Statistical Research Report on Diabetes

-By Chakshu Grover

(1918303) BTech CS 5Th semester

Table Of Content

“The condition of uncontrollable sugar in the bloodstream, by the mean of unhealthy diet or physical inability to counteract the carbohydrates in the body, causing viscous blood is the state of Diabetes.”

By the word of W.H.O., Diabetes is a chronic disease occurring when there is the inability of our gland called pancreas in creation of insulin, which regulates the blood sugar level.

According to the WHO:

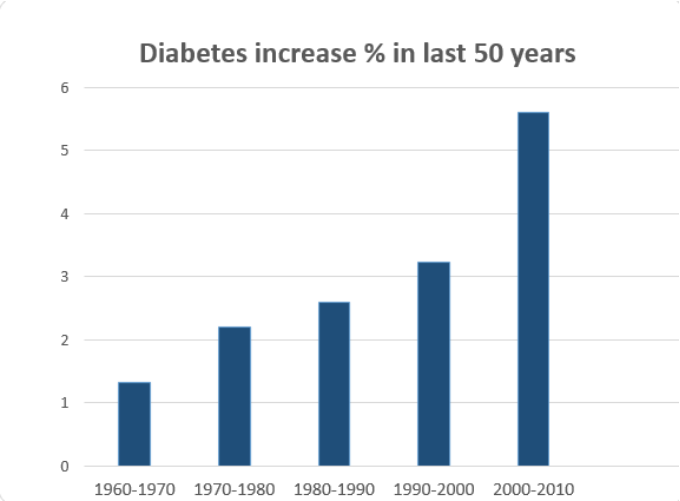
- The Count of Diabetic Patient rose from 108M to 422M people across the world between the time stamp of 1980 and 2014.
- The Scientists saw 5% increase in premature mortality caused by diabetes.

Death count reached 1.5M people directly from Diabetes

Over last 50 years, the world saw a complete change in the technology, food, culture, lifestyle and habits. A lot of these have the direct correlation with the increase in diabetic rate throughout the world. Lets discuss some of the reasons.

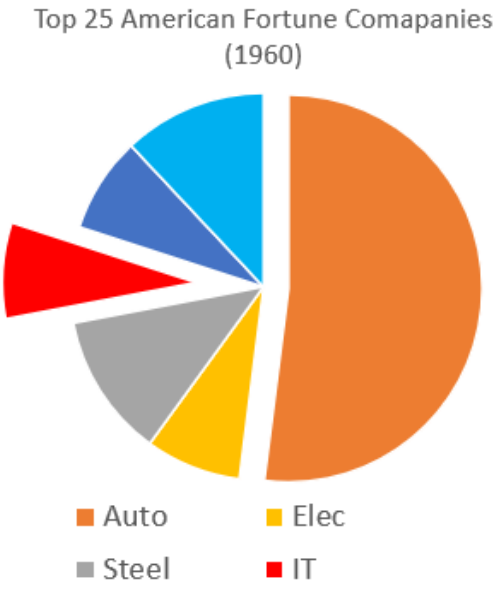
As per the data by WHO, the diabetic patient percentage in USA in 1960 were 1.31% of the total population, whereas this

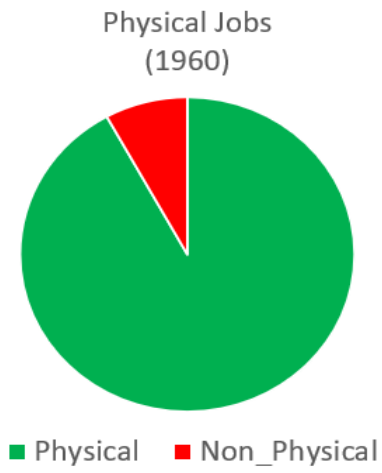
number increased to 5.60% which is approx. 4.25x times of the initial census.

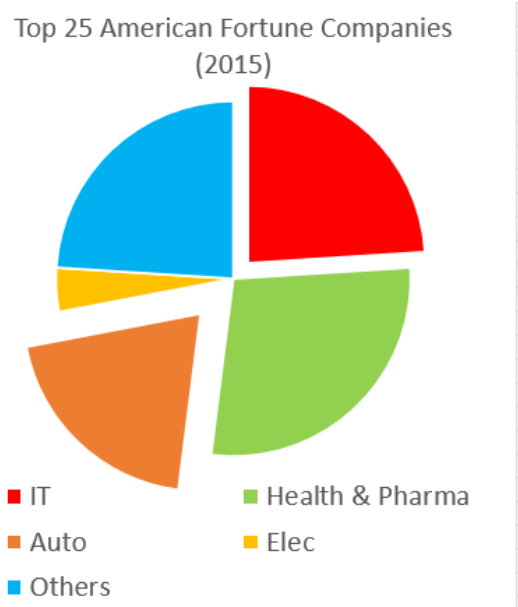


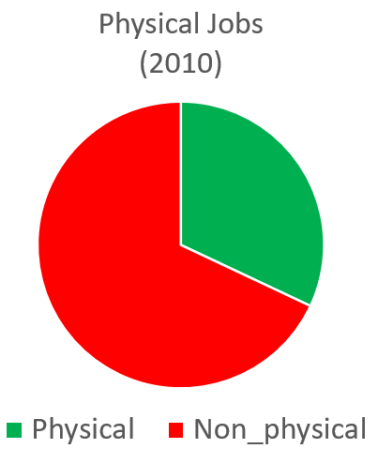
Physical Jobs

Due to the industrial shift in the world, we can see a continuous decline in the physical jobs over last 50 years. Comparing the top 25 fortune companies in 1960 and 2010, these were the industries majoring the market:





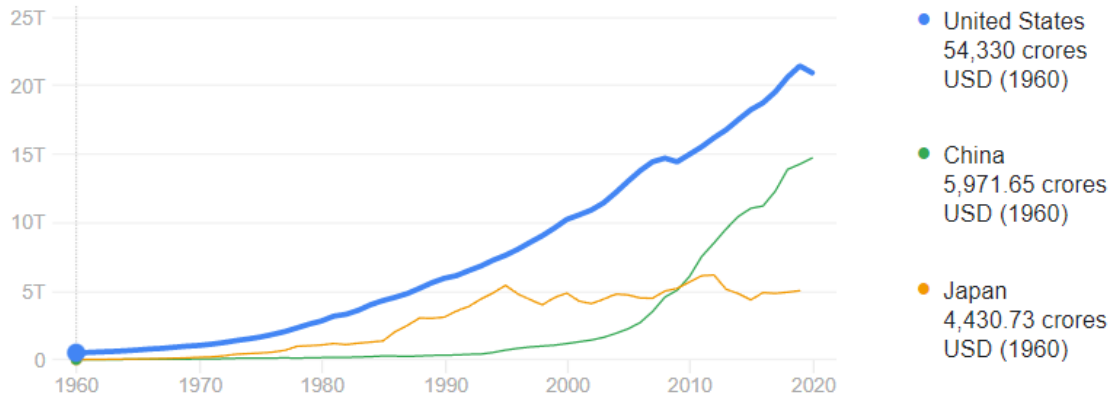


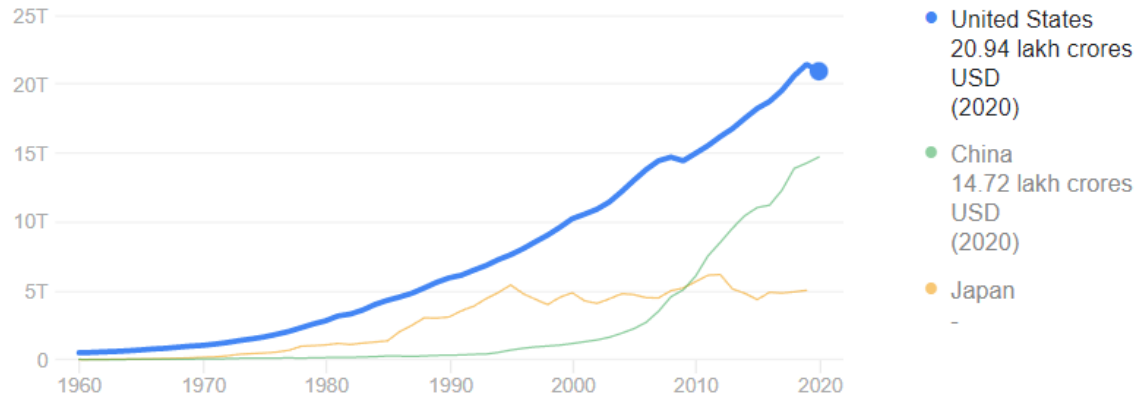


As we can compare the 50-year data, we can conclude there is a positive relation between the rise of no-physical jobs and diabetes.

GDP increase

The world saw an exponential rate of increase in the GDP of the top nation throughout the world. As per World Bank, 1960, the GDP of USA was 543 billion $, and with rising power and economy, the graph saw the exponential growth 210K Billion $ in just 60 years, i.e, 2020.





This increased the living standards, market values, and introduced the modern living which a lot of countries still are far away from. Such increase in half of decade came with technology, telecommunication, business, and modernity in the western world.

Dataset

The Dataset used for the analysis Is the PIMA diabetes dataset, from the National Institute of Diabetes and Digestive and Kidney Diseases. This preprocessed data include recorded dataset and have the primary objective to diagnostically predict the probability of the person's chance of finding positive to Diabetes type 1, based on the certain features (basically 8 features) like:

- pregnancy record,
- glucose concentration,
- Blood Pressure level,
- Skin thickness in mm,
- insulin level in the body,

- body mass index,
- Age, and
- diabetes pedigree function.

The data is labelled with the outcome 0 and 1, depicting negative and positive diabetic status respectively.

The dataset is right skewed, resulting more of lesser data point before mean than more than mean, providing the insights that "*Individuals are more likely to be negative to diabetes*".

Model:

diabetes Prediction system

In [80]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn import svm
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score as accScore, confusion_matrix as conMat
%matplotlib inline
```

In [81]:
```
dataset= pd.read_csv("diabetes.csv")
dataset
```

Out[81]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

In [82]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

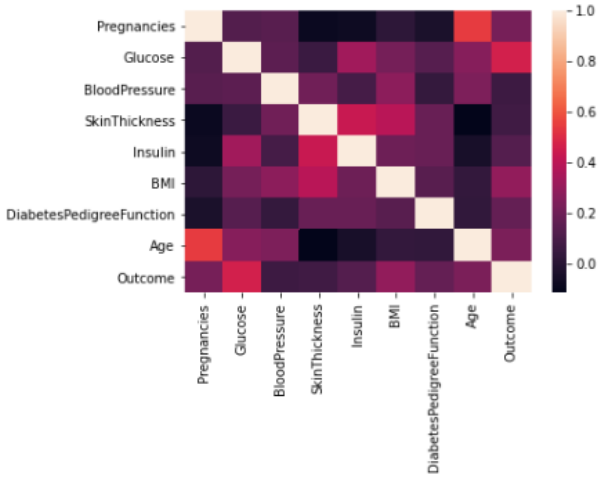
checking for null values

In [83]:
```
dataset.corr()
```

Out[83]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647
DiabetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844

In [84]:
```
sns.heatmap(dataset.corr())
plt.show()
```



feature selection and data preparation

In [93]:
```
x_data= dataset.drop(columns=['Outcome'])
y_data=dataset['Outcome']

x_train, x_test, y_train, y_test = train_test_split(x_data, y_data,
                                                    test_size= 0.15,
                                                    random_state= 100)
```

Model Training

In [94]:
```
Logmodel=LogisticRegression()
Logmodel.fit(x_train, y_train)
```

```
C:\Users\Chakshu\anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarnin
g: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[94]:
```
LogisticRegression()
```

In [95]:
```
svmmodel=svm.SVC(kernel ='linear')
svmmodel.fit(x_train, y_train)
```

Out[95]:
```
SVC(kernel='linear')
```

prediction and comparison

In [96]:
```
y_logres = Logmodel.predict(x_test)
```

In [97]:
```
print("LOGISTIC MODEL PREDICTION:\n")
print("Intercept: ",model.intercept_)
print("Accuracy Score: ",accScore(y_logres, y_test))
print("\nConfusion Matix: \n",conMat(y_logres, y_test))
```

```
LOGISTIC MODEL PREDICTION:

Intercept:  [-8.03932898]
Accuracy Score:  0.7586206896551724

Confusion Matix:
 [[65 18]
 [10 23]]
```

In [98]:
```
y_svmres = svmmodel.predict(x_test)
```

In [99]:
```
print("LOGISTIC MODEL PREDICTION:\n")
print("Accuracy Score: ",accScore(y_svmres, y_test))
print("\nConfusion Matix: \n",conMat(y_svmres, y_test))
```

```
LOGISTIC MODEL PREDICTION:

Accuracy Score:  0.7844827586206896

Confusion Matix:
 [[66 16]
 [ 9 25]]
```

In [100]:
```
if(accScore(y_logres, y_test) > accScore(y_svmres, y_test)):
    print("the Logistic Regression model predicted better than SVM Model")
else:
    print("the SVM model predicted better than logistic Model")
```

```
the SVM model predicted better than logistic Model
```

References:

- https://www.who.int/health-topics/diabetes
- https://www.kaggle.com/uciml/pima-indians-diabetes-database
- https://fortune.com/fortune500/
- https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes#:~:text=Diabetes%20is%20a%20disease%20that,to%20be%20used%20for%20energy.