Introduction:

Due to the recent popularity, automatic Exploratory Data Analysis (EDA) packages were taken into consideration for the pre-processing of the datasets. Automatic EDA packages are recently getting common among data scientists' community to achieve pre-processing in the shortest time possible. In this project, funModelling [3] and DataExplorer [6] packages are analysed on the two datasets namely Bank Marketing Dataset [1] and Human Activity Recognition Using Smartphones Dataset [2]. According to the Staniak EDA report [8] DataExplorer is the second most popular package and funModelling is the fourth most popular package.

Packages:

Only EDA is done in the code since the main objective of the project was to research EDA packages not analysing datasets. However, necessary data manipulation is also applied where necessary. The Bank Marketing dataset is used mainly to run and look at the results of the different functions of both packages whereas functions performance is mostly highlighted in the Human Activity dataset. The reason for this is that Human Activity dataset has 563 features in total. The Bank Marketing dataset is to find useful analysis of the bank customers dataset which gives information such as customer's age, salary etc. in order see weather a customer have subscribed to a term deposit [7]. The Human Activity dataset is used in order to evaluate the activity of a human. The data is fetched from the various sensors of different devices worn by a person such as a smart watch and smart cell phone. The data is given in terms of coordinate positions of acceleration magnitudes in x, y or z axis. Where x-axis is the direction of movement of a person, y-axis is the direction of sideways of the movement and z-axis is the movement in the vertical such as moving up or down. Some functions of funModelling have limited input features requirement which was seen when human activity dataset was used. Due to the same reason, in the Human Activity code, some functions of funModelling were used in for loop.

- In terms of the extent to which the funModelling package is built, the technical validity of the package is extensive and in contrast the DataExplorer is comparatively less technically valid. Main reason for that is that DataExplorer provides fewer functions and most of which are simply for plotting whereas funModelling has better technicality in its functions such as var_rank_info which provides information like information gain and entropy of the data.

- Rigour of DataExplorer is not as extensive as funModelling. Most of the functions of the DataExplorer are for plotting different graphs as compared to the funModelling which have wide variety of functions ranging from statistical functions to plotting functions.

- Performance of both packages were similar on both datasets. However, DataExplorer functions like PCA (plot_prcomp()) are seen to be performing faster than if manually built functions. This point is further discussed in relation to other factors below.

- From the perspective of user, DataExplorer is the easiest automatic EDA package. Beginners will surely find this package useful and easy to work with. It is also extremely helpful when it is known that dataset is completely clean. Learning is easy as most of the plotting functions start with plot name. FunModelling, on the other hand, is not easy at first sight since it has most of its very technical functions. But it offers a wide variety of functions which allows the users to try out different results on their data.

- Although, there are more functions in funModelling but the learning difficulties for new user will be least when funModelling package is used as many of its functions only need dataset as input without any need to mention specifications or limitations such functions are like df_status(), plot_num () and freq(). The package includes its own three different datasets (golf dataset, heart disease and flu country) to try out different functions which is a good start for beginners. Comparatively, DataExplorer also provides results without need to input any other thing other than data, therefore it is similar in terms of learning difficulties.

- Manual quality of the funModelling package is that it is simple to understand with most of the functions having easy examples under each function description which might be particularly a plus point for a beginner in R. Similarly, manual of DataExplorer is also easy to understand.

- For funModelling, most of the functions have their limitations defined in their description where some functions limitations were figured out while using that particular function. For example, using df_status() can only take around 100 columns to run so it was run in for loop for the Human Activity dataset since it has more than 100 features. But this limitation is not generally seen as mostly datasets have features less than 100. However, using big datasets will surely highlight these kinds of issues while using the specific function. In this project the larger dataset was of 7352 rows which is fairly small so in terms of data size limit, it was not an issue. However, using 563 features did highlighted maximum features input for some functions as explained above. DataExplorer did not showed any of these issues when human dataset was used.

- Computational cost of both packages was more elaborated while using the Human Activity dataset. funModelling functions which produced huge range of analysis using only single command took considerable time. Freq() command is the best example of such case which took approximately between 8 and 9 minutes when human dataset was used. DataExplorer functions are comparatively fasters but the difference in times was seen only when there is a need to generate huge number of results. Such as generating summary of each column in dataset.

- Platform requirements are basic for funModelling and DataExplorer both. There is no need for the machine to have any particular hardware update just to run these packages.

Case Studies:

Dataset	Case Study-1	Case Study-2	Case Study-3
funModelling & DataExplorer	Performance change with respect to dataset size	Graph visualization	Can still and moving body can be differentiated in human activity dataset

Case Study-1:

The performance change was not noticeable in most of the commands for funModelling package. Only var_rank_info() and freq() were able to highlight time difference when different number of columns were used. The major reason for the time difference is that the functions of DataExplorer are more advanced and generate extensive results. For example, the plot_str() function evaluate the structure of the entire dataset and visualize it as well. Another example is that plot_bar() function in DataExplorer displays bar plots for all the discrete columns without any need to input discrete columns separately. Similar functions in funModelling which allow to plot different columns does not give flexibility to plot specific kind of plot using single command or without giving necessary conditions.

Case Study-2:

There are 4 plot functions in funModelling package but 11 plot functions in DataExplorer package. It evidently explains that graph visualization power of DataExplorer is much better and it give more flexibility of plotting different kind of functions. The funModelling plotting functions are very low but do fulfill the basic needs of the user such as plotting the bar plots, box plots and correlations graphs. Therefore, it can be concluded that where visualization is not a major part of the project, funModelling will work fine as it can be used for other major analysis as well. But where the user feels the need that visualization is an important part, in most of the steps during project, then DataExplorer is an ideal choice to work with as it can give specifically bar plot, cross plot, boxplot etc.

Case Study-3:

In the human activity dataset, there was a huge number of columns. This dataset can be studied separately to evaluate many different conclusions. Different regression and classification algorithms can be run on the dataset to evaluate better results. In this project, however, machine learning algorithms were avoided to keep the project consistent for EDA. To reduce the number of columns to work with, only those columns were used which had mean values of sensor values. It was noticed that since the dataset was taken from different sensors, one major question that can be answered was that if it can be used to tell what the activity of the person is (standing, walking etc.). There were functions that take only encoded discrete values, so the Activity column was encoded from names to numbers. Then correlation plots were generated for categorical and target data and then correlation between all the variables (to plot continuous variables) was generated. The plot of plotar() function showed that variables of moving body had peaks all >2.3 value while static body variables had <1.5. For better visualization, mean accelerations of all three axis were plotted and it was concluded that z-axis had least variation as presumed where z-axis shows up-down movement geographically. Then to prove the results further, body acceleration magnitude was plotted against activity column. From figure 1, it is evident from the boxplot result that there is a difference of at-least 0.65 acceleration magnitude between static movements and moving body movements. The orange, brown and green box plots in the figure shows the static bodies and turquoise blue, light blue and pink shows the moving bodies. The body is either standing, sitting, and lying on bed/soda when mean value is <-0.8. The body is either walking downstairs or upstairs when the mean value is >-0.25 and >0.1 when body is walking is moving on plane surface.

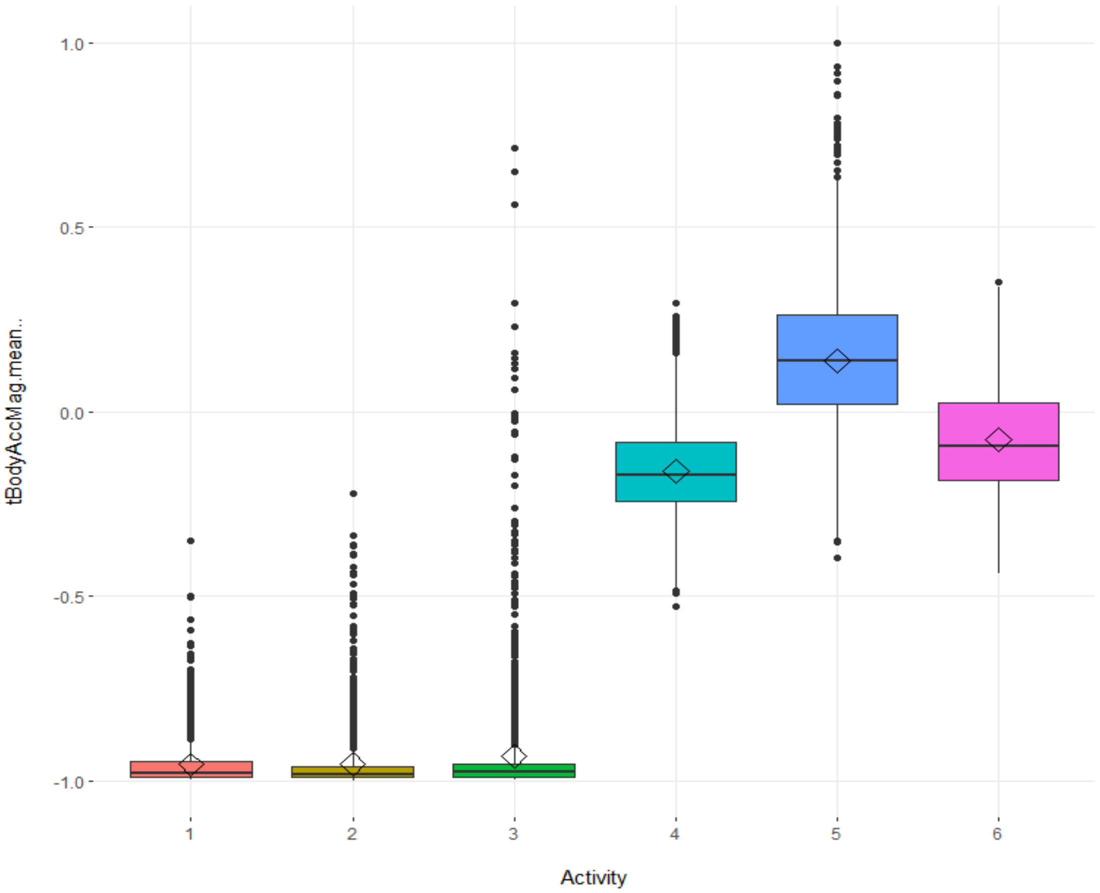

Figure 1: Body acceleration magnitudes of each activity

Conclusion:

Automatic EDA is a very practical solution for fast analysis of the dataset. It gives the ease of using few functions to evaluate the dataset. The visualization can also be done using just one or two functions as the person writing the code do not need to worry about defining each axis variable and scale range etc. Particularly, DataExplorer is for beginners as it has small number of functions with self-explanatory names. Both packages come with some limitations such as limited number of columns or rows input to some functions but that can be handled then and there. As a final verdict funModelling is preferred over DataExplorer to explore the dataset in a short time since it gives variety of functions under one roof and different datasets to explore the package but if the focus of the project is on visualization then DataExplorer is the ultimate winner.

References:

[1] Bank Marketing Data SetFeb 14, 2012-last update. Available: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[2] Human Activity Recognition Using Smartphones Data SetDec 10, 2012-last update. Available: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.

[3] CASAS, P., 2020. *Package 'funModeling'.*

[4] CASAS, P., Jan 24, 2018-last update, Exploratory Data Analysis & Data Preparation with 'funModeling'. Available: https://blog.datascienceheroes.com/exploratory-data-analysis-data-preparation-with-funmodeling/.

[5] CUI, B., Nov 21, 2020a-last update, Introduction to DataExplorer. Available: https://boxuancui.github.io/DataExplorer/articles/dataexplorer-intro.html.

[6] CUI, B., 2020b. *Package 'DataExplorer'.*

[7] MORO, S., CORTEZ, P. and RITA, P., 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, **62**, pp. 22-31.

[8] STANIAK, M. and BIECEK, P., 2019. The Landscape of R Packages for Automated Exploratory Data Analysis. *arXiv preprint arXiv:1904.02101,* .