Video Game Sales

Katie Christensen, Ian Cullum, and Seth Walloch

About the Dataset

About the Dataset

- Information in the dataset
- Why we chose this dataset
- What we anticipated to find

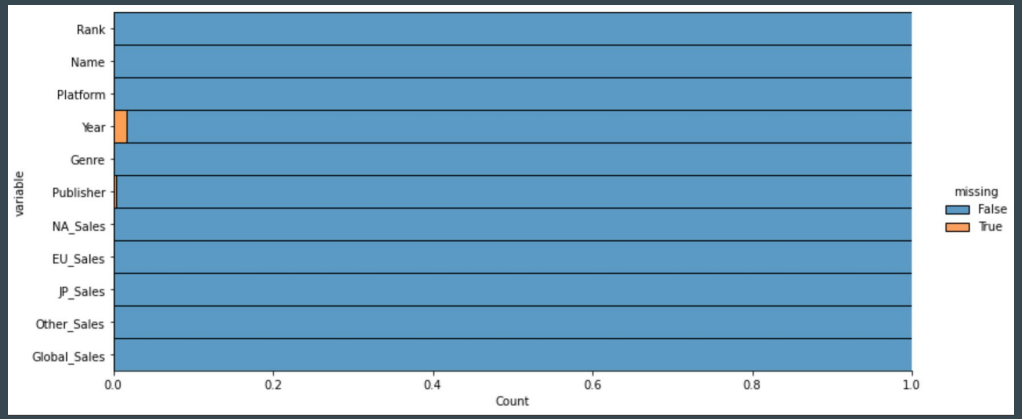
	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
...	...	...	...	...	...	...	...	...	...	...	...
16593	16596	Woody Woodpecker in Crazy Castle 5	GBA	2002.0	Platform	Kemco	0.01	0.00	0.00	0.00	0.01
16594	16597	Men in Black II: Alien Escape	GC	2003.0	Shooter	Infogrames	0.01	0.00	0.00	0.00	0.01
16595	16598	SCORE International Baja 1000: The Official Game	PS2	2008.0	Racing	Activision	0.00	0.00	0.00	0.00	0.01
16596	16599	Know How 2	DS	2010.0	Puzzle	7G//AMES	0.00	0.01	0.00	0.00	0.01
16597	16600	Spirits & Spells	GBA	2003.0	Platform	Wanadoo	0.01	0.00	0.00	0.00	0.01

16598 rows × 11 columns

Data Exploration

Data Exploration

- Visualization show columns containing N/A
- Counted number of N/A
- Checked the attribute types
- Counted number of rows and datapoints



Data Preparation

Data Preparation

- Dropped N/A
- Shuffled the data

- Chose the x-values to use for clustering algorithms:
  - NA_Sales, EU_Sales, JP_Sales, Other_Sales
- Chose the y-value to use for visualizations later:
  - Global_Sales

```
[13] x_values = df[["NA_Sales","EU_Sales","JP_Sales","Other_Sales"]].to_numpy()
     y_values = df['Global_Sales'].to_numpy()
```

Data Training

Data Training - Performance Evaluation

The Silhouette Coefficient

- Bounded between -1 and +1
  - -1 for incorrect clustering
  - +1 for highly dense clustering
  - 0 indicates overlapping clusters
- Higher value when clusters are dense and well separated

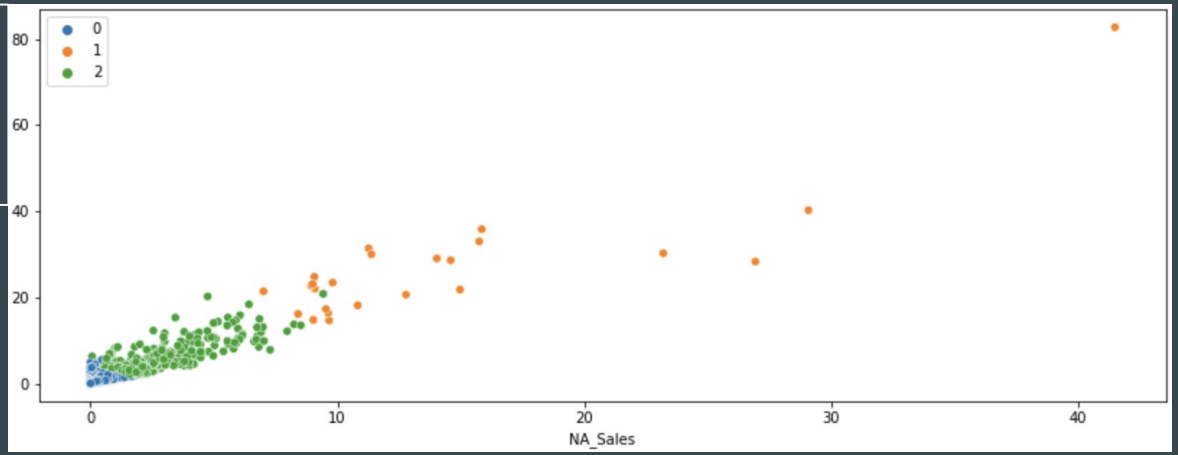
Data Training - KMeans

Silhouette Score: **0.8436**

0 = Low Sales (Blue)

1 = High Sales (Orange)

2 = Medium Sales (Green)



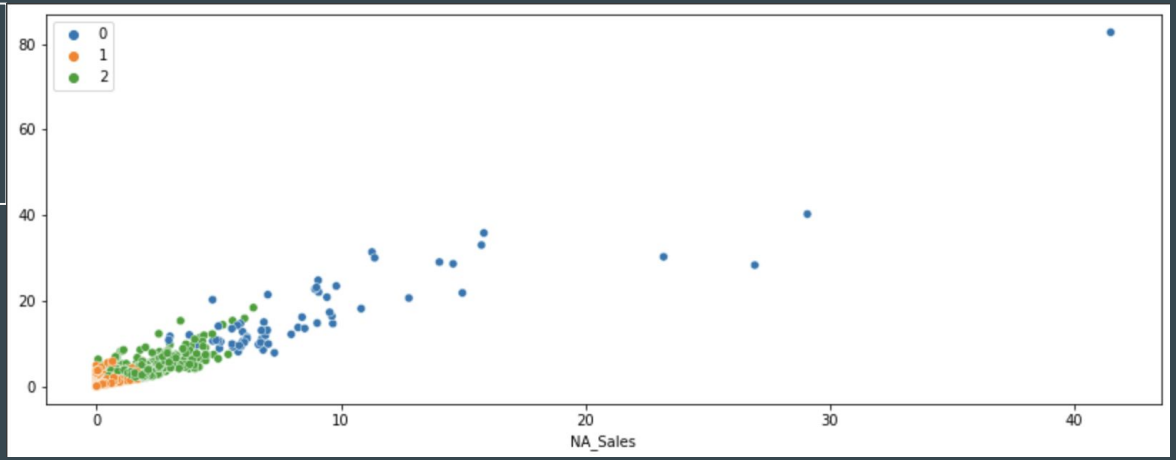
Data Training - Agglomerative

Silhouette Score: **0.8524**

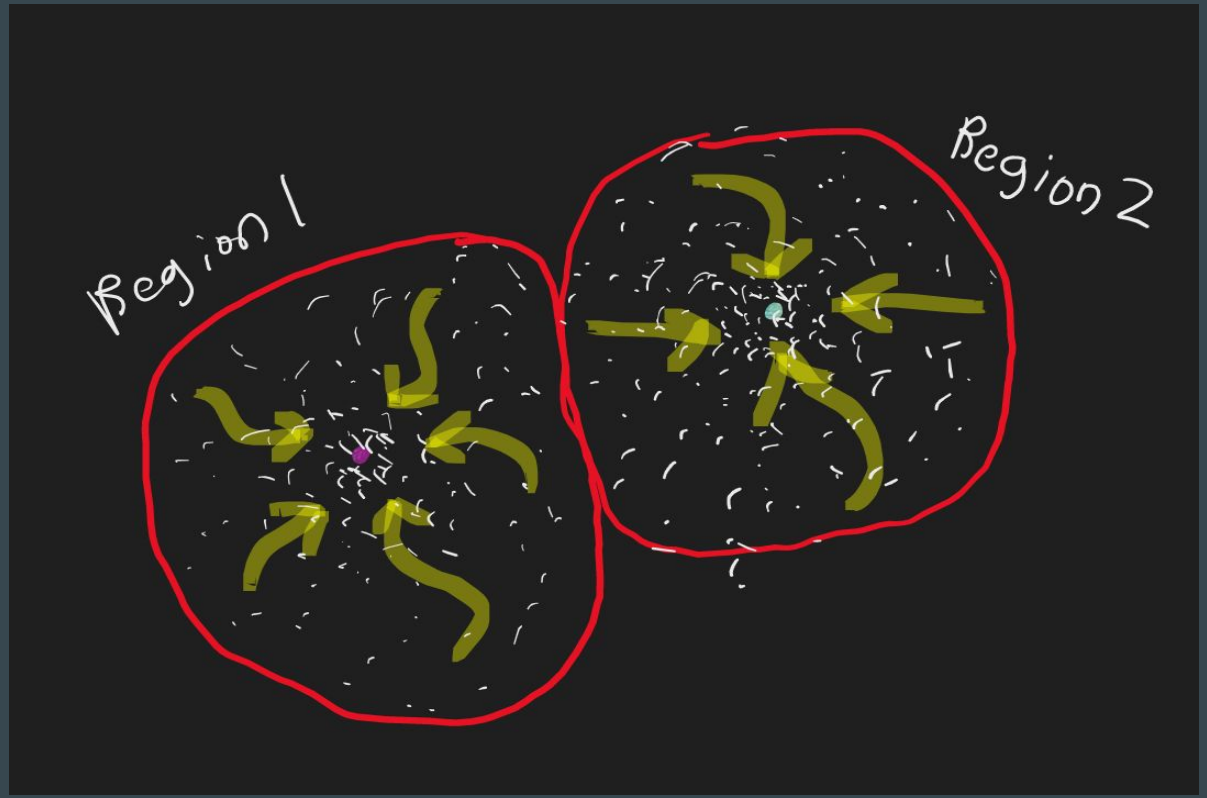
0 = High Sales (Blue)

1 = Low Sales (Orange)

2 = Medium Sales = (Green)

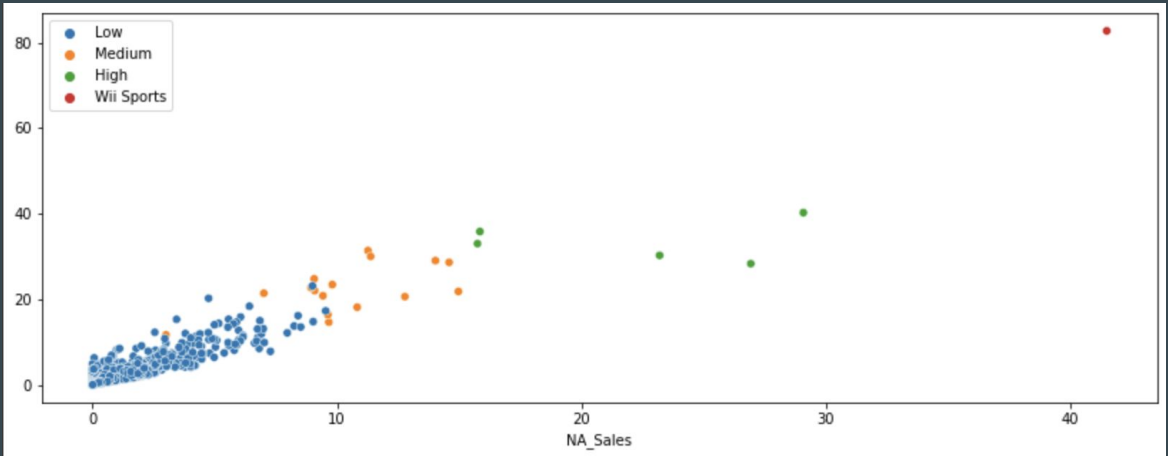


Data Training - MeanShift



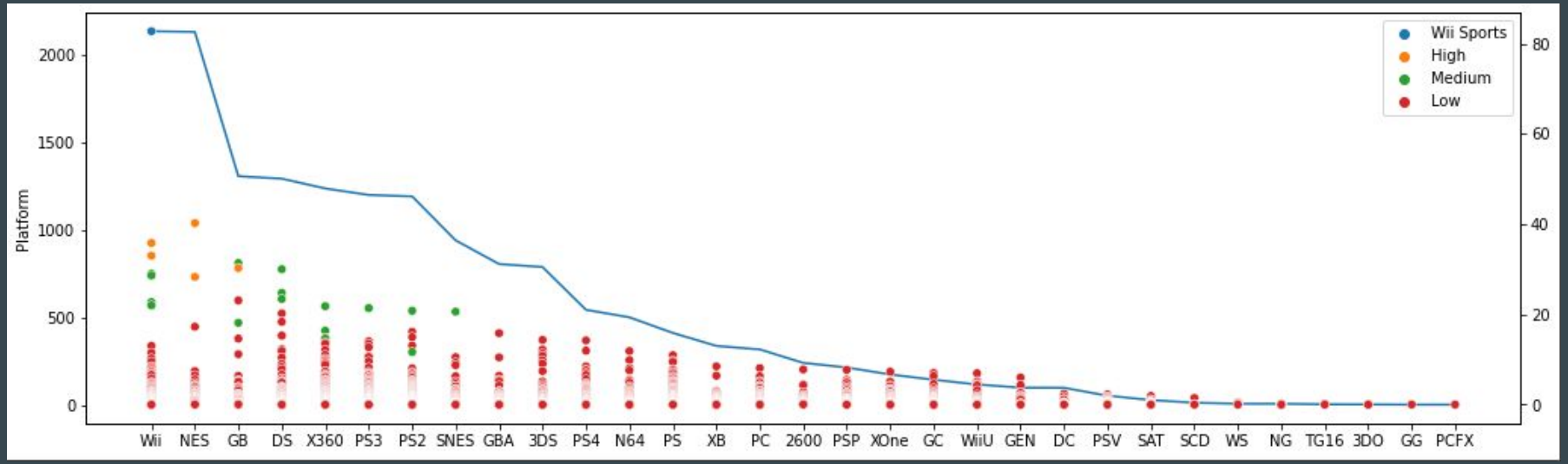
Data Training - MeanShift

Silhouette Score: **0.9567**

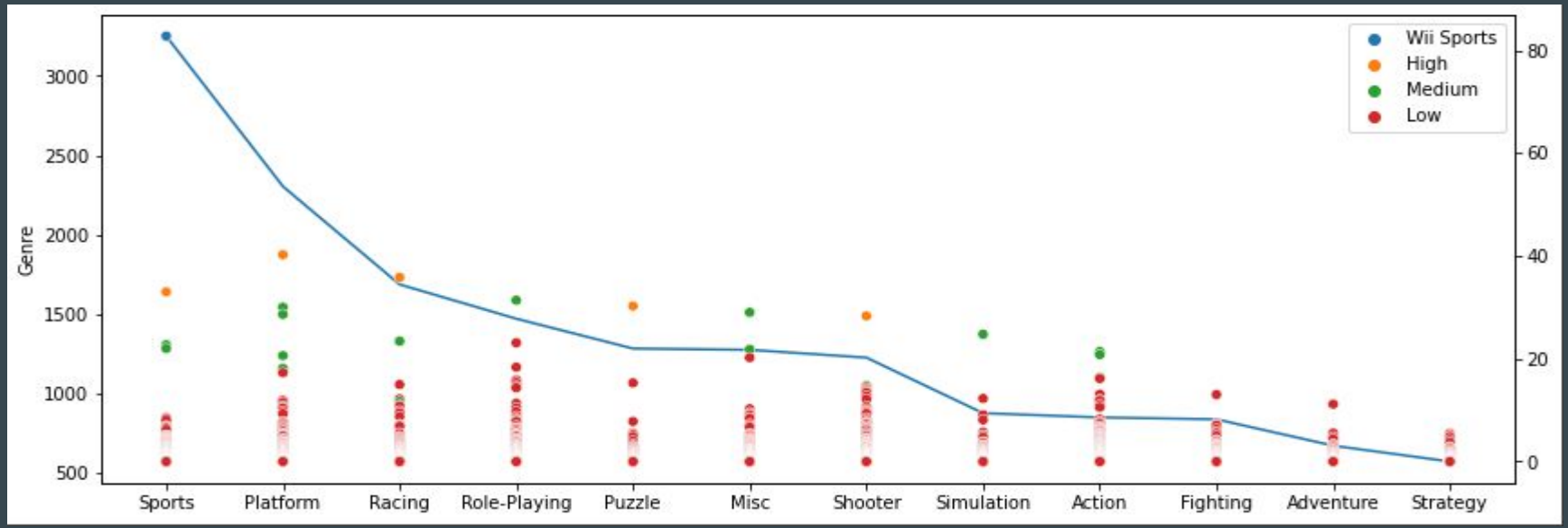


Cluster Analysis

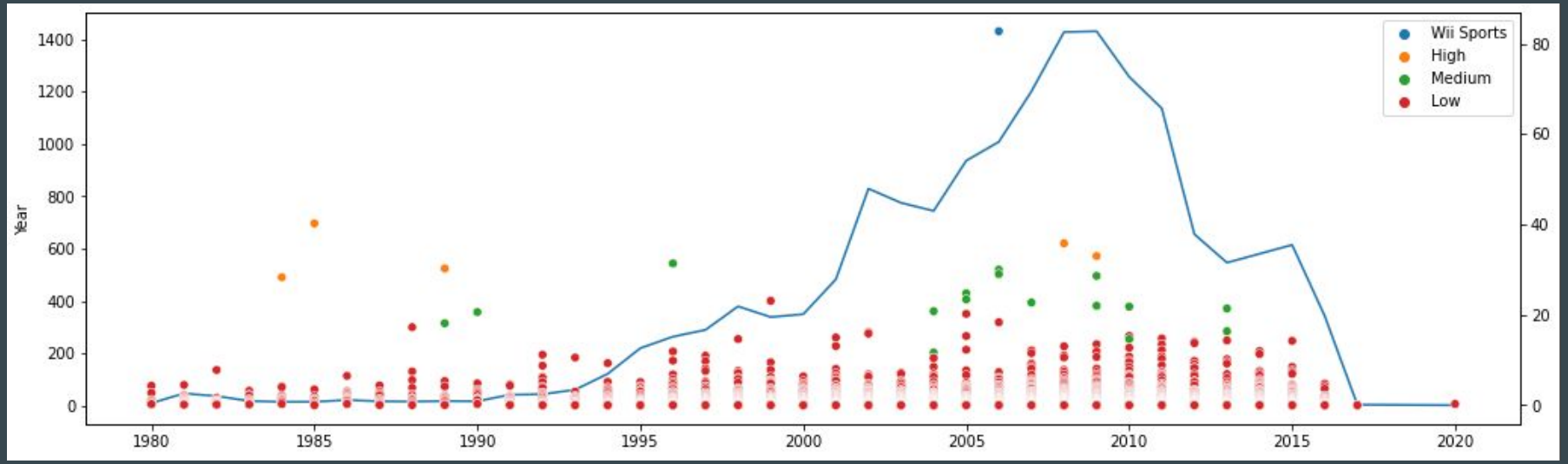
Cluster Analysis - By Platform Using MeanShift



Cluster Analysis - By Genre Using MeanShift



Cluster Analysis - By Year Using MeanShift



Cluster Analysis - Conclusions

1. Platforms with a greater number of successful games are more likely to have games that fall under higher sales clusters.

2. A similar trend holds less strongly for Genres

3. As time goes on, game sales seem to increase, with more titles naturally appearing in higher sales clusters.

Next Steps

Turning our Clustering into Classification

The next steps we would want to take...

- Use clusters as classification labels
- Model can accept the addition of new data
- Be able to classify new data into each cluster based on:
  - Sales
  - Genre
  - Year
  - Publisher

Thank you!

Questions?