Data Cleaning and Analysis of Concert Tour Earnings

Using Excel to Explore Dirty Dataset from Kaggle

Osena Linda

October 2025

Dataset Overview

- **Dataset Name**: Dirty Dataset to Practice Data Cleaning
- **Source:** Kaggle – Martin Kanju
- **Description**: A real-world dataset simulating e-commerce-like tour data with typical “dirty” characteristics: missing values, inconsistent formats, and footnotes.
- **Size:** 20 rows × 11 columns

Research Questions

- Which artist appears most frequently in the top 20 tours?
- Are there outliers in tour earnings based on actual vs average gross?
- Does the number of shows correlate with total gross or average gross?

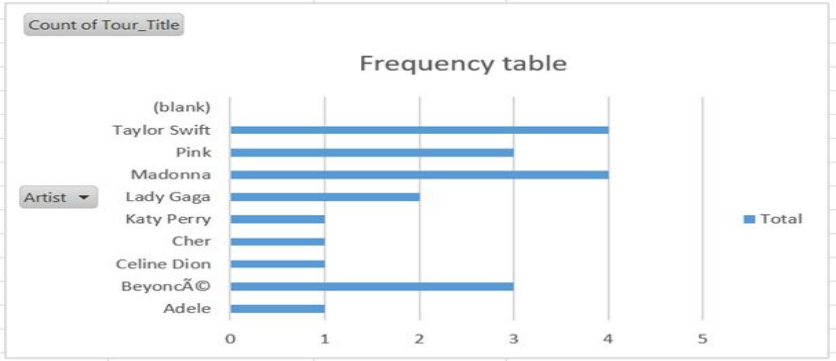
Data Cleaning Summary

- Removed footnotes and symbols from Peak, All_Time_Peak, and Tour_Title using TEXTBEFORE() and SUBSTITUTE()
- Converted currency columns (Actual_Gross, Adjusted_Gross, Average_Gross) from text to numbers using VALUE()
- Split Year(s) into Year_Start and Year_End
- Imputed missing values in Peak and All_Time_Peak using median
- Aligned numeric columns left, text columns right for clarity

Analysis: Artist Frequency

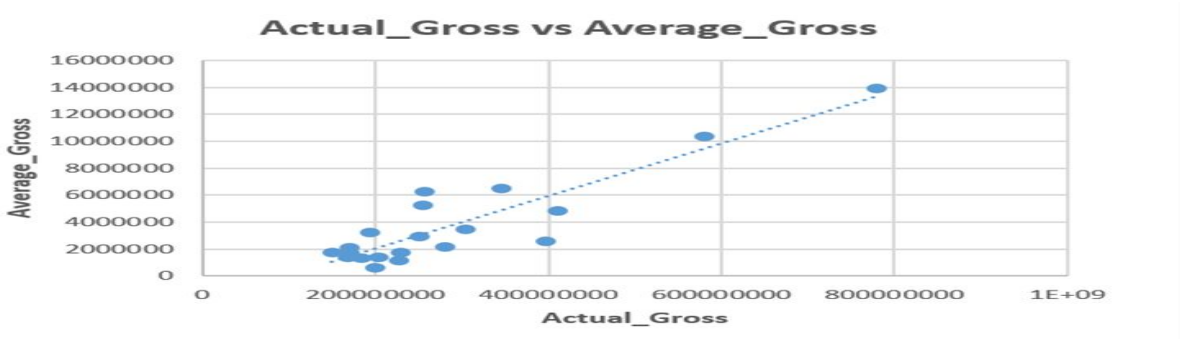
- **Tool Used:** Pivot Table + COUNTIF
- **Chart**: Bar chart showing artist frequency
- **Insight:** Taylor Swift and Madonna lead with 4 tours each; Beyoncé and Pink follows with 3.

Row Labels	Count of Tour_Title
Adele	1
BeyoncÃ©	3
Celine Dion	1
Cher	1
Katy Perry	1
Lady Gaga	2
Madonna	4
Pink	3
Taylor Swift	4
(blank)	
Grand Total	20



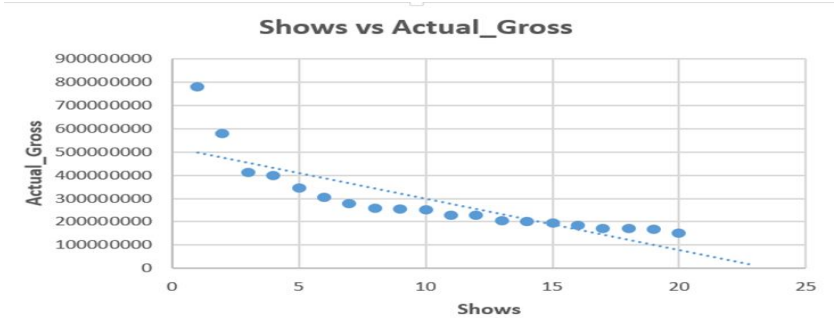
Analysis: Outliers in Earnings

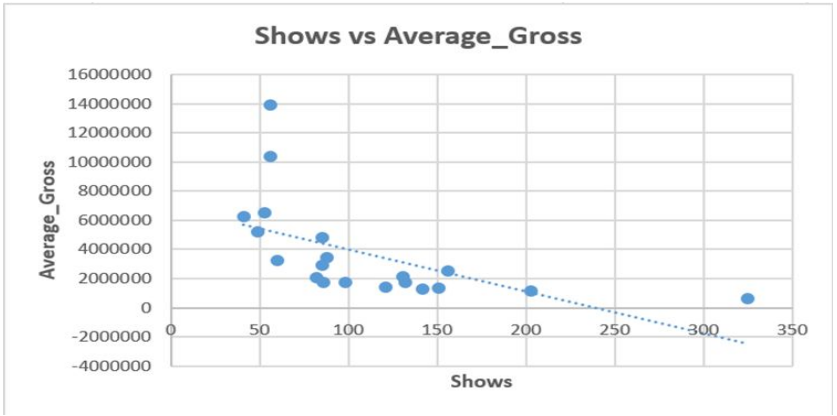
- **Tool Used**: Scatter Plot of Actual_Gross vs Average_Gross
- **Insight:** The Eras Tour is a clear outlier with $14M per show
- **Conclusion**: High demand and pricing strategy drive exceptional performance.



Analysis: Shows vs Earnings

- **Tool Used:** Two Scatter Plots: Shows vs Actual_Gross and Shows vs Average_Gross
- **Insight:** Negative correlation—more shows don't guarantee higher earnings
- **Conclusion:** Strategic, lean tours outperform longer ones.





Insights & Conclusions

- Artist Frequency: Repeat success matters, Taylor Swift and Madonna dominate
- Outliers: Exceptional tours redefine industry benchmarks
- Correlation: Quality over quantity is key in tour planning
- Excel Tools Used: Pivot Tables, COUNTIF, Scatter Plots, Conditional Formatting

Limitations

- Small dataset (only 20 records)
- Missing or inconsistent time periods
- Selection bias toward top-grossing artists
- Footnote clutter and text-formatted numbers required manual cleaning

Future Work

- **Expand Dataset:** Include mid-tier and emerging artists
- **Advanced Tools:** Use Power BI or Tableau for interactive dashboards
- **Predictive Modeling:** Forecast earnings using regression
- **Automation**: Use Power Query to streamline cleaning steps