EXPLORATORY DATA ANALYSIS

Chinook Digital Media Store

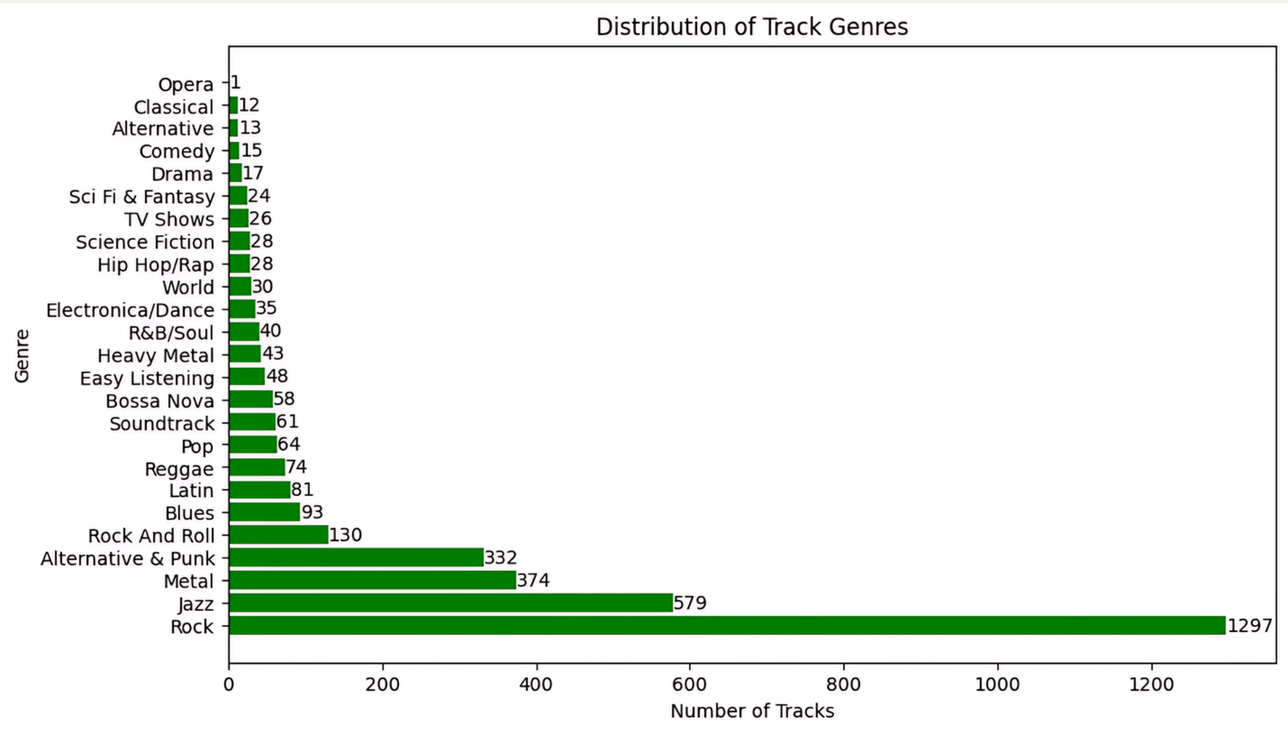
Mahlagha Sadoughi| 2024

INTRODUCTION

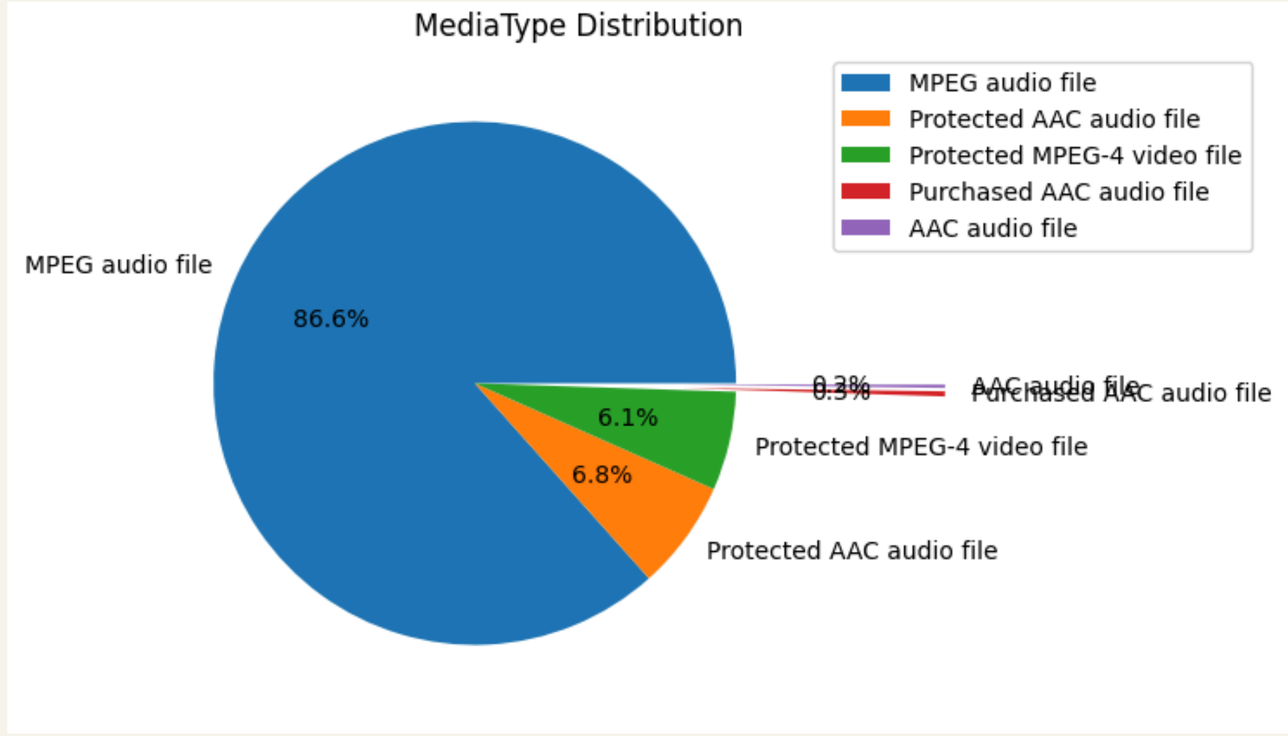
Chinook is a digital Media Store
Contain Information About Artists, Albums, Tracks, Playlists, Geres, MediaTypes, Customers, Employees and Invioces.
EDA Started after connecting MySQL to Python and creating dataframes.
In the Second Step: Null and Duplicated values checked

CATEGORICAL VARIABLES

GENRE TYPE



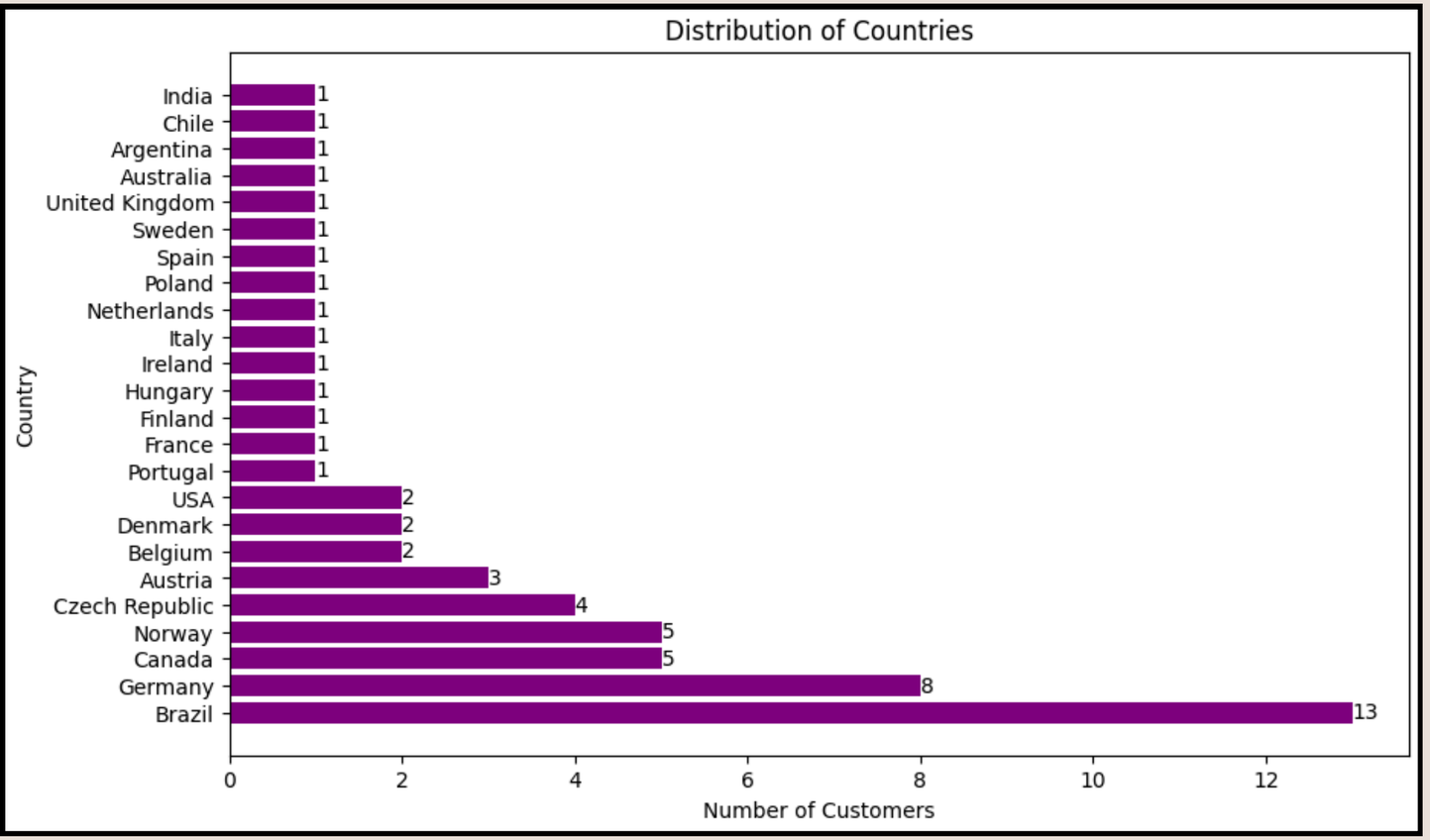
MEDIA TYPE



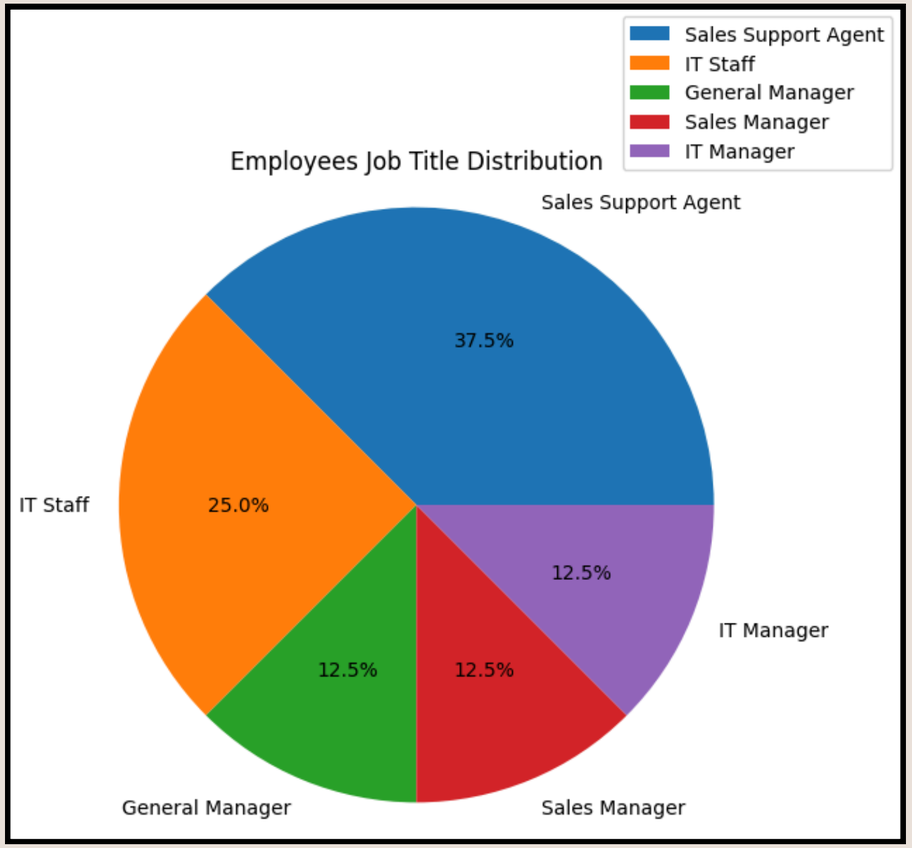
Key Features

CATEGORICAL VARIABLES

CUSTOMER COUNTRY



EMPLOYEE JOB TITLE



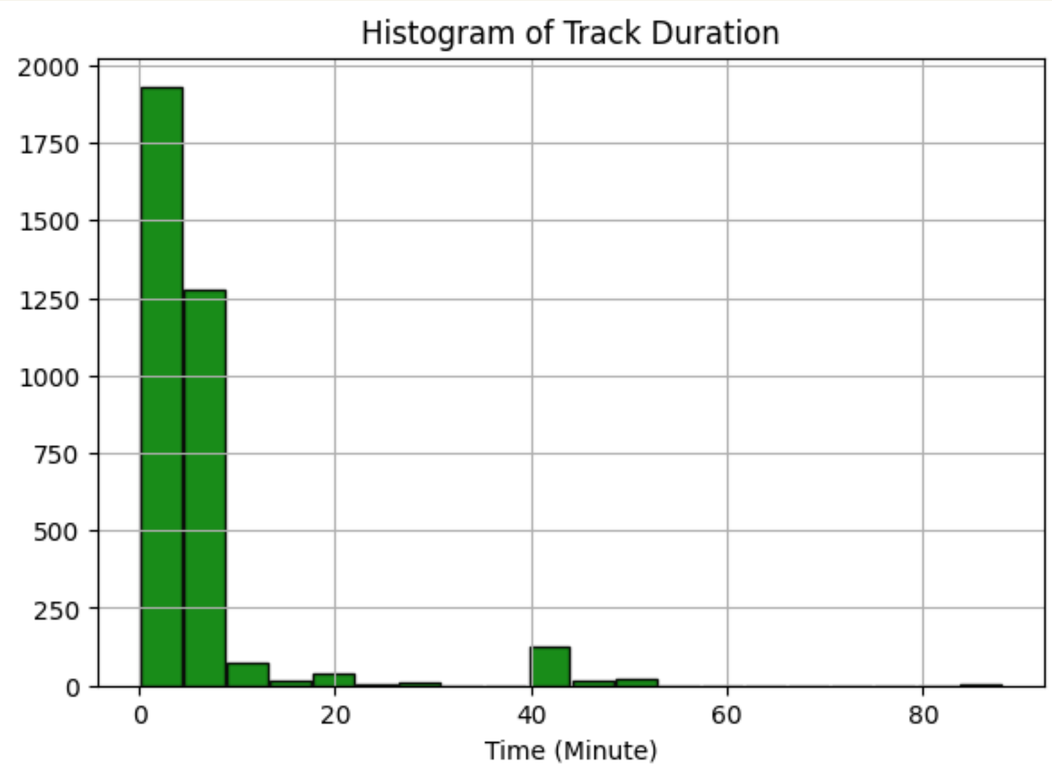
Key Features

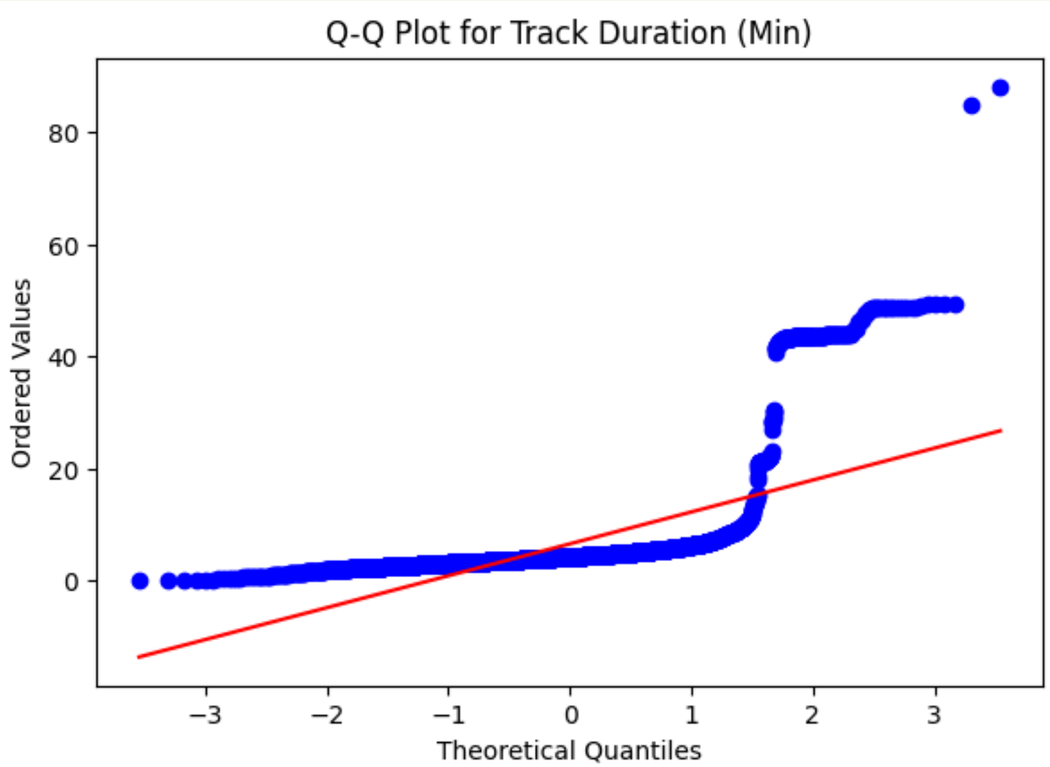
NUMERICAL VARIABLES
TRACK DURATION

Shapiro-Wilk test

H0: Normal
HI: Not Normal

p-value ~ 0
Result: Not Normal

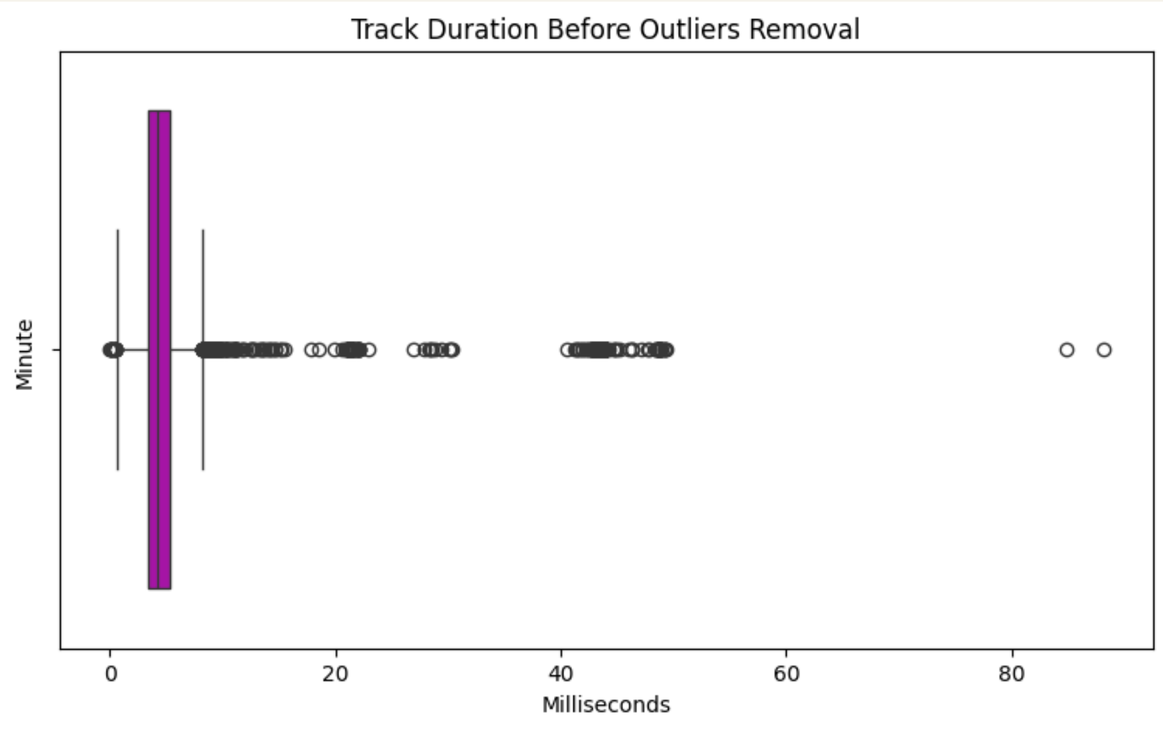


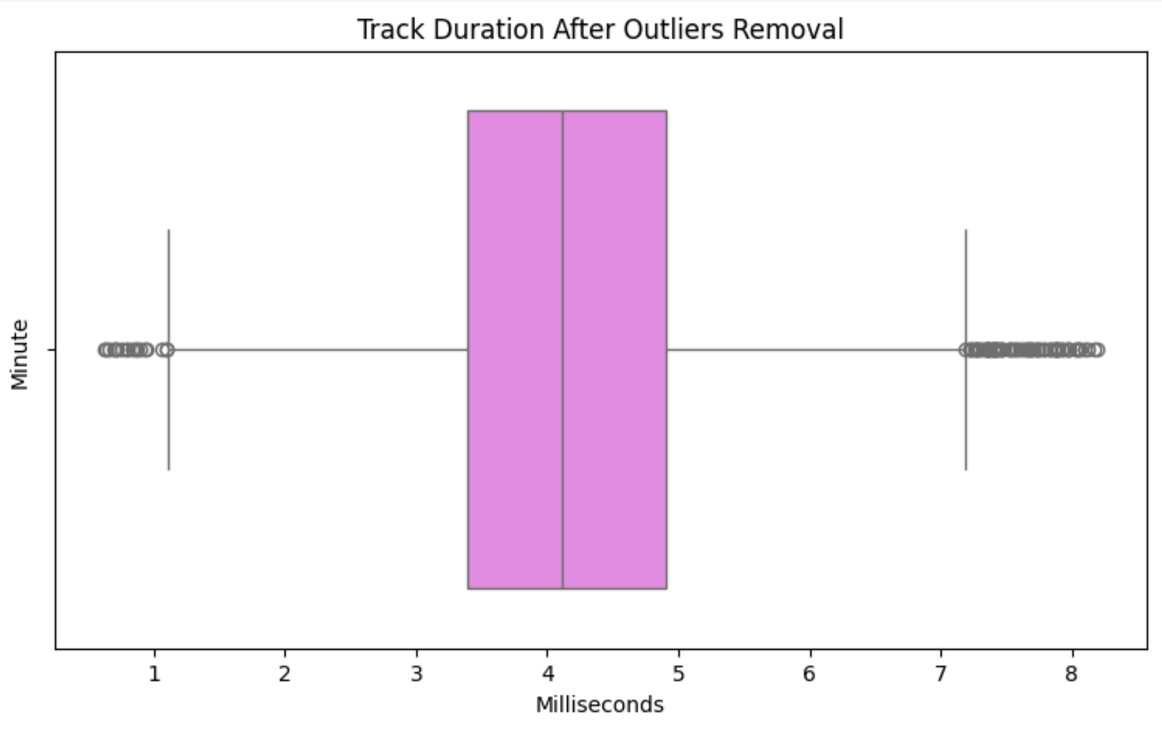


key Features

TRACK DURATION

Max= 88.11 (min)	Q1= 3.45	Q3= 5.3	Upper Bound: 8.21
Min= 0.017 (min)	Q2= 4.2	IQR= 1.9	Lower Bound: 0.59





Numerical variables

TRACK SIZE

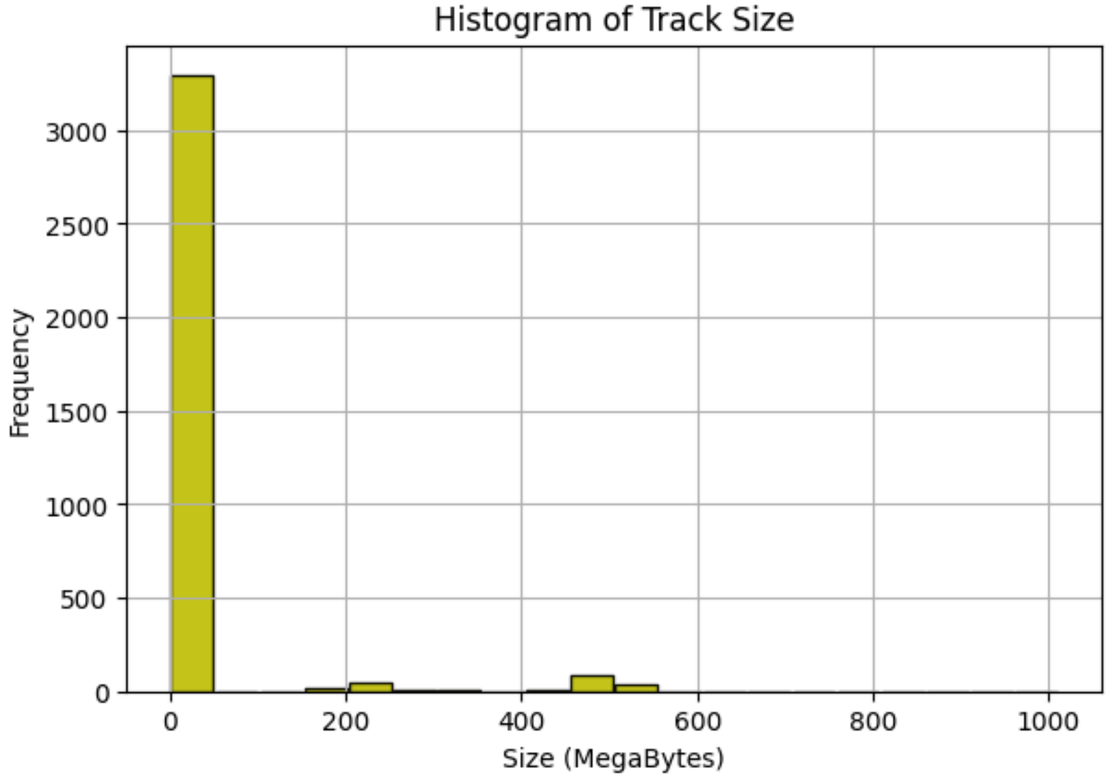
Shapiro-Wilk test

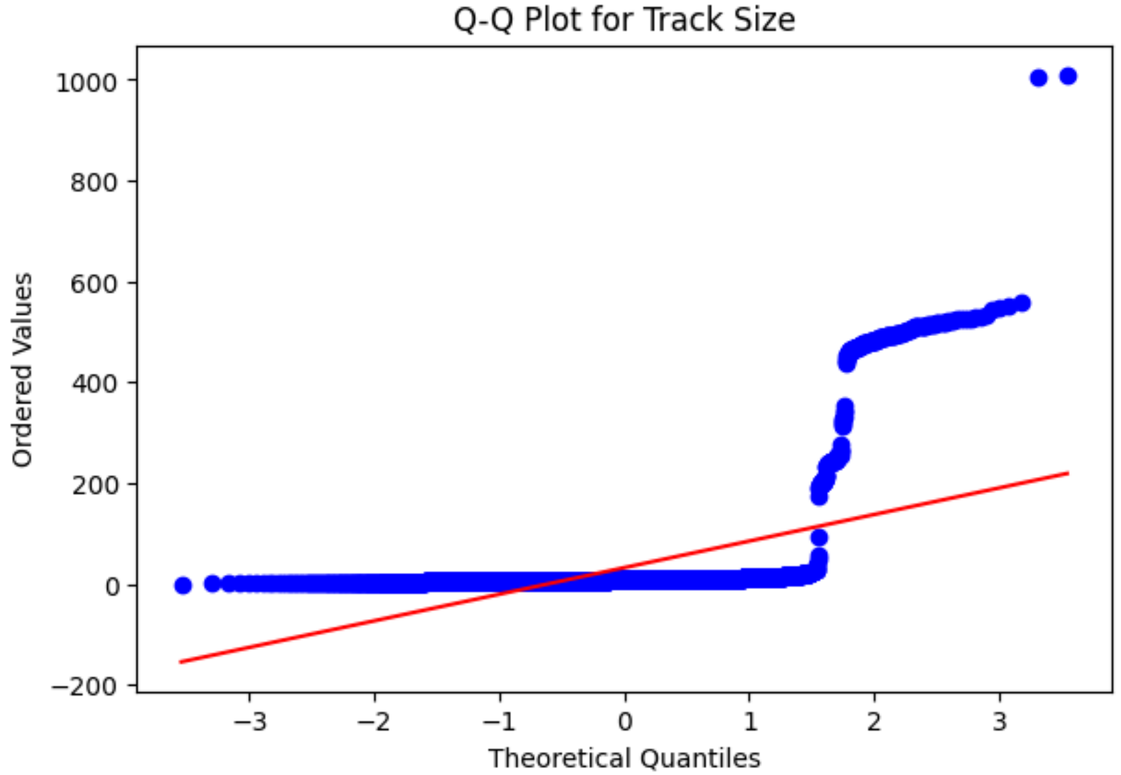
H0: Normal

HI: Not Normal

p Value ~ 0

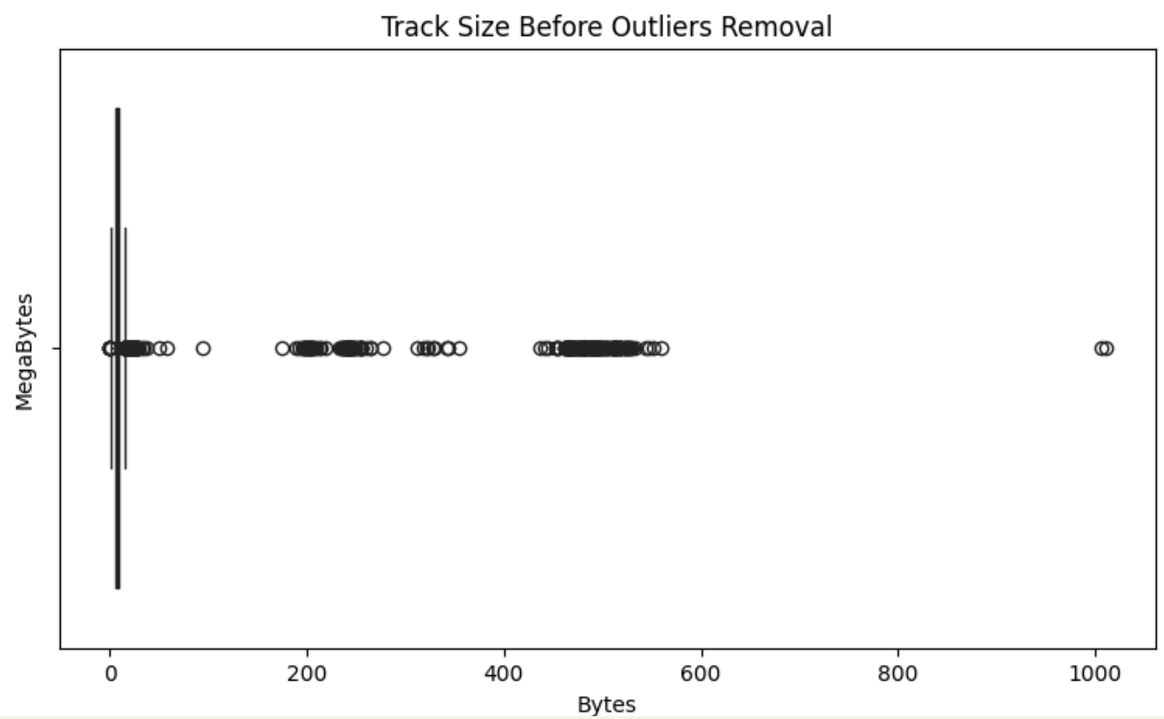
Result: Not Normal

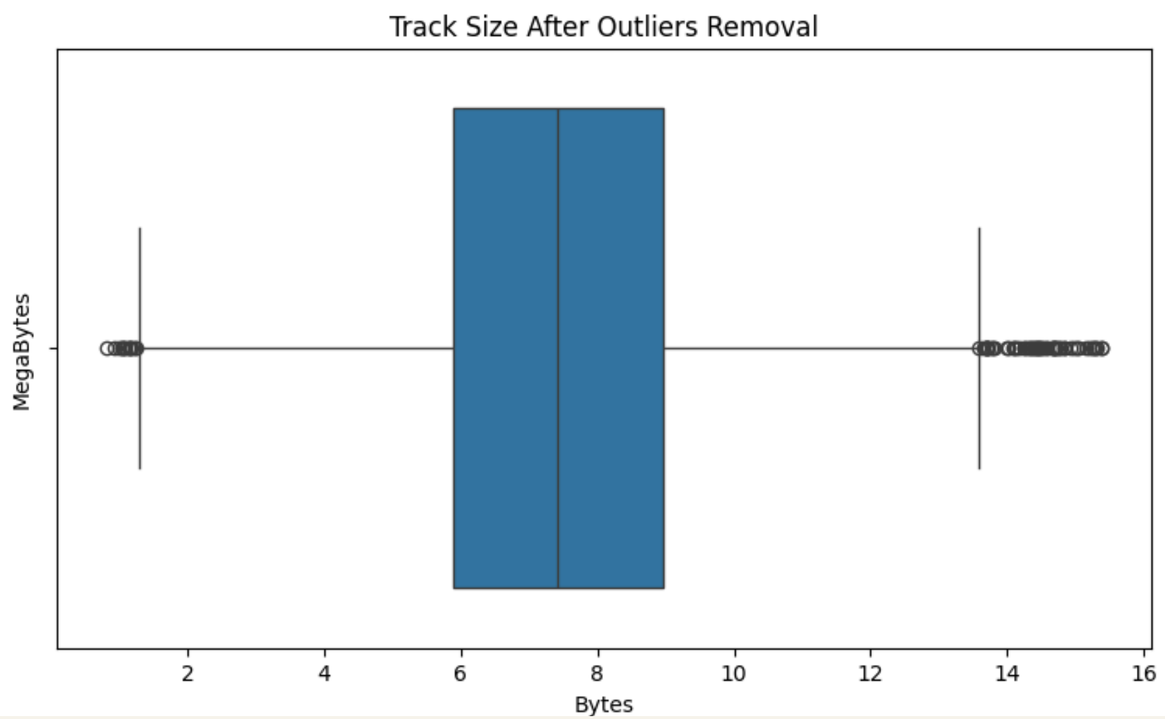




TRACK SIZE

Max= 1010.4 (MB)	Q1= 6.04	Q3= 9.79	Upper Bound: 15.4
Min= 0.037 (MB)	Q2= 7.73	IQR= 3.74	Lower Bound: 0.43





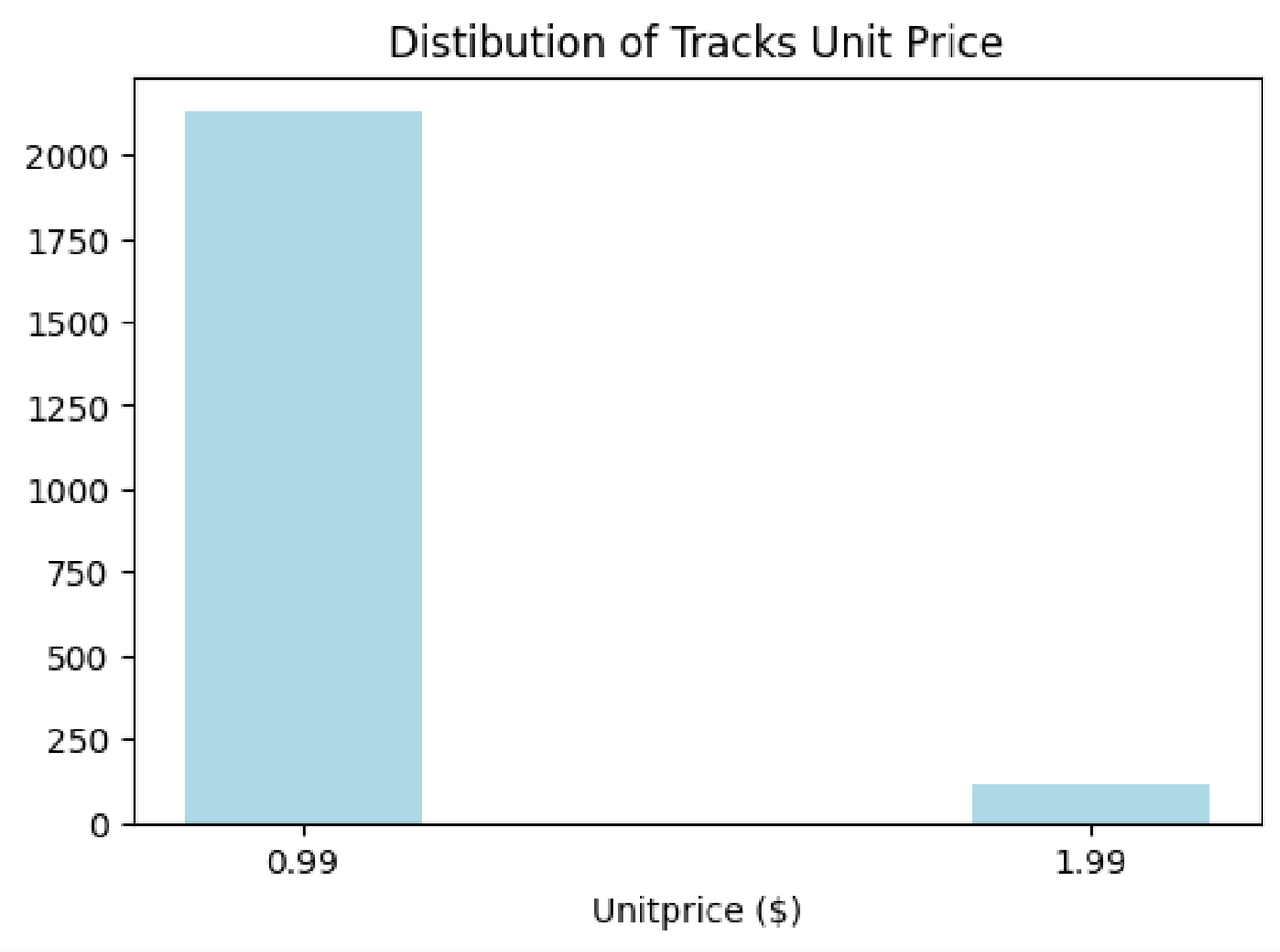
UNIT PRICE AND QUANTITY

- **Unit price**

  0.99 , 1.99

- **Quantity**

  one



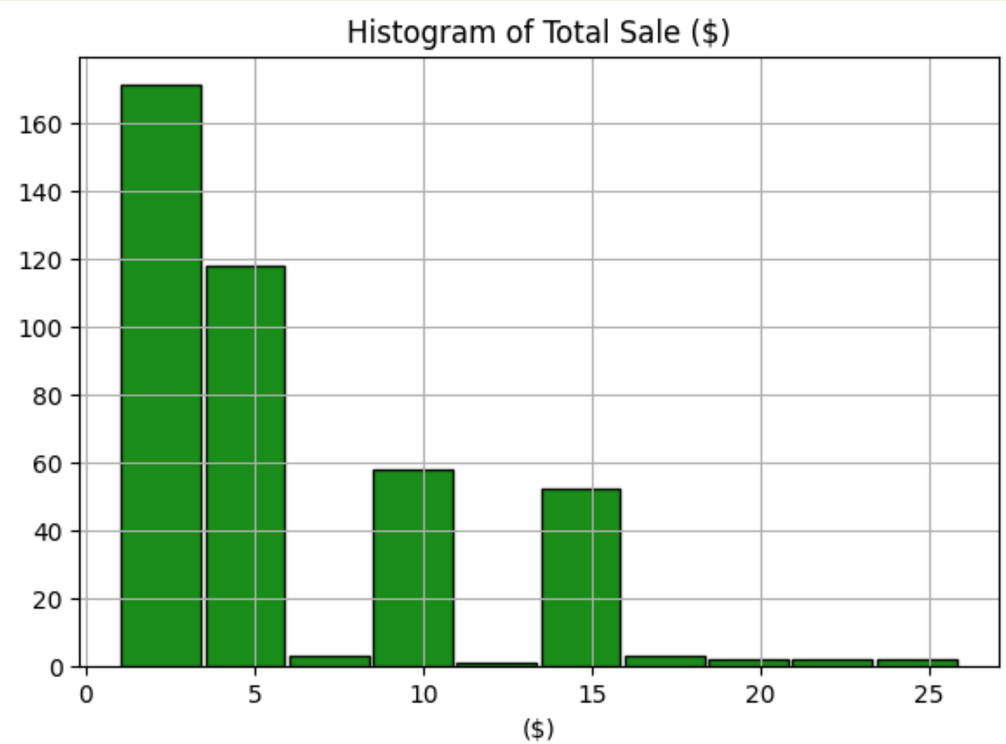
TOTAL SALE

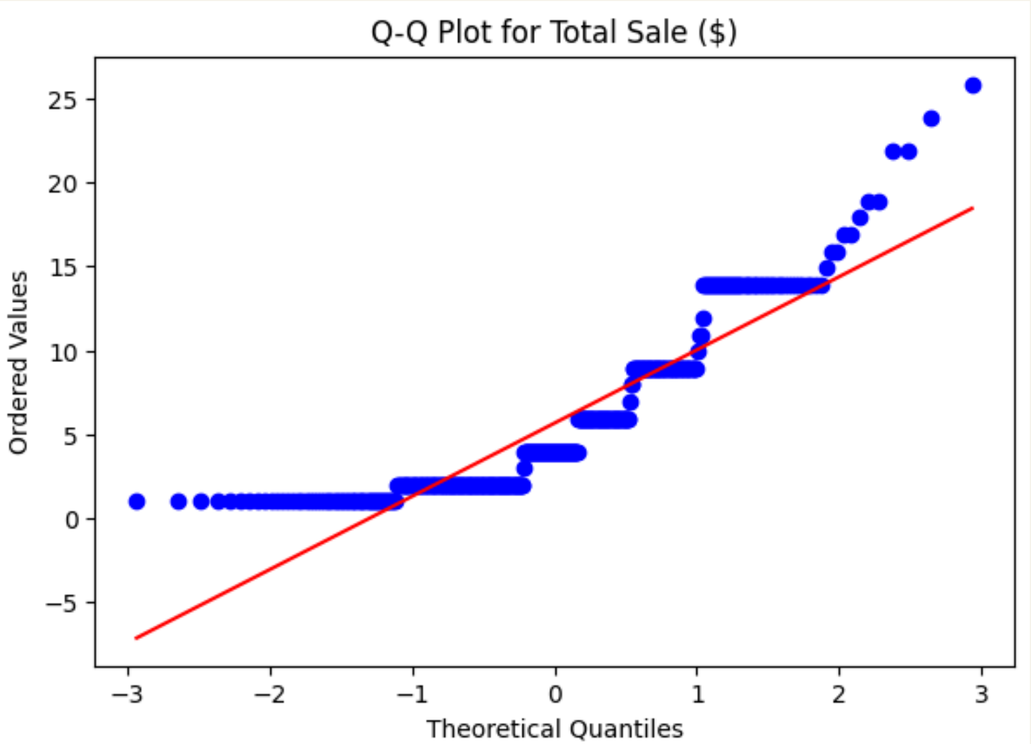
Shapiro-Wilk test

H0: Normal
HI: Not Normal
p Value ~ 0
Result: Not Normal

D'Agostino k^2 test

H0: Normal
HI: Not Normal
p Value~ 0
Result: Not Normal





Track Size

STATISTICAL HYPOTHESIS TESTING

Top 3 Popular genres

Genre Name	Total sale
Rock	826.65
Latin	382.14
Metal	261.36

Difference in Average Price

One Way ANOVA Test

H0: Average Price Similarity
HI: Different Average price
p Value: Nan
Result: Can not perform ANOVA test, Because one or more genres have constant Value

STATISTICAL HYPOTHESIS TESTING

Correlation Between Track Duration and price

Spearman Correlation

H0: Not Related
HI: Related
p Value ~ 0
Result: Monotonic Relationship

Correlation Between Genre and Media Type

Chi-Square Test

H0: Not Related
HI: Related
p Value ~ 0
Result: Dependent

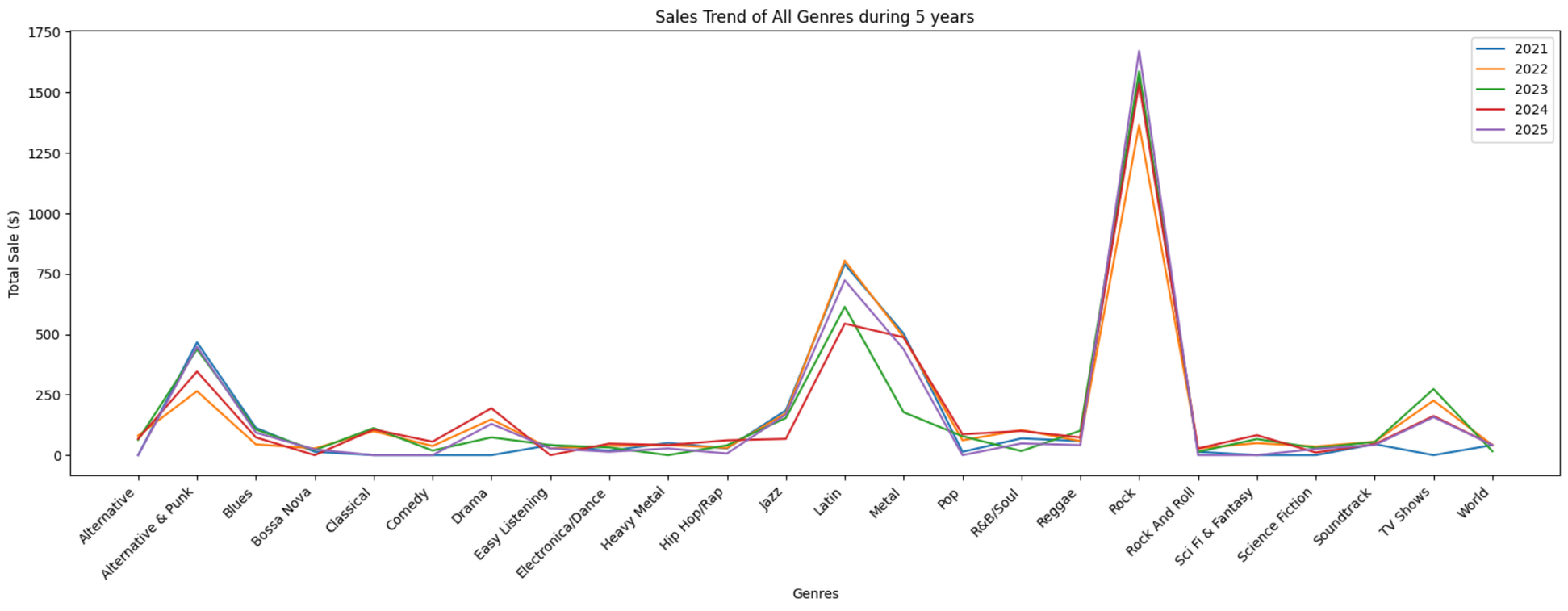
STATISTICAL HYPOTHESIS TESTING

Correlation Between Total Sale and date

Spearman Correlation

H0: Not Related **p Value ~ 0**

HI: Related **Result: No Relation**



STATISTICAL HYPOTHESIS TESTING

Correlation Between Media Type and Track Size

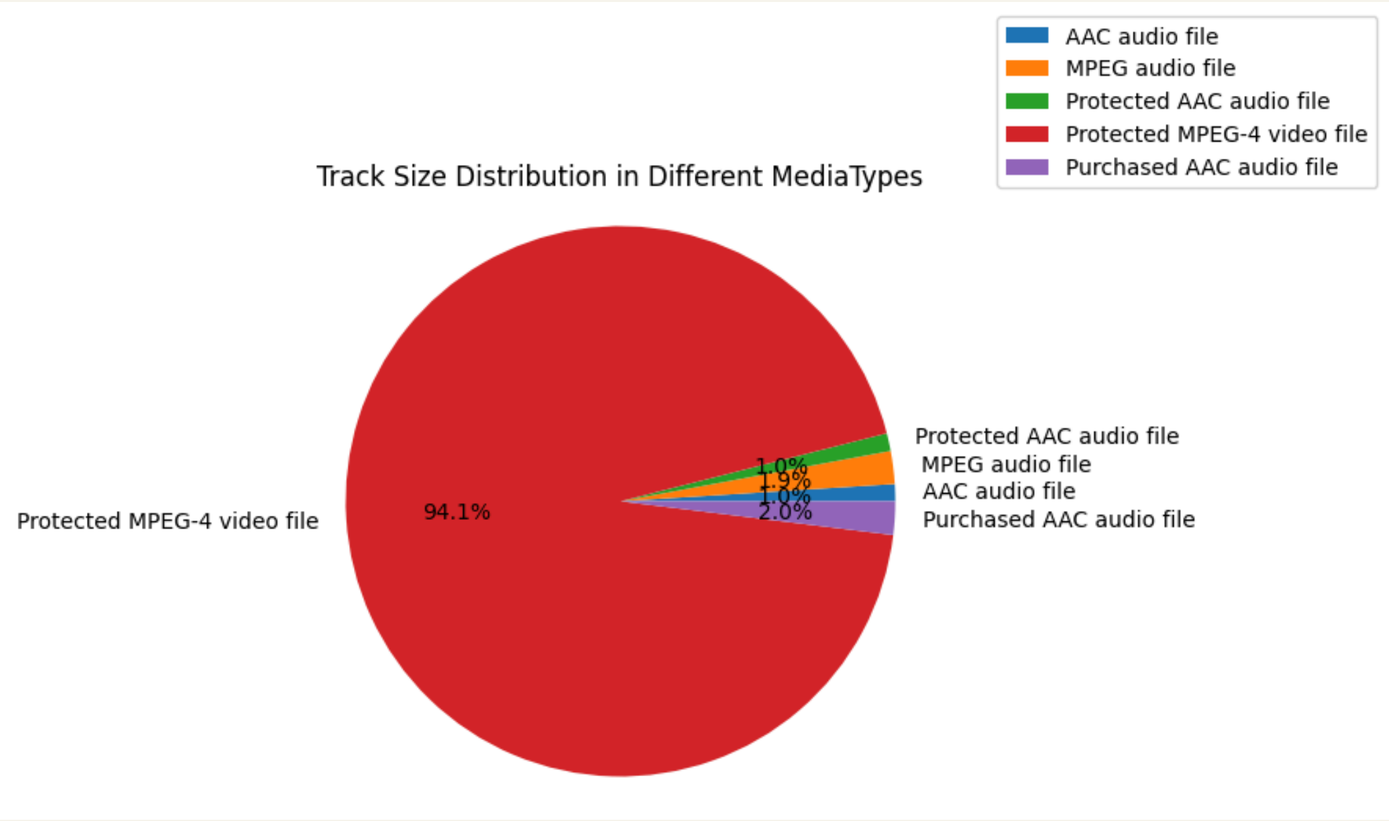
Kendal Rank Correlation

H0: Not Related

HI: Related

p Value ~ 0

Result: There is a Relationship

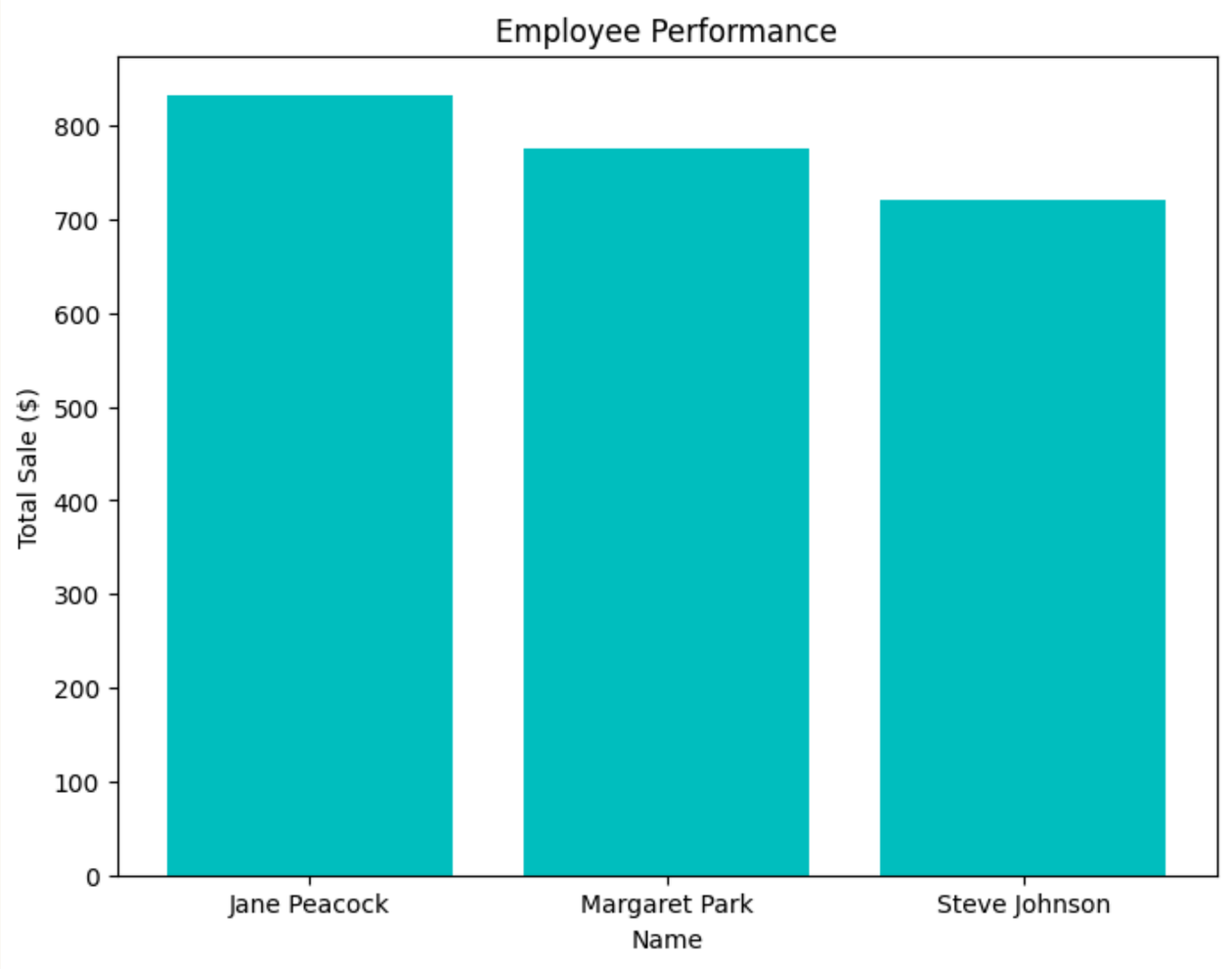


STATISTICAL HYPOTHESIS TESTING

Difference in Sales Between Employees

One Way ANOVA

H0: No Difference
HI: Difference
p Value: 0.94
Result: No different



STATISTICAL HYPOTHESIS TESTING

The average length of Tracks in different genres

95% confidence interval

Kruskal-Wallis Test

H0: Similar
HI: Difference
p Value: 0.46
Result: Different

Name	Mean	Count	STD	Lower Bound	Upper bound
Sci Fi & Fantasy	48.53	26	1.05	48.1	48.9
Sci Fi	43.76	13	0.52	43.4	44
Drama	42.92	64	8.63	40.8	45
TV Shows	35.75	93	11.87	33.3	38.1
Comedy	26.42	17	6.96	23.1	29.7

STATISTICAL HYPOTHESIS TESTING

The average sales in different countries

95% confidence interval

Cuntry	Mean	Coun t	STD	Lower Bound	Upper bound
Chile	6.66	7	6.61	1.76	11.56
Hungry	6.51	7	7.29	1.10	11.92
Ireland	6.51	7	7.29	1.10	11.92
Czech Republic	6.44	14	7.06	2.74	10.15
Austria	6.09	7	6.26	1.44	10.73

STATISTICAL HYPOTHESIS TESTING

The average number of tracks purchased by each user

95% confidence interval

95% Confidence Interval :
35.96 - 39.97

CustomerId	Last name	Count
1	Goncalves	38
45	kovacs	38
33	Sullivan	38
34	Fernandes	38
35	Sampaio	38

THANK YOU

Mahlagha Sadoughi