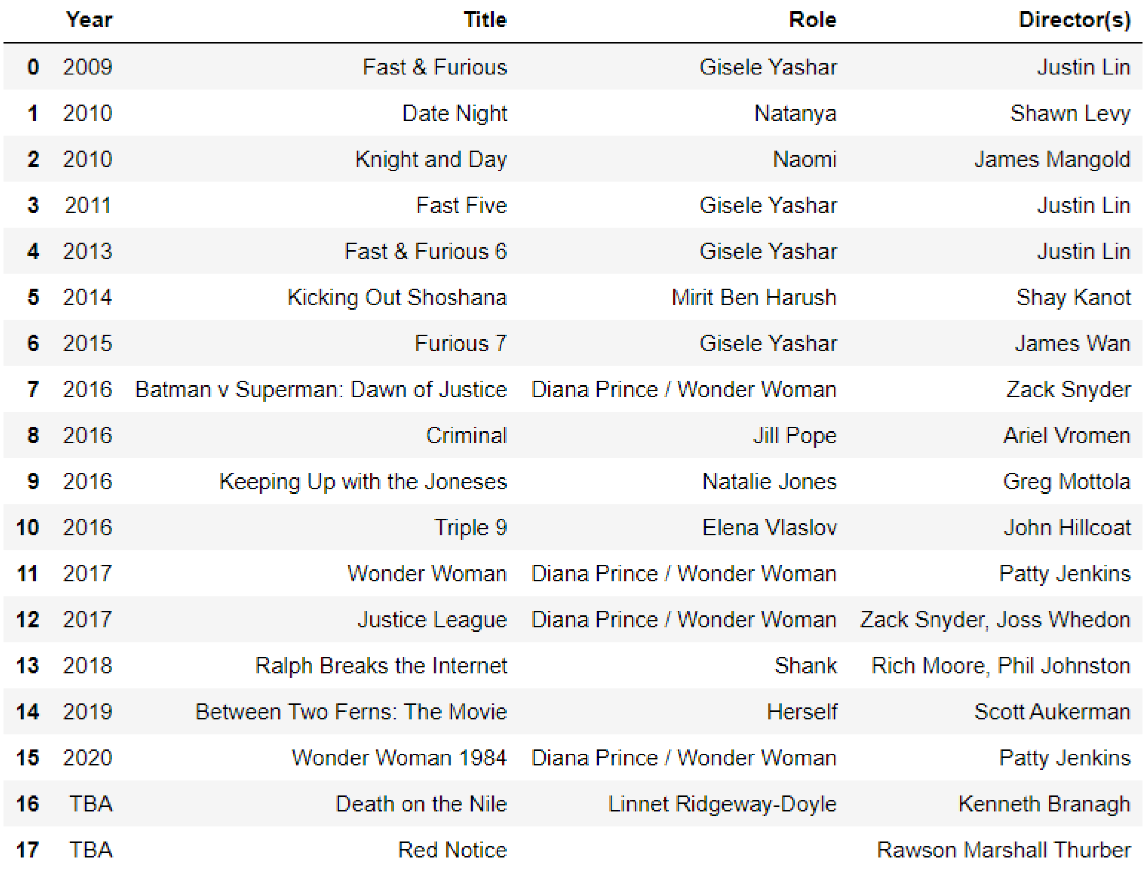
Homework 1: Web Scrapping

Question 1

Table with the films that Gal Gadot participated in (sorted by date):



	Year	Title	Role	Director(s)
0	2009	Fast & Furious	Gisele Yashar	Justin Lin
1	2010	Date Night	Natanya	Shawn Levy
2	2010	Knight and Day	Naomi	James Mangold
3	2011	Fast Five	Gisele Yashar	Justin Lin
4	2013	Fast & Furious 6	Gisele Yashar	Justin Lin
5	2014	Kicking Out Shoshana	Mirit Ben Harush	Shay Kanot
6	2015	Furious 7	Gisele Yashar	James Wan
7	2016	Batman v Superman: Dawn of Justice	Diana Prince / Wonder Woman	Zack Snyder
8	2016	Criminal	Jill Pope	Ariel Vromen
9	2016	Keeping Up with the Joneses	Natalie Jones	Greg Mottola
10	2016	Triple 9	Elena Vlaslov	John Hillcoat
11	2017	Wonder Woman	Diana Prince / Wonder Woman	Patty Jenkins
12	2017	Justice League	Diana Prince / Wonder Woman	Zack Snyder, Joss Whedon
13	2018	Ralph Breaks the Internet	Shank	Rich Moore, Phil Johnston
14	2019	Between Two Ferns: The Movie	Herself	Scott Aukerman
15	2020	Wonder Woman 1984	Diana Prince / Wonder Woman	Patty Jenkins
16	TBA	Death on the Nile	Linnet Ridgeway-Doyle	Kenneth Branagh
17	TBA	Red Notice		Rawson Marshall Thurber

Libraries we used: *urlopen*- to query a website, *BeautifulSoup* - to parse the data returned from the website, *pandas* - to convert list to data frame.

Functions we used: *get_column_values()* - # iterate over the table rows and get the required values, *print_table()* - # prints the table of the films that Gal Gadot participated in.

Explanation + Assumptions: First, we searched for the first table that its class is 'wikitable sortable' (the films table). Second, we iterated over the table rows (without the heading row) and for each row we looked for the 'td' tag. Finally, we split the row cells according to the required columns: year at index 0, title at index 1, roles at index 2 and directors at index 3.

Challenges:

- Some rows don’t have a year value in their tags and that’s because the year is equals to the previous year. In this case, we saved the year’s value of the previous row and used it for the next row (only when the row doesn’t have a year value).
- Some films had more than one director, so we concatenated them using ‘, ‘and we removed unnecessary words, e.g: ‘\n’, ‘(uncredited)\n’.
- We removed the ‘\n’ tag from every cell.

Question 2

Libraries we used: *re*- for regular expressions, and libraries from the first question.

Functions we used:

get_film_info() - get info about each film - while 'i' tag represents a film, 'li' tag represents an actor.

add_player(k,actor) - add actor values to the required lists.

get_player_info() - gets actor info from his page url.

get_year(player_soup) - gets the birth year from the actor’s page - assuming the actor has bday class.

get_country(player_soup) - gets actor's country of birth.

get_awards(player_soup) - get actor's number of awards.
get_number_of_awards_from_list_site(awards_url) - get actor's number of awards from a specific awards page.

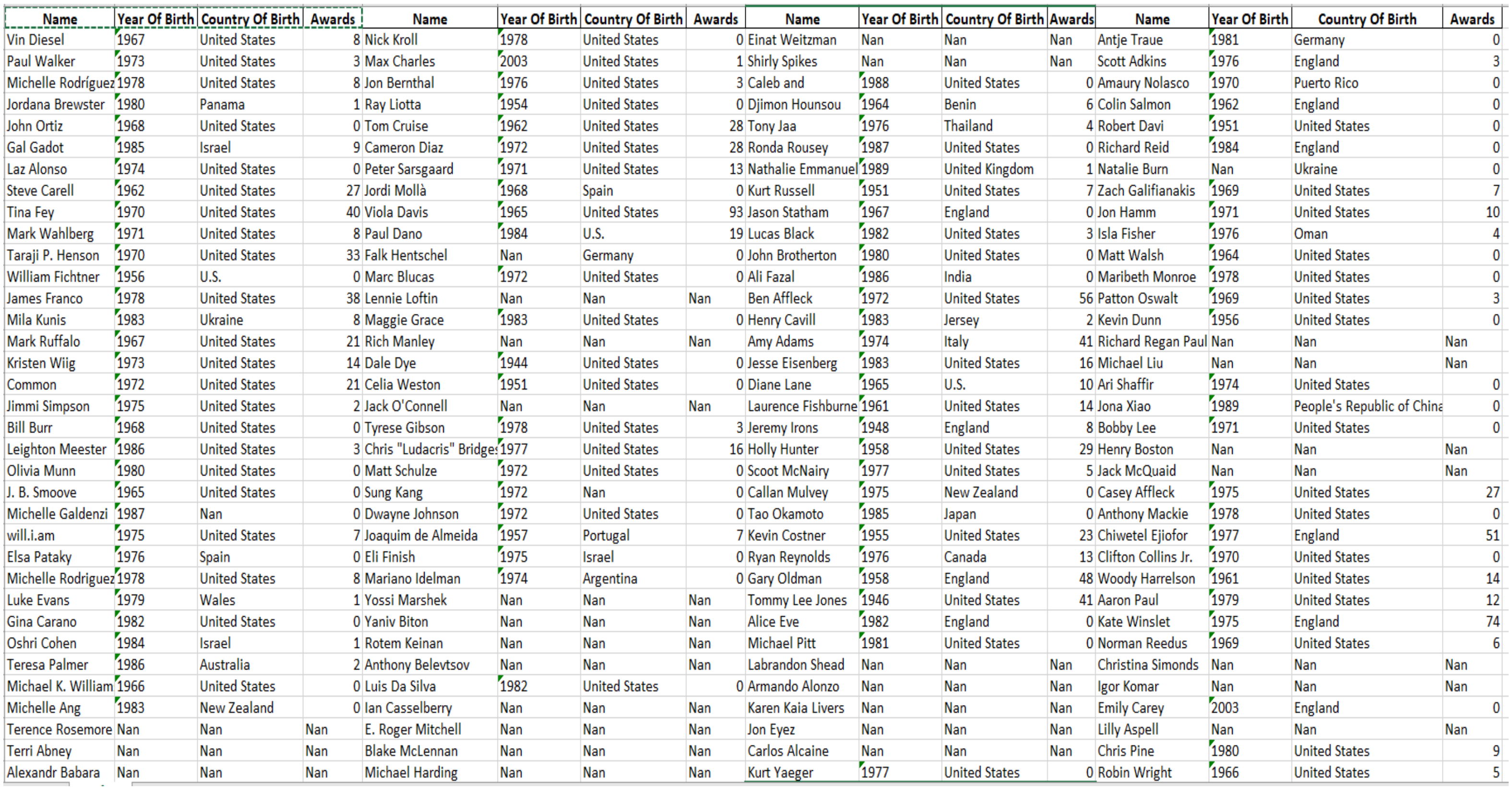
Explanation + Assumptions + Challenges:

1. We saw that each film has the ‘i’ tag, so we got the films_table from the first question and looked for all the ‘i' tags.
2. Then, we Iterated over each film, opened its page by its ‘a’ tag, and looked for the ‘Cast’ string. Each film has more than one ‘cast’ string (the first one is just a link to the cast area), so we needed to take the second object – cast[1].
3. After we found the cast area, we ran over its siblings(actors) and had to face with some cases:
   - Case 1 – the actors under <div><ul><li> tags
   - Case 2 – the actors under <ul><li> tags
   - Special case – movie number 7 has many <ul> tags till the <p> tag
   - Special case = movie number 14 has both of case 1 and case 2.

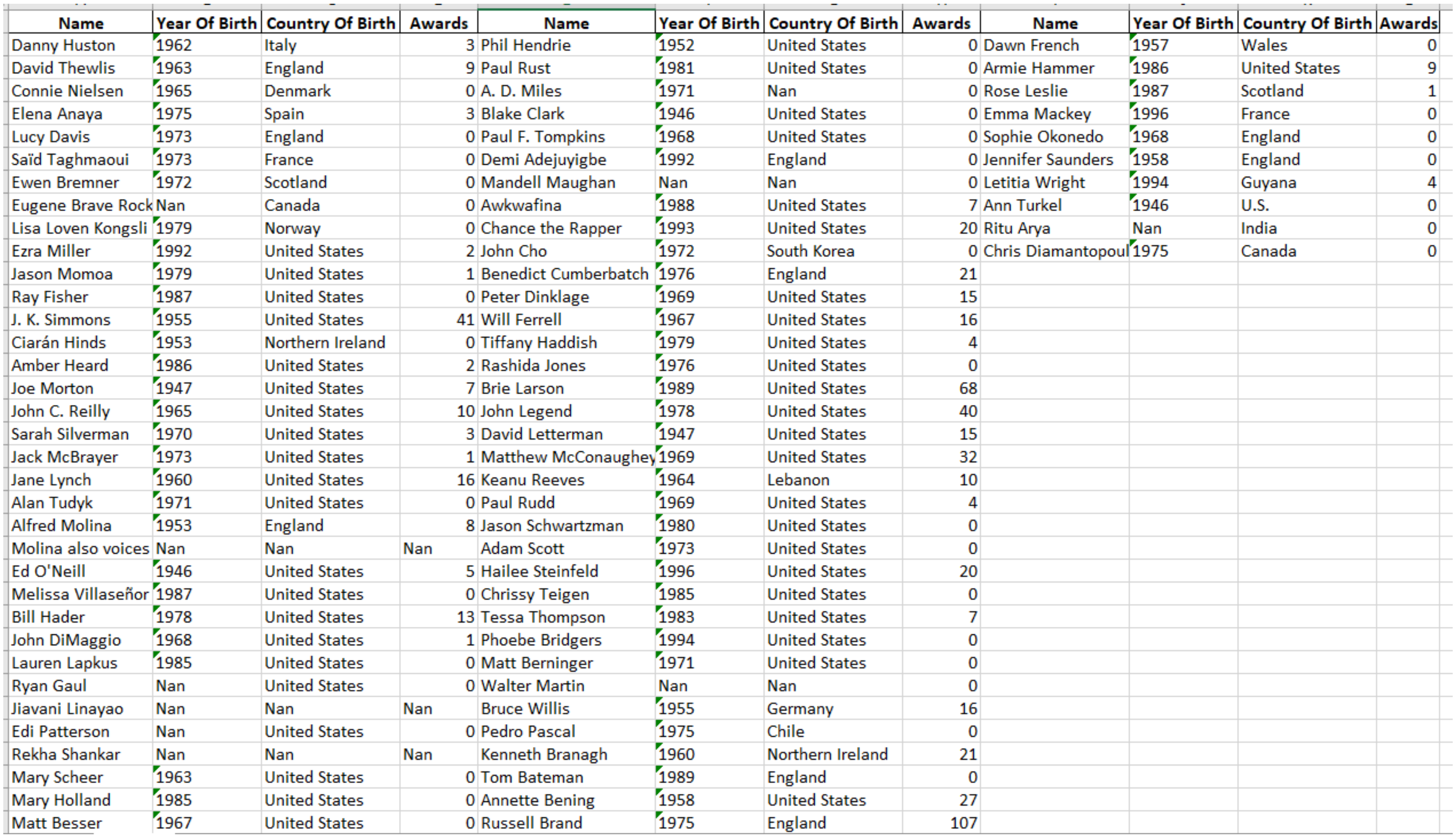
4. For each actor, we found his name, removed the ' as ' string, and added his name to the names array and also added him to the joint array - If the actor's name existed, we just added one to his joint movies.
5. We opened the actor's page (if the actor didn't have page we just added a Nan information about him), and first took his year of birth from the 'span' tag of the 'bday' class. Then we looked for his country of birth. We faced some case here:
    - Case 1 – the actor has 'div' tag of class 'birthplace'
    - Case 1.1 – the actor has url to his country – opened the url, and looked for 'Country' string. If it has that string, we took the country name and added it to the countries array. Else, we went back to the actor page and got the text from the country area by the 'city, country' convention.
    - Case 1.2 – the actor doesn't have url to his country - got the text from the country area by the 'city, country' convention.
    - Case 2 - the actor doesn't have 'div' tag of class 'birthplace', we looked for the 'bday' class by the assumption that the country should be after this class. And again like case 1.1.
6. In order to find the number of awards, we faced with some cases:
    - Case 1 - the awards are on his page in a table of class='yes table-yes2'
    - Case 2 - the actor has a link to a special awards page - assuming the title of the link is 'List of awards':
    - Case 2.1 – the awards page has info box so take the awards from there, then get the total number of awards from the awards tables by the 'td' tag of class 'yes', and finally get the maximum number of awards between the two of them.
    - Case 2.2 – the awards page doesn't contain info box so take number of awards from the awards tables.

    Summary: approximately 25~29 players don't have pages, so we only got their names and the other information about them is Nan. We tried to figure out the common patterns, and out of 219 actors only 15 (approx.) actors had bad scrapping ~ 93%.

    Output:



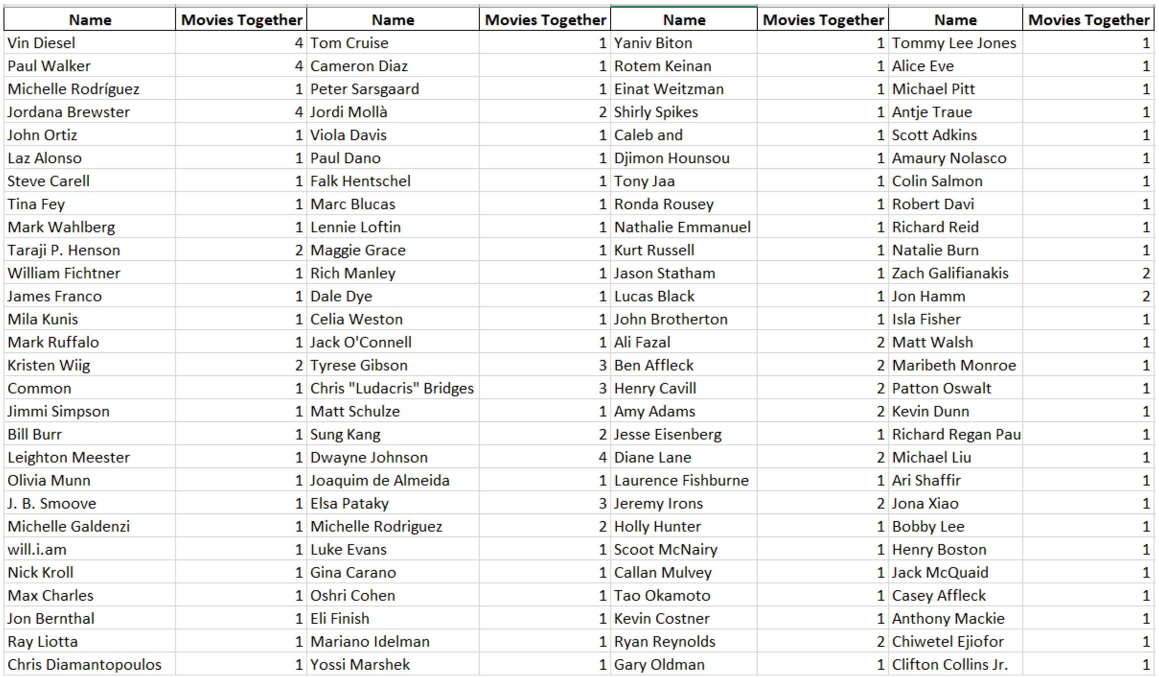
Name	Year Of Birth	Country Of Birth	Awards	Name	Year Of Birth	Country Of Birth	Awards	Name	Year Of Birth	Country Of Birth	Awards	Name	Year Of Birth	Country Of Birth	Awards
Vin Diesel	1967	United States	8	Nick Kroll	1978	United States	0	Einat Weitzman	Nan	Nan	Nan	Antje Traue	1981	Germany	0
Paul Walker	1973	United States	3	Max Charles	2003	United States	1	Shirly Spikes	Nan	Nan	Nan	Scott Adkins	1976	England	3
Michelle Rodríguez	1978	United States	8	Jon Bernthal	1976	United States	3	Caleb and	1988	United States	0	Amaury Nolasco	1970	Puerto Rico	0
Jordana Brewster	1980	Panama	1	Ray Liotta	1954	United States	0	Djimon Hounsou	1964	Benin	6	Colin Salmon	1962	England	0
John Ortiz	1968	United States	0	Tom Cruise	1962	United States	28	Tony Jaa	1976	Thailand	4	Robert Davi	1951	United States	0
Gal Gadot	1985	Israel	9	Cameron Diaz	1972	United States	28	Ronda Rousey	1987	United States	0	Richard Reid	1984	England	0
Laz Alonso	1974	United States	0	Peter Sarsgaard	1971	United States	13	Nathalie Emmanuel	1989	United Kingdom	1	Natalie Burn	Nan	Ukraine	0
Steve Carell	1962	United States	27	Jordi Mollà	1968	Spain	0	Kurt Russell	1951	United States	7	Zach Galifianakis	1969	United States	7
Tina Fey	1970	United States	40	Viola Davis	1965	United States	93	Jason Statham	1967	England	0	Jon Hamm	1971	United States	10
Mark Wahlberg	1971	United States	8	Paul Dano	1984	U.S.	19	Lucas Black	1982	United States	3	Isla Fisher	1976	Oman	4
Taraji P. Henson	1970	United States	33	Falk Hentschel	Nan	Germany	0	John Brotherton	1980	United States	0	Matt Walsh	1964	United States	0
William Fichtner	1956	U.S.	0	Marc Blucas	1972	United States	0	Ali Fazal	1986	India	0	Maribeth Monroe	1978	United States	0
James Franco	1978	United States	38	Lennie Loftin	Nan	Nan	Nan	Ben Affleck	1972	United States	56	Patton Oswalt	1969	United States	3
Mila Kunis	1983	Ukraine	8	Maggie Grace	1983	United States	0	Henry Cavill	1983	Jersey	2	Kevin Dunn	1956	United States	0
Mark Ruffalo	1967	United States	21	Rich Manley	Nan	Nan	Nan	Amy Adams	1974	Italy	41	Richard Regan Paul	Nan	Nan	Nan
Kristen Wiig	1973	United States	14	Dale Dye	1944	United States	0	Jesse Eisenberg	1983	United States	16	Michael Liu	Nan	Nan	Nan
Common	1972	United States	21	Celia Weston	1951	United States	0	Diane Lane	1965	U.S.	10	Ari Shaffir	1974	United States	0
Jimmi Simpson	1975	United States	2	Jack O'Connell	Nan	Nan	Nan	Laurence Fishburne	1961	United States	14	Jona Xiao	1989	People's Republic of China	0
Bill Burr	1968	United States	0	Tyrese Gibson	1978	United States	3	Jeremy Irons	1948	England	8	Bobby Lee	1971	United States	0
Leighton Meester	1986	United States	3	Chris "Ludacris" Bridges	1977	United States	16	Holly Hunter	1958	United States	29	Henry Boston	Nan	Nan	Nan
Olivia Munn	1980	United States	0	Matt Schulze	1972	United States	0	Scoot McNairy	1977	United States	5	Jack McQuaid	Nan	Nan	Nan
J. B. Smoove	1965	United States	0	Sung Kang	1972	Nan	0	Callan Mulvey	1975	New Zealand	0	Casey Affleck	1975	United States	27
Michelle Galdenzi	1987	Nan	0	Dwayne Johnson	1972	United States	0	Tao Okamoto	1985	Japan	0	Anthony Mackie	1978	United States	0
will.i.am	1975	United States	7	Joaquim de Almeida	1957	Portugal	7	Kevin Costner	1955	United States	23	Chiwetel Ejiofor	1977	England	51
Elsa Pataky	1976	Spain	0	Eli Finish	1975	Israel	0	Ryan Reynolds	1976	Canada	13	Clifton Collins Jr.	1970	United States	0
Michelle Rodriguez	1978	United States	8	Mariano Idelman	1974	Argentina	0	Gary Oldman	1958	England	48	Woody Harrelson	1961	United States	14
Luke Evans	1979	Wales	1	Yossi Marshek	Nan	Nan	Nan	Tommy Lee Jones	1946	United States	41	Aaron Paul	1979	United States	12
Gina Carano	1982	United States	0	Yaniv Biton	Nan	Nan	Nan	Alice Eve	1982	England	0	Kate Winslet	1975	England	74
Oshri Cohen	1984	Israel	1	Rotem Keinan	Nan	Nan	Nan	Michael Pitt	1981	United States	0	Norman Reedus	1969	United States	6
Teresa Palmer	1986	Australia	2	Anthony Belevtsov	Nan	Nan	Nan	Labrandon Shead	Nan	Nan	Nan	Christina Simonds	Nan	Nan	Nan
Michael K. William	1966	United States	0	Luis Da Silva	1982	United States	0	Armando Alonzo	Nan	Nan	Nan	Igor Komar	Nan	Nan	Nan
Michelle Ang	1983	New Zealand	0	Ian Casselberry	Nan	Nan	Nan	Karen Kaia Livers	Nan	Nan	Nan	Emily Carey	2003	England	0
Terence Rosemore	Nan	Nan	Nan	E. Roger Mitchell	Nan	Nan	Nan	Jon Eyez	Nan	Nan	Nan	Lilly Aspell	Nan	Nan	Nan
Terri Abney	Nan	Nan	Nan	Blake McLennan	Nan	Nan	Nan	Carlos Alcaine	Nan	Nan	Nan	Chris Pine	1980	United States	9
Alexandr Babara	Nan	Nan	Nan	Michael Harding	Nan	Nan	Nan	Kurt Yaeger	1977	United States	0	Robin Wright	1966	United States	5



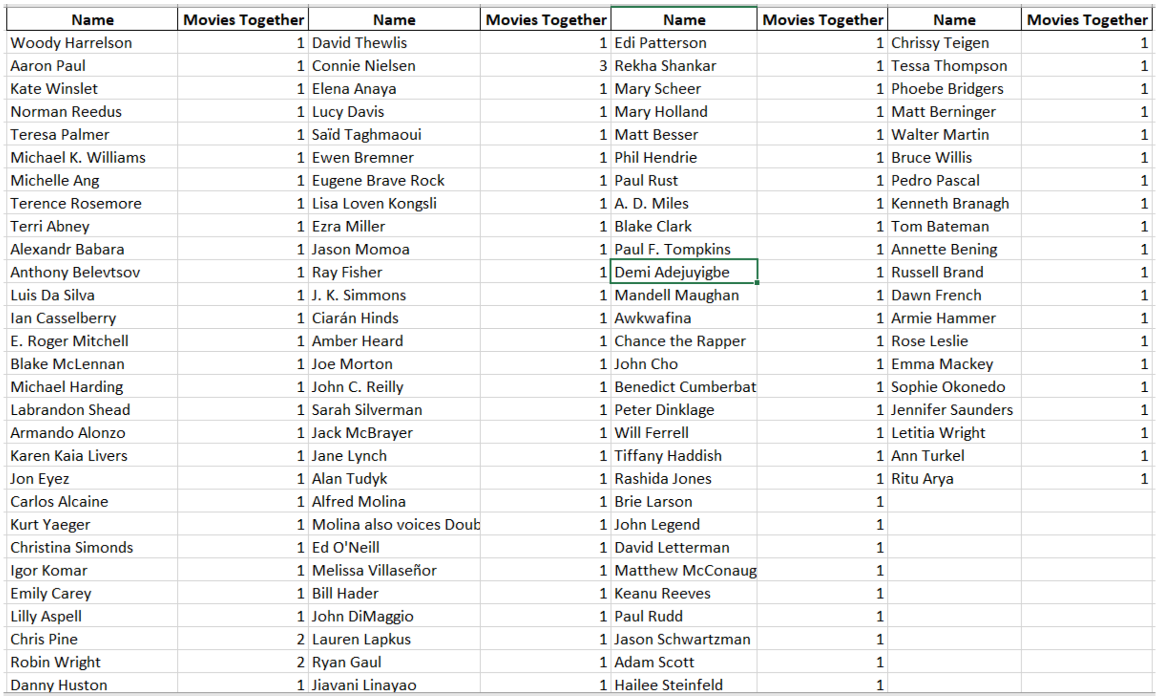
Name	Year Of Birth	Country Of Birth	Awards	Name	Year Of Birth	Country Of Birth	Awards	Name	Year Of Birth	Country Of Birth	Awards
Danny Huston	1962	Italy	3	Phil Hendrie	1952	United States	0	Dawn French	1957	Wales	0
David Thewlis	1963	England	9	Paul Rust	1981	United States	0	Armie Hammer	1986	United States	9
Connie Nielsen	1965	Denmark	0	A. D. Miles	1971	Nan	0	Rose Leslie	1987	Scotland	1
Elena Anaya	1975	Spain	3	Blake Clark	1946	United States	0	Emma Mackey	1996	France	0
Lucy Davis	1973	England	0	Paul F. Tompkins	1968	United States	0	Sophie Okonedo	1968	England	0
Saïd Taghmaoui	1973	France	0	Demi Adejuyigbe	1992	England	0	Jennifer Saunders	1958	England	0
Ewen Bremner	1972	Scotland	0	Mandell Maughan	Nan	Nan	0	Letitia Wright	1994	Guyana	4
Eugene Brave Rock	Nan	Canada	0	Awkwafina	1988	United States	7	Ann Turkel	1946	U.S.	0
Lisa Loven Kongsli	1979	Norway	0	Chance the Rapper	1993	United States	20	Ritu Arya	Nan	India	0
Ezra Miller	1992	United States	2	John Cho	1972	South Korea	0	Chris Diamantopoul	1975	Canada	0
Jason Momoa	1979	United States	1	Benedict Cumberbatch	1976	England	21				
Ray Fisher	1987	United States	0	Peter Dinklage	1969	United States	15				
J. K. Simmons	1955	United States	41	Will Ferrell	1967	United States	16				
Ciarán Hinds	1953	Northern Ireland	0	Tiffany Haddish	1979	United States	4				
Amber Heard	1986	United States	2	Rashida Jones	1976	United States	0				
Joe Morton	1947	United States	7	Brie Larson	1989	United States	68				
John C. Reilly	1965	United States	10	John Legend	1978	United States	40				
Sarah Silverman	1970	United States	3	David Letterman	1947	United States	15				
Jack McBrayer	1973	United States	1	Matthew McConaughey	1969	United States	32				
Jane Lynch	1960	United States	16	Keanu Reeves	1964	Lebanon	10				
Alan Tudyk	1971	United States	0	Paul Rudd	1969	United States	4				
Alfred Molina	1953	England	8	Jason Schwartzman	1980	United States	0				
Molina also voices	Nan	Nan	Nan	Adam Scott	1973	United States	0				
Ed O'Neill	1946	United States	5	Hailee Steinfeld	1996	United States	20				
Melissa Villaseñor	1987	United States	0	Chrissy Teigen	1985	United States	0				
Bill Hader	1978	United States	13	Tessa Thompson	1983	United States	7				
John DiMaggio	1968	United States	1	Phoebe Bridgers	1994	United States	0				
Lauren Lapkus	1985	United States	0	Matt Berninger	1971	United States	0				
Ryan Gaul	Nan	United States	0	Walter Martin	Nan	Nan	0				
Jiavani Linayao	Nan	Nan	Nan	Bruce Willis	1955	Germany	16				
Edi Patterson	Nan	United States	0	Pedro Pascal	1975	Chile	0				
Rekha Shankar	Nan	Nan	Nan	Kenneth Branagh	1960	Northern Ireland	21				
Mary Scheer	1963	United States	0	Tom Bateman	1989	England	0				
Mary Holland	1985	United States	0	Annette Bening	1958	United States	27				
Matt Besser	1967	United States	0	Russell Brand	1975	England	107				

Question 3

- Table of the number of joint movies for each co-actor\actress with Gal Gadot:



Name	Movies Together	Name	Movies Together	Name	Movies Together	Name	Movies Together
Vin Diesel	4	Tom Cruise	1	Yaniv Biton	1	Tommy Lee Jones	1
Paul Walker	4	Cameron Diaz	1	Rotem Keinan	1	Alice Eve	1
Michelle Rodríguez	1	Peter Sarsgaard	1	Einat Weitzman	1	Michael Pitt	1
Jordana Brewster	4	Jordi Mollà	2	Shirly Spikes	1	Antje Traue	1
John Ortiz	1	Viola Davis	1	Caleb and	1	Scott Adkins	1
Laz Alonso	1	Paul Dano	1	Djimon Hounsou	1	Amaury Nolasco	1
Steve Carell	1	Falk Hentschel	1	Tony Jaa	1	Colin Salmon	1
Tina Fey	1	Marc Blucas	1	Ronda Rousey	1	Robert Davi	1
Mark Wahlberg	1	Lennie Loftin	1	Nathalie Emmanuel	1	Richard Reid	1
Taraji P. Henson	2	Maggie Grace	1	Kurt Russell	1	Natalie Burn	1
William Fichtner	1	Rich Manley	1	Jason Statham	1	Zach Galifianakis	2
James Franco	1	Dale Dye	1	Lucas Black	1	Jon Hamm	2
Mila Kunis	1	Celia Weston	1	John Brotherton	1	Isla Fisher	1
Mark Ruffalo	1	Jack O'Connell	1	Ali Fazal	2	Matt Walsh	1
Kristen Wiig	2	Tyrese Gibson	3	Ben Affleck	2	Maribeth Monroe	1
Common	1	Chris "Ludacris" Bridges	3	Henry Cavill	2	Patton Oswalt	1
Jimmi Simpson	1	Matt Schulze	1	Amy Adams	2	Kevin Dunn	1
Bill Burr	1	Sung Kang	2	Jesse Eisenberg	1	Richard Regan Pau	1
Leighton Meester	1	Dwayne Johnson	4	Diane Lane	2	Michael Liu	1
Olivia Munn	1	Joaquim de Almeida	1	Laurence Fishburne	1	Ari Shaffir	1
J. B. Smoove	1	Elsa Pataky	3	Jeremy Irons	2	Jona Xiao	1
Michelle Galdenzi	1	Michelle Rodriguez	2	Holly Hunter	1	Bobby Lee	1
will.i.am	1	Luke Evans	1	Scoot McNairy	1	Henry Boston	1
Nick Kroll	1	Gina Carano	1	Callan Mulvey	1	Jack McQuaid	1
Max Charles	1	Oshri Cohen	1	Tao Okamoto	1	Casey Affleck	1
Jon Bernthal	1	Eli Finish	1	Kevin Costner	1	Anthony Mackie	1
Ray Liotta	1	Mariano Idelman	1	Ryan Reynolds	2	Chiwetel Ejiofor	1
Chris Diamantopoulos	1	Yossi Marshek	1	Gary Oldman	1	Clifton Collins Jr.	1



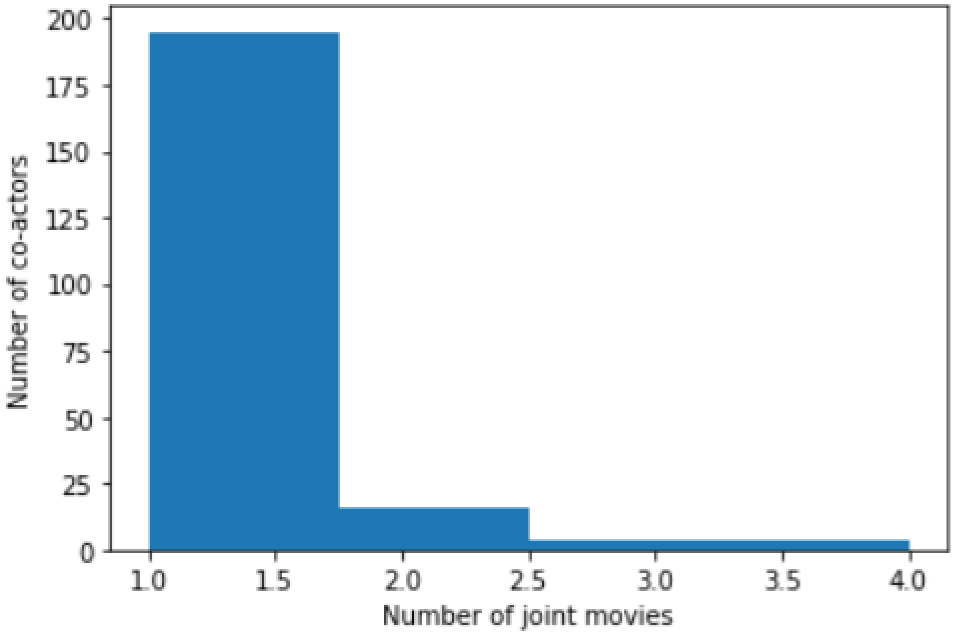
Name	Movies Together	Name	Movies Together	Name	Movies Together	Name	Movies Together
Woody Harrelson	1	David Thewlis	1	Edi Patterson	1	Chrissy Teigen	1
Aaron Paul	1	Connie Nielsen	3	Rekha Shankar	1	Tessa Thompson	1
Kate Winslet	1	Elena Anaya	1	Mary Scheer	1	Phoebe Bridgers	1
Norman Reedus	1	Lucy Davis	1	Mary Holland	1	Matt Berninger	1
Teresa Palmer	1	Saïd Taghmaoui	1	Matt Besser	1	Walter Martin	1
Michael K. Williams	1	Ewen Bremner	1	Phil Hendrie	1	Bruce Willis	1
Michelle Ang	1	Eugene Brave Rock	1	Paul Rust	1	Pedro Pascal	1
Terence Rosemore	1	Lisa Loven Kongsli	1	A. D. Miles	1	Kenneth Branagh	1
Terri Abney	1	Ezra Miller	1	Blake Clark	1	Tom Bateman	1
Alexandr Babara	1	Jason Momoa	1	Paul F. Tompkins	1	Annette Bening	1
Anthony Belevtsov	1	Ray Fisher	1	Demi Adejuyigbe	1	Russell Brand	1
Luis Da Silva	1	J. K. Simmons	1	Mandell Maughan	1	Dawn French	1
Ian Casselberry	1	Ciarán Hinds	1	Awkwafina	1	Armie Hammer	1
E. Roger Mitchell	1	Amber Heard	1	Chance the Rapper	1	Rose Leslie	1
Blake McLennan	1	Joe Morton	1	John Cho	1	Emma Mackey	1
Michael Harding	1	John C. Reilly	1	Benedict Cumberbat	1	Sophie Okonedo	1
Labrandon Shead	1	Sarah Silverman	1	Peter Dinklage	1	Jennifer Saunders	1
Armando Alonzo	1	Jack McBrayer	1	Will Ferrell	1	Letitia Wright	1
Karen Kaia Livers	1	Jane Lynch	1	Tiffany Haddish	1	Ann Turkel	1
Jon Eyez	1	Alan Tudyk	1	Rashida Jones	1	Ritu Arya	1
Carlos Alcaine	1	Alfred Molina	1	Brie Larson	1		
Kurt Yaeger	1	Molina also voices Doub	1	John Legend	1		
Christina Simonds	1	Ed O'Neill	1	David Letterman	1		
Igor Komar	1	Melissa Villaseñor	1	Matthew McConaug	1		
Emily Carey	1	Bill Hader	1	Keanu Reeves	1		
Lilly Aspell	1	John DiMaggio	1	Paul Rudd	1		
Chris Pine	2	Lauren Lapkus	1	Jason Schwartzman	1		
Robin Wright	2	Ryan Gaul	1	Adam Scott	1		
Danny Huston	1	Jiavani Linayao	1	Hailee Steinfeld	1		

Functions we used: *show_joint_table()* - table which presents the number of joint movies for each co-actor\actress with Gal Gadot.

Explanation: in the second question, while we were iterating over the actors we count how many joint movies they have with Gal Gadot – We saved this counter in the ‘joint’ array. we checked every actor if he/she already exists in the 'names' array and if he/she does exist we added +1 to the number of joint movies of this player. In order to create the table in this question, we used that 'joint' array and also the 'names' array that contains the actor's names. We made a Dataframe from both of these arrays and displayed the required table – as you can see upward.

Challenges – We solved the challenges by counting the number of joint movies in the second question.

- Histogram which presents the distribution of joint movies:



Libraries we used: *<u>pyplot</u>*- to plot histograms.

Functions we used: *<u>show_histogram()</u>* - show histogram which presents the distribution of joint movies.

Explanation – First, we calculated the length of the 'joint' array in order to know how to how many bins should we divide the histogram. Finally, we made a histogram from that 'joint' array – as you can see upward.