**Assignment #3**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| This assignment is about analyzing San Francisco restaurant inspection data, which consists of three separate data sets.  The data sets available at the bottom of this page.  ***The columns in these data files are tab separated.***  Below is the information for each of the data sets.  Make sure to clean up the data before you perform analysis o them.   |  |  | | --- | --- | | **Name** | **Columns** | | Business | business\_id, name, address, city, postal\_code, latitude, longitude, phone\_number, tax\_code, business\_certificate, application\_date, owner\_name, owner\_address, owner\_city, owner\_state, owner\_zip | | Inspection | business\_id, score, date, type | | Violation | business\_id, date, violationTypeID, risk\_category, description |   Questions:   1. What is the inspection score distribution like? (inspections\_plus.csv)    * Count score frequency - how many occurrences of score 100, 99, 98, etc. 2. What is the risk category distribution like? (violations\_plus.csv) 3. Which 20 businesses got lowest scores? (inspections\_plus.csv, businesses\_plus.csv)    * "business\_id","name","address","city","postal\_code","latitude",”longitude”, score 4. Which 20 businesses got highest score? (inspections\_plus.csv, businesses\_plus.csv)    * "business\_id","name","address","city","postal\_code","latitude",”longitude”, score 5. Among all the restaurants that got 100 score, what kind of high-risk violations did they get (if any)?    * (inspections\_plus.csv, violations\_plus.csv) 6. Average inspection score by zip code    * (inspections\_plus.csv, businesses\_plus.csv) 7. Compute the proportion of all businesses in each neighborhood (zip code) that have incurred at least one of the violations on this list    * "High risk vermin infestation"    * "Moderate risk vermin infestation"    * "Sewage or wastewater contamination”    * "Improper food labeling or menu misrepresentation"    * "Contaminated or adulterated food”    * "Reservice of previously served foods" 8. Are SF restaurants clean? Justify your answer 9. Extra Credit    * Leverage map feature in spark-notebook to plot the businesses by longitude and latitude   Assignment #4 Databrick notebook Template:   * Copy the following URL and import it into your DataBricks account   + Make sure to copy the complete URL * <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4338404698899132/1148062494309924/2419783655524824/latest.html>   Resources:   * [SF Open Data - Restaurant Scores](https://data.sfgov.org/Health-and-Social-Services/Restaurant-Scores-LIVES-Standard/pyih-qa8i) * [How to Data (Science): Mapping SF Restaurant Inspection Scores](http://www.zipfianacademy.com/blog/post/57158627293/how-to-data-science-mapping-sf-restaurant-inspection) * [SF OPEN DATA: YOUR FAVORITE RESTAURANTS ARE FILTHY](https://reasonengine.wordpress.com/2013/09/17/hells-kitchens/) * [Happy Healthy Hungry -- San Francisco](http://nbviewer.ipython.org/github/Jay-Oh-eN/happy-healthy-hungry/blob/master/h3.ipynb) |

Subpages (1): [Assignment #4 Output](https://sites.google.com/site/continuelearning/Home/introduction-to-spark-with-scala/assignment-3/assignment)