Python for Data Analysis and Visualization

Instructor: Claudia Carroll
Fall 2024

Session 2 (Oct 21)

Transdisciplinary Institute *in* Applied Data Sciences (TRIADS)

Arts & Sciences at Washington University in St. Louis
Signature Initiative

Today's Lesson Plan

1. Dealing with Null Data
2. Aggregating Data
3. Merging Dataframes

Demo 1:
Pandas Review &
Dealing with Null Data

Exercise 1: Calculating null cells

1. Create a new dataframe called SAFI_subset from the SAFI_results.csv that contains the columns respondent_roof_type, respondent_wall_type, respondent_wall_type_other, and respondent_floor_type.
2. Calculate the percentage of cells in the new dataframe that are null.

Hint: For part 2 you will have to remember you mathematical operators!!

Exercise 1 Part 1

```
df_SAFI = pd.read_csv("/content/drive/MyDrive/workshop_data/SAFI_results.csv")

df_SAFI_subset = df_SAFI[["C01_respondent_roof_type",
"C02_respondent_wall_type",
"C02_respondent_wall_type_other", "C03_respondent_floor_type"]]


df_SAFI_subset.columns = df_SAFI_subset.columns.str.replace(r'^.*?_', '',
regex=True)
```

Exercise 1 Part 1 (Alternative)

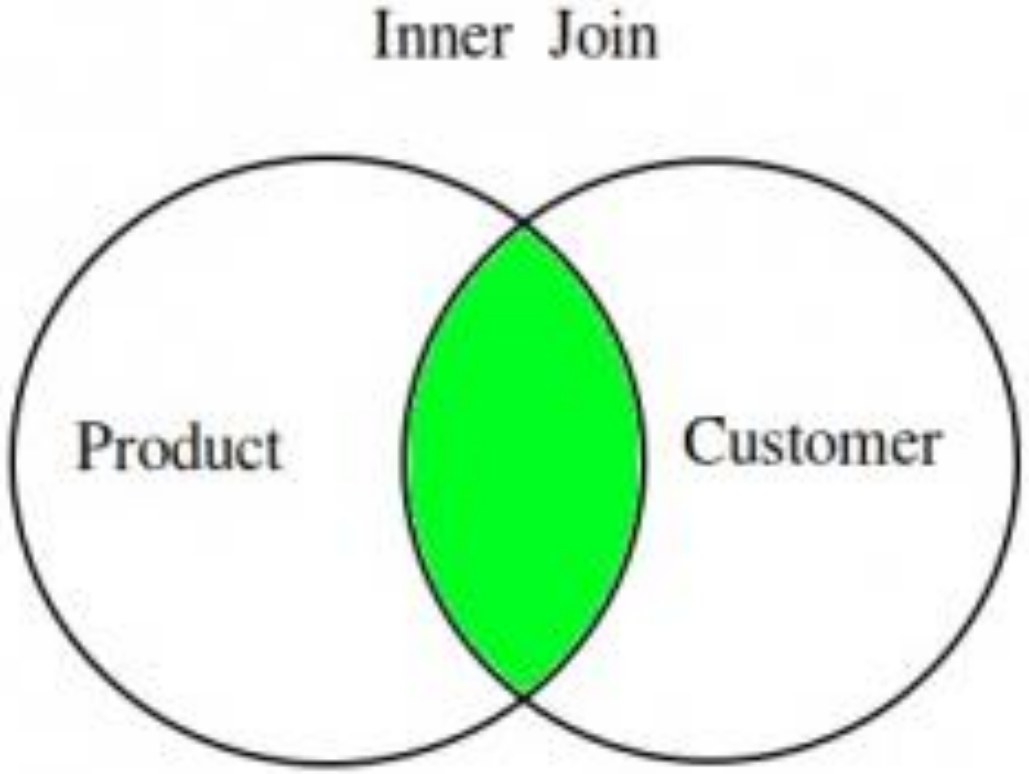
```
df_SAFI = pd.read_csv("/content/drive/MyDrive/workshop_data/SAFI_results.csv")

df_SAFI_subset = df_SAFI[["C01_respondent_roof_type", "C02_respondent_wall_type",
"C02_respondent_wall_type_other", "C03_respondent_floor_type"]]

df_SAFI_subset.rename(columns={'C01_respondent_roof_type':'respondent_roof_type'},
'C02_respondent_wall_type': 'respondent_wall_type'},
'C02_respondent_wall_type_other': 'respondent_wall_type_other'},
'C03_respondent_floor_type': 'respondent_floor_type'}, inplace=True)
```

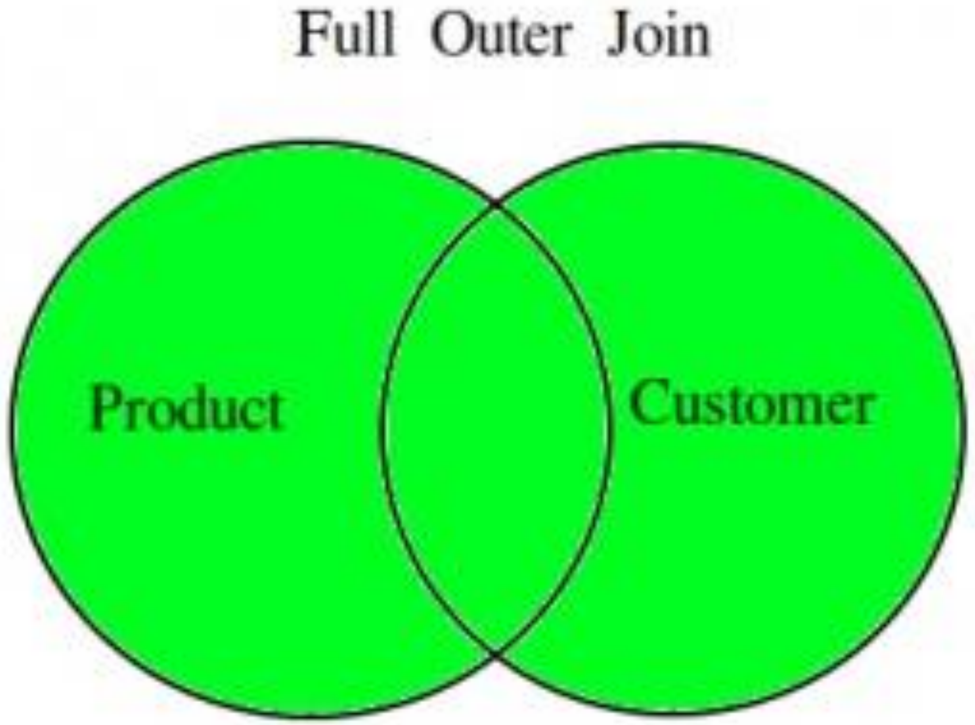
Exercise 1 Part 2

```
col_no = len(df_SAFI_subset.columns)

row_no = len(df_SAFI_subset.index)

total_cells = col_no * row_no

null_cells = 0

for x in df_SAFI_subset.isnull().sum():
    null_cells += x

percentage_null = ((null_cells/total_cells) * 100)

print(percentage_null)
```

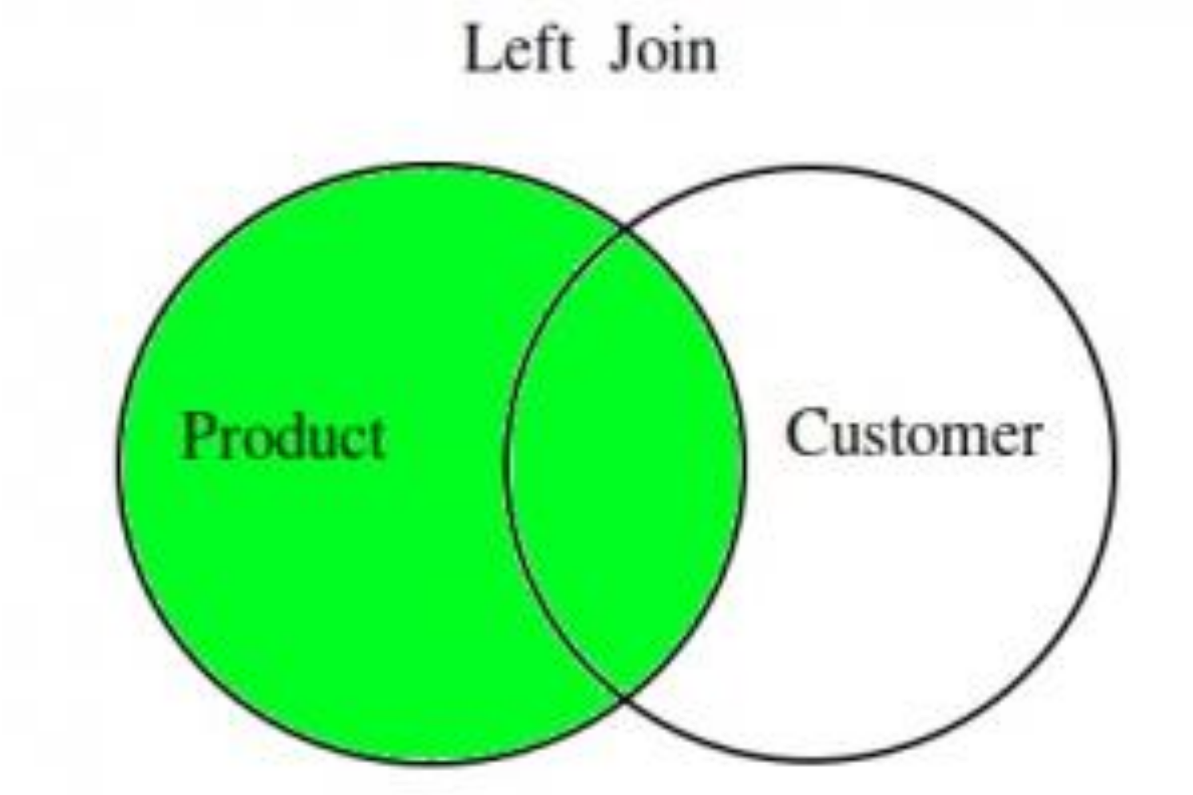
Demo 2: Aggregating Data

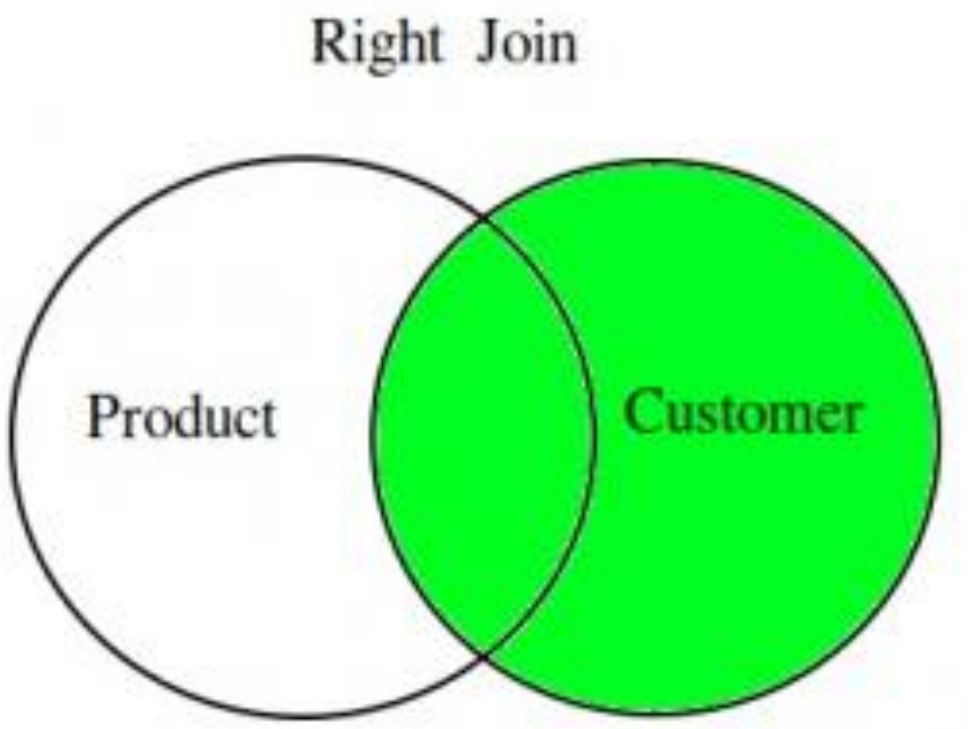
Joins: Inner and Outer





Joins: Left and Right





Exercise 2:

1. Read in the SAFI_results.csv dataset.

2. Get a list of the different respondent_wall_type values.

3. Groupby respondent_wall_type and describe the results.

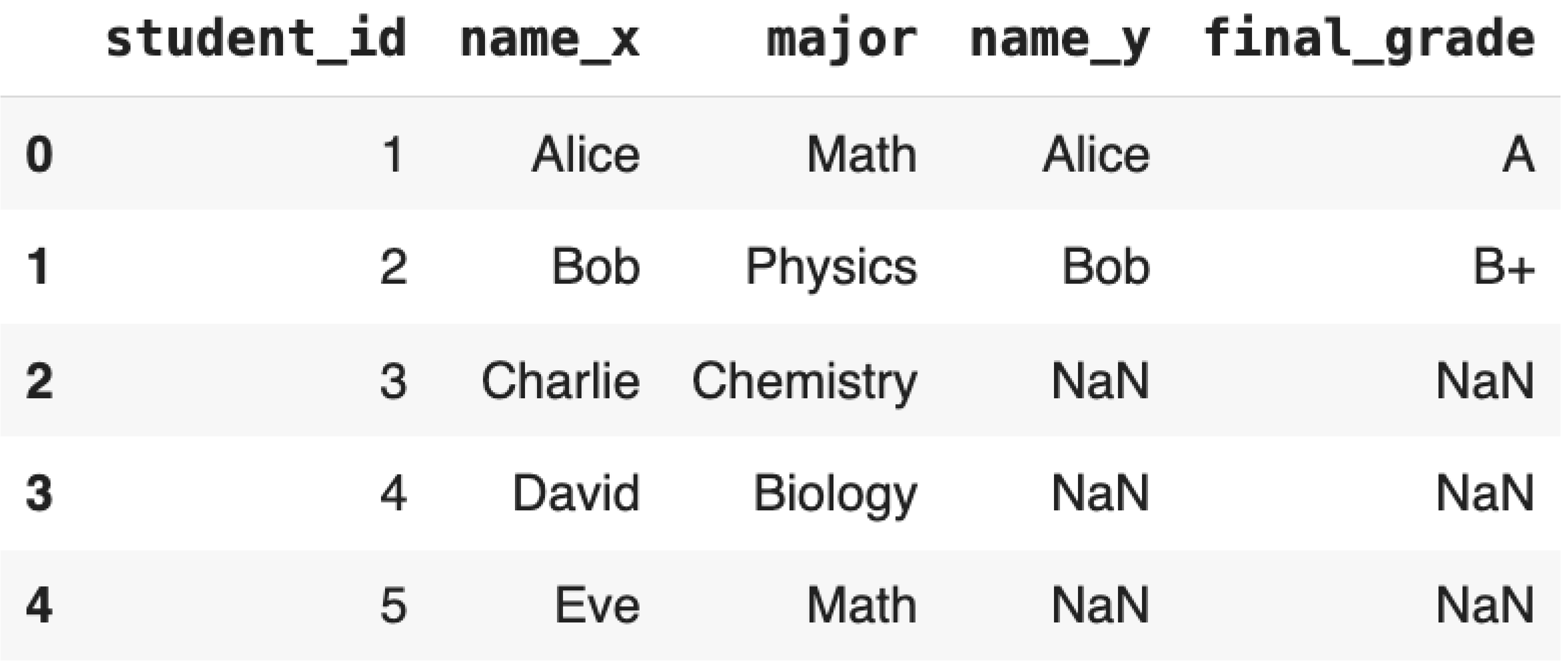
4. Create a new dataframe which is the result of an outer join of the grades and students dataframes using only the student ID column to join on. What do you notice about the column names in the new Dataframe?

Exercise 2 (Parts 1-3): Solution

```
df_SAFI =
pd.read_csv("/content/drive/MyDrive/workshop_data/SAFI_results_cleaned.csv",
index_col=0)

df_SAFI['respondent_wall_type'].unique()

grouped_data = df_SAFI.groupby('respondent_wall_type')

grouped_data.describe()
```

Exercise 2 (Part 4): Solution

```
merged_df = pd.merge(students_df, grades_df, on='student_id', how='left')

merged_df
```



	student_id	name_x	major	name_y	final_grade
0	1	Alice	Math	Alice	A
1	2	Bob	Physics	Bob	B+
2	3	Charlie	Chemistry	NaN	NaN
3	4	David	Biology	NaN	NaN
4	5	Eve	Math	NaN	NaN