Portfolio Project: U.S Medical Insurance Costs

Alfikri Ramadhan - LinkedIn

About The Project

Medical Insurance is a contract that requires an insurer to pay some or all of a person's healthcare costs in exchange for a premium. More specifically, health insurance typically pays for medical, surgical, prescription drug, and sometimes dental expenses incurred by the insured. Health insurance can reimburse the insured for expenses incurred from illness or injury, or pay the care provider directly (Investopedia).

In 2019, the number of people with health insurance in the U.S. was close to 300 million, about 92 percent of the population. Across the United States, Americans pay different premiums monthly for health insurance. The costs for medical insurance can vary based on certain factors. Some of the factor that accounts for costs are age, location, tobacco use, dependents and plan category (healthcare.gov).

In this project, we will analyze individual medical insurance cost and variables that may affect it. This project goal are to:

- Analyze multiple variables in meaningful ways that will provide insight into the dataset
- Implement Data Analyst skills in a real-world scenario
- Utilize Python to model this data in a way that is digestible

Each section of this Notebook will contain code that is being used to analyze the dataset, as well as any findings. The aim is to provide both insights into the data while also showcasing coding and data analysis skills.

The Dataset

The dataset is obtained from Kaggle. The dataset consist of 7 columns:

- `age` : age of primary beneficiary
- `sex` : insurance contractor gender, male or female
- `bmi` : body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight
- `children` : number of children covered by health insurance / number of dependents
- `smoker` : is the beneficiary smoking? yes or no
- `region` : the beneficiary's residential area in the US; northeast, southeast, southwest, northwest.
- `charges` : individual medical costs billed by health insurance, in USD

Import The Dataset

Before begin analyzing our dataset, the first thing we have to do is import the dataset. Then we can see the header and example of row in the dataset. For example, we will take the first row.

In [1]:
```
import csv

with open('insurance.csv', newline='') as insurance_csv:
    csv_reader = csv.reader(insurance_csv)
    csv_headings = next(csv_reader)
    first_line = next(csv_reader)
    print(csv_headings)
    print(first_line)
```

```
['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges']
['19', 'female', '27.9', '0', 'yes', 'southwest', '16884.924']
```

From example above, we can see the insurance.csv file has 7 columns. We can also decide the data type for each columns; `age` and `children` are integers, `bmi` and `charges` are floats, and `sex` , `smoker` and `region` are strings.

We will make separate lists for each column.

In [2]:
```
# create list for each column
age = []
sex = []
bmi = []
children = []
smoker = []
region = []
charges = []

# read the dataset
with open('insurance.csv', newline='') as insurance_csv:
    insurance_dict = csv.DictReader(insurance_csv)
    for data in insurance_dict: # append row from dictionary to list
        age.append(int(data['age']))
        sex.append(data['sex'])
        bmi.append(float(data['bmi']))
        children.append(int(data['children']))
        smoker.append(data['smoker'])
        region.append(data['region'])
        charges.append(float(data['charges']))
```

Investigating the Dataset

Let's see if the dataset has missing data or not.

In [3]:
```
complete_list_dict = {'age': age,
                      'sex': sex,
                      'bmi': children,
                      'children': children,
                      'smoker': smoker,
```

```
                    'region': region,
                    'charges': smoker}

for key in complete_list_dict:
    print("The number of data in {key} list is {list_length}".format(key=key, list_
```

```
The number of data in age list is 1338
The number of data in sex list is 1338
The number of data in bmi list is 1338
The number of data in children list is 1338
The number of data in smoker list is 1338
The number of data in region list is 1338
The number of data in charges list is 1338
```

There are 1338 data in all columns. Since the number is same across all columns, the are no missing data in the dataset

In [4]:
```
total_population = len(age)
print(total_population)
```

```
1338
```

We have imported the dataset, now we can begin analyzing it

Statistical Summary

First, we can perform statistical analysis of data in each columns.

- For categorical columns, we can sum the value in each category.
- For numerical columns, we can evaluate its min, max, and average value.

We will start with **categorical columns**

Sex Distribution

In [5]:
```
# check the unique value in sex column
sex_unique = []
for sx in sex:
    if sx not in sex_unique:
        sex_unique.append(sx)
print(sex_unique)
```

```
['female', 'male']
```

In [6]:
```
# check the distribution of sex
female_num = sex.count('female')
male_num = sex.count('male')
female_pct = round(female_num/total_population*100, 2)
male_pct = round(male_num/total_population*100, 2)
print("There are {count} female data, which contributes to {pct}% of total data"
      .format(count=female_num, pct=female_pct))
print("Ther are {count} male data, which contributes to {pct}% of total data"
      .format(count=male_num, pct=male_pct))
```

```
There are 662 female data, which contributes to 49.48% of total data
Ther are 676 male data, which contributes to 50.52% of total data
```

The sex data is equally distributed to female and male.

Region Distribution

In [7]:
```
# check the unique value in region column
region_unique = []
for reg in region:
    if reg not in region_unique:
        region_unique.append(reg)
print(region_unique)
```

```
['southwest', 'southeast', 'northwest', 'northeast']
```

In [8]:
```
# check the distribution of region
sw_count = region.count('southwest')
se_count = region.count('southeast')
nw_count = region.count('northwest')
ne_count = region.count('northeast')


sw_pct = round(sw_count/total_population*100, 2)
se_pct = round(se_count/total_population*100, 2)
nw_pct = round(nw_count/total_population*100, 2)
ne_pct = round(ne_count/total_population*100, 2)

print("There are {count} people from Southwest region, or {pct}% of total populati
      .format(count=sw_count, pct=sw_pct))
print("There are {count} people from Southeast region, or {pct}% of total populati
      .format(count=se_count, pct=se_pct))
print("There are {count} people from Northwest region, or {pct}% of total populati
      .format(count=nw_count, pct=nw_pct))
print("There are {count} people from Northeast region, or {pct}% of total populati
      .format(count=ne_count, pct=ne_pct))
```

```
There are 325 people from Southwest region, or 24.29% of total population
There are 364 people from Southeast region, or 27.2% of total population
There are 325 people from Northwest region, or 24.29% of total population
There are 324 people from Northeast region, or 24.22% of total population
```

Like sex, the distribution in region is also quite equal.

Smoker Distribution

In [9]:
```
# check the unique value in smoker column
smoker_unique = []
for smk in smoker:
    if smk not in smoker_unique:
        smoker_unique.append(smk)
print(smoker_unique)
```

```
['yes', 'no']
```

In [10]:
```
smoker_yes = smoker.count('yes')
smoker_yes_pct = round(smoker_yes/total_population*100, 2)
print("{smoker} people are smoker, which is {pct}% of total population".format(smol
```

```
274 people are smoker, which is 20.48% of total population
```

Moving forward to **numerical columns**, let's start with age:

Age Distribution

In [11]:
```
age_min = min(age)
age_max = max(age)
```

```
age_avg = round(sum(age)/total_population, 2)
print("The average age in the population is {age} years old".format(age=age_avg))
print("The minimum age in the population is {age} years old".format(age=age_min))
print("The maximum age in the population is {age} years old".format(age=age_max))
```

```
The average age in the population is 39.21 years old
The minimum age in the population is 18 years old
The maximum age in the population is 64 years old
```

Children Distribution

In [12]:
```
no_child = children.count(0)
has_child = total_population - no_child
no_child_pct = round(no_child/total_population*100, 2)
has_child_pct = round(has_child/total_population*100, 2)
print("{pct}% of population has no children ({count} in total)".format(pct=no_chil
print("{pct}% of population has child or children ({count} in total)".format(pct=h
```

```
42.9% of population has no children (574 in total)
57.1% of population has child or children (764 in total)
```

Let's break it down to be more detailed

In [13]:
```
# check number of child
child_range = []
for child in children:
    if child not in child_range:
        child_range.append(child)
        child_range.sort()
print(child_range)
```

```
[0, 1, 2, 3, 4, 5]
```

Because the number of unique values in children is not too high, we can also perceive it as numerical and categorical columns. Let's check the average first

In [14]:
```
average_children = round(sum(children)/total_population, 2)
print(average_children)
```

```
1.09
```

On average, each individual has 1 children

In [15]:
```
no_child = children.count(0)
child_1 = children.count(1)
child_2 = children.count(2)
child_3 = children.count(3)
child_4 = children.count(4)
child_5 = children.count(5)

no_child_pct = round(no_child/total_population*100, 2)
child_1_pct = round(child_1/total_population*100, 2)
child_2_pct = round(child_2/total_population*100, 2)
child_3_pct = round(child_3/total_population*100, 2)
child_4_pct = round(child_4/total_population*100, 2)
child_5_pct = round(child_5/total_population*100, 2)

print("{pct}% of population has no children ({count} in total)".format(pct=no_chil
print("{pct}% of population has 1 child ({count} in total)".format(pct=child_1_pct
print("{pct}% of population has 2 children ({count} in total)".format(pct=child_2_
print("{pct}% of population has 3 children ({count} in total)".format(pct=child_3_
```

```
print("{pct}% of population has 4 children ({count} in total)".format(pct=child_4_
print("{pct}% of population has 5 children ({count} in total)".format(pct=child_5_
```

```
42.9% of population has no children (574 in total)
24.22% of population has 1 child (324 in total)
17.94% of population has 2 children (240 in total)
11.73% of population has 3 children (157 in total)
1.87% of population has 4 children (25 in total)
1.35% of population has 5 children (18 in total)
```

Majority of individual have 0 children. Also, as the number of children increase, the number of individual decreases.

Charges Distribution

In [27]:
```
charges_min = round(min(charges), 2)
charges_max = round(max(charges), 2)
charges_avg = round(sum(charges)/total_population, 2)
print("The average insurance cost in the population is {cost} dollars".format(cost
print("The minimum insurance cost in the population is {cost} dollars".format(cost
print("The maximum insurance cost in the population is {cost} dollars".format(cost
```

```
The average insurance cost in the population is 13270.42 dollars
The minimum insurance cost in the population is 1121.87 dollars
The maximum insurance cost in the population is 63770.43 dollars
```

Analyzing the Dataset

The first and most crucial step that defines how the data science project progress is determining the question. **What question we want to answer with this dataset?**

We are looking at medical insurance cost dataset and some variables that *may* affect it. Early on this project, we have a theory that age is one of the variable that determines insurance cost. The other factor is tobacco use, which we can identify in our dataset as smoker variable.

Lets define some question we will answer from our dataset:

1. Does age factors in one's medical insurance cost?
2. Does BMI affects one's medical insurance cost?
3. Smoking is major cause of cardiovascular disease, such as heart disease and stroke. Does a person being smoker increase their medical insurance charges?

Let's analyze!!

Effect of Age on Medical Insurance Charges

In [17]:
```
print(age_min)
print(age_max)
```

```
18
64
```

The population age are ranging from 18 to 64. Let's classify it to small category to make

analyzing easier. We will divide it into 5 groups:

- Group 1 : 18 - 27 years old
- Group 2 : 28 - 37 years old
- Group 3 : 38 - 47 years old
- Group 4 : 48 - 57 years old
- Group 4 : 58 years old or above

In [28]:
```
# create list for each age group
age_g1 = []
age_g2 = []
age_g3 = []
age_g4 = []
age_g5 = []

# create insurance charges list for each age group
age_g1_charges = []
age_g2_charges = []
age_g3_charges = []
age_g4_charges = []
age_g5_charges = []

num = 0
while num < len(age):
    if age[num] >= 18 and age[num] <= 27:
        age_g1.append(age[num])
        age_g1_charges.append(charges[num])
    elif age[num] >= 28 and age[num] <= 37:
        age_g2.append(age[num])
        age_g2_charges.append(charges[num])
    elif age[num] >= 38 and age[num] <= 47:
        age_g3.append(age[num])
        age_g3_charges.append(charges[num])
    elif age[num] >= 48 and age[num] <= 57:
        age_g4.append(age[num])
        age_g4_charges.append(charges[num])
    else:
        age_g5.append(age[num])
        age_g5_charges.append(charges[num])
    num += 1

# count the individiuals in each group
len_g1 = len(age_g1)
len_g2 = len(age_g2)
len_g3 = len(age_g3)
len_g4 = len(age_g4)
len_g5 = len(age_g5)


print("There are {} people with age from 18 to 27 years old".format(len_g1))
print("There are {} people with age from 28 to 37 years old".format(len_g2))
print("There are {} people with age from 38 to 47 years old".format(len_g3))
print("There are {} people with age from 48 to 57 years old".format(len_g4))
print("There are {} people with age 58 years old or above".format(len_g5))
```

```
There are 362 people with age from 18 to 27 years old
There are 262 people with age from 28 to 37 years old
There are 272 people with age from 38 to 47 years old
There are 278 people with age from 48 to 57 years old
There are 164 people with age 58 years old or above
```

Number of people on with age group 2, 3, and 4 are quite similar. While group 1 is about 100 people more, and group 5 is about 100 people less.

In [19]:
```
age_g1_avg_charges = round(sum(age_g1_charges)/len_g1, 2)
age_g2_avg_charges = round(sum(age_g2_charges)/len_g2, 2)
age_g3_avg_charges = round(sum(age_g3_charges)/len_g3, 2)
age_g4_avg_charges = round(sum(age_g4_charges)/len_g4, 2)
age_g5_avg_charges = round(sum(age_g5_charges)/len_g5, 2)

print("Average insurance cost for individual with age from 18 to 27 years old is {}
print("Average insurance cost for individual with age from 28 to 37 years old is {}
print("Average insurance cost for individual with age from 38 to 47 years old is {}
print("Average insurance cost for individual with age from 48 to 57 years old is {}
print("Average insurance cost for individual with age 58 years old or above is {} d
```

```
Average insurance cost for individual with age from 18 to 27 years old is 9098.19
dollars
Average insurance cost for individual with age from 28 to 37 years old is 11661.81
dollars
Average insurance cost for individual with age from 38 to 47 years old is 13730.04
dollars
Average insurance cost for individual with age from 48 to 57 years old is 15937.66
dollars
Average insurance cost for individual with age 58 years old or above is 19766.12 d
ollars
```

Our analysis shows that individuals in the age group with lowest age (18 to 27 years old) has the lowest average insurance charges of all age groups at 9,098 dollars. On the contrary, the individuals in the age group with highest age (58 years old or above) has the highest average insurance charges at 19,766 dollars. **The increment is more than twice the cost.**

We can also see that an increase from group 1 to group 4 is roughly the same at 2000 dollars. But the increase from group 4 to group 5 shows quite a jump at 4000 dollars. Also about twice the difference.

One major factor in determining charges in health insurance is age, as this article says. This is because as one ages, it means a higher chance of mortality, hospitalization, and medical expenses. Also, as the older one gets, the more difficult it is to determine their risk factors and health-related expenses.

So in general, the higher the age of a person, the higher insurance cost he/she will be charged.

Effect of BMI on Medical Insurance Charges

BMI or Body Mass Index is a measurement that uses the height of the person to determine the suitable weight for the person. More often than not, BMI also accurately predicts the body fat percentage for that person. Therefore, a person with a high BMI has more body fat and vice versa. While using BMI as health indicator is debatable, for the sake of analysis, we will analyze the factor of BMI on insurance cost.

We will divide BMI to several groups:

1. Underweight : bmi below 18.5
2. Normal : bmi 18.5 - 24.9
3. Overweight : bmi 25.0 - 29.9
4. Obese : bmi above 30.0

In [20]:
```
# create list for each bmi group

bmi_underweight = []
bmi_normal = []
bmi_overweight = []
bmi_obese = []

# create insurance charges list for each bmi group
bmi_underweight_charges = []
bmi_normal_charges = []
bmi_overweight_charges = []
bmi_obese_charges = []

num = 0
while num < len(bmi):
    if bmi[num] <= 18.5:
        bmi_underweight.append(bmi[num])
        bmi_underweight_charges.append(charges[num])
    elif bmi[num] > 18.5 and bmi[num] <= 24.9:
        bmi_normal.append(bmi[num])
        bmi_normal_charges.append(charges[num])
    elif bmi[num] > 24.9 and bmi[num] <= 29.9:
        bmi_overweight.append(bmi[num])
        bmi_overweight_charges.append(charges[num])
    else:
        bmi_obese.append(bmi[num])
        bmi_obese_charges.append(charges[num])
    num += 1

# count the individuals in each group
len_underweight = len(bmi_underweight)
len_normal = len(bmi_normal)
len_overweight = len(bmi_overweight)
len_obese = len(bmi_obese)

print("There are {} people with underweight status".format(len_underweight))
print("There are {} people with normal weight status".format(len_normal))
print("There are {} people with overweight status".format(len_overweight))
print("There are {} people with obese status".format(len_obese))
```

```
There are 21 people with underweight status
There are 221 people with normal weight status
There are 380 people with overweight status
There are 716 people with obese status
```

More than half of the population belong to obese status by BMI!!

In [21]:
```
underweight_avg_charges = round(sum(bmi_underweight_charges)/len_underweight, 2)
normal_avg_charges = round(sum(bmi_normal_charges)/len_normal, 2)
overweight_avg_charges = round(sum(bmi_overweight_charges)/len_overweight, 2)
obese_avg_charges = round(sum(bmi_obese_charges)/len_obese, 2)
```

```
print("Average insurance cost for individual with underweight status is {} dollars
print("Average insurance cost for individual with normal weight status is {} dolla
print("Average insurance cost for individual with overweight status is {} dollars"
print("Average insurance cost for individual with obese status is {} dollars".form
```

```
Average insurance cost for individual with underweight status is 8657.62 dollars
Average insurance cost for individual with normal weight status is 10404.9 dollars
Average insurance cost for individual with overweight status is 11006.81 dollars
Average insurance cost for individual with obese status is 15491.54 dollars
```

We can see that people in obese status have highest insurance cost compared to other weight groups. Underweight has the lowest average cost, and normal and overweight status has pretty close average insurance cost.

We mentioned before that using bmi as health indicator is debatable. According to this article from Harvard School of Public Health, BMI is not a perfect measure because it does not directly assess body fat. Muscle and bone are denser than fat, so an athlete or muscular person may have a high BMI, yet not have too much fat. But most people are not athletes, and for most people, BMI is a very good gauge of their level of body fat.

Why an individual with higher BMI (especially in the obese weight category) pays more insurance charges? With BMI in the extremes, the individual has a higher chance of making frequent visits to the hospitals owing to ill health, which translates to higher medical expenditure on the part of the insurance company. Hence, the higher the estimated expenditure on health-related issues, the higher will be the life insurance premiums. For further reading on how BMI and its effect on insurance cost, please check this link.

Our analysis proves that individual with higher bmi likely has higher insurance cost. To have lower insurance cost, it is advised to keep bmi at normal level and avoid the overweight and obese range.

Effect of Smoking Status on Insurance Charges

In [22]:
```
print("There are of {smoker} smoker in our dataset, which is {pct}% of population"
    smoker=smoker_yes,pct=smoker_yes_pct))
```

```
There are of 274 smoker in our dataset, which is 20.48% of population
```

In [23]:
```
# create lists for smoker and non-smoker
smoker_count = []
non_smoker_count = []

# create lists for smoker and non-smoker charges
smoker_charges = []
non_smoker_charges = []

num = 0
while num < len(smoker):
    if smoker[num] == "yes":
        smoker_count.append(smoker[num])
        smoker_charges.append(charges[num])
    else:
        non_smoker_count.append(smoker[num])
        non_smoker_charges.append(charges[num])
    num += 1

# count individuals in each group
```

```
len_smoker = len(smoker_count)
len_non_smoker = len(non_smoker_count)

smoker_pct = round(len(smoker_count)/total_population*100, 2)
non_smoker_pct = round(len(non_smoker_count)/total_population*100, 2)

print("There are {count} smoker in our dataset, which is {pct}% of population".for
     (count=len_smoker, pct=smoker_pct))
print("There are {count} non-smoker in our dataset, which is {pct}% of population"
     (count=len_non_smoker, pct=non_smoker_pct))
```

```
There are 274 smoker in our dataset, which is 20.48% of population
There are 1064 non-smoker in our dataset, which is 79.52% of population
```

In [24]:

```
average_smoker_charges = round(sum(smoker_charges)/len_smoker, 2)
average_non_smoker_charges = round(sum(non_smoker_charges)/len_non_smoker, 2)
smoke_charges_difference = round(average_smoker_charges / average_non_smoker_charg

print("Average insurance cost for smoker is {} dollars".format(average_smoker_char
print("Average insurance cost for non-smoker is {} dollars".format(average_non_smol
print("Individual whom smoke has {}% higher charges than those who doesn't".format
```

```
Average insurance cost for smoker is 32050.23 dollars
Average insurance cost for non-smoker is 8434.27 dollars
Individual whom smoke has 380% higher charges than those who doesn't
```

Our early hypothesis that smoker has higher medical cost than non-smoker is found to be true. **Smokers can pay up to three times more for insurance cost** (Costs & Consequences of Tobacco Use: Connecticut State Department of Public Health%20in,premiums%20than%20non%2Dtobacco%20users.&text=Smokers%20can%20pay%

For next analysis we will look at insurance cost difference for male and female who smoke and doesn't smoke.

In [25]:

```
# create lists for male and female smoker and non-smoker
male_smoker = []
female_smoker = []
male_non_smoker = []
female_non_smoker = []

# create lists for each category charges
male_smoker_charges = []
female_smoker_charges = []
male_non_smoker_charges = []
female_non_smoker_charges = []

num = 0
while num < len(sex):
    if sex[num] == "male":
        if smoker[num] == "yes":
            male_smoker.append(sex[num])
            male_smoker_charges.append(charges[num])
        else:
            male_non_smoker.append(sex[num])
            male_non_smoker_charges.append(charges[num])
    elif sex[num] == "female":
        if smoker[num] == "yes":
            female_smoker.append(sex[num])
            female_smoker_charges.append(charges[num])
        else:
            female_non_smoker.append(sex[num])
```

```
            female_non_smoker_charges.append(charges[num])
    num += 1

print("{} individual are male smoker".format(len(male_smoker)))
print("{} individual are female smoker".format(len(female_smoker)))
print("{} individual are male non-smoker".format(len(male_non_smoker)))
print("{} individual are female non-smoker".format(len(female_non_smoker)))
```

```
159 individual are male smoker
115 individual are female smoker
517 individual are male non-smoker
547 individual are female non-smoker
```

In [26]:
```
average_male_smoker_charges = round(sum(male_smoker_charges)/len(male_smoker_charg
average_female_smoker_charges = round(sum(female_smoker_charges)/len(female_smoker_
average_male_non_smoker_charges = round(sum(male_non_smoker_charges)/len(male_non_
average_female_non_smoker_charges = round(sum(female_non_smoker_charges)/len(femal

print("Average charges for male smoker is {}".format(average_male_smoker_charges))
print("Average charges for female smoker is {}".format(average_female_smoker_charg
print("Average charges for male non-smoker is {}".format(average_male_non_smoker_cl
print("Average charges for female non-smoker is {}".format(average_female_non_smok
```

```
Average charges for male smoker is 33042.01
Average charges for female smoker is 30679.0
Average charges for male non-smoker is 8087.2
Average charges for female non-smoker is 8762.3
```

The average cost for male smokers is higher than that of female smoker. Interestingly, though, for non-smoker the average insurance cost is **higher** for female than male. One possible reason for this is because women are more likely to go to the doctor and take prescriptions, especially during their reproductive ages between 15 and 44 (How Health Insurance Premium varies by Gender?).

This also raise a question. Though female non-smoker has higher insurance cost than male non-smoker due to possible reason above, the opposite happens for male and female smoker. Does this means smoking has greater effect to insurance cost than sex? Or smoking causes more diseases for male than female? This is an example of hypotheses that can be formed by looking at our dataset.

Conclusions

Through working on this dataset, we've found some interesting conclusions:

- Age factors in one's medical insurance cost. In general, higher age individual will pay more insurance cost.
- The average insurance cost is lower for those with normal bmi opposed to overweight and obese bmi.
- Smoking significantly affect medical insurance cost. The average cost difference can reach up to 380%.

Future Directions

Some interesting questions to look in the future might be:

- Does people in different region has difference in insurance cost?
- Why female non-smokers has higher insurance cost than male non-smokers?
- Is smoking has more significant effect to insurance cost than sex?