CKME 136 Data Description

Dataset

```
data <- read.csv(file="https://data.ontario.ca/dataset/f4112442-bdc8-45d2-be3c-12efae72fb27/resource/455fd63b-603d-4608-8216-7d8647f43350/download/conposcovidloc.csv",header=T,sep=",")

#summary statistics of dataset
summary(data)
```

```
##      Row_ID      Accurate_Episode_Date   Age_Group        Client_Gender
##  Min.   :    1   2020-04-15:  691      50s    :5074   FEMALE     :16813
##  1st Qu.: 7716   2020-04-13:  679      40s    :4398   MALE       :13793
##  Median :15430   2020-04-16:  649      20s    :4300   OTHER      :    8
##  Mean   :15430   2020-05-29:  634      30s    :4157   TRANSGENDER:    5
##  3rd Qu.:23145   2020-04-17:  627      60s    :3612   UNKNOWN    :  241
##  Max.   :30860   2020-04-14:  616      80s    :3288
##                  (Other)   :26964      (Other):6031
##                   Case_AcquisitionInfo         Outcome1
##  Contact of a confirmed case:11419     Fatal       : 2450
##  Information pending        : 5913     Not Resolved: 3918
##  Neither                    :11814     Resolved    :24492
##  Travel-Related             : 1714
##
##
##
##  Outbreak_Related                              Reporting_PHU
##     :18558        Toronto Public Health                :11554
##  Yes:12302        Peel Public Health                   : 5010
##                   York Region Public Health Services: 2629
##                   Ottawa Public Health                 : 2001
##                   Durham Region Health Department      : 1582
##                   Region of Waterloo, Public Health    : 1154
##                   (Other)                              : 6930
##                     Reporting_PHU_Address   Reporting_PHU_City
##  277 Victoria Street, 5th Floor:11554     Toronto    :11554
##  7120 Hurontario Street        : 5010     Mississauga: 5010
##  17250 Yonge Street            : 2629     Newmarket  : 2629
##  100 Constellation Drive       : 2001     Ottawa     : 2001
##  605 Rossland Road East        : 1582     Whitby     : 1582
##  99 Regina Street South        : 1154     Waterloo   : 1154
##  (Other)                       : 6930     (Other)    : 6930
##  Reporting_PHU_Postal_Code
##  M5B 1W2:11554
##  L5W 1N4: 5010
##  L3Y 6Z1: 2629
```

```
##  K2G 6J8: 2001
##  L1N 0B2: 1582
##  N2J 4V3: 1154
##  (Other): 6930
##                                                    Reporting_PHU_Website
##  www.toronto.ca/community-people/health-wellness-care/        :11554
##  www.peelregion.ca/health/                                    : 5010
##  www.york.ca/wps/portal/yorkhome/health/                      : 2629
##  www.ottawapublichealth.ca                                    : 2001
##  www.durham.ca/en/health-and-wellness/health-and-wellness.aspx: 1582
##  www.regionofwaterloo.ca                                      : 1154
##  (Other)                                                      : 6930
##  Reporting_PHU_Latitude Reporting_PHU_Longitude
##  Min.   :42.31          Min.   :-94.49
##  1st Qu.:43.65          1st Qu.:-79.71
##  Median :43.66          Median :-79.38
##  Mean   :43.77          Mean   :-79.44
##  3rd Qu.:43.90          3rd Qu.:-79.38
##  Max.   :49.77          Max.   :-74.74
##
```

```
str(data)
```

```
## 'data.frame':    30860 obs. of  14 variables:
##  $ Row_ID                  : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Accurate_Episode_Date   : Factor w/ 123 levels "12:00:00 AM",..: 4 3 5
7 12 15 19 20 15 19 ...
##  $ Age_Group               : Factor w/ 10 levels "<20","20s","30s",..: 5
5 2 2 6 6 5 3 8 5 ...
##  $ Client_Gender           : Factor w/ 5 levels "FEMALE","MALE",..: 1 2 1
1 1 2 2 1 2 1 ...
##  $ Case_AcquisitionInfo    : Factor w/ 4 levels "Contact of a confirmed
case",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Outcome1                : Factor w/ 3 levels "Fatal","Not
Resolved",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Outbreak_Related        : Factor w/ 2 levels "","Yes": 1 1 1 1 1 1 1 1
1 1 ...
##  $ Reporting_PHU           : Factor w/ 34 levels "Algoma Public Health
Unit",..: 31 31 16 31 31 31 31 34 31 4 ...
##  $ Reporting_PHU_Address   : Factor w/ 34 levels "100 Constellation
Drive",..: 23 23 28 23 23 23 23 14 23 29 ...
##  $ Reporting_PHU_City      : Factor w/ 34 levels
"Barrie","Belleville",..: 31 31 11 31 31 31 31 14 31 33 ...
##  $ Reporting_PHU_Postal_Code: Factor w/ 34 levels "K2G 6J8","K6J 5T1",..:
16 16 24 16 16 16 16 11 16 9 ...
##  $ Reporting_PHU_Website   : Factor w/ 34 levels
"www.algomapublichealth.com",..: 31 31 8 31 31 31 31 34 31 4 ...
##  $ Reporting_PHU_Latitude  : num  43.7 43.7 43 43.7 43.7 ...
##  $ Reporting_PHU_Longitude : num  -79.4 -79.4 -81.3 -79.4 -79.4 ...
```

Approach

Step 1: Data Preparation

```r
#Find missing values in Outcome
sum(is.na(data$Outcome1) == TRUE)
```

```
## [1] 0
```

```r
length(data$Outcome1)
```

```
## [1] 30860
```

```r
#Find missing values in Outbreak_Related
sum(is.na(data$Outbreak_Related) == TRUE)
```

```
## [1] 0
```

```r
length(data$Outbreak_Related)
```

```
## [1] 30860
```

```r
#Find missing values in Age_Group
sum(is.na(data$Age_Group) == TRUE)
```

```
## [1] 0
```

```r
length(data$Age_Group)
```

```
## [1] 30860
```

```r
#Find missing values in Client_Gender
sum(is.na(data$Client_Gender) == TRUE)
```

```
## [1] 0
```

```r
length(data$Client_Gender)
```

```
## [1] 30860
```

```r
#Remove all remaining records with missing values
dataclean <- na.omit(data)
nrow(dataclean)
```

```
## [1] 30860
```

```r
#There is no missing data as all rows remain in the dataframe,
```