AIG150- Week 4

Working With Missing Data

Reading Text: Chap 09, Pandas for everyone

Agenda

- Find and count missing data
- Cleaning missing data
- Calculations with missing data Outlier detection

Working With Missing Data

How missing values are represented in Pandas ?

How can data go missing during the processing phase ?

How to fill in the missing data ?

NaN

- NULL in databases, NA in programming languages
- In Pandas they are NaN, NAN, or nan (numpy) Different than all the other data types
- They are not equal to anything, no two NaN values are equal
- Need a special comparison function called isnull() Example

  pandas.isnull(NAN) should return TRUE

Where Do They Come From ??

- If data is missing at the time of loading, Pandas fill the empty cells with nan

- The following three options are available with the read_csv() file function:
  - ***na_values*** example: *na_values="Unknown"*

  - ***keep_default_na*** Boolean parameter, by default this is True –> any additional missing values specified with na_value will be appended to the list of missing value.
  - When set to False → it will use only the missing values specified in na_values

  - ***na_filter*** Boolean parameter, by default this is True –> missing values will be coded as NAN, set it to false if you don't want to record the missing values

- From merging

  data User input

Reindexing

- When you reindex your dataframe, it will introduce missing values
- Sometimes we need to add more data and would like to retain the original data at the same time

Working with Missing Data

- Find and count missing data
- Clean missing data
- Calculations with missing data

Pandas- Missing Data

- Understand how Pandas represents missing data (NaN, pd.NA)
- Identify common sources of missing data (load, merge, reindex, manual entry)
- Detect and quantify missing data with isnull, value_counts, and boolean math
- Clean missing values with fillna, forward/backward fill, interpolation, and dropna
- Know how aggregations behave with missing values (skipna)

Let us review pandas_missing_data_lecture.ipynb

Conclusion

- Missing data can come from files, merges, reindexing, or manual input
- Use .isnull(), .notnull(), .sum(), and .value_counts() to explore
- Fill values with .fillna(), .ffill(), .bfill(), .interpolate()
- Drop missing data with .dropna() if appropriate
- Calculations can ignore missing values using skipna=True
- pd.NA (experimental) offers a more consistent way to handle missingness