Project Maker: Yusa Lin

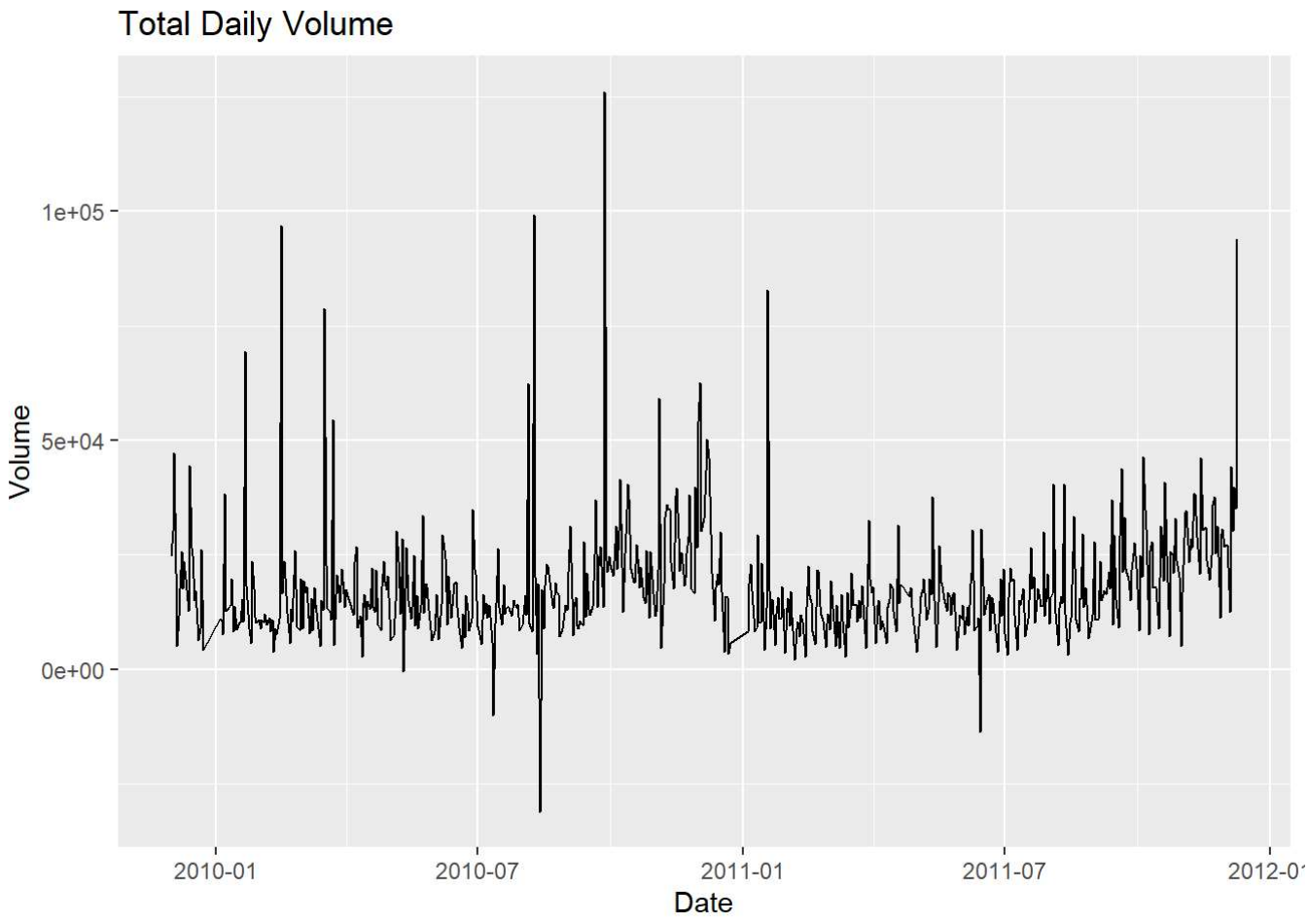
Part 1

Loading data, Packages and preparing Data

```
# Loading packages and data
library(ggplot2) # Package for data visualization
library(plotly) # Package for data visualization
library(forecast)
library(s20x)
library(chillR)
mydata <- read.csv("online_retail_II.csv")
mydata$InvoiceDate <- as.Date(mydata$InvoiceDate)
mydata$Customer.ID <- as.factor(mydata$Customer.ID)
mydata <- mydata[!grepl("C", mydata$Invoice),] # Delete all canceled orders from the dataset
rownames(mydata) <- 1:nrow(mydata)
# str(mydata)
# head(mydata)
```

Part 2

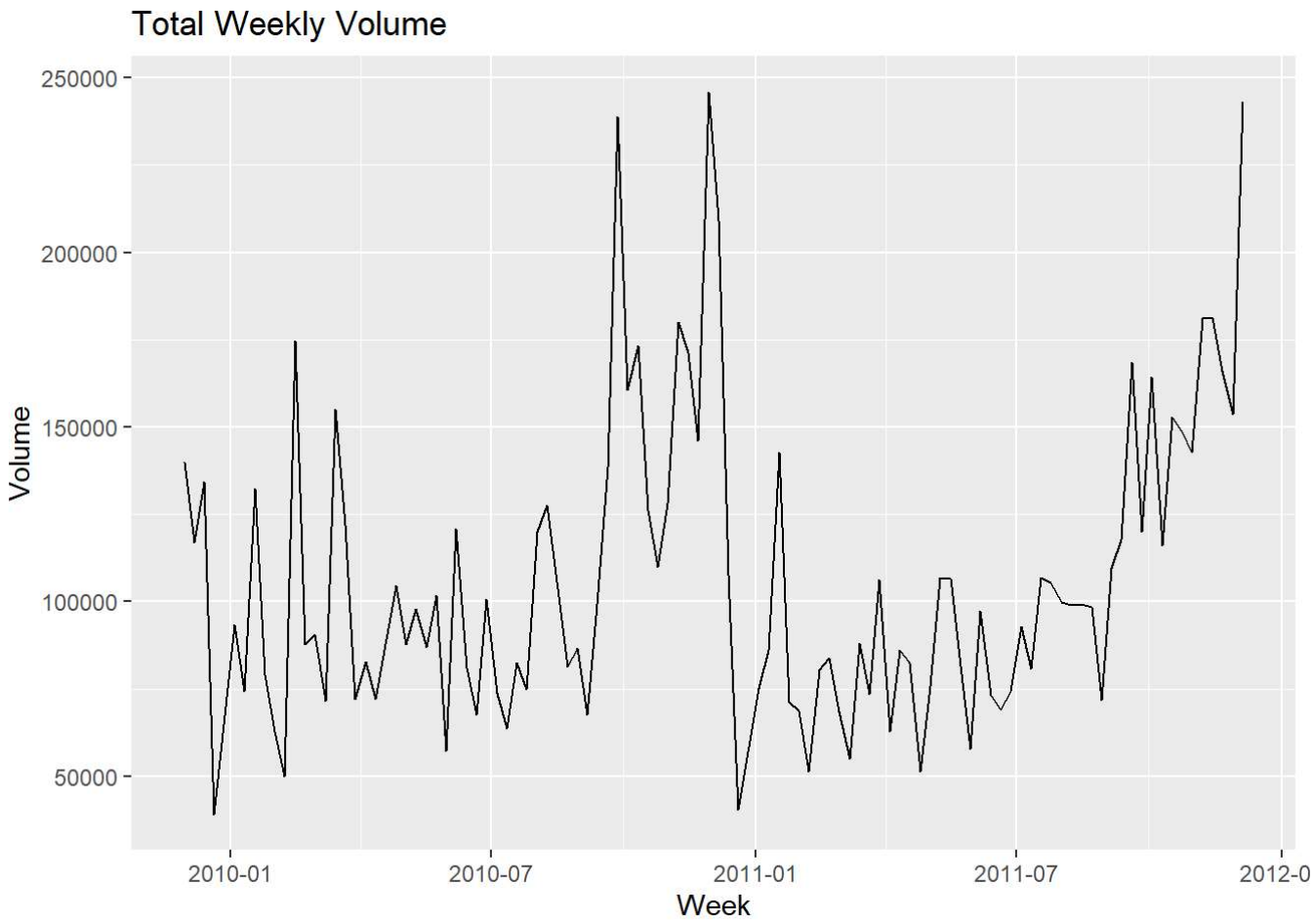
Make a time series plot for total daily volume



According to the total daily volume plot, we can conclude that the series is non-stationary. The variation is non-constant. There are some extreme values in the first half of the series. For example, the total daily volume on 2010-09-27 is 126,146, which is quite high but can still be considered as reasonable since we also got some

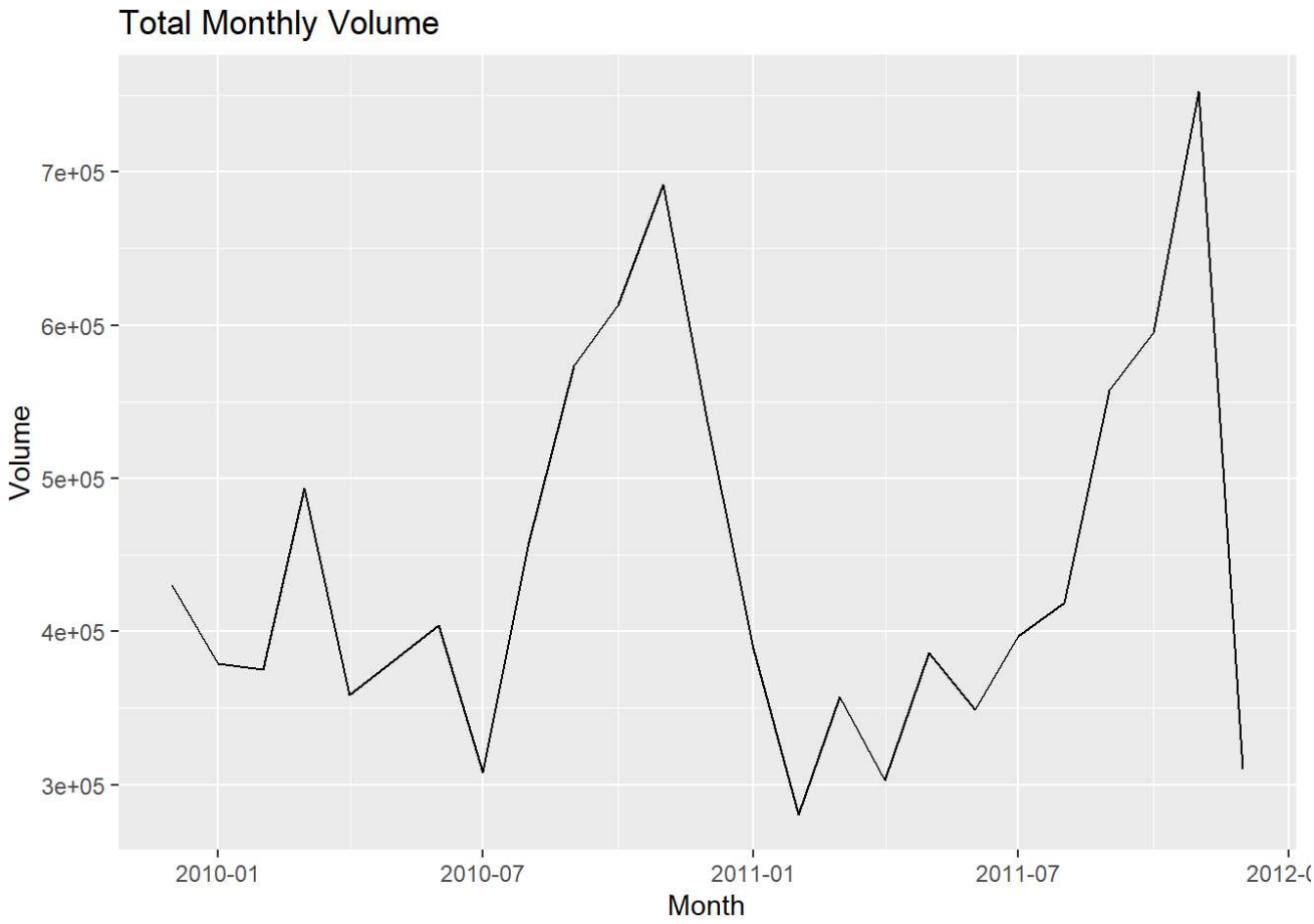
large numbers on 2010-08-09 (total volume is 99,121) and 2010-02-15 (total volume is 96,825). However, we also got some negative values, which is quite unusual. For example, the total daily volume on 2010-08-13 is -31,112. There is no clear evidence shows that the series contains cycles or seasonality.

Make a time series plot for total weekly volume



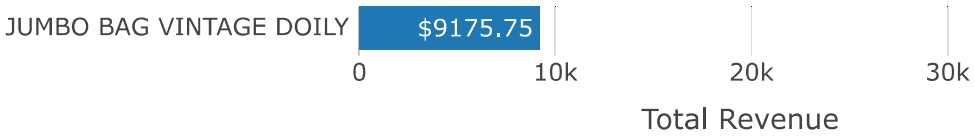
According to the total weekly volume plot, we can clear see that the series is non-stationary. The variation is non-constant and we can see some large values at the middle of the series. The peak occurs on observation 52 (the week of 2010-11-29) but the total volume for the following week is much lower. The plot also shows evidence of seasonality.

Make a time series plot for total monthly volume



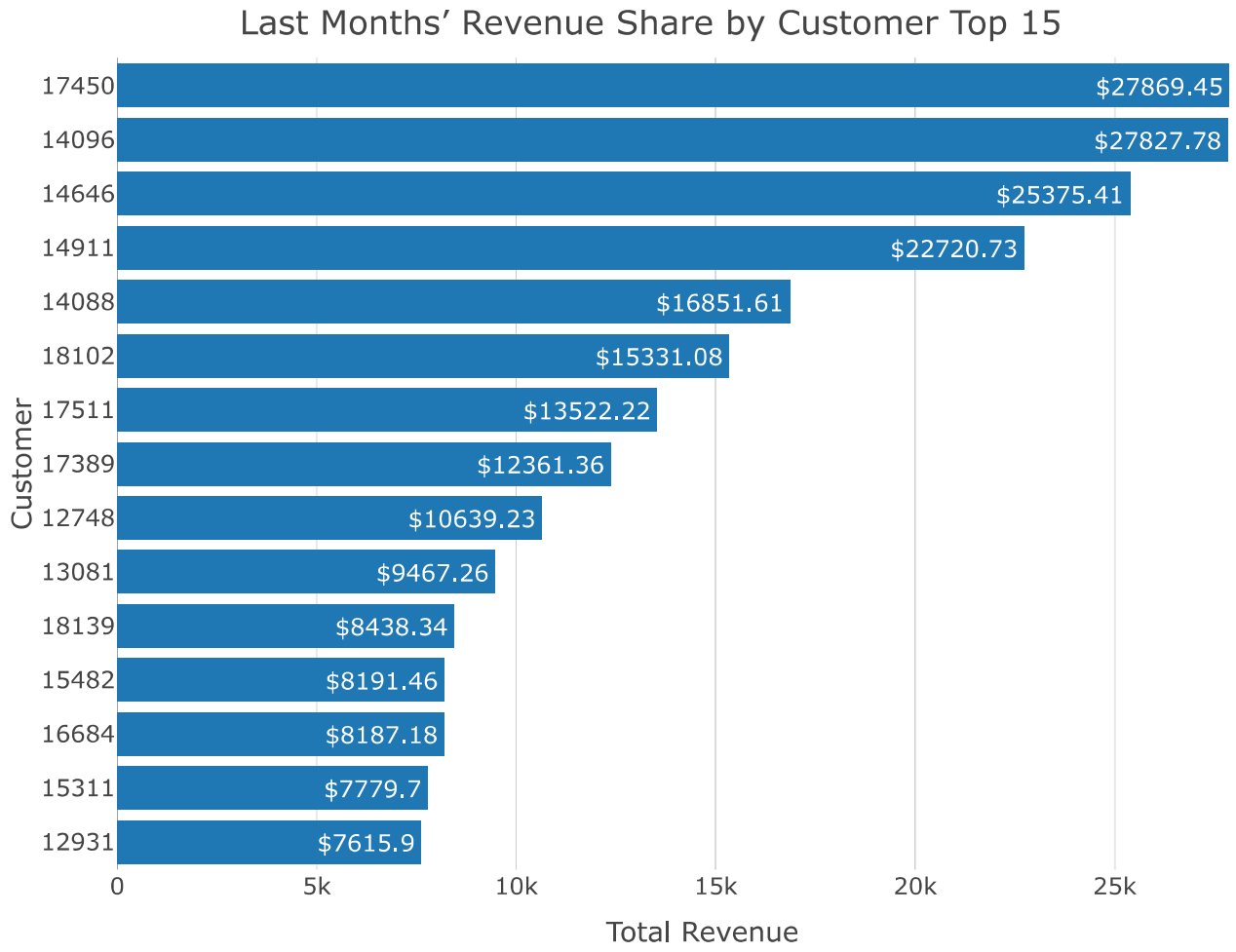
According to the total monthly volume plot, we can clearly see that the series is non-stationary, the variation is non-constant. The first peak occurs at obs 12 (the month of 2010-11-01) with volume of 691,574 and the second peak occurs at obs 24 (the month of 2011-11-01) with volume of 752,907. The total volume for the last month (the month of 2011-12-01) in the dataset is very low. This happens because we only have the data of first 9 days for that month.

Last months' revenue share by product



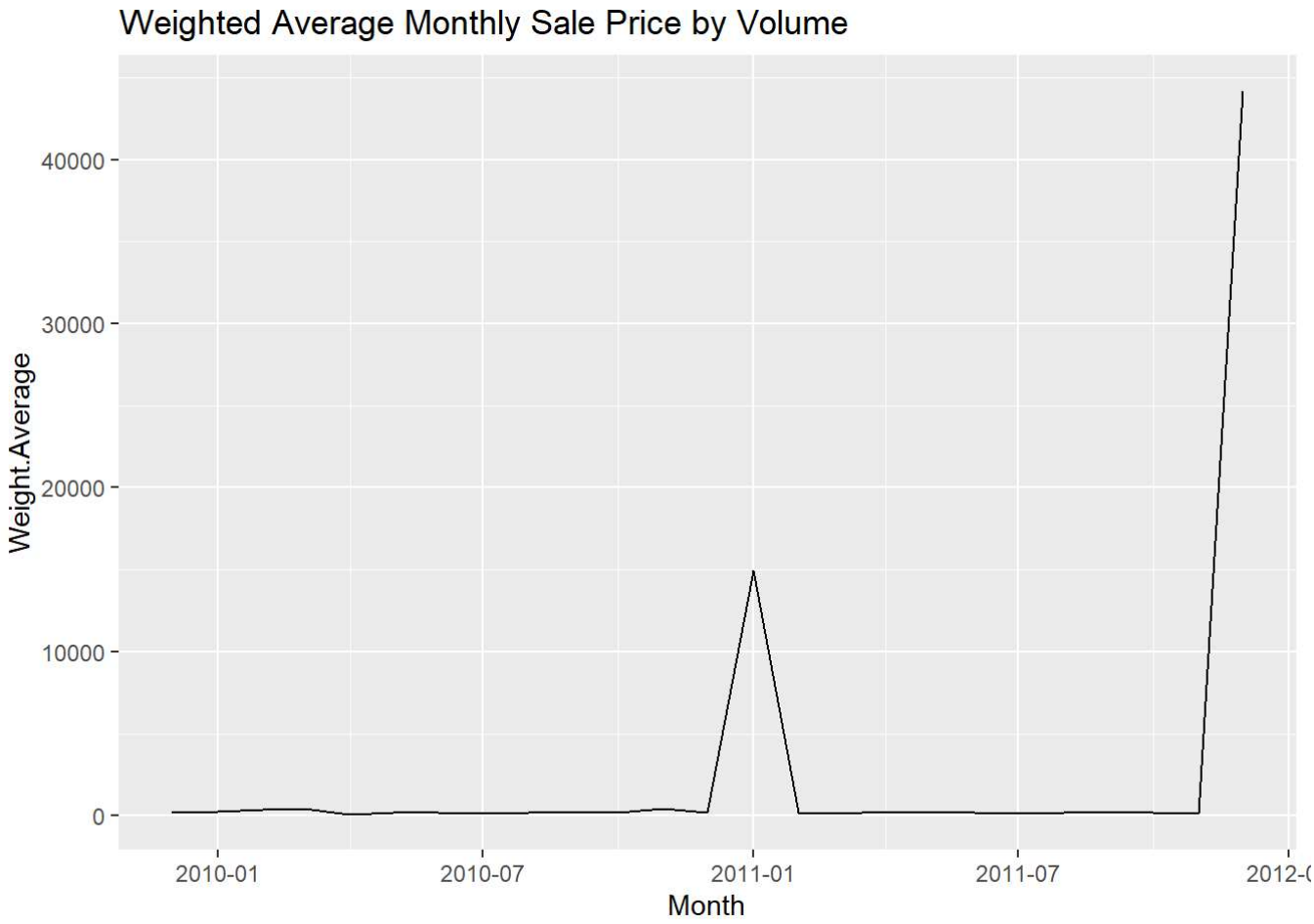
There are 2,955 unique products were sold in last month and the total revenue for last month is 1,509,496 pound sterling. The bar chart shows the top 15 products with the highest revenue in last month. The total revenue of those 15 products is 241,248.1 pound sterling which shares about 15.98% of the total revenue for last month.

Last months' revenue share by customer



There are 1,666 customers made purchase in last month and the total revenue for last month is 1,509,496 pound sterling. The bar chart shows the top 15 customer producing the highest revenue in last month. The total revenue produced by those 15 cuntomers is 222,178.7 pound sterling which shares about 14.72% of the total revenue for last month.

Weighted average monthly sale price by volume



According to the plot, we can see the weighted average sale price for the month of 2011-01-01 is quite large. This happened may because some customers purchase a huge amount of certain products at that month. The weighted average sale price for the month of 2011-12-01 is even larger. However this result may not accurate since we only have the data of first 9 days for that month.

Part 3

Removing negative values

We have some negative values in the variable of quantity which refer to sales returns. Since some of the returns are related to products sold before the data collection date, It is very hard for us to know which return corresponding to which purchase in the past. Thus, just deleting those data may be a quick and resaonable way in order to build the model.

```
modeling.data <- mydata[mydata$Quantity >= 0, ]
```

Part 4

Possible Solutions

There are some different ways to build the model in order to forecas the revenue for this month.

$A\ straight forward\ way$

We can use daily/weekly/monthly total revenue data to forecast the revenue for this month. Since we already know the unit price and sold volume of each product, we can easily calculate the total revenue for each day/week/month. Then we can fit a time series model such as seasonal ARIMA model.

However, since we want to forecast the whole month revenue value, we will need to predict 22 new observations if we use the daily total revenue data to build the model. This is not a very good idea since the variation of new observations will be larger and lagre, which means the accuracy of our prediction will be low. Also, if we use the minthly total revenue data to forecast the future, since the degree freedom is very small (df = 25), the prediction may also be inaccurate.

By contrast, using the weekly total revenue to build the model might be a better idea since we have more degree freedom (df = 104) and we only need to forecast 4 new observations (4 weeks) in order to estimate the total revenue for this month.

A more complex way

We can try to fit a time series model base on the products. That is, we forecast the future for each individual product and then put the results together in order to get the final result. Base the on the data we have, it is possible for us to calcuate the daily/weekly/monthly revenue by each product.

However, since the data set contains more then 5,000 unique products, means we need to build time series models for each of them, which is impossible. A possible way to solve this problem is to bulid a machine learning algorithm to automatically classify those products into different categories, then build time series models for each category to forecast the future. Even though this approach looks possible, we need a lot of time to build a proper classification algorithm and test it's performance to make the prediction be more accurate. Thus, it will be much easier to apply this approach if the client have the category information of their products. Personally, I like this approach since the more information we have, the more accuracy we can get for the forecasting.

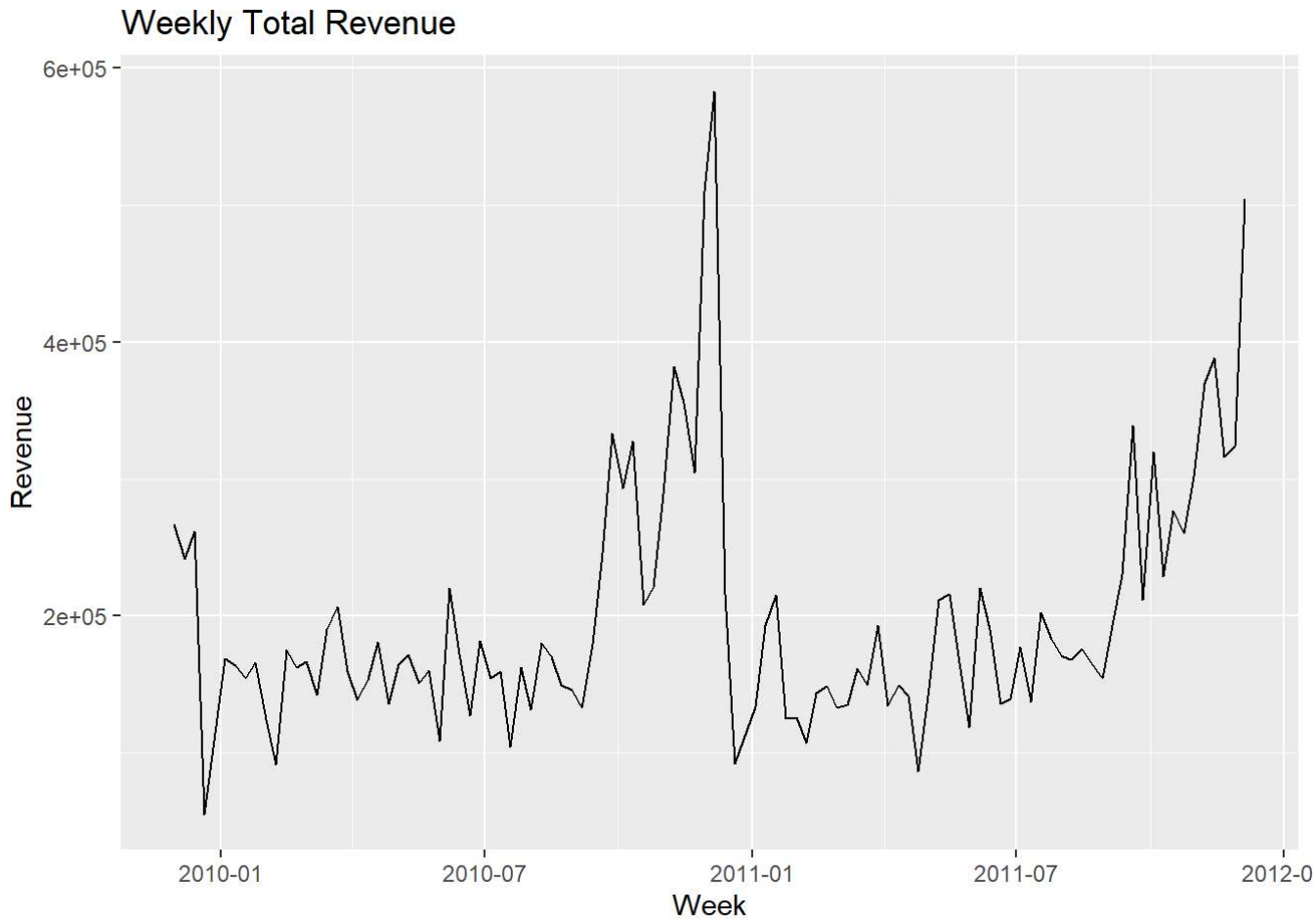
An even more complex way

If it is possible, we can collect the production data from the factories which produce the products in our data set, then we can do some cross-correlation analysis and build a lagged regression model to forecast the future. This approach might work because many manufacturers will estimate the marketing environment to decide the production volume of a certain product. To be specific, a chocolete factory may start to produce more white chocolete from this month if they believe white chocolete will become more popular in the following months. This approach reminds me a very interesting story. Back to the 2016 United States presidential election, a lot of American people believed that Hilary Clinton would become the president at that time. But at the same time, many Chinese factories were producing more products realted to Donald Trump. And now we already knew the election result, Donald Trump became the president of the United States. So some people joked that if we want to know who will be the next president of the United States, just go to ask the owns of those Chinese factories.

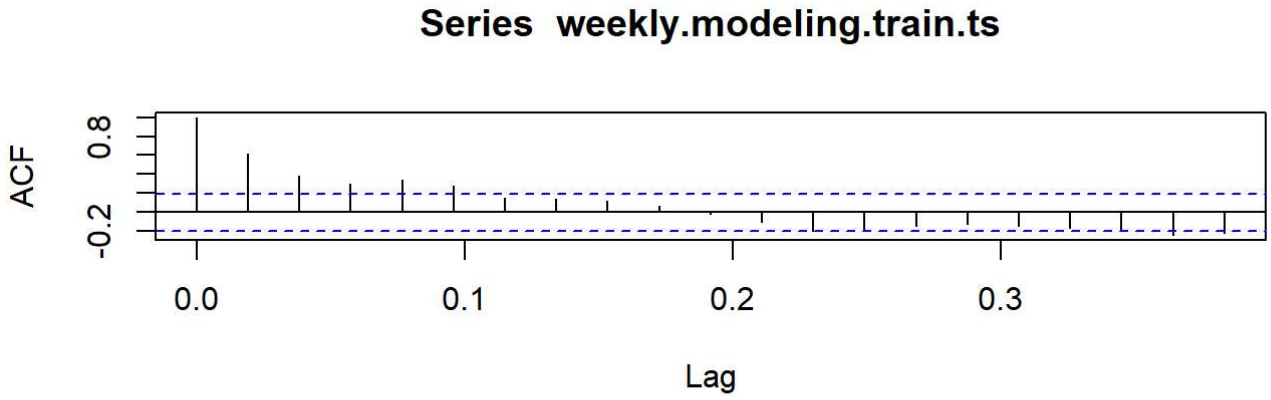
However, this cross-correaltion approach has a lot of uncertainties. For example, we don't know if those factories would like of share their production data with us. Even they do, there is no way for us to know whether their estimation for the future is accurate enough. And collecting data from those factories may take us a lot of time.

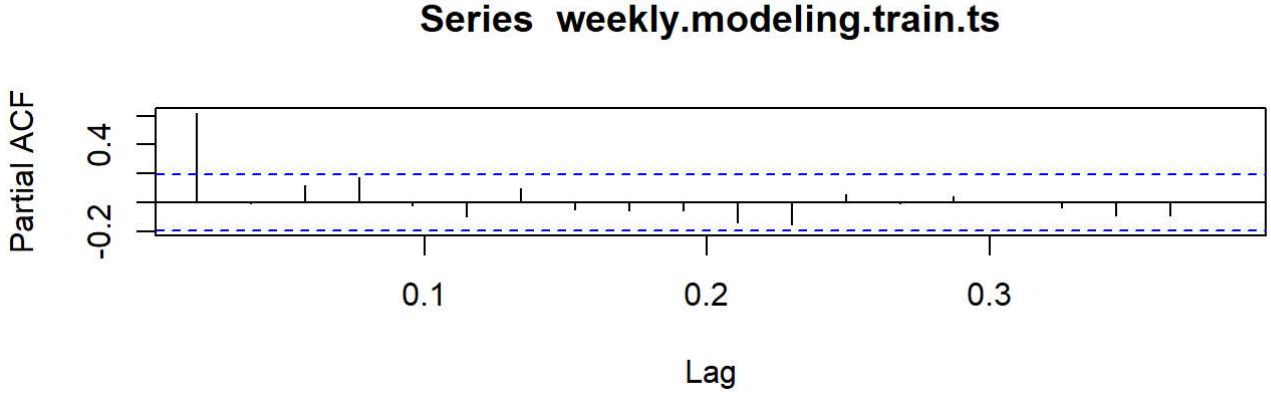
Under current situation, I think fitting a SARIMA model is a good start point.

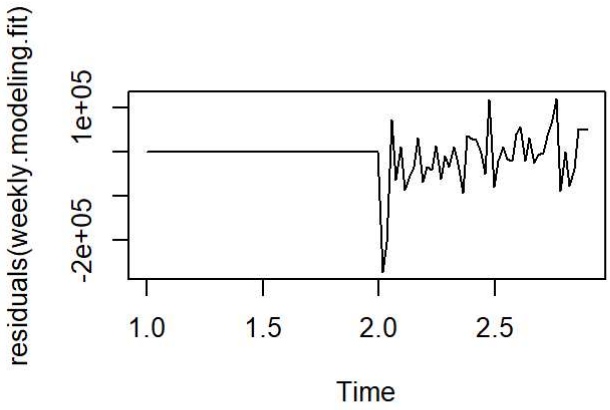
Model Fitting For Weekly Total Revenue(SARIMA)

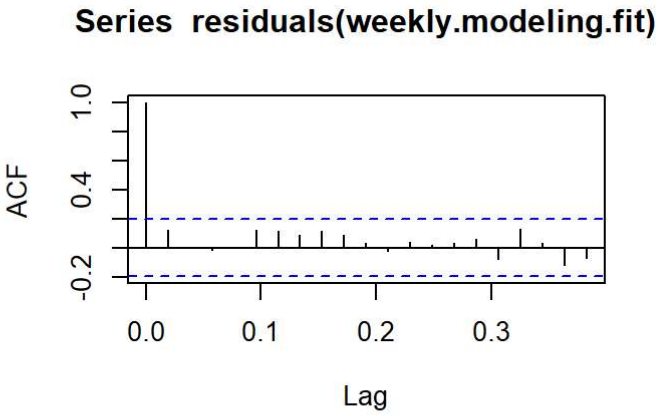


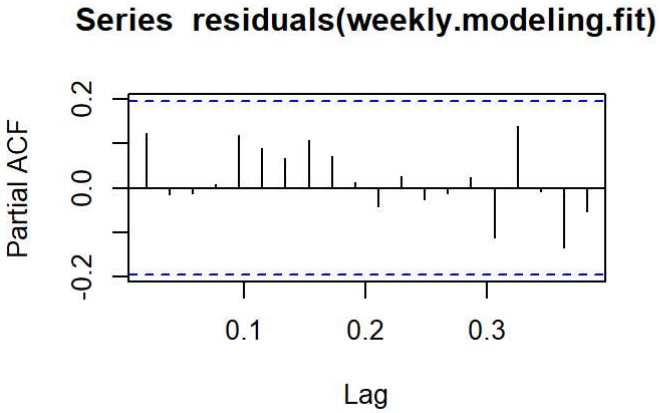
According to the weekly total revenue plot, we can see there are some large values at the middle and the end of the series. Also, the plot shows a non-linear trend in the second half of the series. We can also find some evidence of seasonality in this plot.

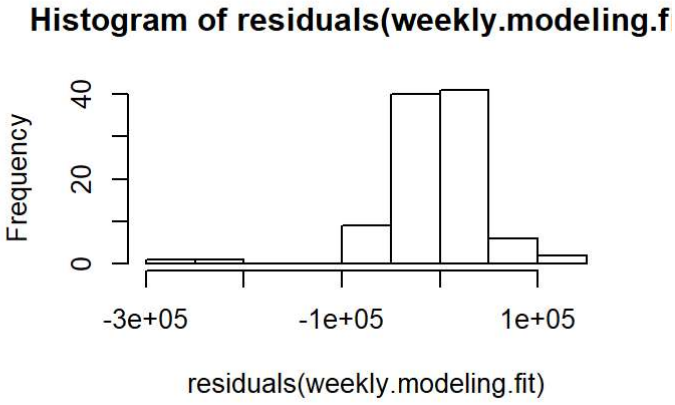












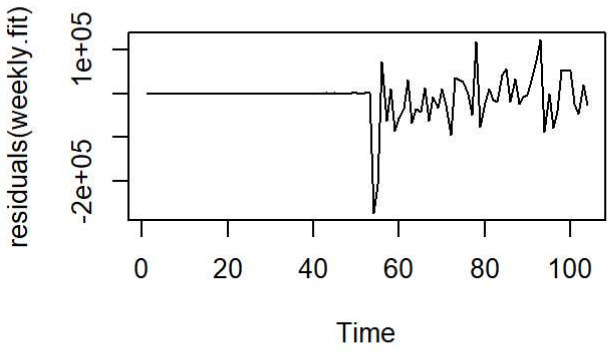

```
## Time Series:
## Start = 2.91649555099247
## End = 2.97399041752225
## Frequency = 52.1785714285714
##            actual Pred.sarima$pred
## 2.916496 388622.4         413580.7
## 2.935661 316412.2         370979.5
## 2.954825 324370.8         319844.5
## 2.973990 504327.9         525568.4
```

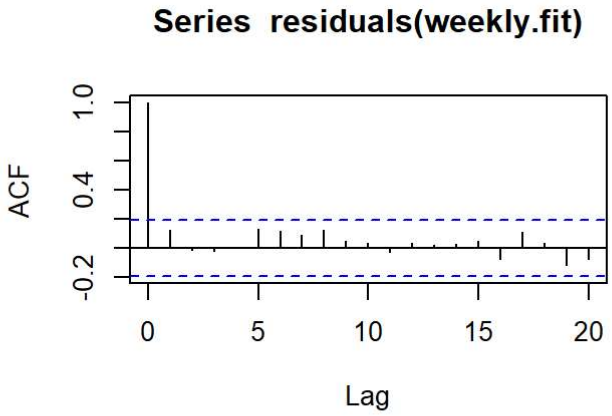
```
## [1] "The RMSEP for predicted values is 31906.71"
```

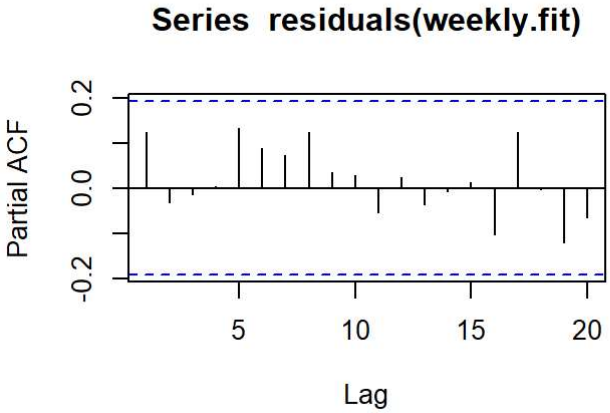
I removed the last four observations from weekly total revenue data set to build a training data set. The fitted model is SARIMA model with $ARIMA(0,1,2) \times (0,1,0)_{52}$. Although the residual plot shows some large values at the middle of the residual series, the there is no significant lag in ACF and PACF plot, which means the residuals are white noise. The histogram for residuals shows left skewness but the main part of the residual is symmetric. Thus, we can canclude that the model assumptions are roughly satisfied and ready to make prediction. I also tried some other SARIMA models but nothing is better than this one since they either have significant lags in ACF/PACF plot or have larger AIC values.

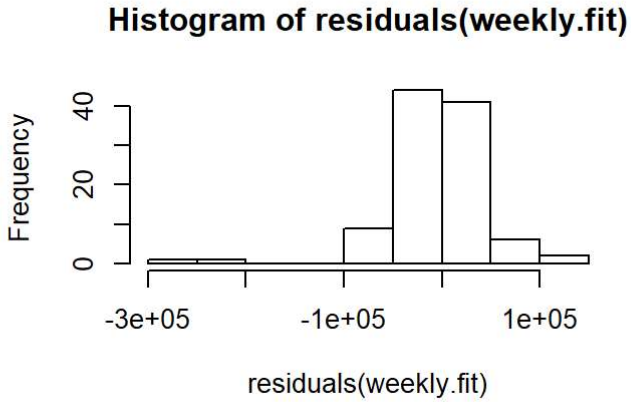
The predicted total revenue for next four weeks are 413580.7, 370979.5, 319844.5 and 525568.4 pound sterling respectively. Comparing them with the actual values, we can get the RMSEP value which is 31906.71, which is not a bad number. Actually the smallest RMSEP value I got from another model is about 29000 but that model seriously violates the model assumptions.

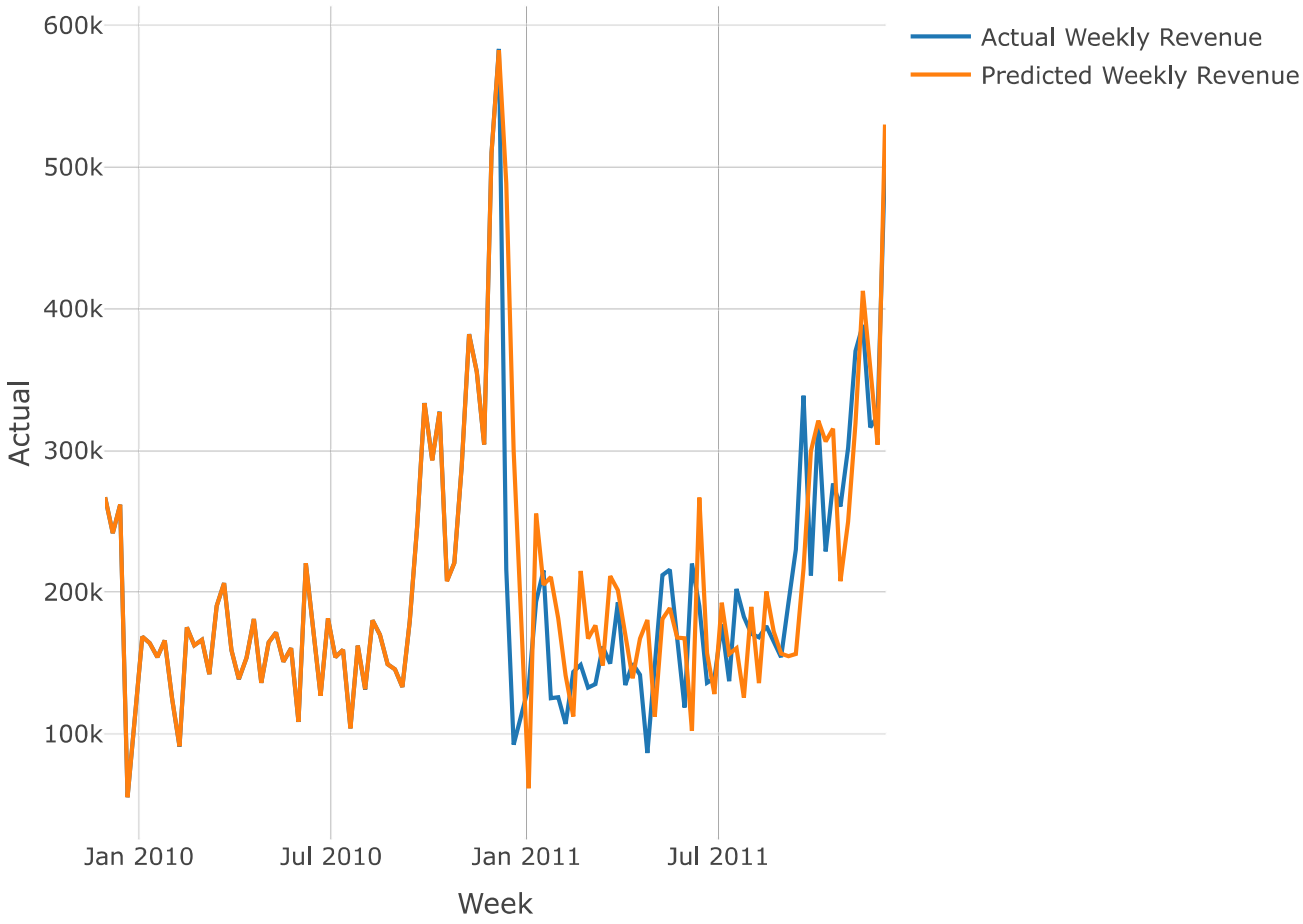
Now we can use the full weekly revenue data set to fit the model and make predictions.

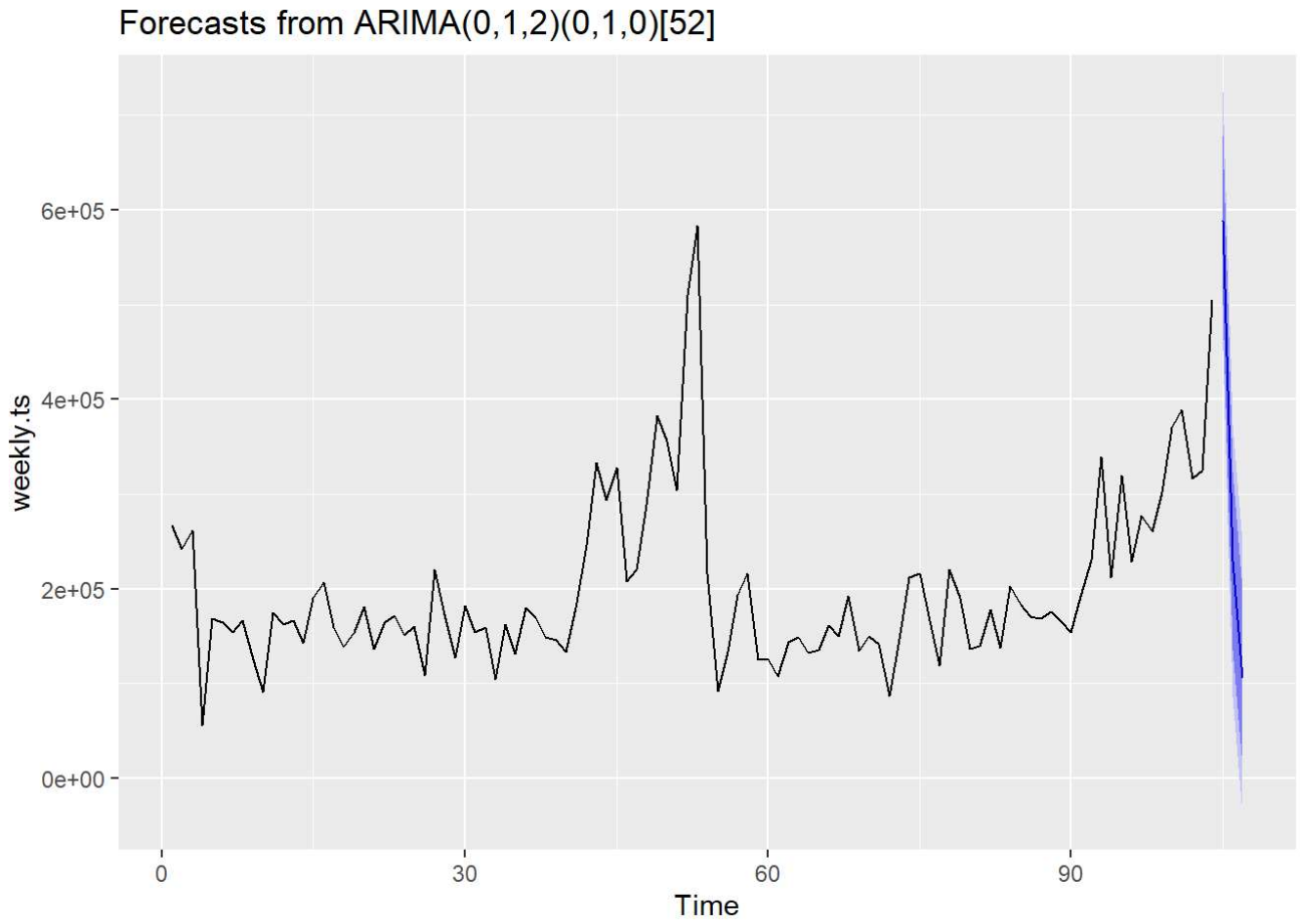

residuals(weekly.fit)
1e+05
-2e+05
0
20
40
60
80
100
Time


Series residuals(weekly.fit)
ACF
1.0
0.4
-0.2
0
5
10
15
20
Lag


Series residuals(weekly.fit)
Partial ACF
0.2
0.0
-0.2
5
10
15
20
Lag


Histogram of residuals(weekly.fit)
Frequency
40
20
0
-3e+05
-1e+05
1e+05
residuals(weekly.fit)


Actual Weekly Revenue
Predicted Weekly Revenue
600k
500k
400k
300k
200k
100k
Actual
Jan 2010
Jul 2010
Jan 2011
Jul 2011
Week



According to the diagnostic plots, we can conclude that all model assumptions are satisfied. The predited weekly total revenue values are*.

Week Number	**Revenue**
2011-12-05	504327.9**
2011-12-12	588850.4
2011-12-19	229368.1
2011-12-26	105725.2
Total for Dec. 2011	1428272.6

[*] Please note, I treat Monday as the first day of a week.

[**] The weekly total revenue for the week of 2011-12-05 comes from the actual data instead of model predictions.

Base on my analytical results and forecasting, the estimated total revenue for December 2011 is about 1428272.6 pound sterling and I am confident with my results. Thus, it is a considerable option for the owner of the online retailer to buy a new Ferrari for his partner.

Since the variance of the data we have seems changes over time. I will also try some other time series methods such as ARCH model if I have additional time.