CISC 3225: Midterm Review

Pandas Fundamentals

- Basic object types: Series vs. DataFrame
  - Creating Series and DataFrame objects
  - The role of indexes in Series and DataFrame objects
- Accessing data
  - Basic indexing
  - Fancy indexing
  - .loc and .iloc
  - Relational operators and Boolean expressions for querying

Analysis

- Detecting and handling missing data
  - Problems with missing data: Pandas vs. NumPy
  - Finding missing values
  - Strategies for handling missing data: drop, fill, ignore
- Interpreting visualizations
  - Box plots
  - Correlation matrices
- Asking and answering questions about data

Pandas Features

- Hierarchical indexes
- Concatenation
- Merging
- Grouping and aggregation

Loading data

- CSV files

1. True/False

____ A DataFrame is similar to a 3-dimensional NumPy array.

____ Pandas will allow you to concatenate two DataFrames with the same row indices.

____ By default, a NumPy array, and a Pandas Series will accept None values as-is and attempt to perform math and other operations on them.

____ A correlation matrix can show you if there are outliers in your data.

____ NaN is a floating point number.

Multiple Choice

2. If a Series is like a dictionary that maps keys to values, a DataFrame is like:

a) A dictionary that maps keys to rows

b) A dictionary that maps keys to Series objects

c) A DataFrame is not like a dictionary

3. You want to retrieve the 500 most recent errors in a DataFrame of logged Web server errors sorted by time. You should use (select all that apply):

a) loc b) iloc c) an IndexSlice object d) Fancy indexing

4. You have a DataFrame of coffee data. Assuming an index appropriate for your query, you want to retrieve the caffeine content of arabica coffee from Ethiopia. You should use (select all that apply):

a) loc b) iloc c) an IndexSlice object d) Fancy indexing

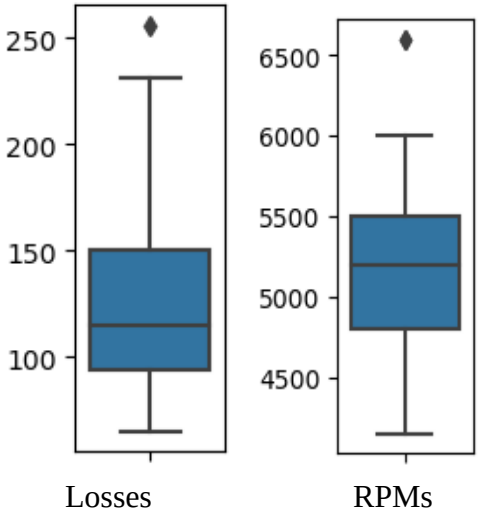
5. Consider the pair of box plots on the right from a dataset about cars. Which of the following facts can we **not** infer from them? **Select all that apply**.

a) The median value of insurance losses is about 112.

b) The mean value of RPMs is about 5250.

c) Neither losses nor RPMs contain values less than 0.

d) The cars with RPMs above 6500 experience losses above 250.



Short Answer and Data Questions

Use the DataFrame below for questions 6 and 7.

	employee_id	department	role	employee_level
0	194482	accounting	accountant	1
1	855324	accounting	accountant	2
2	855324	oversight	auditor	1
3	947734	oversight	manager	4
4	023352	hr	recruiter	2
5	321145	executive	assistant	3

6. Show how to compute the following queries in a single Pandas operation:

- What is the highest employee level in each department?
- What is each employee's highest level in any department?

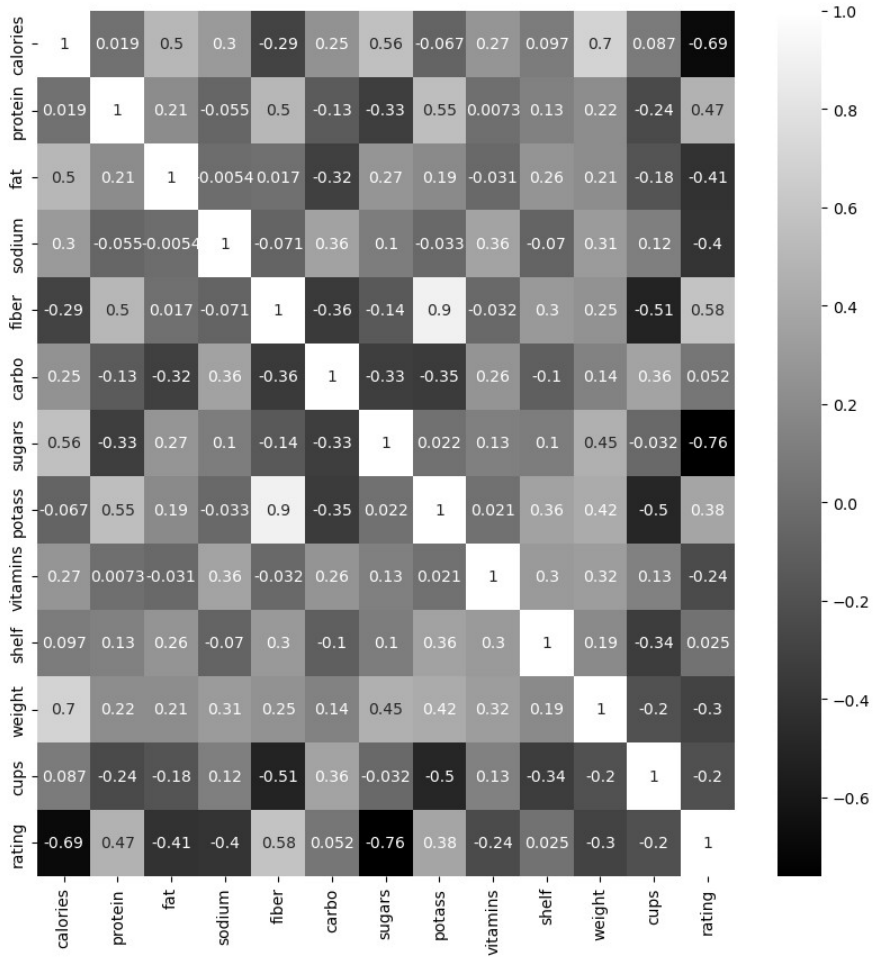
7. Could this DataFrame be multi-indexed to make accessing data easier? Based solely on the data shown in the DataFrame, how could you reindex it and why?

Show how you can compute the following values using the new index:

- The average level of the accounting department
- Employee 855324's average level.

8. Below is a correlation matrix from a dataset of breakfast cereal nutrition. Aside from variables related to basic nutritional information, other variables include:

- vitamins: The percent daily recommended vitamins found in the cereal. Limited to 0%, 25%, and 100%.
- shelf: The display shelf in the grocery store. 1 is near the floor, 2 is mid-level, and 3 is high.
- weight: The weight in ounces of one serving, as determined by the manufacturer.
- cups: The number of cups in one serving, as determined by the manufacturer.
- rating: A score from an organization that reviews manufactured foods. Higher is better.



a) Based on the strength of the correlations shown in this chart, what factors do you think the rating institution uses to determine the rating of breakfast cereals?

b) Identify an interesting correlation that you believe warrants additional research. What does it suggest about the data? Make a guess as to what is going on.

9. Merge the DataFrames shown below. Show the output from a right merge. Is it one-to-one, many-to-one, or many-to-many?

	a	b
0	1	9
1	2	8
2	3	7

	b	c
0	9	5
1	9	4
2	6	6

10. Concatenate the DataFrames shown below along axis=1.

	a	b
0	1	2
1	2	3
2	3	4

	c
0	5
1	6
6	7

1. True/False

F___ A DataFrame is similar to a 3-dimensional NumPy array.

T___ Pandas will allow you to concatenate two DataFrames with the same row indices.

F___ By default, a NumPy array, and a Pandas Series will accept None values as-is and attempt to perform math and other operations on them.

F___ A correlation matrix can show you if there are outliers in your data.

T___ NaN is a floating point number.

Multiple Choice

2. If a Series is like a dictionary that maps keys to values, a DataFrame is like:

a) A dictionary that maps keys to rows

b) A dictionary that maps keys to Series objects

c) A DataFrame is not like a dictionary

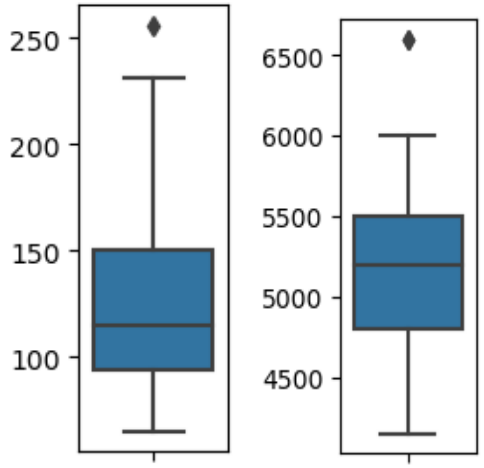
3. You want to retrieve the 500 most recent errors in a DataFrame of logged Web server errors sorted by time. You should use (select all that apply):

a) loc **b) iloc** c) an IndexSlice object

4. You have a DataFrame of coffee data. Assuming an index appropriate for your query, you want to retrieve the caffeine content of arabica coffee from Ethiopia. You should use (select all that apply):

a) loc b) iloc **c) an IndexSlice object**

5. Consider the pair of box plots on the right from a dataset about cars. Which of the following facts can we **not** infer from them? **Select all that apply**.



a) The median value of insurance losses is about 112.

b) The mean value of RPMs is about 5250.

c) Neither losses nor RPMs contain values less than 0.

d) The cars with RPMs above 6500 experience losses above 250.

Short Answer

6.

Maximum employee level by department:

```
df.groupby('department')['employee_level'].max()
```

Maximum employee level by employee:

```
df.groupby('employee_id')['employee_level'].max()
```

7.

In the DataFrame, each employee can be in multiple departments, but only has one role within each department. This means we can make a MultiIndex of the employee_id and department column. If we do this, each employee will still have a unique index value.

The average level of the accounting department:

```
df.loc[idx[:, 'accounting'], 'employee_level'].mean()
```

Employee 855324's average level:

```
df.loc[idx[855324, :], 'employee_level'].mean()
```

8. Answers may vary.

a) Main answer:

The organization is primarily looking for low-calorie, low-sugar cereal. There is a strong negative relationship between rating/calories and rating/sugars, indicating that high calories or high sugar typically results in very low ratings. The next strongest correlation is a positive relationship between fiber and rating, indicating that the organization prefers high-fiber cereal, but it is not as strong as that of calories and fat.

Some other observations:

There is a negative correlation between fat and ratings, but it is not as strong as that for calories, sugar, or fiber.

Potassium has a weak positive correlation with rating, suggesting that the organization views some potassium as favorable, but it does not have a major impact on the score.

There is a weak negative correlation between cups/rating and weight/rating, suggesting the organization slightly prefers smaller serving sizes.

The organization does not appear to consider carbohydrates (possibly because all breakfast cereal is essentially made of carbohydrates) or shelf position (where it is stocked in grocery stores has nothing to do with the organization's quality rating).

b) Here are a few possibilities:

Potassium and fiber are almost perfectly correlated, meaning that cereals high in fiber are also high in potassium. It could be because foods high in fiber are also naturally high in potassium, or because fiber additives are high in potassium. Alternatively, it might be possible that potassium and fiber are good to eat together, so cereal manufacturers add potassium to high-fiber cereals.

The relationships between calories and various nutrition variables are interesting. For example, there is a positive relationship between calories/fat and calories/sugar, which makes sense. There is a negative relationship between calories and fiber, suggesting that we can't extract useful energy from fiber, and it just takes up space in a cereal. There is a small positive relationship between calories and sodium, which is interesting because sodium doesn't contriubte any calories – this suggests that sodium might be a common additive to foods that are already high-calorie.

9.

It is a many-to-one merge.

	a	b	c
0	1.0	9	5
1	1.0	9	4
2	NaN	6	6

10.

	a	b	c
0	1.0	2.0	5.0
1	2.0	3.0	6.0
2	3.0	4.0	NaN
6	NaN	NaN	7.0