

# Preliminary Analysis of the player logs dataset

*Nitish Rangarajan*

*November 22, 2017*

## Player Log Analysis

Load the library and the dataset

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.3.3
```

```
## -- Attaching packages -----  
----- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1      v purrr   0.2.4  
## v tibble  1.3.4      v dplyr   0.7.4  
## v tidyr   0.7.2      v stringr 1.2.0  
## v readr   1.1.1      v forcats 0.2.0
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
## Warning: package 'tibble' was built under R version 3.3.3
```

```
## Warning: package 'tidyr' was built under R version 3.3.3
```

```
## Warning: package 'readr' was built under R version 3.3.3
```

```
## Warning: package 'purrr' was built under R version 3.3.3
```

```
## Warning: package 'dplyr' was built under R version 3.3.3
```

```
## Warning: package 'stringr' was built under R version 3.3.3
```

```
## Warning: package 'forcats' was built under R version 3.3.3
```

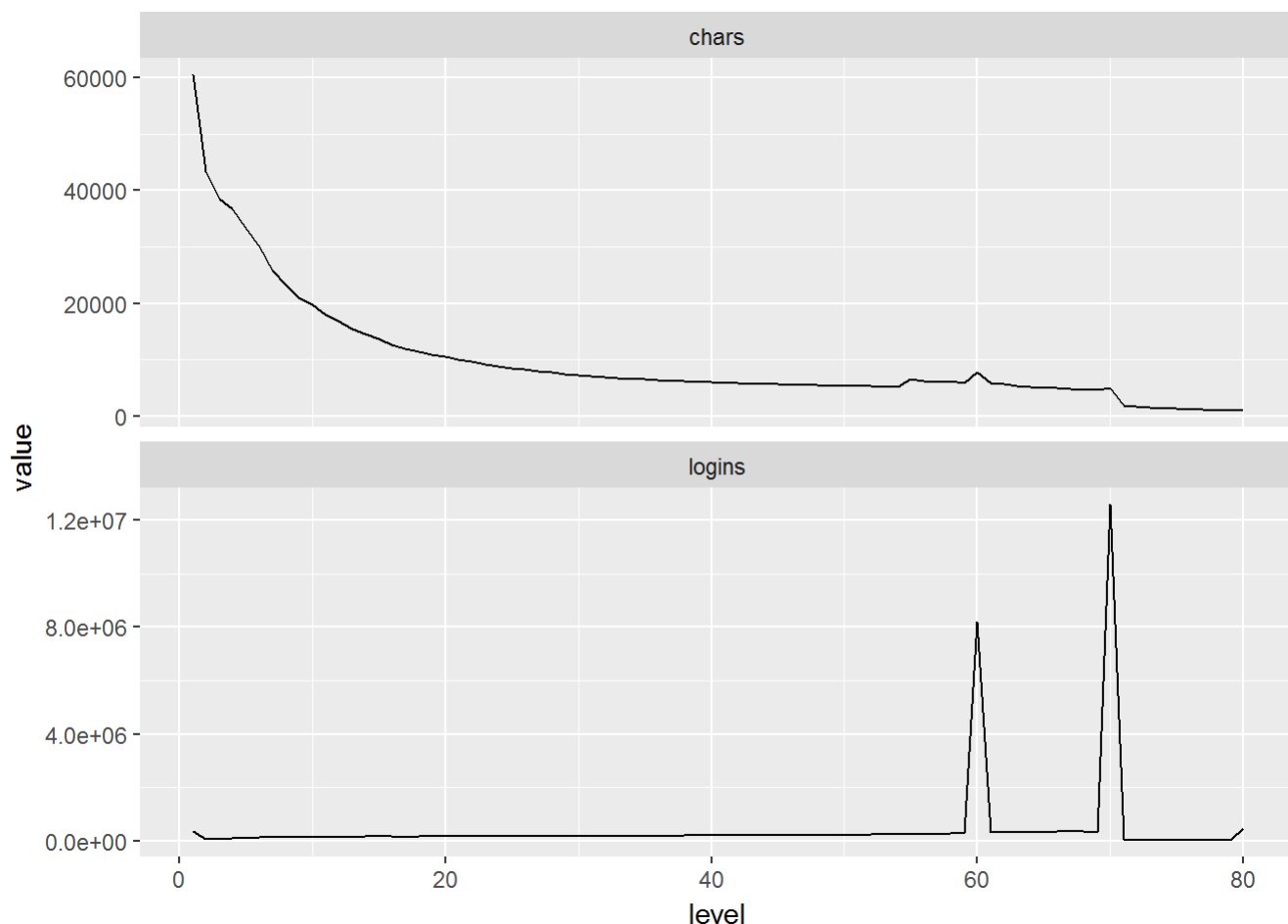
```
## -- Conflicts -----  
--- tidyverse_conflicts() ---  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()
```

```
setwd("C:/Users/nrangara/Downloads/WorldOfWarcraft/output")  
data<- read_csv("wowah_data.csv")
```

```
## Parsed with column specification:  
## cols(  
##   char = col_integer(),  
##   level = col_integer(),  
##   race = col_character(),  
##   charclass = col_character(),  
##   zone = col_character(),  
##   dummy1 = col_character(),  
##   dummy2 = col_integer(),  
##   guild = col_integer(),  
##   timestamp = col_character()  
## )
```

For the preliminary analysis, the number of characters at each level and number of logins at each level were checked.

```
data %>%  
  group_by(level) %>%  
  summarise(chars = n_distinct(char),  
            logins = n()) %>%  
  gather(var, value, -level) %>%  
  ggplot(aes(x = level, y = value)) +  
  geom_line() +  
  facet_wrap(~var, ncol = 1, scales = "free_y")
```



The results of the number of characters at each level wasn't unsurprising as there were lots of people in the first few levels, but the results show that there lots of people who frequently login when they are at the level ~70.

Convert the timestamp to date format. Then generate the player hours. The playing patterns of players over the years can be found by converting the number of observations of each player to hours by dividing the count for each user by 6 since we know that the observations were collected every 10 minutes.

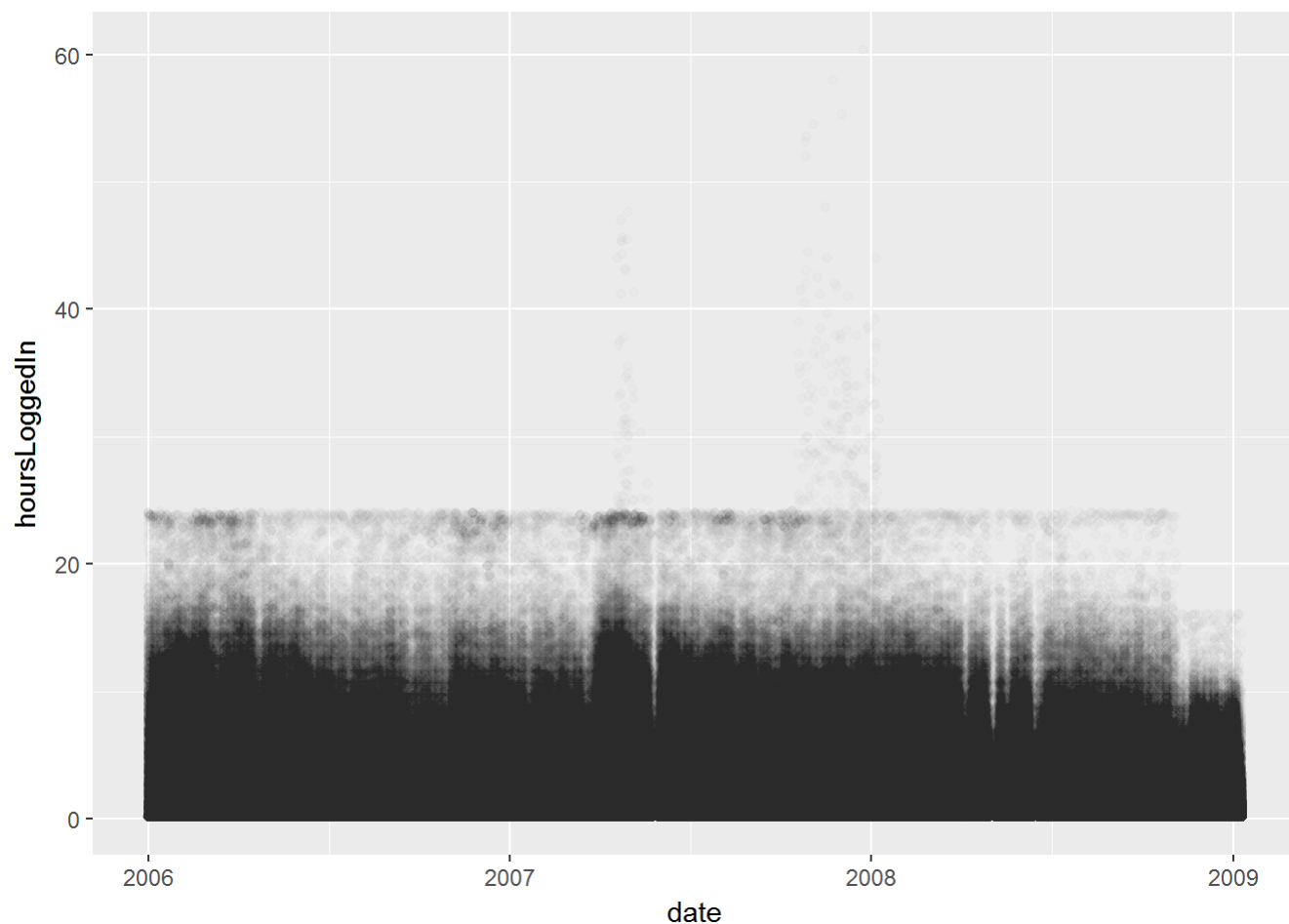
```
df <- tbl_df(data)

data$date <- as.Date(gsub(" .+", "", df$timestamp), format = "%m/%d/%y")

data %>%
  group_by(date, char) %>%
  summarize(hoursLoggedIn=n()/6) -> hours
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

```
hours %>%
  ggplot(aes(x=date, hoursLoggedIn))+geom_point(alpha=0.01)
```



On checking how many average players are logged in each day on a hourly basis to check if there are actually hardcore players, we can see a gradual dropoff as expected in the number of players. Login time of 47 hours might be because of duplicate observations.

```
hours$hoursLoggedIn %>%  
  table %>% data.frame %>%  
  plot()
```

