PROJECT REPORT

INDEX

Serial No.	Content	Page No.
1.	Summary	3
2.	Project Motivation	4
3.	Data Description	5
4.	Data Pre-processing	6
5.	Data Visualizations	7
6.	Prediction Models and Findings	11
7.	Managerial Implications	14
8.	Conclusion	15
9.	References	17

SUMMARY

StockX is an online platform where people who buy limited edition shoes sell their collections to the highest bidder. The shoe listed in StockX dataset belongs to the brands Nike.

One of the main challenges that our client Nike was facing when it comes to the consumer market was that in the case of these exclusive categories of shoes, they could not figure out which shoes to group together to make decisions for a cluster of products, which would help solve their business problem.

Their initial approach to this problem was conducting traditional marker surveys, using focus groups, and conducting questionnaires. This turned out to be very resource intensive. So, as professional data scientists, we told them we'd give them the same answers by analyzing the dataset by the application of various tools and techniques required- saving them a lot of their resources and efforts that would have otherwise been spent on conducting a full-fledged market survey.

The factors influencing the sales of these shoes included the shoe size, order date, release date, sale price, retail price, and the state in which the sale has been made. This analysis would be considered unsupervised learning. We used the methods of K-means clustering and hierarchical clustering that were implemented using Python. Using the elbow method, we obtained the K value to be 4 for K-means clustering and 3 for hierarchical clustering. But it made more sense for us to choose 4 for the K value because more clusters would help give Nike more options to help choose from for their business purposes.

PROJECT MOTIVATION

StockX, even though an e-commerce site, is quite different from the usual ones. It is a real-time marketplace for limited-edition sneakers, watches, handbags, and streetwear. Bids are placed by buyers, and Asks are placed by sellers, and when a Bid and an Ask meet, the transaction is completed immediately - at a true market price.

Buyers and Sellers can use the StockX historical data to see how much an item on the market has sold for in the past and what the current lowest Ask is. The users may browse for the things they want wiser than ever before because of this transparency.

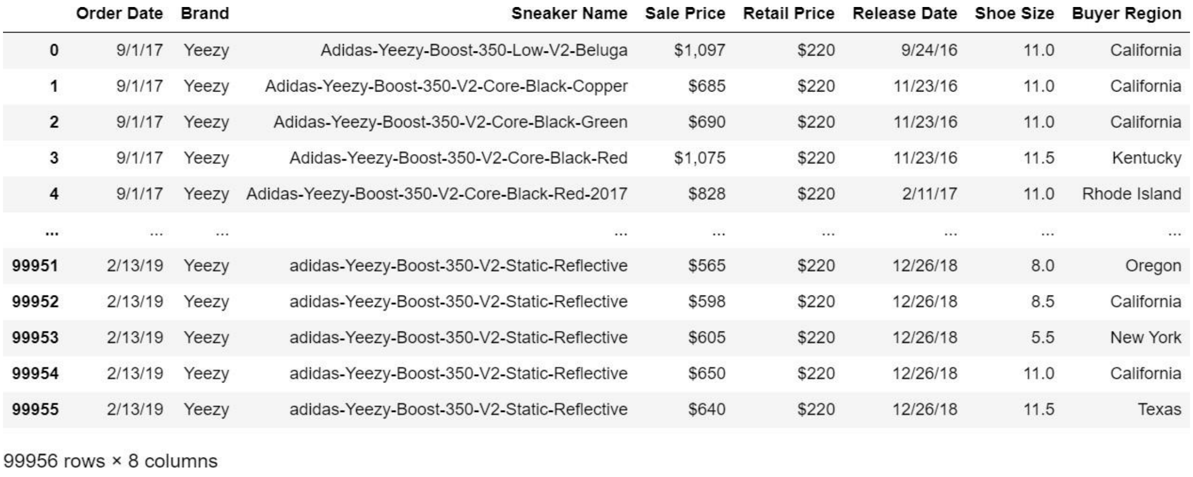
Companies rarely focus on a single product while using them for their business purposes like advertising campaigns or TV ads. Nike has thirty-five limited edition shoes and they wanted to know which shoes to group together, to make it easier to make any business decisions.

They conducted traditional marker surveys, using focus groups, took questionnaires, and surveys, and also observed buyer behavior. But the disadvantage of each of these methods of data collection and analysis is that they can extremely get resource-intensive and time-intensive. We as data scientists working for Nike suggested that we would get them the same insights they hoped to obtain from the traditional methods for a much cheaper cost and considering how we would use data to get these insights, it would be science and data backed.

2-D and 3-D views of these clusters give us information on how each shoe has similarities between them, including intangible factors such as the time it took for the shoe to be sold from the time that it was released. These insights gave us a clear idea of the similarities between the various shoe models, and which can be clustered together to perform ad campaigns and make better business decisions that would help the company.

DATA DESCRIPTION

The dataset in its raw form had 99,955 rows and 8 columns. The columns included Order Date, Brand, Sneaker Name, Sale Price, Retail Price, Release Date, Shoe Size, and Buyer Region. Some columns did not seem significant at the first glance. But the time took for the shoe to be told is very important to how the show can be grouped with other shoes. So we mixed Order Date and Release Date. Similarly, the difference between Sale Price and Retail Price would give us the profit, the shoe size, and buyer region were other significant columns that would affect the shoes would be grouped. A screenshot of the raw dataset can be seen below:

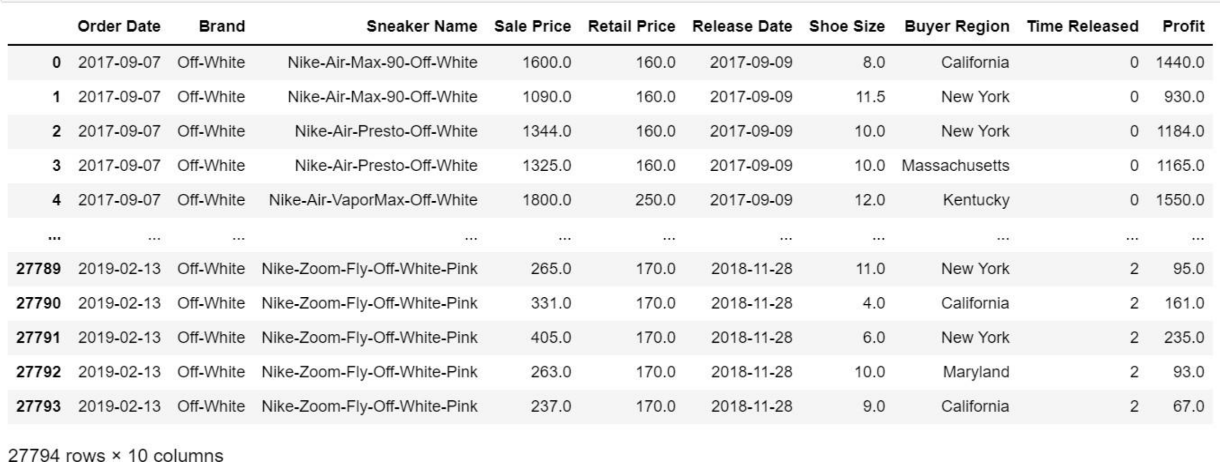


	Order Date	Brand	Sneaker Name	Sale Price	Retail Price	Release Date	Shoe Size	Buyer Region
0	9/1/17	Yeezy	Adidas-Yeezy-Boost-350-Low-V2-Beluga	$1,097	$220	9/24/16	11.0	California
1	9/1/17	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Copper	$685	$220	11/23/16	11.0	California
2	9/1/17	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Green	$690	$220	11/23/16	11.0	California
3	9/1/17	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Red	$1,075	$220	11/23/16	11.5	Kentucky
4	9/1/17	Yeezy	Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017	$828	$220	2/11/17	11.0	Rhode Island
...	...	...	...	...	...	...	...	...
99951	2/13/19	Yeezy	adidas-Yeezy-Boost-350-V2-Static-Reflective	$565	$220	12/26/18	8.0	Oregon
99952	2/13/19	Yeezy	adidas-Yeezy-Boost-350-V2-Static-Reflective	$598	$220	12/26/18	8.5	California
99953	2/13/19	Yeezy	adidas-Yeezy-Boost-350-V2-Static-Reflective	$605	$220	12/26/18	5.5	New York
99954	2/13/19	Yeezy	adidas-Yeezy-Boost-350-V2-Static-Reflective	$650	$220	12/26/18	11.0	California
99955	2/13/19	Yeezy	adidas-Yeezy-Boost-350-V2-Static-Reflective	$640	$220	12/26/18	11.5	Texas

99956 rows × 8 columns

DATA PRE-PROCESSING

The entire preprocessing was done using the Pandas library in Python. The dataset initially had shoes of both Nike and Adidas brands. From there we filtered out only the shoes belonging to Nike. So from 99956 columns it then dropped to 27794 columns. There were two numeric columns Sales Price and Retail Price that were combined into a single column Profit by taking the difference between both. The Time Released column was also added by taking the difference between the columns Release Date and Order Date. Shown below is a screenshot of the same:

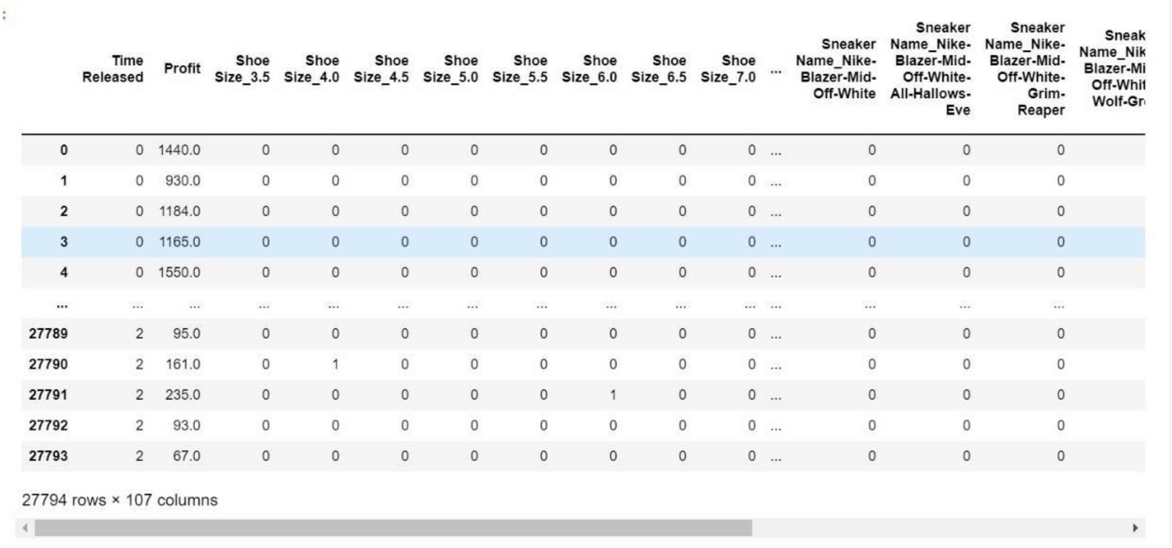


	Order Date	Brand	Sneaker Name	Sale Price	Retail Price	Release Date	Shoe Size	Buyer Region	Time Released	Profit
0	2017-09-07	Off-White	Nike-Air-Max-90-Off-White	1600.0	160.0	2017-09-09	8.0	California	0	1440.0
1	2017-09-07	Off-White	Nike-Air-Max-90-Off-White	1090.0	160.0	2017-09-09	11.5	New York	0	930.0
2	2017-09-07	Off-White	Nike-Air-Presto-Off-White	1344.0	160.0	2017-09-09	10.0	New York	0	1184.0
3	2017-09-07	Off-White	Nike-Air-Presto-Off-White	1325.0	160.0	2017-09-09	10.0	Massachusetts	0	1165.0
4	2017-09-07	Off-White	Nike-Air-VaporMax-Off-White	1800.0	250.0	2017-09-09	12.0	Kentucky	0	1550.0
...	...	...	...	...	...	...	...	...	...	...
27789	2019-02-13	Off-White	Nike-Zoom-Fly-Off-White-Pink	265.0	170.0	2018-11-28	11.0	New York	2	95.0
27790	2019-02-13	Off-White	Nike-Zoom-Fly-Off-White-Pink	331.0	170.0	2018-11-28	4.0	California	2	161.0
27791	2019-02-13	Off-White	Nike-Zoom-Fly-Off-White-Pink	405.0	170.0	2018-11-28	6.0	New York	2	235.0
27792	2019-02-13	Off-White	Nike-Zoom-Fly-Off-White-Pink	263.0	170.0	2018-11-28	10.0	Maryland	2	93.0
27793	2019-02-13	Off-White	Nike-Zoom-Fly-Off-White-Pink	237.0	170.0	2018-11-28	9.0	California	2	67.0

27794 rows × 10 columns

Following this, the Sale Price and the Retail Price column, and the Order Date, as well as the Release Date, were dropped. The final data frame hence consisted of only 8 columns, therefore.

When it comes to data enrichment, we had to convert our categorical columns into numeric ones by using the One Hot Encoding technique. Specifically, we encoded the Shoe Size and the Sneaker Name into binary variables from categorical ones by using this method. The encoded data that is ready to be fed to the model is shown as follows:



:

	Time Released	Profit	Shoe Size_3.5	Shoe Size_4.0	Shoe Size_4.5	Shoe Size_5.0	Shoe Size_5.5	Shoe Size_6.0	Shoe Size_6.5	Shoe Size_7.0	...	Sneaker Name_Nike-Blazer-Mid-Off-White	Sneaker Name_Nike-Blazer-Mid-Off-White-All-Hallows-Eve	Sneaker Name_Nike-Blazer-Mid-Off-White-Grim-Reaper	Sneak Name_Nik Blazer-Mi Off-Whit Wolf-Gr
0	0	1440.0	0	0	0	0	0	0	0	0	...	0	0	0	
1	0	930.0	0	0	0	0	0	0	0	0	...	0	0	0	
2	0	1184.0	0	0	0	0	0	0	0	0	...	0	0	0	
3	0	1165.0	0	0	0	0	0	0	0	0	...	0	0	0	
4	0	1550.0	0	0	0	0	0	0	0	0	...	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
27789	2	95.0	0	0	0	0	0	0	0	0	...	0	0	0	
27790	2	161.0	0	1	0	0	0	0	0	0	...	0	0	0	
27791	2	235.0	0	0	0	0	0	1	0	0	...	0	0	0	
27792	2	93.0	0	0	0	0	0	0	0	0	...	0	0	0	
27793	2	67.0	0	0	0	0	0	0	0	0	...	0	0	0	

27794 rows × 107 columns

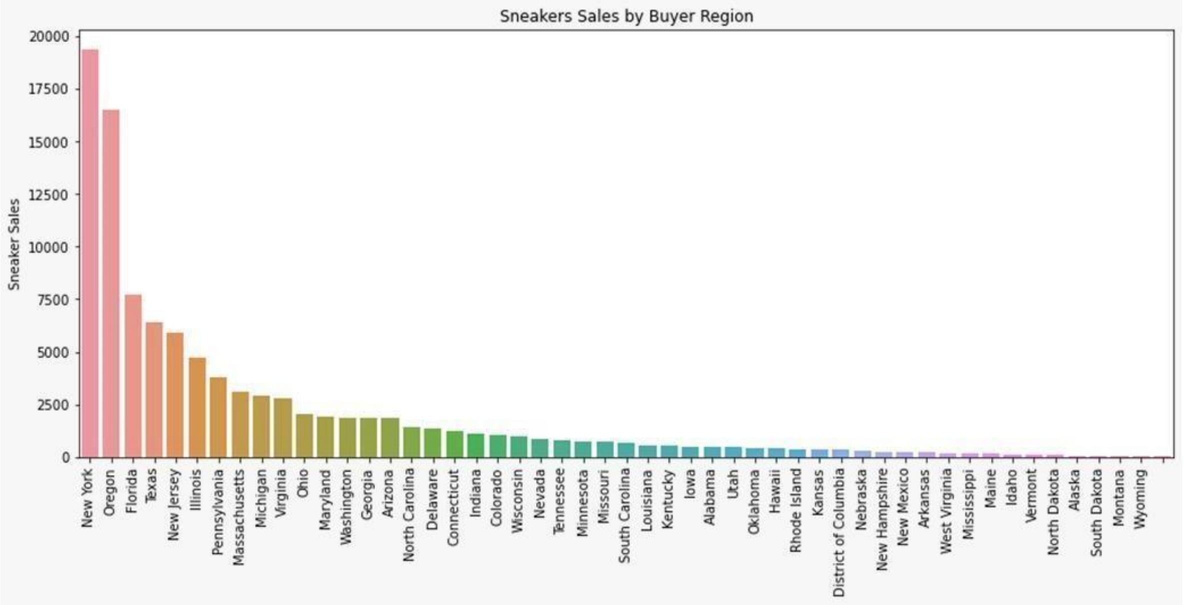
DATA VISUALIZATIONS

The visualization techniques have been used in the project to analyze the selling pattern of the shoe Nike. We have primarily done our preliminary exploratory data analysis to find out more about the data and its relevant properties. We used:

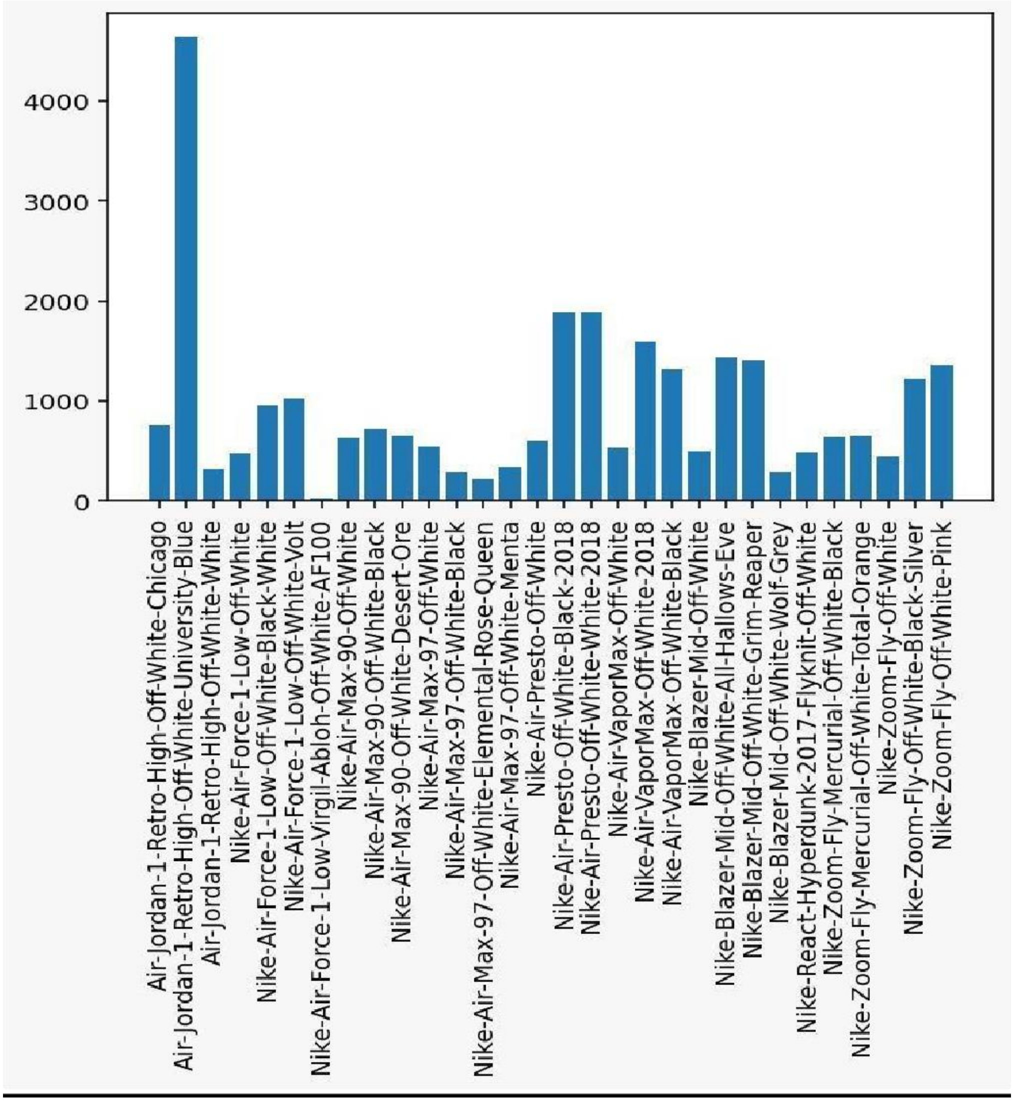
1. Bar graphs to study the:
   a. State-wise sales of the shoes - It gave us insight into the count of shoes sold in every state of the USA. The state-wise sales of the shoes were one of the important variables which had in our dataset.

   b. Model-wise sales of the shoes - It gave us insight into the sales of Nike shoes but based on the model number of all the shoes. This was another important variable used in the project.
2. Line graphs to know the mean price - Several visualizations indicated the average price of the Nike shoes sold on the StockX platform. The mean price of the sneakers was needed to bring the output of the business problem which we were solving.

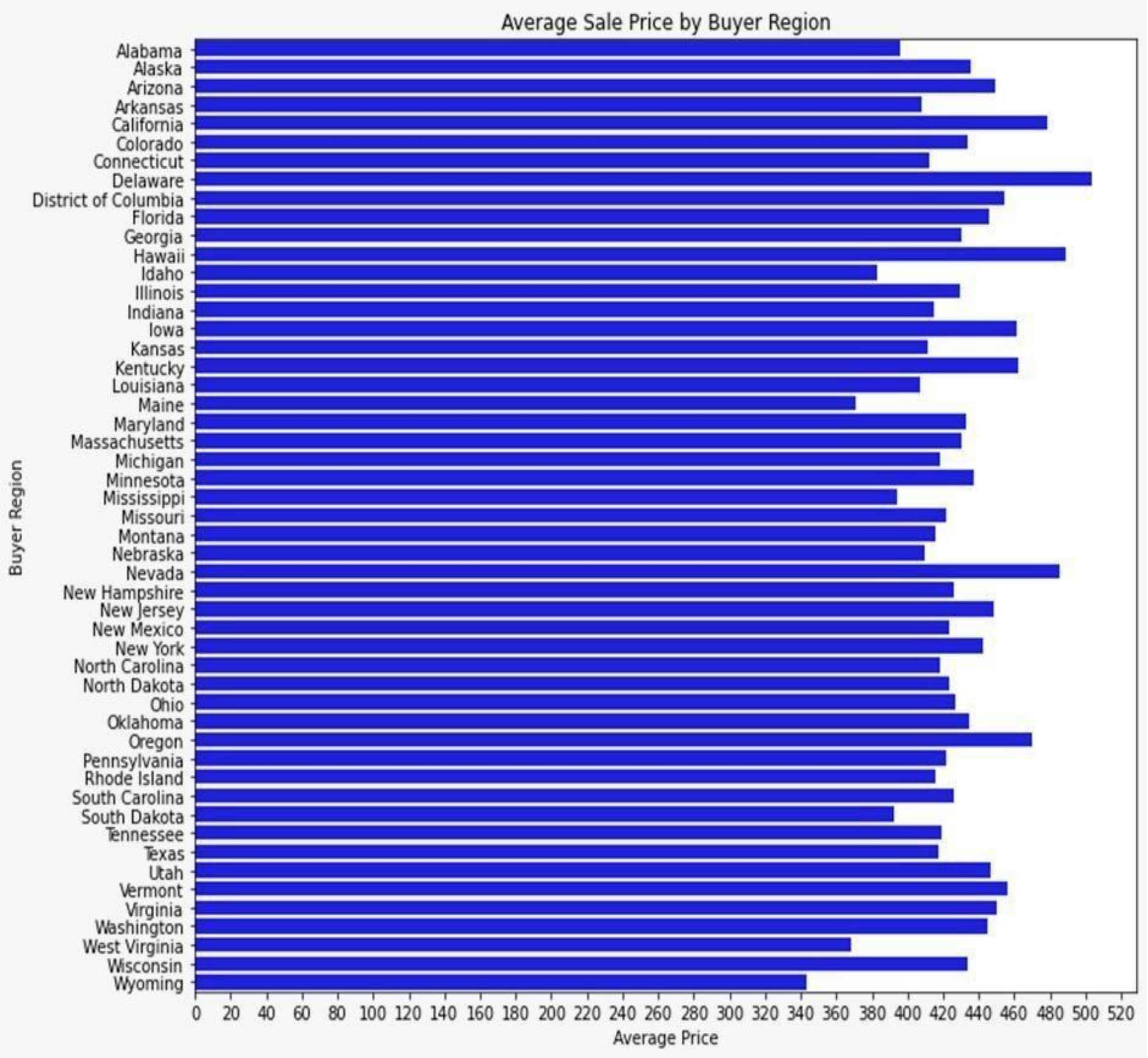
The below bar graph shows the sale price of Nike shoes across various states in the USA. The x-axis indicates the different states of the USA with sneakers sales on the y-axis. It can be noted that the shoes were the most sold in the state of New York and the second state with the highest sales in Oregon. Observing the least sales, Wyoming is the state which has the lowest sneaker sales, and in every other state, they were at least twice as low, when compared to the sales in the state, New York. On observing the graph closely, it shows declining sales as we move forward on the right side of the x-axis. The sneaker sales are high at the beginning, that is, around 20,000, and keep on declining until it reaches around 0 number of sneakers sold, which is the state of Wyoming.



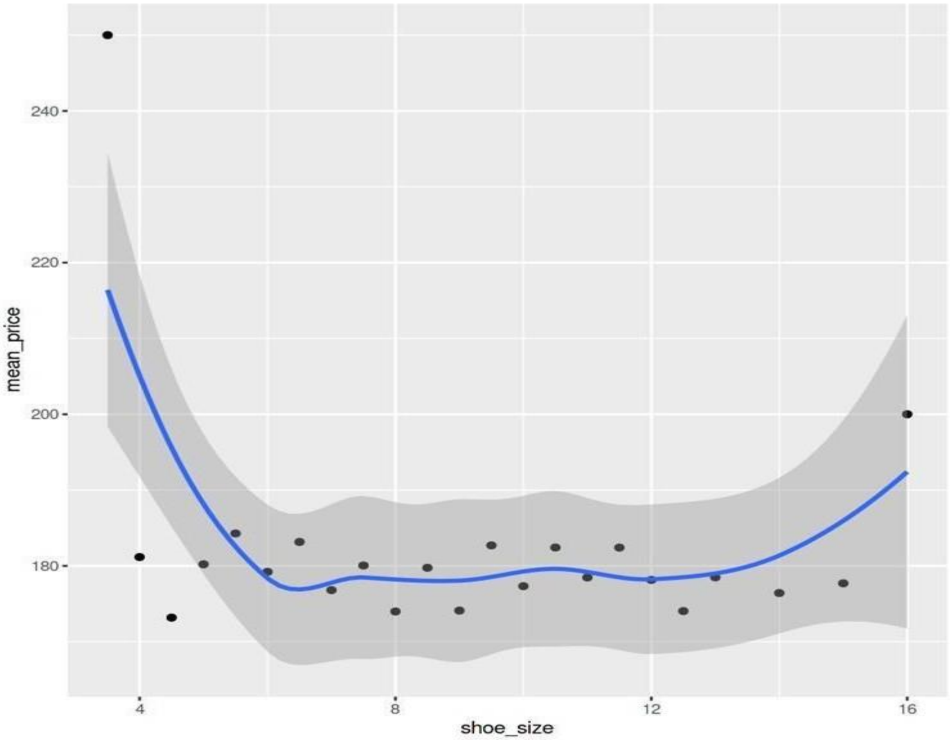
The bar chart below shows the sales of sneakers based on the shoe model. Here, we have a number of shoes sold on the y-axis and the different Nike shoe models on the x-axis. It is clear from the diagram that the Air Jordan Retro High model of the University Blue color sells the most in the market. The graph shows that the number of sneakers sold varies by the larger amount when comparing various shoe models. It can be observed that the lowest sneaker model sold is Nike Air Max 97 off-white Elemental Rose Queen. The highest sales of the sneaker model were above 4000, which is the Air Jordan Retro High model of the University Blue color and the lowest was below 1000. The majority of the shoe model sales were below 1000. There were only a few models which got their sales above 2000. The data gives detail about the color of the particular shoe model with their shoe model number. Also, few of the models include the launching year in their model’s name whereas few had the state name.



The below graph represents the average sale price of the Nike X Off-White shoes in various states of the USA. The graph has names of the state, that is, buyer region on its y-axis and average prices, that is, average price, on its x-axis. Noticeably, the highest average price of the sneaker shoes Nike X-off white was in Delaware, which was around 500. The average sales price by buyer region shows fluctuations in the given below graph. This means, based on the buyer region the sneaker shoe Nike X-off white has different average prices in all the given buyer regions. On comparing the average price of the sneaker, it is very apparent that the lowest average price is in the state of Wyoming, that is, around 340.



The line graph below represents the mean price of the Nike X Off-White shoes depending on the size of the shoe and it is the best fit line. The x-axis indicates shoe size and has the mean price on its y-axis. The line graphs make a curve type of shape. The dark gray shadow part shows us one standard deviation above and below the original line. Almost all the points fall within these lines.



PREDICTION MODELS AND FINDINGS

We have initially drafted and came up with the following two modeling techniques:

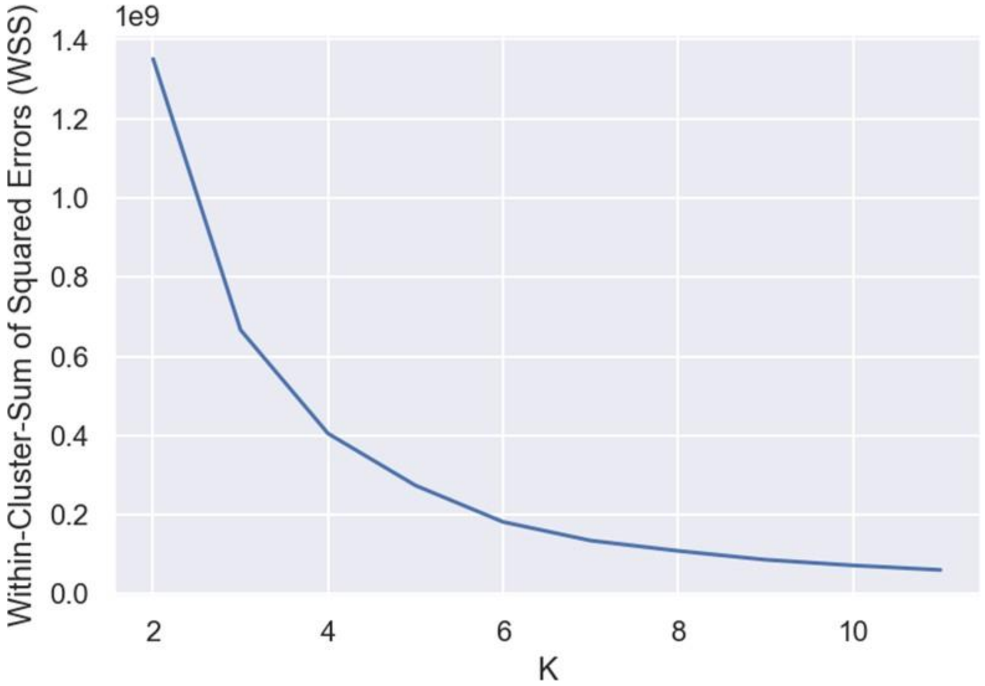
1. **K Means Clustering:** K Means Clustering would be well suited for this kind of prediction because it is a fairly accurate method when it comes to doing categorical predictions like the one that we are doing here.

2. **Hierarchical Clustering**: Hierarchical Clustering is also a similar type of unsupervised learning method that is particularly applicable in this type of dataset where we have categorical variables.

The below visualization shows the number of clusters with the help of the elbow method:

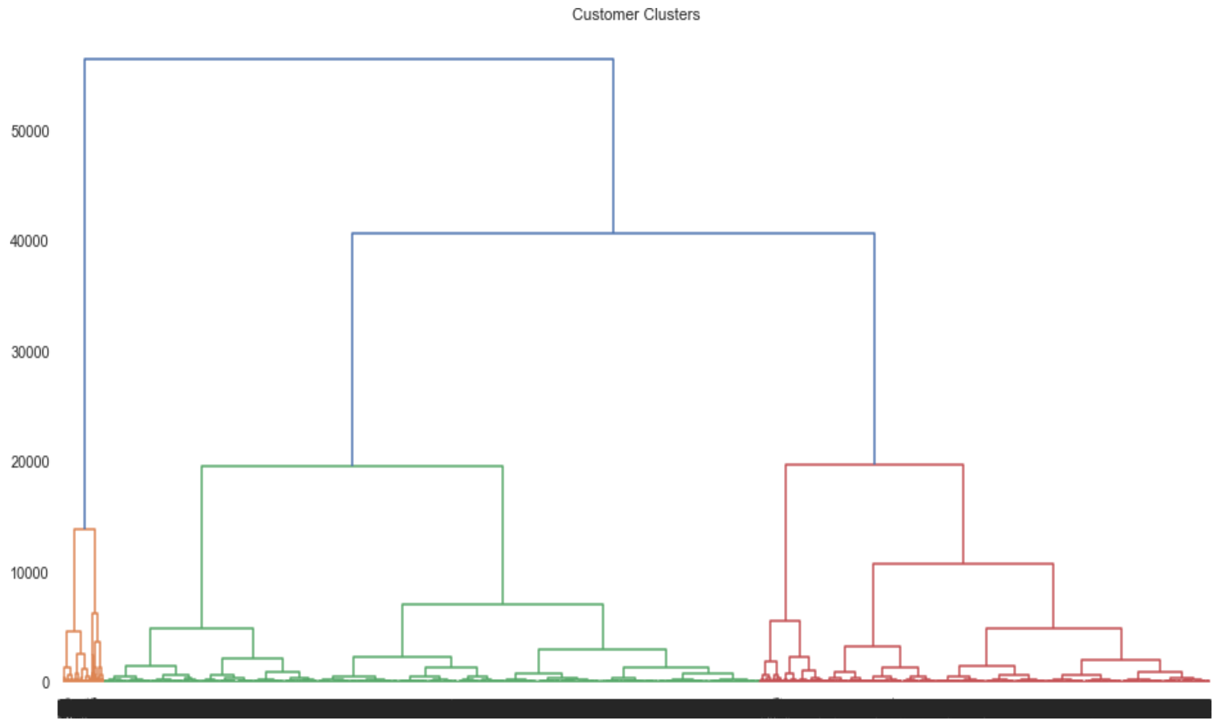
1. **K-means**: The number of clusters we got from the K-means using the elbow method is 4. The below graph shows the output. The x-axis indicates the number of clusters, that is, K whereas the y-axis indicates WSS, that is, Within-Cluster-Sum of Squared Errors.

   For checking, we have also used the silhouette score obtained from the different numbers of clusters and validated our number of clusters.



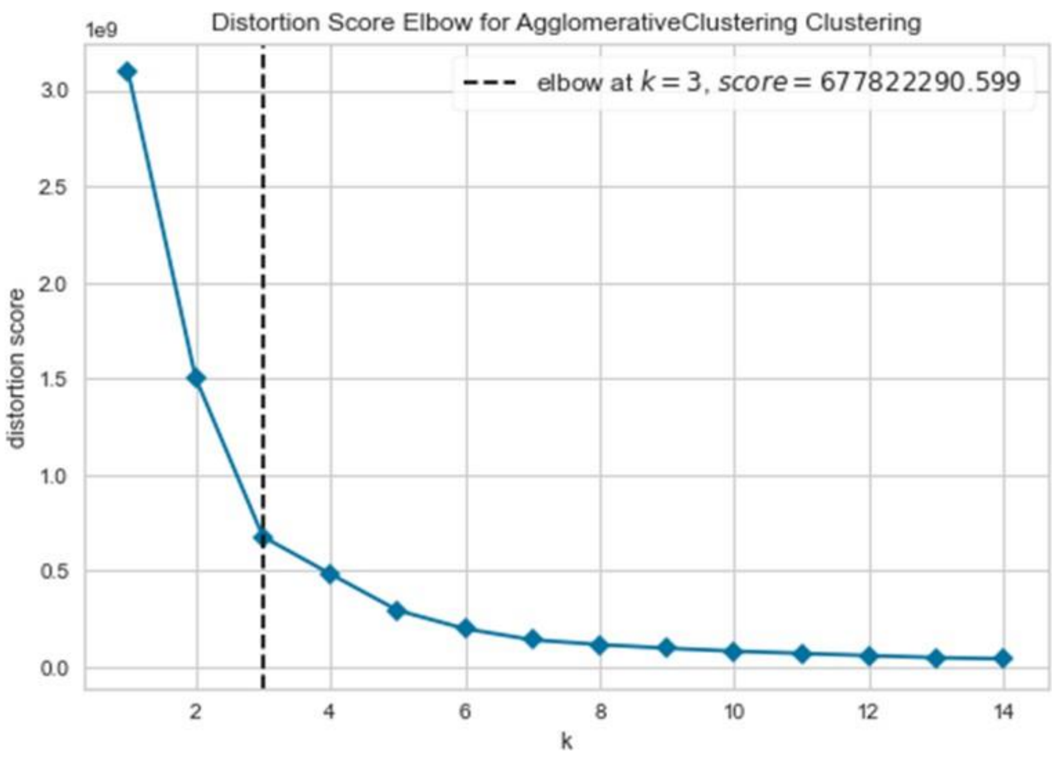
2. **Hierarchical clustering**:

The type of hierarchical clustering technique used here is the agglomerative type which means that it follows the bottom-up approach. The below diagram shows the dendrogram that we obtained from implementing the K Means clustering algorithm. We see in the below figure that the number of clusters formed is 3.



The above diagram shows the dendrogram that we obtained from implementing the K Means clustering algorithm. We see in the above figure that the number of clusters formed is 3.

For confirmation, we also performed the elbow or the knee method and got confirming results. The diagram for the same is attached below:



In the above graph, the x-axis shows the number of clusters (K) and the y-axis indicates the distortion score.

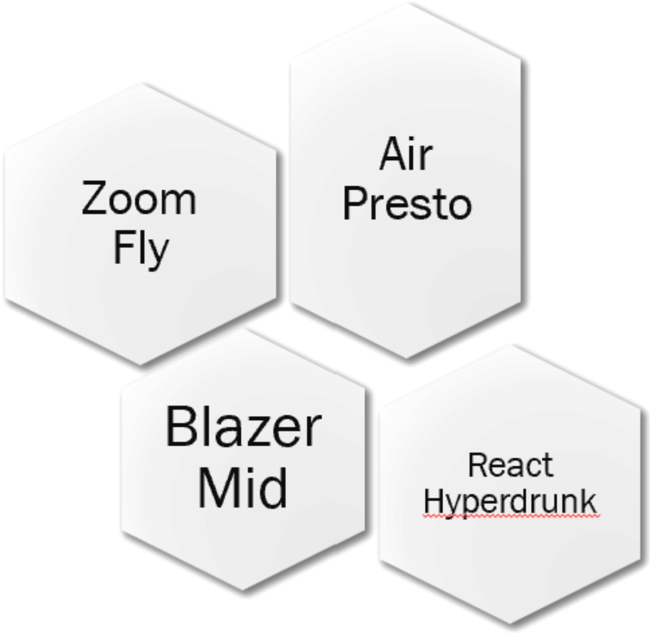
Even though we got 3 as the number of clusters but ultimately, we went ahead with 4 as the number of clusters as it made more sense to the type of the business case that we were dealing with.

MANAGERIAL IMPLICATIONS

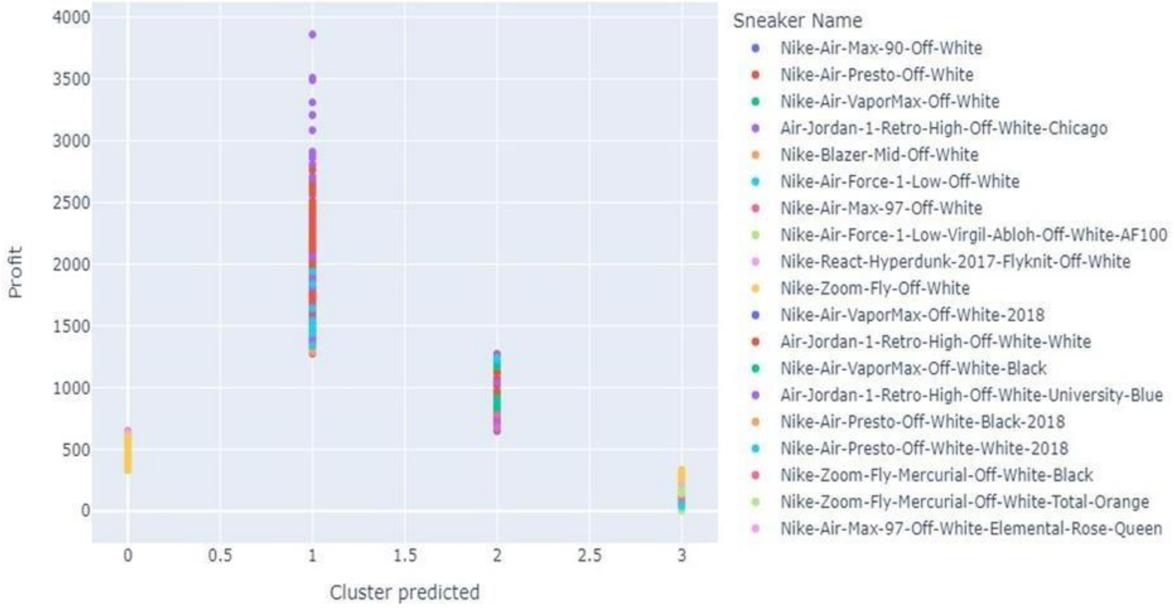
The picture below shows one of the clusters that we obtained after running the analysis. This is specifically the shoes belonging to the zeroth numbered cluster. We see that the cluster comprises the shoe models named Zoom Fly, Air Presto, Blazer Mid, and React Hyperdunk. So tomorrow if our client Nike wants to launch a new model of shoe and wants us to give us a selection of shoes of their own brand which they can use as a reference, they can use the following models.

Nike can also use the shoes that are predicted to be in a particular cluster as a group that they can promote or market together during their various sales or marketing campaigns. This is because of the fact that they have been placed into groups by virtue of having similar properties that have been derived from the column attributes.

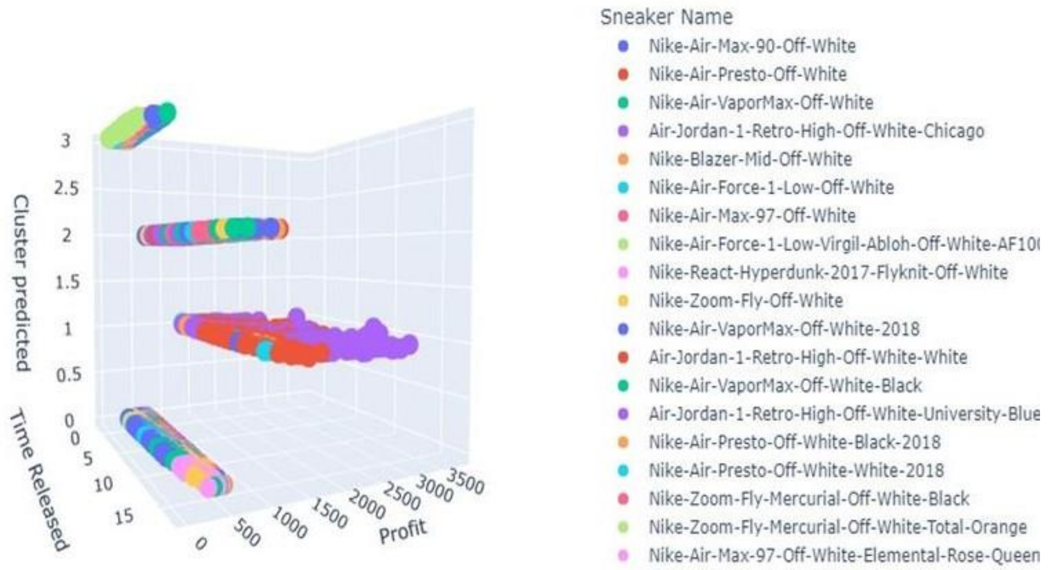
Our client can also use the parameter of profit margin to understand which group of shoes to select. For example, if they want their profit margins to be between 250 to 700 USD, they can select cluster number 0, if they want their profit margins to be between 600 to 1300 USD they can go for cluster number, cluster number 3 for any profit margin below 700 USD and cluster number 1 for any profits above 1300USD.



CONCLUSION



The above graph describes the shoe model in each cluster and the profit obtained for each cluster. We can notice that there are four clusters, and the profit value of shoes gives us the difference between the retail price and sale price. This gives us an insight into how we can cluster the shoes together, and the profit that would be obtained from each shoe in each cluster.



The above graph is the 3-D representation of the previously mentioned clusters. This graph also includes another dimension, that is, time released. This gives us an idea about how the clusters are similar with reference to the time released.

To summarize, the most frequently occurring shoes associated with different clusters are listed as follows:

Cluster 0: Zoom Fly, Air Presto, Blazer Mid, React Hyperdunk, AirMax 97 Off White Elemental Rose Queen.

Cluster 1: Air Jordan 1 Retro High Off-White White, Air Presto Off White, Air VaporMax Off White Black.

Cluster 2: Air Max 90, Zoom Fly Off White, Air VaporMax Off White 2018, Nike Air Presto Off-White 2018.

Cluster 3: Air Force 1 Low Virgil Abloh Off White AF101, Air Max 90 Off White, Air VaporMax Off White.

Hence Nike can use these clusters of shoes while carrying out their various business endeavors like promoting together with a select group of shoes or using them as a baseline for the creation of a new shoe model, additionally having the estimated profit margins as a reference.

REFERENCES

1. https://www.kaggle.com/datasets/hudsonstuck/stockx-data-contest
2. https://stockx.com/news/the-2019-data-contest/
3. https://hands-on.cloud/implementation-of-k-means-clustering-algorithm-using-python/
4. https://nickmccullum.com/python-machine-learning/k-means-clustering-python/
5. https://medium.com/@sametgirgin/hierarchical-clustering-model-in-5-steps-with-python-6c45087d4318
6. https://plotly.com/python/line-and-scatter/
7. https://seaborn.pydata.org/