Legislation Effects on Cigarette/E-Cigarette Usage

Ju Kim, David Hong, Fuxin Tian, and Jialin Shi

Author contributions

Ju Kim contributed to tidying the dataset in order to display graphs in a presentable manner and clearly show each variable's effect on one another.

David Hong contributed to writing the code for correlation map and linear regression.

Fuxin Tian contributed to contructing density plots, Scatter plot for data compare, linear regression and plotting, residual plot, choropleth map, methods, and results.

Jialin Shi contributed to combining other group members works together, making modifications to the works, and writing the abstract, introduction, aims, conclusion.

Abstract

As the implementation of the new policy that forbid indoor smoking, we are interested in whether the law would have any effects on smoking frequencies of smokers within the United States; we believe the frequency of "everyday" is the most representative, and as a result, we utilize it to conduct the analysis. In this project, we intend to investigate the relationship between the legistration percentage of different states and usages of cigarette and e-cigarette. Further, we aim to examine whether the policy has different effects on different gender and different type of cigarette. Last but least, we would like to compare the impact of the policy on each states. After constructing various plots, we conclude that there a negative linear relationship between legistration percentage and smoking frequency, which means the with the increasing of legistration percentage there's a decaresing in smoking frequency; in addition, the results show that gender was not a significant predictor of smoking prevalence and each state has difference legistration percentage and corresponding smoking frequency.

Background

Datasets in this project are collected from the CDC and captured through the BRFSS - the Behavioral Risk Factor Surveillance System. Adults around the US are surveyed through the telephone and asked about their smoking habits/frequencies, and observations are grouped by gender in each of the states. Years 2013-2016 are captured in this dataset due to limited availability of data on e-cigarette usage and policies revolving smoke-free areas.

Aims

This project focuses on solving the questions "how usages of cigarette and e-cigarette related with legistration percentage and gender" and "how do smoking rates vary by state and legislation percentage". To tackle with the first question, we created a bar chart showing smoking prevalence by state, two density plots to visualize the distribution of the 'Data Value' and 'legislation percentage' variables, and two scatter plots of 'Legislation Percentage vs. Data Value (Every Day)'. Then, we performed a sensitivity analysis to determine the impact of the Gender variable on the model's performance, and finally, we constructed a scatter plot with a regression line to visualize the relationship between the legislation percentage and smoking prevalence, colored by Topic Description. For the second question, we construct choropleth maps to visualize the smoking prevalence and legislation percentage by state in the United States. In particular, we filtered the dataset to select the year we were interested in, and merged this filtered data with an ANSI table that contained the state abbreviations and names; then, we used the United States map data from the Altair library and created a base map with light gray fill and black borders, and created two choropleth maps, one for legislation percentage and one for smoking prevalence.

In [1]:
```
pwd
```

Out[1]:
```
'C:\\Users\\StevenTian\\Desktop\\final\\CP2\\CP2'
```

In [2]:
```
pip install vega_datasets
```

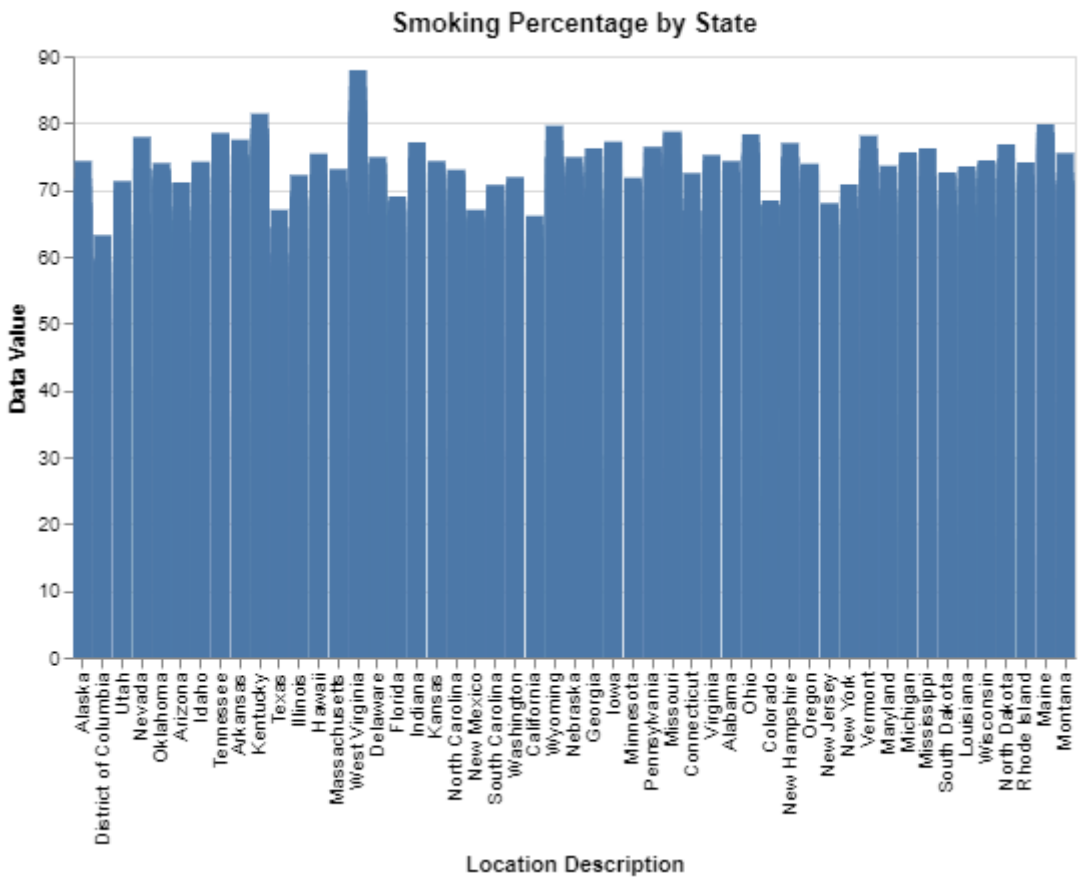
```
Requirement already satisfied: vega_datasets in c:\users\steventian\anaconda3\lib\site-packages (0.9.0)
Requirement already satisfied: pandas in c:\users\steventian\anaconda3\lib\site-packages (from vega_datasets) (1.4.4)
Requirement already satisfied: numpy>=1.18.5 in c:\users\steventian\anaconda3\lib\site-packages (from pandas->vega_datasets) (1.
21.5)
Requirement already satisfied: pytz>=2020.1 in c:\users\steventian\anaconda3\lib\site-packages (from pandas->vega_datasets) (202
2.1)
Requirement already satisfied: python-dateutil>=2.8.1 in c:\users\steventian\anaconda3\lib\site-packages (from pandas->vega_data
sets) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\steventian\anaconda3\lib\site-packages (from python-dateutil>=2.8.1->pandas-
>vega_datasets) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [3]:
```
import numpy as np
import pandas as pd
import altair as alt
import sklearn.linear_model as lm
```

```
import warnings
from sklearn.preprocessing import add_dummy_feature
from vega_datasets import data
warnings.simplefilter(action='ignore', category=FutureWarning)
```

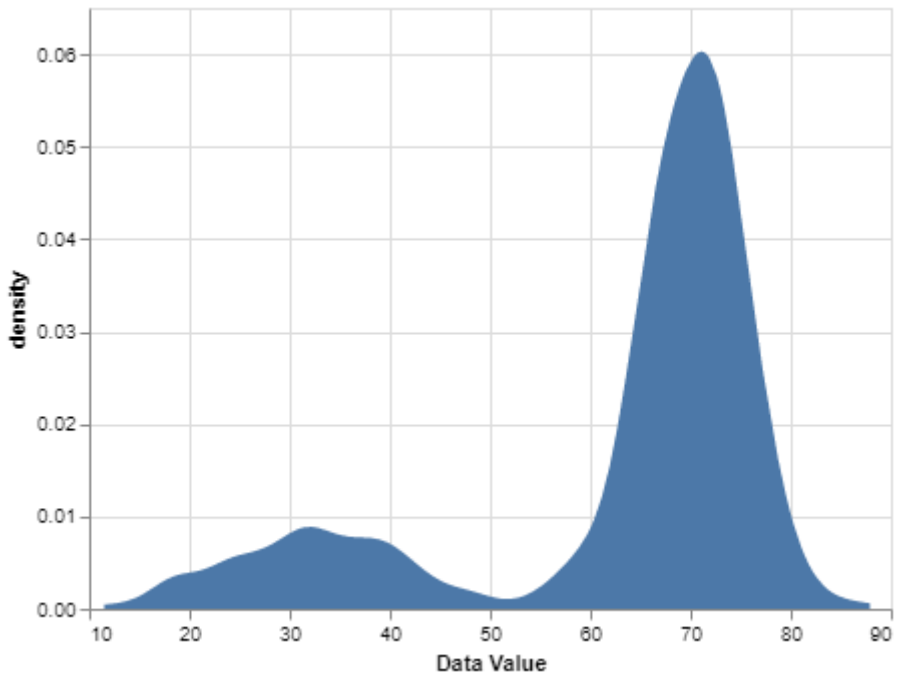
In [4]:
```
dataset=pd.read_csv('data/tidy-data copy.csv')
dataset=dataset.drop(columns=['Unnamed: 0'])
dataset_every_day = dataset.loc[dataset['Response'] == 'Every Day']
dataset_every_day=dataset_every_day[dataset_every_day['Gender']!='Overall']
dataset_every_day=dataset_every_day[dataset_every_day['Data Value']!='*']
legislation_by_state = dataset_every_day.groupby('Location Description')['legislation percentage'].mean().sort_values().reset_in

alt.Chart(dataset_every_day).mark_bar().encode(
    x=alt.X('Location Description:N', sort='-y'),
    y=alt.Y('Data Value:Q'),
    tooltip=['Location Description', 'Data Value']
).properties(
    width=500,
    height=300,
    title='Smoking Percentage by State'
)
```

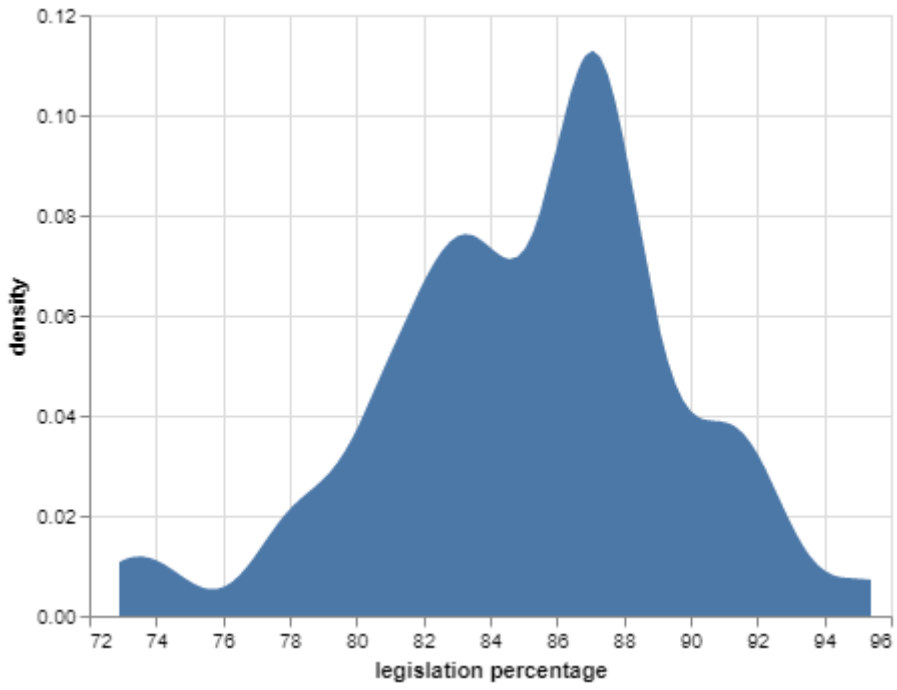
Out[4]:



In [5]:
```
# Density plots
dataset_every_day["Data Value"] = pd.to_numeric(dataset_every_day["Data Value"])
# Create a density plot of Data Value
data_value_plot = alt.Chart(dataset_every_day).transform_density(
    'Data Value',
    as_=['Data Value', 'density'],
).mark_area().encode(
    x='Data Value:Q',
    y='density:Q'
)
# Create a density plot of legislation percentage
legislation_plot = alt.Chart(dataset_every_day).transform_density(
    'legislation percentage',
    as_=['legislation percentage', 'density'],
).mark_area().encode(
    x='legislation percentage:Q',
    y='density:Q'
)

# Combine the three plots into a single chart
alt.vconcat(data_value_plot,  legislation_plot)
```

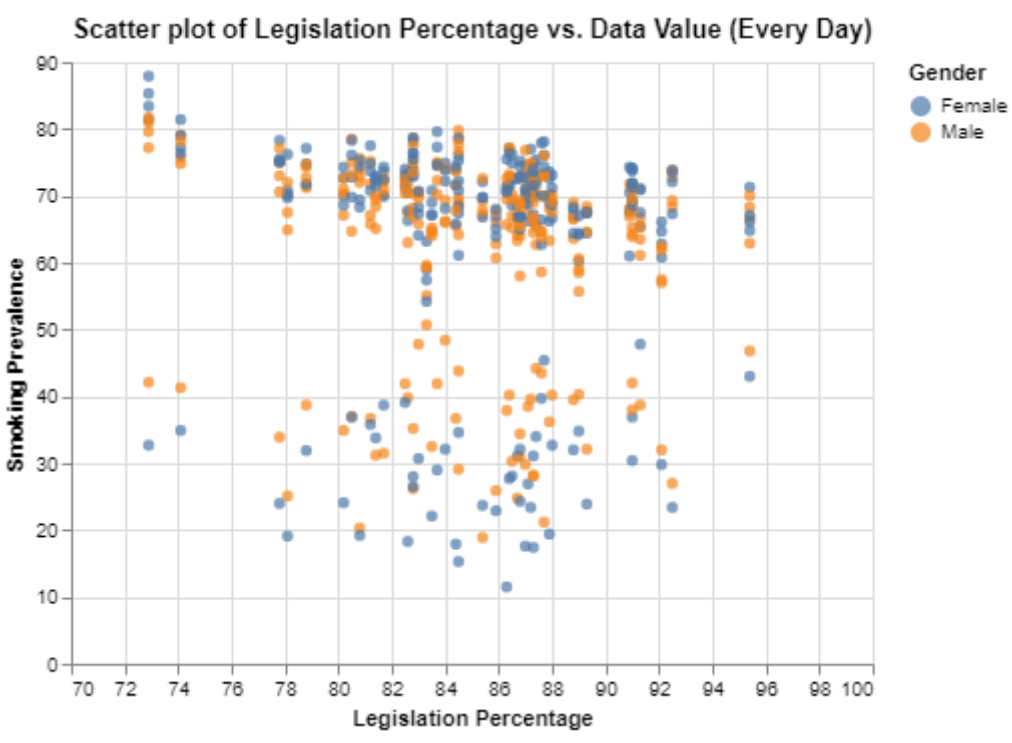
Out[5]:





In [6]:
```
# Scatter plot of Legislation Percentage vs. Data Value (Every Day)
alt.Chart(dataset_every_day).mark_circle().encode(
    alt.X('legislation percentage:Q',title='Legislation Percentage',
          scale=alt.Scale(domain=[70, 100])),
    alt.Y('Data Value:Q', title='Smoking Prevalence'),
    alt.Color('Gender:N')
).properties(
    width=400,
    height=300,
    title='Scatter plot of Legislation Percentage vs. Data Value (Every Day)'
)
```

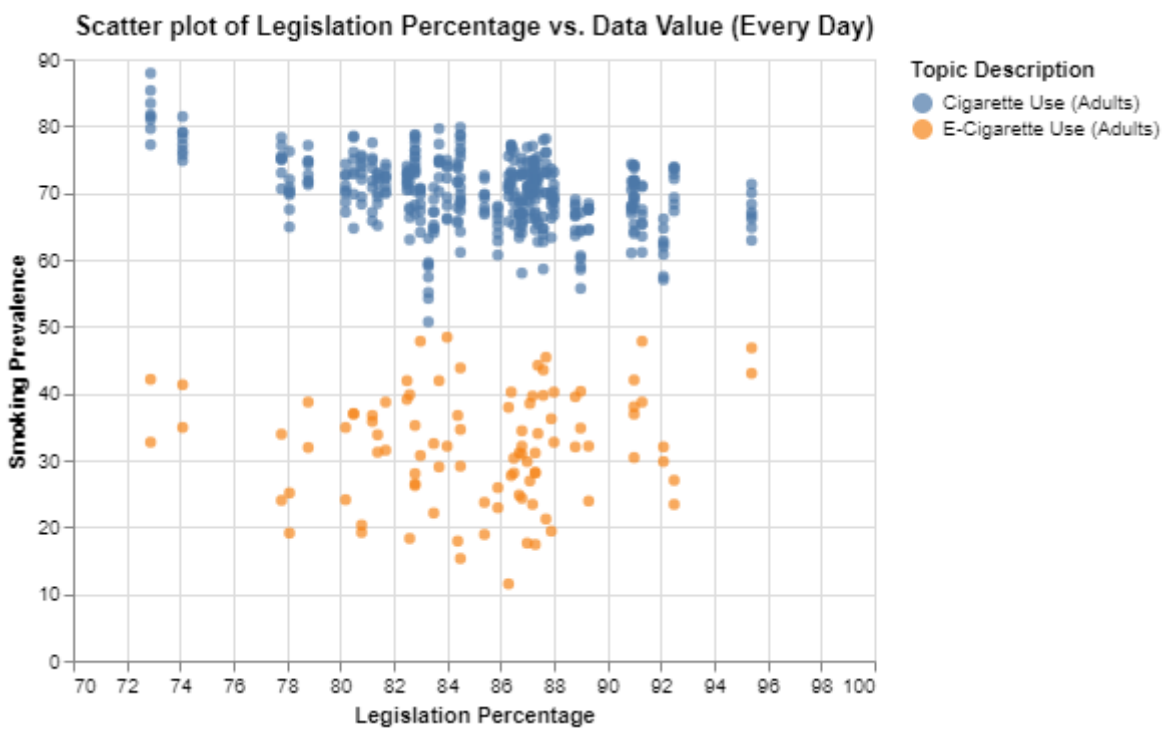
Out[6]:



In [7]:
```
#Scatter plot of Legislation Percentage vs. Data Value (Every Day)
alt.Chart(dataset_every_day).mark_circle().encode(
    alt.X('legislation percentage:Q',title='Legislation Percentage',
          scale=alt.Scale(domain=[70, 100])),
    alt.Y('Data Value:Q', title='Smoking Prevalence'),
    alt.Color('Topic Description:N'),
    tooltip=['Location Description', 'legislation percentage',
             'Data Value', 'Topic Description']
).properties(
    width=400,
    height=300,
```

```
    title='Scatter plot of Legislation Percentage vs. Data Value (Every Day)'
)
```

Out[7]:



In [8]:

```
# Fit the linear regression
x_dummy=dataset_every_day.drop(columns=['Year','Data Value','Low Confidence Limit','High Confidence Limit','Sample Size','legisla
                                        ,'Measure Description','Location Description','Response'])
x_df=pd.get_dummies(x_dummy,drop_first=True)
x_df['legislation']=dataset_every_day['legislation percentage']
x_mx=add_dummy_feature(x_df, value = 1)
y=dataset_every_day['Data Value']
y=pd.to_numeric(y) #y is stored as string in the dataframe
y=y.to_numpy()
mlr = lm.LinearRegression(fit_intercept = False)
mlr.fit(x_mx,y)
# store dimensions
n=x_mx.shape[0]
p=x_mx.shape[1]
# compute x'x
xtx = x_mx.transpose().dot(x_mx)
          # compute x'x inverse
xtx_inv = np.linalg.inv(xtx)
          # compute residuals
resid = y-mlr.predict(x_mx)
          # compute error variance estimate
sigmasqhat = ((n-1)/(n-p)) * resid.var()
          # compute variance-covariance matrix
v_hat = xtx_inv * sigmasqhat
          # compute standard errors
coef_se = np.sqrt(v_hat.diagonal())
coef_se = np.append(coef_se,float('nan'))
          # coefficient labels
newarray=x_df.columns.insert(0,'intercept')
coef_labels=list(newarray)
coef_labels.append('error_variance')
# estimates
coef_estimates= np.append(mlr.coef_,sigmasqhat)
          # summary table
coef_table = pd.DataFrame(data={'coef_estimates':coef_estimates,'coef_se':coef_se},index=coef_labels)
coef_table
```

Out[8]:

	coef_estimates	coef_se
intercept	102.758916	4.731984
Topic Description_E-Cigarette Use (Adults)	-38.190083	0.627385
Gender_Male	-0.379051	0.495846
legislation	-0.379752	0.055362
error_variance	31.101691	NaN

In [9]:

```
# R-squared 'by hand'
R_2 =(y.var() - resid.var())/y.var()

R_2
```

Out[9]:

```
0.881662333662007
```

The metric is simply the difference between the raw variation in the response and residual variation, as a proportion of the variation in the response. If the model fits well, the residual variation will be small, in which case this proportion will be closer to 1. Here the R square is 0.88166, which means the model interprets 88.17% of the variance in the response variable using the predictor.

From the coefficient estimation and coefficient se we can find that the gender is not significant (the estimate is not over 2 times of the standard error.) So we can try to fit a model without the gender and check how the R square change:

In [10]:
```
# Fit the linear regression without the gender:
x_dummy=dataset_every_day.drop(columns=['Year','Data Value','Low Confidence Limit','High Confidence Limit','Sample Size','legisla
                                        ,'Measure Description','Location Description','Response','Gender'])
x_df=pd.get_dummies(x_dummy,drop_first=True)
x_df['legislation']=dataset_every_day['legislation percentage']
x_mx=add_dummy_feature(x_df, value = 1)
y=dataset_every_day['Data Value']
y=pd.to_numeric(y) #y is stored as string in the dataframe
y=y.to_numpy()
mlr = lm.LinearRegression(fit_intercept = False)
mlr.fit(x_mx,y)
# store dimensions
n=x_mx.shape[0]
p=x_mx.shape[1]
# compute x'x
xtx = x_mx.transpose().dot(x_mx)
          # compute x'x inverse
xtx_inv = np.linalg.inv(xtx)
          # compute residuals
resid = y-mlr.predict(x_mx)
          # compute error variance estimate
sigmasqhat = ((n-1)/(n-p)) * resid.var()
          # compute variance-covariance matrix
v_hat = xtx_inv * sigmasqhat
          # compute standard errors
coef_se = np.sqrt(v_hat.diagonal())
coef_se = np.append(coef_se,float('nan'))
          # coefficient labels
newarray=x_df.columns.insert(0,'intercept')
coef_labels=list(newarray)
coef_labels.append('error_variance')
# estimates
coef_estimates= np.append(mlr.coef_,sigmasqhat)
          # summary table
coef_table = pd.DataFrame(data={'coef_estimates':coef_estimates,'coef_se':coef_se},index=coef_labels)
coef_table
```

Out[10]:

	coef_estimates	coef_se
intercept	102.569390	4.723527
Topic Description_E-Cigarette Use (Adults)	-38.190083	0.627125
legislation	-0.379752	0.055339
error_variance	31.075922	NaN

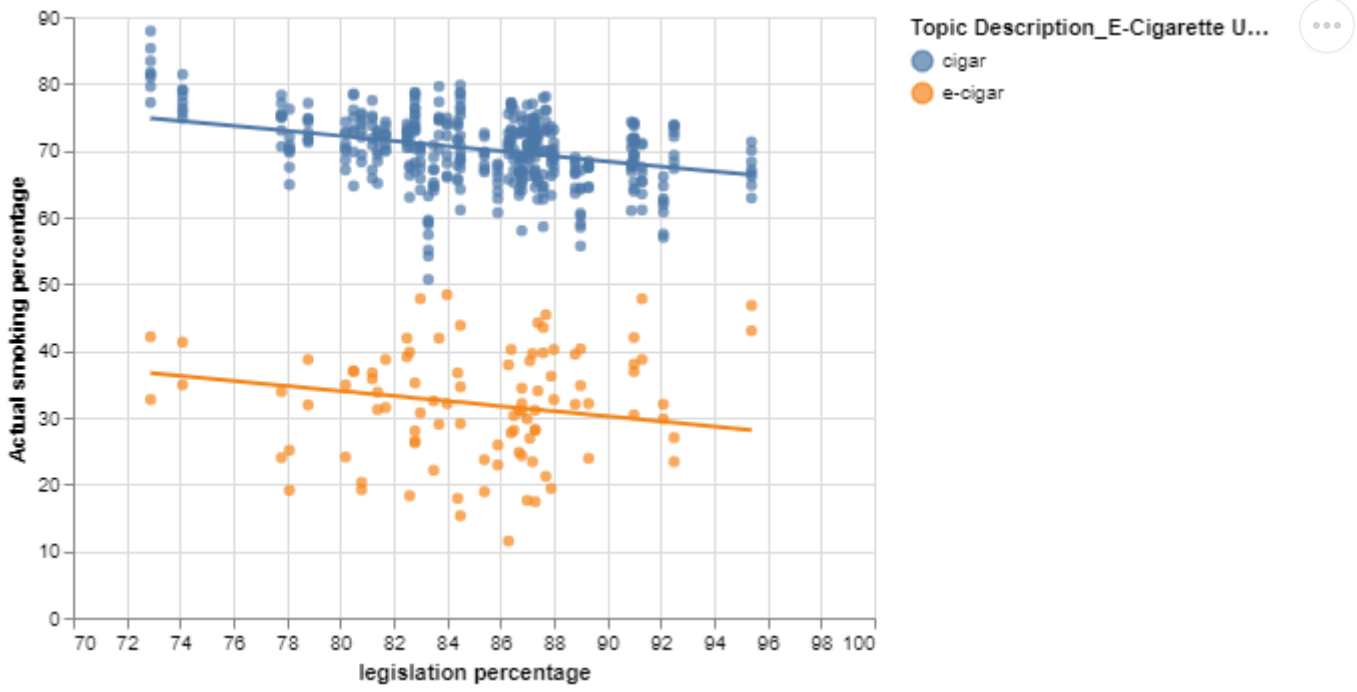
In [11]:
```
# R-squared 'by hand'
R_2 =(y.var() - resid.var())/y.var()

R_2
```

Out[11]:
```
0.8815248462819016
```

The R square only decrease from 88.17% to 88.15%, which suggest it's not a very singificant change. So including gender may not be so informatic, and we can fit the model without gender. This result is similar to the scatter plot colored with gender where the gender does not seem to be separate.

In [12]:
```
# construct Regression scatter plot + line
visuals=x_df
for i in range(0,len(visuals)):
    if visuals.iloc[i,0] ==1:
        visuals.iloc[i,0]='e-cigar'
    elif visuals.iloc[i,0] ==0:
        visuals.iloc[i,0]='cigar'
for i in range(0,len(visuals)):
    if visuals.iloc[i,1] ==1:
        visuals.iloc[i,1]='Male'
    elif visuals.iloc[i,1] ==0:
        visuals.iloc[i,1]='Female'
visuals=visuals.reset_index()
visuals=visuals.drop(columns=['index'])
visuals['predicted values']=mlr. predict(x_mx)
visuals['actual values']=y
visuals['residuals'] = visuals['actual values'] - visuals['predicted values']
scatter_chart = alt.Chart(visuals).mark_circle().encode(
    x=alt.X('legislation', axis=alt.Axis(title='legislation percentage'),
           scale=alt.Scale(domain=[70, 100])),
    y=alt.Y('actual values', axis=alt.Axis(title='Actual smoking percentage')),
    color=alt.Color('Topic Description_E-Cigarette Use (Adults)')
)
# construct Regression scatter plot + line
mlr_lines = scatter_chart.mark_line(color = 'red').encode(y = 'predicted values')
scatter_chart + mlr_lines
```

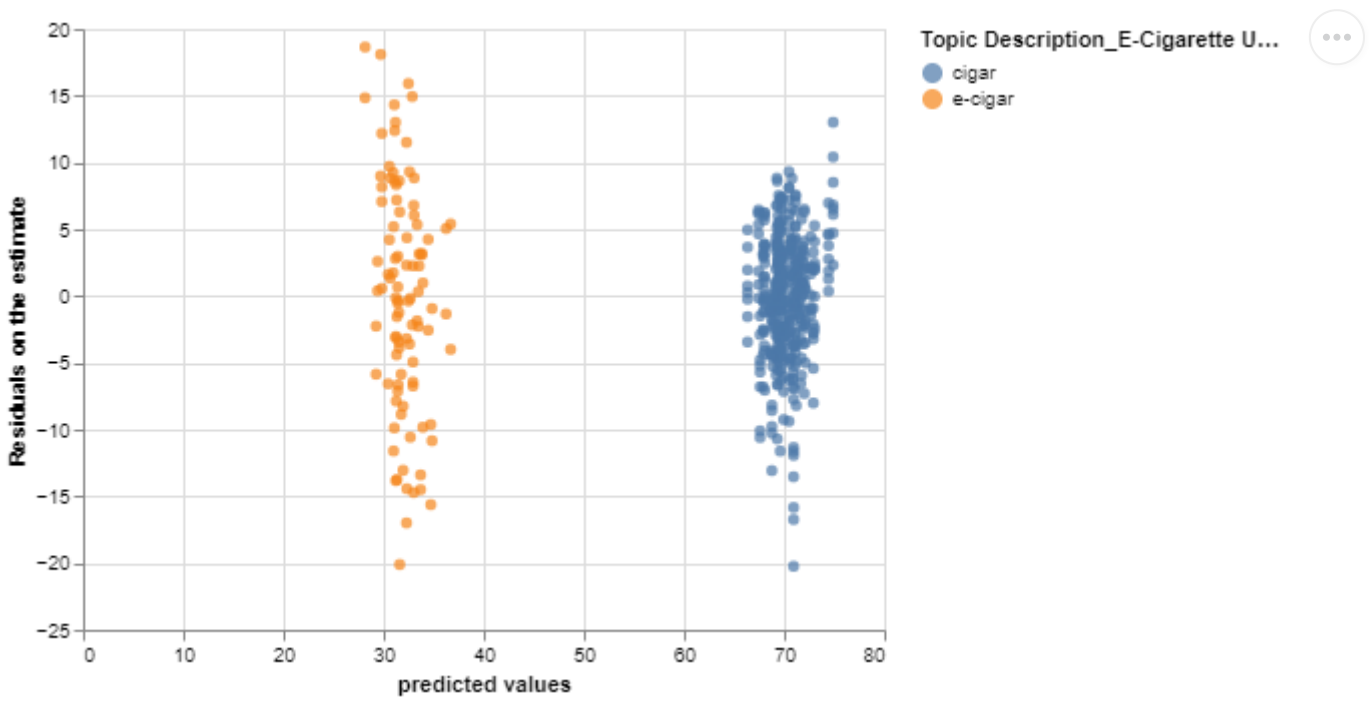
Out[12]:



In [13]:

```
# The residual vs fitted plot
lines1 = alt.Chart(visuals).mark_circle().encode(
x=alt. X('predicted values',axis= alt. Axis(title= 'predicted values')),
y=alt.Y('residuals',axis=alt.Axis(title='Residuals on the estimate')),
color= alt. Color('Topic Description_E-Cigarette Use (Adults)')
)
lines1
```

Out[13]:



Map analysis part code

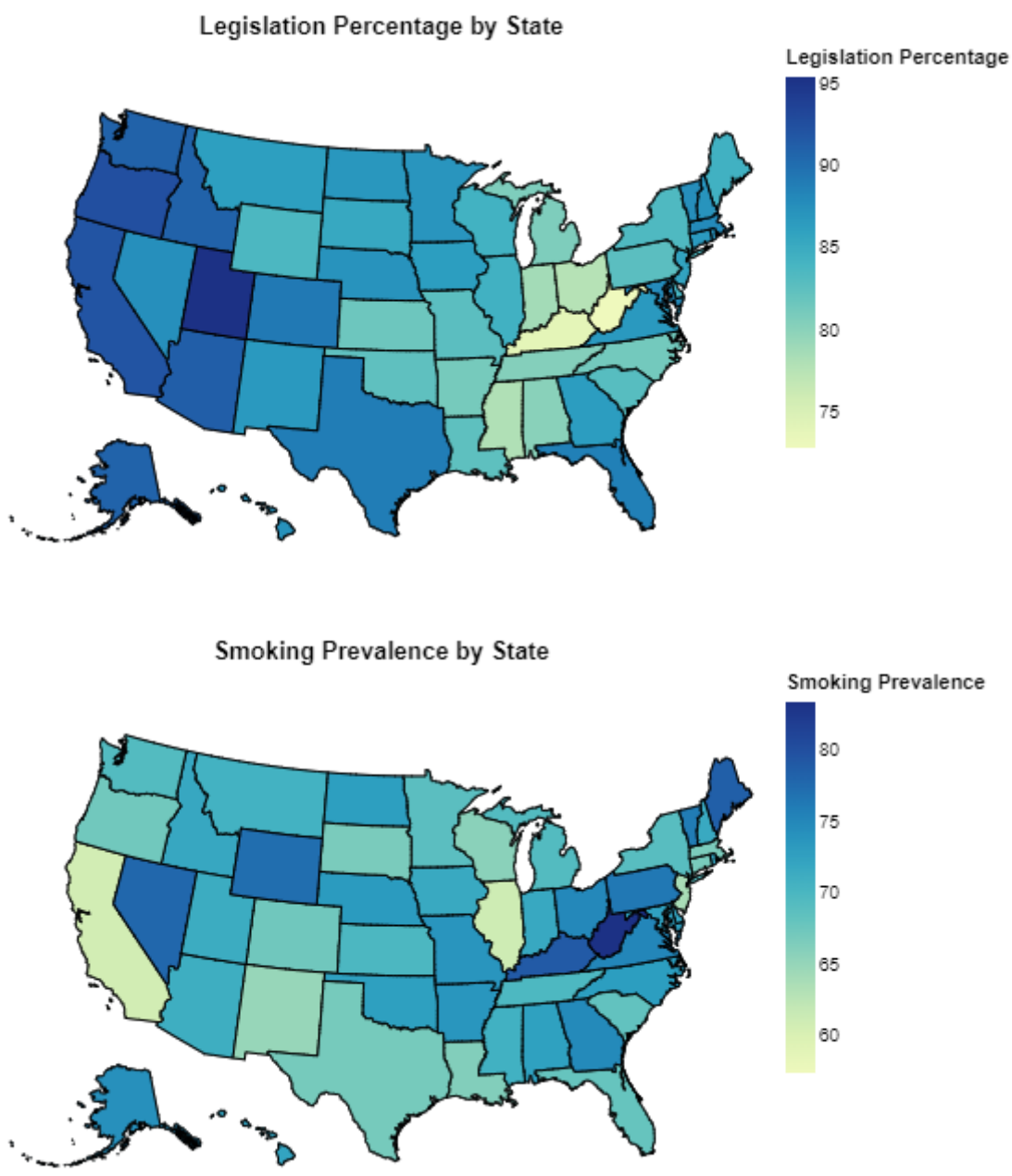
We can create a choropleth map of the United States, where each state is shaded according to the smoking prevalence, with the shade of color varying based on the level of legislation percentage in that state. This could help visualize any potential patterns or trends in smoking rates and legislation percentage across different states. Additionally, a regression analysis could be conducted to see if there is a significant relationship between smoking rates and legislation percentage, while controlling for potential confounding variables such as age and gender.

In [14]:

```
# Filter the dataset to the year you're interested in
year = 2015  # Replace with the desired year
filtered_data = dataset_every_day[dataset_every_day['Year'] == year]
ansi = pd.read_csv('https://www2.census.gov/geo/docs/reference/state.txt', sep='|')
ansi.columns = ['id', 'abbr', 'state', 'statens']
ansi = ansi[['id', 'abbr', 'state']]

filtered_data1 = pd.merge(filtered_data, ansi, how='left',
                          left_on='Location Description', right_on='state')

# Load the United States map data
states = alt.topo_feature(data.us_10m.url, feature='states')

# Create the choropleth map of legislation
base = alt.Chart(states).mark_geoshape(fill='lightgray', stroke='black', strokeWidth=0.5)

chart = alt.Chart(states).mark_geoshape(stroke='black').encode(
    alt.Color('legislation percentage:Q', legend=alt.Legend(title="Legislation Percentage")),
).transform_lookup(
    lookup='id',
    from_=alt.LookupData(filtered_data1, 'id', ['legislation percentage'])
).properties(
    width=400,
    height=300,
    title='Legislation Percentage by State'
).project('albersUsa')
chart1 = base + chart
# Create the choropleth map of legislation
```

```
base = alt.Chart(states).mark_geoshape(fill='lightgray', stroke='black', strokeWidth=0.5)

chart = alt.Chart(states).mark_geoshape(stroke='black').encode(
    alt.Color('Data Value:Q', legend=alt.Legend(title="Smoking Prevalence")),
).transform_lookup(
    lookup='id',
    from_=alt.LookupData(filtered_data1, 'id', ['Data Value'])
).properties(
    width=400,
    height=300,
    title='Smoking Prevalence by State'
).project('albersUsa')
chart2 = base + chart
alt.vconcat(chart1, chart2).resolve_scale(
    color='independent'
).configure_view(
    stroke=None
)
```

Out[14]:



In [15]:

```
x_dummy=dataset_every_day.drop(columns=['Year','Data Value','Low Confidence Limit','High Confidence Limit','Sample Size','legisla
                                        ,'Measure Description','Location Description','Response'])
x_df=pd.get_dummies(x_dummy,drop_first=True)
x_df['legislation']=dataset_every_day['legislation percentage']
x_mx=add_dummy_feature(x_df, value = 1)
y=dataset_every_day['Data Value']
y=pd.to_numeric(y) #y is stored as string in the dataframe
y=y.to_numpy()
mlr = lm.LinearRegression(fit_intercept = False)
mlr.fit(x_mx,y)
# store dimensions
n=x_mx.shape[0]
p=x_mx.shape[1]
# compute x'x
xtx = x_mx.transpose().dot(x_mx)
          # compute x'x inverse
xtx_inv = np.linalg.inv(xtx)
          # compute residuals
resid = y-mlr.predict(x_mx)
          # compute error variance estimate
sigmasqhat = ((n-1)/(n-p)) * resid.var()
          # compute variance-covariance matrix
v_hat = xtx_inv * sigmasqhat
          # compute standard errors
coef_se = np.sqrt(v_hat.diagonal())
coef_se = np.append(coef_se,float('nan'))
          # coefficient labels
newarray=x_df.columns.insert(0,'intercept')
coef_labels=list(newarray)
coef_labels.append('error_variance')
```

```
# estimates
coef_estimates= np.append(mlr.coef_,sigmasqhat)
          # summary table
coef_table = pd.DataFrame(data={'coef_estimates':coef_estimates,'coef_se':coef_se},index=coef_labels)
coef_table
```

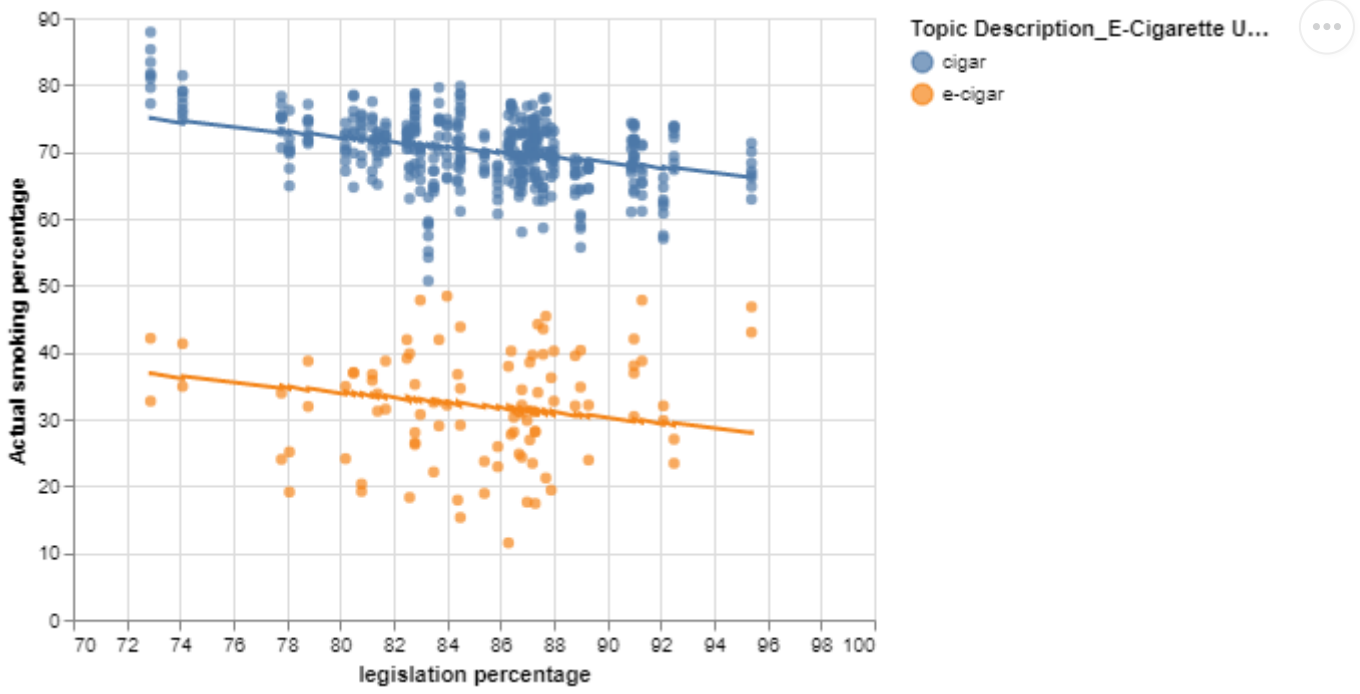
Out[15]:

	coef_estimates	coef_se
intercept	102.758916	4.731984
Topic Description_E-Cigarette Use (Adults)	-38.190083	0.627385
Gender_Male	-0.379051	0.495846
legislation	-0.379752	0.055362
error_variance	31.101691	NaN

In [16]:

```
visuals=x_df
for i in range(0,len(visuals)):
    if visuals.iloc[i,0] ==1:
        visuals.iloc[i,0]='e-cigar'
    elif visuals.iloc[i,0] ==0:
        visuals.iloc[i,0]='cigar'
for i in range(0,len(visuals)):
    if visuals.iloc[i,1] ==1:
        visuals.iloc[i,1]='Male'
    elif visuals.iloc[i,1] ==0:
        visuals.iloc[i,1]='Female'
visuals=visuals.reset_index()
visuals=visuals.drop(columns=['index'])
visuals['predicted values']=mlr. predict(x_mx)
visuals['actual values']=y
visuals['residuals'] = visuals['actual values'] - visuals['predicted values']
scatter_chart = alt.Chart(visuals).mark_circle().encode(
    x=alt.X('legislation', axis=alt.Axis(title='legislation percentage'),
           scale=alt.Scale(domain=[70, 100])),
    y=alt.Y('actual values', axis=alt.Axis(title='Actual smoking percentage')),
    color=alt.Color('Topic Description_E-Cigarette Use (Adults)')
)
# construct Regression scatter plot + line
mlr_lines = scatter_chart.mark_line(color = 'red').encode(y = 'predicted values')
scatter_chart + mlr_lines
```

Out[16]:



In [17]:

```
for i in range(0,506):
    if visuals.iloc[i,0]=='cigar':
        visuals.iloc[i,0]=1
    else:
        visuals.iloc[i,0]=0
for i in range(0,506):
    if visuals.iloc[i,1]=='Male':
        visuals.iloc[i,1]=1
    else:
        visuals.iloc[i,1]=0
visuals=visuals.drop(columns=['actual values','residuals'])
visuals.head()
```

Out[17]:

	Topic Description_E-Cigarette Use (Adults)	Gender_Male	legislation	predicted values
0	1	1	80.2	71.923742
1	1	0	80.2	72.302794
2	1	1	80.2	71.923742
3	1	0	80.2	72.302794
4	1	1	80.2	71.923742

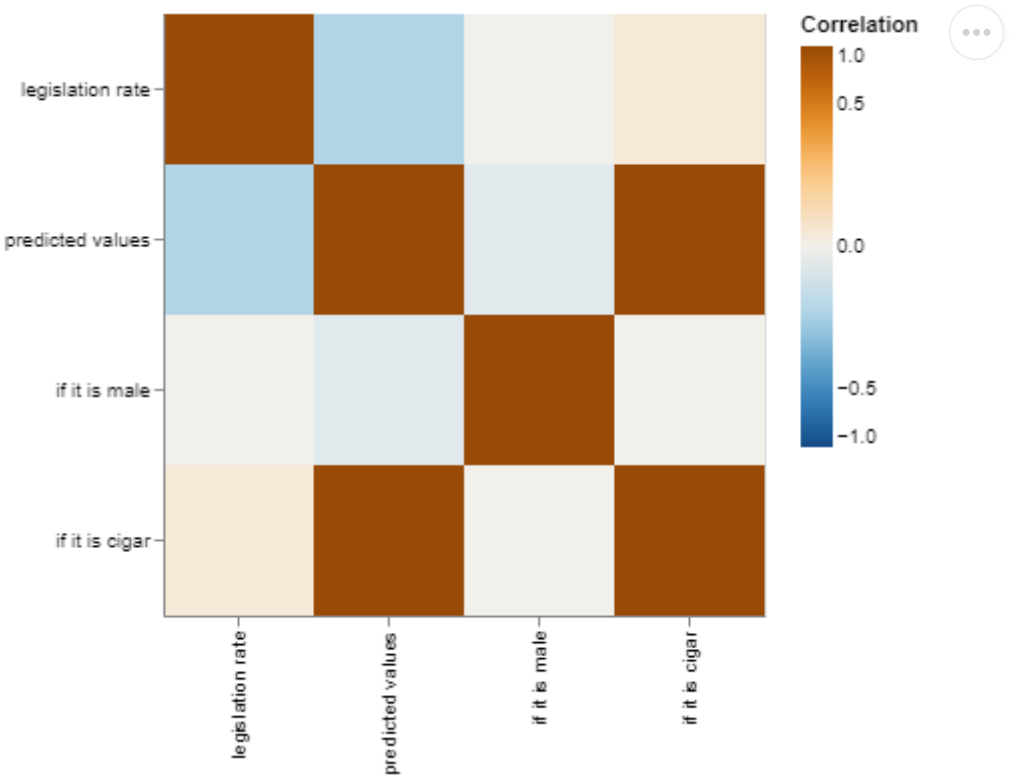
In [18]:
```
a=visuals.to_numpy()
a=a.astype(float)
corrmatrix=np.corrcoef(a.T)
corrtable=pd.DataFrame(corrmatrix)
corrtable=corrtable.rename(columns={0:"if it is cigar", 1:"if it is male",2:"legislation rate",3:"predicted values" })
corrtable=corrtable.rename(index={0:"if it is cigar", 1:"if it is male",2:"legislation rate",3:"predicted values" })
corrtable
```

Out[18]:

	if it is cigar	if it is male	legislation rate	predicted values
if it is cigar	1.000000	0.000000	0.006808	0.993624
if it is male	0.000000	1.000000	0.000000	-0.012488
legislation rate	0.006808	0.000000	1.000000	-0.105284
predicted values	0.993624	-0.012488	-0.105284	1.000000

In [19]:
```
corr_mx_long = corrtable.reset_index().rename(
columns = {'index': 'row'}
).melt(
id_vars = 'row',
var_name = 'col',
value_name = 'Correlation')
alt.Chart(corr_mx_long).mark_rect().encode(
)
# construct plot
alt.Chart(corr_mx_long).mark_rect().encode(
    x = alt.X('col', title = '', sort = {'field': 'Correlation', 'order': 'ascending'}),
    y = alt.Y('row', title = '', sort = {'field': 'Correlation', 'order': 'ascending'}),
    color = alt.Color('Correlation',
                      scale = alt.Scale(scheme = 'blueorange', # diverging gradient
                                        domain = (-1, 1), # ensure white = 0
                                        type = 'sqrt'), # adjust gradient scale
                      legend = alt.Legend(tickCount = 5)) # add ticks to colorbar at 0.5 for reference
).properties(width = 300, height = 300)
```

Out[19]:



Materials and methods

The goal of this section is to describe your dataset(s) and sketch out your analysis.

Datasets

Name	Variable description	Type	Units of measurement
Year	the year that the data was recored	numeric	year
Location Description	the state's name that the data was recorded	characters	state-wise
Topic Description	it describes either the data is about cigar or e-cigar	characters	cigar/e-cigar
Measure Description	it describes the literal meaning of the data value	characters	Smoking frequency
Response	it describes how frequent does the person smoke	characters	Either everyday or someday
Data Value	it describes how much percent of population is consistent with the data value	numeric	percentage
Low confidence limit\High confidence limit	it describes the confidence interval of populatioin falling inside the interval	numeric	percentage
Sample Size	it describes the size of sample	numeric	people
Legislation Percentage	it describes how much percentage of places have indoor smoking policies	numeric	percentage

The example rows you showed in your planning report.

In [20]:
```
dataset_every_day.iloc[0:5,:]
```

Out[20]:

	Year	Location Description	Data Source	Topic Description	Measure Description	Response	Data Value	Low Confidence Limit	High Confidence Limit	Sample Size	Gender	legislation percentage
1	2016	Alabama	BRFSS	Cigarette Use (Adults)	Smoking Frequency	Every Day	67.1	62.0	72.2	574	Male	80.2
2	2016	Alabama	BRFSS	Cigarette Use (Adults)	Smoking Frequency	Every Day	70.0	65.6	74.4	686	Female	80.2
7	2015	Alabama	BRFSS	Cigarette Use (Adults)	Smoking Frequency	Every Day	70.3	65.4	75.2	609	Male	80.2
8	2015	Alabama	BRFSS	Cigarette Use (Adults)	Smoking Frequency	Every Day	72.7	68.6	76.8	784	Female	80.2
13	2014	Alabama	BRFSS	Cigarette Use (Adults)	Smoking Frequency	Every Day	71.0	66.0	76.0	613	Male	80.2

Methods

In this analysis, we first imported the dataset and dropped an unnecessary column, and then filtered it to only include data where the response was 'Every Day'. We also excluded the 'Overall' category for gender to focus on individual gender breakdowns, and removed any missing values from the 'Data Value' column. We then calculated the average legislation percentage by state using a groupby operation.

For our data visualization, we created a bar chart showing smoking prevalence by state. We also created two density plots to visualize the distribution of the 'Data Value' and 'legislation percentage' variables, respectively. Finally, we created two scatter plots of 'Legislation Percentage vs. Data Value (Every Day)', one showing the breakdown by gender and the other showing the breakdown by topic description.

The scatter plots reveal a slightly negative relationship between 'Data Value' and 'legislation percentage', suggesting that higher levels of indoor smoking policies are associated with lower smoking prevalence. Additionally, we can see that smoking prevalence tends to be higher for males compared to females, and that smoking prevalence for E-Cigarette Use tends to be higher compared to cigar use.

Our EDA analysis suggests that indoor smoking policies may play a role in reducing smoking prevalence, particularly among males and in the context of E-Cigarette Use.

The next step of our analysis was to perform linear regression analysis to model the relationship between the smoking prevalence (Data Value) and other variables in the dataset. We started by selecting the relevant columns from the dataset, dropping irrelevant columns, and then creating dummy variables for categorical variables like Topic Description and Gender. We also added a dummy feature for the intercept. After that, we fit the linear regression model using the scikit-learn library in Python. We then computed the x'x matrix, x'x inverse, residuals, error variance estimate, variance-covariance matrix, and standard errors. These values were used to generate a summary table of coefficient estimates and their standard errors.

We also calculated the R-squared value for the model, which is a measure of how well the model fits the data. It is simply the difference between the raw variation in the response and residual variation, as a proportion of the variation in the response. If the model fits well, the residual variation will be small, and this proportion will be closer to 1. Here, we found that the R-squared value was 0.88166, which means that the model interprets 88.17% of the variance in the response variable using the predictor.

Next, we performed a sensitivity analysis to determine the impact of the Gender variable on the model's performance. We fit a new linear regression model without the Gender variable and compared its R-squared value with that of the previous model. We found that the R-squared

value only decreased slightly from 88.17% to 88.15%, which suggests that the Gender variable may not be very informative. Therefore, we decided to fit the model without Gender.

Finally, we constructed a scatter plot with a regression line to visualize the relationship between the legislation percentage and smoking prevalence, colored by Topic Description. We also generated a residual vs. fitted plot to check for any patterns in the residuals that might indicate problems with the model. The plots showed a clear negative relationship between the legislation percentage and smoking prevalence, with the regression line fitting the data well. The residual vs. fitted plot did not show any patterns, indicating that the model was adequate. Overall, our linear regression analysis provided valuable insights into the relationships between the variables in our dataset.

The next part of the analysis involved creating choropleth maps to visualize the smoking prevalence and legislation percentage by state in the United States. To do this, we filtered the dataset to select the year we were interested in, which was 2015. We then merged this filtered data with an ANSI table that contained the state abbreviations and names.

We used the United States map data from the Altair library and created a base map with light gray fill and black borders. We then created two choropleth maps, one for legislation percentage and one for smoking prevalence. For each map, we used the 'mark_geoshape' mark to represent each state as a polygon and used 'transform_lookup' to match each state in the data to its corresponding state in the map using the state abbreviation. We also specified the 'Color' encoding to map the data value to a color scale, and added a title to the plot to indicate the variable being represented.

Finally, we combined the two choropleth maps using the 'alt.vconcat' function to vertically concatenate the plots. We also used the 'resolve_scale' function to specify that the color scales should be independent, and set the view's stroke to None to remove the default border. The resulting plot allowed us to quickly compare the smoking prevalence and legislation percentage by state, providing a clear picture of the distribution of smoking-related policies and behaviors in the United States in 2015.

Results:

We conducted an exploratory data analysis of the dataset, which included descriptive statistics and data visualization techniques. We filtered the dataset to focus on smoking prevalence in the United States in the year 2015. We found that the smoking prevalence varied widely across the states. We also analyzed the relationship between smoking prevalence and indoor smoking policies across states. Our analysis revealed that states with higher indoor smoking policies tended to have lower smoking prevalence. This relationship is shown in the scatter plot of Legislation Percentage vs. Smoking Prevalence (Every Day), where the data points are colored by the type of smoking (cigar or e-cigarette). Furthermore, we conducted a linear regression analysis to model the relationship between smoking prevalence and other variables in the dataset. The analysis revealed that gender was not a significant predictor of smoking prevalence. Finally, we visualized the smoking prevalence and indoor smoking policies on a choropleth map of the United States, highlighting the state-level differences in these measures.

Discussion

After performing an exploratory data analysis of the dataset, We found that the smoking prevalence is different for each state. Furthermore, the results show that states with higher legistration percentage would result in a lower smoking frequency. Additionally, the linear regression analysis revealed that gender was not a significant predictor of smoking prevalence. Last but not least, the smoking prevalence and indoor smoking policies on a choropleth map of the United States manifests the state-level differences in these measures. Before conducting the analysis, we expected to see a negative relationship, ie. the increase of legistration percentage will lead to a decrease of cigarette and e-cigarette usage, because the policy refines smoker's smoking area; also, we assumed that gender would not be a significant predictor as the legistration have effects on both male and female. For this project, we analyzed the relationship between smoking frequency and legistration percentage and constructed choropleth map of the US to show the smoking policy and smoking prevalence in each state, and we believe we can explore the dataset one step further by considering different smoking freqency and various years to make our results more convincing.