D214 Capstone Project:

PA2 – Healthcare Providers and Their Impact on Overall Hospital Ratings

Jason Willis

College of Information Technology,

Western Governors University

Dr. Daniel Smith

September 08, 2022

Table of Contents for Each Rubric

Part 1: Research Question

Describe Purpose, Summarize Research Question and Define Objectives: ***3***

Part 2: Data Collection

Summarize Data Preparation, Perform Exploratory Data Analysis: ***3***

Part 3: Data Extraction and Preparation

Describe Data Extraction and Preparation: ***3***

Part 4: Analysis

Describe Analysis Performed: ***3***

Part 5: Data Summary and Implications

Outcomes, Implications and Recommendations: ***3***

Part F: Sources

List Sources and Bibliography: ***8***

Hospital Ratings

The Patient Survey – Hospital Consumer Assessment of Healthcare Providers and Systems is a dataset provided by the Centers for Medicare and Medicaid Services. This survey poses questions asked of patients and their ratings over a few different clinical perspectives. What can a hospital learn from this survey? Can they affect the outcome, and if so, what services could they focus on? According to Schmocker (2015) "Readiness for discharge appears to be a clinically useful patient-reported metric, as those RFD have higher satisfaction with the hospital and physicians." Is this the only or best metric to use or can a hospital focus on provider care and strengthen their overall service ratings?

A – Research Question

Is communication from a doctor more statistically significant to a patient's overall hospital rating than a nurse?

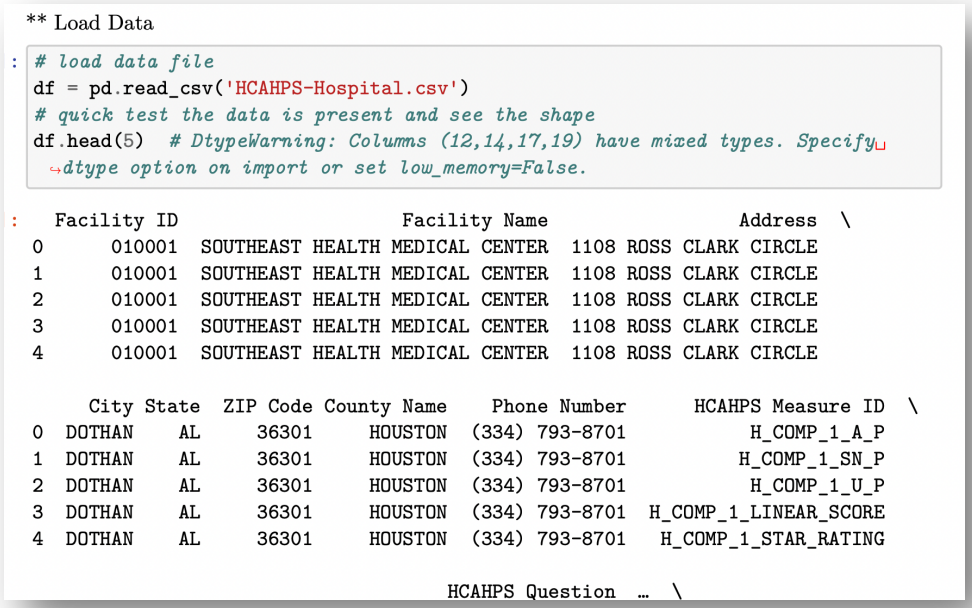
Null hypothesis – Doctor communication does not have a more statistically significant impact on the overall hospital rating when compared to a nurse.

Alternate Hypothesis - Doctor communication has a more statistically significant impact on the overall hospital rating when compared to a nurse.

B – Data Collection

The Patient Survey – Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) dataset (2022) was selected to provide questions posed to patients about their care. The questions focused on were: "Nurse Communication", "Doctor Communication" and "Overall Hospital Rating". A 5-star rating system was utilized.

This survey dataset captures all three measures compared in this analysis for over 4,000 clinics providing over 450,000 rows of data. Most of the time spent was to prepare the data for analytical techniques applied in this study. This effort will be discussed in section C – Data Extraction and Preparation. After the dataset was identified, it was downloaded from the Centers for Medicare & Medicaid Services (CMS, 2022) and then loaded into a data frame (Figure 1).



** Load Data

```
# load data file
df = pd.read_csv('HCAHPS-Hospital.csv')
# quick test the data is present and see the shape
df.head(5)  # DtypeWarning: Columns (12,14,17,19) have mixed types. Specify
↪dtype option on import or set low_memory=False.
```

	Facility ID	Facility Name	Address	\
0	010001	SOUTHEAST HEALTH MEDICAL CENTER	1108 ROSS CLARK CIRCLE	
1	010001	SOUTHEAST HEALTH MEDICAL CENTER	1108 ROSS CLARK CIRCLE	
2	010001	SOUTHEAST HEALTH MEDICAL CENTER	1108 ROSS CLARK CIRCLE	
3	010001	SOUTHEAST HEALTH MEDICAL CENTER	1108 ROSS CLARK CIRCLE	
4	010001	SOUTHEAST HEALTH MEDICAL CENTER	1108 ROSS CLARK CIRCLE	

	City	State	ZIP Code	County Name	Phone Number	HCAHPS Measure ID	\
0	DOTHAN	AL	36301	HOUSTON	(334) 793-8701	H_COMP_1_A_P	
1	DOTHAN	AL	36301	HOUSTON	(334) 793-8701	H_COMP_1_SN_P	
2	DOTHAN	AL	36301	HOUSTON	(334) 793-8701	H_COMP_1_U_P	
3	DOTHAN	AL	36301	HOUSTON	(334) 793-8701	H_COMP_1_LINEAR_SCORE	
4	DOTHAN	AL	36301	HOUSTON	(334) 793-8701	H_COMP_1_STAR_RATING	

	HCAHPS Question	…	\

*Figure 1 - Load Dataset from *.csv File*

C – Data Extraction and Preparation

Once the data is loaded into a data frame, unnecessary columns were dropped and renamed for easier processing. A data type warning was provided for columns with multiple data types since they would cause errors during the analysis process. Some columns had "Not Applicable" and "Not Available" mixed in the same column as the ratings provided by patients. These non-numerical data points were removed, and the data series was converted to an integer

data type. (Figure 2) While processing the data isn't as intuitive as using a graphical user interface, this approach is very efficient as data scales.

```
# Remove Unecessary Data Series
df_clean = df[['HCAHPS Answer Description', 'Patient Survey Star Rating']]

# Rename Columns
df_clean = df_clean.rename(columns={'HCAHPS Answer Description':'Questions',
                                    'Patient Survey Star Rating':'Ratings'})
```

```
# DtypeWarning: Columns (12,14,17,19) have mixed types. Specify dtype option on
 ↪import or set low_memory=False.

df_clean = df_clean.drop(df_clean[df_clean['Ratings'].isin(['Not Applicable',
 ↪'Not Available'])].index)  # Index --> of row
```

3

```
df_clean['Ratings'] = df_clean['Ratings'].astype(int)

df_clean.sample(20)
```

	Questions	Ratings
195019	Recommend hospital - star rating	4
167032	Nurse communication - star rating	4
265440	Doctor communication - star rating	3
135494	Overall hospital rating - star rating	4
136835	Staff responsiveness - star rating	3
43512	Quietness - star rating	2
255371	Overall hospital rating - star rating	2
241706	Summary star rating	3
342851	Discharge information - star rating	5
286237	Cleanliness - star rating	4
33233	Staff responsiveness - star rating	2
277574	Care transition - star rating	2
112362	Doctor communication - star rating	4
6881	Summary star rating	4
41466	Quietness - star rating	2
406193	Care transition - star rating	3
330854	Discharge information - star rating	4
216027	Quietness - star rating	4
188220	Quietness - star rating	2
15535	Nurse communication - star rating	4

Figure 2 - Remove Unnecessary Columns, Clean up Mixed Data Types and Rename Columns

D – Analysis

Exploratory data analysis was performed on the refined data frame. Here are a few explanations of what was performed:

- Info() method to verify column names, null-value counts, and data types; see Figure 3.
- Shape and describe() methods were used to understand the data frame's shape, count, unique categorical entries, most frequent with count, mean, standard deviation, minimal, maximum and quantiles 25%, 50% and 75% of the rating values. See Figure 4.
- Head() method was used to understand the layout of the data frame. An attribute of -5 showed the first and last 5 rows. See Figure 5.
- A Seaborn heatmap was used to show any null values graphically. Additionally, Pandas dropna() and .isnull() methods were used to help verify. See Figure 6.
- Questions and ratings were counted. Then the questions were grouped using the .groupby() method to show each question's mean rating value. See Figure 7.
- Ratings histogram was created, providing visual distribution. See Figure 8.
- Boxplots were created to display the minimum, first quartile, median, third quartile, and maximum values of each grouped question. See Figure 9.
- One-way Analysis of Variance (ANOVA) was calculated. See Figure 10. According to Norman, 2010 "Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of ''coming to the wrong conclusion''. These findings are consistent with empirical literature dating back nearly 80 years." One disadvantage of choosing ANOVA to analyze Likert scale data seemed to be within the limitation of the survey interpretations themselves. The questions to be rated are still able to be interpreted by the individual which may differ when compared to the research objectives.

```
df_clean.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 36520 entries, 4 to 449747
Data columns (total 2 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Questions  36520 non-null  object
 1   Ratings    36520 non-null  int64 
dtypes: int64(1), object(1)
memory usage: 855.9+ KB
```

Figure 3 - Pandas .info() Method

```
print("*******"*5)
print("* DataFrame Shape: ", df_clean.shape)
print("*******"*5)
df_clean.describe(include='all')
```

```
**********************************
* DataFrame Shape:  (36520, 2)
**********************************
```

	Questions	Ratings
count	36520	36520.000000
unique	11	NaN
top	Nurse communication - star rating	NaN
freq	3320	NaN
mean	NaN	3.158050
std	NaN	1.008707
min	NaN	1.000000
25%	NaN	2.000000
50%	NaN	3.000000
75%	NaN	4.000000
max	NaN	5.000000

Figure 4 - Pandas .describe() Method

```
df_clean.head(-5)
```

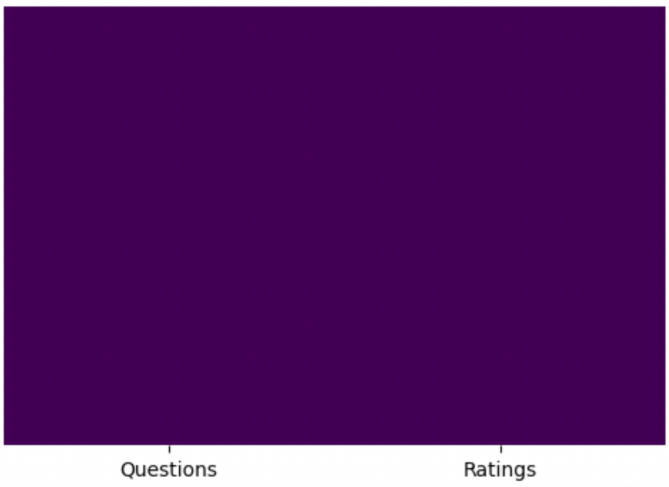
	Questions	Ratings
4	Nurse communication - star rating	3
18	Doctor communication - star rating	3
32	Staff responsiveness - star rating	2
43	Communication about medicines - star rating	3
53	Discharge information - star rating	4
...	...	...
449673	Doctor communication - star rating	4
449687	Staff responsiveness - star rating	3
449698	Communication about medicines - star rating	3
449708	Discharge information - star rating	4
449717	Care transition - star rating	4

36515 rows × 2 columns

Figure 5 - Pandas head() Method

Check for Missing or Null Values

```
# Mapping to view missing data...none present.
fig, ax = plt.subplots(figsize=(6,4))         # Sample figsize in inches
sns.heatmap(df_clean.isnull(), yticklabels=False, cbar=False, cmap='viridis');
```




```
# Drop any null columns
df_clean = df_clean.dropna()

print("*******"*5)
print("* Any Rows Missing: ",df_clean.isnull().all(axis=1).any())
print("*******"*5)
print("Any Null Values:\n", df_clean.isnull().any())
```

```
**********************************
* Any Rows Missing:  False
**********************************
Any Null Values:
 Questions    False
Ratings      False
dtype: bool
```

Figure 6 - Check for Missing or Null Values

```
print('********'*5)
print('*** Describe Data ***')
print('********'*5)
print('* Median: ',df_clean.median())
print('********'*5)

print('Mode: ' + str(df_clean['Questions'].value_counts(ascending=True).loc[lambda x : x>1].to_
      '\n\n' + str(df_clean['Ratings'].value_counts(ascending=True).loc[lambda x : x>1].to_fram
```

```
****************************************
*** Describe Data ***
****************************************
* Median:  Ratings    3.0
dtype: float64
****************************************
Mode:                                        Questions
Nurse communication - star rating           3320
Doctor communication - star rating          3320
Overall hospital rating - star rating       3320

   Ratings
1      374
5      995
2     1903
4     3189
3     3499
```

```
df_grouped = df_clean.groupby(['Questions'],as_index=False).mean() #["Patient Survey Star Ratin
print(df_grouped)
```

```
                              Questions   Ratings
0     Doctor communication - star rating  3.238253
1      Nurse communication - star rating  3.259940
2  Overall hospital rating - star rating  3.263253
```

Figure 7 - Group Question and Rating Data to Aggregate

```
# Ratings Distribution
df_clean['Ratings'].plot.hist();

plt.xlabel('# of Star Rating')
plt.ylabel('# of Patients')
plt.title('Rating Distribution');
```

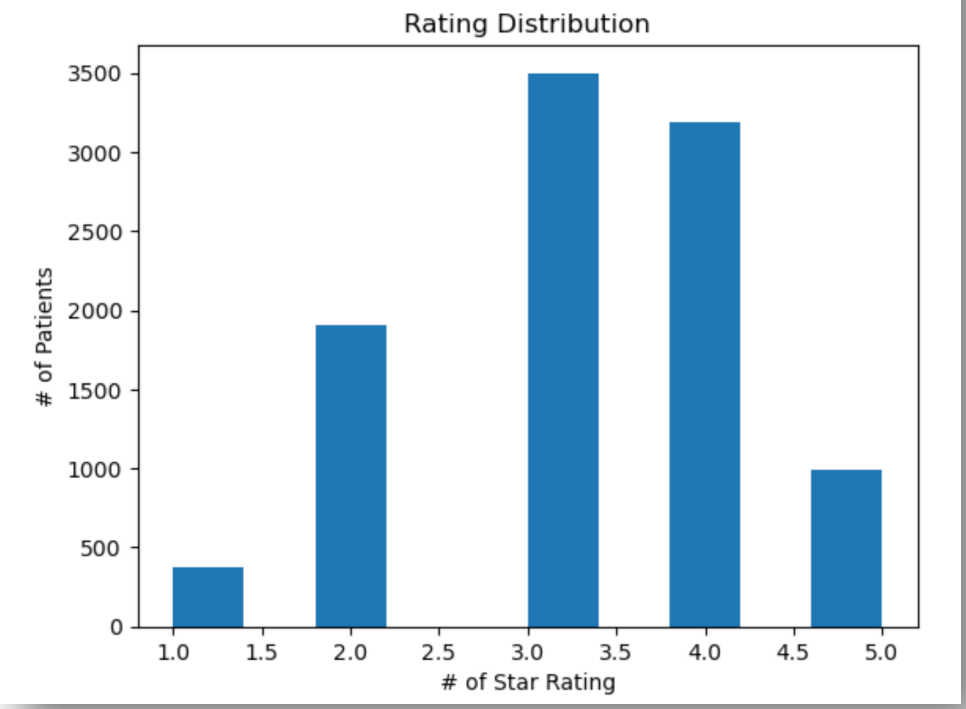


Figure 8 - Ratings Distribution

```
# Create Boxplot
df_clean.boxplot('Ratings', by='Questions', figsize=(10, 6))

control = df_clean['Ratings'][df_clean['Questions'] == 'Overall hospital rating - star rating']

grps = pd.unique(df_clean['Questions'].values)
d_data = {grp:df_clean['Ratings'][df_clean['Questions'] == grp] for grp in grps}

k = len(pd.unique(df_clean['Questions']))  # number of conditions
N = len(df_clean.values)  # conditions times participants
n = df_clean.groupby('Questions').size()[0] #Participants in each condition
```

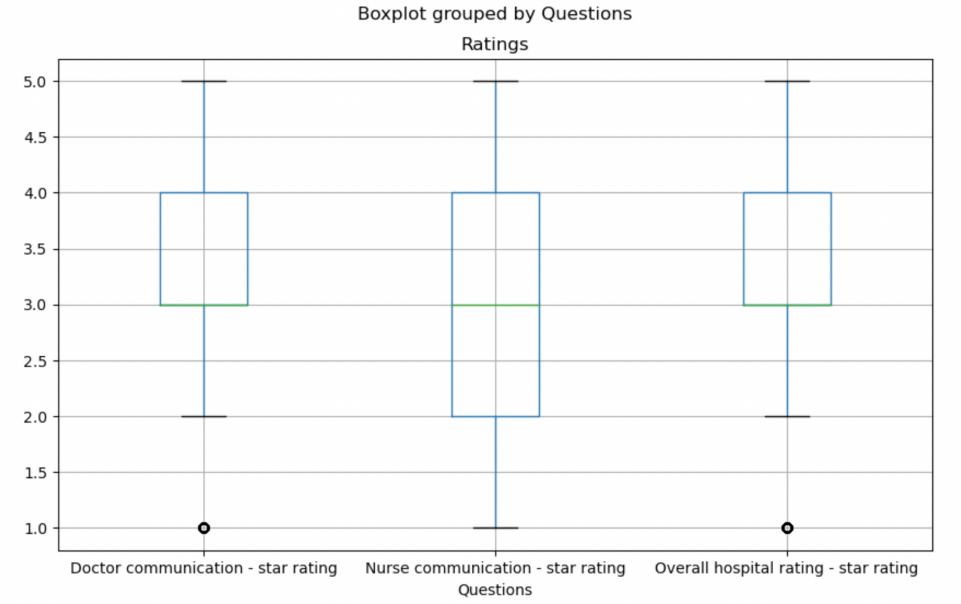


Figure 9 - Boxplot of Questions

```
# Set up ANOVA Model
mod = ols('Ratings ~ C(Questions)', # Note the Catigorical Data C()
                data=df_clean).fit()

# Carry out the ANOVA
aov_table = sm.stats.anova_lm(mod)
print(aov_table)
```

```
                   df       sum_sq   mean_sq          F    PR(>F)
C(Questions)      2.0     1.224297  0.612149   0.614114  0.541141
Residual       9957.0  9925.130723  0.996799        NaN       NaN
```

```
print(mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Ratings   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.6141
Date:                Thu, 08 Sep 2022   Prob (F-statistic):              0.541
Time:                        00:24:30   Log-Likelihood:                -14115.
No. Observations:                9960   AIC:                         2.824e+04
Df Residuals:                    9957   BIC:                         2.826e+04
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================================================
====
                                                                 coef    std err          t      P>|t|      [0.025      0.
975]
--------------------------------------------------------------------------------------------------------------------------------
----
Intercept                                                    3.2383      0.017    186.886      0.000       3.204
3.272
C(Questions)[T.Nurse communication - star rating]            0.0217      0.025      0.885      0.376      -0.026
0.070
C(Questions)[T.Overall hospital rating - star rating]        0.0250      0.025      1.020      0.308      -0.023
0.073
==============================================================================
Omnibus:                      241.290   Durbin-Watson:                   0.859
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              147.053
Skew:                          -0.151   Prob(JB):                     1.17e-32
Kurtosis:                       2.487   Cond. No.                         3.73
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure 10 - One-Way ANOVA

E – Data Summary and Implications

Review of hypothesis: Is communication from a doctor more statistically significant to a patient's overall hospital rating than a nurse?

Null hypothesis – Doctor communication does not have a more statistically significant impact on the overall hospital rating when compared to a nurse.

Alternate Hypothesis - Doctor communication has a more statistically significant impact on the overall hospital rating when compared to a nurse.

Analysis of Variance, ANOVA for convenience, was performed on the data set to ascertain if "…a significant difference among the groups tested" as stated by Dr. Sewell during hist Lecture:

D207 T2 – Welcome to D207 EDA Webinar. (n.d.) ANOVA uses an F-statistic which measures mean equality of a group and a p-value to measure probability under the assumed hypotheses. The F-statistic of the data was 0.6141 and the p-value was 0.541; thus, we fail to reject the null hypotheses. To express this in another way, the result states the means are at least as far apart as observed; given there are no underlying differences between said means. Analysis shows a tight range between doctor, nurse and overall ratings. Both independent variables seem to be important to a hospitals overall rating.

Recommendations: Since a patients experience with their doctor and nurse are both important to the hospital's overall rating, continued training and improved provider/patient relations should be strived for. Additionally, more specific questions could be added to the patient survey to dig deeper into understanding what key behaviors could be championed to improve. A limitation within the current survey point's to how providers communicate, but this really isn't the whole story. Trying to understand why a patient provided a certain rating will help illuminate where focus is needed.

F – Sources

- Help using Markdown: https://www.markdownguide.org/basic-syntax/
- MacTeX: https://tug.org/mactex/mactex-download.html
- Matplotlib Help: https://matplotlib.org/2.1.2/api/_as_gen/matplotlib.pyplot.plot.html
- Numpy Help: https://numpy.org/doc/stable/
- Pandas Help: https://pandas.pydata.org/docs/user_guide/index.html#user-guide
- Python Help: https://docs.python.org/3.9/library/index.html
- Scipy.stats Help: https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html
- Matplotlib: https://matplotlib.org/stable/index.html
- Seaborn: https://seaborn.pydata.org/api.html
- References: See the references section.

References

Sewell, W. (n.d.). Lecture: D207 T2 – Welcome to D207 EDA Webinar. Western Governors University. Found Here: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=fcf752f1-6ff7-4286-9100-ad1f016a98d6

Patient Survey – Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS). (2022). Centers for Medicare & Medicaid Services (CMS). Found Here: https://data.cms.gov/provider-data/dataset/dgck-syfz

Norman G. Likert scales, levels of measurement and the "laws" of statistics. Adv Health Sci Educ Theory Pract. 2010 Dec;15(5):625-32. doi: 10.1007/s10459-010-9222-y. Epub 2010 Feb 10. PMID: 20146096.

Schmocker R.K., Holden S.E., Vang X, Leverson G.E., et. al., Association of Patient-Reported Readiness for Discharge and Hospital Consumer Assessment of Health Care Providers and Systems Patient Satisfaction Scores: A Retrospective Analysis. J Am Coll Surg. 2015 Dec;221(6):1073-82.e1-3. doi: 10.1016/j.jamcollsurg.2015.09.009. Epub 2015 Sep 25. PMID: 26474513; PMCID: PMC4662900.