2131 Assignment 2

5588124

Task1

```r
Cali_UI<-read.csv("Un_Insurance1.csv")
```

1.

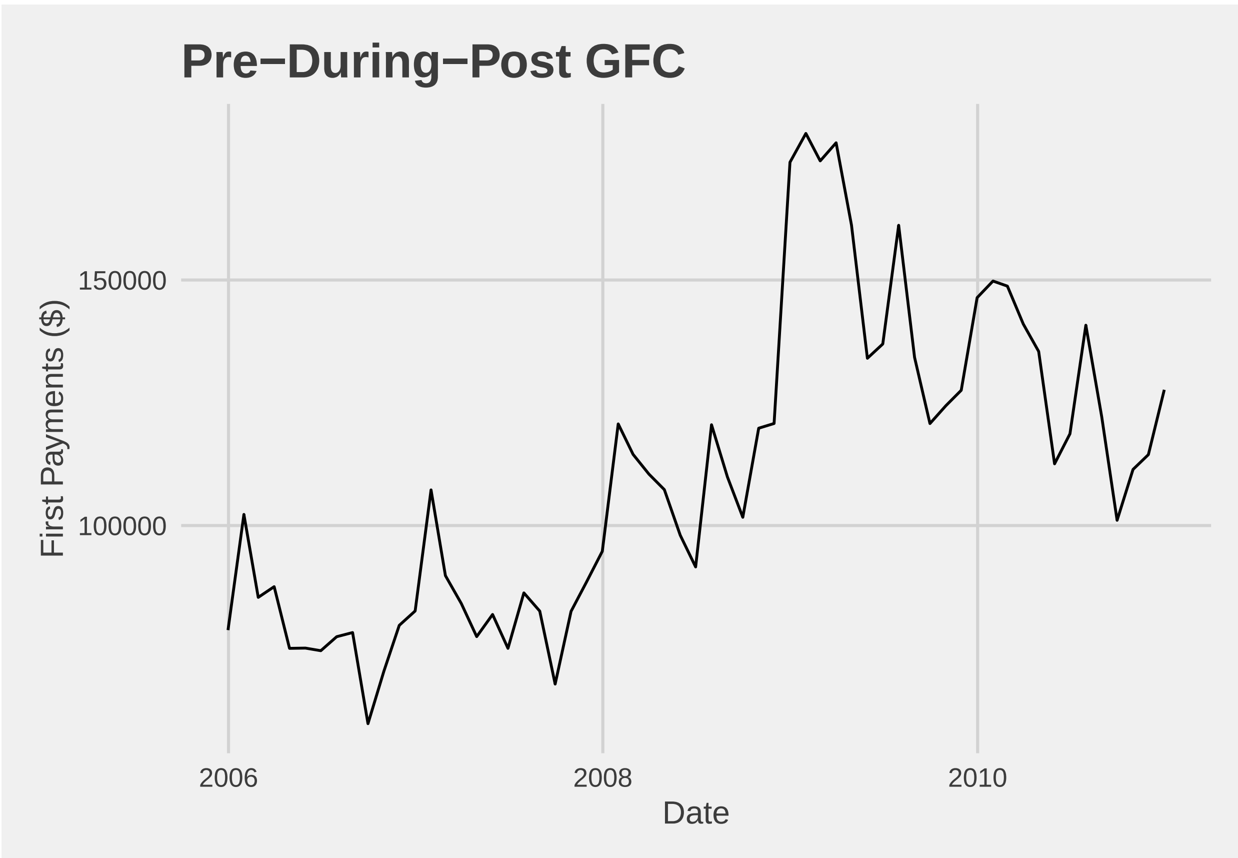
```r
options(scipen=999)
variables<-c("Initial.Claims", "First.Payments", "Weeks.Claimed",
               "Weeks.Compensated", "Avg..Wkly.Benefit",
               "Benefits.Paid", "Final.Payments")
# Calculate summary stats
summary_table<-data.frame(
  Variable=variables,
  Mean=sapply(Cali_UI[variables], mean, na.rm=TRUE),
  Variance=sapply(Cali_UI[variables], var, na.rm=TRUE),
  Skewness=sapply(Cali_UI[variables], skewness, na.rm=TRUE),
  Kurtosis=sapply(Cali_UI[variables], kurtosis, na.rm=TRUE)
)

# Print summary table
print(summary_table, digits=3)
```

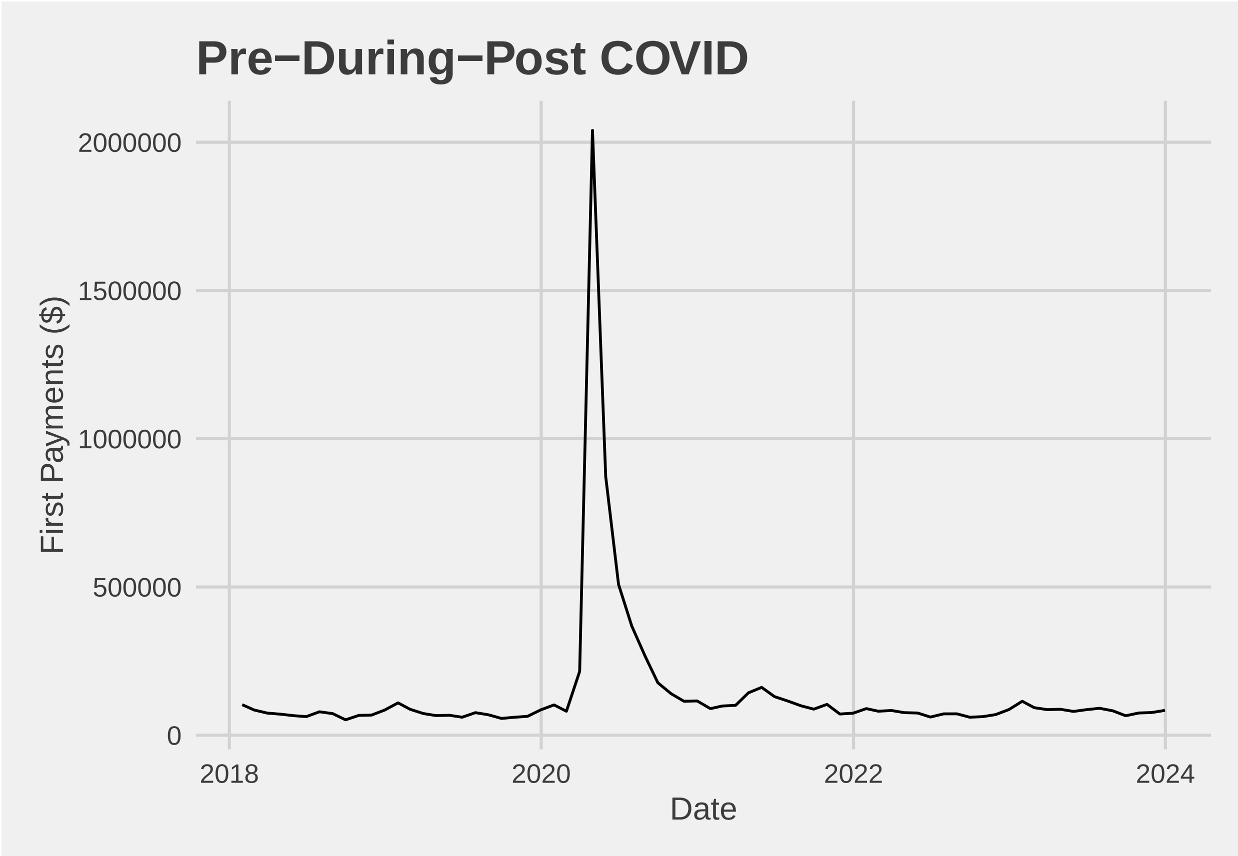
```
##                            Variable      Mean           Variance Skewness
## Initial.Claims       Initial.Claims    245069        18402118247    9.241
## First.Payments       First.Payments     99712         7726298159   17.829
## Weeks.Claimed         Weeks.Claimed   1826786      1379582938800    6.434
## Weeks.Compensated Weeks.Compensated   1693128      1227221643600    7.292
## Avg..Wkly.Benefit Avg..Wkly.Benefit       196              10315    0.207
## Benefits.Paid         Benefits.Paid 345876987 127502141976783952    5.539
## Final.Payments       Final.Payments     43990         1865636317   11.677
##                   Kurtosis
## Initial.Claims       118.1
## First.Payments       376.2
## Weeks.Claimed         56.7
## Weeks.Compensated     66.9
## Avg..Wkly.Benefit      1.5
## Benefits.Paid         46.4
## Final.Payments       171.0
```

2.

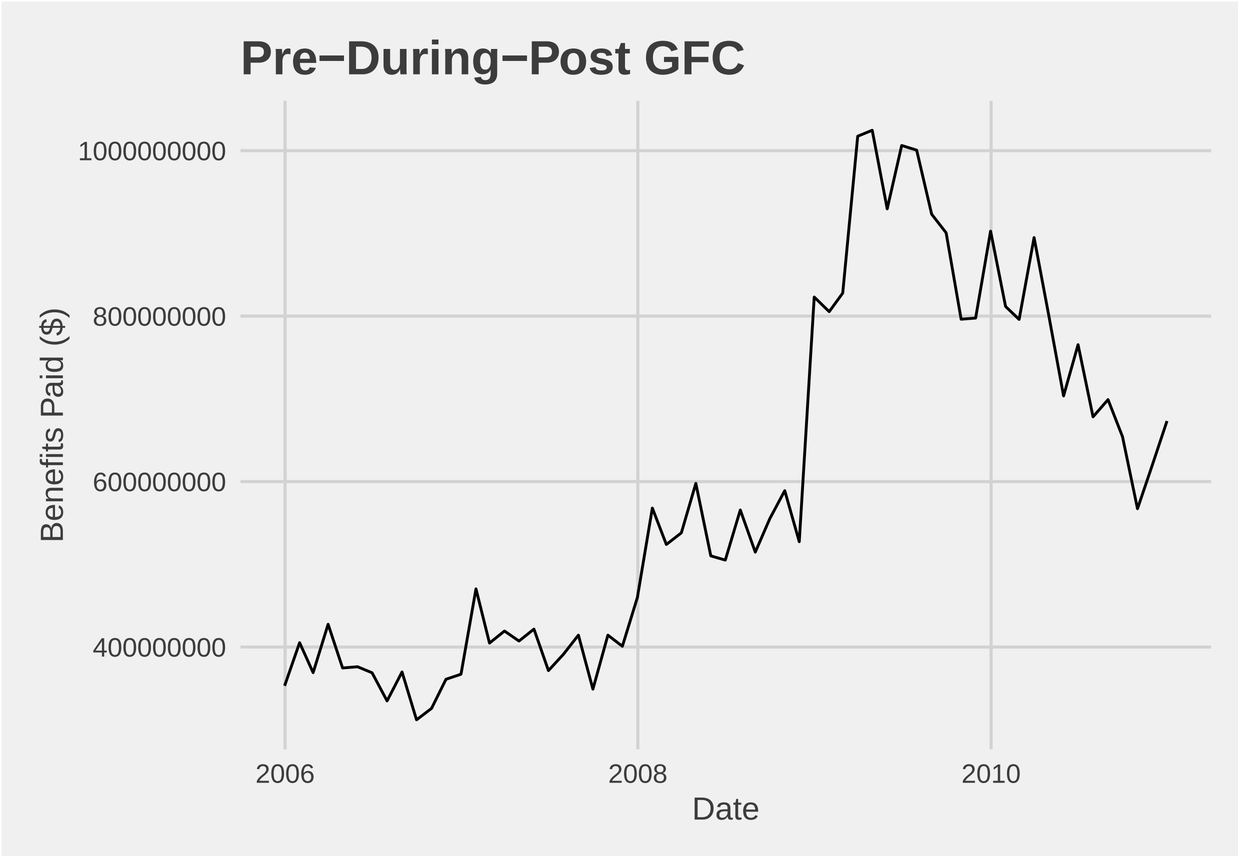
```r
options(scipen=999)
#First select data range of first payments for GFC
Cali_UI$Date<-as.Date(Cali_UI$Date, format="%m/%d/%Y")
data_FP_GFC<-subset(Cali_UI, Date>=as.Date("2005-12-31") & Date <= as.Date("2010-12-31"))
data_FP_GFC<-data_FP_GFC[,c("Date","First.Payments")]
ggplot(data_FP_GFC, aes(x=Date, y=First.Payments)) +
  geom_line()+
  labs(title="Pre-During-Post GFC", x="Date",y="First Payments ($)")+
  theme_fivethirtyeight()+
  theme(axis.title=element_text())
```



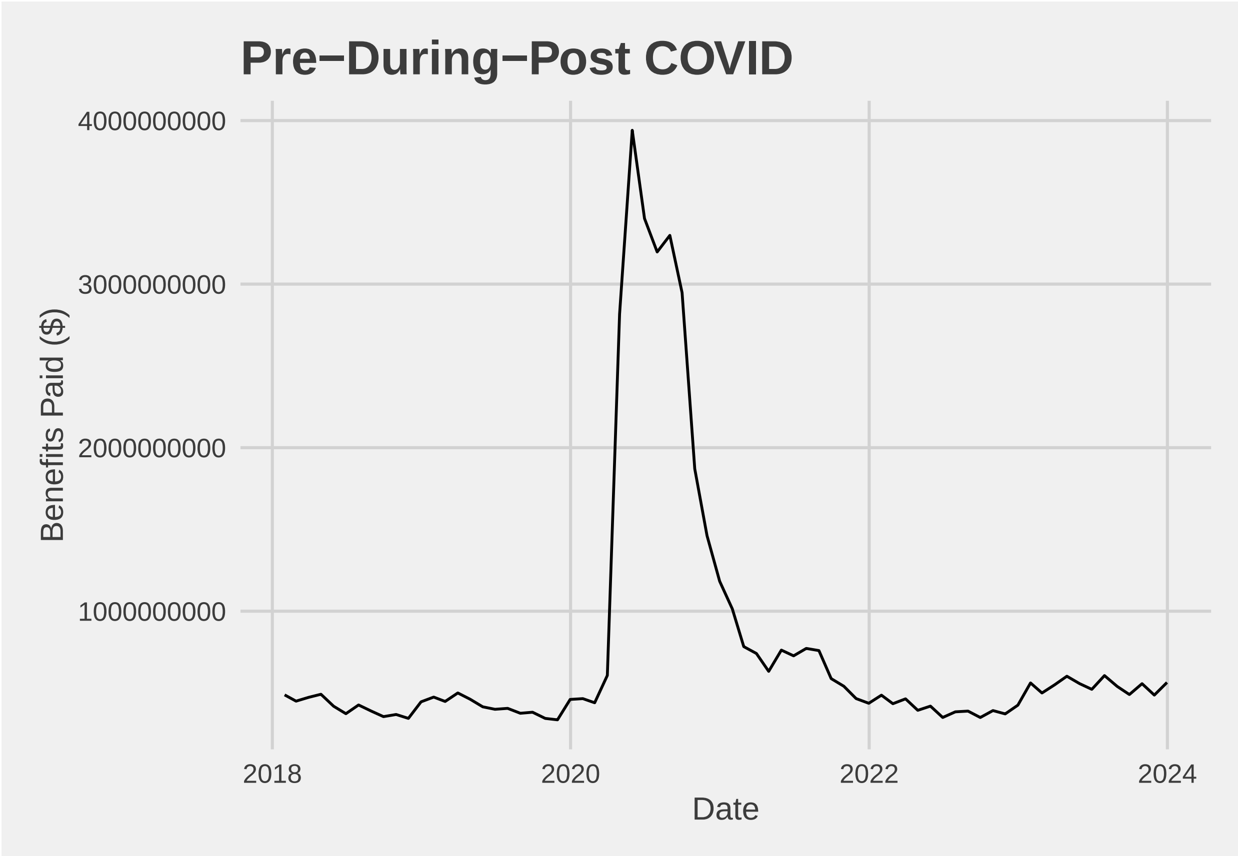

```r
#First Payments during COVID
Cali_UI$Date<-as.Date(Cali_UI$Date, format="%m/%d/%Y")
data_FP_CVD<-subset(Cali_UI, Date>=as.Date("2018-01-31") & Date<=as.Date("2023-12-31"))
data_FP_CVD<-data_FP_CVD[,c("Date","First.Payments")]
ggplot(data_FP_CVD, aes(x=Date, y=First.Payments)) +
  geom_line()+
  labs(title="Pre-During-Post COVID", x="Date",y="First Payments ($)")+
  theme_fivethirtyeight()+
  theme(axis.title=element_text())
```




```r
#Benefits Paid during GFC
data_BP_GFC<-subset(Cali_UI, Date>=as.Date("2005-12-31") & Date<=as.Date("2010-12-31"))
data_BP_GFC<-data_BP_GFC[,c("Date","Benefits.Paid")]
ggplot(data_BP_GFC, aes(x=Date, y=Benefits.Paid)) +
  geom_line()+
  labs(title="Pre-During-Post GFC", x="Date",y="Benefits Paid ($)")+
  theme_fivethirtyeight()+
  theme(axis.title=element_text())
```




```r
#Benefits Paid during COVID
data_BP_CVD<-subset(Cali_UI, Date>=as.Date("2018-01-31") & Date<=as.Date("2023-12-31"))
data_BP_CVD<-data_BP_CVD[,c("Date","Benefits.Paid")]
ggplot(data_BP_CVD, aes(x=Date, y=Benefits.Paid)) +
  geom_line()+
  labs(title="Pre-During-Post COVID", x="Date",y="Benefits Paid ($)")+
  theme_fivethirtyeight()+
  theme(axis.title=element_text())
```



3.

```
#Re-scale as values for initial claims are too large
clean_IC_scaled<-Cali_UI$Initial.Claims
lnorm_fit<-fitdistr(clean_IC_scaled, densfun="log-normal")
print(lnorm_fit)
```

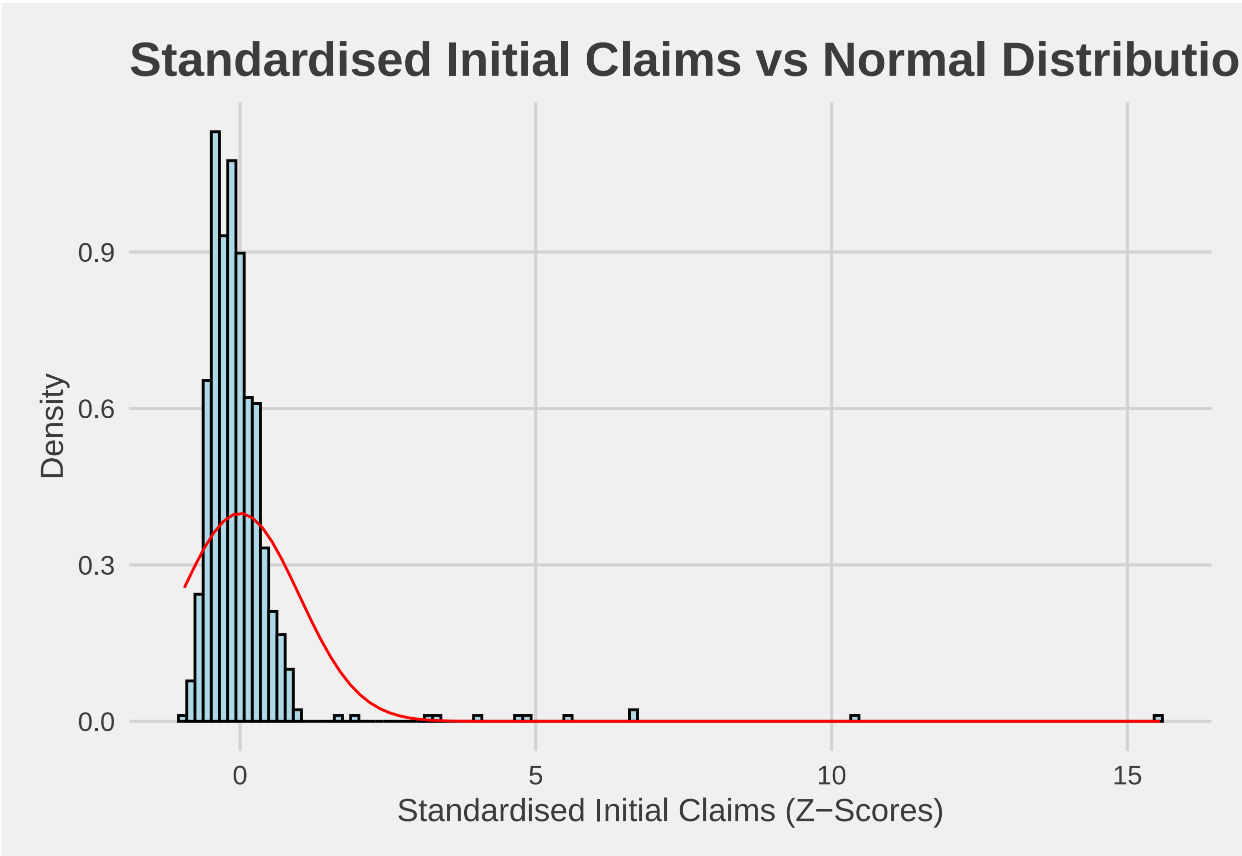
```
##      meanlog         sdlog
##   12.351678613    0.288627370
##  ( 0.011312199) ( 0.007998932)
```

Task 2

1. (In report)
2.

```
#First standardising our dataset for Initial Claims
IC_std<-scale(Cali_UI$Initial.Claims)
df_IC_std<-data.frame(IC_std = IC_std)
ggplot(df_IC_std, aes(x=IC_std)) +
  geom_histogram(aes(y=after_stat(density)), bins = 120, fill ="lightblue", color = "black") +
  stat_function(fun=dnorm, args=list(mean=0, sd=1), color ="red", linewidth = 0.5) +
  labs(title="Standardised Initial Claims vs Normal Distribution",
```

```
      x="Standardised Initial Claims (Z-Scores)", y="Density") +
 theme_fivethirtyeight()+
  theme(axis.title=element_text())
```



3.

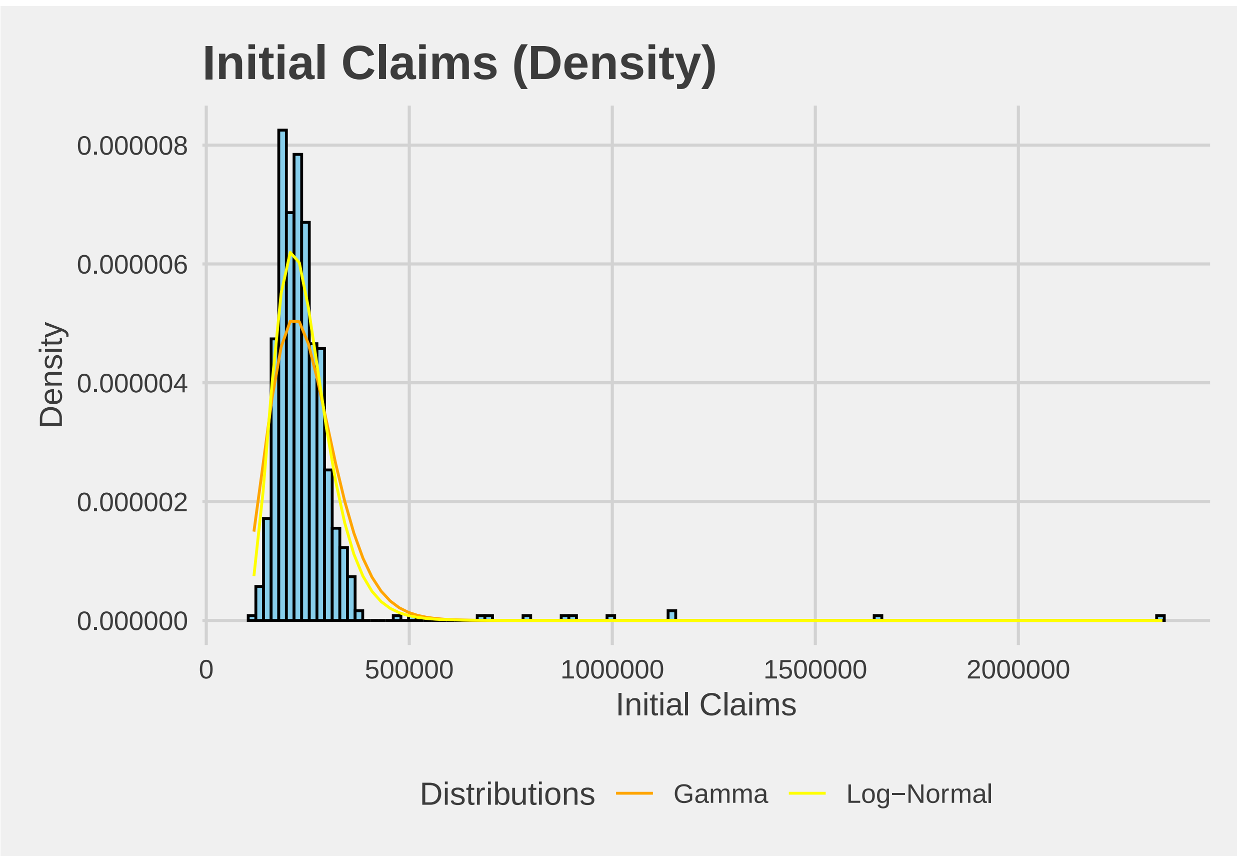
```
Cali_UI_IC<-Cali_UI$Initial.Claims
df_IC<-data.frame(Cali_UI_IC=Cali_UI_IC)
meanlog_IC <- mean(log(Cali_UI_IC))
sdlog_IC <- sd(log(Cali_UI_IC))
#Getting parameters for a Gamma distribution to check fit
clean_IC_scaled<-Cali_UI$Initial.Claims/1000
gamma_fit<-fitdistr(clean_IC_scaled, densfun="gamma")
print(gamma_fit)
```

```
##       shape         rate
##   8.838655993   0.036066668
##  (0.478854714) (0.002009752)
```

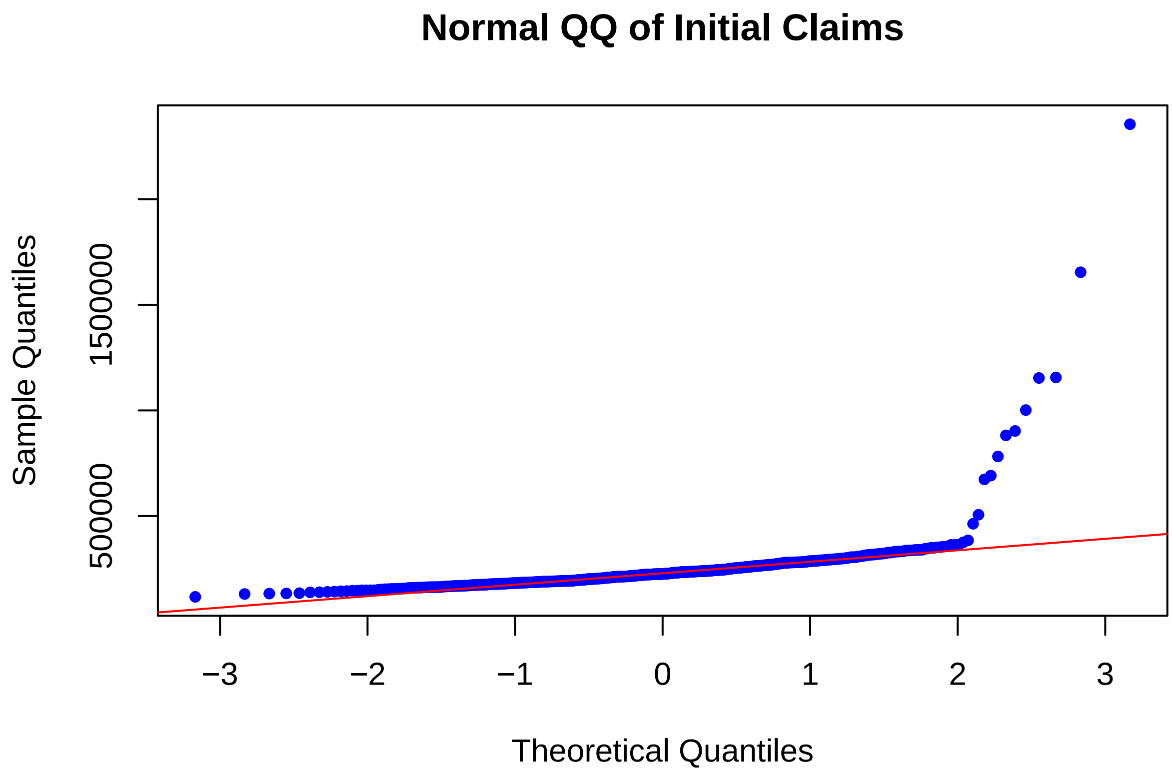
```
#Note the rate parameter is a scaled down value
#So we divide it back by 1000 to get its original rate parameter.
print(0.036066668/1000)
```

```
## [1] 0.00003606667
```

```r
#We can compare this with the shape of the non-standardised form of initial claims with the same number
UI_IC<-Cali_UI$Initial.Claims
df_IC<-data.frame(UI_IC=UI_IC)
ggplot(df_IC, aes(x=UI_IC,colour = Distributions)) +
  geom_histogram(aes(y=after_stat(density)),
                 fill = "skyblue", color = "black", bins = 120) +
  stat_function(fun=dgamma, aes(colour= "Gamma"),
                args=list(shape=8.838655993, rate=0.036066668/1000),
                linewidth = 0.5) +
  stat_function(fun=dlnorm,aes(color = "Log-Normal"),
                args = list(meanlog = meanlog_IC, sdlog = sdlog_IC), linewidth = 0.5)+
  scale_color_manual(name = "Distributions",
                     values = c("Gamma" = "orange", "Log-Normal" = "yellow"))+
  labs(title = "Initial Claims (Density)",
       x="Initial Claims",
       y="Density") +
  theme_fivethirtyeight() + theme(axis.title = element_text())
```




```r
#Construct Q-Q plot to compare Initial Claims with the Normal distribution
qqnorm(Cali_UI_IC, main= "Normal QQ of Initial Claims", pch = 20, col = "blue")
qqline(Cali_UI_IC, col = "red", lwd = 1)
```



Task 3

1.

```
breaks_IC <- c(-Inf, -2, -1, 0, 1, 2, Inf)
observed_IC<-table(cut(IC_std, breaks = breaks_IC))
#Calculating expected proportions under standard normal
expected_probs_IC<-c(pnorm(-2),
                     pnorm(-1)-pnorm(-2),
                     pnorm(0)-pnorm(-1),
                     pnorm(1)-pnorm(0),
                     pnorm(2)-pnorm(1),
                     1-pnorm(2))
expected_counts_IC<-sum(observed_IC)*expected_probs_IC
norm_IC<-chisq.test(x=observed_IC, p=expected_probs_IC)
print(norm_IC)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed_IC
## X-squared = 376.71, df = 5, p-value < 0.00000000000000022
```

```r
#Checking if IC_std fits a student-t distribution
IC1<-as.numeric(Cali_UI$Initial.Claims)
IC1_std<-scale(IC1)
t_fit_IC<- fitdistr(IC1,densfun = "t", start=list(m=mean(IC1),s=sd(IC1), df=3))
df_t<-t_fit_IC$estimate["df"]
breaks<-c(-Inf, -2, -1, 0, 1, 2, Inf)
observed_counts_t<-table(cut(IC1_std, breaks=breaks))
expected_probs_t<-c(pt(-2, df=df_t),
                    pt(-1, df=df_t)-pt(-2, df=df_t),
                    pt(0, df=df_t)-pt(-1, df=df_t),
                    pt(1, df=df_t)-pt(0, df=df_t),
                    pt(2, df=df_t)-pt(1, df=df_t),
                    1-pt(2, df=df_t))
expected_counts<-sum(observed_counts_t)*expected_probs_t
gof_ttest_IC<-chisq.test(x = observed_counts_t, p = expected_probs_t)
print(gof_ttest_IC)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed_counts_t
## X-squared = 429.99, df = 5, p-value < 0.00000000000000022
```

2.

```r
w_comp<-log(Cali_UI$Weeks.Compensated)
w_claim<-log(Cali_UI$Weeks.Claimed)
diff_log<-w_claim-w_comp
diff_std<- scale(diff_log)
observed_diff<-table(cut(diff_std, breaks = breaks))
expected_probs_diff<- c(
  pnorm(-2),
  pnorm(-1) - pnorm(-2),
  pnorm(0) - pnorm(-1),
  pnorm(1) - pnorm(0),
  pnorm(2) - pnorm(1),
  1 - pnorm(2))

expected_counts_diff<-sum(observed_diff)*expected_probs_diff
gof_result_diff<-chisq.test(x = observed_diff, p = expected_probs_diff)
print(gof_result_diff)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed_diff
## X-squared = 1155.8, df = 5, p-value < 0.00000000000000022
```

```r
wilcox.test(w_comp,w_claim, paired=TRUE)
```

```
##
```

```
##  Wilcoxon signed rank test with continuity correction
##
## data:  w_comp and w_claim
## V = 45976, p-value < 0.00000000000000022
## alternative hypothesis: true location shift is not equal to 0
```

3.

```r
bp<-Cali_UI$Benefits.Paid
log_bp<-log(bp)

mean_log_bp<-mean(log_bp)
q60_log_bp<-quantile(log_bp, probs = 0.60)
print(q60_log_bp)
```

```
##      60%
## 19.72133
```

```r
#One sample t-test

# Perform one-sample t-test
t.test(log_bp, mu = q60_log_bp, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  log_bp
## t = -11.908, df = 650, p-value = 1
## alternative hypothesis: true mean is greater than 19.72133
## 95 percent confidence interval:
##  19.30443      Inf
## sample estimates:
## mean of x
##  19.35509
```

```r
#Test Skew and kurtosis
print(skewness(log_bp))
```

```
## [1] -0.104435
```

```r
print(kurtosis(log_bp))
```

```
## [1] 3.0751
```