PA3\_report

Sidney Beck

George Mason University

AIT 580

Ray Hong

November 13, 2022

PA3\_report

In PA3\_1, I made 10 API calls to the Kaggle API; I found a multitude of data on healthcare using this method. All of my API calls were merely downloads to CSV. However, two were zip files I had to unzip. I had to create a Kaggle account to get the API JSON token to download on my computer for authentication; I will include this token in my zip file and delete it after class. Most of the CSV I downloaded from the Kaggle API were all state-based or missing a significant amount of data, and none seemed to have unique keys to join on or iterate over; However, three of my CSV files were different versions of the same original dataset, the Medical Costs Public dataset from a book written in 2013, called “Machine Learning with R by Brett Lantz (CHOI, 2017).” I picked the best of the three and shaped it to what I wanted. I found myself spending much time trying to combine things that didn’t go together, so in the end, I decided it was more efficient to use calculated columns to adjust the 2013 dataset from healthcare costs from 2013 to what healthcare costs would be in 2022 by changing the prices for inflation in a “for loop.” I researched the annual healthcare costs percentage per capita to create a total “Primary Care expenses 2022,” primary care, preventative care, and prescription columns. I summed all three to get the Total PPP Spending for all people in the dataset. The next step was to compare the loss or gain with the $149/month or $1788 annual total price that GoForward.com is currently charging to what people are being charged for PPP care in 2022through insurance or other systems. Finally, We created an age and BMI category and adjusted the Forward Pricing from 149 to tiered pricing by age group, BMI and smoker or not categories (older people charged a little more than younger, but not as much as 149/month, smokers assigned a multiple of two, and those with a low BMI pay a little less than those with a higher BMI). Suppose Forward switched to this pricing model and received government incentives like tax breaks. Next, I created two final calculated columns for gain or loss on the proposed price, one with actual costs and the last a Boolean column for yes they did benefit or no they did now on our new suggested prices.

My findings for PA3\_1 are that sometimes data just doesn’t fit well together, and you have to find other means to find insights. In my approach, when the data didn’t fit, I did more and better math to gain the insights I needed. Sometimes the best method is simple and right in front of you, but it isn’t easy to see.

For PA3\_2, I created a box, bar, pie, scatter, and histogram with the mean center and a density Plot. The box and bar chart uses color encoding for who benefited from the new proposed price and who didn’t benefit. The first set of box and bar graphs show column encoding for who is a smoker and who is not and clearly show smokers have higher medical bills and will benefit most from a reduction in PPP care spending. In contrast, the bar chart shows which regions spend more on the proposed price and how many people benefit using a blue and orange color encoding. The next set is a box and bar chat combo that shows the same thing, and only the y-axis encoding is changed to show total PPP expenses. The Histogram shows counts of benefits/costs from the proposed pricing for smokers with a mean line overlay; I was attempting to show smokers’ pricing on the new categorized price proposal; while some smokers’ expenses increase on the proposed pricing because of the BMI and age calculation, many other smokers decrease their spending per capita. The density plot clearly shows how original PPP expenses are much higher than PPP spending on our proposed BMI, age, and smoker-categorized pricing.

In conclusion, on our proposed pricing model, America could move from an overpriced insurance-based service healthcare system to a new “forward” thinking product-based healthcare system, improving PPP care at a reduced price. For Go Forward to adopt our pricing model, the U.S. government would have to incentivize new medical startups like Forward with the exemption. This would allow more people to afford PPP care through companies like Forward and eventually switch to digital product-based healthcare. This is the first step to moving America toward a universal healthcare system that will ultimately trump any other healthcare system in the world. Assuming startups like forwarding would grow their care to programs with specialty care provider networks Medicaid, Medicare, and CHIP, funds could morph into funding to support something new and better, digital Universal Healthcare for America. People would literally have access to their healthcare from a smartphone with offices to visit if needed. Most PPP care would go digital, and U.S. spending on PPP care (the largest portion next to hospital care) would drastically decrease. Another benefit would be with this new digital healthcare system, people would literally have a doctor in the palm of their hand, and all they would need is U.S. citizenship or a legal reason for being in America. It seems far-fetched, but something like this system could drastically change healthcare spending and health in America.

Hypothesis testing 1 (ANOVA With Repeated Measures of age, BMI, and Benefit from Proposed Pricing):

| Null hypothesis | Alternative hypothesis |
| --- | --- |
| There is no significant difference between the groups of the first factor, age and BMI (measurement repetition), to the dependent variable. | There is a significant difference between the groups of the first factor, age and BMI (measurement repetition), about the dependent variable. |
| There is no significant difference between the groups of the second factor, Benefit from Proposed Price Yes/No to the dependent variable. | There is a significant difference between the groups of the second factor, Benefit from Proposed Price Yes/No, about the dependent variable. |
| There is no interaction effect between the factor age and BMI and Benefit from the Proposed Price Yes/No. | There is an interaction effect between age and BMI and Benefit from the Proposed Price Yes/No. |

A two-factor analysis of variance with measurement repetition was performed to test whether there was a significant difference between the groups of the first factor," age and BMI " (repeated measures), concerning the dependent variable “Benefit from Proposed Price Yes/No.” There was a statistically significant difference between the groups of the second factor, “Benefits from Proposed Price Yes/No,” with the dependent variable. There is an interaction between the factors "age and BMI " and the “Benefit from the Proposed Price Yes/No” or the dependent variable.

* " Age and BMI " concerning the dependent variable, p=<.001
* The First factor, “Benefit from Proposed Price Yes/No,” concerning the dependent variable, p=<.001,
* There was an interaction between the variables “Benefit from Proposed Price Yes/No and " age and BMI " to the dependent variable, p=<.001.

|  | age | bmi | Total |
| --- | --- | --- | --- |
| True | 43.42 | 30.22 | 36.82 |
| False | 31.35 | 31.49 | 31.42 |
| Total | 39.21 | 30.66 | 34.94 |  | Sum of squares | df | Mean Squares | F | p |
|  |  |  |  | age, bmi | 48,819.17 | 1 | 48,819.17 | 556.27 | <.001 |
|  |  |  |  | Benefit from Proposed Price Yes/No | 17,734.57 | 1 | 17,734.57 | 156.29 | <.001 |
|  |  |  |  | A x B | 27,052.81 | 1 | 27,052.81 | 308.25 | <.001 |
|  |  |  |  | Between | 169,338.3 | 1,337 | 126.66 |  |  |
|  |  |  |  | Within the sample | 151,603.73 | 1,336 | 113.48 |  |  |
|  |  |  |  | Residuum | 117,250.43 | 1,336 | 87.76 |  |  |
|  |  |  |  | Within | 193,122.41 | 1,338 | 144.34 |  |  |
|  |  |  |  | Total | 362,460.71 | 2,675 | 135.5 |  |  |

Hypothesis Testing 2: Kruskal-Wallis Test (Age VS Total PPP Spending)

| Null hypothesis | Alternative hypothesis |
| --- | --- |
| There is no difference between the four categories (<18, 18-35, 36-50, 50+) of the independent variable Age Category in terms of the dependent variable Total 2022 PPP expenses. | There is a difference between the four categories of the independent variable, Age Category, in terms of the dependent variable Total 2022 PPP expenses. |

Chart, box and whisker chart

Description automatically generated

| Groups | N | Mean Rank |
| --- | --- | --- |
| 19 - 35 | 505 | 484.66 |
| 18 or younger | 69 | 327.35 |
| 36 - 50 | 408 | 724.71 |
| 51+ | 356 | 934.75 |
|  |  |  |
| Total | 1338 |  |

A Kruskal-Wallis test showed that there is a significant difference between the categories of the independent variable, Age Category, concerning the dependent variable Total 2022 PPP expenses, *p*=<.001. Thus, with the available data, the null hypothesis is rejected.

Post hoc Test

The Kruskal-Wallis test showed that there was a significant difference. A Dunn-Bonferroni test was used to compare the groups in pairs to determine which was significantly different.

The Dunn-Bonferroni test revealed that the pairwise group comparisons of 19 - 35 - 18 or younger, 19 - 35 - 36 - 50, 19 - 35 - 51+, 18 or younger - 36 - 50, 18 or younger - 51+ and 36 - 50 - 51+ have an adjusted p-value less than 0.05. Thus, based on the available data, it can be assumed that these groups are significantly different in each significantly different pair.

Hypothesis Test #3 Chi-Squared for Smoker and Region.

Null: There is no relationship

Alternate: There is a statistically significant relationship between smokers and region.

Each cell in the expected observations table has five or more observations, thus fulfilling the assumptions for the Chi2 test.

Expected:

|  |  | region | | | |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | southwest | southeast | northwest | northeast | Total |
| smoker | True | 66.55 | 74.54 | 66.55 | 66.35 | 274 |
|  | False | 258.45 | 289.46 | 258.45 | 257.65 | 1064 |
|  | Total | 325 | 364 | 325 | 324 | 1338 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Observed:

|  |  | region | | | |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | southwest | southeast | northwest | northeast | Total |
| smoker | True | 58 | 91 | 58 | 67 | 274 |
|  | False | 267 | 273 | 267 | 257 | 1064 |
|  | Total | 325 | 364 | 325 | 324 | 1338 |

Results: A Chi2 test was performed between the smoker and the region. No expected cell frequencies were less than 5. There was no statistically significant relationship between smoker and region, χ²(3) = 7.34, p = .062, Cramer’s V = 0.07

This results in a p-value of .062, which is above the defined significance level of 5%. The Chi2 test is, therefore, not significant, and the null hypothesis is confirmed.

References

CHOI, M. (2017, November 1). Medical Cost Personal Datasets. Kaggle.com. Retrieved November 13, 2022, from <https://www.kaggle.com/datasets/mirichoi0218/insurance>

GoForward.com. (n.d.). GoForward.com. Retrieved November 13, 2022, from <https://goforward.com/why-forward>