# Sprint 1 User Story Progress Review for Sprint 2 Deliverables

*$ values may add to over $100 as acceptance criteria can be combined.*

## Story 1

*Story:* As a user I want to be able to upload PDFs containing tables and extract the data from it, so that I don't have to manually enter data into my computer.

*Acceptance Criteria:* users can drag and drop or select PDFs from their file system, upload them and receive the same data in a digital format.

*$ Value Test:* $10

**Review:** a user can use our software to upload a PDF document, and the software will scan it using OCR then output the data extracted into a structured CSV file format.

**Evidence:**

**Issues/Challenges in Completion:** N/A

**Solution:** N/A

**Status:** expected to be delivered successfully in sprint 2

## Story 2

*Story:* As a researcher I want data to be rearranged in a usable, malleable format, so that I can perform data analysis on the information.

*Acceptance Criteria:* extracted data output in a csv file format, restructured into the original table structure from the PDF document.

*$ Value Test:* $20

**Review:** the software currently generates a CSV file as output from the OCR process, where the data is structured the same was as the original document allowing a user to process the data digitally

**Evidence:**

**Issues/Challenges in Completion:** N/A

**Solution:** N/A

**Status:** expected to be delivered successfully in sprint 2

## Story 3

*Story:* As a user I want to make sure that our information is 99% accurate, so that I don't have to spend so much time manually checking data.

*Acceptance Criteria:* OCR library has a requirement to achieve a 99% of accuracy. Employ multiple OCR models.

*$ Value Test:* $30

**Review:** software implementing several OCR models and uses comparison verification to ensure high degree of accuracy, implements confidence scores to determine accuracy of the OCRs too.

**Evidence:**

**Issues/Challenges in Completion:** N/A

**Solution:** N/A

**Status:** expected to be delivered successfully in sprint 2

## Story 4

*Story:* As a user I want to make sure errors in extraction are visualised, so that I don't have to manually review extracted data.

*Acceptance Criteria:* detected errors in extraction of the data must be flagged for the user to manually review.

*$ Value Test:* $10

**Review:** software currently outputs the cells which have been determined to have a low confidence score after being processed by OCR. There is not yet a user-friendly GUI output for this story yet, only a command line interface output.

**Evidence:**

**Issues/Challenges in Completion:** output of extraction errors is shown in the command line rather than in a user-friendly GUI

**Solution:** team is building the GUI currently to support the visual display in a user-friendly way

**Status:** command line functionality expected to be delivered in sprint 2, GUI visualisation may need to be revaluated and moved to sprint 3 deliverables

## Story 5

*Story:* As a user I want to extract data from many PDFs at once, so that I don't have to scan 100s of documents individually.

*Acceptance Criteria:* allow users to upload many PDF documents at once for batch processing.

*$ Value Test: -*

**Review:** client has suggested the process of extracting data from PDF documents will occur in a stream lined process with user's at several points of the system checking the software, therefore it is not necessary for a user to upload many PDFs at a single time, rather they will upload one PDF at a time and then the next after reviewing the first.

**Evidence:**

**Issues/Challenges in Completion:**

**Solution:**

**Status:** can potentially be shifted to be delivered in sprint 3 if project is progressing ahead of expectations

## Story 6

*Story:* As a user I want to view relevant tables about census data, so that I don't have to manually search for the information I'm looking for.

*Acceptance Criteria:* arrange extracted data into logical structure (original tables), allow it to be searchable (include table title in csv)

*$ Value Test:* $20

**Review:** extracted data is structured in the same logical format as the original tables from the uploaded PDF, a specific search tool is out of scope for the project, however, including the table title in the output CSV file can be investigated

**Evidence:**

**Issues/Challenges in Completion:** automatic table detection errors can result in columns of some tables being excluded in the final output CSV file

**Solution:** team is prioritising automatic table detection during sprint 2

**Status:** on track to be delivered in sprint 2

## Story 7

*Story:* As a future developer I want appropriate documentation, so that I can improve and maintain the system.

*Acceptance Criteria:* technical documentation for the software should be at least minimal.

*$ Value Test: -*

**Review:** documentation has not been prioritised by the team due to client $100 test indicating it is not a priority deliverable for the client. Documentation for code will be written, however.

**Evidence:**

**Issues/Challenges in Completion:**

**Solution:**

**Status:** on track for completion during sprint 2

## Story 8

*Story:* As a user I want an easy way to install the software onto my computer, so that I don't have to learn how to use a command line interface.

*Acceptance Criteria:* create installation software so that a user can directly download it rather than manually from GitHub.

*$ Value Test:* $10

**Review:** out of scope for sprint 2,

**Evidence:**

**Issues/Challenges in Completion:** functionality of the software has been prioritised by the team during sprint 2, it is not important for sprint 2 that the software can be demonstrated on the client's choice of hardware.

**Solution:** software will be demonstrated on the development platform, story will be delivered in sprint 3.

**Status:** moved to be delivered in sprint 3

## Story 8

*Story:* As a user I want the tables I upload to be automatically detected, so that I don't have to manually specify every row and column for the 1000s of pages I upload.

*Acceptance Criteria:* implement automatic table detection so that users don't have to specify table rows and columns.

*$ Value Test:* $20

**Review:** table detection uses luminosity values in the PDF documents to detect rows and columns, errors occurring in detection which may result in OCR extraction issues.

**Evidence:**

**Issues/Challenges in Completion:** errors occurring in detection which may result in OCR extraction issues.

**Solution:** team putting priority and assigning more team members to work on automatic table detection so it can be delivered during sprint 2

**Status:** functionality can be implemented for sprint 2, client will be asked to review PDF document inputs to check if document quality will be an ongoing issue. Improvements expected to be made during sprint 3.