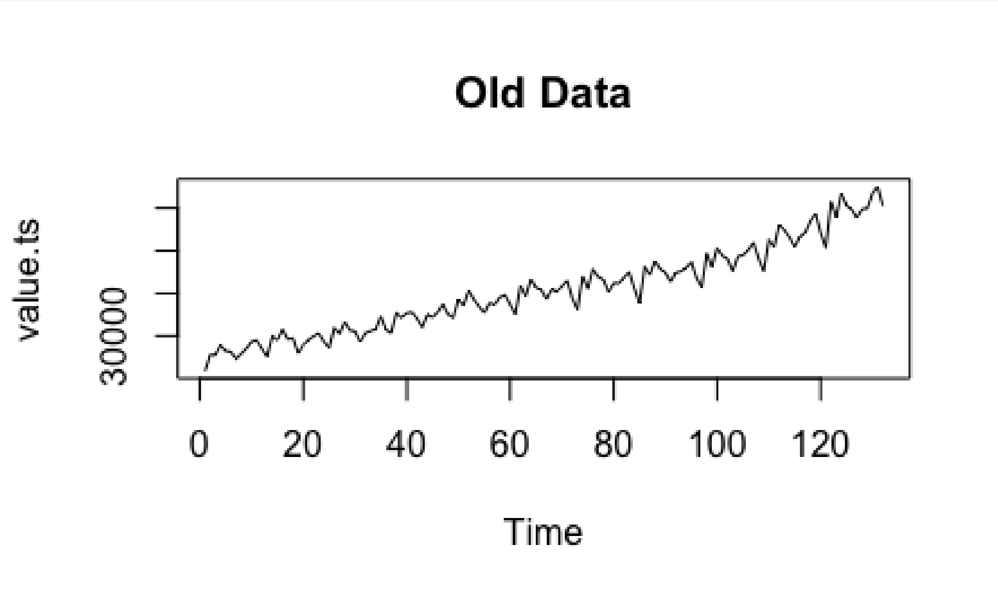
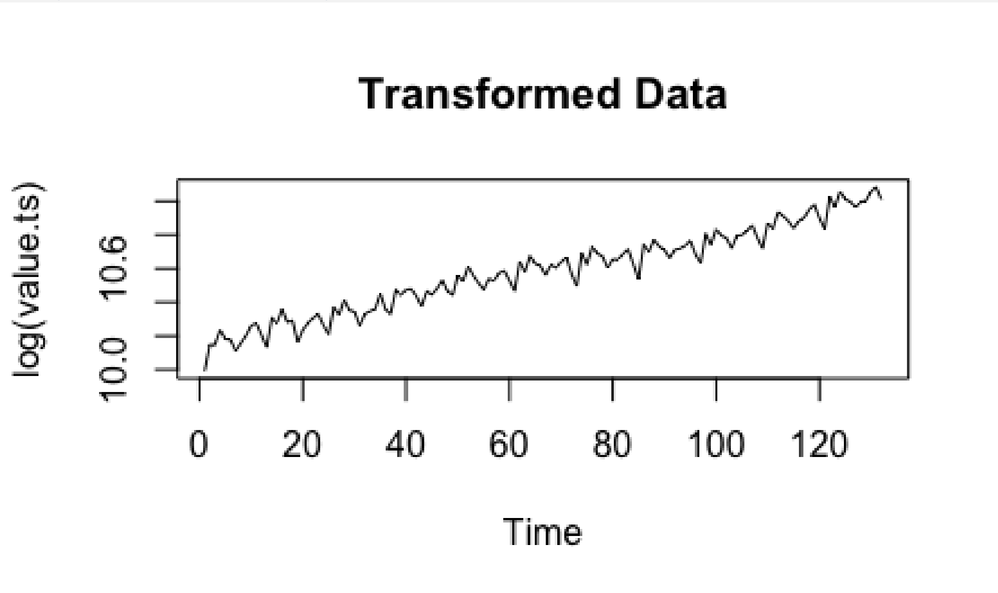
Question 6:

Part a.

```
mypath = "/Users/safurasuleymanovs/Desktop/4B/Stat/A2/"
data.set ="telephone_y.txt"
value.y = scan(paste(mypath,data.set,sep=""))
value.ts = ts(scan(paste(mypath,data.set,sep="")))
# When we plot the data we see that there is an upward trend.
# Also, we notice a seasonal component with increasing seasonal variation
# and period m = 12.
plot(value.ts,type="l")
title("Old Data")
```




```
# To stabilize the variance we decide to transform the data to log(Y(t)).
# According to the graph it worked well.
plot(log(value.ts),type="l")
title("Transformed Data")
```



Model 1.

To capture the trend and seasonal component with the period m=12 we can consider the following model:

$\log Y_t = \beta_0 + \beta_1 t + \sum_{k=2}^{12} \beta_k X_{t,k} + \epsilon_t$, $where\ t = 1, \dots \dots ,132$ (11 years in months)

where $X_{t,k}$ are indicator variables:

$$X_{t,k} = \begin{cases} 1\ if\ t\ corresponds\ to\ month\ k \\ 0\ if\ t\ does\ not\ correspond\ to\ month\ k \end{cases}$$

here $\beta_0 + \beta_1 t$ is the trend component and $\sum_{k=2}^{12} \beta_k X_{t,k}$ is the seasonal component, t is the time in months, ϵ_t is the noise and we will assume that it is OLS for now.

Part b.

Creating Matrix for explanatory variable X

```
b <-diag(12)
x <-rbind(b,b,b,b,b,b,b,b,b,b,b)
x[,1] <- 1:132
telephone.ls <-lsfit(x,log(value.y))
ls.print(telephone.ls)
```

```
Residual Standard Error=0.0303
R-Square=0.9871
F-statistic (df=12, 119)=759.8323
p-value=0

          Estimate Std.Err  t-value Pr(>|t|)
Intercept  10.0410  0.0101 997.4196   0e+00
X1          0.0065  0.0001  93.0617   0e+00
X2          0.1617  0.0129  12.5218   0e+00
X3          0.1126  0.0129   8.7198   0e+00
X4          0.1899  0.0129  14.7064   0e+00
X5          0.1409  0.0129  10.9043   0e+00
X6          0.1159  0.0129   8.9722   0e+00
X7          0.0472  0.0129   3.6504   4e-04
X8          0.1001  0.0129   7.7444   0e+00
X9          0.1021  0.0129   7.8974   0e+00
X10         0.1313  0.0129  10.1544   0e+00
X11         0.1583  0.0129  12.2436   0e+00
X12         0.0743  0.0129   5.7449   0e+00
```

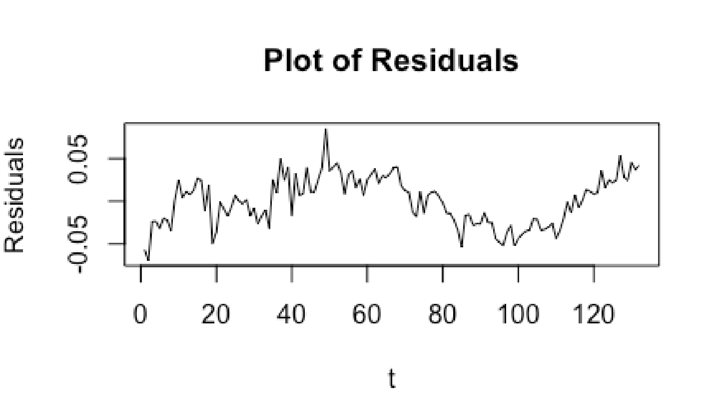
Part c.

- Looking at the p-values of individual t-statistics, we can say that all parameters are statistically significant at the 0.05 level. Hence, we should not eliminate any of them.
- The F-test considers all parameters simultaneously, besides Beta0. Since F-statistic is too large and p-value = 0, we can reject the null hypothesis and conclude that the model 1 has explanatory value.
- We see that R-square = 0.9871 is close to one which indicates that the model has excellent predictive power.

Part d.

Residual Analysis

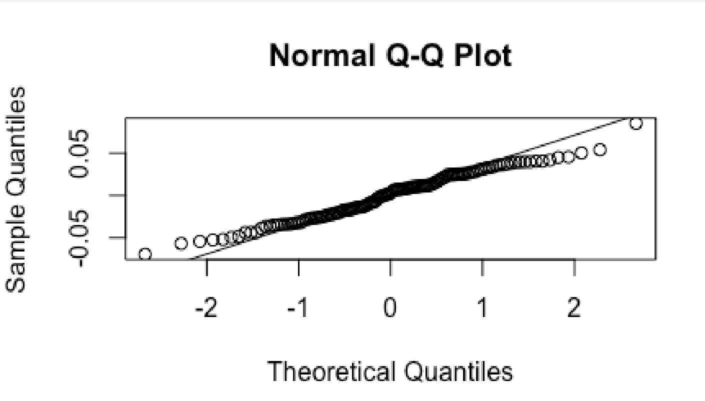
\# plot(c(1:132),telephone.ls$residuals,type="l",xlab="t",ylab="Residuals",main="Plot of Residuals")



The residuals plot shows curvature, which indicates that the modeling of the trend could be improved by using different function of *t*. Also, we still notice peak-and-valley behavior indicating that not all seasonal effect has been removed by our model 1. This possibly means that there are non-zero correlations.

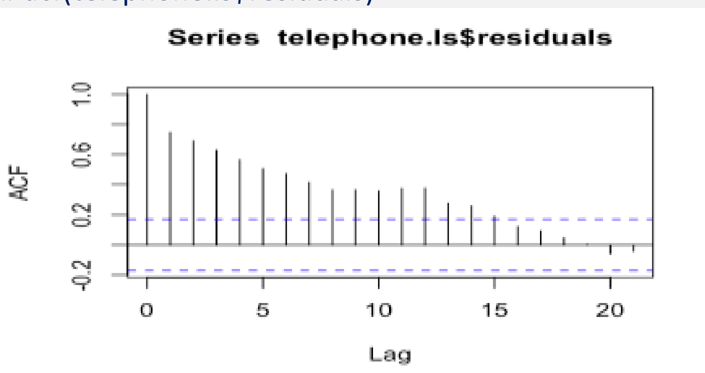
\# qqnorm(telephone.ls$residuals)

\# qqline(telephone.ls$residuals)



The QQ plot does not show any evidence against normality since it the middle part perfectly fits the line, however, the presence of correlations invalidates the use of QQ plot.

\# acf(telephone.ls$residuals)



We see that all vertical lines till lag 15 pass the dashed blue lines which indicates that residuals are not independent.

Overall, our OLS assumptions do not hold, the reported p-values and intervals are unlikely to be correct. However, the model has a good predictive value.

Part e.
Model 2.
To capture the trend and seasonal component with the period m=12 we can consider the following model:
$Y_t = \beta_0 + \beta_1 t^2 + \sum_{k=2}^{12} \beta_k X_{t,k} + \epsilon_t \,, where\ t = 1, \dots \dots ,60$ (5 years in months)
where $X_{t,k}$ are indicator variables:

$$X_{t,k} = \begin{cases} 1\ if\ t\ corresponds\ to\ month\ k \\ 0\ if\ t\ does\ not\ correspond\ to\ month\ k \end{cases}$$

here $\beta_0 + \beta_1 t^2$ is the trend component and $\sum_{k=2}^{6} \beta_k X_{t,k}$ is the seasonal component, $t^2$is squared value of time denoted in months, ϵ_t is the noise and we will assume that it is OLS for now.

Part f.
Model 2

```
Y2 <- value.y[73:132]
plot(c(1:60),Y2,type="l")
b2 <- diag(12)
X2 <- rbind(b2,b2,b2,b2,b2)
X2[,1] <- (c(1:60))^2

model2.ls <-lsfit(X2,Y2)
ls.print(model2.ls)
```

```
Residual Standard Error=646.9566
R-Square=0.9933
F-statistic (df=12, 47)=577.062
p-value=0
```

	Estimate	Std.Err	t-value	Pr(>|t|)
Intercept	36807.8828	297.9866	123.5219	0
X1	5.9359	0.0781	75.9970	0
X2	8231.0668	409.1907	20.1155	0
X3	5231.2618	409.2519	12.7825	0
X4	9530.5849	409.3597	23.2817	0
X5	7266.2360	409.5189	17.7433	0
X6	5929.4153	409.7347	14.4714	0
X7	3179.7227	410.0121	7.7552	0
X8	5186.1582	410.3564	12.6382	0
X9	5165.9218	410.7732	12.5761	0
X10	6692.8136	411.2679	16.2736	0
X11	7821.4334	411.8464	18.9911	0
X12	3313.5813	412.5142	8.0326	0

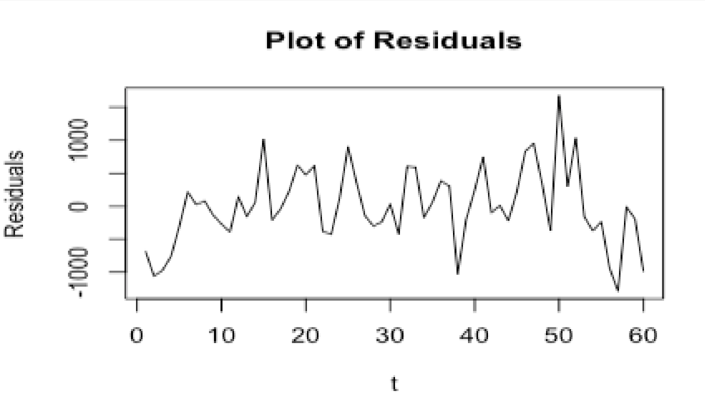
Part g.
- Looking at the p-values of individual t-statistics, we can say that all parameters are statistically significant at the 0.05 level. Hence, we should not eliminate any of them.

- The F-test considers all parameters simultaneously, besides Beta0. Since F-statistic is too large and p-value = 0, we can reject the null hypothesis and conclude that the model 1 has explanatory value.
- We see that R-square = 0.9933 is very close to one which indicates that the model has excellent predictive power.

Part h.

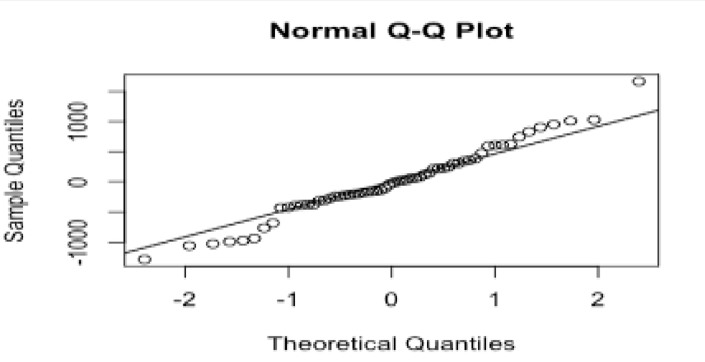
Residual Analysis

```
# plot(c(1:60),model2.ls$residuals,type="l",xlab="t",ylab="Residuals",main="Plot of Residuals")
```



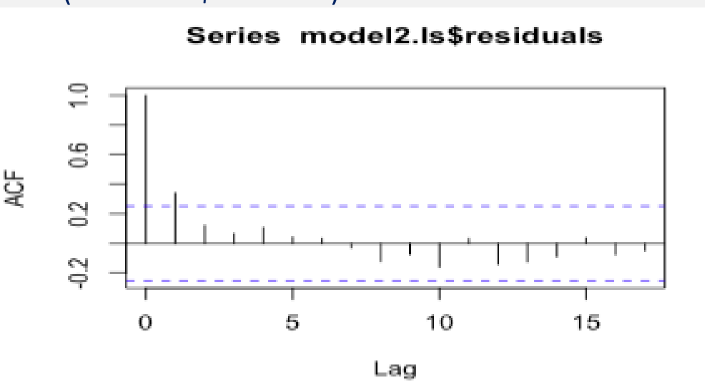
We still notice peak-and-valley behavior indicating that not all seasonal effect has been removed by our model 2. This possibly means that there are some non-zero correlations.

```
# qqnorm(model2.ls$residuals)
# qqline(model2.ls$residuals)
```



The QQ plot does not show any evidence against normality since its middle part perfectly fits the line.

```
# acf(model2.ls$residuals)
```



We see that only vertical line at 1 passes the dashed blue line which indicates that a few residuals are not independent.
Overall, the OLS assumptions seem not to hold, the reported p-values and intervals are unlikely to be correct. However, the model 2 has an excellent predictive value and certainly performs better than model 1 according to the analysis of residuals and values of R-squared.

Part i.

```
# Prediction Interval 1
new.X1 <-c(1,63^2,0,1,0,0,0,0,0,0,0,0,0)
Y.hat1 <-sum(new.X1*model2.ls$coefficients)
XTX.inv <-ls.diag(model2.ls)$cov.unscaled
est.stdev1 <- ls.diag(model2.ls)$std.dev*sqrt(new.X1%*%XTX.inv%*%new.X1 +1)
PI1 <- c(Y.hat1 -2*est.stdev1,Y.hat1 + 2*est.stdev1)
# [1] 64108.37 67089.45

# Prediction Interval 2
new.X2 <-c(1,75^2,0,1,0,0,0,0,0,0,0,0,0)
Y.hat2 <-sum(new.X2*model2.ls$coefficients)
XTX.inv <-ls.diag(model2.ls)$cov.unscaled
est.stdev2 <- ls.diag(model2.ls)$std.dev*sqrt(new.X2%*%XTX.inv%*%new.X2 +1)
PI2 <- c(Y.hat2 -2*est.stdev2,Y.hat2 + 2*est.stdev2)
# [1] 73839.11 77018.56
```

Part j.

```
data.set2 ="telephone_future.txt"
Y.future = scan(paste(mypath,data.set2,sep=""))

# SOS for Model 1
X.forecast1 <- diag(12)
X.forecast1[,1] <- c(133:144)
X.forecast1 <- cbind(rep(1,12),X.forecast1)
Y.hat.forecast1 <-exp(colSums(t(X.forecast1)*telephone.ls$coefficients))
SOS1 <- sum((Y.future-Y.hat.forecast1)^2)
# [1] 322448346 <- SOS for Model 1

# SOS for Model 2
X.forecast2 <- diag(12)
X.forecast2[,1] <- (c(61:72))^2
X.forecast2 <- cbind(rep(1,12),X.forecast2)
Y.hat.forecast2 <- colSums(t(X.forecast2)*model2.ls$coefficients)
SOS2 <- sum((Y.future-Y.hat.forecast2)^2)
# [1] 10543662 <- SOS for Model 2
```

As we can see SOS of Model 2 is much lower than SOS for Model 1, so Model 2 performs better than Model 1.