Data Science

answer of assignments 1

***Professor** : Dr. Kherad Pishe*

***Assistant Professor** : Mohammad Reza Khanmohammadi*

Hasan Roknabady – 99222042

TASK 1: HOUSE PRICES DATASET

INTRODUCTION:

The house prices dataset provides valuable insights into the factors influencing residential property values. Our analysis aims to explore housing market trends and understand the relationships between various features and sale prices.

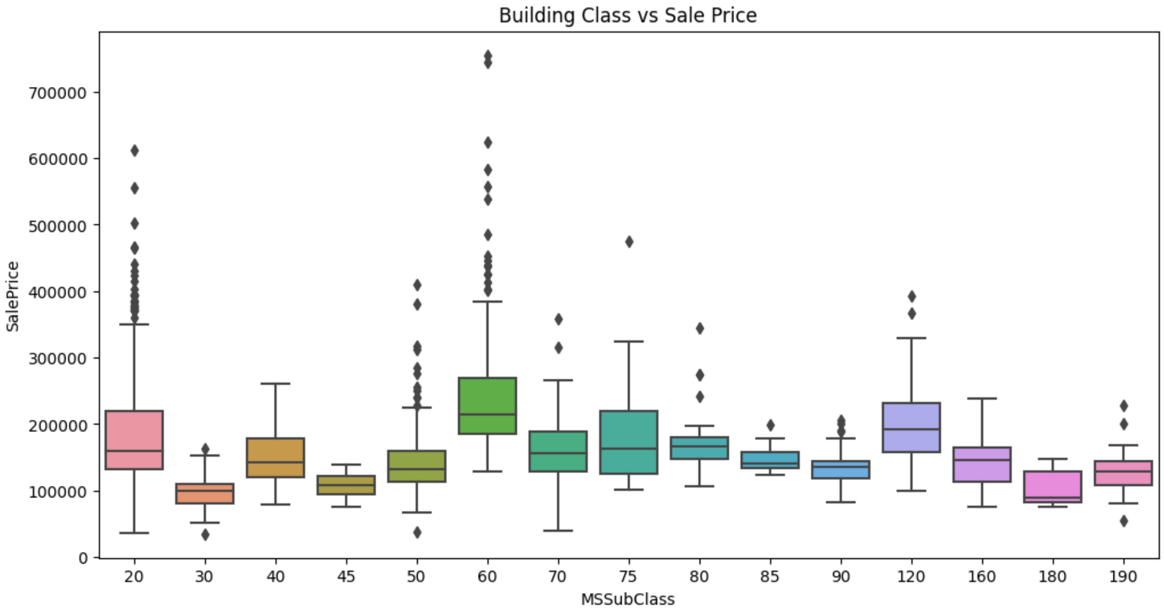
EXPLORATORY DATA ANALYSIS (EDA):

1. OVERVIEW OF DATASET:

We began our analysis by examining the dataset's structure, checking for missing values, and providing summary statistics. The dataset comprises 81 columns, including features such as building class, overall quality, heating types, and utilities.

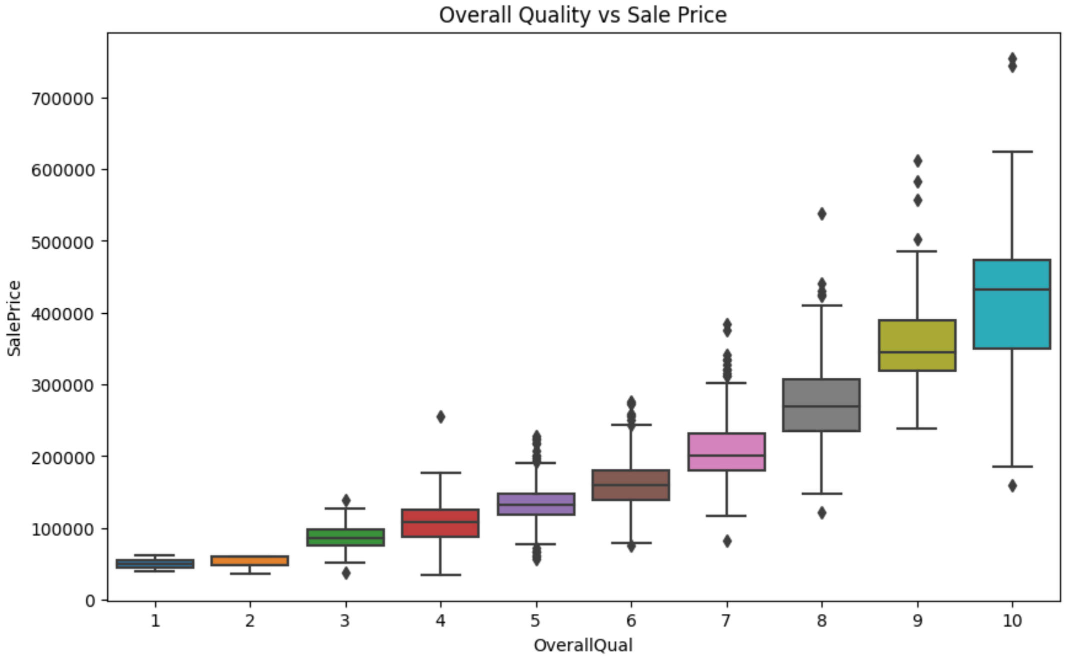
2. BUILDING CLASS AND IMPORTANCE:

The distribution of building classes reveals that class 20 is the most common, followed by classes 60 and 50. A boxplot visualization highlights variations in sale prices across different building classes. Further analysis shows that building class is a significant factor in determining property values.



3. OVERALL QUALITY AND SALE PRICES:

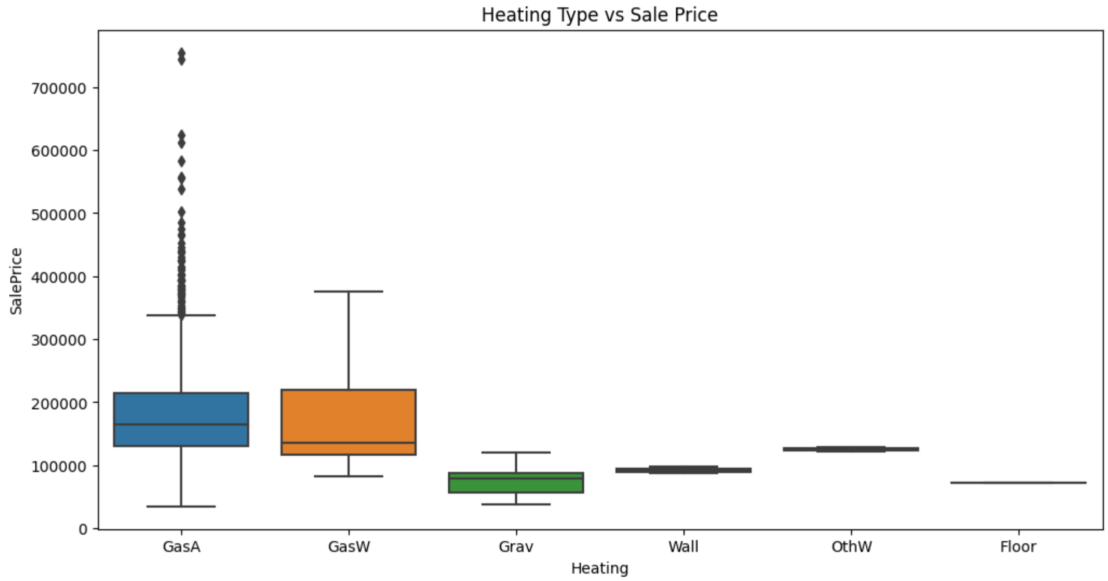
The overall quality rating exhibits a strong positive correlation with sale prices, indicating that higher-quality houses tend to have higher values. A boxplot visualization reinforces this trend, showing a clear relationship between overall quality and sale prices.



The average sale prices by overall quality provide valuable insights, and the correlation coefficient of approximately 0.79 indicates a strong positive correlation between overall quality and sale prices. As overall quality increases, the average sale price tends to increase as well.

4. HEATING TYPES AND SALE PRICES:

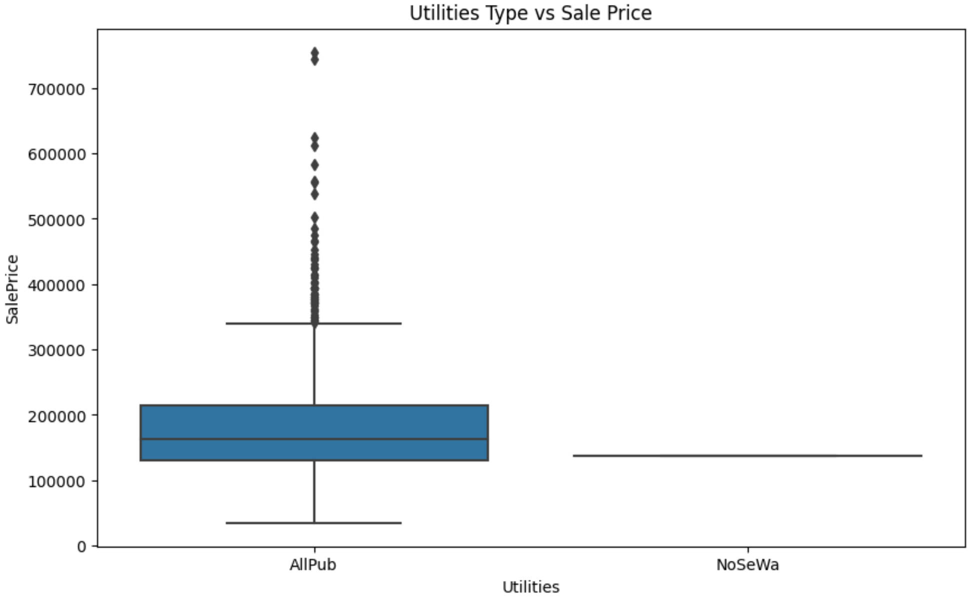
Exploring different heating types and their impact on sale prices reveals variations in average prices. The ANOVA test suggests that the type of heating does have a statistically significant effect on property values.



The average sale prices by heating type provide insights into how different heating systems relate to property values. The ANOVA test result with a p-value of 0.00075 suggests that there are statistically significant differences in sale prices among different heating types, that visualizing show that GasA and GasW has more sale price

5. UTILITIES TYPES AND SALE PRICES:

While there are variations in average sale prices based on utilities types, the ANOVA test indicates that these differences are not statistically significant. The presence of utilities does not appear to be a significant factor in determining property values.



It appears that the average sale prices do show some variation between different utility types. However, the ANOVA test result with a p-value of 0.5847 suggests that there may not be a statistically significant difference in sale prices among different utility types. This means that, based on the available data, the type of utilities in a property may not have a significant impact on its sale price and visualization shows that AllPub has more Sale price value.

ADDITIONAL CREATIVE QUESTIONS:

6. FEATURE IMPORTANCE ANALYSIS:

A correlation analysis identified key features strongly correlated with sale prices. Overall quality, above-ground living area, and garage features emerged as critical factors influencing property values.

7. OUTLIER DETECTION AND ANALYSIS:

Twenty-two outliers were identified based on Z-scores for sale prices. Further investigation into these outliers can provide insights into unique characteristics or anomalies in the dataset.

8. CATEGORICAL FEATURE ANALYSIS:

A Chi-square test for independence was performed on SaleCondition and Fireplaces, revealing a significant association between these two categorical variables.

9. TEMPORAL TRENDS IN SALE PRICES:

Exploring temporal trends in sale prices over time may provide insights into how the real estate market has evolved. Visualization of the trend can help identify patterns.

TASK2 : SPOTIFY TOP SONGS DATASET ANALYSIS

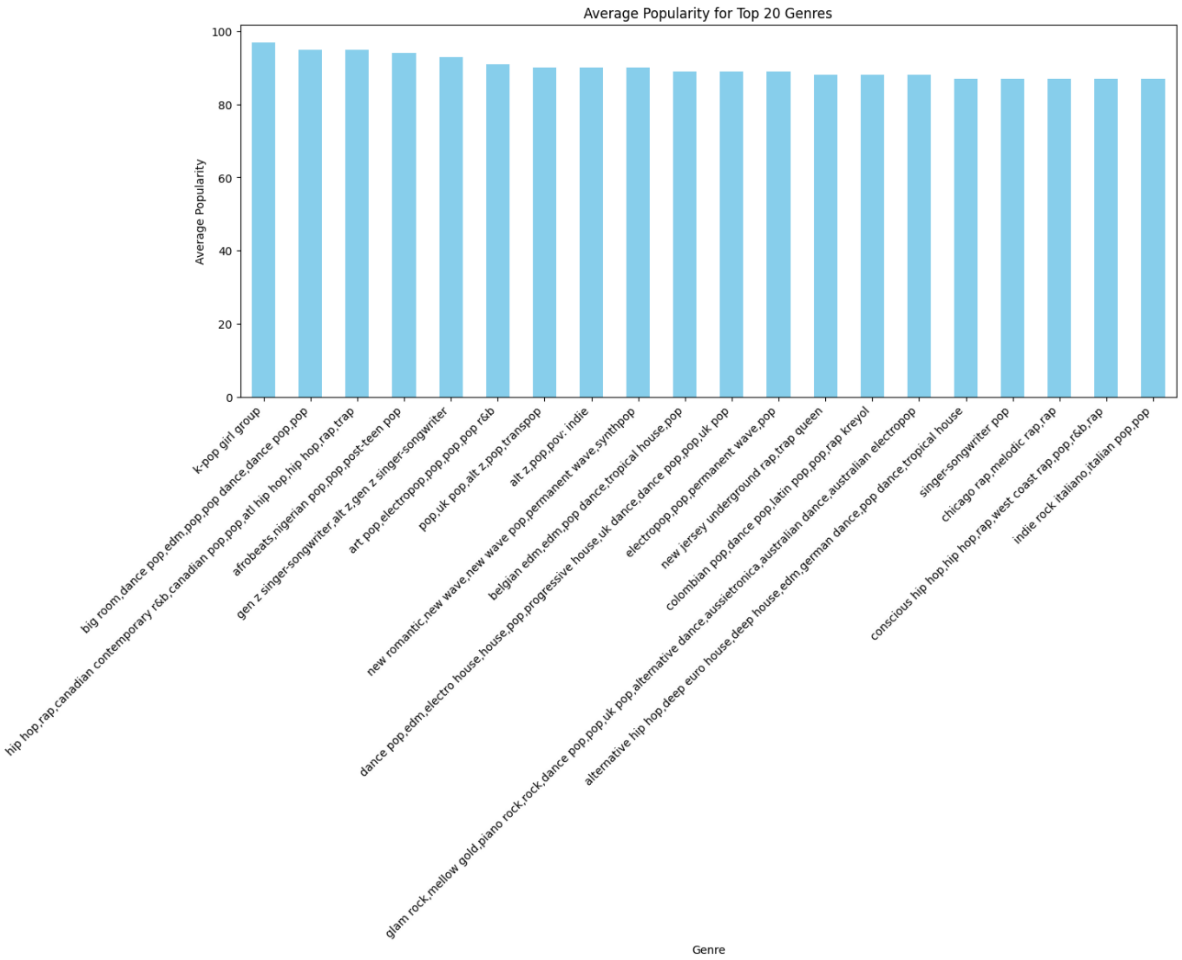
EXECUTIVE SUMMARY

The Spotify Top Songs dataset, consisting of 10,000 popular songs spanning from 1960 to the present day, was subjected to a comprehensive analysis. The focus was on exploring key features, relationships, and trends within the dataset. The analysis involved data cleaning, visualization, and statistical testing.

EDA (EXPLORATORY DATA ANALYSIS)

QUESTION 1: GENRE POPULARITY

Visualization: A bar chart was created to showcase the average popularity of different genres. •

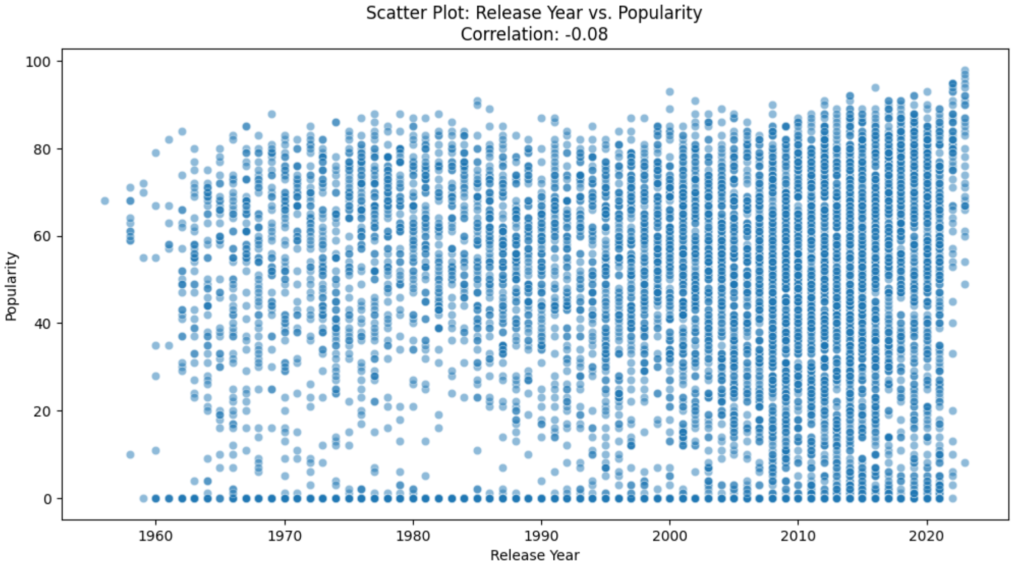


•

- **Insights:** Acoustic pop and neo mellow emerged as among the most popular genres.

QUESTION 2: RELEASE YEAR VS. POPULARITY

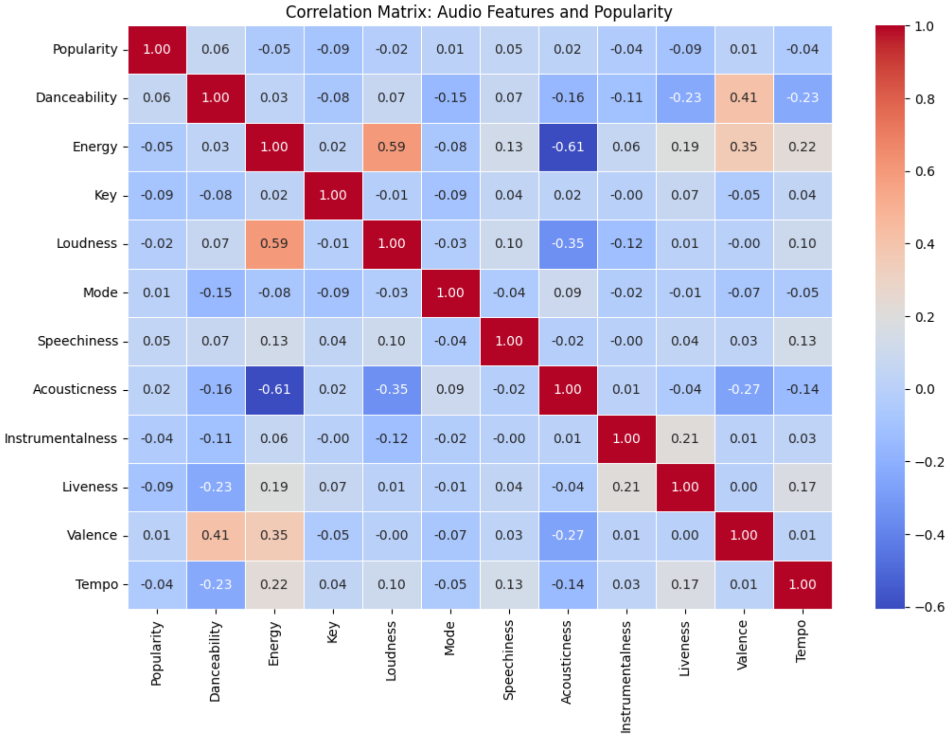
- **Visualization:** A scatter plot was generated to examine the relationship between release year and popularity.



- **Insights:** A slight negative correlation was observed, suggesting that, on average, older songs tend to have slightly lower popularity.

CORRELATION HEATMAP

- **Visualization 4:** A heatmap was generated to visually represent the correlation matrix of key features.

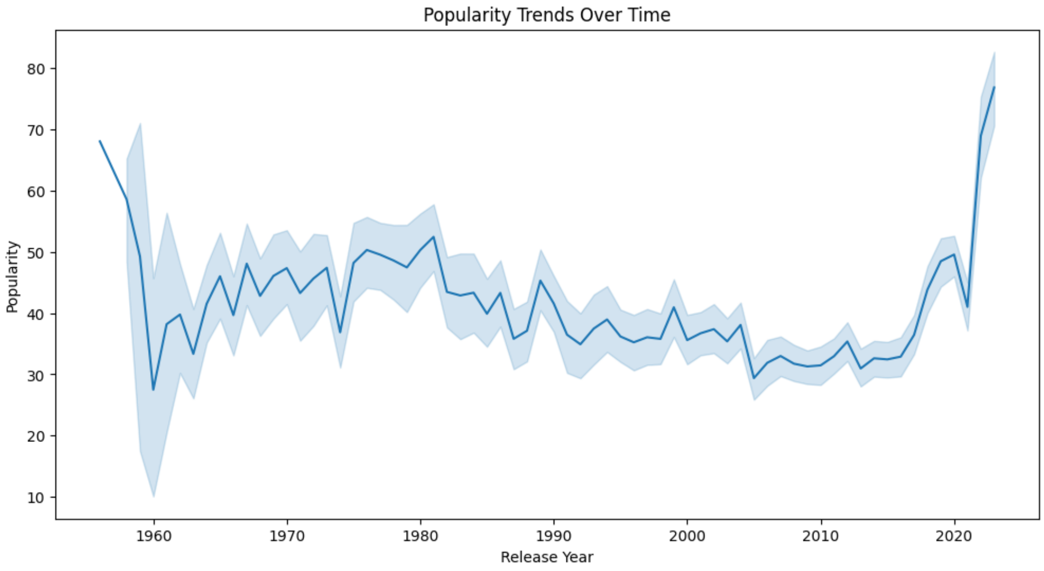


	Popularity	Danceability	Energy	Key	Loudness	Mode	Speechiness	Acousticness	Instrumentalness	Liveness	Valence	Tempo
Popularity	1.00	0.06	-0.05	-0.09	-0.02	0.01	0.05	0.02	-0.04	-0.09	0.01	-0.04
Danceability	0.06	1.00	0.03	-0.08	0.07	-0.15	0.07	-0.16	-0.11	-0.23	0.41	-0.23
Energy	-0.05	0.03	1.00	0.02	0.59	-0.08	0.13	-0.61	0.06	0.19	0.35	0.22
Key	-0.09	-0.08	0.02	1.00	-0.01	-0.09	0.04	0.02	-0.00	0.07	-0.05	0.04
Loudness	-0.02	0.07	0.59	-0.01	1.00	-0.03	0.10	-0.35	-0.12	0.01	-0.00	0.10
Mode	0.01	-0.15	-0.08	-0.09	-0.03	1.00	-0.04	0.09	-0.02	-0.01	-0.07	-0.05
Speechiness	0.05	0.07	0.13	0.04	0.10	-0.04	1.00	-0.02	-0.00	0.04	0.03	0.13
Acousticness	0.02	-0.16	-0.61	0.02	-0.35	0.09	-0.02	1.00	0.01	-0.04	-0.27	-0.14
Instrumentalness	-0.04	-0.11	0.06	-0.00	-0.12	-0.02	-0.00	0.01	1.00	0.21	0.01	0.03
Liveness	-0.09	-0.23	0.19	0.07	0.01	-0.01	0.04	-0.04	0.21	1.00	0.00	0.17
Valence	0.01	0.41	0.35	-0.05	-0.00	-0.07	0.03	-0.27	0.01	0.00	1.00	0.01
Tempo	-0.04	-0.23	0.22	0.04	0.10	-0.05	0.13	-0.14	0.03	0.17	0.01	1.00

- **Insights 4:** The heatmap highlights significant correlations among various features, aiding in understanding the internal relationships within the dataset.

Additional Visualization: Popularity Over Time

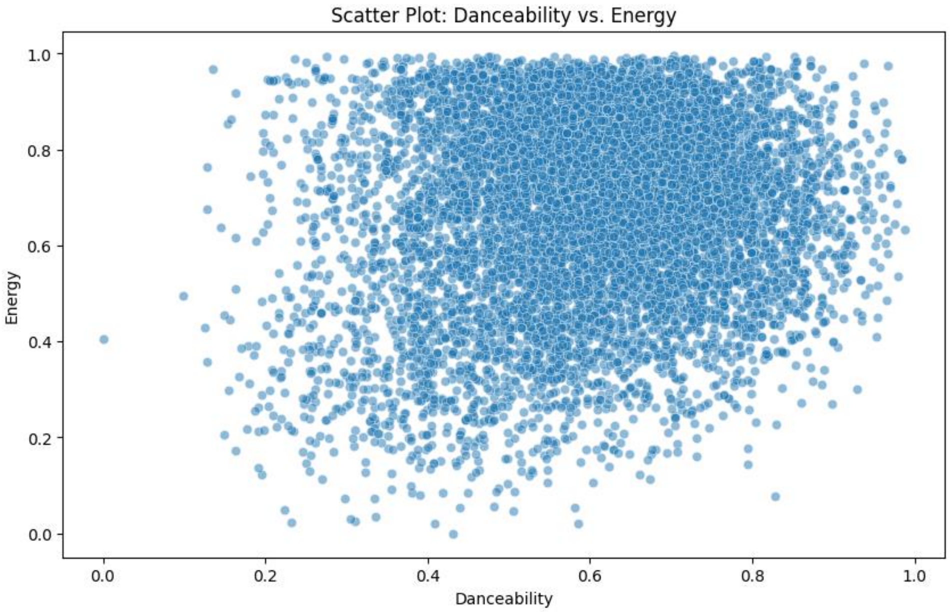
- **Visualization 5:** A line chart was created to depict the trend of average popularity over time.



- **Insights 5:** The chart shows fluctuations in popularity, with certain periods witnessing a surge in average song popularity, that shows that before 1960 and after 2020 we have more trends songs over other periods.

DANCEABILITY VS. ENERGY

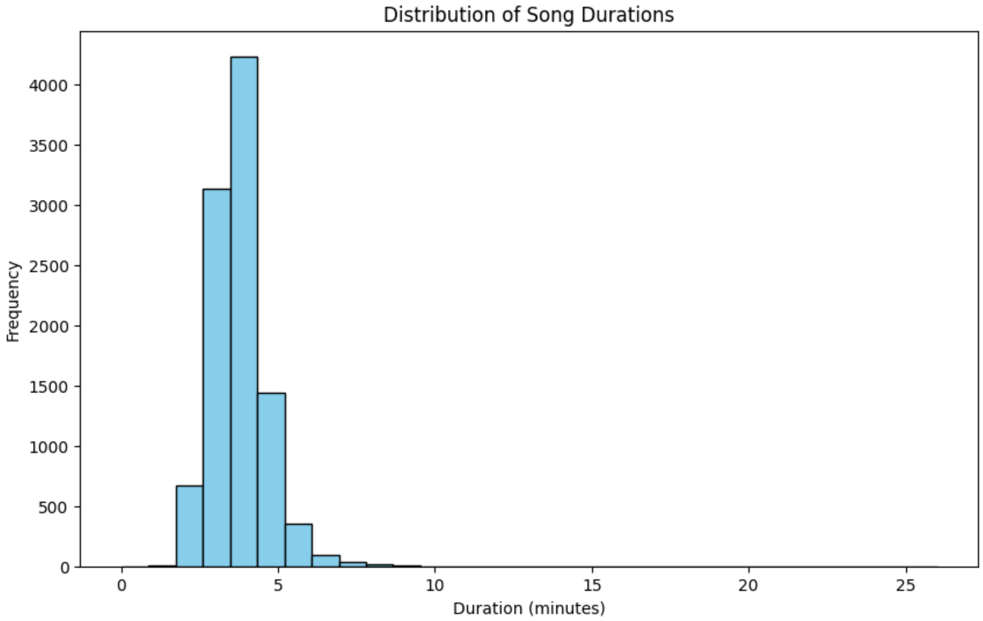
- **Visualization:** Another scatter plot was created to explore the correlation between danceability and energy.



- **Insights:** A weak positive correlation (0.13) was found, indicating that songs with higher danceability tend to have slightly higher energy.

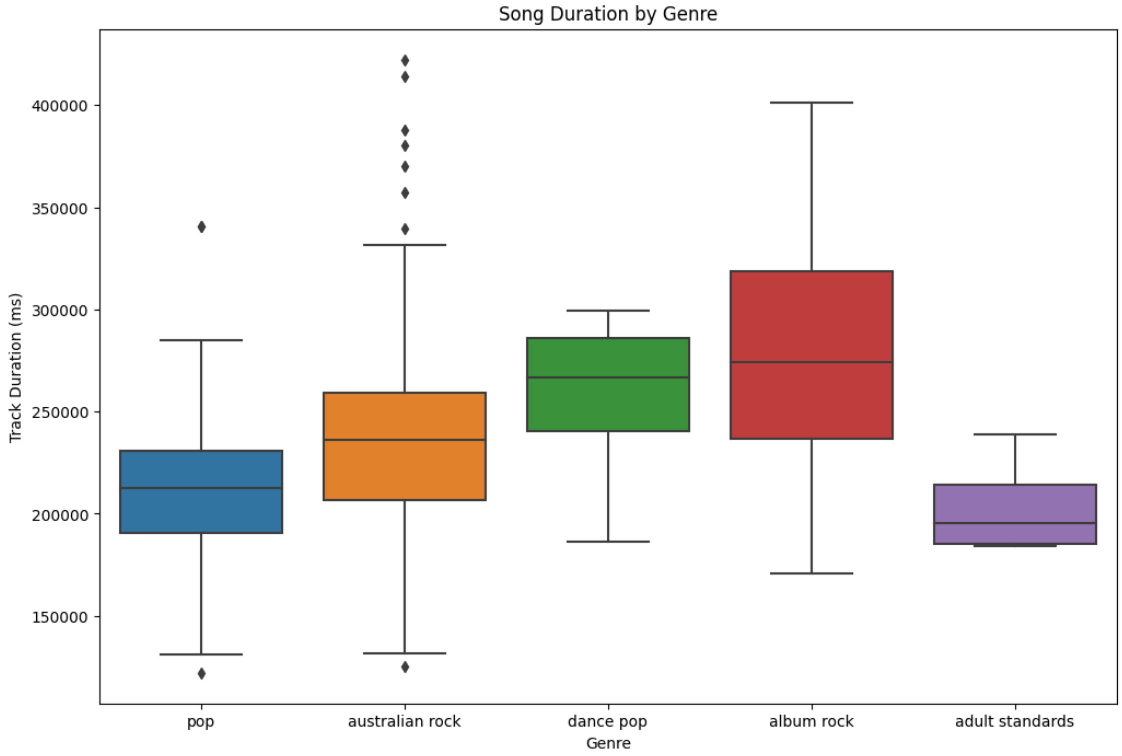
DISTRBUTION OF SONG DURATION

- **Visualization:** It looks like the ANOVA test for song duration by genre has a significant result, indicating that there are differences in song durations among the top genres. Now, let's create a boxplot to visually represent these differences:



- **Insights:** The boxplot provides a clear visualization of the differences in song durations among the top genres. It seems like there are variations in song durations, with some genres having a wider range of durations compared to others, that shows that most durations are between 0 – 10 and most of them are between 2 – 6 minutes.

Now, let's create a boxplot to visually represent these differences:



The boxplot provides a clear visualization of the differences in song durations among the top genres. It seems like there are variations in song durations, with some genres having a wider range of durations compared to others.

Now, let's move on to another statistical test. We can perform a t-test to compare the mean popularity of songs with explicit content and songs without explicit content.

The independent samples t-test results indicate a statistically significant difference in the mean popularity between explicit and non-explicit songs. The t-statistic of 4.72 suggests that the difference is substantial, and the p-value of approximately 2.98e-06 is well below the typical significance level of 0.05. Therefore, we reject the null hypothesis and conclude that there is a significant difference in popularity between explicit and non-explicit songs.

MORE STATISTICAL ANALYSES

ANOVA TEST: TOP GENRES

- **Hypothesis:** The mean popularity differs significantly among the top genres.
- **Results:**
  - F-statistic: 29.99
  - P-value: 4.03e-51
- **Conclusion:** The null hypothesis is rejected, suggesting that there is a significant difference in mean popularity among the top genres.

DANCEABILITY VS. ENERGY

- **Hypothesis:** There is a significant correlation between danceability and energy.
- **Results:**
  - Correlation coefficient: 0.13
- **Conclusion:** A weak positive correlation was found between danceability and energy.

CONCLUSION

The analysis provided valuable insights into the Spotify Top Songs dataset. From genre popularity to the relationship between release year and popularity, the document captures key trends and findings. The statistical analyses added a robust quantitative layer to the exploratory process.