Traffic stops in the State of Rhode Island:
An analysis of the factors contributing to the arrest following a Police vehicle stopping

Simone Zanetti IS71082A Individual Project

Research Question

This assignment aims to provide an analysis of the factors which contribute to the arrest of an individual in the context of traffic stops performed by Police. The analysis is made possible thanks to the inspirational initiative of the *The Stanford Open Policing project*, an interdisciplinary team of researchers and journalists at Stanford University, who define their goal as 'committed to combining the academic rigour of statistical analysis with the explanatory power of data journalism' (openpolicing.stanford.edu), and provides rich dataset to the public for their own purposes. Moreover, this project is encouraged by the moving work of *E. Pierson, C. Simoiu, J. Overgoor, S. Corbett-Davies, V. Ramachandran, C. Phillips, S. Goel. (2017)* "*A large-scale analysis of racial disparities in police stops across the United States*", which investigates the possibility of racial disparities in Police stops across the United States. Not for nothing, the question of racial disparities in relationship with Police attitude is very actual, and studies which tend to verify effective differences of behaving by authorities towards different races seem to be numerous. One of them is the '*Contacts Between Police and the Public, 2015*', released by the *Bureau of Justice Statistics* (bjs.gov/content/pub/pdf/cpp15.pdf) which observed a different attitude of Police towards different races, both in traffic stops and street stops[1].

For the occasion, my personal analysis has been narrowed to the State of Rhode Island, due to the fact that the variables provided by the Police for this State were exhaustive, differently from other States, and the number of observations and structure of the data allowed to obtain an introductive sense of the analysis, leaving space for future investigations. It is important to specify that the analysis is only inspired by the above-mentioned topic, but not forced into those premises. In fact, the study aims not only to focus the attention on recognising the existence of racial disparities in the event of police stopping but to a more generic analysis whose questions can be summarised as follows:

- To what extent demographic factors influence the possibility of being arrested in the event of a vehicle stopping by Police.

- How time variables have influenced/ are influencing this phenomenon over the course of the time.

To conclude, the quantity of variables available and the lack of knowledge about contingents factors such as cause of arrest, data related to criminal activities on the State and so on, imposes the necessity to remind the reader about the important boundaries between correlation and causation: although in most of the cases the two concepts can be linked together, it is not the purpose of the analysis to focus on this effort. In fact, whenever a correlation is observed in this investigation, it will be important to bear in mind that establishing causation between two events would require a deeper analysis.

[1] The research stated that Black residents were more likely to be stopped by police than white or Hispanic residents, both in traffic stops and street stops. Black and Hispanic residents were also more likely to have multiple contacts with police.

Data

The data was found after a series of online researches on the possibilities offered around the aforementioned topic. In this context, an online course offered by the e-learning platform Datacamp has allowed me to discover the Stanford Open Policing Project, which provides through its website collected and standardised data on vehicle and pedestrian stops from law enforcement departments across 31 countries (openpolicing.stanford.edu). The data have been provided into a csv[1] format, which contain 509,681 observations, corresponding to also numbers of Police stops ranging from 1st January, 2005 to 31st December, 2015[2]. The dataset has been provided with a README document, which allowed a clear overview of the dataset and description of each variable contained into it. A summarised description of them is shown in the table below[3]:

Variable Name	Type	Notes
id	String	The unique ID assigned for each stop
state	String	two-letter code for the state in which the stop occurred
stop_date	Date	The date of the stop (YYYY-MM-DD format)
stop_time	Date	The 24-hour time of the stop (HH:MM format)
county_name	float	standardised name of the county in which the stop occurred
county_fips	float	Standardised code in which the stop occurred
fine_grained_location	NaN	Nan
police_department	Integer	Police department that made the stop
driver_gender	String	Gender of the driver
driver_age	float	The age of the driver when the were stopped
driver_race	String	The standardised driver race
violation	String	Violation committed by the driver (standardised into categories)
search_conducted	Boolean	True/False whether the search was performed
search_type	String	The normalised justification for the search
contraband_found	Boolean	True/False indicating whether a contraband was found following a search
stop_outcome	String	The outcome of the stop
is_arrested	Boolean	True/False indicated whether an arrest was made
stop_duration	String	The duration of the stop (0-15, 16-30, 30-60, 1+)

[1] comma-separated-value

[2] In fact, the data is structured in a way where each observation corresponds to a Police stop

[3] The type do not refer to the format in which the file has been presented (only *string* and *float*). In the table it is labelled with the right value in order to ease the understanding.

out_of_state	Boolean	Not described (it will be dropped)
drugs_related_stop	Boolean	True/False indicating whether the stop was related to drug
district	String	The state is broken into six police districts, also known as zones
location_raw	The variables belonging to this section represent the original (raw) data. They have been put in this section since a standardised copy of each of them is already available in the dataset. From this point of view, they are useless and their description represents a duplicate of above mentioned variables.	
driver_age_raw		
driver_race_raw		
violation_raw		
search_type_raw		

. Initial processing

The first step in the so-called process of data wrangling was the necessity to narrow the period of analysis, excluding the observation recorded before 2007[1]. In order to ease the process, the variable `stop_date`, recognised by Python as a *string,* was converted into a *datetime* object and set as index of the dataframe. This allowed to easily filter the data, which now presents observation from the 2007 to the end of 2015.

The second phase begun with the analysis of the *missing values* for each of the variables. In this context, having an interesting and promising variable does not represent a sufficient condition, since any variable needs to be consistent during the whole period considered in the dataset, and the amount of *Nan* values[2] should not be too high. For this reason, the columns `county_name`, `county_fips` and `fine_grained_location` have been dropped, as the number of missing values for them corresponded to the number of observation of the dataset (`len(df)`). On the other side, although the variable `search_type` contained a numerous amount of missing values, it was necessary to be cautious, and verify whether the lack of value was due to the absence of any search for that particular observation[3]. Consequently, each missing values of `search_type` was replaced by the category *no search conducted*. Moreover, rows with *Nan's* distributed over different variables have been removed, making the dataset empty of any missing value.

The third phase regarded the necessity to verify the meaning and the utility of each variable, in the optics of considering the possibility to remove useless columns. In this perspective, all the raw variables which owned a standardised copy of them on the dataset, were removed (`location_raw`, `driver_age_raw`, `driver_race_raw`, `violation_raw`, `search_type_raw`, `out_of_state`). The variable `id` was removed and replaced with a simpler version of it (identifying each case with a unique integer number x_i (`i = 1, .. , tot number of observations`)). In addition, the variable `state` was removed, as it - inevitably - contained the only State of the analysis, which is Rhode Island. To conclude, it was necessary to verify the data type for each variable in order to - eventually - convert them into the best data type

[1] From README: 'Stop counts are considerably lower in 2005/2006'

[2] Technical language which identifies Missing values

[3] Technically speaking, a *Nan* in search type occurred any time *search_conducted == False*

for the analysis. For the occasion, a function has been created to ease the process (Appendix 1)[1] of conversion. Moreover, with the aim of setting a variable of datetime type as index of the dataframe, the two separate columns `stop_time` and `stop_date` have been concatenated into a new variable called `stop_datetime`. The new variable has been then converted into a *datetime* format and then set as the index of the DataFrame. Specifically, this new situation provides rich advantages to the analysis, since it allows to filter and gather observations based on any temporal unit of interest. Moreover, it made useless the creation different variables such as `stop_year`, `stop_month`, `stop_weekday`, `stop_time`, since the `df.resample(x)` and `df.groupby(df.index.x)`[2] functions make the process immediate, without the need to create new variables. Further analysis of the variables proved that the dataset was clean and ready to be used for the analysis. At this purpose, it is important to observe that the data provided by the *Stanford Open Policing Project* is already partially cleaned by their team. In fact, as previously mentioned they are fully committed to provide a dataset which can serve as source of analysis across different States in the USA. As a consequence, the necessity to grant the possibility of comparison between data gathered from different department/with different recording systems leads them to the need of standardising variables which can be different across the States. This has allowed me to avoid the constraint of deepening the process of cleaning of data (gathering raw observations into few categories, etc.), giving me a sense of (even bigger) gratitude towards the Team and their mission.

.Analysis

Concluded the necessary modifications, the final DataFrame covered 409,374 observations (i.e. Police stops) and 14 variables [3], which are summarised below:

Variable Name	Python Type	Variable Name	Python Type
id	int64	search_conducted	bool
driver_age	int64	search_type	category
driver_gender	category	contraband_found	bool
driver_race	category	stop_outcome	category
violation	category	is_arrested	bool
police_department	category	stop_duration	category
district	category	drugs_related_stop	bool
stop_datetime	datetime		

The variable `stop_datetime` is no longer a variable of the the DataFrame, since it now represents the index of it. However, it was worth mention it on the table above.

[1] Refer to Appendix 1 to also observe which variable were converted, and in which data type

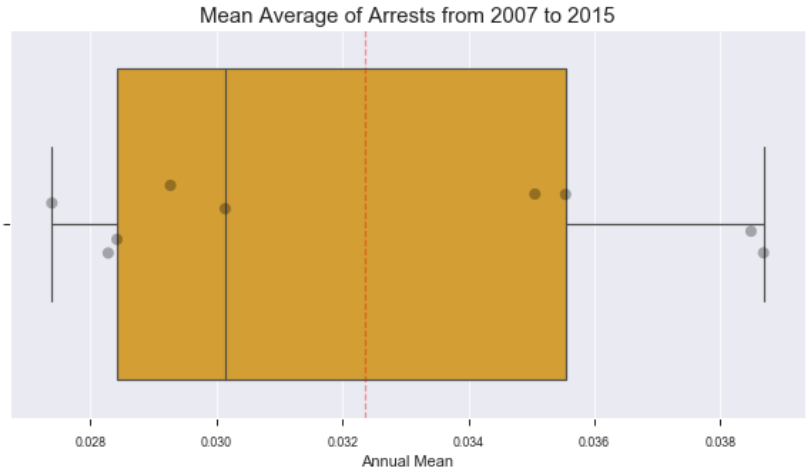
[2] Where x represents the unit of time of interest, such as 'year', 'month', 'hour', etc. See documentation for details:
`pandas.DataFrame.resample` (pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.resample.html)
`Indexing and selecting data` (pandas.pydata.org/pandas-docs/stable/indexing.html)

[3] `df.shape = (409374, 14)`

EXPLORATORY ANALYSIS

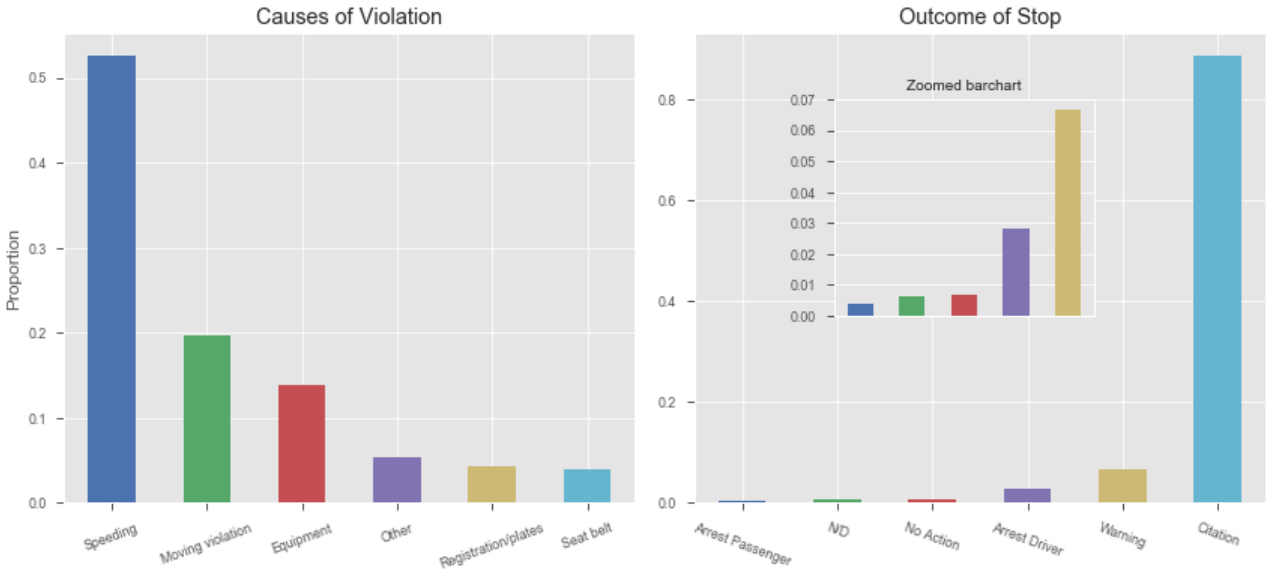
The phenomenon of Arrest after a vehicle stop by Police

Before starting the investigation, it is important to verify the numbers behind the phenomenon in order to obtain a sense of the dimension of the analysis. In particular, during the period included between 2007 and 2015 the average percentage of people arrested after a stopping by the Police has been 3.24 %, as indicated by the red line in the Boxplot, which corresponds to 45,486 people per year. The shape of the box plot seems to indicate a trend in which values rarely go far below the median of 3%, but they can be relatively far above. In fact, in some cases the annual mean reached 3.9 %, which means that the number of people arrested has reached the pick of 54,752.

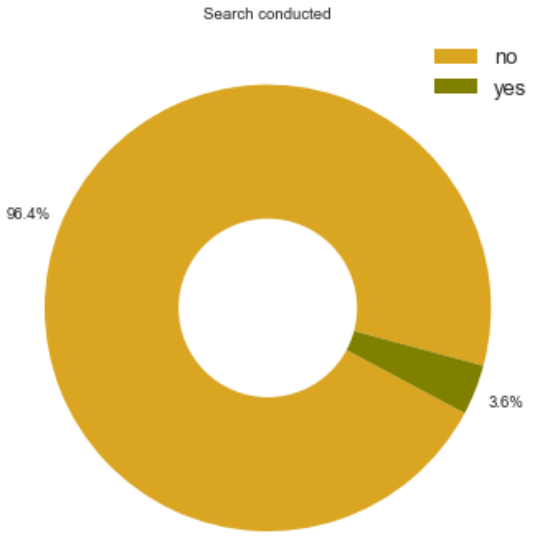


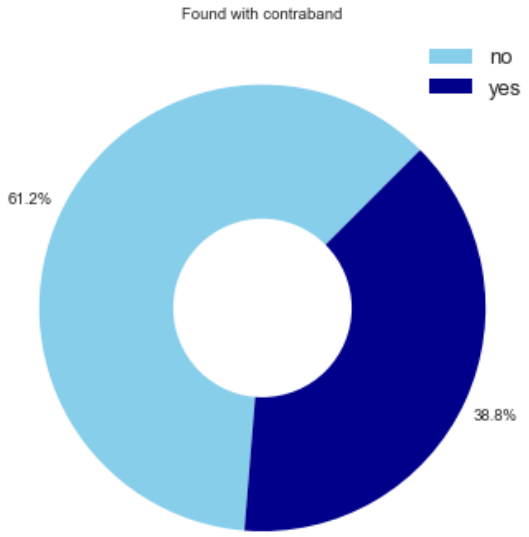
Introduction to the Variables

The following section moves into the premises presented in the previous paragraph, which introduced the necessity of visualise the trend and key points related to the main variables of the investigation. This brief introductory section will help the reader to increase their awareness about the phenomena during the analysis.

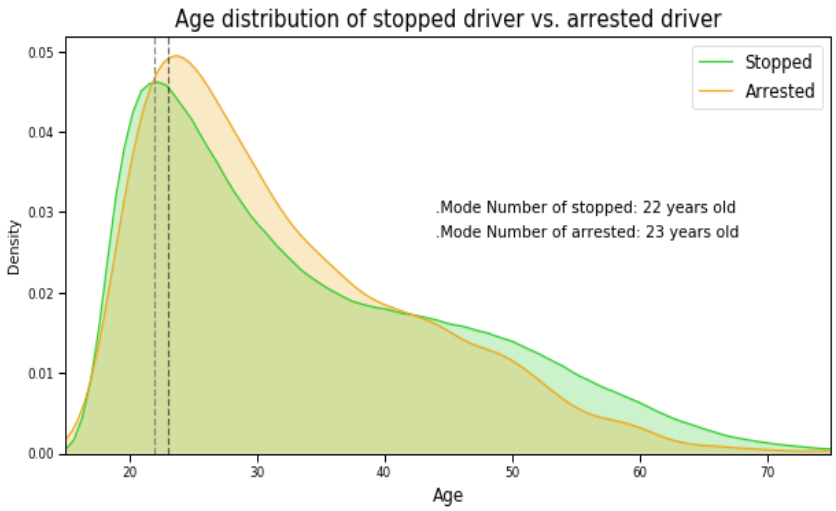


1. The bar charts exposed above shows the different causes of violation for which the drivers are pulled out by the Police and the outcomes of their stop. It is possible to observe an evident predominance of people stopped due to Speeding violations: in fact, for over 50% of the cases this is the reason, while the second and the third causes are due to moving violations and equipment issues. With regards to the outcome of the stop, the great majority of the cases ends with a citation. Specifically, almost 90% of the time. In this context, as previously mentioned, the arrest regards only about 3% of the cases, and it is curious to observe that 0.5 % of the time is the passenger the victim of arrest.

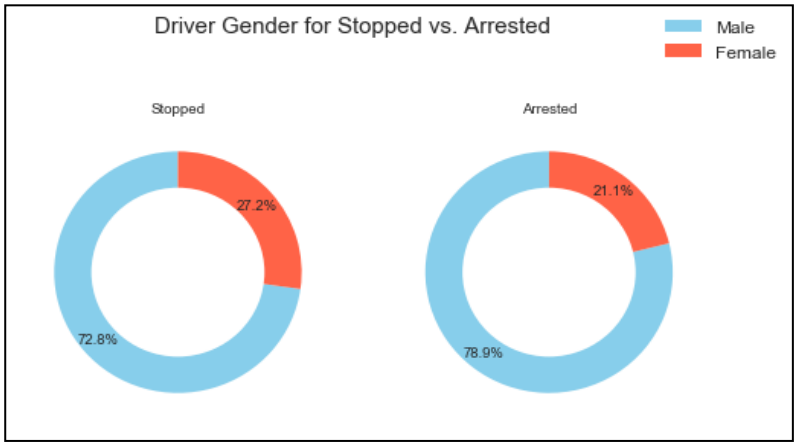




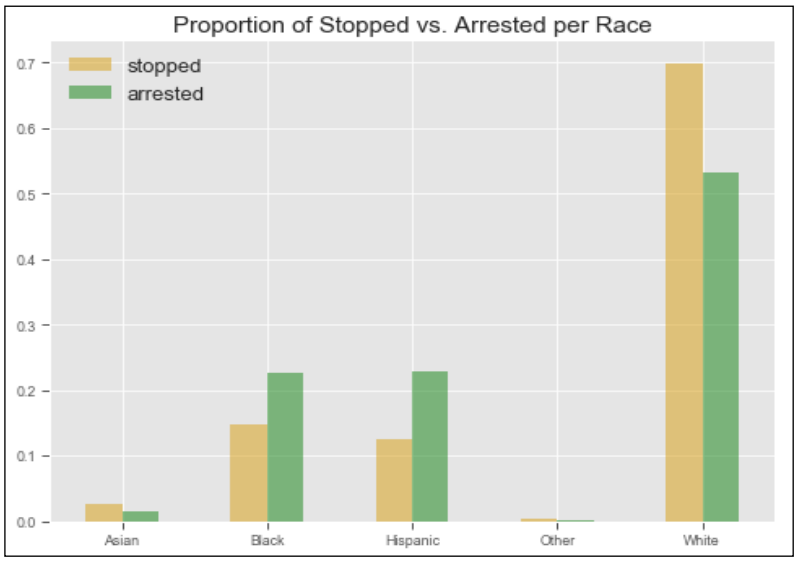
In addition, it is possible to observe from the doughnut charts above that only 3.6 % of the time Police conducts a search, and when it does 38.8 % of the time contraband material is found.



Moving towards a generic overview of the different demographic factors, the density distribution of ages for those who have been stopped, compared to those who have been arrested, shows a trend not too different from each other. In particular, despite both the curves are right skewed, the distribution relative to arrested people, indicated in yellow, seems to expire earlier than the other. This indicates that more people is arrested during the age between 20 and 50, while people of every age is likely to be stopped by the police.



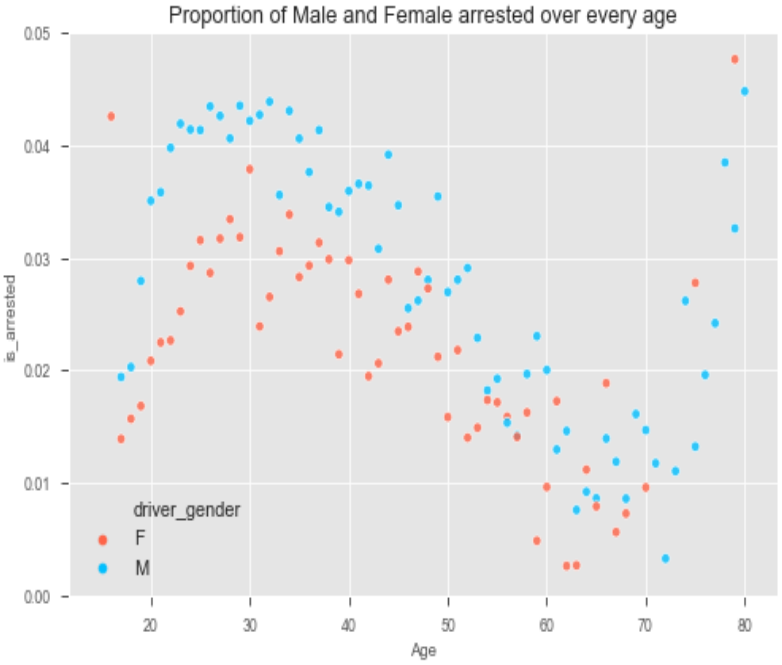
Comparing the number of Male and Female drivers in occasion of Police stops in general, and in occasion of Police stops with an arrest, it is possible to observe a slightly different tendency, as shown in the doughnut charts.
In fact, while the percentage of Female driver generally stopped is around 27.2 %, the one of those arrested decreases to 21.1%. At this point, it is only worth notice this fact, while in the next section the phenomenon will be taken in consideration, and an attempt of digging into these premises will be performed.



The graph presents the distribution of races in relationship to the proportion of people stopped and the one of people arrested. From this point of view, it is very significant to observe a diverging trend between the races across the two situations analysed. In fact, while white represents the dominant category in both cases, in the case of arrest, its value tends to decrease, while the proportion of Black and Hispanic people tends to increase. In this situation it is easy to fall into superficial conclusions. From this point of view, in order to verify the risk of a discriminatory attitude from the Police it would be necessary to perform a deeper analysis.

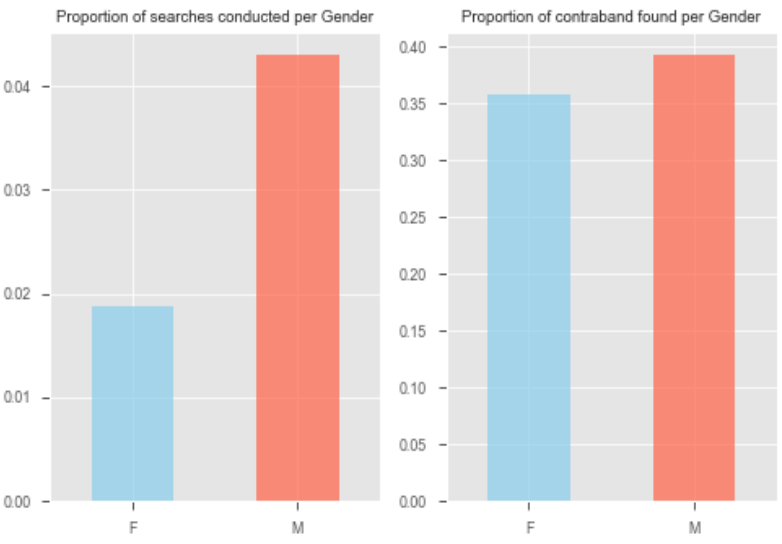
Part 1: Analysis of the demographic factors (Age, Gender, Race)

The following analysis inspects the proportion of people arrested across the different ages, observing the trend for both Male and Female.



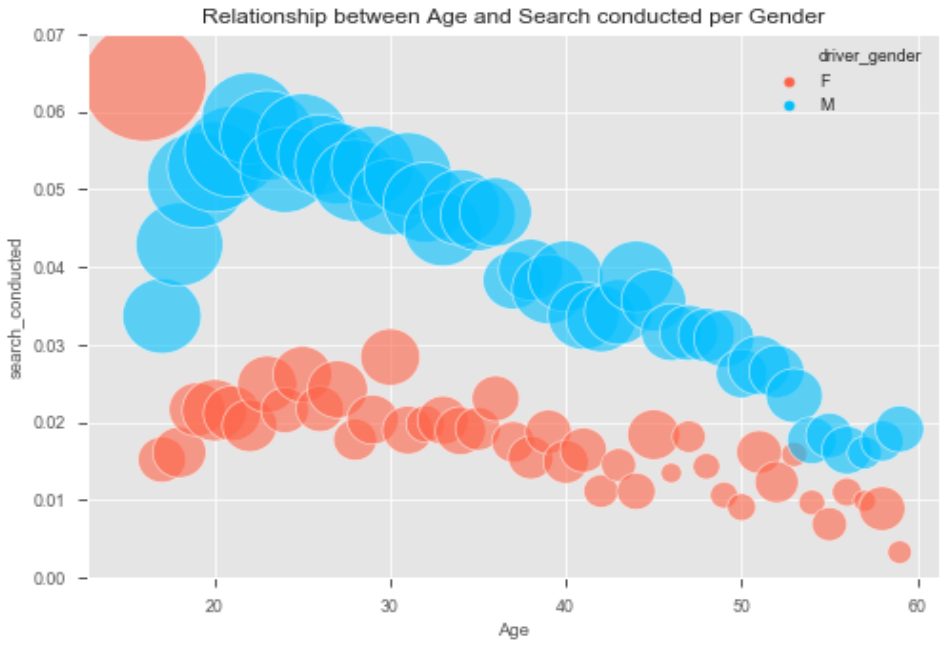
The graph seems to confirm the previously mentioned relationship between age and number of people arrested. In this context, it is possible to verify that the trend seems to be stable for both Male and Female, with a tendency of Female to have a lower proportion between 20 and 50 years, where the average of arrested Male is superior. As a consequence, formulating a series of hypothesis in order to investigate the different attitude can be licit.

- May it be because Female tend to be searched less by the Police?
- May it be because Males are more likely to be involved in illicit activities ?

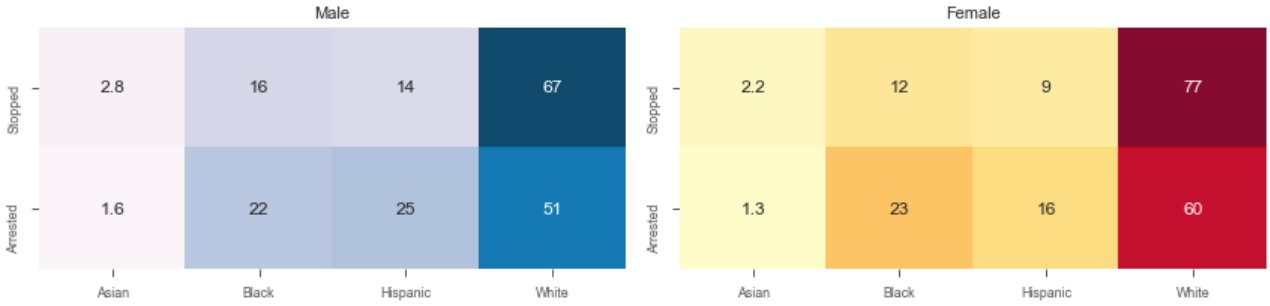


The graph seems to show that searches for Males are basically conducted more often than for girls. Slightly more than 4 % for males against less than 2% for Females. However, the proportion of people found with contraband is almost balanced between the two genders. The graph below is significative in summarising the key concept mentioned. In fact, through this indicative visualisation, it is immediately possible to compare the difference between Male and Female in relationship with the search conducted. Over the entire period of analysis (i.e. during the age between 15 and 60) Male are searched more than Female. The different size of the dots is proportional to the average of contraband found for that specific category of age. It is possible to

verify how the search conducted are actually justified and coherent with the average of contraband found, same if establishing the causation between these two facts (which causes what) is not easy, and probably it is also not fully possible.

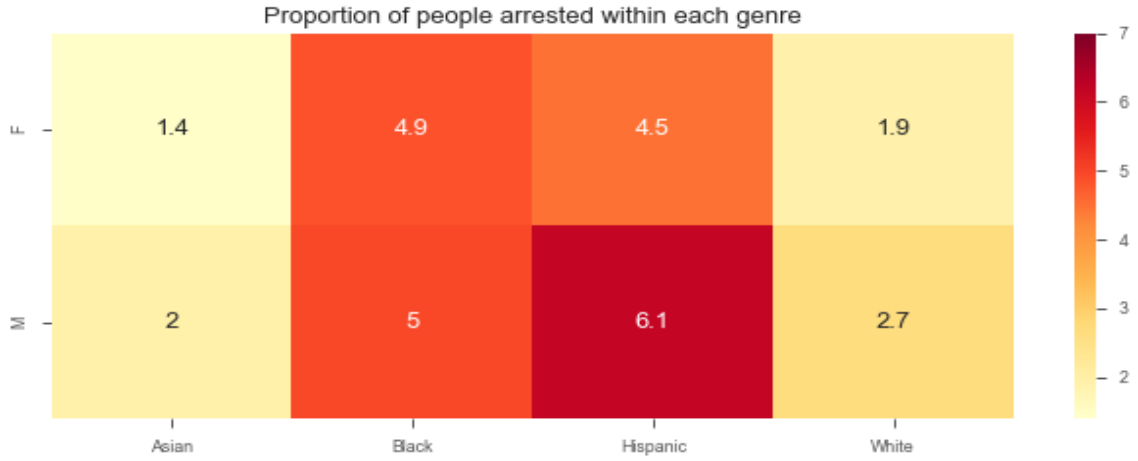


An interesting analysis regards the one which has been conducted analysing the contribute of the race of the driver in the phenomenon of arrest.



The heat map confirms consistency between the above mentioned trend, which observed a tendency of the proportion of Black people and Hispanic people to increase in occasion of an arrest. In this contest, it is vital to observe how white people still represent the biggest percentage of people both stopped and arrested, but the results of this analysis have inspired me to further investigate in this direction. With this purpose, an investigation of the percentages of arrest regarding each race has been conducted.

Moreover, while on the previous analysis the percentage was in relationship to the other races (e.g. 51% of male arrested are White, 22% are Black), curiosity has led me to verify the percentage of arrested people belonging to one race in comparison to the total of people from the same race. In other words, what is the proportion of people stopped which are arrested across the different races?[1]. I personally consider the results surprising.



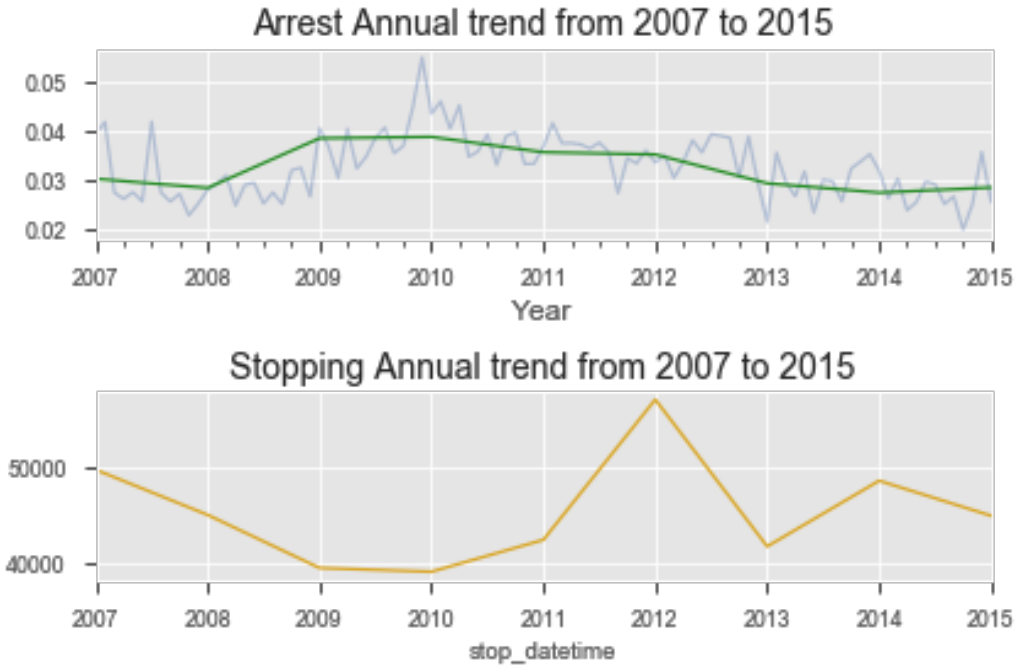
We can observe how Hispanic and Black Males have almost double tendency to be arrested[2]. The same situation is verified for Females, with a percentage sharply lower for White and Asian people.

[1] This is the proportion between the number belonging to one race divided the total number of people of the same race which have been stopped

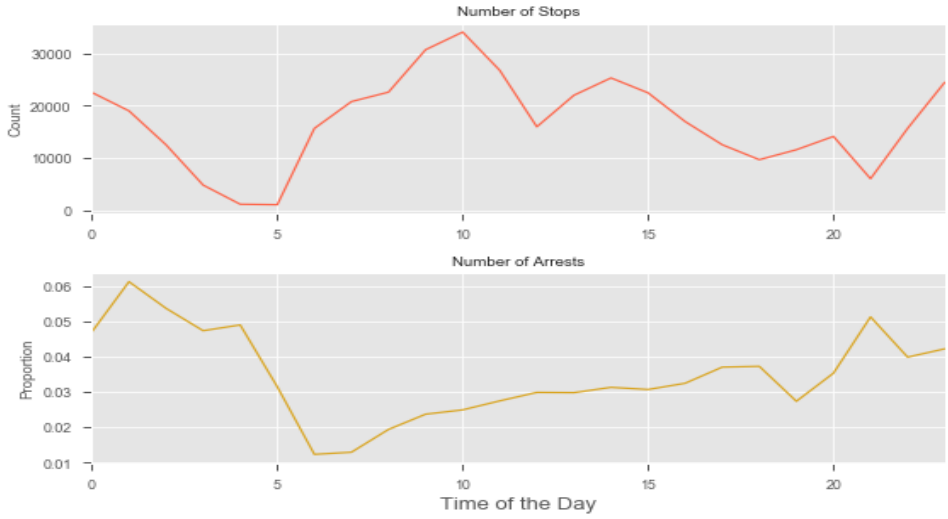
[2] Higher percentages on the heat map indicated with darker colours

Part 2: Analysis of time variables

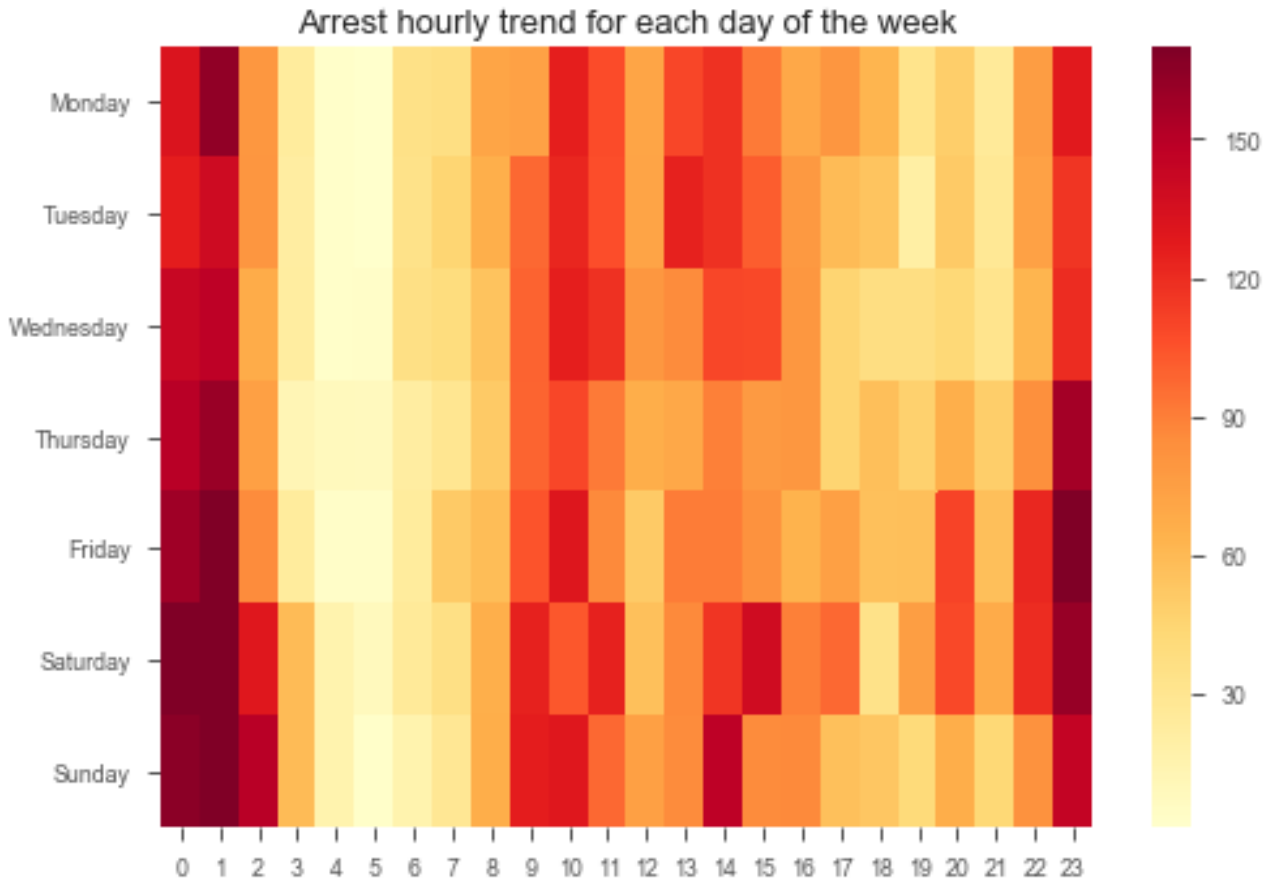
This section intends to give an overview of the relationship between temporal concepts such as year, month, hour and days of the week and the phenomena of arrest. Differently from the previous section, the graphs tend to be more self explicative, so the comments will be limited to the key point of the analysis.



One of them regards the tendency observed over the years of an increasing of the proportion of arrest in occasion of periods of decreasing in the number of stopping, as observed in the period between 2008 and 2011. This situation can suggest a tendency of the number of arrests to remain constant, independently from the number of stops performed by the Police.



Significantly more interesting is the analysis which regards the number of arrests in relationship with a combination between the time of the day and day of the week, as shown in the heat map. This powerful and suggestive visualisation allows to observe and dig into the process of understanding of the moments of the week with the higher number of arrests, allowing to identify significative trends. One of these is that the higher amount of arrest happens during night time, in a range which widens from 11 pm to 2 am. Moreover, it is possible to verify that the weekend seems to have a slightly more intensity of the colours (i.e. more arrests), while the moment between 4 am and 7 am is the one with a trend generically lower than all the other moments of the day. The Line graph shown in *the* line graph above, which however presents values aggregated (not distinguishing week day) seems to provide consistency to the visual interpretation of the heat map.



Conclusion / Evaluation

I performed this analysis with the aim of deepen my knowledge about a topic which is very actual and significant for me. In this context, I had the possibility to verify the influence of demographic factors, such as age, gender and race, into the event of arresting after a vehicle stop by Police. In particular, I had possibility to verify how the trend between people stopped and people arrested seems to be roughly similar if compared with the age, with a right skewed distribution and higher values contained between 18 and 35 years old. On the other side, I could observe how the number of Male both stopped and arrested is definitely higher for Male than Female, with Female having a percentage of arrest lower than the one of stopping. In this context, I verified that Male are stopped more frequently than Females, while the proportion of contraband found between the genders is roughly similar.

In addition, analysing the composition of the different races in this phenomenon I observed how the proportion of Black and Hispanic people arrested tends to increase in comparison with the one of people stopped, although the proportion of white is the higher in both cases. On the contrary, this proportion tends to be completely opposite when I analysed the proportion of the people of a race arrested over the total number of people of that race stopped. In this case, results showed me that Black and Hispanic people have almost double arrested rate, then White and Asian. Moreover, the analysis of time variables has shown that the annual trends between 2007 and 2015 seemed to be relatively constant for the arrest averages, while the amount of amount of stops has shown a slightly discontinue trend with a period of decrease between 2008 and 2010 and a peak in 2012. Combining the week day and the time of the day allowed me to verify that the trend of arrest seems more focused between 23 and 2 am, with a slightly higher intensity distributed over the weekend.

To conclude, the structure of the dataset obtained was good enough to lead some good investigations and experiment a various range of visualisations, such as Box Plot, Bar chart, Doughnut chart, Density plot, Heat map, and Scatter plot. Moreover, the possibility to deal with *datetime* objects allowed me to work with time series. However, I consider the main lack of this dataset to be the lack of any significant location variable[1] or geographical coordinate, which would have significantly increased the value of the investigation and visualisations, allowing me to experiment interactive maps and verify the trends over the territory.

[1] The district value lacked any geographical indication