화물 배차 빅데이터 분석

수간야 셀바라지, 최은미*,1,2
[1]금융정보보안학과 / [2]소프트웨어학부,
국민대학교
e-mail : {suganya, emchoi}@kookmin.ac.kr

Analysis of Freight Big Data using R-Language

Suganya Selvaraj and Eunmi Choi*,1,2
[1] Department of Financial Information Security / [2] College of Computer Science,
Kookmin University, South Korea
{suganya, emchoi}@kookmin.ac.kr

Abstract

Data analysis is a process of generating useful information by evaluating real-world raw data for making better decisions in business development. In the freight transport logistics companies, the analysis of freight data is increasingly garnering considerable importance among the users for making better decisions regarding freight cost reductions. Consequently, in this study, we used R programming language to analyze the freight data that are collected from freight transport logistics company. Usually, the freight rate varies based on chosen day of the week. In here, we analyzed and visualized the results such as frequency of cost vs days, frequency of requested goods in ton vs days, frequency of order vs days, and frequency of order status vs days for the last one-year freight data. These analysis results are beneficial in the viewpoint of the users in ordering process.

1. Introduction

In Big Data, data analysis plays an important role in making better decisions regarding business development. Data analysis is a process of making useful information by evaluating real-world raw data. In the freight transport logistics companies, the analysis of freight data is increasingly garnering considerable importance among the users for making better decisions regarding freight cost reductions. In this paper, we analyzed freight data to find a frequency of cost vs days, frequency of requested goods in ton vs days, frequency of order vs days, and frequency of order status vs days for the past few years using R-Language. In case of freight transport logistics, freight rate defers based on a requested day of the week.

From the results of our experiment, users can compare the cost and easily identify the day in which have the low freight cost. User also can compare the cost of each day with requested frequency and status frequency of each day. Our main aim is to display these analysis results on the freight Logistics company website. These results will be useful for the users to make a better decision to reduce freight cost.

The next section shows the related work. The Section 3 presents the freight data analysis process. The Section 4 expresses the experimental results in the features of freight order. We conclude this paper in the Section 5.

2. The Related Work

Many languages are available for the data analysis. Few popular languages are R-Language, Python, SQL, JAVA, Scala, Julia, and MATLAB. Particularly R and Python are more popular languages. Python is a multipurpose language but R is a specific language and an environment for the statistical computing and graphics introduced by Ihaka R *et al* in 1996[1]. Additionally, R is a free, open source, powerful and highly extensible programming language. R-Language is similar to S language and released under the GNU General Public License. R can be installed as a package and it supports Linux, Windows, and MacOS[2]. In here, we are using R-Language for freight data analysis and visualization.

There are many researches and works using R-Language for the data analysis[3]. Additionally, Tochukwu and Udeh analyzed the twitter trend using R-Language[4]. Robin Lovelace *et al*. proposed a library package stplanr for Transport Planning in R[5]. W. Cho *et al*. proposed interactive visualization for Big Data analysis using Hadoop and R[6]. Ahmed Imran KABIR *et al.* analyzed the power of social media based on sentiment analysis and word clouds on R[7].

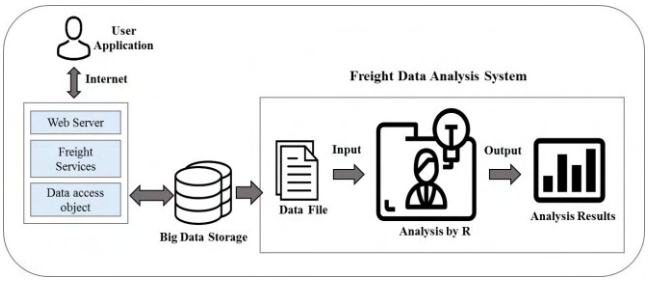
3. Freight Data Analysis Process

Big data analysis requires massive data storages and stepwise analysis by using statistical observation with data regression, data mining mechanisms, and semantic analysis. Figure 1 shows our simplified process of freight data analysis. This project takes data from excel files exported from the freight service system and provides freight data

* Corresponding Author

analysis results as an output.

This project follows three steps for analyzing the data. Those are data preprocessing, data analysis, and data visualization. Usually, real-world data consists of noisy, missing, and redundant data. So data preprocessing is necessary for changing real-world raw data into clean and complete data. Data preprocessing usually contains data cleaning, data normalization, data transformation, missing values imputation, data integration, and noise identification tasks[8]. In data analysis process data will be converted into the useful format by calculations and other operations. Data visualization usually refers the presentation of data in a pictorial or graphical format. In here data visualization is used to display the analyzed freight data into chart format for the better understanding.



(Figure 1) The overall system architecture

4. Experimental Results

4.1 Experimental Environment

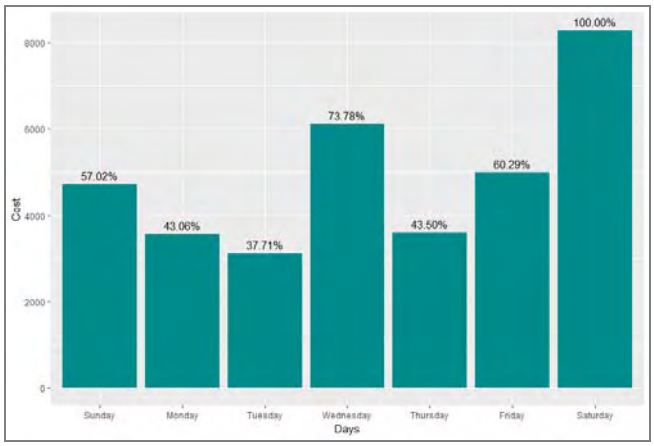
The analysis is performed by using R library packages on the R-Studio. R-Language works after installation of R software which is available in the Comprehensive R Archive Network (CRAN) and R provides graphical user interface called RGui[9]. R-Studio is an integrated development environment (IDE) which provides the user-friendly interface for R and it contains R console, R syntax-highlighting script editor, environment history, plots, packages and others[10]. In this experiment, the R library packages like readxl, data.table, ggplot2 and others are installed for performing the tasks in an experiment. It imports excel file, performs data analysis, creates a data table, draws charts and so on[11].

4. 2 Analysis Results

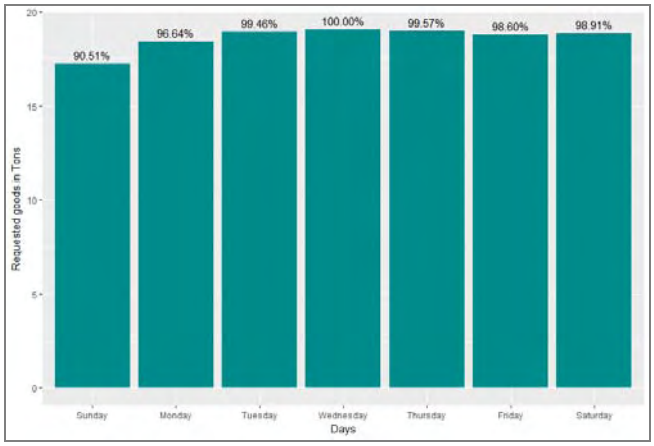
In this paper, we analyzed around 1,050,000 records which are given as an excel file. While executing the R project, first a given excel file will be imported into the project as a table and then imported data will be cleaned by removing rows which contains invalid data. The preprocessing detects invalid data from the raw data set, and the rate of invalid data occupies around 22% of the whole set of raw data. After cleaning process, required data will be calculated for the analysis by the mathematical operations.

Because of the features of freight data, we apply the daily observation of trends. As shown in Figure 2, cost per unit and name of the day needs to be calculated from the given data. Cost per unit will be calculated using requested goods weight, cost, and distance. Day's name will be calculated by using shipping start date. freight data analysis charts are visualized using ggplot library by calculating aggregate mean for all valid unit cost for each day as shown in Figure 2. This result clearly shows that in a week Saturday have highest freight cost and Tuesday have the lowest freight cost comparing others. The freight rate of the package sent by Tuesday is only 37.7% comparing Saturday. Monday and Thursday have around 43% second lowest freight rate than Saturday as shown in Figure 2.

Figure 3 shows the analysis result of requested goods quantity frequency vs days. From the result, requested goods quantity frequency is not having much difference between all days. The lowest requested quantity of goods frequency is only around 10% comparing the highest requested quantity of goods frequency.



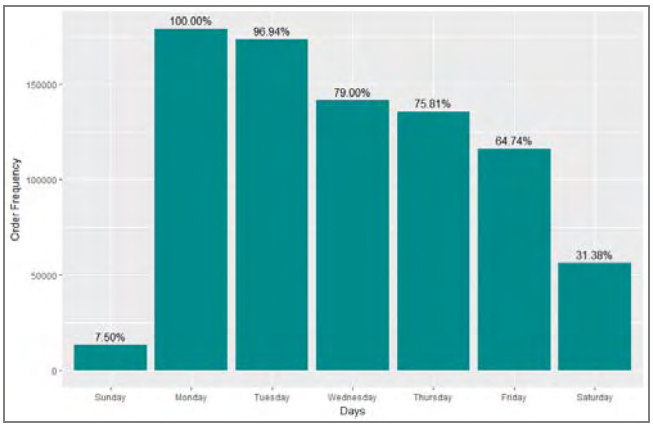
(Figure 2) Analysis result for day vs frequency of cost



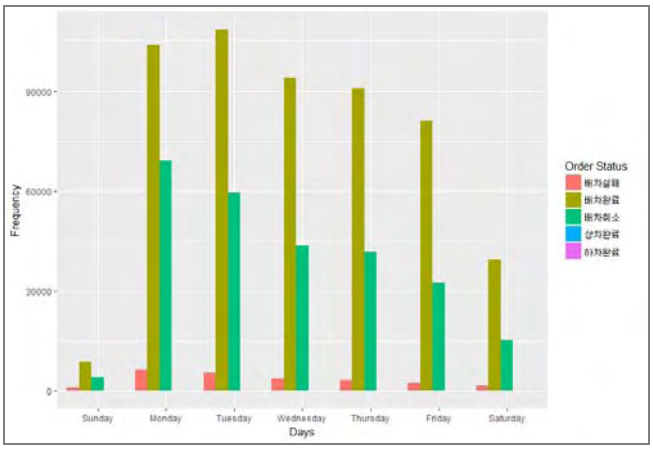
(Figure 3) Analysis result for day vs frequency of Requested goods quantity in Tons

Figure 4 shows that Sunday has the lowest request numbers 7.5% and Saturday has second lowest numbers 31.38% comparing with the highest request on Monday. Figure 5 shows the result of order status frequency vs days. Each record in a given data has anyone of the status like dispatch success, dispatch cancel, dispatch failure and so on. From the result, the successful dispatch frequency is higher than others. Comparison of Figure 2, Figure 4 and Figure 5 shows, average weekdays freight rate is lower than

weekends, so that entire orders on the weekdays are higher than the weekends. These kinds of analysis results are beneficial in the viewpoint of ordering process of the users for effective decision.



(Figure 4) Analysis result for frequency of order vs days



(Figure 5) Analysis result for day vs frequency of order status

5. Conclusion

The real-world data needs to be analyzed to make decisions for the business development. In this study, R-Language, R-Studio, and few required R library packages are installed and used in this experimental analysis. We analyzed and visualized the results as four different results such as frequency of cost vs days, frequency of requested goods in ton vs days, frequency of order vs days, and frequency of order status vs days for the last one-year freight transport data from the user's point of view. These results will be useful for the users to make a better decision in future bookings to reduce the freight cost.

Acknowledgement

This work was supported by the Technological Innovation R&D Program (A2017-0293) funded by the Ministry of SMEs and Startups (MSS, Korea).

References

[1] Ihaka R, Gentleman R. R: A language for data analysis and graphics. J Comput Graph Stat. 1996;5:299–314.
[2] https://www.r-project.org/about.html
[3] http://makemeanalyst.com/companies-using-r/
[4] Tochukwu, Udeh. (2015). Trend Analysis Using R.
[5] Robin Lovelace et al. stplanr: A Package for Transport Planning. https://cran.r-project.org/web/packages/stplanr/vignettes/stplanr-paper.html
[6] W. Cho, Y. Lim, H. Lee, M. K. Varma, M. Lee, and E. Choi, "Big Data Analysis with Interactive Visualization using R packages," in Proc. of the 2014 International Conference on Big Data Science and Computing (BigDataScience '14), Beijing, China, 2014, Article 18, 6 pages
[7] Ahmed Imran KABIR et al. The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R. Informatica Economică vol. 22, no. 1 (2018).
[8] García, S., Ramírez-Gallego, S., Luengo, J. et al. Big Data Anal (2016) 1: 9. https://doi.org/10.1186/s41044-016-0014-0
[9] https://cran.r-project.org/
[10] https://www.rstudio.com/products/rstudio/
[11] https://cran.r-project.org/web/packages/