National University of Computer and Emerging Sciences, CFD Campus

Course Name:	**Data Science**	**Course Code:**	**CS4048**
Program:	**Computer Science**	**Semester:**	**Fall 2025**
Submission Date:	**20-10-2025**	**Total Marks:**	
Section:	**BS(CS)**	**Weight:**	
Exam Type:	**Assignment-2**		

Instructions/Note:

1. **Make single .ipynb file and use this format to name it 23F-1234.pynb.**
2. **Do not submit your assignment after the deadline.**
3. **Do not copy code from any source otherwise you will be penalized with negative marks.**
4. **Datasets (pakistan-media-dataset-synthetic.csv) is provided with this Assignment.**
5. **There Will be Viva for this Assignment.**

Dataset Overview: This dataset represents synthetic media data from Pakistani news outlets, combining television, newspaper, and social media attributes. It simulates journalist activity, coverage topics, sentiment, bias, and audience engagement metrics — useful for analyzing media trends, bias detection, or content influence.

Column Name	**Data Type**	**Description / Meaning**	**Example Value**
ID	Integer	Unique identifier for each record or media entry.	0
Journalist	String	Name of the journalist or anchorperson associated with the story.	Najam Sethi
Channel	String	Name of the TV news channel broadcasting the story.	ARY News
Newspaper	String	Name of the newspaper publishing the story.	The News
Region	String	Province or major region of Pakistan where the report originates.	Sindh, Islamabad
City	String	Specific city associated with the news report.	Multan, Quetta
Topic	String	Main category or subject of the news piece.	Politics, Sports, Health
Headline	String	The headline or summary of the article or broadcast.	"Polio cases reported in KPK"
Ratings	Float	Viewership or popularity rating of the report (numeric score).	27.54
Revenue	String	Estimated monetary revenue (may include text like “million”, “crore”).	5 million, 14072656

Column Name	Data Type	Description / Meaning	Example Value
Airtime	Float	Broadcast time or duration in seconds/minutes.	45.6
SentimentScore	Float	Sentiment analysis score (negative–positive scale).	-0.45 to +0.85
BiasScore	Float	Degree of bias detected in the story (0 = neutral, 10 = highly biased).	5.0
Viewership	Float	Number of viewers who watched the story or broadcast.	1,756,573
Shares	Float	Number of times the content was shared on social platforms.	136,184
AdSpend	String	Advertising spend associated with the report (numeric or textual).	50 lakh, 4,452,987.876
ControversyFlag	String	Indicates whether the content is controversial.	Yes, No, 1
MissingDataFlag	Float	Marks missing or incomplete records (1 = missing, 0 = complete).	1.0, NaN
Date	String	Publication or broadcast date (YYYY-MM-DD).	2024-05-07
Language	String	Language in which the content was published.	English, Urdu
PoliticalAffiliation	String	Political leaning of the journalist or outlet.	Pro-Govt, Opposition, Neutral
SocialMediaInteractions	Float	Total number of engagements (likes, comments, shares, etc.) on social media.	33,436

The Media Accountability Network (MAN), a non-profit organization dedicated to promoting transparency in journalism, has launched an extensive data-driven investigation into how news narratives are shaped in Pakistan’s media ecosystem. Over the past decade, questions of media independence and political interference have become increasingly pressing, as evidence suggests that certain television channels, newspapers, and digital news portals may be offering biased or selectively filtered information to the public. MAN’s mission is to examine how topics are covered, who covers them, and whether the distribution of airtime, sentiment, and advertising revenue reflects objective reporting or deliberate influence.

The organization has compiled a comprehensive dataset integrating information from multiple media outlets. Each record in the dataset includes details such as the news channel or newspaper, the journalist responsible, the topic or category of coverage, sentiment indicators, TRP ratings, airtime allocation, and the corresponding advertising revenue. However, early inspection of this dataset has revealed substantial inconsistencies, suggesting that the data itself may mirror the very biases and irregularities MAN seeks to expose. Channel names, journalist identities, and city labels are found in duplicate or with inconsistent spellings, complicating any direct comparison. Many headlines appear misclassified by topic—for example, incidents of terrorism have been wrongly tagged under entertainment, and political stories have been mislabeled as sports. In several cases, journalists are shown covering beats entirely outside their professional domain, hinting either at editorial control, administrative manipulation, or poor data entry practices.

Further examination reveals severe numerical inconsistencies: airtime values are occasionally negative, TRP ratings exceed the theoretical maximum of 100, and advertisement revenues are recorded using mixed monetary units such as PKR, lakh, or crore. The bias scores—intended to represent political leaning—sometimes surpass the logical range of -1 to +1, while the provided “Controversy” and “Missing Data” flags often do not match observable facts. Interestingly, some pro-government channels appear to receive unusually high advertising revenues during politically sensitive periods, while their coverage of corruption or scandal diminishes, suggesting a potential exchange between favorable coverage and financial incentives.

As a data science investigator working with the Media Accountability Network, you are tasked with conducting a full-scale forensic analysis of this dataset to determine the extent to which systematic bias, editorial influence, and data manipulation can be identified and verified. Your first responsibility is to examine the dataset’s structural integrity, identifying the major sources of inconsistency and noise. This requires designing strategies for standardizing categorical values such as channel and journalist names, correcting or flagging impossible numerical entries, and reconciling monetary values into a single consistent unit. Missing values must be handled carefully, with clear justification for the imputation techniques used, and potential distortions caused by outliers must be critically evaluated before any modeling or visualization is attempted.

Once a reliable foundation has been established, your investigation should shift from data cleaning to interpretation. MAN’s central hypothesis is that financial incentives may correlate with ideological bias. Your role is to explore whether channels that show pro-government sentiment systematically receive higher advertising revenues, whether independent or opposition-leaning journalists experience restricted airtime, and which topics dominate national coverage at the expense of others. For example, are sensitive political or corruption stories being underreported compared to entertainment or sports segments? Does the pattern of coverage change during election seasons or major political events? Similarly, you are expected to examine the autonomy of journalists—whether they consistently report on their assigned beats or are redirected toward topics that align with institutional interests. Each of these lines of inquiry must be supported by quantitative analysis and appropriate visualization, but the emphasis should remain on the interpretation of evidence rather than the mere production of charts.

As part of this case study, you are expected to perform a complete exploratory and statistical analysis of the dataset. Begin by conducting a thorough data cleaning and preprocessing phase, where you identify and correct inconsistent categories, resolve unit mismatches, handle missing values through justified imputation, and detect or treat outliers. Once the data is reliable, proceed to exploratory data analysis (EDA) — compute descriptive statistics (mean, median, variance etc.) and visualize the data distributions of numerical features such as airtime, TRP ratings, bias scores, and advertising revenue. Use bar plots, box plots, histograms, heatmaps, and time-series visualizations to explore coverage patterns, advertising trends, and potential bias indicators. Move beyond surface-level visualization by performing comparative analysis to determine whether pro-government and independent media outlets differ significantly in TRP, revenue, or sentiment. Each visualization or analysis must be accompanied by a clear interpretation explaining what it reveals about media bias, influence, or data integrity. Your submission should include the cleaned dataset, your reproducible code (e.g., a Jupyter notebook), and a concise analytical report summarizing your key findings, and limitations.

Throughout the investigation, a critical question must guide your analysis: can the dataset itself be trusted? The Media Accountability Network recognizes that apparent patterns of bias may be genuine reflections of manipulation, but they might also be artifacts of data corruption or incomplete reporting. Your final report should therefore not only highlight findings of bias or influence but also discuss the limitations of these conclusions. You are expected to reflect on the reliability of the provided metrics, the validity of the controversy and bias flags, and the extent to which human or algorithmic manipulation might still be hidden beneath the data's surface. The investigation concludes with a formal watchdog report summarizing your analytical process, major insights, and ethical reflections.

In this report, you should clearly articulate which channels appear systematically biased, what evidence suggests a link between advertising revenue and favorable coverage, and which topics or journalists seem misrepresented. You should also acknowledge the uncertainties and data limitations that remain, explaining how inconsistencies or missing information affect the confidence of your findings. Finally, propose recommendations for improved future data collection—such as rigorous metadata standards, automated flag validation, or transparent bias scoring systems—to strengthen the reliability of future investigations.

This case study challenges you to move beyond technical proficiency and engage in critical, reflective inquiry. The goal is not to produce perfect models but to demonstrate the reasoning, skepticism, and ethical awareness required when working with real-world data that may itself be a product of manipulation. By the end of the project, you should be able to explain not only what the data shows, but also how and why you chose to interpret it the way you did—and what that reveals about truth, bias, and accountability in modern journalism.

This case study contains multiple layers of information — some visible, others deliberately subtle or hidden within the data. As a data scientist, your responsibility is not just to perform standard cleaning or analysis, but to think critically, question patterns, and uncover what lies beneath the surface. Every irregularity or inconsistency may point toward deeper insights about bias, manipulation, or data integrity. Treat this dataset as a real-world forensic investigation rather than a classroom exercise: explore it creatively, validate your assumptions, and build evidence for every conclusion you draw. The quality of your analysis will be judged not only by technical accuracy, but also by your ability to detect, interpret, and explain the unseen dynamics shaping the data.