COVID19 Investigation

Amy Riffe

9/8/2020

Introduction

The year 2020 began with concerns over a newly emerging disease named Novel Corona Virus, or COVID19. Since this was a new disease, there was little information about its virulity and ability to spread. As the year continued, the spread of this virus became pandemic affecting countries all over the world. Insights about the virus came through collection of case information and evaluation of case investigations. Those insights inform the community about the health risks, offer opportunities for improved treatments, and provide direction for policy makers looking to mitigate the spread and effects of the disease.

For the purpose of the HarvardX Data Science PH125.9x course capstone, a local data file with information about COVID19 cases was obtained. The data file was de-identified for use in the course. There are many cases, but fewer are hospitalized because of the disease. The object of this capstone is to look at the case data and identify predictors for hospitalization.

Analysis

Two .csv data files were used during this project. One contained aggregate case counts over time and the other contained some information about each case. R version 3.6.1 was used for data analysis. Some general cleaning of the data files were done. Descriptive statistics in the manner of tables and charts were done to understand the data better. The data file with individual case information was used for inferential statistics. Both a linear regression model and a logistic regression model were developed to compare which model better predicted if a case would be hospitalized.

The data files were read into R from GitHub.

```
urlfile<-"https://raw.githubusercontent.com/AmyRiffe/Data-Science-
Capstone/master/Cases_Over_Time.csv"
cases<-read.csv(urlfile)

urlfile<-"https://raw.githubusercontent.com/AmyRiffe/Data-Science-
Capstone/master/case_list.csv"
caselist<-read.csv(urlfile)
```

The ‘cases’ file was examined and some descriptive information was identified. To see what the data file looked like, the first six lines were printed.

```
head(cases)
```

```
##   ï..Day.of.Report.Date Cases Day.number
## 1              9/3/2020    44        173
## 2              9/2/2020    54        172
## 3              9/1/2020    33        171
## 4             8/31/2020    35        170
## 5             8/30/2020    13        169
## 6             8/29/2020    38        168
```

Some additional information about the case file came from looking at the structure. The first variable was the date when a case was reported to the local health department. It was a factor variable instead of a date variable. It was left as such since it wasn't used in analysis. The Cases variable was the number of cases reported on that date. The Day.number variable was the number of days from the original case in the geography, which was March 14, 2020.

```
str(cases)
```

```
## 'data.frame':    168 obs. of  3 variables:
##  $ ï..Day.of.Report.Date: Factor w/ 168 levels "3/14/2020","3/17/2020",..:
168 167 166 159 158 156 155 154 153 152 ...
##  $ Cases                : int  44 54 33 35 13 38 56 45 47 35 ...
##  $ Day.number           : int  173 172 171 170 169 168 167 166 165 164 ...
```

Number of days in the file?

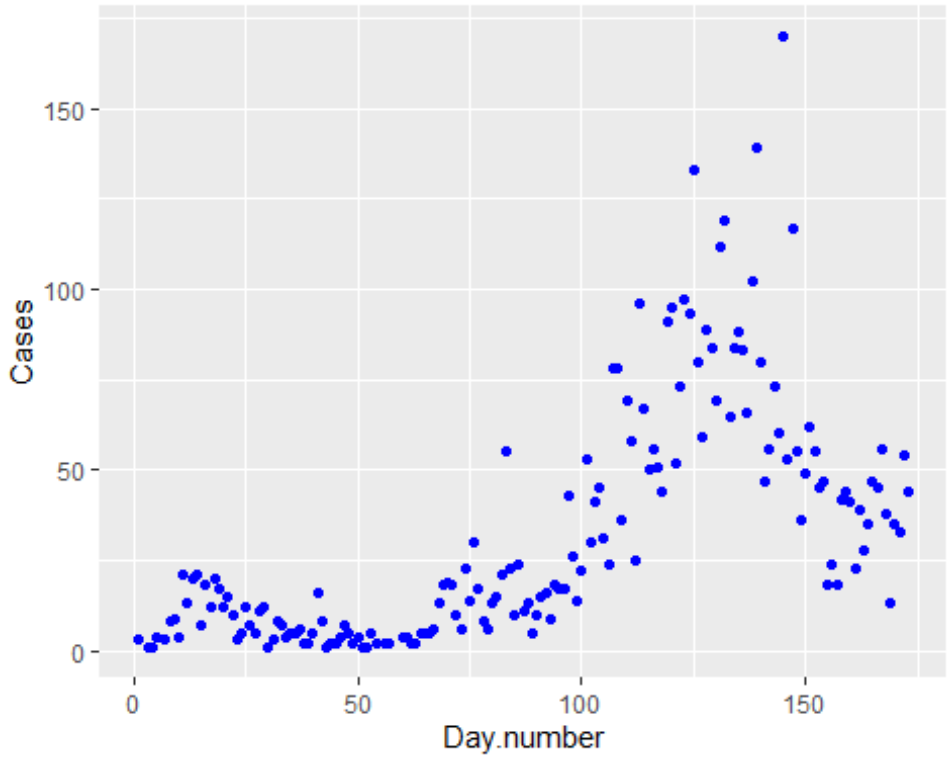
```
## [1] 173
```

The smallest and largest number of cases on a day were:

```
## [1] 1
```

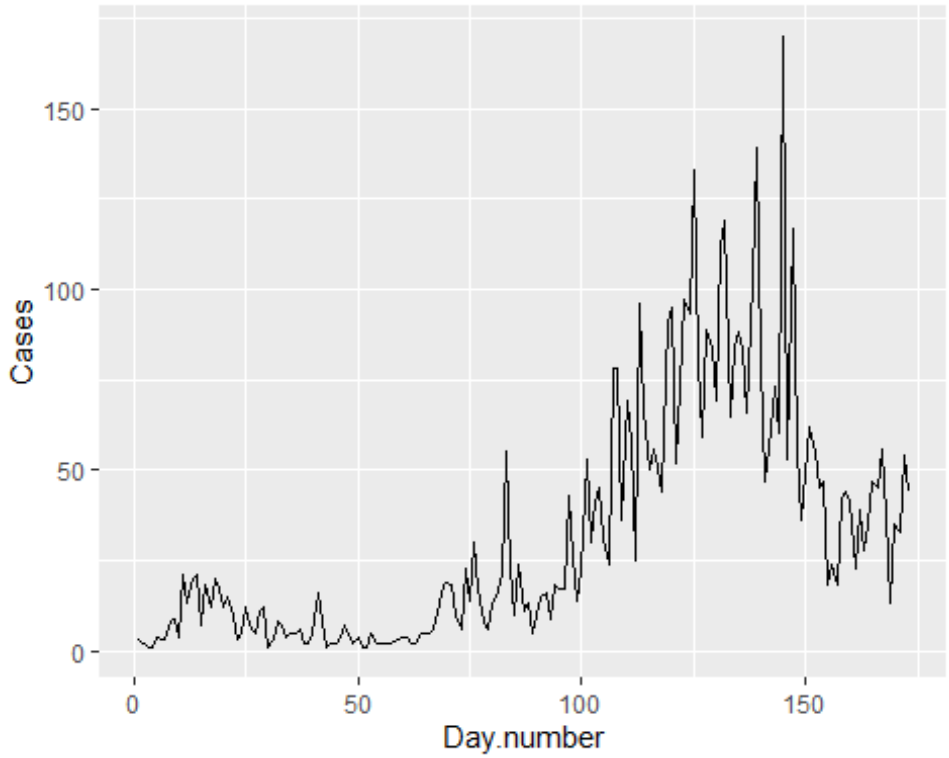
```
## [1] 170
```

Number of cases by day.



There has clearly been an increase in cases over time.

Number of cases by day.



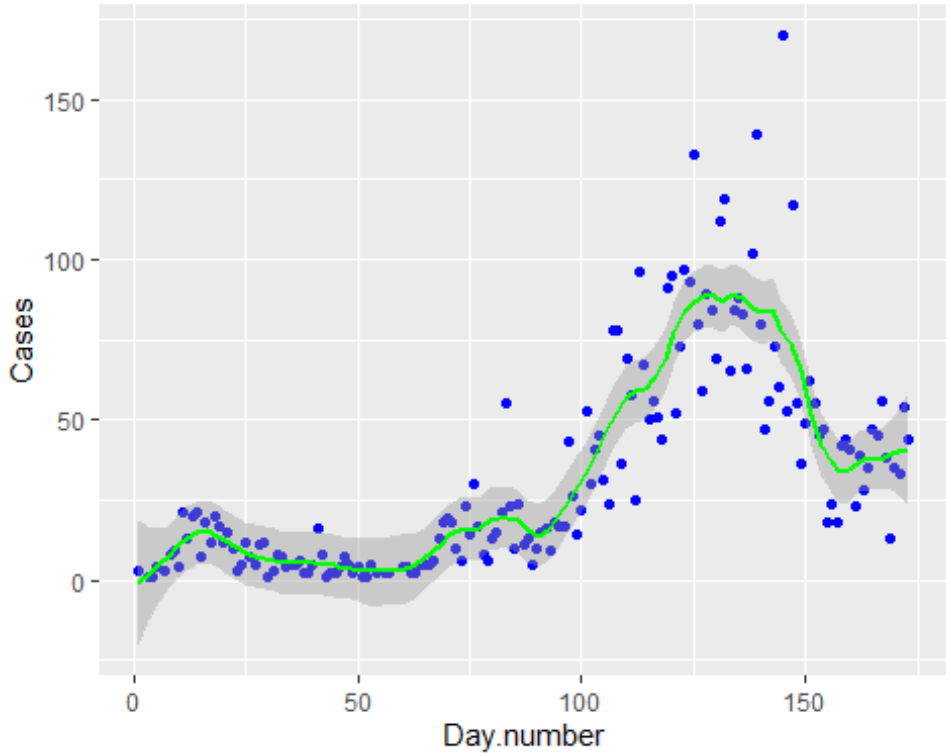
The daily line chart is much too jagged to provide information beyond a general increase over time. It is hard to tell how much the high counts impact the trend?

The next chart was done using a locally weighted smoothed line (LOESS). This takes smaller windows of data and plots a trend, then continues for the next window. It builds a smoother trend based on those window trends.

```
cases%>% ggplot(aes(Day.number, Cases))+
  geom_point(color="blue")+
  geom_smooth(color="green", span=.1, method.args=list(degree=1))
```

Number of cases by day, Smoothed using LOESS.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

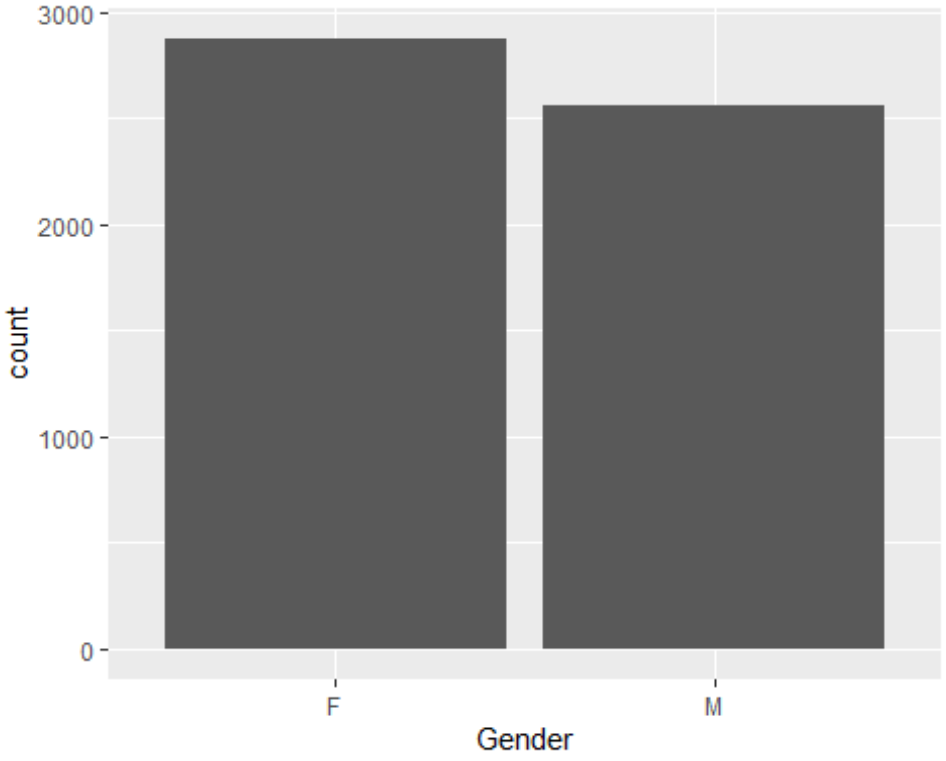


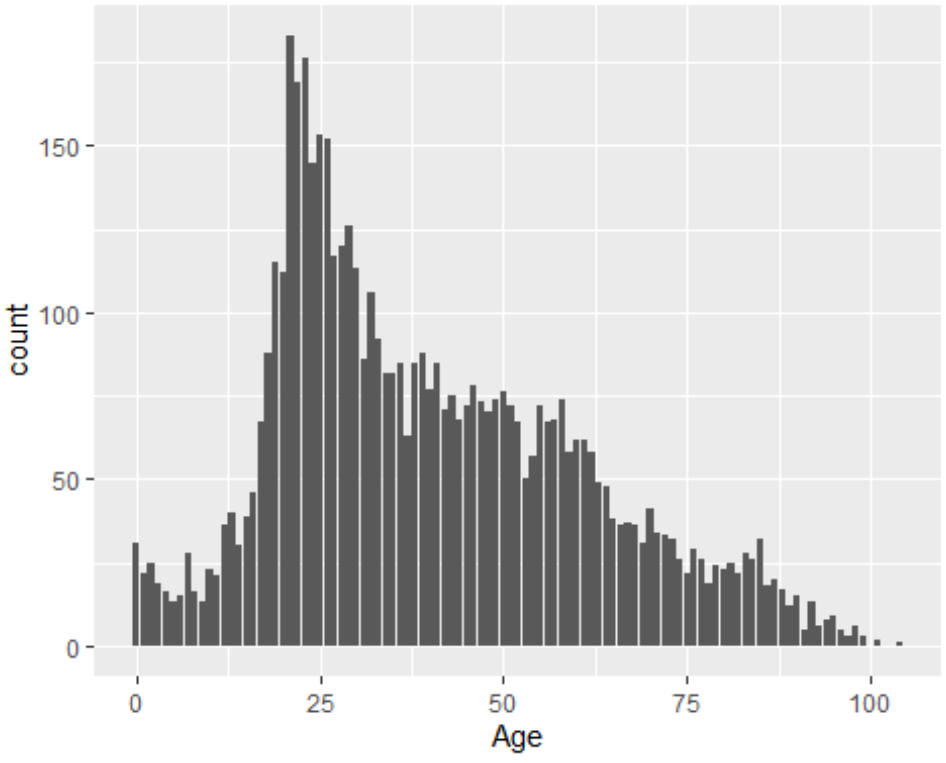
The LOESS trend shows that the days with really high counts don't affect the trend as much when smoothed with days around them. It appears the trend is beginning to level out again, but at a higher count than seen earlier in the year. Memorial day was on day 72 where there begins to be an increase. The fourth of July was on day 112, where the counts continue to rise. Where the counts begin to level off again is mid-August.

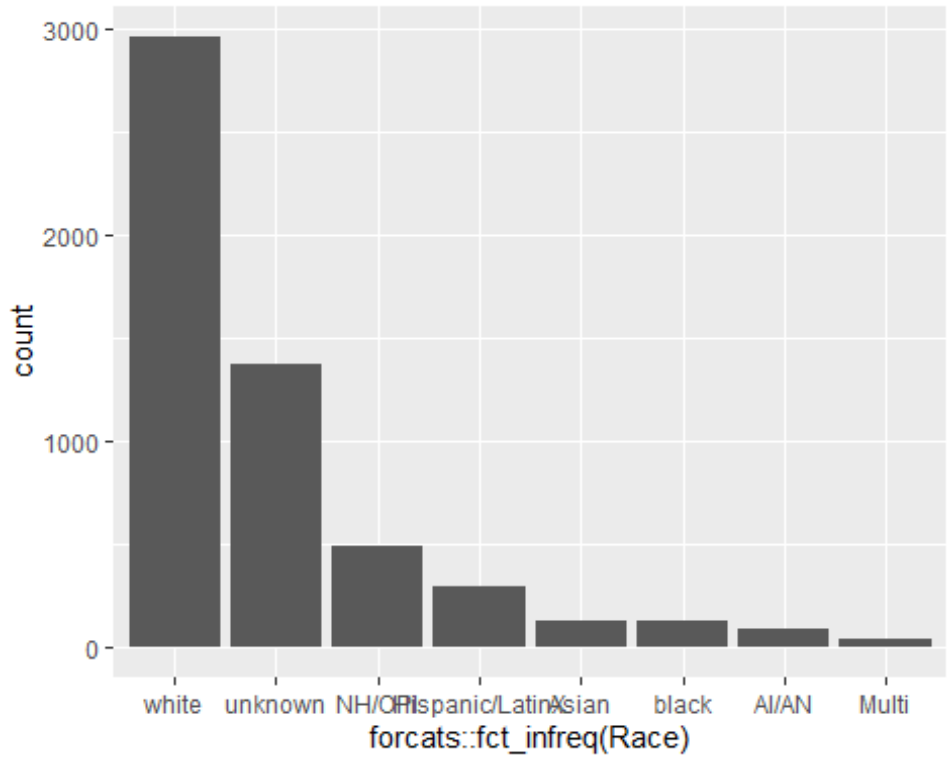
The caselist data file was used to analyze the association of case factors to cases being hospitalized. Some data cleaning was done to correct for misspellings in data, and the like. Once clean, charts were made to look at the demographic splits.

```
str(caselist)
```

```
## 'data.frame':    5484 obs. of  7 variables:
##  $ Age                   : int  46 50 26 41 72 62 22 42 88 72 ...
##  $ Gender                : Factor w/ 2 levels "F","M": 1 2 2 2 2 2 1 1 1 2
...
##  $ Race                  : Factor w/ 8 levels "white","AI/AN",..: 2 1 1 1
8 1 1 1 3 1 ...
##  $ Report.Date           : Factor w/ 168 levels
"3/14/2020","3/17/2020",..: 1 1 1 2 3 4 4 4 4 5 ...
##  $ Collection.Date       : Factor w/ 175 levels "","3/11/2020",..: 2 3 3 4
4 5 7 8 5 8 ...
##  $ Hospitalized.for.COVID: Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1
1 1 ...
##  $ DayNumber             : int  1 1 1 4 5 6 6 6 6 8 ...
```


3000
2000
1000
0
count
F
M
Gender


150
100
50
0
count
0
25
50
75
100
Age



The outcome variable is whether a case was hospitalized.

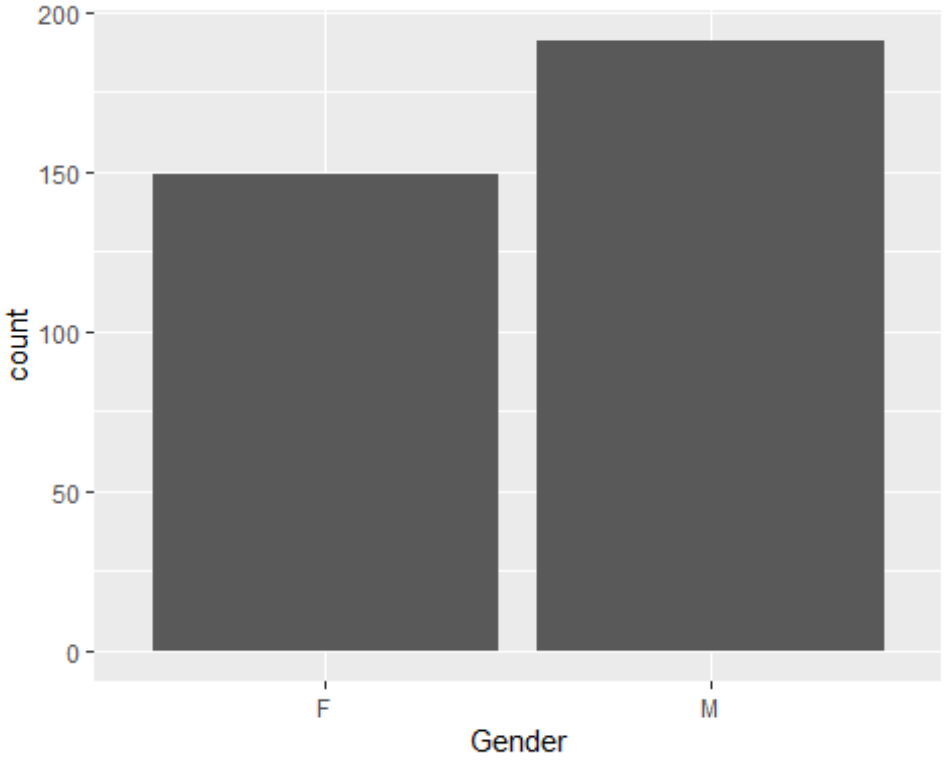
Out of the 5,484 cases, there were the following number of cases that were hospitalized.

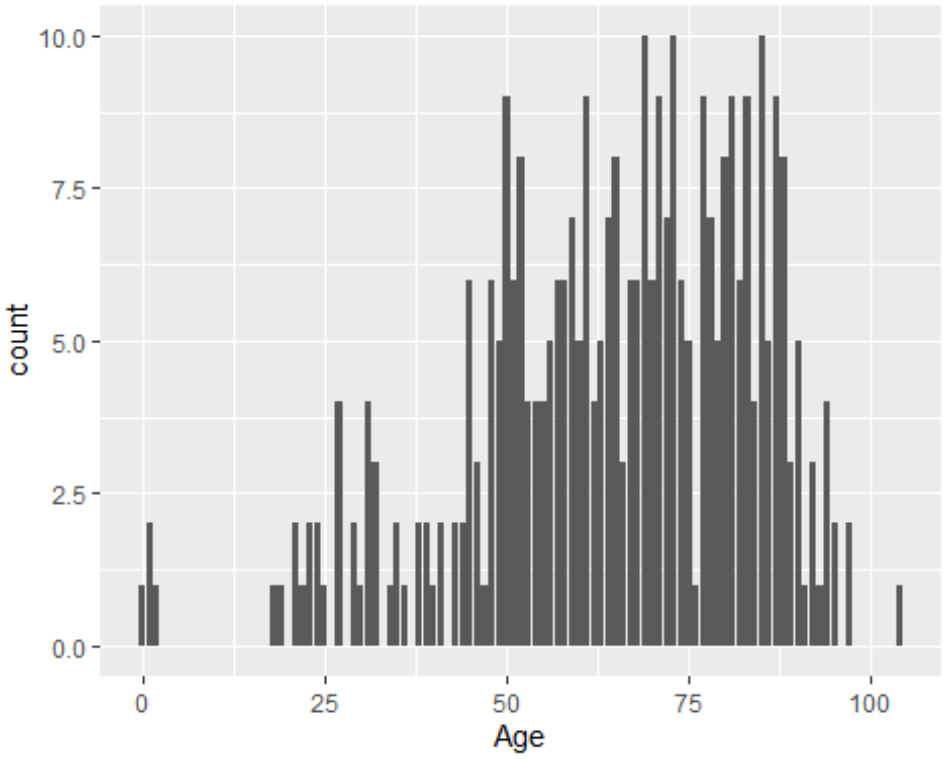
```
## yes
## 341
```

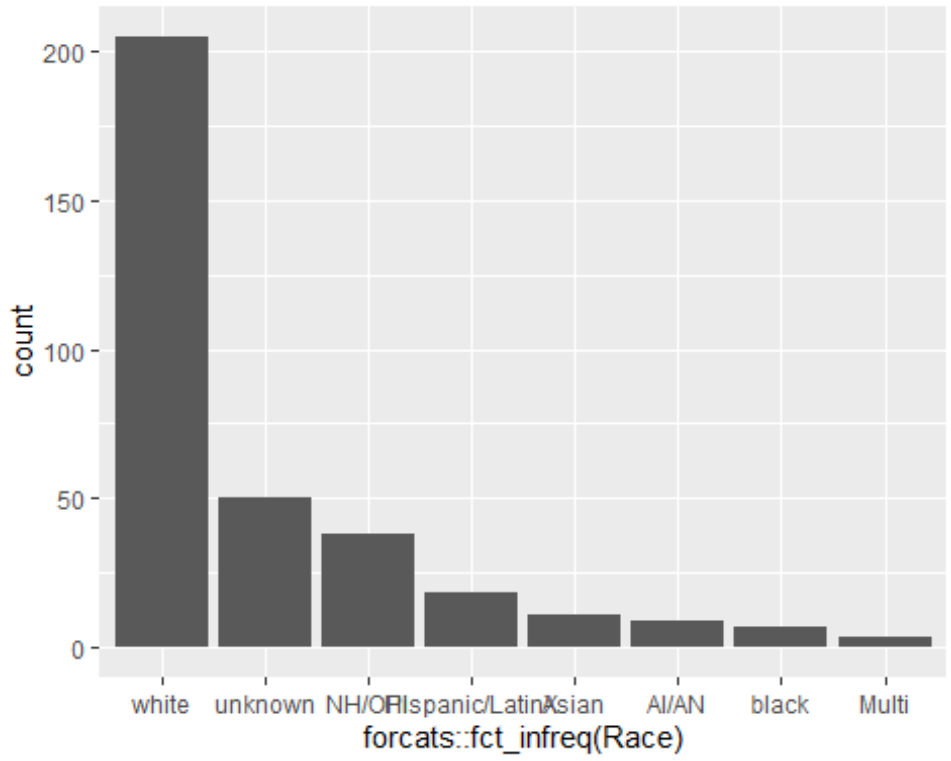
And the percent of cases that were hospitalized was the following.

```
##      yes
## 6.218089
```

Demographics for hospitalized cases looked like the following.


200
150
100
50
0
count
F
M
Gender


10.0
7.5
5.0
2.5
0.0
count
0
25
50
75
100
Age



There were some differences between all cases and hospitalized cases. More males were hospitalized. More older adults were hospitalized where as there were more younger adults among all cases. The racial distribution looked similar, excepting for fewer unknown race among hospitalizations.

The caselist data file was split into a training set and test set using the createDataPartition function of the caret package. The file was split into two files with a 50% partition. This level was chosen because the number of hospitalizations is few compared to the number of cases. This would ensure enough hospitalized cases were in both the training and the test sets for analysis to happen.

```
set.seed(1)
test_index <- createDataPartition(y=caselist$Hospitalized.for.COVID, times =
1, p = 0.5, list = FALSE)
train_cases <- caselist%>% slice(-test_index)
test_cases <- caselist%>% slice(test_index)
```

Model 1

The first model used for predicting hospitalization was a linear regression mode. This is the most basic model to start with. All were used regardless of the correlation as they are socially important to adjust for in a model. The outcome variable is hospitalized = yes and the predictive or independent variables were age, gender, race, and day number. All were used regardless of correlation as they are socially important to adjust for in a model.

```
lm_fit <- mutate(train_cases, y = as.numeric(Hospitalized.for.COVID ==
"yes")) %>% lm(y ~ Age+Gender+Race+DayNumber, data = .)
p_hat <- predict(lm_fit, test_cases)
cutoff<-seq(p_hat[which.min(p_hat)], p_hat[which.max(p_hat)], 0.05)
```

A linear regression model was fitted on the training set, then was used to predict the outcome on the test set. Cutoffs for determining accuracy were between the minimum predictive value and the maximum predictive value in increments of 0.05. They were: -0.12 -0.08 -0.03 0.02 0.07 0.12 0.17 0.22 0.27 0.32 0.37

After testing each cutoff, the one with the best accuracy was 0.37 with an accuracy of 0.9375.

```
cbind(cutoff, Accuracies)
```

```
##            cutoff   Accuracy
##  [1,] -0.12840773 0.06357957
##  [2,] -0.07840773 0.07607497
##  [3,] -0.02840773 0.16611540
##  [4,]  0.02159227 0.38478501
##  [5,]  0.07159227 0.61705255
##  [6,]  0.12159227 0.79015068
##  [7,]  0.17159227 0.89048144
##  [8,]  0.22159227 0.93311283
##  [9,]  0.27159227 0.93421536
## [10,]  0.32159227 0.93715546
## [11,]  0.37159227 0.93752297
```

This was the model for the linear regression model.

```
##
## Call:
## lm(formula = y ~ Age + Gender + Race + DayNumber, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32883 -0.09891 -0.03840  0.00911  1.06188
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.0031810  0.0201507  -0.158  0.87458
## Age                  0.0034747  0.0002140  16.238  < 2e-16 ***
## GenderM              0.0158517  0.0087881   1.804  0.07138 .
## RaceAI/AN            0.0703560  0.0361589   1.946  0.05179 .
## RaceAsian            0.0302180  0.0310919   0.972  0.33119
## Raceblack            0.0424825  0.0306286   1.387  0.16555
## RaceHispanic/LatinX  0.0271749  0.0200861   1.353  0.17620
## RaceMulti            0.0334925  0.0528061   0.634  0.52597
## RaceNH/OPI           0.0420026  0.0169588   2.477  0.01332 *
## Raceunknown         -0.0290105  0.0108107  -2.684  0.00733 **
## DayNumber           -0.0006631  0.0001316  -5.040 4.96e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2283 on 2707 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.1097, Adjusted R-squared:  0.1064
## F-statistic: 33.37 on 10 and 2707 DF,  p-value: < 2.2e-16
```

Model 2

The second model to be used was logistic regression. This method was chosen because the outcome variable was categorical and binary. Similar to the prior model, it was fitted on the training set and the model was used to predict using the test set. Cutoffs for determining accuracy were between the minimum predictive value (0) and the maximum predictive value (0.9) in increments of 0.1.

```
glm_fit <- train_cases %>%
  mutate(y = as.numeric(Hospitalized.for.COVID == "yes")) %>%
  glm(y ~ Age+Gender+Race+DayNumber, data=., family = "binomial")
p_hat_logit <- predict(glm_fit, newdata = test_cases, type = "response")
y_hat_logit <- ifelse(p_hat_logit > 0.37, "yes", "no") %>% factor
confusionMatrix(y_hat_logit,
test_cases$Hospitalized.for.COVID)$overall[["Accuracy"]]

## [1] 0.9331128

p_hat_logit[which.min(p_hat_logit)]

##         2639
## 0.0008037974

#0.0008
p_hat_logit[which.max(p_hat_logit)]

##       100
## 0.8889796

#0.89


cutoff_logit<-seq(0, 0.9, .1)

## Warning in confusionMatrix.default(y_hat_logit1,
## test_cases$Hospitalized.for.COVID): Levels are not in the same order for
## reference and data. Refactoring data to match.
```

After testing each cutoff, the one with the best accuracy was 0.7 with an accuracy of 0.9375.

```
cbind(cutoff_logit, Accuracies_logit)

## Warning in cbind(cutoff_logit, Accuracies_logit): number of rows of result
is
## not a multiple of vector length (arg 1)
```

```
##       cutoff_logit   Accuracy
##  [1,]          0.0 0.06284454
##  [2,]          0.1 0.83866226
##  [3,]          0.2 0.91216465
##  [4,]          0.3 0.92907019
##  [5,]          0.4 0.93348034
##  [6,]          0.5 0.93421536
##  [7,]          0.6 0.93531790
##  [8,]          0.7 0.93752297
##  [9,]          0.8 0.93752297
```

This was the model for the logistic regression model.

```
##
## Call:
## glm(formula = y ~ Age + Gender + Race + DayNumber, family = "binomial",
##     data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4710  -0.3334  -0.1926  -0.1238   3.4431
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -5.309876   0.404421 -13.130  < 2e-16 ***
## Age                  0.063129   0.004662  13.540  < 2e-16 ***
## GenderM              0.298445   0.174333   1.712 0.086911 .
## RaceAI/AN            1.100328   0.548541   2.006 0.044865 *
## RaceAsian            0.409488   0.501657   0.816 0.414345
## Raceblack            1.270511   0.506451   2.509 0.012119 *
## RaceHispanic/LatinX  0.752707   0.377025   1.996 0.045886 *
## RaceMulti            0.684828   1.179876   0.580 0.561629
## RaceNH/OPI           1.125541   0.292722   3.845 0.000121 ***
## Raceunknown         -0.848579   0.266894  -3.179 0.001476 **
## DayNumber           -0.007487   0.001977  -3.788 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1266.15  on 2717  degrees of freedom
## Residual deviance:  977.06  on 2707  degrees of freedom
##   (23 observations deleted due to missingness)
## AIC: 999.06
##
## Number of Fisher Scoring iterations: 7
```

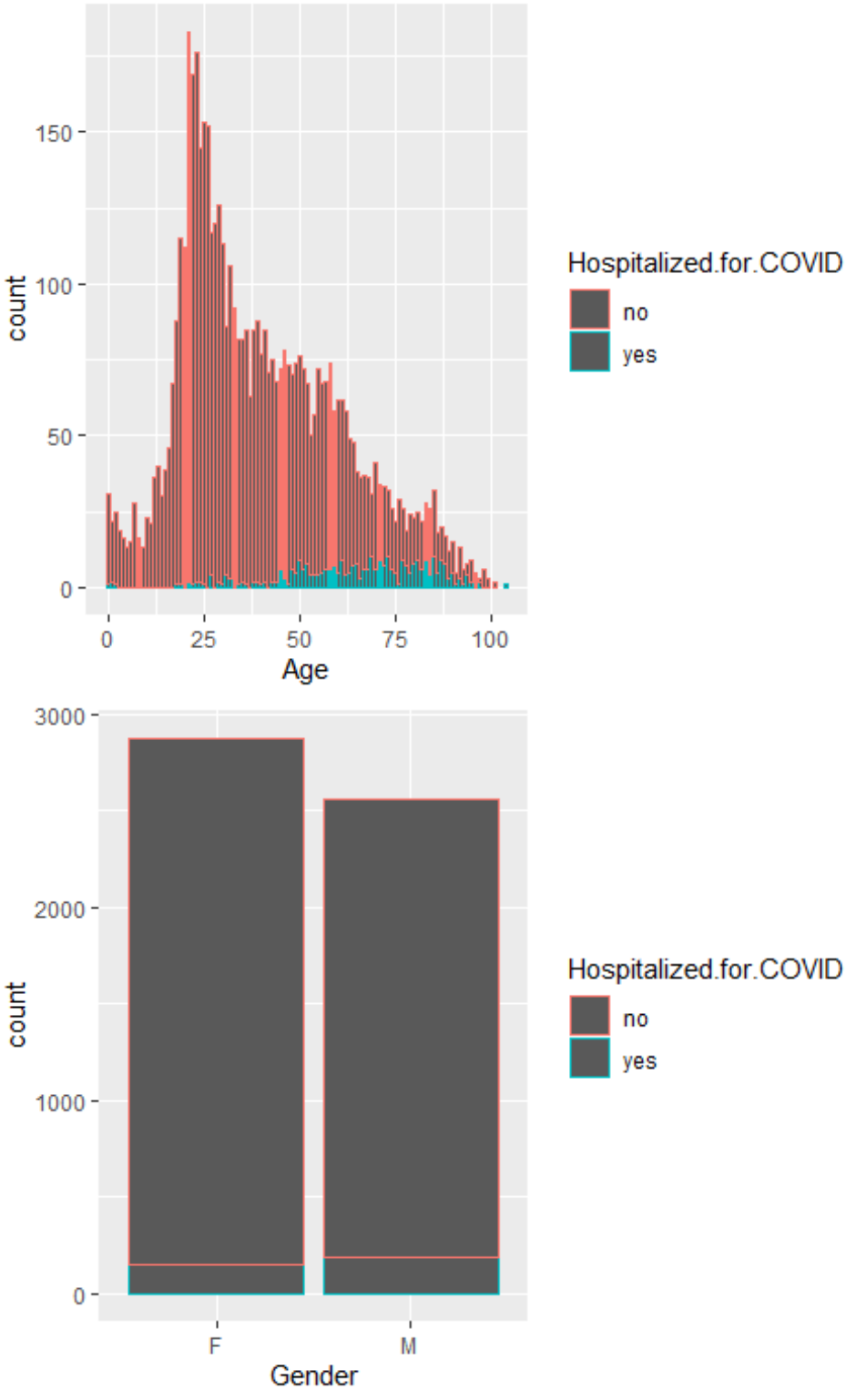
For better interpretability, the logistic regression model predictor coefficents were converted to an odds ratio.

```
or_glm(data=train_cases,
       model = glm_fit,
       incr = list(DayNumber=7, Age=10))
```

```
##              predictor oddsratio ci_low (2.5) ci_high (97.5)
increment
## 1                  Age     1.880        1.719          2.065
10
## 2              GenderM     1.348        0.958          1.900 Indicator
variable
## 3            RaceAI/AN     3.005        0.925          8.240 Indicator
variable
## 4            RaceAsian     1.506        0.515          3.774 Indicator
variable
## 5            Raceblack     3.563        1.172          8.878 Indicator
variable
## 6  RaceHispanic/LatinX     2.123        0.963          4.279 Indicator
variable
## 7            RaceMulti     1.983        0.093         13.290 Indicator
variable
## 8           RaceNH/OPI     3.082        1.702          5.393 Indicator
variable
## 9          Raceunknown     0.428        0.248          0.709 Indicator
variable
## 10           DayNumber     0.949        0.924          0.975
7
```

Results

COVID19 cases in this geography started out slowly, but increased greatly between days 90-140 (mid-June to the end of July). Case counts since then have decreased and appear to be leveling off. The data file includes cases through day 173, 09/03/2020. Descriptive analysis of COVID19 cases showed differences in demographics between all cases and those that have been hospitalized. There are more cases among young adults, but hospitalizations occur more among older adults. And there are more females among cases, but more males among hospitalizations.



Two models were used to determine which would best predict cases that would be hospitalized. Demographic factors of age, gender, race, and the temporal factor of day number of outbreak were used in the predictive models. Both models ended up with the

same accuracy of 0.9375. However, the logistic regression model ended up with more factors that were significant.

Age was very significant. For every change of 10 years and adjusting for all other factors the odds ratio was 1.88. This means that for each increase in age, the likelihood of being hospitalized with COVID19 increased 88%.

Compared to Whites and adjusting for all other factors:

- American Indians/Alaska Natives had an odds ratio of 3.01
- Blacks had an odds ratio of 3.56
- Hispanics/LatinX had an odds ratio of 2.12
- Native Hawaiian/Other Pacific Islander had an odds ratio of 3.08.

This means that compared to whites, these races were 2-3.5 times more likely to be hospitalized with COVID19.

Alternatively, unknown race had an odds ratio of 0.43. This means that cases with an unknown race were more than two time less likely to be hospitalized with COVID19. This may be a reflection of data entry completion in that those cases that are hospitalized may have more investigation time committed to them and may have more complete data in the case file.

The day number had an odds ratio of 0.949 for every change in 7 days and adjusting for all other factors. This means that as time of the outbreak continued, there was a small decrease in the likelihood of a case being hospitalized.

Gender, Asian race, and multi-racial were not significant predictive variables in the logistic regression model.

Conclusion

A small proportion of COVID19 cases end up being hospitalized. But older adults and most non-White race groups are at an increased risk for progressing to the point of needing hospitalization care if they contract COVID19. It may be that older adults are more medically fragile than younger adults due to co-morbidities. Differences by race may also be because of differences in health status between racial groups. But it may also be because of differences in behavior, such as where they live or what their occupation is.

For the purpose of the capstone project, these data were sufficient. The logistic regression model better fit the data for predicting hospitalization among COVID19 cases. A limitation of the analysis is that some identifiers were removed prior to use that may have been useful in the predictive model, such as geographic location and exposure risks. Further analysis with COVID19 case data that is not being made publicly available may find a better fitting model or may find additional or different predictive variables. At a minimum, these finding support the discussion in the community that older adults and people of color are disproportionately affected by COVID19.