Georgia State University

CSC6780 – Fundamentals of Data Science

*Fall 2021*

Project Progress Report

Healthcare Reform Predictions

Reformed AI

Devin O’keefe

Jinsu Hong

Ji Zhang

Li Gan

**Table of Contents**

[**1 Business Understanding**](#_gjdgxs)3

[1.1 Business Problem](#_bnkyogcs40q6) 3

[1.2 Dataset](#_hwr14ta6o0cx) 3

[1.3 Proposed Analytics Solution](#_iy5juclrub6j) 3

[**2 Data Exploration and Preprocessing**](#_913hk1g9ex9q)4

[2.1 Data Quality Report](#_vrbdicq01d9p) 4

[2.2 Missing Values and Outliers](#_a9e5lucq94ha) 12

[2.3 Normalization](#_4f3b3tkc9ksx) 12

[2.4 Transformations](#_iy31tdg94rhj) 13

# 1 Business Understanding

The United States has multiple healthcare systems. Veterans have access to a comprehensive government-run health service; People over 65 are covered by federally funded Medicare system and low-income people are covered by Medicaid system which is also federal funded.

Most people who are employed may have private health insurances which are included in their employment packages. However, during Covid pandemic, massive laid off cause them lost insurance and fall deep into financial crisis.

We, the Reformed AI team, in favor of Healthcare Reform, would like to find out what is the cause of people who are against Universal Healthcare. Find their region and targeted ads toward this area to shift their views towards Public Healthcare system.

## 1.1 Business Problem

Whether it’s political, sports, fashion taste or food preference, people have different views. For instance, the left wants some form of free health care for everyone while the right wants it completely privatized.

What causes people choose one side over the other? We all agree that people are different. Does people’s age, gender, education level, occupation, family income and the city they live determine opinions they bear? Do their religion and cultural environment impact the choices they make?

We would like to analyze the data and find out the cause. Also, we would like to find the zip code of the people who are against public healthcare living in order to target ads to them.

## 1.2 Dataset

The dataset we choose <https://dataverse.harvard.edu/file.xhtml?fileId=5251021&version=1.0>

includes survey responses on health care access during the COVID-19 pandemic.

The data was collected between June 3rd and 8th, 2020. The survey included the following sections: message framing experiment, health reform, COVID-19 stimulus, policy response, personal response, personal worry about COVID-19, media, mental health, impact of COVID-19 on child/elderly care, ideology, cultural cognition, rumors, and demographic questions.

## 1.3 Proposed Analytics Solution

Based on the analytics data, we would like

* Target ads more toward the area where people favor private Healthcare, reduce ads budge in the area that people already favor public healthcare.
* Based on media outlets where people get information from, add more ads on the media that people who favor private health care watch or listen

# 2 Data Exploration and Preprocessing

The dataset is survey result table. It contains 1211 rows and 166 columns. In 166 columns, 158 columns’ type is object, 1 column type is Boolean, 7 columns’ type is int64 or float64. After study the data, we move 3 columns from Continuous Features to Categorical Features. There are total 162 Categorical features and 4 Continuous Features.

## 2.1 Data Quality Report

**Table 1**. Data Quality Report for Categorical Features

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Feature** | **Count** | **% of Missing** | **Card.** | **Mode** | **Mode Freq.** | **Mode %** | **2nd Mode** | **2nd Mode Freq.** | **2nd Mode Perc** |
| 0 | STARTDATE | 1211 | 0.00% | 590 | 6/3/2020 1... | 19 | 1.57% | 6/3/2020 1... | 13 | 1.07% |
| 1 | ENDDATE | 1211 | 0.00% | 735 | 6/3/2020 1... | 9 | 0.74% | 6/3/2020 1... | 7 | 0.58% |
| 2 | RECORDEDDATE | 1211 | 0.00% | 733 | 6/3/2020 1... | 9 | 0.74% | 6/3/2020 1... | 7 | 0.58% |
| 3 | DISTRIBUTIONCHANNEL | 1211 | 0.00% | 1 | anonymous | 1211 | 100.00% |  |  |  |
| 4 | USERLANGUAGE | 1211 | 0.00% | 1 | EN | 1211 | 100.00% |  |  |  |
| 5 | CONSENT | 1211 | 0.00% | 1 | Yes, I con... | 1211 | 100.00% |  |  |  |
| 6 | COMPREHENSION1\_ARM1 | 382 | 68.46% | 2 | Coronaviru... | 358 | 29.56% | Hotel felt... | 24 | 1.98% |
| 7 | COMPREHENSION1\_ARM2 | 382 | 68.46% | 2 | He had emp... | 325 | 26.84% | His state ... | 57 | 4.71% |
| 8 | COMPREHENSION2\_ARM1 | 391 | 67.71% | 2 | Hotel felt... | 283 | 23.37% | He was dis... | 108 | 8.92% |
| 9 | COMPREHENSION2\_ARM2 | 391 | 67.71% | 2 | He had emp... | 307 | 25.35% | His state ... | 84 | 6.94% |
| 10 | HR2\_1 | 1211 | 0.00% | 6 | Very Posit... | 333 | 27.50% | Somewhat P... | 288 | 23.78% |
| 11 | HR2\_2 | 1211 | 0.00% | 6 | Somewhat P... | 286 | 23.62% | Very Posit... | 272 | 22.46% |
| 12 | HR2\_3 | 1211 | 0.00% | 6 | Very Posit... | 404 | 33.36% | Somewhat P... | 365 | 30.14% |
| 13 | HR2\_4 | 1211 | 0.00% | 6 | Somewhat P... | 373 | 30.80% | Very Posit... | 347 | 28.65% |
| 14 | HR2\_5 | 1211 | 0.00% | 6 | Very Posit... | 354 | 29.23% | Somewhat P... | 318 | 26.26% |
| 15 | HR3 | 1211 | 0.00% | 5 | Very satis... | 481 | 39.72% | Somewhat s... | 437 | 36.09% |
| 16 | HR4 | 1211 | 0.00% | 5 | Good | 420 | 34.68% | Excellent | 295 | 24.36% |
| 17 | HR5 | 1211 | 0.00% | 5 | Very satis... | 422 | 34.85% | Somewhat s... | 417 | 34.43% |
| 18 | HR6 | 1211 | 0.00% | 5 | Poor | 325 | 26.84% | Good | 308 | 25.43% |
| 19 | HR7\_1 | 1211 | 0.00% | 5 | Strongly F... | 380 | 31.38% | Favor | 377 | 31.13% |
| 20 | HR7\_2 | 1211 | 0.00% | 5 | Favor | 375 | 30.97% | Strongly F... | 302 | 24.94% |
| 21 | HR7\_3 | 1211 | 0.00% | 5 | Favor | 358 | 29.56% | Strongly F... | 280 | 23.12% |
| 22 | HR7\_4 | 1211 | 0.00% | 5 | Favor | 445 | 36.75% | Strongly F... | 330 | 27.25% |
| 23 | HR8 | 1211 | 0.00% | 5 | Strongly f... | 442 | 36.50% | Somewhat f... | 372 | 30.72% |
| 24 | HR10 | 298 | 75.39% | 286 | Idk | 5 | 0.41% | No | 4 | 0.33% |
| 25 | HR11 | 1209 | 0.17% | 4 | Creating a... | 534 | 44.10% | Incrementa... | 449 | 37.08% |
| 26 | HR11\_6\_TEXT | 34 | 97.19% | 34 | People sho... | 1 | 0.08% | Really clo... | 1 | 0.08% |
| 27 | HR12\_1 | 1211 | 0.00% | 3 | Mostly Pos... | 756 | 62.43% | Mostly Neg... | 257 | 21.22% |
| 28 | HR12\_2 | 1211 | 0.00% | 3 | Mostly Pos... | 704 | 58.13% | Mostly Neg... | 321 | 26.51% |
| 29 | HR12\_3 | 1211 | 0.00% | 3 | Mostly Pos... | 678 | 55.99% | Mostly Neg... | 268 | 22.13% |
| 30 | HR12\_4 | 1211 | 0.00% | 3 | Mostly Pos... | 601 | 49.63% | Mostly Neg... | 387 | 31.96% |
| 31 | HR12\_5 | 1211 | 0.00% | 3 | Mostly Pos... | 746 | 61.60% | Mostly Neg... | 258 | 21.30% |
| 32 | HR13 | 1211 | 0.00% | 6 | No | 586 | 48.39% | Yes, someb... | 218 | 18.00% |
| 33 | HR14 | 1211 | 0.00% | 9 | Private he... | 427 | 35.26% | Private he... | 202 | 16.68% |
| 34 | HR14\_10\_TEXT | 8 | 99.34% | 8 | Nothing | 1 | 0.08% | Federal ci... | 1 | 0.08% |
| 35 | COVID\_HR1 | 726 | 40.05% | 2 | Some- i.e,... | 507 | 41.87% | All | 219 | 18.08% |
| 36 | COVID\_HR2 | 807 | 33.36% | 4 | C. A combi... | 450 | 37.16% | B. Health ... | 267 | 22.05% |
| 37 | COVID\_HR2\_11\_TEXT | 11 | 99.09% | 11 | People hav... | 1 | 0.08% | Nothing | 1 | 0.08% |
| 38 | COVID\_HR3 | 1211 | 0.00% | 3 | Yes, I am ... | 657 | 54.25% | Has not af... | 333 | 27.50% |
| 39 | STIMULUS1 | 1211 | 0.00% | 72 | Nothing ch... | 375 | 30.97% | Work from ... | 346 | 28.57% |
| 40 | STIMULUS2 | 1211 | 0.00% | 3 | No | 610 | 50.37% | Yes, someo... | 333 | 27.50% |
| 41 | STIMULUS3 | 1211 | 0.00% | 4 | Adequate | 509 | 42.03% | Not adequa... | 312 | 25.76% |
| 42 | STIMULUS4 | 1211 | 0.00% | 4 | Adequate | 499 | 41.21% | Not adequa.. | 320 | 26.42% |
| 43 | STIMULUS5 | 1211 | 0.00% | 30 | Yes, check... | 605 | 49.96% | No | 278 | 22.96% |
| 44 | STIMULUS5\_8\_TEXT | 18 | 98.51% | 17 | Direct Dep... | 2 | 0.17% | My husband... | 1 | 0.08% |
| 45 | STIMULUS6 | 1211 | 0.00% | 5 | More direc... | 384 | 31.71% | More stimu... | 340 | 28.08% |
| 46 | STIMULUS6\_7\_TEXT | 45 | 96.28% | 44 | Unsure | 2 | 0.17% | Another st... | 1 | 0.08% |
| 47 | POLICY1 | 1211 | 0.00% | 3 | A little Ã¯... | 523 | 43.19% | Not at all... | 453 | 37.41% |
| 48 | POLICY2 | 1211 | 0.00% | 3 | Very conce... | 461 | 38.07% | Somewhat c... | 445 | 36.75% |
| 49 | POLICY3 | 1211 | 0.00% | 3 | Coronaviru... | 649 | 53.59% | The diseas... | 313 | 25.85% |
| 50 | POLICY4 | 1211 | 0.00% | 2 | People sho... | 628 | 51.86% | While Coro... | 583 | 48.14% |
| 51 | POLICY5 | 1211 | 0.00% | 351 | Issue a st... | 131 | 10.82% | Encourage ... | 88 | 7.27% |
| 52 | POLICY5\_13\_TEXT | 36 | 97.03% | 36 | Use your b... | 1 | 0.08% | DO NOTHING... | 1 | 0.08% |
| 53 | POLICY6 | 1211 | 0.00% | 264 | Encourage ... | 132 | 10.90% | Issue a st... | 112 | 9.25% |
| 54 | POLICY6\_10\_TEXT | 35 | 97.11% | 35 | DO NOTHING... | 1 | 0.08% | Do nothing... | 1 | 0.08% |
| 55 | PERSONAL1 | 1211 | 0.00% | 59 | Staying ho... | 205 | 16.93% | Working fr... | 179 | 14.78% |
| 56 | PERSONAL2 | 1211 | 0.00% | 4 | Most of th... | 540 | 44.59% | All of the... | 388 | 32.04% |
| 57 | PERSONAL3 | 1211 | 0.00% | 5 | Highly unl... | 389 | 32.12% | Likely | 285 | 23.53% |
| 58 | PERSONAL4 | 1211 | 0.00% | 5 | Strongly a... | 605 | 49.96% | Agree | 360 | 29.73% |
| 59 | PERSONAL5 | 1211 | 0.00% | 203 | Wash hands... | 74 | 6.11% | Wash hands... | 53 | 4.38% |
| 60 | PERSONAL6 | 1211 | 0.00% | 169 | Experienci... | 87 | 7.18% | Experienci... | 75 | 6.19% |
| 61 | PERSONAL6\_13\_TEXT | 37 | 96.94% | 35 | none | 3 | 0.25% | non above | 1 | 0.08% |
| 62 | PERSONAL7 | 1211 | 0.00% | 167 | None of th... | 286 | 23.62% | Reduced in... | 134 | 11.07% |
| 63 | PERSONAL7\_11\_TEXT | 14 | 98.84% | 14 | Possibly b... | 1 | 0.08% | Igift urcy... | 1 | 0.08% |
| 64 | WORRY\_1 | 1211 | 0.00% | 4 | Very Worri... | 504 | 41.62% | Somewhat W... | 380 | 31.38% |
| 65 | WORRY\_2 | 1211 | 0.00% | 4 | Somewhat W... | 456 | 37.65% | Very Worri... | 411 | 33.94% |
| 66 | WORRY\_3 | 1211 | 0.00% | 4 | Very Worri... | 505 | 41.70% | Somewhat W... | 384 | 31.71% |
| 67 | WORRY\_4 | 1211 | 0.00% | 4 | Very Worri... | 458 | 37.82% | Somewhat W... | 451 | 37.24% |
| 68 | WORRY\_5 | 1211 | 0.00% | 4 | Very Worri... | 382 | 31.54% | Somewhat W... | 369 | 30.47% |
| 69 | WORRY\_6 | 1211 | 0.00% | 4 | Somewhat W... | 461 | 38.07% | Very Worri... | 409 | 33.77% |
| 70 | WORRY\_7 | 1211 | 0.00% | 4 | Very Worri... | 393 | 32.45% | Somewhat W... | 382 | 31.54% |
| 71 | WORRY\_8 | 1211 | 0.00% | 4 | Somewhat W... | 451 | 37.24% | Very Worri... | 440 | 36.33% |
| 72 | MEDIA1 | 1211 | 0.00% | 4 | A lot (dai... | 449 | 37.08% | Frequent (... | 421 | 34.76% |
| 73 | MEDIA2 | 1211 | 0.00% | 4 | When cover... | 403 | 33.28% | When cover... | 379 | 31.30% |
| 74 | MEDIA3 | 1211 | 0.00% | 207 | TV (e.g., ... | 119 | 9.83% | TV (e.g., ... | 89 | 7.35% |
| 75 | MEDIA3\_11\_TEXT | 23 | 98.10% | 23 | Radio | 1 | 0.08% | CNN online... | 1 | 0.08% |
| 76 | MEDIA4 | 1211 | 0.00% | 201 | TV (e.g., ... | 87 | 7.18% | TV (e.g., ... | 77 | 6.36% |
| 77 | MEDIA4\_11\_TEXT | 36 | 97.03% | 29 | None | 6 | 0.50% | none | 3 | 0.25% |
| 78 | MEDIA5 | 1211 | 0.00% | 6 | CNN | 403 | 33.28% | Fox News | 291 | 24.03% |
| 79 | MEDIA5\_8\_TEXT | 73 | 93.97% | 56 | None | 6 | 0.50% | local news... | 5 | 0.41% |
| 80 | MEDIA5\_1 | 291 | 75.97% | 4 | Both equal... | 142 | 11.73% | Sean Hanni... | 78 | 6.44% |
| 81 | MEDIA5\_1\_3\_TEXT | 19 | 98.43% | 17 | Neither | 2 | 0.17% | None | 2 | 0.17% |
| 82 | MEDIA6\_1 | 1211 | 0.00% | 6 | No confide... | 286 | 23.62% | Very littl... | 250 | 20.64% |
| 83 | MEDIA6\_2 | 1211 | 0.00% | 6 | Some confi... | 295 | 24.36% | A great de... | 279 | 23.04% |
| 84 | MEDIA6\_3 | 1211 | 0.00% | 6 | Some confi... | 352 | 29.07% | A great de... | 255 | 21.06% |
| 85 | MEDIA6\_4 | 1211 | 0.00% | 6 | A great de... | 322 | 26.59% | Complete c... | 291 | 24.03% |
| 86 | MEDIA6\_5 | 1211 | 0.00% | 6 | Some confi... | 269 | 22.21% | Very littl... | 257 | 21.22% |
| 87 | MEDIA6\_6 | 1211 | 0.00% | 6 | Some confi... | 315 | 26.01% | Very littl... | 247 | 20.40% |
| 88 | MEDIA6\_7 | 1211 | 0.00% | 6 | Some confi... | 322 | 26.59% | Very littl... | 294 | 24.28% |
| 89 | MEDIA7 | 1211 | 0.00% | 5 | Good | 510 | 42.11% | Very good | 407 | 33.61% |
| 90 | MEDIA8 | 1211 | 0.00% | 85 | None | 630 | 52.02% | Hypertensi... | 102 | 8.42% |
| 91 | MEDIA8\_9\_TEXT | 72 | 94.05% | 69 | Depression... | 2 | 0.17% | Osteoporos... | 2 | 0.17% |
| 92 | MEDIA9 | 1211 | 0.00% | 3 | No | 582 | 48.06% | Yes | 581 | 47.98% |
| 93 | MEDIA10 | 1211 | 0.00% | 5 | Strongly ... | 358 | 29.56% | Somewhat ... | 324 | 26.75% |
| 94 | MEDIA11 | 1211 | 0.00% | 5 | Never smok... | 560 | 46.24% | Current Sm... | 240 | 19.82% |
| 95 | MEDIA12 | 1211 | 0.00% | 3 | Not at all... | 727 | 60.03% | Some days | 278 | 22.96% |
| 96 | MEDIA13 | 1211 | 0.00% | 3 | No | 852 | 70.36% | Yes | 201 | 16.60% |
| 97 | MEDIA14 | 1211 | 0.00% | 2 | No | 952 | 78.61% | Yes | 259 | 21.39% |
| 98 | MEDIA15 | 1211 | 0.00% | 4 | No, I have... | 710 | 58.63% | Yes, I was... | 320 | 26.42% |
| 99 | MEDIA16 | 1211 | 0.00% | 47 | No, have n... | 563 | 46.49% | Yes, from ... | 96 | 7.93% |
| 100 | MEDIA17 | 1211 | 0.00% | 2 | No | 761 | 62.84% | Yes | 450 | 37.16% |
| 101 | MH1 | 1211 | 0.00% | 4 | Not at all... | 455 | 37.57% | Several da... | 426 | 35.18% |
| 102 | MH2 | 1211 | 0.00% | 4 | Not at all... | 458 | 37.82% | Several da... | 398 | 32.87% |
| 103 | CHILDCARE1 | 1211 | 0.00% | 2 | No | 635 | 52.44% | Yes | 576 | 47.56% |
| 104 | CHILDCARE1\_1 | 576 | 52.44% | 175 | 5 | 30 | 2.48% | <4 | 28 | 2.31% |
| 105 | CHILDCARE1\_2 | 576 | 52.44% | 2 | Yes | 553 | 45.66% | No | 23 | 1.90% |
| 106 | CHILDCARE1\_3 | 576 | 52.44% | 5 | Kids are h... | 440 | 36.33% | Kids are s... | 67 | 5.53% |
| 107 | CHILDCARE1\_3\_6\_TEXT | 9 | 99.26% | 8 | Aunt | 2 | 0.17% | Watched by... | 1 | 0.08% |
| 108 | CHILDCARE1\_3\_7\_TEXT | 18 | 98.51% | 17 | None | 2 | 0.17% | Child is o... | 1 | 0.08% |
| 109 | CHILDCARE2 | 1211 | 0.00% | 3 | Yes | 586 | 48.39% | Don't Know... | 340 | 28.08% |
| 110 | CHILDCARE3 | 1211 | 0.00% | 2 | No | 971 | 80.18% | Yes | 240 | 19.82% |
| 111 | CHILDCARE3\_1 | 240 | 80.18% | 2 | Yes | 204 | 16.85% | No | 36 | 2.97% |
| 112 | CHILDCARE3\_2 | 204 | 83.15% | 2 | Yes | 175 | 14.45% | No | 29 | 2.39% |
| 113 | ELDERLYCARE1 | 1211 | 0.00% | 2 | No | 880 | 72.67% | Yes | 331 | 27.33% |
| 114 | IDEOLOGY1 | 1211 | 0.00% | 4 | A. Democra... | 507 | 41.87% | B. Republi... | 433 | 35.76% |
| 115 | IDEOLOGY1\_6\_TEXT | 20 | 98.35% | 16 | None | 5 | 0.41% | Non politi... | 1 | 0.08% |
| 116 | IDEOLOGY2 | 1211 | 0.00% | 9 | Democratic... | 500 | 41.29% | Republican... | 409 | 33.77% |
| 117 | IDEOLOGY2\_9\_TEXT | 4 | 99.67% | 4 | Ã½Ã½Ã½Ã½Ã½Ã½Ã½Ã½Ã½Ã½... | 1 | 0.08% | free the p... | 1 | 0.08% |
| 118 | IDEOLOGY3 | 1211 | 0.00% | 4 | Yes | 693 | 57.23% | No, not pl... | 323 | 26.67% |
| 119 | IDEOLOGY4 | 888 | 26.67% | 3 | Joe Biden | 568 | 46.90% | Bernie San... | 243 | 20.07% |
| 120 | IDEOLOGY4\_5\_TEXT | 76 | 93.72% | 56 | Neither | 4 | 0.33% | Trump | 4 | 0.33% |
| 121 | IDEOLOGY5 | 1211 | 0.00% | 5 | Neither Ri... | 463 | 38.23% | Center Rig... | 311 | 25.68% |
| 122 | IDEOLOGY6 | 1211 | 0.00% | 5 | Somewhat p... | 470 | 38.81% | Deeply pol... | 285 | 23.53% |
| 123 | IDEOLOGY7 | 1211 | 0.00% | 6 | Donald Tru... | 439 | 36.25% | Hillary Cl... | 373 | 30.80% |
| 124 | IDEOLOGY7\_7\_TEXT | 22 | 98.18% | 17 | Jill Stein... | 3 | 0.25% | Gary Johns... | 2 | 0.17% |
| 125 | IDEOLOGY8 | 1211 | 0.00% | 6 | Joe Biden ... | 456 | 37.65% | Donald Tru... | 415 | 34.27% |
| 126 | IDEOLOGY8\_10\_TEXT | 17 | 98.60% | 16 | Bernie San... | 2 | 0.17% | Nobody is ... | 1 | 0.08% |
| 127 | CULTURE1\_1 | 1211 | 0.00% | 5 | Agree Stro... | 425 | 35.09% | Agree | 327 | 27.00% |
| 128 | CULTURE1\_2 | 1211 | 0.00% | 5 | Agree | 476 | 39.31% | Agree Stro... | 328 | 27.09% |
| 129 | CULTURE1\_3 | 1211 | 0.00% | 5 | Neither ag... | 392 | 32.37% | Agree | 265 | 21.88% |
| 130 | CULTURE1\_4 | 1211 | 0.00% | 5 | Agree | 364 | 30.06% | Neither ag... | 358 | 29.56% |
| 131 | CULTURE1\_5 | 1211 | 0.00% | 5 | Neither ag... | 343 | 28.32% | Agree | 299 | 24.69% |
| 132 | CULTURE1\_6 | 1211 | 0.00% | 5 | Neither ag... | 340 | 28.08% | Agree | 316 | 26.09% |
| 133 | CULTURE2\_1 | 1211 | 0.00% | 5 | Agree Stro... | 360 | 29.73% | Neither ag... | 270 | 22.30% |
| 134 | CULTURE2\_2 | 1211 | 0.00% | 5 | Agree | 409 | 33.77% | Agree Stro... | 368 | 30.39% |
| 135 | CULTURE2\_3 | 1211 | 0.00% | 5 | Agree Stro... | 413 | 34.10% | Agree | 344 | 28.41% |
| 136 | CULTURE2\_4 | 1211 | 0.00% | 5 | Agree Stro... | 442 | 36.50% | Agree | 357 | 29.48% |
| 137 | CULTURE2\_5 | 1211 | 0.00% | 5 | Agree Stro... | 447 | 36.91% | Agree | 359 | 29.64% |
| 138 | CULTURE2\_6 | 1211 | 0.00% | 5 | Agree Stro... | 476 | 39.31% | Agree | 359 | 29.64% |
| 139 | CULTURE2\_7 | 1211 | 0.00% | 5 | Neither ag... | 306 | 25.27% | Agree Stro... | 300 | 24.77% |
| 140 | RUMORS1 | 1211 | 0.00% | 70 | It acciden... | 179 | 14.78% | It was pur... | 137 | 11.31% |
| 141 | RUMORS1\_15\_TEXT | 44 | 96.37% | 42 | Don't know... | 2 | 0.17% | Na | 2 | 0.17% |
| 142 | RUMORS2 | 1211 | 0.00% | 7 | It acciden... | 324 | 26.75% | It was pur... | 295 | 24.36% |
| 143 | RUMORS2\_9\_TEXT | 56 | 95.38% | 46 | I don't kn... | 5 | 0.41% | Not sure | 4 | 0.33% |
| 144 | DEMOGRAPHIC1 | 1209 | 0.17% | 714 | retired | 47 | 3.88% | Retired | 38 | 3.14% |
| 145 | DEMOGRAPHIC2 | 1211 | 0.00% | 2 | No | 651 | 53.76% | Yes | 560 | 46.24% |
| 146 | DEMOGRAPHIC4 | 1211 | 0.00% | 51 | NY | 223 | 18.41% | CA | 152 | 12.55% |
| 147 | DEMOGRAPHIC5 | 1211 | 0.00% | 941 | 10001 | 69 | 5.70% | 90001 | 20 | 1.65% |
| 148 | DEMOGRAPHIC6 | 1211 | 0.00% | 2 | Yes | 1145 | 94.55% | No | 66 | 5.45% |
| 149 | DEMOGRAPHIC7 | 1211 | 0.00% | 2 | No | 1076 | 88.85% | Yes | 135 | 11.15% |
| 150 | DEMOGRAPHIC8 | 1211 | 0.00% | 11 | Protestant... | 447 | 36.91% | Roman Cath... | 347 | 28.65% |
| 151 | DEMOGRAPHIC8\_13\_TEXT | 125 | 89.68% | 77 | Christian | 17 | 1.40% | Christian ... | 6 | 0.50% |
| 152 | DEMOGRAPHIC9 | 1211 | 0.00% | 5 | Very impor... | 615 | 50.78% | Somewhat i... | 277 | 22.87% |
| 153 | DEMOGRAPHIC10 | 1211 | 0.00% | 3 | Female | 630 | 52.02% | Male | 577 | 47.65% |
| 154 | DEMOGRAPHIC10\_4\_TEXT | 4 | 99.67% | 4 | Cozuvpuud ... | 1 | 0.08% | Non confor... | 1 | 0.08% |
| 155 | DEMOGRAPHIC11 | 1211 | 0.00% | 8 | $20,001-$5... | 266 | 21.97% | $75,001-$1... | 230 | 18.99% |
| 156 | DEMOGRAPHIC12 | 1211 | 0.00% | 8 | White | 879 | 72.58% | Black or A... | 158 | 13.05% |
| 157 | DEMOGRAPHIC13 | 1211 | 0.00% | 5 | Middle cla... | 502 | 41.45% | Lower midd... | 260 | 21.47% |
| 158 | FINISHED | 1211 | 0.00% | 1 | TRUE | 1211 | 100.00% |  |  |  |
| 159 | HR1\_10SCALE | 1211 | 0.00% | 10 | 1 | 350 | 28.90% | 10 | 190 | 15.69% |
| 160 | ARM | 1211 | 0.00% | 3 | 0 | 438 | 36.17% | 1 | 391 | 32.29% |
| 161 | HR1 | 1211 | 0.00% | 1 | nan | 1211 | 100.00% |  |  |  |

**Table 2**. Data Quality Report for Continuous Features

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Feature** | **Count** | **% of Missing** | **Card.** | **Min.** | **Q1** | **Median** | **Q3** | **Max.** | **Mean** | **Std. Dev.** |
| 0 | PROGRESS | 1211 | 0.00% | 1 | 100 | 100 | 100 | 100 | 100 | 100 | 0 |
| 1 | DURATION | 1211 | 0.00% | 983 | 232 | 717.5 | 1168 | 1881.5 | 45050 | 1698.93 | 2354.99 |
| 2 | DEMOGRAPHIC3 | 1211 | 0.00% | 80 | 1924 | 1971 | 1982 | 1990 | 2010 | 1979.17 | 15.84 |
| 3 | ELDERLYCARE1\_1 | 331 | 72.67% | 30 | 0 | 1 | 1 | 2 | 9.87E+09 | 29811439 | 5.42E+08 |

**Table 3**. Data Quality Plan

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Feature** | **Number of Missing values** | **Data Quality Issue** | **Potential Handling Strategies** |
| 0 | COMPREHENSION1\_ARM1 | 829 | Missing, 68.46% | Delete features |
| 1 | COMPREHENSION1\_ARM2 | 829 | Missing, 68.46% |
| 2 | COMPREHENSION2\_ARM1 | 820 | Missing, 67.71% |
| 3 | COMPREHENSION2\_ARM2 | 820 | Missing, 67.71% |
| 4 | HR10 | 913 | Missing, 75.39% |
| 5 | HR11\_6\_TEXT | 1177 | Missing, 97.19% |
| 6 | HR14\_10\_TEXT | 1203 | Missing, 99.34% |
| 7 | COVID\_HR1 | 485 | Missing, 40.05% |
| 8 | COVID\_HR2 | 404 | Missing, 33.36% |
| 9 | COVID\_HR2\_11\_TEXT | 1200 | Missing, 99.09% |
| 10 | STIMULUS5\_8\_TEXT | 1193 | Missing, 98.51% |
| 11 | STIMULUS6\_7\_TEXT | 1166 | Missing, 96.28% |
| 12 | POLICY5\_13\_TEXT | 1175 | Missing, 97.03% |
| 13 | POLICY6\_10\_TEXT | 1176 | Missing, 97.11% |
| 14 | PERSONAL6\_13\_TEXT | 1174 | Missing, 96.94% |
| 15 | PERSONAL7\_11\_TEXT | 1197 | Missing, 98.84% |
| 16 | MEDIA3\_11\_TEXT | 1188 | Missing, 98.10% |
| 17 | MEDIA4\_11\_TEXT | 1175 | Missing, 97.03% |
| 18 | MEDIA5\_8\_TEXT | 1138 | Missing, 93.97% |
| 19 | MEDIA5\_1 | 920 | Missing, 75.97% |
| 20 | MEDIA5\_1\_3\_TEXT | 1192 | Missing, 98.43% |
| 21 | MEDIA8\_9\_TEXT | 1139 | Missing, 94.05% |
| 22 | CHILDCARE1\_1 | 635 | Missing, 52.44% |
| 23 | CHILDCARE1\_2 | 635 | Missing, 52.44% |
| 24 | CHILDCARE1\_3 | 635 | Missing, 52.44% |
| 25 | CHILDCARE1\_3\_6\_TEXT | 1202 | Missing, 99.26% |
| 26 | CHILDCARE1\_3\_7\_TEXT | 1193 | Missing, 98.51% |
| 27 | CHILDCARE3\_1 | 971 | Missing, 80.18% |
| 28 | CHILDCARE3\_2 | 1007 | Missing, 83.15% |
| 29 | ELDERLYCARE1\_1 | 880 | Missing, 72.67% |
| 30 | IDEOLOGY1\_6\_TEXT | 1191 | Missing, 98.35% |
| 31 | IDEOLOGY2\_9\_TEXT | 1207 | Missing, 99.67% |
| 32 | IDEOLOGY4 | 323 | Missing, 26.67% |
| 33 | IDEOLOGY4\_5\_TEXT | 1135 | Missing, 93.72% |
| 34 | IDEOLOGY7\_7\_TEXT | 1189 | Missing, 98.18% |
| 35 | IDEOLOGY8\_10\_TEXT | 1194 | Missing, 98.6% |
| 36 | RUMORS1\_15\_TEXT | 1167 | Missing, 96.37% |
| 37 | RUMORS2\_9\_TEXT | 1155 | Missing, 95.38% |
| 38 | DEMOGRAPHIC8\_13\_TEXT | 1086 | Missing, 89.68% |
| 39 | DEMOGRAPHIC10\_4\_TEXT | 1207 | Missing, 99.67% |
| 40 | HR12\_1 | 0 | High Card., 62.43% |
| 41 | HR12\_5 | 0 | High Card., 61.60% |
| 42 | MEDIA12 | 0 | High Card., 60.03% |
| 43 | MEDIA13 | 0 | High Card., 70.36% |
| 44 | MEDIA14 | 0 | High Card., 78.61% |
| 45 | MEDIA17 | 0 | High Card., 62.84% |
| 46 | CHILDCARE3 | 0 | High Card., 80.18% |
| 47 | DEMOGRAPHIC6 | 0 | High Card., 94.55% |
| 48 | DEMOGRAPHIC7 | 0 | High Card., 88.85% |
| 49 | DEMOGRAPHIC12 | 0 | High Card., 72.58% |

Chart, bar chart, box and whisker chart

Description automatically generated

**Figure 1**. Visualizations of Categorical and Continuous Features in Dataset

## 2.2 Missing Values and Outliers

* Missing Values: As our data set includes survey responses, the data set is mainly composed of ordinal and nominal data types. Our group investigated the percentage of missing over 60 percent. We found 37 features that have a missing percent over 60. We consider high missing percentages as unmeaningful data. Thus, we excluded the features that have a missing percent over 60.

* Outliers: There are 4 column data are

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Feature** | **Min.** | **Q1** | **Median** | **Q3** | **Max.** | **IQR** | **Outliers** | **Solution** |
| 0 | PROGRESS | 100 | 100 | 100 | 100 | 100 | 0 | 0.00% | Nothing |
| 1 | DURATION | 232 | 717.5 | 1168 | 1881.5 | 45050 | 1164 | 6.85% | Set to mean |
| 2 | DEMOGRAPHIC3 | 1924 | 1971 | 1982 | 1990 | 2010 | 29 | 1.73% | Set to mean |
| 3 | ELDERLYCARE1\_1 | 0 | 1 | 1 | 2 | 9.87E+09 | 1 | 4.87% | Set to mean |

## 2.3 Normalization

* Identify single value columns. Single value column has no statistics value and should be removed from dataset.
* Identify column with too many missing data. Any column contains more than 50% missing data are removed from dataset.
* Identify none business related columns. Drop survey columns: STARTDATE, ENDDATE, PROGRESS, DURATION, FINISHED and RECORDEDDATE

## 2.4 Transformations

Although Survey is a great way to collect data, It can be very hard to clean up. Survey Data often contains “Level of Agreement”, “Level of Quality” and “Frequency” that can be change to ordinal data. For this Survey data, it contains a lot of long text answers, so we did the following

* Select One Questions: Replace long text answer with answer index.
* Select all questions: Create one extra column for every answer option. If the answer appears in the answer text, set the value in the new column to be 1, otherwise to be 0. Drop the original column after all new columns are created and data are inserted.
* Matrix questions: Replace long text answers with “Level of Agreement”, “Level of Quality” and “Frequency” numbers.
* Text entry questions: This one is the hardest one to clean. Respondents can enter whatever they want, some answers are nonsensical feedback. We have to look at the data one at a time to decide if data is valid. We divide data in “DEMOGRAPHIC1” column into 5 groups: EMPLOYED, UNEMPLOYED, RETIRED, STUDENTS and OTHER. Create one column for each group and insert 1 or 0 based on answer text content.
* Add a column AGE by calculating age by giving birth year.
* Fill all missing data with “NA”