Data Collection and Preprocessing Phase

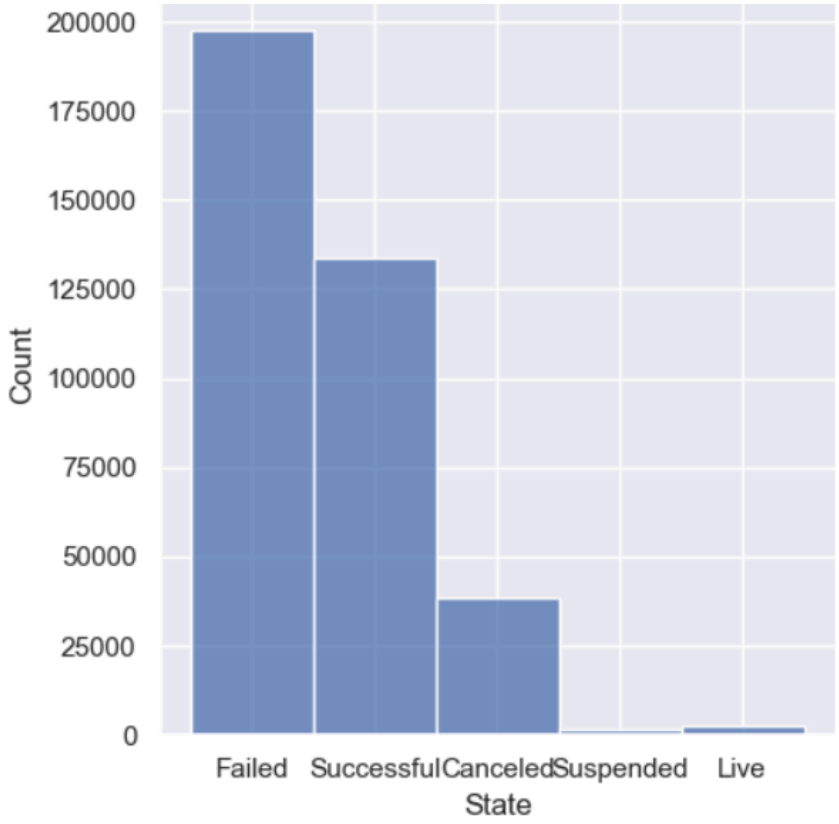
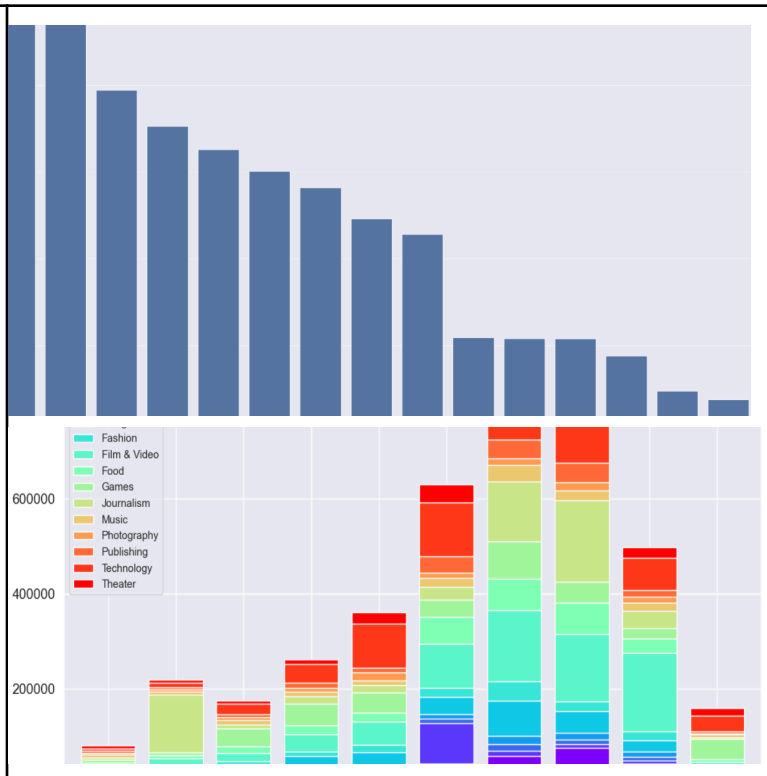
Date	15 July 2024
Team ID	739696
Project Title SmartLender -	Automotive Kickstart
Maximum Marks	6 Marks

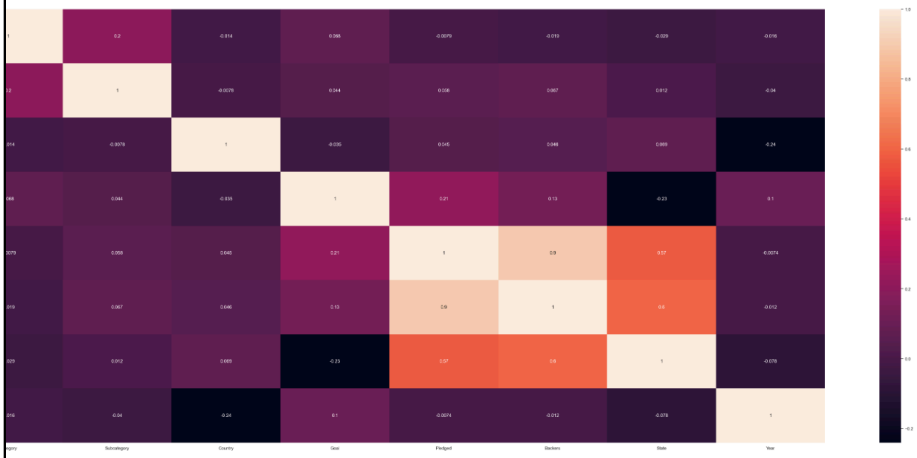
Data Exploration and Preprocessing Report

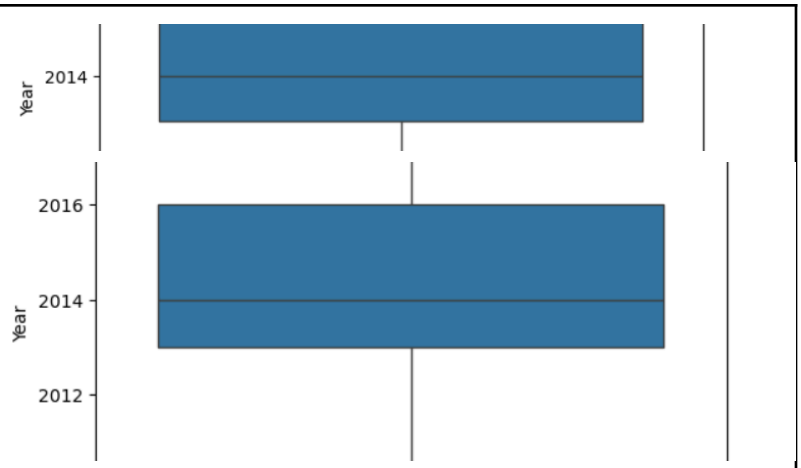
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

Section	Description
Data Overview	Dimension: 374853 rows × 11 columns Descriptive statistics:

ame	Category	Subcategory	Country	Launched	Deadline	Goal	Pledged	
853	374853	374853	374853	374853	374853	3.748530e+05	3.748530e+05	3748
061	15	159	22	374297	3164	NaN	NaN	
usic ent	Film & Video	Product Design	United States	2014-06-06 16:16:32	2014-08-08	NaN	NaN	
13	62694	22310	292618	2	702	NaN	NaN	
NaN	NaN	NaN	NaN	NaN	NaN	4.586378e+04	9.121073e+03	1(
NaN	NaN	NaN	NaN	NaN	NaN	1.158778e+06	9.132054e+04	9
NaN	NaN	NaN	NaN	NaN	NaN	0.000000e+00	0.000000e+00	
NaN	NaN	NaN	NaN	NaN	NaN	2.000000e+03	3.100000e+01	

Univariate Analysis	
Bivariate Analysis	

Multivariate Analysis	

Outliers and Anomalies	

Data Preprocessing Code Screenshots

Loading Data

```
[7]: df = pd.read_csv('automative_kickstart.csv')
[9]: df.head(5)
```

[9]:

	ID	Name	Category	Subcategory	Country	Launched	Deadline	Goal	Pledged	Backers	State
0	1860890148	Grace Jones Does Not Give A F$#% T-Shirt (limi...	Fashion	Fashion	United States	2009-04-21 21:02:48	2009-05-31	1000	625	30	Failed
1	709707365	CRYSTAL ANTLERS UNTITLED MOVIE	Film & Video	Shorts	United States	2009-04-23 00:07:53	2009-07-20	80000	22	3	Failed
2	1703704063	drawing for dollars	Art	Illustration	United States	2009-04-24 21:52:03	2009-05-03	20	35	3	Successful
3	727286	Offline Wikipedia iPhone app	Technology	Software	United States	2009-04-25 17:36:21	2009-07-14	99	145	25	Successful
4	1622952265	Pantshirts	Fashion	Fashion	United States	2009-04-27 14:10:39	2009-05-26	1900	387	10	Failed

Finding & Handling Missing Data	[17]: `df.isnull().sum()` [17]: ID 0 Name 0 Category 0 Subcategory 0 Country 0 Launched 0 Deadline 0 Goal 0 Pledged 0 Backers 0 State 0 dtype: int64

Data Transformation	[162]: ```df1['Category'] = lb.fit_transform(df1['Category']) df1['Subcategory'] = lb.fit_transform(df1['Subcategory']) df1['Country'] = lb.fit_transform(df1['Country']) df1['State'] = lb.fit_transform(df1['State'])``` : ```from sklearn.model_selection import train_test_split``` : ```x_train, x_test, y_train, y_test = train_test_split(Scaled_x, y, test_size=0.2, random_state=42)```
Feature Engineering	Attached the codes in final submission.
Save Processed Data	Saved Processed Data 76]: ```X_standard``` 76]: ```array([[-0.63721416, -0.6340869 , 0.37522943, ..., -0.14902586, -0.58204005, -2.71397736], [-0.38179634, 1.06591714, 0.37522943, ..., -0.71078936, -0.58204005, -2.71397736], [-1.91430326, -0.23668336, 0.37522943, ..., -0.71078936, 1.25700291, -2.71397736], ..., [-0.12637851, -0.50161905, 0.37522943, ..., -0.71078936, 0.33748143, 1.94616419], [-1.91430326, -1.56136184, 0.37522943, ..., -0.75240147, 0.33748143, 1.94616419], [0.12903931, 1.22046297, -0.38238921, ..., 0.93288901, 0.33748143, 1.94616419]])```