In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn import ensemble
from sklearn.model_selection import train_test_split, GridSearchCV, PredefinedSp
from sklearn.metrics import mean_squared_error, make_scorer

from xgboost import XGBRegressor
from xgboost import plot_tree
```

Data Prep

Repositioning Column "INFECTION RATE" to the end of the dataset

In [2]:

```
data = pd.read_csv('final_data_cleaned.csv')
data['INFECTION RATE'].isnull().any()
# no null value in 'INFECTION RATE' --> treat 0.0 as missing values
```

Out[2]:

```
False
```

In [3]:

```
infection_rate = data['INFECTION RATE']
infection_rate
```

Out[3]:

```
0       0.034420
1       0.031714
2       0.034908
3       0.057709
4       0.020817
          ...
3142    0.015867
3143    0.229845
3144    0.019409
3145    0.061508
3146    0.000000
Name: INFECTION RATE, Length: 3147, dtype: float64
```

In [4]:

```
data = data.drop('INFECTION RATE', axis = 1)
data
```

Out[4]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	A
0	0500000US01001	Autauga County	Alabama	55200.0	7.0	6.0	6.6	
1	0500000US01003	Baldwin County	Alabama	208107.0	6.1	5.2	5.7	
2	0500000US01005	Barbour County	Alabama	25782.0	5.6	6.4	7.2	
3	0500000US01007	Bibb County	Alabama	22527.0	6.7	6.6	6.9	
4	0500000US01009	Blount County	Alabama	57645.0	6.4	5.5	5.9	
...	...	...	...	...	...	...	...	
3142	0500000US56037	Sweetwater County	Wyoming	44117.0	7.2	6.2	7.1	
3143	0500000US56039	Teton County	Wyoming	23059.0	4.0	4.8	9.3	
3144	0500000US56041	Uinta County	Wyoming	20609.0	7.2	5.5	5.4	
3145	0500000US56043	Washakie County	Wyoming	8129.0	7.1	5.0	4.6	
3146	0500000US56045	Weston County	Wyoming	7100.0	4.5	6.7	5.1	

3147 rows × 101 columns

In [5]:

```
1 data.insert(101, "INFECTION RATE", infection_rate)
2 data
```

Out[5]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	A
0	0500000US01001	Autauga County	Alabama	55200.0	7.0	6.0	6.6	
1	0500000US01003	Baldwin County	Alabama	208107.0	6.1	5.2	5.7	
2	0500000US01005	Barbour County	Alabama	25782.0	5.6	6.4	7.2	
3	0500000US01007	Bibb County	Alabama	22527.0	6.7	6.6	6.9	
4	0500000US01009	Blount County	Alabama	57645.0	6.4	5.5	5.9	
...	...	...	...	...	...	...	...	
3142	0500000US56037	Sweetwater County	Wyoming	44117.0	7.2	6.2	7.1	
3143	0500000US56039	Teton County	Wyoming	23059.0	4.0	4.8	9.3	
3144	0500000US56041	Uinta County	Wyoming	20609.0	7.2	5.5	5.4	
3145	0500000US56043	Washakie County	Wyoming	8129.0	7.1	5.0	4.6	
3146	0500000US56045	Weston County	Wyoming	7100.0	4.5	6.7	5.1	

3147 rows × 102 columns

Identify Prediction Set (Counties with 0.0% Infection Rate)

In [6]:

```
pred_set = data.loc[data['INFECTION RATE'] == 0.0]
data = data.drop(pred_set.index)
data
```

Out[6]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	A
0	0500000US01001	Autauga County	Alabama	55200.0	7.0	6.0	6.6	
1	0500000US01003	Baldwin County	Alabama	208107.0	6.1	5.2	5.7	
2	0500000US01005	Barbour County	Alabama	25782.0	5.6	6.4	7.2	
3	0500000US01007	Bibb County	Alabama	22527.0	6.7	6.6	6.9	
4	0500000US01009	Blount County	Alabama	57645.0	6.4	5.5	5.9	
...	...	...	...	...	...	...	...	
3141	0500000US56035	Sublette County	Wyoming	9951.0	6.6	3.4	4.7	
3142	0500000US56037	Sweetwater County	Wyoming	44117.0	7.2	6.2	7.1	
3143	0500000US56039	Teton County	Wyoming	23059.0	4.0	4.8	9.3	
3144	0500000US56041	Uinta County	Wyoming	20609.0	7.2	5.5	5.4	
3145	0500000US56043	Washakie County	Wyoming	8129.0	7.1	5.0	4.6	

2618 rows × 102 columns

In [7]:

```
1 pred_set
```

Out[7]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	AG
30	0500000US01061	Geneva County	Alabama	26491.0	6.3	5.3	5.4	
67	0500000US02013	Aleutians East Borough	Alaska	3425.0	5.0	6.9	8.1	
68	0500000US02016	Aleutians West Census Area	Alaska	5750.0	4.8	9.2	8.6	
71	0500000US02060	Bristol Bay Borough	Alaska	890.0	5.1	4.3	5.8	
72	0500000US02068	Denali Borough	Alaska	2232.0	5.6	13.4	5.5	
...	...	...	...	...	...	...	...	
3114	0500000US55123	Vernon County	Wisconsin	30516.0	6.6	4.7	4.6	
3117	0500000US55129	Washburn County	Wisconsin	15689.0	5.0	4.0	3.9	
3125	0500000US56003	Big Horn County	Wyoming	11901.0	7.0	4.5	5.0	
3139	0500000US56031	Platte County	Wyoming	8673.0	5.7	4.9	4.3	
3146	0500000US56045	Weston County	Wyoming	7100.0	4.5	6.7	5.1	

529 rows × 102 columns

Identify the X and y from Data and Prediction Set

In [8]:

```
X = data.iloc[:, 4:101]
X = X.drop("LAND AREA(SQMI)", axis = 1) # No total population or land area
y = data["INFECTION RATE"]
X
```

Out[8]:

	AGE15_19	AGE20_24	AGE25_29	AGE30_34	AGE55_59	AGE60_64	TOTAL_MALE	TOTAL_
0	7.0	6.0	6.6	6.2	7.5	4.9	26874.0	
1	6.1	5.2	5.7	5.5	6.9	7.1	101188.0	
2	5.6	6.4	7.2	7.0	6.4	6.4	13697.0	
3	6.7	6.6	6.9	6.7	6.6	5.4	12152.0	
4	6.4	5.5	5.9	5.9	6.8	6.3	28434.0	
...	...	...	...	...	...	...	...	
3141	6.6	3.4	4.7	6.0	7.3	6.8	5367.0	
3142	7.2	6.2	7.1	7.7	7.0	6.1	22882.0	
3143	4.0	4.8	9.3	9.1	6.4	6.6	11911.0	
3144	7.2	5.5	5.4	6.5	8.5	5.7	10505.0	
3145	7.1	5.0	4.6	5.0	7.1	7.5	4137.0	

2618 rows × 96 columns

In [9]:

```
pred_X = pred_set.iloc[:, 4:101]
pred_X = pred_X.drop("LAND AREA(SQMI)", axis = 1) # No total population or land
pred_X
```

Out[9]:

	AGE15_19	AGE20_24	AGE25_29	AGE30_34	AGE55_59	AGE60_64	TOTAL_MALE	TOTAL_
30	6.3	5.3	5.4	5.8	7.7	6.4	12971.0	
67	5.0	6.9	8.1	6.3	7.7	7.2	1989.0	
68	4.8	9.2	8.6	8.8	7.4	6.6	3432.0	
71	5.1	4.3	5.8	4.4	11.7	8.4	526.0	
72	5.6	13.4	5.5	6.5	7.6	14.2	1117.0	
...	...	...	...	...	...	...	...	
3114	6.6	4.7	4.6	5.0	8.0	7.5	15302.0	
3117	5.0	4.0	3.9	4.7	8.8	9.4	7771.0	
3125	7.0	4.5	5.0	5.6	6.2	7.4	6033.0	
3139	5.7	4.9	4.3	5.4	6.4	9.3	4499.0	
3146	4.5	6.7	5.1	6.0	5.3	11.4	3768.0	

529 rows × 96 columns

Split Data into 2/3 Training and 1/3 Testing

In [10]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random
```

SKLearn GradientBoostingRegressor

Baseline SKLearn GradientBoostingRegressor Model

In [11]:

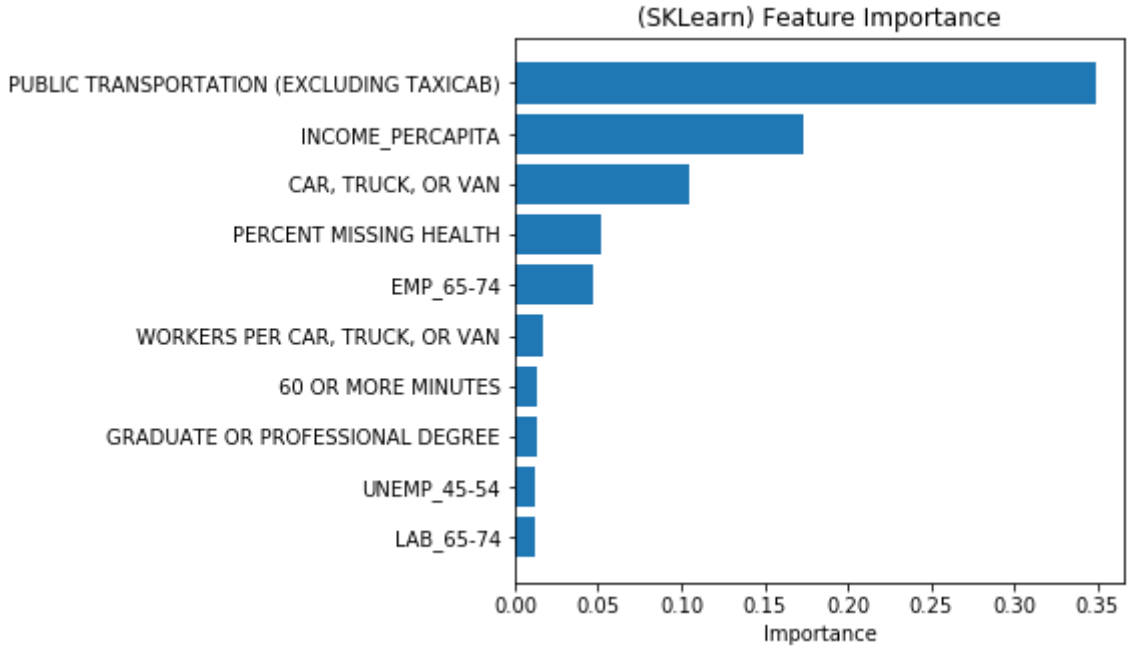
```
gbr_model = ensemble.GradientBoostingRegressor()
gbr_model.fit(X_train, y_train)
mean_squared_error(y_test, gbr_model.predict(X_test))
```

Out[11]:

```
0.02117833305438754
```

In [12]:

```
feature_importance = gbr_model.feature_importances_
#feature_importance = 100.0 * (feature_importance / feature_importance.max())
sorted_idx = np.argsort(feature_importance)
top_10 = sorted_idx[-10:]
pos = np.arange(top_10.shape[0]) + .5
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 2)
plt.barh(pos, feature_importance[top_10], align='center')
plt.yticks(pos, X.columns[top_10])
plt.xlabel('Importance')
plt.title('(SKLearn) Feature Importance')
plt.show()
```



SKLearn: Hyperparameter Tuning

In [14]:

```
test_fold = [-1]*len(X_train) + [0]*len(X_test)
predefined_split = PredefinedSplit(test_fold=test_fold)
predefined_split
```

Out[14]:

```
PredefinedSplit(test_fold=array([-1, -1, ...,  0,  0]))
```

In [15]:

```
param_grid = {
    'max_depth': [3, 5, 7],
    'min_samples_split': [2, 4, 6],
    'n_estimators': [300, 500, 700],
    'learning_rate': [0.001, 0.01, 0.1]
}
gbr = ensemble.GradientBoostingRegressor()
grid = GridSearchCV(
    gbr,
    param_grid,
    scoring = make_scorer(mean_squared_error, greater_is_better = False),
    return_train_score=True
)
grid.fit(X_train, y_train)

# pd.set_option('display.max_rows', 5)
df = pd.DataFrame(grid.cv_results_)
# Flip sign of score back, because GridSearchCV likes to maximize,
# so it flips the sign of the score if "greater_is_better = FALSE"
df['mean_test_score'] = -df['mean_test_score']
df['mean_train_score'] = -df['mean_train_score']
cols_to_keep = ["param_learning_rate", "param_max_depth","param_min_samples_spli
                "mean_test_score","mean_train_score"]
df_toshow = df[cols_to_keep].fillna('-')
df_toshow = df_toshow.sort_values(by=["mean_test_score"])
df_toshow.head()
```

Out[15]:

	param_learning_rate	param_max_depth	param_min_samples_split	param_n_estimators	mean
78	0.1	7	6	300	
60	0.1	3	6	300	
65	0.1	5	2	700	
64	0.1	5	2	500	
79	0.1	7	6	500	

In [24]:

```
best_params = df_toshow.head(1)

params = {
    'n_estimators': best_params['param_n_estimators'].item(),
    'max_depth': best_params['param_max_depth'].item(),
    'min_samples_split': best_params['param_min_samples_split'].item(),
    'learning_rate': best_params['param_learning_rate'].item(),
    'loss': 'ls', 'criterion': 'mse'
}
gbr_model = ensemble.GradientBoostingRegressor(**params)
gbr_model.fit(X_train, y_train)
mean_squared_error(y_test, gbr_model.predict(X_test))

# Worse than baseline model --> Return to the baseline model
```

Out[24]:

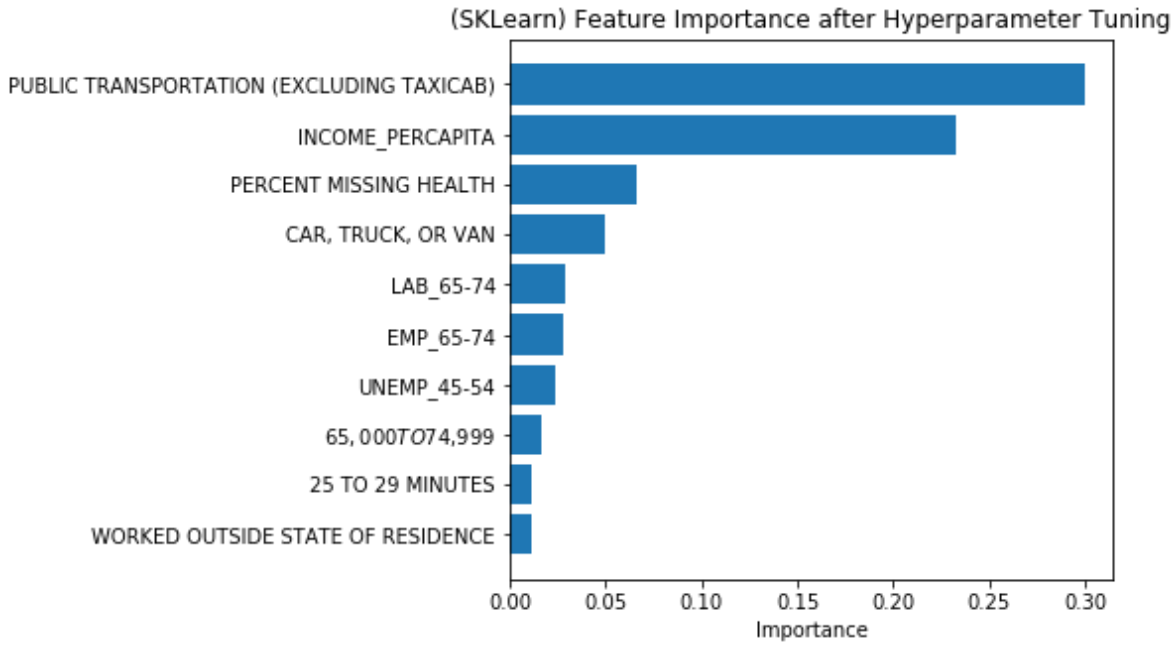
```
0.02180453198751297
```

In [ ]:

```
# baseline default parameter values
params = {
    'n_estimators': 100,
    'max_depth': 3,
    'min_samples_split': 2,
    'learning_rate': 0.1,
    'loss': 'ls', 'criterion': 'mse'
}
gbr_model = ensemble.GradientBoostingRegressor(**params)
gbr_model.fit(X_train, y_train)
mean_squared_error(y_test, gbr_model.predict(X_test))
```

In [25]:

```
feature_importance = gbr_model.feature_importances_
#feature_importance = 100.0 * (feature_importance / feature_importance.max())
sorted_idx = np.argsort(feature_importance)
top_10 = sorted_idx[-10:]
pos = np.arange(top_10.shape[0]) + .5
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 2)
plt.barh(pos, feature_importance[top_10], align='center')
plt.yticks(pos, X.columns[top_10])
plt.xlabel('Importance')
plt.title('(SKLearn) Feature Importance after Hyperparameter Tuning')
plt.show()
```



SKLearn Infection Rate Prediction

In [30]:

```
infection_rate_predictions = gbr_model.predict(pred_X)
pred_set["INFECTION RATE"] = infection_rate_predictions
pred_set
```

```
//anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:2: Setti
ngWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas
-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)
```

Out[30]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	AG
30	0500000US01061	Geneva County	Alabama	26491.0	6.3	5.3	5.4	
67	0500000US02013	Aleutians East Borough	Alaska	3425.0	5.0	6.9	8.1	
68	0500000US02016	Aleutians West Census Area	Alaska	5750.0	4.8	9.2	8.6	
71	0500000US02060	Bristol Bay Borough	Alaska	890.0	5.1	4.3	5.8	
72	0500000US02068	Denali Borough	Alaska	2232.0	5.6	13.4	5.5	
...	...	...	...	...	...	...	...	
3114	0500000US55123	Vernon County	Wisconsin	30516.0	6.6	4.7	4.6	
3117	0500000US55129	Washburn County	Wisconsin	15689.0	5.0	4.0	3.9	
3125	0500000US56003	Big Horn County	Wyoming	11901.0	7.0	4.5	5.0	
3139	0500000US56031	Platte County	Wyoming	8673.0	5.7	4.9	4.3	
3146	0500000US56045	Weston County	Wyoming	7100.0	4.5	6.7	5.1	

529 rows × 102 columns

XGBRegressor

Baseline XGBRegressor Model

In [31]:

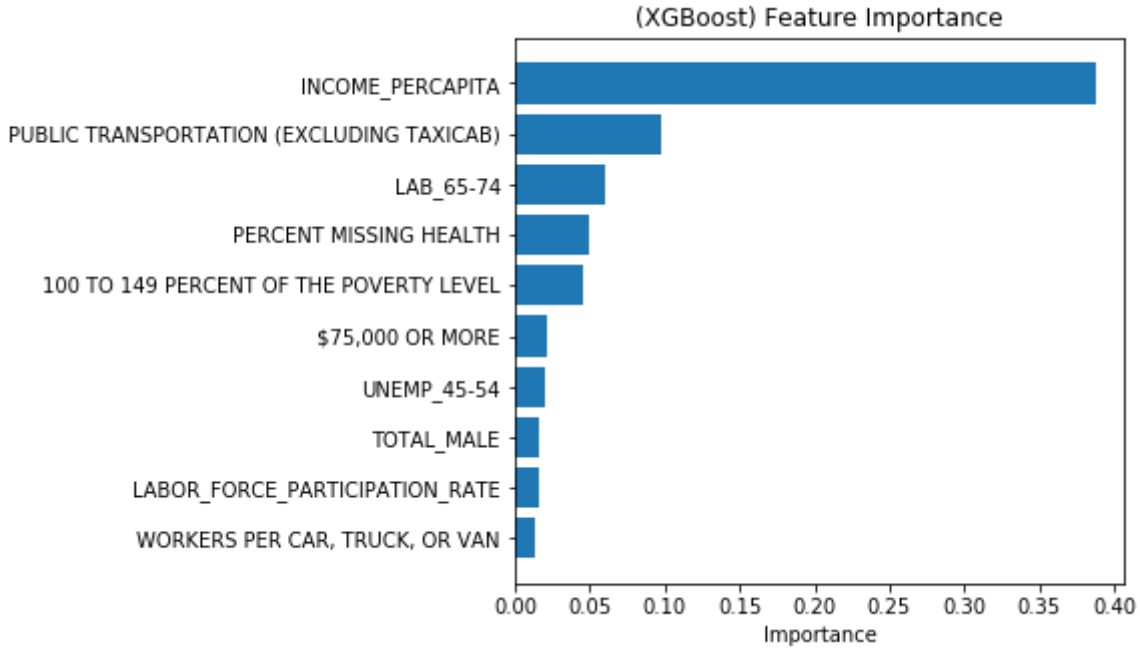
```
xgbr_model = XGBRegressor()
xgbr_model.fit(X_train, y_train)
mean_squared_error(y_test, xgbr_model.predict(X_test))
```

Out[31]:

```
0.021640949145811064
```

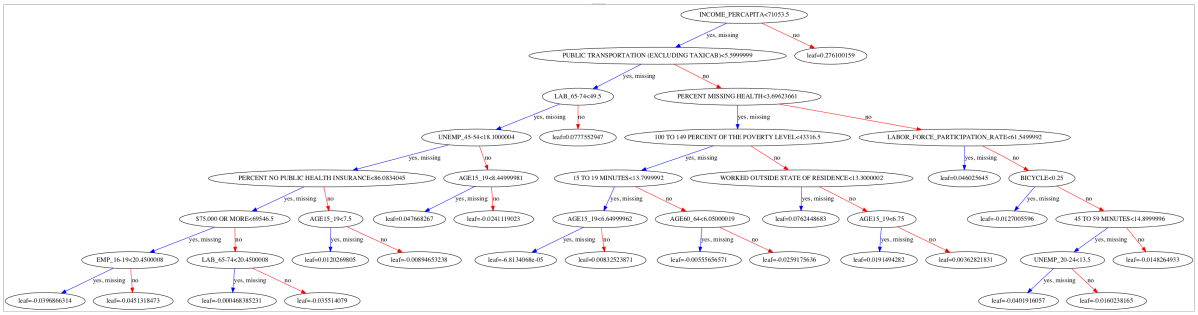
In [32]:

```
feature_importance = xgbr_model.feature_importances_
sorted_idx = np.argsort(feature_importance)
top_10 = sorted_idx[-10:]
pos = np.arange(top_10.shape[0]) + .5
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 2)
plt.barh(pos, feature_importance[top_10], align='center')
plt.yticks(pos, X.columns[top_10])
plt.xlabel('Importance')
plt.title('(XGBoost) Feature Importance')
plt.show()
```



In [42]:

```
plot_tree(xgbr_model)
fig = plt.gcf()
fig.set_size_inches(300, 400)
plt.title('(XGBoost) Gradient Boosting Tree')
plt.show()
```



XGBRegressor: Hyperparameter Tuning

In [34]:

```
xparam_grid = {
    'max_depth': [3, 5, 7],
    'min_child_weight': [2, 4, 6],
    'gamma': [0, 0.1, 0.2],
    'n_estimators': [300, 500, 700],
    'learning_rate': [0.001, 0.01, 0.1]
}
xgbr = XGBRegressor()
xgrid = GridSearchCV(
    xgbr,
    xparam_grid,
    scoring = make_scorer(mean_squared_error, greater_is_better = False),
    return_train_score=True
)
xgrid.fit(X_train, y_train)

# pd.set_option('display.max_rows', 20)
xdf = pd.DataFrame(xgrid.cv_results_)
# Flip sign of score back, because GridSearchCV likes to maximize,
# so it flips the sign of the score if "greater_is_better = FALSE"
xdf['mean_test_score'] = -xdf['mean_test_score']
xdf['mean_train_score'] = -xdf['mean_train_score']
cols_to_keep = ["param_gamma", "param_learning_rate", "param_max_depth",
                "param_min_child_weight", "param_n_estimators",
                "mean_test_score","mean_train_score"]
xdf_toshow = xdf[cols_to_keep].fillna('-')
xdf_toshow = xdf_toshow.sort_values(by=["mean_test_score"])
xdf_toshow.head()
```

Out[34]:

	param_gamma	param_learning_rate	param_max_depth	param_min_child_weight	param_n_e
48	0.0	0.01	7	4	
49	0.0	0.01	7	4	
129	0.1	0.01	7	4	
120	0.1	0.01	5	4	
40	0.0	0.01	5	4	

In [35]:

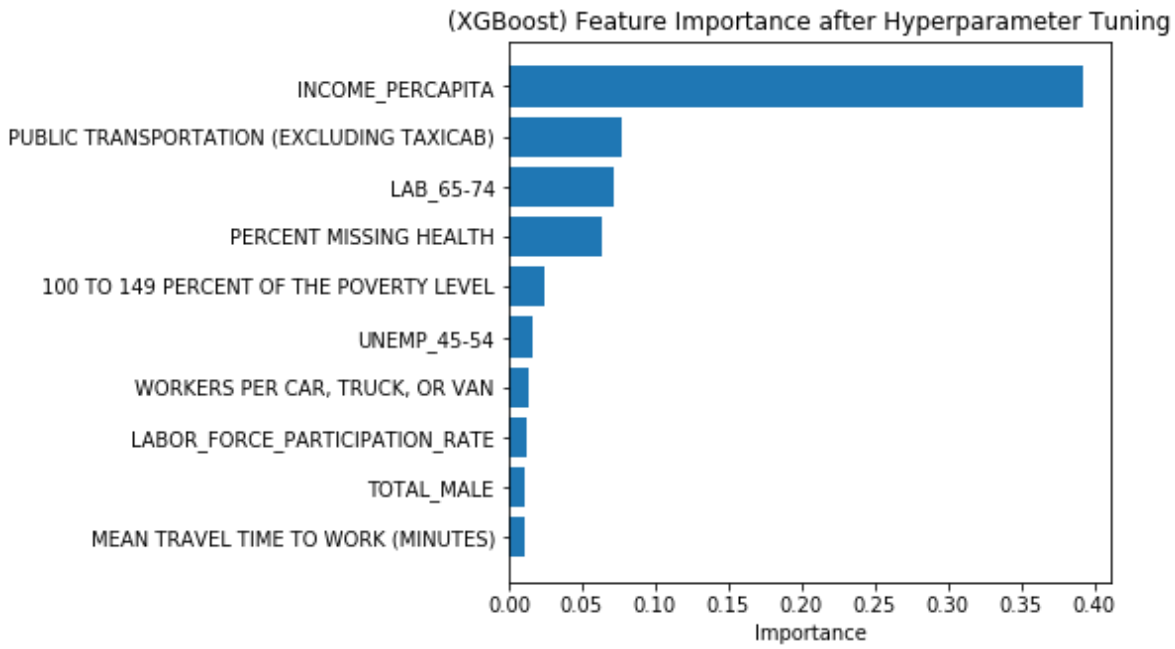
```
xbest_params = xdf_toshow.head(1)

xparams = {
    'n_estimators': xbest_params['param_n_estimators'].item(),
    'max_depth': xbest_params['param_max_depth'].item(),
    'min_child_weight': xbest_params['param_min_child_weight'].item(),
    'learning_rate': xbest_params['param_learning_rate'].item(),
    'gamma': xbest_params['param_gamma'].item(),
    'loss': 'ls', 'criterion': 'mse'
}
xgbr_model = XGBRegressor(**params)
xgbr_model.fit(X_train, y_train)
mean_squared_error(y_test, xgbr_model.predict(X_test))

# Better than baseline model
```

Out[35]:

```
0.02097232781096379
```

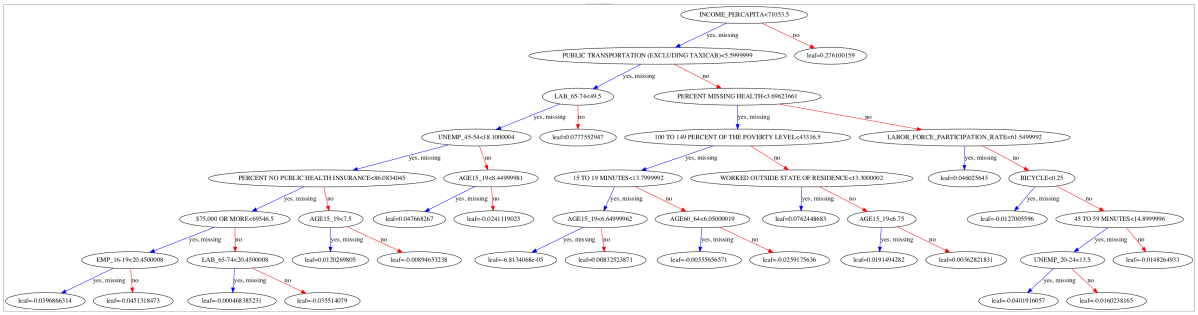
In [36]:

```
feature_importance = xgbr_model.feature_importances_
sorted_idx = np.argsort(feature_importance)
top_10 = sorted_idx[-10:]
pos = np.arange(top_10.shape[0]) + .5
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 2)
plt.barh(pos, feature_importance[top_10], align='center')
plt.yticks(pos, X.columns[top_10])
plt.xlabel('Importance')
plt.title('(XGBoost) Feature Importance after Hyperparameter Tuning')
plt.show()
```



In [41]:

```
plot_tree(xgbr_model)
fig = plt.gcf()
fig.set_size_inches(300, 400)
plt.title('(XGBoost) Gradient Boosting Tree after Hyperparameter Tuning')
plt.show()
```



XGBoost Infection Rate Prediction

In [37]:

```
xinfection_rate_predictions = xgbr_model.predict(pred_X)
pred_set["XGBOOST INFECTION RATE"] = xinfection_rate_predictions
pred_set
```

```
//anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:2: Setti
ngWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas
-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)
```

Out[37]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	AG
30	0500000US01061	Geneva County	Alabama	26491.0	6.3	5.3	5.4	
67	0500000US02013	Aleutians East Borough	Alaska	3425.0	5.0	6.9	8.1	
68	0500000US02016	Aleutians West Census Area	Alaska	5750.0	4.8	9.2	8.6	
71	0500000US02060	Bristol Bay Borough	Alaska	890.0	5.1	4.3	5.8	
72	0500000US02068	Denali Borough	Alaska	2232.0	5.6	13.4	5.5	
...	...	...	...	...	...	...	...	
3114	0500000US55123	Vernon County	Wisconsin	30516.0	6.6	4.7	4.6	
3117	0500000US55129	Washburn County	Wisconsin	15689.0	5.0	4.0	3.9	
3125	0500000US56003	Big Horn County	Wyoming	11901.0	7.0	4.5	5.0	
3139	0500000US56031	Platte County	Wyoming	8673.0	5.7	4.9	4.3	
3146	0500000US56045	Weston County	Wyoming	7100.0	4.5	6.7	5.1	

529 rows × 103 columns

In [44]:

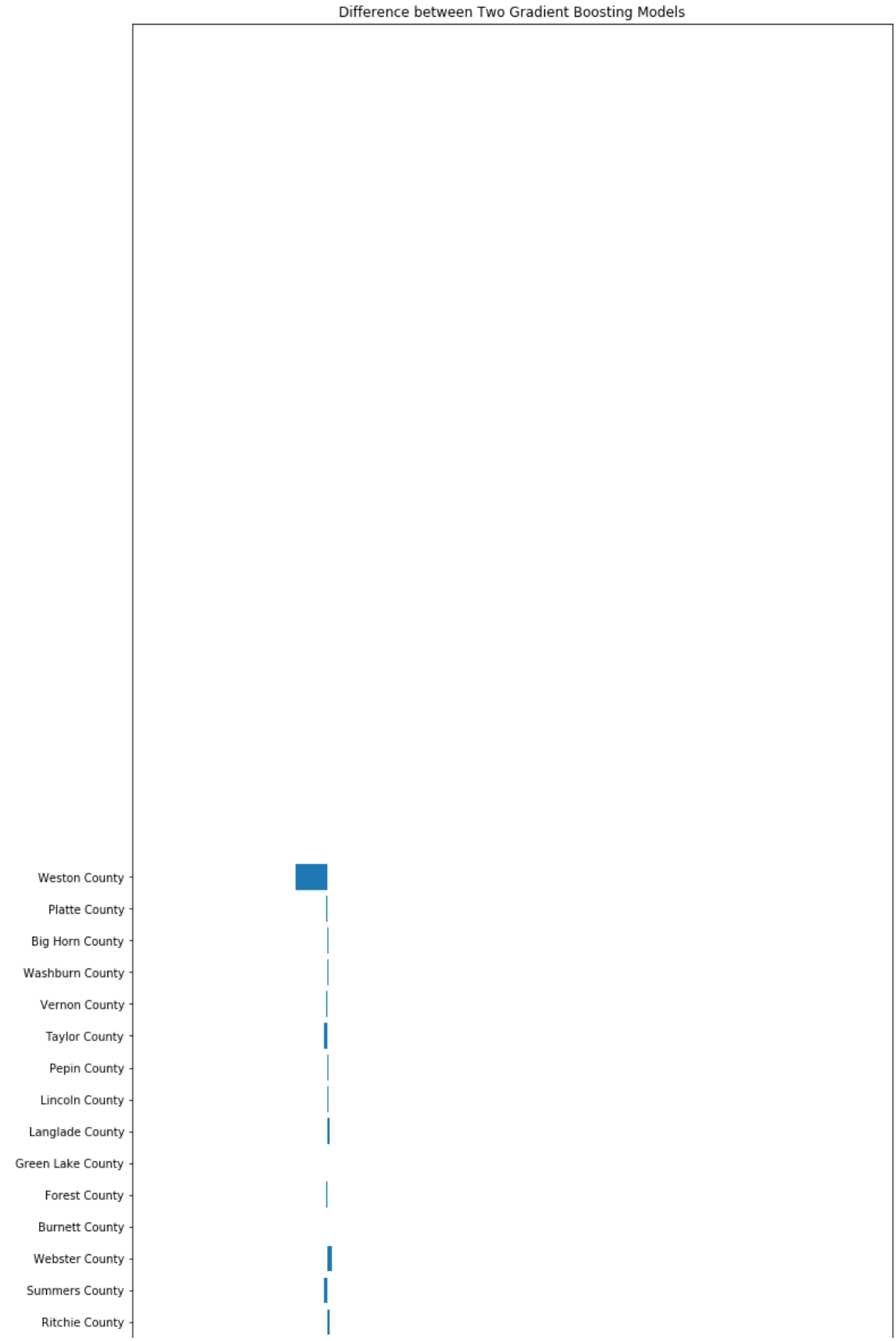
```
1 sum(pred_set["INFECTION RATE"] - pred_set["XGBOOST INFECTION RATE"])/len(pred_se
```

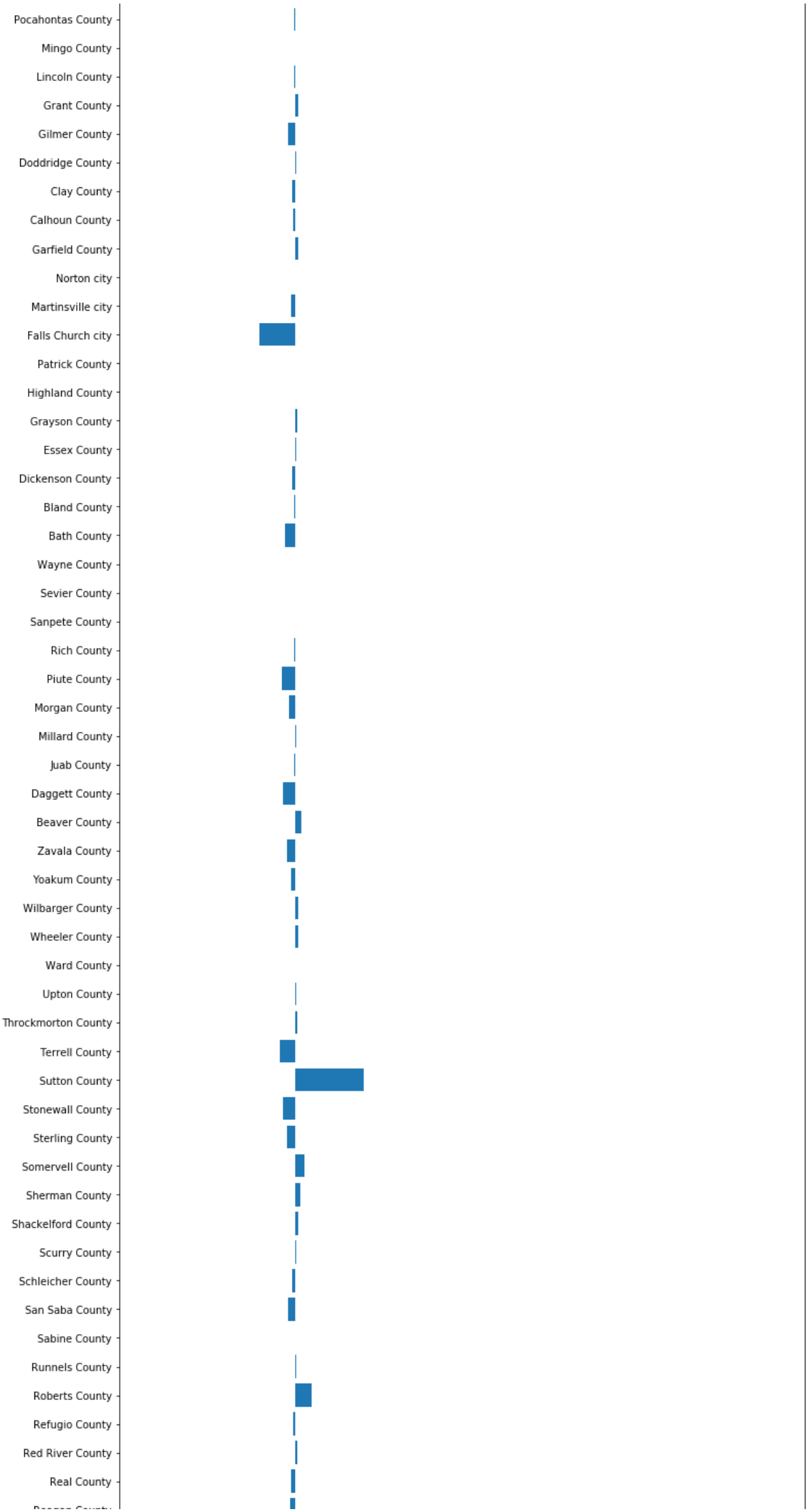
Out[44]:

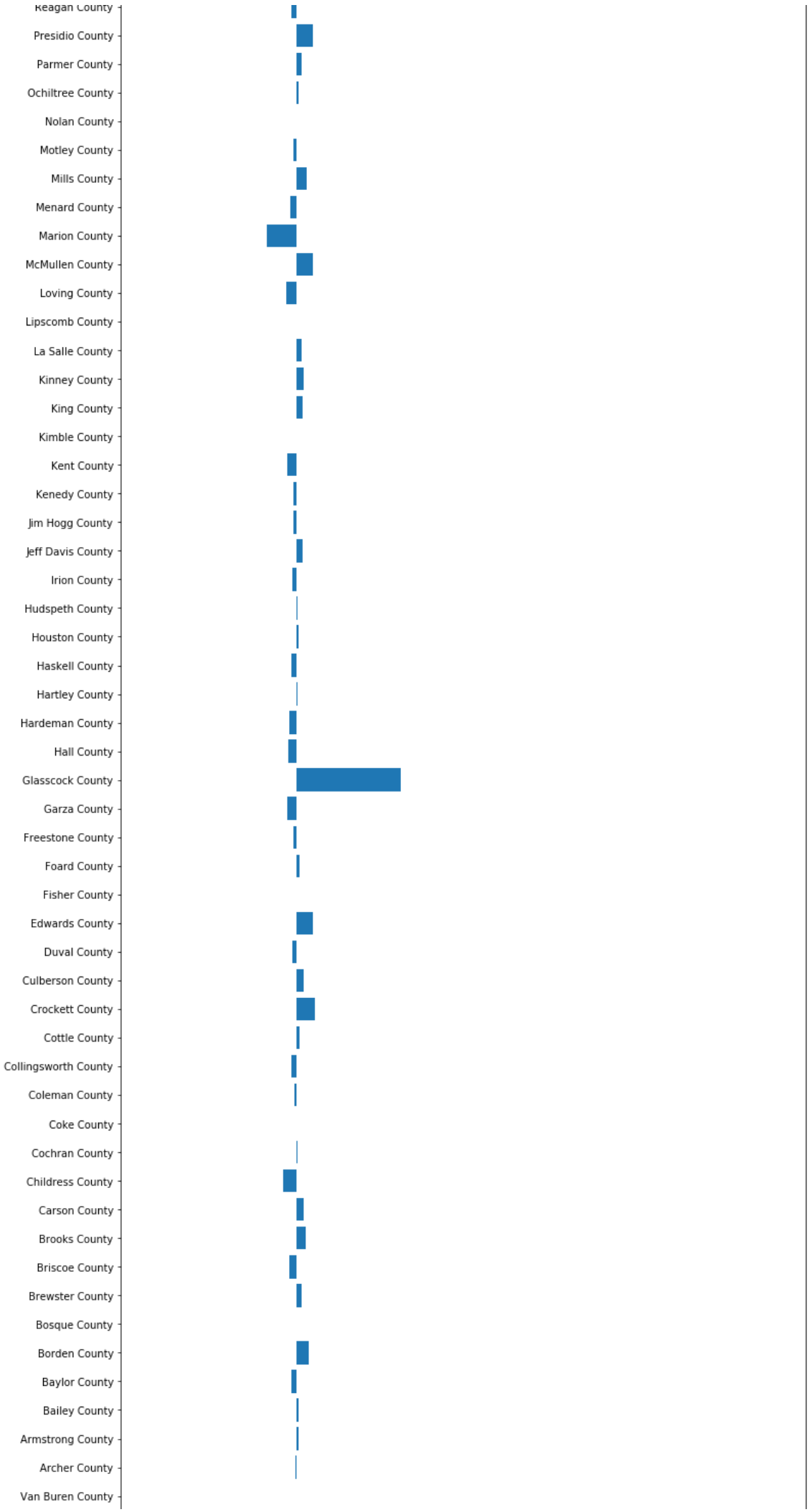
```
0.03650171111683398
```

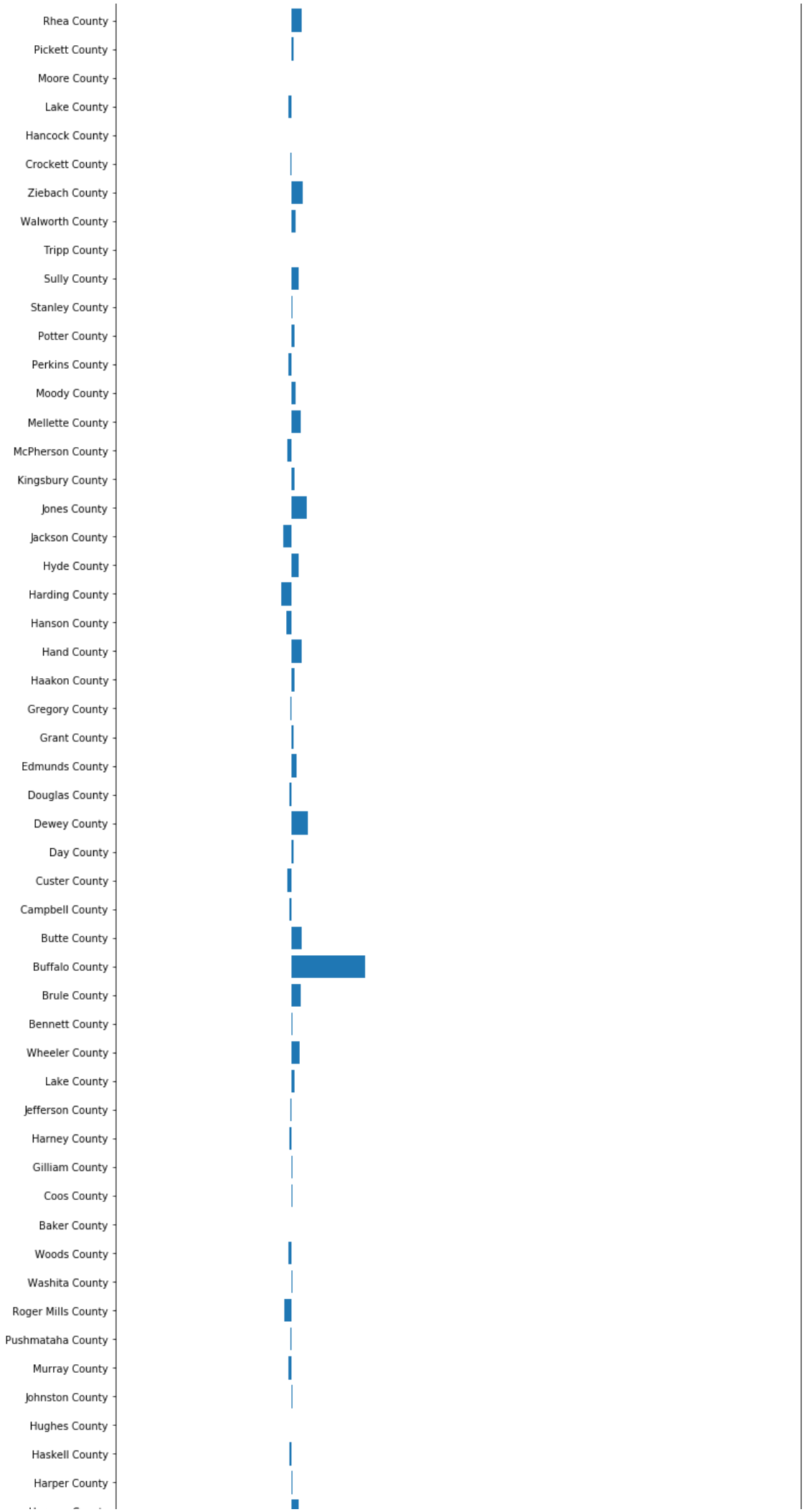
In [64]:

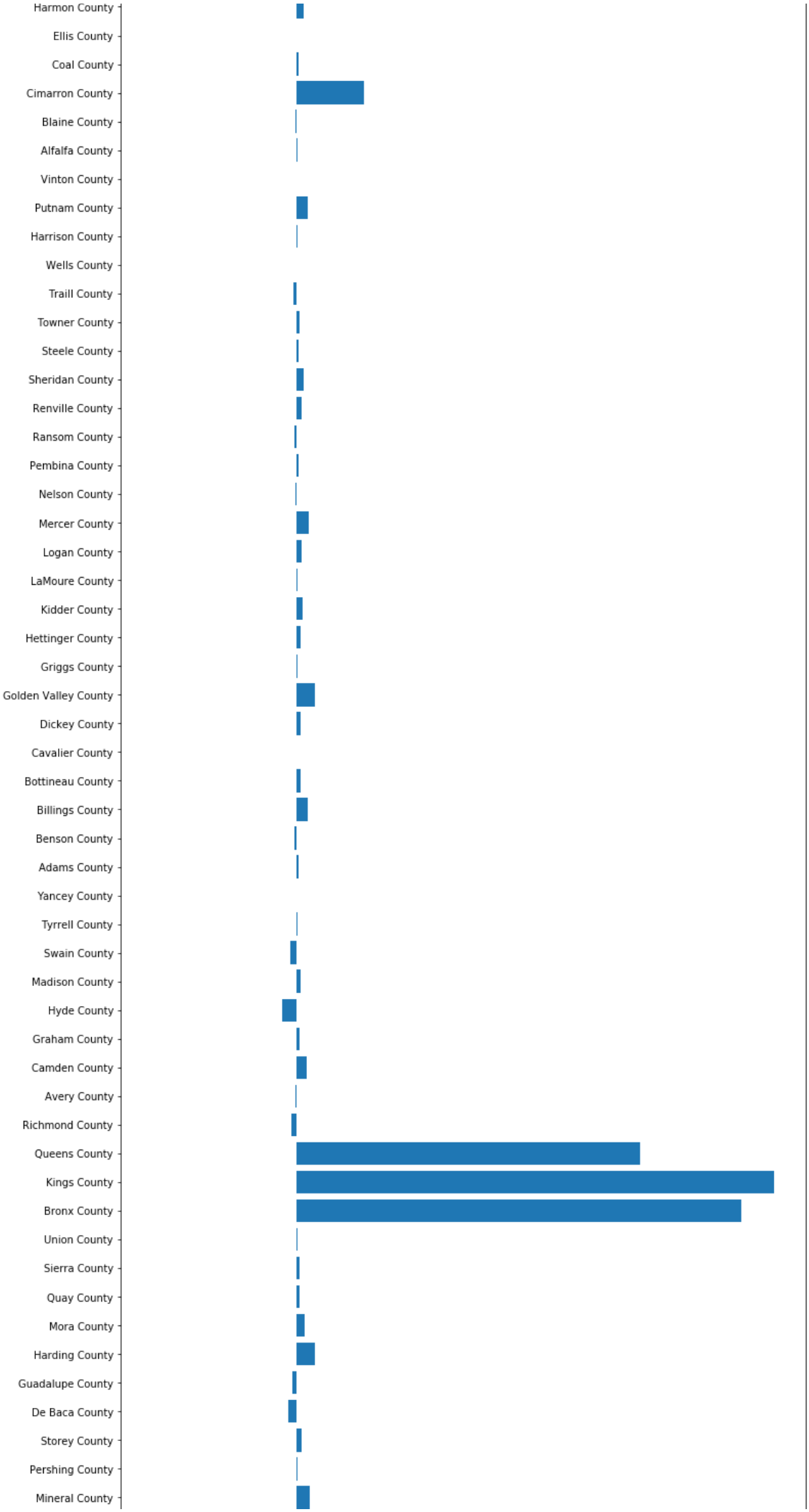
```
difference = pred_set["INFECTION RATE"] - pred_set["XGBOOST INFECTION RATE"]

plt.figure(figsize=(12, 300))
plt.barh(np.arange(len(pred_set)), difference, align='center')
plt.yticks(np.arange(len(pred_set)), pred_set["COUNTY"])
plt.xlabel('Difference')
plt.title('Difference between Two Gradient Boosting Models')

plt.show()
```

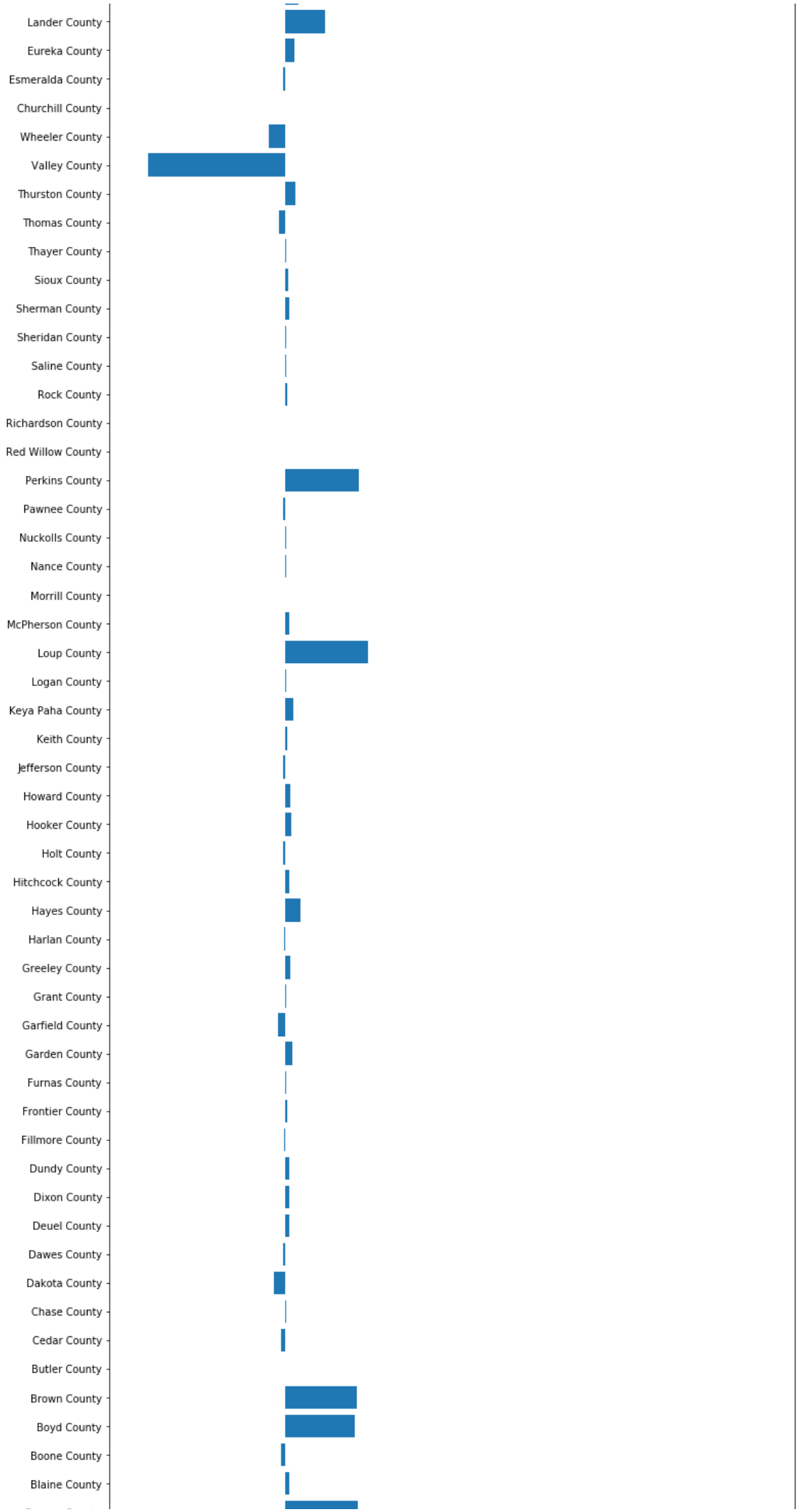


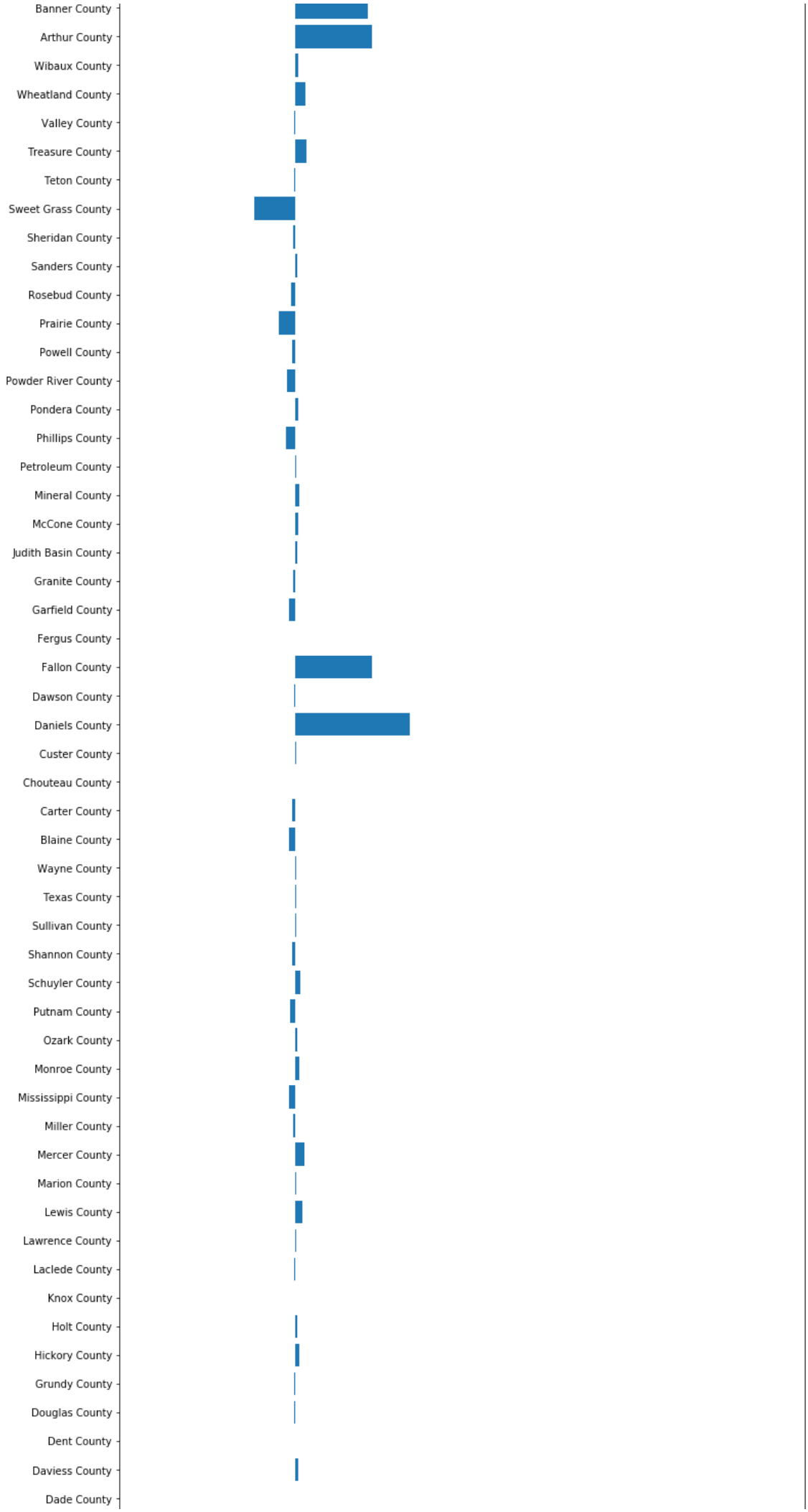

Pocahontas County
Mingo County
Lincoln County
Grant County
Gilmer County
Doddridge County
Clay County
Calhoun County
Garfield County
Norton city
Martinsville city
Falls Church city
Patrick County
Highland County
Grayson County
Essex County
Dickenson County
Bland County
Bath County
Wayne County
Sevier County
Sanpete County
Rich County
Piute County
Morgan County
Millard County
Juab County
Daggett County
Beaver County
Zavala County
Yoakum County
Wilbarger County
Wheeler County
Ward County
Upton County
Throckmorton County
Terrell County
Sutton County
Stonewall County
Sterling County
Somervell County
Sherman County
Shackelford County
Scurry County
Schleicher County
San Saba County
Sabine County
Runnels County
Roberts County
Refugio County
Red River County
Real County

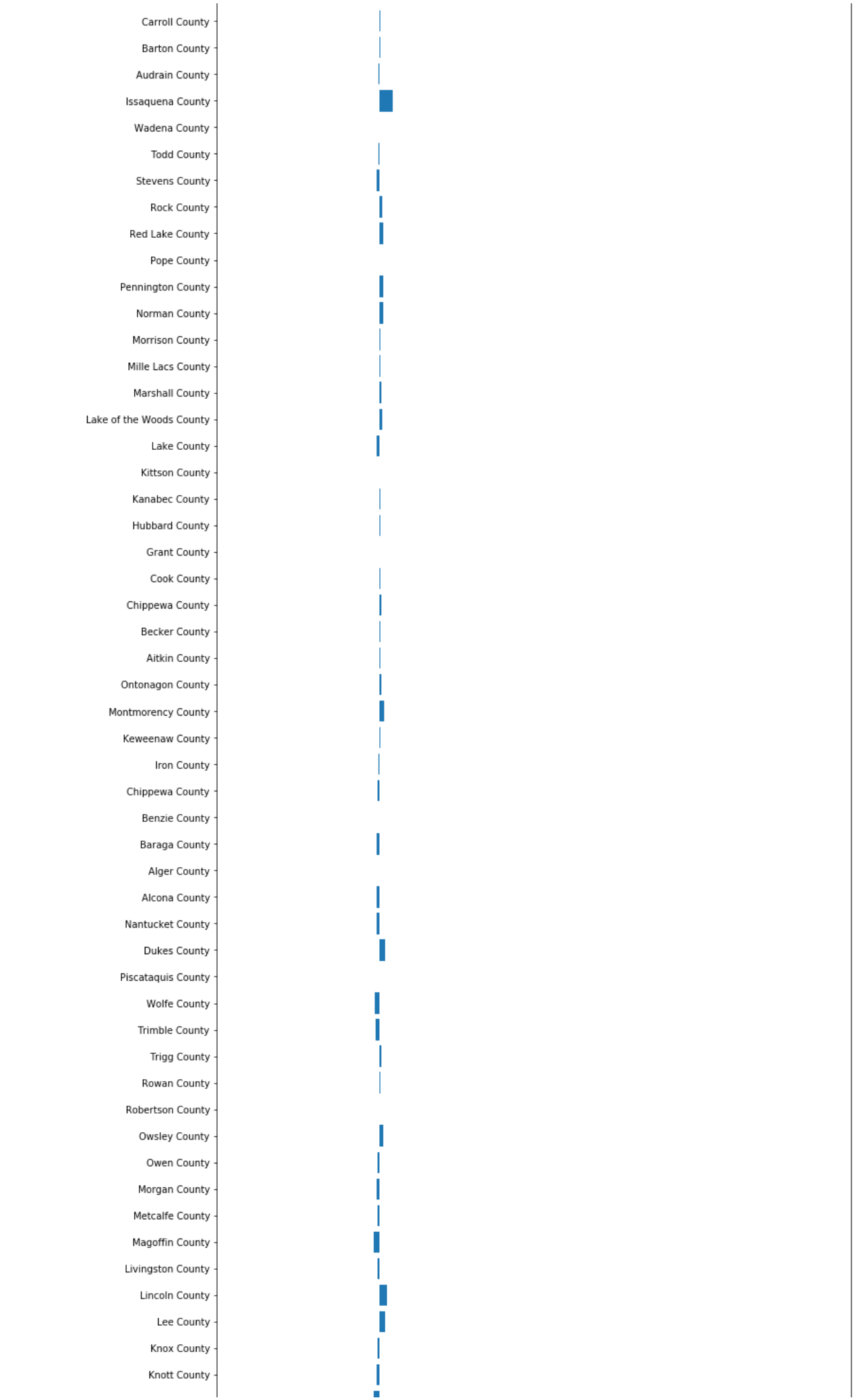

Reagan County
Presidio County
Parmer County
Ochiltree County
Nolan County
Motley County
Mills County
Menard County
Marion County
McMullen County
Loving County
Lipscomb County
La Salle County
Kinney County
King County
Kimble County
Kent County
Kenedy County
Jim Hogg County
Jeff Davis County
Irion County
Hudspeth County
Houston County
Haskell County
Hartley County
Hardeman County
Hall County
Glasscock County
Garza County
Freestone County
Foard County
Fisher County
Edwards County
Duval County
Culberson County
Crockett County
Cottle County
Collingsworth County
Coleman County
Coke County
Cochran County
Childress County
Carson County
Brooks County
Briscoe County
Brewster County
Bosque County
Borden County
Baylor County
Bailey County
Armstrong County
Archer County
Van Buren County

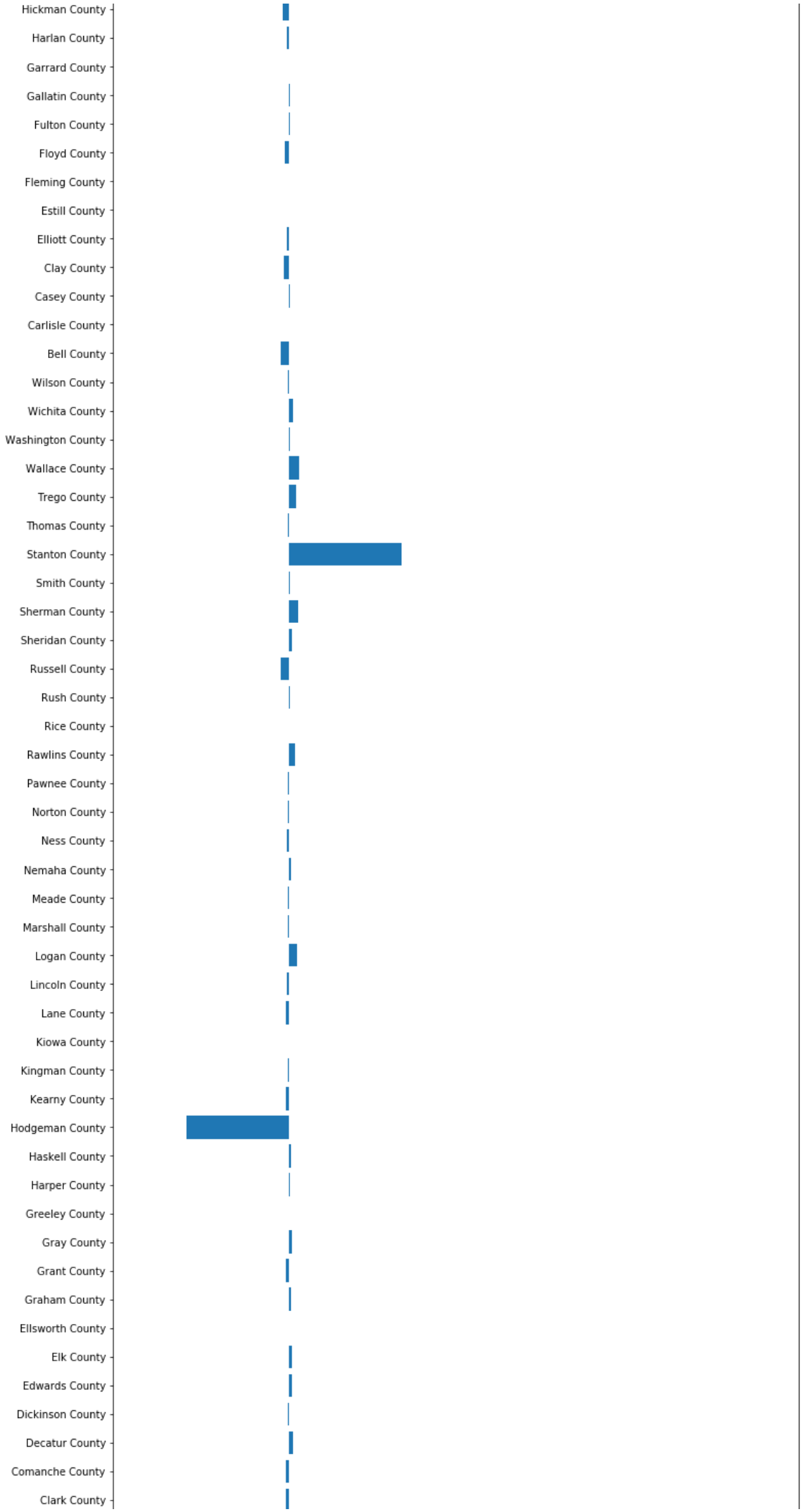

Rhea County
Pickett County
Moore County
Lake County
Hancock County
Crockett County
Ziebach County
Walworth County
Tripp County
Sully County
Stanley County
Potter County
Perkins County
Moody County
Mellette County
McPherson County
Kingsbury County
Jones County
Jackson County
Hyde County
Harding County
Hanson County
Hand County
Haakon County
Gregory County
Grant County
Edmunds County
Douglas County
Dewey County
Day County
Custer County
Campbell County
Butte County
Buffalo County
Brule County
Bennett County
Wheeler County
Lake County
Jefferson County
Harney County
Gilliam County
Coos County
Baker County
Woods County
Washita County
Roger Mills County
Pushmataha County
Murray County
Johnston County
Hughes County
Haskell County
Harper County

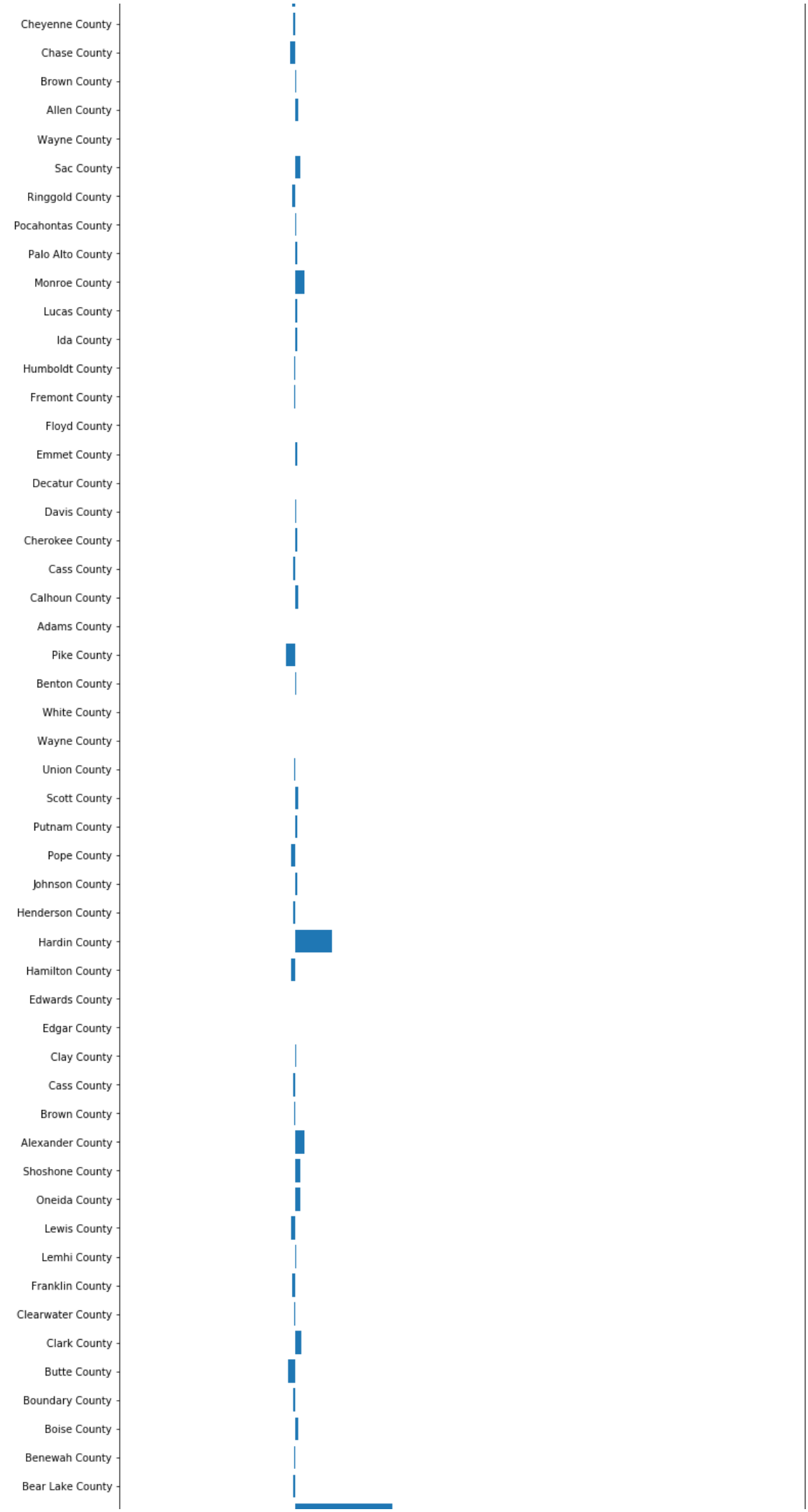

Harmon County
Ellis County
Coal County
Cimarron County
Blaine County
Alfalfa County
Vinton County
Putnam County
Harrison County
Wells County
Traill County
Towner County
Steele County
Sheridan County
Renville County
Ransom County
Pembina County
Nelson County
Mercer County
Logan County
LaMoure County
Kidder County
Hettinger County
Griggs County
Golden Valley County
Dickey County
Cavalier County
Bottineau County
Billings County
Benson County
Adams County
Yancey County
Tyrrell County
Swain County
Madison County
Hyde County
Graham County
Camden County
Avery County
Richmond County
Queens County
Kings County
Bronx County
Union County
Sierra County
Quay County
Mora County
Harding County
Guadalupe County
De Baca County
Storey County
Pershing County
Mineral County

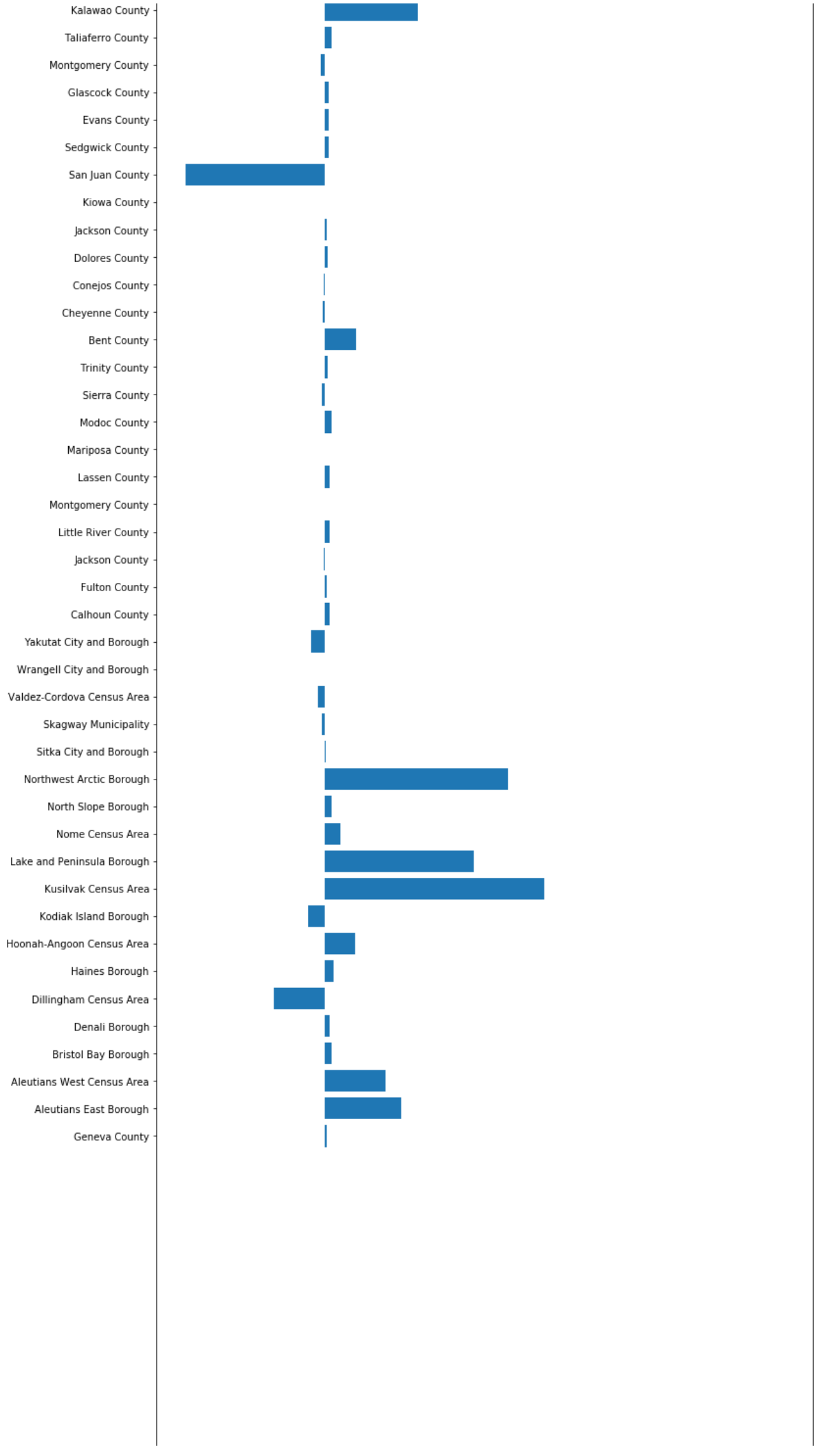

Lander County
Eureka County
Esmeralda County
Churchill County
Wheeler County
Valley County
Thurston County
Thomas County
Thayer County
Sioux County
Sherman County
Sheridan County
Saline County
Rock County
Richardson County
Red Willow County
Perkins County
Pawnee County
Nuckolls County
Nance County
Morrill County
McPherson County
Loup County
Logan County
Keya Paha County
Keith County
Jefferson County
Howard County
Hooker County
Holt County
Hitchcock County
Hayes County
Harlan County
Greeley County
Grant County
Garfield County
Garden County
Furnas County
Frontier County
Fillmore County
Dundy County
Dixon County
Deuel County
Dawes County
Dakota County
Chase County
Cedar County
Butler County
Brown County
Boyd County
Boone County
Blaine County

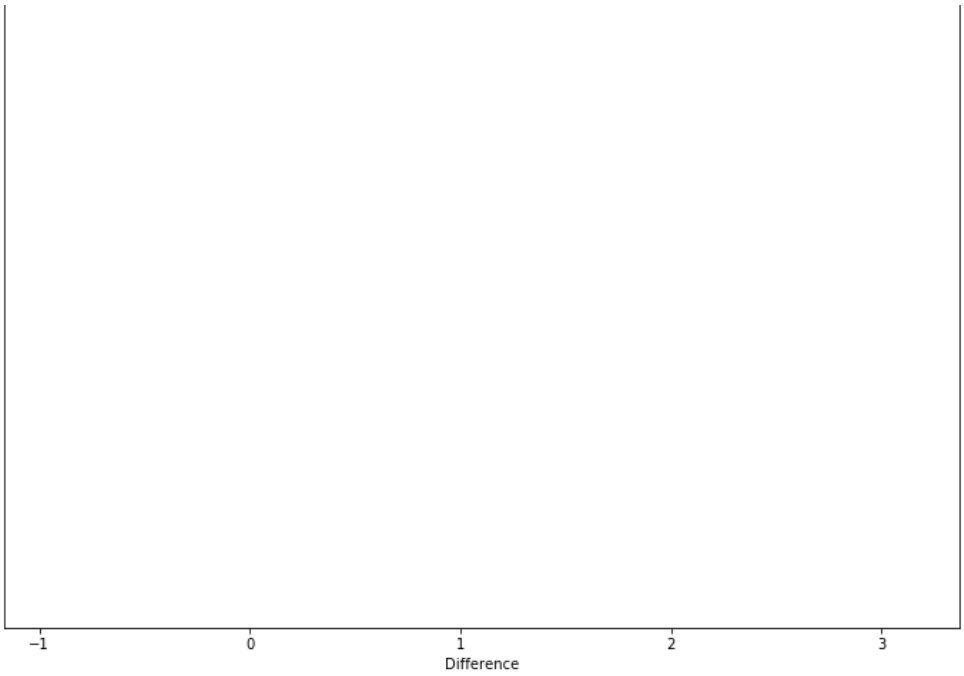

Banner County
Arthur County
Wibaux County
Wheatland County
Valley County
Treasure County
Teton County
Sweet Grass County
Sheridan County
Sanders County
Rosebud County
Prairie County
Powell County
Powder River County
Pondera County
Phillips County
Petroleum County
Mineral County
McCone County
Judith Basin County
Granite County
Garfield County
Fergus County
Fallon County
Dawson County
Daniels County
Custer County
Chouteau County
Carter County
Blaine County
Wayne County
Texas County
Sullivan County
Shannon County
Schuyler County
Putnam County
Ozark County
Monroe County
Mississippi County
Miller County
Mercer County
Marion County
Lewis County
Lawrence County
Laclede County
Knox County
Holt County
Hickory County
Grundy County
Douglas County
Dent County
Daviess County
Dade County


Carroll County
Barton County
Audrain County
Issaquena County
Wadena County
Todd County
Stevens County
Rock County
Red Lake County
Pope County
Pennington County
Norman County
Morrison County
Mille Lacs County
Marshall County
Lake of the Woods County
Lake County
Kittson County
Kanabec County
Hubbard County
Grant County
Cook County
Chippewa County
Becker County
Aitkin County
Ontonagon County
Montmorency County
Keweenaw County
Iron County
Chippewa County
Benzie County
Baraga County
Alger County
Alcona County
Nantucket County
Dukes County
Piscataquis County
Wolfe County
Trimble County
Trigg County
Rowan County
Robertson County
Owsley County
Owen County
Morgan County
Metcalfe County
Magoffin County
Livingston County
Lincoln County
Lee County
Knox County
Knott County


Hickman County
Harlan County
Garrard County
Gallatin County
Fulton County
Floyd County
Fleming County
Estill County
Elliott County
Clay County
Casey County
Carlisle County
Bell County
Wilson County
Wichita County
Washington County
Wallace County
Trego County
Thomas County
Stanton County
Smith County
Sherman County
Sheridan County
Russell County
Rush County
Rice County
Rawlins County
Pawnee County
Norton County
Ness County
Nemaha County
Meade County
Marshall County
Logan County
Lincoln County
Lane County
Kiowa County
Kingman County
Kearny County
Hodgeman County
Haskell County
Harper County
Greeley County
Gray County
Grant County
Graham County
Ellsworth County
Elk County
Edwards County
Dickinson County
Decatur County
Comanche County
Clark County


Cheyenne County
Chase County
Brown County
Allen County
Wayne County
Sac County
Ringgold County
Pocahontas County
Palo Alto County
Monroe County
Lucas County
Ida County
Humboldt County
Fremont County
Floyd County
Emmet County
Decatur County
Davis County
Cherokee County
Cass County
Calhoun County
Adams County
Pike County
Benton County
White County
Wayne County
Union County
Scott County
Putnam County
Pope County
Johnson County
Henderson County
Hardin County
Hamilton County
Edwards County
Edgar County
Clay County
Cass County
Brown County
Alexander County
Shoshone County
Oneida County
Lewis County
Lemhi County
Franklin County
Clearwater County
Clark County
Butte County
Boundary County
Boise County
Benewah County
Bear Lake County


Kalawao County
Taliaferro County
Montgomery County
Glascock County
Evans County
Sedgwick County
San Juan County
Kiowa County
Jackson County
Dolores County
Conejos County
Cheyenne County
Bent County
Trinity County
Sierra County
Modoc County
Mariposa County
Lassen County
Montgomery County
Little River County
Jackson County
Fulton County
Calhoun County
Yakutat City and Borough
Wrangell City and Borough
Valdez-Cordova Census Area
Skagway Municipality
Sitka City and Borough
Northwest Arctic Borough
North Slope Borough
Nome Census Area
Lake and Peninsula Borough
Kusilvak Census Area
Kodiak Island Borough
Hoonah-Angoon Census Area
Haines Borough
Dillingham Census Area
Denali Borough
Bristol Bay Borough
Aleutians West Census Area
Aleutians East Borough
Geneva County


−1
0
1
2
3
Difference

In [57]:

```
1 pred_set["DIFF"] = difference
2 pred_set
```

```
//anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:1: Setti
ngWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas
-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
 (https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
l#returning-a-view-versus-a-copy)
  """Entry point for launching an IPython kernel.
```

Out[57]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	AC
30	0500000US01061	Geneva County	Alabama	26491.0	6.3	5.3	5.4	
67	0500000US02013	Aleutians East Borough	Alaska	3425.0	5.0	6.9	8.1	
68	0500000US02016	Aleutians West Census Area	Alaska	5750.0	4.8	9.2	8.6	
71	0500000US02060	Bristol Bay Borough	Alaska	890.0	5.1	4.3	5.8	
72	0500000US02068	Denali Borough	Alaska	2232.0	5.6	13.4	5.5	
...	...	...	...	...	...	...	...	
3114	0500000US55123	Vernon County	Wisconsin	30516.0	6.6	4.7	4.6	
3117	0500000US55129	Washburn County	Wisconsin	15689.0	5.0	4.0	3.9	
3125	0500000US56003	Big Horn County	Wyoming	11901.0	7.0	4.5	5.0	
3139	0500000US56031	Platte County	Wyoming	8673.0	5.7	4.9	4.3	
3146	0500000US56045	Weston County	Wyoming	7100.0	4.5	6.7	5.1	

529 rows × 104 columns

In [71]:

```
# 0.3% is the current national infection rate
counties_bigDiff = pred_set.loc[abs(pred_set["DIFF"]) >= 0.3]
counties_bigDiff.sort_values(by=['DIFF'], ascending=False)
```

Out[71]:

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	A
1851	0500000US36047	Kings County	New York	2600747.0	5.4	6.6	9.5	
1830	0500000US36005	Bronx County	New York	1437872.0	6.8	7.9	8.5	
1868	0500000US36081	Queens County	New York	2298513.0	5.1	6.2	8.3	
81	0500000US02158	Kusilvak Census Area	Alaska	8198.0	9.1	8.0	8.8	
86	0500000US02188	Northwest Arctic Borough	Alaska	7734.0	8.1	7.7	9.0	
82	0500000US02164	Lake and Peninsula Borough	Alaska	1375.0	6.7	9.6	7.6	
1607	0500000US30019	Daniels County	Montana	1753.0	5.4	2.1	4.5	
981	0500000US20187	Stanton County	Kansas	2063.0	8.0	3.8	5.5	
2609	0500000US48173	Glasscock County	Texas	1430.0	9.2	11.0	2.1	
548	0500000US15005	Kalawao County	Hawaii	75.0	2.7	0.0	2.7	
1711	0500000US31115	Loup County	Nebraska	585.0	3.6	5.8	5.8	
67	0500000US02013	Aleutians East Borough	Alaska	3425.0	5.0	6.9	8.1	
1610	0500000US30025	Fallon County	Montana	2838.0	4.8	5.9	7.6	
1656	0500000US31005	Arthur County	Nebraska	418.0	4.8	6.2	2.6	
2369	0500000US46017	Buffalo County	South Dakota	2053.0	10.8	8.0	5.9	
1721	0500000US31135	Perkins County	Nebraska	2907.0	5.3	3.5	5.9	
1657	0500000US31007	Banner County	Nebraska	696.0	4.3	3.3	3.4	
1662	0500000US31017	Brown County	Nebraska	2988.0	3.8	3.6	4.4	

	GEO_ID	COUNTY	STATE	POPULATION	AGE15_19	AGE20_24	AGE25_29	A
1661	0500000US31015	Boyd County	Nebraska	2042.0	5.3	3.3	3.5	
2740	0500000US48435	Sutton County	Texas	3865.0	4.3	6.0	7.9	
2143	0500000US40025	Cimarron County	Oklahoma	2189.0	4.3	3.9	8.0	
68	0500000US02016	Aleutians West Census Area	Alaska	5750.0	4.8	9.2	8.6	
73	0500000US02070	Dillingham Census Area	Alaska	4975.0	7.6	7.9	7.9	
929	0500000US20083	Hodgeman County	Kansas	1842.0	5.6	6.2	6.1	
1741	0500000US31175	Valley County	Nebraska	4224.0	5.6	5.0	5.1	
300	0500000US08111	San Juan County	Colorado	544.0	1.8	2.0	11.0	

26 rows × 104 columns

In [ ]:

```
1
```