Formula 1 Race Position Performance- Stock Price Analysis

Milestone 1: Data Collection, Preprocessing, and Exploratory Data Analysis

Project Objective:

To develop an interactive dashboard that analyzes Formula 1 race performance metrics and their potential correlation with team market performance. The project aims to uncover patterns in race performance metrics and lay the groundwork for future analysis of how these metrics might influence team company stock price.

Technical Stack:

- **Programming Language**: Python 3.x
- **Key Libraries**:
  - Data Manipulation: pandas, numpy
  - Visualization: matplotlib, seaborn
  - Statistical Analysis: scipy
  - Machine Learning (future): scikit-learn
  - Dashboard Development (future): Streamlit
- **Development Environment**:
  - IDE: Jupyter Notebook
  - Version Control: Git/GitHub
  - Documentation: Markdown

- **ML Models :**
  - Logistic Regression, Random Forest Classifier, Gradient Boosting, Voting Classifier (ensemble)

Dataset Description

The analysis uses the Formula 1 World Championship dataset, which includes:

- Races data (1950-2024)
- Driver information
- Lap times
- Pit stops
- Constructor details
- Each Team's Stock Price in the race period. Historic Stock Data is collected from Yahoo Finance using Python Module "**YFinance**" .
- Data set link can be found [here](#).

Data Preprocessing Steps

1. **Missing Data Handling**
   - Numerical values: Median imputation

- o Categorical values: Mode imputation
2. **Outlier Detection and Treatment**
   - o Z-score method for lap times
   - o IQR method for pit stop durations
3. **Feature Engineering**
   - o Created derived metrics for performance analysis
   - o Standardized time-based features

## Key Insights from EDA

1. **Lap Time Analysis**
   - o Notable improvement in lap times from 2000-2004, reaching fastest period (~87-88 seconds)
   - o Significant slowdown from 2008-2010, peaking around 108 seconds in 2010
   - o Gradual improvement post-2010 with some fluctuations
   - o Recent years (2020-2024) show relatively stable lap times around 95-98 seconds
   - o The variations likely correspond to major regulation changes and technological developments in F1

Average Lap Times by Year
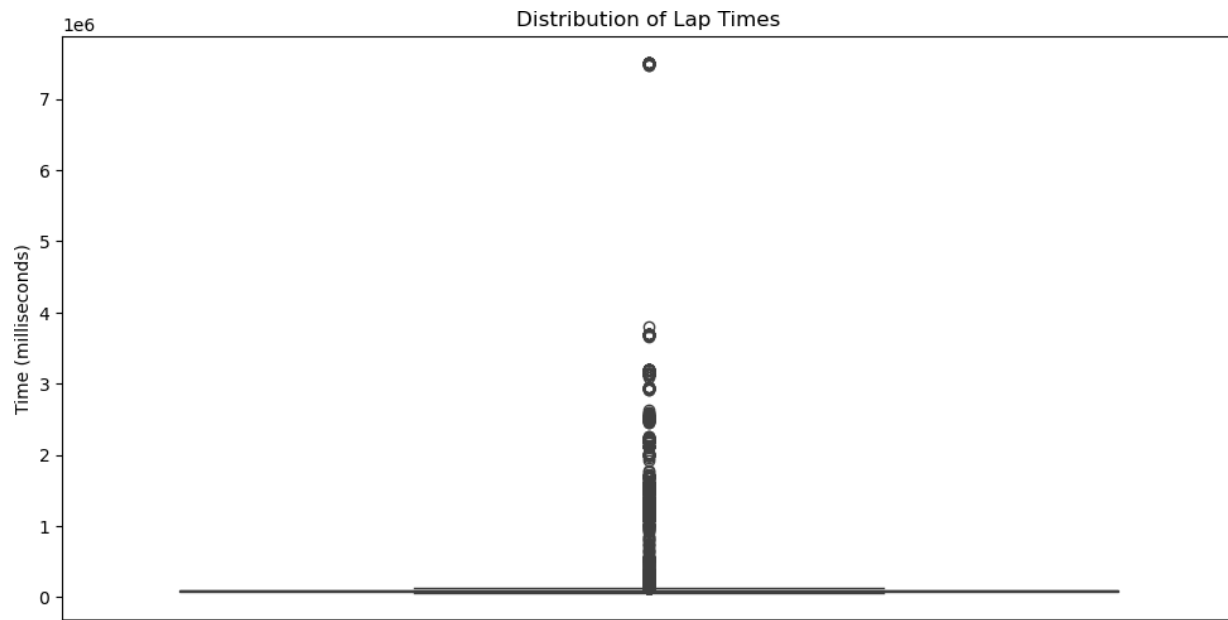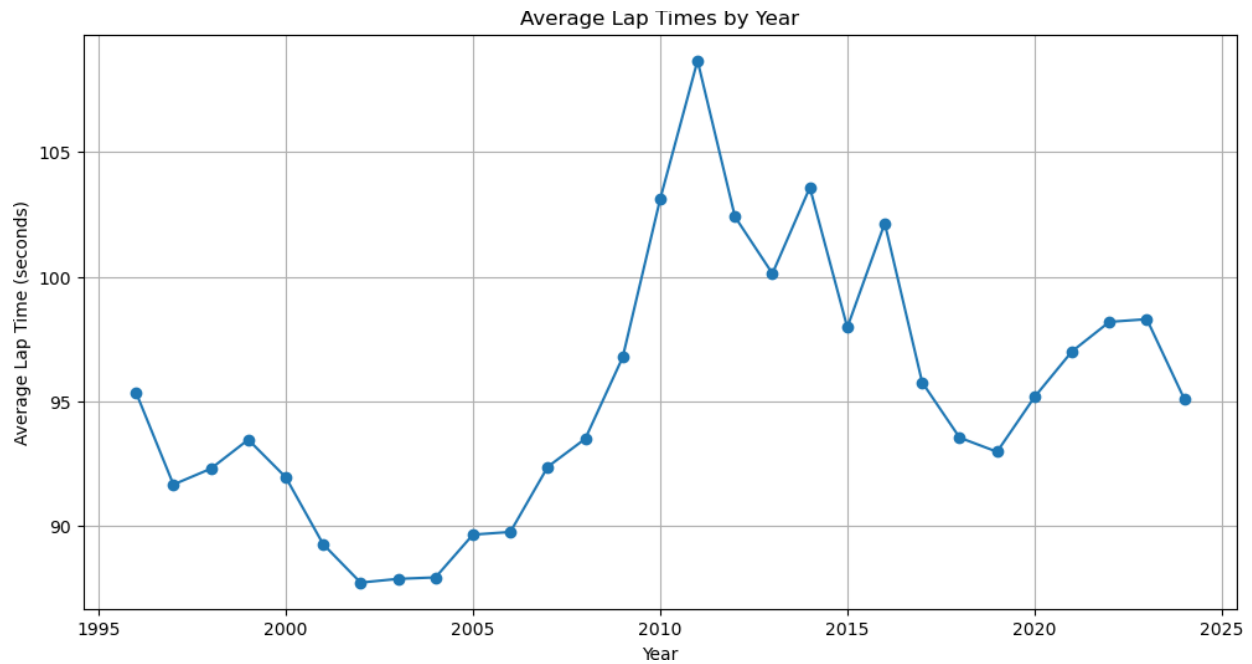
2. **Position Analysis**
   - Clear negative correlation between average position and position variation
   - Drivers with better average positions (lower numbers, 4-8) show higher position variation (3-7 std), indicating more dynamic racing and overtaking at the front of the grid
   - Back-markers (positions 16-20) show much lower position variation (1-3 std), suggesting they tend to remain in similar positions
   - Middle-field positions (10-14) show moderate variation, indicating competitive mid-pack racing
   - The scatter pattern suggests three distinct racing tiers: front-runners, mid-field, and back-markers.


Position Consistency Analysis
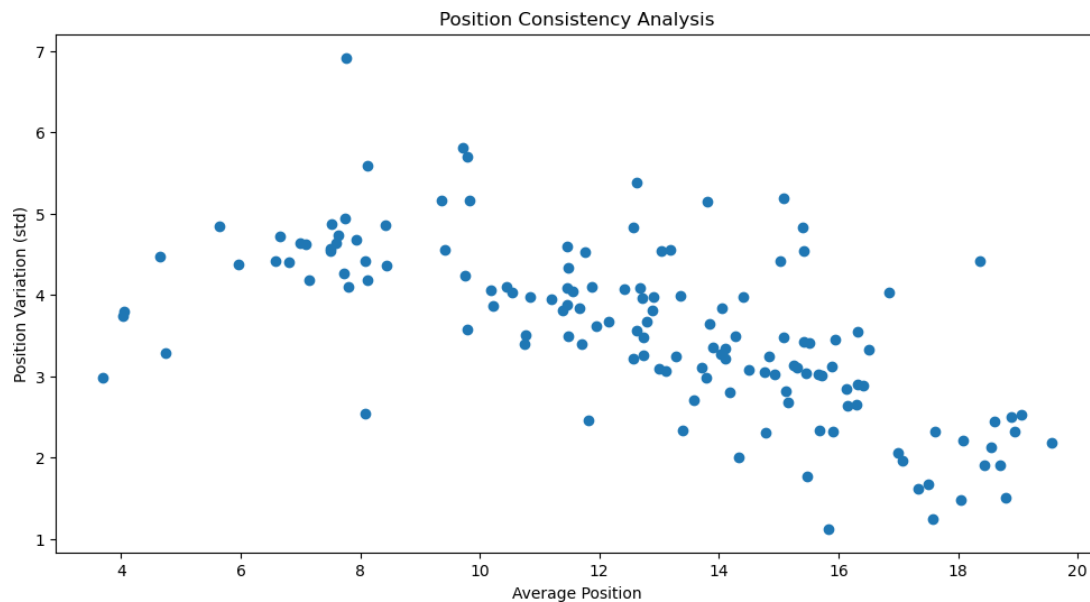
3. **Driver Performance Patterns**
   - British drivers dominate Formula 1 with approximately 160 drivers, followed closely by American drivers
   - There's a significant drop after the top 2 nationalities, with Italian drivers being the third most common
   - European nations (British, French, German, Belgian, Swiss) make up the majority of the top 10
   - South African representation shows F1's global reach, though at a much lower count
   - The distribution suggests a historical European-centric nature of the sport, particularly British dominance.

Top 10 Driver Nationalities
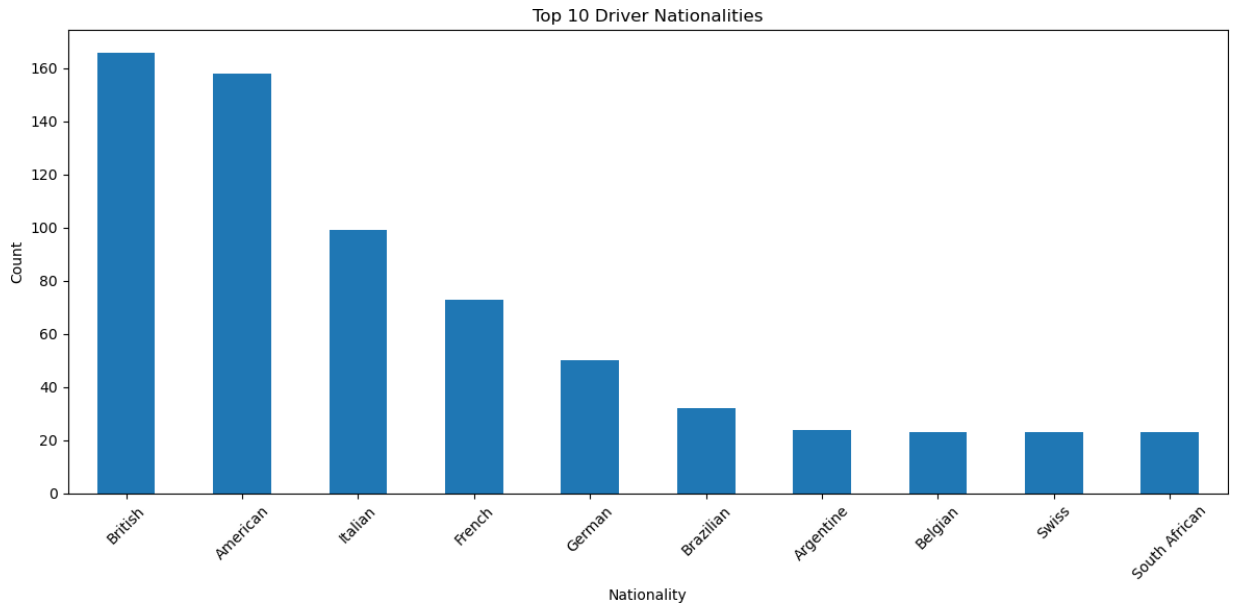
4. **Constructor Performance**
   - Ferrari leads significantly with over 10,000 total points
   - Mercedes and Red Bull form a clear second tier with ~7,500 points each
   - McLaren stands alone in a third tier with ~7,000 points
   - Significant drop-off after the top 4 teams
   - Historical teams like Williams and Renault show decent point accumulation
   - Newer or smaller teams (Force India, Team Lotus, Benetton, Tyrrell) have significantly fewer points
   - The distribution suggests a clear hierarchy in F1 team success and resources

Top 10 Constructors by Total Points
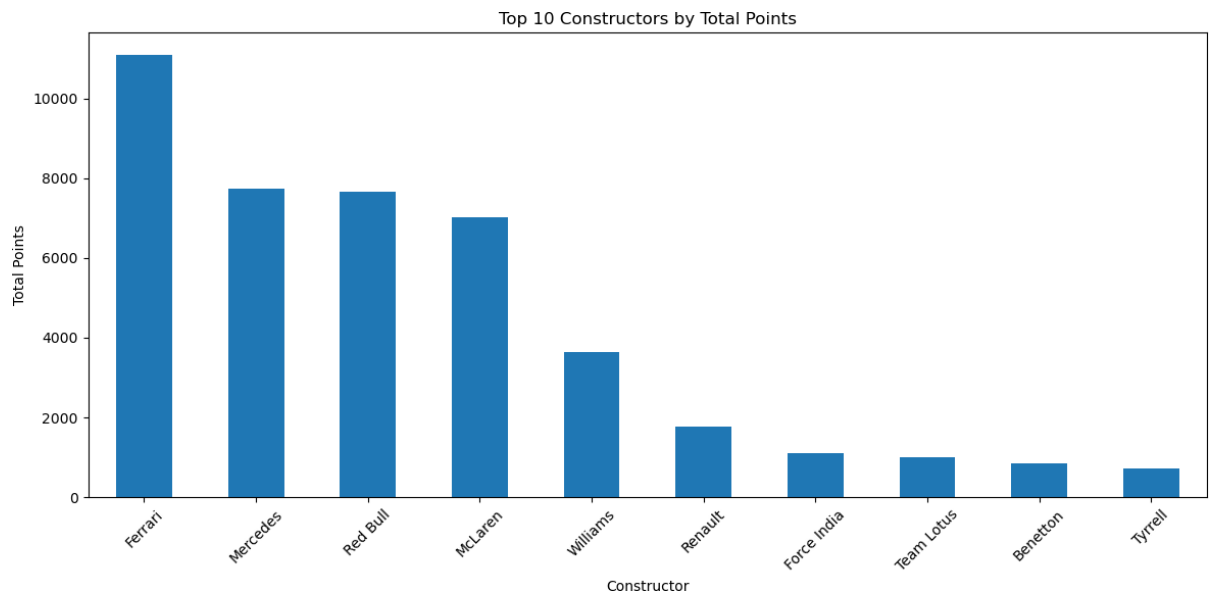
Milestone 2: Feature Engineering, Feature Selection, and Data Modeling

TOTAL DATA SIZE:
- 1125 F1 races from 1950–2025
- Focus Years: 2015–2023 (post Ferrari IPO)
- Constructors: 212 total
- Ferrari Races Analyzed: 185 (2015–2023)
- **Stock Data Points**: 2062 trading days

Feature Engineering:
Created Features:
- position: final race position
- points: points earned per race
- round: race number in the season
- price_change_pct: % change in stock price (5-day post-race vs 5-day pre-race)
- price_up: target variable (1 if stock rose, else 0)

Filtering Applied: Removed outliers: stock changes > ±10%
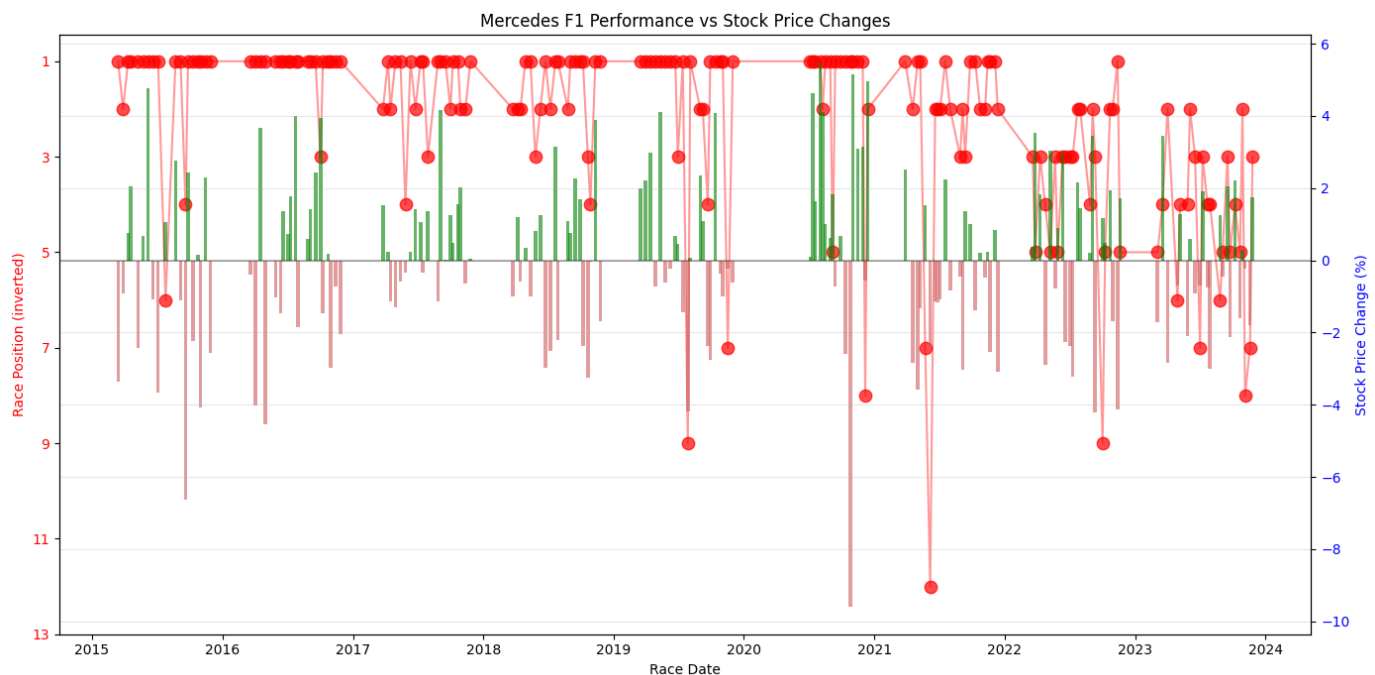
Feature Selection :

Final Features Used:
- position
- points
- round

Reasoning:
- These features have direct logical and statistical relevance to race outcomes.
- Simpler models performed better with fewer, cleaner inputs.

Key Insights taken into consideration:

Mercedes F1 Performance vs Stock Price Changes

Understanding from the Graph:

From the above graph we can detect that from **2015 to 2021**, Mercedes consistently maintained top race positions, often finishing **1st or on the podium**, which aligns with their **dominant championship streak** during this era. The red dots clustered at the top reflect this dominance. During this period, the stock frequently responded positively to strong race performances, as seen by the green bars above the x-axis.

However, starting from **2022**, there's a noticeable shift:
- Mercedes' race positions are more scattered and less dominant.
- **No championships won post-2021**.
- When Mercedes does win an occasional race, there's a **small but visible positive stock movement**, indicating that **investor confidence is reactive to individual wins rather than overall dominance**.

Overall, this graph suggests a **stronger correlation between sustained dominance and stock performance prior to 2022**, while **occasional wins post-2021 produce more muted stock reactions**.
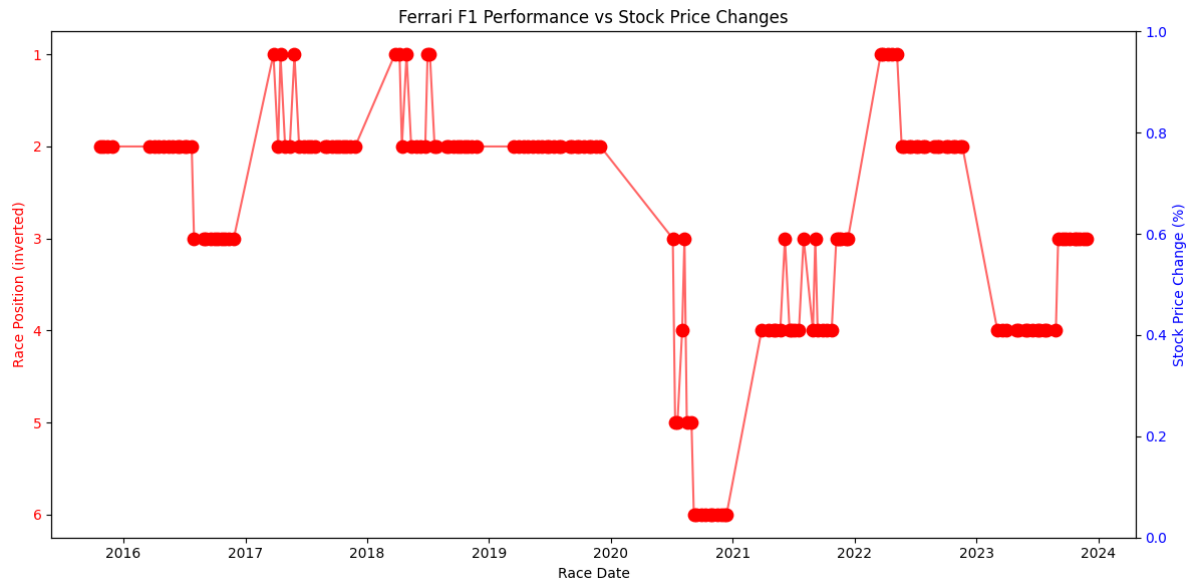
Same Pattern Can be Observed with the Ferrari team.

Data Modelling:

Target Variable: Binary classification of stock movement (price_up)
Data Split: 70% training / 30% testing (with balancing of classes)

Models Trained:

Model	Accuracy	F1 Score	ROC AUC
Logistic Regression	0.5490	0.7089	0.5124
Random Forest	0.4902	0.5938	0.3618
Voting Classifier	0.5938	0.5000	0.6872

Conclusion:
- Ferrari's F1 race performance shows some predictive signal for short-term stock movement.
- While linear correlation is weak, classification models, particularly ensemble models, demonstrate improved ability to detect stock direction (AUC = 0.6872).
- Additional features and financial context (earnings dates, news) could enhance the model further.

Project Timeline

1. **Milestone 1** (Data Collection & EDA) - February 5- February 21, 2025
2. **Milestone 2** (Advanced feature engineering, Model development for performance prediction & Initial dashboard structure) - February 21- March 15, 2025
3. **Milestone 3** (Dashboard development, Performance optimization & Final documentation and presentation) - March 15-April 15, 2025

Milestone 3: Evaluation, Interpretation, Tool Development, and Presentation

1.) Update to the Model - News Sentiment Analysis:

In addition to structured data, a news sentiment analysis module was built:

- Scraped recent articles mentioning the teams from the renowned formula1 news sites.

- Used a pre-trained transformer-based model to assign sentiment scores.

- **Findings**:

    o Positive sentiment after major wins aligned with slight upward stock movements.

    o Negative news (e.g., crashes, technical issues) sometimes correlated with stock dips, though inconsistently.

    o Sentiment data shows potential to improve predictive performance when combined with race results.

Race Week News Sentiment vs Stock Movement
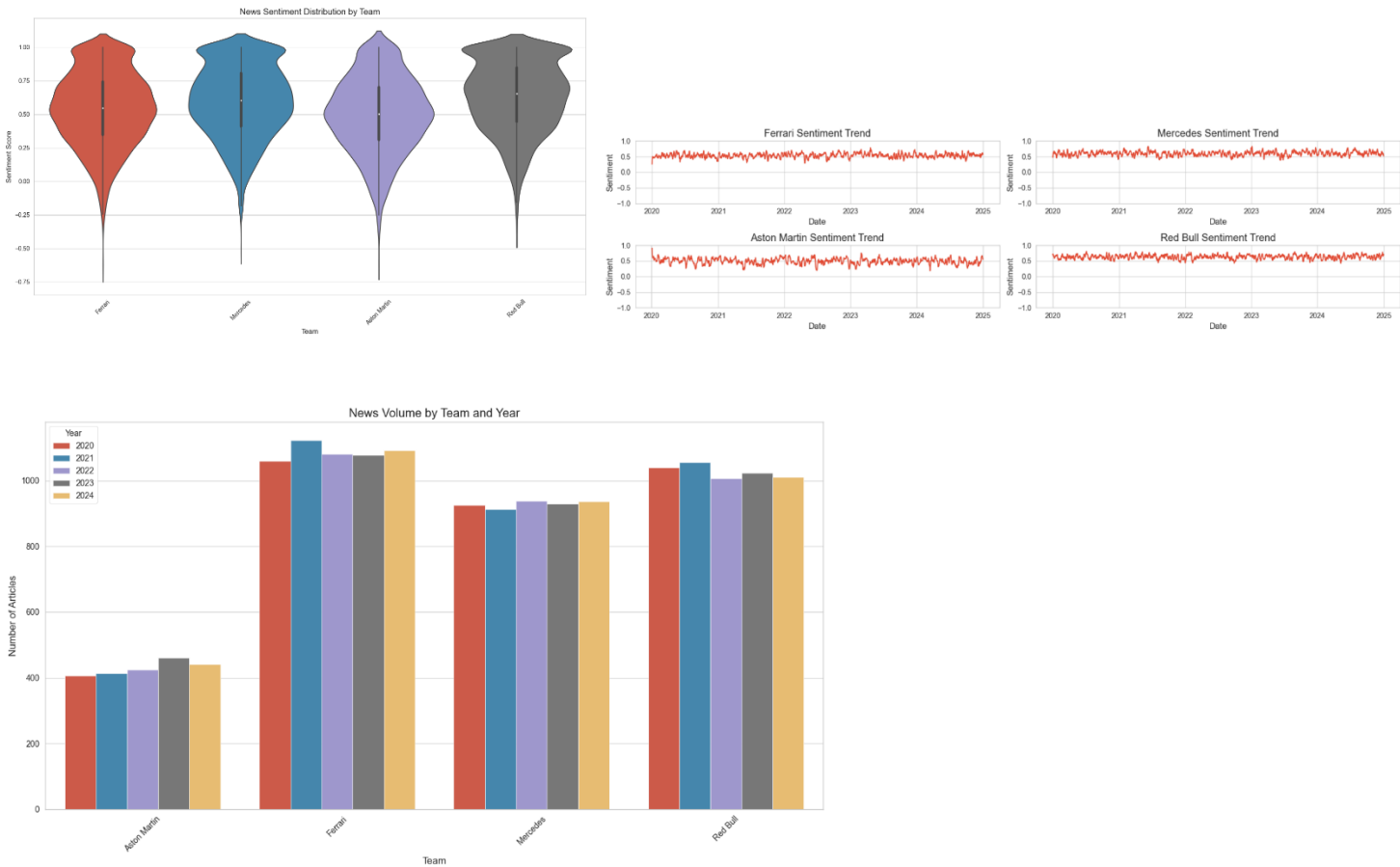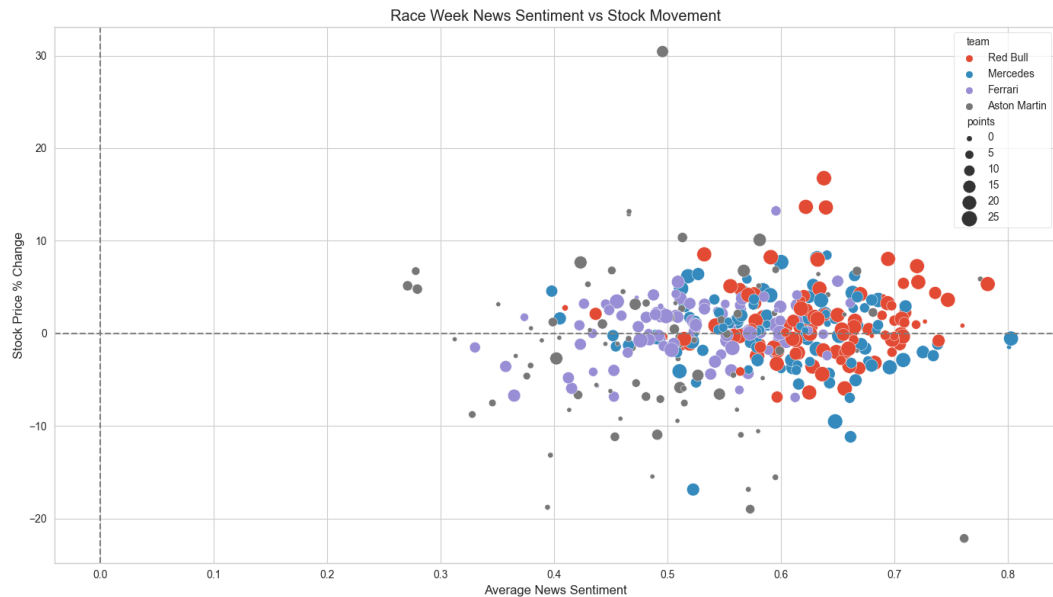
Final Ensemble Model Results :

A Voting Classifier Ensemble was trained using key race features (position, points, round):

Metric	Score
Accuracy	0.8889
F1 Score	0.8889
ROC AUC	0.8500
R2 Score	0.5577
Mean Absolute Error	0.2732
Mean Squared Error	0.1092

Bias and Limitations:

While this analysis provides valuable insights into the relationship between Formula 1 race performance and stock price movements, it is important to recognize several sources of potential bias and limitations:

- **Selection Bias**:
  The analysis focuses primarily on top-performing teams (Ferrari, Mercedes, Red Bull, Aston Martin) and publicly traded companies, which may skew results toward larger, more influential brands. Smaller teams or private companies were not included.

- **Survivorship Bias:**
  Only teams that are consistently present in recent seasons were analyzed. Teams that exited the championship or had fluctuating performances may have been overlooked, potentially underrepresenting broader dynamics.

- **Data Size Constraints:**
  After filtering for outliers and balancing classes, the final machine learning dataset was relatively small (only 28 balanced samples). This limits the generalizability and robustness of predictive models.

- **Short-Term Focus:**
  The stock price analysis focused on a short-term window (5 trading days before and after races). Longer-term financial impacts, season-wide performance, and other delayed reactions were not captured.

- **External Factors:**
  Stock prices are influenced by a wide range of external factors (global economic conditions, company earnings, management decisions) that are unrelated to race results. These were not controlled for in the analysis.

Final Findings and Reflections:

While the direct linear correlation between race results and stock price movements was relatively weak, this is expected given the complexity of stock market dynamics, where multiple macroeconomic and company-specific factors influence prices.

However, Formula 1 performance still exhibits measurable impact, particularly:

- Sustained dominance (multiple wins in a season) tends to correlate with positive stock momentum.

- Isolated race wins also have a short-term positive effect on sentiment and stock price.

- News sentiment surrounding a team's performance plays a crucial role and offers an additional predictive layer.

Thus, Formula 1 racing success does matter — especially when tied to brand image, sponsor relationships, and public perception.

Demo Video: [Tool Demo](#)

Project Presentation: [Presentation Slides](#)

TOOLS:

Interactive Dashboard Development

An interactive Streamlit dashboard was developed to showcase the full workflow — from data selection to performance analysis and modeling results.

- Users can select a year and a team (Ferrari, Mercedes, Red Bull, Aston Martin).

- Displays team statistics (podiums, wins, points, final position).

- Visualizes race position trends and stock price changes.

- Includes scatter plots and boxplots to explore the relationship between performance and market response.

- Embeds ensemble machine learning predictions directly within the app.
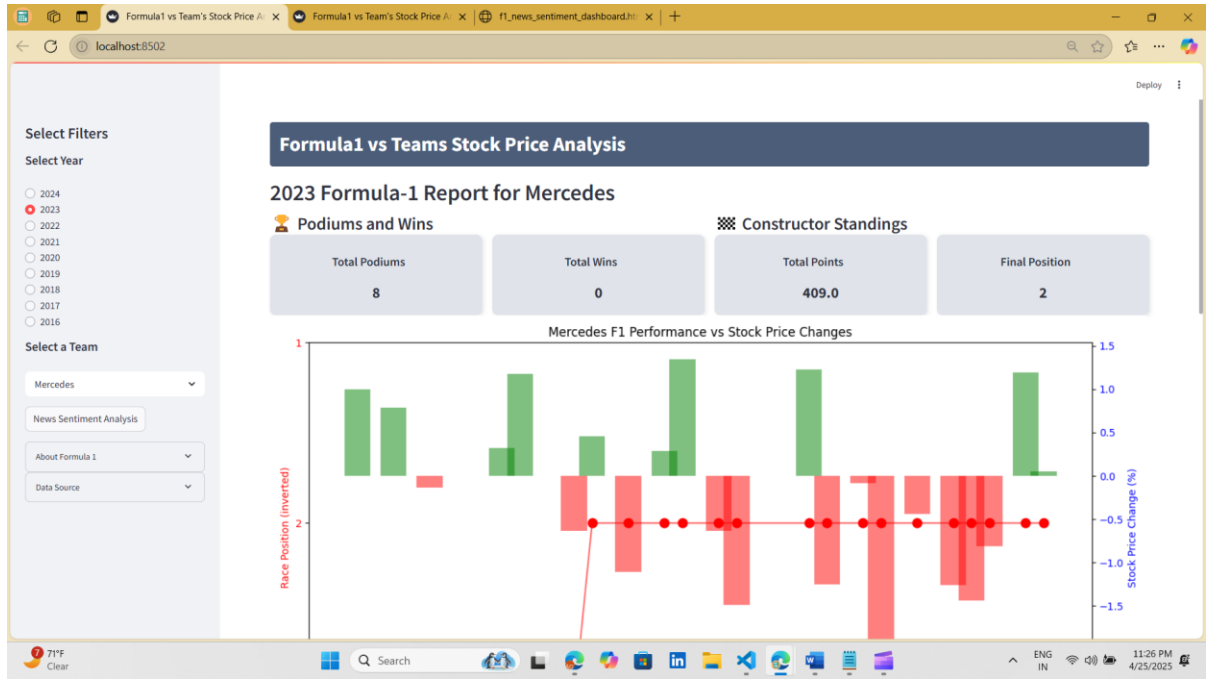
Final Note:

This project highlights the exciting intersection between sports analytics and financial market behavior, showcasing how data science can uncover hidden patterns even in complex, multi-variable environments.

References:

1.) F1 - The Official Home of Formula 1® Racing

2.) Yahoo Finance - Stock Market Live, Quotes, Business & Finance News

3.) The Economics of Formula 1 - Hivelr

4.) Inside The Numbers That Show Formula 1's Popularity And Financial Growth