אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

Data Visualization project

"UK Car Accidents 2005-2015"

Aviv Ohana

040819237

2/23/2019

Dr Bak, Peter

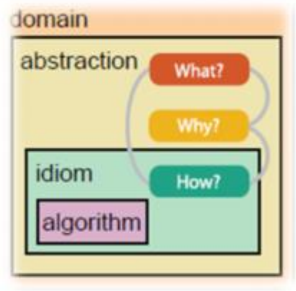
List of content:

1. Domain

The domain is filed of road safety. The data I have is about car accidents in the UK between the years of 2005 and 2015.

I will use a visualization in order to better communicate and analyze this huge amount of data. Visualization will "tell a story" of data , allowing to reveal facts and insights out of the raw data in a way that is clear and available to the end user.

I used the model we learned in the class – in order to build the visualization



2. Data and Task

What?

I am using a dataset taken from Kaggle web site that contains all the accidents reported in the UK during 2005 and 2015. UK police forces collect data on every vehicle collision in the UK on a form called Stats19. Data from this form ends up at the DfT and is published at https://data.gov.uk/dataset/road-accidents-safety-data .

The is tabular and extracted from relational DB, The Kaggle data contained 3 major CSV –

- Accidents – data about the accident such as date, location etc.
- Vehicles – data about the vehicle\s involve in the accident like vehicle type, junction type, sex of driver etc.
- Casualty – data about the casualties in the accident like sex, age, casualty severity etc.

Kaggle also provide Zip file containing context data with 29 files. I found that many columns were still missing the context data. I searched the UK government web site and found all the context data that was missing.

The files contained data of 1,780,653 accidents, 3,520,115 vehicles and 2,589,098 casualties!

Table 1 – **Accidents**

#	Attributes	Type	Description
1	Accident Index	Number	PK to identify record
2	Police Force	Categorial	The police station that handle the accident
3	Accident Severity	Categorial	Fatal, Serious, Slight
4	Number of Vehicles	Number	Number of vehicles involved in the accident
5	Number of Casualties	Number	Number of casualties involved in the accident
6	Date (DD/MM/YYYY)	Continuous	
7	Day of Week	Categorial	Sun-Sat
8	Time (HH:MM)	Continuous	

9	Location Easting OSGR (Null if not known)	Continuous	
10	Location Northing OSGR (Null if not known)	Continuous	
11	Longitude (Null if not known)	Continuous	WGS84
12	Latitude (Null if not known)	Continuous	WGS84
13	Local Authority (District)	Categorial	On which authority accident happen at
14	Local Authority (Highway Authority - ONS code)	Categorial	
15	1st Road Class	Categorial	Based on the UK roads numbering (letter +1-4 number)
16	1st Road Number	Categorial	Based on the UK roads numbering (letter +1-4 number)
17	Road Type	Categorial	Roundabout, one way street, dual\single carriageway etc.
18	Speed limit	Continuous	
19	Junction Detail	Categorial	Type of the junction – T, Crossroad, etc.
20	Junction Control	Categorial	Traffic signal, stop sign, authorized person etc.
21	2nd Road Class	Categorial	Based on the UK roads numbering (letter +1-4 number)
22	2nd Road Number	Categorial	Based on the UK roads numbering (letter +1-4 number)
23	Pedestrian Crossing-Human Control	Categorial	The way pedestrian is controlled – school patrol, authorized person etc
24	Pedestrian Crossing-Physical Facilities	Categorial	Type of pedestrian – Zebra, footbridge, subway etc.
25	Light Conditions	Categorial	Daylight, Darkness – lights lit, unlit, no lightning
26	Weather Conditions	Categorial	Fine, raining snow, wind\no wind, fog or mist
27	Road Surface Conditions	Categorial	Dry, wet, snow, frost or ice, flood, oil, mud
28	Special Conditions at Site	Categorial	Same as above + define if auto traffic signal is out or defective, roadworks
29	Carriageway Hazards	Categorial	Indicates any interruption – object on road, dog on road, previous accident etc.
30	Urban or Rural Area	Categorial	Urban/Rural
31	Did Police Officer Attend Scene of Accident	Categorial	Yes/No
32	Lower Super Ouput Area of Accident_Location (England & Wales only)	Missing	

Table 2: **Vehicle**

#	Attributes	Type	Description
1	Accident Index	Number	PK to identify record
2	Vehicle Reference	Number	For multiple vehicles involve in certain accident
3	Vehicle Type	Categorial	Motorcycle, taxi, car, minibus etc.
4	Towing and Articulation	Categorial	None, Caravan, Single trailer, other tow etc.
5	Vehicle Manoeuvre	Categorial	What manoeuvre vehicle perform during the accident – reversing, parking, U-turn etc.
6	Vehicle Location-Restricted Lane	Categorial	Indicate if vehicle was on restricted location such as: bus lane, cycle lane, Tram etc.
7	Junction Location	Categorial	Approaching junction, leaving roundabout, leaving main road etc.
8	Skidding and Overturning	Categorial	Skidded, jackknifed, overturned etc.
9	Hit Object in Carriageway	Categorial	Kerb, road works, previous accident etc.
10	Vehicle Leaving Carriageway	Categorial	
11	Hit Object off Carriageway	Categorial	If vehicle hit some objects during the accident like Road sign, lamp post, tree, bus stop etc
12	1st Point of Impact	Categorial	First point of impact to the vehicle involved – front, back, offside, nearside, did not impact
13	Was Vehicle Left Hand Drive	Categorial	1- No, 2- Yes, -1 – missing or out of range
14	Journey Purpose of Driver	Categorial	Part of work, commuting from\to work, pupil riding from\to school etc.

15	Sex of Driver	Categorial	1- Male , 2- Female, 3-Not known, -1 – data missing or out of range
16	Age of Driver	Number	
17	Age Band of Driver	Categorial	Data was splited to range of ages into 11 groups, 0-5→1, 6-10->2… over 75 ->11
18	Engine Capacity	Number	Engine capacity in CC
19	Vehicle Propulsion Code	Categorial	What fuel runs the engine – petrol, electric, steam, gas etc.
20	Age of Vehicle (manufacture)	Number	
21	Driver IMD Decile	Categorial	The English Index of Multiple Deprivation
22	Driver Home Area Type	Categorial	1 – Urban area, 2- Small town, 3- Rural, -1- missing or out of range

Table 3: **Casualties**

#	Attributes	Type	Description
1	Accident Index	Number	PK to identify record
2	Vehicle Reference	Number	For multiple vehicles involve in certain accident
3	Casualty Reference	Number	For multiple casualties involve in certain accident
4	Casualty Class	Categorial	Indicate if casualty is: 1 - Driver or rider, 2- Passenger, 3- Pedestrian
5	Sex of Casualty	Categorial	1- Male , 2- Female, -1 – data missing or out of range
6	Age of Casualty	Number	
7	Age Band of Casualty	Categorial	Data was split to range of ages into 11 groups, 0-5→1, 6-10->2… over 75 ->11
8	Casualty Severity	Categorial	Casualty severity of injury: 1 – Fatal, 2 – Serious, 3- Slight
9	Pedestrian Location		
10	Pedestrian Movement		
11	Car Passenger	Categorial	Indicate where passenger located in a car: 0- Not car passenger, 1 - Front seat passenger, 2- Rear seat passenger, -1 - Data missing or out of range
12	Bus or Coach Passenger		
13	Pedestrian Road Maintenance Worker (From 2011)		
14	Casualty Type	Categorial	See details below under table – *"Casualties per vehicle type"*
15	Casualty_Home_Area_Type		

After all the below analysis was done and after understanding my user tasks (see chapter:""), I decided to remove the columns I'm not going to need for my visualization (mark red on the above table)

Descriptive statistics

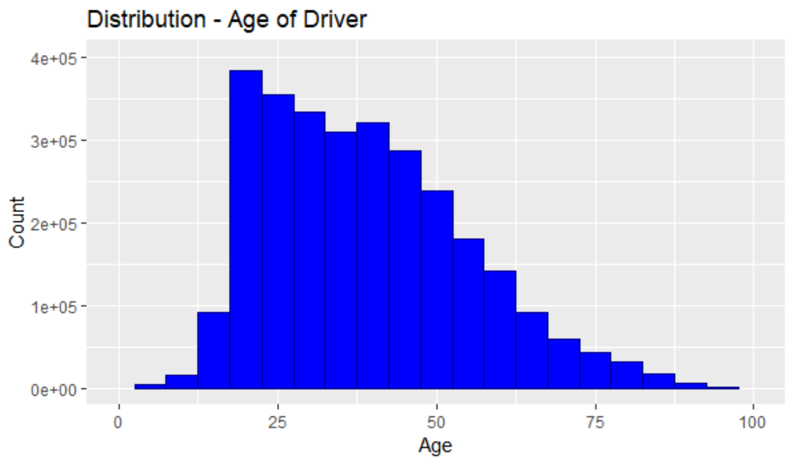
Driver and Vehicle analysis:

Age of Driver – checked the distribution of the raw data. Found that there are ages which are not relevant for a car permit - Remove all age<17 (age of license in the UK is 17), on second thought the file also contain all kind vehicle which kids are also able to use, so I decided to just get everything >0:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   1.00   26.00   37.00   38.81   49.00  100.00  257845 
```

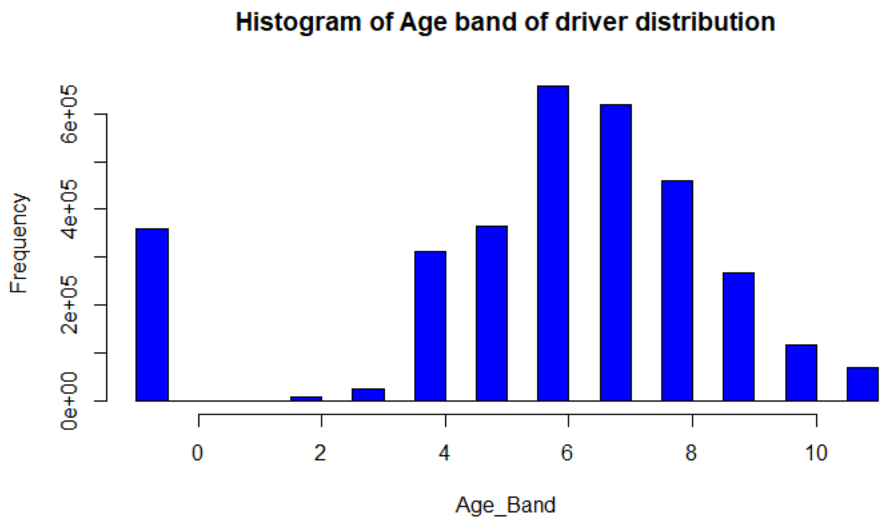
Then I built histogram of the age:

```
AgeHistogram <- ggplot(data = vehic, aes(x = Age_of_Driver)) +
  geom_histogram(fill ="blue",col="darkblue",binwidth =5)+
  labs(title="Distribution - Age of Driver")+
  labs(x="Age", y="Count")+
  xlim (c(0,100))+
  ylim(c(0,400000))
```



The DB also provide **age band** (Age_Band_of_Driver) of all driver by the following key:

code	label
1	0 - 5
2	6 - 10
3	11 - 15
4	16 - 20
5	21 - 25
6	26 - 35
7	36 - 45
8	46 - 55
9	56 - 65
10	66 - 75
11	Over 75
-1	Data missing or out of range



Gender of Driver – I analyzed the data about the gender of the **driver**. Raw data contained 4 categories 1-male, 2 – Female, 3 – Not known, -1 – missing data.

That was the raw data distribution –

```
 -1       1       2       3 
 52 2147401  924565  190252 
```

I cleaned out the -1 and 3 to get the exact numbering per gender: (only few were missing or unknown)

```
gender
      1       2 
2147401  924565 
```

Well, obviously men's are better drivers but involved in more accidents 😉(70% men's, 30% women's)

Analyzing the Vehicle_Type – I can see that safest way to commute in the UK will be the Tram. Most risky is obviously a car, but the surprise is that pedal bikes are next!

Used this R code to get these insights:

```
##Vehicle type
vic.type <- read.csv("./Vehicle_Type.csv", header = T)
vehic <-merge(vehic,vic.type, by.x="Vehicle_Type", by.y = "code", all.x=TRU
table(vehic$label)
vehic <- vehic[-c(1)] #drop un needed column
colnames(vehic)[colnames(vehic)=="label"] <- "Vehicle_Type" #Rename a colum

sort(table(vehic$Vehicle_Type))
```

I checked if there is any relation to the fact that in the UK driving on left lane and having cars that have right handed drive. I did not found any correlation.

```
> table(vehic$Was_Vehicle_Left_Hand_Drive.)

     -1        1        2
```

```
24068 3223341    14861
```

Casualties analysis

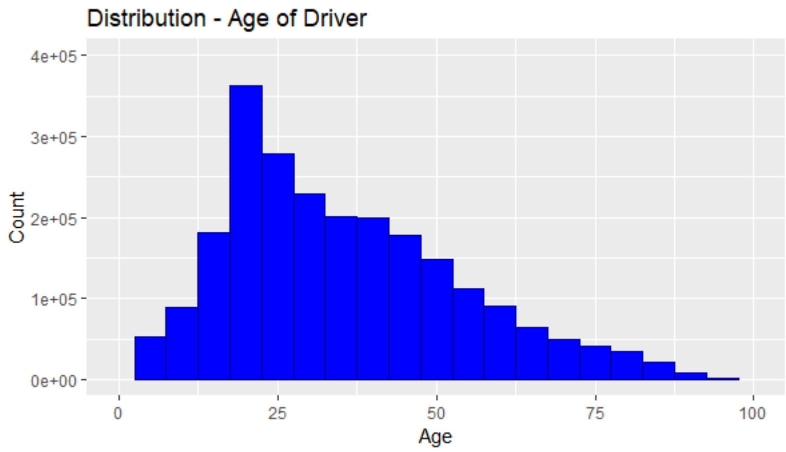
Age of casualty – checked the distribution and percentile of the age of casualty

```
> summary(Casualties$Age_of_Casualty)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  -1.00   20.00   31.00   34.49   47.00  104.00  186189 
```

Checking accurate percentile:

```
> quantile(Casualties$Age_of_Casualty, probs = c(0, 0.25, 0.5, 0.75, 1),na.rm = T
RUE)
  0%  25%  50%  75% 100% 
  -1   20   31   47  104 
```

```
#calculate the precntile of age casualty
quantile(Casualties$Age_of_Casualty, probs = c(0, 0.25, 0.5, 0.75, 1),na.rm
AgeHistogram <- ggplot(data = Casualties, aes(x = Age_of_Casualty)) +
  geom_histogram(fill ="blue",col="darkblue",binwidth =5)+
  labs(title="Distribution - Age of Driver")+
  labs(x="Age", y="Count")+
  xlim (c(0,100))+
  ylim(c(0,400000))
#scale_y_discrete(labels=c("20,000","100,000","400,000"))
AgeHistogram
```



It seems like have of casualties are under the age of 31, so like the age of driver youngers are also involved more as casualties. (I also checked the quantile of age of driver is almost exactly the same, 25%-22, 50%-34, 75%-47. 100%-100)

Gender of casualty

```
####Gender of casualty
table(Casualties$Sex_of_Casualty)
gender <-Casualties$Sex_of_Casualty
gender <-subset(gender, gender==1 | gender==2)
table(gender)
```

```
gender
      1       2
1402561  999657
```

Still more men's (60%) are involved also as casualty in car accident comparing to women's (40%).

I used this code the check casualties per vehicle type to create this summary table

Casualties per vehicle type:

Code	Type	Number of accidents
9	Car occupant	1,477,077
0	Pedestrian	299,926
1	Cyclist	197,373
5	Motorcycle over 500cc rider or passenger	92,652
3	Motorcycle 125cc and under rider or passenger	74,963
11	Bus or coach occupant (17 or more pass seats)	67,874
19	Van / Goods vehicle (3.5 tonnes mgw or under) occupant	54,575
2	Motorcycle 50cc and under rider or passenger	36,559
8	Taxi/Private hire car occupant	33,234

4	Motorcycle over 125cc and up to 500cc rider or passenger	26,412
21	Goods vehicle (7.5 tonnes mgw and over) occupant	12,961
90	Other vehicle occupant	11,103
10	Minibus (8 - 16 passenger seats) occupant	7,885
20	Goods vehicle (over 3.5t. and under 7.5t.) occupant	6,521
16	Horse rider	1,263
17	Agricultural vehicle occupant	1,250
22	Mobility scooter rider	521
97	Motorcycle - unknown cc rider or passenger	431
98	Goods vehicle (unknown weight) occupant	191
18	Tram occupant	111
23	Electric motorcycle rider or passenger	27

By this table we can see that obviously most of the casualties are from cars accident which make sense as they are most involved with accidents (see above). What supprised me is that pedestrians and cyclist are most vulnerable after car casualties! Also, if I sum up all motorcycles casualties together they are become third on list (after pedestrians).

Why?

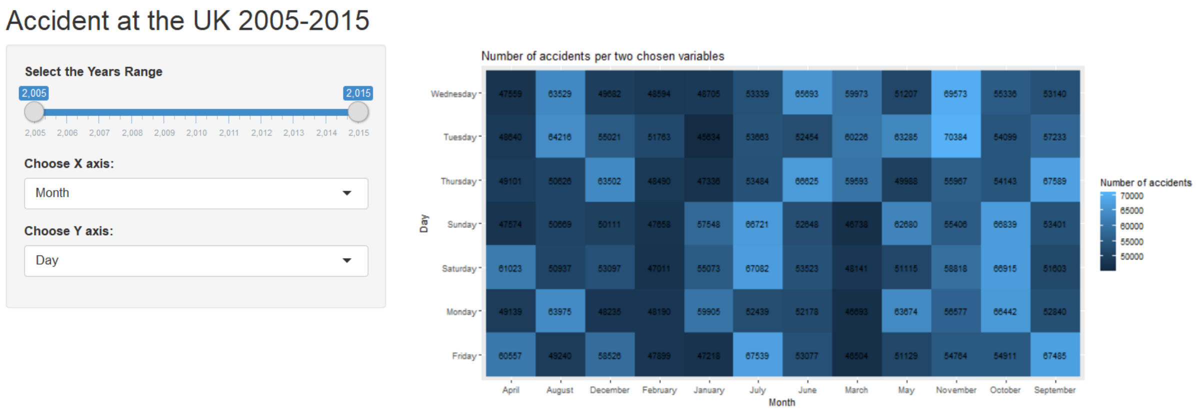
We would like to help the user to find correlation between the attributes (factors), using a visualization I can indicate if there is a strong correlation between the attributes. {discover, correlation)

The user tasks are: {action, target}

1. User will be able to find risky days of year by seeing a heatmap that represent the data.

   User selections are: X->"Month", Y->"Day" , data-> number of accidents per day and month

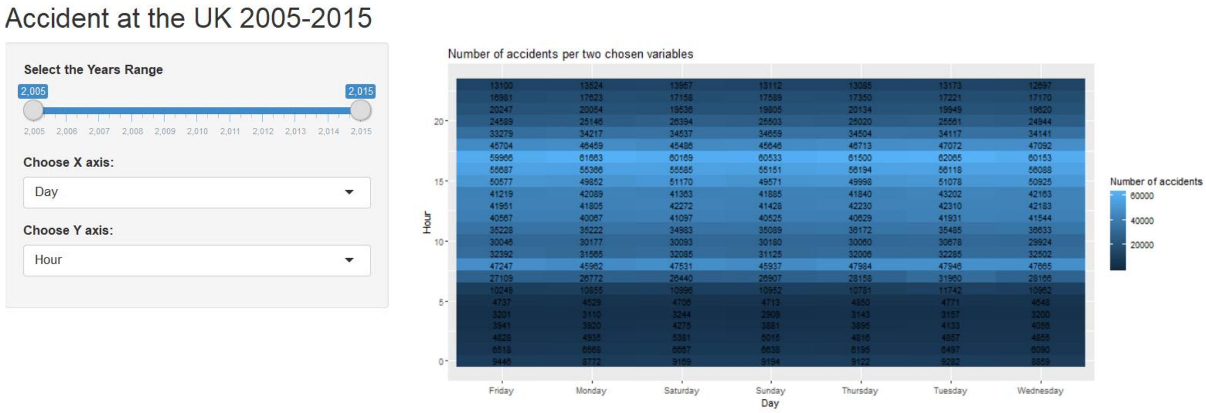
   Screenshot from the application:



2. User will be able to identify risky hours by seeing a heatmap that represent the data.

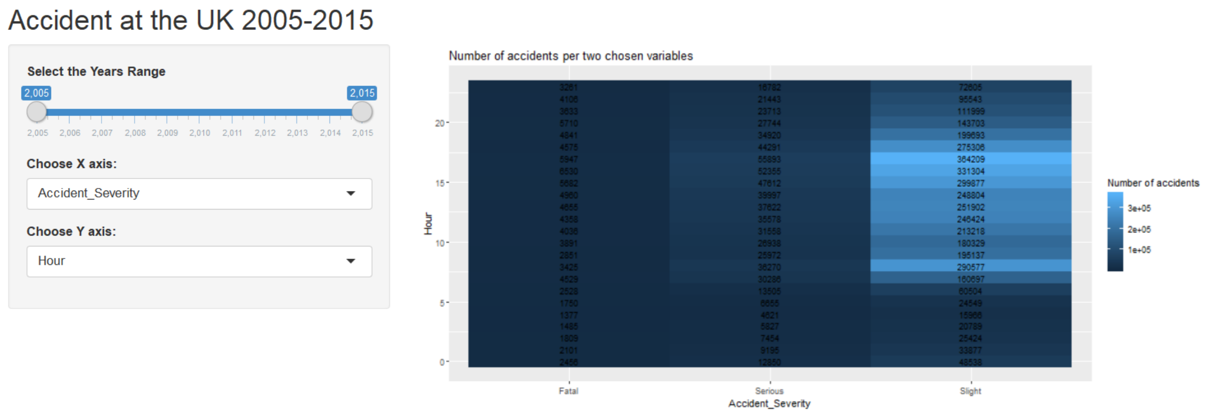
   User selections are: X->"Days", Y->"Hour", data -> number of accidents per hour per day.

   Screenshot from the application:



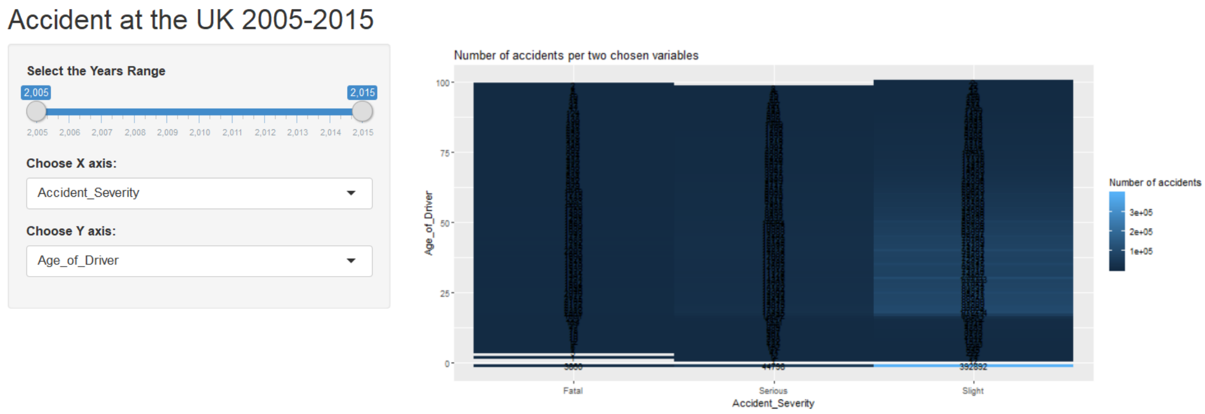
3. User will be able to identify risky hours of the day. User selections are: X ->"Accident_Severity", Y -> "Hour" Data-> number of accidents per accident severity per hour of the data.

   Screenshot from application:



4. User wants to find if there is a correlation between the driver's age and the severity of accident. User selections are: X->"Accident_Severity", Y->"Age_of_Driver", Data->count of accidents per age per severity.

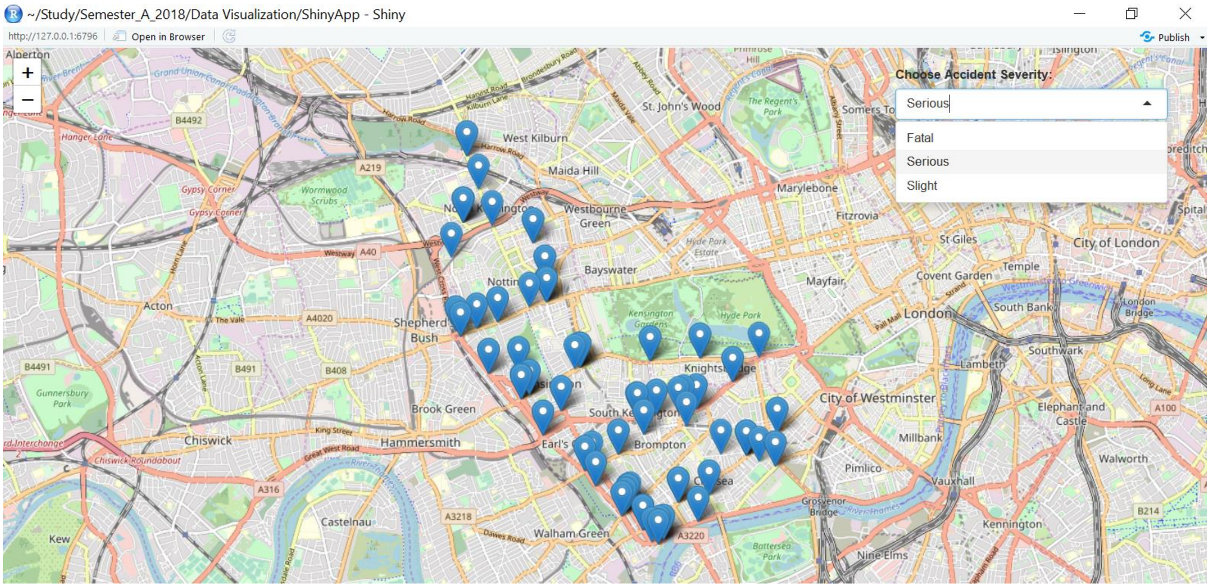
   Screenshot from the application:



The next user tasks are written for implementation of Shiny UI using a map (Leaflet), I was working on it and couldn't get it to a level that it can be submitted for grade. (file is attached and working for sample data only ("Leaf_Acc.R").

5. User will be able to view accident location on map per accident severity. Data file contains all the accident Longitude and Latitude as well as the Accident_Severity index (1-Fatal, 2-serious, 3-slight) and the application will present the data on the map.

   This is a screenshot from the application:



6. User will be able to identify amount of accidents per Area-District

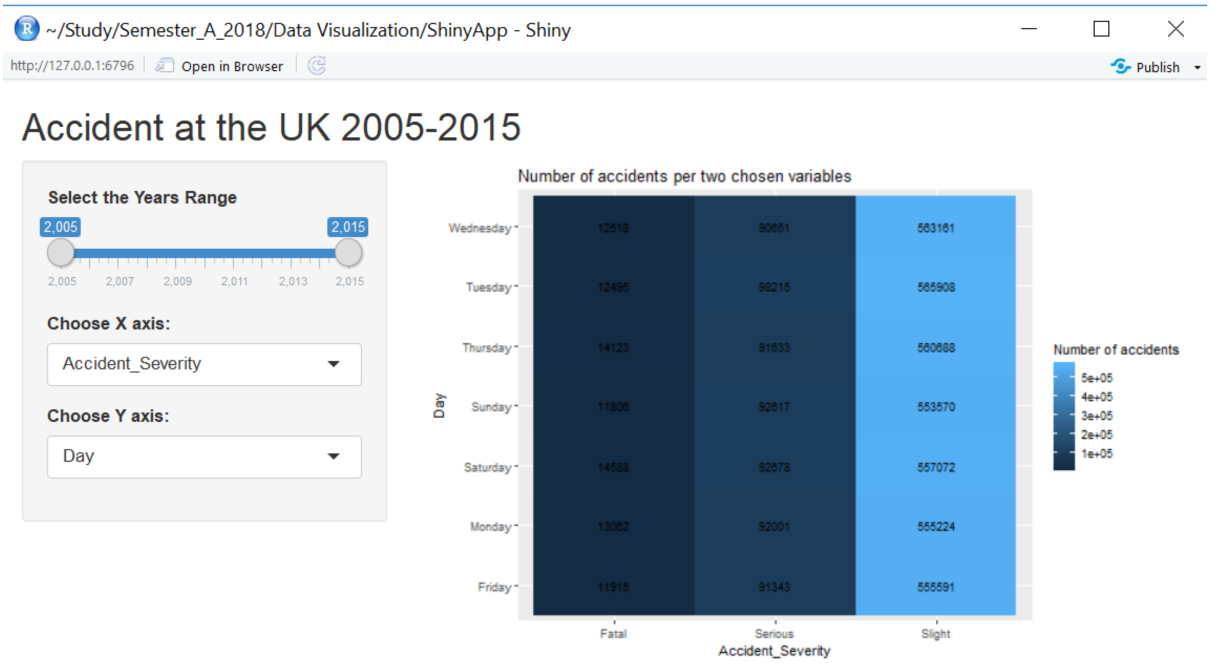
7. User will be able to identify High risk roads. ((Map, Severity, amount, Day, road num?) I will create a new derived attribute called: "Risky_Road" – this is how I will calculate it:

   For each Road number – sum{number of accident*accident severity(fatal-3, serious-2, light-1)}.

   This will create an index for dangerous roads which then I can highlight on map.

How?

This is the default screen of my application:



I took into consider two aspects – Channels and Marks.

Marks – since I know that the user wants to find correlation between two attributes that will explain him the number of accidents, I decided to use visual mapping using Heatmap.

Channels –

- **Position** – I have a **sidebar panel** the contain the arguments user will select (X\Y axis), then I have the **main panel** – which contains the plot of the heatmap.
- **Shape** - NA
- **Size –** heatmap gets most of the screen with a small legend of colors
- **Color –** since it's a heatmap, I'm using the color to represent a continues variable, so each range of number gets different color (actually not different Hue but different saturation\intensity)
- **Tilt** - NA

3. Visual mapping

When using heatmap – I can map 3 attributes.

Axis X → 1st categorial attribute, for example: Accident severity, Day, Month, Casualty severity.

Axis Y → 2nd categorial attribute, for example: Day, Hour, Day number.

Data → it's the sum of accidents per axis X\Y. mapped to a **color** by ranges.

On my application I also provided the user the ability to **filter range** of years he would like to study, so, he can decide to view single year and range up to 2005-2015.

In order to be able to use understand the data I'm working on I run some analysis on my raw data (see attached "Raw_Data_prep.R" file), most of the steps and learning I showed here under section: "Descriptive Statistics". Finally, once I understood the data I could narrow it down to the exact columns I need for my visualization and created a new CSV file which is attached part of this project – "Accident_Vehic_Cas.CSV".

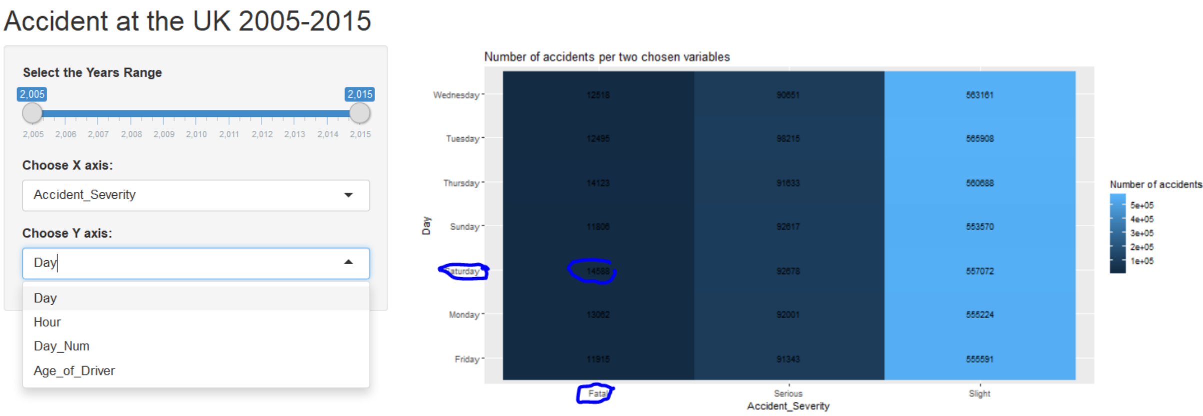
In order to run my Shiny application, you need the above CSV file and "App.R" file.

4. Results

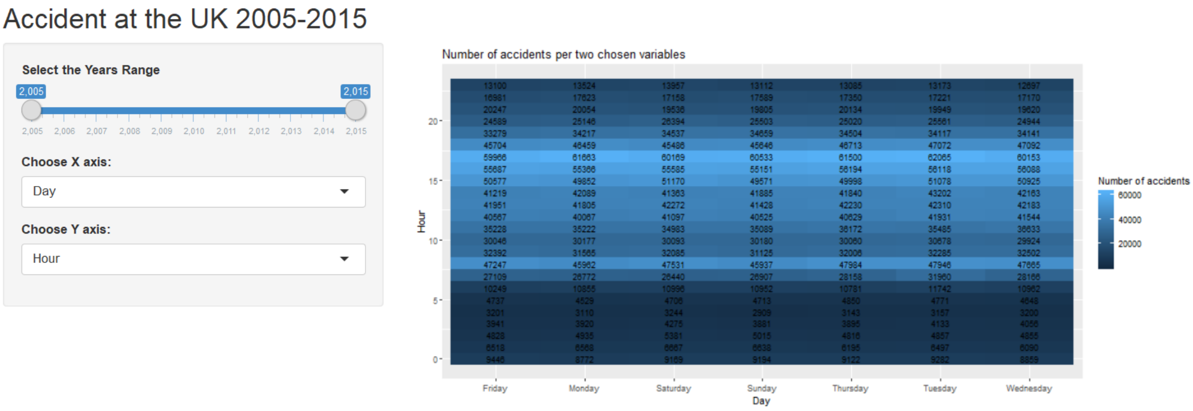
I learned a lot from this projects about accidents at the UK and was happy it revealed me some surprising insights.
I will start with the descriptive statistics that teaches me a lot about the raw data and the attributes. Some of the finding were "expected" like men's are involved in more accidents, Tram is the safest transportation. I did find some surprising facts like: cycles are most dangers vehicle after a car, on the other hand, pedestrians are second in number of casualties! I also expected the mean age of drivers to be younger than 38…

Insights from the visualizations –

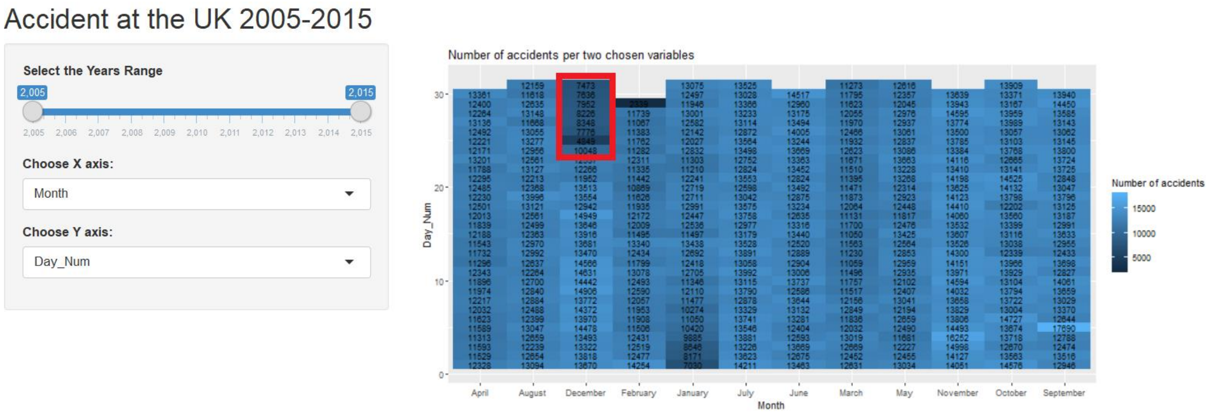
- Found that most fatal accidents happen on Friday's. could be explain by the fact that this the time people are going out on the weekends (and drinking?)



- Most of the accidents occurring during the day, but you can see peak on early morning (when people commute to work) and the afternoon, around 15-19 when people are traveling back home.



- At the end of the year there are significantly less accidents



Unfortunately, I failed to achieve some results I wanted:

1. I couldn't present the data I wanted on top of a map (as I mentioned above, I have a Shiny application that can only show part of the data on a map).
2. failed to present well one heatmap – casualty_severity\Age_of_Driver, since I put labels of the data it hides the number and was too crowed to be able to read it…
3. since my code was very generic in order to be able to present any two X\Y slices, I could play much with the graph or the axis since it need to fit all. This caused some issues like: days or month not ordered,

6. Evaluation

I will evaluate this visualization as we learned in class – “The Value of Visualization”:

- **Time -** since this visualization simply present huge amount of data, we are saving the user endless time of trying to figure it out of a table. User can easily choose whatever slice and dice of the data he likes in a split second.
- **Insights -** the ability of the visualization to give new insights, as I mentioned on the Results chapter I found many insights I couldn’t possibly know without this visualization.
- **Essence –** this visualization indeed summarizes for us huge amount of data which was almost impossible to research in traditional methods. (or cost us a lot of time and resources)
- **Confidence –** I think this visualization results are supporting some pre knowledge we all have about car accidents so this only strength our confidence of this visualization.

Advantages of the visualization –

- Maybe the main advantage of this visualization is simply its ability to contain vast amount of data – this case, 10 years of accidents in the UK alone created a raw data file with ~4.6M records!
- The visualization using the Shiny package gives the user simple UI where he can interact and research the data as he wishes to and on real-time.
- I created the visualization only after I had the user tasks I want to achieve clear and written. So, it’s simply “tailor made” for the user’s tasks.
- Scalability – I see how fast this UI works with 4.6M records, I’m sure that even double of the records or adding more attribute will not affect the performance. R appear to be very strong tool for analyzing this kind of data and Shiny make it very “handy” for almost everyone to create his own UI.

Dis-advantages of the visualization -

- Since the code is generic some visualization are missing adjustments like order axis, choose different color etc.
- One of the visualization - casualty_severity\Age_of_Driver, showing the data badly, labels are over plotting and you can’t read the data.
- Not all combinations on the dropdown make sense, but, I didn’t want to mess with hardcoding the possible combinations.