Preliminary Analysis

Kanika Chopra

2022-11-03

Predicting Parkinson's Based on Auditory Data

a) Import Data

First, we want to read in our data.

```
library(RCurl)
data <- read.csv(text = getURL(
  "https://raw.githubusercontent.com/kanikadchopra/Parkinsons-Prediction/main/parkinson_data.csv"))
```

```
attach(data)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::complete() masks RCurl::complete()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
```

```
dim(data)
```

```
## [1] 195  24
```

```
glimpse(data)
```

```
## Rows: 195
## Columns: 24
## $ name             <chr> "phon_R01_S01_1", "phon_R01_S01_2", "phon_R01_S01_3",~
## $ MDVP.Fo.Hz.      <dbl> 119.992, 122.400, 116.682, 116.676, 116.014, 120.552,~
## $ MDVP.Fhi.Hz.     <dbl> 157.302, 148.650, 131.111, 137.871, 141.781, 131.162,~
## $ MDVP.Flo.Hz.     <dbl> 74.997, 113.819, 111.555, 111.366, 110.655, 113.787, ~
## $ MDVP.Jitter...   <dbl> 0.00784, 0.00968, 0.01050, 0.00997, 0.01284, 0.00968,~
## $ MDVP.Jitter.Abs. <dbl> 0.00007, 0.00008, 0.00009, 0.00009, 0.00011, 0.00008,~
## $ MDVP.RAP         <dbl> 0.00370, 0.00465, 0.00544, 0.00502, 0.00655, 0.00463,~
## $ MDVP.PPQ         <dbl> 0.00554, 0.00696, 0.00781, 0.00698, 0.00908, 0.00750,~
## $ Jitter.DDP       <dbl> 0.01109, 0.01394, 0.01633, 0.01505, 0.01966, 0.01388,~
## $ MDVP.Shimmer     <dbl> 0.04374, 0.06134, 0.05233, 0.05492, 0.06425, 0.04701,~
## $ MDVP.Shimmer.dB. <dbl> 0.426, 0.626, 0.482, 0.517, 0.584, 0.456, 0.140, 0.13~
## $ Shimmer.APQ3     <dbl> 0.02182, 0.03134, 0.02757, 0.02924, 0.03490, 0.02328,~
## $ Shimmer.APQ5     <dbl> 0.03130, 0.04518, 0.03858, 0.04005, 0.04825, 0.03526,~
```

```
## $ MDVP.APQ         <dbl> 0.02971, 0.04368, 0.03590, 0.03772, 0.04465, 0.03243,~
## $ Shimmer.DDA      <dbl> 0.06545, 0.09403, 0.08270, 0.08771, 0.10470, 0.06985,~
## $ NHR              <dbl> 0.02211, 0.01929, 0.01309, 0.01353, 0.01767, 0.01222,~
## $ HNR              <dbl> 21.033, 19.085, 20.651, 20.644, 19.649, 21.378, 24.88~
## $ status           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ RPDE             <dbl> 0.414783, 0.458359, 0.429895, 0.434969, 0.417356, 0.4~
## $ DFA              <dbl> 0.815285, 0.819521, 0.825288, 0.819235, 0.823484, 0.8~
## $ spread1          <dbl> -4.813031, -4.075192, -4.443179, -4.117501, -3.747787~
## $ spread2          <dbl> 0.266482, 0.335590, 0.311173, 0.334147, 0.234513, 0.2~
## $ D2               <dbl> 2.301442, 2.486855, 2.342259, 2.405554, 2.332180, 2.1~
## $ PPE              <dbl> 0.284654, 0.368674, 0.332634, 0.368975, 0.410335, 0.3~
```

b) Missing Values

Next, we want to conduct a few quick quality checks, such as checking if there are any missing values.

```
data %>% summarise_all(~ sum(is.na(.)))
```

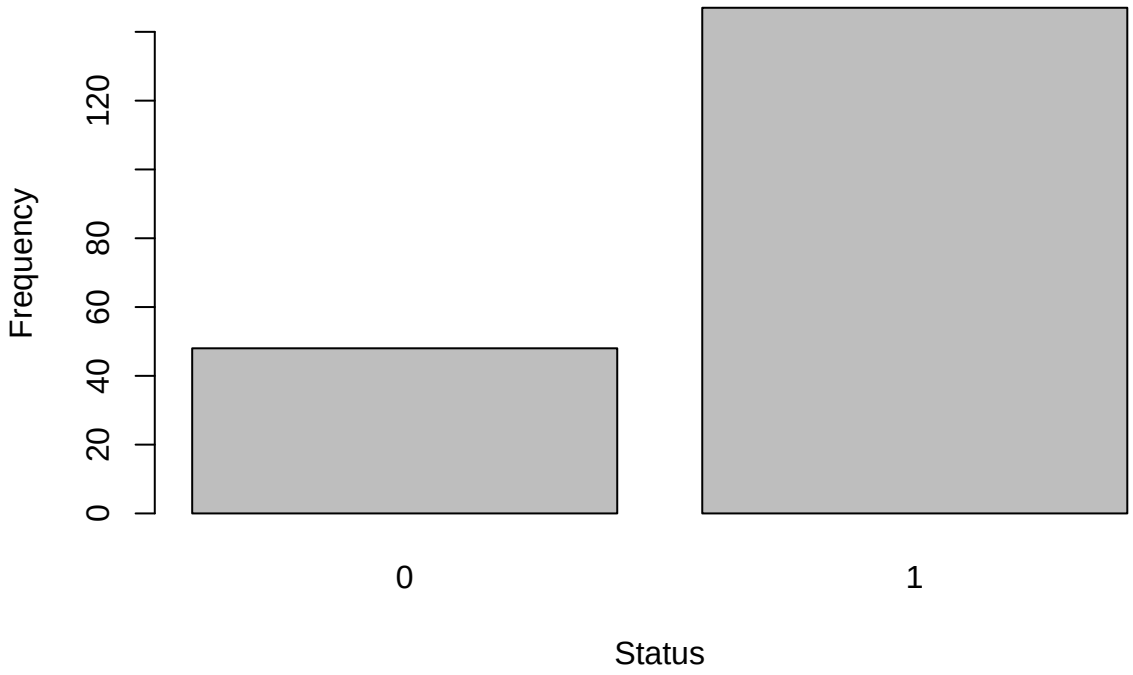
```
##   name MDVP.Fo.Hz. MDVP.Fhi.Hz. MDVP.Flo.Hz. MDVP.Jitter... MDVP.Jitter.Abs.
## 1    0           0            0            0              0                0
##   MDVP.RAP MDVP.PPQ Jitter.DDP MDVP.Shimmer MDVP.Shimmer.dB. Shimmer.APQ3
## 1        0        0          0            0                0            0
##   Shimmer.APQ5 MDVP.APQ Shimmer.DDA NHR HNR status RPDE DFA spread1 spread2 D2
## 1            0        0           0   0   0      0    0   0       0       0  0
##   PPE
## 1   0
```

We can see that we have zero missing values along all of our columns which is good news as we don't have to handle them in the analysis.

c) Preliminary Plots

First, let's look at our predictor variable.

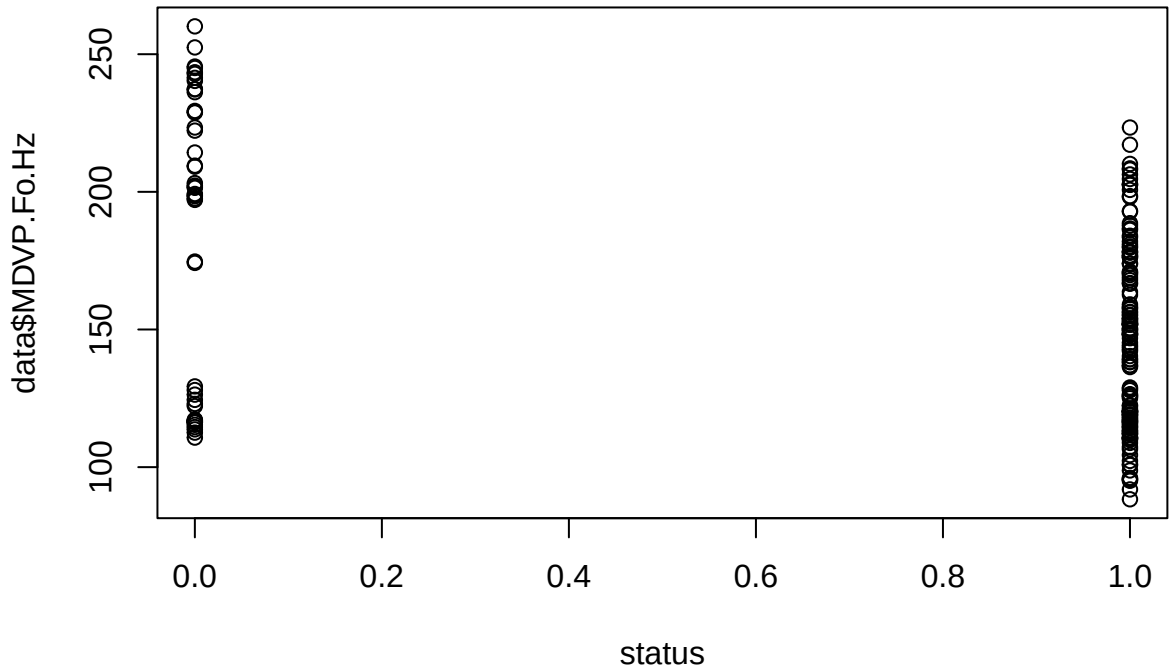
```
barplot(table(status), xlab='Status', ylab='Frequency')
```



Firstly, we can see that our data is skewed in that we have more data on patients with Parkinson's (`status=1`) than patients who do not have Parkinson's (`status=0`).

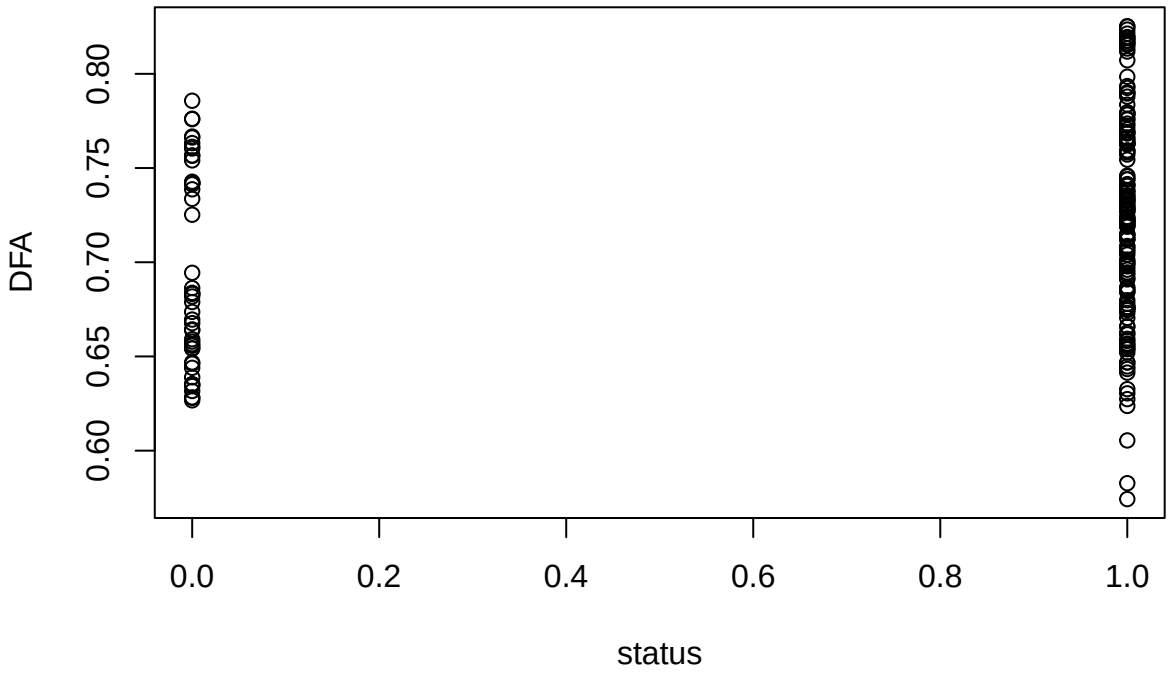
Next, we look at some of our variables for a preliminary analysis. We pick a few of our variables to investigate:

```
plot(status, data$MDVP.Fo.Hz)
```



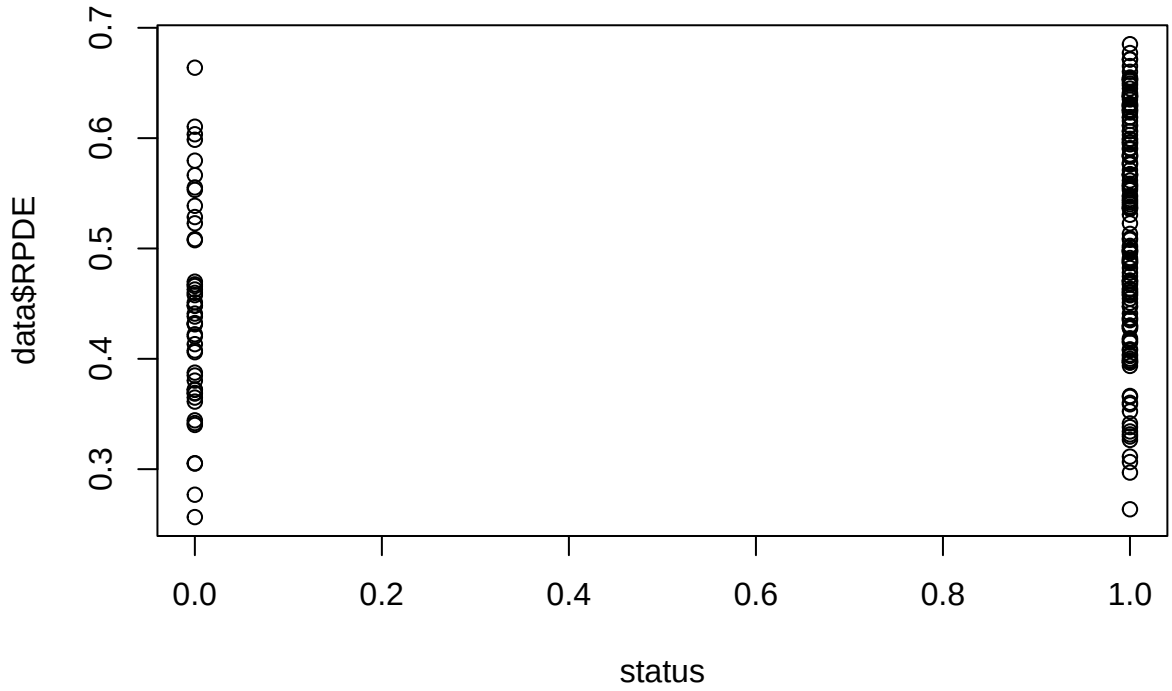
We have a higher distribution of MDVP values when we have status=0 so no Parkinson's.

```
plot(status, DFA)
```

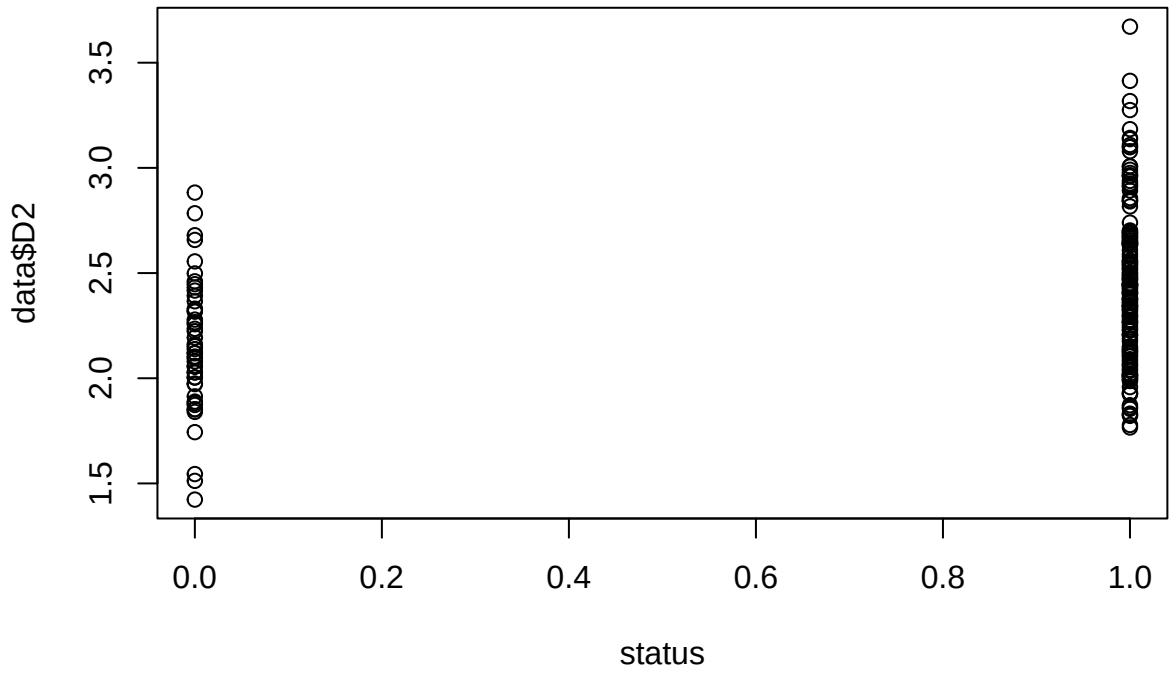


We see higher DFA values concentrated for when status=1 (Parkinson's)

```
plot(status, data$RPDE)
```

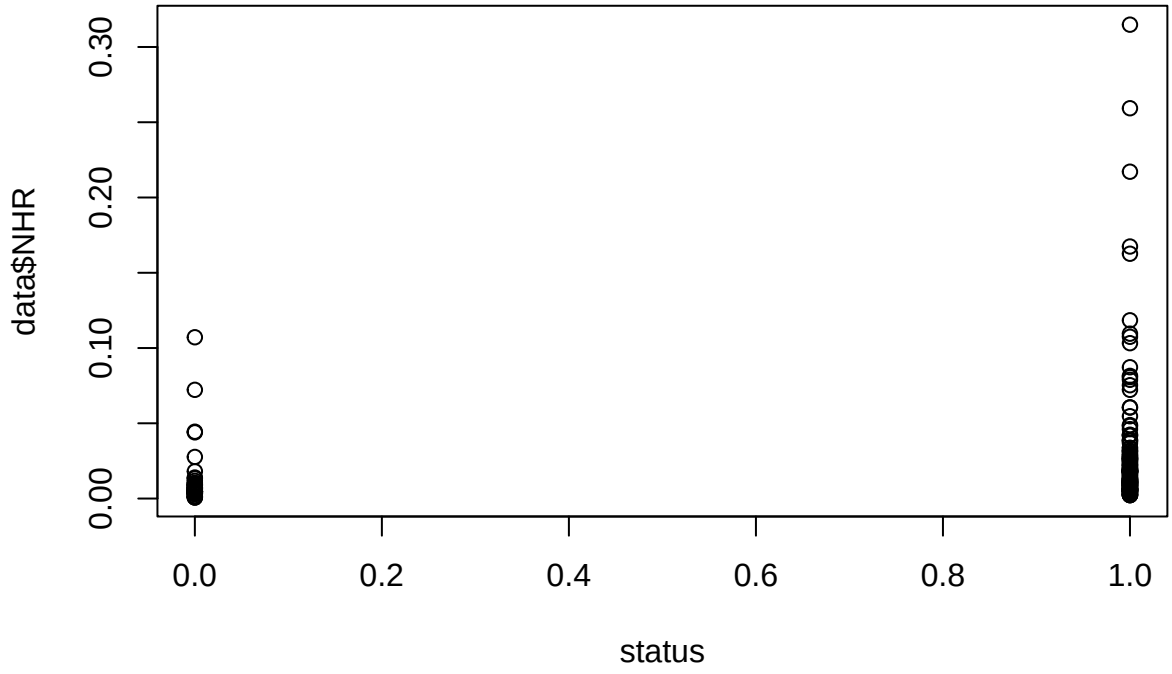



```
plot(status, data$D2)
```



For D2 and RPDE, we see a higher distribution of values for patients with Parkinson's.

```
plot(status, data$NHR)
```



For NHR, healthy patients seem to have a concentrated low value whereas patients with Parkinson's seem to have some values spread out with higher NHRs. There may be some anomalies here with the values higher than 0.2 for NHR.