MIDDLE EAST TECHNICAL UNIVERSTIY

DEPARTMENT OF COMPUTER ENGINEERING

SUMMER PRACTICE REPORT

CENG 300

Student Name: Yusuf Eren Tunç

Organization Name: OTTOO Yazılım Aş.

Address: Esenler/İstanbul

Start Date: 15/08/2019

End Date: 20/09/2019

Total Working Days: 30 days

STUDENT'S SIGNATURE **ORGANIZATION APPROVAL**

TABLE OF CONTENTS

1. INTRODUCTION

OTTOO Software Company interests with car fleet tracking systems. Company make products for the insurance, rent a car and logistics companies. In GPS systems, connection may be interrupt under special conditions. This interruptions occur because of device cannot receive information from satellite. In this interruptions, device's time deviation occur and route of device is lost. Project that I participate is about to solve this problem with correcting route and fixing time deviation with speed-distance guessing. I fixed this bug under the R & D's data scientist staff.
I will first mentions about the information about project then will be move on to organization.

2. INFORMATION ABOUT PROJECT

ANALYSIS PHASE-1

As the beginning of summer practice, I don't have any experience about R language and Data Science. My mentor make an education program in starter week for me. First of all, I start to learn about R language and it's libraries that I use in project which are ggplot for graphical examination of data, tidyverse for manipulating data and leaflet for realizing data to the map. I learn how to use R Studio and R language's functions efficiently from datacamp website. After that, I move on to learn to ggplot functions. Ggplot library is using for the visualizing data frame to graph. I understand that examining data frame with visual objects are much easier then examining from data frame's table. With that visualizing, one can be easily understand the data and find possible solutions. I learn ggplot from r-statistics.co web site. Moreover, Tidyverse is another library of R that contains several vectoral operations. This vectoral operations provide manipulating data frame. As I learned from my researches vectoral operations are fastest then loop operations in R and more effective too. At the beginning, I confused to how to use these vectoral operations in my solution because I never used vectoral operations before. I learn tidyverse library from datacarpentry.org website. Finally, leaflet library contains functions that take lat and long coordinates and realize them on the real world map. Leaflet library have several options like making circles on the map, drawing vectors from one point to another, using different type of point image and point labels and saving map as a html. I learn leaflet library from its offical webpage leafletjs.com. At the continuous weeks of my summer practice, these libraries

was very beneficial to solving problems of project. After that learning process, my mentor introduced me with the project. Project is about to missing time scales in the gps systems and its effects on the path of vehicle. That cause to disordered locations on the map and when vectors drawing between these locations to show path that doesn't look like a vehicle's movement.

With my mentor watch, I examined the disordered data. I used leaflet function to see how it looks like on the map. My first impression is locations of gps are correct on the map but because of time deflections, entries orders was wrong. I thought that if we can correct these disorder somehow, we can solve this issue. So, I started to check disorder parts of data to find how it look like. As I look these parts, I found that time deflection is constant time at each entry of disorder part. In other words, gps recorded the movements correctly but recorded time incorrect and that incorrect time was always a constant in disorder parts. I use ggplot and make a graph with order number of entries in the vertical axis and time as horizontal axis. I see that after disordered part finish there is a big time jump on the clock and actually that time jump represent that time pass at the disorder part. That time jump was seem as anomaly in data frame because time jumps are too much then normal entries time differences between each other. I thought that if we can detect that time anomalies somehow then we can detect the finish of the disorder part too. To conclusion, I learn R language and discovered that at disorder part there is a constant time deflection and finishing of disordered part is a time anomaly.

DESIGN PHASE-1

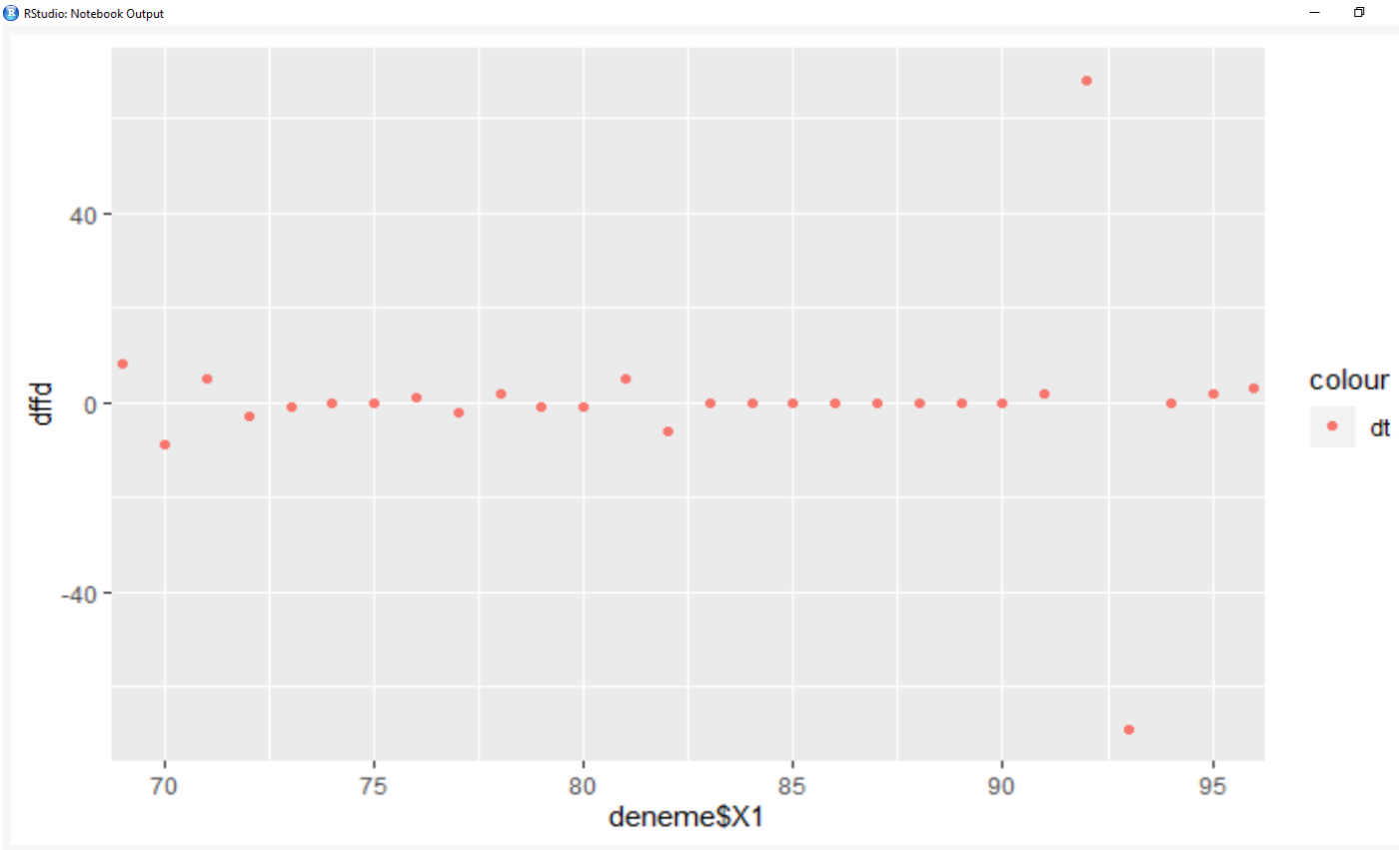
At part analysis-1, I explain that detect the disorder part of data frame by time jump which is a time anomaly on data frame and time deflection which is a constant time pattern. I thought that if I can find the time anomalies in the gps data frame somehow, we can find the finish points of the all disorder parts in data frame. Moreover, we can find the start of the disorder part with time deflection which is a constant pattern. When we have start and finish point of disorder part then we have the disorder part. After that analysis, I start to think about how to implement these ideas to code. I talked with my mentor about my ideas and he gave me an equation that use to detect anomalies in dataframe. Moreover, I took sequential extraction of time data and keep them as a new columns because I wanted to make constant pattern of time deflection to a zero pattern. Therefore, time jumps occur as anomaly in that column. I use that my mentor's equation on the new column that I created and it actually gave me the all-time jumps. With that, I have list of the starting points of all disorder parts in gps data frame. After that, I use ggplot to create a graph of that sequential extraction data and examine how long does zero pattern exists and is there any normal data between time jump which is finish point of

disorder data and time deflection. That time deflections which are zero patterns, exists more than three entry and sometimes there is normal entries which is not time deflection, between time jump and zero pattern. I create a loop to detect zero pattern from the time jump point starting and I record this data as Lost Path. Lost Path contains two more entries from start and end of disorder data because it is needed as references. To sum up, in this design phase, I create an algorithm that find start point of disorder data then record all disorder data as Lost Path.

IMPLEMENTATION PHASE-1

```
(abs(mean(deneme$ffd,na.rm=TRUE)-deneme$ffd)/sd(deneme$ffd,na.rm=TRUE)>1.5)
```

Anomaly Detect Equation (ffd represents time differences between sequential entries)



Time Difference Graph

Dot at 92 represent time jump and dots between 81 to 90 represent zero pattern with finish. Lost Path data frame contains entries from 94 to 80.

ANALYSIS PHASE-2

After part-1, I have all entries of disorder data and it is time to make order in disorder data. I analyse data frame with my mentor. We check several different data with graphs and map. In the end, we decide to correct data with respect to distance between each entry. In other words, algorithm start at the top of the disorder data and connect it to the closest entry. Then algorithm move on to that closest entry and connect it to the closest entry. With a loop, that work around to all disorder data and in the end there will be order. We couldn't find alternative solutions because there is not one. Therefore, gps data frame also keep speed of vehicle. I think that we can use this speed to find a possible range of vehicle at each entry. To find range of vehicle, we also need to know how much time passed between two entries. From part-1, I discover that time jump is time that passed at disorder part. Since I have also how many entries are in disorder part, I found a mean time for every entry. With famous speed, time and distance formula, I can find disorder part's all entries possible ranges. These ranges like a reference point to improve distance solutions accuracy. Therefore, distance solution still not be accurate in every cases. For example, if there is a disorder in circular paths, solution may not be give exactly order. I talked about that with my mentor and he says that there isn't any circular lost path and if there is, it would be a rare case. Since it is not a %100 percent solution, we decide to go on distance solution. In summary, I figure out that distances between entries can be used for solve project's problem with an estimated range.

DESIGN PHASE-2

At the beginning of the design process, I create an average time from time jump and length of the data frame division. I use this average time to find an estimated range of each entries with speed production. This range was my reference point. After that, I move on to distance solution. My first problem is coordinate of entries are lats and longs. I try to find a distance equation of between two lats longs coordinates. At my research, I found a function in web called to earth.dist from conservationecology.wordpress.com. I use these function to find distance. Another problem that I faced was how to create a data frame that keep all entries' distance between each other. I make another research to how to create that effectively. I thought to use a for loop for that problem and I talked with my mentor. My mentor told me to for and while loops are too slow in R language. I need to find a vectorial solution. I discovered that apply functions in R is best solution to my problem. I learn apply functions and use mapply function to create a data frame that contains distances between all entries. At that data frame, from starting point each entries represent a column. In other words, lost path's each entries represented in data frame by a column. Also each row is represented the entries too. In the end of the rows, I add estimated range of each entries. I called that data frame to msfdata. In the msfdata, I subtract each column's range from column's entry. Moreover, I change diagonal entries of this data frame to a very little number because each diagonal entry represent the column's distance to itself and diagonal entries must be ignored in the future. Each of these operations done by vectorial functions to increase speed and stability of project. With msfdata, I try to make a list of ordered entries. My algorithm start with first column of msfdata. Then, it find smallest entry and record that entries name to a list. After that my algorithm delete column that it look and move on to smallest entries column. Algorithm repeat that process until it reach to last column. In the end, I have a list of correct order of lost path. Only problem is all this process done in a for loop and a bit slower. I examine algorithm with my mentor several times and we couldn't find a vectorial algorithm because algorithm travel msfdata without a certain path. All vectorial operations needed a certain path. Therefore, I have an order list and I change the lost path data frame. At that time, my algorithm work correct and path was fixed but time was still a problem. To conclusion, I fixed disorder part with a distance data frame creation and a loop.

IMPLEMENTATION PHASE-2

```
zaman = (as.integer(ano_dt$ffd[1]))/(as.integer(count(kayip)))
```

Average time equation for estimated range

```
```{r}
#Long-Lat Koordinat Sisteminde Mesafe Bulma Fonksiyonu
#https://conservationecology.wordpress.com/2013/06/30/distance-between-two-points-i?-r/
  earth.dist <- function (long1, lat1, long2, lat2){
    rad <- pi/180
    a1 <- lat1 * rad
    a2 <- long1 * rad
    b1 <- lat2 * rad
    b2 <- long2 * rad
    dlon <- b2 - a2
    dlat <- b1 - a1
    a <- (sin(dlat/2))^2 + cos(a1) * cos(b1) * (sin(dlo?/2))^2
    c <- 2 * atan2(sqrt(a), sqrt(1 - a))
    R <- 6378.145
    d <- R * c
    d <- d*1000
    return(d)
  }
```
```

Earth.dist function from conservationecology.wordpress.com

```
```{r}
#Noktalar Arasi Mesafe Datasını Olusturma

msf<- kayip %>% mutate(range = kayip$spd*(zaman/3.6)+((kayip$spd*(zaman/3.6)))/4) %>%
select(X1,lng,lat,range)
longs<-kayip %>% select(lng)
lats<- kayip %>% select(lat)
x<-mapply(earth.dist,rowwise(longs),rowwise(lats),msf$lng,msf$lat)
msfdata <- as.data.frame(x)
name <- kayip %>% select(X1)
name <- as.list(name)
colnames(msfdata)<-unlist(name)
rownames(msfdata)<-unlist(name)
msfdata <-msfdata %>% rbind(range = msf$range)

```
```

Algorithm that create msfdata

At that algorithm, mapply is equivalent to function that contain two for loop.

I didn’t share algorithm that create order list of msfdata due to company privacy.

ANALYSIS PHASE-3

I start to work on to correct time at Lost Path data frame. I examine time differences and as I mentioned before time that pass in the Lost Path entries show in the time jump. Also, I have speed of all entries and distance between each other. With speed and distance objects, I can apply a function that extract time from speed and distance division. Therefore, time may flow out cause of estimation. We can avoid that with a rate of total time and estimated time for each entry.

DESIGN PHASE-3

First of all, my algorithm found total real time that pass in the Lost Path. After that, algorithm create a column that keeps estimated time for each entry with vectorial operations. That column gives us all entries estimated times in order. However, in some cases total of these estimated times can be bigger than total real time. To solve this problem I add some extra codes to algorithm. Total time that we have divide by total time that we found from speed and distance division. It is a rate for us. Then, for each entry we take product of entries estimated time and rate and change entries time data with data that we found. These operations provide that total estimated time can’t get bigger than total real time. After all of this operations, my algorithm was ready for testing.

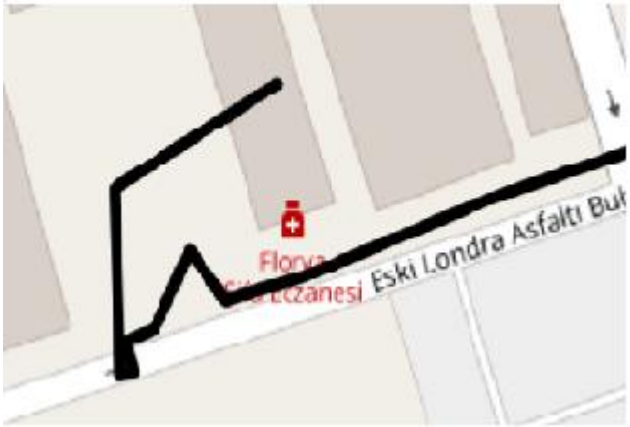
IMPLEMENTATION PHASE-3

There is only one function that correcting time. However, I didn’t share it due to company privacy.

TESTING PHASE

I tested my algorithm with several cases. In test cases, I realize that if vehicle starting or finishing its path from an underground positon (such as garages or under bridges) then algorithm failed. So I added a control function that name is Checking_Underground. That function checks time differences and movements of vehicle and if both of them same algorithm doesn't try to correct that situation.

At the continuous of there is images from tested cases by my algorithm.



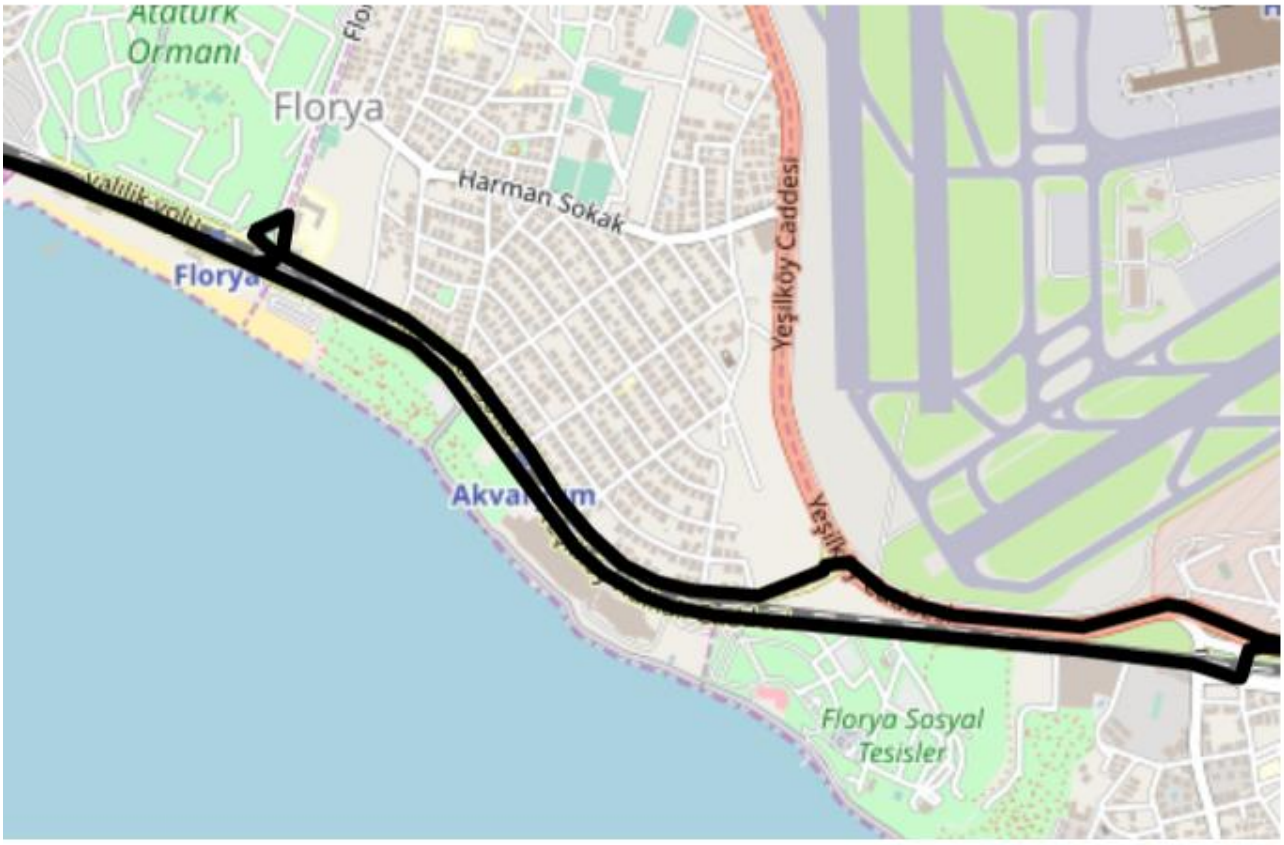

D100
Kartaltepe Mahallesi
Metrobüs
Küçükçekmece

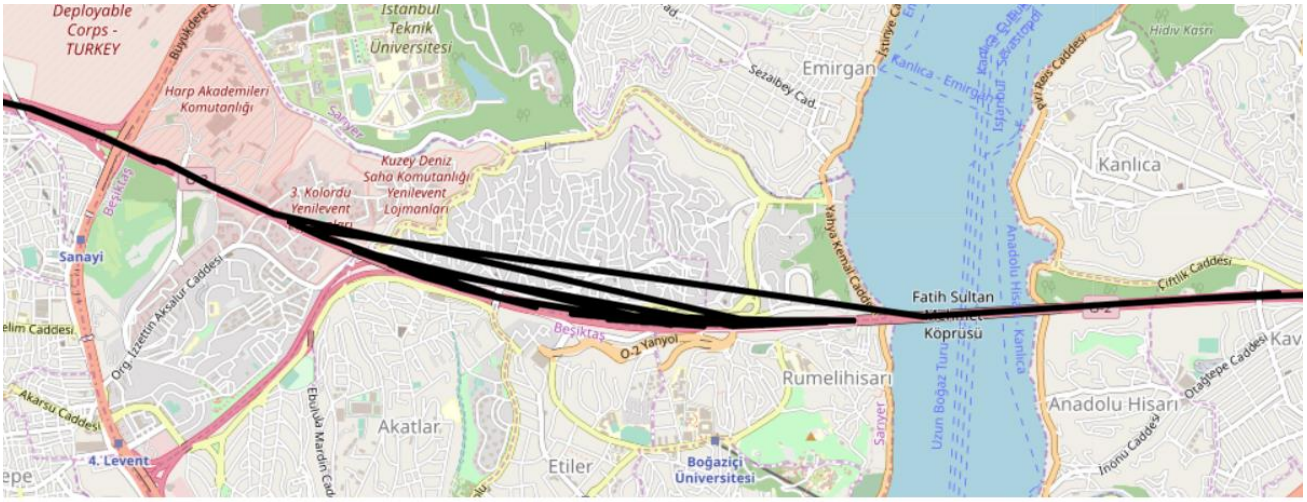

D100
Kartaltepe Mahallesi
Metrobüs

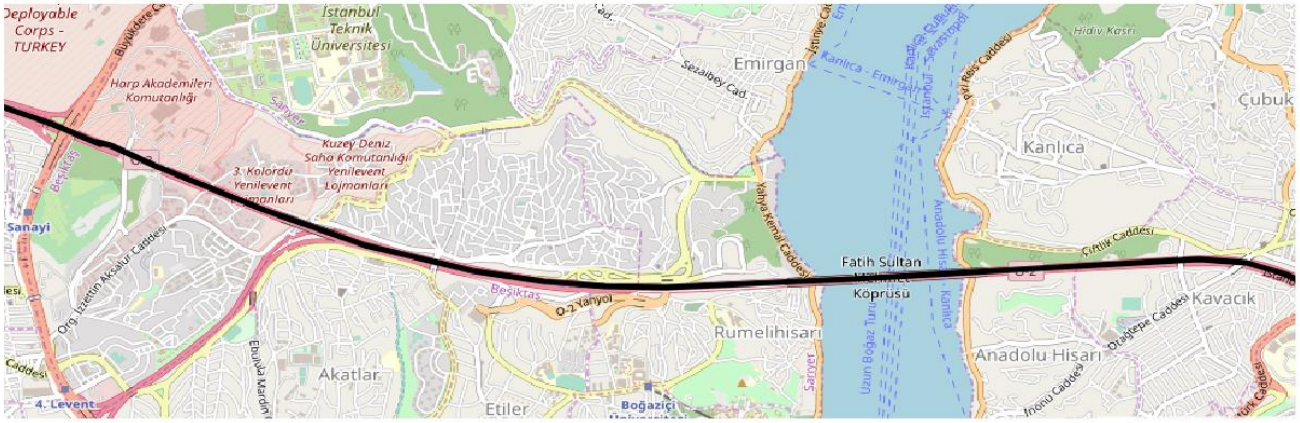

D-100
Defterdar
Metrobüs
Namazgâh Sokağı
Güldalı Sokak
Şişhane Caddesi

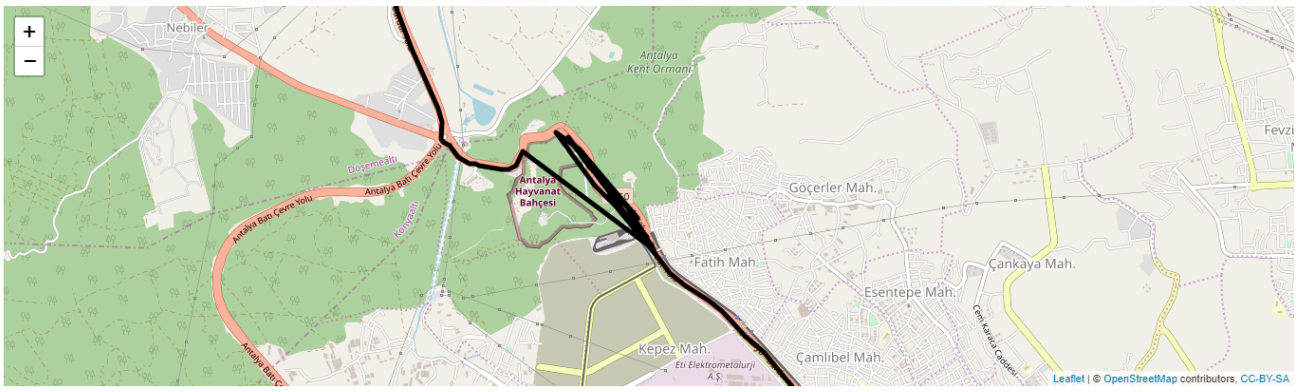

D-100
Defterdar
Metrobüs
Namazgâh Sokağı
Güldalı Sokak
Meşatlık Sokağı
Şişhane Caddesi

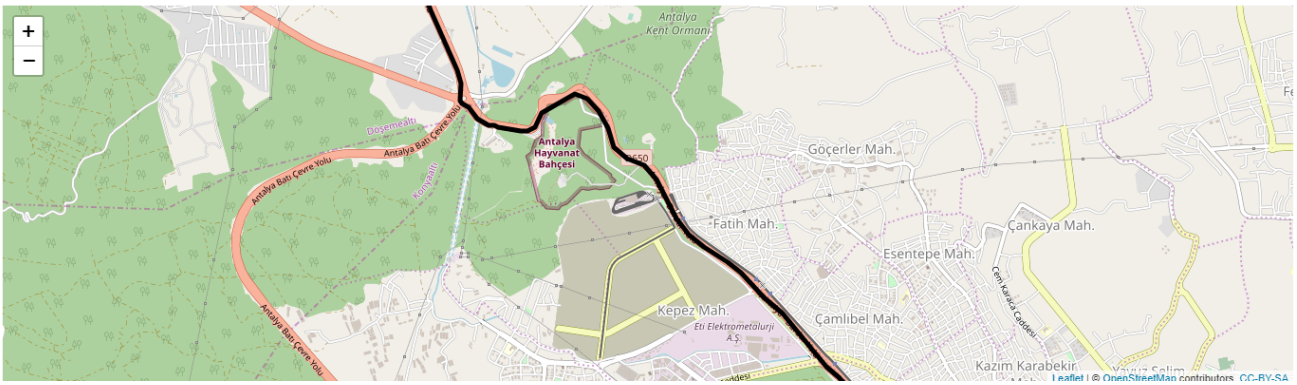

Atatürk
Ormanı
Florya
Harman Sokak
Yeşilköy Caddesi
Florya Sosyal
Tesisler

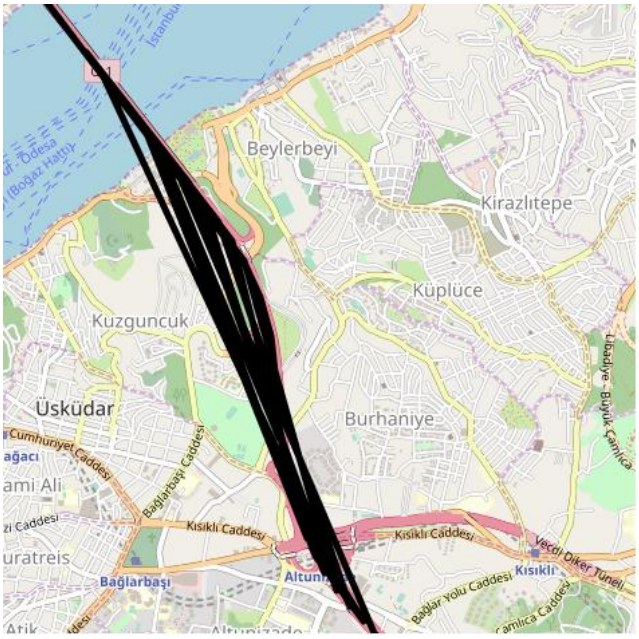

Atatürk
Ormanı
Florya
Harman Sokak
Yeşilköy Caddesi
Florya Sosyal
Tesisler

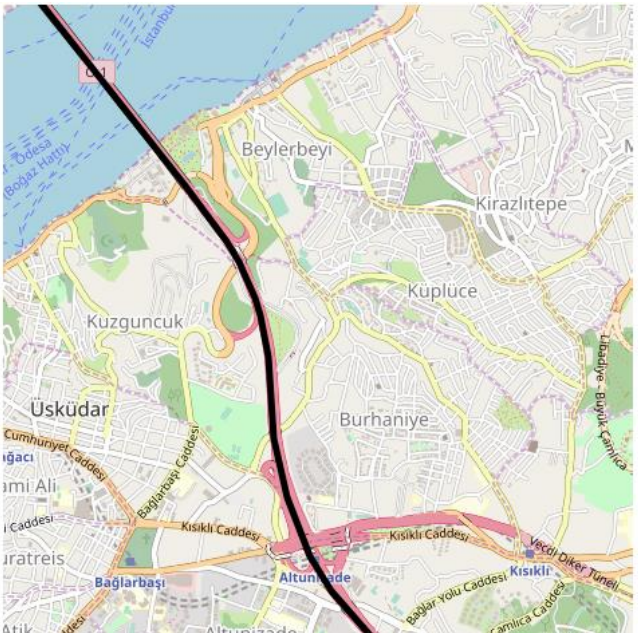

Deployable Corps - TURKEY
İstanbul Teknik Üniversitesi
Harp Akademileri Komutanlığı
Kuzey Deniz Saha Komutanlığı Yenilevent Lojmanları
3. Kolordu Yenilevent
Emirgan
Sezaibey Cad.
Kanlıca
Hıdiv Kasrı
Sanayi
Org. İzzettin Aksalur Caddesi
Beşiktaş
O-2 Yanyol
Fatih Sultan Mehmet Köprüsü
Çınlık Caddesi
Rumelihisarı
Anadolu Hisarı
Akatlar
Etiler
4. Levent
Boğaziçi Üniversitesi
İnönü Caddesi
Sarıyer
Uzun Boğaz Turu


Deployable Corps - TURKEY
İstanbul Teknik Üniversitesi
Harp Akademileri Komutanlığı
Kuzey Deniz Saha Komutanlığı Yenilevent Lojmanları
3. Kolordu Yenilevent
Emirgan
Sezaibey Cad.
Çubuk
Kanlıca
Hıdiv Kasrı
Sanayi
Org. İzzettin Aksalur Caddesi
Beşiktaş
O-2 Yanyol
Fatih Sultan Mehmet Köprüsü
Çınlık Caddesi
Kavacık
Rumelihisarı
Anadolu Hisarı
Akatlar
Etiler
4. Levent
Boğaziçi
İnönü Caddesi
Sarıyer
Uzun Boğaz Turu


Nebiler
Antalya Kent Ormanı
Döşemealtı
Antalya Batı Çevre Yolu
Antalya Hayvanat Bahçesi
Göçerler Mah.
Fevzi
Fatih Mah.
Çankaya Mah.
Esentepe Mah.
Cem Karaca Caddesi
Kepez Mah.
Çamlıbel Mah.
Eti Elektrometalurji A.Ş.
Leaflet | © OpenStreetMap contributors, CC-BY-SA


Antalya Kent Ormanı
Döşemealtı
Antalya Batı Çevre Yolu
Antalya Hayvanat Bahçesi
Göçerler Mah.
Fatih Mah.
Çankaya Mah.
Esentepe Mah.
Cem Karaca Caddesi
Kepez Mah.
Çamlıbel Mah.
Eti Elektrometalurji A.Ş.
Kazım Karabekir
Leaflet | © OpenStreetMap contributors, CC-BY-SA


Beylerbeyi
Kirazlıtepe
Küplüce
Kuzguncuk
Üsküdar
Burhaniye
Bağlarbaşı
Kısıklı Caddesi
Bağlar Yolu Caddesi
Kısıklı
Vecdi Diker Tüneli


Beylerbeyi
Kirazlıtepe
Küplüce
Kuzguncuk
Üsküdar
Burhaniye
Bağlarbaşı
Kısıklı Caddesi
Bağlar Yolu Caddesi
Kısıklı
Vecdi Diker Tüneli


Lütfi Aykaç Bulvarı
TEM Avrupa Konutları
Mevlana
Eyyap Caddesi
Fatih Sultan Mehmet Bulvarı
5. Levent
Atatürk Caddesi

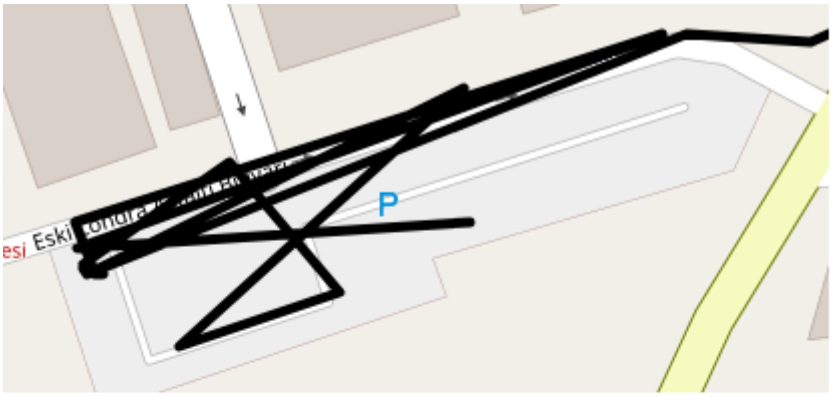

Lütfi Aykaç Bulvarı
Sultangazi
Gaziosmanpaşa
TEM Avrupa Konutları
Mevlana
Atatürk Caddesi
Eyyap Caddesi
Fatih Sultan Mehmet Bulvarı
5. Levent

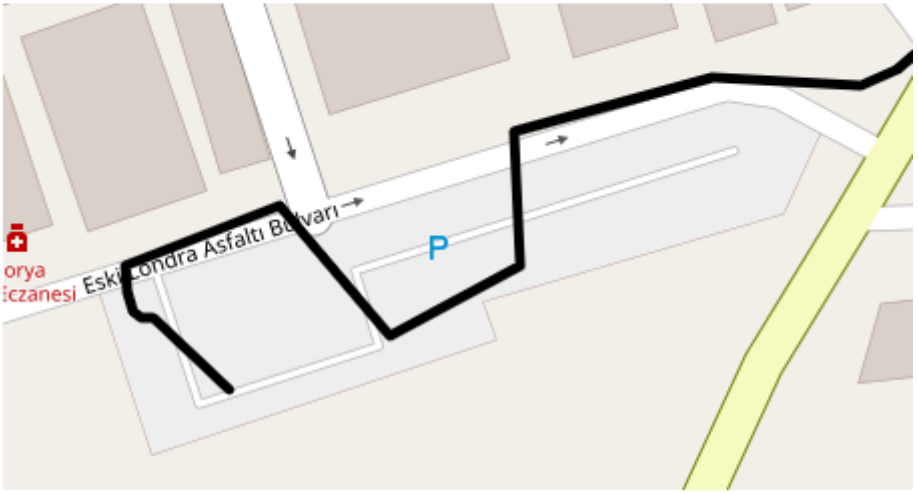

Yıldız Parkı
Yıldız
Palanga Caddesi
Müvezzi Caddesi
Asâriye Caddesi
Ziya Kalkavan
Denizcilik
Anadolu
Teknik
Lisesi
Yıldız Mahallesi
Beşiktaş


Yıldız Parkı
Yıldız
Palanga Caddesi
Müvezzi Caddesi
Asâriye Caddesi
Ziya Kalkavan
Denizcilik
Anadolu
Teknik
Lisesi
Yıldız Mahallesi

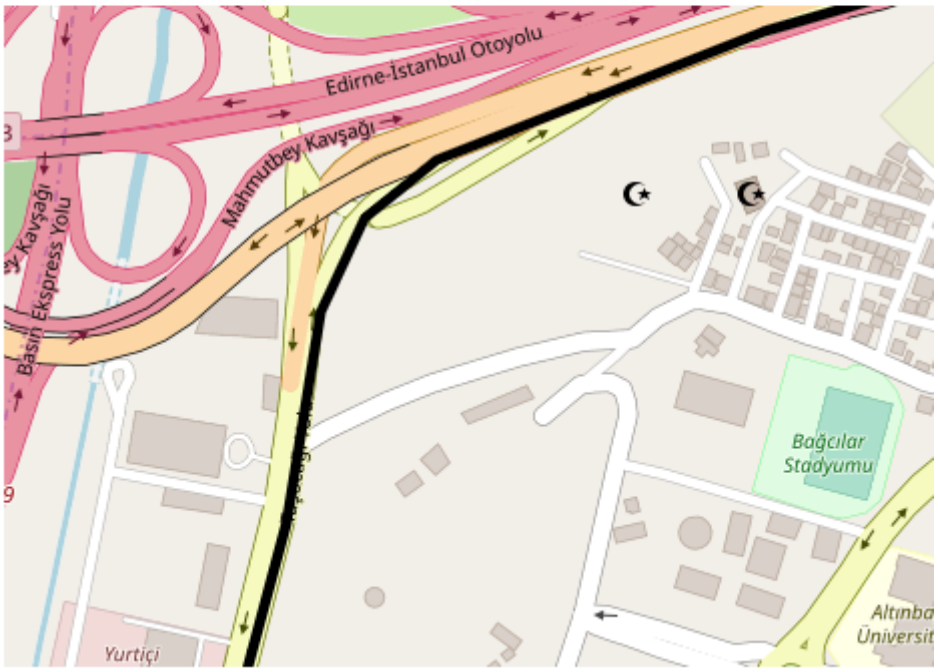

Beşiktaş

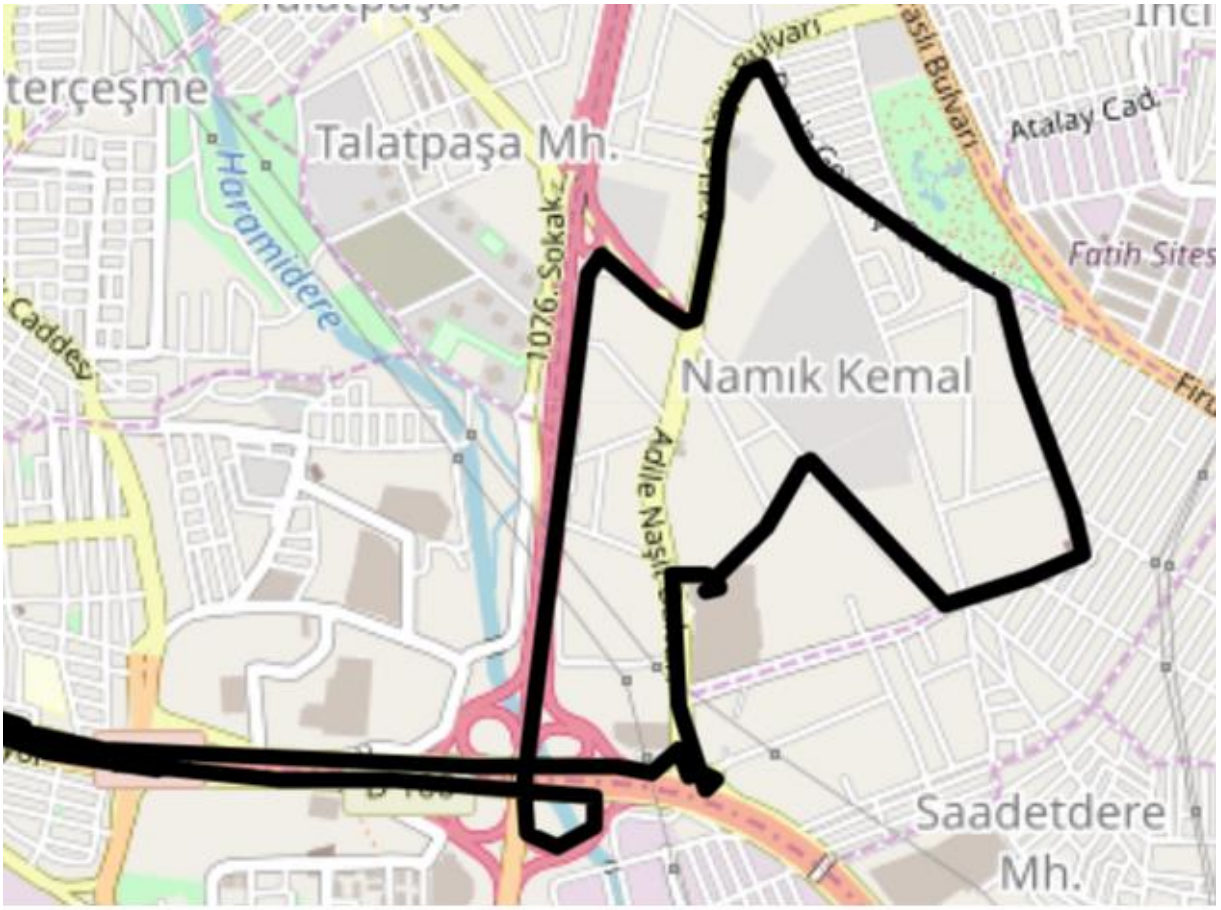

Beşiktaş
Bebek Arnavutköy Caddesi
Kolej

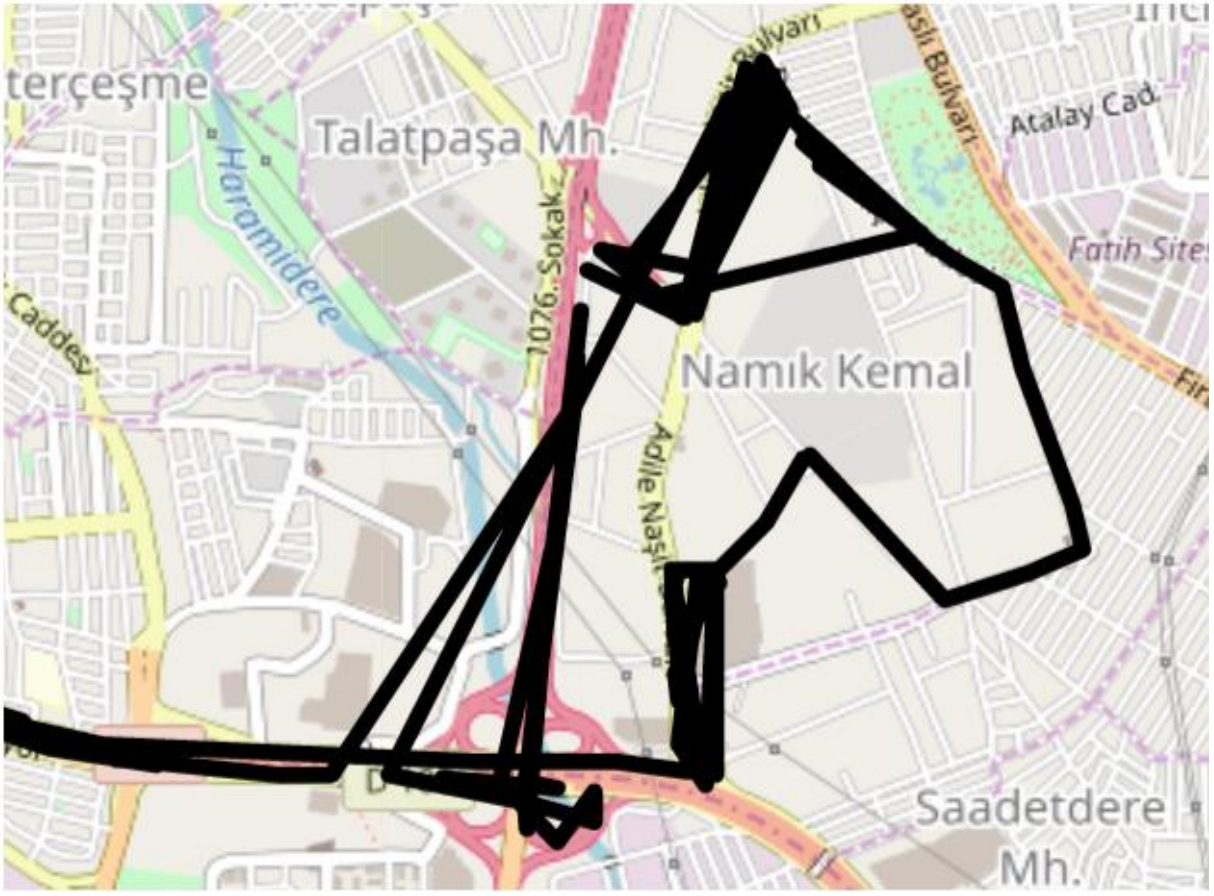

P


Eski Londra Asfaltı Bulvarı
P


Edirne-İstanbul Otoyolu
Mahmutbey Kavşağı
Basın Ekspres Yolu
Yurtiçi
Kargo İstanbul
Avrupa
Bölge Müdürlüğü


Edirne-İstanbul Otoyolu
Mahmutbey Kavşağı
Basın Ekspres Yolu
Bağcılar
Stadyumu
Yurtiçi


terçeşme
Talatpaşa Mh.
Haramidere
1076. Sokak
Atalay Cad.
Namık Kemal
Saadetdere
Mh.


terçeşme
Talatpaşa Mh.
Haramidere
1076. Sokak
Atalay Cad.
Namık Kemal
Saadetdere
Mh.

3. ORGANIZATION

ORGANIZATION AND STRUCTURE

Otto Yazılım is a start-up company. As I mentioned in Introduction part, company interests with car fleet tracking systems. All of employees are computer engineering and they are young. Although, it is a small and start-up company, there is a hierarchy. There is CEO, CTO, Product Manager etc. CTO was my mentor and ambience of company likes a friends environment.

METHODOLOGIES AND STRATEGIES USED IN THE ORGANIZATION

I observed that company hire many interns to educate them. Most of interns was long term interns and they was at bottom level when they started to work at company. Especially, CTO be interested in interns, encourage them and educate them. With that strategy, they plan to train educated employee for future of their company.

4. CONCLUSION

To sum up, my internship was at OTTO Yazılım which is a company at İstanbul. I learned about data science, data manipulation and R language. After that intern, I interests to data science because I enjoyed it.Therefore, it was my first job experience and it was satisfying due to learning R language and creating a product. In the end of my internship, my mentor said that they will use my algorithm in their product and that make me a bit proud and confident.