Software Engineering Revision Notes

Concepts Covered in the Class

---

1. Handling Missing Data in Pandas

Difference between `None` and `NaN`:

- `None` and `NaN` indicate missing values in dataframes but are used differently depending on the data type of the column.
- `None` is used for non-numeric data, while `NaN` is used for numeric data to denote missing entries【4:6†transcript.txt】【4:9†transcript.txt】.

Detecting Missing Values:

- Functions like `isna()` and `isnull()` help in identifying missing values. Both return a boolean DataFrame indicating the presence of missing values【4:7†typed.md】.

Removing and Imputing Missing Data:

- `dropna()` can remove rows or columns with missing values. 【4:15†typed.md】
- `fillna()` replaces missing values with specified values. Common strategies include filling with zeros, mean, median, or mode 【4:15†typed.md】【4:5†typed.md】.

2. Aggregation and Grouping in Pandas

Basic Grouping and Aggregation:

- Group data based on columns to apply functions such as `max()`, `min()`, `count()`, etc., to understand data distribution. For instance, finding the movies' release span for directors【4:0†transcript.txt】 【4:4†transcript.txt】.

- Multi-indexes help in managing hierarchical data, where you can perform complex aggregations and manage results with multi-index structures 【4:2†transcript.txt】.

Practical Application of Aggregates:

- Calculate productivity by counting occurrences, for example, counting movie titles per director and analyzing their productivity over the years 【4:4†transcript.txt】【4:13†transcript.txt】.

3. Advanced Data Manipulation:

Pivoting and Melting DataFrames:

- Reshape dataframes for ease of analysis using `melt()` to convert wide format data into a long format, whereas `pivot()` is used to reverse the operation【4:18†transcript.txt】【4:19†transcript.txt】.

Data Transformation and Preparation:

- Transform data for better accessibility and efficiency, useful in cases where future data may vary in structure, such as time intervals that might change periodically【4:18†transcript.txt】.

4. Practical Exercises and Assignments

Exercises involving Real Data Sets:

- Calculate statistics such as the highest average pressure for drugs or most productive directors, employing data manipulation techniques discussed【4:17†transcript.txt】.

5. Visualization and Future Topics

Transition to Visualization:

- Future classes will cover visualization libraries like Matplotlib and Seaborn, crucial for presenting data findings visually【4:12†transcript.txt】【4:17†transcript.txt】.

- Time series operations and other advanced topics will be covered in subsequent sessions, building on foundational data handling skills 【4:16†transcript.txt】.

This concludes your revision of the discussed topics. Ensure you practice the exercises as it reinforces your understanding of these concepts.