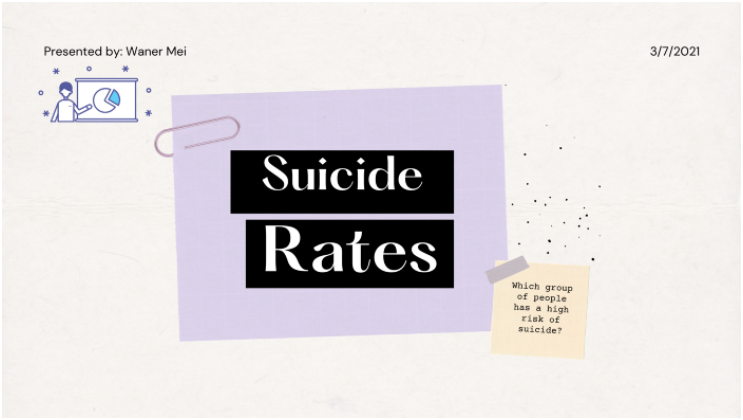
Exploratory Data Analysis (EDA) in Python on Suicide Rates

Waner Mei Just now · 7 min read



Created by Waner Mei on Canvas

Introduction

Suicide is a health issue that has to be addressed and analyzed in the United States. Although the U.S. is wealthy and prosperous, its suicide rate is still about 10 to 11 per 100,000 population according to World Health Organization. Base on my experience, I met many friends and family members that suffer from depressions and other mental health issues. In addition to that, the pandemic during the past year causes much stress on people. Therefore, for my very first EDA, I decided to research suicide rates and trying to find out which group and portion of the population has a higher risk when facing suicidal thoughts.

Part 1: Cleaning the dataset

```
[2] url = "https://raw.githubusercontent.com/wanermelon/EDA-on-Suicide-Rates/main/master.csv"
    Suicide_rate = pd.read_csv(url)
    # Dataset is now stored in a Pandas Dataframe and defined as Suicide_rate
```

importing dataset

I found a perfect and simple dataset for me to use on Kaggle. It is an overview of suicide rates from 1985 to 2016. Although there are only 12 columns, it is still necessary to clean up the dataset. Here is a quick demonstration of the first 5 rows of this dataset.

```
# Top 5 obs
Suicide_rate.head()
```

	country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation

4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

Showing the first 5 rows of the dataset

After importing the dataset successfully, I have to clean up this dataset and make it even more organized and straightforward. To do that, I first check the property of the dataset, and then I see that some names for the column could potentially be an issue due to the syntax. Therefore, I changed two column names. Then, I dropped the columns that I will not use in this EDA. After that, I dropped the duplicate rows and rows with missing values.

Cleaning up data set (left 1), before and after cleaning(right 2 and right 1)

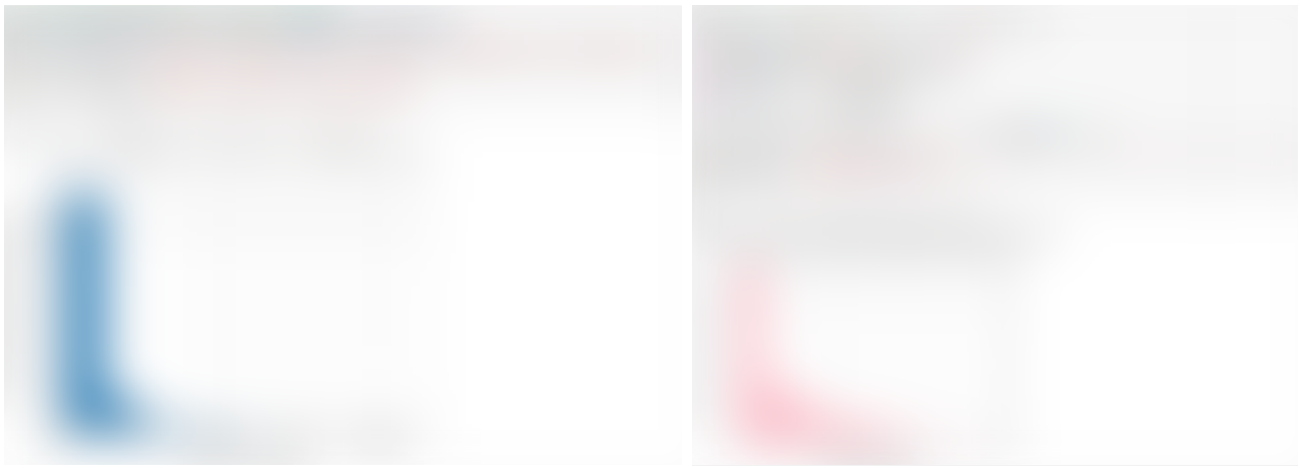
Next, I used the “describe” method to find the count, unique value, top value, and frequency, mean, standard deviation, minimum, 25% percentiles, 50% percentiles, 75% percentiles, and maximum if available for the columns. NaN indicates that this statistical analysis cannot be operated on that column.

“describe” method

However, the “describe” method does not show us which value is an outlier, therefore, I found some functions that can indicate the outliers. After that, I created a new feature that categorizes the values base on the Suicide_per_Hundred_K_Pop statistical analysis into “outlier” and “normal”.

Outlier functions I found and the new column “outlier” I created

Here is a comparison of the histogram including outliers and excluding outliers. There isn’t much difference but the pink graph is more detailed and we can see that most of the countries and states have a suicide rate below 5 and there are still about 4,200 countries and states that have a suicide rate between 5 and 10. Whereas in the blue graph we can only roughly assume that most of the countries and states have a suicide rate below 50. But we can’t tell how those data are distributed.



With outliers (blue), without outliers (pink)

Part 2: Analyzing the dataset

After testing some graphs and diagrams, I decided to do a set of boxplots and distribution for GDP per capita and another set for the amount of suicide per year. In the set for GDP per capita, we can see that people that earn below 25,000 USD have a high risk of suicide. In the set for the amount of suicide per year, we can see from the distribution that there is an increase in suicide per year. There were more suicides after the 2000s. Within the distribution, there is a bar graph showing each year’s suicides. For each bar group, there are thinner bars that are indicating the age groups that suicided that year. So far we can’t tell if age plays a big role here, but from 2000 to 2010, different age groups have a similar amount of suicide.

boxplot and normal distribution

Before we get to some more-detailed analysis, I wanted to check if all of the countries and states have the same amount of values. The graph below showed that some countries and states are missing data for certain years and certain ages. Therefore, there would be some errors due to the incompleted dataset.

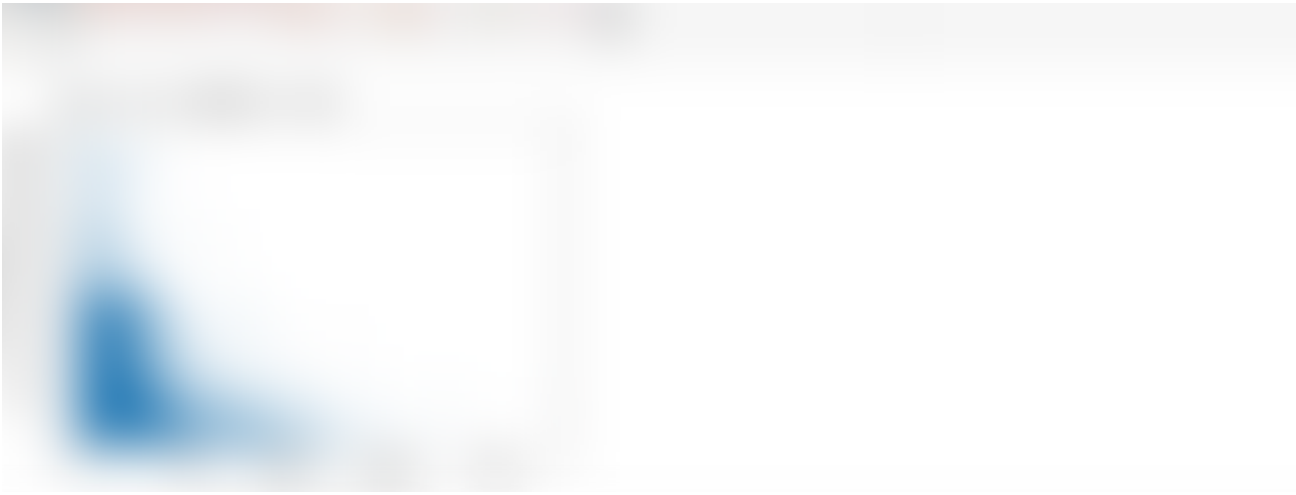
Number of unique values per countries and states

Next, I made a heat map that demonstrates the correlations between columns that have numeral values. Blue means that there is a positive

correlation whereas red means that there is a negative correlation. For example, Suicide number and population have a strong positive correlation. This heat map helped me determine which columns I should analyze. Because the dark blue and dark red colors indicate that the two variables are possibly related.

Heat map that shows the correlation between different columns

For instance, the GDP per capita and the Suicide per 100K have a strong negative correlation. This means that when the suicide rate increases, the GDP per capita decreases. It was proved by the scatter plot below. We can see that there most of the values are placed on the spot where GDP is high and the suicide rate is low or the opposite.

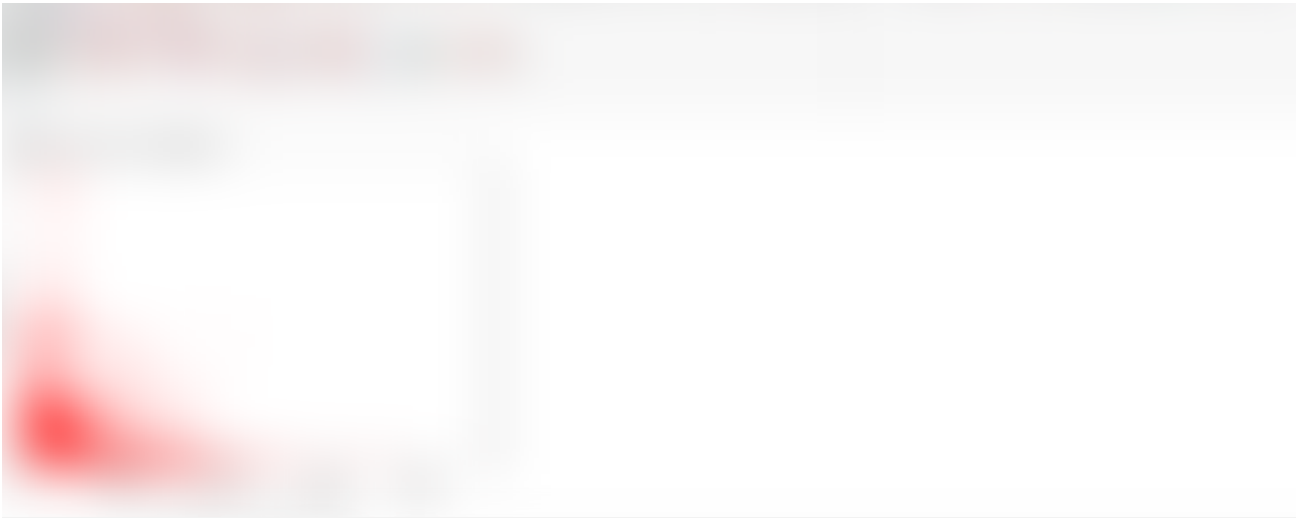


Suicide Rate vs. GDP per capita

Besides using raw values, I also used the p-value means for the suicide rate and try to double-check the correlation. The p-value of the graph is about 3.02, which means that it is statistically significant. Therefore there is a strong negative correlation between the Suicide rate and GDP.

P-value means for suicide rate vs. GDP per capita

Next, I decided to see the correlation between the suicide rate and the population. It seems like that there is a weak negative correlation between the two columns. However, this was not what I expected. My hypothesis was that increase in population would lead to an increase in the suicide rate. Therefore I decided to analyze the relationship between these two factors. In the scatter plot below, there is a negative correlation indicated. But, there are also values that scattered against the correlation. So, we can say that the suicide rate and population do not affect each other as much.



There is a typo on the comment, it should be Suicide rate and population.

After checking the correlations, I think that for the rest of the columns, I can trust the analysis that the heat map provided. Thus, for the following analysis, I will compare the categorical value with the suicide rate.

Since I mentioned earlier that some countries and states are missing some data. Therefore, I decided to compare the average suicide rate over the years for each country. The graph below showed that the more developed countries are more likely to have a higher suicide rate. From the graph, some top results including Australia, South Korea, and Spain.

Mean of suicide rate per country

After that, I wanted to look at how gender influences the suicide rate. Which surprisingly, turns out that males have a higher risk of suicide than females.

mean of the suicide rate for each sex

Next, I graphed the mean suicide rate for each generation. I decided to leave out Generation Z and the Millenials since this dataset mainly focused on generations that lived before 2010. So, there might not be enough data for Millenials and Generation Z. Focusing on the earlier generations, we see that G.I Generation has the highest suicide rate. G.I Generation is people that were born between 1900 and 1927. After that is the Silent Generation and the Boomers and then the Generation X. Therefore, we see that people with older age have a higher risk of suicide.

mean value of suicide rate for each generation

The following graphs that shows the mean suicide rate for each age group also supported that statement.

mean value of suicide rate for each age groups

Treemap that shows the distribution of suicide rate of different generations

Conclusion

In conclusion, the following groups have a greater risk of suicide in the past decades: the low-income population, people who live in developed countires, males, and older generations.

Thank you for reading my blog!

Data Science Eda Suicide Health Society

More from Waner Mei

Follow

Student studying in data science

More From Medium

Customizing Pandas-Profiling Summaries

Ian Eaves in Towards Data Science

The Death of New York City Is Clickbait

Jada Gomez in LEVEL

Word Embeddings

S. T. Lanier

Apartment Market, Web-Scrapping and EDA using Python

Hurmet Noka

It's Time to Get Rid of the Filibuster.

Jake Wilder

Data Engineering — Google Bigquery to Pretty Email Templates using Apache Airflow

Nicholas Leong

Python for exploratory data analysis and association rules applied to an e-commerce data set

Bruno Argollo

The Fate Of Elon Musk's Assistant Is A Cautionary Tale For Negotiating Salary

Nick Wolny in Entrepreneur's Handbook