CITADEL | CITADEL | Securities

APAC Datathon Report

Kim Hyun Bin, Yeo Fu Kai Marcus, Agrawal Naman, Jasraj Singh

April 16, 2022

Contents

1. Guiding Question

In 2018, in the United States, 1,708,921 people were diagnosed with cancer, and the number of deaths was recorded at 599,265, as reported by Centers for Disease Control and Prevention [1] (CDC). Cancer treatments are extremely expensive and most people require monetary support to undergo chemotherapy. This motivated us to propose the following topic question - are the payments from Medicaid and Medicare doing a fair and efficient job of allocating resources.

It is important to attack these questions in order to help steer policy making towards eliminating the factors responsible for Cancer infection as far as possible, and limiting their impact thereon. They can also help the state governments in rebalancing the budget to accommodate economic disparities in the population and reach more people in need. Specifically, we explore the following questions -

1. Is any state an outliers in terms of total amount paid? Are they outliers in terms of value and volume? How do their payment details reflect that divergence?

2. Is there any correlation between the death rate due to Cancer and the amount of money claimed through Medicaid for it?

3. Which states have the highest rate of Cancer infection, and how is it related to different factors determining quality of life?

We hope these questions can help improve the support for Cancer patients and make treatment opportunities more equitable.

2. Background

In 2018, for every 100,000 people in the population, 436 new cases were reported and 149 people died of cancer.

Causes of Cancer include, but are not limited to, tobacco use, obesity, poor diet, lack of physical activity, excessive drinking of alcohol, exposure to ionizing radiation and environmental pollutants.[2] This implies that quality of life is an important determinant of Cancer infection rate in a region.

As of 2021, no type of cancer has a cure. However, there are treatments for cancer patients.[3] Our improved understanding of molecular biology and cellular biology due to cancer research has led to new treatments for it. Since 1971, the US has spent over $200 billion on cancer research.[2] The quality of treatments has improved overtime, and the cancer death rate (adjusting for size and age of the population) declined by five percent between 1950 and 2005[2], and by 27%, from 196.5 to 144.1 deaths per 100,000 population, between 2001 and 2020.[4]

Despite advances in treatments, there are still unexplained disparities in terms of death rates between states[5] (see Figure 1). These differences can be due to a multitude of factors, including state healthcare expenditure, quality of life, genetics, etc. For states with high death rates, it is important to recognize the causes behind these numbers and make policy changes with the goal of countering these causes.

In this report, we will be discussing our study on investigating whether Medicaid & Medicare spending on cancer reaps its benefits of a low cancer mortality rate and look at other possible factors that could be contributing to the cancer mortality rate such as political, economic and socio-cultural determinants.

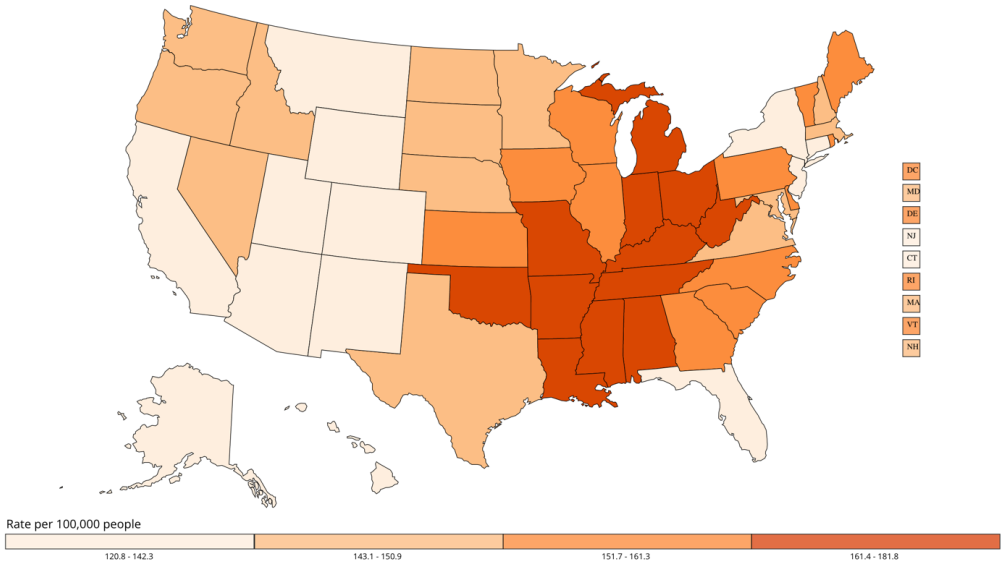

Figure 1: Death rate per 100,000 people by the state (2018). Source: Cancer Today

3. Executive Summary

We have produced an significant set of actionable insights for the Centers for Medicare & Medicaid Services (CMS) through the use of general payment data as well as social-economic indicators in the United States, with the hopes of ensuring justifiable healthcare treatment subsidies or payouts with respect to cancer treatment across different states.

Our recommendation is to extract distinct outliers in the open payment system through the use of a multivariable anomaly detection algorithm, the Isolation Forest model. This would be able to determine the over or under allocation of funds regarding different treatments. Coupled with the analysis of trends with social-economic factors, it is crucial for the government to distribute financial resources to states that are in need.

In our report, we report our findings of a relatively significant flow of payments to the state of California, mostly for cancer related medication and treatments. Contrastingly, the states with lower payments with respect to cancer suffer from high cancer mortality rates, unlike the state of California. We then proceeded on to investigate whether correlation implies causation in this scenario by employing external socio-economic features. Our findings suggest that there is a weak correlation between individual external factors, which could be a strong indication that our initial finding of correlation between cancer medical spending and cancer mortality rates does imply causation.

4. Technical Exposition

4.1 Model Architecture

Isolation forest is an anomaly detection algorithm that can be used to identify outliers in a data set by deriving an anomaly score for each data point. It's an unsupervised algorithm, and therefore does not rely on the pre-labelling of data points as normal and outliers. Unlike several outlier detection algorithms that work by profiling normal points (such as DBScan Clustering and Boxplot analysis), isolation forests explicitly identify anomalies by isolating the outliers through the construction of an ensemble of proper binary trees that are based on the notion of density. The process of isolating data points involves the randomized selection of a feature and an initialization of a split between the maximum and minimum value of the selected feature (a random hyperplane, along one of the variable axes, is introduced in the multidimensional feature space containing the data points). The process is repeated till all the data points are isolated from one another.

The algorithm computes the number of partitions required to isolate every point in the feature space. Since the outliers are more likely to stay away from other data points, they are more susceptible to being isolated faster. This is quantified using the path length $h(x)$ of a point x, measured by the number of edges x traverses an isolation tree from the root node until the traversal is terminated at an external node. [6] Since outliers require a lesser number of partitions to isolate them, they are likely to have a lower path length compared to other data points. Once an isolation forest has been constructed, the anomaly score of point x in a data set of n observations can be calculated as follows:

$$\alpha(x, n) = 2^{\frac{E[h(x)]}{c(n)}}$$

Where $h(x)$ is the path length of point x, $E(h(x))$ is the average search height of x from all the isolation trees constructed, and $c(n)$ is the average value of $h(x)$, that is the mean path length required to find any general node across all the trees (independent of x). Outliers tend to have lower $h(x)$ and consequently higher anomaly score. Instances that have an anomaly score very close to one are regarded as perfect anomalies, while those with a score much less than 0.5 can be safely assumed as normal points:

$$E[h(x)] \gg c(n) \implies \alpha(x, n) \approx 1$$
$$E[h(x)] \approx c(n) \implies \alpha(x, n) \approx 0.5$$

Among the myriad outlier detection algorithms available for our analysis, we decided to pursue isolation forests because of several reasons:

1. Supports multidimensional outlier analysis: Allows us to easily find outliers in a multidimensional space (unlike methods such as z-score computations, boxplot analysis and standard deviation calculations).
2. Does not require any assumptions about the prior distribution of the data set (unlike algorithms such as minimum covariance determination, which require the feature space to be a multi-dimensional Gaussian distribution).
3. Ability to handle big data (unlike algorithms such as DBScan Clustering) with high processing capabilities. [6]
4. Even though the binary trees are constructed based on the notion of density, the algorithm eliminates any major computational cost arising from determining the density or distance between points. [6]
5. Linear time complexity with a low constant and low memory requirement (compared to many other methods such as Local Outlier Factor analysis or One Class Support Vector Machines). [6]

4.2 Modelling and Analysis

When it comes to the Isolation Forest Algorithm, the question that we need to address is the contamination hyperparameter which is one of the most important parameters. The contamination is defined as 'the amount of contamination of the data set, i.e. *the proportion of outliers in the data set'*.
As we can see, the general payments dataset's *'Total_Amount_of_Payment_USDollars'* feature is spread out over a wide range, with the median being only 17 dollars and the upper quartile being 25 dollars despite the fact that the maximum value is 1.6 million dollars. Now in order to obtain an objective contamination rate, we initially proceeded on with attempting to remove some of the extreme values to see whether the curve resembles a normal distribution. This was done by removing different percentages from the dataset and using a statistical analysis, based on D'Agostino and Pearson's, to test the null hypothesis that the remaining sample comes from a normal distribution.

The p-value was obtained from a 2-sided chi-squared (χ^2) test for the hypothesis and the statistic is a value derived from the calculation $s^2 + k^2$, where s is the z-score returned by skewtest and k is the z-score returned by kurtosistest.

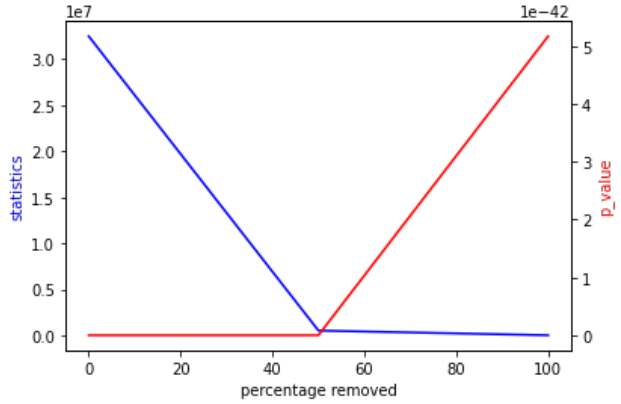

Figure 2: Testing for Gaussian Distribution

Table 1: Testing our null hypothesis that it's a normal distribution

Percentage_Removed	0%	50%	99.9995%
Statistic	32508096.00	508956.88	407.73
p_value	0.000000e+00	0.000000e+00	5.170398e-42

As observed, even after removing a substantial portion of the dataset, we were not able to show that the distribution comes from a normal distribution.

We realized this may be due to the high standard deviation of the dataset in the first place, 2.07^4, even though the mean is a mere value of 300. Therefore, we decided that attempting to obtain an objective contamination rate through statistical analysis was inappropriate for this instance.

With the knowledge that this dataset contains some extreme values, we decided to only observe the top 1% of all payments in the general dataset. The motivation behind this decision was that by observing and investigating the top 1% of payments, we would be able to pick out the payments that were the most influential and observe any disparity in treatments and medications.

Along with our contamination rate as 1%, we've also decided to utilize multi-dimensional inputs to our isolation forest model. This was to ensure that our model was able to obtain the outliers after they observed multiple features that we deemed relevant. Out of the general payments dataset, we chose [*'Recipient_State', 'Physician_Primary_Type', 'Physician_Specialty', 'Total_Amount_of_Payment_USDollars',*

'Number_of_Payments_in_Total_Amount'] as the relevant features for our model to consider. We believed that any outliers would be heavily influenced by what state the recipient was in, their respective practicing industry/nature and the amount of monetary value that was flowing through. After label encoding our dataset appropriately and passing into our model, we obtained 57656 rows of data. We were interested in checking out to see which states appeared the most as an outlier.

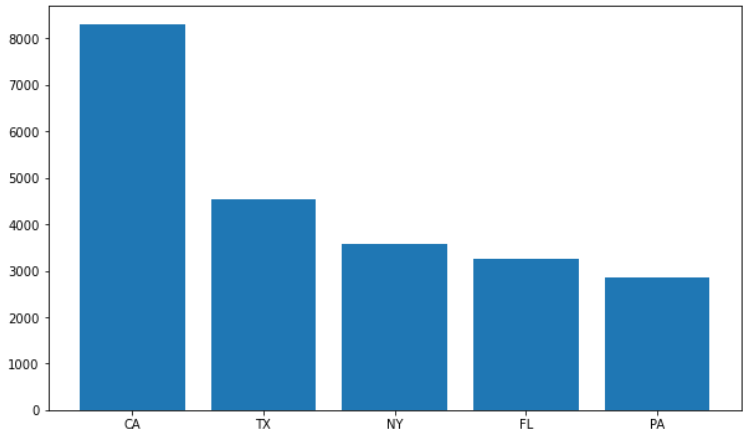

Figure 3: Total amount of payment on different products in California

Table 2: States with the greatest number of outliers

Recipient_State	No. of Outliers
CA	8299
TX	4356
NY	3563
FL	3263
PA	2848

Surprisingly, California was experiencing twice as much inflow as the next highest state, Texas, 8 times the mean value (1067.5) and away from it by about 5 standard deviations.

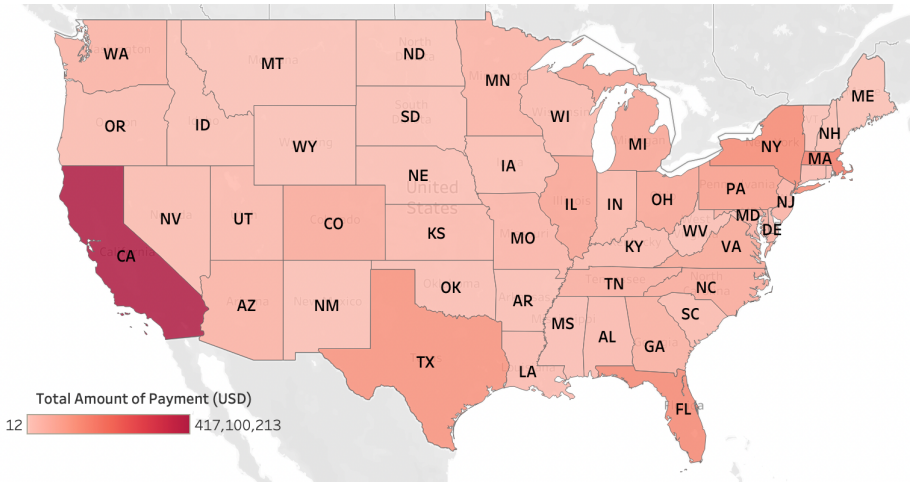

Figure 4: California with a very large total amount of payment

It is made even clearer through this visualization that California is experiencing an extremely large amount of payment, compared to the rest of the states.

Digging deeper into California, we went on to investigate the nature behind the large amount of payments and we decided to look into the individual products money was being spent on. We managed to clean and engineer a dataset such that the last few columns containing majority null values were all combined and sorted by the amount of money spent.

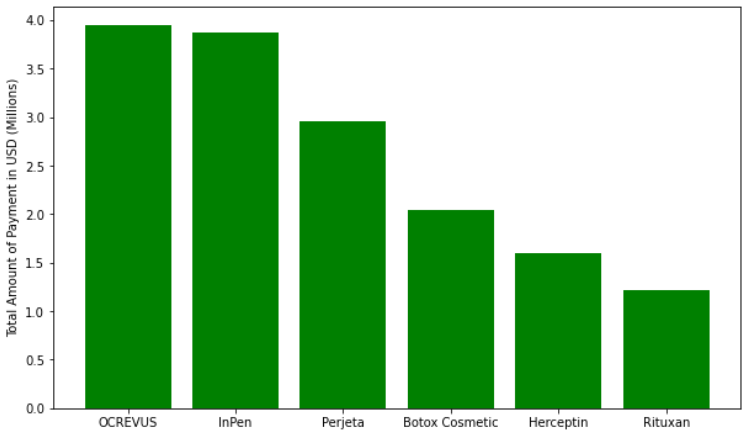

Figure 5: States having the most outliers

Table 3: Name of products California has spent most on

Product_Name	**Total_Amount_of_Payment_USDollars**
OCREVUS	39461440
InPen	38779320
Perjeta	29634030
BOTOX COSMETIC	20464440
Herceptin	15934570
Rituxan	12115960

These are the names of the products most money was spent on and after searching these medications individually on the web, we found out that Ocrevus, Perjeta, Herceptin, and Rituxan are medications related to cancer treatments, Chemotherapy, and immunosuppressive drugs. Thus, the total amount of payment related to cancer was roughly 96 million dollars in total, while money spent for treatment related to diabetes was 38 million dollars and cosmetics were 20 million dollars.

With this fact, the question that came to our minds was, with such high spending, what is the cancer situation in California. With this question in our minds, we utilized some external datasets to observe the effects of such high spending on cancer.[7]

As it turns out, California had the 7th lowest cancer mortality rate as shown below.

Table 4: States in order of increasing cancer mortality rates

State	Cancer Mortality Rate
UT	119.5
HI	123.8
CO	127.2
AZ	127.7
NY	128.8
NM	129.8
CA	130.3
NJ	133.4
CT	133.8
MA	135.2

We wanted to check whether the opposite was true. We were able to find out that Kentucky had the highest mortality rate and searched for which products Kentucky spent the most amount of money on.

Table 5: Name of products Kentucky has spent most on

Product	Total_Amount_of_Payment_USDollars
REUNION	1075485.0
Bridle	836033.3
Aptis DRUJ	735075.2
MAKO	351881.8
Da Vinci Surgical System	343442.8

As we can see, the medications that had the highest amount of payment are very diverse. Reunion is a medication related to bone problems, Bridle is related to a respiratory problem, and Aptis DRUJ and MAKO are related to joint problems.

This strongly suggests that our hypothesis that the opposite/contrary is true, at least for the states that we observed, which had one of the lowest and the other having the highest cancer mortality rates in the United States.

However, we believe there is a deeper analysis to be done here as the amount of money spent on cancer might not be the only contributing factor to the cancer mortality rate figures.

4.3 External Data Analysis

An important factor which may be associated with the cancer mortality rate is the quality of life. While it is undeniable that the factors constituting the quality of life are subjective, we shortlisted some features such as education level, economic level, healthcare, poverty rate and physical health for our evaluation.

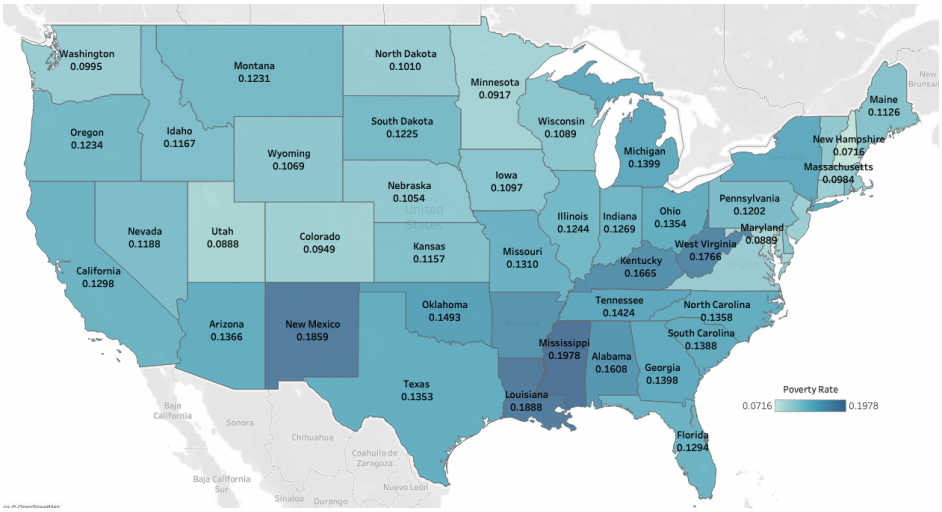

Figure 6: Poverty Rate demographics in U.S.

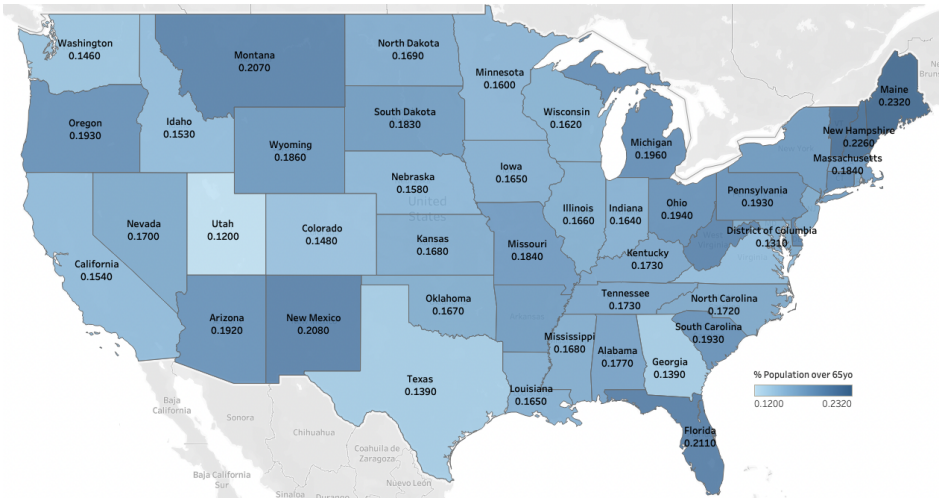

Figure 7: Elderly population (over 65 years old) demographics in U.S

To better understand individual states, we decided to granularize our analysis. Using the various quality of life and poverty rate statistics provided by the World Population Review [8], we want to find out contributory factors for the cancer mortality rates among the different states. From Figure 6, Kentucky (high cancer mortality rate) has a relatively higher poverty rate as compared to California (low cancer mortality rate). This could be attributed to why Kentucky spent more on inexpensive healthcare treatments such as joins or respiratory issues rather than investing in cancer treatments which explains the high cancer mortality rate.

In addition, from Figure 7, the elderly population in Kentucky (17.3%) is slightly higher than in California (15.4%), which could translate to a higher mortality rate for cancer since the elderly are more likely to develop cancer as compared to the younger population.

To uncover the relationship between features, we also decided to perform Exploratory Factor Analysis (EFA) on cleaned data to understand if any other features might explain the mortality rate for cancer apart from total expenditure on its treatments.

	State	lifeQualityRank	healthCareRank	educationRank	economyRank
0	Washington	1	4	4	3
1	New Hampshire	2	16	5	13
2	Minnesota	3	10	17	18
3	Utah	4	9	10	2
4	Vermont	5	11	8	29
5	Maryland	6	8	13	26
6	Virginia	7	18	7	25
7	Massachusetts	8	2	1	7
8	Nebraska	9	27	6	21
9	Colorado	10	12	11	1
10	Wisconsin	11	14	14	24

Figure 8: Discrete values of quality of life rankings according to states

Out of the Pearson, Spearman and Kendall correlation, we decided to use the Kendall correlation since it assumes that the features need not follow a predetermined distribution, unlike the normal distribution for Pearson correlation, which is beneficial in our case as our feature variables have a very small p-value (7.1^{-18}), suggesting stronger evidence for the alternative hypothesis. Moreover, Kendall correlation is more tolerant and lenient of outliers in its feature variables. Kendall correlation (τ) is calculated as follows:

$$\tau = \frac{\text{Number of Concordant Pairs} - \text{Number of Discordant Pairs}}{\binom{n}{2}}$$

Upon closer inspection of Figures 9 and 10, the result shows that these features could possibly not have a direct correlation with the cancer mortality rate, since the highest score is only 0.46 which is linked to the obesity rate as part of physical health.

Granted, while the features might not show a strong bivariate correlation, there is still a possibility of a strong association in regression once other features are controlled for.

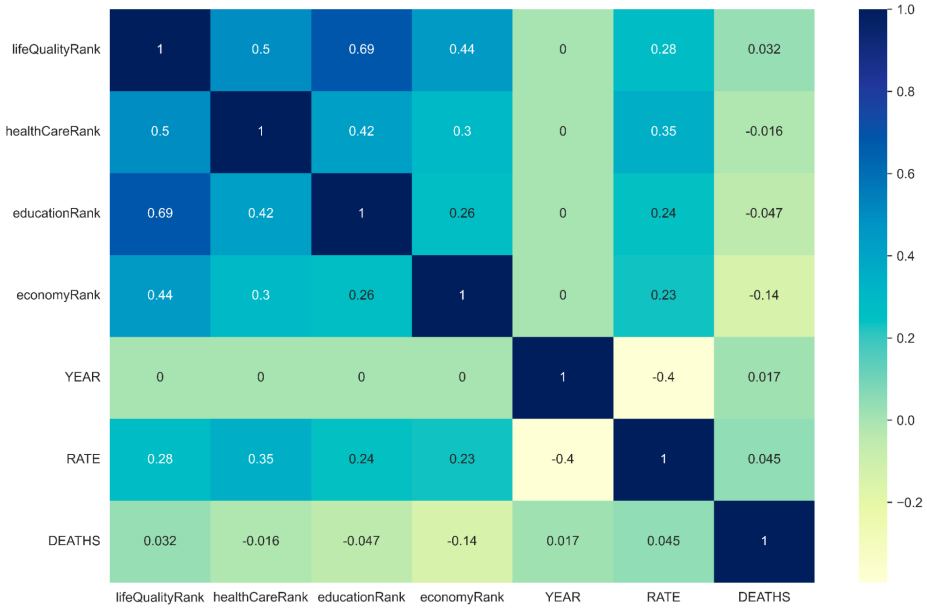

Figure 9: Correlation of life quality rankings and cancer mortality rates (in U.S.)

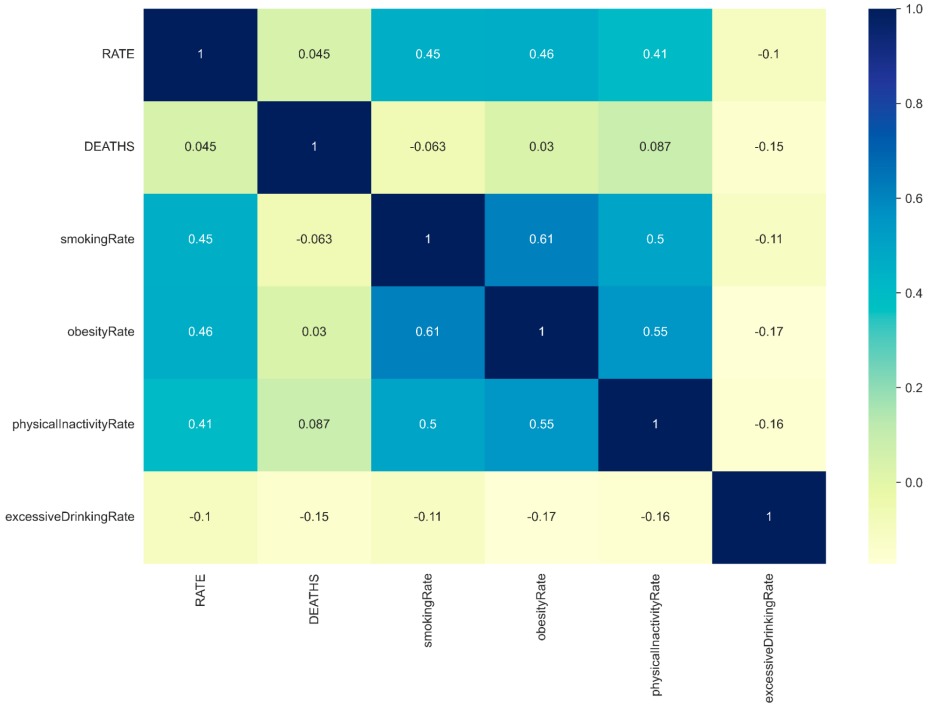

Figure 10: Correlation of physical health and cancer mortality rates (in U.S.)

5. Conclusion

As we observed, there is a weak correlation between various socio-economic features and cancer mortality rates. We acknowledge that investigating further with the help of a multivariate regression analysis could have yielded a more concrete result and would like to leave this part as a potential area for future further research.

Thus far, we have spotted California as a strong outlier, especially in cancer-related issues. Digging deeper and accounting for other states, we found a strong linear relationship between the large spending and cancer mortality rates. Given that external factors don't have such a strong correlation with cancer mortality rates, we believe the magnitude of payments and interest in cancer-related treatments greatly determine and influence the cancer mortality rates in that region.

Therefore, we believe this to be a significant discovery for the whole country so that relevant parties can pay heed to this issue of inequality and divergence and hopefully address it in the near future.

6. References

[1] Centers for Disease Control and Prevention. (2021, June 8). Cancer Data and statistics. Centers for Disease Control and Prevention. Retrieved April 17, 2022, from https://www.cdc.gov/cancer/dcpc/data/index.htm#:~:text=In%20the%20United%20States%20in,which%20incidence%20data%20are%20available

[2] Wikimedia Foundation. (2022, April 10). Cancer. Wikipedia. Retrieved April 17, 2022, from https://en.wikipedia.org/wiki/Cancer

[3] Bonvissuto, D. (n.d.). Is there a cure for cancer? WebMD. Retrieved April 17, 2022, from https://www.webmd.com/cancer/guide/cure-for-cancer#

[4] National Center for Health Statistics, National Vital Statistics System, Mortality Data . Retrieved April 17, 2022, from https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm

[5] U.S. Cancer Statistics Working Group. U.S. Cancer Statistics Data Visualizations Tool, based on 2020 submission data (1999-2018): U.S. Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute; www.cdc.gov/cancer/dataviz, released in June 2021.

[6] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 413–422). IEEE.

[7] National Center for Health Statistics. (2022, February 28). Stats of the States - Cancer Mortality. Centers for Disease Control and Prevention. Retrieved April 17, 2022, from https://www.cdc.gov/nchs/pressroom/sosmap/cancer_mortality/cancer.htm

[8] US states - ranked by population 2022. (n.d.). Retrieved April 17, 2022, from https://worldpopulationreview.com/states