MCIS6273 Data Mining (Prof. Maull) / Fall 2023 / HW1

Points Possible	Due Date	Time Commitment (estimated)
20	Tuesday, Oct 24 @ Midnight	*up to* 24 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Learn more about data science tools in the wild for practitioners
- Perform data engineering in Pandas
- Perform exploratory data analysis (EDA) in Pandas
- Perform pattern mining with the `mlxtend` library

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw0`. Put all of your files in that directory. Then tar that directory, rename it with your name as the first part of the filename (e.g. `maull_hw0_files.zip`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using tar in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(20%) Learn more about data science tools in the wild for practitioners

Lot's of interesting things are going on in the data science landscape. You should make sure that you are aware of innovative changes in the field and how data is being used to move and shake even the most tech-resistant industries.

You will listen to the approx. 38 minute podcast Making Data Simple **September 13, 2023**: *Data-Driven Apartment Innovation: A Conversation with Mike Kaeding | Part 1* featuring an interview with Mike Kaeding, CEO of Norhart, who goes into detail about innovation in the apartment industry.

You can listen to / watch the show from one of the links below:

- (main page) Player.fm: Making Data Simple | *Data-Driven Apartment Innovation: A Conversation with Mike Kaeding | Part 1*
- (mp3 direct) MP3 file direct download
- Apple Podcasts
- Spotify

§ Task: Listen to the podcast / watch the video and write a 3-5 sentence *reaction* to the podcast. State in your own words what you learned, what expanding your knowledge of the topic and what you found *interesting* about the information you received.

§ Task: What did you learn about data, data science or data mining that you found surprising? (Please no more than 3 sentenses.)

(25%) Perform data engineering in Pandas

As data scientists, especially in a small organization, you will be tasked with doing exploratory analysis of data. Sometimes you will not know much about the data and will need to warm up to it so that you can then choose appropriate secondary methods of analysis. I often call this "taking the data out for coffee" to get to know it a little better.

Often times, basic statistical tools go a long way to quickly size up your data and understand which tools make the most sense moving forward.

In this part of the assignment we will be working with a very interesting dataset from the University of Arkansas, Fayetteville published just 6 weeks ago on August 13, 2023:

> Johansson, Emily (2023). Mammalian Camera Trap Data; Northwest Arkansas [Dataset]. Dryad. https://doi.org/10.5061/dryad.kd51c5bb6

This dataset includes the camera trap data from several months of data over two years, mostly during the late spring and summer months. A *camera trap* is

> "a device usually comprised of a DSLR camera, equipped with either infrared or motion sensors, and possibly an external flash, set up on a tripod or secured to a tree; When an animal, unaware of the setup, crosses the sensor's path, the camera captures a natural moment of the creature in its routine, without disrupting its life."

Please read short the abstract and description of this dataset so you can learn more about it. We will be uncovering some basic relationships in the data, which will later help us understand which algorithms we might find interesting.

§ Task: (*Data Extraction, Selection and Transformation*)

The data is mostly in the form we will be working with it, but you will need to do some transformation so that we can work with it in the next part a little more easily.

Your Python program / notebook must do the following:

(1) fetch to your local file system on Jupyter Hub the main remote data file (there is only one given above)
(2) once you have fetched the file, make 6 new CSV files from it which will act in a similar way as binary databases (similar to those you have read about in Zaki, Ch. 8 and in lecture). We are going to use these in the next part:

- **file 1 (`dataset_2000_2359.csv`):**

  group all rows by date, filter on only the times of day between 2000 (8p) and 2359 (1159p), the columns should be the species and the data instance should be `True` if there are 1 or more of the species, otherwise `False`

- **file 2 (`dataset_0000_0459.csv`):**

  similar to file 1 group by date, but the times of day will be 0000 (midnight) and 0459 (459a)

- **file (`dataset_2000_2359_urban.csv`):**

  use file 1 and filter on the columns which have `Forest Cover within 1.5km of camera (km^2)` value less than (<) 0.30.

- **file 4 (`dataset_2000_2359_rural.csv`):**

  use file 1 and filter on the columns which have `Forest Cover within 1.5km of camera (km^2)` value greater than or equal to (≥) 0.30.

- **file 5 (`dataset_0000_0459_urban.csv`):**

  similar to file 3, except use file 2 instead of file 1

- **file 6 (`dataset_0000_0459_rural.csv`):**

  similar to file 4, except use file 2 instead of file 1

You will (minimally) need to study the following Pandas functions to complete this task:

- `to_csv()`
- `groupby()`
- `unstack()`

- `fillna()`
- `map()`

You do not want to overthink this, but you **must use the filenames provided**.

(25%) Perform exploratory data analysis (EDA) in Pandas

Now that we have data, let's perform additional analysis on it.

We want to get some ideas about what's in the data, and for now, we will not come back to the files we generated in the prior part until the next part of this homework.

Let's find out about our data to learn a few things about what is there.

You will need to use the original dataset (not the one's you just created) to answer the following questions (each is work 2 points).

You will (minimally) need to study the following Pandas functions to complete this task: * `groupby()` * `unstack()` * `mode()`

You will also need to use the Seaborn (SNS) heatmap function here:

- `heatmap()`

In your notebook, you will also need to execute to install seaborn:

```
!conda install -y seaborn -c conda-forge # you must use -y
```

Each question **must be accompanied by the corresponding Pandas code to earn full credit**:

§ Task: What are the top 5 species detected over the entire dataset?

§ Task: Which 5 species over all data are most frequent at midnight (12am)?

§ Task: Use the library `sns.heatmap()` to generate a heatmap with `Species Detected` on the y-axis and `Hour of Detection` on the x-axis.

§ Task: Compare the result of your prior answer with running `heatmap()` with the `robust=True` parameter. Which of the two output graphs looks more interpretable. Why?

§ Task: Compute the *mode* of all species in the dataset.

§ Task: Compute the probabilities of all species, but make two columns: (1) with those in the group with forest density < 0.3 and (2) those with forest density ≥ 0.3. Label them "urban" and "rural".

Your table will look something like this:

Species Detected	rural	urban
Virginia Opossum	0.034703	0.0295551
Groundhog	0.00567261	0.0122004
Fox Squirrel	0.00155719	0.00117891
Nine-banded Armadillo	0.0635109	0.0263474
Person	0.0537229	0.10199
…	…	…

And in this fake table, you might notice the probability of *Virginia Opossum* is about the same in urban and rural, but the *Nine-banded Armadillo* is 2.4 times more likely in rural than urban environments.

(30%) Perform pattern mining with the `mlxtend` library

Now that we have data, let's perform additional analysis on it.

In the prevous part we use the *proximity* of forest density as a proxy for *urbanization*. While this may not fully be a founded assumption, for the purpose of this assignment that assumption is sufficient.

One interesting area to explore, is which species are out *earlier* in the evening, versus *later* and whether the setting (rural or urban) makes a difference. Thus, we split our data into a quadrant:

	early evening (*9p-1159p*)	late evening (*12a-4a*)
urban	?	?
rural	?	?

You might start to wonder a number of things and already be thinking maybe you have the data to answer this or that question. And you might be right but ... we are going to throw a twist into our inquiry. We are going to consider this a *frequent pattern mining* problem and think about this using tools for mining frequent patterns.

If we can convert the data into a binary database (binary table, binarized matrix, etc.) we can easily learn which species occur frequently with one another. To researchers in this area, this might be a very valuable thing to know.

The setup goes something like this:

- if every evening was grouped such that the rows were the dates and the columns the species, we could consider it a transaction $t \in T$ with items $i \in I$, where i are just the species occuring in that date (transaction t)
- once we have such a representation it is a trivial matter to use the libraries designed to do these types of pattern mining

You will (minimally) need to study the frequent patterns functions in mlxtend to complete this task:

- `apriori()`

You will also need to execute to install seaborn:

```
!conda install -y mlxtend -c conda-forge # you must use -y
```

Once we have everything set up, we would ultimately be able to answer the questions below. Remember each question **must be accompanied by the corresponding Pandas and mlxtend code to earn full credit**:

§ Task: (*Frequent Pattern Mining*)

Which *early evening* species are frequent together in groups of 2 or more regardless of rural or urban? (Use a support of 0.50 when answering).

§ Task: (*Frequent Pattern Mining*)

Show the *early evening* species frequent patterns for both urban and rural. Compare and contrast them. Do you see any surprises or are these what you expect?

§ Task: (*Frequent Pattern Mining*)

Show the *late evening* species frequent patterns for both urban and rural. Compare and contrast them. Do you see any surprises or are these what you expect?

§ Task: (*Reflection*)

Fill in the table with the most frequent itemset (with length $>$ 1) of highest support for each:

	early evening (*9p-1159p*)	late evening (*12a-4a*)
urban	(A, B); *sup* $= 0.45$	(A, B, G); *sup* $= 0.45$
rural	(A, H, K, N); *sup* $= 0.45$	(C, B); *sup* $= 0.65$