Openrefine Airbnb Submission and Autograding

What to submit

You need to submit two files:

1. `airbnb_clean.csv`: Export clean file to from Openrefine and name it airbnb_clean.csv.
2. `airbnb_recipe.json`: Export cleaning recipe as json to file named airbnb_recipe.json.

How to submit

We have created a new assignment called openrefine_02. Fetch it like you have fetched other assignments. Fetching it will copy the assignment to your home directory at path /home/<netid>/openrefine_02. To submit you need to upload those two files in this folder and submit like you have submitted previous assignments.
Please note that this assignment has a dummy notebook called IgnoreMe.ipynb. You can ignore this notebook. This is just there because our infrastructure needs each assignment to have at least one notebook.

Autograding

Autograding for this assignment will work differently from other assignments.

Our autograder will test two aspects of your submission:

1. **Clean file:** We will compare your airbnb_clean.csv with expected clean file and count number of correct cells for each column. Your score for that column will be weighted by fraction of correct cells in that column. Scores for each column are different based on hardness level of the task. Table below lists maximum score for each column.
2. **Recipe file:** We will also run airbnb_recipe.json against the original dirty file to produce new clean file. Ideally, it should produce same clean file as submitted by you but sometimes it doesn't. As long as it is very close if not identical to clean file submitted by you, we will not deduct any score. If it is significantly different then that is an indicator of something wrong. We will review it case by case in such scenarios. You will get to know beforehand what clean file your recipe is producing and if it is different from clean file submitted by you.

Daily autograder run

Openrefine is a very resource intensive tool. Our server will not be able to sustain 30 students running Openrefine simultaneously. As a workaround, we ask you to work on assignment by installing Openrefine on your personal computer like previous assignment and submit clean file and json recipe file. We will then grade it once a day at 6pm and share scores and other feedback with you. We will do enough grader runs before deadline so you can see how you are doing daily. We will then do a final run after submission deadline that will determine your final score.

Autograder will write your scores and other feedback to files in the same folder i.e. in folder /home/<netid>/openrefine_02. Following are the specific files that will be written by autograder.

1. airbnb_cleanfile_grade.csv: This will contain your grade for each column and also your cumulative grade.
2. airbnb_recipe_clean.csv: This is the file produced by running your recipe on the dirty file.
3. airbnb_recipe_diff.xlsx: Ideally, airbnb_recipe_clean.csv should be identical to airbnb_clean.csv. In case it is not then airbnb_recipe_diff.xlsx will contain columns from both files juxtaposed next to each other and color highlighted to show differences. Green color will be for cells from airbnb_clean.csv and red will be for cells airbnb_recipe_clean.csv. Last row of this file will also list total number of mismatches. In case you see even a single highlighted cell in this file, please contact course staff for further assistance.
4. airbnb_recipe_match.csv: This file will contain column wise statistics for mismatch between airbnb_recipe_clean.csv and airbnb_clean.csv

In case there was an error while running autograder, it will write to files airbnb_cleanfile_run.err and airbnb_recipe_run.err. Please contact us if you see these files after autograder run and we will investigate further.

column names	scores	associated tasks
name	2	trim spaces
name_grel	10	GREL: remove outer parentheses
name_grel_star	10	GREL: remove exclamation marks and asterisks
host_id	1	to Number
host_name	2	trim spaces
host_name 1	6	Split column using regex And
host_name 2	6	Split column using regex And
neighbourhood	2	trim spaces
neighbourhood_case	5	to titlecase
neighbourhood_loop	6	edit 'Loop'
neighbourhood_cluster	10	cluster 'O'Hare' and 'West Garfield Park'
latitude	1	to Number
longitude	1	to Number
room_type	2	trim spaces
price	1	to Number
price_crazy	10	Numeric facets >$5000
minimum_nights	1	to Number
minimum_nights_long	10	Numeric facets > 300 nights
number_of_reviews	1	to Number
last_review	4	to Date
last_review_timeless	6	GREL: edit time
reviews_per_month	1	to Number
calculated_host_listings	1	to Number
availability_365	1	to Number
Total	**100**	