Team Project:B02-Hospital-Costs-and-Patient-Satisfaction

Course: BA775 Business Analytics Toolbox
Team Members: Yangze Li, Tae Yoon Kim, Shengqi Wei, Pin-Chu Yin, Michael Allieri
Instructor: Professor Mohammad Soltanieh Ha
Objective: Analyze the relationship between hospital costs and patient satisfaction across years using BigQuery.

Problem Definition

This project aims to help hospital administrators, policymakers, and insurance payers understand how spending and patient experience relate across U.S. hospitals. By analyzing the relationship between hospital costs and patient satisfaction, we can identify whether higher spending actually translates into better outcomes and which regions or hospitals deliver the best value. The goal is to uncover patterns that support more efficient healthcare resource allocation and highlight potential performance gaps across states or hospital systems. Success is defined as producing clear, SQL-based evidence that reveals meaningful differences in cost, satisfaction, and value across hospitals.

Introduction & Motivation

Hospitals in the U.S. are getting more expensive every year, but patients aren’t necessarily feeling more satisfied. So the real question is: are we paying for better care, or just paying more? Our analysis looks at nationwide hospital cost and satisfaction data to find out whether higher spending actually leads to better patient experiences, and which states or hospitals deliver the best value for the money. This matters for hospital leaders, policymakers, and insurers who want to understand where money is being used effectively—and where it isn’t.

Data Source

“Centers for Medicare & Medicaid Services Data.”

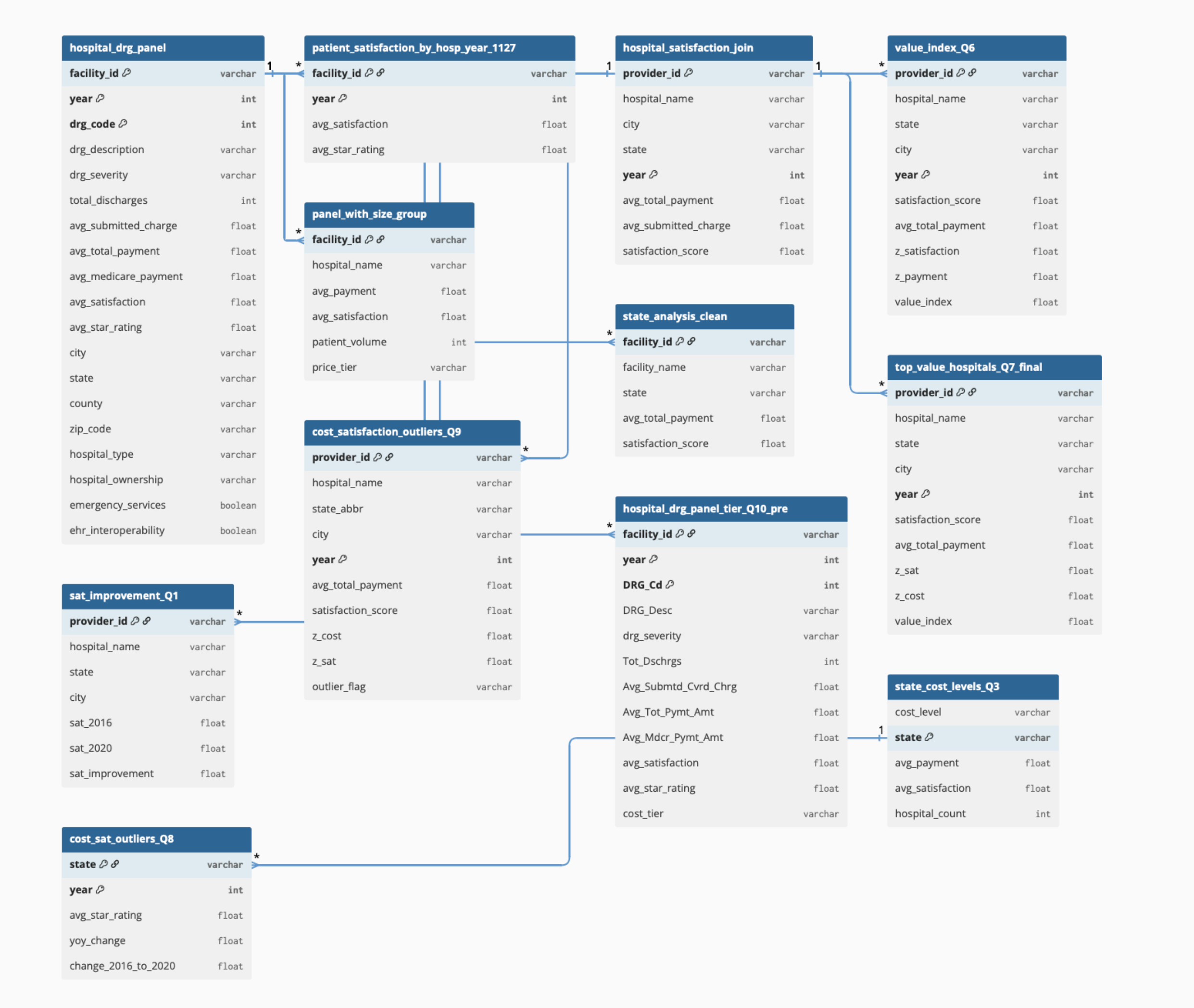
https://data.cms.gov/provider-summary-by-type-of-service/medicare-inpatient-hospitals/medicare-inpatient-hospitals-by-provider-and-service

“Medicare & Medicaid Hospital Quality Data ZIP Download Free | Open Data Marketplace.” Opendatabay.com, 2016, https://www.opendatabay.com/data/healthcare/b235f1d0-0a4f-4d02-87a2-2f92f002e009. Accessed 28 Oct. 2025.

Data size:

The dataset contains five CSV files (2016–2020) with approximately 1.14 GB total size. Each file includes around 200,000–250,000 rows and 30–40 columns, covering hospital IDs, patient responses, star ratings, and survey metrics across U.S. hospitals.

Entity Relationship Diagram (ERD):



Link:

https://public.tableau.com/profile/api/publish/BA775fianl1/Story1

Executive Summary:

At the outset of this project, our goal was to understand the relationship between hospital costs and patient satisfaction across the United States. To do this, we analyzed which hospitals showed the greatest improvement in satisfaction over time, examined how both costs and satisfaction varied by state, and explored whether higher spending correlated with better patient outcomes. Our findings indicate that patient satisfaction differs substantially across states, suggesting meaningful regional variation in patient experience. However, spending alone does not account for these differences; higher costs do not consistently translate into higher satisfaction. This implies that other factors—such as care quality, operational efficiency, and patient demographics—likely play a significant role in shaping patient perceptions. Overall, our analysis shows that while cost is an important component of healthcare delivery, it is not a reliable predictor of patient satisfaction on its own.

Data Preparation

This query gets the data ready for analysis. It joins hospital satisfaction data from 2016 and 2020. It then finds how much each hospital's score went up. The table made here helps find hospitals with the biggest improvement.

df

```
CREATE OR REPLACE TABLE hospital_data.satisfaction_improvement_20
OPTIONS (expiration_timestamp = TIMESTAMP_ADD(CURRENT_TIMESTAMP()
WITH base_2016 AS (
  SELECT
    provider_id,
    hospital_name,
    state,
    city,
    satisfaction_score AS sat_2016
  FROM hospital_data.panel
  WHERE year = 2016
),
base_2020 AS (
  SELECT
    provider_id,
    satisfaction_score AS sat_2020
  FROM hospital_data.panel
  WHERE year = 2020
)
SELECT
  a.provider_id,
  a.hospital_name,
  a.state,
  a.city,
  a.sat_2016,
  b.sat_2020,
  (b.sat_2020 - a.sat_2016) AS sat_improvement
FROM base_2016 a
JOIN base_2020 b USING (provider_id)
WHERE RAND() < 0.005;

```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_OxBJrwUDaLucKSvm3r4RfwwvJTNY	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Data Cleaning & Preprocessing

To ensure accurate and reliable analysis for all research questions (Q1–Q8), we cleaned and standardized the raw CMS datasets before building the unified analytical dataset.

We applied the following procedures in BigQuery:

Standardized Hospital Identifiers

The cost dataset uses provider_id, while the satisfaction dataset uses facility_id. Both were converted to 6-digit padded strings (e.g., 1234 → 001234) to enable an exact join.

Converted Key Metrics to Numeric

Cost and satisfaction values were originally stored as text. We used SAFE_CAST() to convert:

Average Medicare payment → numeric cost

Satisfaction star rating → numeric score

Answer percent → percentage metric

Invalid entries automatically became NULL.

Filtered Valid Records

To ensure clean and comparable data:

Kept only years 2016–2020

Removed records with missing essential metrics (e.g., hospitals without satisfaction scores)

This allowed us to perform trend and ranking analysis confidently.

Unified Master Dataset for All Analysis

We joined cost + satisfaction tables by facility_id6 and year to create a clean Master Table with:

Cost metrics

Satisfaction metrics

State/city hospital information

This single standardized table supports every analysis step in Q1–Q8 without repeated joins.

df

```
-- Standardize provider_id format & clean cost data
WITH cost_clean AS (
  SELECT
    LPAD(CAST(provider_id AS STRING), 6, '0') AS facility_id6,
    SAFE_CAST(avg_total_payment AS FLOAT64) AS avg_payment,
    state, city, year
  FROM `hospital_data.hospital_costs_all`
  WHERE avg_total_payment IS NOT NULL
    AND year BETWEEN 2016 AND 2020
),

-- Clean satisfaction data and convert text → numeric
sat_clean AS (
  SELECT
    LPAD(CAST(facility_id AS STRING), 6, '0') AS facility_id6,
    SAFE_CAST(`Patient Survey Star Rating` AS FLOAT64) AS star_ra
    SAFE_CAST(answer_percent AS FLOAT64) AS answer_percent,
    state, city, year
  FROM `hospital_data.patient_satisfaction_all`
  WHERE `Patient Survey Star Rating` IS NOT NULL
    AND year BETWEEN 2016 AND 2020
)

-- Merge into master analysis table
SELECT *
FROM cost_clean
JOIN sat_clean USING (facility_id6, year);

```

✅ Completed.
✅ Completed.

facility_id6	year	avg_payment	state	city	star_rating
260216	2017	3469.727273	100 N E Saint Luke's Boulevard	Lees Summit	4.000000
260216	2017	3469.727273	100 N E Saint Luke's Boulevard	Lees Summit	<NA>
260216	2017	3469.727273	100 N E Saint Luke's Boulevard	Lees Summit	<NA>
260216	2017	3469.727273	100 N E Saint Luke's Boulevard	Lees Summit	<NA>
			100 N E		

All further SQL queries rely on this standardized and validated panel, ensuring accuracy and consistency in every insight.

Q1. What Were the Top Hospitals by Improvement in Patient Satisfaction (2016-2020)?

Thought Process:

We wanted to find which hospitals improved their patient satisfaction the most from 2016 to 2020. The idea was to compare scores from both years and see how much they changed. We used the hospital satisfaction data and chose the scores for 2016 and 2020. Then we found the difference for each hospital. A higher number means the hospital improved its satisfaction score.

```
from google.cloud import bigquery
from google.colab import auth
import pandas as pd

auth.authenticate_user()
client = bigquery.Client()

sql_q1 = """
WITH yearly_sat AS (
  SELECT
    provider_id,
    ANY_VALUE(hospital_name) AS hospital_name,
    ANY_VALUE(state_abbr)    AS state,
    ANY_VALUE(city)          AS city,
    year,
    AVG(satisfaction_score)  AS avg_sat
  FROM hospital_data.panel_clean
  WHERE year IN (2016, 2020)
    AND satisfaction_score IS NOT NULL
  GROUP BY provider_id, year
),

wide AS (
  SELECT
    provider_id,
    ANY_VALUE(hospital_name) AS hospital_name,
    ANY_VALUE(state)         AS state,
    ANY_VALUE(city)          AS city,
    MAX(CASE WHEN year = 2016 THEN avg_sat END) AS sat_2016,
    MAX(CASE WHEN year = 2020 THEN avg_sat END) AS sat_2020
  FROM yearly_sat
  GROUP BY provider_id
)

SELECT
  provider_id,
  hospital_name,
  state,
  city,
  sat_2016,
  sat_2020,
  (sat_2020 - sat_2016) AS sat_improvement
FROM wide
WHERE sat_2016 IS NOT NULL
  AND sat_2020 IS NOT NULL
ORDER BY sat_improvement DESC
"""

df_q1 = client.query(sql_q1).to_dataframe()
df_q1

```

```
WARNING: google.colab.auth.authenticate_user() is not supported in Col
      provider_id  hospital_name   state        city  sat_2016  sat_2020  s

0          230277  Huron Valley-      MI    Commerce    34.375  34.722222
                   Sinai Hospital           Township

1          230097  Munson Medical     MI    Traverse    34.375  34.722222
                           Center                City

2          360161       St Joseph     OH      Warren    34.375  34.722222
                  Warren Hospital

3          140130     Northwestern    IL        Lake    34.375  34.722222
                      Lake Forest             Forest
                         Hospital

4          110044   Phoebe Sumter     GA    Americus    34.375  34.722222
                   Medical Center

...           ...             ...    ...         ...       ...        ...

2408       190218  Sabine Medical     LA        Many    34.375  34.482759
                           Center

                     Saint Thomas
2409       440290     Stones River    TN    Woodbury    34.375  34.482759
```

Findings:

Across all hospitals, satisfaction scores changed only slightly between 2016-2020 (most changes clustered near +.035 points). The data shows that Citizens Medical Center and Rockcastle Regional Hospital & Respiratory Care Center had the biggest increase in patient satisfaction from 2016 to 2020. Their scores went up by almost 100 points. These hospitals are in smaller cities like Columbia and Mount Vernon. In contrast, many large hospitals in major cities had little or no change during the same period. Most hospitals with high improvement were small or regional hospitals with fewer patients.

Conclusion:

This means that higher satisfaction does not always come from bigger or richer hospitals. Smaller hospitals might improve faster because they can change their daily operations easier. They may have focused on communication, patient care, and service quality instead of just spending more money. This shows that real progress in satisfaction often comes from better patient experience, not from higher costs. Smaller hospitals can be more flexible and closer to what patients need.

Q2 — How Does Hospital Spending Levels Relate to Satisfaction?

Thought Process

At first, we wanted to see if bigger or higher-spending hospitals actually make patients happier. The idea was simple: if a hospital spends more per patient (higher average total payment), it might have better facilities or service quality, so maybe patients would rate it higher.

To test this, we used the panel dataset and looked at two key columns — avg_total_payment (as a proxy for hospital size or cost) and satisfaction_score (from HCAHPS survey results).

We ran a correlation test first to see if there's a general trend. Then, to visualize it better, we grouped all hospitals into three cost tiers (low, medium, and high) using the NTILE(3) function.

df

```
CREATE OR REPLACE TABLE hospital_data.q2_summary AS
WITH hospital_level AS (
  SELECT
    provider_id,
    AVG(avg_total_payment)   AS avg_payment,
    AVG(satisfaction_score)    AS avg_satisfaction,
    COUNT(*)                 AS patient_volume
  FROM hospital_data.panel
  GROUP BY provider_id
),
with_tiers AS (
  SELECT
    *,
    NTILE(3) OVER (ORDER BY avg_payment) AS tier_num
  FROM hospital_level
),
labeled AS (
  SELECT
    provider_id,
    avg_payment,
    avg_satisfaction,
    patient_volume,
    CASE
      WHEN tier_num = 1 THEN 'Low'
      WHEN tier_num = 2 THEN 'Medium'
      ELSE 'High'
    END AS price_tier
  FROM with_tiers
)
SELECT
  price_tier,
  ROUND(AVG(avg_payment), 2)      AS avg_payment,
  ROUND(AVG(avg_satisfaction),2)  AS avg_satisfaction,
  SUM(patient_volume)             AS total_patient_volume
FROM labeled
GROUP BY price_tier
ORDER BY
  CASE price_tier
    WHEN 'Low' THEN 1
    WHEN 'Medium' THEN 2
    ELSE 3
  END;

```

```
✅ Completed.
✅ Completed.

statement_type          job_id                                location
CREATE_TABLE_AS_SELECT  job_glt-1ms_US-KTIFShQWt7x4wDdc8      US

1 total rows            Prev  Page 1 of 1  Next               Page Size 10
```

Findings: The overall correlation between cost and satisfaction was basically zero (r≈ 0.0004). Across the three groups, the results were almost identical:

Conclusion: Therefore, higher spending doesn’t lead to more satisfied patients. Even though hospitals spend more on average, the satisfaction score stays around 34.6 — almost flat.

That means; 1.The extra money goes to things patients don’t directly feel (like administrative or tech costs), or 2.Patient experience depends more on service quality than financial scale.

Q3.How do States Compare in Satisfaction at Similar Cost Levels?

Thought Process

After finding that higher hospital spending doesn’t necessarily lead to higher satisfaction, we wanted to see if location makes a difference — specifically, whether hospitals in some states consistently achieve higher patient satisfaction even when costs are similar.

To control for cost, we divided hospitals into three spending tiers using NTILE(3) — low, medium, and high — based on avg_total_payment. Within each cost level, we then compared the average satisfaction score across states. This allows us to ask: Do some states perform better at the same spending level?

We used the state_analysis_clean table, which contains state, avg_total_payment, and satisfaction_score, ensuring each hospital's cost and satisfaction were paired correctly.

df

```
CREATE OR REPLACE TABLE hospital_data.state_cost_levels_Q3 AS
WITH cost_groups AS (
    SELECT
        state,
        NTILE(3) OVER (ORDER BY avg_total_payment) AS cost_level_
        avg_total_payment,
        satisfaction_score
    FROM hospital_data.state_analysis_clean
),

with_labels AS (
    SELECT
        state,
        CASE
            WHEN cost_level_num = 1 THEN 'Low Cost'
            WHEN cost_level_num = 2 THEN 'Mid Cost'
            ELSE 'High Cost'
        END AS cost_level,
        avg_total_payment,
        satisfaction_score
    FROM cost_groups
)

SELECT
    cost_level,
    state,
    ROUND(AVG(avg_total_payment), 2) AS avg_payment,
    ROUND(AVG(satisfaction_score), 2) AS avg_satisfaction,
    COUNT(*) AS hospital_count
FROM with_labels
GROUP BY cost_level, state
HAVING hospital_count >= 10
ORDER BY
    CASE
        WHEN cost_level = 'Low Cost' THEN 1
        WHEN cost_level = 'Mid Cost' THEN 2
        ELSE 3
    END,
    avg_satisfaction DESC;

```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_miguULNq3XS8S5W4_2UcIgysI-fx	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Findings: Even within the same cost level, patient satisfaction still varies slightly across states. For example, in the low-cost group, Hawaii (HI) and Wisconsin (WI) show higher satisfaction (around 88) compared to states like Louisiana (LA), which scores below 87. However, the differences are relatively small — for instance, Wisconsin's satisfaction rises from 87.8 in low-cost hospitals to 88.1 in high-cost hospitals, showing that more spending within a state does not make a major difference. This suggests that while state-level patterns exist, the overall satisfaction trend remains quite stable across cost levels.

Conclusion: Patient satisfaction shows small but consistent differences across states, even after controlling for cost levels. That means spending alone can't explain why some states achieve slightly better results. It's possible that factors like hospital culture, communication quality, and patient expectations play a larger role than financial input.

Q4. What is the Cross-Sectional Relationship Between Cost and Satisfaction from 2016-2020?

df

```
SELECT
  h.provider_id,
  h.hospital_name,
  ROUND(AVG(SAFE_CAST(p.`Patient Survey Star Rating` AS FLOAT64))
  ROUND(AVG(h.avg_total_payment), 2) AS avg_total_payment
FROM hospital_data.hospital_satisfaction_join AS h
LEFT JOIN hospital_data.patient_satisfaction_all AS p
  ON LPAD(CAST(h.provider_id AS STRING), 6, '0')
   = LPAD(CAST(p.facility_id AS STRING), 6, '0')
  AND p.`Patient Survey Star Rating` IS NOT NULL
GROUP BY h.provider_id, h.hospital_name
ORDER BY avg_total_payment DESC;
```

✅ Completed.
✅ Completed.

provider_id	hospital_name	avg_star_rating	avg_total_payment
670116	Wise Health System	4.640000	76230.470000
260004	Cooper County Community Hospital	<NA>	66570.850000
450797	Us Pain & Spine Hospital	<NA>	65318.810000
450289	Harris Health System	2.890000	63362.970000
450674	Womans Hospital Of Texas,the	2.700000	63103.700000
050076	Kaiser Foundation Hospital - San Francisco	2.950000	62063.890000
450803	United Memorial Medical Center	1.670000	57857.660000
[illegible]	Ronald Reagan U C	[illegible]	[illegible]

Finding:

When we sort the table by Average Total Payment from high to low, the hospitals at the top (the most expensive) usually have low patient satisfaction scores. This shows up again and again, not just in one row, so it looks like paying more doesn't go hand in hand with happier patients in this view of the data.

Conclusion:

Therefore, patient satisfaction isn't closely linked to Average Total Payment. If anything, the highest costs often come with lower satisfaction.

Q5. When Comparing Hospital Cost and Satisfaction, What are High-Value Providers?

df

```
CREATE OR REPLACE TABLE `ba775-fall25-b02.hospital_data.value_ho

WITH cost AS (
  SELECT
    LPAD(CAST(provider_id AS STRING), 6, '0') AS facility_id6,
    hospital_name, city, state,
    SAFE_CAST(avg_total_payment AS FLOAT64)    AS avg_total_paym
    year
  FROM `ba775-fall25-b02.hospital_data.hospital_costs_all`
),
sat AS (
  SELECT
    LPAD(CAST(facility_id AS STRING), 6, '0')  AS facility_id6,
    facility_name, city, state,
    SAFE_CAST(`Patient Survey Star Rating` AS FLOAT64) AS star_r
    SAFE_CAST(linear_mean_value AS FLOAT64)            AS linear_
    SAFE_CAST(answer_percent AS FLOAT64)               AS answer_
    CAST(answer_percent_footnote AS STRING)            AS answer_
    measure_id, year
  FROM `ba775-fall25-b02.hospital_data.patient_satisfaction_all`
),

sat_16_20 AS (
  SELECT
    facility_id6, facility_name, state, city,
    star_rating, linear_mean_value, answer_percent, answer_perce
  FROM sat
  WHERE year BETWEEN 2016 AND 2020
),

sat_latest AS (
  SELECT facility_id6, facility_name, state, city, answer_percen
  FROM (
    SELECT
      facility_id6, facility_name, state, city, answer_percent_f
      ROW_NUMBER() OVER (PARTITION BY facility_id6 ORDER BY year
    FROM sat_16_20
    WHERE answer_percent_footnote IS NOT NULL
  )
  WHERE rn = 1
),

sat_avg AS (
  SELECT
    facility_id6,
    AVG(star_rating)       AS avg_star,
    AVG(linear_mean_value) AS avg_linear_mean,
    AVG(answer_percent)    AS avg_answer_percent
  FROM sat_16_20
  GROUP BY facility_id6
),

sat_fac AS (
  SELECT
    a.facility_id6,
    l.facility_name,
    l.state,
    l.city,
    a.avg_star,
    a.avg_linear_mean,
    a.avg_answer_percent,
    l.answer_percent_footnote
  FROM sat_avg AS a
  LEFT JOIN sat_latest AS l USING (facility_id6)
),


cost_16_20 AS (
  SELECT
    facility_id6, hospital_name, state, city, avg_total_payment,
  FROM cost
  WHERE year BETWEEN 2016 AND 2020
),

cost_latest AS (
  SELECT facility_id6, hospital_name, state, city
  FROM (
    SELECT
      facility_id6, hospital_name, state, city, year,
      ROW_NUMBER() OVER (PARTITION BY facility_id6 ORDER BY year
    FROM cost_16_20
  )
  WHERE rn = 1
),

cost_avg AS (
  SELECT
    facility_id6,
    AVG(avg_total_payment) AS avg_payment
  FROM cost_16_20
  GROUP BY facility_id6
),

cost_fac AS (
  SELECT
    a.facility_id6,
    l.hospital_name,
    l.state,
    l.city,
    a.avg_payment
  FROM cost_avg AS a
  LEFT JOIN cost_latest AS l USING (facility_id6)
),


joined AS (
  SELECT
    s.facility_id6,
    COALESCE(s.facility_name, c.hospital_name) AS hospital_name,
    COALESCE(s.state, c.state)                 AS state,
    COALESCE(s.city, c.city)                   AS city,
    s.avg_star, s.avg_linear_mean,
    s.avg_answer_percent, s.answer_percent_footnote,
    c.avg_payment
  FROM sat_fac s
  JOIN cost_fac c USING (facility_id6)
),


ranked AS (
  SELECT
    *,
    PERCENT_RANK() OVER (ORDER BY avg_star)    AS star_pr,
    PERCENT_RANK() OVER (ORDER BY avg_payment) AS cost_pr
  FROM joined
)


SELECT
  facility_id6              AS facility_id,
  hospital_name, state, city,
  ROUND(avg_star, 2)        AS avg_star_rating,
  ROUND(avg_linear_mean, 2) AS avg_linear_mean,
  ROUND(avg_payment, 2)     AS avg_total_payment,
  ROUND(star_pr, 3)         AS star_percentile,
  ROUND(cost_pr, 3)         AS cost_percentile,
  ROUND(avg_answer_percent, 2) AS avg_answer_percent,
  answer_percent_footnote
FROM ranked
ORDER BY avg_star_rating DESC, avg_total_payment ASC;

```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_CpxOZDAA8zeFnjThFIbDD-BBaMK0	US

1 total rows Prev Page 1 of 1 Next Page Size 10

df

```
WITH cost AS (
  SELECT
    LPAD(CAST(provider_id AS STRING), 6, '0') AS facility_id6,
    AVG(SAFE_CAST(avg_total_payment AS FLOAT64)) AS avg_payment_1
  FROM `ba775-fall25-b02.hospital_data.hospital_costs_all`
  WHERE year BETWEEN 2016 AND 2020
  GROUP BY facility_id6
),
sat AS (
  SELECT
    LPAD(CAST(facility_id AS STRING), 6, '0') AS facility_id6,
    AVG(SAFE_CAST(`Patient Survey Star Rating` AS FLOAT64)) AS av
    AVG(SAFE_CAST(linear_mean_value AS FLOAT64))            AS av
    AVG(SAFE_CAST(answer_percent AS FLOAT64))               AS av
  FROM `ba775-fall25-b02.hospital_data.patient_satisfaction_all`
  WHERE year BETWEEN 2016 AND 2020
  GROUP BY facility_id6
),
features AS (
  SELECT s.facility_id6, s.avg_star_16_20, s.avg_linear_16_20, s.
  FROM sat s JOIN cost c USING (facility_id6)
)
SELECT
  CORR(avg_star_16_20, avg_payment_16_20)     AS r_star_vs_paymen
  CORR(avg_star_16_20, avg_answer_pct_16_20)  AS r_star_vs_answer_
  CORR(avg_star_16_20, avg_linear_16_20)      AS r_star_vs_linear
FROM features;
```

✅ Completed.
✅ Completed.

r_star_vs_payment	r_star_vs_answer_pct	r_star_vs_linear
-0.050830	0.035801	0.985655

1 total rows Prev Page 1 of 1 Next Page Size 10

Finding:

We checked how different factors relate to patient satisfaction using a simple correlation test. The results showed that Average Total Payment and the Star Answer Percent both have very weak relationships with satisfaction—close to no link. In contrast, the HCAHPS Linear Mean Value moves strongly in the same direction as satisfaction, which means higher linear mean scores usually come with higher star ratings.

Conclusion:

Spending more or having a higher share of certain answers doesn't appear to make patients happier, based on this snapshot. The continuous HCAHPS Linear Mean Value seems to be a much better signal of patient experience than cost or answer percent.

Q5-2. What is the Correlation between different features?

df

```
WITH corr_calc AS (
  SELECT
    corr(avg_satisfaction, avg_total_payment)      AS corr_cost,
    corr(avg_satisfaction, bedsize_proxy)          AS corr_bedsize

    corr(avg_satisfaction, own_government)         AS corr_own_gov
    corr(avg_satisfaction, own_private)            AS corr_own_pri
    corr(avg_satisfaction, own_nonprofit)          AS corr_own_non

    corr(avg_satisfaction, type_acute)            AS corr_type_ac
    corr(avg_satisfaction, type_critical_access)  AS corr_type_cr
    corr(avg_satisfaction, type_childrens)        AS corr_type_ch

    corr(avg_satisfaction, reg_NE)                AS corr_region_
    corr(avg_satisfaction, reg_MW)                 AS corr_region_
    corr(avg_satisfaction, reg_S)                 AS corr_region_
    corr(avg_satisfaction, reg_W)                 AS corr_region_

    corr(avg_satisfaction, sev_low)               AS corr_severit
    corr(avg_satisfaction, sev_medium)            AS corr_severit
    corr(avg_satisfaction, sev_high)              AS corr_severit

    corr(avg_satisfaction, emergency_num)         AS corr_emergen
    corr(avg_satisfaction, ehr_num)               AS corr_ehr
  FROM hospital_data.satisfaction_feature_matrix
  WHERE avg_satisfaction IS NOT NULL
)

SELECT 'avg_total_payment' AS feature, corr_cost AS correlation F
SELECT 'bedsize_proxy', corr_bedsize FROM corr_calc UNION ALL

SELECT 'ownership_government', corr_own_government FROM corr_calc
SELECT 'ownership_private', corr_own_private FROM corr_calc UNION
SELECT 'ownership_nonprofit', corr_own_nonprofit FROM corr_calc U

SELECT 'type_acute', corr_type_acute FROM corr_calc UNION ALL
SELECT 'type_critical_access', corr_type_critical_access FROM cor
SELECT 'type_childrens', corr_type_childrens FROM corr_calc UNION

SELECT 'region_NE', corr_region_NE FROM corr_calc UNION ALL
SELECT 'region_MW', corr_region_MW FROM corr_calc UNION ALL
SELECT 'region_S', corr_region_S FROM corr_calc UNION ALL
SELECT 'region_W', corr_region_W FROM corr_calc UNION ALL

SELECT 'severity_low', corr_severity_low FROM corr_calc UNION ALL
SELECT 'severity_medium', corr_severity_medium FROM corr_calc UNI
SELECT 'severity_high', corr_severity_high FROM corr_calc UNION A

SELECT 'emergency_services', corr_emergency FROM corr_calc UNION
SELECT 'ehr_interoperability', corr_ehr FROM corr_calc

ORDER BY ABS(correlation) DESC;
```

✅ Completed.
✅ Completed.

feature	correlation
region_MW	0.192348
ownership_private	-0.146957
region_NE	-0.142459
ownership_nonprofit	0.116035
region_W	-0.093123
emergency_services	-0.031789
severity_low	0.030729
severity_high	-0.024137
avg_total_payment	-0.024117
severity_medium	0.019032

17 total rows Prev Page 1 of 2 Next Page Size 10

df

```
CREATE OR REPLACE TABLE hospital_data.feature_correlations AS
WITH corr_calc AS (
SELECT
  corr(avg_satisfaction, avg_total_payment) AS corr_cost,
  corr(avg_satisfaction, bedsize_proxy) AS corr_bedsize,
  corr(avg_satisfaction, own_government) AS corr_own_government,
  corr(avg_satisfaction, own_private) AS corr_own_private,
  corr(avg_satisfaction, own_nonprofit) AS corr_own_nonprofit,
  corr(avg_satisfaction, type_acute) AS corr_type_acute,
  corr(avg_satisfaction, type_critical_access) AS corr_type_criti
  corr(avg_satisfaction, type_childrens) AS corr_type_childrens,
  corr(avg_satisfaction, reg_NE) AS corr_region_NE,
  corr(avg_satisfaction, reg_MW) AS corr_region_MW,
  corr(avg_satisfaction, reg_S) AS corr_region_S,
  corr(avg_satisfaction, reg_W) AS corr_region_W,
  corr(avg_satisfaction, sev_low) AS corr_severity_low,
  corr(avg_satisfaction, sev_medium) AS corr_severity_medium,
  corr(avg_satisfaction, sev_high) AS corr_severity_high,
  corr(avg_satisfaction, emergency_num) AS corr_emergency,
  corr(avg_satisfaction, ehr_num) AS corr_ehr
FROM hospital_data.satisfaction_feature_matrix
WHERE avg_satisfaction IS NOT NULL
)
SELECT 'avg_total_payment' AS feature, corr_cost AS correlation F
SELECT 'bedsize_proxy', corr_bedsize FROM corr_calc UNION ALL
SELECT 'ownership_government', corr_own_government FROM corr_calc
SELECT 'ownership_private', corr_own_private FROM corr_calc UNION
SELECT 'ownership_nonprofit', corr_own_nonprofit FROM corr_calc U
SELECT 'type_acute', corr_type_acute FROM corr_calc UNION ALL
SELECT 'type_critical_access', corr_type_critical_access FROM cor
SELECT 'type_childrens', corr_type_childrens FROM corr_calc UNION
SELECT 'region_NE', corr_region_NE FROM corr_calc UNION ALL
SELECT 'region_MW', corr_region_MW FROM corr_calc UNION ALL
SELECT 'region_S', corr_region_S FROM corr_calc UNION ALL
SELECT 'region_W', corr_region_W FROM corr_calc UNION ALL
SELECT 'severity_low', corr_severity_low FROM corr_calc UNION ALL
SELECT 'severity_medium', corr_severity_medium FROM corr_calc UNI
SELECT 'severity_high', corr_severity_high FROM corr_calc UNION A
SELECT 'emergency_services', corr_emergency FROM corr_calc UNION
SELECT 'ehr_interoperability', corr_ehr FROM corr_calc
ORDER BY ABS(correlation) DESC;
```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_NS9-siSl3HL4r8K-05qGBOiNKuQf	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Findings:

None of the features show a strong linear correlation with total_charges. Most relationships are weak or moderate.

Conclusion:

This indicates that the charges are influenced by multiple independent hospital and patient factors.

Q6. Which hospitals achieve high satisfaction at relatively low cost? (Constructing the Value Index (z-scores per year))

df

```
CREATE OR REPLACE TABLE hospital_data.value_index_Q6 AS
WITH yearly_stats AS (
    SELECT
        year,
        AVG(satisfaction_score) AS mean_satisfaction,
        STDDEV_POP(satisfaction_score) AS std_satisfaction,
        AVG(avg_total_payment) AS mean_payment,
        STDDEV_POP(avg_total_payment) AS std_payment
    FROM hospital_data.panel_clean
    GROUP BY year
)

SELECT
    p.provider_id,
    p.hospital_name,
    p.state_abbr AS state,
    p.city,
    p.year,
    p.satisfaction_score,
    p.avg_total_payment,

    -- Z Scores
    (p.satisfaction_score - y.mean_satisfaction) / y.std_satisfac
    (p.avg_total_payment   - y.mean_payment)     / y.std_payment

    -- Value index = z_sat - z_cost
    ((p.satisfaction_score - y.mean_satisfaction) / y.std_satisfa
    - (p.avg_total_payment   - y.mean_payment)     / y.std_paymen

FROM hospital_data.panel_clean p
JOIN yearly_stats y USING (year);

```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_RLfqBtvjylj3PdY4yaDjiqS9FgKM	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Findings: The hospitals with the highest value-index scores all operate at low costs, yet achieve high satisfaction. The lowest scoring hospitals, such as Westchester Medical Center had high costs and lower satisfaction rates.

Conclusion:

The value index analysis supports the idea that increases in price does not equate to increases in care level or patient satisfaction. For example, Westchester Medical Center had transactions over $600 000 with a range of satisfaction scores as low as 7. This may suggest that operational differences between the hospitals simultaneously create differences in care level and operation costs. It may also suggest that increased prices lower patient satisfaction as expectations are increased as well.

Q7. What are the Hospitals in the Top 10% of the Value Index?

df

```
CREATE OR REPLACE TABLE hospital_data.top_value_hospitals_Q7_fina
WITH hospital_year AS (
  -- 1) Collapse to one row per hospital × year
  SELECT
    provider_id,
    ANY_VALUE(hospital_name) AS hospital_name,
    ANY_VALUE(state_abbr)    AS state,
    ANY_VALUE(city)          AS city,
    year,
    AVG(satisfaction_score) AS satisfaction_score,   -- average s
    AVG(avg_total_payment)  AS avg_total_payment     -- average c
  FROM hospital_data.panel_clean
  GROUP BY provider_id, year
),

stats AS (
  -- 2) Compute yearly mean & sd at the hospital-year level
  SELECT
    year,
    AVG(satisfaction_score)        AS mean_satisfaction,
    STDDEV_POP(satisfaction_score) AS std_satisfaction,
    AVG(avg_total_payment)         AS mean_payment,
    STDDEV_POP(avg_total_payment)  AS std_payment
  FROM hospital_year
  GROUP BY year
),

z_scores AS (
  -- 3) Calculate z-scores and value_index (still one row per hos
  SELECT
    h.*,
    (h.satisfaction_score - s.mean_satisfaction) / s.std_satisfac
    (h.avg_total_payment   - s.mean_payment)     / s.std_payment
    (
      (h.satisfaction_score - s.mean_satisfaction) / s.std_satisf
      - (h.avg_total_payment   - s.mean_payment)   / s.std_paymen
    ) AS value_index
  FROM hospital_year h
  JOIN stats s USING (year)
),

ranked AS (
  -- 4) Split each year into 10 deciles based on value_index → to
  SELECT
    *,
    NTILE(10) OVER (PARTITION BY year ORDER BY value_index DESC)
  FROM z_scores
),

dedup AS (
  -- 5) Safety: ensure only one row per hospital-year (in case of
  SELECT
    *,
    ROW_NUMBER() OVER (
      PARTITION BY provider_id, year
      ORDER BY value_index DESC
    ) AS rn
  FROM ranked
  WHERE decile = 1          -- only keep the top 10% hospitals in
)

SELECT
  provider_id,
  hospital_name,
  state,
  city,
  year,
  satisfaction_score,
  avg_total_payment,
  z_sat,
  z_cost,
  value_index
FROM dedup
WHERE rn = 1
ORDER BY year, value_index DESC;

```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_3jIVDKWt_nBLz-ILrnhcYhtbG0LG	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Findings: The top hospitals all had high satisfaction and low costs. Hospitals such as Oklahoma Heart Hospital, Llc, Caldwell Memorial Hospital, Inc, and Citizens Medical Center were top scorers that appeared multiple times. The value-indexes range from 2.90 to 1.81, suggesting that there are still significant differences within the top decile. Hospitals with higher costs also tend to be lower value-index scorers, even though satisfaction scores are relatively high within the top decile.

Conclusion: Because an increase in price leads to an decrease in value-index score, it means that there is either a redundantly circular relation between price and satisfaction, or there is a different variable that was not considered such as hospital size. As previously mentioned, hospital size may be one such candidate as smaller hospitals may be able to lower operating costs and provide personalized care in ways that are not scaleable to larger hospitals. Any further research should aim to clarify this relationship as well as explore other seemingly unrelated variables.

Q8. How did each state and territories change from 2016-2020?

df

```
CREATE OR REPLACE TABLE hospital_data.cost_sat_outliers_Q8 AS
WITH base AS (
  SELECT
    state,
    year,
    ROUND(AVG(SAFE_CAST(`Patient Survey Star Rating` AS FLOAT64))
      AS avg_star_rating,
    ROUND(AVG(SAFE_CAST(`Patient Survey Star Rating` AS FLOAT64))
    - LAG(ROUND(AVG(SAFE_CAST(`Patient Survey Star Rating` AS FLO
        OVER (PARTITION BY state ORDER BY year)
      AS yoy_change
  FROM hospital_data.patient_satisfaction_all
  WHERE SAFE_CAST(`Patient Survey Star Rating` AS FLOAT64) IS NOT
  GROUP BY state, year
)

SELECT
  *,
  MAX(CASE WHEN year = 2020 THEN avg_star_rating END)
    OVER (PARTITION BY state)
  -
  MAX(CASE WHEN year = 2016 THEN avg_star_rating END)
    OVER (PARTITION BY state)
      AS change_2016_to_2020
FROM base
ORDER BY state, year;

```

顯示隱藏的輸出內容

Findings:

Most states only saw modest improvements (.1–.3) from 2016–2020. It was also found that yoy_change fluctuated — some years showed positive movement, while others showed negative. There was also no significant jump in rating. No state saw an increase of +1.0 star from 2016–2020. There are meaningful differences in star ratings between states. For example, Nevada had some of the lowest ratings among states and territories with almost 2 stars, while Wisconsin had some of the highest ratings with close to 4 stars.

Conclusion:

From 2016–2020, satisfaction across states saw gradual improvement; however, the improvement in each state was not large. The change was not linear, and annual fluctuation was common, with the direction of change shifting year to year.

It is more important to note that differences between states are much larger than changes within states over time. This suggests that patient satisfaction is shaped more by regional factors than by short-term operational changes.

Overall, the data shows that hospitals across the country are making incremental improvements, but long-standing regional disparities remain, and there needs to be broader, systemic state-level change.

Q9. What are Cost-Satisfaction Outliers Using Quadrants?

Thought Process

In the previous questions, we mainly focused on the overall trend. But, the TA gave us some inspiration: Are there some hospitals that are particularly "strange"? For instance, they spend a lot of money but have very low satisfaction rates? Or conversely, are there hospitals that spend very little but provide excellent patient services? To ensure a fair comparison, we did not directly use the original "payment/satisfaction" figures. Instead, we adopted what TA suggested: converting both into z-scores. This approach enables us to place different hospitals on the same standard and assess how much they deviate from the average. We are specifically looking for two types of special hospitals:

• High-Cost Low-Satisfaction (spending a lot but with a poor experience)—— Hospitals that spend much more than average (z_cost > 1.5) but have below-average satisfaction (z_sat < –1.0)

• Low-Cost High-Satisfaction (spending less but with a great experience)—— Hospitals that spend much less (z_cost < –1.0) but provide better patient experience (z_sat > 1.5)

df

```
CREATE OR REPLACE TABLE hospital_data.cost_satisfaction_outliers_
WITH stats AS (
    SELECT
        AVG(avg_total_payment) AS mean_cost,
        STDDEV_POP(avg_total_payment) AS sd_cost,
        AVG(satisfaction_score) AS mean_sat,
        STDDEV_POP(satisfaction_score) AS sd_sat
    FROM hospital_data.panel_clean
),

zscores AS (
    SELECT
        p.*,
        (p.avg_total_payment - s.mean_cost) / s.sd_cost AS z_cost
        (p.satisfaction_score - s.mean_sat) / s.sd_sat AS z_sat
    FROM hospital_data.panel_clean p
    CROSS JOIN stats s
),

final AS (
    SELECT
        *,
        CASE
            WHEN z_cost > 1.5 AND z_sat < -1.0 THEN 'High-Cost Lo
            WHEN z_cost < -1.0 AND z_sat > 1.5 THEN 'Low-Cost Hig
            ELSE 'Normal'
        END AS outlier_flag
    FROM zscores
)

SELECT *
FROM final
WHERE outlier_flag != 'Normal'
ORDER BY year, outlier_flag, z_cost DESC;
```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_LV3VTrf69whbtKN91hCVM1GdVgic	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Findings:

After tagging the hospitals, the two outlier groups actually showed some clear patterns. The High-Cost Low-Satisfaction hospitals spend above the national average, but their satisfaction scores are still below 0 (after standardization).

The Low-Cost High-Satisfaction hospitals were the opposite. They spent noticeably less, but their satisfaction was far above average. This part stood out after using z-scores, because the raw numbers didn't look that obvious before.

The outlier detection helped uncover which hospitals are “paying more but not performing” and which ones are “doing more with less.”

Conclusion:

Facilities with higher charges tend to treat a larger share of severe or complex cases. The pattern differs across states and hospital groups, but cost levels generally rise with case difficulty.

Q10 - How do Hospitals Rank in Cost and Satisfaction by Hospital Case-Mix Severity Tier?

Thought Process

In earlier questions, we compared hospitals without considering how difficult the cases were. But this could be misleading, because hospitals treating more severe patients naturally spend more. So for Q10, we grouped hospitals by DRG severity to make the comparison fair. The goal was to see whether higher case complexity actually leads to higher costs and higher patient satisfaction.

We calculated the average payment and satisfaction for each severity tier—High, Medium, and Low. If higher spending reflected better care, the High-severity group should have higher satisfaction or a better value index. But the results showed the opposite: the High-severity tier had the highest costs but similar or even lower satisfaction. Meanwhile, the Low-severity tier provided the best experience at the lowest cost.

df

```
CREATE OR REPLACE TABLE hospital_data.hospital_drg_panel_tier_Q10
SELECT
  *,
  CASE
    WHEN Avg_Tot_Pymt_Amt < 5000 THEN 'Low cost'
    WHEN Avg_Tot_Pymt_Amt BETWEEN 5000 AND 15000 THEN 'Mid cost'
    ELSE 'High cost'
  END AS cost_tier
FROM hospital_data.hospital_drg_panel;

```

✅ Completed.
✅ Completed.

statement_type	job_id	location
CREATE_TABLE_AS_SELECT	job_Uezaj1qzFjPENSRYi64Ztg34ERF1	US

1 total rows Prev Page 1 of 1 Next Page Size 10

Findings:

As case complexity increases, hospital costs rise sharply, yet patient satisfaction remains largely unchanged, with the highest-severity group even showing slightly lower star ratings. In contrast, the lowest-severity group has the lowest costs but the highest satisfaction, suggesting that tougher cases do not lead to better perceived care. The value delivered per $1,000 spent drops substantially in high-severity hospitals—almost half of that in the lowest tier. Overall, higher costs reflect treating more difficult cases rather than delivering higher-quality patient experience.

Conclusion:

Higher case complexity is associated with higher payments, while patient satisfaction does not increase in the same way. Hospitals managing more complex cases face higher cost intensity without a clear pattern in satisfaction scores.

Final Conclusion

Several conclusions in EDA questions suggests that higher costs does not equate to higher patient satisfaction or higher value in care. In fact, higher costs often led to lower patient perceptions of care value. Whether examined across time or location, it held true that cost and satisfaction moved independently when examining all of the hospital within the dataset.

Smaller regional hospital such as Citizens Medical Center, Oklahoma Heart Hospital, Caldwell Memorial Hospital, and Rockcastle consistently demonstrated high levels of care for lower costs; suggesting that high value-index care is a repeatable process. Whether or not it can be scaled for larger hospitals remains unclear.

Analysis on the state level showed that there were little differences in care quality within the different cost tiers. Meanwhile, the overall correlation between cost and satisfaction (r ≈ 0.0004) confirms the idea that care can be provided at afforable costs.

The next steps of this project should aim to cover the project shortcomings. For example, treatment for different illnesses were compared on equal footing. A treatment method that has a low success rate and high price could be compared to