# Project 2 Requirements

The goal of this project is to redo the linear modeling of only the best model from your Project 1 but on an Apache Spark platform. You should use R's sparklyr package to talk to an Apache Spark instance. The two tasks are:

1. Install sparklyr and Apache Spark on your computer
2. Run R code that uses Spark to redo the linear modeling

Note that it is not required to redo the exploratory data analysis in Project 1.

## Installing sparklyr and Apache Spark :

It is easiest to install from within RStudio (assuming that the tidyverse library is also installed).

1. Install package "sparklyr"
2. Load the sparklyr library and install Apache Spark using sparklyr:
    - `library(sparklyr)`
    - `spark_install()`

This should work in either Windows or Linux. More detailed instructions for installing on Linux from scratch is included at the end.

## Linear modeling in Spark

This code will be very similar to the code shown in class. Use the same data file from Project 1.

```
> mylocaldata <- read_csv
  ("http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.csv")
> library(sparklyr)
> sc <- spark_connect(master = "local")
> myremotedata <- copy_to(sc, mylocaldata)
> mymodel <- ml_linear_regression(x=myremotedata , formula =
  bodyfat ~ Weight + Height)
> summary(mymodel)
```

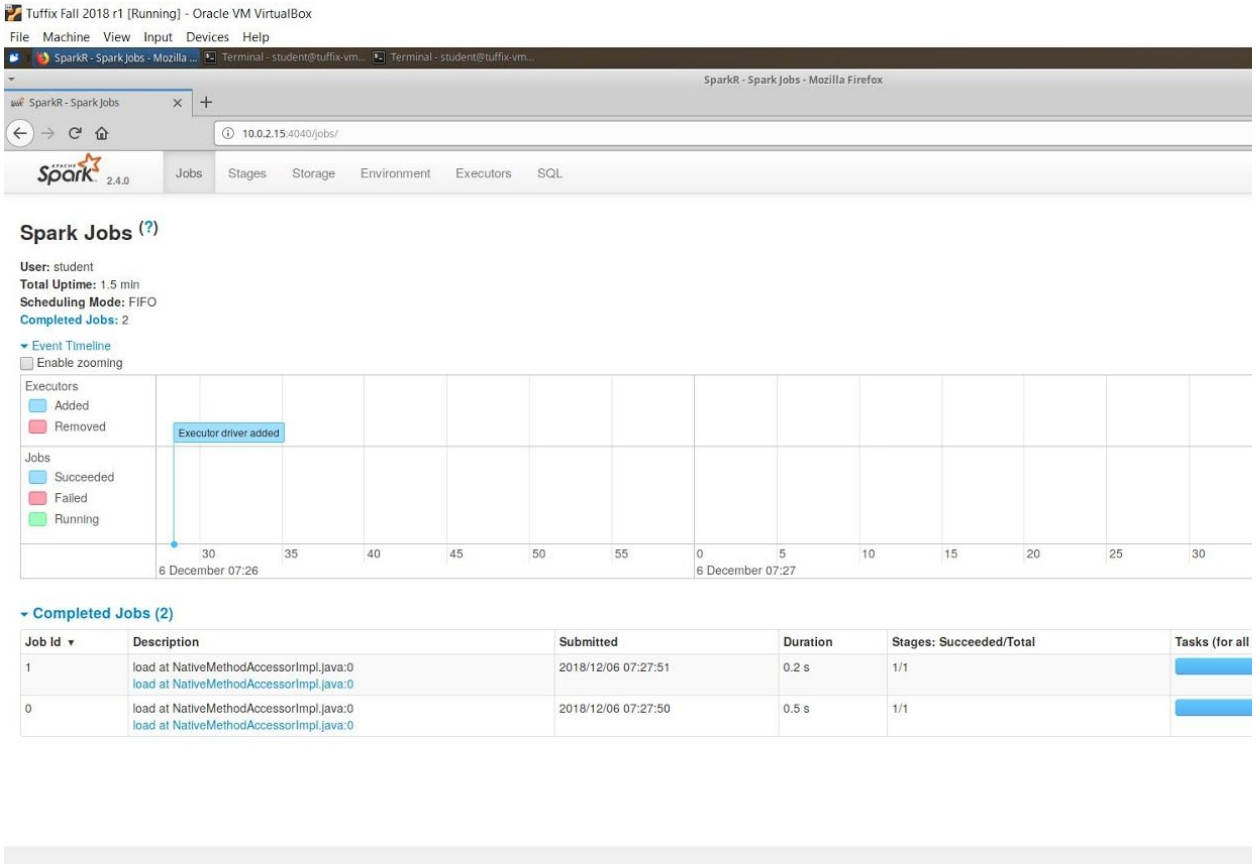
You will edit the above code to use the specific variables you used and perform any transformations that you used.

## Submission:

1. Prepare a short report containing only the following:
    a. Your R code
    b. Output of `summary(model)`
    c. A screen capture image of the running Apache Spark web UI. You can go to this webpage from R:

```
> spark_web(sc)
```

An example is shown below:




## Due date:

Friday 12/20, 11:55 pm on Titanium. Submit a single PDF file.

## Group work:

You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group needs to submit to Titanium.

## **Appendix**. Installation on Linux/Tuffix:

You may want to try out "Tuffix", the Titan-branded version of Ubuntu 18.04. Instructions on how to install Tuffix or a Tuffix-based VM are in the Tuffix Titanium Community for Students, https://communities.fullerton.edu/course/view.php?id=1547 (also the best venue to receive help with Tuffix). It is easiest to install into a Linux (virtual) machine. Then install R or Rstudio, install the sparklyr package inside R/Rstudio, and then use sparklyr's install.spark() function to do the Spark installation.

1. To install R, from the Linux command line:
   > `sudo apt install r-base`
2. To install RStudio:
   > Download the latest version (as a .deb file) from https://www.rstudio.com/products/rstudio/download/#download
   > `sudo apt install gdebi`
   > `sudo gdebi <location of downloaded rstudio .deb file>`
3. Make sure Java 8 is used.
   > `sudo apt install openjdk-8-jdk`
   > `sudo update-alternatives --config java`
   - This will show all currently installed versions of Java. Select Java 8 (openjdk-8-jdk)
4. Install the sparklyr package inside R/Rstudio:
   > `install.packages("tidyverse")`
   - If there are errors during installation of tidyverse, make sure these libraries are installed
     1. `sudo apt-get install libxml2-dev`
        - (for package xml2)
     2. `sudo apt-get install libcurl4-gnutls-dev`
        - (for package curl)

   > `install.packages("sparklyr")`

5. Install spark from inside R/Rstudio[1]
   > `library(sparklyr)`
   > `install.spark()`

[1] The above instructions install spark to a local folder. You can also install to a system-wide folder, a install a specific version of Spark, or use an existing installation of Spark.