Task2 Analysis

```
library(ggplot2)
library(rjson)
library(rvest)
library(stringr)
library(dplyr)
library(plotly)
```

Import Dataset

```
df<-read.csv("positionTS.csv")
head(df)
```

```
####Import csv
#### I have converted the json file into csv for later data analysis.
#### Since Excel automatically converts positionTS column into scientific notation, I have to extract data from
json file which related to timestamp and save it as the 'positionTS' column.
```

Tidy Data

```
df<- df %>%     #### I mutate the"positionTS" column to get rid of unecessary information.
  mutate(positionTS=str_replace(positionTS,'"positionTS":',"")) %>%
  mutate(positionTS=str_replace(positionTS,',',""))

z<-as.numeric(df[,"positionTS"])
z<-as.POSIXct((z+0.1)/1000, origin = "1970-01-01")
df[,"positionTS"]<-z  ##### Converting original timestamp data into more readable format.

df<-df %>%
 janitor::clean_names()  #Clean column names
```

The records of this dataset were collected between“2017-01-24 10:49:15 EST” and “2017-01-24 11:08:16 EST”

```
range(df$position_ts)
```

```
## [1] "2017-01-24 10:49:15 EST" "2017-01-24 11:08:16 EST"
```

Does every tag being detected at same frequency? Are the number of detections for each tag equivalent to each other?

Does every tag being detected during the same period of time? When?

```
tbl_freq_detected=df %>%
  group_by(id,name) %>%
  summarize(n=n())
tbl_freq_detected
```

```
## # A tibble: 26 x 3
## # Groups:   id [?]
##    id           name        n
##    <fct>        <fct>   <int>
##  1 b4994c876dbb 024 res   172
##  2 b4994c876dcb 002 dev  5417
##  3 b4994c876de6 019 res  5439
##  4 b4994c877897 003 dev  5359
##  5 b4994c877aa1 005 dev  5456
##  6 b4994c877cb8 006 dev  5508
##  7 b4994c877d82 021 res   295
##  8 b4994c877eca 007 dev  3874
##  9 b4994c877ee8 012 pub  5468
## 10 b4994c877fa2 017 res  5474
## # ... with 16 more rows
```

According to table shown above,most tags had been detected more than 5000 times whereas a small amount of tags were detected less than 300 times.

Does it mean that some tags were detected in shorter period? Let's check out the timestamps between first detection and last detection for each tag.

```
tbl_interval_detection=df %>%
  group_by(id,name) %>%
  summarize(start=min(position_ts),end=max(position_ts))

tbl_interval_detection
```

```
## # A tibble: 26 x 4
## # Groups:   id [?]
##    id           name    start               end
##    <fct>        <fct>   <dttm>              <dttm>
##  1 b4994c876dbb 024 res 2017-01-24 10:50:11 2017-01-24 11:08:10
##  2 b4994c876dcb 002 dev 2017-01-24 10:49:19 2017-01-24 11:08:16
##  3 b4994c876de6 019 res 2017-01-24 10:49:19 2017-01-24 11:08:16
##  4 b4994c877897 003 dev 2017-01-24 10:49:19 2017-01-24 11:08:16
##  5 b4994c877aa1 005 dev 2017-01-24 10:49:19 2017-01-24 11:08:16
##  6 b4994c877cb8 006 dev 2017-01-24 10:49:19 2017-01-24 11:08:16
##  7 b4994c877d82 021 res 2017-01-24 10:49:15 2017-01-24 11:08:08
##  8 b4994c877eca 007 dev 2017-01-24 10:49:19 2017-01-24 11:08:16
##  9 b4994c877ee8 012 pub 2017-01-24 10:49:19 2017-01-24 11:08:16
## 10 b4994c877fa2 017 res 2017-01-24 10:49:19 2017-01-24 11:08:16
## # ... with 16 more rows
```

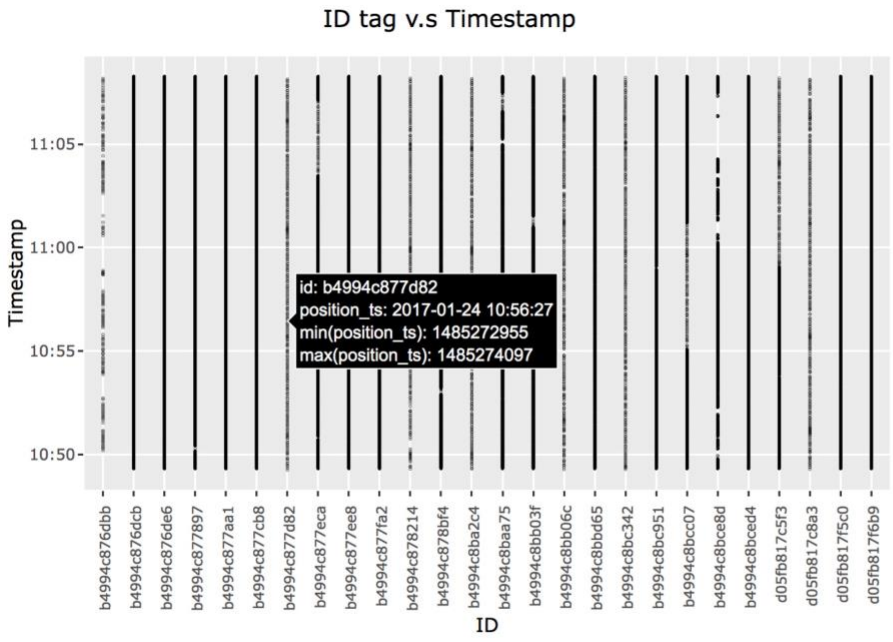
According to the table 'tbl_interval_detection', we found that the interval for each tag is about the same. Starting from around 10:49AM and ends around 11:08AM.

Maybe tags were detected at different frequencies? Let's plot ID vs. Timestamp for each ID tag to get a bigger picture.

Plotly View enable us to determine the exact timestamp of ID tag in our graph. Nevigate your cursor on plotly plot and relative details will be shown.

```
ggplot_timestamp=df %>%
  ggplot(aes(x = id, y = position_ts, ymin = min(position_ts),
             ymax = max(position_ts))) +
    geom_point(size=0.08) +
    theme(axis.text.x = element_text(angle = 90,size = 8)) +
  theme(legend.position = "none")+
  labs(x="ID",y="Timestamp",title="ID tag v.s Timestamp")

ggplotly(ggplot_timestamp)
```



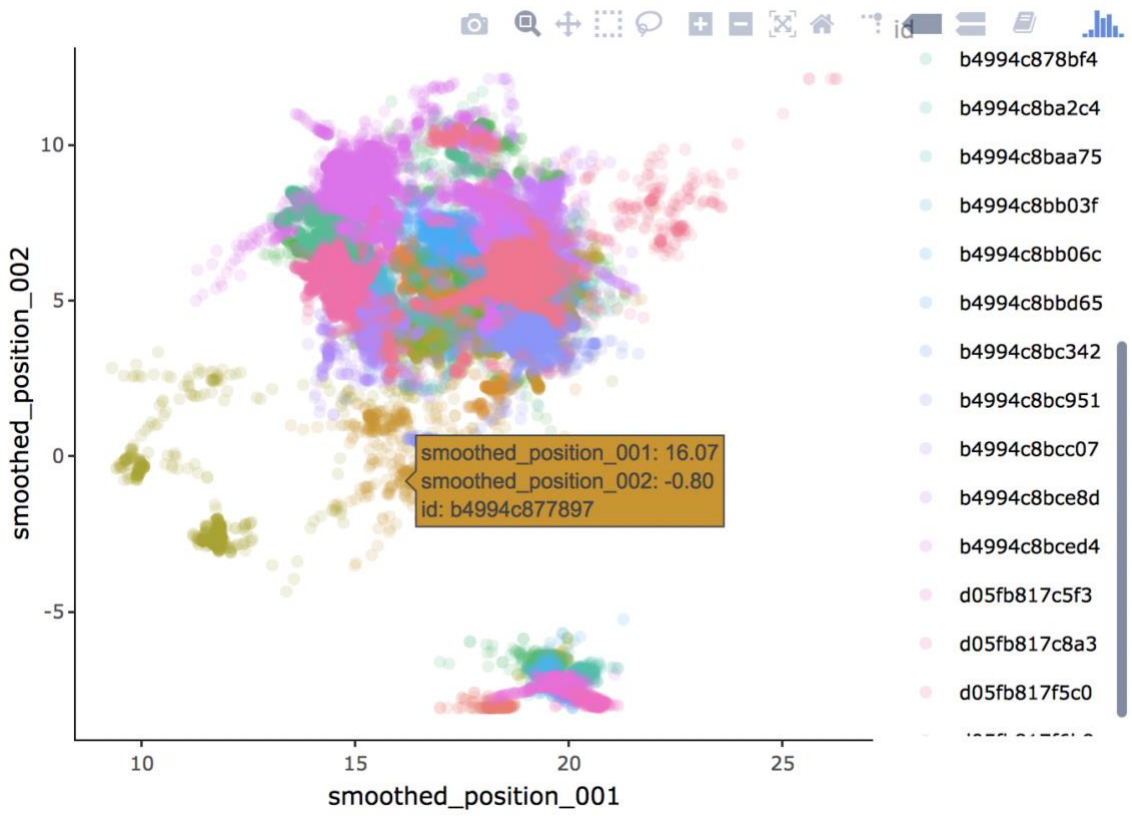
####According the the "ggplot_timestamp", we found that several tags were not detected as frequently as most of others. For example,"b4994c876dbb","b4994c877d82",etc.

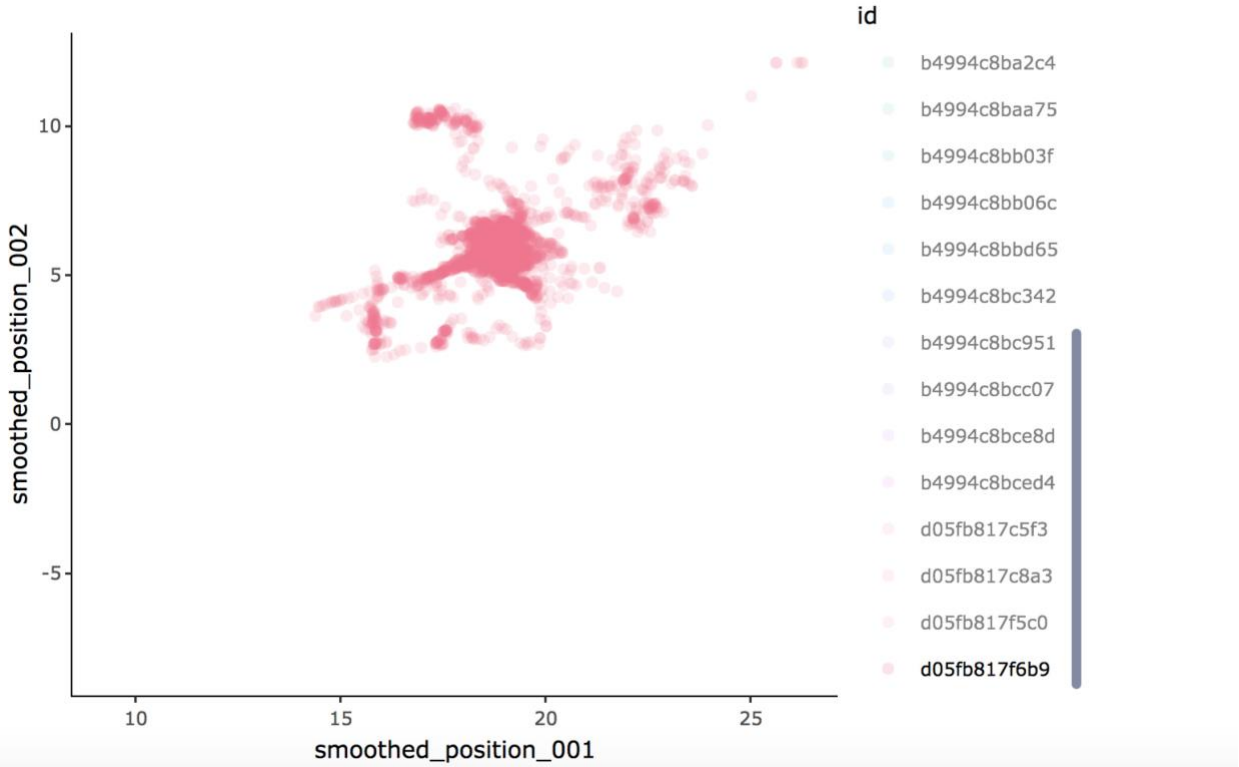
What about location plots for each tag? Since all values in smooth_position_003(z-coordinate) are zero, I plot 2D diagram instead of 3D.

By clicking id on the right, we can observe the x and y positions for specific ID independently.

```
scatter_position=df %>%
  ggplot(aes(x = smoothed_position_001, y = smoothed_position_002, color = id)) +
  geom_point(alpha = 0.15) +
  theme_classic()

ggplotly(scatter_position)
```


id
b4994c878bf4
b4994c8ba2c4
b4994c8baa75
b4994c8bb03f
b4994c8bb06c
b4994c8bbd65
b4994c8bc342
b4994c8bc951
b4994c8bcc07
b4994c8bce8d
b4994c8bced4
d05fb817c5f3
d05fb817c8a3
d05fb817f5c0
smoothed_position_001: 16.07
smoothed_position_002: -0.80
id: b4994c877897
smoothed_position_002
smoothed_position_001
10
5
0
-5
10
15
20
25


id
b4994c8ba2c4
b4994c8baa75
b4994c8bb03f
b4994c8bb06c
b4994c8bbd65
b4994c8bc342
b4994c8bc951
b4994c8bcc07
b4994c8bce8d
b4994c8bced4
d05fb817c5f3
d05fb817c8a3
d05fb817f5c0
d05fb817f6b9
smoothed_position_002
smoothed_position_001
10
5
0
-5
10
15
20
25

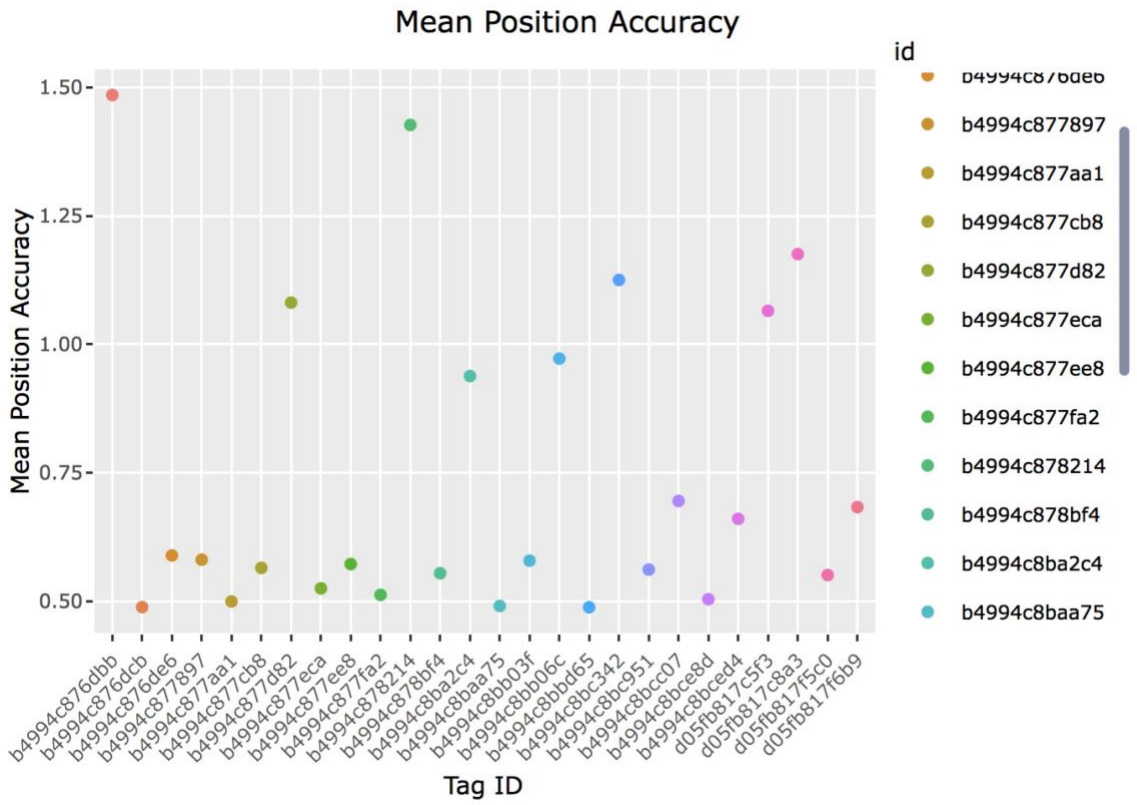
What about position accuracy for each tag?

Let's plot the mean position accuracy for each tag.

```
tbl_position_accuracy=df %>%
  group_by(id,name) %>%
  summarize(avg_accuracy= mean(position_accuracy))

position_accuracy=tbl_position_accuracy %>%
  ggplot(aes(x=id,y=avg_accuracy,color=id))+
  geom_point()+theme(axis.text.x = element_text(angle = 45))+labs(x="Tag ID",y="Mean Position Accuracy"
,title="Mean Position Accuracy")

ggplotly(position_accuracy)
```

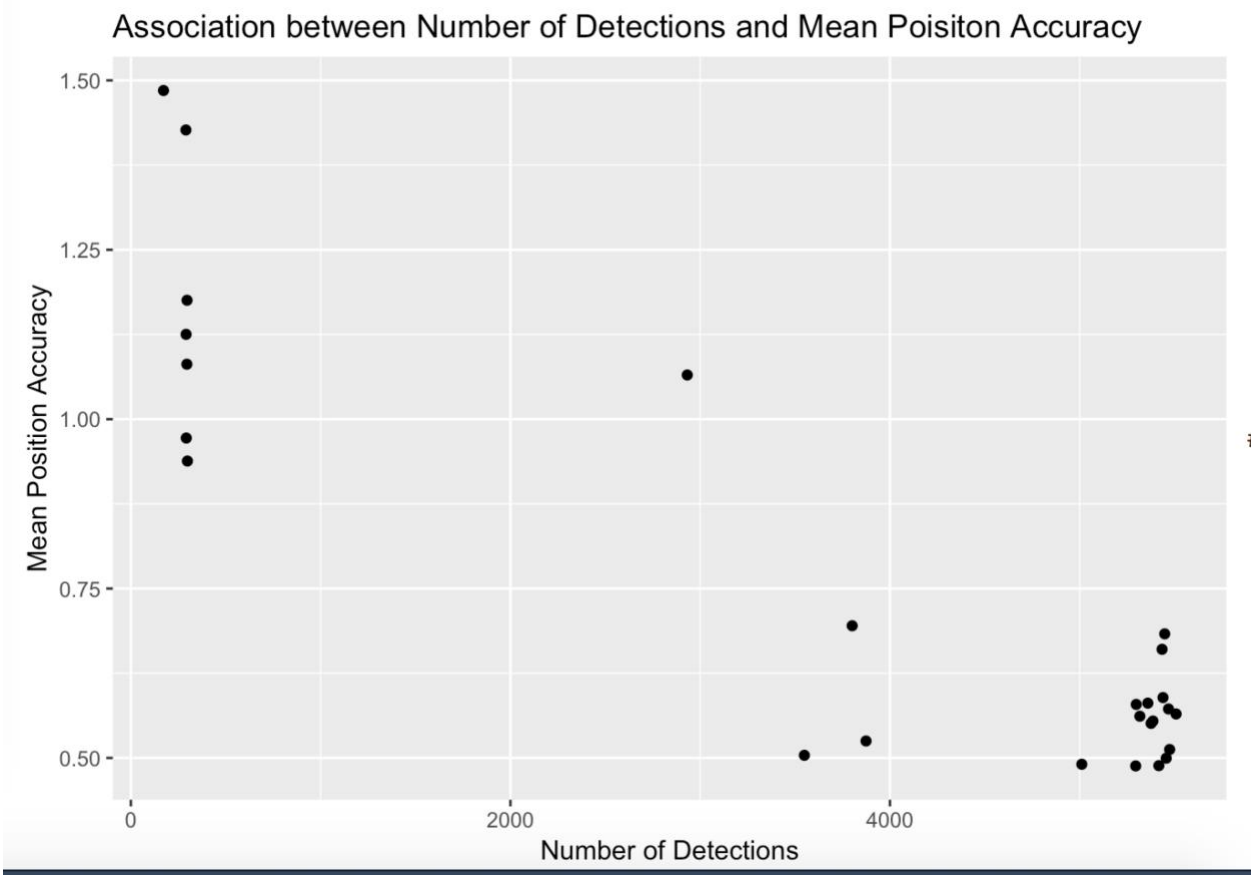


Based on the 'Mean Position Accuracy' plot, each tag seems to have very different accuracy position. Is that a consequence of different detection frequency?

I merge the tbl_postion_accuracy and tbl_freq_detected together for comparison.

```
tbl_freq_acc=merge(tbl_freq_detected,tbl_position_accuracy)

tbl_freq_acc %>%
  ggplot(aes(x=n,y=avg_accuracy))+geom_point()+labs(x="Number of Detections",y="Mean Position Accuracy"
,title="Association between Number of Detections and Mean Poisiton Accuracy ")
```



According to the plot shown above, tags which got detected in lower frequency have relatively higher position accuracies whereas tags which got detected in higher frequency (especially those more than 5000 times) possess much lower position accuracies.