1 Practica

```
require("knitr")
opts_chunk$set(cache=TRUE, cache.path = 'test_files/', fig.path='figure/')
```

Chosen dataset.

I have chose Kickstarter Projects dataset from Kaggle's datasets repository because I am interested in becoming familiar with it due to a project that I am developing with a friend and which I want to publish in Kickstarter.\

Taking it's definition from Wikipedia, Kickstarter is an American public-benefit corporation that maintains a global crowdfunding platform focused on creativity and merchandising.\

The original dataset has 378661 registered projects with 15 attributes per instance. Before using this data set I'm going to remove instances (projects) in order to reduce the large amount of data and only use between 5000 and 10000 projects in this practice.

```
set.seed(1)
# reading dataset and removing instances which has NA values
data = na.omit(read.csv("KickStarterProjects2018.csv"))
```

Let's remove columns that are not interesting or redundant...

```
data[,c("ID","category","goal","pledged","usd.pledged")] = NULL

# renaming columns...
names(data) = c("Name","Category","Currency","Deadline","Launched","State","Backers","Country","Pledged
```

Let's modify Launched column in order to remove the launching hour which is not necessary

```
data$Launched = factor(substr(data$Launched,1,10))
```

Now let's transform Deadline and Launched from factor to Data class

```
data$Launched = as.Date(data$Launched,format = "%Y-%m-%d")
data$Deadline = as.Date(data$Deadline,format = "%Y-%m-%d")
```

We are interested in having the launching month and the days that the project was available in kickstarter, so let's do the needed operations to reach this information:

```
# Setting lifetime days in Kickstarter
data$LifetimeDays = as.numeric(data$Deadline-data$Launched)

# Setting launch month
data$Launched = factor(substr(data$Launched,6,7))
names(data)[5] = "LaunchMonth"
```

WE don't need anymore Deadline and Launched attributes...

```
data$Launched = NULL
data$Deadline = NULL
```

Now that we have the information that we want we take only a random sample of 7000 projects of the total length of projects in the data set

```
data = data[sample(c(1:nrow(data)),7000),]
```

To use “character” as attribute delimiter in the csv file i’ll create later, I’m going to replace these characters in the projects names that are using it with “:” character, because there are few projects using “;” character and I think it is a mistake.

```
data$Name = gsub(";",":",data$Name)
data$Name = gsub("\"","'",data$Name)
data$Name = gsub("~","-",data$Name)
```

Now let’s transform every attribute in the type in which we are interested to upload the database in our repository

```
data$LaunchMonth = as.numeric(data$LaunchMonth)
data$Backers = as.integer(data$Backers)
data$LifetimeDays = as.integer(data$LifetimeDays)
```

Finally we export de dataset into a new csv file:

```
write.table(data, "kickstarterProjects.csv", row.names = FALSE, col.names = FALSE, quote = FALSE,sep =
```

select country, currency, avg(pledged), avg(goal) from kp group by country, currency order by avg(goal) desc;
select country, count(state) from kp where state = ‘successful’ group by country order by count(state) desc;