‘CleanlockHolmes’, Unraveling Data Mysteries with Python

**Rebecca Bronfeld, Sarah Gore, Kieran Black**

<https://github.com/kierblk/NU-DS5010-Project>

**Summary:** Summarize the overall purpose of the package (i.e., what problem does it solved), any related work or similar libraries, and a brief, non-technical description of any notable modules, classes, or functions.

**Project Purpose:**

What is clean data , how do you get data sets that aren’t “clean”, why do we want a package that can do something about it

**Related Work**

Identify similar libraries such as pandas – using pandas to help with underlying data frame not utilizing their own built in cleaning functionalities

**Structure Overview**

Single class that creates an instance of the data wrapped in the object and interact with that via a number of methods that identify what is considered “invalid” or “valid” data

**Design:** A technical description of the modules, classes, and functions implemented in the package. This does not need to be exhaustive, and should only cover the most important aspects of the package for a user to understand. The description should be clear and concise so that a user can understand the overall design and organization of the package.

Note design choices that were made and justify why

Explain that backbone of class is cleanloclholmes object. One to one mapping between a dataset and an object . represents only a single cleaning project.

Break down what the methods are within the class by organizing them by which stage ion the workflow they belong in .

Eg creating the object specifying a notion of valid values, running a search for invalid values, running a method to remediate these values and then running a method to produce the data as an output file.

Specify which methods are public and private ( public – user calls direct, private – internal to the class)

What are class attributes and how is it doing the job / justifying why I stored what I did as a class attribute

Eg :

self.file\_name = file\_name

self.data\_object = self.read\_data(file\_name)

* Pandas data frame – why is it appropriate to use a pandas data fram e

self.columns = self.data\_object.columns

self.invalid\_dictionary = {}

* Why have a valid and invalid dictionary -> why are they structured this way

self.valid\_dictionary = {}

self.col\_types\_dictionary = {}

self.data\_ranges = {}

**Usage:** Description and examples of how the package should be used, how it fits into a workflow. This section may include code snippets, but it should primarily be a written description.

Insert snippets of user workflows (cleaned up and concise)

Write out workflow 1

5 step process :

Create object

Specific valid?invalid

Clean using method of choice

Use and output

1 interactive

2. non interactive

- why? Give the user the option to customize . some users will like that interface and involves writing less code but other uses prefer to write the specification themselves incase they want to reproduce it .

**Discussion:** Discuss how the package compares to related libraries, and how it fits into the overall ecosystem. Why should people use this package? How could the package be improved?

REBECCA TO DO

**Statement of contributions:** List the full names of the authors and how each member contributed to the completion of the project.

TO BE FINALIZED

**References:** Cite any external libraries used by the project, and any sources that were used as a reference. Use a consistent format and numbering scheme

REBECCA TO DO