Data Mining Project 1

Project Group 10

2024-09-30

# 2. Data Exploration

## Dataset 1:- 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)  
tx\_covid\_cases

## county\_fips\_code county\_name state state\_fips\_code date  
## 1 0 Statewide Unallocated TX 48 2020-01-22  
## confirmed\_cases deaths  
## 1 0 0  
## [ reached 'max' / getOption("max.print") -- omitted 94349 rows ]

### 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 ]

Summary of he COVID cases and Death by County

tx\_covid\_cases\_county = tx\_covid\_cases %>% group\_by(county\_name) %>%  
 summarise(total\_confirmed\_cases = sum(confirmed\_cases),   
 total\_deaths = sum(deaths),   
 .groups = 'drop')  
tx\_covid\_cases\_county$death\_perc=tx\_covid\_cases\_county$total\_deaths\*100/tx\_covid\_cases\_county$total\_confirmed\_cases   
tx\_covid\_cases\_county <- tx\_covid\_cases\_county[order(tx\_covid\_cases\_county$death\_perc,  
 decreasing = TRUE), ]  
tx\_covid\_cases\_county

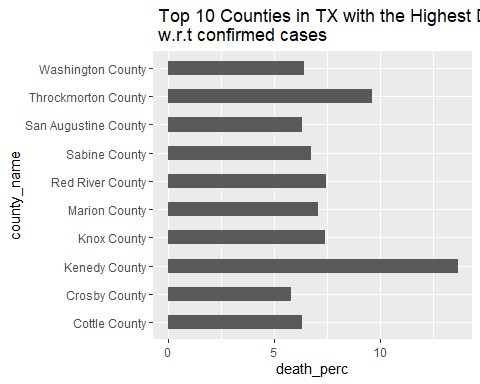
## # A tibble: 254 × 4  
## county\_name total\_confirmed\_cases total\_deaths death\_perc  
## <chr> <int> <int> <dbl>  
## 1 Kenedy County 2244 307 13.7   
## 2 Throckmorton County 3951 381 9.64  
## 3 Red River County 50991 3808 7.47  
## 4 Knox County 20479 1521 7.43  
## 5 Marion County 37648 2668 7.09  
## 6 Sabine County 29743 2018 6.78  
## 7 Washington County 182682 11751 6.43  
## 8 Cottle County 11376 722 6.35  
## 9 San Augustine County 54882 3469 6.32  
## 10 Crosby County 30712 1789 5.83  
## # ℹ 244 more rows

### Data Analysis

Bubble chart, Relation between Confirmed vs dead counts of every county

1. Top 10 which counties have more death %, FOr those, WHy fro ther data set?
2. Top 10 which counties have lesser death %, FOr those, WHy fro ther data set?
3. Based on season, WHen the confirmed deaths are higher
4. Co-relation between confirmed vs death follow the same pattern ( Heat map)

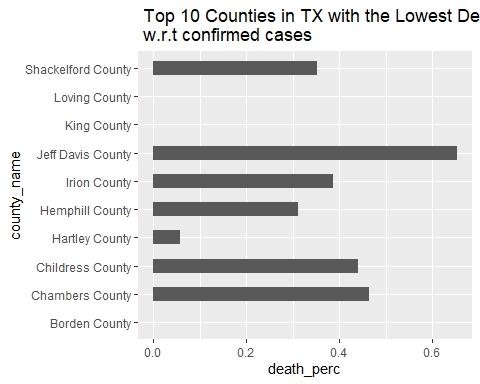
ggplot(head(tx\_covid\_cases\_county,10),  
 aes(x=county\_name, y=death\_perc)) +   
 geom\_bar(stat="identity",width=0.5) +  
 ggtitle(" Top 10 Counties in TX with the Highest Death Percentage \n w.r.t confirmed cases")+  
 coord\_flip()



head

## function (x, ...)   
## UseMethod("head")  
## <bytecode: 0x000001efde57a300>  
## <environment: namespace:utils>

ggplot(tail(tx\_covid\_cases\_county,10),  
 aes(x=county\_name, y=death\_perc)) +   
 geom\_bar(stat="identity",width=0.5) +  
 ggtitle(" Top 10 Counties in TX with the Lowest Death Percentage \n w.r.t confirmed cases")+  
 coord\_flip()



head

## function (x, ...)   
## UseMethod("head")  
## <bytecode: 0x000001efde57a300>  
## <environment: namespace:utils>

tx\_covid\_cases\_date = tx\_covid\_cases %>% group\_by(date) %>%  
 summarise(total\_confirmed\_cases = sum(confirmed\_cases),   
 total\_deaths = sum(deaths),   
 .groups = 'drop')  
tail(tx\_covid\_cases\_date,10)

## # A tibble: 10 × 3  
## date total\_confirmed\_cases total\_deaths  
## <chr> <int> <int>  
## 1 2021-01-16 2098586 31817  
## 2 2021-01-17 2115975 32024  
## 3 2021-01-18 2126601 32070  
## 4 2021-01-19 2138454 32379  
## 5 2021-01-20 2166895 32828  
## 6 2021-01-21 2191076 33269  
## 7 2021-01-22 2213428 33691  
## 8 2021-01-23 2230519 34099  
## 9 2021-01-24 2240949 34307  
## 10 2021-01-25 2248927 34378

ggplot(tx\_covid\_cases\_date, aes(x=date,y=total\_deaths)) +   
 geom\_line()

## `geom\_line()`: Each group consists of only one observation.  
## ℹ Do you need to adjust the group aesthetic?

