**Subsetting rows by categorical variables**

Subsetting data based on a categorical variable often involves using the "or" operator (|) to select rows from multiple categories. This can get tedious when you want all states in one of three different regions, for example. Instead, use the .isin() method, which will allow you to tackle this problem by writing one condition instead of three separate ones.

colors = ["brown", "black", "tan"]

condition = dogs["color"].isin(colors)

dogs[condition]

homelessness is available and pandas is loaded as pd.

# Subset for rows in South Atlantic or Mid-Atlantic regions

south\_mid\_atlantic = homelessnesshomelessness["region"].isin(["South Atlantic"]) | homelessness["region"].isin(["Mid-Atlantic"])

# See the result

print(homelessness[south\_mid\_atlantic]) # homelessness[condition]

Mutating, transforming, feature engineering ==== names for adding columns to a dataframe

Sorting

pd.sort\_values(“col\_name”)

or pd.sort\_values([“col\_name”])

* Add a column to homelessness, indiv\_per\_10k, containing the number of homeless individuals per ten thousand people in each state.
* Subset rows where indiv\_per\_10k is higher than 20, assigning to high\_homelessness.
* Sort high\_homelessness by descending indiv\_per\_10k, assigning to high\_homelessness\_srt.
* Select only the state and indiv\_per\_10k columns of high\_homelessness\_srt and save as result. *Look at the result.*

# Create indiv\_per\_10k col as homeless individuals per 10k state pop

homelessness["indiv\_per\_10k"] = 10000 \* homelessness["individuals"] / homelessness["state\_pop"]

# Subset rows for indiv\_per\_10k greater than 20

high\_homelessness = homelessness[homelessness["indiv\_per\_10k"] > 20]

# Sort high\_homelessness by descending indiv\_per\_10k

high\_homelessness\_srt = high\_homelessness.sort\_values("indiv\_per\_10k", ascending = False)

# From high\_homelessness\_srt, select the state and indiv\_per\_10k cols

result = high\_homelessness\_srt[["state", "indiv\_per\_10k"]]

# See the result

print(result)

Notes from above -> homelessness[condition] gives the rows

Condition gives true or false

Type(homelessness[“indiv\_per\_10k”] > 20) ---🡪 pandas.core.series.Series

Type(high\_homelessness) --🡪 pandas.core.frame.DataFrame

Also note -> sort\_values by descending --🡪 use ascending = False

* Aggregating .agg()
* Update the aggregation functions called by .agg(): include iqr and np.median in that order.
* # Import NumPy and create custom IQR function
* import numpy as np
* def iqr(column):
* return column.quantile(0.75) - column.quantile(0.25)
* # this works def npmedian(column):
* ## return np.median(column.to\_numpy())
* # Update to print IQR and median of temperature\_c, fuel\_price\_usd\_per\_l, & unemployment
* print(sales[["temperature\_c", "fuel\_price\_usd\_per\_l", "unemployment"]].agg([iqr, np.median])) # used npmedian
* # my code worked but the tester requires to pass np.median as a function into .agg

Sort\_values -> to sort the entire dataframe values by a column

Use df.sort\_values(“column\_name”)

Drop\_duplicates(subset = “column\_name”)

df.drop\_duplicates(subset=”column\_name\_you\_want\_to\_drop\_duplicates\_from”)

Dropping Duplicates

* Remove rows of sales with duplicate pairs of store and type and save as store\_types and print the head.
* Remove rows of sales with duplicate pairs of store and department and save as store\_depts and print the head.
* Subset the rows that are holiday weeks using the is\_holiday column, and drop the duplicate dates, saving as holiday\_dates.
* Select the date column of holiday\_dates, and print.
* # Drop duplicate store/type combinations
* store\_types = sales.drop\_duplicates(["store", "type"])
* print(store\_types.head())
* # Drop duplicate store/department combinations
* store\_depts = sales.drop\_duplicates(["store", "department"])
* print(store\_depts.head())
* # Subset the rows where is\_holiday is True and drop duplicate dates
* holiday\_dates = sales[sales["is\_holiday"] == True].drop\_duplicates(["date"])
* # Print date col of holiday\_dates
* print(holiday\_dates["date"])
* Count the number of stores of each store type in store\_types.
* Count the proportion of stores of each store type in store\_types.
* Count the number of different departments in store\_depts, sorting the counts in descending order.
* Count the proportion of different departments in store\_depts, sorting the proportions in descending order.

# Count the number of stores of each type

store\_counts = store\_types["type"].value\_counts()

print(store\_counts)

# Get the proportion of stores of each type

store\_props = store\_types["type"].value\_counts(normalize=True)

print(store\_props)

# Count the number of each department number and sort

dept\_counts\_sorted = store\_depts["department"].value\_counts()

print(dept\_counts\_sorted)

# Get the proportion of departments of each number and sort

dept\_props\_sorted = store\_depts["department"].value\_counts(sort=True, normalize=True)

print(dept\_props\_sorted)

* Import numpy with the alias np.
* Get the min, max, mean, and median of weekly\_sales for each store type using .groupby() and .agg(). Store this as sales\_stats. Make sure to use numpy functions!
* Get the min, max, mean, and median of unemployment and fuel\_price\_usd\_per\_l for each store type. Store this as unemp\_fuel\_stats.
* # Import numpy with the alias np
* import numpy as np
* # For each store type, aggregate weekly\_sales: get min, max, mean, and median
* sales\_stats = sales.groupby("type")["weekly\_sales"].agg([np.min, np.max, np.mean, np.median])
* # Print sales\_stats
* print(sales\_stats)
* # For each store type, aggregate unemployment and fuel\_price\_usd\_per\_l: get min, max, mean, and median
* unemp\_fuel\_stats = sales.groupby("type")["unemployment", "fuel\_price\_usd\_per\_l"].agg([np.min, np.max, np.mean, np.median])
* # Print unemp\_fuel\_stats
* print(unemp\_fuel\_stats)

PIVOT TABLE

* Print the mean weekly\_sales by department and type, filling in any missing values with 0.
* # Print mean weekly\_sales by department and type; fill missing values with 0
* print(sales.pivot\_table(values = "weekly\_sales", index = "type", columns = "department", fill\_value = 0))
* Print the mean weekly\_sales by department and type, filling in any missing values with 0 and summing all rows and columns.
* # Print the mean weekly\_sales by department and type; fill missing values with 0s; sum all rows and cols
* print(sales.pivot\_table(values="weekly\_sales", index="department", columns="type", fill\_value = 0, margins = True))

Setting and Removing indexes

* *Look at temperatures*.
* Set the index of temperatures to "city", assigning to temperatures\_ind.
* *Look at temperatures\_ind. How is it different from temperatures?*
* Reset the index of temperatures\_ind, keeping its contents.
* Reset the index of temperatures\_ind, dropping its contents.
* # Look at temperatures
* print(temperatures)
* # Index temperatures by city
* temperatures\_ind = temperatures.set\_index(["city"])
* # Look at temperatures\_ind
* print(temperatures\_ind)
* # Reset the index, keeping its contents
* print(temperatures\_ind.reset\_index())
* # Reset the index, dropping its contents
* print(temperatures\_ind.reset\_index(drop = True))
* Create a list called cities that contains "Moscow" and "Saint Petersburg".
* Use [] subsetting to filter temperatures for rows where the city column takes a value in the cities list.
* Use .loc[] subsetting to filter temperatures\_ind for rows where the city is in the cities list.
* # Make a list of cities to subset on
* cities = ["Moscow", "Saint Petersburg"]
* # Subset temperatures using square brackets
* print(temperatures[temperatures["city"].isin(cities)])
* # Subset temperatures\_ind using .loc[]
* print(temperatures\_ind.loc[["Moscow", "Saint Petersburg"]])
* Set the index of temperatures to the "country" and "city" columns, and assign this to temperatures\_ind.
* Specify two country/city pairs to keep: "Brazil"/"Rio De Janeiro" and "Pakistan"/"Lahore", assigning to rows\_to\_keep.
* Print and subset temperatures\_ind for rows\_to\_keep using .loc[]
* # Index temperatures by country & city
* temperatures\_ind = temperatures.set\_index(["country", "city"])
* # List of tuples: Brazil, Rio De Janeiro & Pakistan, Lahore
* rows\_to\_keep = [("Brazil", "Rio De Janeiro"), ("Pakistan", "Lahore")]
* # Subset for rows to keep
* print(temperatures\_ind.loc[rows\_to\_keep])
* Sort temperatures\_ind by the index values.
* Sort temperatures\_ind by the index values at the "city" level.
* Sort temperatures\_ind by ascending country then descending city.
* # Sort temperatures\_ind by index values
* print(temperatures\_ind.sort\_index())
* # Sort temperatures\_ind by index values at the city level
* print(temperatures\_ind.sort\_index(level = "city"))
* # Sort temperatures\_ind by country then descending city
* print(temperatures\_ind.sort\_index(level = ["country", "city"], ascending = [True, False]))
* Sort the index of temperatures\_ind.
* Use slicing with .loc[] to get these subsets:
  + from Pakistan to Russia.
  + from Lahore to Moscow. (*This will return nonsense.*)
  + from Pakistan, Lahore to Russia, Moscow.
* # Sort the index of temperatures\_ind
* temperatures\_srt = temperatures\_ind.sort\_index()
* # Subset rows from Pakistan to Russia
* print(temperatures\_srt.loc["Pakistan":"Russia"])
* # Try to subset rows from Lahore to Moscow
* print(temperatures\_srt.loc["Lahore":"Moscow"])
* # Subset rows from Pakistan, Lahore to Russia, Moscow
* print(temperatures\_srt.loc[("Pakistan", "Lahore"):("Russia", "Moscow")])
* Use .loc[] slicing to subset rows from India, Hyderabad to Iraq, Baghdad.
* Use .loc[] slicing to subset columns from date to avg\_temp\_c.
* Slice in both directions at once from Hyderabad to Baghdad, and date to avg\_temp\_c.
* # Subset rows from India, Hyderabad to Iraq, Baghdad
* print(temperatures\_srt.loc[("India", "Hyderabad"):("Iraq", "Baghdad")])
* # Subset columns from date to avg\_temp\_c
* print(temperatures\_srt.loc[:,"date":"avg\_temp\_c"])
* # Subset in both directions at once
* print(temperatures\_srt.loc[("India", "Hyderabad"):("Iraq", "Baghdad"), "date":"avg\_temp\_c"])
* Use Boolean conditions, not .isin() or .loc[], and the full date "yyyy-mm-dd", to subset temperatures for rows in 2010 and 2011 and print the results.
* Set the index to the date column and sort it.
* Use .loc[] to subset temperatures\_ind for rows in 2010 and 2011.
* Use .loc[] to subset temperatures\_ind for rows from Aug 2010 to Feb 2011.

Note when using Boolean and date time -> the full date needs to be used eg, “2010-01-01’ , just using the year will not work

# Use Boolean conditions to subset temperatures for rows in 2010 and 2011

temperatures\_bool = temperatures[(temperatures["date"] >= "2010-01-01") & (temperatures["date"] <= "2011-12-31")]

print(temperatures\_bool)

# Set date as the index and sort the index

temperatures\_ind = temperatures.set\_index("date").sort\_index()

# Use .loc[] to subset temperatures\_ind for rows in 2010 and 2011

print(temperatures\_ind.loc["2010":"2011"])

# Use .loc[] to subset temperatures\_ind for rows from Aug 2010 to Feb 2011

print(temperatures\_ind.loc["2010-08-01":"2011-02-01"])

Note above, also a thing about loc, it is end inclusive, so you’ll need the exact label for start and end.

Use .iloc[] on temperatures to take subsets.

* Get the 23rd row, 2nd column (index positions 22 and 1).
* Get the first 5 rows (index positions 0 to 5).
* Get all rows, columns 3 and 4 (index positions 2 to 4).
* Get the first 5 rows, columns 3 and 4.
* ####################
* # Get 23rd row, 2nd column (index 22, 1)
* print(temperatures.iloc[22, 1])
* # Use slicing to get the first 5 rows
* print(temperatures.iloc[:5])
* # Use slicing to get columns 3 to 4
* print(temperatures.iloc[:,2:4])
* # Use slicing in both directions at once
* print(temperatures.iloc[:5,2:4])
* Add a year column to temperatures, from the year component of the date column.
* Make a pivot table of the avg\_temp\_c column, with country and city as rows, and year as columns. Assign to temp\_by\_country\_city\_vs\_year, and *look at the result*.

# Add a year column to temperatures

temperatures["year"] = temperatures["date"].dt.year

# Pivot avg\_temp\_c by country and city vs year

temp\_by\_country\_city\_vs\_year = temperatures.pivot\_table(values = "avg\_temp\_c", index = ['country', 'city'], columns = 'year')

# See the result

print(temp\_by\_country\_city\_vs\_year)

Use .loc[] on temp\_by\_country\_city\_vs\_year to take subsets.

* From Egypt to India.
* From Egypt, Cairo to India, Delhi.
* From Egypt, Cairo to India, Delhi, and 2005 to 2010.

# Subset for Egypt to India

temp\_by\_country\_city\_vs\_year.loc["Egypt":"India"]

# Subset for Egypt, Cairo to India, Delhi

temp\_by\_country\_city\_vs\_year.loc[("Egypt", "Cairo"):("India", "Delhi")]

# Subset in both directions at once

temp\_by\_country\_city\_vs\_year.loc[("Egypt", "Cairo"):("India", "Delhi"), "2005":"2010"]

Pivot tables are filled with summary statistics, but they are only a first step to finding something insightful. Often you'll need to perform further calculations on them. A common thing to do is to find the rows or columns where the highest or lowest value occurs.

Recall from Chapter 1 that you can easily subset a Series or DataFrame to find rows of interest using a logical condition inside of square brackets. For example: series[series > value].

pandas is loaded as pd and the DataFrame temp\_by\_country\_city\_vs\_year is available.

##### Instructions

**70 XP**

* Calculate the mean temperature for each year, assigning to mean\_temp\_by\_year.
* Filter mean\_temp\_by\_year for the year that had the highest mean temperature.
* Calculate the mean temperature for each city (across columns), assigning to mean\_temp\_by\_city.
* Filter mean\_temp\_by\_city for the city that had the lowest mean temperature.
* # Get the worldwide mean temp by year
* mean\_temp\_by\_year = temp\_by\_country\_city\_vs\_year.mean(axis = "index")
* # Filter for the year that had the highest mean temp
* print(mean\_temp\_by\_year[mean\_temp\_by\_year == mean\_temp\_by\_year.max()])
* # Get the mean temp by city
* mean\_temp\_by\_city = temp\_by\_country\_city\_vs\_year.mean(axis = "columns")
* # Filter for the city that had the lowest mean temp
* print(mean\_temp\_by\_city[mean\_temp\_by\_city == mean\_temp\_by\_city.min()])
* Print the head of the avocados dataset. *What columns are available?*
* For each avocado size group, calculate the total number sold, storing as nb\_sold\_by\_size.
* Create a bar plot of the number of avocados sold by size.
* Show the plot.
* # Import matplotlib.pyplot with alias plt
* import matplotlib.pyplot as plt
* # Look at the first few rows of data
* print(avocados.head())
* # Get the total number of avocados sold of each size
* nb\_sold\_by\_size = avocados.groupby("size")["nb\_sold"].sum()
* # Create a bar plot of the number of avocados sold by size
* nb\_sold\_by\_size.plot(kind = "bar")
* # Show the plot
* plt.show()

# Import matplotlib.pyplot with alias plt

import matplotlib.pyplot as plt

# Get the total number of avocados sold on each date

nb\_sold\_by\_date = avocados.groupby("date")["nb\_sold"].sum()

# Create a line plot of the number of avocados sold by date

nb\_sold\_by\_date.plot(kind = "line")

# Show the plot

plt.show()

* Create a scatter plot with nb\_sold on the x-axis and avg\_price on the y-axis. Title it "Number of avocados sold vs. average price".
* # Scatter plot of nb\_sold vs avg\_price with title
* avocados.plot(x = "nb\_sold", y = "avg\_price", title = "Number of avocados sold vs. average price", kind = "scatter")
* # Show the plot
* plt.show()
* Subset avocados for the conventional type, and the average price column. Create a histogram.
* Create a histogram of avg\_price for organic type avocados.
* Add a legend to your plot, with the names "conventional" and "organic".
* Show your plot.
* # Histogram of conventional avg\_price
* avocados[avocados["type"] == "conventional"]["avg\_price"].hist()
* # Histogram of organic avg\_price
* avocados[avocados["type"] == "organic"]["avg\_price"].hist()
* # Add a legend
* plt.legend(["conventional", "organic"])
* # Show the plot
* plt.show()

Missing values

* Print a DataFrame that shows whether each value in avocados\_2016 is missing or not.
* Print a summary that shows whether *any* value in each column is missing or not.
* Create a bar plot of the total number of missing values in each column.
* # Import matplotlib.pyplot with alias plt
* import matplotlib.pyplot as plt
* # Check individual values for missing values
* print(avocados\_2016.isna())
* # Check each column for missing values
* print(avocados\_2016.isna().any())
* # Bar plot of missing values by variable
* avocados\_2016.isna().sum().plot(kind = "bar")
* # Show plot
* plt.show()
* In this exercise, you'll see how replacing missing values can affect the distribution of a variable using histograms. You can plot histograms for multiple variables at a time as follows:
* dogs[["height\_cm", "weight\_kg"]].hist()
* pandas has been imported as pd and matplotlib.pyplot has been imported as plt. The avocados\_2016 dataset is available.
* Replace the missing values of avocados\_2016 with 0s and store the result as avocados\_filled.
* Create a histogram of the cols\_with\_missing columns of avocados\_filled.
* # From previous step
* cols\_with\_missing = ["small\_sold", "large\_sold", "xl\_sold"]
* avocados\_2016[cols\_with\_missing].hist()
* plt.show()
* # Fill in missing values with 0
* avocados\_filled = avocados\_2016.fillna(0)
* # Create histograms of the filled columns
* avocados\_filled[cols\_with\_missing].hist()
* # Show the plot
* plt.show()
* For each airline group, select the nb\_bumped, and total\_passengers columns, and calculate the sum (for both years). Store this as airline\_totals.
* # From previous step
* airline\_bumping = pd.read\_csv("airline\_bumping.csv")
* print(airline\_bumping.head())
* # For each airline, select nb\_bumped and total\_passengers and sum
* airline\_totals = airline\_bumping.groupby("airline")[["nb\_bumped", "total\_passengers"]].sum()

Inner join

* Merge taxi\_owners with taxi\_veh on the column vid, and save the result to taxi\_own\_veh.
* # Merge the taxi\_owners and taxi\_veh tables
* taxi\_own\_veh = taxi\_owners.merge(taxi\_veh, on="vid")
* # Print the column names of the taxi\_own\_veh
* print(taxi\_own\_veh.columns)

suffix

# Merge the taxi\_owners and taxi\_veh tables setting a suffix

taxi\_own\_veh = taxi\_owners.merge(taxi\_veh, on='vid', suffixes =("\_own", "\_veh"))

# Print the column names of taxi\_own\_veh

print(taxi\_own\_veh.columns)

value counts

# Print the value\_counts to find the most popular fuel\_type

print(taxi\_own\_veh['fuel\_type'].value\_counts())

* Starting with the licenses table on the left, merge it to the biz\_owners table on the column account, and save the results to a variable named licenses\_owners.
* Group licenses\_owners by title and count the number of accounts for each title. Save the result as counted\_df
* Sort counted\_df by the number of **accounts** in **descending order**, and save this as a variable named sorted\_df.
* Use the .head() method to print the first few rows of the sorted\_df.
* # Merge the licenses and biz\_owners table on account
* licenses\_owners = licenses.merge(biz\_owners, on = "account")
* # Group the results by title then count the number of accounts
* counted\_df = licenses\_owners.groupby("title").agg({'account':'count'})
* # Sort the counted\_df in desending order
* sorted\_df = counted\_df.sort\_values("account", ascending = False)
* # Use .head() method to print the first few rows of sorted\_df
* print(sorted\_df.head())

Your goal is to find the total number of rides provided to passengers passing through the Wilson station (station\_name == 'Wilson') when riding Chicago's public transportation system on weekdays (day\_type == 'Weekday') in July (month == 7)

# Merge the ridership, cal, and stations tables

ridership\_cal\_stations = ridership.merge(cal, on=['year','month','day']) \

                            .merge(stations, on='station\_id')

# Create a filter to filter ridership\_cal\_stations

filter\_criteria = ((ridership\_cal\_stations['month'] == 7)

                   & (ridership\_cal\_stations['day\_type'] == "Weekday")

                   & (ridership\_cal\_stations['station\_name'] == "Wilson"))

# Use .loc and the filter to select for rides

print(ridership\_cal\_stations.loc[filter\_criteria, 'rides'].sum())

Three table merge

* Starting with the licenses table, merge to it the zip\_demo table on the zip column. Then merge the resulting table to the wards table on the ward column. Save result of the three merged tables to a variable named licenses\_zip\_ward.
* Group the results of the three merged tables by the column alderman and find the median income.
* # Merge licenses and zip\_demo, on zip; and merge the wards on ward
* licenses\_zip\_ward = licenses.merge(zip\_demo, on = "zip") \
* .merge(wards, on = "ward")
* # Print the results by alderman and show median income
* print(licenses\_zip\_ward.groupby("alderman").agg({'income':'median'}))
* Merge land\_use and census on the ward column. Merge the result of this with licenses on the ward column, using the suffix \_cen for the left table and \_lic for the right table. Save this to the variable land\_cen\_lic.
* # Merge land\_use and census and merge result with licenses including suffixes
* land\_cen\_lic = land\_use.merge(census, on = "ward")\
* .merge(licenses, on = "ward", suffixes = ("\_cen", "\_lic"))
* Group land\_cen\_lic by ward, pop\_2010 (the population in 2010), and vacant, then count the number of accounts. Save the results to pop\_vac\_lic.
* # Group by ward, pop\_2010, and vacant, then count the # of accounts
* pop\_vac\_lic = land\_cen\_lic.groupby(("ward", "pop\_2010", "vacant"),
* as\_index=False).agg({'account':'count'})
* Sort pop\_vac\_lic by vacant, account, andpop\_2010 in descending, ascending, and ascending order respectively. Save it as sorted\_pop\_vac\_lic.
* # Sort pop\_vac\_lic and print the results
* sorted\_pop\_vac\_lic = pop\_vac\_lic.sort\_values(["vacant", "account", "pop\_2010"],
* ascending=[False, True, False])

LEFT JOIN

* Count the number of rows in movies\_financials with a null value in the budget column.
* # Merge the movies table with the financials table with a left join
* movies\_financials = movies.merge(financials, on='id', how='left')
* # Count the number of rows in the budget column that are missing
* number\_of\_missing\_fin = movies\_financials['budget'].isnull().sum()
* # Print the number of movies missing financials
* print(number\_of\_missing\_fin)

USING left\_on and right\_on

* Merge movies and scifi\_only using the id column in the left table and the movie\_id column in the right table with an inner join.
* # Merge the movies and scifi\_only tables with an inner join
* movies\_and\_scifi\_only = movies.merge(scifi\_only, left\_on="id", right\_on="movie\_id")
* Merge movie\_to\_genres and pop\_movies using a right join. Save the results as genres\_movies.
* Group genres\_movies by genre and count the number of id values.
* # Use right join to merge the movie\_to\_genres and pop\_movies tables
* genres\_movies = movie\_to\_genres.merge(pop\_movies, how='right',
* left\_on = "movie\_id",
* right\_on = "id")
* # Count the number of genres
* genre\_count = genres\_movies.groupby('genre').agg({'id':'count'})
* # Plot a bar chart of the genre\_count
* genre\_count.plot(kind='bar')
* plt.show()