Introduction to sparklyr

We will largely follow chapters **2** and **3** of **Mastering Spark with R**, https://therinspark.com.

First install the following packages if you do not already have them already, and load them with the `library()` function:

```
library(sparklyr)
library(dplyr)
library(ggplot2)
library(knitr)
```

Preliminiaries

If you are working on EIDF, first make sure that the default working directory in RStudio is your folder. In RStudio, select Tools -> Global Options. Change the default working directory to be /work/eidf071/eidf071/.

To confirm the change has taken effect, close and then reopen RStudio, and type getpw() into the console. It should show your working directory correctly as above.

Connecting

```
sc = spark_connect(master = 'local')
```

Task 3

Which dplyr functions can be used to have a first look at the data? Run them.

```
#install.packages("nycflights13", "Lahman")
library(dplyr)
library(ggplot2)
library(jsonlite)
```

```
## Warning: package 'jsonlite' was built under R version 4.4.1
```

```
iris_tbl <- copy_to(sc, iris, overwrite = TRUE)
flights_tbl <- copy_to(sc, nycflights13::flights, "flights", overwrite = TRUE)
batting_tbl <- copy_to(sc, Lahman::Batting, "batting", overwrite = TRUE)
src_tbls(sc)
```

```
## [1] "batting" "flights" "iris"
```

```
flights_tbl %>% filter(dep_delay == 2)
```

```
## # Source:   SQL [?? x 19]
## # Database: spark_connection
##     year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
##  1  2013     1     1      517            515         2      830            819
##  2  2013     1     1      542            540         2      923            850
##  3  2013     1     1      702            700         2     1058           1014
```

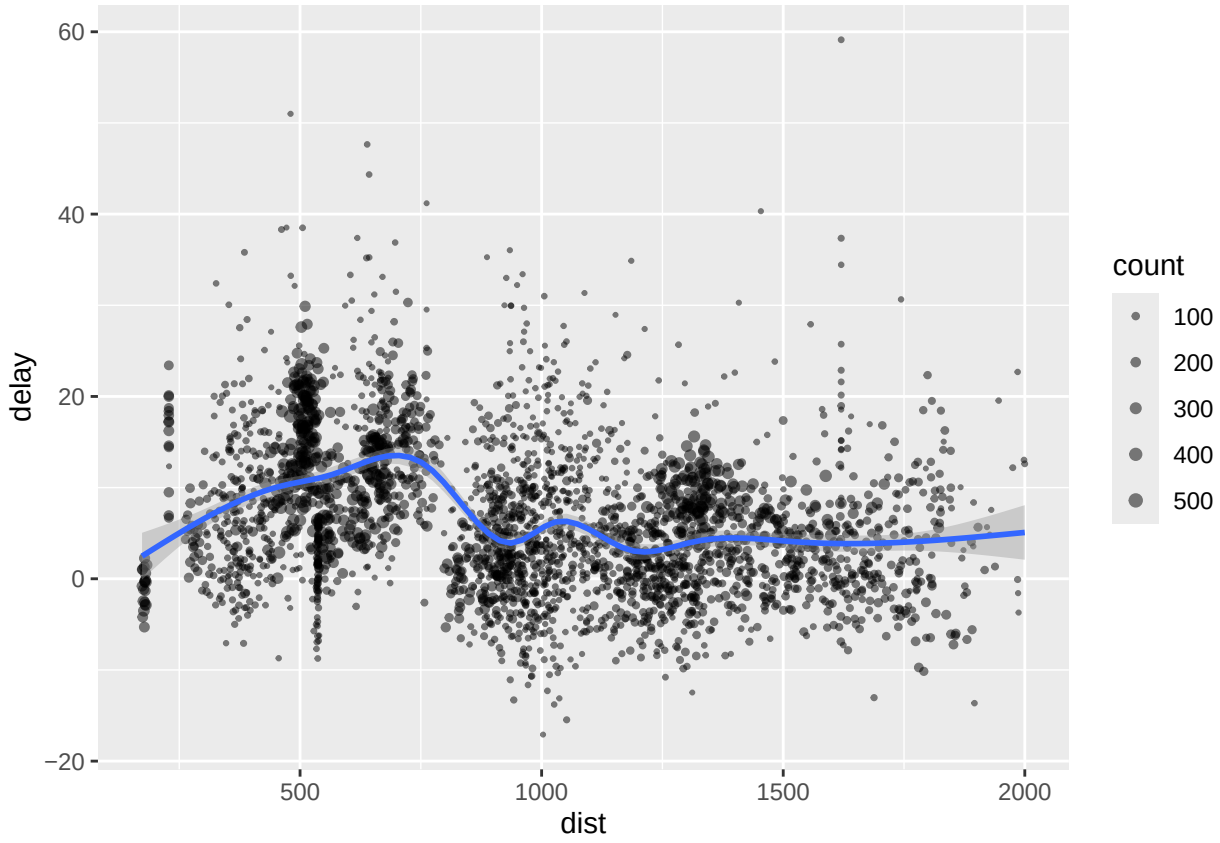
```
##  4  2013     1     1      715            713         2      911            850
##  5  2013     1     1      752            750         2     1025           1029
##  6  2013     1     1      917            915         2     1206           1211
##  7  2013     1     1      932            930         2     1219           1225
##  8  2013     1     1     1028           1026         2     1350           1339
##  9  2013     1     1     1042           1040         2     1325           1326
## 10  2013     1     1     1231           1229         2     1523           1529
## # i more rows
## # i 11 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
## #   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
## #   hour <dbl>, minute <dbl>, time_hour <dttm>
```

```r
delay <- flights_tbl %>%
  group_by(tailnum) %>%
  summarise(count = n(), dist = mean(distance, na.rm = TRUE), delay = mean(arr_delay, na.rm = TRUE)) %>%
  filter(count > 20, dist < 2000, !is.na(delay)) %>%
  collect()

# Saving as json
#write_json(delay, "file.json")

ggplot(delay, aes(dist, delay)) +
  geom_point(aes(size = count), alpha = 1/2) +
  geom_smooth() +
  scale_size_area(max_size = 2)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Task 4

Count the occurrence of all the different values for the order variable that are present in the dataset. Use kable to display any tables.

```
# your code here
```

Task 5

Select the columns name, order and brainwt. Filter to only those rows where the order variable is Primates. Finally, arrange the result by brainwt. Try to use the pipe operator %>% to cut down on how many lines of code you write.

```
# your code here
```

Task 6

In this task, we will create a table of mean brain weight to body weight ratio, grouped by the variable order. Create a new column called brain_body_wt_ratio that is equal to brainwt/bodywt. Group by the variable order, and then calculate the mean for each group.

```
# your code here
```

Task 7

The total number of missing values in each column can be viewed using the following code

```
sleep %>%
  summarise_all(~sum(as.integer(is.na(.)))) %>%
  kable()
```

extra	group	ID
0	0	0

Impute missing values for brainwt. Do this by creating a new column called brainwt_imputed, where NA values are replaced with the mean value for brainwt. Verify the result by displaying the head of a table with the columns name, brainwt and brainwt_imputed. You may find the ifelse or case_when functions useful (Google them).

```
# your code here
```

Task 8

Use collect() and ggplot to make a horizontal bar chart with order on the vertical axis and average sleep_total on the horizontal axis. You may find geom_col from ggplot useful.

```
# your code here
```