amunategui@gmail.com

amunategui.github.io

Explorations continued:

summarize (or agg) / groupBy (or group_by)

```
head(summarize(groupBy(diabetes_sparkdf, diabetes_sparkdf$insulin
== 'Up' & diabetes_sparkdf$readmitted==1),
               count = n(diabetes_sparkdf$insulin == 'Up' &
diabetes_sparkdf$readmitted==1)))

head(summarize(groupBy(diabetes_sparkdf, diabetes_sparkdf$insulin
== 'Up' & diabetes_sparkdf$readmitted==1),
               total_readmits =
sum(diabetes_sparkdf$readmitted)))



head(arrange(agg(groupBy(diabetes_sparkdf, diabetes_sparkdf$age),
total_diagnoses = sum(diabetes_sparkdf$number_diagnoses)),
     diabetes_sparkdf$age))
```

magrittr

It's all about pipes! Here is a great introduction from the author: http://www.r-statistics.com/2014/08/simpler-r-coding-with-pipes-the-present-and-future-of-the-magrittr-package/

```
# install.packages('magrittr')
library(magrittr)
groupBy(diabetes_sparkdf, diabetes_sparkdf$age) %>%
  summarize(total_diagnoses =
sum(diabetes_sparkdf$number_diagnoses)) %>%
  arrange(diabetes_sparkdf$age) %>%
        head
```

SparkSQL

```
# http://spark.apache.org/docs/latest/sql-programming-guide.html

# register data frame as an SQL table
registerTempTable(diabetes_sparkdf, "diabetes_sparktbl")

# SELECT
payor_code <- sql(sqlContext, "SELECT payer_code FROM
diabetes_sparktbl")
head(payor_code)

# SELECT DISTINCT
payor_code <- sql(sqlContext, "SELECT DISTINCT payer_code FROM
diabetes_sparktbl")
collect(payor_code)

# count
young_at_risk <- sql(sqlContext, "SELECT count(*) FROM
diabetes_sparktbl WHERE num_lab_procedures >= 30 AND age ='[10-
20)'")
head(young_at_risk)
```