# HDSC 2022 PREMIERE PROJECT

**PP22/H602**

* Scope: **Health and Medicare**
* Topic: **US Hospital Customer Satisfaction 2016- 2020: Is Patient Satisfaction Correlated With Overall Hospital Performance?**

## PROJECT DESCRIPTION

One of the many ways a hospital is ranked is the satisfaction of its patients. By collecting data from the United States’ Centers for Medicare and Medicaid Services (CMS), it is possible to analyse how satisfied US patients are. See more about the provided datasets [here](https://www.kaggle.com/abrambeyer/us-hospital-customer-satisfaction-20162020).

Hospital Compare provides data on over 4,000 Medicare-certified hospitals, including acute care hospitals, critical access hospitals (CAHs), children’s hospitals, Veterans Health Administration (VHA) Medical Centers, and hospital outpatient departments"

The Centers for Medicare & Medicaid Services (CMS) uses a five-star quality rating system to measure the experiences Medicare beneficiaries have with their health plan and health care system — the Star Rating Program. Health plans are rated on a scale of 1 to 5 stars, with 5 being the highest.

One part of a hospital's overall rating is it's patient satisfaction survey scores. CMS attempts to take into consideration how well patients are treated by the provider. A description of HCAHPS can be found here \*\*\*[HCAHPS Description](https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/HospitalHCAHPS).

#### Dataset description:

| Column Name | Data Type | Description |
| --- | --- | --- |
| Facility ID | Char(6) | Facility Medicare ID |
| Facility Name | Char(72) | Name of the facility |
| Address | Char(51) | Facility street address |
| City | Char(20) | Facility City |
| State | Char(2) | Facility State |
| ZIP Code | Num(8) | Facility ZIP Code |
| County Name | Char(25) | Facility County |
| Phone Number | Char(14) | Facility Phone Number |
| HCAHPS Measure ID | Char(25) | HCAHPS Patient Survey Measure Name |
| HCAHPS Question | Char(138) | HCAHPS Patient Survey Question |
| HCAHPS Answer Description | Char(118) | HCAHPS Patient Survey Answer |
| Patient Survey Star Rating | Char(14) | Overall rating for survey item |
| Patient Survey Star Rating Footnote | Char(7) | n/a |
| HCAHPS Answer Percent | Char(14) | Percent of surveys with question answered |
| HCAHPS Answer Percent Footnote | Char(8) | n/a |
| HCAHPS Linear Mean Value | Char(14) | HCAHPS Patient Survey question linear mean value |
| Number of Completed Surveys | Char(13) | Number of completed surveys for hospital. N-size. |
| Number of Completed Surveys Footnote | Char(8) | n/a |
| Survey Response Rate Percent | Char(13) | Hospital survey response rate. |
| Survey Response Rate Percent Footnote | Char(8) | n/a |
| Start Date | Date | Survey collection period start date |
| End Date | Date | Survey collection period end date |
| Year | Char(4) | cms data release year |
| Hospital Type | Char(34) | What type of facility is it? |
| Hospital Ownership | Char(43) | What type of ownership does the facility have? |
| Emergency Services | Char(3)) | Does the facility have emergency services Yes/No? |
| Meets criteria for promoting interoperability of EHRs | Char(1) | Does facility meet government EHR standard Yes/No? |
| Hospital overall rating | Char(13) | Hospital Overall Star Rating 1=Worst; 5=Best. Aggregate measure of all other measures |
| Hospital overall rating footnote | Num(8) |  |
| Mortality national comparison | Char(28) | Facility overall performance on mortality measures compared to other facilities |
| Mortality national comparison footnote | Num(8) |  |
| Safety of care national comparison | Char(28) | Facility overall performance on safety measures compared to other facilities |
| Safety of care national comparison footnote | Num(8) |  |
| Readmission national comparison | Char(28) | Facility overall performance on readmission measures compared to other facilities |
| Readmission national comparison footnote | Num(8) |  |
| Patient experience national comparison | Char(28) | Facility overall performance on pat. exp. measures compared to other facilities |
| Patient experience national comparison footnote | Char(8) |  |
| Effectiveness of care national comparison | Char(28) | Facility overall performance on effect. of care measures compared to other facilities |
| Effectiveness of care national comparison footnote | Char(8) |  |
| Timeliness of care national comparison | Char(28) | Facility overall performance on timeliness of care measures compared to other facilities |
| Timeliness of care national comparison footnote | Char(8) |  |
| Efficient use of medical imaging national comparison | Char(28) | Facility overall performance on efficient use measures compared to other facilities |
| Efficient use of medical imaging national comparison footnote | Char(8) |  |

#### Data Transformation:

The dataset was constructed in such a way that each hospital had 55 rows of data, each row corresponding to one of the 55 HCAHPS questions on the patient survey.

The patient’s responses to the HCAHPS questions were used to calculate the HCAHPS linear mean value, which were in turn used to generate 10 star ratings:

1. Communication with Nurses
2. Communication with Doctors
3. Responsiveness of Hospital Staff
4. Communication about Medicines
5. Discharge Information
6. Care Transition
7. Cleanliness of Hospital Environment
8. Quietness of Hospital Environment
9. Hospital Rating
10. Recommend the Hospital

These 10 star ratings were then used to generate a **Summary Star Rating**.This summary star rating can be used to represent the overall patient satisfaction.

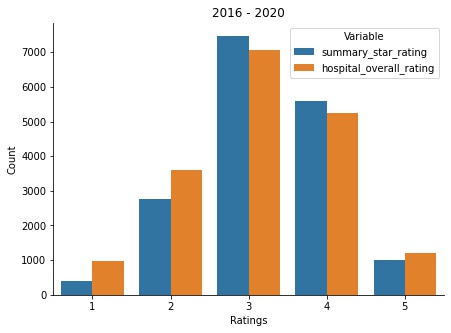
The above information can be found [here](https://www.hcahpsonline.org/globalassets/hcahps/star-ratings/tech-notes/january_2022_star-ratings_tech-notes.pdf).

So the entire data was transformed such that each hospital only had one row, with extra columns added for the summary star rating, the other star ratings and the HCAHPS linear mean value.

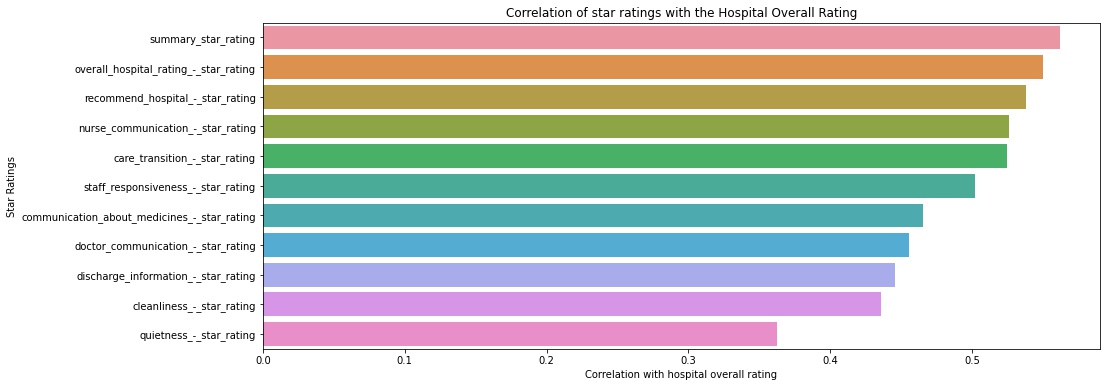
The code used to generate the new dataset can be found [here](https://colab.research.google.com/drive/1o13JOZzm3lGYaIkCZXW12AelCWBiGldc?usp=sharing).

#### Exploration and Visualization:

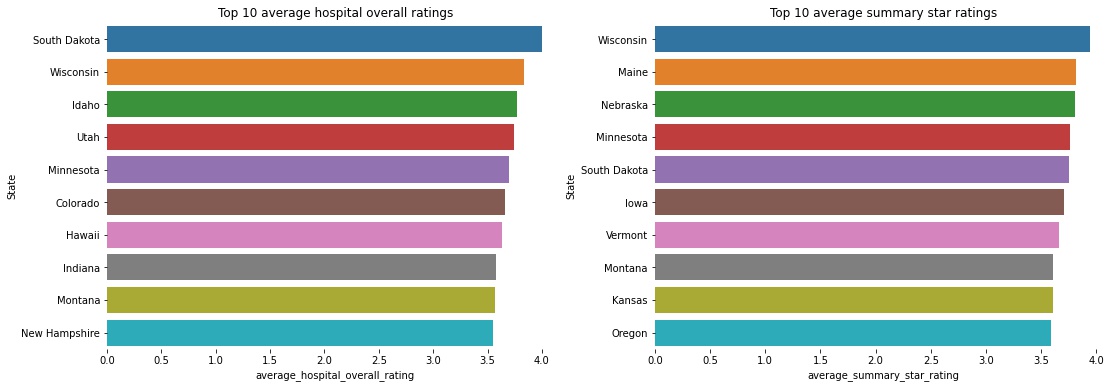
Some Exploration of the data was done.



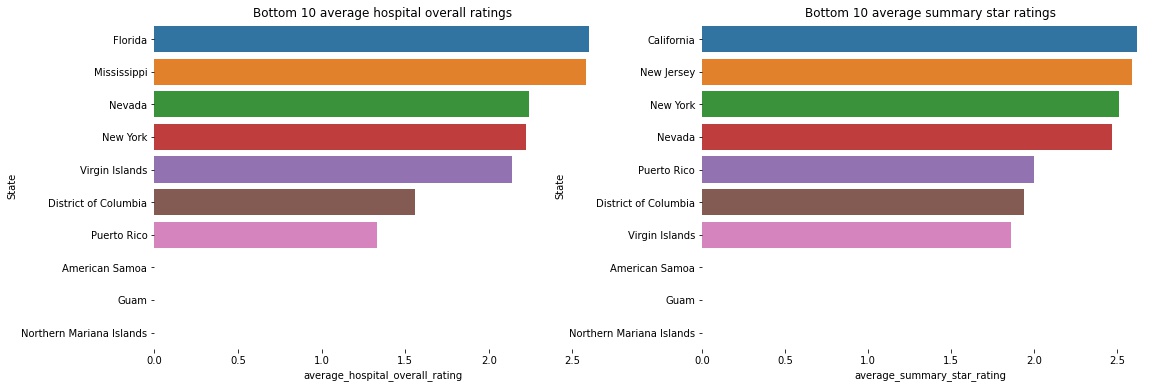
**Count of the different ratings for both the hospital overall rating and the summary star rating over the five year period (2016-2020).**



**Correlation of the star ratings with the hospital overall rating over the five year period (2016-2020).**



**Top 10 states with the highest average hospital overall rating and average summary star rating across the five year period (2016-2020).**



**Bottom 10 states with the lowest average hospital overall rating and average summary star rating across the five year period (2016-2020).**

Code for the above can be found [here](https://colab.research.google.com/drive/14IEkCMWMpc2GRws3u4F_VR8FtCo2kXRc" \l "scrollTo=1cHOaVJQnEZV).

**\*\*more exploration and visualization to be done.**

#### Statistical Analysis:

**Hypothesis**: There is no significant relationship between both variables, they are independent of each other.

**Test**: If p-value is less than 0.05, reject hypothesis and declare both variables to be dependent but if p-value is greater than 0.05 accept hypothesis and declare both variables to be independent.

**Result**: Analysis of the summary star rating and the hospital overall ratingresulted in a \*\*p-value of 0.0, therefore both variables can be said to have significant relationship.

The correlation between both variables, about 0.56, also indicates a significant positive relationship.

\*\*p-value of exactly 0.0 seems suspicious, the value may have been approximated.

#### Regression Analysis:

A regression analysis is to be carried out to obtain the importance of the features.

The dependent variable or target is the hospital overall rating.

Some data preprocessing steps are to be carried out before proceeding with the analysis.

Some of which are:

* All categorical features are to be encoded with numerical variables.
* All rows with it’s hospital overall rating value missing is to be dropped.
* All the ratings columns with missing values are filled with 0.
* All other categorical columns with missing values are to be filled in randomly using forward and backward filling (pandas.DataFrame.ffill and pandas.DataFrame.bfill).
* Hospital Ownership was not used in training the models, it was concluded that this feature could introduce some bias to the model. To ensure fairness, it was removed.
* Pain management linear mean score and star rating columns are both dropped due to the high percentage of missing values in them (about 70%).

Control features: These are to be used to determine the influence of the star rating in predicting the target.

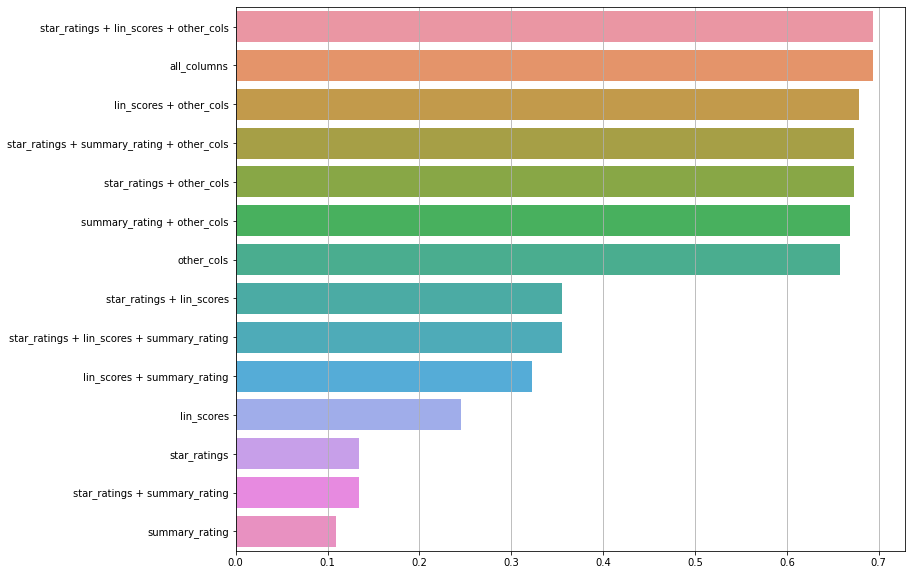
1. 'hospital\_type',
2. 'emergency\_services',
3. 'meets\_criteria\_for\_promoting\_interoperability\_of\_ehrs',
4. 'mortality\_national\_comparison',
5. 'safety\_of\_care\_national\_comparison',
6. 'readmission\_national\_comparison',
7. 'patient\_experience\_national\_comparison',
8. 'effectiveness\_of\_care\_national\_comparison',
9. 'timeliness\_of\_care\_national\_comparison',
10. 'efficient\_use\_of\_medical\_imaging\_national\_comparison'

The data is to be split randomly, using the sklearn\_model\_selection.train\_test\_split with a random state value of 0, 70% for training and 30% for testing.

The Ridge Regression model is to be trained using the following feature combinations:

1. Control\_features
2. Summary\_rating
3. Star\_ratings
4. Linear\_mean\_scores
5. Control\_features + Summary\_rating
6. Control\_features + Star\_ratings
7. Control\_features + Linear\_mean\_scores
8. Linear\_mean\_scores + Summary\_rating
9. Linear\_mean\_scores + Star\_ratings
10. Control\_features + Summary\_rating + Star\_ratings
11. Control\_features + Linear\_mean\_scores + Star\_ratings
12. Control\_features + Linear\_mean\_scores + Summary\_rating
13. Star\_ratings + Linear\_mean\_scores + Summary\_rating
14. All\_columns

The models were evaluated using the Coefficient of Determination (R-squared) score.



**Results of training the models using the various feature combinations.**