COMP 5070
Statistical Programming for Data Science
New York Movies Scene

Assignment 1 DUE by 11:00pm (CST), Sunday 17 April

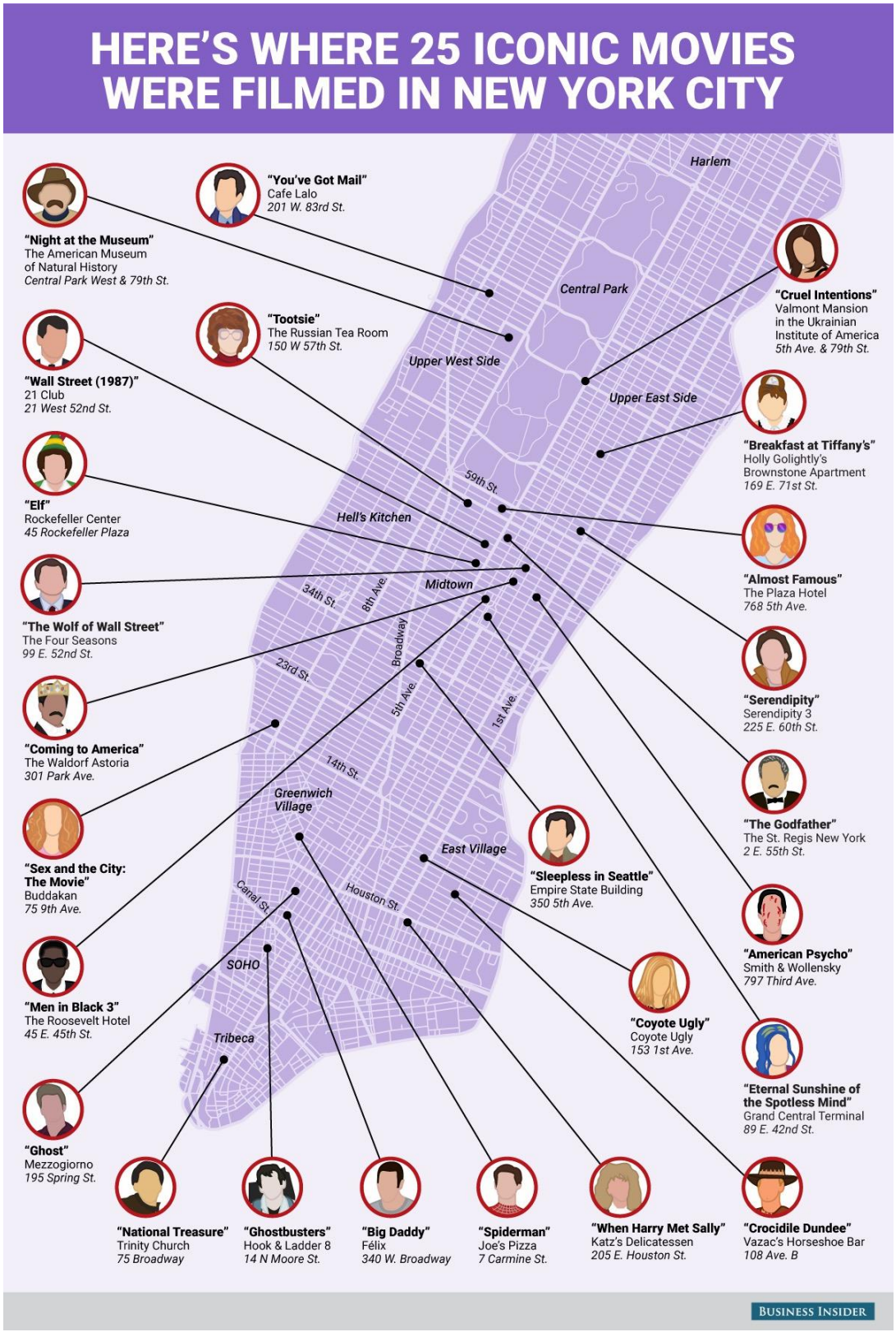
- This assignment is worth 25% of your overall grade.
- You must prepare your assignment in Jupyter Notebook. The report should include your code, results from running this code, all data visualisations, and your discussion and explanations as Markdown cells. Hint: google for "*jupyter notebook markdown example*" and you get a lot of examples on how to format your text.
- You must submit two files (!!!): (1) Jupyter Notebook file as described above; (2) PDF print out of the Notebook document. Go in menu File -> Print Preview, you get a new browser window, press Ctrl+P (or Command + P on MacOSX) to print the window and select "Save as PDF" as printing destination. If any of the above two files is missing, 50% deduction would apply.
- You do NOT need to include the data files provided to you because I have them too.
- The assignment is out of 100 marks. To obtain the maximum available marks you should aim to:

    1. Code all requested components (60%).
    2. Aim for optimised code in terms of computational overhead (5%). It is not always possible to avoid loops, however you should aim to avoid loops where possible.
    3. Use a clear coding style (5%). Code clarity is an important part of your submission. Thus, you should choose meaningful variable names and adopt the use of comments - you don't need to comment every single line, as this will affect readability - however you should aim to comment at least each section of code.
    4. Have the code run successfully (5%). Don't use hard-coded path to data file, learn how to use **working directory**.
    5. Write a short report within Jupyter Notebook (25%).

- This assignment can be openly discussed on the forum and in the class, students are welcome to share tips and tricks (not entire program or report, however).
- Assignments submitted late, without an extension being granted, will attract a penalty of 10 marks per each day or any part thereof beyond the due date and time.
- **Plagiarism is a specific form of academic misconduct.** Although the University encourages discussing work with others and the Social Forum will support this, ultimately this assignment is to represent your individual work. If plagiarism is found, all parties will be penalised. You should retain copies of all assignment computer files used during development of the solution to Assignment 1. These files must remain unchanged after submission, for the purpose of checking if required.

As usual, if you have any questions, problems, concerns - feel free to contact me.

New York City Filming Permits



Data and Background Information

When filming on location, permits are required for the exclusive use of city property such as streets, parks and even footpaths. There are many film locations around the world however one of the most iconic is New York City. To film in New York City permission is required from the Mayor's Office of Media and Entertainment.

Data on these filming permits is hosted by the open data platform of the City of New York. https://opendata.cityofnewyork.us/ You don't need this website for the assignment but it's a great link to follow for lots of interesting data sets.

The data for this question is in the file **File_Permits.csv**. It contains 74,449 rows of data across 14 columns. This is the data you will analyse. The file **Film_Permits_Data_Dictionary.docx** contains the data dictionary for each variable in the data set.

For this assignment, you will need to produce a report summarising a collection of requested statistical analyses and visualisations of the data. See the below for details.

Assignment tasks

Your report should contain:

1. An introduction outlining the analysis to follow/background information and data set description. The introduction can be 2-4 paragraphs. For the purposes of this assignment a paragraph is 6-8 sentences.
2. Calculate and present a monthly based statistics for the number of film shootings (by start date). You need a numerical summary and graphical representation. Discuss trends in the number of shootings over time. Compare the number of film shootings during COVID-19 period and outside it. **Hint**: you might need package **datetime** for date/time manipulation or you can use **datetime** functionality in **numpy**/**pandas**.
3. Analyse the *Duration of filming*. To determine filming duration, you will need to use the variables *StartDateTime* and *EndDateTime*.
   The statistical summary should consist of a numerical analysis and visual representation of your calculated duration. In an addition to the overall distribution, you should analyse duration split by *Borough* and by *Category* independently. What areas get short or long durations? What filming categories get long or short durations?
4. A numerical and visual analysis of *Category* – overall and then broken down by *Country*. Include a numerical and visual analysis. Are there any patterns you can detect in the output? A discussion of one to two paragraphs in total is fine.
5. Repeat analysis of *Category* but broken down by *PolicePrecinct*. What precincts are the most popular for what filming categories? If there are more than one precinct registered for a category, then each precinct gets a count in this category.
6. Conclusions