|  |  |
| --- | --- |
| A Comparative Analysis of Suicide Rates | |
|  | |
| Runtime Errors | Project Contributions made by Adwait Deshpande, Vansh Kumar Singh, Vasu Verma, Shrirang Joshi, Suraj Kumar Konda |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | Introduction**Suicide is a major issue with multiple countries, figuring out the rise for the rise in suicide rates is a puzzler. A study was conducted on suicide rates based on sex, race, Hispanic origin, and age. our task would be to figure out which group of people have the highest suicide rate. The group could overlap categories i.e., sex, race etc.** Provided data set:   1. <https://tinyurl.com/bdccs4zm>   **Our aim of this project was to analyze the given data and to identify which demographic of people is most affected by this ‘act.’**  **We went beyond the datasets given by the organizers and referred to an external dataset which collectively contained more than 2.6 lac data points.**    National Crime Records Bureau (NCRB), Govt of India has shared this [dataset](https://data.gov.in/dataset-group-name/accidental-deaths-and-suicides) under [Govt. Open Data License - India](https://data.gov.in/government-open-data-license-india). | |  |
|  | The resources we used | |  |



**To facilitate access to data and point toward useful tool to work with the data we fed the data points to Kleagle to obtain datasets which could be integrated with tools and APIs to solve the problem given at hand. By using the platform Kleagle all our datasets and information and resources are compiled into interactive and downloadable metadata**

## Data tools

**Here we explain the data tools and frameworks that we used to create understandable and interactive charts and data tables using the metadata we were given that we generated using Kleagle.**

**Pandas library**

**Pandas is a powerful package that is commonly used for data analysis.**

**It is convenient use this platform as it does not require writing code in the Python language. We used the *class*pandas.DataFrame.**

**which is a 2D data structure which organises data into rows and columns where every row depicts observation, and every column depicts variables**

**Matplotlib**

**At its core Matplotlib Is an object oriented library which directly allows the user to work with the data and convert it into plots and graphs** [**matplotlib.pyplot**](https://matplotlib.org/stable/api/pyplot_summary.html#module-matplotlib.pyplot) **is a collection of functions which lets the user create, edit and develop interactive graphs**

**Seaborn**

**Seaborn as a Python data visualization library built on top of matplotlib the import seaborn portion of the code tells Python to bring the seaborn library into your current environment. The as sns portionof the code then tells Python to give Seaborn the alias of sns.]**

# Methodology

**Our task was to figure out which group of people have the highest suicide rates.**

**Looking at the problem statement we realized we don’t need to use any AI/ML to resolve data, as we only needed to pull data from the given set.**

**We used Jupiter-Notebooks for initial code debugging , but soon ran into issues when we wanted to use dasher to generate graphs , as it required a production server due to WSDL , and Jupyter-Notebooks only supported Debugging Server.**

**We switched to Kaggle for ease of use in re-rerun of code.**

**As we were not able to resolve the given dataset (we did later on), so we used a data set provided by the government of India.**

**After successfully resolving such a large dataset of over 2.5 lakh inputs , we decided to go back to the main problem statement.**

**It needed a bit of clean-up using some of the libraries**

**After getting the graphs, we verified their accuracy , and the error was in null to negligible.**

**After identifying the reasons of suicide , we worked on ways to predict and prevent it.**

**We also have an website for suicide-prevention , but it’s an already working venture of ours , and minimal work has been done to it during the datathon.**

# GRAPH ANALYSIS

Chart, bar chart

Description automatically generatedChart, scatter chart

Description automatically generated

**The main goal of our model was to be as accurate in data representation as possible, so we tested our code output against a testing model that we knew was accurate.**

**We see a gradual increase in the number of suicides every year, however,  in 2012 we have a dip in the numbers as compared to 2011.**

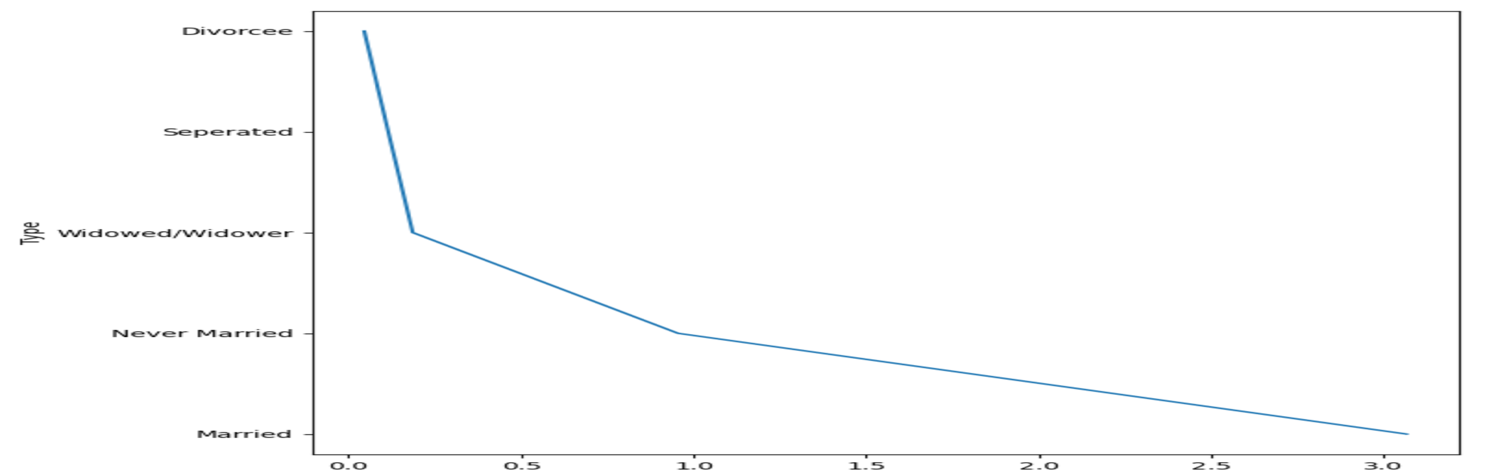
A picture containing waterfall chart

Description automatically generatedChart, bar chart

Description automatically generated

**Next, we look at the state-wise suicide numbers to find out that the states with the highest number of suicides are Maharashtra, West Bengal, Andhra Pradesh, Tamil Nadu, and Karnataka.**

**Exploring the type code feature gives us the below graph**



The type code gives us information on the basis of

a) Suicide causes b) Education status c) By means adopted d) Professional profile e) Social status

Let’s look at each of them one by one.

Chart, pie chart

Description automatically generated

**The huge spike was noticed in the graph below and what we essentially did is divide the aspects of that spike into a pie chart that you see above**

Chart

Description automatically generated

**The given graphs below are the graph renders of the problem statement. We converted the given datasets to .*csv* files and ran it through a graph maker and the results are as follows.**

Chart, bar chart, line chart

Description automatically generated

**The graphs above show us the deaths occurred across a range of age demographic. As you can see the spikes in this graph are explained in the graph bellow. For e.g., in the graph above the red dots displaying the age of 85 years and above, shows the highest number of deaths and in the graph below notice that the range of red dots on the top represent the age of 85 and above.**

Chart, histogram, scatter chart

Description automatically generated

**From our research what we found is that throughout the age demography Mental illness was a common reason a person committed suicide**

**We went further to prove this with another graph given below which shows an exponential rise in the mental illness range**

Line chart

Description automatically generated

# Solution to the problem

*First solution to the problem is recognising the problem here is how we do it.*

**Semantics Recognition as a way to identify trends in a workplace.**

NLP.py can be used by organisations to detect semantics of the mails, text-communication , and thus getting enough of an data to predict the feeling in the workplace and better treat their employees.

This can serve as a very good indicator to understand and analyse the general working conditions of the employees in the organisation , and thus it can help… one to focus on the key areas of hotspots to curb the issues

The model used here used Supervised learning , but feedback loop is yet to be worked out.

**Using Big data to curb suicidal tendencies.**

We can use today’s interconnected world to analyze the behavioral patterns of a person to actively predict what type of sentiment is the person having, such as sudden changes in ones’ sleep pattern , high intake or no intake at all of food , or not coming close to family members , can be good indicators of suicidal tendencies , but not only limited to it. Thus, IOT and Big Data, can be seemingly interconnected with advanced algorithms to get a hold on the pattern , and thus in turn…. The person can be shown adverts or recommendations in his feed to help him out of his situation without them ever knowing it.

# Our references

GitHub

Pyplot

CSV.plot

Dasher

<https://www.analyticsvidhya.com/>

Web-page- SELF from an another project , not created during datathon , used as a resource only.

Reference: for sucide prediction model

<https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=7a75b29b6668>

“”””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””””

Basics of data processing :

- Data Processing Steps : https://www.talend.com/resources/what-is-data-processing/

https://monkeylearn.com/blog/data-preprocessing/

- Libraries for data prep : https://www.simplilearn.com/top-python-libraries-for-data-science-article

a) Pandas : https://pandas.pydata.org/docs/

b) Pandas crash course : https://www.youtube.com/watch?v=tRKeLrwfUgU

c) Numpy : https://numpy.org/doc/

d) Numpy Crash Course : https://www.youtube.com/watch?v=uRsE5WGiKWo

- Machine Learning Libraries :

a) Tensorflow :

- https://www.tensorflow.org/api\_docs

- https://www.youtube.com/watch?v=6g4O5UOH304

- https://www.tensorflow.org/lite/guide/roadmap

b) Scikit-learn :

- https://scikit-learn.org/stable/

- https://www.youtube.com/watch?v=0Lt9w-BxKFQ

- https://scikit-learn.org/stable/modules/preprocessing.html

Picking and building a Model:

The first step of picking a model is to identify your problem.

An ML Problem can broadly split into two categories:

1) Supervised Learning: Supervised learning is a type of machine learning where an algorithm learns from labeled data. Labeled data means that each example in the data has an associated label or target value. The algorithm uses this labeled data to learn how to make predictions or classifications on new, unlabeled data.

2) Unsupervised Learning: Unsupervised learning is a type of machine learning where an algorithm learns from unlabeled data. Unlike supervised learning, there are no target values or labels associated with each example in the data. Instead, the algorithm tries to find patterns or structure in the data on its own.

Other Resources for Model Selection:

- Natural Language Processing:

https://www.datarobot.com/blog/what-is-natural-language-processing-introduction-t o-nlp/ https://towardsdatascience.com/roadmap-to-natural-language-processing-nlp-38a81d cff3a6

- Computer Vision:

https://www.ibm.com/in-en/topics/computer-vision

https://www.kaggle.com/getting-started/158267

- Classification:

https://dotnettutorials.net/lesson/classification-and-its-use-cases-in-machine-learning/

https://www.analyticsinsight.net/top-10-ml-classification-algorithms-for-data-scientists/

- Clustering:

https://www.analyticsinsight.net/top-10-ml-classification-algorithms-for-data-scientist s/ https://www.advancinganalytics.co.uk/blog/2022/6/13/10-incredibly-useful-clustering-al gorithms-you-need-to-know.

Data Visualization:

Basics

https://www.tableau.com/learn/articles/data-visualization-tips https://www.tableau.com/learn/articles/data-visualization

Tableau

https://www.youtube.com/watch?v=TPMlZxRRaBQ https://evolytics.com/blog/six-favorite-tableau-tips-tricks-and-hacks-to-enhance-dashboards/

MatPlotLib

https://matplotlib.org/stable/index.html https://www.youtube.com/watch?v=MU73e\_\_Luo0 https://matplotlib.org/cheatsheets/

Seaborn

https://www.youtube.com/watch?v=GcXcSZ0gQps https://seaborn.pydata.org

Report Preparation:

https://www.researchgate.net/publication/303326261\_Machine\_Learning\_Project

Note : Also has basics of machine learning and a lot of key topics.

Google Collab Cheatsheet :

https://towardsdatascience.com/cheat-sheet-for-google-colab-63853778c09