# **OVERVIEW**

Our goal was to see if wages for the restaurant/hospitality industry are keeping up with inflation by way of CPI analysis, second is to see if grocery prices impact restaurant employment/wages, and third is to see if particular neighborhoods in Minneapolis have more restaurants than others and does this correlate with lake/pond area within those neighborhoods.

# **SOURCE SELECTION & CURATION**

## Datasets Selected

* For the primary dataset, which is our **restaurant employment** data, we went to mn.gov, as they provide employment census data at the city,county, and state level which can be broken down by industry and time.
* To compare this with **grocery price** data, we got average price of rice (per lb) and average price of flour (per lb) by month and year from the U.S Bureau of Labor Statistics (data.bls.gov). This made sense as rice and flour are two common ingredients used in every cuisine.
* Finally, for the geospatial data we got Minneapolis **neighborhood data** from opendata.minneapolis.gov to visualize the neighborhoods seeing growth in the food service industry.

## Selection Rationale

* The employment by zip, industry, and year is to see which Minneapolis zips have seen growth relative to employment in other industries within the zip over time. This data was selected to analyze trends and inform workers of neighborhoods seeing growth in restaurant establishments and payroll.

## Data Source Metadata

* Please refer to [SEIS 732 - Project - Data Dictionary & Data Source Metadata Document](https://docs.google.com/spreadsheets/d/1Bb6dgRYRnbZnGMlwcVaGi9D5wJaOacBCp9bQ5biK6vE/edit#gid=1762163530)

## Data Dictionary

* Please refer to [SEIS 732 - Project - Data Dictionary & Data Source Metadata Document](https://docs.google.com/spreadsheets/d/1Bb6dgRYRnbZnGMlwcVaGi9D5wJaOacBCp9bQ5biK6vE/edit#gid=1762163530)

## Caveats, Quality and Biases

* In terms of caveats, our data for **average price of rice and flour** were not in the same granularity as our **employment data by industry (primary)**. The average price data was broken down by month and year whereas the employment data was annual. To solve this, we ended up taking the average for each year and joining on the year column.
* We also noticed the employment data had numbers combined for several zip codes as one. For example, the zips 55401-04;15,54,87,88 all had numbers combined into one. In terms of biases, this **could mean over inflated numbers** for zip codes that have combined data.
* Since there are some years where there is missing data, calculating percentage change year over year yields “infinity”. In terms of analysis, we would exclude rows where we encounter this would be misleading.

# 

# **FEATURE ENGINEERING**

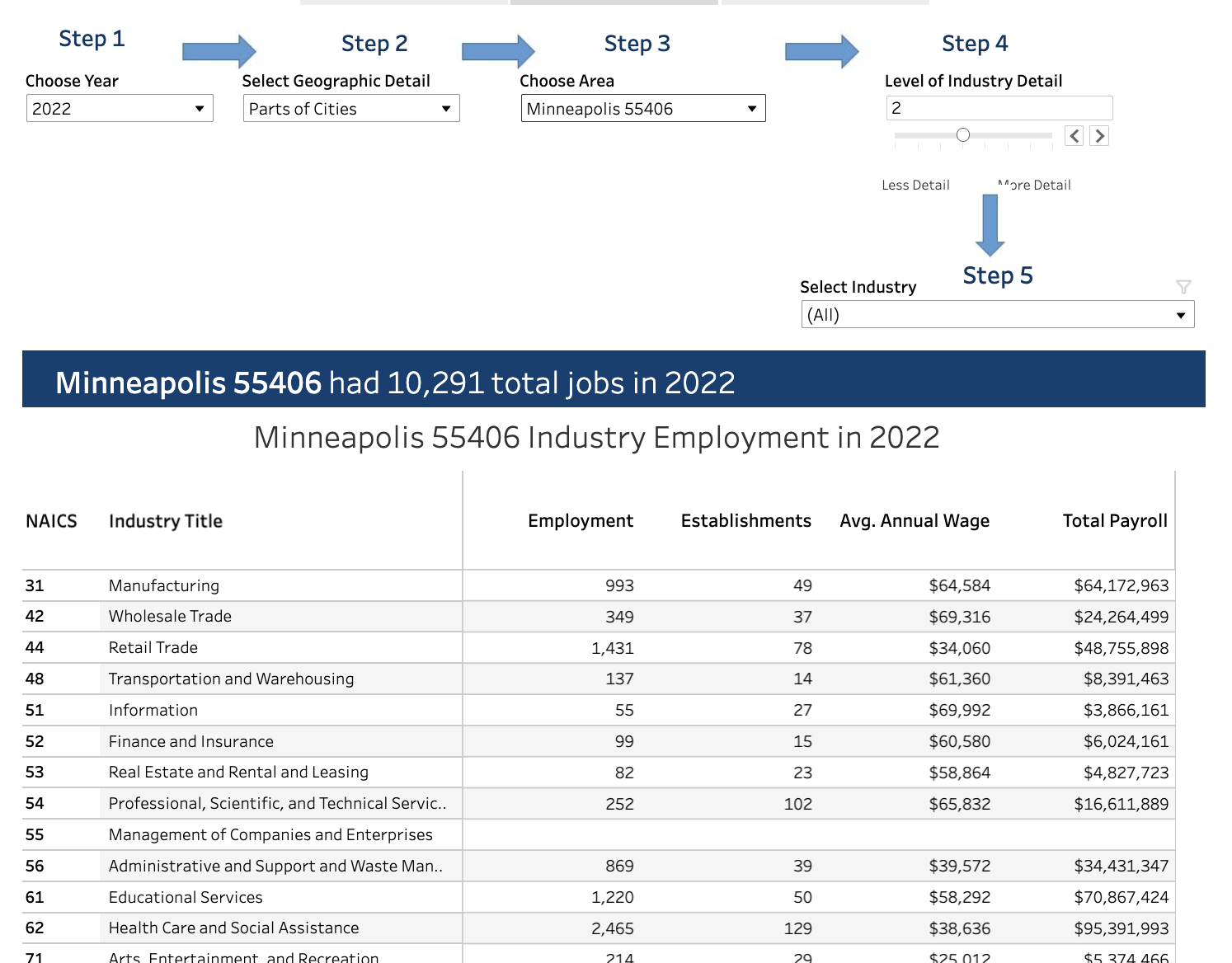
## Feature Requirements

From the **primary source** data, we had to consolidate our source data further after extracting. The data we were able to pull was employment by industry for each year and zip code. That means we had the following number of csv files that we downloaded:

* 23 years x 10 zip codes = 230 csv files.

Each csv file came with the following fields for each zip code and year:

* Employment
* Establishments
* Avg. Annual Wage
* Total Payroll



From this we wrote code to summarize this data into one file with 230 rows and created the following features, these features were **not** part of the source data and we are **not** counting it towards the feature engineering portion but documenting it for completeness sake just to showcase the different challenges one can face when trying to extract data:

* **Total Employment in Zipcode**: Sum of the **Employment** column in the file
* **Total Establishments in Zipcode**: Sum of the **Establishments** column in the file
* **Total Payroll in Zipcode**: Sum of the **Payroll** column in the file
* **Employment for NAICS 722**: NAICS 722 is the industry code for “**Food and Drinking Places**”, this is employment for our industry in question specifically in the zip code that year.
* **Establishments for NAICS 722:** Number of Establishments for “**Food and Drinking Places**” in the zip code that year.
* **Avg. Annual Wage for NAICS 722:** Avg. Annual Wage for “**Food and Drinking Places**” in the zip code that year.
* **Total Payroll for NAICS 722:** Total Payroll in the zip code that year.

With our consolidated data, we **sorted** the data by year and did an aggregation by zipcode to come up with the following **engineered features**. These features **were created to look at yearly trends and analyze which zip codes are ideal for restaurant workers to look towards**.

* **Total Employment in Zipcode Yearly % Change**: Will be used to assess the overall employment “health” of the zip code by year
* **Total Establishments in Zipcode Yearly % Change**: Will be used to assess the overall prosperity of the zip code by year. Will help us see if there is potential to open up new business in the zip code if establishments opening are seeing a positive trend.
* **Total Payroll in Zipcode Yearly % Change**: Will be used to see if in totality the zip code is seeing a positive trend in wages.
* **Employment for NAICS 722 Yearly % Change**: Will be used to assess the employment trend for the restaurant industry in the zip code by year. This feature can help in providing recommendations to workers in the industry looking for employment.
* **Establishments for NAICS 722 Yearly % Change**: Will be used to assess the overall prosperity of the “Food and Drinking” places industry in the zip code by year. Will help us see if there is potential to open up new restaurants in the zip code if establishments opening are seeing a positive trend.
* **Avg. Annual Wage for NAICS 722 Yearly % Change**: Will help us see where “food and drinking” places are seeing a positive/negative change in wages over time. This will help us in providing recommendations for our restaurant workers
* **Total Payroll for NAICS 722 Yearly % Change**:
* **Adjusted Avg. Annual Wage for NAICS 722**: CPI adjusted average annual wage for the food industry in the zip code that year.
* **Adjusted Flour Avg Price per lb:** CPI adjusted average price of flour per lb that year.
* **'Adjusted Rice Avg Price per lb:** CPI adjusted average price of rice per lb that year

## Feature Rationale

The percentage change features will be used in analyzing yearly trends within the different zip codes and help inform restaurant workers where they will be faced with positive wage and employment changes.

Used in conjunction with the average rice and flour data, we can also confirm if the zip codes that do see increases in wages are doing so in proportion to inflation. For example, if we see the average price of rice and flour has increased by 4%-5% year-over-year, we would want to look at zip codes that have seen a similar increase, rather than looking at zip codes that have seen any kind of positive increase.

## Source Data Joins/Merges

We joined our **restaurant employment** data with the **average price of rice and flour** data over time. This was so we can see if grocery (in this case, essentials like rice and flour) prices impact restaurant employment/wages. We can use the rice and flour data to recommend to our restaurant workers which zips are seeing employment and wage growth proportional to rice and flour prices.

We used the zip code field in the **restaurant employment** data to create a new field, “**Neighborhoods**”, to be able to join with our geospatial data.

## 

## Transformation of Features

* Mapping the zip code values to neighborhoods. This was done by doing a manual lookup of the zip codes and creating a configuration within our code to map to a neighborhood.
* For the **average price of flour** and **rice data**, we transformed the monthly data into an average to get a singular value for the year.
* As mentioned above, we did a sum of employment and establishments for each zip code by year to get **Total Employment and Total Establishments in Zip code** in the source file.
* Use CPI to seasonally adjust our wage and price columns to get “adjusted” columns. These include:
  + Adjusted Avg. Annual Wage for NAICS 722
  + Adjusted Flour Avg Price per lb
  + Adjusted Rice Avg Price per lb

## Missing/Bad data

* We had several zip codes that did not have any employment numbers for the restaurant (NAICS 722) industry for particular years. We filled the missing data with 0’s. Therefore, these zip codes will not be part of analyzing employment trends, we acknowledge this leaves potential on the table for informing restaurant workers.
* For the average price of rice data by month and year, we had one full year where we did not have data. We manually filled in that year of data by taking the average of the previous and next year.
  + The data missing was for 2001, we filled in the data by taking the average of 2000 and 2002.