1. CSE 482 FINAL PROJECT (Cover Page)

Project Title: **Predicting Happiness Index by Country**

Summary of Team Member Participation:

Fill out the following table for each team member of the group.

Name	Participate in data collection	Participate in preprocessing	Participate in data analysis/ experiment	Participate in writing the final report	Participate in creating video presentation	Completed Assigned Tasks
Becky Henning	Yes	Yes	Yes	Yes	N/A	Yes
Elena Komesu	Yes	Yes	Yes	Yes	N/A	Yes

Team Member Roles and Contributions:

Name	Roles and Contributions
Becky Henning	Responsible for collecting and preprocessing data; helped writing the final report; Created models and did general data analysis
Elena Komesu	Performed data analysis; helped writing the final report; Responsible for PCA Preprocessing and predictions

I approve the content of the final report (please add your signature below):

Becky Henning: ---------Rebecca Henning----------------

Elena Komesu: ----------Elena Komesu--------------------

Predicting Happiness Index by Country

Becky Henning, Elena Komesu
https://github.com/elekom/482_final_project

ABSTRACT

The goal of our project is to be able to predict the "Cantril Life Ladder" otherwise known as the "Happiness Index" of a Country based on certain attributes of the Country. There is an annual report called the World Happiness Report, which surveys a random sample of people from each available Country with the question, "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"

Datasets used in this project were analyzed under the assumption that the happiness index of a country is influenced by a number of factors affecting each country. We hoped to determine what these factors are in an effort to be able to predict the index of a country in any given year. We were also interested to see if countries clustered by trend in life ladder and other metrics would have higher regression predicting power than the regression model trained with all country data.

The motivation behind this project was pure curiosity. It is an interesting question and getting to the root of a population's level of happiness is something many professionals spend their lives researching. We formulated the problem as a regression problem and applied a multiple linear regression model to each of the resulting clusters as well as to the full data set to solve it.

After thorough analysis, we were able to achieve our goal. We determined the best attributes to use in our regression model to increase our r-squared and lower root mean squared error. We found that clustering the data by life ladder coefficient increased regression accuracy for 77 of the 164 countries (+0.084). Clustering by population mean increased regression accuracy for a whopping 156 of the 164 countries or over 95% of the data points (+0.042 average). Clustering by the correlation coefficient between Life Ladder and Delivery Quality increased regression accuracy for 100% of the tested countries (+0.05).

2. INTRODUCTION

Happiness is not binary; a person is not simply "happy" or "unhappy". Happiness can change from moment to moment, day to day. It lies on a spectrum and because of this, happiness is quantifiable. In the case of this project, variables increasing the cantril life ladder, "positive affects" include frequency of laughter and enjoyment on a given day. Conversely, "negative affects" include frequency of worry, sadness, and anger on a given day. These two attributes show clear cause and effect for happiness, but we theorized that these are not the only important factors, in fact they might not be the most important factors at all. Economic, Sociocultural, and Governmental factors of a given country may set the groundwork resulting in the aforementioned affects. This is what we set out to find.

The goal of the project was to create a regression model trained with the 1562 rows and 17 columns of data that we initially collected. This data contains 164 countries of the recognized 195 countries total in the world. Each of these 164 countries has a number of years of surveyed cantril life ladder data, ranging between 2005 and 2017. However, the range of years differs significantly between each country. We hoped to fill the intermediate year gaps and find trends in the data through clustering the countries in different ways and tracking the regression model's predictive power. We hypothesized that clusters would increase the model's accuracy for the majority of the countries, compared to the model used on the full dataset.

We started our analysis by iteratively creating a multiple linear regression model using different attributes from the data set. Those attributes that are determined to create the most accurate predictions and will be used in our clustering when creating the groupings for the regression models. This will allow us to create the most strategic groupings.

We have collected data from the World Happiness Report itself along with population data from WorldBank. This data ranges from information about the Country's economic standing to social support to population density. All attributes are factors that we believe may have a hand in determining the general level of happiness that a Country maintains. Some data is surveyed, and other data is collected from world data banks. We believe our source is reliable, and the surveys unbiased.

Collecting the data was challenging only because of the lack of availability of public data surrounding our particular problem. We really only had one source for about half of the attributes in our dataset, so finding the same or similar data for countries not listed or years of countries left out was nearly impossible. There was also a significant amount of missing data for some of the listed attributes. We overcame both of these main challenges by using predictive modeling techniques such as the Matrix Factorization model described in class to fill in missing data. We reasoned that the country row of data would affect the datapoint along with other similar data points, creating a statistic that uniquely matches the data.

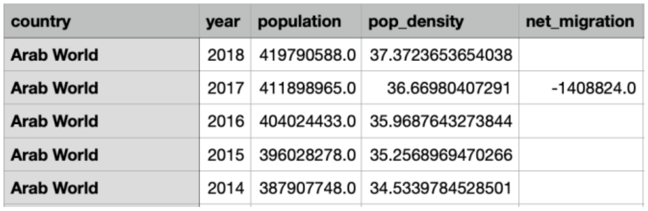
We tried many clustering techniques in the process of analyzing the data, and most were unsuccessful, averaging r-squared scores between 0.5-0.6 and root mean squared errors in the same range. The three most successful clustering techniques were using the life ladder coefficients, population mean, and correlation coefficient between life ladder and delivery quality. Our data supports our hypothesis with the latter technique providing a way to increase regression accuracy for every data point over the model fit with the entire dataset.

3. DATA

For this project, we used 2 datasets, both originally CSV files, consisting of Migration and Population Density from WorldBank and an assortment of surveyed indexes and financial data from the 2019 World Happiness Report. Both datasets contain empty data points, but we were able to maintain the majority of the data for all calculations. The data was merged through a left join on a combined key of Country name and year. Any and all precalculated attributes were removed in the cleansing process along with attributes with a majority of null values.

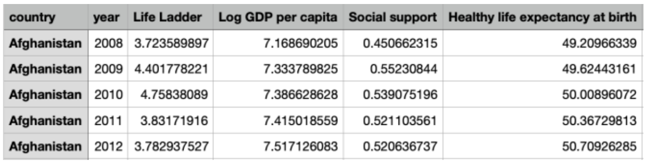
Final attributes used from the World Happiness Report include: Country, Year, Life Ladder, Log GDP Per Capita, Social Support, Life Expectancy, Freedom of Choice, Generosity, Corruption, Positive Affect, Negative Affect, Confidence in National Government, Democratic Quality, Delivery Quality, and Gini of Household Income. Life Ladder can be considered a measure of subjective well-being, containing values from 0 to 10. Social support was calculated from binary responses of each population of having healthy social support from relatives or friends. Generosity is based on the frequency of donations to charity. Corruption relates to perception on the individual level of both corporate and government corruption. Positive affect was determined from an individual's previous day overall level of happiness, laughter, and enjoyment. Negative affect was determined similarly with level of worry, sadness, and anger.

Two attributes: Population and Population Density were taken from the WorldBank dataset. These data points were collected to test our initial hypothesis that population density would play a significant part in the happiness index score.



country	year	population	pop_density	net_migration
Arab World	2018	419790588.0	37.3723653654038	
Arab World	2017	411898965.0	36.66980407291	-1408824.0
Arab World	2016	404024433.0	35.9687643273844	
Arab World	2015	396028278.0	35.2568969470266	
Arab World	2014	387907748.0	34.5339784528501	

Table 1: First four attributes of the raw data acquired from source WorldBank.



country	year	Life Ladder	Log GDP per capita	Social support	Healthy life expectancy at birth
Afghanistan	2008	3.723589897	7.168690205	0.450662315	49.20966339
Afghanistan	2009	4.401778221	7.333789825	0.55230844	49.62443161
Afghanistan	2010	4.75838089	7.386628628	0.539075196	50.00896072
Afghanistan	2011	3.83171916	7.415018559	0.521103561	50.36729813
Afghanistan	2012	3.782937527	7.517126083	0.520636737	50.70926285

Table 2: First four attributes of the raw data acquired from source World Happiness Report.

When determining what data points to keep, we operated under the assumption that we would not be using all of the given data for the regression models. We removed those columns with too many missing values as the predictive model would add a significant amount of noise if those were left in. The data itself includes 1562 rows and 17 attributes following our initial preprocessing steps with years ranging between 2005 and 2017 per country for 164 distinct countries.

Number of observations	1562 (~360 kB)
Number of attributes	17
Number of countries	164
Range of observed years	2005-2017
Number of original datasets	2
% missing values	5.63%

Table 3: Summary statistics of the merged raw data

This is a regression problem, so there are really 16 predictor attributes in our dataset and the target attribute is the life ladder shown in column 2 of Table 2 provided above. We determined that we needed to reduce the dimensionality of the data so as to allow the data to be more easily visualized in our Jupyter Notebooks as well as to help eliminate irrelevant features and reduce noise. We created the groups for each cluster strategy, applied PCA with 2 components, and fit the regression model to each group when iteratively testing solutions.

4. METHODOLOGY

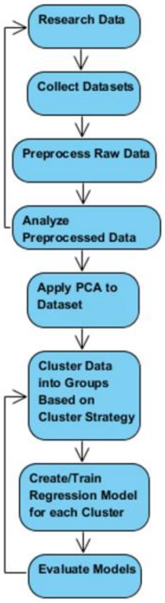

Research Data
Collect Datasets
Preprocess Raw Data
Analyze Preprocessed Data
Apply PCA to Dataset
Cluster Data into Groups Based on Cluster Strategy
Create/Train Regression Model for each Cluster
Evaluate Models

Figure 1: High-level flowchart summarizing process

We fit the data using the PCA model with 2 components, and we used the multiple linear regression model on each cluster to determine whether the clustering strategy increased the r-squared value and decreased the root mean square error value. If positive results were not seen, the clustering method was tossed out, and we analyzed the data to determine other possible methods to try next.

This is a clustering problem. For this problem we used DataComplete.csv as our preprocessed data to be analyzed and reduced. DataComplete.csv was created in Preprocessing.ipynb and is a compilation of all 16 predictive attributes and the target attribute.

Jupyter Notebooks:

- Preprocess.ipynb: this is the Jupyter notebook file that we used to drop attributes, merge datasets, and fill in missing values. This file also creates the merged csv data file "DataComplete.csv"

- Analyze.ipynb: this is the Jupyter notebook file that we used to analyze the attributes in the data and play around to determine possible clustering strategies
- Model.ipynb: this is the Jupyter notebook file that we used to reduce the dimensionality of the data and created the multiple linear regression models to analyze

5. EXPERIMENTAL EVALUATION

5.1 Experimental Setup

We used local machines to do all of the preprocessing and analysis of our data.

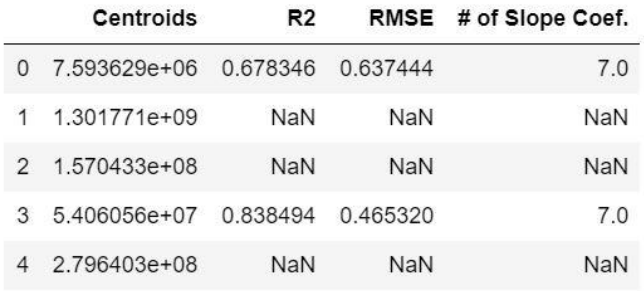
Our baseline model was a linear regression model trained with the whole dataset. On every model we created, PCA was applied to reduce the dimensions of the data from 16 attributes to the number of principal components that represented 85% of the explained variance of the data. Each dataset was divided with a test-train split of 30-70.

Out evaluation metric was root-mean-square error (RMSE) and R-squared (R2) values. When comparing the different linear regression models, we mostly relied on the R2 values since it is conveniently scaled to be at most 1. R2 values are said to be easily overfit, but as we applied PCA to each of our datasets, we believe the R2 score accurately reflects the success of our models.

5.2 Experimental Results

The baseline model (full dataset) was reduced to seven principal components and had results of RMSE = 0.6334 and R2 = 0.6645. From there, we aimed to find more accurate regression models by using clustering to find subsets within the dataset.

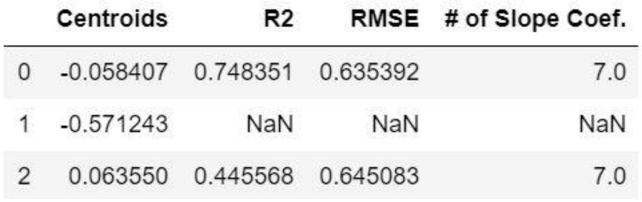
We looked at various attributes of the dataset and attempted to find clusters that formed the most accurate linear regression models. Interestingly, we found that clustering by the average population created models with the highest R2 values. To ensure that there was enough data in the dataset to train and test the model, we only used clusters that had over 100 lines in the subset of the data. As shown in the table below, both models formed from the largest clusters had R2 values higher than the baseline model. Since most of the clusters did not have enough data to train, it is difficult to determine whether population mean really is a good way to cluster countries when looking to predict life ladder scores.



	Centroids	R2	RMSE	# of Slope Coef.
0	7.593629e+06	0.678346	0.637444	7.0
1	1.301771e+09	NaN	NaN	NaN
2	1.570433e+08	NaN	NaN	NaN
3	5.406056e+07	0.838494	0.465320	7.0
4	2.796403e+08	NaN	NaN	NaN

Table 4: Centroid, R2, RMSE, and the number of attributes for each of the five clusters found when using average population. The clusters with less than 100 datasets contain NaN for all values other than centroid location.

Our next clustering method was based on trend in life ladder scores over the years reported. From the coefficients of these linear regression models, we applied K-means++ clustering to find three groups of countries that had similar trends in life ladder score. Then, using the data for each subset of countries, we formed new linear regression models to compare to our baseline model. Again, we only used clusters that had over 100 lines in the subset of the data.

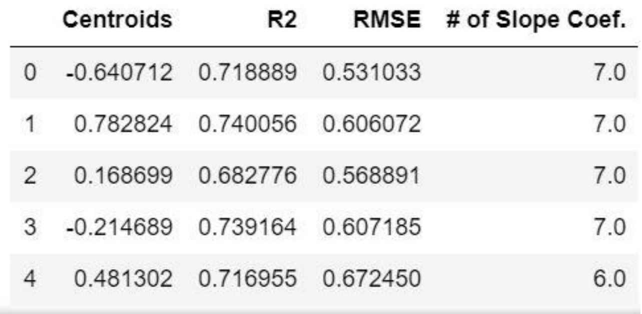


	Centroids	R2	RMSE	# of Slope Coef.
0	-0.058407	0.748351	0.635392	7.0
1	-0.571243	NaN	NaN	NaN
2	0.063550	0.445568	0.645083	7.0

Table 5: Centroid, R2, RMSE, and the number of attributes for each of the three clusters found when using life ladder score trends. The clusters with less than 100 datasets contain NaN for all values other than centroid location.

Of the two clusters we used to create linear regression models, only one had a higher R2 value than our baseline model. This was the cluster of countries with a very small negative trend in life ladder value. This could be because countries that don't have much change in life ladder score from year to year would be easier to predict. We can also see that each cluster that had enough data had its dimensionality reduced to seven attributes, much like the baseline model.

Lastly, we tried clustering based on the correlation coefficient between life ladder and delivery quality. We found these attributes to have strong correlation during our preprocessing steps, and were curious to see how the clusters would do when turned into linear regression models.



	Centroids	R2	RMSE	# of Slope Coef.
0	-0.640712	0.718889	0.531033	7.0
1	0.782824	0.740056	0.606072	7.0
2	0.168699	0.682776	0.568891	7.0
3	-0.214689	0.739164	0.607185	7.0
4	0.481302	0.716955	0.672450	6.0

Table 6: Centroid, R2, RMSE, and the number of attributes for each of the five clusters found when using correlation coefficient between life ladder and delivery quality.

These clusters were more evenly distributed, and we can see that the centroids for the clusters are more evenly spaced out. This is probably because correlation coefficient is limited to values between -1 and 1. The linear regression model for each cluster had a higher R2 value than the baseline. All but one of the clusters had a lower RMSE value as well. We can conclude that clustering the dataset by countries with a similar correlation coefficient between life ladder and delivery quality produced the best models for future prediction.

All in all, the project was successful because we were able to use clustering to form various linear regression models that improved upon the baseline model. We may have been able to get improve our models and get higher R2 values if we focused more on optimization. We found what appeared to be the best number of clusters when doing K-Means++ by manually trying different values. If we created a better way to test different cluster numbers, we may have had better results. Similarly, we chose the number of principal components when applying PCA by setting the percent of variance that needs to be explained to 85%. Trying more values and optimizing this may have contributed to a more robust model.

6. CONCLUSIONS

Through our analysis, we were able to find linear regression models that were optimized for certain clusters of countries. We formed these clusters based on average population, trends in life ladder scores, and correlation coefficients between the life ladder score and delivery quality. We found the linear regression models formed from clusters based on the correlation coefficients between the life ladder score and delivery quality had the most effective results, with five percent higher R2 values on average.

In this project, we were limited by the amount of available data on every country. Ideally, we would have liked to find more attributes over a larger span of years to more thoroughly develop our models. For example, when clustering, we found that some country clusters did not have enough data to make robust linear regression models. Furthermore, we would like to experiment with more clustering methods. Most of the clustering we did was with one variable, and it would be interesting to see if creating clusters based on a larger set of attributes effects the accuracy of the linear regression models.

7. REFERENCES

[1] Bronshtein, Adi. *Simple and Multiple Linear Regression in Python.* Towards Data Science, May 8, 2017.

[2] Bronshtein, Adi. *Train/Test Split and Cross Validation in Python.* Towards Data Science, May 17, 2017.

[3] Helliwell, John F., Huang, Haifang and Wang, Shun. *Changing World Happiness.* World Happiness Report. March 20, 2019.

[4] World Bank Group. *Migration Population.* 2019.