EDA: 911 Calls

Code

Cho Sungin, Jang Yoonseo

Nov 9, 2016

- File : `911.csv`
- Used package: `dplyr` , `tidyr` , `xts` , `lubridate` , `qtlcharts` , `forecast` , `tseries` , `leaflet` , `ggplot2` , `plotly` , `dygraphs` , `viridis` , `graphics`

---

1 Introduction

This is an exploratory analysis for data collected from **911 calls in Montgomery County**.

Members of rescue team always risk many people’s lives while they are working. Their work requires great deal of concentration for a very long time, as one little mistake can result in death. Hence, sufficient time for rest and relaxation is extremely important. However, a certain number of rescue workers on call are always necessary because it is impossible to know what kind of accident would happen in the future. We thought, if we can predict and forecast a general number of accidents in the future, it would be easier to allocate proper number of rescue workers at the right time. Which may also lead to a proper amount of rest time for them.

The source of the `911.csv` data is “http://montcoalert.org/ (http://montcoalert.org/)”. This page provides information of 911 calls in Montgomery County. Montgomery County is located in Commonwealth of Pennsylvania. The data are collected from Dec 10, 2015 to Oct 25, 2016.

- These are the index of our script:
  - Brief Explanation on “Accident Distribution”, by **Mapping**
  - 911 Calls in Montgomery County, **Generally**
    - **Percentage** of Types 911 Calls
    - How the number of 911 Calls varies as the **time goes by**
    - **Correlation** among Types of 911 Calls
    - **911 Call Trend** by Month, Day, Hour
    - Summarized Result, by **Heat Map**
  - 911 Calls in Montgomery County, by **Townships**
    - **Mosaic Plot** by Types of 911 Calls
    - **Mosaic plot** by Subtypes of 911 Calls
    - Summarized Result, by **Heat Map for Each Townships**
  - **Forecasting** the number of accidents in the future

1.1 Loading packages and Attributes of data

Now that our packages are loaded, let’s read in and check the attributes of data.

Code

Code

lat	lng	desc	zip	title	timeStamp	twp	addr	e

lat	lng	desc	zip	title	timeStamp	twp	addr	e
40.29788	-75.58129	REINDEER CT & DEAD END; NEW HANOVER; Station 332; 2015-12-10 @ 17:10:52;	19525	EMS: BACK PAINS/INJURY	2015-12-10 17:40:00	NEW HANOVER	REINDEER CT & DEAD END	1
40.25806	-75.26468	BRIAR PATH & WHITEMARSH LN; HATFIELD TOWNSHIP; Station 345; 2015-12-10 @ 17:29:21;	19446	EMS: DIABETIC EMERGENCY	2015-12-10 17:40:00	HATFIELD TOWNSHIP	BRIAR PATH & WHITEMARSH LN	1
40.12118	-75.35198	HAWS AVE; NORRISTOWN; 2015-12-10 @ 14:39:21-Station:STA27;	19401	Fire: GAS-ODOR/LEAK	2015-12-10 17:40:00	NORRISTOWN	HAWS AVE	1

Code

```
## 'data.frame':    123884 obs. of  9 variables:
##  $ lat      : num  40.3 40.3 40.1 40.1 40.3 ...
##  $ lng      : num  -75.6 -75.3 -75.4 -75.3 -75.6 ...
##  $ desc     : Factor w/ 123847 levels " ; ; 2016-03-09 @ 05:15:47;",..: 87476 13086 47179 3373 19732 167
89 58508 22429 64201 12311 ...
##  $ zip      : int  19525 19446 19401 19401 NA 19446 19044 19426 19438 19462 ...
##  $ title    : Factor w/ 117 levels "EMS: ABDOMINAL PAINS",..: 10 21 88 16 23 38 46 54 62 115 ...
##  $ timeStamp: Factor w/ 91782 levels "2015-12-10 17:40:00",..: 1 1 1 2 2 2 2 2 2 2 ...
##  $ twp      : Factor w/ 69 levels "","ABINGTON",..: 36 20 37 37 30 23 21 49 32 42 ...
##  $ addr     : Factor w/ 24150 levels "",".","10TH AVE",..: 17261 2302 9195 569 3713 3088 11374 4272 1239
2 2084 ...
##  $ e        : int  1 1 1 1 1 1 1 1 1 1 ...
```

These are the names, class type of variables. We can also check first few observations. In total, there are 123884 observations and 9 variables. Simple description of the variables is as follows. :

Variable Name	Description
lat	Latitude
lng	Longitude
desc	Description of the Emergency Call (EMS: Emergency Medical Service, Fire: Fire Accident, Traffic: Traffic Accident)
zip	Zipcode
title	Title
timeStamp	YYYY-MM-DD HH:MM:SS
twp	Township
addr	Address
e	Dummy variable (always 1)

1.2 Handling Data

We had to handle the data. We created some new variables from existing variables and removed variables that we did not need.

Code

lat	lng	desc	zip	Types	Subtypes	timeStamp	twp	addr	Date	Year	Mo
40.29788	-75.58129	REINDEER CT & DEAD END; NEW HANOVER; Station 332; 2015-12-10 @ 17:10:52;	19525	EMS	BACK PAINS/INJURY	2015-12-10 17:40:00	NEW HANOVER	REINDEER CT & DEAD END	2015-12-10	2015	12

lat	lng	desc	zip	Types	Subtypes	timeStamp	twp	addr	Date	Year	Mo
40.25806	-75.26468	BRIAR PATH & WHITEMARSH LN; HATFIELD TOWNSHIP; Station 345; 2015-12-10 @ 17:29:21;	19446	EMS	DIABETIC EMERGENCY	2015-12-10 17:40:00	HATFIELD TOWNSHIP	BRIAR PATH & WHITEMARSH LN	2015-12-10	2015	12
40.12118	-75.35198	HAWS AVE; NORRISTOWN; 2015-12-10 @ 14:39:21-Station:STA27;	19401	Fire	GAS-ODOR/LEAK	2015-12-10 17:40:00	NORRISTOWN	HAWS AVE	2015-12-10	2015	12

Code

```
## 'data.frame':    123884 obs. of  14 variables:
##  $ lat      : num  40.3 40.3 40.1 40.1 40.3 ...
##  $ lng      : num  -75.6 -75.3 -75.4 -75.3 -75.6 ...
##  $ desc     : Factor w/ 123847 levels " ; ; 2016-03-09 @ 05:15:47;",..: 87476 13086 47179 3373 19732 167
89 58508 22429 64201 12311 ...
##  $ zip      : Factor w/ 105 levels "17752","18036",..: 103 83 70 70 NA 83 40 76 79 88 ...
##  $ Types    : Factor w/ 3 levels "EMS","Fire","Traffic": 1 1 2 1 1 1 1 1 1 3 ...
##  $ Subtypes : Factor w/ 84 levels " ABDOMINAL PAINS",..: 10 22 38 16 25 42 50 60 69 78 ...
##  $ timeStamp: POSIXlt, format: "2015-12-10 17:40:00" "2015-12-10 17:40:00" ...
##  $ twp      : Factor w/ 69 levels "","ABINGTON",..: 36 20 37 37 30 23 21 49 32 42 ...
##  $ addr     : Factor w/ 24150 levels "","."10TH AVE",..: 17261 2302 9195 569 3713 3088 11374 4272 1239
2 2084 ...
##  $ Date     : Date, format: "2015-12-10" "2015-12-10" ...
##  $ Year     : Factor w/ 2 levels "2015","2016": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Month    : Factor w/ 11 levels "1","2","3","4",..: 11 11 11 11 11 11 11 11 11 11 ...
##  $ Day      : Factor w/ 31 levels "1","2","3","4",..: 10 10 10 10 10 10 10 10 10 10 ...
##  $ Hour     : Factor w/ 24 levels "0","1","2","3",..: 18 18 18 18 18 18 18 18 18 18 ...
```

2 Exploratory Data Analysis

2.1 Brief Explanation on "Accident Distribution"

Code

- As we explained above, data are lacated in **Montgomery County, PA** .

- Most of the 911 calls are made in **Norristown(36,264)** , which is followed by **Abington(17,911)** and **Lansdale(17,504)** .

2.2 911 Calls in Montgomery County, Generally

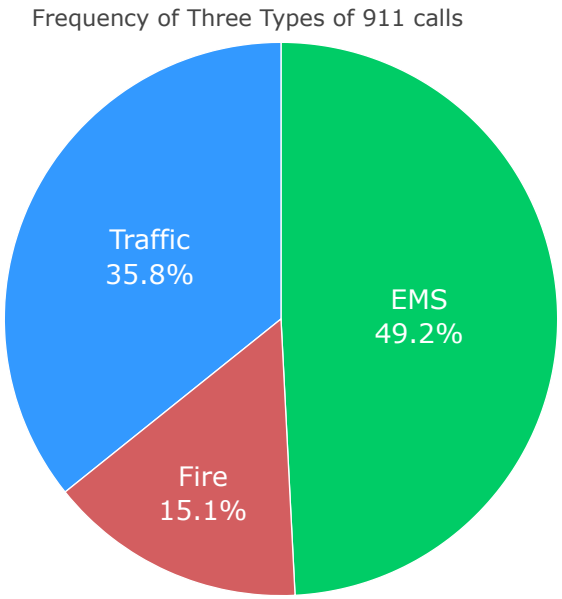
2.2.1 Percentage of Types of 911 Calls

Different types of accidents require different types of experts. We checked the percentage of each three types of calls by pie chart.

Code

```
##      Var1  Freq
## 1     EMS 60939
## 2    Fire 18651
## 3 Traffic 44294
```

Code

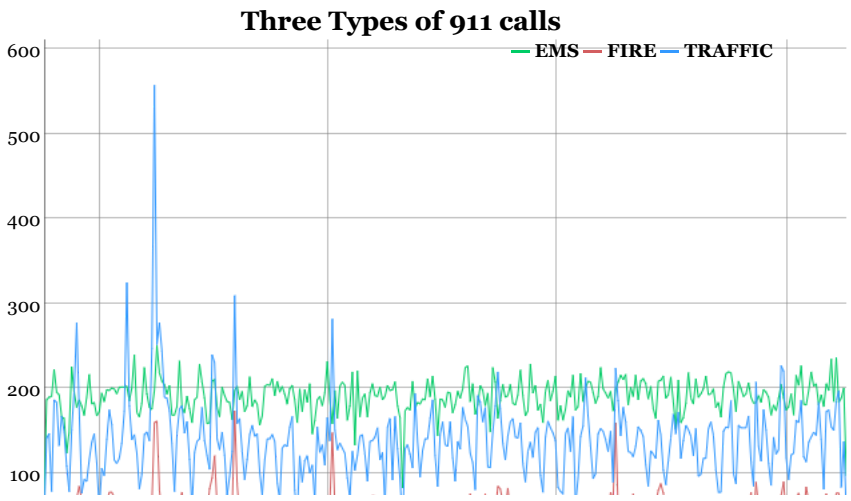


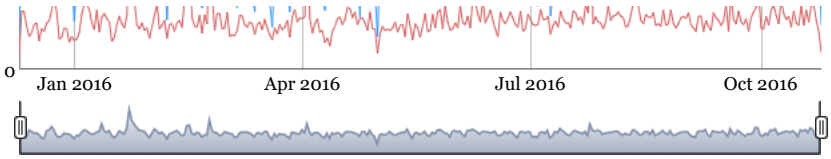
- **EMS** has accounted for **49.2%** and is followed by **Traffic(35.8%)** and **Fire(15.1%)** .
- EMS showed the largest number. Perhaps this is because people gets hurt in most of the accident, regardless of its types.

2.2.2 How the number of 911 Calls varies as the time goes by (for each types)

We were curious whether number of calls for EMS would be higher than the other two variables, even within a certain period of time. We used the time series graph to figure it out.

Code



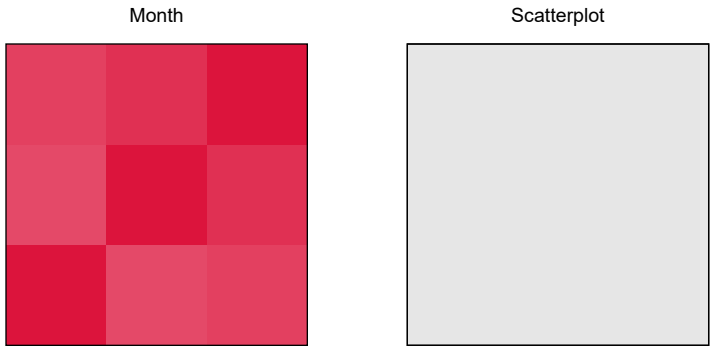


- Overall, number of calls for EMS is higher than the others.
- Number of calls for Traffic, rarely have exceeded the number of calls for EMS.
- Especially, number of calls for traffic in **"23, Jan"** was significantly high.
- Numbers of each three types of 911 calls seem correlated.
- The graph did not show any trend.

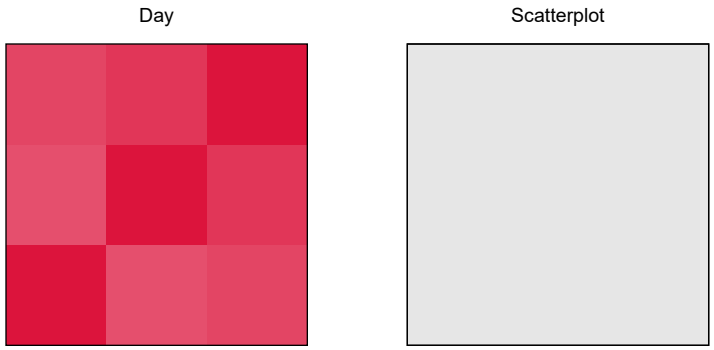
2.2.3 Correlation among Types of 911 Calls

In order to see correlations between three types, we made a correlation matrix. By **mousing over** on a square box, we can see a **correlation coefficient** and by **clicking** it, we can see a **scatter plot**.

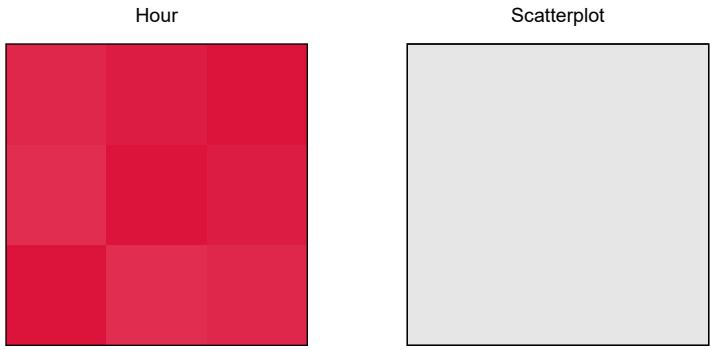
Code



Code



Code

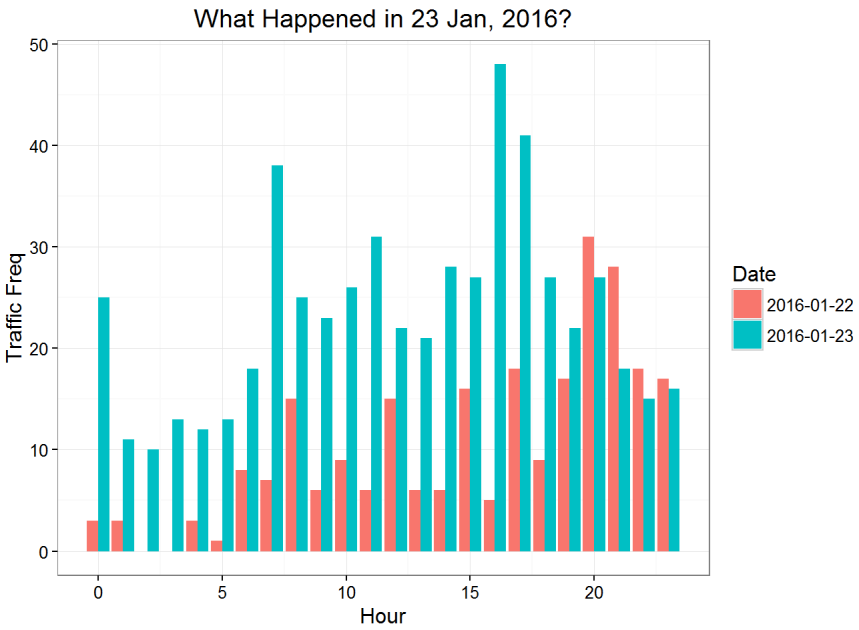


- All types of 911 calls showed quite large correlation coefficient regardless of month, day, and hour.
- We can assume that all types of accidents are **mutually correlated** .

2.2.3.1 What happened in 2016-01-23?

Big accident might have occured in 23, Jan. We compared the number of accidents happened in 22nd and 23rd by hours. If there were a critical accident in 23rd, then a number of 911 calls for traffic accidents at a specific time would be significantly large.

Code

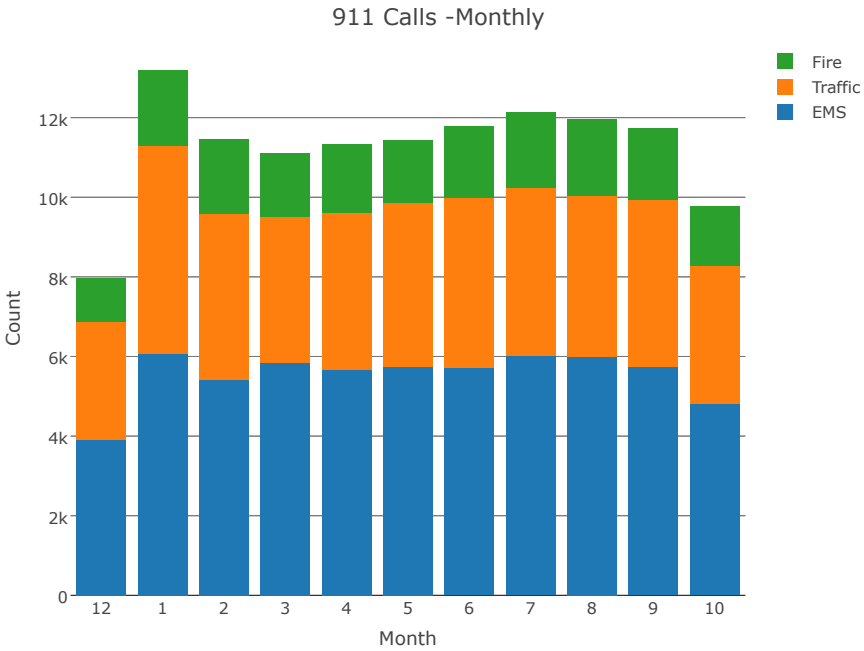


- The number of 911 calls for traffic in 23rd is generally larger than 22nd.
- It seems like there were no critical accident on that day.

2.2.4 911 Call Trend by Month, Day, Hour

2.2.4.1 Calls - Monthly

Code

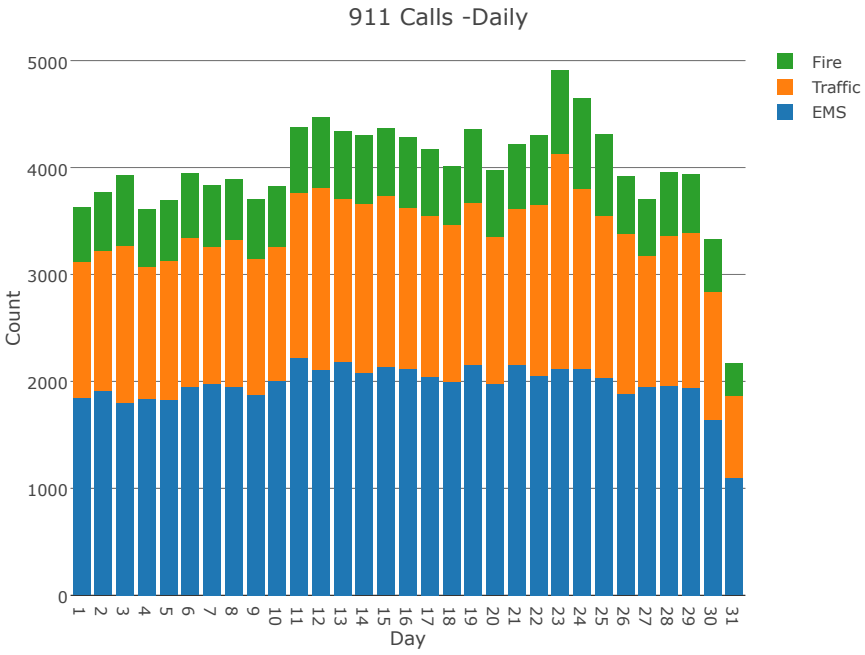


- Frequency of 911 calls are lower in December and October because there are less data in those months (Oct : 25, Dec : 22) and frequency of 911 calls in January is slightly higher than others as there were significantly large number of calls on 23

January.
- Except for these months, frequency of 911 calls of the months are very alike.

2.2.4.2 Calls - Daily

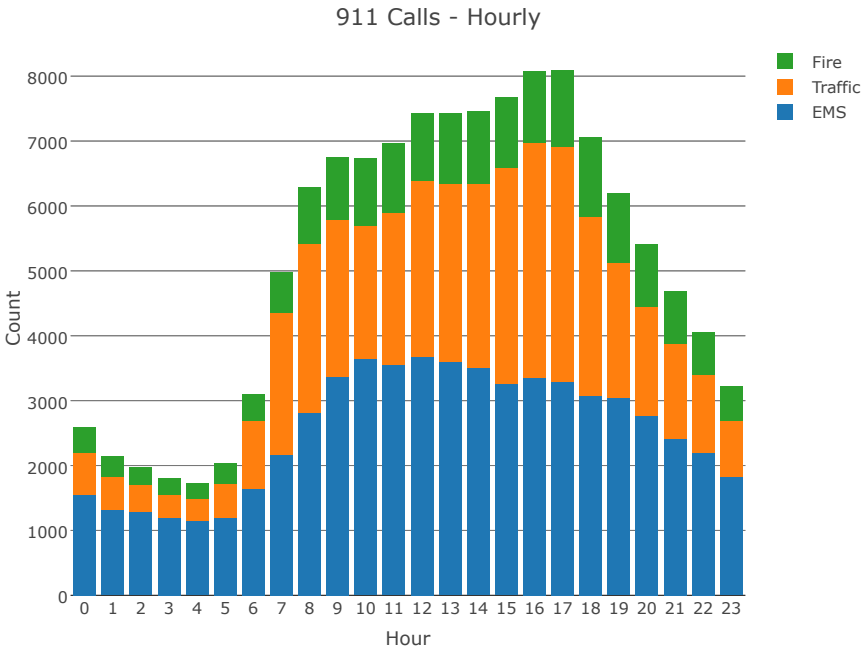
Code



- Frequency of 911 calls are low from 1st to 9th and from 26th to 31st. Frequency of 911 calls in 31st is significantly low. This may be due to less data in October, December, and months that do not consist of 31 days. Furthermore, frequency of 911 calls is very high on 23rd as there were significantly large number of 911 calls for traffic on 23 January.
- Considering these factors, no trend is seen, when we look at the frequency of the 911 calls of the days.

2.2.4.3 Calls - Hourly

Code



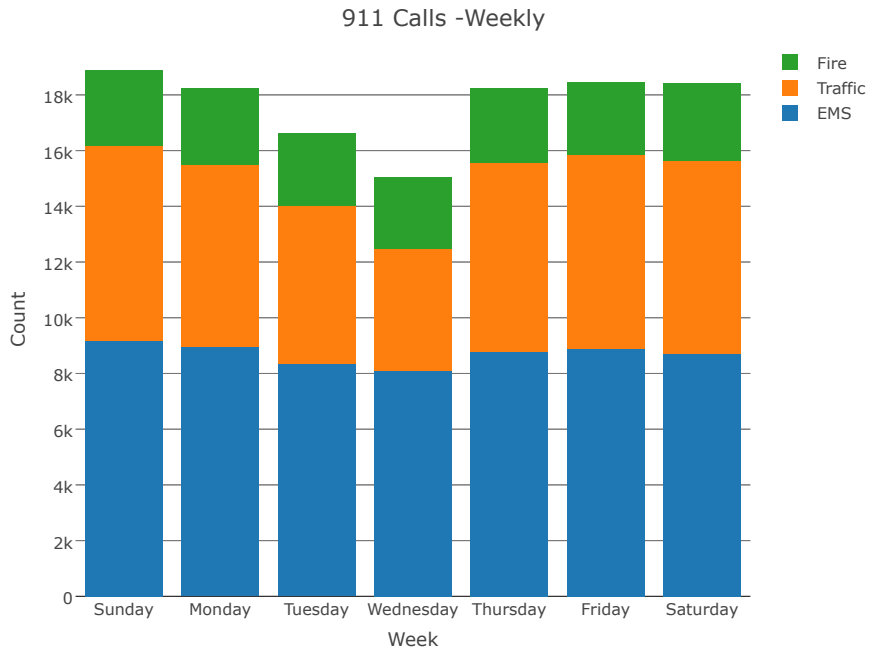
- Frequency of 911 calls are certainly higher from 7am to 9pm.
- People lead their lives in the daytime. Therefore, accidents are likely to occur more often in this time-period.
- It seems like the frequency of 911 calls differs by hours rather than month and days.

2.2.4.4 Calls - Weekly

Code

```
##  Factor w/ 7 levels "Sunday","Monday",..: 5 5 5 5 5 5 5 5 5 5 ...
```

Code



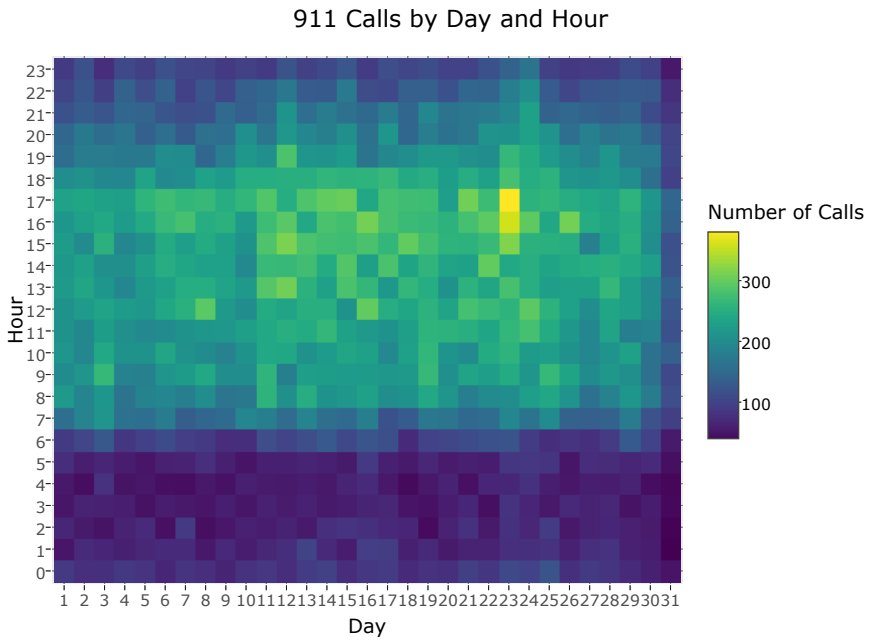
- As you can see, during weekdays, frequency of total 911 calls and their ratio of types are all alike.
- However, frequency of total 911 calls on weekend is lower than that is on weekdays.
- We can see that the numbers of 911 calls for fire and EMS on weekend are almost identical with the numbers of 911 calls for fire an EMS on weekdays.
- Only the frequency of 911 calls for traffic showed the difference. It decreased from about 7,000 calls to about 5,000 calls.
- It seems reasonable to assume that people have less traffic accident on weekend as they do not have to go to work.

2.2.5 Summarized Result, by Heat Map

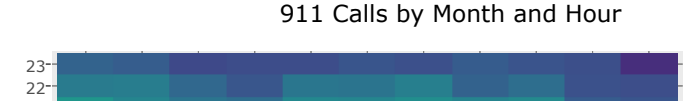
A heat map is a three-dimensional representation of data in which values are represented by colors. Let's see heatmaps of 911 calls by Month, Day and Hour. The number of 911 calls gets larger as a color of square goes closer to yellow.

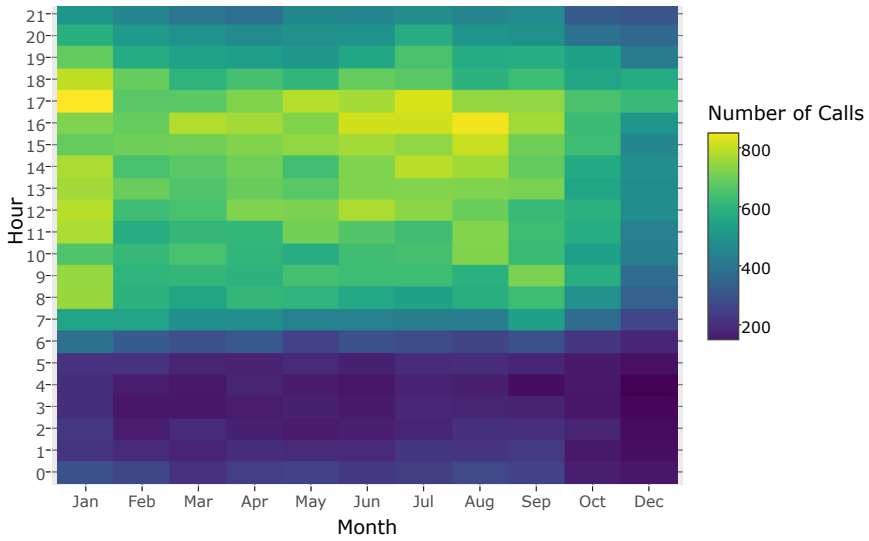
2.2.5.1 911 Calls by "Day and Hour" & "Month and Hour"

Code



Code



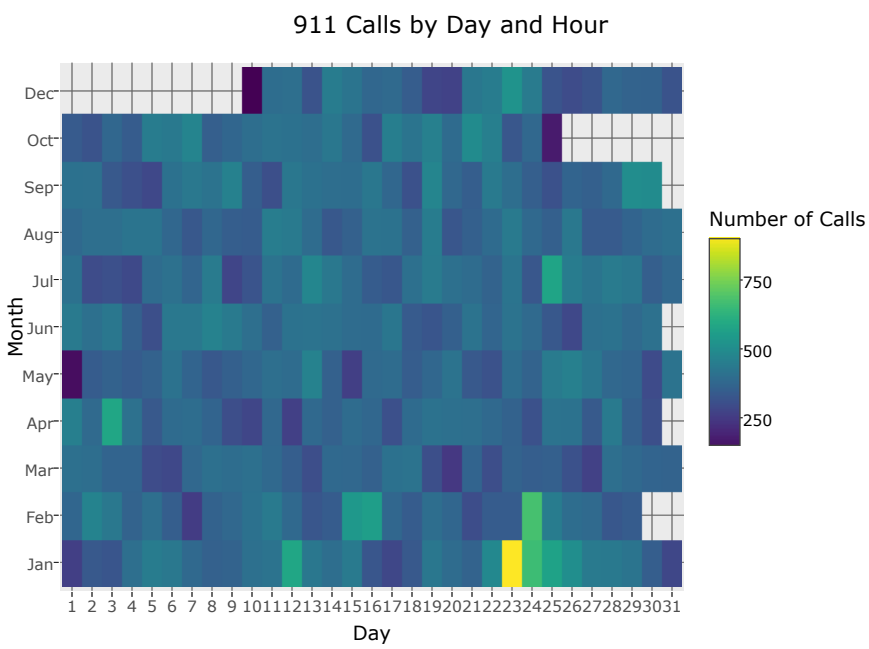


- By looking at the two heat maps above, we can assume that majority of the calls are during daytime, as most of the yellow squares are concentrated in the middle of the plot, horizontally.

2.2.5.2 911 Calls by Day and Month

In the following Heatmap, There are blanks in February, April, June, September because they do not consist of 31 days. Blanks in October and December are because there are less data in these months, as explained above.

Code

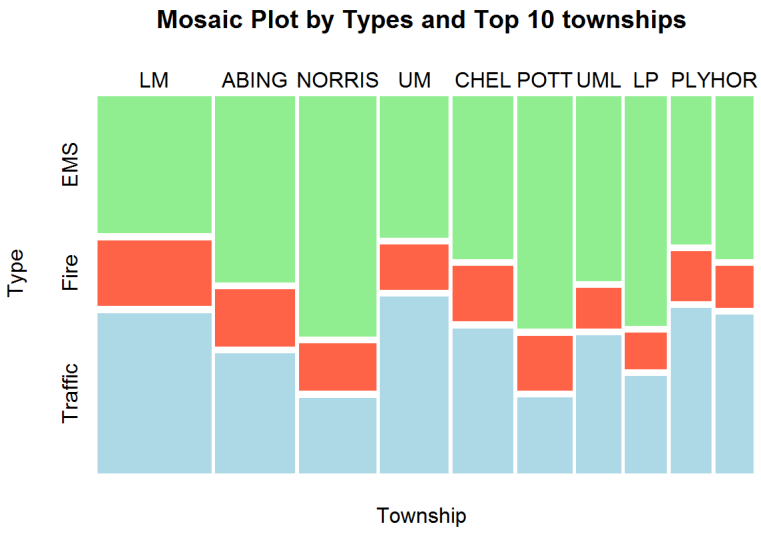


- No pattern can be seen in the heatmap of day and month.
- Rather than day and month, it is obvious that the numbers of 911 calls are mainly depended on the hours.

2.3 911 Calls in Montgomery County, by Townships

2.3.1 Mosaic Plot by Types of 911 Calls

Code



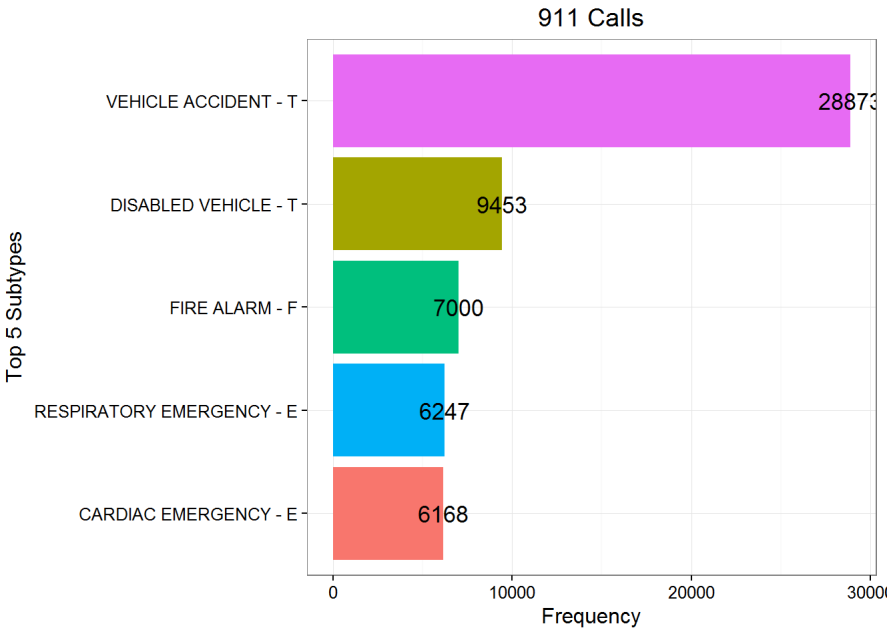
- Simple description of the abbreviated word is as follows. :
  - `LM` : LOWER MERION
  - `ABING` : ABINGTON
  - `NORRIS` : NORRISTOWN
  - `UM` : UPPER MERION
  - `CHEL` : CHELTENHAM
  - `POTT` : POTTSTOWN
  - `UML` : UPPER MORELAND
  - `LP` : LOWER PROVIDENCE
  - `PLY` : PLYMOUTH
  - `HOR` : HORSHAM

- The above plot represents the ratio of types of 911 calls that top 10 townships have.
- Generally, EMS calls have a highest frequency followed by Traffic and Fire.
- However, their ratio shows difference among each townships.
- The county council should allocate rescue workers into each townships proportionally referencing this plot.

2.3.2 How does the Top 5 subtypes vary among top 10 township?

2.3.2.1 Top 5 Subtypes

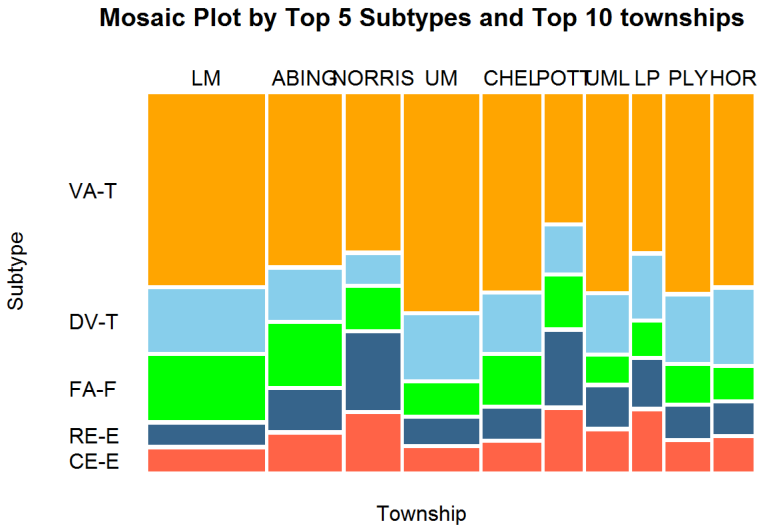
Code



- These are the top 5 subtypes that have largest frequency.
- Subtypes related to vehicles are ranked in 1st and 2nd.
- Subtypes from EMS are even lower than subtype from Fire. Subtypes of EMS were evenly distributed when we checked the data. We thought that it might be the reason.

2.3.2.2 Mosaic plot by Subtypes of 911 Calls

Code



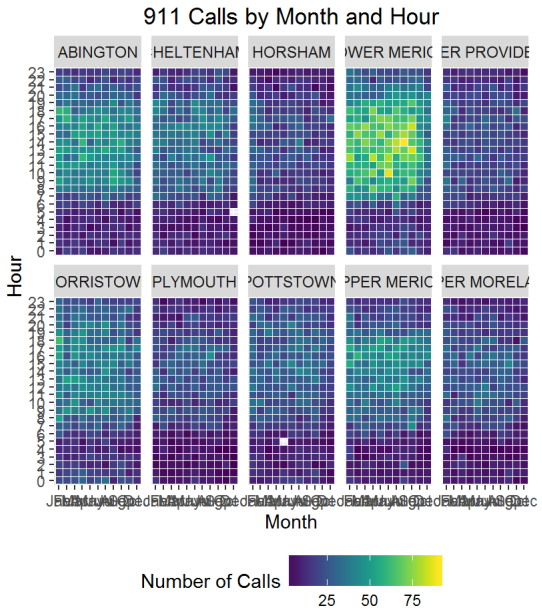
- Simple description of the abbreviated word is as follows. :
  - `VA-T` : VEHICLE ACCIDENT (Traffic)
  - `DV-T` : DISABLED VEHICLE (Traffic)
  - `FA-F` : FIRE ALARM (Fire)
  - `RE-E` : RESPIRATORY EMERGENCY (EMS)
  - `CE-E` : CARDIAC EMERGENCY (EMS)

- Frequency of 911 calls for Traffic is highest when you look at the plot.
- Norristown and Pottstown showed relatively low ratio for 911 calls for Traffic.
- Pottstown is located at very edge of the county. Hence, low traffic accidents may be reasoned.
- However, Norristown is located at the center of county and has pretty large number of a floating population. It was unexpected result.

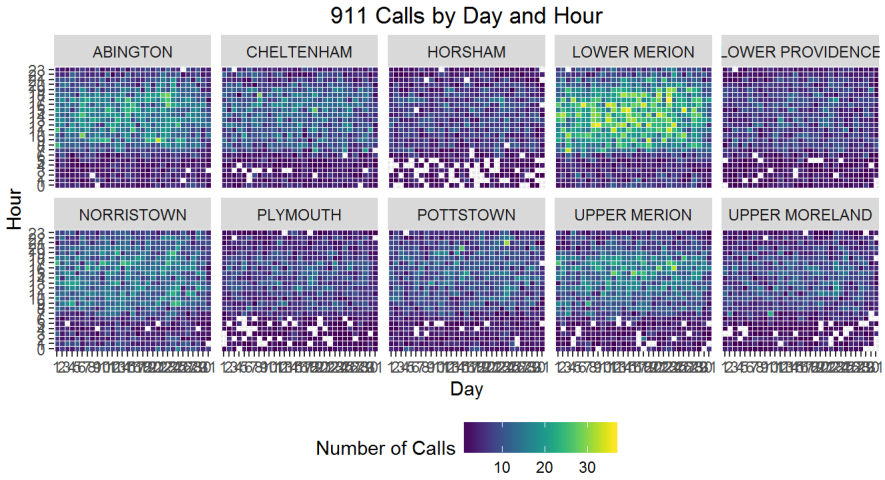
2.3.3 Summarized Result, by Heat Map for Each Townships

2.3.3.1 911 Calls by "Month and Hour" & "Day and Hour" (without transformation)

Code



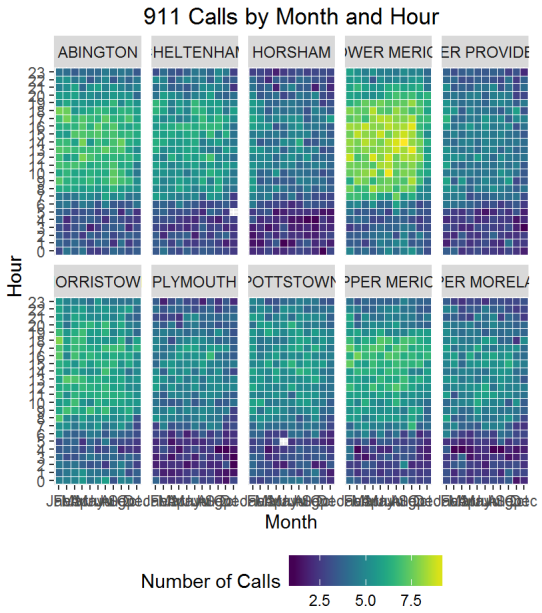
Code



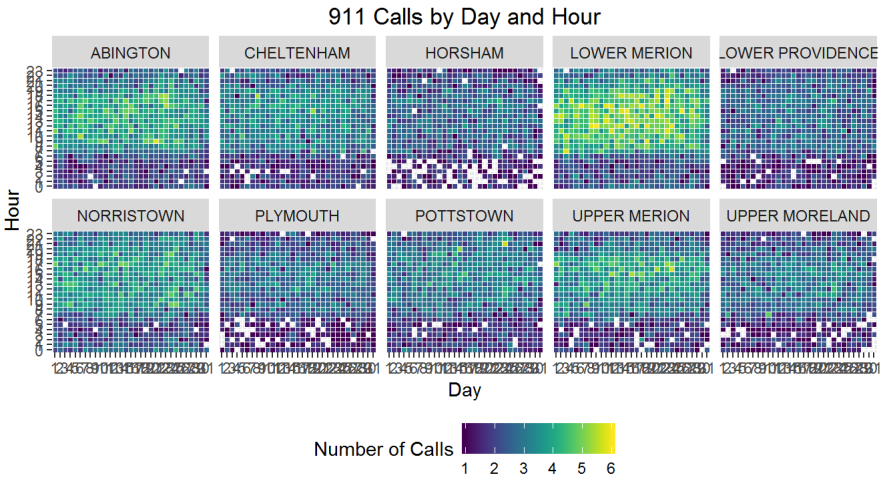
- Frequency of total 911 calls is relatively too high in Lower Merion. Therefore, other township's heat map were too dark in overall plot.
- Pattern could not be seen distinctively.

2.3.3.2 911 Calls by "Month and Hour" & "Day and Hour" (with square-rooted frequency)

Code



Code



- To lower the differences between frequencies of 911 calls among townships, we square rooted the frequencies of each townships.
- We could see an explicit pattern. All heat maps were more deeply coloured in yellow at the center, regardless of the townships.
- When we look at the heat map drawn for each townships, frequency of 911 calls differs mostly by hours as well.

2.4 Forecasting the number of accidents in the future

Now let's forecast the number of accidents that might happen in the future. Time-series analysis is used for analyzation.

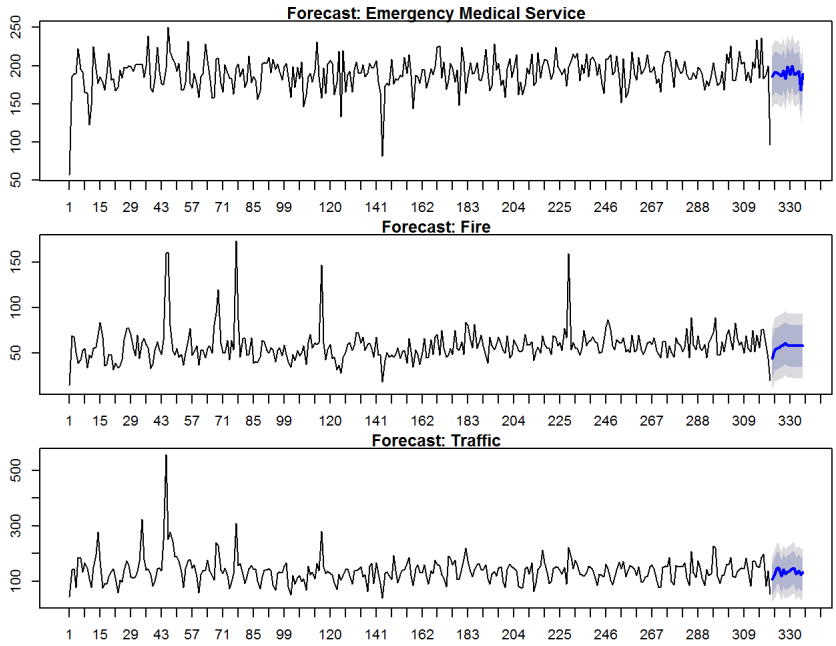
First, we computed **autocorrelation function and partial autocorrelation function** to check briefly whether the data are stationary time-series.

Second, we checked whether differences are necessary by using **ndiffs** function.

Third, we did Dickey-Fuller test for null-hypothesis, data are non-stationary. We could reject null-hypothesis.

Lastly, we tried to find the best ARIMA model by their AIC, AICc and BIC values.

Code



Code

Date	EMS	Fire	Traffic
2016-10-26	185.46	44.36	106.18
2016-10-27	191.32	54.11	126.79
2016-10-28	191.29	55.05	147.67
2016-10-29	189.47	56.07	147.29
2016-10-30	186.79	57.90	118.06
2016-10-31	193.85	59.03	141.82
2016-11-01	183.12	61.27	125.71

- These are the forecasted number of 911 calls of three types of data for 7 days(1 week), respectively.
- Honestly, the result of forecasting is very digressed from our early purpose. We wanted to use forecasted result to allocate rescue workers properly. However, this forecast is not sufficient to do that.
- What we analyzed were time-series data, collected daily. Also, it is data about accidents. Although, it satisfied the non-stationary, there were no patterns and there were unexplainable fluctuation in the data.
- Hence, we thought time-series analysis is meaningless for analyzing this data.

3 Conclusion

- It's impossible to predict frequency of 911 calls using this data.
- We should do Baysian analysis to predict the accidents.
- We ought to collect more variables that are related closely to the occurence of each types of accidents(such as floating population, weather, and etc.) for Bayesian analysis. This data is not enough.
- By analyzing this data, we were highly surprised once again by rescue workers' devotion to the society. Most of the work has its busy season and off-season. We figured out that ,for rescue workers, every day was busy season. Even at night, some of them have to be on night duty. I want to claim that, we must have a deep respect for them. Furthermore, the state should provide them with high quality of welfare and benefit.
- All types of accidents are laregly correlated with each others. What we have to focus on, is putting our effort in preventing accidents from happening, as it may result in more additional accidents.

```
* Writer: Cho Sungin, Jang Yoonseo
* Creation date: Nov 9, 2016
```