# USING VISUAL ANALYTICS AND K-MEANS CLUSTERING FOR MONETISING LOGISTICS DATA, A CASE STUDY WITH MULTIPLE E-COMMERCE COMPANIES

**Hamzah Qabbaah, George Sammour, Koen Vanhoof**

***Abstract***:

*Logistics companies possess and collect a large amount of data on the shipments they perform while at the same time facing a challenge to understand their complicated market better. Therefore, investigating whether large databases gathered by logistics companies on their e-commerce partners could be monetised as a business service and how this could eventually be achieved is an important research venture. In this paper we used visual analytics and k-means clustering to see whether the data could be structured and presented in a monetisable way, while at the same time adhering to the quality characteristics necessary for doing so: reliable, accurate, relevant, segmented, secured and anonymized. Results show that is clearly the case for the database we investigated and contained 85989 transactions. Using a semi-structured interview with several key managers of both the logistics company and its e-commerce partners, a business-model canvass was developed that indicates the necessary elements for this venture and the right mindset to manage the process. We can confidently conclude that all elements are present to answer the monetisability question positively and to pretend that given the right visualization and confidence between the companies the process could very well be profitable.*

***Keywords***: *Data monetisation, Visual analytics, K-means clustering, Logistics, E-commerce.*

## 1. Introduction

The concept of data monetisation is to be situated in the broader context of data sharing between companies Business to Business (B2B) Data Sharing. The exchange of data between organisations producing and holding data constitutes the 'data market' (data supplier companies, representing the supply side of the market) and organisations using or re-using data (data user companies, representing the demand side of the market) is situated at the core of B2B data sharing, representing the exchange of data resources with multiple applications or users in different organisations (Carnelley 2018).

Many companies use the data produced in their value chain only internally to improve productivity, manage costs, or improve customer relationships. They do not realise that data are also an asset, when they are used to create new revenue externally. Data monetisation is all about this way of exploiting data (Carnelley 2018).

Findling et al. (Serge Findling 2018) identified three primary paths towards realising that value for data: (1) data sale or licensing creating immediate revenue; (2) bundling data with other services or products thus creating extra value for these services; and (3) exchanging premiums/trade advantages or discounts for data. Our research only deals with the first two ways of monetising data.

## 2. Background

Several definitions of data monetisation have been put forward in literature. Moore (Moore 2015) stresses that through data monetisation data are transformed into a source of profit. Najjar and Kettinger (Najjar and Kettinger 2014) describe data monetisation more clearly by stating that "Data monetisation happens when the intangible value of data is converted into real value, usually by selling them". The definition by Fred (FRED 2017) can be considered as being more generic, since in his eyes the revenue will not only be obtained by selling the data but also by creating different applications and services. His definition identifies data monetisation as: "The revenue generation

with and out of data and data-derived information-based products and services". Finally, Thomas and Leiponen (Thomas and Leiponen 2016) have developed a categorisation scheme for monetising data. They identify three categories : (1) Selling data; (2) Providing insights or analyses and (3) Creating new services.

Scientific literature indicates a number of factors favouring the creation of value through data sharing (Liu, Ren et al. 2009, Najjar and Kettinger 2014, Bonneau 2015, Bataineh, Mizouni et al. 2016). They can be subdivided in three areas: technological factors, business factors and execution factors.

All three areas may have relevance either to the quality of the data that can be shared and monetised or to the mind-set and processes going on in the company.

Companies have to assess the value of their data carefully. From a theoretical point of view, four characteristics or considerations are relevant (platform 2018) (Bataineh, Mizouni et al. 2016).

The data have to be: (1) Reliable and accurate, (2) Relevant, (3) Segmented and (4) Secure and anonymised.

Data monetisation creates a new business model for the company involved. Data are not only used to run the business anymore, they become a product/service that can be sold to partners and generates an income larger than the cost of creating and gathering the data (Liu, Ren et al. 2009, Najjar and Kettinger 2014). This signifies that not only a correct way of connecting and sharing the data with the partners has to be developed, but also a route to sell the data in the best possible way. The connecting and sharing of the data is a technological factor in terms of the conditions for success, whereas the selling mode is both a business and execution factor (Najjar and Kettinger 2014).

Connecting and sharing information with suppliers necessitates an improved technical capability by developing a supplier portal (hosted by a third-party analytics firm) that allows to share information with partners (Najjar and Kettinger 2014, Bonneau 2015).

## 3. Literature Review

Gottlieb and Rifai (Company 2017) have indicated the rapid development of the data sharing and monetisation market. An enquiry among 530 executives and senior level managers revealed that data monetisation was widely seen as a new means of generating revenue. 41% of the respondents had begun doing so in recently. Everis (Catarina Arnaut 2018) indicated that among companies starting to share their data with other companies a majority of 52% asked some form of remuneration for this sharing, while 40% were willing to share their data to a limited number of partners for free. 8% shared data freely to a wider audience. The generated income (Catarina Arnaut 2018) is very diverse however and ranges from 5000 euros (for one third of all respondents) to more than one million euros (in one fifth of these cases), depending also on the size of the companies in the sample. These investigations prove that data monetisation is slowly becoming part of the business world.

Literature mentions a number of examples. The most explicit one is Dawex (Carnelley 2018). Founded in 2015, this French data marketplace operates a website by connecting companies selling or buying data, using standard enterprise software. Dawex has managed in two years to progressively enlarge its data offer to a wide array of industries, from automotive, to energy, from agriculture to retail, healthcare and, more recently, financial services. 2,000 companies are connected to the platform. 45% of them are situated in Europe, 38% in the United States, and 17% in Asia.

All other studies are situated in specific sectors of industry. Perrons and Jenssen examined existing data management practices in the upstream oil and gas industry and compare them to practices and philosophies that have emerged in organisations that are leading the way in Big Data. The comparison showed that this kind of data can be regarded as a valuable asset, although they are frequently just regarded as descriptive information about a physical asset (Perrons and Jensen 2015). Bataineha et al. (Bataineh, Mizouni et al. 2016) investigated the use of data gathered from customers in the mobile phone market. 'Mobile phone-based sensing' is a new business practice aiming at using smart phones to answer sensing requests and collect useful data. A

wide variety areas ranging from health-care applications (think actually of the contact tracing apps developed in covid-times) to pollution monitoring are benefiting from these data. Xu et al. (Xu, Qiu et al. 2019) have studied the use of data-driven logistics in commerce from the perspective of risk management. The paper focuses on quantitative operational risks in E-commerce. These operational risks mainly refer to the risks owing to supply/ demand uncertainties, human mistakes and accidents that would decrease the service level or threat the normal operations.

To the best of our knowledge, monetisation of personal data has not been studied extensively, except from the angle of privacy concerns (Laudon 1996, Bélanger and Crossler 2011, Li and Raghunathan 2014) with a focus on organisations as data owners. Authors in (Li and Raghunathan 2014) adopted an economics-based approach which addresses the issue of disseminating sensitive data to a third party data user (Bataineh, Mizouni et al. 2016).

## 4. Problem statement and Research question

Logistics companies possess large amounts of data because they face the challenge to understand their complicated market better (Qabbaah, Sammour et al. 2019). This research focuses on the possibility of monetising the large databases by logistics companies gathered on the shipments of their partner companies that are mostly e-commerce companies. The economical relevance of such an approach is that this may allow e-commerce companies to allocate some of their limited resources to manage their data more effectively.

Knowing how logistics companies that possess such huge data bases can effectively offer the service of improving the knowledge of their business partners with respect to their business and customers is an interesting field of research. In other words, can they really market their data to these partners? For this purpose the data have to be instrumental in nature (giving customers a better idea about some business questions). Visual analytics using statistics and charts can certainly be used to describe the data and to understand the

market of logistics companies better. The lack of previous research in this field makes this effort very important.

The contribution of this paper is consequently to investigate whether value can be derived in an international context from extensively looking into the monetisation possibility of specific logistics data of e-commerce companies (a field and combination that has not been studied before). The following research question will be addressed **'What is the possibility of data to be monetised in the logistics sector?'**

## 5. Methodology

In order to answer our research question, we start from a real life dataset. Different methodological strategies were applied. First we will develop in depth statistics and visualization charts on the data aimed at showing in which way they could help in getting a clearer understanding about the dataset. We will use visual analytics in doing so. The most important example of this is e-companies' market analyses. The purpose is to see whether this could enable logistics partner companies to understand their competitiveness in more depth and help them in improving their image and sales volume by looking at creative ways to develop new unique selling propositions.

Second, we will use K-means clustering on the data for segmentation of transactions purposes. This is useful for the e-commerce companies as well as for the customers since the clustering results show the prevalent logistics variable combinations.

Third, we will create a semi-structured interview to be held with the senior decision makers of the partner companies and of the logistics company involved to get their opinion about the data monetisation concept. We will confront the interviewees with some results to find out their reaction and eventually some evidence of the monetisability of our data.

In the next paragraphs we will explain these methodologies.

Data visualization involves "Presenting data in graphical or pictorial form which makes the information easy to understand. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information" (Sadiku, Shadare et al. 2016). Whereas, visual analytics is defined as "The science of analytical reasoning facilitated by visual representations used within a personal context" (Huang, Tory et al. 2015). Therefore, visual analytics goes one step further than data visualization. It can be considered as an integral approach combining visualization and data analysis. Visual analytics integrates data visualization methodology with information analytics, geospatial analytics, and scientific analytics. Therefore, visual analytics benefits from methodologies developed in the fields of statistical analytics, data management, knowledge representation, and knowledge discovery (Pak Chung Wong 2004). It is not likely to become a separate field of study, but its influence will spread over the research areas it comprises

K-means clustering is used often for segmentation purposes. This method offers a real-time solution for the development of distributed interactive decision supports since it allows the consumer to model his/her preferences along multiple dimensions, such as product information and logistics route and then produces data clusters of the products-logistics combinations retrieved to enhance marketing decisions (Papamichail and Papamichail 2007).

The main objective of this algorithm is to partition the dataset into k clusters in which each instance belongs to the cluster with the nearest mean. It is suitable for large datasets and offers ease of implementation and high speed performance (Carmona, Ramírez-Gallego et al. 2012). The K-Means algorithm starts from k central point's chosen randomly. Every instance is assigned to the closest central point. Next, the heuristic performs a reassignment of the central points. The algorithm finally stop when the assignments of the individual instances no longer change (Kotu and Deshpande 2019). The algorithm follows five steps.

The first step initiates k random centroids. The number of clusters k should be specified by the user.

Step 2 consists in assigning data points. Once centroids have been initiated, all the data points are assigned to the nearest centroid to form a cluster. In this context the 'Nearest' is calculated by a proximity measure. The Euclidean distance measurement is the most common proximity measure used in this respect. The Euclidean distance between two data points X (x1, x2,…xn) and C (c1, c2,…cn) with n attributes is given by:

$$\text{Distance d} = \sqrt{(\text{x1} - \text{c1})^2 + (\text{x2} - \text{c2})^2 + \cdots + (\text{xn} - \text{cn})^2} \qquad (1)$$

In a third step new centroids are calculated. For each cluster, a new centroid is calculated, which is also the prototype of each cluster group. This new centroid is the most representative data point of all data points in the cluster. Mathematically, this step can be expressed as minimizing the sum of squared errors (SSE) of all data points in a cluster to the centroid of the cluster. The overall objective of the step is to minimize the sum of squared errors of individual clusters. The SSE of a cluster can be calculated by the following equation:

$$\text{SSE} = \sum_{\text{i=1}}^{\text{k}} \sum_{\text{xj} \in \text{ci}}^{\text{k}} ||\text{xj} - \mu\text{i}||^2 \qquad (2)$$

where Ci is the ith cluster, j are the data points in a given cluster, μi is the centroid for ith cluster, and xj is a specific data object. The centroid with minimal SSE for the given cluster i is the new mean of the cluster. The mean of the cluster can be calculated by:

$$\mu\text{i} = \frac{1}{\text{ji}} \sum_{\text{x} \in \text{ci}}^{\text{k}} \text{X} \qquad (3)$$

where X is the data object vector (x1, x2,… xn).

Step 4 is a repeated assignment and calculation of new centroids. Once the new centroids have been identified, assigning data points to the nearest centroid is repeated until all the data points are reassigned to new centroids.

Step 3 and step 4 are iterative until no further change in assignment of data points happens or, in other words, no significant change in centroids is noted anymore. The final centroids are declared the prototype data objects or vectors and they are used to describe the whole clustering model. Each data point in the dataset is now tied with a new clustering ID attribute that identifies the cluster.

Since monetising data is relatively new so that data on the revenue generated and the marketing for it do not exist readily in the logistics sector, we finally wanted to investigate the potential for monetisation. We used a semi-structured interview technique to develop some insight into this point. We opted for this research methodology as it can provide qualitative data, which in this case is more relevant than quantitative ones.

Our enquiry was subdivided in two parts, one for e-companies and one for the logistics provider (which possesses the database). In our interviews we spoke with the manager responsible for the transactions and relationship management with the logistics company. The e-companies were selected by the logistics company on the basis of some criteria: being situated in Jordan, important as a partner and using the logistics services for a number of different items or product categories. These criteria were selected because they offered a variety of different situations in which these companies are operating. For reasons of privacy, their names and activity fields are not disclosed. Our questions were based on previous research (Laitila 2017, Derwisch 2019). They reflect the reality of the data used by the e-companies and the use they can make of it.

We have to stress that this selection is not without danger. The relationship with the logistics company of the managers of the e-companies and the involvement of the logistics company in selecting the interviewees may influence the results in a positive way. Nevertheless, we believe that this bias is limited as we did not mention to the interviewed managers of the e-companies that they were selected because of their relationship with the logistics company and the

questions were phrased in a more general way than to mention logistics databases.

## 6. Data

Logistics companies that ship products sold online have a huge amount of detailed market basket data available. They contain information on sets of items that buyers acquire, whether the items are bought together or at different times, the physical characteristics of the products they buy, the type of products they like to buy, the suppliers they like to buy from, the payment mode they use, the logistics route used (country of origin and country of destination), the brands they select and price levels that trigger their buying and so on. Moreover, the online buying behaviour of the customers is also registered in the data base. They are however rarely taken into account in a combined product/ customer categorisation effort.

The data used in this research were obtained from a logistics services company situated in the Middle East. Cleaning, merging tables and pre-processing of the data were applied in order to obtain the final data set. The total number of transactions in the final dataset is equal to the size of the sample (n=85959). Table 1 below shows the variables, the type of data they represent and the description of each of the variables used in our research.

Table 1. Data Descriptions

| **Variable** | **Data type** | **Data Description** |
|---|---|---|
| **Variables in the original dataset** | | |
| **ID** | Integer | The ID of the order |
| **Weight** | Double | The weight of shipment |

| | | |
|---|---|---|
| **CODValueUSD** | Double | The amount of cash on delivery |
| **Payment** | String | Type of payment. P: prepaid, C:cash, 3:third party, F:free |
| **Destination** | String | Destination of shipment |
| **Origin Country** | String | Country of origin of the shipment |
| **DestCountry** | String | Country of destination of the shipment |
| **ShipperID** | Integer | The ID of the E-commerce companies |
| **CODFlag** | Boolean | Cash on delivery flag |
| **'Consignee Tel'** | Integer | The telephone number of the customers |
| **Added variables** | | |
| **Weight In KG** | Double | Total weight in KG |
| **Value USD** | Double | The price of the goods in the shipment in US Dollar |
| **Product Group name** | String | Product group name of the shipment |
| **Product Group ID** | Integer | Product group ID |
| **Customs class** | Boolean | If the shipment subject to customs or not |

Tableau Software' was used to visualize the data. We wanted to find the distributions of transactions in our data on the bases of cities and countries of shipments, how the product categories and returned products distributed on the basis on countries, customers, e-commerce companies and customs class.

Therefore, visualizing the different attributes dimensions, such as location, products, customs class, customers and e-commerce companies was necessary to prepare the data in a format that might be monetisable.

## 7. Visual analytics results

Table 2 represents the summary of the data we will visualize in this section. It indicates the number of the transactions according to the dimensions (variables) mentioned.

Table 2: The visualization of the different dimensions in this section.

| Number of transactions according to | Dimensions (variables) |
|---|---|
| Destination countries, origin countries and destination cities. | Country (origin, destination),<br>Destination cities. |
| Products transferred to the country of destination | Products←→ Destination countries |
| Products transferred to the city of destination | Products ←→ Destination cities |
| Products transferred from country of origin | Products←→ Origin countries |
| E-commerce companies have orders transferred to destination countries | e-commerce companies←→ Destination countries |
| Customers have orders transferred to | Customers←→ Destination |

| | |
|---|---|
| destination countries | countries |
| Product categories shipped by the customers | Customers ←→ Products |
| Product categories transferred to destination countries from the countries of origin. | Origin countries→ Destination countries→ Products |
| Retuned orders by the country of destination | Return products ←→ Destination countries |
| Retuned orders by the city of destination | Return products ←→ Destination cities |
| Retuned orders by the e-commerce companies | Return products ←→ E-commerce companies |
| Retuned orders by the customers | Return products←→ Customers |
| Customs class by country of origin | Origin countries → Customs |
| Customs class by country of destination | Destination countries → Customs |
| Customs class by Product categories | Products→ Customs |
| Customs class by product categories transferred from country of origin to destination country | Destination countries→ Origin countries→Products→Customs |

We will show few examples from Table 2 in the following Figures.

Figure 1 presents the top 10 products transferred to all destination countries together. We can see that 'Apparel' has the highest percentage with 32%, followed by 'Beauty supplies' and 'watches' with 7% and 5%, respectively. Companies can learn from this figure what are the most products ordered by the customers in the our destination countries.

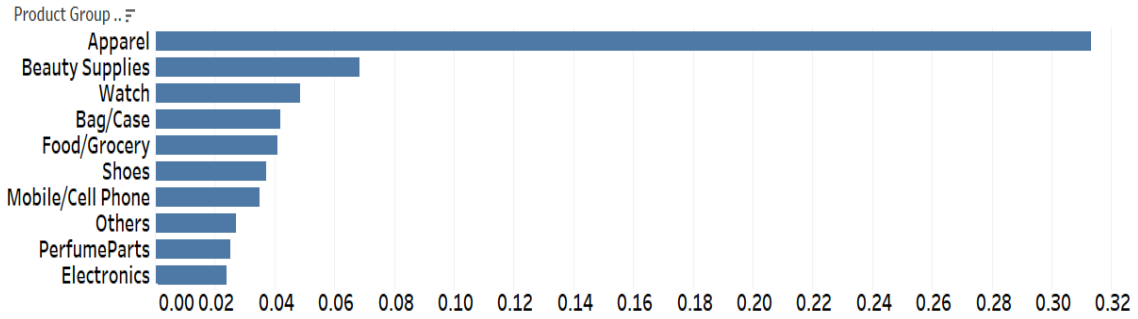


Figure 1: Top 10 products shipped.

Figure 2 presents the percentages of the transactions according to the product categories transferred to Jordan for the top four countries of origin. 'Apparel' is mostly transferred from the 'US' to Jordan (29%), 'Beauty supplies' with 9% is also transferred most frequently from 'US', while shipments from 'GB' are mostly 'Apparel' and 'Home supplies' with 4% and 2%, respectively. Companies can learn from this figure that the customers in Jordan prefer to order their products such as 'Apparel' and ' Beauty supplies' from 'US'.

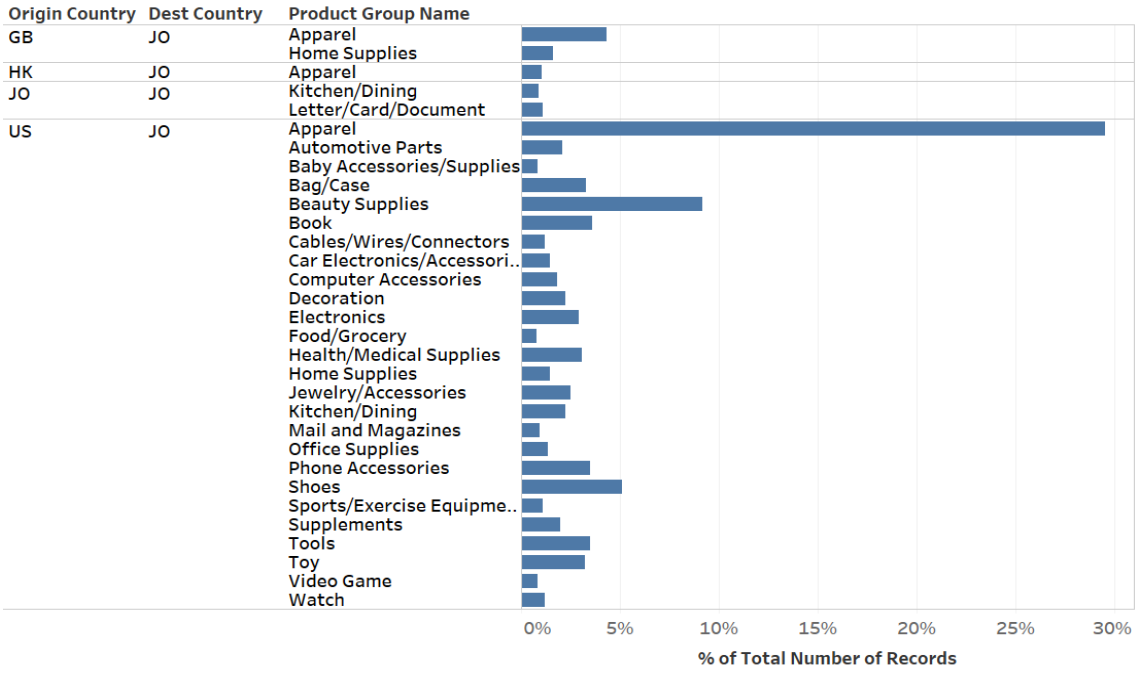


Figure 2: Product categories shipped to Jordan according to the top four countries of origin (expressed in percentages).

Figure 3 shows the percentages of the transactions according to the returned orders by the country of destination and city of destination. 'Jeddah' and 'Riyadh' in Saudi Arabia represent 35% and 34%, respectively, while 'Dubai' represents more returned orders than 'Abu Dhabi' in the 'UAE'.

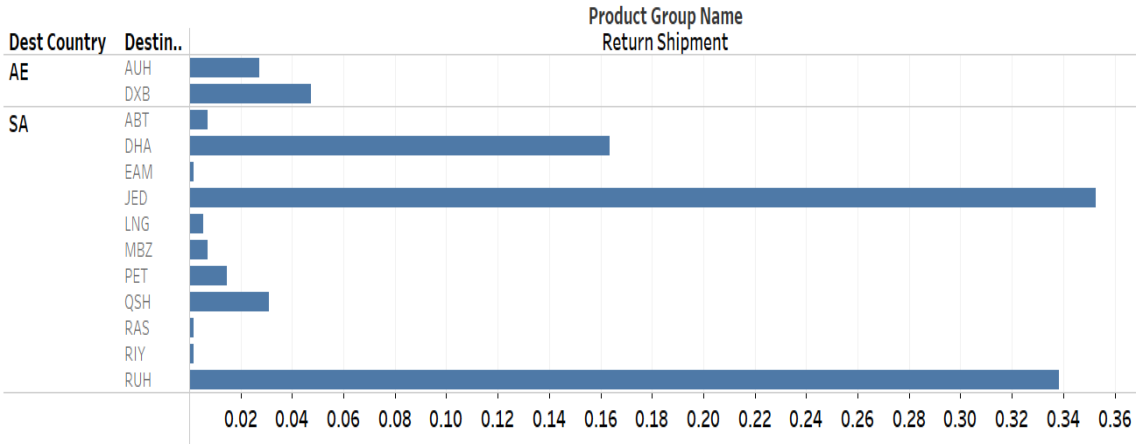


Figure 3: The returned orders according to the city of destination.

Customs variable has been transformed using the 'Customs Value' variable. The variable contains two classes: 'Yes': if the order is subject to customs tariffs and 'No': if the order is not subject to customs tariffs.

Figure 4 shows the percentages of the transactions according to the customs (Class 'Yes') for the most frequently send product categories to Jordan for several countries of origin. Companies can see that 'Apparel' and 'Beauty supplies' transferred from the 'US' to Jordan have the highest percentage of customs application class with 27% and 9%, respectively, while 'Apparel' has the highest percentage of customs transferred from the 'UK' with 4%.

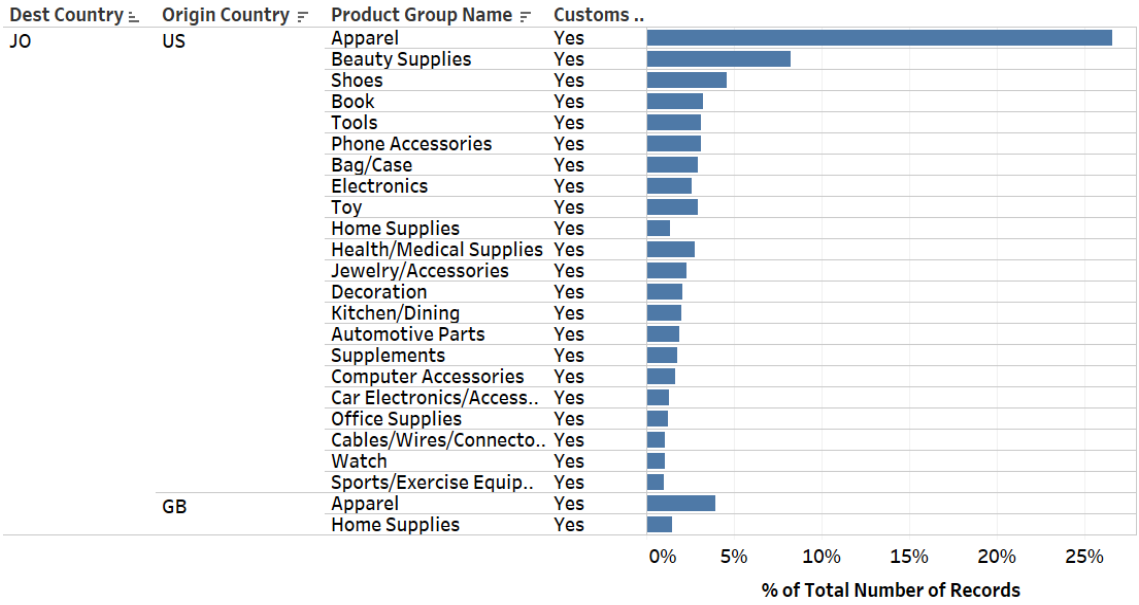


Figure 4: Customs (class 'Yes') according to the most product categories transferred to Jordan (expressed in percentages).

## 8. E-commerce companies market analyses and results

The visual analytics of our data can fully be understood when the weight of the different variables (products, countries and customers) relative to the total number of transactions is clearly mentioned. This is much like saying that the importance of a product in a certain market can be grasped when compared to

the total market of competitive alternatives. This concept must be visualized as well. Therefore, we have tried to explain the relative weight of the different products-countries-customers in as much detail as possible by clearly describing what have been visualized.

The values were calculated as follows. First, the time period of all the data in the dataset was the same for the companies, products and customers, which makes our comparison accurate on the time level. Second, as an example, we calculate for instance the company's total transactions for the different types of products. Third, we divide the number of transactions of each company by the total number of the product transactions. We also apply this procedure to calculate the countries market for the different products categories.

Some examples are shown below.

Figure 5 presents the destination countries market analysis according to the eight most common products transferred. Section 5A shows that Saudi Arabia has the biggest market for 'Apparel' product with 90.34% following by Jordan and UAE with 6.79% and 2.87%, respectively. Section 5G shows that UAE has bigger market than Jordan for Product 'Letter/card/Document' with 9.57% and 5.54%, respectively, but still Saudi Arabia has the highest percentages with 84.9%. and so on.

Figure 6 presents the e-commerce companies relative importance on the market on the basis of the products transactions for one of the destination countries, namely Jordan.

E-company '15037' has the highest market importance for 'Apparel', 'Bag/Case', 'Beauty supplies', 'Book', 'Food/Grocery', 'Jewellery Accessories' and 'shoes' with 69%, 86%, 92%, 82%, 77%, 85% and 79%, respectively. Whereas e-company '197483' has the highest market importance for 'letter/ card/ document' product with 40%. Companies can learn from this figure how the results of the market importance of the products are distributed between the top five e-companies in Jordan.

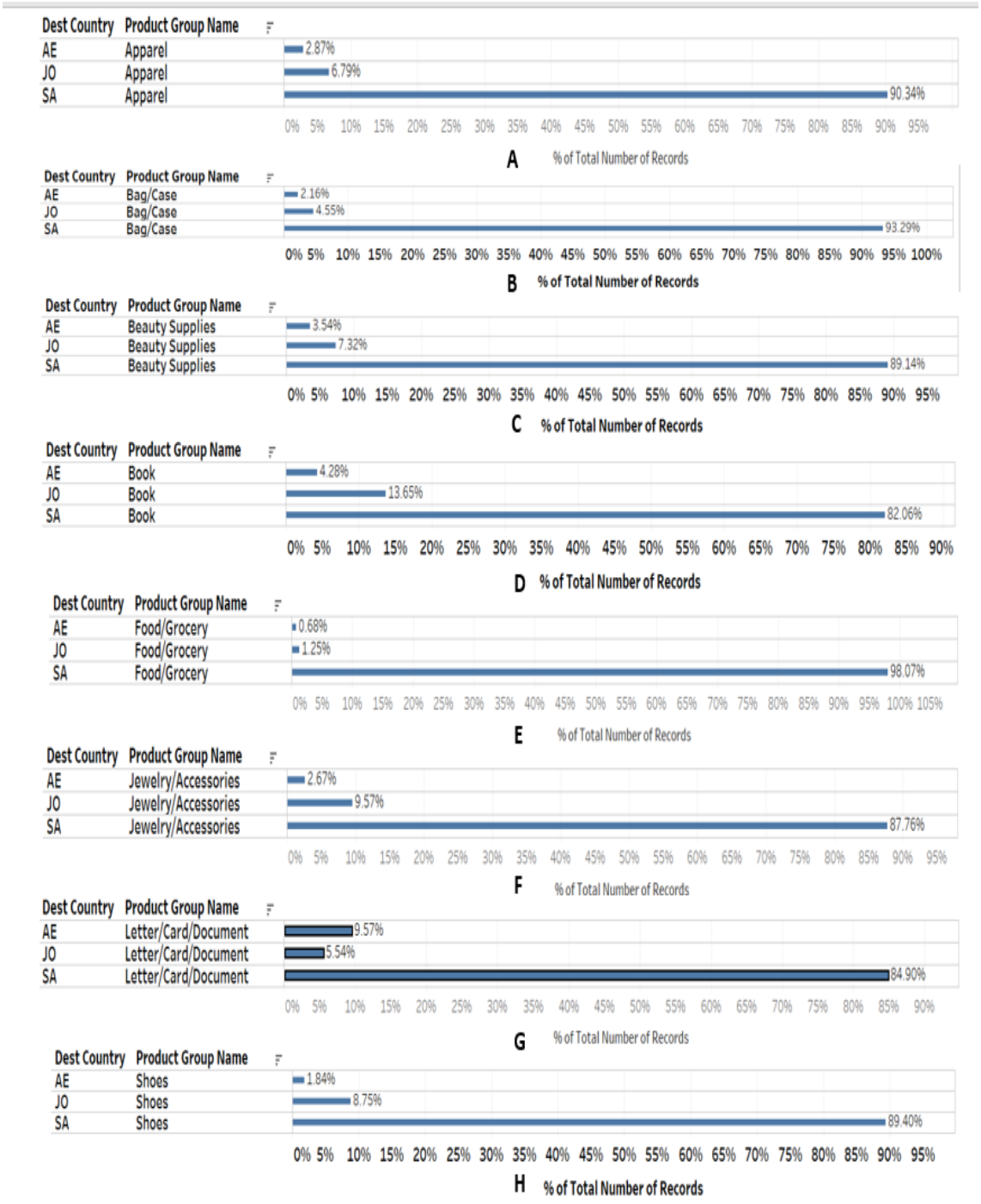


Figure 5: Countries of destination market according of the eight products most commonly ordered (expressed in percentages).

These results could provide e-companies with a clear vision on their position in the market according to products and locations. E-companies can identify and

know their competitors in the market in general or on the basis of some products particularly.

A

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 15037 | Apparel | 69.34% |
| | 64450 | Apparel | 16.50% |
| | 565631 | Apparel | 7.92% |
| | 546181 | Apparel | 1.42% |
| | 565363 | Apparel | 1.20% |

B

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 15037 | Bag/Case | 85.98% |
| | 565631 | Bag/Case | 7.32% |
| | 64450 | Bag/Case | 4.27% |
| | 503568 | Bag/Case | 1.22% |
| | 391989 | Bag/Case | 0.61% |

C

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 15037 | Beauty Supplies | 92.09% |
| | 64450 | Beauty Supplies | 5.58% |
| | 502065 | Beauty Supplies | 0.70% |
| | 580389 | Beauty Supplies | 0.47% |
| | 599148 | Beauty Supplies | 0.47% |

D

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 15037 | Book | 81.91% |
| | 64450 | Book | 16.49% |
| | 366180 | Book | 0.53% |
| | 502065 | Book | 0.53% |
| | 548247 | Book | 0.53% |

E

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 15037 | Food/Grocery | 77.27% |
| | 64450 | Food/Grocery | 20.45% |
| | 586537 | Food/Grocery | 2.27% |

F

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 15037 | Jewelry/Accessories | 84.50% |
| | 64450 | Jewelry/Accessories | 12.40% |
| | 601479 | Jewelry/Accessories | 1.55% |
| | 557424 | Jewelry/Accessories | 0.78% |
| | 565363 | Jewelry/Accessories | 0.78% |

G

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 197483 | Letter/Card/Document | 39.77% |
| | 362662 | Letter/Card/Document | 13.64% |
| | 64450 | Letter/Card/Document | 11.36% |
| | 15037 | Letter/Card/Document | 7.95% |
| | 594857 | Letter/Card/Document | 6.82% |

H

| Dest Country | Shipper ID | Product Group Name | % of Total Number of Records |
|---|---|---|---|
| JO | 15037 | Shoes | 78.93% |
| | 64450 | Shoes | 14.29% |
| | 565631 | Shoes | 2.50% |
| | 565363 | Shoes | 2.14% |
| | 585583 | Shoes | 1.07% |

Figure 6: E-commerce companies market importance on the basis of the products transactions for Jordan (expressed in percentages).

## 9. Clustering Results

The clustering model focuses on clustering the logistics transactions on the basis of logistics variables such as product type purchased, customs classes, weight of the shipment, value of goods bought and location of shipments from origin countries to destination cities. This segmentation shows the prevalent logistics variable combinations. These data have shown to be of particular interest to the e-commerce partners and could be monetisable since an improved focus on customers interested in these combinations is possible. They can be served better.

In order to find the best cluster fit experiments we analysed the data for 2 to 5 clusters. We used ANOVA and the Calinski-Harabasz criterion to assess the clustering quality. Calinski-Harabasz has the option that if a user does not specify the number of clusters, the number of clusters will be picked corresponding to the first local maximum of the Calinski-Harabasz index automatically. Then the best fit in each of these models was selected by comparing the clusters results from k=2 till k=5 taking into consideration the evaluation results.

The variables used in our model are 'Value USD', 'Weight In KG', 'Product Group Name', 'Customs', 'Country of origin', 'Country of Destination', 'Destination' and 'Payment type'. The data include all three countries of destination (Jordan, UAE and Saudi Arabia) (N=85,959).

Table 3 shows the results of the 5-cluster solution. The customs class in clusters-1,2 and 4 has been identified as 'Subject to customs' while the customs class in cluster-3 and 5 is 'Not subject to customs'. Both clusters-1 and 2 have the 'Apparel' product as the most frequently shipped product to 'Riyadh' in Saudi Arabia, while the most frequently transferred products in clusters 3, 4 and 5 are 'Food/Grocery', 'Mobile/Cell Phone' and 'Letter/Card/Document'. These transactions are shipped respectively to 'Jeddah' and 'Amman'. The most common transactions in cluster-3 and 5 with a customs class 'Not subject to customs' are shipped within Saudi Arabia and Jordan. Cluster-3 has the highest average of the weight with 9.1 KG, and cluster-2 has the highest average of the price of the shipment with 156 dollar.

Table 3: The results of the 5-cluster solution.

| | **Attributes/ Clusters** | **Cluster1** | **Cluster2** | **Cluster3** | **Cluster4** | **Cluster5** |
|---|---|---|---|---|---|---|
| | Number of Items | 39882 | 25073 | 7731 | 13010 | 262 |
| | Avg. Weight In KG | 1.3013 | 3.8994 | 9.0847 | 1.6742 | 1.2353 |
| | Avg. Value USD | 75.161 | 155.5 | 107.98 | 110.49 | 37.12 |
| | Customs | 1 | 1 | 0 | 1 | 0 |
| **Most Common** | Product Group Name | Apparel | Apparel | Food/ Grocery | Mobile/Cell Phone | Letter/Card/ Document |
| | Country of Destination | SA | SA | SA | SA | JO |
| | Destination | RUH | RUH | JED | JED | AMM |
| | Payment | C | P | P | P | F |
| | Country of origin | US | HK | SA | SA | JO |

Notes: RUH: Riyadh, JED: Jeddah, AMM: Amman. SA: Saudi Arabia, US: USA, HK: Hong Kong, JO: Jordan. Customs class: 1- Subject to customs. 0- Not subject to customs.

Payment: C: cash, P:prepaid, F:Free

Table 4 shows the results of the analysis of the variance (ANOVA) test for our clustering Model.

Table 4: The results of the analysis of the variance test

| Number of clusters | Variable | F-statistic | P-value |
|---|---|---|---|
| **2-clusters** | Avg. Weight In KG | 1689 | 0.000 |
| | Avg. Value USD | 698.3 | 0.000 |
| **3-clusters** | Avg. Weight In KG | 601.9 | 0.000 |
| | Avg. Value USD | 445.4 | 0.000 |
| **4-clusters** | Avg. Weight In KG | 1503 | 0.000 |
| | Avg. Value USD | 423 | 0.000 |
| **5-clusters** | Avg. Weight In KG | 1169 | 0.000 |
| | Avg. Value USD | 321.3 | 0.000 |

The results of the analysis of variance (ANOVA) of the different cluster solutions show that the p-value is <0.001 for the continuous variable 'Avg.Total Value USD' and 'Avg. Weight In KG'. The results of the Calinski-Harabasz test indicate that k=5 is the best cluster fit. We can see that the number of items in cluster-5 is lower than in the other four clusters. Cluster-5 is the only cluster showing 'Jordan' as a country of destination while the other four clusters have 'Saudi Arabia' as a country of destination. Moreover, in cluster-5, the most frequently product shipped is 'Letter/Card/Document', which is different from the other clusters.

The model has two clusters with a customs class 'Not subject to customs' and three clusters with a customs class 'Subject to customs'. The model has clusters with three different product groups, three different destination cites, three different payment methods and four different country of origins.

The following two figures show the percentages of the highest number of transactions of the 5-cluster solution in detail, four variables are shown in these Figures. The x-axis contains the total number of transactions, whereas the y-axis contains the following shipment variables: customs class, product categories, country of origin and destination city, while the percentages of the total number of transactions are presented per y-axis variables for the cluster group. We will show the results of two clusters only. Cluster-1 with the highest number of orders subject to customs and cluster-3 with the highest number of orders not subject to customs. From Figure 7 companies can learn that in cluster-1 the highest percentages of the transactions were shipped from 'US' to 'RUH', 'JED' and 'AMM', respectively for the category 'Apparel' with a custom class 'Subject to customs'.

Companies can learn from Figure 8 that in cluster-3 the highest percentages of the transactions are transferred within 'SA' for the category 'Food/Grocery', and the customs application class was 'Not subject to customs'.

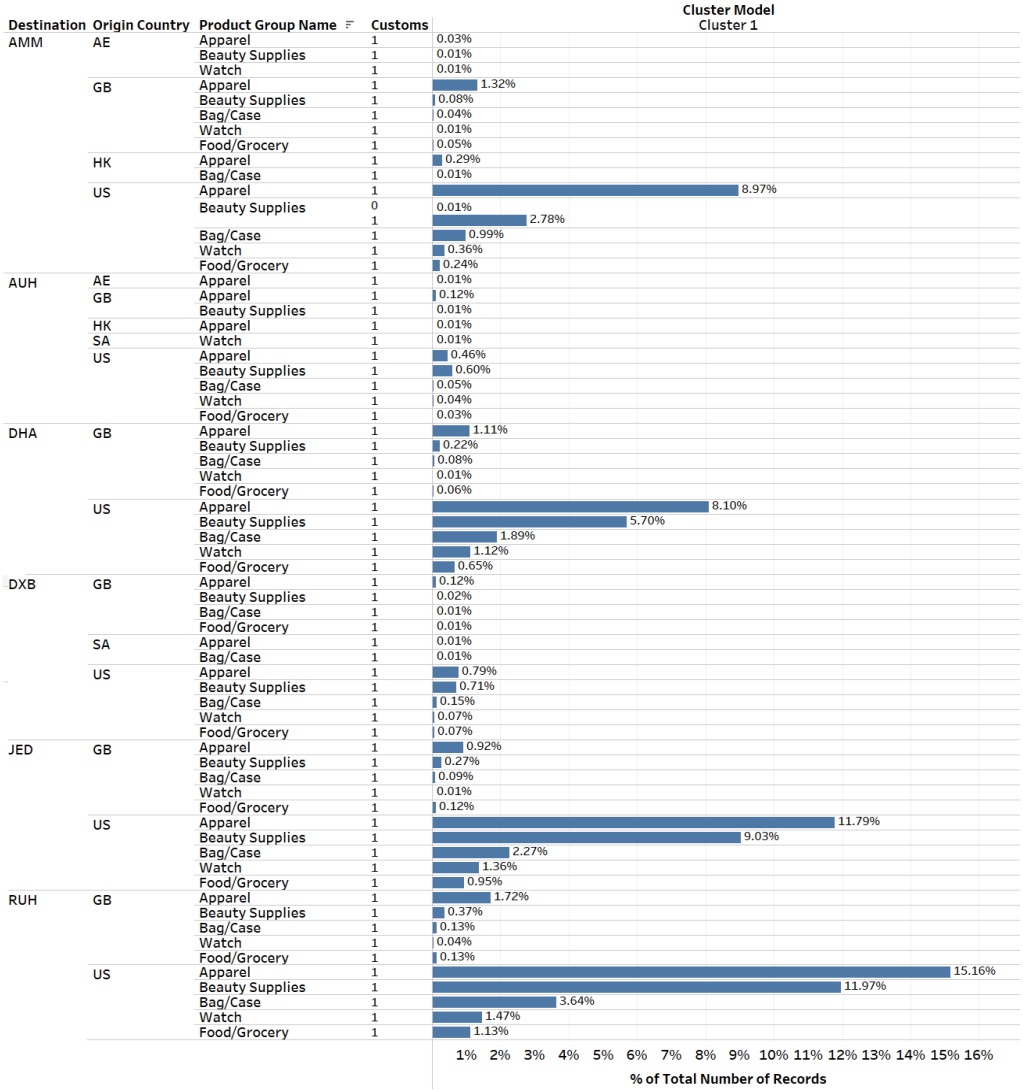


Figure 7: The distribution of the highest number of transactions in cluster-1 in detail of the 5-cluster solution expressed in percentages.

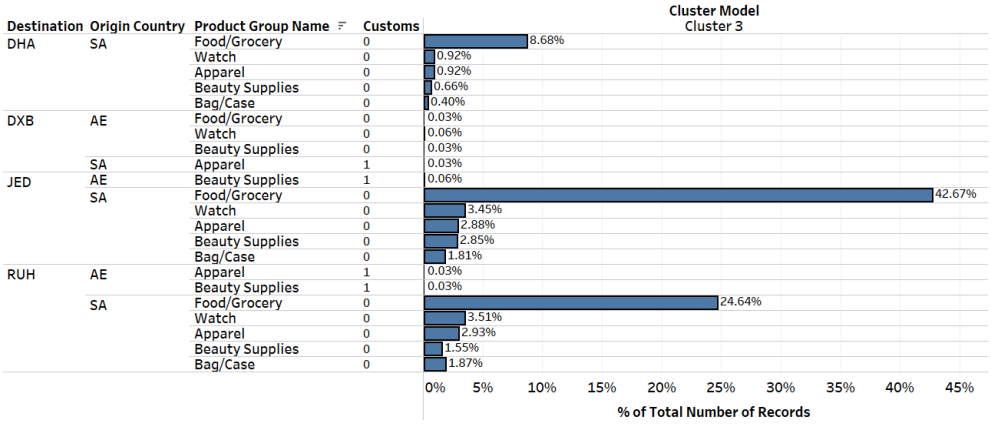


Figure 8: The distribution of the highest number of transactions in cluster-3 in detail of the 5-cluster solution expressed in percentages.

## 10. Results of the Semi-structured interviews

Appendix A and B respectively enumerate the questions asked during the semi-structured interviews with the key managers of both the e-company partners of our logistics company and with the manager of the logistics company.

### 10.1 Results for the E-companies interviews

In these companies we summarised the answers of the four companies involved in the semi-structured interviews.

Market data are of extreme importance to each of these companies in order to know their customers and the market better and to assess their market position compared to competitors in a more detailed way (more specifically with regards to market share (company 2) and building a customer data base (company 4). All four managers expect market data to help them in developing and offering customer centric solutions (company 1) and complementary services (company 2 and 4). The hope is that this will be beneficial for both their company (company 3 and 4) and customers (company 1 and 3).

Since the managers told us that they actually only possess customer characteristic data (companies 2,3 and 4) and transaction data (all companies), a lot of data are lacking. This starts with data on market shares (company 1) and customer preferences on products and sources (company 2 and 3) and ends with bases for customer clustering and segmentation like income, location and physiological trends (company 4). In general managers thus lack deeper insight in and complementary data on customers. The simple data they actually possess are clearly insufficient.

These data are mainly gathered via transactional data. Two companies possess a specific department or some people responsible for this. One company (company 4) gathers the data in a more creative way, namely by asking customers to upload their data on an online blog, posted on their website and actually attached a customer referral marketing campaign to it by adding a gift when new contacts were suggested.

However, companies 1,2 and 4 also purchase some customer contact data from third party sources, but never data on preferences and e-company sources. This has some advantages: it helps them in starting new campaigns (company 1 and 2) or in developing customer profiles (in the hope of segmenting more efficiently – company 4). Selection of these third parties is done on the basis of reliability (quality, delivery time, service – company 4), integrity (company 1 and 4), security (company 1 and 4).

Later in the enquiry the same characteristics were also mentioned for the data companies would eventually be willing to buy and for the companies they would buy them from. On this last element, also the fact that the third party should never be a competitor (company 3) was explicitly mentioned. Three of the four companies are indeed interested in getting the lacking insights in their market, customers and trends by buying the information from a third party. Only company 4 indicated that investing in market research also offers advantages over buying data from third party sources. It helps better in getting very specific information and learn from the experience by being capable of understanding the market needs better. If data would have to be bought from third party sources, their format should be capable of offering additional insights easily, that is in tabular of graphical format, dashboards being explicitly mentioned by (company 2).

This overview tells us that in general data are very valuable for the e-companies in developing a more effective and efficient marketing. Moreover, they are willing to buy the lacking data from third party sources taking into consideration the characteristics of the data mentioned in the literature are fulfilled. This is a first indication that data monetisation on logistics data could be very well possible. With regards to the type of data monetisation, selling is not the only way in which the logistics data could be monetised as our respondents clearly mentioned that deeper insights, generated through visualization and further analysis via eventual dashboards would be very useful.

The last question wanted to test whether the data generated from the actual data base we are using through visualization and further analysis would be extremely useful for our respondents. We presented some of our results as data, our respondents could potentially buy. They all four were very excited by

the possibility to buy these data. They were interested in all formats we offered them: customer preferences, payment, customs service data, market and competitor data and clustering results. It proves that carefully managed and presented data are monetisable with these respondents in an actual test environment, maybe even more than the general answers before mentioned. All respondents immediately came up with the benefits they could make from all these data, such as initiating campaigns (companies 1,2,3 and 4), predicting customer reactions to products (companies 2, 3 and 4), predicting customs application (companies 1, 2 and 3), focusing on areas competitors (companies 1,2 and 3), detailed market segmentation and clustering (companies 1,3 and 4) and offering complementary services (companies 1, 2 and 3). This is a sign of the value of the data base the logistics company actually possesses and of the necessity to visualize and analyse them deeper before monetising their data. They could from this small experiment be considered as a source of income.

Six months after we conducted our enquiry, we contacted the e-commerce companies again to check whether they had continued talking to the logistics company about the data. Three of them confirmed having bought the data, with the fourth one still in the process of negotiating about it. This proves the monetisability of the data, although we could not yet verify which value the data exactly represented as we did not get access to their financial statements, which would have clarified the impact of the use of the data on their marketing and sales efforts.

## 10.2 Results for the logistics company interview

Contrary to the enquiry with the e-companies, we showed our respondent in the logistics company the data they possess in the format we developed by visualizing and analysing them.

Our respondent told us that actually no data at all are used by, shared with or sold to partners or customers at the moment, although many of their partners indicated to be interested in them, particularly in the data on the number of returns based on a successful 'COD' or not, and most importantly on detailed

customs values. Therefore, our respondent really believes that making some of these data available to their partners and making them easy to be accessible, would activate business and improve the quality of their services.

It would offer the company new market opportunities and gain them "easy money". Engaging in making these data available and selling them would interest customers as most of the data, in the eyes of our respondent, would cover missing market information for their customers/partners. It would be a win-win situation for both parties as it would help the partners in solving unaddressed issues. Our respondent considers this to be a realistic extra service to be offered given the quality of the data and the absent availability on the partners' side. Future services could also include investing in solutions for the challenge of the rising number of returns.

The logistics company also thinks the data we prepared for them would also be extremely valuable to their own business and can be monetised. The reasons are more specifically:

1- The customer data contain detailed information about the purchases and the commodity of each purchase. This would be of huge help in initiating campaigns to target customers based on their interests.
2- Preferred payment info and detailed customs values per destination would, if shared within the data, give the company estimates and forecasts about how much returns can be expected, and why certain destinations are having drops in volumes.
3- Data about other e-commerce competitors will give the company insight into the current market situation. It would indicate whether they are doing well compared to competitors. This would let them analyse their performance better and signal improvement areas.
4- The detailed clustering results were considered to be the most interesting ones. They could give great flexibility in visualizing the data and manipulating needed dimensions as per our interests.

Given these answers, we can conclude that the logistics company indeed sees the opportunity of data monetisation with the data base they possess, but more specifically if these data would be transformed through deeper analysis and visualization so as to make it easier for them and for customers/ partners to gain deeper insight. They consequently realise that probably the category of

offering services which provide deeper knowledge is more interesting than just selling the data in terms of the categorisation theory mentioned before. Therefore, the second and third data monetisation techniques are preferred to monetise the data. They are: Providing insights or analyses and creating platform for new services. This is very congruent with our conclusions for the e-companies.

Our respondent developed a business model for monetising the data using the business model canvas proposed by Osterwalder et al. (Osterwalder, Pigneur et al. 2010). Table 5 shows the results.

From the business model canvas we can read that the data monetisation efforts will provide the company with the possibility to offer improved and enlarged service packages to their partners. This will help them in increasing sales and improving their market share as their knowledge of the market will be enhanced. Our logistics company mentioned that from the cost perspective, the company already has R&D and Marketing departments. They can treat this new market easily. Consequently, a monetisation project will not cost much since it does not need special or new IT infrastructure. The revenue is based on sharing the data with their partners as an extra service. This revenue could be collected directly.

The conditions relating to the process in the company mentioned before can be drawn from the business model canvas. For instance, connecting and sharing the data with the e-companies and partners accomplished using the key activities and channels segments. This is relevance to the technology factor. Whereas, the selling mode chosen by the company is double: it contains both providing insights or analyses and creating a platform for new services. This is related to both business and execution factors.

Table 5: Data monetisation business model canvas.

<table>
<tr><th colspan="5">Business Model Canvas</th></tr>
<tr><td>Key Partners</td><td>Key Activities</td><td>Value Propositions</td><td>Customer Relationships</td><td>Customer Segments</td></tr>
<tr><td rowspan="3">E-commerce companies / partners</td><td>R&D (Research and Development), building customers database on the base of products interest</td><td rowspan="3">Data monetisation, improved marketing activities, increased sales and market share</td><td>Analytics to help partners. Trust in the data integrity and security. Partnership to serve one goal</td><td rowspan="3">Partners, e-commerce companies</td></tr>
<tr><td>Key Resources</td><td>Channels</td></tr>
<tr><td>Employees, Data, IT infrastructure</td><td>Visits, meetings, marketing campaigns, conference calls</td></tr>
<tr><td colspan="3">Cost Structure</td><td colspan="2">Revenue Structure</td></tr>
<tr><td colspan="3">R&D costs, Marketing costs</td><td colspan="2">Service sales, Data share revenue from partners</td></tr>
</table>

## 11. Discussion and Conclusion

Visual analytics gives the logistic companies a clearer image allowing them to understand their market in more detail and adapt their marketing accordingly. All customers can be grouped in different ways, such as by 'Products', 'E-companies', 'Destination cities' and so on. The results vary according to the different variables used in combination with each other which is proven by our results of the market analyses. Moreover, logistics companies would benefit from the results of the e-companies' market analyses and better understand the competitiveness of the market of e-companies, eventually looking at ways to develop a more unique selling proposition (for instance via e-cards for special occasions or document safety services) as this seems to be a highly competitive market. It can therefore improve the market image and sales volume of the e-companies.

Every e-commerce company or customer can therefore indeed be assessed by logistics companies according to the results of the visualization. Knowing the distribution pattern of the shipments according to 'Product types', 'Customers', 'Cities' and so on is indeed highly valuable. It can direct the attention of the logistic companies to specific patterns which allows them to better target their marketing efforts. This type of information is highly significant to all participating partners.

Visual analytics, while being highly useful to manage the transactions, is however only a first step in answering our research question. For instance, it could help the logistics companies to monetise their data and selling it to other e-commerce companies.

To be able to potential monetise the data, the data should not only be well presented, but also adhere to the characteristics of monetisability. Are these two elements present?

With regards to these four criteria of data quality we can be fairly sure that they are present. The data are reliable and accurate since they are taken from the real dataset of logistics transactions by the logistics company. They are relevant to the customer companies as the e-companies are their partners and are all situated in the same sector and region. The encryption of the name of the

customers and e-companies makes the data anonymous, moreover, the results are well segmented, as the clustering results show. Therefore, all four criteria for data quality previously mentioned are fulfilled.

We can indeed say that the data are clearly ordered and presented in such a way that some of the data have become interesting to many partners.

In order to find out whether the company has the right mind-set to really market the data, the question was investigated by looking deeper into the business model the company could use in doing so by having semi-structured interviews with some of the partner's managers involved and with a manager of the logistics company. The results prove that the partner companies consider the data to be valuable enough to be invested into and the logistics company confirms this idea. Moreover, the logistic company agreed with its E-partners on what we proposed, namely that the second and third data monetisation techniques were the best to monetise our data, namely providing insights or analyses and creating a platform for new services.

All this evidence points to the fact that our research question "**What is the possibility of data to be monetised in the logistics sector?"** can be answered positively given the right visual analytics and business model used. The fact that three of the four contacted companies have since our investigation bought the data proves the value of them and that they are monetisable.

## Bibliography

Bataineh, A. S., R. Mizouni, M. E. Barachi and J. Bentahar (2016). "Monetizing Personal Data: A Two-Sided Market Approach." Procedia Computer Science 83: 472-479.

Bélanger, F. and R. E. Crossler (2011). "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems." MIS Quarterly 35(4): 1017-1041.

Bonneau, V. (2015). "Data Monetisation: Opportunities beyond OTT: finance, retail, telecom and connected objects." Communications & Strategies 97: 123-126,151.

Carmona, C. J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del Jesus and S. García (2012). "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com." Expert Systems with Applications 39(12): 11243-11249.

Carnelley, G. M. H. S. P. (2018). Update of the European Data Market: Data monetisation. Luxembourg.

Catarina Arnaut, M. P., Elizabeth Scaria, Arnaud Berghmans,Sophie Leconte (2018). Study on data sharing between companies in Europe, everis.

Company, M. (2017). Fueling growth through data monetization

Derwisch, S. (2019). Data Monetization – Use Cases, Implementation and Added Value. Germany, The Business Application Research Center (BARC).

FRED, J. (2017). Data Monetization – How an Organization Can Generate Revenue with Data? , Tampere University of Technology.

Huang, D., M. Tory, B. A. Aseniero, L. Bartram, S. Bateman, S. Carpendale, A. Tang and R. Woodbury (2015). "Personal Visualization and Personal Visual Analytics." IEEE Transactions on Visualization and Computer Graphics 21(3): 420-433.

Kotu, V. and B. Deshpande (2019). Chapter 7 - Clustering. Data Science (Second Edition). V. Kotu and B. Deshpande, Morgan Kaufmann: 221-261.

Laitila, M. (2017). Data monetization: Utilizing data as an asset to generate new revenues for firms, Aalto University.

Laudon, K. C. (1996). "Markets and privacy." Commun. ACM 39(9): 92-104.

Li, X.-B. and S. Raghunathan (2014). "Pricing and disseminating customer data with privacy awareness." Decision support systems 59: 63-73.

Liu, Y. H., Y. Ren and R. Dew (2009). "Monetising user generated content using data mining techniques." AusDM.

Moore, S. (2015). "How to Monetize Your Customer Data."

Najjar, M. and W. Kettinger (2014). "Data Monetization: Lessons from a Retailer's Journey." MIS Quarterly Executive 12.

Osterwalder, A., Y. Pigneur and T. Clark (2010). Business model generation : a handbook for visionaries, game changers, and challengers. Hoboken, NJ, Wiley.

Pak Chung Wong, J. T. (2004). "Visual Analytics." IEEE Computer Graphics and Applications 24(05).

Papamichail, G. P. and D. P. Papamichail (2007). "The k-means range algorithm for personalized data clustering in e-commerce." European Journal of Operational Research 177(3): 1400-1408.

Perrons, R. K. and J. W. Jensen (2015). "Data as an asset: What the oil and gas sector can learn from other industries about "Big Data"." Energy Policy 81: 117-121.

platform, L. s. d. m. (2018). "How to Monetize Your Data." How to Monetize Your Data.

Qabbaah, H., G. Sammour and K. Vanhoof. (2019). Using K-Means Clustering and Data Visualization for Monetizing logistics Data. 2nd International Conference on new Trends in Computing Sciences (ICTCS), IEEE.

Sadiku, M., A. Shadare, S. Musa, C. Akujuobi and R. Perry (2016). "DATA VISUALIZATION." International Journal of Engineering Research and Advanced Technology (IJERAT) 12: 2454-6135.

Serge Findling, M. S., Lynne Schneider, Dan Vesset (2018) "IDC PlanScape: Data Monetization."

Thomas, L. D. W. and A. Leiponen (2016). "Big data commercialization." IEEE Engineering Management Review 44(2): 74-90.

Xu, G., X. Qiu, M. Fang, X. Kou and Y. Yu (2019). "Data-driven operational risk analysis in E-Commerce Logistics." Advanced Engineering Informatics 40: 29-35.

## Appendix A E-commerce companies interview

How valuable are market data for you ?

- How valuable are market data for you?

Which data about your market do you possess and use ?

Which data are lacking ?

Which of the data you use do you gather yourselves ?

- How do you do so ?

Which of the data you use are you currently buying from a third source?

- Which advantages does this have for you ?
- Which criteria do you use to select such a third source ?

Which ways would you like to use to obtain the lacking data ?

- Would you prefer to invest in further research yourselves or be willing to buy these data from a third source ?
- If you would be buying these data from a third source, what characteristics should the data have?
- How should they be presented ?
- Which characteristics should this third source have ?

We are now going to present some types of data that logistics partners could eventually present and offer to you?

please indicate whether you are interested to buy these data, if yes (what makes these data valuable for your business?

## Appendix B Our logistics company interview

We are presenting you some of the data you possess.

Do some of your partners/customers actually use some of your data ?

- Which data are involved ?
- How many partners/customers are interested ? (many, some, only a few)

Do you actually share data with partners/customers?

- Which data are involved ?
- How many partners/customers are interested ? (many, some, only a few)

Did you already get questions by partners about the data you possess and that could be interesting to them ?

- If so, About which data did these questions handle?
- How many partners/customers asked about this ? (many, some, a few,..)

What is in your eyes the value of the data shared/used with your partners for them?

Why are they willing to engage in these activities according to you ?

- What are the advantages for them?

Do you currently sell data to partners/customers or provide them for free?

- If yes, which ones? At which value?
- In which way do these data contribute to your revenue?
- Do you think this could be improved ? In which way ?
- Which processes inside your company should/could be improved to be more effective/successful in this respect ?
- If no, how realistic do you think it would be to provide your partners/customers with these data for a certain price (selling them) ?
- How realistic would providing extra services on the basis of these data can be?

- Would offering new services be a good form of getting revenue from them for your partners/customers? Why/why not?
- Which services could be involved ?

Do you think these data can either be sold/or-and/offered as an additional service to your partners? If yes. What makes these data valuable to your business? What is your business model to monetise these data?

## Authors' Information

***Hamzah Qabbaah*** *- PhD student at Research group of business informatics, Faculty of Business Economics, Hasselt University, B6b, Campus Diepenbeek, B-3590 Diepenbeek, Limburg,Belgium e-mail: hamzah.qabbaah@uhasselt.be*

*Major Fields of Scientific Research: Data mining, Digital marketing, Business Intelligence, Machine Learning, Knowledge Management*

***George Sammour*** *- Director of Quality Assurance and Accreditation Centre, Business Information Technology Department, Princess Sumaya University for Technology, Amman, Jordan. e-mail: George.Sammour@psut.edu.jo*

*Major Fields of Scientific Research: Data mining, Digital learning, lifelong learning, professional development*

***Koen Vanhoof*** *- Professor Dr., Head of the discipline group of Quantitative Methods, Faculty of Business Ecnomics, Universiteit Hasselt; Campus Diepenbeek; B-3590 Diepenbeek, Limburg,Belgium e-mail:koen.vanhoof@uhasselt.be*
*Major Fields of Scientific Research: data mining, knowledge retrieval*