Chinook Database Analysis

Negar Ghiasi Far
March 2025

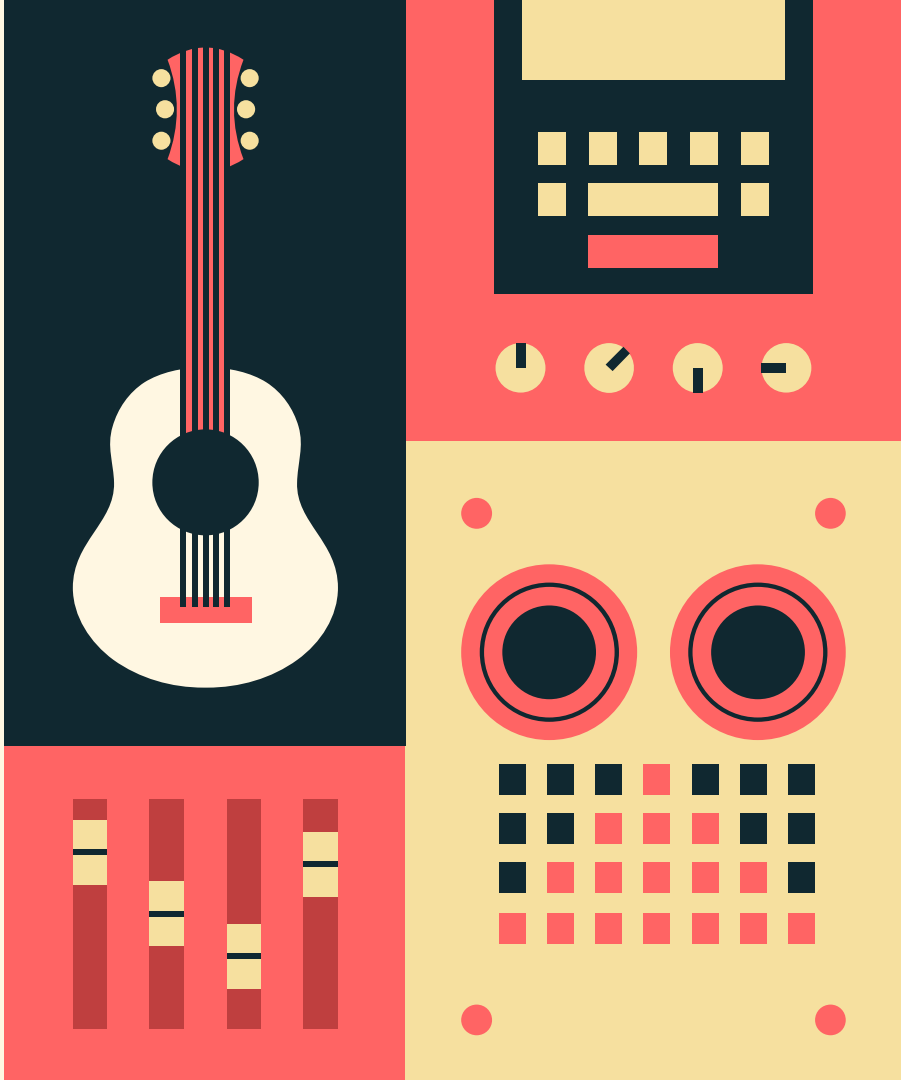
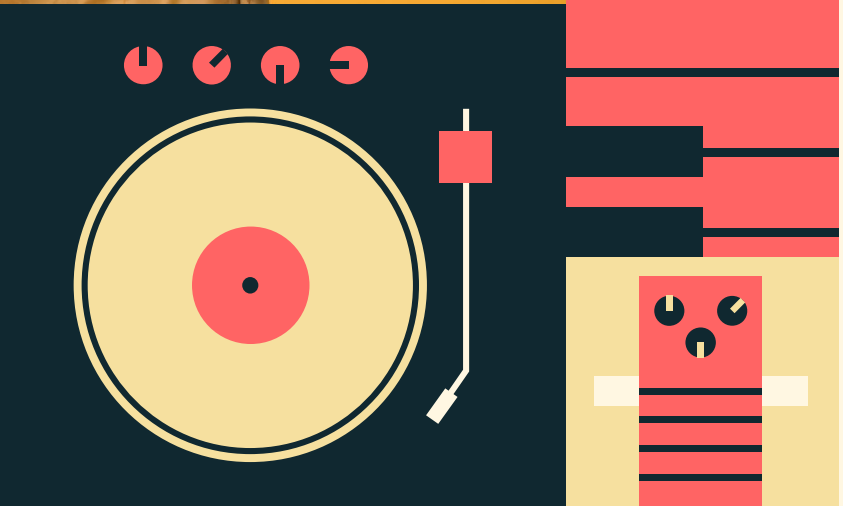
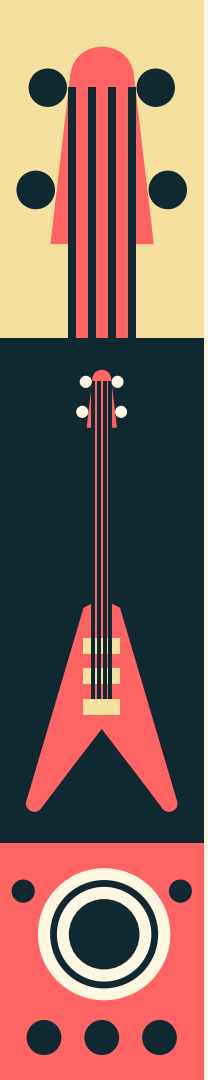


TABLE OF CONTENTS ★



01 & 02 ★

Loading Database & Initial Review



Loading Database

Using **SQL alchemy** python library I have loaded **Chinook Database** into data frames and started the analysis.

The Chinook Database holds information about a music store, containing tables for artists, albums, media tracks, invoices, and customers.

Initial Review

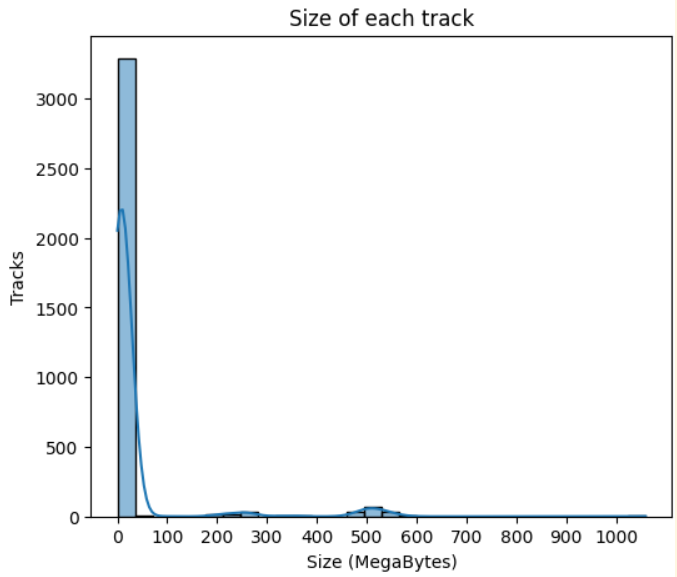
Checked data types, missing values, and basic statistics to ensure data quality.

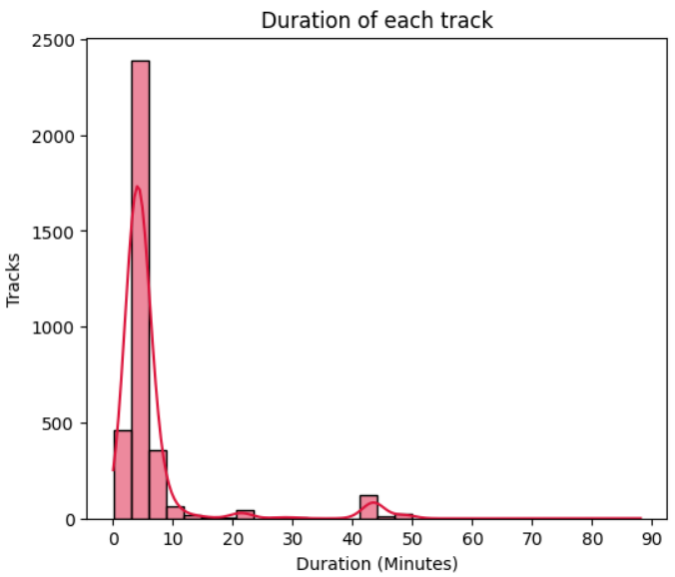
03 ★

Key Variables

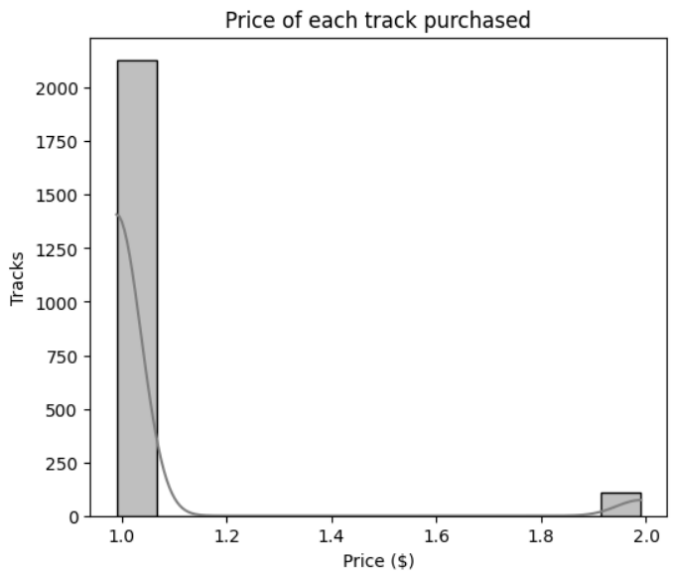
Key Variables

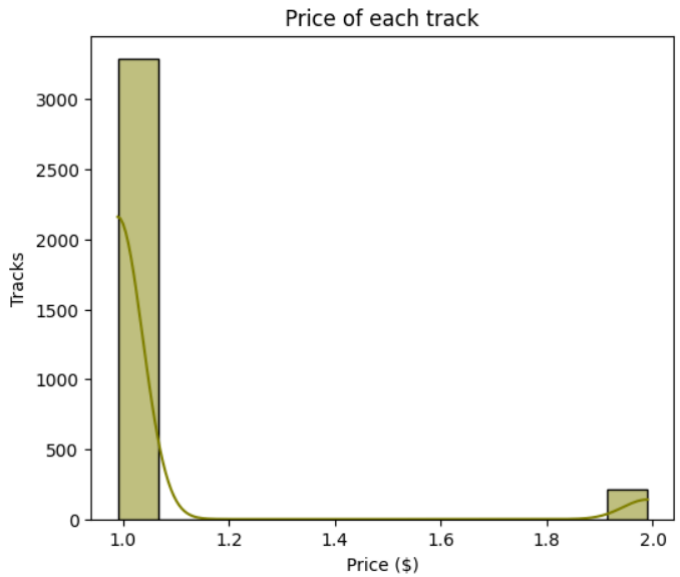
There are many variables that describe each **Track**. Below are the distrubutions.





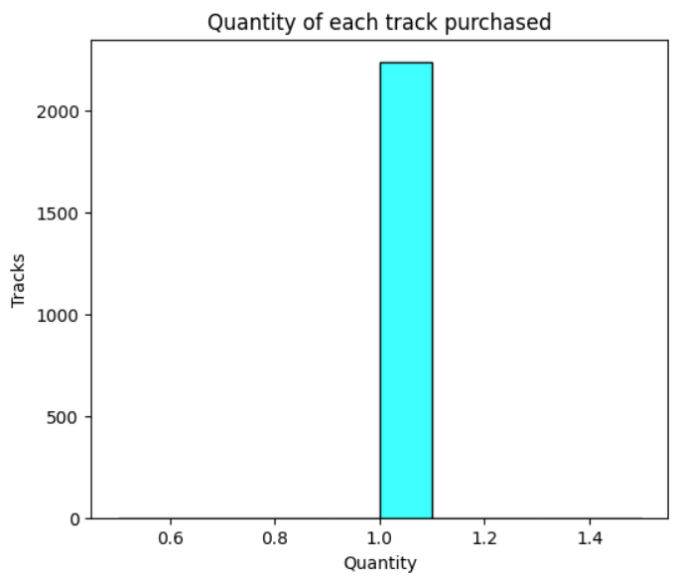
★ Key Variables

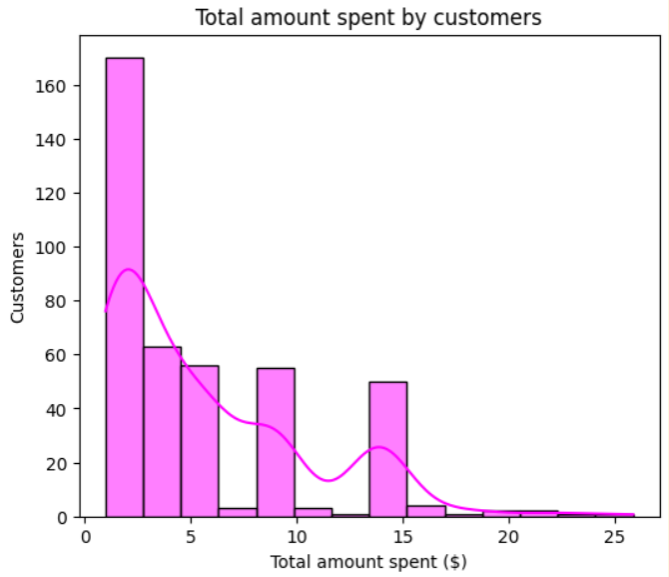


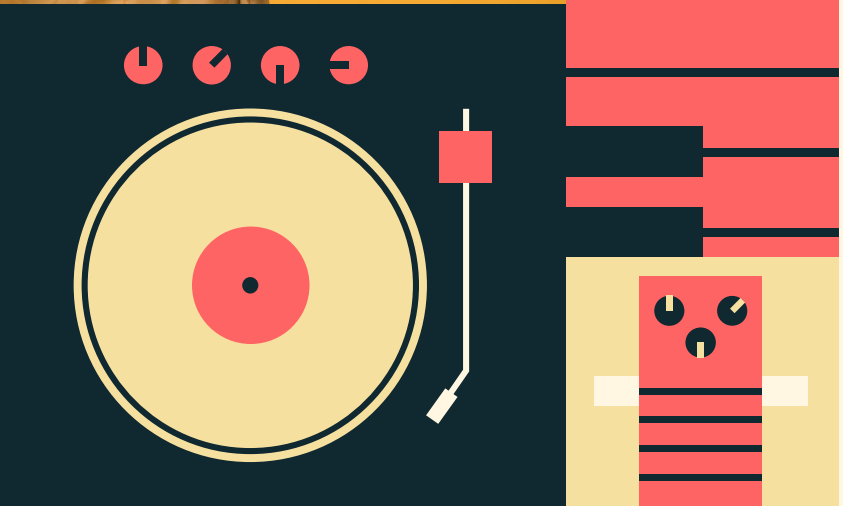


Key Variables

Plots below show key variables about purchases.







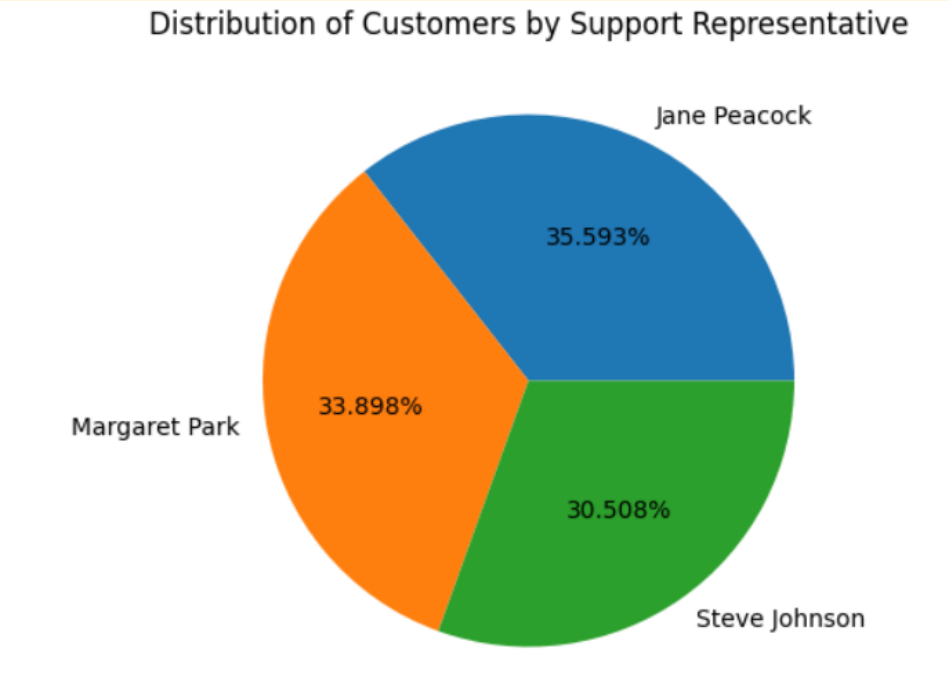
04 ★

Understanding Data

Understanding Data

Customer / Support Rep

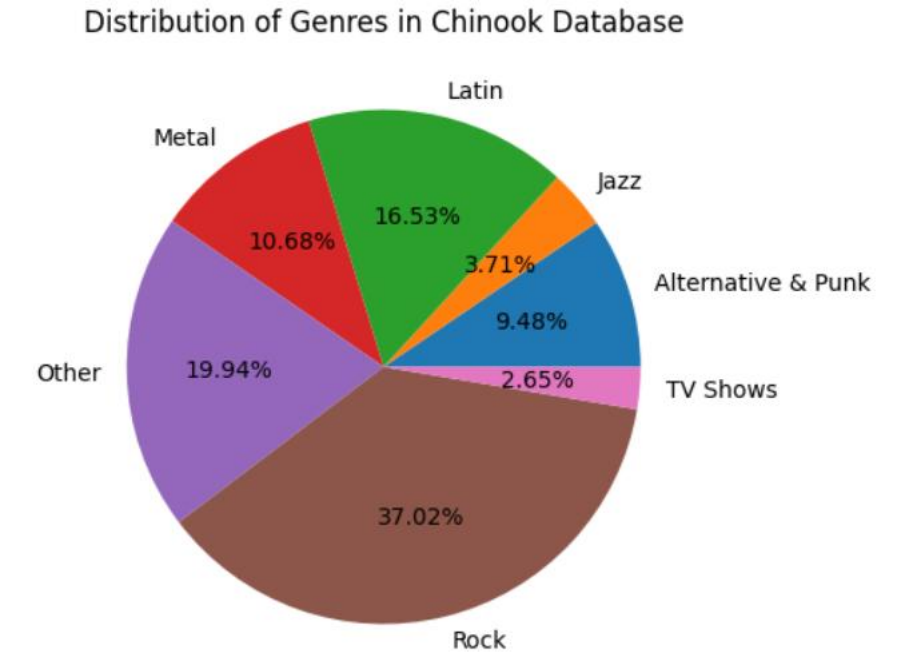
This plot shows distribution of support representatives across customers.



Understanding Data

Genre

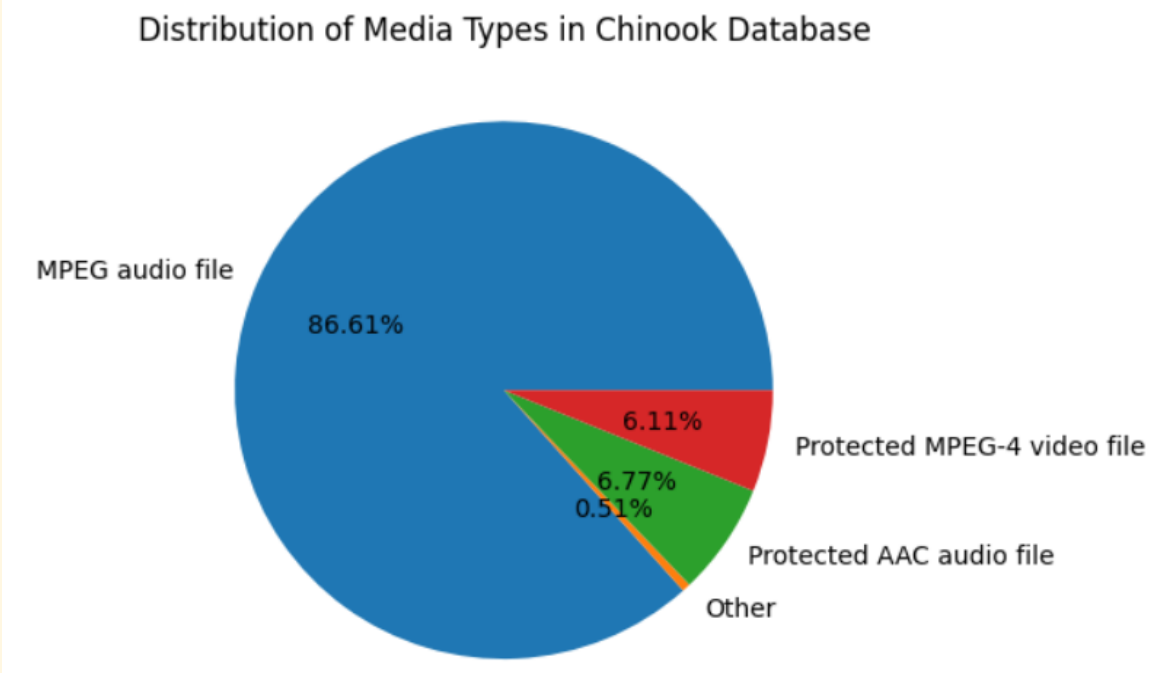
This plot shows distribution of genres of each track.



Understanding Data

Media Type

This plot shows distribution of Media Types of each track.



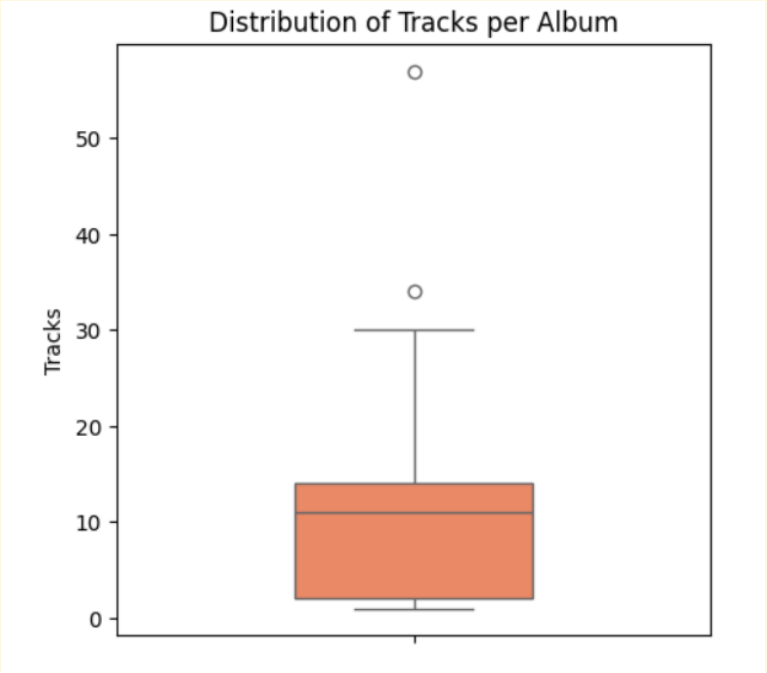
Understanding Data

Track / Album

This plot shows distribution of Tracks per album.

The average number of tracks per album is 10.

50% of albums have 2 to 14 tracks.



05 ★

Normality

Normality

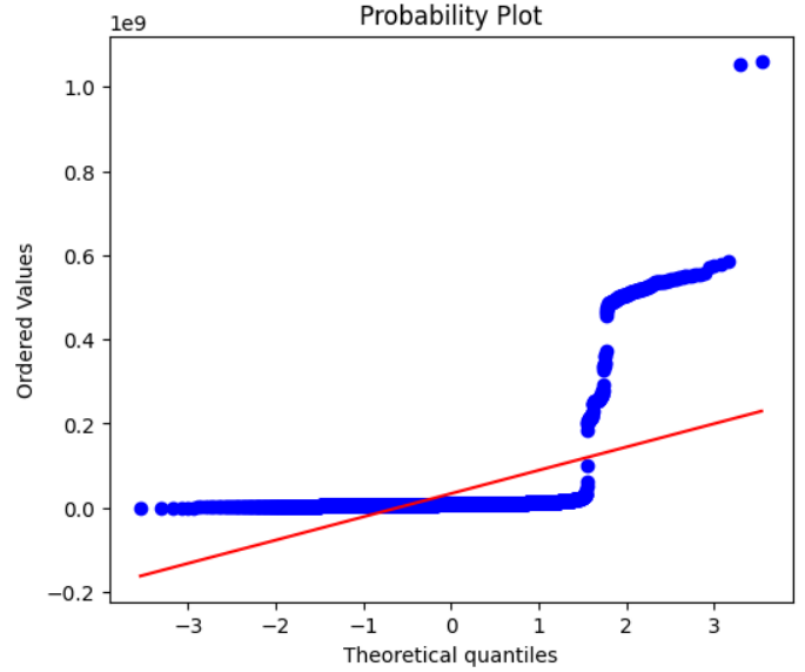
Size of each track

Shapiro-Wilk Test

Statistic = 0.275,
p-value = 1.72e-79

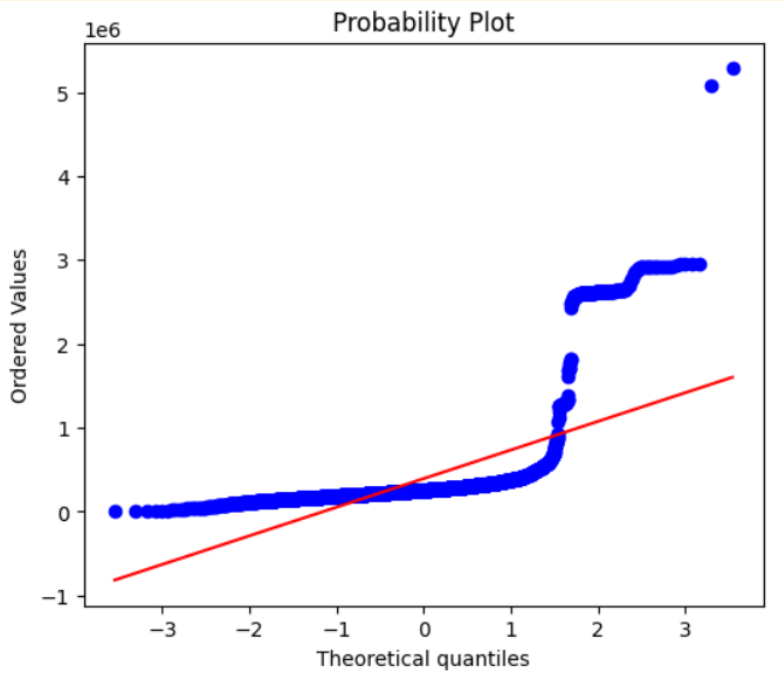
Normality Disproved

Q-Q PLOT



Normality

Q-Q PLOT



Duration of each track

Shapiro-Wilk Test

Statistic = 0.406,
p-value = 2.72e-75

Normality Disproved

Normality

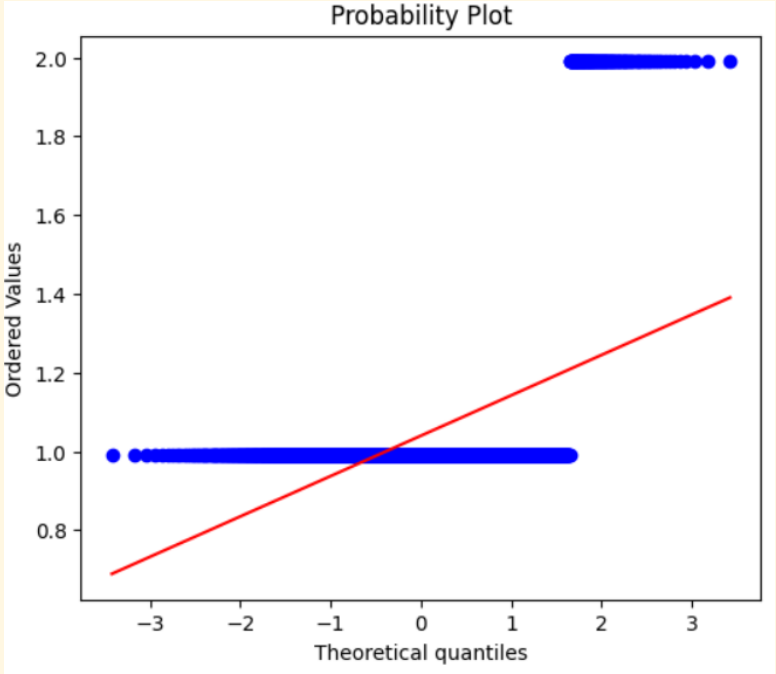
Price of each track purchased

Shapiro-Wilk Test

Statistic = 0.222,
p-value = 2.10e-70

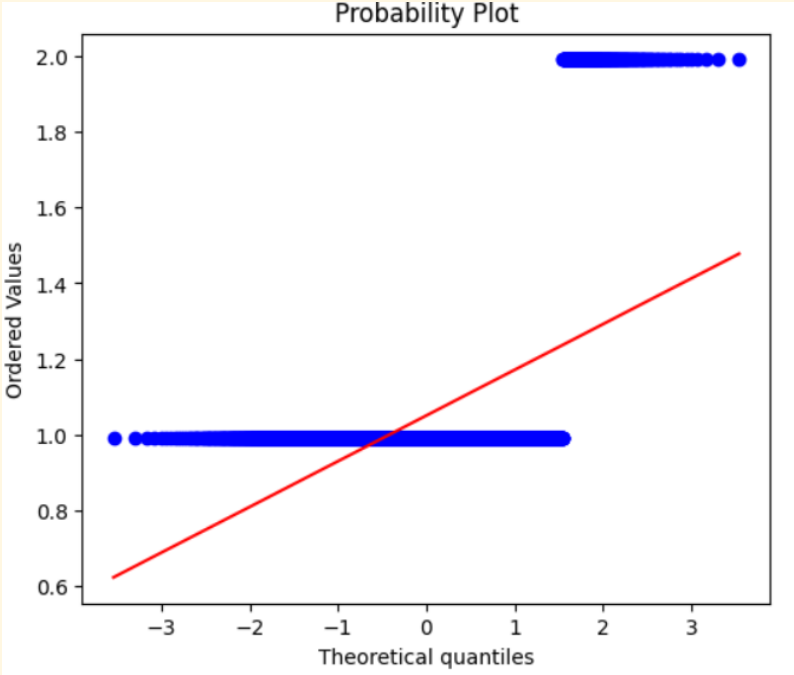
Normality Disproved

Q-Q PLOT



Normality

Q-Q PLOT



Price of each track

SHAPIRO-WILK TEST

Statistic = 0.253,
p-value = 3.98e-80

Normality Disproved

Normality

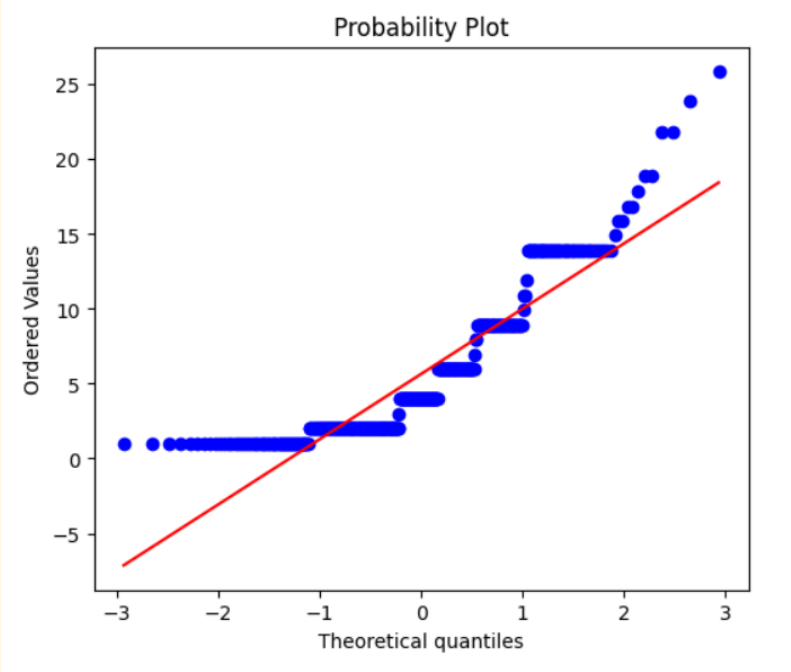
Total amount spent by customers

Shapiro-Wilk Test

Statistic = 0.836,
p-value = 3.4e-20

Normality Disproved

Q-Q PLOT



06 ★

Outliers

Outliers

Number of Outliers

- Size of each track (343 points)
- Duration of each track (357 points)
- Price of each track purchased (111 points)
- Price of each track (213 points)
- Total amount spent by customers (4 points)

Given the high number of detected outliers, removing them could introduce bias and distort the true distribution of the data. So I did not remove the outliers.

07 ★

Hypothesis Test

★ Hypothesis Test

1) 3 Most popular genres

GenreId	Name_genre	Count
1	Rock	835
7	Latin	386
3	Metal	264

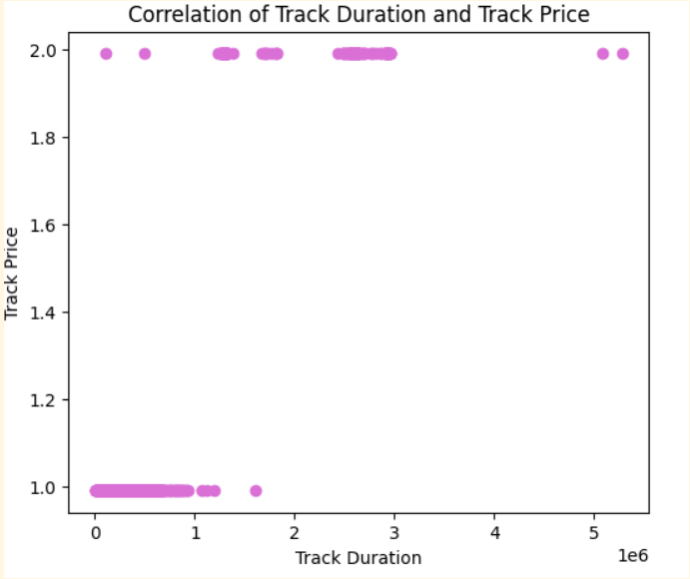
Kruskal-Wallis Test

Null hypothesis: No difference in price between most popular genres.
Alternative hypothesis: Significant difference in price between most popular genres.

There is no need for a hypothesis test because all the prices in these genres are the same.

★ Hypothesis Test

2) Dependency of Track duration and price



Spearman correlation

Spearman Correlation Coefficient: 0.409,
p-value: 4.57e-142

The track duration and price are dependent.
There is a medium dependency between duration and price of each track.

Hypothesis Test

4) Dependency of Genres and Media Types

Chi-square test

Null hypothesis: No difference in Genre and Media Type.
Alternative hypothesis: Significant difference in Genre and Media Type.

Reject the null hypothesis.
There is a significant association between Genre and Media Type.

Hypothesis Test

5) Dependency of the Customer's Total spent and their Country

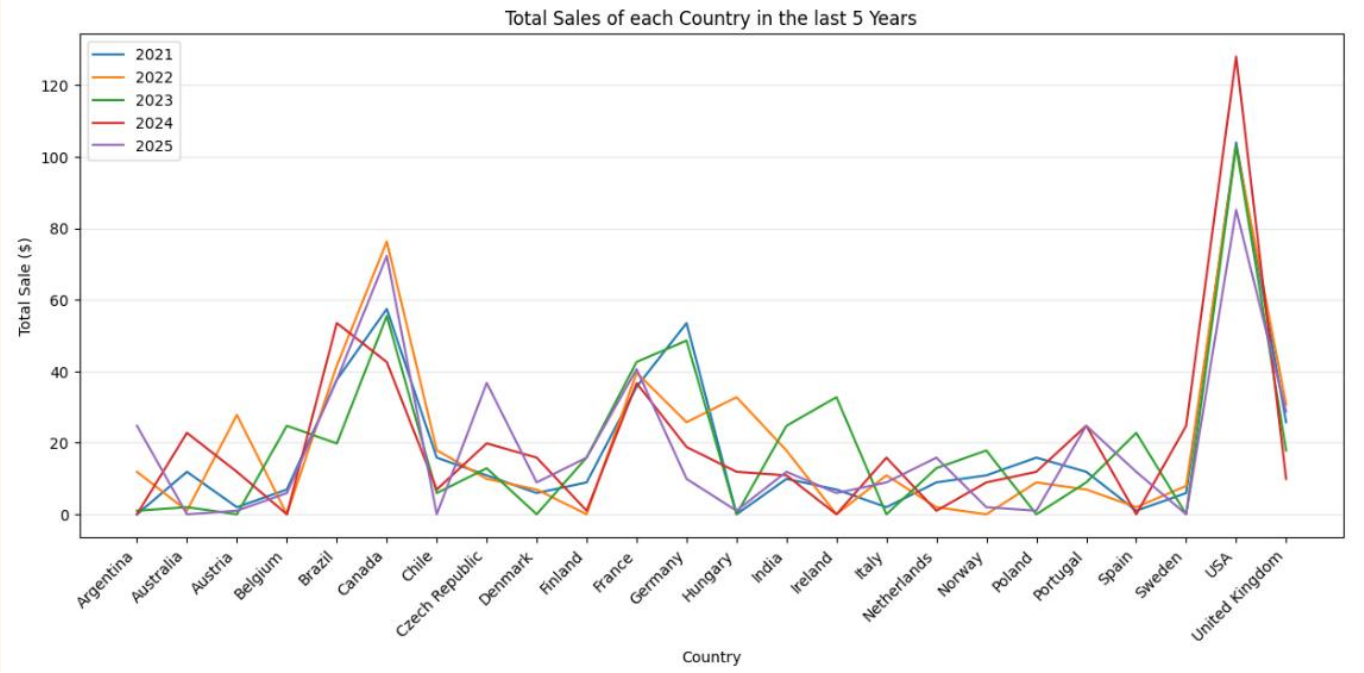
Kruskal-Wallis Test

Null hypothesis: No difference in the Total spending of customers across different countries.
Alternative hypothesis: Significant difference in the Total spending of customers across different countries

Fail to reject the null hypothesis.
There is no significant difference in the Total spending of customers across different countries.

★ Hypothesis Test

5) Dependency of the Customer's Total spent and their Country



★ Hypothesis Test

6) Dependency of track Media Types and track Size

Kruskal-Wallis Test

Null hypothesis: No difference in track Size between different Media Types.
Alternative hypothesis: Significant difference in track Size between different Media Types.

Reject the null hypothesis.
There is a significant difference in track Size between different Media Types.

Hypothesis Test

7) Dependency of Support Representative and Customer Total spent

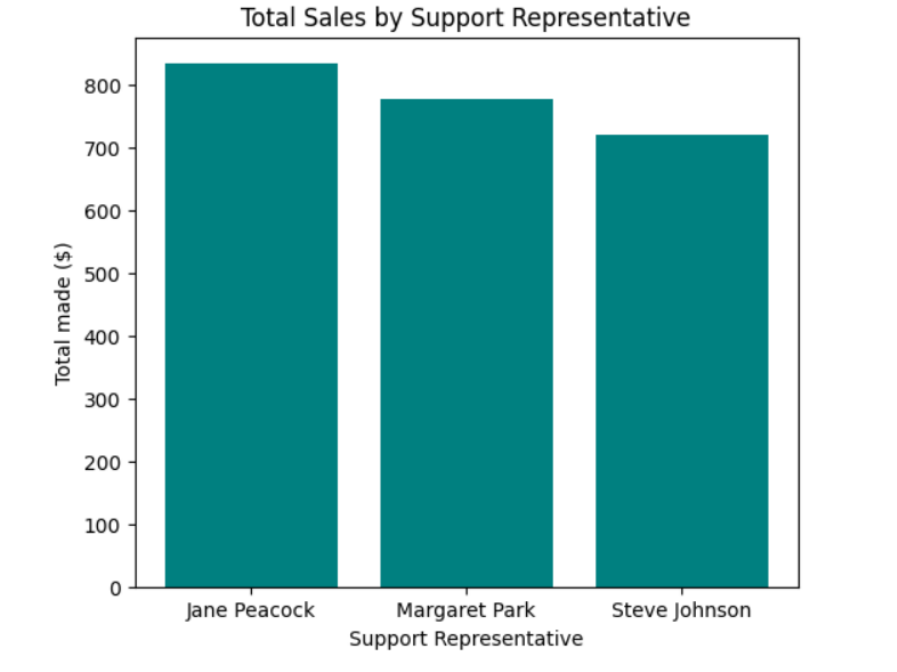
Kruskal-Wallis Test

Null hypothesis: No difference in customer Total spending between different Support Representatives.
Alternative hypothesis: Significant difference in customer Total spending between different Support Representatives.

Fail to reject the null hypothesis.
There is no significant difference in customer Total spending between different Support Representatives.

★ Hypothesis Test

7) Dependency of Support Representative and Customer Total spent



08 ★

Confidence Interval

Confidence Interval

1) Average Duration of Tracks in different genres

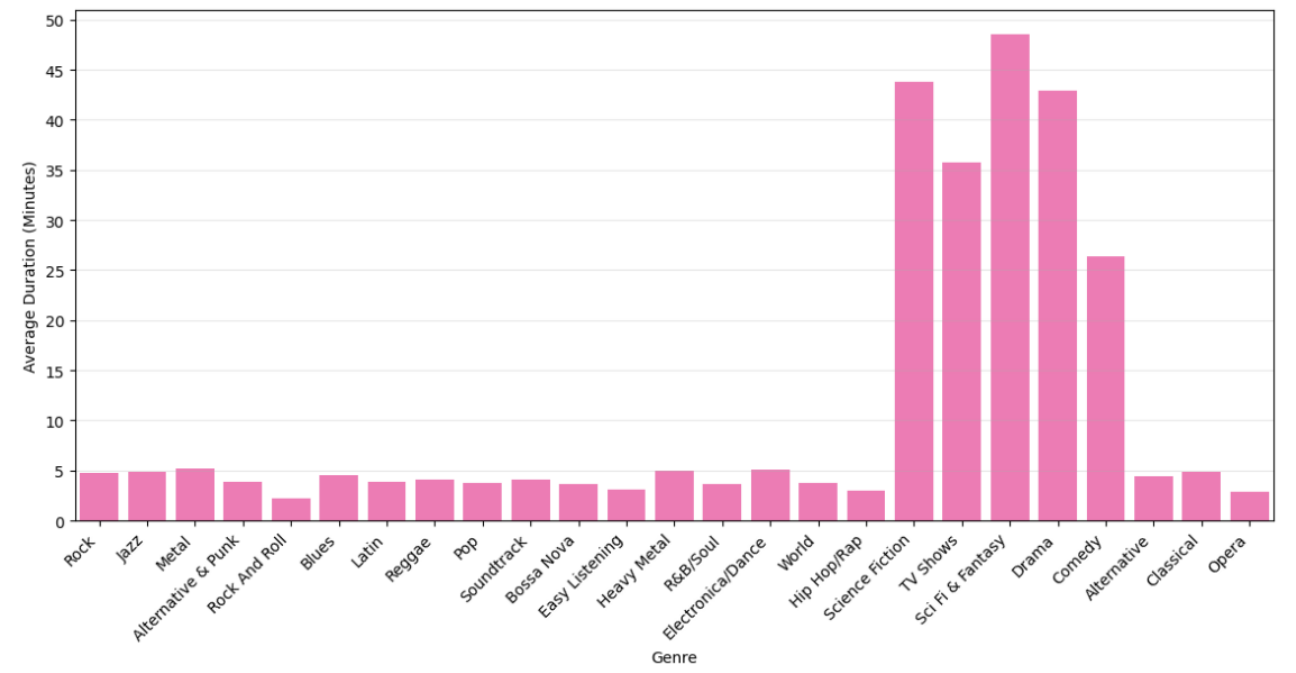
Kruskal-Wallis Test

Null hypothesis: No difference in Track durations between different Genres.
Alternative hypothesis: Significant difference in Track durations between different Genres.

Reject the null hypothesis.
There is a significant difference in Track durations between different Genres.

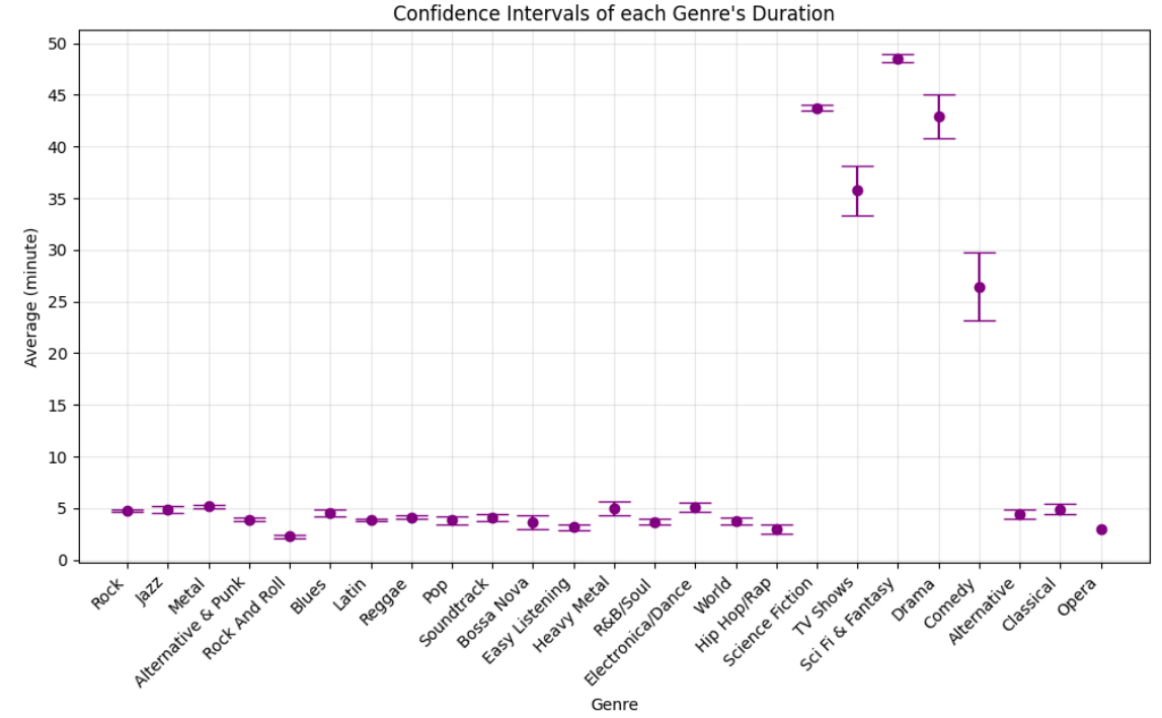
Confidence Interval

1) Average Duration of Tracks in different genres



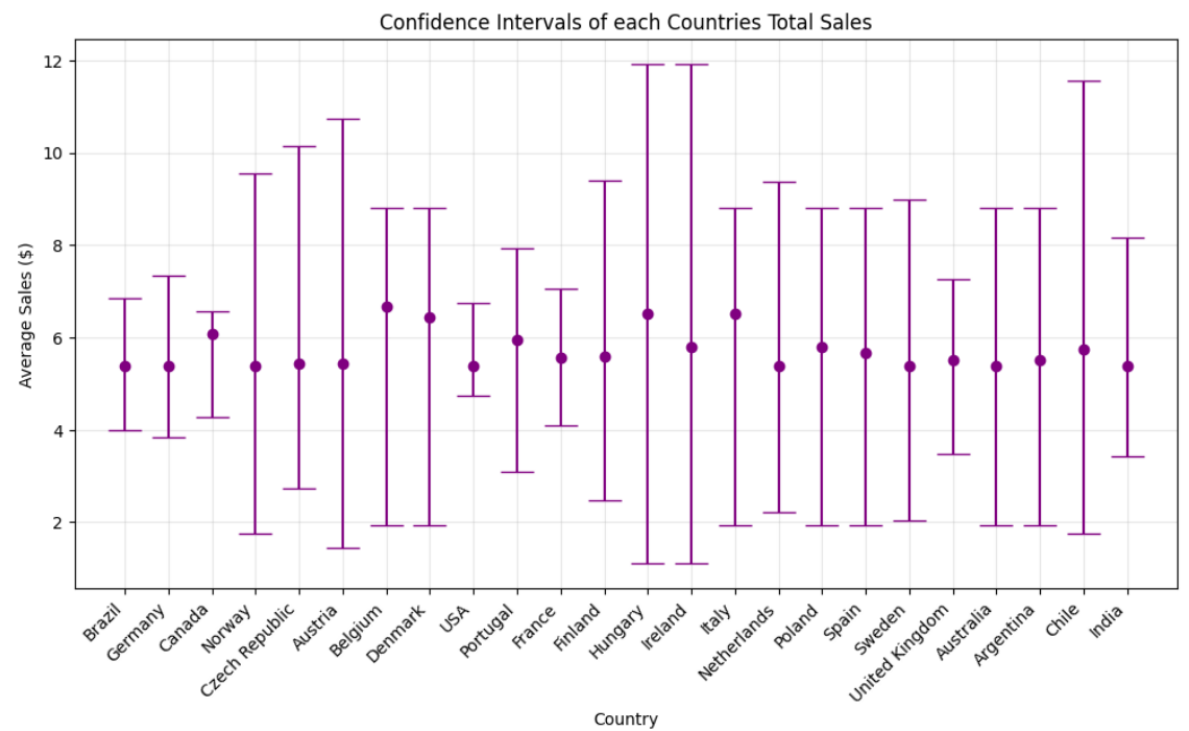
Confidence Interval

1) 95% Confidence Interval of genre durations



Confidence Interval

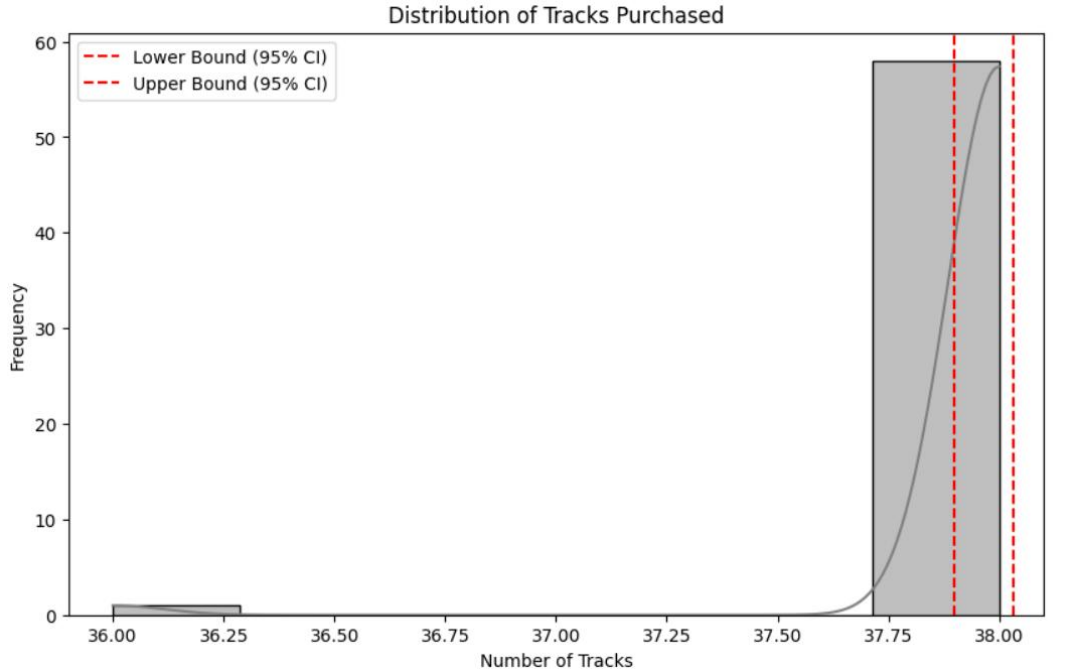
2) 95% Confidence Interval of different countries sales

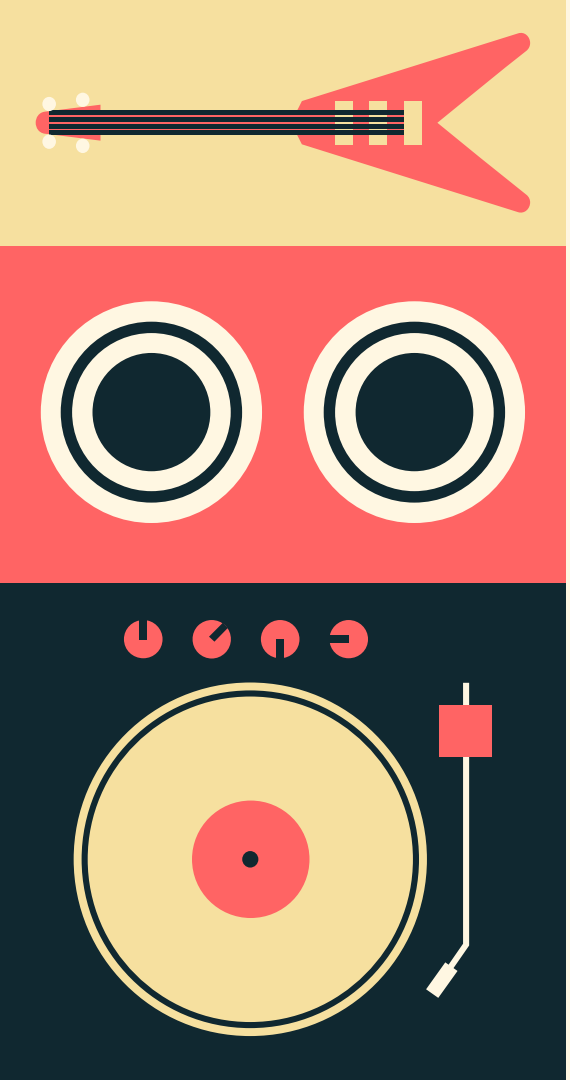


Confidence Interval

3) Average number of tracks purchased by each customer

The average number of tracks purchased by each customer is 38 tracks.





THANKS!