Data Mining Project 1

Project Group 10

2024-10-03

# 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=100)  
  
county\_census <-  
read.csv("C:\\Users\\nilee\\Documents\\AbhilashStudy\\DataMining\\COVID-19\\COVID-19\_cases\_plus\_census.csv",header =  
TRUE)  
options(max.print=100)

### 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$date <- as.Date(tx\_covid\_cases$date,format = "%Y-%m-%d")  
options(max.print=10)  
tx\_covid\_cases

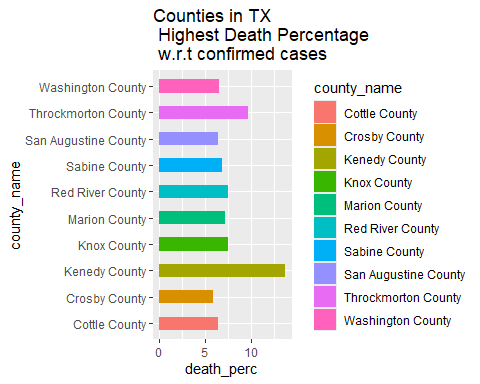
## county\_fips\_code county\_name state state\_fips\_code date  
## 371 48001 Anderson County TX 48 2020-01-22  
## confirmed\_cases deaths  
## 371 0 0  
## [ reached 'max' / getOption("max.print") -- omitted 93979 rows ]

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), ]

### Data Analysis

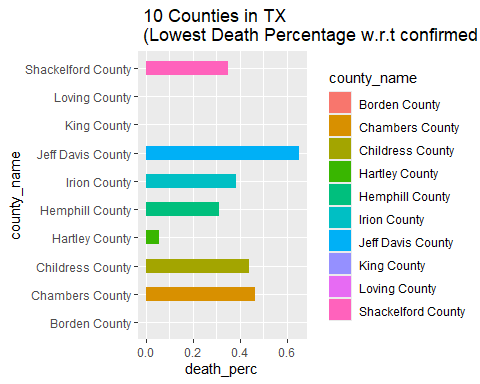
ggplot(head(tx\_covid\_cases\_county,10),  
 aes(x=county\_name, y=death\_perc,fill=county\_name)) +   
 geom\_bar(stat="identity",width=0.5) +  
 ggtitle("Counties in TX \n Highest Death Percentage \n w.r.t confirmed cases")+  
 coord\_flip()



head

## function (x, ...)   
## UseMethod("head")  
## <bytecode: 0x000001a64ab7c300>  
## <environment: namespace:utils>

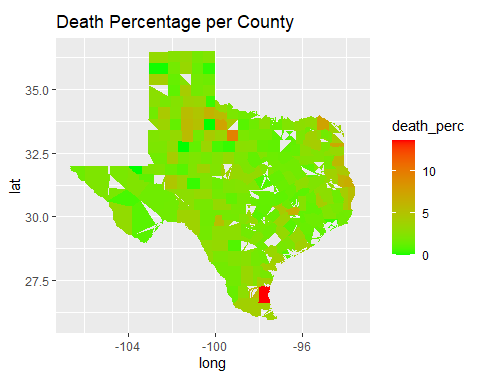
ggplot(tail(tx\_covid\_cases\_county,10),  
 aes(x=county\_name, y=death\_perc,fill=county\_name)) +   
 geom\_bar(stat="identity",width=0.5) +  
 ggtitle(" 10 Counties in TX \n (Lowest Death Percentage w.r.t confirmed cases)")+  
 coord\_flip()



head

## function (x, ...)   
## UseMethod("head")  
## <bytecode: 0x000001a64ab7c300>  
## <environment: namespace:utils>

counties <- as\_tibble(map\_data("county"))  
counties\_TX <- counties %>% dplyr::filter(region == "texas") %>% rename(c(county = subregion))  
  
tx\_covid\_cases\_county <- tx\_covid\_cases\_county %>% mutate(county = county\_name %>% str\_to\_lower() %>%   
 str\_replace('\\s+county\\s\*$', ''))  
  
counties\_w\_map<-merge(x = counties\_TX, y = tx\_covid\_cases\_county, by = "county")  
  
  
ggplot(counties\_w\_map, aes(long, lat, label = county)) +   
 geom\_polygon(aes(group = group, fill = death\_perc)) +  
 scale\_fill\_gradient(low="green", high="red") +  
 coord\_quickmap() +  
 labs(title = "Death Percentage per County")



### Filtering only TX Counties for Census data

county\_census <- subset(county\_census,state == 'TX')  
county\_census <- county\_census %>% mutate(county = county\_name %>% str\_to\_lower() %>%   
 str\_replace('\\s+county\\s\*$', ''))

### Joining the Covid data with the Census data

covid\_cases\_w\_census<-merge(x = county\_census, y = tx\_covid\_cases\_county, by = "county")

### Is there a relation between median income and death percentage of a county

ggplot(covid\_cases\_w\_census, mapping = aes(x= median\_income, y = death\_perc, label = county\_name.x)) +   
 geom\_smooth(method = lm) +  
 geom\_point(mapping = aes(size = total\_confirmed\_cases), color = "blue") +   
 geom\_text\_repel(data = subset(covid\_cases\_w\_census, death\_perc > quantile(death\_perc, .95))) +  
 labs(x = "Median Income", y = "Death Percentage", size = "Confirmed Cases")

## `geom\_smooth()` using formula = 'y ~ x'

## Warning: The following aesthetics were dropped during statistical transformation: label.  
## ℹ This can happen when ggplot fails to infer the correct grouping structure in  
## the data.  
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical  
## variable into a factor?

## Warning: ggrepel: 4 unlabeled data points (too many overlaps). Consider  
## increasing max.overlaps

