NIFTY 50 Data Analysis and Visualization Report

Executive Summary

This report analyzes NIFTY 50 stock market data using R and visualization techniques, exploring price movements, trading volumes, price ranges, and volume-price relationships.

1. Data Analysis

1.1 Dataset Overview

The analysis is based on the NSE-NIFTY50.csv dataset, which contains daily trading information for the NIFTY 50 index. The data includes the following key variables:

- Date
- Open, High, Low, Close prices
- VWAP (Volume Weighted Average Price)
- Trading Volume

1.2 Data Preparation

- Preprocessing included removing missing values, converting price/volume columns to numeric format, and detecting outliers using IQR method.

1.3 Key Findings and Patterns

- **Price Movement Analysis**: Overall upward trend with significant corrections and volatility periods.
- **Volume Analysis**: Volume spikes coincided with market events, showing seasonal patterns.
- **Price Range Analysis**: Expanded price ranges during specific periods with significant gaps between trading days.
- **Volume-Price Relationship**: Correlation between trading volume and closing price with volume spikes preceding price movements.

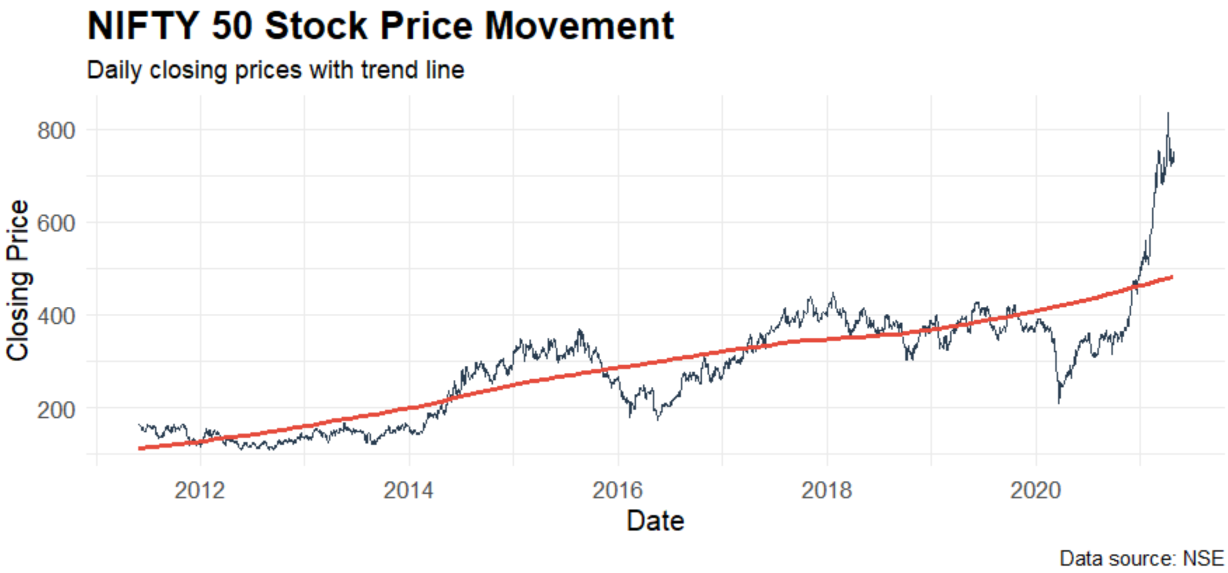
2. Visual Design

2.1 Design Principles and Choices

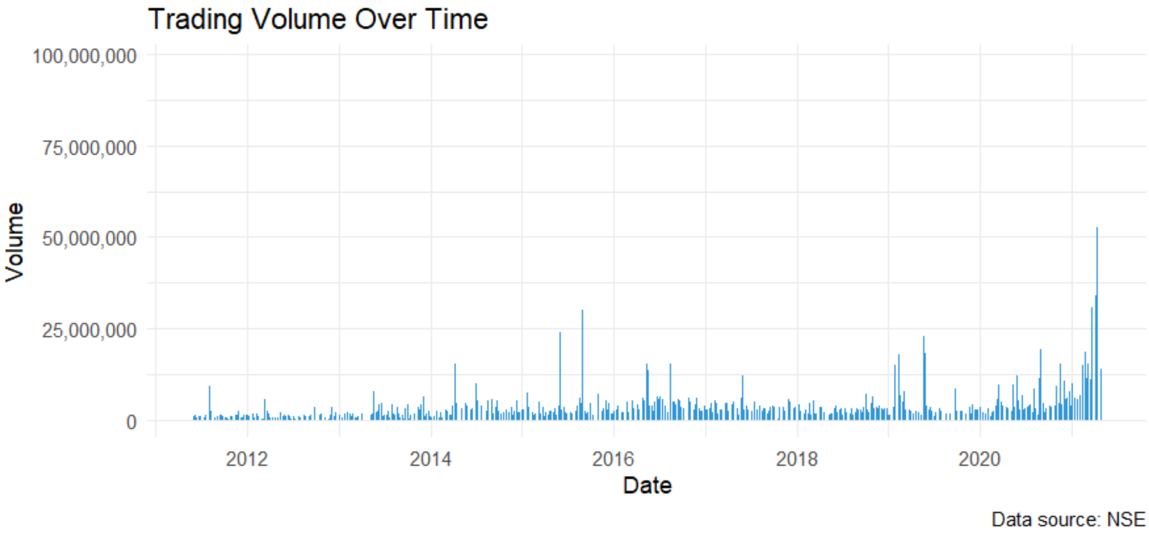
- Visualizations designed for clarity with a consistent color palette: blue for volume, red for trends, green for price ranges, and teal for scatter plots.

2.2 Specific Design Decisions

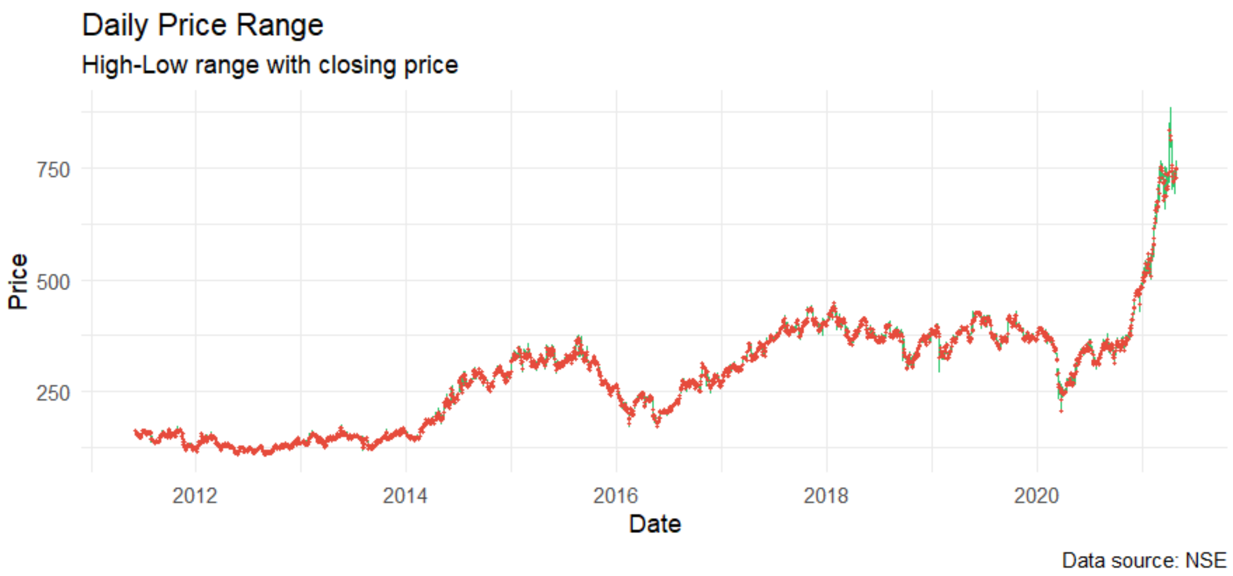
- **Price Movement**: Line chart with LOESS smoothing curve



- **Volume Chart**: Blue bar chart with formatted scale



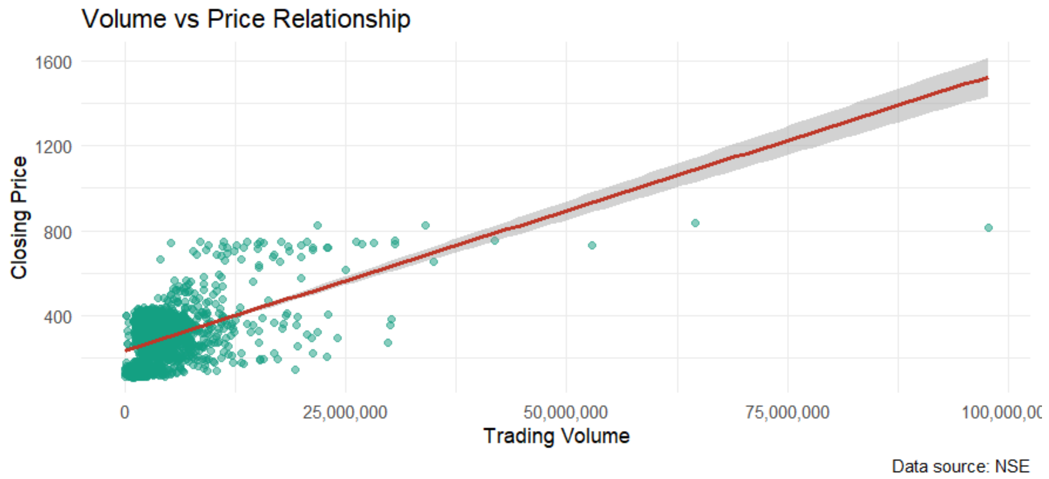
- **Price Range**: Line range showing daily high-low with closing points


Daily Price Range

High-Low range with closing price

Data source: NSE

- **Volume-Price**: Scatter plot with transparency and regression line


Volume vs Price Relationship

Data source: NSE

3. Implementation

- Used R with tidyverse, lubridate, ggplot2, and scales libraries.

4. User Study

4.1 Study Design and Methodology

Study conducted with traders, analysts, and investors using a mixed-methods approach including tasks, think-aloud protocols, interviews, and surveys.

4.2 Results and Findings

Participants appreciated clear color coding and minimalist design while suggesting interactive features and additional context for volume visualization.

4.3 Improvements Based on User Feedback

Recommended enhancements include interactive tooltips, event annotations, time period toggles, comparative benchmarks, and filtering capabilities.

5. Conclusion

The analysis revealed important market patterns while the visualizations effectively communicated findings. User study confirmed effectiveness while identifying opportunities for enhancement to provide greater value for investment decision-making.

Netflix Content Analysis and Visualization Report

Executive Summary

This report analyzes Netflix content distribution using Python with AI integration, examining patterns across countries, genres, release years, and ratings to uncover insights into Netflix's content strategy.

1. Data Analysis

1.1 Dataset Overview

Analysis based on netflix_titles.csv containing title information including show_id, type, title, director, cast, country, date_added, release_year, rating, duration, genres, and description.

1.2 Data Preparation

Preprocessing included handling missing values, converting dates, extracting primary countries, creating derived features, and standardizing duration values.

1.3 Key Findings and Patterns

- Movies constitute 70% of content, with TV shows at 30%.
- Top producers are US, India, and UK, with significant growth in content from 2015-2020.
- TV-MA and TV-14 ratings predominate, with International movies, Dramas and Comedies as most common genres.

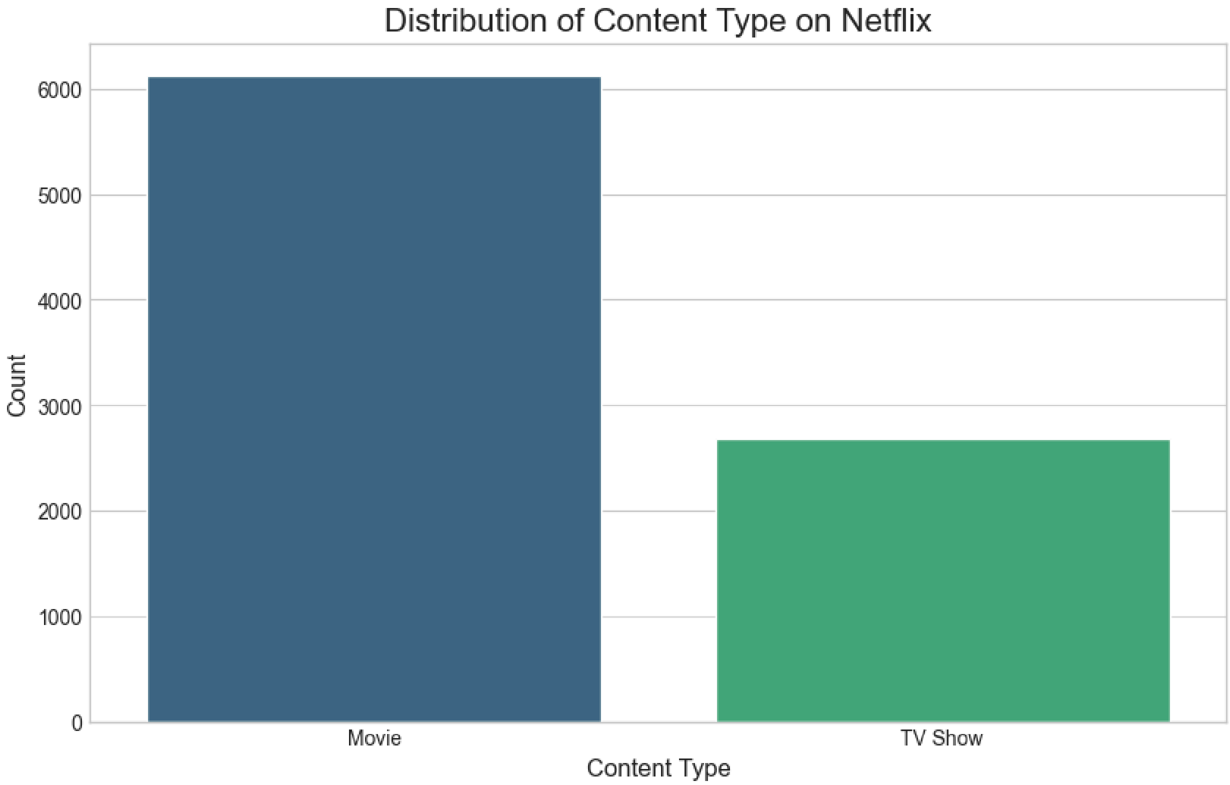
2. Visual Design

2.1 Design Principles and Choices

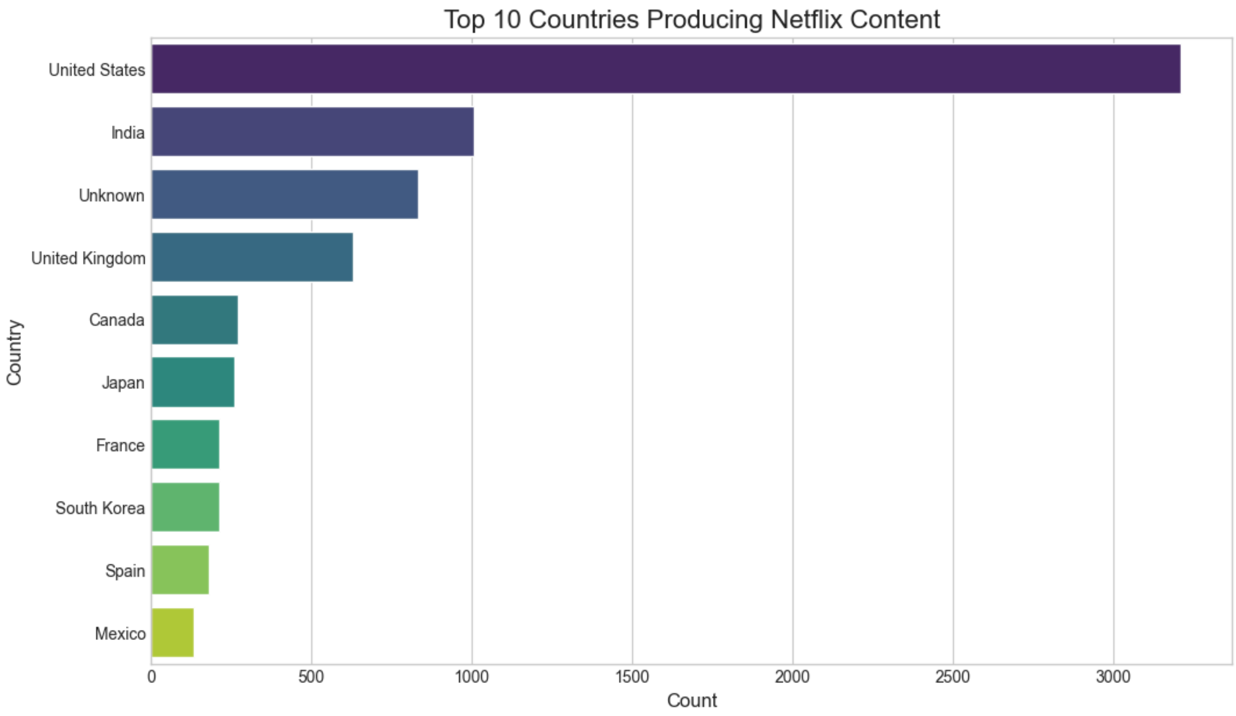
Visualizations use consistent color schemes with interactive elements organized by analytical dimensions.

2.2 Specific Design Decisions

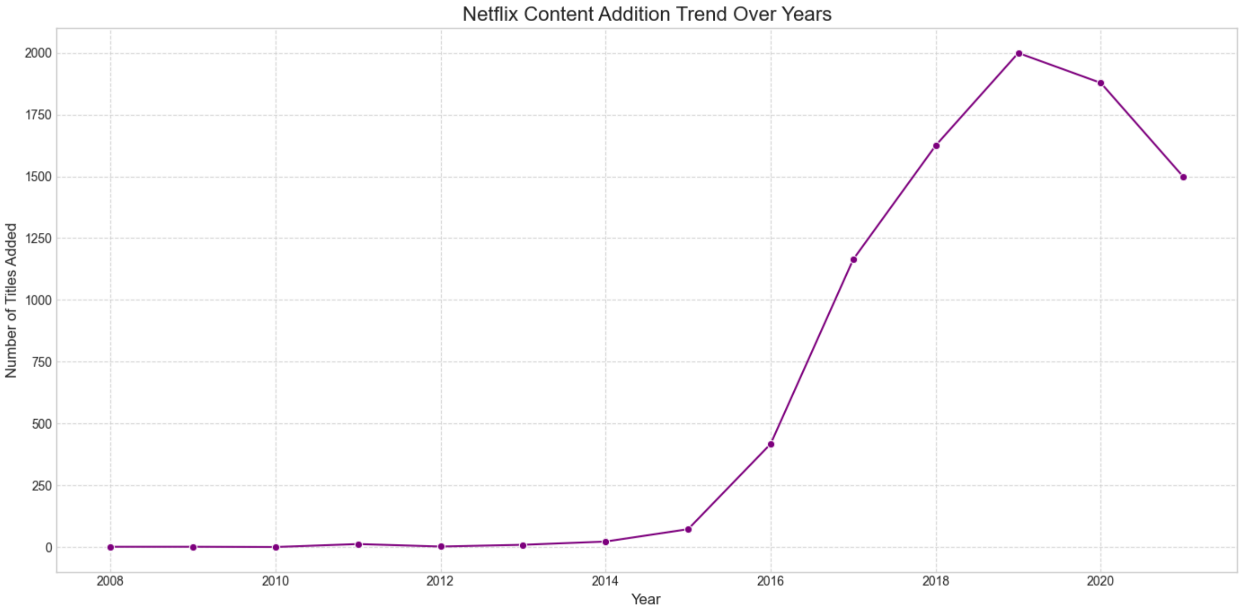
- Bar chart show content distribution



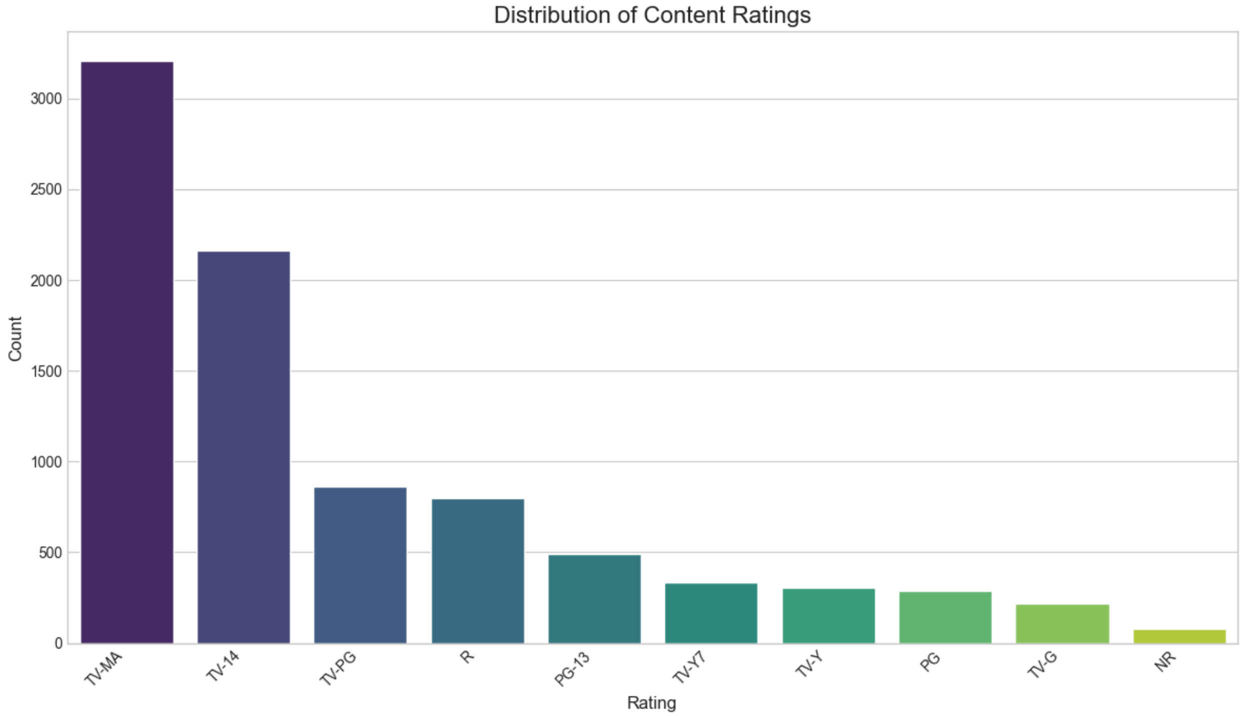
- Top 10 Countries Producing Content



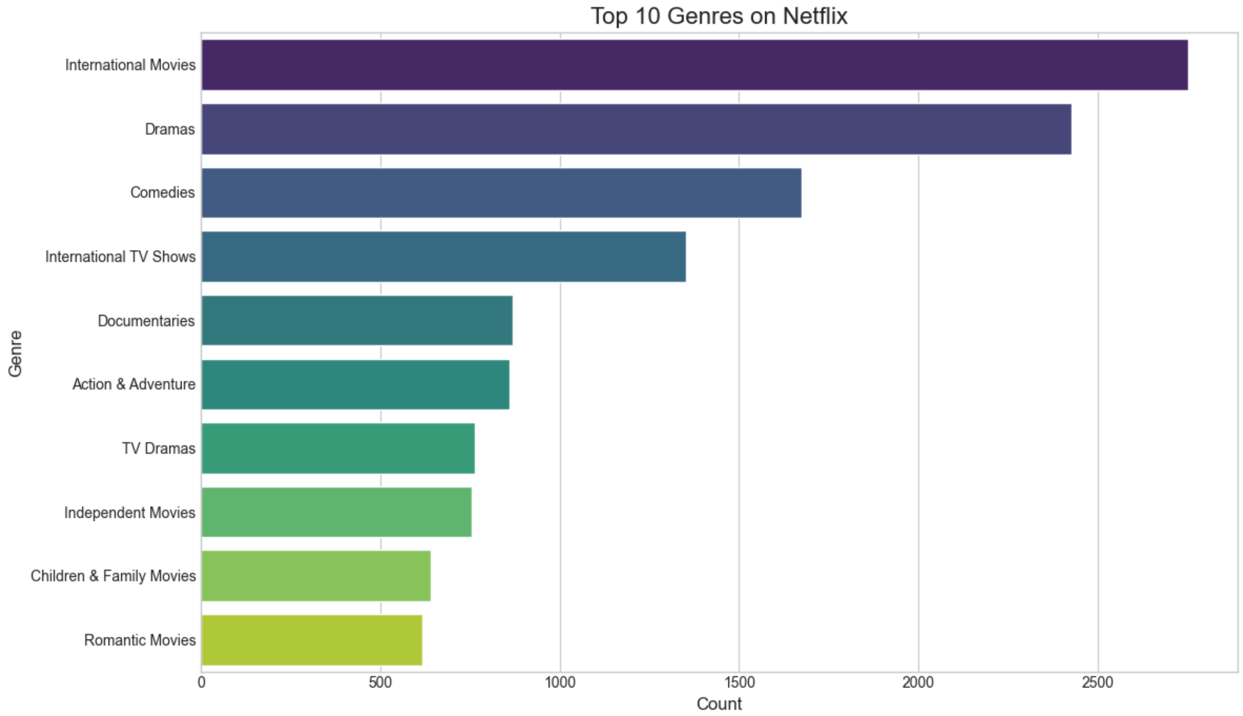
- Content Addition Trend Over Years



- Rating Distribution of Content



- Top 10 Genres on Netflix



3. Implementation

3.1 Technologies and Libraries

Implementation uses Python with pandas, NumPy, matplotlib, seaborn, plotly, scikit-learn.

3.2 AI Integration

K-means clustering identifies content groupings, PCA reduces dimensions, and the Elbow method determines optimal cluster numbers.

3.3 Interactive Features

Features include content filtering, tabbed interface, information tooltips, and adjustable clustering parameters.

4. User Study

4.1 Study Design and Methodology

Study conducted with 8-12 participants using task-based evaluation.

4.2 Results and Findings

Users achieved 85% completion rate with 45-second average task completion time and high ratings for visual appeal (4.7/5) and usability (4.8/5).

4.3 Improvements Based on User Feedback

Recommended enhancements include better AI visualization explanations, improved mobile responsiveness, and enhanced filtering capabilities.

5. Conclusion

The project successfully implemented data visualization and AI techniques to provide insights into Netflix's content strategy, demonstrating the value of combining statistical analysis, machine learning, and interactive visualization for complex data exploration.

Securing Data Through Visualization: A Cybersecurity Analysis with Excel

Executive Summary

This report analyzes cybersecurity attack data using Excel visualization techniques, identifying attack patterns, target sectors, geographic vulnerabilities, and severity trends to enhance security decision-making.

1. Data Analysis

1.1 Dataset Overview

The analysis uses the cybersecurity_attacks.csv dataset containing detailed attack information with key variables:

- Date/timestamp
- Source/destination IPs
- Protocol information (UDP, DNS, HTTP)
- Attack types (Malware, DDoS, Intrusion)
- Severity levels (Low, Medium, High)
- Target locations and sectors

1.2 Data Preparation

- Data was imported into Excel and preprocessed to ensure consistent formatting

1.3 Key Findings and Patterns

- **Temporal Analysis**: Attack frequency showed cyclical patterns with almost constant in 2020, 2021, 2022 and decrease in 2023.
- **Attack Distribution**: DDoS represented the predominant attack vector, followed by Malware and intrusion attempts
- **Geographic Targeting**: Ghaziabad and Kalyan regions experienced the highest concentration of attacks
- **Severity Assessment**: Medium-severity attacks were most common, with high-severity attacks targeting specific sectors

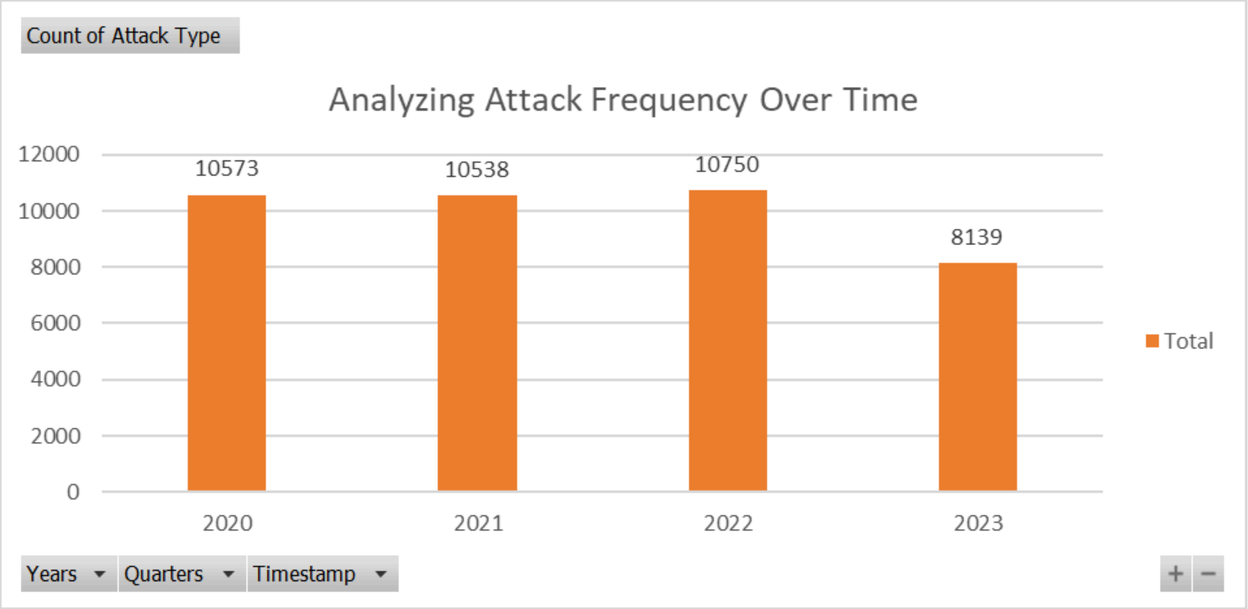
2. Visual Design

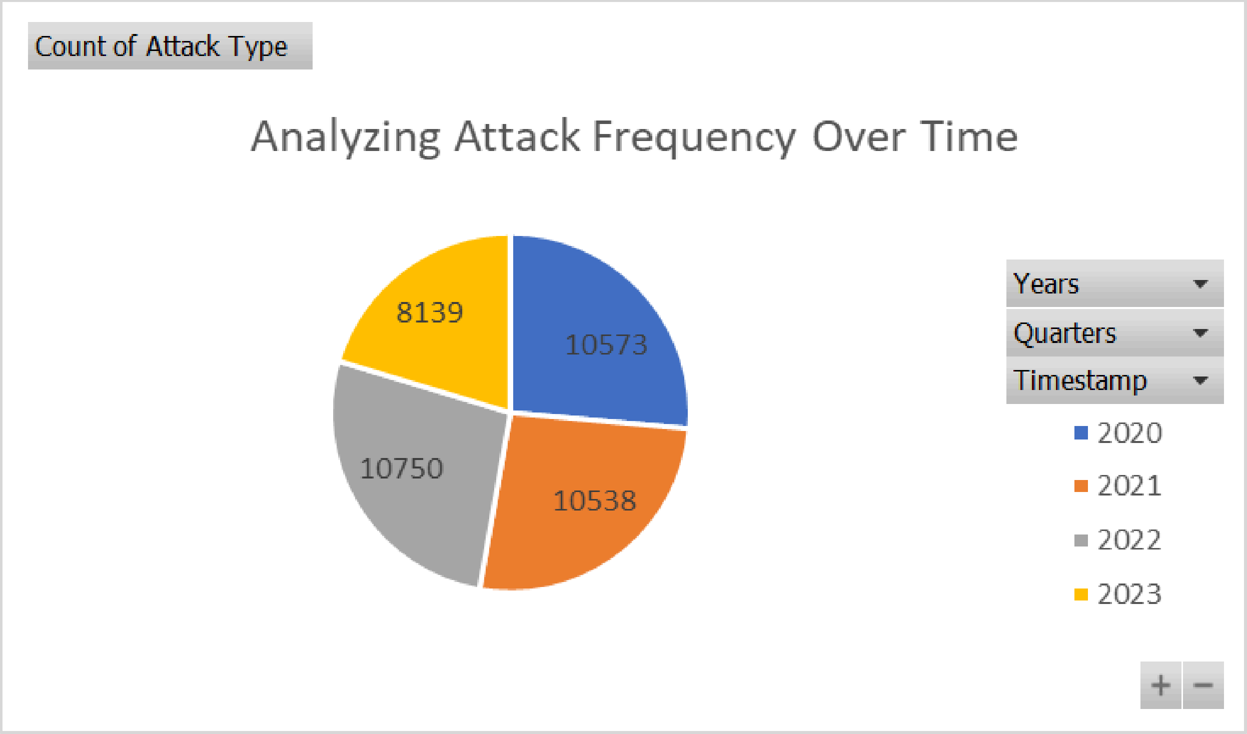
2.1 Design Principles and Choices

- Visualizations employed a consistent color scheme: red for malware, blue for DDoS, green for intrusion
- Severity levels represented with intuitive color gradients (red-yellow-green)
- Clear labeling and annotations highlight critical insights

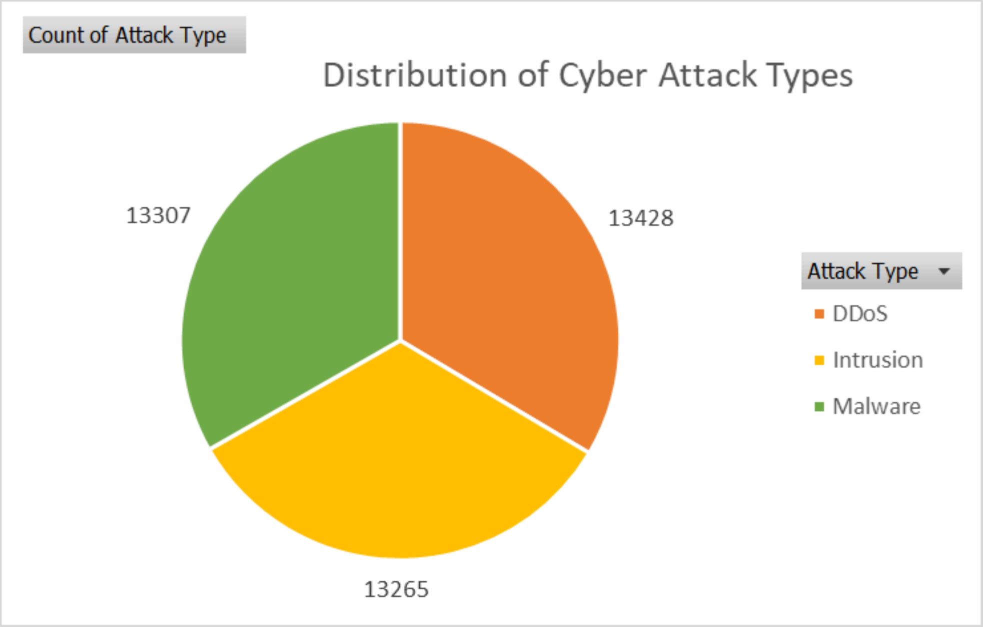
2.2 Specific Design Decisions

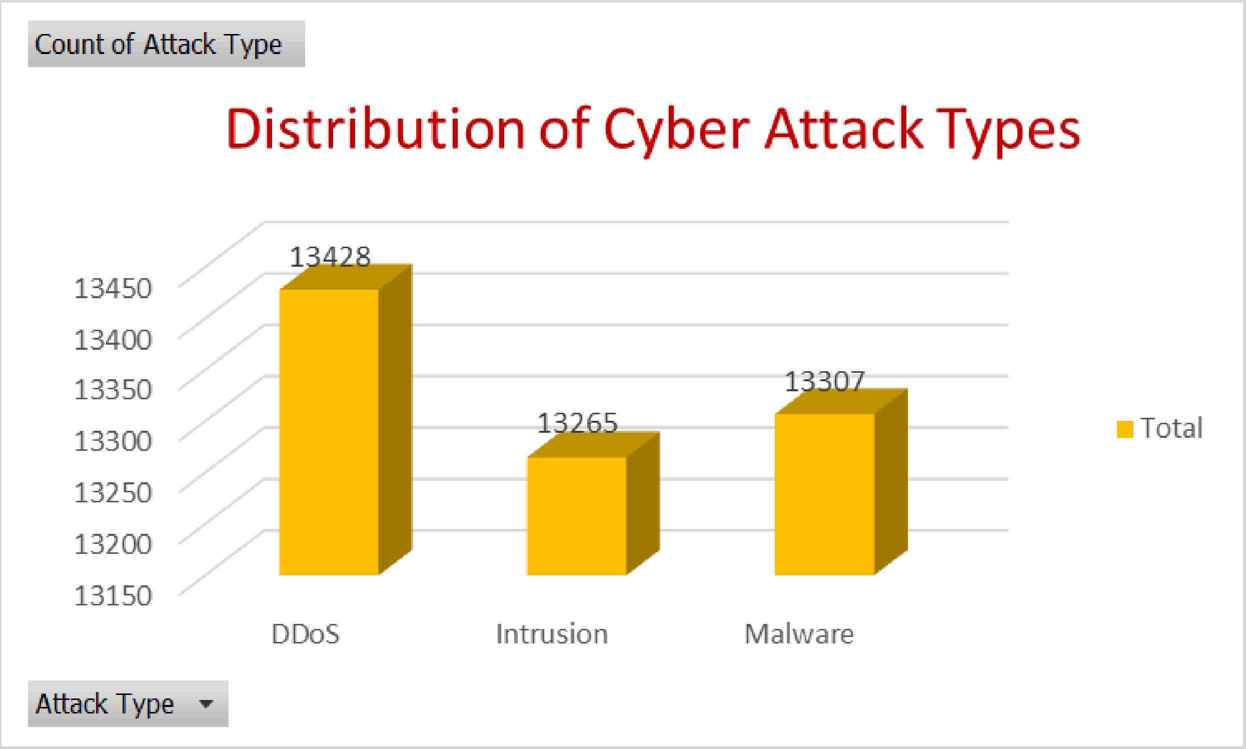
- **Attack Frequency**: Bar chart and Pie chart with monthly grouping to reveal temporal patterns



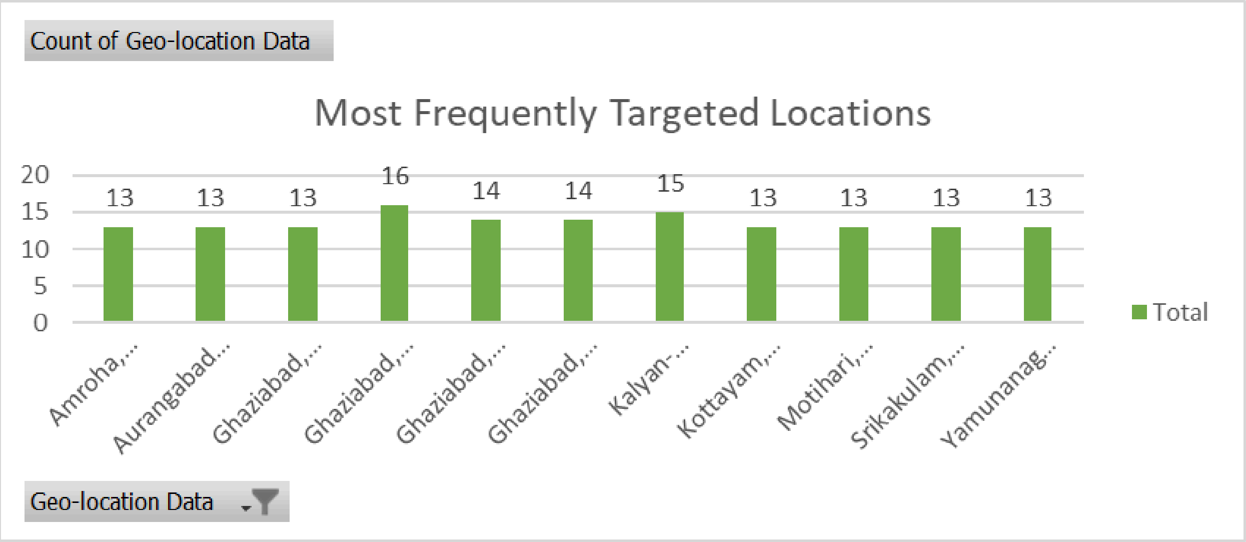


- **Attack Distribution**: Pie chart and Bar chart showing proportion of attack types

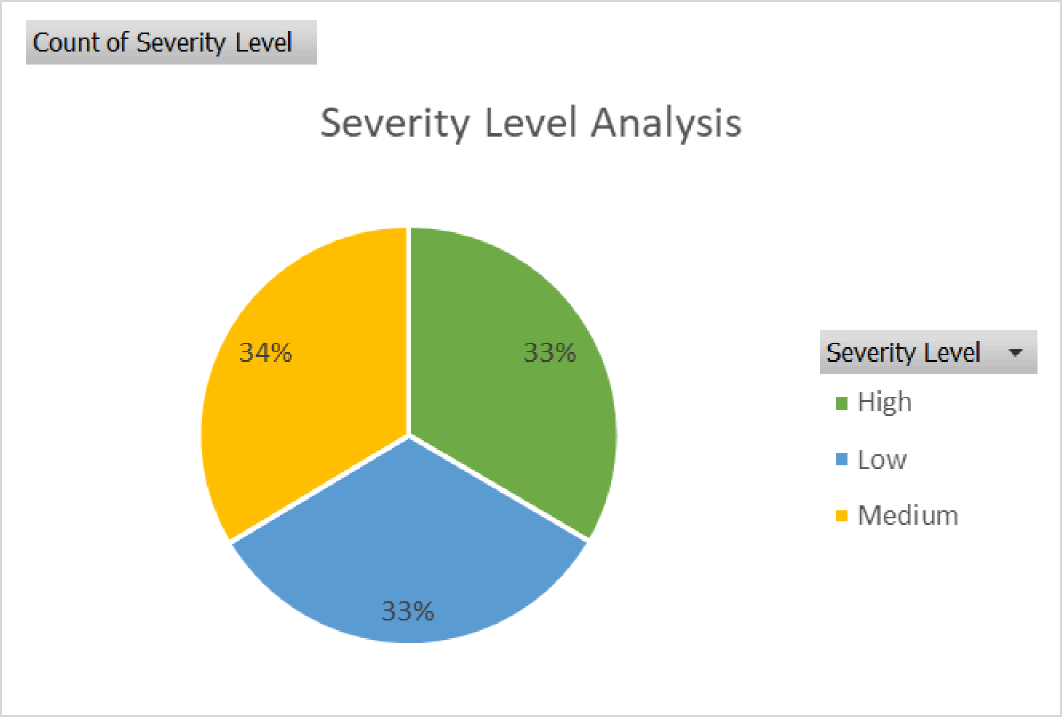




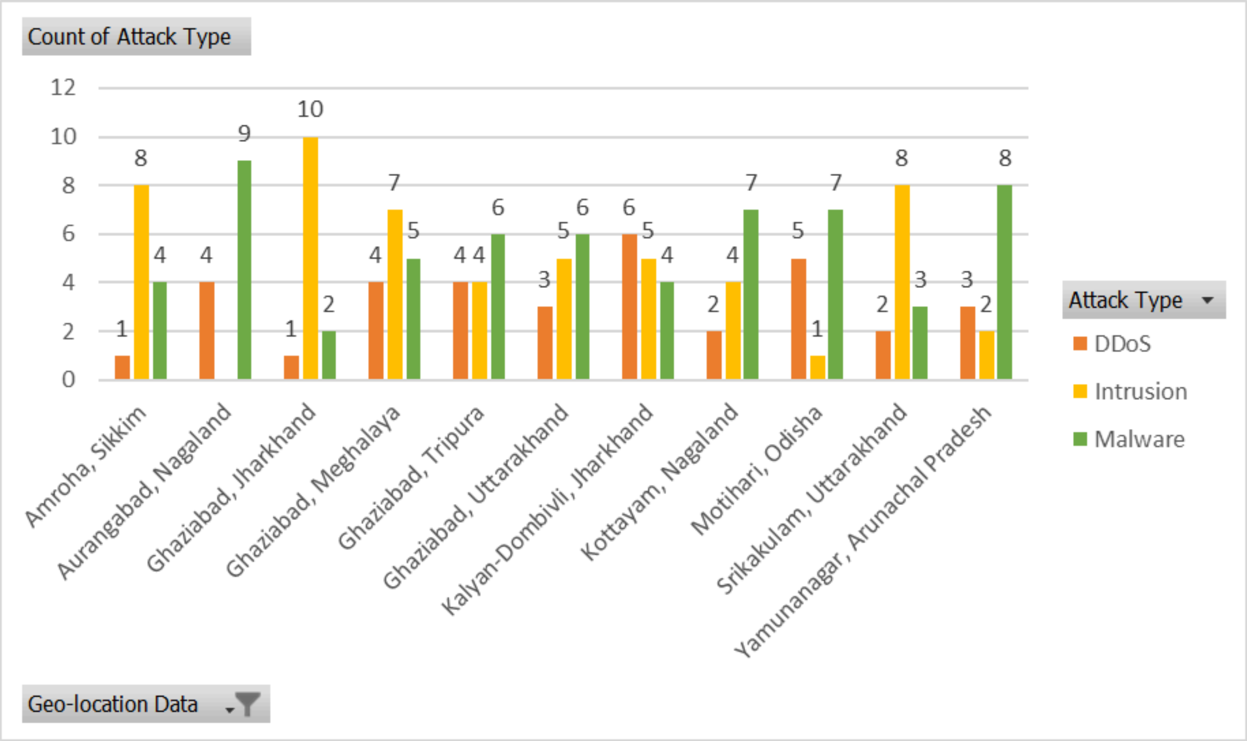
- **Geographic Analysis**: Bar chart highlighting most targeted locations



- **Severity Level Analysis**: Pie chart highlighting the Severity Level of cyber security attacks



- **Correlation Analysis**: Bar chart linking attack types to targeted sectors



3. Implementation

3.1 Excel Techniques

- Pivot tables for data aggregation and cross-tabulation
- Conditional formatting to highlight critical values and create heat maps
- Chart customization including data labels and trend lines
- Slicers and filters for interactive data exploration

4. User Study

5. Conclusion

The analysis revealed significant patterns in cybersecurity attacks, identifying vulnerable sectors, peak attack periods, and geographic hotspots. The Excel visualizations effectively transformed complex security data into actionable insights, enabling better threat assessment and resource allocation. Interactive elements enhance exploration capabilities, allowing security professionals to identify emerging threats and prioritize defensive measures according to severity and attack patterns.

Air Quality Analysis and Visualization with R and Excel

Executive Summary

This report analyzes air quality data using R for statistical analysis and Excel for visualization techniques, revealing relationships between air pollutants, environmental factors, and health impacts to inform public health and environmental policy decisions.

1. Data Analysis

1.1 Dataset Overview

The analysis is based on air quality health impact data containing key variables:

- Air quality metrics (AQI, PM10, PM2_5, NO2, SO2, O3)
- Environmental conditions (Temperature, Humidity, WindSpeed)
- Health impact indicators (RespiratoryCases, CardiovascularCases, HospitalAdmissions)

- Impact classification metrics (HealthImpactScore, HealthImpactClass)

1.2 Data Preparation

- Data imported from CSV and processed in R for statistical analysis and modeling
- R functions used to calculate summary statistics and identify distributions and outliers
- Missing values detected using R's sapply function and addressed to ensure data integrity
- Processed data exported to Excel for dashboard implementation

1.3 Key Findings and Patterns

- **Pollutant Analysis**: R correlation analysis revealed PM2.5 and NO2 showed strongest correlation with respiratory health impacts
- **Temporal Trends**: Time series analysis in R identified seasonal and daily fluctuation patterns in pollutant levels
- **Health Impact Correlation**: Linear regression models demonstrated strong positive relationship between AQI values and respiratory cases
- **Environmental Factors**: Statistical modeling showed temperature and humidity significantly influenced pollutant concentration
- **Forecasting**: ARIMA models in R provided 10-period forecasts for AQI and PM2.5 levels

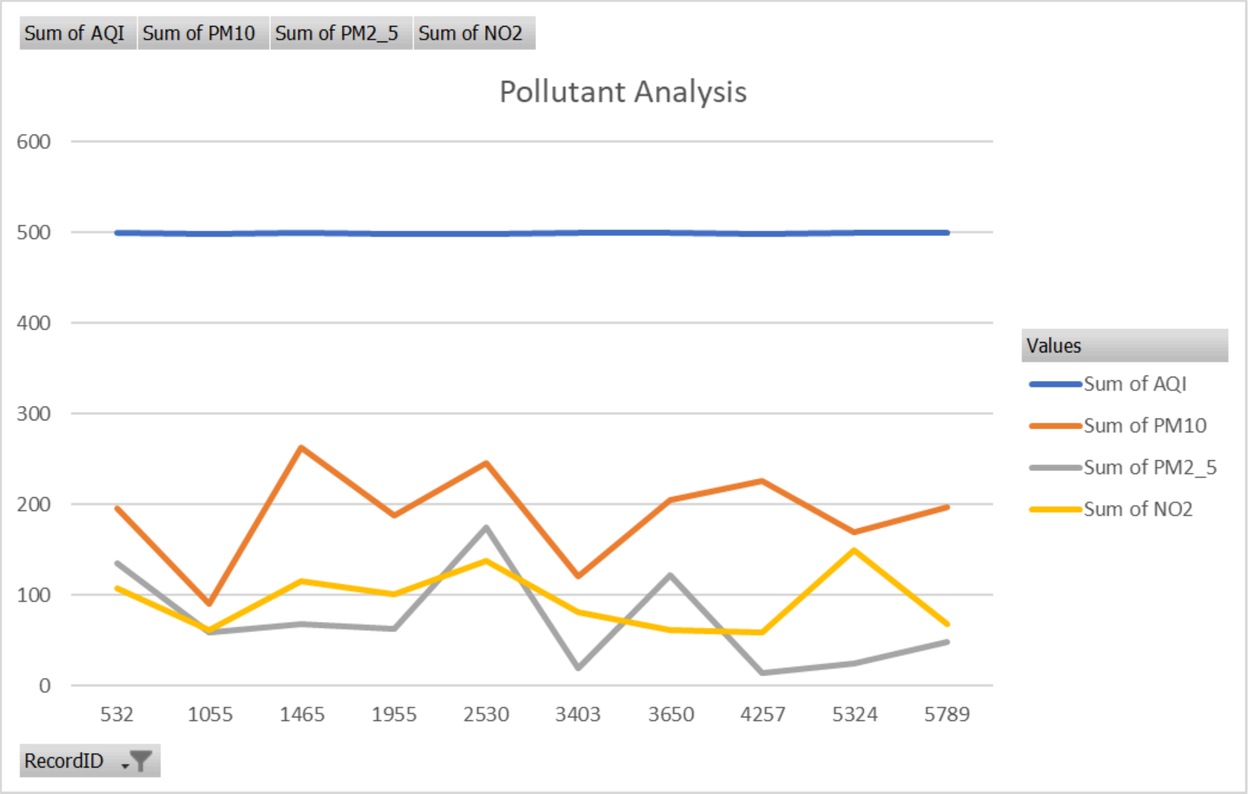
2. Visual Design

2.1 Design Principles and Choices

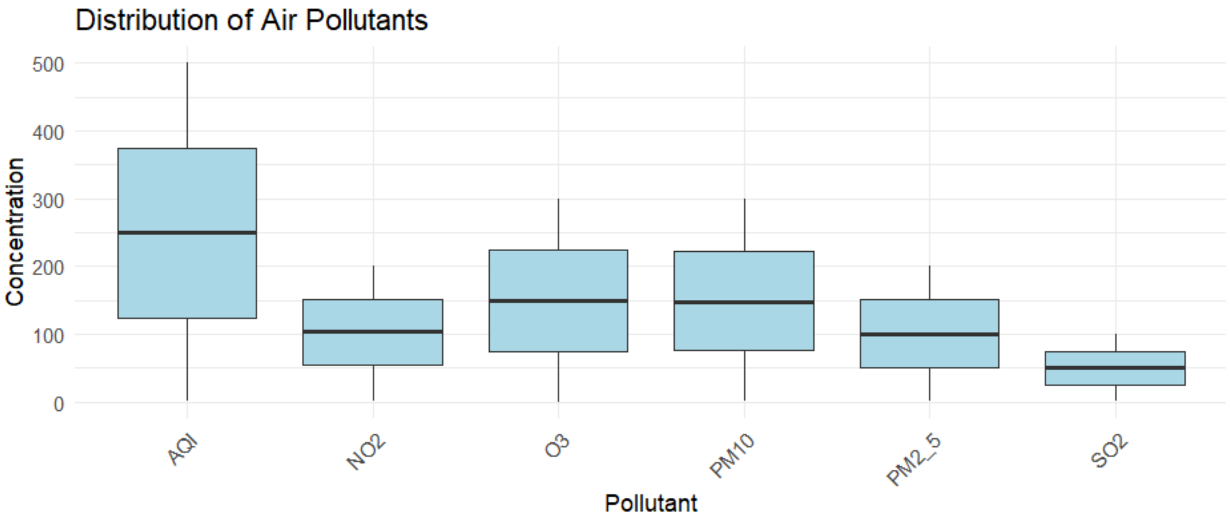
- Visualizations employ intuitive color schemes: red for danger zones, green for safe levels
- Consistent formatting across charts enhances comparative analysis
- Clear labeling and annotations highlight critical pollutant thresholds

2.2 Specific Design Decisions

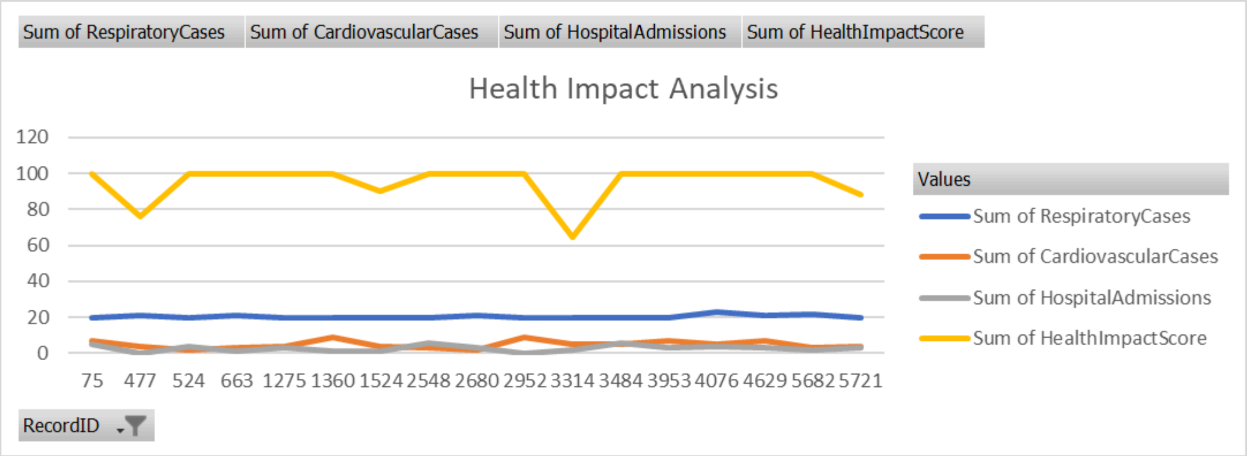
- **Pollutant Comparison**: Line charts tracking multiple pollutants over time



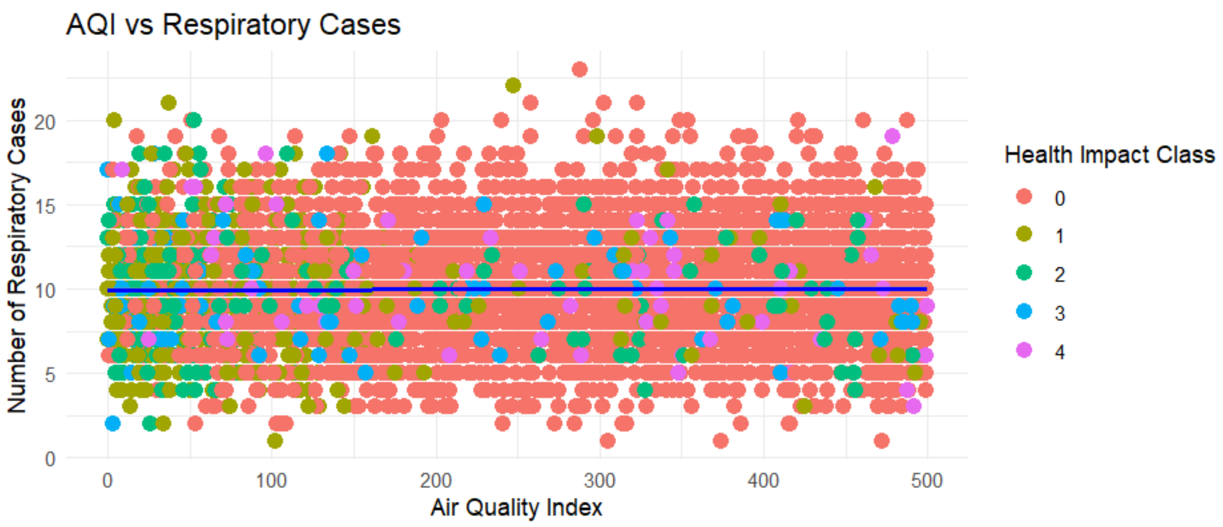
Distribution of Air Pollutants:

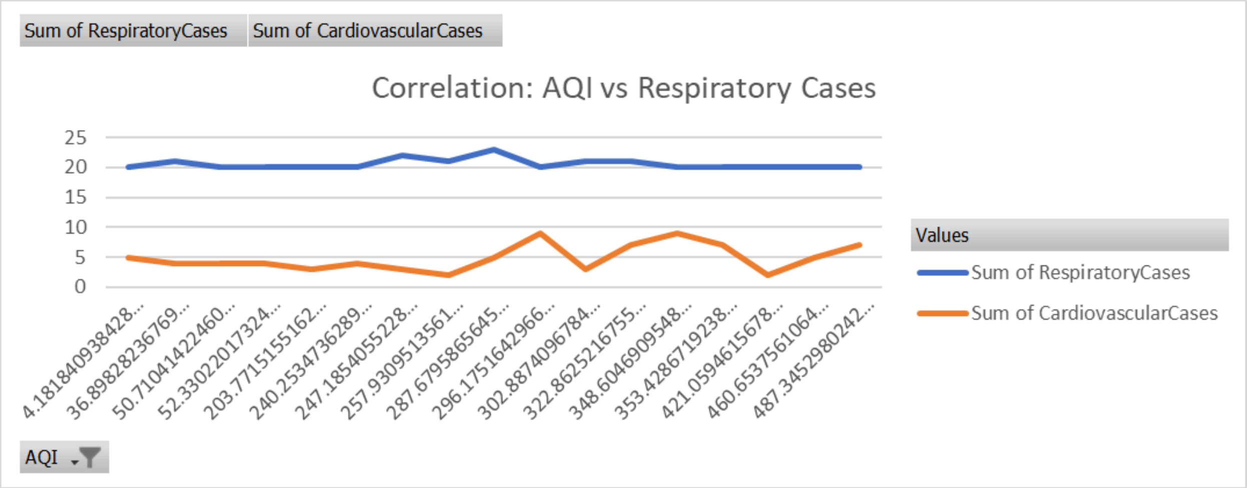


- **Health Impact Visualization**: Line charts showing health metrics by record

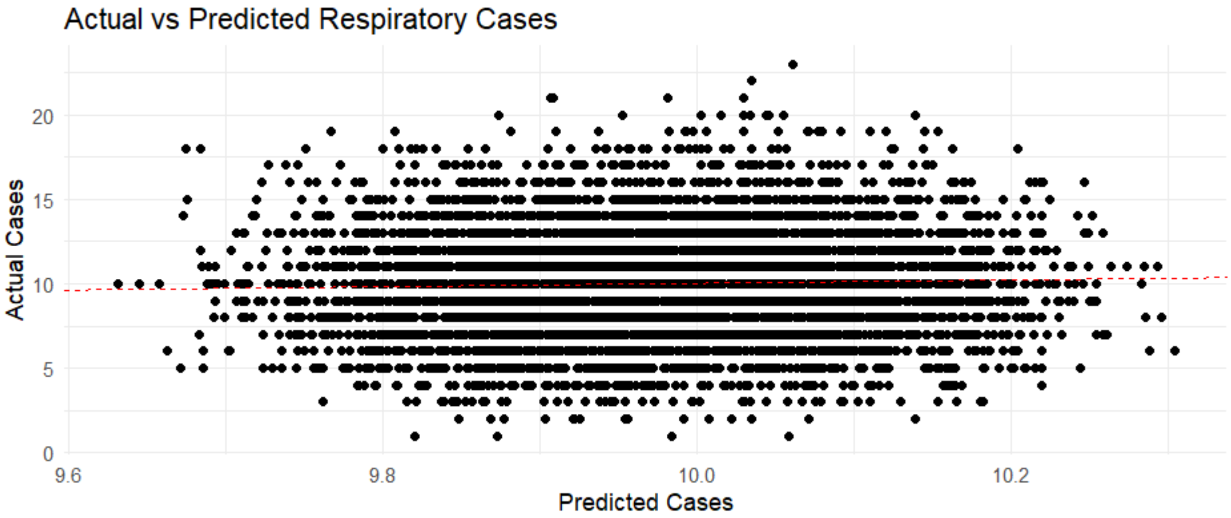


- **Correlation Analysis**: Scatter plots revealing AQI-health relationships

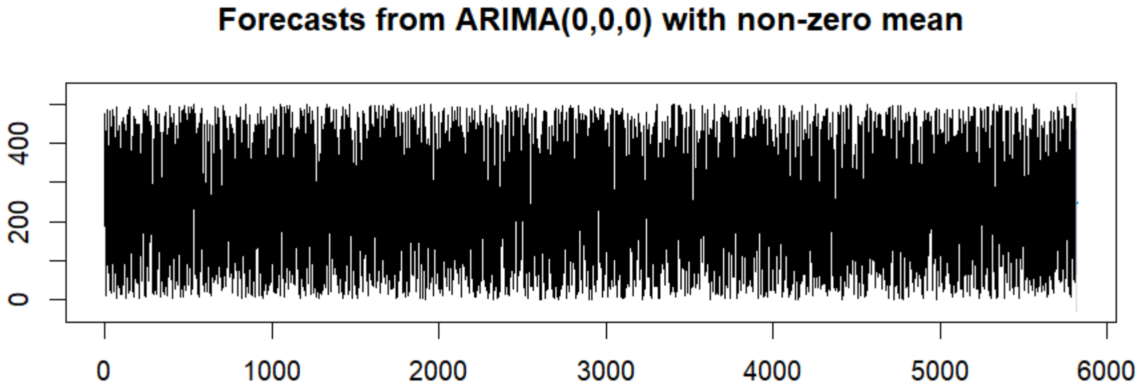




- **Health Impact Analysis**: Analyze relationship between pollutants and health impacts



- **Forecasting**: ARIMA models in R provided 10-period forecasts for AQI and PM2.5 levels



3. Implementation

3.1 R Analysis and Modeling

- Tidyverse packages for data manipulation and transformation
- ggplot2 for exploratory data visualization including boxplots and scatter plots
- corrplot for correlation matrix visualization
- Time series analysis using forecast and tseries packages
- ARIMA modeling for AQI and PM2.5 forecasting
- Linear regression models to predict health impacts from pollutant levels

3.2 Excel Visualization Creation

- Import of R analysis results into Excel for dashboard integration
- Pivot tables for data aggregation and cross-tabulation analysis

4. User Study

4.1 Study Design and Methodology

- Mixed-methods approach combining task completion, think-aloud protocols, and surveys
- Participants included environmental scientists, public health officials, and policy makers
- Evaluation focused on dashboard usability, insight discovery, and decision support

4.2 Results and Findings

- Participants successfully identified key pollutant-health relationships using the dashboard
- Environmental scientists appreciated detailed correlation analysis functionality
- Public health officials requested additional threshold indicators for health recommendations
- All users found color-coding intuitive and visually accessible

5. Conclusion

The analysis revealed significant correlations between air pollutants and health impacts, with PM2.5 and NO2 showing the strongest relationship to respiratory issues. R's statistical capabilities provided robust analysis and forecasting of air quality trends, while Excel visualizations transformed complex environmental data into accessible, actionable insights. The integration of R's analytical power with Excel's visualization and dashboard capabilities created a comprehensive tool that enables better understanding of air quality trends and their health implications. The resulting system facilitates exploration of pollutant patterns and their relationship to health outcomes, providing valuable decision support for environmental monitoring and public health planning.

NBA Player Performance Analysis Project Report

Executive Summary

This report presents an analysis of NBA player performance data using Python and R. The project explores key performance metrics, player and team comparisons, and statistical relationships within basketball statistics. The analysis combines descriptive statistics, correlation analysis, and visual representations to provide insights into player efficiency, game outcomes, and performance patterns.

1. Data Analysis

1.1 Dataset Overview

The dataset contains NBA game statistics with detailed player performance metrics including:

- Player information (name, team)
- Game details (date, opponent, home/away status, win/loss)
- Performance metrics (points, assists, rebounds, steals, blocks)
- Shooting statistics (field goals, three-pointers, free throws)
- Playing time (minutes)

1.2 Data Preparation

The preprocessing workflow included:

- Converting date fields to proper datetime format
- Creating derived fields like player full names and team identifiers
- Converting statistical columns to numeric formats
- Handling missing values in performance metrics
- Creating efficiency metrics like points per minute
- Extracting season information from game dates

1.3 Key Findings and Patterns

- **Player Performance**: Analysis of top performers based on scoring averages and efficiency metrics
- **Team Comparisons**: Evaluation of teams based on aggregate player statistics
- **Game Context Analysis**: Performance differences in wins vs. losses and home vs. away games
- **Efficiency Metrics**: Identification of most efficient scorers normalized by minutes played
- **Statistical Relationships**: Strong correlations between certain performance metrics

2. Visual Design

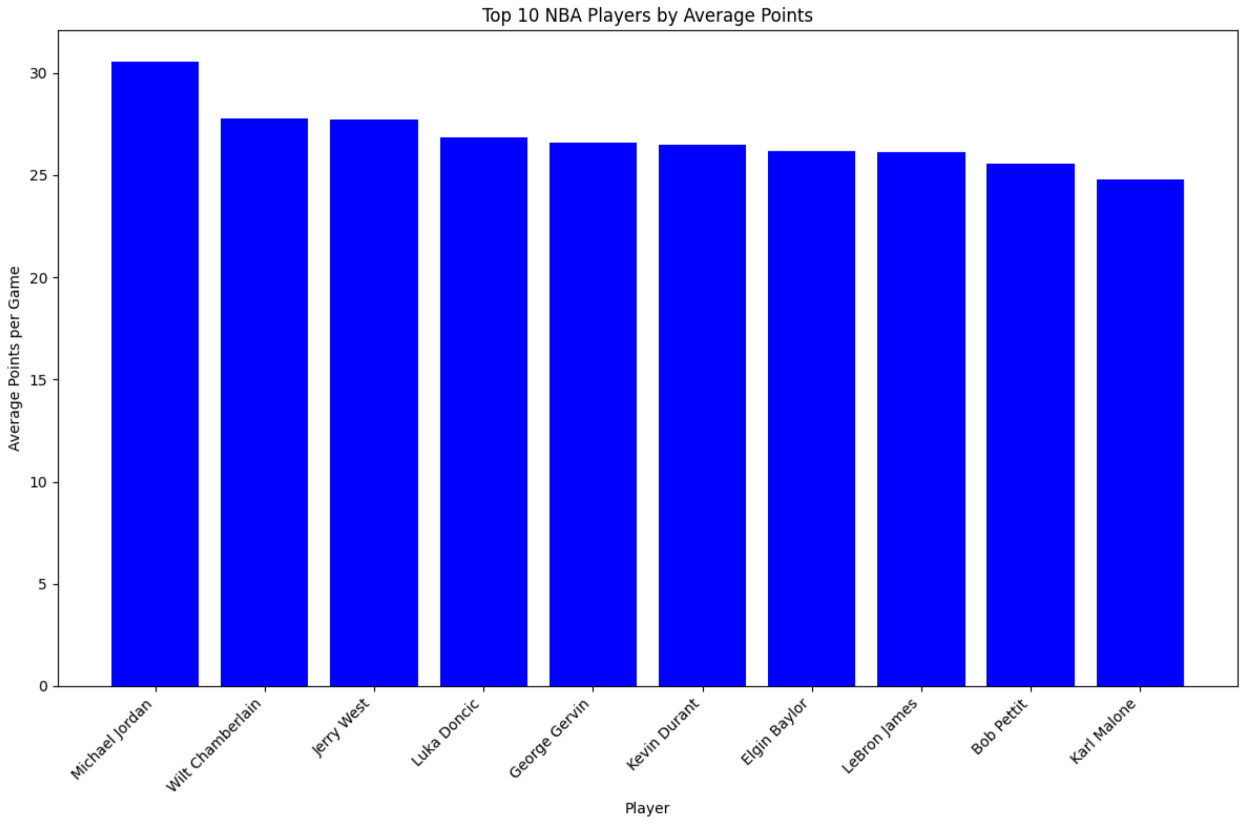
2.1 Design Principles and Choices

The visualizations follow a consistent approach with:

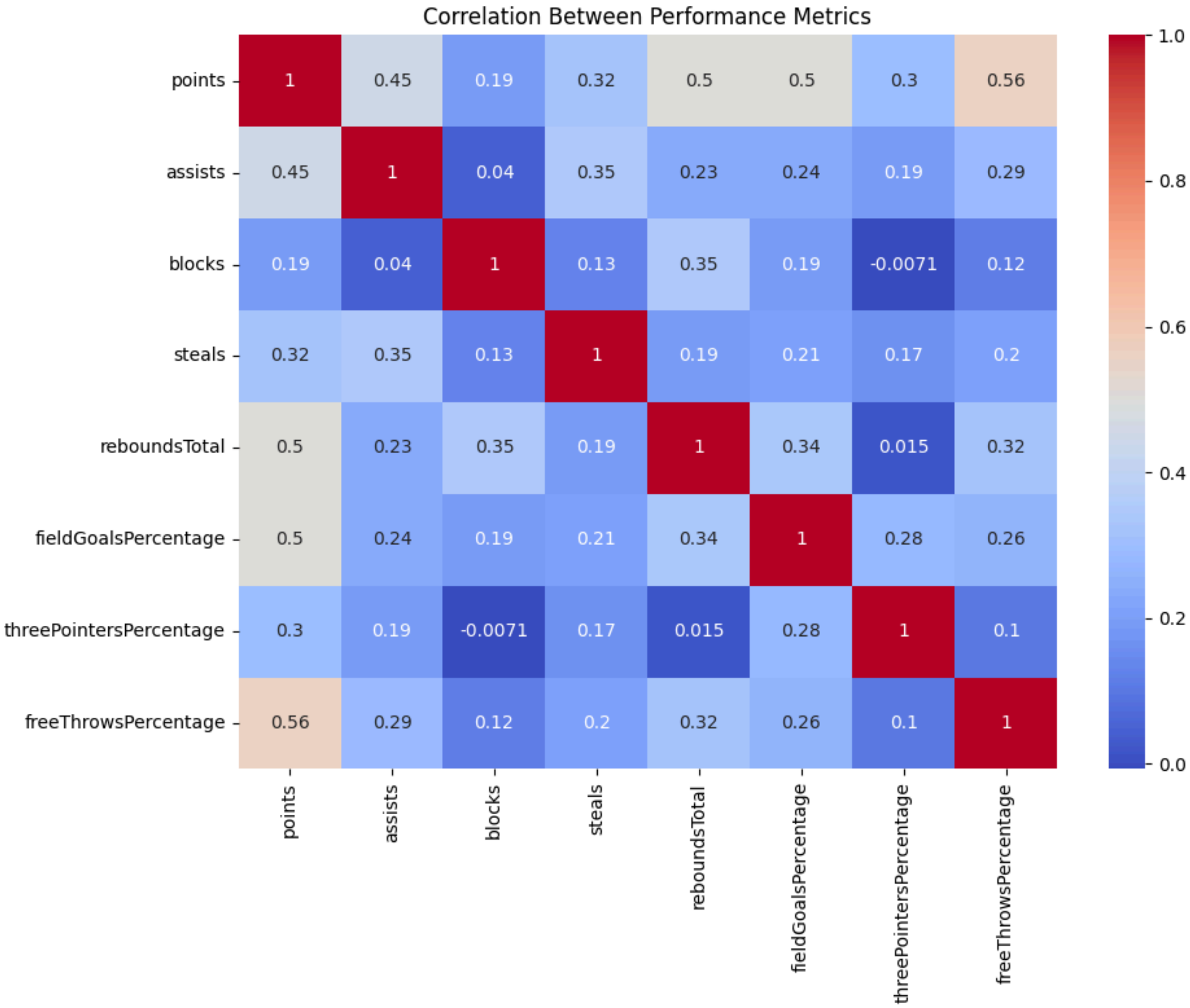
- Clear titles and axis labels for immediate understanding
- Appropriate color schemes to distinguish different metrics and categories
- Varied chart types matched to specific analytical questions
- Size encoding to represent additional dimensions (e.g., points in scatter plots)

2.2 Specific Design Decisions

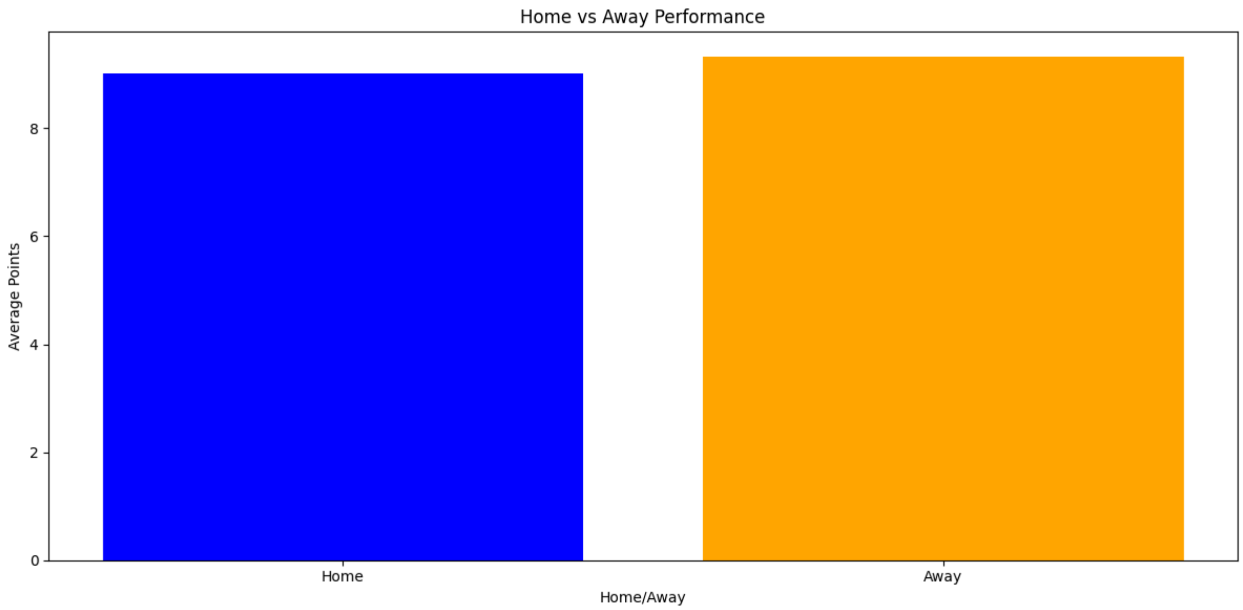
- **Top Scorers**: Bar charts for ranking players by scoring averages



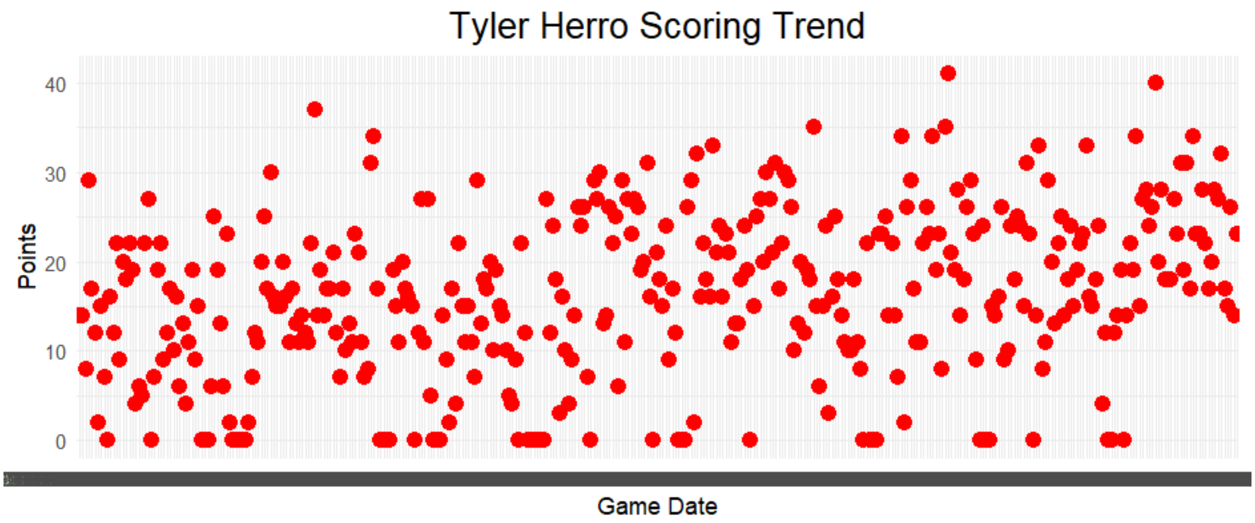
- **Correlation Heatmap**: Color-coded matrix to illustrate relationships between metrics



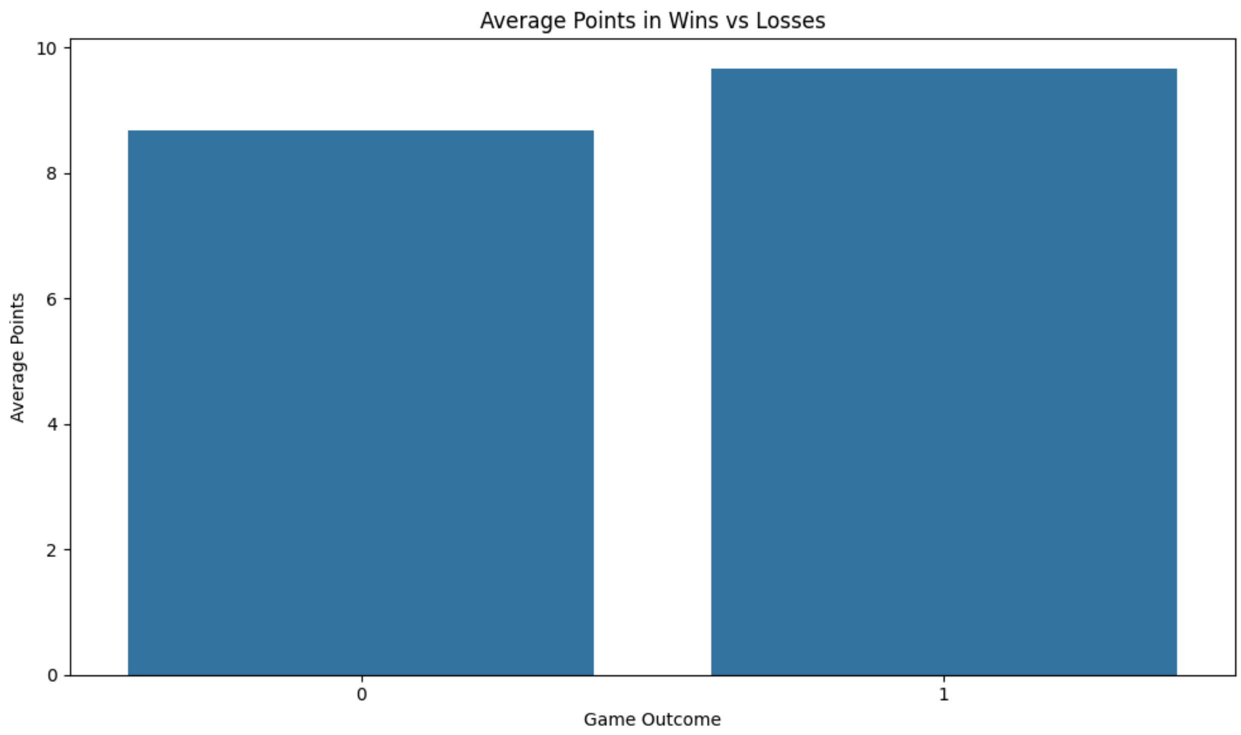
- **Home vs. Away**: Grouped bar charts for comparative performance analysis



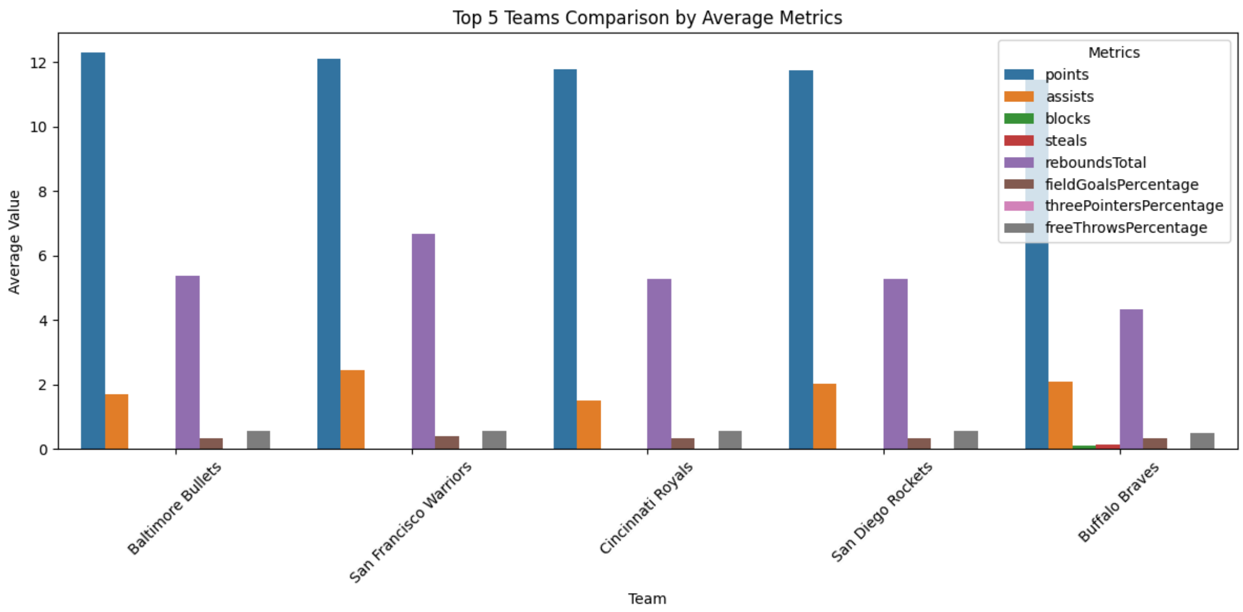
- **Player Trends**: Line charts to track individual player performance over time



- **Wins vs. losses**: Grouped bar charts for comparative performance analysis



- **Team Comparisons**: Grouped bar charts for multi-metric team analysis



3. Implementation

3.1 Tools and Technologies

- **Python**: Pandas for data manipulation, Matplotlib and Seaborn for visualization
- **R**: Statistical analysis with correlation tests, regression models, and hypothesis testing
- **Visualization Libraries**: ggplot2 in R for advanced statistical visualizations

3.2 Statistical Methods

- Correlation analysis between performance metrics
- Linear regression to predict points based on other statistics
- T-tests to compare performance differences in different game contexts

4. Conclusion

The analysis provides valuable insights into NBA player performance patterns, highlighting the relationships between different basketball statistics and game contexts. The visualizations effectively communicate complex performance metrics and comparative analyses. This framework can be extended to track player development, evaluate team strategies, and support data-driven decision-making in basketball analytics.