Depression and life satisfaction among youths in Oslo

Roy Lachica

March 3, 2020

Coursera IBM Data Science Professional capstone project

1 Introduction

1.1 Background

In 2017, an estimated 264 million people in the world experienced depression [1]. The Nordic countries top the polls as the happiest in the world [2], at the same time these countries also have some of largest numbers of young who are not happy with their life [3].

1.2 Problem

Risk factors and determinants for depression are already widely known in the field of clinical psychology but in this report we will take a closer look at intra city differences focusing on boroughs in Oslo Norway and explore various features and their correlations with depression and life satisfaction among the young.

The project had 6 research goals:

1. Find out if income is associated with depression or life satisfaction.
2. Find out if education is associated with depression or life satisfaction.
3. Find out if the prevalence of social security welfare benefits is associated with depression or life satisfaction.
4. Find out if the number of venues in a borough is associated with depression of life satisfaction.
5. Find out if the share of child protection services cases in a borough is associated with depression of life satisfaction.
6. Find out if nature closeness is associated with depression or life satisfaction.

1.3 Interest

The target audience of this report is public health services in Norway, Oslo city government as well as others interested in the topic of depression and mental health.

2 Data

2.1 Data sources

We used three different sources of data.

1. Foursquare API [4] to get the number of venues for all boroughs.

2. Oslo municipality statistics database [5] with access to aggregated datasets on population income, education, social security and depression by borough.
   We define the following abbreviated features which will be used throughout our IBM Watson Studio Jupyter notebook[8]:
   a. PctTopIncome: Share of population with top income interval (800 000+ NOK)
   b. PctLowIncome: Share of population with low early wages EU-scale.
   c. PctTopEducation: Share of population with top education (higher university degree)
   d. PctSocialSecurity: Share of population age 50-59 on social security welfare benefits.
   e. PctChildProt: Share of children involved in child protection services.
   f. PctLifeSatis: Averge satisfaction rating of own health among the youth.
   g. PctDepression: Share of youth population with depressive symptoms.

3. An array of manually coded nature closeness levels and borough locations to be able to analyse nature closeness and connect Foursquare venues to boroughs as no public datasets for borough geo positions or nature closeness could be found online.

2.2 Definitions

Youths: Youths are here defined as participants in the survey "Young In Oslo 2015" [7] which consist of pupils in 8th, 9th and 10th grade at the secondary school and 1st, 2nd and 3rd levels in the upper secondary school in Oslo [7].

2.3 Data wrangling

The municipality statistics datasets are available as auto generated downloadable Excel files. A report wizard [5] provided by the municipality website was used to setup datasets with the data we needed such as the most recent years, the correct variables, combining both men and women etc. with the correct filters and groupings etc. Some of the numbers were represented as percentages while others not so in order to normalize the data we pulled other datasets on population numbers to be able to calculate percentages.

Using Pandas read_excel we loaded datasets and merged them into a single dataframe for analysis. Fortunately the data from the Oslo municipality statistics database has already gone through considerable quality checks so there was not a need to clean the data. We did however need to rename some boroughs that had minor naming inconsistencies. Boroughs did not have a unique code so we had to marge datasets on borough name. A lot of work went into repeated steps such as:

- Skipping the correct number of intro rows in the Excel file.
- Setting meaningful column headers.
- Removing the “Borough ”-prefix in front of all borough names.
- Stripping whitespace on borough column.
- Indexing the borough column to be able to merge datasets using the borough as the key.

3 Methodology

3.1 Process

The following list is an overview of our analysis process:

1. Load and prepare necessary data.
2. Display depression levels on map in order to get a geographical overview that could guide further research and if necessary focus further analysis on specific boroughs.
3. Create a correlation matrix to get an overview of associations between features.
4. Investigate and validate top correlations.
5. Cluster boroughs in order to build an overview of similar boroughs in terms of depression and life satisfaction that would be simple enough for communication purposes.

3.2 Analysis

We first visualize depression and life satisfaction using matplotlib horisonal barchart plots.

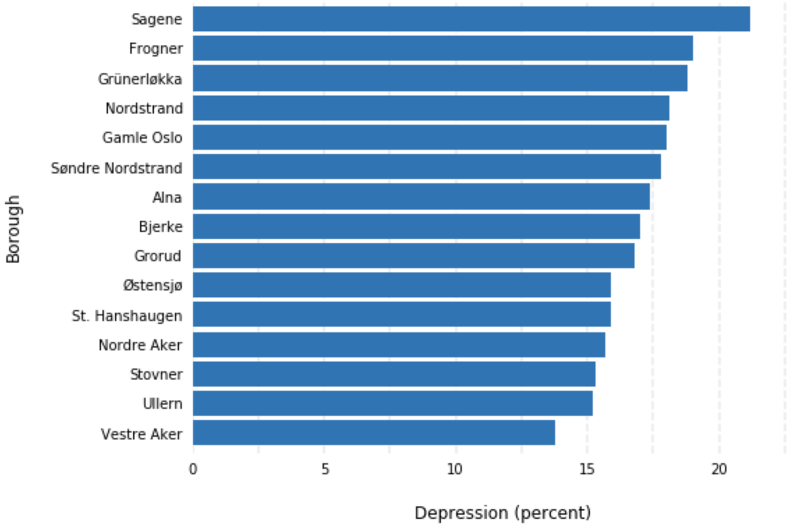

Fig.1. Depression levels among the young in Oslo

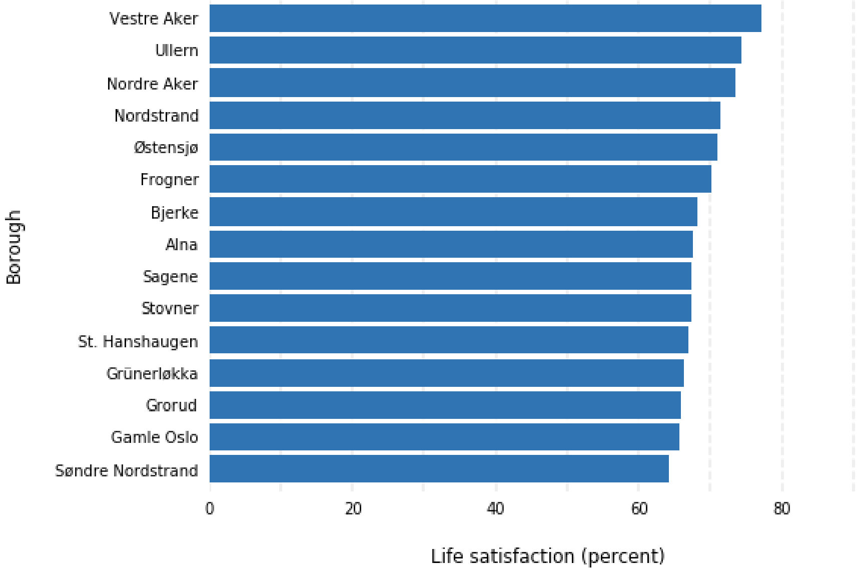

Fig.2. Life satisfaction levels among the young in Oslo

By using the Folium map visualization python library and plotting boroughs and depression level (Green: low depression, yellow: mid-level, red: high levels of depression) we can see one borough “Sagene” slightly north of the city centre struggling with high depression levels.

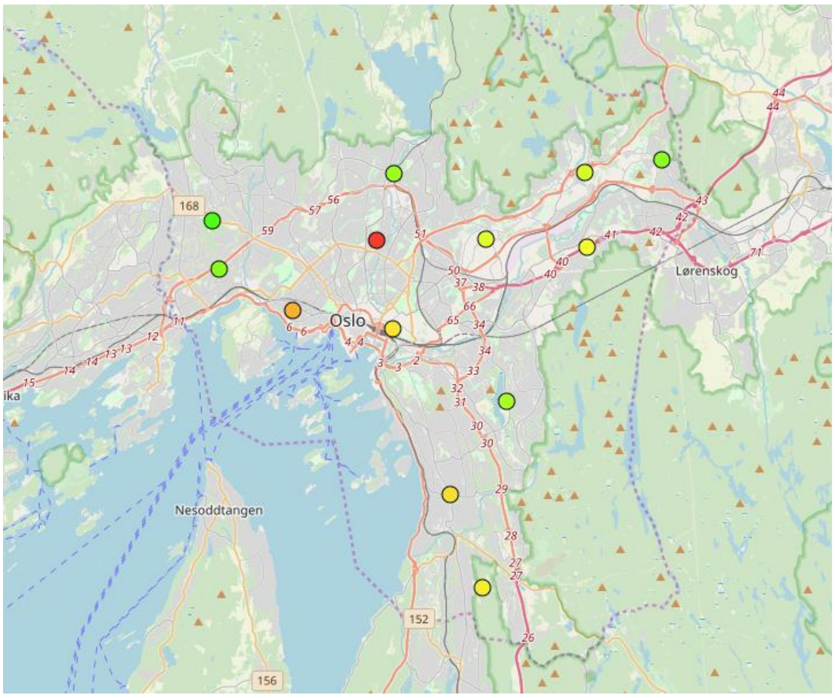

Fig 3. Depression levels in Oslo

Using Seaborn heatmaps we get the following correlation matrix.

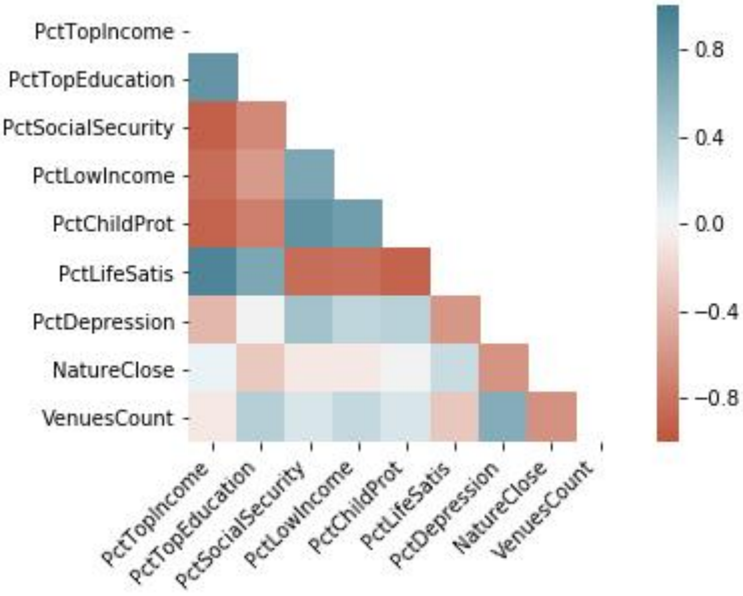

Fig 4. Correlation matrix

3.3 Table of most highly correlated variables

With the SciPy statistical package using stats.pearsonr() function and some custom code we iterate over all features to build a list of correlations along with p-numbers. Our code filters out associations with low p-value or low correlation. Inclusion criteria: Variable pairs with correlation (abs(corr[0])>0.4) AND statistical significant p-value (ρ< 0.05). Duplicate associations (reversed) are also removed.

Table headers

- ρ: Correlation coefficient using Pearson product-moment correlation coefficient [6].
- df: Number of paired data.
- p-value: Probability value [9].

Association	ρ	df	p-value
PctTopIncome-PctSocialSecurity	-0.92	11	0.000008
NatureClose-VenuesCount	-0.61	11	0.026954
PctDepression-NatureClose	-0.60	11	0.029002
PctLifeSatis-PctDepression	-0.58	11	0.036467
PctTopIncome-PctTopEducation	0.82	11	0.000676
PctTopIncome-PctLowIncome	-0.83	11	0.000453
PctTopIncome-PctChildProt	-0.90	11	0.000033
PctTopIncome-PctLifeSatis	0.93	11	0.000004
PctTopEducation-PctSocialSecurity	-0.68	11	0.011228
PctTopEducation-PctLowIncome	-0.56	11	0.046869
PctTopEducation-PctChildProt	-0.73	11	0.004826
PctTopEducation-PctLifeSatis	0.67	11	0.012984
PctSocialSecurity-PctLowIncome	0.66	11	0.013213
PctSocialSecurity-PctChildProt	0.83	11	0.000505
PctSocialSecurity-PctLifeSatis	-0.83	11	0.000404
PctLowIncome-PctChildProt	0.75	11	0.003355
PctLowIncome-PctLifeSatis	-0.83	11	0.000484
PctChildProt-PctLifeSatis	-0.89	11	0.000055
PctDepression-VenuesCount	0.61	11	0.026584

3.4 Correlations and regression modelling

Now with this table we investigate correlations further by using the Seaborn lmplot function to plot data and regression model fit across a FacetGrid.

We repeat this step for the hypotheses we wanted to test with our research goals.

For associations such as depression and high income we see a negative correlation but the points are too spread out to be statistical significant.

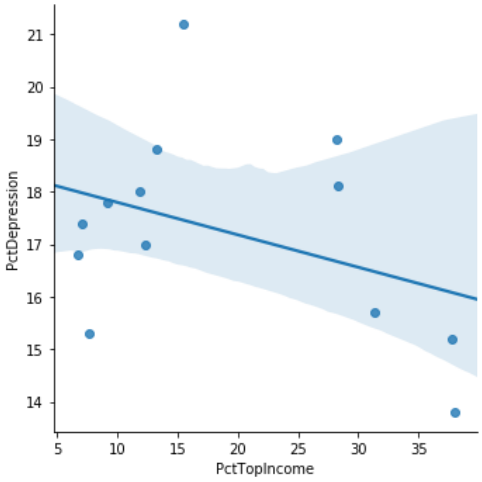

Fig.5. High income vs depression

To further check the correlation coefficient and p-value of this feature pair we used the SciPy stats pearsonr function which gives us a correlation of -0.378 and a p-value of 0.202.

An example of a correlation plot with a much stronger p-value is the association between low Income and life satisfaction.

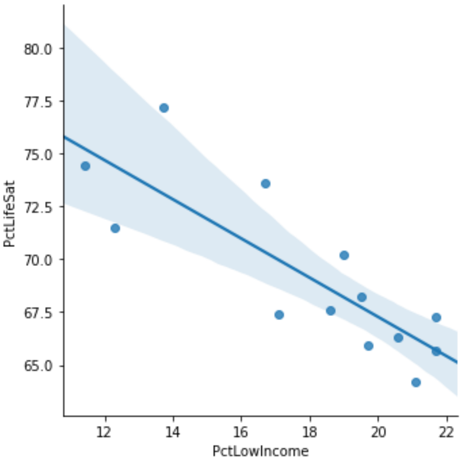

Fig.6. Low income vs life satisfaction

Pearsonr function here gives us a correlation coefficient of -0.827 and a p-value of 0.000484.

4 Results

4.1 Research goals

Returning to our research goals we have found:

1. High average income level (all ages) is very strongly correlated with life satisfaction among the young (ρ 0.93, p-value 0.000004).

2. A high average education level (all ages) is moderately strong correlated with life satisfaction among the young (ρ 0.67, p-value 0.012984).

3. The use of social security welfare benefits among ages 50 to 59 has a strong negative correlation with life satisfaction among the young (ρ -0.83, p-value 0.000404).

4. The number of venues in a borough is moderately correlated with depression among young (ρ 0.61 p-value 0.026584).

5. High share of child protection services cases has a strong negative correlation with life satisfaction among the young (ρ -0.89, p-value 0.000055).

6. High number of venues is moderately correlated with depression (ρ 0.62, p-value: 0.022743).

In the above summary we have used the following correlation coefficient interpretation scale by Schober et al.[10]

Correlation coefficient	**Interpretation**
0.00-0.09	Negligible correlation
0.10-0.39	Weak
0.40-0.69	Moderate
0.70-0.89	Strong
0.90-1.00	Very strong

4.2 Other surprising findings

High income has a stronger negative correlation with child protection cases than high education level.

Life satisfaction is only moderately negatively correlated with depression which may suggest that depression can also be found in high income families.

4.3 Other not surprising findings

Some not surprising findings include the following:

- Low income is correlated with child protection service cases.
- Top income is highly correlated with top education level.
- High income is negatively correlated with the use of social security welfare benefits.

- The number of venues in a borough is negatively correlated with the degree of nature and green spaces nearby.

4.4 Clustering

The following clustering analysis shows similarity of boroughs. We create 3 clusters using the *k*-means clustering algorithm [11]. 3 clusters were chosen in order to grade boroughs from low, medium to high when it comes to life satisfaction and depression.

Red boroughs warrant attention. It also clearly shows three boroughs stand out positively at the North West side of Oslo.

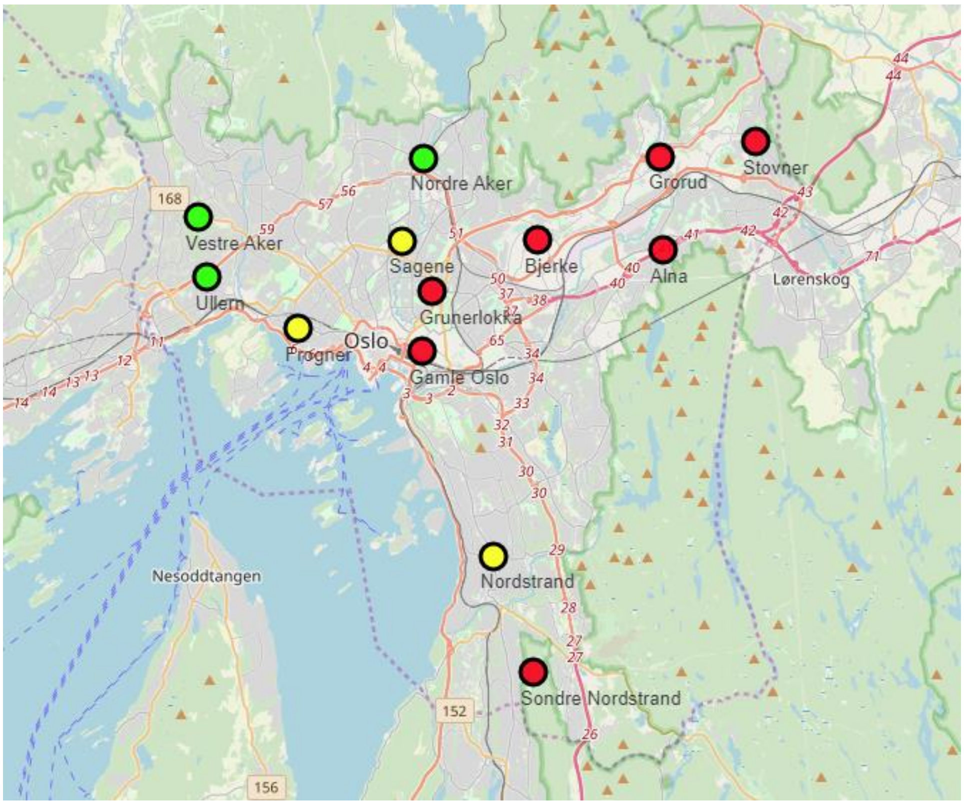

Fig.7. Depression and life satisfaction clusters Oslo

5 Discussion

Depression had more variations than life satisfaction and so we were not able to find a lot of correlations on depression. Correlations on life satisfaction on the other hand were found on several variables. We interpret this as depression being more randomly distributed with more complex underlying causal structures.

5.1 Limitations

There are considerable limitations and weaknesses in this analysis. However for educational purposes we still feel the overall process and methodology is adequate.

- As we are not associated with a public institution we were not able to get raw data access on a more granular level other than on an aggregated mean statistical level. Although we were able to achieve statistical significance in many places a more thorough analysis on a more granular level would provide a better basis for running the regression analysis.
- The analysis was not guided by or supported by any prior research, framework or theory. We were simply checking correlations on random seemingly meaningful variables.
- We did not have exact borough geo positions and were checking radius from the centre of the borough instead of within defined boundaries.
- As this analysis was not intended for real use we have not done a very thorough job in finding alternative variables or variations of variables to analyse. We could for example rerun and adjust various threshold levels for various indicators such as what we want to define as top income levels etc.
- Some data is older than others. Most datasets are from 2018 surveys and some data as depression levels are as old as from 2015.

5.2 Recommendations

Based on our findings we recommend policy actions aimed towards the following boroughs: Stovner, Grorud, Alna, Grunderløkka, Bjerke, Gamle Oslo and Søndre Norstrand.

Policy actions such as low-threshold in neighbourhood health services for those with mental challenges should be considered. Also school programs, campaigns, classes etc. should be considered to tackle the issue of poor life satisfaction.

6 Conclusion

By using a correlation matrix we were able to explore and find correlations, some of which were surprising. Our top research goal was to test the hypothesis that income is associated with depression and life satisfaction. We could not find evidence of income being associated with depression. We did on the other hand find strong evidence of higher life satisfaction being associated with high income (strong correlation, p-value: 0.000004).

This paper has shown how machine learning can be used in analysing associations between variables. First linear regression was used to explore correlations, and then *k*-means clustering was used to display similar boroughs and produce a simple easy to understand map overview of depression and life satisfaction.

We conclude this research summarising our main findings which can be interpreted as follows:

> *Higher income levels (at least up to a certain level) in the family make children happier, but living in a high income family will not necessarily save the children from depression.*

7 References

[1] Our World In Data, Global Change Data Lab https://ourworldindata.org/mental-health

[2] The World Happiness Report https://worldhappiness.report/ed/2019/

[3] Andreasson, U., Birkjær, M. (2018). In the Shadow of Happiness. Copenhagen https://doi.org/10.6027/ANP2018-799

[4] Foursquare API https://developer.foursquare.com/docs/api

[5] Oslo municipality statistics database http://statistikkbanken.oslo.kommune.no/webview

[6] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[7] Young in Oslo survey, OsloMet http://www.hioa.no/Om-OsloMet/Senter-for-velferds-og-arbeidslivsforskning/NOVA/Publikasjonar/Rapporter/2015/Ung-i-Oslo-2015

[8] Assignment Github Jupyter notebook https://github.com/roylac/Coursera_Capstone/blob/master/Capstone%20week%204.ipynb

[9] p-value wikipedia https://en.wikipedia.org/wiki/P-value

[10] Correlation Coefficients: Appropriate Use and Interpretation, Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA Anesthesia & Analgesia: May 2018 - Volume 126 - Issue 5 - p 1763-1768 doi: 10.1213/ANE.0000000000002864

[11] *k*-means clustering wikipedia https://en.wikipedia.org/wiki/K-means_clustering