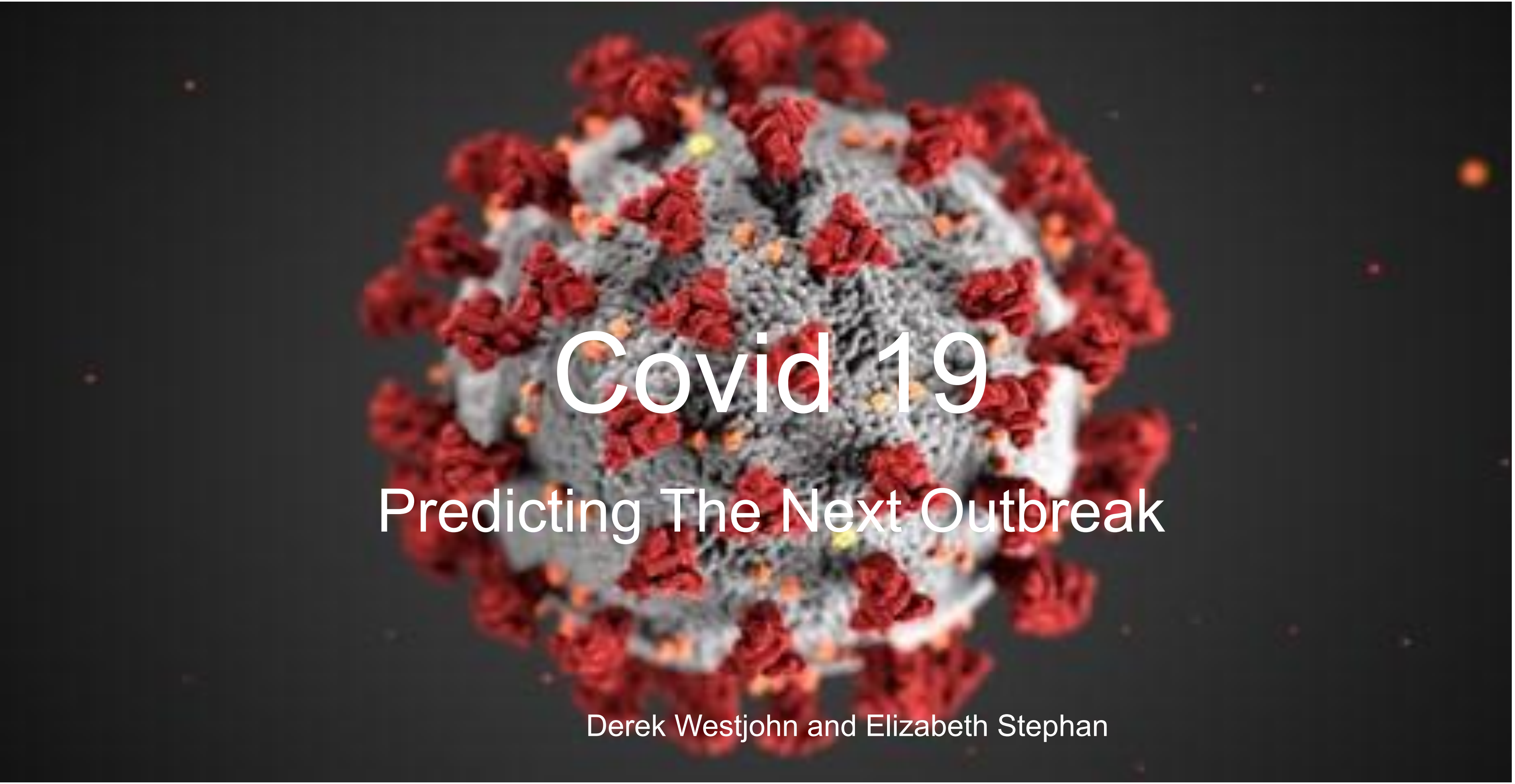


Covid 19

Predicting The Next Outbreak

Derek Westjohn and Elizabeth Stephan

Why This Topic

- World's deadliest viral pandemic since the Spanish Flu
- Some areas seemed to be more devastated by the outbreak than others
- Understand factors that contributed to the unequal devastation

Questions to Answer

- Is there an underlying correlation between the areas that with hit harder with Covid.
- Can we use this information to predict future outbreaks and subsequently can the information be used to prevent future outbreaks

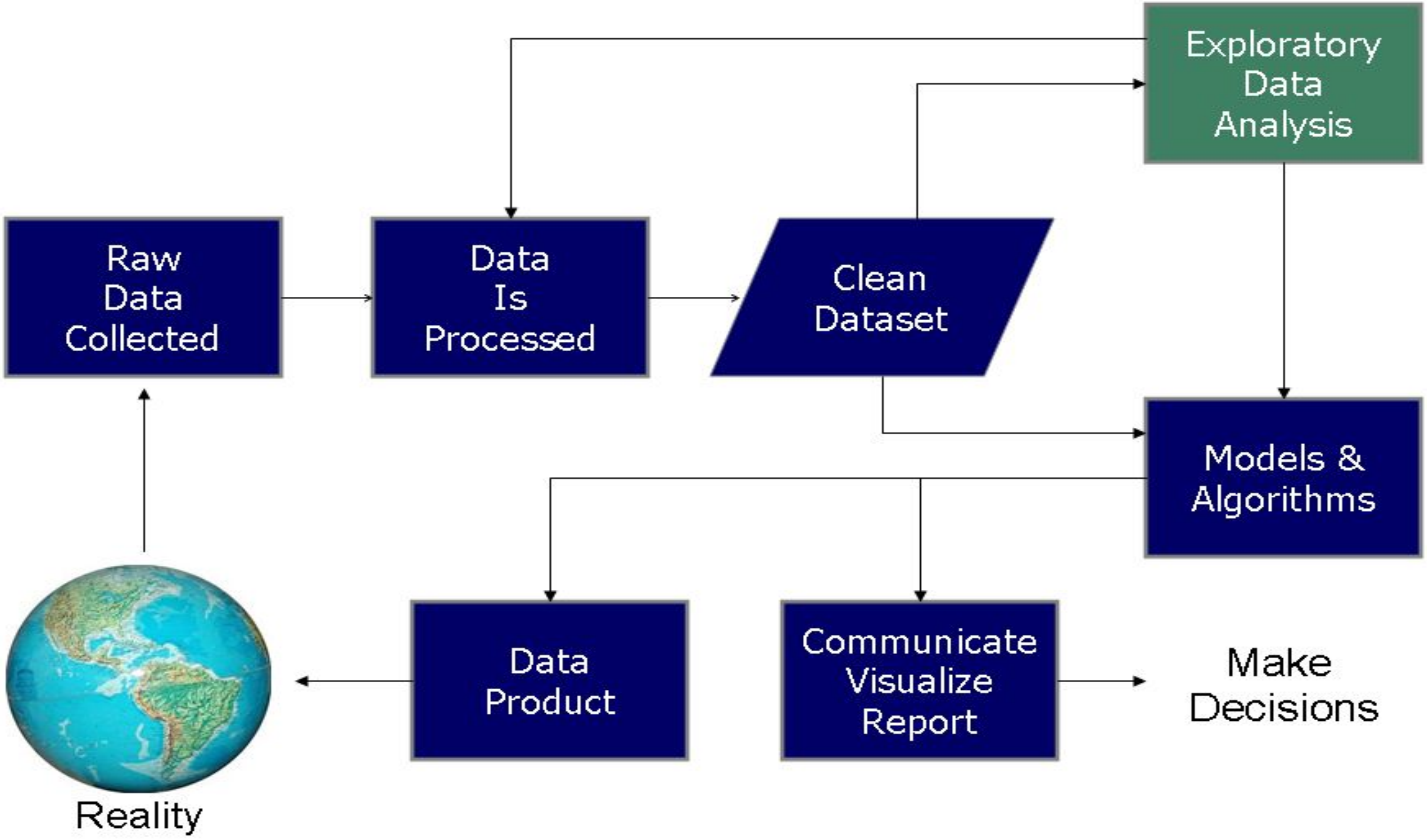
Data Sources

NYT - county level Covid19 case and fatality data (github)

USDA Economic Research SErvice

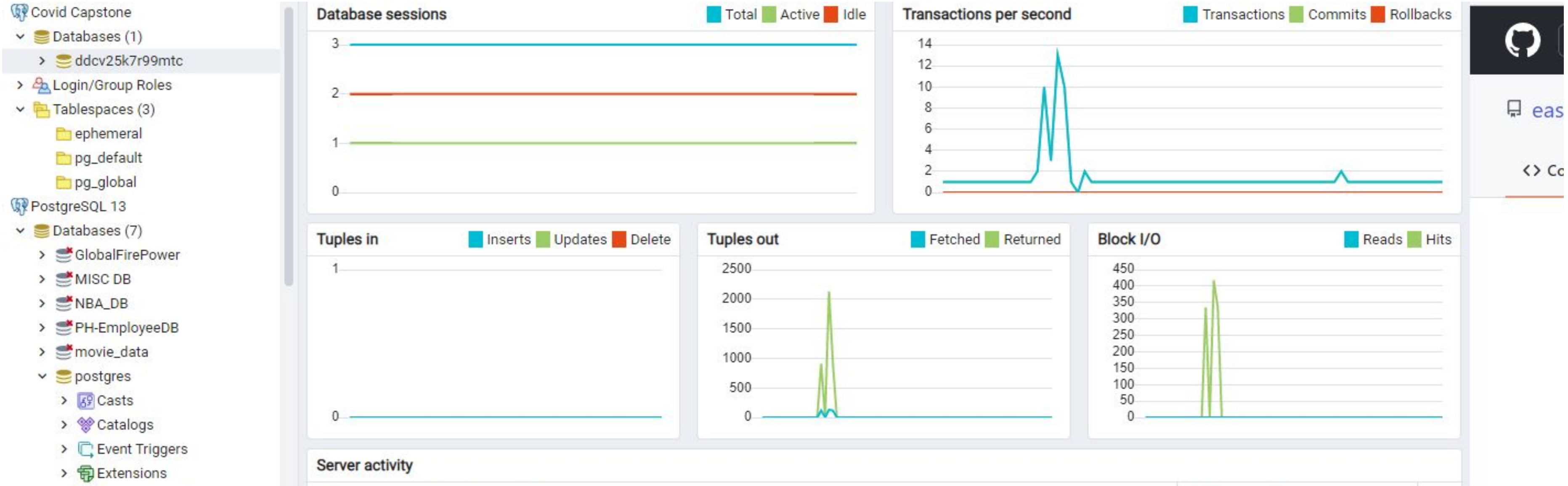
- Unemployment and median household income for US, States and Counties 2000-20
- Education attainment for US, States, Counties 1970-2019
- Population estimates for the US, States and Counties 2010-2020

Data Science Process



Data Exploration and Analysis

- Raw data was taken from NYT Covid 19 fatality data set and also key socio-economic factors from the USDA.
- Raw data was stored in Haroku database and linked to PostgreSQL - where tables were created.
- The tables were cleaned, null values were investigated and either were changed to 0 or were dropped.
- Dataframes were created from the tables and merged on “FIPS” location code,




```
In [1]: #Import dependencies
        from path import Path
        import pandas as pd
```

```
In [ ]: #Connect Python to Postgres
        import psycopg2
        engine=psycop2.connect(
                database ="ddcv25k7r99mtc",
                user= "lkoibcunpocdfy",
                password = "xxxxx",
                host ="ec2-34-199-200-115.compute-1.amazonaws.com",
                port ="5432
        )
```

In [4]:
```
deaths_df= pd.read_csv("covid deaths.csv")
deaths_df.head()
```

Out[4]:

	fips	state	county	number_of_deaths
0	1001	Alabama	Autauga	791.0
1	1003	Alabama	Baldwin	2967.0
2	1005	Alabama	Barbour	472.0
3	1007	Alabama	Bibb	471.0
4	1009	Alabama	Blount	1085.0

In [7]:
```
locations_df=pd.read_csv("locations.csv")
locations_df.head()
```

Out[7]:

	fips	state	county	latitude	longitude	total_population	area_per_square_mile	population_density_per_square_mile
0	1001	Alabama	Autauga	32.534928	-86.642748	55049	594.446120	92.605533
1	1003	Alabama	Baldwin	30.727489	-87.722575	199510	1589.807425	125.493187
2	1005	Alabama	Barbour	31.869589	-85.393213	26614	884.875776	30.076538
3	1007	Alabama	Bibb	32.998634	-87.126480	22572	622.582355	36.255444
4	1009	Alabama	Blount	33.980878	-86.567383	57704	644.806508	89.490412

In [8]:
```
education_df=pd.read_csv("Education.csv",dtype=str)
education_df.head()
```

Out[8]:

	FIPS Code	State	Area name	Less than a high school diploma, 2015-19	High school diploma only, 2015-19	Some college or associate's degree, 2015-19	Bachelor's degree or higher, 2015-19
0	0	US	United States	26,472,261	59,472,748	63,756,905	70,920,162
1	1000	AL	Alabama	458922	1022839	993344	845772

In [9]:
```
covid_depend_df=pd.read_csv("covid dependents.csv")
covid_depend_df.head()
```

Out[9]:

	fips	state	county	minority_percentage	limited_english	multi_unit_housing	mobile_homes	overcrowding_rank	no_vehicle_household	institutionalized_
0	1001	Alabama	Autauga	0.6339	0.5355	0.6791	0.7268	0.2477	0.3298	
1	1003	Alabama	Baldwin	0.5253	0.5282	0.9733	0.5387	0.2639	0.0872	
2	1005	Alabama	Barbour	0.9042	0.6979	0.2814	0.9370	0.4438	0.8816	
3	1007	Alabama	Bibb	0.6450	0.3553	0.4072	0.9249	0.0248	0.5645	
4	1009	Alabama	Blount	0.4238	0.7482	0.1344	0.8465	0.5056	0.1907	

In [37]:
```python
covid_complete_df=reduce(lambda left,right:pd.merge(left,right, on=['fips'], how='outer'), data_frames)
```

In [38]:
```python
covid_complete_df.head()
```

Out[38]:

	fips	state	county	number_of_deaths	latitude	longitude	total_population	area_per_square_mile	population_density_per_square_mile	minority_perc
0	1001	Alabama	Autauga	791.0	32.534928	-86.642748	55049.0	594.446120	92.605533	
1	1003	Alabama	Baldwin	2967.0	30.727489	-87.722575	199510.0	1589.807425	125.493187	
2	1005	Alabama	Barbour	472.0	31.869589	-85.393213	26614.0	884.875776	30.076538	
3	1007	Alabama	Bibb	471.0	32.998634	-87.126480	22572.0	622.582355	36.255444	
4	1009	Alabama	Blount	1085.0	33.980878	-86.567383	57704.0	644.806508	89.490412	

5 rows × 22 columns

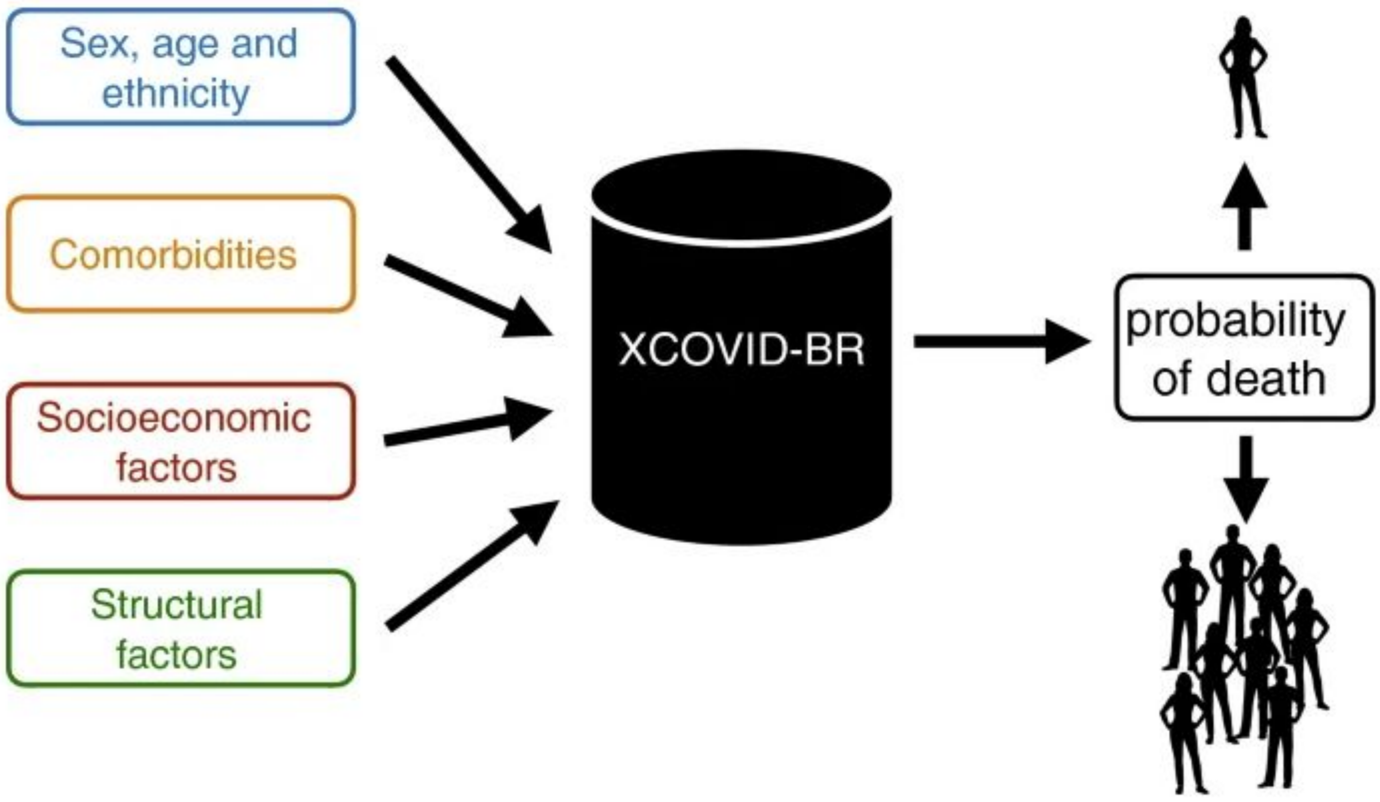
In [41]:
```python
#Find null values
for column in covid_complete_df.columns:
    print(f"Column {column} has {covid_complete_df[column].isnull().sum()} null values")
```

```
Column fips has 0 null values
Column state has 140 null values
Column county has 140 null values
Column number_of_deaths has 433 null values
Column latitude has 140 null values
Column longitude has 140 null values
Column total_population has 140 null values
Column area_per_square_mile has 140 null values
Column population_density_per_square_mile has 140 null values
Column minority_percentage has 140 null values
Column limited_english has 140 null values
Column multi_unit_housing has 140 null values
Column mobile_homes has 140 null values
Column overcrowding_rank has 140 null values
Column no_vehicle_household has 140 null values
Column institutionalized_ranker has 140 null values
Column housing_and_transportation has 140 null values
Column social_vulnerability has 140 null values
Column Less than a high school diploma, 2015-19 has 10 null values
Column High school diploma only, 2015-19 has 10 null values
Column Some college or associate's degree, 2015-19 has 10 null values
Column Bachelor's degree or higher, 2015-19 has 10 null values
```

In [42]:
```python
covid_complete_df["number_of_deaths"].fillna(0,inplace=True)
```

In [43]:
```python
#Find null values
for column in covid_complete_df.columns:
    print(f"Column {column} has {covid_complete_df[column].isnull().sum()} null values")
```

Figure 1


Sex, age and ethnicity
Comorbidities
Socioeconomic factors
Structural factors
XCOVID-BR
probability of death

In [46]:
```
#To create a correlation matrix to see what columns (variables) could present a good variables to test against
corr=covid_complete_df.corr()
corr
```

Out[46]:

	fips	number_of_deaths	latitude	longitude	total_population	area_per_square_mile	population_density_per_square_m
fips	1.000000	-0.058086	0.060077	0.121840	-0.055953	-0.092250	0.0237
number_of_deaths	-0.058086	1.000000	-0.083293	0.036395	0.973784	0.025095	0.3112
latitude	0.060077	-0.083293	1.000000	-0.293782	-0.057794	0.265048	0.0072
longitude	0.121840	0.036395	-0.293782	1.000000	0.002150	-0.377956	0.1034
total_population	-0.055953	0.973784	-0.057794	0.002150	1.000000	0.026403	0.3332
area_per_square_mile	-0.092250	0.025095	0.265048	-0.377956	0.026403	1.000000	-0.0307
population_density_per_square_mile	0.023786	0.311260	0.007280	0.103499	0.333226	-0.030771	1.0000
minority_percentage	-0.057613	0.241787	-0.437147	-0.143856	0.226665	0.111257	0.1345
limited_english	0.020930	0.276481	-0.248990	-0.151361	0.271044	0.050584	0.1481
multi_unit_housing	0.004688	0.363160	0.211813	-0.001236	0.341998	0.004216	0.1918
mobile_homes	0.000314	-0.271149	-0.410686	0.040638	-0.275804	-0.021542	-0.1839
overcrowding_rank	-0.054611	0.134660	-0.237876	-0.277427	0.142246	0.141417	0.0877
no_vehicle_household	-0.065233	0.155547	-0.103838	0.236776	0.117206	0.026210	0.1248
institutionalized_ranker	-0.015489	-0.017983	0.040895	0.022841	-0.025154	0.052320	0.0062
housing_and_transportation	-0.054778	0.162899	-0.204223	0.001801	0.135399	0.084906	0.0975
social_vulnerability	-0.092763	0.072261	-0.486962	0.027865	0.033129	0.060175	0.0215
Less than a high school diploma, 2015-19	-0.056385	0.906951	-0.081702	-0.020504	0.950822	0.034522	0.3095

```
In [31]: data_df.corr()['number_of_deaths']

Out[31]: number_of_deaths                                      1.000000
         total_population                                      0.973784
         area_per_square_mile                                  0.025095
         population_density_per_square_mile                    0.311260
         minority_percentage                                   0.241787
         limited_english                                       0.276481
         multi_unit_housing                                    0.363160
         mobile_homes                                         -0.271149
         overcrowding_rank                                     0.134660
         no_vehicle_household                                  0.155547
         institutionalized_ranker                             -0.017983
         housing_and_transportation                            0.162899
         social_vulnerability                                  0.072261
         Less than a high school diploma, 2015-19              0.906951
         High school diploma only, 2015-19                     0.986901
         Some college or associate's degree, 2015-19           0.977690
         Bachelor's degree or higher, 2015-19                  0.922002
         Name: number_of_deaths, dtype: float64
```

Unsupervised Learning

Unsupervised Learning

```
In [57]: from sklearn.preprocessing import StandardScaler
         data_scaler=StandardScaler()

In [58]: covid_data_scaled=data_scaler.fit_transform(covid_complete_update_df)

In [59]: covid_data_scaled[:5]

Out[59]: array([[-0.13052159, -0.1425129 , -0.09752616,  0.46369547,  0.12595086,
                  0.62016999, -0.56785283, -0.73379171, -0.11289334, -0.12503346,
                 -0.15894355, -0.15396663],
                [ 0.54685283,  0.30177745, -0.07923043,  0.08761386,  0.10083111,
                  1.63729524, -1.40505023, -0.58038006,  0.14930449,  0.44836329,
                  0.44231868,  0.31147732],
                [-0.22982419, -0.22996485, -0.13231164,  1.39974389,  0.68477931,
                 -0.75478144,  1.33637438,  1.49327456, -0.09866661, -0.24570832,
                 -0.25599008, -0.24955441],
                [-0.23013548, -0.24239603, -0.12887426,  0.50213475, -0.49412849,
                 -0.31985841,  0.24208215,  0.57592134, -0.1376058 , -0.22884717,
                 -0.26956348, -0.25445011],
                [-0.03900134, -0.13434743, -0.09925913, -0.26388047,  0.85786473,
                 -1.26299834, -1.04787821, -1.09879371, -0.01808488, -0.11036818,
                 -0.11102683, -0.21143621]])
```

In [62]:
```python
# Initialize PCA model
pca = PCA(n_components=2)
```

In [63]:
```python
# Get two principal components for the covid data.
covid_pca = pca.fit_transform(covid_data_scaled)
```

In [64]:
```python
#Transform PCA data to a Dataframe
covid_pca_df=pd.DataFrame(data=covid_pca, columns=["principal component 1","principal component 2"])
covid_pca_df.head()
```

Out[64]:

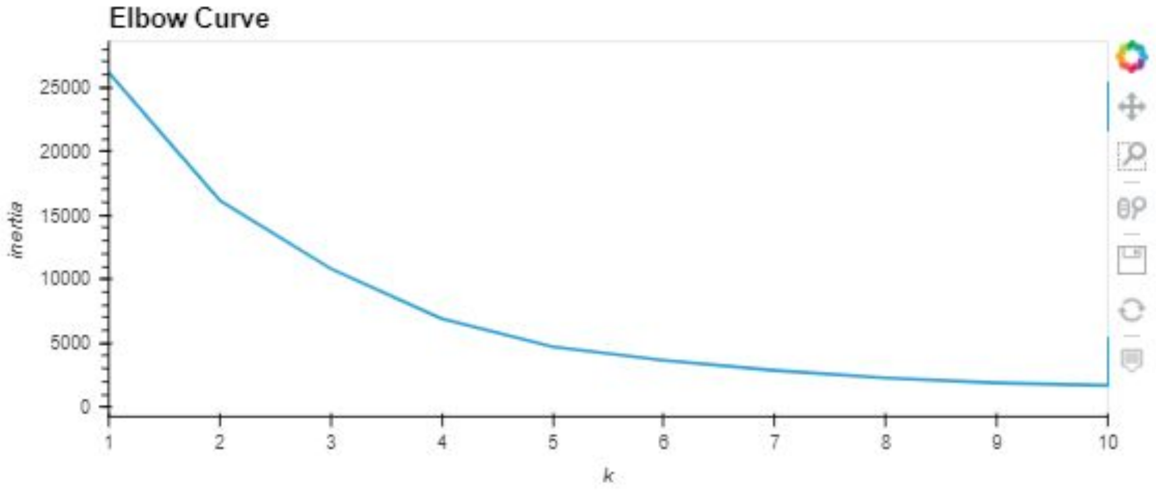
	principal component 1	principal component 2
0	-0.267109	0.246536
1	0.967681	0.882183
2	-0.094863	-2.392524
3	-0.526244	-0.617686
4	-0.581242	1.151701

In [66]:
```python
# Get the explained variance
pca.explained_variance_ratio_
```

Out[66]:
```
array([0.52307586, 0.17062066])
```

In [67]:
```python
# Find the best value for K
inertia = []
k = list(range(1, 11))

# Calculate the inertia for the range of K values
for i in k:
    km = KMeans(n_clusters=i, random_state=0)
    km.fit(covid_pca_df)
    inertia.append(km.inertia_)

# Create the elbow curve
elbow_data = {"k": k, "inertia": inertia}
df_elbow = pd.DataFrame(elbow_data)
df_elbow.hvplot.line(x="k", y="inertia", xticks=k, title="Elbow Curve")
```

Out[67]:



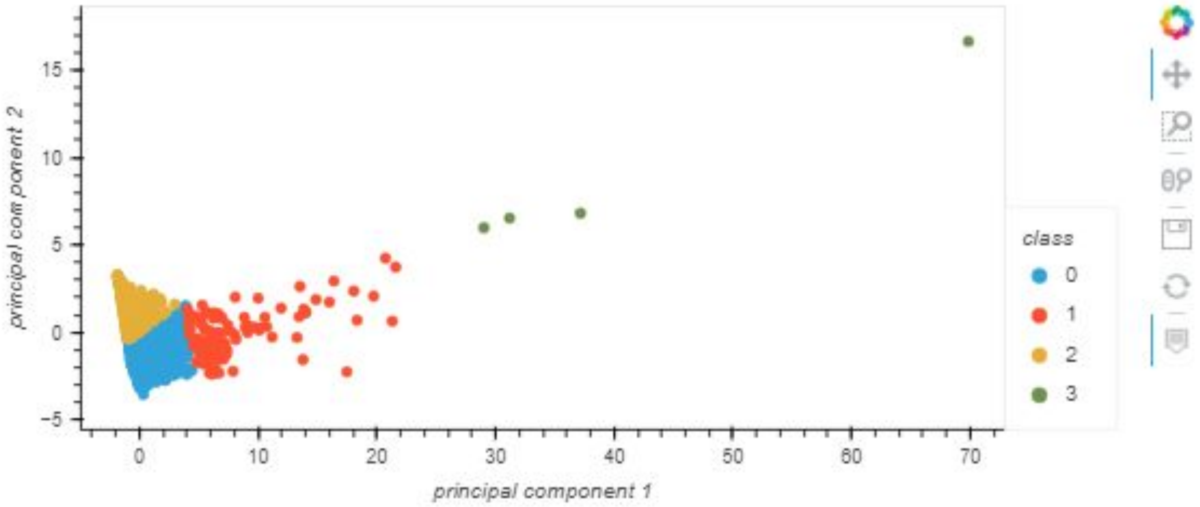
In [68]:
```
# Initialize the K-means model
model = KMeans(n_clusters=4, random_state=0)

# Fit the model
model.fit(covid_pca_df)

# Predict clusters
predictions = model.predict(covid_pca_df)

# Add the predicted class columns
covid_pca_df["class"] = model.labels_
covid_pca_df.head()
```

Out[68]:

	principal component 1	principal component 2	class
0	-0.267109	0.246536	2
1	0.967681	0.882183	2
2	-0.094863	-2.392524	0
3	-0.526244	-0.617686	0
4	-0.581242	1.151701	2

In [70]:
```
covid_pca_df.hvplot.scatter(
    x="principal component 1",
    y="principal component 2",
    hover_cols=["class"],
    by="class",
)
```

Out[70]:



```
In [86]: covid_pca_complete_df.groupby(["class"])[['number_of_deaths']].agg(['sum'])
```

Out[86]:

	number_of_deaths
	sum
class	
0	1665359.0
1	1193805.0
2	713642.0
3	229919.0

		number_of_deaths
		sum
state	class	
Alabama	0	47568.0
	1	11214.0
	2	23009.0
Alaska	0	6954.0
	2	1133.0
...	...	...
Wisconsin	0	23122.0
	1	12467.0
	2	27981.0
Wyoming	0	1021.0
	2	5905.0

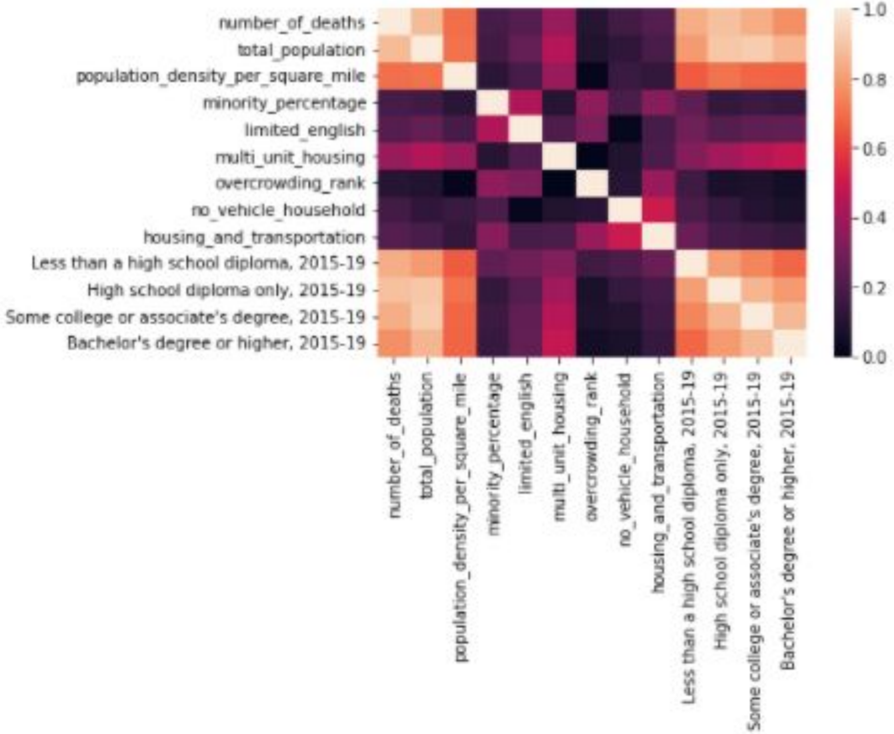
135 rows × 1 columns

Supervised Learning

```
In [37]: # Create our features
         X = data_df.drop(columns=['number_of_deaths'], axis=1)
         # Create our target
         y = data_df["number_of_deaths"]

In [38]: # Scale the data
         Scaler = StandardScaler().fit(X)
         X = Scaler.transform(X)

In [40]:
         from sklearn.model_selection import train_test_split
         X_train,X_test,y_train,y_test = train_test_split(X,y, test_size=0.2, random_state = 0)

In [41]: from sklearn.linear_model import LinearRegression
         from sklearn.metrics import r2_score
         regressor = LinearRegression()
         regressor.fit(X_train,y_train)

Out[41]: LinearRegression()

In [43]: y_pred = regressor.predict(X_test)
         print(r2_score(y_test,y_pred))

         0.9731904050032136
```

```
In [45]: print('Train Score: ', regressor.score(X_train, y_train))
         print('Test Score: ', regressor.score(X_test, y_test))

         Train Score:  0.9823416444527159
         Test Score:  0.9731904050032136
```

Out[47]: <AxesSubplot:>



Results

- This model confirmed a lot of what is already known about COVID 19 and it's ultimate fatality
    - Areas with larger populations are prone to have higher death rates
    - Areas with larger populations of minorities are more likely to have higher death rates
    - Populations that have limited english have higher death rates
    - Multi-unit housing also is more inclined to have higher death rates
    - Alternatively the level of education (higher education) leads to lower death rates