

Effects of Buying Black Shoes

Karan Mahendra Singh
Data Intensive Architectures
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
karanmsingh10@gmail.com

Abstract—This paper describes a project which aims at using data intensive architectures to deal with two sufficiently large datasets. The data quality of the data being used in the analysis is addressed according to the norms described in [3]. Python was used to remove columns irrelevant to the project and deal with missing values and different formats of similar columns in the datasets. After accessing Hadoop through Amazon EC2, MapReduce was employed to count the number of accidents that took place on same dates and corresponded with the purchase of black men's shoes. Further, it was used to make joins of two other pairs of columns to form a baseline for analysis of a total of three pairs of variables. Thus, this project has successfully found three insights which could only be found by putting these two datasets together. Pearson's correlation coefficient was used to test the correlation among these 3 pairs of variables which were later plotted on graphs. Therefore, the project has led to gaining in-depth knowledge of how Amazon EC2 could be used along with Hadoop and MapReduce for scaling computing power.

Index Terms—Python, Amazon EC2, Hadoop, MapReduce

I. INTRODUCTION

The color black is often associated with misfortune and negativity in many cultures across the world. Thus, the aim of the project is to find whether there is really any statistical proof, or it is just a myth. The objectives are as follows:

A. Black shoes and number of accidents

Checking if there is any relation between buying black shoes and meeting with an accident on the road. Black shoes being an independent variable and Number of accidents: dependent.

B. Black shoes and police investigation of an accident

Investigating if there is a correlation between black shoes and the accident not being investigated.

C. Number of cars involved in accidents and prices of shoes

Crashing a car incurs expense. This damage could be damage to other cars, public or private property. Probing the effects of cars crashing on the prices of shoes.

II. DATA

A. Men's Shoe Prices:

Contains data of men's shoes with information such as the date on which an item was bought, its color, brand and so on. Therefore, the data of black shoes according to the project's scope could be extracted [1].

B. Road Safety Data – Accidents 2017:

Data such as the date, time and place of an accident, the number of cars, accident severity and so on. Thus, relevant columns such as date and number of cars were extracted from this [2].

III. METHODOLOGY

A. Data quality:

According to the characteristics of data quality in [3], the datasets were accessed as follows:

- *Accuracy*

- *Semantic Accuracy:*

Date formats, colors of shoes, numbers of cars involved in an accident and so on, all displayed proper values under appropriate column names.

- *Syntactic Accuracy:*

The [2] had some unrealistic values in the binary variable: "DidPoliceInvestigate." Thus, the sum of positive (1) and negative (0) cases of the variable were taken to check the higher number of cases and were included in the case with the higher aggregate: (1) Positive.

Furthermore, [1] had some data where some shoes had prices: 0. All such cases were discarded.

Thereafter, prices of shoes in [1] were in Dollars which were converted to Euros.

- *Completeness*

The [1] had low completeness as it contained numerous missing values. All missing values had to be thus removed to have a complete dataset.

- *Consistency*

Apart from instances where there were multiple accidents on the same day, which were counted to obtain the number of accidents taking place on that day, both the data sets were found to be consistent with no incongruity. However, despite having a good level of consistency in the two datasets, the date formats were distinct and had to be changed into a common new format of ('%d/%m/%y').

- *Credibility*

The [1] is from the DataInfinity's product database, freely available on Kaggle meant for analyzing trends in luxury shoes, strategies of their pricing and so on. The [2] is obtained from the authentic website of data.go.uk published by the Department of Transport meant for the reference of researchers, public and other stakeholders.

- *Currentness*

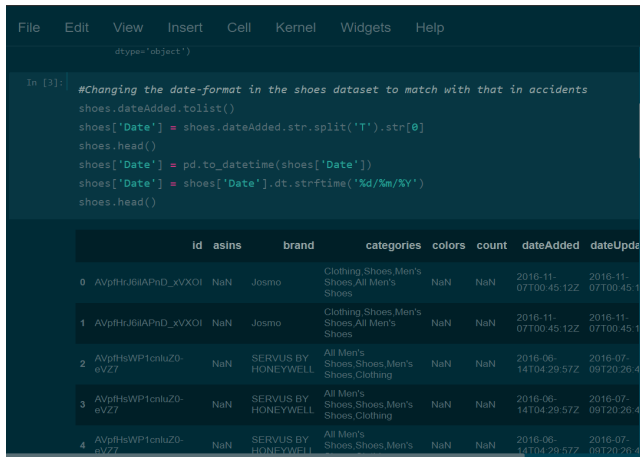
The presence of date columns in both datasets shows currentness of the years 2016 and 2017. Thereafter the presence of brands such as Gucci, Saint Laurent, Nike and so on which were in the top 10 rankings in 2017 asper [4] adds to the currentness.

IV. IMPLEMENTATION AND ARCHITECTURE

A. Data Cleaning

The project began with data cleaning. Given the flaws stated above, data cleaning was inevitable and it was done with Python.

Fig. 1 is an example of the data cleaning performed on the datasets. It shows changing the format of the date column in a dataset used in the project.



```

In [3]: #Changing the date-format in the shoes dataset to match with that in accidents
shoes.dateAdded.tolist()
shoes['Date'] = shoes.dateAdded.str.split('T').str[0]
shoes.head()
shoes['Date'] = pd.to_datetime(shoes['Date'])
shoes['Date'] = shoes['Date'].dt.strftime('%d/%m/%Y')
shoes.head()

```

	id	asins	brand	categories	colors	count	dateAdded	dateUpdate
0	AvpRrJ8iAPhD_vX0i	NaN	Josmo	Clothing, Shoes, Men's Shoes, All Men's Shoes	NaN	NaN	2016-11-07T00:45:12Z	2016-11-07T00:45:12Z
1	AvpRrJ8iAPhD_vX0i	NaN	Josmo	Clothing, Shoes, Men's Shoes, All Men's Shoes	NaN	NaN	2016-11-07T00:45:12Z	2016-11-07T00:45:12Z
2	AvpRrJ8iAPhD_vX0i	NaN	SERVUS BY HONEYWELL	All Men's Shoes, Shoes, Men's Shoes, Clothing	NaN	NaN	2016-06-14T04:29:57Z	2016-07-09T20:28:40Z
3	AvpRrJ8iAPhD_vX0i	NaN	SERVUS BY HONEYWELL	All Men's Shoes, Shoes, Men's Shoes, Clothing	NaN	NaN	2016-06-14T04:29:57Z	2016-07-09T20:28:40Z
4	AvpRrJ8iAPhD_vX0i	NaN	SERVUS BY HONEYWELL	All Men's Shoes, Shoes, Men's Shoes	NaN	NaN	2016-06-14T04:29:57Z	2016-07-09T20:28:40Z

Fig. 1. Data cleaning in Python

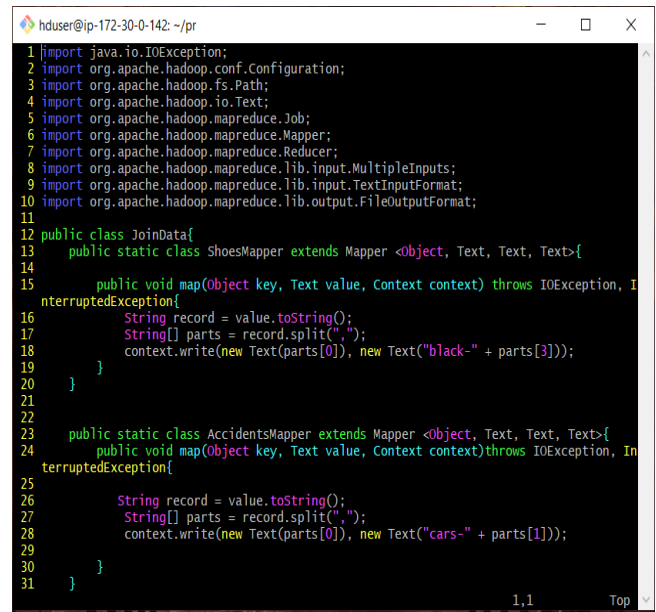
B. Application workflow

1) Mappers

The application initiates with mapping input files. The input files are datasets of (.csv) format. Mapping in this project starts with splitting inputs and passing the key which is *date*, the 0th index of both the input files. Thereafter, a tag: "black" and "cars", marking the start of another column is used. See Fig. 2.

2) Reducers

Reducers in this project work differently on different



```

1 import java.io.IOException;
2 import org.apache.hadoop.conf.Configuration;
3 import org.apache.hadoop.fs.Path;
4 import org.apache.hadoop.io.Text;
5 import org.apache.hadoop.mapreduce.Job;
6 import org.apache.hadoop.mapreduce.Mapper;
7 import org.apache.hadoop.mapreduce.Reducer;
8 import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;
9 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
10 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
11
12 public class JoinData {
13     public static class ShoesMapper extends Mapper<Object, Text, Text, Text> {
14
15         public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
16             String record = value.toString();
17             String[] parts = record.split(",");
18             context.write(new Text(parts[0]), new Text("black-" + parts[3]));
19         }
20     }
21
22     public static class AccidentsMapper extends Mapper<Object, Text, Text, Text> {
23         public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
24             String record = value.toString();
25             String[] parts = record.split(",");
26             context.write(new Text(parts[0]), new Text("cars-" + parts[1]));
27         }
28     }
29 }
30
31

```

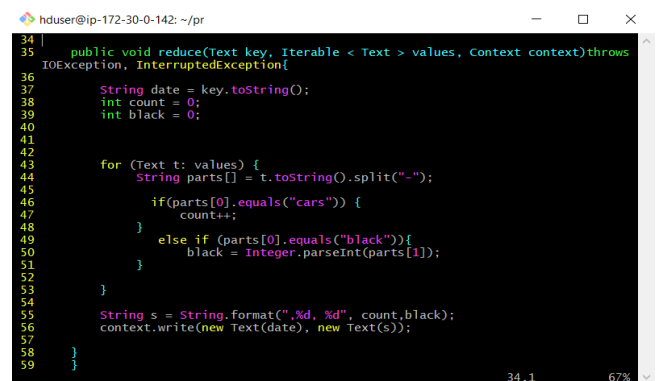
Fig. 2. Mapper used for fetching input from dataset

insights.

On the insight stated in Section I, A, a counter is used to count number of accidents on a particular date. This counter increments on the sight of "cars" in the parts array defined for inputs and prints the occurrence of 1 or 0, 1 indicating the sight of a black shoe and 0, otherwise. Thereafter, on the Section I, B insight, it prints the column contents on sighting "cars". Here, 1 stands for investigations done and 0 for non-investigated accidents; spotting of black shoes too is printed as it is from the parts array.

Lastly, for the insight in Section I, C, the reducer aggregates the number of cars crashed in the column to count the total number of cars involved in accidents on that date along with the prices of shoes correspondingly bought on that day.

The output from the Reducers formed the base for further



```

34 public void reduce(Text key, Iterable<Text> values, Context context) throws
35 IOException, InterruptedException {
36     String date = key.toString();
37     int count = 0;
38     int black = 0;
39
40     for (Text t : values) {
41         String parts[] = t.toString().split(",");
42         if (parts[0].equals("cars")) {
43             count++;
44         } else if (parts[0].equals("black")) {
45             black = Integer.parseInt(parts[1]);
46         }
47     }
48     String s = String.format("%d, %d", count, black);
49     context.write(new Text(date), new Text(s));
50 }
51
52
53
54
55
56
57
58
59

```

Fig. 3. Reducer used for counting the number of accidents and show black shoes

analysis. See Fig. 5.

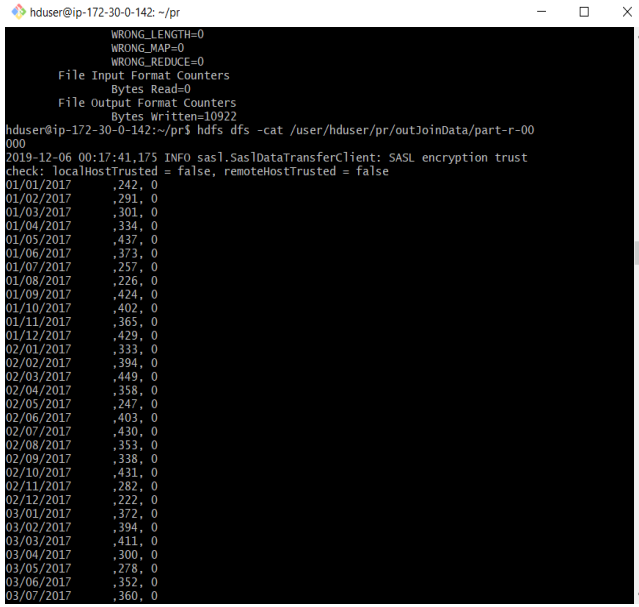


Fig. 4. Output of the first MapReduce showing Date, No of accidents and occurrence of black shoes

C. Analytics

The entire analytics was done with Python using the Pearson's Coefficient of correlation. For the result of Section I, A which involved correlating a dependent variable which was continuous and a binary independent variable, Point-Biserial Correlation Coefficient was used in addition to Pearson's correlation for verifying the correlation since Point-Biserial Correlation is generally used for variables of the aforementioned kind.

V. RESULTS

A. Black shoes and Number of Accidents

Null Hypothesis: Buying black shoes results in accidents. The Pearson's Correlation Coefficient for buying Black Shoes and number of accidents is shown in Fig. 5.

	NoOfAccidents	BlackShoes
NoOfAccidents	1.000000	-0.333695
BlackShoes	-0.333695	1.000000

Fig. 5. Pearson's Correlation Coefficient for Black shoes and the number of accidents.

Since the correlation is -0.33 which is greater than -0.50 (thus further from -1 and closer to 0), it may be concluded

that the variables are moderately related. Moreover, the correlation is negative which means with the increase in the number of black shoes, the accidents would decrease. The output is verified with Point-biserial Correlation Coefficient which is used for correlations between one binary one continuous variable, the result is in Fig. 6.

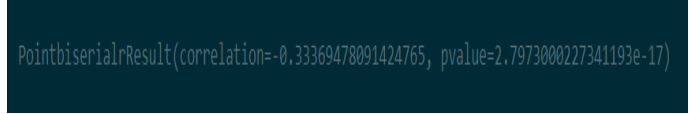


Fig. 6. Point Bi-serial Correlation Coefficient for Black shoes and the number of accidents.

The Point Bi-serial Correlation Coefficient has the same result as the Pearson's Correlation Coefficient and hence the result is held.

Finally, the correlation coefficient is lesser than 0.50 which means it is not large enough to defeat the null hypothesis and thereby, there could still be a chance of meeting with an accident after buying black shoes.

To better understand the relationship, it was plot as a scatter plot since there was a binary independent variable as shown in Fig. 7.

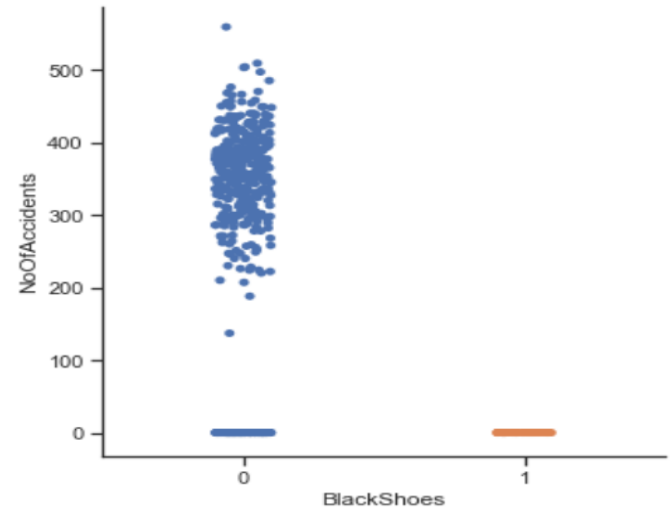


Fig. 7. Scatter plot of Black shoes and number of accidents.

B. Black shoes and police investigation of an accident

Null Hypothesis: Buying black shoes leads to no investigation of the accident.

The obtained correlation result is as shown in Fig. 8. The Pearson's correlation coefficient for the above null hypothesis is -0.34. Although the coefficient indicates a negative correlation in favour of the null hypothesis, its value is greater than -0.50 and closer to 0 implying that there is not enough evidence in support of the null hypothesis therefore,

	DidPoliceInvestigate	BlackShoes
DidPoliceInvestigate	1.000000	-0.341187
BlackShoes	-0.341187	1.000000

Fig. 8. Pearson's Correlation Coefficient for Black shoes and Investigation of the accident.

it has to be rejected. Thus, there is no correlation between buying black shoes and the accident scene being investigated by the police.

The scatter plot for the correlation is Fig. 9.

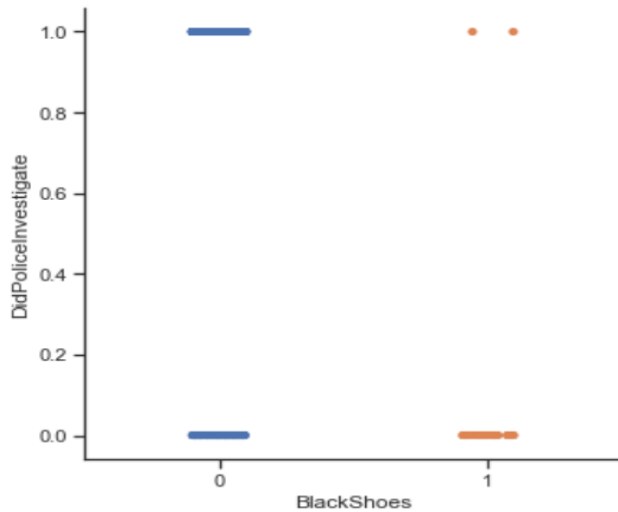


Fig. 9. Pearson's Correlation Coefficient for Black shoes and Investigation of the accident.

C. Number of cars involved in accidents and prices of shoes

Null Hypothesis: Higher the number of cars crashed in accidents, higher the prices of shoes.

Pearson's Correlation Coefficient is shown in Fig. 10.

	NoOfCarsInAccidents	PricesOfShoes
NoOfCarsInAccidents	1.000000	0.094048
PricesOfShoes	0.094048	1.000000

Fig. 10. Pearson's Correlation Coefficient for number of cars involved in accidents and prices of shoes.

From Fig. 10, it can be seen that the coefficient has the value 0.094 which is very close to 0. Thus, it can be concluded that there is almost no correlation between number of cars crashed in accidents and the prices of shoes.

Since both variables are continuous, the variables must be plot on a line graph to see if there is any linearity.

It is seen that the relationship has no linearity thus, a scatter plot could better depict the relationship.

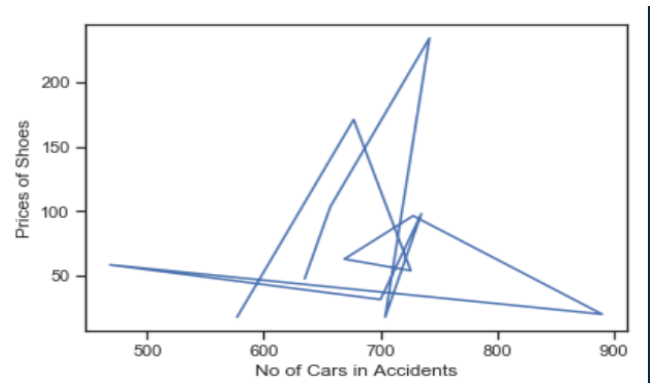


Fig. 11. Line graph for number of cars involved in accidents and prices of shoes.

The dispersed data points plotted on the scatter plot

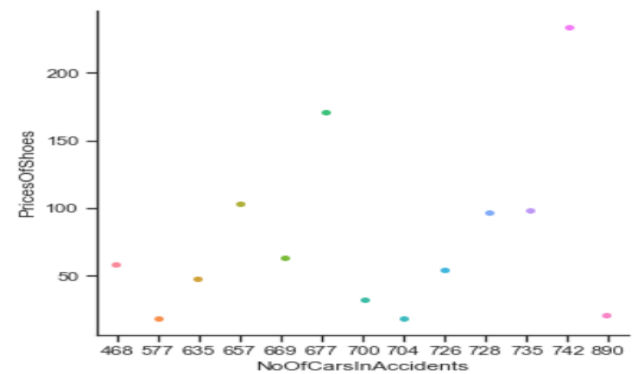


Fig. 12. Scatter plot for number of cars involved in accidents and prices of shoes.

show there is too much variation and little correlation between the number of cars crashed in accidents and the prices of shoes.

D. Challenges faced

- A key challenge faced in establishing a correlation between variables was the variation in the spread of variables. From the low correlation it is apparent that the data was widely spread which affected the overall output and thus, no good correlation was successfully found.
- Finding black shoes from the dataset containing shoes of numerous colors was a challenge at first. Applying a lambda sum function on the colors column of [1] which incremented on sight of the word "black" temporarily solved the problem until it was realized that the lambda function counted multiple "black" words in a row and instead of making the column binary, made it categorical. This was fixed by replacing values greater than 1 by 1. Fig.13 shows the challenge.

```

File Edit View Insert Cell Kernel Widgets Help Python 3
In [8]: #Adding a column to indicate whether the color of shoes is black or not
shoe_f['colors'] = shoe_f['colors'].str.lower()
shoe_f['Black_shoes'] = shoe_f['colors'].apply(lambda x: sum(1 in r".black" for i in x.split()))
shoe_f.head()

   Date                colors ShoesPrices Black_shoes
13  15/11/2016  red.black sketchy slant yellow.blackblue.black    9.99          0
14  15/11/2016  red.black sketchy slant yellow.blackblue.black   19.99          0
15  15/11/2016  red.black sketchy slant yellow.blackblue.black    25          0
16  15/11/2016  red.black sketchy slant yellow.blackblue.black   15.99          0
17  15/11/2016  red.black sketchy slant yellow.blackblue.black    9.99          0

In [9]: #cleaning values
shoe_f['Black_shoes'].loc[shoe_f['Black_shoes'] > 1] = 1
shoe_f.head()

```

Fig. 13. Challenge: Finding black shoes

VI. CONCLUSIONS AND FUTURE WORK

Thus, the project has greatly helped in gaining in-depth knowledge of all procedures such as cleaning the data, transferring it to a cloud platform to scale up the computing power and obtain results from heavy datasets, coding programs that handle such data in a highly parallel and time efficient manner and so on, which are required to be performed while dealing with data intensive architectures. It is interesting to not only know how such architectures work, but also run them to achieve the objectives of an individual project such as this. It would certainly be interesting to dig deeper into the data sets and obtain insights that would be put to use for the betterment of current systems. For instance, derive such results that would help gain information about why accidents occur in those places where the observed frequency of accidents is high and improve the situations there and as a result, minimize mishaps in those areas. Thereafter, even in the men's shopping dataset, trying to analyse the success of particular brands, their pricing trends and their impact on the overall economy could possibly be a part of the future work.

REFERENCES

- [1] Kaggle.com. (2019). Men's Shoe Prices. [online] Available at: <https://www.kaggle.com/datafiniti/mens-shoe-prices> [Accessed 5 Dec. 2019].
- [2] Data.gov.uk. (2019). Road Safety Data - data.gov.uk. [online] Available at: <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data> [Accessed 28 Nov. 2019].
- [3] eCOMITIA 2.0, n., Olucaro Dashboard 1.0, n., Prometheus IDS Core 1.0, n., Cibersad and SIXA, n. and ProEducative 3.0, f. (2019). ISO 25012. [online] Iso25000.com. Available at: <https://iso25000.com/index.php/en/iso-25000-standards/iso-25012> [Accessed 5 Dec. 2019].
- [4] Evening Standard. (2019). 10 most popular fashion brands of 2017. [online] Available at: <https://www.standard.co.uk/fashion/news/10-most-popular-fashion-brands-of-2017-a3724121.html> [Accessed 5 Dec. 2019].