

Data Collection and Preprocessing Phase

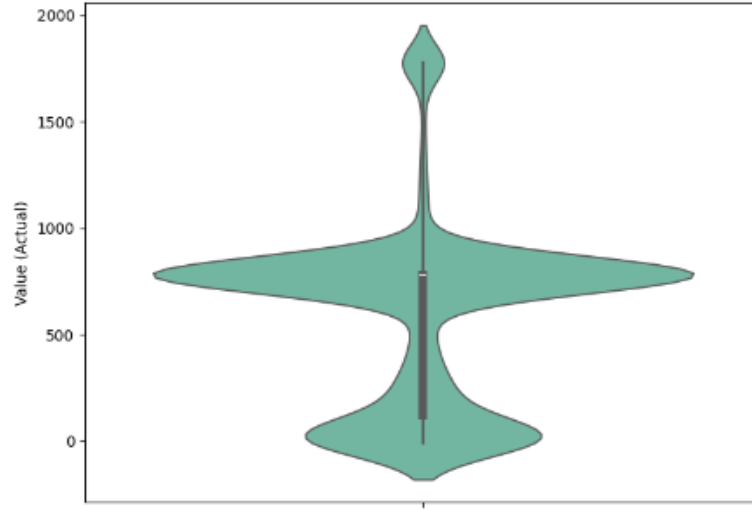
Date	8 July 2024
Team ID	739928
Project Title	Rhythmic Revenue: Unveiling The Future Of Music Sales With Machine Learning
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

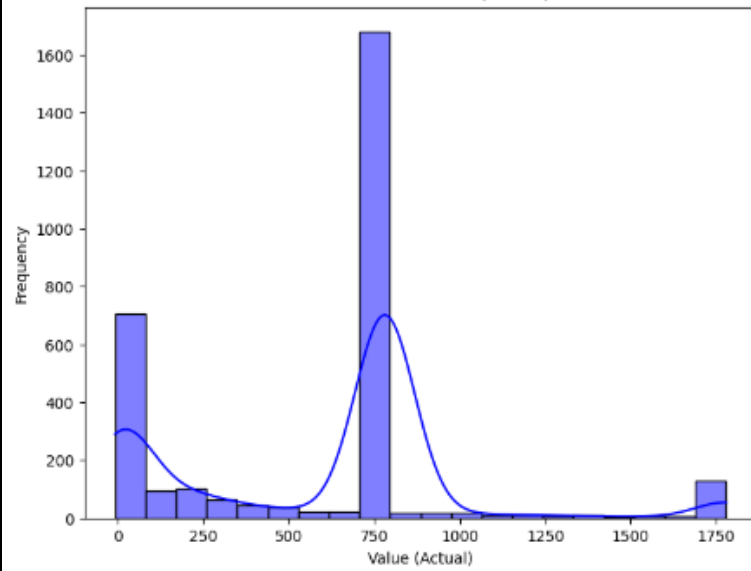
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																														
Data Overview	<u>Dimension:</u> 3008rows x 6columns																																																														
	<u>Descriptive statistics:</u>																																																														
	<pre>df.describe()</pre>																																																														
	<table><thead><tr><th></th><th>index</th><th>Format</th><th>Metric</th><th>Year</th><th>Number of Records</th><th>Value (Actual)</th></tr></thead><tbody><tr><td>count</td><td>3008.00000</td><td>3008.000000</td><td>3008.000000</td><td>3008.000000</td><td>3008.0</td><td>3008.000000</td></tr><tr><td>mean</td><td>1503.50000</td><td>10.890625</td><td>1.078125</td><td>1996.000000</td><td>1.0</td><td>594.928843</td></tr><tr><td>std</td><td>888.47913</td><td>6.998959</td><td>0.796701</td><td>13.586915</td><td>0.0</td><td>431.109519</td></tr><tr><td>min</td><td>0.00000</td><td>0.000000</td><td>0.000000</td><td>1973.000000</td><td>1.0</td><td>-7.650944</td></tr><tr><td>25%</td><td>751.75000</td><td>5.000000</td><td>0.000000</td><td>1984.000000</td><td>1.0</td><td>116.560241</td></tr><tr><td>50%</td><td>1503.50000</td><td>10.000000</td><td>1.000000</td><td>1996.000000</td><td>1.0</td><td>781.291237</td></tr><tr><td>75%</td><td>2255.25000</td><td>17.000000</td><td>2.000000</td><td>2008.000000</td><td>1.0</td><td>781.291237</td></tr><tr><td>max</td><td>3007.00000</td><td>23.000000</td><td>2.000000</td><td>2019.000000</td><td>1.0</td><td>1778.387731</td></tr></tbody></table>		index	Format	Metric	Year	Number of Records	Value (Actual)	count	3008.00000	3008.000000	3008.000000	3008.000000	3008.0	3008.000000	mean	1503.50000	10.890625	1.078125	1996.000000	1.0	594.928843	std	888.47913	6.998959	0.796701	13.586915	0.0	431.109519	min	0.00000	0.000000	0.000000	1973.000000	1.0	-7.650944	25%	751.75000	5.000000	0.000000	1984.000000	1.0	116.560241	50%	1503.50000	10.000000	1.000000	1996.000000	1.0	781.291237	75%	2255.25000	17.000000	2.000000	2008.000000	1.0	781.291237	max	3007.00000	23.000000	2.000000	2019.000000	1.0
	index	Format	Metric	Year	Number of Records	Value (Actual)																																																									
count	3008.00000	3008.000000	3008.000000	3008.000000	3008.0	3008.000000																																																									
mean	1503.50000	10.890625	1.078125	1996.000000	1.0	594.928843																																																									
std	888.47913	6.998959	0.796701	13.586915	0.0	431.109519																																																									
min	0.00000	0.000000	0.000000	1973.000000	1.0	-7.650944																																																									
25%	751.75000	5.000000	0.000000	1984.000000	1.0	116.560241																																																									
50%	1503.50000	10.000000	1.000000	1996.000000	1.0	781.291237																																																									
75%	2255.25000	17.000000	2.000000	2008.000000	1.0	781.291237																																																									
max	3007.00000	23.000000	2.000000	2019.000000	1.0	1778.387731																																																									
Univariate Analysis																																																															

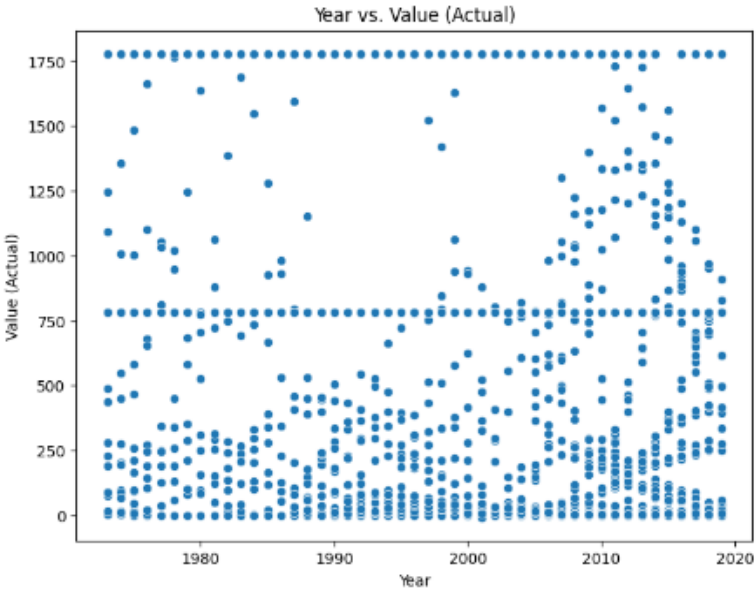
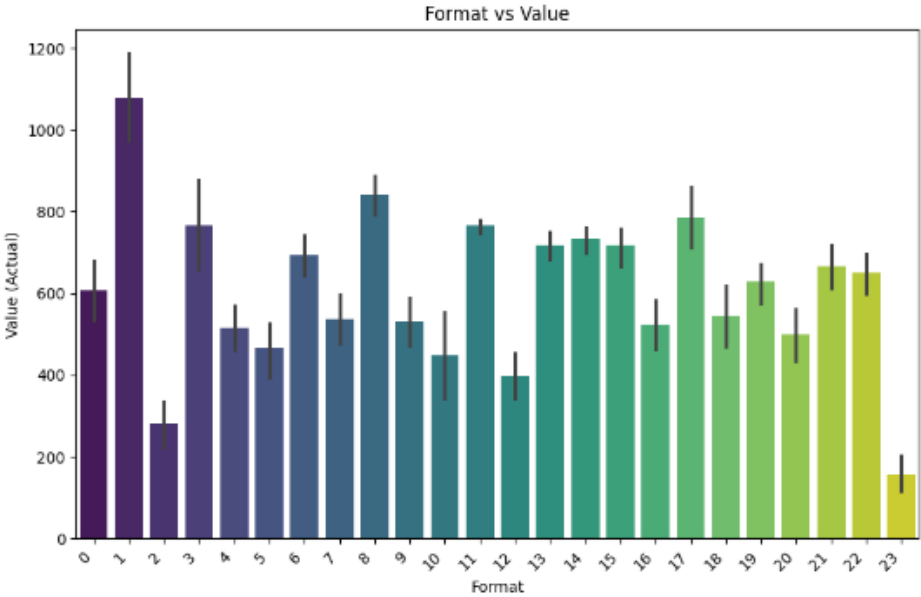
Violin Plot of Value (Actual)

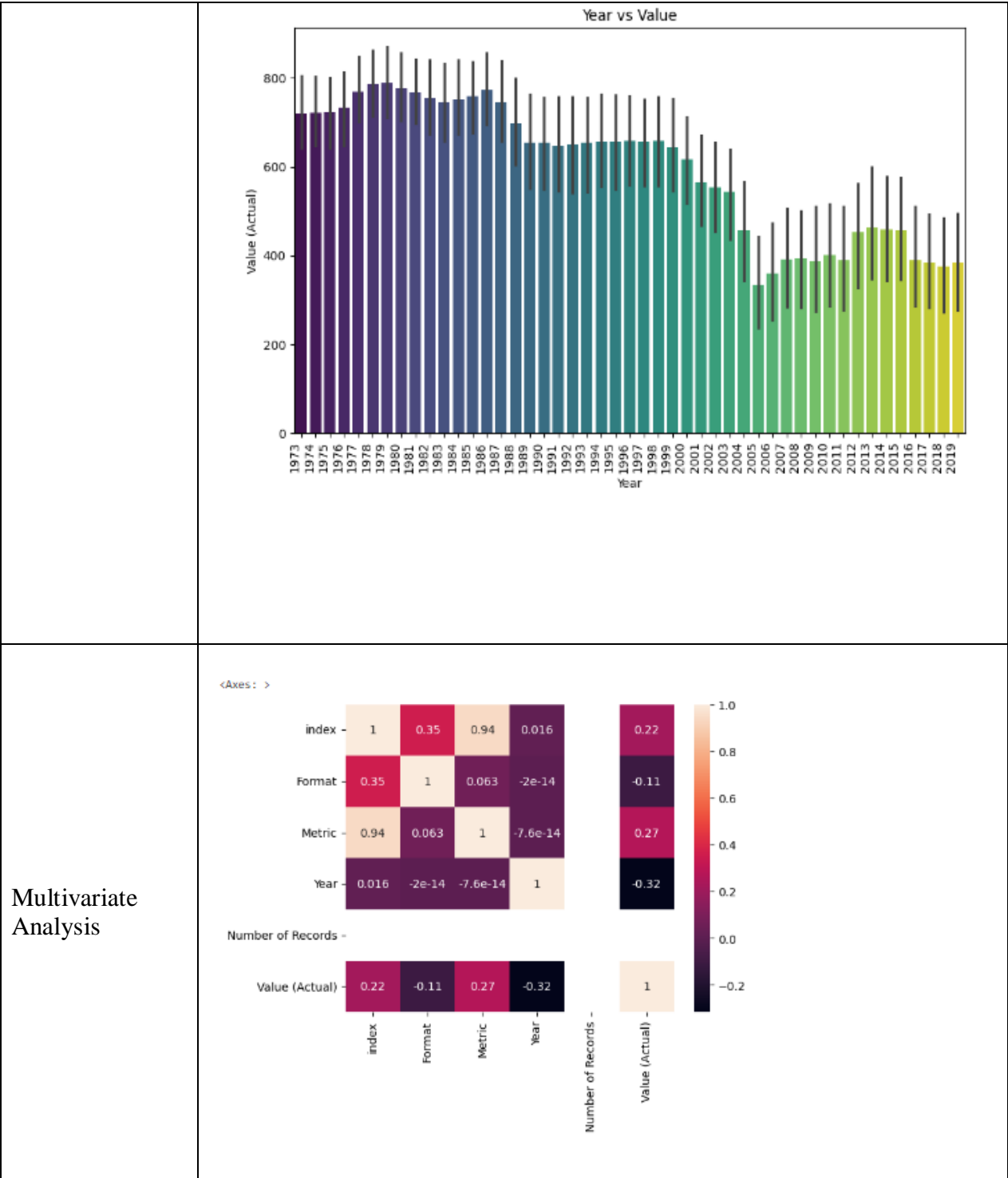


Distribution of Value (Actual)

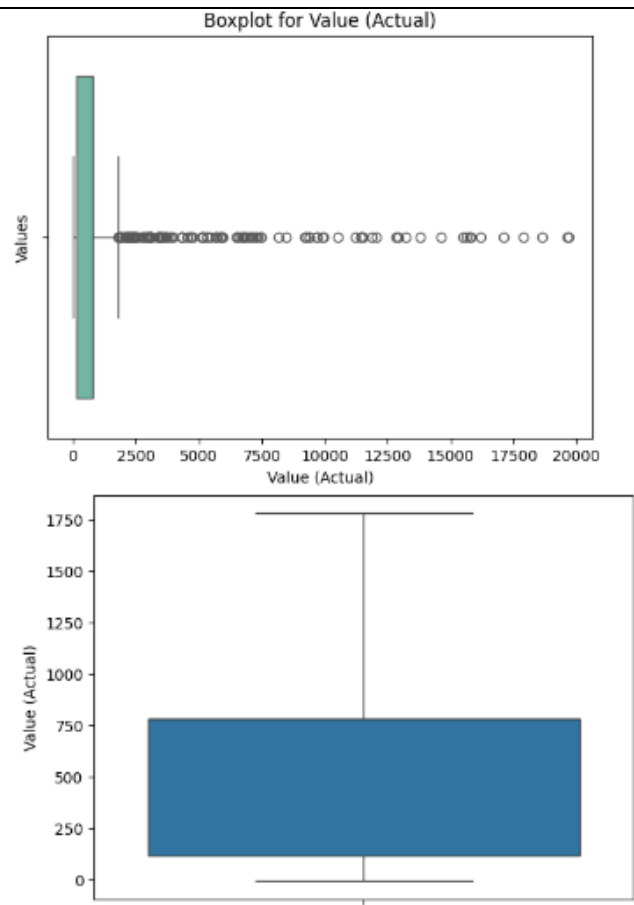


Bivariate Analysis





Outliers and Anomalies



Data Preprocessing Code Screenshots

Loading Data

```
[ ] df = pd.read_csv('/content/MusicData.csv')

[ ] df
```

	index	Format	Metric	Year	Number of Records	Value (Actual)
0	0	CD	Units	1973	1	NaN
1	1	CD	Units	1974	1	NaN
2	2	CD	Units	1975	1	NaN
3	3	CD	Units	1976	1	NaN
4	4	CD	Units	1977	1	NaN
...
3003	3003	Vinyl Single	Value (Adjusted)	2015	1	6.205390
3004	3004	Vinyl Single	Value (Adjusted)	2016	1	5.198931
3005	3005	Vinyl Single	Value (Adjusted)	2017	1	6.339678
3006	3006	Vinyl Single	Value (Adjusted)	2018	1	5.386197
3007	3007	Vinyl Single	Value (Adjusted)	2019	1	6.795948

3008 rows x 6 columns

Handling Missing Data	<pre>df.isnull().sum()</pre> <pre>index 0 Format 0 Metric 0 Year 0 Number of Records 0 Value (Actual) 1657 dtype: int64</pre> <pre>[] df['Value (Actual)'].mean() 781.2912371175493 df['Value (Actual)'].fillna(781.2912371175493,inplace=True)</pre> <pre>[] df.isnull().sum()</pre> <pre>index 0 Format 0 Metric 0 Year 0 Number of Records 0 Value (Actual) 0 dtype: int64</pre>
Data Transformation	<pre>[] from sklearn.preprocessing import LabelEncoder label_encoder = LabelEncoder() df['Format']=label_encoder.fit_transform(df['Format'])</pre> <pre>[] from sklearn.preprocessing import LabelEncoder label_encoder = LabelEncoder() df['Metric']=label_encoder.fit_transform(df['Metric'])</pre>
Feature Engineering	Attached the code in final submission
Save Processed Data	-