IMDB Dataset Cleaning – Report

A comprehensive data cleaning case study transforming a real-world messy IMDB dataset into analysis-ready outputs.

https://github.com/oliviadita/messy-imdb

SAMPLE BY OLIVIA D.K.

1. Project Overview

This project focuses on cleaning and standardizing a messy IMDB movie dataset sourced from Kaggle. The dataset presents realistic data quality issues commonly found in production environments, including encoding problems, malformed dates, inconsistent categorical values, and mixed numeric formats.

Data Source:

https://www.kaggle.com/datasets/davidfuenteherraiz/messy-imdb-dataset

2. Raw Dataset Summary

- Original shape: 101 rows × 12 columns
- Expected (per Kaggle description): 100 movies, 11 columns
- File characteristics:
  - Non-UTF8 encoding
  - Semicolon (;) delimiter

Key issues identified:

- Messy and inconsistent column headers
- Encoding artifacts (mojibake)
- Invalid and malformed date values
- Inconsistent country and genre labels
- Empty column and fully empty row

3. Structural Cleaning

Encoding and Delimiter Handling:

The dataset required explicit handling of non-UTF8 encoding and semicolon delimiters to ensure correct parsing.

Column Name Standardization:
All column names were cleaned by fixing encoding artifacts, removing extra spaces, replacing spaces with underscores, and converting names to snake_case.

Column and Row Removal:
- Dropped one column ('Unnamed: 8') that contained no data
- Removed one row with all values missing

4. Column-by-Column Cleaning Details

imdb_title_id:
- Validated IMDB ID format ('tt' followed by 7 digits)
- No missing values or outliers detected

original_title:
- Verified all titles are valid strings
- Retained legitimate numeric titles (e.g., '1917') after verification
- Fixed encoding issues such as WALLÂ·E → WALL·E and LÃ©on → Léon

release_year and release_date:
- Separated year and full release date into two columns
- Standardized release_date to YYYY-MM-DD format
- Corrected malformed dates and translated non-English month names
- Manually verified and corrected invalid dates using external references

genre:
- Normalized genre names (e.g., Sci-Fi → Science Fiction)
- Removed duplicates
- Applied One-Hot Encoding for analytical output

duration:
- Filled one missing value using the rounded mean duration

SAMPLE BY OLIVIA D.K.

country:

- Standardized country names
- Fixed typos, noise, and historical references (e.g., West Germany → Germany)

content_rating:

- Unified 'Not Rated', 'Unrated', and 'Approved' into a single category
- Final total: 26 entries labeled as 'Not Rated'

director:

- Corrected encoding artifacts and standardized name formatting

Numeric columns (income, votes, score):

- Cleaned formatting issues and typos
- Converted income and votes to integers
- Ensured score is stored as a float

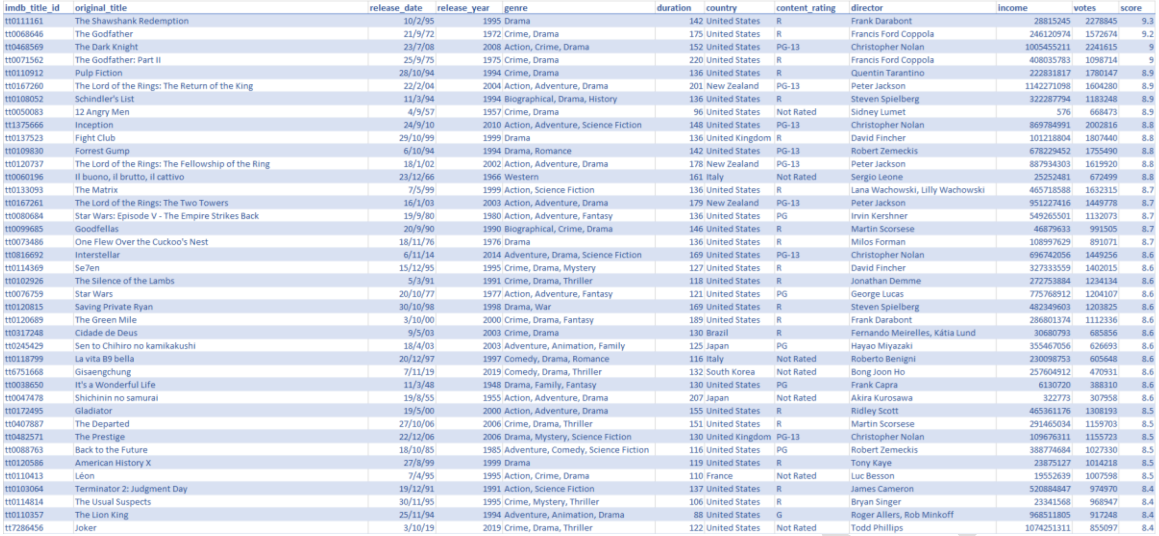
5. Final Outputs

Two cleaned datasets were produced for different use cases:

1. cleaned_for_report.csv:

- Human-readable formatting
- Accents and special characters preserved
- Country names in Title Case
- Suitable for reports and presentations

SAMPLE BY OLIVIA D.K.

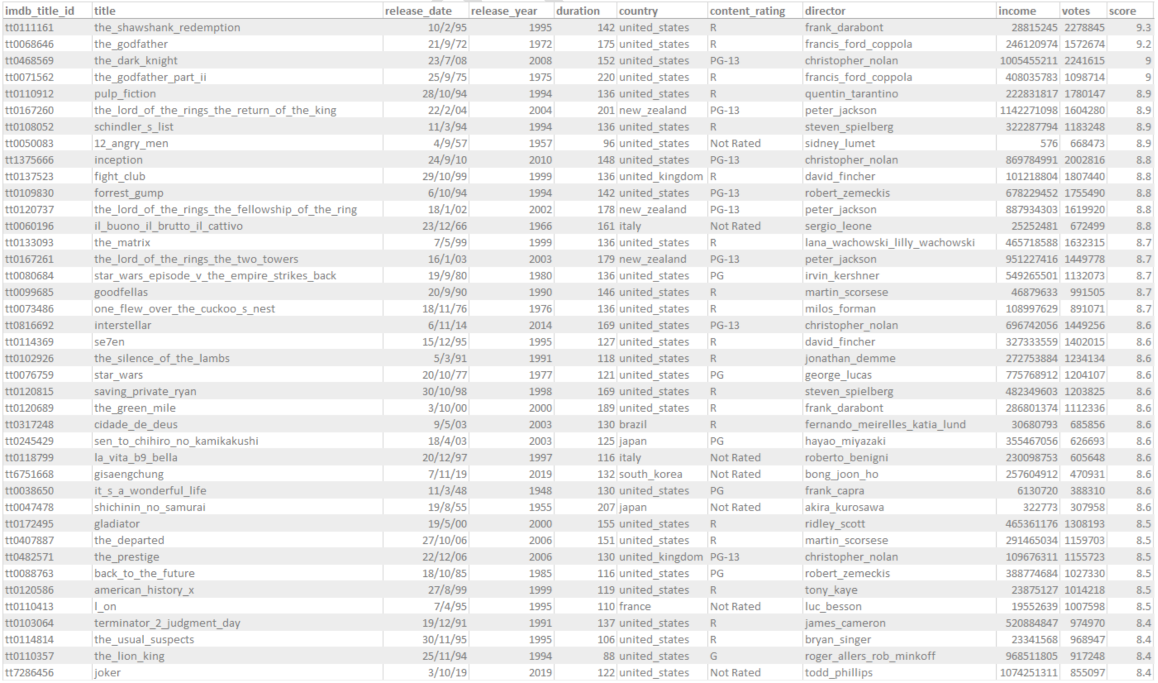


imdb_title_id	original_title	release_date	release_year	genre	duration	country	content_rating	director	income	votes	score
tt0111161	The Shawshank Redemption	10/2/95	1995	Drama	142	United States	R	Frank Darabont	28815245	2278845	9.3
tt0068646	The Godfather	21/9/72	1972	Crime, Drama	175	United States	R	Francis Ford Coppola	246120974	1572674	9.2
tt0468569	The Dark Knight	23/7/08	2008	Action, Crime, Drama	152	United States	PG-13	Christopher Nolan	1005455211	2241615	9
tt0071562	The Godfather: Part II	25/9/75	1975	Crime, Drama	220	United States	R	Francis Ford Coppola	408035783	1098714	9
tt0110912	Pulp Fiction	28/10/94	1994	Crime, Drama	136	United States	R	Quentin Tarantino	222831817	1780147	8.9
tt0167260	The Lord of the Rings: The Return of the King	22/2/04	2004	Action, Adventure, Drama	201	New Zealand	PG-13	Peter Jackson	1142271098	1604280	8.9
tt0108052	Schindler's List	11/3/94	1994	Biographical, Drama, History	136	United States	R	Steven Spielberg	322287794	1183248	8.9
tt0050083	12 Angry Men	4/9/57	1957	Crime, Drama	96	United States	Not Rated	Sidney Lumet	576	668473	8.9
tt1375666	Inception	24/9/10	2010	Action, Adventure, Science Fiction	148	United States	PG-13	Christopher Nolan	869784991	2002816	8.8
tt0137523	Fight Club	29/10/99	1999	Drama	136	United Kingdom	R	David Fincher	101218804	1807440	8.8
tt0109830	Forrest Gump	6/10/94	1994	Drama, Romance	142	United States	PG-13	Robert Zemeckis	678229452	1755490	8.8
tt0120737	The Lord of the Rings: The Fellowship of the Ring	18/1/02	2002	Action, Adventure, Drama	178	New Zealand	PG-13	Peter Jackson	887934303	1619920	8.8
tt0060196	Il buono, il brutto, il cattivo	23/12/66	1966	Western	161	Italy	Not Rated	Sergio Leone	25252481	672499	8.8
tt0133093	The Matrix	7/5/99	1999	Action, Science Fiction	136	United States	R	Lana Wachowski, Lilly Wachowski	465718588	1632315	8.7
tt0167261	The Lord of the Rings: The Two Towers	16/1/03	2003	Action, Adventure, Drama	179	New Zealand	PG-13	Peter Jackson	951227416	1449778	8.7
tt0080684	Star Wars: Episode V - The Empire Strikes Back	19/9/80	1980	Action, Adventure, Fantasy	136	United States	PG	Irvin Kershner	549265501	1132073	8.7
tt0099685	Goodfellas	20/9/90	1990	Biographical, Crime, Drama	146	United States	R	Martin Scorsese	46879633	991505	8.7
tt0073486	One Flew Over the Cuckoo's Nest	18/11/76	1976	Drama	136	United States	R	Milos Forman	108997629	891071	8.7
tt0816692	Interstellar	6/11/14	2014	Adventure, Drama, Science Fiction	169	United States	PG-13	Christopher Nolan	696742056	1449256	8.6
tt0114369	Se7en	15/12/95	1995	Crime, Drama, Mystery	127	United States	R	David Fincher	327333559	1402015	8.6
tt0102926	The Silence of the Lambs	5/3/91	1991	Crime, Drama, Thriller	118	United States	R	Jonathan Demme	272753884	1234134	8.6
tt0076759	Star Wars	20/10/77	1977	Action, Adventure, Fantasy	121	United States	PG	George Lucas	775768912	1204107	8.6
tt0120815	Saving Private Ryan	30/10/98	1998	Drama, War	169	United States	R	Steven Spielberg	482349603	1203825	8.6
tt0120689	The Green Mile	3/10/00	2000	Crime, Drama, Fantasy	189	United States	R	Frank Darabont	286801374	1112336	8.6
tt0317248	Cidade de Deus	9/5/03	2003	Crime, Drama	130	Brazil	R	Fernando Meirelles, Kátia Lund	30680793	685856	8.6
tt0245429	Sen to Chihiro no kamikakushi	18/4/03	2003	Adventure, Animation, Family	125	Japan	PG	Hayao Miyazaki	355467056	626693	8.6
tt0118799	La vita B9 bella	20/12/97	1997	Comedy, Drama, Romance	116	Italy	Not Rated	Roberto Benigni	230098753	605648	8.6
tt6751668	Gisaengchung	7/11/19	2019	Comedy, Drama, Thriller	132	South Korea	Not Rated	Bong Joon Ho	257604912	470931	8.6
tt0038650	It's a Wonderful Life	11/3/48	1948	Drama, Family, Fantasy	130	United States	PG	Frank Capra	6130720	388310	8.6
tt0047478	Shichinin no samurai	19/8/55	1955	Action, Adventure, Drama	207	Japan	Not Rated	Akira Kurosawa	322773	307958	8.6
tt0172495	Gladiator	19/5/00	2000	Action, Adventure, Drama	155	United States	R	Ridley Scott	465361176	1308193	8.5
tt0407887	The Departed	27/10/06	2006	Crime, Drama, Thriller	151	United States	R	Martin Scorsese	291465034	1159703	8.5
tt0482571	The Prestige	22/12/06	2006	Drama, Mystery, Science Fiction	130	United Kingdom	PG-13	Christopher Nolan	109676311	1155723	8.5
tt0088763	Back to the Future	18/10/85	1985	Adventure, Comedy, Science Fiction	116	United States	PG	Robert Zemeckis	388774684	1027330	8.5
tt0120586	American History X	27/8/99	1999	Drama	119	United States	R	Tony Kaye	23875127	1014218	8.5
tt0110413	Léon	7/4/95	1995	Action, Crime, Drama	110	France	Not Rated	Luc Besson	19552639	1007598	8.5
tt0103064	Terminator 2: Judgment Day	19/12/91	1991	Action, Science Fiction	137	United States	R	James Cameron	520884847	974970	8.4
tt0114814	The Usual Suspects	30/11/95	1995	Crime, Mystery, Thriller	106	United States	R	Bryan Singer	23341568	968947	8.4
tt0110357	The Lion King	25/11/94	1994	Adventure, Animation, Drama	88	United States	G	Roger Allers, Rob Minkoff	968511805	917248	8.4
tt7286456	Joker	3/10/19	2019	Crime, Drama, Thriller	122	United States	Not Rated	Todd Phillips	1074251311	855097	8.4

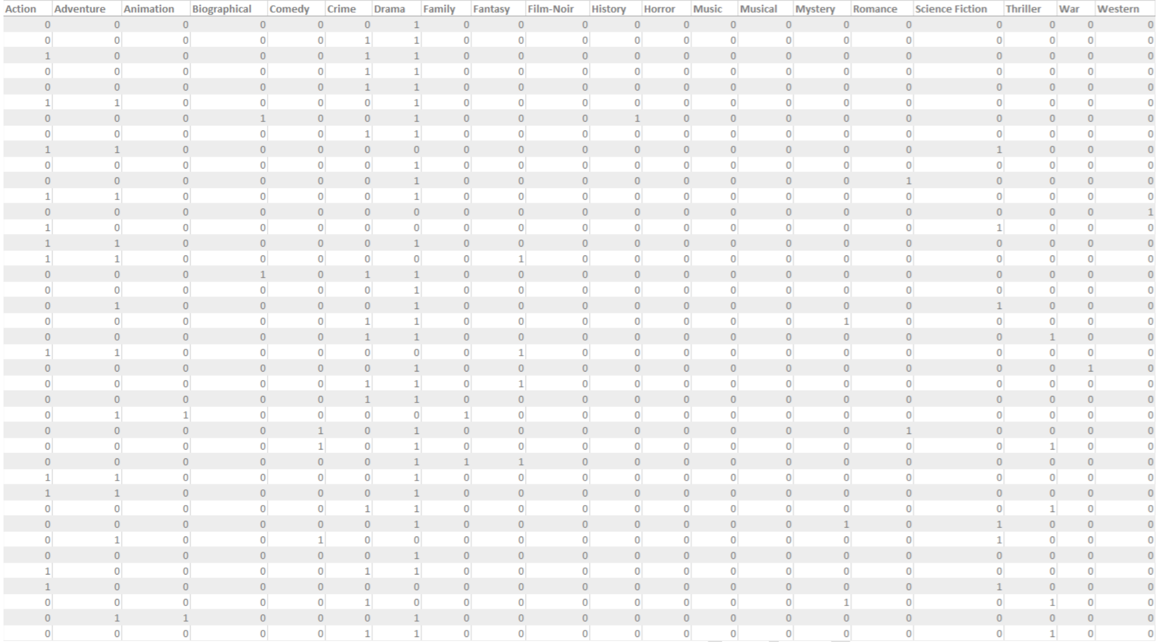
2. cleaned_for_analysis.csv:

- Fully standardized snake_case format
- No encoding artifacts
- Genre columns One-Hot Encoded
- Optimized for analysis and machine learning workflows

OLIVIA



imdb_title_id	title	release_date	release_year	duration	country	content_rating	director	income	votes	score
tt0111161	the_shawshank_redemption	10/2/95	1995	142	united_states	R	frank_darabont	28815245	2278845	9.3
tt0068646	the_godfather	21/9/72	1972	175	united_states	R	francis_ford_coppola	246120974	1572674	9.2
tt0468569	the_dark_knight	23/7/08	2008	152	united_states	PG-13	christopher_nolan	1005455211	2241615	9
tt0071562	the_godfather_part_ii	25/9/75	1975	220	united_states	R	francis_ford_coppola	408035783	1098714	9
tt0110912	pulp_fiction	28/10/94	1994	136	united_states	R	quentin_tarantino	222831817	1780147	8.9
tt0167260	the_lord_of_the_rings_the_return_of_the_king	22/2/04	2004	201	new_zealand	PG-13	peter_jackson	1142271098	1604280	8.9
tt0108052	schindler_s_list	11/3/94	1994	136	united_states	R	steven_spielberg	322287794	1183248	8.9
tt0050083	12_angry_men	4/9/57	1957	96	united_states	Not Rated	sidney_lumet	576	668473	8.9
tt1375666	inception	24/9/10	2010	148	united_states	PG-13	christopher_nolan	869784991	2002816	8.8
tt0137523	fight_club	29/10/99	1999	136	united_kingdom	R	david_fincher	101218804	1807440	8.8
tt0109830	forrest_gump	6/10/94	1994	142	united_states	PG-13	robert_zemeckis	678229452	1755490	8.8
tt0120737	the_lord_of_the_rings_the_fellowship_of_the_ring	18/1/02	2002	178	new_zealand	PG-13	peter_jackson	887934303	1619920	8.8
tt0060196	il_buono_il_brutto_il_cattivo	23/12/66	1966	161	italy	Not Rated	sergio_leone	25252481	672499	8.8
tt0133093	the_matrix	7/5/99	1999	136	united_states	R	lana_wachowski_lilly_wachowski	465718588	1632315	8.7
tt0167261	the_lord_of_the_rings_the_two_towers	16/1/03	2003	179	new_zealand	PG-13	peter_jackson	951227416	1449778	8.7
tt0080684	star_wars_episode_v_the_empire_strikes_back	19/9/80	1980	136	united_states	PG	irvin_kershner	549265501	1132073	8.7
tt0099685	goodfellas	20/9/90	1990	146	united_states	R	martin_scorsese	46879633	991505	8.7
tt0073486	one_flew_over_the_cuckoo_s_nest	18/11/76	1976	136	united_states	R	milos_forman	108997629	891071	8.7
tt0816692	interstellar	6/11/14	2014	169	united_states	PG-13	christopher_nolan	696742056	1449256	8.6
tt0114369	se7en	15/12/95	1995	127	united_states	R	david_fincher	327333559	1402015	8.6
tt0102926	the_silence_of_the_lambs	5/3/91	1991	118	united_states	R	jonathan_demme	272753884	1234134	8.6
tt0076759	star_wars	20/10/77	1977	121	united_states	PG	george_lucas	775768912	1204107	8.6
tt0120815	saving_private_ryan	30/10/98	1998	169	united_states	R	steven_spielberg	482349603	1203825	8.6
tt0120689	the_green_mile	3/10/00	2000	189	united_states	R	frank_darabont	286801374	1112336	8.6
tt0317248	cidade_de_deus	9/5/03	2003	130	brazil	R	fernando_meirelles_katia_lund	30680793	685856	8.6
tt0245429	sen_to_chihiro_no_kamikakushi	18/4/03	2003	125	japan	PG	hayao_miyazaki	355467056	626693	8.6
tt0118799	la_vita_b9_bella	20/12/97	1997	116	italy	Not Rated	roberto_benigni	230098753	605648	8.6
tt6751668	gisaengchung	7/11/19	2019	132	south_korea	Not Rated	bong_joon_ho	257604912	470931	8.6
tt0038650	it_s_a_wonderful_life	11/3/48	1948	130	united_states	PG	frank_capra	6130720	388310	8.6
tt0047478	shichinin_no_samurai	19/8/55	1955	207	japan	Not Rated	akira_kurosawa	322773	307958	8.6
tt0172495	gladiator	19/5/00	2000	155	united_states	R	ridley_scott	465361176	1308193	8.5
tt0407887	the_departed	27/10/06	2006	151	united_states	R	martin_scorsese	291465034	1159703	8.5
tt0482571	the_prestige	22/12/06	2006	130	united_kingdom	PG-13	christopher_nolan	109676311	1155723	8.5
tt0088763	back_to_the_future	18/10/85	1985	116	united_states	PG	robert_zemeckis	388774684	1027330	8.5
tt0120586	american_history_x	27/8/99	1999	119	united_states	R	tony_kaye	23875127	1014218	8.5
tt0110413	l_on	7/4/95	1995	110	france	Not Rated	luc_besson	19552639	1007598	8.5
tt0103064	terminator_2_judgment_day	19/12/91	1991	137	united_states	R	james_cameron	520884847	974970	8.4
tt0114814	the_usual_suspects	30/11/95	1995	106	united_states	R	bryan_singer	23341568	968947	8.4
tt0110357	the_lion_king	25/11/94	1994	88	united_states	G	roger_allers_rob_minkoff	968511805	917248	8.4
tt7286456	joker	3/10/19	2019	122	united_states	Not Rated	todd_phillips	1074251311	855097	8.4



Action	Adventure	Animation	Biographical	Comedy	Crime	Drama	Family	Fantasy	Film-Noir	History	Horror	Music	Musical	Mystery	Romance	Science Fiction	Thriller	War	Western
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0
0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0

6. Value Delivered

- Improved data reliability and consistency
- Reduced risk of analytical errors
- Created transparent and reproducible cleaning logic
- Delivered datasets ready for reporting, visualization, and modeling

7. Conclusion

This project demonstrates a practical, end-to-end data cleaning workflow applied to a real-world dataset. By addressing encoding issues, structural inconsistencies, and content-level noise, the final outputs provide a strong foundation for accurate analysis and downstream data science tasks.

SAMPLE BY OLIV