Analyzing Correlation between Well-Being and Health Loss

by Nitesh Surtani

Dataset

Well-Being OECD data

	Country	INDICATOR	Year	Value
0	Australia	HO_BASE	2013	1.2
1	Austria	HO_BASE	2013	1.2
2	Belgium	HO_BASE	2013	1.4
3	Canada	HO_BASE	2013	0.2
4	Czech Republic	HO_BASE	2013	0.7

Health Loss GBD data

	Country	Deaths	Year
23	China	690.013983	2013
24	China	702.629150	2014
25	China	702.004329	2015
26	China	707.098323	2016
50	Indonesia	606.007202	2013

Data Cleaning

1. Removed duplicates from OECD data and average multiple values for same year.
2. Removed columns with Null values, as they couldn't be imputed.
3. These two datasets have only 8 countries in common.
4. Renamed columns for merging the two datasets.

	Country	INDICATOR	Year	Value	Deaths
0	Australia	HO_BASE	2013	1.2	658.103782
1	Australia	HO_HISH	2013	19.0	658.103782
2	Australia	HO_NUMR	2013	2.3	658.103782
3	Australia	IW_HADI	2013	28884.0	658.103782
4	Australia	IW_HADI	2013	58409.0	658.103782

Merged data

```
{'Australia',
 'Brazil',
 'Germany',
 'Japan',
 'Mexico',
 'Poland',
 'South Africa',
 'United States'}
```

Common Countries

Data Split

1. Training Data: 7 countries, Year: 2013, 2014, 2015

   Rows: 7 X 3 = 21, Columns = 21 features

2. Test Data: 8 countries, Year: 2016

   Rows: 8 X 1 = 8, Columns = 21 features

	Country	Year	Deaths	CG_VOTO	EQ_AIRP	EQ_WATER	ES_EDUA	ES_EDUEX	ES_STCS	HO_BASE	...	IW_HADI	IW_HNFW	JE_EMPL	JE_LTU
0	Australia	2013	658.103782	93.0	14.0	91.000000	73.333333	18.500000	519.200000	1.2	...	32538.666667	32178.0	72.40	1.01146
1	Australia	2014	664.537658	93.0	13.0	91.000000	74.333333	18.700000	513.600000	1.1	...	35144.666667	38482.0	71.80	1.12097
2	Australia	2015	675.651352	93.0	13.0	91.000000	76.333333	19.300000	512.600000	1.1	...	35250.666667	47657.0	70.80	1.13010
3	Brazil	2013	587.344164	79.5	19.0	78.500000	41.000000	16.266667	404.800000	6.7	...	13623.000000	5861.0	70.40	2.68732
4	Brazil	2014	594.874998	79.5	18.0	71.000000	43.333333	16.300000	406.200000	6.7	...	13620.666667	6875.0	69.00	2.07492
5	Brazil	2015	604.891721	80.0	18.0	72.333333	44.666667	16.300000	404.000000	6.7	...	15480.666667	6844.0	69.20	1.78917
6	Germany	2013	1096.438309	70.4	16.0	94.800000	85.666667	17.866667	509.400000	0.9	...	31773.666667	44938.0	70.40	3.16461
7	Germany	2014	1089.524152	71.0	16.0	95.000000	86.333333	18.066667	515.800000	0.9	...	33894.000000	49484.0	70.60	2.86721
8	Germany	2015	1103.744693	71.0	16.0	95.000000	86.333333	18.200000	516.400000	0.1	...	34757.333333	50394.0	70.60	2.73347
9	Japan	2013	1021.032003	69.0	25.0	85.500000	92.000000	18.733333	529.600000	6.4	...	26755.333333	74966.0	69.20	1.80278
10	Japan	2014	1030.930593	59.5	24.0	85.500000	93.000000	16.200000	539.333333	6.4	...	27773.666667	85309.0	72.40	1.63546
11	Japan	2015	1049.830073	52.5	24.0	84.500000	93.500000	15.866667	540.250000	6.4	...	28931.000000	86764.0	71.25	1.65912
12	Mexico	2013	505.077088	63.0	33.0	78.250000	36.333333	14.866667	421.400000	4.2	...	16019.333333	9946.0	61.80	0.10373
13	Mexico	2014	510.444746	63.0	30.0	69.500000	36.333333	15.200000	418.800000	4.2	...	16168.000000	10449.0	63.00	0.09707

Training dataset

TASK 1: Machine Learning Approach

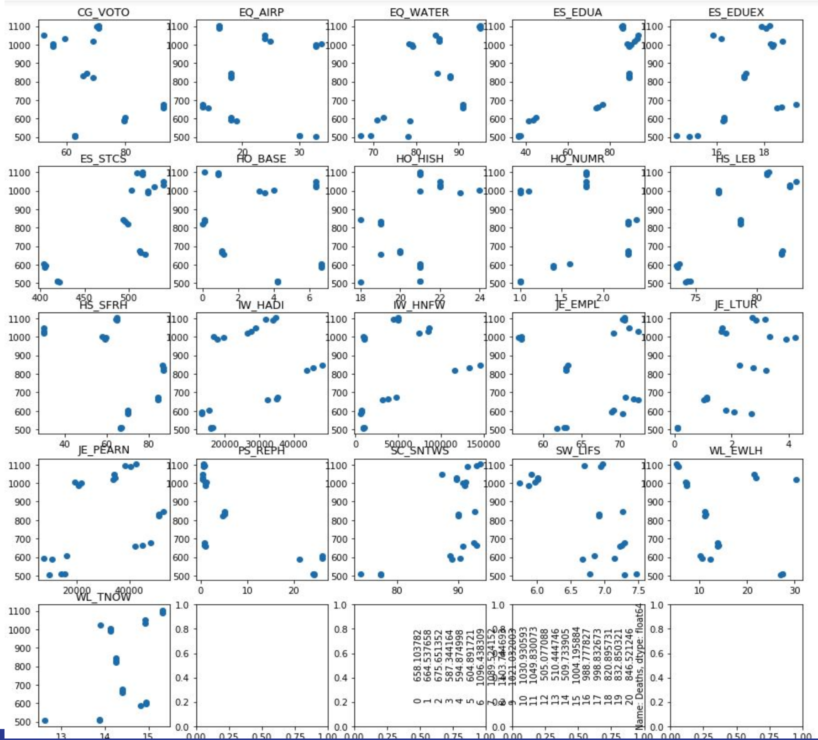
1. Feature Exploration

Some features (or independent variables) are positively correlated, some negatively correlated while others are not correlated to the dependent variable.

Positively Correlated:
ES_EDUA: Educational attainment
ES_EDUEX: Years in education
JE_EMPL: Employment rate

Negatively Correlated:
HO_NUMR: Rooms per person
CG_VOTO: Voter turnout
WL_EWLH: Employees working very long hours

Note: We want the negatively correlated variables as they reduce the health loss.



TASK 1: Machine Learning Approach

1. Dimensionality Reduction

- Grid Search to find optimal parameters for Ridge regression and PCA. Dimension (n=7) works best for the model, achieving R^2 error of 0.96 on training and 0.86 on test set.
- 95% of the information is stored in 1st principal component.

Best Params : PCA dim = 7, alpha = 0.001 ¶

```
pipe = make_pipeline(StandardScaler(), PCA(n_components=7), StandardScaler(), Ridge(alpha=0.001))

ridge01 = pipe.fit(X_train, y_train)
print("Training set R2 score: {:.2f}".format(ridge01.score(X_train, y_train)))
print("Test set R2 score: {:.2f}".format(ridge01.score(X_test, y_test)))
```

```
Training set R2 score: 0.96
Test set R2 score: 0.86
```

```
### 95% of information can be represented by 1 feature.
pca = PCA(n_components=7)
pca.fit(X_train)
print(pca.explained_variance_ratio_)
```

```
[   9.56844904e-01   4.20318030e-02   1.12273786e-03   4.47058023e-07
    4.64281397e-08   3.58441575e-08   1.86334007e-08]
```

TASK 1: Machine Learning Approach

Grid Search

- Grid Search to find optimal parameters for Ridge regression and PCA. Dimension (n=7) and alpha = 0.001 works best for the model, achieving R^2 error of 0.96 on training and 0.86 on test set.
- 95% of the information is stored in 1st principal component.
- K-Fold cross-validation with folds=10.

```
pipe = make_pipeline(StandardScaler(), PCA(), StandardScaler(), Ridge())

param_grid = {'ridge__alpha': [0.001, 0.01, 0.1, 1, 10, 100], 'pca__n_components': range(2,11)}
grid = GridSearchCV(pipe, param_grid, cv=10, scoring='r2')
grid.fit(X_train, y_train)
print("Best estimator:\n{}".format(grid.best_estimator_))
```

```
Best estimator:
Pipeline(steps=[('standardscaler-1', StandardScaler(copy=True, with_mean=True, with_std=True)), ('pca', PCA(copy=Tru
e, iterated_power='auto', n_components=7, random_state=None,
  svd_solver='auto', tol=0.0, whiten=False)), ('standardscaler-2', StandardScaler(copy=True, with_mean=True, with_std
=True)), ('ridge', Ridge(alpha=0.001, copy_X=True, fit_intercept=True, max_iter=None,
   normalize=False, random_state=None, solver='auto', tol=0.001))])
```

TASK 1: Machine Learning Approach

1. Performance based on selecting top K features from high correlation features from Scatter plots.

   Top K = 16 features

   Training set performance: 0.98
   Test set performance: 0.84

2. Negative coefficients represent negative correlation with deaths, while positive coefficients show positive correlation. p<0.05 is statistically significant.

Positive:
JE_EMPL: Employment rate
ES_EDUA: Educational attainment
ES_EDUEX (p high): Years in education

Negative:
HO_NUMR: Rooms per person
WL_EWLH (p somewhat high) : Employees working very long hours

	Coefficients	Standard Errors	t-values	p-values	0
0	-71.3305	24903.947	-0.003	0.998	CG_VOTO
1	-21.1180	33.511	-0.630	0.536	EQ_AIRP
2	41.1070	102.032	0.403	0.691	EQ_WATER
3	39.1750	16.870	2.322	0.031	ES_EDUA
4	29.2858	26.006	1.126	0.273	ES_EDUEX
5	-2.4290	144.639	-0.017	0.987	ES_STCS
6	-49.0564	5.459	-8.986	0.000	HO_NUMR
7	22.8467	410.559	0.056	0.956	HS_LEB
8	-77.1882	202.194	-0.382	0.707	HS_SFRH
9	36.1360	32.375	1.116	0.278	IW_HADI
10	6.0398	0.021	287.070	0.000	IW_HNFW
11	17.8815	0.007	2700.870	0.000	JE_EMPL
12	28.8534	122.479	0.236	0.816	JE_LTUR
13	-4.1222	204.101	-0.020	0.984	PS_REPH
14	4.7103	53.836	0.087	0.931	SC_SNTWS
15	-46.3292	25.034	-1.851	0.079	WL_EWLH
16	818.7730	36.435	22.472	0.000	NaN

TASK 2: Statistical Approach

1. F_regression: Computes the correlation between each regressor and the target.

2. Mutual_info_regression: Mutual information (MI) measures the dependency between the independent and the target variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

```
f_test, _ = f_regression(X_train, y_train)
f_test /= np.max(f_test)
print(f_test)
```

```
[  7.26405407e-02   4.83384234e-05   1.56161550e-01   1.00000000e+00
   1.07968881e-01   5.39318669e-01   1.32486699e-02   7.40786894e-02
   7.47315741e-03   1.82469086e-01   9.60477354e-02   4.36753514e-02
   4.99680673e-02   5.95390102e-05   2.58055036e-01   6.57741064e-02
   5.39187793e-01   1.65057585e-01   2.05351880e-01   5.90182622e-02
   8.31705605e-02]
```

```
mi = mutual_info_regression(X_train, y_train)
mi /= np.max(mi)
print(mi)
```

```
[ 0.29842856  0.          0.33806028  1.          0.          0.68636311
  0.          0.68813703  0.6617069   0.53383858  0.707954    0.58837669
  0.25326791  0.32820513  0.33578287  0.81627698  0.81655875  0.26368736
  0.68192114  0.35996316  0.32674854]
```

TASK 2: Statistical Approach

1. Selected top 18 features with maximum mutual information gain.

2. Grid search with K-fold cross validation. Best param: alpha=1.0.

3. Achieved training performance of R^2 = 0.98 and testing R^2 error = 0.93.

```
pipe = make_pipeline(StandardScaler(), Ridge(alpha=1.0))

ridge01 = pipe.fit(X_train_new, y_train)
pred = ridge01.predict(X_train_new)
params = np.append(ridge01.named_steps["ridge"].coef_, ridge01.named_steps["ridge"].intercept_)
# print(summ(X_train_new, y_train, pred, params))


print("Training set R2 score: {:.2f}".format(ridge01.score(X_train_new, y_train)))
print("Test set R2 score: {:.2f}".format(ridge01.score(X_test_new, y_test)))
```

```
Training set R2 score: 0.98
Test set R2 score: 0.93
```

TASK 2: Statistical Approach

1. Positively Correlated
   HO_HISH: Housing expenditure
   EQ_WATER: Water quality
   ES_EDUA (very high p): Educational attainment

2. Negatively Correlated:
   HO_NUMR: Rooms per person
   CG_VOTO (very high p): Voter turnout

OBSERVATION

1. Since Number of rows M ~ N (Number of columns), therefore training a regression model is facing issues, getting arbitarily high p-value when M ~ N (dimensions not reduced). When the dimensions are reduced, the p -values become more stable.

	Coefficients	Standard Errors	t-values	p-values	0
0	40.9324	11901.653	0.003	0.997	ES_EDUA
1	-6.7955	53.017	-0.128	0.899	PS_REPH
2	16.9051	31.064	0.544	0.592	JE_PEARN
3	11.3744	0.047	239.476	0.000	HO_HISH
4	-76.1730	78.484	-0.971	0.343	HS_SFRH
5	4.3016	32.916	0.131	0.897	ES_STCS
6	18.3144	12.820	1.429	0.169	SW_LIFS
7	24.1036	355.825	0.068	0.947	IW_HADI
8	-42.0467	0.059	-709.789	0.000	HO_NUMR
9	17.3410	1575.436	0.011	0.991	HS_LEB
10	-19.1050	155.402	-0.123	0.903	WL_EWLH
11	56.4453	72.501	0.779	0.445	EQ_WATER
12	35.4676	82.638	0.429	0.672	JE_LTUR
13	-12.4199	298.219	-0.042	0.967	JE_EMPL
14	29.9893	226.684	0.132	0.896	WL_TNOW
15	-48.5697	1012.083	-0.048	0.962	CG_VOTO
16	20.8368	44.341	0.470	0.643	SC_SNTWS
17	-0.1131	62.166	-0.002	0.999	IW_HNFW
18	818.7730	0.010	80573.048	0.000	NaN

TASK 3: Correlation Analysis

Combining the correlations from various approaches:

Positive:
JE_EMPL: Employment rate
ES_EDUA: Educational attainment
ES_EDUEX (p high): Years in education

Negative:
HO_NUMR: Rooms per person
WL_EWLH (p somewhat high) : Employees working very long hours
CG_VOTO: Voter turnout

Therefore, the below mentioned variables should result in the maximum reduction in health loss in 2016.

HO_NUMR: Rooms per person
WL_EWLH (p somewhat high) : Employees working very long hours
CG_VOTO: Voter turnout