10/29/2020 Matt Ring
PPOL 564 – Data Science Foundations Final Project Proposal

1. High-level statement of the problem or analysis

How might work, including union membership, wages, and working time, relate to happiness and life satisfaction?

2. Data sources

World Happiness Report (Gallup Poll data): Individual datasets of happiness by country from 2015-2019 sourced from Kaggle. These data come from Gallup World Polls and are used by the World Happiness Report. Variables include, but are not limited to, country name, region, happiness score, freedom, generosity, and trust in government. These are stored in .csv files, separated by year, which would need to be merged.

Our World in Data (“Self-Reported”): A .csv file containing country, country abbreviation, year, and happiness, as assessed using the Cantril’s Ladder question. Years range from 2005 to 2018, dependent on the country. Our World in Data’s article Happiness and Life Satisfaction (Ortiz-Ospina and Roser) may provide a useful framework for interpreting this data.

International Labor Organization (ILO) (“Home”): Statistics from the ILO include employment, wages, working time, inequality, social protection, productivity, inflation, union membership, and more. Most of these data span the years 2000 to 2016, though this varies.

Polity IV Project (Marshall and Gurr): These data were used in class but had also been merged with data on GDP per capita and population. The information on regime types and GDP per capita should be included to reduce omitted variable bias.

3. Plan to obtain data

Most of these data can be downloaded in useable formats from the sites above. The World Happiness Report data, Polity IV data from class, and Our World in Data happiness data are in .csv formats. The ILO data may be downloaded as either a .xlsx or a zipped .csv file. Both the World Happiness Report and ILO data come in a series of files split by year and particular statistic, respectively. The World Happiness Report data would be drawn from Kaggle.

4. Methods

Data wrangling:

Data wrangling for this project would involve merging of each dataset, dealing with missing values, and selecting features. Lack of data for given years and countries will necessitate filling or removing various observations and taking note of how this is done. Creation of new features, assignment of dummy variables, and interaction variables may also be necessary. These features would be scaled and normalized before any analysis.

Visualizations:

As many of these features are likely to be related, a correlation matrix might be necessary. This correlation matrix would likely utilize Pearson’s r and be colored by the level of correlation. Should this prove too difficult or unsatisfactory, Seaborn’s pairplot function may also work.

Interpretability of the coefficients of this analysis is paramount, as I hope to determine the impact of a set of labor statistics on happiness and life satisfaction. A simple plot for each model showing a point for each scaled variable representing its coefficient would suffice. These points could be colored by significance levels.

Each model would have a learning curves for cross validation. This would show performance on training and validation sets of each model, revealing potential issues with overfitting should the training score be particularly high while the validation score remains low. The x-axis would display increasing sizes of the training set. Line plots of the flexibility as related to Mean Squared Error (MSE), as shown in Chapter 2 of "An Introduction to Statistical Learning" would be useful for showing overfitting should learning curves prove unacceptable or not useful.

Countries with high happiness metrics or labor statistics, such as productivity, could be highlighted using radar plots. A single radar plot could be displayed for each year of data.

Statistical Learning:

I will construct at least two supervised regression models using labor statistics, polity data, and other control variables to predict happiness by country and year. One model will be as interpretable as possible (linear regression or random forest), and one will be more accurate (SVM). Scikit-learn will be used for creating and tuning these models. Cross validation using k-fold validation and bootstrapping would be used for both models. Model fit will be assessed using MSE, and this will subsequently be minimized by balancing the bias-variance trade off as discussed in Chapter 2.2.2 of "An Introduction to Statistical Learning".

5. A definition for "success"

"Success" for this project entails showing statistically whether any work statistic, such as union membership, wages, and working time, have a meaningful impact on happiness. Furthermore, significant coefficients would need to be described in understandable terms as to their impact on happiness. Comparison of coefficients, both in terms of magnitude and significance, would constitute the final component of "success" for this project.

Websites and Books Referenced:

1. "Home - ILOSTAT - The Leading Source of Labour Statistics." *ILOSTAT*, ilostat.ilo.org/.
2. James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. Springer, 2017.
3. Marshall, Monty G, and Ted R Gurr. "Polity IV Project: Political Regime Characteristics and Transitions, 1800-2013." *Polity IV Project: Country Reports 2010*, University of Maryland, 6 June 2014, www.systemicpeace.org/polity/polity4.htm.
4. Ortiz-Ospina, Esteban, and Max Roser. "Happiness and Life Satisfaction." *Our World in Data*, 14 May 2013, ourworldindata.org/happiness-and-life-satisfaction.
5. "Self-Reported Life Satisfaction." *Our World in Data*, ourworldindata.org/grapher/happiness-cantril-ladder.

Packages Referenced:

- Scikit Learn
- Pandas
- Seaborn