу **розподілу та підготовкою датасету**:

1) перш за все необхідно виділити прибзино 70-75% даних із вибірки в тренувальний сет;
2) далі важливо змаштабувати дані, тобто привести їх всіх до єдиного діапазону значень;
3) також потрібно відділити цільову змінну (в нашому випадку - Category) окремо від залежних змінних.

```
ran <- sample(1:nrow(test4), 0.75 * nrow(test4))
scale <- function(x) { (x -min(x))/(max(x)-min(x))   }
data_sc <- as.data.frame(lapply(test4[, c(2:ncol(test4))], scale))
summary(data_norm)
```

```
## < table of extent 0 x 0 >
```

```
rem <- nearZeroVar(data_norm)
data_norm <- data_norm[,-rem]
data_train <- data_norm[ran,]
data_test <- data_norm[-ran,]
data_target_category <- test4[ran,1]
data_test_category <- test4[-ran,1]
```

Наступний етап - **побудова моделі кластерного аналізу**:

* за допомогою методу knn() створюємо нашу модель, де параметрами будуть раніше відібрані вибірки та змінні;
* одразу порівняємо отримані результати зі спостережуваними даними;
* як видно, модель не дуже точно визначає потрібні категорії, тому спробуймо щось виправити.

```
set.seed(1234)
knn_method <- knn(train = data_train, test = data_test, cl = data_target_category, k=10)
knn_outcome <- data.frame(data_test_category)
class_comparison <- data.frame(knn_method, knn_outcome)
```

```
names(class_comparison) <- c("Predicted", "Observed")
head(class_comparison)
```

```
##       Predicted       Observed
## 1 LARCENY/THEFT OTHER OFFENSES
## 2 LARCENY/THEFT  VEHICLE THEFT
## 3 LARCENY/THEFT   NON-CRIMINAL
## 4 LARCENY/THEFT  LARCENY/THEFT
## 5 LARCENY/THEFT        ROBBERY
## 6 LARCENY/THEFT   NON-CRIMINAL
```

Застосуємо іншу функцію - train з пакета caret і спробуймо спрогнозувати значення з цією моделлю.
На жаль, ми отримали невисоку точність, що свідчить про непридатність використання методу knn на цьому датасеті.

```
knn_pred_class <- train(data_train, data_target_category, method = "knn")
prediction <- predict(knn_pred_class, newdata = data_test)
ACC <- 100 * sum(data_test_category == prediction)/NROW(data_test_category)
cat("Accuracy: ", ACC)
```

```
## Accuracy:  27.30924
```