Does national health care spending correlate to patient satisfaction and ranking of hospitals in the U.S?

The American Hospital Association (AHA), a national organization that represents hospitals and their patients, and acts as a source of information on health care issues and trends. The dataset is the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey results for the last 9 years provided by Maven Analytics.

The surveys contains questions to evaluate the following measures:

```
Communication with Nurses - H_COMP_1
Communication with Doctors - H_COMP_2
Responsiveness of Hospital Staff - H_COMP_3
Communication about Medicines - H_COMP_5
Discharge Information - H_COMP_6
Care Transition - H_COMP_7
Cleanliness of Hospital Environment - H_CLEAN_HSP
Quietness of Hospital Environment - H_QUITE_HSP
Overall Hospital Rating - H_HSP_RATING
Willingness to Recommend the Hospital - H_RECMND
```

Purpose of analysis

The purpose of this report is to test the hypothesis if the national spendig on healthcare influence the ranking of the hospitals. To test the hypothesis, the inflation adjusted national health expentditure from 1970 to 2021 from Peterson-KFF Health System Tracker sourced from the National Health Expenditure (NHE). This data was chosen in particular because the expenditure was recorded on a per capita basis and inflation adjusted which makes the year or year comparision equal.

Assumptions

The assumption in the analysis is that the national health spend is for the calendar year. The periods the surveys were reported on was from October to September the following year. The assumption is made to help match the spending and survey period.

In [1]:
```
# importing libraries and data
import pandas as pd
import numpy as np
from matplotlib import pyplot

health_data = pd.read_csv('./GOOD_DATA/final/data-PhiPo.csv')
national_results = pd.read_csv('./HCAHPS+Patient+Survey/data_tables/national_results.csv')
```

Approach

The approach taken to test the hypothesis is to filter the health expenditure data to start from 2014 in order to align with the survey periods which start from October 2013. The survey results in the format

```
Bottom-box Answer   Middle-box Answer   Top-box Answer
Sometimes or never   Usually             Always
```

Each record of the national survey results is represented in percentage and sums up to 100%. Therefore, "Top-box Answer" of the result is chosen to test the hypothesis. The dataset is cleaned to merge with the expenditure data.

In [2]:
```
# filtering health spend data to start from 2014 since the survey periods start from October 20
health_data = health_data[health_data.Year>=2014]
health_data
```

Out[2]:

	Year	Total national health expenditures	Constant 2021 dollars
44	2014	3002.6	3374.3
45	2015	3165.4	3549.3
46	2016	3307.4	3671.7
47	2017	3446.5	3757.4
48	2018	3604.4	3847.5
49	2019	3757.4	3951.8
50	2020	4144.1	4311.0
51	2021	4255.1	4255.1

In [2]:
```
# preparing survey data
results = national_results
results['Release Period'] = results['Release Period'].replace({'07_' : ''}, regex=True)
results['Release Period'] = results['Release Period'].astype(int)
results['Release Period'] = results['Release Period'] -1
results = results[(results['Release Period']>=2014) & (results['Release Period']<=2021)]
results
```

Out[2]:

	Release Period	Measure ID	Bottom-box Percentage	Middle-box Percentage	Top-box Percentage
0	2014	H_CLEAN_HSP	8	18	74
1	2014	H_COMP_1	4	17	79
2	2014	H_COMP_2	4	14	82
3	2014	H_COMP_3	9	23	68
4	2014	H_COMP_5	18	17	65
...	...	...	...	...	...
75	2021	H_COMP_6	14	0	86
76	2021	H_COMP_7	6	42	52
77	2021	H_HSP_RATING	8	20	72
78	2021	H_QUIET_HSP	10	27	63
79	2021	H_RECMND	6	23	71

80 rows × 5 columns

Hypothesis Testing

The hypothesis will be testing if there is a relationship between 2 continous variables - the national health expenditure and the overall national ranking in the US. For this hypothesis testing, Spearman's Correlation test will be used to test the correlation between as the variables' distribution is not normal and Spearman's Correlation assumes non-Gaussian distribution.

In addition to the overall hospital ranking, the realtionship between the health expenditure and each of the measure is tested to prove if there is a statistically significant relationship between the ranking and the evaluation measures.

Hypothesis

```
Null Hypothesis: There is no correlation between the 2 variables
Alternate Hypothesis: There is a linear relationship between the 2 variables
i.e, the measure increases and/or decreases with the health spend.
```

In [13]:

```
from scipy.stats import spearmanr

# list the measures to iterate to get the Spearman correlation for each of the measure
measures = ['H_COMP_1','H_COMP_2','H_COMP_3','H_COMP_5','H_COMP_6','H_COMP_7','H_CLEAN_HSP','H_

for measure in measures :
    measure_df = results[results['Measure ID'] == measure]

    df = pd.merge(health_data, measure_df, how='inner',  left_on='Year', right_on='Release Peri
    df = df[['Year', 'Constant 2021 dollars', 'Top-box Percentage']]

    x1 = df[['Constant 2021 dollars']]
    x2 = df[['Top-box Percentage']]

    corr, _ = spearmanr(x1, x2)
    print('Spearman correlation for measure ' + measure + ' is ' + str(round(corr,3)) )
    print('Spearman p value for measure ' + measure + ' is ' + str(round(_,3)) )

    if(_<0.05):
        print(measure)
    print("\n")
```

```
Spearman correlation for measure H_COMP_1 is 0.73
Spearman p value for measure H_COMP_1 is 0.04
H_COMP_1


Spearman correlation for measure H_COMP_2 is -0.405
Spearman p value for measure H_COMP_2 is 0.319


Spearman correlation for measure H_COMP_3 is 0.358
Spearman p value for measure H_COMP_3 is 0.385


Spearman correlation for measure H_COMP_5 is 0.326
Spearman p value for measure H_COMP_5 is 0.431


Spearman correlation for measure H_COMP_6 is 0.126
Spearman p value for measure H_COMP_6 is 0.766


Spearman correlation for measure H_COMP_7 is 0.643
Spearman p value for measure H_COMP_7 is 0.086


Spearman correlation for measure H_CLEAN_HSP is 0.432
Spearman p value for measure H_CLEAN_HSP is 0.285


Spearman correlation for measure H_QUIET_HSP is 0.126
Spearman p value for measure H_QUIET_HSP is 0.766


Spearman correlation for measure H_HSP_RATING is 0.481
```

```
Spearman p value for measure H_HSP_RATING is 0.227


Spearman correlation for measure H_RECMND is 0.394
Spearman p value for measure H_RECMND is 0.334
```

The Spearman's test's p-value for each of the measure, expcept for H_COMP_1, are greater than 0.05 indicating that we fail to reject the null hypothesis. Therefore, there is no correlation between those measures and the healthcare spendings.

Since the p value score for H_COMP_1 is less than 0.05, we reject the null hypothesis and conclude that there is a statistically significant evidence to show that there is a linear positive relationship between the national healthcare spending and the communication with nurses and the patients, i.e, more health care spend increases the satisfaction of patient's communication with nurses.

We can assume that since the nation spends more on the health care and hospitals, more nurses are hired with the right qualifications who can promptly attend the patients which ultimately leads to good communication between patiendts and nurses but does not increase the ranking of the hospitals.

Further analysis

To understand if the measures are correlated with each other, the Spearman's correlation test is conducted within measures to test the hypothesis.

In [4]:

```
measures = ['H_COMP_1', 'H_COMP_2', 'H_COMP_3', 'H_COMP_5', 'H_COMP_6', 'H_COMP_7',
            'H_CLEAN_HSP', 'H_QUIET_HSP', 'H_HSP_RATING', 'H_RECMND']

for m1 in range(0, len(measures)):
    for m2 in range(m1):

        m1_df = results[results['Measure ID'] == measures[m1]]
        m1_df = m1_df.rename(columns={"Top-box Percentage": "M1"})
        m2_df = results[results['Measure ID'] == measures[m2]]
        m2_df = m2_df.rename(columns={"Top-box Percentage": "M2"})
        df = pd.merge(m1_df, m2_df, how='inner',  on='Release Period')
        df = df[['Release Period', 'M1', 'M2']]

        x1 = df[['M1']]
        x2 = df[['M2']]

        corr, _ = spearmanr(x1, x2)

        # show the measures that have statistical significant relationship.


        if _ <= 0.05:
            print("p value less than 0.05")
            print('Speaman correlation for ' +  measures[m2] + ' and ' +  measures[m1] + ' is '
            print('Spearman p value for for ' +  measures[m2] + ' and ' +  measures[m1] + ' is
            print("\n")


#         if corr < 0.05  :
#             print("correlation less than 0")
#             print('Speaman correlation for ' +  measures[m2] + ' and ' +  measures[m1] + ' is
#             print('Spearman p value for for ' +  measures[m2] + ' and ' +  measures[m1] + ' i
#             print("\n")
```

```
p value less than 0.05
Speaman correlation for H_COMP_1 and H_COMP_3 is 0.727
Spearman p value for for H_COMP_1 and H_COMP_3 is 0.041
```

p value less than 0.05
Speaman correlation for H_COMP_1 and H_COMP_5 is 0.707
Spearman p value for for H_COMP_1 and H_COMP_5 is 0.05

p value less than 0.05
Speaman correlation for H_COMP_3 and H_COMP_5 is 0.979
Spearman p value for for H_COMP_3 and H_COMP_5 is 0.0

p value less than 0.05
Speaman correlation for H_COMP_3 and H_COMP_6 is 0.743
Spearman p value for for H_COMP_3 and H_COMP_6 is 0.035

p value less than 0.05
Speaman correlation for H_COMP_1 and H_COMP_7 is 0.831
Spearman p value for for H_COMP_1 and H_COMP_7 is 0.011

p value less than 0.05
Speaman correlation for H_COMP_3 and H_COMP_7 is 0.883
Spearman p value for for H_COMP_3 and H_COMP_7 is 0.004

p value less than 0.05
Speaman correlation for H_COMP_5 and H_COMP_7 is 0.901
Spearman p value for for H_COMP_5 and H_COMP_7 is 0.002

p value less than 0.05
Speaman correlation for H_COMP_1 and H_CLEAN_HSP is 0.765
Spearman p value for for H_COMP_1 and H_CLEAN_HSP is 0.027

p value less than 0.05
Speaman correlation for H_COMP_3 and H_CLEAN_HSP is 0.928
Spearman p value for for H_COMP_3 and H_CLEAN_HSP is 0.001

p value less than 0.05
Speaman correlation for H_COMP_5 and H_CLEAN_HSP is 0.947
Spearman p value for for H_COMP_5 and H_CLEAN_HSP is 0.0

p value less than 0.05
Speaman correlation for H_COMP_7 and H_CLEAN_HSP is 0.961
Spearman p value for for H_COMP_7 and H_CLEAN_HSP is 0.0

p value less than 0.05
Speaman correlation for H_COMP_1 and H_HSP_RATING is 0.738
Spearman p value for for H_COMP_1 and H_HSP_RATING is 0.037

p value less than 0.05
Speaman correlation for H_COMP_3 and H_HSP_RATING is 0.863
Spearman p value for for H_COMP_3 and H_HSP_RATING is 0.006

p value less than 0.05
Speaman correlation for H_COMP_5 and H_HSP_RATING is 0.753
Spearman p value for for H_COMP_5 and H_HSP_RATING is 0.031

p value less than 0.05
Speaman correlation for H_COMP_6 and H_HSP_RATING is 0.8
Spearman p value for for H_COMP_6 and H_HSP_RATING is 0.017

```
p value less than 0.05
Speaman correlation for H_COMP_7 and H_HSP_RATING is 0.713
Spearman p value for for H_COMP_7 and H_HSP_RATING is 0.047


p value less than 0.05
Speaman correlation for H_CLEAN_HSP and H_HSP_RATING is 0.713
Spearman p value for for H_CLEAN_HSP and H_HSP_RATING is 0.047


p value less than 0.05
Speaman correlation for H_COMP_3 and H_RECMND is 0.907
Spearman p value for for H_COMP_3 and H_RECMND is 0.002


p value less than 0.05
Speaman correlation for H_COMP_5 and H_RECMND is 0.802
Spearman p value for for H_COMP_5 and H_RECMND is 0.017


p value less than 0.05
Speaman correlation for H_COMP_6 and H_RECMND is 0.745
Spearman p value for for H_COMP_6 and H_RECMND is 0.034


p value less than 0.05
Speaman correlation for H_COMP_7 and H_RECMND is 0.73
Spearman p value for for H_COMP_7 and H_RECMND is 0.04


p value less than 0.05
Speaman correlation for H_CLEAN_HSP and H_RECMND is 0.76
Spearman p value for for H_CLEAN_HSP and H_RECMND is 0.029


p value less than 0.05
Speaman correlation for H_HSP_RATING and H_RECMND is 0.976
Spearman p value for for H_HSP_RATING and H_RECMND is 0.0
```

Communication with Doctors and Quietness of Hospital Environment does not affect the Rating and Recommendation of the hospitals or any other metrics as there is no significant relationship between these variables. There is a high probability that the observed correlation between the other measures and the ranking is unlikely to have occurred by random chance alone and could rather be due to a true relationship in the population.

Communication with Nurses have a strong positive relationsip between Responsiveness of Staff, Communication about Medicines, Care Transition and Cleanliness of the Hospital whihc in turn have a strong relationship with each other and hence the overall rating and recommendation of the hospital.

It can be concluded that the Communication with Nurses correlates to the rating of the hospitals and there is no significant relationship or correlation between the healthcare expenditure and the rating.