Project 5: Predicting Covid Deaths in Nursing Homes

Shirley Lin

Jamie Squires

Sam Waldner

Background:
Nursing Home Covid Outbreaks

- The communal nature of nursing homes puts elderly and vulnerable residents at an increased risk of infection and severe illness from covid

- In the US, long-term care facilities accounted for almost **40%** of covid deaths in 2020 [2]

1: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-in-nursing-homes.html
2: https://www.nytimes.com/2020/12/31/opinion/sunday/covid-nursing-homes.html

Background:
Medicare & Medicaid Data

US Centers for Medicare & Medicaid house an extensive database on various attributes for nursing homes, including:

- Size, location, financial data of facilities
- Scores for health inspections, staffing, quality
- Number of complaints, citations, fines

- Covid vaccinations, cases, deaths

Could any of these attributes be predictors of covid deaths in nursing homes?

Problem Statement

Government oversight is comprehensive in how they score and evaluate nursing homes. However, they have yet to incorporate any COVID related data into these scores. This project will aim to predict COVID deaths in nursing homes based on a wide variety of features including the star ratings and their component parts.

We hypothesize that because these factors are indicative of the overall safety, sanitation, and operational efficiency of a nursing home, they should help represent and understand COVID related incidents since the virus is spread through known health factors. Any information relating to COVID is an important decision-making tool for people looking to find a safe nursing home for their loved ones, and so based on our results we will make recommendations on the nursing home selection process.

1. **Can we create a reliable model to predict Covid-19 deaths amongst nursing homes?**

2. **Which features are relatively most important in determining covid deaths?**
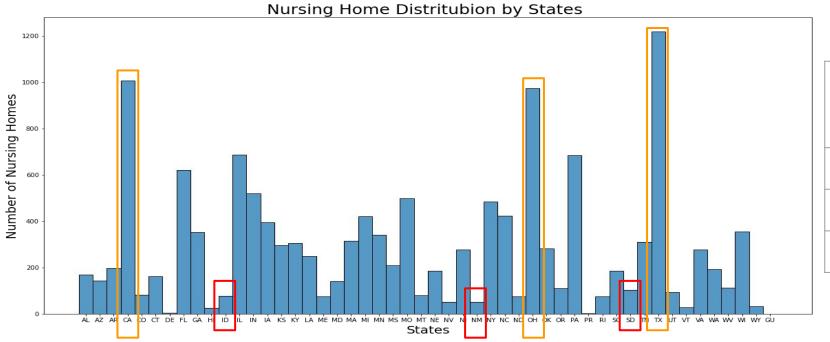
Action Plan

- We will perform a **variety of regression & modeling techniques** to predict covid deaths per 1,000 nursing home residents from 2020-2021.

Success Metrics:
  - Optimize for the highest RMSE value
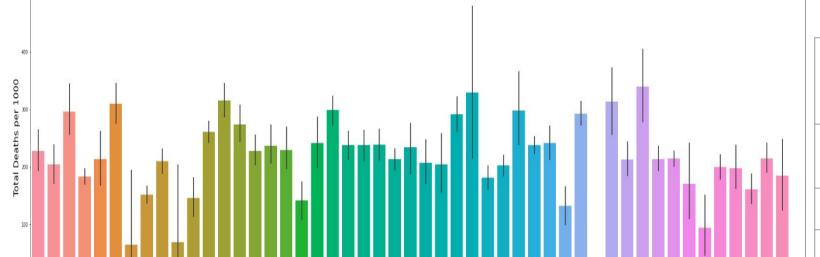 - How well does our model generalize to new data?

Methodology

1. Explore the CMS database, gain understanding of data structure
2. Merge different data sources, clean data
3. Generate visualizations & perform EDA
4. Perform various modeling techniques
5. Provide recommendations to prospective residents on nursing home selection

Nursing Home Distribution by States

Top Three States with the most nursing homes:

State	Nursing Homes	Mean Deaths per 1000
Texas	1218	213
California	1007	183
Ohio	974	240
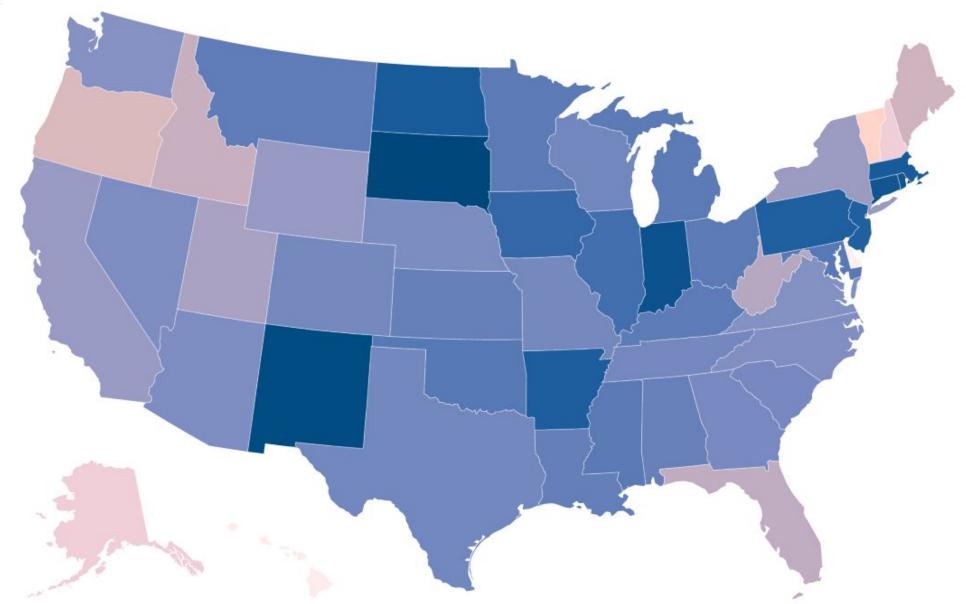
States Vs. Total Deaths From 2020-2021

Top Three States with the highest mean deaths per 1000:
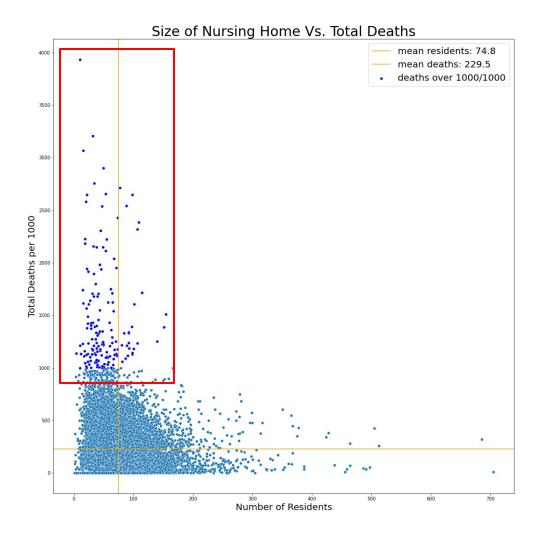
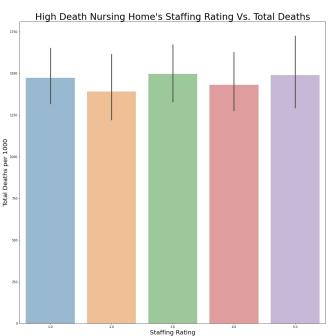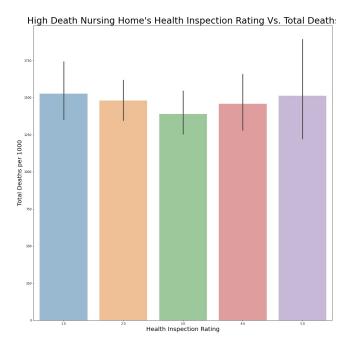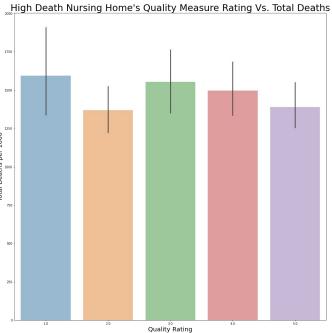| State	Nursing Homes	Mean Deaths per 1000
South Dakota	104	342
New Mexico	51	329
Indiana	519	316

Nursing Home Total Deaths
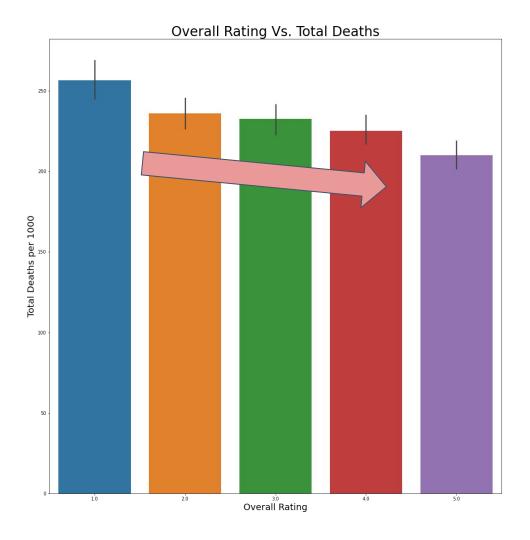
64.88　　　　340.16
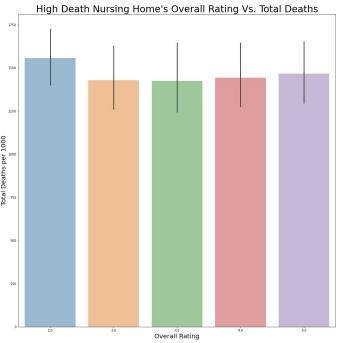
EDA - distribution of deaths

EDA: Different Ratings Vs. Total Deaths

Modeling & Evaluation
Model Features & Variables

Features:
- General Attributes:
  - *Rural vs urban*
  - *Size*
  - *Financials*
- 5 Star Ratings:
  - *Health Inspection Ratings*
  - *Overall Rating*
  - *Staffing Ratings*
- Violations:
  - *Fines, Citations, Penalties, Abuse history*
- Staff and resident vaccination rates

Target:
Total covid deaths per 1,000 residents (2020-2021)
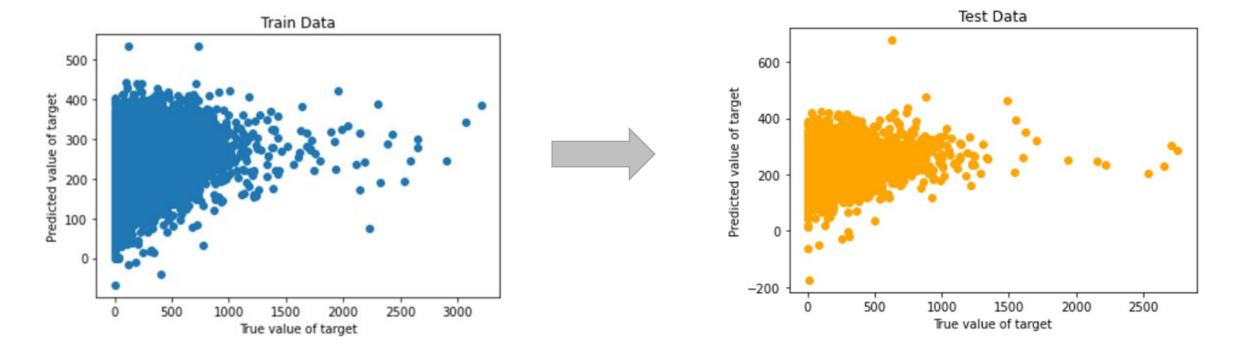
Modeling & Evaluation
Model Selection

Model	RMSE	MAE
Linear Regression	237.83	165.90
XGBoost	238.24	165.70
Random Forests Regressor	240.17	169.15
Bagging Regressor	240.66	169.47
KNN Regressor	241.54	168.20
PCA	244.06	172.07
Decision Tree Regressor	245.58	172.42
Neural Network	248.98	164.11
AdaBoost Regressor	419.87	362.00

- **Baseline RMSE**: 242.81

- 5 models performed *slightly* better than baseline

Modeling & Evaluation
Best Model: Linear Regression

Model	RMSE	MAE	R2
Training	233.62	162.73	6.6%
Testing	237.83	165.90	6.5%

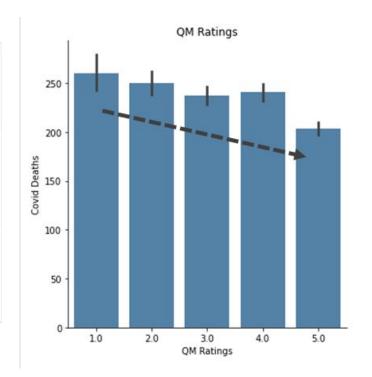**Target:** Total Covid Deaths per 1,000 Residents (2020-2021)

Modeling & Evaluation
Inference

Which rating system had the strongest coefficients in our regression model?

Ratings with the highest negative coefficients

1. QM Rating
2. RN Staffing Rating
3. Overall Rating
4. Health Inspection Rating

Insights:

- Compared to the other scores, the QM Rating had the highest relative negative correlation with covid deaths *(the higher the score, the lower the deaths)*
- Recommended to use **QM Rating** to select a nursing home if no other information given
- However, given the low predictive performance of our model, these scores clearly do not explain the whole picture of a nursing home's ability to protect residents from Covid

Conclusion

Results:

- Government Health, Safety, and Wellness Measures
- COVID
- Predictability

Recommendations/Actionable Takeaways:

- Update Score Metrics
- National Reporting Oversight
- New Tools!

Thank you!

Appendix

Sources

Image on Page 1:
- https://data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAANsAAADmCAMAAABruQABAAAAkFBMV

Image on Page 2:
- https://www.nytimes.com/2020/12/31/opinion/sunday/covid-nursing-homes.html

Image on Page 3:
- https://435x081fqey8xkeiq1hyznh5-wpengine.netdna-ssl.com/wp-content/uploads/2017/05/5-star-edit-300x207.jpg

image on Page 8:
- https://app.datawrapper.de/map/Fj5DY/visualize#refine

Covid 19 Data:
- https://data.cms.gov/covid-19/covid-19-nursing-home-data

Vaccine Data:
- https://data.cms.gov/provider-data/dataset/avax-cv19

Nursing Home Provider Data:
- https://data.cms.gov/provider-data/dataset/4pq5-n9py

Financial Data:
- https://data.cms.gov/provider-compliance/cost-report/skilled-nursing-facility-cost-report

Scoring Information:
- https://www.cms.gov/Medicare/Provider-Enrollment-and-Certification/CertificationandComplianc/FSQRS