EDA_Covid_19_Prevention_National_04

Group K

2025-09-09

```
library(readr)
library(ggplot2)
library(dplyr)
library(here)

getwd()
```

```
## [1] "C:/Users/Caitlin/Documents/GitHub/BIN381-Project/scripts"
```

```
list.files("data/raw_data")
```

```
## character(0)
```

Load data and show the structure

```
cop_df <- read_csv(here("raw_data", "covid-19-prevention_national_zaf.csv"))
```

```
## Rows: 35 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): ISO3, DataId, Indicator, Value, Precision, DHS_CountryCode, Countr...
## dbl  (8): IndicatorOrder, CharacteristicId, CharacteristicOrder, IsTotal, Is...
## lgl  (4): RegionId, CILow, CIHigh, LevelRank
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

#Dislay the first and last 10 rows

```
head(cop_df, 5)
```

```
## # A tibble: 5 x 29
##   ISO3   DataId Indicator Value Precision DHS_CountryCode CountryName SurveyYear
##   <chr>  <chr>  <chr>     <chr> <chr>     <chr>           <chr>       <chr>
## 1 #coun~ #meta~ #indicat~ #ind~ #indicat~ <NA>            #country+n~ #date+year
## 2 ZAF    795844 Populati~ 83.5  1         ZA              South Afri~ 1998
```

```
## 3 ZAF    795750 Populati~ 36    1         ZA              South Afri~ 1998
## 4 ZAF    795755 Populati~ 23.1  1         ZA              South Afri~ 1998
## 5 ZAF    795740 Populati~ 19.3  1         ZA              South Afri~ 1998
## # i 21 more variables: SurveyId <chr>, IndicatorId <chr>, IndicatorOrder <dbl>,
## #   IndicatorType <chr>, CharacteristicId <dbl>, CharacteristicOrder <dbl>,
## #   CharacteristicCategory <chr>, CharacteristicLabel <chr>,
## #   ByVariableId <chr>, ByVariableLabel <chr>, IsTotal <dbl>,
## #   IsPreferred <dbl>, SDRID <chr>, RegionId <lgl>, SurveyYearLabel <dbl>,
## #   SurveyType <chr>, DenominatorWeighted <dbl>, DenominatorUnweighted <dbl>,
## #   CILow <lgl>, CIHigh <lgl>, LevelRank <lgl>
```

#Dimenison of the data

```
dim(cop_df)
```

```
## [1] 35 29
```

#Provide a summary of the datset

```
summary(cop_df)
```

```
##      ISO3               DataId           Indicator            Value
##  Length:35          Length:35          Length:35          Length:35
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Precision         DHS_CountryCode    CountryName         SurveyYear
##  Length:35          Length:35          Length:35          Length:35
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    SurveyId          IndicatorId        IndicatorOrder      IndicatorType
##  Length:35          Length:35          Min.   :250162010   Length:35
##  Class :character   Class :character   1st Qu.:250162190   Class :character
##  Mode  :character   Mode  :character   Median :250252010   Mode  :character
##                                        Mean   :252040162
##                                        3rd Qu.:250292085
##                                        Max.   :260831120
##                                        NA's   :1
##  CharacteristicId CharacteristicOrder CharacteristicCategory
##  Min.   :1000     Min.   :0           Length:35
##  1st Qu.:1000     1st Qu.:0           Class :character
##  Median :1000     Median :0           Mode  :character
##  Mean   :1000     Mean   :0
##  3rd Qu.:1000     3rd Qu.:0
##  Max.   :1000     Max.   :0
```

```
##  NA's   :1        NA's   :1
##  CharacteristicLabel ByVariableId       ByVariableLabel       IsTotal
##  Length:35           Length:35          Length:35          Min.   :1
##  Class :character    Class :character   Class :character   1st Qu.:1
##  Mode  :character    Mode  :character   Mode  :character   Median :1
##                                                            Mean   :1
##                                                            3rd Qu.:1
##                                                            Max.   :1
##                                                            NA's   :1
##   IsPreferred    SDRID           RegionId       SurveyYearLabel
##  Min.   :1    Length:35          Mode:logical   Min.   :1998
##  1st Qu.:1    Class :character   NA's:35        1st Qu.:1998
##  Median :1    Mode  :character                  Median :2016
##  Mean   :1                                      Mean   :2009
##  3rd Qu.:1                                      3rd Qu.:2016
##  Max.   :1                                      Max.   :2016
##  NA's   :1                                      NA's   :1
##   SurveyType        DenominatorWeighted DenominatorUnweighted  CILow
##  Length:35          Min.   :11066       Min.   :11066         Mode:logical
##  Class :character   1st Qu.:37205       1st Qu.:37925         NA's:35
##  Mode  :character   Median :37205       Median :37925
##                     Mean   :38815       Mean   :39353
##                     3rd Qu.:52007       3rd Qu.:52465
##                     Max.   :52007       Max.   :52465
##                     NA's   :3           NA's   :3
##   CIHigh        LevelRank
##  Mode:logical   Mode:logical
##  NA's:35        NA's:35
##
##
##
##
##
```

#Convert Value to a numeric

```
cop_df$Value <- as.numeric(cop_df$Value)
```

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

#Find duplicated values

```
sum(duplicated(cop_df))
```

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

#Inspect columns for missing/empty data as values and pecentages respectively

```
colSums(is.na(cop_df))
```

```
##                   ISO3                 DataId              Indicator
```

```
##                     0                     0                     0
##                 Value             Precision       DHS_CountryCode
##                     1                     0                     1
##           CountryName            SurveyYear              SurveyId
##                     0                     0                     0
##           IndicatorId        IndicatorOrder         IndicatorType
##                     0                     1                     1
##       CharacteristicId   CharacteristicOrder CharacteristicCategory
##                     1                     1                     1
##    CharacteristicLabel          ByVariableId       ByVariableLabel
##                     1                     0                    34
##               IsTotal           IsPreferred                 SDRID
##                     1                     1                     1
##              RegionId       SurveyYearLabel            SurveyType
##                    35                     1                     1
##    DenominatorWeighted DenominatorUnweighted                 CILow
##                     3                     3                    35
##                CIHigh             LevelRank
##                    35                    35
```

```
round(colMeans(is.na(cop_df)) * 100, 2)
```

```
##                  ISO3                DataId             Indicator
##                  0.00                  0.00                  0.00
##                 Value             Precision       DHS_CountryCode
##                  2.86                  0.00                  2.86
##           CountryName            SurveyYear              SurveyId
##                  0.00                  0.00                  0.00
##           IndicatorId        IndicatorOrder         IndicatorType
##                  0.00                  2.86                  2.86
##       CharacteristicId   CharacteristicOrder CharacteristicCategory
##                  2.86                  2.86                  2.86
##    CharacteristicLabel          ByVariableId       ByVariableLabel
##                  2.86                  0.00                 97.14
##               IsTotal           IsPreferred                 SDRID
##                  2.86                  2.86                  2.86
##              RegionId       SurveyYearLabel            SurveyType
##                100.00                  2.86                  2.86
##    DenominatorWeighted DenominatorUnweighted                 CILow
##                  8.57                  8.57                100.00
##                CIHigh             LevelRank
##                100.00                100.00
```

Check for unique values

```
sapply(cop_df, function(x) length(unique(x)))
```

```
##                  ISO3                DataId             Indicator
##                     2                    35                    21
##                 Value             Precision       DHS_CountryCode
```

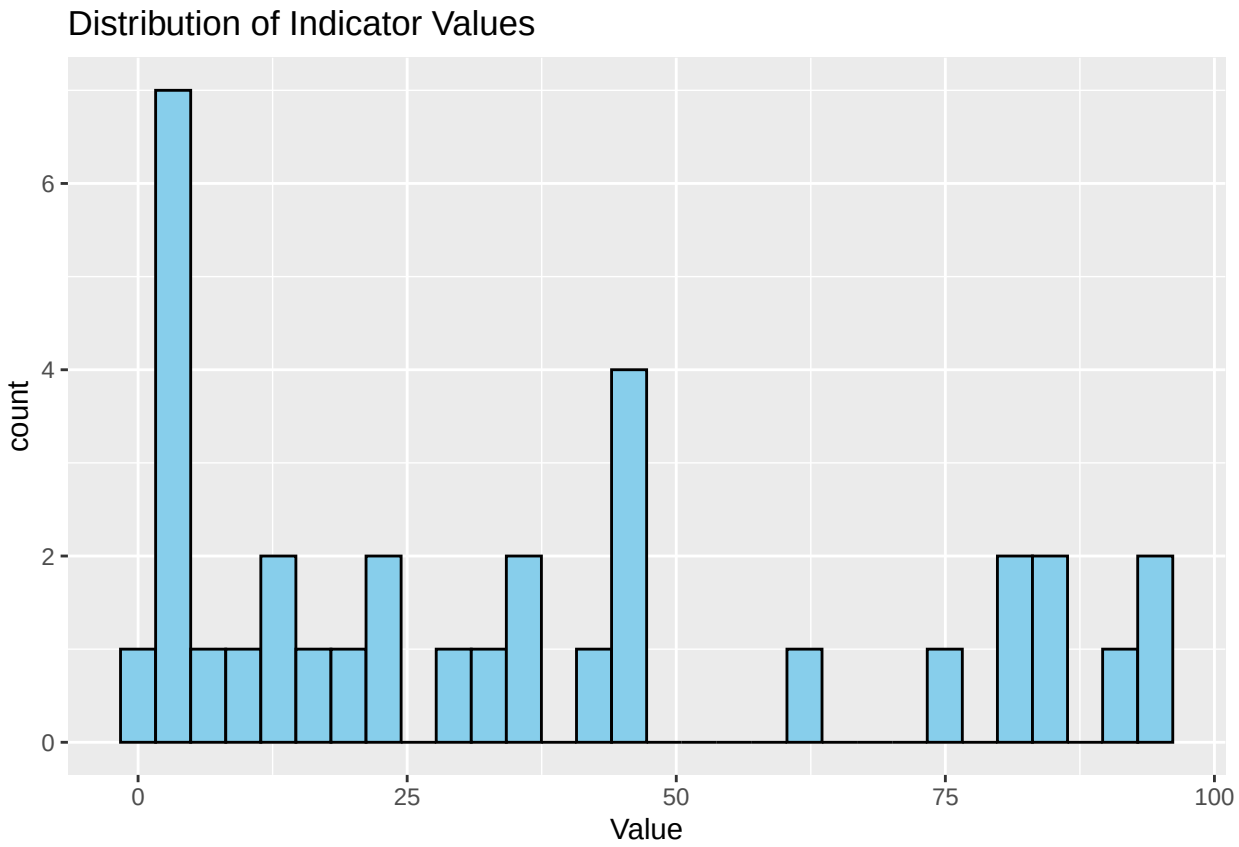
```
##                     35                      3                      2
##            CountryName             SurveyYear               SurveyId
##                      2                      3                      3
##            IndicatorId         IndicatorOrder          IndicatorType
##                     21                     21                      2
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                      2                      2                      2
##    CharacteristicLabel           ByVariableId        ByVariableLabel
##                      2                      2                      2
##                IsTotal            IsPreferred                  SDRID
##                      2                      2                     21
##               RegionId        SurveyYearLabel             SurveyType
##                      1                      3                      2
##    DenominatorWeighted  DenominatorUnweighted                  CILow
##                      9                      9                      1
##                 CIHigh             LevelRank
##                      1                      1
```

#Plots

```
#Histogram of Values

ggplot(cop_df, aes(x = Value)) +
geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Indicator Values")
```

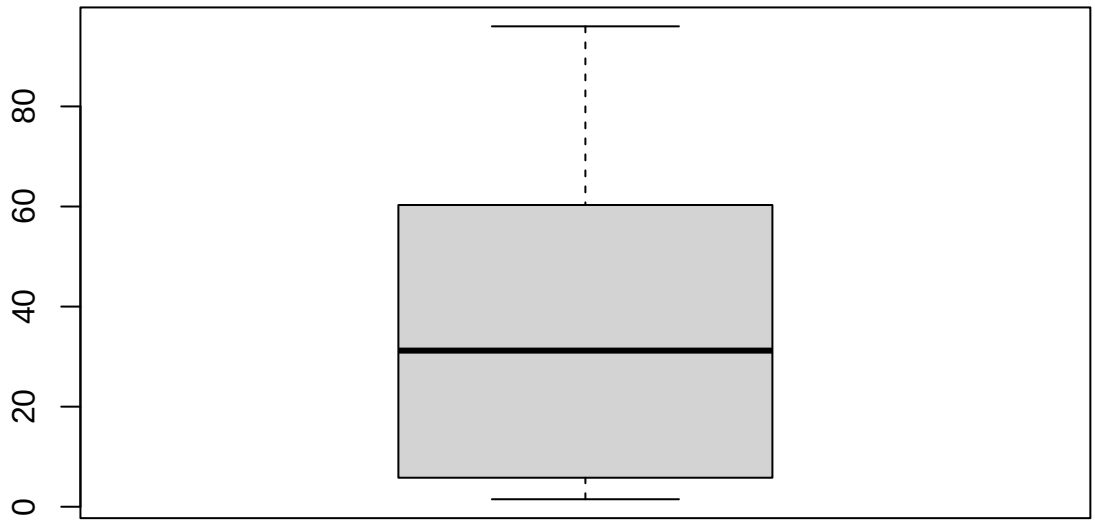
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_bin()‘).
```




```
ggsave("../outputs/visuals/covid_rate_histo.png", width = 6, height = 4)
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (‘stat_bin()‘).
```

```
#Outliers
boxplot(cop_df$Value, main = "Outlier check for Values")
```


Outlier check for Values
0
20
40
60
80