LOYALiST COLLEGE

Arbit Research Project – Modelling Sneaker Resale Industry Data

AIGS1006 – Deep Learning

<u>Group 2</u>

Rohit Lovers, Abhishek Rajendra Mahale, Brenden Clarke, Ilan Goldfarb, Olabode Aremu, Vijay Bhaskar Bonthu

TABLE OF CONTENT

Pages

1 INTRODUCTION ……………………..……………………………… 3

2 LITERATURE REVIEW ………………….……..……………………. 3

3 METHODS …………………………………..………………………. 4

- 3.1 – Segmentation & Regression
- 3.2 – Clustering
- 3.3 – LSTM

4 RESULTS & ANALYSES ………………………………..…………… 5-14

- 4.1 – Segmentation & Regression
- 4.2 – Clustering
- 4.3 – LSTM

6 CONCLUSIONS ……………………………………………………… 14

7 CITATIONS ………………………………………………………..…… 15

8 APPENDIX ..…..…………..……………………………………..…… 15

1 – INTRODUCTION:

The entire sneaker resale market is expected to reach $11.5 billion by the end of 2023, an increase of 8.5% compared to last year – this is equivalent to approximately 15% of total worldwide sneaker sales. In the United States alone, the sneaker resale market is projected to generate $2 billion worth of sales by the end of 2023 and is forecasted to reach $30 billion in 2030, indicating a 47.2% compound annual growth rate (CAGR) [1,2].

In the realm of business and sales, the application of machine learning (ML) and deep learning (DL) has become increasingly pivotal. These technologies offer sophisticated means to sift through vast volumes of data, providing invaluable insights into consumer behavior, market trends, and revenue patterns. Leveraging ML/DL techniques empowers businesses to decipher intricate patterns, forecast demand, optimize pricing strategies, and identify potential growth opportunities.

This report delves into the dynamic domain of the sneaker resale market. By employing multivariate-type methods (Clustering & Outlier detection) and time-series forecasting (LSTM), Group 2 aims to provide relevant business insights and preliminary model development for our project partner – Arbit.

2 – LITERATURE REVIEW:

In the broader retail context, clustering models have been extensively used for customer segmentation, product categorization, and market basket analysis. Smith et al. (2017) demonstrated the effectiveness of K-means clustering in identifying customer segments based on purchasing behaviour in online retail. While the focus was on general e-commerce, the principles and methodologies are applicable to the sneaker resale market [3].

Machine learning methods trained on raw numerical time series data often face challenges such as heightened sensitivity to hyperparameters and the initial random weights. In their research at Cornell University, Steven Elsworth and Stefan Güttel proposed an innovative approach combining recurrent neural networks with a dimension-reducing symbolic representation for time series forecasting. Their findings demonstrate that this symbolic representation not only mitigates certain issues but also potentially expedites training while maintaining forecast accuracy. This study addresses the drawbacks of feeding raw data into LSTM models, leading to suboptimal model performance. To address this, Elsworth and Güttel employed the ABBA symbolic representation, converting features into alphabetic clusters. However, this conversion and clustering process significantly prolonged dataset creation for LSTM utilization [4]. To accommodate the time and reduce the complexity of the project we have taken an approach to create the data out of lagged difference mentioned in the article, it suggests using previous input/output values to predict the future values and the parallel architecture to process the data faster.

3 – METHODS:

3.1 – Segmentation & Regression:

The regression analysis aimed to ascertain the relationship between the count and sold price, employing linear regression with the slope-intercept formula to quantify the association between the number of units sold and their resulting prices.

3.2 – Clustering:

The first step in this modelling approach was data filtering and feature selection. Null values were removed from a subset of 10K rows using code in AWS services, and the data was downloaded to be processed in the individual users Jupyter Notebook. The data was loaded using a Pandas DataFrame, Matplotlib, Scikit-learn, and Seaborn libraries were imported for statistical analyses and data visualization.

3.3 – LSTM:

Input Feature Generation: Employed sliding window methodology to extract preceding price data (first 5 or 6 prices) for prediction of the subsequent 6th or 7th price.

Model Architecture: Constructed an LSTM model with 128 nodes, using the number of variables in the input list and the count of elements within each index position (e.g., 7, 1 or 6, 5) as the model's input structure.

Normalization Techniques: Applied Standard Scalar to normalize data, ensuring a mean of 0 and standard deviation of 1, reducing outlier influence on models reliant on gradient descent.

Data Split: Segregated the dataset into training (70%), validation (10%), and test (20%) sets. Emphasized that the analysis was limited to a specific sneaker with a unique SKU.

Model Configuration: Leveraged LSTM's memory capabilities by setting tanh as the activation function, aiding in data preservation within a range of -1 to 1. Employed a dropout rate of 0.2% and included an additional layer with 8 hidden units using the RELU activation function. The output layer was set to linear.

Model Training: Compiled the model using Mean Squared Error as the loss function, Adam optimizer with a learning rate of 0.001, and Root Mean Squared Error as the evaluation metric.

Data Transformation: Utilized inverse scalar transformation to revert predictions back to their original dataset values.

4 – RESULTS & ANALYSIS:

This report delves into the analysis of a dataset encompassing sneaker sales,

To start, we performed basic statistical analyses focusing on key metrics such as the top selling sneakers, highest priced sales, and the correlation between retail price and sales price. By documenting the basic relationships in the dataset, we will be better informed for more complex analysis.

4.1 – Segmentation & Regression:

Table 1. Top 5 Total Counts

Sneaker Type	Average Price	Total Count	Highest Value Sold
Dunk SB	$356.47	363,517	$65,000
Yeezy Boost 350	$315.42	440,566	$2,875
Air Jordan 4	$336.89	555,868	$35,701
Dunk	$181.65	1,494,555	$15,001
Air Jordan 1	$225.17	2,738,068	$35,000

Table 2. Top 10 Maximum Sold Prices (with respective Model, Size, Brand, Color)

Model	Size	Brand	Color	Price sold
Nike MAG Back to the Future (2016)	11	Nike	green	$108,500
Nike MAG Back to the Future (2016)	11	Nike	green	$108,000
Nike MAG Back to the Future (2016)	11	Nike	green	$105,000
Nike MAG Back to the Future (2016)	13	Nike	green	$76,925
Nike MAG Back to the Future (2016)	9	Nike	green	$65,000
Nike SB Dunk Low Paris	8.5	Nike	gray	$65,000
Nike SB Dunk Low Paris	7	Nike	gray	$62,000
Nike MAG Back to the Future (2016)	13	Nike	green	$56,500
Nike Dunk High Pro SB FLOM	9	Nike	white	$56,139
Nike SB Dunk Low Paris	11	Nike	gray	$51,950

Insights from initial investigation reveal compelling trends: the dominance of iconic models like the Dunk and Air Jordan 1, boasting both high counts and substantial average prices. Following closely are the Dunk SB and Yeezy Boost 350, demonstrating impressive average prices and noteworthy highest values. Despite its slightly lower count, the Air Jordan 4 yields a significant average price, solidifying its standing in the sneaker hierarchy. Understanding these dynamics is crucial for building more sophisticated ML/DL models, as they signify not just popularity but also the value distribution across these models within this dynamic market. Moreover, uncovering the top 10 maximum sold prices sneakers offers valuable insights into exceptionally sought-after and high-value items, guiding the development of models to predict such rare and high-value transactions.

Having determined some basic relationships, regression was employed to examine the relationship between average sold price and count of sneakers sold, for the top 25 most sold silhouettes.

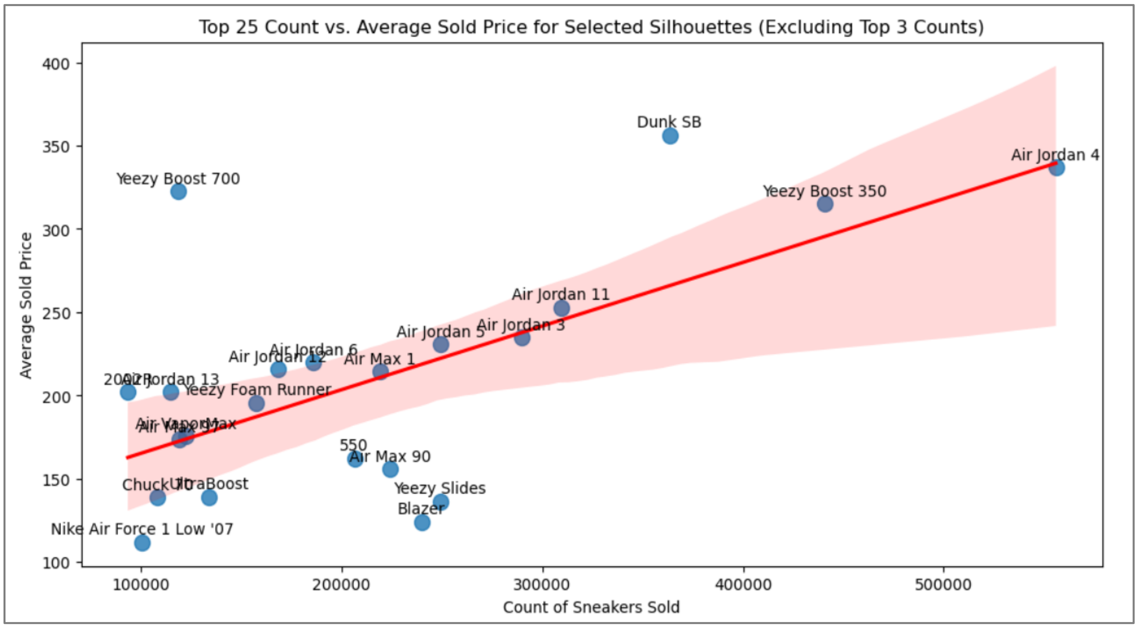


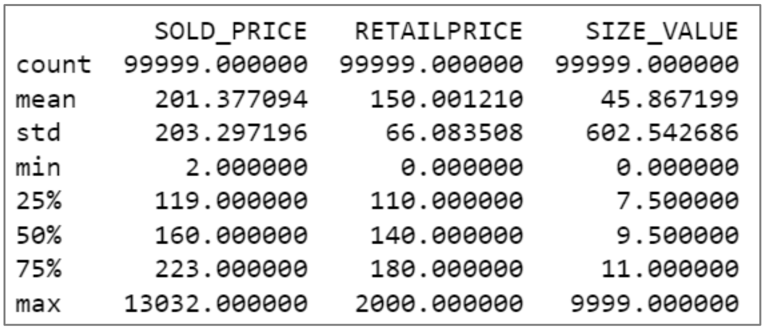
Figure 1. Depicting the positive correlation between count and sold price within the top 25 most sold sneaker silhouettes.

Equation 1: $Price = \$162.83 + 164.62x\ , where\ x = \frac{Count}{100{,}000}$

The interpretation of the regression states that for every increase of 100,000 units in the count of a top 25 most sold silhouette, there is an associated increase of $164.62 in the price. This indicates a positive correlation between count and sold price, suggesting that in the case of the top 25 most sold sneaker, the popularity of a sneaker drives the price.

4.2 – Clustering:

The descriptive statistics that summarize the central tendency, dispersion, and shape of the numerical columns (SOLD_PRICE, RETAILPRICE, SIZE_VALUE) were investigated, with the following output:



```
         SOLD_PRICE    RETAILPRICE    SIZE_VALUE
count  99999.000000   99999.000000  99999.000000
mean     201.377094     150.001210     45.867199
std      203.297196      66.083508    602.542686
min        2.000000       0.000000      0.000000
25%      119.000000     110.000000      7.500000
50%      160.000000     140.000000      9.500000
75%      223.000000     180.000000     11.000000
max    13032.000000    2000.000000   9999.000000
```

Figure 2. Descriptive statistical outputs from numerical columns of interest (SOLD_PRICE, RETAILPRICE, SIZE_VALUE).

From this Figure, we can deduce that on average, the sold price is 125% the retail price, suggesting that the overall average profitability of sneaker resale is approximately 25%. However, if we examine the 25-75% percentiles of the data, we observed an average profitability of approximately 8%, 12%, and 20%, respectively. This trend makes sense, as exclusive sneakers are likely to fetch a higher retail price and therefore a higher resell price. The mean size data is skewed because of differing size scales implemented in the dataset, however, the 75% percentile of data showing size 11 is most likely reflective of the dataset, because men were the greatest contributors to sneaker resale:

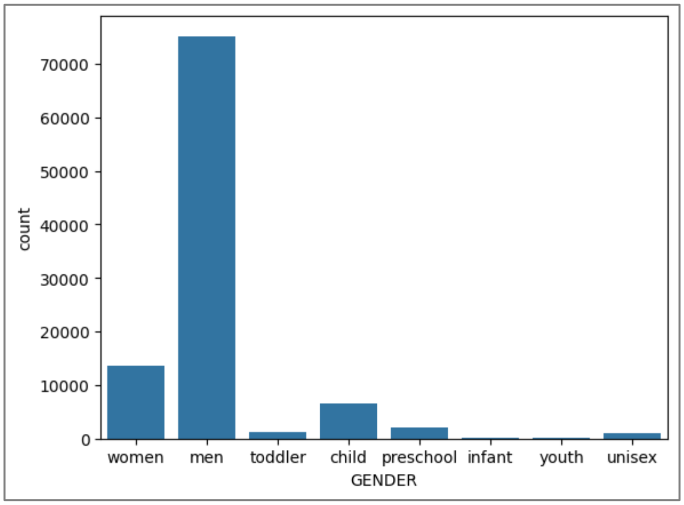


Figure 3. Gender contribution to total dataset counts.

The following figures display the influence of Brand, Collaborators, and Color on sneaker resale:

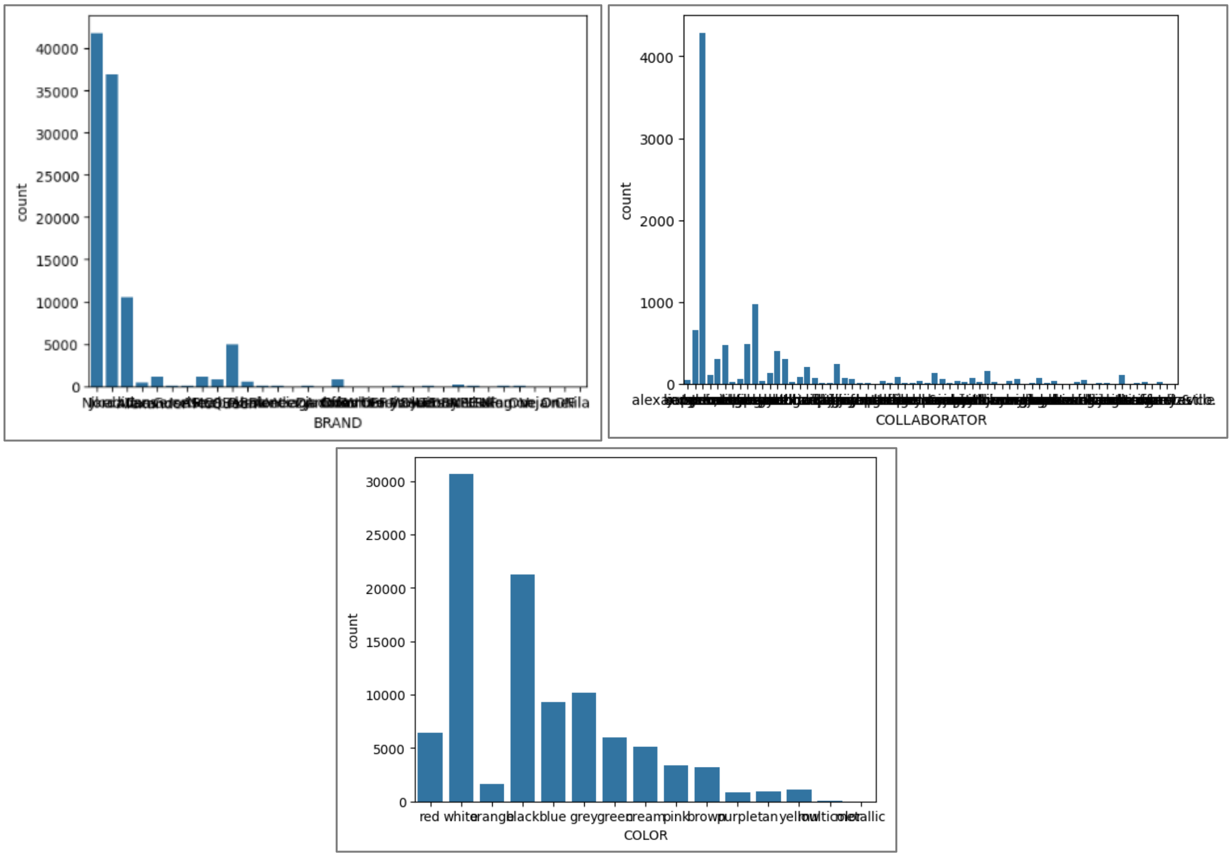
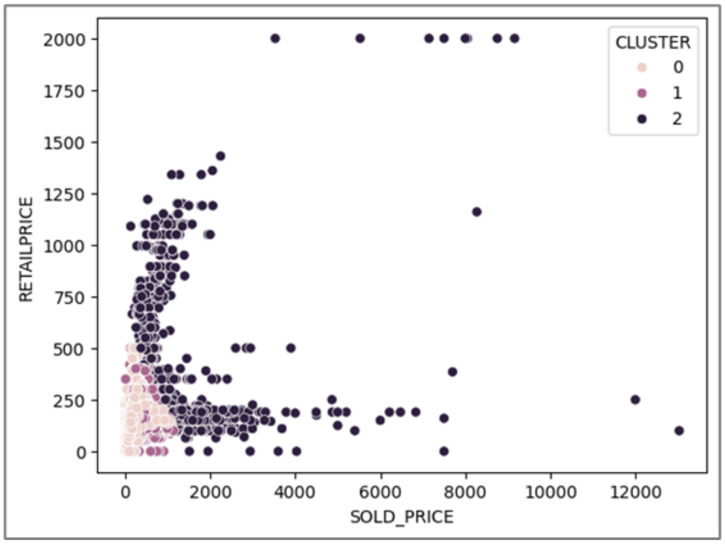
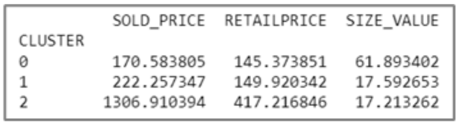


Figure 4. Brand, Collaborator, and Color contributions to total dataset counts.

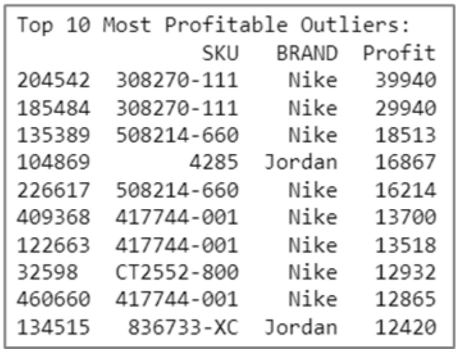
Figure 4 reveals that Nike, Jordan, and Adidas are the top brands, that the presence of collaborators is not as important as the specific collaborator (Kanye West & Supreme are greatest), and that black and white sneakers sell 2-3x more than any other color. These visualizations are crucial for initial understanding of the data's characteristics, identifying potential patterns, and informing further analysis and modeling decisions. These trends will be taken into consideration when implementing more predictive models.

The following outlines results of some K-Means clustering on scaled features, evaluating the quality of clusters, visualizing the clusters, interpreting the cluster characteristics, and ultimately allowing us to conduct further analysis and derive insights based on domain specific knowledge.





```
         SOLD_PRICE  RETAILPRICE  SIZE_VALUE
CLUSTER
0        170.583805   145.373851   61.893402
1        222.257347   149.920342   17.592653
2       1306.910394   417.216846   17.213262
```

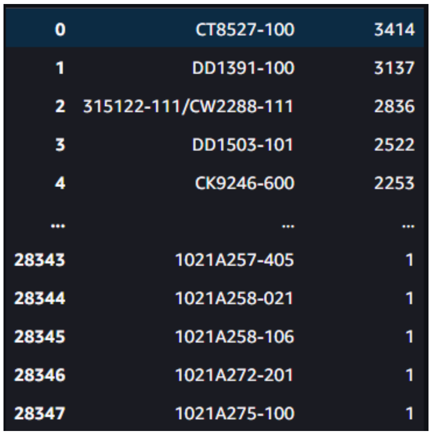


```
Top 10 Most Profitable Outliers:
               SKU   BRAND  Profit
204542  308270-111    Nike   39940
185484  308270-111    Nike   29940
135389  508214-660    Nike   18513
104869        4285  Jordan   16867
226617  508214-660    Nike   16214
409368  417744-001    Nike   13700
122663  417744-001    Nike   13518
32598   CT2552-800    Nike   12932
460660  417744-001    Nike   12865
134515   836733-XC  Jordan   12420
```

Figure 5. K-means clustering examining the relationship between interest (SOLD_PRICE & RETAILPRICE)

The K-means clustering summarizes the trends between retail and resale price, visualization of this data clearly outlines the presence of outliers, allowing for specific insight into the most profitable sneakers. Clustering by brand was attempted, however, due to the number of brands, this produced a poor visualization – the table of Top 10 most profitable outliers was generated instead. This table reveals that Nike SKU's 308270-111 and 417744-001 should be of particular interest, as they are four of the top ten most profitable sneakers in the (subset) data. This model could prove useful for prediction if filtering by brand, as the outliers belong to a few specific brands. As well, it is interesting to note that retail price doesn't correlate with resale price – it is generally assumed that more expensive sneakers would innately have a higher resale value, but most of the sneakers resell at or around the par price. Data points low on the y-axis and high on the x-axis provide high profitability and should be examined.

4.3 – LSTM:



```
0               CT8527-100   3414
1               DD1391-100   3137
2   315122-111/CW2288-111    2836
3               DD1503-101   2522
4               CK9246-600   2253
...                    ...    ...
28343         1021A257-405      1
28344         1021A258-021      1
28345         1021A258-106      1
28346         1021A272-201      1
28347         1021A275-100      1
```

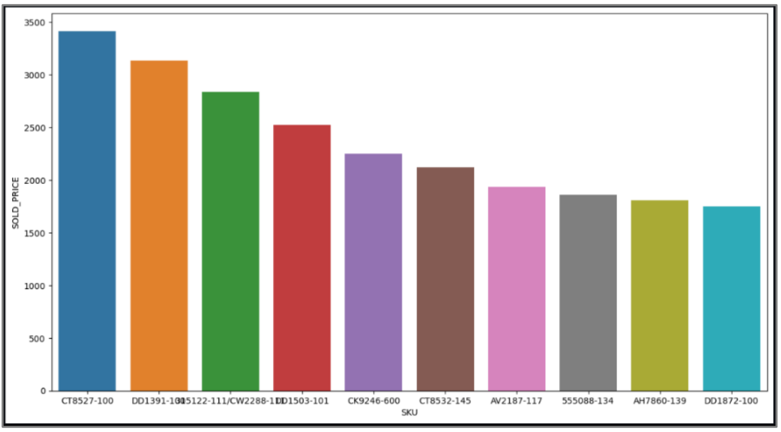


Figure 6. Number of sold items per SKU

The objective was to identify the highest number of sneaker sales for a specific sneaker in the dataset, enabling focused analysis. To achieve this, we aggregated the data by day, aiming to mimic stock market transactions with open and close prices. The plotted graphs represent the daily minimum and maximum prices for individual sneakers.

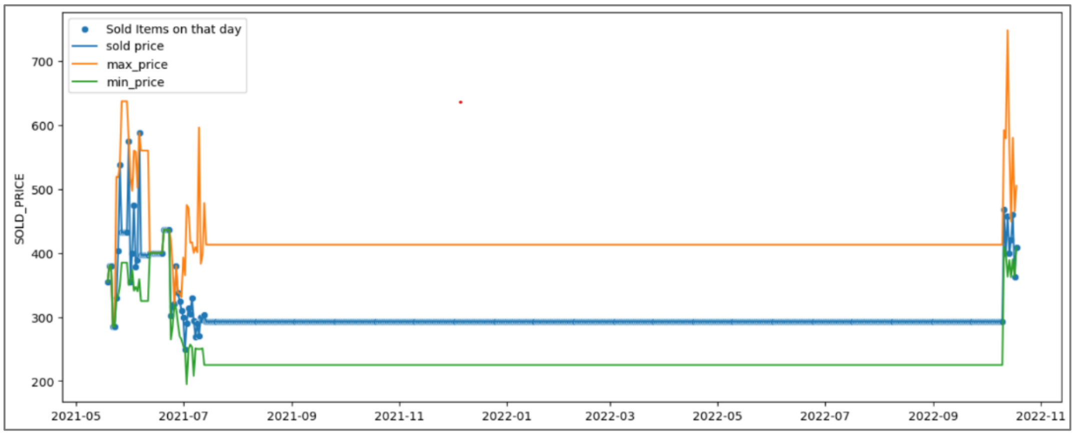


Figure 7. Shows max, min price for a day along with the transactions.

Figure 7 showcases the distribution of data concerning the maximum sale price, a crucial metric for online sales evaluation. It provides insight into the number of items sold for a specific SKU and size (Figure 8), an essential visualization for understanding sales trends.

Given online sales largely focus on the highest selling price, we've selected this as our key output variable:

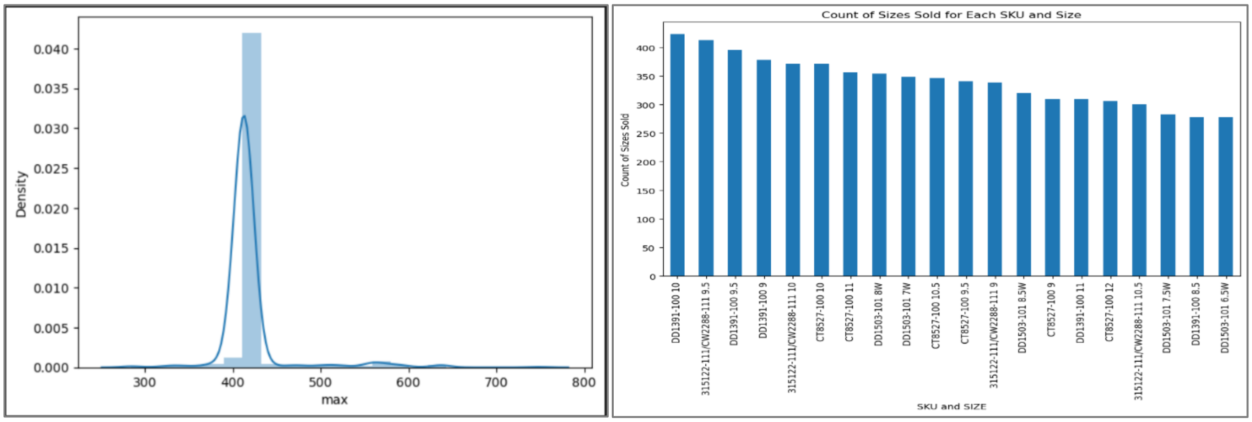


Figure 8. Number of items sold for a particular SKU and Size

Data Manipulation for Input Features:

We used sliding window as a base and taking the first 5 or 6 (depending on the window size we have setup in the code) prices as input and predicting the 6th / 7th

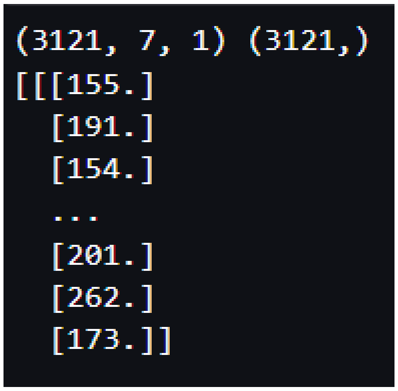
We added Dropout to LSTM to reduce overfitting.

1st scenario.

We picked price and using sliding window, we have window size input variable and predicting the output.

We pass that to our LSTM model with 128 nodes as input is set to number of variables in input list and

number of items in each index position example below has 7,1, cause 7 rows and each row has only 1 element in it.



```
(3121, 7, 1) (3121,)
[[[155.]
  [191.]
  [154.]
  ...
  [201.]
  [262.]
  [173.]]
```

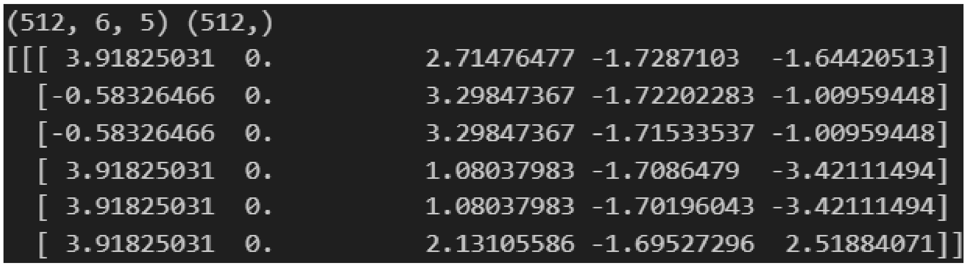
Figure 9. Data Transformation Explanation

The above figure shows the transformation of data using sliding window with combine 7 data points into 1.

Doing dropout to mitigate overfitting,

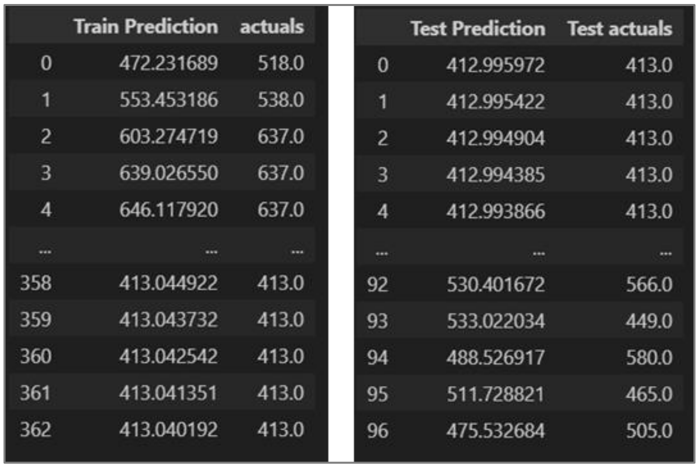
Number of nodes will help in splitting the data and depending on the

For the below it will be 6,5 because we have 6 rows, and each row has 5 elements.



```
(512, 6, 5) (512,)
[[[ 3.91825031  0.          2.71476477 -1.7287103  -1.64420513]
  [-0.58326466  0.          3.29847367 -1.72202283 -1.00959448]
  [-0.58326466  0.          3.29847367 -1.71533537 -1.00959448]
  [ 3.91825031  0.          1.08037983 -1.7086479  -3.42111494]
  [ 3.91825031  0.          1.08037983 -1.70196043 -3.42111494]
  [ 3.91825031  0.          2.13105586 -1.69527296  2.51884071]]
```

Figure 10. Data Transformation Explanation The above figure shows the transformation of data using sliding window with combine 6 data points



	Train Prediction	actuals
0	472.231689	518.0
1	553.453186	538.0
2	603.274719	637.0
3	639.026550	637.0
4	646.117920	637.0
...	...	...
358	413.044922	413.0
359	413.043732	413.0
360	413.042542	413.0
361	413.041351	413.0
362	413.040192	413.0

	Test Prediction	Test actuals
0	412.995972	413.0
1	412.995422	413.0
2	412.994904	413.0
3	412.994385	413.0
4	412.993866	413.0
...	...	...
92	530.401672	566.0
93	533.022034	449.0
94	488.526917	580.0
95	511.728821	465.0
96	475.532684	505.0

Figure 11. Actuals vs Predicted outputs for Training (left) and Test (right).

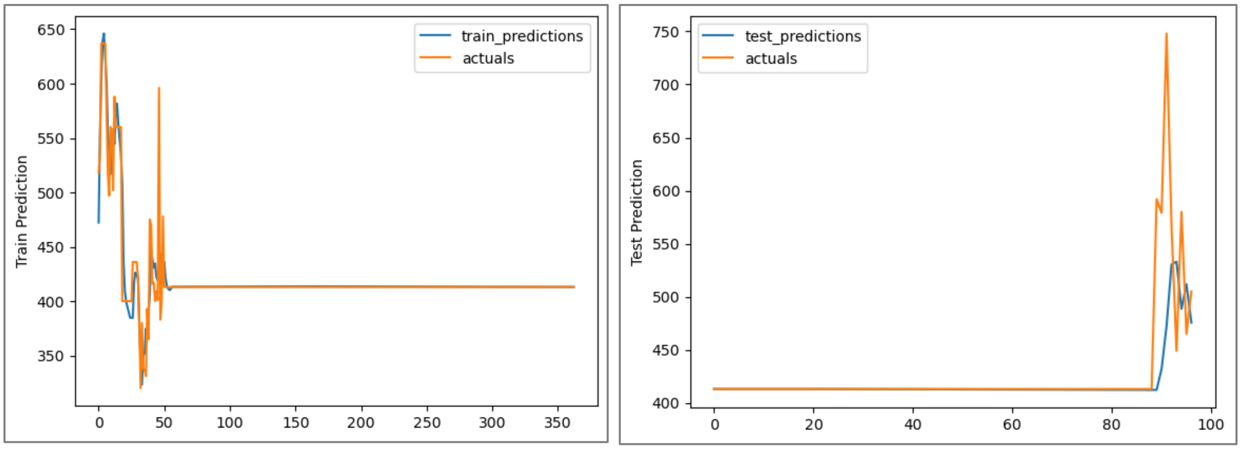


Figure 12. Model predictions with Training (left) and Test (right) data versus Actual labels.

6 – CONCLUSIONS:

In this project, we attempted to derive useful insights from the dataset provided by our project partner Arbit. Our aim was to offer information that can assist the organization in basing their decisions on the data gathered over the years.

Our general analysis of the top 25 most sold sneaker silhouettes reveals a positive correlation between the count of the item and its price, inferring that popularity drives the price. This suggests that consumers are willing to pay a premium for popular sneakers. Clustering analysis revealed that "Nike SKU's 308270-111 and 417744-001" are four of the top ten most profitable sneakers in the dataset, making them of special interest. Clustering reveals that mean sneakers contribute more than twice as much as all the other demographic groups combined. This finding can offer insight into future focus on other demographic groups as an opportunity for marketing companies to increase profit.

Comprehensive analysis has been done with the data provided to us to gather insights on the sneaker data using Segmentation & Regression, Clustering analysis which can be used to drive profits to the sneakers business. We have used K mean to identify different clusters (adding something maybe like 1 line). With LSTM we tried to predict the price in the future which helps business to encourage/recommend its customers to buy and sell certain sneakers.

<u>7 – CITATIONS</u>:

[1] Young, S. (2023, April 6). Inside the growing sneaker-resale market. Leaders.com. https://leaders.com/news/business/inside-the-growing-sneaker-resale-market/

[2] Sneakers - worldwide: Statista market forecast. Statista. https://www.statista.com/outlook/cmo/footwear/sneakers/worldwide

[3] Smith, J., et al. (2017). "Customer Segmentation in E-commerce: A K-means Approach." Journal of Retail Analytics, 12(3), 201-215.

[4] Elsworth, Steven & Güttel, Stefan. (2020). ABBA: adaptive Brownian bridge-based symbolic aggregation of time series. Data Mining and Knowledge Discovery. 34. 10.1007/s10618-020-00689-6.

<u>8 – APPENDIX</u>:

<u>Libraries</u>:
Pandas Library:
Pandas is a third-party library used for data structuring, analysis, and description.
Documentation: <u>https://pandas.pydata.org/pandas-docs/stable/</u>

Matplotlib Library:
Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.
Documentation: <u>https://matplotlib.org/</u>

Seaborn Library:
Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
Documentation: <u>https://seaborn.pydata.org/</u>

SciPy Library:
SciPy is algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics, and many other classes of problems.
Documentation: <u>https://scipy.org/</u>