Data Analyst Intern Assignment

Data sets:

1. Legally Operating Businesses: https://data.cityofnewyork.us/Business/Legally-Operating-Businesses/w7w3-xahh
2. License Applications: https://data.cityofnewyork.us/Business/License-Applications/ptev-4hud

Questions:

1. What are the mean DCA license turnaround times for issued applications (incl. renewal) for each of the 5 NYC boroughs in 2017? If you need additional reference data sets, please provide a link / description for each and include as part of your submission.
2. For the borough with the highest value determined in the previous question, which license category is responsible for the longest median turnaround time for application decisions in 2016?
3. Find the borough with the highest rate of application denials and the one with the highest rate of renewal denials in 2018. List the rates for all boroughs.
4. For currently active garage or parking-related businesses that are going to have their license expire in the 2nd half of 2018 or in 2019, give the last license application activity date. Save the output as a CSV file. The output report should contain license number, license expiration date, license status, business category and last license activity date.

Notes:

1. Your submission should consist of Python code (Python 3.6), associated documentation, output files as detailed in the questions and any additional datasets required.
2. Do not include the 2 source data sets listed above in your submission. It is assumed that they will be located in your project directory under "./data/" with no filename changes.
3. The use of numpy and pandas is allowed.
4. The use of notebooks (Jupyter or Zeppelin) is allowed, if so desired.
5. An application is considered "in" a year if its submission/start date falls in that year.

Assessment criteria:

1. Data analysis:
    - Describe assumptions (if any)
    - Explain rationale for handling missing data (if any)
    - Deliver accurate analysis and responses to assessment questions
    - Draw observations and/or conclusions from results (if any)

2. Coding fluency:
    - Document code (inline comments and a README file)
    - Use descriptive variable and function names
    - Write modular functions
3. Creativity and initiative:
    - Leverage data sets not provided by the assignment, if required, with reference and rationale
    - Demonstrate ability to join dirty data sets in the absence of explicit primary keys
    - Highlight other issues identified in the source datasets