Modeling and processing of resources

Assignment One

Researcher:

Mohammad Rezaei Kalantary

401422087

Teacher:

Dr Saeidreza Kheradpishe

Shahid Beheshti University
Mathematics Department
Fall 2023

Contents

Abstract

This study analyses two datasets, "US Accidents" and "Top Songs on Spotify," using both descriptive and inferential statistics. The investigation reveals some trends in American music tastes and traffic accidents, offering insights into both current music trends and traffic safety.

Introduction

This study uses a strong combination of descriptive and inferential statistics to conduct a thorough analysis of two different datasets: "US Accidents" and "Top Songs on Spotify." Our goal is to use these datasets to extract insightful information, recognise patterns, and make significant inferences.

We completed a thorough phase of exploratory data analysis (EDA) and data pretreatment before beginning the statistical analysis. This was a critical step in guaranteeing the accuracy, integrity, and usefulness of the data for our study. We were able to successfully manage missing values, identify outliers, and visualise the distributions of important variables thanks to EDA.

For every dataset, we have developed a set of five to six general questions that will direct our statistical research. Our investigation is based on these questions, which allow us to investigate a variety of factors and aspects. Our inquiries for the "US Accidents" dataset range from figuring out accident trends and variables impacting traffic safety to estimating the seriousness of each event. Conversely, our inquiries concerning the "Top Songs on Spotify" dataset centre on identifying prevailing music trends, evaluating the impact of diverse music attributes, and forecasting song popularity.

The significance of this research comes from its twin focus on two different but equally important facets of American culture today. We hope to offer useful insights for experts in the music industry as well as road safety management by utilising descriptive and inferential statistics. Those who are interested in learning more about general society trends and preferences may also find this study useful.

Our analysis aims to provide a comprehensive understanding of these datasets, illuminating the variables influencing traffic safety and the dynamic landscape of Spotify users' musical preferences. As we proceed, we will examine each dataset, responding to the questions we posed, and coming to insightful findings that advance our understanding of these crucial aspects of contemporary American life.

1 Methodology

An overview of the statistical procedures and tests used to examine the accident dataset is given in this section. These tests are essential for examining patterns, correlations, and relationships in the data.

1.1 T-Test

A statistical technique called a **t-test** is employed to compare the means of two groups and ascertain whether there are any noteworthy differences between them. We evaluated the relationship between visibility (measured in miles) and accident severity within the framework of our analysis using the t-test. We assessed whether there were statistically significant variations in accident severity according to visibility levels by comparing the means of severity between various visibility conditions. The t-test offers important information about how visibility affects the results of accidents.

1.2 Kruskal-Wallis Test

A non-parametric test for comparing the medians of three or more groups is the **Kruskal-Wallis test**. The Kruskal-Wallis test was used in our analysis to determine whether any counties or geographic areas were linked to more severe accidents. We assessed whether there were statistically significant variations in accident severity throughout these regions by comparing the median accident severity among various counties. When working with non-normally distributed data or comparing several groups at once, the Kruskal-Wallis test is helpful.

1.3 Chi-Squared Test

A statistical test called the **Chi-squared test** is used to ascertain whether two categorical variables significantly correlate with one another. The Chi-squared test was employed in our analysis to investigate the temporal distribution of weather conditions during accidents. We assessed whether there was a statistically significant variation in the distribution of weather conditions between years by comparing the observed distribution of weather conditions with the expected distribution. An effective method for determining trends and evaluating the independence of categorical variables is the Chi-squared test.

1.4 Linear Regression Analysis

Linear regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. In our analysis, we employed linear regression to examine trends in accident severity over time. By fitting a linear regression model to the data and analyzing the p-value associated with the slope of the model, we determined whether there was a significant trend in accident severity over the years. Linear regression is a powerful tool for assessing trends and associations in continuous data.

Investigating and comprehending the trends and contributing variables in the accident dataset requires the use of statistical tests and methods. They offer a strict framework for deriving insightful conclusions and making deliberations based on facts.

2 Top songs on spotify

2.1 Data Overview

With 9,999 rows and 35 columns, the ”Top Songs on Spotify” dataset offers a wealth of data on well-known songs. The structure and content of the dataset can be quickly examined:

- There are 35 columns in total, each representing a specific attribute related to the songs.
- The dataset contains a mix of data types, including object, integer, float, boolean, and datetime.
- Notably, several columns contain missing values, and some columns, such as ”Album Genres,” are entirely devoid of data.

2.2 Data Cleaning

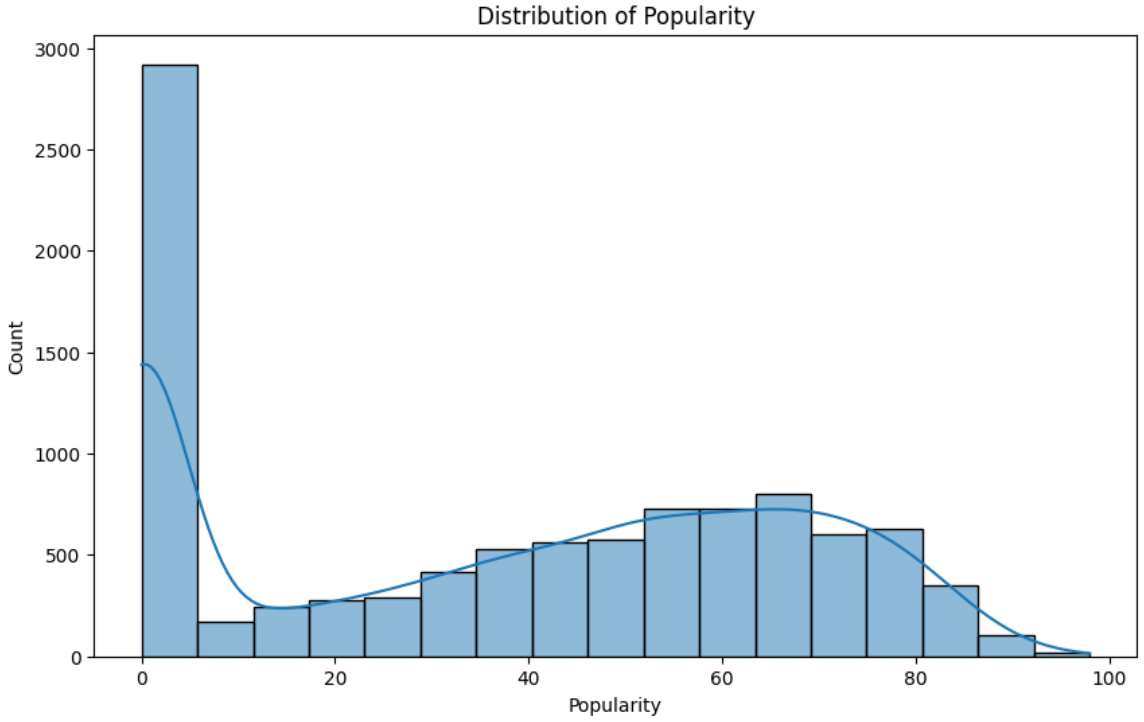
To ensure data quality and consistency, we performed the following data cleaning steps:

1. Duplicate Rows: We identified and removed 48 duplicate rows from the dataset.
2. Missing Values: We addressed missing values in various columns, either by imputing them or by excluding rows with missing critical data.

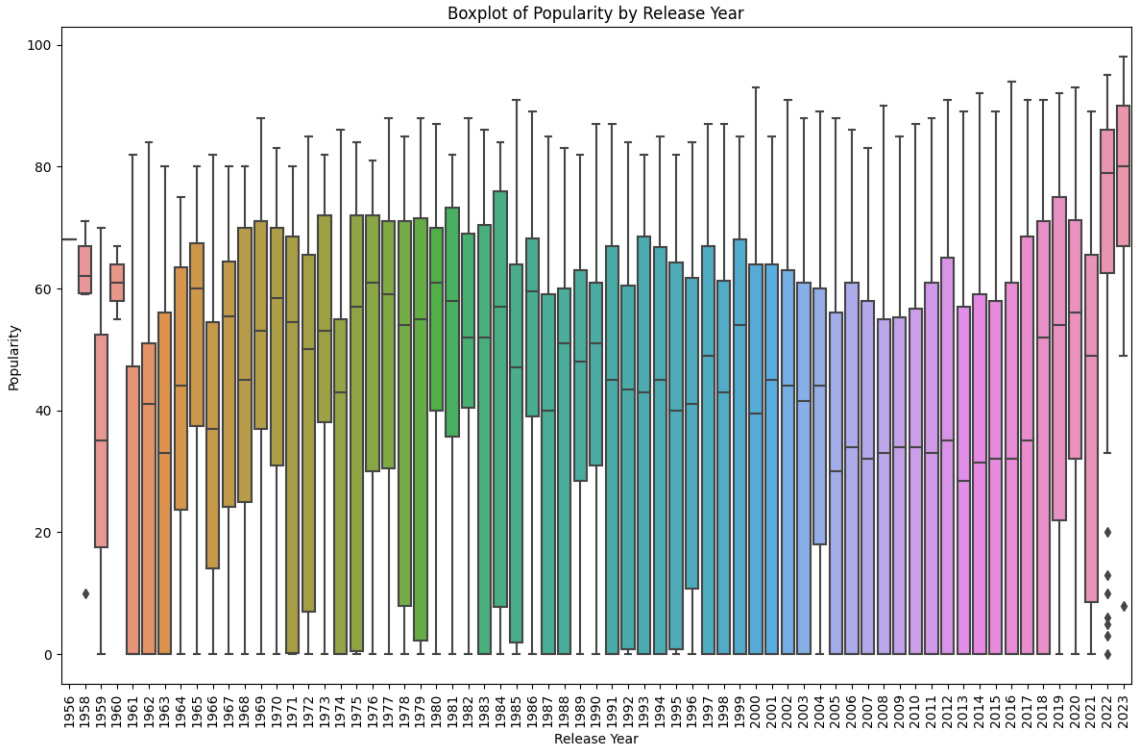
2.3 Data Exploration

Our exploration of the dataset encompassed several key aspects:

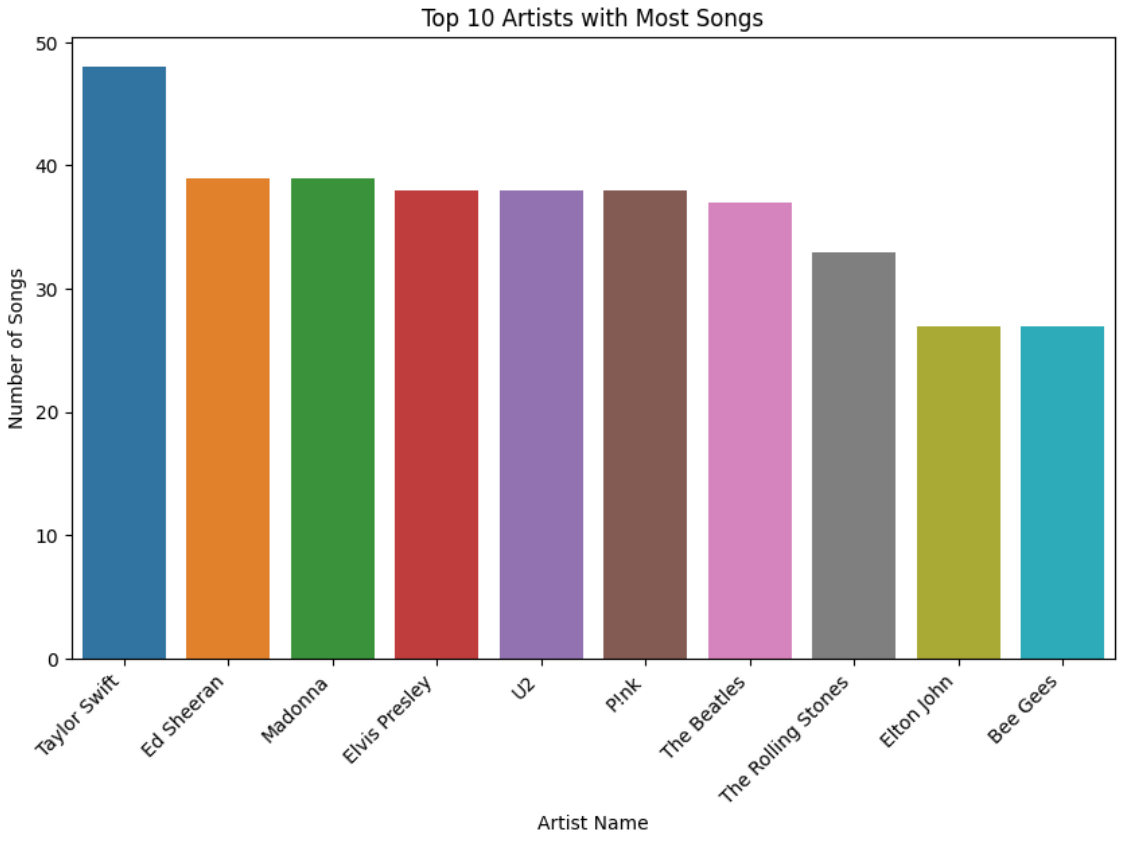
1. **Popularity Distribution:** We visualized the distribution of song popularity, a vital metric for understanding the popularity of tracks in the dataset.



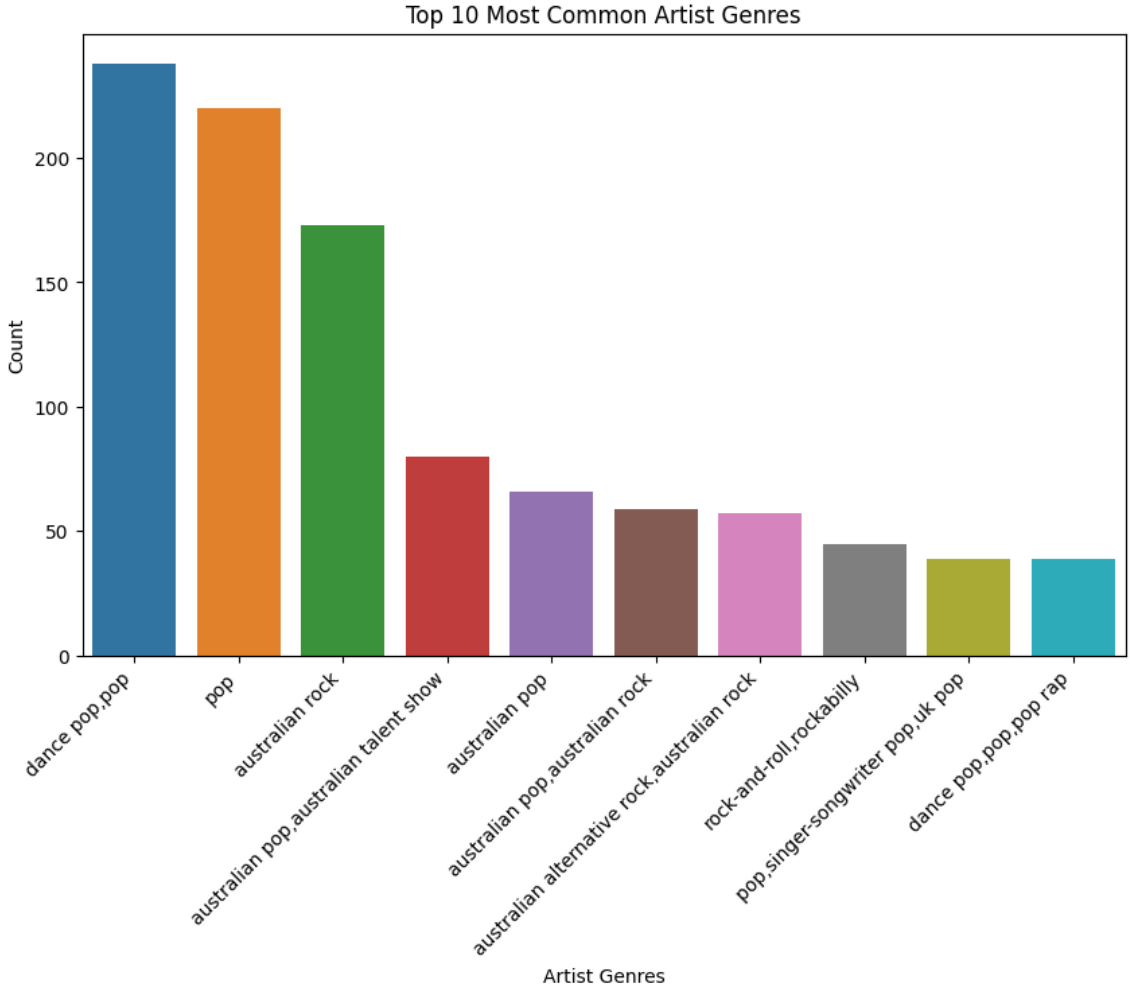
2. **Release Year and Popularity:**We examined the relationship between the release year of songs and their popularity. This analysis helped us uncover trends in song popularity over time.



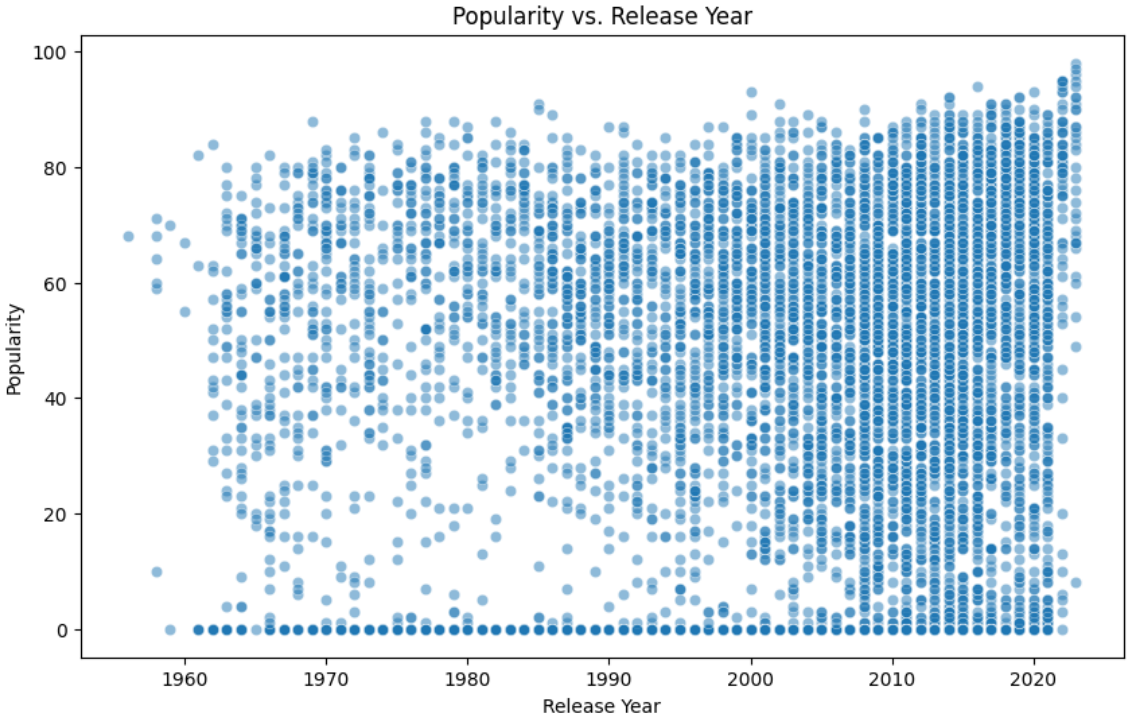
3. **Top Artists:**We identified the top 10 artists with the most songs in the dataset, shedding light on the most prolific contributors to the Spotify platform.



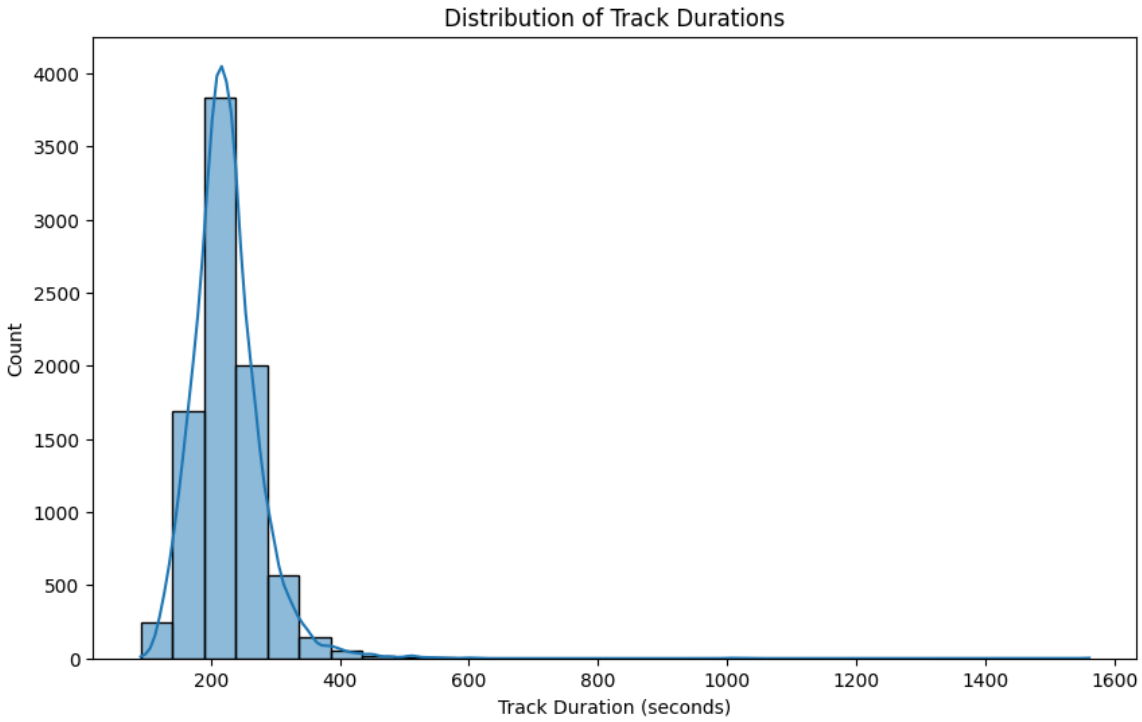
4. **Artist Genres:**We explored the most common artist genres present in the dataset, revealing the musical diversity encompassed in the collection.



5. **Popularity vs. Release Year:**A scatterplot helped us visualize how song popularity varies with release year, providing insights into the evolving tastes of Spotify users over time.



6. **Track Durations:**We examined the distribution of track durations, allowing us to understand the typical length of popular songs in the dataset.

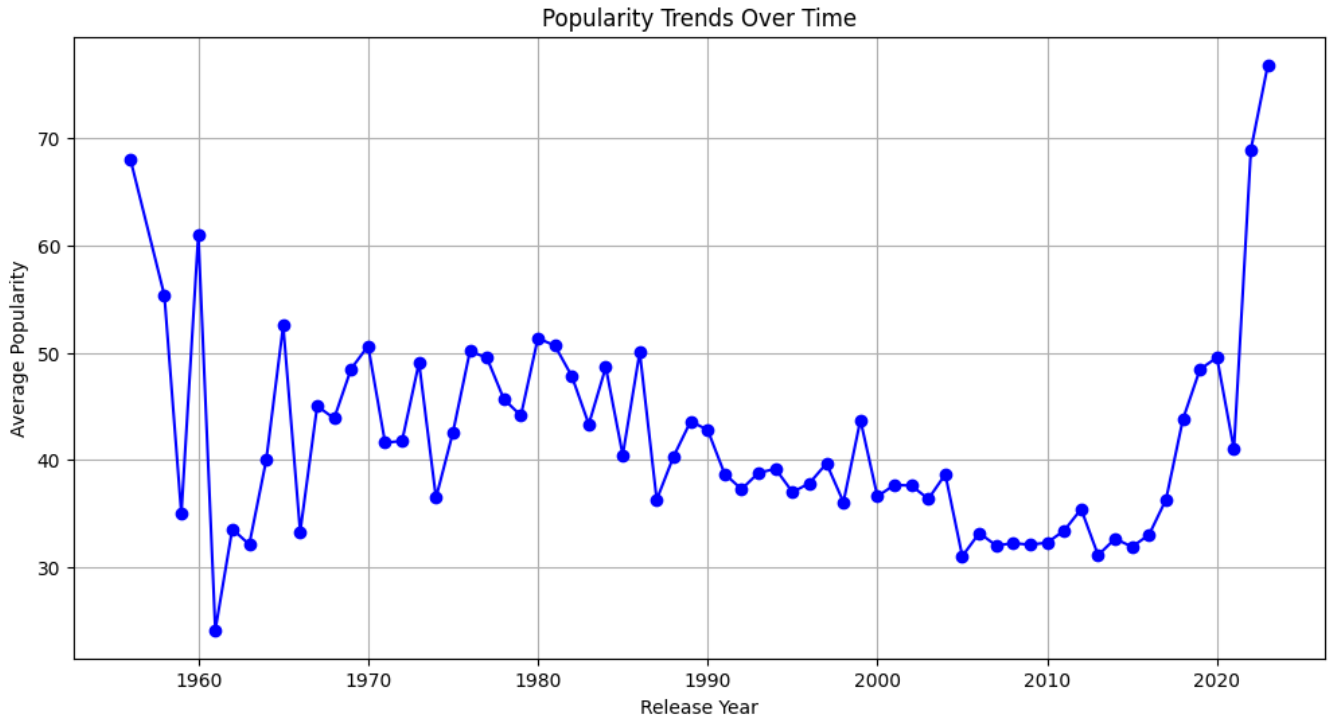


2.4 descriptive statistics and inferential statistics

2.4.1 Analysis 1: Popularity Trends Over Time

Question: Are there significant trends in song popularity (Popularity) over time?

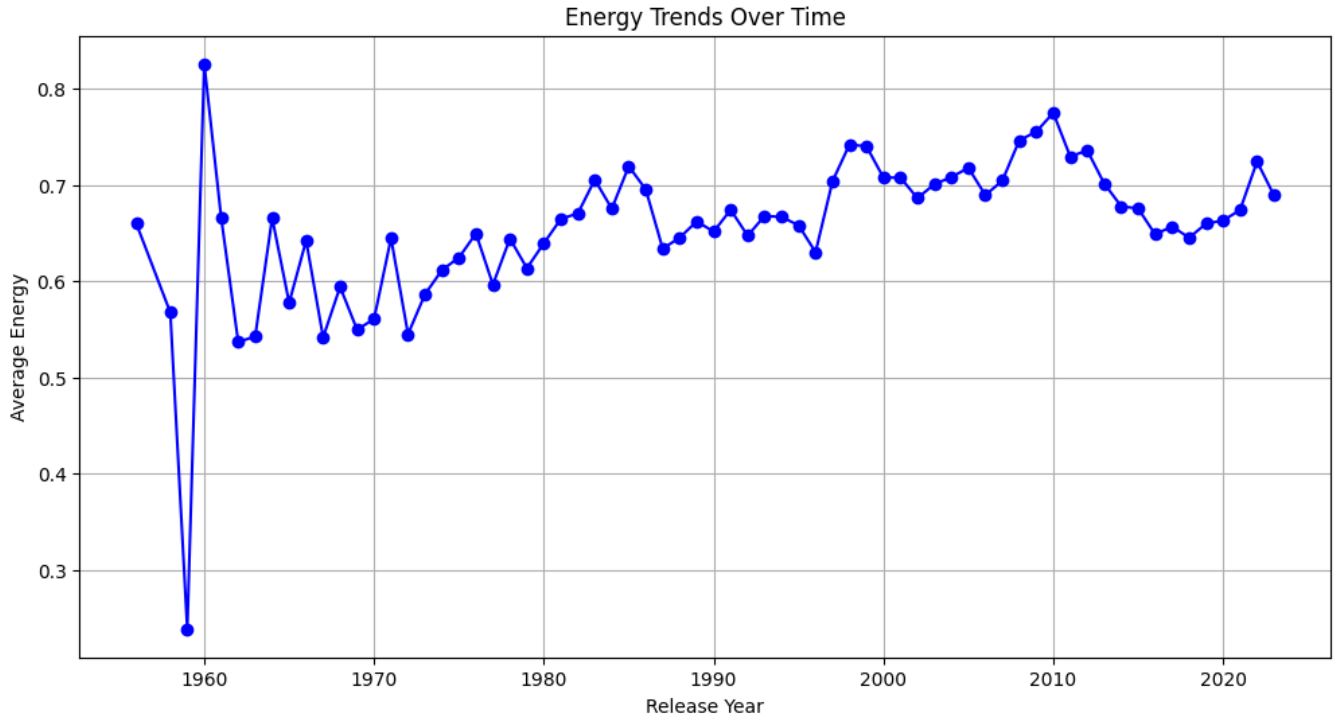
In this analysis, we look into whether song popularity has shown any notable trends over time. We do this by calculating the mean popularity for each year by grouping the data according to the year of release. The popularity trends are visualised using a line plot. In order to assess the statistical significance of the trend, we also run a linear regression analysis and look at the p-value connected to the regression model's slope. The findings suggest that there isn't a clear pattern in the popularity of songs over time.



2.4.2 Analysis 2: Energy Trends Over Time

Question: Are there noticeable trends in the energy levels of songs over time?

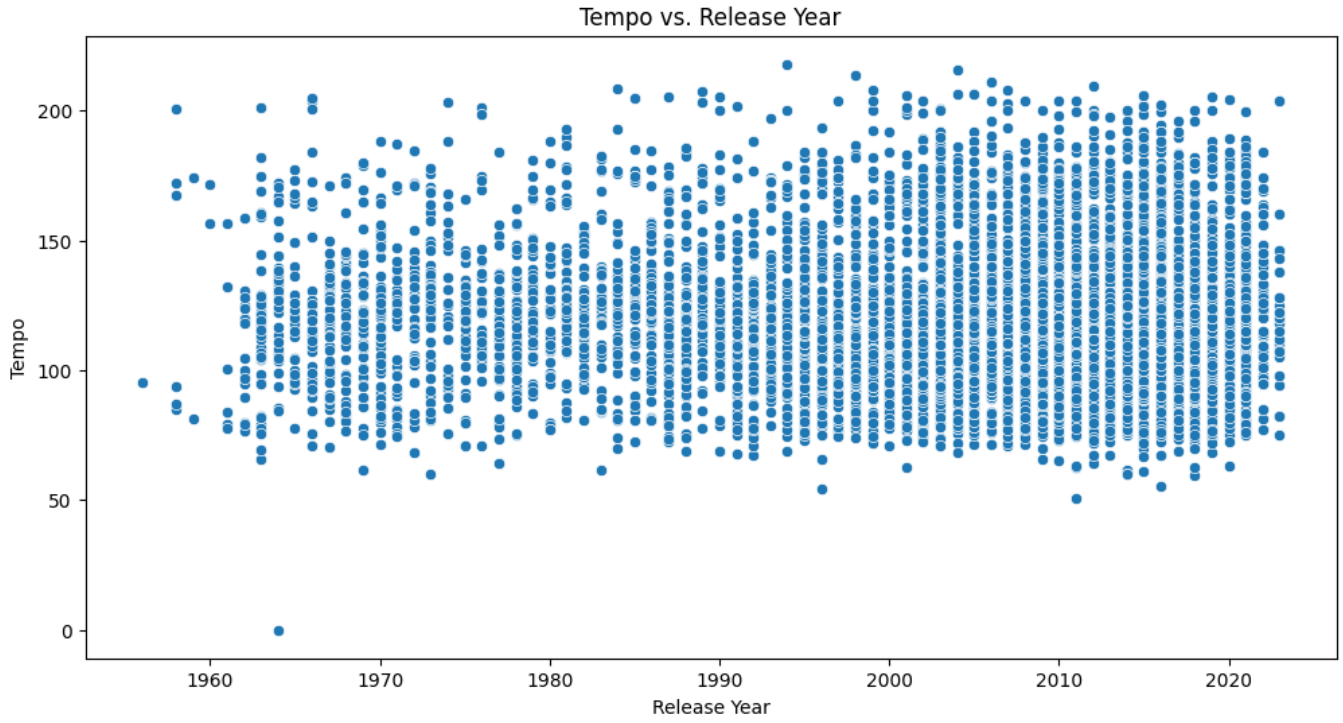
The goal of this analysis is to comprehend how song energy levels have changed over time. As in the previous analysis, we compute the mean energy for each year by grouping the data according to the year of release. To see the energy trends over time, a line plot is created. A linear regression model is utilised to evaluate the trend's significance. In this instance, the data show a noteworthy trend in energy levels over time.



2.4.3 Analysis 3: Tempo vs. Release Year

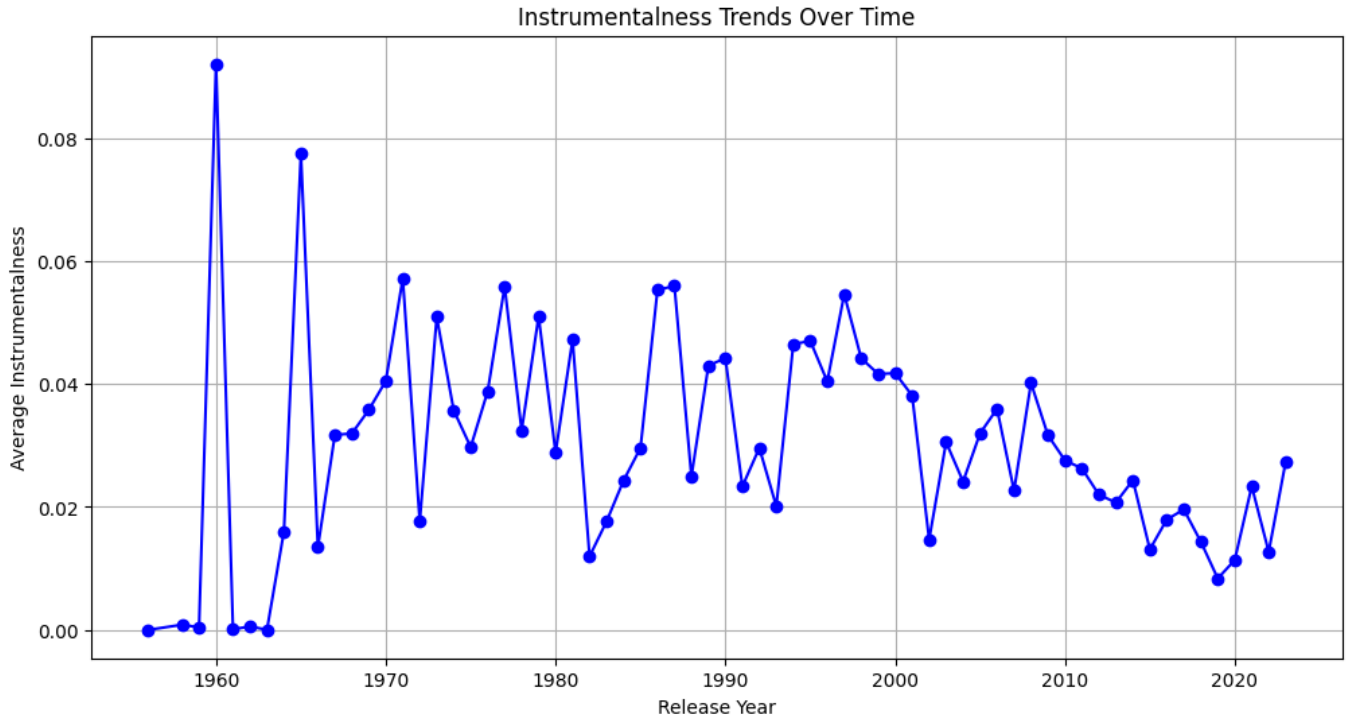
Question: Is there a correlation between the tempo of songs and their release year?

Here, we investigate any possible relationship between a song's tempo and the year it was released. To see the connection between release year and tempo, a scatter plot is made. We compute the correlation coefficient and run a statistical test to see if the correlation is statistically significant in order to quantify this relationship. The analysis comes to the conclusion that the relationship between release year and tempo is not statistically significant.



2.4.4 Analysis 4: Instrumentalness Trends Over Time

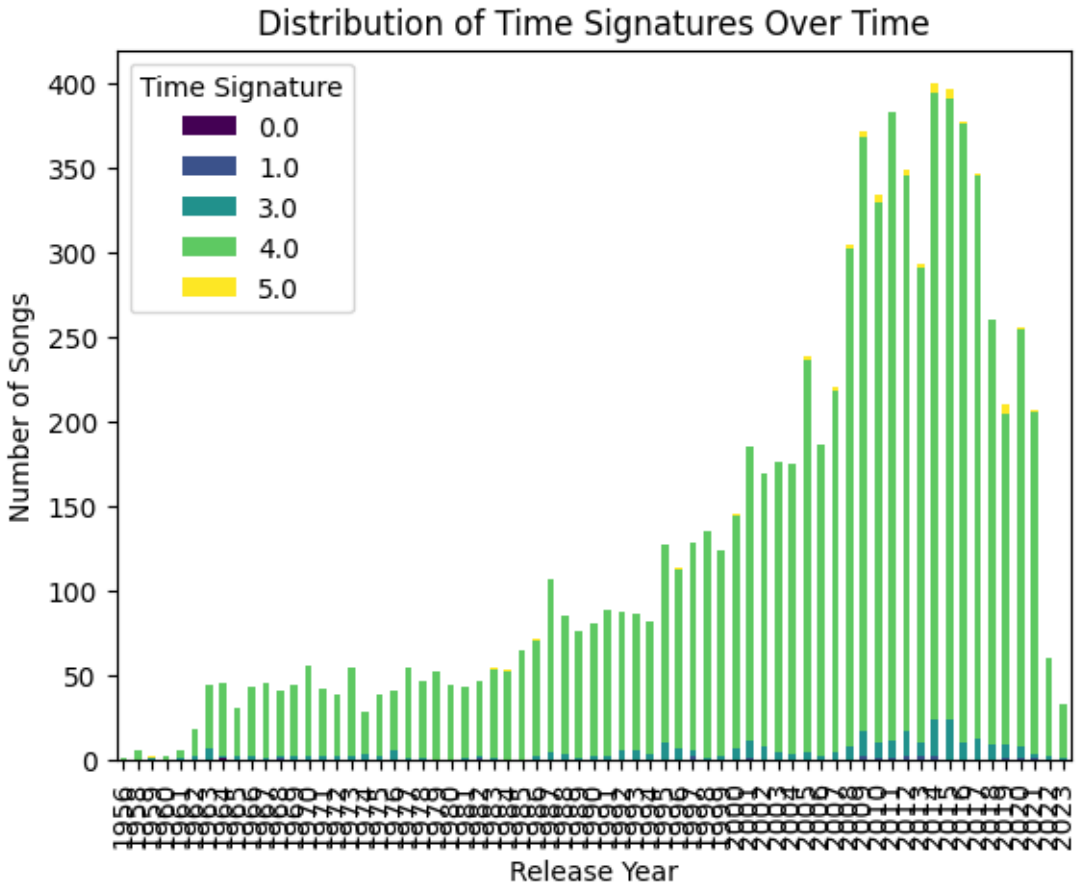
Question: How has the instrumentalness of songs evolved over time?
This analysis looks at how songs have changed in instrumentality over time. We compute the mean instrumentalness for each year by classifying the data according to the year of release. To see the trend of instrumentalness over time, a line plot is utilised. To determine if the trend has statistical significance, a linear regression model is utilised. The findings show that instrumentalness has not shown any noticeable trend over time.



2.4.5 Analysis 5: Time Signature Distribution Over Time

Question: Are certain time signatures more prevalent in recent music compared to earlier decades?

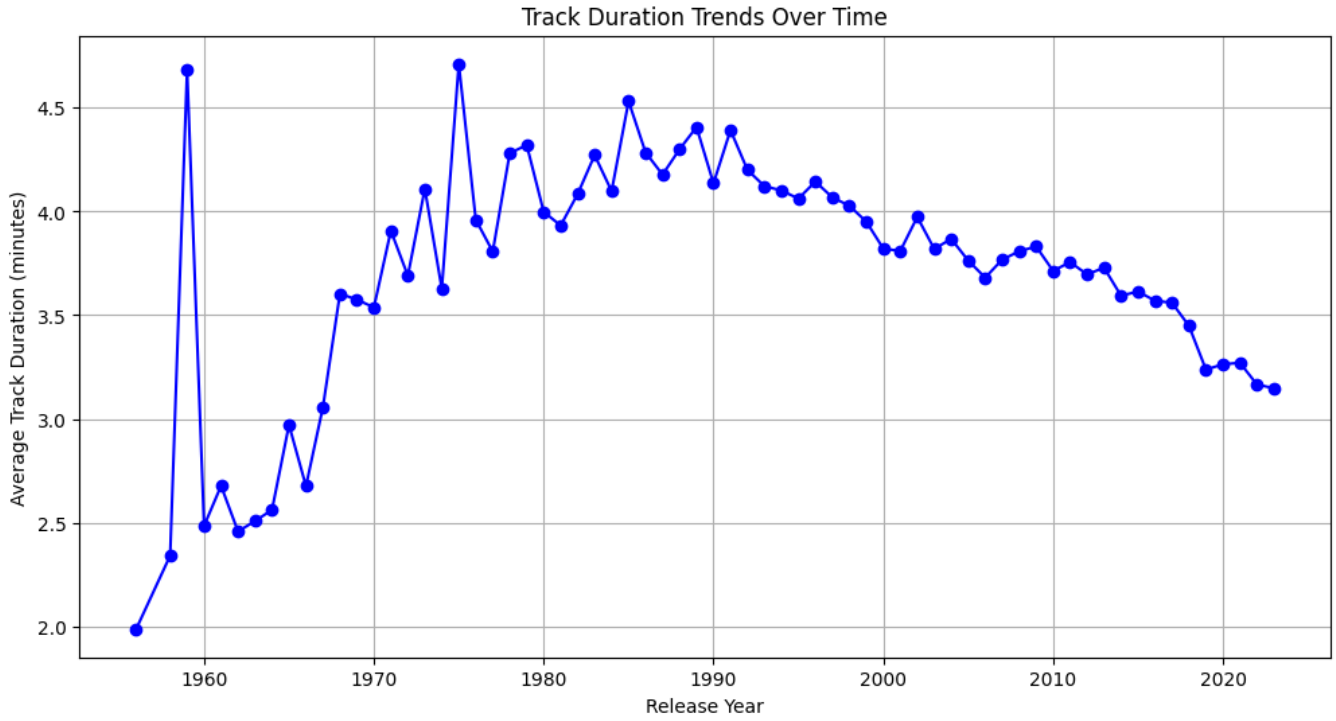
In this analysis, we look into whether recent music uses a particular time signature more often than music from previous decades. We count the occurrences of each time signature after classifying the data by release year. To compare the distribution of time signatures over time, a stacked bar chart is utilised. In addition, a chi-squared test is run to see if the time signature distribution changes over time in a way that is statistically significant. A statistically significant variation in the time signature distribution over time is revealed by the analysis.



2.4.6 Analysis 6: Track Duration Trends Over Time

Question: Is there a trend in the duration (Track Duration) of songs over time?

This analysis explores the historical trend of song durations. We compute the mean track duration for each year by dividing the data according to the year of release. For easier reading, the track duration is also converted from milliseconds to minutes. To see the trend of track duration over time, a line plot is made. The statistical significance of the trend is tested using a linear regression model, in line with earlier studies, and the results indicate that there is no noticeable pattern in track duration over time.



3 US Accidents

3.1 Data Overview

The ”US Accidents” dataset consists of 7,728,394 rows and 46 columns, offering a comprehensive view of traffic accidents in the United States. An initial examination of the dataset reveals its structure and content:

- There are 46 columns in total, each representing various attributes related to traffic accidents.
- The dataset encompasses a range of data types, including object, integer, float, boolean, and datetime.
- Some columns contain missing values, necessitating data cleaning procedures.

3.2 Data Cleaning

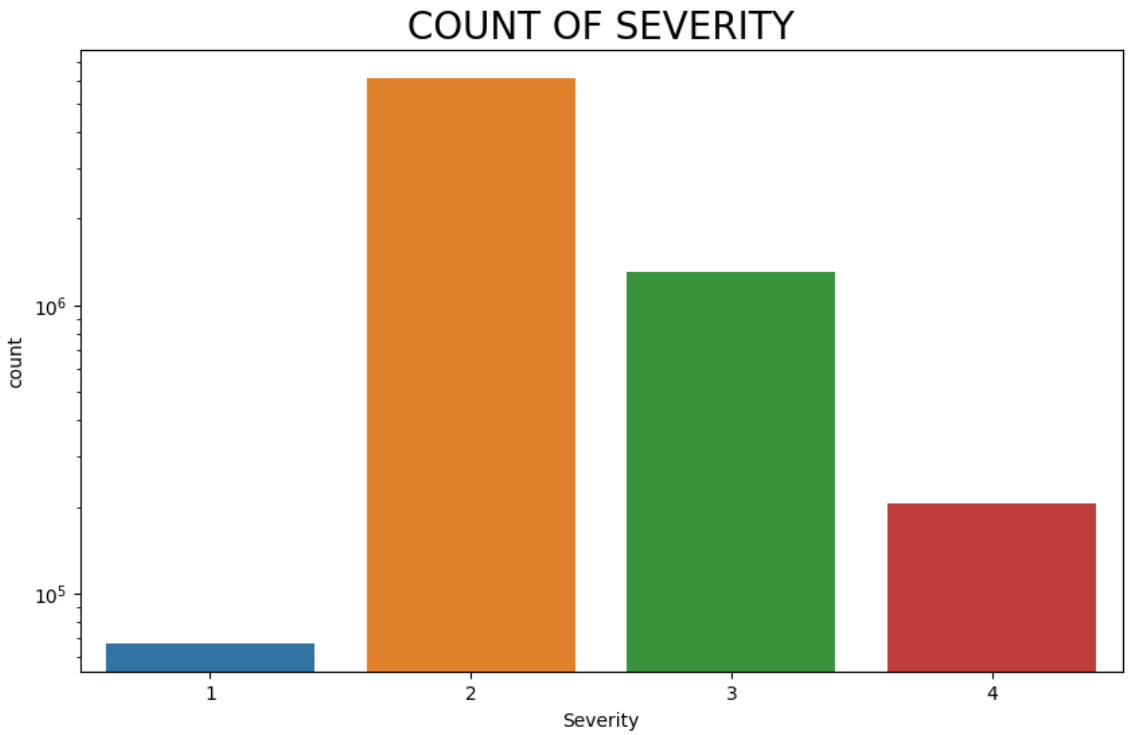
To ensure data quality and consistency, we performed the following data cleaning steps:

1. Duplicate Rows: No duplicate rows were identified in the dataset.
2. Missing Values: We addressed missing values in various columns, either by imputing them or by excluding rows with missing critical data.

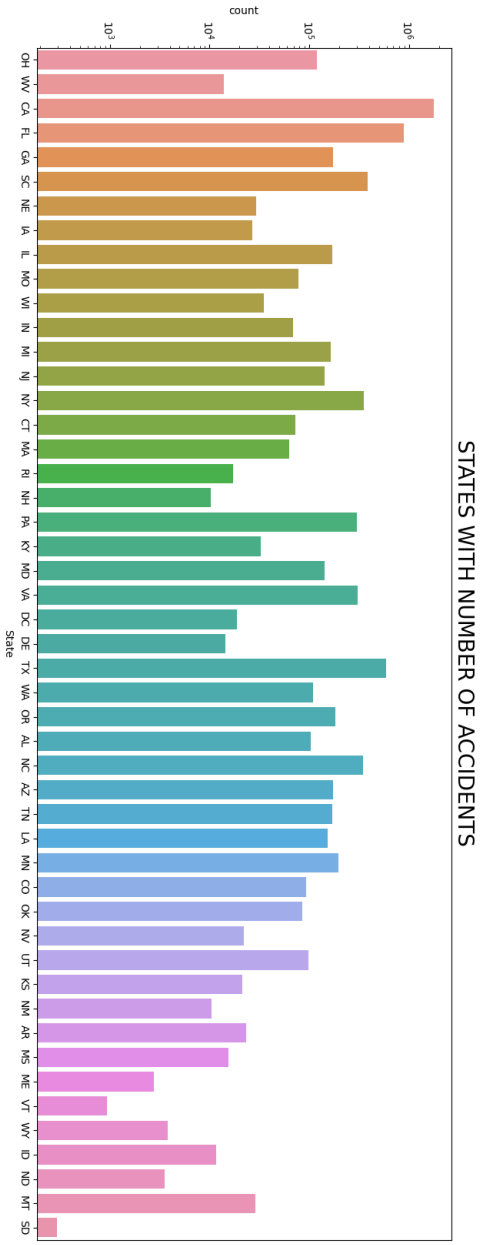
3.3 Data Exploration

Our exploration of the dataset encompassed several key aspects:

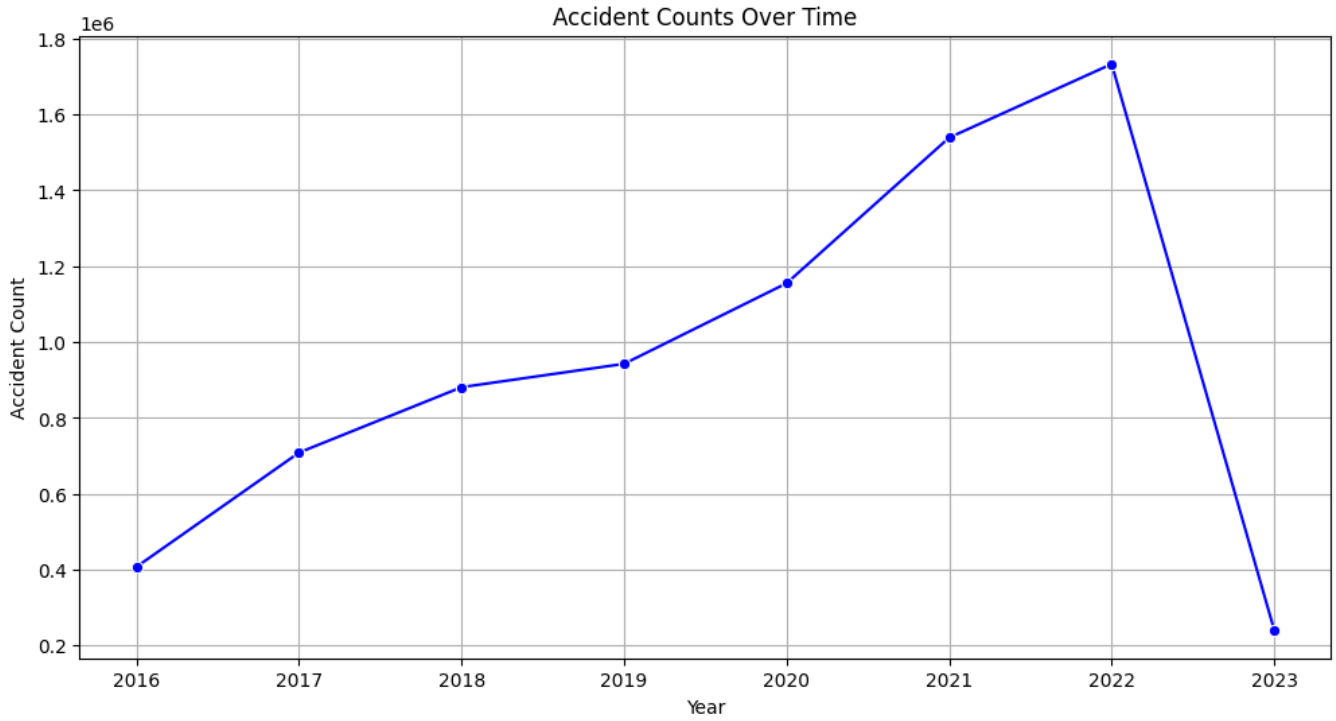
1. **Severity Distribution:** We visualized the distribution of accident severity, a crucial metric for understanding the impact of accidents.



2. **Accident Locations:**We identified the top states with the highest accident counts, shedding light on regions with frequent accidents.



3. **Accident Time Trends:**We analyzed how accident counts vary with the year, allowing us to understand temporal trends in accident data.

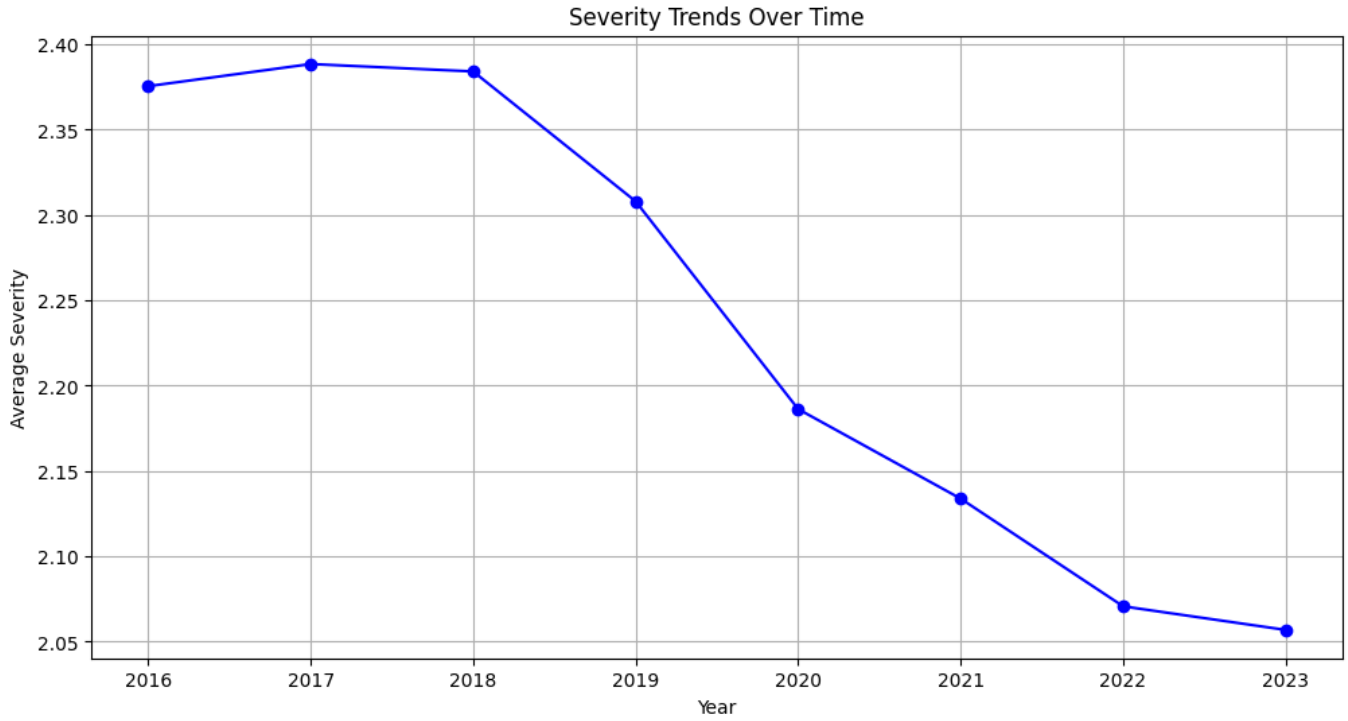


3.4 Descriptive Statistics and Inferential Statistics

In this section, we delve into several key analyses to gain insights into accident data, exploring trends, correlations, and significant factors related to accidents. Each analysis is framed as a specific question, and we provide a summary of the findings and the methods used for the analysis.

3.4.1 Analysis 1: Accident Severity Trends Over Time

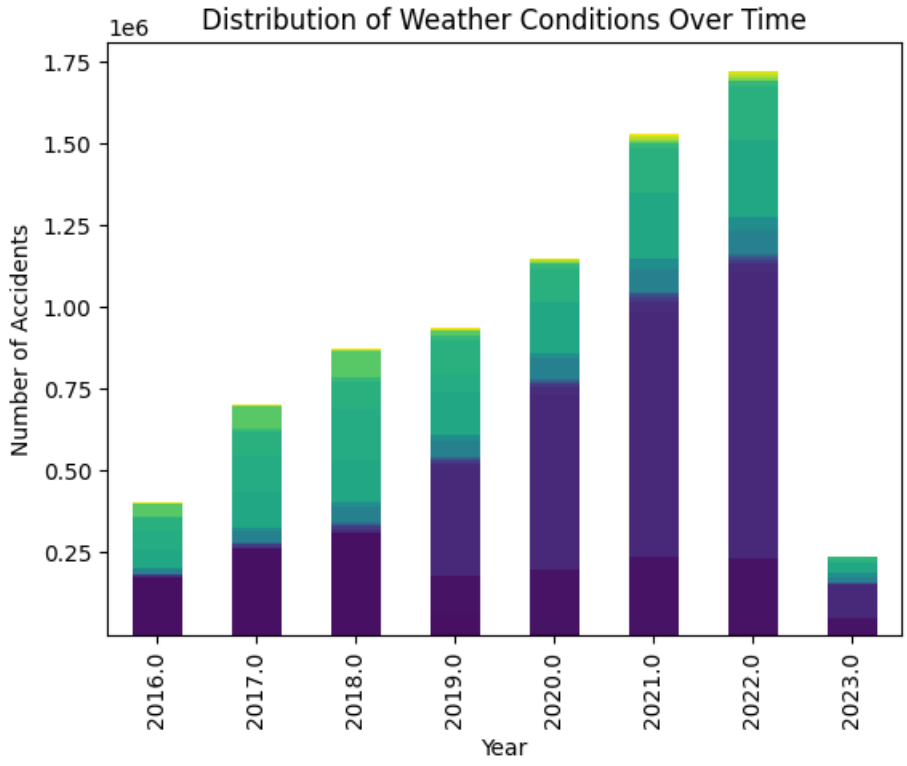
Question: Are there significant trends in accident severity over time?
We look at how accident severity has changed over time in this analysis. The average accident severity for each year is computed after the data is sorted by year. Severity trends are visualised using a line plot. In addition, a linear regression model is used to evaluate the trend's statistical significance. The findings show that the severity of accidents has been trending significantly over time.



3.4.2 Analysis 2: Weather Conditions During Accidents Over Time

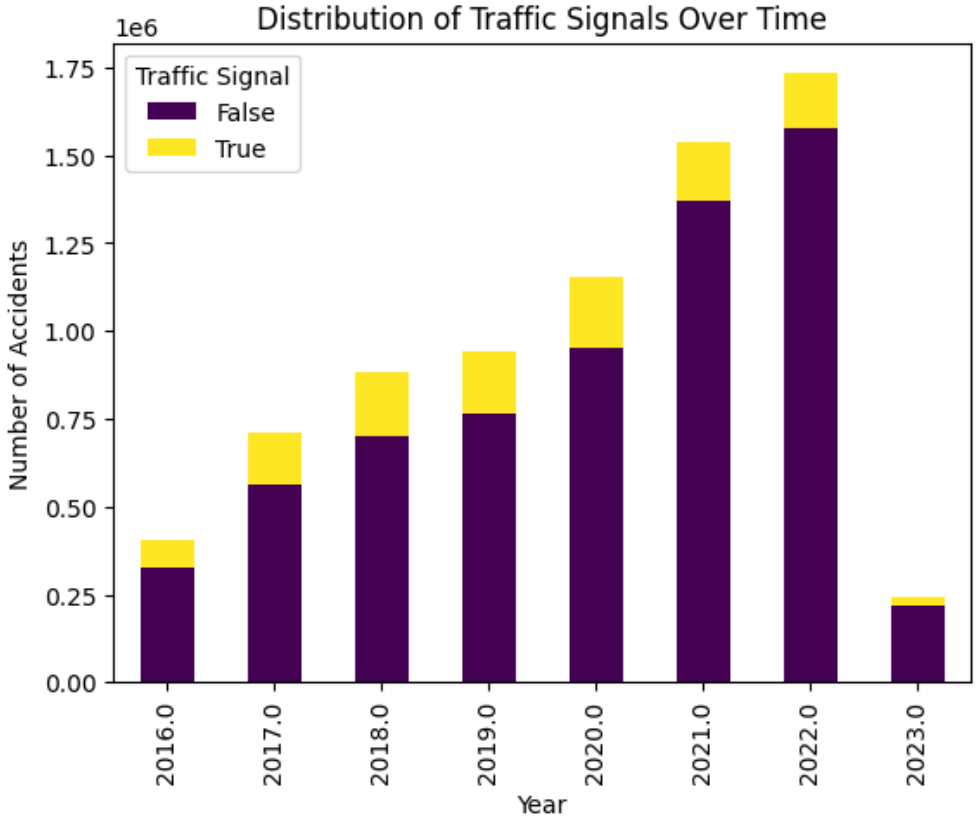
Question: Are there noticeable trends in weather conditions during accidents over time?

In this analysis, we explore the variations in weather conditions during accidents across different years. The data is grouped by year, and we count the occurrences of each unique weather condition. A stacked bar chart is created to compare the distribution of weather conditions over time. A chi-squared test is performed to determine if there is a statistically significant difference in weather condition distribution over time. And we found out that There is a statistically significant difference in weather condition distribution over time.



3.4.3 Analysis 3: Traffic Signal Distribution Over Time

Question: How has the traffic signal distribution evolved over time?
In this analysis, we investigate how the presence of traffic signals during accidents has changed over the years. The data is grouped by year, and the counts of accidents with and without traffic signals are recorded. A stacked bar chart is used to visualize the distribution of traffic signals over time. A chi-squared test is conducted to assess the statistical significance of changes in traffic signal distribution. And the findings show us that There is a statistically significant difference in traffic signal distribution over time.

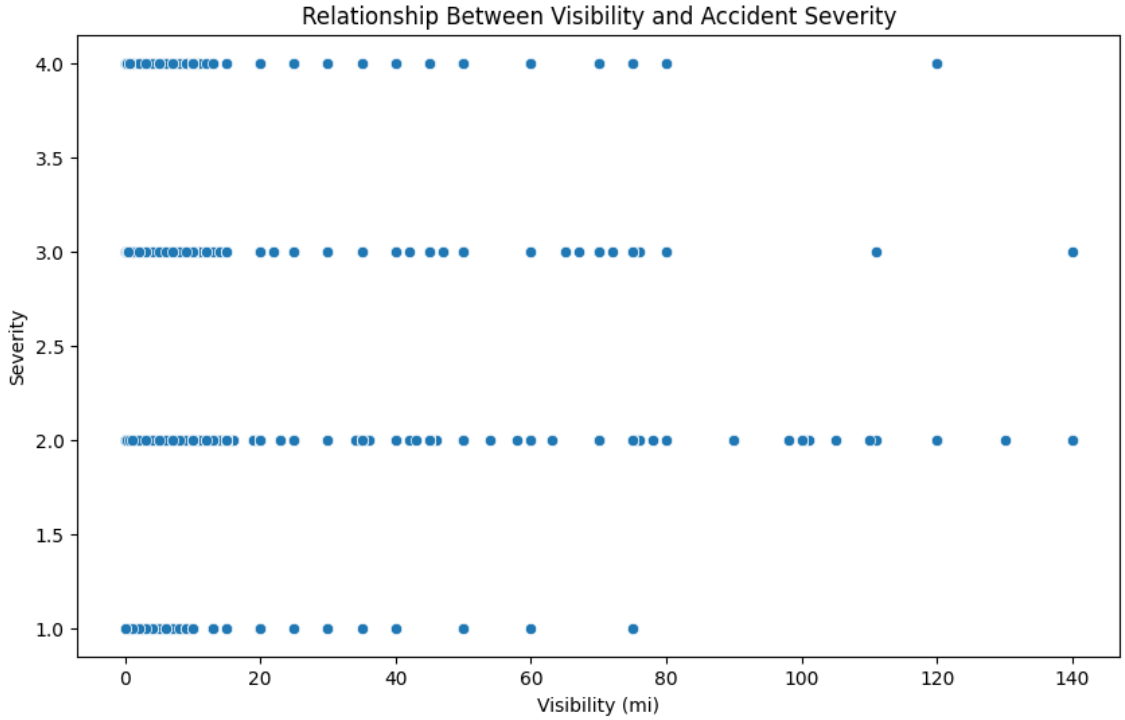


3.4.4 Analysis 4: Geographic Regions and Accident Severity

Question: Are certain geographic regions (counties) associated with higher accident severity?
We investigate the connection between geographical areas (counties) and accident severity in this analysis. Each county's average accident severity is determined, and the highest and lowest average severity counties are noted. To find out if there are statistically significant variations in accident severity between counties, a Kruskal-Wallis test is also run. We knew that higher accident severity is related to certain geographic regions (counties).

3.4.5 Analysis 5: Relationship Between Visibility and Accident Severity

Question: Is there a relationship between visibility and accident severity?
We investigate the possible relationship between visibility (measured in miles) and the seriousness of accidents in this analysis. This relationship is visualised using a scatter plot. The means of severity for the various visibility conditions are compared using a t-test, and the results are evaluated for statistical significance. We discover that the severity of accidents does not significantly vary depending on the visibility conditions.



4 Conclusion

In conclusion, our analyses have provided valuable insights into two distinct datasets: Spotify song data and accident data.

We found special patterns in the energy levels of songs using Spotify song data, which points to a significant change in the songs' musical qualities over time. On the other hand, no discernible patterns in song popularity, instrumentality, or track length were found. Furthermore, no statistically significant relationship was found between the songs' release year and tempo. Certain time signatures have become more common in recent music, according to the analysis of time signature distribution.

Our analysis of accident data showed noteworthy patterns in the severity of accidents over time, emphasising the need for ongoing observation and safety precautions. The distribution of weather conditions at the time of accidents showed a substantial variation over time, suggesting possible alterations in the climate or driving habits. Additionally, we found a noteworthy distinction in the traffic signal distribution during collisions, indicating changes in traffic control over time. The relevance of targeted safety initiatives is highlighted by the correlation between higher accident severity and specific geographic regions (counties). Lastly, there was no discernible correlation between visibility and the seriousness of the accidents.

Overall, our analyses provide a comprehensive understanding of these datasets, shedding light on trends, correlations, and significant factors related to songs and accidents. These insights can inform future decisions in both the music industry and road safety.