MARCH 5, 2021

US MEDICAL INSURANCE COSTS

CODECADEMY – PYTHON PORTFOLIO PROJECT

CORIE TYLER

Table of Contents

Objectives

Personal Objectives:

1. Organize data into variables for analysis.
2. Learn how to present my findings in basic tables, charts, and graphs primarily using tabulate and matplotlib.
3. Practice using Python class methods (one minute I understand, the next I do not 😅).
4. Write clearly defined code using what I have learned and summarize findings.

Codecademy Project Objectives:

- Work locally on your own computer.
- Import a dataset into your program.
- Analyze a dataset by building out functions or class methods.
- Use libraries to assist in your analysis.
- Optional: Document and organize your findings.
- Optional: Make predictions about a dataset’s features based on your findings.

Overview

The Codecademy Data Science Pathway offers a Python Fundamentals course for the student to get started with coding and datasets right away. This project is the culmination of the Python Fundamentals course and is intended to provide some insight into the development of the student's coding abilities without the usual prompts, questions, and answers provided.

Except where indicated, all work is my own.

Methodology

Before importing the datafile, I looked over the data to get an idea of what is offered for analysis. The dataset is presented in .csv format with 7 columns representing age, sex, bmi, number of children, smoking status, region, and annual insurance cost.

Table 1: Sample of dataset rows and columns

```
   age     sex   bmi  children smoker     region  charges

0   19  female 27.90         0    yes  southwest 16884.92

1   18    male 33.77         1     no  southeast  1725.55

2   28    male 33.00         3     no  southeast  4449.46

3   33    male 22.70         0     no  northwest 21984.47

4   32    male 28.88         0     no  northwest  3866.86
```

The dataset has no errors or omissions.

I will be looking for basic statistical values present among the dataset and comparing those values by breaking down the data by column. I would like to know if certain factors affect the actual cost of insurance more than others.

The methodology for collecting the data is unknown.

Integrity

The original dataset has not been manipulated in any way. The process for pulling different parts of the data has been carefully performed to keep the data in original order and fully intact.

Much of the numerical data contains decimal values (floats). As much as I was able, I limited the decimal values to 2 places to keep the numbers uniform and simple. In the future, I look forward to learning what the conventions are for this type of data.

Insights and Analysis

Population Statistics

Population Statistical Data

Table 2: Population Statistics

	age	bmi	children	charges	sex	smoker	region
count	1338.00	1338.00	1338.00	1338.00	-	-	-
mean	39.21	30.66	1.09	13270.42	-	-	-
std	14.05	6.10	1.21	12110.01	-	-	-
min	18.00	15.96	0.00	1121.87	-	-	-
25%	27.00	26.30	0.00	4740.29	-	-	-
50%	39.00	30.40	1.00	9382.03	-	-	-
75%	51.00	34.69	2.00	16639.91	-	-	-
max	64.00	53.13	5.00	63770.43	-	-	-
mode	18.00	32.30	0.00	1639.56	male	no	southeast
mode freq	69.00	13.00	574.00	2.00	676	1064	364
range	46.00	37.17	5.00	62648.55	-	-	-

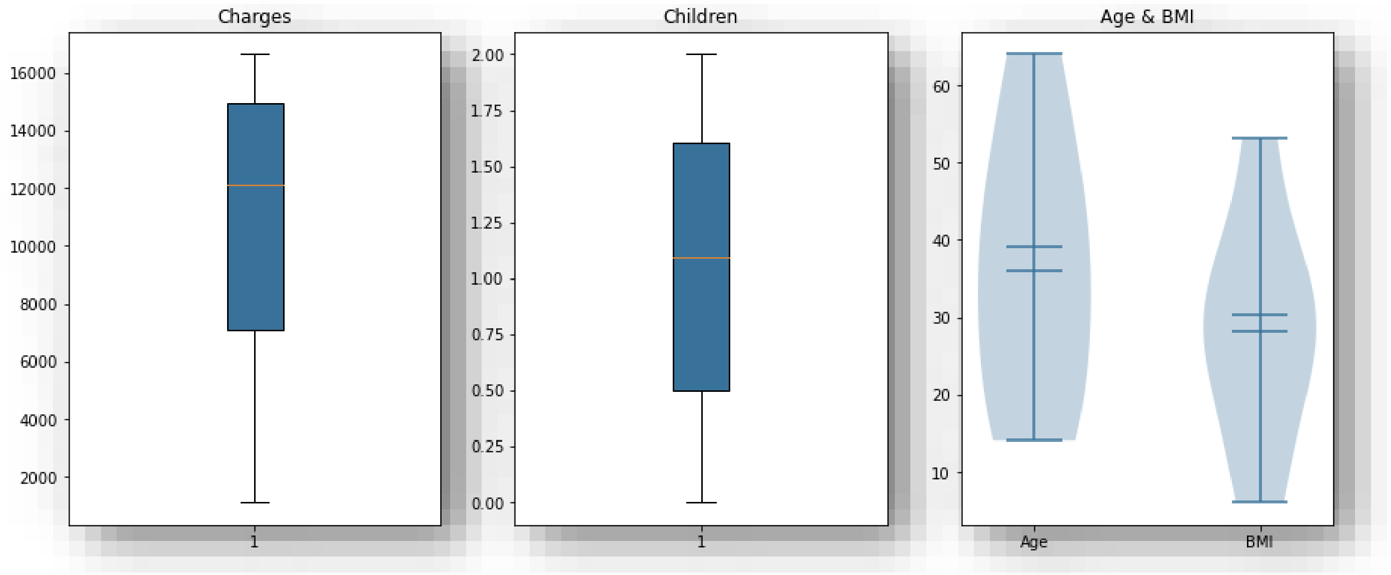
There are 1338 rows representing 1338 individual patients, insurance cost factors, and annual charges. There are 1064 nonsmokers represented and there are slightly more males than females.

The average patient is 39 years old with a BMI of 30.66. They have approximately 1 child and were charged $13,270.42 for insurance.

There are 3% more patients in the Southeast region than any other region.

Visualize individual statistics of numerical variables.

Figure 1: Box Plot and Violin Plot for Charges, Children, Age, and BMI



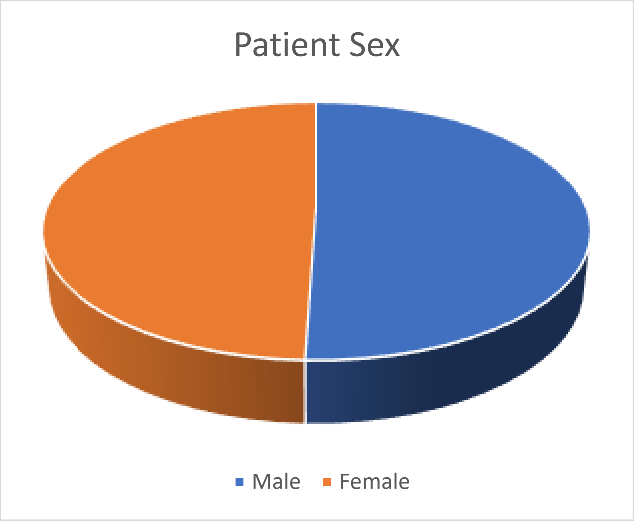
Visualize individual statistics of categorical variables.

Figure 2: Patient Sex

Figure 3: Patient Smoker Status

Males	676	50.52%
Females	662	49.48%

Smokers	1064	79.52%
Nonsmokers	274	20.48%



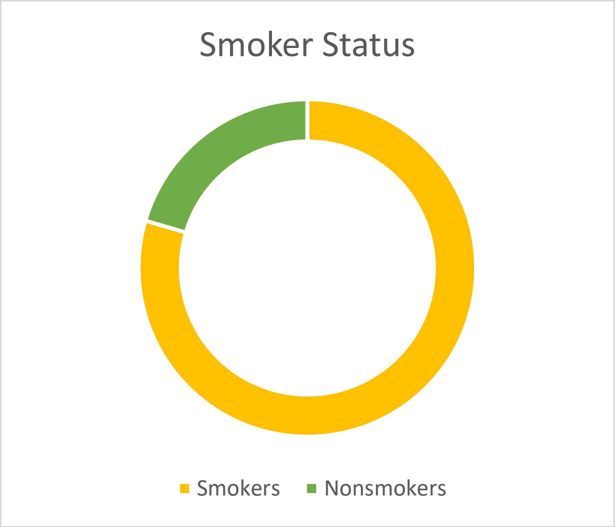
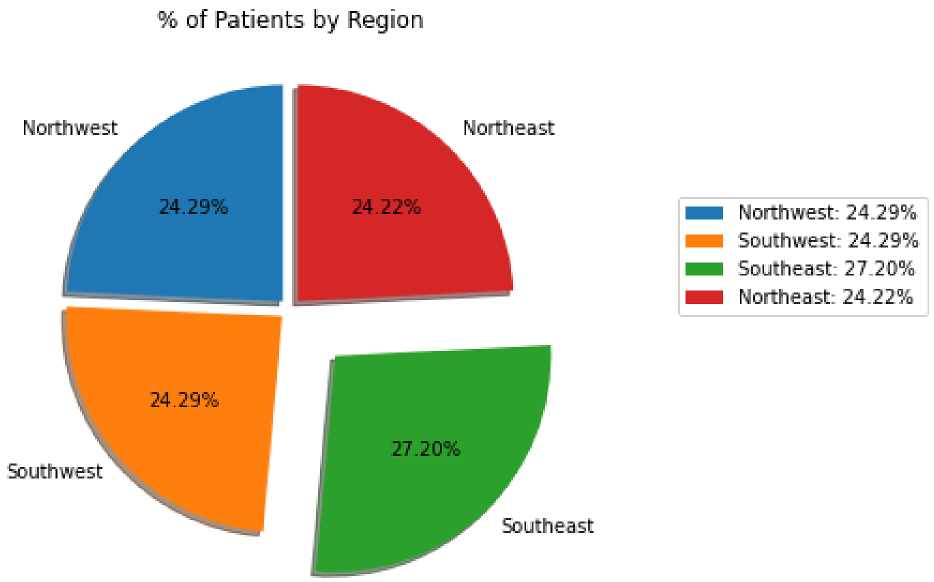


Figure 4: Percentage of Patients by Region

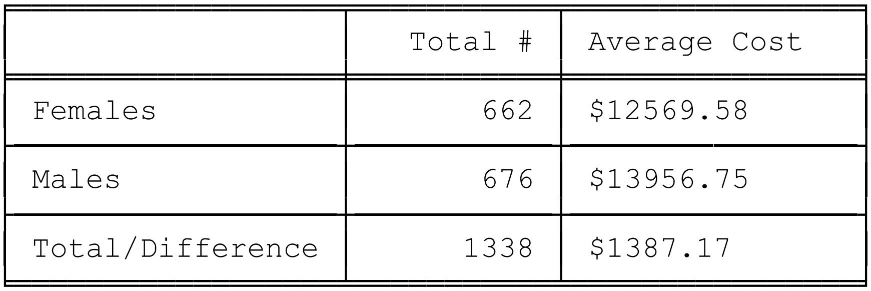


Northwest	Southwest	Southeast	Northeast
325	325	364	324

Average Cost for Each Insurance Variable

Males are charged an average of $1,387.17 more for insurance than females.

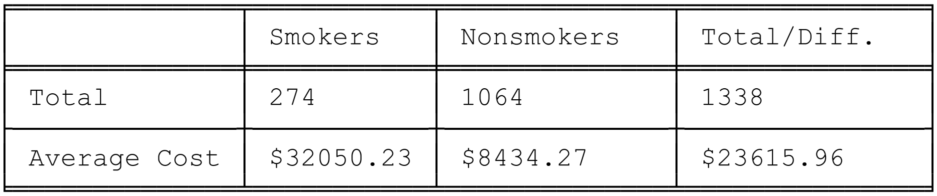
Table 3: Average Annual Cost for Females v Males



	Total #	Average Cost
Females	662	$12569.58
Males	676	$13956.75
Total/Difference	1338	$1387.17

Smokers are charged an average of $23,615.96 more for insurance than nonsmokers.

Table 4: Average Annual Cost for Smokers v Nonsmokers



	Smokers	Nonsmokers	Total/Diff.
Total	274	1064	1338
Average Cost	$32050.23	$8434.27	$23615.96

Age Groups and Average Charges

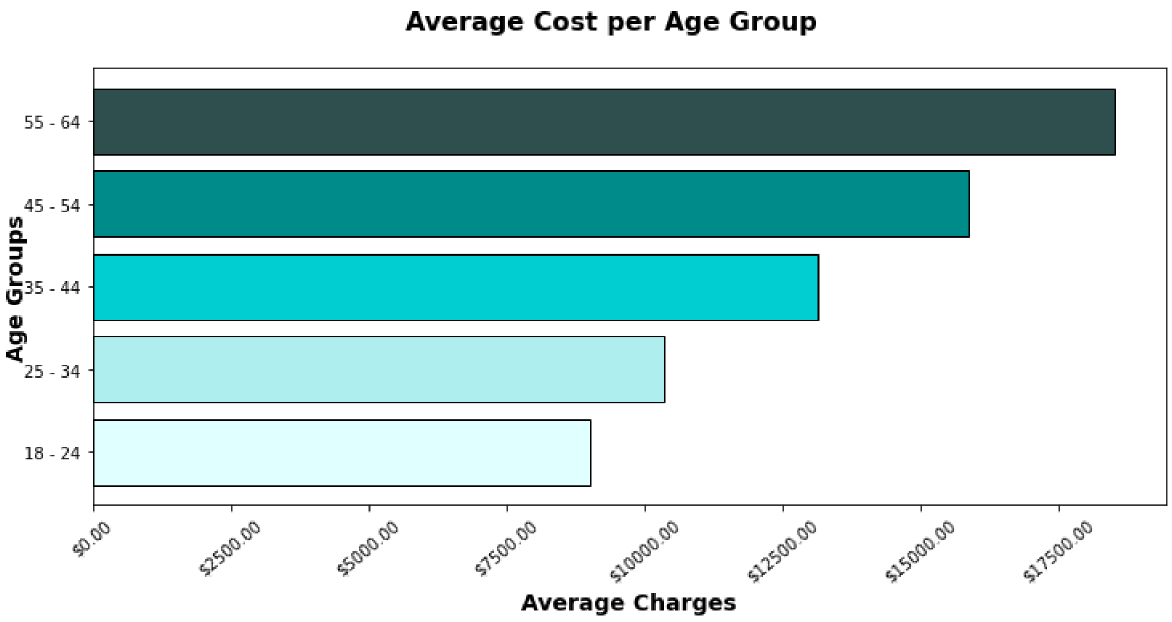
The range for ages is 46 years with a maximum of 64 and minimum of 18. Divide the patients into age groups and calculate the average charges for each group.

In Table 5, the median age group of 35 – 44 aligns closely with the average charges in the population. The average cost of insurance increases as age increases.

Table 5: Age Groups and Average Charges

Age Group (yrs)	Number in Group	Average Charges
18 - 24	278	$9011.34
25 - 34	271	$10352.39
35 - 44	260	$13134.17
45 - 54	287	$15853.93
55 - 64	242	$18513.28
Population Avg		
39.21	25	$13270.42

Figure 5: Average Cost of Insurance per Age Group



Average Cost per Number of Children

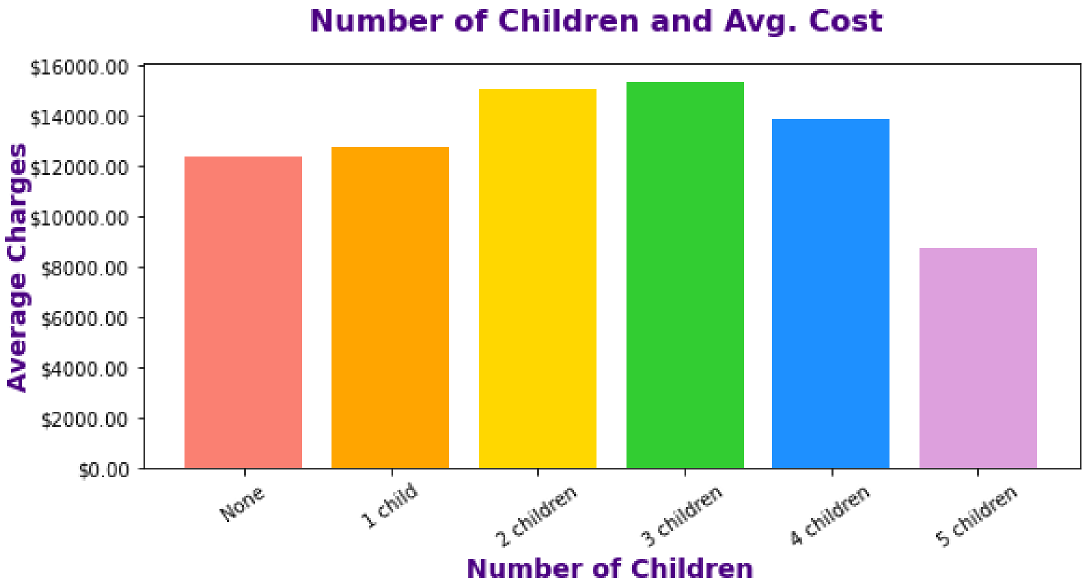
The range for number of children is 5 with close to half of the population having no children. Calculate the average charges per number of children.

Table 6 shows that the cost of insurance increases with each additional child up to 3, then begins to decrease. This is better visualized with the bar chart in Figure 5. However, there is not enough data to support an accurate analysis on number of children as there are only 25 patients with 4 children and 18 patients with 5 children. The sample is not large enough to make a deduction.

Table 6: Average Cost for Number of Children

Number of Children		Avg. Cost
No children	574	$12365.98
1 child	324	$12731.17
2 children	240	$15073.56
3 children	157	$15355.32
4 children	25	$13850.66
5 children	18	$8786.04
Average	1.09492	

Figure 6: Bar chart: Average Cost for Number of Children

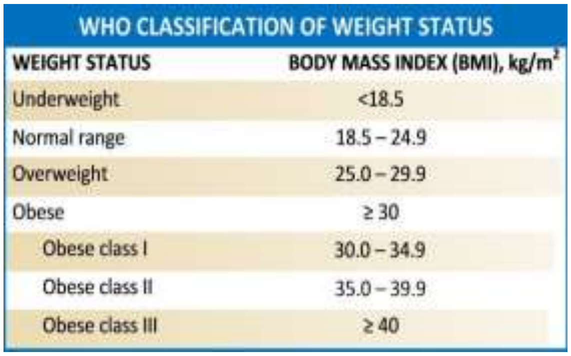


Average Cost per Weight Status

The WHO (World Health Organization) along with the CDC (Centers for Disease Control, U.S.) classify BMI into 6 weight status categories. BMI calculations are standardized and categorized for adults over the age of 19. The BMI for children and teens aged 5-19 is calculated similarly to adults but scaled based on standard deviation of the BMI measurement. For this project, I will include those measurements in my analysis as if they are in the adult age range of 20 years or older. What are the average charges per category?

Figure 7: WHO Classification of Weight Status

Table 7: BMI Weight Status and Average Cost



WHO CLASSIFICATION OF WEIGHT STATUS

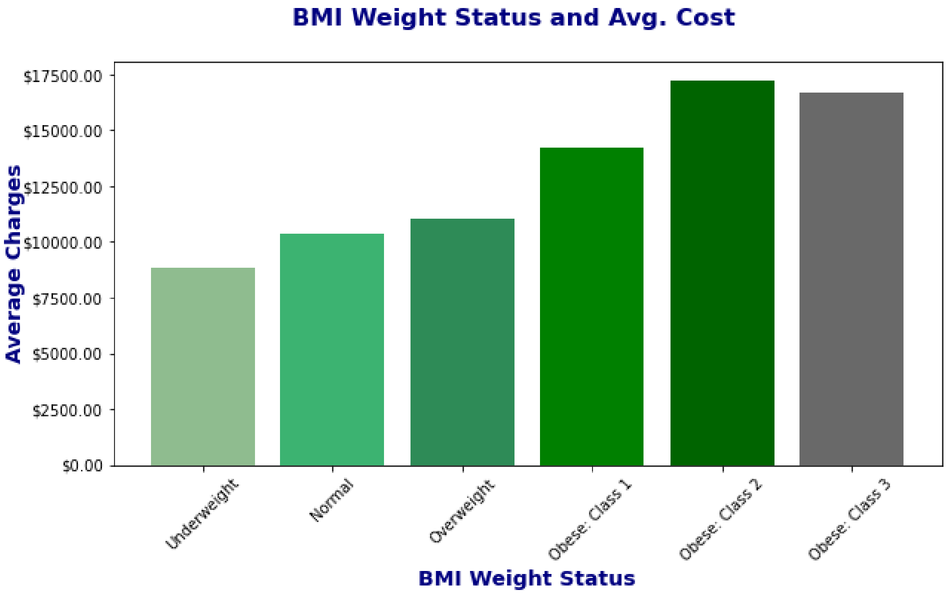
WEIGHT STATUS	BODY MASS INDEX (BMI), kg/m^2
Underweight	<18.5
Normal range	18.5 – 24.9
Overweight	25.0 – 29.9
Obese	≥ 30
Obese class I	30.0 – 34.9
Obese class II	35.0 – 39.9
Obese class III	≥ 40

(BMIChart)

BMI Category	Total	Avg. Cost	Cost Diff.
Underweight	20	$8852.20	$1527.30
Normal	222	$10379.50	-
Overweight	380	$11006.81	$627.31
Obese: Class 1	397	$14217.62	$3838.12
Obese: Class 2	226	$17245.41	$6865.91
Obese: Class 3	93	$16667.61	$6288.11
Average	30.66	$13270.42	

As expected, as BMI increases so do average charges. However, Table 7 shows that patients in the Obese: Class 3 category paid less than those in the Obese: Class 2 category.

Figure 8: Bar Chart: BMI Weight Status and Average Cost



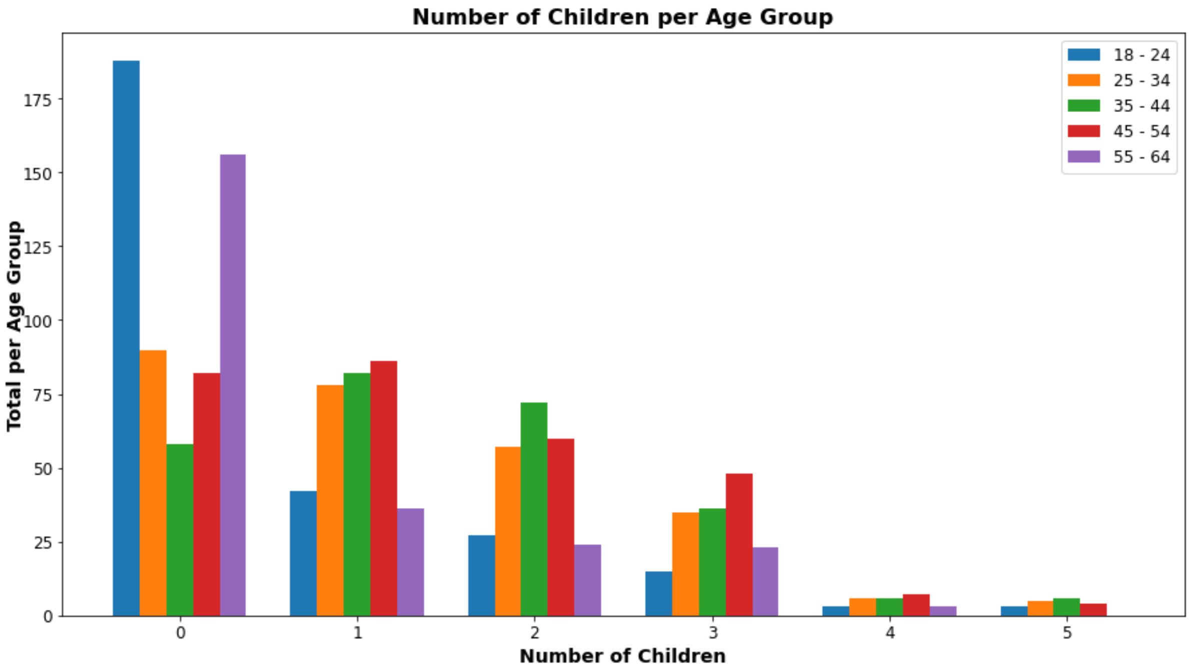
Number of children per age group

How are the number of children distributed among the population?

Table 8: Number of Children per Age Group

Age Groups	0	1	2	3	4	5	Avg	Total
18 - 24	188	42	27	15	3	3	0.6	168
25 - 34	90	78	57	35	6	5	1.28	346
35 - 44	58	82	72	36	6	6	1.49	388
45 - 54	82	86	60	48	7	4	1.39	398
55 - 64	156	36	24	23	3	0	0.68	165

Figure 9: Bar chart: Number of Children per Age Group



Patients ages 35 – 44 years old have the highest average number of children, but a lower total number of children than patients ages 45 – 54. The average and total number of children for patients ages 18 – 24 and 55 – 64 are similar.

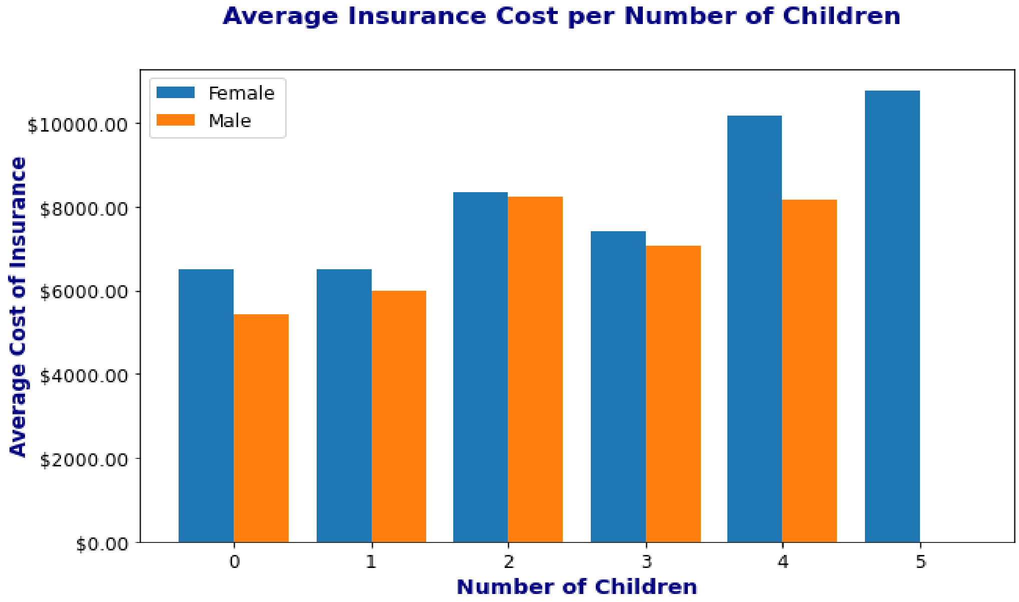
Affects of Number of Children and Insurance Cost

I am interested to see how having children impacts insurance costs for males and females.

Using the average BMI and Age in the population and filtering the data by smoker status the affects of number of children on insurance costs can be closely examined.

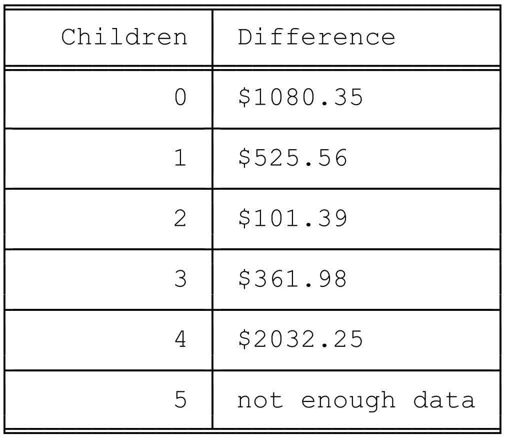
Average Age = 39.21 I will use my previously defined age group of 35 - 44 years old.
Average BMI = 30.66 This is between overweight and obese: class 1 categories, so I will use patients from both the overweight (25 - 29.9) and obese: class 1 (30 - 34.9) categories.

Figure 10: Bar chart: Average Insurance Cost per Number of Children



I expected males to be charged more for insurance overall, however it looks like females are charged more for insurance when considering the affects of number of children. Females with 2 children are charged an average of only $101.39 more than males with 2 children. While females with 4 children are charged an average of $2,032.25 more than males with 4 children.

Table 9: Difference in Cost per Number of Children



Children	Difference
0	$1080.35
1	$525.56
2	$101.39
3	$361.98
4	$2032.25
5	not enough data

Smoking Cessation Incentive Program

Smokers are charged an average of $23,615.96 more for insurance each year than nonsmokers. Insurance companies take great risk insuring smokers knowing the health issues that could arise from smoking. Creating an incentive program for smokers to quit smoking will lower risk for insurance companies, save smokers some money, and possibly lead them to a healthier lifestyle!

In previous Python projects we estimated insurance costs using the following formula:

estimated insurance cost:

250 * age - 128 * sex + 370 * bmi + 425 * num_of_children + 24000 * smoker – 12500

What happens if we decrease the factor for smoker by 1/3? How much money could smokers potentially save?

proposed estimated insurance cost formula:

250 * age - 128 * sex + 370 * bmi + 425 * num_of_children + 24000 - 1/3(24000) * smoker - 12500

Table 10: Difference in Estimated Costs after Smoker Incentive Program

Avg. Actual Cost	$32050.23
Avg. Estimated	$32889.58
Avg. Proposed	$24889.58
Avg. Estimated Savings	$8000.00
Diff in Actual v Proposed	$-7160.65

Smokers who complete a smoking cessation program and successfully quit smoking could save an average of $7,160.65 on their annual insurance costs with the new Smoker Incentive Program. This program could save smokers money and make them healthier and save insurance companies money by lowering their risk on insured smokers. Perhaps those smokers could further be incentivized to stay smoke free by lowering the variable for smokers in small increments for each year that they stay smoke free.

Conclusions

There are several variables that affect the cost of insurance. Some of those variables can be controlled like BMI weight status and whether a person chooses to smoke tobacco products. Some variables cannot be controlled such as age.

In my analysis of insurance costs, I found that the number of children a person has affects the amount they have to pay for insurance. I do not know what method was used to calculate the insurance costs in the dataset. Suppose the insurance costs in the dataset included insurance coverage for children? That might explain the discrepancy in costs of patients with children and patients without children. If the costs do not include coverage for children, then there are some issues with the method for calculating insurance costs. Females do not have an advantage in the estimated insurance cost formula (- sex factor *128, sex factor for females is 0) and it appears that they are charged more for the number of children they have than males. The question remains: why are the number of children a patient has factored into the cost of insurance if the total cost does not include the price of coverage for the children? More transparency on the method for calculating insurance costs is needed to further analyze this issue.

I created a scenario where an insurance company could incentivize smokers to complete a smoking cessation program to save them money and improve their health. The results show that smokers could save over $7,000 each year on insurance if they quit smoking. Insurance companies might lower their risk of insuring smokers if they implemented such a program. However, insurance companies might also stand to lose money by lowering the rates of smokers who complete the program. If insurance companies and doctors work together on the implementation of a smoking cessation program, then both entities might make up for their losses by taking a cut of the profits from the program.

Regional Statistical Data – extra practice

Table 11: Regional Statistics by Age

	count	mean	std	min	25%	50%	75%	max	mode	range
Northwest	325	39.1969	14.0516	19	26	39	51	64	19	45
Southwest	325	39.4554	13.9599	19	27	39	51	64	19	45
Southeast	364	38.9396	14.1646	18	26.75	39	51	64	18	46
Northeast	324	39.2685	14.069	18	27	39.5	51	64	18	46

Table 12: Regional Statistics by BMI

	count	mean	std	min	25%	50%	75%	max	mode	range
Northwest	325	29.1998	5.13676	17.385	25.745	28.88	32.775	42.94	28.31	25.555
Southwest	325	30.5966	5.69184	17.4	26.9	30.3	34.6	47.6	25.8	34.8
Southeast	364	33.356	6.47765	19.8	28.5725	33.33	37.8125	53.13	33.33	38.06
Northeast	324	29.1735	5.93751	15.96	24.8663	28.88	32.8937	48.07	32.3	32.11

Table 13: Regional Statistics by Num. Children

	count	mean	std	min	25%	50%	75%	max	mode	range
Northwest	325	1.14769	1.17183	0	0	1	2	5	0	5
Southwest	325	1.14154	1.27595	0	0	1	2	5	0	5
Southeast	364	1.04945	1.17728	0	0	1	2	5	0	5
Northeast	324	1.0463	1.19895	0	0	1	2	5	0	5

Table 14: Regional Statistics by Charges

	count	mean	std	min	25%	50%	75%	max	mode	range
Northwest	325	12417.6	11072.3	1621.34	4719.74	8965.8	14711.7	60021.4	1639.56	58400.1
Southwest	325	12346.9	11557.2	1241.57	4751.07	8798.59	13462.5	52590.8	1241.57	1242.26
Southeast	364	14735.4	13971.1	1121.87	4440.89	9294.13	19526.3	63770.4	1121.87	1131.51
Northeast	324	13406.4	11255.8	1694.8	5194.32	10057.7	16687.4	58571.1	1694.8	1702.46

Table 15: Population v Region Stats for Sex & Smoker

Population v Region: Sex				
Population	Northwest	Southwest	Southeast	Northeast
sex count 1338 unique 2 top male freq 676	sex count 325 unique 2 top female freq 164	sex count 325 unique 2 top male freq 163	sex count 364 unique 2 top male freq 189	sex count 324 unique 2 top male freq 163
Population v Region: Smoker				
Population	Northwest	Southwest	Southeast	Northeast
smoker count 1338 unique 2 top no freq 1064	smoker count 325 unique 2 top no freq 267	smoker count 325 unique 2 top no freq 267	smoker count 364 unique 2 top no freq 273	smoker count 324 unique 2 top no freq 257

References

(n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK535456/bin/bmi__WHO.jpg

Code Sources

Personal Notes & Jupyter Notebooks

https://docs.python.org/3/library/statistics.html

https://pypi.org/project/tabulate/

https://numpy.org/doc/stable/user/absolute_beginners.html#how-to-create-a-basic-array

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html

https://matplotlib.org/stable/gallery/statistics/boxplot_demo.html

https://matplotlib.org/stable/gallery/index.html

https://matplotlib.org/stable/gallery/statistics/boxplot_vs_violin.html#sphx-glr-gallery-statistics-boxplot-vs-violin-py

https://stackoverflow.com/questions/40278845/suppress-name-dtype-from-python-pandas-describe

https://stackoverflow.com/questions/48595445/is-it-possible-to-add-range-ie-max-min-to-the-pandas-describe-function-in-py

https://matplotlib.org/stable/gallery/pyplots/dollar_ticks.html

Review of Objectives

Personal Goals:

☑ 1. Organize data into usable variables for analysis.

I created variables for each column in the dataset and was able to call those variables when I needed to.

☑ 2. Learn how to present my findings in basic tables, charts, and graphs using tabulate and matplotlib.

The bulk of my time on this project was spent learning how to tabulate results and plot bar charts. I think I could have used a histogram for age groups or a density graph to show changes in charges over a certain variable, but overall, I am happy with the results of my effort and feel confident plotting bar charts and tabulating results.

☑ 3. Practice using Python class methods.

I think I did a good job using classes correctly in this project. I think creating new classes is a good way to analyze data, but I feel like there is a broader use of classes than what this project needed. Most of the analysis could be done with basic functions rather than classes.

☑ 4. Write clearly defined code using what I have learned and summarize findings.

I tried to vary using classes, functions, and for loops. I feel confident the code is easy to follow, however, I feel that I could do better with writing global functions that could be used outside of this project. I started to get the hang of it towards the end of my analysis, but I will focus on this for future learning goals.

Codecademy Project Objectives:

* Work locally on your own computer ☑

* Import a dataset into your program ☑

* Analyze a dataset by building out functions or class methods ☑

* Use libraries to assist in your analysis ☑

* Optional: Document and organize your findings ☑

* Optional: Make predictions about a dataset's features based on your findings. ☑