project-guidelines

June 16, 2023

1 Course project guidelines

Your assignment for the course project is to formulate and answer a question of your choosing based on one of the following datasets:

1. ClimateWatch historical emissions data: greenhouse gas emissions by U.S. state 1990-present
2. World Happiness Report 2023: indices related to happiness and wellbeing by country 2008-present
3. Any dataset from the class assignments or mini projects

A good question is one that you want to answer. It should be a question with contextual meaning, not a purely technical matter. It should be clear enough to answer, but not so specific or narrow that your analysis is a single line of code. It should require you to do some nontrivial exploratory analysis, descriptive analysis, and possibly some statistical modeling. You aren't required to use any specific methods, but it should take a bit of work to answer the question. There may be multiple answers or approaches to contrast based on different ways of interpreting the question or different ways of analyzing the data. If your question is answerable in under 15 minutes, or your answer only takes a few sentences to explain, the question probably isn't nuanced enough.

1.1 Deliverable

Prepare and submit a jupyter notebook that summarizes your work. Your notebook should contain the following sections/contents:

- **Data description**: write up a short summary of the dataset you chose to work with following the conventions introduced in previous assignments. Cover the sampling if applicable and data semantics, but focus on providing high-level context and not technical details; don't report preprocessing steps or describe tabular layouts, etc.
- **Question of interest**: motivate and formulate your question; explain what a satisfactory answer might look like.
- **Data analysis**: provide a walkthrough with commentary of the steps you took to investigate and answer the question. This section can and should include code cells and text cells, but you should try to focus on presenting the analysis clearly by organizing cells according to the high-level steps in your analysis so that it is easy to skim. For example, if you fit a regression model, include formulating the explanatory variable matrix and response, fitting the model, extracting coefficients, and perhaps even visualization all in one cell; don't separate these into 5-6 substeps.
- **Summary of findings**: answer your question by interpreting the results of your analysis, referring back as appropriate. This can be a short paragraph or a bulleted list.

1.2 Evaluation

Your work will be evaluated on the following criteria:

1. Thoughtfulness: does your question reflect some thoughtful consideration of the dataset and its nuances, or is it more superficial?
2. Thoroughness: is your analysis an end-to-end exploration, or are there a lot of loose ends or unexplained choices?
3. Mistakes or oversights: is your work free from obvious errors or omissions, or are there mistakes and things you've overlooked?
4. Clarity of write-up: is your report well-organized with commented codes and clear writing, or does it require substantial effort to follow?

2 Data Description:

- We are choosing the data from homework 2 - SEDA data.

The Stanford Educational Data Archive (SEDA) includes a range of detailed data on educational conditions, contexts, and outcomes in school districts and counties across the United States in 2018. It includes measures of academic achievement and achievement gaps for school districts and counties, as well as district-level measures of racial and socioeconomic composition, racial and socioeconomic segregation patterns, and other features of the schooling system. This dataset is probably data from a typical sample. The scope of inference for this dataset is limited to the specific schools and districts included in the sample, which is not drawn using a probability sampling method, and the sampling frame is not clearly identified. After tidying the data, we have the following columns:

District ID: represents a unique identifier assigned to each district in the dataset.

Locale: provides information about the classification of each district based on its location. It indicates whether the district is rural, urban, suburban, or falls into another category.

log(Median income): represents the natural logarithm of the median income in each district. It provides an indicator of the economic prosperity or wealth in the district.

Poverty rate: indicates the percentage of individuals living below the poverty line in each district. It helps assess the level of poverty within the district.

Unemployment: provides the percentage of unemployed individuals in each district. It reflects the level of joblessness within the district.

SNAP rate: represents the proportion of the population in each district that receives benefits from the Supplemental Nutrition Assistance Program (SNAP). It gives insights into the prevalence of food assistance needs in the district.

Socioeconomic index: contains a composite index that measures the overall socioeconomic status of each district. It takes into account various socioeconomic factors and provides an overall assessment of the district's well-being.

3 Question of Interest:

Our question of Interest would be that, based on the Seda dataset in 2018, what is the relationship between the Unemployment rate and factors such as Median income, Locale, Poverty rate, SNAP rate, and Socioeconomic index? The motivation is to understand the relationship between the Unemployment rate and socioeconomic factors is crucial for analyzing the dynamics of labor markets and identifying key determinants of employment outcomes. By examining the impact of the Median income, Locale, Poverty rate, SNAP rate, and Socioeconomic index on the Unemployment rate, we can gain insights into the complex interplay between these variables and their influence on employment opportunities and economic well-being.

The Formulation of the question would be how do Unemployment rates relate to Median income, Locale, Poverty rate, SNAP rate, and Socioeconomic index? Are specific locales, higher poverty rates, lower Socioeconomic index scores, or the SNAP rate influential in impacting the Unemployment rate? A satisfactory answer to our question of interest would involve a relatively comprehensive analysis of the relationships between the Unemployment rate and factors like median income, Locale, Poverty rate, SNAP rate, and Socioeconomic index. The analysis should include exploratory analysis, descriptive statistics, and potential hypothesis testing or regression modeling to examine the strength and significance of the relationships. The answer should provide insights into the associations between Unemployment rates and these socioeconomic factors, highlighting any significant patterns, trends, or correlations. It should also address potential confounding factors and limitations, providing a nuanced understanding of how median income, Locale, Poverty rate, SNAP rate, and Socioeconomic index contribute to unemployment outcomes. Ultimately, a satisfactory answer would enhance our understanding of the socioeconomic determinants of Unemployment rates and inform policy discussions aimed at reducing unemployment and promoting economic well-being.

[1]:
```
import numpy as np
import pandas as pd
import altair as alt
# disable row limit for plotting
alt.data_transformers.disable_max_rows()
# uncomment to ensure graphics display with pdf export
alt.renderers.enable('mimetype')
```

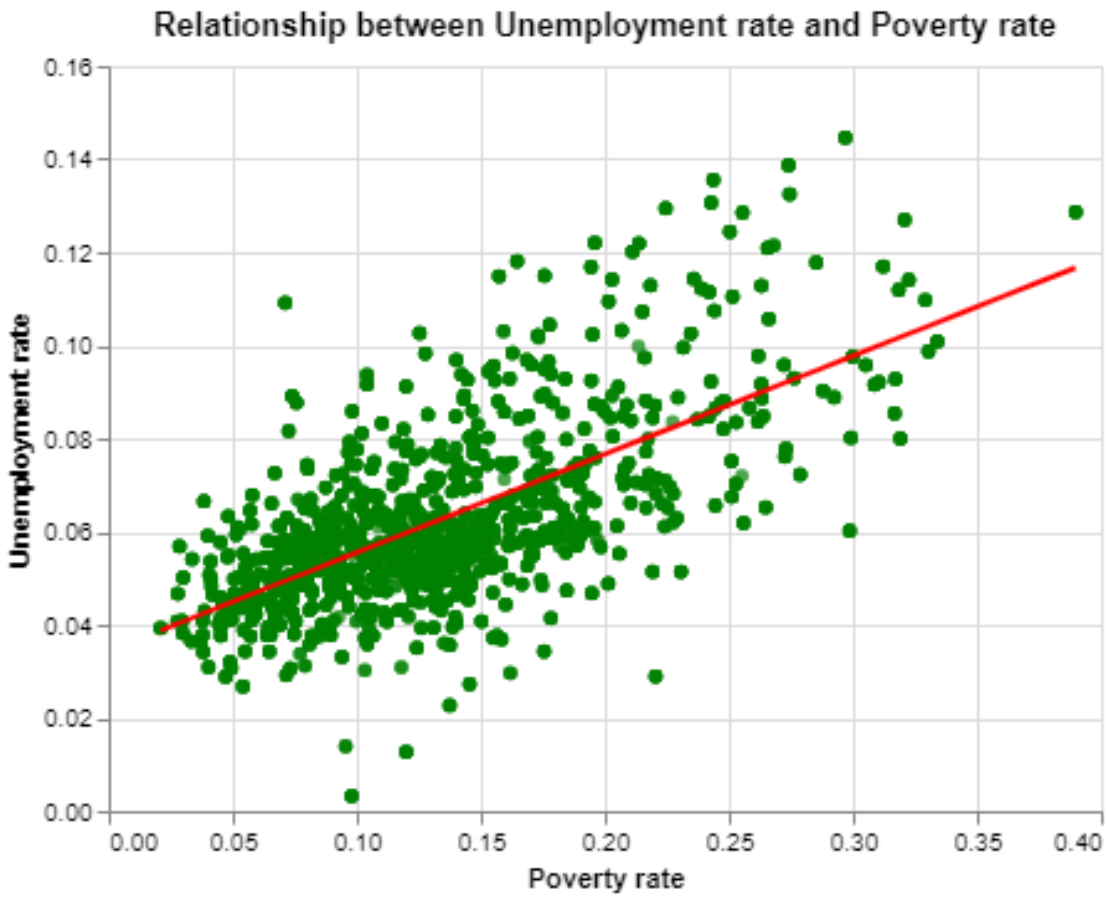
[1]:
```
RendererRegistry.enable('mimetype')
```

[2]:
```
data = pd.read_csv('cleaned_data.csv')
```

[3]:
```
# What unemployment rate is associated with?

# What is the relationship between Unemployment rate and Poverty rate?

poverty_plot = alt.Chart(data).mark_circle(color = 'green').encode(
    x=alt.X('Poverty rate'),
    y=alt.Y('Unemployment rate'),
    tooltip = ['Poverty rate','Unemployment rate']
).properties(
    title = 'Relationship between Unemployment rate and Poverty rate'
```

```
)

line_of_fit = poverty_plot.transform_regression(
    'Poverty rate', 'Unemployment rate', method='poly', order=1
).mark_line(color='red')

poverty = poverty_plot + line_of_fit

poverty
```

[3]:

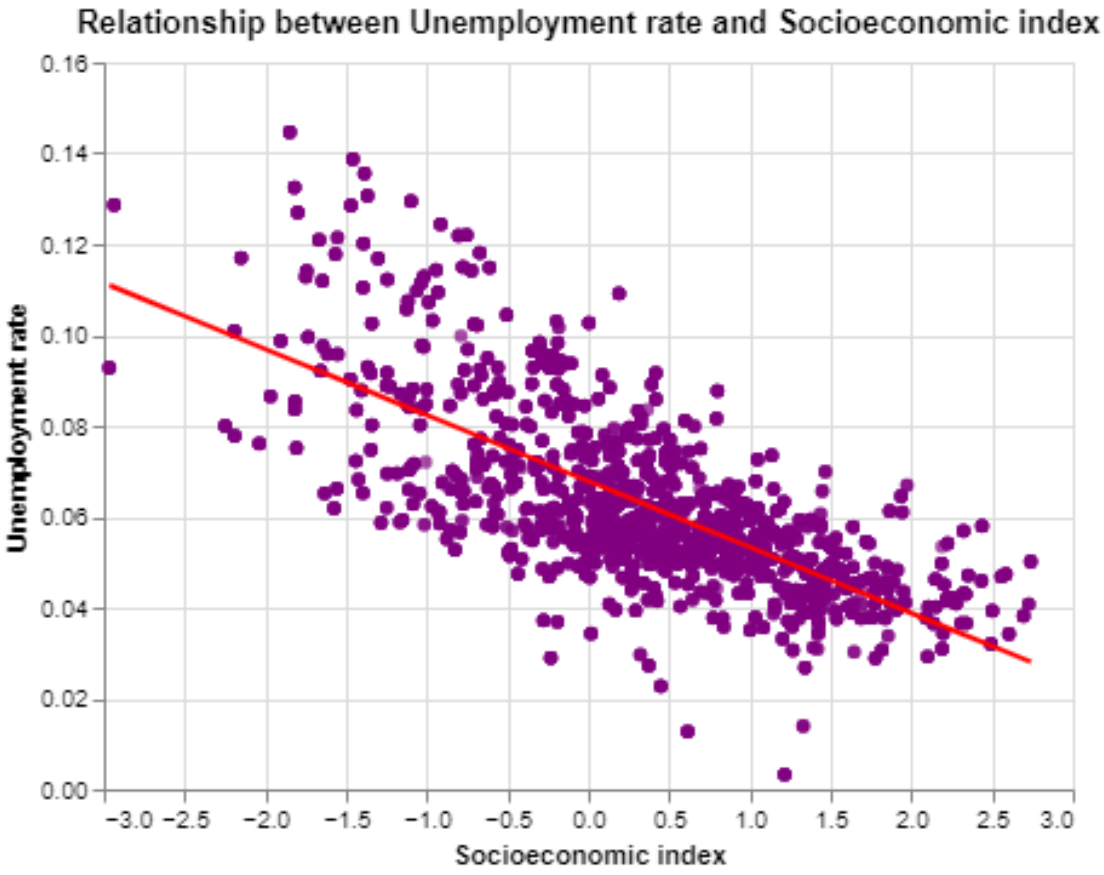


[4]:

```
# What is the relationship between Unemployment rate and socioeconomic index?

si_plot = alt.Chart(data).mark_circle(color = 'purple').encode(
    x=alt.X('Socioeconomic index'),
    y=alt.Y('Unemployment rate'),
    tooltip = ['Socioeconomic index','Unemployment rate']
).properties(
    title = 'Relationship between Unemployment rate and Socioeconomic index'
)
```

```
line_of_fit = si_plot.transform_regression(
    'Socioeconomic index', 'Unemployment rate', method='poly', order=1
).mark_line(color='red')

si = si_plot + line_of_fit

si
```

[4]:

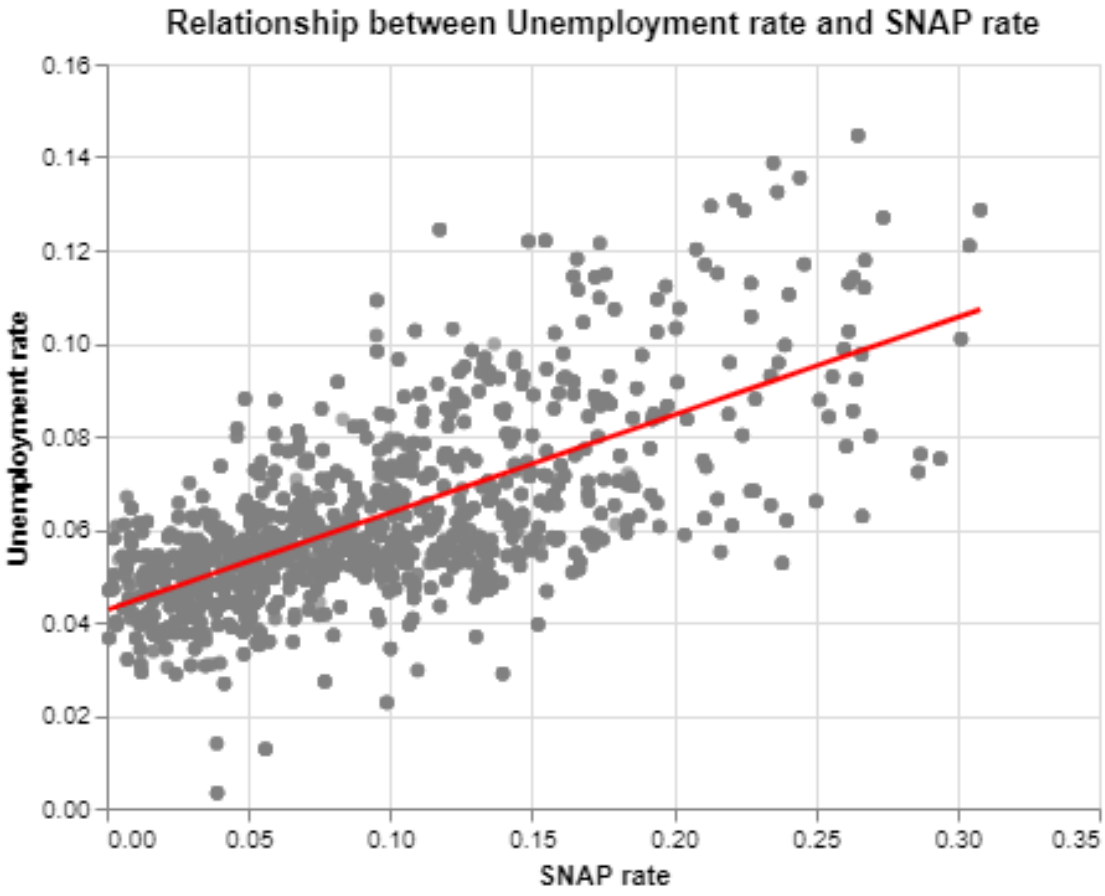


[5]:

```
# What is the relationship between Unemployment rate and SNAP rate?

snap_plot = alt.Chart(data).mark_circle(color = 'grey').encode(
    x=alt.X('SNAP rate'),
    y=alt.Y('Unemployment rate'),
    tooltip = ['SNAP rate','Unemployment rate']
).properties(
    title = 'Relationship between Unemployment rate and SNAP rate'
)
```

```
line_of_fit = snap_plot.transform_regression(
    'SNAP rate', 'Unemployment rate', method='poly', order=1
).mark_line(color='red')

snap = snap_plot + line_of_fit

snap
```

[5]:

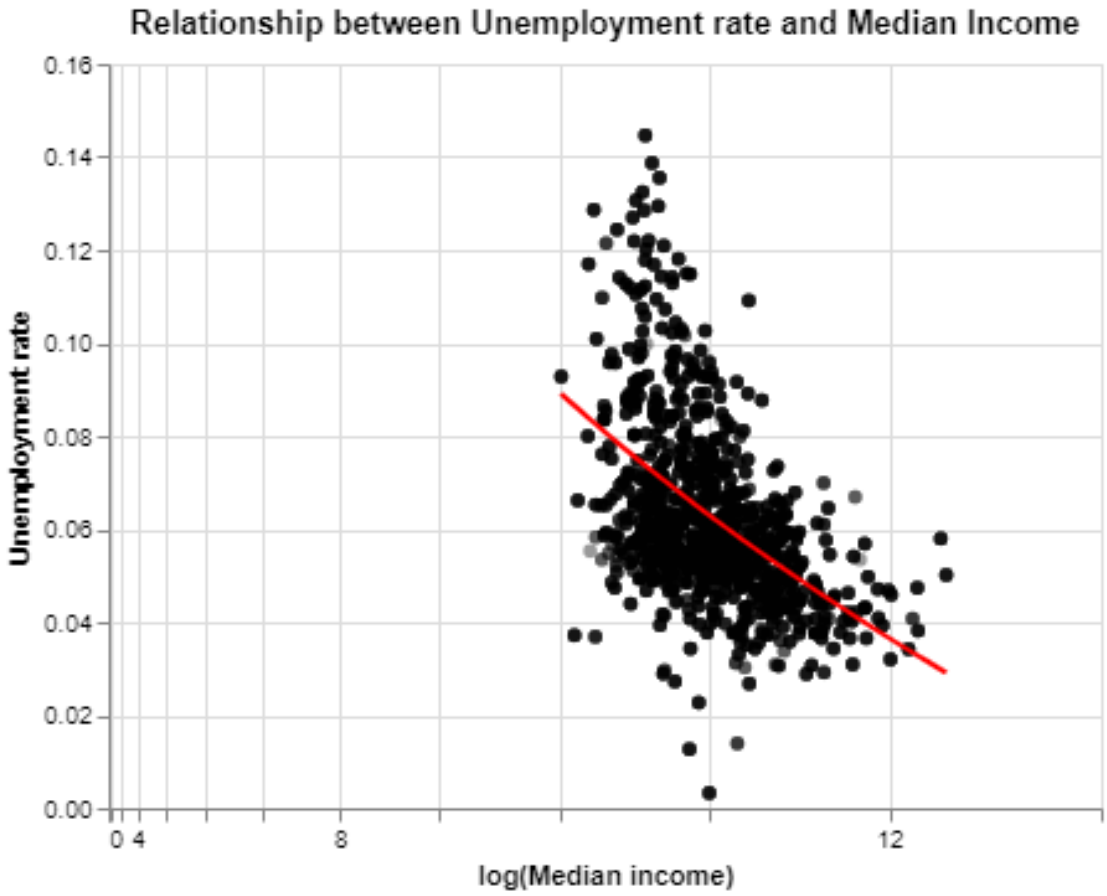


[6]:

```
# What is the relationship between Unemployment rate and median income?

income_plot = alt.Chart(data).mark_circle(color='black', opacity=0.4).encode(
    x=alt.X('log(Median income)', scale=alt.Scale(type='pow',exponent=3)),
    y=alt.Y('Unemployment rate', scale=alt.Scale(zero=False)),
    tooltip=['log(Median income)', 'Unemployment rate']
).properties(
    title='Relationship between Unemployment rate and Median Income'
)

# Add a line of best fit
```

```
line_of_fit = income_plot.transform_regression(
    'log(Median income)', 'Unemployment rate', method='poly', order=1
).mark_line(color='red')

# Combine the scatter plot and line of best fit
income = income_plot + line_of_fit

income
```

[6]:

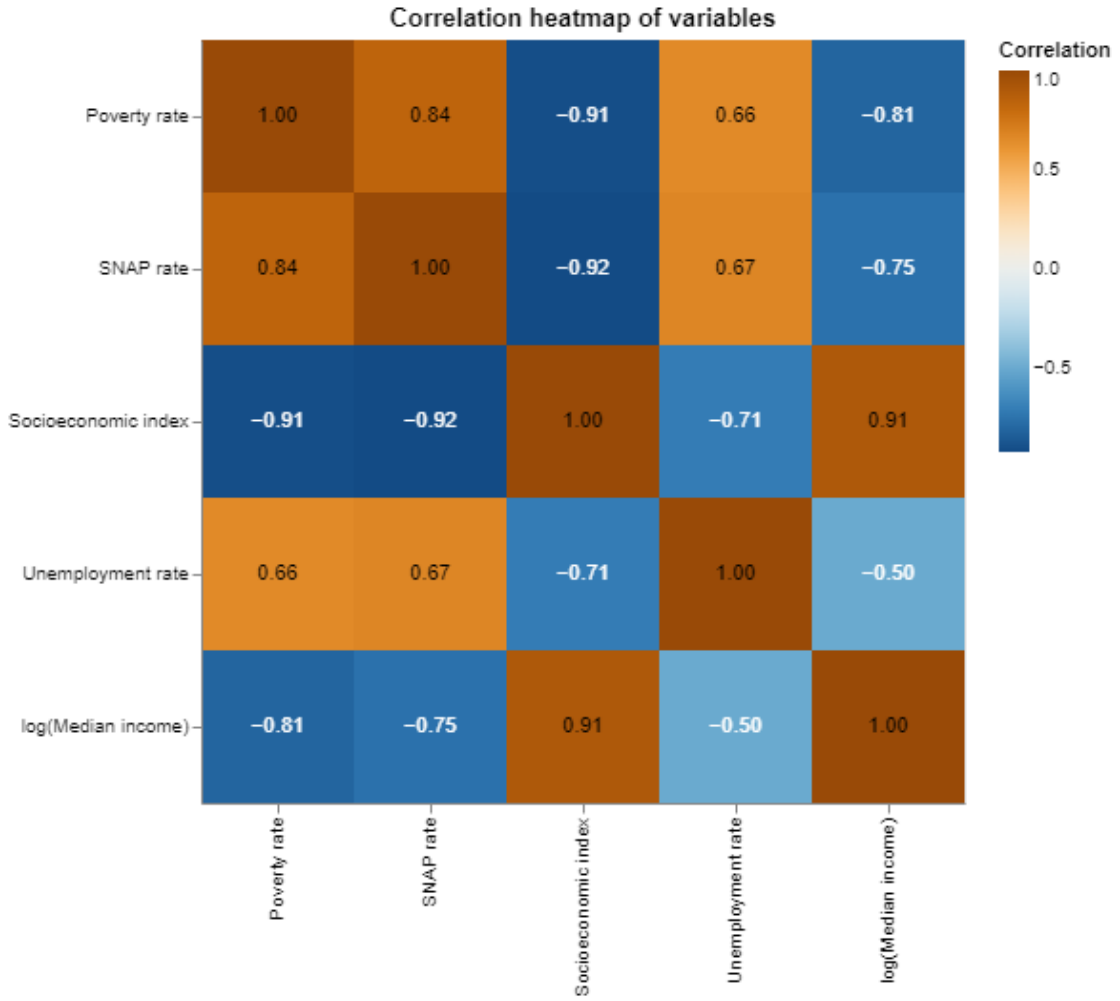


[7]:

```
# Any correlations between these factors?

# Calculating correlations of each variables
correlations = data[['log(Median income)', 'Unemployment rate', 'Poverty rate',␣
 ↪'SNAP rate', 'Socioeconomic index']].corr()

# generate Correlation matrix
correlations = correlations.reset_index().melt('index')
```

[8]:
```
correlations.columns = ['Variable 1', 'Variable 2', 'Correlation']

# heatMap
heatmap = alt.Chart(correlations).mark_rect().encode(
    x=alt.X('Variable 1:O', title = ''),
    y=alt.Y('Variable 2:O', title = ''),
    color=alt.Color('Correlation:Q', scale=alt.Scale(scheme='blueorange')),
).properties(
    width=400,
    height=400,
    title='Correlation heatmap of variables'
)

# adding text to each block show correlation value
text = heatmap.mark_text(baseline='middle').encode(
    text=alt.Text('Correlation:Q', format='.2f'),
    color=alt.condition(
        alt.datum.Correlation > 0.5,
        alt.value('black'),
        alt.value('white')
    )
)

HeatMap = heatmap + text

HeatMap
```

[8]:



In the first part of the analysis, we have ploted Unemployment rate against each dependent variable and created a heat map. We may see that there is a strong positive relationship between Unemployment and Poverty rate, indicating that as Poverty rate increases, the Unemployment rate will increase. It implies that areas with higher poverty rates are more likely to have higher unemployment rates. This observation aligns with the common understanding that economic disadvantage and limited access to resources can contribute to higher unemployment rates. We may see that there is a strong negative relationship between Unemployment and Socioeconomic index, indicating that as Socioeconomic index increases, the Unemployment rate will decrease. Socioeconomic index, which serves as a measure of overall socioeconomic well-being or advantage, increases, it implies a higher level of economic stability and access to resources. Consequently, areas or districts with higher Socioeconomic index scores are more likely to have lower unemployment rates. We may see that there is a strong positive relationship between Unemployment and SNAP rate, indicating that as SNAP rate increases, the Unemployment rate will increase. This positive relationship indicates that areas or districts with higher unemployment rates are more likely to have higher SNAP participation rates. We may see that there is a strong negative relationship between Unemployment and log(Median Income), indicating that as log(Median Income) increases, the Unemployment rate

will decrease. It implies that areas with higher log(Median Income) are more likely to have lower unemployment rates.

From the heatmap, we can see that Poverty rate is positively correlated with SNAP rate and Unemployment rate. Socioeconomic index is negatively correlated with Poverty rate and SNAP rate and Unempoyment rate. Income is negatively correlated with Poverty rate, SNAP rate, and Unemployment rate. Overall, high unemployment rate probably would coincide with a high poverty rate, a high SNAP rate, and a low socioeconomic index, and a low income. The observation aligns with economic common sense.

[9]:
```
import statsmodels.api as sm

# areate a DataFrame with the relevant columns
df = pd.DataFrame({
    'Unemployment rate': data['Unemployment rate'],
    'Poverty rate': data['Poverty rate']
})

# add a constant column for the intercept term
df = sm.add_constant(df)

# fit the linear regression model
model = sm.OLS(df['Unemployment rate'], df[['const', 'Poverty rate']])
results = model.fit()
```

[10]:
```
coef_tbl = pd.DataFrame({'estimate': results.params.values,
              'standard error': results.bse})
coef_tbl.loc['error variance', 'estimate'] = results.scale

coef_tbl
```

[10]:

	estimate	standard error
const	0.034385	0.000589
Poverty rate	0.211202	0.003951
error variance	0.000228	NaN

Lastly, a standard metric often reported with linear models is the ^2 score, which is interpreted as the proportion of variation in the response captured by the model.

[11]:
```
results.rsquared
```

[11]:
```
0.4319475169769662
```

This indicates that the model explains or predicts 43.2% of the relationship between the dependent (Unemployment rate) and independent variables.

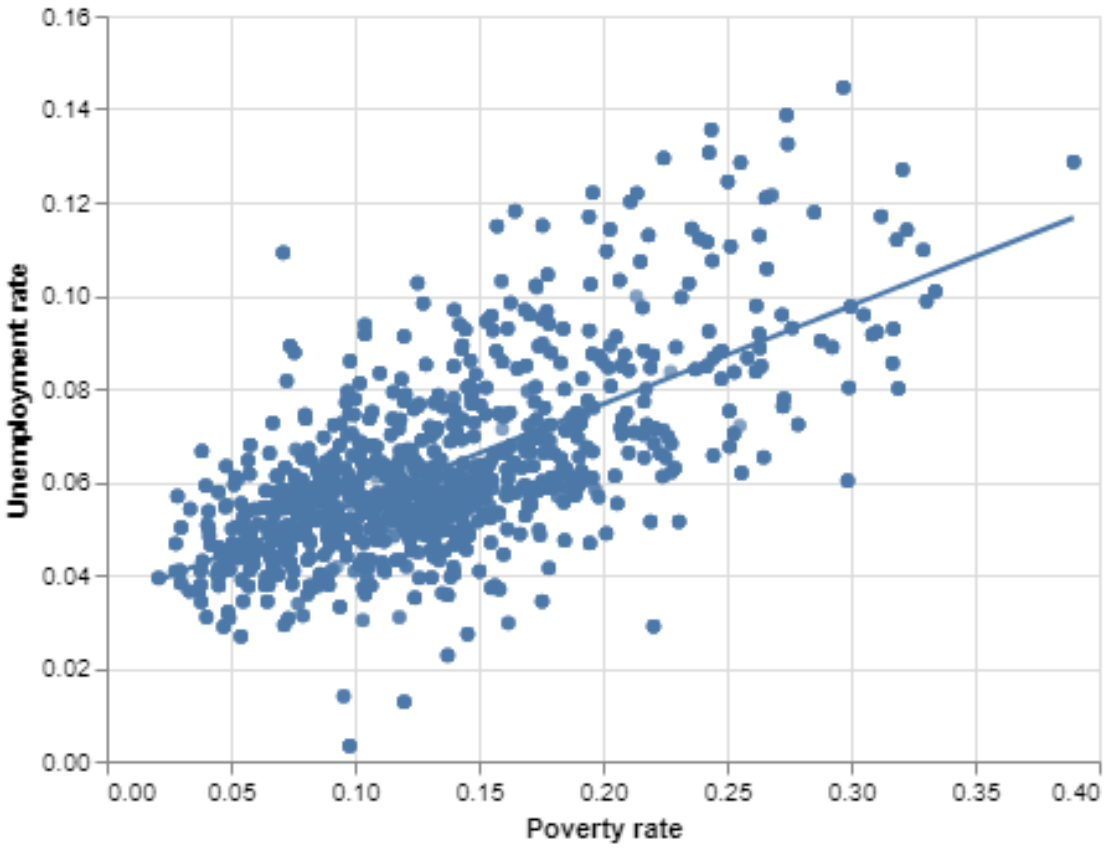
[12]:
```
# append fitted values and residuals
data['fitted_slr'] = results.fittedvalues
data['resid_slr'] = results.resid
```

```
data.head(3)
```

[12]:
```
   District ID          Locale  log(Median income)  Poverty rate
0       600001  Rural, Distant           11.392048      0.091894  \
1       600001  Rural, Distant           11.392048      0.091894
2       600001  Rural, Distant           11.392048      0.091894

   Unemployment rate  SNAP rate  Socioeconomic index  fitted_slr  resid_slr
0           0.048886   0.035165             1.237209    0.053793  -0.004907
1           0.048886   0.035165             1.237209    0.053793  -0.004907
2           0.048886   0.035165             1.237209    0.053793  -0.004907
```

[13]:
```
# construct line plot

scatter_pov = alt.Chart(data).mark_circle().encode(
    x=alt.X('Poverty rate:Q', title='Poverty rate', scale=alt.
 ↪Scale(zero=False)),
    y=alt.Y('Unemployment rate:Q', title='Unemployment rate', scale=alt.
 ↪Scale(zero=False))
)

slr_line = alt.Chart(data).mark_line().encode(
    x = 'Poverty rate',
    y = 'fitted_slr'
)

# layer
scatter_pov + slr_line
```

[13]:

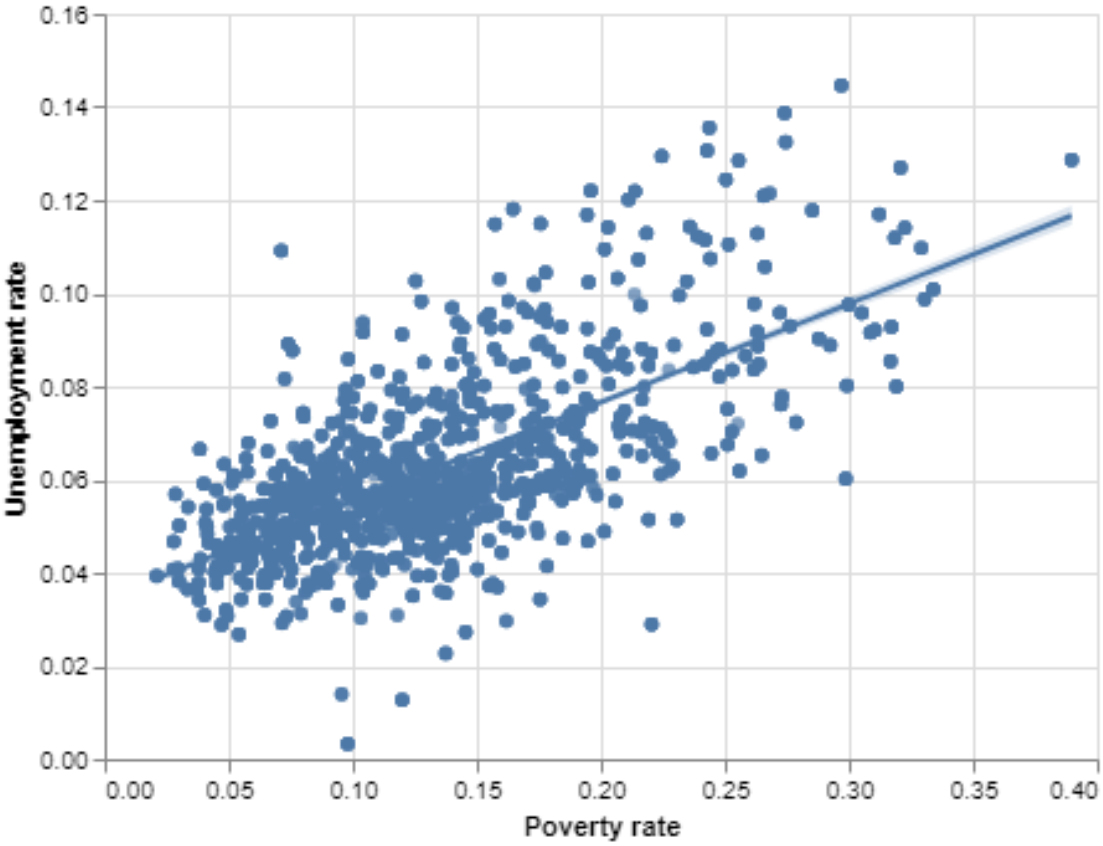


[14]:
```
df = df.drop(['Unemployment rate'], axis=1)
```

[15]:
```
preds = results.get_prediction(df)
data['lwr_mean'] = preds.predicted_mean - 2*preds.se_mean
data['upr_mean'] = preds.predicted_mean + 2*preds.se_mean
```

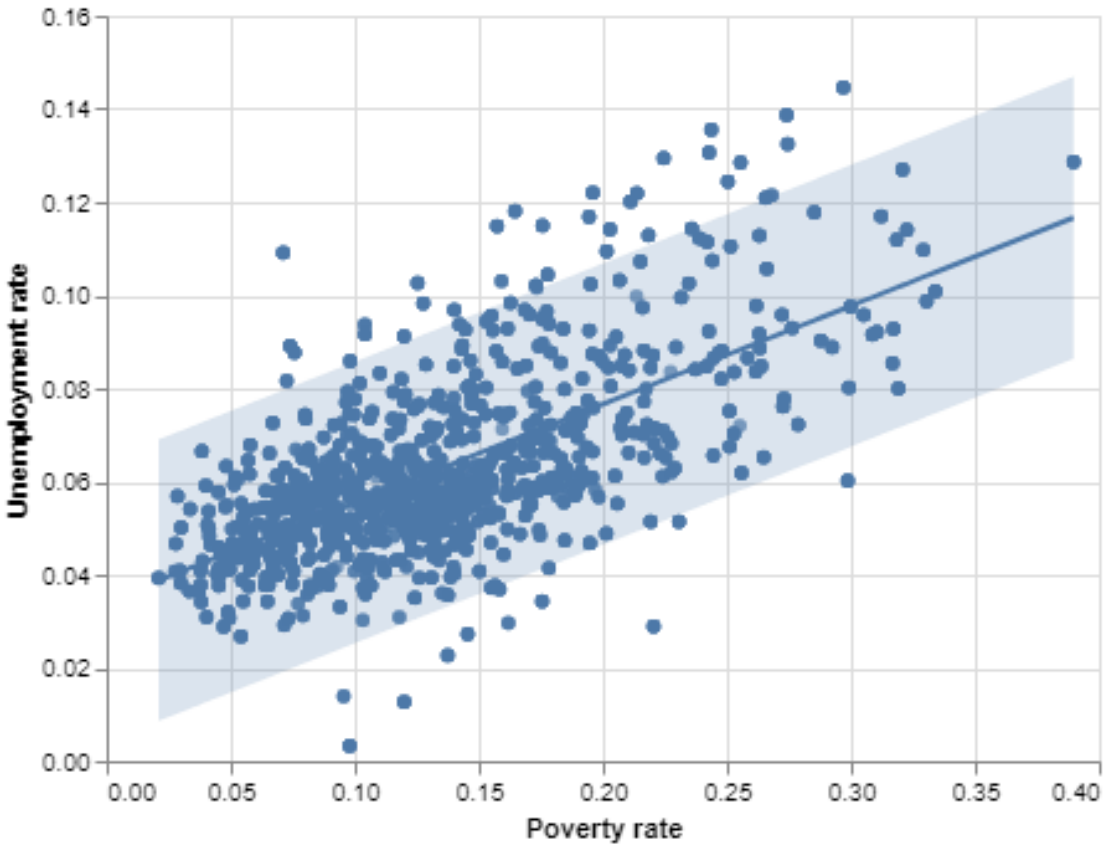
[16]:
```
# create bandwidth
band = alt.Chart(data).mark_area(opacity = 0.2).encode(
    x = 'Poverty rate',
    y = 'lwr_mean',
    y2 = 'upr_mean'
)

scatter_pov + slr_line + band
```

[16]:



[17]:
```
# compute prediction uncertainty bounds

data['lwr_obs'] = preds.predicted_mean - 2*preds.se_obs
data['upr_obs'] = preds.predicted_mean + 2*preds.se_obs

# construct plot showing prediction uncertainty
band = alt.Chart(data).mark_area(opacity = 0.2).encode(
    x = 'Poverty rate',
    y = 'lwr_obs',
    y2 = 'upr_obs'
)

scatter_pov + slr_line + band
```

[17]:



[18]:
```
# create a df with the relevant columns
df1 = pd.DataFrame({
    'Unemployment rate': data['Unemployment rate'],
    'Poverty rate': data['Poverty rate'],
    'Socioeconomic index': data['Socioeconomic index'],
    'Income':data['log(Median income)'],
    'SNAP rate':data['SNAP rate']
})

df1 = sm.add_constant(df1)
mlr = sm.OLS(df1['Unemployment rate'], df1[['const', 'Poverty rate',␣
 ↪'Socioeconomic index','Income','SNAP rate']])
results1 = mlr.fit()
results1.summary()
```

[18]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:      Unemployment rate   R-squared:                       0.633
```

```
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     1618.
Date:                Fri, 16 Jun 2023   Prob (F-statistic):               0.00
Time:                        07:36:43   Log-Likelihood:                 11253.
No. Observations:                3760   AIC:                        -2.250e+04
Df Residuals:                    3755   BIC:                        -2.246e+04
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
=======
                           coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
-------
const                   -0.4786      0.015    -32.068      0.000      -0.508
-0.449
Poverty rate             0.0120      0.008      1.597      0.110      -0.003
0.027
Socioeconomic index     -0.0380      0.001    -37.363      0.000      -0.040
-0.036
Income                   0.0512      0.001     36.123      0.000       0.048
0.054
SNAP rate               -0.1017      0.009    -11.541      0.000      -0.119
-0.084
==============================================================================
Omnibus:                      129.290   Durbin-Watson:                   0.456
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              234.399
Skew:                           0.273   Prob(JB):                     1.26e-51
Kurtosis:                       4.095   Cond. No.                         872.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

[19]:
```
coef_tbl = pd.DataFrame({
    'estimate': results1.params.values,
    'standard error': results1.bse
    }
)

coef_tbl.loc['error variance', 'estimate'] = results1.scale

coef_tbl
```

[19]:

	estimate	standard error
const	-0.478568	0.014924
Poverty rate	0.012028	0.007530
Socioeconomic index	-0.037981	0.001017
Income	0.051161	0.001416
SNAP rate	-0.101708	0.008813
error variance	0.000147	NaN

[20]:
```
results1.rsquared
```

[20]:
```
0.6328719924692203
```

In the second part of the analysis, a Multiple Linear Regression (MLR) model was constructed to investigate the relationship between unemployment and all variables except for Locale. The variables included in the model were Poverty rate, Socioeconomic index, Income, and SNAP rate. A constant term was also included in the model.

The MLR model would be the following:

$$\text{Unemployment rate} = -0.4786 + 0.0120 \times \text{Poverty rate} - 0.0380 \times \text{Socioeconomic index} + 0.0512 \times \text{Income} - 0.1017 \times \text{SN}$$

The results of the MLR model are as follows:

1. **Poverty rate**: The coefficient for the Poverty rate is 0.0120. This suggests that for every unit increase in the Poverty rate, the Unemployment rate increases by 0.0120 units, holding all other variables constant. However, the p-value associated with the Poverty rate is 0.110, which is greater than the typical significance level of 0.05. This indicates that the effect of Poverty rate on Unemployment rate is not statistically significant at the 5% level.
2. **Socioeconomic index**: The coefficient for the Socioeconomic index is -0.0380. This suggests that for every unit increase in the Socioeconomic index, the Unemployment rate decreases by 0.0380 units, holding all other variables constant. The p-value associated with the Socioeconomic index is less than 0.05, indicating that the effect of Socioeconomic index on Unemployment rate is statistically significant at the 5% level.
3. **Income**: The coefficient for Income is 0.0512. This suggests that for every unit increase in Income, the Unemployment rate increases by 0.0512 units, holding all other variables constant. The p-value associated with Income is less than 0.05, indicating that the effect of Income on Unemployment rate is statistically significant at the 5% level.
4. **SNAP rate**: The coefficient for the SNAP rate is -0.1017. This suggests that for every unit increase in the SNAP rate, the Unemployment rate decreases by 0.1017 units, holding all other variables constant. The p-value associated with the SNAP rate is less than 0.05, indicating that the effect of SNAP rate on Unemployment rate is statistically significant at the 5% level.

The R-squared value of the model is 0.633, suggesting that about 63.3% of the variation in the Unemployment rate can be explained by the included variables. This is a relatively high R-squared

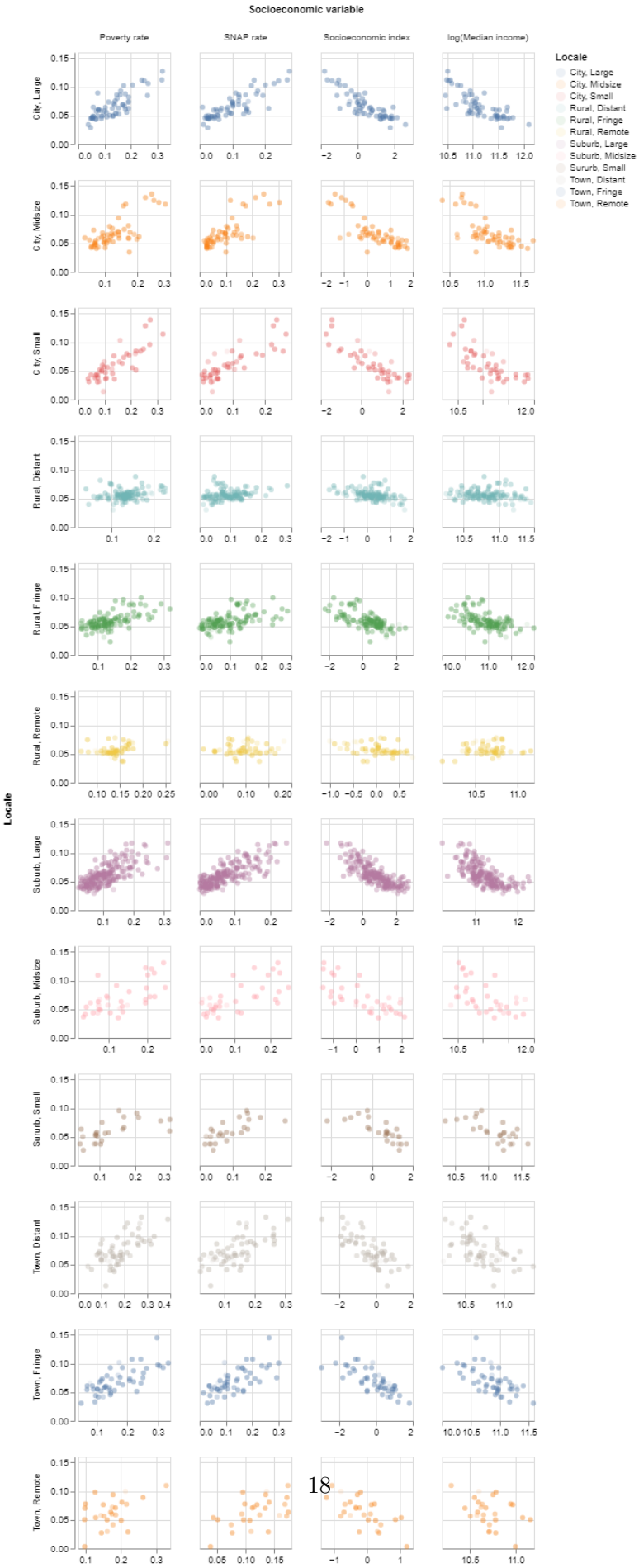
value, indicating that the model fits the data well. The prediction interval also shows that the majority of the data points will fall into the range.

These results provide valuable insights into the factors influencing unemployment.

[21]:
```
# ABout Locale
```

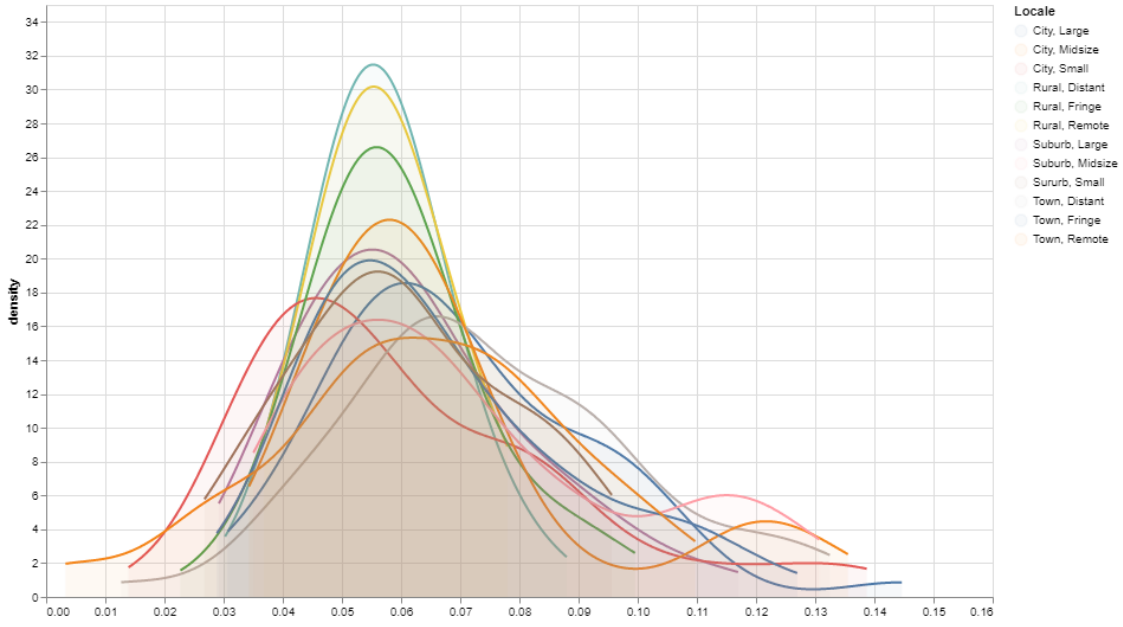
[22]:
```
# melt the dataframe for visualization of unemployment rate vs socioeconomic␣
 ↪variables by Locale
plot_df = data.melt(id_vars=['Locale', 'Unemployment rate'],
                    value_vars=['log(Median income)', 'Poverty rate', 'SNAP␣
 ↪rate', 'Socioeconomic index'],
                    var_name='Socioeconomic variable',
                    value_name='Measure')

# create plot and facet by Locale
locale_plot = alt.Chart(plot_df).mark_circle(opacity = 0.1).encode(
    y=alt.Y('Unemployment rate', title=''),
    x=alt.X('Measure:Q', scale = alt.Scale(zero = False), title = ''),
    color='Locale:N'
).properties(
    width=100,
    height=100
).facet(
    column=alt.Column('Socioeconomic variable:N'),
    row=alt.Row('Locale:N')
).resolve_scale(x='independent')

locale_plot
```

[22]:


Socioeconomic variable
Poverty rate
SNAP rate
Socioeconomic index
log(Median income)
Locale
City, Large
City, Midsize
City, Small
Rural, Distant
Rural, Fringe
Rural, Remote
Suburb, Large
Suburb, Midsize
Surburb, Small
Town, Distant
Town, Fringe
Town, Remote
Locale
City, Large
City, Midsize
City, Small
Rural, Distant
Rural, Fringe
Rural, Remote
Suburb, Large
Suburb, Midsize
Surburb, Small
Town, Distant
Town, Fringe
Town, Remote

[23]:
```
# Then create the KDE plot

kde = alt.Chart(data).transform_density(
    density = 'Unemployment rate',
    groupby=['Locale'],
    as_=['Unemployment rate', 'density'],
    bandwidth = 0.01,
    steps =1000
).mark_line().encode(
    y='density:Q',
    color='Locale:N',
    x=alt.X('Unemployment rate:Q',
        title=None,
    ),
).properties(
    width = 800,
    height = 500
)

chart2 = kde + kde.mark_area(opacity = 0.05)

chart2
```

[23]:



In the third part of the analysis, the focus is on the 'Locale' variable. The analysis involves visualizing the relationship between unemployment rate and various socioeconomic variables, separated

by Locale. A scatter plot is created for these variables, with the variable on the x-axis and the Unemployment rate on the y-axis.

Log(Median income): The scatter plots show a downward trend, indicating a negative relationship between unemployment rate and log(Median income). This suggests that higher median income is associated with lower unemployment rates. The trend appears to be consistent across different locales.

Poverty rate: The scatter plots show an upward trend, indicating a positive relationship between unemployment rate and poverty rate. This suggests that higher poverty rates are associated with higher unemployment rates. The trend appears to be consistent across different locales.

SNAP rate: The scatter plots show an upward trend, indicating a positive relationship between unemployment rate and SNAP rate. This suggests that higher SNAP rates are associated with higher unemployment rates. The trend appears to be consistent across different locales.

Socioeconomic index: The scatter plots show a downward trend, indicating a negative relationship between unemployment rate and socioeconomic index. This suggests that higher socioeconomic index scores are associated with lower unemployment rates. The trend appears to be consistent across different locales.

Rural Remote region has the least variation for all the variables, and Rural Distant has the second-least variations (not vary too much as x increases). The slope for all the other plots are quite distinct (large variations between different x values).

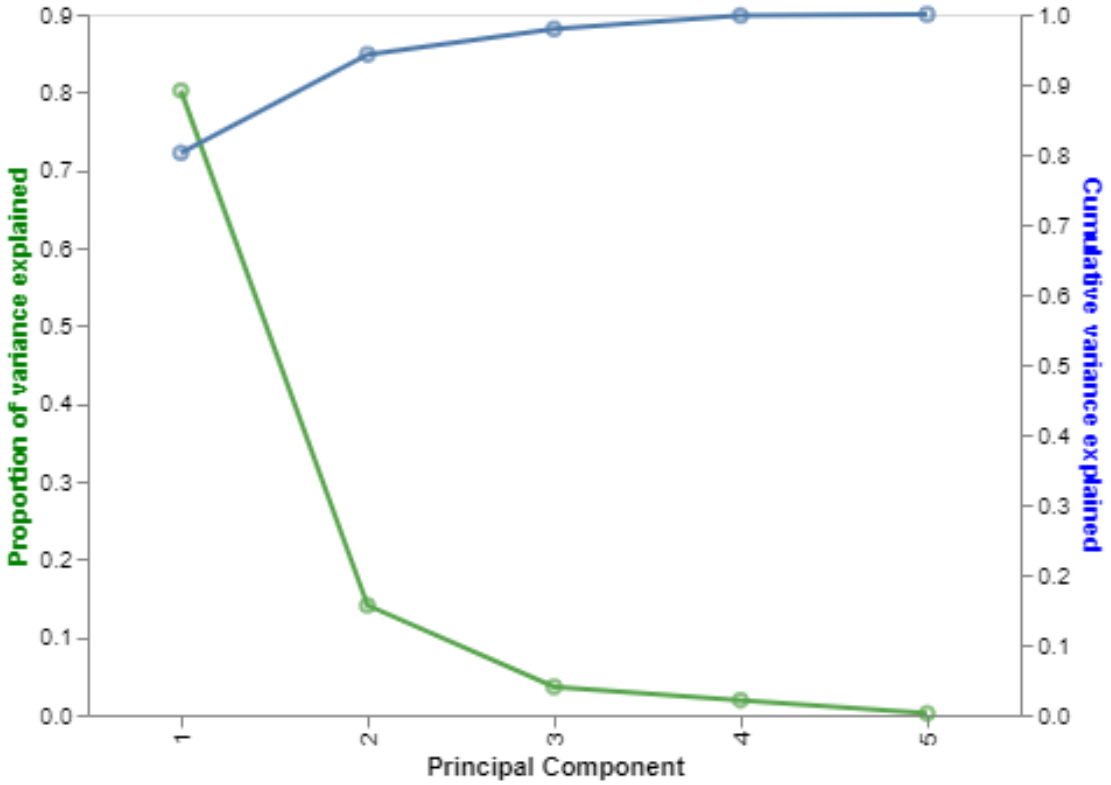
In addition to the scatter plots, kernel density plots are also created for each Locale. The plots provide a detailed view of how the relationship between Unemployment rate and socioeconomic variables might differ across different Locales. They help to visually identify trends and patterns in the data, which can then be further investigated using statistical methods.

[24]:
```
# PCA

from sklearn.decomposition import PCA
import statsmodels.api as sm
# helper variable pcdata_raw; set District ID as indices
pcdata_raw = data.dropna().drop(columns = ['Locale']).set_index(['District ID'])

# center and scale the relative abundances
pcdata = sm.add_constant(pcdata_raw)
pcdata = (pcdata - pcdata.mean()) / pcdata.std()

pcdata = pcdata.drop(columns = 'const')
pcdata

# compute pcs
pca = PCA(n_components = 5)
pca.fit(pcdata)

#retrieve variance info
```

```
# store proportion of variance explained as a dataframe
pcvars = pca.explained_variance_ratio_
pcvars = pd.DataFrame(pcvars)
pcvars = pcvars.rename(columns = {0:'Proportion of variance explained'})

# add component number as a new column
pcvars['Component'] = [1,2,3,4,5]

# add cumulative variance explained as a new column
pcvars['Cumulative variance explained'] = pcvars['Proportion of variance␣
 ↪explained'].cumsum()


## plot variance explained

# encode component axis only as base layer
base = alt.Chart(pcvars).encode(
    alt.X('Component:O', title='Principal Component')
)

# make a base layer for the proportion of variance explained
prop_var_base = alt.Chart(pcvars).encode(
    alt.X('Component:O', title='Principal Component'),
    alt.Y('Proportion of variance explained:Q', title='Proportion of Variance␣
 ↪Explained')
).properties(
width=400,
).encode(
    y=alt.Y('Proportion of variance explained', axis=alt.
 ↪Axis(titleColor='green'))
)

# make a base layer for the cumulative variance explained
cum_var_base = alt.Chart(pcvars).encode(
    alt.X('Component:O', title='Principal Component'),
    alt.Y('Cumulative variance explained:Q', title='Proportion of Variance␣
 ↪Explained')
).properties(
width=400,
).encode(
    y=alt.Y('Cumulative variance explained', axis=alt.Axis(titleColor='blue'))
)

# add points and lines to each base layer
prop_var = prop_var_base.mark_line(stroke = '#57A44C') + prop_var_base.
 ↪mark_point(color = '#57A44C')
cum_var = cum_var_base.mark_line() + cum_var_base.mark_point()
```

```
fig2 = alt.layer(prop_var, cum_var).resolve_scale(y = 'independent')

fig2
```

[24]:



The purpose of making this PCA plot is to quickly determine the fewest number of principal components that capture a considerable portion of variation and covariation. 'Considerable' here is a bit subjective. As seen in this graph, we can see that the first component explains 80% of variation, which is a considerable portion of variation and covariation. So we may conclude that one principal component is enough.

3.1 Summary:

Based on the SEDA dataset in 2018, the relationship between the Unemployment rate and factors such as Median income, Locale, Poverty rate, SNAP rate, and Socioeconomic index was investigated. We have the following findings:

*Median Income: Based on the heat map and the MLR model, there is a strong negative relationship between Unemployment and log(Median Income). This indicates that as the log of Median Income increases, the Unemployment rate decreases. This suggests that areas with higher median incomes tend to have lower unemployment rates.

*Locale: Based on the heat map and KDE plot, the relationship between unemployment rate

and socioeconomic variables differs across different locales. For example, urban areas might show different trends compared to rural areas.

*Poverty Rate: Based on the heat map and the MLR model, there is a strong positive relationship between Unemployment and Poverty rate. This indicates that as the Poverty rate increases, the Unemployment rate also increases. This suggests that areas with higher poverty rates are more likely to have higher unemployment rates.

*SNAP Rate: Based on the heat map and the MLR model, there is a strong positive relationship between Unemployment and SNAP rate. This indicates that as the SNAP rate increases, the Unemployment rate also increases. This suggests that areas with higher SNAP participation rates are more likely to have higher unemployment rates.

*Socioeconomic Index: Based on the heat map and the MLR model, there is a strong negative relationship between Unemployment and Socioeconomic index. This indicates that as the Socioeconomic index increases, the Unemployment rate decreases. This suggests that areas with higher socioeconomic index scores are more likely to have lower unemployment rates.