Harshal Talele & Harshalkumar Loya

In [ ]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from sklearn.metrics import r2_score
```

In [ ]:
```
train_data = pd.read_csv('training_data.csv')
test_data = pd.read_csv('test_data.csv')
```

In [ ]:
```
train_data['date_int'] = train_data.date_index_converted.str.replace('day_','').astype
test_data['date_int'] = test_data.date_index_converted.str.replace('day_','').astype('
```

In [ ]:
```
county_cols = ['date_int', 'county', 'cases', 'deaths', 'date_index_converted', 'count
```

In [ ]:
```
corr = train_data[county_cols].corr()
corr.style.background_gradient(cmap='coolwarm').set_precision(2)
```

```
C:\Users\harsh\AppData\Local\Temp\ipykernel_3148\1680320004.py:2: FutureWarning: this
method is deprecated in favour of `Styler.format(precision=..)`
  corr.style.background_gradient(cmap='coolwarm').set_precision(2)
```

Out[ ]:

	date_int	cases	deaths	county_data_length	total_pop	percent_25_34	pe
date_int	1.00	0.26	0.29	0.01	-0.01	0.01	
cases	0.26	1.00	0.77	0.24	0.34	0.24	
deaths	0.29	0.77	1.00	0.12	0.28	0.15	
county_data_length	0.01	0.24	0.12	1.00	0.89	0.59	
total_pop	-0.01	0.34	0.28	0.89	1.00	0.63	
percent_25_34	0.01	0.24	0.15	0.59	0.63	1.00	
percent_highschool	-0.03	0.05	0.06	0.09	0.17	-0.05	
labor_force_rate	-0.01	0.07	0.09	0.20	0.29	0.01	
unemployment_rate	0.01	0.06	0.07	0.16	0.14	0.18	
median_housing_cost	-0.03	0.10	0.07	0.20	0.35	0.11	
median_household_earnings	-0.03	0.03	0.01	-0.00	0.11	-0.10	
median_worker_earnings	-0.03	0.06	0.04	0.11	0.22	-0.03	
percent_insured	-0.02	0.02	0.03	0.02	0.07	-0.02	
percent_married	-0.02	-0.19	-0.18	-0.44	-0.49	-0.55	
poverty_rate	0.01	0.03	0.02	0.12	0.06	0.19	
median_property_value	-0.02	0.06	0.04	0.09	0.21	-0.02	
percent_white	0.01	-0.31	-0.27	-0.75	-0.88	-0.66	

In [ ]:
```
len(county_cols)
```

Out[ ]:
```
19
```

In [ ]:
```
county_cols_to_use = ['date_int', 'total_pop', 'deaths', 'county_data_length', 'percer
len(county_cols_to_use)
```

Out[ ]:
```
8
```

In [ ]:
```
social_features_string = 'core_cosine, core_cosine_normalized, core_intersection, core

social_features_string = social_features_string.replace(" ","")
social_features = social_features_string.split(",")
social_features = [each for each in social_features_string.split(",")]

for i in range(len(social_features)):
    if social_features[i] == 'median_worker_earning':
        social_features[i] = 'median_worker_earnings'
    elif social_features[i] == 'percent_insure':
        social_features[i] = 'percent_insured'
    else:
        pass
len(social_features)
```

Out[ ]:
```
137
```

In [ ]:
```
awareness_cols_to_use = ['core_intersection','health_technology_cosine_normalized', 'p
```

In [ ]:
```
cols_to_use = county_cols_to_use + awareness_cols_to_use
```

In [ ]:
```
X = train_data[cols_to_use]
y = train_data['cases']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_stat
```

In [ ]:
```
from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor()
rf.fit(X_train, y_train)

y_pred_rf = rf.predict(X_test)

print('Random Forest Regression -',r2_score(y_test, y_pred_rf))
```

```
Random Forest Regression - 0.9032000839831487
```

In [ ]:
```
from sklearn.ensemble import GradientBoostingRegressor

gbr = GradientBoostingRegressor()
gbr.fit(X_train, y_train)

y_pred_gbr = gbr.predict(X_test)

print('Gradient Boosting Regression -',r2_score(y_test, y_pred_gbr))
```

```
Gradient Boosting Regression - 0.9113667056127134
```

In [ ]:
```
import xgboost as xg

xgb = xg.XGBRegressor()

xgb.fit(X_train, y_train)

y_pred_xgb = xgb.predict(X_test)

print('XGBoost Regression -',r2_score(y_test, y_pred_xgb))
```

```
XGBoost Regression - 0.9127768979626822
```

In [ ]:
```
from sklearn.ensemble import ExtraTreesRegressor

etr = ExtraTreesRegressor()

etr.fit(X_train, y_train)

y_pred_etr = etr.predict(X_test)

print('Extra Trees Regression -',r2_score(y_test, y_pred_etr))
```

```
Extra Trees Regression - 0.9755185854707977
```

In [ ]:
```
gbr.fit(X,y)
xgb.fit(X,y)
etr.fit(X,y)
```

Out[ ]:
```
ExtraTreesRegressor()
```

In [ ]:
```
# gbr.fit(train_data[cols_to_use],y)
xgb.fit(train_data[cols_to_use],y)
# etr.fit(train_data[cols_to_use],y)

# test_pred = gbr.predict(test_data[cols_to_use])
test_pred = xgb.predict(test_data[cols_to_use])
# test_pred = etr.predict(test_data[cols_to_use])


out = pd.DataFrame(test_pred, columns=['Cases'])

for i in range(len(out)):
    if out['Cases'][i] < 0:
        out['Cases'][i] = 0

out['Index'] = out.index

out['Cases'] = np.floor(pd.to_numeric(out['Cases'], errors='coerce')).astype('Int64')

out = out[['Index', 'Cases']]
out
```

Out[ ]:

	Index	Cases
0	0	1
1	1	428
2	2	0
3	3	0
4	4	0
...	...	...
7326	7326	0
7327	7327	103
7328	7328	0
7329	7329	517
7330	7330	0

7331 rows × 2 columns

In [ ]:
```
for i in range(len(xgb.feature_names_in_)):
    print(xgb.feature_names_in_[i], "       ",xgb.feature_importances_[i])
```

```
date_int          0.0019541495
total_pop          0.033885073
deaths          0.5588625
county_data_length          0.028772343
percent_25_34          0.28685516
labor_force_rate          0.009185454
unemployment_rate          0.0651842
median_household_earnings          0.008705765
core_intersection          0.0009018636
health_technology_cosine_normalized          0.0010103182
politics_democratic_hate_intersection          0.004145311
race_cosine          0.00053787866
```

In [ ]:
```
print('Zeros', list(out['Cases']).count(0))
print('Avg', out['Cases'].mean())
print('Std Dev', out['Cases'].std())
print(out['Cases'][34])
```

```
Zeros 4376
Avg 178.00968489974082
Std Dev 950.2225028665334
12555
```

In [ ]:
```
out.to_csv('submission.csv', index=False)
```

In [ ]:
```
xgb.fit(train_data[social_features],y)
```

Out[ ]:
```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=None, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=None, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             n_estimators=100, n_jobs=None, num_parallel_tree=None,
             predictor=None, random_state=None, ...)
```

In [ ]:
```
for i in range(len(xgb.feature_names_in_)):
    print(xgb.feature_names_in_[i], "       ",xgb.feature_importances_[i])
```

core_cosine 0.024615863
core_cosine_normalized 0.0
core_intersection 0.09556902
core_intersection_normalized 0.0
core_jaccard 0.022157174
core_jaccard_normalized 0.0
domestic_cosine 0.0021220807
domestic_cosine_normalized 0.0
domestic_intersection 0.00500702
domestic_intersection_normalized 0.0
domestic_jaccard 0.00028898945
domestic_jaccard_normalized 0.0
economy_cosine 0.00994675
economy_cosine_normalized 0.0
economy_intersection 0.0029225317
economy_intersection_normalized 0.0
economy_jaccard 0.025123214
economy_jaccard_normalized 0.0
education_cosine 0.0062960936
education_cosine_normalized 0.0
education_intersection 0.0002282981
education_intersection_normalized 0.0
education_jaccard 0.005131746
education_jaccard_normalized 0.0
entertainment_cosine 0.030499658
entertainment_cosine_normalized 0.0
entertainment_intersection 0.0019923614
entertainment_intersection_normalized 0.0
entertainment_jaccard 0.00063093565
entertainment_jaccard_normalized 0.0
foreign_cosine 0.0068207425
foreign_cosine_normalized 0.0
foreign_intersection 0.0020403892
foreign_intersection_normalized 0.0
foreign_jaccard 0.0041365153
foreign_jaccard_normalized 0.0
gender_cosine 0.009844095
gender_cosine_normalized 0.0
gender_intersection 0.0030140837
gender_intersection_normalized 0.0
gender_jaccard 0.0076313647
gender_jaccard_normalized 0.0
health_cosine 0.00071302574
health_cosine_normalized 0.0
health_intersection 0.0026228642
health_intersection_normalized 0.0
health_jaccard 0.0
health_jaccard_normalized 0.0
health_technology_cosine 0.00044385757
health_technology_cosine_normalized 0.0
health_technology_intersection 0.0034077147
health_technology_intersection_normalized 0.0
health_technology_jaccard 0.0024322523
health_technology_jaccard_normalized 0.0
ideology_cosine 0.00026810556
ideology_cosine_normalized 0.0
ideology_intersection 4.274941e-05
ideology_intersection_normalized 0.0
ideology_jaccard 0.0
ideology_jaccard_normalized 0.0

```
illness_cosine         0.006805097
illness_cosine_normalized         0.0
illness_intersection         0.024313843
illness_intersection_normalized         0.0
illness_jaccard         0.0015105379
illness_jaccard_normalized         0.0
labor_force_rate         0.015310176
median_household_earnings         0.0010157936
median_housing_cost         0.0042637493
median_property_value         0.0035040753
median_worker_earnings         0.0016336194
nationalistic_cosine         0.00060253905
nationalistic_cosine_normalized         0.0
nationalistic_intersection         0.00092368975
nationalistic_intersection_normalized         0.0
nationalistic_jaccard         6.15813e-05
nationalistic_jaccard_normalized         0.0
percent_25_34         0.061960787
percent_highschool         0.0008704953
percent_insured         0.0022402084
percent_married         0.004328926
percent_white         0.06454977
politics_cosine         0.008607632
politics_cosine_normalized         0.0
politics_democratic_hate_cosine         0.003822595
politics_democratic_hate_cosine_normalized         0.0
politics_democratic_hate_intersection         0.26147783
politics_democratic_hate_intersection_normalized         0.0
politics_democratic_hate_jaccard         0.00068259286
politics_democratic_hate_jaccard_normalized         0.0
politics_democratic_love_cosine         0.008728886
politics_democratic_love_cosine_normalized         0.0
politics_democratic_love_intersection         0.030199183
politics_democratic_love_intersection_normalized         0.0
politics_democratic_love_jaccard         0.00035059484
politics_democratic_love_jaccard_normalized         0.0
politics_intersection         0.0012166788
politics_intersection_normalized         0.0
politics_jaccard         0.00058649445
politics_jaccard_normalized         0.0
politics_republican_hate_cosine         0.00273341
politics_republican_hate_cosine_normalized         0.0
politics_republican_hate_intersection         0.00048914045
politics_republican_hate_intersection_normalized         0.0
politics_republican_hate_jaccard         0.001410888
politics_republican_hate_jaccard_normalized         0.0
politics_republican_love_cosine         0.0033076252
politics_republican_love_cosine_normalized         0.0
politics_republican_love_intersection         0.031491783
politics_republican_love_intersection_normalized         0.0
politics_republican_love_jaccard         0.01820632
politics_republican_love_jaccard_normalized         0.0
poverty_rate         0.0007284084
race_cosine         0.092853084
race_cosine_normalized         0.0
race_intersection         0.0009174826
race_intersection_normalized         0.0
race_jaccard         0.00020526536
race_jaccard_normalized         0.0
religion_cosine         0.0024786734
```

```
religion_cosine_normalized         0.0
religion_intersection         0.0044941716
religion_intersection_normalized         0.0
religion_jaccard         0.0022873005
religion_jaccard_normalized         0.0
social_cosine         0.007970839
social_cosine_normalized         0.0
social_intersection         0.024642438
social_intersection_normalized         0.0
social_jaccard         0.0012810383
social_jaccard_normalized         0.0
sports_cosine         0.01622521
sports_cosine_normalized         0.0
sports_intersection         0.001685279
sports_intersection_normalized         0.0
sports_jaccard         0.0010748423
sports_jaccard_normalized         0.0
```

core_intersection 0.09556902 core_jaccard 0.022157174 labor_force_rate 0.015310176 politics_democratic_hate_intersection 0.26147783 politics_democratic_love_intersection 0.030199183 politics_republican_love_intersection 0.031491783 race_cosine 0.092853084