Individual Portion-JMP Analysis of COVID data by Kasey Moran: Link to file (3.34GB)

Group Portion-Business Summary of COVID analysis:

Executive Summary

This study aims to evaluate data from the Centers for Disease Control and Prevention (CDC) to determine which variables are the most significant contributors to mortality in Coronavirus disease 2019 (COVID-19) cases. This analysis provides a firsthand assessment of real-world COVID-19 death data. Our goal is to produce a predictive model that would help individuals evaluate the likelihood that a particular patient would not survive their case of COVID-19.

Our initial approach was to look for a comprehensive data set that related demographics, social situations, and preexisting medical conditions to factors such as hospitalization and treatment path for COVID-19. However, as we looked deeper into the CDC data sets, we began to experience challenges. They were disaggregated, troublesome for this type of analysis, and contained missing data. Once the required cleanup was completed, the team opted to use one set of data that could relate multiple factors to the outcome of death from COVID-19 cases (COVID-19_Case_Surveillance_Public_Use_Data_with_Geography_less_CLMS).

With the final dataset determined, the data distributions were analyzed to look for missing data and understand the first-order trends. The initial distribution showed that age was a significant factor in the simple aggregate tabulate numbers. This led to the need for logistic regression to further understand the significance of this contributor. The independent variables we evaluated were age group, sex, and race/ethnicity, with death as the binary (yes or no) dependent variable.

The analysis took two forms: logistic regression and decision tree. The logistic regression required the evaluation of each independent variable for usable contributing value. In the case of preexisting conditions, we determined that the missing data fraction was too large and could not be used as a trustworthy variable. After reviewing the data, we selected the following three independent variables in our analysis: age group, race/ethnicity, and sex. We chose to evaluate the aggregated data set for the years 2020 and 2021 to capture the most extensive viewpoint possible.

The results of our analysis provide compelling evidence that age group is the highest contributing factor to death. For example, we found that a COVID-19 patient over the age of 65 is almost 43,000 times more likely to die than someone under 17 years of age. Our Logworth analysis ranked our chosen variables with age group being the highest, followed by race/ethnicity, and lastly, sex. The logic tree analysis also confirmed that age is the highest contributing factor. Our logistic regression parameter estimates are all significant to p-value <0.0001. The parameter estimates showed that the group aged 50-64 has a positive value of (+1.98), while the lower age groups have negative values (-0.84, -5.90). The positive value in the older group indicates an increased contribution to an outcome of death by COVID-19, a result we expected based on our knowledge of the pandemic.

The data exploration and analysis process gave us key insights as we reviewed such a large data set. First, every data set requires careful evaluation to determine if it will adequately provide the information needed to support or not support an idea. While we were able to find evidence to support our question, we also learned that the data categories could have been segmented even further to narrow the scope for variables like age group.

Our conclusion based on analysis is that age is the highest contributing factor in the outcome of death for those infected with COVID-19. This is a probabilistic analysis, and there are many factors that we could not include in the scope of this analysis. Nevertheless, our analysis supports the general concern of the CDC that vaccines should be allocated to the elderly first.

Background

As COVID-19 spread across the globe in 2020, it became apparent that data could play a pivotal role in understanding who was vulnerable to the worst of its symptoms. As a result, the CDC led a massive effort to collect information on individual cases in the United States for analysis and research. This effort is ongoing as public leaders and healthcare providers evaluate what public health in the US should be after the pandemic. Utilizing the data, they have provided to the public, we attempt to find what demographic factors or variables are most significant in predicting death in COVID-19 cases. Our analysis aims to provide a predictive model that will calculate the probability of a person's likelihood of death upon contracting COVID-19 based on age group, gender, and race/ethnicity.

Data Set Used

We selected an extensive, public data set from the CDC containing 69.7 million rows.[1] [2] Each row represented a COVID-19 case in the United States that became known from January 2020 to March 2022. Following our research question, we selected the variables sex, race,

[1] Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Public Use Data with Geography (version date: April 04, 2022).

[2] The final data set we used for our analysis can be found at this link: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4.

ethnicity, and age to compare against death. From this dataset, we excluded and removed any rows where sex, age, death, race, or ethnicity values were entered as “unknown” or “missing.” For example, if ethnicity was labeled as NA, it was categorized as non-Hispanic. The ethnicity column and race column were then combined into one column. For the new race/ethnicity column, any label that was “race,” Hispanic was recoded into Hispanic/Latino. For the remaining non-Hispanic/Latino combinations, the label was recoded to “race,” NH[3].

Analysis

We began our analysis by performing a logistic regression of sex, age, and race/ethnicity against death. Using the effect summary for the LogWorth of each variable to indicate individual significance to the outcome, we then reviewed the parameter estimates and odds ratios. The results below indicate a protective factor for age groups 0-17, 18-49, females, and Multiple/other, NH and Native Hawaiian/Other Pacific Islander, NH.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-6.40956	0.0822745	6069.1	<.0001*
age_group[0 - 17 years]	-5.9032018	0.2080431	805.14	<.0001*
age_group[18 to 49 years]	-0.8447427	0.0700159	145.56	<.0001*
age_group[50 to 64 years]	1.98278657	0.0695384	813.02	<.0001*
sex[Female]	-0.2259042	0.0018766	14491	<.0001*
Race/Ethnicity 2[American Indian/Alaska Native,NH]	0.20386851	0.0535029	14.52	0.0001*
Race/Ethnicity 2[Asian,NH]	1.17787058	0.0453027	676.00	<.0001*
Race/Ethnicity 2[Black,NH]	0.27030136	0.0444844	36.92	<.0001*
Race/Ethnicity 2[Hispanic/Latino]	1.14434528	0.0444817	661.84	<.0001*
Race/Ethnicity 2[Multiple/Other,NH]	-0.8096472	0.0514842	247.31	<.0001*
Race/Ethnicity 2[Native Hawaiian/Other Pacific Islander,NH]	-1.749942	0.2621304	44.57	<.0001*

For log odds of Yes/No

[3] NH = Non-Hispanic/Latino

Odds Ratios for age_group

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
18 to 49 years	0 - 17 years	157.34788	<.0001*	91.312265	271.13941
50 to 64 years	0 - 17 years	2659.7525	<.0001*	1544.2313	4581.1034
65+ years	0 - 17 years	42974.391	<.0001*	24952.793	74011.686
50 to 64 years	18 to 49 years	16.903644	<.0001*	16.447242	17.372712
65+ years	18 to 49 years	273.11707	<.0001*	266.22633	280.18616
0 - 17 years	18 to 49 years	0.0063553	<.0001*	0.0036881	0.0109514
65+ years	50 to 64 years	16.157289	<.0001*	15.96908	16.347717
0 - 17 years	50 to 64 years	0.000376	<.0001*	0.0002183	0.0006476
18 to 49 years	50 to 64 years	0.0591588	<.0001*	0.0575615	0.0608005
0 - 17 years	65+ years	2.327e-5	<.0001*	1.3511e-5	4.0076e-5
18 to 49 years	65+ years	0.0036614	<.0001*	0.0035691	0.0037562
50 to 64 years	65+ years	0.0618916	<.0001*	0.0611706	0.062621

Odds Ratios for sex

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
Male	Female	1.571151	<.0001*	1.5596355	1.5827515
Female	Male	0.6364761	<.0001*	0.6318111	0.6411754

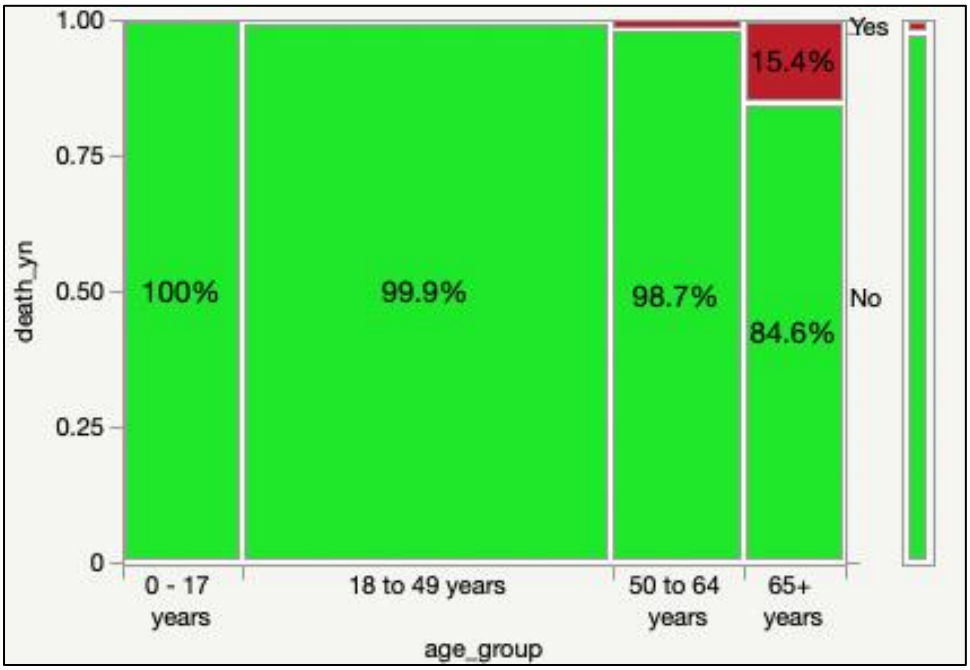
Odds Ratios for Race/Ethnicity

Level1	/Level2	Odds Ratio	Prob>Chisq	Lower 95%	Upper 95%
White,NH	Native Hawaiian/Other Pacific Islander,NH	4.5409921	<.0001*	2.4941562	8.2675695
White,NH	Multiple/Other,NH	1.7733149	<.0001*	1.6679795	1.8853025
White,NH	American Indian/Alaska Native,NH	0.6436083	<.0001*	0.6001346	0.6902311
White,NH	Black,NH	0.6022408	<.0001*	0.5952641	0.6092994
White,NH	Hispanic/Latino	0.2512915	<.0001*	0.248363	0.2542545
White,NH	Asian,NH	0.2430065	<.0001*	0.2374536	0.2486892
Native Hawaiian/Other Pacific Islander,NH	Multiple/Other,NH	0.3905127	0.0022*	0.2138297	0.7131851
Native Hawaiian/Other Pacific Islander,NH	White,NH	0.2202162	<.0001*	0.1209545	0.4009372
Native Hawaiian/Other Pacific Islander,NH	American Indian/Alaska Native,NH	0.141733	<.0001*	0.0775341	0.2590891
Native Hawaiian/Other Pacific Islander,NH	Black,NH	0.1326232	<.0001*	0.0728378	0.2414804
Native Hawaiian/Other Pacific Islander,NH	Hispanic/Latino	0.0553384	<.0001*	0.0303925	0.10076
Native Hawaiian/Other Pacific Islander,NH	Asian,NH	0.053514	<.0001*	0.0293808	0.0974701
Multiple/Other,NH	Native Hawaiian/Other Pacific Islander,NH	2.5607364	0.0022*	1.4021606	4.6766189
Multiple/Other,NH	White,NH	0.5639156	<.0001*	0.5304189	0.5995278
Multiple/Other,NH	American Indian/Alaska Native,NH	0.3629408	<.0001*	0.3307958	0.3982094
Multiple/Other,NH	Black,NH	0.339613	<.0001*	0.3191875	0.3613456
Multiple/Other,NH	Hispanic/Latino	0.1417072	<.0001*	0.1331853	0.1507744
Multiple/Other,NH	Asian,NH	0.1370352	<.0001*	0.1283933	0.1462586
Hispanic/Latino	Native Hawaiian/Other Pacific Islander,NH	18.070618	<.0001*	9.9245763	32.902891
Hispanic/Latino	Multiple/Other,NH	7.0568054	<.0001*	6.632427	7.5083377
Hispanic/Latino	White,NH	3.9794428	<.0001*	3.9330672	4.0263652
Hispanic/Latino	American Indian/Alaska Native,NH	2.5612023	<.0001*	2.3866023	2.7485756
Hispanic/Latino	Black,NH	2.3965829	<.0001*	2.3602571	2.4334677
Hispanic/Latino	Asian,NH	0.9670304	0.0088*	0.9430939	0.9915745
Black,NH	Native Hawaiian/Other Pacific Islander,NH	7.5401599	<.0001*	4.1411231	13.729129
Black,NH	Multiple/Other,NH	2.944528	<.0001*	2.7674335	3.1329552
Black,NH	White,NH	1.6604653	<.0001*	1.6412294	1.6799267
Black,NH	American Indian/Alaska Native,NH	1.0686892	0.0652	0.9958254	1.1468844
Black,NH	Hispanic/Latino	0.4172608	<.0001*	0.4109362	0.4236827
Black,NH	Asian,NH	0.4035039	<.0001*	0.3934995	0.4137625
Asian,NH	Native Hawaiian/Other Pacific Islander,NH	18.686711	<.0001*	10.259559	34.035884
Asian,NH	Multiple/Other,NH	7.2973973	<.0001*	6.8372027	7.7885664
Asian,NH	White,NH	4.1151163	<.0001*	4.0210827	4.211349
Asian,NH	American Indian/Alaska Native,NH	2.6485229	<.0001*	2.4611787	2.8501276
Asian,NH	Black,NH	2.478291	<.0001*	2.4168453	2.541299
Asian,NH	Hispanic/Latino	1.0340936	0.0088*	1.0084971	1.0603398
American Indian/Alaska Native,NH	Native Hawaiian/Other Pacific Islander,NH	7.0555218	<.0001*	3.8596756	12.897557
American Indian/Alaska Native,NH	Multiple/Other,NH	2.7552706	<.0001*	2.5112414	3.0230134
American Indian/Alaska Native,NH	White,NH	1.5537402	<.0001*	1.44879	1.6662928
American Indian/Alaska Native,NH	Black,NH	0.9357257	0.0652	0.8719275	1.0041921
American Indian/Alaska Native,NH	Hispanic/Latino	0.3904416	<.0001*	0.3638248	0.4190057
American Indian/Alaska Native,NH	Asian,NH	0.377569	<.0001*	0.3508615	0.4063094

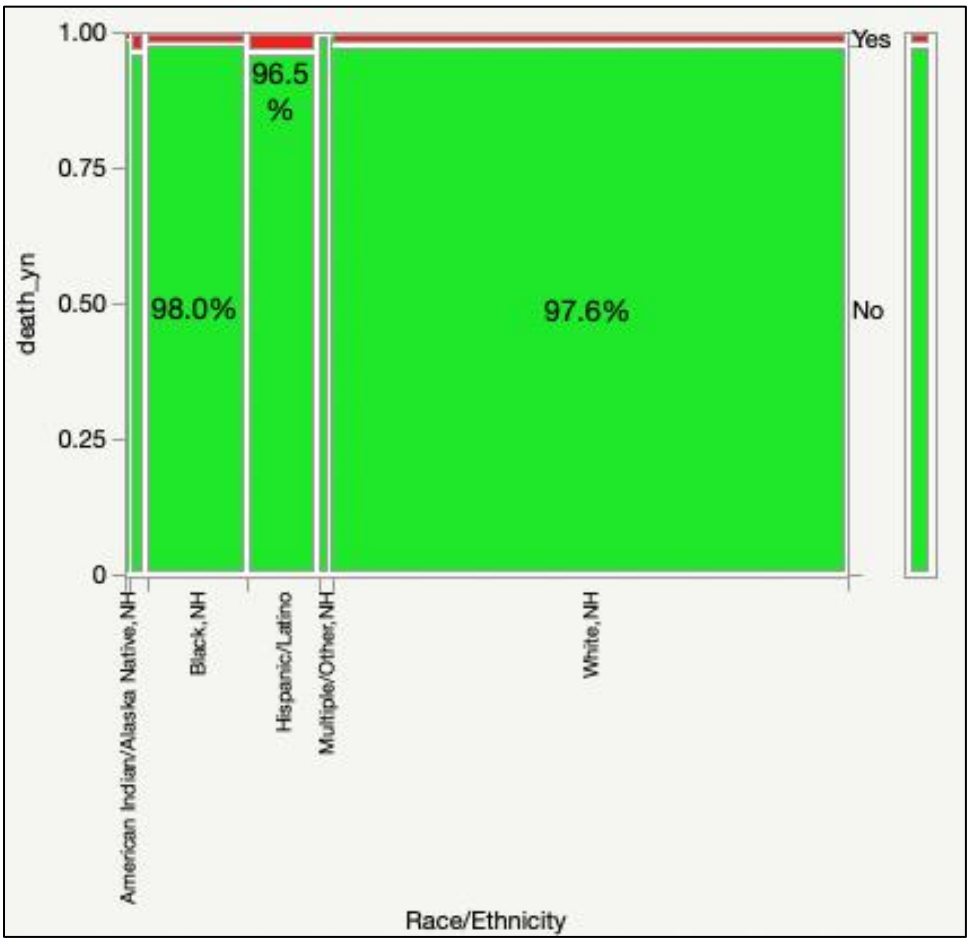
*NH=Non-Hispanic/Latino

The following mosaic plots show each variable in order of significance (most to least) and include the percentages of death and survival:

Contingency Analysis of Death by Age

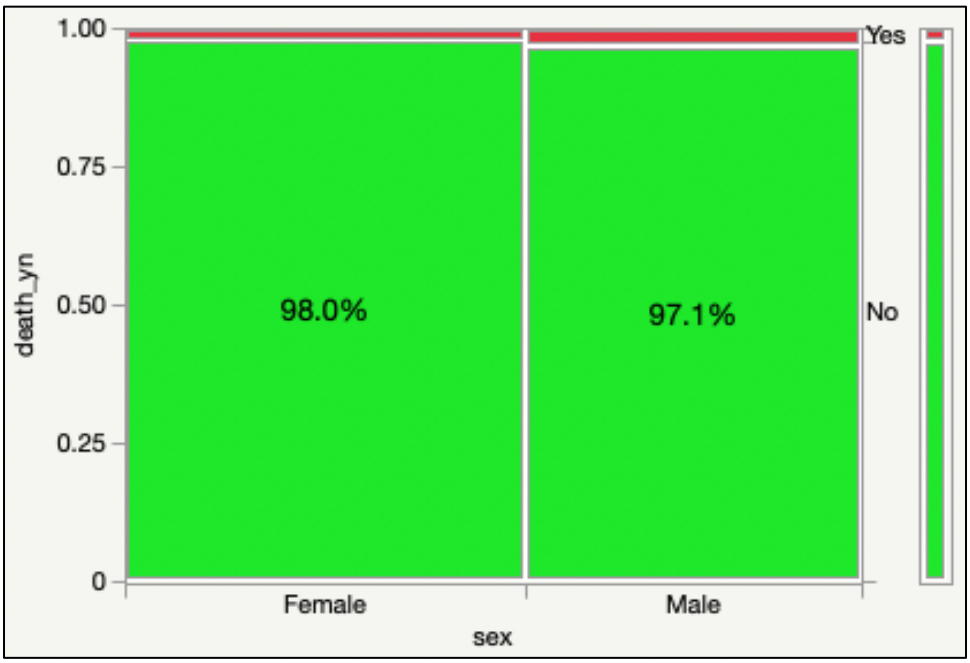


Contingency Analysis of Death by Race/Ethnicity



*NH=Non-Hispanic/Latino

Contingency Analysis of Death by Sex



Discussion and Results

Our analysis shows that age is the most significant measured variable that resulted in COVID-19 deaths. The age group with the highest predictor of death was the 65+ age group, while the group with the least likelihood of death was the 0-17 age group. Compared to the 0-17 years group, the odds of death for the 65+ years age group was 42974.4, and for 50-64 years, it was 2659.8. Sex was the least significant of the three variables. However, the odds of death for males compared to females was still 1.57.

Race/Ethnicity was the second most significant variable that contributed to death. Native Hawaiian/Other Pacific Islander, NH consistently had the lowest odds for death compared to the other races and ethnicities. Multiple/Other, NH also had lower odds for death than all others except for Native Hawaiian/Other Pacific Islander, NH. Asian, NH had the highest odds compared to all other races and ethnicities. Asian, NH had 18.7 greater odds of death compared to Native Hawaiian/Other Pacific Islander, NH, and 7.3 to Multiple/other, NH.

Compared to White, NH their odds were 4.1 greater, to Black, NH 2.5 greater and Hispanic/Latino 1.03 greater Hispanic/Latino had the second-highest odds of death compared to others except for the Asian, NH group. Compared to the Native Hawaiian/Other Pacific Islander, NH group, they had 18.1 greater odds of death than Multiple/Other, Non-Hispanic, 7.1 greater odds of death. The odds of death compared to White, NH, American Indian/Alaska Native, NH, and Black, NH were 3.98, 2.6, and 2.4 greater, respectively. Black, NH had greater death odds than all except Hispanic/Latino and Asian, NH. The odds of death were 7.5 greater than Native Hawaiian/Other Pacific Islander, NH, 2.9 greater compared to Multiple/Other, NH, 1.7 greater than White, NH, and 1.07 greater compared to American Indian/Alaska Native, NH. White, NH only had greater odds than Native Hawaiian/Other Pacific Islander, NH and Multiple/Other, NH, which was 4.5 and 1.8 greater. American India/Alaska Native, NH only had greater odds compared to Native Hawaiian/Other Pacific Islander, NH, Multiple/Other, NH and White, NH, which were 7.1, 2.8, and 1.6, respectively.

The implications of these findings can help individuals and healthcare providers understand the potential health consequences of contracting COVID-19. The model we produced will allow them to make better decisions based on a particular demographic profile. For example, a person whose race is Asian may want to consider being more careful regarding high exposure environments and using preventative measures such as masks since their odds are higher compared to other races/ethnicities. Also, knowing the demographic factors that are significant contributors to death by COVID-19 can help guide health professionals' advice for patients and the treatment they recommend. For example, a doctor may offer a more

aggressive treatment for patients 65+ and recommendations for reducing their risk knowing the risk for death is more significant for that age group.

Limitations

Several items could have improved our analysis of the data set provided by the CDC. The enormous size made it difficult to download and use in its raw format for initial exploratory evaluations. This led to limitations regarding the amount of computing power available for analysis. A visual scan of the data revealed there many data fields missing from rows. This allowed us to narrow the data and make it more manageable for the computing power available. Because there was not a unique identifier for any of the rows of data, we had to assume there were not any duplicates for the purposes of this project. Since the data was suppressed to protect individual privacy, some available data fields limited the analysis. For example, the age groups were broad and could be comparing people at different stages of life. Providing narrower bands of age groups should have been possible to allow a more detailed analysis without compromising individual privacy. With the lack of computing power and the lack of a unique identifier for each case, other significant variables that could have contributed to death from COVID-19 were not included as a part of this assessment. Underlying conditions could not be included due to missing data and the inability to relate it to other data available from the CDC. If there were an effective way to relate the data across tables, a more complete analysis could be performed with other variables that could lead to death from COVID-19.

Lessons Learned

This assignment provided an opportunity to dive into an extensive data set and attempt to solve a question around demographic variables that might impact mortality when

contracting COVID-19. However, our initial download contained tens of millions of rows that needed to be carefully analyzed, reviewed, and tested in iterations to determine the best way to measure critical variables and determine a solution. When analyzing and narrowing the data set, we had to carefully review rows and columns for missing information or ways to combine variables like race and ethnicity. We also learned that one could end up with a meaningful solution, even with imperfect data categorization. One key example of this is that some age groups were narrow, like 0-17, while other age groups like 18-49 were extremely broad. It would have been more interesting to see this age group split in two to further segment and see the influence of specific age groups on survival when contracting COVID-19.

Conclusions

Throughout the analysis of this data set, the team learned that age, race/ethnicity, and sex all contributed to the likelihood of death by COVID-19. Of all the variables we evaluated, age was the most significant and had the biggest demographic impact on survivability. People in the 64+ age group are most likely to die from COVID-19, with an odds ratio of 42974.4. As such, the team is in line with the CDC's recommendations that the elderly should exercise the most caution regarding preventative measures, and they should receive vaccinations before the remaining population. We also learned that Asian, Non-Hispanic individuals have a higher risk of death than other races and ethnicities. This is an opportunity for additional research that could explore why specific demographics are more at risk than others.