Fall 2018: Introduction to Data Science

GIRI NARASIMHAN, SCIS, FIU
The DataFrame

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>foo</td>
<td>one</td>
<td>small</td>
</tr>
<tr>
<td>1</td>
<td>foo</td>
<td>one</td>
<td>large</td>
</tr>
<tr>
<td>2</td>
<td>foo</td>
<td>one</td>
<td>large</td>
</tr>
<tr>
<td>3</td>
<td>foo</td>
<td>two</td>
<td>small</td>
</tr>
<tr>
<td>4</td>
<td>foo</td>
<td>two</td>
<td>small</td>
</tr>
<tr>
<td>5</td>
<td>bar</td>
<td>one</td>
<td>large</td>
</tr>
<tr>
<td>6</td>
<td>bar</td>
<td>one</td>
<td>small</td>
</tr>
<tr>
<td>7</td>
<td>bar</td>
<td>two</td>
<td>small</td>
</tr>
<tr>
<td>8</td>
<td>bar</td>
<td>two</td>
<td>large</td>
</tr>
</tbody>
</table>

- Rows -> Axis 0
- Columns -> Axis 1
- `df["C"]`
- `df.iloc[3]`
- `df.iloc[6]["A"]`
Chain Indexing

- `df.iloc[6]["A"]` is an example of chain indexing and is considered bad Python practice
Missing Values

- Python uses NaN to indicate missing values as it reads in.
- This feature can be turned off.
- Missing values can be filled in with other default values.
- ForwardFill and BackwardFill propagate next or previous values in table.
Scales

- **Ratio** Scale: equally spaced with valid +/-1; e.g. height
- **Interval** Scale: equally spaced, but zero has specific meaning; e.g. temp
- **Ordinal** Scale: ordered values, but not equally spaced; e.g. grades
- **Nominal** Scale: categorized, no order; e.g., Countries

- Can convert one to another
  - Grades could be nominal/categorical
  - Can be converted to ordinal or ratio
- Can also convert numerical values to categorical
  - Discretization
  - Histograms
- Use cut feature in pandas
SQL is a query language used to query relational databases

SELECT operation

- SELECT [ ] FROM [ ] WHERE [ ]

Python notebooks allow for SQL queries to be incorporated

query = """SELECT fields FROM Rel WHERE conds """

df = Rel.query_to_pandas(query)
Google’s BigQuery

- Google's serverless enterprise data warehouse with security
- Infrastructure by Google to create logical data warehouse
- Allows scalable data analytics and ML tools at good price-performance
- Uses SQL without need for database administrator
- Allows relational DB, spreadsheets, objects DB, and ODBC/JDBC drivers
- Makes it easy to join public or commercial datasets with local datasets
- Columnar & cloud storage, parallel execution, automatic optimizations
- Supports popular BI tools like Tableau, MicroStrategy, Looker, and Data Studio BETA out of the box
Let’s try BigQuery

- BigQuery is a database that lets you use SQL to work with very large datasets.
- After logging in, upload the Python notebook sql2py.ipynb and run it.
- The code, loads the Chicago_crime database.
- It then shows how to convert SQL queries into python code.
Blogs

- Planetpython.org
- Dataskeptic.com