Happiness_Project_Final

December 14, 2017

1 Kaela Nelson

1.1 Background

Each year, the Gallop World Poll ranks countries based on their happiness. This is a relevant research topic that inform future leaderson how they can adjust policies to improve the overall well-being of their country. Since happiness is a difficult quantity to measure, I am interested in analyzing how factors such as average income, education, and network support levels contribute to a country's happiness level. My research questions are: "Does health, personal care, and income level affect a country's happiness?", "Is there a relationship between education and life satisfaction?", and "What are the main features that distinguish happy countries from unhappy countries?". To investigate these questions, I will analyze a data set collected by The Organisation for Economic Co-operation and Development (stats.oecd.org), which contains features such as life expectancy, job security, and life satisfaction. I will use the life satisfaction feature as my measurement for happiness.

```
In [55]: import numpy as np
         import pandas as pd
         from matplotlib import pyplot as plt
         import statsmodels.api as sm
         import importlib

         import feature_engineering
         importlib.reload(feature_engineering) #necessary to update function
         from feature_engineering import create_features
         plt.rcParams["figure.dpi"] = 200

         import personal_care_earnings_subplot
         importlib.reload(personal_care_earnings_subplot) #necessary to update function
         from personal_care_earnings_subplot import plot_earnings_vs_care

         import health_network_subplot
         importlib.reload(health_network_subplot) #necessary to update function
         from  health_network_subplot import plot_health_vs_network

         import plot_leisure_earnings_educ
         importlib.reload(plot_leisure_earnings_educ) #necessary to update function
```

```
from  plot_leisure_earnings_educ import create_leisure_earnings_educ

import earningscare_subplot
importlib.reload(earningscare_subplot) #necessary to update function
from  earningscare_subplot import plot_earningscare
```

1.2 Data Collection and Feature Engineering

The data set from OECD is called the "Better Life Index 2016". It contains government collected data (i.e. from Labour Force Statistics database and Statistics New Zealand) as well as survey based data for 38 different countries. The categories include life expectancy, job security, housing expenditure, etc. I was able to download this data set from stats.oecd.org as a csv file. I removed the columns and rows with NaN values.

```
In [50]: #read in data set
         oecd16 = pd.read_csv("well_being_2016.csv")
         #remove nan's and save cleaned data
         oecd16.drop(["Unnamed: 1", "Unnamed: 2"], axis=1 , inplace=True)
         oecd16.dropna(axis=0, inplace=True)
         oecd16.to_csv("oecd16_cleaned.csv")
```

I removed the following columns from the data set: 'rooms per person', 'student skills', 'educational attainment', 'air pollution', 'Stakeholder engagement for developing regulations', and 'water quality'. I removed those columns based these three initial assumptions:

1. I am assuming that Housing expenditure, household net adjusted disposable income, and household net financial wealth can adequately describe an average family's financial well being. Thus, I removed 'rooms per bedroom' because I am assuming it is a subset of household expenditures.

2. I am mainly interested in years in education as a measurement of eduation level. Thus I removed "education attainment".

3. I am less interested in evaluating environmental factors for the purpose of this project, thus I removed air pollution and water quality.

4. I am less interested in evaluating 'Stakeholder engagement for developing regulations' for the purpose of this project.

```
In [51]: oecd16.drop(["Rooms per person","Student skills", "Educational attainment",
                      "Air pollution","Water quality",
                      "Stakeholder engagement for developing regulations"],
                     axis=1, inplace=True)
```

Since I am interested in how happier countries differ from unhappy countries, I created a new column that binned countries into "low life satisfaction"(4.8,5.8), "medium life satisfaction"(5.8,6.7), and "high life satisfaction"(6.7,7.6) categories. This will help me visualize how features affect each group differently. Note that, all countries fall within a range of (4.8,7.6) out of a (0,10) scale based on the average score of a survey asking people to rank their life satisfaction. I am also looking to analyze the direct relationship between life satisfaction and (personal

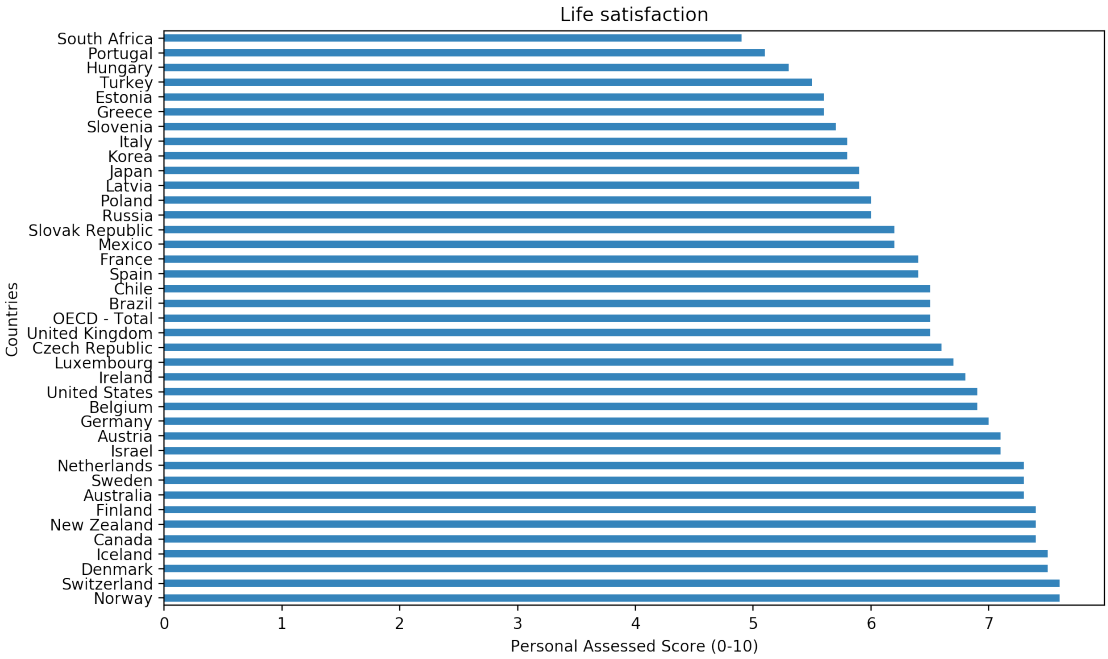
care)*(personal earnings), as well as with (personal earnings)*(years in education), so I added those columns to my data set. The rest of my columns have accurate measurements, and so I leave them alone. This is done in my function create_features.

```
In [57]: oecd16, low16, medium16, high16 = create_features(oecd16)
```

1.3 Data Visualization and Analysis

I initially visualize life satisfaction rank among all 38 countries. According to the OECD metadata, life satisfaction is a feature that measures people's evaluation of their life as a whole. Specifically, "it is a weighted-sum of different response categories based on people's rates of their current life relative to the best and worst possible lives for them on a scale from 0 to 10, using the Cantril Ladder" (stats.oecd.org).

```
In [5]: #sort barchart via life satisfaction, set index to country, plot
        sorted16 = oecd16.sort_values(by=['Life satisfaction'], ascending=False)
        sorted16.index = sorted16["Country"]
        sorted16['Life satisfaction'].plot(kind="barh", alpha=0.9, figsize=(10,6))
        plt.title("Life satisfaction")
        plt.xlabel("Personal Assessed Score (0-10)")
        plt.ylabel("Countries")
        plt.tight_layout()
        plt.show()
```



Notice that South Africa has the lowest satisfaction, while Norway and Switzerland are tied for the highest satisfied countries. I now want to visualize if there is a correlation between life satisfaction, and the following features:"time devoted to personal care", "personal earnings", and

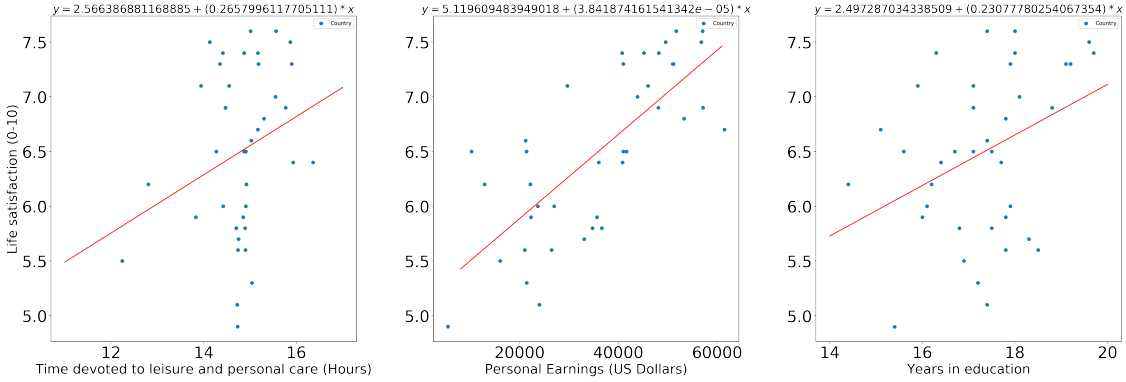
years in education". I use an ordinary least squares (OLS) model to create a line of best fit for all of my visualizations.

```
In [54]: """Time devoted to leisure and personal care: This indicator measures the amount
         of minutes (or hours) per day that, on average, full-time employed people spend
         on leisure and on personal care activities.

         Personal earnings:This indicator refers to the average annual wages per full-
         time equivalent dependent employee

         Years in education:This indicator is the average duration of education in
         which a 5 year old child can expect to enrol during his/her lifetime until
         the age of 39.

         Plots the following along life satisfaction: Time devoted to leisure and
         personal care, Personal earnings, Years in education
         """
         create_leisure_earnings_educ(oecd16)
```

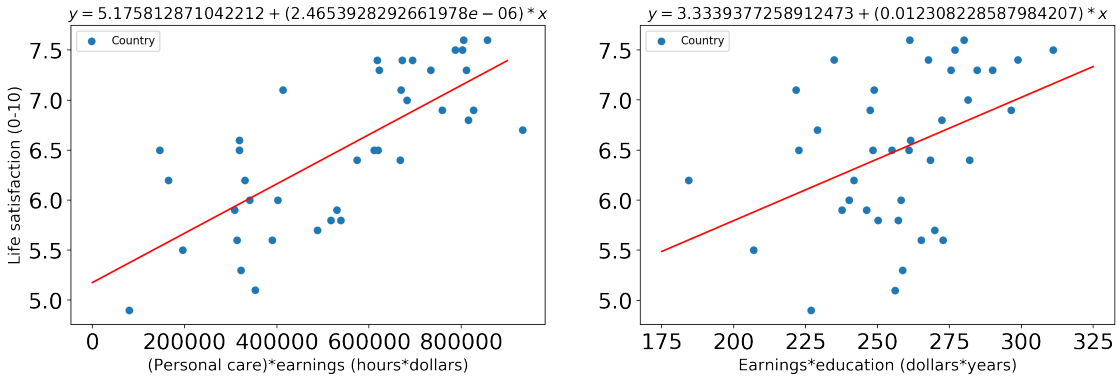


It is important to note that these three features are on different scales, and so the middle visualization will look different if put on the same scale as the first and third plots.

From this, we can see from the R^2 values that a more distinct positive correlation between life satisfaction and personal earnings (R^2: .550), while years in education (R^2 0.146) seems to be more scattered. I initially expected the model to capture a more defined relationship between years in education and life satisfaction. However, there could very well be other ommitted variables that affect life satisfaction, aside from years in education. Note that all three features have low pvalue scores (Time devoted to personal care:0.095, Personal Earnings: 0.000, Years in education:0.016), so they still seem to be statistically significant enough to not omit them.

From these scatter plots, I am not implying that personal care, personal earnings, or years in education alone are the only indicators of happiness - I am analyzing each of their correlation behavior with happiness. I now investigate the relationship between personal care and personal earnings, as well as years in education with personal earnings even further.

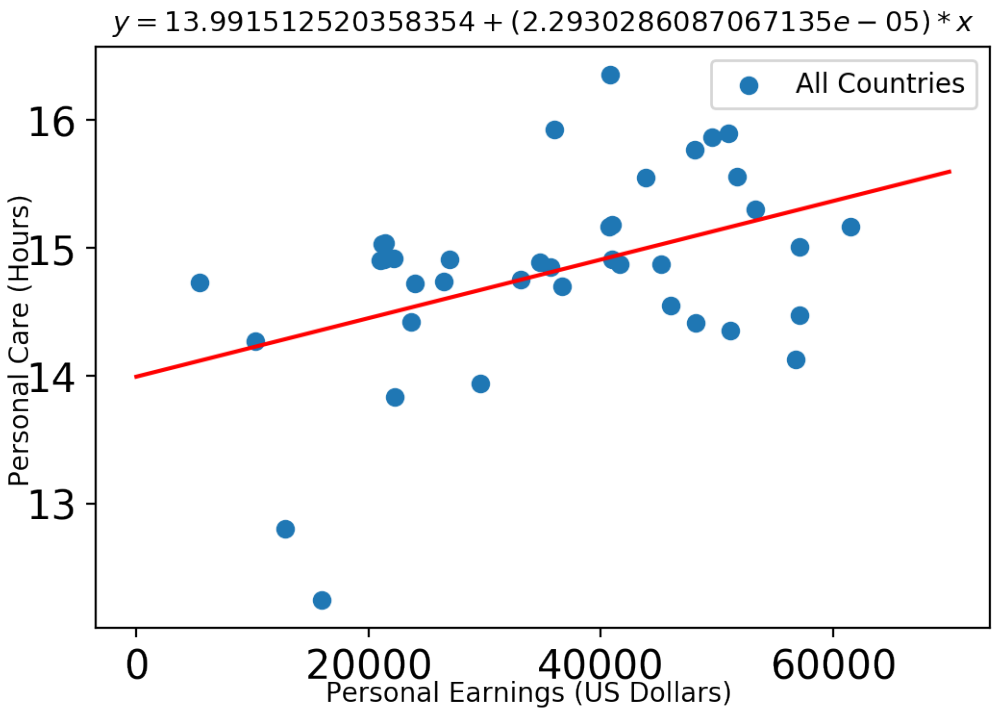
```
In [59]: """First look at combined relationship of earnings with personal care
         and earnings with education"""
         plot_earningscare(oecd16)
```



Notice that earnings*education in relation to life satisfaction has a lower R^2 value (0.176). However, its pvalue (0.008) is still low enough to imply it has statistical significance. In addition, (personal care)*personal earnings has a positive correlation with life satifaction (R^2: 0.543) and a pvalue of 0.000. So we will further analyze personal care along with personal earnings for low, medium, and highly satisfied countries.

```
In [47]: """Personal Earnings: This indicator refers to the average annual wages
         per full-time equivalent dependent employee"""
         #create subplot
         fig, ax1 = plt.subplots(1, 1, sharex='col', sharey='row', figsize=(6,4))
         X = oecd16["Personal earnings"]
         y = oecd16["Time devoted to leisure and personal care"]
         X = sm.add_constant(X)
         #fit model
         results1 = sm.OLS(y, X).fit()
         x_lim = np.linspace(0, 70000)
         #plot line of best fit
         ax1.plot(x_lim, results1.params[0] + results1.params[1]*x_lim, color="r")
         #plot data
         ax1.scatter(oecd16["Personal earnings"],
                     oecd16["Time devoted to leisure and personal care"],
                     label="All Countries")
         ax1.legend(loc="upper right")

         #set appropriate paramers
         ax1.set_title(f"$y = {results1.params[0]} + ({results1.params[1]})*x$",
                       fontsize = 10.5)
         fig.text(0.5, 0.04, 'Personal Earnings (US Dollars)', ha='center',
                  va='center', fontsize = 10)
         fig.text(0.06, 0.5, 'Personal Care (Hours)', ha='center', va='center',
                  rotation='vertical', fontsize = 10)
         plt.show()
```

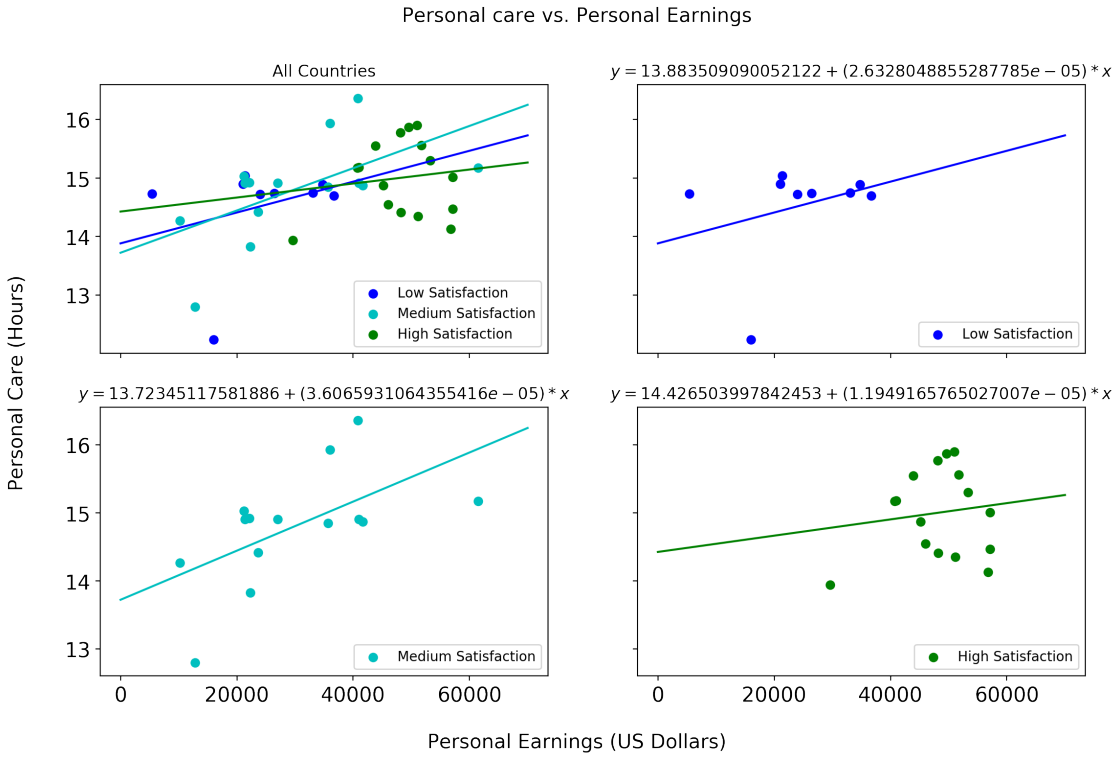


Personal earnings in relation to personal care has a pvalue of 0.006, and an R^2 value of 0.188. The lowest outlier country is Mexico (medium satisfaction), and the second lowest outlier country is Turkey (low satisfaction). Since these outliers are categorized in different levels of life satisfaction, according to my binning system, we will investigate this further by analyzing the trend separately for low, medium, and highly satisfied countries.

```
In [9]: """Plots Personal earnings vs. Time devoted to personal care:
        Personal Earnings: This indicator refers to the average annual wages per
        full-time equivalent dependent employee.

        Time devoted to personal care: This indicator measures the amount of
        minutes (or hours) per day that, on average, full-time employed people
        spend on leisure and on personal care activities.

        In order from left to right, top to bottom:
        First plot contains all three categories on the same plot, with all of their
        lines of best fit from OLS model.
        Second plot shows low life satisfaction countries with line of best fit
        from OLS model.
        Third plot shows low life satisfaction countries with line of best fit
        from OLS model.
        Fourth plot shows low life satisfaction countries with line of best fit
        from OLS model."""
        plot_earnings_vs_care(low16, medium16, high16)
```



Note that the low satisfied countries have a flater trend in regards to personal care vs. personal earnings (personal earnings has pvalue of 0.429), and the regression line is affected by the outlier country, Turkey. Turkey is described as a "developing" country - not quite 3rd world or 1st world. It could be the case that people in Turkey are less likely to spend time on personal care than the other countries with similar personal earnings.

It is even more interesting that within highly satisfied countries (outlier Israel) personal earnings are less statistically significant in relation to personal care (personal earnings has pvalue of .619) as it is in medium satisfied countries (personal earnings has pvalue of .028). From this visualization, we see that medium satisfied countries are more likely to spend more time for personal care as they earn more. This seems to imply that there is might be a certain threshold where wealth stops affecting the amount of leisure time one spends.

I now investigate other research question, and am interested in how health and quality of network support are correlated in relation to life satisfaction.

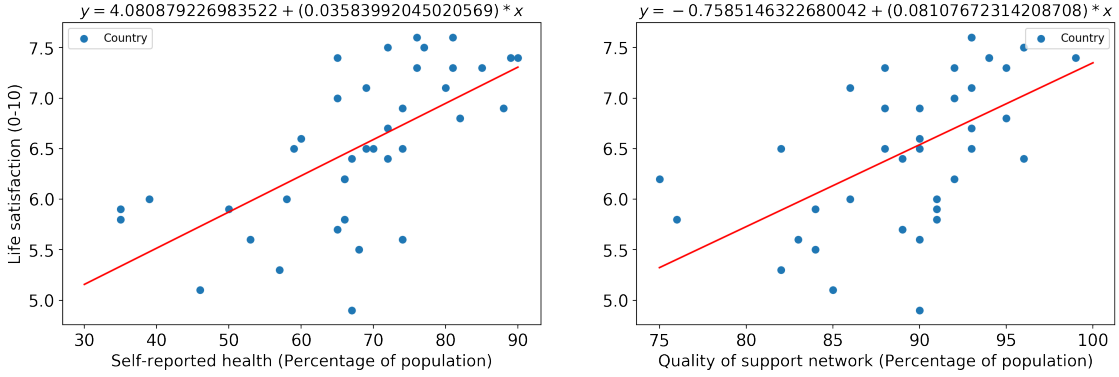
```
In [10]: #create subplot
         fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(17,5))
         X1 = oecd16["Self-reported health"]
         y1 = oecd16["Life satisfaction"]
         #fit OLS model
         X1 = sm.add_constant(X1)
         results1 = sm.OLS(y1, X1).fit()

         #set parameters and plot Self-reported health vs. Life satisfaction
         x_lim1 = np.linspace(30, 90)
```

```
ax1.plot(x_lim1, results1.params[0] + results1.params[1]*x_lim1, color="r")
ax1.scatter(oecd16["Self-reported health"], oecd16["Life satisfaction"],
            label="Country")
ax1.set_xlabel("Self-reported health (Percentage of population)",
               fontsize = 14.5)
ax1.legend()
ax1.set_ylabel("Life satisfaction (0-10)",fontsize = 14.5)
ax1.set_title(f"$y = {results1.params[0]} + ({results1.params[1]})*x$",
              fontsize = 14.5)

#set parameters and plot Quality of support network vs. Life satisfaction
X2 = oecd16["Quality of support network"]
y2 = oecd16["Life satisfaction"]
X2 = sm.add_constant(X2)
#fit model
results2 = sm.OLS(y2, X2).fit()
x_lim2 = np.linspace(75, 100)
ax2.plot(x_lim2, results2.params[0] + results2.params[1]*x_lim2, color="r")
ax2.scatter(oecd16["Quality of support network"], oecd16["Life satisfaction"],
            label="Country")
ax2.legend()
ax2.set_xlabel("Quality of support network (Percentage of population)",
               fontsize = 14.5)
ax2.set_title(f"$y = {results2.params[0]} + ({results2.params[1]})*x$",
              fontsize = 14.5)

plt.show()
```



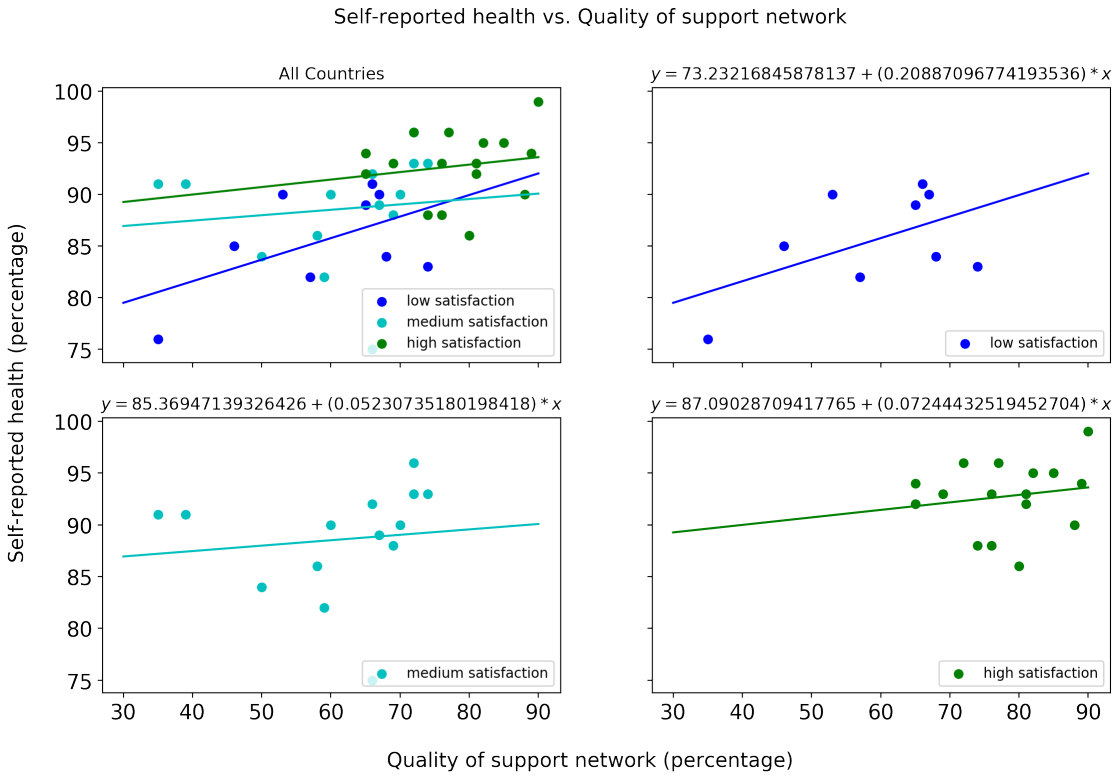
Because both self reported health(pvalue: 0.000) and quality of network support(pvalue:0.000) both have very low pvalues, we will investigate both features furthor in how they relate to eachother in the different life satisfaction groups.

In [11]:
```
"""Plots Self-reported health vs. Quality of support network:
Self-reported health : This indicator refers to the percentage of the
population aged 15 years old and over who report good or better health.
```

```
Quality of support network: The indicator is based on the question: If
you were in trouble, do you have relatives or friends you can count on
to help you whenever you need them, or not? and it considers the
respondents who respond positively.

First plot contains all three categories on the same plot, with all of their
lines of best fit from OLS model.

Second plot shows low life satisfaction countries with line of best fit
from OLS model.
Third plot shows medium life satisfaction countries with line of best fit
from OLS model.
Fourth plot shows high life satisfaction countries with line of best fit
from OLS model."""
plot_health_vs_network(low16, medium16, high16)
```



For low life satisfaction, quality of network support has a pvalue of .144 and R^2 0.279. For medium life satisfaction, quality of network support has a pvalue of 0.687 and R^2 0.014. For high life satisfaction, quality of network support has a pvalue of 0.534 and R^2 0.028.

Note that the R^2 value (0.279) of network of support quality vs. self-reported health is the highest in low satisfied countried. Also, quality of network support vs. life satisfaction has a lower pvalue (.144), implying this feature is more likely to affect lower satisfied countries than medium and high satisfied countries.

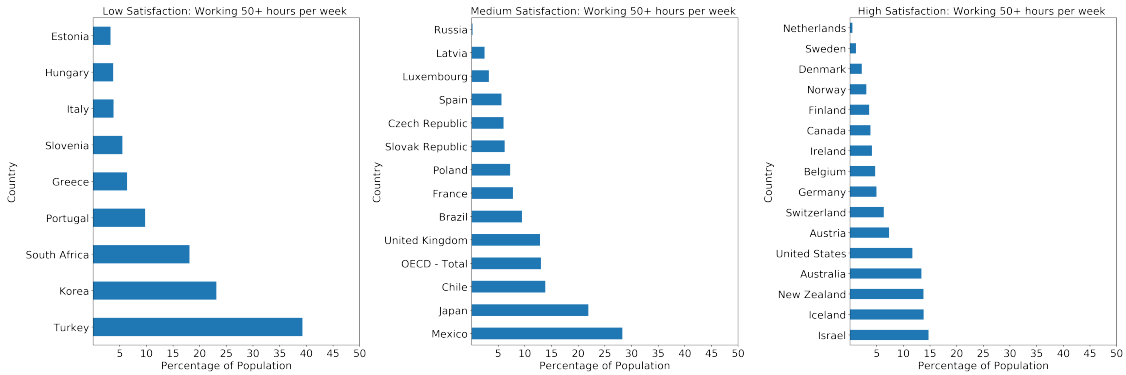
For my last research question, we analyze how unemploymnet rates and work hours/conditions affect each level of life satisfaction of countries.

In [39]:
```
"""Employees working very long hours: This indicator measures the proportion of
dependent employed whose usual hours of work per week are 50 hours or more.
"""
#create subplots
fig, axes = plt.subplots(1, 3, figsize=(30,10))

#sort the barcharts by Employees working very long hours
sort1 = low16.sort_values(by=['Employees working very long hours'],
                          ascending=False)
sort2 = medium16.sort_values(by=['Employees working very long hours'],
                             ascending=False)
sort3 = high16.sort_values(by=["Employees working very long hours"],
                           ascending=False)

sort_ls = [sort1, sort2, sort3]
title_ls = ["Low Satisfaction: Working 50+ hours per week ",
            "Medium Satisfaction: Working 50+ hours per week ",
            "High Satisfaction: Working 50+ hours per week "]
index = ["Employees working very long hours", "Employees working very long hours",
         "Employees working very long hours"]

#create 3 subplots, for each satisfaction group
for sort, ax, title, i in zip(sort_ls, axes, title_ls, index):
    sort.index = sort["Country"]
    sort[i].plot(kind="barh",ax=ax)
    ax.set_title(title, fontsize=20)
    ax.set_xlabel("Percentage of Population",fontsize = 20)
    ax.set_ylabel("Country",fontsize = 20)
    if ax != ax1:
        ax.set_xticks([(i+1)*5 for i in range(10)])
plt.rcParams['xtick.labelsize']=50
plt.rcParams['ytick.labelsize']=50
plt.tight_layout()
plt.show()
```



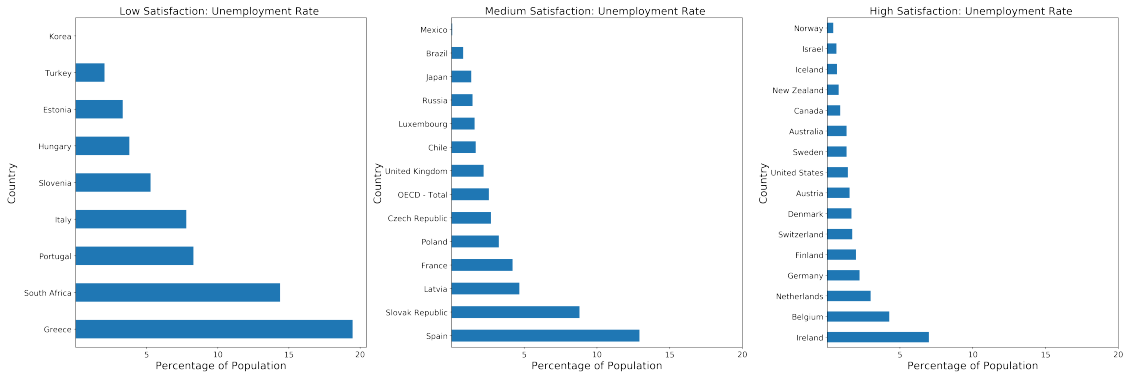
In [46]:
```
"""Long-term unemployment rate: This indicator refers to the number of persons
who have been unemployed for one year or more as a percentage of the labour force
(the sum of employed and unemployed persons).
"""
#create subplot
fig, axes = plt.subplots(1, 3, figsize=(30,10))

#sort according to unemployment rate
sort1 = low16.sort_values(by=['Long-term unemployment rate'], ascending=False)
sort2 = medium16.sort_values(by=['Long-term unemployment rate'], ascending=False)
sort3 = high16.sort_values(by=["Long-term unemployment rate"], ascending=False)

sort_ls = [sort1, sort2, sort3]
title_ls = ["Low Satisfaction: Unemployment Rate ",
            "Medium Satisfaction: Unemployment Rate ",
            "High Satisfaction: Unemployment Rate "]
index = ["Long-term unemployment rate", "Long-term unemployment rate",
         "Long-term unemployment rate"]

#plot three plots by life satisfaction level
for sort, ax, title, i in zip(sort_ls, axes, title_ls, index):
    sort.index = sort["Country"]
    sort[i].plot(kind="barh",ax=ax)
    ax.set_title(title, fontsize=20)
    ax.set_xlabel("Percentage of Population",fontsize = 20)
    ax.set_ylabel("Country",fontsize = 20)
    if ax != ax1:
        ax.set_xticks([(i+1)*5 for i in range(4)])

#set appropriate parameters
plt.rcParams['xtick.labelsize']=15
plt.rcParams['ytick.labelsize']=15
plt.tight_layout()
plt.show()
```



Note that from these bar charts, countries with higher satisfaction have on average a lower percentage of employees working more than 50 hours per week, and a lower percentage of the unemployed employees.

Lastly, we run an ordinary least squares regression model on the combined features that we analyzed above to see their significance to life satisfaction.

```
In [14]: features= ["Self-reported health", "Years in education", "Quality of support network"
                    "Personal earnings", "Time devoted to leisure and personal care"]
         X = oecd16[features]
         y = oecd16["Life satisfaction"]
         X = sm.add_constant(X)
         #fit model
         results = sm.OLS(y, X).fit()
         #print pvalues
         print ("Pvalues:\n", results.pvalues)
```

```
Pvalues:
 const                                        0.238980
Self-reported health                         0.016102
Years in education                           0.425995
Quality of support network                   0.392765
Personal earnings                            0.001411
Time devoted to leisure and personal care    0.760905
dtype: float64
```

From this model, we can see that time devoted leisure and personal care is the least statistically significant, and personal earnings seem to be the most statistically significant. In our model that uses personal care as our only feature, which implied it was statistically significant. This tells us that personal care is less statistically significant than the other features. Based on these results, I would ommit the following variables as they relate to predicting life satisfaction/happiness: time devoted to personal care, years in education, and quality of support network.

1.4 Conclusion

Based on the visualizations, we can conclude that personal earnings have little correlation to personal care in highly satisfied countries. Also, years in education is less likely to impact life satisfaction than the other examined feathtures. Lastly, higher satisfied countries have lower percentages of the population that are unemployed, or over worked in their jobs. These are a few observations about traits of happy and unhappy countries.