ASSIGNMENT 3

DEADLINE APRIL 5, 2017

NYC VEHICLE COLLISION ANALYSIS

Q1_PART ONE

- Use ‘vehicle_collisions’ data set.
- For each month in 2016, find out the percentage of collisions in Manhattan out of that year's total accidents in New York City.
- Display a few rows of the output **use df.head().**
- Generate a csv output with four columns (‘Month’, ‘Manhattan’, ‘NYC’, ‘Percentage’)
- Example output:

MONTH	MANHATTAN	NYC	PERCENTAGE
Jan	3178	18101	0.17557041
Feb	3195	15985	0.199874883

NYC VEHICLE COLLISION ANALYSIS

Q1_PART TWO

- Use ‘vehicle_collisions’ data set.
- For each borough, find out distribution of each collision scale. (One car involved? Two? Three? or more?) (From 2015 to present)
- Display a few rows of the output **use df.head().**
- Generate a csv output with five columns ('borough', 'one-vehicle', 'two-vehicles', 'three-vehicles', 'more-vehicles')
- Example output:

BOROUGH	ONE_VEHICLE_INVOLVED	TWO_VEHICLES_INVOLVED	THREE_VEHICLES_INVOLVED	MORE_VEHICLES_INVOLVED
QUEENS	12962	70260	4498	1935

EMPLOYEE COMPENSATION ANALYSIS

Q2_PART ONE

- Use 'employee_compensation' data set.
- Find out the highest paid departments in each organization group by calculating mean of total compensation for every department.
- Output should contain the organization group and the departments in each organization group with the total compensation **from highest to lowest** value.
- Display a few rows of the output **use df.head().**
- Generate a csv output.
- Example output:

		Total Compensation
Community Health	**Public Health**	96190.190140

EMPLOYEE COMPENSATION ANALYSIS

Q2_PART TWO

- Use 'employee_compensation' data set.
- Data contains fiscal and calendar year information. Same employee details exist twice in the dataset. Filter data by calendar year and find average salary (you might have to find average for each of the columns for every employee. Eg. Average of Total Benefits, Average of total compensation etc.) for every employee.
- Now, find the people whose overtime salary is greater than 5% of salaries (salaries refers to 'Salaries' column)
- For each 'Job Family' these people are associated with, calculate the percentage of total benefits with respect to total compensation (so for each job family you have to calculate average total benefits and average total compensation). Create a new column to hold the percentage value.
- **Display** the top 5 Job Families according to this percentage value using **df.head()**.
- Write the output (jobs and percentage value) to a csv.
- Example output:

	Total Benefits	Total Compensation	Percent_Total_Benefit
Job Family			
Public Service Aide	10000.000000	100000.000000	10.00

CRICKET MATCHES ANALYSIS

Q3_PART ONE

- Use ‘cricket_matches’ data set.
- Calculate the average score for each team which host the game and win the game.
- Remember that if a team hosts a game and wins the game, their score can be innings_1 runs or innings_2 runs. You have to check if the host team won the game, check which innings they played in (innings_1 or innings_2), and take the runs scored in that innings. The final answer is the average score of each team satisfying the above condition.
- Display a few rows of the output **use df.head()**
- Generate a csv output
- Example output:

	home	Score
0	Abahani Limited	200.000000

MOVIE AWARDS ANALYSIS

Q4_PART ONE

- Use ‘movies_awards’ data set.
- You are supposed to extract data from the awards column in this dataset and split it into several columns. An example is given below.
- The awards has details of wins, nominations in general and also wins and nominations in certain categories(e.g. Oscar, BAFTA etc.)
- You are supposed to create a win and nominated column (these 2 columns contain total number of wins and nominations) and other columns that extract the number of wins and nominations for each category of award.
- If a movie has 2 Oscar nominations and 4 Oscar won, the columns Oscar_Awards_Won should have value 4 and Oscar_Awards_Nominated should have value 2. You should also have a total won and nominated column which aggregates all the awards (won or nominated).
- Create two separate columns for each award category (won and nominated).
- Write your output to a csv file. (Sample output is given in next page)

SAMPLE OUTPUT FOR Q4

E	F	G	H	I	J	K	L	M	N	O
Awards	Awards_won	Awards_nominated	Prime_Awards_nominated	Oscar_Awards_nominated	Golden_Glob	BAFTA_Awar	Prime_Awards_won	Oscar_Awards_won	Golden_Globe_Awards_won	BAFTA_Awards_v
1 win & 2 nominations.	1	2	0	0	0	0	0	0	0	0
1 win.	1	0	0	0	0	0	0	0	0	0
1 nomination.	0	1	0	0	0	0	0	0	0	0
3 wins & 2 nominations.	3	2	0	0	0	0	0	0	0	0
1 win & 1 nomination.	1	1	0	0	0	0	0	0	0	0
5 nominations.	0	5	0	0	0	0	0	0	0	0
2 wins & 5 nominations.	2	5	0	0	0	0	0	0	0	0
Nominated for 1 Oscar. Another 13 wins & 13 nominations.	13	14	0	1	0	0	0	0	0	0
5 wins & 4 nominations.	5	4	0	0	0	0	0	0	0	0
Nominated for 1 Golden Globe. Another 14 wins & 34 nominations.	14	35	0	0	1	0	0	0	0	0
3 wins & 2 nominations.	3	2	0	0	0	0	0	0	0	0
1 nomination.	0	1	0	0	0	0	0	0	0	0
5 nominations.	0	5	0	0	0	0	0	0	0	0
1 win & 2 nominations.	1	2	0	0	0	0	0	0	0	0
22 wins & 44 nominations.	22	44	0	0	0	0	0	0	0	0
1 win & 3 nominations.	1	3	0	0	0	0	0	0	0	0
1 win & 8 nominations.	1	8	0	0	0	0	0	0	0	0
2 nominations.	0	2	0	0	0	0	0	0	0	0
6 wins & 1 nomination.	6	1	0	0	0	0	0	0	0	0
Won 2 Golden Globes. Another 68 wins & 201 nominations.	70	201	0	0	0	0	0	0	2	0
Won 1 Golden Globe. Another 31 wins & 143 nominations.	32	143	0	0	0	0	0	0	1	0
Nominated for 4 Primetime Emmys. Another 16 nominations.	0	20	4	0	0	0	0	0	0	0
3 wins & 2 nominations.	3	2	0	0	0	0	0	0	0	0
3 wins & 6 nominations.	3	6	0	0	0	0	0	0	0	0
2 wins.	2	0	0	0	0	0	0	0	0	0
6 wins & 4 nominations.	6	4	0	0	0	0	0	0	0	0
7 wins & 6 nominations.	7	6	0	0	0	0	0	0	0	0
9 wins & 30 nominations.	9	30	0	0	0	0	0	0	0	0
Won 1 Primetime Emmy. Another 1 win & 6 nominations.	2	6	0	0	0	0	1	0	0	0
1 nomination.	0	1	0	0	0	0	0	0	0	0
Nominated for 3 BAFTA Film Awards. Another 11 wins & 39 nomina	11	42	0	0	0	3	0	0	0	0
6 nominations.	0	6	0	0	0	0	0	0	0	0
1 win & 4 nominations.	1	4	0	0	0	0	0	0	0	0
2 wins & 5 nominations.	2	5	0	0	0	0	0	0	0	0
14 wins & 11 nominations.	14	11	0	0	0	0	0	0	0	0
1 win & 1 nomination.	1	1	0	0	0	0	0	0	0	0

BEFORE YOU SUBMIT, PLEASE CHECK BELOW !!!

- You MUST use pandas DataFrame to process your work.
- Please create a new file for each of the question. You should have SIX code files.
- You have to generate one csv file for each question. You should have SIX csv files. And your output csv file MUST be similar as the given example.
- You need to print some proper output in your code by using df.head().
- You have to write proper comments for each question.
- Please use relative path to read csv files.
- Your submissions are suppose to be as .py files. [No .ipynb files are acceptable] and your file names should have the question and part number (e.g. Q1_Part_1)
- Your submissions should be executable without any errors.
- No submission after the deadline will be evaluated.