# COMP1204: Data Management
# Coursework One: Hurricane Monitoring

Lukas Kakogiannos
32158998

April 3, 2023

## 1 Introduction

The aim of this coursework is to describe the data cleaning process, in this case, to develop and explain a bash script that efficiently cleans data given in .kml files, making it useful for other scripts and readable to the human eye. Those .kml files contain useful information describing the trajectory of different storms throughout different times of the year, which are surrounded by useless tags and other bits of information.

## 2 Create CSV Script

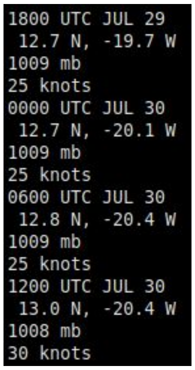
Listing 1: final script

```
#!/bin/bash

#message for the user
echo "Converting "$1" -> "$2""

#text runs through a pipeline of commands and is modified until it is
#presentable (see page 2)
grep -w ’UTC\|N\|mb\|knots’ "$1" | sed -e ’/name/d’ -e ’/dtg/d’ -e ’s/;.*//g’ -e ’s/<tr><td>//g’ -e ’s/<.td><.tr>//g’ -e ’s/<B>//g’ -e ’s/<.B>//g’ | awk ’(NR%4==1){time=$1; month=$3; day=$4} (NR%4==2){latitude=$1; longitude=$3} (NR%4==3){pressure=$1} (NR%4==0){knots=$1} (NR%4==0){print time" UTC "month" "day","latitude" N,"longitude" W,"pressure" mb,"knots" knots"}’ | sed ’1 i Timestamp,Latitude,Longitude,MinSeaLevelPressure,MaxIntensity’ > "$2"

#message for the user
echo "Done!"
```

1. 'grep -w' is used to print only lines that match one of the following four patterns (time, location, pressure and intensity), which leaves mostly useful information.

2. 'sed' is used to delete useless information (e.g. mph, kph and Hg) and useless tags in the .kml files.

3. the content of the pipeline at this point has the following format:



```
1800 UTC JUL 29
 12.7 N, -19.7 W
1009 mb
25 knots
0000 UTC JUL 30
 12.7 N, -20.1 W
1009 mb
25 knots
0600 UTC JUL 30
 12.8 N, -20.4 W
1009 mb
25 knots
1200 UTC JUL 30
 13.0 N, -20.4 W
1008 mb
30 knots
```

Figure 1: output after the use of sed

   Upon examining the pattern, it can be seen that every four lines match a different point in the hurricane's trajectory. 'awk' is used in the following way: the script goes through the input and assigns the important information into different variables, and every four lines, those variables are printed in the required format.

4. The final 'sed' inserts a single header line which details the five columns of data.

5. Finally, the output is stored in a file taken as a parameter from the shell.

## 3 Storm Plots

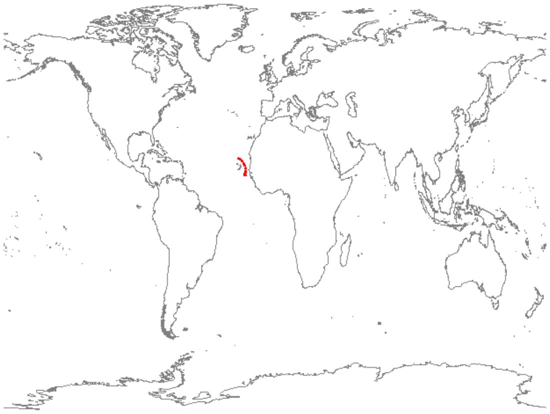


Figure 2: al102020.kml storm plot

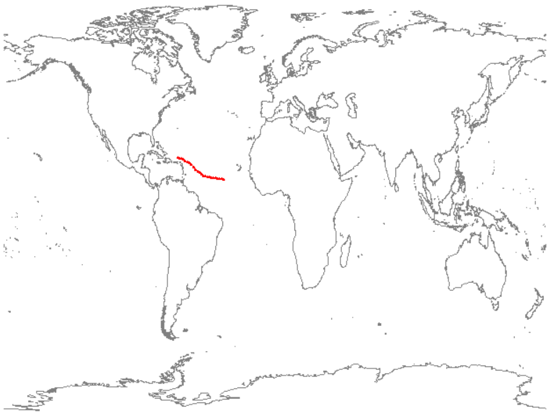


Figure 3: al112020.kml storm plot

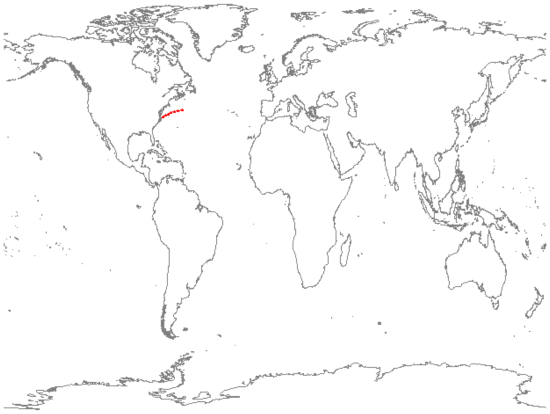


Figure 4: al122020.kml storm plot