# Final Project - Utah Programmer Job Market Analysis

Tyler Johnston & Kaelyn Hughes

Git Repo: https://github.com/Tyler-Johnston/cs5830-final

## Introduction

For our final project, we analyzed different counties' livability based on the cost of living, the local job market, and the expected salary in the area. We want to focus on Utah and the field of programming in particular because this is what Kaelyn and I (along with the rest of our CS5830 class) are going into. We munged our own wages dataset using information from the department of workforce services (jobs.utah.gov), and we also utilized a 'comfortability' dataset which talks about the cost of living per different familial types in different counties of Utah. We were able to identify the affordability index of different job types and locations in Utah using this information, and we were able to form a compelling story informing our stakeholders about the best choices they should make to make the most money for after they graduate.

## Dataset

First, we munged a dataset from the department of workforce services. We were able to obtain one's job title, county name, hourly pay (for inexperienced devs), hourly pay (for the median dev), annual pay (for inexperienced devs), and annual pay (for the median dev). This site we obtained this from didn't show an option to download any of this information, so we manually parsed the data. However, I only grabbed information that showed information for an employer with a bachelor's degree. I decided this because our stakeholders are mainly students who would be graduating from USU with this degree type. Any associates/high school/PHD/Doctorate as their highest form of education was not considered. Thus, we considered Software Developers, Computer Systems Analysts, Computer Programmers, Software Quality Assurance Analysts and Testers, Web Developers, and Network and Computer Systems Administrators in this analysis.

The supplemental dataset we utilized was a cost of living per county in Utah based on family type (1 person 0 kids, 2 person 3 kids, etc), taxes, housing costs, childcare costs, transportation costs, healthcare costs, etc. We took this dataset as-is and utilized it for our later analyses.

We were planning on web scraping additional data. However, sites such as Indeed, Handshake, Linkedin, etc did not allow web scraping data, and it was challenging to find current, accurate datasets that contained information from these sites. Obtaining skills sought after from current employers was thus abandoned due to these constraints.

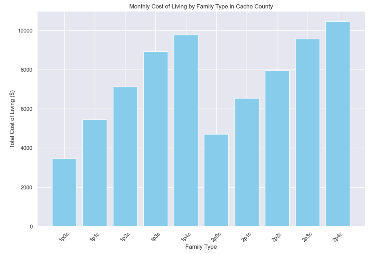
## Analysis Technique

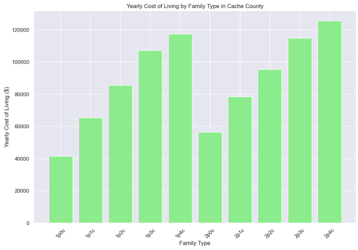
First, we tried to understand the data by plotting different family types by comparing the cost of living per county. Then we analyzed the most/least expensive counties by 'average' monthly cost. Each family type had a different monthly cost of living, so to make a comparison against other places and job types, we decided to average the cost of all family types. To get the best paying/worst paying jobs, highest cost of living/worst cost of living, and any analysis similar to this, we simply sorted the data and grabbed the first 'x' or so data points in either ascending or descending order. We utilized this technique with the other factor we were considering, whether that was job title or county.

Additionally, because the munged dataset and the cost of living dataset had different ways of describing areas/counties, we needed to map each general area to their respective counterparts during later parts of our analysis. This 'mapped' column was utilized throughout this project. After this, we computed an 'affordability index' to compare the actual cost of living compared with the average salary in that general area/county. This is utilized because, while you may be making a lot of money in a specific county, if the cost of living was really high, your take-home income is much less. We sought to find the 'true' best place to live for Utah programmers.

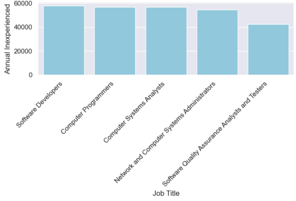
Some libraries were utilized, such as itertools to allow for easier computation of combinations of job titles, and ColumnTransformer/OneHotEncode (along with the Linear Regression built-in models) to allow for easier computation of one hot encoding for our linear regression analysis.

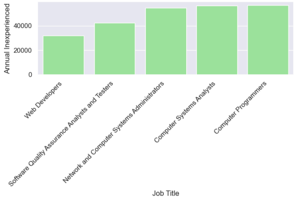
## Results

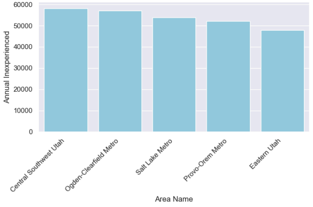


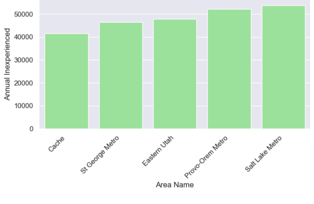


Depending on one's familial type, it costs more per month/year the more kids you have, and it decreases when you have a dual-income family.

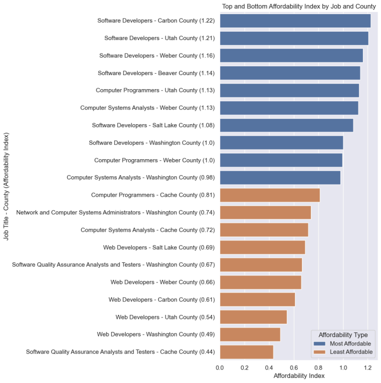


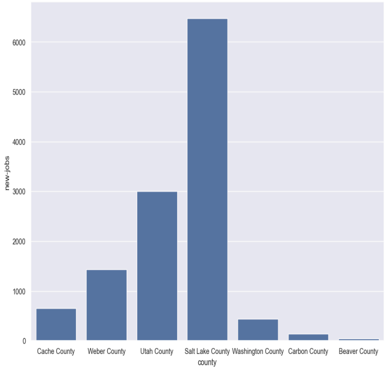


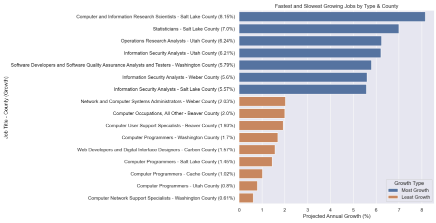




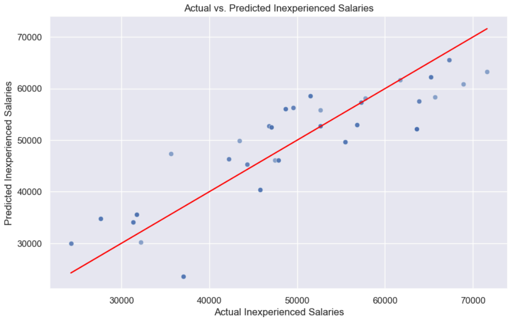
Central Southwest Utah (aka Beaver County area) has the highest salary on average for programmers, while Cache county has the lowest average salary on average. It should be noted that college towns tend to have the lower programming salaries, such as Cache county corresponding to USU, Provo area having BYU, St. George having Utah Tech, etc. This likely isn’t a coincidence, as there would be a higher supply of programmers in these areas.

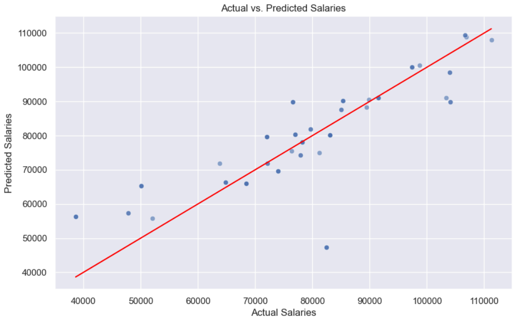






Notably, software developers in Carbon, Utah, Weber, County, etc make the most compared to salary and cost of living. Web Developers make the least in all these counties. Salt Lake County is also projected to have the most job openings in this field. A prospective programmer should keep these job types / areas in mind when making life changing decisions.





The MSE error for inexperienced dev salaries compared with Job Title and County information was $6308.25, while for the median dev it was $10091.06. The correlation between actual and predicted salaries for inexperienced salaries was 0.86, while this for the median salary was 0.81.

In summary, a prospective programmer should strongly consider software development or computer programming and NOT quality assurance or web development if they want the most money. Living in smaller, less dense areas such as Carbon County will give them the most money for cost of living, but jobs are often not hiring in these areas.

## Technical

Data was pre-processed by eliminating NaN values, only Utah areas were considered. The Affordability Index was calculated by taking the average annual median income of a specific job type and dividing it by the average cost of living for that area. The higher the number, the more affordable it is.