DATA QUALITY AND DOCUMENTATION

- *Set data quality check pipeline*
- *Improve data documentation*

SET DATA QUALITY CHECK PIPELINE & IMPROVE DOCUMENTATION

Problem Statement

- Repeated questions from various constituents
- Non - automated data quality checks
- Delayed information on data quality issues to stakeholders

Objective

- Automate data quality checks
- Inform stakeholders of reported issues
- Improve data documentation for constituents to make information easily accessible and interpretable

SET DATA QUALITY CHECK PIPELINE & IMPROVE DOCUMENTATION

Action

- Use **approved building permits** dataset
- Browse dataset rows and cols and populate potential data checks

- Understand the structure of great expectations library (python compatible) used for conducting data quality checks

- Create a draft data quality checklist compatible with the library to iteratively finalize data quality checks to run with the client (output: link)

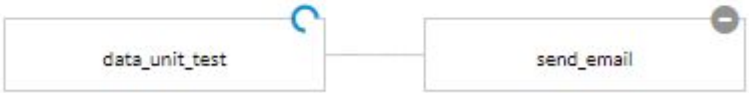
- Use the above document to generate an expectation/ data quality suite (json script) to run the data quality check (created via jupyter notebook)

SET DATA QUALITY CHECK PIPELINE & IMPROVE DOCUMENTATION

Action

- Fed the expectation suite and the batch data to run the data unit test script template
- Returns a data quality report generated by the GE library (output: link)

- If error reported in the report, send automatic emails to relevant stakeholders



- Updated the website with more detailed information on building permits with data definitions, data types, values and more details on types of permits present

SET DATA QUALITY CHECK PIPELINE & IMPROVE DOCUMENTATION

Action

- Used google sheet to coordinate with client on what checks to run

  Advantages

  - Good tool to collaborate and make changes
  - Allows multiple stakeholders to engage in the same document

  However, ...

  - Could not be directly fed into the workflow
  - Easy to overwrite changes with multiple stakeholder engagement
  - Difficult to save so many sheets on git or cloud

SET DATA QUALITY CHECK PIPELINE & IMPROVE DOCUMENTATION

Action

- Used google sheet to coordinate with client on what checks to run

  Advantages

  - Good tool to collaborate and make changes
  - Allows multiple stakeholders to engage in the same document

  However, ...

  - Could not be directly fed into the workflow
  - Easy to overwrite changes with multiple stakeholder engagement
  - Difficult to save so many sheets on git or cloud

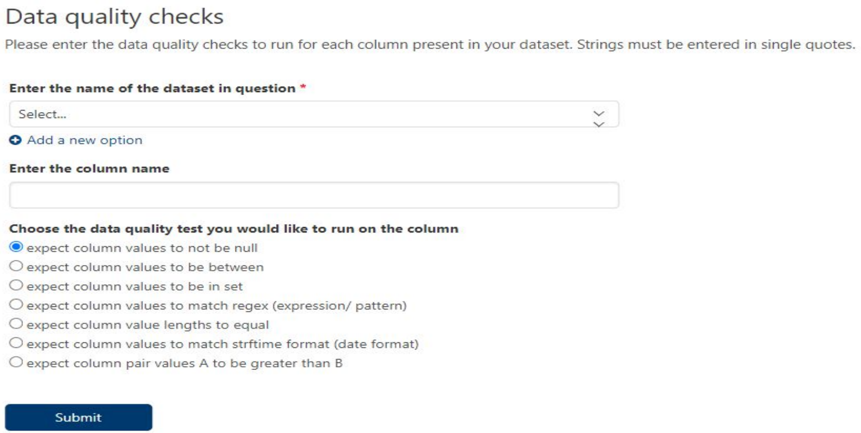
- We tested if Knack app would be a potential alternative to collaborate

  - Allows to take the advantages of google doc
  - Allows stakeholders to interact using Knack Forms
  - Allows them to update changes in the table directly
  - Can authorize users to certain views
  - Can directly connect it to workflow
  - Could be potentially used to generate expectation suite directly & automatically
  - Users/ owners could update checks without disturbing the workflow

While this is still internally being discussed and other platforms are being explored, it has the potential to automate data quality pipeline right from Step 1 of client engagement.

SET DATA QUALITY CHECK PIPELINE & IMPROVE DOCUMENTATION

Action



Data quality checks

Please enter the data quality checks to run for each column present in your dataset. Strings must be entered in single quotes.

Enter the name of the dataset in question *

Select...

Add a new option

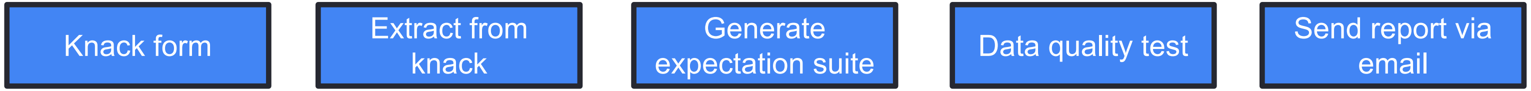
Enter the column name

Choose the data quality test you would like to run on the column

● expect column values to not be null
○ expect column values to be between
○ expect column values to be in set
○ expect column values to match regex (expression/ pattern)
○ expect column value lengths to equal
○ expect column values to match strftime format (date format)
○ expect column pair values A to be greater than B

Submit

Link



Knack form

Extract from knack

Generate expectation suite

Data quality test

Send report via email

SET DATA QUALITY CHECK PIPELINE & IMPROVE DOCUMENTATION

Impact

- **Documentation**: Makes data more accessible and interpretable to end users/ constituents/
- **Data quality**: Enable stakeholders to take prompt actions to improve data quality
- **Data collaboration**: Enables stakeholders to work more collaboratively and smoothly to both create and update data quality checks

  Scalability: Potential to plug in the pipeline to other datasets, potentially before data are published in Analyze Boston by running the tests on internal data

A BIG SHOUT OUT TO BOSTON CITYWIDE ANALYTICS TEAM!

Especially Kevin, Andrew, Daniel, Courtney and John and others

- For always being so approachable and ready to help!
- For providing a great learning ecosystem despite a remote internship and despite having to transition back to office.

YOU ALL ARE GREAT MENTORS!

- Thanks to Kelly from the Inspectional Services department for always providing prompt feedback on all outputs and encouraging to take ownership of the work

- Thanks to the **awesome** Git Book documentation! ☺️ (I couldn’t have survived the fellowship without it!)

THANK YOU