AlphaSimplex - Movie Data

By: Jason Q Huang
Date: 9/9/20

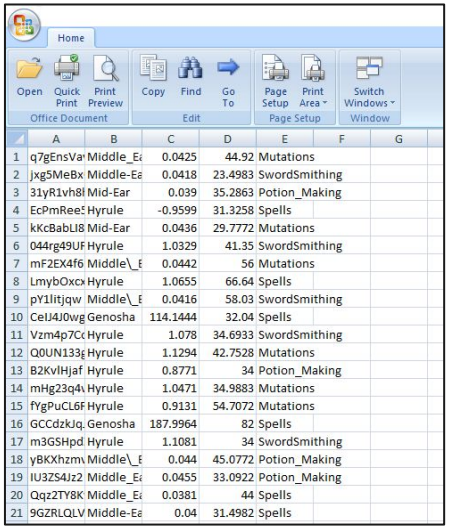
Agenda

1. Data Overview
2. Data Cleaning / EDA
3. Insights
4. Functions
5. Future Steps

Data Overview

Data Overview

Original



	A	B	C	D	E	F	G
1	q7gEnsVa	Middle_Ea	0.0425	44.92	Mutations		
2	jxg5MeBx	Middle-Ea	0.0418	23.4983	SwordSmithing		
3	31yR1vh8	Mid-Ear	0.039	35.2863	Potion_Making		
4	EcPmRee	Hyrule	-0.9599	31.3258	Spells		
5	kKcBabLI	Mid-Ear	0.0436	29.7772	Mutations		
6	044rg49U	Hyrule	1.0329	41.35	SwordSmithing		
7	mF2EX4f6	Middle_E	0.0442	56	Mutations		
8	LmybOxcx	Hyrule	1.0655	66.64	Spells		
9	pY1litjqw	Middle_E	0.0416	58.03	SwordSmithing		
10	CeIJ4J0wg	Genosha	114.1444	32.04	Spells		
11	Vzm4p7C	Hyrule	1.078	34.6933	SwordSmithing		
12	Q0UN133	Hyrule	1.1294	42.7528	Mutations		
13	B2KvlHjaf	Hyrule	0.8771	34	Potion_Making		
14	mHg23q4	Hyrule	1.0471	34.9883	Mutations		
15	fYgPuCL6F	Hyrule	0.9131	54.7072	Mutations		
16	GCCdzkJq	Genosha	187.9964	82	Spells		
17	m3GSHpd	Hyrule	1.1081	34	SwordSmithing		
18	yBKXhzm	Middle_E	0.044	45.0772	Potion_Making		
19	IU3ZS4Jz2	Middle_Ea	0.0455	33.0922	Potion_Making		
20	Qqz2TY8K	Middle_Ea	0.0381	44	Spells		
21	9GZRLQLV	Middle-Ea	0.04	31.4982	Spells		

Combined Dataframe

	name	country	ticket_price	avg_rating	genre
0	d19rpnckGv	Hyrule	1.0935	62.0000	SwordSmithing
1	KaWRtnToJC	Hyrule	0.7481	28.0000	SwordSmithing
2	fJkJTvK9pT	Middle_Earth	0.0437	38.4386	SwordSmithing
3	lbDBE8KYaU	Middle_Earth	0.0417	73.0533	SwordSmithing
4	sBRX9lLull	Hyrule	0.9455	57.3400	Spells
...	...	...	...	...	...
3328	NqQBsiArt	Genosha	131.5168	37.1521	SwordSmithing
3329	G6OInlu5F	Genosha	187.0604	77.8402	Spells
3330	f24B8j0ss	Hyrule	0.9642	69.5299	SwordSmithing
3331	kvGl6XnoF	Middle_Earth	0.0400	22.8600	Spells
3332	GuHWFz1MB	Hyrule	1.0368	79.7148	Spells

9999 rows × 5 columns

Data Overview

	name	country	ticket_price	avg_rating	genre
0	d19rpnckGv	Hyrule	1.0935	62.0000	SwordSmithing
1	KaWRtnToJC	Hyrule	0.7481	28.0000	SwordSmithing
2	fJkJTvK9pT	Middle_Earth	0.0437	38.4386	SwordSmithing
3	lbDBE8KYaU	Middle_Earth	0.0417	73.0533	SwordSmithing
4	sBRX9lLull	Hyrule	0.9455	57.3400	Spells
...	...	...	...	...	...
3328	NqQBsiArt	Genosha	131.5168	37.1521	SwordSmithing
3329	G6OInlu5F	Genosha	187.0604	77.8402	Spells
3330	f24B8j0ss	Hyrule	0.9642	69.5299	SwordSmithing
3331	kvGl6XnoF	Middle_Earth	0.0400	22.8600	Spells
3332	GuHWFz1MB	Hyrule	1.0368	79.7148	Spells

9999 rows × 5 columns

Columns:

1. name: Movie name
2. country: Country of origin
3. ticket_price: Last known ticket price
4. avg_rating: Average critic rating
5. genre: movie genre

Data Overview

	name	country	ticket_price	avg_rating	genre
0	d19rpnckGv	Hyrule	1.0935	62.0000	SwordSmithing
1	KaWRtnToJC	Hyrule	0.7481	28.0000	SwordSmithing
2	fJkJTvK9pT	Middle_Earth	0.0437	38.4386	SwordSmithing
3	lbDBE8KYaU	Middle_Earth	0.0417	73.0533	SwordSmithing
4	sBRX9lLull	Hyrule	0.9455	57.3400	Spells
...	...	...	...	...	...
3328	NqQBsiArt	Genosha	131.5168	37.1521	SwordSmithing
3329	G6OInlu5F	Genosha	187.0604	77.8402	Spells
3330	f24B8j0ss	Hyrule	0.9642	69.5299	SwordSmithing
3331	kvGl6XnoF	Middle_Earth	0.0400	22.8600	Spells
3332	GuHWFz1MB	Hyrule	1.0368	79.7148	Spells

9999 rows × 5 columns

Countries:

1. Genosha
2. Hyrule
3. Middle Earth

Ticket Price:

Unknown currencies

Genres:

1. SwordSmithing
2. Spells
3. Mutations
4. Potion Making

Average Rating:

Ratings from 0-100

Data Cleaning / EDA

Data Cleaning / EDA - Duplicates

	name	country	ticket_price	avg_rating	genre
23	5xUIWm4jnH	Middle_Earth	0.0422	47.3764	Spells
25	JqMFZNOId0	Middle_Earth	0.0410	51.1800	Mutations
53	LmijTifuUV	Hyrule	0.9894	59.6746	Mutations
106	kDR3dpTkVC	Hyrule	1.0400	64.5252	Mutations
145	CkGj32GGCu	Hyrule	1.0361	33.6472	Mutations
...	...	...	...	...	...
3082	HEy2Cao2qc	Middle_Earth	0.0445	54.9200	SwordSmithing
3102	UFiifyZ6y1	Hyrule	0.9885	29.9875	SwordSmithing
3127	5Filgaug69	Genosha	157.6355	56.5093	Spells
3255	IdghRbHrf8	Hyrule	1.0123	75.7800	Potion_Making
3277	ixFuwQ4RzY	Hyrule	1.0227	58.5615	Mutations

132 rows × 5 columns

	name	country	ticket_price	avg_rating	genre
23	5xUIWm4jnH	Hyrule	0.9589	49.9253	Mutations
23	5xUIWm4jnH	Middle_Earth	0.0422	47.3764	Spells

Duplicates:

132 cases of shared movie names.

Verified same movie name but unique entry.

Data Cleaning / EDA - Typo's

Check for Typo's in Data

In [6]:
```
data_combined.genre.unique()
```

Out[6]:
```
array(['SwordSmithing', 'Spells', 'Mutations', 'Potion_Making'],
      dtype=object)
```

No typo's in Genre column.

In [7]:
```
data_combined.country.unique()
```

Out[7]:
```
array(['Hyrule', 'Middle-Earth', 'Middle_Earth', 'Middle\\_Earth',
       'Mid-Ear', 'Genosha'], dtype=object)
```

Found a couple of typo's in the country column.

In [9]:
```
# Checking
data_combined.country.unique()
```

Out[9]:
```
array(['Hyrule', 'Middle_Earth', 'Genosha'], dtype=object)
```

Typo's:

Corrected typo's for Middle Earth.

Data Cleaning / EDA - Overview

Ticket Prices Grouped by Country

	ticket_price							
	count	mean	std	min	25%	50%	75%	max
country								
Genosha	497.0	1.398716e+09	1.556169e+10	-213.2527	120.4246	142.1965	161.3713	1.838831e+11
Hyrule	5941.0	9.216631e+06	9.505584e+07	-1.2845	0.9269	0.9972	1.0684	1.250900e+09
Middle_Earth	3462.0	4.672733e+05	4.384331e+06	-0.0482	0.0404	0.0419	0.0434	4.550000e+07

Action Items:

1. Negative Numbers
2. Outliers
3. Empty / Null values

Data Cleaning / EDA - Negative Values

	name	country	ticket_price	avg_rating	genre
42	MFsjlY22c8	Genosha	-128.6223	33.4917	SwordSmithing
77	Q04jpRscMu	Middle_Earth	-0.0425	65.0400	SwordSmithing
112	eEa2ZQUVxV	Middle_Earth	-0.0430	55.0000	Mutations
169	O9MEuAZDIR	Middle_Earth	-0.0394	83.6464	Spells
178	kK4umWUVRq	Hyrule	-0.8454	42.9900	Spells
...	...	...	...	...	...
3258	s1xxwolRi	Genosha	-157.6664	55.7200	Potion_Making
3263	boTM4vOZ1	Middle_Earth	-0.0418	74.4341	Spells
3266	taycnnMMK	Middle_Earth	-0.0416	50.9029	Mutations
3268	x3F5nDICl	Hyrule	-1.2058	62.7968	Mutations
3304	rHhKxBOuD	Hyrule	-0.9091	67.0000	Mutations

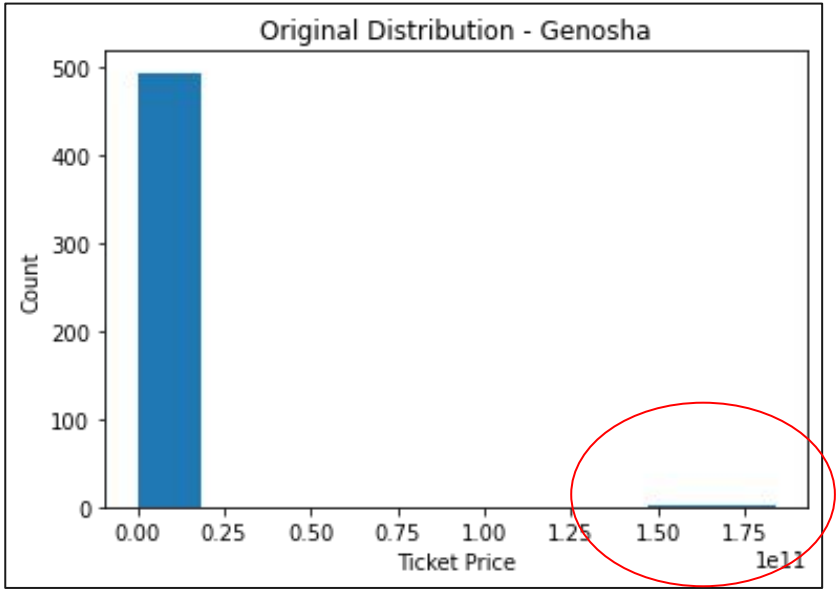
297 rows × 5 columns

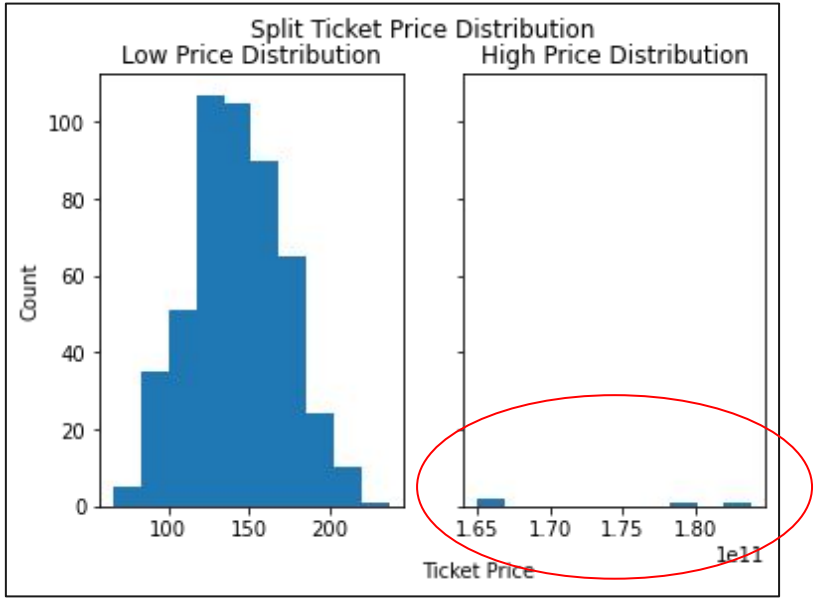
Assumption:

Absolute value of negative ticket prices.

Data Cleaning / EDA - Outliers

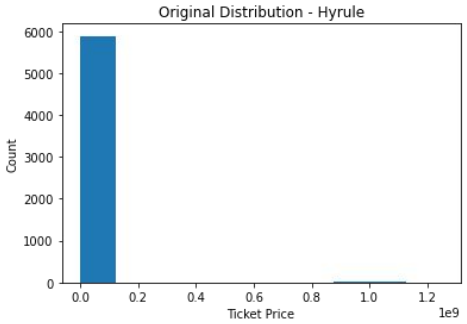
Genosha

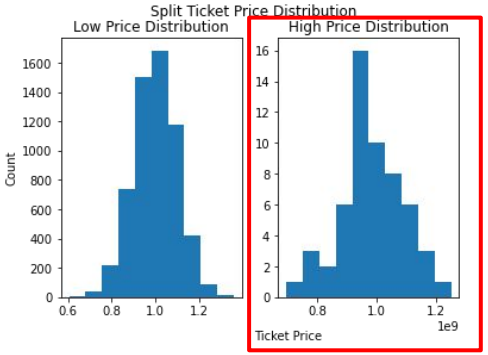


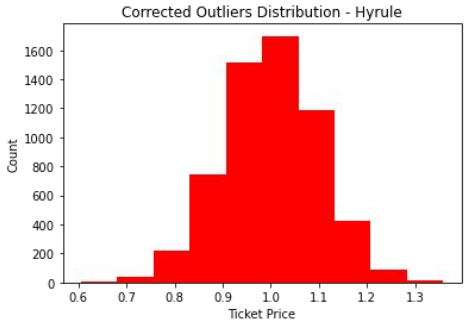


Data Cleaning / EDA - Outliers

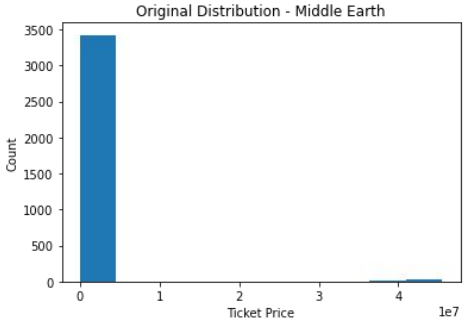
Hyrule

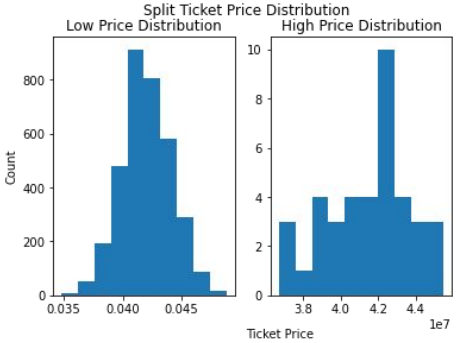


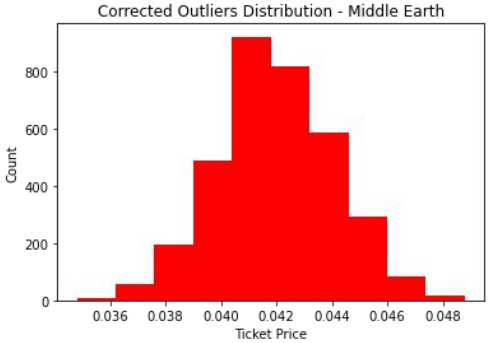




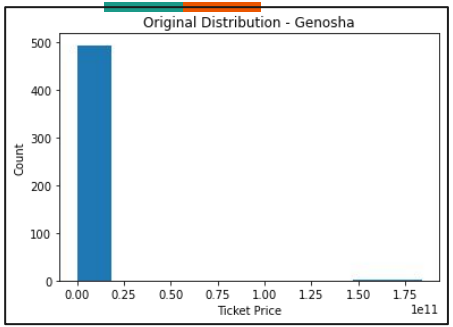
Middle Earth

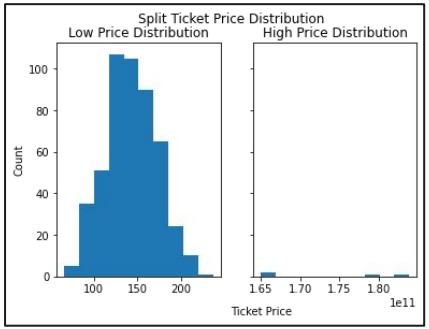


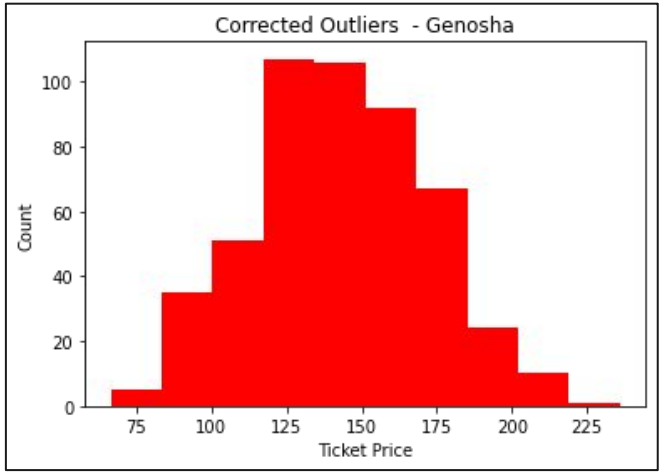




Data Cleaning / EDA - Outliers







Assumptions:

Pattern in high end prices.

Prices were inflated by ~1e11 times original price.

Adjusted per country basis to account for currency differences.

Data Cleaning / EDA - Null Values

```
data_combined.isnull().sum()
```

```
name            0
country         0
ticket_price   99
avg_rating      0
genre           0
dtype: int64
```

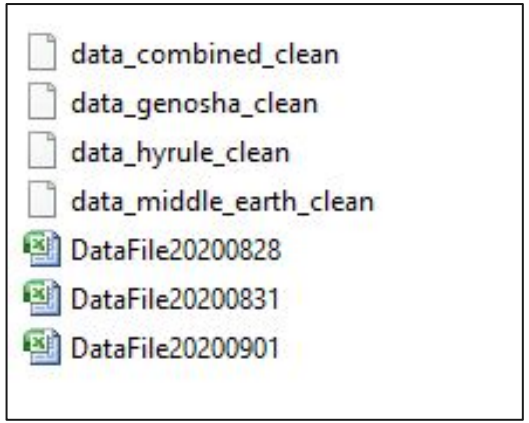
```
Genosha: 143.02870905432596
Hyrule: 1.0006156707624951
Middle Earth: 0.04196984402079717
```

Assumptions:

Nulls were imputed with country specific mean values.

Accounts for currency differences, outliers, and negative values.

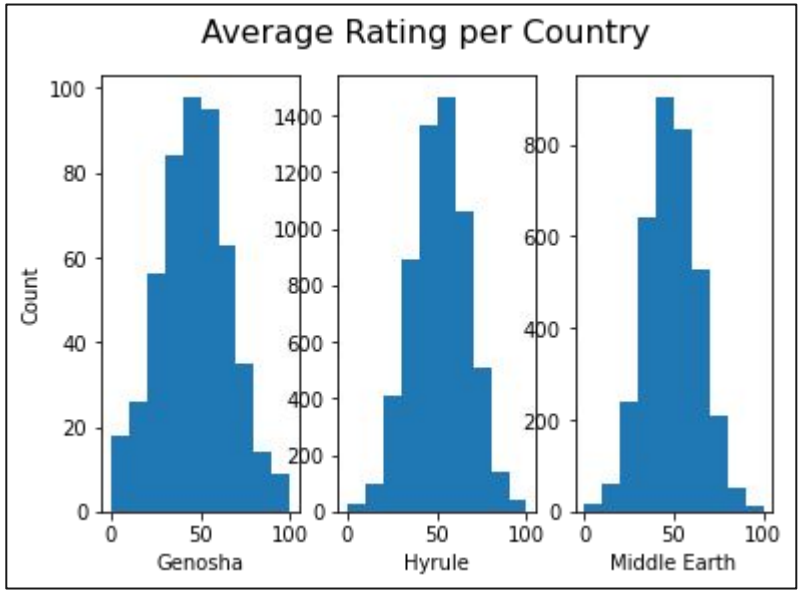
Data Cleaning / EDA - Export to CSV

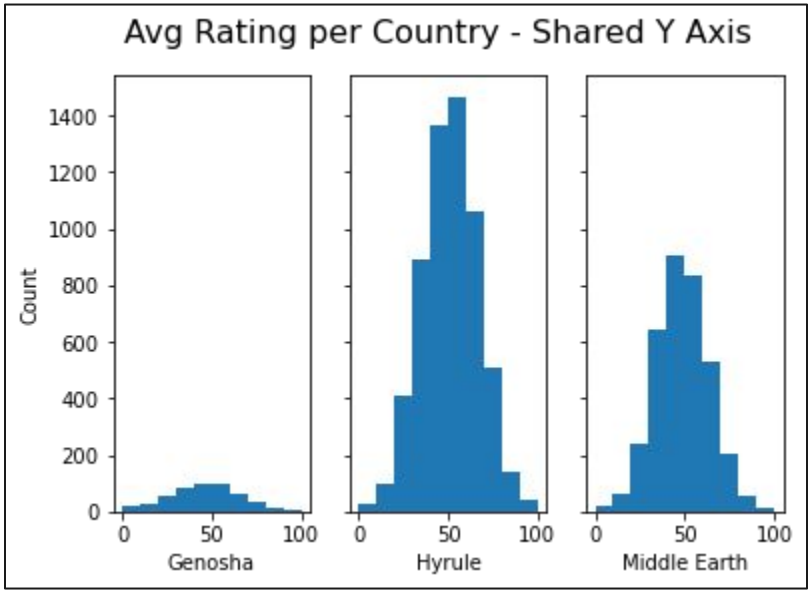


Can be exported to .CSV / .EXLS format.

Insights

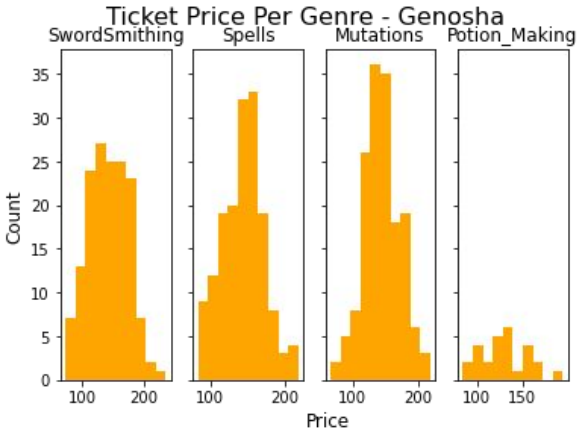
Insights - Average Rating per Country

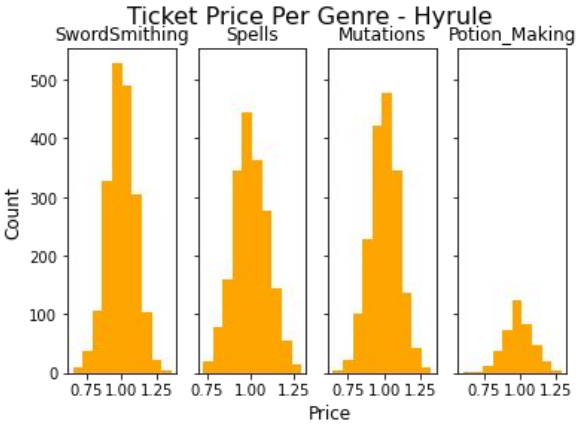


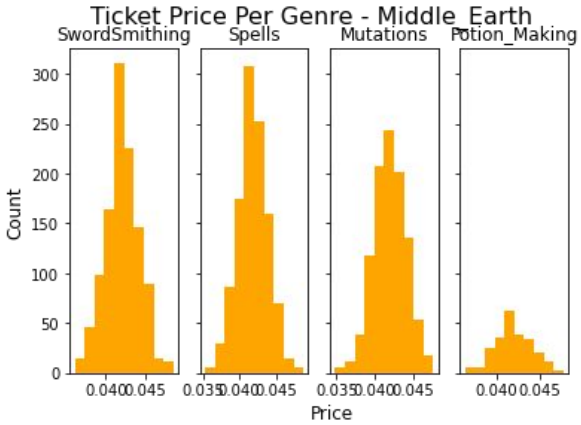


Observation:
All average ratings are normally distributed between 0-100.

Insights - Price and Number of Tickets per Genre & Country

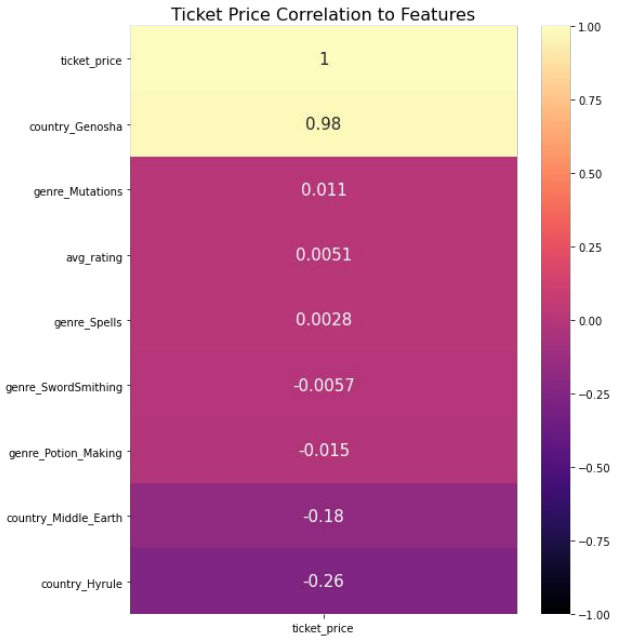


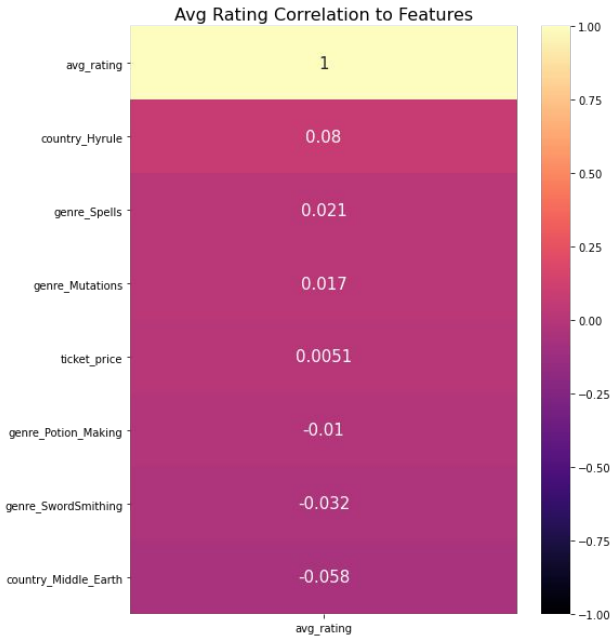




Observation:
Potion_Making isn't popular in any of the 3 countries.
Hyrule is the largest market and spends the most on SwordSmithing films.

Insights - Correlation Heatmap





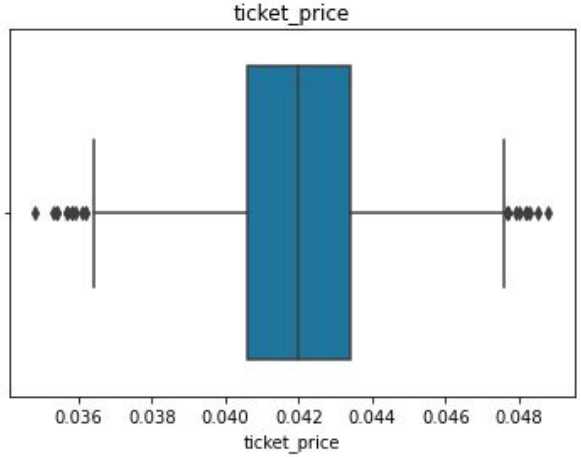
Observation:
Strong ticket price correlation is explained by Genosha's unadjusted currency.

Functions

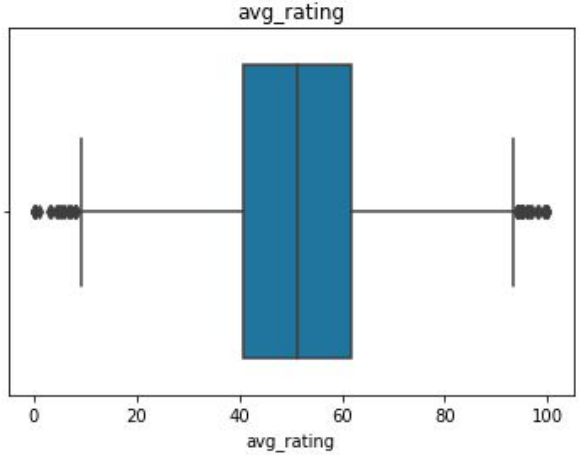
Functions - Boxplot for Country

```python
def boxplot_country(country, metric):
    temp_df = data.groupby(by='country').get_group(country)
    ax = sns.boxplot(x=temp_df[metric]).set_title(metric)
    return ax
```

```python
boxplot_country('Middle_Earth', 'ticket_price');
```




```python
boxplot_country('Hyrule', 'avg_rating');
```



Functions - Dataframe for Average Rating Ranges per Country

Dataframe for Average Rating Ranges per Country

```python
def country_per_rating_df(rating, operator, country):
    if operator == '>':
        return data.loc[(data.avg_rating > rating)&(data.country == country), :]
    elif operator == '>=':
        return data.loc[(data.avg_rating >= rating)&(data.country == country), :]
    elif operator == '<':
        return data.loc[(data.avg_rating < rating)&(data.country == country), :]
    elif operator == '<=':
        return data.loc[(data.avg_rating <= rating)&(data.country == country), :]
    else:
        return print('Check Operator. Spelling must be '>', '>=', '<', or '<=''')
```

```python
# Dataframe for Genosha movies that have ">=" 50 rating.
country_per_rating_df(50, '>=', 'Genosha')
```

	name	country	ticket_price	avg_rating	genre
1	xk8b9MocTE	Genosha	176.2547	62.8047	SwordSmithing
4	ThpLWmUBkv	Genosha	175.4893	62.1337	SwordSmithing
9	HrSU41PPs7	Genosha	166.9742	56.9700	Mutations
10	IQGrNcwfD1	Genosha	173.0252	61.0000	Spells
11	Jo3AOCQT1w	Genosha	176.6632	63.1297	Mutations
...	...	...	...	...	...
491	5Filgaug69	Genosha	157.6355	56.5093	Spells
492	VQt8334ia	Genosha	159.0684	56.8655	SwordSmithing
494	iAebg1Wst	Genosha	155.5371	54.4200	Spells
495	s1xxwolRi	Genosha	157.6664	55.7200	Potion_Making
497	G6OInlu5F	Genosha	187.0604	77.8402	Spells

216 rows × 5 columns

Functions - Dataframe for Genre per Country

```
def groupby_country_genre_df(country_genre):
    if (country_genre == 'Genosha') or (country_genre == 'Hyrule') or (country_genre == 'Middle_Earth'):
        return data.groupby(by='country').get_group(country_genre)
    elif (country_genre == 'SwordSmithing') or (country_genre == 'Spells') or (country_genre == 'Mutations') or (country_gen
        return data.groupby(by='genre').get_group(country_genre)
    else:
        return print('Check Spelling')
```

```
groupby_country_genre_df('Hyrule')
```

	name	country	ticket_price	avg_rating	genre
498	d19rpnckGv	Hyrule	1.0935	62.0000	SwordSmithing
499	KaWRtnToJC	Hyrule	0.7481	28.0000	SwordSmithing
500	sBRX9lLull	Hyrule	0.9455	57.3400	Spells
501	O7B9AjXpFt	Hyrule	0.9657	48.6407	SwordSmithing
502	1HpLXjq9Dz	Hyrule	0.9404	42.5596	SwordSmithing
...	...	...	...	...	...
6496	VKYFLLMCa	Hyrule	0.9951	42.6700	Mutations
6497	BXud2exyC	Hyrule	0.9163	46.1300	Spells
6498	Z1qfZ5oWC	Hyrule	1.0227	44.0000	Potion_Making
6499	f24B8j0ss	Hyrule	0.9642	69.5299	SwordSmithing
6500	GuHWFz1MB	Hyrule	1.0368	79.7148	Spells

6003 rows × 5 columns

Future Steps

Future Steps

1. Reassess Assumptions
    a. Impute nulls with median vs. mean.
    b. Determine conversion rate on currency to isolate purchasing power.
2. Further EDA
    a. Explore data without outliers and data adjustment.
3. Functions can be easily tailored for specific queries.

Questions?