codecademy

Medical Insurance

Python project

Knox Maclean Bather

30/10/2023

Scope of Analysis – The Infulence of BMI on Insurance Cost.

We will be analysing a large .csv dataset of people's medical insurance data.
In the medical insurance dataset, each row is representative of one person.

Each person has the following information stored about them:

- age (person's age)
- sex (male or female)
- bmi (body mass index)
- children (number of children that person has)
- smoker (smoker or non-smoker)
- region (area they live – southeast/west, northeast/west)
- charges (the insurance cost they are charged)

BMI is known to be an inaccurate measure of health, but it is also easy to measure and assign to individual people. **This analysis will be looking at how BMI scores affect medical insurance costs.**

<u>Assumptions about the data</u>

i. The data is clean and does not need further work to correct it.
ii. The data is representative of real medical insurance data but is itself fictional.
iii. All people in the dataset are customers of the same medical insurance company.
iv. All persons within the dataset have active medical insurance.
v. The data is up to date and an accurate representation of each individual.

Project Index

SYNOPSIS and COMMENTARY ON DATASET

1. Extract the data from the .csv file. What is the average BMI for this pool of people?
2. What are the average BMIs of each 10-year age range? (10-19, 20-29, 30-39, etc)?
3. What are the average insurance cost of these age ranges? Compare the results to the average BMIs, is there a correlation? What other correlation might be present?
4. Is there any correlation between region, BMI and insurance cost?
5. How does the number of children affect BMI? Work out the average BMI for each number of children.

5b. Explore the reason behind the result of the '5 children' average BMI.

6. Does smoking and sex affect BMI numbers? Compare the average BMIs of smokers and non-smokers, males and females, then compare the BMIs of male smokers/ non-smokers and female smokers/ non-smokers.
7. Find two people with the same information but different BMIs, what is the difference in their costs? Are there any other pairings? Are the differences consistent?
8. Work out the highest and lowest BMI scores. Use these values to evenly split the BMI scores into groups. How many people are in each group?
9. What are the average costs of each BMI group?
10. Are any other correlations present in these BMI groups? Average ages/children/etc, split sexes and smokers by percentages.

CONCLUSION

Synopsis

In this analysis **no clear link was found between BMI value and insurance cost**.
I explored results that appeared to show clear links between BMI and cost but also other avenues of the data which ultimately lead me to the conclusion that, while BMI may be a factor in insurance cost, it does not have a large influence over the costing.

Other links found with BMI and degrees of influence over insurance cost.

- **Age** appear to be a **significant factor** in the formation of **insurance cost** and analysis shows a positive correlation with BMI. Particularly accounting for the trend of muscle mass and bone density decreasing as people age, this correlation with BMI suggests a greater percentage of body fat in people as age increases. Something that is linked with negative health implications.
- **Region** appears to have a **strong link with BMI** with different regions showing very **different correlations**, suggesting that the areas might have **different typical lifestyles**. The Southeast region being noted as a hotspot for people in and above the obese BMI category. While this data cannot provide any greater depths of insights into the difference between the region's typical lifestyles, further studies would be useful in developing understanding of these correlations.
- Men skew towards the higher ranges of the BMI scale while women skew towards the lower ranges but there is not a significant difference between the average BMIs of each sex.

<u>Commentary on Dataset</u>
This dataset was provided by CodeCademy as part of the Data Science Career Path. This dataset has many short comings, particularly surrounding its context; it is not known how or where this data was gathered; it is not known if each person provided additional information that was considered in the insurance cost; it is not known if this data is all from the same insurance company or whether it is a mixture of companies with different prices, tiers of insurance, or methods of calculating insurance costs.
As such this dataset is assumed to be fictional but based on real datasets, it will be treated in this analysis as representative of real-life data and so taken in the context of real people and general health trends seen in the general population.

1. Extract the data from the .csv medical insurance file

Set up

Main code

```
import csv
ages = []
sexes = []
bmis = []
children = []
smokers = []
region = []
charges = []
indexes=list(range(len(ages)))
full_med_records = []

with open("insurance.csv", 'r') as medical_records:
    personel_records = csv.DictReader(medical_records)
    for row in personel_records:
        ages.append(row["age"])
        sexes.append(row["sex"])
        bmis.append(row["bmi"])
        children.append(row["children"])
        smokers.append(row["smoker"])
        region.append(row["region"])
        charges.append(row["charges"])
        full_med_records.append(row)

total_bmi = 0
for person in full_med_records:
    total_bmi += float(person["bmi"])
average_bmi = total_bmi/int(len(full_med_records))
print(round(average_bmi, 2))
```

Result

```
30.66
```

On inspecting the .csv file the following data columns were identified for each person:
Age, sex, bmi, number of children, smoker or not, the region they lived, and the amount they were charged for insurance.
These were all extracted into individual lists to make specific analysis more accessible.
A full medical records list was also created in which the elements are a directory of each person's details.

The first bit of analysis done was to calculate the average bmi of all 1338 patients.
The **average bmi** of this pool of people is **30.66**

The standard BMI scale categorises people like so:

BMI	Categorisation
18.5	Underweight
18.5-24.9	Normal
25.0-29.9	Overweight
30.0-34.9	Obese
35<	Extremely Obese

2. Average BMIs for 10 year age ranges.

```
# split the records into age groups
age_groups = {'teens':[], 'twenties':[], 'thirties':[], 'fourties':[], 'fifties':[], 'sixties':[]}
for row in full_med_records:
    age = int(row["age"])
    if  19 >= age >= 10:
        age_groups['teens'].append(row)
    elif 29  >= age >= 20:
        age_groups['twenties'].append(row)
    elif 39  >= age >= 30:
        age_groups['thirties'].append(row)
    elif 49 >= age >= 40:
        age_groups['fourties'].append(row)
    elif 59 >= age >= 50:
        age_groups['fifties'].append(row)
    elif 69 >= age >= 60:
        age_groups['sixties'].append(row)
    else: print("age 70+ catagory needed")
for keys, values in age_groups.items():
    print("The number of people in their "+keys+" is: "+str(len(values)))

# work out the average bmi for each age group
age_gps_bmi_avgs = {}
for keys, values in age_groups.items():
    bmi_total = 0
    for i in values:
        bmi_total += float(i["bmi"])
    bmi_average = bmi_total/(len(values))
    age_gps_bmi_avgs[keys+"_avg_bmi"] = [round(bmi_average, 2)]
print(age_gps_bmi_avgs)
```

In this calculation I used an if/elif conditionals to split each person into a preprepared dictionary (age_groups) and then performed calculations on these lists using a for loop and inputted the average value into a new list with renamed keys.

Results:

```
The number of people in their teens is: 137
The number of people in their twenties is: 280
The number of people in their thirties is: 257
The number of people in their fourties is: 279
The number of people in their fifties is: 271
The number of people in their sixties is: 114
teens_avg_bmi is [29.97]
twenties_avg_bmi is [29.79]
thirties_avg_bmi is [30.44]
fourties_avg_bmi is [30.71]
fifties_avg_bmi is [31.51]
sixties_avg_bmi is [32.02]
```

We can see that there is a **positive trend between BMI and age** with a difference of 2 BMI points from teenage years to sixties.

One point of BMI is the equivalent of ~0.5 stone (6-7lbs) of weight so these two points of difference could show a large difference in body composition, particularly accounting for the fact that people tend to lose muscle and bone density as they age. This suggests that for the same person with consistent weight; their older-self is proportionally less muscle and bone, and more body fat.

This correlation between BMI and age suggests that **the older a person is**, the more likely they are to have a **greater percentage of body fat**, a characteristic that is linked with negative health implications.

3. Calculate the average insurance for each age group

```
age_gps_insurance_avgs = {}
for keys, values in age_groups.items():
    insurance_total = 0
    for i in values:
        insurance_total += float(i["charges"])
    insurance_average = insurance_total/(len(values))
    age_gps_insurance_avgs[keys+"_avg_insurance"] = [round(insurance_average, 2)]
print(age_gps_insurance_avgs)
```

```
{'teens_avg_insurance': [8407.35], 'twenties_avg_insurance': [9561.75], 'thirties_avg_insurance': [11738.78], 'fourties_avg_ins
urance': [14399.2], 'fifties_avg_insurance': [16495.23], 'sixties_avg_insurance': [21248.02]}
```

In this calculation I used the age_groups dictionary (the values contain list all persons in the data set split into age ranges of ten years, each person is individually inputted as a dictionary with all their data) and performed calculations on each person's insurance cost which were under the 'charges' key. I did this in the same way as working out he average bmi for each age group, by using a for loop to total up the charges in each list and dividing it by the number of value in that list.

For the results we can clearly see a **great increase in insurance cost as age increases**. This is much more significant rate than the rise of BMI as age increases which **lead us to believe**, at this early stage, that **age is a much bigger factor in insurance cost than BMI** is. This is unsurprising since someone with a greater age is more likely to be susceptible to health issues such as illness and mobility impediments and is much closer to the average mortality age (in the UK this figure was 81.8 years for males and 85.5 years for females in 2020) than someone with a much younger age who is typically more active and has a stronger immune system.

4. Correlation between region, bmi and insurance cost.

```
locations_bmi_insurance = {}
# splitting all medical records by region.
locations = {}
for person in full_med_records:
    if person['region'] not in locations:
        locations[person['region']] = [person]
        locations_bmi_insurance[person['region']] =[]
    elif person['region'] in locations:
        locations[person['region']].append(person)
    else: print("error")
# calculating the average insurance costs for each region.
locations_insurance_avgs = {}
insurance_avg_list = []
for keys, values in locations.items():
    insurance_total = 0
    for i in values:
        insurance_total += float(i["charges"])
    insurance_average = insurance_total/(len(values))
    insurance_avg_list.append(insurance_average)
    locations_insurance_avgs[keys] = [round(insurance_average, 2)]
# calculating the average bmi for each region.
locations_bmi_avgs = {}
for keys, values in locations.items():
    bmi_total = 0
    for i in values:
        bmi_total += float(i["bmi"])
    bmi_average = bmi_total/(len(values))
    locations_bmi_avgs[keys] = [round(bmi_average, 2)]

# combining the locations and calaulations of averages into a dictionary.
for region in locations_bmi_insurance:
    locations_bmi_insurance[region] = ({"insurance": locations_insurance_avgs[region], "bmi": locations_bmi_avgs[region]})
for keys, values in locations_bmi_insurance.items():
    print(str(keys)+" has values: "+str(values))
```

In this calculation I split all the medical records by **region** into lists within dictionaries using a for loop and if/elif statements to populate the key name and values.
I then performed calculations on the insurance cost and BMI values of each of these people by fetching the “charges” and “bmi” values of each person in the lists of the dictionary values and populated separate dictionaries with these values.
I then combined the keys in these dictionaries to create a dictionary with the region as the key and insurance cost and BMI averages as the values.

Results:

```
southwest has values: {'insurance': [12346.94], 'bmi': [30.6]}
southeast has values: {'insurance': [14735.41], 'bmi': [33.36]}
northwest has values: {'insurance': [12417.58], 'bmi': [29.2]}
northeast has values: {'insurance': [13406.38], 'bmi': [29.17]}
```

From the results, we can see that the SW, NW, and NE all have very similar average BMIs and insurance costs, but the **SE** has much **greater average BMI and slightly higher average costs**. The cost on its own could be explained by a different cost of living in that area, however the **significant difference in the BMI suggests there is more to investigate** within the regional split of people. Most likely there are **complicated social factors** at play and **further studies and data would be needed** to fully explain the greater average BMI in the Southeast.

5a. Does the number of children a person has affect BMI? Find the average BMI of people with the same number of children.

```
def avg_bmi_of_parents(data):
    bmi_parents ={}
    for person in data:
        num_kids = person["children"]
        bmi = float(person["bmi"])
        if num_kids not in bmi_parents:
            if num_kids=="0":
                bmi_parents[num_kids] = [bmi]
            elif num_kids=="1":
                bmi_parents[num_kids] = [bmi]
            elif num_kids=="2":
                bmi_parents[num_kids] = [bmi]
            elif num_kids=="3":
                bmi_parents[num_kids] = [bmi]
            elif num_kids=="4":
                bmi_parents[num_kids] = [bmi]
            elif num_kids=="5":
                bmi_parents[num_kids] = [bmi]
            else: print("The catagory for "+str(num_kids)+
                        " children need to be added.")
        elif num_kids in bmi_parents:
            bmi_parents.get(num_kids).append(bmi)
    for key, value in bmi_parents.items():
        total_bmi = 0
        for bmi in value:
            total_bmi += bmi
        avg_bmi = total_bmi/len(value)
        bmi_parents[key] = round(avg_bmi, 2)
    bmi_parents_avgs = dict(sorted(bmi_parents.items()))
    for key, value in bmi_parents_avgs.items():
        print("People with "+str(key)+" children have an average BMI of "+str(value))
avg_bmi_of_parents(full_med_records)
```

Results:

```
People with 0 children have an average BMI of 30.55
People with 1 children have an average BMI of 30.62
People with 2 children have an average BMI of 30.98
People with 3 children have an average BMI of 30.68
People with 4 children have an average BMI of 31.39
People with 5 children have an average BMI of 29.61
```

From the results we can see **no clear correlation between the number of children and BMI**. There is a slight upwards trend of BMI from 0 children to 4 children, 'slight' because the difference between these averages is small (<1 BMI point). In addition, **all average values are very close to the overall average BMI** (30.66). With these points, we can dismiss causation links of the number of children a person has and their BMI values.

People with **5 children was an interesting result** as it was lower than all other BMI averages and over 1 BMI point below the overall BMI average. This appears to be an anomalous result and lead to further analysis to see if we can find any reasons for this. **This analysis is continued on the next slide.**

In this calculation I used the full_medical_records to isolate each person's "bmi" and "children" values. Conditional statements inside of a for loop allowed easy splitting of the data into a dictionary where the keys were the number of children a person has, and the values were a list of BMIs. A 'not in' conditional was used to add new keys and value lists every time a new number of children was acted on in the loop. Performing calculations on the value lists gave us the average BMI for each number of children people have.

5b. Explore the reason behind the result of the '5 children' average BMI. [5a Continued]

```python
def avg_age_region_of_parents(data):
    age_parents ={}
    region_parents ={}
    for person in data:
        num_kids = person["children"]
        age = float(person["age"])
        region = person["region"]
        if num_kids not in age_parents:
            if num_kids=="0":
                age_parents[num_kids] = [age]
            elif num_kids=="1":
                age_parents[num_kids] = [age]
            elif num_kids=="2":
                age_parents[num_kids] = [age]
            elif num_kids=="3":
                age_parents[num_kids] = [age]
            elif num_kids=="4":
                age_parents[num_kids] = [age]
            elif num_kids=="5":
                age_parents[num_kids] = [age]
            else: None
        elif num_kids in age_parents:
            age_parents.get(num_kids).append(age)
        if region not in region_parents:
            if num_kids=="5":
                region_parents[region] = [region]
            else: None
        elif region in region_parents and num_kids=="5":
            region_parents.get(region).append([region])
# age print
    for key, value in age_parents.items():
        total_age = 0
        for i in value:
            age = float(i)
            total_age += age
        avg_age = total_age/len(value)
        age_parents[key] = round(avg_age, 2)
    age_parents_avgs = dict(sorted(age_parents.items()))
    for key, value in age_parents_avgs.items():
        print("People with "+str(key)+" children have an average age of "+str(value))
# region print
    for key, value in region_parents.items():
        print("There are "+str(len(value))+" people with 5 children in the "+str(key)+" region.")
avg_age_region_of_parents(full_med_records)
```

Results:

```
People with 0 children have an average age of 38.44
People with 1 children have an average age of 39.45
People with 2 children have an average age of 39.45
People with 3 children have an average age of 41.57
People with 4 children have an average age of 39.0
People with 5 children have an average age of 35.61
There are 8 people with 5 children in the southwest region.
There are 3 people with 5 children in the northeast region.
There are 6 people with 5 children in the southeast region.
There are 1 people with 5 children in the northwest region.
```

From this further analysis we can see that the **average age of people with 5 children is noticeably lower** than all other average ages of people with similar number of children. **This links back to questions 2 and 3**. Back in question 2 we noticed a correlation between age and BMI that suggested higher body fat percentages in older people and in question 3 we noticed a strong correlation between age and insurance cost. Hence, **lower age tends to equal lower BMI and cost**. With the results here in 5b we can see that people with 0/1/2/3/4 children have an average age of ~ 39.6, nearly half a decade greater than those with 5 children. This could very well account for the difference in BMI seen in question 5a.

It was also noted in question 4 that people in the SE region tend to have greater BMIs compared to all other regions. The category of 5 children is made up of 33.3% of people in the SE region which has increased the average BMI. If we remove the SE people, the 5 children avg BMI falls to 28.72, an even greater difference from the average age of the 0/1/2/3/4 number of children people.

This extra analysis has confirmed that the **number of children does have a positive correlation with BMI**, but that **age has a significantly stronger positive correlation with BMI**. And so**, in the case of people with five children, the lower average age has undercut expected increase in BMI caused by the number of children**.

I adapted the code in 5a which sorted the people by the number of children. This time I extracted peoples ages and the regions of people with five children (separately).

6. How does smoking and the person's sex affect BMI? Split people into sections based on their smoking habits, their sex, and then both.

```
smoke_bmi = {key: value for key, value in zip(bmis, smokers)}
sex_bmi = {key: value for key, value in zip(bmis, sexes)}

def smoking_avg_bmi_calculator(smoke_bmi):
    smoker_bmi_list =[]
    non_smoker_bmi_list = []
    for key, value in smoke_bmi.items():
        if value == "yes":
            smoker_bmi_list.append(float(key))
        elif value == "no":
            non_smoker_bmi_list.append(float(key))
        else: print("error")
    smoker_avg_bmi = (sum(smoker_bmi_list))/(len(smoker_bmi_list))
    non_smoker_avg_bmi = sum(non_smoker_bmi_list)/(len(non_smoker_bmi_list))
    print("The average bmi for smokers is: "+str(round(smoker_avg_bmi, 2)))
    print("The average bmi for non-smokers is: "+str(round(non_smoker_avg_bmi, 2)))
smoking_avg_bmi_calculator(smoke_bmi)

def sex_avg_bmi_calculator(sex_bmi):
    male_bmi_list =[]
    female_bmi_list = []
    for key, value in sex_bmi.items():
        if value == "male":
            male_bmi_list.append(float(key))
        elif value == "female":
            female_bmi_list.append(float(key))
        else: print("error")
    male_avg_bmi = (sum(male_bmi_list))/(len(male_bmi_list))
    female_avg_bmi = sum(female_bmi_list)/(len(female_bmi_list))
    print("The average bmi for men is: "+str(round(male_avg_bmi, 2)))
    print("The average bmi for women is: "+str(round(female_avg_bmi, 2)))
sex_avg_bmi_calculator(sex_bmi)
```

In these calculations I created **two dictionaries** with each person's bmi as the key and their answer to answers to whether they smoked and what their sex is as the values.
I created functions that used for loops and if conditionals to **split the dictionary** of people into groups with the same value element (whether they smoked/ same sex) by **extracting their BMIs** (the key of each value) and inputting them into a list.
I then performed **calculations on this list**, summing the BMIs and dividing by the length to get the average BMI of each list. These were then all outputted as a using print() sentence to improve readability.

Code continued on the next slide.

6. How does smoking and the person's sex affect BMI? Split people into sections based on their smoking habits, their sex, and then both. [Continued]

```python
def smokers_sexes_bmi_avgs(records):
    sex_smoking_bmis = {"male smoker":[],"male non-smoker":[],"female smoker":[],"female non-smoker":[]}
    for person in records:
        if person["sex"]=="male" and person["smoker"]=="yes":
            sex_smoking_bmis["male smoker"].append(float(person["bmi"]))
        elif person["sex"]=="male" and person["smoker"]=="no":
            sex_smoking_bmis["male non-smoker"].append(float(person["bmi"]))
        elif person["sex"]=="female" and person["smoker"]=="yes":
            sex_smoking_bmis["female smoker"].append(float(person["bmi"]))
        elif person["sex"]=="female" and person["smoker"]=="no":
            sex_smoking_bmis["female non-smoker"].append(float(person["bmi"]))
        else: print("error")
    for key, value in sex_smoking_bmis.items():
        sex_smoking_bmis[key] = sum(value)/len(value)
    for key, value in sex_smoking_bmis.items():
        print("The average bmi for a "+str(key)+" is "+str(round(value, 2)))

smokers_sexes_bmi_avgs(full_med_records)
```

In this calculation I created a **new dictionary** for the **four possible combinations** of sex and smoker with an empty list as the values. I **extracted the BMI of each person** and put them into the value (empty lists).
This allowed me to easily perform **calculations on each group's BMIs**, just like the functions in the previous slide.

Results:

```
The average bmi for smokers is: 31.03
The average bmi for non-smokers is: 31.11
The average bmi for men is: 31.48
The average bmi for women is: 30.72
The average bmi for a male smoker is 31.5
The average bmi for a male non-smoker is 30.77
The average bmi for a female smoker is 29.61
The average bmi for a female non-smoker is 30.54
```

From the results we can see that there is **no discernible correlation between smoking and BMI, and sex and BMI**. Each result shown on the left is close in value to one another and to the average BMI value of the whole dataset (30.66).

7. Find people with the same information but different bmis.

```
similar_persons = []
for person in full_med_records:
    iteration = person
    if person in similar_persons:
        continue
    elif person not in similar_persons:
        for person in full_med_records:
            if person['bmi'] == iteration['bmi']:
                continue
            elif person['age']== iteration['age'] and \
                person['sex']== iteration['sex'] and \
                person['children']== iteration['children'] and \
                person['smoker']== iteration['smoker']:
                similar_persons.append([iteration])
                similar_persons.append([person])
            else: None
    else: None
similar_persons_list = []
for i in similar_persons:
    if i not in similar_persons_list:
        similar_persons_list.append(i)
        print(i)
print(len(similar_persons_list))
```

In this code I used an embedded for statement to **iterate through each person** and **compare them to every other person**, using if/elif statements to compare people's information to others. The information we were looking to get the same was the age, sex, number of children, and smoker status.
I then used conditionals to remove all duplicate values in that list. This new list has **1089** values which shows that the vast majority of people have near identical information with someone else.

Results: I selected the results below due to the relationship with region, BMI and insurance cost seen in question 4.
Of the **19-year-old female smokers with no children**; the three SW had low costs; the four NW, two of the have low cost and two had much higher costs; the one SE person had high costs. All of these people's costs have the differentiator of BMI.
Those with higher costs had BMIs over 30 and all with much **lower costs had BMIs under 30**. For this small gap in bmi (28/29 – 30/33) the insurance cost is approximately doubled. On the standard BMI scale, 30 is the point at which someone is considered obese.
Given the difference in insurance cost between these people, who's only major difference is which side of 30 their BMI score falls, **this grouping** of similar people would **suggest that BMI can greatly contribute to a person's insurance cost**.

>> We will look at more pairings on the next slide to further this analysis.

```
There are 1089 people in this data that have a similar pairing of information but differnt bmis
[{'age': '19', 'sex': 'female', 'bmi': '27.9', 'children': '0', 'smoker': 'yes', 'region': 'southwest', 'charges': '16884.92
4'}]
[{'age': '19', 'sex': 'female', 'bmi': '28.3', 'children': '0', 'smoker': 'yes', 'region': 'southwest', 'charges': '17081.0
8'}]
[{'age': '19', 'sex': 'female', 'bmi': '21.7', 'children': '0', 'smoker': 'yes', 'region': 'southwest', 'charges': '13844.50
6'}]
[{'age': '19', 'sex': 'female', 'bmi': '28.31', 'children': '0', 'smoker': 'yes', 'region': 'northwest', 'charges': '17468.98
39'}]
[{'age': '19', 'sex': 'female', 'bmi': '33.11', 'children': '0', 'smoker': 'yes', 'region': 'southeast', 'charges': '34439.85
59'}]
[{'age': '19', 'sex': 'female', 'bmi': '28.88', 'children': '0', 'smoker': 'yes', 'region': 'northwest', 'charges': '17748.50
62'}]
[{'age': '19', 'sex': 'female', 'bmi': '32.49', 'children': '0', 'smoker': 'yes', 'region': 'northwest', 'charges': '36898.73
308'}]
[{'age': '19', 'sex': 'female', 'bmi': '30.02', 'children': '0', 'smoker': 'yes', 'region': 'northwest', 'charges': '33307.55
08'}]
```

7. Find people with the same information but different bmis. [Continued]

2.
```
[{'age': '23', 'sex': 'female', 'bmi': '28.31', 'children': '0', 'smoker': 'yes', 'region': 'northwest', 'charges': '18033.96
79'}]
[{'age': '23', 'sex': 'female', 'bmi': '31.4', 'children': '0', 'smoker': 'yes', 'region': 'southwest', 'charges': '34166.27
```

3.
```
[{'age': '50', 'sex': 'female', 'bmi': '27.83', 'children': '3', 'smoker': 'no', 'region': 'southeast', 'charges': '19749.383
38'}]
[{'age': '50', 'sex': 'female', 'bmi': '28.16', 'children': '3', 'smoker': 'no', 'region': 'southeast', 'charges': '10702.642
4'}]
[{'age': '50', 'sex': 'female', 'bmi': '28.12', 'children': '3', 'smoker': 'no', 'region': 'northwest', 'charges': '11085.586
8'}]
```

4.
```
[{'age': '54', 'sex': 'male', 'bmi': '33.63', 'children': '1', 'smoker': 'no', 'region': 'northwest', 'charges': '10825.253
7'}]
[{'age': '54', 'sex': 'male', 'bmi': '39.6', 'children': '1', 'smoker': 'no', 'region': 'southwest', 'charges': '10450.552'}]
[{'age': '54', 'sex': 'male', 'bmi': '29.2', 'children': '1', 'smoker': 'no', 'region': 'southwest', 'charges': '10436.096'}]
[{'age': '54', 'sex': 'male', 'bmi': '25.46', 'children': '1', 'smoker': 'no', 'region': 'northeast', 'charges': '25517.1136
3'}]
```

5.
```
[{'age': '37', 'sex': 'male', 'bmi': '30.8', 'children': '0', 'smoker': 'no', 'region': 'southwest', 'charges': '4646.759'}]
[{'age': '37', 'sex': 'male', 'bmi': '29.64', 'children': '0', 'smoker': 'no', 'region': 'northwest', 'charges': '5028.146
6'}]
[{'age': '37', 'sex': 'male', 'bmi': '36.19', 'children': '0', 'smoker': 'no', 'region': 'southeast', 'charges': '19214.7055
3'}]
[{'age': '37', 'sex': 'male', 'bmi': '29.8', 'children': '0', 'smoker': 'no', 'region': 'southwest', 'charges': '20420.6046
```

It should first be noted that, on inspection of **all the paired people**, most pairings have **very similar insurance costs despite differences in BMI**.
Examples 2, 3, 4, 5 have been chosen because of the differences in insurance cost.

2. **Upholds the initial conclusion** that the first example suggested that BMI categories play a significant role in insurance cost.
3. **Opposes the initial conclusion**. The person in this set of pairings with the highest insurance cost has the lowest BMI. It is worth noting that all of these BMIs are under 30.
4. **Strongly opposes the initial conclusion**. Like example 3 the person with the highest insurance cost in this pairing set has the lowest BMI of 25.46. Unlike example 3, the three people with lower insurance in example 4 have BMIs of 33.63, 39.60, and 29.2. The first falls into the obese category and the second falls into the extremely obese category!
5. **Strongly opposes the initial conclusion**. Similar ages males, similar BMIs but two have much higher insurance costs.

This all allows us to dismiss the initial suggestion that BMI contributes greatly to insurance cost due to the inconsistent costs between pairings of similar people who have different BMIs. While some might support the suggestion, many do not support it and some - like examples 3, 4, 5 - completely oppose it.

8. Create a range between the highest and lowest BMIs, separating all people in the data into these sections.

```
def highest_bmi(records):
    top_bmi = float("-inf")
    for person in records:
        bmi = float(person["bmi"])
        if top_bmi < bmi:
            top_bmi = float(person["bmi"])
        else: None
    print("Highest BMI in the records is: "+str(top_bmi))
def lowest_bmi(records):
    bottom_bmi = float("inf")
    for person in records:
        bmi = float(person["bmi"])
        if bmi < bottom_bmi:
            bottom_bmi = float(person["bmi"])
        else: None
    print("Lowest BMI in the records is: "+str(bottom_bmi))
highest_bmi(full_med_records)
lowest_bmi(full_med_records)
```

Results:

```
Highest BMI in the records is: 53.13
Lowest BMI in the records is: 15.96
```

We can see in the sections below that the upper end of the scale **(>30) contains many more people than the lower end (<25)**. This is unsurprising given the average BMI is 30.66 but does classify more people as 'obese' than 'underweight' or 'normal' **suggesting overweight issues in a large proportion of the population of this dataset**.

```
bmi_range = list(range(15, 55, 5))
bmi_sections ={}
bmi_sections_full={}
for i in bmi_range:
    bmi_sections_full[i] = []
for person in full_med_records:
    bmi = float(person["bmi"])
    if bmi_range[0]< bmi <bmi_range[1]:
        bmi_sections_full[bmi_range[0]].append(person)
    elif bmi_range[1]< bmi <bmi_range[2]:
        bmi_sections_full[bmi_range[1]].append(person)
    elif bmi_range[2]< bmi <bmi_range[3]:
        bmi_sections_full[bmi_range[2]].append(person)
    elif bmi_range[3]< bmi <bmi_range[4]:
        bmi_sections_full[bmi_range[3]].append(person)
    elif bmi_range[4]< bmi <bmi_range[5]:
        bmi_sections_full[bmi_range[4]].append(person)
    elif bmi_range[5]< bmi <bmi_range[6]:
        bmi_sections_full[bmi_range[5]].append(person)
    elif bmi_range[6]< bmi <bmi_range[7]:
        bmi_sections_full[bmi_range[6]].append(person)
    elif bmi_range[7]< bmi:
        bmi_sections_full[bmi_range[7]].append(person)

for key, value in bmi_sections_full.items():
    new_key = str(key)+"_"+str(key+4)
    bmi_sections[new_key]=len(value)
for key, value in bmi_sections.items():
    print(key+": "+str(value))
```

```
{'15_19': 41, '20_24': 204, '25_29': 384, '30_34': 389, '35_39': 225, '40_44': 71, '45_49': 17, '50_54': 3}
```

Green line shows the break in code.

In this calculation I mimicked the extended BMI category numbers, splitting the people in the data into sections with a range of 5. All ranges are labelled with a range of 4 to improve readability. Eg the first category is 15 – 19.99 but has been shown as 15-19.

Using a list(range) I created a dictionary whose keys were the elements of the list(range). The intention was to make this code suitable for other ranges of numbers.

Once the people had been sorted, I renamed the keys to more closely resemble the ranges of values they represent.

9. What are the average costs of each BMI group?

```
bmi_sections_costs ={}
for key, value in bmi_sections_full.items():
    total_cost = 0
    new_key = str(key)+"_"+str(key+4)
    for person in value:
        total_cost += float(person["charges"])
    avg_cost = total_cost/len(value)
    bmi_sections_costs[new_key] = round(avg_cost, 2)
print("The average cost of insurance for BMI sections are:")
for key, value in bmi_sections_costs.items():
    print(key+": "+str(value))
```

Results:

```
The average cost of insurance for BMI sections are:
15_19: 8838.56
20_24: 10572.37
25_29: 10989.85
30_34: 14429.42
35_39: 17022.26
40_44: 16569.6
45_49: 17815.04
50_54: 16034.31
```

In this calculation I used the same BMI sections produced in question 8 and used imbedded for loops to create a new dictionary of the average insurance cost for each section.

I used the original "bmi_sections" dictionary to perform the calculations, extracting the 'charges' for each person, performing the calculations and then inputting that average into the "bmi_sections_costs" dictionary with a re-named key.

We can see that there is a **positive correlation between BMI and average insurance cost** in the results above. The 15.00-19.99 bmi section's average cost being $8,839 compared to the 35.00-39.99 bmi section's $17,022, that's a near doubling of cost!

>> The **lower ranges**, 15-19, 20-24, 25-29 which cover the normal and overweight BMI categories, are the **most similar in cost** and on the lower side of the scale. This is **likely due to the relationship between age, BMI, and cost** – those people with a lower age, tend to have a "healthier" BMI and lower insurance costs.

>> 30.00-34.99 marks the obese category for BMI and the average cost jumps up greatly (~$3,500) from the 25.00-29.99 range

>> Above 35.00 BMI value (extremely obese category) the average costs are similar to each other, around the $17,000 mark which is ~$6,000 greater than the 25-29 category. A significant jump up in price.

The **greater average costs of BMI above the values of 30 and 35 suggests there BMI is considered in insurance cost**. Relationships between age and cost and BMI have already been discussed, and **age may be a factor in the insurance price** in each of these BMI sections. The average ages of each BMI section will be calculated in question 10. But given the results above, it is likely that age cannot fully account for these differences, suggesting that BMI has a positive relationship with insurance cost.

10. Do these BMI groups have any correlations with other data groups?

```
# age
bmi_sections_ages = bmi_sections_costs
for key, value in bmi_sections_full.items():
    total_age = 0
    for person in value:
        total_age += float(person["age"])
    avg_age = total_age/len(value)
    bmi_sections_ages[key] = round(avg_age, 2)
print("The average ages of each BMI section is as follows: ")
print(bmi_sections_ages)
```

```
# children
bmi_sections_children = bmi_sections_costs
for key, value in bmi_sections_full.items():
    total_number = 0
    for person in value:
        total_number += float(person["children"])
    avg_kids = total_number/len(value)
    bmi_sections_children[key] = round(avg_kids, 2)
print("The average number of children of each BMI section is as follows: ")
print(bmi_sections_children)
```

```
# smoking
bmi_sections_smokers = bmi_sections_costs
for key, value in bmi_sections_full.items():
    total_smokers = 0
    for person in value:
        if person["smoker"] == "yes":
            total_smokers += 1
        else: None
    avg_yes = total_smokers/len(value)
    bmi_sections_smokers[key] = str(round(avg_yes*100, 2))+"%"
print("The percentage of smokers in each BMI section is as follows: ")
print(bmi_sections_smokers)
```

For each of these categories I **adapted the code in question 9** which worked out the average costs of each BMI group to work out the average **ages**, number of **children**, percentage of **smokers**, split of **sexes**, and the split of **regions**.
Each BMI section was then renamed for readability. Eg, the 15.00-19.99 BMI section is labelled as 15-19.

- The ages and number of children were calculated as an average value.
- The number of smokers was calculated as a percentage of that group that do smoker.
- The split of sexes was calculated as pairs of male and female percentages for each BMI group.
- The split of regions was calculated as ratios between the four regions.

10. Do these MBI groups have any correlations with other data groups? [Continued]

```
# sexes
bmi_sections_sexes = bmi_sections_costs
for key, value in bmi_sections_full.items():
    total_men = 0
    total_women = 0
    for person in value:
        if person["sex"] == "male":
            total_men += 1
        elif person["sex"] == "female":
            total_women += 1
    dec_male = round(total_men/len(value), 2)
    dec_female = round(total_women/len(value), 2)
    bmi_sections_smokers[key] ="M:F "+str(dec_male)+":"+str(dec_female)
print("The proportions of males and females in each BMI section is as follows: ")
print(bmi_sections_smokers)
```

```
# locations
bmi_sections_regions = bmi_sections_costs
for key, value in bmi_sections_full.items():
    total_nw = 0
    total_ne = 0
    total_sw = 0
    total_se = 0
    for person in value:
        if person["region"] == "northwest":
            total_nw += 1
        elif person["region"] == "northeast":
            total_ne += 1
        elif person["region"] == "southwest":
            total_sw += 1
        elif person["region"] == "southeast":
            total_se += 1
    nw_dec = round(total_nw/len(value), 2)
    ne_dec = round(total_ne/len(value), 2)
    sw_dec = round(total_sw/len(value), 2)
    se_dec = round(total_se/len(value), 2)
    bmi_sections_smokers[key] ="NW:"+str(nw_dec)+" NE:"+str(ne_dec)+" SW:"+str(sw_dec)+" SE:"+str(se_dec)
print("The proportions of regions in each BMI section is as follows: ")
print(bmi_sections_regions)
```

For each of these categories I **adapted the code in question 9** which worked out the average costs of each BMI group to work out the average **ages**, number of **children**, percentage of **smokers**, split of **sexes**, and the split of **regions**.
Each BMI section was then renamed for readability. Eg, the 15.00-19.99 BMI section is labelled as 15-19.

- The ages and number of children were calculated as an average value.
- The number of smokers was calculated as a percentage of that group that do smoker.
- The split of sexes was calculated as pairs of male and female percentages for each BMI group.
- The split of regions was calculated as ratios between the four regions.

10. Do these MBI groups have any correlations with other data groups? [Continued]

Results:

```
Avg ages of each BMI section:
15_19: 33.73
20_24: 36.91
25_29: 38.85
30_34: 39.41
35_39: 41.92
40_44: 40.9
45_49: 42.71
50_54: 21.0

Avg children in each BMI section:
15_19: 1.15
20_24: 1.12
25_29: 1.06
30_34: 1.15
35_39: 1.03
40_44: 1.01
45_49: 1.59
50_54: 0.67

Smokers(%) in each BMI section:
15_19: 21.95%
20_24: 22.55%
25_29: 19.27%
30_34: 18.77%
35_39: 22.22%
40_44: 22.54%
45_49: 23.53%
50_54: 33.33%
```

```
Males:females in each BMI section:
15_19: M:46.3%, F:53.7%
20_24: M:47.5%, F:52.5%
25_29: M:48.7%, F:51.3%
30_34: M:51.9%, F:48.1%
35_39: M:52.4%, F:47.6%
40_44: M:54.9%, F:45.1%
45_49: M:52.9%, F:47.1%
50_54: M:100.0%, F:0.0%

Regional split in each BMI section:
15_19: NW:0.34 NE:0.44 SW:0.2 SE:0.02
20_24: NW:0.27 NE:0.32 SW:0.21 SE:0.2
25_29: NW:0.28 NE:0.26 SW:0.26 SE:0.21
30_34: NW:0.27 NE:0.23 SW:0.26 SE:0.24
35_39: NW:0.16 NE:0.17 SW:0.26 SE:0.42
40_44: NW:0.11 NE:0.2 SW:0.14 SE:0.55
45_49: NW:0.0 NE:0.06 SW:0.18 SE:0.76
50_54: NW:0.0 NE:0.0 SW:0.0 SE:1.0
```

From these results we can see that:

- **Age has a positive correlation with BMI**. A higher age will statistically mean a higher change of having a greater BMI. However, the difference between average ages is less than 10 from the lowest to the second highest suggesting that this is **not a strong correlation**.

- **The number of children does not appear to have any correlation with BMI**. The 45-49 BMI range does have a much higher average than the other ranges suggesting parents with multiple kids are more likely to be higher up the BMI scale.

- **Smoking has no correlation with BMI**, it has consistent percentage across all BMI ranges. The only exception being the 50-54 range which can be discounted as it only has three people in it, so skewing the %.

- The **male to female ratio has slight correlations with BMI** but is relatively consistent through the BMI ranges with **females being more prominent in the lower ranges** and **males more prominent in the upper ranges**.

- **Regional split shows some strong correlations with BMI**. **SE** has a **strong positive correlation** with its proportion rising greatly as BMI range goes up. **NW and NE have negative correlations**, weaker than SE but inverse; the proportion of people in these regions decreases as BMI increases. **SW is the only exception** with little correlation, its proportion staying consistent around 0.20 through all BMI ranges.
This **suggests** that **region is playing a large part in BMI values** with the NW and NE having lower BMI values, suggesting better eating/exercise habits and high BMI values for the SE suggesting people are generally unhealthier in this region.
Further research and additional studies into daily exercise and dietary habits of people in this region should be explored.

Conclusions

BMI Sections: Average Costs

```
15_19: 8838.56
20_24: 10572.37
25_29: 10989.85
30_34: 14429.42
35_39: 17022.26
40_44: 16569.6
45_49: 17815.04
50_54: 16034.31
```

BMI Sections: Average Ages

```
15_19: 33.73
20_24: 36.91
25_29: 38.85
30_34: 39.41
35_39: 41.92
40_44: 40.9
45_49: 42.71
50_54: 21.0
```

Age Groups: Average BMIs

```
teens_avg_bmi is [29.97]
twenties_avg_bmi is [29.79]
thirties_avg_bmi is [30.44]
fourties_avg_bmi is [30.71]
fifties_avg_bmi is [31.51]
sixties_avg_bmi is [32.02]
```

Age Groups: Average Costs

```
teens_avg_insurance: [8407.35]
twenties_avg_insurance: [9561.75]
thirties_avg_insurance: [11738.78]
fourties_avg_insurance: [14399.2]
fifties_avg_insurance: [16495.23]
sixties_avg_insurance: [21248.02]
```

In this analysis we have looked at BMI in many different context, particularly from the grouping of other data and of other data in BMI groupings. Of these, the results most conducive for this BMI analysis have been those with data of age and cost.

The average cost of BMI sections suggests a clear relationship between cost and BMI, as does the average BMIs of age groups. Both of these positive correlations suggesting as age rises, BMI rises and cost rises. The average cost for BMI sections particularly showcases the difference between BMI categories with sharp increases at BMI values of 30 (categorised as obese) and >35 (categorised as extremely obese) with costs levelling out at BMIs above 35 promoting the idea that people with BMIs above 35 have been categorise together during costing.

Conversely, when looking at the average costs of age groups and the average ages in BMI sections we see a very different picture; costs has a very strong, positive relationship with age and looking at the average ages in each BMI section we can see a positive relationship between age and BMI, though only a few years between each section this would account for a large increases in cost between each BMI section. This even accounts for the levelling out of costs above 35 BMI value. Comparing the BMI sections above 35 (underlined above), we can see that the ages dip and rise as the costs dip and rise! However, the 50_54 range is still much higher than expected given the much lower age. On inspection of the three people in this BMI section we can see that they are all; male, aged 18-23, 0 or 1 child, in the southeast region, but the charges are very different with two people having an average charge of ~$2,000 and the other man having a charge of ~$44,500! That is a huge difference and accounts for the average cost being larger than expected given the average age of the group.

Both this and inspection of the full list of results from question 7 (pairing people with similar information but different BMIs) **suggests that there is no true link between BMI and insurance costs**, and any link is overshadowed by other factors, most notably 'age'. We can also assume that some of these large differences in costs between individuals with very similar information is down to additional factors/ information gathered during the process of producing an insurance cost or possibly even different insurance plans (depending on the individuals extent of medical requirements and history)