**ETL Project – Technology Companies in Phoenix**

**By: Julie Staling and Rachel Groves**

**Introduction**

Arizona used to be known for the 5 'C''s: copper, cattle, cotton, citrus and climate.

Today, the greatest areas of economy come from the manufacturing of aerospace, electronics, and semiconductor. This is where technology and software come in.

**Finding Data**

We wanted to look at two separate lists:

Software companies in Phoenix (IE: Enterprise Software)

Software companies in Phoenix that have a woman as the CEO (IE: service technologies)

**The two sources we extracted from:**

1.) SaaS Companies with offices in Phoenix, AZ that have female CEO

<https://gregslist.com/phoenix/software-companies-other/woman-ceo/>

Located in Women\_CEO webscraping.ipynb

2.) Gregslist of Software Companies in Phoenix, AZ

<https://gregslist.com/phoenix/>

Located in SoftwareCompany.ipynb

**Data Cleanup & Analysis**

See Jupyter Notebook on both Women CEO and in SoftwareCompany.

In both notebooks, we have exported the data to DataFrame and then written it to CSV.

Software company Jupyter notebook is where we joined the two and cleaned the NA values.

It took quite a bit of time extracting the data from the HTML. Eventually, we found that looking within one company at a time was the best way to accurately extract info without leaving missing spaces and messing up the order of extracted info.

Here we took 3 major things: NAME, DETAILS, & WEBSITE from both the female CEO list as well as the Software company list and plan on joining them into a concise list. We anticipate there may be duplicates and we will clean those up using the Company column.

![A screenshot of a social media post

Description automatically generated]()

**Extract**

Data was in HTML and we extracted it using web scraping techniques.

You can see many of these in the notebook. It required isolating various parts and beautiful soup commands.

A screenshot of a cell phone

Description automatically generated

**Transform**

We cleaned and joined these lists using the SoftwareCOMPANY Jupyter notebook. We merged both DFs getting and preserving the integrity of the data by doing a loop that allows us to see any missing values in the data scraped.

**Load**

We then loaded our CSVs into pandas and merged them. From there, we printed the combined CSV with both sets of information.

![A screenshot of a cell phone

Description automatically generated]()

**Final Thoughts**

This list is always expanding by at least 10 companies a day, and this way we can keep adding to it.

Hopefully we can look at this info and better understand and get more familiar with the tech landscape of Arizona and know where we might be a good fit.