OCTOBER, 2023

EXTRACTING INSIGHTS FROM US DATA WITH PYTHON

GROUP PROJECT

PROGRAMMING FOR DATA SCIENCE 2023/2024

I. INTRODUCTION

Working with data is of paramount importance in the modern world. Data drives our decisions, shapes our understanding of the world, and influences our future. In this course, we have covered, and will continue to cover, different ways to leverage the power of programming to extract insights from data. Our projects will provide a hands-on opportunity to apply the concepts and techniques learned throughout the course.

II. PROJECT GOALS

Working in data science is more than just technical ability, as it requires a great deal of critical thinking and problem-solving skills to succeed. You have the option to choose one of three projects for this course:
Project 1: Gun Crime Data in the US
Project 2: Airbnb Data in the US
Project 3: Tripadvisor Data on US Museums

Each end-of-course group project is divided into three main components, each with its own set of objectives:

Preprocessing and Exploratory Data Analysis (EDA): Clean and preprocess your dataset, then perform an exploratory analysis to uncover patterns and relationships.
Working with Advanced Topics: The second component requires you to apply one or more of the advanced topics covered in the course. This could involve working with text, datetime objects, geographical coordinates or any other advanced data type relevant to your project.
Data Integration: The final component involves combining the information present in your dataset with data from other sources (including data from the other projects). This section is more open-ended, allowing you to choose the level of complexity you wish to pursue. We provide some thought-provoking questions that could be a starting point for your analysis.

III. PROJECT 1: GUN CRIME

Tackling the issue of gun violence in America requires a comprehensive strategy. The National Institute of Justice (NIJ) suggests that scrutinizing data from past incidents can reveal patterns and trends, thereby guiding policy decisions and providing a solid framework to address this pressing issue. As a team working for 'Brady: United Against Gun Violence', your task is to delve into, analyze, and correlate past gun incidents to provide valuable insights for policymakers aiming to mitigate gun violence in America.

III.1. PROJECT OBJECTIVES

The goals for this project are three-fold:

Communicate the data: Your first task is to import, comprehend, and preprocess the dataset. This will enable you to create compelling visualizations and identify key aspects of these incidents.

Time Intelligence: Leveraging your knowledge of datetime and timeseries, you should manipulate the data to answer specific questions such as:

1. What's the mean time between shootings by state?
2. Present a weekly time-series visualization for that state.
3. Identify seasonal and trend components while drawing some conclusions.
4. Is there any seasonal correlation with the number of persons killed or injured?

Open-source Intelligence: The final challenge involves integrating data from other sources with the Gun Crime dataset you've been working on. You should formulate relevant questions and use the data to answer them. Here follows a suggestion that could be (but does not have to be) a starting point:

- *Within a given state, is there a relationship between the locations of mass murders (10 kills or more) and the location of culturally relevant sites such as museums?*

III.2. PROJECT DATA

ATTRIBUTE	DESCRIPTION
incident_id	Incident unique identifier
date	Recorded date of the incident
state	State in which the incident took place
city	City within the state where the incident took place
address	Address where the incident took place
n_killed	Number of persons killed during the incident
n_injured	Number of persons injured during the incident

IV. PROJECT 2: AIRBNB LOCATIONS

Airbnb operates a progressive data education initiative known as Data University, with the goal of enabling all employees to make decisions informed by data. This initiative is Airbnb's effort to cultivate "citizen data scientists" and enhance data literacy across its workforce. As part of this program, your group is tasked with exploring, analyzing, and correlating an Airbnb dataset that includes information about properties throughout the United States, along with details about the property and its reviews.

IV.1. PROJECT OBJECTIVES

The goals for this project are three-fold:

Communicate the data: Your first task is to import, comprehend, and preprocess the dataset. This will enable you to create compelling visualizations and identify key aspects that may be relevant for the business.

Geospatial Intelligence: Leveraging your knowledge of geopandas, you should manipulate the data to answer specific questions about Airbnb operations in California:

1. Can you determine the geographical boundary encompassing all the listings of each room-type?
2. Which boundary has the biggest area?
3. What is the closest 'Private room' from the USS Midway? Can you plot both points on a map?
4. How many rooms exist, by type, within 100km of the USS Midway?

Open-source Intelligence: The final challenge involves integrating data from other sources with the Airbnb dataset you've been working on. You should formulate relevant questions and use the data to answer them. Here follows a suggestion that could be a starting point:

- *Within a given state, is there a relationship between the number of reviews (hence the number of stays) on a specific location and the amount of gun crimes reported in the area?*

IV.2. PROJECT DATA

ATTRIBUTE	DESCRIPTION
room_id	Airbnb's unique identifier for the listing
host_id	Airbnb's unique identifier for the host/user
neighbourhood	Location of the listing
latitude	Latitude of the listing (uses WGS84 projection)
longitude	Longitude of the listing (uses WGS84 projection)
room_type	Room type of the listing
price	Daily price of the listing in USD
minimum_nights	Minimum number of night stay
total_reviews	Total number of reviews
availability_365	The availability of the listing 365 days in advance
number_of_reviews_ltm	Total number of reviews in the last 12 months
state	State of the listing (abbreviated)
city	City/County of the listing

V. PROJECT 3: MUSEUM DATA

As a comprehensive resource for travel recommendations, TripAdvisor is seeking to optimize its current database on US attractions. To achieve this, the company has launched a **Museum Data Science Challenge**, inviting teams to delve into the data and uncover patterns that could enhance their museum recommendations. If your team's work yields promising results, you'll be invited to extend your analysis to data related to other points of interest, such as parks or restaurants. This is an exciting opportunity to contribute to improving the travel experiences of countless TripAdvisor users.

V.1. PROJECT OBJECTIVES

The goals for this project are three-fold:

Communicate the data: Your first task is to import, comprehend, and preprocess the dataset. This will enable you to create compelling visualizations and identify key aspects that may be relevant for the business.

Visitor Preferences Intelligence: Leveraging your knowledge of pandas, you should manipulate the data to answer specific questions about TripAdvisor museum attributes and visits across the US:

1. Are there types of museums whose majority of visitors are families? What are the most relevant tags associated with these museums?
2. What are the most highly-rated museums? Where are they located?
3. How many highly-rated museums can have their rating justified by a small number of reviews?
4. If a couple wants to visit a museum in New York, what would your recommendation be?

Open-source Intelligence: The final challenge involves integrating data from other sources with the museum dataset you've been working on. You should formulate relevant questions and use the data to answer them. Here follows a suggestion that could be (but does not have to be) a starting point:

- *Is there a relationship between the popularity of a museum and the number of Airbnb places to stay within a radius of, for example, 20 km?*

V.2. PROJECT DATA

ATTRIBUTE	DESCRIPTION
Address	The address of the museum
Description	A small text description introducing the museum
FeatureCount	The number of additional perks of the museum (bars, libraries, etc...)
MuseumName	Name of the museum
Rank	Position in the TripAdvisor Ranking
Rating	Average rating of the museum
ReviewCount	Number of reviews given to the museum by TripAdvisor users
TotalThingsToDo	Number of total activities that can be done at the museum
Museumid	Identifier tag of the museum

V.3. ADDITIONAL DATA

DATASET	DESCRIPTION
Museum Category	Includes one Description column with the type of museum
Tags	Includes tags associated with attributes associated with the museum
Traveller Type	Includes a count of the type of visitors the museum receives (number of families, number of couples, etc...)
Traveler Rating	Includes the number of votes in each tier of rating (Excellent, Very good, etc...)

VI. OUTLINE

Your project deliverables (especially the Jupyter Notebook) should respect the following outline:

Summary
A small overview of what you did in your project. The abstract should give an overview of your work in 200 to 300 words. What project did you choose? What questions did you explore at each level of analysis? What did you explore to tackle each of the 3 levels of analysis requested?

I. Level 1: Preprocessing and Exploratory Data Analysis (EDA):

- Description of data received
- Steps taken to clean and prepare the data

II. Level 2: Advanced Topics

- Additional preprocessing steps adopted
- Explain your rationale for addressing the questions that were asked

III. Level 3: Data Integration

- Present the questions you formulated
- Explain the strategy you adopted to address those questions

IV. Conclusion

- Discussion of your main insights
- Discussion of limitations of your work (e.g. what could you have done differently)
- Recommendations and suggestions for what the data owners can do based on your work.

VII. DELIVERABLES

Upon the project's deadline, you will be required to submit:

- A Jupyter notebook (or a zip of multiple notebooks) featuring all the code you used throughout the project to:
  a. Decide on your final solution for the problem at hand
  b. Obtain your final results (code that helped you make decisions but does not directly contribute to reaching the final solution should be included but commented).
  c. Any additional datasets required to run the notebook should also be present in the zip file.
- A presentation (meant for 15 minutes) describing what problems you wanted to tackle, how you went about tackling those problems and your main findings.
  - The file naming format should follow ***PDS_GroupXX_Presentation.pdf***, where ***GroupXX*** should be your group number.

VIII. EVALUATION

Your work will be evaluated according to the following criteria:

CRITERIA	PERCENTAGE (%)	MAXIMUM GRADE (OUT OF 20)
Notebook Quality	10	2
Preprocessing and EDA	25	5
Advanced Topics	20	4
Data Integration	15	3
Presentation	10	2
Discussion	20	4

10

Your grade will reflect our assessment of the quality of your work in terms of the correctness of your code, its clarity and efficiency. Please find below more details about what is taken into account for each topic:

- **Notebook Quality (2v):** A good notebook should be understandable to someone reading it for the first time. Having a good structure, with appropriate comments showing that you understand what is being done at every block of code goes a long way into that. It makes it easier for us to understand your rationale when tackling a problem, it will also help you when you wish to reuse some code or function later in your academic journey. Having an unorganized notebook will result in a lower grade on this component.
- **Preprocessing and EDA (5v):** Describe the data and extract meaningful insights that you consider helpful. Avoid adding visualizations and elements that add nothing to address the problem at hand. This section also covers the initial preprocessing of your original dataset. In essence, it should unambiguously explain what you did on a specific step and what are the reasons for your choice of approach.
- **Advanced Topics (4v):** This section covers your mastery of one or more of the “advanced topics” that we cover later in the course. Each project calls for a specific set of approaches to answer each question.
- **Data Integration (3v):** This component evaluates your approach to the final section of each project. The criteria for this section include the following questions: How pertinent were the open-source questions you formulated? How difficult is it to tackle these questions? How adequate is the strategy you adopted? Did you search for additional data, or did you only stick with the datasets we provided?
- **Presentation (2v):** In this section, we will evaluate the overall quality of your presentation as a standalone document.
- **Discussion (4v):** This section evaluates how you, as a group, deliver the message you want to convey whilst presenting it. How comfortable you are in answering questions about your data and methods is also under evaluation.

IX. PARTING NOTES

1. Please don't provide long theoretical explanations of topics covered in class in your report.
2. The trustworthiness of the information you provide is key. If you look for information outside the materials we provided, you should cite the source of the materials appropriately.
3. Before submitting, run your notebook from the start one last time, please comment unnecessarily lengthy cells that take too much time to run.
4. All the code you used that is unneeded to highlight the points you want to convey **should be part of your submitted notebook(s), but it should be commented.**
5. We will run your Jupyter Notebooks. So, please make sure we can run the notebook from start to finish in one go. Notebooks that do not fulfil this condition will be penalized.
6. **The notebook code will pass through a process of plagiarism and AI generation checking.**
7. **To avoid situations where we have conflicting versions, please make sure that you show, in the notebook, the version of the package you are using for each package you use.**
8. When determining the grade for your work, there will be a comparative component between your work and the works presented by your peers.

Friendly Reminders:

1. Attendance at the presentation is mandatory for approval of the project. The presentation and discussion have a group component and an individual component.
2. As questions are individualized, every group member should be able to understand what was done at every step of the way.
3. **If something is good enough to be mentioned, it is also good enough to know.** DO NOT include techniques/algorithms/steps you cannot explain in your report: we may (**and probably will**) ask about them in the defense.
4. **Finished is better than perfect.**