Automated Correlation Analysis

Overview

- User inputs data and the column they want to analyze
- Calculates correlations between a target column and all other columns in the data

Motivation

- Tools don't exist to automatically find correlation between variables given input data
- Allows for correlation analysis without extensive statistical experience from the user
- Available tools do not have the ability to determine which tests should be used based on the datatype of each column

Why Correlation?

- 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, guide decision making, and making inferences.
- Examples: Find out which variables impact sales the most, which variables determine to apple quality, which lifestyle factors impact cardiovascular disease, etc.

Implementation

- Created in a Python Jupyter Notebook, Statistics performed using SciPy and NumPy functions
- Data inputted by user through .csv file path, data is stored in pandas dataframe
- Correlation is categorized by comparing the results of correlation tests to the accepted cutoffs for each type of test, with these cutoffs being established by the statistical community
- Given the datatypes that my program identifies, the correct correlation tests are used
- Tests of Correlation Used: Chi^2, Spearman's, ANOVA, Binned Chi^2
- My program determines the strength of this correlation, since some correlation tests can only tell if correlation exists
- Tests of Correlation Strength Used: Cramér's V, Eta-Squared

Example Scenario

	A	B	C	D	E	F	G	H	I
1	A_id	Size	Weight	Sweetness	Crunchine	Juiciness	Ripeness	Acidity	Quality
2	0	-3.97005	-2.51234	5.34633	-1.01201	1.8449	0.32984	-0.49159	good
3	1	-1.19522	-2.83926	3.664059	1.588232	0.853286	0.86753	-0.72281	good
4	2	-0.29202	-1.35128	-1.73843	-0.34262	2.838636	-0.03803	2.621636	bad
5	3	-0.6572	-2.27163	1.324874	-0.09787	3.63797	-3.41376	0.790723	good
6	4	1.364217	-1.29661	-0.38466	-0.55301	3.030874	-1.30385	0.501984	good
7	5	-3.4254	-1.40908	-1.91351	-0.55577	-3.85307	1.914616	-2.98152	bad
8	6	1.331606	1.635956	0.875974	-1.6778	3.106344	-1.84742	2.414171	good
9	7	-1.99546	-0.42896	1.530644	-0.74297	0.158834	0.974438	-1.47013	good

Example data about Apple Quality, this data has ratings of an apple's attributes. My program will take in a target column (Quality) and then determine correlation between the target and all other columns

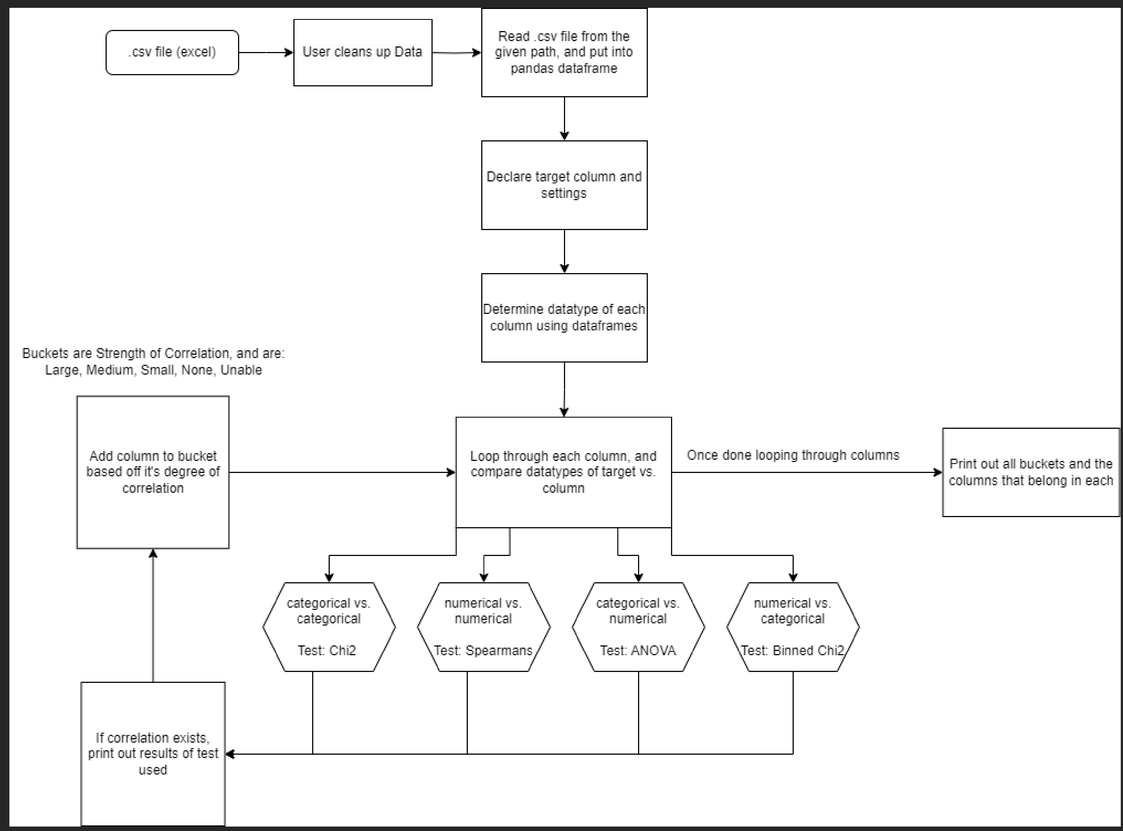
```
Test: Binned Chi-Square (X²) and Cramers V
Target: Quality (object)
Column: Sweetness (float64)

chi2: 78.444
cv: 68.669
p: 0.008
Cramers V: 0.302

Correlation: Medium

----------------------------------------------------


Test: Binned Chi-Square (X²) and Cramers V
Target: Quality (object)
Column: Crunchiness (float64)

chi2: 102.028
cv: 65.171
p: 0.000
Cramers V: 0.226

Correlation: Small
```

Outputs Results of Tests, mainly for Advanced Users

```
----------------------------------------------------


Correlation Between Target and Columns

Target: Quality

Medium
 - Sweetness
Small
 - Crunchiness
 - Ripeness
None
 - A_id
 - Size
 - Weight
 - Juiciness
 - Acidity
```

Outputs all the column names and which degree of correlation they fall into

Design Diagram



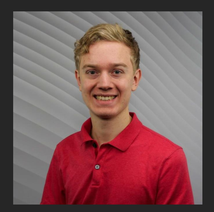
Future Work

- Integrate with a visualization tool for dynamically generated visualizations
- Support computing correlation between existing data and aggregated data values
- More extensive research and testing to better understand strengths and limitations of my approach over large data
- Developing (graphical) user interface (GUI), e.g., using webapp

Technologies

NumPy, git, pandas, SciPy, jupyter

Ted Cordonnier
Computer Science 2024
Advisor: Prof. Seokki Lee



University of CINCINNATI