**University Of Illinois- Urbana Champaign**

**Master of Science in Data Science**

**Mid-Term Project**

**2019-2020**

**Submitted By:**

**Kieyn Parks Lavanya Rao Sarang Deotale**

**Title Of Contents**

[Abstract](#_y14tby3vzzxn)

[Introduction](#_25io1e8u52wr)

[Motivation and Background](#_qkx6m9ocpcpz)

[Methodology(algorithm or Analysis)](#_gg8jjy2ql39a)

[Summary of Research Questions and Results](#_3nk4s1ym2xzv)

[Research Questions 1 & 2:](#_u90k3i8yd6in)

[In order to find the best selling sneaker brands, the raw data was stored in HDFS and processed using the Spark framework.](#_gypp80if4lun)

[Importing Data to HDFS via Spark:](#_twypk82m4v3l)

[Processing in SPARK:](#_hs0qxhardi5q)

[Using SPARK SQL to find the best brands:](#_jxkqmmsnyboq)

[References](#_qp02jwy0ze0v)

# **Abstract**

This project presents an exploratory and comparative study of the people who are more than enthusiastic in buying athletic shoes. These people are called ‘SneakerHeads’, who are ready to pay multiples of original retail price in order to buy their favourite pair of latest sneakers. Nike seems to be the leader of this industry followed by Adidas. This project will use a dataset provided by Stockx to analyze the demographics of the SneakerHeads along with drawing useful conclusions like below.

1. Which sneaker brand/names top this list?
2. Comparison of sales of different brands in course of the past 3 years.
3. Predict the socio economic status of these people based on the demographics and their average income.

# **Introduction**

In the United States and around the world there are many subcultures. One of these cultures is the shoe culture also known as sneakerheads. A sneakerhead is someone who collects and trades athletic footwear (usually basketball shoes). They treat their sneakers like pets and are willing to spend thousands of dollars to procure the latest pair.

# **Motivation and Background**

Sneakers are a $55 billion global industry. According to sneakerhead data website [StockX](https://stockx.com/), the secondary market for rare and limited-release sneakers is estimated to be [worth over $1 billion](https://stockx.com/sneaker-blog/2014/10/15/sneaker-resell-market-hits-1-billion/) **(TRIOMPHE, 2018)**. This is fueled by Nike and its long standing [Jordan brand](https://www.nike.com/us/en_us/c/jordan). People wait in lines for hours and spend tens of thousands of dollars just to get their hands on the latest pair of sneakers [BusinessInsider]. The growth of online retailing and auction sites (e.g ebay) has provided sneaker collectors with new methods to find the rarest shoes. Shops ,websites and events offer rare and exclusive sneakers.

In order to understand the world of sneaker collectors, we need to find out who are these people? Where do they live? What is their socioeconomic standing, average income, etc.

We hope to answer some of these questions, based on the analysis of StockX data set and supporting information.

**Dataset**

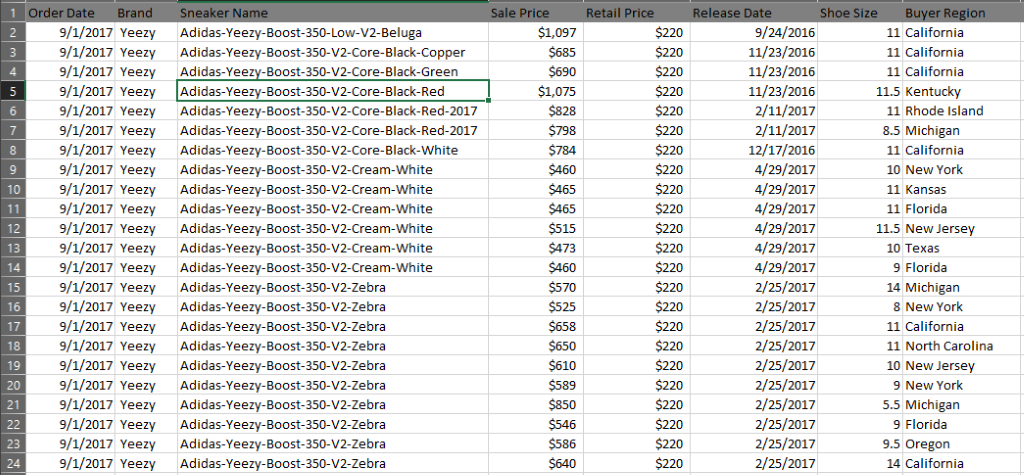


Figure 1: example of dataset in csv format – Source www.stocks.com

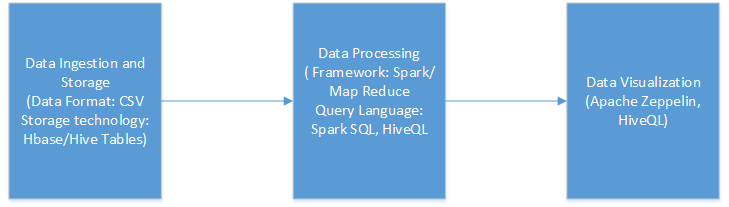
The dataset proposed in this paper contains sales data on popular culture luxury shoe brands adidas and Nike from across the united states. The dataset contains information on Off-White x Nike and Yeezy 350 sales from between 9/1/2017 to present day. The data is furnished by shoe retailer stock X, a large online consumer apparel trading site with main emphasis on luxury sneakers for use or collectors, and contains 100K records of random daily sales data; this is a random sample. The collection contains information on order date, Brand, Sneaker Name, Sales Price($), Retail Price($), Release Date, Shoe Size and Buyer State(USA). This is raw data, not calculated averages.



Figure 2: Average household income by county in each State in csv format.. Compiled by: Golden Oaks Research.

The data set of average household income was compiled by Golden Oaks Research and contains 30,000 records. The raw data is from the US census bureau and represents a time frame between 2011-2015 for all 50 states. Each record contains a unique id, state names, state code, state abbreviation, city, county, place (geographic location), type, primary (block), zip code, area code, aland (land area), awater (water area), Lat, lon, Mean income, Median (income), stdev, household (number of homes in the sample). Average household income per state can be calculated from the data.

# **Methodology(algorithm or Analysis)**



**Tools and Technologies**

1. **Hadoop**. Java software framework to support data-intensive distributed applications.
2. **MapReduce**. A flexible parallel data processing framework for large data sets.
3. **HDFS**. Hadoop Distributed File System
4. **HBase**. Key-value database
5. **Hive**. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.
6. **Spark**: An open-source distributed general-purpose cluster-computing framework.
7. **Spark-SQL**: Fast Data queries using SQL syntax or program abstract on in-memory data (dataset or dataframes)

# **Research Questions 1 & 2:**

## In order to find the best selling sneaker brands, the raw data was stored in HDFS and processed using the Spark framework.

## **Importing Data to HDFS via Spark:**

Apache Spark is a modern processing engine that is focused on in-memory processing. Spark’s primary data abstraction is an immutable distributed collection of items called a resilient distributed dataset (RDD). RDDs can be created from Hadoop input formats (such as HDFS files) or by transforming other RDDs. Each dataset in an RDD is divided into logical partitions, which may be transparently computed on different nodes of the cluster.

Spark DataFrames can be created from different data sources such as the following:

Existing RDDs  
Structured data files  
JSON datasets  
Hive tables  
External databases

Due to its flexibility and friendly developer API, Spark is often used as part of the process of ingesting data into Hadoop. With Spark, you can read data from a CSV file, external SQL or NO-SQL data store, or another data source, apply certain transformations to the data, and store it onto Hadoop in HDFS or Hive.

The following steps were performed in order to store the Sneaker sales data into HDFS.

1. Data Clean up: The data was cleaned using the Excel worksheet, to remove mis-formatting of the data.
2. The file was copied to the hdfs using the *dfs -put* command.

The stored data was then read using the Spark framework and the created Data Frame was queried to get the desired result.

## 

## **Evaluation/Results:**

Top 5 Adidas Brands:

|  |  |  |
| --- | --- | --- |
| **Brand** | **SneakerName** | **OrderCount** |
| Yeezy | adidas-Yeezy-Boost-350-V2-Butter | 11423 |
| Yeezy | Adidas-Yeezy-Boost-350-V2-Beluga-2pt0 | 10395 |
| Yeezy | Adidas-Yeezy-Boost-350-V2-Zebra | 10110 |
| Yeezy | Adidas-Yeezy-Boost-350-V2-Blue-Tint | 9297 |
| Yeezy | Adidas-Yeezy-Boost-350-V2-Cream-White | 9097 |
|  |  |  |

Top 5 Nike Brands:

|  |  |  |
| --- | --- | --- |
| **Brand** | **SneakerName** | **OrderCount** |
| Off-White | Air-Jordan-1-Retro-High-Off-White-University-Blue | 4635 |
| Off-White | Nike-Air-Presto-Off-White-Black-2018 | 1884 |
| Off-White | Nike-Air-Presto-Off-White-White-2018 | 1883 |
| Off-White | Nike-Air-VaporMax-Off-White-2018 | 1591 |
| Off-White | Nike-Blazer-Mid-Off-White-All-Hallows-Eve | 1435 |

**Best Time to buy the best selling Adidas brand:**

|  |  |  |
| --- | --- | --- |
| **SneakerName** | **MONTH** | **OrderCount** |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017 | December | 132 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017 | November | 125 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017 | October | 114 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017 | January | 106 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Red-2017 | September | 106 |

**Worst Time to buy the best selling Adidas brand:**

|  |  |  |
| --- | --- | --- |
| **SneakerName** | **MONTH** | **OrderCount** |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Green | April | 7 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Green | March | 9 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Green | August | 9 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Copper | April | 10 |
| Adidas-Yeezy-Boost-350-V2-Core-Black-Green | June | 10 |

**Best Time to buy the best selling Nike brand:**

|  |  |  |
| --- | --- | --- |
| **SneakerName** | **MONTH** | **OrderCount** |
| Air-Jordan-1-Retro-High-Off-White-Chicago | November | 215 |
| Nike-Air-Max-90-Off-White | November | 207 |
| Nike-Blazer-Mid-Off-White | November | 195 |
| Nike-Air-Presto-Off-White | November | 169 |
| Nike-Air-VaporMax-Off-White | November | 143 |

**Worst Time to buy the best selling Nike brands:**

|  |  |  |
| --- | --- | --- |
| **SneakerName** | **MONTH** | **OrderCount** |
| Nike-Air-VaporMax-Off-White | June | 10 |
| Nike-Air-Max-90-Off-White | August | 11 |
| Nike-Blazer-Mid-Off-White | July | 12 |
| Nike-Blazer-Mid-Off-White | June | 12 |
| Nike-Air-Max-90-Off-White | June | 13 |

**Forecast for the current year (Feb 2018- Feb 2019) based on the historical trends**:

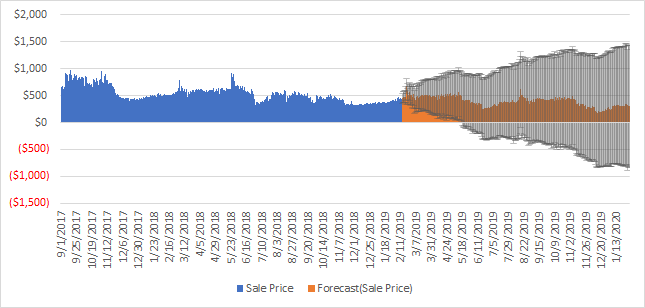


Figure3: Avg household income by state

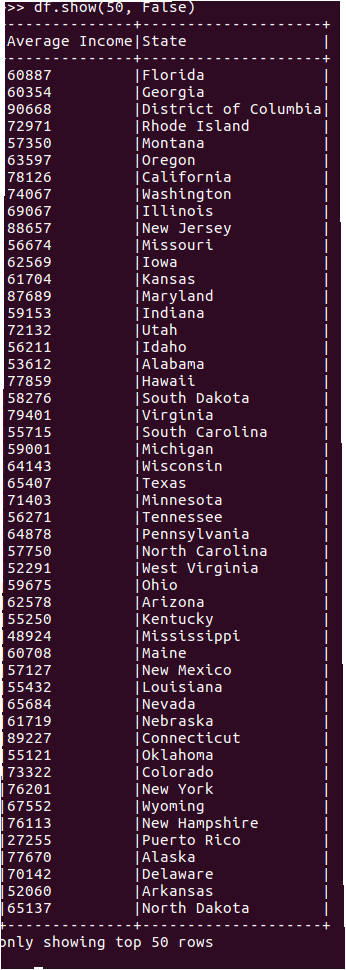


Figure 3: The above table is a summary of the data from figure 2. The table shows the average household income for each state.

# 

# 

# 

# 

Discussion Section:

* Relationships

# Related Works:

# **Conclusion:**

A detailed analysis of the SneakerHeads Data led to the following conclusions.

* There are certain time periods during a year, where the latest pairs of sneakers are sold the most. This was identified to be the holiday season (Nov, Dec).
* Identified the most popular pair and brand of sneakers which the sneakerheads are interested in.

**Future Work:**

The analysis done in the project can be exposed via web services to the online shopping sites to launch their new products at appropriate times of the year and at identified locations. This can help them focus their resources on particular places in order to have a higher return.

# **Summary of Research Questions and Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Research Question | Analysis Method/ Tools and Technologies | Result | Conclusion/visualization description | Owner |
| Which is the top 5 best selling Adidas and Nike sneakers. | Data storage: Using Hive on Hadoop and HiveQL/spark to query the dataset and visualize the result. | The top selling brands were identified using the Spark SQL | The data was stored as a structured data frame. The result was derived by applying SPARK sql queries on the data frame. | Lavanya Rao |
| Which is the Best/Worst Time to Sell the top 5 sneakers. | Data storage: Using Hive on Hadoop and HiveQL/spark to query the dataset and visualize the result. | Based on the result of the queries, the best time seems to Nov and Dec. The sales during Jan and Feb is less as compared to other months. | The data was stored as a structured data frame. The result was derived by applying SPARK sql queries on the data frame. | Lavanya Rao |
| Total Sales of Adidas vs. Nike brands for 2017,2018 and 2019 | Data storage: Using Hive on Hadoop and HiveQL to query the dataset and visualize the result. | Total sales of Adidas and Nike for year 2017 to 2019 was reported using Spark SQL/SCALA | The data was stored as a structured data frame. The result was derived by applying SPARK sql queries on the data frame  Some preprocessing was done. | Sarang Deotale |
| Customer willingness to pay by demographics. | Data Storage using HBASE.. Data processing using Hadoop Map Reduce with Java. Spark SQL (data frames) for fast queries. Zepplin for visualizations. | Successfully statewise customer willingness to pay and which type of shoes then can buy was determined | Two data sources 1> shoes sales 2> state wise income details were used. Spark RDD, Spark SQL and Scala functions was used | Sarang Deotale |
| When Should I Buy Yeezys/Off-Whites on StockX? Visualization showing how prices change before and after release date, and when is the optimal time to buy. | Data Storage using HBASE with Thrift for python. Data processing using Hadoop Map Reduce with Python. Spark SQL for fast queries. Jupyter Notebook/Zepplin for visualizations. |  |  | Kieyn Parks |
| From figure 2 calculate the states average household income from the means column. Evaluation of household income to find any correlation with sneaker sales. | Data Storage using HBASE with Thrift for python. Data processing using Hadoop Map Reduce with Python. Spark SQL (data frames) for fast queries. Jupyter NoteBook/Zepplin for visualizations. |  |  | Kieyn Parks |

# **References**

* “Designer shoe sales data”. StockX. Retrieved January 28, 2019.

<https://www.reddit.com/r/datasets/comments/as34yj/sneakers_sales_dataset_from_stockx/>.

* “US Household incomes”. Golden Oaks Research Group. Retrieved March 3, 2019. <https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations>.
* “Luxury sneakers: High Time in A Booming Market”,Catherine TRIOMPHE. Retrieved March 30, 2019.

<https://news.yahoo.com/luxury-sneakers-high-style-booming-market-033449512.html>.

* Importing data into HDFS: <http://www.informit.com/articles/article.aspx?p=2756471&seqNum=5>

**APPENDIX:**

## **Processing in SPARK:**

Below is the step by step process that shows the processing of the data in Spark engine.

**scala> case class SneakerDataClass (order\_date: String, brand:String, sneakername:String,saleprice: Double, retailprice:Double,release\_date: String,shoesize:Double,buyerregion:String)**

**defined class SneakerDataClass**

**scala> val salesRDD= sc.textFile("/user/udfs/ProjectData/SneakerHeadsRawData.csv")**

**scala> val salesDF= salesRDD.map(sale=> { SneakerDataClass(sale.split(",")(0),sale.split(",")(1),sale.split(",")(2),sale.split(",")(3).toDouble,sale.split(",")(4).toDouble,sale.split(",")(5),sale.split(",")(6).toDouble,sale.split(",")(7))}).toDF()**

**19/05/04 16:34:37 INFO HiveContext: Initializing HiveMetastoreConnection version 0.13.1 using Spark classes.**

**salesDF: org.apache.spark.sql.DataFrame = [order\_date: string, brand: string, sneakername: string, saleprice: double, retailprice: double, release\_date: string, shoesize: double, buyerregion: string]**

**scala> salesDF.printSchema**

**root**

**|-- order\_date: string (nullable = true)**

**|-- brand: string (nullable = true)**

**|-- sneakername: string (nullable = true)**

**|-- saleprice: double (nullable = false)**

**|-- retailprice: double (nullable = false)**

**|-- release\_date: string (nullable = true)**

**|-- shoesize: double (nullable = false)**

**|-- buyerregion: string (nullable = true)**

**scala> salesDF.registerTempTable("SneakerSalesTable")**

**scala> sqlContext.sql("select order\_date,brand,sneakername from SneakerSalesTable").show(5)**

**+----------+------+--------------------+**

**|order\_date| brand| sneakername|**

**+----------+------+--------------------+**

**| 9/1/2017| Yeezy|Adidas-Yeezy-Boos...|**

**| 9/1/2017| Yeezy|Adidas-Yeezy-Boos...|**

**| 9/1/2017| Yeezy|Adidas-Yeezy-Boos...|**

**| 9/1/2017| Yeezy|Adidas-Yeezy-Boos...|**

**| 9/1/2017| Yeezy|Adidas-Yeezy-Boos...|**

**+----------+------+--------------------+**

## **Using SPARK SQL to find the best brands:**

**Top 5 best selling Adidas:**

**scala> sqlContext.sql("select brand,sneakername, count(\*)as count from SneakerSalesTable group by brand , sneakername having brand=' Yeezy' order by count desc").show(5)**

**+------+--------------------+-----+**

**| brand| sneakername|count|**

**+------+--------------------+-----+**

**| Yeezy|adidas-Yeezy-Boos...|11423|**

**| Yeezy|Adidas-Yeezy-Boos...|10395|**

**| Yeezy|Adidas-Yeezy-Boos...|10110|**

**| Yeezy|Adidas-Yeezy-Boos...| 9297|**

**| Yeezy|Adidas-Yeezy-Boos...| 9097|**

**+------+--------------------+-----+**

**Top 5 best selling Nike:**

**scala> sqlContext.sql("select brand,sneakername, count(\*)as count from SneakerSalesTable group by brand , sneakername having brand='Off-White' order by count desc").show(5)**

**+---------+--------------------+-----+**

**|brand| sneakername|count|**

**+---------+--------------------+-----+**

**|Off-White|Air-Jordan-1-Retr...| 4635|**

**|Off-White|Nike-Air-Presto-O...| 1884|**

**|Off-White|Nike-Air-Presto-O...| 1883|**

**|Off-White|Nike-Air-VaporMax...| 1591|**

**|Off-White|Nike-Blazer-Mid-O...| 1435|**

**+---------+--------------------+-----+**

**Best time to buy the top brands:**

**scala> sqlContext.sql("select order\_date, count(\*) as count from SneakerSalesTable group by order\_date order by count desc").show()**

**+----------+-----+**

**|order\_date|count|**

**+----------+-----+**

**|11/16/2018| 1388|**

**|11/23/2018| 1262|**

**|12/27/2018| 1034|**

**| 6/30/2018| 1033|**

**| 2/7/2019| 893|**

**|12/14/2018| 890|**

**| 2/6/2019| 879|**

**|11/26/2018| 817|**

**|12/28/2018| 763|**

**|12/26/2018| 745|**

**|11/25/2017| 657|**

**|12/16/2017| 635|**

**|12/19/2018| 630|**

**| 1/4/2019| 594|**

**|10/31/2018| 585|**

**| 1/3/2019| 552|**

**| 5/30/2018| 547|**

**|11/30/2018| 538|**

**| 1/9/2019| 536|**

**|12/18/2018| 530|**

**+----------+-----+**