Contents

1 week 10: analysis and visualization

Introduction to pandas

- viewing data
  - info(), head(), tail(), shape, columns, describe()
  - selecting column syntax
  - .values()
  - slicing rows
- sorting data
  - value_counts('title') # department, artist display name
  - sort_values('accessionYear')
- filtering data
  - df['title'] == 'Woman'
    - creates series of booleans
  - woman = df['title'] == 'Woman'
    - creates new dataframe, df[woman]
  - highlights = df['isHighlight'] == 1.0
    - by boolean condition
  - df[highlights].info()
  - courbet = df['artistDisplayName'].str.contains('Courbet', na=False)
    - using str.contains
  - df[courbet]['title']

Intermediate pandas

- sorting values
  - df.sort_values('Bill Type', inplace=True)

  - see "pandas.DataFrame.sort_values" in docs to understand other parameters
  - sorted = df.sort_values([’Bill Type’, ’State’])
    - can sort by multiple values, create new df.

- filtering by values
  - Which states have the most book bans?
    - df[’Bill Type’] == ’Book Ban’
    - books = df[’Bill Type’] == ’Book Ban’
      - df[books]
    - df[books].value_counts(’State’)
  - `str.contains`: filtering by words
    - df[’Bill Type’].str.contains(’Bathroom’)
    - bathrooms = df[’Bill Type’].str.contains(’Bathroom’)
    - df[bathrooms].info()
    - df[bathrooms].value_counts(’State’)

- plotting data
  - df.plot(kind=’bar’)
  - df.value_counts(’Bill Type’).plot(kind=’bar’)
  - adding nlargest(10)
    - df.value_counts(’Bill Type’).nlargest(10).plot(kind=’bar’)
  - df.value_counts(’Bill Type’).nlargest(10).plot(kind=’barh’, xlabel=’Number of Bills’, title=’Most Frequent Categories for Anti-Trans Bills’)
  - df.value_counts(’Bill Type’).plot(kind=’pie’)