October 14, 2017  
Data Science Assignment  
John G Schwitz

Executive Summary: My objective is to rapidly achieve the competitive advantage of technology companies through exploiting Data Science and ‘streaming’ infrastructure transition. Execution requires a process such as:

1. Formulation of Business Objectives and Success Metrics
2. Rapid DS discovery from Idea -> Code -> Experiment
3. Develop processes to Assure the integrity of data
4. Competitive Advantage from DS discovery to rapid implementation at scale
5. Continuous Learning and Optimization

Treating this work as a beta test I; 1) Formulate a Process, 2) Identify Anomalies and Data Transformation, 3) Outline Business Objectives, and discuss 4) Implementation at Scale.

I would like to emphasize Critical Success Factors of my Analysis

1. It is just analysis unless it is actionable  
   This requires teaming with Line Executives to perform analysis that is actionable
2. Data Science Insights must be immediately transferable to production  
   All the analysis is captured into a reproducible Notebook which is ‘finished’ for Executive consumption and sharable with other Analysts. Behind the firewall I would use Shiny to create a website.
3. What I executed for 2 stores I can execute for 5,000 for 1 to several years on the Cloud with Spark.

A summary of the findings:  
  
There are problems with the data. Only using GMT does the data span July 22 & 23.

1. There is missing data by certain hours for store 3333.
2. Significant data preprocessing is required for analysis
   1. Scan errors should be examined for fraud and eliminated to prevent skewing any analysis by purchase size. I collected the largest errors.
   2. Finance should be segmented in a distinct file and analyzed as a Risk Management Practice.
   3. Gas was segmented and did not contain amounts.
3. Spending by hour can be used to manage store manpower schedules.
4. Store 3333 generates less than half the volume of store 2222.
5. A third of the departments generate 80% of the dollar volume
6. A preponderance of high-dollar baskets had only 1 item. Why no complementary items?
7. The reproducible analysis is attached (WalmartReproducableCode).

I conclude with what I can rapidly deliver for Walmart.

***Process & Substantive Findings***

1. Clean data and identify anomalies
   1. Best to eliminate and identify for inspection large items that were scanned and canceled. These transactions should be examined.

Anomalies to Investigate

Store Dept Reg UPC Amount Date

2222 33 15 2000000000000000000033 +/ $40004.00 7/22 10:32 PM

2222 72 18 0000000000888143002400 +/- $268.00 2x 7/23 5:46 PM

3333 61 7 0000000000081312502746 +/- $398.00 7/22 8:46 PM

2222 16 74 0000000006937419204870 +/- $147.00 7/23 7:50 PM

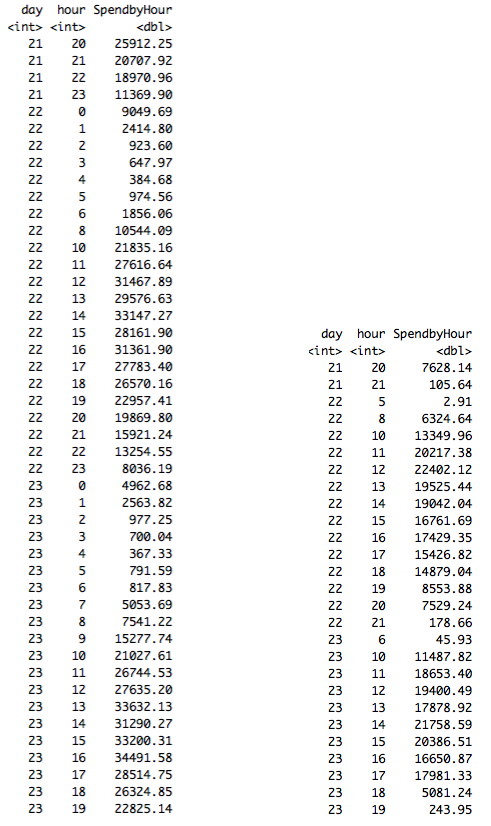
2222 16 74 0000000006937419204870 +/- $147.00 7/23 3:48 PM

2222 16 74 0000000006937419204870 +/- $147.00 7/23 3:51 PM

2222 23 15 2000000000000000000023 +/- $5005.00 7/23 6:52 PM

3333 64 6 0000000000000000924474 +/- $179.88 7/21 8:48 PM

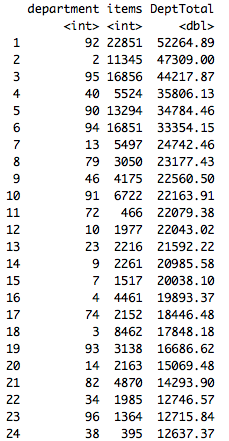
1. Assure data collected is complete and has no duplicates. Data collection from stores is prone to these errors.
   1. Duplicates: There are no duplicated lines.
   2. Missing Data: Assuming stores on East coast (simply shift times if incorrect); I found the following problems. Time format is POSIX in microseconds (13 Ints).
      1. Unless the stores report on GMT (which would hinder staffing and store comparisons), there is Friday data for US Time zones, and considerable missing data. See page 2 of Walmart Analytics. I generated min/max quartiles for data validation. Min Time is 1500681613000. Entering in EpochConverter: <https://www.epochconverter.com/> gives Fri July 21, 8:00 PM GMT -4:00 DST.
      2. GMT time means the store is in England and has great business from 1 to 4 AM. Any US stores west of NYC have even more hours on Friday.
      3. Store 3333 clearly demonstrates the problem. There is missing data for the 22’nd at 6,7,9 and 23’d at 7,8,9.   
         See charts next page
2. I segmented out Financial transactions (Money Orders etc) and Gasoline purchases into separate Files for analysis. Finance requires audit and extra vigilance, thus I have a file for analysis. I discovered the large financial transactions when examining outlier amounts over $100. Gasoline should not skew basket analysis, and additionally the amounts are not captured on gas transactions. Gas amounts are typically 2.0 & 2.3.
3. In the future I would analyze Tires and segment out the pharmacy.

Store 2222 Spending by Data and Time Store 3333 Spending by Date and Time  


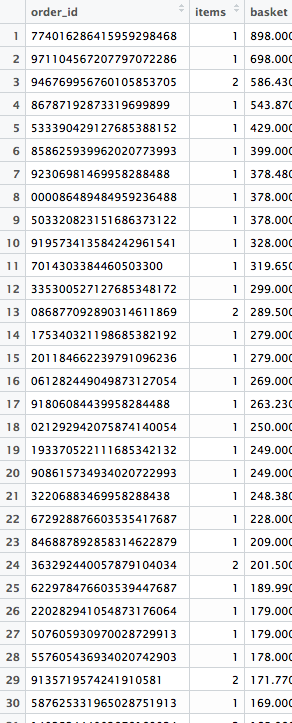
Analysis

1. Store to Store comparison demonstrates 3333 does less than half the volume of 2222.  
   Excludes Finance transactions and Gasoline. Store 3333 is not open the same hours as 2222.

|  |  |  |
| --- | --- | --- |
| Store | July 22 Spending | July 23 Spending |
| 2222 | $364,356 | $324,739 |
| 3333 | $181,623 | $149,569 |

1. Analysis of Store 2222.   
     
   24 of 66 (36%) of the departments generate 80% of volume  
   

The high-dollar baskets are dominated by single item purchases.



Analysis must be actionable!



I have a driving ambition to deliver key technologies to Walmart enabling a robust competitive response to online Retailers. I achieve actionable results through facilitating collaborative working sessions with Line Executives. This elicits the desired results and success metrics.

Four areas where I can make immediate contributions using ‘technology innovations’ coupled with Walmart presence. This turns back the ‘weapon’ on technology companies!

1. Recommender Systems
2. Countering Amazon by Exploiting Walmart Presence
3. Delivering Data Science at Scale
4. Supply Chain Optimization through Blockchain

***Recommender Systems*** are such a critical component that in 2006 NetFlix offered $1 Million to the team that could improve their movie recommendations by 10 %. Winning required openly publishing the technologies. The prize was awarded in 2009. This technology presents the recommendations on Amazon.

I have used this technology, collaborative filtering, and believe Walmart can turn this to Walmart advantage. Walmart has the customer present, not internet attention, and millions of baskets from similar demographic stores. I can exploit this.

***Countering Amazon by exploiting Walmart Presence***

Inventory cost/requirements scale as the square root of outlets maintaining inventory. Therefore; 1,000 stores require 32 times the inventory as a central store (Amazon). Amazon is moving into the high-dollar large/heavy items. Walmart can combine its presence with direct-to-store delivery perhaps followed by store to household delivery (the last mile) to win in this space. What Dominos can do Walmart can do.

Winning requires Data Science combined with an understanding of optimizing logistics. I worked this problem on two previous assignments.

***Delivering Data Science at Scale***

I have command of the Technology Trilogy which Walmart must exploit to rapidly scale Data Science to Production. The Technology Trilogy is Kafka (messaging), Cassandra (NoSQL), and Spark (distributed analysis). This is integrated to scale using Akka/Scala/Java and exploited using Scala/Python/R/Julia.

***Supply Chain Optimization through Blockchain***

I have a recent Masters in Science and Technology with a Cybersecurity/Cryptography focus. Blockchain exploits Cryptography to provide distributed processing/validation against an immutable ledger. I can serve on or lead a committee to exploit this technology for Walmart.

These are only four of the areas where I could contribute to Walmart’s transformation to meet and beat the competition. I would like to be part of the team accomplishing this.