Data Mining Project 1

Project Group 10

2024-09-26

# 2. Data Exploration

## TX COVID Cases

library("tidyverse")

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library("ggplot2")  
library("ggrepel")  
library("ggcorrplot")  
library("DT")

### Data Collection

Loading the TX COVID Cases to a data frame

tx\_covid\_cases <-  
read.csv("C:\\Users\\nilee\\Documents\\AbhilashStudy\\DataMining\\COVID-19\\COVID-19\_cases\_TX.csv",header =  
TRUE)  
options(max.print=10)

### Data Cleaning

Removing the records which are Un-allocated to a county as these may not have a value while performing a county level analysis

tx\_covid\_cases <- subset(tx\_covid\_cases,county\_name !='Statewide Unallocated')   
tx\_covid\_cases

## county\_fips\_code county\_name state state\_fips\_code date  
## 371 48001 Anderson County TX 48 2020-01-22  
## confirmed\_cases deaths  
## 371 0 0  
## [ reached 'max' / getOption("max.print") -- omitted 93979 rows ]

ggplot(subset(tx\_covid\_cases,county\_fips\_code<48004) , aes(x=county\_name, y=confirmed\_cases)) +   
geom\_bar(stat="identity",width=0.5) +  
theme(plot.margin = unit(c(0.5, 0.5, 0.5, 0.5),   
 "inches")) +  
  
coord\_flip()

