ELMA EDISON

A00279906

Interpretation of Data – (AL_KDATA_9_1)

30-11-2020

Assignment 1

Introduction

Healthcare, due to its privacy laws and federal/state legislation, provides one of the most difficult forms of data in any sector. Yet the amount of available data is growing with the rapid growth of enabling technology tools that collect healthcare data. Pharmacy gathers and preserves vast amounts of patient and transaction-based data for a period of 10 years. As customer prescription spending increases and margins grow tighter, competition for retail pharmacies everywhere is getting more and more intense. In order to remain competitive in the ever-growing pharmaceutical industry, creating organized, maximized operations that keep the company running smoothly day in and day out is essential for pharmacy owners to allow them to spend more time focused on patient health. Pharmacy information is given in all sorts of formats and from different sources, including data from patients, data on drug purchases, data on the supply chain, clinical evidence from therapeutics, data on insurance data, data from the external marketplace. This data may be included in various databases, financial systems, systems of the product category, systems of the supply chain including pharmacy management systems.

Considering the case where a pharmacist is about to start up her own retail pharmacy business and she is unsure of what data analytics tools can be used for the development of her business. Most of her company's pharmaceutical data is stored in one database management system that contains many files in multiple formats including csv, txt, and Excel formats. The data is stored in various locations including the Cloud, three desktop PC's, a Microsoft laptop, a MacBook and some supplier information on her iPhone. There will be some redundant data that needs to be removed and cleaned and some would require merging into other dataset.

Importing, Cleaning and Merging of Data using Excel

Excel is one of the easiest data analytic tool which can be used for importing, cleaning and merging data. External data can be imported properly using 'From Text' option under Data tab. The text import wizard will appear and the delimiter can be selected accordingly. To separate the contents of one excel cell to different columns, 'Text to Column' wizard can be used. For the sales data of the pharmacy, Date column can be split into day, month and year. So that it becomes easier when depicting trends based on time period. Formulas like

MID(text, start_num, num_chars), number formats, addition of decimal places etc can be used in Excel to clean this data. She can use the 'Get Data' option under the Data tab to merge multiple excel datasheets. If her data is completely fine, then she can combine and load the data or else if changes are required, then she can Transform data in which she can filter out files accordingly. Duplicate rows in excel can be removed by 'Remove Duplicate' option under Data tab. To fill in the missing values, firstly the count of blank cells in each column need to be obtained. So for that "=COUNTBLANK(range)" function can be used. According to the percent of blanks in each column, we can determine whether to fill in the missing values. If the percent is low, then she can fill in the blanks with "=IF" function after finding the mean (=AVERAGE) or median (=MEDIAN).

Data Manipulation and Exploration using R Studio

For a much more detailed cleaning, tidying and visualisation of data, a data analytics tool called "R Studio" can be used. Tidyverse is a collection of packages in R Studio required for data analytics. It contains readr, dplyr, tidyr, ggplot2 etc but it's better to always install and load packages separately. Firstly readr Package can be used to read and load differnt forms of data into R. Install the package first using install.packages('readr') and load it using library(readr). To read delimited files, read_csv(), read_delim(),read_tsv(), read_csv2() can be used and to read fixed width files, read_table(), read_fwf() can be used. After reading the data into R, exploration of data can be done using 'dplyr' and 'tidyr' Packages.

Dplyr Package is used to explore and transform the data. It is highly adaptive to use as it has a chaining syntax. There are some major data manipulation commands in this package which she can use to manipulate her data in order to generate the medical expenses report. The initial step is to install and load the package. There are few major data manipulation commands in R which can be very helpful for her. Filter() command can be used to filter out data with required conditions. She can use Select() command to pick up columns of her interest from the dataset. The command Arrange() is used to arrange the values of the dataset in ascending or descending order. Mutate() can be used from existing variables to construct new variables. Summarise() or Summarise_each() with group_by() can be used to evaluate widely used operations such as min, max, mean count, etc and to find insights from data.

Reshape2 Package is a very helpful package in reshaping data with two functions namely melt and cast. The pharmacist has her data in many forms such as csv, txt and Excel. So she has to tame it according to her need. The 'melt' function converts wide-format data to long-format data. Several categorical columns are melted into special rows here which is a type of data restructuring. The function 'cast' is the opposite of the function 'melt' that transforms data from long format to large format. It has two features, dcast (gets back output as a data frame) and acast (gets back output as a vector/matrix/array). Splitting the date into day, month and year can be also be done using lubridate Package which helps with date time variables in R.

To make her data look tidy, tidyr Package can be used with its four major functions which are gather(), spread(), separate() and unite(). As always, downloading and loading the package is the initial step. Gather() can be found in the reshape package as an alternative to'melt '. Multiple columns are also collected and translated to key:value pairs. Spread(), which takes the key:value pair and transforms it into columns, is the opposite of gather(). To divide a column into several columns, separate() is used and unite() is the reverse of separate.

Using dplyr with tidyr can make the data exploration phase a lot easier. R Studio, along with the packages mentioned above will help the pharmacist to generate the medical expense report as the patients require it for their end-of-year tax purposes. As pharmacy dispensaries are busy places, R Studio can be of great help for her to complete the operation quickly and accurately. Large quantities of drug supply data documentation are obtained in raw text format from the data suppliers which contains drug composition, clinical indications (i.e., uses), side-effects and dosages. This data is then imported, cleaned and merged. From this dataset, it is possible to filter out the medicines dispensed under state covered schemes. This can be used to calculate how much HSE owes back to her pharmacy. Through this way, monitoring the sales of medicines can be done for the purpose of reimbursement from HSE.

Visualisation of data can be done using ggplot2 Package which offers lots of patterns and colours. The package 'ggplot' with combination of other packages produce better visualization. Complex chain commands can be used to plot graphs with different criteria.

Various graphs such as Scatter Plot, Bar Plot, Histogram etc can be plotted using this package. But a much more simpler data analytics tool for visualisation which I think would be Rapid Miner.

Data Visualisation using Rapid Miner Studio

Data visualisation can be done using Rapid Miner Studio for plotting various graphs to depict trends in stock movements. Since the pharmacist have plans for biannual meetings with other company healthcare staff to discuss various topics and what she mainly wants is to depict the type of medicines dispensed in winter and summer months. Some antibiotics are highly dispensed in the winter months when more people are sick and hay fever medicines are dispensed more in summer months when the pollen count is more. In Rapid Miner Studio, we can import excel or csv files and under Results Tab, visualisation is shown where different types of graphs are available to visualise data.

Conclusion

As the pharmacist is about to start a new retail pharmacy, she should care about her customer's needs and medicine management should be dealt with importance. There are various data analytics tools used for data collection, manipulation and visualisation. Few among them are Excel, R Studio and Rapid Miner Studio which can help her to improve her business.