CapstoneThree Final Report

HOW MUCH DO YOU HAVE TO PAY FOR HEALTH INSURANCE?

by Junko Takasawa

July 2021

Problem Statement

USA TODAY states that according to the World Health Care Organization (WHO), health care costs are growing faster than the rest of the global economy.

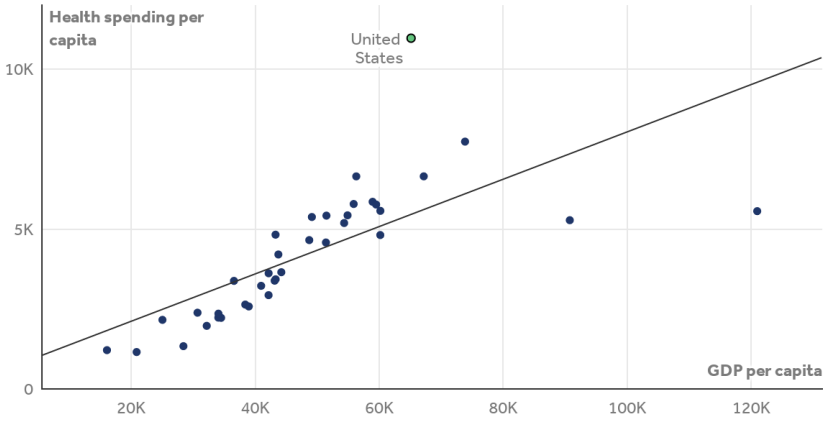
Every country has its unique political, economic, and social climate affecting its health care policies and spending. 24/7 Tempo reviewed health care expenditure data from the Organization for Economic Co-operation and Development (OECD), a group of 34 predominantly rich countries, because health spending is associated with a nation's wealth.

A third of OECD countries spend more than $2,000 per person each year on health care. The 12 countries with the highest health care costs, spend about twice that amount. The differences between countries is staggering, ranging from more than $8,000 per person in the country with the most expensive health care system to $541 in the OECD country with the lowest health care expenses per capita.

The United States has the highest healthcare expenditure per capita in the world. Wikipedia and other source show that it is more than $11,000 in 2019 (approx. $1,000/month) and is growing rapidly.

Rank	Country	2019 Expenditure/capita
1	USA	$11,072
2	Switzerland	$7,732
3	Norway	$6,647
4	Germany	$6,646
5	Austria	$5,851
6	Sweden	$5,782
7	Netherland	$5,765
8	Denmark	$5,568
9	Luxembourg	$5,558
10	Belgium	$5,428
11	Canada	$5,418
12	France	$5,376
13	Ireland	$5,276
14	Australia	$5,187
15	Japan	$4,823

GDP per capita and health consumption spending per capita, 2019 (U.S. dollars, PPP adjusted)



Notes: U.S. value obtained from National Health Expenditure data. Health consumption does not include investments in structures, equipment, or research.

Source: KFF analysis of OECD and National Health Expenditure (NHE) data
• PNG

Peterson-KFF Health System Tracker

HealthCare.gov lists some examples of how much certain health care may cost in the US :

- Fixing a broken leg can cost up to $7,500
- The average cost of a 3-day hospital stay is around $30,000
- Comprehensive cancer care can cost hundreds of thousands of dollars

This is the reason why having health insurance to protect you from high, unexpected costs like these is very important. It is even mandatory to have health insurance in some states, like California, Rode Island, Washington D.C., to name a few.

Unfortunately, health insurance is also very expensive. Thus it is very helpful to know how much you are likely to pay for insurance based on some fundamental features, such as age, gender, region you live in, and what factors are more likely to affect the insurance fee.

Data

Data source from kaggle : (https://www.kaggle.com (https://www.kaggle.com))
Links of datasheet : (https://www.kaggle.com/mirichoi0218/insurance/version/1 (https://www.kaggle.com/mirichoi0218/insurance/version/1))
Sampling methods : Random sampling

I downloaded the CSV file from kaggle, and imported it using pandas. This sample dataset contains 1338 rows of those insured with attributes of fundamental features.

Data Definition

- **age** - age of the insured
- **sex** - gender of the insured
- **bmi** - BMI (Body Mass Index) of the insured
- **children** - number of children of the insured
- **smoker** - smoking status of the insured
- **region** - region where insured lives in
- **charges** - annual insurance charge

Data Wrangling

The dataset is relatively clean, and there are no missing or undefined values in the dataset.

To make a few attributes more legible for data analysis, I created the following additional attributes from "charges", "age", and "bmi".

- **monthly_charge** : charges / 12 (months)
  * **charges** column is for annual premium charge. For analysis, monthly insurance charge is calculated.
- **age_group** : broke them to into groups of "under 20", "20's", "30's", "40's", "50's" and "over 60's".

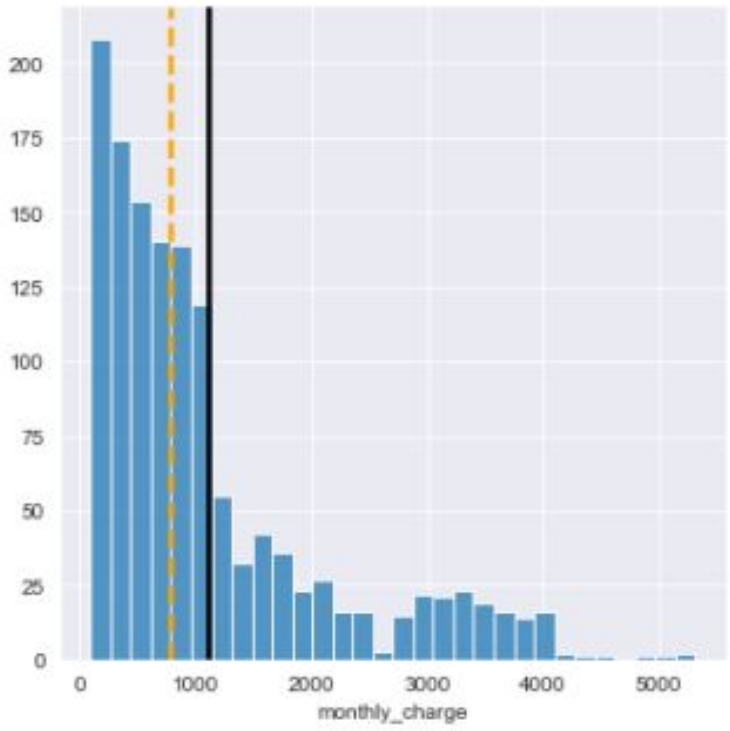
- **weight_status** : grouped them into weight status "Underweight", "Normal", "Overweight" and "Obese" based on CDC Data for BMI standard chart for reference Data Source: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html (https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html).

BMI	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Normal

BMI	Weight Status
25.0 – 29.9	Overweight
30.0 and Above	Obese

Exploratory Data Analysis

This sample dataset coinside with the average monthly health insurance fee found on various reliable sources referenced in the Problem Statement.

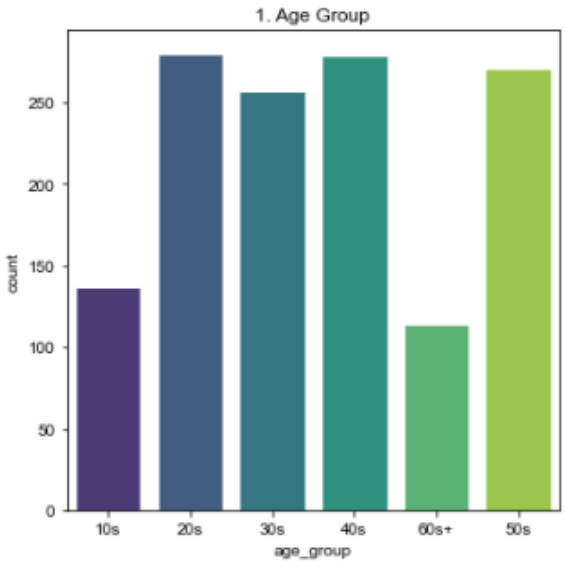


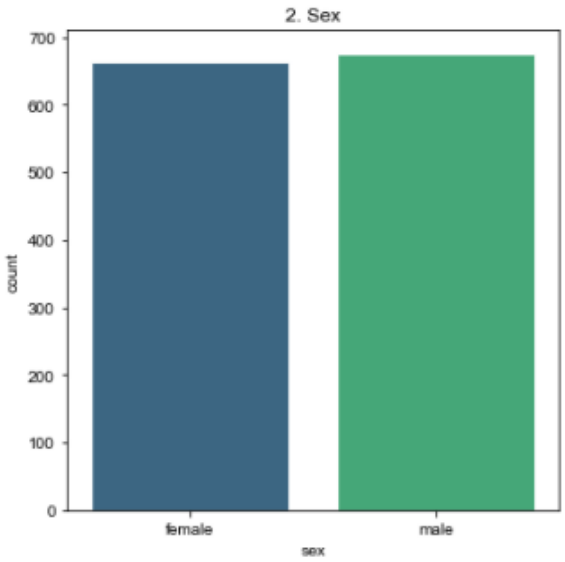
Average monthly fee (black line) : **$1,105.87**

Median (yellow line) : **$781.84**

Data Distribution

This data sample seems fairly well distributed among the age and gender groups.



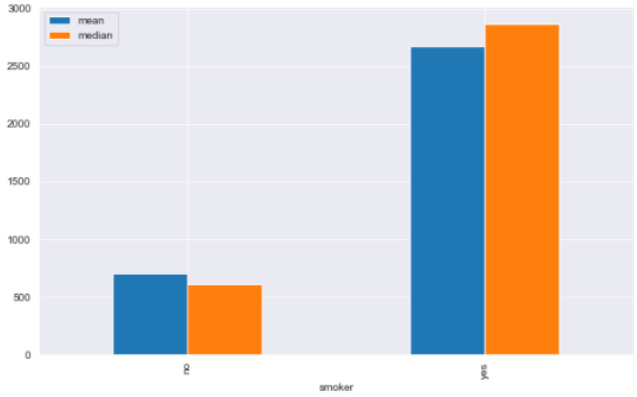


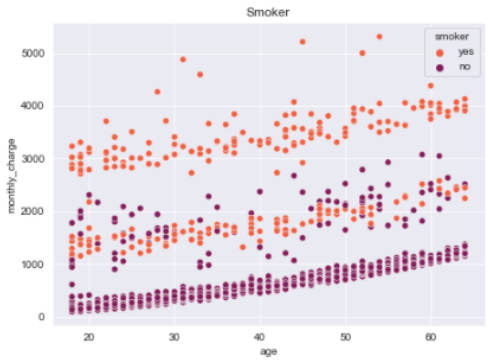
Except for 10's and 60's groups that are more likely not employed and are dependent member of other family members' health insurance policy, 20's, 30's, 40's, and 50's age groups all have around 250 sample data.

Gender groups seem equally distributed between male and female.

Feature - SMOKER

- EDA reveals that "smoker" status seems to have the most significant effect on the price of insurance.
- Left chart indicates that Mean (average) and Median of the insurance price for Smoker is almost five times higher than those who don't smoke.
- As shown on the chart on the right, a group of those who pay significantly higher insurance price in each group are dominantly smokers.





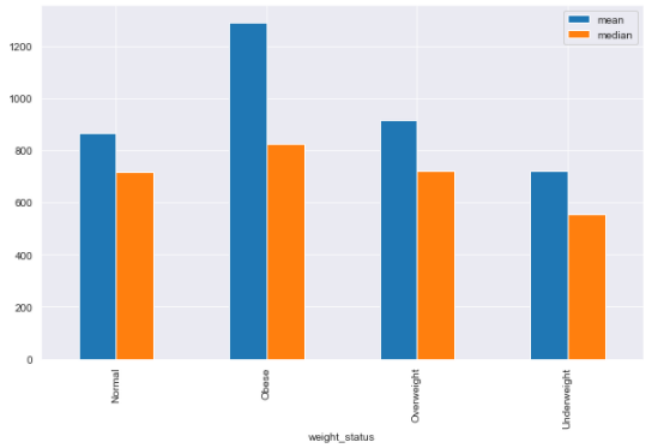
Health Insurance Monthly Fee by Smoking Status

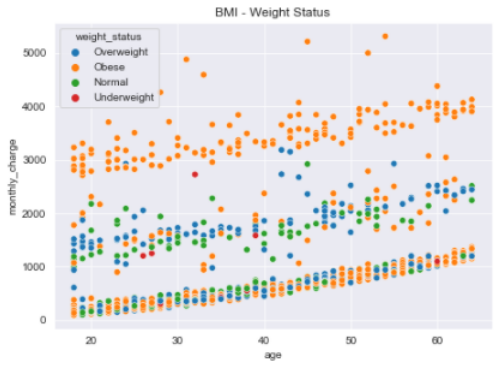
Smoker?	Mean	Median
Yes	$2,670.85	$2,871.36
No	$702.85	$612.11

In [ ]:

Feature - WEIGHT STATUS

- Second most important feature on the health insurance price seems to be the weight status.
- Left chart indicates that these weight status, "Underweight", "Normal", "Overweight", and "Obese", increases the health insurace price in that order.
- Just as what we saw for smoker, the chart on the right shows a group of those who pay significantly higher insurance price in each group are dominantly those in the "Obese" weight status.





Health Insurance Monthly Fee by Weight Status

Weight Status	Mean	Median
Obese	$1,290.96	$824.60
Overweight	$917.23	$721.61
Normal	$867.07	$717.04
Underweight	$721.46	$553.38

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [10]:
```
# import necessary modules

import pandas as pd
import numpy as np

from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# load file prepared in the "Data-Wrangling" stage

df = pd.read_excel('insurance_data.xlsx', index_col=0)
df.head()
```

Out[10]:

	age	sex	bmi	children	smoker	region	charges	monthly_charge	age_group	wei
0	19	female	27.900	0	yes	southwest	16884.92400	1407.08	10s	
1	18	male	33.770	1	no	southeast	1725.55230	143.80	10s	
2	28	male	33.000	3	no	southeast	4449.46200	370.79	20s	
3	33	male	22.705	0	no	northwest	21984.47061	1832.04	30s	
4	32	male	28.880	0	no	northwest	3866.85520	322.24	30s	

In [11]:
```
# load file prepared in the "Data-Wrangling" stage

df = pd.read_excel('insurance_data.xlsx', index_col=0)

df.head()
```

Out[11]:

	age	sex	bmi	children	smoker	region	charges	monthly_charge	age_group	wei
0	19	female	27.900	0	yes	southwest	16884.92400	1407.08	10s	
1	18	male	33.770	1	no	southeast	1725.55230	143.80	10s	
2	28	male	33.000	3	no	southeast	4449.46200	370.79	20s	
3	33	male	22.705	0	no	northwest	21984.47061	1832.04	30s	
4	32	male	28.880	0	no	northwest	3866.85520	322.24	30s	

In [12]:
```
# convert 'sex' and 'smoker' to binary

df_prep = df[['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'monthly_charg
df_prep['gender'] = np.where(df['sex']=='female', 1, 0)  # female = 1, male = 0
df_prep['smoking'] = np.where(df['smoker']=='yes', 1, 0)  # smoker = 1, non-smoke
```

```
<ipython-input-12-3d9b891546fa>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
  df_prep['gender'] = np.where(df['sex']=='female', 1, 0)  # female = 1, male =
0
<ipython-input-12-3d9b891546fa>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
  df_prep['smoking'] = np.where(df['smoker']=='yes', 1, 0)  # smoker = 1, non-s
moker = 0
```

In [13]:
```
df_prep.drop(['sex', 'smoker'], axis = 1, inplace = True)
df_prep.head()
```

```
C:\Users\junko\Anaconda3\lib\site-packages\pandas\core\frame.py:4163: SettingWi
thCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy (https://pandas.pyd
ata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-c
opy)
  return super().drop(
```

Out[13]:

	age	bmi	children	region	monthly_charge	gender	smoking
0	19	27.900	0	southwest	1407.08	1	1
1	18	33.770	1	southeast	143.80	0	0
2	28	33.000	3	southeast	370.79	0	0
3	33	22.705	0	northwest	1832.04	0	0
4	32	28.880	0	northwest	322.24	0	0

In [14]:
```
# change 'region' to binary

df_dummy = pd.get_dummies(df_prep)
df_dummy.head()
```

Out[14]:

	age	bmi	children	monthly_charge	gender	smoking	region_northeast	region_northwest	re
0	19	27.900	0	1407.08	1	1	0	0	
1	18	33.770	1	143.80	0	0	0	0	
2	28	33.000	3	370.79	0	0	0	0	
3	33	22.705	0	1832.04	0	0	0	1	
4	32	28.880	0	322.24	0	0	0	1	

Data Splitting and Scaling

In [15]:
```
X = df_dummy.drop(['monthly_charge'], axis = 1)
y = df_dummy['monthly_charge']
```

In [16]:
```
# split data into 80% training and 20% testing

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 0, test_
```

In [17]:
```
# apply StandardScaler

sc = StandardScaler()

X_train_sc = sc.fit_transform(X_train)
X_test_sc = sc.transform(X_test)
```

In [18]:
```
print(X_train_sc)
```

```
[[-0.14853305  0.54530479  2.41394802 ... -0.55791981 -0.6155209
   1.70862925]
 [-1.49780784  0.59867181 -0.89219519 ... -0.55791981  1.6246402
  -0.58526447]
 [-1.14273553  0.96092064  0.76087642 ... -0.55791981 -0.6155209
  -0.58526447]
 ...
 [ 0.06451033 -0.91339361 -0.89219519 ... -0.55791981  1.6246402
  -0.58526447]
 [-1.42679338  0.77656186 -0.89219519 ...  1.79237229 -0.6155209
  -0.58526447]
 [-0.4325909  -1.97749955 -0.06565939 ... -0.55791981 -0.6155209
   1.70862925]]
```

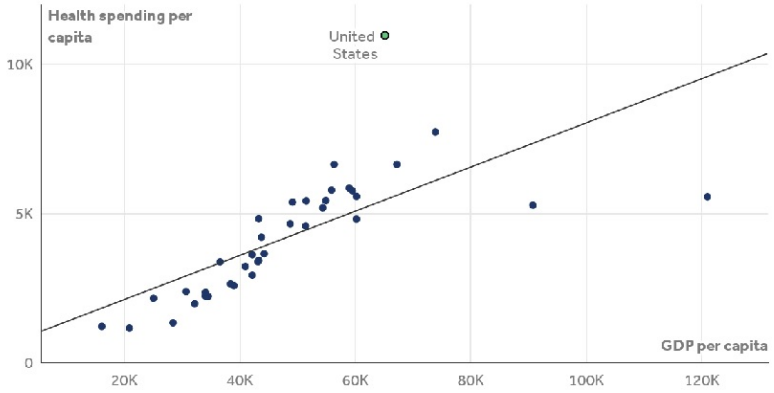
In [ ]:
```

```

Rank	Country	2019 Expenditure/capita

Rank	Country	2019 Expenditure/capita
1	USA	$11,072
2	Switzerland	$7,732
3	Norway	$6,647
4	Germany	$6,646
5	Austria	$5,851
6	Sweden	$5,782
7	Netherland	$5,765
8	Denmark	$5,568
9	Luxembourg	$5,558
10	Belgium	$5,428
11	Canada	$5,418
12	France	$5,376
13	Ireland	$5,276
14	Australia	$5,187
15	Japan	$4,823

GDP per capita and health consumption spending per capita, 2019
(U.S. dollars, PPP adjusted)



Notes: U.S. value obtained from National Health Expenditure data. Health consumption does not include investments in structures, equipment, or research.

Source: KFF analysis of OECD and National Health Expenditure (NHE) data
• PNG

Peterson-KFF
Health System Tracker

In [ ]:

In [ ]:

USA TODAY
https://www.usatoday.com/story/money/2019/04/11/countries-that-spend-the-most-on-public-health/39307147/ (https://www.usatoday.com/story/money/2019/04/11/countries-that-spend-the-most-on-public-health/39307147/)

Wikipedia
https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita (https://en.wikipedia.org/wiki/List_of_countries_by_total_health_expenditure_per_capita)

https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/#item-spendingcomparison_gdp-per-capita-and-health-consumption-spending-per-capita-2019 (https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-

countries/#item-spendingcomparison_gdp-per-capita-and-health-consumption-spending-per-capita-2019)

https://www.statista.com/statistics/184955/us-national-health-expenditures-per-capita-since-1960/ (https://www.statista.com/statistics/184955/us-national-health-expenditures-per-capita-since-1960/)

High Cost https://www.healthcare.gov/why-coverage-is-important/protection-from-high-medical-costs/ (https://www.healthcare.gov/why-coverage-is-important/protection-from-high-medical-costs/)

In [ ]: