A brief healthcare development indicators overview for developed and underdeveloped countries.

DAB 304 Healthcare Analytics: Project Report

Akhil Nandakumar (0775505)
Data Analytics for Business
St Clair College
Windsor, Canada

Akshay Joshi (0773818)
Data Analytics for Business
St Clair College
Windsor, Canada

Vijayalaxmi Rohane (0775039)
Data Analytics for Business
St Clair College
Windsor, Canada

Shrikant Narawane (0775146)
Data Analytics for Business
St Clair College
Windsor, Canada

Abstract:

The overall countries development worldwide depends on how the country is progressing in the field of healthcare. Most of the countries end up spending maximum percentage of their overall GDP in healthcare field. Healthcare profusely affects the country's expenditure and based on how much the country is ready to spend on healthcare, the human development index of one's country will go higher or lower. Here we provide a brief analysis on what those factors are which affect the healthcare domain. We have used analytical tool such as tableau and implemented several machine learning models like Random Forest and ARIMA (Auto-Regressive Integrated Moving Average) to observe the future trends.

Introduction:

The motivation behind this project is to investigate factors that determine various countries' government expenditure on healthcare and doing a comparison of the same with Developed, Developing and Underdeveloped countries. Unlike other public goods and services, the healthcare industry exhibits uncertainty which gives opportunity to data scientist to analyse the trends and come up with strategies to benefit the government on providing insights for budgeting healthcare domain. Another goal behind this project is studying if unemployment and reductions in government healthcare expenditure are associated with significant increases in different diseases globally.

Whenever any kind of crisis emerges where the country's economy is failing this analysis will tell on how far the adverse effects on healthcare that country will suffer from.

The main object of the project is to compare various variables across different countries and determine how countries are budgeting their finances in healthcare field. Thus, we can consider the project complete if we can differentiate the finances and draw conclusions based on the different comparison graphs that we are planning to build. The different factors that we will be studying are total population, healthcare expenditure, GDP, healthcare workers, etc. The success of the project will be estimated if it can deliver the correlation between the healthcare and different economics of various countries.

Related Work:

Government Healthcare Spending, Unemployment and Cerebrovascular Mortality, worldwide. [1]

Data was obtained from the World Bank and World Health Organization. Multivariate regression analysis was used to assess the effect of changes in unemployment and government healthcare expenditure on cerebrovascular mortality. For every 1% increase in unemployment, was associated with a significant increase in cerebrovascular mortality. Every 1% rise in government healthcare expenditure, across both genders, was associated with significant decreases in cerebrovascular deaths.

Recession and its effects on health and related activities of Americans. [3]

The above paper studies the impact of current crisis on the individual's health related activities and how macroeconomy affects health. The paper determines the relation between rising unemployment levels and the probability that they have no health insurance. Statistically significant decline in probability of visiting a hospital in the 12 months following the recession, as a direct result of rising unemployment.

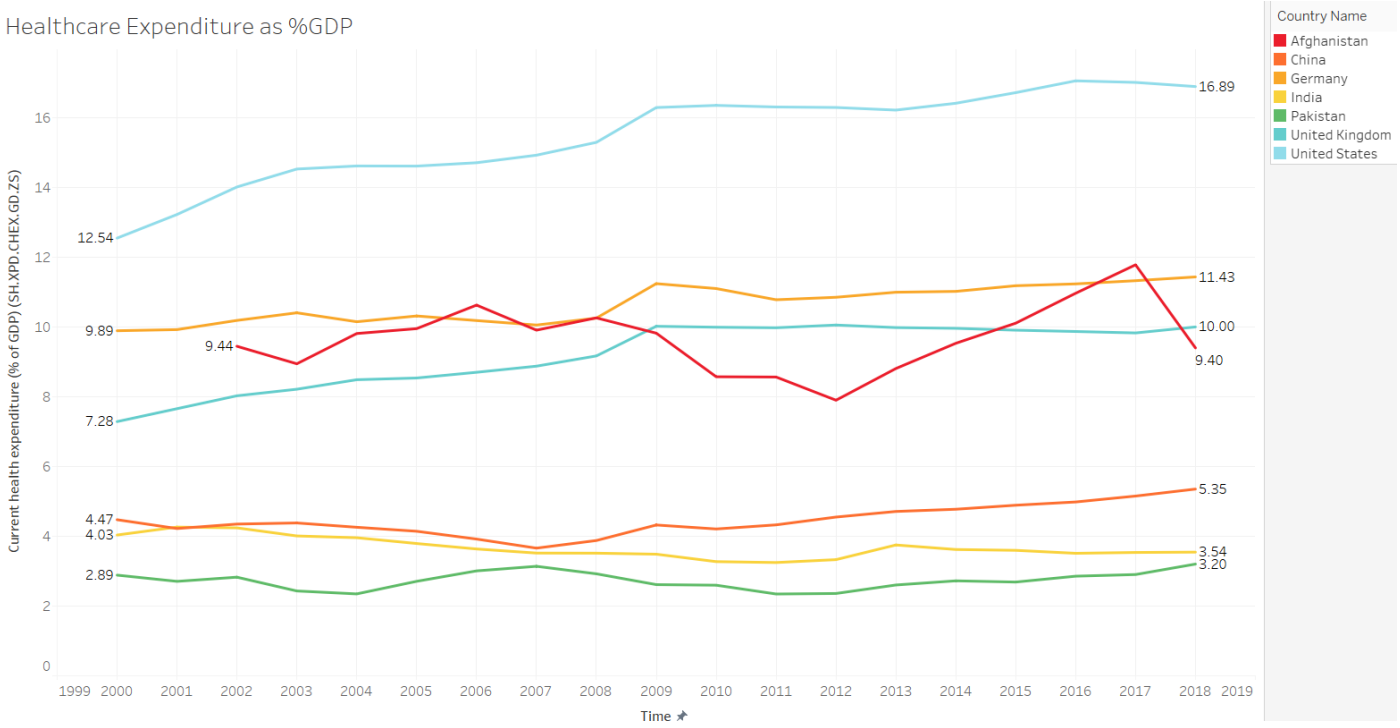
Determinants of healthcare expenditures in GCC countries. [2]

Identifying factors that determine GCC countries' government expenditure on healthcare. Used econometric models and carried out regression analysis on the data. Simultaneous cross-sectional analysis (integrating data from different countries) and the analysis of time series involving data from different periods. Healthcare expenditures had a positive and significant effect on variables related to government revenues, population, and government debt. Foreign exchange reserves to maintain the expected level of spending on the healthcare sector because their public revenues depend mainly on the oil revenues.

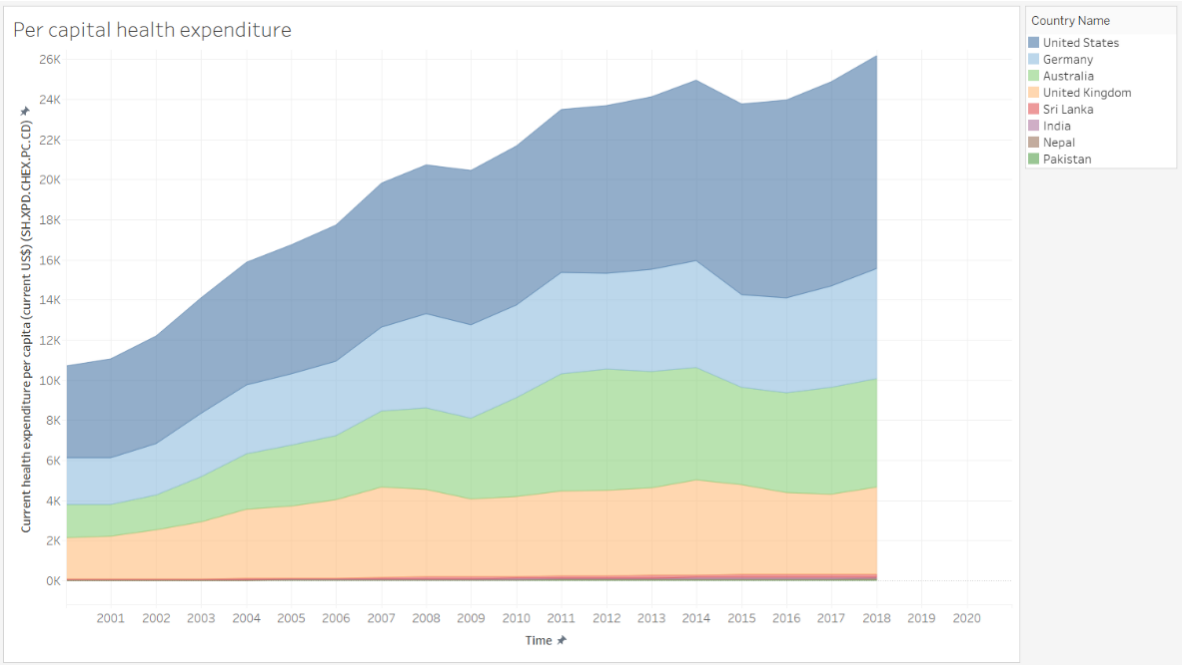
Methods

This project is intended to discover the effects of GDP on healthcare and the various events that happened in the past that had a massive effect in the healthcare industry starting from the past 20 to 30 years. The dataset is taken from The World Bank – World Development Indicator dataset. The dataset contains various demographics about the economics of different countries such as various mortality, population, health expenditure etc.

After extracting the data, we have selected the features, based on domain knowledge of healthcare industry, that we thought would be relevant to perform trend analysis (Refer EDA code from appendices). We have taken care of null values from the columns, and converted the year column to date time. We have dropped all the highly corelated columns to avoid overfitting in the model. To understand the final dataset, we have plotted some visualization for different features. We have plotted trend lines for all these features up-to the year 2030 to analyse them.



The graph here shows the Healthcare expenditure as percentage of the overall GDP. United states has seen gradual increase in healthcare expenditure growth over the last 20 years, and levels in 2018 stands significantly higher than the listed countries at 16.68. Pakistan is the lowest here among the countries selected, at a level of 3.20 in 2018. Afghanistan has seen a fluctuation over the past 20 years with a significant drop from 2017 to 2018.



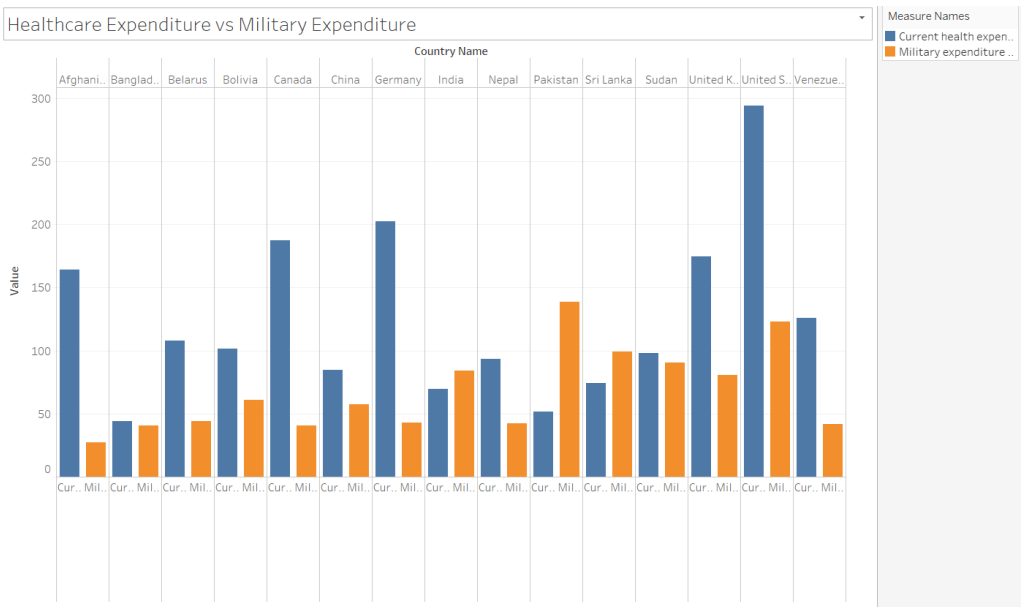
The above graph is an area chart that shows the per capita health expenditure from the year 2001 to 2018. Here the US, Germany and Australia are the top 3 countries in this visualization. India, Nepal and Pakistan are the lowest here which seem to be almost nonexistent in this graph.

We have implemented two algorithms, Random Forest and ARIMA on the processed data and compared their accuracies.

Results

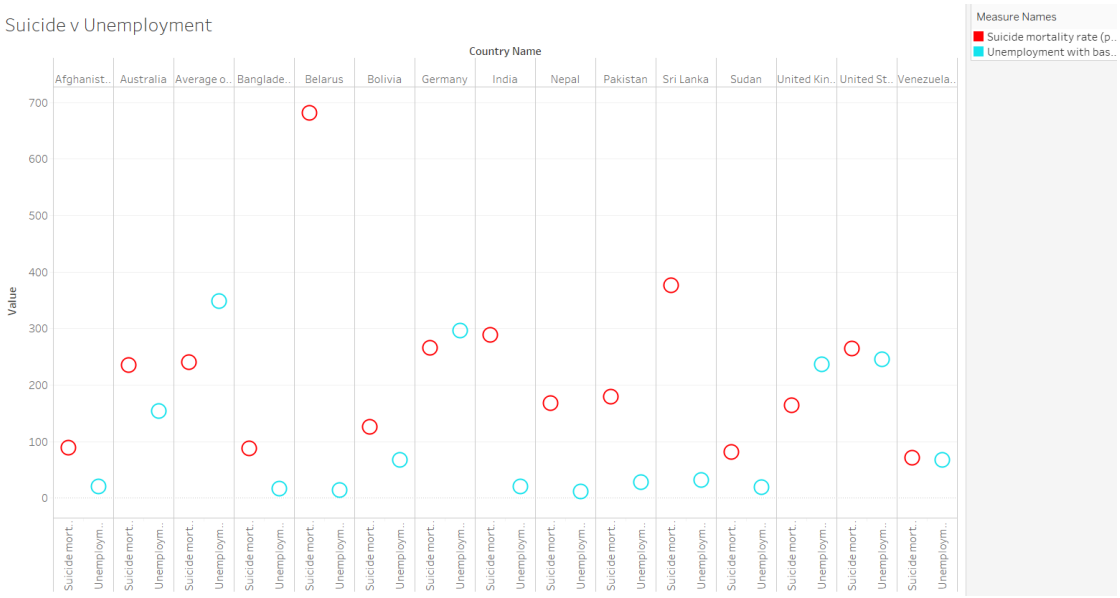
Below are the plots of some data insights and result of trend analysis for world healthcare indicators:

Healthcare vs military expenditure



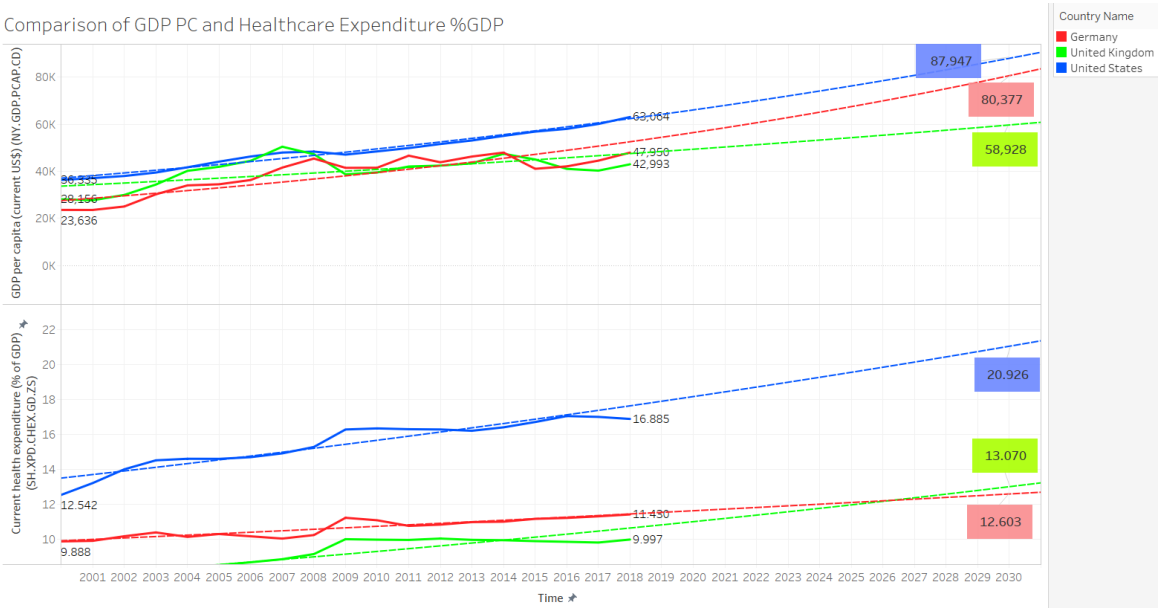
This visualization is a comparison between the country's military expenditure and healthcare expenditure as a percentage of their overall GDP. The United States and Germany spends a significantly higher percentage of their overall GDP on healthcare while Pakistan is the only country among the selected here that spends a greater percentage of their GDP on military when compared to healthcare.

Suicide vs Unemployment



This graph shows a comparison between a countries suicide and unemployment rates. The unemployment described here is the percentage of the population that has got past high school education. Belarus seems to have the highest differential between suicide and unemployment, there doesn't seem to by any firm correlation that we can identify based on this visualization.

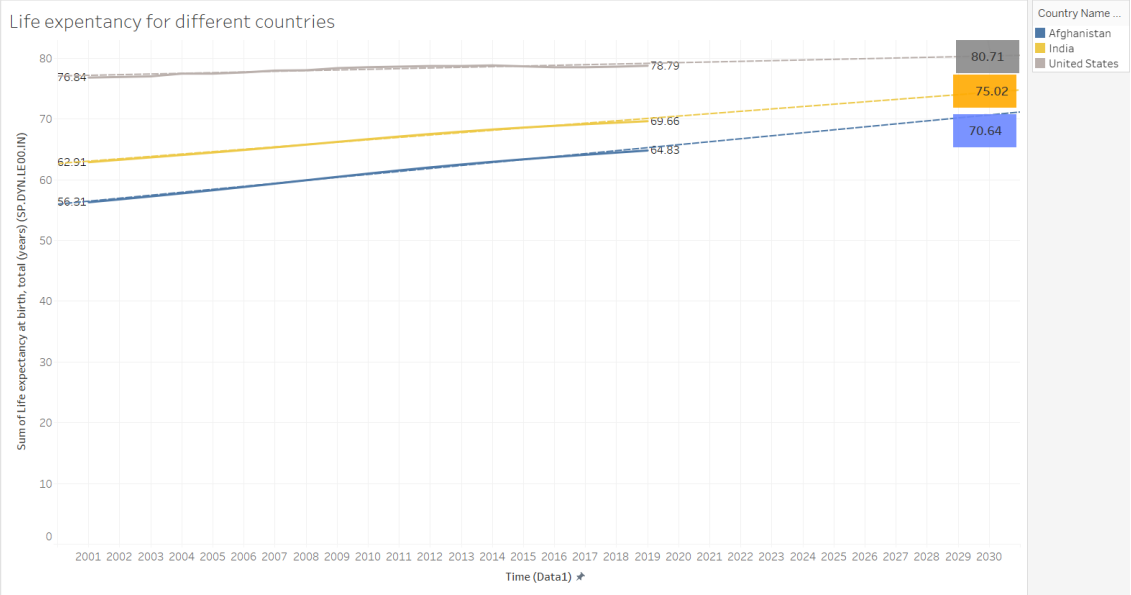
Comparison of GDP and Healthcare expenditure % GDP



The above graph shows a trend prediction between 3 countries Germany, United states and United Kingdom. The GDP and healthcare have been predicted till the year 2030. The United

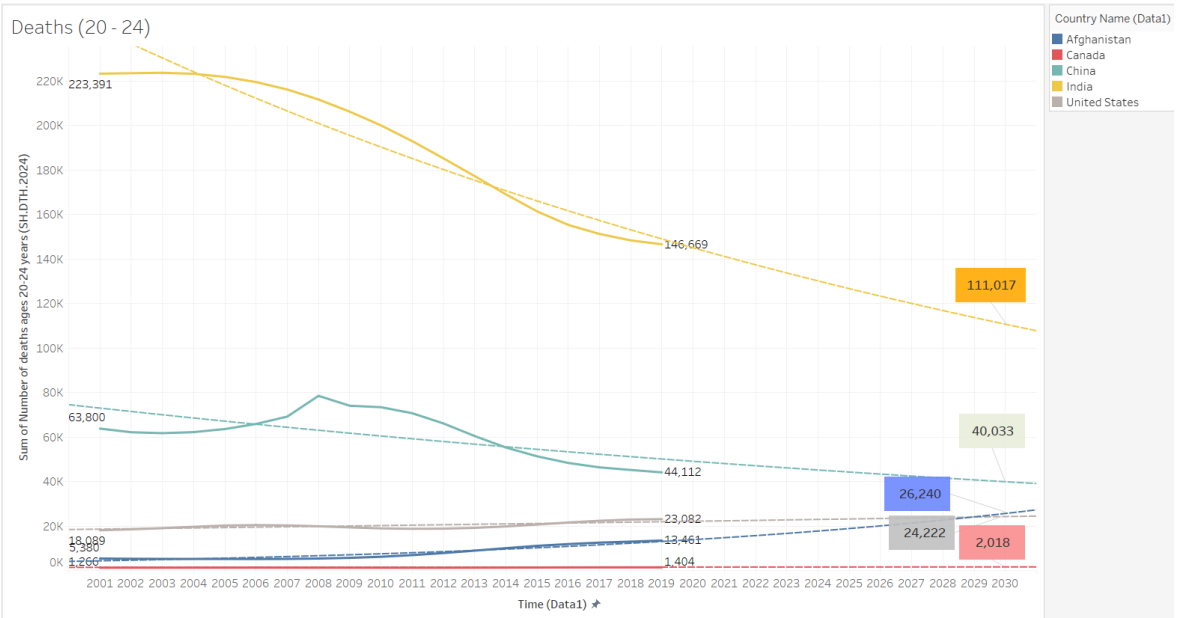
States seems to be significantly higher than the other 2 countries. The UK is currently set to pass Germany in current health expenditure by 2030 based on current upward trend.

Life Expectancy



The United States has been stagnant in terms of life expectancy over the last 20 years while India and Afghanistan have been steadily climbing over the last decade, but based on current growth trend, even by the year 2030 it does not look like it shall reach the level at which the United States currently stands.

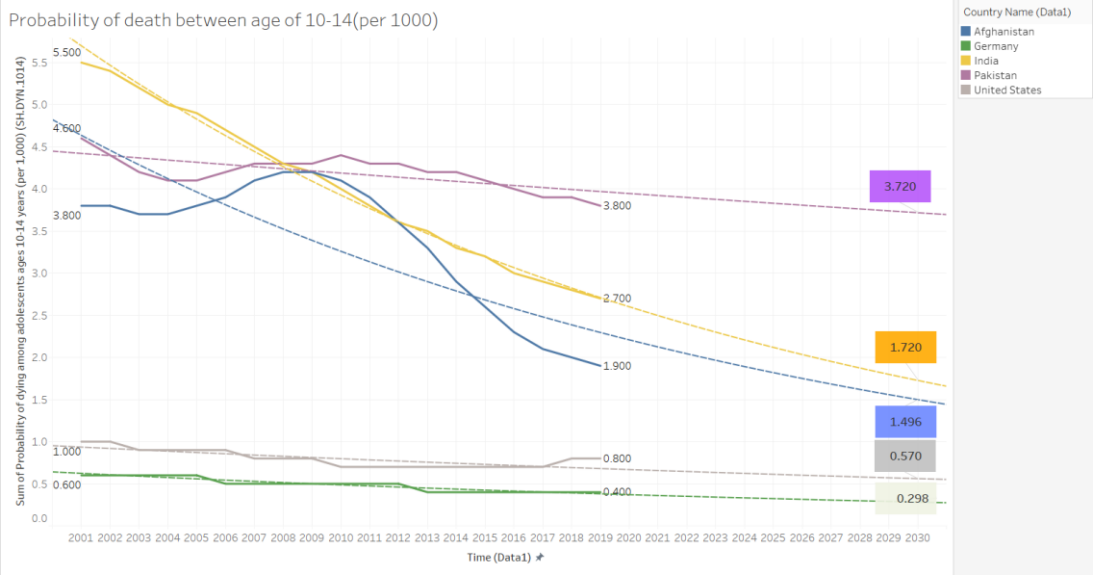
Deaths between the age 20-24



The above visualization shows the average deaths that occur between 5 countries of the population between the ages of 20 to 24. India has seen a dramatic decline over the last 20 years, but the number still stands significantly higher than all the other countries shown

here. The lowest country depicted here is Canada. Afghanistan seems to be the only country here that seems to be in an upward trajectory in this category.

Probability of death between 10-14



This visualization depicts the probability of a child between the age of 10-14 dying per 1000 of the overall population, in the listed countries. Pakistan stands the highest here while Germany is by far the lowest. Afghanistan showed a spike between the years 2007 to 2011 and Pakistan saw a spike from 2003 to 2009. All other listed countries show signs of a steady decline over the last 20 years and this trend continues till 2030.

Model outputs:

```
In [20]: #Creating RF model
         rf = RandomForestRegressor(n_estimators=100, n_jobs=-1, oob_score=True)
         rf.fit(X_train, y_train)

Out[20]: RandomForestRegressor(n_jobs=-1, oob_score=True)

In [21]: # evaluating the models with test data

         val_preds = rf.predict(X_test)
         print(r2_score(y_test, val_preds))
         #  regression score function r2_score

         0.9604676807797856
```

We could achieve 96% accuracy on the data to predict the healthcare expenditure values from Random Forest algorithm. With ARIMA model the accuracy was very low as time series data needs better pre-processing with respect to singularity.

Discussion

Based on our objectives, we have identified trends in healthcare indicators of major countries and projected their trends till the year 2030. We have tried to find out the point of intersection for features of developing countries with that of developed countries. We were successful in identifying for features such as, “Life expectancy”, “Birth rate & Fertility rate”, “Healthcare expenditure % GDP”. But for some features like “Probability of Death between 10-14 years”, “Deaths between age 20-24”, it is difficult to find an intersection point even in the next 15 years.

We have implemented a model that predicts current healthcare expenditure for a specific country based on previous 15 years of healthcare features (total population, community health workers, number of hospital beds, GDP, etc.). Though, we have implemented a model, we were only able to achieve less than 40% of accuracy in time trend analysis.

Conclusion

Based on the visualization plots, we could observe that the healthcare trend is globally increasing for all developed, under-developed and developing countries. For some of the countries like India, Pakistan, Sri Lanka, Afghanistan these trends are very inclined exponentially in the past very years and will continue to increase in the next 10 years. Even though it would be difficult to say that they could achieve the same development as of countries like USA, Germany, Australia, but based on the trends they are surely in the process of getting better. China has had the most development in healthcare indicators including their total GDP and healthcare expenditure in the past 20 years among all the developing countries. The model which we used (ARIMA) to predict the healthcare expenditure based on number of years can be studied further to achieve better results. More data pre-processing and advanced algorithm such as SARIMA can be implemented to enhance the results of the model.

Contributions

Akshay Joshi	Research on related work, Project proposal, EDA on Python, Model building, Random Forest, Final presentation, Project Report.
Akhil Nandakumar	Research on related work, Project proposal, EDA on Tableau/Excel, Tableau data visualization, trend analysis, Final presentation, Project Report.
Vijayalaxmi Rohane	Research on related work, Finding dataset, Project proposal, EDA on Python, Model building, ARIMA, Final presentation, Project Report.
Shrikant Narawane	Research on related work, Project proposal, EDA on Tableau/Excel, Tableau data visualization, trend analysis, Final presentation, Project Report.

References:

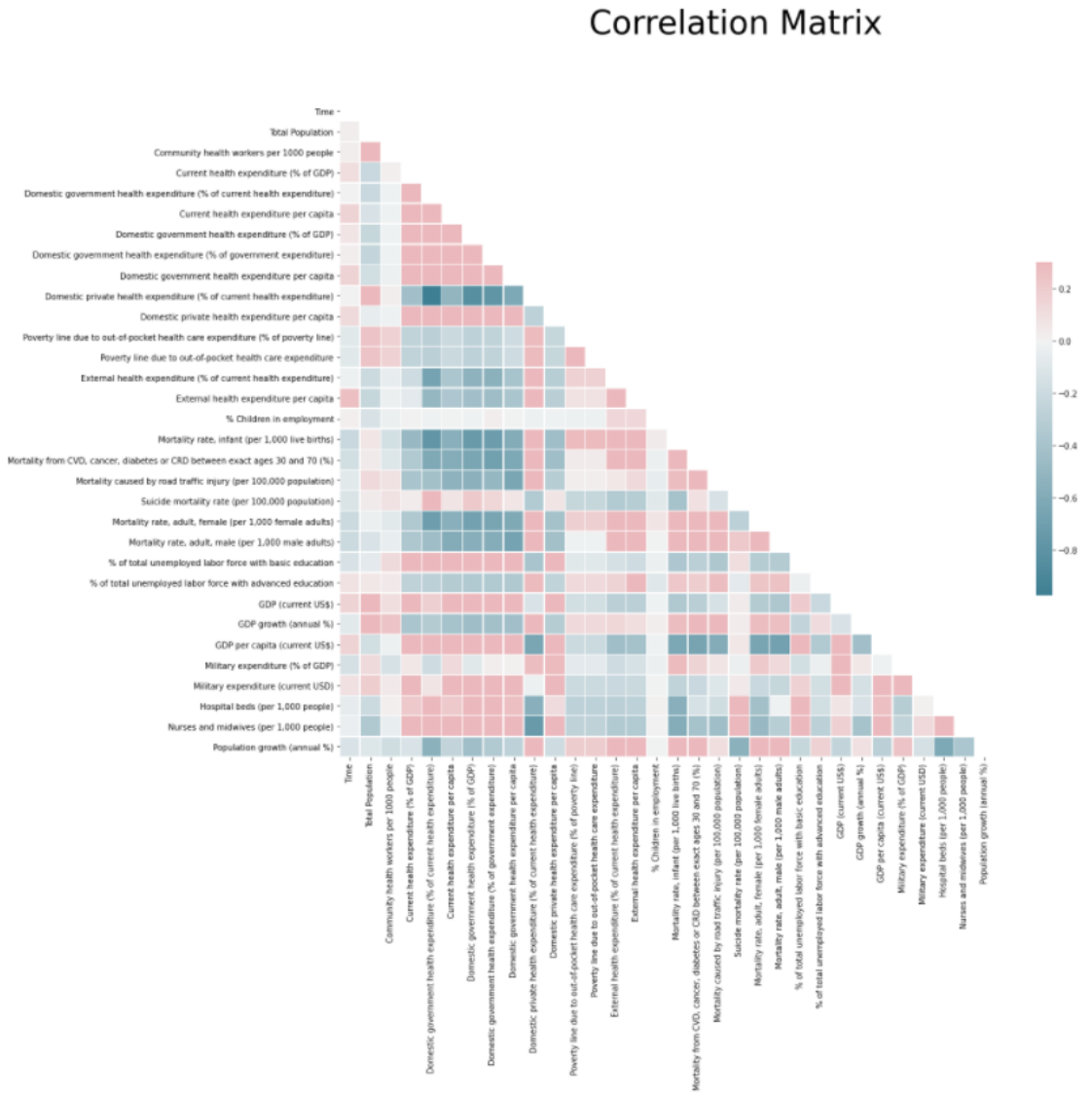
1. https://journals.sagepub.com/doi/abs/10.1111/ijs.12408
2. https://www.koreascience.or.kr/article/JAKO202026061031779.view?orgId=kodisa
3. https://thew.web.unc.edu/wp-content/uploads/sites/5246/2013/09/simon.pdf
4. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf
5. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
6. https://www.kaggle.com/redwankarimsony/time-series-forecasting-with-arima/notebook
7. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
8. https://medium.com/swlh/temperature-forecasting-with-arima-model-in-python-427b2d3bcb53
9. https://seaborn.pydata.org/generated/seaborn.diverging_palette.html
10. https://stackoverflow.com/questions/53822194/python-generate-a-mask-for-the-lower-triangle-of-a-matrix
11. https://machinelearningmastery.com/random-forest-ensemble-in-python/
12. https://stackabuse.com/change-figure-size-in-matplotlib/
13. https://youtu.be/2XGSIlgUBDI
14. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
15. https://towardsdatascience.com/can-machine-learning-be-used-to-forecast-poverty-c7a54bbd6e6c

Appendices

Code:

```
# importing all the required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

import seaborn as sns
import plotly.graph_objs as go#visualization
import plotly.offline as py#visualization

from sklearn.model_selection import train_test_split # for Train-Test data
split
from sklearn.ensemble import RandomForestRegressor # for creating Random
forest model
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import accuracy_score

from sklearn.metrics import mean_absolute_error, r2_score

from datetime import datetime
warnings.filterwarnings('ignore', 'statsmodels.tsa.arima_model.ARMA',

import warnings
warnings.filterwarnings('ignore', 'statsmodels.tsa.arima_model.ARMA',
                        FutureWarning)

# loading the data, wrt sheet name and skipping the footer

data = pd.read_excel ('Data_Extract_From_World_Development_Indicators
(1).xlsx',sheet_name='Data', skipfooter=5)

# checking the rows and columns of data
data.shape

# checking sample data
data.head()


# filtering out data from 2000 , coz lots of missing values before year
2000
data=data[(data['Time'] >=2000) & (data['Time'] <=2018)]


# dropping unneccessary columns
drop_cols=["Time Code",
           "Domestic general government health expenditure per capita, PPP
(current international $) [SH.XPD.GHED.PP.CD]",
           "Domestic private health expenditure per capita, PPP (current
international $) [SH.XPD.PVTD.PP.CD]",
           "External health expenditure per capita, PPP (current
international $) [SH.XPD.EHEX.PP.CD]",
           "Adequacy of unemployment benefits and ALMP (% of total welfare
of beneficiary households) [per_lm_alllm.adq_pop_tot]",
```

```
           "Coverage of unemployment benefits and ALMP (% of population)
[per_lm_alllm.cov_pop_tot]"
          ]

data.drop(drop_cols, axis=1, inplace=True)


# replacing empty values by nan
data = data.replace('..', np.NaN)


# fixing null values with mean values
data=data.fillna(data.mean())


# Calculating correlation matrix
#corr = data.corr()
corr_matrix = data.corr()#.abs()


# plotting heat map on the basis of correlation matrix
mask = np.zeros_like(corr_matrix, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(18, 15))
f.suptitle("Correlation Matrix", fontsize = 40)

# Generate a custom diverging colormap
cmap = sns.diverging_palette(220, 10, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr_matrix, mask = mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

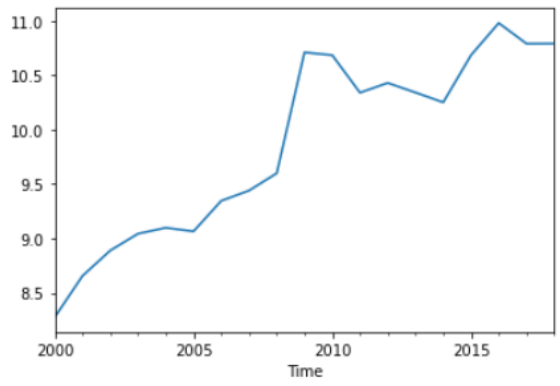



```
# Select upper triangle of correlation matrix
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape),
k=1).astype(np.bool))


# Find features with correlation greater than 0.95
to_drop = [column for column in upper.columns if any(upper[column] > 0.95)]


# dropping columns with high correlations
data.drop(to_drop, axis=1, inplace=True)


# converting the year column from int to datetime
data['Time_dt'] = data.Time.map(lambda x: pd.to_datetime(f'{x}-01-01'))


# saving the cleaned file
data.to_csv('Cleaned_data_final.csv', index=False)


# encoding country column
```

```
from sklearn.preprocessing import LabelEncoder
labelencoder=LabelEncoder()

data["Country Code"]=labelencoder.fit_transform(data["Country Code"])


# Model building

X = data[[  "Country Code"
            ,"Total Population"
            ,"Community health workers per 1000 people"
            ,"Domestic government health expenditure (% of current health
expenditure)"
            ,"Current health expenditure per capita"
            ,"Domestic government health expenditure (% of GDP)"
            ,"Domestic government health expenditure (% of government
expenditure)"
           # ,"Domestic government health expenditure per capita"
            ,"Domestic private health expenditure (% of current health
expenditure)"
            ,"Domestic private health expenditure per capita"
            ,"Poverty line due to out-of-pocket health care expenditure (%
of poverty line)"
            #,"Poverty line due to out-of-pocket health care expenditure"
            ,"External health expenditure (% of current health
expenditure)"
            ,"External health expenditure per capita"
            ,"% Children in employment"
            ,"Mortality rate, infant (per 1,000 live births)"
            ,"Mortality from CVD, cancer, diabetes or CRD between exact
ages 30 and 70 (%)"
            ,"Mortality caused by road traffic injury (per 100,000
population)"
            ,"Suicide mortality rate (per 100,000 population)"
            ,"Mortality rate, adult, female (per 1,000 female adults)"
            ,"Mortality rate, adult, male (per 1,000 male adults)"
            ,"% of total unemployed labor force with basic education"
            ,"% of total unemployed labor force with advanced education"

            ,"GDP (current US$)"
            ,"GDP growth (annual %)"
          #  ,"GDP per capita (current US$)"
            ,"Military expenditure (% of GDP)"
            ,"Military expenditure (current USD)"
            ,"Hospital beds (per 1,000 people)"
            ,"Nurses and midwives (per 1,000 people)"
            ,"Population growth (annual %)"
               ]]

y = data["Current health expenditure (% of GDP)"]

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2,random_state=55)


#Creating RF model
rf = RandomForestRegressor(n_estimators=100, n_jobs=-1, oob_score=True)
rf.fit(X_train, y_train)
```

```
# evaluating the models with test data

val_preds = rf.predict(X_test)
print(r2_score(y_test, val_preds))
#  regression score function r2_score


# to parse date column
def parser(x):
    return datetime.strptime(str(x), '%Y')


# creating countriwise records.
test = pd.read_csv('Cleaned_data_final.csv')
#test.rename(columns={"Country Code": "Country_Code"}, inplace=True)

for i in test["Country Code"].unique():
    temp=pd.DataFrame()
    #print(i)
    temp=test[(test["Country Code"]==i)]
    temp.to_csv('data'+i+'.csv', index=False)


# reading the Canada data
cols = ["Time","Current health expenditure (% of GDP)"]
country_data = pd.read_csv('dataCAN.csv',
                      header=0, parse_dates=[0], index_col=0,
                      usecols=cols,
                      squeeze=True,
                      date_parser=parser)
country_data


# making sure that no null cells are there
country_data.dropna(inplace=True)


# printing the records for Canada
print(country_data.head())
country_data.plot()
plt.show()
```

```
Name: Current health expenditure (% of GDP), dtype: float64
```




```
# Checking For Stationarity  - ADF(Augmented Dickey Fuller Test)
```

```
from statsmodels.tsa.stattools import adfuller
def adf_test(dataset):
    dftest = adfuller(dataset, autolag = 'AIC')
    print("1. ADF : ",dftest[0])
    print("2. P-Value : ", dftest[1])
    print("3. Num Of Lags : ", dftest[2])
    print("4. Num Of Observations Used For ADF Regression:",
dftest[3])
    print("5. Critical Values :")
    for key, val in dftest[4].items():
        print("\t",key, ": ", val)


adf_test(country_data)


# importing arima library
from statsmodels.tsa.arima.model import ARIMA
from sklearn.metrics import mean_squared_error


X = country_data.values
size = int(len(X) * 0.66)


train, test = X[0:size], X[size:len(X)]
history = [x for x in train]
predictions = []


for t in range(len(test)):
    model = ARIMA(history, order=(3,1,0))
    model_fit = model.fit()
    output = model_fit.forecast()
    yhat = output[0]
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))


# Checking the error rate of
import math
error = mean_squared_error(test, predictions)
print('Test Root Mean Squared Error: %.3f' % math.sqrt(error))
```