JUNE 18, 2024

CAPSTONE D214

A Regression Analysis Project

ROBERT PATTON
WESTERN GOVERNORS' UNIVERSITY
Rpatt33@wgu.edu

Table of Contents

Part I: Research

A. Research Question

1. To what extent does the county, OSHPDID, and hospital system variables affect the number of adverse events?
2. This data analysis project aims to interpret if there is a correlation between specific hospitals and their locations throughout the state of California, and whether those factors have an influence on the number of adverse events that take place after medical procedures. Adverse events, in excess, should raise concern for hospital executives as this can impact the hospital rating. Hospitals with a poor rating may indicate that there is poor quality of care being provided to the local community, showing that there is an increased risk for receiving care there. In addition, another assumption for poor hospital ratings may be that a hospital lacks specific medical specialties, such as neurological services, that may contribute to an increase in stroke adverse events. However, it may not be lack of services or poor quality of care at all, rather a hospitals location and higher adverse event rates could be related to the population majority that the hospital serves, such as the elderly.
3. Such assumptions, and potentially many more, may be able to explain the correlation between hospitals, their locations, and adverse event rates. This correlation analysis could then help to improve hospital ratings which would improve overall operations. Most people derive their public opinions about hospitals from friends, new articles, or hearing about the general reputation of a hospital throughout the community, and the public uses this information to make sound judgments on where to get care for themselves and family members (DeAngelis 2016). A hospital lacking trust in their local community may be subject to a decline in hospital visits, lowering hospital wide census metrics that impact revenue. On the other hand, higher adverse event rates may also contribute to longer hospital admission, or an increase in the amount of hospital readmissions. This should be very important to hospital executives because the Center for Medicare and Medicaid services evaluates these metrics and can fine hospitals for having too many readmissions within a certain time frame. For example, in 2017 imposed fines by the CMS for hospital readmissions equated to over half a billion dollars (Upadhyay et al. 2019). Overall, this analysis project could benefit California hospitals by improving their hospital ratings by improving metrics on adverse events. Therefore, the proposed hypothesis is that the county, hospital, and hospital system variables do have statistical significance that affects the number of adverse events, whereas the null hypothesis would state that the county, hospital, and hospital system variables do not have statistical significance in the number of adverse events occurring.

Part II: Data Collection

B. Data Collection Process

1. Data relevant to this project will be collected from data.gov, from the California Department of Health Care Access and Information. It is a public data set named California Hospital Performance Ratings. The data was downloaded in a CSV format and will be imported into a Jupyter notebook, for use with Python, to manipulate the data for the analysis project. It has been updated as recently as March 30th, 2024. The data set identifies hospital performance ratings based on adverse events for various performance measures, or medical conditions/procedures. The data set has 25,974 entries and 13 columns. The data set looks at hospital ratings for 336 hospitals throughout the state of California, calculating each hospital's rating based on 18 performance measures and the adverse events of total cases for each measure. The hospital rating is categorized into three classes, "As expected", "Better", and "Worse", and is determined by calculating a risk-adjusted rate based on the total number of cases and the adverse events observed for those total cases.
2. Advantages to collecting this data is that the research and data collection was done by the California Department of Health Care Access and Information. They had also already constructed a CSV file of the data which makes it very easy to obtain and import into Jupyter notebook. The one disadvantage of data collection for this project is that the data dictionary was not very elaborate. For example, there was missing values in the CSV file and no explanation as to why. This provided a challenge because removing values from the data, without context of why they are missing, could impact analysis and provide an inaccurate result.
3. To overcome this temporary roadblock, the California Department of Health Care Access and Information had their contact information on the databases page via data.gov. An email was sent regarding questions about the missing values, how the risk-adjusted rates were calculated, and to get access to any supporting material relevant to the CSV file. A prompt response by the authors highlighted that the missing values meant that either the hospital did not report any cases or cases equated to less than 3 and were not included in the research. In addition, supplemental information was provided that elaborated on the study that can be used for referencing.

Part III: Data Extraction and Preparation

C. Data Extraction Preparation and Process

1. This data analysis project is being submitted in a Jupyter notebook file, using Python as the primary programming language, and is where all data extraction and preparation takes place. Commentary was added before each line of code to provide emphasis on what is being performed during data preparation. Two Jupyter notebooks were used to perform this data analysis, one for performing all the exploratory data analysis and one for executing the regression model. This was done for clarity purposes and to help convey the importance

of data pre-processing in the presentation part of this project. The first task for this project was to import all necessary packages into the Jupyter notebook shell, and then import the California hospital performance ratings CSV file. Once imported, exploratory data analysis can be performed, and the data can be examined/transformed.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

```
#Read in the csv file
df=pd.read_csv('/Users/robertpatton/Desktop/Desktop - Robert's MacBook Pro/D214/Hospital_Ratings.csv')
```

```
#Examine data info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25974 entries, 0 to 25973
Data columns (total 13 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Year                 25974 non-null  int64
 1   County               25974 non-null  object
 2   hospital             25974 non-null  object
 3   OSHPDID              25974 non-null  int64
 4   system               25974 non-null  object
 5   Type of Report       25974 non-null  object
 6   Performance Measure  25974 non-null  object
 7   # of Adverse Events  18552 non-null  float64
 8   # of Cases           18552 non-null  float64
 9   Risk-adjusted Rate   18438 non-null  float64
 10  Hospital Ratings     18552 non-null  object
 11  Longitude            25876 non-null  float64
 12  Latitude             25876 non-null  float64
dtypes: float64(5), int64(2), object(6)
memory usage: 2.6+ MB
```

```
#Examine first five rows
df.head(10)
```

	Year	County	hospital	OSHPDID	system	Type of Report	Performance Measure	# of Adverse Events	# of Cases	Risk-adjusted Rate	Hospital Ratings	Longitude	Latitude
0	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	Pneumonia	2.0	76.0	3.0	As Expected	-122.253	37.76266
1	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	Heart Failure	2.0	111.0	2.1	As Expected	-122.253	37.76266
2	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	GI Hemorrhage	5.0	83.0	4.6	As Expected	-122.253	37.76266
3	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	PCI	NaN	NaN	NaN	NaN	-122.253	37.76266
4	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	Acute Stroke Subarachnoid	NaN	NaN	NaN	NaN	-122.253	37.76266
5	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	Acute Stroke Hemorrhagic	5.0	9.0	48.9	Worse	-122.253	37.76266
6	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	AMI	3.0	17.0	16.1	As Expected	-122.253	37.76266
7	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	Acute Stroke	9.0	74.0	23.6	Worse	-122.253	37.76266
8	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	Acute Stroke Ischemic	4.0	65.0	15.2	As Expected	-122.253	37.76266
9	2016	Alameda	Alameda Hospital	106010735	Alameda Health System	IMI	Carotid Endarterectomy	NaN	NaN	NaN	NaN	-122.253	37.76266

2. An advantage of using Python is the pre-built in packages used for data analysis, such as Pandas, NumPy, Matplotlib, Stats models, SKlearn, SciPy and Seaborn. Used for reading the data, cleaning the data, renaming data columns, dummy variables, and creating linear models, the packages are powerful tools for handling all kinds of data. Using the Pandas package, missing values and duplicates were easily identifiable and managed by removing the NaN values. Such values were removed as they either represented no reported cases or case totals amounted to less than three. Therefore, removing these values, rather than replacing them with the mean, median, or mode, maintains the integrity of the data by retaining only actual reported cases. Being a common programming language, Python is a

powerful tool for data management and makes it very easy to deal with large data sets. On the other hand, the disadvantage of Python is that it can perform slowly during execution of certain lines of code, taking up a lot of memory (Joy 2019).

```
#Check for missing values
df.isnull().sum()
```

```
Year                      0
County                    0
hospital                  0
OSHPDID                   0
system                    0
Type of Report            0
Performance Measure       0
# of Adverse Events    7422
# of Cases             7422
Risk-adjusted Rate     7536
Hospital Ratings       7422
Longitude                98
Latitude                 98
dtype: int64
```

```
#Drop missing values
df.dropna(inplace=True)
```

```
#Confirm missing values managed
df.isnull().sum()
```

```
Year                   0
County                 0
hospital               0
OSHPDID                0
system                 0
Type of Report         0
Performance Measure    0
# of Adverse Events    0
# of Cases             0
Risk-adjusted Rate     0
Hospital Ratings       0
Longitude              0
Latitude               0
dtype: int64
```

```
#Check for duplicate values
df.duplicated()
```

```
0        False
1        False
2        False
5        False
6        False
         ...
25968    False
25969    False
25970    False
25971    False
25973    False
Length: 18386, dtype: bool
```

```
#Drop columns irrelevent to study
df=df.drop(['Year', 'hospital', 'Type of Report', 'Performance Measure', 'Risk-adjusted Rate', 'Hospital Ratings', 'Longitude', 'Latitude', '# of Cases'], axis=1
```

```
#Confirm variables dropped
df.head()
```

	County	OSHPDID	system	# of Adverse Events
0	Alameda	106010735	Alameda Health System	2.0
1	Alameda	106010735	Alameda Health System	2.0
2	Alameda	106010735	Alameda Health System	5.0
5	Alameda	106010735	Alameda Health System	5.0
6	Alameda	106010735	Alameda Health System	3.0

```
#Rename # of adverse events column
df.rename(columns={'# of Adverse Events': '#_Adverse_Events'}, inplace=True)
df
```

	County	OSHPDID	system	#_Adverse_Events
0	Alameda	106010735	Alameda Health System	2.0
1	Alameda	106010735	Alameda Health System	2.0
2	Alameda	106010735	Alameda Health System	5.0
5	Alameda	106010735	Alameda Health System	5.0
6	Alameda	106010735	Alameda Health System	3.0
...	...	...	...	...
25968	Yuba	106580996	Adventist Health Systems	13.0
25969	Yuba	106580996	Adventist Health Systems	3.0
25970	Yuba	106580996	Adventist Health Systems	19.0
25971	Yuba	106580996	Adventist Health Systems	16.0
25973	Yuba	106580996	Adventist Health Systems	0.0

18386 rows × 4 columns

```
#Re-express County column to numerical values and use one-hot encoding
df=pd.get_dummies(df, columns=['County'], prefix=None, drop_first=True)
df.head()
```

	OSHPDID	system	#_Adverse_Events	County_Amador	County_Butte	County_Calaveras	County_Colusa	County_Contra Costa	County_Del Norte	County_El Dorado	...	County_Solano	County_Sonoma	County_Stanislaus	County_Tehama	County_Trinity	County_T
0	106010735	Alameda Health System	2.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	106010735	Alameda Health System	2.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	106010735	Alameda Health System	5.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
5	106010735	Alameda Health System	5.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
6	106010735	Alameda Health System	3.0	0	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 57 columns

```
#Get unique value count for system variable
print(df['system'].value_counts().sort_index(ascending=True))
```

```
AHMC Healthcare, Inc.                326
Adventist Health Systems             808
Alameda Health System                197
Alta Hospitals System                210
Avanti Hospitals                     173
Cedars-Sinai Health System           209
County of Los Angeles                234
Dignity Health                      1790
HCA Healthcare Corporation           384
Independent/Other                   5153
John Muir Health                     144
KPC Healthcare, Inc.                 354
Kaiser Foundation Hospitals         2192
Kindred Healthcare, Inc.               3
Loma Linda University                155
MemorialCare                         237
Prime Healthcare Services            752
Providence St. Joseph Health        1175
Scripps Health                       295
Sharp Healthcare                     282
Sutter Health                       1193
Tenet Healthcare Corporation         824
Universal Health Services, Inc.      266
University of California             476
University of Southern California    133
Verity Health System                 405
Vibra Healthcare                      16
Name: system, dtype: int64
```

```
#Get unique value counts for system
df.system.nunique()
```

```
27
```

```
#Rename system to System
df.rename(columns={'system': 'System'}, inplace=True)
df
```

	OSHPDID	System	#_Adverse_Events	County_Amador	County_Butte	County_Calaveras	County_Colusa	County_Contra Costa	County_Del Norte	County_El Dorado	...	County_Solano	County_Sonoma	County_Stanislaus	County_Tehama	County_Trinity	C
0	106010735	Alameda Health System	2.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	106010735	Alameda Health System	2.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	106010735	Alameda Health System	5.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
5	106010735	Alameda Health System	5.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
6	106010735	Alameda Health System	3.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
25968	106580996	Adventist Health Systems	13.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
25969	106580996	Adventist Health Systems	3.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
25970	106580996	Adventist Health Systems	19.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
25971	106580996	Adventist Health Systems	16.0	0	0	0	0	0	0	0	...	0	0	0	0	0	
25973	106580996	Adventist Health Systems	0.0	0	0	0	0	0	0	0	...	0	0	0	0	0	

3. Exploratory data analysis (EDA) will be used to examine the data set's shape, size, characteristics, and values. During this phase, missing values, outliers, duplicate data, number of hospitals, number of adverse events, and various other metrics will be identified and transformed to answer the research question. With these insights, a regression model will be produced to help determine if hospital location has any correlation to higher rates of adverse events. Each of the mentioned packages within Python will allow for the creation of visualizations, statistical metric retrieval, and the creation/implementation of a regression model. During EDA, the data will be transformed in a way that helps determine which variables are best for a regression model. The EDA phase also provides insights for how to transform the data to run in a regression model most effectively.

Part IV: Analysis

D. Data Analysis Process

1. Using linear regression allows a data analyst to assess the strength of a relationship between a dependent variable and multiple independent variables, also called predictor variables. In addition, linear regression can help determine how significant each predictor variable is to a dependent variable, thus allowing an analyst to eliminate those that have little to no correlation. Justification of the regression model will come from the R-squared metric, which evaluates model performance in regression analysis by representing a proportion of variance in the dependent variable (Chugh 2020). This data analysis project will specifically use a multiple linear regression model to answer the research question. Being an extension of linear regression, multiple linear regression is used when there are two or more independent/predictor variables and one dependent variable (Badole 2024). The project uses the County, OSHPDID, and System variables as the independent variables and the # of Adverse Events as the dependent variable, justifying the need and use of a multiple linear regression model.
2. Before running a multiple linear regression model, the first step is to ensure that all the variables are of numeric value. In order to do this, one-hot encoding, also called re-expression of variables, can be implemented to meet this requirement. One-hot encoding re-expresses the categorical/text values of a column into each of their own columns and gives a 0 or 1 numeric value for each new column, a 0 means that data from other variables is not related to that column and a 1 means that column is related to the data. For example, the county variable needed to be one-hot encoded to meet the linear regression numeric requirement, turning that variable into 54 new columns. Below a screen shot provides a visual reference for this process. The system variable was also one-hot encoded and created 26 new variables. With each of the variables in the data set now having a numeric value, the regression model can be created to answer the research question.
3. After cleaning the data, dropping insignificant variables to the study, manipulating the data with one-hot encoding, and performing all other EDA tasks, the regression model can be fed the appropriate variables and model performance can be assessed. An initial regression

model is run on the independent/predictor variables. These are compared against the dependent variable to determine correlation. The ordinary least-squared results can be seen below. The initial regression model has an R squared value of just 0.091, meaning that 0.91% of the model can explain variation. This infers that the predictor variables influence # of adverse events. Following the initial model, a heatmap is constructed to look for values that have higher correlation. A backwards stepwise elimination method is then used to reduce the initial model, using p-values to determine which variables should be removed. A correlation heatmap helps give a more detailed insight and visualization of which predictor variables may be correlated to the dependent variable. In addition, a heatmap will allow an analyst to see where variables with high multicollinearity may be present. Looking at the output of the initial regression model, four variables are identified as not having significance to the model, with their p-values being greater than 0.05. These will be removed one at a time and a new regression model will be ran after removing each one until the model is left with only variables having a p-value less than or equal to 0.05. Variables removed from the model were 'System_Kindred Healthcare, Inc.', 'System_Verity Health System', 'System_Vibra Healthcare', and 'System_KPC Healthcare, Inc.'. The final and reduced regression model has an R-squared value of 0.090, meaning that the model explains 90% of the variation. This model was very accurate and based on these results it can confidently be said that the independent variables do influence adverse events. Furthermore, future adverse events and standards of error residuals are created and added to the data frame as variables from the reduced regression model results, using the predict() function from Sklearn. Residuals are an important metric for regression analysis, where residual standard of error helps measure how well a regression model fits a dataset, measuring the standard deviation of residuals in a regression model (Bobbitt 2021). Reduced residuals are plotted in a scatter plot, looking for points to be densely packed around the regression line and then the standard of error is calculated. Reduced residuals standard of error should be close to zero, where the smaller the standard of error the better the model, this regression analysis has a standard of error of only 11.38 and indicates that the regression model fits the dataset well.

4. Linear regression models are great and effective tools for determining correlation between variables, however. Just as advantageous as they may be, they also have their disadvantages. For the purposes of this project, it is required to specify one of each per the rubric. First, computational efficiency is important when performing data analysis. Not only does it make an analyst's job easier to be able to run an analysis project faster, but not having to over work a devices computational limit is also beneficial. That is why using a linear regression model is an advantage for this project, it does not require complicated calculations, runs predictions fast on large amounts of data, and the modeling speed is quite fast (Satyavishnumolakala 2020). One of the bigger disadvantages of linear regression modelling is the assumption of linearity. Regression analysis itself assumes that there is linearity, independence, and constant variance, which may be untrue in most real-world

situations and an analyst may need to pay great deal of attention to this matter in order to produce a truly efficient model (Simplilearn 2024). Again, these two examples are not expressive of the full pros and cons to regression analysis.

5. Screen shots below for reference of the above data analysis process.

```
#Re-express County column to numerical values and use one-hot encoding
df=pd.get_dummies(df, columns=['County'], prefix=None, drop_first=True)
df.head()
```

	OSHPDID	system	#_Adverse_Events	County_Amador	County_Butte	County_Calaveras	C
0	106010735	Alameda Health System	2.0	0	0	0	
1	106010735	Alameda Health System	2.0	0	0	0	
2	106010735	Alameda Health System	5.0	0	0	0	
5	106010735	Alameda Health System	5.0	0	0	0	
6	106010735	Alameda Health System	3.0	0	0	0	

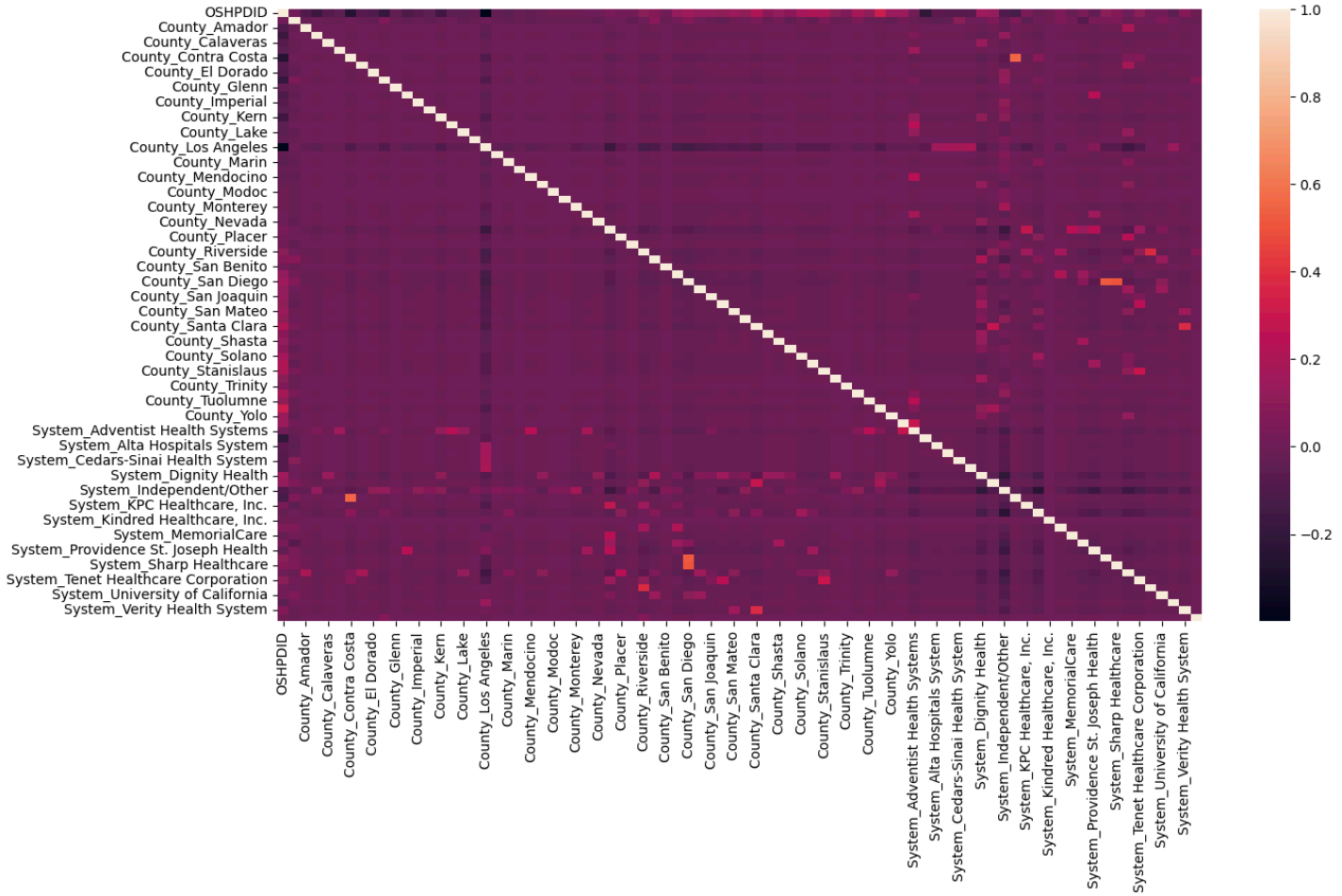
5 rows × 57 columns

```
#Define dependent and predictor variables and create initial regression model
y= df['#_Adverse_Events']
X= df.drop(['#_Adverse_Events'], axis=1).assign(const=1)
```

```
#Create initial regression model
df_model=sm.OLS(y, X)
df_results= df_model.fit()
print(df_results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       #_Adverse_Events   R-squared:                       0.091
Model:                            OLS   Adj. R-squared:                  0.087
Method:                 Least Squares   F-statistic:                     22.54
Date:                Thu, 13 Jun 2024   Prob (F-statistic):          2.00e-310
Time:                        09:55:05   Log-Likelihood:                -70763.
No. Observations:               18386   AIC:                         1.417e+05
Df Residuals:                   18304   BIC:                         1.423e+05
Df Model:                          81
Covariance Type:            nonrobust
==============================================================================
```

```
#Create inital heatmap
clean_df_initial_heatmap= df
plt.figure(figsize=(15,8))
ax= sns.heatmap(clean_df_initial_heatmap.corr(), annot=False)
plt.show()
```




```
#Remove System_KPC Healthcare, Inc. for final model
y= df['#_Adverse_Events']
X= df.drop(['#_Adverse_Events', 'System_Kindred Healthcare, Inc.', 'System_Verity Health System', 'System_Vibra Healthcare', 'System_KPC Healthcare, Inc.'], axis=1).assign(const=1)

reduced_model=sm.OLS(y, X)
reduced_results= reduced_model.fit()
print(reduced_results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       #_Adverse_Events   R-squared:                       0.090
Model:                            OLS   Adj. R-squared:                  0.087
Method:                 Least Squares   F-statistic:                     23.63
Date:                Sun, 16 Jun 2024   Prob (F-statistic):          8.26e-312
Time:                        11:01:47   Log-Likelihood:                -70766.
No. Observations:               18386   AIC:                         1.417e+05
Df Residuals:                   18308   BIC:                         1.423e+05
Df Model:                          77
Covariance Type:            nonrobust
==========================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
```

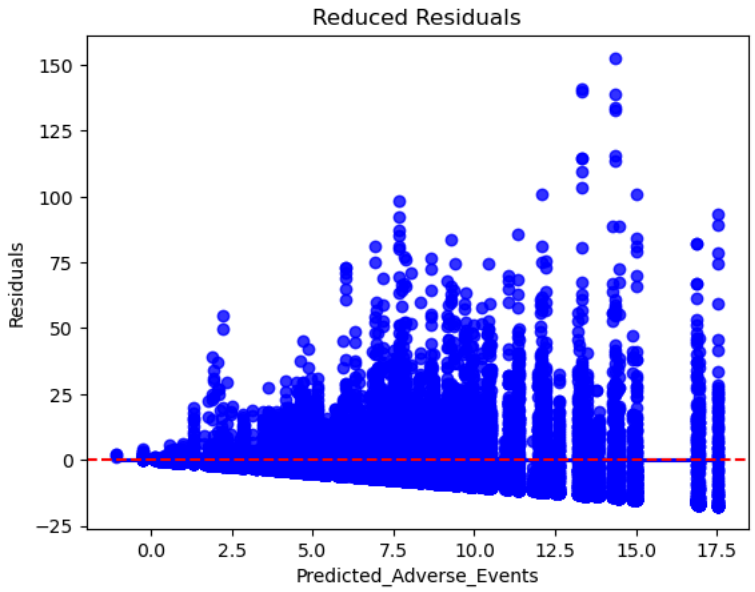
```
#Create dataframe of predicted values and residuels

df["Predicted_Adverse_Events"] = reduced_results.predict(X)
df["Residuals"] = reduced_results.resid
df.head()
```

	OSHPDID	#_Adverse_Events	County_Amador	County_Butte	County_Calaveras	County_Colusa
0	106010735	2	0	0	0	0
1	106010735	2	0	0	0	0
2	106010735	5	0	0	0	0
5	106010735	5	0	0	0	0
6	106010735	3	0	0	0	0

5 rows × 84 columns

```
#Create scatterplot of residuals
df['Intercept']=1
sns.regplot(x='Predicted_Adverse_Events', y='Residuals', data=df, color='blue', ci=None)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Reduced Residuals')
plt.show()
```



Predicted_Adverse_Events	Residuals
3.701752	-1.701752
3.701752	-1.701752
3.701752	1.298248
3.701752	1.298248
3.701752	-0.701752

```
#Residual standard of error
model = sm.OLS(y, X).fit()
model.resid.std(ddof=X.shape[1])
```

```
11.382601686854398
```

Part V: Data Summary and Implications

E. Summarize Data Analysis

1. This project began with establishing three points of reference for data analysis:
   A research question: To what extent does the county, OSHPDID, and hospital system variables affect the number of adverse events?
   A null hypothesis: The county, OSHPDID, and hospital system variables do not have statistical significance in the number of adverse events occurring.
   An alternate hypothesis: The county, OSHPDID, and hospital system variables do have statistical significance that affect the number of adverse events.

   The reduced final regression model, using multiple linear regression, had an R-squared value of 0.090 or 90%. As a result, the null hypothesis is rejected in favor of the alternate hypothesis. Creating of residuals was done for confirmation of model performance, it can be seen in the residuals scatter plot that the model does indeed produce accurate results for answering the research question. Furthermore, the residual standard of error model produced a result of 11.38, also supporting the regression model performance. With these results, it can be said that the extent to which the hospital location variables affect the number of adverse events is statistically significant. Therefore, it can be hypothesized that this correlation could be impacting hospital ratings and further research could help improve those ratings.
2. Although the results of this analysis are pleasing, there could be limitations that affect this analysis. The biggest limitation for this analysis is that it only includes four years' worth of data, and some of the data had to be removed because either the hospitals did not report any cases, or the values were less than 3. However, even if cases were less than 3, it could still be of value to the study and produced very different results had those been included. If there was ten years' worth of data for this project, the results would probably be a lot more accurate and there would probably be data available for each of those that were dropped.
3. Based on the results from this analysis, adverse events are correlated to hospital location. This could be because of many factors, such as patient population type, average distance from residences to a hospital, population size that creates more traffic, health hazards in the community, etc. Whatever the influencing factors, hospital ratings reflect an increase in adverse events. When hospitals have high rates of the events, community members are more likely to seek care elsewhere, hospitals may receive more fines, and business operations and revenue may become impacted. Therefore, based on these results, it would be wise for hospital executive to conduct further research into factors outside of the hospital that me contribute to an increase in adverse events. Improving overall hospital ratings can improve the way people perceive the facility and instill trust in the local community.
4. Future study on this matter can be improved in many ways, however. I propose that one method for improving this study would be to look at ambulance transport times for those

patients experiencing the performance measures/medical conditions listed in this data set. Adding that metric to this data set would most certainly help explain a lot of the adverse event outcomes. Communities where hospitals have negative ratings and higher adverse events rates may see some correlation between what happens before patients arrive to the hospitals, via emergency vehicles, and that of measures taken in hospital. Another measure that could improve this study would be combine some of the performance measures. For example, there are four stroke categories in this study: Acute Stroke Subarachnoid, Acute Stroke Hemorrhagic, Acute Stroke, and Acute Stroke Ischemic. Subarachnoid and hemorrhagic strokes could be combined into one category as both infer that a patient had a "bleeding brain stroke". Whereas an acute stroke and acute ischemic stroke infer a "blood clot brain stroke". By combining the categories, the study can be simplified into the two main types of strokes and research could go into identifying what is the root cause of adverse events related to hemorrhagic and ischemic strokes, one of the most debilitating and life threating medical conditions.

F. Resources

Badole, M. (2024, January 17). Multiple linear regression : Definition , example, and applications. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/multiple-linear-regression-using-python-and-scikit-learn/

Bobbitt, Z. (2021, May 11). How to interpret residual standard error. Statology. https://www.statology.org/how-to-interpret-residual-standard-error/

Chugh, A. (2024, January 18). Mae, MSE, RMSE, coefficient of determination, adjusted R squared - which metric is better?. Medium. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

DeAngelis, C. D. (2016, December). How helpful are hospital rankings and ratings for the public's health? The Milbank quarterly. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5192961/

Department of Health Care Access and Information. (2024, March 30). State of California - California Hospital Performance Ratings. Catalog. https://catalog.data.gov/dataset/california-hospital-performance-ratings-91d9b

Dowell, E. K. P. (2021, October 8). Census Bureau's 2018 County Business Patterns provides data on over 1,200 Industries. Census.gov. https://www.census.gov/library/stories/2020/10/health-care-still-largest-united-states-employer.html

Joy, A. (2019, September 14). 5 main disadvantages of python programming language. Pythonista Planet. https://pythonistaplanet.com/disadvantages-of-python/#

Satyavishnumolakala. (2020, June 12). Linear regression -Pros & Cons. Medium. https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0

Seabury, S., Bognar, K., Xu, Y., Huber, C., Commerford, S. R., & Tayama, D. (2017, March 19). Regional disparities in the quality of stroke care. The American Journal of Emergency Medicine. https://www.sciencedirect.com/science/article/pii/S0735675717302127#:~:text=Non%2Dmetropolitan%20hospitals%20performed%20worse,versus%2082.7%25%2C%20respectively).

Simplilearn. (2024, March 5). What is regression analysis? types: Examples: Uses. Simplilearn.com. https://www.simplilearn.com/tutorials/excel-tutorial/regression-analysis

Upadhyay, S., Stephenson, A. L., & Smith, D. G. (2019). Readmission Rates and Their Impact on Hospital Financial Performance: A Study of Washington Hospitals. Inquiry: a journal of medical care organization, provision, and financing, 56, 46958019860386. https://doi.org/10.1177/0046958019860386