PRACTICAL TEST | Núcleo de Produção de Dados

About the practical test:

This practical and technical test is intended to assess the following skills:

- Good coding and documentation practices
- Ability to analyze and interpret a dataset
- Solve problem with creativity

About the database:

The database used in this practical test is a public sample of the Information System on Live Births (SINASC) available on the Ministry of Health/DATASUS website. The purpose of SINASC is to gather information regarding births that occurred throughout the national territory.

Download the sinasc_2016.zip file

Mission:

Explore a database, structure a Jupyter Notebook in R or Python or an RMarkdown in R with the code and documentation of answers and solutions for the proposed challenge.

Challenge:

Information:

- Change column names to lowercase.
- At the end of the challenge, show the information regarding the number of records and columns, as well as the list of variables contained in the database after the proposed solutions.
- The final database must be optimized, containing only the necessary information, to occupy as little storage space as possible..

From the information provided, explore the SINASC dataset and answer the following questions:

1. Considering that the first two digits of the municipality code represent its Federative Unit (UF):

   a) Create two variables, one with the code and the other with the name of the UF of birth and show the distribution.
   b) What is the name of the five most frequent UFs, where the municipality of residence is the same as the one at birth?
   c) How many municipalities appear in each UF of birth?

2. Create an integer variable with the age of the mother using the date 12/07/2021 as a reference.

   a) What are the three lowest ages and three highest ages of mothers?
   b) What is the frequency distribution of the age of mothers in Bahia?
   c) If there is any inconsistent age, take the necessary treatment and show.

3. In relation to the birth weight:

   a) Create an integer type variable with the length of the 'weight' field and show the distribution.
   b) How many babies were born weighing less than 2000 grams? (If necessary, remove inconsistencies from the variable.)
   c) What is the percentage of this value considering the size of the database?

4. How many health facilities does the database contain in each FU?

5. In relation to UF of birth:

   a) How many children were born in a hospital or at home in Bahia?
   b) Show the distribution of births per month in Bahia

6. In relation to the mothers of the Born:

   a) Show the distribution of schooling of single, black and brown mothers for each UF.
   b) What percentage of married women have had a vaginal birth at less than 37 weeks' gestation? And in relation to women who had Cesarean delivery, what percentage??

Data dictionary - SINASC:

- `dtnasc` : Date of birth: dd mm yyyy
- `codmunnasc` : City of birth code
- `codmunres` : City of residence code
- `dtnascmae` :Mother's date of birth (dd mm yyyy)
- `codestab` : Code of the health establishment (CNES) where the birth took place.
- `racacormae` : Type of race and mother's color: 1 – White; 2- Black; 3- Yellow; 4 – Brown; 5 – Indigenous.
- `locnasc` : Place of birth: 1 – Hospital; 2 – Other health establishments; 3 – Home; 4 – Others.
- `parto` : Type of delivery: 1- Vaginal; 2- Cesarean; 9- Ignored.
- `peso` :birth weight in grams.
- `escmae2010` : Schooling (2010). Values: 0 – No education; 1 – Elementary I; 2 – Elementary II; 3 – High School; 4 – Incomplete higher education; 5 – Complete Superior; 9 – Ignored.
- `estcivmae` : Mother's marital status: 1- Single; 2- Married; 3- Widow; 4 – Legally separated/divorced; 5- Stable union; 9- Ignored.
- `gestacao` : Weeks of gestation: 1– Less than 22 weeks; 2-22 to 27 weeks; 3– 28 to 31 weeks; 4– 32 to 36 weeks; 5-37 to 41 weeks; 6– 42 weeks and more; 9- Ignored.
- `semagestac` : Number of gestation weeks
- `idadepai` : father's age
- `seriescmae` : mother's school grade
- `tpapresent` : Type of presentation of the RN. Values: 1- Cephalic; 2- Pelvic or podal; 3- Transverse; 9- Ignored.
- `sttrabpart` : Induced labor? Values: 1- Yes; 2- No; 3- Not applicable; 9– Ignored.
- `stcesparto` : Did cesarean occur before labor started? Values: 1- Yes; 2- No; 3- Not applicable; 9– Ignored.
- `tpnascassi` : Birth was attended by? Values: 1 – Physician; 2-Nurse/midwifery; 3- Midwife; 4- Others; 9- Ignored.
- `consprenat` :Number of prenatal appointments.

What do we expect from the candidate?

We want to understand your technical skills, analytical reasoning and presentation of results so we have a clear insight into how you solve problems. Therefore, we hope that you will deliver a Jupyter Notebook in R/Python or an RMarkdown in R demonstrating your exploratory analysis process and your conclusions.

Answers, code documentation and comments can be written in Portuguese.