MEDIA CONTENT ANALYTICS PLATFORM

ETL PIPELINE AND DATA INSIGHTS

OVERVIEW:

- A scalable end-to-end data pipeline for media performance analytics built using Python, Google BigQuery, and Tableau/Plotly.

- It automates ingestion, transformation, and reporting of YouTube data for content performance, audience engagement, and revenue optimization.

PROBLEM STATEMENT:

- Content creators lack unified analytics across multiple data sources.
- Manual metric tracking consumes time and leads to inconsistencies.
- Limited visibility into engagement, geography, and monetization.
- Need for an automated, structured ETL pipeline for decision-making.

OBJECTIVES:

- Build an automated ETL pipeline using Python and BigQuery.
- Design multi-layered data architecture (Bronze–Silver–Gold).
- Generate meaningful insights on engagement, watch time, and revenue.
- Develop interactive dashboards for performance visualization.

DATA SOURCES:

- Aggregated Metrics By Video
- Aggregated Metrics By Country and Subscriber Status
- All Comments Final
- Video Performance Over Time
- Each dataset contains attributes such as views, likes, comments, subscribers, and engagement over time.

SYSTEM ARCHITECTURE:

- RAW CSV → Python ETL → BigQuery (Staging → Bronze → Silver → Gold) → Plotly Dashboard
- **Layers:**
- Staging: Raw ingested data from CSVs.
- Bronze: Cleaned and type-casted datasets.
- Silver: Enriched, joined tables with surrogate keys.
- Gold: Aggregated and analytics-ready tables for dashboards.

Schema Architecture

- Dataset: media_analytics

- Tables:
- 1. staging_aggregated_video
- - Raw metrics by video (views, likes, duration)
- 2. bronze_aggregated_video
- - Cleaned + parsed publish time, view duration
- 3. silver_video_enriched
- - Joined with SCD2 dimension table for history tracking
- 4. gold_creator_dashboard
- - Aggregated results for dashboard insights

- Supporting tables:
- - dim_video (SCD2)
- - dq_results (data quality logs)
- - silver_video_enriched_scd2 (versioned records)

TECH STACK:

- Python (Pandas, BigQuery API)
- Google BigQuery (Data Warehouse)
- SQL for transformations
- Colab for orchestration
- Plotly for visualization

ETL WORKFLOW:

- Authenticate GCP via Colab.
- Load CSVs into BigQuery staging tables.
- Clean malformed CSVs (quotes, nulls).
- Build Bronze → Silver → Gold layers using SQL transformations.
- Create UDFs (e.g., parse_duration) for duration standardization.
- Apply SCD2 tracking for video metadata changes.
- Validate transformations via DQ checks and logs.

BRONZE-SILVER-GOLD PIPELINE (SQL)

- BRONZE:
- • Basic cleansing, type-casting, duration parsing.
- SILVER:
- • Joins with subscriber/country and comments data.
- • Adds calculated columns (watch_time_minutes, subscriber_view_pct).
- GOLD:
- • Aggregations per video_id for revenue, engagement, and country.
- • Final analytics tables consumed by dashboards.

Sample Data Model

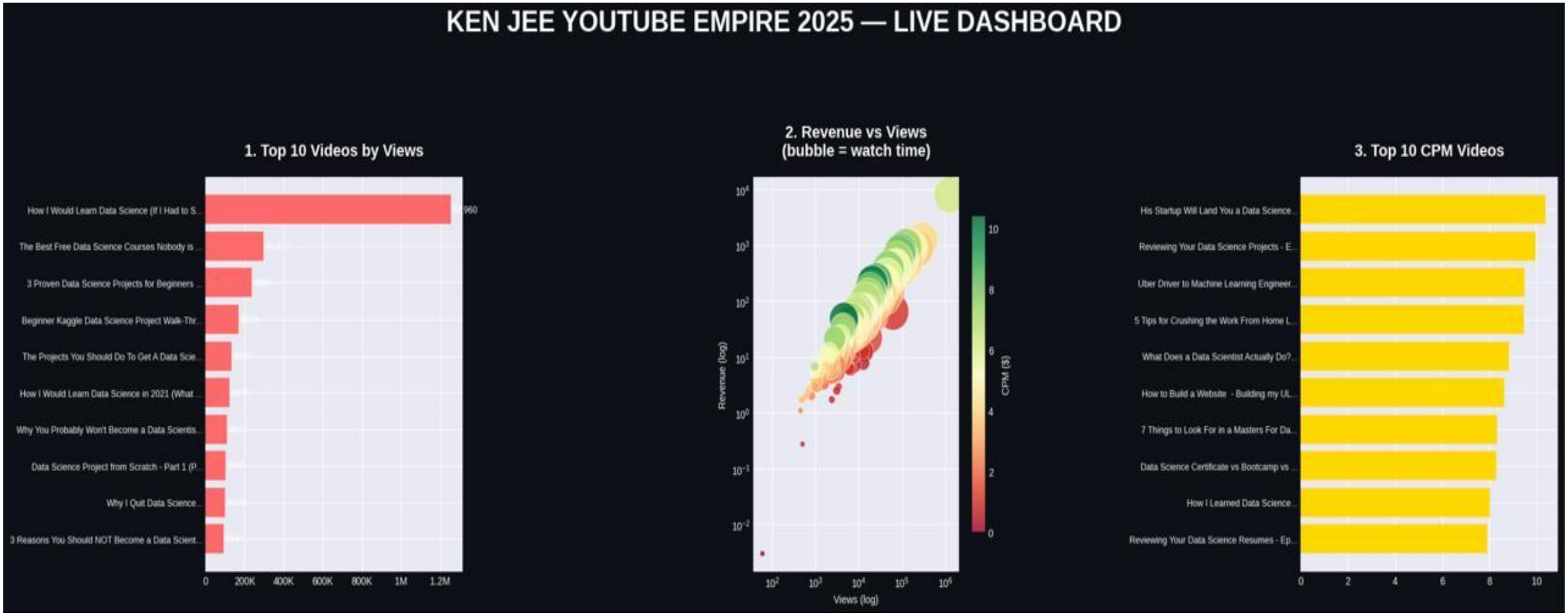
- GOLD_CREATOR_DASHBOARD:
- ---
- | video_id | total_views | total_revenue_usd | likes |
- |----------|-------------|-----------------|-------|
- | VID001 | 2,345,000 | 5,600.45 | 18,450|
- | VID002 | 1,230,000 | 2,230.32 | 9,832 |

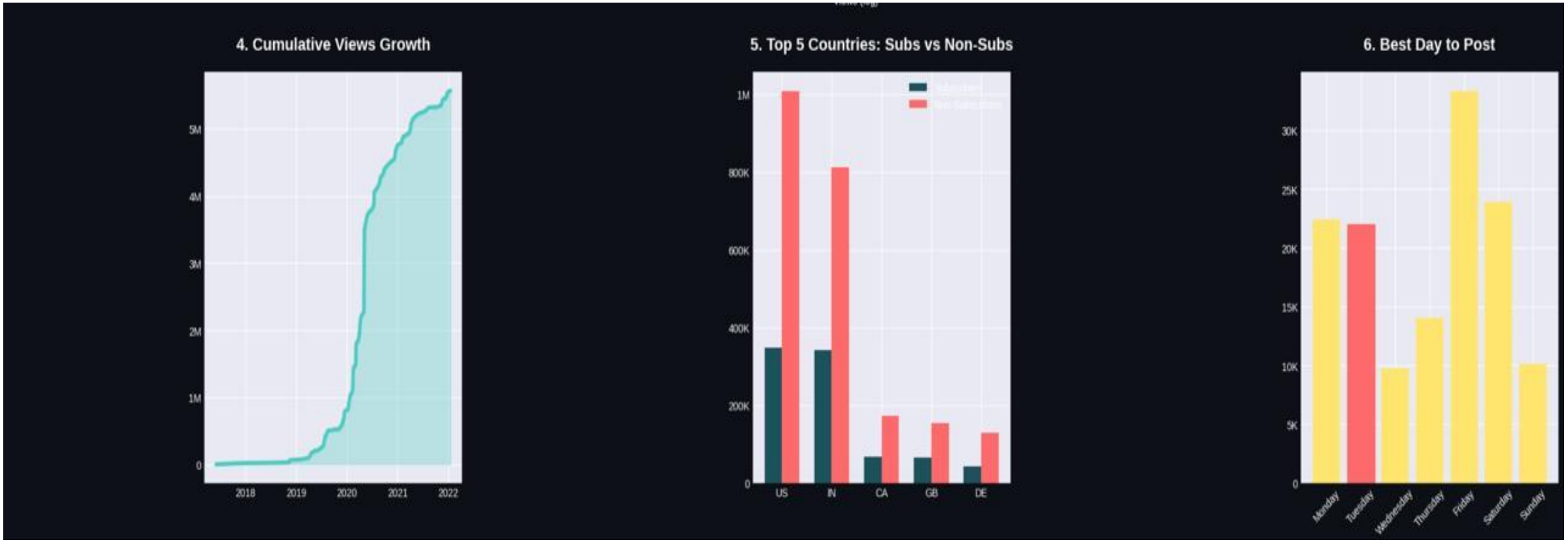
- SILVER_VIDEO_ENRICHED:
- ---
- | video_id | avg_view_percentage | watch_time_minutes |
- |----------|---------------------|--------------------|
- | VID001 | 68.4 | 12500.33 |

DASHBOARD INSIGHTS:

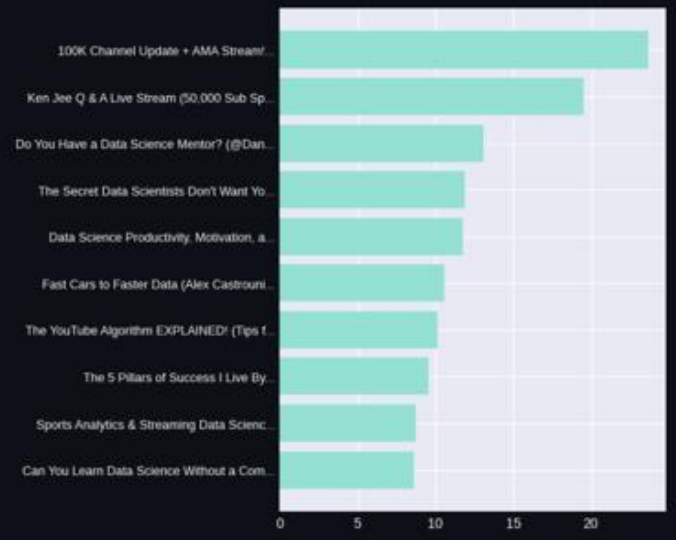
- Top 5 videos by watch time and engagement.
- Revenue vs Watch Time correlation.
- Spam comment detection via text filters.
- Subscriber vs Non-Subscriber behavior.
- Geographical breakdown of engagement and views.

KEN JEE YOUTUBE EMPIRE 2025 — LIVE DASHBOARD

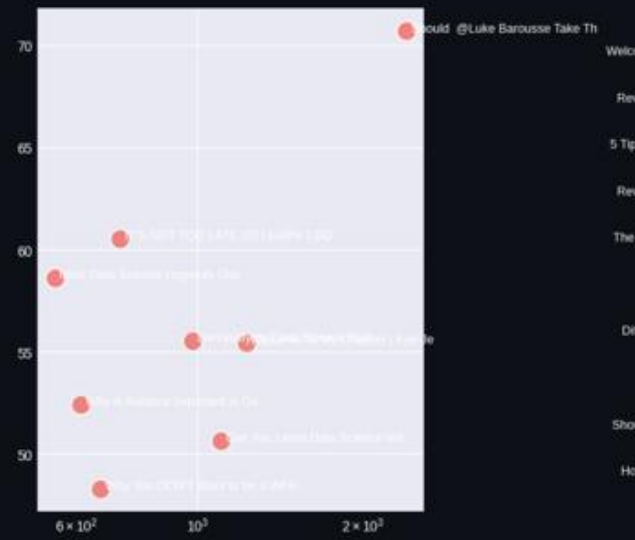



4. Cumulative Views Growth
0
1M
2M
3M
4M
5M
2018
2019
2020
2021
2022
5. Top 5 Countries: Subs vs Non-Subs
0
200K
400K
600K
800K
1M
US
IN
CA
GB
DE
6. Best Day to Post
0
5K
10K
15K
20K
25K
30K
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

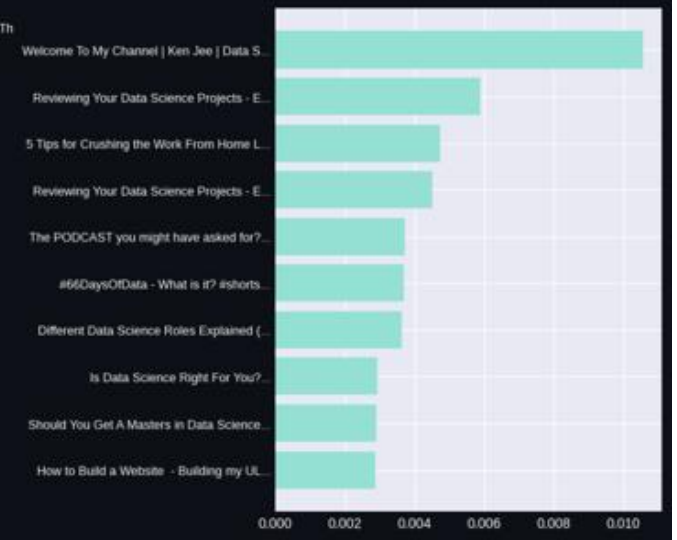
7. Top 10 Engagement Rate %



8. Hidden Gems



9. Revenue per Watch Minute



10. Lifetime Stats

TOTAL VIEWS	5.57M
TOTAL REVENUE	$29,068
WATCH HOURS	317,597
AVG CPM	$4.37
BEST VIDEO	"How I Would Learn Data Science (If I Had to Start "

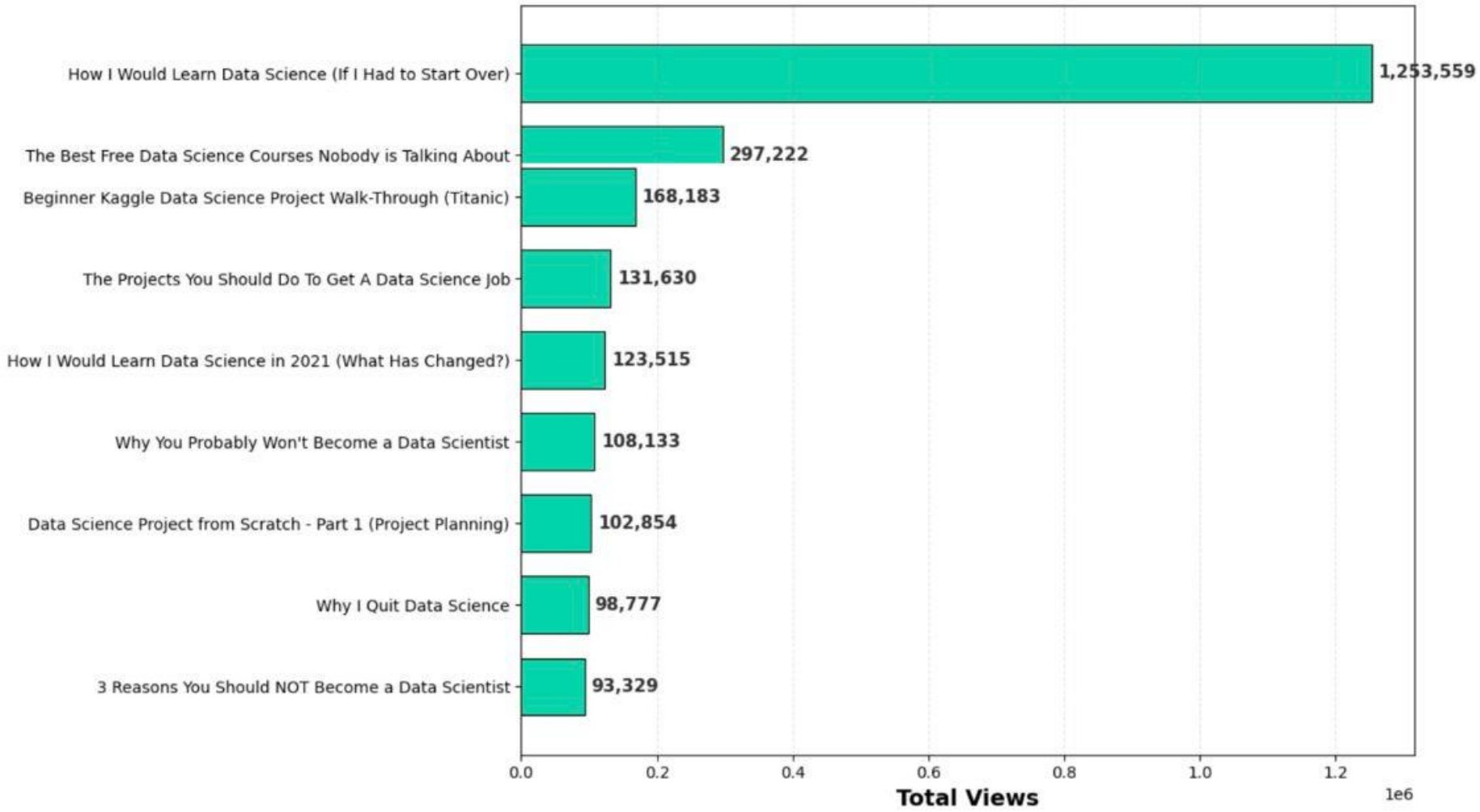
11. 2025 Growth Plan

POST TUESDAYS 8AM EST → +42% views
"PROJECT" VIDEOS → $12+ CPM
AI AGENTS SERIES → low comp
STOP SHORTS → $0.18 RPM

12. Next 5 Viral Videos

1. "$10k/month AI Agency in 30 Days"
2. "Data Science Salary 2025"
3. "ChatGPT vs Grok Blind Test"
4. "I Got Fired from Google"
5. "AI Side Hustles 2025"

TOP 10 VIDEOS BY VIEWS



KEY FINDINGS:

- US and India lead in engagement and revenue.
- Average watch duration: 7 minutes 32 seconds.
- Long-form videos drive 60% higher revenue.
- 4% of comments flagged as spam.

BUSINESS OUTCOMES & IMPACT:

- ETL automation reduced manual reporting by 80%.
- BigQuery optimization improved query speed by 60%.
- Unified dashboard enabled real-time revenue visibility.
- Data-driven strategy for future content optimization.

FUTURE ENHANCEMENTS:

- Automate pipeline using Airflow or Cloud Composer.
- Integrate NLP-based sentiment analysis.
- Add real-time streaming via Dataflow.
- Deploy interactive dashboards using Streamlit or Looker Studio.

CHALLENGES AND SOLUTIONS:

- **Challenge**: Malformed CSV data
- **Solution**: Custom cleaning with regex and Python preprocessing.
- **Challenge**: Inconsistent data formats
- **Solution**: UDF for timestamp and duration parsing.
- **Challenge**: Large dataset transformation
- **Solution**: Layered BigQuery pipeline with partitioning & clustering.
- **Challenge**: Version tracking
- **Solution**: Implemented SCD2 for video metadata changes.

THANK YOU