A3Data

LOUCOS POR DADOS, APAIXONADOS POR GENTE.
#VEMPRAA3
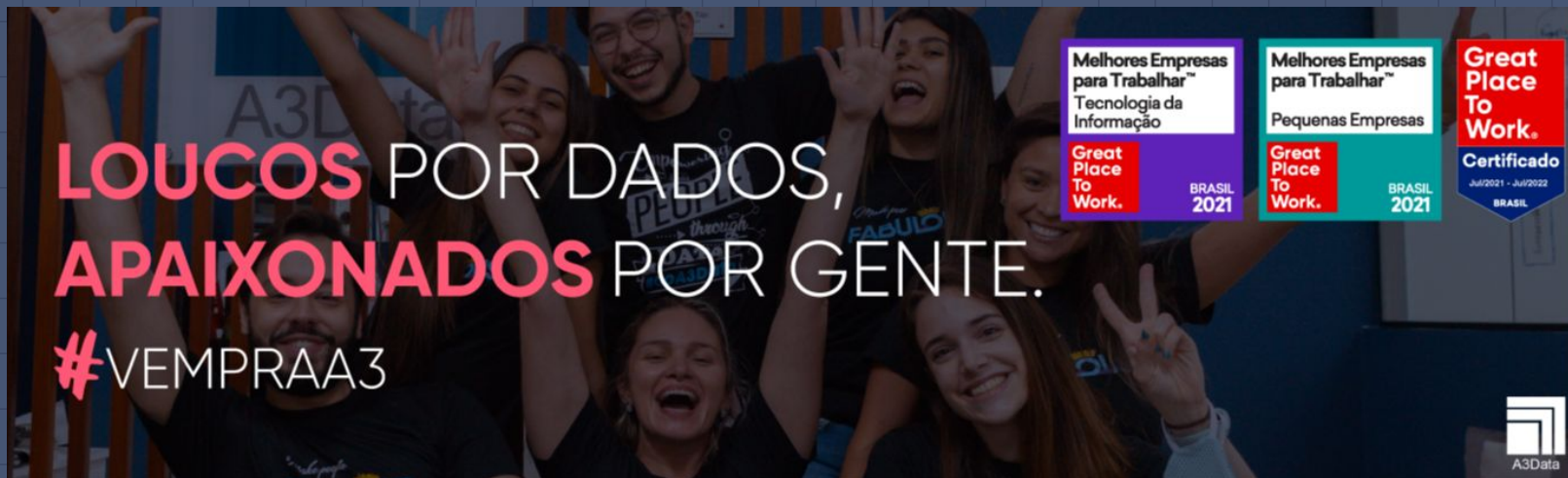
Job Application Case

Case context

Given the **available dataset** and a **deadline of 48 hours**, the candidate has to:

- Explore the **dataset** to **find insights**
- Make a pdf **presentation** with:
  - (1) a description of the **challenge**;
  - (2) an explanation of the **process** used;
  - (3) **hypotheses raised**;
  - (4) **exploratory analysis**; and
  - (5) **conclusions** and **insights** generated
- Make the **code** used for this project **available on GitHub**

Project Structuring
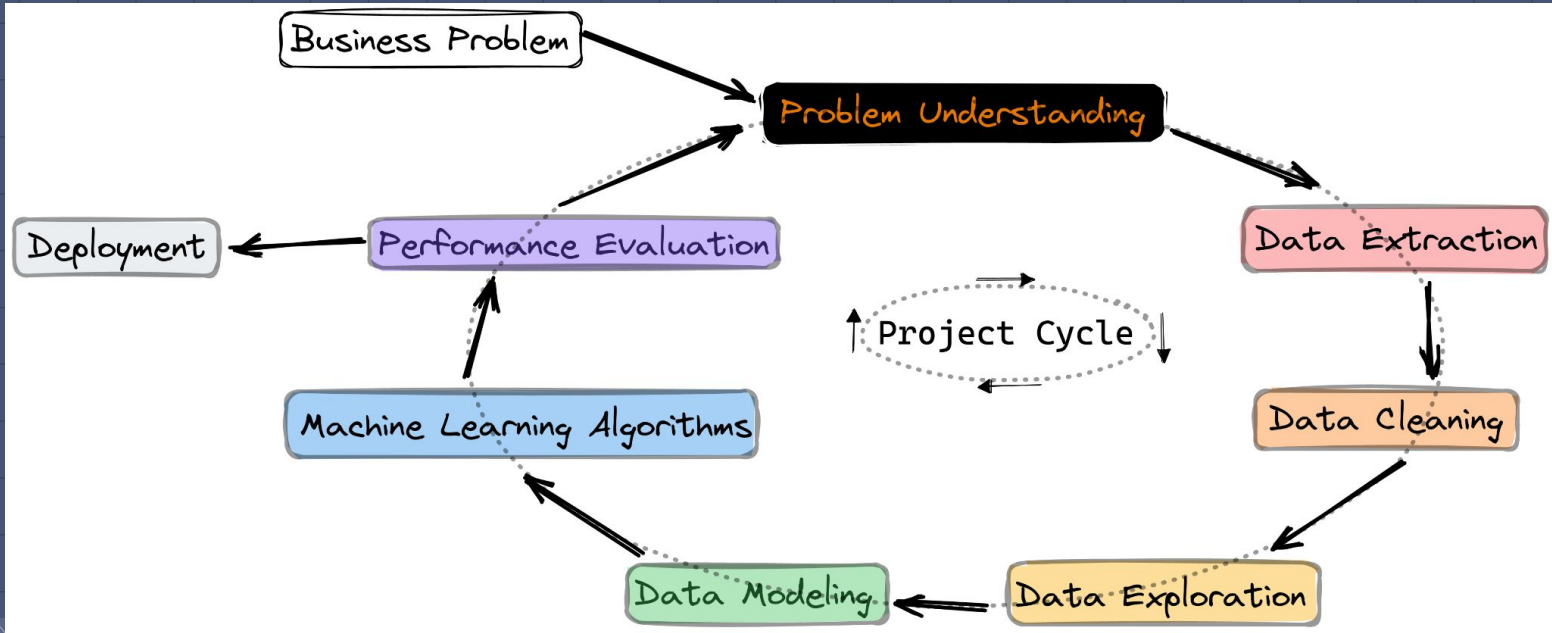
CRISP-DS

Business problem

What is the company? **A3Data**

What is its business model? **A3Data** is one of the **leading consultancies** specialising in **data** and **artificial intelligence** in Brazil.

What is the business problem the company is facing?

A3Data was given the task by one of its clients to find out what are the **main factors associated** with **aeronautical accidents given** the **public database** from the Brazilian Aeronautical Accident Investigation and Prevention Center. These factors (and **insights**) will be used for **public campaigns** to advise the population, politicians and companies about what should be done to **reduce the number of accidents**.

Problem understanding

What is the business solution that this project has to deliver?

Based on the given dataset, this project has to **deliver insights** that could be used in **public campaigns** to **reduce the number of aeronautical accidents**. These **insights** will be delivered in the **format** of a **presentation** with concise explanations about the **process** used to **explore** the **data** available, about what were the **raised** and **validated hypotheses**, and what were the **concluding insights**. The **code** used for this project should also be made **available on Github**.

References:

https://dados.gov.br/dataset/ocorrencias-aeronauticas-da-aviacao-civil-brasileira

Main assumptions of this cycle

Business assumptions:

- The main goal of this project is to find out what are the most main factors regarding aeronautical accidents

Data assumption:

- All *** values mean that data is not available (missing data)

Solution Strategy

Solution Strategy: IoT method

INPUT → **OUTPUT** → **TASK**
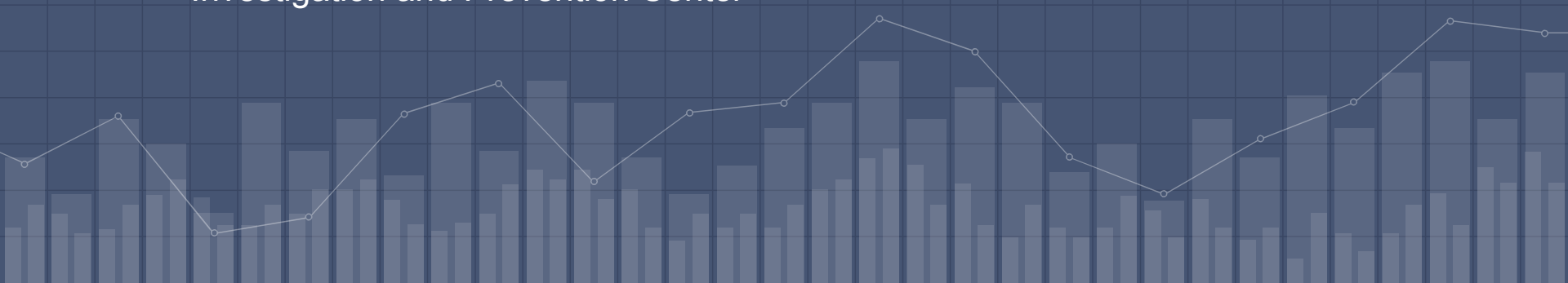
-> Business problem

-> Business questions

-> Available data

-> What is the best solution display
for the final user ???
E.g. dashboard, email, spreadsheet....

-> What is the best deployment format
for the final user ???
E.g. API, online dashboard, Telegram bot...

-> Is this solution straightforward enough
to be readily used by the final user ???

-> Does this solution solve
the business problem ???

-> What is the step-by-step to solve
each business question???

-> What tools do we need to use ???
E.g. Python, SQL, Spark...

-> What tools does the company use ???
E.g. Python, Postgresal...

IoT method

Input

- **Business problem**: data exploration to raise intelligence from available data
- **Business questions**: what are the main factors associated with aeronautical accidents that could be used in public campaigns to reduce the number of aeronautical accidents
- **Available data**: a public database from the Brazilian Aeronautical Accident Investigation and Prevention Center

IoT method

Output

- A **presentation** with:
  - concise explanations about the process used to explore the data available
  - what were the raised and the validated hypotheses
  - what were the concluding insights
- A **Github repository** with the **code** used

IoT method

Task

What are the main factors related to aeronautical accidents:

1. Understand what is the main goal of this project
2. Define the scope of the solution (for this project cycle)
3. Understand the data available (at a high level)
4. Load the data and merge the required tables
5. Clean and prepare the data for analysis
6. Define hypotheses that will be tested in data exploration
7. Search for misleading data (regarding business understanding)
8. Explore the data to get insights
9. Prepare the storytelling regarding what is required for the final product

Hypotheses Mind Map
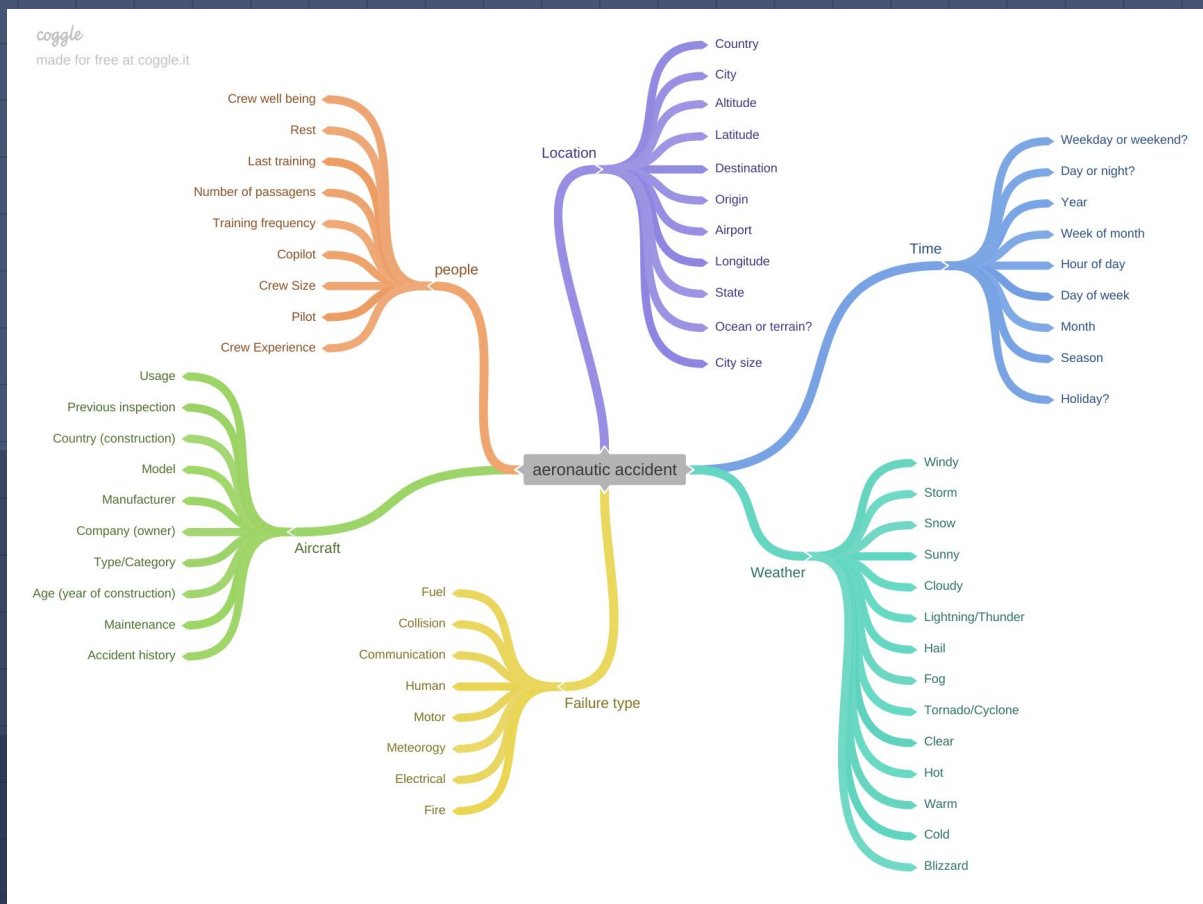
Hypothesis Mind Map

- The hypothesis mind map is the product of **a brainstorm that took into consideration factors that could contribute to an aeronautic accident**.

- This mind map is a great help when trying to **raise hypotheses** that could **lead to insights**. It is also helpful to **guide feature engineering** (create new relevant features) and when there is a need to **look for more data elsewhere**.

Code (notebook) structure

- Step 01. Business Problem
- Step 02. Problem Understanding
- Step 03. Imports
- Step 04. Data Extraction
- Step 05. Data Description
- Step 06. Feature Engineering
- Step 07. Data Filtering
- Step 08. Exploratory Data Analysis
- Step 09. Deploy
- Step 10. Restart the cycle

Top 3 Insights

Hypothesis I: the number of accidents in the state with most accidents is at least 50% higher than the average of the remaining states.
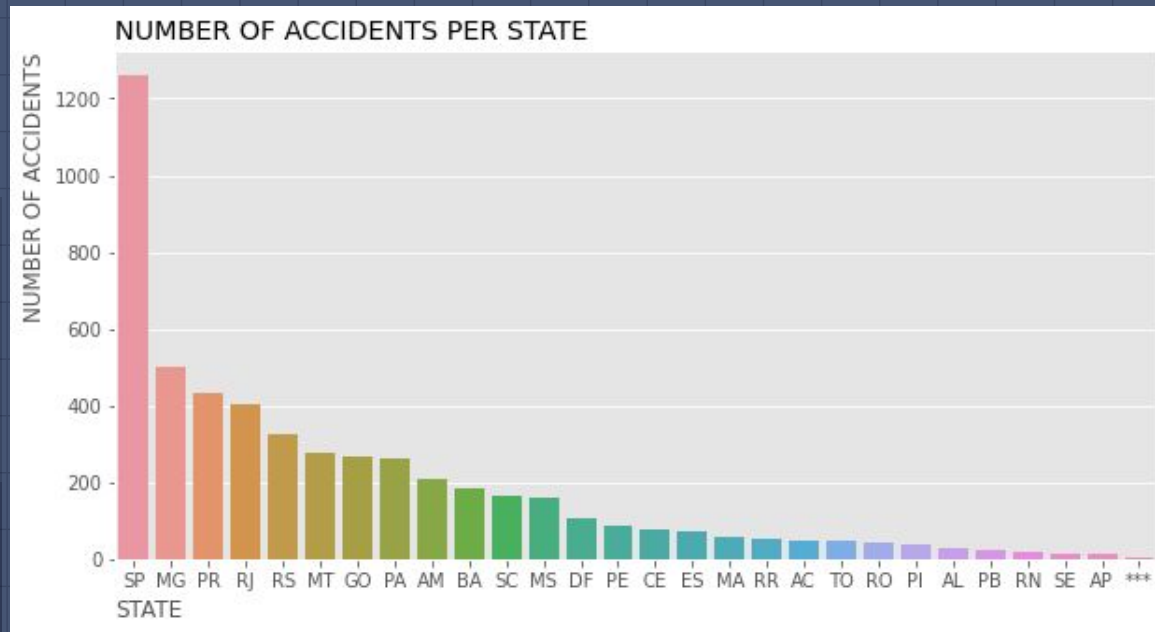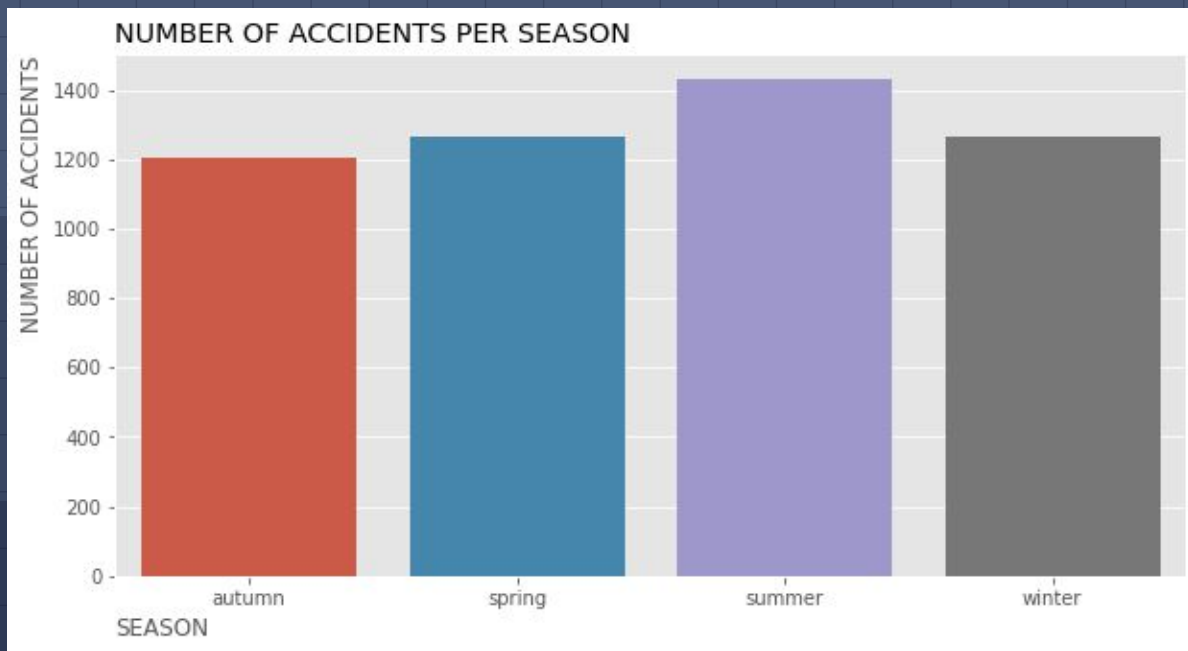
Top 3 Insights

- *Hypothesis I is TRUE.* SP is the state with the highest number of accidents and the number of accidents in SP (alone) is 8.72 times the average number of accidents in all remaining other states.

- *Possible actionable:* SP is a key state when trying to reduce the total number of accidents in Brazil. As a first solution, the testing campaigns ought to be tested in this state (as the campaign performance could be more readily apparent).

Top 3 Insights

Hypothesis II: summer has a number of accidents at least 10% larger than the average of remaining seasons.
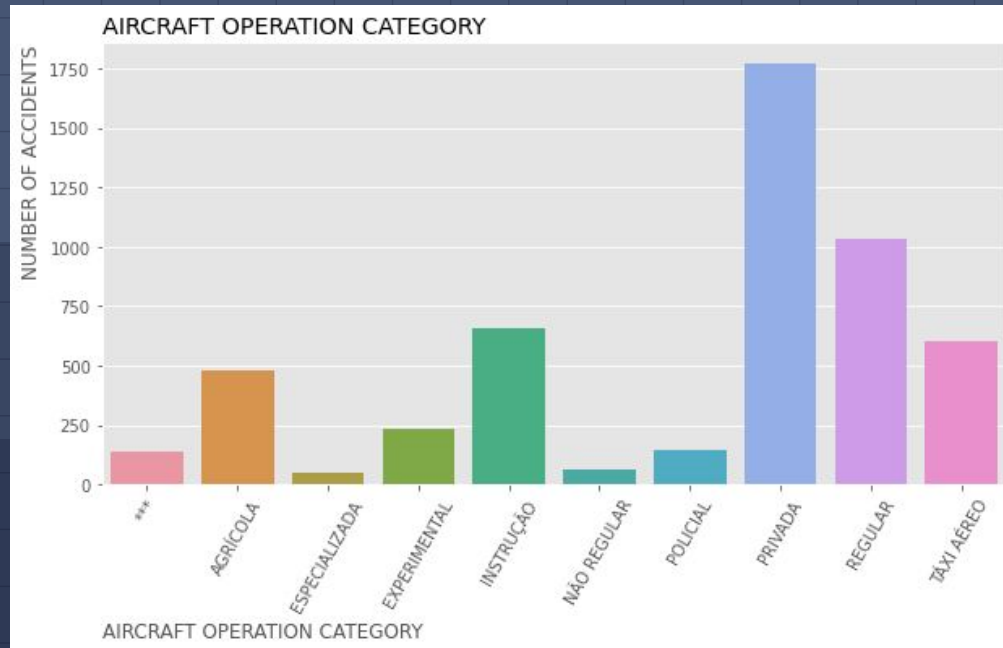
Top 3 Insights

- **_Hypothesis II is TRUE._** The number of accidents in the summer is about 15% higher than the average number of accidents in the remaining seasons.

- **_Possible actionable:_** _There is some difference between the number of accidents in different seasons; however, this difference is not very large. This could indicate that campaigns ought to be consistent throughout the year in order to achieve meaningful results in terms of accident reduction._

Top 3 Insights

Hypothesis III: the number of accidents with private aircraft is more than 70% of the total number of accidents.

Top 3 Insights

Hypothesis III is FALSE. The number of accidents with private aircraft is about 34% of the total number of accidents

Possible actionable: *Private aircrafts alone correspond to more than 30% of the total number of accidents. So, campaigns ought to also focus on private aircraft and not only 'companies/business' aircraft to achieve meaningful results.*

Business solution

- the **code** used for this case on **GitHub**

- a pdf **presentation** with:
  - (1) the presentation of the **challenge**;
  - (2) an explanation of the **process** used;
  - (3) **hypotheses raised**;
  - (4) **exploratory analysis**; and
  - (5) **conclusions** and **insights** generated

Conclusions

- Not all data science projects require modelling, this job application case is a clear example.

- Doing a project with a very strict deadline (48 hours) not only requires good a **time management** skills but also a good **task prioritization** skills.

- Without a clear **business problem** and **business context**, it is much harder to **guide** what are the best **decisions** to make, and what **assumptions to raise** and **explore** the data to achieve **meaningful insights**. To overcome all of this, we decided to create a business context around this project.

Lessons Learned

- **Task prioritization** due to the very strict deadline of this case (48 hours)

- How to **set** a **business context** for an open case to keep the solution as close as possible to a real scenario and also to have a **guide** to make **assumptions** and make **decisions** throughout the project.

- How to do a data science project where the deliverable is not a model to make some predictions.

- How to plan the **storytelling** to make the project accessible to others.

Next steps to improve

- **Exploratory Data Analysis:** raise more hypotheses to validate and explore the relationship between variables deeper.
- **Business Understanding:** collect feedback from the last project cycle and use it to improve business understanding. Revise the planned solution for this project (according to the IoT method) and update it according to feedback.
- **Database:** research for a better description of every column in the database to further understand (and use) available data.
- **Feature Engineering:** explore feature creating to find out new relevant features (and metrics) for the project solution.
- **Storytelling:** get feedback for the first presentation and use it to improve the storytelling template for the next cycle.

Thanks!

- Firstly, thanks for the opportunity of joining this admission process.

- Secondly, thanks for giving me the opportunity of going so far in this admission process.

- Last but not least, I'd like to thank you for being a part of my first technical interview in data science [I will always remember these moments...]

Contact

- Portfolio page: https://ds-gustavo-cunha.github.io/projects-portfolio/

- Linkedin: https://www.linkedin.com/in/ds-gustavo-cunha/

- Github: https://github.com/ds-gustavo-cunha

- Medium: https://medium.com/@ds-gustavo-cunha

- Email: gcunhaj@gmail.com

Questions?