Chinook database analysis

Niloofar Moosavi

Load DataFrames

DataFrames:

- df_album
- df_artist
- df_customer
- df_employee
- df_genre
- df_invoice
- df_invoiceline
- df_mediatype
- df_playlist
- df_playlisttrack
- df_track

```
conn = mysql.connector.connect(
    host = db_config['host'],
    port = db_config['port'],
    user = db_config['user'],
    password = db_config['password'],
    database = db_config['database'],
    auth_plugin='mysql_native_password'
)
```

```
query_album = """
select * from album
"""
```

Initial data review with pandas

Functions:

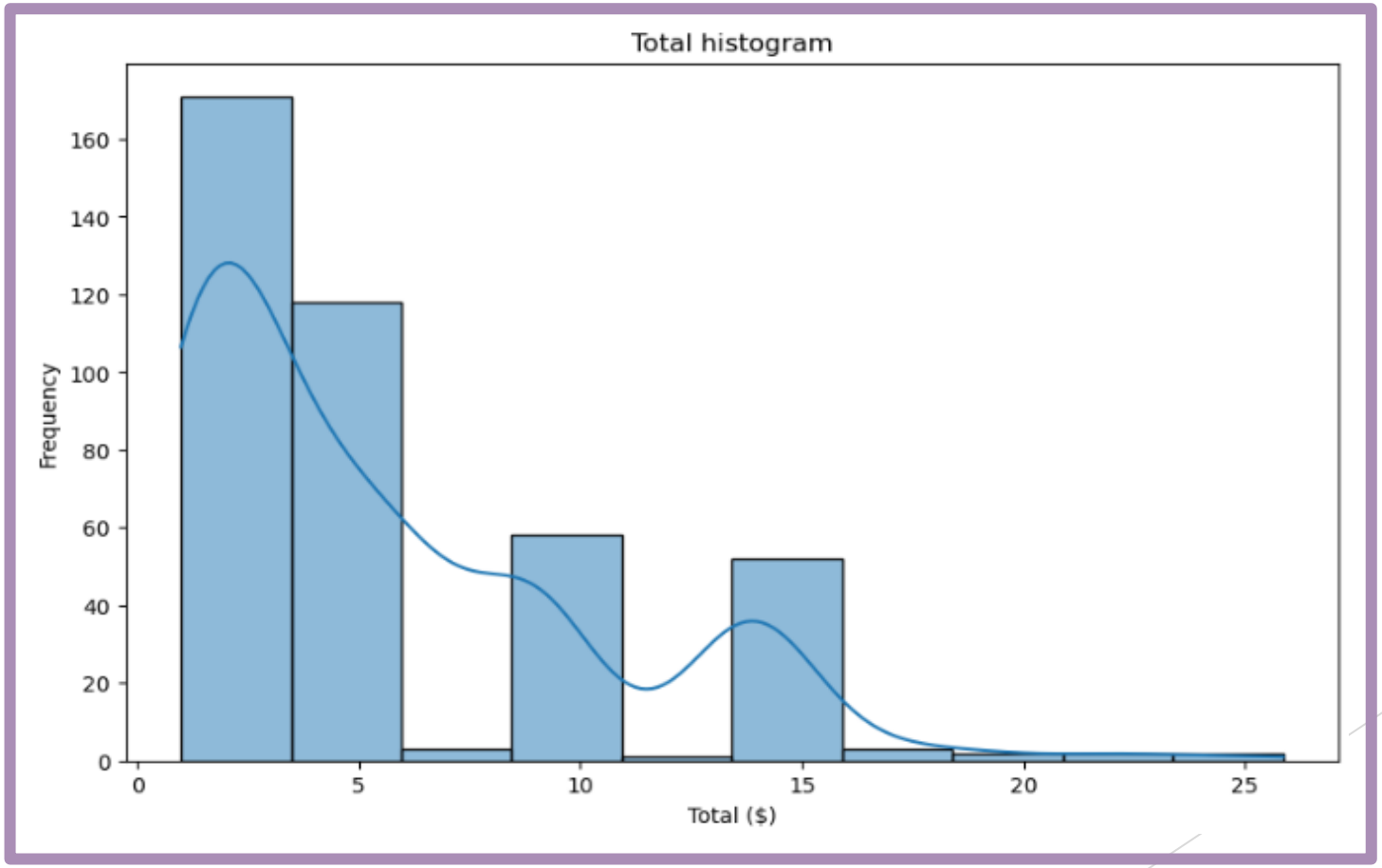
- Info ()
- Describe ()
- Duplicated ()
- Isnull ()

DataFrame	Num of row	Num of Column	Duplicated	Null
Album	347	3	0	0
Artist	275	2	0	0
Customer	59	13	0	Company: 49 State: 29 Fax: 47
Employee	8	15	0	ReportTo: 1
Genre	25	2	0	0
Invoice	412	9	0	BillingState: 202 BillingPostalCode: 28
Invoiceline	2240	5	0	0
MediaType	5	2	0	0
Playlist	18	2	0	0
Playlisttrack	8715	5	0	0
track	3503	9	0	Composer: 977

Key variables

numerical variables:

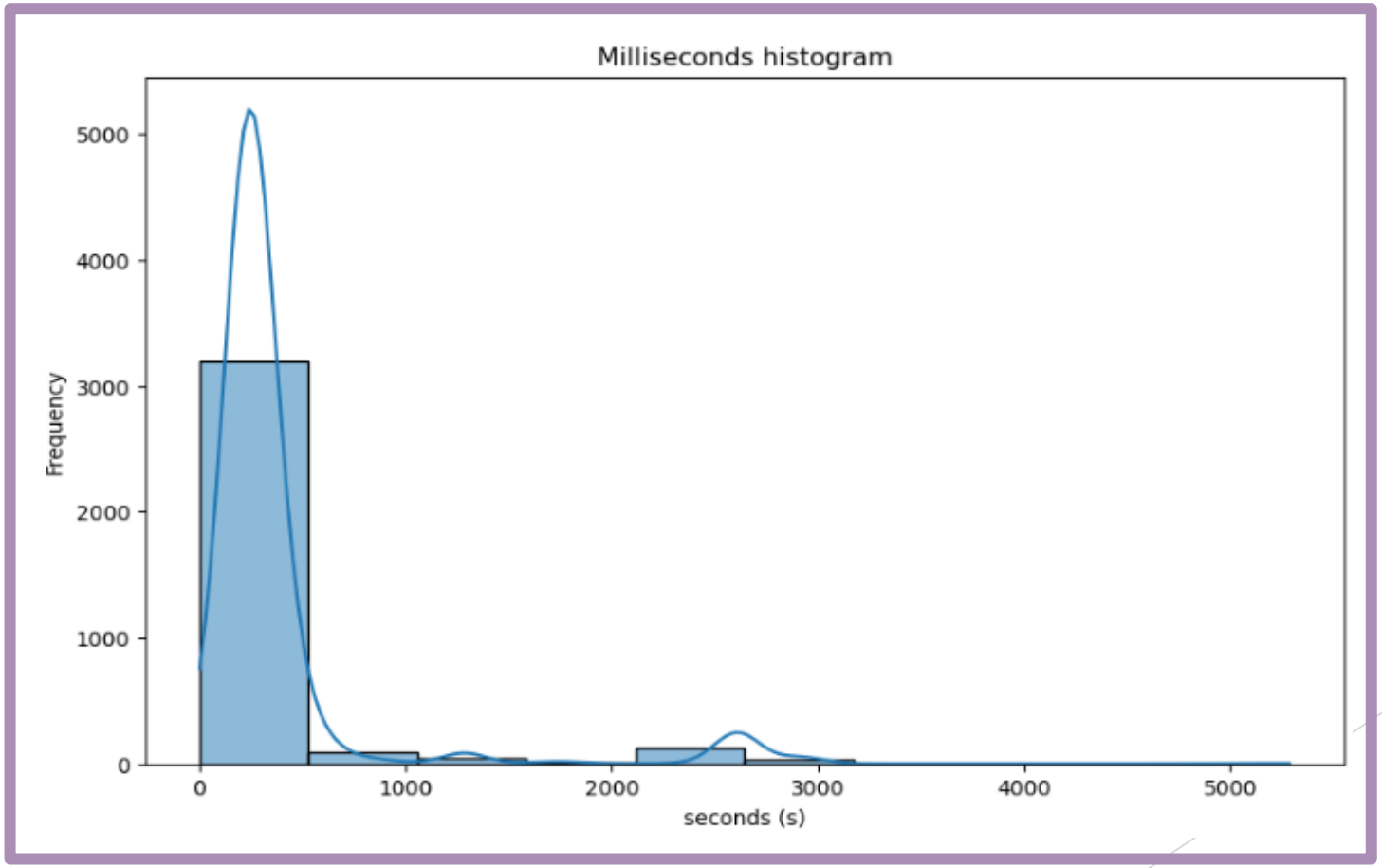
- Total



Key variables

numerical variables:

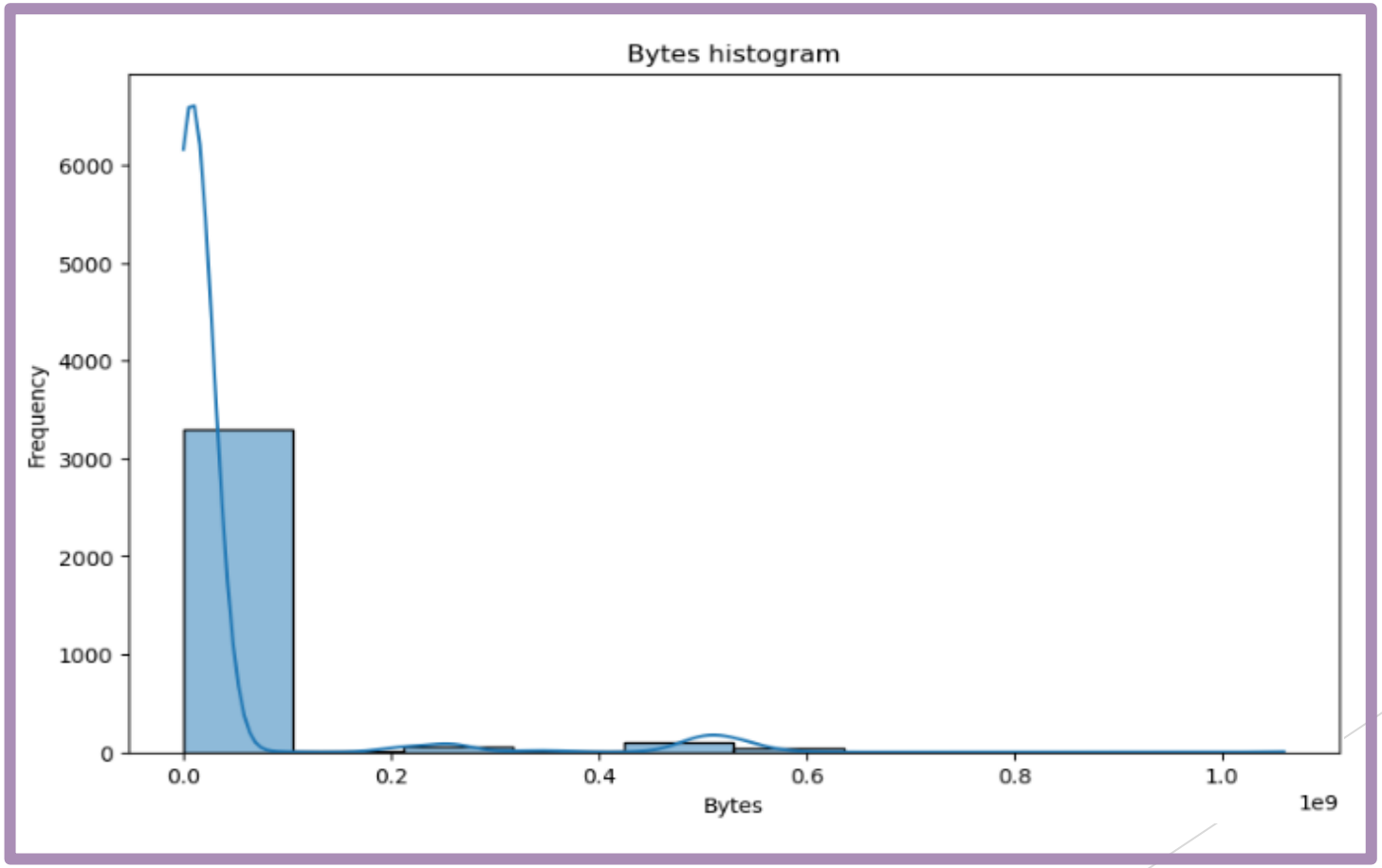
- Milliseconds



Key variables

numerical variables:

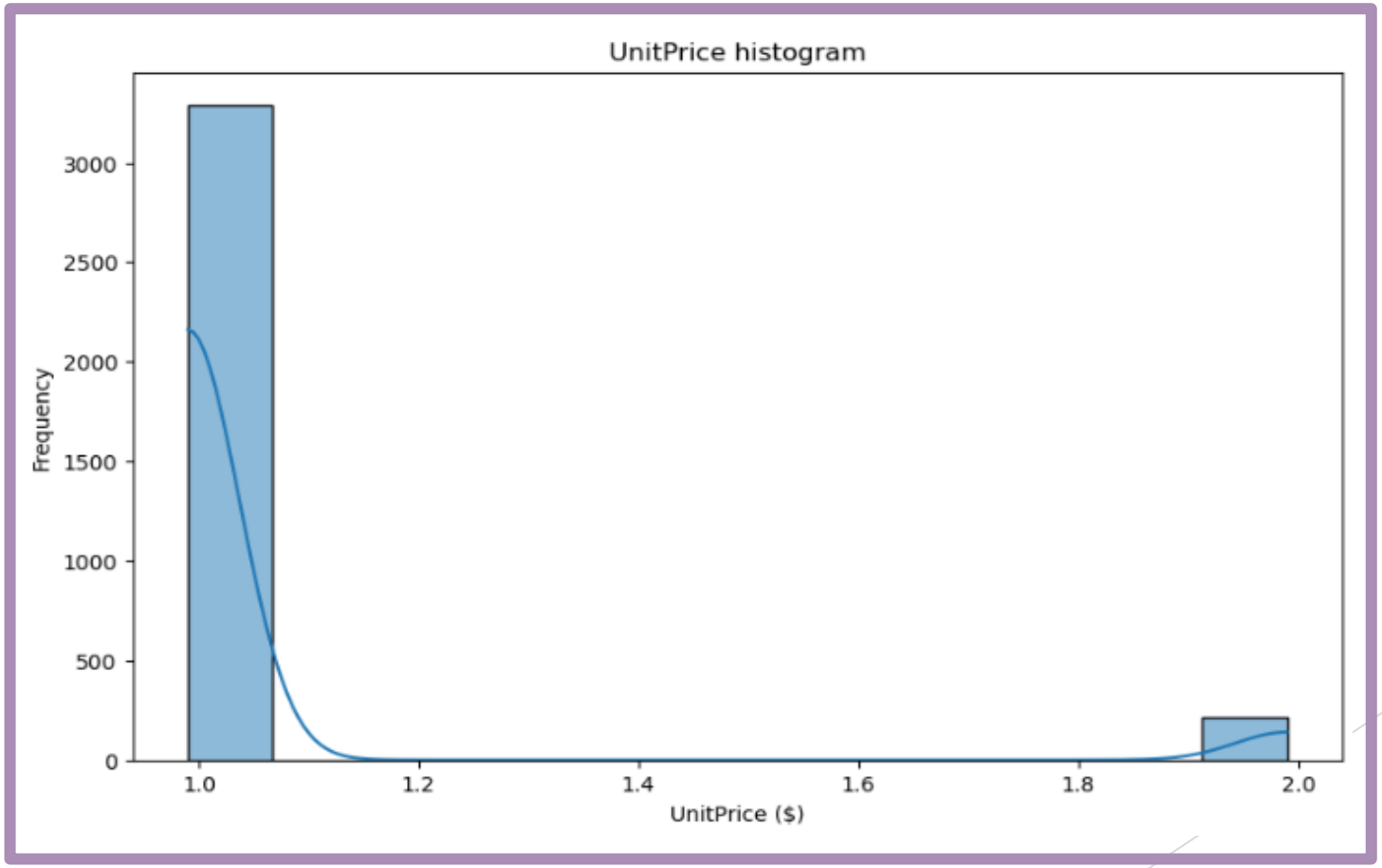
- Bytes



Key variables

numerical variables:

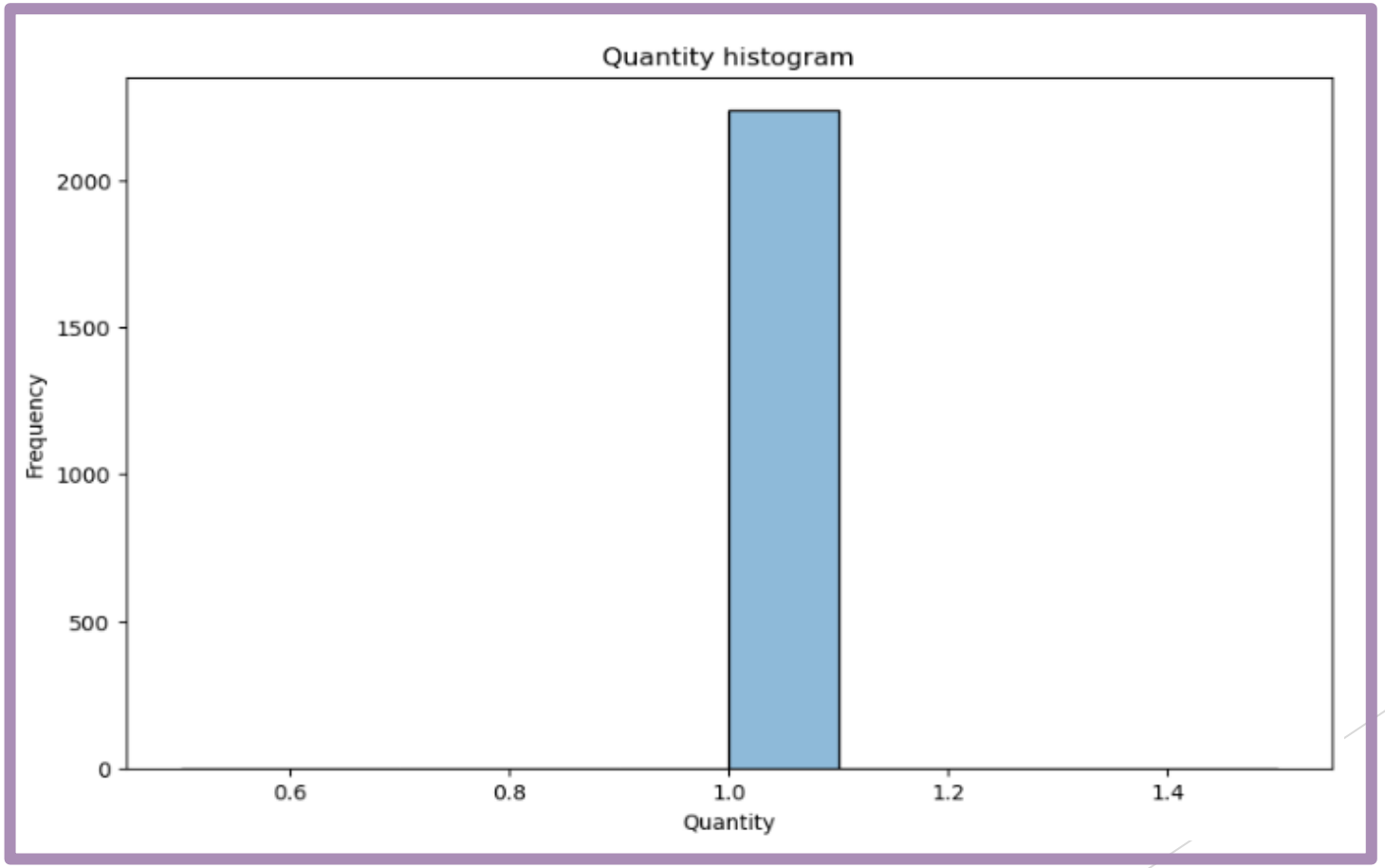
- UnitPrice



Key variables

numerical variables:

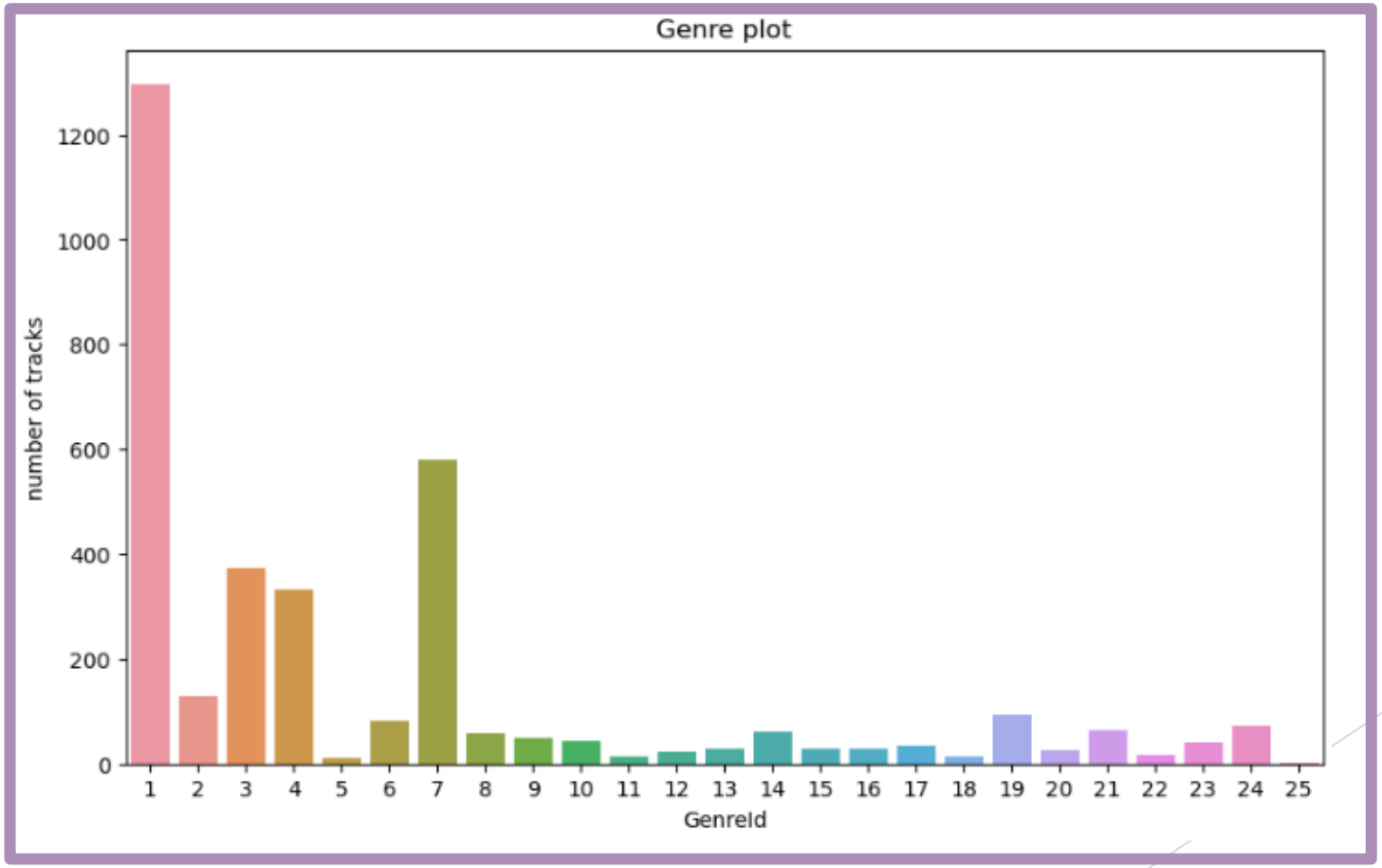
- Quantity



Key variables

Categorical variables:

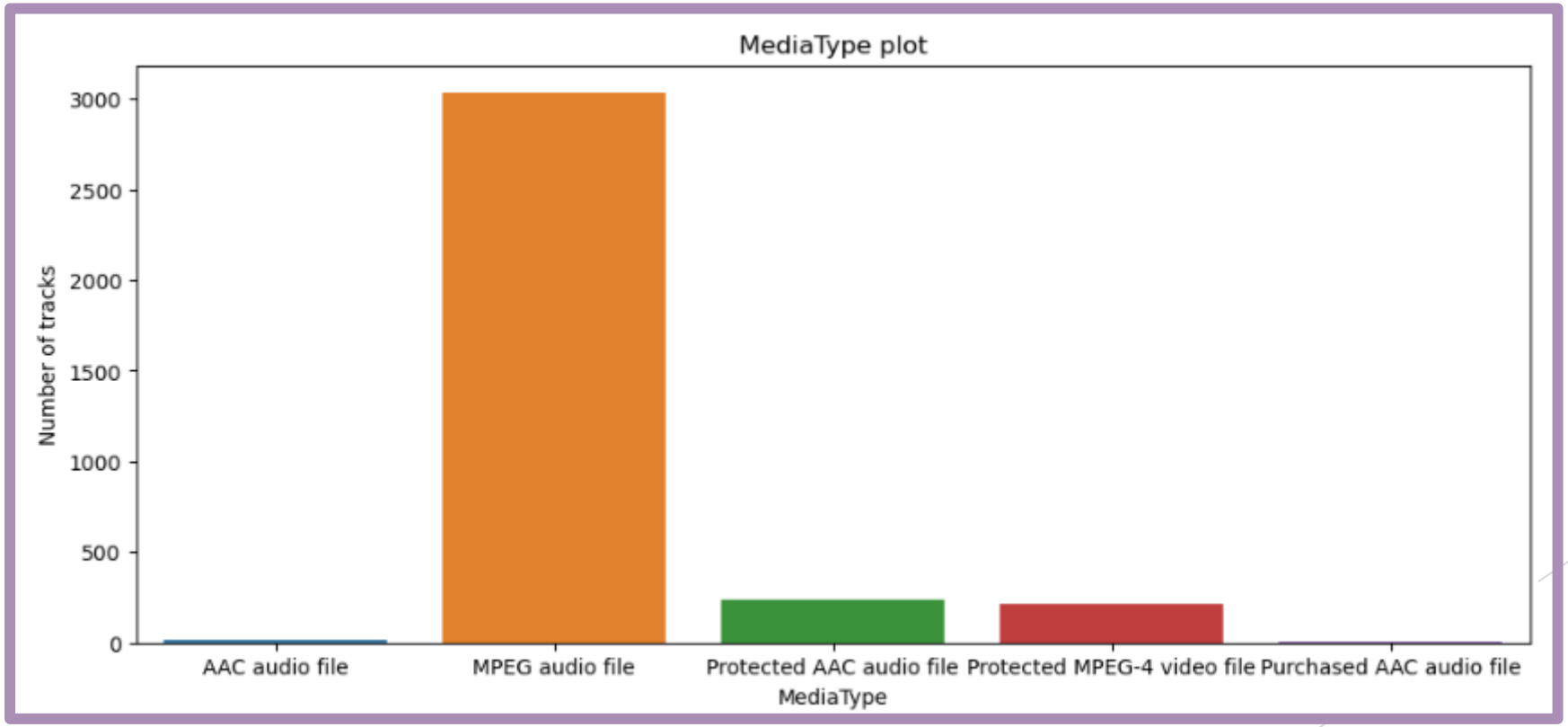
- GenreId



Key variables

Categorical variables:

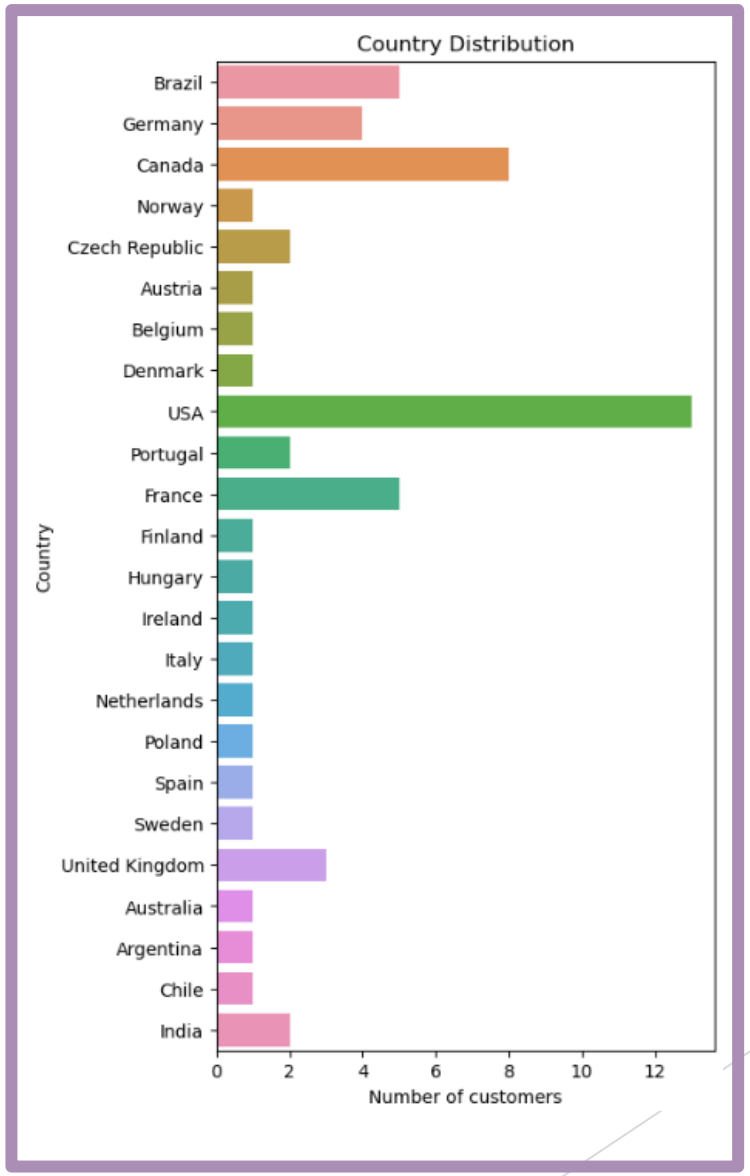
- MediaType



Key variables

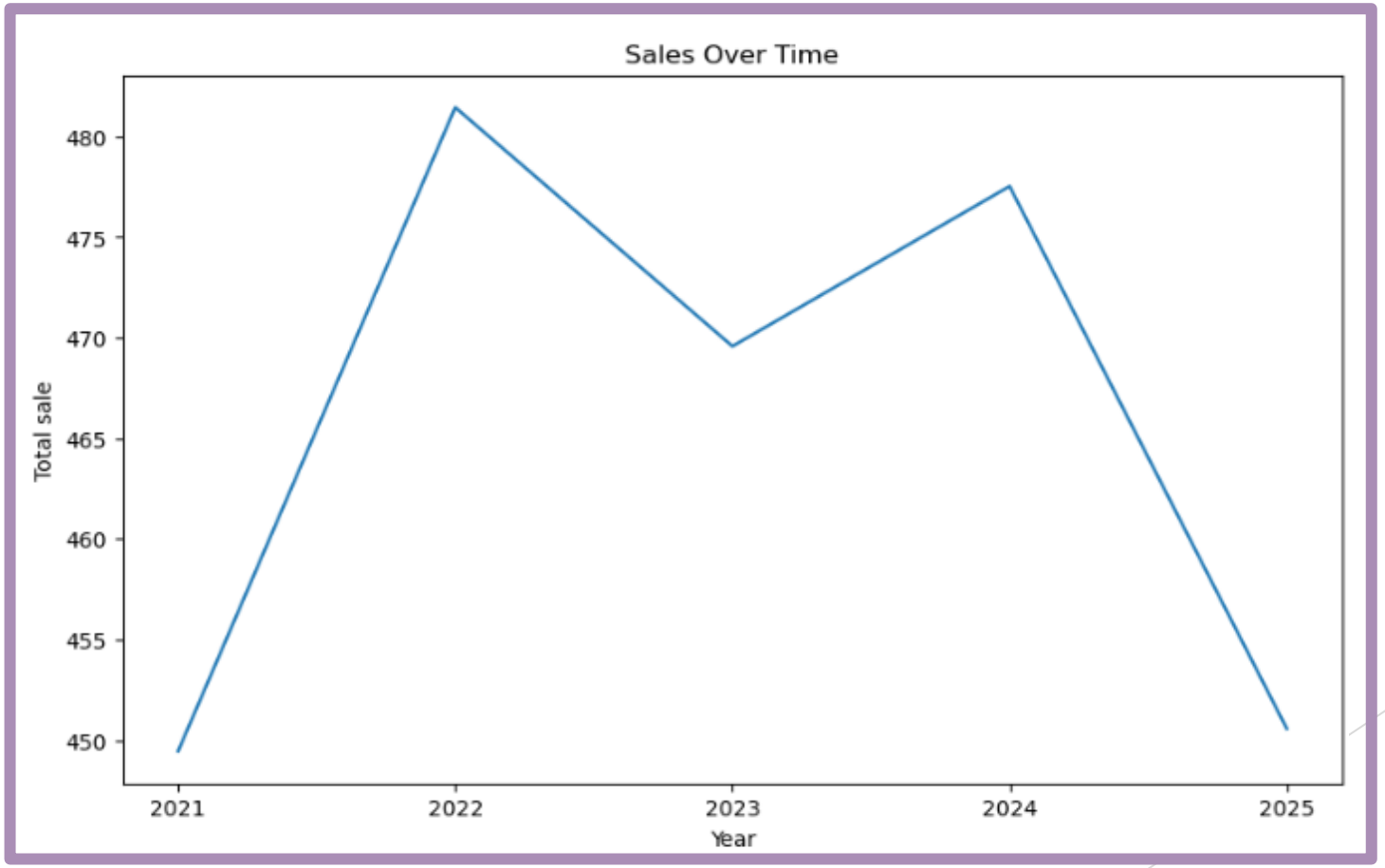
Categorical variables:

- Country



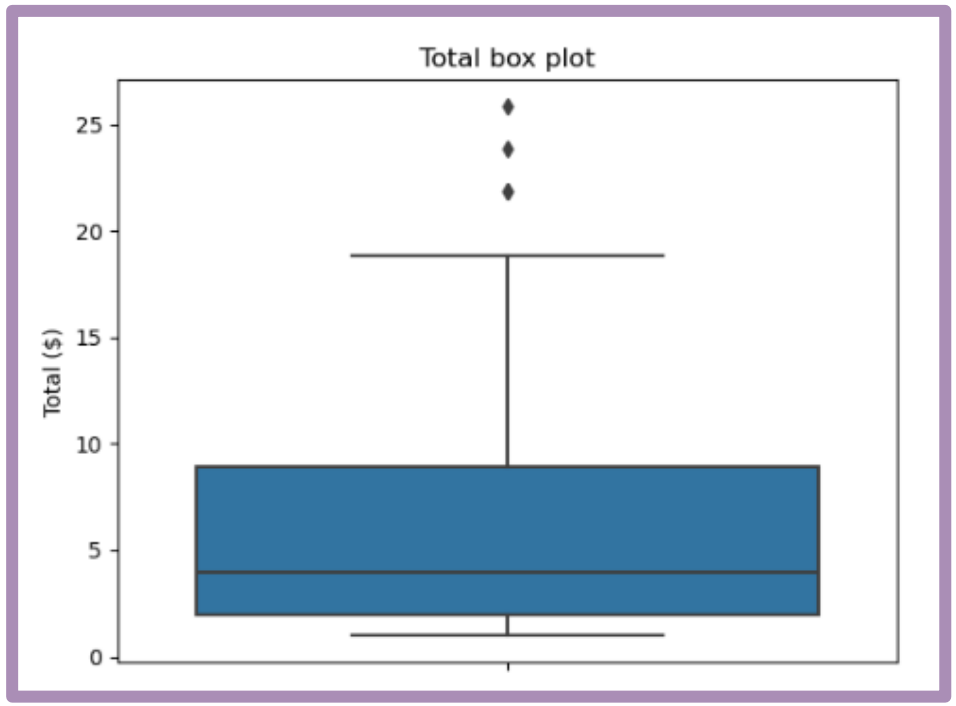
Key variables

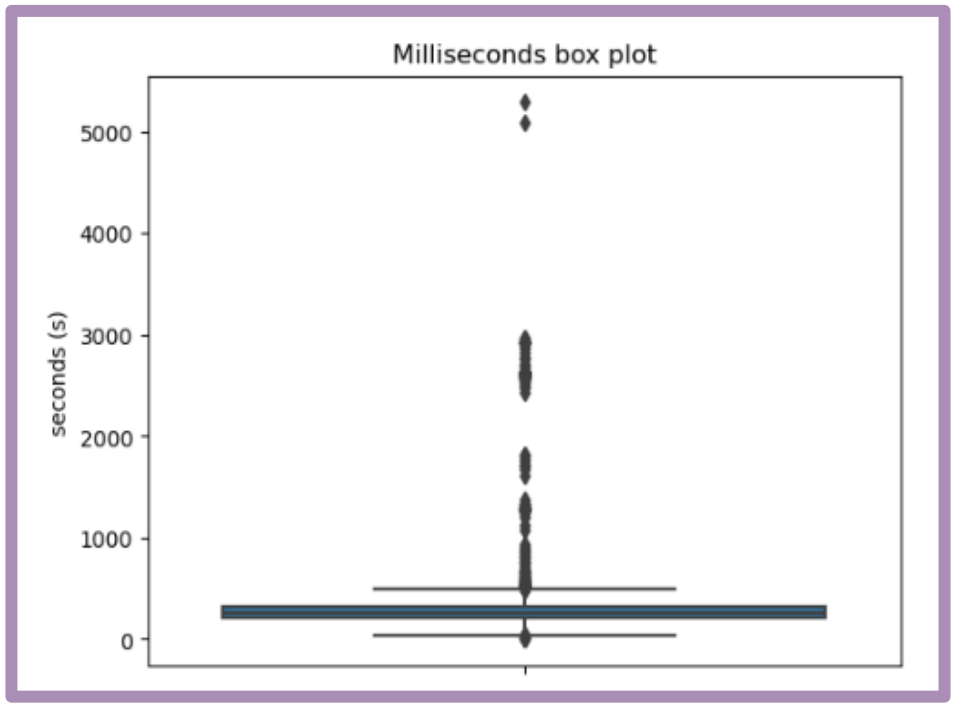
- InvoiceDate



Others Plot

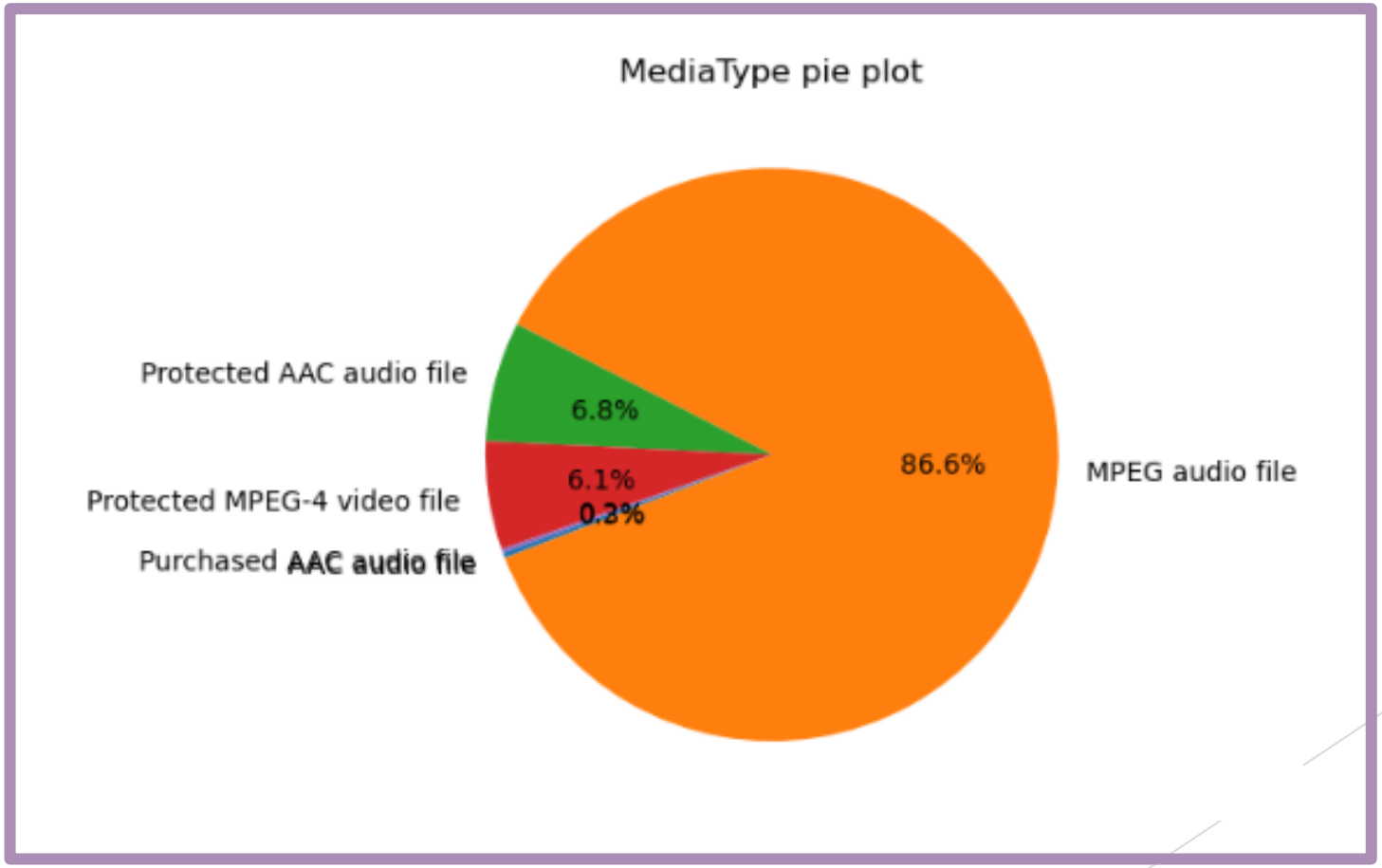
- Box Plot





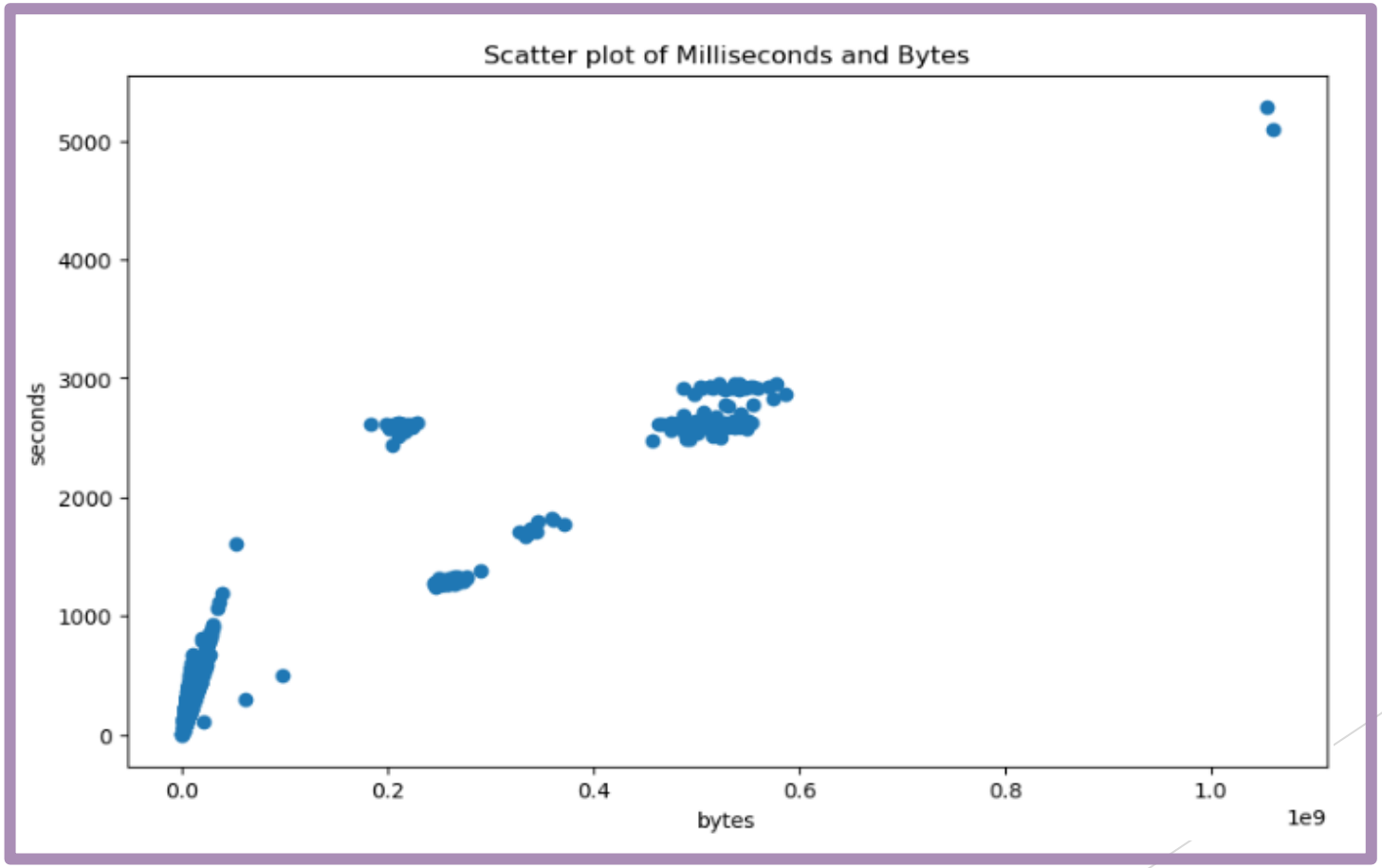
Others Plot

- Pie Plot



Others Plot

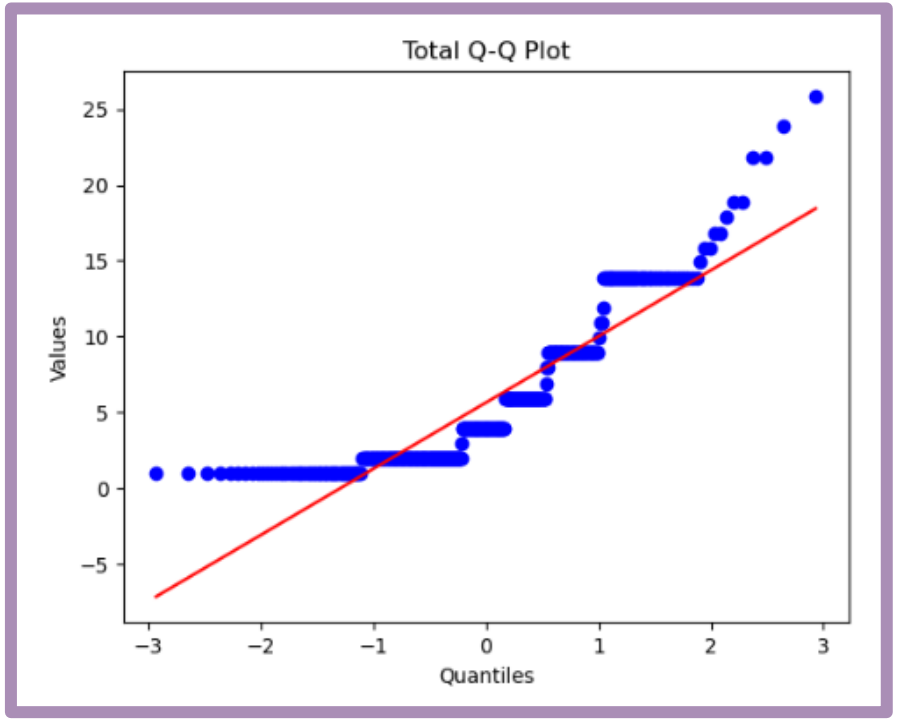
- Scatter Plot

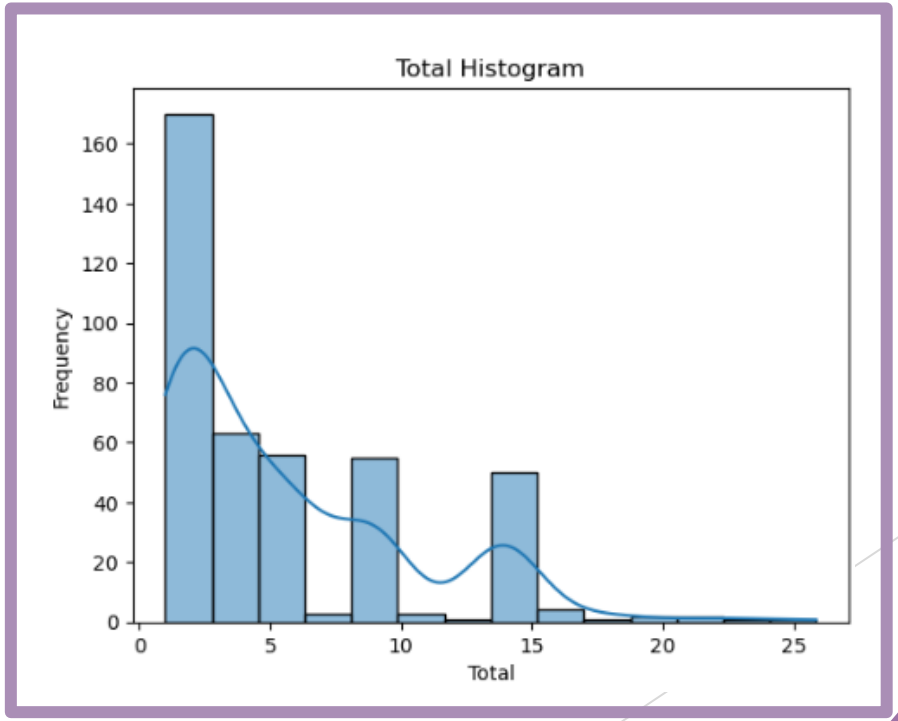


Numerical variables normality: Shapiro-Wilk test, Q-Q Plot, Histogram

- Total normality

```
T-statistic: 0.8367117643356323, p-value: 3.4004587188848802e-20
Reject the null hypothesis. Total is not normally distributed.
```

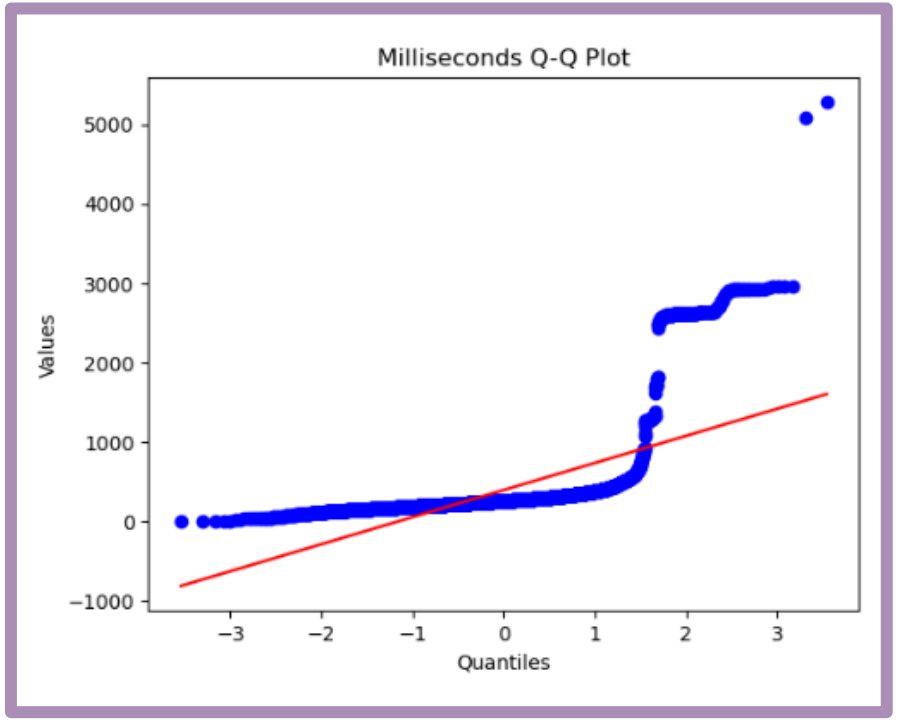


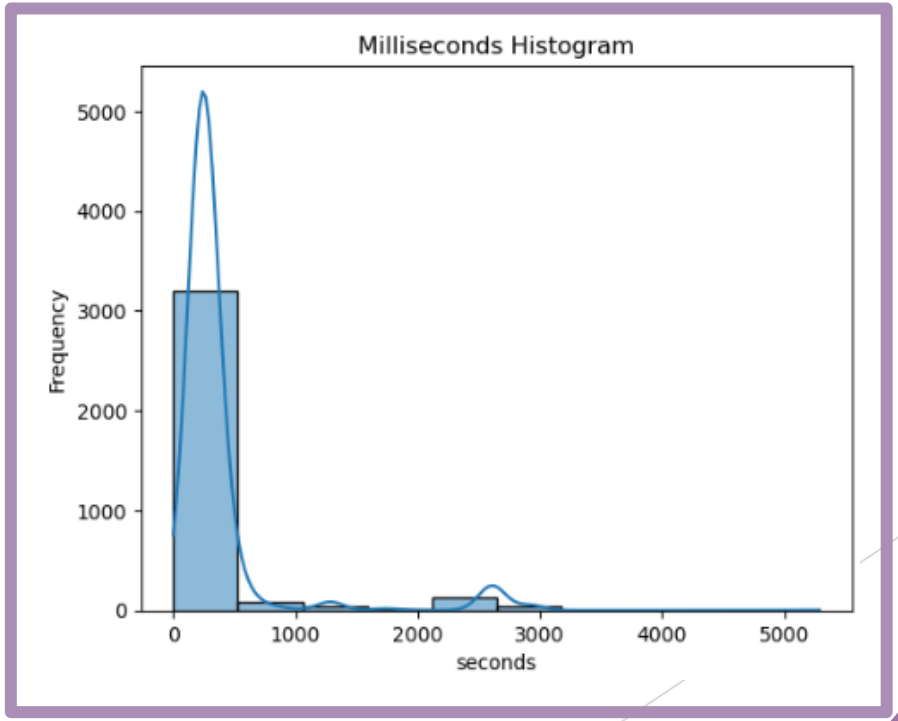


Numerical variables normality: Shapiro-Wilk test, Q-Q Plot, Histogram

- Milliseconds normality

```
T-statistic: 0.4069346785545349, p-value: 0.0
Reject the null hypothesis. Milliseconds is not normally distributed.
```

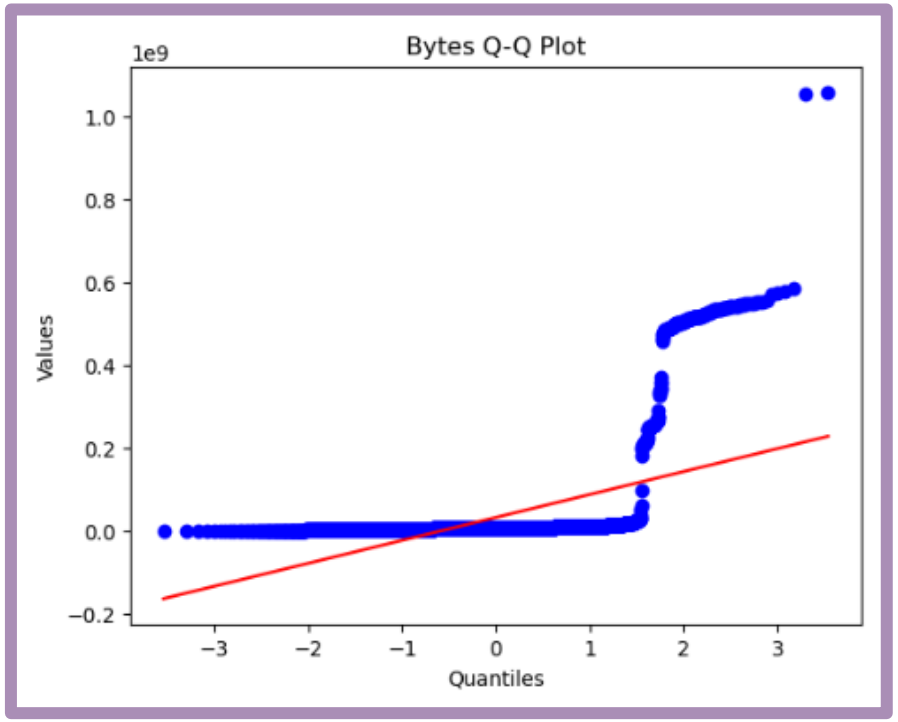


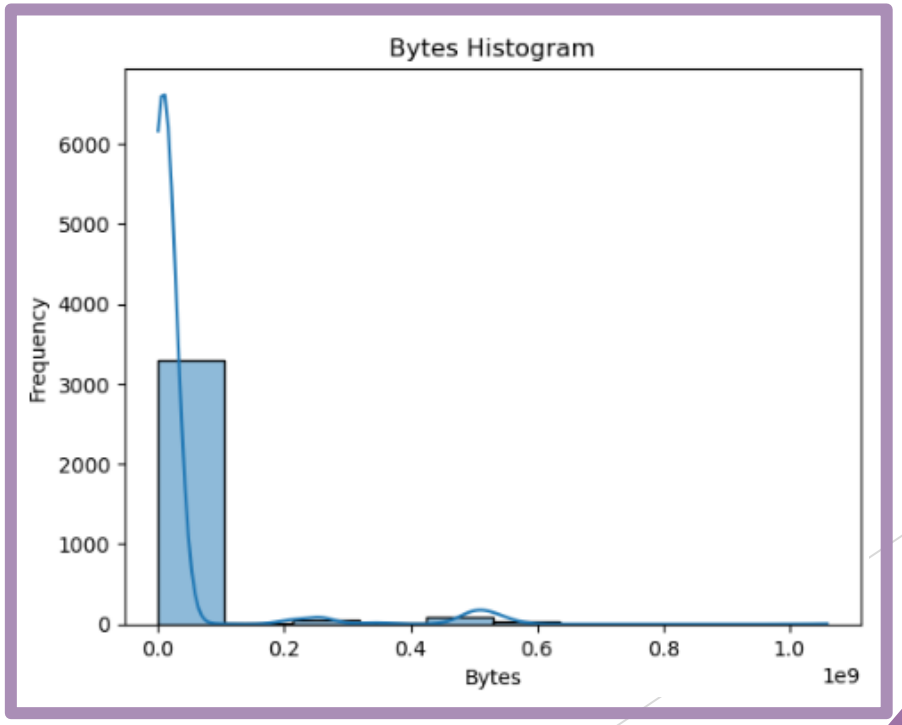


Numerical variables normality: Shapiro-Wilk test, Q-Q Plot, Histogram

- Bytes normality

```
T-statistic: 0.27564042806625366, p-value: 0.0
Reject the null hypothesis. Bytes is not normally distributed.
```

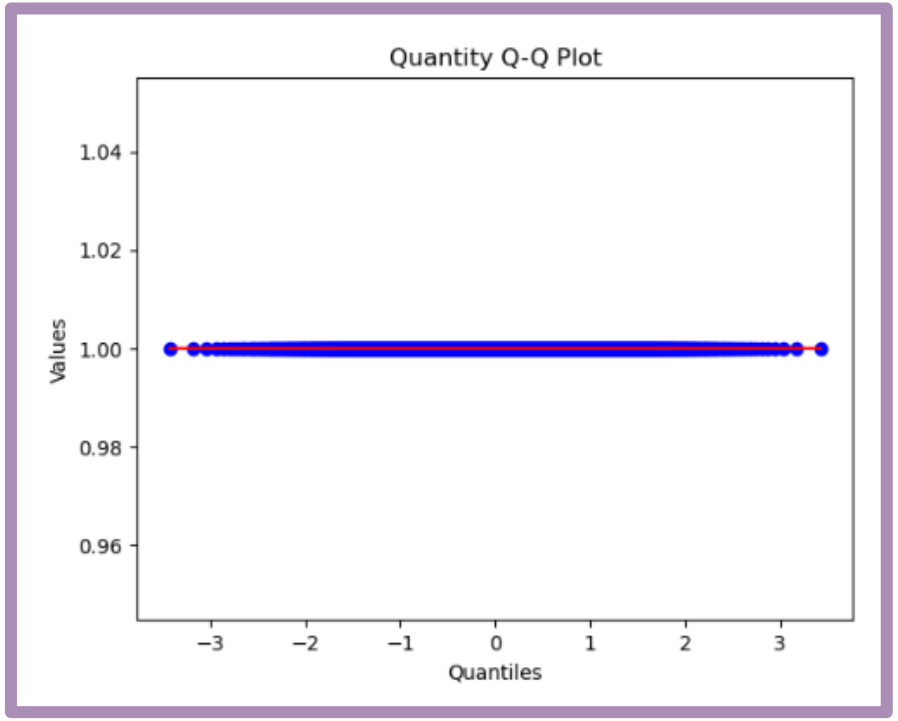


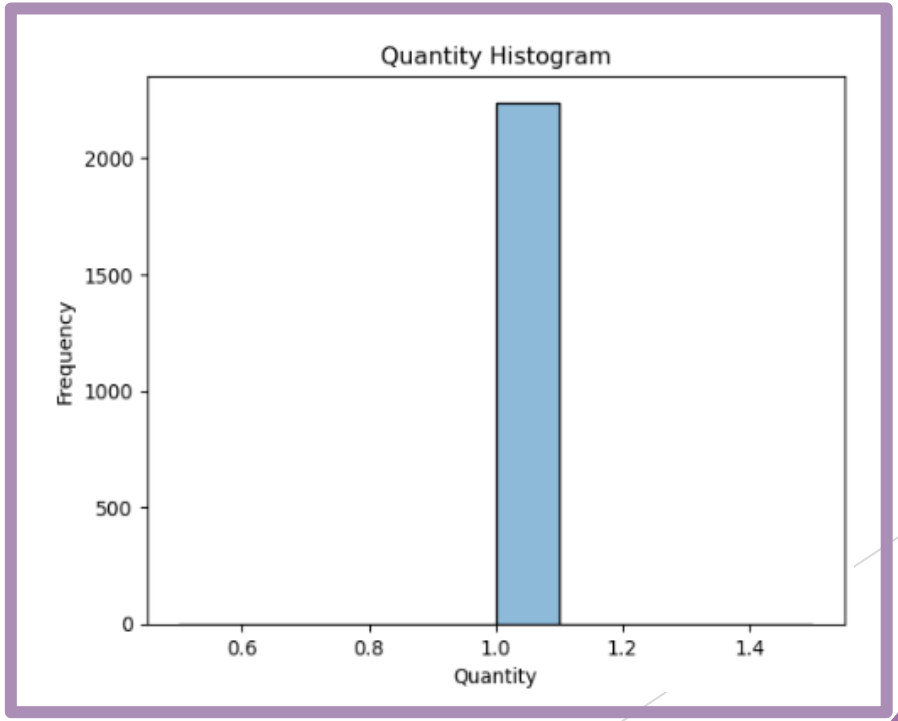


Numerical variables normality: Shapiro-Wilk test, Q-Q Plot, Histogram

- Quantity normality

```
T-statistic: 1.0, p-value: 1.0
Fail to reject the null hypothesis. Quantity is normally distributed.
```

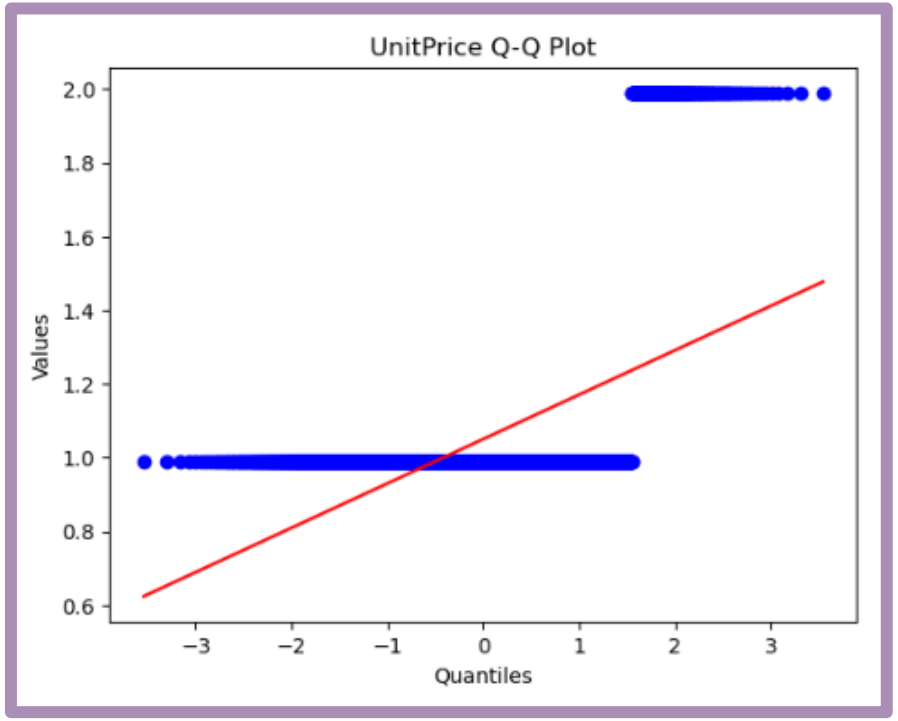


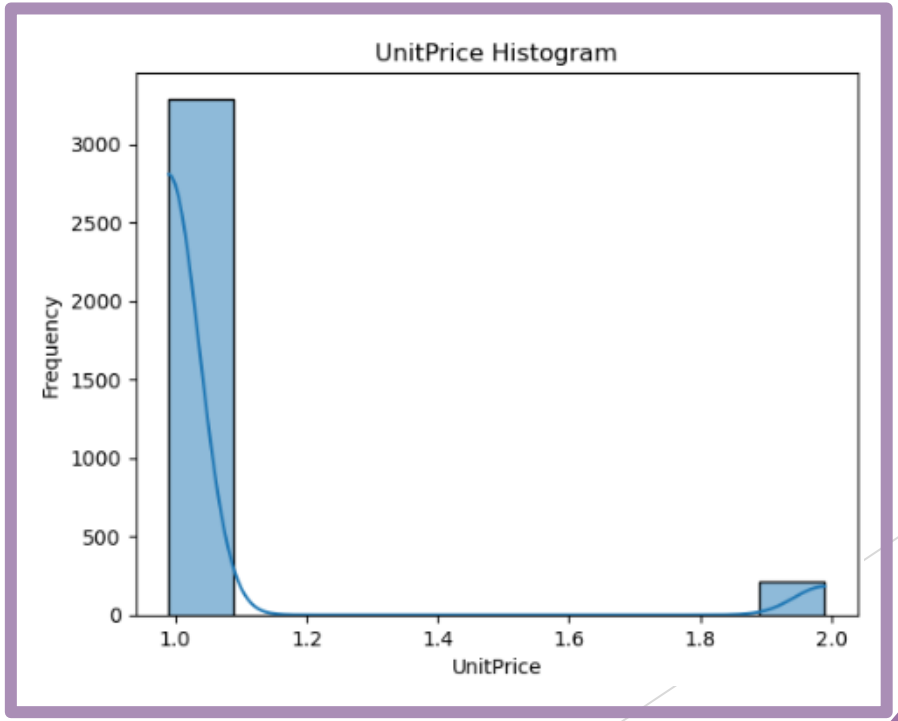


Numerical variables normality: Shapiro-Wilk test, Q-Q Plot, Histogram

- UnitPrice normality

```
T-statistic: 0.2536665201187134, p-value: 0.0
Reject the null hypothesis. UnitPrice is not normally distributed.
```

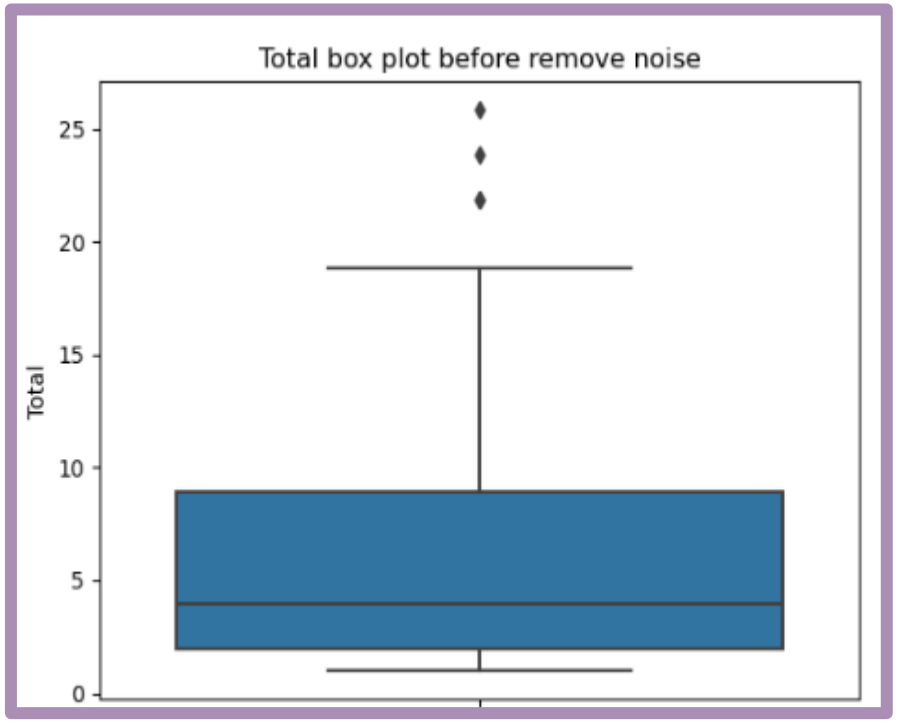


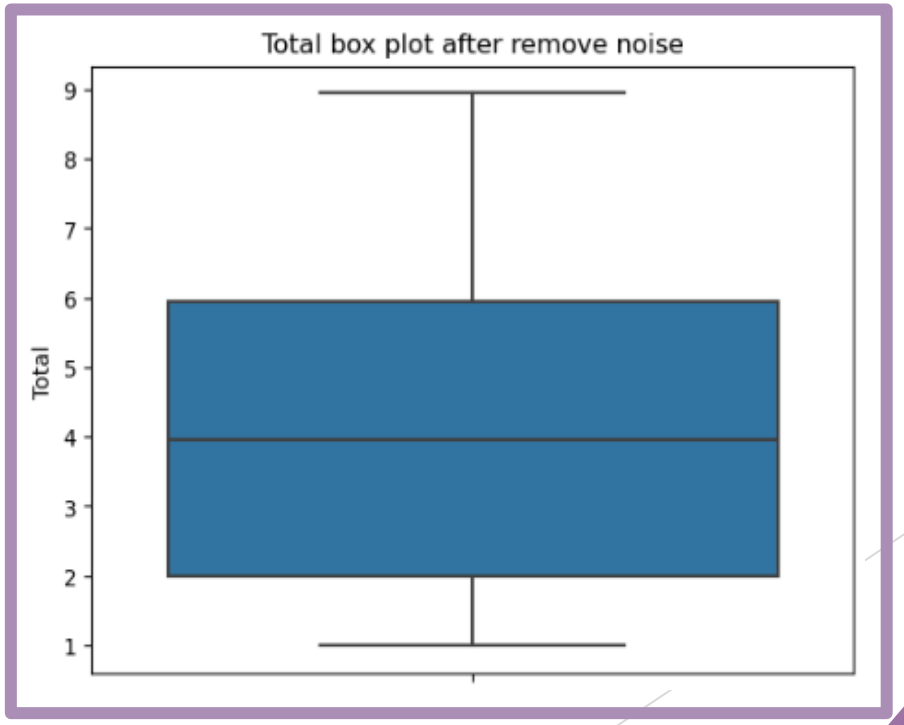


Outliers: Z-Score

- Total Outliers

```
Original data size: (412, 10)
Cleaned data size: (347, 10)
```

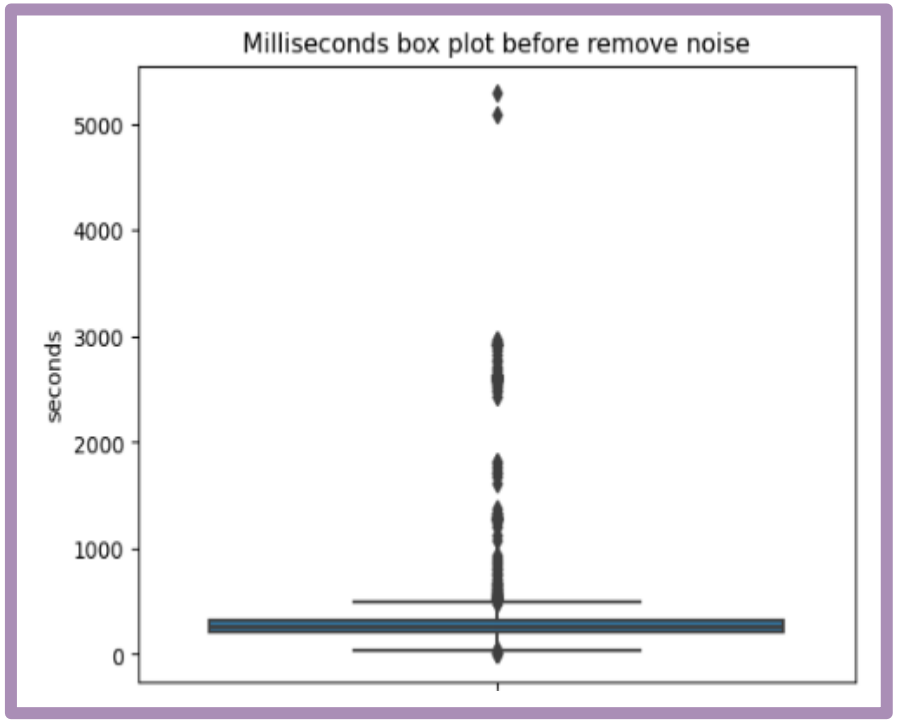


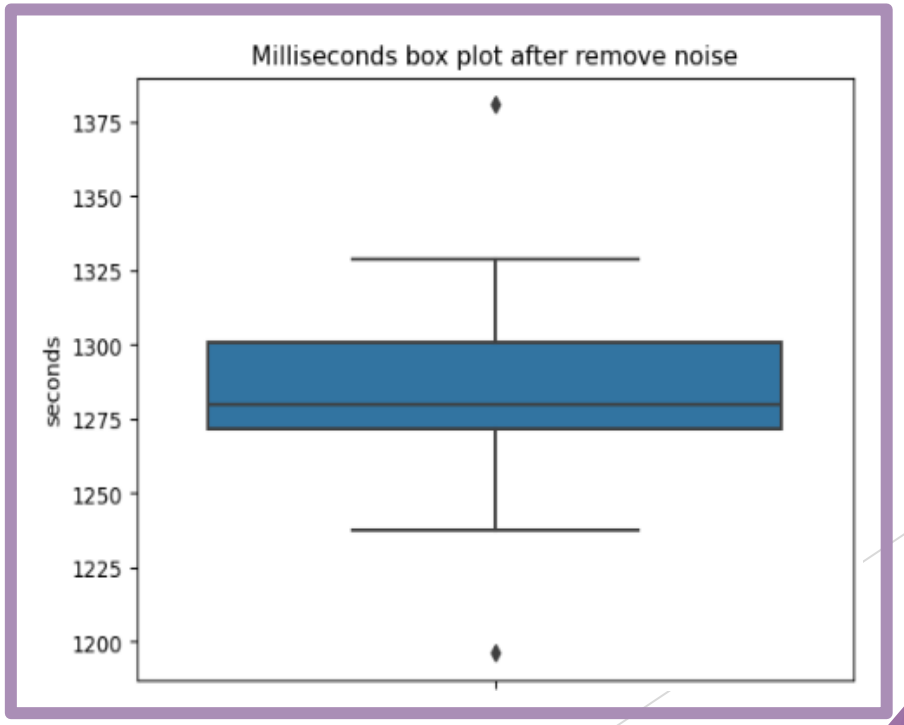


Outliers: Z-Score

- Milliseconds Outliers

```
Original data size: (3503, 10)
Cleaned data size: (43, 10)
```





Q7-1: Find three popular genres with pandas. Check the difference between the average price of these two distributions.

Popular 3 Genres:

- Rock
- Metal
- Latin

Test: Anova

```
T-statistic: nan, p-value: nan
Fail to reject the null hypothesis. There is no significant difference in mean price between genre 1, genre 3 and genre 7.
```

Q7-2: Check independence between the length of the Tracks and its price.

Test: T-test

```
T-statistic: 43.42652221765946, p-value: 0.0
Reject the null hypothesis. There is no significant realation between Track lenght and Track price.
```

Q7-3: Check independence between the Genre and MediaType.

Test: Chi-square

```
Chi-square statistic: 5650.1804373219775, p-value: 0.0
Reject the null hypothesis. There is no significant realation between MediaTypeId and GenreId categories.
```

Q7-4: Is the average revenue from Canada and America different?

Test: T-test

```
T-statistic: 5650.1804373219775, p-value: [0.68671557]
Fail to reject the null hypothesis. There is no significant difference in mean Total between USA and Canada.
```

Q7-5: Is there a significant relationship between the genre and the country of each user?

Test: Chi-square

```
Chi-square statistic: 1247.948598374529, p-value: 5.32766628545981e-60
Reject the null hypothesis. There is a significant association between the genre of music and the country of the customer.
```

Q7-6: Is there a significant relationship between album genre and album artist?

Test: Chi-square

```
Chi-square statistic: 210.25000000000003, p-value: 0.4433092625431917
Fail to reject the null hypothesis. There is no association between the genre of an album and the artist of an album.
```

Q8-1: Is the average length of tracks in different genres the same? Compute a 95% confidence interval for the mean length of track in each genre.

Test: Anova

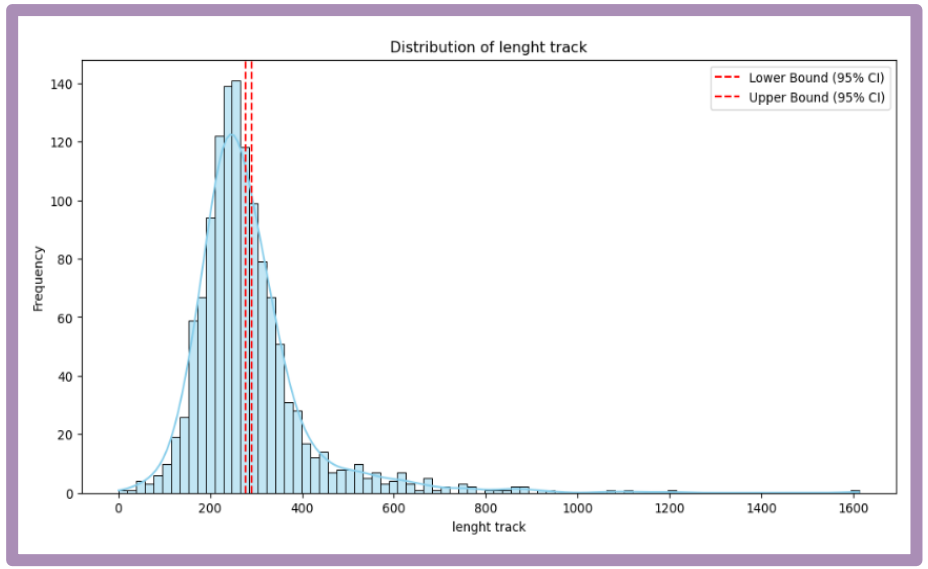
```
statistic: 5650.1804373219775, p-value: 0.0
Reject the null hypothesis. The average length of Tracks in genres is different from each other.
```

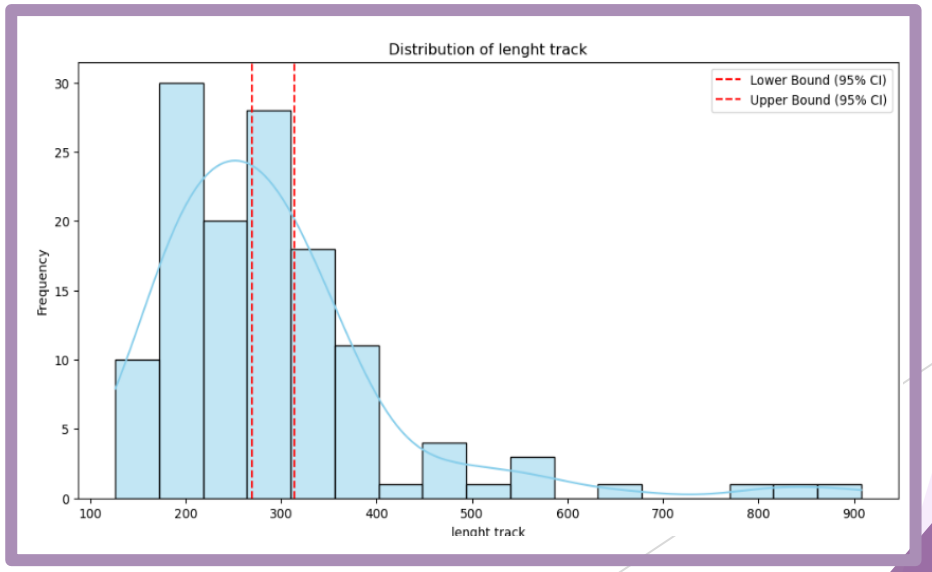
Genre 1:

```
95% Confidence interval for mean lenght track: (277.01217842040177, 290.80790793272087)
```

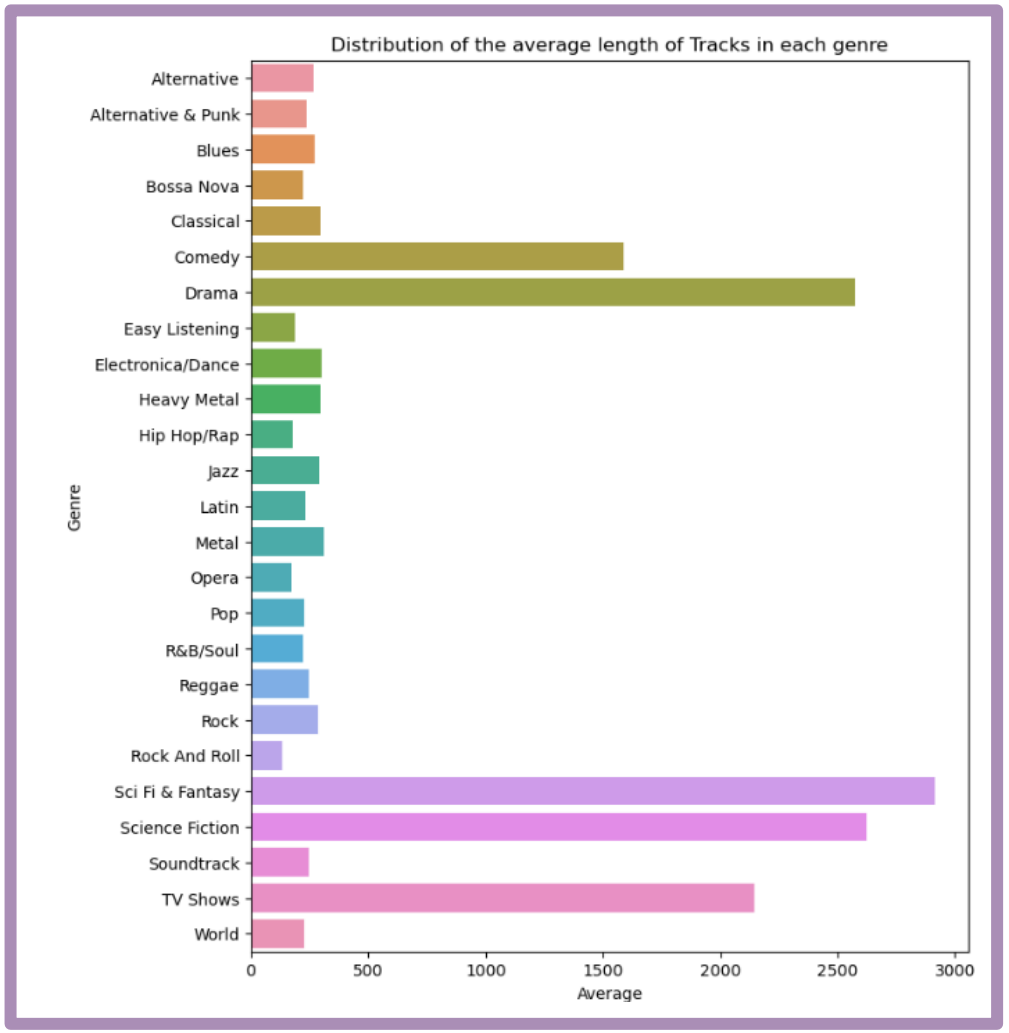
Genre 2:

```
95% Confidence interval for mean lenght track: (269.64445899973623, 313.86629484641765)
```





Q8-1: Is the average length of tracks in different genres the same? Compute a 95% confidence interval for the mean length of track in each genre.

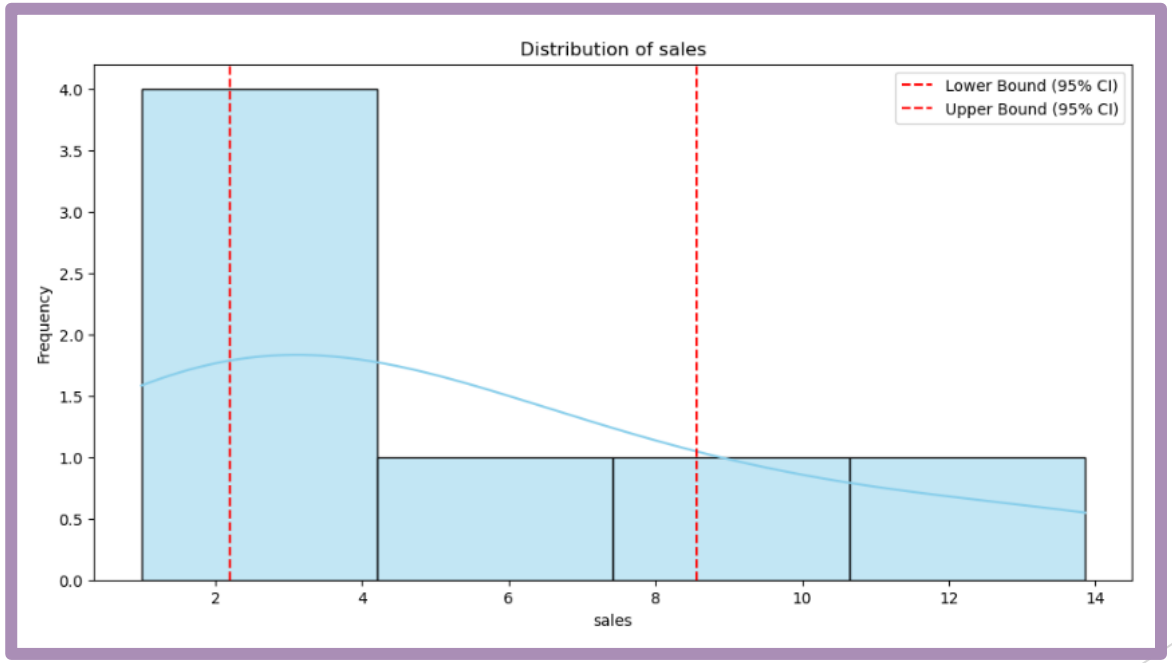


Q8-2: What is the average sales in different countries? Calculate the 95% confidence interval for the mean sales in each country.

0	Argentina	5.37
1	Australia	5.37
2	Austria	6.09
3	Belgium	5.37
4	Brazil	5.43
5	Canada	5.43
6	Chile	6.66
7	Czech Republic	6.45
8	Denmark	5.37
9	Finland	5.95
10	France	5.57
11	Germany	5.59
12	Hungary	6.52
13	India	5.79
14	Ireland	6.52
15	Italy	5.37
16	Netherlands	5.80
17	Norway	5.66
18	Poland	5.37
19	Portugal	5.52
20	Spain	5.37
21	Sweden	5.52
22	USA	5.75
23	United Kingdom	5.37

Argentina:

```
95% Confidence interval for mean sales: (2.193007245584535, 8.555564182986894)
```



Q8-3: What is the average number of Tracks purchased by each user? Calculate a 95% confidence interval for the mean number of Tracks purchased by each user.

0	Almeida	1.0
1	Barnett	1.0
2	Bernard	1.0
3	Brooks	1.0
4	Brown	1.0
5	Chase	1.0
6	Cunningham	1.0
7	Dubois	1.0
8	Fernandes	1.0
9	Francis	1.0
10	Girard	1.0
11	Gonçalves	1.0
12	Gordon	1.0
13	Goyer	1.0
14	Gray	1.0
15	Gruber	1.0
16	Gutiérrez	1.0
17	Hansen	1.0
18	Harris	1.0
19	Holý	1.0
20	Hughes	1.0

```
95% Confidence interval for mean Quantity: (1.0, 1.0)
```