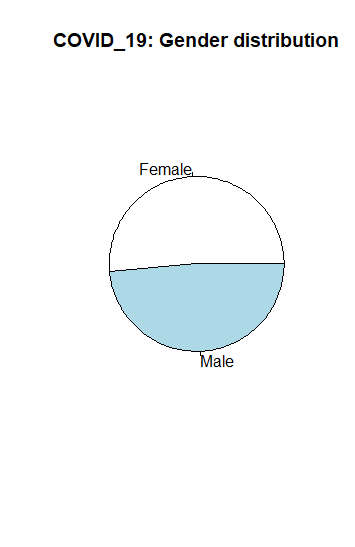
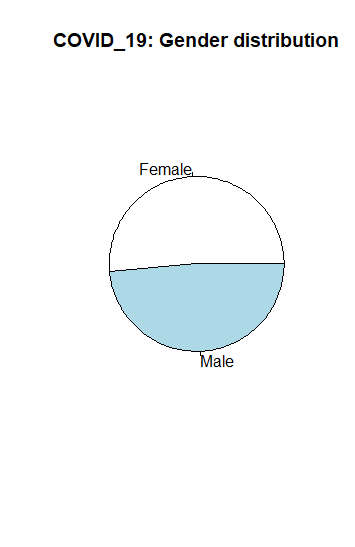
**COVID-19 Data Analysis**

* Hypothesis: The risk factor of COVID-19 can be predicted by statistical learning algorithm.
* Data set source: COVID-19 data set (from 2020-01-22 to 2020-05-14) provided by data.gov
* Source URL:

<https://data.cdc.gov/api/views/vbim-akqf/rows.csv?accessType=DOWNLOAD>

* Raw data set: The size of the data set was extremely large (1111016 obs. Of 10 variables), and there were lots of unknown values that cannot be described (1066688 unknowns). I removed all rows of the data of the data if there is at least one missing value. The size of the raw data set was able to reduce to 44328 rows. In addition, non-laboratory confirmed cases were excluded from this analysis. Total amount of data set was reduced to 44138 objects with 10 variables.
* Time data format: Time data such as “cdc\_report\_date”, “positive\_specimen\_collection\_date”, and “onset\_date” were formatted into POSIXct.
* Data Analysis:

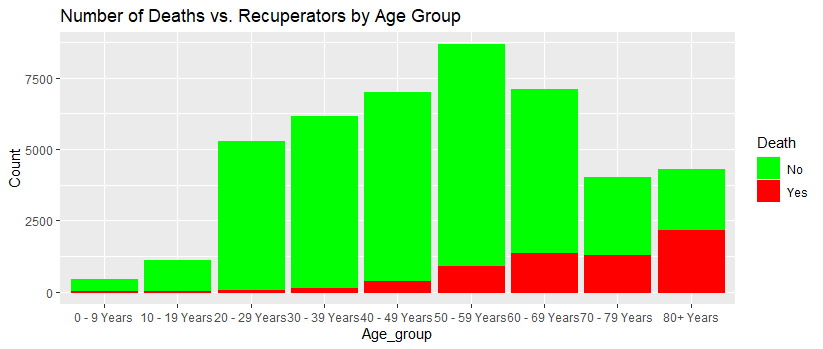


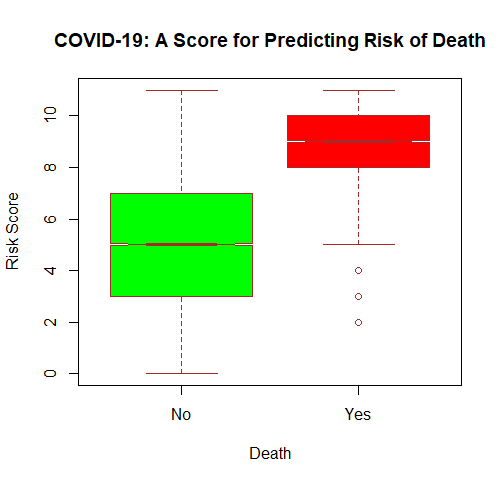
* The gender distribution of COVID-19 was very similar to each other. Number of the female group was slightly larger but we can conclude that there is no difference.

|  |  |
| --- | --- |
| Female | Male |
| 22690 | 21448 |

* Hospitalization rate of COVID-19 subjects was 33%.
* The death rate of COVID-19 from this data set was 14.1%. The death rate seems too large comparing to the CDC’s report. This is because unknown data was excluded from the analysis.

|  |  |  |
| --- | --- | --- |
| Age\_Group | Death | |
| No | Yes |
| 0 - 9 Years | 435 | 6 |
| 10 - 19 Years | 117 | 5 |
| 20 - 29 Years | 5251 | 45 |
| 30 - 39 Years | 6032 | 135 |
| 40 - 49 Years | 6661 | 356 |
| 50 - 59 Years | 7781 | 891 |
| 60 - 69 Years | 5748 | 1347 |
| 70 - 79 Years | 2761 | 1278 |
| 80+ Years | 2119 | 2170 |

* Age group versus number of the death was calculated as shown in the table. Maximum number of death occurred in the group of 80+ years. Overall, number of death mostly increased as the age group gets older. This tendency can be easily seen from the histogram drawn by ggplot and geom\_bar functions.
*  To use age group as a numerical value, we converted group names into corresponding number (age\_group\_number) such as "10 - 19 Years" into 1 and "20 - 29 Years" into 2. Also, “Yes” and “No” binary logics were converted into 1 and 0 to find a tendency that leads to the death. Then, summations of those converted variables (hosp\_yn, icu\_yn, medcond\_yn, age\_group\_number) were calculated as scores that can predict the death risk.



* The graph of risk score vs. death shows the good threshold around 8 that can predict the death of the subject.
* Logistic regression with multiple predictors were used to predict the probability of death. Data set was randomly divided into two group to create training and test set to perform regression. All variables that have binary logic (hosp\_yn, icu\_yn, medcond\_yn, age\_group\_number) were used to predict the death probability model. For all variables, P-values to death were less than 2e-16 which indicate strong correlations between individual variables to the death.
* Using test data set, regression coefficients were tested to predict the probability of death.
* Conclusion: Logistic regression using 4 medical records was able to predict the probability of death of COVID-19.