SARB Analysis

Buhlali Mhlanzi

2025-11-04

This analysis focuses on the Rand per US Dollar published by the South African Reserve Bank (SARB).

The Rand per US Dollar column gives us the "Rand per US Dollar, Weighted average of the banks`daily Rand per US Dollars at approximately 10:30 am. Weights are based on the banks` foreign exchange transactions." as provided at https://www.resbank.co.za/en/home/what-we-do/statistics/key-statistics/selected-historical-rates, on the South African Reserve Bank website.
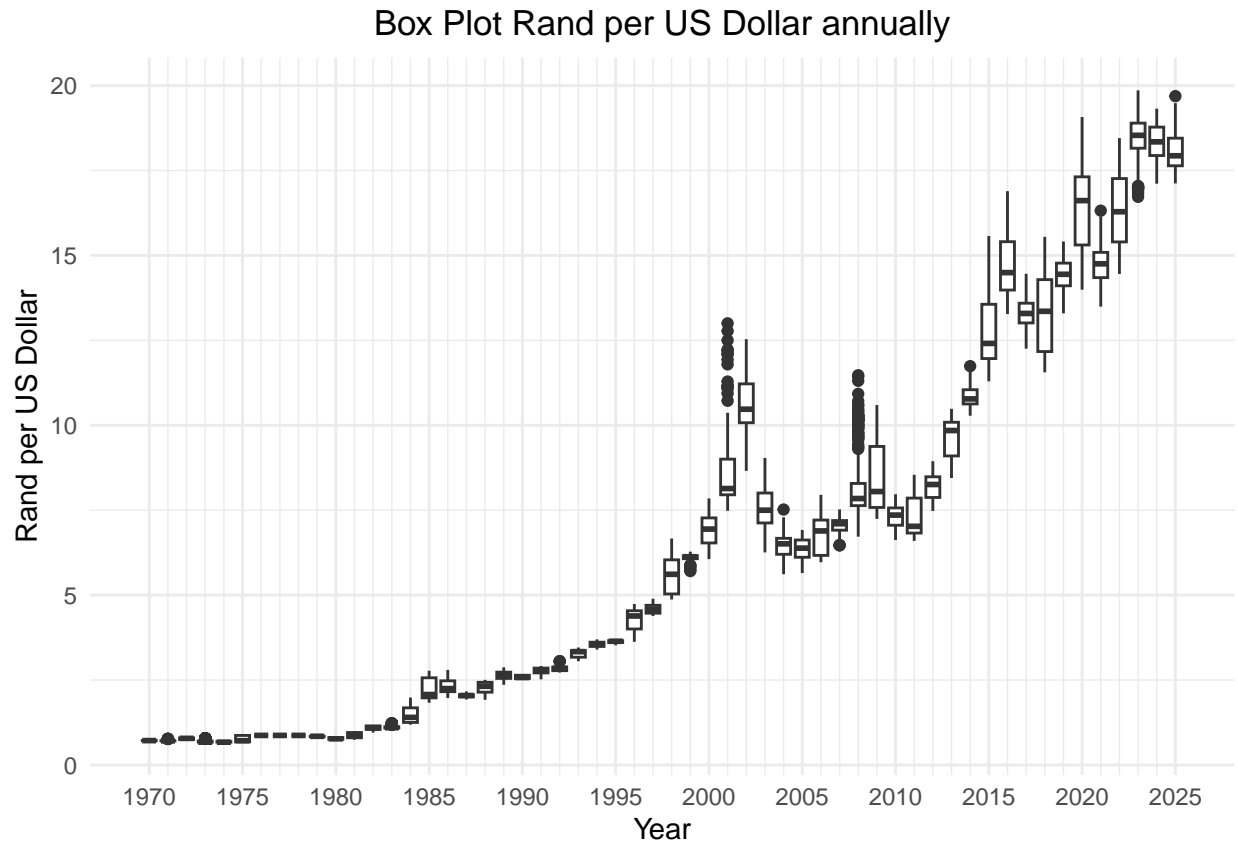
1. Basic analysis:

The first levels of analysis, will be a summary of the overall data, then annually.

```
summary(data$`Rand per US Dollar`) |> as.table()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6667  1.1950  4.8135  6.2910  9.3878 19.8618
```

```
data$Year<- as.numeric(data$Year)
ggplot(data) +
  geom_boxplot(aes(x = Year, y = `Rand per US Dollar`, group = Year)) +
  scale_x_continuous(
    breaks = seq(min(sort(unique(data$Year))), max(sort(unique(data$Year))), by = 5),
    minor_breaks =sort(unique(data$Year))
  ) +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x = "Year", y = "Rand per US Dollar", title = "Box Plot Rand per US Dollar annually")
```

Box Plot Rand per US Dollar annually

In the above plot we pick up on some general trends:

1. There is some general upward trend.

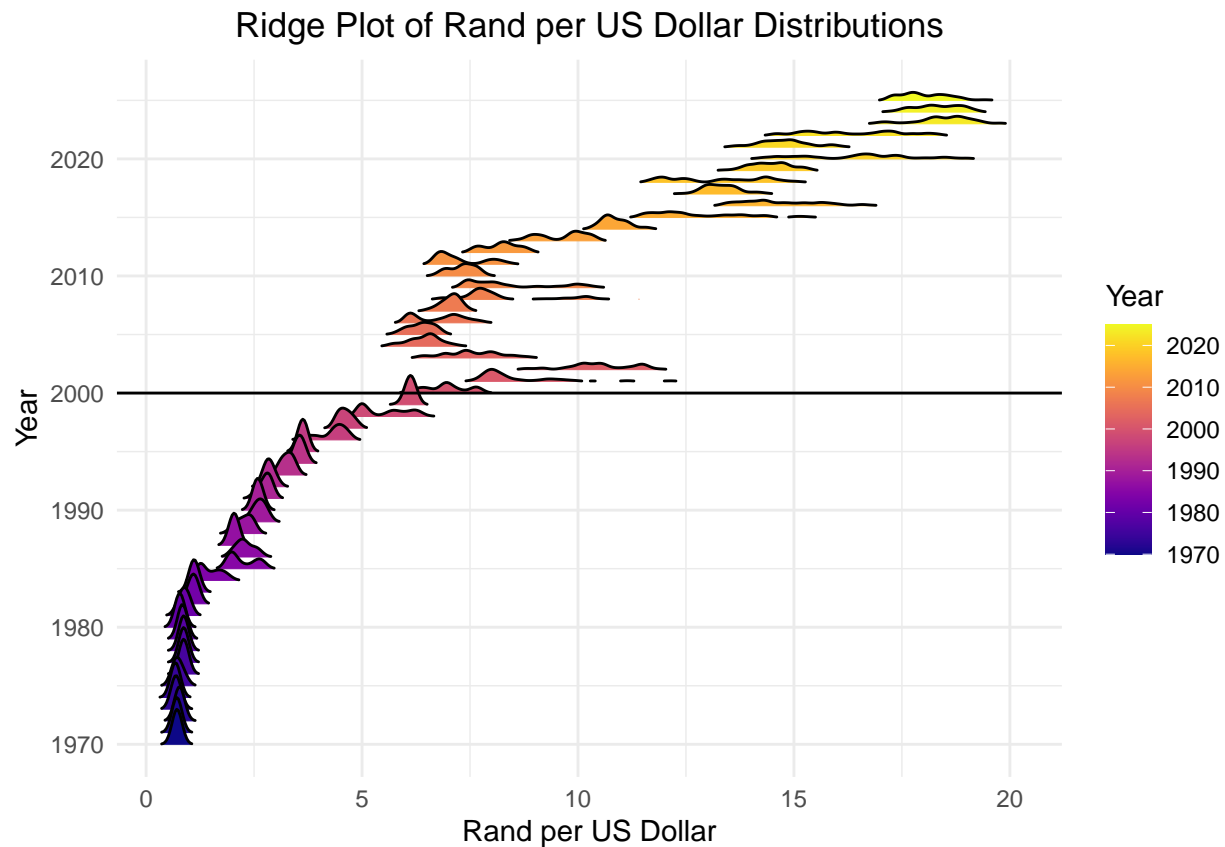2. The volatility is an increasing with time

From 1), we see as time passes, our Rand per US Dollar is increasing, this makes sense as we would expect the Rand per US Dollar to increase with time. In 2), we see increasing volatility with time. The volatility seems to begin showing drastic volatility from about 2000.

1.1. Volatility Investigation:

1.2. Graphic analysis:

```
ggplot(data, aes(x = `Rand per US Dollar`, y = Year, fill = Year, group = Year)) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
  scale_fill_viridis(name = "Year", option = "C")+theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5)) +geom_hline(yintercept = 2000)+
  labs(x ="Rand per US Dollar", y = "Year", title = "Ridge Plot of Rand per US Dollar Distributions")
```

```
## Picking joint bandwidth of 0.116
```

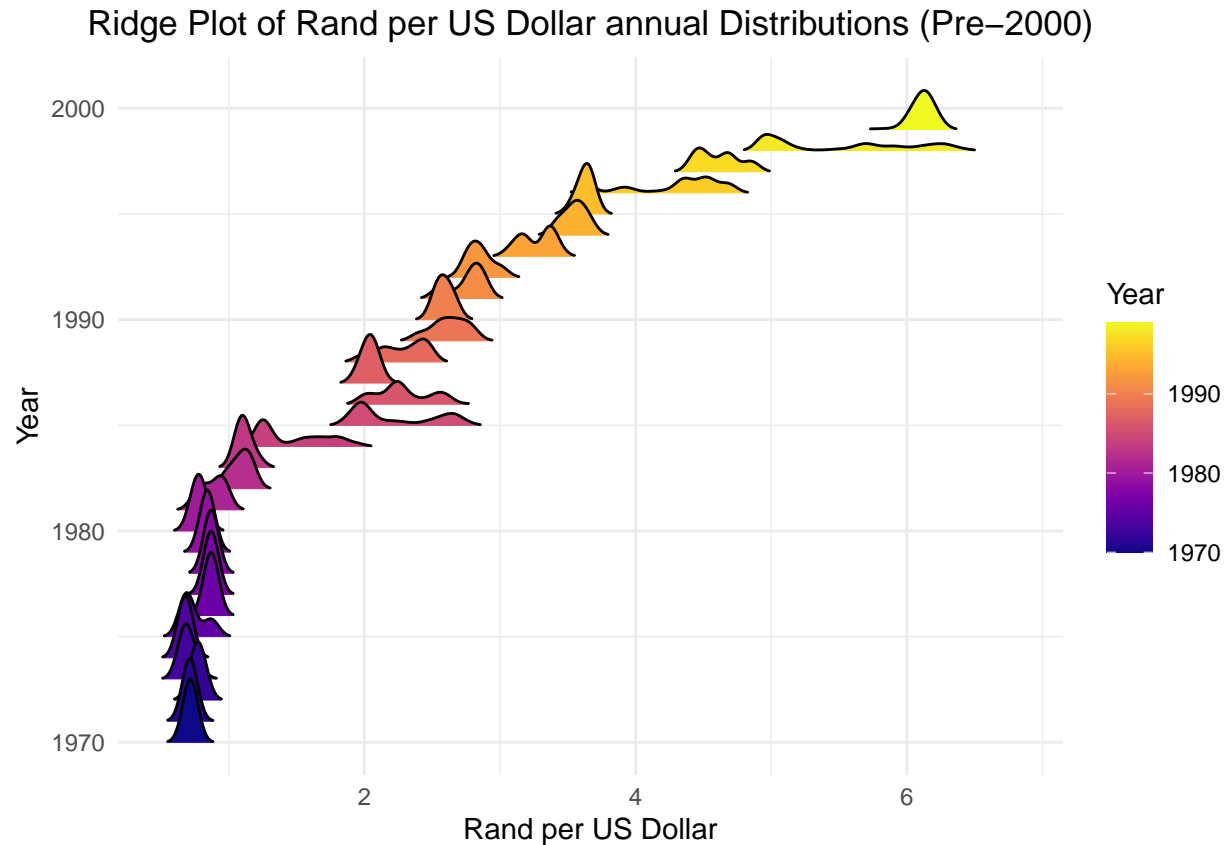## Ridge Plot of Rand per US Dollar Distributions

The graph above is consistent with what we notice, the volatility drastically increases after the year 2000. The distributions also seem to take very different shapes before and after this point (the year 2000). Next we analyse the partitions.

Pre-2000

```r
Part1 <- subset(data, Year < 2000)
ggplot(Part1, aes(x = `Rand per US Dollar`, y = Year, fill = Year, group = Year)) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
  scale_fill_viridis(name = "Year", option = "C") +theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x = "Rand per US Dollar", y = "Year", title = "Ridge Plot of Rand per US Dollar annual Distributi
```

```
## Picking joint bandwidth of 0.0558
```

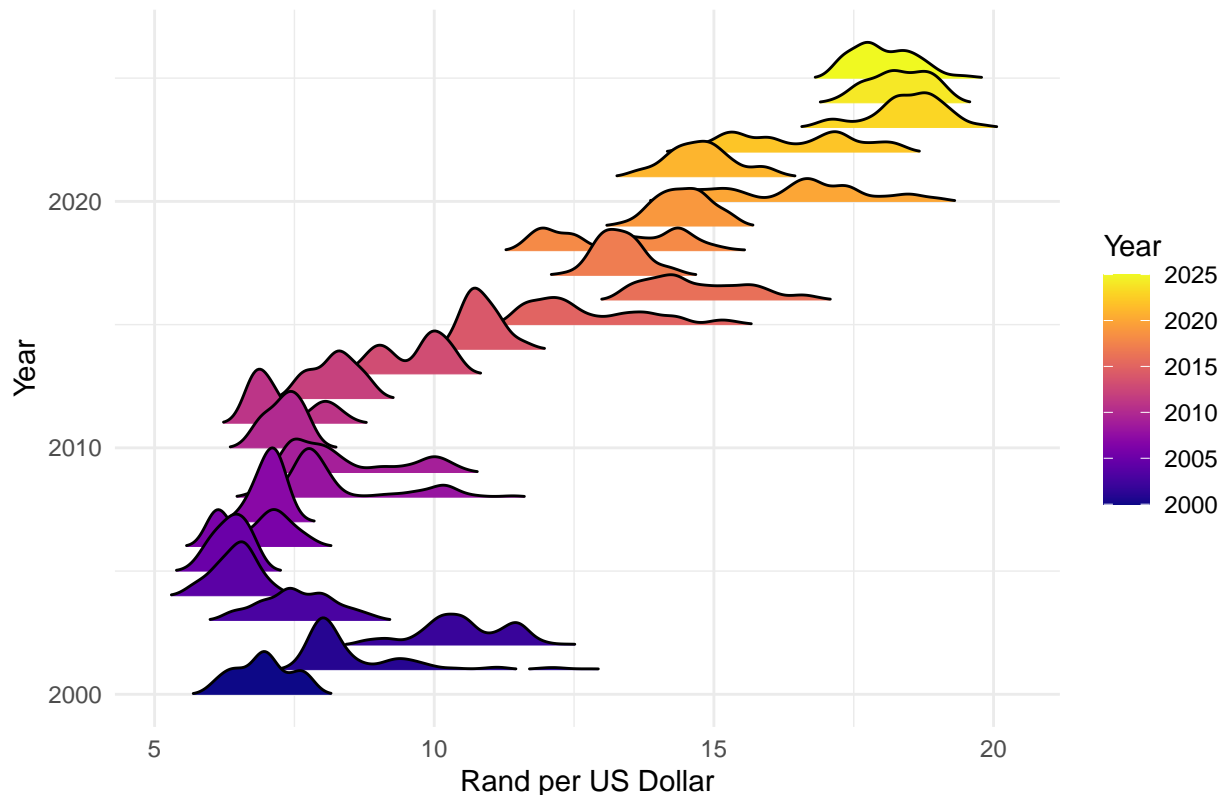## Ridge Plot of Rand per US Dollar annual Distributions (Pre−2000)

From years 1970, we see a very similar distribution till about 1980, from there our variance increases and the distributions become progressively more heavy tailed, even shifting to bi-modal distributions. This holds relevance since it shows the Rand per US Dollar indeed seems to be an increasingly varying measure, and seeing a bi-modal distribution, shows the value does not just vary, but it does not show "clumping" around only one point.

Post-2000

```
Part2 <- subset(data, Year >= 2000)
ggplot(Part2, aes(x = `Rand per US Dollar`, y = Year, fill = Year, group = Year)) +
  geom_density_ridges_gradient(scale = 3, rel_min_height = 0.01) +
  scale_fill_viridis(name = "Year", option = "C") +theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))+
  labs(x = "Rand per US Dollar", y = "Year", title = "Ridge Plot of Rand per US Dollar annual Distributi
```

```
## Picking joint bandwidth of 0.185
```

Ridge Plot of Rand per US Dollar annual Distributions (Post−2000)

The pattern of drastic variance seems to not be completely consistent, if we look at the distributions from 2000 to 2010, we see positively skewed long tails, then from 2010 to about 2015, the distributions still show large variance, but they seem notably less varied.

Now the observation made at the beginning of section 1.1, still holds, we see distributions with greater variance Post-2000, compared to Pre-2000.

2. Mean Groupings

Notice in our graphs, we see periods of increase in Rand per US Dollar Rand per US Dollars, where our distributions shift rapidly from year to year, but we also see periods of plateau, where the mean seems somewhat constant. To investigate this, we will use One-way ANOVA, paired with Scheffe groupings test.

2.1. One-Way ANOVA with Scheffe Grouping

```
ANOVA_data <- data
ANOVA_data$Year <- as.factor(ANOVA_data$Year)
fitAN <- aov(`Rand per US Dollar` ~ Year, data = ANOVA_data)
scheffe <- ScheffeTest(fitAN)
pvals <- scheffe$Year[,"pval"]
cld <- multcompLetters(pvals, threshold = 0.05)
ANOVA_data$Group <- cld$Letters[as.character(data$Year)]
df <- data.frame(Year = names(cld$Letters),Group = cld$Letters,stringsAsFactors = FALSE)
```

```
group_table <- group_by(df, Group) |> summarise(Years = paste(Year, collapse = ", "), .groups = "drop")
group_table <- group_table |> arrange(Group)
kable(group_table, booktabs = TRUE, longtable = TRUE) |>
  kable_styling(latex_options = c("hold_position"))
```

Group	Years
a	1973, 1974
ab	1971, 1970
abc	1972, 1975, 1976, 1977, 1978, 1979, 1980, 1981
bc	1982
cd	1983
d	1984
e	1987
ef	1985
efg	1986, 1988
fgh	1990
gh	1989
h	1991, 1992
i	1993, 1994, 1995
j	1996, 1997
k	1998
l	1999, 2005
lr	2004
mn	2000, 2007
mr	2006
nq	2010, 2011
o	2001, 2008, 2009, 2012
p	2002, 2014
q	2003
s	2013
t	2015
u	2016, 2019, 2021
v	2017, 2018
w	2020, 2022
x	2023
xy	2024
y	2025

The table shows us, there are indeed periods of "plateau", where the Rand per US Dollar means do not significantly differ, however, many years fall into their own group, these years would represent the years where we are seeing rapid change.
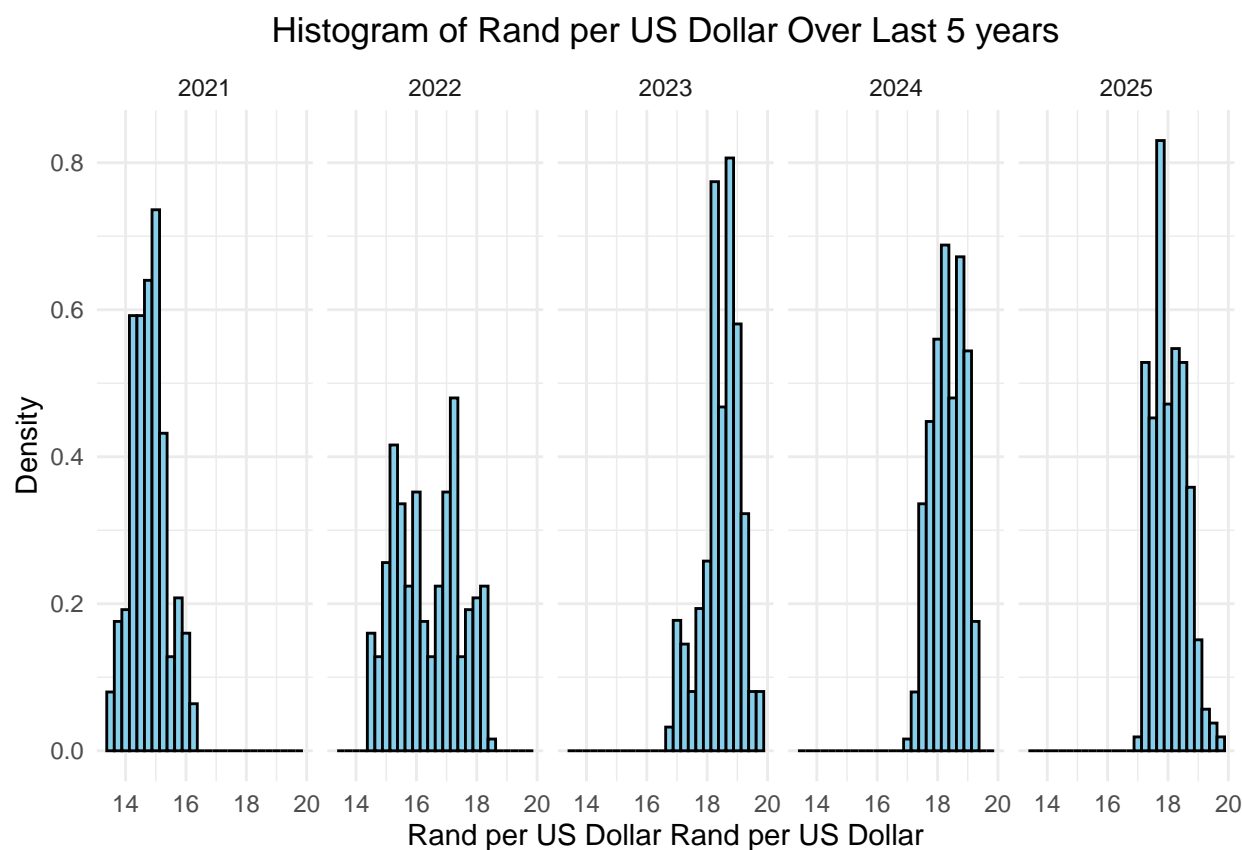
3. Modeling

Now we are going to focus on essentially adding statistics to our observations, by answering, what distributions have we been seeing, what is the underlying process we see in the data?

3.1 Annual Rand per US Dollar distribution

For this section, we will focus in on the last 5 years data, this analysis could be extended to include more years.

```
data$Year <- data$Year |> as.numeric()
Part1 <- subset(data, Year >= 2021)
Hist2023 <- ggplot(Part1, aes(x = `Rand per US Dollar`)) +
  geom_histogram(aes(y = after_stat(density)), binwidth = 0.25, fill = "skyblue", color = "black") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(x = "Rand per US Dollar Rand per US Dollar", y = "Density", title = "Histogram of Rand per US Do:
Hist2023
```

From our histogram, we see a light tailed distributions, with generally one mode. Some suggestions are the Exponential, or Weibull Distribution. Another important observation is the increasing trend, our datas cluster seems to be shifting with time, in 2021, we see the cluster around 14 and 15, then 2022, we see the cluster around 16 and 17, and so on.

3.1. Fitting Exponential and Weibull distributions

From inspection, I propose Exponential and Weibull distributions, we will fit these distributions and overlay the distributions and perform a Chi-Squared Goodness Of Fit test. Now the issue is our data does not start at 0, our observations have a minimum value of at least 12, so I will need to truncate the distributions to effectively fit the distribution. This will be done for the 2024 data, but can be extended for as many years necessary.

```
fit2024 <- subset(data, Year == 2024)
summary(fit2024$`Rand per US Dollar`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.11   17.94   18.35   18.33   18.78   19.32
```

From this, I will truncate the data with a lower bound of 15.

```
dexp_trunc <- function(x, RpUD, a = 15) {
  ifelse(x < a, 0,RpUD * exp(-RpUD * (x - a)))
}
pexp_trunc <- function(q, RpUD, a = 15) {
  F_a <- pexp(a, RpUD = `Rand per US Dollar`)
  F_q <- pexp(q, RpUD = `Rand per US Dollar`)
  ifelse(q < a, 0,(F_q - F_a) / (1 - F_a))
}
dweibull_trunc <- function(x, shape, scale, a = 15) {
  fx <- dweibull(x, shape = shape, scale = scale)
  S_a <- 1 - pweibull(a, shape = shape, scale = scale)
  ifelse(x < a, 0, fx / S_a)
}
pweibull_trunc <- function(q, shape, scale, a = 15) {
  F_a <- pweibull(a, shape = shape, scale = scale)
  F_q <- pweibull(q, shape = shape, scale = scale)
  ifelse(q < a, 0,(F_q - F_a) / (1 - F_a))
}
```

Now we fit the proposed distributions.

```
fit_exp <- fitdist(fit2024$`Rand per US Dollar`, "exp_trunc",start = list(RpUD = 0.5))
fit_wei <- fitdist(fit2024$`Rand per US Dollar`, "weibull_trunc",start = list(shape = 1, scale = 20))
```

```
ggplot(fit2024, aes(x = `Rand per US Dollar`)) +
  geom_histogram(aes(y = ..density..),
                 fill = "lightblue",alpha = 0.6) +

  stat_function(aes(color = "Truncated Weibull"),
    fun = dweibull_trunc,args = list(shape = fit_wei[["estimate"]][["shape"]],
                                     scale = fit_wei[["estimate"]][["scale"]],
                                     a = 15),
    size = 1) +

  stat_function(aes(color = "Truncated Exponential"),
    fun = dexp_trunc,args = list(RpUD = fit_exp[["estimate"]][["RpUD"]],a = 15),size = 1) +

  scale_color_manual(name = "Fitted Models",
    values = c("Truncated Weibull" = "red",
               "Truncated Exponential" = "purple")) +theme_minimal()+
  labs(x = "Rand per US Dollar",
       y = "Density",
       title = "Histogram of 2024 Rand per US Dollar Overlayed by Proposed Densities")+
  theme(plot.title = element_text(hjust = 0.5))
```
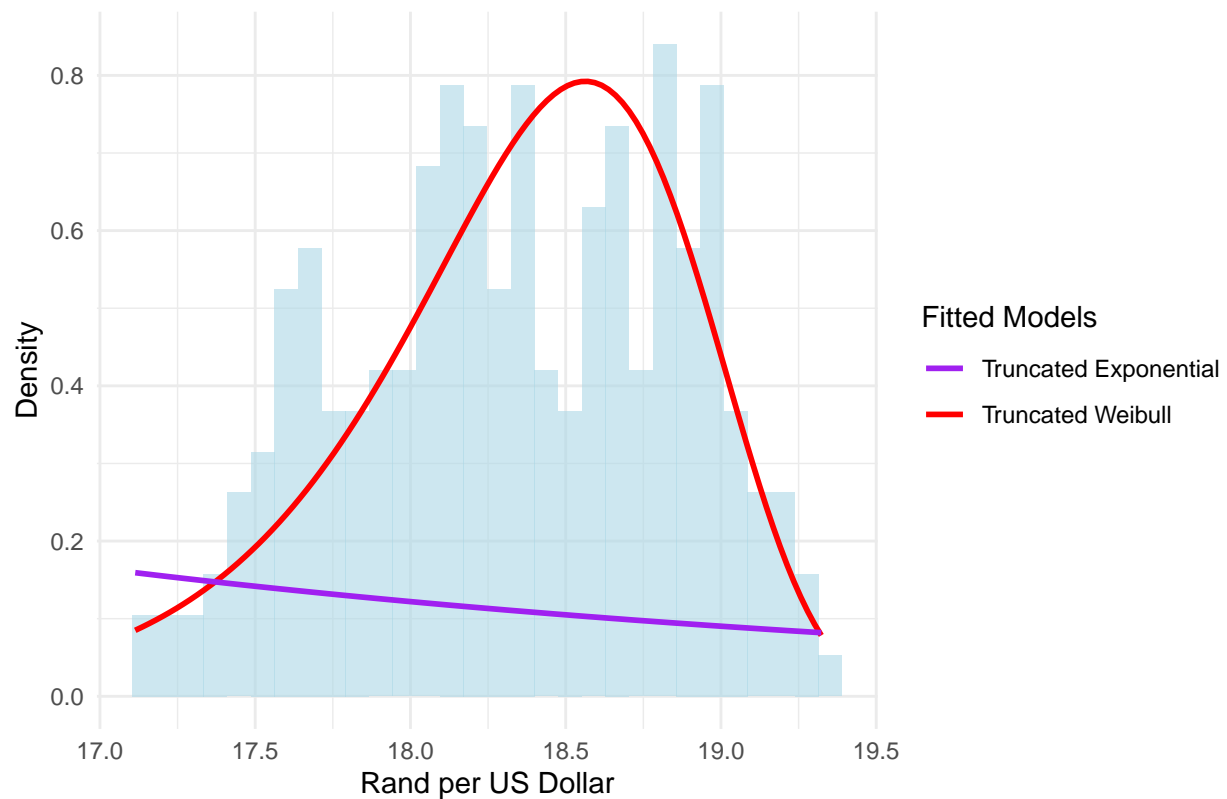
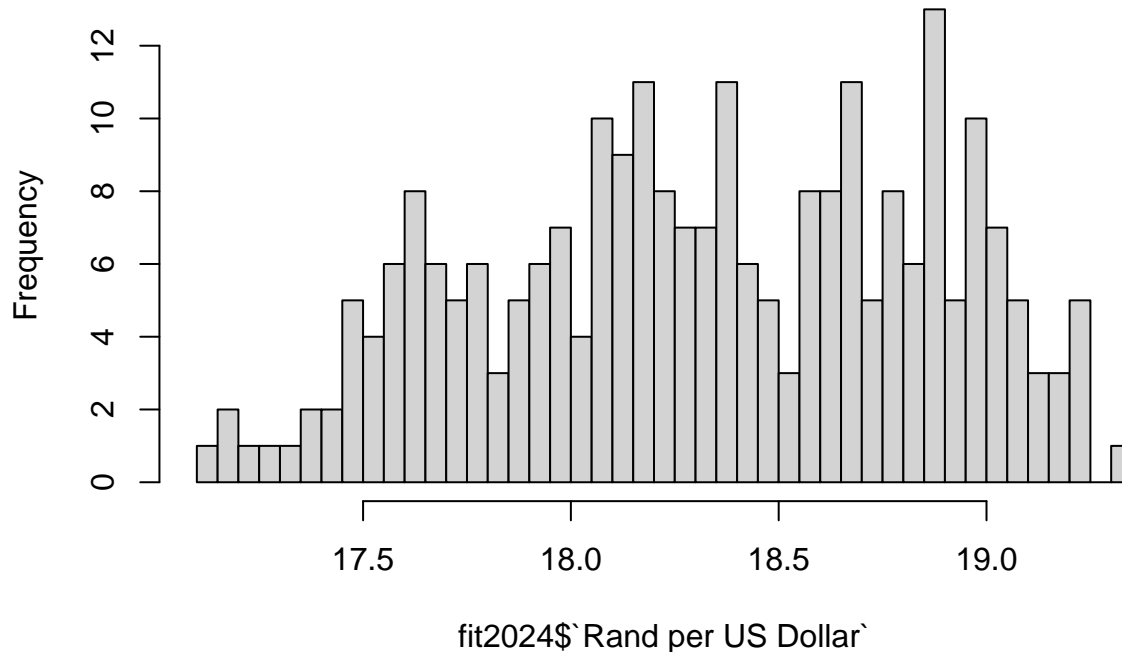gram of 2024 Rand per US Dollar Overlayed by Proposed Densities

From our histogram, we see the truncated Weibull, visually, fits very well, however the exponential is very poor. We proceed to a Goodness of Fit test for the Weibull. We define the null hypothesis:

H_0 : The Truncated Weibull Density is the true underlying distribution for the Rand per US Dollar of 2024.

H_A : The Truncated Weibull Density is not the true underlying distribution for the Rand per US Dollar of 2024.

```
histogram <- hist(fit2024$`Rand per US Dollar`,breaks=50)
```

Histogram of fit2024$`Rand per US Dollar`


```
breaks_cdf <- pweibull_trunc(histogram$breaks,fit_wei$estimate[1],fit_wei$estimate[2], a=15)
null.probs <- rollapply(breaks_cdf, 2, FUN = function(x) x[2]-x[1])
Chi <- chisq.test(
  histogram$counts, p = null.probs, rescale.p = TRUE, simulate.p.value = TRUE)
ifelse(Chi$p.value < 0.05,
"At a 5% level of significance, we reject the null hypothesis. The Truncated Weibull density is not the
"At a 5% level of significance, we do not reject. The Truncated Weibull density, is the true underlying
```

```
## [1] "At a 5% level of significance, we do not reject. The Truncated Weibull density, is the true und
```

3.2. Model Diagnostics

In this section, we want to analyse the time series. The motivation for this is, previously, we had a look at the actual distributions of the annual Rand per US Dollar, this is interesting, but not incredibly useful, with the reason being, the distribution tells us about the Random Variable, not about the behavior of the random variable over time.

3.2.1. Stationarity test

We are going to start by testing for unit roots, using the Augmented Dickey Fuller test, with an Auto Regressive Underlying Data Generating Processes of order determined using Alkaline Information Criteria (AIC). From visual inspection of our graph in 1, we see an upward trend

For an ARIMA(p,d,q) process, we test:
$$H_0 : d = 1$$
$$H_A : d = 0$$

```
adf_test <- ur.df(data$`Rand per US Dollar`, type="trend", selectlags="AIC")
ifelse(adf_test@teststat[1] < adf_test@cval[1, "5pct"],"Reject the null hypothesis. The series is stati
```

```
## [1] "Do not reject the null hypothesis. The series has a unit root."
```

Next we estimate parameters:

```
auto_fit <- auto.arima(data$`Rand per US Dollar`, stationary = F)
auto_fit$coef
```

```
##          ar1          ma1         drift
##   0.841789212 -0.855887188   0.001191155
```
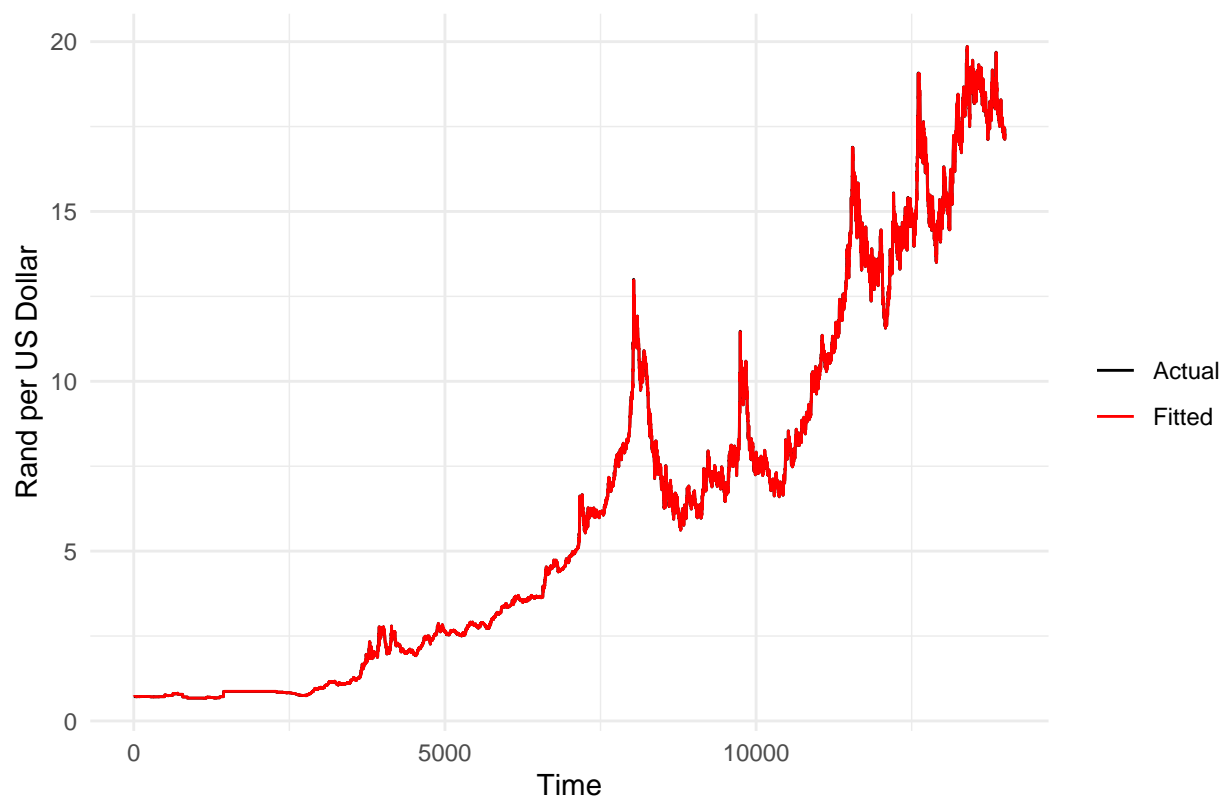
Our model is thus:

$$\Delta Z_t = 0.0011912 + 0.841789\Delta Z_{t-1} - 0.855887a_{t-1} + a_t$$
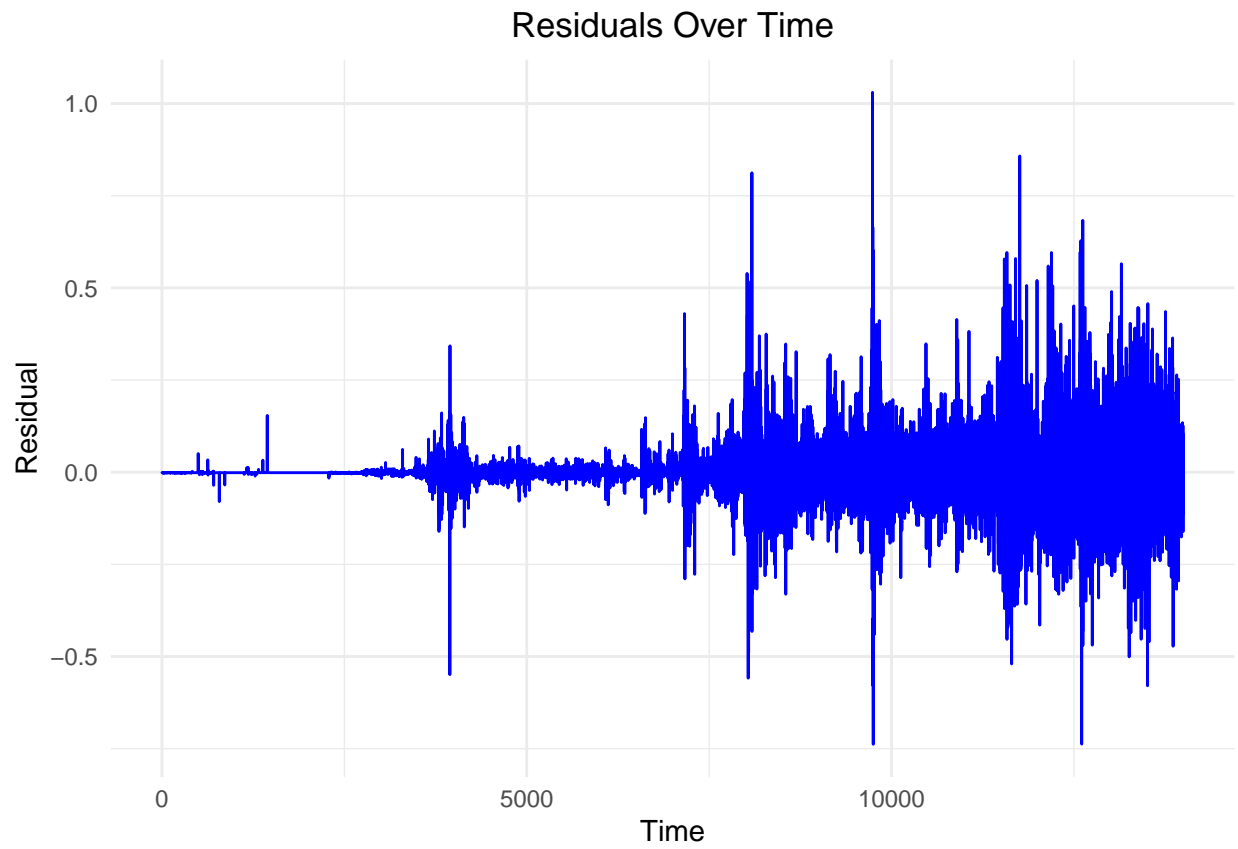
3.3. Plotting ARIMA Process and Residuals

```
fitted_vals <- fitted(auto_fit)
drift_trend <- cumsum(rep(coef(auto_fit)["drift"], length(data$`Rand per US Dollar`)))
plot_df <- data.frame(
  Time = 1:length(data$`Rand per US Dollar`),
  Actual = as.numeric(data$`Rand per US Dollar`),
  Fitted = as.numeric(fitted_vals),
  Drift = drift_trend
)
ggplot(plot_df, aes(x=Time)) +
  geom_line(aes(y=Actual, color="Actual")) +
  geom_line(aes(y=Fitted, color="Fitted")) +
  scale_color_manual("", values=c("Actual"="black", "Fitted"="red")) +
  labs(title="Observed data overlayed with fitted ARIMA(1,1,1) process", y="Rand per US Dollar", x="Tim
  theme_minimal()+theme(plot.title = element_text(hjust = 0.5))
```
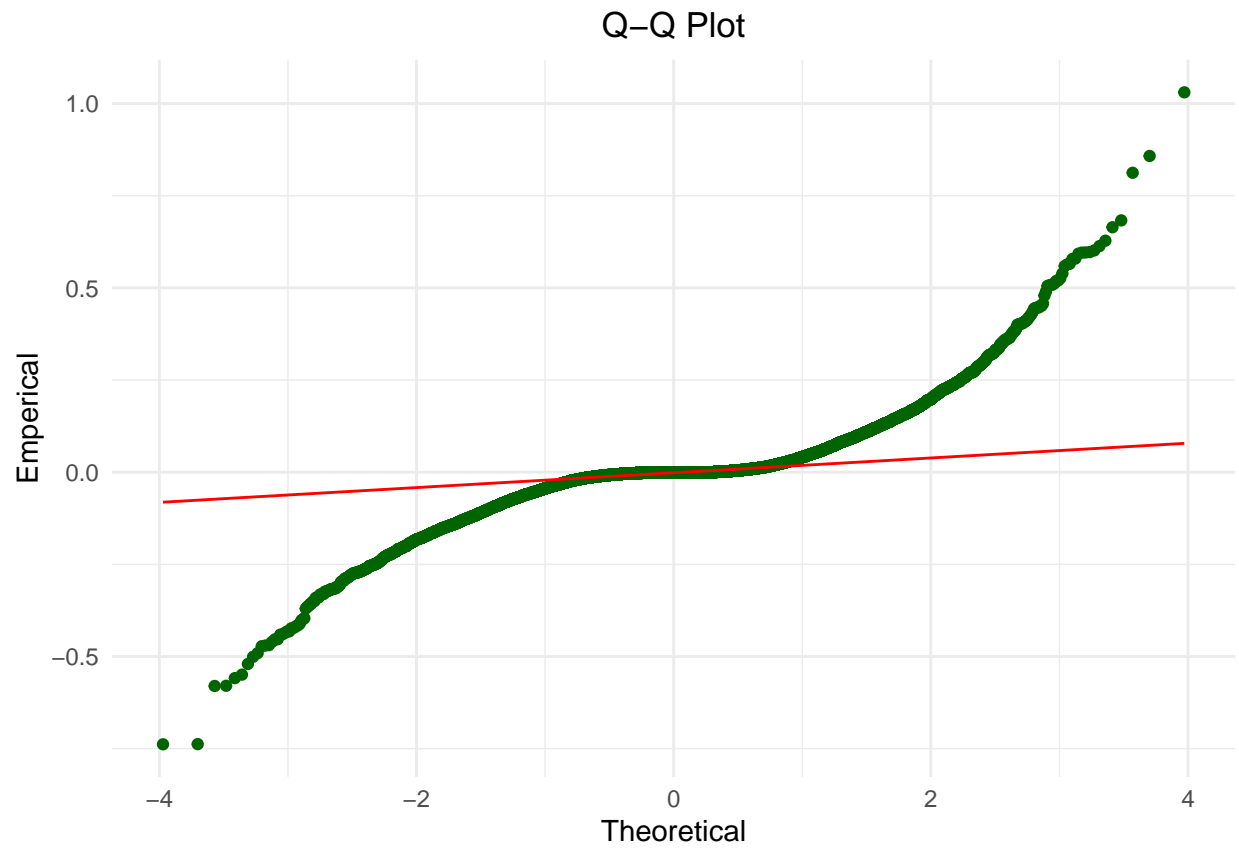
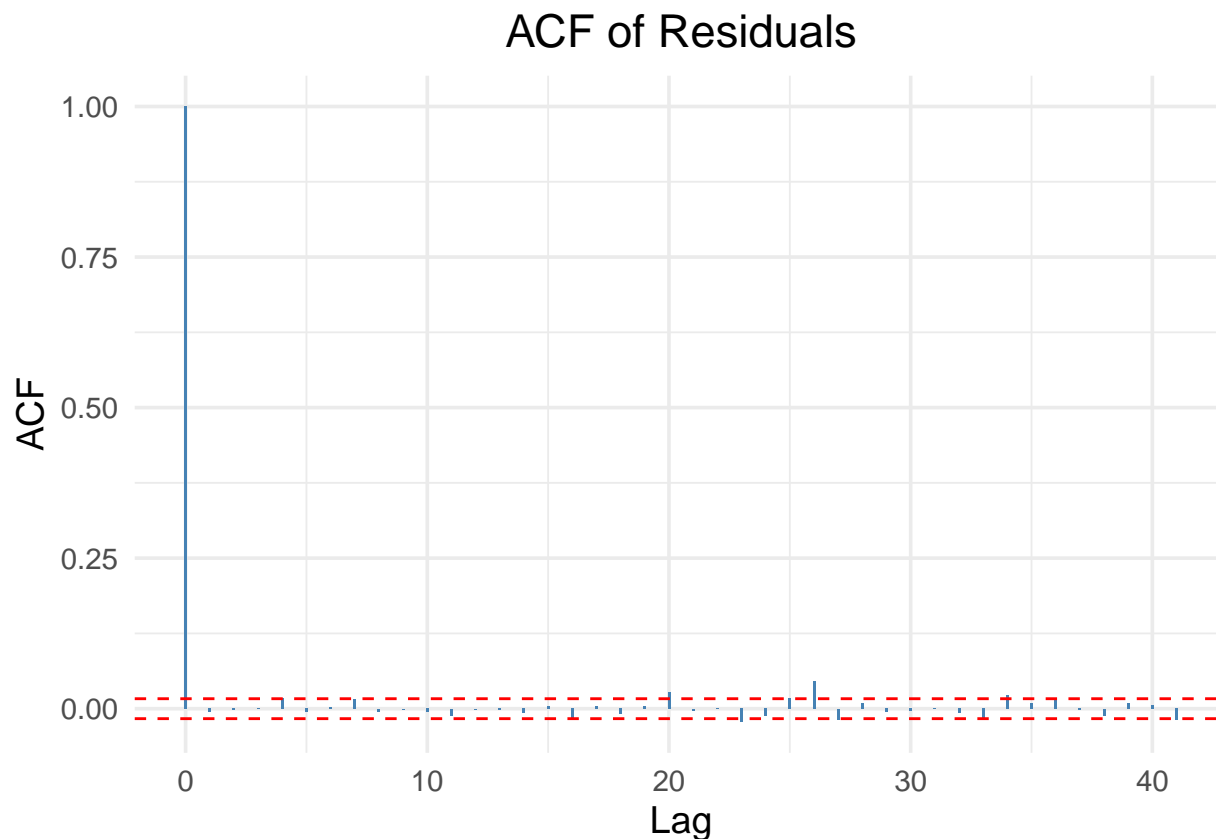## Observed data overlayed with fitted ARIMA(1,1,1) process


```
resid <- residuals(auto_fit)
p1 <- ggplot(data.frame(Time=1:length(resid), resid=resid), aes(x=Time, y=resid)) +
      geom_line(color="blue") +labs(title="Residuals Over Time", x="Time", y="Residual")+ theme_minimal
p2 <- ggplot(data.frame(resid=resid), aes(sample=resid)) +
      stat_qq(color="darkgreen") + stat_qq_line(color="red") + theme_minimal() +theme(plot.title = eleme
acf_res <- acf(resid, plot=FALSE)
acf_df <- with(acf_res, data.frame(Lag = lag, ACF = acf))
conf_limit <- qnorm((1 + 0.95)/2)/sqrt(nrow(data))
p3 <- ggplot(acf_df, aes(x=Lag, y=ACF)) +geom_bar(stat="identity", fill="steelblue", width=0.1) +geom_h
p1
```

Residuals Over Time

p2

Q–Q Plot

p3

ACF of Residuals

From our residual analysis, our model has correlated residuals, and increasingly varying over time.

The most concerning issue is the increasing residuals over time, this means when forecasting our Rand per US Dollar, our forecast becomes increasingly incorrect. This makes the model unreliable. I believe the reason for the increasing residuals, is the increasing variance, this introduces complexity to the time series.

With regards to the correlated residuals, this is problematic, however, when looking at 40 lags, we only see a handful of statistically significant correlations at a 5% level of significance, this is not good, but could be acceptable. We also see our residuals are not normally distributed. This is not ideal, but is not necessarily an issue.

```
R2 <- cor(data$`Rand per US Dollar`, fitted_vals)^2
cat("Our Coefficient of Determination (CoD) is",R2)
```

```
## Our Coefficient of Determination (CoD) is 0.999757
```

From the CoD, we see our model explains about 99.13% of the variation in the Rand per US Dollar.

3.4. Conclusion

Our model is not perfect, however, with such a high CoD, we can accept the model. It is worth noting, our CoD is very high, and with correlated residuals, we enter the territory of over-fitting the model.

```
n <- nrow(plot_df)
h <- 10*365
```

```
fc <- forecast(auto_fit, h = h)
forecast_df <- data.frame(Time = (n+1):(n+h),Forecast = as.numeric(fc$mean),
  Lower80 = fc$lower[,"80%"],Upper80 = fc$upper[,"80%"],
  Lower95 = fc$lower[,"95%"],Upper95 = fc$upper[,"95%"])
ggplot()+geom_line(data = plot_df, aes(x = Time, y = Actual, color = "Actual")) +
geom_line(data = plot_df, aes(x = Time, y = Fitted, color = "Fitted")) +
geom_line(data = forecast_df, aes(x = Time, y = Forecast, color = "Forecast")) +
geom_ribbon(data = forecast_df,aes(x = Time, ymin = Lower95, ymax = Upper95, fill = "95% CI"),alpha = 0
geom_ribbon(data = forecast_df,aes(x = Time, ymin = Lower80, ymax = Upper80, fill = "80% CI"),alpha = 0
scale_color_manual("",values = c("Actual" = "black","Fitted" = "red","Forecast" = "blue")) +
scale_fill_manual("",values = c("80% CI" = "lightblue","95% CI" = "lightgray")) +
labs(title = "ARIMA(1,1,1) Forecasted for the Next 10 Years",x = "Time",y = "Rand per US Dollar")+
theme_minimal() +theme(plot.title = element_text(hjust = 0.5))
```


ARIMA(1,1,1) Forecasted for the Next 10 Years