SUICIDE RATE OVERVIEW

DATA 603

SUBMITTED TO: PROFESSOR ANDREW ENKEBOLL

SUBMITTED BY: ANUJA YAWAR

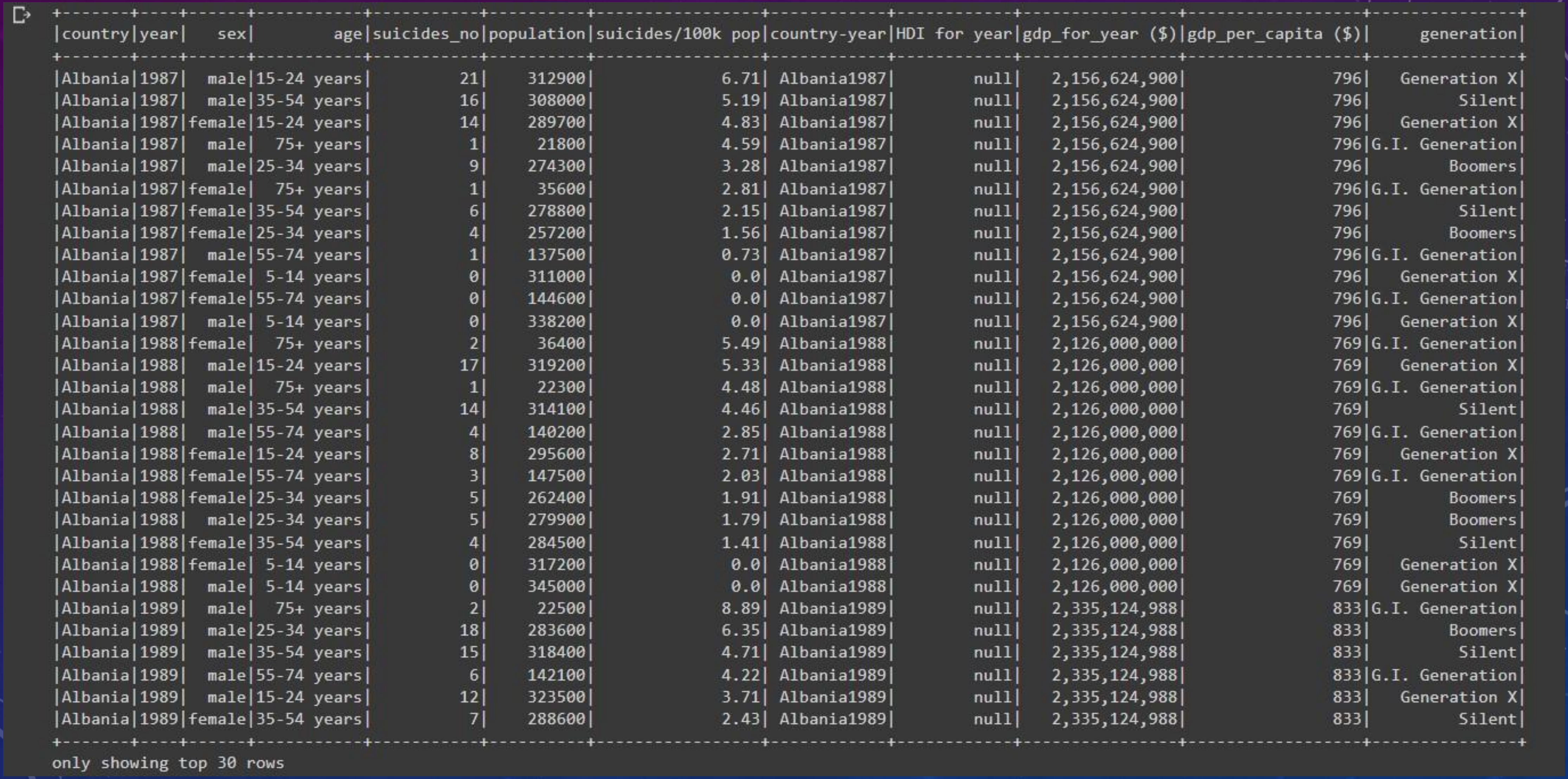
ABOUT THE DATA

- Globally, the availability and quality of data on suicide and suicide attempts is poor. Only some 80 Member States have good-quality vital registration data that can be used directly to estimate suicide rates. This problem of poor-quality mortality data is not unique to suicide, but given the sensitivity of suicide – and the illegality of suicidal behavior in some countries – it is likely that under-reporting and misclassification are greater problems for suicide than for most other causes of death.
- The data set is compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum from the year 1985-2016.
- The data set contains 27820 data points and 12 features such as sex, age-group, year, gross domestic product, population etc.
- Data Set Source – Kaggle

WHAT ARE WE TRYING TO DO HERE?

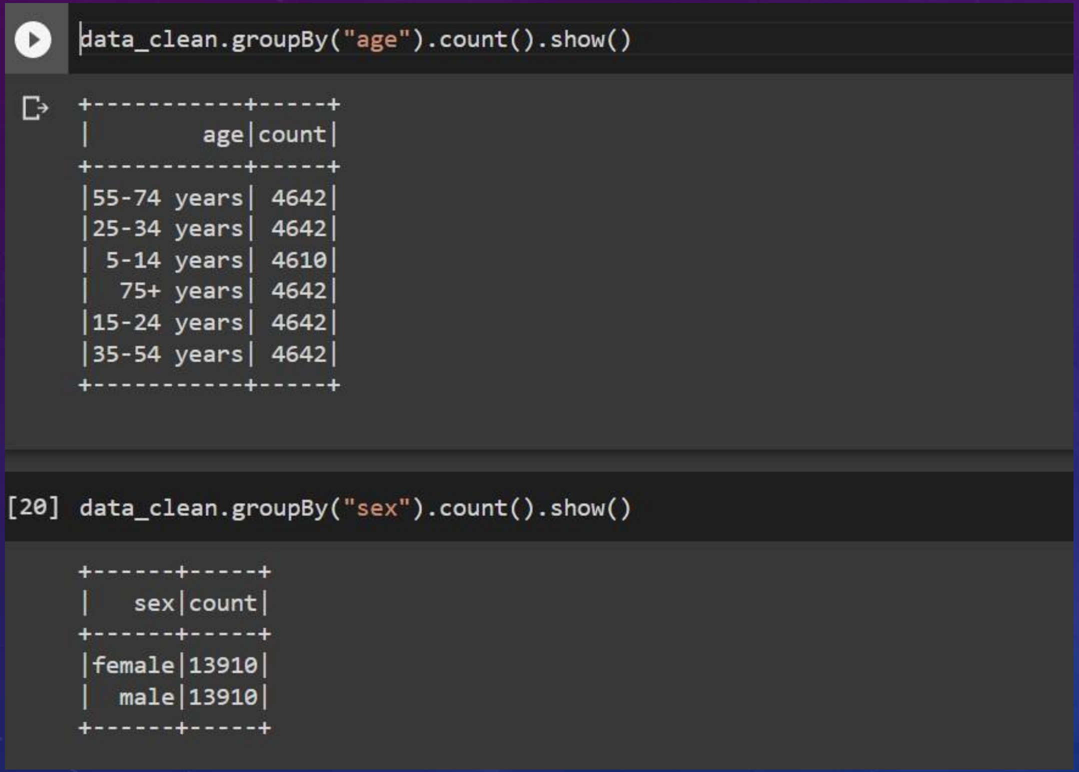
- Perform Exploratory Data Analysis to pre-process the data and gathering insights from it using PySpark.
- Building Machine Learning model to predict the rate of Suicide gobally using PySpark Machine Learning libraries
- Finally, using Tableau, Matplotlib and Seaborn for creating Visualization

PYSPARK DATAFRAME



```
+-------+----+------+-----------+-----------+----------+-----------------+------------+------------+----------------+------------------+---------------+
|country|year|   sex|        age|suicides_no|population|suicides/100k pop|country-year|HDI for year|gdp_for_year ($)|gdp_per_capita ($)|     generation|
+-------+----+------+-----------+-----------+----------+-----------------+------------+------------+----------------+------------------+---------------+
|Albania|1987|  male|15-24 years|         21|    312900|             6.71| Albania1987|        null|   2,156,624,900|               796|   Generation X|
|Albania|1987|  male|35-54 years|         16|    308000|             5.19| Albania1987|        null|   2,156,624,900|               796|         Silent|
|Albania|1987|female|15-24 years|         14|    289700|             4.83| Albania1987|        null|   2,156,624,900|               796|   Generation X|
|Albania|1987|  male|  75+ years|          1|     21800|             4.59| Albania1987|        null|   2,156,624,900|               796|G.I. Generation|
|Albania|1987|  male|25-34 years|          9|    274300|             3.28| Albania1987|        null|   2,156,624,900|               796|        Boomers|
|Albania|1987|female|  75+ years|          1|     35600|             2.81| Albania1987|        null|   2,156,624,900|               796|G.I. Generation|
|Albania|1987|female|35-54 years|          6|    278800|             2.15| Albania1987|        null|   2,156,624,900|               796|         Silent|
|Albania|1987|female|25-34 years|          4|    257200|             1.56| Albania1987|        null|   2,156,624,900|               796|        Boomers|
|Albania|1987|  male|55-74 years|          1|    137500|             0.73| Albania1987|        null|   2,156,624,900|               796|G.I. Generation|
|Albania|1987|female| 5-14 years|          0|    311000|              0.0| Albania1987|        null|   2,156,624,900|               796|   Generation X|
|Albania|1987|female|55-74 years|          0|    144600|              0.0| Albania1987|        null|   2,156,624,900|               796|G.I. Generation|
|Albania|1987|  male| 5-14 years|          0|    338200|              0.0| Albania1987|        null|   2,156,624,900|               796|   Generation X|
|Albania|1988|female|  75+ years|          2|     36400|             5.49| Albania1988|        null|   2,126,000,000|               769|G.I. Generation|
|Albania|1988|  male|15-24 years|         17|    319200|             5.33| Albania1988|        null|   2,126,000,000|               769|   Generation X|
|Albania|1988|  male|  75+ years|          1|     22300|             4.48| Albania1988|        null|   2,126,000,000|               769|G.I. Generation|
|Albania|1988|  male|35-54 years|         14|    314100|             4.46| Albania1988|        null|   2,126,000,000|               769|         Silent|
|Albania|1988|  male|55-74 years|          4|    140200|             2.85| Albania1988|        null|   2,126,000,000|               769|G.I. Generation|
|Albania|1988|female|15-24 years|          8|    295600|             2.71| Albania1988|        null|   2,126,000,000|               769|   Generation X|
|Albania|1988|female|55-74 years|          3|    147500|             2.03| Albania1988|        null|   2,126,000,000|               769|G.I. Generation|
|Albania|1988|female|25-34 years|          5|    262400|             1.91| Albania1988|        null|   2,126,000,000|               769|        Boomers|
|Albania|1988|  male|25-34 years|          5|    279900|             1.79| Albania1988|        null|   2,126,000,000|               769|        Boomers|
|Albania|1988|female|35-54 years|          4|    284500|             1.41| Albania1988|        null|   2,126,000,000|               769|         Silent|
|Albania|1988|female| 5-14 years|          0|    317200|              0.0| Albania1988|        null|   2,126,000,000|               769|   Generation X|
|Albania|1988|  male| 5-14 years|          0|    345000|              0.0| Albania1988|        null|   2,126,000,000|               769|   Generation X|
|Albania|1989|  male|  75+ years|          2|     22500|             8.89| Albania1989|        null|   2,335,124,988|               833|G.I. Generation|
|Albania|1989|  male|25-34 years|         18|    283600|             6.35| Albania1989|        null|   2,335,124,988|               833|        Boomers|
|Albania|1989|  male|35-54 years|         15|    318400|             4.71| Albania1989|        null|   2,335,124,988|               833|         Silent|
|Albania|1989|  male|55-74 years|          6|    142100|             4.22| Albania1989|        null|   2,335,124,988|               833|G.I. Generation|
|Albania|1989|  male|15-24 years|         12|    323500|             3.71| Albania1989|        null|   2,335,124,988|               833|   Generation X|
|Albania|1989|female|35-54 years|          7|    288600|             2.43| Albania1989|        null|   2,335,124,988|               833|         Silent|
+-------+----+------+-----------+-----------+----------+-----------------+------------+------------+----------------+------------------+---------------+
only showing top 30 rows
```

COUNT OF CATEGORY COLUMNS



```
data_clean.groupBy("age").count().show()
```

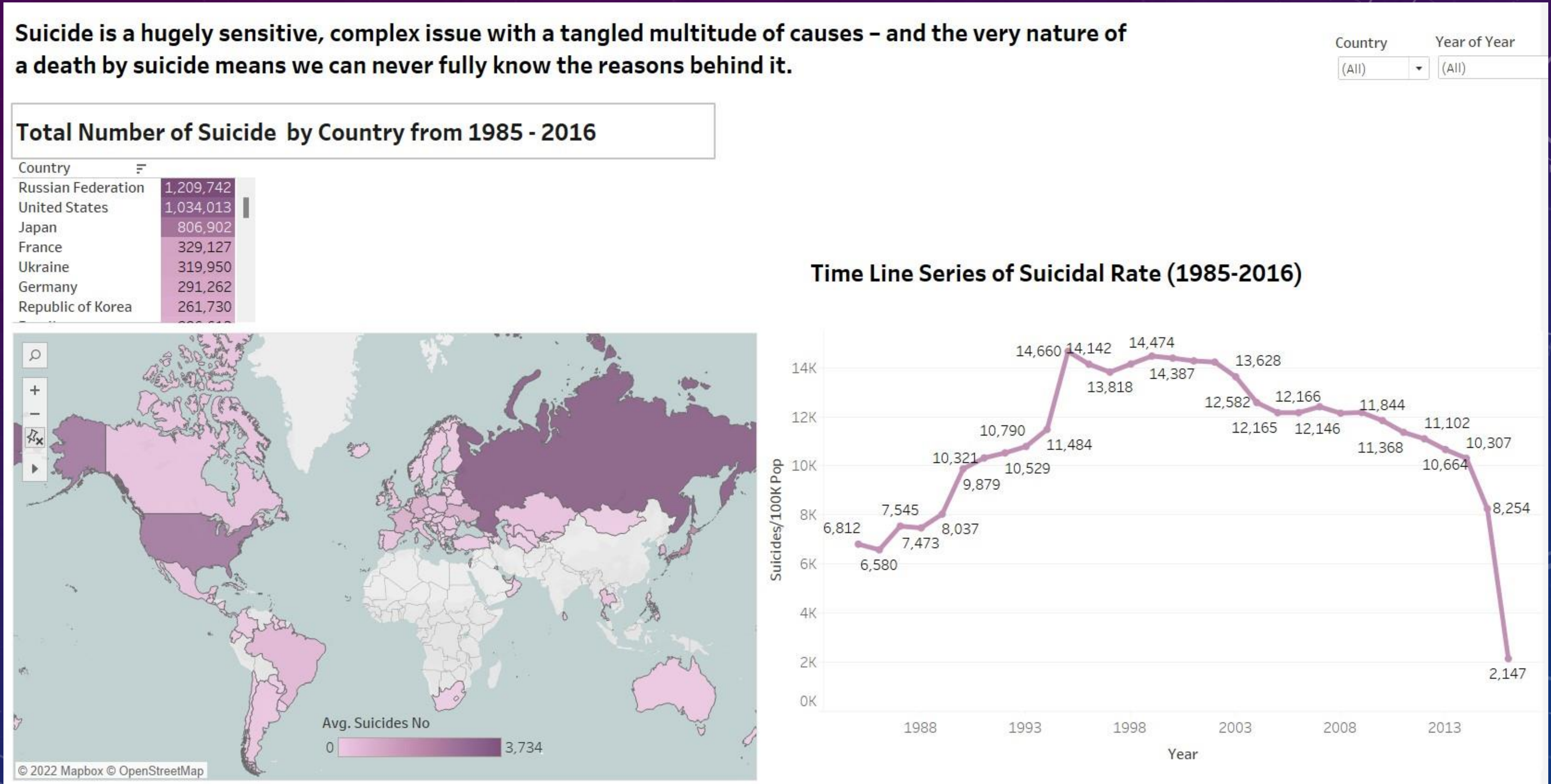
```
+-----------+-----+
|        age|count|
+-----------+-----+
|55-74 years| 4642|
|25-34 years| 4642|
| 5-14 years| 4610|
|  75+ years| 4642|
|15-24 years| 4642|
|35-54 years| 4642|
+-----------+-----+
```

```
[20] data_clean.groupBy("sex").count().show()
```

```
+------+-----+
|   sex|count|
+------+-----+
|female|13910|
|  male|13910|
+------+-----+
```

RESEARCH QUESTIONS

- Research Question1: How did the Rate of Suicide changed over time?
- Research Question 2: What is the relationship between the gender and the number of suicides?
- Research Question 3: Countries with highest rate of Suicides
- Research Question 4: How different generation have affected Suicides?
- Research Question 5: Are certain age groups more inclined to suicide?



Suicide is a hugely sensitive, complex issue with a tangled multitude of causes – and the very nature of a death by suicide means we can never fully know the reasons behind it.

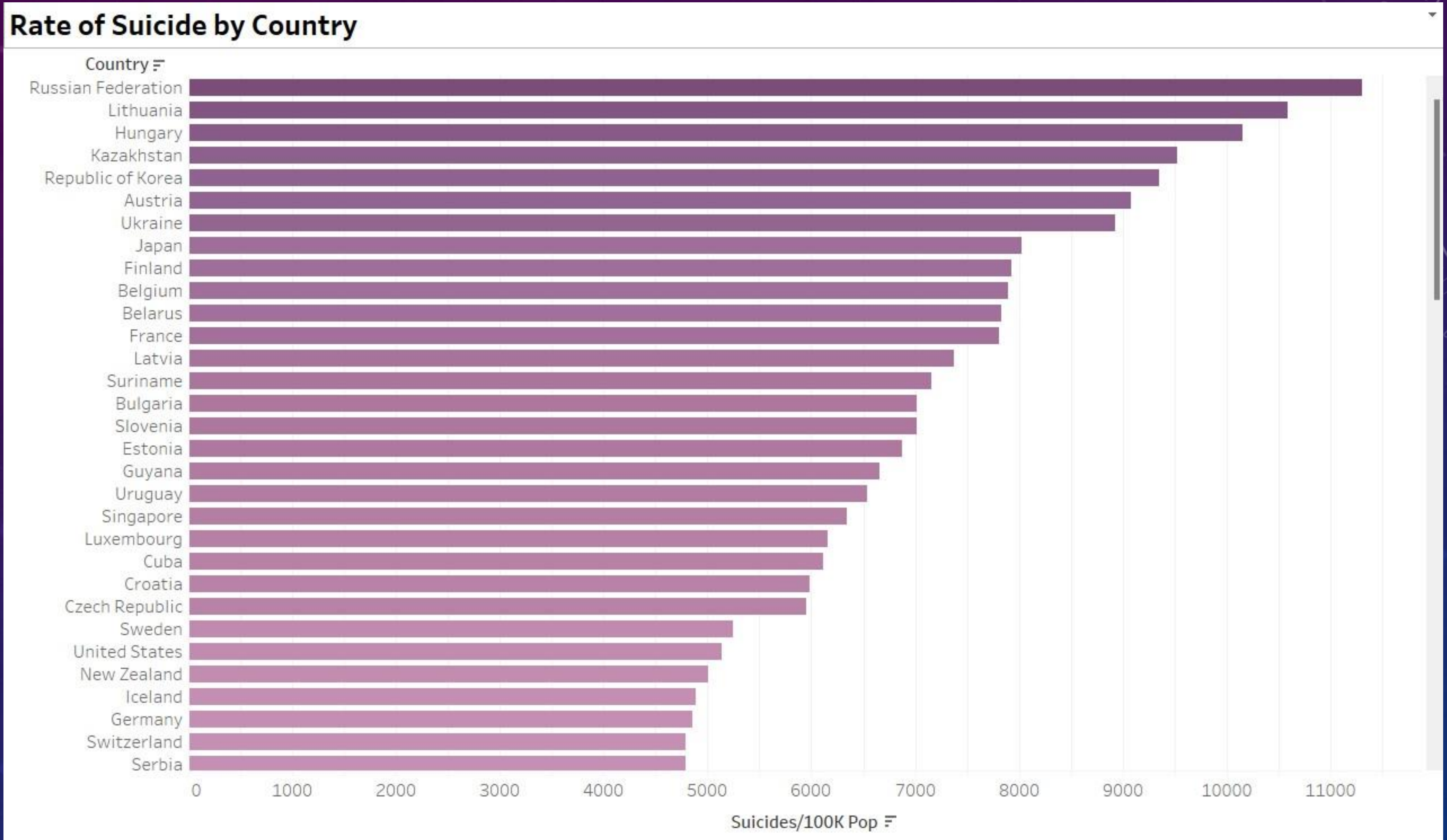
Country (All) ▾

Year of Year (All)

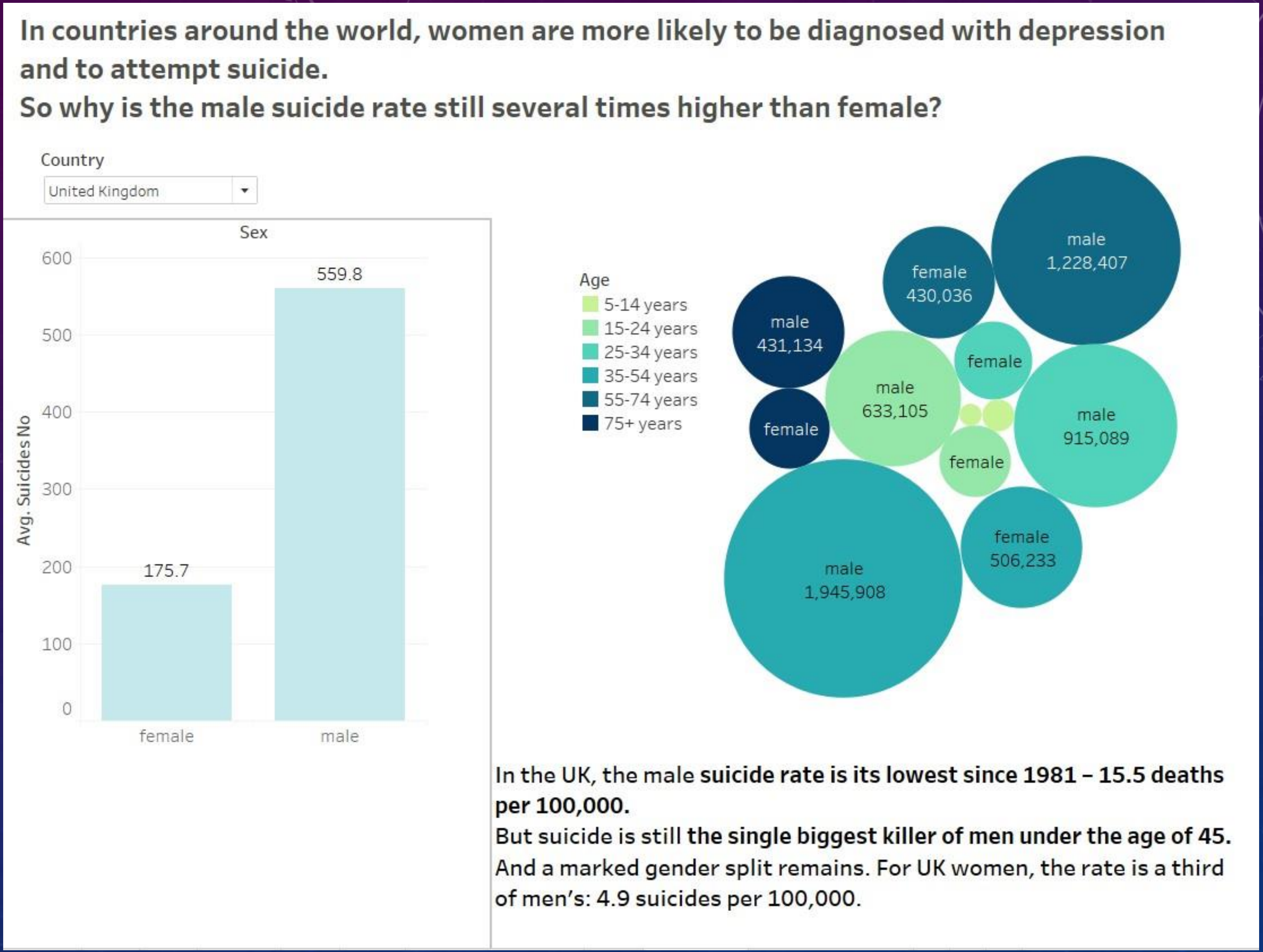
Total Number of Suicide by Country from 1985 - 2016

Country	
Russian Federation	1,209,742
United States	1,034,013
Japan	806,902
France	329,127
Ukraine	319,950
Germany	291,262
Republic of Korea	261,730

Time Line Series of Suicidal Rate (1985-2016)



Rate of Suicide by Country



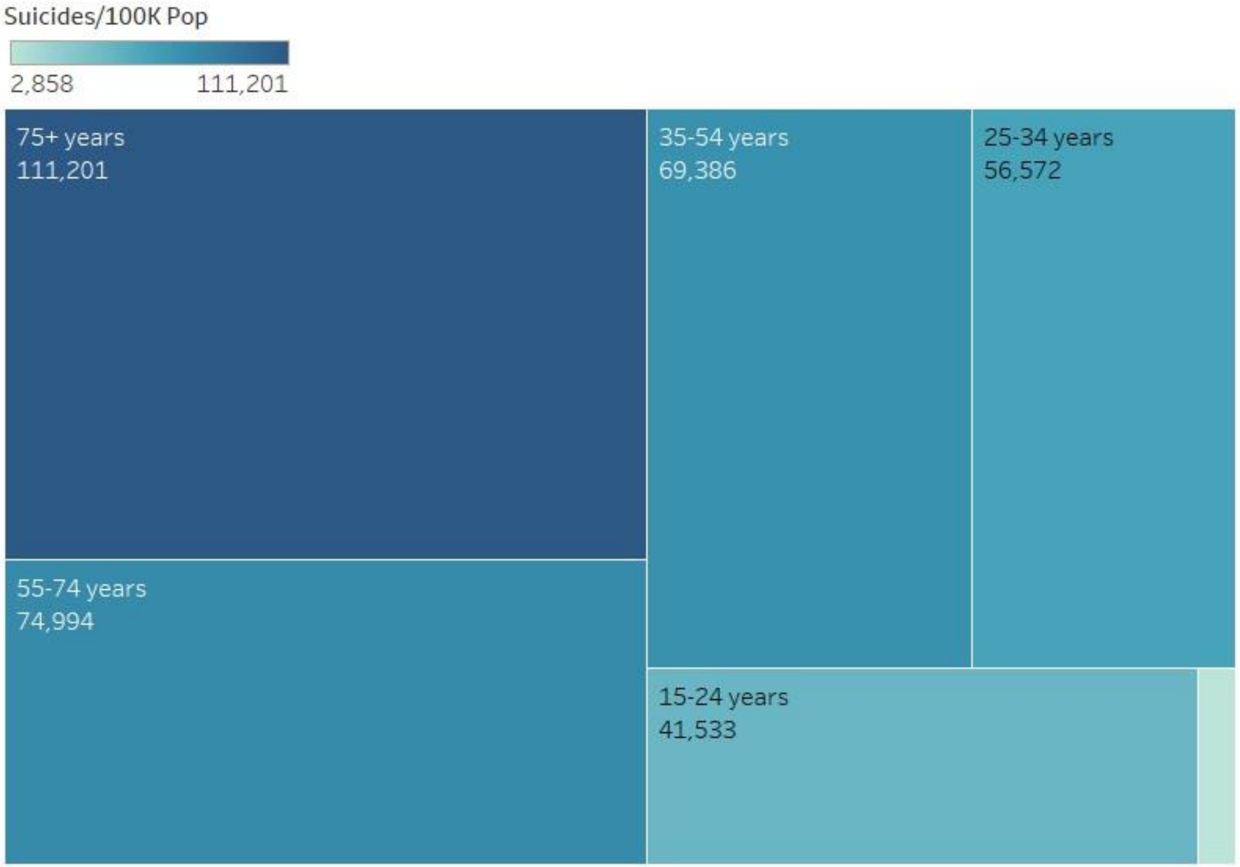
In countries around the world, women are more likely to be diagnosed with depression and to attempt suicide.
So why is the male suicide rate still several times higher than female?

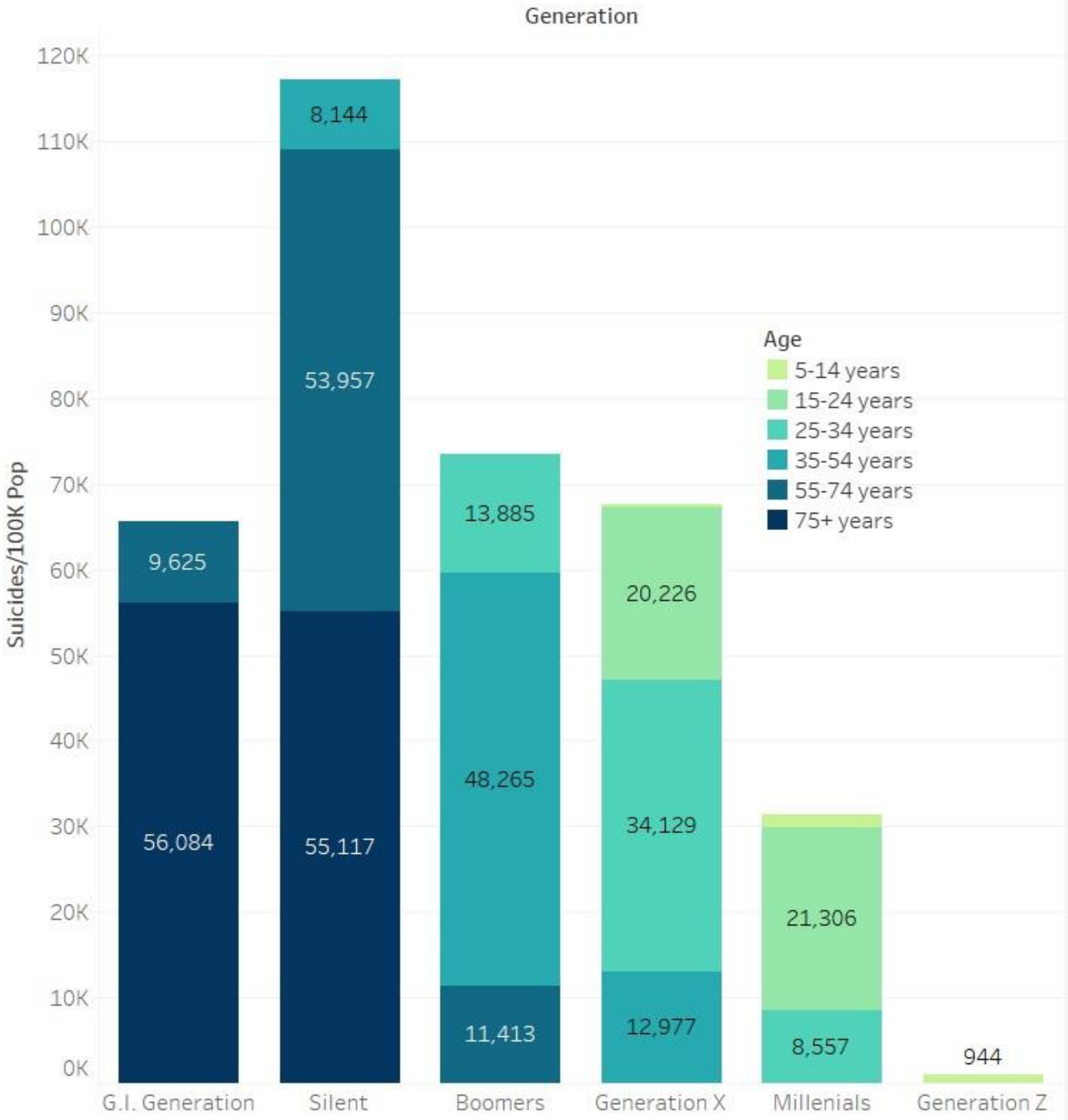
Country: United Kingdom

In the UK, the male **suicide rate is its lowest since 1981 – 15.5 deaths per 100,000.**
But suicide is still **the single biggest killer of men under the age of 45.**
And a marked gender split remains. For UK women, the rate is a third of men’s: 4.9 suicides per 100,000.

As person gets older it tends to be more suicidal.
Some of the causes could be:

1. Breaking medical regimens
2. Loss of interest in things or activities that are usually found enjoyable
3. Experiencing or expecting a significant personal loss (spouse or other)
4. Stock-piling medication or obtaining other lethal means

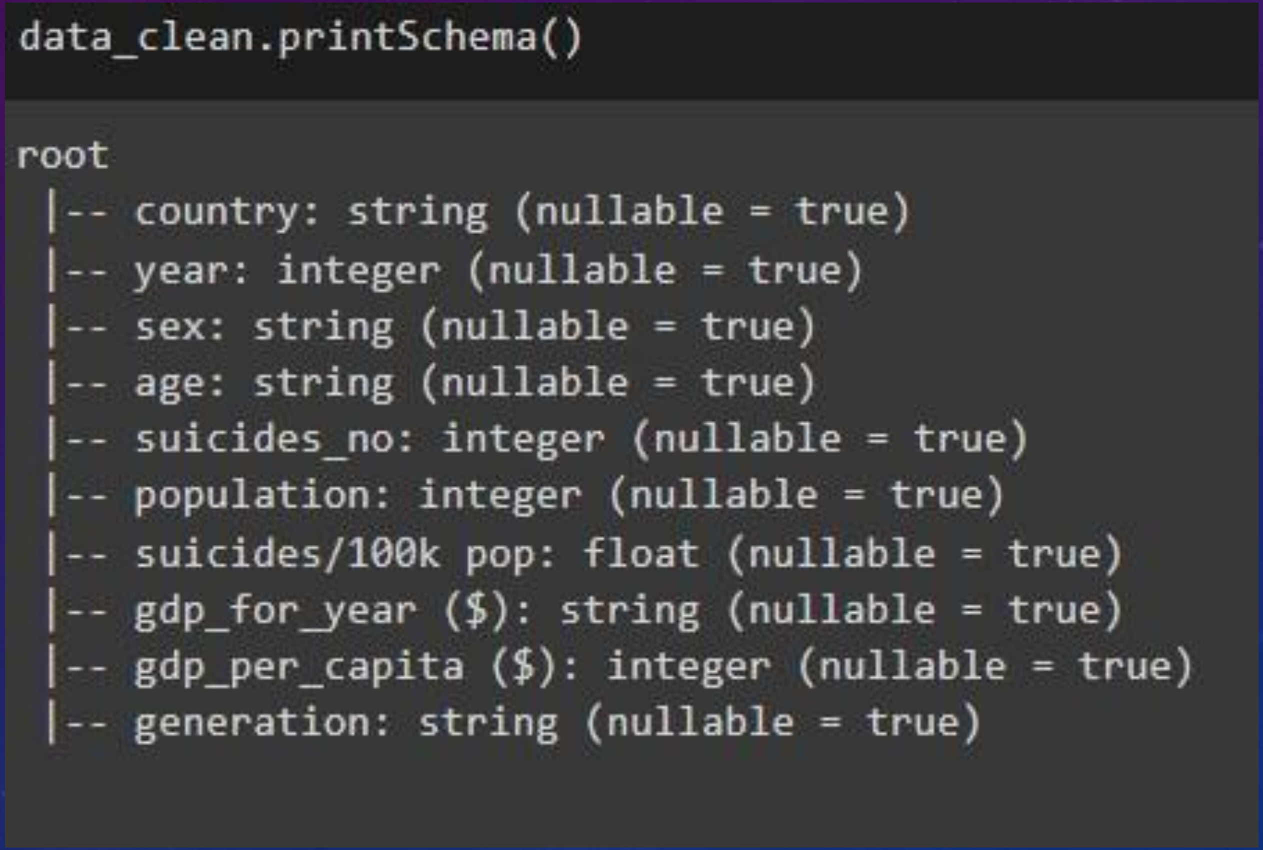




INSIGHTS

- There was a decrease in suicide towards the 80's. This could be due to the awareness of suicides and mental health in 80s as well as improved recognition if those at risk.
- Russian levels of alcohol consumption plays an immense role in it's large suicide count, but their is a lack of data to support this due to Soviet secrecy.
- Data show alarming differences in suicide for different sexes. It's evident that males are more inclined to suicide, than females.
- the G.I. Generation or the Greatest Generation (the generation who lived during the WWII) has the highest suicide rate with almost 25 suicides per 100,000 person. This is a very big number compared to younger generations, this might be due to the fact that this generation suffered a lot during the WWII, many of them lost their loved ones and experienced different traumatic events. The suicide rates decrease from a generation to another, where Generation Z has the lowest suicide rates with 1 suicide per 100,000 person.
- We can see that as the person gets older it tends to be more suicidal. This could be explained by the fact that important life changes that happen as we get older may cause feelings of uneasiness, stress, and sadness. But this might be due to the fact that old people (75+ years) belong to the G.I. Generation which already has the highest suicide rates. To further explore this, we must check the number of people that committed suicide within each age category with respect to their generation. This way we can find out the distribution of ages of suicidal people within each generation. This will help us to identify if suicide is due to the age factor or to the generation.

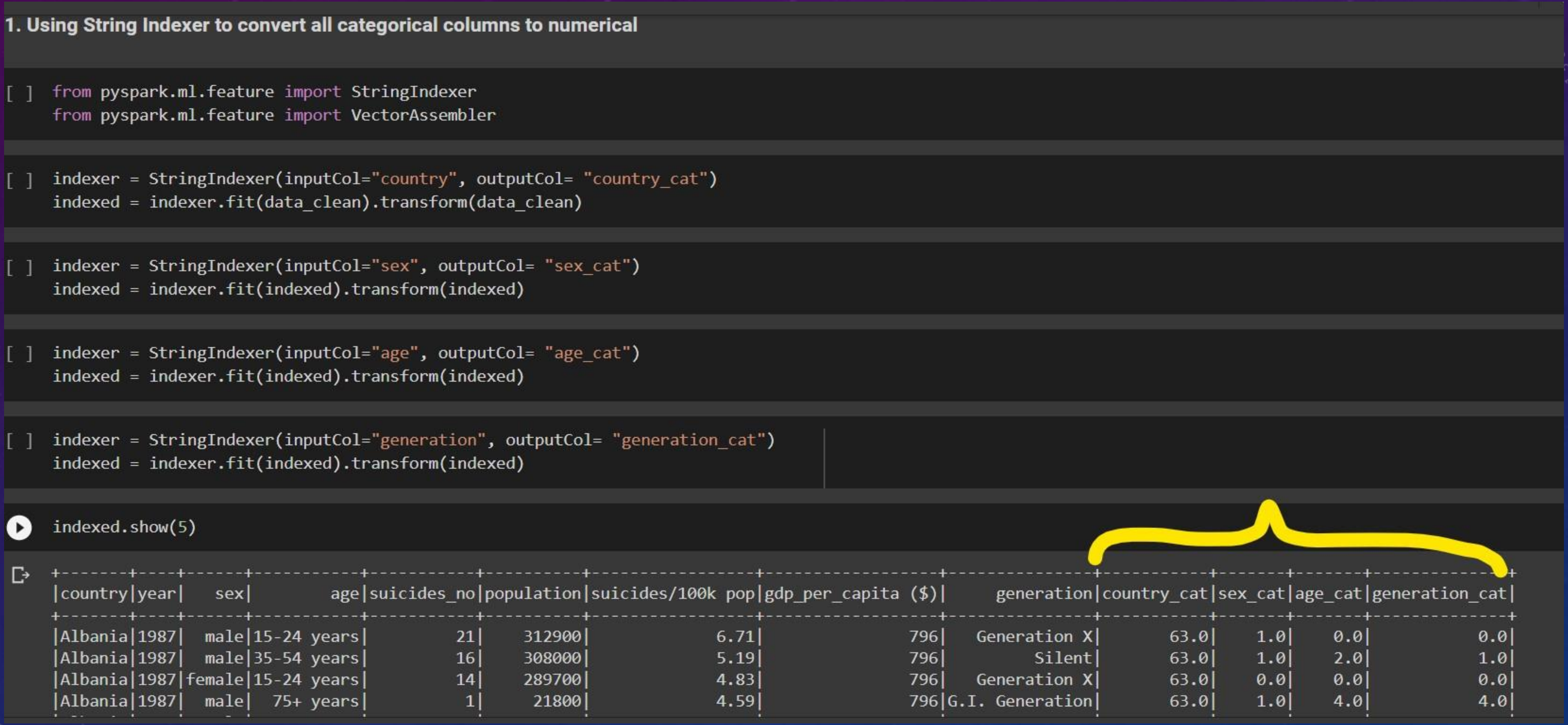
MODEL BUILDING USING PYSPARK MACHINE LEARNING LIBRARIES



```
data_clean.printSchema()
```

```
root
 |-- country: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- sex: string (nullable = true)
 |-- age: string (nullable = true)
 |-- suicides_no: integer (nullable = true)
 |-- population: integer (nullable = true)
 |-- suicides/100k pop: float (nullable = true)
 |-- gdp_for_year ($): string (nullable = true)
 |-- gdp_per_capita ($): integer (nullable = true)
 |-- generation: string (nullable = true)
```

STRING INDEXER



1. Using String Indexer to convert all categorical columns to numerical

```
[ ] from pyspark.ml.feature import StringIndexer
    from pyspark.ml.feature import VectorAssembler
```

```
[ ] indexer = StringIndexer(inputCol="country", outputCol= "country_cat")
    indexed = indexer.fit(data_clean).transform(data_clean)
```

```
[ ] indexer = StringIndexer(inputCol="sex", outputCol= "sex_cat")
    indexed = indexer.fit(indexed).transform(indexed)
```

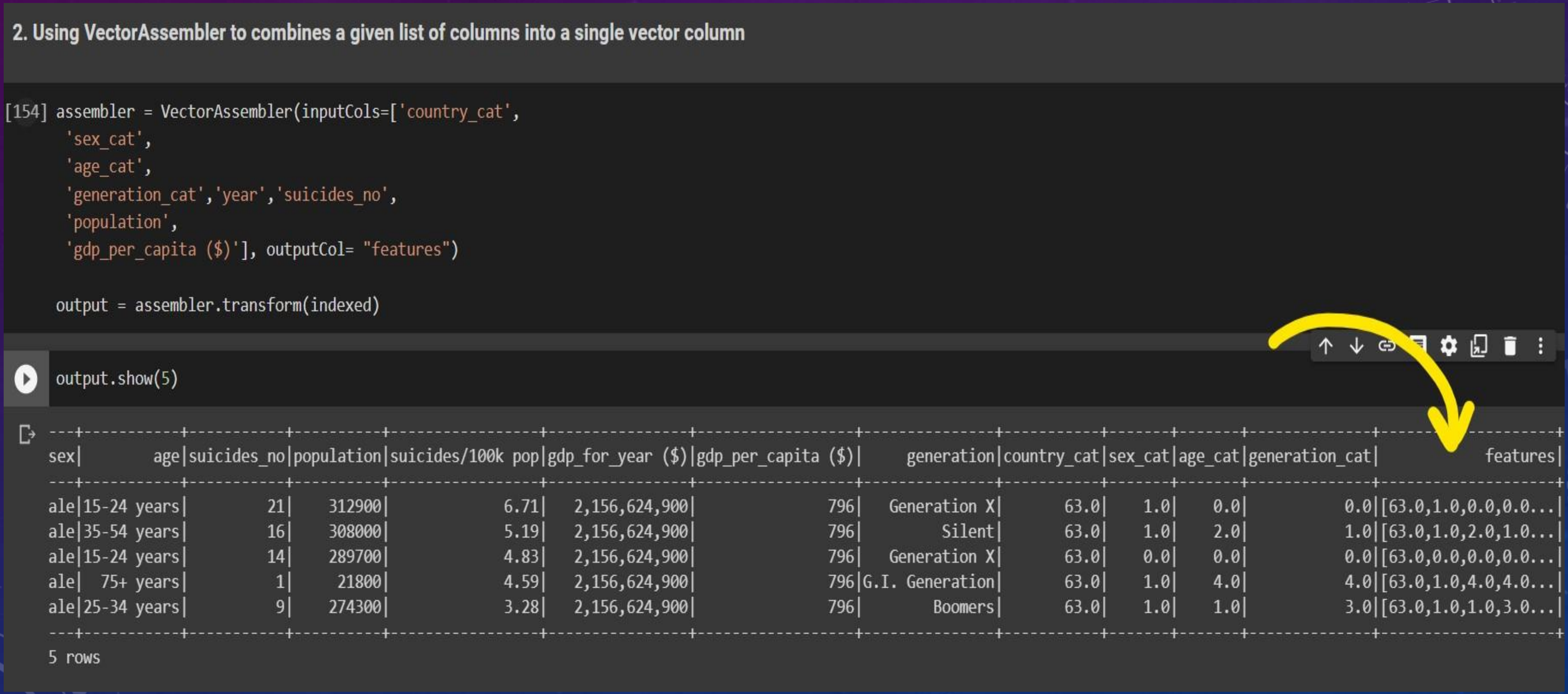
```
[ ] indexer = StringIndexer(inputCol="age", outputCol= "age_cat")
    indexed = indexer.fit(indexed).transform(indexed)
```

```
[ ] indexer = StringIndexer(inputCol="generation", outputCol= "generation_cat")
    indexed = indexer.fit(indexed).transform(indexed)
```

```
indexed.show(5)
```

```
+-------+----+------+-----------+-----------+----------+-----------------+------------------+---------------+-----------+-------+-------+--------------+
|country|year|   sex|        age|suicides_no|population|suicides/100k pop|gdp_per_capita ($)|     generation|country_cat|sex_cat|age_cat|generation_cat|
+-------+----+------+-----------+-----------+----------+-----------------+------------------+---------------+-----------+-------+-------+--------------+
|Albania|1987|  male|15-24 years|         21|    312900|             6.71|               796|   Generation X|       63.0|    1.0|    0.0|           0.0|
|Albania|1987|  male|35-54 years|         16|    308000|             5.19|               796|         Silent|       63.0|    1.0|    2.0|           1.0|
|Albania|1987|female|15-24 years|         14|    289700|             4.83|               796|   Generation X|       63.0|    0.0|    0.0|           0.0|
|Albania|1987|  male|  75+ years|          1|     21800|             4.59|               796|G.I. Generation|       63.0|    1.0|    4.0|           4.0|
```

VECTOR ASSEMBLER



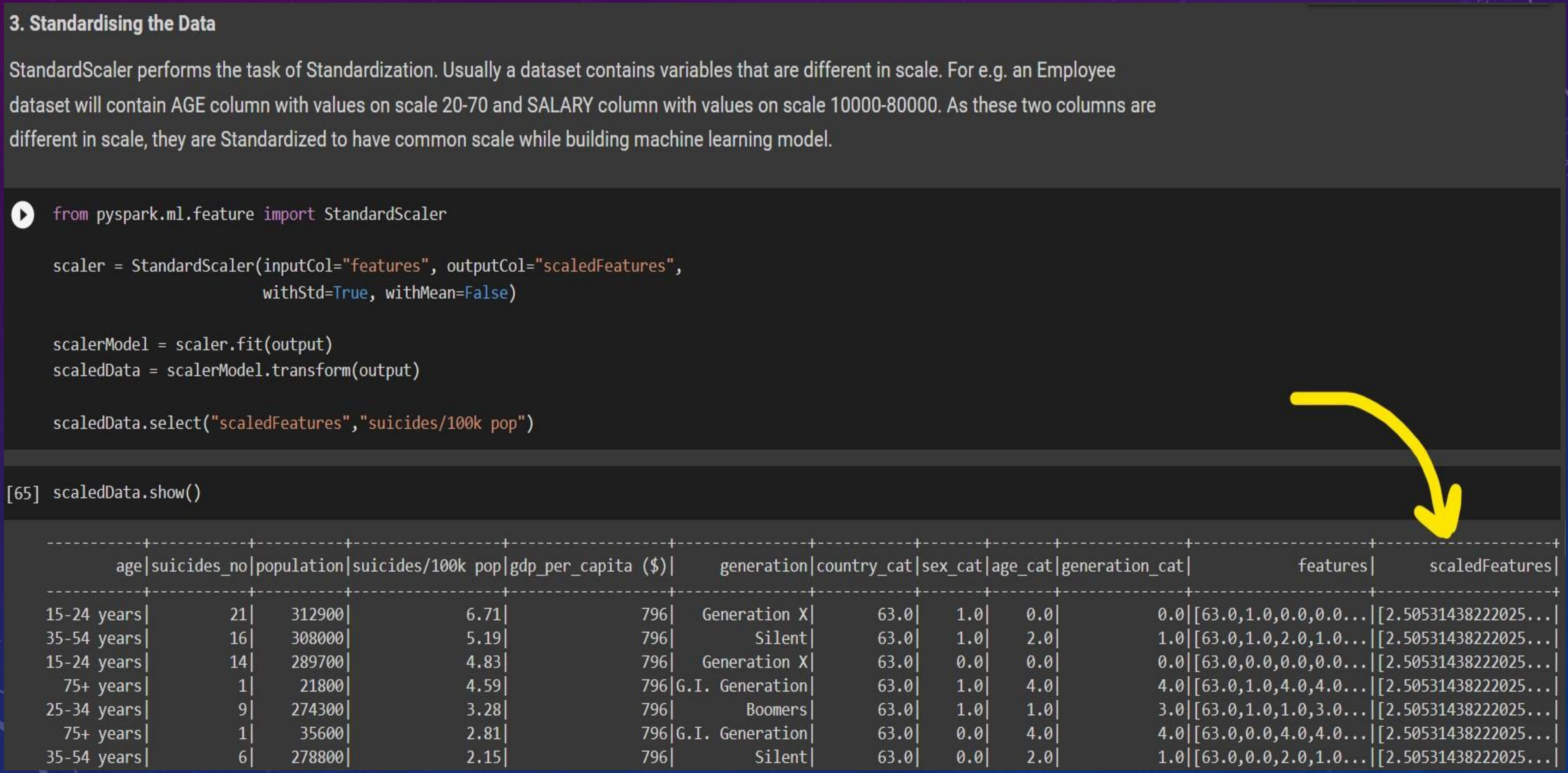
2. Using VectorAssembler to combines a given list of columns into a single vector column

```
[154] assembler = VectorAssembler(inputCols=['country_cat',
        'sex_cat',
        'age_cat',
        'generation_cat','year','suicides_no',
        'population',
        'gdp_per_capita ($)'], outputCol= "features")

      output = assembler.transform(indexed)
```

```
output.show(5)
```

```
---+-----------+-----------+----------+-----------------+----------------+------------------+---------------+-----------+-------+-------+--------------+--------------------+
sex|        age|suicides_no|population|suicides/100k pop|gdp_for_year ($)|gdp_per_capita ($)|     generation|country_cat|sex_cat|age_cat|generation_cat|            features|
---+-----------+-----------+----------+-----------------+----------------+------------------+---------------+-----------+-------+-------+--------------+--------------------+
ale|15-24 years|         21|    312900|             6.71|   2,156,624,900|               796|   Generation X|       63.0|    1.0|    0.0|           0.0|[63.0,1.0,0.0,0.0...|
ale|35-54 years|         16|    308000|             5.19|   2,156,624,900|               796|         Silent|       63.0|    1.0|    2.0|           1.0|[63.0,1.0,2.0,1.0...|
ale|15-24 years|         14|    289700|             4.83|   2,156,624,900|               796|   Generation X|       63.0|    0.0|    0.0|           0.0|[63.0,0.0,0.0,0.0...|
ale|  75+ years|          1|     21800|             4.59|   2,156,624,900|               796|G.I. Generation|       63.0|    1.0|    4.0|           4.0|[63.0,1.0,4.0,4.0...|
ale|25-34 years|          9|    274300|             3.28|   2,156,624,900|               796|        Boomers|       63.0|    1.0|    1.0|           3.0|[63.0,1.0,1.0,3.0...|
---+-----------+-----------+----------+-----------------+----------------+------------------+---------------+-----------+-------+-------+--------------+--------------------+
5 rows
```

STANDARDIZING THE DATA



3. Standardising the Data

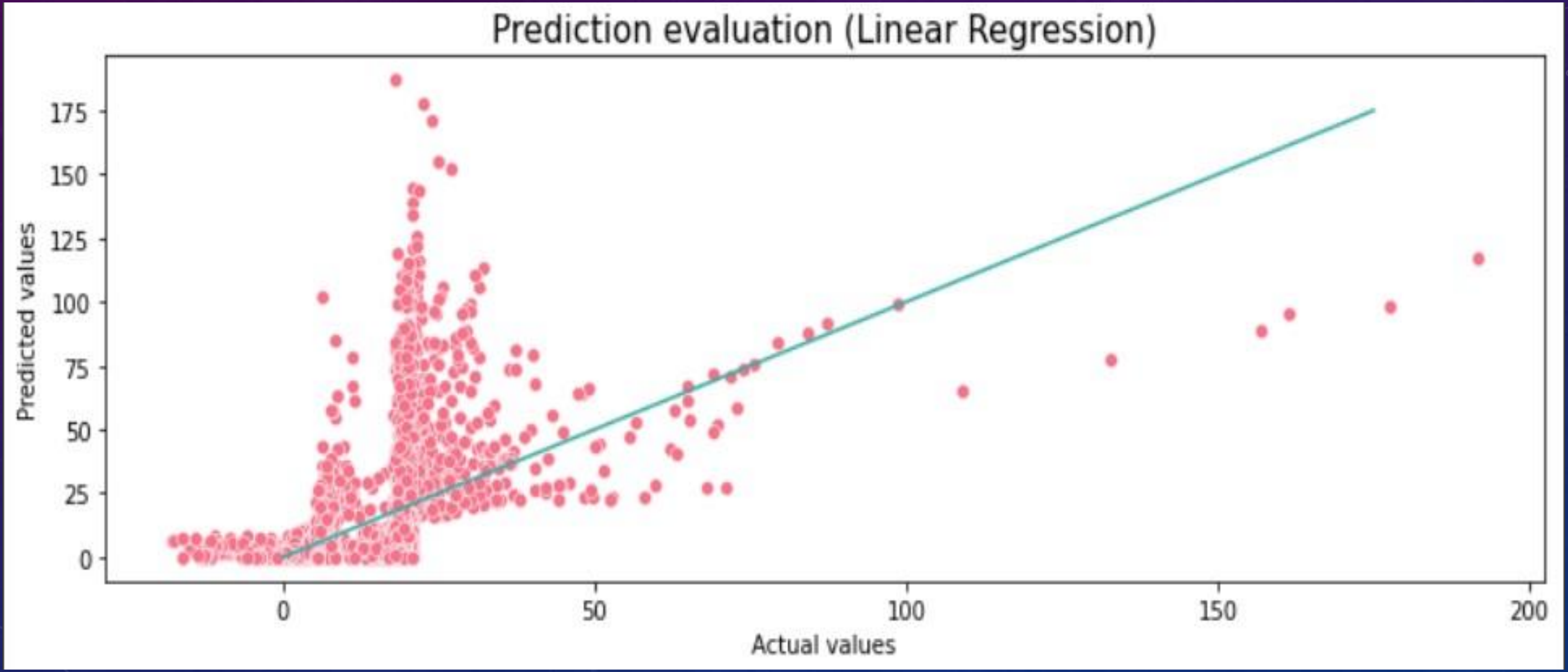
StandardScaler performs the task of Standardization. Usually a dataset contains variables that are different in scale. For e.g. an Employee dataset will contain AGE column with values on scale 20-70 and SALARY column with values on scale 10000-80000. As these two columns are different in scale, they are Standardized to have common scale while building machine learning model.

```
from pyspark.ml.feature import StandardScaler

scaler = StandardScaler(inputCol="features", outputCol="scaledFeatures",
                        withStd=True, withMean=False)

scalerModel = scaler.fit(output)
scaledData = scalerModel.transform(output)

scaledData.select("scaledFeatures","suicides/100k pop")
```

```
[65] scaledData.show()
```

```
-----------+-----------+----------+-----------------+------------------+---------------+-----------+-------+-------+--------------+--------------------+--------------------+
        age|suicides_no|population|suicides/100k pop|gdp_per_capita ($)|     generation|country_cat|sex_cat|age_cat|generation_cat|            features|      scaledFeatures|
-----------+-----------+----------+-----------------+------------------+---------------+-----------+-------+-------+--------------+--------------------+--------------------+
15-24 years|         21|    312900|             6.71|               796|   Generation X|       63.0|    1.0|    0.0|           0.0|[63.0,1.0,0.0,0.0...|[2.50531438222025...|
35-54 years|         16|    308000|             5.19|               796|         Silent|       63.0|    1.0|    2.0|           1.0|[63.0,1.0,2.0,1.0...|[2.50531438222025...|
15-24 years|         14|    289700|             4.83|               796|   Generation X|       63.0|    0.0|    0.0|           0.0|[63.0,0.0,0.0,0.0...|[2.50531438222025...|
  75+ years|          1|     21800|             4.59|               796|G.I. Generation|       63.0|    1.0|    4.0|           4.0|[63.0,1.0,4.0,4.0...|[2.50531438222025...|
25-34 years|          9|    274300|             3.28|               796|        Boomers|       63.0|    1.0|    1.0|           3.0|[63.0,1.0,1.0,3.0...|[2.50531438222025...|
  75+ years|          1|     35600|             2.81|               796|G.I. Generation|       63.0|    0.0|    4.0|           4.0|[63.0,0.0,4.0,4.0...|[2.50531438222025...|
35-54 years|          6|    278800|             2.15|               796|         Silent|       63.0|    0.0|    2.0|           1.0|[63.0,0.0,2.0,1.0...|[2.50531438222025...|
```

LINEAR REGRESSION MODEL – COMPARING ACTUAL AND PREDICTED VALUES



CONCLUSION

- Conclusion Our model doesn't seem to be doing a good job, this might be due to fact that the features we selected aren't good enough, or it might be due the fact that the data we have isn't linear so a similar model won't be any good to estimate the values.
- Future Scope: Maybe the decision tree will perform better.