DATA301 EDA Report

The structure of the report is based on the "Report Marking Scheme" on page 4 of the project_briefing_2021_edit.pdf on Blackboard

Part 1. Background and Data [1-3 pages, 6 marks]

1.1 State which dataset(s) your group worked on, and their source

We work on the analysis of handwritten spirals, whose data were collected as part of a study with patients of Parkinson disease.

1.2 Explain briefly why the dataset is of interest, or what questions it could be used to answer; assume that the reader has never heard of your dataset

To be filled.

1.3 State the types of data in the dataset(s) and the structure of the dataset(s). Are the data numerical, categorical, or both? Time series? Coordinates? Diagnostic categories? This does NOT need to be an exhaustive list of every variable, just a few comments on the overall types.

Let's choose one example from each of the two groups (control and ill patients) to have a look at the datasets.

```
control_data <- read_table("./Data/Controles30jun14/C01/session00002/u00003s00002_hw000012.svc",
                           col_names = FALSE, skip = 1)
names(control_data) <- c("x", "y", "Time", "On/Off", "Azimuth", "Altitude", "Pressure")

patients_data <- read_table("./Data/Protocolo temblor/T001/session00001/u00005s00001_hw00001.svc",
                           col_names = FALSE, skip = 1)
names(patients_data) <- c("x", "y", "Time", "On/Off", "Azimuth", "Altitude", "Pressure")
```

Above section reads the control data and patients data. We already know that each dataset has seven variables with the time-ordered x and y coordinates, the time stamp, the on/off state of the pen, the azimuth, the altitude, and the pressure.

```
cat(class(control_data$x[1]),class(control_data$y[1]),class(control_data$Time[1]),
    class(control_data$`On/Off`[1]),class(control_data$Azimuth[1]),
    class(control_data$Altitude[1]),class(control_data$Pressure[1]))
```

```
## numeric numeric numeric numeric numeric numeric numeric
```

```
cat(class(patients_data$x[1]),class(patients_data$y[1]),class(patients_data$Time[1]),
    class(patients_data$`On/Off`[1]),class(patients_data$Azimuth[1]),
    class(patients_data$Altitude[1]),class(patients_data$Pressure[1]))
```

```
## numeric numeric numeric numeric numeric numeric numeric
```

```
summary(control_data)
```

```
##        x               y             Time             On/Off
##  Min.   :4053   Min.   :7214   Min.   :1863797   Min.   :0.0000
##  1st Qu.:4703   1st Qu.:7833   1st Qu.:1867733   1st Qu.:1.0000
##  Median :5226   Median :8356   Median :1871575   Median :1.0000
##  Mean   :5180   Mean   :8367   Mean   :1871575   Mean   :0.9907
##  3rd Qu.:5632   3rd Qu.:8843   3rd Qu.:1875418   3rd Qu.:1.0000
##  Max.   :6360   Max.   :9593   Max.   :1879260   Max.   :1.0000
##     Azimuth        Altitude        Pressure
##  Min.   :1530   Min.   :770.0   Min.   :   0
##  1st Qu.:1870   1st Qu.:810.0   1st Qu.: 963
##  Median :1960   Median :820.0   Median :1029
##  Mean   :1942   Mean   :817.5   Mean   :1004
##  3rd Qu.:2030   3rd Qu.:830.0   3rd Qu.:1089
##  Max.   :2210   Max.   :850.0   Max.   :1445
```

```
summary(patients_data)
```

```
##        x               y              Time             On/Off
##  Min.   : 557   Min.   : 9752   Min.   :2098403   Min.   :0.000
##  1st Qu.:1114   1st Qu.:10476   1st Qu.:2103405   1st Qu.:1.000
##  Median :1675   Median :11018   Median :2108490   Median :1.000
##  Mean   :1663   Mean   :11022   Mean   :2108588   Mean   :0.776
##  3rd Qu.:2220   3rd Qu.:11616   3rd Qu.:2113785   3rd Qu.:1.000
##  Max.   :2723   Max.   :12317   Max.   :2118975   Max.   :1.000
##     Azimuth        Altitude        Pressure
##  Min.   :1870   Min.   :650.0   Min.   :   0
##  1st Qu.:2030   1st Qu.:680.0   1st Qu.:1237
##  Median :2110   Median :710.0   Median :2048
##  Mean   :2099   Mean   :707.8   Mean   :1520
##  3rd Qu.:2180   3rd Qu.:740.0   3rd Qu.:2048
##  Max.   :2270   Max.   :760.0   Max.   :2048
```

From the above code outputs, we see that all seven variables are numerical. The only exception is the 'On/off' variable, which is a data type with a value of 0 or 1, meaning that it is also categorical. For 'x' and 'y', they represent the x-axis and y-axis coordinate values along the time series variable 'Time'.

1.4 State how complete the dataset(s) are (i.e. how many missing, any structure to the missing data, whether there are errors in the data)

```
cat(sum(is.na(control_data$x)),sum(is.na(control_data$y)),sum(is.na(control_data$Time)),
    sum(is.na(control_data$`On/Off`)),sum(is.na(control_data$Azimuth)),
    sum(is.na(control_data$Altitude)),sum(is.na(control_data$Pressure)))
```

```
## 0 0 0 0 0 0 0
```

```
cat(sum(is.na(patients_data$x)),sum(is.na(patients_data$y)),sum(is.na(patients_data$Time)),
    sum(is.na(patients_data$`On/Off`)),sum(is.na(patients_data$Azimuth)),
    sum(is.na(patients_data$Altitude)),sum(is.na(patients_data$Pressure)))
```

```
## 0 0 0 0 0 0 0
```

There are no missing values for both healthy dataset and patients dataset from the above two code outputs.

1.5 If you used more than one dataset, state what steps you had to take to integrate the datasets

To be filled.

Part 2. Ethics, Privacy and Security [1-2 pages, 6 marks]

2.1 Brief discussion of any ethical considerations that apply to your project

2.2 Brief discussion of any privacy concerns that might arise connected to your project

2.3 Brief discussion of what steps you could take to keep your project data and results secure (you do NOT need to carry this out, you just need to talk about it in the report)

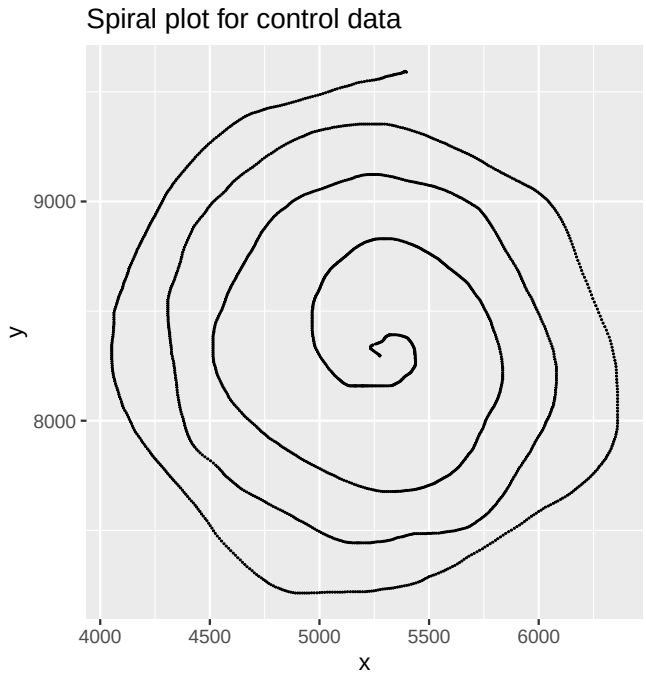
Mason is working on this part.

Part 3. Exploratory Data Analysis [3-8 pages, 15 marks]

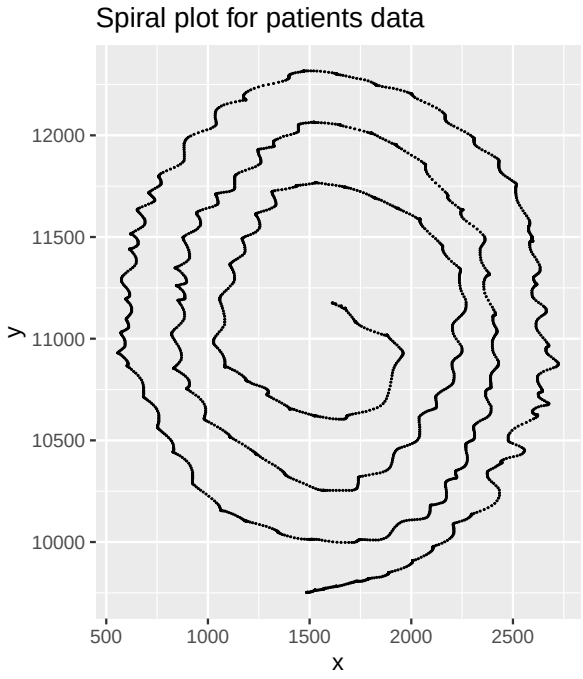
For this section, do NOT try to summarize everything you can find in the dataset(s). Select a subset, highlighting features that you thought were interesting in the data. The plots do not have to be complicated; simple bar charts and scatter plots are fine.

- Several summary tables and/or plots, each describing one, two or three variables in the data that you thought were interesting
- Explain the definitions of the variables in each table/plot
- Comment on the main features of each plot
- Include suitable labels and keys for each plot
- adjust the point sizes and/or line thicknesses to improve readability
- Lay out all tables so that they are clearly readable and clearly labelled, and do not use excessive significant figures

```
ggplot(control_data, aes(x=x, y=y)) +
geom_point(size=.01) +
coord_fixed()+
ggtitle("Spiral plot for control data")
```

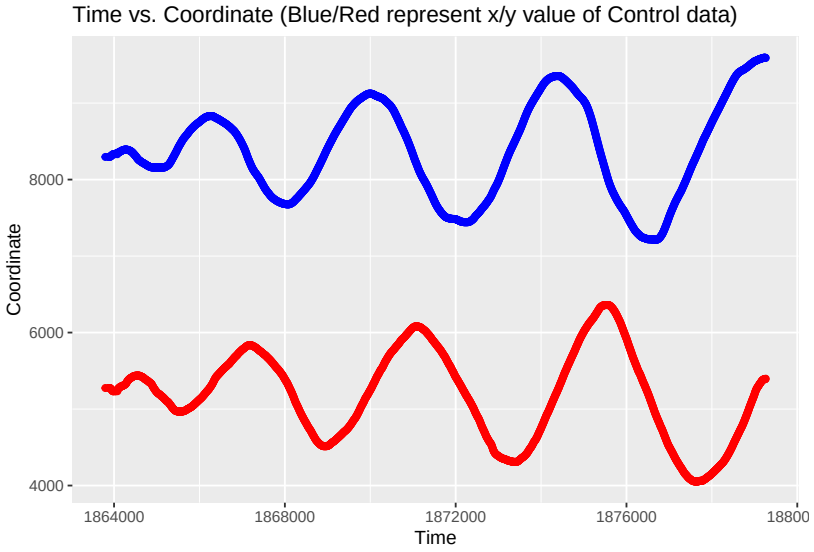



```
ggplot(patients_data, aes(x=x, y=y)) +
geom_point(size=.01) +
coord_fixed()+
ggtitle("Spiral plot for patients data")
```

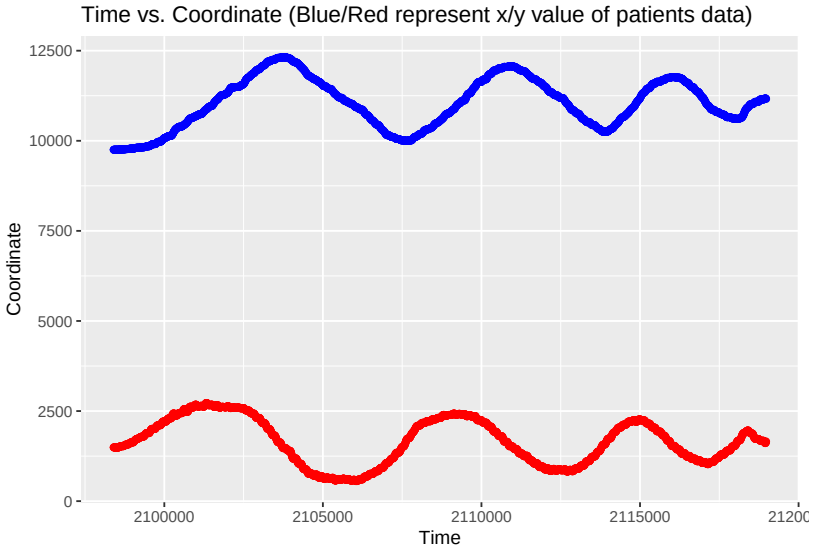


Above are two spiral plots of x and y coordinates for control data and patients data respectively. We can intuitively see the movement trajectory of the corresponding handwritten spiral from the figure. The spiral drawn by normal control was more smooth and regular than that of Parkinson's Disease patient. This indicated that stiffness and tremor features were reflected in the hand movement of the patient.

```
ggplot(control_data, aes(x=Time, y=x)) +
  geom_point(col="red") +
  geom_point(aes(x=Time, y=y), col="blue")+ xlab("Time") + ylab("Coordinate") +
  ggtitle("Time vs. Coordinate (Blue/Red represent x/y value of Control data)")
```

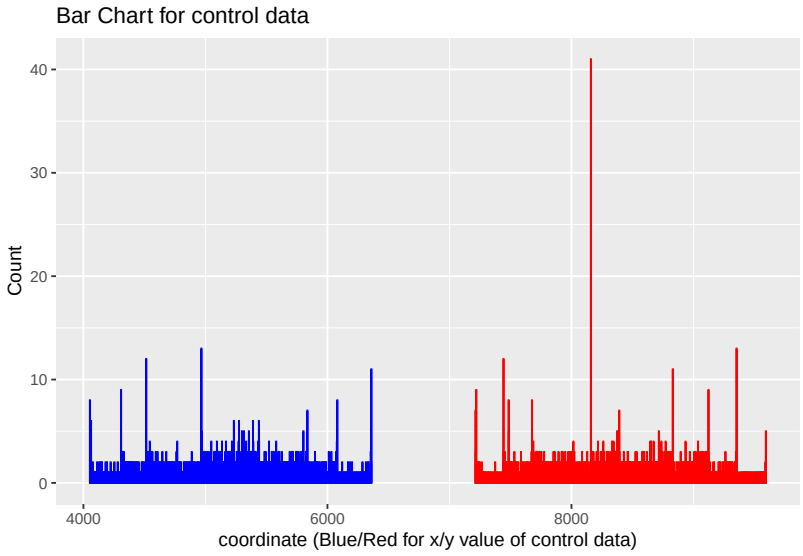



```
ggplot(patients_data, aes(x=Time, y=x)) +
  geom_point(col="red") +
  geom_point(aes(x=Time, y=y), col="blue")+ xlab("Time") + ylab("Coordinate") +
  ggtitle("Time vs. Coordinate (Blue/Red represent x/y value of patients data)")
```

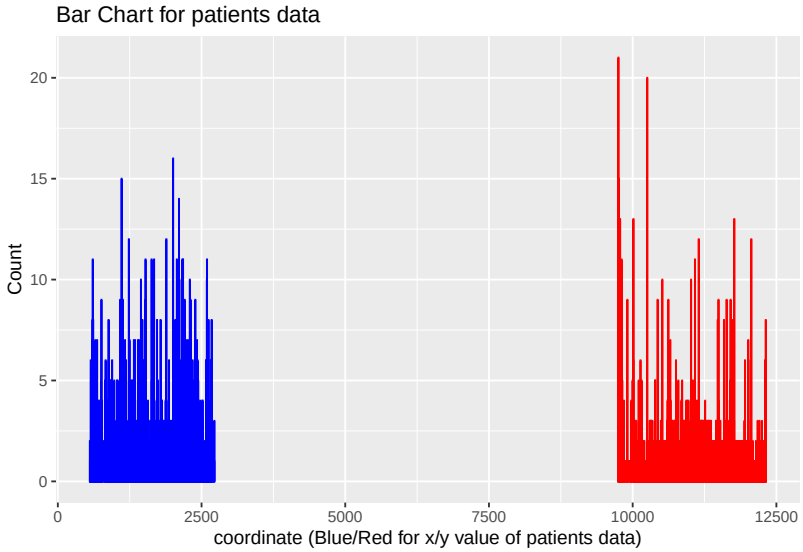


Above are two scatter plots of x/y coordinate vs Time for control data and patients data respectively. Blue points represent the x coordinate value change over time while red points represent the y coordinate value change over time. We see that the patterns of x and y values are very similar. But for different datasets, the patterns tend to be different. The line of control data is smoother with higher amplitude, while the line of patients data has more local changes and less amplitude.

```
ggplot(control_data) +
  geom_bar(aes(x=y),col="red") + geom_bar(aes(x=x),col="blue")+
  xlab("coordinate (Blue/Red for x/y value of control data)") + ylab("Count") +
  ggtitle("Bar Chart for control data")
```




```
ggplot(patients_data) +
  geom_bar(aes(x=y),col="red") + geom_bar(aes(x=x),col="blue")+
  xlab("coordinate (Blue/Red for x/y value of patients data)") + ylab("Count") +
  ggtitle("Bar Chart for patients data")
```



Above are two bar charts of x coordinate and y coordinate count for control data and patients data respectively. Blue points represent the x coordinate value change over time while red points represent the y coordinate value change over time. We see that the patterns of x and y values are very similar. But for different datasets, the patterns tend to be different. The patient data are concentrated with a higher average count, while control data have a lower average count, which reflects that patients with Parkinson's disease tend to draw a non-smooth spiral.