Create a report containing the following information:

1. **A description of what your data is, and how you prepared it.**
    a. I looked at a lot of different datasets, searching for things like “transaction” and “receipts”
    b. Lots of them were about land or government taxes
    c. Ultimately, I downloaded search results from ACM for several different terms to get a broad span of the literature
    d. After deleting duplicates, there are about 20,300 different entries
    e. Preprocessing included the following steps:
        i. Merge the data from different search terms into one file
        ii. Divide the authors’ names up into separate data points
        iii. Clean the results:
            1. The downloaded data didn’t handle Unicode characters well (e.g. ü, ñ, é). As a result, some of the names got divided up. I tried to clean it as much as possible but this can only really be done by hand and as of now I have about 3000 cells left to clean. As it stands, it will run in the algorithm. It will just generate some false associations because some names will always occur with the other parts of their names that got divided up
2. **Any interesting associations you uncovered, along with possible reasons for their presence.**
3. **Recommendations for the owners of the data or other interested parties (e.g. some products that should be placed together).**

We chose to use web scraping to gather our data. After researching web scraping, we discovered Google doesn’t allow it. To get around that we used the ACM search machine for several different terms, which includes a feature to export search results to a CSV file. This showed to be very useful. By implementing this we were able to generate results providing a broad span of literature. After deleting duplicates, there were about 20,300 different entries left. Preprocessing the data consisted of first merging the data from different search terms into a single file. We then divided the authors’ names up into separate data points. Cleaning the results was a bit tricky since the downloaded data didn’t handle the Unicode characters well (e.g. ü, ñ, é). Because of this, some author names got divided up. After trying to sanitize the data as much as possible, we realized this would be most efficiently done by hand.