**Quality of Life**
**Exploratory Data Analysis**
DATS 6101 Group Project 1
March 7, 2018

Ye-in Jeon
Sam Kupfer
Amy Mykityshyn
Krystin Sinclair
Vishnupriya Venkateswaran

Background of Dataset

Our team members share common interests in the areas of public health and the environment. After researching several publicly-available datasets, we found a data set online from the Organisation for Economic Cooperation and Development (OECD), an international forum with 35 member countries from around the globe. Founded in 1961, the OECD's mission is "to promote policies that will improve the economic and social well-being of people around the world" (Organisation for Economic Cooperation and Development, About the OECD, 2018). Representatives from member countries provide input to specialized committees that review current developments in several policy areas, like economics, trade, science, employment, and education. These policy recommendations and collected data are utilized by the OECD to develop and promote policies that improve the quality of people's lives.

Our data set was created from annual data available through the OECD web site concerning Social Protection and Well-being. It is called the Better Life Index and it includes data on housing, income, jobs, community support, education, environment (air and water quality), civic engagement, health, life satisfaction, safety, and work-life balance, for the 35 OECD member countries, plus three key partners, Brazil, South Africa, and Russia. It was our goal to investigate this collection of data to see which of these variables contributed most significantly to overall life satisfaction of the people who live across the globe over the past three years.

There are some limitations to this dataset. Life satisfaction is a survey result, therefore subjective in nature. Many of the other variables included in this survey are also subjective questions. Due to cultural differences around the world, what is perceived as a "satisfying life" or "good health" may differ. We also have only three years of data. This will allow for minimal time series comparisons.

To graphically show the countries for which OECD records quality of life data and to see which countries had high or low life satisfaction scores, we used a library in R, “rworldmap”, that allowed us to depict the 2017 measures of life satisfaction pictorially as shown in Figure 1.

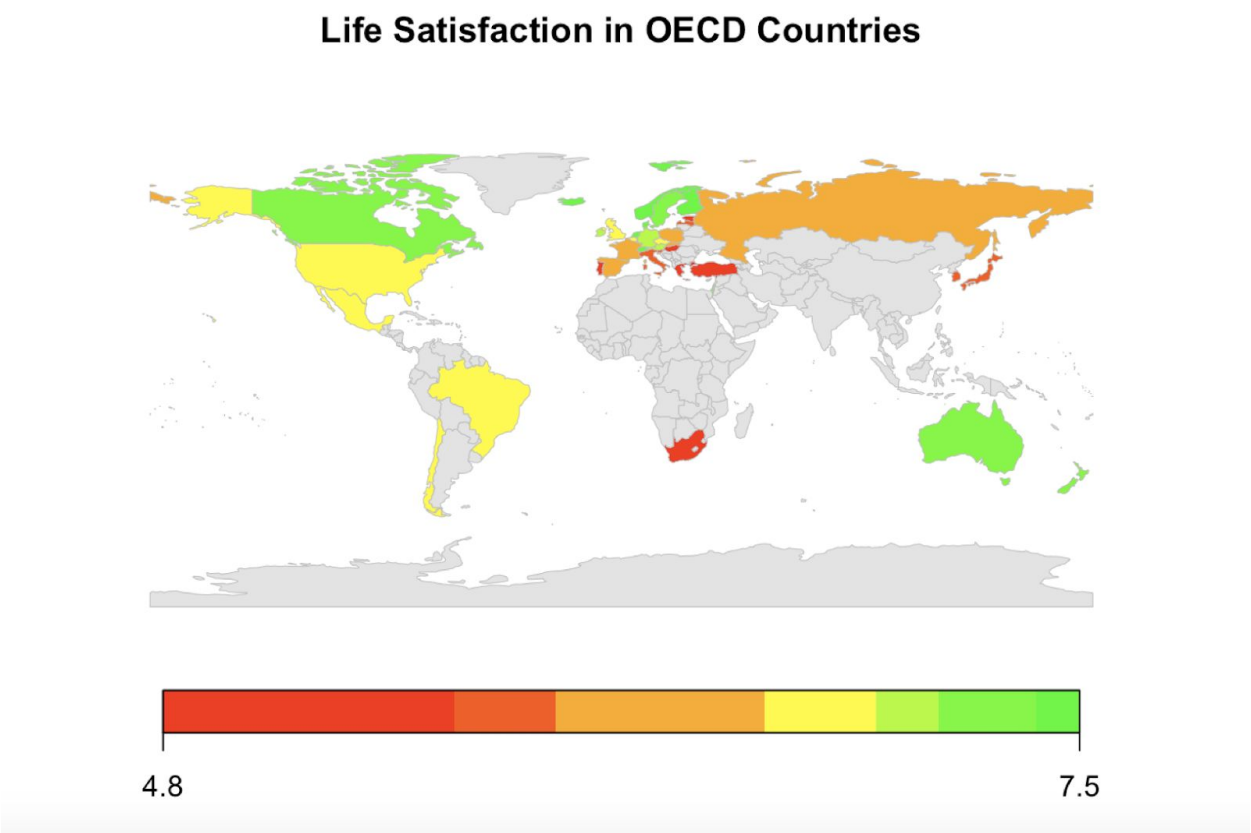


Figure 1. Life Satisfaction in OECD Countries for 2017

Data Exploration

Better Life Index data were provided on the OECD web site, by year (Organisation for Economic Cooperation and Development, 2015-2017). It could be exported from a static web page in Excel or .csv format. According to the OECD, the data that comprise 11 high-level topics of well-being are sourced primarily from official sources like the OECD or national accounts, United Nations statistics, and National Statistics offices. A few indicators were from the Gallup World Poll (Organisation for Economic Cooperation and Development, What’s the Better Life Index?, 2018).

We first created a complete data frame by merging the annual data provided for 2015, 2016, and 2017, as these were the most recent years and they most closely aligned in terms of the provided variables. We utilized str() in R to get a view of overall structure of the variables in our dataset of 112 observations and 29 variables. This shows that the majority of the variables are ratio and are grouped by country, a nominal variable. Then we viewed descriptive statistics for all variables to develop a greater understanding of the data. This allowed us to discuss methods of grouping variables and for selection of the main variables of interest as shown in Table 1.

Table1. Descriptive Statistics for Variables of Interest

| Variable | Count | Mean | Std. Dev | Min | 25% | Median | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| Life satisfaction | 112 | 6.54 | 0.77 | 4.80 | 5.90 | 6.70 | 7.30 | 7.60 |
| Personal earnings | 112 | $36,482 | $13,965 | $5,429 | $23,593 | $36,504 | $48,308 | $62,636 |
| Employment rate | 112 | 66.84% | 7.97% | 43.00% | 62.00% | 67.50% | 73.00% | 86.00% |
| Household net adjusted disposable income | 112 | $24,294 | $7,776 | $8,712 | $18,252 | $23,965 | $29,567 | $44,049 |
| Rooms per person | 112 | 1.64 | 0.45 | 0.7 | 1.2 | 1.7 | 1.9 | 2.5 |
| Time devoted to leisure and personal care | 112 | 14.84 | 0.70 | 12.24 | 14.48 | 14.90 | 15.17 | 16.36 |
| Homicide rate | 112 | 2.95 | 5.64 | 0.20 | 0.60 | 1.00 | 1.53 | 27.60 |
| Life expectancy | 112 | 79.51 | 4.23 | 56.80 | 78.15 | 81.05 | 81.92 | 83.90 |
| Air pollution | 112 | 15.63 | 7.05 | 3.00 | 11.00 | 15.00 | 18.00 | 46.00 |
| Years of education | 112 | 17.41 | 1.31 | 14.40 | 16.40 | 17.40 | 18.12 | 21.20 |

The main variable of interest to us is Life Satisfaction. The measure of life satisfaction for each country was a compilation of citizen's responses to several different survey questions, on a 0-to-10 scale. We noticed that for the United States life satisfaction had a rating of 6.9 which is slightly above the median 6.7. The maximum score was seen in Switzerland, Hungary and Denmark at 7.6. We calculated descriptive statistics on all the other variables in the data set and our overall purpose was to understand which of the other variables correlated with Life Satisfaction ratings. The descriptive statistics showed that for the variables of interest some had relatively small standard deviations, others had wider distributions with outliers. Homicide rate, for example, had some outliers on the high end, those countries include Brazil, Mexico, Russia, and South Africa.

In the table above we showcase our variables of interest. We narrowed down the variables to focus on based on background research. There have been many quantitative and qualitative studies done in regards to life satisfaction. From the articles that we read,

we noticed patterns of certain variables appearing multiple times (more on this in the “Background Research” section below). We decided to make these variables or categories of variables our main interest groups and find out if our research project would find similar results to prior studies.

As part of data exploration we created box plots and histograms in R, as seen in Figures 2 through 4. This allowed us to visually assess the normality of the variables and see some trends. For example, personal earnings has a bimodal distribution. Self reported health is left skewed and air pollution is right skewed. Self reported health had some outliers on the lower end and air pollution had an outlier on the higher end. For example, in 2015 Chile declared a state of environmental emergency due to air pollution near its capital, Santiago.


Figure 2. Personal Earnings Data

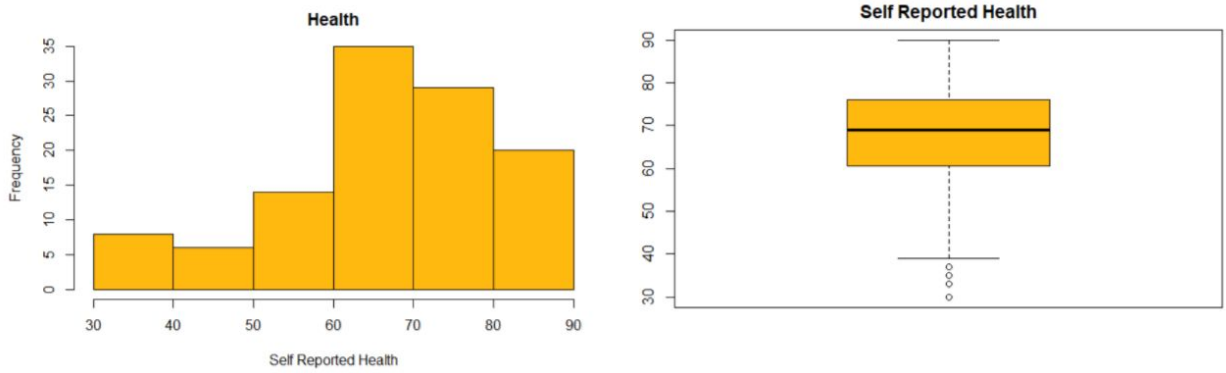

Figure 3. Self Reported Health Data

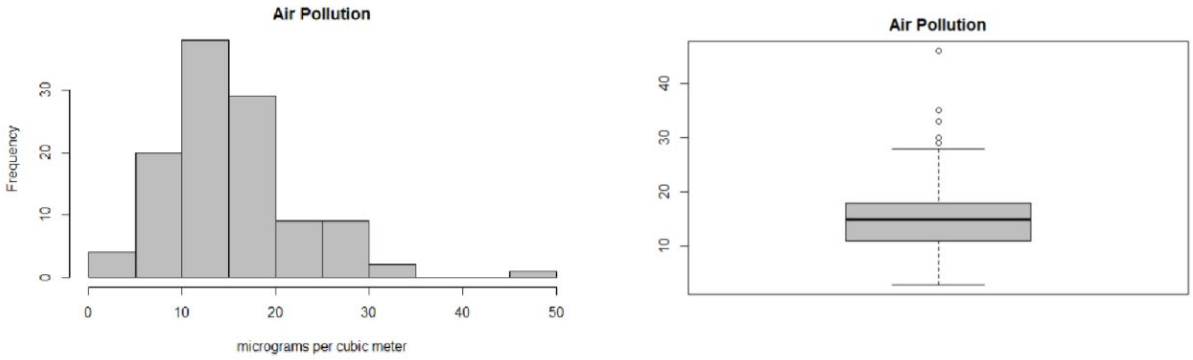


Figure 4. Air Pollution Data

## Background Research

Researchers have long been studying what factors contribute to measures of life satisfaction. In 1972, Palmore and Luikart published an early journal article on the subject, and they reached the conclusion that, “not only that health is the strongest variable, but that it alone accounts for the large majority of explained variance in life satisfaction” (Palmore and Luikart, 1972). Since then, many psychologists and sociologists have conducted investigations to study the attributes that contribute to the quality of life for people around the world. Researchers have found that both economic and civic engagement are necessary for happiness or life satisfaction. Simon Rogers, a writer for The Guardian, found that financial factors are the top contributors to life satisfaction (Rogers, 2012). A more recent article from the Positive Psychology Program states that health and healthcare are the main variables that contribute to life satisfaction (Chompoo, 2015). These readings showed that there have been many studies that tried to assess life satisfaction and its contributing factors quantitatively. Most of the studies we found came to different conclusions on which variables contribute most to life satisfaction. We felt that this OECD dataset would be ideal for answering this question on our own for the countries surveyed by the OECD over the past three years.

## SMART Question Development

Keeping in mind the OECD mission and the information contained in our dataset, we were interested in what affects life satisfaction. We started with a very general question: “What characteristics contribute most to life satisfaction?”. This is a very broad, almost

philosophical question that can be answered or discussed in many ways, both qualitatively and quantitatively. We clearly needed to make this question more specific for quantitative study. This began with grouping the variables, that made it easier to talk about and provided a way to compare variables. The next iteration of our question became: “Do economic, mental and physical health, and environment and education contribute to life satisfaction?” This question still did not meet all of the SMART criteria.

Taking this question and making it a bit more quantitative, we came up with: “How much do economic, mental and physical health, and environment and education contribute to life satisfaction?” We found “contribute” to be a bit vague and hard to measure, so for more specification, we chose to use “correlation” instead. We chose to calculate the correlation using Pearson’s method of correlation. After further discussion, we decided to define “strong correlation” as having an absolute value of Pearson correlation greater than .6. We also chose to reframe our question, so that it’s answer would be a straightforward list of variables. This answer would be a better reflection of the question we originally asked (“What characteristics contribute most to life satisfaction?”). After all this, we discussed again and decided on a final, most specific question: “Which of the variables in the dataset correlate strongly ($|r| > .6$) with life satisfaction in countries considered by the OECD from 2015-2017?”

This question meets the SMART criteria for research question development:

a) It is **S**pecific, because it is about a specific set of countries, time frame and focuses on 10 variables.
b) It is **M**easurable: it can be measured with a correlation coefficient. We decided that $|r| > .6$ is considered a strong correlation for our set of data. The rho is from a Pearson correlation which is the measure of the strength of a linear association between two variables.
c) This question is **A**chievable; the dataset is small enough to be analyzed in R.
d) The question is **R**elevant to researchers who want to understand factors that drive global life satisfaction. It also aligns with the OECD goals for this initiative, as the OECD Better Life Initiative focuses on developing statistics to capture aspects of life that matter to people and that shape the quality of their lives. This allows for a better understanding of what drives the well-being of people and nations, and what needs to be done to achieve greater progress for all. (1)
e) Lastly, it is **T**ime oriented. We believe that this question is straightforward enough to be answered in the given time and the data is already collected.

Statistical Analyses (Correlation)

We focused on our SMART question and examined which of the other variables in the Better Life Index correlate strongly(|r| > .6) with life satisfaction in countries considered by the OECD from 2015-2017? The five variables that strongly correlate with life satisfaction are Personal earnings in US dollars with r =.72, Employment Rate as a percentage of working age population with r=.7, household net adjusted disposable income in US dollars with r=.64, and rooms per person as the total number of rooms divided by number of people in dwelling with r=.62, and self reported health as a percentage reporting good or better health based on survey data with r=.64. Economic variables had the strongest correlations to life satisfaction. Health also has a strong correlation. These results agree with the results of some of our background research. The variables that we were interested in that did not strongly correlated was life expectancy in number of years with r=.48, air pollution as micrograms per cubic meter with r = -.3, years in education as number of years in school since age 5 with r=.36 ad homicide rate as the rate of death by assault per 100,000 people with r =-.11. The correlation matrix is shown below.

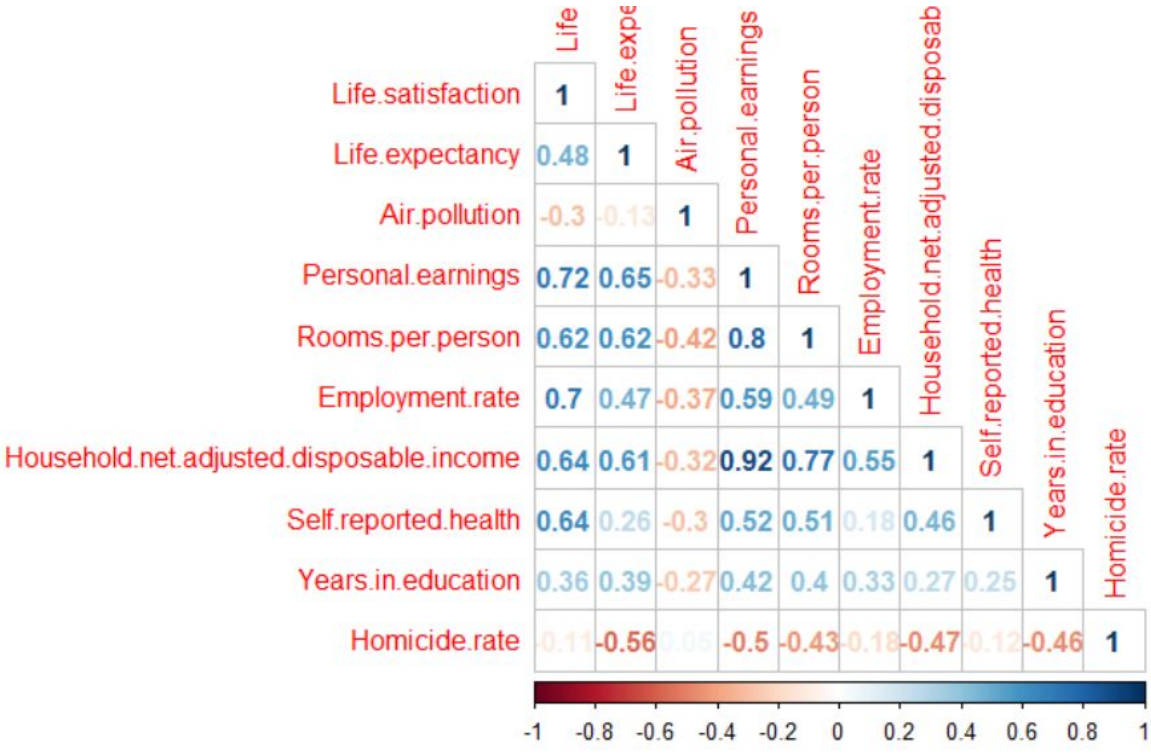


Figure 5. Correlation Values for Variables in the Better Life Index

## Conclusions from Exploratory Data Analysis and Future Analyses

Our expectations matched with the data for the most part. From our background research, we expected that Personal Earnings and Employment Rate would correlate highly with overall Life Satisfaction of the people who live in OECD countries. Not all of the nine variables of interest we selected are strongly correlated with overall Life Satisfaction, but five are. This aligned with the readings that we had done prior to data analysis. The financial variables show stronger correlation than the education and environment variables.

With more information collected in the future, we may be able to look at trends for specific countries or small groupings of countries. We could try to cluster countries together based on these variables. This research project could be continued into another stage to further our knowledge of the variables that affect global Life Satisfaction.

As stage two of this project, we intend to build predictive models to assess what the best case scenario of all attributes are to have maximum life satisfaction. These models could also be used to investigate the effect of different cultural expectations on what survey respondents consider a “satisfying life”. This could be done by using the model to estimate an expected value of life satisfaction for a country, and comparing this to the actual value based on the survey answers from people in that country.

References

Chompoo (2015). Life Satisfaction and Its 7 Contributing Factors. February 24, 2018 from https://positivepsychologyprogram.com/life-satisfaction/

Organisation for Economic Cooperation and Development. About the OECD. February 10, 2018 from http://www.oecd.org/about/

Organisation for Economic Cooperation and Development. *Better Life Index - concatenated, 2015-2017* [Data set]. Retrieved from http://stats.oecd.org/Index.aspx?DataSetCode=BLI

Organisation for Economic Cooperation and Development. What's the Better Life Index?. February 11, 2018 from http://www.oecdbetterlifeindex.org/about/better-life-initiative/

Palmore, E. & Luikart, C. (1972). Health and Social Factors Related to Life Satisfaction. [Electronic version]. *Journal of Health and Social Behavior, 13.1,* 68-80.

Rogers, S. (2012). Better Life: Relaunching the Happiness Index. February 24, 2018 from https://www.theguardian.com/news/datablog/2012/may/22/better-life-index-oecd