EDA of Traffic Announcements (Istanbul-UYM)

13 12 2020

Contents

EDA of Traffic Announcements

This dataset contains set of traffic announcements entered into the system by the Transport Management Center operators from 2018 to 2020. There are 12 variables and 58422 observations.

Project Proposal Link

Preprocessing

The first step was combining the excel files, since there were two separate datasets.

The name of variables were converted to ENG.

The data type of the *Announcement_Types* were converted from char to factor. The values of the *Announcement_Types* were converted to ENG.

The data type of the *Accident_Types* were converted from char to factor. The values of the *Accident_Types* were converted to ENG.

EDA

The various data analysis and visualizations will be made on EDA part. In further analysis *dplyr* and *tidyverse* package will be used for data manipulation, *ggplot2* package will be used for creating plots and visuals and *lubridate* package will be used to make it easier to work with dates and times, *data.table*

package will be used to make it easier to work with data tables, *leaflet* package will be used to mapping. In order to get healthy and reliable results, care has been taken to ensure that the ending time of the project/construction etc. is greater than the starting time of it. But, since the date and time structure has been changed since 2020, it is observed that in some records ending time is smaller than starting time -Especially in recordings between 11 am to 2 pm -. In order to make this point clear, we want to give an example. Announcement ID: 602730, Landscaping project, starting time is 10/15/2020 10:49 AM and ending time is 10/15/2020 3:17:00 AM. Ending time of this project looks smaller than starting time, the reason why it is wrong is the difference in am/pm. Since there are many such erroneous records, these records were filtered out in calculations but were kept in count operations.

Besides, there were negative values for *Effected_Lanes*, these were excluded.

SET-1

The traffic is assumed to be caused by two variables in our data set. These are, number of lanes affected from the incident and the duration of it. So "Problem Index" is created from multiplication of these variables .

```
P1<- DS %>%
mutate(Diff_Time_Hours = round(difftime
(DS$Ending_Time,DS$Starting_Time)/3600,digits = 2))
P1 <- P1 %>%
mutate(Problem_Index = P1$Diff_Time_Hours*P1$Effected_Lanes)
P1$Diff_Time_Hours = as.numeric(P1$Diff_Time_Hours)
P1$Problem_Index = as.numeric(P1$Problem_Index)
```

The outliers were detected with statistical methods. Records that were 3 standard deviations away from the mean is excluded from the data set. Approximately %99 confidence interval is created. Data is grouped by announcement type. Summarized and sorted according to the problem index.

Date-related recording errors are filtered out from the subset because of the calculations.

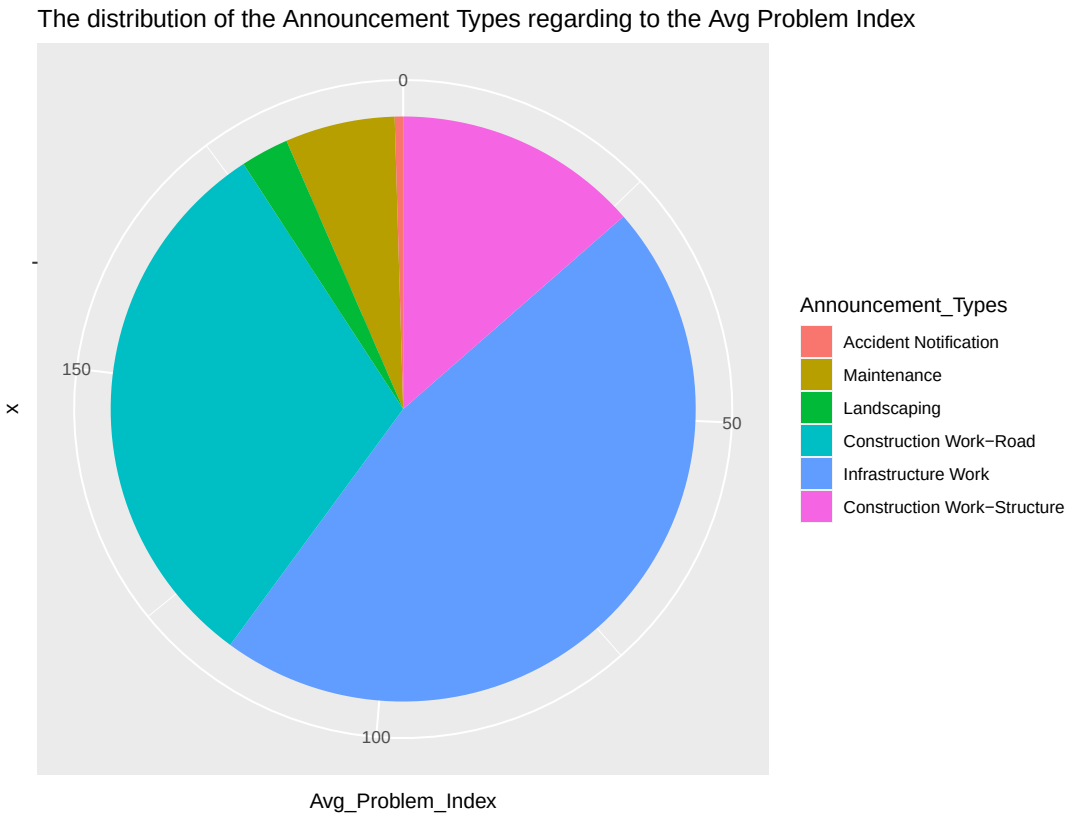
```
P1_1 <- P1 %>% filter(Problem_Index >= 0) %>%
  filter(Problem_Index<= 3*sd(Problem_Index)+
  mean(Problem_Index) )%>%
  group_by(Announcement_Types) %>%
  summarise(Avg_Problem_Index = mean(Problem_Index),
  Count_Announcement_Types= n(),
  Sum_Problem_Index=sum(Problem_Index))%>%
  filter(Avg_Problem_Index > 0)
P1_1 <- data.frame(P1_1)
P1_1<- P1_1[order(P1_1$Avg_Problem_Index,decreasing = TRUE), ]
P1_1
```

```
##            Announcement_Types Avg_Problem_Index Count_Announcement_Types
## 5         Infrastructure Work        90.6383582                       67
## 4      Construction Work-Road        59.9131707                       41
## 6 Construction Work-Structure        26.4447059                       17
## 2                 Maintenance        11.8293100                     5290
## 3                 Landscaping         5.1812475                      505
## 1       Accident Notification         0.9064472                    26250
```

```
##   Sum_Problem_Index
## 5           6072.77
## 4           2456.44
## 6            449.56
## 2          62577.05
## 3           2616.53
## 1          23794.24
```

The distribution of the announcement types regarding to the average problem index, are shown on the pie chart.

```
ggplot(P1_1,aes(x="", y=Avg_Problem_Index, fill=Announcement_Types)) +
geom_bar(stat='identity',width = 1)+
coord_polar("y")+
labs(title =
"The distribution of the Announcement Types regarding to the Avg Problem Index")
```



- It is shown that infrastructural or road constructional incidences were the worst traffic makers on average. However, car accidents and maintenance works occurred more frequently.Therefore, when it was summed up, these were the main traffic generators in total.

In order to identify top 5 locations regarding problem index, the calculations are done.

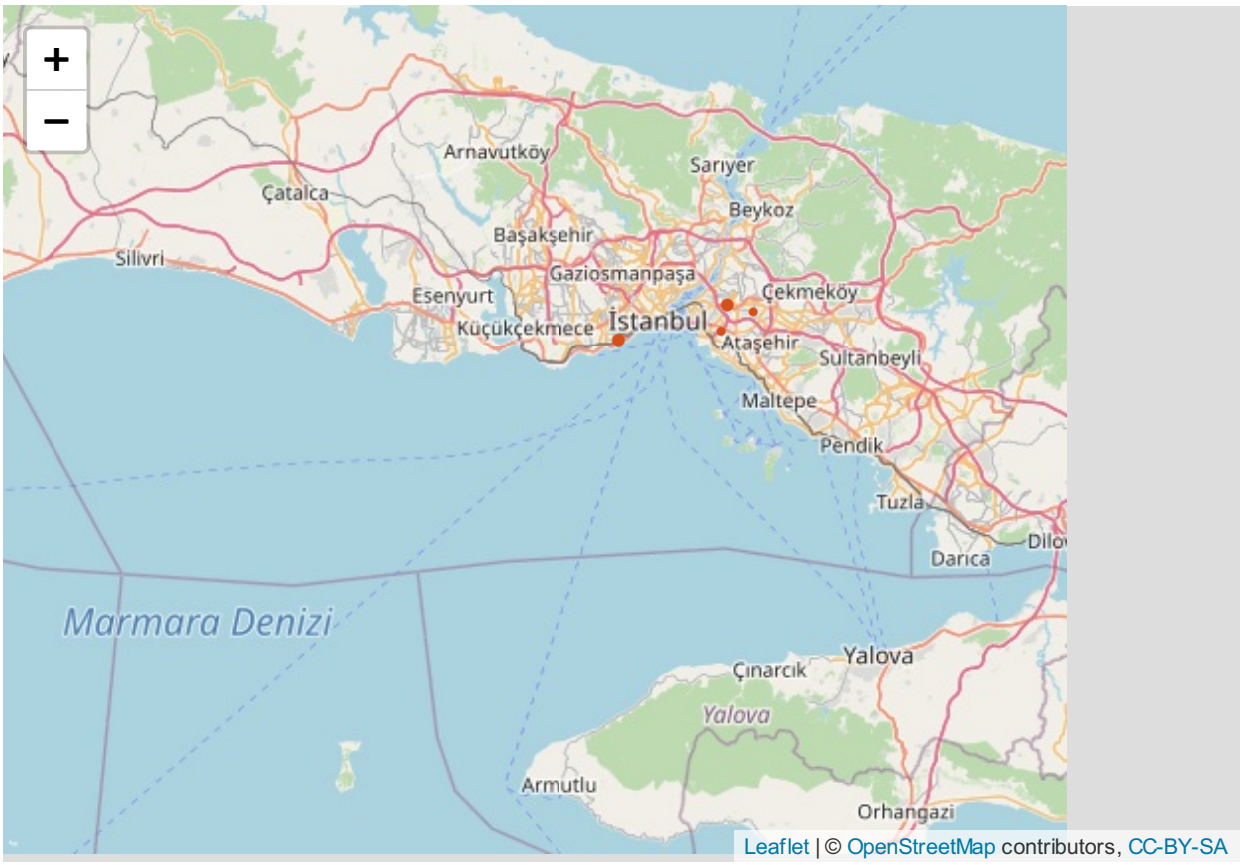
Date-related recording errors are filtered out from the subset because of the calculations.

```
P1_2 <- P1 %>% filter(Problem_Index >= 0) %>% select(Location,Problem_Index,Coordinates)
P1_2 <- data.frame(P1_2)
P1_2 <- P1_2[order(P1_2$Problem_Index,decreasing = TRUE), ]
P1_2 <- head(P1_2,5)
P1_2 <- data.frame(P1_2,do.call(rbind,strsplit(P1_2$Coordinates,",")))
P1_2$X2=as.double(P1_2$X2)
P1_2$X1=as.double(P1_2$X1)
P1_2  %>% select(Location,Problem_Index)
```

```
##                                          Location Problem_Index
## 11553                  S.Yolu Z.Burnu-Bakırköy yönü      23801.00
## 1347    Tünel Vecdi Diker Altunizade - Ümraniye yönü      18630.00
## 5030                        D100 Kınalı-Silivri Yönü      13202.00
## 6734  D100 Kadıköy-Çamlıca D100 Çamlıca-Kadıköy Yönü      13069.77
## 3687         Şile Yolu Tepeüstü-Ümraniye Kavşak Yönü      10921.15
```

These 5 locations are shown on the map.

```
leaflet(P1_2) %>%
  setView(lng = 28.70, lat = 41, zoom = 9.1)  %>%
  addTiles() %>%
  addCircles(data = P1_2, lat = ~ X1,
  lng = ~ X2, weight = 1,
  radius = ~sqrt(Problem_Index)*5,
  popup = ~as.character(Location),
  color = "#d3501d", fillOpacity = 20)
```



SET-2

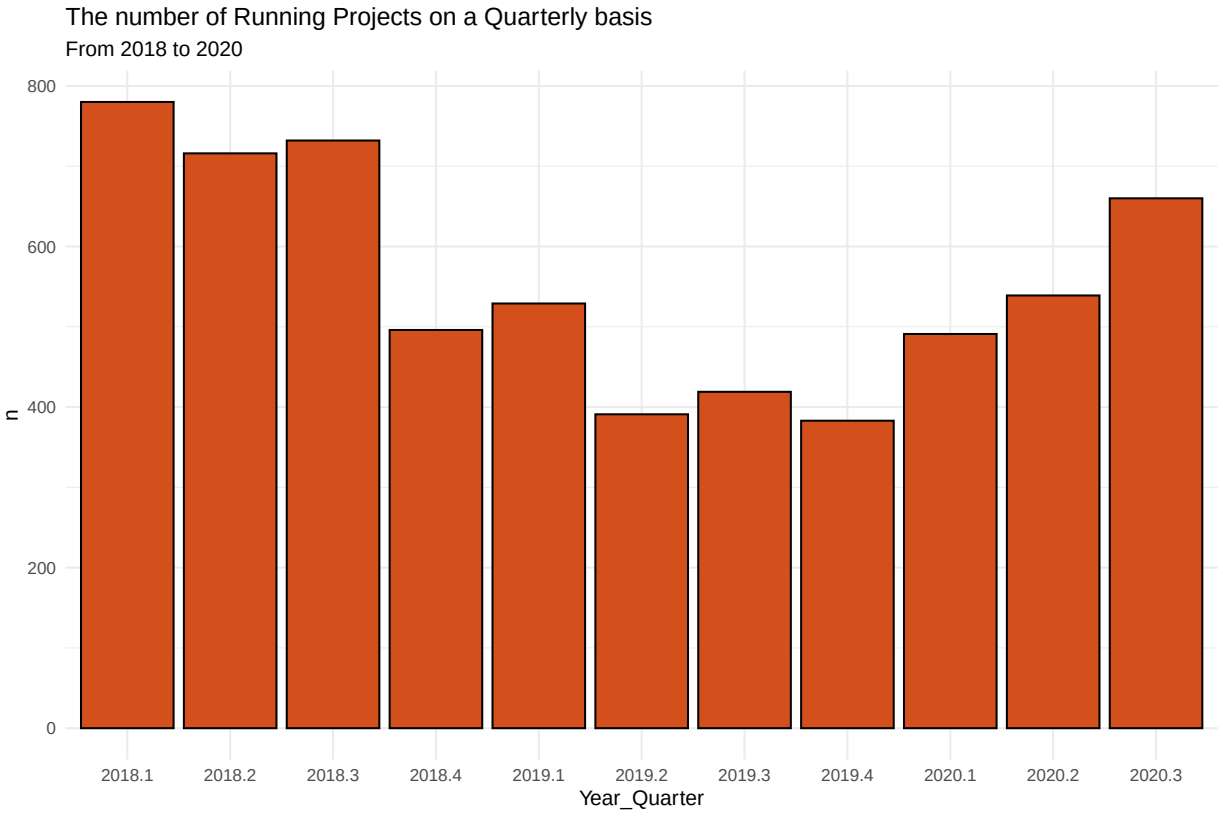
In the sub-study below, it is examined whether there is a relationship between the municipality elections and the number of running projects/constructions etc.

Since the November of 2020 is not completed yet, is not included in the study and date-related recording errors are kept on the subset.

```
P2 = DS %>%
filter(Announcement_Types %in% c('Landscaping',
'Maintenance','Infrastructure Work',
'Construction Work-Structure',
'Construction Work-Road')
& Effected_Lanes >= 0)
P2_1 = P2 %>%
group_by(Year_Quarter= lubridate::quarter(Starting_Time, with_year = TRUE)) %>%
summarise(n = n())
P2_1$Year_Quarter = factor(P2_1$Year_Quarter)
P2_1=P2_1 %>% filter(Year_Quarter != "2020.4")
```

The number of the running projects such as construction, landscaping are shown on the bar chart on a quarterly basis.

```
ggplot(P2_1, aes(x=Year_Quarter, y=n)) +
geom_bar(stat="identity", position="dodge",fill="#d3501d", colour="black")+
theme_minimal()+
labs(title = "The number of Running Projects on a Quarterly basis",
subtitle ="From 2018 to 2020")
```



- The number of projects carried out in 2018 decreased except in the third quarter.
- Considering that there will be an election in March 2019, the number of projects carried out compared to the last quarter of the previous year has increased, but it was considerably lower than the same quarter of the previous year.
- After the election repetition decision, the existing administration reduced the number of projects it carried out and most probably allocated resources for the second tour of the election.
- In the second quarter of 2019, the number of projects carried out according to the same content of last year has almost halved. And it was one of the least studied quarters of the 3-year period.
- With the the new administration and the elimination of election uncertainty, the number of projects carried out by the municipality has started to increase regularly since the last quarter of 2019.
- Although, the number of projects carried out in 2020 is above the projects carried out in 2019 – on a quarterly basis -, is below the 2018 values. The reason for this may be pandemic and economic problems.

SET-3

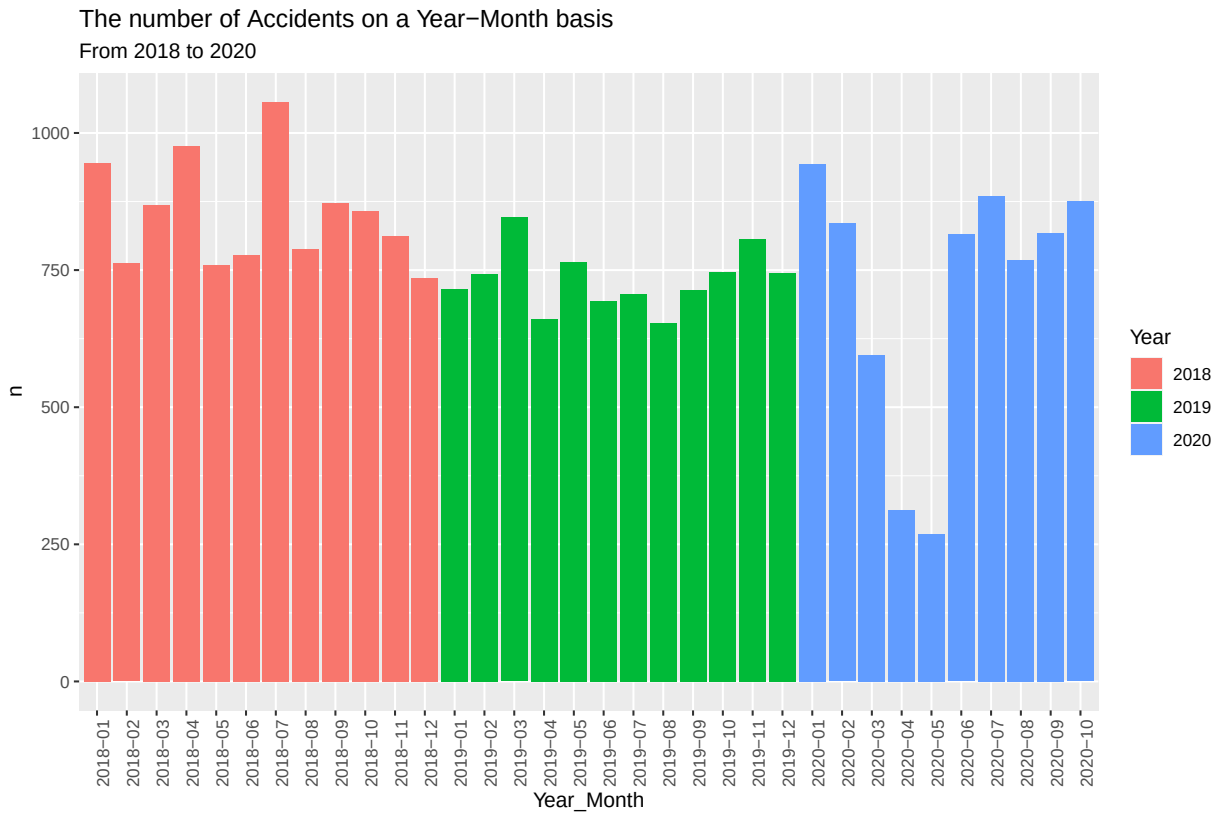
As a third section, it is investigated whether there is an effect of the pandemic on the number of Accidents.

Since the November of 2020 is not completed yet, is not included in the study and date-related recording errors are kept on the subset.

```
P3=setDT(DS)[, Year_Month := format(as.Date(‘Starting_Time‘), "%Y-%m") ]
P3= P3 %>% mutate(Year=year(Starting_Time),Day=day(Starting_Time))
P3_1 = P3 %>%
filter(Announcement_Types == "Accident Notification" &
Effected_Lanes >= 0 & Year_Month != "2020-11") %>%
group_by(Year_Month,Year) %>% summarise(n = n())
P3_1$Year=as.factor(P3_1$Year)
```

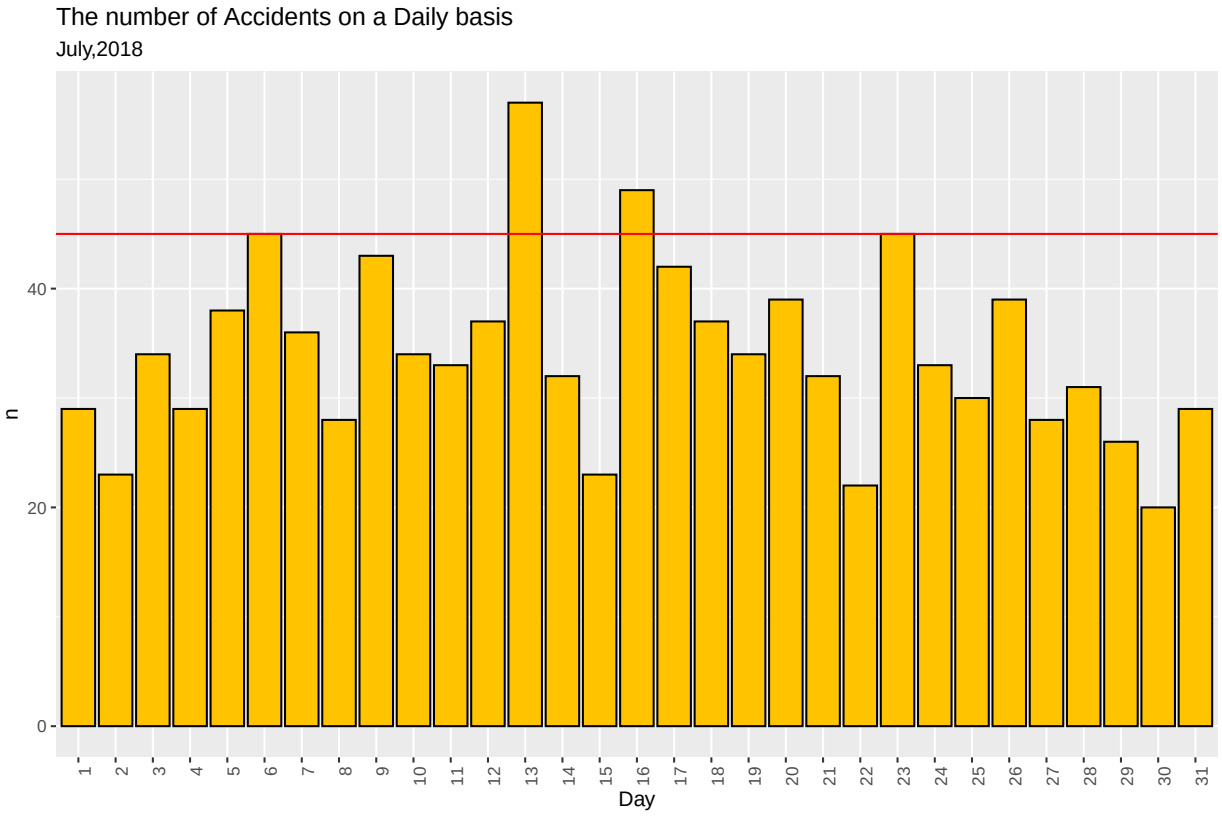
The number of Accidents are shown on the bar chart on a year-month basis. The bar chart has been colored on an annual basis for a clear view.

```
ggplot(P3_1,aes(x=Year_Month,y = n)) +
geom_bar(stat="identity",aes(fill=Year))+
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
labs(title = "The number of Accidents on a Year-Month basis",
subtitle ="From 2018 to 2020")
```



- There is a clear effect of the pandemic that can be seen in 2020. The first pandemic case occurred in March in Turkey. In April and May, there were a lot of lockdowns were applied. The number of Accidents significantly decreased in that period. After rule bendings and starting of the summer period, the numbers increased again.
- Besides, there is a peak on 2018-07 that is seen on the chart. To investigate in more detail, the daily number of the Accidents in 2018-07 is plotted.

```
P3_2 = P3 %>%
filter(Announcement_Types == "Accident Notification"
& Effected_Lanes >= 0 & Year_Month == "2018-07") %>%
group_by(Day) %>%
summarise(n = n())
P3_2$Day=as.factor(P3_2$Day)
ggplot(P3_2,aes(x=Day,y = n)) +
geom_bar(stat="identity",fill="#FFC300", colour="black")+
theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
geom_hline(yintercept=45, color="red")+
labs(title = "The number of Accidents on a Daily basis",subtitle = "July,2018")
```



- 13th and 16th of July had the highest number of Accidents. 15'th of July is a national holiday in Turkey. 15'th of July was Sunday so the increase of these days could be because of that people tend to combine holidays and travel more before and after national holidays.

SET-4

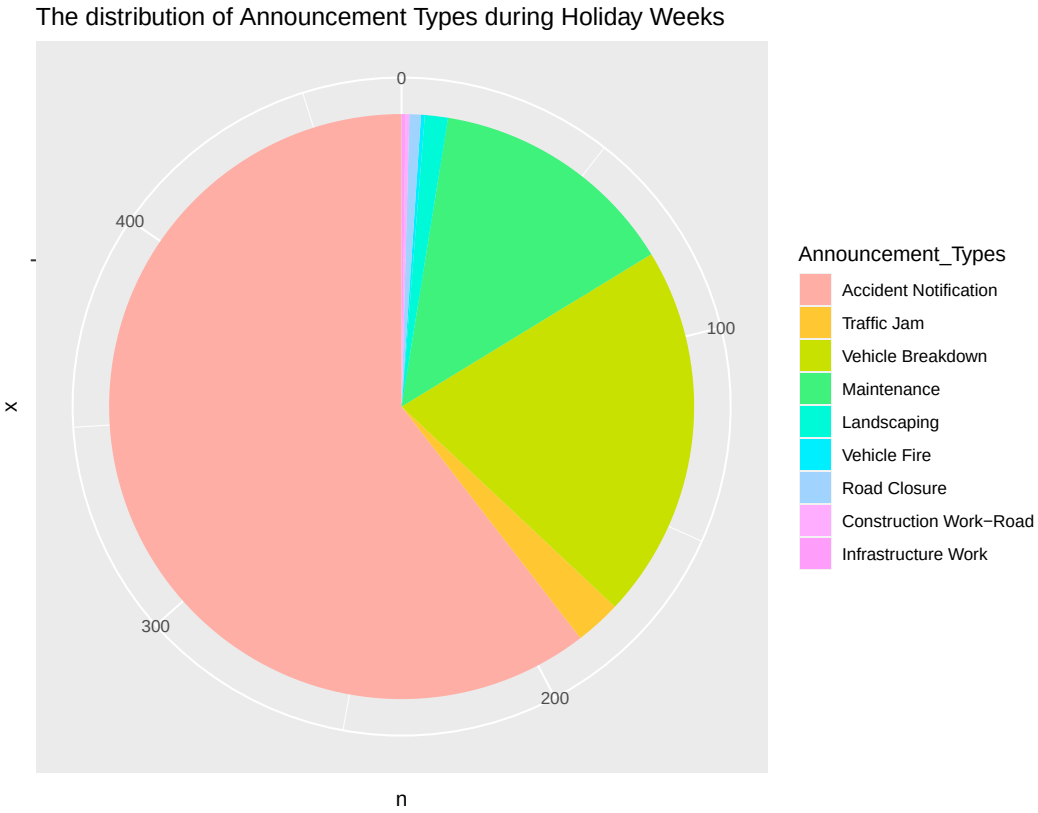
In the final section, official holiday events are introduced to the data set in order to define holiday periods in the data. For instance, 'New Year', 'Labor Day', 'Ramadan', 'National Sovereignty and Children's Day' and 'May 19 Commemoration of Atatürk and Youth and Sports Day'. In the data set, weeks that are includes any of the holiday event are taken into consideration. Announcements that are reported during the holiday week are counted. Then, total number of announcements are shown on the pie chart in order to interpret the most common announcement reasons in the holiday term. The external data source is added for holiday events.

Date-related recording errors are kept on the subset.

```
holidayCalendar <- read_excel("CalendarEvents.xlsx")
P4 <- DS %>%
mutate(weekStarting_Time = lubridate::week(DS$Starting_Time), yearDS =
lubridate::year(DS$Starting_Time))
P4_1 <- mutate(P4, holidayWeek =
(P4$weekStarting_Time==holidayCalendar$WeekHolidayStartDate &
P4$yearDS==holidayCalendar$Year |
P4$weekStarting_Time==holidayCalendar$WeekHolidayEndDate &
P4$yearDS==holidayCalendar$Year)) %>%
filter(Effected_Lanes >= 0 & holidayWeek==TRUE) %>%
group_by(Announcement_Types) %>% count(Announcement_Types)
```

The distribution of Announcement Types during holiday weeks are shown in the below pie chart.

```
ggplot(P4_1,aes(x="", y=n, fill=Announcement_Types)) +geom_bar(stat='identity',width =1)+
coord_polar("y")+scale_fill_hue(l=85)+
labs(title = "The distribution of Announcement Types during Holiday Weeks")
```



- In the holiday periods, majority of traffic announcements are occurred by the accident notification records. Presumably, increased intercity travels are caused to accidents during holiday week.
- Similarly, vehicle breakdown is affected by heavy travel demand and has got the second highest percentage among announcement types.
- Remarkable results belong to the number of Maintenance. Municipality may prefer to complete road works when the city is relatively empty because traffic on the road will be less than the regular term. Under this circumstances, working on road maintenance will be more productive.

References

- Data Source - İBB Open Data Portal
- External Data Source - Holiday Events
- Mapping Link

You may click here *to reach other items of A.K.A - R Group's Progress Journal.*