Data Processing[¶](#gjdgxs)

In [1]:

**import** pandas **as** pd  
**import** numpy **as** np  
**import** os  
**import** cpi  
**import** rasterio  
**from** rasterio.mask **import** mask  
**import** zipfile  
  
**import** geopandas **as** gpd  
**import** matplotlib.pyplot **as** plt  
pd**.**options**.**display**.**float\_format **=** '{:,}'**.**format

/var/folders/ql/2vqj8p0d1gq1gggdp2dycp5c0000gn/T/ipykernel\_10430/1685008173.py:4: StaleDataWarning: CPI data is out of date. To accurately inflate to today's dollars, you must run `cpi.update()`.  
 import cpi

In [2]:

*# Helper function to convert source data to right format*  
**def** currency\_to\_number(currency\_val):  
 **if** isinstance(currency\_val, str):  
 **return** pd**.**to\_numeric(currency\_val**.**replace('$', '')**.**replace(',', ''), errors**=**'coerce')  
 **return** currency\_val

In [3]:

os**.**chdir('/Users/hashim/Library/Mobile Documents/com~apple~CloudDocs/ust/seis732-data stores and feature design/project/city\_employment')  
path\_to\_data\_folder **=** os**.**listdir('/Users/hashim/Library/Mobile Documents/com~apple~CloudDocs/ust/seis732-data stores and feature design/project/city\_employment')  
path\_to\_data\_folder**.**remove('.DS\_Store')

In [4]:

*# Here is the directory structure of our data*  
path\_to\_data\_folder

Out[4]:

['55407',  
 'mpls\_zipcode\_employment\_summary.csv',  
 '55406',  
 '55413,18,21',  
 '55412,30',  
 '55405,08,16',  
 '55414,55',  
 '55401-04;15,54,87,88',  
 '55409,10,19',  
 '55411',  
 '55417']

In [5]:

*# This is the naming convention we have employed*   
*# so we are able to consolidate the data easily*  
os**.**listdir('55407/')

Out[5]:

['2008.csv',  
 '2020.csv',  
 '2021.csv',  
 '2009.csv',  
 '2022.csv',  
 '2019.csv',  
 '2018.csv',  
 '2001.csv',  
 '2015.csv',  
 '2014.csv',  
 '2000.csv',  
 '2016.csv',  
 '2002.csv',  
 '2003.csv',  
 '2017.csv',  
 '2013.csv',  
 '2007.csv',  
 '2006.csv',  
 '2012.csv',  
 '2004.csv',  
 '2010.csv',  
 '2011.csv',  
 '2005.csv']

In [6]:

df **=** pd**.**read\_csv('55406/2003.csv', delimiter**=**'\t', encoding**=**'utf-16-le', skiprows**=**2)

Preview of one source file for our primary data (employment by industry and zip)[¶](#30j0zll)

### Below is one source file. I had to extract 230 csv files. We are looking to use Python to consolidate this into one Pandas DataFrame consisting of 230 rows.[¶](#1fob9te)

In [7]:

df**.**head(3)

Out[7]:

|  | **NAICS** | **Industry Title** | **Employment** | **Establishments** | **Avg. Annual Wage** | **Total Payroll** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | 238 | Specialty Trade Contractors | 353 | 38.0 | $42,744 | $15,276,341 |
| **1** | 311 | Food Manufacturing | 260 | 5.0 | $37,596 | $9,763,310 |
| **2** | 323 | Printing and Related Support Activities | 247 | 9.0 | $50,128 | $12,379,716 |

Let's consolidate these[¶](#3znysh7)

In [8]:

*# Initialize a list to store the summary data*  
summary\_data **=** []  
  
*# Iterate over each folder*  
**for** zipcode\_folder **in** path\_to\_data\_folder:  
 zipcode **=** zipcode\_folder   
 folder\_path **=** zipcode\_folder  
  
 *# Check if it's a directory*  
 **if** os**.**path**.**isdir(folder\_path):  
 *# Iterate over each file in the folder*  
 **for** file **in** os**.**listdir(folder\_path):  
 **if** file**.**endswith('.csv'):  
 year **=** file**.**split('.')[0]  
 file\_path **=** os**.**path**.**join(folder\_path, file)  
  
 *# Read the CSV file*  
 df **=** pd**.**read\_csv(file\_path, delimiter**=**'\t', encoding**=**'utf-16-le', skiprows**=**2)  
 df **=** df**.**fillna(0)  
   
 **try**:  
 df['Employment'] **=** df['Employment']**.**astype(int)  
 **except** Exception:  
 df['Employment'] **=** df['Employment']**.**str**.**replace(',','')**.**fillna(0)**.**astype(int)  
   
 **try**:  
 df['Establishments'] **=** df['Establishments']**.**astype(int)  
 **except** Exception:  
 df['Establishments'] **=** df['Establishments']**.**str**.**replace(',','')**.**fillna(0)**.**astype(int)  
   
 df['Avg. Annual Wage'] **=** df['Avg. Annual Wage']**.**str**.**replace('$', '')**.**str**.**replace(',', '')**.**fillna(0)**.**astype(float)  
 df['Total Payroll'] **=** df['Total Payroll']**.**str**.**replace('$', '')**.**str**.**replace(',', '')**.**fillna(0)**.**astype(float)  
   
   
 *# Total Employment in Zipcode*  
 total\_employment **=** df['Employment']**.**sum()  
 total\_establishments **=** df['Establishments']**.**sum()  
 total\_payroll **=** df['Total Payroll']**.**sum()  
  
 *# Filter for NAICS 722*  
 df\_722 **=** df[df['NAICS'] **==** 722]  
  
 *# Check if NAICS 722 data is available*  
 **if** **not** df\_722**.**empty:  
 employment\_722 **=** df\_722['Employment']**.**sum()  
 establishments\_722 **=** df\_722['Establishments']**.**sum()  
 avg\_wage\_722 **=** df\_722['Avg. Annual Wage']**.**sum()  
 total\_payroll\_722 **=** df\_722['Total Payroll']**.**sum()  
 **else**:  
 employment\_722 **=** 0  
 establishments\_722 **=** 0  
 avg\_wage\_722 **=** 0  
 total\_payroll\_722 **=** 0  
  
 *# Append the data to the summary list*  
 summary\_data**.**append({  
 'Zipcode': zipcode,  
 'Year': year,  
 'Total Employment in Zipcode': total\_employment,  
 'Total Establishments in Zipcode': total\_establishments,  
 'Total Payroll in Zipcode': total\_payroll,  
 'Employment for NAICS 722': employment\_722,  
 'Establishments for NAICS 722': establishments\_722,  
 'Avg. Annual Wage for NAICS 722': avg\_wage\_722,  
 'Total Payroll for NAICS 722': total\_payroll\_722  
 })  
  
*# Create a DataFrame from the summary data*  
summary\_df **=** pd**.**DataFrame(summary\_data)

In [9]:

summary\_df**.**sort\_values('Zipcode')**.**head()

Out[9]:

|  | **Zipcode** | **Year** | **Total Employment in Zipcode** | **Total Establishments in Zipcode** | **Total Payroll in Zipcode** | **Employment for NAICS 722** | **Establishments for NAICS 722** | **Avg. Annual Wage for NAICS 722** | **Total Payroll for NAICS 722** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **140** | 55401-04;15,54,87,88 | 2021 | 148201 | 4159 | 16,292,541,001.0 | 6243 | 405 | 30,004.0 | 190,986,359.0 |
| **139** | 55401-04;15,54,87,88 | 2020 | 148898 | 4189 | 15,380,483,441.0 | 5617 | 418 | 27,456.0 | 153,933,482.0 |
| **141** | 55401-04;15,54,87,88 | 2009 | 146088 | 4314 | 9,960,902,372.0 | 8652 | 373 | 18,148.0 | 157,317,173.0 |
| **142** | 55401-04;15,54,87,88 | 2022 | 153292 | 4379 | 17,291,375,031.0 | 7993 | 403 | 32,708.0 | 262,605,818.0 |
| **143** | 55401-04;15,54,87,88 | 2019 | 164357 | 4239 | 15,465,370,734.0 | 10854 | 429 | 26,832.0 | 291,761,862.0 |

In [ ]:

*#summary\_df.to\_csv('mpls\_zipcode\_employment\_summary.csv',index=False)*

Summary data fields[¶](#2et92p0)

In [10]:

summary\_df**.**columns

Out[10]:

Index(['Zipcode', 'Year', 'Total Employment in Zipcode',  
 'Total Establishments in Zipcode', 'Total Payroll in Zipcode',  
 'Employment for NAICS 722', 'Establishments for NAICS 722',  
 'Avg. Annual Wage for NAICS 722', 'Total Payroll for NAICS 722'],  
 dtype='object')

* 'Zipcode',
* 'Year',
* 'Total Employment in Zipcode',
* 'Total Establishments in Zipcode',
* 'Total Payroll in Zipcode',
* 'Employment for NAICS 722',
* 'Establishments for NAICS 722',
* 'Avg. Annual Wage for NAICS 722',
* 'Total Payroll for NAICS 722'

This is the data we have described in the data dictionary

Data Merging[¶](#tyjcwt)

In [11]:

os**.**chdir('/Users/hashim/Library/Mobile Documents/com~apple~CloudDocs/ust/seis732-data stores and feature design/project')

Read in our summary data (primary) that we created above[¶](#3dy6vkm)

In [12]:

df\_employment **=** pd**.**read\_csv('city\_employment/mpls\_zipcode\_employment\_summary.csv')

Read in our time series data[¶](#1t3h5sf)

* Average price of rice per lb by month and year
* Average price of flour per lb by month and year

In [13]:

df\_flour **=** pd**.**read\_excel('raw data/flour\_midwest\_avg\_price\_not\_seasonally\_adjusted.xlsx', skiprows**=**9)  
df\_rice **=** pd**.**read\_excel('raw data/rice\_midwest\_avg\_price\_not\_seasonally\_adjusted.xlsx', skiprows**=**9)

/opt/anaconda3/envs/py39/lib/python3.9/site-packages/openpyxl/styles/stylesheet.py:226: UserWarning: Workbook contains no default style, apply openpyxl's default  
 warn("Workbook contains no default style, apply openpyxl's default")  
/opt/anaconda3/envs/py39/lib/python3.9/site-packages/openpyxl/styles/stylesheet.py:226: UserWarning: Workbook contains no default style, apply openpyxl's default  
 warn("Workbook contains no default style, apply openpyxl's default")

In [14]:

df\_flour**.**head(3)

Out[14]:

|  | **Year** | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** | **Jul** | **Aug** | **Sep** | **Oct** | **Nov** | **Dec** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1990 | 0.253 | 0.249 | 0.252 | 0.251 | 0.254 | 0.259 | 0.26 | 0.255 | 0.249 | 0.244 | 0.235 | 0.236 |
| **1** | 1991 | 0.244 | 0.235 | 0.235 | 0.237 | 0.235 | 0.237 | 0.239 | 0.238 | 0.232 | 0.23 | 0.218 | 0.223 |
| **2** | 1992 | 0.241 | 0.244 | 0.239 | 0.238 | 0.245 | 0.244 | 0.252 | 0.256 | 0.249 | 0.241 | 0.234 | 0.233 |

In [15]:

df\_rice**.**head()

Out[15]:

|  | **Year** | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** | **Jul** | **Aug** | **Sep** | **Oct** | **Nov** | **Dec** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 1990 | 0.501 | 0.476 | 0.502 | 0.497 | 0.496 | 0.494 | 0.494 | 0.497 | 0.505 | 0.497 | 0.507 | 0.491 |
| **1** | 1991 | 0.494 | 0.492 | 0.498 | 0.502 | 0.502 | 0.506 | 0.503 | 0.494 | 0.506 | 0.508 | 0.517 | 0.51 |
| **2** | 1992 | 0.516 | 0.515 | 0.515 | 0.522 | 0.524 | 0.52 | 0.54 | 0.537 | 0.542 | 0.543 | 0.536 | 0.525 |
| **3** | 1993 | 0.526 | 0.53 | 0.525 | 0.522 | 0.518 | 0.518 | 0.527 | 0.507 | 0.494 | 0.495 | 0.49 | 0.495 |
| **4** | 1994 | 0.515 | 0.543 | 0.556 | 0.575 | 0.562 | 0.555 | 0.566 | 0.546 | 0.534 | 0.534 | 0.541 | 0.534 |

Caveat Note:[¶](#4d34og8)

There is no data for 2001 and some months are missing for 2000 and 2002 for rice data.

I am going to take the average of 2000 and 2002 and just add it in there

In [16]:

df\_rice[df\_rice['Year']**.**isin([2000, 2001 ,2002])]

Out[16]:

|  | **Year** | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** | **Jul** | **Aug** | **Sep** | **Oct** | **Nov** | **Dec** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | 2000 | 0.499 | 0.492 | 0.481 | 0.49 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **11** | 2002 | NaN | NaN | NaN | NaN | 0.472 | 0.47 | 0.475 | 0.482 | 0.471 | 0.466 | 0.461 | 0.458 |

In [17]:

df\_rice**.**loc[10:11]

Out[17]:

|  | **Year** | **Jan** | **Feb** | **Mar** | **Apr** | **May** | **Jun** | **Jul** | **Aug** | **Sep** | **Oct** | **Nov** | **Dec** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **10** | 2000 | 0.499 | 0.492 | 0.481 | 0.49 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **11** | 2002 | NaN | NaN | NaN | NaN | 0.472 | 0.47 | 0.475 | 0.482 | 0.471 | 0.466 | 0.461 | 0.458 |

Workaround[¶](#2s8eyo1)

Take mean of this

In [18]:

df\_rice**.**loc[10:11]**.**mean()['Jan':'Dec']**.**mean()

Out[18]:

0.4764166666666667

In [19]:

*# New row data for the year 2001 with all months having a value of 0.476*  
new\_row **=** {'Year': 2001, 'Jan': 0.476, 'Feb': 0.476, 'Mar': 0.476, 'Apr': 0.476, 'May': 0.476,   
 'Jun': 0.476, 'Jul': 0.476, 'Aug': 0.476, 'Sep': 0.476, 'Oct': 0.476, 'Nov': 0.476, 'Dec': 0.476}

In [20]:

df\_rice **=** pd**.**concat([df\_rice, pd**.**DataFrame([new\_row])], ignore\_index**=True**)

In [21]:

df\_rice**.**sort\_values('Year', inplace**=True**)  
df\_rice**.**reset\_index(drop**=True**, inplace**=True**)

Consolidate the rice and flour data[¶](#17dp8vu)

Since this data is not in the same granularity (i.e broken down by month) as our primary data (broken down by year), we need to aggregate this so we can join on the Year field.

### How are we going to aggregate ?[¶](#3rdcrjn)

* Take the mean of each year

In [22]:

years **=** pd**.**Series(list(range(1990,2024)))  
  
*# We are going to take the mean of each year and that*   
flour\_avg\_price **=** df\_flour**.**T**.**loc['Jan':'Dec',]**.**mean()  
rice\_avg\_price **=** df\_rice**.**T**.**loc['Jan':'Dec', ]**.**mean()

In [23]:

years**.**name **=** 'Year'  
flour\_avg\_price**.**name **=** 'Flour Avg Price per lb'  
rice\_avg\_price**.**name **=** 'Rice Avg Price per lb'

In [24]:

df\_avg\_price **=** pd**.**concat([years, flour\_avg\_price, rice\_avg\_price],axis**=**1)

In [25]:

pd**.**options**.**display**.**float\_format **=** '{:,.2f}'**.**format

In [26]:

df\_avg\_price**.**head()

Out[26]:

|  | **Year** | **Flour Avg Price per lb** | **Rice Avg Price per lb** |
| --- | --- | --- | --- |
| **0** | 1990 | 0.25 | 0.50 |
| **1** | 1991 | 0.23 | 0.50 |
| **2** | 1992 | 0.24 | 0.53 |
| **3** | 1993 | 0.23 | 0.51 |
| **4** | 1994 | 0.23 | 0.55 |

Join primary and time series data[¶](#26in1rg)

In [27]:

df\_merged **=** pd**.**merge(df\_employment, df\_avg\_price, on**=**['Year'], how**=**'inner')

In [28]:

df\_merged**.**head()

Out[28]:

|  | **Zipcode** | **Year** | **Total Employment in Zipcode** | **Total Establishments in Zipcode** | **Total Payroll in Zipcode** | **Employment for NAICS 722** | **Establishments for NAICS 722** | **Avg. Annual Wage for NAICS 722** | **Total Payroll for NAICS 722** | **Flour Avg Price per lb** | **Rice Avg Price per lb** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 55407 | 2008 | 6077 | 416 | 350,163,127.00 | 473 | 53 | 14,404.00 | 6,822,473.00 | 0.51 | 0.73 |
| **1** | 55406 | 2008 | 5685 | 449 | 206,421,620.00 | 925 | 62 | 13,884.00 | 12,881,469.00 | 0.51 | 0.73 |
| **2** | 55413,18,21 | 2008 | 25961 | 1211 | 1,291,854,067.00 | 0 | 0 | 0.00 | 0.00 | 0.51 | 0.73 |
| **3** | 55412,30 | 2008 | 1545 | 116 | 60,056,841.00 | 0 | 0 | 0.00 | 0.00 | 0.51 | 0.73 |
| **4** | 55405,08,16 | 2008 | 9958 | 1069 | 383,932,010.00 | 0 | 0 | 0.00 | 0.00 | 0.51 | 0.73 |

Feature Engineering[¶](#lnxbz9)

Going to calculate number of food service jobs in each zip code by year.

In [29]:

df\_merged['% Food Service Jobs in Zip'] **=** df\_merged['Employment for NAICS 722']**/**df\_merged['Total Employment in Zipcode']  
df\_merged['% Food Service Establishments in Zip'] **=** df\_merged['Establishments for NAICS 722']**/**df\_merged['Total Establishments in Zipcode']  
  
df\_merged **=** df\_merged**.**sort\_values(by**=**['Zipcode', 'Year'])**.**reset\_index(drop**=True**)

For our hypothetical analysis we are going to calculate percentage change year over year grouped by Zip code. Below are the fields where we are going to calculate percentage change.

In [30]:

columns\_to\_calculate **=** [  
 'Total Employment in Zipcode',   
 'Total Establishments in Zipcode',   
 'Total Payroll in Zipcode',   
 'Employment for NAICS 722',   
 'Establishments for NAICS 722',   
 'Avg. Annual Wage for NAICS 722',   
 'Total Payroll for NAICS 722'  
]

In [31]:

df\_merged**.**fillna(0,inplace**=True**)

In [32]:

**for** column **in** columns\_to\_calculate:  
 df\_merged[column **+** ' Yearly % Change'] **=** df\_merged**.**groupby('Zipcode')[column]**.**pct\_change() **\*** 1

In [33]:

df\_merged**.**head()

Out[33]:

|  | **Zipcode** | **Year** | **Total Employment in Zipcode** | **Total Establishments in Zipcode** | **Total Payroll in Zipcode** | **Employment for NAICS 722** | **Establishments for NAICS 722** | **Avg. Annual Wage for NAICS 722** | **Total Payroll for NAICS 722** | **Flour Avg Price per lb** | **Rice Avg Price per lb** | **% Food Service Jobs in Zip** | **% Food Service Establishments in Zip** | **Total Employment in Zipcode Yearly % Change** | **Total Establishments in Zipcode Yearly % Change** | **Total Payroll in Zipcode Yearly % Change** | **Employment for NAICS 722 Yearly % Change** | **Establishments for NAICS 722 Yearly % Change** | **Avg. Annual Wage for NAICS 722 Yearly % Change** | **Total Payroll for NAICS 722 Yearly % Change** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 55401-04;15,54,87,88 | 2000 | 153650 | 4880 | 8,205,509,715.00 | 8089 | 368 | 15,132.00 | 122,613,294.00 | 0.29 | 0.49 | 0.05 | 0.08 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **1** | 55401-04;15,54,87,88 | 2001 | 153482 | 4752 | 8,618,403,063.00 | 7971 | 363 | 15,392.00 | 123,097,435.00 | 0.30 | 0.48 | 0.05 | 0.08 | -0.00 | -0.03 | 0.05 | -0.01 | -0.01 | 0.02 | 0.00 |
| **2** | 55401-04;15,54,87,88 | 2002 | 142347 | 4462 | 8,143,669,190.00 | 8185 | 369 | 15,340.00 | 125,700,410.00 | 0.31 | 0.47 | 0.06 | 0.08 | -0.07 | -0.06 | -0.06 | 0.03 | 0.02 | -0.00 | 0.02 |
| **3** | 55401-04;15,54,87,88 | 2003 | 138532 | 4402 | 8,094,680,577.00 | 8059 | 379 | 15,964.00 | 128,937,572.00 | 0.31 | 0.45 | 0.06 | 0.09 | -0.03 | -0.01 | -0.01 | -0.02 | 0.03 | 0.04 | 0.03 |
| **4** | 55401-04;15,54,87,88 | 2004 | 144339 | 4358 | 9,171,613,599.00 | 8397 | 386 | 16,276.00 | 136,867,030.00 | 0.30 | 0.54 | 0.06 | 0.09 | 0.04 | -0.01 | 0.13 | 0.04 | 0.02 | 0.02 | 0.06 |

Map Zipcode to neighborhood to join with geojson data[¶](#35nkun2)

Just Google'd the zip codes and found the relevant neighborhoods.

In [34]:

map\_zipcode\_to\_neighborhood **=** {  
 '55401-04;15,54,87,88':'North Loop',  
 '55405,08,16': 'Lowry Hill',  
 '55406': 'Howe',  
 '55407': 'Powderhorn Park',  
 '55409,10,19': 'King Field',  
 '55411': 'Near - North',  
 '55412,30': 'Webber - Camden',  
 '55413,18,21': 'Beltrami',  
 '55414,55': 'University of Minnesota',  
 '55417': 'Wenonah'  
   
}

In [35]:

df\_merged['Neighborhood'] **=** df\_merged['Zipcode']**.**map(map\_zipcode\_to\_neighborhood)  
df\_merged**.**fillna(0,inplace**=True**)  
df\_merged **=** df\_merged**.**replace(np**.**inf, 1)  
*#df\_merged.to\_csv('employment\_and\_grocery\_merged\_updated.csv', index=False)*

More feature engineering[¶](#1ksv4uv)

We are going to add fields to seasonally adjust our wage and price columns

In [36]:

df\_merged['Adjusted Avg. Annual Wage for NAICS 722'] **=** df\_merged**.**apply(**lambda** x: cpi**.**inflate(x['Avg. Annual Wage for NAICS 722'], x['Year']), axis**=**1)  
df\_merged['Adjusted Flour Avg Price per lb'] **=** df\_merged**.**apply(**lambda** x: cpi**.**inflate(x['Flour Avg Price per lb'], x['Year']), axis**=**1)  
df\_merged['Adjusted Rice Avg Price per lb'] **=** df\_merged**.**apply(**lambda** x: cpi**.**inflate(x['Rice Avg Price per lb'], x['Year']), axis**=**1)  
  
  
df\_merged['Adjusted Avg. Annual Wage for NAICS 722 Yearly % Change'] **=** df\_merged**.**groupby('Zipcode')['Adjusted Avg. Annual Wage for NAICS 722']**.**pct\_change() **\*** 1  
df\_merged['Adjusted Avg. Flour Avg Price per lb Yearly % Change'] **=** df\_merged**.**groupby('Zipcode')['Adjusted Flour Avg Price per lb']**.**pct\_change() **\*** 1  
df\_merged['Adjusted Avg. Rice Avg Price per lb Yearly % Change'] **=** df\_merged**.**groupby('Zipcode')['Adjusted Rice Avg Price per lb']**.**pct\_change() **\*** 1

Rationale[¶](#44sinio)

We are going to calculate the "delta" between adjusted wages for restaurant workers and rice and flour prices.

Hypothesis here is, every restaurant spends on these essential ingredients, if the expenditure on these items (and the implication here is on other grocery items) is going up, we want to give recommendations to our restaurant workers where adjusted wages are increasing in proportion to this

In [37]:

df\_merged['% Difference Avg. Wage NAICS 722 - Flour'] **=** df\_merged['Adjusted Avg. Annual Wage for NAICS 722 Yearly % Change'] **-** df\_merged['Adjusted Avg. Flour Avg Price per lb Yearly % Change']  
df\_merged['% Difference Avg. Wage NAICS 722 - Rice'] **=** df\_merged['Adjusted Avg. Annual Wage for NAICS 722 Yearly % Change'] **-** df\_merged['Adjusted Avg. Rice Avg Price per lb Yearly % Change']

A preview of what zip codes are looking favorable for workers

In [38]:

pd**.**options**.**display**.**float\_format **=** '{:,.2f}'**.**format

In [39]:

(df\_merged[(df\_merged['% Difference Avg. Wage NAICS 722 - Flour']**>**0)  
 **&**(df\_merged['% Difference Avg. Wage NAICS 722 - Rice']**>**0)  
 **&**(df\_merged['Year']**>**2018)  
 ]['Neighborhood']**.**unique()**.**tolist())

Out[39]:

['North Loop',  
 'Howe',  
 'Powderhorn Park',  
 'King Field',  
 'Near - North',  
 'University of Minnesota']

Join with Geospatial[¶](#2jxsxqh)

In [40]:

os**.**getcwd()

Out[40]:

'/Users/hashim/Library/Mobile Documents/com~apple~CloudDocs/ust/seis732-data stores and feature design/project'

In [41]:

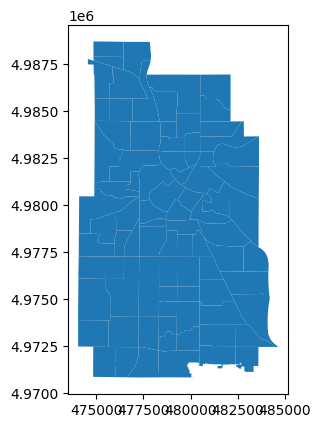
*# read in dataset*  
mpls\_nlcd **=** rasterio**.**open('mpls\_nlcd.tif')  
  
*# read in Minneapolis neighborhoods data*  
mpls\_neighborhoods\_df **=** gpd**.**read\_file('Minneapolis\_Neighborhoods.geojson')  
  
*# reproject the vector data to be the same as the raster*  
mpls\_neighborhoods\_df **=** mpls\_neighborhoods\_df**.**to\_crs(crs**=**mpls\_nlcd**.**crs)

In [42]:

mpls\_neighborhoods\_df**.**plot()

Out[42]:

<Axes: >



Filter Employment Data to 10 Year Increments[¶](#z337ya)

In [43]:

mpls\_emp\_df\_2002 **=** df\_merged**.**loc[(df\_merged['Year'] **==** 2002)]  
mpls\_emp\_df\_2012 **=** df\_merged**.**loc[(df\_merged['Year'] **==** 2012)]  
mpls\_emp\_df\_2022 **=** df\_merged**.**loc[(df\_merged['Year'] **==** 2022)]

Merge combined employment (primary), avg price of rice and flour (time series) with neighborhood data (geospatial)[¶](#3j2qqm3)

In [44]:

geo\_merge\_df\_2002 **=** pd**.**merge(mpls\_emp\_df\_2002, mpls\_neighborhoods\_df, how**=**'outer', left\_on**=**'Neighborhood', right\_on**=**'BDNAME')  
*# set the active geometry*  
geo\_merge\_df\_2002 **=** geo\_merge\_df\_2002**.**set\_geometry("geometry")  
*#view the head*  
geo\_merge\_df\_2002**.**head()

Out[44]:

|  | **Zipcode** | **Year** | **Total Employment in Zipcode** | **Total Establishments in Zipcode** | **Total Payroll in Zipcode** | **Employment for NAICS 722** | **Establishments for NAICS 722** | **Avg. Annual Wage for NAICS 722** | **Total Payroll for NAICS 722** | **Flour Avg Price per lb** | **...** | **INT\_REFNO** | **PREFIX** | **UDI** | **SYMBOL\_NAM** | **BDNAME** | **BDNUM** | **TEXT\_NBR** | **SHAPE\_Length** | **SHAPE\_Area** | **geometry** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 55401-04;15,54,87,88 | 2,002.00 | 142,347.00 | 4,462.00 | 8,143,669,190.00 | 8,185.00 | 369.00 | 15,340.00 | 125,700,410.00 | 0.31 | ... | -2,144,131,400.00 | REFNO | 23186.00 | WARDAREA | North Loop | 86 | 86 | 0.07 | 0.00 | MULTIPOLYGON (((478550.326 4982112.307, 478592... |
| **1** | 55405,08,16 | 2,002.00 | 13,922.00 | 1,465.00 | 500,040,639.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | ... | -2,144,135,700.00 | REFNO | 23143.00 | WARDAREA | Lowry Hill | 43 | 43 | 0.06 | 0.00 | MULTIPOLYGON (((477293.021 4980440.894, 477292... |
| **2** | 55406 | 2,002.00 | 6,155.00 | 419.00 | 215,393,905.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.31 | ... | -2,144,133,700.00 | REFNO | 23163.00 | WARDAREA | Howe | 63 | 63 | 0.08 | 0.00 | MULTIPOLYGON (((484138.698 4976477.210, 484136... |
| **3** | 55407 | 2,002.00 | 4,010.00 | 324.00 | 162,742,399.00 | 466.00 | 42.00 | 12,792.00 | 5,956,071.00 | 0.31 | ... | -2,144,134,500.00 | REFNO | 23155.00 | WARDAREA | Powderhorn Park | 55 | 55 | 0.06 | 0.00 | MULTIPOLYGON (((480482.625 4976064.000, 480482... |
| **4** | 55409,10,19 | 2,002.00 | 9,343.00 | 1,176.00 | 266,852,787.00 | 883.00 | 66.00 | 12,272.00 | 10,842,984.00 | 0.31 | ... | -2,144,132,900.00 | REFNO | 23171.00 | WARDAREA | King Field | 71 | 71 | 0.06 | 0.00 | MULTIPOLYGON (((478327.523 4976072.079, 478327... |

5 rows × 40 columns

In [45]:

geo\_merge\_df\_2012 **=** pd**.**merge(mpls\_emp\_df\_2012, mpls\_neighborhoods\_df, how**=**'outer', left\_on**=**'Neighborhood', right\_on**=**'BDNAME')  
*# set the active geometry*  
geo\_merge\_df\_2012 **=** geo\_merge\_df\_2012**.**set\_geometry("geometry")  
*#view the head*  
geo\_merge\_df\_2012**.**head()

Out[45]:

|  | **Zipcode** | **Year** | **Total Employment in Zipcode** | **Total Establishments in Zipcode** | **Total Payroll in Zipcode** | **Employment for NAICS 722** | **Establishments for NAICS 722** | **Avg. Annual Wage for NAICS 722** | **Total Payroll for NAICS 722** | **Flour Avg Price per lb** | **...** | **INT\_REFNO** | **PREFIX** | **UDI** | **SYMBOL\_NAM** | **BDNAME** | **BDNUM** | **TEXT\_NBR** | **SHAPE\_Length** | **SHAPE\_Area** | **geometry** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 55401-04;15,54,87,88 | 2,012.00 | 154,014.00 | 4,284.00 | 11,567,873,852.00 | 9,460.00 | 416.00 | 20,280.00 | 192,059,097.00 | 0.52 | ... | -2,144,131,400.00 | REFNO | 23186.00 | WARDAREA | North Loop | 86 | 86 | 0.07 | 0.00 | MULTIPOLYGON (((478550.326 4982112.307, 478592... |
| **1** | 55405,08,16 | 2,012.00 | 17,660.00 | 1,454.00 | 732,456,085.00 | 3,806.00 | 164.00 | 18,304.00 | 69,644,617.00 | 0.52 | ... | -2,144,135,700.00 | REFNO | 23143.00 | WARDAREA | Lowry Hill | 43 | 43 | 0.06 | 0.00 | MULTIPOLYGON (((477293.021 4980440.894, 477292... |
| **2** | 55406 | 2,012.00 | 6,773.00 | 536.00 | 225,661,072.00 | 930.00 | 53.00 | 15,288.00 | 14,275,447.00 | 0.52 | ... | -2,144,133,700.00 | REFNO | 23163.00 | WARDAREA | Howe | 63 | 63 | 0.08 | 0.00 | MULTIPOLYGON (((484138.698 4976477.210, 484136... |
| **3** | 55407 | 2,012.00 | 4,462.00 | 348.00 | 268,482,063.00 | 555.00 | 55.00 | 17,524.00 | 9,751,158.00 | 0.52 | ... | -2,144,134,500.00 | REFNO | 23155.00 | WARDAREA | Powderhorn Park | 55 | 55 | 0.06 | 0.00 | MULTIPOLYGON (((480482.625 4976064.000, 480482... |
| **4** | 55409,10,19 | 2,012.00 | 6,993.00 | 896.00 | 249,987,110.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.52 | ... | -2,144,132,900.00 | REFNO | 23171.00 | WARDAREA | King Field | 71 | 71 | 0.06 | 0.00 | MULTIPOLYGON (((478327.523 4976072.079, 478327... |

5 rows × 40 columns

In [46]:

geo\_merge\_df\_2022 **=** pd**.**merge(mpls\_emp\_df\_2022, mpls\_neighborhoods\_df, how**=**'outer', left\_on**=**'Neighborhood', right\_on**=**'BDNAME')  
*# set the active geometry*  
geo\_merge\_df\_2022 **=** geo\_merge\_df\_2022**.**set\_geometry("geometry")  
*#view the head*  
geo\_merge\_df\_2022**.**head()

Out[46]:

|  | **Zipcode** | **Year** | **Total Employment in Zipcode** | **Total Establishments in Zipcode** | **Total Payroll in Zipcode** | **Employment for NAICS 722** | **Establishments for NAICS 722** | **Avg. Annual Wage for NAICS 722** | **Total Payroll for NAICS 722** | **Flour Avg Price per lb** | **...** | **INT\_REFNO** | **PREFIX** | **UDI** | **SYMBOL\_NAM** | **BDNAME** | **BDNUM** | **TEXT\_NBR** | **SHAPE\_Length** | **SHAPE\_Area** | **geometry** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 55401-04;15,54,87,88 | 2,022.00 | 153,292.00 | 4,379.00 | 17,291,375,031.00 | 7,993.00 | 403.00 | 32,708.00 | 262,605,818.00 | 0.49 | ... | -2,144,131,400.00 | REFNO | 23186.00 | WARDAREA | North Loop | 86 | 86 | 0.07 | 0.00 | MULTIPOLYGON (((478550.326 4982112.307, 478592... |
| **1** | 55405,08,16 | 2,022.00 | 16,045.00 | 1,467.00 | 1,083,205,670.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.49 | ... | -2,144,135,700.00 | REFNO | 23143.00 | WARDAREA | Lowry Hill | 43 | 43 | 0.06 | 0.00 | MULTIPOLYGON (((477293.021 4980440.894, 477292... |
| **2** | 55406 | 2,022.00 | 8,333.00 | 730.00 | 377,395,691.00 | 1,068.00 | 72.00 | 27,144.00 | 29,030,047.00 | 0.49 | ... | -2,144,133,700.00 | REFNO | 23163.00 | WARDAREA | Howe | 63 | 63 | 0.08 | 0.00 | MULTIPOLYGON (((484138.698 4976477.210, 484136... |
| **3** | 55407 | 2,022.00 | 7,015.00 | 635.00 | 512,076,518.00 | 691.00 | 78.00 | 26,520.00 | 18,358,914.00 | 0.49 | ... | -2,144,134,500.00 | REFNO | 23155.00 | WARDAREA | Powderhorn Park | 55 | 55 | 0.06 | 0.00 | MULTIPOLYGON (((480482.625 4976064.000, 480482... |
| **4** | 55409,10,19 | 2,022.00 | 7,079.00 | 1,219.00 | 376,029,972.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.49 | ... | -2,144,132,900.00 | REFNO | 23171.00 | WARDAREA | King Field | 71 | 71 | 0.06 | 0.00 | MULTIPOLYGON (((478327.523 4976072.079, 478327... |

5 rows × 40 columns

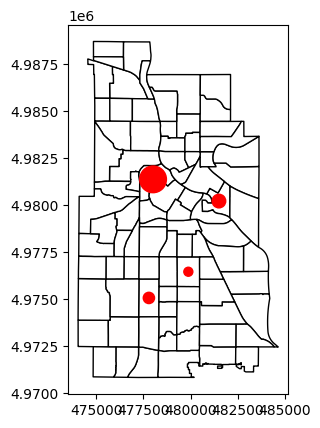
Plot NAICS 722 Establishment Counts as Centroids[¶](#1y810tw)

2002[¶](#4i7ojhp)

### Number of "food and drinking place" establishments by neighborhood¶

In [47]:

*# add the centroids column*  
geo\_merge\_df\_2002['neighborhood\_centroid'] **=** geo\_merge\_df\_2002**.**centroid  
  
*# set the base layer for multilayered map plot*  
base **=** geo\_merge\_df\_2002**.**plot(color**=**'white', edgecolor**=**'black')  
  
*# plot 2002 NAICS 722 Establishments*  
geo\_merge\_df\_2002['neighborhood\_centroid']**.**plot(ax**=**base, marker**=**'o', color**=**'red', markersize**=**geo\_merge\_df\_2002['Establishments for NAICS 722']);

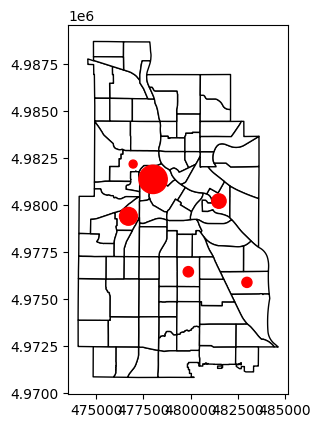


2012[¶](#2xcytpi)

### Number of "food and drinking place" establishments by neighborhood¶

In [48]:

*# add the centroids column*  
geo\_merge\_df\_2012['neighborhood\_centroid'] **=** geo\_merge\_df\_2012**.**centroid  
  
*# set the base layer for multilayered map plot*  
base **=** geo\_merge\_df\_2012**.**plot(color**=**'white', edgecolor**=**'black')  
  
*# plot 2012 NAICS 722 Establishments*  
geo\_merge\_df\_2012['neighborhood\_centroid']**.**plot(ax**=**base, marker**=**'o', color**=**'red', markersize**=**geo\_merge\_df\_2012['Establishments for NAICS 722']);

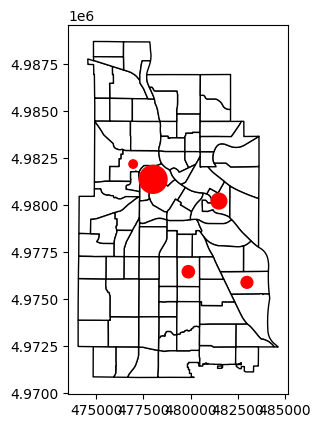


2022[¶](#1ci93xb)

### Number of "food and drinking place" establishments by neighborhood¶

In [49]:

*# add the centroids column*  
geo\_merge\_df\_2022['neighborhood\_centroid'] **=** geo\_merge\_df\_2022**.**centroid  
  
*# set the base layer for multilayered map plot*  
base **=** geo\_merge\_df\_2022**.**plot(color**=**'white', edgecolor**=**'black')  
  
*# plot 2012 NAICS 722 Establishments*  
geo\_merge\_df\_2022['neighborhood\_centroid']**.**plot(ax**=**base, marker**=**'o', color**=**'red', markersize**=**geo\_merge\_df\_2022['Establishments for NAICS 722']);



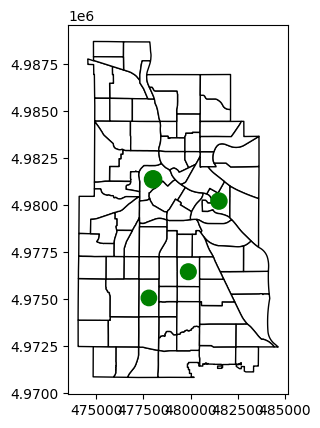
Plot NAICS 722 Establishment Avg Annual Wage as Centroids[¶](#3whwml4)

2002[¶](#4i7ojhp)

### Avg. Annual Wage of "food and drinking place" industry by neighborhood¶

In [50]:

*# add the centroids column*  
geo\_merge\_df\_2002['neighborhood\_centroid'] **=** geo\_merge\_df\_2002**.**centroid  
  
*# set the base layer for multilayered map plot*  
base **=** geo\_merge\_df\_2002**.**plot(color**=**'white', edgecolor**=**'black')  
  
*# plot 2002 NAICS 722 Avg Annual Wage*  
geo\_merge\_df\_2002['neighborhood\_centroid']**.**plot(ax**=**base, marker**=**'o', color**=**'green', markersize**=**geo\_merge\_df\_2002['Avg. Annual Wage for NAICS 722']**/**100);

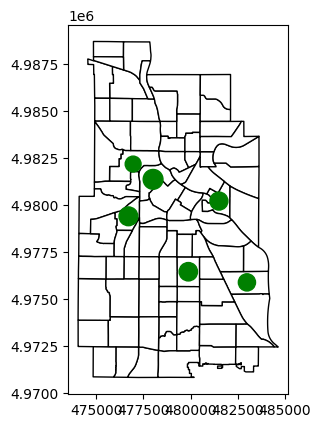


2012[¶](#2xcytpi)

### Avg. Annual Wage of "food and drinking place" industry by neighborhood¶

In [51]:

*# add the centroids column*  
geo\_merge\_df\_2012['neighborhood\_centroid'] **=** geo\_merge\_df\_2012**.**centroid  
  
*# set the base layer for multilayered map plot*  
base **=** geo\_merge\_df\_2012**.**plot(color**=**'white', edgecolor**=**'black')  
  
*# plot 2012 NAICS 722 Avg Annual Wage*  
geo\_merge\_df\_2012['neighborhood\_centroid']**.**plot(ax**=**base, marker**=**'o', color**=**'green', markersize**=**geo\_merge\_df\_2012['Avg. Annual Wage for NAICS 722']**/**100);

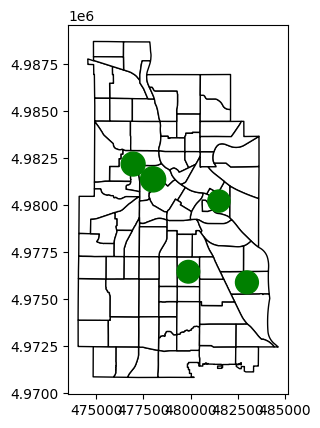


2022[¶](#1ci93xb)

### Avg. Annual Wage of "food and drinking place" industry by neighborhood¶

In [52]:

*# add the centroids column*  
geo\_merge\_df\_2022['neighborhood\_centroid'] **=** geo\_merge\_df\_2022**.**centroid  
  
*# set the base layer for multilayered map plot*  
base **=** geo\_merge\_df\_2022**.**plot(color**=**'white', edgecolor**=**'black')  
  
*# plot 2012 NAICS 722 Avg Annual Wage*  
geo\_merge\_df\_2022['neighborhood\_centroid']**.**plot(ax**=**base, marker**=**'o', color**=**'green', markersize**=**geo\_merge\_df\_2022['Avg. Annual Wage for NAICS 722']**/**100);



In [ ]: