CPSC 375 Project 2 Report

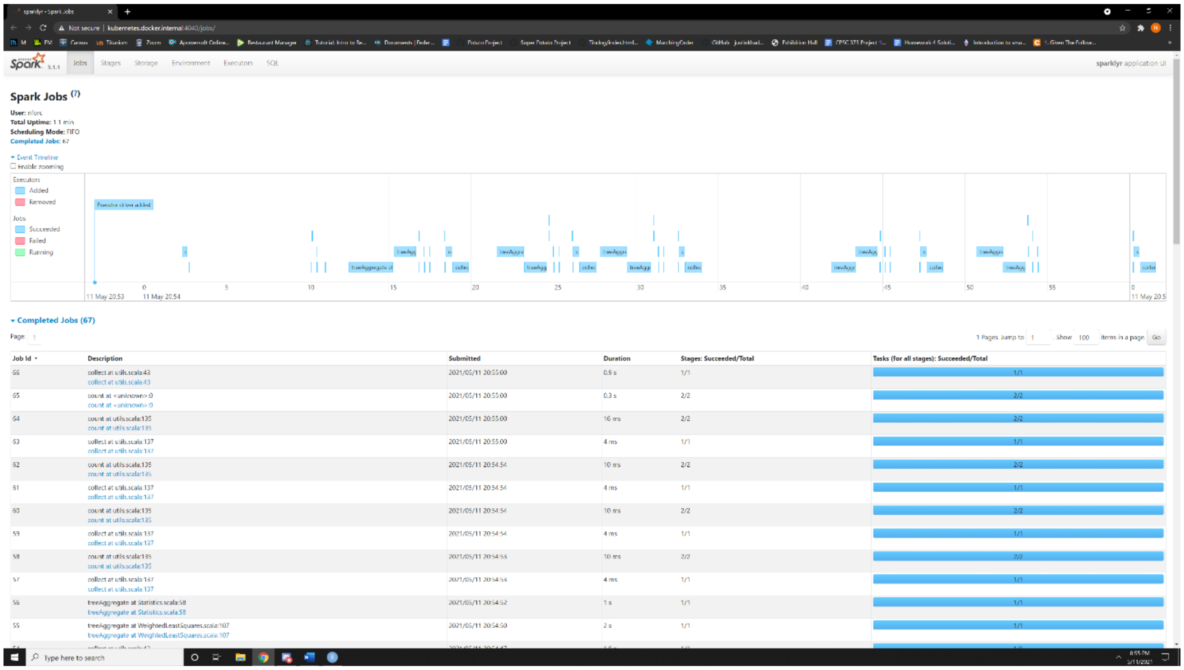
GROUP MEMBERS:

NATHAN MARCOS

```
#Using the saved R code from Project 1 to recreate the data

library(tidyverse)
library(ggplot2)
library(sparklyr)

#Covid Dataset
deaths <- read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-
19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_
deaths_global.csv")
confirmed <-
read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-
19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_
confirmed_global.csv")
deaths <- deaths %>% pivot_longer(-(1:4), names_to="Day", values_to="Deaths")
deaths <- deaths %>% rename(`Country`= `Country/Region`)
deaths <- deaths %>% group_by(Country, Day) %>%
summarize(Deaths=sum(Deaths))
deaths <- deaths[order(as.Date(deaths$Day, format="%m/%d/%Y")),]
deaths <- deaths %>% arrange(Country)

confirmed <- confirmed %>% pivot_longer(-(1:4), names_to="Day",
values_to="Confirmed")
confirmed <- confirmed %>% rename(`Country`= `Country/Region`)
confirmed <-  confirmed %>% group_by(Country, Day) %>%
summarize(Confirmed=sum(Confirmed))
confirmed <- confirmed[order(as.Date(confirmed$Day, format="%m/%d/%Y")),]
confirmed <- confirmed %>% arrange(Country)

covid <- deaths %>% full_join(confirmed)

#Hospital Beds Dataset
beds <- read_csv("C:/Users/nforc/Downloads/hospitalbeds.csv")
beds <- beds %>% group_by(Country) %>% filter(Year == max(Year))
beds <- beds %>% select(-Year)

#Demographics Dataset
demographics <- read_csv("C:/Users/nforc/Downloads/demographics.csv")
demographics <- demographics %>% select(`Country Name`, `Series Code`,
YR2015)
demographics <- demographics %>% pivot_wider(names_from = `Series Code`,
values_from = YR2015)
demographics <- demographics %>% rename(`Country`= `Country Name`)
demographics <- demographics %>% mutate("Population
(Total)"=SP.POP.TOTL.FE.IN+SP.POP.TOTL.MA.IN, "Population
(Urban)"=SP.URB.TOTL, "Population (80+)"=SP.POP.80UP.FE+SP.POP.80UP.MA,
"Population (65+)"=SP.POP.65UP.FE.IN+SP.POP.65UP.MA.IN, "Population (15-
64)"=SP.POP.1564.MA.IN+SP.POP.1564.FE.IN, "Population (0-
14)"=SP.POP.0014.MA.IN+SP.POP.1564.FE.IN)
demographics <- demographics %>% select(-(2:16))

#Join Datasets
covid <- covid %>% mutate(Country = replace(Country, Country == "Korea,
South", "South Korea")) %>% mutate(Country = replace(Country, Country ==
"US", "United States")) %>% mutate(Country = replace(Country, Country ==
"Congo (Brazzaville)", "Congo")) %>% mutate(Country = replace(Country,
Country == "Congo (Kinshasa)", "Congo"))
beds <- beds %>% mutate(Country = replace(Country, Country == "Republic of
Korea", "South Korea")) %>% mutate(Country = replace(Country, Country ==
"Iran (Islamic Republic of)", "Iran")) %>% mutate(Country = replace(Country,
Country == "United Kingdom of Great Britain and Northern Ireland", "United
Kingdom")) %>% mutate(Country = replace(Country, Country == "Bolivia
(Plurinational State of)", "Bolivia")) %>% mutate(Country = replace(Country,
Country == "United States of America", "United States")) %>% mutate(Country =
replace(Country, Country == "Democratic Republic of the Congo", "Congo"))
%>% mutate(Country = replace(Country, Country == "Venezuela (Bolivarian
Republic of)", "Venezuela"))
demographics <- demographics %>% mutate(Country = replace(Country, Country
== "Korea, Dem. People's Rep.", "South Korea")) %>% mutate(Country =
replace(Country, Country == "Korea, Rep.", "South Korea")) %>% mutate(Country
= replace(Country, Country == "Iran, Islamic Rep.", "Iran")) %>% mutate(Country
= replace(Country, Country == "Congo, Dem. Rep.", "Congo")) %>%
mutate(Country = replace(Country, Country == "Congo, Rep.", "Congo")) %>%
mutate(Country = replace(Country, Country == "Venezuela, RB", "Venezuela"))
mydata <- covid %>% full_join(beds) %>% full_join(demographics)

#Spark
mydata <- na.exclude(mydata)
sc <- spark_connect(master = "local")
myremotedata <- copy_to(sc, mydata)

#Model 1
mymodel1<- ml_linear_regression(x=myremotedata, formula = Deaths ~
Confirmed + Hospital_beds_per_10_000_population)

summary(mymodel1)
Deviance Residuals (approximate):
   Min      1Q  Median      3Q     Max
-190953   -3325   -2645   -1490  170644

Coefficients:
                  (Intercept)
                 3715.47716167
```

```
                        Confirmed
                       0.01902129
Hospital_beds_per_10_000_population
                     -26.78712697

R-Squared: 0.883
Root Mean Squared Error: 14350

#Model 2
mymodel2<- ml_linear_regression(x=myremotedata, formula = Deaths ~
Confirmed + Population_Urban)
summary(mymodel2)
Deviance Residuals (approximate):
    Min      1Q  Median      3Q     Max
-191234   -2826   -2444   -1999  170947

Coefficients:
     (Intercept)       Confirmed Population_Urban
    2.402685e+03    1.888106e-02     1.094594e-05

R-Squared: 0.8833
Root Mean Squared Error: 14340

#Model 3
mymodel3<- ml_linear_regression(x=myremotedata, formula = Deaths ~
Confirmed + Population_65 + Population_Total)
summary(mymodel3)
Deviance Residuals (approximate):
    Min      1Q  Median      3Q     Max
-165945   -2929   -2383   -1291  173124

Coefficients:
     (Intercept)       Confirmed    Population_65
    2.364015e+03    1.892911e-02     6.061368e-04
Population_Total
   -5.056071e-05

R-Squared: 0.8892
Root Mean Squared Error: 13970

#Model 4
myremotedata2 <- myremotedata %>% mutate (z =
Population_Urban/Population_Total)
mymodel4<- ml_linear_regression(x=myremotedata2, formula = Deaths ~
Confirmed + z)
summary(mymodel4)
```

```
Deviance Residuals (approximate):
     Min        1Q    Median        3Q       Max
-185789.1   -4198.8   -2259.2     279.2  170363.0

Coefficients:
  (Intercept)     Confirmed             z
-4.354539e+03  1.899646e-02  1.024845e+04

R-Squared: 0.8849
Root Mean Squared Error: 14240

#Model 5
mymodel5<- ml_linear_regression(x=myremotedata, formula = Deaths ~
Confirmed + Population_014)
summary(mymodel5)
Deviance Residuals (approximate):
    Min      1Q  Median      3Q     Max
-187735   -2961   -2839   -1949  171414

Coefficients:
   (Intercept)      Confirmed Population_014
  2.880775e+03   1.914719e-02  -8.845855e-06

R-Squared: 0.8831

#Web UI
spark_web(sc)
```



Spark Jobs (?)

User: nforc
Total Uptime: 1.1 min
Scheduling Mode: FIFO
Completed Jobs: 67

Completed Jobs (67)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
66	collect at utils.scala:43 collect at utils.scala:43	2021/05/11 20:55:00	0.9 s	1/1	1/1
65	count at <unknown>:0 count at <unknown>:0	2021/05/11 20:55:00	0.3 s	2/2	2/2
64	count at utils.scala:135 count at utils.scala:135	2021/05/11 20:55:00	16 ms	2/2	2/2
63	collect at utils.scala:137 collect at utils.scala:137	2021/05/11 20:55:00	4 ms	1/1	1/1
62	count at utils.scala:135 count at utils.scala:135	2021/05/11 20:54:54	10 ms	2/2	2/2
61	collect at utils.scala:137 collect at utils.scala:137	2021/05/11 20:54:54	4 ms	1/1	1/1
60	count at utils.scala:135 count at utils.scala:135	2021/05/11 20:54:54	10 ms	2/2	2/2
59	collect at utils.scala:137 collect at utils.scala:137	2021/05/11 20:54:54	4 ms	1/1	1/1
58	count at utils.scala:135 count at utils.scala:135	2021/05/11 20:54:53	10 ms	2/2	2/2
57	collect at utils.scala:137 collect at utils.scala:137	2021/05/11 20:54:53	4 ms	1/1	1/1
56	treeAggregate at Statistics.scala:58 treeAggregate at Statistics.scala:58	2021/05/11 20:54:52	1 s	1/1	1/1
55	treeAggregate at WeightedLeastSquares.scala:107 treeAggregate at WeightedLeastSquares.scala:107	2021/05/11 20:54:50	2 s	1/1	1/1

8:55 PM
5/11/2021