Florida_Covid19_Analysis_script.R

user

2020-11-17

```
library(readxl)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(forcats)
library(ggplot2)
Florida_COVID19_Case_Line_Data_1_ <- read_excel ("C:/Users/user/Desktop/Florida_COVID19_Case_Line_Data
                                                 skip = 1)
View(Florida_COVID19_Case_Line_Data_1_)
str(Florida_COVID19_Case_Line_Data_1_)
```

```
## tibble [42,778 x 16] (S3: tbl_df/tbl/data.frame)
##  $ County        : chr [1:42778] "Orange" "Orange" "Orange" "Orange" ...
##  $ Age           : chr [1:42778] "42" "32" "28" "20" ...
##  $ Age_group     : chr [1:42778] "35-44 years" "25-34 years" "25-34 years" "15-24 years" ...
##  $ Gender        : chr [1:42778] "Female" "Male" "Male" "Male" ...
##  $ Jurisdiction  : chr [1:42778] "FL resident" "FL resident" "FL resident" "FL resident" ...
##  $ Travel_related: chr [1:42778] "Unknown" "No" "No" "Unknown" ...
##  $ Origin        : chr [1:42778] "NA" "NA" "NA" "NA" ...
##  $ EDvisit       : chr [1:42778] "UNKNOWN" "UNKNOWN" "NO" "UNKNOWN" ...
##  $ Hospitalized  : chr [1:42778] "UNKNOWN" "UNKNOWN" "NO" "UNKNOWN" ...
##  $ Died          : chr [1:42778] "NA" "NA" "NA" "NA" ...
##  $ Case_         : chr [1:42778] "Yes" "Yes" "Yes" "Yes" ...
##  $ Contact       : chr [1:42778] "YES" "UNKNOWN" "YES" "YES" ...
##  $ Case1         : chr [1:42778] "2020/06/23 05:00:00+00" "2020/07/01 05:00:00+00" "2020/06/24 05:00
##  $ EventDate     : chr [1:42778] "2020/06/23 00:00:00+00" "2020/07/01 00:00:00+00" "2020/06/18 00:00
##  $ ChartDate     : chr [1:42778] "2020/06/23 05:00:00+00" "2020/07/01 05:00:00+00" "2020/06/24 05:00
##  $ ObjectId      : num [1:42778] 45 71 127 128 129 130 137 138 139 140 ...
```

```
summary(Florida_COVID19_Case_Line_Data_1_)
```

```
##     County              Age             Age_group           Gender
##  Length:42778       Length:42778       Length:42778       Length:42778
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Jurisdiction       Travel_related        Origin            EDvisit
##  Length:42778       Length:42778       Length:42778       Length:42778
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Hospitalized           Died              Case_              Contact
##  Length:42778       Length:42778       Length:42778       Length:42778
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##     Case1             EventDate          ChartDate            ObjectId
##  Length:42778       Length:42778       Length:42778       Min.   :    45
##  Class :character   Class :character   Class :character   1st Qu.:177899
##  Mode  :character   Mode  :character   Mode  :character   Median :366653
##                                                           Mean   :369623
##                                                           3rd Qu.:560501
##                                                           Max.   :744986
```

```
county = as_factor(Florida_COVID19_Case_Line_Data_1_$County)
Age=as.numeric(Florida_COVID19_Case_Line_Data_1_$Age)
```

```
## Warning: NAs introduced by coercion
```

```
Age_group=as_factor(Florida_COVID19_Case_Line_Data_1_$Age_group)
Gender=as_factor(Florida_COVID19_Case_Line_Data_1_$Gender)
Jurisdiction=as_factor(Florida_COVID19_Case_Line_Data_1_$Jurisdiction)
Travel_related=as_factor(Florida_COVID19_Case_Line_Data_1_$Travel_related)
Origin=as_factor(Florida_COVID19_Case_Line_Data_1_$Origin)
EDvisit = as_factor(Florida_COVID19_Case_Line_Data_1_$EDvisit)
Hospitalized = as_factor(Florida_COVID19_Case_Line_Data_1_$Hospitalized)
Died = as_factor(Florida_COVID19_Case_Line_Data_1_$Died)
Case_ = as_factor(Florida_COVID19_Case_Line_Data_1_$Case_)
Contact = as_factor(Florida_COVID19_Case_Line_Data_1_$Contact)
ObjectId = as_factor(Florida_COVID19_Case_Line_Data_1_$ObjectId)
Case1 = as.POSIXct(Florida_COVID19_Case_Line_Data_1_$Case1)
month_case1 = as_factor (format(Case1, "%B"))
EventDate = as.POSIXct  (Florida_COVID19_Case_Line_Data_1_$EventDate)
month_EventDate = as_factor (format(EventDate, "%B"))
head(month_EventDate)
```

```
## [1] June July June June June June
## 10 Levels: June July May August September March October April ... January
```

```
ChartDate = as.POSIXct  (Florida_COVID19_Case_Line_Data_1_$ChartDate)
month_ChartDate = as_factor (format(ChartDate, "%B"))
head(month_ChartDate)
```

```
## [1] June July June June June June
## Levels: June July September August April October May March
```

```
Florida_COVID19_cases = data.frame(county, Age, Age_group, Gender, Jurisdiction,
                                   Travel_related, Origin, EDvisit, Hospitalized,Died, Case_,
                                   Contact,Case1, EventDate, ChartDate, month_case1,
                                   month_EventDate,
                                   month_ChartDate, ObjectId)
str(Florida_COVID19_cases)
```

```
## 'data.frame':    42778 obs. of  19 variables:
##  $ county         : Factor w/ 1 level "Orange": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age            : num  42 32 28 20 28 43 29 35 63 23 ...
##  $ Age_group      : Factor w/ 11 levels "35-44 years",..: 1 2 2 3 2 1 2 1 4 3 ...
##  $ Gender         : Factor w/ 3 levels "Female","Male",..: 1 2 2 2 1 1 2 2 2 2 ...
##  $ Jurisdiction   : Factor w/ 3 levels "FL resident",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Travel_related : Factor w/ 3 levels "Unknown","No",..: 1 2 2 1 1 2 1 1 1 1 ...
##  $ Origin         : Factor w/ 195 levels "NA","MOROCCO",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ EDvisit        : Factor w/ 4 levels "UNKNOWN","NO",..: 1 1 2 1 1 2 1 1 1 1 ...
##  $ Hospitalized   : Factor w/ 4 levels "UNKNOWN","NO",..: 1 1 2 1 1 2 1 1 1 1 ...
##  $ Died           : Factor w/ 2 levels "NA","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Case_          : Factor w/ 1 level "Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Contact        : Factor w/ 5 levels "YES","UNKNOWN",..: 1 2 1 1 1 3 1 1 1 1 ...
##  $ Case1          : POSIXct, format: "2020-06-23 05:00:00" "2020-07-01 05:00:00" ...
##  $ EventDate      : POSIXct, format: "2020-06-23 00:00:00" "2020-07-01 00:00:00" ...
##  $ ChartDate      : POSIXct, format: "2020-06-23 05:00:00" "2020-07-01 05:00:00" ...
##  $ month_case1    : Factor w/ 8 levels "June","July",..: 1 2 1 1 1 1 1 1 1 1 ...
##  $ month_EventDate: Factor w/ 10 levels "June","July",..: 1 2 1 1 1 1 1 1 1 1 ...
##  $ month_ChartDate: Factor w/ 8 levels "June","July",..: 1 2 1 1 1 1 1 1 1 1 ...
##  $ ObjectId       : Factor w/ 42778 levels "45","71","127",..: 1 2 3 4 5 6 7 8 9 10 ...
```

```
summary(Florida_COVID19_cases)
```

```
##     county           Age               Age_group        Gender
##  Orange:42778   Min.   :  0.00   25-34 years:9634   Female :21818
##                 1st Qu.: 25.00   15-24 years:8293   Male   :20571
##                 Median : 36.00   35-44 years:7410   Unknown:  389
##                 Mean   : 38.99   45-54 years:6194
##                 3rd Qu.: 51.00   55-64 years:4612
##                 Max.   :106.00   65-74 years:2419
##                 NA's   :49       (Other)    :4216
##                        Jurisdiction   Travel_related          Origin
##  FL resident                 :42316   Unknown:24848   NA          :42033
##  Non-FL resident             :  461   No     :17185   NY          :   53
##  Not diagnosed/isolated in FL:    1   Yes    :  745   FL; NY      :   53
##                                                      GA          :   39
##                                                      FL; GA      :   31
##                                                      FL; UNKNOWN:   23
```

```
##                                                    (Other)    :  546
##     EDvisit       Hospitalized    Died       Case_          Contact
##  UNKNOWN:27507   UNKNOWN:28005   NA :42244   Yes:42778   YES    :11659
##  NO     :11356   NO     :13081   Yes:  534               UNKNOWN:17050
##  YES    : 2783   YES    : 1519                           Yes    : 5585
##  NA     :   13   NA     :    2                           NO     : 6434
##  NA's   : 1119   NA's   :  171                           NA     : 2050
##
##
##      Case1                        EventDate
##  Min.   :2020-03-13 05:00:00   Min.   :2020-01-10 00:00:00
##  1st Qu.:2020-06-30 05:00:00   1st Qu.:2020-06-27 00:00:00
##  Median :2020-07-16 05:00:00   Median :2020-07-15 14:02:39
##  Mean   :2020-07-22 05:32:43   Mean   :2020-07-18 20:00:47
##  3rd Qu.:2020-08-11 05:00:00   3rd Qu.:2020-08-06 20:51:39
##  Max.   :2020-10-14 05:00:00   Max.   :2020-10-14 23:11:03
##
##    ChartDate                      month_case1     month_EventDate
##  Min.   :2020-03-13 05:00:00   July     :18483   July     :17522
##  1st Qu.:2020-06-30 05:00:00   June     : 8795   June     :10350
##  Median :2020-07-16 05:00:00   August   : 6931   August   : 6618
##  Mean   :2020-07-22 05:32:43   September: 4197   September: 4144
##  3rd Qu.:2020-08-11 05:00:00   October  : 2349   October  : 1746
##  Max.   :2020-10-14 05:00:00   April    : 1042   May      :  859
##                                (Other)  :  981   (Other)  : 1539
##   month_ChartDate     ObjectId
##  July     :18483   45     :    1
##  June     : 8795   71     :    1
##  August   : 6931   127    :    1
##  September: 4197   128    :    1
##  October  : 2349   129    :    1
##  April    : 1042   130    :    1
##  (Other)  :  981   (Other):42772
```

```
#ANALYSIS
#FREQUENCIES
#NO.OF PEOPLE INFECTED IN ORANGE COUNTY.
summary(Case_)
```

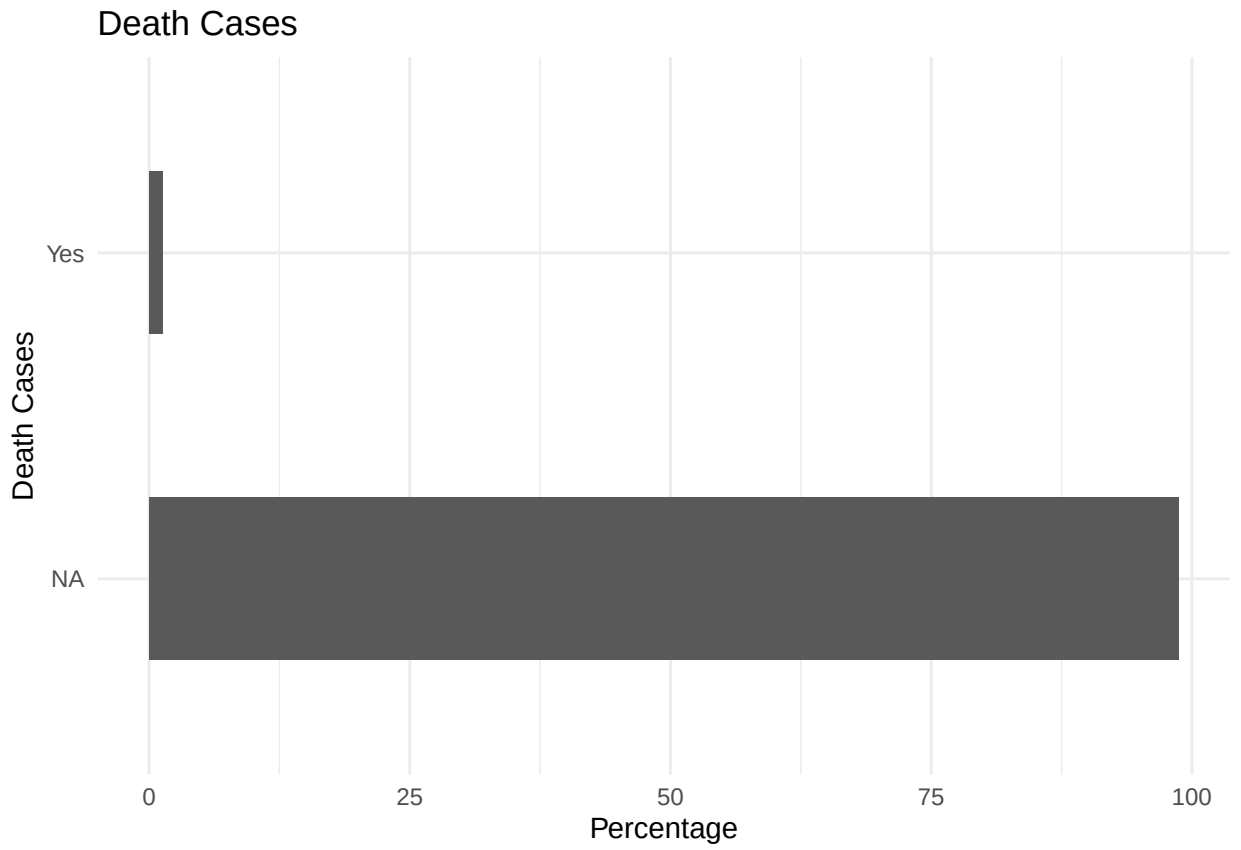
```
##   Yes
## 42778
```

```
#We had 42778 cases of Covid_19 infections from the Orange County in Florida.


#No. of infections that resulted to death
plot1 <- ggplot(Florida_COVID19_cases, aes(x=Died)) + ggtitle("Death Cases") +
  xlab("Death Cases") +
       geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + theme_minimal()

summary(Died)
```

```
##    NA   Yes
```

```
## 42244   534
```

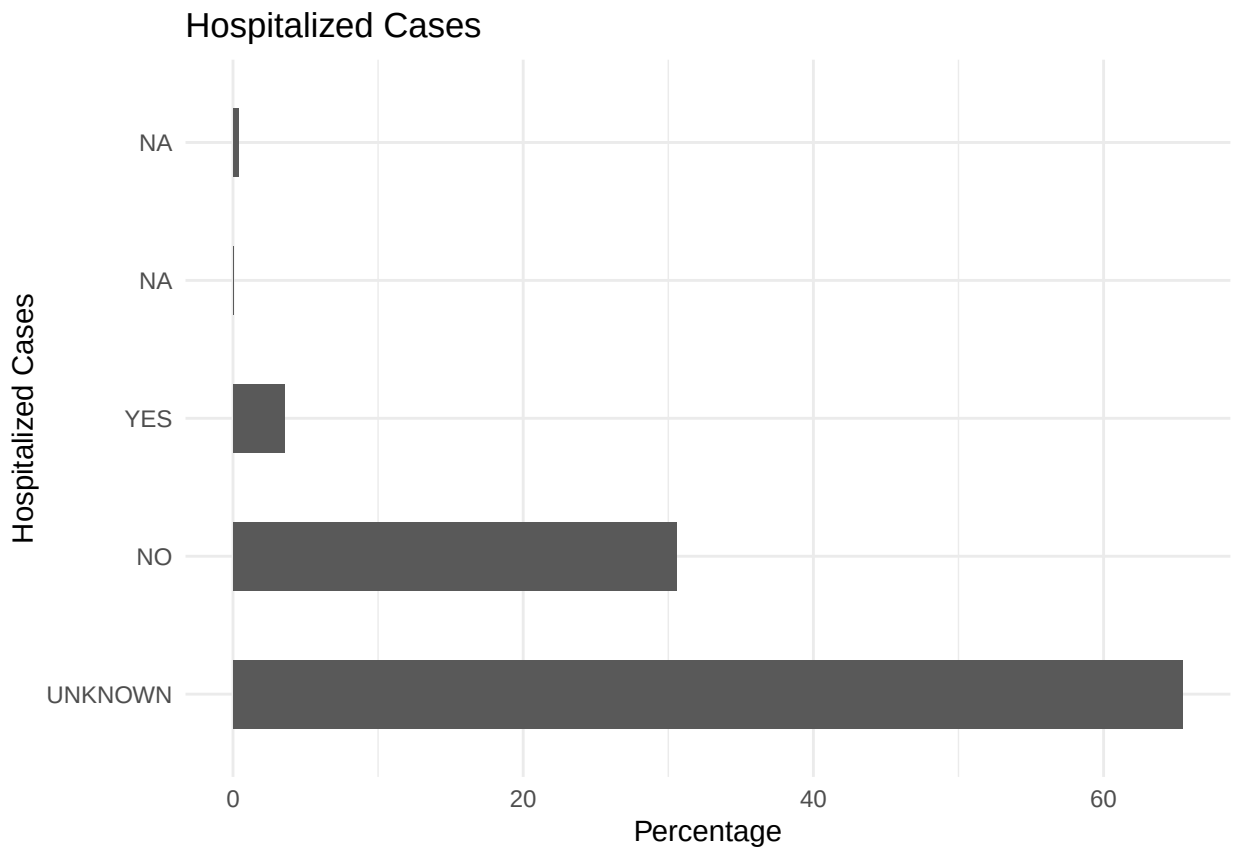
```
#Out of the 42778 people observed, 534 cases resulted to death.
plot1
```




```
#No.of hospitalized people
plot2 <- ggplot(Florida_COVID19_cases, aes(x=Hospitalized)) + ggtitle("Hospitalized Cases") +
  xlab("Hospitalized Cases") +
       geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip() + theme_minimal()

summary(Hospitalized)
```

```
## UNKNOWN      NO     YES      NA    NA's
##   28005   13081    1519       2     171
```

```
plot2
```




```
#Summary statistics for the age
summary(Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   25.00   36.00   38.99   51.00  106.00      49
```

```
hist(Age)
```


```
#The histogram of age shows the distribution of Age of people infected with Covid_19
#From the histogram the distribution is skewed to the right, implying that majority of
#....people who were infected were of Age.
#It is also right to conclude that there were few young people infected as compared to
#....the old people


mean(Age, na.rm = TRUE)
```

```
## [1] 38.99071
```

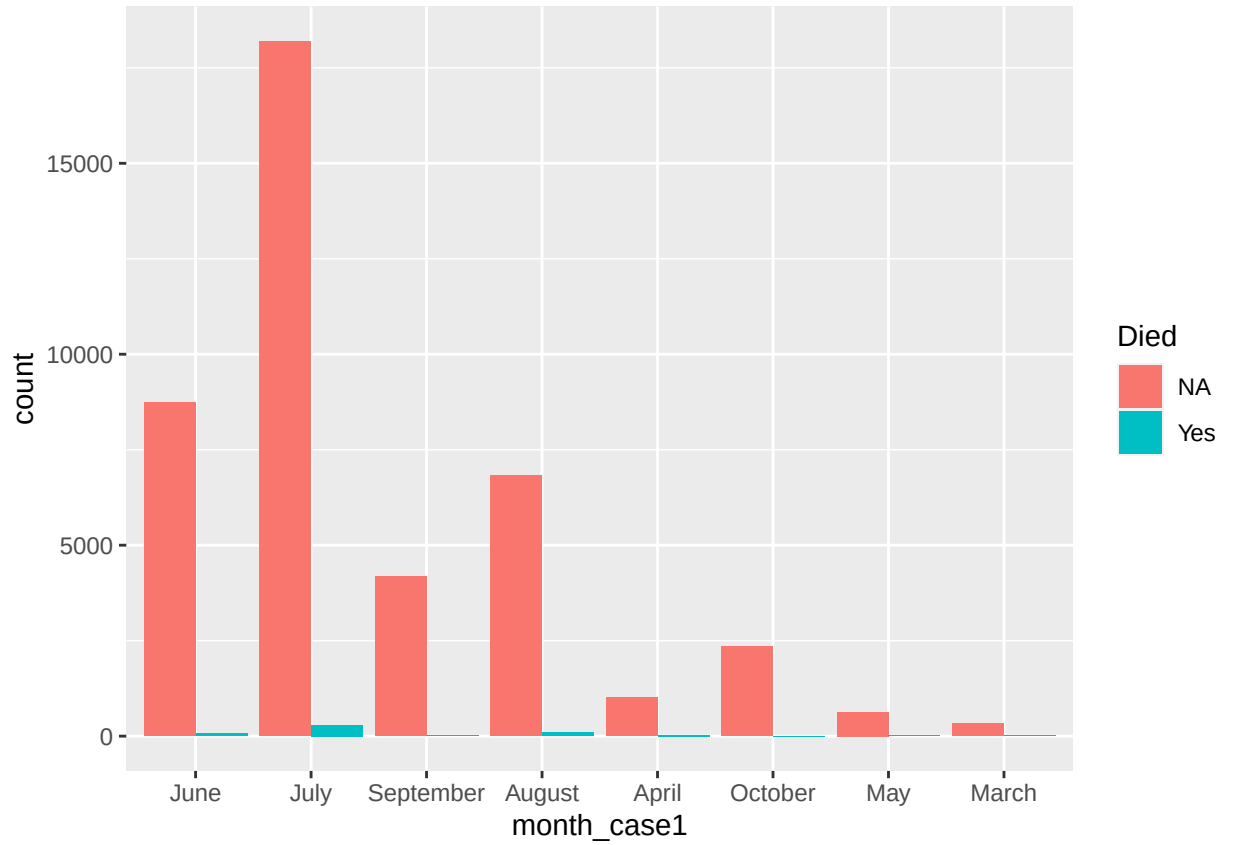
```
#As observed, the average age of people infected with Covid_19 was 38.99 years
#....approximately 39 years.


sd(Age, na.rm = TRUE)
```

```
## [1] 18.59037
```

```
#The standard deviation that we observe here is 18.59, which is less than the observed mean
#We therefore conclude that majority of the individuals observed were aged close to 39 years.

#Chi-Square test of Independence
chisq.test(table(Gender, Died))
```

```
## Warning in chisq.test(table(Gender, Died)): Chi-squared approximation may be
## incorrect

##
##  Pearson's Chi-squared test
##
## data:  table(Gender, Died)
## X-squared = 23.249, df = 2, p-value = 8.945e-06
```

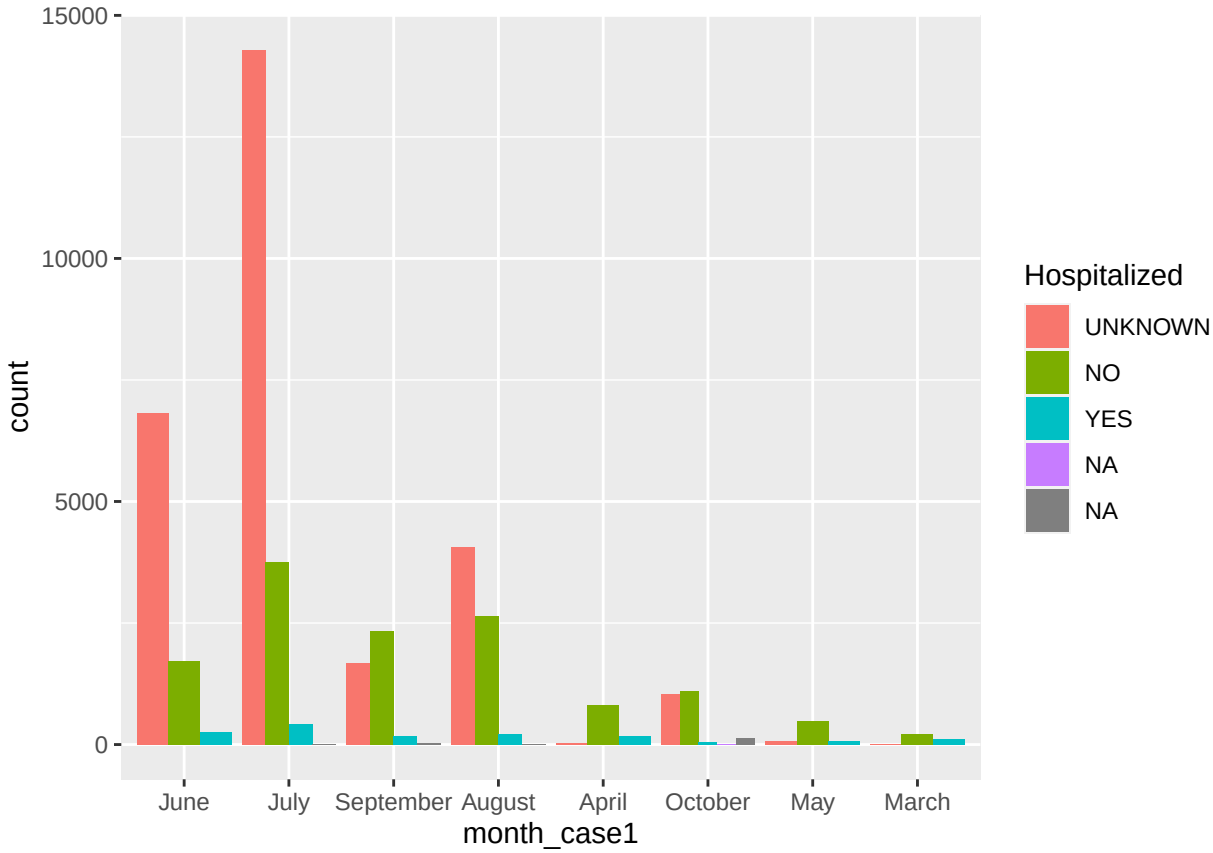
```
#We used Chi- Square test to test the significance of the relationship between the two
#....variables that is Gender and Death
#The null hypothesis in this test was that there is no relationship between Gender and Death
#The test was conducted at 0.05 level of significance. The computed p-value being less than
#....our level of significance, we proceed to reject the null hypothesis and accept the
#....alternative hypothesis.
#From the Chi-Square test, we obtain a p-value that is less than our level of significance
#....i.e. 0.05
#This means that there is statistically sufficient evidence to conclude that there is a
#....relationship between Gender and Death
#In other words, its true that to some extent, Gender Influenced Death


#A comparison of the month that had most deaths
plot3 = ggplot(Florida_COVID19_cases, aes(month_case1, ..count..)) + geom_bar(aes(fill = Died),
                                                                               position = "dodge")
plot3
```



```r
#From this graph, its clear that majority of deaths happened during the months of July,
#...June and August.
#March was among the three months that recorded the lowest number of deaths.

#The table below shows the number of deaths as observed during each month.
table(month_case1, Died)
```

```
##             Died
## month_case1    NA   Yes
##    June       8729    66
##    July      18193   290
##    September  4173    24
##    August     6833    98
##    April      1015    27
##    October    2347     2
##    May         629     9
##    March       325    18
```

```r
#A comparison of the month that had most hospitalization
plot4 = ggplot(Florida_COVID19_cases, aes(month_case1, ..count..)) + geom_bar(aes(fill = Hospitalized),
                                                                               position = "dodge")
plot4
```



```r
#From this barplot it is evident that most of the people infected were hospitalized in july

#The table below shows the number of people hospitalized for each month
table(month_case1, Hospitalized)
```

```
##            Hospitalized
## month_case1 UNKNOWN    NO   YES    NA
##   June         6812  1719   264     0
##   July        14293  3762   425     0
##   September    1672  2333   168     0
##   August       4062  2641   226     0
##   April          37   823   182     0
##   October      1041  1107    57     2
##   May            76   482    80     0
##   March          12   214   117     0
```

```r
#March was the third-last month in terms of people who were hospitalized
#July, June, August and April were the months that recorded the highest number of Hospitalizations




#Logistic Regression
#Predicting Death
glm(Died ~ Age + Gender + EDvisit + EventDate, data = Florida_COVID19_cases, family = binomial)
```

```
##
## Call:  glm(formula = Died ~ Age + Gender + EDvisit + EventDate, family = binomial,
##     data = Florida_COVID19_cases)
##
## Coefficients:
##   (Intercept)            Age     GenderMale  GenderUnknown      EDvisitNO
##     6.631e+01      1.012e-01      6.944e-01     -1.144e+01     -6.619e-01
##    EDvisitYES      EDvisitNA      EventDate
##     1.729e+00     -1.270e+01     -4.838e-08
##
## Degrees of Freedom: 41614 Total (i.e. Null);  41607 Residual
##   (1163 observations deleted due to missingness)
## Null Deviance:       5652
## Residual Deviance: 3470  AIC: 3486
```

```r
summary(glm(Died ~ Age + Gender + EDvisit + EventDate, data = Florida_COVID19_cases, family = binomial))
```

```
##
## Call:
## glm(formula = Died ~ Age + Gender + EDvisit + EventDate, family = binomial,
##     data = Florida_COVID19_cases)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1035  -0.0947  -0.0428  -0.0233   3.7837
##
```

```
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    6.631e+01  2.016e+01   3.289 0.001004 **
## Age            1.012e-01  3.259e-03  31.042  < 2e-16 ***
## GenderMale     6.944e-01  1.001e-01   6.935 4.07e-12 ***
## GenderUnknown -1.144e+01  2.067e+02  -0.055 0.955868
## EDvisitNO     -6.619e-01  1.725e-01  -3.838 0.000124 ***
## EDvisitYES     1.729e+00  1.086e-01  15.922  < 2e-16 ***
## EDvisitNA     -1.270e+01  9.853e+02  -0.013 0.989719
## EventDate     -4.838e-08  1.266e-08  -3.823 0.000132 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5652.2  on 41614  degrees of freedom
## Residual deviance: 3469.7  on 41607  degrees of freedom
##   (1163 observations deleted due to missingness)
## AIC: 3485.7
##
## Number of Fisher Scoring iterations: 16
```

```
#From the logistic Regression, its evident that Age, Gender and EventDate were the best predictors of
#.....whether a patient would die from the infection
#Age and EDvisit-YES had the lowest p-values suggesting a strong association of the Age of patient
#....with the probability of having died. The same also applies to EDvisit-YES
#The positive coefficient for age, indicates that a unit increase in age will increase the log odds by
#...0.1012 or in other words, all other predictors being constant, Aging patients are more likely
#...to die. Also from the output, being male increases the chance of dying by 0.6944
#Also those who did not visit ED their chance of dying was reduced by 0.6619 while a visit to ED
#....increased thechance of dying by 1.729.

#Predicting Hospitalization
glm(Hospitalized ~ Age + Gender + EDvisit + EventDate, data = Florida_COVID19_cases, family = binomial)
```

```
##
## Call:  glm(formula = Hospitalized ~ Age + Gender + EDvisit + EventDate,
##     family = binomial, data = Florida_COVID19_cases)
##
## Coefficients:
##   (Intercept)            Age     GenderMale  GenderUnknown      EDvisitNO
##     2.089e+02      2.304e-02     -1.733e-01     -1.550e+00      9.612e+00
##    EDvisitYES      EDvisitNA      EventDate
##     8.084e+00      2.670e+00     -1.341e-07
##
## Degrees of Freedom: 41614 Total (i.e. Null);  41607 Residual
##   (1163 observations deleted due to missingness)
## Null Deviance:       53840
## Residual Deviance: 6052  AIC: 6068
```

```
summary(glm(Hospitalized ~ Age + Gender + EDvisit + EventDate, data = Florida_COVID19_cases,
            family = binomial))
```

```
##
```

```
## Call:
## glm(formula = Hospitalized ~ Age + Gender + EDvisit + EventDate,
##     family = binomial, data = Florida_COVID19_cases)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8809  -0.1905  -0.1522   0.0843   3.2914
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.089e+02  2.450e+01   8.526  < 2e-16 ***
## Age            2.304e-02  2.123e-03  10.851  < 2e-16 ***
## GenderMale    -1.733e-01  8.249e-02  -2.101 0.035619 *
## GenderUnknown -1.550e+00  8.495e-01  -1.824 0.068113 .
## EDvisitNO      9.612e+00  1.462e-01  65.728  < 2e-16 ***
## EDvisitYES     8.084e+00  1.605e-01  50.371  < 2e-16 ***
## EDvisitNA      2.670e+00  7.866e-01   3.394 0.000689 ***
## EventDate     -1.341e-07  1.537e-08  -8.721  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53837.0  on 41614  degrees of freedom
## Residual deviance:  6052.4  on 41607  degrees of freedom
##   (1163 observations deleted due to missingness)
## AIC: 6068.4
##
## Number of Fisher Scoring iterations: 8
```

```
#The output indicates that Age, EventDate and EDvisit(a visit to emergency department) were the best
#....predictors of whether a patient would be hospitalized.
#The predictor Age has a positive coefficient implying that a unit increase in age will increase the
#...likelihood of being hospitalized by 0.0204. Suggesting that older patients were more likely to
#....be hospitalized.
#The GenderMale is a dummy variable, and from the output it suggests that being male reduced the
#...chances of a patient being hospitalized.
#Additionally, those who did not visit an ED had higher chances of being hospitalized than those
#....who did.
#As days progressed from march to october, less and lesser people got hospitalized i.e. the
#...chances of being hospitalized reduced with advancement in time.
```