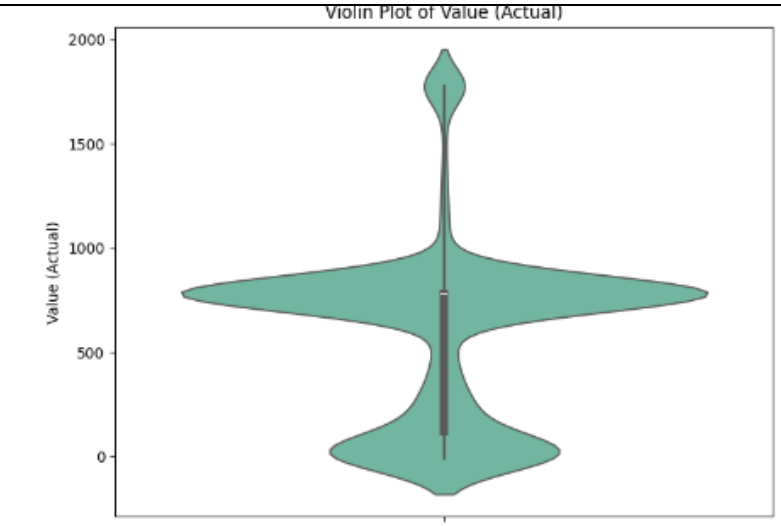
Data Collection and Preprocessing Phase

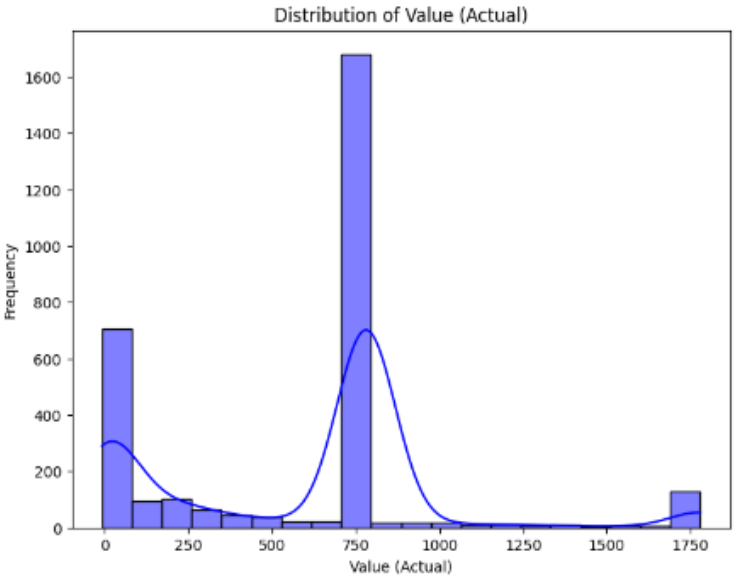
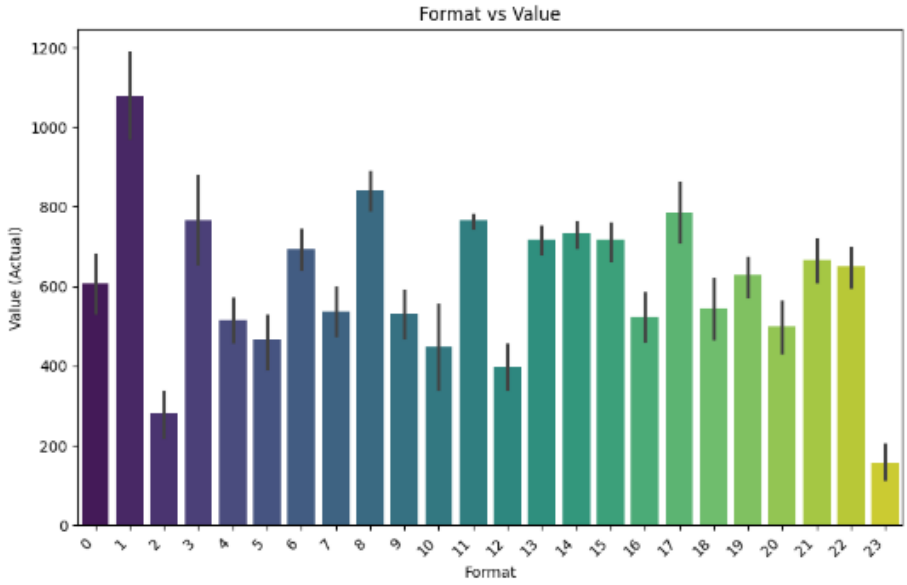
Date	8 July 2024
Team ID	740074
Project Title	Rhythmic Revenue: Unveiling The Future Of Music Sales With Machine Learning
Maximum Marks	6 Marks

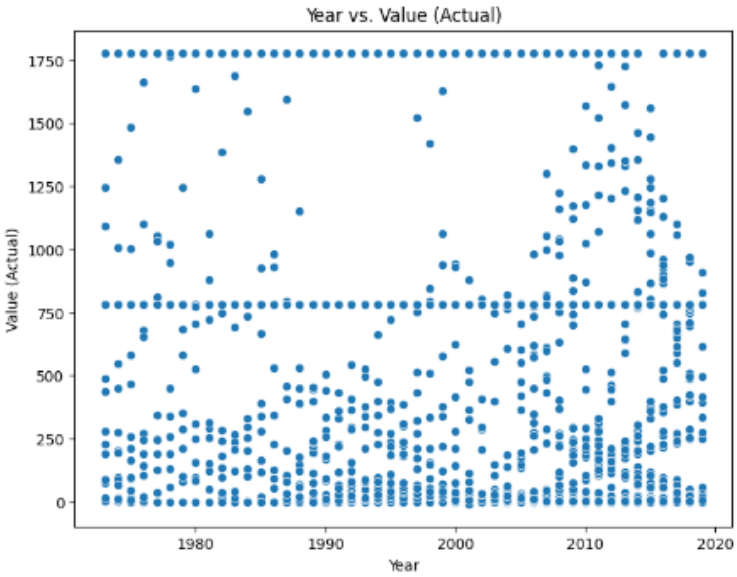
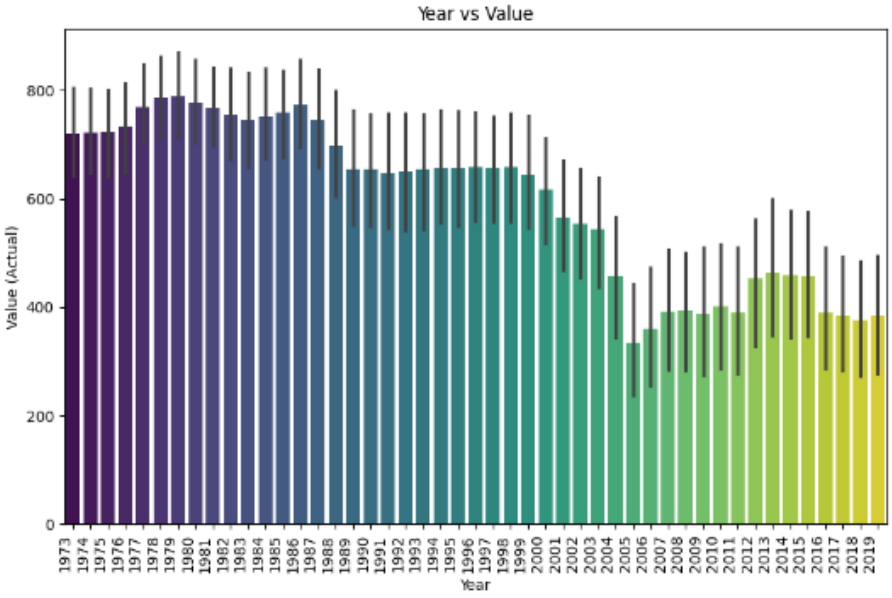
Data Exploration and Preprocessing Template

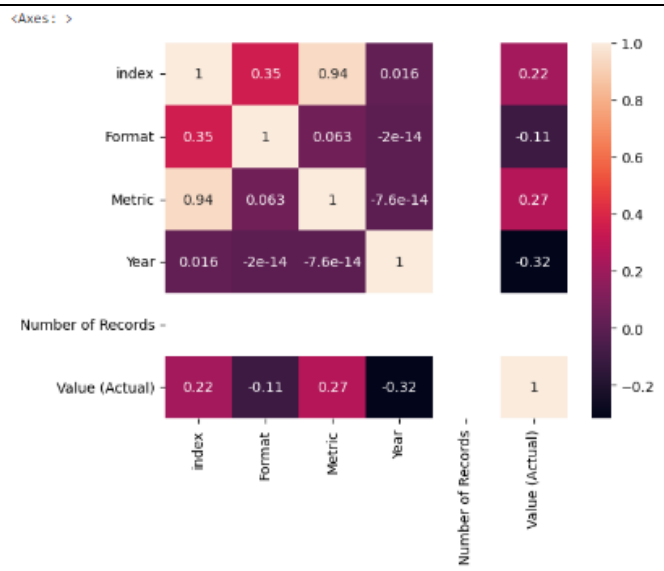
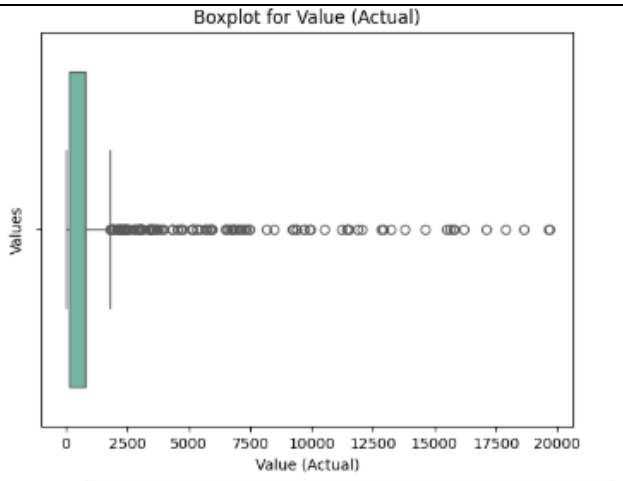
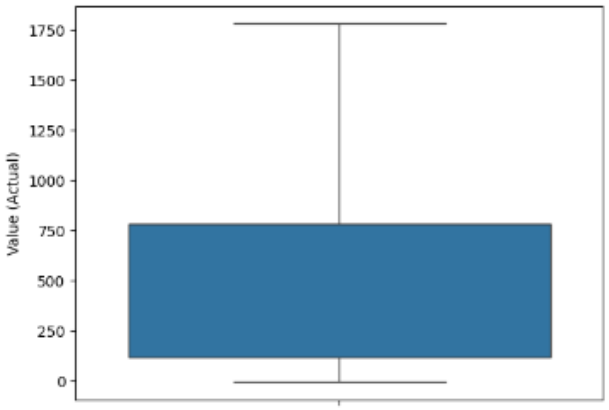
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description
Data Overview	Dimension: 3008rows x 6columns Descriptive statistics: `df.describe()` index: count 3008.00000, mean 1503.50000, std 868.47913, min 0.00000, 25% 751.75000, 50% 1503.50000, 75% 2255.25000, max 3007.00000 Format: count 3008.000000, mean 10.890625, std 6.996959, min 0.000000, 25% 5.000000, 50% 10.000000, 75% 17.000000, max 23.000000 Metric: count 3008.000000, mean 1.078125, std 0.796701, min 0.000000, 25% 0.000000, 50% 1.000000, 75% 2.000000, max 2.000000 Year: count 3008.000000, mean 1996.000000, std 13.566915, min 1973.000000, 25% 1984.000000, 50% 1996.000000, 75% 2008.000000, max 2019.000000 Number of Records: count 3008.0, mean 1.0, std 0.0, min 1.0, 25% 1.0, 50% 1.0, 75% 1.0, max 1.0 Value (Actual): count 3008.000000, mean 594.928843, std 431.109519, min -7.650944, 25% 116.560241, 50% 781.291237, 75% 781.291237, max 1778.387731
Univariate Analysis	

<table>
<tr>
<td></td>
<td> 


</td>
</tr>
<tr>
<td>Bivariate Analysis</td>
<td> 
</td>
</tr>
</table>

	 
Multivariate 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 6.795946
	3008 rows × 6 columns

Handling Missing Data	`df.isnull().sum()`
	
	index 0
	Format 0
	Metric 0
	Year 0
	Number of Records 0
	Value (Actual) 1657
	dtype: int64
	
	`df['Value (Actual)'].mean()`
	`781.2912371175493`
	`df['Value (Actual)'].fillna(781.2912371175493,inplace=True)`
	
	`df.isnull().sum()`
	
	index 0
	Format 0
	Metric 0
	Year 0
	Number of Records 0
	Value (Actual) 0
	dtype: int64

Data Transformation	`from sklearn.preprocessing import LabelEncoder`
	`label_encoder = LabelEncoder()`
	`df['Format']=label_encoder.fit_transform(df['Format'])`
	
	`from sklearn.preprocessing import LabelEncoder`
	`label_encoder = LabelEncoder()`
	`df['Metric']=label_encoder.fit_transform(df['Metric'])`

Feature Engineering	Attached the code in final submission
Save Processed Data	-