What contributes to Happiness?

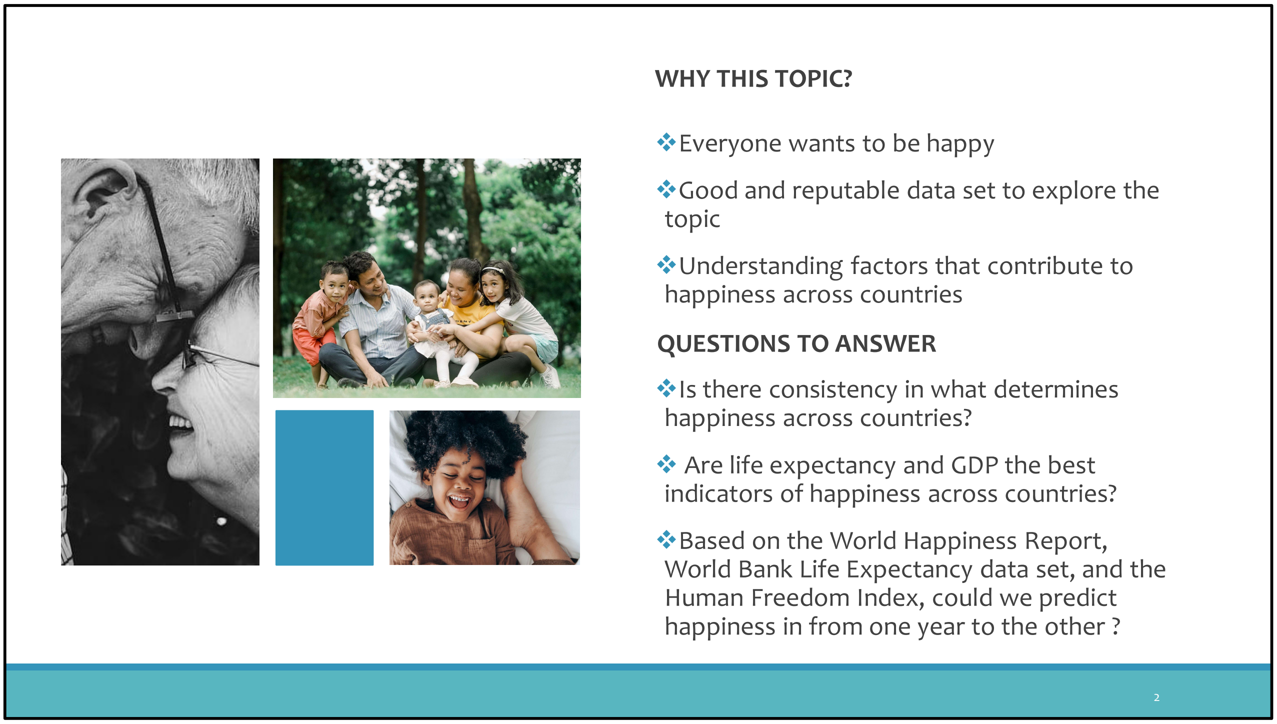
HAPPINESS AND ITS CONTRIBUTING FACTORS
BY
ASSITAN CISSE, DIANA BORKAR, MERELYNN (LYNN) OKANG & GLORIA YAHOUEDEOU

1

Lynn –

Good evening everyone! Our group will be presenting its findings on the topic of happiness and its contributing factors . We know that everyone wants to be happy, and so we wanted to explore what defined happiness.

Our group was composed of **Assitan Cisse, Diana Borkar, Gloria Yahouedeou, and myself, Merelynn (Lynn) Okang .**



WHY THIS TOPIC?

- Everyone wants to be happy
- Good and reputable data set to explore the topic
- Understanding factors that contribute to happiness across countries

QUESTIONS TO ANSWER

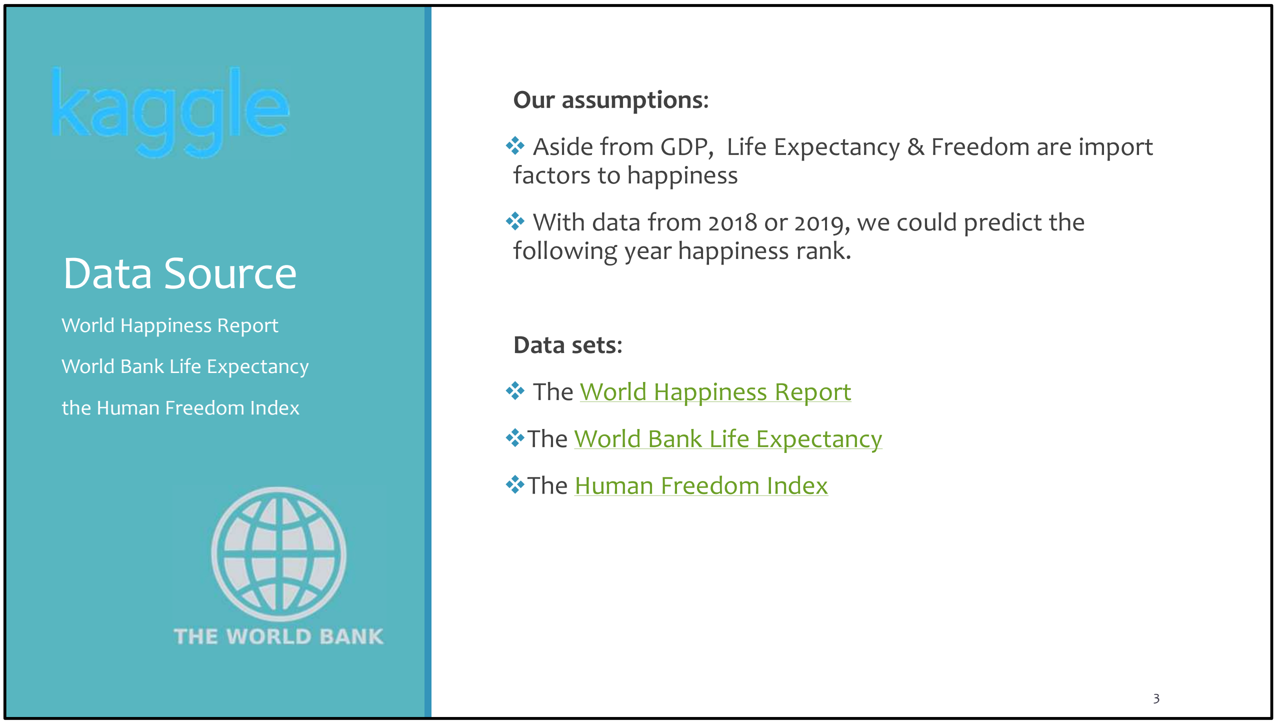
- Is there consistency in what determines happiness across countries?
- Are life expectancy and GDP the best indicators of happiness across countries?
- Based on the World Happiness Report, World Bank Life Expectancy data set, and the Human Freedom Index, could we predict happiness in from one year to the other ?

2

Lynn – We chose the topic of happiness because it was something that we could all agree that we want to be, and also, between you and I, we were out of ideas.

Another reason for choosing this topic was the readily available data set from the World Happiness Report. It was a clean, extensive, and is reputable data set that we could dive in and learn from.

Once we were clear on the topic we wanted to discuss, we had a few predictions of the outcome. From the get-go, we assumed that GDP, would play significantly role but also wanted to consider Life Expectancy and Freedom and how they impacted happiness.



kaggle

Data Source

World Happiness Report

World Bank Life Expectancy

the Human Freedom Index

THE WORLD BANK

Our assumptions:

- ❖ Aside from GDP, Life Expectancy & Freedom are import factors to happiness
- ❖ With data from 2018 or 2019, we could predict the following year happiness rank.

Data sets:

- ❖ The World Happiness Report
- ❖The World Bank Life Expectancy
- ❖The Human Freedom Index

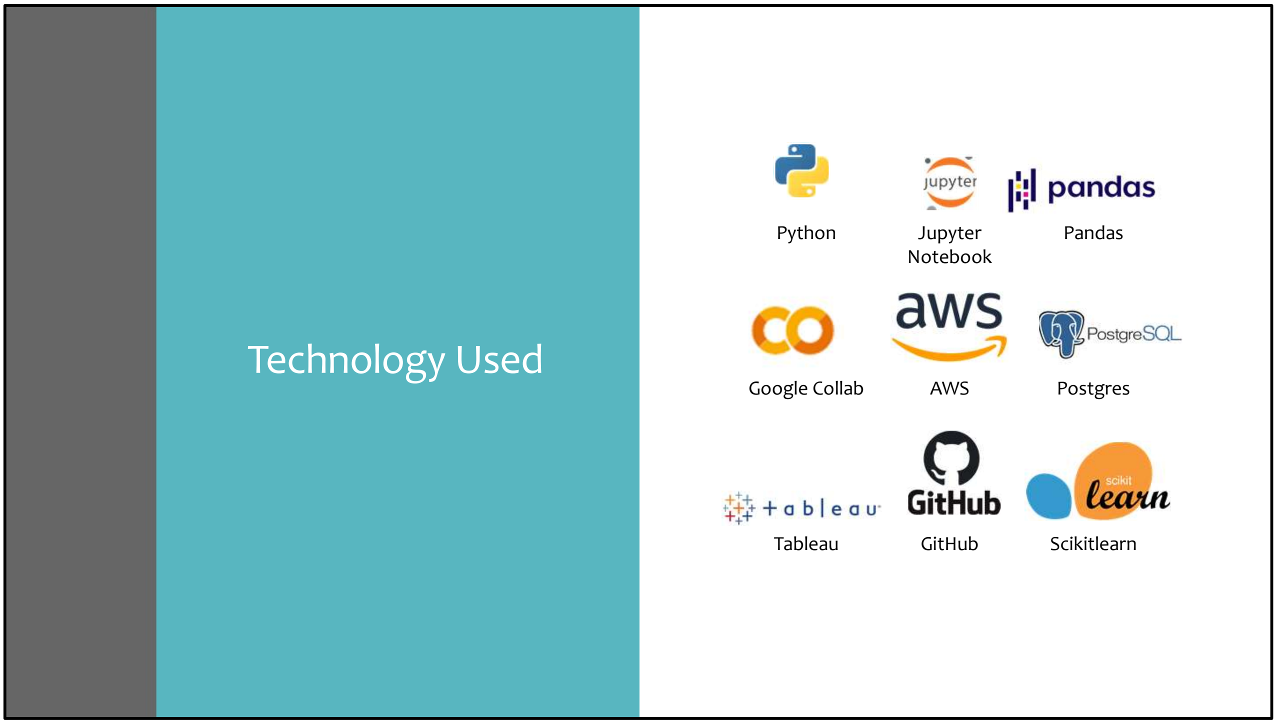
Lynn –

As mentioned earlier we predicted that GDP would play a role, but we wanted know how important Life Expectancy & Freedom would contribute to the happiness score. We also intended to use data from 2018 and 2019, so that we could predict the following year happiness score… at least that was the plan.

In order to address our main questions, we used various data sets from Kaggle and the World Bank.

The main dataset was the World Happiness Report from Kaggle. It is a landmark survey of the state of global happiness. The data set had a collection of indicators on more than 140 countries around the world including happiness rank, happiness score on a scale of 0 to 10, standard error, and more…

We also used the World Bank's Life Expectancy data set and from Kaggle, the Human Freedom Index.



Technology Used

Python

Jupyter Notebook

Pandas

Google Collab

AWS

Postgres

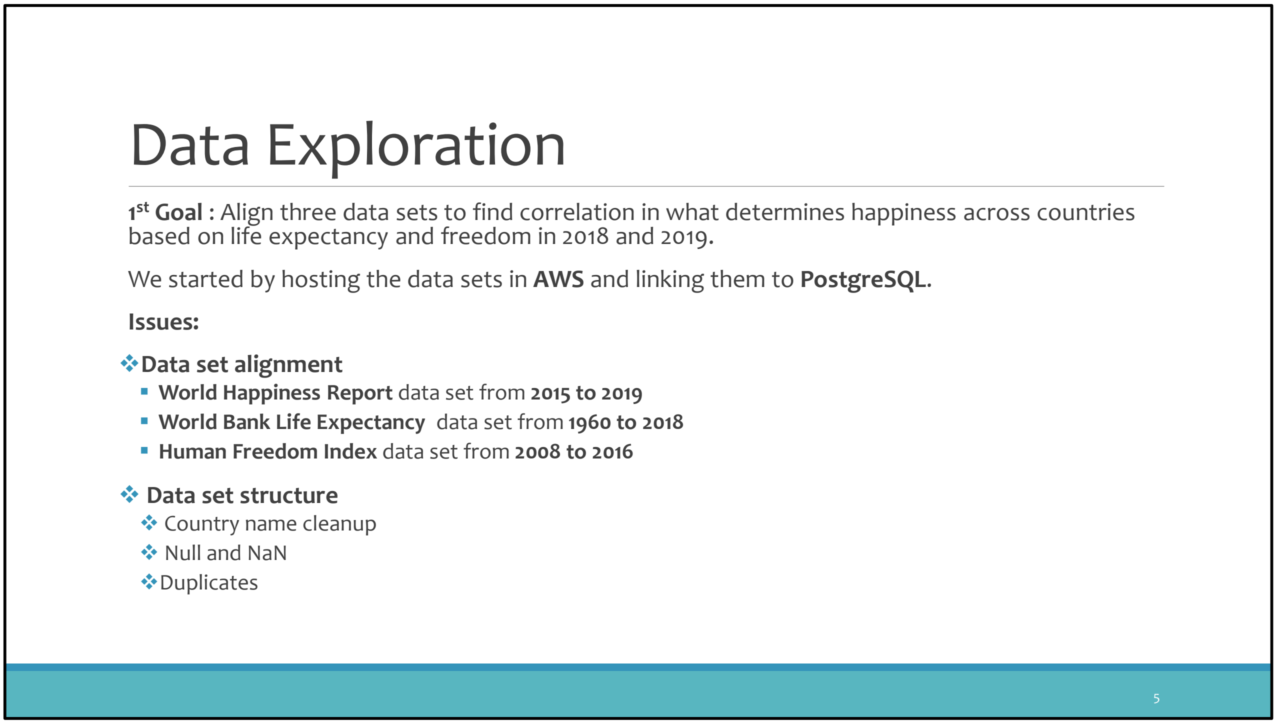
Tableau

GitHub

Scikitlearn

Lynn – To come up with our answers during this project we used:

- Python, Jupyter Notebook, Pandas, Scikitlearn, and Google Collab, to clean the data and create the machine learning model
- AWS, and Postgres to store the data
- Tableau to create the dashboard
- And GitHub to submit our work.



Data Exploration

1st Goal : Align three data sets to find correlation in what determines happiness across countries based on life expectancy and freedom in 2018 and 2019.

We started by hosting the data sets in **AWS** and linking them to **PostgreSQL**.

Issues:

- **Data set alignment**
  - **World Happiness Report** data set from **2015 to 2019**
  - **World Bank Life Expectancy** data set from **1960 to 2018**
  - **Human Freedom Index** data set from **2008 to 2016**
- **Data set structure**
  - Country name cleanup
  - Null and NaN
  - Duplicates

Gloria – When we started our analysis, we had for goal to align all three data sets on a specific year, such as 2018 and 2019, and we wanted to find the correlation between the features that we identified and the happiness score.

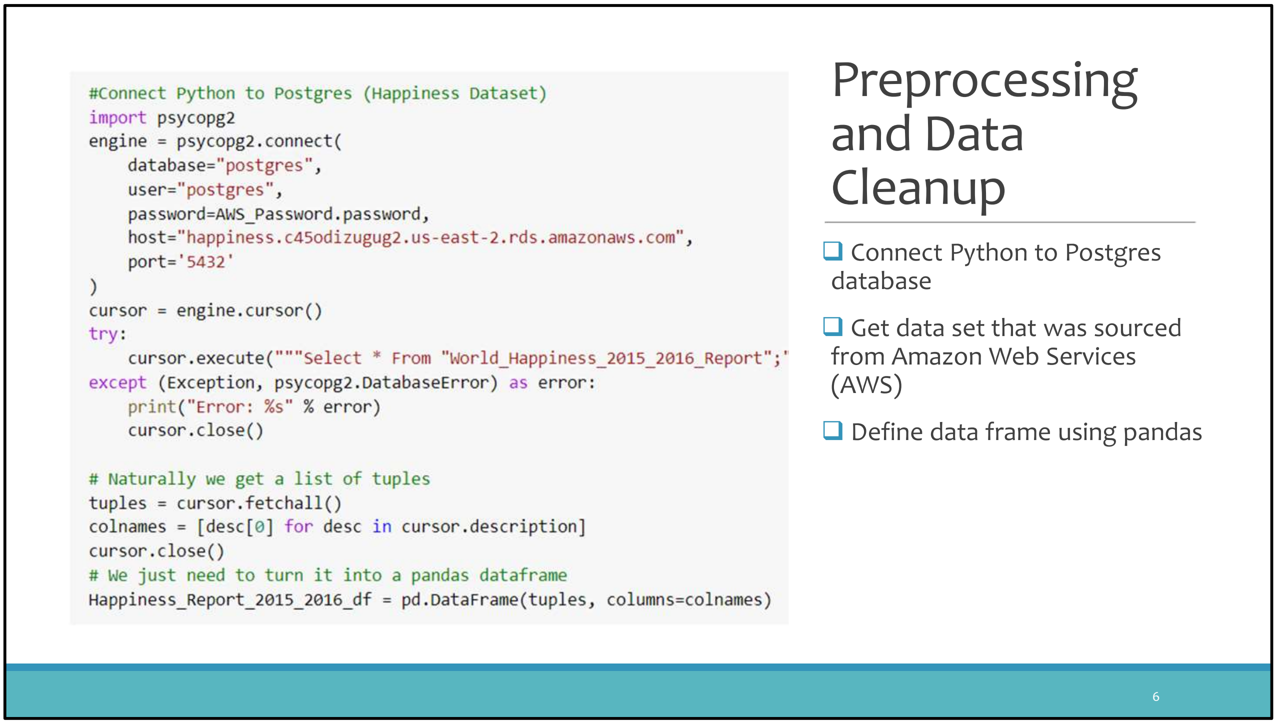
We started by hosting the csv files in AWS, so that we could easily pull information at any time into PostgreS and create the necessary tables we needed.

As we started manipulating the data in PGAdmin, we noticed that each data set had various time range:

- **World Happiness Report** data set from **2015 to 2019**
- **World Bank Life Expectancy** data set from **1960 to 2018**
- **Human Freedom Index** data set from **2008 to 2016**

Additionally, some of the data structure were drastically different from one another with duplicates countries, different countries name formatting etc… We knew that there would be a few changes to the plan that would need to take place.

One of those changes was to now focus on the year 2015 and 2016.



```
#Connect Python to Postgres (Happiness Dataset)
import psycopg2
engine = psycopg2.connect(
    database="postgres",
    user="postgres",
    password=AWS_Password.password,
    host="happiness.c45odizugug2.us-east-2.rds.amazonaws.com",
    port='5432'
)
cursor = engine.cursor()
try:
    cursor.execute("""Select * From "World_Happiness_2015_2016_Report";'
except (Exception, psycopg2.DatabaseError) as error:
    print("Error: %s" % error)
    cursor.close()

# Naturally we get a list of tuples
tuples = cursor.fetchall()
colnames = [desc[0] for desc in cursor.description]
cursor.close()
# We just need to turn it into a pandas dataframe
Happiness_Report_2015_2016_df = pd.DataFrame(tuples, columns=colnames)
```

Preprocessing and Data Cleanup

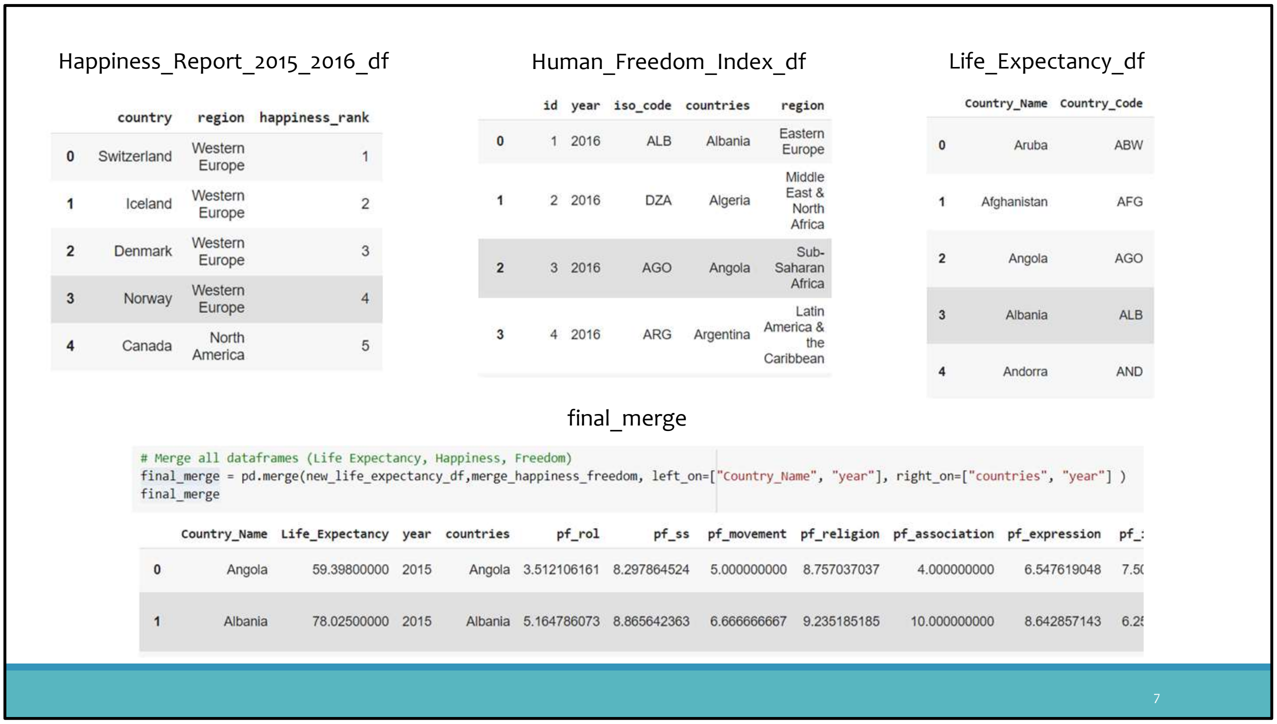
- ❑ Connect Python to Postgres database
- ❑ Get data set that was sourced from Amazon Web Services (AWS)
- ❑ Define data frame using pandas

Gloria – After creating the tables for 2015 and 2016, for each data set, in PGAdmin, and creating a merged table for the data we went to our Jupyter notebook file and started to do a bit of preprocessing using **psycopg2**. In Jupyter notebook, we connected to Postgres, which had data that was stored on the AWS server, and then we defined data frames using pandas.

All through the preprocessing stage, we systematically cleaned up each of the tables we wanted to use to focus on the minimum required fields to create a merged data set in Jupyter Notebook incorporating:

- The same year (2015-2016)
- Same country name
- Same amount of data with no null or duplicates
- Same data type

First, we cleaned up and merged the 2015 and 2016 world happiness data frame, then did the same for the human freedom index data frame, followed by the life expectancy data frame, and finally we merged all three into one data frame.



Happiness_Report_2015_2016_df

	country	region	happiness_rank
0	Switzerland	Western Europe	1
1	Iceland	Western Europe	2
2	Denmark	Western Europe	3
3	Norway	Western Europe	4
4	Canada	North America	5

Human_Freedom_Index_df

	id	year	iso_code	countries	region
0	1	2016	ALB	Albania	Eastern Europe
1	2	2016	DZA	Algeria	Middle East & North Africa
2	3	2016	AGO	Angola	Sub-Saharan Africa
3	4	2016	ARG	Argentina	Latin America & the Caribbean

Life_Expectancy_df

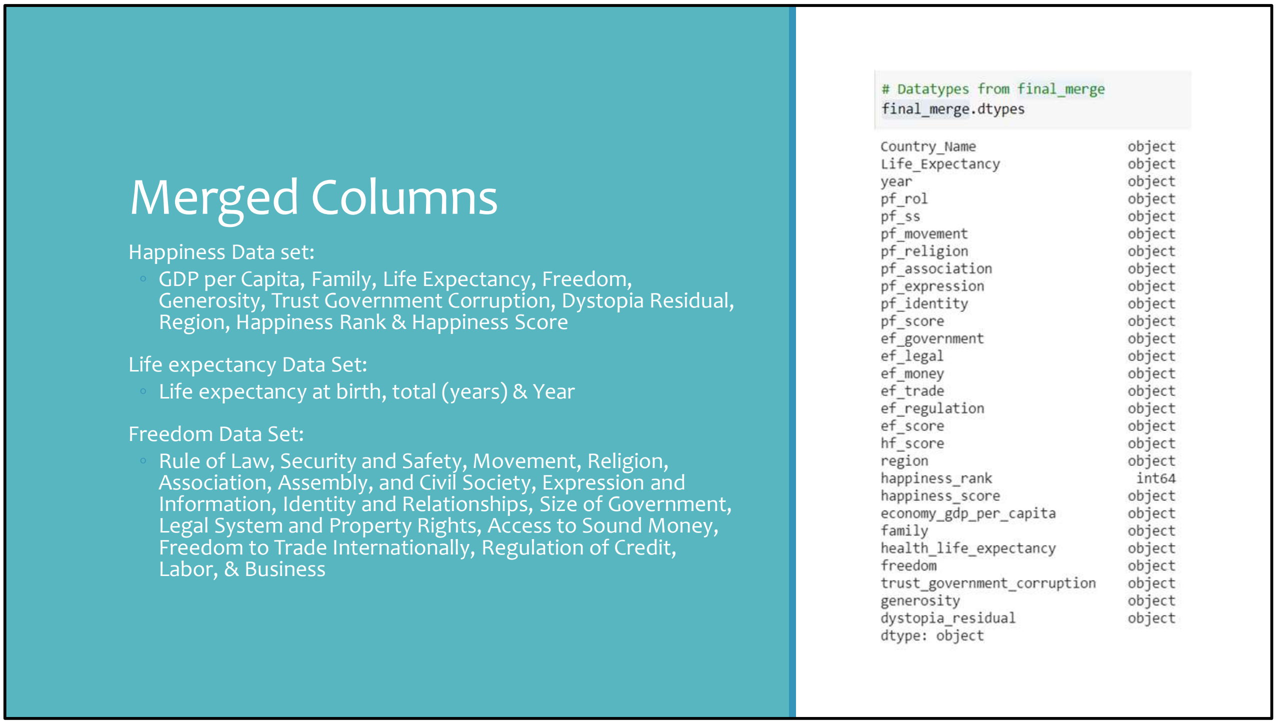
	Country_Name	Country_Code
0	Aruba	ABW
1	Afghanistan	AFG
2	Angola	AGO
3	Albania	ALB
4	Andorra	AND

final_merge

```
# Merge all dataframes (Life Expectancy, Happiness, Freedom)
final_merge = pd.merge(new_life_expectancy_df,merge_happiness_freedom, left_on=["Country_Name", "year"], right_on=["countries", "year"] )
final_merge
```

	Country_Name	Life_Expectancy	year	countries	pf_rol	pf_ss	pf_movement	pf_religion	pf_association	pf_expression	pf_
0	Angola	59.39800000	2015	Angola	3.512106161	8.297864524	5.000000000	8.757037037	4.000000000	6.547619048	7.5
1	Albania	78.02500000	2015	Albania	5.164786073	8.865642363	6.666666667	9.235185185	10.000000000	8.642857143	6.2

This is an example of our a final-merge data frame



Merged Columns

Happiness Data set:

- GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption, Dystopia Residual, Region, Happiness Rank & Happiness Score

Life expectancy Data Set:

- Life expectancy at birth, total (years) & Year

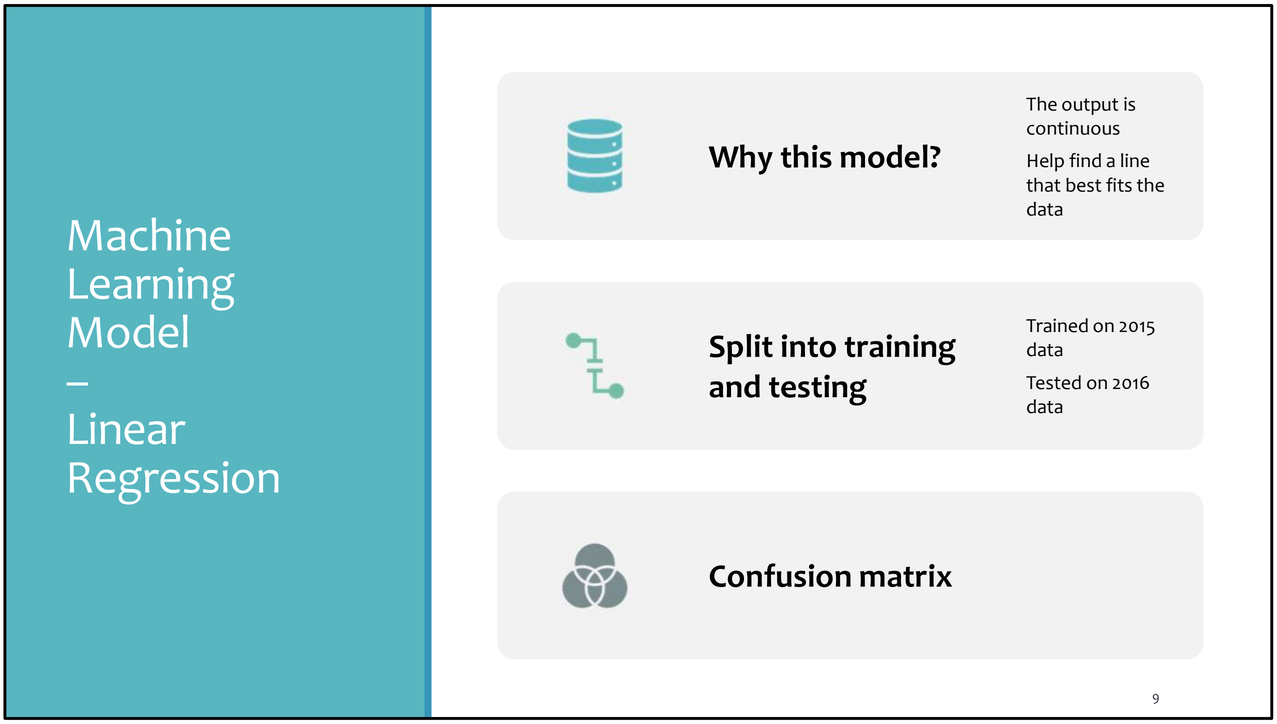
Freedom Data Set:

- Rule of Law, Security and Safety, Movement, Religion, Association, Assembly, and Civil Society, Expression and Information, Identity and Relationships, Size of Government, Legal System and Property Rights, Access to Sound Money, Freedom to Trade Internationally, Regulation of Credit, Labor, & Business

```
# Datatypes from final_merge
final_merge.dtypes
```

```
Country_Name                     object
Life_Expectancy                  object
year                             object
pf_rol                           object
pf_ss                            object
pf_movement                      object
pf_religion                      object
pf_association                   object
pf_expression                    object
pf_identity                      object
pf_score                         object
ef_government                    object
ef_legal                         object
ef_money                         object
ef_trade                         object
ef_regulation                    object
ef_score                         object
hf_score                         object
region                           object
happiness_rank                    int64
happiness_score                  object
economy_gdp_per_capita           object
family                           object
health_life_expectancy           object
freedom                          object
trust_government_corruption      object
generosity                       object
dystopia_residual                object
dtype: object
```

Gloria - We had 28 columns in our final merged dataframe. Out of those, we were going to pay special attention to Life expectancy, economy gdp per capita, and pf_expression features. We also wanted to see the correlation between all of these features and this data set was going to feed into our machine learning model.



Machine Learning Model – Linear Regression

Why this model?

The output is continuous

Help find a line that best fits the data

Split into training and testing

Trained on 2015 data

Tested on 2016 data

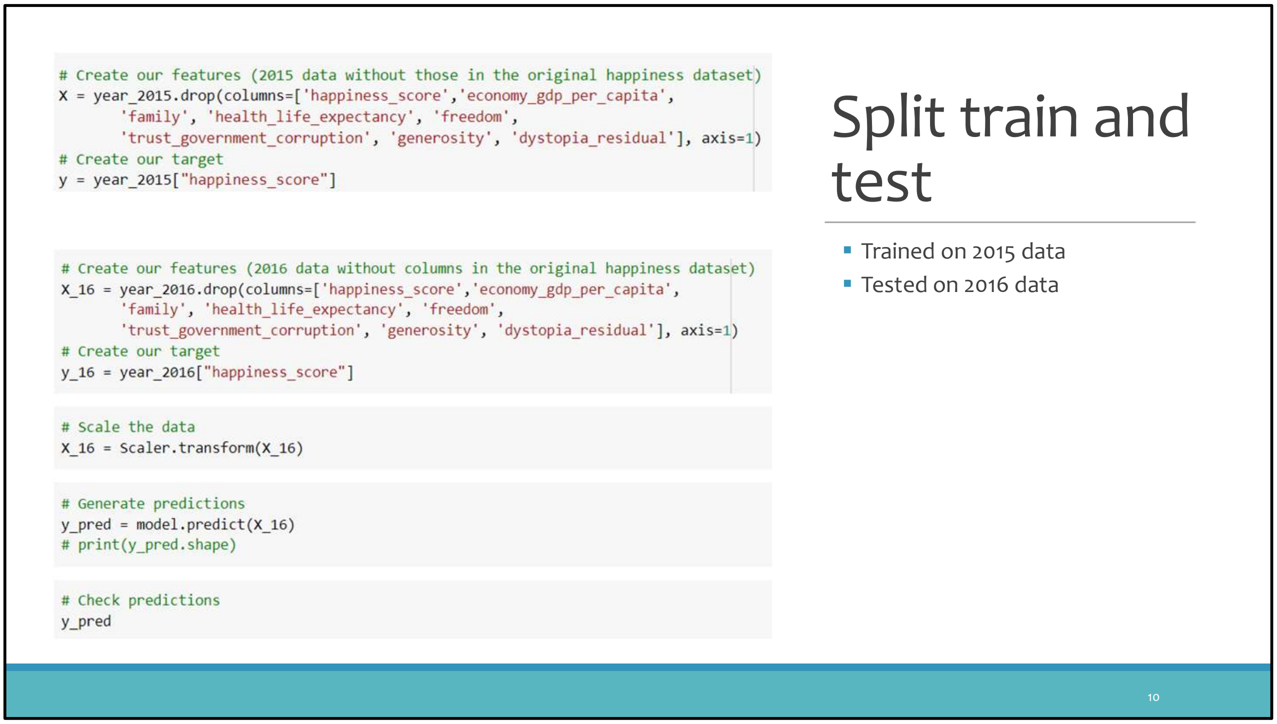
Confusion matrix

Diana – For our data, our group chose to use a supervised linear regression machine learning model to predict the happiness score of a country based on the features in the dataset. We needed a supervised model because we knew the outcome and our data was numerical.

The benefits of using a linear regression model are that it will be able to predict the happiness score, a continuous variable, based on the features of the data. A limitation of the linear regression model is that is assumes a linear relationship between the features and target and could miss some outliers or results that are not directly correlated.

The target, y, is meant to be the output that the machine learning model will try to predict. For this project, the target was the happiness score, which was created by using just that column from our final data frame. The features for our machine learning model were selected by dropping the year and all columns from the happiness dataset and using all the other columns as features, specifically those from the Life Expectancy and freedom datasets.

We had some trial and error to select features with different levels of correlation with the target, but the best results came from using all columns other than those originally used to determine the happiness score in the happiness dataset.



```
# Create our features (2015 data without those in the original happiness dataset)
X = year_2015.drop(columns=['happiness_score','economy_gdp_per_capita',
       'family', 'health_life_expectancy', 'freedom',
       'trust_government_corruption', 'generosity', 'dystopia_residual'], axis=1)
# Create our target
y = year_2015["happiness_score"]
```

```
# Create our features (2016 data without columns in the original happiness dataset)
X_16 = year_2016.drop(columns=['happiness_score','economy_gdp_per_capita',
       'family', 'health_life_expectancy', 'freedom',
       'trust_government_corruption', 'generosity', 'dystopia_residual'], axis=1)
# Create our target
y_16 = year_2016["happiness_score"]
```

```
# Scale the data
X_16 = Scaler.transform(X_16)
```

```
# Generate predictions
y_pred = model.predict(X_16)
# print(y_pred.shape)
```

```
# Check predictions
y_pred
```

Split train and test

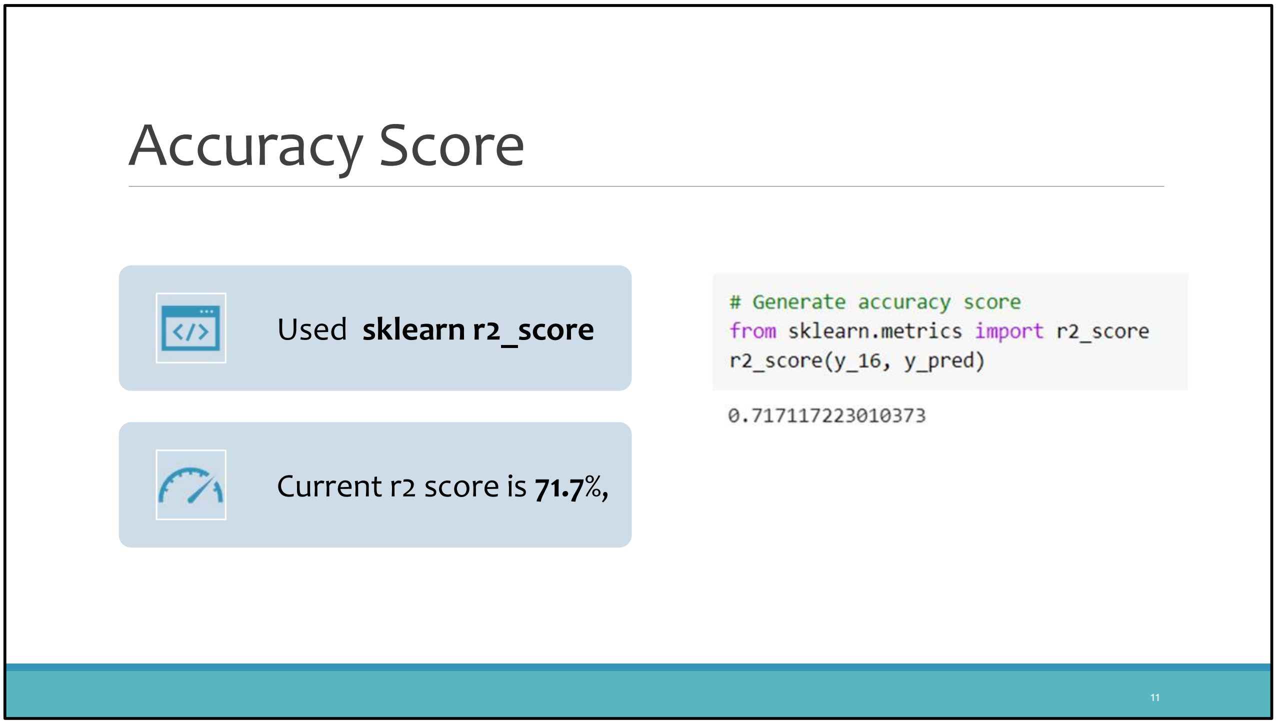
- Trained on 2015 data
- Tested on 2016 data

10

Diana - We split the final data frame by year into 2015 and 2016.

The training for the linear regression model was conducted by fitting the 2015 data to the model, with the features mentioned previously, and then the model to be tested with the 2016 data that the model had not previously seen. Future iterations of training the model can be done by changing the training parameters and/or changing the model features to include different columns, such as and those with a high correlation to the happiness score.

The method of training with 2015 data and testing on 2016 data that the model had not seen before, produced a much higher accuracy score compared to using Scikit-Learn's train_test_split on the whole merged dataset.



Accuracy Score

Used **sklearn r2_score**

Current r2 score is **71.7**%,

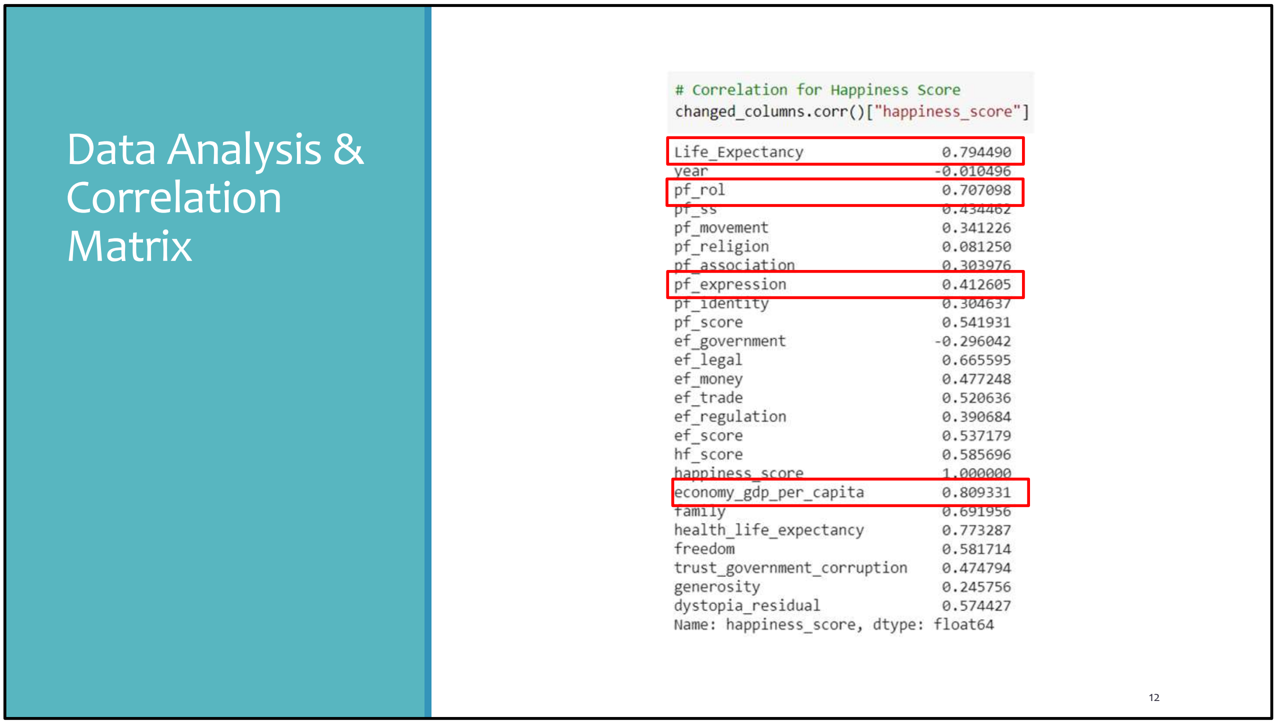
```
# Generate accuracy score
from sklearn.metrics import r2_score
r2_score(y_16, y_pred)
```

```
0.717117223010373
```

Diana - Our machine learning model used the sklearn r2_score, coefficient of determination, to assess the accuracy of the model.

The current r2 score is **71.7%**, which reveals that nearly 72% of the predicted data fits the linear regression model. This also means that the happiness score is correctly predicted nearly 72% of the time.

The current accuracy score can be improved upon, but it remains the most accurate so far, especially compared to alternate features that were tested using sklearn's train-test-split and received accuracy scores under 45%.



Data Analysis & Correlation Matrix

```
# Correlation for Happiness Score
changed_columns.corr()["happiness_score"]
```

```
Life_Expectancy                0.794490
year                          -0.010496
pf_rol                         0.707098
pf_ss                          0.434462
pf_movement                    0.341226
pf_religion                    0.081250
pf_association                 0.303976
pf_expression                  0.412605
pf_identity                    0.304637
pf_score                       0.541931
ef_government                 -0.296042
ef_legal                       0.665595
ef_money                       0.477248
ef_trade                       0.520636
ef_regulation                  0.390684
ef_score                       0.537179
hf_score                       0.585696
happiness_score                1.000000
economy_gdp_per_capita         0.809331
family                         0.691956
health_life_expectancy         0.773287
freedom                        0.581714
trust_government_corruption    0.474794
generosity                     0.245756
dystopia_residual              0.574427
Name: happiness_score, dtype: float64
```

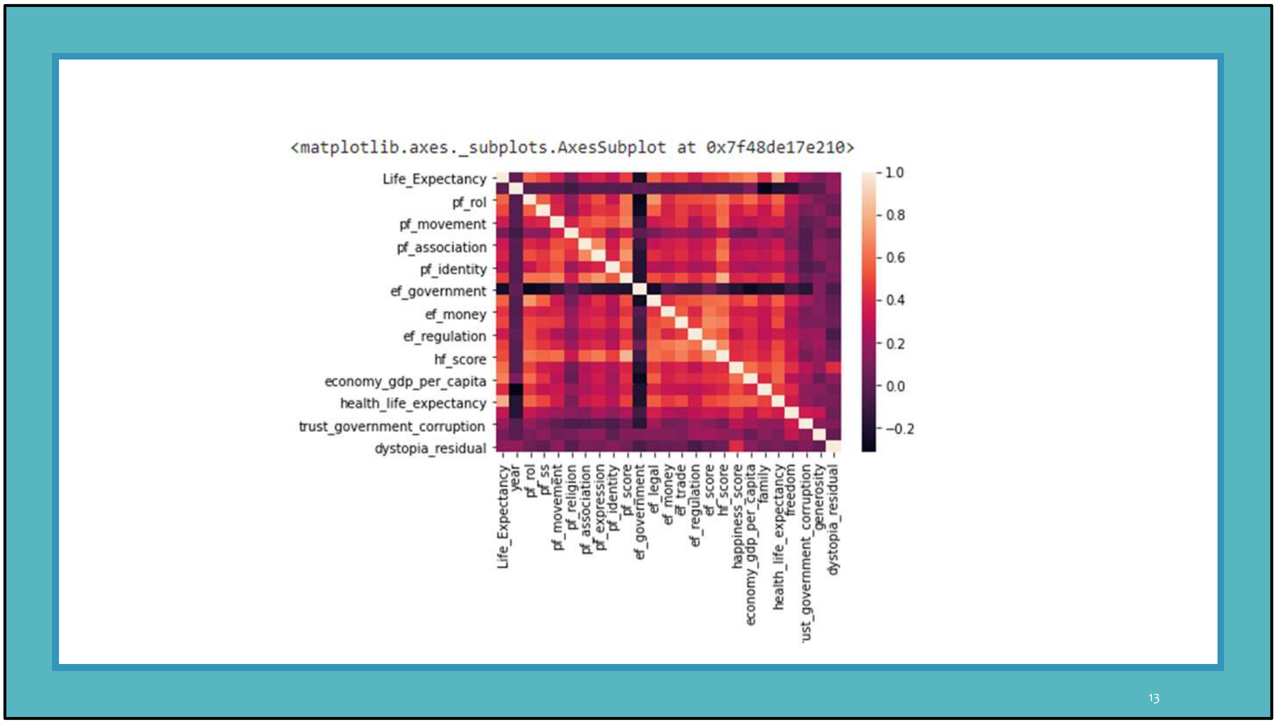
12

Diana – This image show the correlation between the happiness score and our features.

We found that there was a strong correlation of between ‘Life Expectancy’ and ‘Happiness Score’, as the score was 0.79, which is closer to the maximum correlation of 1. ‘Rule of Law’ and the ‘Happiness Score’ also are strongly correlated with a score of 0.70.

There was also a strong correlation of 0.81 between ‘Economy GDP per Capita’ and ‘Happiness Score’, and a low correlation of 0.41 between “Freedom of expression” and ‘Happiness Score”, which we were not expecting in our predictions.

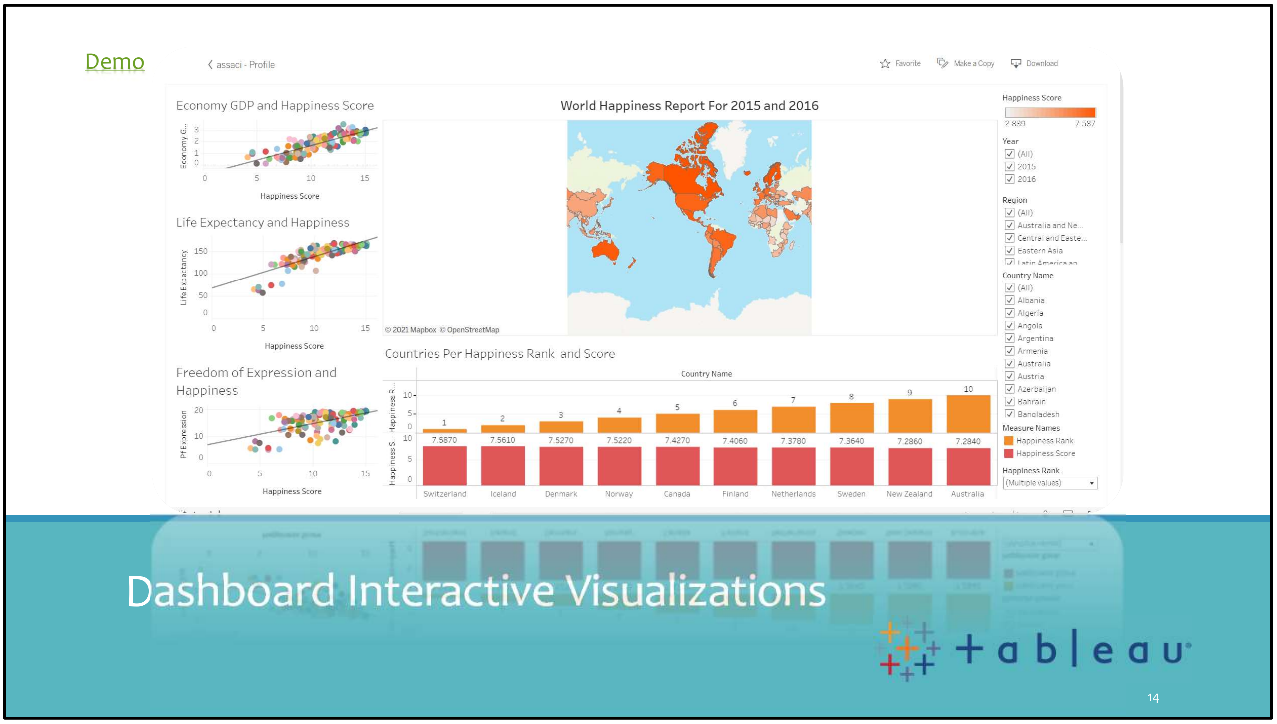
The least relevant factors seemed to be religion, generosity, and trust in the government.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f48de17e210>
```

Diana – This is a heatmap of the correlation matrix with all the columns. The lighter square are most strongly correlated and the darker squares at the least correlated.

In addition to this visual, we used Tableau to create our dashboard, which Assitan will demonstrate.



Demo

‹ assaci - Profile

☆ Favorite   Make a Copy   Download

Economy GDP and Happiness Score
Economy G..
Happiness Score
Life Expectancy and Happiness
Life Expectancy
Happiness Score
Freedom of Expression and Happiness
Pf Expression
Happiness Score
World Happiness Report For 2015 and 2016
© 2021 Mapbox © OpenStreetMap
Countries Per Happiness Rank and Score
Country Name
Happiness R..
Happiness S..
1 2 3 4 5 6 7 8 9 10
7.5870 7.5610 7.5270 7.5220 7.4270 7.4060 7.3780 7.3640 7.2860 7.2840
Switzerland Iceland Denmark Norway Canada Finland Netherlands Sweden New Zealand Australia
Happiness Score
2.839 7.587

Year
☑ (All)
☑ 2015
☑ 2016

Region
☑ (All)
☑ Australia and Ne...
☑ Central and Easte...
☑ Eastern Asia
☑ Latin America an...

Country Name
☑ (All)
☑ Albania
☑ Algeria
☑ Angola
☑ Argentina
☑ Armenia
☑ Australia
☑ Austria
☑ Azerbaijan
☑ Bahrain
☑ Bangladesh

Measure Names
🟧 Happiness Rank
🟥 Happiness Score

Happiness Rank
(Multiple values) ▾

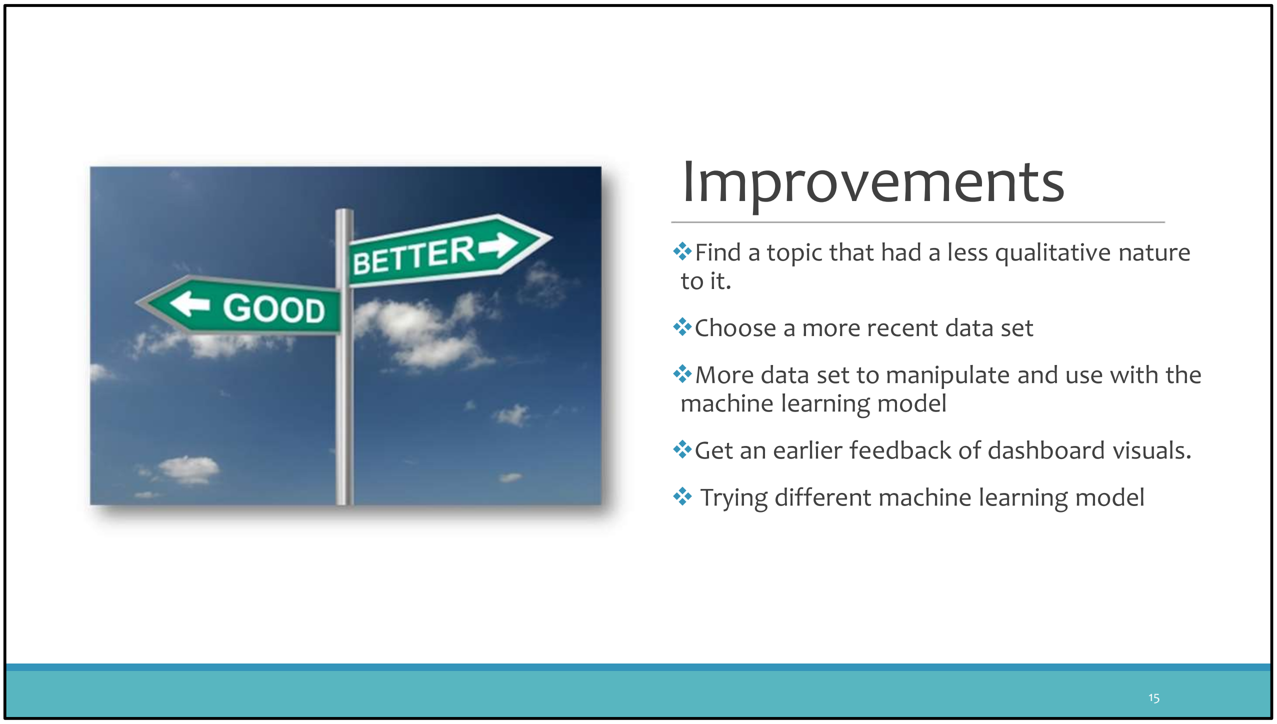
Dashboard Interactive Visualizations

+ableau

14

Assitan – Using Tableau, we have been able to visualize a few findings regarding some our initial assumptions and we also picked up a few interesting bit of information.

We created an interactive dashboard that provided the happiness score for each countries across a map, filterable by year, by region, and by rank. We also showed correlation between Economy GDP, Life Expectancy, Freedom of expression and Happiness score. The visuals for the correlation confirm, what Diana shared earlier. There is a strong correlation between Economy GDP, Life Expectancy and the Happiness score, but for freedom of expression there is a lower correlation.

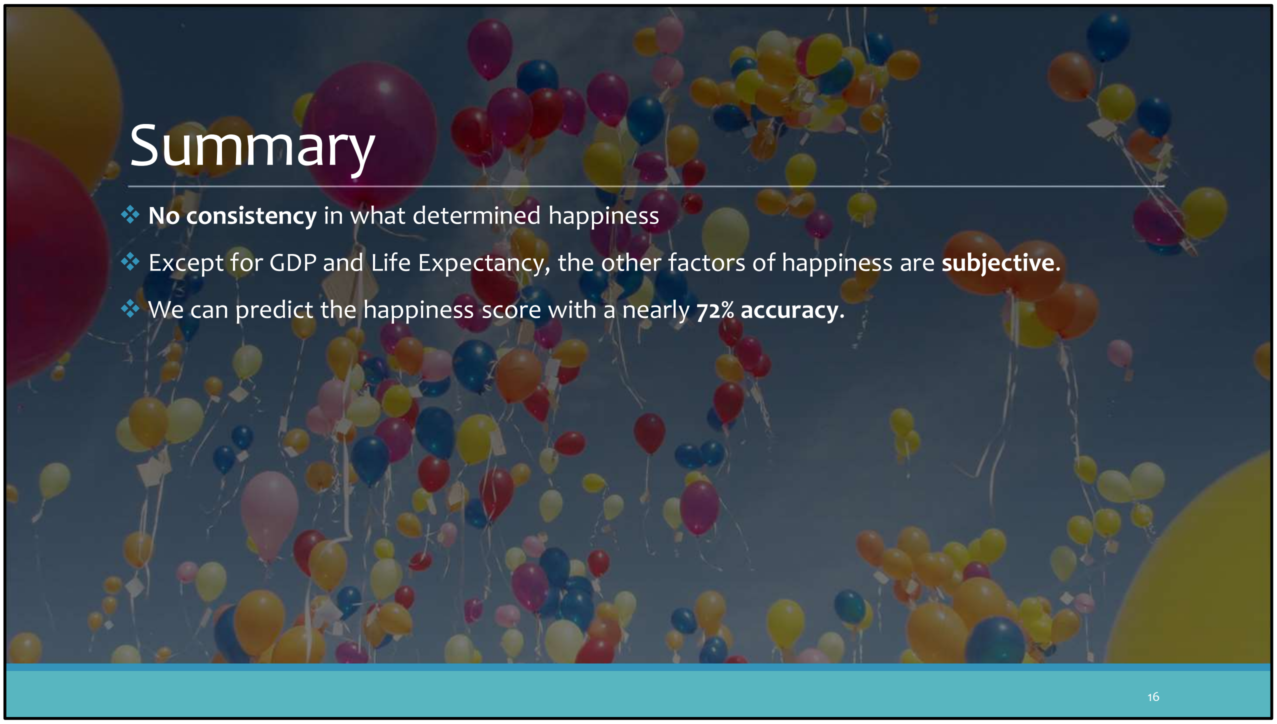

GOOD
BETTER

Improvements

- ❖Find a topic that had a less qualitative nature to it.
- ❖Choose a more recent data set
- ❖More data set to manipulate and use with the machine learning model
- ❖Get an earlier feedback of dashboard visuals.
- ❖ Trying different machine learning model

Lynn – Overall, we were proud of this work, and we were able to apply many of the principles we've learned during this bootcamp. However, if we had to do this project over, we would:

- Maybe find a Topic that had a less qualitative nature to it. With this topic finding factors with a solid correlation was difficult.
- Choose a more recent data set
- Get even more data set to manipulate and use with the machine learning model
- Get an earlier feedback of dashboard visuals
- And Try different model to see if the accuracy score would increase



Summary

- **No consistency** in what determined happiness
- Except for GDP and Life Expectancy, the other factors of happiness are **subjective.**
- We can predict the happiness score with a nearly **72% accuracy.**

16

Lynn – We started this project with questions about consistency in what determined happiness, whether or not factors such as life expectancy or GDP were the best indicators of happiness and trying to find if we could predict happiness scoring from year to the other.

Through our work, we concluded that happiness is a subjective topic that it is hard to predict, at the very least hard for us to predict. Features have a different weight and importance, for different people in different countries. We logically came to the same conclusion that everyone keeps on saying “everyone has a different definition for happiness”.

Our analysis did show one thing though: life expectancy and a country's GDP are the two features most impactful on happiness. The higher of both these categories were, the higher was the happiness score, so maybe money does make happiness?

Thank you!