Solving a Healthcare Problem for Improved Service Delivery

Problem Description:

The facility problem is arising in United States of America to meet certain type of service example (maternity). The facility area zip codes in which the problems are to be solved are 98007, 98290, 98065, 98801 and 98104. Since the problem states that there is a shortage of facility needs in the above described zip codes therefore, people have to travel to the another facility which is too much far from their respective area in order to solve this problem we have to determine the population etc. on that basis we could determine maternity population (or the number of births) in the particular area.

Assumptions:

For the problem solution predict the maternity population of the respective given zip codes so that government could figure out how much facility is required in the given respective zip codes.

The following steps are followed to calculate maternity in the respective zip codes.

1) Making of Data from online websites.

2) Data set cleaned.

3) Data set analyzed.

4) Data set visualized.

5) Training of Data.

6) A to B mapping or Prediction.

All the points are described below

First Step:

Check the data required for solving the problem is available online that is maternity population data for a given zip code is available online.

Second Step:

Gathered Data from sites and build the data set, cleaned and analyzed.

Current Population in States:

```
In [129]: new_data=cdf.head(45)
          df = new_data[['Birth_n','Population_2018']]
          df
```

Out[129]:

	Birth_n	Population_2018
0	58941.0	4887871.0
1	10445.0	737438.0
2	37520.0	3013825.0
3	64382.0	5695564.0
4	35221.0	3572665.0
5	10855.0	967171.0
6	9560.0	702455.0
7	38430.0	3156145.0
8	36519.0	2911505.0
9	54752.0	4468402.0
10	61018.0	4659978.0
11	12298.0	1338404.0
12	71641.0	6042718.0
13	70702.0	6902149.0
14	111426.0	9998915.0

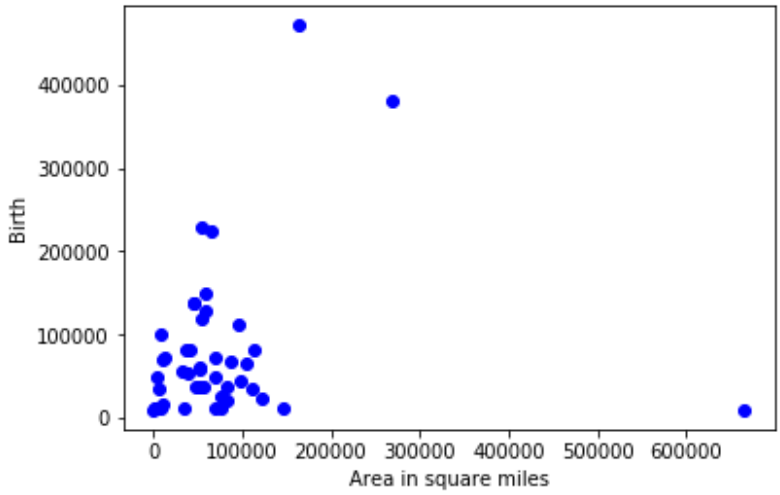
Third Step:

To calculate the maternity population for a zip code checked which features are required and maternity population is gathered through website (link is given at end).

Area is not a feature.

Data Analyze:

```
In [209]: #plt.plot(df.Population_2018, df.Birth_n,  color='blue')
          plt.scatter(df.sq_mi, df.Birth,  color='blue')
          plt.xlabel("Area in square miles")
          plt.ylabel("Birth")
          plt.show()
```



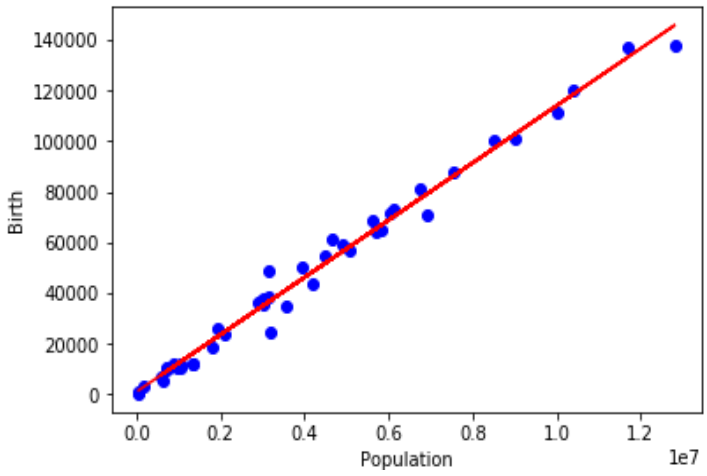
Fourth Step:

After checking every feature to have an effect on maternity population .The input which effects the most maternity population is the population of that area. The data pattern is shown below

Using Data Visualization:

```
In [100]: plt.scatter(df.Population_2018, df.Birth_n,  color='blue')
          plt.plot(train_x, regr.coef_[0][0]*train_x + regr.intercept_[0], '-r')
          plt.xlabel("Population")
          plt.ylabel("Birth")
```

```
Out[100]: Text(0, 0.5, 'Birth')
```



Fifth Step:

1) Training the model required data of population as input and Maternity Population as output.
2) Since the number of births data for zip codes is not available online so Linear Regression model is used for training of data of territory/States of Us. Using the features set and find that population is linearly related with maternity population.

```
In [99]: import numpy as np
         from sklearn import linear_model
         regr = linear_model.LinearRegression()
         train_x = np.asanyarray(df[['Population_2018']])
         train_y = np.asanyarray(df[['Birth_n']])
         regr.fit (train_x, train_y)
         # The coefficients
         print ('Coefficients: ', regr.coef_)
         print ('Intercept: ',regr.intercept_)

         Coefficients:  [[0.01130373]]
         Intercept:  [997.15056806]
```

Sixth Step:

Using year as input and population as output another model is trained for zip code to predict population of future and then by using the future predicted population putting the value into the previous train model it could easily calculate the maternity population or the number of births of present as well as future for that zip code.

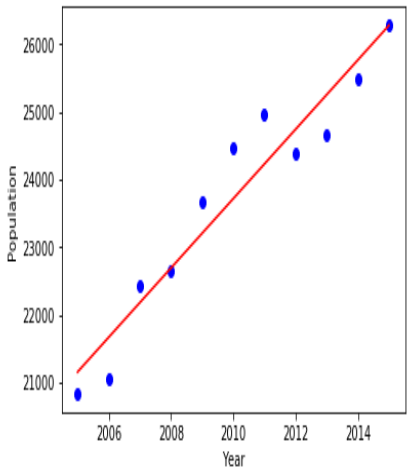
Data Frame for Zip Code (98007):

Out[133]:

	Year	Population	Maternity_Population
0	2005	20840.00000	1232
1	2006	21059.00000	1235
2	2007	22436.00000	1250
3	2008	22649.00000	1253
4	2009	23674.00000	1264
5	2010	24467.00000	1273
6	2011	24977.00000	1279
7	2012	24390.00000	1272
8	2013	24650.00000	1275
9	2014	25490.00000	1285
10	2015	26280.00000	1294
11	2016	26792.96360	1300
12	2017	27305.24500	1305
13	2018	27817.52727	1311
14	2019	28329.80900	1317
15	2020	28842.09000	1323

```
In [58]: plt.scatter(df_98007.Year, df_98007.Population,  color='blue')
         plt.plot(train_x, regr.coef_[0][0]*train_x + regr.intercept_[0], '-r')
         plt.xlabel("Year")
         plt.ylabel("Population")

Out[58]: Text(0, 0.5, 'Population')
```



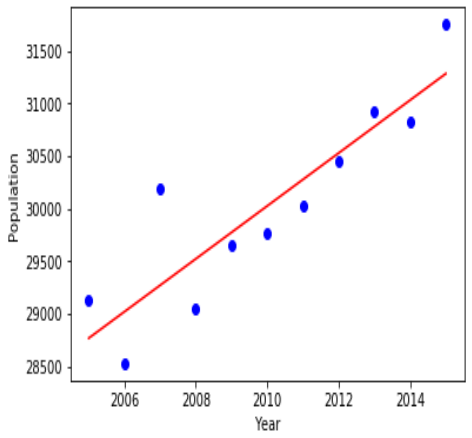
Data Frame for zip code (98290):

Out[136]:

	Year	Population	Maternity_Population
0	2005	29125	1326
1	2006	28527	1319
2	2007	30189	1338
3	2008	29049	1325
4	2009	29646	1332
5	2010	29766	1333
6	2011	30023	1336
7	2012	30450	1341
8	2013	30920	1346
9	2014	30820	1345
10	2015	31750	1356
11	2016	31533	1353
12	2017	31784	1356
13	2018	32036	1359
14	2019	32287	1362
15	2020	32539	1364

```
plt.scatter(df_98290.Year, df_98290.Population,  color='blue')
plt.plot(train_x, regr.coef_[0][0]*train_x + regr.intercept_[0], '-r')
plt.xlabel("Year")
plt.ylabel("Population")
```

```
Text(0, 0.5, 'Population')
```

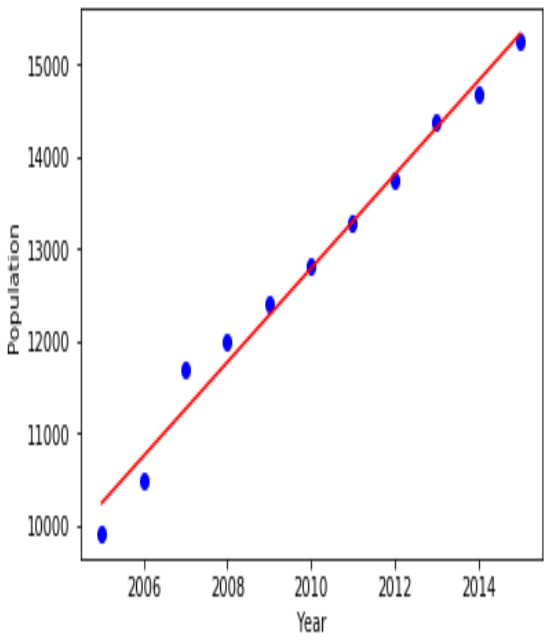


Data Frame for Zip code (98065):

Out[120]:

	Year	Population	Maternity_Population
0	2005	9922.00000	1109
1	2006	10495.00000	1115
2	2007	11692.00000	1129
3	2008	11992.00000	1132
4	2009	12398.00000	1137
5	2010	12825.00000	1142
6	2011	13282.00000	1147
7	2012	13750.00000	1152
8	2013	14380.00000	1159
9	2014	14680.00000	1163
10	2015	15250.00000	1169
11	2016	15833.85450	1176
12	2017	16341.52727	1181
13	2018	16849.20000	1187
14	2019	17356.87270	1193
15	2020	17864.54540	1199

Out[6]: Text(0, 0.5, 'Population')

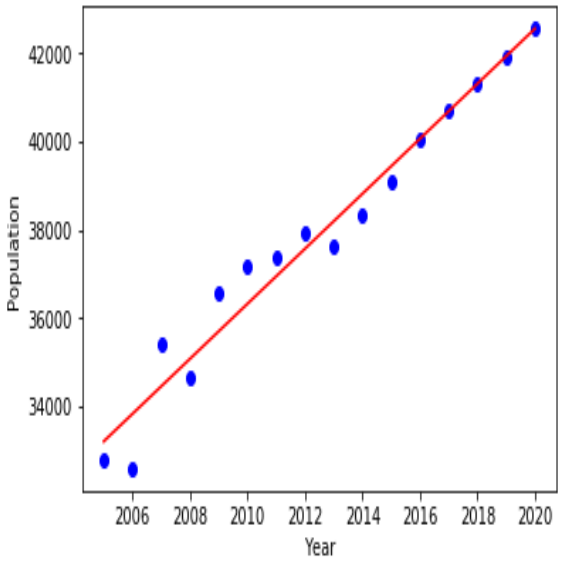


Data Frame for Zip Code (98801):

Out[125]:

	Year	Population	Maternity_Population
0	2005	32794.0000	1367
1	2006	32601.0000	1365
2	2007	35400.0000	1397
3	2008	34685.0000	1389
4	2009	36575.0000	1410
5	2010	37153.0000	1417
6	2011	37397.0000	1419
7	2012	37920.0000	1425
8	2013	37640.0000	1422
9	2014	38320.0000	1430
10	2015	39110.0000	1439
11	2016	40061.4363	1449
12	2017	40683.8270	1457
13	2018	41306.3090	1464
14	2019	41928.7454	1471
15	2020	42551.1818	1478

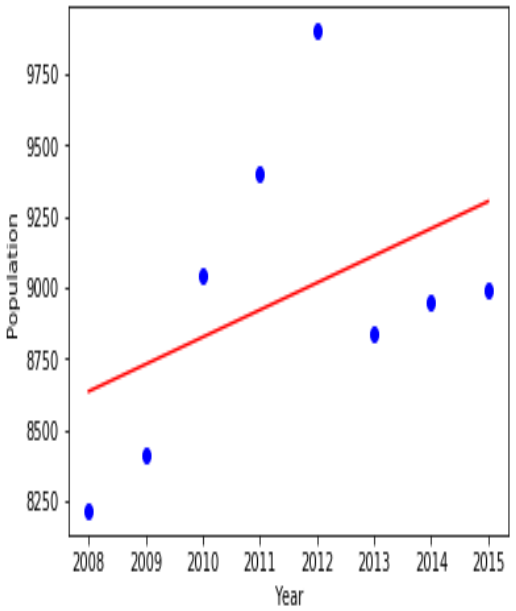
Out[15]: Text(0, 0.5, 'Population')



Data Frame for Zip Code (98104):

Out[128]:

	Year	Population	Maternity_Population
0	2008	8217.0000	1090
1	2009	8412.0000	1092
2	2010	9041.0000	1099
3	2011	9396.0000	1103
4	2012	9900.0000	1109
5	2013	8840.0000	1097
6	2014	8950.0000	1098
7	2015	8990.0000	1098
8	2016	9396.9285	1103
9	2017	9492.1904	1104
10	2018	9587.4523	1105
11	2019	9682.7140	1106
12	2020	9777.9760	1107



Libraries:

1) Numpy

2) Matplotlib

3) Scikit Learn

4) Pandas

References:

Maternity population of each state is gathered through this link:

https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_01-508.pdf

All the possible reasons which effect maternity:

https://www.cdc.gov/reproductivehealth/emergency/pdfs/PregnacyEstimatoBrochure508.pdf

Population data gathered for zip code 98007 using website:

https://www.unitedstateszipcodes.org/98007/

Population data gathered for zip code 98290 using website:

https://www.unitedstateszipcodes.org/98290/

Population data gathered for zip code 98065 using website:

https://www.unitedstateszipcodes.org/98065/

Population data gathered for zip code 98801 using website:

https://www.unitedstateszipcodes.org/98801/

Population data gathered for zip code 98104 using website:

https://www.unitedstateszipcodes.org/98104/

Territory/States area and population is gathered through website:

https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population

https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_area

Zip codes population and nearby facility id (98007):

https://www.zipdatamaps.com/98007

Zip codes population and nearby facility id (98290):

https://www.zipdatamaps.com/98290

Zip codes population and nearby facility id (98065):

https://www.zipdatamaps.com/98065

Zip codes population and nearby facility id (98801):

https://www.zipdatamaps.com/98801

Zip codes population and nearby facility id (98104):

https://www.zipdatamaps.com/98104

Current Population Counts:

For the current population of the respective area zip codes the website is used for this purpose. Since the population was in the graph form so by manually entering data from the website and storing in the data frame using pandas.

The data could be extracted using web scraping if the data is acquired through the website but in this case data is stored in excel file so acquiring data from excel using pandas library.

```
In [205]: import pandas as pd
          data_f=pd.read_excel(r"C:\Users\hp\Downloads\dojo.xlsx")
          data_f
```

For data acquiring request or soap could be used. The websites from which the data could be acquired publically are Kaggle and Github.

Future Population Counts:

The data is acquired through websites. Then Features are selected means which column to be as input. Then Output is selected from the data set which is called target variable. Machine Learning models are used to predict the future population as the population is a continuous value so supervised learning algorithm regression is used. In Supervised Learning there are two types of algorithm regression and classification. Regression is used for predicting continuous values which is in this case and classification is used for predicting discrete labels. The linear regression is used because population is linearly related with maternity population. For training the model the dataset is required and coefficient is determined. That coefficients determine the line of equation to be fit. If job investments are planned in an area there would be an increase in population and migration rate in an area by considering the change in migration rate we could predict the future population however there are many sites available through which data could be gathered and on kaggle there are many public and free datasets available.

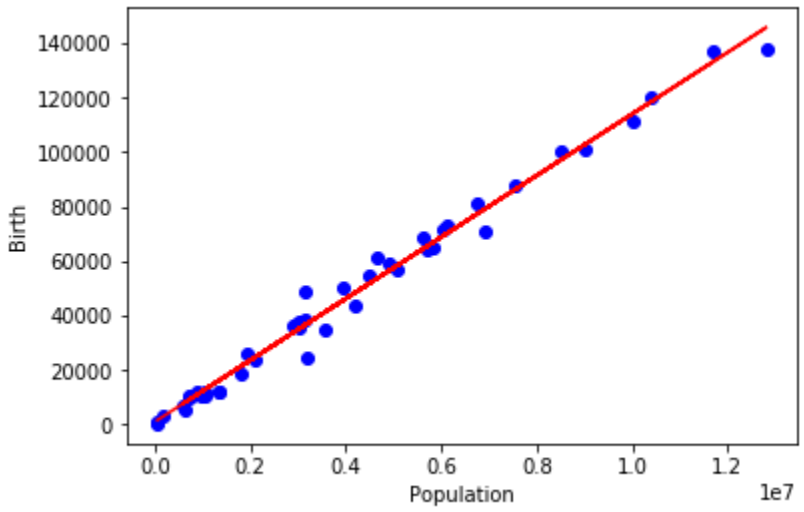
The code is shown at the next page.

In [99]:
```
import numpy as np
from sklearn import linear_model
regr = linear_model.LinearRegression()
train_x = np.asanyarray(df[['Population_2018']])
train_y = np.asanyarray(df[['Birth_n']])
regr.fit (train_x, train_y)
# The coefficients
print ('Coefficients: ', regr.coef_)
print ('Intercept: ',regr.intercept_)
```

```
Coefficients:  [[0.01130373]]
Intercept:  [997.15056806]
```

In [100]:
```
plt.scatter(df.Population_2018, df.Birth_n,  color='blue')
plt.plot(train_x, regr.coef_[0][0]*train_x + regr.intercept_[0], '-r')
plt.xlabel("Population")
plt.ylabel("Birth")
```

Out[100]: Text(0, 0.5, 'Birth')



Area ID	Nearby Facility ID	Area Population	Maternity Population within Area (2015)	Predicted Future Maternity Population (2016)	Current Capacity For Maternity In Facility
98007	98033,98008,98006,98005,98052	26280	1294	1300	200
98290	98252,98208,98205,98296,98258,	31750	1356	1353	36000
98065	98024,98027,98045	15250	1169	1176	12000
98801	98815,98843,98847,98828,98822, 98926,98802,98826	39110	1439	1449	0
98104	98144,98109,98134,98102,98122, 98121,98101	8990	1098	1103	58000

This table is generated from the above described information.From the above table it is cleared that in which areas facilities are needed to be increased.