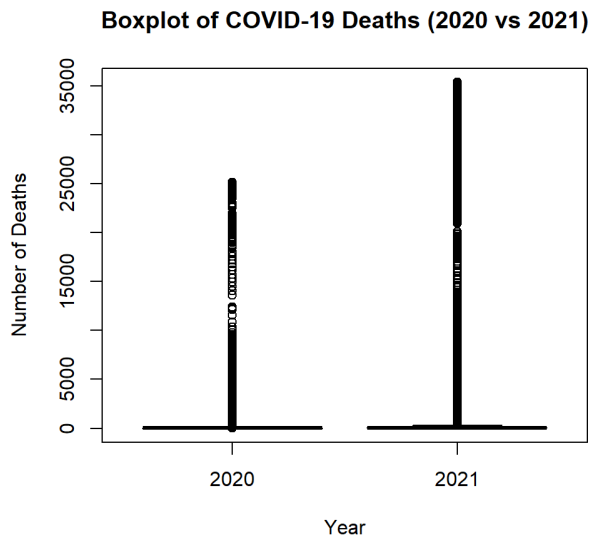
Tatiana Leonovich

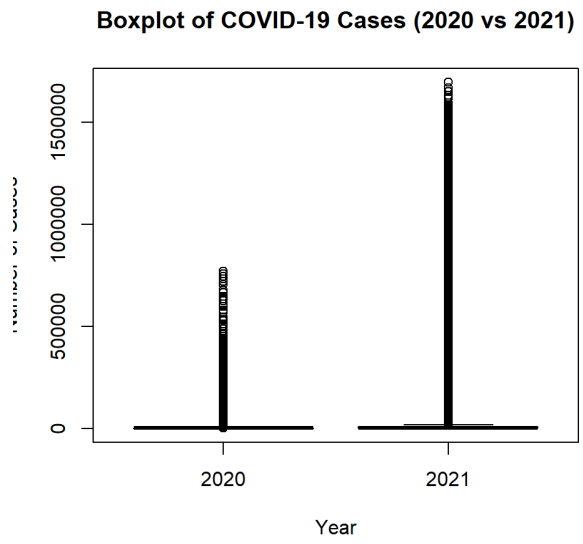
Dr. Ahmed Eleish

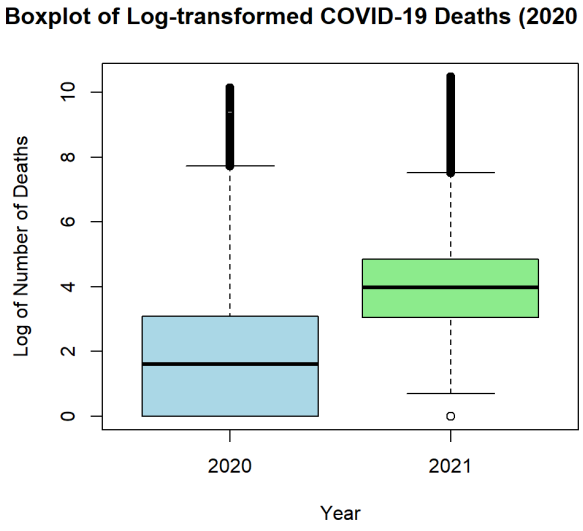
Data Analytics

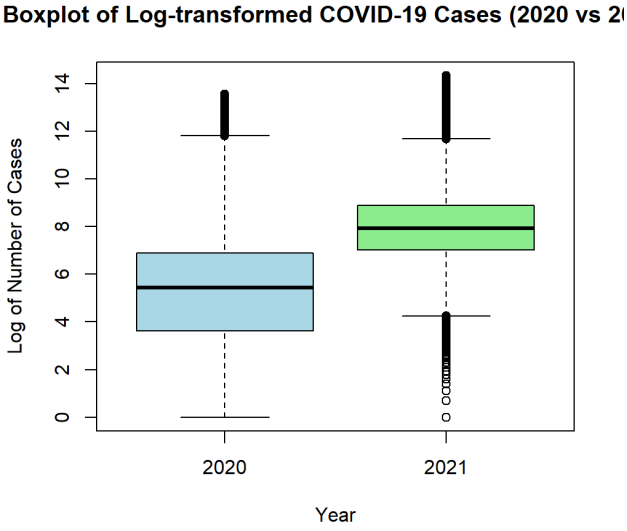
Assignment 3

(a)









Code for Part A:

```
# Combine the two datasets by stacking them vertically
combined_data <- rbind(counties2020, counties2021)

# Create boxplot for "Cases"
boxplot(cases ~ year, data = combined_data,
        main = "Boxplot of COVID-19 Cases (2020 vs 2021)",
        xlab = "Year",
        ylab = "Number of Cases",
        col = c("lightblue", "lightgreen"))
# Create boxplot for "Deaths"
boxplot(deaths ~ year, data = combined_data,
        main = "Boxplot of COVID-19 Deaths (2020 vs 2021)",
        xlab = "Year",
        ylab = "Number of Deaths",
        col = c("lightblue", "lightgreen"))

combined_data$log_cases <- log1p(combined_data$cases)  # log1p handles 0 values well
combined_data$log_deaths <- log1p(combined_data$deaths)

# Create boxplot for log-transformed "Cases"
boxplot(log_cases ~ year, data = combined_data,
```

```
    main = "Boxplot of Log-transformed COVID-19 Cases (2020 vs 2021)",
    xlab = "Year",
    ylab = "Log of Number of Cases",
    col = c("lightblue", "lightgreen"))

# Create boxplot for log-transformed "Deaths"
boxplot(log_deaths ~ year, data = combined_data,
    main = "Boxplot of Log-transformed COVID-19 Deaths (2020 vs 2021)",
    xlab = "Year",
    ylab = "Log of Number of Deaths",
    col = c("lightblue", "lightgreen"))

# Summary statistics for "Cases"
summary(counties2020$cases)
summary(counties2021$cases)

# Summary statistics for "Deaths"
summary(counties2020$deaths)
summary(counties2021$deaths)
```

Written Answer for Part (a)

Log transformation was applied to reduce skewness in house prices, which are typically right-skewed, and to minimize the impact of outliers. This helps improve the interpretability of

the box plots by compressing large values and providing a clearer view of the data's central tendency and spread. Additionally, it ensures better visualization and prepares the data for more robust modeling.

Summary Statistics:

```
 summary(counties2020$cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.
      0      36     228    1952     993
   Max.
 770915
> summary(counties2021$cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.
      0    1136    2778   11160    7340
   Max.
1697286
> # Summary statistics for "Deaths"
> summary(counties2020$deaths)
   Min. 1st Qu.  Median    Mean 3rd Qu.
    0.0     0.0     4.0    53.6    21.0
   Max.    NA's
25144.0   18761
> summary(counties2021$deaths)
   Min. 1st Qu.  Median    Mean 3rd Qu.
    0.0    20.0    52.0   193.6   125.0
```

Max. NA's

35382.0 28470

Cases:

In 2020, the minimum number of cases in a county was 0, with a median of 228, and a mean of 1,952, indicating a right-skewed distribution driven by a few counties with very high cases (maximum: 770,915). The majority of counties reported fewer than 993 cases (3rd quartile).

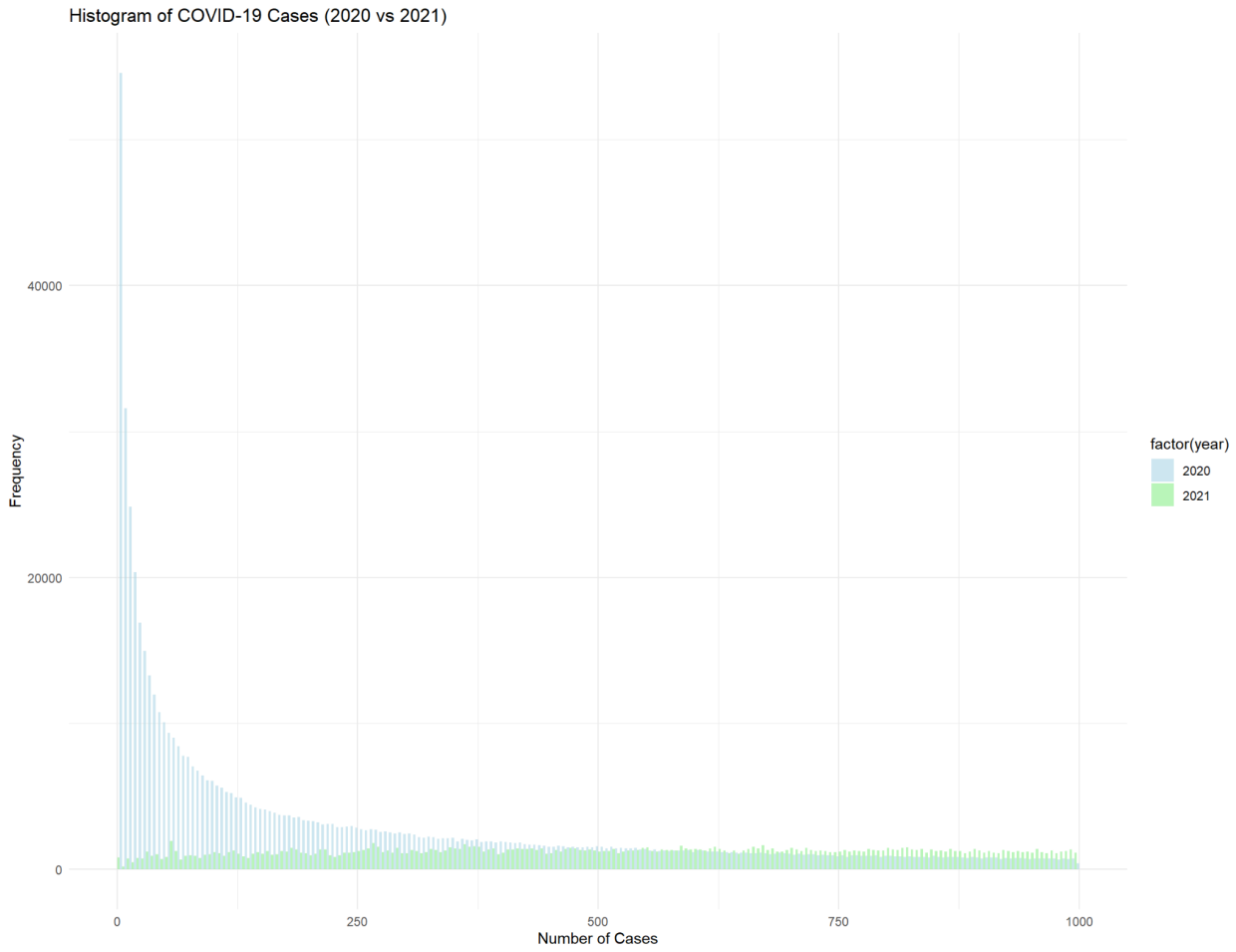
In 2021, the case counts increased significantly, with a median of 2,778 and a mean of 11,160, showing higher case rates and a more skewed distribution (maximum: 1,697,286). The larger spread reflects the pandemic's progression and increased testing/reporting.

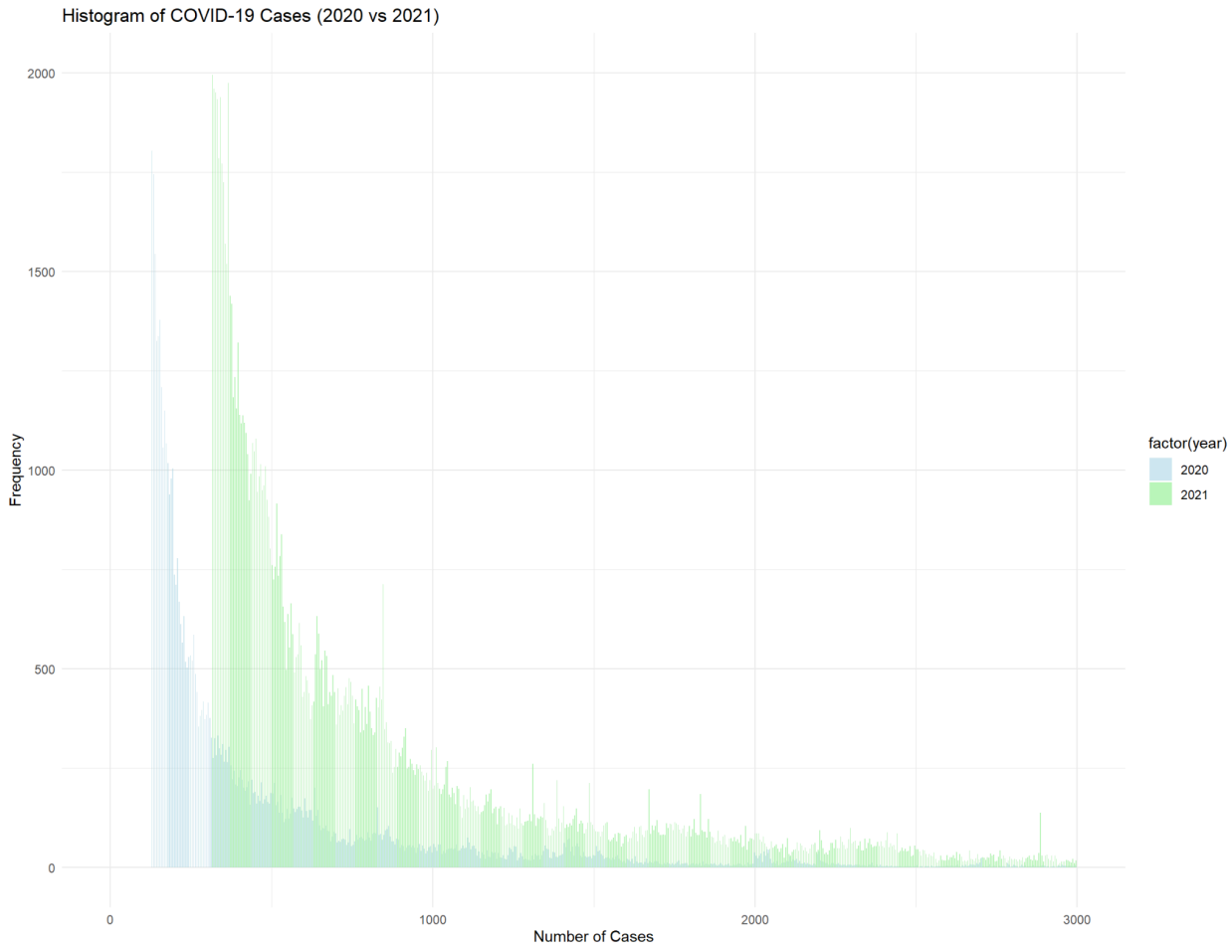
Deaths:

In 2020, many counties reported no deaths (minimum: 0, 1st quartile: 0), with a median of only 4. The mean of 53.6 and a maximum of 25,144 deaths highlight the concentration of high mortality in a few counties.

In 2021, deaths increased, with a median of 52 and a mean of 193.6, suggesting more widespread impact, while the maximum reached 35,382. The 1st quartile rising to 20 also indicates more counties were affected by deaths in 2021.

(b) Histograms


Histogram of COVID-19 Cases (2020 vs 2021)
Frequency
40000
20000
0
0
250
500
750
1000
Number of Cases
factor(year)
2020
2021



Code for Part (b)

```
clean_combined_data <- combined_data[complete.cases(combined_data), ]
clean_counties2021 <- counties2021[complete.cases(counties2021), ]
clean_counties2021 <- clean_counties2021[clean_counties2021$cases > 0 &
clean_counties2021$deaths > 0, ]

clean_counties2020 <- counties2020[complete.cases(counties2020), ]
clean_counties2020 <- clean_counties2020[clean_counties2020$cases > 0 &
clean_counties2020$deaths > 0, ]
```

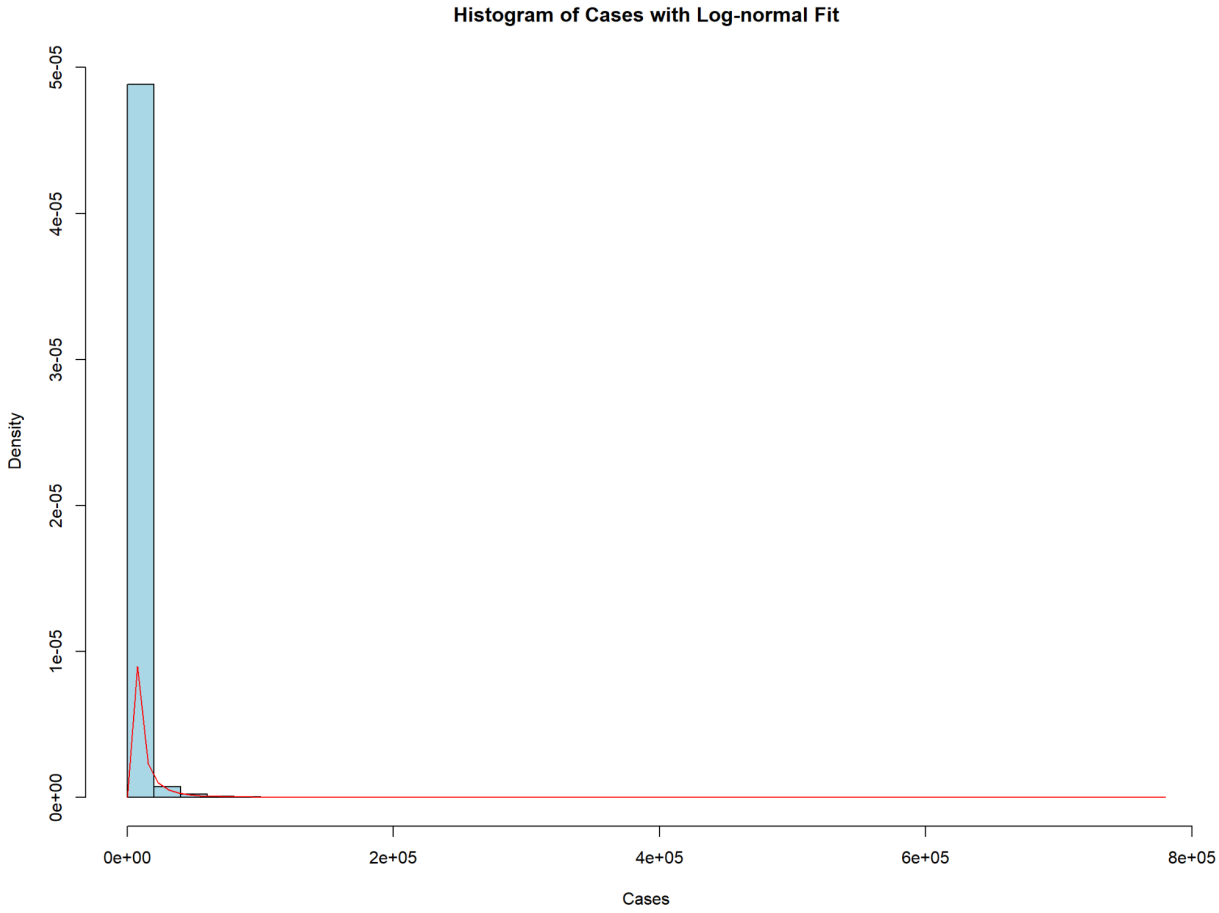
```
# Histograms for COVID-19 Cases in 2020 and 2021# counties2020Histograms for COVID-19 Cases in 2020 and 2021
ggplot(clean_combined_data, aes(x = cases ,fill = factor(year))) +
  geom_histogram(binwidth = 5, position = "dodge", alpha = 0.6) +
  xlim(0,1000)+
  labs(title = "Histogram of COVID-19 Cases (2020 vs 2021)",
       x = "Number of Cases",
       y = "Frequency") +
  scale_fill_manual(values = c("lightblue", "lightgreen")) +
  theme_minimal()

#histogram for deaths
ggplot(clean_combined_data, aes(x = deaths ,fill = factor(year))) +
  geom_histogram(binwidth = 5, position = "dodge", alpha = 0.6) +
  xlim(0,3000)+
  ylim(0,2000)+
  labs(title = "Histogram of COVID-19 Cases (2020 vs 2021)",
       x = "Number of Cases",
       y = "Frequency") +
  scale_fill_manual(values = c("lightblue", "lightgreen")) +
  theme_minimal()
```

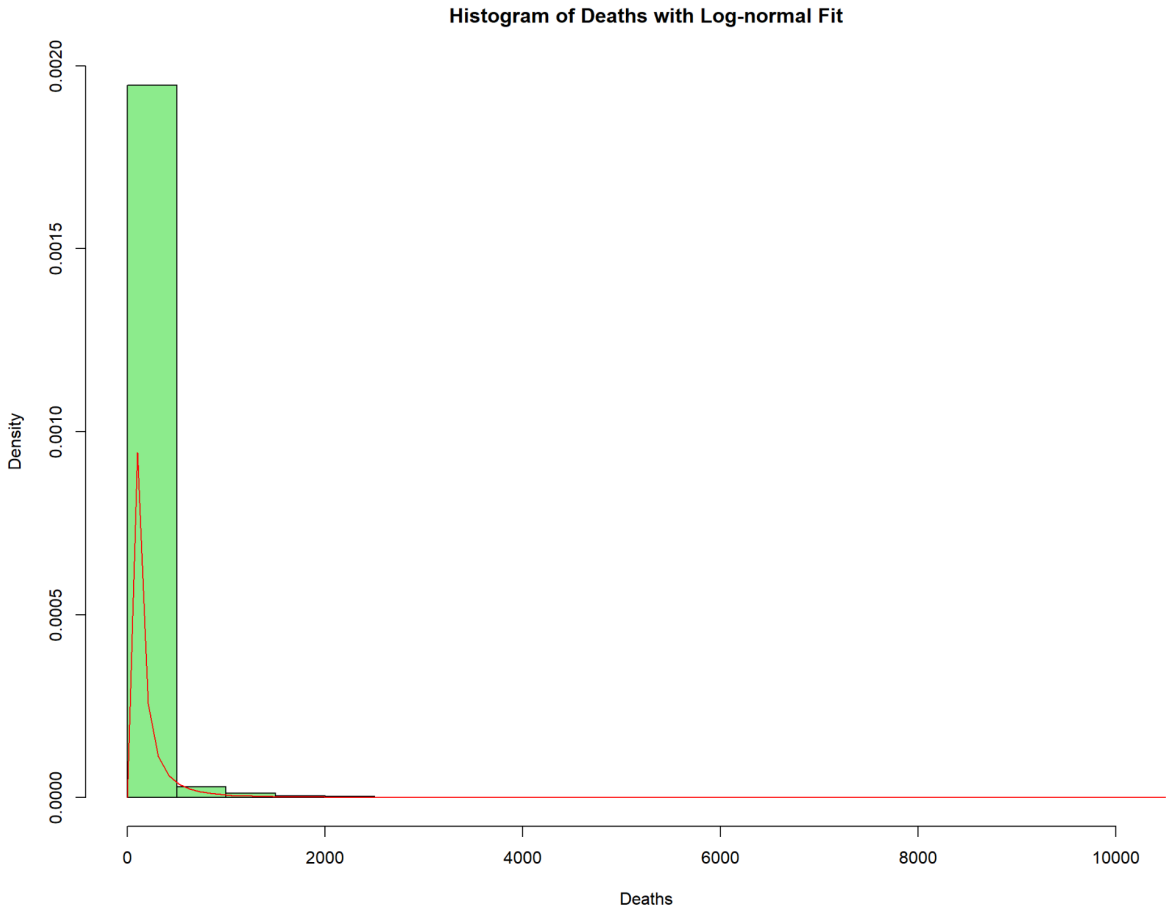
Applying Log-normal Distribution Fit model

```
# Fit Log-normal distribution to Cases for 2020
fit_cases_2020_lognormal <- fitdistr(clean_counties2020$cases, "lognormal")

# Fit Log-normal distribution to Deaths for 2020
fit_deaths_2020_lognormal <- fitdistr(clean_counties2020$deaths, "lognormal")

# Fit Log-normal distribution to Cases
fit_cases_2021_lognormal <- fitdistr(clean_counties2021$cases, "lognormal")

# Fit Log-normal distribution to Deaths
fit_deaths_2021_lognormal <- fitdistr(clean_counties2021$deaths, "lognormal")

dev.off()
# Plot histogram for 'cases' with Log-normal fit
hist(clean_counties2020$cases, probability = TRUE, breaks = 30, main = "Histogram of Cases
with Log-normal Fit", xlab = "Cases", col = "lightblue")
curve(dlnorm(x, meanlog = fit_cases_2020_lognormal$estimate["meanlog"], sdlog =
fit_cases_2020_lognormal$estimate["sdlog"]), col = "red", add = TRUE)

# Plot histogram for 'deaths' with Log-normal fit
```

hist(clean_counties2020$deaths, probability = TRUE, breaks = 30, main = "Histogram of Deaths with Log-normal Fit", xlab = "Deaths", col = "lightgreen")

curve(dlnorm(x, meanlog = fit_deaths_2020_lognormal$estimate["meanlog"], sdlog = fit_deaths_2020_lognormal$estimate["sdlog"]), col = "red", add = TRUE)





Explanation:

The distributions of COVID-19 cases and deaths in 2020 and 2021 seem to match a log-normal distribution because they are positively skewed, with most counties reporting lower values and a few counties having very high numbers. When we overlay the log-normal curve on the histograms, it fits well, showing the long tail and variability in the data.

The differences between 2020 and 2021, like the wider spread and higher peaks in 2021, suggest changes in how cases and deaths were distributed across counties over time.

(c) Plotting ECDFs

Code:

```
# ECDF for Cases

plot(ecdf(clean_counties2020$cases), main = "ECDF of Cases", xlab = "Cases", ylab = "ECDF",
col = "blue")



# ECDF for Deaths

plot(ecdf(clean_counties2020$deaths), main = "ECDF of Deaths", xlab = "Deaths", ylab =
"ECDF", col = "green")



# Q-Q plot for Cases with Log-normal distribution

qqnorm(clean_counties2020$cases, main = "Q-Q plot for Cases")



# Add Log-normal Q-Q line manually

meanlog <- fit_cases_2020_lognormal$estimate["meanlog"]

sdlog <- fit_cases_2020_lognormal$estimate["sdlog"]
```

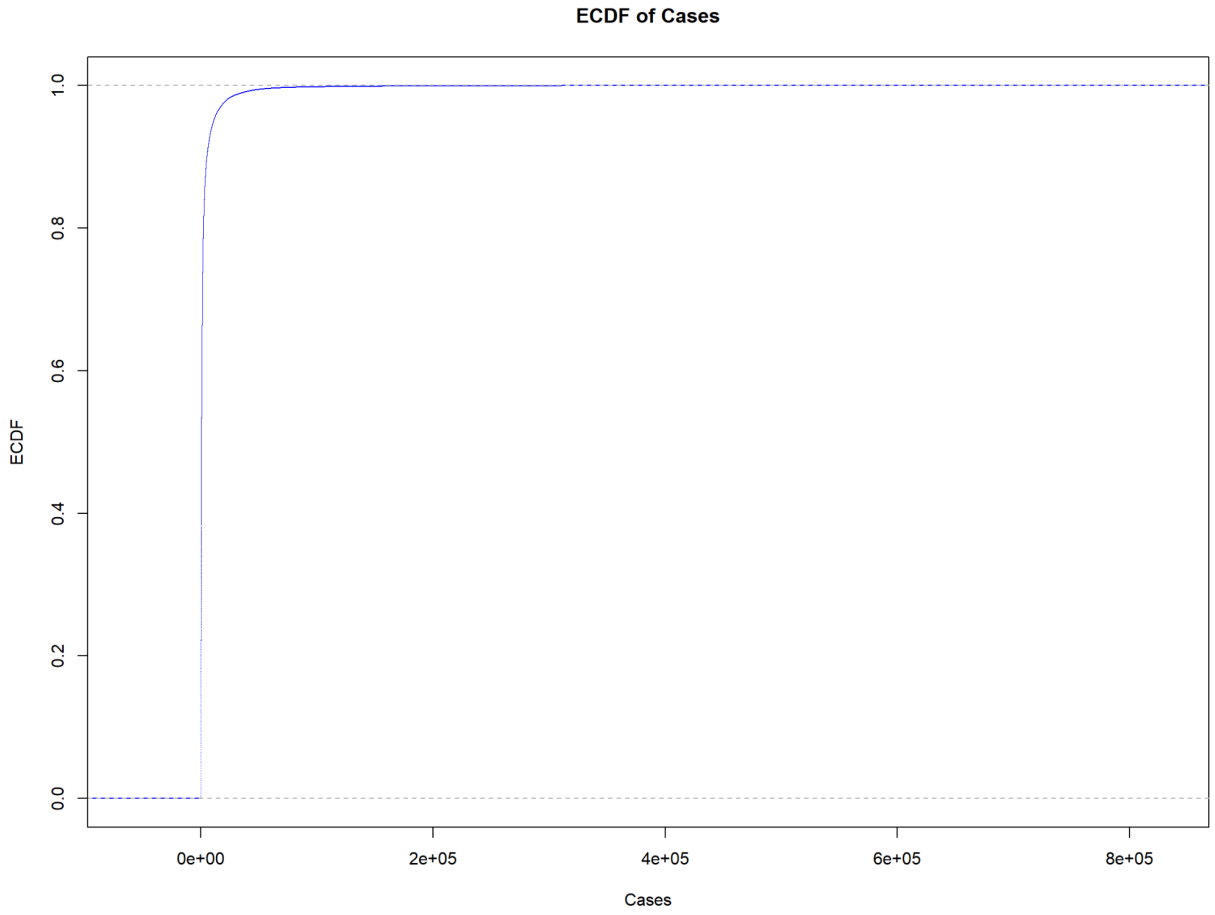
```
# Generate theoretical Log-normal quantiles

theoretical_quantiles <- qlnorm(ppoints(length(clean_counties2020$cases)), meanlog = meanlog,
sdlog = sdlog)



# Add the Log-normal Q-Q line

lines(sort(clean_counties2020$cases), theoretical_quantiles[order(clean_counties2020$cases)],
col = "red", lwd = 2)



# Q-Q plot for Deaths with Log-normal distribution

qqnorm(clean_counties2020$deaths, main = "Q-Q plot for Deaths")

# Add Log-normal Q-Q line manually

meanlog_deaths <- fit_deaths_2020_lognormal$estimate["meanlog"]

sdlog_deaths <- fit_deaths_2020_lognormal$estimate["sdlog"]



# Generate theoretical Log-normal quantiles

theoretical_quantiles_deaths <- qlnorm(ppoints(length(clean_counties2020$deaths)), meanlog =
meanlog_deaths, sdlog = sdlog_deaths)
```

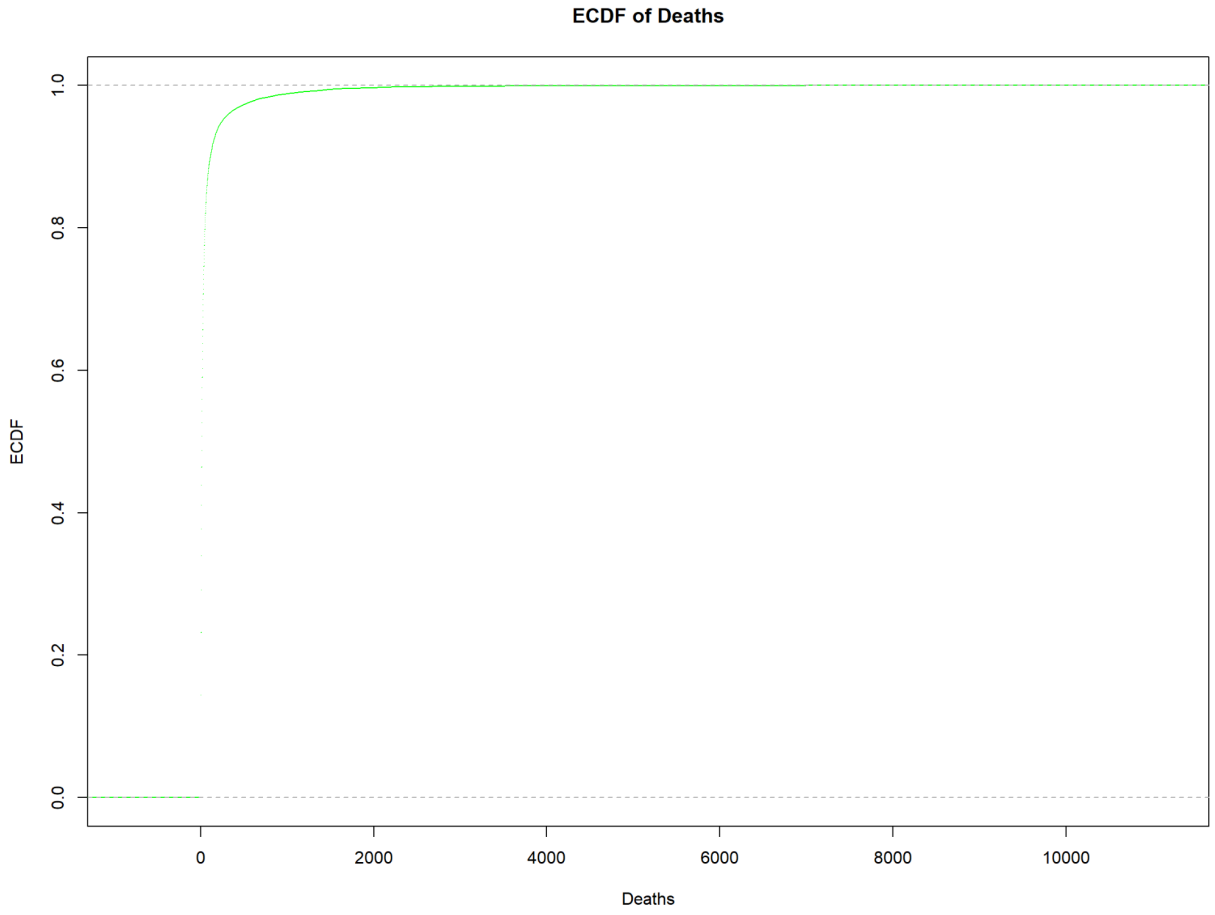
Add the Log-normal Q-Q line

lines(sort(clean_counties2020$deaths),

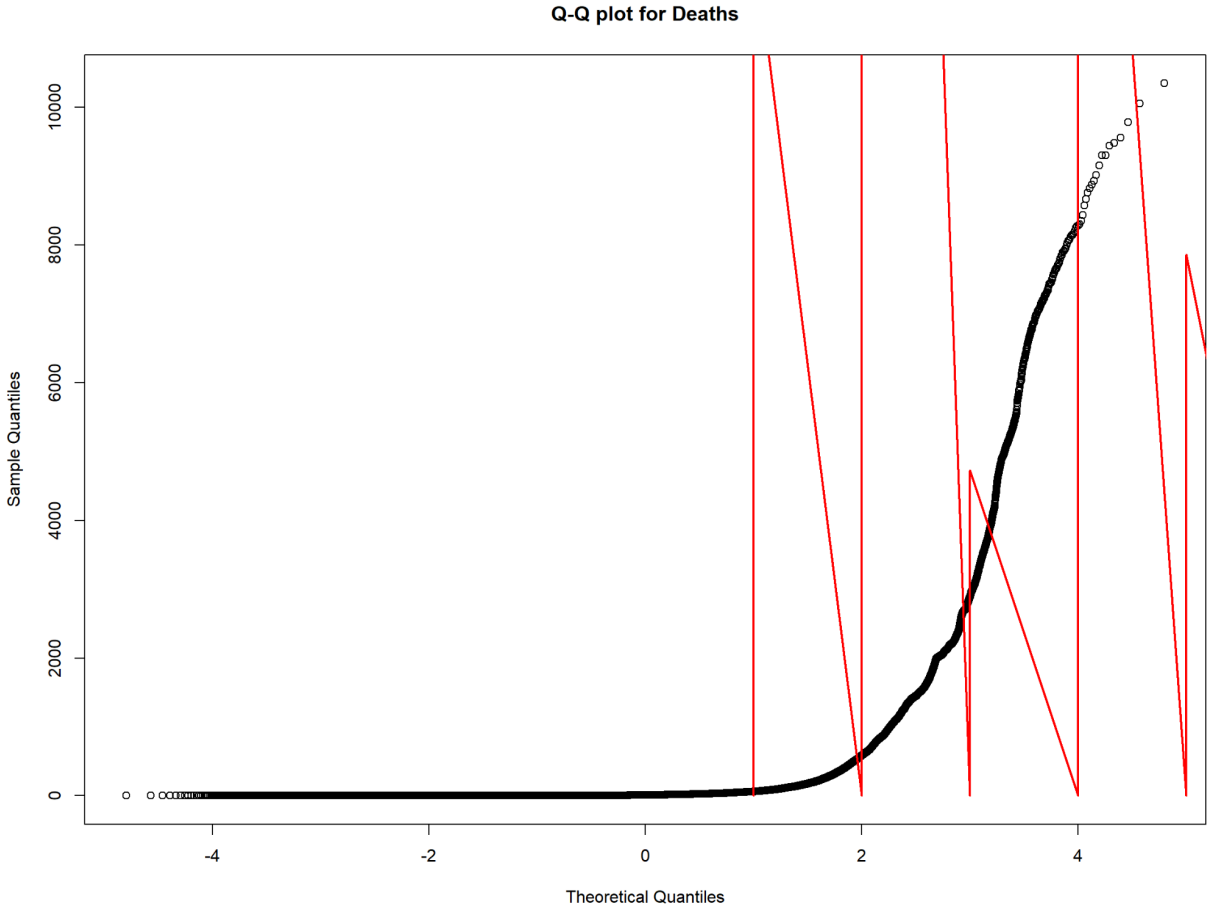
theoretical_quantiles_deaths[order(clean_counties2020$deaths)], col = "red", lwd = 2)

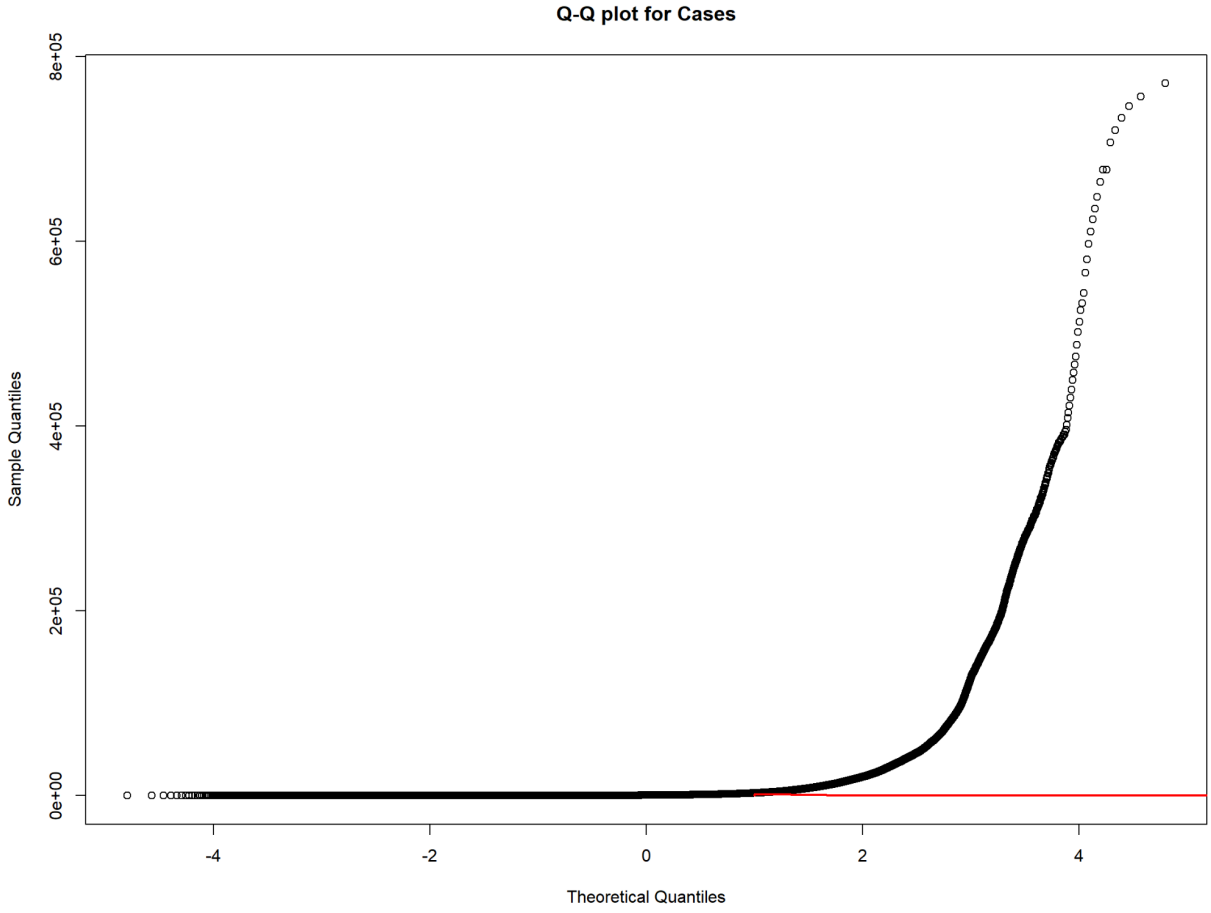
ECDF Plots:




ECDF of Deaths
ECDF
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Deaths

Q-Q Plots:


Q-Q plot for Deaths
Sample Quantiles
Theoretical Quantiles
0
2000
4000
6000
8000
10000
-4
-2
0
2
4



The ECDF (Empirical Cumulative Distribution Function) plots for COVID-19 cases and deaths provide insight into the cumulative distribution of the data. The ECDF plots for 2020 and 2021 cases show that the majority of counties fall within lower case counts, with steep initial rises indicating concentrated data in the lower range. However, the curves level off more gradually in 2021 compared to 2020, reflecting the higher range of case counts in 2021. Similarly, the ECDF for deaths shows a comparable pattern but with smaller overall values, indicating the more limited distribution of death counts.

The Q-Q (Quantile-Quantile) plots assess how well the data aligns with a theoretical log-normal distribution. For both cases and deaths, the points follow the 45-degree line closely in the lower

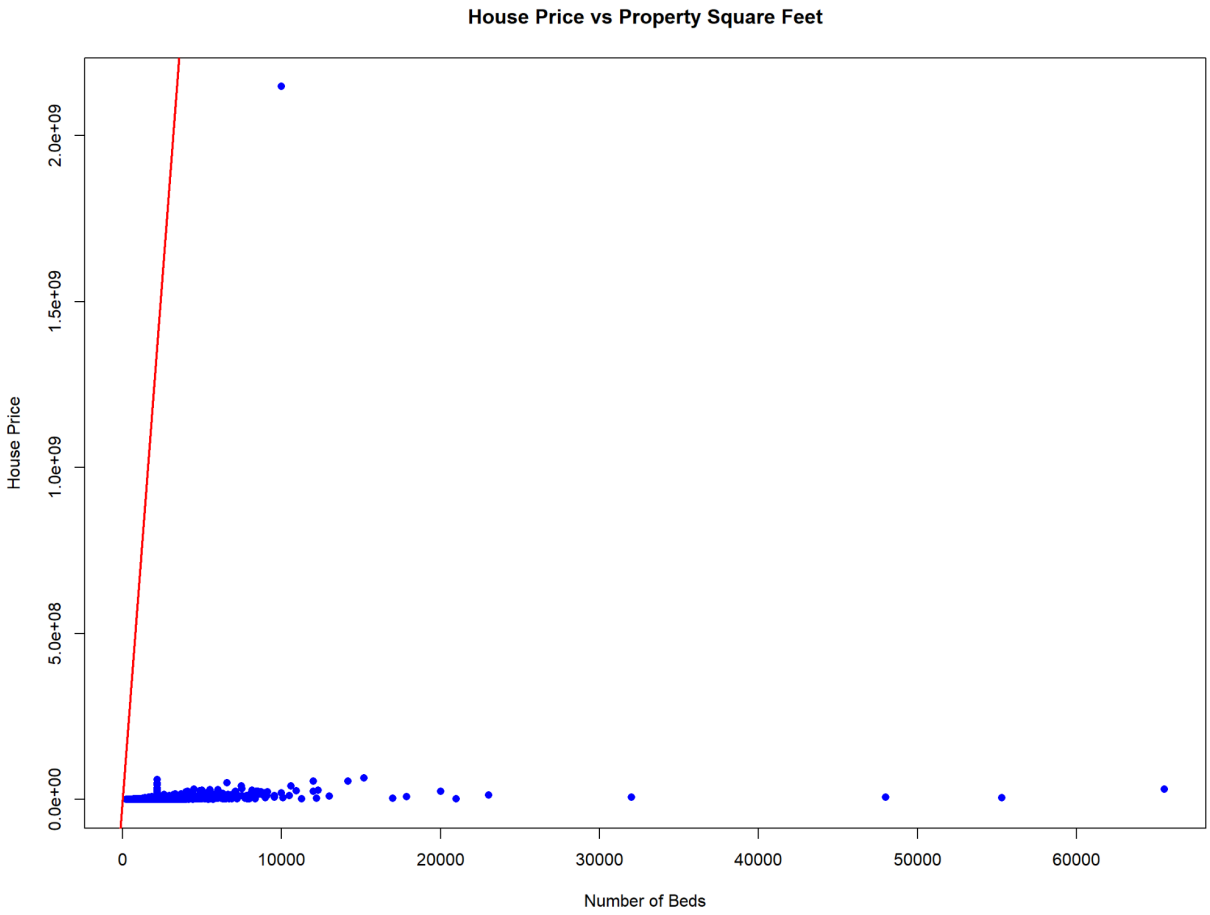
quantiles, indicating a good fit in these ranges. However, deviations occur in the upper tails, with points diverging above the line, reflecting the influence of extreme values or outliers in the data. This suggests that while the log-normal distribution approximates the data well, it may not fully capture the heavy-tailed nature of these distributions.

Question 2 NY Houses Data Set

(a)

```
ny_house <- read.csv("NY-House-Dataset.csv")

head(ny_house)

summary(ny_house)



#cleaning the data

ny_house_clean <- ny_house %>%

  filter(!is.na(PRICE) & !is.na(BEDS) & !is.na(BATH) & !is.na(PROPERTYSQFT),

         PRICE > 0, BEDS < 10, BATH < 10, PROPERTYSQFT > 0)

#fitting linear model

model_full <- lm(PRICE ~ BEDS + BATH + PROPERTYSQFT, data = ny_house_clean)

summary(model_full)
```

```
#scatter plot with a best fit line

plot(ny_house_clean$BEDS, ny_house_clean$PRICE,

     main = "House Price vs Beds",

     xlab = "Number of Beds", ylab = "House Price",

     col = "blue", pch = 16)

abline(lm(PRICE ~ BEDS, data = ny_house), col = "red", lwd = 2)



#plotting the residuals

plot(model_full$residuals,

     main = "Residual Plot",

     ylab = "Residuals", xlab = "Index",

     col = "darkgreen", pch = 16)

abline(h = 0, col = "red", lwd = 2)
```



We can see in the first scatter plot that the most influencing house price is the dot that appears towards the left side of the graph in the upper left quadrant with a price near 2.5E09. We can consider this point an outlier. Determined by the formula given, the variable that influences the house price variable the most are both BATH and PROPERTY SQFT.

The formula given results in:

Call:

lm(formula = PRICE ~ BEDS + BATH + PROPERTYSQFT, data = ny_house_clean)

Residuals:

Min	1Q	Median	3Q	Max
-59664190	-1466832	-246508	955234	2129687042

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3675591	1001121	-3.671	0.000244	***
BEDS	-182460	372891	-0.489	0.624644	
BATH	1820745	495951	3.671	0.000244	***
PROPERTYSQFT	1182	228	5.187	2.23e-07	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31340000 on 4687 degrees of freedom

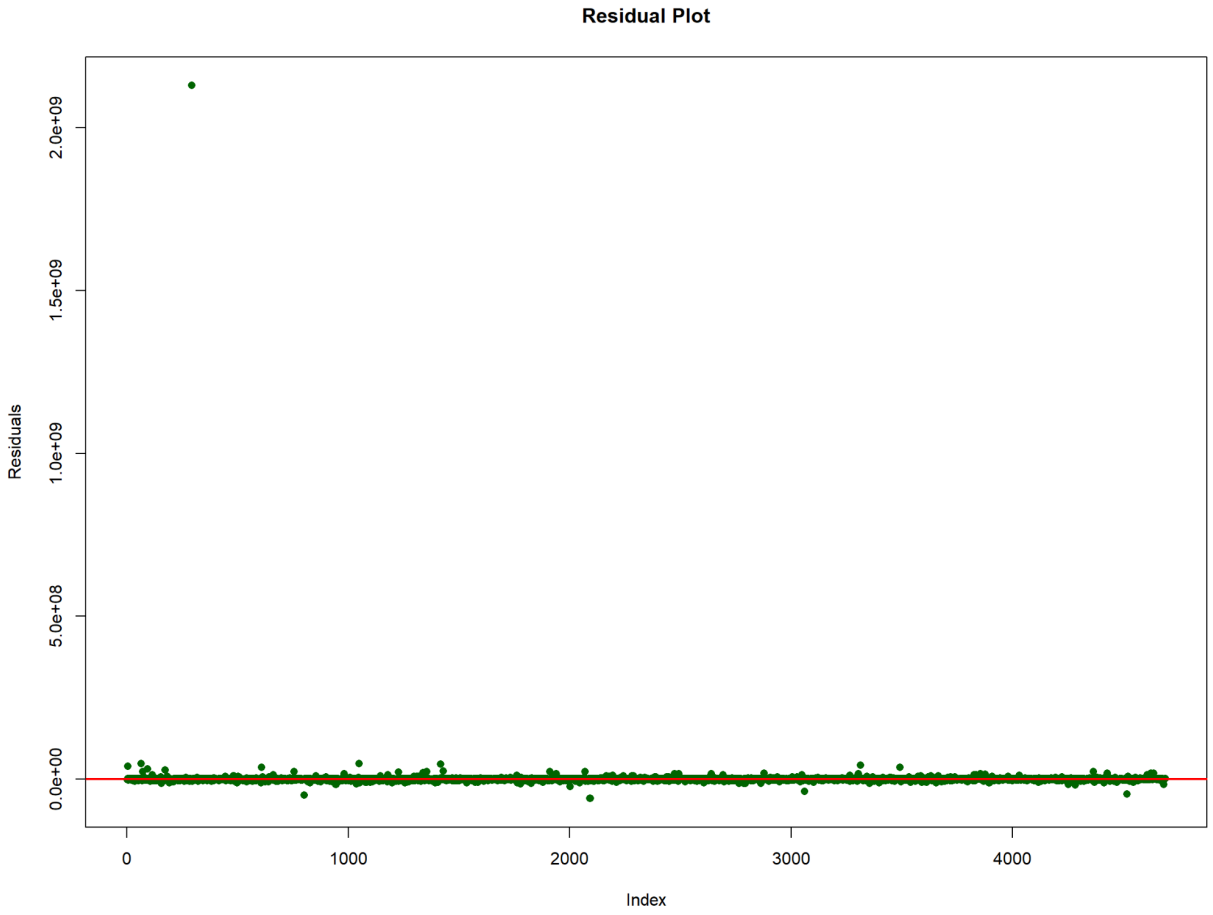
Multiple R-squared: 0.01495, Adjusted R-squared: 0.01432

F-statistic: 23.72 on 3 and 4687 DF, p-value: 3.142e-15

BATH has a significant positive coefficient (1,820,745) with a **p-value** of 0.000244, indicating that more bathrooms significantly increase house price.

PROPERTYSQFT has an even stronger significant effect (coefficient = 1,182), with a **very small p-value** (2.23e-07), suggesting that the size of the property (in square feet) has a strong positive relationship with house price.

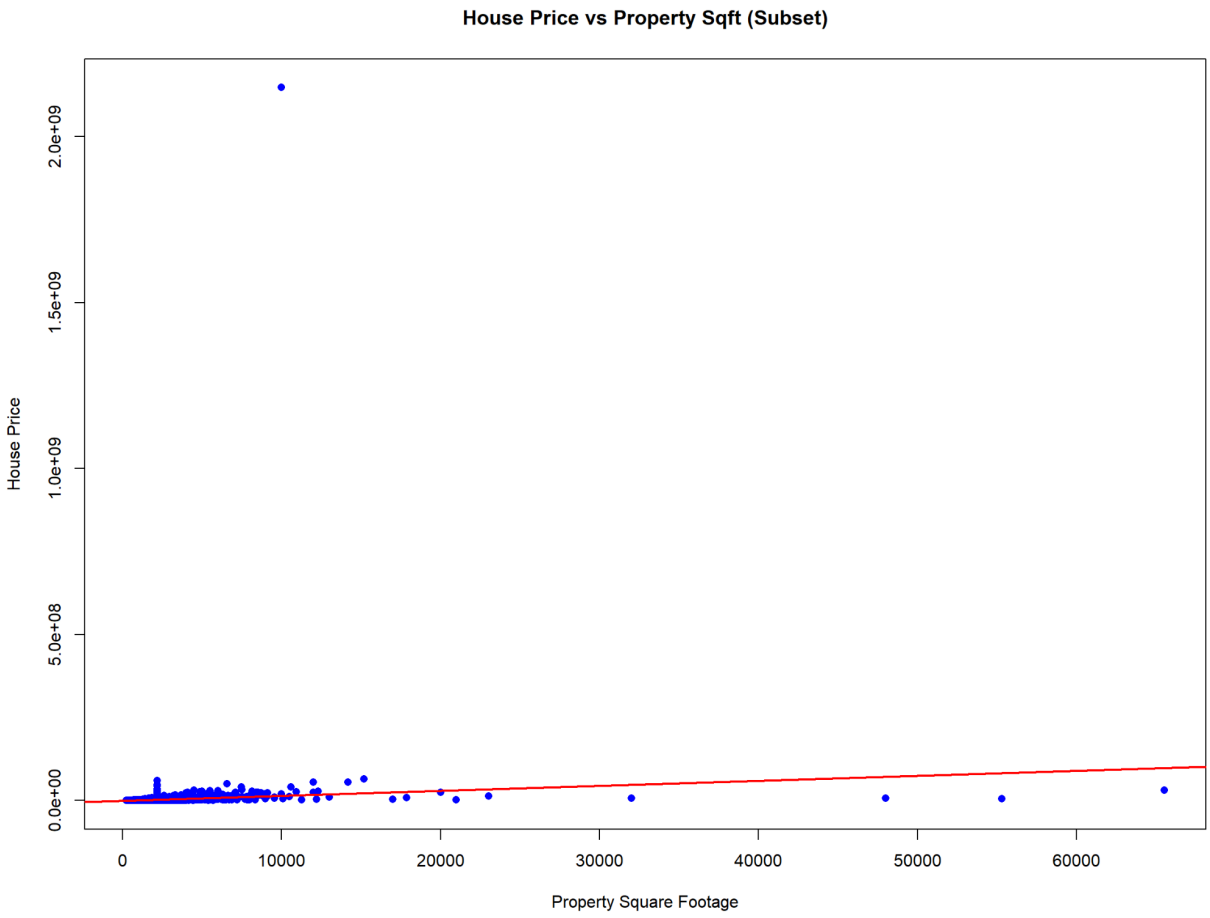
Plotting the residuals

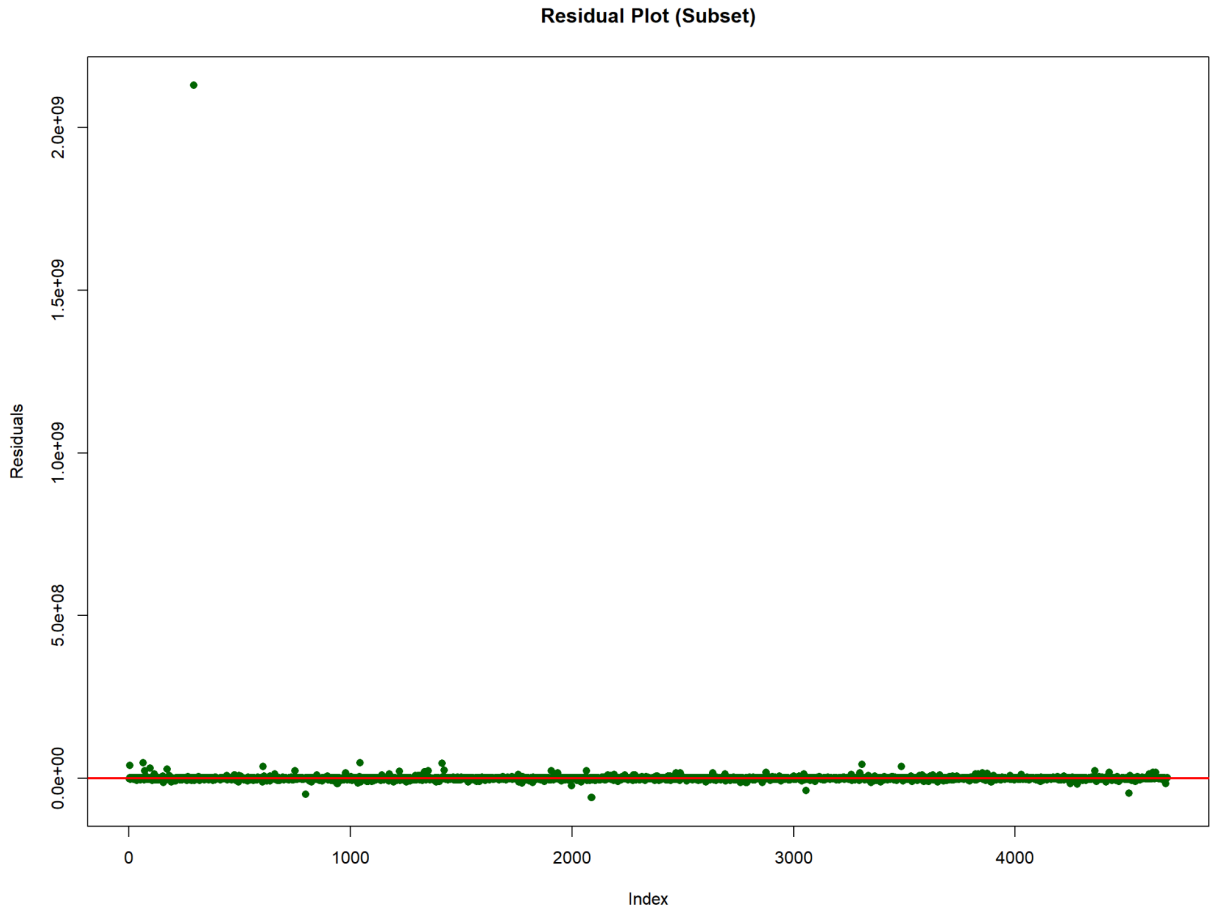


(b) Repeating the Plots with a subset of PRICE >50000

```
ny_house_subset <- subset(ny_house_clean  , PRICE > 50000)

#fitting the new linear model

model_subset <- lm(PRICE ~ BEDS + BATH + PROPERTYSQFT, data = ny_house_subset)

summary(model_subset)



#repeating scatter plots and residuals

plot(ny_house_subset$PROPERTYSQFT, ny_house_subset$PRICE,

    main = "House Price vs Property Sqft (Subset)",

    xlab = "Property Square Footage", ylab = "House Price",

    col = "blue", pch = 16)

abline(lm(PRICE ~ PROPERTYSQFT, data = ny_house_subset), col = "red", lwd = 2)



# Residuals

plot(model_subset$residuals,

    main = "Residual Plot (Subset)",

    ylab = "Residuals", xlab = "Index",
```

```
    col = "darkgreen", pch = 16)

abline(h = 0, col = "red", lwd = 2)
```





Summary of the subset model:

Call:

lm(formula = PRICE ~ BEDS + BATH + PROPERTYSQFT, data = ny_house_subset)

Residuals:

Min	1Q	Median	3Q	Max
-59662878	-1467868	-247482	961234	2129686465

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3674531.5	1001806.1	-3.668	0.000247	***
BEDS	-182557.1	373076.0	-0.489	0.624631	
BATH	1820847.5	496285.8	3.669	0.000246	***
PROPERTYSQFT	1182.5	228.1	5.184	2.26e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31360000 on 4683 degrees of freedom

Multiple R-squared: 0.01495, Adjusted R-squared: 0.01432

F-statistic: 23.69 on 3 and 4683 DF, p-value: 3.256e-15

In the linear model for the subset of the dataset, the significance of the input variables has not drastically changed. The **BEDS** variable remains statistically insignificant with a high p-value (0.6246), which indicates that the number of bedrooms does not have a meaningful relationship

with house price in this subset. However, **BATH** and **PROPERTYSQFT** maintain their significance, with both having p-values well below 0.05. This suggests that the number of bathrooms and the property's square footage are still strong predictors of house price, even within the subset of data. The adjusted R-squared remains low (0.01432), indicating that the model does not explain a large portion of the variation in house prices.