BIST100 & Spotify Data Analysis Project Report

1. Introduction

In today's data-driven world, uncovering hidden patterns and correlations in diverse datasets provides fascinating insights into both personal and professional realms. This project, conducted as part of the Sabancı University CS210 Introduction to Data Science course, explores the connection between financial market behavior and personal music preferences. By analyzing two disparate datasets—one capturing the weekly historical data of BIST100 (Borsa Istanbul) and the other detailing a year of personal Spotify usage—the project delves into the interplay between investment performance and music listening habits.

Project Motivation

As a budding investor and avid music listener, I have often pondered whether fluctuations in my investment portfolio influence my daily choices, including music preferences. This project stems from that curiosity, seeking to investigate whether market trends and mood, reflected through music listening habits, have an underlying connection. By combining these datasets, I aim to gain a deeper understanding of potential trends and relationships that could offer both personal and analytical insights.

Objectives

The primary goals of this project are:

1. To explore trends within the BIST100 financial dataset and Spotify usage data independently.
2. To identify potential correlations or patterns between stock market behavior and music preferences.
3. To apply statistical and machine learning techniques to assess the predictability of these relationships.
4. To summarize findings in a comprehensive report, providing visual and analytical interpretations.

Datasets

- **BIST100 Dataset**: Weekly historical financial data for one year, sourced via the Yahoo Finance API. This dataset includes key metrics such as opening and closing prices, highs and lows, and trading volumes.
- **Spotify Dataset**: One year of personal music consumption data, exported through Spotify's user data request feature. Key features include track names, artists, listening durations, and timestamps.

By integrating these datasets and leveraging data analysis techniques, this project embarks on a unique journey to bridge financial and behavioral data, offering a novel perspective on the impact of market behavior on personal habits.

2. Data Collection and Preparation

BIST100 Dataset

The BIST100 data was obtained using the Yahoo Finance API. The dataset spans one year, encompassing weekly records of key metrics such as:

- **Open Price**: The price at which the index started trading each week.
- **Close Price**: The final trading price of the week.
- **High and Low Prices**: The highest and lowest prices recorded during the week.
- **Volume**: The total number of shares traded.

These metrics were extracted into a structured format and saved as a CSV file for further analysis. After importing, initial checks were performed to ensure data quality, including handling missing values and converting data types as needed. Rows with incomplete records were dropped, and the "Date" column was set as the index for temporal alignment with the Spotify dataset.

Spotify Dataset

The Spotify data was sourced via the platform’s user data request feature, capturing detailed records of music listening habits over a year. Key features extracted include:

- **Track Name**: The title of the song.
- **Artist Name**: The performing artist.
- **End Time**: The timestamp indicating when a track was played.
- **Listening Duration**: Measured in milliseconds.

This data was converted from JSON format into a Pandas DataFrame, cleaned, and saved as a CSV file. Additional preprocessing steps included:

1. **Timestamp Conversion**: The "End Time" column was converted into datetime format.
2. **Daily Aggregation**: Listening durations were aggregated to calculate total daily listening time (in minutes).
3. **Data Visualization**: Summary statistics and initial plots provided an overview of listening behavior across different time intervals.

Merging the Datasets

To explore potential relationships, the two datasets were merged on their respective "Date" columns. The merged dataset retained only overlapping dates, ensuring alignment between stock

market trends and music listening behavior. This integrated dataset served as the foundation for further analysis, including hypothesis testing and predictive modeling.

Quality Assurance

Data quality checks ensured that:

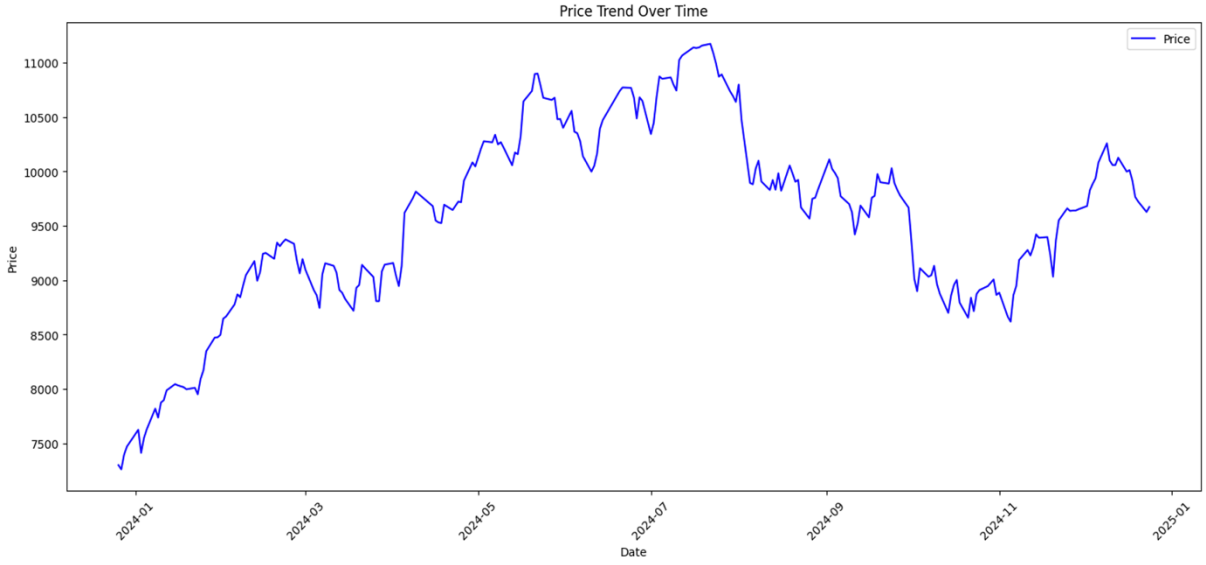
- All entries were free of null values post-cleaning.
- Numerical columns were correctly formatted for analysis.
- The merged dataset covered a meaningful time interval for hypothesis testing and modeling.

3. Exploratory Data Analysis (EDA)

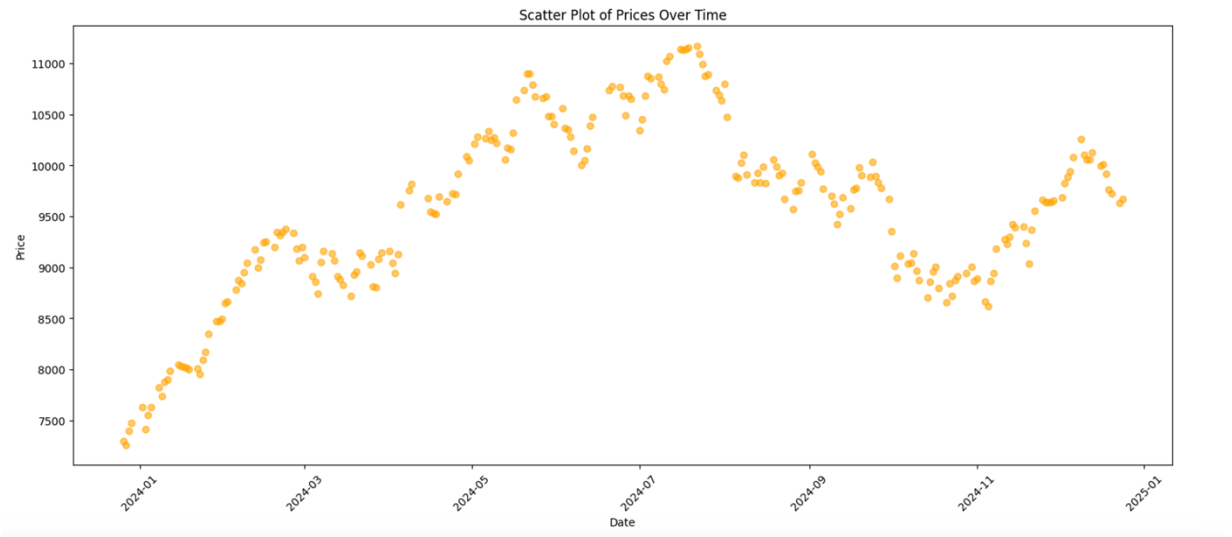
BIST100 Dataset

The BIST100 dataset was explored to understand key trends and characteristics of the financial market. Key visuals included:

- **Line Plot of Price Trends**: A line plot was used to visualize the fluctuations in BIST100 prices over time, highlighting significant upward or downward movements.



- **Scatter Plot of Prices**: A scatter plot revealed the distribution of prices, emphasizing clustering or anomalies.

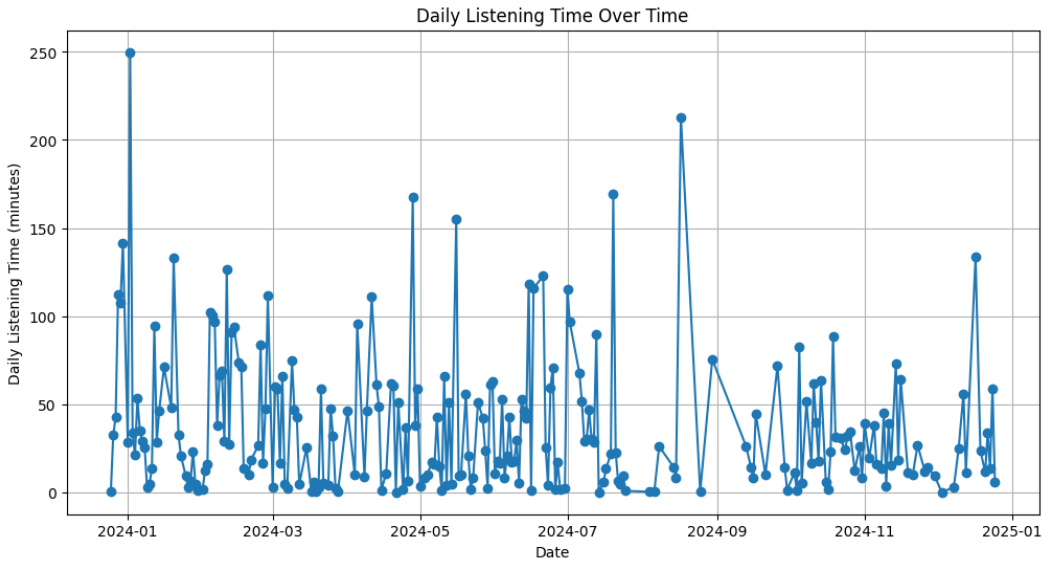


These visualizations offered a clear picture of the financial market's behavior during the one-year period.

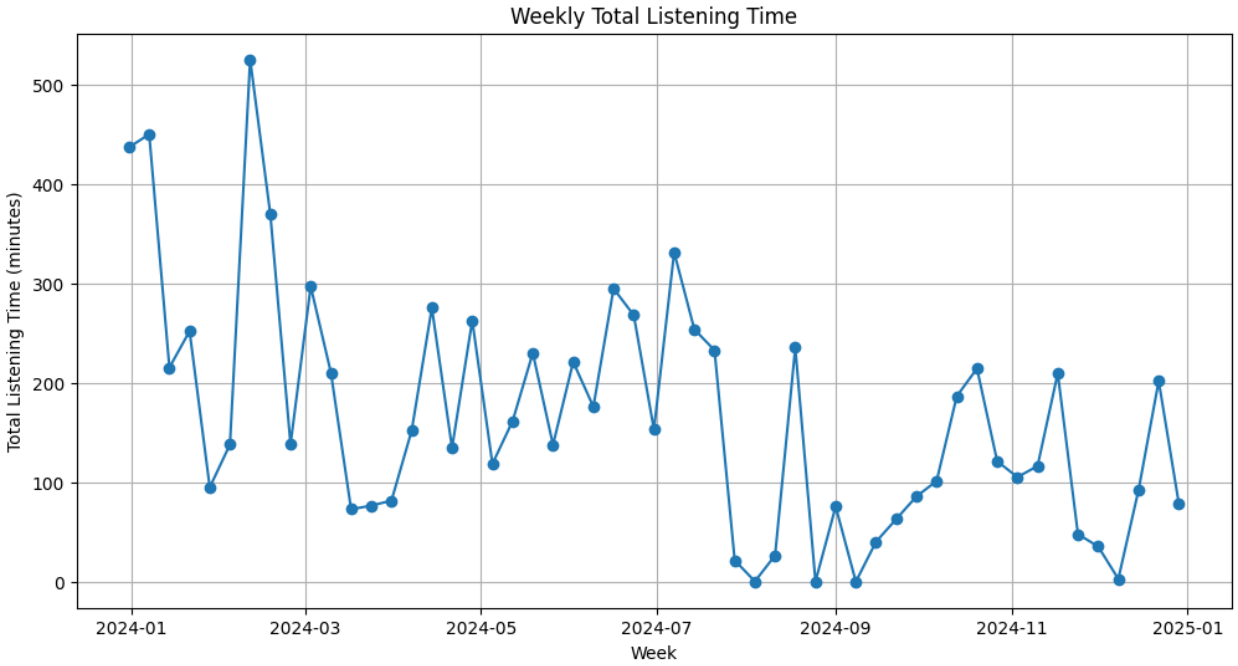
Spotify Dataset

The Spotify dataset was analyzed to uncover patterns in music listening behavior. Key visuals included:

- **Daily Listening Time Over Time**: A line plot illustrated fluctuations in daily listening time, showcasing peaks and troughs that may align with significant events or personal routines.



- **Rolling Average of Daily Listening Time**: A 7-day rolling average smoothed out short-term variations, revealing overall trends.

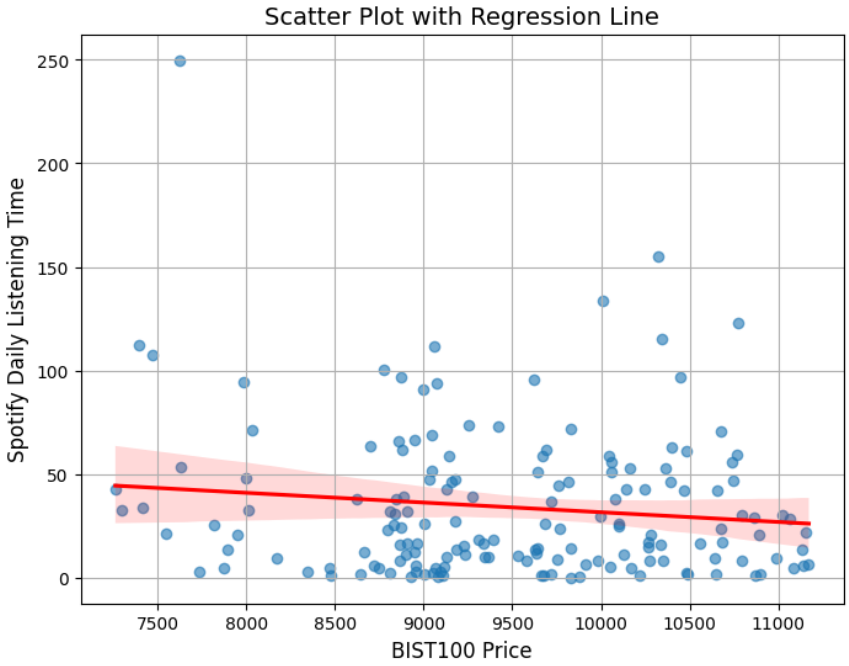


These plots helped to understand the user's music consumption habits across different time scales.

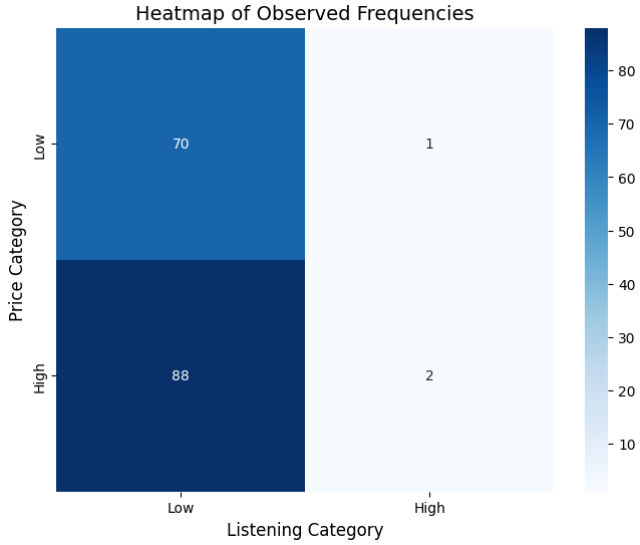
Combined Insights

Using the merged dataset, combined insights were derived through:

- **Scatter Plot with Regression Line**: This plot examined the relationship between BIST100 prices and daily listening time, providing a visual representation of their correlation.



- **Heatmap of Observed Frequencies**: A heatmap captured the interaction between price categories and listening time categories, aiding categorical comparison.



The EDA phase laid the groundwork for deeper statistical analysis and modeling by identifying key trends and potential relationships within and across datasets.

4. Hypothesis Testing

Hypotheses

To investigate the potential relationship between BIST100 prices and Spotify listening behavior, the following hypotheses were formulated:

- **H_0 (Null Hypothesis):** There is no correlation between BIST100 prices and Spotify listening time.
- **H_1 (Alternative Hypothesis):** There is a correlation between BIST100 prices and Spotify listening time.

Statistical Tests and Results

1. Pearson Correlation

A Pearson correlation test was conducted to measure the strength and direction of the relationship between the two variables. The results indicated:

- **Correlation Coefficient:** -0.124 (weak negative correlation)
- **P-value:** 0.117 (greater than 0.05, indicating no statistical significance)

This suggests that there is no meaningful linear relationship between BIST100 prices and daily Spotify listening time.

2. T-Test

A T-test was performed to compare BIST100 prices between groups with high and low Spotify listening times (categorized based on the median). The results showed:

- **T-statistic:** -0.971
- **P-value:** 0.333 (greater than 0.05, failing to reject the null hypothesis)

The analysis concluded no significant difference in BIST100 prices between high and low listening groups.

3. Chi-Square Test

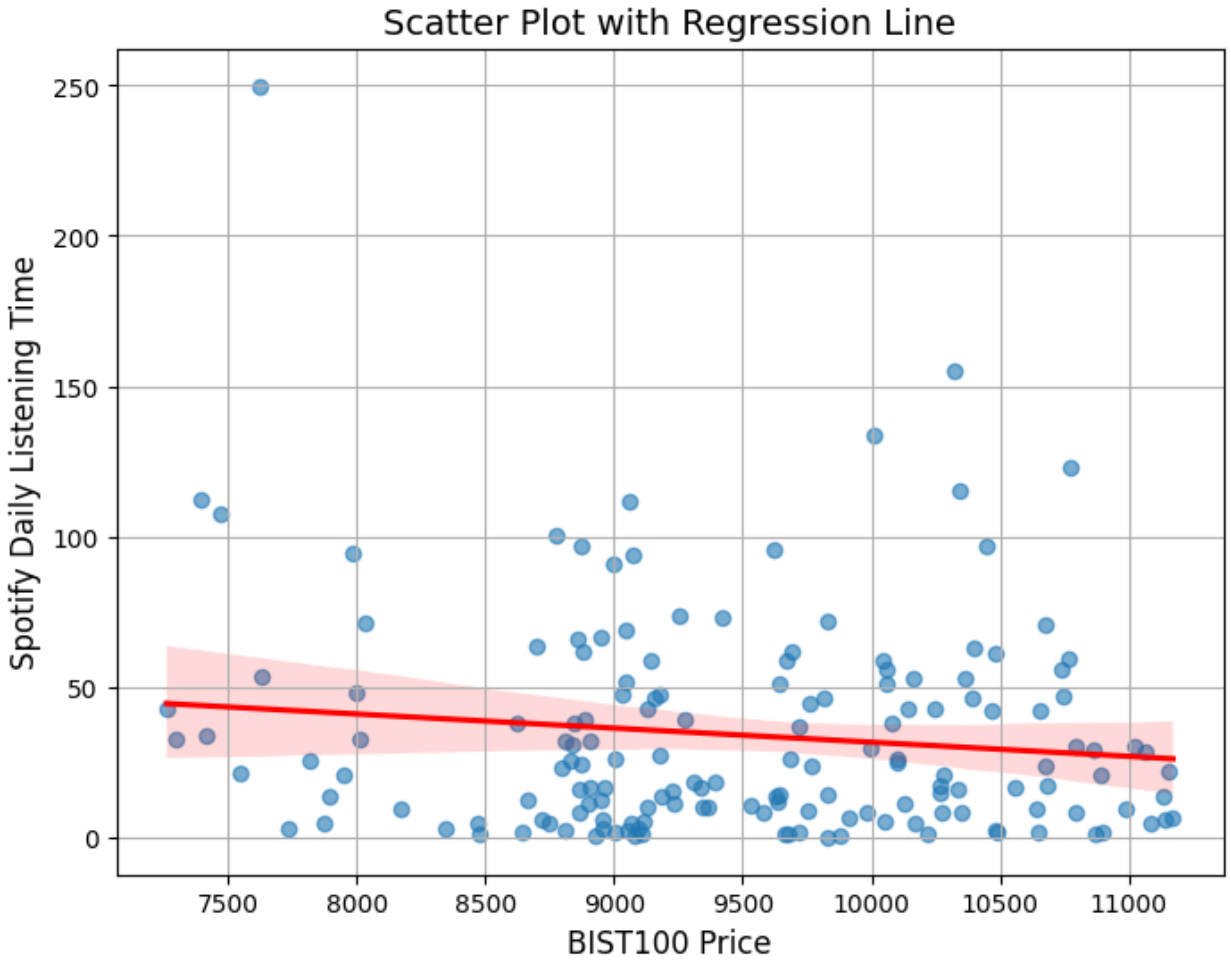
The Chi-Square test examined the association between categorical variables (price and listening time categories). The results were:

- **Chi-Square Statistic:** 0.0
- **P-value:** 1.0

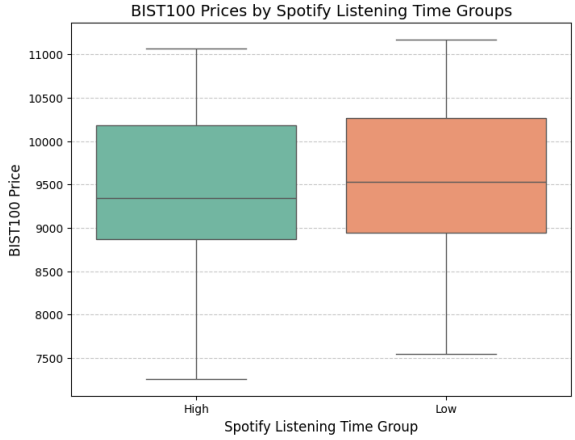
This indicates no association between the two variables, as their categorical distributions are independent.

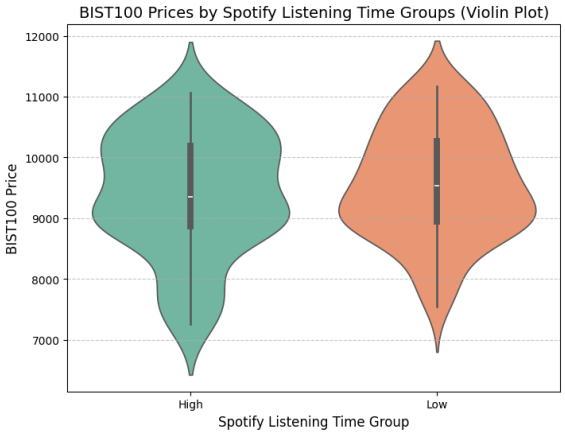
Visual Representations

- **Scatter Plot with Regression Line:** Illustrated the weak and non-significant relationship between the variables.

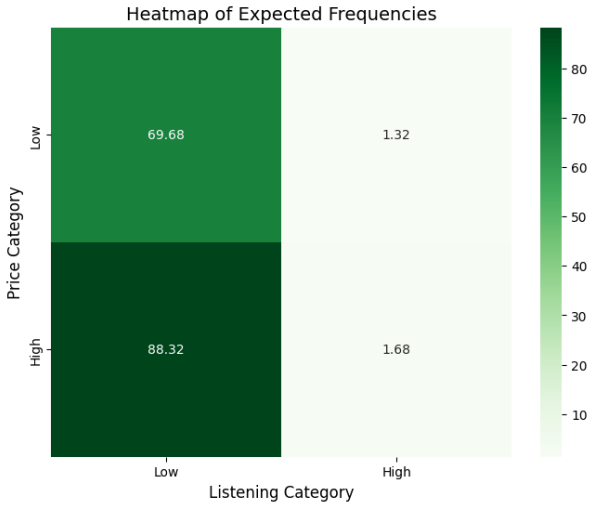


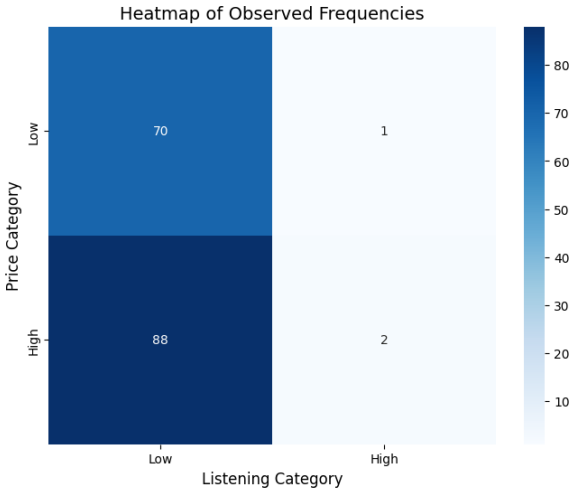
- **Box Plot and Violin Plot:** Highlighted the distribution of BIST100 prices across listening time groups.





- **Heatmaps:** Showed observed and expected frequencies of categorical variables, confirming independence.





Conclusion

The hypothesis testing phase revealed no statistically significant relationship or difference between BIST100 prices and Spotify listening behavior. These findings suggest that external factors may play a larger role in influencing these variables independently.

5. Predictive Modeling

Linear Regression

A linear regression model was applied to explore the relationship between daily listening time and BIST100 prices. The model's performance metrics included:

- **Intercept:** 9650.64
- **Coefficient:** -4.84
- **Mean Squared Error (MSE):** 984,482.20
- **R-squared:** -0.071

The negative R-squared and high MSE suggest that daily listening time is not a meaningful predictor of BIST100 prices. The linear model failed to capture any significant patterns between the variables.

Decision Tree Regression

A decision tree regression model was implemented to predict BIST100 prices based on daily listening time. Results showed:

- **Mean Squared Error (MSE):** 960,549.23
- **R-squared:** -0.045

The tree structure revealed no significant splits that could explain variability in BIST100 prices using listening time, reinforcing the weak predictive relationship.

Random Forest Regression

The random forest regression model aimed to enhance predictive accuracy by combining multiple decision trees. Results included:

- **Mean Squared Error (MSE):** 1,159,505.06
- **R-squared:** -0.261

The random forest model also underperformed, indicating that listening time data alone does not provide enough information to predict stock prices.

Conclusion

Predictive modeling efforts across linear regression, decision tree regression, and random forest regression consistently demonstrated poor performance, with high error metrics and negative R-squared values. These results suggest that the relationship between BIST100 prices and Spotify listening time is either non-existent or influenced by external, unaccounted factors

6. Conclusion

This project sought to bridge the gap between financial market behavior and personal music preferences by analyzing BIST100 and Spotify datasets. Despite initial hypotheses suggesting a potential relationship, the analyses revealed no significant correlations or predictive links between the two variables. This outcome underscores the importance of hypothesis validation and the challenges of finding meaningful connections across disparate data domains.

Through comprehensive exploratory analysis, hypothesis testing, and predictive modeling, the study demonstrated the value of rigorous data examination. The lack of significant findings does not diminish the project's contributions but rather highlights the need for multidimensional approaches and the inclusion of diverse variables to uncover nuanced patterns. Future studies incorporating broader datasets and alternative methodologies may yield richer insights into the complex interplay of behavioral and financial data.