Data Analytics: Exam
June 2021

In this exam, we'll examine different types of business cases in which data analytics are useful to provide meaningful insights to the business. Some of these cases are to be worked conceptually and others will require some R coding and the adequate interpretation of the results.

You are required to deliver a document in word, PDF or html format. If you decide to answer these questions in Rmarkdown, you can generate the html file. For the questions that require R code, you also need to provide the .Rmd file or the .R file. Thus, you need to submit a zip file with the final document (in PDF, html or word) + the .R or .Rmd document.

Case 1. German Credit Dataset (5 points)

Dataset: **credit-g.csv**

The German Credit dataset classifies people described by a set of attributes as good or bad credit risks. The list of attributes of this dataset is the following:

1. **checking_status:** Status of existing checking account, in Deutsche Mark.
2. **duration**: Duration in months
3. **credit_history**: Credit history (credits taken, paid back duly, delays, critical accounts)
4. **purpose**: Purpose of the credit (car, television,...)
5. **credit_amount**: Credit amount
6. **saving_status**: Status of savings account/bonds, in Deutsche Mark.
7. **employment**: Present employment, in number of years.
8. **installment_commitment**: Installment rate in percentage of disposable income
9. **personal_status**: Personal status (married, single,...) and sex
10. **other parties**: Other debtors / guarantors
11. **residence_since**: Present residence since X years
12. **property_mangitude**: Property (e.g. real estate, life insurance, car...)
13. **age**: Age in years
14. **other_payment_plans**: Other installment plans (banks, stores)
15. **housing:** Housing (rent, own, "for free"...)
16. **existing_credits**: Number of existing credits at this bank
17. **job**: Type of job ("skilled", "high qualif", ...)
18. **num_depenents**: Number of people being liable to provide maintenance for
19. **own_telephone**: Telephone (yes,no)
20. **foreign worker** (yes,no)
21. **class** (good, bad)

Imagine that you are a data analytics consultant. Your customer is this German Bank. After examining the dataset, you are required to explain to the bank in which ways this dataset might be useful.

Question 1 (1.5 points)

Based on the provided data, and after having a first look at its structure and contents, write and explain three main research questions (data analysis questions) that can be asked to this dataset and whose result might be obtained with a proper analysis.

Then, for each of these three research questions, explain how you would answer them based on the dataset and the algorithms that you know. Also, explain why this analysis is useful for the business.

For example, a possible main research question is:

- *What are the attributes and values of a good customer and a bad customer regarding credit risk?*

And the approach is:

- *We'll use a classification algorithm, a decision tree, where class is the dependent variable and the other variables are the independent variables.*

Thus, you need to think three more research questions that can be answered with one of the approaches (algorithms) that we saw in class. To answer this exercise, fill out a table like the one that is depicted below. The first example provided is already included in the table.

To get a good grade in this question, don't do variations of the same questions. Try to ask different research questions, of different types, to give different types of insights to the bank's experts.

Data Analytics Question	Approach	Why is this analysis useful for the business
What are the attributes and values of a good customer and a bad customer regarding credit risk?	*We'll use a classification algorithm, a decision tree, where class (risk) is the dependent variable and the other variables are the independent variables.*	*1) Understand current customers' risk* *2) To predict new customers' risk, based on the model obtained with the decision tree.*

Question 2 (3.5 points)

You are required to do the analysis of the research question exemplified in question 1. To do that, you need to upload dataset "credit-g.csv" and apply a classification algorithm to the dataset to investigate whether it is possible to predict credit risk. To do that, you need to follow the steps of the data analysis:

1. Upload the file. Create train and test splits.
2. Train the algorithm **decision tree**, using as dependent variable the variable *class* and the other variables as independent variables. Be sure that the variable "class" is a factor before you run the algorithm.
3. **Interpret** the information conveyed by the tree.
4. **Test the algorithm** and calculate its **performance** (quality).
5. Interpret the **quality of the algorithm** in technical terms. You can use **several metrics** to evaluate and explain how the algorithm performs to predict credit risk.
6. Train a **K-NN algorithm**, using the same dataset splits.
7. **Evaluate** the performance of the K-NN algorithm.
8. **Compare** the results of the K-NN algorithm with the ones obtained with the decision tree. Do it in terms of: interpretation provided by the algorithms and performance (quality).
9. Write a paragraph in which you explain to your customer (German Bank) the analysis that you did. You need to explain it very clearly, without many technical terms, as if it was a kind of executive summary for business experts.

To answer this question, you can provide plots, tables, etc. Do it as well as possible, as if it was a real case for a real customer. You can also iterate with different versions of the algorithm, if you wish, until you find a satisfactory solution and explanation. You can use all variables or a selection of the most important variables.

In a way, you are the analyst and you have freedom to lead the analysis throughout the best path to get the best results, considering accuracy and interpretability. As you have seen, some of the questions are open, so that you can decide what is best in each situation.

Case 2. Insurance Company (3 points)

Dataset: **insurance.csv**

An insurance company has gathered information about worker compensation insurance policies, when the worker had an accident. For each record there is demographic and worker related information, as well as a text description of the accident. This is the list of attributes:

1	**ClaimNumber**: Unique policy identifier
2	**DateTimeOfAccident**: Date and time of accident
3	**DateReported**: Date that accident was reported
4	**Age**: Age of worker
5	**Gender**: Gender of worker
6	**MaritalStatus**: Martial status of worker. (M)arried, (S)ingle, (U)nknown.
7	**DependentChildren**: The number of dependent children
8	**DependentsOther**: The number of dependants excluding children
9	**WeeklyWages**: Total weekly wage
10	**PartTimeFullTime**: Binary (P) or (F)
11	**HoursWorkedPerWeek**: Total hours worked per week
12	**DaysWorkedPerWeek**: Number of days worked per week
13	**ClaimDescription**: Free text description of the claim
14	**InitialIncurredClaimCost**: Initial estimate by the insurer of the claim cost (log scale)
15	**UltimateIncurredClaimCost**: Total claims payments by the insurance company (log scale).

The aim of this company is to understand the relevant factors that determine the final amount of the compensation (as stored in variable **UltimateIncurredClaimCost**), with the final aim of predicting this cost and thus, being able to improve their policies and insurance prices.

As analysts, we know that a regression analysis is appropriate for this problem. Thus, follow the next steps to provide this analysis to the insurance company:

1. **Explain why a regression analysis** is appropriate here.
2. In which ways is **this analysis different** from the one performed in case 1?
3. **Upload** the dataset in R. Do some **descriptive analysis** by showing some plots of the most relevant variables.

   Tip: Read the file using this instruction:

   ```
   df<-read.csv("insuranceBig.csv", sep=",", stringsAsFactors=TRUE)
   ```

4. **Train a regression algorithm** with the full dataset, by using **UltimateIncurredClaimCost** as the dependent variable and all the other variables as independent variables. Don't include the variables that are singular for each user or descriptive: that is, skip "ClaimNumber", "DateTimeOfAccident", "DateReported", "ClaimDescription".
5. Identify and **select the relevant independent variables** by looking at the model. Justify your selection.
6. **Refine the model** (retrain the model) by using only the relevant variables identified in step 5.
7. **How can we predict** the variable **UltimateIncurredClaimCost**, based on the selected variables? Explain the formula behind this model. Also explain the contribution of every variable to the model.
8. Is the **model precise (good) enough**? Explain.
9. Provide an example, with an invented worker accident: what would be the **prediction** of the model for this invented accident? Calculate with R.

Case 3. Customer Segmentation at RetailMart (2 points)

(no dataset needed for this exercise)

RetailMart is very happy with your work on predicting families that are expecting a baby. Thus, they ask whether we could also help in the understanding of the customer profiles at their supermarket. The dataset they provided contains the summary of sales records that summarise the customer value in the so-called RFM model, which contains three attributes: Recency, Frequency and Monetary Value:

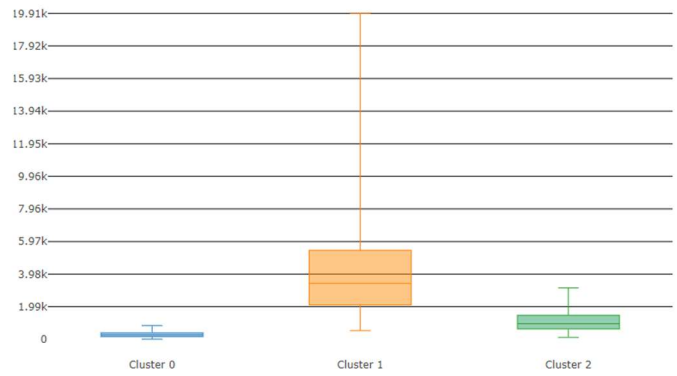
- Recency: The value of how recently a customer purchased at the establishment. This information is useful for the business, because it has been proved that the most recent the customer has visited the supermarket, the more likely will be receptive to communications from the brand.
- Frequency: How frequent the customer's transactions are at the establishment. Also, customers with frequent activities are more receptive to the brand's products.
- Monetary value: The dollar value of all the transactions that the customer made at the establishment. It is useful to identify big spenders and small spenders.

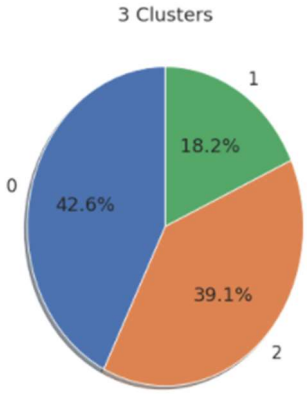
We aim at investigating whether there are different types of customers regarding these three attributes, and also to be able to allocate each customer of the dataset to one particular type of customer. To do that, we apply a clustering approach. Answer the following questions:

1. Explain why a **clustering** analysis is appropriate here.
2. In which ways is this **analysis different** from the ones performed in cases 1 and 2?
3. After applying the clustering algorithm, we obtained the following results. **Explain these results** to RetailMart, by explaining which type of customers they have. Name each cluster with an appropriate label.

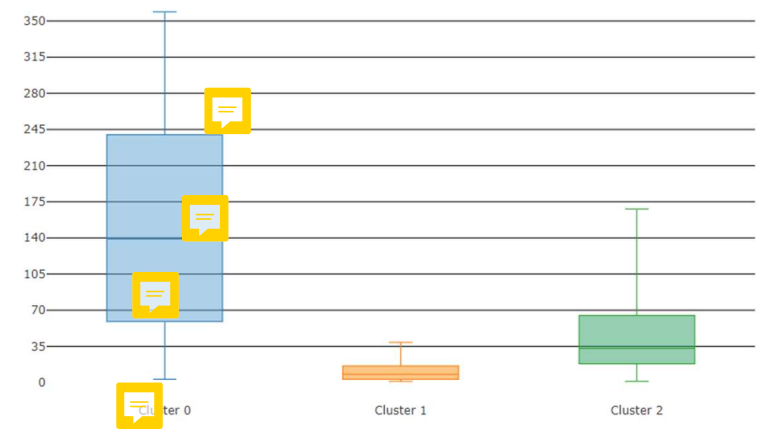
```
Centroids:
        amount     recency  frequency
0   261.952265  116.604917   1.190876
1  3967.994380    7.236580  10.044493
2  1006.914317   33.819966   3.152227
```

Boxplot of amount in the three clusters:





Boxplot of recency in the three clusters:



	Boxplot of frequency in the three clusters: