**ITD105 Laboratory Exercises #2**

**Data Collections**

**Name**: Mohammad Jibreel Y. Hamdain

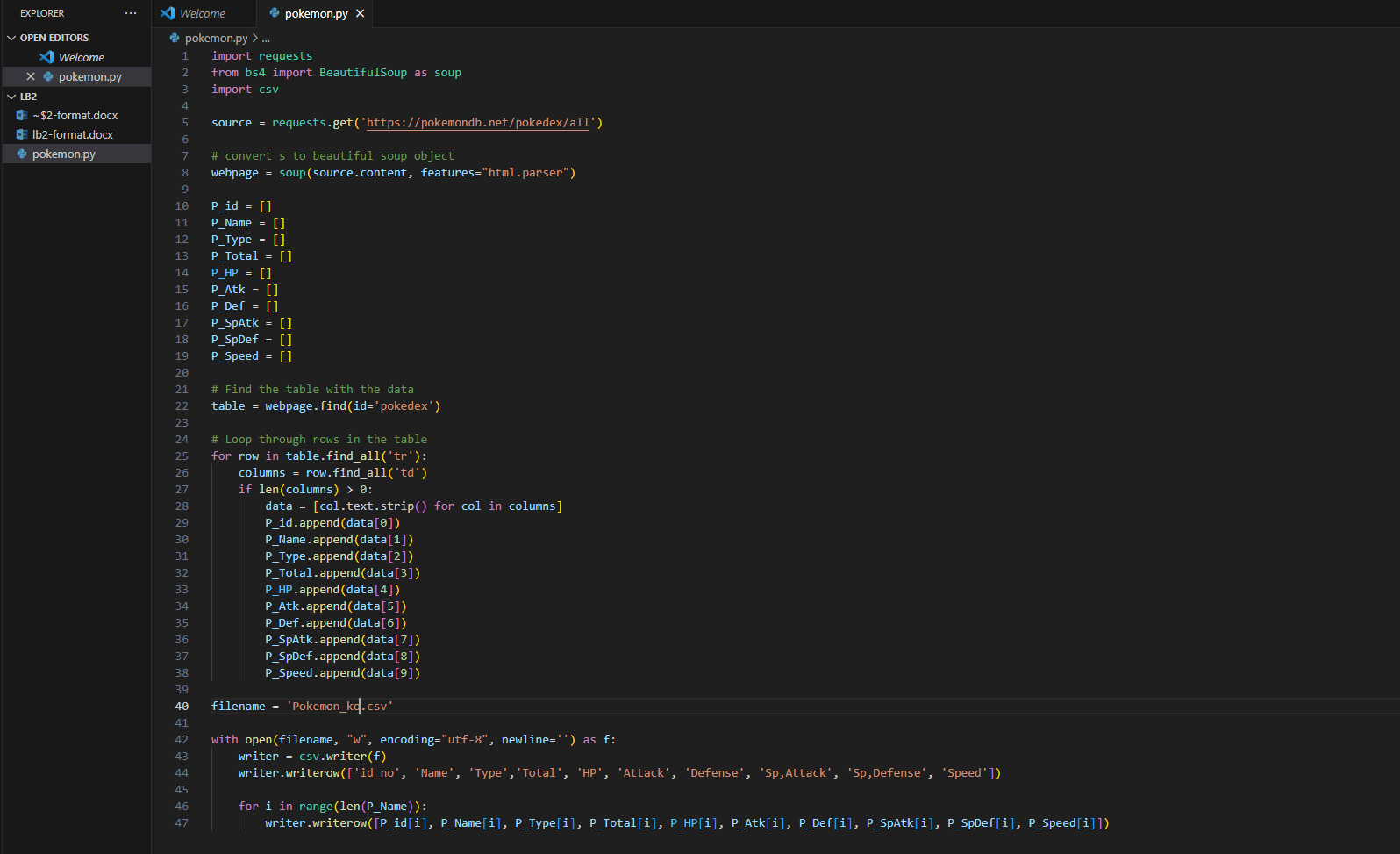
1. **Public Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Title | Type | Description | Source Link |
| Medical Cost Personal Datasets | Regression | This dataset was inspired by Brett Lantz's book Machine Learning with R. The information includes medical information as well as charges billed by health insurance firms. It includes 1338 rows of data including the columns age, gender, BMI, children, smoker, region, and insurance charges. | https://www.kaggle.com/datasets/mirichoi0218/insurance |
| American House Prices | Regression | A dataset comprising various variables around housing and demographics for the top 50 American cities by population. | https://www.kaggle.com/datasets/jeremylarcher/american-house-prices-and-demographics-of-top-cities |
| Student Performance Data Set | Classification | This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). In [Cortez and Silva, 2008], the two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details). | https://www.kaggle.com/datasets/larsen0966/student-performance-data-set |
| Fake News Classification | Classification | (WELFake) is a dataset of 72,134 news articles with 35,028 real and 37,106 fake news. For this, authors merged four popular news datasets (i.e. Kaggle, McIntire, Reuters, BuzzFeed Political) to prevent over-fitting of classifiers and to provide more text data for better ML training. Dataset contains four columns: Serial number (starting from 0); Title (about the text news heading); Text (about the news content); and Label (0 = fake and 1 = real). | https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification |

1. **Web Scraping**
2. PokemonDB

Video Link:

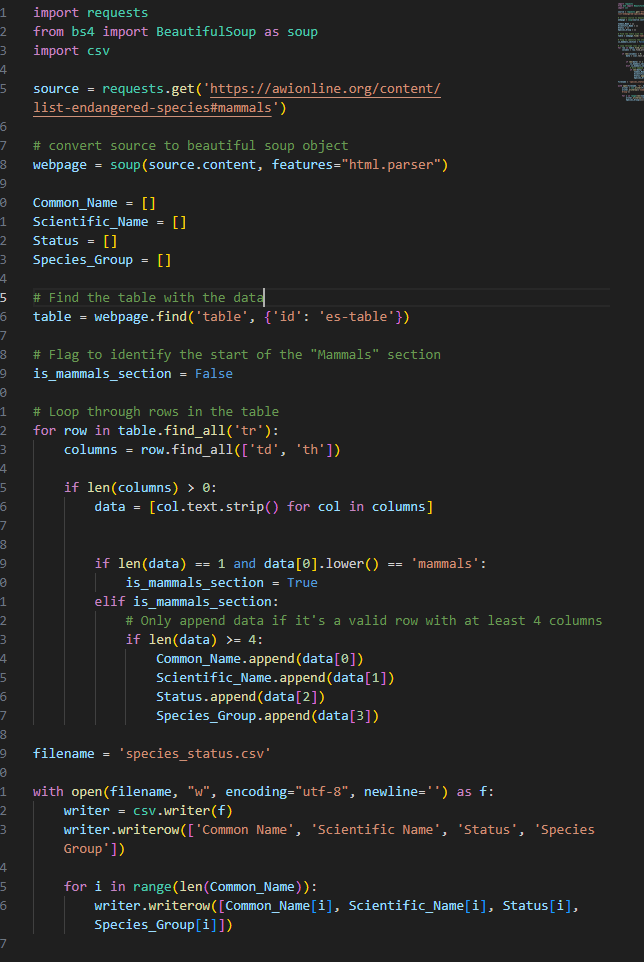
Python Code Snippet:



1. Type: Classification

Species\_Status, https://awionline.org/content/list-endangered-species#mammals

Video Link: https://drive.google.com/file/d/19CNk9CCDblsHoet7ETDzhIjVSKl0TK5X/view?usp=sharing

Python Code Snippet: https://drive.google.com/file/d/1dzZBOHuO9d18MEE61-AILM0FtbP-5UnO/view?usp=sharing

Excel File Link: <https://drive.google.com/file/d/1Z3jArbppmsoRvbpgqLlE6fkkursDzkUr/view?usp=sharing>

1. Survey Questionnaire
   1. Age: How old are you? (Numeric)
   2. Gender: What is your gender? (Male, Female, Other)
   3. Education Level: What is your highest level of education? (High School, Bachelor's Degree, Master's Degree, Doctorate, Other)
   4. Employment Status: What is your current employment status? (Employed, Unemployed, Student, Other)
   5. Income: What is your annual income? (Numeric)
   6. Marital Status: What is your marital status? (Single, Married, Divorced, Widowed)
   7. Hobbies: Select your hobbies (Checkbox options: Reading, Sports, Travel, Music, Cooking, Other)
   8. Health Status: How would you rate your overall health? (Poor, Fair, Good, Very Good, Excellent)
   9. Social Media Usage: How many hours per day do you spend on social media? (Numeric)
   10. Preferred Communication Channel: What is your preferred communication channel? (Phone, Email, Messaging Apps, In-person)

Sample Dataset Link: https://drive.google.com/file/d/1pwaHQR1SGkqtZHwWQ9HHcPniUfJTqZJO/view?usp=sharing  
Type of Dataset: Classification - the goal is to predict a categorical label (e.g., preferred communication channel) based on the survey responses.