[[1]](#footnote-2)

Data Cleaning Using Python to Reduce Manual Inspection

Zackary Gill (zgill@smu.edu) and Mai Loan Tran (mltran@smu.edu)

*Abstract*—A key practice of successful data cleaning is to utilize tools and programming efforts to reduce manual inspection. Company X requires monthly reporting to its senior management for travel and reimbursed expenses by vendor. The vendor names need to be standardized to satisfy the reporting requirements. These vendor names come from multiple sources with different representations that overlap and contradict one another. While Company X has been able to develop a database model to manage different names for the same object on the schema level, data conflicts persist on the instance level with different representations of a vendor name resulting in inconsistent aggregating of spending for specific vendors. Although Company X has a designated resource to clean this data manually using SQL DML, this project seeks to provide a repeatable alternative in Python to reduce the manual inspection and manipulation of the variations in vendor names.

*Index Terms*—TBD.

# INTRODUCTION

T

HIS document describes the use of Python to perform data cleaning towards reducing manual inspection and manipulation of data in standardizing vendor names from multiples sources. The requirements for data cleaning increases significantly when data is sourced from varying credit card processors, agencies, and free text input from numerous users, where only a small subset of values come from lists. Manual inspection to find all the representations of a vendor name then manipulating a range of 21,000 to 132,000 tuples using SQL statements to standardize takes 2-3 days per month. The first objective of this project is to determine whether Python is an effective alternative to reduce the current duration of data cleaning enabling more time for analysis. The second is to create a repeatable method to clean and standardize the data.

# METHODOLOGY

This project plan in which deliverables will be created then subsequently submitted are given below.

Week 5: Data exploration and resources identified.

Week 6-7: Python solution in development.

Week 8: Lightning Talk given during class time.

Week 10: Python solution testing completed.

Week 12: Complete draft due.

Week 15: Final presentation.

Week 16: Final paper due.

As seen above this research is a 12 week project with 7 steps along the way. Week 5 will consist of data exploration where we will look at the available data and see what if any other resources are required. This involves planning out the specific portions of the Python solution. In weeks 6-7 a working skeleton of the project will be created that will allow the discussion of the topic. A five minute presentation of this project will be presented in Week 8 to give our solution approach. Week 10 will test the solution for preliminary results and viability of the repeatable method. Week 12 will see the draft of the project towards the final paper to be delivered in Week 16. Week 15 includes a fifteen minute final presentation to demonstrate our solution. Each milestone will ensure that the research is done consistently with quality.

In data cleaning literature there are many different criteria for data including validity, integrity, precision and accuracy, consistency, and uniformity. The definitions of these vary but this project will ensure that these are carried out in the cleaning process. For the actual cleaning comparison we have procured a dataset from Company X with approximately 500k rows. The data to be cleaned will be stored in a MySQL database which will be accessed from a Python script. The data manipulation within Python will be done using the pandas library. Other libraries or tools may in the future be added to this list as the need for them is found. To improve the understandability of the process the modification history will be logged. Finally, the resulting solution will be compared to the manual SQL table manipulation. The comparison will include the difference in the data cleaning criteria and in the time taken for the process.

1. This draft is submitted on February 3 2019. It is a proposal for the MSDS 7330-405 class project.

   This work is supported by non-financial data provided by Company X, whose name shall not be disclosed in publications without written approval from its Legal representatives.

   Z. Gill and M. Tran are graduate students of the Master of Sciences Data Science Program within Southern Methodist University, Dallas, TX. (e-mail: zgill@smu.edu; mltran@smu.edu). [↑](#footnote-ref-2)