P1(Q1) EDA exploratory data analysis

Initial EDA:

Identification of variables and data types:

Using 'Data info', 'Data table' , 'Feature statistic'

Regional indicator: Nominal

Ladder Score, economic production, social support, life expectancy, freedom, absence of corruption, and generosity: Numerical

>155 instance and 13 variables (2017 data)

>156 instance and 10 variables (2019 and 2018 data)

>Variable scale is different — might require normalisation.

>Data size is small — might need cross validation.

>Missing values in regional indicator— does not effect majority of data (except 1c).

>In 2018 corrupting is seen as string— domain is edited to numerical.

S1: Generate questions about your data

S2: Search for answers by visualising, transforming, and modelling your data

S3: Use what you learn to refine your questions and/or generate new questions

S1a: What are the correlation between the variable and are there any important patterns we can identify?

S2a: Using tools to 'Preprocess' data by imputing missing values and normalising the numerical value to intervals [0,1], 'Select Columns' to extract ladder score and the 6 variables, identify 'Correlations' [Figure 1 in appendix] by creating non-graphical visualisation and 'Scatter plot' to create a graphical visualisation in orange [1a.ows].

All correlation ranking above 0.5 (moderately correlated) has been consistent between 2017-2019. The highest correlation is with GDP and life expectancy, this is followed closely by GDP and ladder score then life expectancy and ladder score.

GDP and Life expectancy has a positive correlation with ladder score and GDP has a positive correlation with life expectancy. This may indicate that GDP has the strongest relationship with ladder scores as an increase in GDP directly and indirectly (through life expectancy) contribute to increases in ladder scores.

S3a: We have determined the correlation between variables and identified an important patterns. To further investigate we should determine which combination of variable have the greatest influence on ladder score.

S1b: What countries consistently score high or low across variables?

S2b: Using tools to 'Preprocess' data by imputing missing values and normalising numerical value to intervals [0,1], 'Select Columns' to extract ladder score and the 6 variables, 'Select row' to extract countries who scored consistently high (above 0.6 for all variable excluding corruption and below 0.35 for corruption) and low (below 0.7 for all variables excluding corruption and above 0.35 for corruption), identify countries using 'Data Table' to create non-graphical visualisation [Figure 2] [1b.ows].

>Corruption is inverted as I assume the higher the corruption the lower it contributes to happiness. Above and below value adjusted until results were found.

High

Malta scored consistently high between 2017-2019

Thailand scored consistently high in 2017 and 2019

Iceland scored consistently high in 2019

Low

Georgia scored consistently low in 2017-2019

Gambia consistently low in 2019

Burundi consistently low in 2019

Ethiopia consistently low in 2017

S3b: We have determined which countries score consistently high and low across variables. How does scoring consistently high or low across variables effect a country's happiness?

S1c: Are their discrepancy in happiness across regions and years?

S2c: Using tools to 'Preprocess' the data by removing rows with missing values, 'Select Columns' regions and ladder score (displayed using violin graph to create graphical visualisation [Figure 4]), 'Pivot Table' to group a region to its ladder score average and 'Merge data' to produce a data table with average ladder scores by region between 2017-2019. [Figure 3 in appendix, 1c.ows]

By regions there are no significant discrepancy between the average ladder score by regions with the different of ladder score per-year by region not exceeding 0.15 points.

However when observing the violin graph, the distribution of happiness score is spread out with few points being at the start and end of each regions box plot. This is not a confirmation of discrepancy but an indicator that they may exist.

Hence over regions there are no significant discrepancy, however countries within regions may produce a different result.

S3c: We have determined that there is no significant discrepancy when observing the happiness score across regions and years. To better investigate we should identify if there are countries within these regions which may have a discrepancy in happiness.

>EDA helped answer these question by summarising, visualising, and identifying important characteristics of the data. It ensured we were familiar with the data using the initial pre-process, helped us answer questions through graphical and non graphical outputs and better understand patterns by challenging us to refine/generate new questions.

P1(Q2a) Business understanding questions/business objectives

1. Which factors contributes the least to the happiness score?
2. Is there noticeable geographical pattern on the variable which contributes the most to ladder score across years?

P1(Q2b) Machine learning techniques

Regression is used to find the relationship between two variables one dependent and one independent. This technique is best used to identify what factor contributes the most and least to happiness score as it allows a machine to determine the strongest/weakest relationship between happiness score (independent) and a variable (dependent).

>To determine the factor that contributes most to the happiness score, we can leverage a regression algorithm, which falls under the umbrella of supervised learning techniques. In this instance we are trying to understand the relationships between the 6 different dependent factors, and the independent "ladder score" of the country. Using linear regression, we can model each factor(y) to the "ladder score"(x) to try to establish a relation, which can be visualized through scatter plots. With these six created plots, while a "line of best fit" can visually express the relationship, we ultimately want to calculate the "correlation coefficient" which indicates the strength of the association between the 2 variables, up to the highest of 1 for strongest correlation, -1 for strongest inverse correlation & 0 for no association. Acquiring 6 correlation coefficients for the 6 factor's relationship to the ladder score. We can rank them by how far the correlation coefficients are away from 0(taking an "absolute" value of the coefficient).
The highest rank will correspond to the factor that contributes most to the happiness score.

>To determine the factor that contributes least to the happiness score, we set up Linear Regression models and rank the correlation coefficients as described above for our first question. But the lowest ranked coefficient will correspond to the factor that contributes least to the happiness score.

> (We can set up Orange Data Mining to do this, if we use the "scatter plot" widgets to do this and plot 1 axis as the "ladder score" and the other as one of the "factors". The regression line will show an "r" value which will be the correlation coefficient. We can also see the value of these 6 "r" values being ranked in the "correlations" widget if we set the "ladder score" as the target variable)

Clustering group or cluster observations that have similar characteristics. This technique is best used to identify patterns in a region as it groups instances with similar characteristics together allowing for ease of analysis.

>To determine the geographical pattern on the variable which contributes most to ladder score across years, we set up Hierarchical Clustering. This falls under the umbrella of unsupervised learning techniques. In this instance we are trying to cluster instances with similar results for a specified variable and display them on a dendrogram which will allow us to observe geographic patterns.

> (We can set up Orange to do this by calculating the distance of the 3 set of data and running that through the 'Hierarchical Clustering' tool)

(P1)(Q3) Merge Dataset

[merge.cvs] [3.ows]

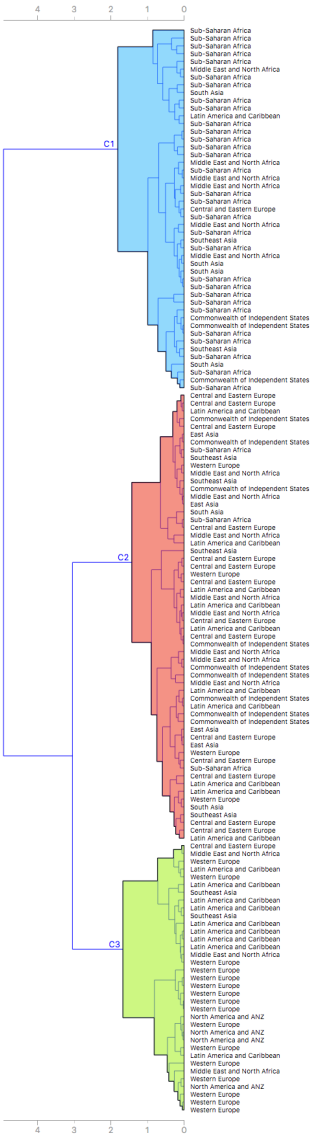
The domain for 2018 data is changed to depict Corruption as numerical variable.
Data from each year is preprocessed — this removes instances with missing values and normalise numerical value to intervals [0,1].
Select column is then used to collect important variables such as ladder score, rank and the 6 variables.
Merge data is then used to collect all the important variables from each year into one sheet.
Aggregate column is used on each variable to merge the value of the variable from each year to produce an average per instance.
All the aggregate columns of the variables are collected using select column which is then displayed on a data table.

The data is ready to be used as an initial dataset for machine learning task— it is a reflection of each instances performance in each variable over the 3 years. The dataset contains 10 variables, 8 are numerical and depict the 6 variables, ladder score and rank, 1 is text which depicts the country the instance belongs to and 1 is categorical which depicts the region which an instance belongs to.

(P1)(Q4) Hierarchical Clustering

S1: is there a noticeable geographical pattern on the happiness score across years?
S2: Using tools to 'Preprocess' the data by removing rows with missing values and normalising the numerical value to intervals [0,1], 'Select Column' to collect ladder score and country/area, 'Merge' 2017, 2018, 2019 data, calculate 'Distance' by row and produce a graphical visualisation using 'Hierarchical Clustering' [4.ows].
From the Hierarchical Cluster we can observe that there is a geographical pattern in the dendrogram on happiness score across years.
Countries in Western Europe tend to score higher in happiness score, whereas countries in sub-saharan Africa tend to score lower in happiness score. For the placement of countries in our hierarchical clusters we can observe that they tend to be from the same regions.
This may suggest that geographical placement of a country may influence its happiness score.

S3: We have established a geographic pattern on happiness score, to further investigate we should determine which highs and lows in variable these regions share in common.



>The venogram helped us answer the question by creating a visual output of the hierarchical cluster of happiness scores. This made it easier to identify patterns of regions in relations to happiness score.

P2(Q1) Data mining question and Machine learning technique

Data mining question: Predict a country's ladder score based on a set of factors of happiness.

ML Techniques:

>Regression forms predictions from data by learning the relationship between features of the data and some observed, continuous-valued response.

Regression will be used to find out which factor is the most contributing to happiness score and which factor the least.

>Clustering is a type of unsupervised machine learning algorithm. It is used to group data points having similar characteristics as clusters.

Cluster will be used to group countries of similar happiness scores together into groups, then run a clustering algorithm based on strongest or weaker factors to see which cluster it belongs to. Evaluate test data to see if its predictions are accurate.

>Classification is a type of supervised machine learning algorithm. For any given input, the classification algorithms help in the prediction of the class of the output variable.

Classification will be used to run a classification algorithm with the raw inputs of the 6 variables of happiness to try and classify them into groups based on a prediction of their happiness score.

Evaluate and score results.

We have selected these machine learning techniques as they center on collecting instances into groups. This is necessary for us to predict a country's happiness score based on a set of factors of happiness as we need to train our machine to be able to identify what group an instance belongs to solely based on an instance combination of variables.

P2(Q2) Preprocessing

Preprocess steps taken:

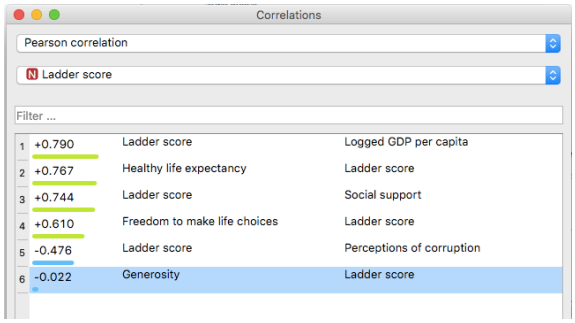
>impute missing data, however looking at 2021 data source we have seen that there is no missing data
>The values scale differently (age 52-74) (GDP 7-12) (corruption 0-1) etc… so we normalised it to make the variables measured in a common scale to allow easier comparisons and determination of high & low values,
>Removing outliers countries which have exceptionally high or low scores in 1 of its factors. This removes noise and contamination in our prediction algorithms.
>Discretise the ladder score values to 7 groups, ranging from “very low”->"very high” as this is how we group countries of similar happiness scores together.

[Processeds.csv][Part2.ows]

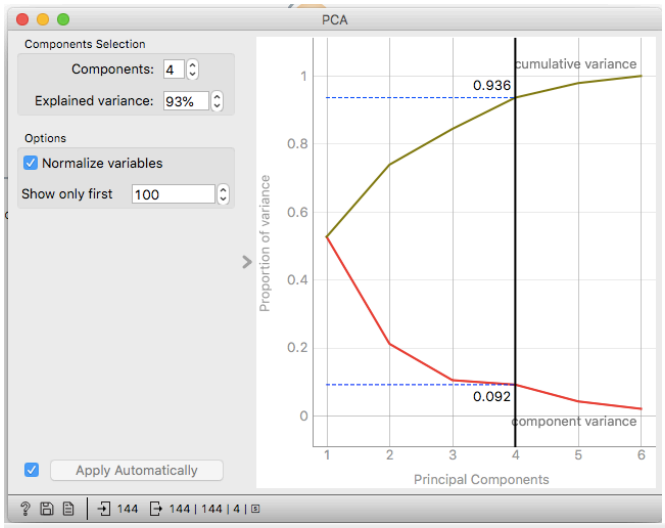
P2(Q3) Dimensionality reduction

Using domain knowledge of the dataset, we can remove the duplicate features in our dataset which will affect the performance of our algorithms (explained by: columns, high and low whiskers of the ladder score + residual dystopia).

One dimensionality reduction technique we used was High Correlation Filter. In which we find the correlations between the factors and the ladder score by comparing the Pearson’s Correlation Coefficient. Thankfully this is all done under the hood and displayed in the “correlations” widget of Orange



Another dimensionality reduction technique we used was Principal Component Analysis. PCA in this case helps us through the “feature extraction” concept, which will highlight the strongest and weakest features contributing to the ladder scores. Configuring 4 components in our PCA gives us a 93% variance which covers an acceptable amount of variability in our dataset.
Looking at our first and main Principal Component which covers the greatest variance, we can extract the “factors” that define said component the most, indicating the strongest factor towards “ladder score”.



Components to variables, for feature extraction

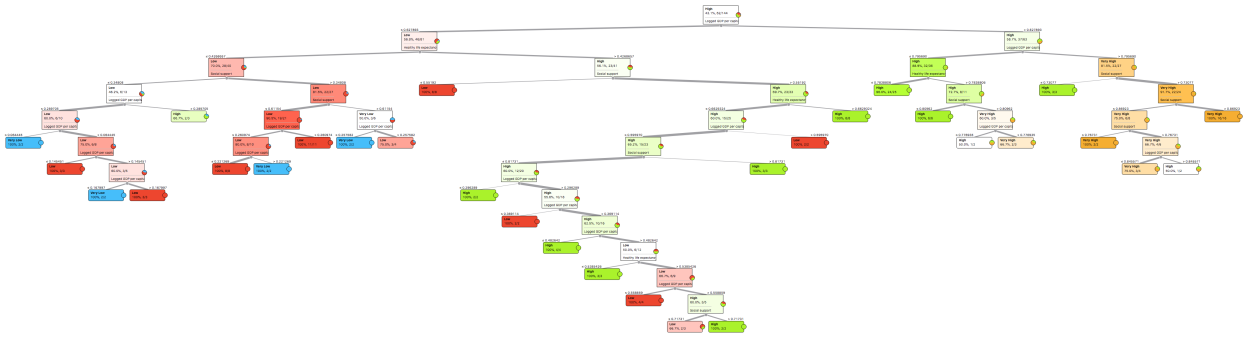
	components	Social support	althy life expectar	gged GDP per cap	om to make life ch	Generosity	ceptions of corrup
1	PC1	-0.485329	-0.509826	-0.520191	-0.368389	0.083487	0.302083
2	PC2	-0.162242	-0.108974	-0.1461	0.340704	0.763329	-0.491636
3	PC3	-0.269866	-0.0032077	0.0303525	-0.411684	-0.374091	-0.785375
4	PC4	0.205199	0.243742	0.236881	-0.753874	0.516224	0.0869328

[Part2.ows]

P2(Q4)
Dimensionality reduction techniques alongside the filtering of redundant features leverage in question 3 already identified the most and least important features that are part of the happiness score.

With Correlation Filters, using the “Correlations” widget, the calculated Pearson’s Correlation Coefficient shows that GDP, High Life Expectancy and Social Support had the highest weights as performed via linear regression.

With the help of Decision Trees for both our PCA transformed data and High Correlation Filtered Data. We can confirm that the strong factors are PC1 for PCA transformed data and Social Support for Filtered data.



[Part2.ows]

P2(Q5a)
For Classification we used 7 classification algorithms (Neural Network, Gradient Boosting, Random Forest, SVM, Naive Bayes, Logistic Regression and KNN). Of all of these algorithms we found that Neural Network (79% accuracy) and Logistic Regression (75% accuracy) performed the best across both transformed datasets - This accuracy rate was determined using the ‘Evaluation’ widgets (Test and score and confusion matrix).
Furthermore we can observe which instance way misclassified using the ‘Scatter plot’ widget.

For Clustering we used 2 clustering algorithms K means Clustering and Louvian Clustering. Of these algorithms we found that K mean performed better when observing the mosaic - however we were not able to produce a test and score outcome. Overall both clustering techniques did not produce the most accurate grouping as we can observe from the mosaic a majority of the graphed instances where clustered into the incorrect group.

Clustering ended up having weaker results than classification techniques. The difference lies in the fact that classification uses predefined classes in which objects are assigned, while clustering identifies similarities between objects, which it groups according to those characteristics in common and which differentiate them from other groups of objects. Classification performed better as class-action technique identifies the input data as a part of a specific category or group which allows it to more accurately predict a discrete value (ladder score) whereas clustering algorithms use distance measures to group or separate data points. This produces homogeneous groups that differ from one another which does not necessarily correlate with ladder score.

P2(Q5b)
Classification we used 7 classification algorithms (Neural Network, Gradient Boosting, Random Forest, SVM, Naive Bayes, Logistic Regression and KNN).
Clustering we used 2 clustering algorithms K means Clustering and Louvian Clustering.

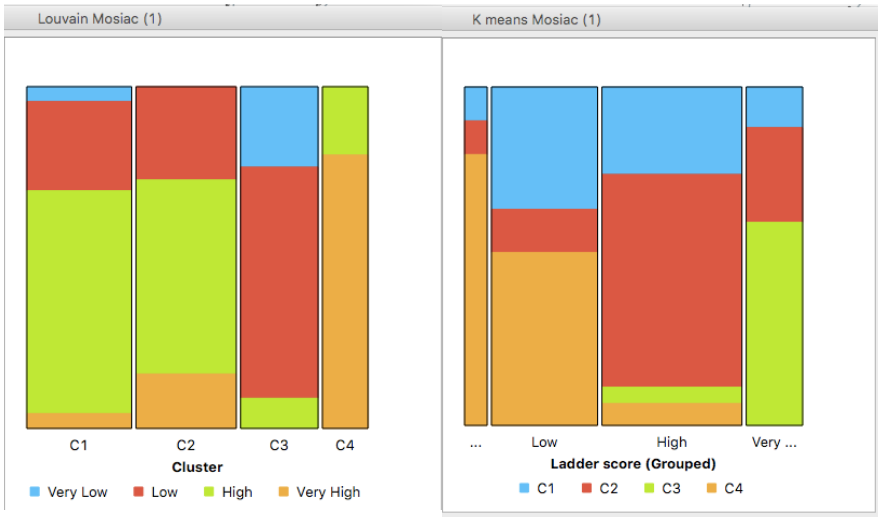
In best cases, Classification was better with 79% accuracy with Neural Network algorithm and over 70% accuracy for all other classification algorithms.
The confusion Matrix and Scatter Plots show clearly which instance (country) was misclassified, and as seen in the plot, around the borders of the clusters. Most incorrectly plotted instances were them being a boundary case (e.g between very high and high) giving us confidence that classification is a very accurate algorithm for estimating an instance’s ladder score based on its variables.

Test and Score (1)

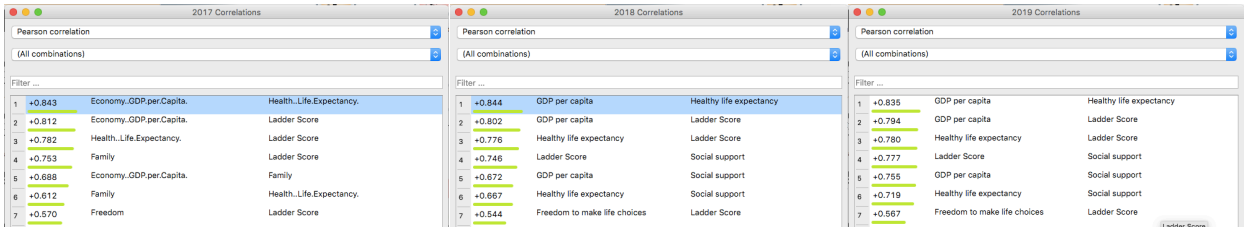
Evaluation Results

Model	AUC	CA	F1	Precision	Recall
kNN	0.880	0.736	0.720	0.705	0.736
SVM	0.883	0.757	0.728	0.709	0.757
Random Forest	0.902	0.750	0.747	0.747	0.750
Neural Network	0.866	0.764	0.745	0.783	0.764
Naive Bayes	0.888	0.688	0.695	0.708	0.688
Logistic Regression	0.899	0.785	0.770	0.762	0.785
Gradient Boosting	0.828	0.694	0.697	0.700	0.694

Clustering did not do well in neither PCA and Correlation Filters, seeing the graph and the Mosiac display, we can see that while there are 4 clusters created by K-means and Louvain, to correspond to our established 4, they were grouped very different to how the ladder score would've been.



APPENDIX
[FIGURE 1]



[FIGURE 2]

Consistently low

2019 | 2017 | 2018

	Country/Area	Regional indicator	GDP per capita	Social support	althy life expectar	om to make life ch	Generosity	:eptions of corrup	Ladder Score
1	Georgia	Commonwea...	0.526128	0.410099	0.659071	0.54834	0.07597	0.36203	0.338893
2	Gambia	Sub-Sahara...	0.182898	0.578202	0.375110	0.60539	0.47527	0.36865	0.338283
3	Burundi	Sub-Sahara...	0.027316	0.275246	0.333041	0.34865	0.31095	0.39735	0.187551

Consistently high

2017 | 2019 | 2018

	Country/Area	Regional indicator	GDP per capita	Social support	althy life expectar	om to make life ch	Generosity	:eptions of corrup	Ladder Score
1	Iceland	Western Eur...	0.819477	1.0000	0.899211	0.93661	0.62544	0.26049	0.944060
2	Malta	Western Eur...	0.771971	0.935961	0.875548	0.89382	0.66254	0.33333	0.787836
3	Thailand	Southeast A...	0.623515	0.867611	0.725679	0.88273	0.63428	0.06181	0.641782

[FIGURE 3]

1c

	Regional indicator	AVG Ladder Score 2017	AVG Ladder Score 2018	AVG Ladder Score 2019
6	North America and ANZ	7.22675	7.2025	7.17625
10	Western Europe	6.70371	6.8291	6.8984
4	Latin America and Caribbean	5.9474	5.93775	5.94255
1	Central and Eastern Europe	5.54338	5.6575	5.77694
3	East Asia	5.6816	5.7204	5.7406
5	Middle East and North Africa	5.42276	5.32565	5.27453
8	Southeast Asia	5.44488	5.31344	5.27367
2	Commonwealth of Independent States	5.25158	5.22917	5.25633
7	South Asia	4.56467	4.52417	4.43433
9	Sub-Saharan Africa	4.14294	4.2457	4.36056

[FIGURE 4]