Microsoft Excel Data Cleaning Project Report
Presented by Rennie Mancho

1. Project Overview

This project involved cleaning and preparing a messy dataset for analysis using Microsoft Excel. The dataset contained inconsistent entries, spacing issues, missing or incorrect values, and formatting problems. The objective was to ensure the data was accurate, consistent, and ready for further analysis.

2. Data Overview

This dataset as seen below in fig 1 contained retail transaction records constituting of dates, IDs, customer names, product details, quantities, unit prices, regions and ratings.

- Source file: Excel_DirtyData.xlsx
- Initial number of rows: 32
- Initial number of columns: 8
- Issues observed: extra letter spacing, blank spaces, numbers stored as text, mixed ID numbering, spelling errors, duplicate information, improper column/row width, Name casing, no column indicating total of purchase.

Date	ID	Name	Region	Rating	Product	Quantity	Price Per Unit
######	1	John Smith	North	Good	Magic Wand	10	$20.00
######	2	Jane Doe	East	Excelent	Unicorn Horn	15	$10.00
######	3	Mike Tyson	West	Poor	Boxing Gloves	0	inf
######	4	Anna Belle	South	Average	Fairy Dust	25	$10.00
######	5	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67
######	6	Peter Parker		Excelent	Web Shooter	0	inf
######	7	Mary Jane	West	Poor	Potent Potion	35	$10.00
######	8	Bruce Wayne	South	Average	Bat Signal	40	$15.00
######	9	Clark Kent	East	Good	Glasses with X-ray Vision	45	$12.22
######	10	Diana Prince	North	Excelent	Lasso of Truth	50	$14.00
######	11	Tony Stark	West	Poor	Iron Man Suit	5	$160.00
######	12	Steve Rogers	South	Average	Captain America Shield	20	$45.00
######	13	Natasha Romanoff	East	Good	Black Widow's Bite	0	inf
######	14	Bruce Banner		Excelent	Gamma Radiation Serum	30	$36.67
######	4	Anna Belle	South	Average	Fairy Dust	25	$10.00
######	5	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67
######	6	Peter Parker		Excelent	Web Shooter	0	inf
######	15	Nick Fury	West	Poor	Eye Patch	35	$34.29
######	16	Phil Coulson		Average	Agent ID Card	0	inf
######	17	Peggy Carter	East	Good	Vintage Pistol	40	$35.00
######	18	Howard Stark	North	Excelent	Arc Reactor	45	$33.33
######	19	Hank Pym	West	Poor	Ant-Man Suit	50	$32.00
######	20	Janet van Dyne	South	Average	Wasp's Wings	55	$30.91
######	21	Kurt Busiek	East	Good	Comic Book	60	$30.00
######	22	George Perez	North	Excelent	Drawing Pad	0	inf
######	23	Roger Stern	West	Poor	Notepads	65	$30.77
######	24	Tom DeFalco	South	Average	Pen Set	70	$30.00
######	25	Loki Laufeyson	Asgard	Mischief	Trickster's Hat	75	$29.33
######	26	Thor Odinson	Asgard	Worthy	Mjolnir	80	$28.75
######	27	Natasha Romanoff	East	Spy	Spy Kit	0	inf
######	28	Steve Rogers	South	Leader	Leadership Manual	85	$29.41

Fig 1: Messy data set in Microsoft excel

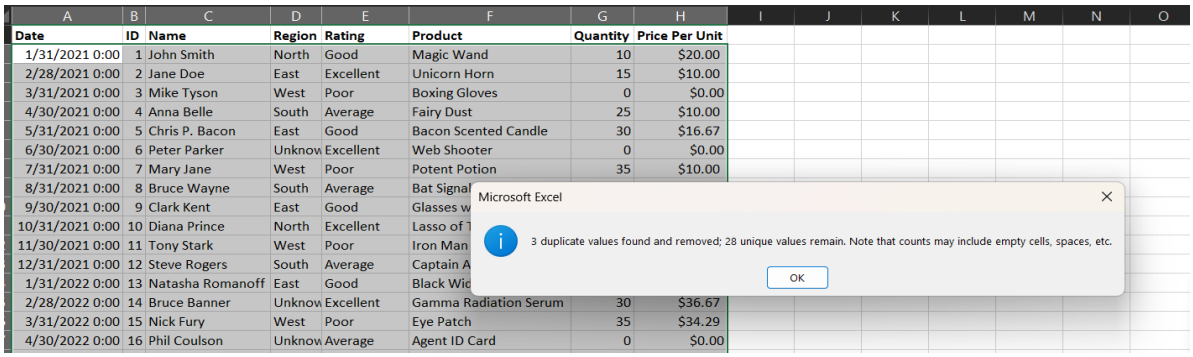
3. Data Cleaning Steps/Methodology

The following steps were taken to clean and prepare the dataset:

1. Autofit columns and rows to improve readability.

2. Identified and removed duplicate rows using 'Remove Duplicates' tool.

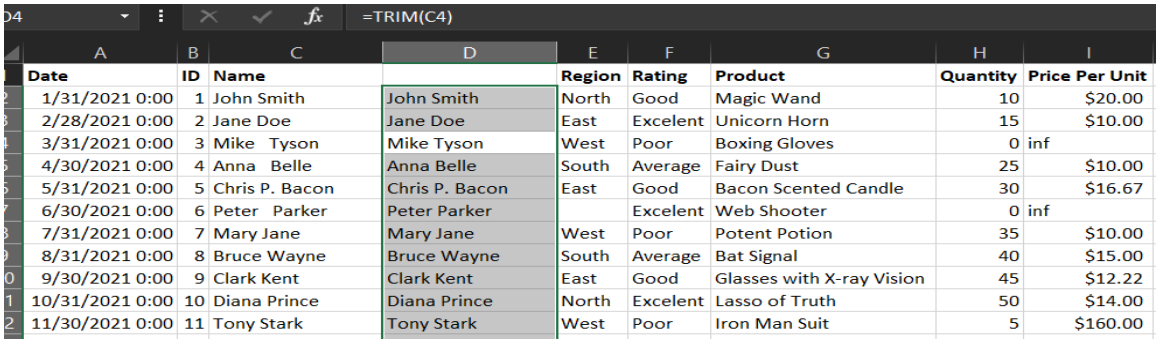
- Removed 3 duplicate rows as seen in fig 2.



Date	ID	Name	Region	Rating	Product	Quantity	Price Per Unit
1/31/2021 0:00	1	John Smith	North	Good	Magic Wand	10	$20.00
2/28/2021 0:00	2	Jane Doe	East	Excellent	Unicorn Horn	15	$10.00
3/31/2021 0:00	3	Mike Tyson	West	Poor	Boxing Gloves	0	$0.00
4/30/2021 0:00	4	Anna Belle	South	Average	Fairy Dust	25	$10.00
5/31/2021 0:00	5	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67
6/30/2021 0:00	6	Peter Parker	Unknow	Excellent	Web Shooter	0	$0.00
7/31/2021 0:00	7	Mary Jane	West	Poor	Potent Potion	35	$10.00
8/31/2021 0:00	8	Bruce Wayne	South	Average	Bat Signal		
9/30/2021 0:00	9	Clark Kent	East	Good	Glasses w		
10/31/2021 0:00	10	Diana Prince	North	Excellent	Lasso of T		
11/30/2021 0:00	11	Tony Stark	West	Poor	Iron Man		
12/31/2021 0:00	12	Steve Rogers	South	Average	Captain A		
1/31/2022 0:00	13	Natasha Romanoff	East	Good	Black Wid		
2/28/2022 0:00	14	Bruce Banner	Unknow	Excellent	Gamma Radiation Serum	30	$36.67
3/31/2022 0:00	15	Nick Fury	West	Poor	Eye Patch	35	$34.29
4/30/2022 0:00	16	Phil Coulson	Unknow	Average	Agent ID Card	0	$0.00

Fig 2: Finding & Removing Duplicates

3. Trimmed extra spaces in the Names column using the =TRIM() function by creating a helper column, and later pasting the corrected values as seen in Fig 3.

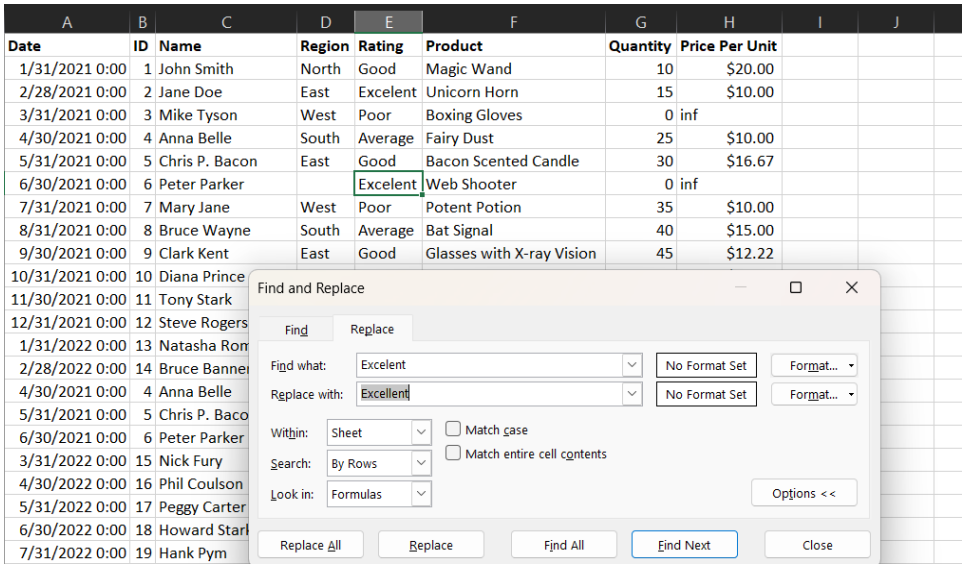


Date	ID	Name		Region	Rating	Product	Quantity	Price Per Unit
1/31/2021 0:00	1	John Smith	John Smith	North	Good	Magic Wand	10	$20.00
2/28/2021 0:00	2	Jane Doe	Jane Doe	East	Excelent	Unicorn Horn	15	$10.00
3/31/2021 0:00	3	Mike Tyson	Mike Tyson	West	Poor	Boxing Gloves	0	inf
4/30/2021 0:00	4	Anna Belle	Anna Belle	South	Average	Fairy Dust	25	$10.00
5/31/2021 0:00	5	Chris P. Bacon	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67
6/30/2021 0:00	6	Peter Parker	Peter Parker		Excelent	Web Shooter	0	inf
7/31/2021 0:00	7	Mary Jane	Mary Jane	West	Poor	Potent Potion	35	$10.00
8/31/2021 0:00	8	Bruce Wayne	Bruce Wayne	South	Average	Bat Signal	40	$15.00
9/30/2021 0:00	9	Clark Kent	Clark Kent	East	Good	Glasses with X-ray Vision	45	$12.22
10/31/2021 0:00	10	Diana Prince	Diana Prince	North	Excelent	Lasso of Truth	50	$14.00
11/30/2021 0:00	11	Tony Stark	Tony Stark	West	Poor	Iron Man Suit	5	$160.00

Fig 3: Using the Trim() function

4. Replaced blank cells with appropriate default values (e.g. Unknown for blank cells in the regions and ratings column),

5. Used find and replace to correct data entry errors as seen in fig 4.



A	B	C	D	E	F	G	H	I	J
Date	**ID**	**Name**	**Region**	**Rating**	**Product**	**Quantity**	**Price Per Unit**		
1/31/2021 0:00	1	John Smith	North	Good	Magic Wand	10	$20.00		
2/28/2021 0:00	2	Jane Doe	East	Excelent	Unicorn Horn	15	$10.00		
3/31/2021 0:00	3	Mike Tyson	West	Poor	Boxing Gloves	0	inf		
4/30/2021 0:00	4	Anna Belle	South	Average	Fairy Dust	25	$10.00		
5/31/2021 0:00	5	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67		
6/30/2021 0:00	6	Peter Parker		Excelent	Web Shooter	0	inf		
7/31/2021 0:00	7	Mary Jane	West	Poor	Potent Potion	35	$10.00		
8/31/2021 0:00	8	Bruce Wayne	South	Average	Bat Signal	40	$15.00		
9/30/2021 0:00	9	Clark Kent	East	Good	Glasses with X-ray Vision	45	$12.22		
10/31/2021 0:00	10	Diana Prince							
11/30/2021 0:00	11	Tony Stark							
12/31/2021 0:00	12	Steve Rogers							
1/31/2022 0:00	13	Natasha Ron							
2/28/2022 0:00	14	Bruce Banne							
4/30/2021 0:00	4	Anna Belle							
5/31/2021 0:00	5	Chris P. Baco							
6/30/2021 0:00	6	Peter Parker							
3/31/2022 0:00	15	Nick Fury							
4/30/2022 0:00	16	Phil Coulson							
5/31/2022 0:00	17	Peggy Carter							
6/30/2022 0:00	18	Howard Stark							
7/31/2022 0:00	19	Hank Pym							

Fig 4: Find & Replace

6. Capitalized properly text in the Name column using =PROPER()

7. Converted the cleaned dataset into an Excel Table for easier filtering and analysis.

8. Validated the data using Data Validation rules to ensure correctness and consistency. Made use of "lists":

- Regions: North, West, South, East, Asgard, Unknown
- Ratings: Poor, Average, Good, Excellent, Unknown

8. Added a total column to ease further analysis.

4. Key Formulas Used

Formulas used during cleaning:

- =TRIM(A2) – to remove extra spaces.
- =PROPER(A2) – to capitalize names or text properly.
- Used Find and Replace to correct data entry errors.
- Remove Duplicates.

5. Challenges and Resolutions

1. Finding the duplicates due to how clustered the row width was.
   - Solution: made use of autofit row & column, easing readability

6. Final Dataset Summary

- File Name: Excel_CleanData.xlsx
- Final shape: 30 rows x 9 columns
- Records removed: 3 rows of duplicate data

- 7 Blank cells replaced with value “unknown”
- Data is now in a table and can be filtered
- Validation done on: Region, Ratings columns
- Seen in figure 5 below.

RETAIL SALES AND CUSTOMER TRANSACTION DATA- CLEANED DATASET								
Date	ID	Name	Region	Rating	Product	Quantity	Price Per Unit	Total
1/31/21 12:00 AM	1	John Smith	North	Good	Magic Wand	10	$20.00	$200.00
2/28/21 12:00 AM	2	Jane Doe	East	Excellent	Unicorn Horn	15	$10.00	$150.00
3/31/21 12:00 AM	3	Mike Tyson	West	Poor	Boxing Gloves	0	$0.00	$0.00
4/30/21 12:00 AM	4	Anna Belle	South	Average	Fairy Dust	25	$10.00	$250.00
5/31/21 12:00 AM	5	Chris P. Bacon	East	Good	Bacon Scented Candle	30	$16.67	$500.10
6/30/21 12:00 AM	6	Peter Parker	Unknown	Excellent	Web Shooter	0	$0.00	$0.00
7/31/21 12:00 AM	7	Mary Jane	West	Poor	Potent Potion	35	$10.00	$350.00
8/31/21 12:00 AM	8	Bruce Wayne	South	Average	Bat Signal	40	$15.00	$600.00
9/30/21 12:00 AM	9	Clark Kent	East	Good	Glasses with X-ray Vision	45	$12.22	$549.90
10/31/21 12:00 AM	10	Diana Prince	North	Excellent	Lasso of Truth	50	$14.00	$700.00
11/30/21 12:00 AM	11	Tony Stark	West	Poor	Iron Man Suit	5	$160.00	$800.00
12/31/21 12:00 AM	12	Steve Rogers	South	Average	Captain America Shield	20	$45.00	$900.00
1/31/22 12:00 AM	13	Natasha Romanoff	East	Good	Black Widow's Bite	0	$0.00	$0.00
2/28/22 12:00 AM	14	Bruce Banner	Unknown	Excellent	Gamma Radiation Serum	30	$36.67	$1,100.10
3/31/22 12:00 AM	15	Nick Fury	West	Poor	Eye Patch	35	$34.29	$1,200.15
4/30/22 12:00 AM	16	Phil Coulson	Unknown	Average	Agent ID Card	0	$0.00	$0.00
5/31/22 12:00 AM	17	Peggy Carter	East	Good	Vintage Pistol	40	$35.00	$1,400.00
6/30/22 12:00 AM	18	Howard Stark	North	Excellent	Arc Reactor	45	$33.33	$1,499.85
7/31/22 12:00 AM	19	Hank Pym	West	Poor	Ant-Man Suit	50	$32.00	$1,600.00
8/31/22 12:00 AM	20	Janet Van Dyne	South	Average	Wasp's Wings	55	$30.91	$1,700.05
9/30/22 12:00 AM	21	Kurt Busiek	East	Good	Comic Book	60	$30.00	$1,800.00
10/31/22 12:00 AM	22	George Perez	North	Excellent	Drawing Pad	0	$0.00	$0.00
11/30/22 12:00 AM	23	Roger Stern	West	Poor	Notepads	65	$30.77	$2,000.05
12/31/22 12:00 AM	24	Tom DeFalco	South	Average	Pen Set	70	$30.00	$2,100.00
1/31/23 12:00 AM	25	Loki Laufeyson	Asgard	Unknown	Trickster's Hat	75	$29.33	$2,199.75
2/28/23 12:00 AM	26	Thor Odinson	Asgard	Unknown	Mjolnir	80	$28.75	$2,300.00
3/31/23 12:00 AM	27	Natasha Romanoff	East	Unknown	Spy Kit	0	$0.00	$0.00
4/30/23 12:00 AM	28	Steve Rogers	South	Unknown	Leadership Manual	85	$29.41	$2,499.85

Fig 5: Cleaned Final data set

7. Conclusion

The dataset is now clean, consistent, and ready for further analysis or visualization. All identified issues have been addressed, and appropriate formatting has been applied to ensure clarity.