Jared Castaneda
Khoa Do
Kurt Prutsman
CPSC 375

Project 2

1a. proj2.R:

```
# install.packages("sparklyr")
# library(sparklyr)
# spark_install()

library(tidyverse)
library(sparklyr)

setwd("c:/temp/cpsc375proj2/")

# mylocaldata <- read_csv ("http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Bodyfat.csv")
mylocaldata <- read_csv("Bodyfat.csv")

sc <- spark_connect(master = "local")
myremotedata <- copy_to(sc, mylocaldata, overwrite = TRUE)
# our group's formula for the best model that describes bodyfat
# bodyfat ~ Wrist + log(Abdomen) + Weight^2
# unfortunately, ml_linear_regression has a problem with the log and exponent functions \
# so we have to simplify the formula
mymodel <- ml_linear_regression(x=myremotedata , formula =
bodyfat~Wrist+Abdomen+Weight)

summary(mymodel)

spark_web(sc)
```

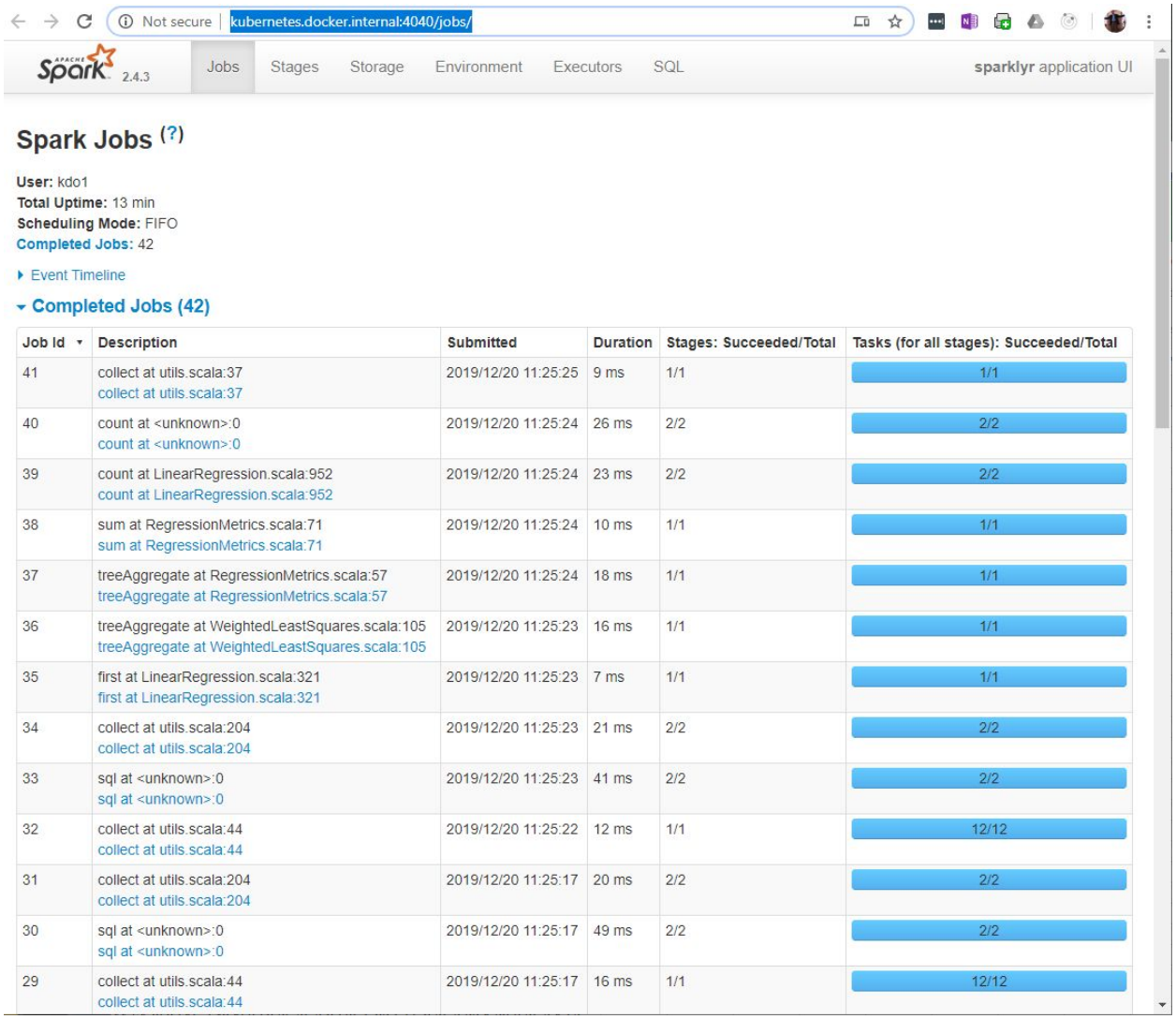
1b. Output of summary(model):

```
> summary(mymodel)
Deviance Residuals:
    Min      1Q  Median      3Q     Max
-13.0803  -3.2463  -0.2175   3.2472   9.8018

Coefficients:
(Intercept)       Wrist     Abdomen      Weight
-27.9299169  -1.2448589   0.9751296  -0.1144609

R-Squared: 0.7277
Root Mean Squared Error: 4.358
```

1c. A screen capture image of the running Apache Spark web UI.



Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
41	collect at utils.scala:37 collect at utils.scala:37	2019/12/20 11:25:25	9 ms	1/1	1/1
40	count at <unknown>:0 count at <unknown>:0	2019/12/20 11:25:24	26 ms	2/2	2/2
39	count at LinearRegression.scala:952 count at LinearRegression.scala:952	2019/12/20 11:25:24	23 ms	2/2	2/2
38	sum at RegressionMetrics.scala:71 sum at RegressionMetrics.scala:71	2019/12/20 11:25:24	10 ms	1/1	1/1
37	treeAggregate at RegressionMetrics.scala:57 treeAggregate at RegressionMetrics.scala:57	2019/12/20 11:25:24	18 ms	1/1	1/1
36	treeAggregate at WeightedLeastSquares.scala:105 treeAggregate at WeightedLeastSquares.scala:105	2019/12/20 11:25:23	16 ms	1/1	1/1
35	first at LinearRegression.scala:321 first at LinearRegression.scala:321	2019/12/20 11:25:23	7 ms	1/1	1/1
34	collect at utils.scala:204 collect at utils.scala:204	2019/12/20 11:25:23	21 ms	2/2	2/2
33	sql at <unknown>:0 sql at <unknown>:0	2019/12/20 11:25:23	41 ms	2/2	2/2
32	collect at utils.scala:44 collect at utils.scala:44	2019/12/20 11:25:22	12 ms	1/1	12/12
31	collect at utils.scala:204 collect at utils.scala:204	2019/12/20 11:25:17	20 ms	2/2	2/2
30	sql at <unknown>:0 sql at <unknown>:0	2019/12/20 11:25:17	49 ms	2/2	2/2
29	collect at utils.scala:44 collect at utils.scala:44	2019/12/20 11:25:17	16 ms	1/1	12/12