COMP 309/AIML421 | Machine Learning Tools and Techniques

Assignment 2: Data Exploration, Manipulation and Modelling

Corvin Idler - Student ID (300598312)

Part 1: Understanding Data From World Happiness Report [40 marks]

1.

(a) the correlation between variables, check any important pattern:

I decided to measure correlation between the ladder score and the explanatory variables on a dataset that combines the years 2017-2019. There is an implicit assumption that patterns didn't change considerably over the years. I haven't validated that assumption as the main goal of that decision was to keep the analytical output at bay (given the restrictions on the report length).
The pearson correlation coefficients are as follows.

+0.797 GDP per capita vs. Ladder Score
+0.758 Social support vs. Ladder Score
+0.751 Healthy life expectancy vs. Ladder Score
+0.550 Freedom to make life choices vs. Ladder Score
+0.405 Perceptions of corruption vs. Ladder Score
+0.117 Generosity vs. Ladder Score

It appears that GDP, social support and healthy life expectancy are all univariately important for the ladder score.

There are also some interesting correlation patterns present amongst the explanatory variables. Particularly GDP per capita, healthy life expectancy and social support seem to be quite correlated with each other. This lines up with all three also being highly correlated with the ladder score.
Generosity sticks out as something uncorrelated with the above group of 3 variables, which also lines up with generosity barely being correlated with the ladder score.

0.782	GDP per capita	Healthy life expectancy
0.696	GDP per capita	Social support
0.644	Healthy life expectancy	Social support
0.457	Freedom to make life choices	Perceptions of corruption
0.423	Freedom to make life choices	Social support
0.345	Freedom to make life choices	GDP per capita
0.325	GDP per capita	Perceptions of corruption
0.324	Generosity	Perceptions of corruption
0.319	Freedom to make life choices	Healthy life expectancy

0.270	Healthy life expectancy	Perceptions of corruption
0.264	Freedom to make life choices	Generosity
0.208	Perceptions of corruption	Social support
-0.030	Generosity	Healthy life expectancy
-0.006	GDP per capita	Generosity
0.002	Generosity	Social support

(b) figure out any countries consistently scoring high or low across the variables:

I performed an equal-frequency discretisation across all variables (with three intervals for high, medium, low). I then filtered those countries that score either low across all variables or high across all variables. Via a pivot table I then selected countries that scored high or low across all variables for all years. That was my interpretation of “consistently scoring high or low across the variables”.
No country scored low across all variables for all years. but the following countries scored high across all variables for all years:

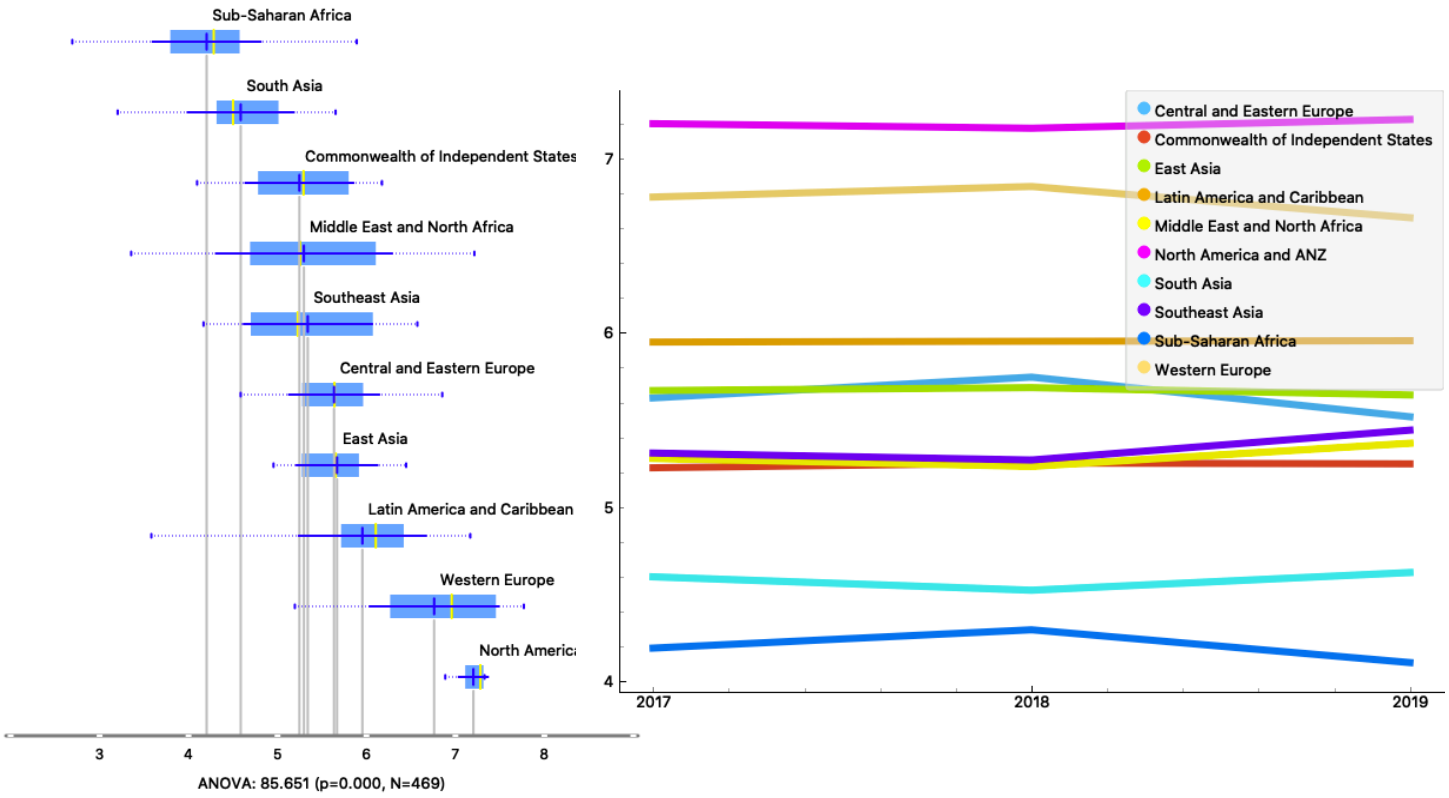
	Country/Area
1	Australia
2	Canada
3	Denmark
4	Iceland
5	Ireland
6	Malta
7	Netherlands
8	New Zealand
9	Norway
10	Sweden
11	Switzerland

(c) any happiness discrepancy across regions and years:

I standardised the variables and fixed missing regional information and harmonised the spelling of different countries across time.

As one can see in the boxplot representation, there are considerable differences in the mean happiness across regions with North America + ANZ and Western Europe leading the way and sub-Saharan Africa and South Asia trailing the field.
Further more one can see that some region contain very extreme cases of happiness and unhappiness (e.g. Middle East and Latin America/Caribbean).

Over time some regions improved (e.g. North America and ANZ, South Asia, South East Asia and Middle East/North Africa, Middle East and North Africa) while some other regions declined (e.g. Western Europe, Central and Eastern Europe and Sub-Saharan Africa)



2.

(a) Identify and describe two other business questions that could be asked of the happiness data as well.

Another question one could ask is if there are regional differences in happiness levels or in what drives happiness.
A third question one could ask, if there are changes over time in happiness levels or in what drives happiness. Visually some of these questions are partly answered already in 1. c)

(b) For these three business understanding questions, decide which machine learning techniques can be applied to meet them and provide your justifications in the report.

Not sure if the question relates only to the data at hand which seems to be the result of some sort of regression already or if one would assume to have access to the full set of variables (e.g. like in part 2). In the first case, there isn't too much sophisticated machine learning left to do. In the latter case:
One could use pooled regression to quantify the dependence of the ladder score on the explanatory variables. One could use regression with fixed effects per region and find out e.g. via ANOVA if there are significant regional differences in happiness that can't be explained with the explanatory variables. One can do the same with fixed effects for the years. One could also estimate 3 pooled regressions for the three years and see if the regression coefficients significantly differ. Last but not least, one could perform regression with interaction terms for regional dummy variables or temporal dummy variables and see which explanatory variables depend in their behaviour on regions or time. One would have to check with ANOVA the significance of these terms or at least "reign in" the number of variables by regularisation or

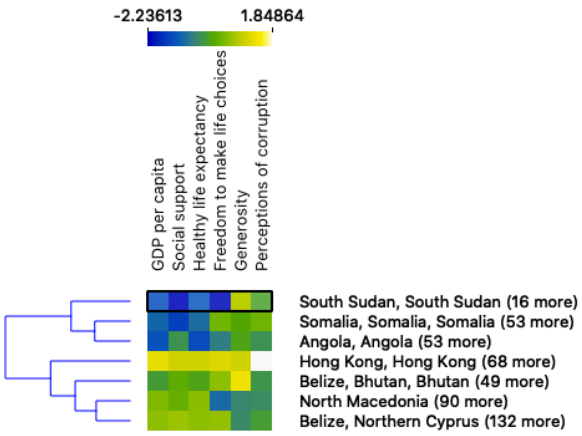
variable subset selection schemes. One could also apply regression trees to the problem and determine variable importance

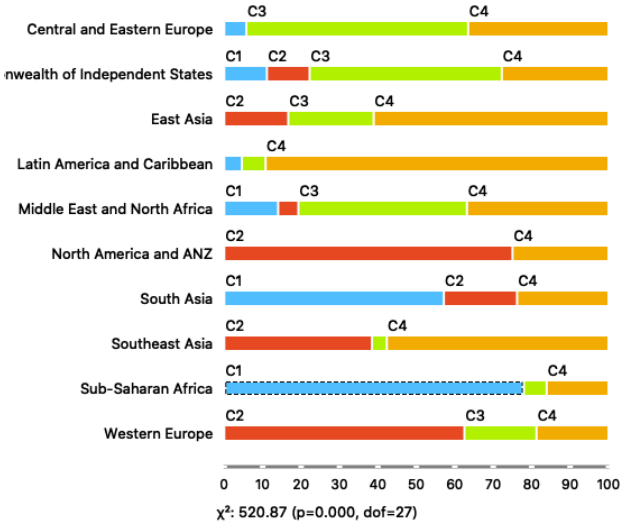
3.

I concatenated the data sets and added a year column to it, to distinguish the years.
The new data set has 476 rows and 12 columns, with one target variable (ladder score), 8 numerical features, 3 categorical features. That said, the ladder rank should not be included as a feature (see later parts of the assignment).

4.

I performed at heat map analysis with 7 cluster and we can see there are certain patterns or country profiles with regards to the variables. E.g. High across the board but exceptionally well for corruption perception (e.g. Finland, Norway, Denmark, Hong Kong etc). Or bad across the board with the exception of Generosity and corruption perception (e.g. South Sudan, Burundi, Afghanistan). Some of these countries line up geographically… but from the hierarchical cluster analysis with even only 4 cluster we can see that there is a fair amount of geographical variety over which certain cluster spread





Part 2: Data Preparation and Machine Learning on Happiness Report Data [60 marks]

1.

In my view this constitutes a regression problem as the ladder score is a continuous variable, unless one discretises the ladder score into intervals like “high”, “medium”, “low”.
Given the potential multicolinearity risk in the data set, one could use dimensionality reduction techniques with regression but unless the factor decomposition pans out in a way that makes the factors meaningful (e.g. dimension of the real world drivers is captured by a factor) the interpretability would suffer and the business question might not be answered to the satisfaction of the business. Maybe regularisation techniques like Lasso or Elastic net or variable subset selection methods might be better placed. Not sure these Feature Selection methods count as a “machine learning technique”. Strictly speaking the business question is more specific than just “what drives happiness”, namely “what makes the world's happiest countries so happy?”, one

would have to define which countries to be considered the world's happiest (e.g. top 10 or top 10%) and then restrict the analysis to these countries (conceptually similar to quantile regression). But not sure that is really desired here.

Besides regression one could also apply a clustering algorithm and hope that some clusters form in a fashion that one contains the most happiest countries.

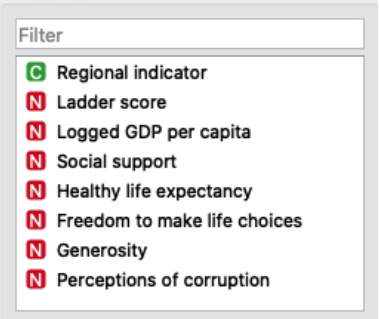
2.

The RadViz plot shows some outlier (Afghanistan and Rwanda) that might be worth exploring further. Afghanistan shows very low levels for social support and freedom to make choices, extremely high levels of corruption perception as well as low levels of the ladder score. Might very well be an outlier compared to the rest of the world, but then again, not too surprising for a war torn nation.
Rwanda scores pretty well across the board (for a sub-saharan african nation), but has a rather low ladder score and an extremely low perception of corruption. The relative unhappiness might still be a ripple effect from the genocide/civil war.

I standardised the data and I did a multivariate outlier analysis (e.g. one class svm or single class tree) and by definition (depending on the settings for the SVM) there will always be some outlier. E.g. Rwanda and Afghanistan. It would require a lot more digging to understand why the SVM considered some of the other countries outliers (e.g. are the sadder than they should be or happier than they should be). I decided not exclude these countries from further analysis.
I couldn't fine missing data or unnecessary instances. I also one-hot encoded the regional indicator variable.

3.

I only kept the following variables (with country as a meta attribute) as the other variables were more or less already the result of a regression analysis.



4.

I used the feature ranking building block from orange data miner and as one can see below there is a considerable difference between the assessment when variable interactions are taken into account (RReliefF) and when they are not taken into account (Univar reg). So the question of which features are most important depends wether or not one wants a model with interaction terms.

	#	Univar. reg.	RReliefF
N Regional indicator=Latin America and Caribbean		2.855	0.200
N Regional indicator=South Asia		7.935	0.163
N Regional indicator=Sub-Saharan Africa		62.955	0.160
N Freedom to make life choices		86.098	0.157
N Generosity		0.047	0.142
N Perceptions of corruption		31.693	0.136
N Regional indicator=Middle East and North Africa		1.638	0.129
N Social support		197.165	0.129
N Logged GDP per capita		243.667	0.128
N Healthy life expectancy		211.516	0.126
N Regional indicator=Central and Eastern Europe		3.455	0.119
N Regional indicator=Southeast Asia		0.130	0.101
N Regional indicator=Western Europe		55.355	0.079
N Regional indicator=Commonwealth of Independent States		0.049	0.060
N Regional indicator=North America and ANZ		9.602	0.044
N Regional indicator=East Asia		0.416	0.039

5.+6.

Ordinary regression tries to estimate a global model of what drives happiness. All instances are used to estimate the regression coefficients and for each instance all explanatory variables are used to estimate the ladder score. This method might be suited to make statements “on average” about the whole population, it might be less suited to behave “localised”. E.g. a regression tree could ignore certain features for certain instances if these features provide no additional insight. A regression model could do that only if we estimate a different model for subgroups of instances like in the case of quantile regression. I divided the group into 5 quantiles and performed one set of analyses on the whole population and one set of analysis on the top quantile only. Furthermore, I performed in each case a regression analysis including regional dummy variables as well as excluding them.

When regional dummy variables are included they represent regional fixed effects. So everything a region has in common that isn’t captured by any of the other features. One can see that GDP, freedom to make life choices and social support lead the field in terms of importance based on the magnitude of their regression coefficient. As all features were standardised the regression coefficient tells us something about the feature importance. We need to be careful though with the dummy variables as they are all one, so not really a standardised feature.
When we look at the results for “the most happiest countries in the world” = top 20% of ladder score, we find, that the field is led by social support, healthy life expectancy and freedom to make life choices. GDP is comparatively “unimportant”. They way I interpret these results is that GDP is probably generally high amongst the top 20% of happy countries. So it’s a given “hygiene factor”. The focus therefore shifts onto other variables in which these top 20% countries might differ more and which therefore can explain differences in happiness within the top 20% group.

All instances		Top 20%	
Regional indicator=South Asia	-0.557	Regional indicator=Central and Eastern Europe	-0.165
Regional indicator=Southeast Asia	-0.460	Perceptions of corruption	-0.126
Regional indicator=Commonwealth of Independent States	-0.189	Generosity	-0.012
Regional indicator=Sub-Saharan Africa	-0.134	Regional indicator=East Asia	-0.010
Regional indicator=Middle East and North Africa	-0.108	Regional indicator=North America and ANZ	-0.004
Perceptions of corruption	-0.059	Regional indicator=Commonwealth of Independent States	0.000
Regional indicator=East Asia	-0.048	Regional indicator=Middle East and North Africa	0.000
Generosity	0.075	Regional indicator=South Asia	0.000
Healthy life expectancy	0.095	Regional indicator=Southeast Asia	0.000
Regional indicator=Central and Eastern Europe	0.171	Regional indicator=Sub-Saharan Africa	0.000
Social support	0.223	Regional indicator=Latin America and Caribbean	0.029
Freedom to make life choices	0.256	Regional indicator=Western Europe	0.030
Regional indicator=Latin America and Caribbean	0.300	Logged GDP per capita	0.085
Logged GDP per capita	0.309	Freedom to make life choices	0.181
Regional indicator=Western Europe	0.506	Healthy life expectancy	0.185
Regional indicator=North America and ANZ	0.520	Social support	0.344
intercept	5.504	intercept	6.118

One can perform the same analysis while excluding the regional dummy variables and the result are similar. For the all country data set, GDP is still the most important variable, followed by social support and freedom to make life choices (these two swapped places). For the top 20% the order is the same as with regional indicators.

All instances		Top 20%	
Perceptions of corruption	-0.108	Perceptions of corruption	-0.154
Generosity	0.055	Generosity	0.004
Healthy life expectancy	0.204	Logged GDP per capita	0.059
Freedom to make life choices	0.227	Freedom to make life choices	0.150
Social support	0.284	Healthy life expectancy	0.211

Logged GDP per capita	0.323	Social support	0.276
intercept	5.533	intercept	6.185

With regards to variable importance there is a further caveat: Above statements are only valid in the context of a full model (all variables included). The picture might change once one or many variables are excluded. So across all possible subsets we find that the order of importance is different. Once could try to use methods as described in this paper to find the variables that are most often included in variable subsets during subset selection processes https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3316704/

Last but not least, one can perform a cluster analysis to see what the happiest countries have in common.
One can see that the cluster means of the happiest 17 countries cluster are high across all variables except for perception of corruption, for which a low score is obviously a good thing.