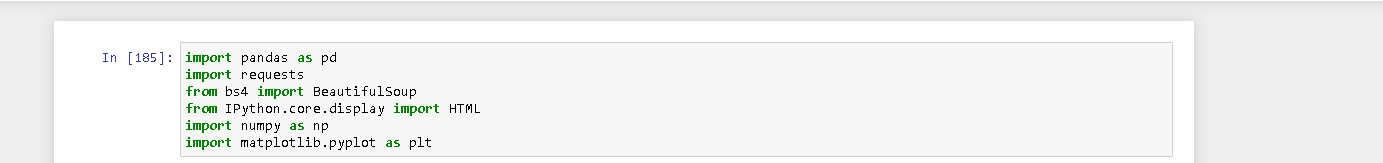
Omar Mohsen 179771 /Omar Haitham 182468

Rally race 1985

**First “Importation”**



needed libraries:

-Pandas: for creating tables

-Requests: for getting data from web Page (Html)

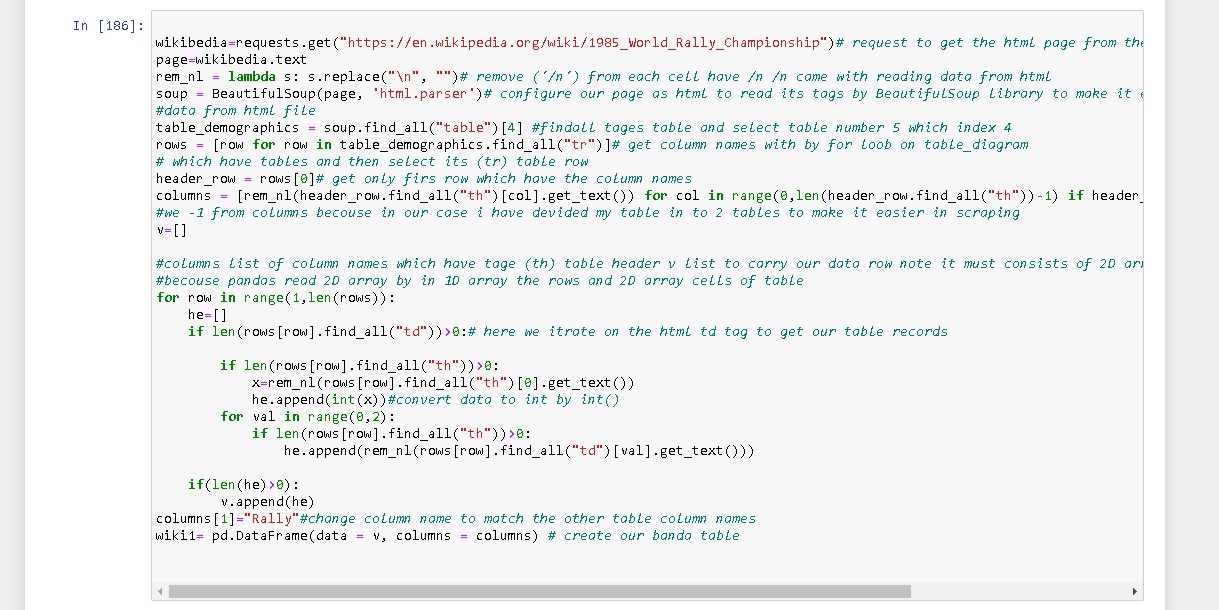
-BeatifulSoup: for reading and retrieving data from Html file

-Ipython.core.display HTML: for displaying html file in anaconda

-Numpy: we use it here to check on its data type that it has int64 which is an integer in pandas and it also used in data analyses and calculations

-Matplotlib. pyplot: it used to create charts and calculations

**Second “Get data from a web and Create Table”**



-get data from a web: by request, library get function and pass the URL (Wikipedia variable)

-get HTML: by request get we get a text that retrieved from the web page as HTML text (String, page variable)

- configure text as HTML variable soup by BeatifulSoup library to read it and retrieve data from it (soup variable)

-now we read our HTML by BeatifulSoup allow us to get any data by its HTML tag with function find() or find\_all() so we will get tables by find\_all(“HTML tag (which is the table)”) and select the needed table we need table number 5 which in index 4(table\_demographics variable)

- get our table columns: so we will make for loop on table\_demographics we will get data of tr(table row tag) which will carry all the rows in the table so the column will be located in a row[0] (index 0 in a row) and row[:1](data of row after 0 will carry our records)( variables row and columns and header\_row)

-to create a pandas table we need to give it parameters of columns and data record. data record must be a 2D array that the first array carries the rows and second array carry the cells and the second array must be the same length as the column length

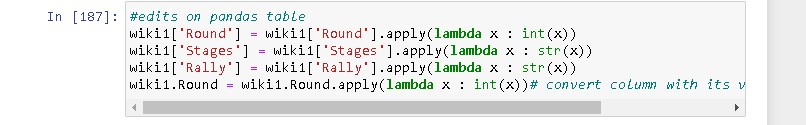
- to I have made for loop to search for td tag from row variable and insert its data in the array(list in python) I have divided the table into 2 parts tables the Podium finishers column rows more than the rest of the table rows and has inner columns

-finally, we have created panda DataFrame and gave it parameter data=” our rows” columns= “our columns names”

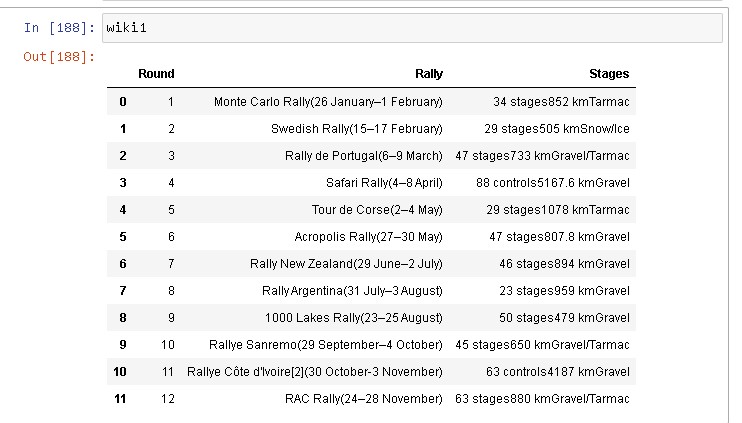
-

**Third “edit data”**

-here we have set every row its suitable datatype



**“Our pandas table of Wikipedia”**

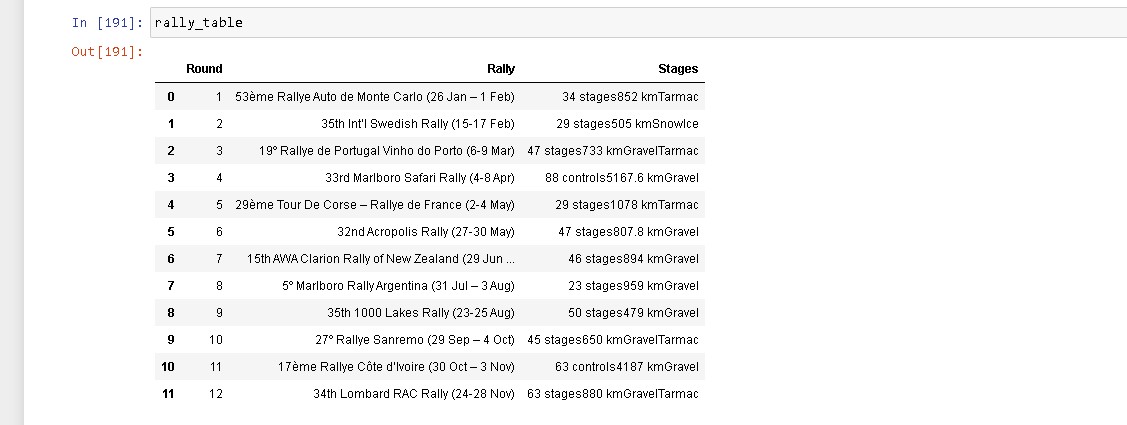


**Second Link:**



The same as before bus the table in index 0

**“Our panda's table of rallygroupbshrine”**

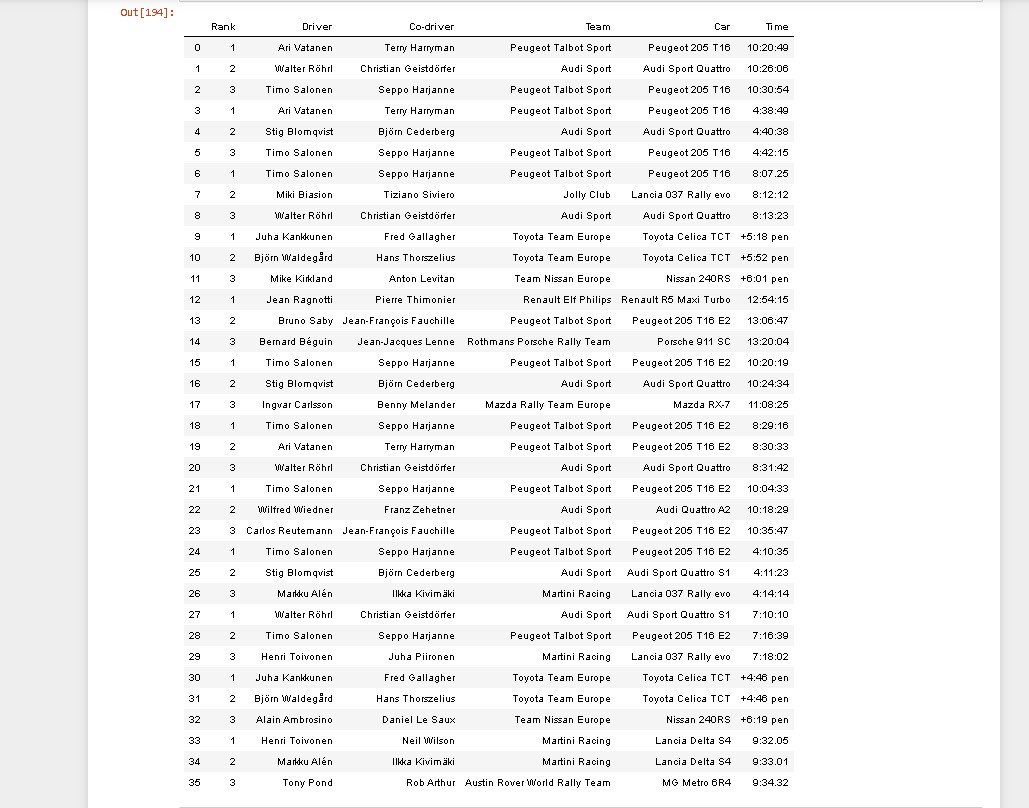


**Get the second part of the table**

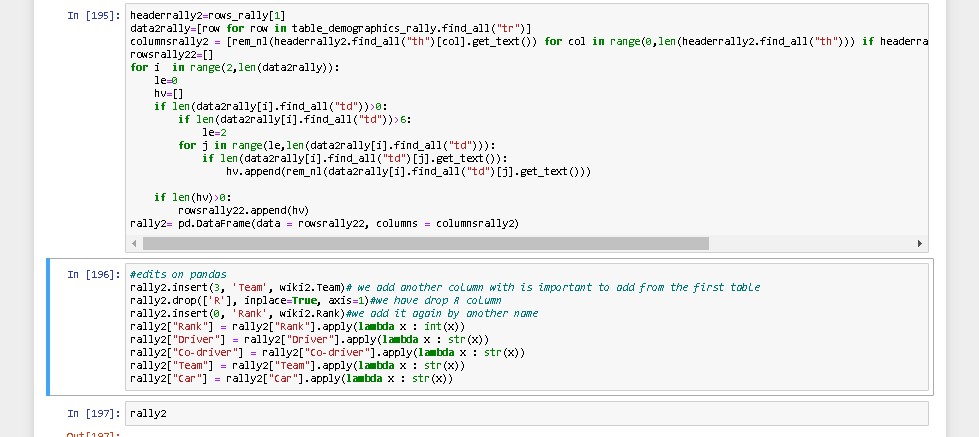


The same base bur we get the td tag after the first table date by jumping every row 2 cells

**“Second part by pandas of Wikipedia table”**

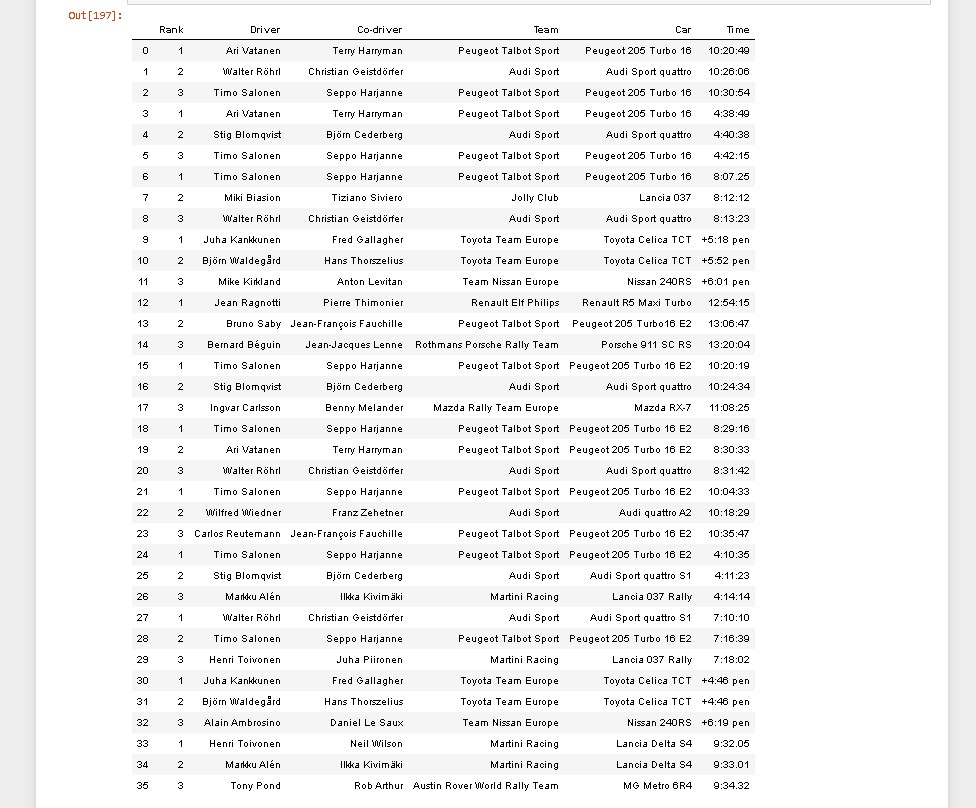


**Second Link second part of the table:**

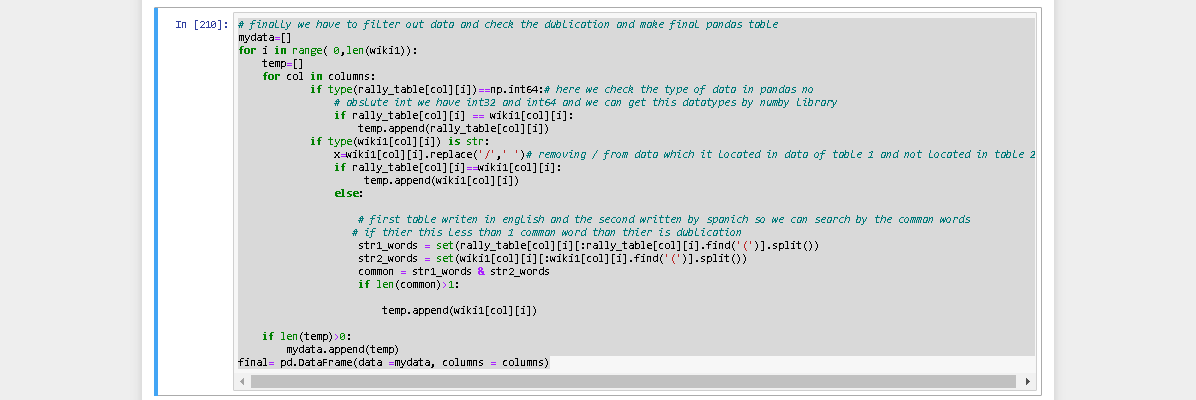


The same as before

**“Second part by pandas of rallygroupbshrine”**



**“Combining date for first part”**



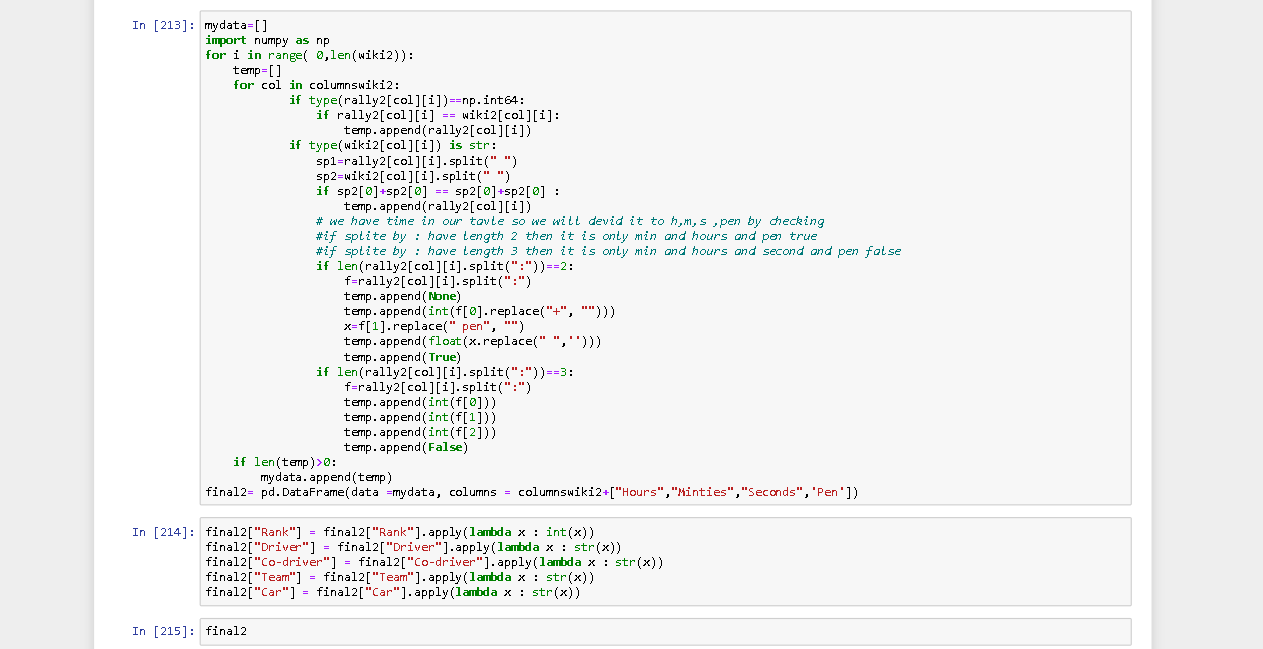
-to check duplication we must make a loop my first loop iterate on row data records, second loop iterate on columns. First, if condition check data type of cell data if its type is NumPy.int64( I have converted rank col to int but it does not convert to python Integer datatype it converted to Numby int64) will push it to 2d array. Else if type string we will make other functions because in the Wikipedia table the Rally name with English but in the second web with Spanish so we must search about the common words in each row so I have sliced the first row and the second and make variable common to get the list of the common words and I have removed the circle brackets and check if common words more the 1 push to the list

Then all the same as before

**“Final result of the first part”**

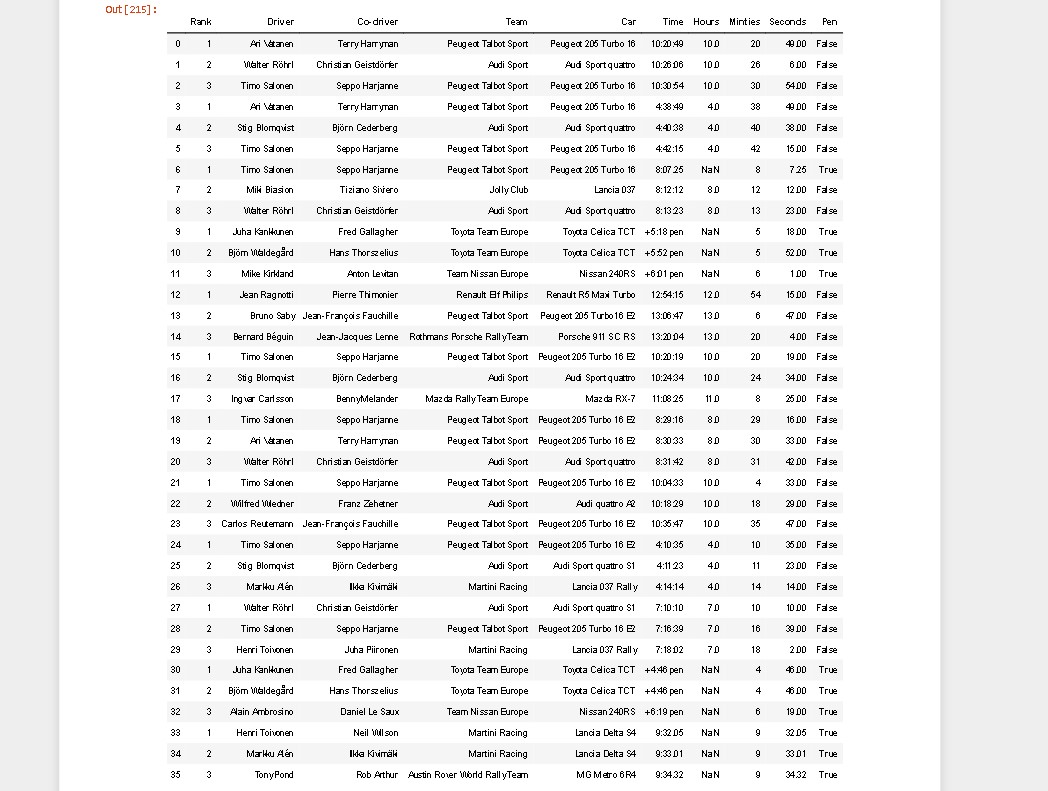


**“Combining date for second part”**



* In the second part the same as first but in the string checker if condition the( car column in the first table deferent in form of the second table) to we get the first two string in table 1 to first two string in table 2 that the car name and model in the first 2 string and the rest the car specs
* To divide the time column into hours, minutes, second, then we have to condition pen time(minutes, second) and not pen time (hours, minutes, second) so we add condition if in split by(:)= length 2 so it is pen so we will push split list index[0] hours is none to float we will push split list index[1] minutes to float and will push split list index[2] second to float and pen is true else if in split by(:) length 3 so it is not ben we will push split list index[0] hours to float we will push split list index[1] minutes to float and will push split list index[2] second to float and pen is false
* Finally, add validation to remove + and spaces to convert to number datatype

**“Final result of the second part”**

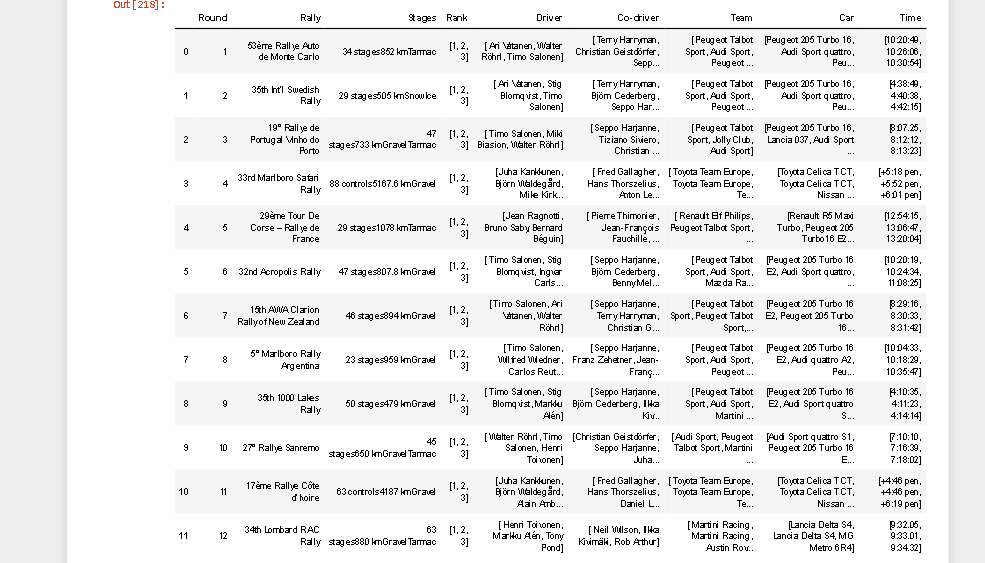


**“Combining 2 parts”**

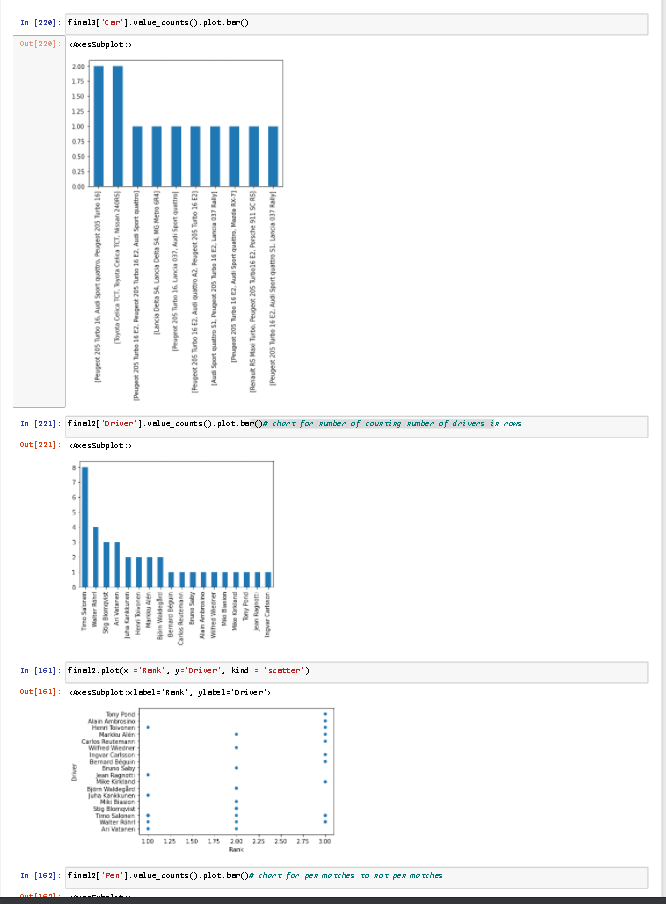


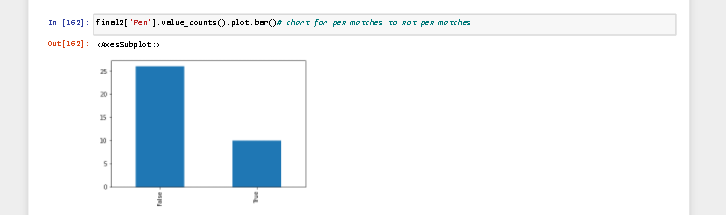
* I have make a list for the second part for example every 1 row in the first part matches with 3 from the second part so I have to make list for every 3 rows from the second

**“Final result of the combining parts”**

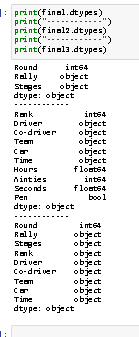


“Charts”





“Tables datatypes”



Note: please read code comments

Thanks for reading