Senior Design Spring Presentation

AUTHOR: TED CORDONNIER

Team

- Project Name: Automated Correlation Analysis
- Individual Project by Ted Cordonnier
  - 5th Year Computer Science Major, Graduating Spring 2024
- Team: LOL
- Email: cordontd@mail.uc.edu
- Advisor: Seokki Lee

Goals

► Find Correlation between target column and all other columns in a .csv file

► Automates statistical tests of correlation as well as allows any analyst to find strength of association between all types of variables

► All without extensive statistical experience from the user

Intellectual Merits

► Myself and my advisor Seokki Lee were not able to think of or find tools that are able to automatically find correlation between variables given a .csv file input.

► Tools do exist (IBM SPSS, minitab) that allow users to run different statistical tests on a given set of data. However, it requires that the user has a solid foundation on statistics.

► Available tools do not have the ability to do do all these tests automatically

► Available tools do not have the ability to determine which test should be used based on the datatype of each column.

► Given the datatype that my program identifies, the correct correlation method is used

► Some correlation tests only identify if there is correlation and can't say anything about its strength. My program uses further tests to determine the strength of this correlation, something that I have not seen other programs do

Broader Impacts

- Correlation is useful because it quantifies the degree to which two variables are related. Understanding correlation helps in predicting one variable from another, identifying trends, and making inferences.

- Correlation is a fundamental statistical tool that enables data analysts to draw insights from data as well as guide decision-making.
  - Examples: Find out which variables impact sales the most, which variables determine to apple quality, which lifestyle factors impact cardiovascular disease, etc.

- My program allows a user without extensive statistical experience to be able to calculate correlation from the target they want to analyze to their entire .csv file.

Design Specifications

- Program was created inside of a Jupyter Notebook

- Input: .csv file

- Output: Strength of correlation between target and all other columns

- Takes the .csv file and puts it into a Pandas dataframe. Dataframe automatically determines the datatype of each column.

- Loop through each column and run one of 4 correlation tests depending on the datatypes of the target vs. column.

Design Specifications
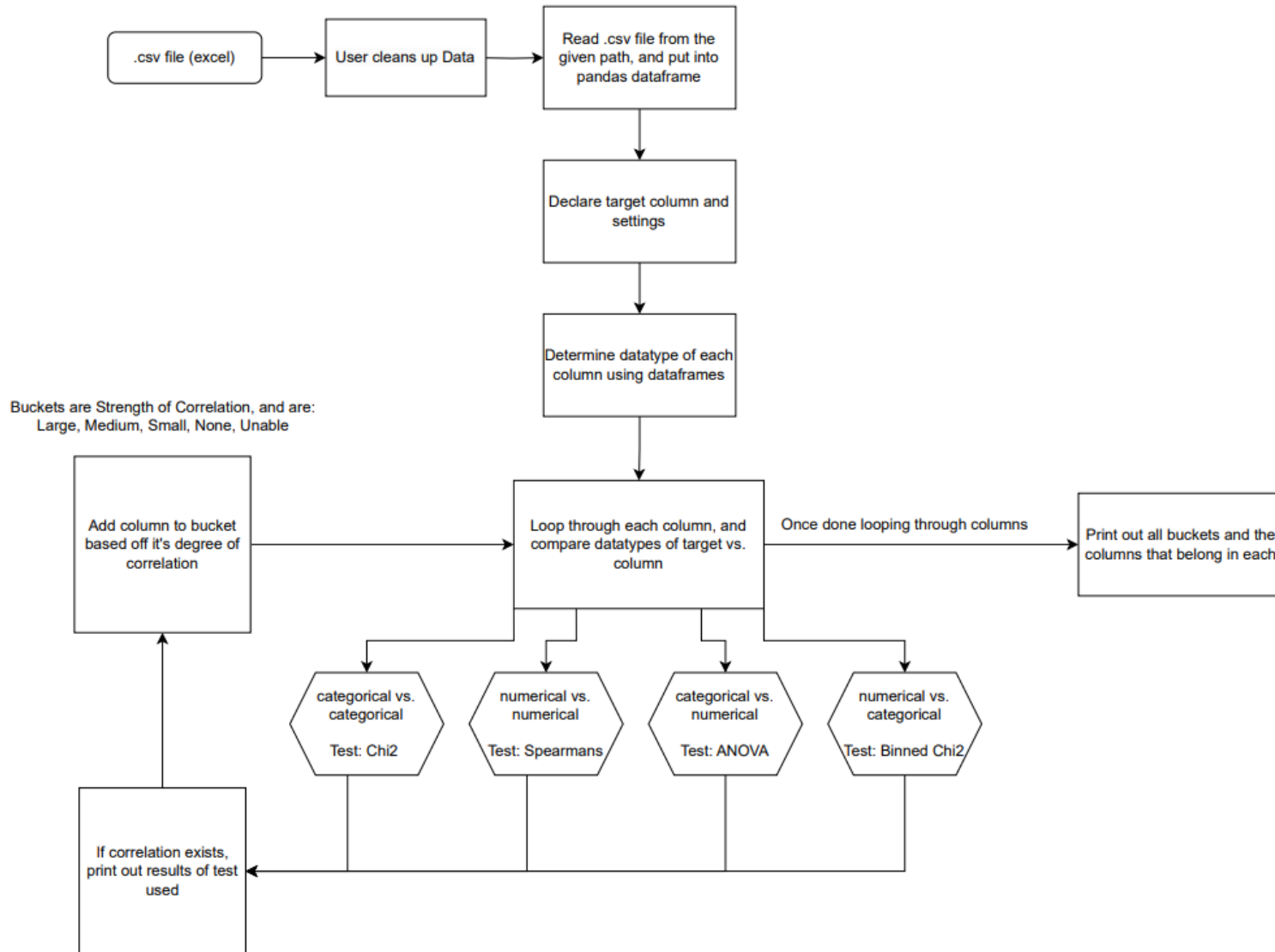
Technologies

- Jupyter Notebook
  - Python Data Science Workspace
- Git
  - Version Control
- Data Handling and Analysis Libraries
  - NumPy, Pandas Dataframes
  - Handling data from the .csv
- Statistical Analysis Libraries
  - SciPy, Statsmodels
  - Running the tests for correlation and strength of correlation

Milestones

- 1/15: Project plan, environment setup, data collection
- 1/22: Initial data analysis report outlining which types of .csv should be used
- 1/30: Research which tests should be used for each correlation datatypes
- 2/15: Prototype for functions for ANOVA, chi-square, and correlation tests integrated into the project.
- 2/25: Initial analysis of correlation results, keep working on tests
- 2/30: Ensure correlation results are correct with IBM SPSS
- 3/25: Complete final presentation, poster, deliverables, etc.
- 4/14: Final project presentation at Expo, collection of feedback, documentation of lessons learned, and discussion of potential future work

Results

- Achieved all milestones

- Currently have a working Correlation Analysis Program

- Demo will consist of showing the execution of the program live with 3 datasets I have picked

- There are things that could be added in the future, but with time constraints, this is the place that I have decided to stop at

- I am happy with how things turned out and believe that my program demonstrates the research and implementation of Computer Science/Statistics

Challenges

- Project direction changed somewhat sharply right before the end of the Fall Semester
  - My advisor and I believed that this was the correct choice and would lead to a more meaningful project
- Lots of research had to be done on statistics, correlation, etc., since I did not have an extensive background
  - I was able to learn a considerable amount from resources around the internet.
  - There was a shockingly low number of resources on programs to calculate correlation. I learned there are many formulas for calculating correlation
  - I found out that the method and formula used depends on the datatypes of the variables being analyzed