Domain.

Bank Direct Marketing Campaign

Dataset. This dataset contains information about direct marketing campaigns of a Portuguese banking institution.

a link to the dataset(s) that you have identified.

https://www.kaggle.com/berkayalan/bank-marketing-data-set

what information in it is relevant to your project (there may be lots of irrelevant extra data too)

The dataset includes some useful personal information about the customer, such as age, job, marital, education, existence credit, housing and loan.

Data relevant to the previous campaigns is also useful, eg. outcome, the number of contacts.

Lastly, customer's social and economic attributes are also relevant to our project, such as employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate, existence subscription of term deposit.

Any learning you will have to do in order to interpret the data

We will have to search online to interpret some of the "Social and economic context attributes", such as emp.var.rate(employment variation rate), cons.price.idx(consumer price index), cons.conf.idx(consumer confidence index), euribor3m(number of employees).

Any cleaning up you think you will have to do in order to use the data

There are missing values labelled "unknown"(existence credit in default). When we investigate attributes that have a missing value, we will remove the whole observation. For example, for the first question that we will investigate, we will remove any row with an unknown value in any of the attributes of age, job, marital, education, existence credit, housing, loan and pcontact. Similarly, when we will investigate other questions, we will remove any row with an unknown value in any of the attributes from those tables.

In addition, We need to check if there is an incorrect format of data and also check outliers.

Questions. Your three investigative questions that you plan to answer using this dataset.

1. Which of the following attributes is associated with the outcome of the previous marketing campaign(poutcome)?
    - age
    - job
    - marital
    - education
    - existence credit
    - Housing
    - Loan
    - pcontact(number of times contacted of the previous campaign)
    - Employment variation rate
    - Consumer price index
    - Consumer confidence index
    - Euribor 3 month rate
    - Existence of subscription of a term deposit

Follow up question:

- if there is an association, does the average poutcome increase or decrease when the attribute increase?
- For each attribute above, what value(group) results in the lowest or highest poutcome?

2.Which kind of people have the lowest or highest average poutcome?

Follow up question:

- What value of personal information(age, job, marital, education, existence credit, Housing, Loan) does the group of people have?
- What value of social and economic attributes does the group of people have?

3. Do existence credit vary by different marital, education and job groups?

Follow up question:

- if there is an association, does the existence credit increase or decrease when the attribute increase?

- For each attribute above, what value(group) results in the lowest or highest existence credit?

Schema.
Relational schema

Person(<u>ID</u>, age, job, marital, education, credit, housing, loan, pcontact)
A tuple in this relation represents a person that this bank is doing marketing campaigns to.
ID is a unique assigned person's ID, and *age, job, education* are obvious.
marital is this person's marital status.
credit indicates whether this person has credit or not. (notice that credit is the "default" attribute in the original dataset, we rename it to "credit")
housing indicates whether this person has a house loan, which can be either yes or no.
loan indicates whether this person has a personal loan.
pcontact is the number of times this bank contacted a certain person of the previous marketing campaign.(notice that pcontact is the "previous" attribute in the original dataset, we rename it to "pcontact").

Outcome(<u>ID</u>, poutcome)
A tuple in this relation represents the outcome of the previous marketing campaign that this bank did for a certain person.
ID is a unique assigned person's ID.
poutcome is the outcome of the previous marketing campaign that this bank did for a certain person with ID.

Index(<u>ID</u>, empVarRate, CPI, CCI, Euribor3m, subscribed)
A tuple in this relation represents the social and economic context for a certain person that this bank is doing marketing campaigns.
ID is a unique assigned person's ID.
empVarRate is the employment variation rate, a quarterly indicator, for a certain person.
CPI is the consumer price index, a monthly indicator, for a certain person.

CCI is the consumer confidence index, a monthly indicator, for a certain person.
Euribor3m(euribor3m in the original dataset) is the Euribor 3 month rate, a daily indicator, for a certain person.
Note: empVarRate is the "emp.var.rate" attribute in the original dataset, we rename it to "empVarRate". CPI is the "cons.price.idx" attribute in the original dataset, we rename it to "CPI". CCI is the "cons.conf.idx" attribute in the original dataset, we rename it to "CCI".
subscribed indicates whether this person has a term deposit, which can be either yes or no.

MaritalCredit(<u>ID</u>, marital, education, job, credit)
A tuple in this relation represents a person's information about marital status and credit status.
ID is a unique assigned person's ID.
marital is this person's marital status, education and job are obvious.
credit indicates whether this person has a credit or not.(notice that credit is the "default" attribute in the original dataset, we rename it to "credit")

Integrity Constraints

- Person[ID] ⊆ MaritalCredit[ID]
- Person[marital] ⊆ MaritalCredit[marital]
- Person[credit] ⊆ MaritalCredit[credit]
- MaritalCredit[ID] ⊆ Index[ID]#deleted b/c dont need these two table in any one question
- Outcome[ID] ⊆ Index[ID]

- Person[ID] ⊆ MaritalCredit[ID]
- Person[ID] ⊆ Outcome[ID]
- Index[ID] ⊆ Outcome[ID]

Data dictionary

Include attribute name, description of representation, data type, whether or not a value will always be known

Person

Attribute	Description	Type	Required	Default
ID	unique assigned person’s ID	INT	yes	
Age	The person’s age	INT	yes	
marital	The person’s marital status	TEXT	yes	None; Allowable values are 'divorced', 'married', 'single', 'unknown'
education	The person’s education	TEXT	yes	None; Allowable values are 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree'
credit	whether this person has credit or not.	TEXT	yes	None; Allowable values are ‘yes’, ‘no’
housing	whether this person has a house loan, which can be either yes or no	boolean	yes	None; Allowable values are ‘yes’, ‘no’
loan	whether this person has a personal loan	boolean	yes	None; Allowable values are ‘yes’, ‘no’
pcontact	number of times this bank contacted a certain person of the previous marketing campaign	INT	yes	

Outcome

Attribute	Description	Type	Required	Default
ID	The unique assigned person’s ID.	INT	Yes	
poutcome	The outcome of the previous marketing campaign that this bank did for a certain person with ID.	TEXT	Yes	None. Allowable values are: “failure”, “nonexistent”, “success”

Index

Attribute	Description	Type	Required	Default
ID	The unique assigned person’s ID.	INT	Yes	
empVarRate	The employment variation rate, a quarterly indicator, for a certain person.	FLOAT	Yes	
CPI	The consumer price index, a monthly indicator, for a certain person.	FLOAT	Yes	
CCI	The consumer confidence	FLOAT	Yes	

	index, a monthly indicator, for a certain person.			
Euribor3m	The Euribor 3 month rate, a daily indicator, for a certain person.	FLOAT	Yes	
subscribed	Whether this person has a term deposit	BOOLEAN	Yes	None. Allowable values are “yes”, “no”.

MaritalCredit

Attribute	Description	Type	Required	Default
ID	unique assigned person’s ID	INT	yes	
marital	The person’s marital status	TEXT	yes	
credit	whether this person has credit or not.	BOOLEAN	yes	

-Justification of design
Firstly, we create *Person, Index* and *MaritalCredit* because they are respectively related to the three questions that we want to investigate. We did not use the original structure of the dataset because there are missing values in several attributes. Also, we plan to delete the whole observation when there is missing value in attributes related to the question. To avoid removing the observation when there is missing value in some attributes that are not related to the question we want to investigate, we choose to create one table for each question.

Secondly, since the first and the second question both are related to the outcome of the previous campaign, to avoid redundancy, instead of putting the outcome in *Person* and *Index* table, we separate it as a new table *Outcome*.

We invent a key ID in each table to represent a unique id of a person because in the original dataset, there is no unique attribute that can make each observation be distinct from others.