SUICIDE ANALYSIS AND PREDICTION

1.0 Introduction

This iteration is a deep dive into the suicide dataset to learn much more about the reasons for the thousands of suicides that occur each year around the world. Even though various studies on suicide have already been done previously, such as John et al. (2018), this study aimed to produce new insights that can help government bodies better grasp the problems that lie beneath them. This research could also benefit them in developing new strategies to minimize mortality rates over time. This research will look at a variety of suicide attributes and predict how many more fatalities will occur in various countries in the next years.

The goal of this research is to figure out why people commit suicide in each country. Every year, 800,000 individuals commit suicide, according to Wikipedia (2012). Suicide, for example, is becoming a more prevalent and serious problem in India, according to the World Health Organization (WHO). To address these issues, we must examine various patterns and clusters in the data and determine what circumstances cause someone to consider suicide. In addition, a web-based system will be developed that may offer dynamically illuminating visualizations of the suicide dataset, as well as opportunities for page administrators to submit new suicides to the dataset. This initiative will have a huge influence on society by allowing the government to identify and assist those who are in need, hence reducing the number of suicides each year in each nation. The government will not only save lives but also make the globe a better and safer place for people to live by implementing suitable steps based on the findings of this study.

2.0 Background Problem

Different social, economic, and cultural contexts exist in different countries. Russia and Ukraine, for example, are two of the most commonly mentioned countries recently. The world is aware that the two countries are involved in a major dispute. When you see that kind of observation and data insights in Explanatory data

analysis (EDA), it's always suspicious (there is a presumption that there is a relationship between the conflict and suicides in two countries); the two countries, among others, have high suicide rates.

In 2011, 554 people in Ireland committed suicide, according to the CSO statistical release (2011). In terms of the country's population, this is a large figure. Each suicide will have its own set of motives. Have you given any thought to the various reasons of these figures? You won't know the answers to these questions unless you start studying and researching suicides, like Zetzsche et al's (2007) research when they sought to figure out why people commit suicide in Western and Central Europe and came up with a few extremely interesting findings. moreover, all of the suicides might be due to a number of factors that we are not aware of. To put it differently, collecting all of that information is challenging, but there are some elements that push individuals to commit suicide in every country. Some of these common factors are included in the suicide dataset as features, which can be used to dig deeper into and analyze data from multiple countries.

2.1.0 Research questions

1. Check relation between GDP Per Capita and suicide rate
2. Which country is affected by the highest number of suicides with respect to population?
3. Which age groups is more likely to suicide?
4. Predict number of suicides going to happen in each continent in next 5 years
5. Predict top 5 countries with least number suicides in coming 5 years.
6. Find out the age group of people who are more likely to suicide?

3.0 Literature Review

ARIMA Model and FBProphet models are normally used for predicting suicide deaths around the world, the study of Kumar and Susan (2020) is a great example of the effective use of these approaches. Covid-19 was a very sensitive topic in recent years. The main goal of this study was to identify the future infected cases and virus spread rate for the preparation of the healthcare services to avoid deaths. In this study, they have used day level information on covid-19 spread for

cumulative cases from the whole world. The top ten most affected counties were the US, Spain, Italy, France, Germany, Russia, Iran, United Kingdom, Turkey, and India. They have used temporal data of coronavirus spread from January 22, 2020 to May 20, 2020. They have used ARIMA and Prophet models are effectively used for forecasting future infected cases and evaluation of the model is done using mean absolute error, root mean square error, root relative squared error, and mean absolute percentage error. This study has proved that ARIMA Model was more effective in forecasting covid-19 prevalence. Many people have been affected by it and lost their life. The dataset is very similar to what I have chosen for this suicide prediction as well. In this view, this is an excellent model to take inspiration from. The paper talks about countries including India and to understand the pattern in deaths that happened around the world. ARIMA and FBProphet models are used for analysis. Data has been split into training and testing. The trend analysis showed a rapid increase in the affected cases and the prediction study showed a great increase in the expected active, recovered, and death cases worldwide. However, as per their research, containment policies and lockdowns might affect the prediction results.

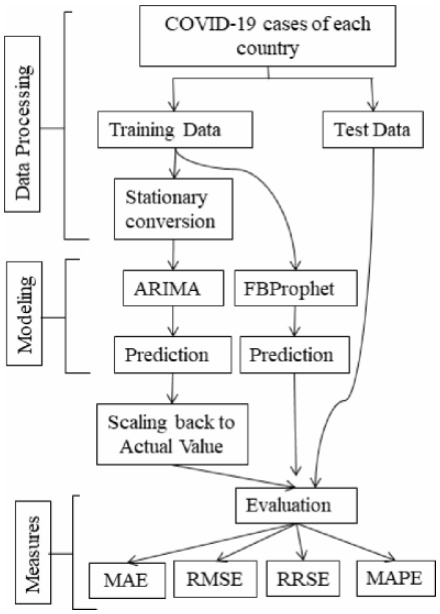


Fig. 1.0: *Framework to evaluate the forecasting model from Kumar and Susan (2020)*

In Qingdao, China, another study was conducted on forecasting cancer-related deaths in 2021. They've also employed the ARIMA Model to forecast fatalities. The ARIMA method integrates the autoregressive and moving average models into one. Airiti and her team conducted another study on exchange rate forecasting. This research makes use of ARIMA and Artificial Neural Networks.

The study of Singh et al. (2020) for predicting the covid-19 pandemic on time series data is also a similar example. But they have used the Support Vector Machine (SVM) model for the prediction. The objective of the research was to produce a real-time forecast using SVM Model. The purpose of this study was to investigate the coronavirus disease in the year 2019 prediction of confirmed, diseased and recovered cases. This prediction is used to plan resources, determine government policy, provide survivors with immunity passports and use the same plasma for care. The findings from the author indicate that Covid-19's daily mortality rate is positively correlated with several confirmed cases. The study also showed that it is dependent on the dietary routine and immune system. As per the author, an emergency can be awakened before the proper vaccine is invented. Another study from Włodarczyk (2021) predicting birth has used the same SVM model, this study was trying to figure out the preterm births. This study has also used machine learning algorithms like the random forest, K-Nearest Neighbor, and Convolutional Neural Network (CNNs) along with SVM.

SVM has been applied in Time series analysis, such as when Huang et al (2017) looked at classification issues in the Breast cancer dataset. Different kernel functions employed in the SVM Classifier were also investigated in this work. RBF kernel-based SVM ensembles based on boosting outperform other classifiers on large size datasets, according to their findings. Cortes and Vapnik (1995) were the first to develop SVM, which proved to be more effective for two-group classification issues.

Working with Multivariate time series data, I was looking for models which can make predictions on more than one variable. For example, the (Vector Autoregressive Models for Multivariate Time Series 2006) showed me how

relevant is VAR (Vector Autoregressive Models) model for suicide analysis. VAR Model is mainly used when we have to deal with Multivariate time series data. The suicide dataset contains more than twenty variables, which needs such a complex model like VAR. VAR is a systematic but flexible approach for dealing with complex real-world behaviour. VAR is also popular amongst data scientists because of its high forecasting performance. Most importantly VAR can capture the intertwined dynamics of time series data. We have to understand the lag using calculating AIC and BIC. AIC is considered in our case and looking at different lags from one to nine, we need to look where the AIC value is dropping quickly, we will be using that particular lag for the model fitting. Filho and Valk (2020) have implemented VAR model-based control charts for batch process monitoring In the field of Statistical Process Control (SPC). There were many approaches to deal with monitoring the batch process. A three-way data structure (batches x variables x time-instants). For each batch, there are multivariate time series data available. In traditional approaches, they do not take the nature of the time series data into account. They used multivariate techniques on the reduced two-way data. Recent developments in SPC have proposed to use the VAR model concerning the original three-way structure. However, they are restricted control approaches focused on VAR Models. This study has suggested a new method to deal with the batch process focusing on VAR.

4.0 Technologies and methods

The whole data was utilized for the visualizations in this project. I solely utilized data from the "Russian Federation" for the time series modeling and forecasting phase. Working with time series forecasting is an important part of this dissertation. I have several different targets in this dissertation including dashboard visualization, forecast modelling, database management etc. I have been looking for ways to predict the number of suicides in upcoming years. This interest in time series and ML made me dive deep into sophisticated time series models like SARIMA and VAR to make models on the suicide data and forecast future suicides

in different countries. The ARIMA model is a combination of multiple models, including the Autoregressive model, the Moving average model, and the Autoregressive Moving Average model. The form of the ARIMA model is represented by ARIMA (p, d, q), where p is the autoregressive order, d is the number of differences, and q is the moving average order.

Vector Auto Regressive Model is mostly used in finance and econometrics because they offer a framework for achieving important modelling goals, including data description, Forecasting, Structural Inference, and Policy Analysis. VAR Model is a workhouse time series multivariate model that relates current observations of a variable with past observations of itself and past observations of other variables in the system.

Thirdly, we need a database server for data to be stored on the server. I will be using PSQL or ThisSQL servers for data storage and management. I want the data in this DB to be updated from time to time and this model has to be updated based on the new data injected in each time. The reason for choosing these DB's is the flexibility of usage and its syntax matching with Structured Query Language (SQL) minute differences. Firstly, I want to talk about the python dashboard. It's always fascinating to see how we can make models and interpret them. But it is also important to note, recently there are many concerns about how well we can make modifications to the existing model and maintain them. So, our model has to work dynamic and make a prediction based on the available data. In recent years programmers used to use VueJS or web-based languages for making dashboards, we now have the most advanced packages like Streamlit has made these processes easier and more efficient. I am going to use some of the python packages like plotly to make an interactive dashboard and make models that can make great predictions.

Finally, I want to talk about the python dashboard. It's always wonderful to see how we can make models and interpret them. But it is also important to note, recently there were some concerns about how well we can make modifications to the existing model and maintain them. So, our model must work dynamically and make

a prediction based on the available data. In recent years programmers used to use VueJS or web-based languages for making dashboards, we now have the most advanced packages like Streamlit has made these processes easier and more efficient. I am going to use some of the python packages like plotly to make an interactive dashboard and make models that can make great predictions.

4.1 Data Preparation

In every data analysis, about seventy percentage the total time is spent on preparing the data and making it ready for doing analysis. Initially, I had to explore the dataset using describe () method as shown in fig 4.1. Also using visualizations and some statistical analysis I have cleaned the dataset.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13276 entries, 1 to 14030
Data columns (total 26 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   country               13276 non-null  category
 1   year                  13276 non-null  int64
 2   sex                   13276 non-null  category
 3   age                   13276 non-null  category
 4   suicides              13276 non-null  int64
 5   population            13276 non-null  int64
 6   sucid_in_hundredk     13276 non-null  float64
 7   country-year          13276 non-null  object
 8   yearly_gdp            13276 non-null  float64
 9   gdp_per_capita        13276 non-null  int64
 10  generation            13276 non-null  category
 11  suicides              13276 non-null  float64
 12  internetusers         13276 non-null  float64
 13  expenses              13276 non-null  float64
 14  employeecompensation  13276 non-null  float64
 15  unemployment          13276 non-null  float64
 16  physician_price       13276 non-null  float64
 17  laborforcetotal       13276 non-null  float64
 18  lifeexpectancy        13276 non-null  float64
 19  mobilesubscriptions   13276 non-null  float64
 20  refugees              13276 non-null  float64
 21  selfemployed          13276 non-null  float64
 22  electricityacess      13276 non-null  float64
 23  continent             13276 non-null  category
 24  country_code          13276 non-null  object
 25  mobilesubscription    13276 non-null  float64
dtypes: category(5), float64(15), int64(4), object(2)
memory usage: 2.3+ MB
```

Fig. 4.0: *showing information about the dataset using info() function in pandas*

Imputation was carefully done based on the time, context, and importance of the variable. I have chosen a dataset which was simple and aggregated. But, later on, thinking about the complexity and wide range of the reasons behind committing

suicide I did a thorough research about how much additional information I can incorporate into the existing dataset. There have been several variables like continent missing in the dataset. So, I have added additional columns for continent names. Also, I have received another dataset which is similar to the suicide master sheet I have previously received and contained much more information. The main reason behind taking this dataset into account is that those variables were very meaningful concerning the context I am working with, for example, I assume there could be some relation between suicide rates and unemployment or the number of internet users and suicides in any country.

```
df.describe(include='all').T
```

✓ 0.2s

	count	unique	top	freq	mean	std	min	25%	50%	75%
country	27820	101	Netherlands	382	NaN	NaN	NaN	NaN	NaN	NaN
year	27820.0	NaN	NaN	NaN	2001.258375	8.469055	1985.0	1995.0	2002.0	2008.0
sex	27820	2	female	13910	NaN	NaN	NaN	NaN	NaN	NaN
age	27820	6	15-24 years	4642	NaN	NaN	NaN	NaN	NaN	NaN
suicides	27820.0	NaN	NaN	NaN	242.574407	902.047917	0.0	3.0	25.0	131.0
population	27820.0	NaN	NaN	NaN	1844793.617398	3911779.441756	278.0	97498.5	430150.0	1486143.25
sucid_in_hundredk	27820.0	NaN	NaN	NaN	12.816097	18.961511	0.0	0.92	5.99	16.62
country-year	27820	2321	Belarus2007	12	NaN	NaN	NaN	NaN	NaN	NaN
yearly_hdi	8364.0	NaN	NaN	NaN	0.776601	0.093367	0.483	0.713	0.779	0.855
yearly_gdp	27820.0	NaN	NaN	NaN	445597926548.398254	1453907394884.571777	46919625.0	8985352832.0	48114688201.0	260000000000.0
gdp_per_capita	27820.0	NaN	NaN	NaN	16866.464414	18887.576472	251.0	3447.0	9372.0	24874.0
generation	27820	6	Generation X	6408	NaN	NaN	NaN	NaN	NaN	NaN

Fig. 4.1: *Describing the dataset using describe() function in pandas*

Many such variables were making this research firm on the ground in terms of working with useful and meaningful information for machine learning modelling. The second main reason is that data visualisation is a major part of this final project. If I had a greater number of variables in the dataset, I would get more opportunities of making more visualizations. Outliers in the data are one main thing we need to carefully do. Replacing the outliers without thinking about why they occur is a dangerous practice.

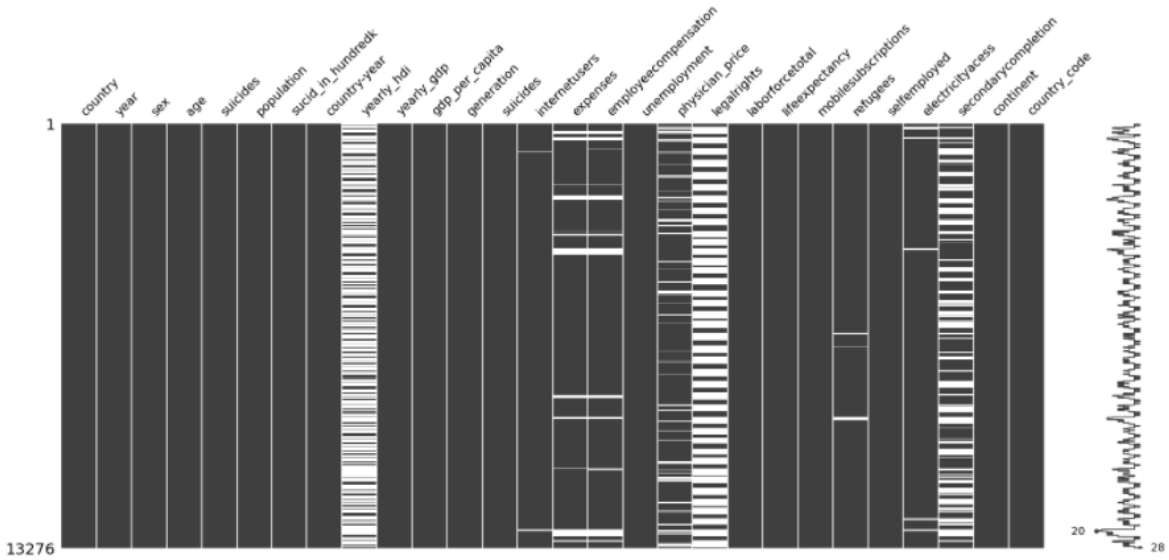


Fig. 4.2: *Checking for NaN or null values in the dataset*

Figure 4.2 gave me a clear idea of how much data is missing in each column of the dataset. Also used visualizations like boxplot and histogram, I have explored the dataset for data preparation.

4.2 Data Visualization

The dashboard about suicide was inspired by Johns Hopkins COVID-19 Dashboard (Coronavirus Resource Center n.d.). This dashboard displays real-time Covid-19 data that is regularly updated, with corresponding graphical representations created automatically. The python dashboard, which visualizes the data and insights for the public, will be the most appealing aspect of this project. Python dash and streamlit have emerged as the most capable frameworks for web-based visualisation projects in recent years. This project provides both static and dynamic visualisations. Before the real web dashboard app, individual static graphs are produced to obtain insight from the data. A final dashboard app with dynamic visualisations will be constructed when the initial static models are completed in Jupyter Notebook.

Suicide per hundred thousand around the world

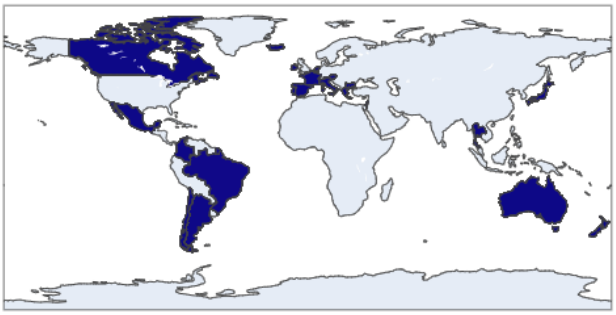


Fig. 4.3: *Suicide per hundred thousand around the world - Timeline in plotly*

Plotly is used to depict the global suicide rate in fig 4.3. Visitors can see information based on the year on an animation frame page. The rate of suicides per 100,000 is shown by colored zones.

Suicide per hundred thousand in Male and Female

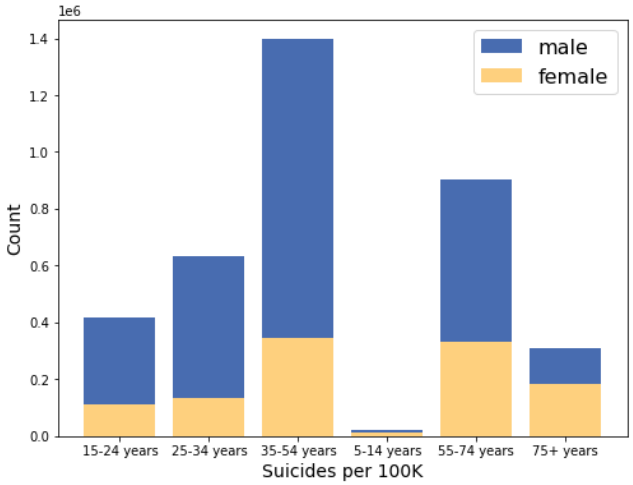


Fig. 4.4: *Suicide per hundred k among males and females in different age groups in different countries*

As per fig 4.4, Most suicides are happening between the age of 35 and 54. And out of the majority are Males. In all the age groups females are less affected groups. Also, we can see from the age of five to fourteen children are less likely to commit suicide.

Suicide per hundred thousand Vs GDP Per capita

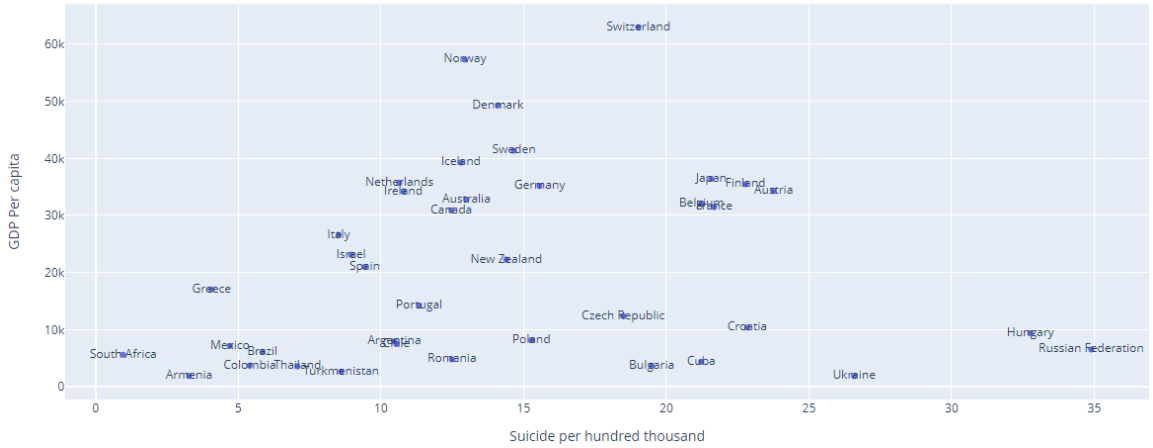


Fig. 4.5: *Suicide per hundred thousand around the world in different countries*

As seen in the fig 4.5, nations such as Russia, Ukraine, and Hungary have some form of relationship in terms of the number of suicides per 100,000 people. According to a BBC News article, the causes of suicide in Ukraine are the consequences of Russia-Ukraine hostilities and the ongoing war. Ukraine is undoubtedly one of the most afflicted countries in terms of suicide, according to our statistics.

Top ten countries with the highest suicide averages

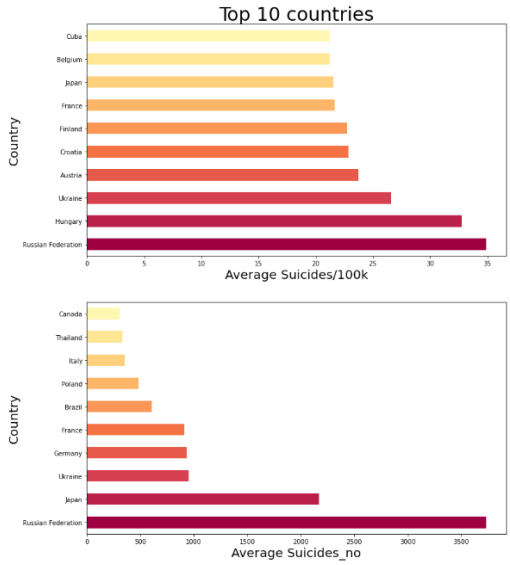


Fig. 4.6: *Top ten countries with the highest suicide rates*

The extent of suicides in various nations is seen in Figure 4.6. It is apparent that the Russian Federation has the highest suicide rate of all the countries studied. According to earlier research by Bellman and Namdev (2022), Russia has a considerable problem with suicidal behavior among men, and their drinking habits have a substantial impact on their decision to commit suicide when compared to other nations.

Population, Suicide, Suicide in Hundred Thousand Vs Gender and Age

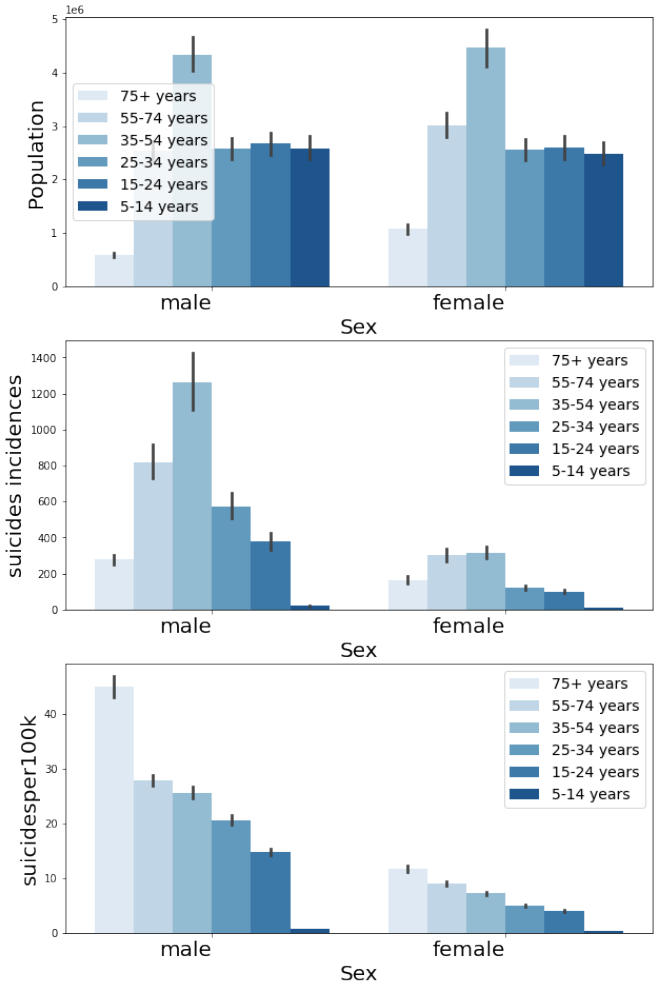


Fig. 4.7: *Suicide, Suicide per hundred thousand, Population in different genders and age in all countries*

As seen in fig. 4.7, Females outnumber men in terms of population. In relation to total suicides among men and women, the majority occurred between the ages of 25 and 34. It is undeniable that the male population has a higher suicide rate per 100,000. Furthermore, the majority of those who died were above the age of 75.

Population, Suicide, Suicide in Hundred Thousand Vs Gender and Age

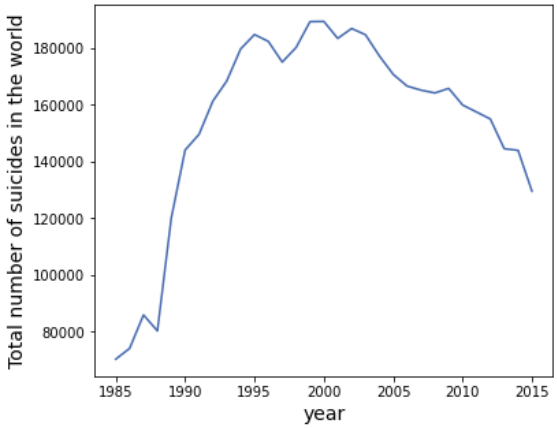


Fig. 4.8: *Total suicides in the world in each year distribution*

Fig 4.8 shows that there is a quick rise in the suicide rate from the year 1990. After that, it continued increasing until the next ten years. From 2000 it started to decline.

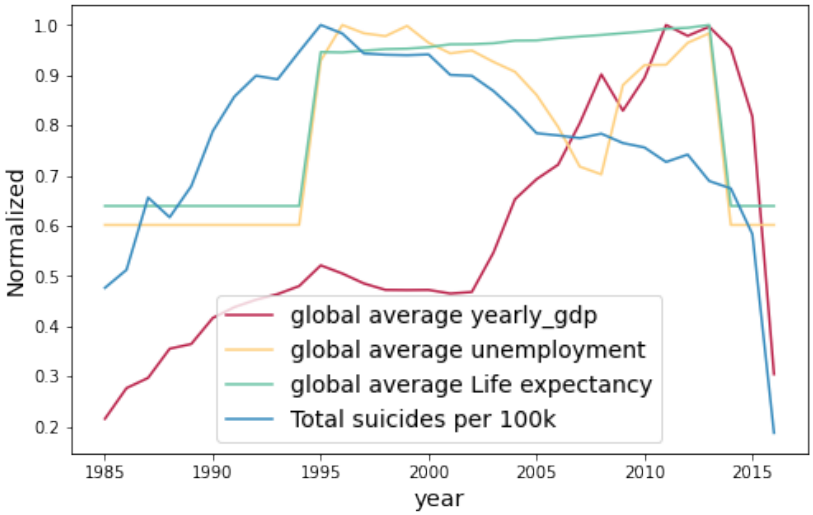


Fig. 4.9: Global Distribution of different features

According to the fig 4.9 above, the number of suicides per 100,000 climbed from 1985 to 1995, then rapidly fell. At the same time, global GDP per capita climbed gradually from 1985 to 1995, remained stable until 2003, then increased abruptly until 2014, before falling precipitously in 2015. (World Economic Situation UN). Life expectancy and unemployment have been influenced by missing data and imputation, making the lines appear irrelevant.

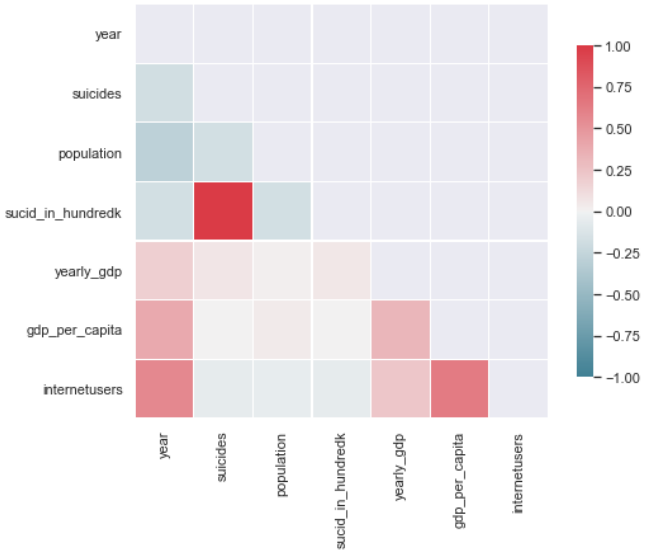


Fig. 4.10: *correlation matrix of suicide dataset*

Normally, checking for features in the correlation matrix is necessary to see whether there are any features with strong correlation. We usually eliminate such variables from the dataset since they might cause the model to overfit. As seen in fig. 4.10, there is a significant correlation between suicide and suicide per hundred thousand. As a result, before modeling, the suicides feature was deleted from the Dataframe.

4.3 Modeling and Forecasting

This project has various objectives, including creating a dashboard for visualizations, forecasting using Machine Learning Models, and creating an Admin Panel Portal for updating new suicides etc. Working with a time series model and projecting future values would be the most intriguing and challenging aspect of this

endeavor. A decision tree classifier was employed to categorize the risk and non-risk groups while dealing with numerical and categorical information. Brunello et al . seem to have effectively utilized them in their research, which prompted me to extend the concept to suicide analysis. Checking seasonality, trends, and stationarity, as well as assessing prediction tests for identifying the optimal model for prediction, such as AIC and BIC, are all part of this research.

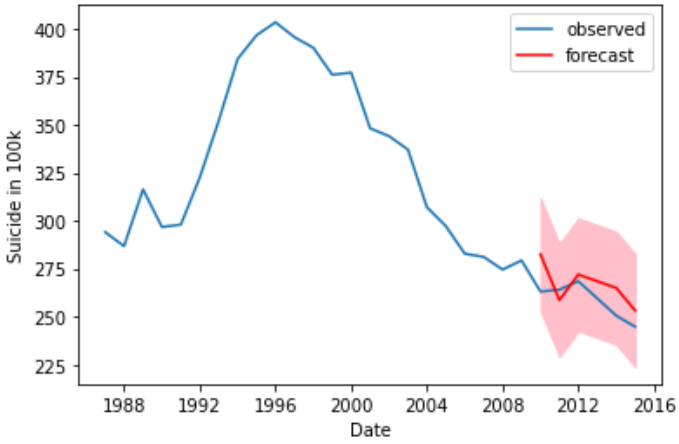


Fig. 4.11: *ARIMA Forecast and observed value graph on Russian data*

AIC and BIC graphs were made as shown in fig 4.11 for checking the order of the ARIMA model. In this research I am trying to work on predictions so, I will be looking at the AIC. A lower AIC score means a better predicting model. If the order is set too high, it could result in a high AIC value, this stops us from overfitting the training data. BIC is similar to AIC, lower BIC indicates a better model. BIC likes to choose a simple model with the lower order. AIC is better at predictive models but BIC is choosing a good explanatory model.

```
print(order_df.sort_values('aic'))
```

✓ 0.9s

	p	q	aic	bic
6	2	0	244.814959	248.811573
5	1	2	244.988533	250.317351
3	1	0	245.265917	247.930327
4	1	1	245.657204	249.653818
8	2	2	246.039121	252.700143
7	2	1	246.823370	252.152188
2	0	2	346.241586	350.238200
1	0	1	372.341183	375.005592
0	0	0	404.953923	406.286127

Fig. 4.12: *AIC and BIC Score in ascending order on Russian data*

Here Fig 4.12, I am looking at a better predicting model so, I will be choosing AIC with the least score. This is an ARMA(2,0) Model

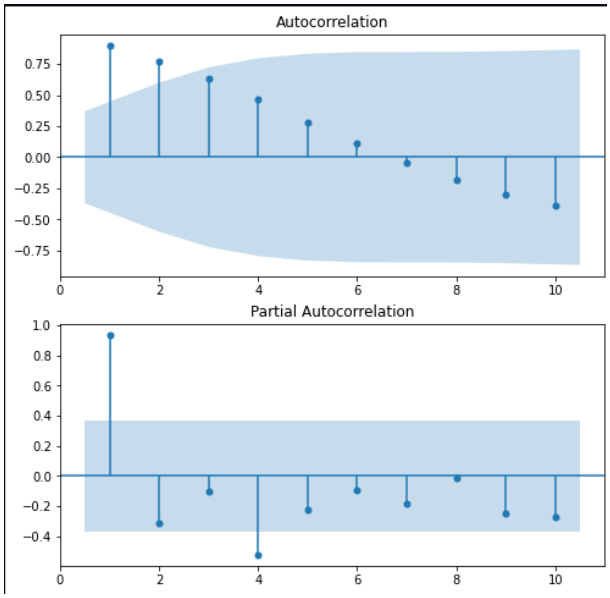
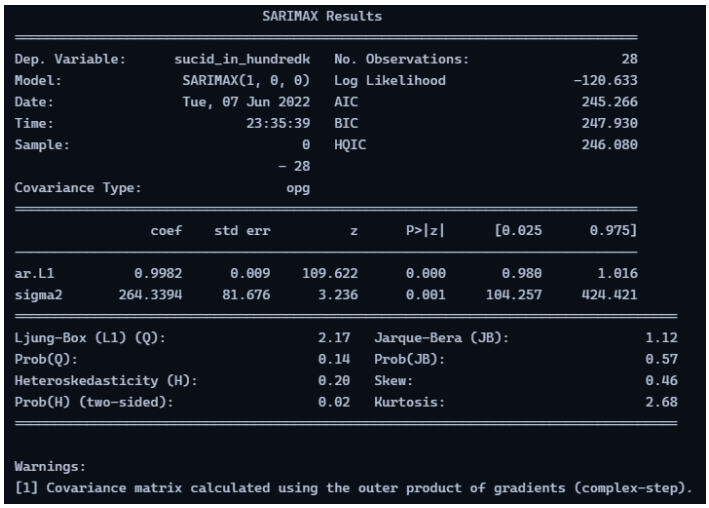


Fig. 4.13: *ACF and PACF to choose the model in ARIMA on Russian data*

From fig 4.13, In the above ACF and PACF, we can see ACF tails off and PACF cuts off since we have a MA(q) model. So this is an AR(1) Model.



SARIMAX Results

Dep. Variable:	sucid_in_hundredk	No. Observations:	28
Model:	SARIMAX(1, 0, 0)	Log Likelihood	-120.633
Date:	Tue, 07 Jun 2022	AIC	245.266
Time:	23:35:39	BIC	247.930
Sample:	0	HQIC	246.080
	- 28		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9982	0.009	109.622	0.000	0.980	1.016
sigma2	264.3394	81.676	3.236	0.001	104.257	424.421

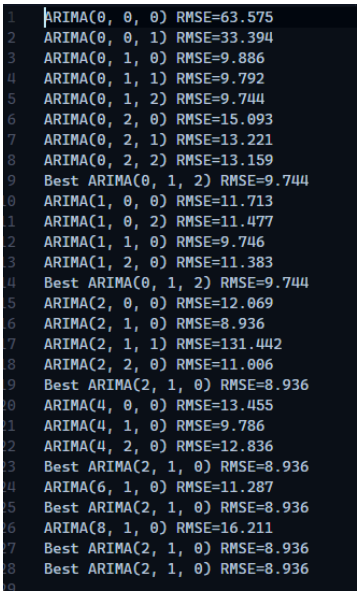
Ljung-Box (L1) (Q):	2.17	Jarque-Bera (JB):	1.12
Prob(Q):	0.14	Prob(JB):	0.57
Heteroskedasticity (H):	0.20	Skew:	0.46
Prob(H) (two-sided):	0.02	Kurtosis:	2.68

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Fig. 4.13: *Sarimax Model Results with order (1,0,0) on Russian data*

4.4.1 grid search ARIMA parameters for time series

We can automate the process of training and evaluating the ARIMA Models on different combinations of model hyperparameters. In machine learning, this is called grid search.



```
ARIMA(0, 0, 0) RMSE=63.575
ARIMA(0, 0, 1) RMSE=33.394
ARIMA(0, 1, 0) RMSE=9.886
ARIMA(0, 1, 1) RMSE=9.792
ARIMA(0, 1, 2) RMSE=9.744
ARIMA(0, 2, 0) RMSE=15.093
ARIMA(0, 2, 1) RMSE=13.221
ARIMA(0, 2, 2) RMSE=13.159
Best ARIMA(0, 1, 2) RMSE=9.744
ARIMA(1, 0, 0) RMSE=11.713
ARIMA(1, 0, 2) RMSE=11.477
ARIMA(1, 1, 0) RMSE=9.746
ARIMA(1, 2, 0) RMSE=11.383
Best ARIMA(0, 1, 2) RMSE=9.744
ARIMA(2, 0, 0) RMSE=12.069
ARIMA(2, 1, 0) RMSE=8.936
ARIMA(2, 1, 1) RMSE=131.442
ARIMA(2, 2, 0) RMSE=11.006
Best ARIMA(2, 1, 0) RMSE=8.936
ARIMA(4, 0, 0) RMSE=13.455
ARIMA(4, 1, 0) RMSE=9.786
ARIMA(4, 2, 0) RMSE=12.836
Best ARIMA(2, 1, 0) RMSE=8.936
ARIMA(6, 1, 0) RMSE=11.287
Best ARIMA(2, 1, 0) RMSE=8.936
ARIMA(8, 1, 0) RMSE=16.211
Best ARIMA(2, 1, 0) RMSE=8.936
Best ARIMA(2, 1, 0) RMSE=8.936
```

Fig. 4.14: *grid search result from ARIMA Model on Russian data*

In fig 4.14, we can see I have implemented the grid search and evaluated the ARIMA Model. Also, I have evaluated a set of different parameters.

4.4.2 Prediction using Vector Auto Regression Models (VAR Model)

Another Model used for the time series data is the VAR model (Vector Auto Regression). The reason behind using this model is that it helps in forecasting models based on multiple variables in time series. Usually, we use single variable and sequential time for time series analysis. But here I was able to include multiple variables in the model as you can see in the figure. Vector Autoregressive Models are one of the best models we could use to choose for time series.

year	sucid_in_hundredk_2d	gdp_per_capita_2d	lifeexpectancy_2d	expenses_2d
2012-01-01	529.412795	-4.469357e+06	-5481.612016	1599.709929
2013-01-01	-348.035624	1.086854e+06	16733.380993	1462.733174
2014-01-01	-393.171701	7.379111e+06	-17050.184437	-7497.633351
2015-01-01	240.134938	-3.715248e+06	833.249851	4506.396883
2016-01-01	905.938622	-1.050009e+07	5435.718296	6079.875040

Fig. 4.15: *predicting future values in VAR Model on Russian data*

You can see in fig 4.15, that we have predicted the number of suicides for the year 2016 using VAR Model on the time series sequential data.

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Wed, 08, Jun, 2022
Time:                     11:43:05
----------------------------------------------------------------------
No. of Equations:         4.00000    BIC:                    69.5772
Nobs:                     21.0000    HQIC:                   67.5521
Log likelihood:          -770.594    FPE:                 2.86088e+29
AIC:                      66.9908    Det(Omega_mle):      4.16352e+28
----------------------------------------------------------------------
Results for equation sucid_in_hundredk
================================================================================
                    coefficient       std. error           t-stat            prob
--------------------------------------------------------------------------------
const                  -7.307751        56.220354           -0.130           0.897
L1.sucid_in_hundredk   -0.182328         0.281278           -0.648           0.517
L1.gdp_per_capita      -0.000153         0.000074           -2.051           0.040
L1.lifeexpectancy      -0.142009         0.058078           -2.445           0.014
L1.expenses             0.334395         0.137784            2.427           0.015
L2.sucid_in_hundredk   -0.553061         0.234854           -2.355           0.019
L2.gdp_per_capita      -0.000121         0.000089           -1.351           0.177
L2.lifeexpectancy       0.018728         0.088536            0.212           0.832
L2.expenses             0.024843         0.196243            0.127           0.899
L3.sucid_in_hundredk   -0.253278         0.155461           -1.629           0.103
L3.gdp_per_capita      -0.000362         0.000096           -3.763           0.000
L3.lifeexpectancy       0.030652         0.062760            0.488           0.625
L3.expenses             0.067628         0.140647            0.481           0.631
================================================================================
```

Fig. 4.16: *Regression result summary from VAR Models on Russian data*

The coefficient scores for multiple variables in the time series analysis are shown in Fig 4.16. Each variable's standard error is also included, and coefficient values reflect the correlation of variables utilized to create the model's statistical equation.

I was able to create predictions on multi-variate time series data using VAR Models, as I described previously. The actual value of and the anticipated value distribution in the VAR Model are shown in Figure 4.17.

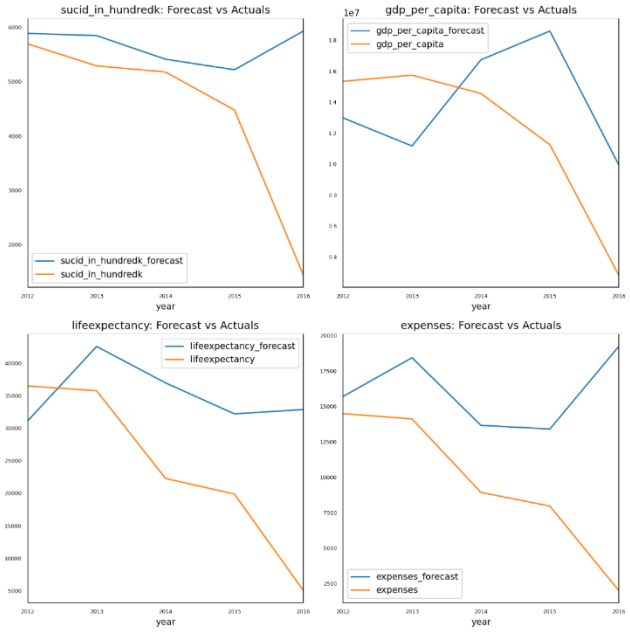


Fig. 4.17 *Plot of Forecast vs Actuals from VAR Models on Russian data*

From fig 4.17 you can see how the model forecasted different variables in the Russian suicide dataset concerning actual values.

4.4.3 Prediction using Auto Regression Models (AR Model)

The next model I have created is the Auto regression model, Train and Test were split into seventy and 30 per cent. Seventy per cent of the data was used for training the model and the rest thirty per cent was used for testing. I have got an 11.792 Root mean squared error. Also, I could save different models to the local and I was able to load the models later and update them accordingly.

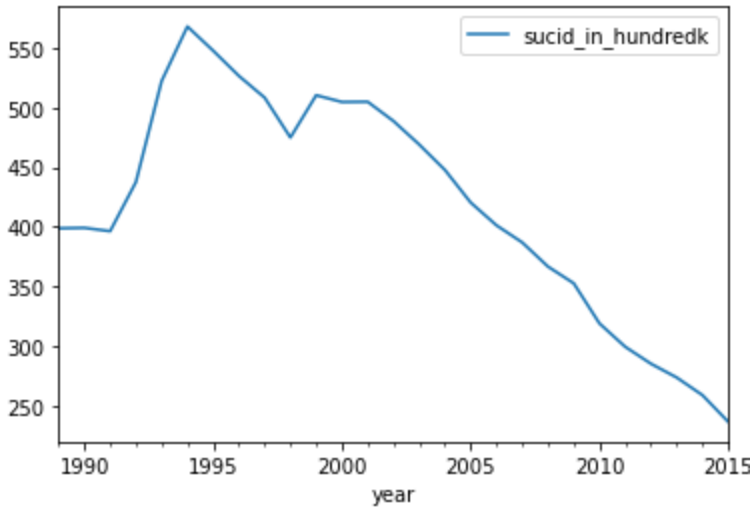


Fig. 4.18 *Plot of Forecast vs Actuals from VAR Models on Russian data*

As you can see in the diagram above, the suicides per hundred thousand are distributed throughout the year in 'Russian Federation' is shown.

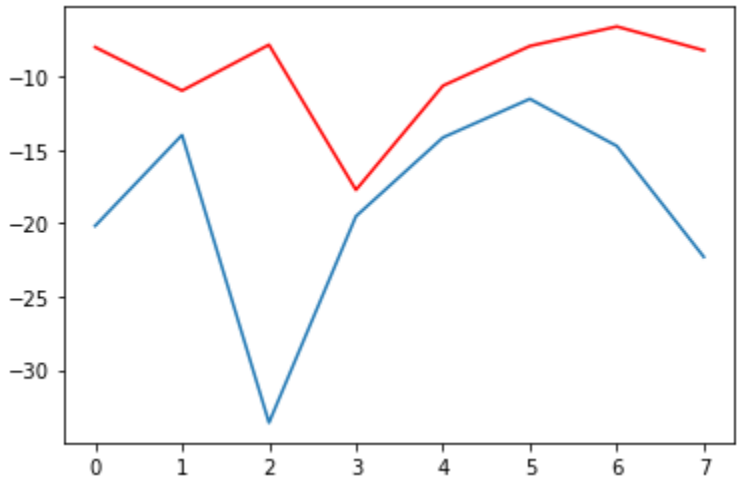


Fig. 4.19 *Plot of Forecast vs Actuals from AR Models on Russian data*

In this fig 4.18, the blue line is the test data and the red line is the predicted values. I have AR 1 Model with one window.

4.4.4 Decision Tree Classifier

I have added a new column called risk where I split the data into two classes, class 1 stands for high risk and class 0 for low risk. Using the decision tree model, I have made a classification.

```
******************Decision Tree classifier**************
Accuracy = 0.9962962962962963
Train Accuracy= 1.0
CM
 [[144   1]
 [  0 125]]
classification report for decision tree
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       145
           1       0.99      1.00      1.00       125

    accuracy                           1.00       270
   macro avg       1.00      1.00      1.00       270
weighted avg       1.00      1.00      1.00       270

# of leaves 4
 Depth 3
```

Fig. 4.20 *Plot of Forecast vs Actuals from AR Models on Russian data*

Fig 4.20 shows the result from the Decision Tree classification algorithm. Lee and Oh (1996) have done studies on Neural networks using a Decision Tree classifier to distinguish between complex features. Here this goals are to separate the risk group and non-risk groups based on the feature called risk. I got 99.62 training accuracy and a hundred per cent testing accuracy.

4.4 Evaluation of models

Evaluation of the model is as important as making the model. I have created 3 models in ARMA, Auto regression and Vector Auto regression. Using mean squared error and R-squared error I took the error rates of different models. Accuracy is also calculated to understand how efficient and precise this model is.

Despite the fact that the model below was constructed using a Russian dataset. The live dashboard and modeling for the final product will be based on the whole country's master dataset.

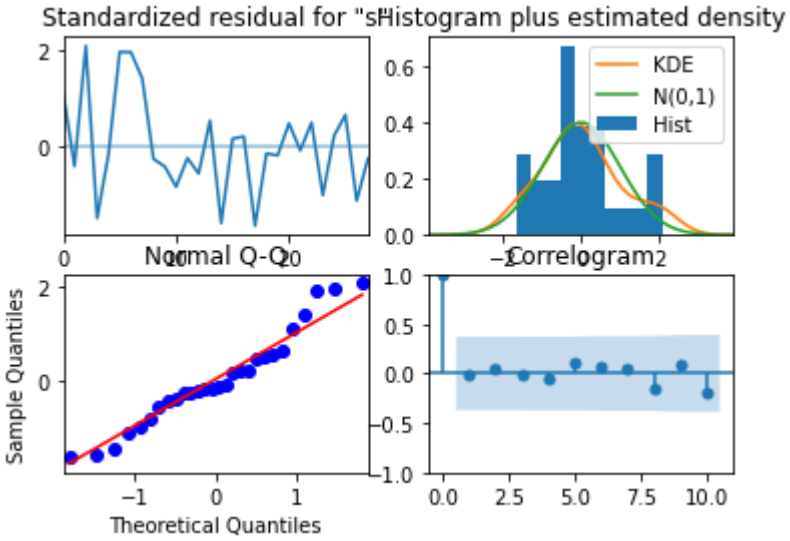


Fig. 4.21 *ARIMA Diagnostic plots on Russian data*

After finalizing this models, I would be able to add more evaluation techniques. Above are the four diagnostic plots I have created after running ARIMA Model. Looking at the plots I can say there is no pattern in the standardized model. Looking at the Histogram there is no Gaussian distribution (green and red lines should be almost the same for gaussian). The QQ Plot seems to be not normally distributed, if it is normally distributed all the blue dots will be aligned over the line (except for some values in either end)

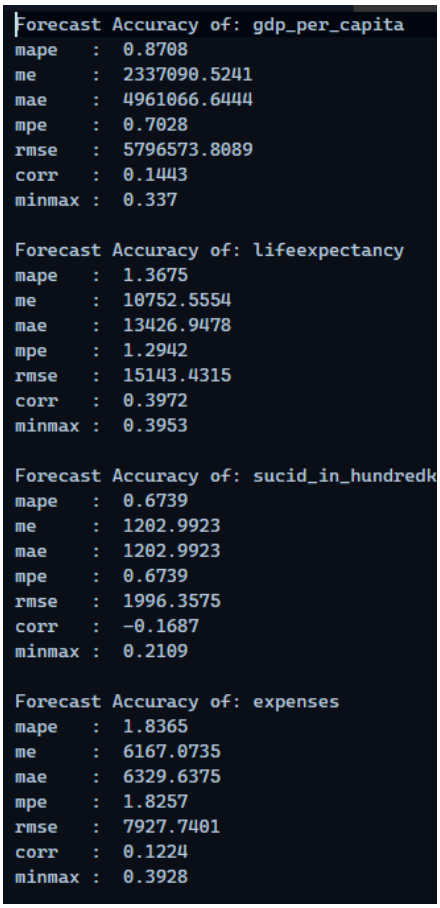


Fig. 4.22 *Plot of Forecast vs Actuals from VAR Models on Russian data*

As you can see in fig 4.22, we have calculated the accuracy of each variable. This helps us understand how well our model is performing.

4.5 Data Storage

Thirdly, we need a database server for data to be stored on the server. I will be using PSQL or ThisSQL servers for data storage and management. I want the data in this DB to be updated from time to time and this model has to be updated based on the new data injected in each time. The reason for choosing these DB's is the flexibility of usage and its syntax matching with Structured Query Language (SQL) minute differences. Initially, data was stored in the CSV format in different files. Later I uploaded them into mochahost Psql server.

4.6 Web Server and Hosting

We know there are thousands of hosting companies providing hosting services, I have chosen Mochahost as one of the best service providers for small business websites. This goal is to make a highly dynamic web application on the server. I have purchased a VPS service which allows running PIP Packages on their server making the server IDE more suitable for the Dash App. Mochahost cPanel will be connected to the Github repository where this application is updated from time to time. Using git technology for the hosting makes the process more sophisticated and professional in terms of version control.

4.7 Data Security

Data Security has become an important concern in this era. Even though the suicide dataset is publicly available on the internet, I have followed the best practices in data security to ensure there is no data leakage. I have used encrypted windows drive to store the data. Whole project codes are updated from time to time to GitHub private repository. Any information related to this study has been considered for data security and ethical practices before actually using them. No personal information is used in this study. For web applications, files are kept in a private repository and used that repo to pull changes to the live mocha host server.

4.8 Applications and software

The most popular IDE for coding is Microsoft Visual Studio. I've been using Github for version control. In addition to Jupyter Notebook, Spyder, and Atom, I've used various coding tools such as Jupyter Notebook, Spyder, and Atom. All of the testing is done in the Anaconda environment on the local machine. For the whole study, Python version 3.8.8 was employed. The PIP package is needed to set up the IDE. Minor csv CSV file checks are also performed using Microsoft Excel. The

built-in git version control facilities in Visual Studio are occasionally used to manage the repository's branches.

MS Word and notepad are used for reporting and notes. PowerPoint is a presentation software that allows you to create slides. PDF files are managed with Adobe Acrobat DC. !pip and git commands are run via the built-in terminal in Visual Studio, Anaconda Prompt, and Windows Terminal. The entire project is run on the Windows operating system. The browsers utilized in this experiment are Google Chrome and Mozilla Firefox.

5.0 Ethical Considerations

To begin, consider some recent unethical events. When a researcher's action has a negative impact on participants or society, it is considered harm. There are a number of reasons why researchers' actions cause so much harm for society or people. Likewise, one of the most obvious examples of such accidents was Human Radiation Experiments (2017) was one of the biggest examples of such incidents. In 1994 US President Clinton created an advisory team to research human radiation that has been conducted over the years. In this study, doctors injected Plutonium into the body of many patients and many of them did not consent to be part of this study. Also, there was a company called Quaker Oats (2020) which is also part of this study included radioactive components in oatmeal and were unknowingly fed to the children.

In this study, no such experiment is done on humans in the process of data collection or analysis. An aggregated suicide dataset only provides information about the country's general population and related detail as features is used throughout the research. No prior experiment is conducted to gather data for this research. No harm is made to any subject in this regard. There are several benefits related to the data. Data provides an overview of how many suicides are happening from time to time. Talking about the societal impact of this research is enormous. For example, Study of Benefits of Electric Cars (2016) has created a significant impact on how this research has benefited society to help understand the carbon footprint reduction and cost-saving. In suicide analysis, I am trying to

make use of data to leverage suicide attempts by helping the government to take measures or policies from the outcome of this study it's going to help create plans to tackle such acts in coming years.

This study of suicide analysis was based on a dataset that is open source and available for download on Kaggle. In terms of data storage and security, I wouldn't call it a particularly sensitive dataset since, for starters, this information is not private on the internet, and the creator has left public access open. Second, this suicide dataset does not contain any personally identifiable information; rather, it is a summary dataset that provides broad statistics on the country's mortality rates. Also, information on who is more prone to commit suicide, such as age group, internet users, human development index, and so on.

5.0.1 SWOT Analysis

The suicide dataset, as previously indicated, did not contain any personal information. I strongly believe that in future research, we should include more humans in the trial in order to get data from individuals in real time. The most critical step is to obtain written consent from everyone who wishes to participate in the study. Consider what would happen if these human individuals who are participating in the study/experiment had not given their consent. They may eventually take us to court (Facebook Data Breach BBC) and file a complaint accusing us of misusing personal information. In addition, we must explicitly disclose what activities or hazards are involved in the research. So that people are aware before they participate in these activities. (Bogod 2004) is one of the real-life examples where in 1942 prisoners were asked to undergo dangerous experiments to understand the survival chance of soldiers sometimes even leading to deaths. Understanding personal, social, and business impacts of data practice. In addition, even sharing information of individual sharing with any other colleagues or third party would be through proper procedure and getting signs on consent forms.

Strength: - In this study, I am trying to see suicide rates in different countries from time to time. This research strength is its dynamic nature. Similar weather forecast

of google or Microsoft, this model will be run from time to time based on the latest data. This research aims at tackling suicide tendencies in every country's population. This research is going to predict how many people are going to commit suicide in the next 5 years in different countries or continents. When working with a socially responsible research project, it is going to stand out in the world of the internet. Similar to the websites showcasing covid trends live, this website is also going to show the same impact of suicide numbers and create respective visualizations for any general audience to easily understand what the trend in data would be.

Weakness: - For forecasting, the data is aggregated, and no unique information about individuals is provided. As a result, I believe the data must have more precise traits that may be used to create reliable suicide predictions. However, if additional particular characteristics could have been added to the dataset, the model would have been more accurate. Things like topic diagnostic information, population happiness index, education index, and each country's happiness index. As a result, this forecast is more of a broad grasp of data trends.

Opportunity: - It's difficult to put into words how much can be learned through suicide data analysis. The government is working to figure out what causes suicides and how to reduce the number of suicides each year. We can build creative strategies to lessen the effect of suicides by studying and interpreting current statistics. Machine Learning models might be used to develop smart apps that help mobile users based on their activity data. Suicide analysis ushers in a new era of artificial intelligence in which we can track who is on the verge of dying.

Now let's look at this data and its opportunities. Have you ever thought of having a suicide prediction model for each country? The wide range of opportunities using AI and the Time Series model on big data is possible using current technologies. Internet of things, cloud computing and Machine Learning are the best examples of state-of-the-art technologies. The suicide prediction model and live dashboard

visualization is a great analysis model which any growing business can take inspiration from. Just imagine a burger selling vendor creating a live predicting model of a specific kind of burger that is sold at a particular season of a year? or maybe checking bestselling milkshakes each month? Wouldn't these analyses make them grow? or even predict how many products are going to be sold in the coming months so they can prepare their store for the coming period to avoid lack of materials. Thus, this model is ultimately showing what kind of predictions or analysis our business and health industry need today to go smarter and do smarter businesses.

Threat: - Data may be utilized in a variety of ways. Some individuals utilized it for good causes, while others exploited it in a different way. It's possible that the suicide data will be abused in some way. However, from this perspective, they are less likely to occur as long as we do not provide detailed information about people. In this scenario, I'd argue that if new features are added to the model in the future, I'll have to change the model statically and then make it dynamic using cron tasks. Furthermore, additional storage space may be necessary in the near future when it comes to keeping individual information, and this model may function badly due to server needs. Even if we have alternative possibilities for purchasing cloud storage space, it will still be more expensive; still, I will have to find ways to enhance the needs. When it comes to the examination of prior years' suicides, the differences in counts over different political administration eras might have a political influence.

6.0 Conclusion

Models such as ARIMA, VAR, and AR were utilized to investigate and forecast the effect of suicide in different nations in this work. All of the models were built using the 'Russian Federation' suicide dataset as a starting point. When compared against other algorithms, the ARIMA model outperformed the others. One of the models, called SVM, has been found as a good fit for working on the

suicide dataset, which must be done alongside other models that have already been completed. In comparison to other models, ARIMA and Decision tree classifiers provided me with greater accuracy. From this point forward, I'll be working on the primary dashboard, which is an online tool that makes real-time forecasts based on data input.

7.0 Proposed Future Analysis

The preliminary analysis of this research reveals that additional specific data is necessary for the project. In order to produce more accurate and meaningful forecasts in the future, daily suicide data with more precise personal information will need to be obtained. Future development will involve expanding the backend options to include a form with more fields particular to the person who died so that individual data may be collected. For instance, the person's blood group, health information, alcohol use, and so on. More models, such as SVM, will be used in future study to assist us understand the best approaches for forecasting and achieving optimal results.

8.0 References

(3) 1/4: What is Streamlit - YouTube. Available from: https://www.youtube.com/watch?v=R2nr1uZ8ffc [accessed 4 June 2022].

Airiti Library_Comparative+Study+of+Artificial+Neural+Network+and+ARIMA+Models+in+Predicting+Exchange+Rate. Available from: https://www.airitilibrary.com/Publication/alDetailedMesh?docid=20407467-201211-201512080011-201512080011-4397-4403 [accessed 8 April 2022].

Bellman, V. and Namdev, V. (2022). Suicidality Among Men in Russia: A Review of Recent Epidemiological Data. *Cureus* [online], 14(3). Available from: https://www.cureus.com/articles/88128-suicidality-among-men-in-russia-a-review-of-recent-epidemiological-data [accessed 8 June 2022].

Bogod, D. (2004). The Nazi Hypothermia Experiments: Forbidden Data?. *Anaesthesia* [online], 59(12), pp.1155–1156.

Brunello, A., Marzano, E., Montanari, A. and Sciavicco, G. (2019). J48SS: A Novel Decision Tree Approach for the Handling of Sequential and Time Series Data. *Computers 2019, Vol. 8, Page 21*

[online], 8(1), p.21. Available from: https://www.mdpi.com/2073-431X/8/1/21/htm [accessed 7 June 2022].

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning* [online], 20(3), pp.273–297.

COVID-19 Map - Johns Hopkins Coronavirus Resource Center. Available from: https://coronavirus.jhu.edu/map.html [accessed 11 June 2022].

Filho, D.M. and Valk, M. (2020). Dynamic VAR model-based control charts for batch process monitoring. *European Journal of Operational Research* [online], 285(1), pp.296–305.

Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Yamaguchi, R., Miyano, S., Sogayar, M.C. and Ferreira, C.E. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology* [online], 1.

How suicide became the hidden toll of the war in Ukraine - BBC News. Available from: https://www.bbc.com/news/world-europe-60318298 [accessed 8 June 2022].

Huang, M.W., Chen, C.W., Lin, W.C., Ke, S.W. and Tsai, C.F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLoS ONE*, 12(1).

Human Radiation Experiments | Atomic Heritage Foundation. Available from: https://www.atomicheritage.org/history/human-radiation-experiments [accessed 15 April 2022].

John, A., Glendenning, A.C., Marchant, A., Montgomery, P., Stewart, A., Wood, S., Lloyd, K. and Hawton, K. (2018). Self-Harm, Suicidal Behaviours, and Cyberbullying in Children and Young People: Systematic Review. *J Med Internet Res 2018;20(4):e129 https://www.jmir.org/2018/4/e129* [online], 20(4), p.e9044. Available from: https://www.jmir.org/2018/4/e129 [accessed 9 June 2022].

John, A., Okolie, C., Eyles, E., Webb, R.T., Schmidt, L., McGuiness, L.A., Olorisade, B.K., Arensman, E., Hawton, K., Kapur, N., Moran, P., O'Connor, R.C., O'Neill, S., Higgins, J.P.T. and Gunnell, D. (2020). The impact of the COVID-19 pandemic on self-harm and suicidal behaviour: a living systematic review. *F1000Research 2020 9:1097* [online], 9, p.1097. Available from: https://f1000research.com/articles/9-1097 [accessed 7 June 2022].

Kumar, N. and Susan, S. (2020a). COVID-19 Pandemic Prediction using Time Series Forecasting Models. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020* [online], 1 July 2020.

Kumar, N. and Susan, S. (2020b). COVID-19 Pandemic Prediction using Time Series Forecasting Models. *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020* [online], 1 July 2020.

Lee, K.C. and Oh, S.B. (1996). An intelligent approach to time series identification by a neural network-driven decision tree classifier. *Decision Support Systems* [online], 17(3), pp.183–197.

Mochahost Review 2022: Mocha Host Details, Pricing & Features | Sitechecker. Available from: https://sitechecker.pro/web-hosting/mochahost.com/ [accessed 7 June 2022].

Qi, F., Xu, Z., Zhang, H., Wang, R., Wang, Y., Jia, X., Lin, P., Geng, M., Huang, Y., Li, S. and Yang, J. (2021). Predicting the mortality of smoking attributable to cancer in Qingdao, China: A time-series analysis. *PLOS ONE* [online], 16(1), p.e0245769. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245769 [accessed 8 April 2022].

Singh, V., Poonia, R.C., Kumar, S., Dass, P., Agarwal, P., Bhatnagar, V. and Raja, L. (2020). Prediction of COVID-19 corona virus pandemic based on time series data using support vector machine. *https://doi.org/10.1080/09720529.2020.1784535* [online], 23(8), pp.1583–1597. Available from: https://www.tandfonline.com/doi/abs/10.1080/09720529.2020.1784535 [accessed 8 June 2022].

Study: Benefits of Electric Cars Add Up—in the Billions! | NRDC. Available from: https://www.nrdc.org/experts/luke-tonachel/study-benefits-electric-cars-add-billions [accessed 15 April 2022].

Suicide Statistics 2011 - CSO - Central Statistics Office. Available from: https://www.cso.ie/en/releasesandpublications/er/ss/suicidestatistics2011/ [accessed 7 June 2022].

Tang, L., Pan, H. and Yao, Y. (2018). K-nearest neighbor regression with principal component analysis for financial time series prediction. *ACM International Conference Proceeding Series* [online], 12 March 2018, pp.127–131.

Värnik, P. (2012). Suicide in the World. *International Journal of Environmental Research and Public Health* [online], 9(3), pp.760–771.

Vector Autoregressive Models for Multivariate Time Series. (2006). *Modeling Financial Time Series with S-PLUS®* [online], 9 October 2006, pp.385–429. Available from: https://link.springer.com/chapter/10.1007/978-0-387-32348-0_11 [accessed 4 June 2022].

When Quaker Oats Fed Children Radioactive Oatmeal | by Calin Aneculaesei | History of Yesterday. Available from: https://historyofyesterday.com/when-quaker-oats-fed-children-radioactive-oatmeal-5e06faf3ce4d [accessed 9 June 2022].

Włodarczyk, T., Płotka, S., Szczepański, T., Rokita, P., Sochacki-Wójcicka, N., Wójcicki, J., Lipa, M. and Trzciński, T. (2021). Machine learning methods for preterm birth prediction: A review. *Electronics (Switzerland)*, 10(5).

Zetzsche, T., Bobes, J., de La Fuente, J.M., Pogarell, O., Norra, C., Schmidtke, A., Wasserman, D., Löhr, C. and Rihmer, Z. (2007). Changing suicide rates in western and central Europe. *European Psychiatry* [online], 22(S1), pp.S35–S35. Available from: https://www.cambridge.org/core/journals/european-psychiatry/article/changing-suicide-rates-in-western-and-central-europe/1B2C943626D6D31150E00FCD3CECFDA9 [accessed 11 June 2022].