The beginner's data science project checklist

Sara Iris Garcia

Content

- Defining requirements
- Outlining a data science project
- Reproducibility and Readability checklist
- Best practices for writing Documentation
- Useful python tools
- Tips on presenting your findings

About me

Sara Iris Garcia

- Born and raised in Guatemala
- Computer science engineer, data analyst by trade
- Community leader: Women in Data Guatemala City chapter
- Still a bit shy -please be patient :)

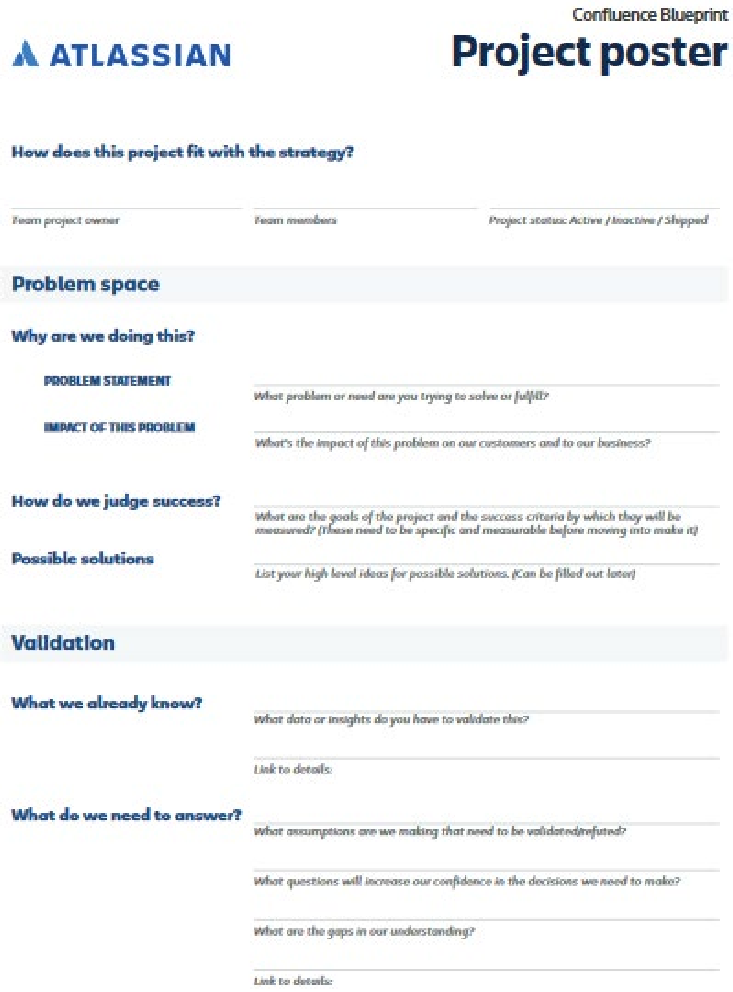
The beginner’s data science project checklist

- Before – During – After

Defining Requirements

In this step we describe what we need to do and why we are doing it.

- Understand the problem is key
- Define the objectives and goals as clear as possible
- Identify possible roadblocks or challenges we foresee
- Define the metrics we are going to evaluate the project



Confluence Blueprint

Project poster

How does this project fit with the strategy?

Team project owner

Team members

Project status: Active / Inactive / Shipped

Problem space

Why are we doing this?

PROBLEM STATEMENT

What problem or need are you trying to solve or fulfill?

IMPACT OF THIS PROBLEM

What's the impact of this problem on our customers and to our business?

How do we judge success?

What are the goals of the project and the success criteria by which they will be measured? (These need to be specific and measurable before moving into make it)

Possible solutions

List your high level ideas for possible solutions. (Can be filled out later)

Validation

What we already know?

What data or insights do you have to validate this?

Link to details:

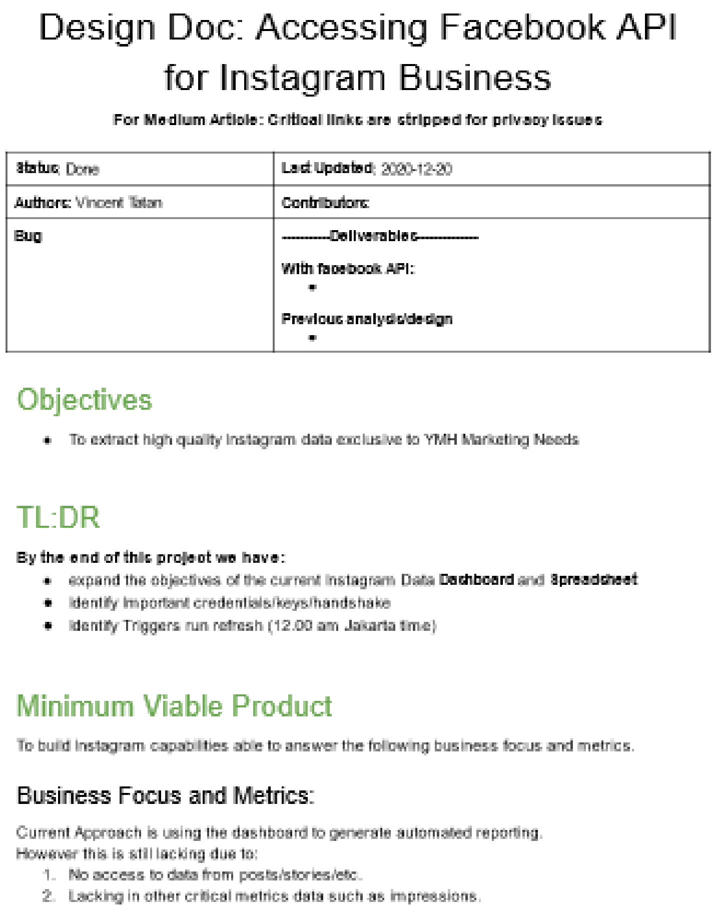
What do we need to answer?

What assumptions are we making that need to be validated/refuted?

What questions will increase our confidence in the decisions we need to make?

What are the gaps in our understanding?

Link to details:



Design Doc: Accessing Facebook API for Instagram Business

For Medium Article: Critical links are stripped for privacy issues

Status: Done	**Last Updated:** 2020-12-20
Authors: Vincent Tatan	**Contributors:**
Bug	**-----------Deliverables-------------** **With facebook API:** • **Previous analysis/design** •

Objectives

- To extract high quality Instagram data exclusive to YMH Marketing Needs

TL:DR

By the end of this project we have:

- expand the objectives of the current Instagram Data **Dashboard** and **Spreadsheet**
- identify important credentials/keys/handshake
- identify Triggers run refresh (12.00 am Jakarta time)

Minimum Viable Product

To build Instagram capabilities able to answer the following business focus and metrics.

Business Focus and Metrics:

Current Approach is using the dashboard to generate automated reporting.
However this is still lacking due to:

1. No access to data from posts/stories/etc.
2. Lacking in other critical metrics data such as impressions.

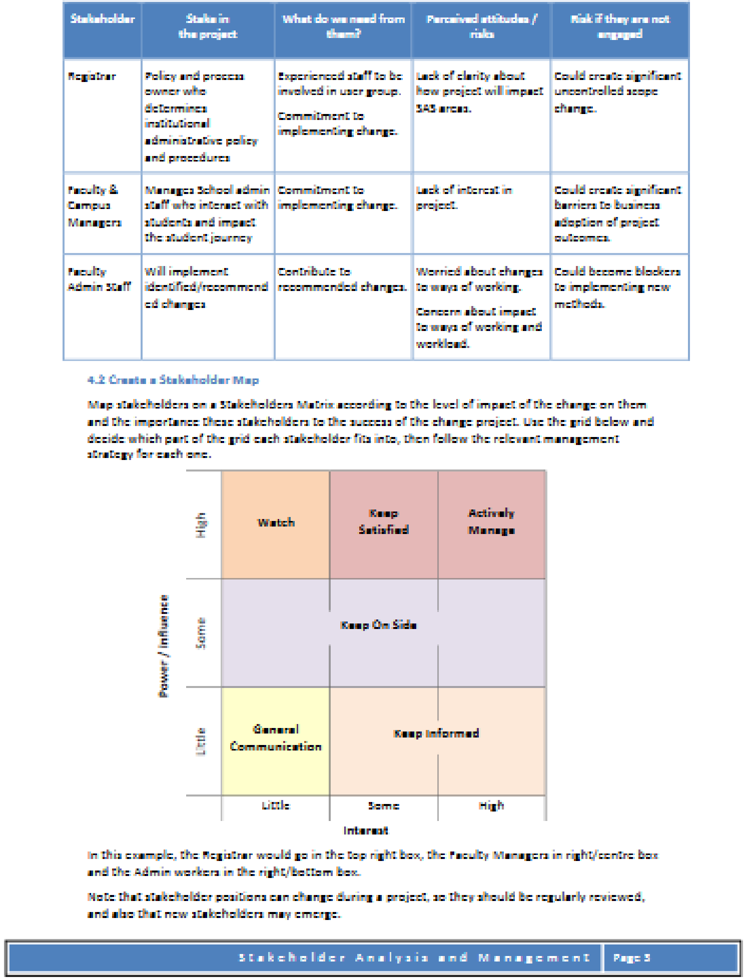
Useful Links

- Project poster template
- Design doc example(Medium blog)
- User stories

Outlining the project

Once the requirements and goals are clearly defined (the what and the why) then we describe how we are going to do it.

- Describe the deliverables
- Identify milestones
- Define the timeline
- Describe the data
- List resources and tools
- Describe the implementation



Stakeholder	Stake in the project	What do we need from them?	Perceived attitudes / risks	Risk if they are not engaged
Registrar	Policy and process owner who determines institutional administrative policy and procedures	Experienced staff to be involved in user group. Commitment to implementing change.	Lack of clarity about how project will impact SAS areas.	Could create significant uncontrolled scope change.
Faculty & Campus Managers	Manages School admin staff who interact with students and impact the student journey	Commitment to implementing change.	Lack of interest in project.	Could create significant barriers to business adoption of project outcomes.
Faculty Admin Staff	Will implement identified/recommended changes	Contribute to recommended changes.	Worried about changes to ways of working. Concern about impact to ways of working and workload.	Could become blockers to implementing new methods.

4.2 Create a Stakeholder Map

Map stakeholders on a Stakeholders Matrix according to the level of impact of the change on them and the importance these stakeholders to the success of the change project. Use the grid below and decide which part of the grid each stakeholder fits into, then follow the relevant management strategy for each one.

Power / Influence

	Little	Some	High
High	Watch	Keep Satisfied	Actively Manage
Some	Keep On Side		
Little	General Communication	Keep Informed	

Interest

In this example, the Registrar would go in the top right box, the Faculty Managers in right/centre box and the Admin workers in the right/bottom box.

Note that stakeholder positions can change during a project, so they should be regularly reviewed, and also that new stakeholders may emerge.

Stakeholder Analysis and Management | Page 5

Each set of analysis will likely need to be validated. Initially, this may be through historical data, and eventually, through some type of a field trial.

	Analysis 1	Analysis 2	Analysis 3
What is the type of analysis? *e.g. description, prediction, detection, causal inference*			
What is the purpose of this analysis? *e.g. understand historical behavior of individuals, estimate risk of disease, identify which actions will increase graduation rates amongst students*			
Which action will this analysis inform? *eg. inspections of compliance regarding handling of hazardous materials*			
How will you validate this analysis using existing data? **What methodology and what metrics will you use?** **How will you compare against existing baselines?** *e.g. creating multiple train and test sets based on time, using precision or positive predictive value at top 10% as a metric, and comparing against random and "existing system" baselines*			

Useful Links

- Data science scoping document
- Stakeholder analysis toolkit
- Data science project template

Reproducibility and Readability checklist

70% of researchers have failed to reproduce another scientist’s experiments, and approximately 60% have failed to reproduce their ownexperiments - Nature’s Survey (2016)

Reproducibility is to ensure that our results can be independently verified and replicated. In involves sharing the data, methodology and code that were used to produce the results.

Checks

- Don't do things manually (i.e. edit data to remove outliers by hand, format data files)
- Use a data validation tool (Great Expectations, Pandera, Pydantic)
- Use a version control (Github, Bitbucket, etc)
- Keep a record of your hardware & software environment (docker, conda env, virtualenv, etc)
- Use test automation tools (Hypothesis, UnitTest, PyTest)
- Set the seed number

Useful Links

- Reproducibility and Why it Matters
- Code Review
- Reproducible research checklist
- Guide for reproducible research

Best practices for writing Documentation

- Use linters (pylint, flake8, prospector)
- Agree on a code style (PEP8, Google style guide)
- Use docstrings to document your functions
- Document your API (Swagger, Sphinx) – *or better, use FastAPI!*

Useful python tools

- CookieCutter
- Mlflow
- DVC
- FastAPI
- Streamlit, Dash

Tips on presenting your findings

- Know your audience
- Create an executive summary (template guide)
- Use bullet points
- Highlight the performance of the deliverables and the main points delivered
- Describe how the project fit in the business planning
- Provide regular updates on the current status of the project
- Discuss the recommended actions to take based on the outcome of the project, their benefits, risks and consequences
- Briefly discuss the results and findings

Conclusion

- Structure from the start
- Document every stage
- Briefly report your findings and progress
- Less jargon and more visuals

Thank you

Sarairis.garcia@gmail.com

linkedin.com/in/sarairisgarcia

https://github.com/montjoile/Europython -2022