Game market: analysis and prediction

Data visualization with Tableau
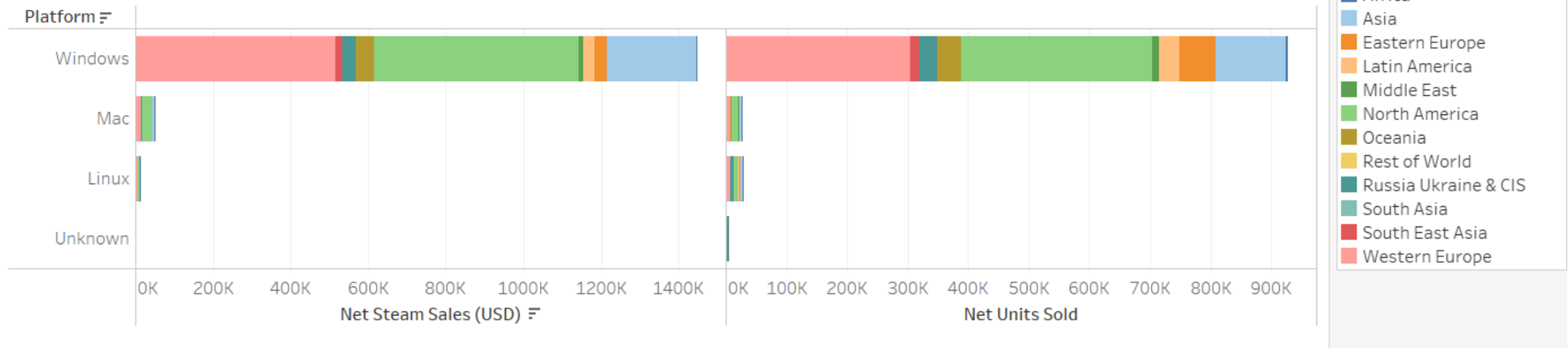

Total sale and sold at difference region
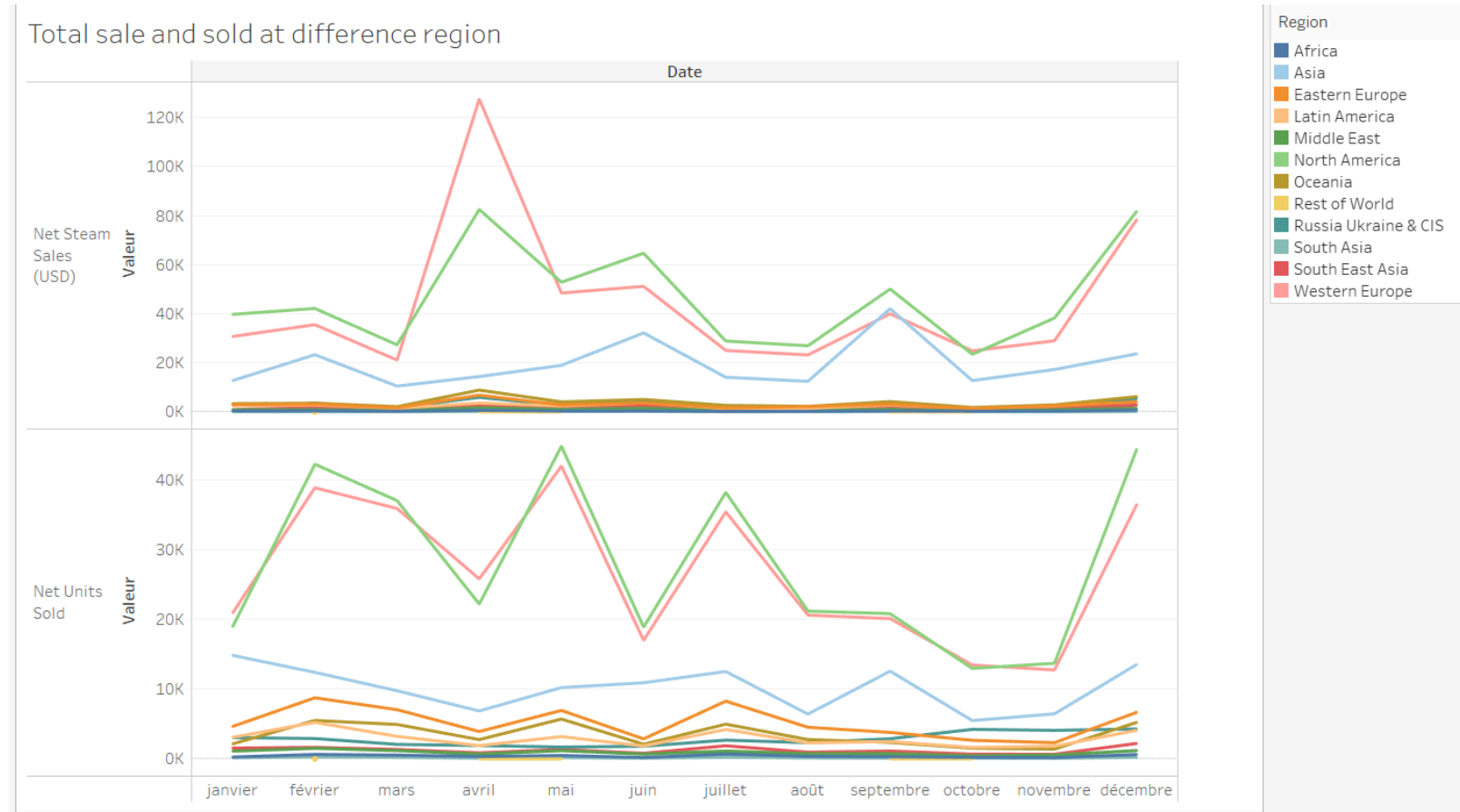
The 2 most region market are Western Europe and North America

Total Sale and Sold

Region
Asia
North America
Oceania
Russia Ukraine & CIS
Western Europe

Platform
Linux
Mac
Unknown
Windows

| Country	Net Steam Sales (USD)	Net Units Sold
United States	466 714	267 665
France	157 050	49 396
Germany	150 195	77 191
China	112 172	66 260
United Kingdom	65 080	57 655
Canada	62 658	48 069
Japan	60 930	
Korea Republic of		
Australia		32 535
Russian Federation		
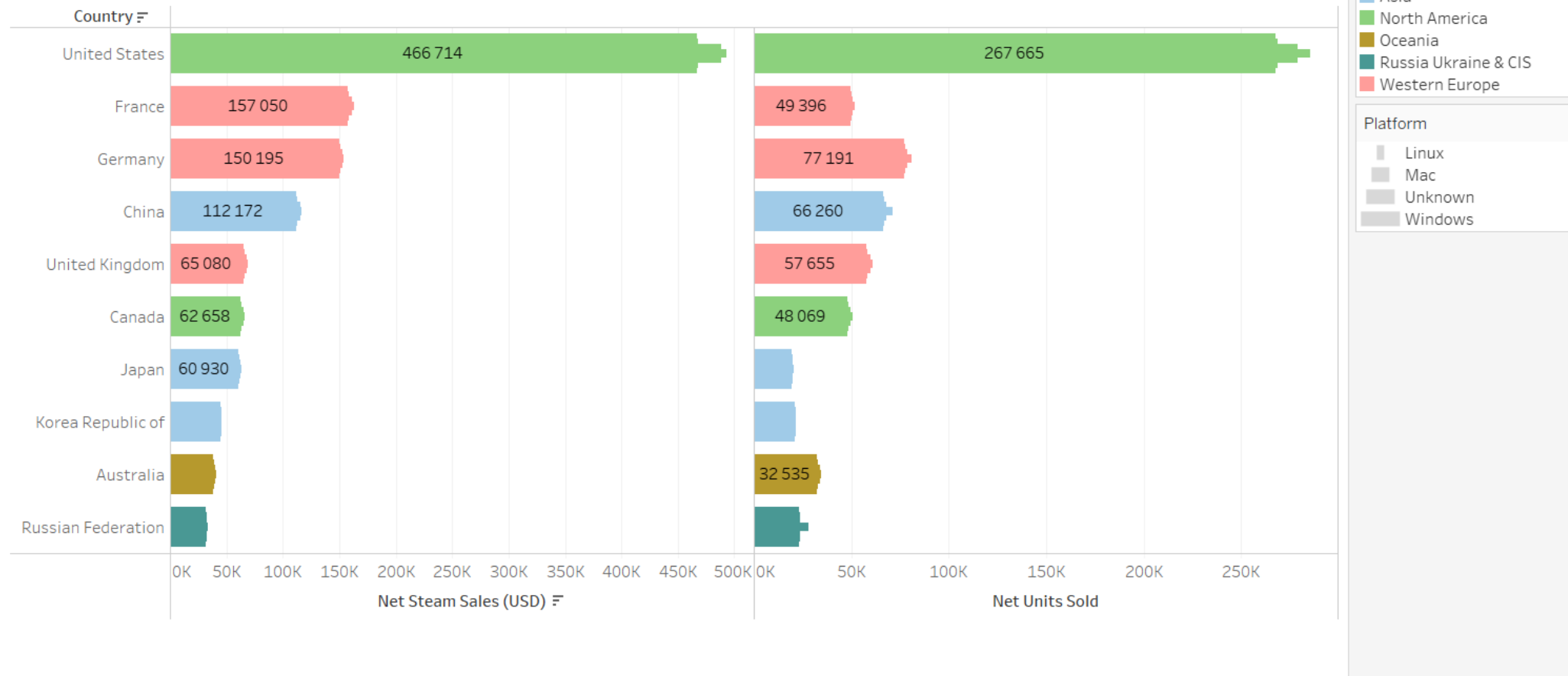
The United States is the biggest market for game publisher.

Total sale and sold

Region	Platform			
	Windows	Mac	Linux	Unknown
Asia	97.44%	2.41%	0.15%	0.00%
Western Europe	96.45%	2.61%	0.93%	0.00%
Latin America	96.39%	2.05%	1.57%	0.00%
Eastern Europe	96.28%	1.87%	1.85%	0.00%
Middle East	95.91%	2.99%	1.10%	0.00%
Russia Ukraine ..	95.07%	3.38%	1.54%	0.00%
North America	94.78%	4.35%	0.86%	0.00%
South East Asia	94.45%	4.93%	0.62%	0.00%
Oceania	94.30%	4.82%	0.88%	0.00%
Africa	93.27%	4.50%	2.23%	0.00%
South Asia	91.03%	6.43%	2.54%	0.00%

Window is the biggest platform for game player with over 90% for all the region

Data visualization with Python
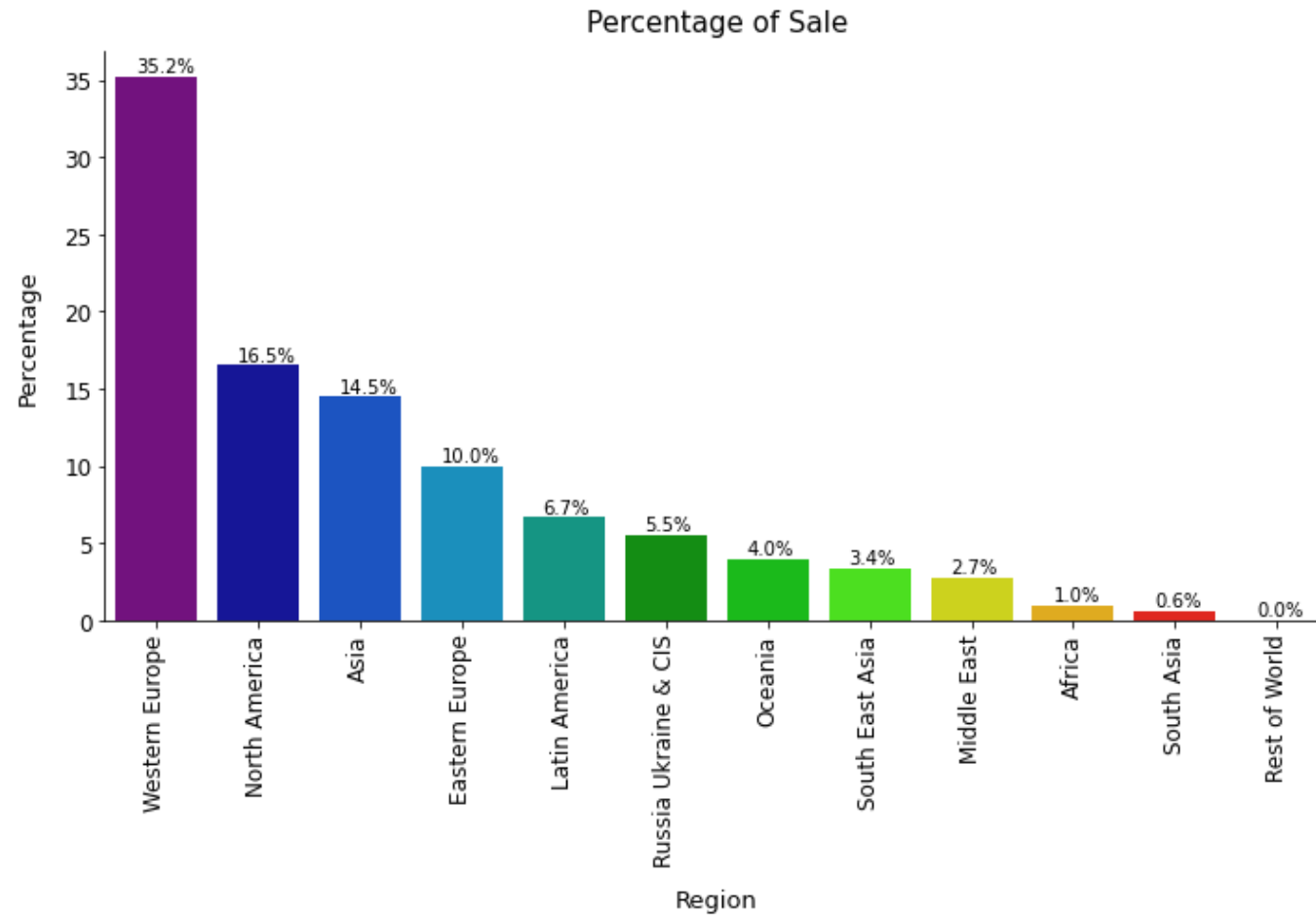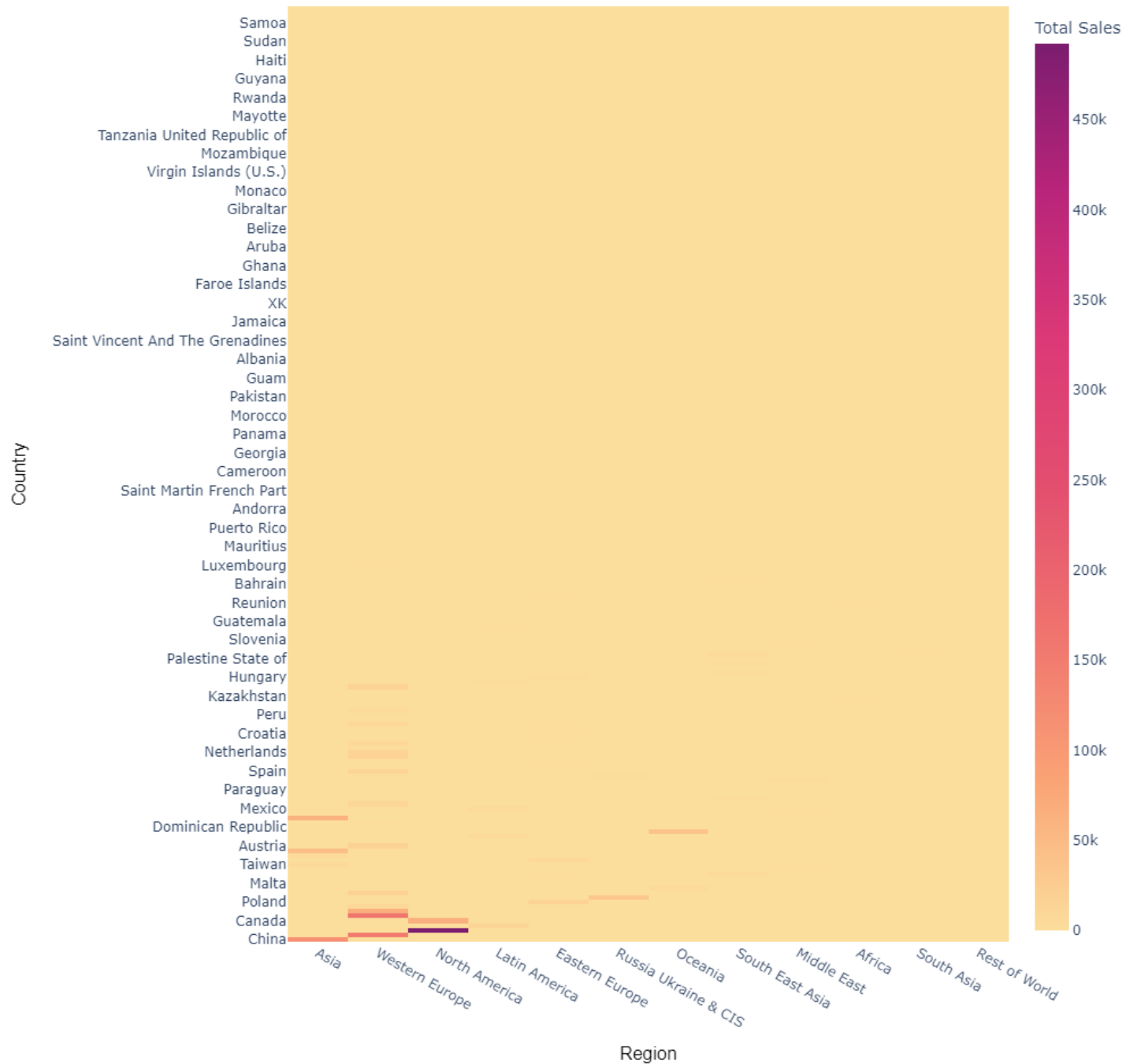
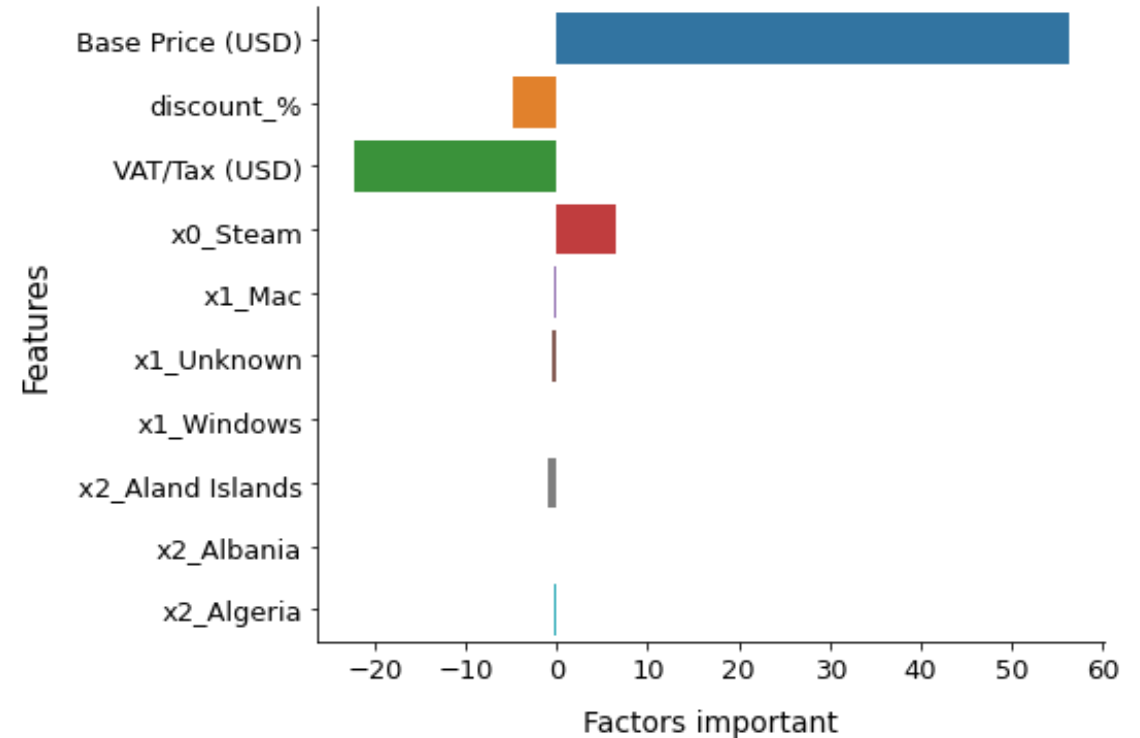The United States is the biggest market for game publisher.

Machine Learning

Prediction total sale

```
# calculate discount percentage
df["discount_%"] = ((df["Base Price"]-df["Sale Price"])/df["Base Price"])*100
df["discount_%"] = df["discount_%"].fillna(0)

# convert Base Price to USD
df["Base Price (USD)"] = df["Base Price"]*(df["Gross Steam Sales
(USD)"]/df["Sale Price"])
df["Base Price (USD)"] = df["Base Price (USD)"].fillna(0)

# all the country having value count < 10 is convert to "other"
mask = df["Country"].map(df["Country"].value_counts()) < 10
df["Country"] =  df["Country"].mask(mask, "other")

numeric features: ['Base Price (USD)', 'discount_%', 'VAT/Tax (USD)']
categorical features ['Type', 'Platform', 'Country']
```

Machine Learning

Prediction total sold

```
# calculate discount percentage
df["discount_%"] = ((df["Base Price"]-df["Sale Price"])/df["Base Price"])*100
df["discount_%"] = df["discount_%"].fillna(0)

# convert Base Price to USD
df["Base Price (USD)"] = df["Base Price"]*(df["Gross Steam Sales
(USD)"]/df["Sale Price"])
df["Base Price (USD)"] = df["Base Price (USD)"].fillna(0)

df_ML1 = df_ML[df_ML["Base Price (USD)"] != 0]


# all the country having value count < 10 is convert to "other"
mask = df_ML1["Country"].map(df_ML1["Country"].value_counts()) < 10
df_ML1["Country"] =  df_ML1["Country"].mask(mask, 'other')
```

numeric features: ['Base Price (USD)', 'discount_%', 'VAT/Tax (USD)']
categorical features ['Type', 'Platform', 'Country']
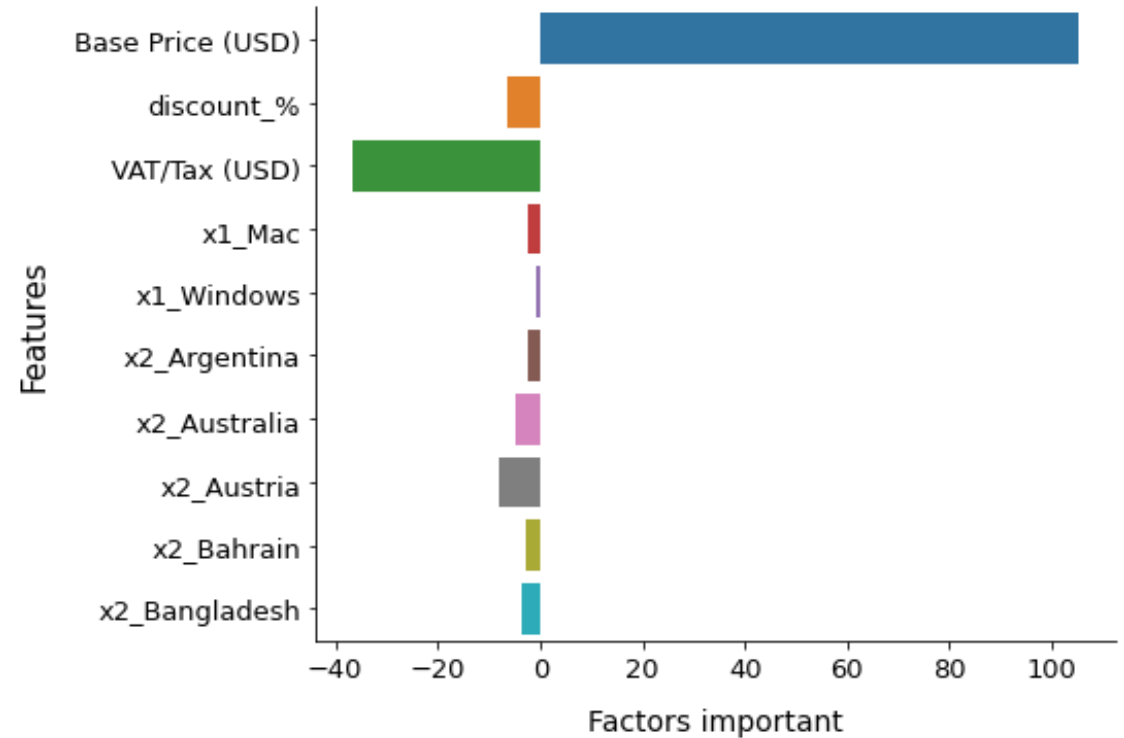
In both case, the most importance feature is « base price » of product.