# Life Expectancy Analysis Using Regression

Kimberly Statham

Manning MEAP project

December 1, 2020

Life Expectancy Analysis Using Regression Analysis

This MEAPS project uses the CRISP-DM methodology to perform data mining and regression analysis on data related to life expectancy. The datasets used in the analysis are provided. The Jones Family Foundations has made a goal this year to find variables related to the life expectancy of Americans. Additionally, the foundation is seeking more knowledge on the geographic distribution of certain demographic factors that may be related to life expectancy.

## Business Understanding

The primary business objective of this project is to identify non-health related characteristics that have a direct relation to life expectancy of American people. The data to be examined is grouped by states, the District of Columbia, and the five inhabited territories in the United States. The demographic information to be analyzed is related to level of education, life expectancy historical data, crime statistics, income, and geographic features.

A secondary business objective is to answer the Board of Directors five questions connected to the demographic data. The five questions to be answered are the following:

1. What percent of the states have a life expectancy greater than 80 years?
2. Which state has the highest life expectancy and which state the lowest?
3. Is life expectancy equally distributed across the different regions of the U.S.?
4. Is education equally distributed in each of the three levels (high school, bachelor, and advanced degrees)? Does any of the educational levels show a greater spread across the states?
5. How does the level of education in the different states relate to life expectancy and income?

The primary business objective success criteria are that useful insights into non-health factors that influence life expectancy are exposed by data mining techniques and regression models. If there is indication that some of the variables in the data do effect life expectancy, then

these insights will be validated by a minimum accuracy of 75% in predicting a states average life expectancy for both females and males.

The secondary business objective success criteria will be that the answers to the BOD questions are easily explained using visuals.

**Goals and Assessments**

The factors to be explored are from provided datasets that are indexed by USA states, districts, or territories. The goal will be to determine if any of the variables in the datasets impact a state's life expectancy. The criteria to be used to confirm the results of the goal is that a regression analysis model accurately predicts a state's life expectancy at least 75% of the time. The validation dataset will be used to assess the accuracy of the model. The validation data will consist of 50% of the data records, randomly selected.

A secondary goal will be to answer the Board of Directors five questions. Questions 1 and 2 will be answered in the data exploration step. Questions 3 and 4 will be addressed with a visual graph. Question 5 will be explained with a regression model.

The Python programming language along with the Jupyter Notebook IDE will be used for the data cleaning, data exploration, and the regression analysis model. The data will need to be transformed from three different file formats, CSV, Excel, and text, and then merged together by state, into a tabular format with the records indexed by state/district/territory. The data to be analyzed is small and there are no performance requirements so the data mining and analysis will be performed on a personal computer.

**Table 1**

*Modeling Risks with Associated Resolution*

| Risk | Resolution |
|---|---|
| Minimal data analytic experience using Python. | Accepted. The goal of doing this project is to improve data analytics and Python skills. |
| Small amount of records available for the regression modeling which could slant accuracy of the model. | Accepted. There is no additional data at this time. |

**Project Plan**

The project plan, including durations and resources, is shown below in table 2. The only known risk to this project plan is lack of experience doing data analytics and minimal experience with the Python programming language so the project schedule may slip.

**Table 2**

*Project Plan*

| Phase | Time | Resources | Notes |
|---|---|---|---|
| Business Understanding | 1 week<br>12/1/20-12/7/20 | One Data Analyst | Deliverable is the Project Plan |
| Data Understanding | 3 days<br>12/8/20-12/12/20 | One Data Analyst | Includes Data Collection and Cleaning. |
| Data Preparation | 1 week<br>12/13/20-<br>12/20/20 | One Data Analyst | Includes merging data from all files into a tabular format and data exploration |

| | | | |
|---|---|---|---|
| Modeling | 2 weeks<br>12/21/20-<br>1/3/20 | One Data Analyst | Includes holiday time off, unusual observations, answering the BOD questions and regression analysis |
| Evaluation | 1 week<br>1/4/20 – 1/11/20 | One Data Analyst | |