🏠 / Assignments / Assignment 1

Assignment 1

This Assignment includes three parts, focusing on everything we have seen in the course so far. Doing this assignment fully will help you in your upcoming midterm. Here is a brief overview of the assignment:

- **Part 1** involves analyzing datasets (ramen, avocado, exams) using Pandas. Tasks include data reading, statistical analysis, column selection, sorting, and basic data manipulation. It also test students' understanding of the datasets.
- **Part 2** delves into theoretical concepts and practical applications of data mining and statistics. It requires you to describe the knowledge discovery process, perform statistical analysis on a given set of data, categorize attributes, and explore correlations in data.
- **Part 3** covers distance matrices and data normalization. It includes exercises on calculating distances between data points with different attribute types and normalizing datasets using various techniques.

Submission Instructions

- Solve Part 1 in 3 separate Jupyter notebooks. Make sure you run all the cells in sequence. Then export them as pdfs (You can do this by printing it to a pdf document)
- Solve Parts 2 and 3 in a word processing program and export as pdf.
- Combine these parts 1, 2 and 3 into a single pdf using a tool like this. **Submit this one pdf file.**
- Also submit the jupyter notebooks from part 1. We may have to reference these if the code is not clear from the pdf alone.
- I recommend not submitting hand written work because it is sometimes hard to read. If you must submit hand written work, to avoid any loss of marks, please make sure it is clearly written and easy to read.

Part 1: Python and Pandas

Some useful links that might be helpful along with the slides and notebooks posted on D2L:

1. **Official Pandas Cheat Sheet**
2. **Recommended YouTube video for pandas**

Part 1 Question 1: Analyze the ramen data (10 marks)

The data file: `ramen-ratings.csv`

- This dataset rates various types of ramen based on the brand variety style and country that sells it.

Tasks

1. Read the data from the CSV file into a DataFrame.
2. Display the first five rows of data.
3. Display the last five rows of data.
4. Display statistical information for the numeric columns using the describe() method.
5. Display the number of unique values for each column.
6. Display only rows where the country is Vietnam.
7. Display only the Brand and Style columns.
8. Display only the Country column.
9. Display the data after it has been sorted by the Stars column from high values to low values.
10. In the Country column replace “USA” with “United States”. Make sure this change is saved in the DataFrame and then display the first five rows to be sure the change was made correctly.

Questions

1. How many countries are represented in the data?
2. Which three countries have the highest average rating?
3. Which three countries have the lowest average rating?
4. Which three countries have the most brands and how many brands does each of these countries have?

Part 1 Question 2: Analyze the avocado data (10 marks)

The data file `avocado.csv`

- This dataset contains historical data on avocado prices and sales volume in multiple U.S. markets. One of the columns in this dataset Unnamed: 0 contains sequential numbers that are irrelevant to analyzing this data. Three of the other columns contain sales for PLU (price look-up) codes 4046 4225 and 4770. These columns will also not be used by the analyses done in these projects.
- If you review the data you’ll see that some of the regions overlap. For example, one of the regions is the entire U.S., and all of the other regions are parts of the U.S. Because of that, you would need to review this data carefully before determining the best way to analyze it. For the purposes of these exercises though, the overlapping regions won’t be taken into consideration.

Tasks

1. Read the data from the CSV file into a DataFrame.
2. Display type memory consumption and null count information using the info() method.
3. Display the number of unique values in each column.
4. Display all the rows of data that JupyterLab displays by default.
5. Display the first and last five rows of data and the first and last four columns of data.
6. Choose any three columns access them with bracket notation and display the first five rows of this data.
7. Select one column and access it with dot notation.
8. Multiply the Total Volume and AveragePrice columns and store the result in a new column called EstimatedRevenue. Then display the first five rows of this data to confirm that the column was added and has the correct values.
9. Create a DataFrame that’s grouped by region and type and that includes the average price for the grouped columns. Then reset the index and display the first five rows.

10. Create a bar plot that shows the mean median and standard deviation of the Total Volume column by year.

Questions

1. How many unique regions are there?
2. What is the average price for each type of avocado (organic and conventional)? Be sure to include just the type and AveragePrice columns in the results.
3. Which region has the lowest average price for organic avocados? Hint: Create wide data from the grouped data that you created in task 8.
4. Have the Total Bags sold per year of each type of avocado become more or less consistent over time?

Part 1 Question 3: Analyze the exam data (10 marks)

The data file: `exams.csv`

This dataset contains math reading and writing scores based on students’ gender race level of parent education whether they receive a free or reduced-cost lunch and whether they completed a test preparation course.

Tasks

1. Read the data from the CSV file into a DataFrame and display the first five rows.
2. Display the basic information for the DataFrame and its columns using the info() method.
3. Display statistical information for the math score reading score and writing score columns using the describe() method.
4. Group the data by the race/ethnicity column and display the mean scores.
5. Display a single column as a DataFrame with bracket notation.
6. Display a single column as a Series with bracket notation.
7. Display a single column as a Series with dot notation.
8. Display only rows for females with a math score greater than or equal to 90.

Questions

1. Does taking a test preparation course improve average scores?
2. Which gender is better on average at math?
3. Which gender is better on average at all three subjects? Hint: Start by adding a column to the DataFrame with the total score.

Part 2: Stats and Attribute Comparison

Part 2 Question 1: The process of knowledge discovery (6 marks)

Describe in brief (max 250 words) the knowledge discovery process in your own terms and explain the role of data mining within it. Please refer to the following figure to answer this question. Use bulleted lists and bold important text if necessary. Figure 1.1 Page 3

Part 2 Question 2: Statistics of Data

Please read section 2.2 Statistics of data

Question 2.1 (6 marks)

Suppose that the data for analysis include the attribute age. The age values are (in ascending order).

```
ages = (13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25,\
     25, 30, 33, 33, 35, 35, 35, 36, 40, 45, 46, 52, 56, 60, 70)
```

- What is the mean of the data? What is the median?
- What is the mode of the data? Comment on the data's modality (i.e., unimodal, bimodal, trimodal, etc.).
- What is the midrange of the data?
- Can you find (roughly) the first quartile (Q1) and the third quartile (Q3) of the data?
- Give the five-number summary of the data.
- Show a boxplot of the data. You can use panda's boxplot method for this

Question 2.2 (2 marks)

If the median of a dataset is significantly different from the mean, what can you infer about the dataset's distribution?

Question 2.3 (2 marks)

What does it imply about a dataset if the variance is zero?

Part 2 Question 3: Attribute Types (4 marks)

Please review section 2.1 Data Types. This video may be helpful.

	0	1	2	3	4
Car_Model	Model A	Model C	Model B	Model B	Model A
Sunroof	0	0	1	0	1
Is_A_Transformer	0	0	0	0	0
Condition	Fair	Good	Excellent	Fair	Good
Engine_Oil_Temp_In_Celsius	69	49	35	37	49
Weight	2145	4832	5197	3278	5561
Owners	1	1	0	3	3
Mileage	10211	65427	60469	18527	41964

For the provided dataset, please match each column to one of the corresponding attribute types listed below:

- continuous

- binary_non_symmetric
- ordinal
- interval_scaled
- ratio_scaled
- binary_symmetric
- discrete
- nominal

Part 2 Question 4: Comparing attributes (10 marks)

As per the exams.csv file, does the parents' level of education have an effect on the average scores?

Perform the following steps:

Step 1 Load the `exams.csv` into a pandas Dataframe

Step 2 Create a new column called `parental education numeric` which maps parent's education level to a number. You can use the `map` method of pandas along with the following mapping:

```
education_mapping = {
    'some high school': 1,
    'high school': 2,
    'some college': 3,
    'associate\'s degree': 4,
    'bachelor\'s degree': 5,
    'master\'s degree': 6
}
```

After this step, if you displayed these columns of your Dataframe, your output should look something like this:

```
df[['parental level of education', 'parental education numeric', 'math score',\
     'reading score', 'writing score']].head()
```

	edu_lvl	edu_lvl_numeric	math score	reading score	writing score
0	bachelor's degree	5	72	72	74
1	some college	3	69	90	88
2	master's degree	6	90	95	93
3	associate's degree	4	47	57	44
4	some college	3	76	78	75

note: I have replaced some of the column headers with shorter names so that they fit on the page

Step 3 Create a correlation matrix on these 4 numeric columns: parental education numeric, math score, reading score, writing score. You can use the `corr()` method of the pandas Dataframe object for this purpose. See docs.

Now answer these questions briefly:

- **Q4.1** Show the correlation matrix and provide your interpretation of it.
- **Q4.2** Describe the relationship between parent's education and marks obtained? Are these features negatively correlated, independent or positively correlated? Is there a strong correlation or weak?
- **Q4.3** Which course has the strongest correlation with the parent's education level?

(No need to provide the python code for the question 4)

Part 2 Question 5: x^2-Square Hypothesis Testing (10 Marks)

Tip: review Example 2.14 for this question

With reference to the following data from the doomed voyage of the Titanic, conduct a chi-square test to investigate whether survival was independent of a passenger's status on the Titanic. Utilize a significance level of **0.001** for your analysis.

Titanic Survival Data:

Status	Lived	Died	Total
Crew	212	673	885
1st Class	202	123	325
2nd Class	118	167	285
3rd Class	178	528	706
Total	710	1491	2201

Tasks:

1. State the null hypothesis for the chi-square test in this context.
2. Calculate the expected counts for each cell in the table if survival were independent of passenger status and create a contingency table similar to table 2.2 on page 37 of the textbook.
3. Compute the chi-square statistic based on the observed counts and the expected counts. χ^2-Square lookup table
4. Conclude whether to reject or not reject the null hypothesis based on the chi-square statistic and the significance level.

Part 3: Distance Matrices and Data Normalization

Part 3 Question 1: Distance Matrices (10 marks)

The following data matrix shows a rating from 1-10 of the level of interest in sports, movies and music for Anna, Bob and Chuck. Build the Manhattan and Euclidean Distance Matrices between each pair of individuals.

Name	Sports	Movies	Music
Anna	3	6	5
Bob	8	4	3
Chuck	1	9	8

Part 3 Question 2: Calculating distance between data with mixed attribute types (10 marks)

Note: For this question, please refer back to section 2.3.6 for review.

In practical scenarios, datasets often include a mix of different types of features such as numeric, nominal, asymmetric binary, symmetric binary, and ordinal. Imagine you have a dataset with the following columns:

- Sex Assigned at Birth (Symmetric Binary: Male, Female)
- Age (Numeric)
- Occupation (Nominal: Engineer, Doctor, Artist, etc.)
- Is an Olympic Medalist (Asymmetric Binary: Yes, No)
- Education Level (Ordinal: High School, Bachelor's, Master's, PhD)

Sample data looks like this:

Index	Name	Sex Assigned at Birth	Age	Occupation	Is an Olympic Medalist	Education Level
1	Alice	F	25	Computer Programmer	No	Bachelor's
2	Bob	M	30	Data Scientist	No	Master's
3	Carol	F	22	ML Expert	No	Bachelor's
4	Dave	M	40	Computer Programmer	No	Doctorate
5	Eve	F	35	Data Scientist	No	Master's
6	Frank	M	29	ML Expert	No	Bachelor's
7	Grace	F	24	Computer Programmer	Yes	Bachelor's
8	Henry	M	50	Data Scientist	No	Master's
9	Irene	F	31	ML Expert	No	Bachelor's
0	Jack	M	26	Computer Programmer	No	Master's

You will compare the distance between two rows. To determine which two rows you need to compare, use the last two digits of your student ID. For example, if your student ID is `20201234`, you will compare rows 3 and 4, which correspond to Carol and Dave in the table. (If the last two digits of your ID are the same, then use the digit before that.)

Show your work for full credit

Part 3 Question 3: Data Normalization (10 marks)

Consider the following small dataset with three attributes: `Age`, `Salary`, and `Temperature`.

Index	Age	Salary	Temperature
1	25	55,000	35
2	30	40,000	28
3	22	80,000	40
4	28	50,000	30
5	35	70,000	25

Tasks:

For indexes 1 and 2, perform:

1. **Min-max normalization**: Normalize the `Age` and `Salary` attributes using min-max normalization to a range of [0, 1].
2. **Z-score normalization**: Apply z-score normalization to the `Temperature` attribute. Calculate the mean and standard deviation first.
3. **Normalization by Decimal Scaling**: Use decimal scaling to normalize the `Salary` attribute.
4. **Summary**: Which normalization technique you would recommend for columns that have:

    - When we know the min and max values and want to preserve the original distribution of the data.
    - When dealing with attributes having unknown future min and max, or when we have outliers.
    - When the range is not known, and is less affected by outliers.

Please show your calculations.