Stats\_task

## “Health-COVID-19-Cases-Source of cases” Indicator

The present report aims to focus on the summary,analysis and visualisation of the “Health-COVID-19-Cases-Status of cases” indicator that is downloaded from the Stats NZ COVID-19 data portal.

This indicator Contains useful information that helps to analyse how the covid has wide spread across different locations in NZ ,the status like deceased,recovered ,active and the number of cases.

## The meta data structure could be further improved by providing:

### 1.Proper navigation to the section to download the Indicators

(as currently when the user opens the portal page, the download section isn’t visible at the very first glance until unless the user scrolls down through the middle of the page.)

### 2.Note/hint near downloaded section

This would help the user understand that when *multiple indicators* are selected at once for download , the data relevant to these indicators will be aggregated and downloaded into one single excel file and the data related to these indicators can be distinguished based on the columns *Resource ID* and the *Source*.

### 3.Meaningful names to the columns in the Excel data so that helps the user to easily interpret and understand the context of the column data.(For instance

1. the column names like *Label1*, *Label2* in the *DATA* Sheet doesn’t give the clear insight of what the column data is about and
2. *Var1* and *Var2* columns that exist in the Meta datasheet aren’t included in the documentation “Columns included” under COVID-19 data download section.)

### 4.Proper Naming convention to the downloaded Indicator’s Excel file -

(currently when the user downloads any indicator file it is downloaded as “covid\_19\_data\_portal” which is difficult to identify when multiple files are download)

# Data Preprocessing

## Step 1: Importing the dataset

## Load libraries  
  
library(readxl) # to import excel file  
library(magrittr) # to use <%> for computing operations  
library(ggplot2) # to use ggplot function to display graph/plot  
library(dplyr) # to use functions like summarise

## imported the excel sheet and assigned to "covid\_19\_df"  
  
covid\_19\_df <- read\_excel("covid\_19\_data\_portal.xlsx",sheet = "data")  
  
#View(covid\_19\_df) # to view the imported Excel sheet

## To view the summary of the dataset  
  
summary(covid\_19\_df) # to see the summary of the dataset  
  
#str(covid\_19\_df) # To see the structure of the covid\_19\_df dataset

## Step 2: Handling the Missing Data

### Removed “NA” from the data set

## Removed "NA" from the columns using omit() function  
  
na.omit(covid\_19\_df)

### Removed redundant data columns

## Removed the unwanted columns \*Period\*,\*Label2\*,\*Multiplier\*,\*Measure\*,\*ResourceID\*, as they contain the same redundant data.  
  
within(covid\_19\_df, rm(Period,Label2,Unit,Multiplier,ResourceID,Measure))

### Renamed the columns for easy interpretation and visualisation of the data

names(covid\_19\_df)[names(covid\_19\_df) == "Geo"] <- "Locations"  
names(covid\_19\_df)[names(covid\_19\_df) == "Label1"] <- "Type\_of\_cases"  
names(covid\_19\_df)[names(covid\_19\_df) == "Value"] <- "No\_of\_cases"

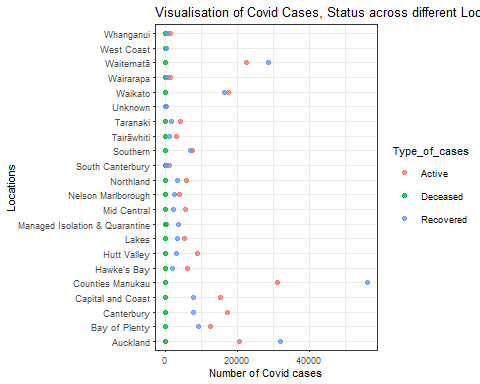
## Performed computation on the dataset

## To find the total number of Active cases in NZ  
  
covid\_19\_df %>% filter(Type\_of\_cases == "Active") %>% summarise(sum(No\_of\_cases))  
  
## to find the maximum number of cases in the whole dataset  
covid\_19\_df[covid\_19\_df$No\_of\_cases == max(covid\_19\_df$No\_of\_cases),]

# Visualisation of the dataset to analyse the number of cases and the type across different Locations in NZ.

The below visualisation depicts the analysis of the no.of cases and the status in different locations of New Zealand (where the Xaxis represents the number of cases while Y axis represents the Locations).

## plotting the graph using ggplot  
  
ggplot(covid\_19\_df,aes(y=Locations,x=No\_of\_cases,colour=Type\_of\_cases)) +  
 theme\_bw(base\_size = 8)+  
 geom\_point(alpha=0.8) +xlab('Number of Covid cases')+ylab('Locations') +ggtitle('Visualisation of Covid Cases, Status across different Locations in NZ')



# Insights from Analysis and Visualisation

The reason behind choosing this dataset is am highly interested to know about the maximum number of recovered and active cases in New Zealand and the above analysis resulted that the Counties Manukau" location is having the maximum number of Active cases ‘31206’ and recovered cases ‘56137’ when compared to different locations.