

# **Assignment 3**

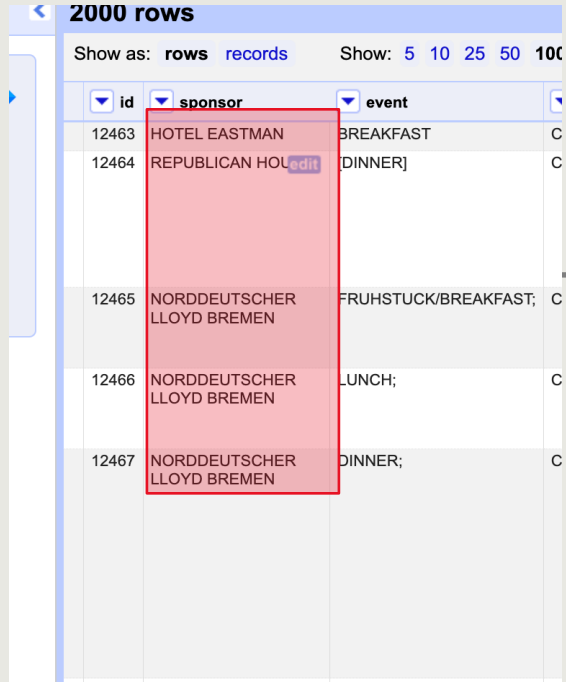
# **Data Cleaning**

Firdovsi Aliyev (UCID:30178471)

Lab B07

# Common Transformations

Before

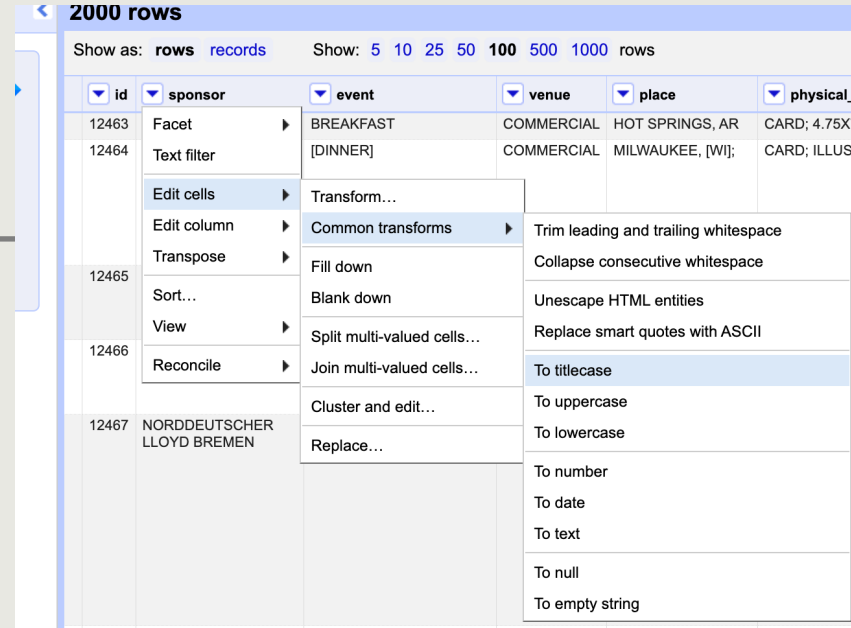


2000 rows

Show as: rows records Show: 5 10 25 50 100

id	sponsor	event	
12463	HOTEL EASTMAN	BREAKFAST	C
12464	REPUBLICAN HOUSE	[DINNER]	C
12465	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST;	C
12466	NORDDEUTSCHER LLOYD BREMEN	LUNCH;	C
12467	NORDDEUTSCHER LLOYD BREMEN	DINNER;	C

After

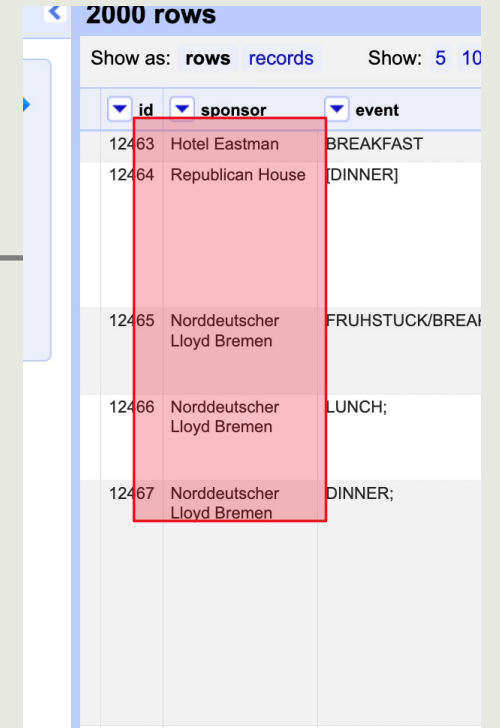


2000 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

id	sponsor	event	venue	place	physical_desc
12463	Facet	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR	CARD; 4.75X7.5;
12464	Text filter	[DINNER]	COMMERCIAL	MILWAUKEE, [WI];	CARD; ILLUS; CO
12465	Edit cells	Transform...			
12465	Edit column	Common transforms			
12465	Transpose	Fill down			
12465	Sort...	Blank down			
12466	View	Split multi-valued cells...			
12466	Reconcile	Join multi-valued cells...			
12467	NORDDEUTSCHER LLOYD BREMEN				

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- Replace smart quotes with ASCII
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- To null
- To empty string



2000 rows

Show as: rows records Show: 5 10

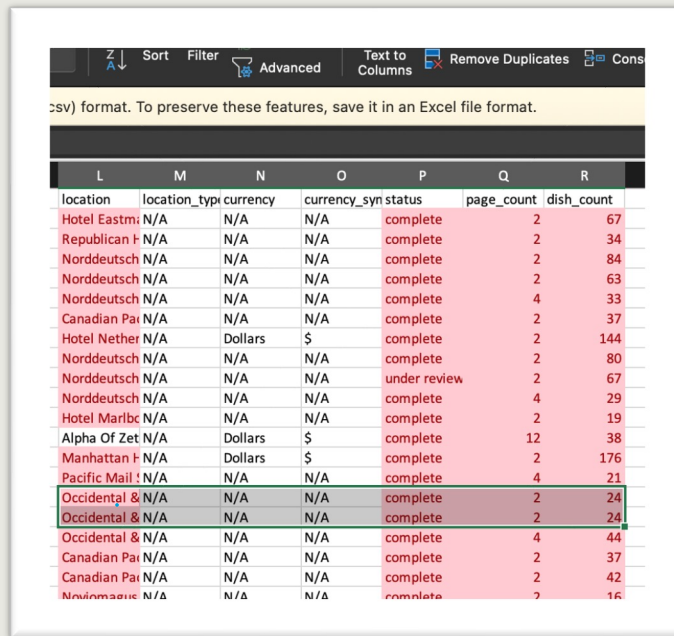
id	sponsor	event	
12463	Hotel Eastman	BREAKFAST	
12464	Republican House	[DINNER]	
12465	Norddeutscher Lloyd Bremen	FRUHSTUCK/BREA	
12466	Norddeutscher Lloyd Bremen	LUNCH;	
12467	Norddeutscher Lloyd Bremen	DINNER;	

I utilized OpenRefine's data transformation capabilities, applying common methods to ensure uniformity in text case. The objective was to standardize data formatting, resulting in all values being presented consistently in either lowercase or uppercase, which facilitates streamlined data analysis

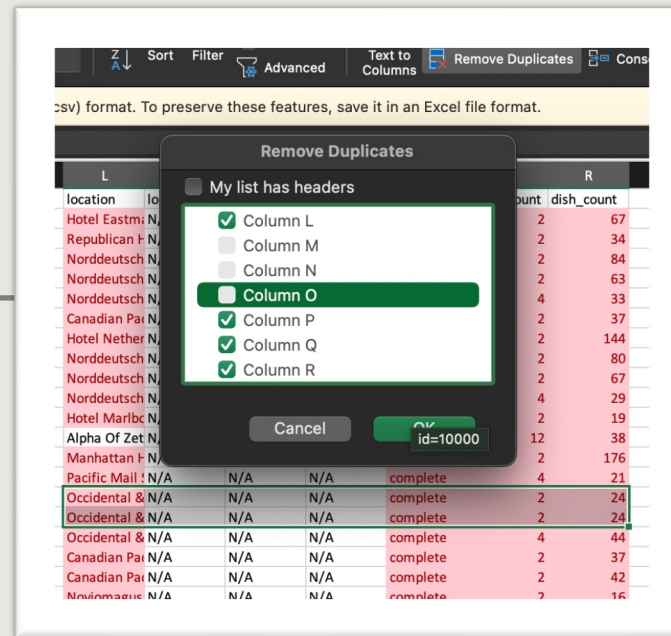
Before

# Removing Duplicates

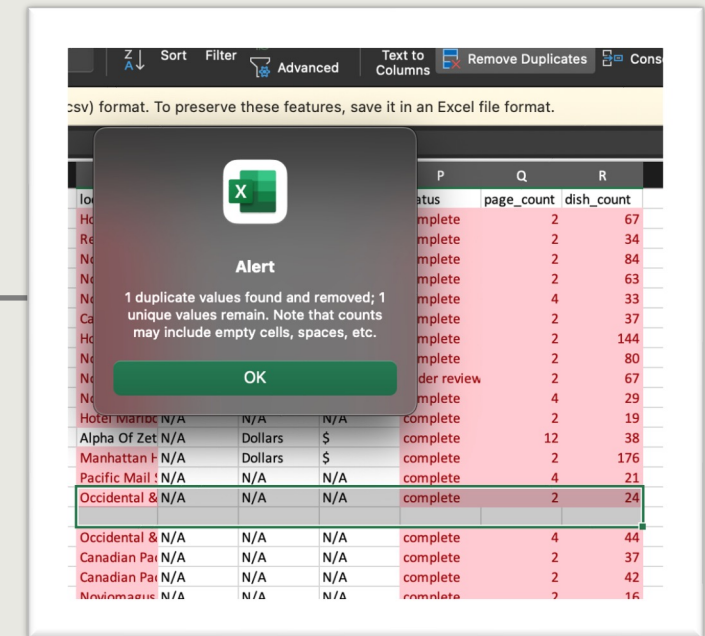
After



L	M	N	O	P	Q	R
location	location_type	currency	currency_symbol	status	page_count	dish_count
Hotel Eastm	N/A	N/A	N/A	complete	2	67
Republican H	N/A	N/A	N/A	complete	2	34
Norddeutsch	N/A	N/A	N/A	complete	2	84
Norddeutsch	N/A	N/A	N/A	complete	2	63
Norddeutsch	N/A	N/A	N/A	complete	4	33
Canadian Pai	N/A	N/A	N/A	complete	2	37
Hotel Nether	N/A	Dollars	\$	complete	2	144
Norddeutsch	N/A	N/A	N/A	complete	2	80
Norddeutsch	N/A	N/A	N/A	under review	2	67
Norddeutsch	N/A	N/A	N/A	complete	4	29
Hotel Marlbc	N/A	N/A	N/A	complete	2	19
Alpha Of Zet	N/A	Dollars	\$	complete	12	38
Manhattan H	N/A	Dollars	\$	complete	2	176
Pacific Mail	N/A	N/A	N/A	complete	4	21
Occidental &	N/A	N/A	N/A	complete	2	24
Occidental &	N/A	N/A	N/A	complete	2	24
Occidental &	N/A	N/A	N/A	complete	4	44
Canadian Pai	N/A	N/A	N/A	complete	2	37
Canadian Pai	N/A	N/A	N/A	complete	2	42
Norinmaque	N/A	N/A	N/A	complete	2	16



L	M	N	O	P	Q	R
location	location_type	currency	currency_symbol	status	page_count	dish_count
Hotel Eastm	N/A	N/A	N/A	complete	2	67
Republican H	N/A	N/A	N/A	complete	2	34
Norddeutsch	N/A	N/A	N/A	complete	2	84
Norddeutsch	N/A	N/A	N/A	complete	2	63
Norddeutsch	N/A	N/A	N/A	complete	4	33
Canadian Pai	N/A	N/A	N/A	complete	2	37
Hotel Nether	N/A	Dollars	\$	complete	2	144
Norddeutsch	N/A	N/A	N/A	complete	2	80
Norddeutsch	N/A	N/A	N/A	under review	2	67
Norddeutsch	N/A	N/A	N/A	complete	4	29
Hotel Marlbc	N/A	N/A	N/A	complete	2	19
Alpha Of Zet	N/A	Dollars	\$	complete	12	38
Manhattan H	N/A	Dollars	\$	complete	2	176
Pacific Mail	N/A	N/A	N/A	complete	4	21
Occidental &	N/A	N/A	N/A	complete	2	24
Occidental &	N/A	N/A	N/A	complete	2	24
Occidental &	N/A	N/A	N/A	complete	4	44
Canadian Pai	N/A	N/A	N/A	complete	2	37
Canadian Pai	N/A	N/A	N/A	complete	2	42
Norinmaque	N/A	N/A	N/A	complete	2	16



L	M	N	O	P	Q	R
location	location_type	currency	currency_symbol	status	page_count	dish_count
Hotel Eastm	N/A	N/A	N/A	complete	2	67
Republican H	N/A	N/A	N/A	complete	2	34
Norddeutsch	N/A	N/A	N/A	complete	2	84
Norddeutsch	N/A	N/A	N/A	complete	2	63
Norddeutsch	N/A	N/A	N/A	complete	4	33
Canadian Pai	N/A	N/A	N/A	complete	2	37
Hotel Nether	N/A	Dollars	\$	complete	2	144
Norddeutsch	N/A	N/A	N/A	complete	2	80
Norddeutsch	N/A	N/A	N/A	under review	2	67
Norddeutsch	N/A	N/A	N/A	complete	4	29
Hotel Marlbc	N/A	N/A	N/A	complete	2	19
Alpha Of Zet	N/A	Dollars	\$	complete	12	38
Manhattan H	N/A	Dollars	\$	complete	2	176
Pacific Mail	N/A	N/A	N/A	complete	4	21
Occidental &	N/A	N/A	N/A	complete	2	24
Occidental &	N/A	N/A	N/A	complete	4	44
Canadian Pai	N/A	N/A	N/A	complete	2	37
Canadian Pai	N/A	N/A	N/A	complete	2	42
Norinmaque	N/A	N/A	N/A	complete	2	16

Duplicate data, whether in rows, values, or columns, can hinder data usability. These duplicates, often the result of inadvertent repetition, were removed from the dataset to prevent inaccuracies in analysis. This "Remove Duplicates" process streamlines data analysis, facilitating quicker decision-making and analysis.

Before

# Refining data via Clustering

id	sponsor	event	venue	place
12463	HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR
12464	REPUBLICAN HOUSE	DINNER	COMMERCIAL	MILWAUKEE, WI
12465	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12466	NORDDEUTSCHER LLOYD BREMEN	LUNCH	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12467	NORDDEUTSCHER LLOYD BREMEN	DINNER;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12468	CANADIAN PACIFIC RAILWAY COMPANY	DINNER	COMMERCIAL	R.M.S. EMPIRE OF CHINA
12469	HOTEL NETHERLAND	SUPPER	COMMERCIAL	NEW YORK, NY
12470	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE

id	sponsor	event	venue	place
12463	HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR
12464	REPUBLICAN HOUSE	DINNER	COMMERCIAL	MILWAUKEE, WI
12465	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12466	NORDDEUTSCHER LLOYD BREMEN	LUNCH	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12467	NORDDEUTSCHER LLOYD BREMEN	DINNER;	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12468	CANADIAN PACIFIC RAILWAY COMPANY	DINNER	COMMERCIAL	R.M.S. EMPIRE OF CHINA
12469	HOTEL NETHERLAND	SUPPER	COMMERCIAL	NEW YORK, NY
12470	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK/BREAKFAST	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE

Cluster and edit column "event"

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: Key collision Keying function: Fingerprint 17 clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
5	357	<ul style="list-style-type: none"><li>DINNER (330 rows)</li><li>[DINNER] (22 rows)</li><li>DINNER; (3 rows)</li><li>DINNER (?)</li><li>[?DINNER?]</li></ul>	<input checked="" type="checkbox"/>	DINNER
5	208	<ul style="list-style-type: none"><li>BREAKFAST (201 rows)</li><li>[BREAKFAST] (4 rows)</li><li>BREAKFAST (?)</li><li>BREAKFAST(?)</li><li>BREAKFAST;</li></ul>	<input checked="" type="checkbox"/>	BREAKFAST
4	40	<ul style="list-style-type: none"><li>SUPPER (35 rows)</li><li>SUPPER (?) (2 rows)</li><li>SUPPER(?) (2 rows)</li><li>SUPPER;</li></ul>	<input checked="" type="checkbox"/>	SUPPER
3	8	<ul style="list-style-type: none"><li>TABLE D'HOTE DINNER (6 rows)</li><li>DINNER TABLE D'HOTE</li><li>TABLE d'HOTE DINNER</li></ul>	<input checked="" type="checkbox"/>	TABLE D'HOTE DINNER
3	102	<ul style="list-style-type: none"><li>LUNCH (90 rows)</li><li>[LUNCH] (8 rows)</li><li>LUNCH;</li></ul>	<input checked="" type="checkbox"/>	LUNCH

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

The "Cluster and Edit" functionality in OpenRefine represents a salient capability, facilitating the transformation of nearly identical words into a singular canonical form. This feature harnesses an algorithm to ascertain potential word equivalence. Cleaning these entities is crucial to ensure data accuracy and consistency, eliminating inaccuracies and redundancies. This, in turn, leads to more efficient and reliable data analysis, enabling the extraction of meaningful insights.

After

id	sponsor	event	venue	place
12463	HOTEL EASTMAN	BREAKFAST	COMMERCIAL	HOT SPRINGS, AR
12464	REPUBLICAN HOUSE	DINNER	COMMERCIAL	MILWAUKEE, WI
12465	NORDDEUTSCHER LLOYD BREMEN	FRUHSTUCK-BREAKFAST	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12466	NORDDEUTSCHER LLOYD BREMEN	LUNCH	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE
12467	NORDDEUTSCHER LLOYD BREMEN	DINNER	COMMERCIAL	DAMPFER KAISER WILHELM DER GROSSE

# Before

# Before

[illegible]

The screenshot shows the Excel ribbon with the 'Home' tab selected. The 'Cells' dropdown menu is open, and the 'Delete' option is highlighted in green. The menu includes options like Cut, Copy, Paste, Paste Special, Insert, Delete, Clear Contents, Format Cells..., Column Width..., Hide, Unhide, AutoFill, and Services. The background shows a spreadsheet with columns labeled 'M' and 'N'.

After

[illegible]

Data cleaning typically employs one or two techniques. In this context, the "location type" column is superfluous as it contains no data. The rationale for removing an empty column is driven by the recognition that this data serves no purpose, and eliminating it enhances the dataset's clarity and efficiency.

# Find and Replace function

Before

After

A	B	C	D	E	F	G	H	I	J
12579	TRUSTEES O 11TH ANNU/ PROF;			SOUTHERN F BROADSIDE; ANNUAL		WINES LISTE 1900-2627			3/31/1900
12583									4/15/1900
12584									4/15/1900
12585									4/16/1900
12586									4/16/1900
12587									4/16/1900
12588									4/16/1900
12589									4/16/1900
12590									4/17/1900
12591									4/17/1900
12592									4/17/1900

A	B	C	D	E	F	G	H	I	J
12579	TRUSTEES O 11TH ANNU/ PROF;			SOUTHERN F BROADSIDE; ANNUAL		WINES LISTE 1900-2627			3/31/1900
12583	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/15/1900
12584	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/15/1900
12585	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/16/1900
12586	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/16/1900
12587	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/16/1900
12588	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/16/1900
12589	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/16/1900
12590	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/17/1900
12591	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	4/17/1900

A	B	C	D	E	F	G	H	I	J
12578	HADDON HA DINNER			COMMERCIA (PHILADELPH BROADSIDE; DAILY		1900-2626		3/31/1900	Haddon
12579	TRUSTEES O 11TH ANNU/ PROF;			SOUTHERN F BROADSIDE; ANNUAL		WINES LISTE 1900-2627		3/31/1900	Trustee
12583								4/15/1900	Hotel E;
12584								4/15/1900	Republi
12585								4/16/1900	Nordde
12586								4/16/1900	Nordde
12587								4/16/1900	Nordde
12588								4/16/1900	Canadia
12589								4/16/1900	Hotel N
12590								4/17/1900	Nordde

Find & Replace

Find

Replace

Find what:

Replace with:

N/A

Options

Replace

Replace All

Find All

Previous

Next

Close

The "Find and Replace" method, an inherent capability of Excel, simplifies the data cleaning and analysis process, ensuring data cleanliness. This technique was selected to promote dataset purity and uniformity, enabling reliable conclusions when analyzing the cleansed data.