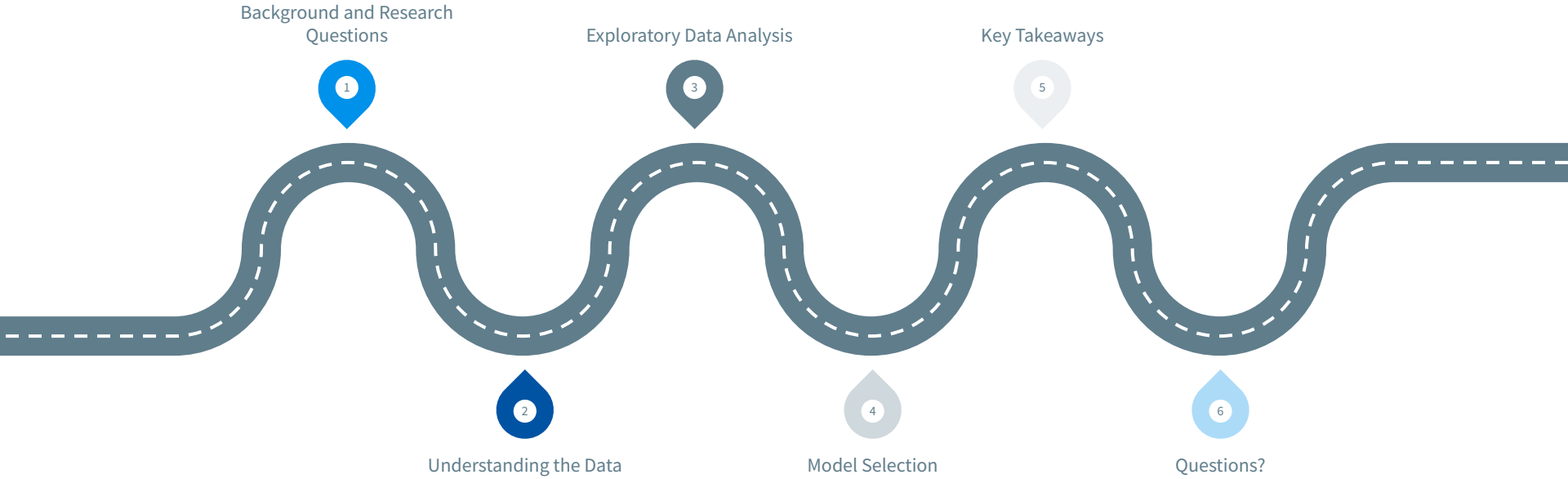
Data Analysis of 2011-2015 California Hospital Ischemic Stroke Patient 30-Day Readmissions/Mortality Risk Adjusted Rates

By Alex McCorriston

Roadmap


Background and Research Questions
1
Exploratory Data Analysis
3
Key Takeaways
5
2
Understanding the Data
4
Model Selection
6
Questions?

1.

Background and Research Questions

Background: What is an Ischemic Stroke?

- Occurs when blood clots or other particles block the blood vessels to the brain, preventing brain tissue from getting oxygen/nutrients
- Most common cause: Atherosclerosis
- Common temporary or permanent symptoms:
  - Inability to move on one side of the body
  - Weakness on one side of the body
  - Problems with thinking, awareness, attention, learning, judgment, and memory
  - Problems understanding or forming speech
  - Problems with controlling or expressing emotions
  - Numbness or strange sensations
  - Pain in the hands and feet that worsens with movement and temperature changes
  - Depression

750,000+

people in the United States have a stroke every year.

~87%

of all strokes are ischemic strokes.

5th

major cause of death and leading cause of adult disability in the United States

Other Important Definitions

30-Day Readmissions Rate

Percentage of admitted patients who return to the hospital within 30 days of initial discharge.

30-Day Mortality Rate

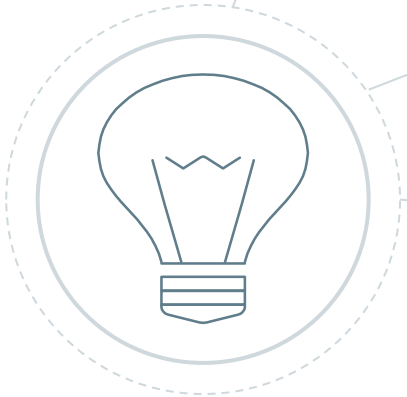
Percentage of admitted patients who die within 30 days of initial admission to hospital.

Risk Adjusted Rate

Number of 30-day Readmissions or Mortality Admissions / Total Number of Admissions is adjusted to account for pre-existing health problems that put some patients at greater risk of readmission/death

"Levels the playing field"

Research Questions:



- What variables, if any, influence California hospitals' risk adjusted rates for ischemic stroke patients?
- Is there a significant difference between the 30-day readmission data and 30-day mortality data?

Importance of Data Analysis

- Identifying any significant predictors of increased hospital readmissions and mortality rates can potentially help hospitals make appropriate changes to lower them and ideally result in…

  - Lower healthcare costs
  - Improved quality of care and patient outcomes
  - Increased patient satisfaction

2.

Understanding the Data

Data Description

Ischemic Stroke 30-Day Mortality and 30-Day Readmissions Risk Adjusted Rates Data

- Data set found on data.world but original source is https://data.chhs.ca.gov
- URL: https://data.world/chhs/06ed38d3-b047-4ae2-aa00-2e43b5491d6e
- Data set contains 2188 rows and 11 columns (288 unique hospitals)
- Panel data: Mix of cross-sectional and time series data
  - Years: 2011-2012, 2012-2013, 2013-2014, and 2014-2015
- No potential predictor variables in data

Field Title	Field Name	Data type	Description
Year	Year	plain text	Year of Discharge
County	County	plain text	County
Hospital	Hospital	plain text	Hospital
OSHPD ID	OSHPDID	plain text	Hospital OSHPD ID
Measure	Measure	plain text	Outcome Measure
Risk Adjusted Rate	Risk Adjusted Rate	number	Risk Adjusted 30-day Mortality/Readmission Rates presented here adjust the observed mortality rates. This statistical methodology takes into account pre-existing health problems that put some patients at greater risk of death/readmission to "level the playing field" and allow fair comparisons across hospitals.
Number of 30-day Deaths/ Readmissions	# of Deaths/Readmissions	number	Number of 30-day Deaths/Readmissions in California acute care hospitals
Number of Total Admissions	# of Cases	number	Number of Total Admissions with Ischemic Stroke in California acute care hospitals
Hospital Ratings	Hospital Ratings	plain text	Hospital Performance Ratings based on a 98% Confidence Interval (CI). If a hospital's upper CI is less than the statewide observed rate, it is designated as performing "better" than the average hospital. If a hospital's lower CI is greater than the state rate, it is designated as a performing "worse" than the average state hospital.
Longitude	Longitude	number	Longitude of the hospital
Latitude	Latitude	number	Latitude of the hospital

DV

Data Description Continued

California Licensed and Certified Healthcare Facility Listings Data

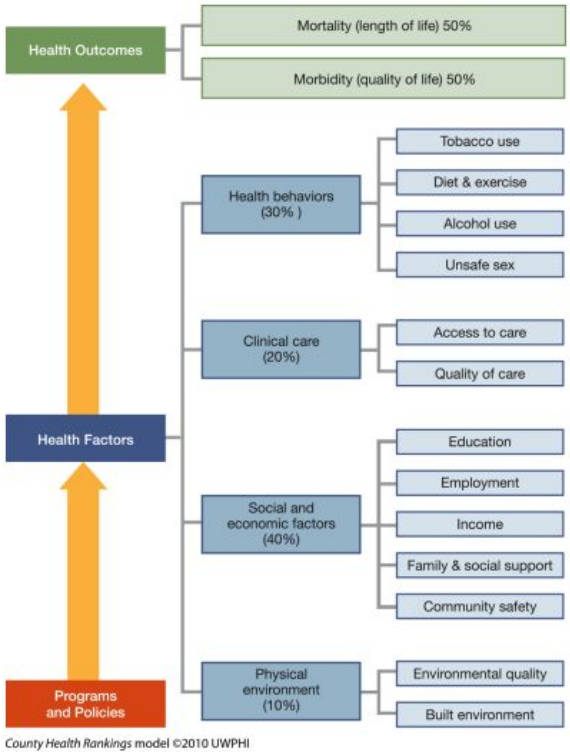
- Data set found on https://data.chhs.ca.gov
- URL: https://data.chhs.ca.gov/dataset/healthcare-facility-locations
- Data set includes 59 variables but only 4 used to merge with original stroke data
- stroke = merge(stroke, hospital_facilities_ca, by.x = "OSHPDID", by.y = "OSHPD_ID", all.x = T)

VARIABLE	FORMAT	DEFINITION
CAPACITY	Numeric	CAPACITY is the same as TOTAL_CAPACITY in the FACILITY table in ELMS, referring to the number of licensed beds, and obtained from the Facility's license application form.
ZIP	String	ZIP is the same as FAC_ZIP5 in the FACILITY table in ELMS, obtained from the facility's license application form.
ENTITY_TYPE_DESCRIPTION	String	ENTITY_TYPE_DESCRIPTION is the type of entity of the facility. See Lookup table.
OSHPD_ID	String	OSHPD_ID is a unique nine-digit identifier assigned to each facility by the Office of Statewide Health Planning and Development. OSHPDs financial and utilization databases begin with a 3-digit number that indicates the type of facilityÂ (106=hospital, 206=long term care, 306 = clinic, 406 = home health/hospice agency). The next two digits indicate the county in which the facility is located. The last four digits are unique within each county. Data Source: https://data.chhs.ca.gov/ Data Definition: http://www.oshpd.ca.gov/HID/Data_Request_Center/Data_Documentation.html

Data Description Continued

California County Health Ratings

- Data sets found on https://countyhealthrankings.org
- URL: https://www.countyhealthrankings.org/explore-health-rankings/california/data-and-resources
- Data sets include rankings on two measures (health factors and health outcomes) by county and year
- Rankings added to original data set



County Health Rankings model ©2010 UWPHI

2011 Rankings:

Rank	Health Outcomes	Rank	Health Factors
1	Marin	1	Marin
2	San Benito	2	Placer
3	Placer	3	Santa Clara
4	Santa Clara	4	San Mateo
5	San Mateo	5	Nevada
6	Orange	6	San Luis Obispo
7	Santa Cruz	7	San Francisco
8	Colusa	8	El Dorado
9	Yolo	9	Napa
10	El Dorado	10	Santa Cruz
11	Nevada	11	Sonoma
12	Sonoma	12	Orange
13	San Luis Obispo	13	Santa Barbara
14	Napa	14	Ventura
15	Monterey	15	Mono
16	San Diego	16	Contra Costa
17	Ventura	17	Yolo
18	Santa Barbara	18	Alameda
19	Contra Costa	19	Amador
20	Calaveras	20	Inyo

Data Description Continued

American Community Survey Data

- Variables obtained through use of tidycensus package (get_acs() function)
- Predictor variables of interest extracted: Median Household Income Estimate, CHCI, Population Estimate, Poverty Rate, Percent Population Ages 65-84
    - CHCI calculated using calculation weights on relevant variables
    - Percent Population Ages 65-84 obtained by adding DPO5_0015PE and DPO5_0016PE variables

Variable	Definition	Calculation Weight (CHCI)
B19013_001	Median Household Income	
DP02_0060PE	Percent Population Ages 25+ Education Attainment - Less than 9th grade	50
DP02_0061PE	Percent Population Ages 25+ Education Attainment - 9th-12 Grade No Diploma	100
DP02_0062PE	Percent Population Ages 25+ Education Attainment - High School Graduate	120
DP02_0063PE	Percent Population Ages 25+ Education Attainment - Some College No Degree	130
DP02_0064PE	Percent Population Ages 25+ Education Attainment - Associate's Degree	140
DP02_0065PE	Percent Population Ages 25+ Education Attainment - Bachelor's Degree	190
DP02_0066PE	Percent Population Ages 25+ Education Attainment - Graduate or Professional Degree	230
B01003_001	Population Estimate	
DP03_0119PE	Poverty Rate	
DP05_0015PE	Percent Population Ages 65-74	
DP05_0016PE	Percent Population Ages 75-84	

CHCI

Data Cleaning/Wrangling

- Edit Hospital column of data so all OSHPDIDs and Hospital names match
  - Raw data has 288 unique OSPHDIDs and 395 unique Hospital names
    - Ex. Mercy Hospital <96> Bakersfield
    - Ex. Community Hospital Monterey PeniAs Expectedula
- Separate Location variable into Latitude and Longitude variables
- After stroke and hospital facilities merge, there are 150 rows with NAs
  - 10 rows have missing Risk Adjusted Rates → drop
  - 140 rows have missing Capacity, Zipcode, and Entity Type → fill manually (mainly due to hospital closures since time of data collection)
    - Missing data found on https://hcai.ca.gov
  - Resulting data has 2170 rows

Data Cleaning/Wrangling Continued

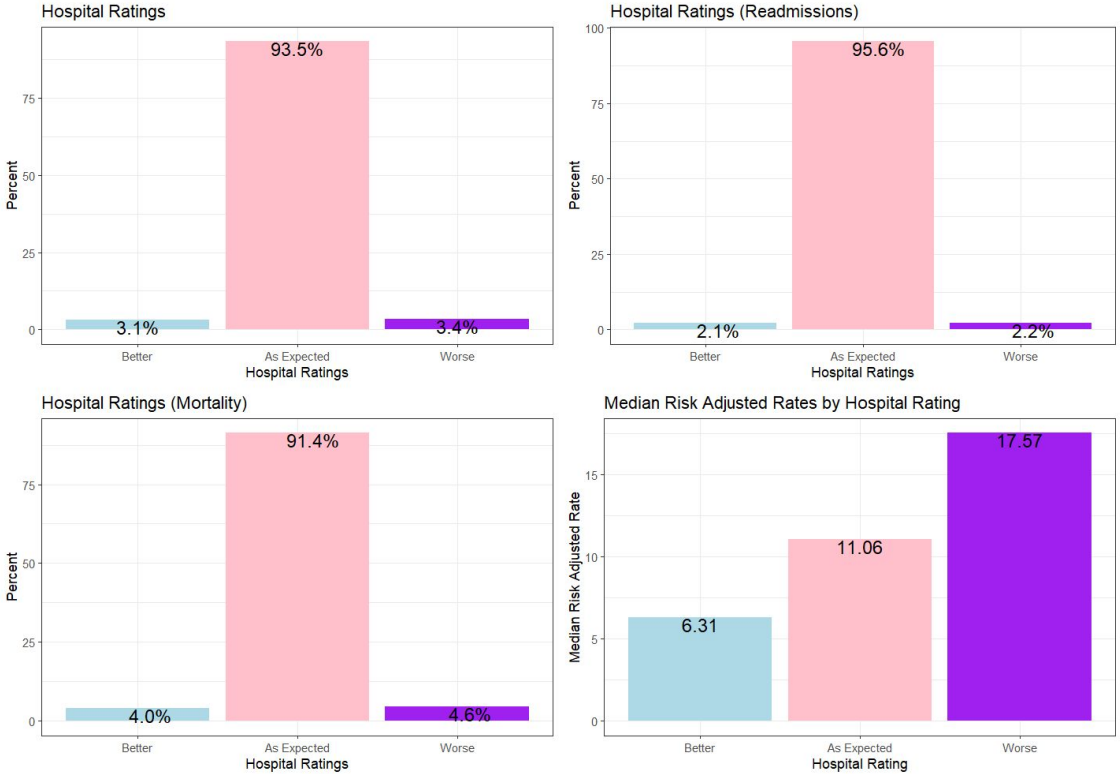
- Make data frame with missing values and merge it with stroke data frame
- Flip Number of 30 Day Readmissions/Deaths and Number of Total Admissions values for 2013-2014
- Subset stroke data frame by year in order to add in County Health data and Census variables since these values differ by year
- Use rbind() function to combine data that was previously split according to year
- Separate data by Measure (readmissions and mortality) → separate models will be run on each

Number of 30 Day Readmissions/Deaths	Number of Total Admissions
500	48
531	57
126	4
122	18
616	90

3.

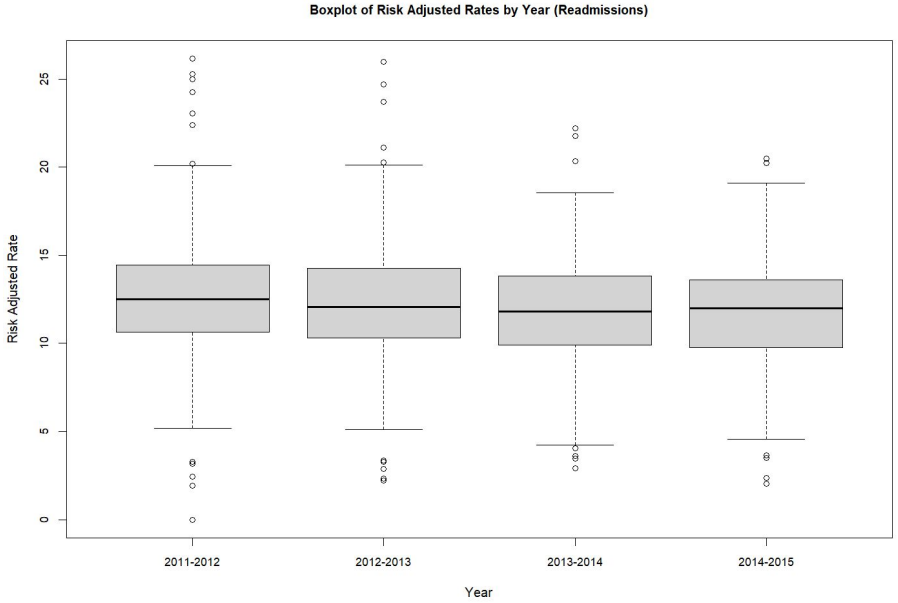
Exploratory Data Analysis

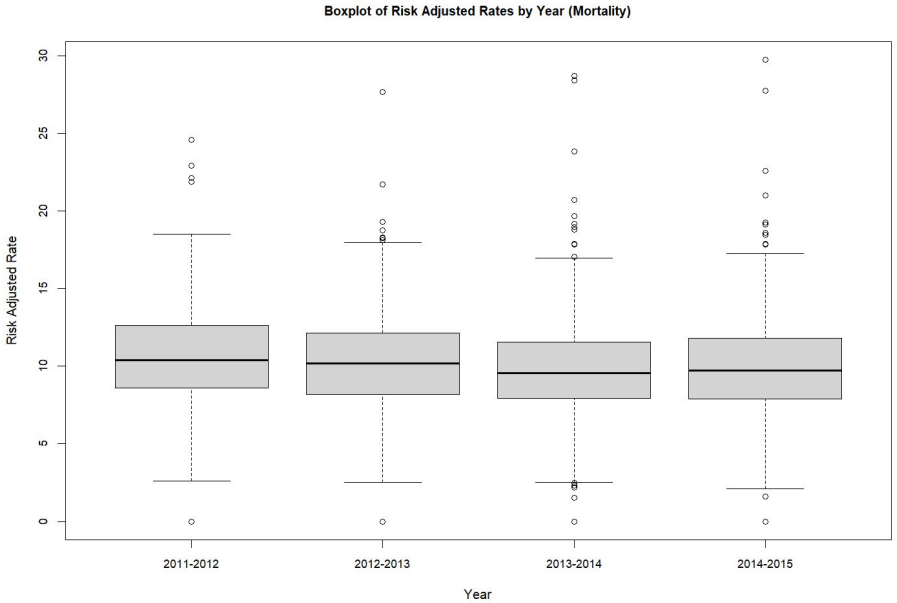
EDA: Hospital Ratings



- Majority of hospitals are rated “As Expected”
- Higher median Risk Adjusted Rate is correlated with “Worse” Hospital Rating
- Lower median Risk Adjusted Rate is correlated with “Better” Hospital Rating

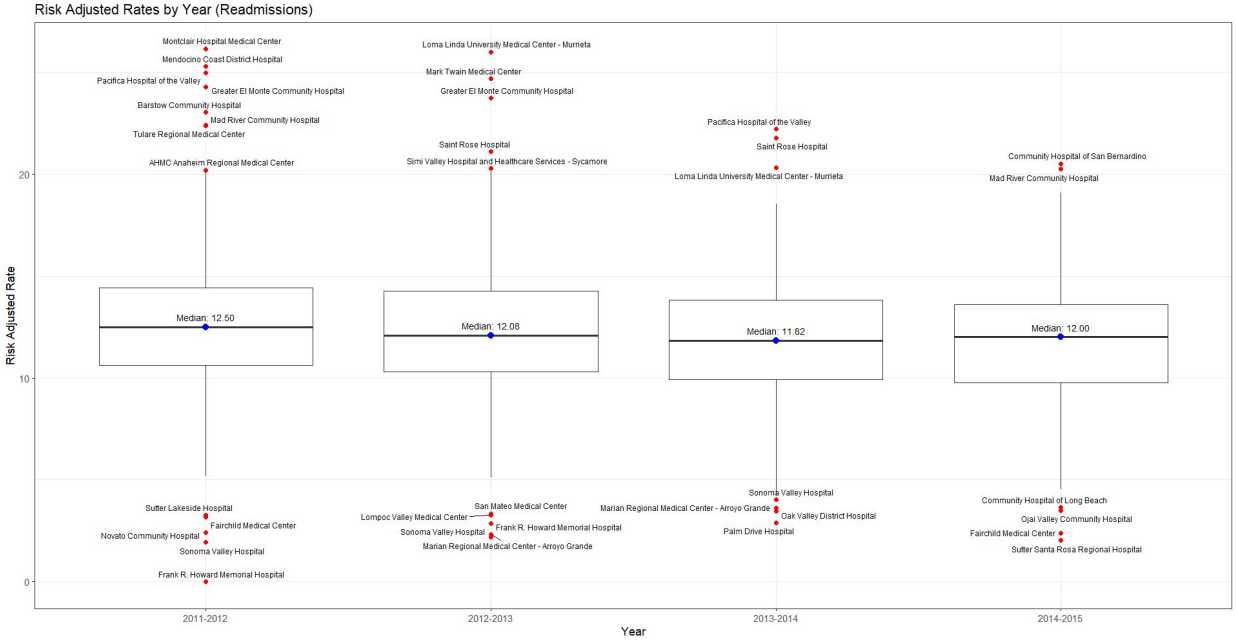
EDA: Boxplot of Hospital Risk Adjusted Rates by Year





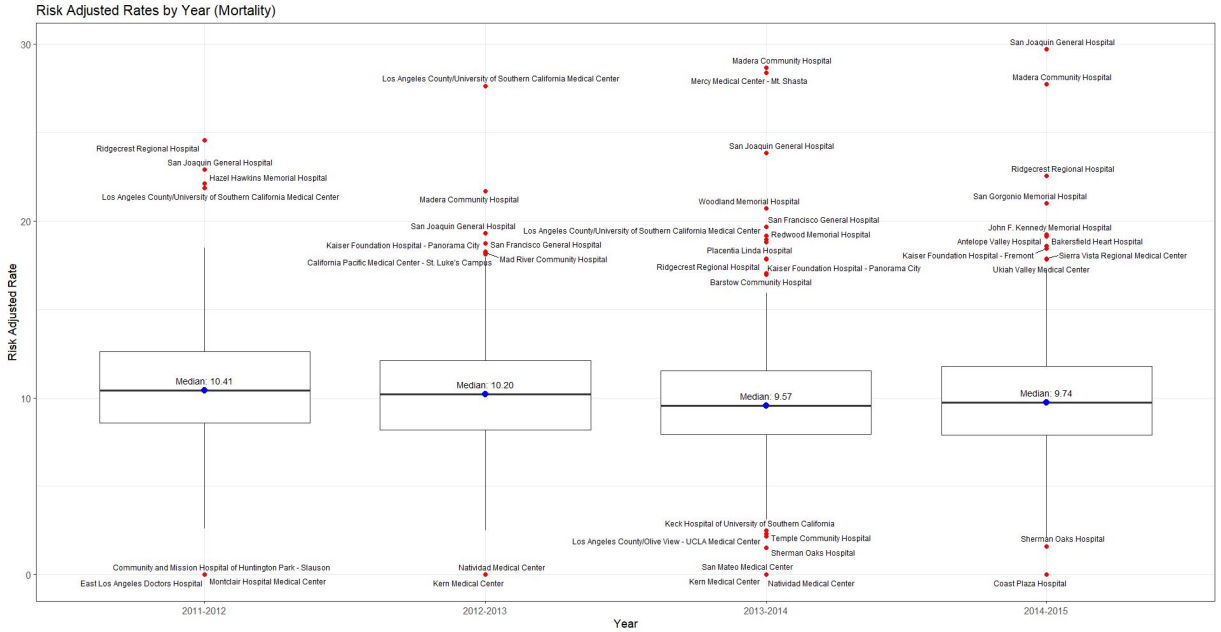
- For both data frames, median Risk Adjusted Rate decreases from 2011-2014 but then slightly increases from 2014-2015
- Median Risk Adjusted Rate is overall lower for the Mortality data than the Readmissions data

EDA: Boxplot of Risk Adjusted Rates by Year



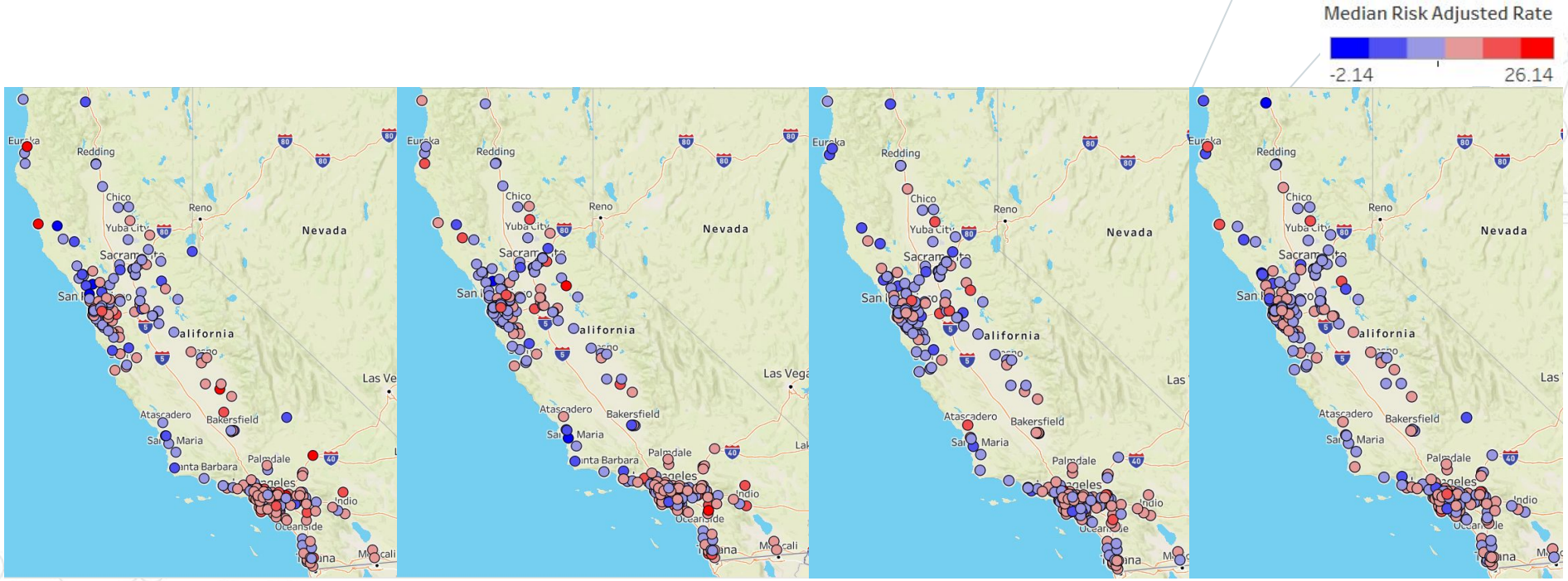
- 18 outliers above the median
- 18 outliers below the median
- Sonoma Valley Hospital in low outliers 3x
- Sonoma Valley Hospital is acute stroke ready certified by Center for Improvement in Healthcare Quality (as of 2019)
- Certification given to hospitals that meet high standards of care for treatment of stroke patients

EDA: Boxplot of Risk Adjusted Rates by Year



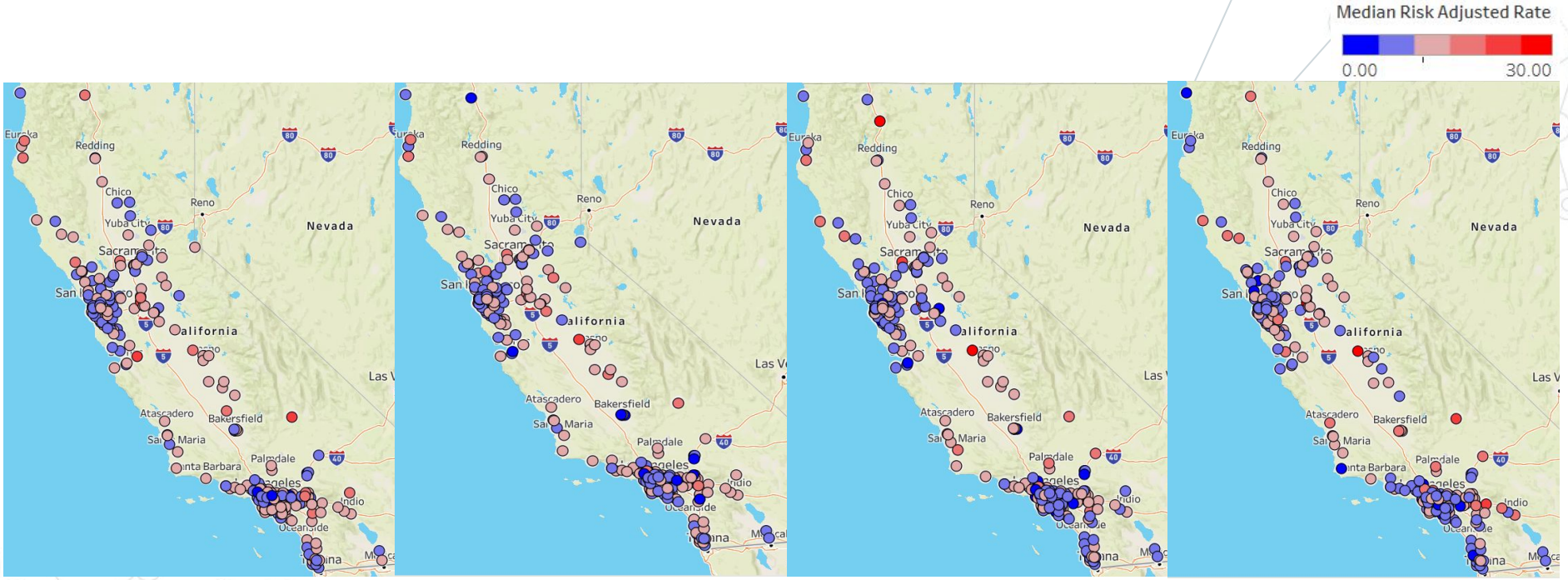
- 32 outliers above the median
- 16 outliers below the median
- LA County/USC Medical Center, Madera Community Hospital, Ridgecrest Regional Hospital in high outliers 3x
- San Joaquin General Hospital in high outliers 4x
- San Joaquin General Hospital is certified by The Joint Commission as dedicated Primary Stroke Center (goal to deliver definitive care to stroke patients within 60 minutes of arrival)
  - Perhaps something lacking in quality of care after initial treatment

EDA: Hospital Locations and Median Risk Adjusted Rates (Readmissions)



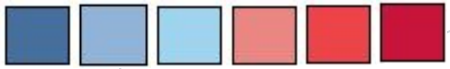
2011-2012 2012-2013 2013-2014 2014-2015

EDA: Hospital Locations and Median Risk Adjusted Rates (Mortality)



2011-2012 2012-2013 2013-2014 2014-2015

California County Health Factor Rankings

Good Rank  Bad Rank

2011

2012

2013

2014

4.

Model Selection

Which Models Can Be Used on the Data?

Dependent Variable

- Risk Adjusted Rate (numeric)

Independent Variables to Consider:

- Capacity (Number of Beds)
- Entity Type
- County Health Factors
- Median Household Income Estimate
- CHCI
- Population Estimate
- Poverty Rate
- Percent Age 65-84
- Factor(Hospital)
- Factor(County)
- Factor(Year)

Model Options

- Linear regression
- Machine learning
  - Regression trees
  - Random Forest
  - XGBoost

Linear Regression Model (Not Including Fixed Effect Variables)

Readmissions

```
Call:
lm(formula = `Risk Adjusted Rate` ~ `Median Household Income Estimate` +
    CHCI + `Population Estimate` + `Poverty Rate` + `Percent Age 65-84` +
    as.numeric(`County Rank Health Factors`), data = readmissions)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2507  -2.0529   0.0771   1.8777  12.9385

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            1.328e+01  1.593e+00   8.336 2.36e-16 ***
`Median Household Income Estimate`     1.963e-05  8.382e-06   2.342  0.01934 *
CHCI                                  -1.102e-02  4.179e-03  -2.638  0.00846 **
`Population Estimate`                 -1.257e-05  6.485e-06  -1.938  0.05285 .
`Poverty Rate`                        -2.597e-02  2.552e-02  -1.017  0.30919
`Percent Age 65-84`                    5.229e-02  4.364e-02   1.198  0.23112
as.numeric(`County Rank Health Factors`)  4.982e-02  6.670e-03   7.469 1.68e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.31 on 1061 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.05943,   Adjusted R-squared:  0.05411
F-statistic: 11.17 on 6 and 1061 DF,  p-value: 4.058e-12
```

- Adjusted R^2 = 0.05
- Median household income estimate and county health factors ranking positively correlated with risk adjusted rate
- CHCI negatively correlated with risk adjusted rate

Mortality

```
Call:
lm(formula = `Risk Adjusted Rate` ~ `Median Household Income Estimate` +
    CHCI + `Population Estimate` + `Poverty Rate` + `Percent Age 65-84` +
    as.numeric(`County Rank Health Factors`), data = mortality)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9087  -2.0017  -0.2236   1.7873  19.3754

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            9.257e+00  1.763e+00   5.252 1.81e-07 ***
`Median Household Income Estimate`    -2.999e-05  9.060e-06  -3.310 0.000964 ***
CHCI                                   8.812e-03  4.590e-03   1.920 0.055163 .
`Population Estimate`                  4.071e-06  7.052e-06   0.577 0.563862
`Poverty Rate`                        -3.495e-02  2.795e-02  -1.250 0.211445
`Percent Age 65-84`                   -1.171e-02  4.836e-02  -0.242 0.808655
as.numeric(`County Rank Health Factors`)  1.771e-02  7.214e-03   2.455 0.014263 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.59 on 1079 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.01819,   Adjusted R-squared:  0.01273
F-statistic: 3.332 on 6 and 1079 DF,  p-value: 0.002951
```

- Adjusted R^2 = 0.01
- County health factors ranking positively correlated with risk adjusted rate
- Median household income negatively correlated with risk adjusted rate

Linear Regression Model (+ factor(County))

Readmissions

```
lm(formula = `Risk Adjusted Rate` ~ `Median Household Income Estimate` +
    CHCI + `Population Estimate` + `Poverty Rate` + `Percent Age 65-84` +
    as.numeric(`County Rank Health Factors`) + factor(County),
    data = readmissions)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6485  -1.7819  -0.0323   1.6106  14.4464

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            1.127e+01  1.973e+00   5.712 1.47e-08 ***
`Median Household Income Estimate`     1.379e-05  8.402e-06   1.641 0.101118
CHCI                                  -7.137e-03  4.307e-03  -1.657 0.097830 .
`Population Estimate`                 -3.895e-06  6.737e-06  -0.578 0.563228
`Poverty Rate`                        -1.558e-02  2.571e-02  -0.606 0.544587
`Percent Age 65-84`                    2.249e-02  4.594e-02   0.490 0.624463
as.numeric(`County Rank Health Factors`) 1.391e-01  5.384e-02   2.583 0.009939 **
factor(County)Amador                  -6.393e-01  1.626e+00  -0.393 0.694288
factor(County)Butte                   -1.223e+00  1.279e+00  -0.957 0.338869
factor(County)Calaveras                2.653e+00  1.649e+00   1.609 0.108001
factor(County)Contra Costa             1.301e-01  7.376e-01   0.176 0.859988
factor(County)Del Norte               -5.440e+00  1.960e+00  -2.775 0.005619 **
factor(County)El Dorado               -2.017e+00  1.566e+00  -1.288 0.197901
factor(County)Fresno                  -5.058e+00  2.101e+00  -2.408 0.016234 *
factor(County)Humboldt                -1.450e+00  1.210e+00  -1.199 0.230923
factor(County)Imperial                -3.810e+00  2.291e+00  -1.663 0.096564 .
factor(County)Kern                    -5.466e+00  2.170e+00  -2.519 0.011938 *
factor(County)Kings                   -5.399e+00  2.364e+00  -2.284 0.022573 *
factor(County)Lake                    -1.006e+01  2.272e+00  -4.428 1.05e-05 ***
factor(County)Los Angeles             -1.388e+00  1.082e+00  -1.283 0.199774
factor(County)Madera                  -4.799e+00  2.325e+00  -2.065 0.039208 *
factor(County)Marin                   -5.555e-01  1.347e+00  -0.412 0.680223
factor(County)Mendocino               -3.976e+00  1.392e+00  -2.857 0.004367 **
factor(County)Merced                  -5.788e+00  2.413e+00  -2.399 0.016610 *
factor(County)Monterey                -3.050e+00  1.191e+00  -2.561 0.010577 *
factor(County)Napa                     6.137e-01  1.246e+00   0.493 0.622331
factor(County)Nevada                   8.361e-01  1.690e+00   0.495 0.620813
factor(County)Orange                   1.099e+00  7.436e-01   1.478 0.139735
factor(County)Placer                  -1.043e+00  1.310e+00  -0.796 0.426034
factor(County)Riverside               -1.469e+00  1.331e+00  -1.103 0.270197
factor(County)Sacramento              -2.548e+00  1.016e+00  -2.508 0.012289 *
factor(County)San Benito               9.859e-02  1.670e+00   0.059 0.952922
factor(County)San Bernardino          -2.857e+00  1.694e+00  -1.687 0.091995 .
factor(County)San Diego                3.846e-02  6.204e-01   0.062 0.950575
factor(County)San Francisco            7.541e-01  9.100e-01   0.829 0.407476
factor(County)San Joaquin             -4.280e+00  1.859e+00  -2.303 0.021502 *
factor(County)San Luis Obispo         -5.805e-01  1.095e+00  -0.530 0.596149
factor(County)San Mateo                2.455e-01  1.092e+00   0.225 0.822158
factor(County)Santa Barbara           -2.717e+00  1.018e+00  -2.669 0.007729 **
factor(County)Santa Clara              9.762e-01  1.011e+00   0.965 0.334656
factor(County)Santa Cruz              -9.541e-01  1.248e+00  -0.765 0.444607
factor(County)Shasta                  -3.914e+00  1.672e+00  -2.341 0.019449 *
factor(County)Siskiyou                -9.474e+00  1.903e+00  -4.979 7.51e-07 ***
factor(County)Solano                  -3.770e+00  1.054e+00  -3.577 0.000364 ***
factor(County)Sonoma                  -3.490e+00  7.703e-01  -4.530 6.58e-06 ***
factor(County)Stanislaus              -5.360e+00  1.838e+00  -2.916 0.003626 **
factor(County)Tehama                  -4.432e+00  2.212e+00  -2.004 0.045343 *
factor(County)Tulare                  -3.498e+00  2.335e+00  -1.498 0.134459
factor(County)Tuolumne                -3.001e+00  1.658e+00  -1.810 0.070545 .
factor(County)Ventura                 -4.425e-01  7.643e-01  -0.579 0.562741
factor(County)Yolo                    -3.928e+00  1.220e+00  -3.219 0.001329 **
factor(County)Yuba                    -6.181e+00  2.523e+00  -2.450 0.014455 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.095 on 1016 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.2127,    Adjusted R-squared:  0.1732
F-statistic: 5.382 on 51 and 1016 DF,  p-value: < 2.2e-16
```

- Adjusted R^2 = 0.17
- County health factors ranking positively correlated with risk adjusted rate
- 20 significant counties
- Out of the significant counties, all of them are negatively correlated with risk adjusted rate
- Out of the significant counties, 13/20 are in Northern CA; 6/20 are in Central CA, and 1/20 is in Southern CA

Linear Regression Model (+ factor(County))

Mortality

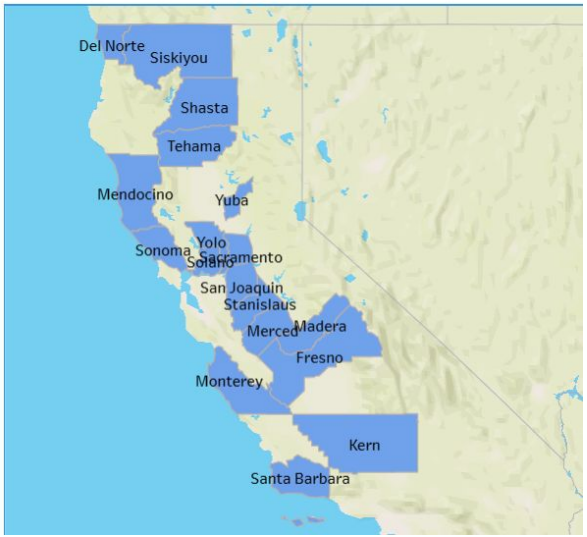
```
Call:
lm(formula = `Risk Adjusted Rate` ~ `Median Household Income Estimate` +
    CHCI + `Population Estimate` + `Poverty Rate` + `Percent Age 65-84` +
    as.numeric(`County Rank Health Factors`) + factor(County),
    data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-12.405  -1.776  -0.014   1.681  18.548

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.098e+01  2.084e+00   5.270 1.66e-07 ***
`Median Household Income Estimate` -2.908e-05  8.822e-06  -3.296 0.001012 **
CHCI                                5.173e-04  4.570e-03   0.113 0.909885
`Population Estimate`              -6.167e-06  6.898e-06  -0.894 0.371508
`Poverty Rate`                     -7.143e-02  2.686e-02  -2.659 0.007961 **
`Percent Age 65-84`                 7.200e-02  4.881e-02   1.475 0.140510
as.numeric(`County Rank Health Factors`)  8.942e-02  5.500e-02   1.626 0.104314
factor(County)Amador                9.505e-01  1.689e+00   0.563 0.573722
factor(County)Butte                -3.304e+00  1.316e+00  -2.510 0.012234 *
factor(County)Calaveras             2.585e+00  1.700e+00   1.521 0.128639
factor(County)Contra Costa         -3.708e-01  7.556e-01  -0.491 0.623674
factor(County)Del Norte            -4.762e+00  2.000e+00  -2.381 0.017454 *
factor(County)El Dorado             1.790e+00  1.508e+00   1.187 0.235528
factor(County)Fresno               -1.916e+00  2.148e+00  -0.892 0.372694
factor(County)Humboldt              2.534e+00  1.212e+00   2.090 0.036831 *
factor(County)Imperial             -5.518e+00  2.343e+00  -2.354 0.018738 *
factor(County)Kern                 -9.929e-01  2.214e+00  -0.448 0.653966
factor(County)Kings                -6.015e-01  2.418e+00  -0.249 0.803637
factor(County)Lake                 -5.557e-01  2.328e+00  -0.239 0.811398
factor(County)Los Angeles          -3.170e+00  1.103e+00  -2.873 0.004151 **
factor(County)Madera                1.102e+01  2.392e+00   4.606 4.61e-06 ***
factor(County)Marin                -9.786e-02  1.386e+00  -0.071 0.943707
factor(County)Mendocino             6.980e-01  1.390e+00   0.502 0.615703
factor(County)Merced               -3.702e+00  2.475e+00  -1.496 0.135019
factor(County)Monterey             -3.282e+00  1.222e+00  -2.686 0.007351 **
factor(County)Napa                  2.204e+00  1.282e+00   1.720 0.085773 .
factor(County)Nevada                2.342e+00  1.742e+00   1.345 0.179009
factor(County)Orange                2.428e-02  7.634e-01   0.032 0.974630
factor(County)Placer                1.853e+00  1.347e+00   1.376 0.169098
factor(County)Riverside            -3.778e-01  1.357e+00  -0.278 0.780736
factor(County)Sacramento           -1.106e+00  1.037e+00  -1.067 0.286435
factor(County)San Benito            5.056e+00  1.717e+00   2.945 0.003300 **
factor(County)San Bernardino       -3.035e+00  1.733e+00  -1.751 0.080168 .
factor(County)San Diego            -7.676e-01  6.387e-01  -1.202 0.229692
factor(County)San Francisco         7.732e-01  9.352e-01   0.827 0.408557
factor(County)San Joaquin           5.624e-01  1.900e+00   0.296 0.767341
factor(County)San Luis Obispo       3.985e+00  1.124e+00   3.546 0.000409 ***
factor(County)San Mateo             2.235e-01  1.124e+00   0.199 0.842355
factor(County)Santa Barbara         1.768e+00  1.054e+00   1.678 0.093627 .
factor(County)Santa Clara           9.239e-01  1.042e+00   0.887 0.375321
factor(County)Santa Cruz            1.662e+00  1.287e+00   1.291 0.196865
factor(County)Shasta               -3.098e-01  1.713e+00  -0.181 0.856547
factor(County)Siskiyou              3.871e+00  1.830e+00   2.115 0.034661 *
factor(County)Solano               -3.552e+00  1.072e+00  -3.313 0.000954 ***
factor(County)Sonoma               -1.562e-02  7.943e-01  -0.020 0.984318
factor(County)Stanislaus           -2.296e+00  1.880e+00  -1.222 0.222124
factor(County)Tehama               -1.673e+00  2.267e+00  -0.738 0.460685
factor(County)Tulare                1.718e-01  2.388e+00   0.072 0.942687
factor(County)Tuolumne              2.389e+00  1.705e+00   1.401 0.161490
factor(County)Ventura               3.934e-01  7.867e-01   0.500 0.617143
factor(County)Yolo                  4.098e+00  1.258e+00   3.259 0.001156 **
factor(County)Yuba                 -1.648e+00  2.587e+00  -0.637 0.524303
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.187 on 1034 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.2585,    Adjusted R-squared:  0.2219
F-statistic: 7.069 on 51 and 1034 DF,  p-value: < 2.2e-16
```

Adjusted R^2 = 0.22
Median household income estimate and poverty rate negatively correlated with risk adjusted rate
12 significant counties
Out of the significant counties, 6 positively correlated with risk adjusted rate and 6 negatively correlated with risk adjusted rate

- +: 4/6 in Northern CA; 2/6 in Central CA
- -: 4/6 in Northern CA; 2/6 in Southern CA

Linear Regression Model (+ factor(County)) - County Visualization

Significant CA Counties After Running Linear Regression Model and Adding County Fixed Effect (Readmissions)



Significant CA Counties After Running Linear Regression Model and Adding County Fixed Effect (Mortality)

- Readmissions:
  - All significant counties spread across far northern to central California
  - Visualization similar to EDA visualization of median risk adjusted rates by hospital
  - Possible variable (unknown) in northern CA counties may explain correlation with lower risk adjusted rates
- Mortality:
  - No discernable pattern

Key: **Blue:** Negatively correlated with risk adjusted rate
Red: Positively correlated with risk adjusted rate

Linear Regression Model (+ factor(County) + factor(Hospital))

Readmissions

```
Call:
lm(formula = `Risk Adjusted Rate` ~ `Median Household Income Estimate` +
    CHCI + `Population Estimate` + `Poverty Rate` + `Percent Age 65-84` +
    as.numeric(`County Rank Health Factors`) + factor(Hospital) +
    factor(County), data = readmissions)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8338 -1.1038 -0.0094  1.1174  9.0231

Residual standard error: 2.395 on 779 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.6384,    Adjusted R-squared:  0.5047
F-statistic: 4.775 on 288 and 779 DF,  p-value: < 2.2e-16
```

- Adjusted R^2 = 0.50
- County health factors ranking positively correlated with risk adjusted rate
- No significant counties

Mortality

```
Call:
lm(formula = `Risk Adjusted Rate` ~ `Median Household Income Estimate` +
    CHCI + `Population Estimate` + `Poverty Rate` + `Percent Age 65-84` +
    as.numeric(`County Rank Health Factors`) + factor(Hospital) +
    factor(County), data = mortality)

Residuals:
   Min     1Q Median     3Q    Max
-7.774 -1.034  0.000  1.062  8.837

Residual standard error: 2.262 on 792 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.714,     Adjusted R-squared:  0.6082
F-statistic: 6.747 on 293 and 792 DF,  p-value: < 2.2e-16
```

- Adjusted R^2 = 0.61
- County health factors ranking positively correlated with risk adjusted rate
- Median household income estimate and poverty rate negatively correlated with risk adjusted rate
- No significant counties

Best/Worst Significant Hospitals from Full Linear Regression Model (Readmissions)

Best: Frank R. Howard Memorial Hospital

County: Mendocino (Northern CA)

Coefficient: -4.86

P-value: 7.79 x 10^-3

<u>Potential reasons for rating:</u>

- Community education efforts - increase proportion of adults ages 20+ who are aware of the symptoms or how to prevent stroke

Worst: Greater El Monte Community Hospital

County: Los Angeles (Southern CA)

Coefficient: 14.14

P-value: 1.84 x 10 x 10^-10

<u>Potential reasons for rating:</u>

- No rehabilitation services for stroke patients
- Low nurse staffing

Best/Worst Significant Hospitals from Full Linear Regression Model (Mortality)

Best: Montclair Hospital Medical Center

County: San Bernardino (Southern CA)

Coefficient: -13.54

P-value: 2.00 x 10^-7

<u>Potential reasons for rating:</u>

- Rehabilitation services - speech/language pathologists

Worst: Madera Community Hospital (closed as of January 2023)

County: Madera (Central CA)

Coefficient: 10.95

P-value: 6.77 x 10^-11

<u>Potential reasons for rating:</u>

- Rural area - not a lot of access to resources
- Only general hospital in area - overloaded

Machine Learning Models (Not Including Fixed Effect Variables)

Models Used:

- Linear Regression
- Tree Regression
- Random Forest Regression
- XGBoost Regression
- 2011-2013 data used as train set
- 2014 data used as test set

- Cross Validation Score: Higher = better model
- MSE and RMSE: Lower = better model

Readmissions Summary Statistics

Model	CV	MSE	RMSE
Linear	9.642	9.815	3.133
Tree	9.550	9.762	3.124
RF	9.698	9.234	3.039
XGB	11.911	10.160	3.187

- XGBoost model has highest CV score
- Random forest model has lowest MSE and RMSE

Mortality Summary Statistics

Model	CV	MSE	RMSE
Linear	14.716	14.092	3.754
Tree	14.284	14.215	3.770
RF	14.190	12.572	3.546
XGB	17.749	11.972	3.460

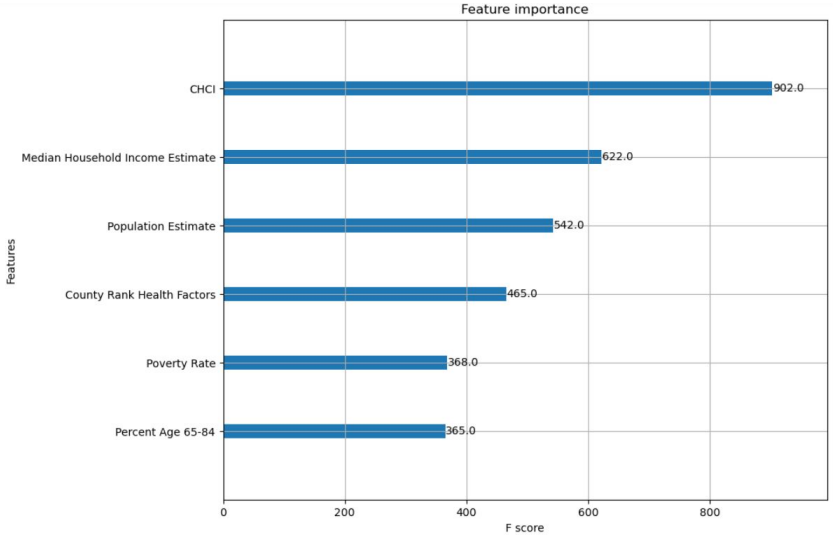
- XGBoost model has highest CV score and lowest MSE and RMSE

Key Takeaway:

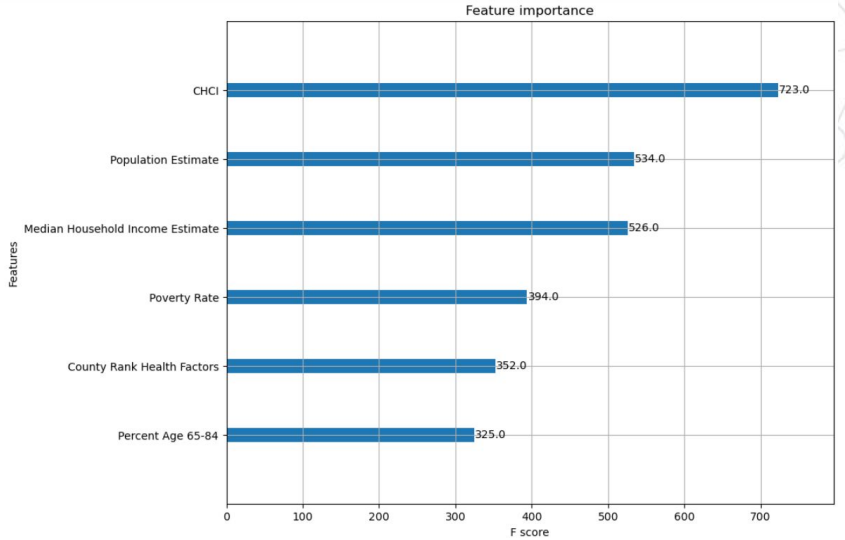
- Explore XGBoost model further as it may be better fit for data rather than a linear model

Feature Importance Charts From XGBoost Regression Models

Readmissions



Mortality



- CHCI is most important feature for both data frames when fixed effect variables are not included in the model
- Note that for linear regression model without fixed effect variables, CHCI, median household income estimate, and county health factors ranking were only significant independent variables across both data frames

Machine Learning Models (Including Fixed Effect Variables)

Models Used:

- Linear Regression
- Tree Regression
- Random Forest Regression
- XGBoost Regression
- 2011-2013 data used as train set
- 2014 data used as test set

- Cross Validation Score: Higher = better model
- MSE and RMSE: Lower = better model

Readmissions Summary Statistics

Model	CV	MSE	RMSE
Linear	11.138	10.118	3.181
Tree	9.550	9.762	3.124
RF	9.254	9.114	3.019
XGB	9.842	8.839	2.973

- Linear model has highest CV score
- XGBoost model has lowest MSE and RMSE

Mortality Summary Statistics

Model	CV	MSE	RMSE
Linear	13.074	9.044	3.007
Tree	14.284	14.215	3.770
RF	13.102	11.072	3.327
XGB	14.856	9.217	3.036

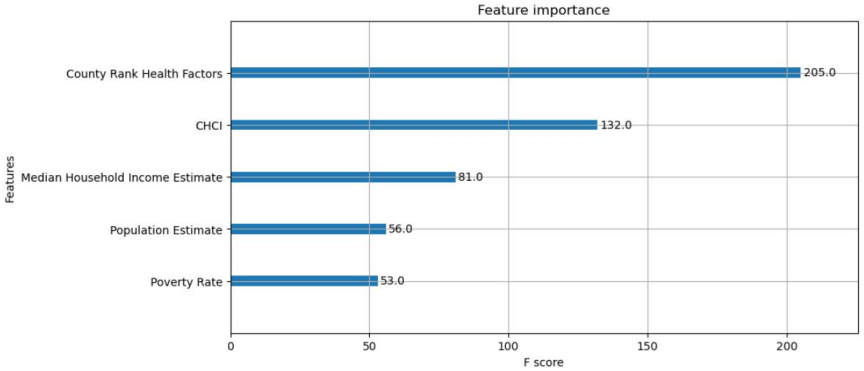
- XGBoost model has highest CV score
- Linear model has lowest MSE and RMSE

Key Takeaway:

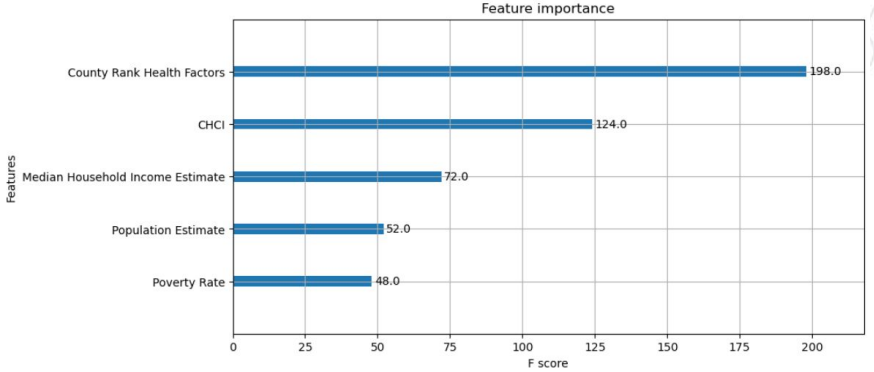
- Explore XGBoost model further but between linear and XGBoost, prefer linear because of simplicity and easier interpretability

Feature Importance Charts From XGBoost Regression Models

Readmissions



Mortality



- County health factors ranking is most important feature for both data frames when fixed effect variables are included in the model
- Note that for linear regression model with fixed effect variables, county health factors ranking, median household income estimate, and poverty rate were only significant independent variables across both data frames

5.

Key Takeaways

Key Takeaways

- County health factors ranking, CHCI, median household income most influential on risk adjusted rate (see this in both linear regression and XGBoost regression models) → improving counties on these measures may help to lower risk adjusted rates

- There does appear to be a difference between the readmissions and mortality data as median risk adjusted rates, significant independent variables, and summary statistics (e.g. adjusted R^2, CV score, MSE, RMSE) differ between the two for same model
  - Potential reasons:
    - Small sample size → may be biasing results
    - Maybe some other variable not included in analysis can account for this difference

- Although independent variable relationships to risk adjusted rate can be explained with linear regression model, there is evidence for potential non-linear relationships between the variables due to good performance of XGBoost regression model as well

Thanks!

Any questions?