Stat 321 Final Project Statistic Analysis

Lin Xueying

2025-03-22

1. Import Data Set

```
suppressMessages(library(tidyverse))
library(readxl)


# 1. COVID case data
covid_data <- read_csv("COVID-19_Weekly_Cases_and_Deaths_by_Age__Race_Ethnicity__and_Sex_-_ARCHIVED_2025

covid_data <- covid_data %>%
  select(-`...1`)

covid_data <- covid_data %>%
 rename(
    Jurisdiction = jurisdiction,
    AgeGroup = age_group,
    Sex = sex,
    Race = race_ethnicity_combined,
    CaseSuppressed = case_count_suppressed,
    DeathSuppressed = death_count_suppressed,
    CaseRate = case_crude_rate_suppressed_per_100k,
    DeathRate = death_crude_rate_suppressed_per_100k
  ) %>%
  # Convert end_of_week as a string to a real Date.
  mutate(Date = as.Date(end_of_week, format = "%m/%d/%Y")) %>%
  # Remove raw columns end_of_week
  select(-end_of_week) %>%
  # arrange date
  arrange(Date)

#move date to first
covid_data <- covid_data %>%
select(Date, everything())

covid_data <- covid_data %>%
  mutate(across(where(is.numeric), ~replace_na(., 0)))

# 2. Initial unemployment claims
claims_data <-  read_csv("ICSA_Initial claims_cleaned.csv", show_col_types = FALSE)

#rename
```

```r
claims_data <- claims_data %>%
  rename(
    Date = observation_date
  )

claims_data <- claims_data %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%Y"))

# 3. Unemployment rate data
unemp_data <- read_csv("unemployment rate data.csv", show_col_types = FALSE)

# 4. VGT
vgt_data <- read_csv("VGT_stock_data_cleaned.csv", show_col_types = FALSE)

# 5. VHT
vht_data <- read_csv("VHT_stock_data_cleaned.csv", show_col_types = FALSE)

# state case and death
state_case_death <- read_csv("Weekly_United_States_COVID-19_Cases_and_Deaths_by_State_-_ARCHIVED_202503
                             show_col_types = FALSE)

state_case_death <- state_case_death %>%
  rename(
    DateUpdate = date_updated,
    State = state,
    StartDate = start_date,
    EndDate = end_date,
    TotalCases = tot_cases,
    NewCases = new_cases,
    TotalDeaths = tot_deaths,
    NewDeaths = new_deaths
  )
```

2. VGT modeling

```r
# Analyze with average_movement of vgt_data
vgt_clean <- vgt_data %>%
  select(Date, vgt_average = average_movement)

# Covid Monthly Data
covid_monthly <- covid_data %>%
  mutate(Month = format(Date, "%Y-%m")) %>%
  group_by(Month) %>%
  summarise(COVID_Cases = sum(CaseSuppressed, na.rm = TRUE)) %>%
  mutate(Date = as.Date(paste0(Month, "-01")))

# combine 2 data set together
vgt_combine_monthly <- left_join(vgt_clean, covid_monthly, by = "Date") %>%
  mutate(Period = case_when(
    Date < as.Date("2020-03-01") ~ "Pre_covid",
    Date <= as.Date("2023-11-01") ~ "During_covid",
     Date > as.Date("2023-11-01") ~ "Post_covid"
```

```
  )
)

# Summarize the values of prices for each period (mean, standard deviation, minimum, maximum)
vgt_summary <- vgt_combine_monthly %>%
  group_by(Period) %>%
  summarise(
    Avg_Price = mean(vgt_average, na.rm = TRUE),
    SD_Price = sd(vgt_average, na.rm = TRUE),
    Min_Price = min(vgt_average, na.rm = TRUE),
    Max_Price = max(vgt_average, na.rm = TRUE),
    Num_Months = n()
  ) %>%
  mutate(period = factor(Period, levels = c("Pre_covid", "During_covid", "Post_covid"))) %>%
  arrange(period)

print(vgt_summary)
```

```
# Difference in VGT average between periods
vgt_period_model <- lm(vgt_average ~ Period, data = vgt_combine_monthly)
summary(vgt_period_model)

# Does the number of cases during Covid affect VGT

model_covid_vgt <- lm(vgt_average ~ COVID_Cases,
                      data = vgt_combine_monthly %>%
                     filter(Period == "During_covid"))
summary(model_covid_vgt)
```

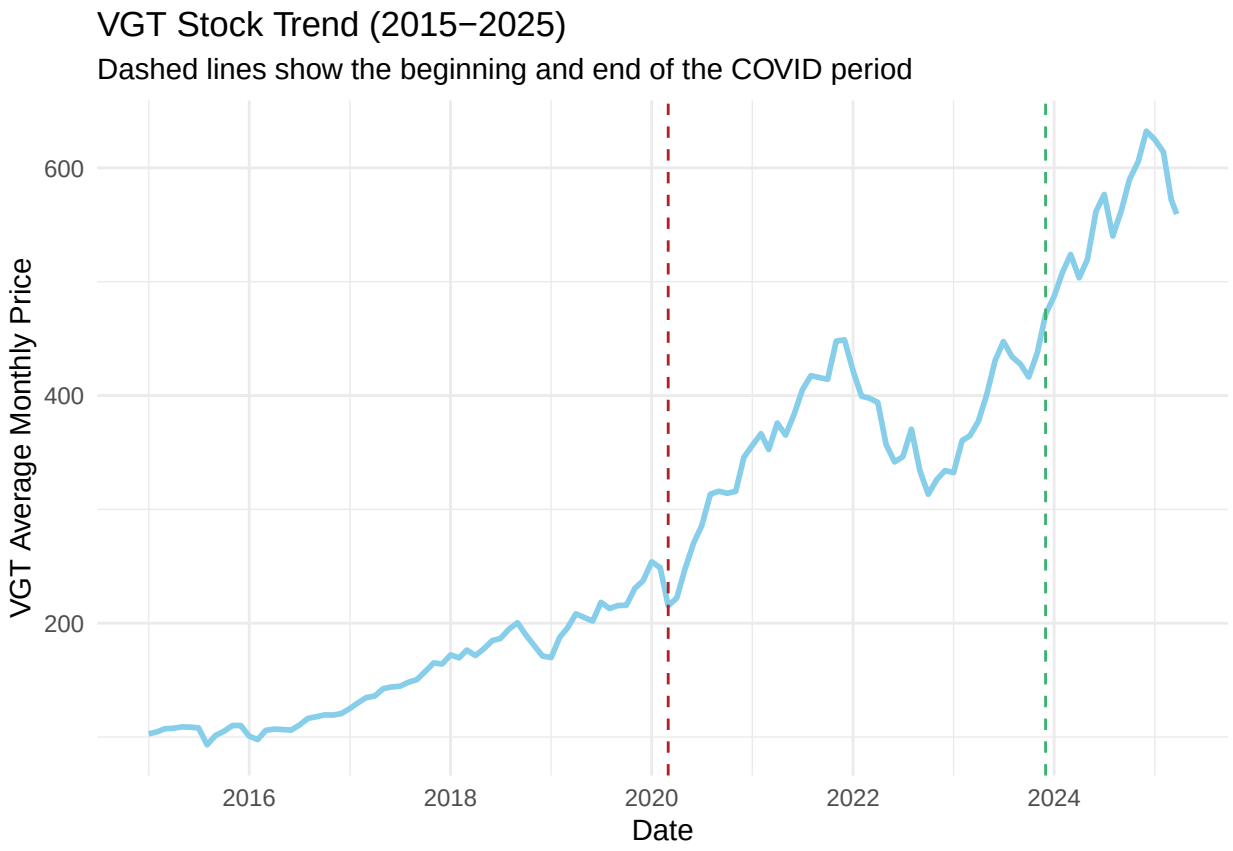
Analysis:

1. The null hypothesis (H0) was that there is no difference in the average VGT price between the three periods: pre-COVID, during COVID, and post-COVID. The alternative hypothesis (H1) was that there is a difference. To test this hypothesis, we used a linear regression model.
2. The model showed significant differences ($p < 0.001$). During the COVID time, the average price of VGT was \$363.55. During the pre-COVID time, the average price was \$153.47, a reduction of \$210.08 from the COVID time. Later, after COVID, the average price went up to \$556.04, an increase of \$192.49 from the COVID time. Both differences are statistically significant ($p < 0.001$), and therefore the null hypothesis is rejected. The adjusted R-squared was 0.892, indicating that the period of the pandemic explains about 89% of the changes in VGT prices. This show the pandemic had a strong long-term impact on tech stocks, especially in terms of overall price growth.
3. Furthermore,We examined the impact of changes in the number of COVID-19 cases on price changes of VGT during the pandemic.The result we got was statistically not significant ($p = 0.257$), and the adjusted R-squared was a mere 0.007. Therefore, we fail to reject the null hypothesis, and we conclude that short-term price movement was not affected significantly by monthly COVID-19 case numbers.

VGT Ploting

```
library(ggplot2)

#VGT vs COVID-19 Monthly Case
# During Covid data
vgt_plot_data <- vgt_combine_monthly %>%
```

```r
  filter(Period == "During_covid")

# Average Price Trend for VGT
ggplot(vgt_combine_monthly, aes(x = Date, y = vgt_average)) +
  geom_line(color = "skyblue", size = 1) +

  # Vertical lines to indicate period breakpoints
  geom_vline(xintercept = as.Date("2020-03-01"), linetype = "dashed", color = "firebrick") +
  geom_vline(xintercept = as.Date("2023-12-01"), linetype = "dashed", color = "mediumseagreen") +

  labs(
    title = "VGT Stock Trend (2015-2025)",
    subtitle = "Dashed lines show the beginning and end of the COVID period",
    x = "Date",
    y = "VGT Average Monthly Price"
  ) +
  theme_minimal()
```



3. VHT modeling

```r
# Analyze with average_movement of vht_data
vht_clean <- vht_data %>%
  select(Date, vht_average = average_movement)

# combine 2 data set together
```

```r
vht_combine_monthly <- left_join(vht_clean, covid_monthly, by = "Date") %>%
  mutate(Period = case_when(
    Date < as.Date("2020-03-01") ~ "Pre_covid",
    Date <= as.Date("2023-11-01") ~ "During_covid",
    Date > as.Date("2023-11-01") ~ "Post_covid"
  )
)

# Summarize the values of prices for each period (mean, standard deviation, minimum, maximum)
vht_summary <- vht_combine_monthly %>%
  group_by(Period) %>%
  summarise(
    Avg_Price = mean(vht_average, na.rm = TRUE),
    SD_Price = sd(vht_average, na.rm = TRUE),
    Min_Price = min(vht_average, na.rm = TRUE),
    Max_Price = max(vht_average, na.rm = TRUE),
    Num_Months = n()
  ) %>%
  mutate(period = factor(Period, levels = c("Pre_covid", "During_covid", "Post_covid"))) %>%
  arrange(period)

print(vht_summary)
```

```r
# Difference in VHT average between periods
vht_period_model <- lm(vht_average ~ Period, data = vht_combine_monthly)
summary(vht_period_model)

# Does the number of cases during Covid affect VHT
model_covid_vht <- lm(vht_average ~ COVID_Cases,
                      data = vht_combine_monthly %>% filter(Period == "During_covid"))
summary(model_covid_vht)
```

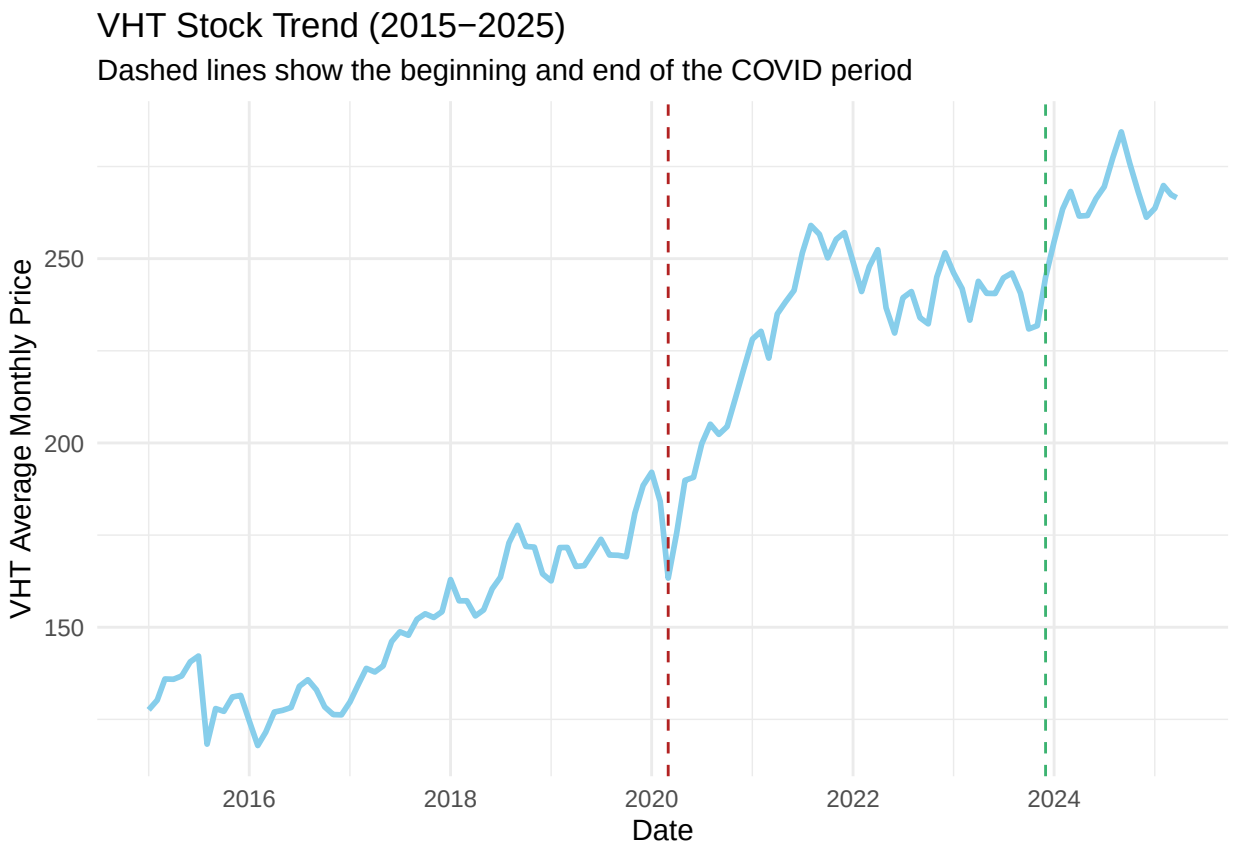
Analysis:

1. The null hypothesis (H0) was that there is no difference in the average VHT price between the three periods: pre-COVID, during COVID, and post-COVID. The alternative hypothesis (H1) was that there is a difference.To test this hypothesis, we used a linear regression model.
2. The regression results showed there were significant differences ($p < 0.001$). During the COVID period, the average price of VHT was \$231.80. Before the pandemic, the average was \$149.30, and it is \$82.50 lower than in COVID. Post-pandemic, the average price went up to \$266.18, which is \$34.38 higher than during COVID. Both of these differences are statistically significant ($p < 0.001$ for pre-COVID and $p < 0.001$ for post-COVID), and we can reject the null hypothesis. The adjusted R-squared of the model was 0.853, indicating that the period explains about 85% of the variability of VHT prices. This showed a strong impact of the pandemic on the healthcare industry's stock prices.
3. Additionally, We tested if the monthly number of COVID-19 cases would be a explanation of VHT price fluctuations during the COVID period. The result was not statistically significant ($p = 0.200$), and the adjusted R-squared was only 0.016. Therefore, we fail to reject the null hypothesis and conclude that short-term case did not have a significant effect on VHT price fluctuations during the pandemic.

VHT Ploting

```r
library(ggplot2)
```

```r
#VHT vs COVID-19 Monthly Case
# During Covid data
vht_plot_data <- vht_combine_monthly %>%
  filter(Period == "During_covid")

# Average Price Trend for VHT
ggplot(vht_combine_monthly, aes(x = Date, y = vht_average)) +
  geom_line(color = "skyblue", size = 1) +

  # Vertical lines to indicate period breakpoints
  geom_vline(xintercept = as.Date("2020-03-01"), linetype = "dashed", color = "firebrick") +
  geom_vline(xintercept = as.Date("2023-12-01"), linetype = "dashed", color = "mediumseagreen") +

  labs(
    title = "VHT Stock Trend (2015-2025)",
    subtitle = "Dashed lines show the beginning and end of the COVID period",
    x = "Date",
    y = "VHT Average Monthly Price"
  ) +
  theme_minimal()
```



4. Unemployment rate in different periods

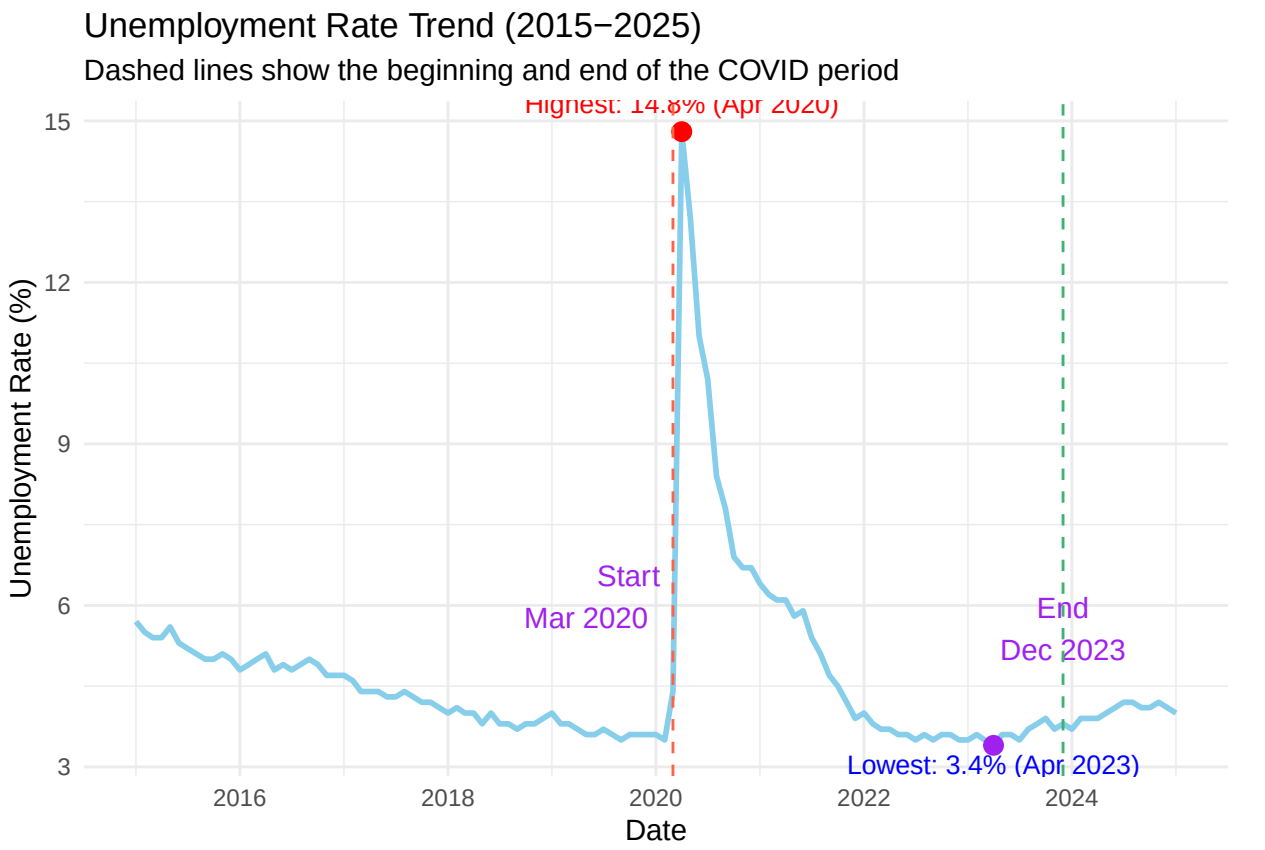
```r
# Rename and add time period
unemp_clean <- unemp_data %>%
```

```
  rename(Date = observation_date, unemployment_rate = UNRATE) %>%
  mutate(
    Date = as.Date(Date, format = "%m/%d/%Y"),
    period = case_when(
      Date < as.Date("2020-03-01") ~ "pre_covid",
      Date <= as.Date("2023-11-30") ~ "during_covid",
      Date > as.Date("2023-11-01") ~ "post_covid"
    )
  )
```

Unemployment rate Ploting

```
# Unemployment rate time trend
# Finding the max, min, start, end
point_start <- unemp_clean %>% filter(Date == as.Date("2020-03-01"))
point_end <- unemp_clean %>% filter(Date == as.Date("2023-12-01"))
point_highest <- unemp_clean %>% filter(unemployment_rate == max(unemployment_rate))
point_lowest <- unemp_clean %>% filter(unemployment_rate == min(unemployment_rate))

#ploting
ggplot(unemp_clean, aes(x = Date, y = unemployment_rate)) +
  geom_line(color = "skyblue", size = 1) +

  #Start
  geom_text(data = point_start,
            aes(label = paste("Start", format(Date, "%b %Y"),sep = "\n")),
            vjust = -1, hjust = 1.2, color = "purple") +

  #End
  geom_text(data = point_end,
            aes(label = paste("End", format(Date, "%b %Y"), sep = "\n")),
            vjust = -1, color = "purple") +

  #Highest
  geom_point(data = point_highest, aes(x = Date, y = unemployment_rate),
             color = "red", size = 3) + geom_text(data = point_highest,
          aes(label = paste0("Highest: ", unemployment_rate,
                             "% (", format(Date, "%b %Y"), ")")),
                  vjust = -1, color = "red", size = 3.5) +

  #Lowest
  geom_point(data = point_lowest, aes(x = Date, y = unemployment_rate),
             color = "purple", size = 3) +geom_text(data = point_lowest,
  aes(label = paste0("Lowest: ", unemployment_rate,
                     "% (", format(Date, "%b %Y"), ")")), vjust = 1.5,
                      color = "blue", size = 3.5) +

  # Vertical lines to indicate period breakpoints
  geom_vline(xintercept = as.Date("2020-03-01"), linetype = "dashed", color = "tomato") +
  geom_vline(xintercept = as.Date("2023-12-01"), linetype = "dashed", color = "mediumseagreen") +

  labs(
    title = "Unemployment Rate Trend (2015-2025)",
```

```
    subtitle = "Dashed lines show the beginning and end of the COVID period",
    x = "Date",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal()
```



5. Initial Claims for Unemployment Insurance

3 weeks with lowest and highest unemployment claims

```
top3 <- head(claims_data[order(-claims_data$ICSA), ], 3)
last3 <- tail(claims_data[order(-claims_data$ICSA), ], 3)

cat(paste0("3 weeks with highest unemployment claims:\n",
           "1. ", top3$Date[1], " - ", top3$ICSA[1], " claims\n",
           "2. ", top3$Date[2], " - ", top3$ICSA[2], " claims\n",
           "3. ", top3$Date[3], " - ", top3$ICSA[3], " claims\n\n",
           "3 weeks with lowest unemployment claims:\n",
           "1. ", last3$Date[1], " - ", last3$ICSA[1], " claims\n",
           "2. ", last3$Date[2], " - ", last3$ICSA[2], " claims\n",
           "3. ", last3$Date[3], " - ", last3$ICSA[3], " claims\n"),
    sep = "")
```

```
## 3 weeks with highest unemployment claims:
## 1. 2020-04-04 - 6137000 claims
## 2. 2020-03-28 - 5946000 claims
```

```
## 3. 2020-04-11 - 4869000 claims
##
## 3 weeks with lowest unemployment claims:
## 1. 2021-12-25 - 195000 claims
## 2. 2024-01-13 - 194000 claims
## 3. 2022-09-24 - 187000 claims
```

Compare ICSA and unemployment rate

```
# Change Date Format
# Monthly Caims data set
monthly_claims <- claims_data %>%
  mutate(Month = format(Date, "%Y-%m")) %>%
  group_by(Month) %>%
  summarise(
    TotalClaims = sum(ICSA, na.rm = TRUE),
    .groups = "drop"
  )

# Add month to unemp_clean
unemp_clean <- unemp_clean %>%
  mutate(Month = format(Date, "%Y-%m"))

# Combine unemp_clean and Monthly Caims
unemp_claims_data <- unemp_clean %>%
  select(Month, unemployment_rate, period) %>%
  left_join(monthly_claims, by = "Month")
```

Simple linear regression: unemployment_rate ~ TotalClaims

```
# Simple linear regression: unemployment_rate ~ TotalClaims
model_claims_unemp <- lm(TotalClaims ~ unemployment_rate, data = unemp_claims_data)
summary(model_claims_unemp)
```

Analysis:

1. The null hypothesis (H0) was that there is no relationship between the unemployment rate and the number of initial unemployment claims, while the alternative hypothesis (H1) was that there is a relationship. To test this hypothesis, we used a simple linear regression model with monthly total unemployment claims and the unemployment rate.
2. The regression model showed a statistically significant result ($p < 0.001$). The estimated slope was 1,044,116, show that, on average, for every 1% increase in the unemployment rate, the number of unemployment claims increases by 1,044,116 people. The intercept was (-3,285,323), but the intercept is less meaningful in this situation because 0% unemployment is unrealistic. The adjusted R-squared value was 0.717, meaning that approximately 71.7% of the variation in the number of claims could be explained by the unemployment rate. Since the p-value is less than 0.05, the null hypothesis is rejected, and one can say that there is strong evidence of a relationship between unemployment rate and unemployment claims.
3. Overall, the model shows that higher unemployment is strongly related to higher unemployment claims.

6. Covid case and death analysis

Covid case and death base on Age

```r
# add a column in covid data contain age group
covid_data <- covid_data %>%
  mutate(
    AgeGroupNum = case_when(
      AgeGroup == "0 - 4 Years"   ~ 1,
      AgeGroup == "5 - 11 Years"  ~ 2,
      AgeGroup == "12 - 15 Years" ~ 3,
      AgeGroup == "16 - 17 Years" ~ 4,
      AgeGroup == "18 - 29 Years" ~ 5,
      AgeGroup == "30 - 39 Years" ~ 6,
      AgeGroup == "40 - 49 Years" ~ 7,
      AgeGroup == "50 - 64 Years" ~ 8,
      AgeGroup == "65 - 74 Years" ~ 9,
      AgeGroup == "75+ Years"     ~ 10,
      AgeGroup == "Overall"       ~ 11
    )
  )

# age summary table (excluding Overall)
age_summary <- covid_data %>%
  filter(AgeGroup != "Overall") %>%
  group_by(AgeGroup, AgeGroupNum) %>%
  summarise(
    TotalCases = sum(CaseSuppressed, na.rm = TRUE),
    TotalDeaths = sum(DeathSuppressed, na.rm = TRUE),
    MeanCaseRate = mean(CaseRate, na.rm = TRUE),
    MeanDeathRate = mean(DeathRate, na.rm = TRUE),
    MaxDeathRate = max(DeathRate, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(AgeGroupNum)

# Total Case
# Highest / Lowest of cases
highest_cases_age <- age_summary[which.max(age_summary$TotalCases), ]

lowest_cases_age <- age_summary[which.min(age_summary$TotalCases), ]

# Mean Case Rate
highest_mean_case_age <- age_summary[which.max(age_summary$MeanCaseRate), ]

lowest_mean_case_age <- age_summary[which.min(age_summary$MeanCaseRate), ]

# Total Death
# Highest /Lowest of death
most_deaths_age <- age_summary[which.max(age_summary$TotalDeaths), ]

least_deaths_age <- age_summary[which.min(age_summary$TotalDeaths), ]
```

```r
# Mean Death Rate
# Highest / Lowest Mean Death Rate
highest_mean_deathrate_age <- age_summary[which.max(age_summary$MeanDeathRate), ]

lowest_mean_deathrate_age <- age_summary[which.min(age_summary$MeanDeathRate), ]

# Max / Min Death Rate
highest_max_deathrate_age <- age_summary[which.max(age_summary$MaxDeathRate), ]

lowest_max_deathrate_age <- age_summary[which.min(age_summary$MaxDeathRate), ]

# Print all
cat(
  " Age Group with the MOST cases:", highest_cases_age$AgeGroup,
  "(", highest_cases_age$TotalCases, "cases)\n",

  "Age Group with the FEWEST cases:", lowest_cases_age$AgeGroup,
  "(", lowest_cases_age$TotalCases, "cases)\n\n",

  "Age Group with the HIGHEST average case rate:", highest_mean_case_age$AgeGroup,
  "(", highest_mean_case_age$MeanCaseRate, "per 100k)\n",

  "Age Group with the LOWEST average case rate:", lowest_mean_case_age$AgeGroup,
  "(", lowest_mean_case_age$MeanCaseRate, "per 100k)\n\n",

  "Age Group with the MOST deaths:", most_deaths_age$AgeGroup,
  "(", most_deaths_age$TotalDeaths, "deaths)\n",

  "Age Group with the FEWEST deaths:", least_deaths_age$AgeGroup,
  "(", least_deaths_age$TotalDeaths, "deaths)\n\n",

  "Age Group with the HIGHEST average death rate:", highest_mean_deathrate_age$AgeGroup,
  "(", highest_mean_deathrate_age$MeanDeathRate, "per 100k)\n",

  "Age Group with the LOWEST average death rate:", lowest_mean_deathrate_age$AgeGroup,
  "(", lowest_mean_deathrate_age$MeanDeathRate, "per 100k)\n\n",

  "Age Group with the HIGHEST death rate:", highest_max_deathrate_age$AgeGroup,
  "(", highest_max_deathrate_age$MaxDeathRate, "per 100k)\n",

  "Age Group with the LOWEST death rate:", lowest_max_deathrate_age$AgeGroup,
  "(", lowest_max_deathrate_age$MaxDeathRate, "per 100k)\n"
)
```

```
##  Age Group with the MOST cases: 18 - 29 Years ( 131665357 cases)
##  Age Group with the FEWEST cases: 16 - 17 Years ( 16895805 cases)
##
##  Age Group with the HIGHEST average case rate: 30 - 39 Years ( 138.9509 per 100k)
##  Age Group with the LOWEST average case rate: 0 - 4 Years ( 73.61458 per 100k)
##
##  Age Group with the MOST deaths: 75+ Years ( 3944452 deaths)
##  Age Group with the FEWEST deaths: 16 - 17 Years ( 248 deaths)
##
##  Age Group with the HIGHEST average death rate: 75+ Years ( 9.078197 per 100k)
```

```
##  Age Group with the LOWEST average death rate: 16 - 17 Years ( 0.0002006679 per 100k)
##
##  Age Group with the HIGHEST death rate: 75+ Years ( 660.68 per 100k)
##  Age Group with the LOWEST death rate: 5 - 11 Years ( 0.59 per 100k)
```

Analysis:

1. We made a comparison across different age groups for the total cases, total deaths, and death rates of COVID-19. The 18–29 age group had the highest number of reported cases (over 131 million), and the 16–17 age group had the lowest (around 17 million). The 30–39 age group had the highest average case rate per 100,000 people (138.95), which was slightly higher than the 18–29 age group.
2. In the case of the number of deaths, the 75+ age group was the most affected, with around 3.94 million deaths.The lowest number of deaths was in the 16–17 age group (a total of 248). The 75+ age group had the highest average death rate (9.0782 per 100,000 people), and the lowest was in the 16–17 age group (0.0002 per 100,000). In addition, the highest weekly death rate was also in the 75 + age group (660.68 per 100,000), andthe lowest in the 5–11 age group (0.59 per 100,000).

Age ~ Death Rate Simple linear regression

```r
# simple linear regression age ~ DeathRate
model_deathdata_age <- covid_data %>%
  filter(AgeGroup != "Overall") %>%
  select(AgeGroup, DeathRate) %>%
  mutate(AgeGroup = as.factor(AgeGroup))

model_death_age <- lm(DeathRate ~ AgeGroup, data = model_deathdata_age)
summary(model_death_age)
```

Analysis:

1. The null hypothesis (H0) was no difference in COVID-19 death rate by age group, and the alternative hypothesis (H1) was that death rates differ by age. We tested it using a linear model with death rate and age group.
2. The model showed statistically significant differences ($p < 0.001$). Compared to the baseline group (ages 0–4), older age groups (particularly those over 50) showed significantly higher death rates than the baseline group. Among the groups, the 75 + group experienced the highest increase, an average of 9.078 more deaths per 100,000 people ($p < 0.001$). The 65–74 age group experienced the second highest increase of 2.741, and the 50–64 age group had an increase of 1.016, both significantly different ($p < 0.001$). The younger age groups did not experience significant differences. The adjusted R-squared was 0.1198, show that age group can be explain approximately 11.98% of the changes in COVID-19 death rates.
3. These results are supported by previous research. In the "COVID-19 in the elderly" section of Kang & Jung (2020)(Link 1), the authors explain that older adults are more susceptible to infections due to the natural decline of immune function with age. Their immune systems are less responsive, making it harder to fight off viruses like COVID-19. The study also highlights that chronic subclinical systemic inflammation, also known as inflammaging, may worsen COVID-19 outcomes in elderly patients. Similarly, in the "Association of Comorbidities with Mortality" section of Biswas et al. (2020)(Link 2), the authors found that older adults were more likely to have chronic health conditions such as hypertension, diabetes mellitus, cardiovascular disease, and chronic respiratory disease. These comorbidities were strongly associated with higher mortality among COVID-19 patients. Although we did not include information in our model, their findings help explain why the 75+ group experienced the highest death rates. It may be a combination of weakened immune response and underlying health conditions.

Age ~ Case Rate Simple linear regression

```
# simple linear regression age ~ CaseRate
model_casedata_age <- covid_data %>%
  filter(AgeGroup != "Overall") %>%
  select(AgeGroup, CaseRate) %>%
  mutate(AgeGroup = as.factor(AgeGroup))

model_case_age <- lm(CaseRate ~ AgeGroup, data = model_casedata_age)
summary(model_case_age)
```

Analysis:

1. We used a linear regression model to explore the relationship between age group and COVID-19 case rate. The null hypothesis (H0) was that there is no difference in case rate by age group, and the alternative hypothesis (H1) was that case rates differ by age. For this model, the 0–4 years age group was used as the baseline.
2. The results showed statistically significant differences for all age groups ($p < 0.001$). Compared to the baseline, the highest case rate increases were in the 30–39 (+65.34 per 100,000) and 18–29 (+63.79 per 100,000) age groups. Their estimated average case rates were around 138.95 and 137.40 per 100,000 respectively. The youngest children (age 5–11) had the lowest increase (+14.33), and the 75+ group had a moderate increase (+35.29). Since the p-value was below 0.05 for all groups, we reject the null hypothesis. It can be inferred that age is actually related to case rate. However, adjusted R-squared was around 0.0121, which means that age only explains about 1.21% of the variance in case rate. It can be say that there are many other variables that can affect the risk of infection.
3. These results are supported by previous research. In the "Age groups that sustain resurging COVID-19 epidemics in the United States" section of Monod et al. (2021)(Link 3), the authors show that most new cases of COVID-19 were from adults aged 20-49. Specifically, at peak times, the largest numbers of infections in most U.S. regions were in the age 20-34 and age 35-49 groups. Their infectivity was due to social mobility, high contact with other adults, and higher opportunities to visit public areas.This supports our finding that rates of cases were highest among young to middle-aged adults, because they expose themselves and have more contact with others during the pandemic.

Covid case and death base on Sex

```
# sex summary table (excluding Overall)
sex_summary <- covid_data %>%
  filter(Sex != "Overall") %>%
  group_by(Sex) %>%
  summarise(
    TotalCases = sum(CaseSuppressed, na.rm = TRUE),
    TotalDeaths = sum(DeathSuppressed, na.rm = TRUE),
    MeanCaseRate = mean(CaseRate, na.rm = TRUE),
    MeanDeathRate = mean(DeathRate, na.rm = TRUE),
    MaxDeathRate = max(DeathRate, na.rm = TRUE),
    .groups = "drop"
  )

# Total Case
# Highest of cases
highest_cases_sex <- sex_summary[which.max(sex_summary$TotalCases), ]

# Lowest of cases
```

```r
lowest_cases_sex <- sex_summary[which.min(sex_summary$TotalCases), ]


# Total Death
# Highest of death
most_deaths_sex <- sex_summary[which.max(sex_summary$TotalDeaths), ]

#Lowest of death
least_deaths_sex <- sex_summary[which.min(sex_summary$TotalDeaths), ]

# Mean Death Rate
# Highest Mean Death Rate
highest_mean_deathrate_sex <- sex_summary[which.max(sex_summary$MeanDeathRate), ]

# Lowest Mean Death Rate
lowest_mean_deathrate_sex <- sex_summary[which.min(sex_summary$MeanDeathRate), ]

# Max Death Rate
highest_max_deathrate_sex <- sex_summary[which.max(sex_summary$MaxDeathRate), ]

# Min Death Rate
lowest_max_deathrate_sex <- sex_summary[which.min(sex_summary$MaxDeathRate), ]


# Total Case
cat(
  "  Sex group with the MOST cases: ", highest_cases_sex$Sex,
  "(", highest_cases_sex$TotalCases, " cases)\n ",

  "Sex group with the FEWEST cases: ", lowest_cases_sex$Sex,
  "(", lowest_cases_sex$TotalCases, " cases)\n\n ",

  # Total Death
  "Sex group with the MOST deaths: ", most_deaths_sex$Sex,
  "(", most_deaths_sex$TotalDeaths, " deaths)\n ",

  "Sex group with the FEWEST deaths: ", least_deaths_sex$Sex,
  "(", least_deaths_sex$TotalDeaths, " deaths)\n\n ",

  # Mean Death Rate
  "Sex group with the HIGHEST average death rate: ", highest_mean_deathrate_sex$Sex,
  "(", highest_mean_deathrate_sex$MeanDeathRate, " per 100k)\n ",

  "Sex group with the LOWEST average death rate: ", lowest_mean_deathrate_sex$Sex,
  "(", lowest_mean_deathrate_sex$MeanDeathRate, " per 100k)\n\n ",

  # Max Death Rate
  "Sex group with the HIGHEST death rate: ", highest_max_deathrate_sex$Sex,
  "(", highest_max_deathrate_sex$MaxDeathRate, " per 100k)\n ",

  "Sex group with the LOWEST death rate: ", lowest_max_deathrate_sex$Sex,
  "(", lowest_max_deathrate_sex$MaxDeathRate, " per 100k)"
)
```

```
##   Sex group with the MOST cases:  Female ( 360931148  cases)
##   Sex group with the FEWEST cases:  Male ( 302206161  cases)
##
##   Sex group with the MOST deaths:  Male ( 3970788  deaths)
##   Sex group with the FEWEST deaths:  Female ( 3282344  deaths)
##
##   Sex group with the HIGHEST average death rate:  Male ( 1.541264  per 100k)
##   Sex group with the LOWEST average death rate:  Female ( 0.9983938  per 100k)
##
##   Sex group with the HIGHEST death rate:  Male ( 660.68  per 100k)
##   Sex group with the LOWEST death rate:  Female ( 369.58  per 100k)
```

Analysis:

1. We analyzed how sex relates to COVID-19 death outcomes. Females had more total cases (360 million), and males had more total deaths (3.97 million). Males also had a higher average death rate (1.54 per 100k) and a higher maximum death rate (660.68 per 100k) compared to females.

Sex ~ Death Rate T-test & Mann-Whitney U test

```
ttest_deathdata_sex <- covid_data %>%
  filter(Sex != "Overall") %>%
  select(Sex, DeathRate) %>%
  mutate(Sex = as.factor(Sex))

# print Hypothese
cat("Hypotheses for T-test and Mann-Whitney U test:\n",
"H0: There is no difference in death rate between males and females.\n",
"H1: There is a difference in death rate between males and females.\n")
```

```
## Hypotheses for T-test and Mann-Whitney U test:
##  H0: There is no difference in death rate between males and females.
##  H1: There is a difference in death rate between males and females.
```

```
# T- Test
t.test(DeathRate ~ Sex, data = ttest_deathdata_sex)

# Mann-Whitney U test
wilcox.test(DeathRate ~ Sex, data = ttest_deathdata_sex)
```

Analysis:

1. The null hypothesis (H0) was that there is no difference in death rate between the two groups, and the alternative hypothesis (H1) was that there is a difference.To test this hypothesis,we used both a Welch Two Sample T-test and a Mann-Whitney U test.

2. The T-test results showed that the average death rate for males was 1.541 and for females it was 0.998. The p-value result was less than 0.001 (p-value $<$ 2.2e-16), and the 95% confidence interval for the difference in means was between (–0.600) and (–0.485),thus not including zero. This result show that the difference is statistically significant, and the null hypothesis could be rejected. The Mann-Whitney U test also gave a p-value less than 0.001 (p-value $<$ 2.2e-16), which supports the same conclusion.Since both tests show strong evidence of a difference, we conclude that males have a significantly higher COVID-19 death rate than females.

3. These findings are supported by previous research. In the "Discussion" of Biswas et al. (2020)(Link 2), the authors observed that male COVID-19 patients fared on average worse than female patients. One possible explanation that researchers examined is whether the biological differences of sex could explain

differences in immune response and susceptibility to severe infection. The authors concluded that the exact processes are unclear, but inference was made that hormonal and genetic variations may be a possible explanation. In addition, In the "4.2 Social factors in variation in male-female disparities in COVID-19 outcomes" section in Danielsen et al. (2022)(Link 4) also highlight the potential contribution of social and behavioral factors in explaining these disparities. Specifically,it has been show that men had lower rates of following with public health recommendations, such as social distancing and wearing a mask. And male were more likely to work in high-risk occupations (manufacturing, transport and construction), that put them at higher risk of exposure to the virus.

Sex ~ Case Rate Simple linear regression

```
## simple linear regression Sex ~ CaseRate
model_casedata_sex <- covid_data %>%
  filter(Sex != "Overall") %>%
  select(Sex, CaseRate) %>%
  mutate(Sex = as.factor(Sex))

model_case_sex <- lm(CaseRate ~ Sex, data = model_casedata_sex)
summary(model_case_sex)
```

Analysis:

1. We used linear regression modeling to compare the case rates for females and males. The baseline group was females with a mean case rate of 118.440 cases per 100,000 population. The case rate for males was 14.197 ($p < 0.001$) lower than that of females, which means that their estimated case rate was approximately 104.243 cases per 100,000 people.
2. The R-squared is 0.0015, which means that gender only explains 0.15% of the difference in case rate. This suggests that gender has little effect on the case rate, although it is statistically significant. In short, females have a slightly higher case rate than males, but the difference is small.

Sex ~ Case Rate T-test & Mann-Whitney U test

```
ttest_casedata_sex <- covid_data %>%
  filter(Sex != "Overall") %>%
  select(Sex, CaseRate) %>%
  mutate(Sex = as.factor(Sex))

# Step 2: Print the hypotheses
cat("Hypotheses for T-test and Mann-Whitney U test:\n",
"H0: There is no difference in COVID-19 case rate between males and females.\n",
"H1: There is a difference in COVID-19 case rate between males and females.\n\n")
```

```
## Hypotheses for T-test and Mann-Whitney U test:
##  H0: There is no difference in COVID-19 case rate between males and females.
##  H1: There is a difference in COVID-19 case rate between males and females.
```

```
# T-Test
t.test(CaseRate ~ Sex, data = ttest_casedata_sex)

# Mann-Whitney U test
wilcox.test(CaseRate ~ Sex, data = ttest_casedata_sex)
```

Analysis:

1. The null hypothesis (H) was that there is no difference in COVID-19 case rate between the two groups (males and females), and the alternative hypothesis (H) was that there is a difference. To test this

hypothesis, we used both a Welch Two Sample T-test and a Mann-Whitney U test.

2. The T-test result showed a statistically significant difference ($t = 20.15$, $p < 0.001$), with females having a higher average case rate (mean = 118.44) compared to males (mean = 104.24). Also,the 95% confidence interval for the difference in the means was (12.82, 15.58), and this interval does not include zero. Because the p-value was below 0.05, we rejected the null hypothesis and concluded that there is a statistically significant difference in case rates for male and female. we also use a Mann-Whitney U test to confirming the result. This test also showed a significant difference ($p < 0.001$), supporting the same conclusion. The results show a relationship between sex and the rate of COVID-19, females experiencing higher case rates on average.

3. These findings are supported by previous research. In the article entitled "Is there a sex difference in COVID-19 Outcomes in the U.S.?" by Danielsen et al. (2022) (Link 4), The authors noted the higher proportion of confirmed cases in females is likely impacted due to testing policies. COVID-19 tests were especially required for pregnant women. Moreover, most of the tests were conducted on women hospitalized to give birth. In addition, health care workers, who were mostly women, were also tested more often because of their occupation and the routine testing policy. ll these factors most probably led to detecting more cases in women, which served to explain the differences in our model.

Age + Sex ~ Death Rate Multiple linear regression

```
# multiple linear regression Age + Sex ~ DeathRate
model_data_agesex <- covid_data %>%
  filter(Sex != "Overall", AgeGroup != "Overall") %>%
  select(AgeGroup, Sex, DeathRate) %>%
  mutate(
    AgeGroup = as.factor(AgeGroup),
    Sex = as.factor(Sex)
  )

# Multiple linear regression model: Age + Sex ~ DeathRate
model_age_sex <- lm(DeathRate ~ AgeGroup + Sex, data = model_data_agesex)
summary(model_age_sex)
```

Analysis:

1. We used a multiple linear regression to examine how age group and sex affect COVID-19 death rate, using females aged 0–4 years as the baseline group.The model showed that people aged over 40 had much higher death rates than those under 40, and males had higher death rates than females.For example, people aged 75+ had 8.9429 more deaths per 100k than the baseline, and males had 0.5772 more deaths than females ($p < 0.0001$).The R-squared of the model is 0.113, meaning age and sex together explain about 11.3% of the difference in death rate.

2. These findings extend our earlier single-variable analyses,and are consistent with general patterns. As we mentioned, older people — particularly those who are 75 years +—have a higher risk of mortality due to declines in immune function, chronic inflammation and increased probability of having comorbid conditions. Male have elevated COVID-19 mortality compared to Female, possibly as a result of biological factors and risk-taking behavior.When these two factors combined in the same model, both factors remain significant. This goes a significant way in explaining why older male population and certain risk factors have the highest mortality rates of all observed groups.

Covid case and death base on Race

```
# Race Summary table (excluding Overall)
race_summary <- covid_data %>%
  filter(Race != "Overall") %>%
```

```r
  group_by(Race) %>%
  summarise(
    TotalCases = sum(CaseSuppressed, na.rm = TRUE),
    TotalDeaths = sum(DeathSuppressed, na.rm = TRUE),
    MeanCaseRate = mean(CaseRate, na.rm = TRUE),
    MeanDeathRate = mean(DeathRate, na.rm = TRUE),
    MaxDeathRate = max(DeathRate, na.rm = TRUE),
    .groups = "drop"
  )

# Total Case
# Highest of cases
highest_cases_race <- race_summary[which.max(race_summary$TotalCases), ]

# Lowest of cases
lowest_cases_race <- race_summary[which.min(race_summary$TotalCases), ]

# Mean Case Rate
highest_mean_case_race <- race_summary[which.max(race_summary$MeanCaseRate), ]

lowest_mean_case_race <- race_summary[which.min(race_summary$MeanCaseRate), ]


# Total Death
# Highest of deaths
most_deaths_race <- race_summary[which.max(race_summary$TotalDeaths), ]

# Lowest of deaths
least_deaths_race <- race_summary[which.min(race_summary$TotalDeaths), ]


# Highest Mean Death Rate
highest_mean_deathrate_race <- race_summary[which.max(race_summary$MeanDeathRate), ]

# Lowest Mean Death Rate
lowest_mean_deathrate_race <- race_summary[which.min(race_summary$MeanDeathRate), ]

# Highest Max Death Rate
highest_max_deathrate_race <- race_summary[which.max(race_summary$MaxDeathRate), ]

# Lowest Max Death Rate
lowest_max_deathrate_race <- race_summary[which.min(race_summary$MaxDeathRate), ]

# Total Case
cat(
  " Race group with the MOST cases:", highest_cases_race$Race,
  "(", highest_cases_race$TotalCases, " cases)\n",

  "Race group with the FEWEST cases:", lowest_cases_race$Race,
  "(", lowest_cases_race$TotalCases, " cases)\n\n",

  # Mean Case Rate
  "Race group with the HIGHEST average case rate:", highest_mean_case_race$Race,
```

```
  "(", highest_mean_case_race$MeanCaseRate, " per 100k)\n",

  "Race group with the LOWEST average case rate:", lowest_mean_case_race$Race,
  "(", lowest_mean_case_race$MeanCaseRate, " per 100k)\n\n",

  # Total Death
  "Race group with the MOST deaths:", most_deaths_race$Race,
  "(", most_deaths_race$TotalDeaths, " deaths)\n",

  "Race group with the FEWEST deaths:", least_deaths_race$Race,
  "(", least_deaths_race$TotalDeaths, " deaths)\n\n",

  # Mean Death Rate
  "Race group with the HIGHEST average death rate:", highest_mean_deathrate_race$Race,
  "(", highest_mean_deathrate_race$MeanDeathRate, " per 100k)\n",

  "Race group with the LOWEST average death rate:", lowest_mean_deathrate_race$Race,
  "(", lowest_mean_deathrate_race$MeanDeathRate, " per 100k)\n\n",

  # Max Death Rate
  "Race group with the HIGHEST death rate:", highest_max_deathrate_race$Race,
  "(", highest_max_deathrate_race$MaxDeathRate, " per 100k)\n",

  "Race group with the LOWEST death rate:", lowest_max_deathrate_race$Race,
  "(", lowest_max_deathrate_race$MaxDeathRate, " per 100k)\n"
)
```

```
##  Race group with the MOST cases: White, NH ( 293941659  cases)
##  Race group with the FEWEST cases: AI/AN, NH ( 5333095  cases)
##
##  Race group with the HIGHEST average case rate: Hispanic ( 136.5546  per 100k)
##  Race group with the LOWEST average case rate: Asian/PI, NH ( 75.52885  per 100k)
##
##  Race group with the MOST deaths: White, NH ( 4369003  deaths)
##  Race group with the FEWEST deaths: AI/AN, NH ( 53139  deaths)
##
##  Race group with the HIGHEST average death rate: Hispanic ( 1.814881  per 100k)
##  Race group with the LOWEST average death rate: Asian/PI, NH ( 0.5773722  per 100k)
##
##  Race group with the HIGHEST death rate: Hispanic ( 660.68  per 100k)
##  Race group with the LOWEST death rate: White, NH ( 185.8  per 100k)
```

Analysis:

1. We looked at COVID-19 cases and deaths by race group. The group with the most cases was White, NH (about 294 million), and the group with the fewest cases was AI/AN, NH (around 5.3 million). However, the highest average case rate was in the Hispanic group (136.555 per 100k), and the lowest was in Asian/PI, NH (75.529 per 100k).
2. For total deaths, White, NH also had the most deaths (over 4.3 million), and AI/AN, NH had the fewest deaths (about 53,000).
3. But when we look at death rates, the results are different. Hispanic had the highest average death rate and also the highest weekly death rate (660.68 per 100k). Asian/PI, NH had the lowest average death rate, and White, NH had the lowest weekly death rate.

4. So even though White, NH had more total deaths, Hispanic had a higher risk of death from COVID-19.

Death Rate ~ Race Simple linear regression

```
# simple linear regression DeathRate ~ Race
model_data_race <- covid_data %>%
  filter(Race != "Overall") %>%
  select(Race, DeathRate) %>%
  mutate(Race = as.factor(Race))

model_race_simple <- lm(DeathRate ~ Race, data = model_data_race)
summary(model_race_simple)
```

Analysis:

1. The null hypothesis (H) was that there is no difference in COVID-19 death rates across racial/ethnic groups. The alternative hypothesis (H) was that death rates vary by race. We tested this using a linear regression model, with death rate and race group. The baseline group was American Indian/Alaska Native, non-Hispanic (AI/AN, NH).

2. The model showed that all race groups had statistically significant differences in death rate compared to the baseline ($p < 0.001$). The model showed that all race groups had statistically significant differences in death rate compared to the baseline ($p < 0.001$). Hispanic had the highest estimated increase in death rate (+1.067 per 100k), followed by Black, NH (+0.795), and White, NH (+0.529). In contrast, Asian/PI, NH had a slight decrease (–0.172) compared to AI/AN, NH. Although these differences were statistically significant, the adjusted R-squared was very low (0.0038), indicating that race alone does not explain much of the variation in death rate.

3. These findings are supported by Bassett et al. (2020)(Link 5) in the Discussion section of their study in PLOS Medicine. The authors note that the existing disparities in COVID-19 mortality were substantially higher in both Black and Hispanic populations, especially in the younger age group. Specifically, they found significantly higher rates of premature mortality and years of life lost (YPLL) compared with similar populations of White populations. In discussing their findings, the authors suggest that aside from comorbidities, the causes of the large disparities were attributed to increased exposure risk due to essential work, less access to health care, and more broadly, structural inequities in society. In part, this might explain why the predicted death rate for Hispanics was the most elevated in our model.

Case Rate ~ Race Simple linear regression

```
## simple linear regression CaseRate ~ Race
model_data_race <- covid_data %>%
  filter(Race != "Overall") %>%
  select(Race, CaseRate) %>%
  mutate(Race = as.factor(Race))

model_case_race <- lm(CaseRate ~ Race, data = model_data_race)
summary(model_case_race)
```

Analysis:

1. We used a linear regression model to compare case rates across racial groups. The baseline group was AI/AN, NH, with an average case rate of 121.988 per 100k.

2. Hispanic had a higher case rate (+14.567), while Asian/PI, NH (−46.459), White, NH (−33.650), and Black, NH (−23.030) all had lower case rates compared to the baseline. All differences were statistically significant ($p < 0.001$).

3. The R-squared was 0.017, which means race explains about 1.7% of the difference in case rate. So, the differences between groups are clear, but race alone does not explain most of the variation.
4. In short, Hispanic had the highest case rate, and Asian/PI had the lowest.
5. According to the Discussion section of Bassett et al. (2020) (Link 5), Hispanic communities in the U.S. faced higher risks of COVID-19 infection because many people worked essential jobs, had less access to healthcare, and lived in more crowded conditions. These challenges likely made it easier for the virus to spread. This helps explain why the Hispanic group had the highest case rate in our model.

Age + Sex + Race ~ Death Rate Multiple linear regression

```
#Multiple linear regression Age + Sex + Race ~ DeathRate
model_deathdata_asr <- covid_data %>%
  filter(AgeGroup != "Overall", Sex != "Overall", Race != "Overall") %>%
  select(AgeGroup, Sex, Race, DeathRate) %>%
  mutate(
    AgeGroup = as.factor(AgeGroup),
    Sex = as.factor(Sex),
    Race = as.factor(Race)
  )

# Multiple linear regression model: Age + Sex + Race ~ DeathRate
model_death_asr <- lm(DeathRate ~ AgeGroup + Sex + Race, data = model_deathdata_asr)
summary(model_death_asr)
```

Analysis:

1. We found earlier that race alone had a significant effect on death rate, but the R-squared was very low (0.00377), which means race by itself didn't explain much. We already found that age and sex had a strong effect on death rate.So, in this final model, we added age and sex to see if race still matters.The effect for Hispanic increased from (+1.066) to (+1.200) per 100k, and Black and White also stayed significantly higher than the baseline. This shows that race still has an independent effect, even after controlling for age and sex.
2. The R-squared of the full model is 0.097, so these three factors together explain about 9.7% of the difference in death rate. Although the R-squared is only 0.097, it still suggests that age, gender, and race together have a measurable effect on deathrate. However, much of the difference may be due to other factors not included in the model.
3. These results are consistent with previous studies. Older adults have a higher risk of dying from COVID-19 because they have a less effective immune response and are more likely to develop severe disease. Male also had higher mortality rates, which may be related to biological factors and lower protective health behaviors. Hispanic and black communities face higher mortality rates, not only because of health risk factors, but also likely because of lower access to health care, higher rates of work exposure. These combined factors help explain the differences in mortality rates across age, gender, and race that we observed in our model.

Age + Sex + Race ~ Case Rate Multiple linear regression

```
# Multiple linear regression: Age + Sex + Race ~ CaseRate
model_casedata_asr <- covid_data %>%
  filter(AgeGroup != "Overall", Sex != "Overall", Race != "Overall") %>%
  select(AgeGroup, Sex, Race, CaseRate) %>%
  mutate(
    AgeGroup = as.factor(AgeGroup),
```

```
    Sex = as.factor(Sex),
    Race = as.factor(Race)
  )

model_case_asr <- lm(CaseRate ~ AgeGroup + Sex + Race, data = model_casedata_asr)
summary(model_case_asr)
```

Analysis:

1. We built a multiple linear regression model to study how age, sex, and race together affect COVID-19 case rate. The baseline group was children aged 0–4, female, and AI/AN.
2. Adults aged 30–39 and 18–29 had the highest increases in case rate, with estimates of (+60.246) and (+58.051) per 100k. Males had a lower case rate than females (−13.218). For race, Hispanic had the highest increase (+16.007), while Asian/PI (−45.568), White (−32.024), and Black (−22.084) had lower case rates than the baseline. All effects were statistically significant ($p < 0.001$).
3. The R-squared was 0.029, meaning the model explains about 2.9% of the difference in case rate. In short, people aged 18–39 and Hispanic groups had the highest case rates, but age, sex, and race together only explain a small part of the variation.
4. These results are consistent with previous studies. The study concludes that because of increased testing, women may have increased case rates, specifically in pregnant women and health care worker. The Hispanic community may be at increased risk for exposure and infection owing to their jobs, crowded living situations, and decreased access to health care. Younger adults also experience higher risks of infection due to higher mobility and job exposures. These factors explain the patterns by age, sex, and race that we observed in our model.

State-Level Analysis

```
# summary tables: total cases, total deaths per state
state_summary <- state_case_death %>%
  group_by(State) %>%
  summarise(
    TotalCases = sum(NewCases, na.rm = TRUE),
    TotalDeaths = sum(NewDeaths, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  mutate(
    Death_Percentage = TotalDeaths / TotalCases * 100  #Calculate Death Percentage
  )

# Highest total cases
most_cases_state <- state_summary[which.max(state_summary$TotalCases), ]

# Lowest total cases
least_cases_state <- state_summary[which.min(state_summary$TotalCases), ]

# Highest total deaths
most_deaths_state <- state_summary[which.max(state_summary$TotalDeaths), ]

# Lowest total deaths
least_deaths_state <- state_summary[which.min(state_summary$TotalDeaths), ]

# Highest death percentage
```

```r
highest_deathrate_state <- state_summary[which.max(state_summary$Death_Percentage), ]

# Lowest death percentage
lowest_deathrate_state <- state_summary[which.min(state_summary$Death_Percentage), ]

# print result
cat(
  "State with the MOST total cases:", most_cases_state$State,
  " (", most_cases_state$TotalCases, " cases)\n",
  "State with the FEWEST total cases:", least_cases_state$State,
  " (", least_cases_state$TotalCases, " cases)\n\n",
  "State with the MOST total deaths:", most_deaths_state$State,
  " (", most_deaths_state$TotalDeaths, " deaths)\n",
  "State with the FEWEST total deaths:", least_deaths_state$State,
  " (", least_deaths_state$TotalDeaths, " deaths)\n\n",
  "State with the HIGHEST death rate:", highest_deathrate_state$State,
  " (", round(highest_deathrate_state$Death_Percentage, 2), " %)\n",
  "State with the LOWEST death rate:", lowest_deathrate_state$State,
  " (", round(lowest_deathrate_state$Death_Percentage, 2), " %)\n",
  sep = ""
)
```

```
## State with the MOST total cases:CA (12251820 cases)
## State with the FEWEST total cases:PW (6000 cases)
##
## State with the MOST total deaths:CA (101886 deaths)
## State with the FEWEST total deaths:PW (9 deaths)
##
## State with the HIGHEST death rate:PA (1.43 %)
## State with the LOWEST death rate:RMI (0.11 %)
```

Analysis:

1. We analyzed the total number of COVID-19 cases and deaths in each US state. CA had the highest number of cases (12,251,820) and deaths (101,886). The RMI had the lowest number of cases and deaths, with only 6,000 and 9 cases respectively.
2. In terms of percentage of death, PA had the highest at 1.430% . The RMI had the lowest at 0.110%.
3. This shows that some states have much higher percentage of death even though the total number of cases not highest.

State Simple linear regression

```r
model_simple_state <- lm(TotalDeaths ~ TotalCases, data = state_summary)
summary(model_simple_state)
```

Analysis:

1. We used linear regression modeling to see if the number of COVID-19 cases in each US state predicted the number of deaths.

The results showed a very strong statistical relationship ($p < 0.0001$). On average, every 1000 confirmed cases were associated with approximately 10 deaths.

2. The R-squared was 0.942, meaning that the number of cases explained 94.2% of the variation in the total number of deaths across states. This shows that the number of cases by itself is a good relative

of the total number of deaths in each state.