Homework 1

Your Name

2022-09-01

```
library(data.table)
library(dplyr)
library(dplyr)
library(tidyr)
library(plotly)
library(lubridate)
```

In this homework you should use plotly unless said otherwise.

To create pdf version of your homework, knit it first to html and then print it to pdf. Interactive plotly plots can be difficult sometimes to convert to static images suitable for insertion to LaTex documents (that is knitting to PDF).

Look for questions in R-chunks as comments and plain text (they are prefixed as Q.).

Part 1. Iris Dataset. (20 points)

> "The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis" https://en.wikipedia.org/wiki/Iris_flower_data_set

```
# Q1.1. Read the iris.csv file  (2 points)
# hint: use fread from data.table, it is significantly faster than default methods
#       be sure to have strings as factors (see stringsAsFactors argument)
```

```
# Q1.2. Show some values from data frame (2 points)
```

```
# Q1.3. Build histogram plot for Sepal.Length variable for each species using plot_ly
# (use color argument for grouping) (2 points)
# should be one plot
```

```
# Q1.4. Repeat previous plot with ggplot2 and convert it to plotly with ggplotly (2 points)
```

```
# Q1.5. Create facet 2 by 2 plot with histograms similar to previous but for each metric
# (2 points)
# hint:
#   following conversion to long format can be useful:
#   iris %>% gather(key = "metric", value = "value",-Species)
#
```

Q1.6. Which metrics has best species separations? (2 points)

```
# Q1.7. Repeat above plot but using box plot (2 points)
```

```
# Q1.8. Choose two metrics which separates species the most and use it to make scatter plot
# color points by species (2 points)
```

```
# Q1.9. Choose three metrics which separates species the most and use it to make 3d plot
# color points by species (2 points)
```

Q1.10. Comment on species separation (2 points):

Part 2. Covid-19 Dataset. (18 points)

Download us-states.csv (there is also a copy in homework assignment) from https://github.com/nytimes/covid-19-data/. README.md for details on file content.

```
# Q2.1. Read us-states.csv (2 points)
```

```
# Q2.2. Show some values from dataframe
```

```
# Q2.3. Create new dataframe with new cases per month for each state (2 points)
# hint:
#    is cases column cumulative or not cumulative?
```

```
# Q2.4.Using previous dataframe plot new monthly cases in states, group by states
# The resulting plot is busy, use interactive plotly capabilities to limit number
# of displayed states
# (2 points)
```

```
# Q2.5.Plot new monthly cases only in NY state
# (2 points)
```

```
# Q2.6. Found the year-month with highest cases in NY state
# (2 points)
```

```
# Q2.7. Plot new cases in determined above year-month
# using USA state map, color each state by number of cases  (2 points)
# hint:
#   there two build in constants in R: state.abb and state.name
#   to convert full name to abbreviation
```

```
# Q2.8. Add animation capability (2 points)
# hint:
#     for variable frame you need either integer or character/factorial so
#     convert date to character or factorial
```

Q2.9. Compare animated plot from Q2.8 to plots from Q2.4/Q2.5 (When you would prefer one or another?) (2 points)