Dataset: Movie Industry

Dataset description

Attribute name	Description
Budget	the budget of a movie. Some movies don't have this, so it appears as 0
company	the production company
country	country of origin
director	the director
genre	main genre of the movie.
gross	revenue of the movie
name	name of the movie
rating	rating of the movie (R, PG, etc.)
released	release date (YYYY-MM-DD)
runtime	duration of the movie
score	IMDb user rating
votes	number of user votes
star	main actor/actress
writer	writer of the movie
year	year of release

Data Preprocessing

DataFrame:

```
[ ] data.head()
```

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0	58853106.0	Columbia Pictures	104.0
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0	538375067.0	Lucasfilm	124.0
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0	83453539.0	Paramount Pictures	88.0
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000.0	39846344.0	Orion Pictures	98.0

Columns

```
[ ] data.columns

Index(['name', 'rating', 'genre', 'year', 'released', 'score', 'votes',
       'director', 'writer', 'star', 'country', 'budget', 'gross', 'company',
       'runtime'],
      dtype='object')
```

Null values

```
data.isnull().sum()
```

```
name          0
rating       77
genre         0
year          0
released      2
score         3
votes         3
director      0
writer        3
star          1
country       3
budget     2171
gross       189
company      17
runtime       4
dtype: int64
```

Filling numerical missing values with mean

```
df_reduced = data.fillna(data.mean())
df_reduced
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future versio
  """Entry point for launching an IPython kernel.
```

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.400000	9.270000e+05	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1.900000e+07	4.699877e+07	Warner Bros.	146.000000
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.800000	6.500000e+04	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4.500000e+06	5.885311e+07	Columbia Pictures	104.000000
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.700000	1.200000e+06	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	1.800000e+07	5.383751e+08	Lucasfilm	124.000000
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.700000	2.210000e+05	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3.500000e+06	8.345354e+07	Paramount Pictures	88.000000
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.300000	1.080000e+05	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6.000000e+06	3.984634e+07	Orion Pictures	98.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.100000	1.800000e+01	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7.000000e+03	7.850054e+07	NaN	90.000000

Filling Categorical data with “Unknown”

```
# filling with Unknown class
df_clean = df_reduced.fillna("Unknown")
df_clean
```

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.400000	9.270000e+05	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	1.900000e+07	4.699877e+07	Warner Bros.	146.000000
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.800000	6.500000e+04	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4.500000e+06	5.885311e+07	Columbia Pictures	104.000000
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.700000	1.200000e+06	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	1.800000e+07	5.383751e+08	Lucasfilm	124.000000
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.700000	2.210000e+05	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3.500000e+06	8.345354e+07	Paramount Pictures	88.000000
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.300000	1.080000e+05	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6.000000e+06	3.984634e+07	Orion Pictures	98.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7663	More to Life	Unknown	Drama	2020	October 23, 2020 (United States)	3.100000	1.800000e+01	Joseph Ebanks	Joseph Ebanks	Shannon Bond	United States	7.000000e+03	7.850054e+07	Unknown	90.000000

Standard Scaling to Normalise the data

```
[ ] from sklearn.preprocessing import StandardScaler
    scaled_features = df_clean.copy()

    col_names = ['gross','budget']
    features = scaled_features[col_names]
    scaler = StandardScaler().fit(features.values)
    features = scaler.transform(features.values)

    scaled_features[col_names] = features
    print(scaled_features)
```

```
1                                The Blue Lagoon      R  Adventure
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action
3                                         Airplane!     PG     Comedy
4                                       Caddyshack      R     Comedy
...                                            ...    ...        ...
7663                                  More to Life  Unknown     Drama
7664                                  Dream Round   Unknown     Comedy
7665                                Saving Mbango   Unknown     Drama
7666                                 It's Just Us   Unknown     Drama
7667                                    Tee em el   Unknown     Horror
```

Adding Success label as class attribute

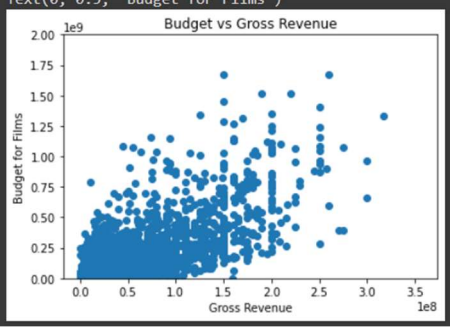
```
[ ] df_clean['success'] = df_clean.apply(lambda row: 1 if row["gross"] > row["budget"] else 0, axis = 1)
    df_clean.success.value_counts(normalize=True)

    1    0.524909
    0    0.475091
    Name: success, dtype: float64
```

EDA

Plots Explaining the data

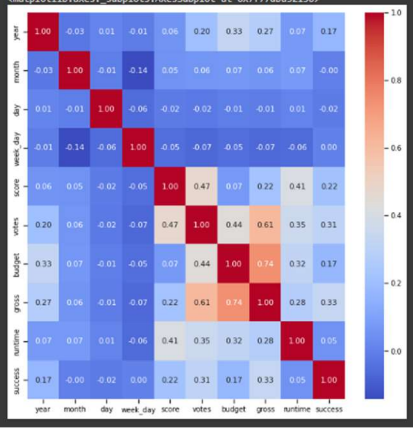
```
import matplotlib.pyplot as plt
plt.scatter(x=data['budget'], y=data['gross'])
plt.title('Budget vs Gross Revenue')
plt.ylim([0,2*10**9])
plt.xlabel('Gross Revenue')
plt.ylabel('Budget for Films')
```

```
Text(0, 0.5, 'Budget for Films')
```



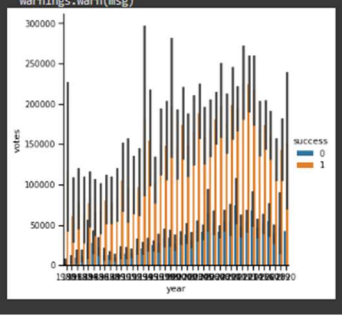

```
sns.set_context('notebook', font_scale=0.9)
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(ax=ax, data=data_reduced.corr(), annot=True, fmt= '.2f', cmap='coolwarm')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f77dba32150>
```

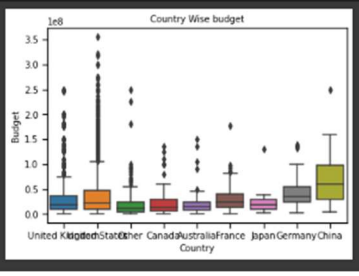


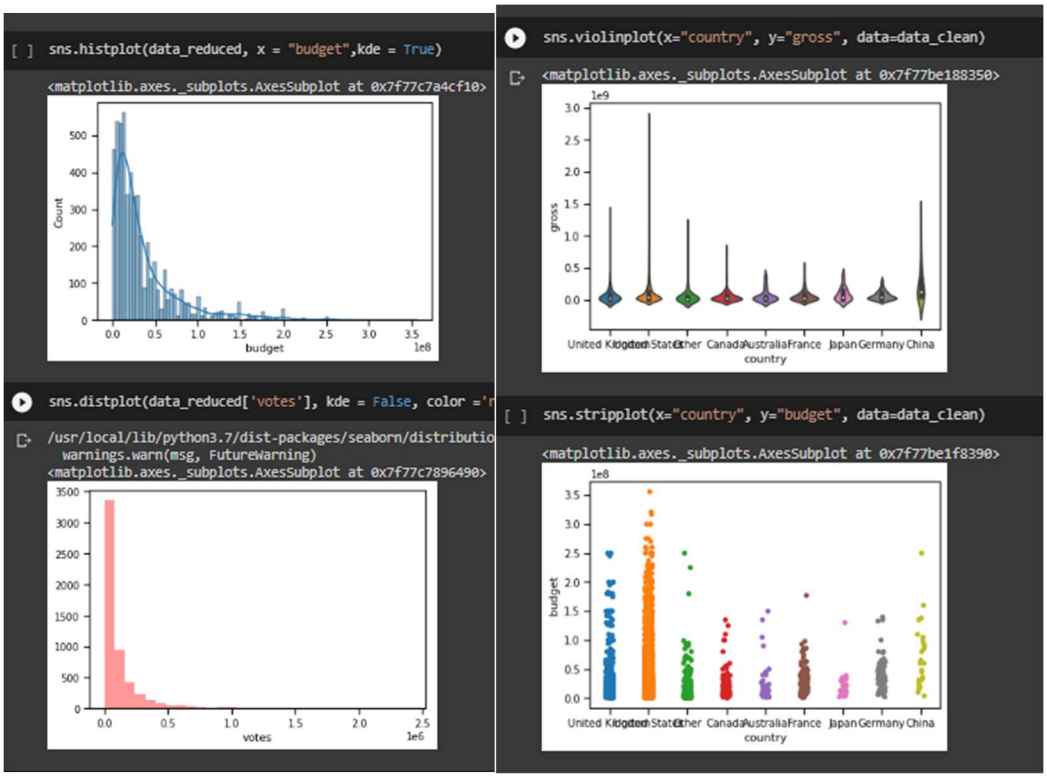

```
sns.factorplot(x="year", y="votes", hue="success", data=data_reduced, kind="bar");
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/categorical.py:3717: UserWarning: The `f
  warnings.warn(msg)
```




```
b=sns.boxplot(x="country", y="budget", data=data_reduced, dodge=False);
b.axes.set_title("Country Wise budget",fontsize=10)
b.set_xlabel("Country",fontsize=10)
b.set_ylabel("Budget",fontsize=10)
b.tick_params(labelsize=10)
```



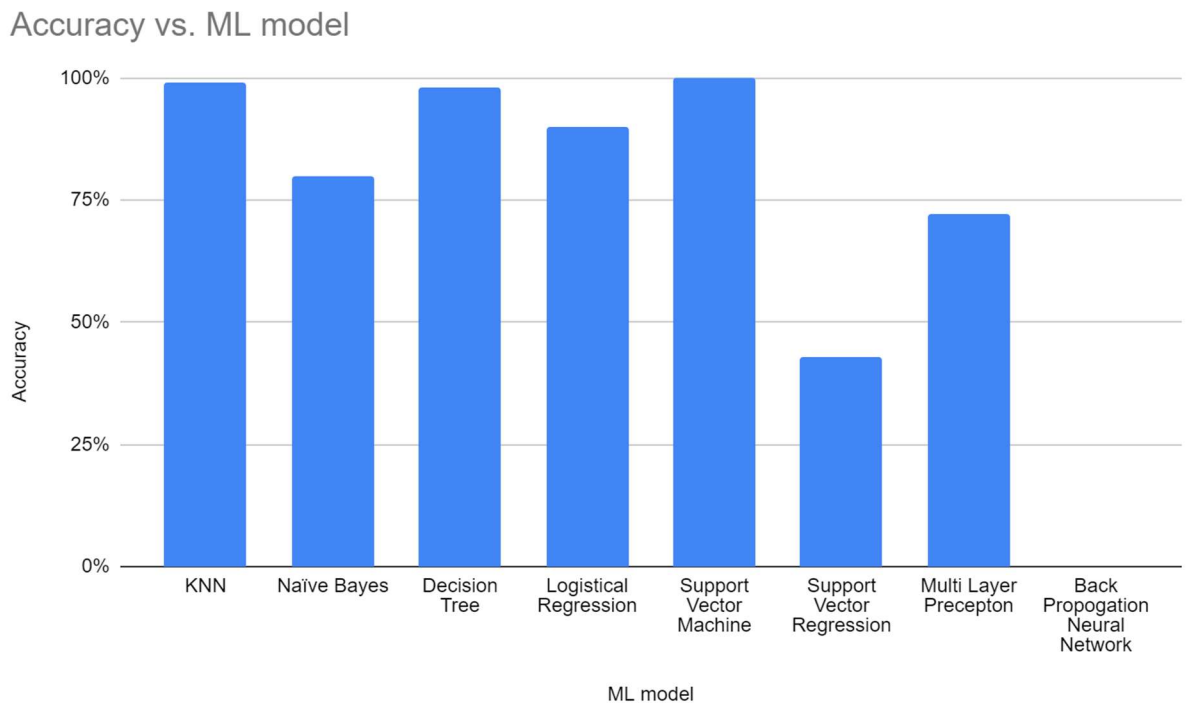


Comparision

Models Comparisions				
Problem Statement	**Independent Attributes**	**Dependent Attributes**	**ML model**	**Accuracy**
To determine Success of a movie based on various attributes	year', 'score', 'votes', 'budget', 'gross', 'runtime'	Success	KNN	99%
To determine Success of a movie based on various attributes	year', 'score', 'votes', 'budget', 'gross', 'runtime'	Success	Naïve Bayes	80%
To form clusters using K Means Clustering	Votes and Budget	6 clusters	K Means Clustering	Measure accuracy through visualisation (scatter plot) for different values of K. Determine the best value of K using elbow plot.
To form clusters using Heirarchal clustering	Votes and Budget	4 clusters	Heirarchical Clustering	Measure accuracy through visualisation (scatter plot). Determine the optimal no of

				clusters using dendogram.
Decision tree is plotted considering score, votes, budget, year,gross,runtime	score, votes, budget, year,gross,runtime	Success	Decision Tree	98%
To determine Success of a movie based on various attributes	Votes and Budget	Success	Logistical Regression	90%
Synthetic dataset is manufactured with the help of make_blobs	Synthetic dataset with 2 clusters	2 clusters	Support Vector Machine	100%
To determine Success of a movie based on various attributes	Votes and Budget	Success	Support Vector Regression	43%
To determine the Success using MLP	Votes and Budget	Success	Multi Layer Precepton	72.00%
To determine the Success using BLP	Votes and Budget	Success	Back Propogation Neural Network	0.00%

Accuracy Comparision



Colab Link:

https://colab.research.google.com/drive/1bFqUE5ynvHVVw87Acn-SSA0snrQTUbCD?usp=sharing

Github Link:

https://github.com/Vaishnavikm/Machine-Learning-Movie-Industry-Analysis