1. Introduction

1.1 Background

In the last decades, healthcare has been one of the major field of studies. There are several reasons. One of them is the increasing healthcare expenditure throughout the world. It is true that people can reach out the healthcare utilization easily if we compare it couples of decades ago. This and other reason such as inflation, technology improvement causes steep rise in healthcare expenditures.

1.2 Business Problem

Countries like Canada, having public healthcare, need to do more research on healthcare demand and utilization in order to allocate limited resources among their citizens.

1.3 Interest

The objective of this project to define healthcare demand in Canada in 2005. The target audience of this project is healthcare providers in Canada. It can be private or public healthcare providers.

The following questions are raised in this project:

1. The average hospital-stay and physician visit in Canada.
2. Defining question 1 problems by using age groups (age 30 and below, 30-64, 65 and more)
3. Do walking as exercise, age, income affect the demand for healthcare?
4. Defining cancer by using age, income and BMI (Body Mass Index)

Machine learning techniques and multiple linear regression model will be used to conduct research.

2. Data acquisition and cleaning

2.1 Data source

The data is Canadian Health Community Survey: cycle 3 - 2005. The data is taken from University of Saskatchewan library through ODESI.

2.2 Data cleaning and feature selection

There are 132221 observations and 1284 variables. In order to run my project, I chose 11 variables which are related to healthcare utilization and healthcare demand.

The row variable description is as following:

DHHEGAE - age

1 = 12-14,
2 = 15-17,
3 = 18-19,
4 = 20-24,
5 = 25-29,
6 = 30-34,
7 = 35-39,
8 = 40-44,
9 = 45-49,
10 = 50-54,
11 = 55-59,
12 = 60-64,
13 = 65-69,
14 = 70-74,
15 = 75-79,
16 = 80 and more

For example, 1 stands for people's age range from 12 to 14. When we replaced categorical values, we took average of 12 to 14 as 13. All 1s were replaced by 13, 2s by 16 and so on.

DHHE_SEX - sex

1 - Male

2 - Female

In sex variable, 2 replaced by 0, while 1 remains the same.

CCCE_131 - cancer

1 - yes

2 - no

6 - not applicable

7 - do not know

8 - refusal

9 - not stated

INCEGPER - Total household income

1 =no income,

2 = less than 15000,

3 = 15000-29999,

4 = 30000-49999,

5 = 50000-79999,

6 = 80000 or more.

96 = not applicable

97 = do not know

98 = refusal

99 = not stated

HWTEGHTM - height in metres

1.27 - 1.257 to 1.282

1.295 - 1.283 to 1.307

...

...

Averages until 2.134

But 9.999 - not stated

HWTEGWTK - weight in kgs

999.99 - not stated

HWTEGBMI - BMI (Body Mass Index)

999.99 - Not Stated - Pregnant women

HCUE_01 - Overnight patient

1 = yes

2 = no

6 = not applicable

7 = do not know

8 = refusal

9 = not stated

If yes in HCUE_01

HCUEG01A - Number of nights as a patient

1 = 1

2 = 2

3 = 3

...

30 = 30

31 = 31 and more

96 = not applicable (1 - Check Notes below)

97 = do not know

98 = refusal

99 = not stated (2 - Check notes below)

HCUEG02A - Physician visit or number of consultations

1 = 1

2 = 2

3 = 3

...

30 = 30

31 = 31 and more

96 = not applicable

97 = do not know

98 = refusal

99 = not stated

PACE_1A - Last three-month walking

1 = yes

2 = no

6 = not applicable

7 = do not know

8 = refusal

9 = not stated (3 check Notes below)

Notes:

(1) They said 'no' in HCUE_01. It implies that we need to replace them with 0.

(2) From description of variables guide, in HCUEG01A 99 = 1-night stay, this is because these people have not been questioned in HCUE_01. There are couple of previous variables which also define hospital stay or number of nights as a patient. Thus, it is useless to ask again this question

(3) People did interview by proxy.

Depending on survey questions, the variable values such as 6, 7, 8, 9 or 96, 97, 98, 99 were replaced by null or zero. By considering number of null values are not substantial in comparison to whole data, we can drop them. There is also alternate way like replacing them with the mean of the variables, however in my analysis, some of the variables are categorical variables. Thus, it doesn't make any sense to replace them with continuous numbers.

3. Exploratory Data Analysis

Question 1 - Average hospital stay and physician visit

Annual average hospital stay is 0.59 day and physician visit 3.2 times. If we convert 0.59 days to hours, it equals to 14 approximately. It implies that average Canadian people spend 14 hours in hospital as an overnight patient and they visit or get a consultations 3 times per person every year.

Question 2 - Average hospital stay, physician visit and having cancer by age groups.

In order to solve question 2, we need to create a new categorical variable which possess only three numbers 1, 2 and 3. Here, 1 stands for ages under 30, 2 for 30 - 64, and 3 for 65 and more. By using Python grouping function we can get the following results.

Table 1 - Defining the variables by age groups

	Hospital Stay	**Physician visit**	**Cancer**
Under 30	0.35	2.8	0.002
30 - 64	0.46	3	0.01
65 and more	1.09	3.86	0.04

According to *Table 1*, senior people spend approximately 26 (24 hours x 1.09) hours as an overnight patient. People who are ages between 30 and 64 spend 15 less hours then senior ones. Under 30 years old people just spend 8 hours annually in hospitals. As we see from the table, physician visits are increasing by age groups increase. The cancer also arises by age groups, so if 2 out of 1000 people have a cancer under 30 years old group, it is 10 for 30 - 64 years old people and 40 for senior people.

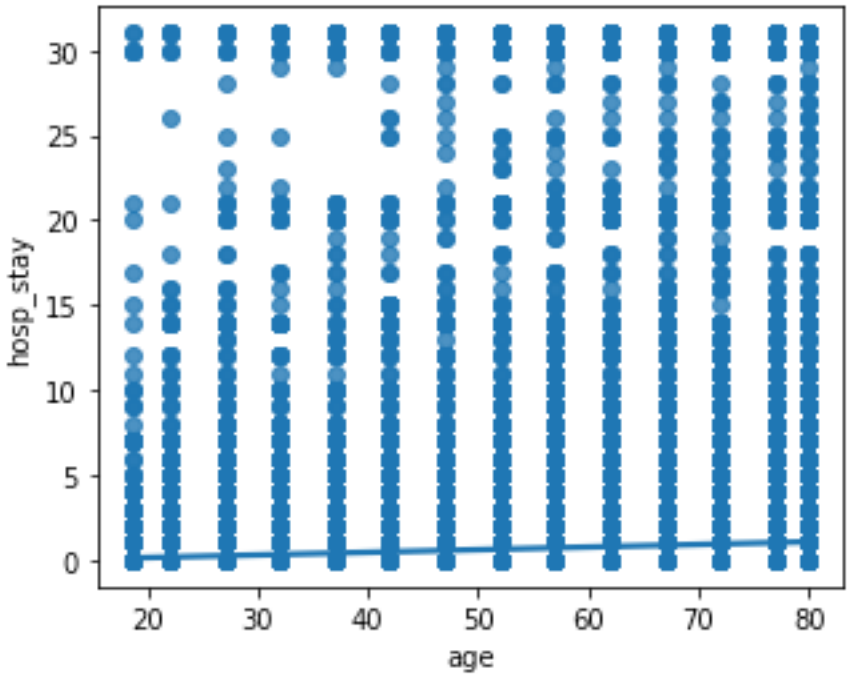
Question 3 - Demand for healthcare

Let's assume demand for healthcare is hospital stay as an overnight patient. The following formula is going to be multiple linear regression as below:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

- Y is dependent variable – hospital stay
- β_0 constant, x_1 - age, x_2 - income, x_3 - walking,
- β_1, β_2, β_3 - are coefficients
- u - disturbance

First, we need to use Pearson correlation method to find out whether there is positive or negative correlation. *Table 2* gives information about Pearson correlation.



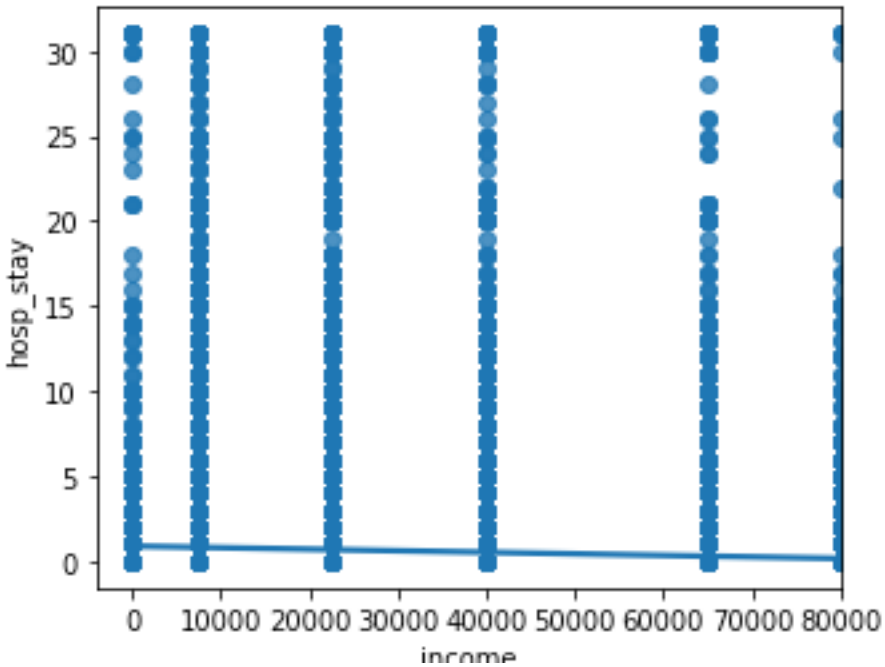


Table 2 - Pearson correlation among the variables.

	Hospital Stay	**Age**	**Income**	**Walking**
Hospital Stay	1.00	0.093	-0.07	-0.03

Age		1.00	-0.14	-0.04
Income			1.00	0.03
Walking				1.00

As *Table 2* demonstrates, there is slightly positive and negative correlation among the variables. Now let's do multiple linear regression and to find out whether there is statistically significant correlation between variables.

By using Stats model library, we get the following results:

Table 3 - R-square and coefficients

R^2	**0.05**
β_1	0.02
β_2	-5.961e-06
β_3	-0.13

Table 4 - T and P values

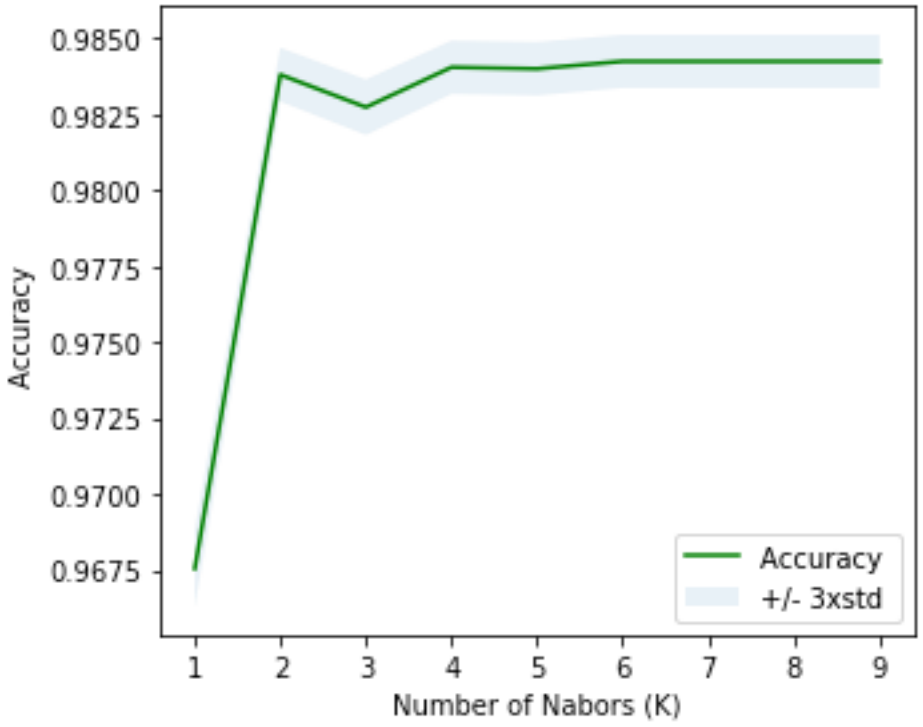
	Age	**Income**	**Walking**
Count	102770	102770	102770
P-Value	0.00	0.00	0.00
T	54	16	6

Table 3 and *Table 4* show that p-values are less than 0.05 and all T values are greater than 1.96. Thus, our coefficients are statistically significant. Let's interpret one of the coefficients. Age coefficient is 0.02 and it is statistically significant. If we increase age one unit, hospital stay will increase 0.02 times. People having 30 years old spend 10 x 0.02 x 24 hours = 4.8 hours less than people who is 40 years old at hospitals.

Walking as an exercise is negatively correlated with hospital stay as expected. There is also negative correlation between hospital stay and income. It is understandable, because people having more income might get qualitative goods and better lifestyles.

Question 4

In this question, we need to define cancer by using the following features age, income and BMI (Body Mass Index). We will apply K nearest neighbor machine learning technique to solve the problem. First, we need to normalize data by making it unit variance and zero mean. After that we need to train data by dividing it 4 parts. Here out-of-sample forecast is used. Let's choose random number for k = 4 and then k = 6. In both cases, train and test set accuracy are approximately 0.98 which is almost 100 percent precision. The best K value for model is k = 6.



4. Conclusion

In conclusion, descriptive and predictive statistics enable us to say that there is statistically significant correlation between healthcare utilization and variables such as age, income and walking as an exercise. The healthcare utilization is positively correlated with age, however negatively correlated with the latter ones.

K nearest neighbor prediction showed 98 percent accuracy which was the best fit when K = 6.