Module 6

ae

12/7/2021

Loading the appropriate packages and connecting to spark

```
library(tidyverse)
library(sparklyr)

sc <- spark_connect(master = "spark://fire.truman.edu:7077")
```

```
## Error in system2(file.path(spark_home, "bin", "spark-submit"), "--version", : '"/bin/spark-submit"' n
```

Connecting to the files. I have commented out two files for the final homework output

```
filename <- "file:///opt/Data/nycflights13.csv"

hdfs_filename_13 <- "hdfs://fire.truman.edu:9000/user/aje3887/nycflights13.csv"

#hdfs_filename_09to12 <- "hdfs://fire.truman.edu:9000/user/aje3887/flights_09-12.csv"

#hdfs_filename_09 <- "hdfs://fire.truman.edu:9000/user/aje3887/2009.csv"

#df_tbl13 <- copy_to(sc, df)
#df_tbl09to12 <- copy_to(sc, df)
#df_tbl09 <- copy_to(sc, df)
```

First part of the homework asking us to calculate the mean arrival delay for each carrier, the mean arrival delay for each origin/destination pair, and a filter to include only the results of a single origin/desination pair.

I have commented out the timing on all code chunks. I have commented out the two files not loaded initially as well.

```
#system.time({
flights <- read.csv(filename) %>%
  select(arr_delay, dep_delay, carrier, origin, dest, distance) %>%
  na.omit()
```

```
## Error in file(file, "rt"): cannot open the connection
```

```
#})
```

#32mb

```
#system.time({
flights_tbl_13 <- spark_read_csv(sc, name="df_tbl13", path=hdfs_filename_13) %>%
  select(arr_delay, carrier, origin, dest, dep_delay, distance) %>%
  na.omit()
```

```
## Error in src_tbls(sc): object 'sc' not found
```

```
#})
```

#2.9gb

```
#system.time({
#flights_tbl_09to12 <- spark_read_csv(sc, name="df_tbl09to12", path=hdfs_filename_09to12) %>%
#  select(OP_CARRIER, ORIGIN, DEST, ARR_DELAY, DEP_DELAY, DISTANCE) %>%
#  na.omit()
#})
```

#756mb

```
#system.time({
#flights_tbl_09 <- spark_read_csv(sc, name="df_tbl09", path=hdfs_filename_09) %>%
#  select(OP_CARRIER, ORIGIN, DEST, ARR_DELAY, DISTANCE, DEP_DELAY) %>%
#  na.omit()
#})
```

```
#system.time({
flts <- flights %>%
  mutate(late = as.numeric(arr_delay > 0)) %>%
  group_by(carrier, dest, origin, distance) %>%
  summarize(Pct_Late = sum(late)/n()) %>%
  filter(origin == "JFK" && dest == "LAX")
```

```
## Error in mutate(., late = as.numeric(arr_delay > 0)): object 'flights' not found
```

```
knitr::kable(flts)
```

```
## Error in knitr::kable(flts): object 'flts' not found
```

```
#})
```

```
#system.time({
#flts09 <- flights_tbl_09 %>%
#  mutate(late = as.numeric(arr_delay > 0)) %>%
#  group_by(OP_CARRIER, DEST, ORIGIN) %>%
#  summarize(Pct_Late = sum(late)/n()) %>%
#  filter(origin == "JFK" && dest == "LAX")
#})
```

```
#system.time({
#flts09to12 <- flights_tbl_09to12 %>%
#  mutate(late = as.numeric(arr_delay > 0)) %>%
#  group_by(OP_CARRIER, DEST, ORIGIN) %>%
#  summarize(Pct_Late = sum(late)/n()) %>%
#  filter(origin == "JFK" && dest == "LAX")
#})
```

```
#system.time({
flts_tbl_13 <- flights_tbl_13 %>%
  mutate(late = as.numeric(arr_delay > 0)) %>%
  group_by(carrier, dest, origin, arr_delay, dep_delay, distance) %>%
  summarize(Pct_Late = sum(late)/n()) %>%
  filter(origin == "JFK" && dest == "LAX")
```

```
## Error in mutate(., late = as.numeric(arr_delay > 0)): object 'flights_tbl_13' not found
```

```
knitr::kable(flts_tbl_13)
```

```
## Error in knitr::kable(flts_tbl_13): object 'flts_tbl_13' not found
```

```
#})
```

Creating the linear model to predict arrival delay as a function of departuer delay, carrier, destination, and distance.

The timing has been commented out. The two files that were never loaded did not get a code chunk because they would follow the same exact format as the two formats below.

```
#system.time({
flights_lm13 <- flights %>%
  filter(origin == "JFK") %>%
  select(arr_delay, dep_delay, carrier, dest, distance)
```

```
## Error in filter(., origin == "JFK"): object 'flights' not found
```

```
fit13 <- lm(arr_delay ~ dep_delay + carrier + distance + dest, data = flights_lm13)
```

```
## Error in is.data.frame(data): object 'flights_lm13' not found
```

```
fit13
```

```
## Error in eval(expr, envir, enclos): object 'fit13' not found
```

```
#})
```

```
#system.time({
flights_lm_tbl13 <- flts_tbl_13 %>%
  filter(origin == "JFK") %>%
  select(arr_delay, dep_delay, carrier, distance, dest)
```

```
## Error in filter(., origin == "JFK"): object 'flts_tbl_13' not found
```

```
fit13_ml <- ml_linear_regression(flights_tbl_13, arr_delay ~ dep_delay + carrier + distance + dest)
```

```
## Error in ml_linear_regression(flights_tbl_13, arr_delay ~ dep_delay + : object 'flights_tbl_13' not
```

```
fit13_ml
```

```
## Error in eval(expr, envir, enclos): object 'fit13_ml' not found
```

```
#})
```

Predicting Arrival Delay from JFK to MIA with the given variables: 1090 Miles, AA, Dep_Delay = 0

```
newDF <- data.frame(
  distance = 1090,
  dep_delay = 0,
  carrier = "AA",
  origin = "JFK",
  dest = "MIA")

newflight_tbl <- copy_to(sc, newDF)
```

```
## Error in copy_to(sc, newDF): object 'sc' not found
```

```
predict(fit13, newdata=newDF)
```

```
## Error in predict(fit13, newdata = newDF): object 'fit13' not found
```

Neither predictive model worked. The error appears to be a mismatch between spark versions.

```
#predict(fit13_ml, newdata=newflight_tbl)

#in case the above predict doesn't work...
#ml_predict(fit13_ml, dataset=newflight_tbl)
```

I had an interesting output. For r,the average loading time was .907 For Spark, the average loading time was .733

However, for calculating the arrival delay... For r, the average time was .019 For Spark, the average time was .078

Then Spark come out on top with the linear model For r, the average time was 2.138 For Spark, the average time was 1.964

All three things considered, Spark outperformed R by .4 seconds on average. If that half-second was worth the extra steps in setup and analysis, I'm not sure. However, it may be worth the effort as the data becomes larger.

Unfortunately, using R through the VPN to access the Spark server made it impossible to knit the file as PDF Latex was not installed. Bringing the code over to my local R-Studio program made it impossible to access the Fire studio where all the files are. As a result the submission is a mix-mash of the two systems given the limitations each provided.