Quality of the insurance company data:

Profiling of insurance company data.
There are 600 clients.
There are 957 unique identifiers.
There are 3 customers without an identifier.

Customer information:

- There are no duplicate clients (the first and last names are unique for each identifier).
- Each client has a single insurance contracted.
- Name:
  - All names are words, they are possible.
- Surnames:
  - There are 21 customers without the second last name and 5 of them have no last name.
  - All surnames are words, they are possible.
- Date of birth:
  - 17 clients have no date of birth.
  - All existing dates are possible (they range from 1991 to 1997).
- Phones:
  - 14 customers do not have phone 1 or phone 2.
  - There are no impossible phone numbers.
  - They all start with 6 and have 9 digits.
  - There are no duplicate phones.
- E-mail:
  - There are no duplicate emails.
  - 7 clients do not have email.
  - 63 has the wrong email.
  - 890 have gmail.com mail.
  - There are 65 emails with the letter ñ and 60 are from gmail.com.

- Contracted insurance:
  - Insurance:
  - There are 12 types of insurance contracted (from 1 to 12).
  - 60 clients do not have a contracted service but they do have a contracting date.
- Insurance contract date:
  - 37 clients do not have a service contracting date and they do have a contracted service.
  - All recruitment dates are possible (they range from 1990 to 2017).
- Country:
  - 12 clients have no country.
  - The only country where customers live is Spain.
  - There is an error with a couple of clients with Italy country and with a Spanish province.

- Province:
  - Correct typographical errors (tildes) of province.
  - All values are provinces of Spain.

- CP Service:
  - Fix errors in data type (string instead of int) of cp service.
  - CP services range from number 1 to 49.
  - Almería, Granada, Malaga and Seville have more than one cp service. The rest have only one cp service.
  - There are only 5 provinces that share cp service. Not knowing the meaning of cp service I can't tell if this is an error or not.

Analysis of the observed data quality problems.
Instead of a joint database (client and type of service contracted) there should be two types of database:

- Dataset with customer data.
- Dataset with the data of the insurance contracted by each client (each client would be an identifying number in this dataset).

Incomplete information:

- Unique identifiers missing per customer
- Last names missing
- Birth dates are missing
- Contact phone numbers missing
- Emails are missing
- The types of insurance contracted for some clients are missing
- Contract dates are missing for some clients
- There are no countries for contracting the service (although we assume that the service is provided only in Spain. Therefore, this variable would be irrelevant)
- The ID number variable is missing for the ID type variable to make sense

Wrong information

- There are incorrect emails
  - Non-existent email providers (example: @com)
  - Emails with letter ñ
- There are wrong countries
  - The country Italy is specified when clients appear to contract the service in Spanish provinces.
- There are duplicated provinces due to tick errors.
- There are cp services with wrong data type (str instead of int).

10 data quality controls.
Two databases must be created:

- One with the client's data: NUM_ID, NAME, SURNAME1, SURNAME2, NAC_DATE, TYPE_ID, TELF1, TELF2, EMAIL, DOM_SERVICIO_PROVINCIA, DOM_SERVICIO_CP.
- Another database of the Service: Service contracted ID, Service contracted, DATE_CONTRATACIÓN and NUM_ID.

Data quality controls:

- NUM_ID quality control
  - There must be a unique identifier per customer, there are no duplications.
  - They must be of type int (numeric).
  - The ratio of correct records to the number of total records must be 100%.
  - The data engineering team must create a form registration that only accepts numeric characters for the NUM_ID field.
  - If there are incidents, contact the data engineering technical team.

- Quality control of NAME, LAST NAME1 and LAST NAME2
  - The composition of the three fields has to be unique in the customer's database.
  - It must be of type str (word).
  - The ratio of correct records to the number of total records must be 100%.
  - The data engineering team must create a form registration that only accepts alphabetic characters for these fields.
  - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- NAC_DATE quality control
  - They must be in date format.
  - Must be possible dates
    - The oldest date must be a maximum of 120 years ago.
    - The most recent date must be at least 18 years old.
  - The ratio of correct records to the number of total records must be 70%.
  - The data engineering team must create a form registration that only accepts dates for this field.
  - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- Quality control of phones:
  - They must be type int (numeric).
  - They must start with 6 (if they are mobile) and have 9 digits.
  - There should be no duplicate phones.
  - The ratio of correct records to the number of total records must be 70%.
  - The data engineering team will need to create a form discharge that only accepts discrete numbers for this field.

    - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- Email quality control:
    - There should be no duplicate emails.
    - Emails must have an existing email provider (example: gmail.com)
    - Emails must not contain the letter ñ.
    - The ratio of correct records to the number of total records must be 60%.
    - The data engineering team must create a form registration that only accepts emails with a valid email provider for this field.
    - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- Quality control of contracted insurance:
    - The insurance must be within the twelve types of existing insurance.
    - Every client must have, at least, a contracted insurance.
    - The ratio of correct records to the number of total records must be 100%.
    - The data engineering team must create a form registration that only accepts existing types of insurance.
    - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- Quality control of the insurance contracting date:
    - They must be in date format.
    - Must be possible dates
        - The oldest date must be the company start date.
        - The most recent date must be today.
    - The ratio of correct records to the number of total records must be 80%.
    - The data engineering team must create a form registration that only accepts date format for this field.
    - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- Country (I assume it refers to the country in which the service is provided, not the client's country of origin).
    - I assume it is a company that only provides service in Spain.
    - The country of residence must be only Spain.
    - It must be of type string (word).

    - Following the assumption of Spain as the only operating country and that it refers to the country in which the service is provided, I would proceed to eliminate this variable, since it would be redundant.

- Quality control of the province
    - It must be type string (word).
    - It must be within the Spanish provinces.
    - You must identify the province whether or not it has typographical errors. In case of not identifying it, it should not be a valid field.
    - The ratio of correct records to the number of total records must be 70%.
    - The data engineering team will have to create a registration form that only accepts between the existing provinces of Spain.
    - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- TIPO_ID quality control
    - It must be of type str (word).
    - The ratio of correct records to the number of total records must be 95%.
    - You should create another variable specifying the ID number for this variable to have meaning and relevance.
    - The data engineering team must create a form registration that only accepts alphabetic characters for these fields.
    - If there are technical incidents, contact the technical engineering team. If there are functional incidents, contact the back-office team that is in charge of making customer records.

- Quality control of CP service (I do not know the meaning of this variable so I cannot establish a quality control for it).
    - It must be type int (numeric).