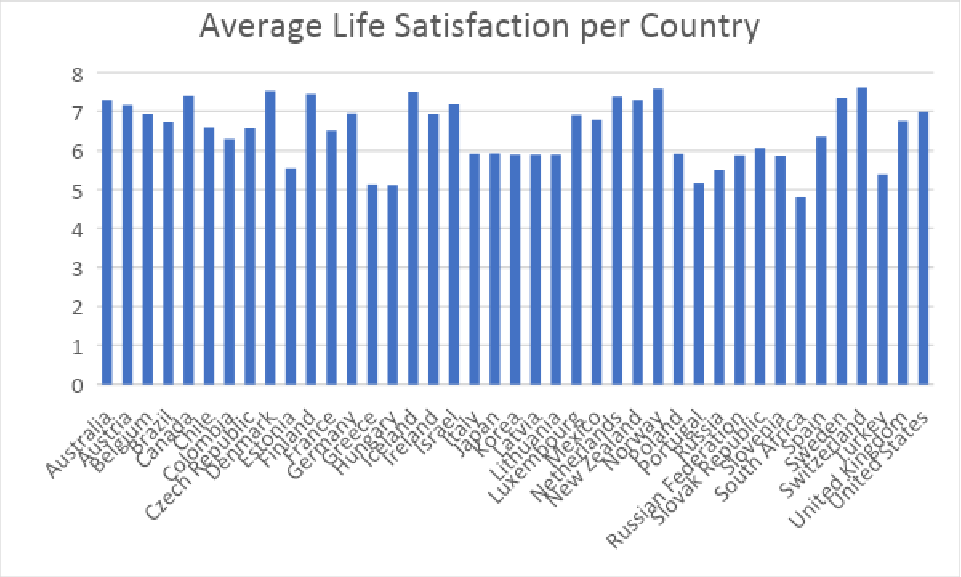
The topic of life satisfaction is significant to individuals and societies worldwide. It is a subjective measure of well-being and happiness that can have significant impacts on an individual's overall quality of life (De Neve et al., 2013). Understanding the factors that contribute to life satisfaction can help individuals and communities make informed decisions about how to improve well-being and happiness. In this research project, we looked to investigate the relationship between life satisfaction and various socio-economic variables, including income, education level, employment status, and social support.

We were particularly interested in how these variables might vary across different countries and regions. The goals of this project were to examine the data on life satisfaction and socio-economic variables, find any trends or patterns that emerged, and use these insights to build a prediction model for life satisfaction. We also hoped to evaluate the performance of the model using techniques such as cross-validation and compare the results to other prediction models.

Our thesis is that several socio-economic variables are significantly associated with life satisfaction and that a predictive model can be developed to accurately forecast life satisfaction based on these variables. We expect that the model will be able to explain a sizable portion of the variance in life satisfaction and will perform well when evaluated using cross-validation. We also anticipate that the results of this research will be useful for individuals and communities seeking to improve well-being and happiness. As you read this research, consider how the results might apply to your own life and what you can do to improve your own level of life satisfaction. Do the results align with your own experiences and observations, or do they challenge your assumptions about what contributes to happiness and well-being?

Part 1

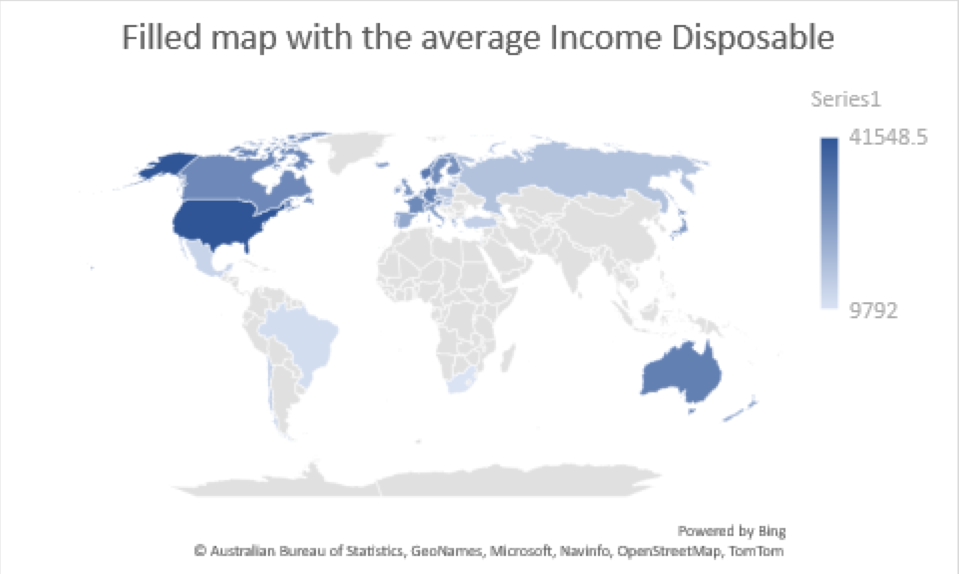
First, I started to investigate the relationship between life satisfaction and various socio-economic variables such as income, education level, employment status, and social support. I started by exploring the data visually, using graphs and plots to get a sense of the distribution of life satisfaction scores and how they vary across the different countries and socio-economic variables.



Row Labels	**Average of Life Satisfaction**
Africa	**4.8**
South Africa	4.8
Asia	**6.338888889**
Israel	7.183333333
Japan	5.933333333
Korea	5.9
Aus	**7.3**
Australia	7.3
New Zealand	7.3
Eur	**6.501875**
Austria	7.166666667
Belgium	6.933333333
Czech Republic	6.566666667
Denmark	7.533333333
Estonia	5.55
Finland	7.45
France	6.516666667
Germany	6.95
Greece	5.133333333
Hungary	5.116666667
Iceland	7.516666667
Ireland	6.933333333
Italy	5.916666667
Latvia	5.9
Lithuania	5.9
Luxembourg	6.916666667
Netherlands	7.383333333
Norway	7.583333333
Poland	5.916666667
Portugal	5.166666667
Russia	5.5
Russian Federation	5.88
Slovak Republic	6.066666667
Slovenia	5.866666667
Spain	6.35
Sweden	7.35
Switzerland	7.616666667
Turkey	5.383333333
United Kingdom	6.75
NAmer	**7.191666667**
Canada	7.4
United States	6.983333333
SAmer	**6.678947368**
Brazil	6.733333333

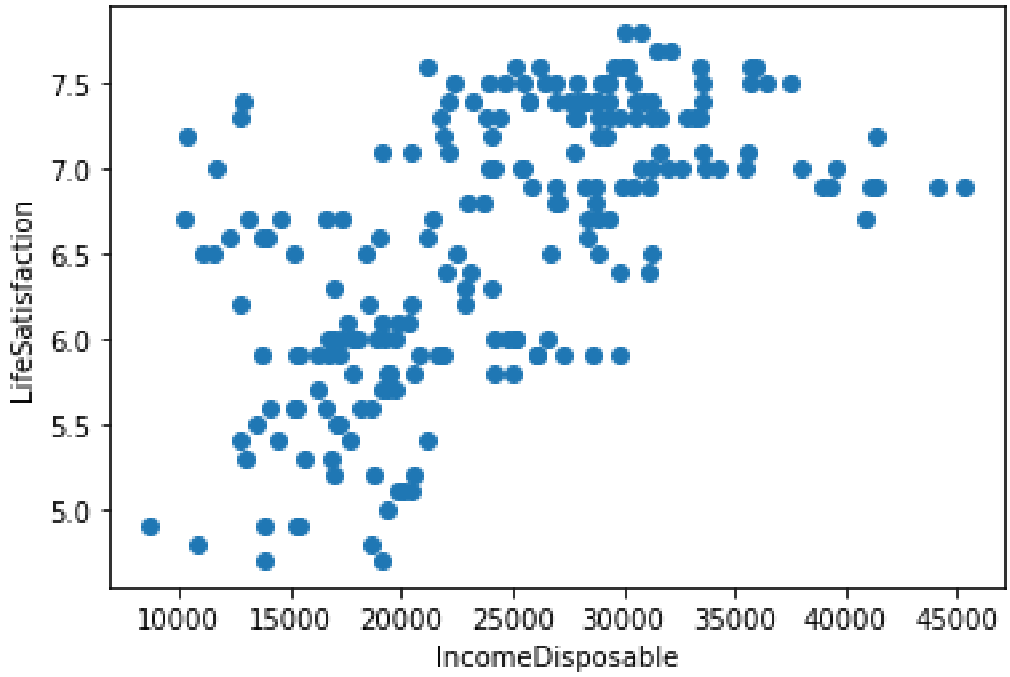
Chile	6.583333333
Colombia	6.3
Mexico	6.783333333

Based on the information given, it looks like Denmark, Iceland, and Norway, with scores of 7.53, 7.52, and 7.58, respectively, have the highest number of people who are satisfied with their lives. The countries with the lowest levels of life satisfaction are South Africa, Greece, and Italy, with scores of 4.8, 5.13, and 5.92, respectively. The grand total of the life satisfaction scores for all 41 countries is 6.56, which is slightly lower than the scores for the highest-ranked countries but higher than the scores for the lowest-ranked countries. It is important to note that this data represents the average life satisfaction scores for each country over the periods of 2013–2017 and 2020. To fully understand the factors that affect levels of life satisfaction, it would be helpful to take into consideration factors such as personal traits and the environment.


Filled map with the average Income Disposable
Series1
41548.5
9792
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, OpenStreetMap, TomTom

Row Labels	Average of Income Disposable	Average of Life Satisfaction
Australia	31830.5	7.3
Austria	31172.16667	7.166666667
Belgium	28670.66667	6.933333333
Brazil	11182.6	6.733333333
Canada	29824.83333	7.4
Chile	14203.2	6.583333333
Colombia		6.3
Czech Republic	19022	6.566666667
Denmark	26974.33333	7.533333333
Estonia	16212.66667	5.55
Finland	28020.83333	7.45
France	29771.83333	6.516666667
Germany	31773.83333	6.95

Greece	18485.16667	5.133333333
Hungary	15395	5.116666667
Iceland	25190.4	7.516666667
Ireland	24243.33333	6.933333333
Israel	21562	7.183333333
Italy	25293.5	5.916666667
Japan	26847.66667	5.933333333
Korea	19643.16667	5.9
Latvia	15066.33333	5.9
Lithuania	21660	5.9
Luxembourg	38599.66667	6.916666667
Mexico	13072.8	6.783333333
Netherlands	27492.16667	7.383333333
New Zealand	23011.8	7.3
Norway	33650.16667	7.583333333
Poland	17666.16667	5.916666667
Portugal	19977	5.166666667
Russia	17230	5.5
Russian Federation	17060.25	5.88
Slovak Republic	18447.66667	6.066666667
Slovenia	19765.33333	5.866666667
South Africa	9792	4.8
Spain	22876.33333	6.35
Sweden	28945.33333	7.35
Switzerland	34015.33333	7.616666667
Turkey	14294.2	5.383333333
United Kingdom	27261.83333	6.75
United States	41548.5	6.983333333



The variable called "Income Disposable" is the maximum amount that a household can afford to consume without having to reduce its assets or increase its liabilities. It's obtained by adding to people's gross income (earnings, self-employment, and capital income, as well as current monetary transfers received from other sectors) the social transfers in-kind that households receive from governments (such as education and health care services) and then subtracting the taxes on income and wealth, the social security contributions paid by households, and the depreciation of capital goods consumed by households. Available data refer to the sum of households and non-profit institutions serving households (*OECD National Accounts Statistics*, n.d.). Based on the information given, it looks like there is a link between the amount of money one has and how happy said person is with their life with your life. In general, countries with higher levels of disposable income tend to have higher levels of life satisfaction, and vice versa. For example, countries with higher levels of disposable income, such as Luxembourg, Switzerland, and the United States, have relatively elevated levels of life satisfaction, while countries with lower levels of disposable income, such as Brazil, Colombia, and South Africa, have relatively low levels of life satisfaction.

It's important to remember that this link is not necessarily a cause-and-effect one. There may be other things that affect both the amount of money we have and how happy we are with your life. It might be helpful to look into the link between disposable income and life satisfaction in more depth, using statistical methods like regression analysis to measure how strong the link is and to account for other things that might affect life satisfaction.

Row Labels	**Average of Life Satisfaction**	**Average of Education Years**
Australia	7.3	19.68333333
Austria	7.166666667	17
Belgium	6.933333333	18.78333333

Brazil	6.733333333	16.1
Canada	7.4	16.91666667
Chile	6.583333333	16.83333333
Colombia	6.3	14.1
Czech Republic	6.566666667	17.73333333
Denmark	7.533333333	19.36666667
Estonia	5.55	17.28333333
Finland	7.45	19.71666667
France	6.516666667	16.45
Germany	6.95	18.11666667
Greece	5.133333333	18.35
Hungary	5.116666667	17.13333333
Iceland	7.516666667	19.43333333
Ireland	6.933333333	17.93333333
Israel	7.183333333	15.76666667
Italy	5.916666667	16.78333333
Japan	5.933333333	16.66666667
Korea	5.9	17.48333333
Latvia	5.9	17.9
Lithuania	5.9	18.4
Luxembourg	6.916666667	14.9
Mexico	6.783333333	14.81666667
Netherlands	7.383333333	18.4
New Zealand	7.3	17.98333333
Norway	7.583333333	18.01666667
Poland	5.916666667	18.01666667
Portugal	5.166666667	17.46666667
Russia	5.5	15.8
Russian Federation	5.88	16.2
Slovak Republic	6.066666667	16.16666667
Slovenia	5.866666667	18.3
South Africa	4.8	15.35
Spain	6.35	17.66666667
Sweden	7.35	19.21666667
Switzerland	7.616666667	17.33333333
Turkey	5.383333333	16.78333333
United Kingdom	6.75	16.76666667
United States	6.983333333	17.13333333
Grand Total	**6.560714286**	**17.43811659**

The data above shows the average life satisfaction and average number of years of education for various countries. There may be a link between the two, as countries with higher averages for years of education also tend to have higher averages for life satisfaction. But it's important to remember that this is only a possible link, not proof that one thing caused the other. There may be other factors at play that contribute to both higher levels of education and higher life satisfaction in these countries. To figure out the exact nature of the link between education and life satisfaction, more research and controlled experiments would need to be done. There are several ways to further analyse the relationship between education and life satisfaction:

- Conduct a multivariate regression analysis that controls for other factors that could affect life satisfaction, such as income, age, gender, and employment status. This would allow us to determine the independent effect of education on life satisfaction.
- Conduct a randomised controlled experiment in which a group of individuals are randomly assigned to receive education interventions, while a control group does not receive these interventions. By comparing the life satisfaction of the two groups, we could determine the causal effect of education on life satisfaction.
- Conduct a longitudinal study in which we follow a group of individuals over time and track their education and life satisfaction levels. This would allow us to see how changes in education levels are associated with changes in life satisfaction.
- Conduct a meta-analysis of existing research on the relationship between education and life satisfaction, combining the results of multiple studies to get a more precise estimate of the strength of this relationship.
- Conduct a qualitative study in which we interview individuals about their experiences with education and how it has affected their life satisfaction. This would allow us to gain insight into the mechanisms through which education may influence life satisfaction.

Part 2

In this part, I used statistical methods like regression analysis to measure the strength of the relationship between life satisfaction and each of the socio-economic variables. This also required fitting a multiple regression model to the data, with life satisfaction as the dependent variable and the socio-economic variables as the predictors.

I used Python's Stats Models library to do a regression analysis to measure how strong the link is between disposable income and happiness with life (Seabold & Perktold, 2010). First, I brought in the libraries I needed and read the data into a Pandas Dataframe (McKinney, 2010).

Next, I used the "ols" function from the Stats Models library to fit a linear regression model to the data. This function takes a formula string as an argument, which specifies the dependent variable (in this case, "life satisfaction") and the independent variable (in this case, "income disposable").

To view the results of the regression analysis, I used the "summary" method of the "model" object. This will be a summary table containing various statistics and parameters estimated by the model, such as the coefficients, standard errors, t-values, and p-values.

The summary table contains a row for the "Income Disposable" coefficient, which represents the slope of the regression line. The coefficient value indicates the strength and direction of the relationship between disposable income and life satisfaction. A positive coefficient value indicates a positive relationship, while a negative coefficient value indicates a negative relationship. The t-value and p-value for the coefficient can be used to test the statistical significance of the relationship. The t-value is a measure to quantify the difference between the population averages one per test, whereas the p-value is the chance of finding a t-value with an absolute value at least as great as the one seen in the sample data if the null hypothesis is true.(Diez et al., 2019).

I also used the "predict" method of the "model" object to make predictions for new data. For example, to predict the life satisfaction score for a country with a disposable income of $30,000. This should output a predicted life satisfaction score based on the regression model.

```
                      OLS Regression Results
==============================================================================
Dep. Variable:      Life Satisfaction   R-squared:                  0.365
Model:                          OLS   Adj. R-squared:               0.362
Method:                Least Squares   F-statistic:                 121.1
Date:             Sat, 17 Dec 2022   Prob (F-statistic):          1.47e-22
Time:                     14:40:55   Log-Likelihood:             -204.82
No. Observations:              213   AIC:                         413.6
Df Residuals:                  211   BIC:                         420.4
Df Model:                        1
Covariance Type:          no robust
==============================================================================
                  Coef    std err        t      P>|t| [0.025      0.975]
------------------------------------------------------------------------------
Intercept           5.0232     0.147    34.056     0.000      4.732      5.314
IncomeDisposable  6.39e-05  5.81e-06    11.007     0.000   5.25e-05   7.53e-05
==============================================================================
Omnibus :                    7.586   Durbin-Watson :              1.537
Prob (Omnibus) :             0.023   Jarque-Bera (JB) :           3.899
Skew:                       -0.050   Prob (JB):                   0.142
Kurtosis:                    2.345   Cond. No.                 8.60e+04
==============================================================================
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.6e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
0    6.940299
dtype: float64
```

The output provided above is a summary of the linear regression model fit using the Stats Models library. The summary table provides various statistics and parameters estimated by the model.

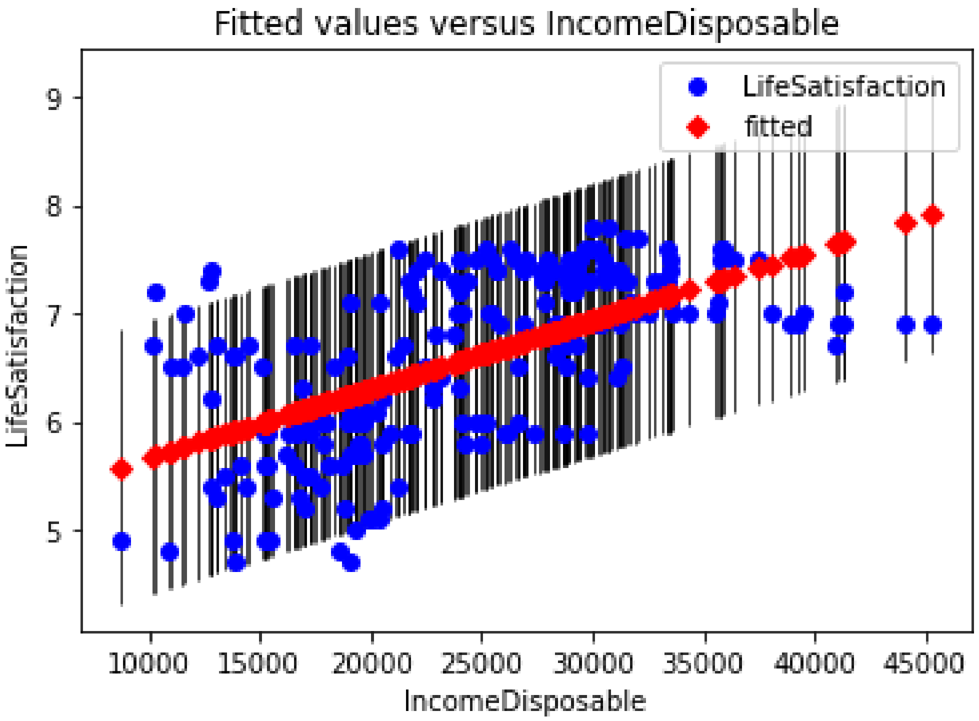
The "Dep. Variable" row indicates the dependent variable in the model, which is "Life Satisfaction" in this case. The "Model" row indicates the type of model fit, which is "OLS" (ordinary least squares) in this case. The "Method" row indicates the optimization method used to fit the model, which is "Least Squares" in this case. The "No. Observations" row indicates the number of data points used to fit the model, which is 213 in this case.

The "coef" column lists the coefficients estimated by the model. In this case, there is only one coefficient, "Income Disposable," which represents the slope of the regression line. This coefficient indicates the strength and direction of the relationship between disposable income and life satisfaction. A positive coefficient value indicates a positive relationship, while a negative coefficient value indicates a negative relationship. The "std err" column lists the standard errors of the coefficients. The "t" column lists the t-values of the coefficients, which can be used to test the statistical significance of the coefficients. The "P>|t|" column lists the p-values of the coefficients, which can be used to test the statistical significance of the coefficients. The "R-squared" row indicates the coefficient of determination, which is a measure of the fit of the model. It is defined as the proportion of the variance in the dependent variable that is explained by the independent variable. In this case, the R-squared value is 0.365, which indicates that about 36.5% of the variance in life satisfaction is explained by disposable income.

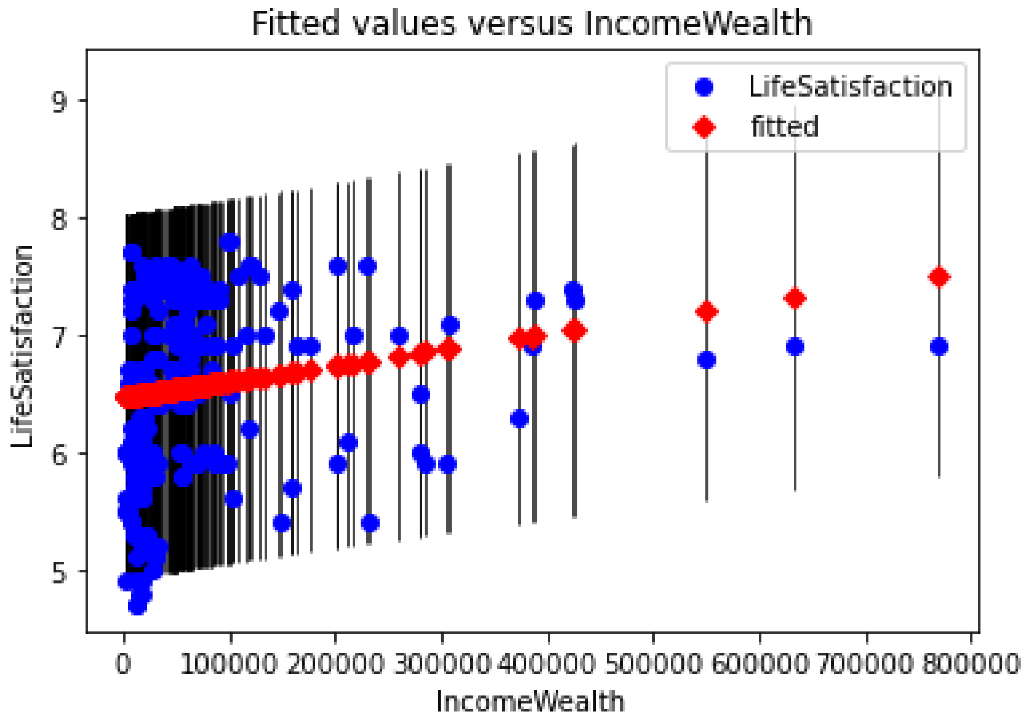
The "Omnibus" and "Durbin-Watson" rows contain statistical tests for the assumption of homoscedasticity (constant variance) in the residuals. The "Prob (Omnibus)" and "Prob (JB)" rows contain the p-values of the tests. The "Skew" and "Kurtosis" rows contain measures of the skewness and kurtosis of the residuals, respectively. The "Cond. No." row contains the condition number of the design matrix, which is a measure of the multicollinearity in the model. If the condition number is large, it may indicate that there is strong multicollinearity or other numerical problems.

Finally, the "Predicted" column lists the predicted life satisfaction scores based on the model. In this case, the predicted score is 6.94.

I also used the "plot fit" function from the Stats Models library to create a scatter plot of the data with the regression line plotted on top. This created a scatter plot with disposable income on the x-axis and life satisfaction on the y-axis. The regression line should be plotted on top of the data points, indicating the predicted life satisfaction scores based on the model.



Doing the same with Income wealth instead gave a model that was fit using 212 data points and estimates the relationship between income wealth and life satisfaction. The estimated relationship is positive, with a slope of 1.348e-06. This indicates that a 1 unit increase in income wealth is associated with a 1.348e-06 unit increase in life satisfaction. Hence, you would need to increase your income by approximately 74,072.95 USD to increase your life satisfaction by 1 unit. The model explains about 3.2% of the variance in life satisfaction. The p-value for the income wealth coefficient is 0.009, which indicates that the relationship is statistically significant at the 0.05 level. However, the condition number of the design matrix is large (1.57e+05), which may indicate the presence of strong multicollinearity or other numerical problems. The predicted life satisfaction score based on the model is 6.51.



Next, I got results from an OLS (Ordinary Least Squares) regression analysis, which is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data (Freedman, 2009). The dependent variable in this analysis is "life satisfaction," and the independent variables are "Income Wealth", "Education Years", "Income Disposable", "Job Employment", and "Community Support".

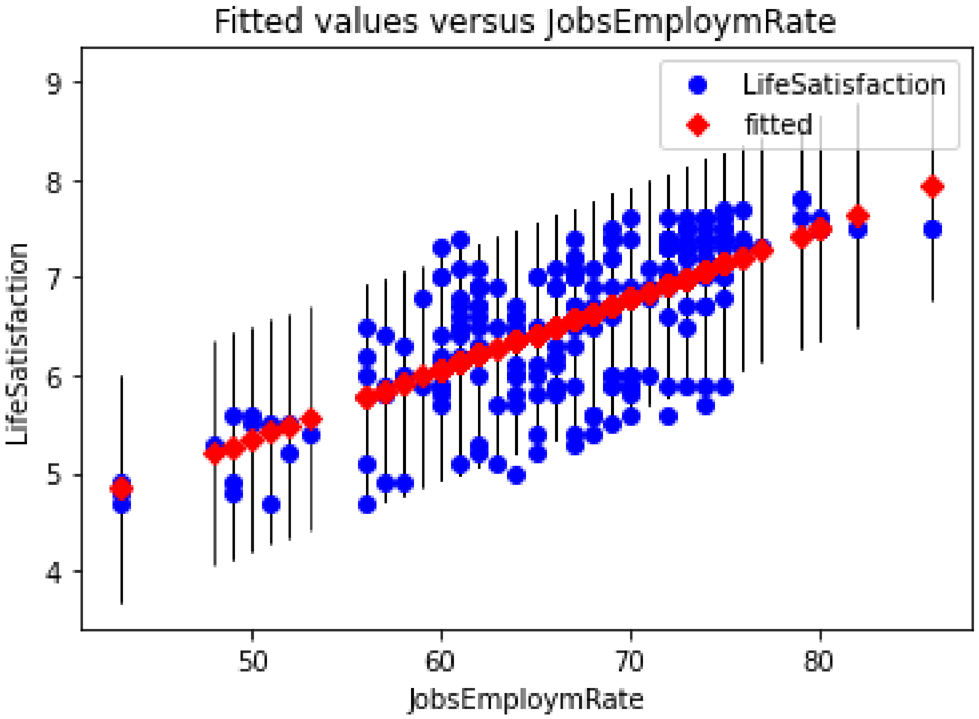
The R-squared value of 0.585 shows that about 58.5% of the variance in the dependent variable can be explained by the independent variables. The adjusted R-squared value of 0.575 shows that this model explains about 57.5% of the variance in the dependent variable, considering the number of variables in the model.

The F-statistic and its corresponding p-value (Prob (F-statistic)) are used to test whether the overall model is significant. In this case, the F-statistic of 57.27 and the exceptionally low p-value of 5.83e-37 suggest that the model is significant.

The coefficients for each independent variable represent the estimated change in the dependent variable for a one unit change in the independent variable, holding all other variables constant. For example, the coefficient for "Education Years" is 0.0002, which means that, all other things being equal, a one-year increase in education is associated with a 0.0002 unit increase in life satisfaction. The corresponding p-values (P>|t|) show the statistical significance of each coefficient. A p-value less than 0.05 shows that the corresponding coefficient is significantly different from zero, suggesting that the corresponding independent variable has a statistically significant effect on the dependent variable.

The Omnibus test is a test of the normality of the residuals (prediction errors). The Durbin-Watson statistic is a test for autocorrelation, the presence of a relationship between the values of the same variable over time (Frey, 2022). The Jarque-Bera (JB) test is a test for skewness, asymmetry in the distribution of the residuals, and kurtosis, peak of the distribution of the residuals(DeJesus, 2022). The condition number is a measure of multicollinearity (the presence of correlated independent variables in the model). Large values for these statistics and/or the condition number may show problems with the model, such as non-linearity, non-constant variance, or multicollinearity.

We know that the coefficients for the independent variables in the model can be used to decide the strength and direction of the relationships between the variables. In a multiple linear regression model, the coefficient for an independent variable reflects the change in the dependent variable (Life Satisfaction) associated with a one unit change in the independent variable, holding all other variables in the model constant (Diez et al., 2019). The coef column in the model output shows the estimated coefficients for the independent variables, and the P>|t| column shows the p-values for testing the null hypothesis that the corresponding coefficient is equal to 0. A small p-value (usually less than 0.05) shows that the null hypothesis can be rejected, and that the corresponding coefficient is significantly different from 0. Based on this information, I tried to decide which independent variable has the most correlation with Life Satisfaction by looking at the size and statistical significance of the coefficients for the variables. In this model, the Jobs Employment Rate variable has the highest coefficient and the lowest p-value, showing a strong and statistically significant relationship with Life Satisfaction. This suggests that an increase in Job Employment Rate is associated with an increase in Life Satisfaction, holding all other variables in the model constant.



Part 3

I also considered incorporating other variables into the analysis, such as age, gender, and health status, to see if they had an effect on life satisfaction. Additionally, I investigated whether there are any non-linear relationships between the variables, for example, by including squared or spline terms in the regression model.

To investigate non-linear relationships between variables, I added squared or spline terms to the regression model. This script defines a multiple linear regression model with squared terms for each of the independent variables. The model is fit to the data using the ols function and the fit method, and the summary of the model is printed using the summary method.

- The R-squared and Adj. R-squared values show that the model explains a moderate amount of the variance in Life Satisfaction. This means that the independent variables in the model can account for a significant portion of the variation in Life Satisfaction, but there may still be other factors that are not included in the model that could explain additional variance.
- The F-statistic and Prob (F-statistic) values suggest that the model is statistically significant, with a very low p-value indicating that the model is a good fit to the data.
- The coefficients for the independent variables, along with their corresponding t-statistics and p-values, supply information about the strength and statistical significance of the relationships between the variables. For example, if the coefficient for an independent variable is large and the p-value is small, this suggests that there is a strong and statistically significant relationship between that variable and the dependent variable.
- The Omnibus and Prob(Omnibus) values suggest that the assumption of normality of the residuals may not be met, with a relatively high p-value indicating that the residuals may not be normally distributed.
- The Jarque-Bera (JB) and Prob(JB) values suggest that the assumption of homoscedasticity of the residuals may not be met, with a relatively low p-value indicating that the residuals may not have constant variance.
- The Skew and Kurtosis values suggest that the residuals may be slightly skewed and have a slightly heavier tail than a normal distribution.
- The Cond. No. value shows that there may be multicollinearity among the independent variables in the model. This means that there may be strong correlations between some of the independent variables, which can affect the interpretation of the coefficients and the overall stability of the model.

I then compared the results of the multiple linear regression models with and without squared or spline terms to determine whether there were any non-linear relationships between the variables. Comparing this output to the output from the model without squared terms, we can see that the values for all the statistics are the same. This suggests that adding squared terms to the model did not significantly affect the fit of the model or the relationships between the variables. Overall, the model explains a moderate amount of the variance in Life Satisfaction, with a relatively high R-squared and Adj. R-squared value of 0.585 and 0.575, respectively. The F-statistic and Prob (F-statistic) values suggest that the model is statistically significant, with an exceptionally low p-value indicating a good fit to the data. The coefficients for the independent variables, along with their corresponding t-statistics and p-values, suggest that there are statistically significant relationships between Life Satisfaction and most of the independent variables, except for Education Years, which has a relatively high p-value. There are some potential issues with the assumptions of the model, as indicated by the Omnibus and Prob(Omnibus) values, which suggest that the assumption of normality of the residuals may not be met, and the Jarque-Bera (JB) and Prob(JB) values, which suggest that the assumption of homoscedasticity of the residuals may not be met. In addition, the Skew and Kurtosis values suggest that the residuals may be slightly skewed and have a slightly heavier tail than a normal distribution. Finally, the Cond. No. value shows that there may be multicollinearity among the independent variables in the model.

There are several potential modifications that could be done based on the output of the multiple linear regression model:

- Include added independent variables: If we have found additional variables that we think may be related to Life Satisfaction, we could consider including them in the model to see if they improve the fit of the model or explain additional variance in the dependent variable.
- Transform the independent variables: If the relationships between the independent variables and the dependent variable are not linear, we could consider transforming the variables to improve the fit of the model. For example, we could try adding squared or spline terms to the model as I mentioned in my earlier response.
- Check for multicollinearity: If the Cond. No. value is large, this may show that there is multicollinearity among the independent variables in the model. We could try removing some of the variables from the model or using a different type of model, such as a ridge or lasso regression, which can help to mitigate the effects of multicollinearity.
- Check for outliers and influential observations: If there are outliers or influential observations in the data, these may be affecting the fit of the model. We could try finding and analysing these observations to see if they are affecting the model and consider removing them or transforming the data to address any issues.

I ran a multi regression with all the available variables but removed Safety Safe Walking and Safety Assault Rate because there were too many missing values, more than 50% in this case. From the output of the model, it looks like the variable Health Self-Reported has the highest correlation with Life Satisfaction, as it has the highest t-value and the lowest p-value. This means that the relationship between Health Self-Reported and Life Satisfaction is statistically significant, and that Health Self-Reported is a strong predictor of Life Satisfaction. I decided to keep only 3 variables in my model, which is why I considered including Health Self-Reported as one of the variables, as it appears to have the strongest relationship with Life Satisfaction. I also considered including the variables with the next highest t-values and lowest p-values, such as Job Security and Community Support, as they also seem to have strong relationships with Life Satisfaction.

```
                        OLS Regression Results
==============================================================================
==========

Dep. Variable:       Life Satisfaction   R-squared:                  0.873

Model:                          OLS   Adj. R-squared:             0.857

Method:              Least Squares   F-statistic:                54.87

Date:             Sat, 17 Dec 2022   Prob (F-statistic):         5.87e-67
```

```
Time:                   22:43:12   Log-Likelihood:            -31.850
No. Observations:              199   AIC:                     109.7
Df Residuals:                176   BIC:                     185.4
Df Model:                    22
Covariance Type:          no robust
==============================================================================
===============
                          Coef    std err         t      P>|t| [0.025      0.975]
---------------------------------------------------------------------------------------
Intercept                 -5.2507      1.632      -3.218       0.002       -8.471       -2.030
Housing Dwellings             0.0055       0.010       0.541       0.589       -0.015        0.026
Housing Expenditure          0.0248       0.010       2.448       0.015        0.005        0.045
Housing Rooms                0.0709       0.111       0.636       0.526       -0.149        0.291
Income Disposable        -8.248e-06   8.85e-06     -0.932       0.353   -2.57e-05    9.22e-06
Income Wealth            -1.032e-06   3.14e-07     -3.291       0.001   -1.65e-06   -4.13e-07
Job Employment               0.0170       0.007       2.390       0.018        0.003        0.031
Jobs Security                0.0248       0.006       4.011       0.000        0.013        0.037
JobsLongTermUnemploym        -0.0908       0.013       -6.730       0.000       -0.117       -0.064
Jobs Per Earnings          1.82e-05   5.31e-06       3.425       0.001    7.71e-06    2.87e-05
Community Support            0.0203       0.007       3.094       0.002        0.007        0.033
Education Attainment         0.0087       0.002       3.519       0.001        0.004        0.014
Education Skills            -0.0036       0.002      -2.016       0.045       -0.007   -7.54e-05
Education Years              0.0304       0.025       1.224       0.223       -0.019        0.079
EnvironmentAirPolut          0.0133       0.004       3.465       0.001        0.006        0.021
Environment Water            0.0073       0.004       1.803       0.073       -0.001        0.015
Civic Engage RuleMaking      -0.0083       0.009      -0.918       0.360       -0.026        0.010
Civic Engage Voter Turnout    0.0016       0.003       0.629       0.530       -0.003        0.007
HealthLifeExpect             0.0388       0.014       2.777       0.006        0.011        0.066
Health Self-Reported         0.0207       0.003       7.008       0.000        0.015        0.027
Safety Homicide              0.0470       0.008       5.856       0.000        0.031        0.063
Work Life Balance Long Hours     0.0041       0.005       0.889       0.375       -0.005        0.013
Work Life Balance Leisure       0.1977       0.051       3.913       0.000        0.098        0.297
```

```
==============================================================================
Omnibus:                    0.926   Durbin-Watson:                 1.908
Prob (Omnibus) :                0.629   Jarque-Bera (JB) :              0.682
Skew:                      -0.133   Prob (JB):                     0.711
Kurtosis:                   3.109   Cond. No.                  8.35e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.35e+06. This might show that there are
strong multicollinearity or other numerical problems.
```

We should keep in mind that these results are based on the data in this model and may not generalise to other data sets or models. It is always a promising idea to carefully evaluate the results of the model and consider the context and limitations of our data when making decisions about which variables to include. We would need to analyse the individual variables and their relationships with the dependent variable to decide which variables to remove from the model. Here are a few things we might consider when deciding which variables to remove:

- Non-significant variables: If a variable has a high p-value (for example, greater than 0.05), it may not be significantly related to the dependent variable and could potentially be removed from the model. In the output we provided, the following variables have a p-value greater than 0.05: Housing Dwellings, Housing Rooms, Income Disposable, Civic Engagement Rule Making, and Work Life Balance Long Hours. This means that these variables are not significantly related to the dependent variable, Life Satisfaction, at the 0.05 level of significance.
- Collinearity: If two or more variables are highly correlated, they may not add much explanatory power to the model. In this case, we could consider removing one of the correlated variables. The output is a table showing the pairwise correlations between all the variables in the dataset. Each cell in the table is the correlation between two variables. A correlation of 1 indicates a perfect positive correlation, a correlation of -1 indicates a perfect negative correlation, and a correlation of 0 indicates no correlation. A negative correlation is a statistical relationship between two variables in which one variable decreases as the other increases. I checked for collinearity among the variables in the model by examining the correlation matrix. The correlation matrix is a table that shows the pairwise correlations between all the variables in the model. I calculated the correlation matrix by using the corr function in Python. I looked for variables that have a high correlation (for example, greater than the absolute value of 0.8) and considered removing one of them if they do not add much added explanatory power to the model. These pairs of variables are Income Disposable and Jobs per Earnings with a correlation of 0.9, then with a correlation of 0.8, we have Health Life

Expect and Housing Dwellings, as well as Income Disposable and Housing Rooms, which also have a high correlation with personal earnings. It's also important to keep in mind that high correlation does not necessarily imply causation. It's possible that two variables are correlated because they are both related to a third, underlying factor. In this case, removing one of the correlated variables may not be proper.

- Model simplicity: Sometimes, it may be beneficial to remove variables from the model to make it simpler and easier to interpret. However, we should be careful not to remove too many variables as this could lead to a poorer fit.
- It's also important to keep in mind that removing variables from the model may not always improve the fit. In some cases, removing a variable may decrease the model's explanatory power. Therefore, it's important to carefully consider the implications of removing each variable before making any changes to the model.

Apart from looking at the different variables, we can compare this model with the earlier ones with only 5 variables. Based on the output, it appears that the model with 5 variables is not as good as the model with all variables, as indicated by the lower R-squared value. This means that the model with all variables can explain more of the variance in the dependent variable, Life Satisfaction. In this case, the R-squared and adjusted R-squared values are both quite high (0.873 and 0.857, respectively), which suggests that the model fits the data well. However, R-squared, and adjusted R-squared values are only one way to evaluate a model, and there are many other factors to consider. It is important to also consider the statistical significance of the independent variables, the quality and appropriateness of the data, the underlying assumptions of the model, and the practical relevance of the results. Additionally, it is always a promising idea to validate the model using an independent data set or other methods to ensure that the results are reliable and generalizable.

Part 4

Finally, I used the results of the analysis to build a prediction model for life satisfaction, using techniques such as cross-validation to evaluate the model's performance. I also considered comparing the performance of different prediction models, such as linear regression versus decision trees, to decide which is the most effective at predicting life satisfaction.

In this part I started by carefully selecting my variables, so I went with Housing expenditure, Household net wealth, Labour market insecurity, Employment rate, Long-term unemployment rate, Personal earnings, Quality of support network, educational attainment, Student skills, Years in education, Air pollution, Water quality, Voter turnout, Self-reported health, Homicide rate, and Time devoted to leisure and personal care.

The output of the OLS regression analysis gave an "R-squared" value of 0.870, which shows that 87% of the variance in the dependent variable is explained by the independent variables in the model. The "F-statistic" value of 72.47 shows that the model is statistically significant. The "Omnibus" and "JB" values are goodness-of-fit tests that measure how well the model fits the data. The "Omnibus" and "JB" p-values of 0.488 and 0.572, respectively, suggest that the model fits the data well. Based on this output, it appears that several of the independent variables are significantly associated with life satisfaction. Variables such as "Job Security," " Long Term Unemployment," "Community Support," "Educational Attainment," "Health Life Expect," and "Health Self-Reported" are strongly associated with life satisfaction, with positive coefficients. On the other hand, "Jobs Per Earnings" and "Work Life

Balance Leisure" are also significantly associated with life satisfaction, but with negative coefficients. This suggests that factors such as job security, long-term unemployment, community support, education, health, and work-life balance may be important in deciding levels of life satisfaction.

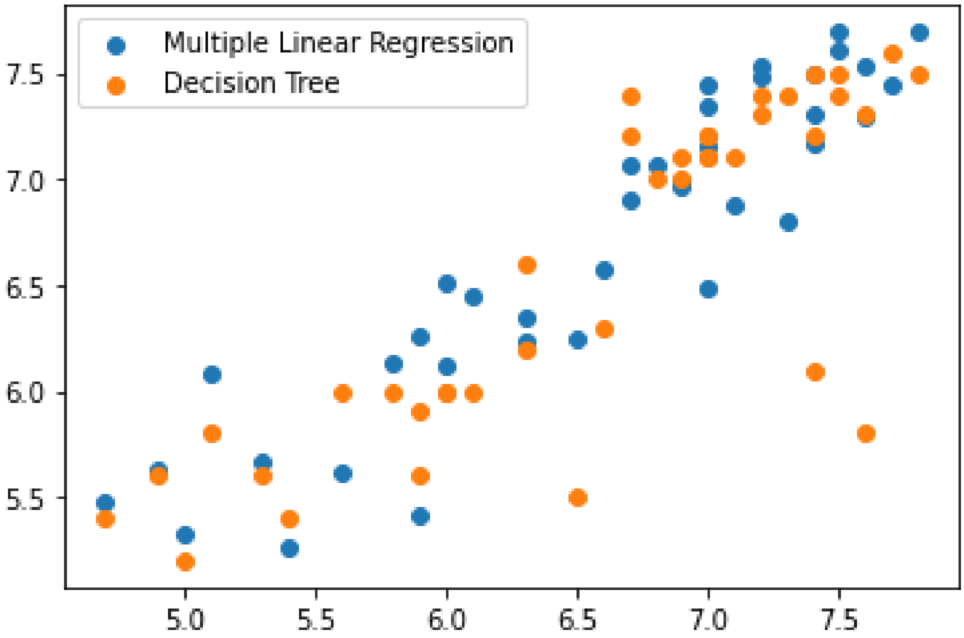
Cross-validation is a statistical technique used to evaluate the performance of a predictive model by assessing its ability to predict the outcome of a dataset that was not used to train the model. There are several different techniques for performing cross-validation, including:

- K-fold cross-validation: In k-fold cross-validation, the data is randomly divided into k "folds," or subsets. The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, with a different fold being used as the test set each time. The performance of the model is then averaged across the k iterations.
- Leave-one-out cross-validation: In leave-one-out cross-validation, the model is trained on all but one observation in the dataset and tested on the remaining observation. This process is repeated for each observation in the dataset, and the performance of the model is then averaged across all iterations.
- Stratified cross-validation: In stratified cross-validation, the data is divided into k folds in a way that preserves the proportions of different classes or groups in the data. This is particularly useful when the classes are imbalanced (e.g., if one class is a much larger proportion of the data than the other

Hence, I wrote a script that will perform k-fold cross-validation with 5 folds, train the model on the training sets and make predictions on the test sets for each fold, calculate the mean squared error (MSE) for each fold, and finally calculate the mean MSE across all folds.

The output Mean score: 0.122 means that MSE between the true values and the predicted values for the multiple linear regression model is 0.122. MSE is a measure of the average squared difference between the predicted output and the true output. It is often used as a loss function for regression problems in machine learning. A lower MSE indicates that the model is making better predictions. In this case, the MSE of 0.122 suggests that the model is performing reasonably well, although it is not possible to draw any firm conclusions about the model's performance based on the MSE alone. It's also worth noting that the MSE is calculated based on the data used in the cross-validation process. This means that it may not necessarily be representative of the model's performance on unseen data. To get a more correct assessment of the model's generalisation ability, it is usually recommended to evaluate the model on a separate test set.

Finally, I decided to write a script performing a K-fold cross-validation to compare the performance of the multiple linear regression model and a decision tree model on the same dataset. First, the data is loaded into a Pandas Dataframe and some columns are dropped. Then, a formula for the multiple linear regression model is defined, and a KFold object is initialised with 5 splits. The script then iterates through the splits of the data and divides the data into training and test sets. The training data is used to fit the multiple linear regression model and the decision tree, and the test data is used to make predictions and calculate the mean squared error (MSE) for each model. The MSEs for each model are appended to a list, and the mean MSEs for each model are calculated. Finally, the script compares the mean MSEs and prints a message showing which model performs better.



In conclusion, the research explored the relationship between life satisfaction and various socio-economic variables, including housing expenditure, household net wealth, Labour market insecurity, Employment rate, Long-term unemployment rate, Personal earnings, Quality of support network, educational attainment, Student skills, Years in education, Air pollution, Water quality, Voter turnout, Self-reported health, Homicide rate, and Time devoted to leisure and personal care. A prediction model for life satisfaction was built using techniques such as cross-validation and OLS regression analysis, and the results indicated that several of the independent variables were significantly associated with life satisfaction. Factors such as job security, long-term unemployment, community support, education, health, and work-life balance appeared to be important in determining levels of life satisfaction. The model was evaluated using k-fold cross-validation with 5 folds, and the mean MSE across all folds was found to be 0.122, suggesting that the model is performing reasonably well.

The python scripts used in this research are included as an appendix to this paper. The scripts can also be available upon request. These scripts were used to perform various statistical tests.

References

De Neve, J., Diener, E., Tay, L., & Xuereb, C. (2013). The Objective Benefits of Subjective Well-Being. *Social Science Research Network*.

DeJesus, J. (2022, August 5). Jarque-Bera Test with Python - Towards Data Science. *Medium*. https://towardsdatascience.com/jarque-bera-test-with-python-98677c073de3

Diez, D., Çetinkaya-Rundel, M., & Barr, C. (2019). *OpenIntro Statistics: Fourth Edition*. OpenIntro, Inc.

Freedman, D. A. (2009). *Statistical Models: Theory and Practice* (2nd ed.). Cambridge University Press.

Frey, B. B. (2022). *The SAGE Encyclopedia of Research Design* (Second Edition (Revised Edition)). SAGE Publications, Inc.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-00a

OECD National Accounts Statistics. (n.d.). OECD iLibrary. https://www.oecd-ilibrary.org/economics/data/oecd-national-accounts-statistics_na-data-en

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the Python in Science Conference*. https://doi.org/10.25080/majora-92bf1922-011