**Michael Perrine**

**DSC 550 Data Mining**

**Professor Werner**

**Final Assignment**

# Analysis of Population Data

Milestone 1

What do demographics, economic growth and school ratings have in common? These factors are important to developers. The cost of building in the wrong location can cost millions in losses. The goal of this analysis is to determine locations of interest for land development. We believe Texas is an attractive place to develop land for the purpose of economic growth. To validate this assumption, I will explore several factors that make a location attractive. These factors are the quality of schools, economic development, and population growth.

To perform my analysis, I will use several libraries found in the python program. Some of the libraries are Pandas, Matplotlib, NumPy, Scikit learn, and Seaborn among others. These libraries will give me the ability to perform my preliminary analysis. The data will come from the Texas Open Data Portal: State of Texas | Open Data Portal | Open Data Portal. This website has csv files for all the data of interest. To begin the project, I will upload the files into my Python program. I will need to perform some data analysis to uncover relationships such as what factors stimulate population growth? What factors drive business? What impact does a good school play in migration decisions? How can I use this information to predict areas of growth?

Some of the work that I will perform will be exploratory data analysis. I will need to transform and clean the data sets. Once I clean the data it will be ready for the visualization process. Visualizing the data is another important tool to analyze and interpret data. This includes building graphs such as histograms, scatter plots, and box plots. These are just some of the visualization tools available.

After I build my graphs, I can determine the relationships involved. The next step is to build a model. Scikit Learn is the library that will build the model. This is probably the most important step in the process because it gives me the ability to make predictions and take actionable steps in determining what areas are prime investment opportunities.

There are challenges to my analysis. I don't know what relationships if any I will find. If I do find relationships in the data, I'm not sure how strong they are or if they will lead to a viable investment decision.

In [1]:
```
# import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

1. The first step is to upload the data and to build some graphs to visualize the data

In [2]:
```
# This code imports the school data and displays the first 10 rows
school = pd.read_csv(r"texas school data.csv")
school.head(10)
```

Out[2]:

| | District\nNumber | District | Campus\nNumber | Campus | Region | County | Scho |
|---|---|---|---|---|---|---|---|
| **0** | 1902 | CAYUGA ISD | NaN | NaN | REGION 07: KILGORE | ANDERSON | |
| **1** | 1902 | CAYUGA ISD | 1902001.0 | CAYUGA H S | REGION 07: KILGORE | ANDERSON | Hig |
| **2** | 1902 | CAYUGA ISD | 1902041.0 | CAYUGA MIDDLE | REGION 07: KILGORE | ANDERSON | Midd |
| **3** | 1902 | CAYUGA ISD | 1902103.0 | CAYUGA EL | REGION 07: KILGORE | ANDERSON | El |
| **4** | 1903 | ELKHART ISD | NaN | NaN | REGION 07: KILGORE | ANDERSON | |
| **5** | 1903 | ELKHART ISD | 1903001.0 | ELKHART H S | REGION 07: KILGORE | ANDERSON | Hig |
| **6** | 1903 | ELKHART ISD | 1903041.0 | ELKHART MIDDLE | REGION 07: KILGORE | ANDERSON | Midd |
| **7** | 1903 | ELKHART ISD | 1903101.0 | ELKHART EL | REGION 07: KILGORE | ANDERSON | El |
| **8** | 1903 | ELKHART ISD | 1903102.0 | ELKHART INT | REGION 07: KILGORE | ANDERSON | El |
| **9** | 1904 | FRANKSTON ISD | NaN | NaN | REGION 07: KILGORE | ANDERSON | |

10 rows × 41 columns

In [3]:
```
# This code drops rows with NaN values
school_1 = school.dropna()
school_1.head()
```

Out[3]:

| | District\nNumber | District | Campus\nNumber | Campus | Region | County |
|---|---|---|---|---|---|---|
| **159** | 10902 | BANDERA ISD | 10902001.0 | BANDERA H S | REGION 20: SAN ANTONIO | BANDERA |
| **278** | 14906 | KILLEEN ISD | 14906044.0 | MANOR MIDDLE | REGION 12: WACO | BELL |
| **280** | 14906 | KILLEEN ISD | 14906048.0 | PALO ALTO MIDDLE | REGION 12: WACO | BELL |
| **517** | 15905 | EDGEWOOD ISD | 15905041.0 | BRENTWOOD MIDDLE | REGION 20: SAN ANTONIO | BEXAR |
| **519** | 15905 | EDGEWOOD ISD | 15905044.0 | E T WRENN MIDDLE | REGION 20: SAN ANTONIO | BEXAR |

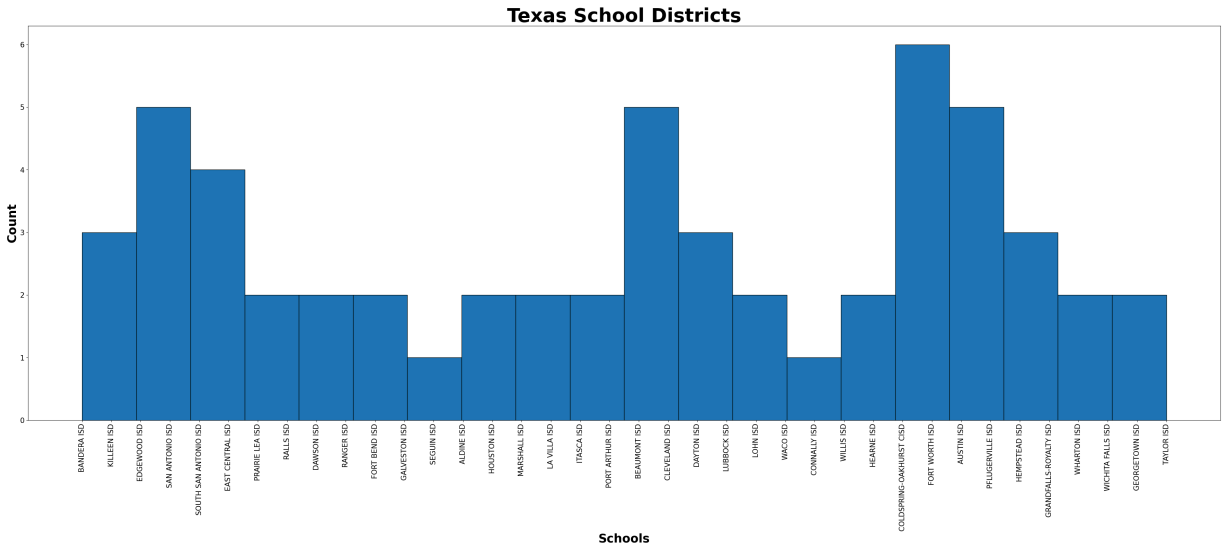
5 rows × 41 columns

In [4]:
```
# This code shows the dimensions of the school data
school_1.shape
```

Out[4]: (56, 41)

In [5]:
```
# This code displays a histogram of the school data
plt.figure(figsize=(45,15))
plt.hist(x = school_1["District"],  bins=20, edgecolor = "black")
plt.title("Texas School Districts", fontsize=40, weight='bold')
plt.xlabel("Schools", fontsize=25, weight='bold' )
plt.ylabel("Count", fontsize=25, weight='bold')
plt.xticks(rotation = 90, fontsize= 15)
plt.yticks(fontsize= 15)

plt.show()
```



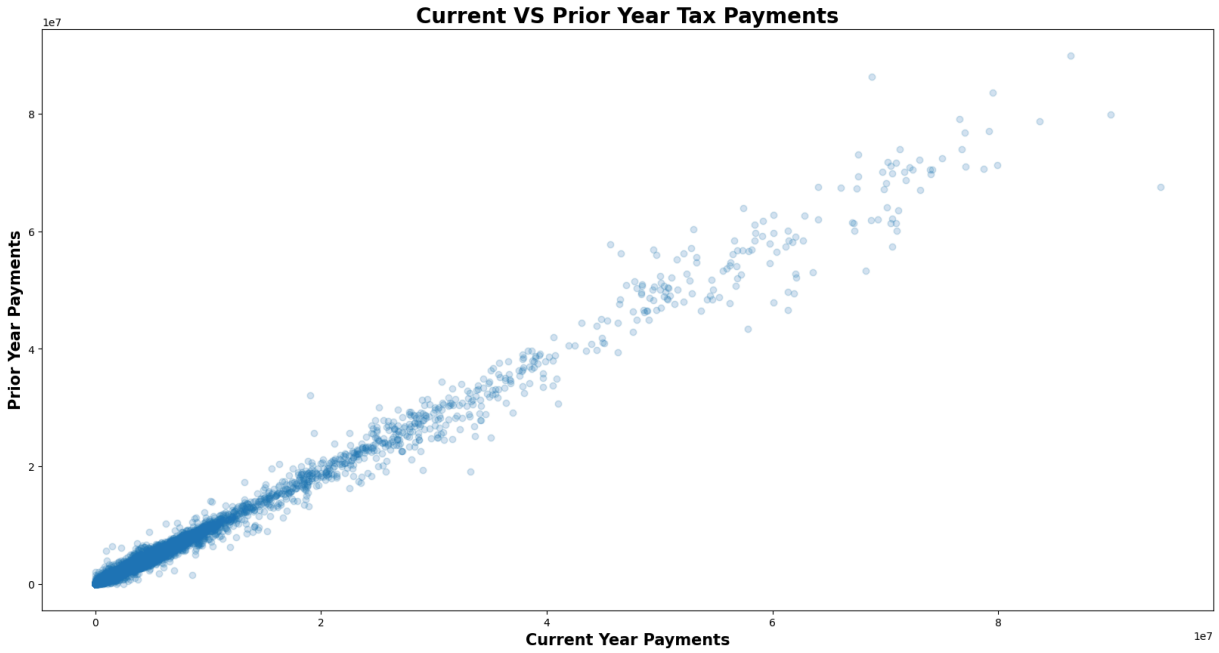

In this first graph I want to visualize the total schools in each district. This is important because larger school districts may have a correlation to population growth and I expect this will play a role in the decision to move to a certain area. After viewing this data I noticed that the Cold Spring/Oakhurst and Ft Worth school districts are the largest districts in Texas. These districts are coming in at approximately 120 schools respectively.

In [ ]:
```
# This code imports the sales tax allocation data set and displays the first 5 rows
sales_tax = pd.read_csv(r"Sales_Tax_Allocation__City_20250420.csv")
sales_tax.head()
```

Out[ ]:

| | City | Net Payment This Period | Comparable Payment Prior Year | Percent Change From Prior Year | Payment to Date | Previous Payments to Date | Percent Change To Date | Report Month |
|---|---|---|---|---|---|---|---|---|
| **0** | Abbott | 14301.88 | 15919.41 | -10.16 | 146889.53 | 152146.26 | -3.45 | 11 |
| **1** | Abernathy | 23516.10 | 23987.13 | -1.96 | 270303.69 | 256849.99 | 5.23 | 11 |
| **2** | Abilene | 5526055.82 | 5477084.93 | 0.89 | 57014074.04 | 53887697.25 | 5.80 | 11 |
| **3** | Ackerly | 8877.41 | 12792.00 | -30.60 | 86833.38 | 139977.10 | -37.96 | 11 |
| **4** | Addison | 1588208.70 | 1316545.16 | 20.63 | 15730647.58 | 14677334.92 | 7.17 | 11 |

In [ ]:
```
# This code creates a scatter plot showing the comparison of the current year and p
plt.figure(figsize = (20,10))
plt.scatter(x = 'Net Payment This Period',y ='Comparable Payment Prior Year',  data
plt.title("Current VS Prior Year Tax Payments", fontsize=20, weight='bold')
plt.xlabel("Current Year Payments", fontsize=15, weight='bold' )
plt.ylabel("Prior Year Payments", fontsize=15, weight='bold')
plt.show()
```




In this graph we are comparing current and prior year sales taxes paid. This is an important metric. Texans don't pay state income tax. All their revenue is derived from sales tax. This will indicate the areas where more spending takes place. The thought is areas that are spending more will have more affluent residents and will be more attractive to individuals when they decide to move.

In [ ]:
```
# This code imports the population data set and displays the first 5 rows
pop = pd.read_csv(r"population data.csv")
pop.head()
```

Out[ ]:

| | migration_scenario | year | year_month | FIPS | area_name | age_in_yrs_num | age_in_yrs_cha |
|---|---|---|---|---|---|---|---|
| **0** | Mid | 2020 | 202004 | 0 | State of Texas | -1 | All Age |
| **1** | Mid | 2020 | 202004 | 0 | State of Texas | 0 | < 1 y |
| **2** | Mid | 2020 | 202004 | 0 | State of Texas | 1 | 1 y |
| **3** | Mid | 2020 | 202004 | 0 | State of Texas | 2 | 2 yr |
| **4** | Mid | 2020 | 202004 | 0 | State of Texas | 3 | 3 yr |

5 rows × 25 columns

In [ ]:
```
# This code creates a subset of the population data and removes the first row which
# respective columns. It also displays the first 5 rows of the data set
pop_sub = pop[['year', 'total_male', 'total_female','total' ]]
```

```
pop_sub = pop_sub.iloc[1: ]
pop_sub.head()
```

Out[ ]:

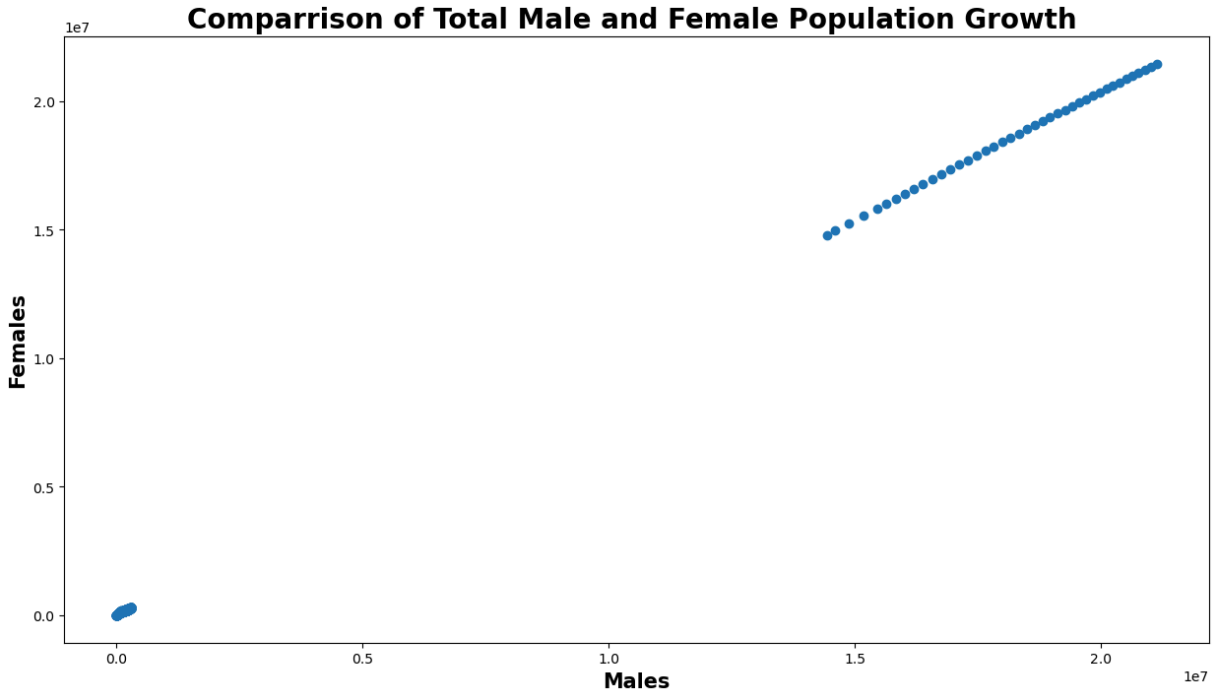
| | year | total_male | total_female | total |
|---|---|---|---|---|
| **1** | 2020 | 173553 | 167009 | 340562 |
| **2** | 2020 | 176404 | 170602 | 347006 |
| **3** | 2020 | 183499 | 177691 | 361190 |
| **4** | 2020 | 192237 | 185157 | 377394 |
| **5** | 2020 | 199902 | 193206 | 393108 |

In [ ]:
```
# This code adds two new columns to show percentage change
# between the respective columns
pop_sub['total_male_Percentage'] = pop_sub['total_male'].apply(lambda x: (x / pop_s
pop_sub['total_female_Percentage'] = pop_sub['total_female'].apply(lambda x: (x / p
pop_sub.head()
```

Out[ ]:

| | year | total_male | total_female | total | total_male_Percentage | total_female_Percentage |
|---|---|---|---|---|---|---|
| **1** | 2020 | 173553 | 167009 | 340562 | 0.011549 | 0.010881 |
| **2** | 2020 | 176404 | 170602 | 347006 | 0.011738 | 0.011115 |
| **3** | 2020 | 183499 | 177691 | 361190 | 0.012211 | 0.011577 |
| **4** | 2020 | 192237 | 185157 | 377394 | 0.012792 | 0.012063 |
| **5** | 2020 | 199902 | 193206 | 393108 | 0.013302 | 0.012588 |

In [ ]:
```
# This code creates a scatter plot comparing total male and female populations
plt.figure(figsize = (15,8))
plt.scatter(x = 'total_male',y ='total_female',  data = pop_sub )
plt.title("Comparrison of Total Male and Female Population Growth ", fontsize=20, w
plt.xlabel("Males", fontsize=15, weight='bold' )
plt.ylabel("Females", fontsize=15, weight='bold')
plt.show()
```




The purpose of this graph is to compare the total growth between male and females in Texas. The graph didn't turn out as expected. I tried different variations and inputs but with no success. I need to spend more time on this data set. It could be a factor of the age break down in the data.

In [ ]:
```
# This code imports the data set into a pandas data frame and displays the first 10
key_eco = pd.read_csv(r"Key_Economic_Indicators_20250420.csv")

key_eco.head(10)
```

Out[ ]:

| | Month | Year | Consumer Confidence Index TX | Consumer Confidence West South Central | Consumer Confidence Index US | PCE Deflator | Consumer Price Index TX | Consumer Price Index U.S. | U.S F Ene |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **1** | 2 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **2** | 3 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **3** | 4 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **4** | 5 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **5** | 6 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **6** | 7 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **7** | 8 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **8** | 9 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |
| **9** | 10 | 2005 | NaN | NaN | NaN | NaN | NaN | NaN | |

10 rows × 31 columns

In [ ]:
```
# This code drops the NaN values in the data set
key_eco_1 = key_eco.dropna()
key_eco_1.head()
```

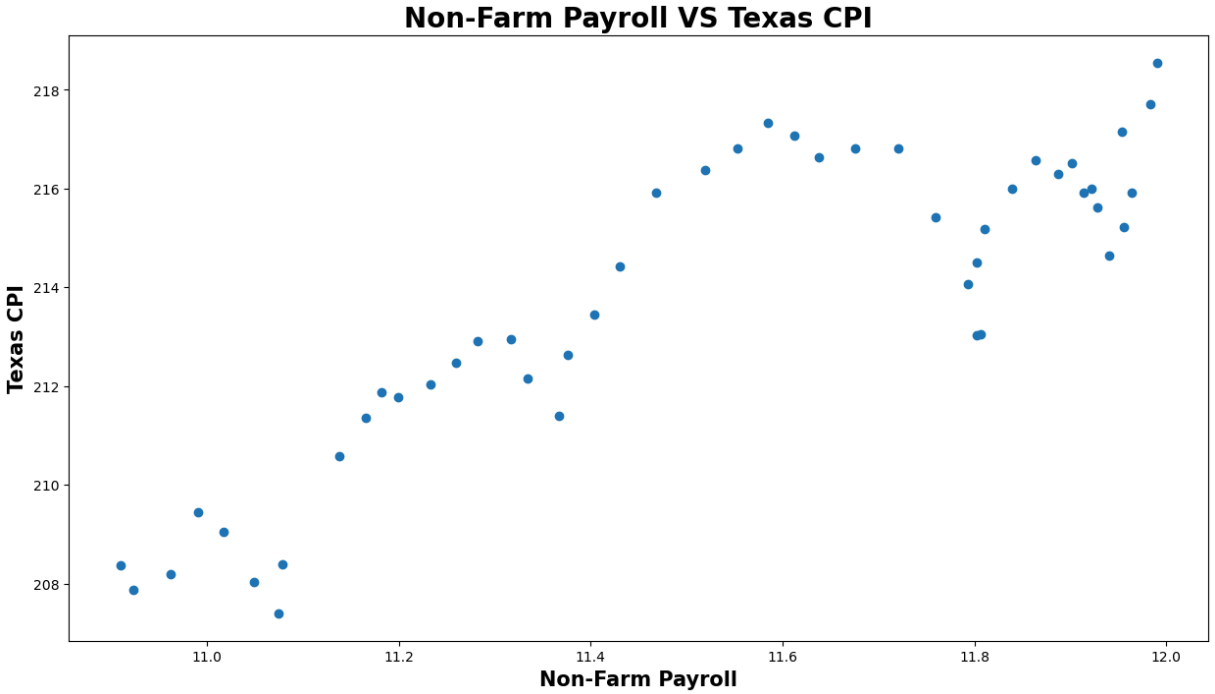
Out[ ]:

| | Month | Year | Consumer Confidence Index TX | Consumer Confidence West South Central | Consumer Confidence Index US | PCE Deflator | Consumer Price Index TX | Consumer Price Index U.S. | U E |
|---|---|---|---|---|---|---|---|---|---|
| **89** | 6 | 2012 | 104.8 | 95.6 | 62.7 | 105.844 | 208.372 | 229.478 | 22 |
| **90** | 7 | 2012 | 93.7 | 92.6 | 65.4 | 105.880 | 207.881 | 229.104 | 22 |
| **91** | 8 | 2012 | 75.1 | 74.9 | 61.3 | 106.238 | 208.201 | 230.379 | 23 |
| **92** | 9 | 2012 | 83.7 | 78.5 | 68.4 | 106.576 | 209.451 | 231.407 | 23 |
| **93** | 10 | 2012 | 92.0 | 84.2 | 73.1 | 106.886 | 209.044 | 231.317 | 23 |

5 rows × 31 columns

In [ ]:
```
#This code create a scatter plot for Non-farm Payroll and Texas CPI
plt.figure(figsize = (15,8))
plt.scatter(x = 'Nonfarm Employment TX',y ='Consumer Price Index TX',  data = key_e
plt.title("Non-Farm Payroll VS Texas CPI", fontsize=20, weight='bold')
```

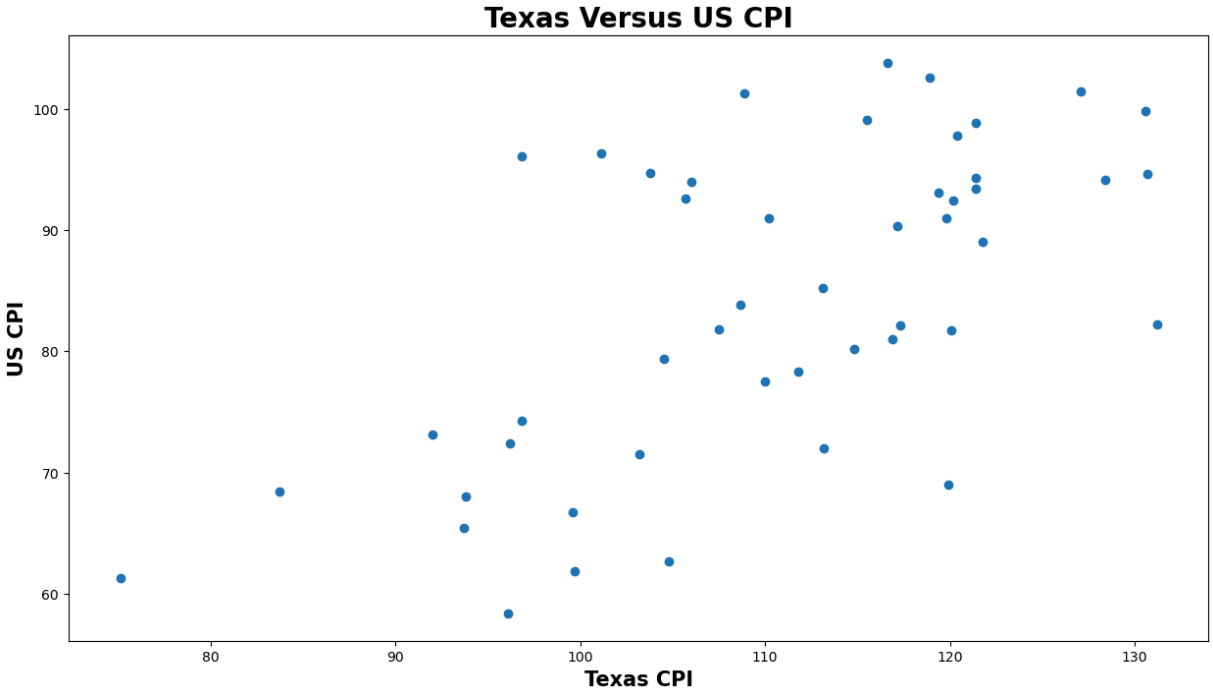
```
plt.xlabel("Non-Farm Payroll", fontsize=15, weight='bold' )
plt.ylabel("Texas CPI", fontsize=15, weight='bold')
plt.show()
```




This graph shows a comparison between non-farm payroll and the Texas CPI. Understanding the correlation between these two features is important because it will help to uncover the economic condition of the cities within Texas. How well do these respective economies respond to changes in price. Is there a correlation between population growth and the elasticity of the economy

In [ ]:
```
sub_eco= key_eco_1[['Year', 'Consumer Confidence Index TX', 'Consumer Confidence In
```

In [ ]:
```
# This code creates the scatter plot for US and Texas CCI
plt.figure(figsize = (15,8))
plt.scatter(x = 'Consumer Confidence Index TX',y ='Consumer Confidence Index US',
plt.title("Texas Versus US Consumer Confidence", fontsize=20, weight='bold')
plt.xlabel("Texas Consumer Confidence", fontsize=15, weight='bold' )
plt.ylabel("US Consumer Confidence", fontsize=15, weight='bold')
plt.show()
```




This graph shows a comparison between the Texas CCI and the US CCI. This graph does show a correlation between the US and Texas CCI markers. The correlation is relatively weak and I see a few outliers. However there is a positive relationship between the two features and I will explore this further.