

时间序列学期报告

应用统计硕士 2020270026 王姿文

这篇使用的数据来自于Kaggle，为真实游戏数据，此份报告使用jupyter完成，为了报告内容清晰故不涵盖coding，coding附于project.py内。

以下为本篇报告的大纲：

1. 数据介绍

2. 移动&平滑

2.1 Moving Average Method

2.2 Smoothing Method

2.2.1 Exponential Smoothing

2.2.2 Double Exponential Smoothing

2.2.3 Triple Exponential Smoothing(Holt-Winters Model)

3. ARIMA

3.1 平稳性检验

3.2 Seasonal Autoregression Moving Average model

1. 数据介绍

- 数据来源：Kaggle
- 数据内容：手游玩家每小时观看的广告量以及手游玩家每天的游戏币消费情况，为两个数据的结合
- 选择原因：两个数据分别为短期时间序列数据和中长期时间序列数据，因此可以同时分析两种类型的数据

1.1 手游戏玩家每小时观看的广告量

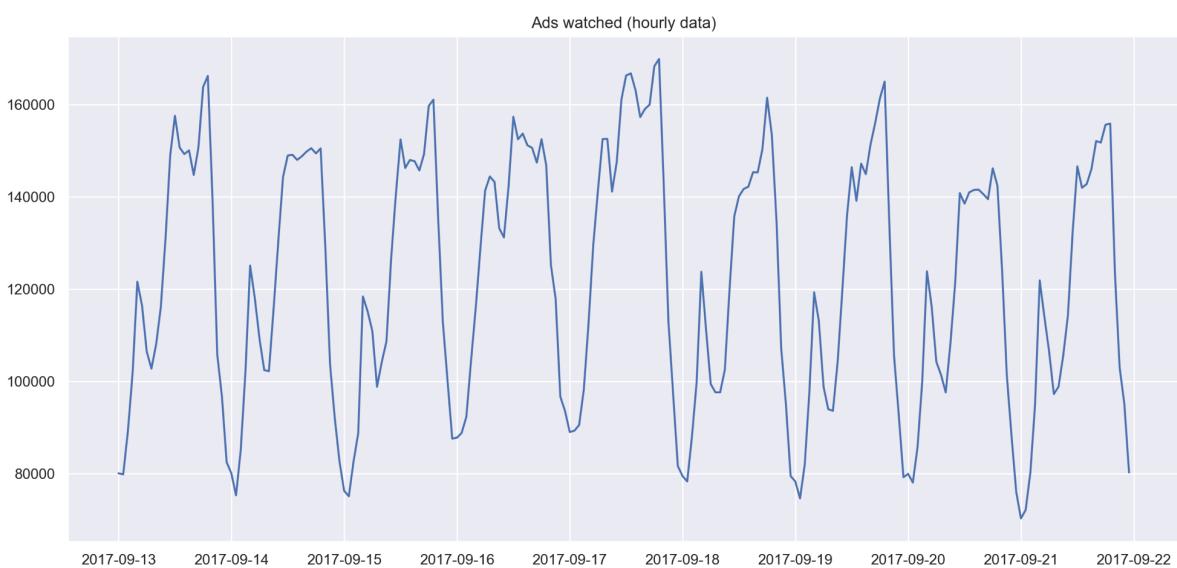
- 时间范围：2017-09-13 00:00:00 ~ 2017-09-21 23:00:00
- 前十笔数据：

Ads

Time

Time	Ads
2017-09-13 00:00:00	80115
2017-09-13 01:00:00	79885
2017-09-13 02:00:00	89325
2017-09-13 03:00:00	101930
2017-09-13 04:00:00	121630
2017-09-13 05:00:00	116475
2017-09-13 06:00:00	106495
2017-09-13 07:00:00	102795
2017-09-13 08:00:00	108055
2017-09-13 09:00:00	116125

- 这九天内手游玩家每小时观看的广告量趋势图：



1.2 手游戏玩家每天的游戏币消费

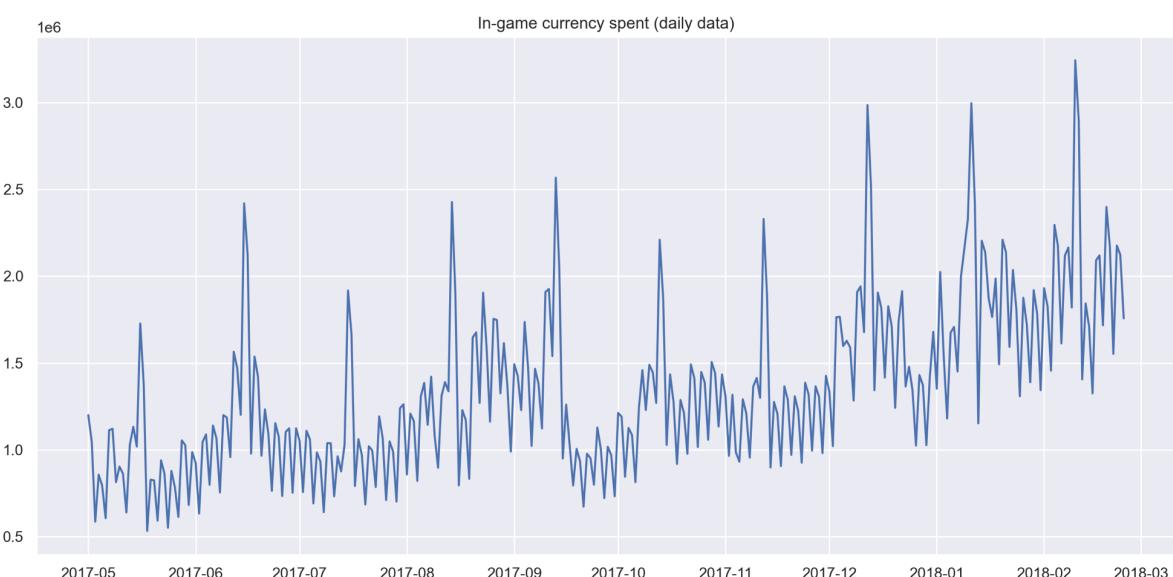
- 时间范围：2017-05-01 ~ 2018-02-24
- 前十笔数据：

GEMS_GEMS_SPENT

Time

Time	GEMS_GEMS_SPENT
2017-05-01	1199436
2017-05-02	1045515
2017-05-03	586111
2017-05-04	856601
2017-05-05	793775
2017-05-06	606535
2017-05-07	1112763
2017-05-08	1121218
2017-05-09	813844
2017-05-10	903343

- 这九个月内手游玩家每天的游戏币消费趋势图：



2. Moving Average Method & Smoothing Method

2.1 Moving Average Method

假设未来某个值的预测，取决于它前面的 n 个数的平均值 \Rightarrow 以*moving average*(移动平均数)作为预测值：

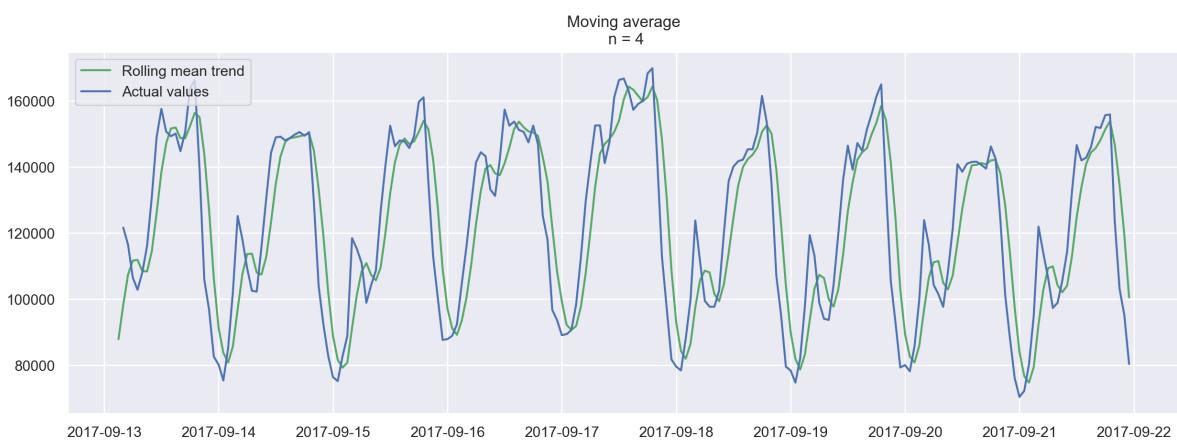
$$\hat{y}_t = \frac{1}{k} \sum_{n=0}^{k-1} y_{t-n}$$

1. *moving average*(移动平均数)比较适合短期预测
2. *moving average*(移动平均数)可对原始的时间序列数据进行平滑处理，以找到数据的变化趋势

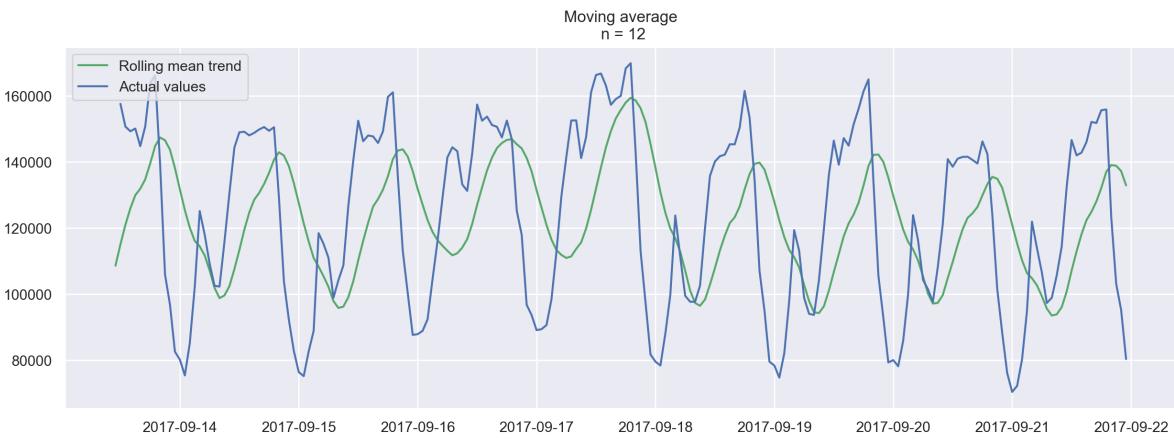
2.1.1 手游戏玩家每小时观看的广告量

由下图可看见， n 越大则数据预测趋势线越平滑，且 n 越小越接近原始数据趋势

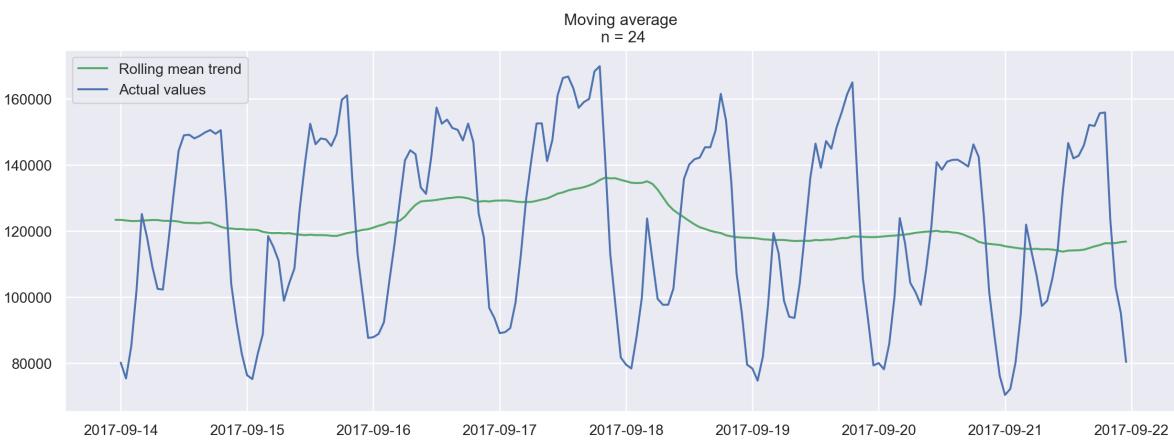
对过去4小时的广告浏览量的预测： 100478.75



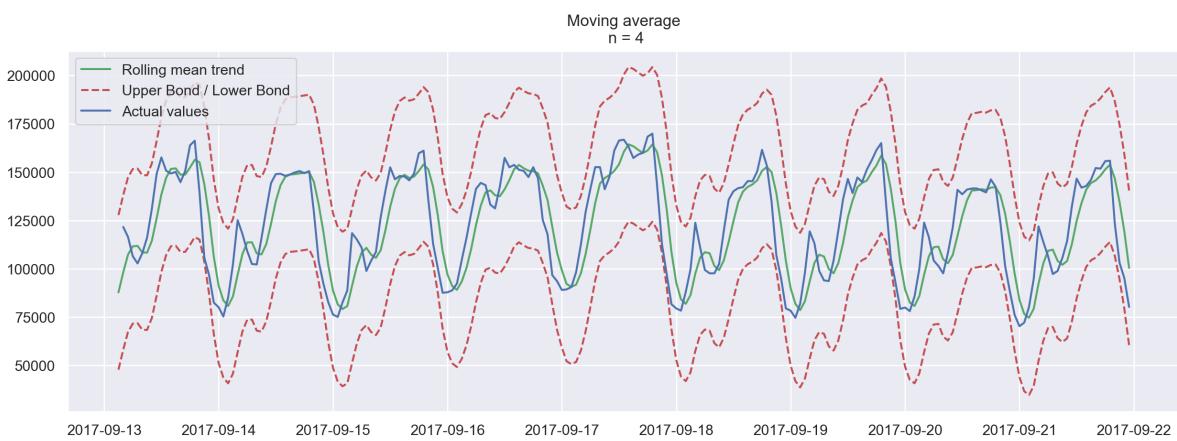
对过去12小时的广告浏览量的预测： 132903.3333333334

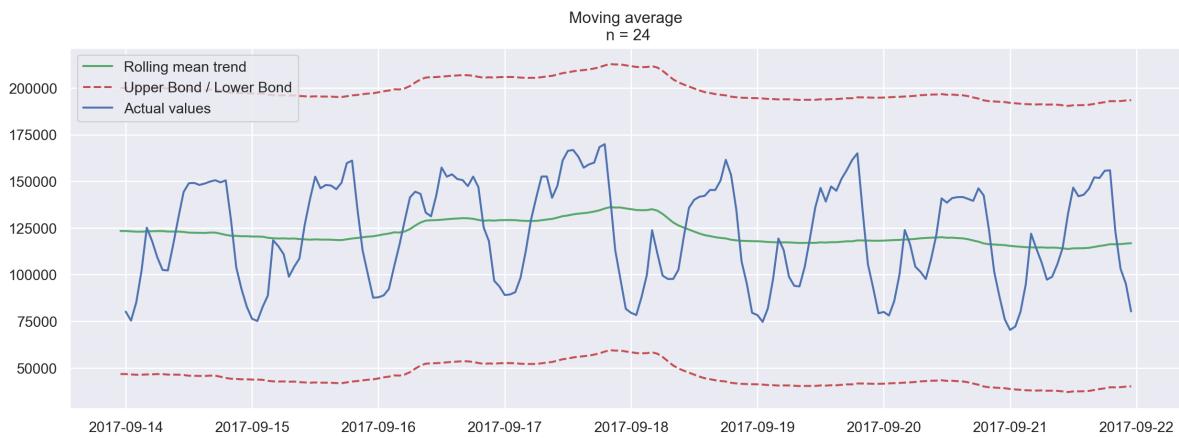
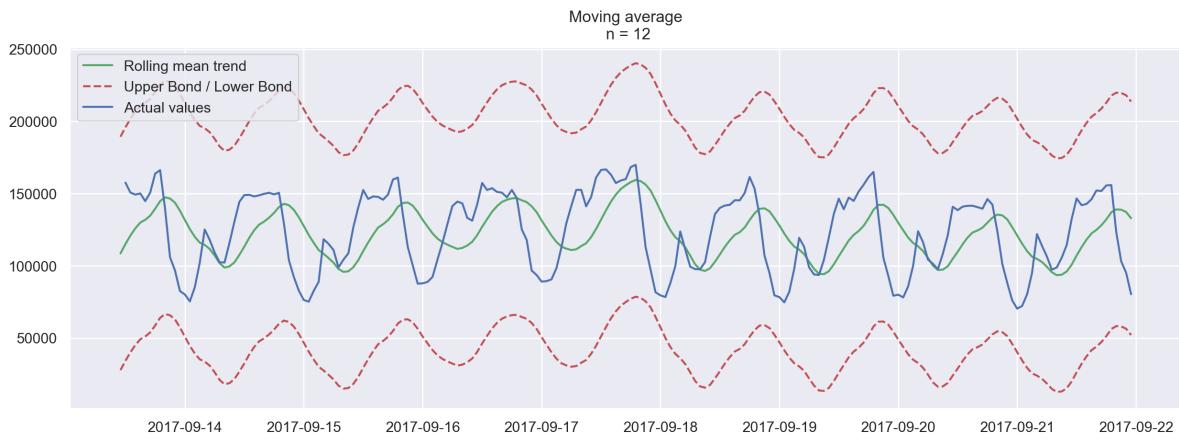


对过去24小时的广告浏览量的预测： 116805.0



此外还能以信赖区间来检验异常值，可见均无异常值

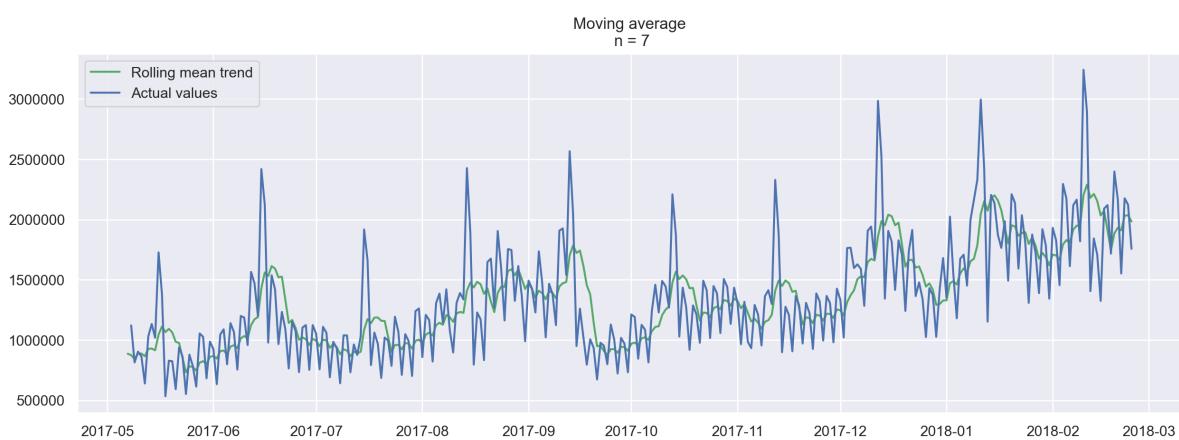




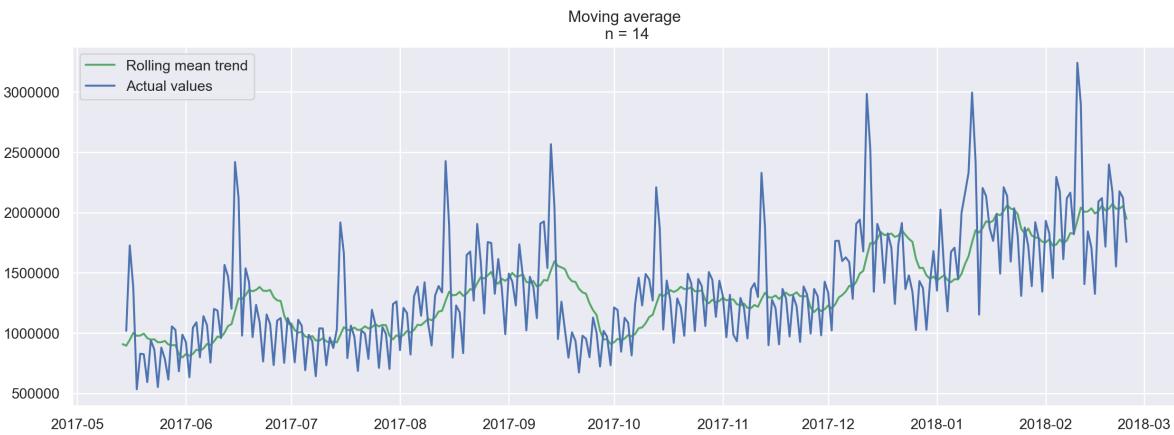
2.1.2 手游戏玩家每天的游戏币消费

一样由下图可看见，n越大则数据预测趋势线越平滑，且n越小越接近原始数据趋势

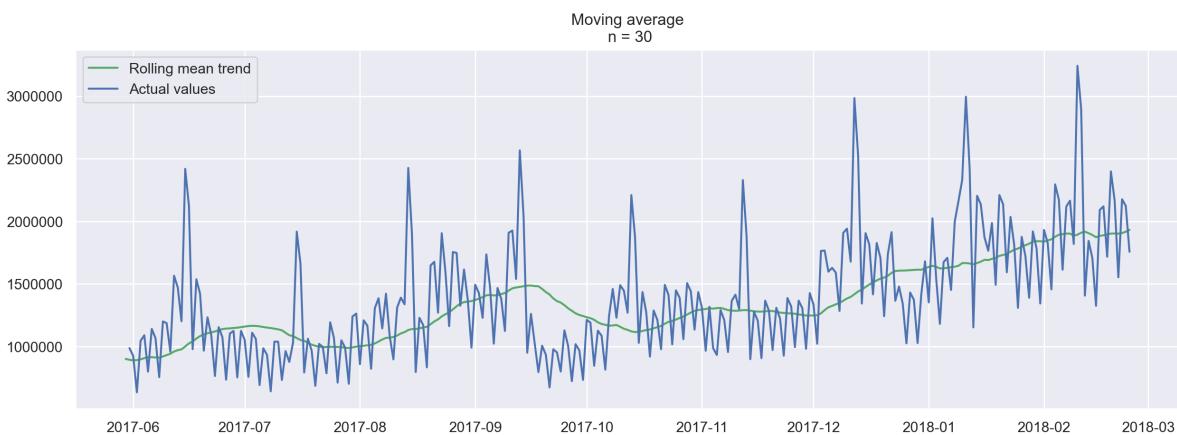
对过去7天游戏币消费的预测： 1983998.2857142857



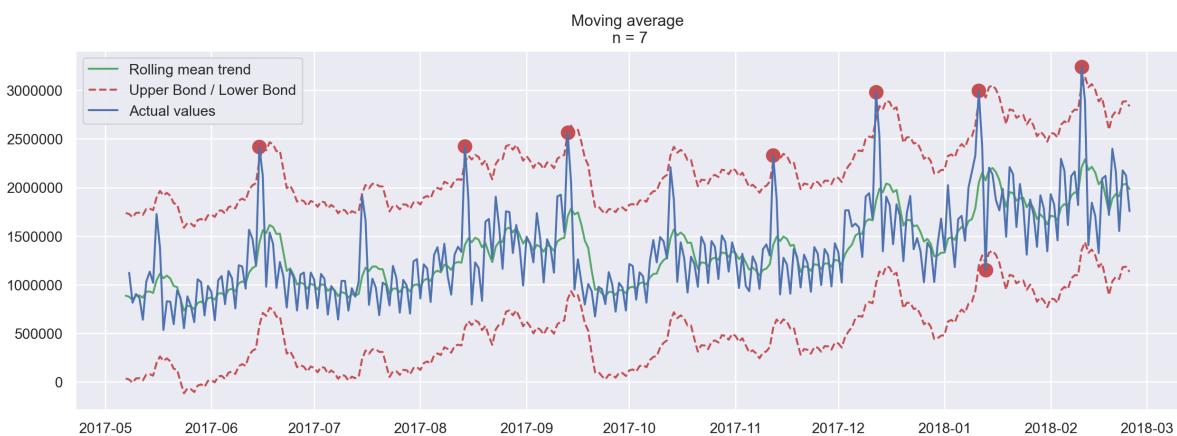
对过去14天游戏币消费的预测： 1948071.5714285714

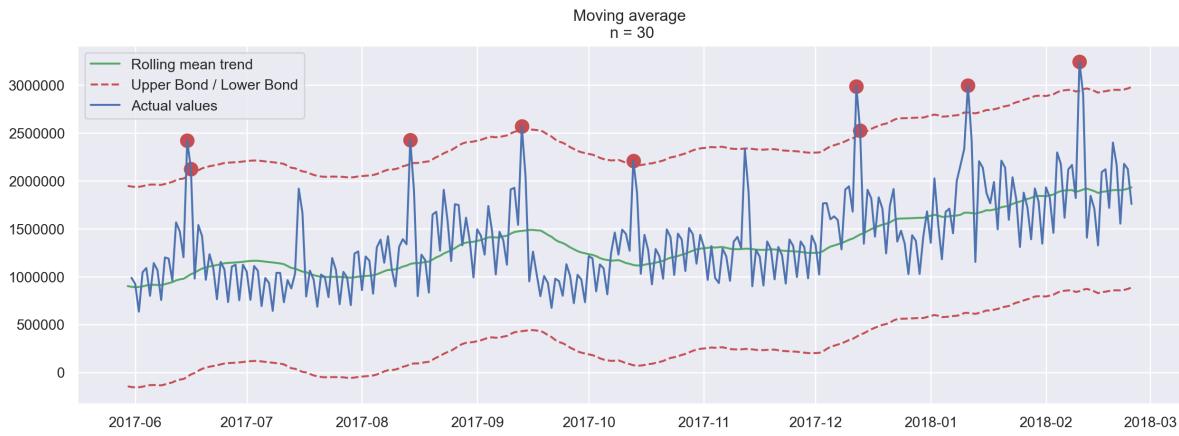
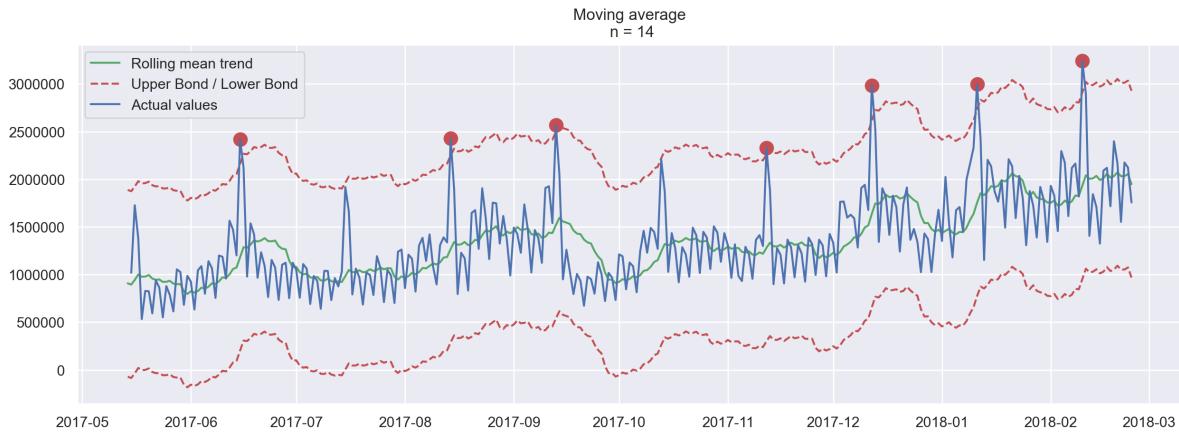


对过去30天游戏币消费的预测： 1931348.8



再来以信赖区间来检验异常值，均见到异常值，它没有在数据中捕捉到每个月中出现的季节性变化，反倒几乎把所有每隔7/14/30天出现的峰值标记为异常





2.1.3 小结

由上述两个数据能看出方法的合适与否与数据特性高度相关，*moving average*的合适度为手游玩家每小时观看的广告量>手游玩家每天的游戏币消费情况

2.2 Smoothing Method

对目前所有的已观测数据进行加权处理，并且每一个数据点的权重，呈指数形式递减

a. Exponential Smoothing

α 表示平滑因子，它定义我们“遗忘”当前真实观测值的速度有多快， α 越小，表示当前真实观测值的影响力越小，而前一个模型预测值的影响力越大，最终得到的时间序列将会越平滑

$$\hat{y}_t = \alpha y_t + (1 - \alpha)\hat{y}_{t-1}$$

- 单指数平滑的特点：能够追踪数据变化。预测过程中，添加了最新的样本数据之后，新数据逐渐取代老数据的地位，最终老数据被淘汰
- 单指数平滑的局限性：第一，预测值不能反映趋势变动、季节波动等有规律的变动；第二，这个方法多适用于短期预测，而不适合中长期的预测；第三，由于预测值是历史数据的均值，因此与实际序列相比，有滞后的现象
- 单指数平滑的系数：Eviews提供两种确定指数平滑系数的方法：自动给定和人工确定。一般来说，如果序列变化比较平缓，平滑系数值应该比较小，比如小于0.1；如果序列变化比较剧烈，平滑系数值可以取得大一些，如0.3~0.5。若平滑系数值大于0.5才能跟上序列的变化，表明序列有很强的趋势，不能采用一次指数平滑进行预测

b. Double Exponential Smoothing

由于单指数平滑在产生新的序列的时候，考虑了前面的K条历史数据，但是仅仅考虑其静态值，即没有考虑时间序列当前的变化趋势 \Rightarrow 如果当前的时间序列数据处于上升趋势，那么当我们对明天的数据进行预测的时候，就不应该仅仅是“对历史数据进行‘平均’”，还应考虑到当前数据变化的上升趋势。同时考虑历史平均和变化趋势，这个就是双指数平滑法

通过序列分解法(series decomposition)，可以得到两个分量， $l = \text{intercept(also, level)}$ ， $b = \text{trend(also, slope)}$ ，根据前面学习的方法，知道了如何预测 l ，可以将同样的指数平滑法应用到 b 上

$$\begin{aligned} l_t &= \alpha y_t + (1 - \alpha)(l_{t-1} - b_{t-1}) \\ b_t &= \beta(l_{t-1} - b_{t-1}) + (1 - \beta)b_{t-1} \\ \hat{y}_{t+1} &= l_t + b_t \end{aligned}$$

- α 决定时间序列数据自身变化趋势的平滑程度， β 决定趋势的平滑程度

c. Triple Exponential Smoothing(Holt-Winters Model)

与前两种平滑方法相比，这次多考虑了一个因素seasonality(季节性)，故如果时间序列数据不存在季节性变化，就不适合用三指数平滑 \Rightarrow 模型中的季节性分量($= s_t$)，用来解释截距和趋势的重复变化，并且由季节长度来描述，也就是变化重复的周期来描述，其中 γ = 指数平滑的权重， \hat{y}_{t+m} = 未来m步之后的预测值

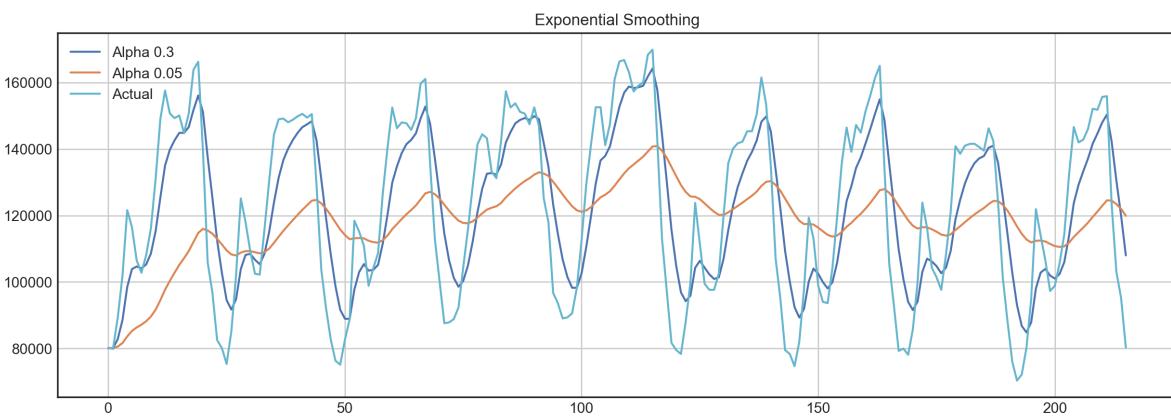
$$\begin{aligned} l_t &= \alpha(y_t - s_{t-L}) + (1 - \alpha)(l_{t-1} - b_{t-1}) \\ b_t &= \beta(l_{t-1} - b_{t-1}) + (1 - \beta)b_{t-1} \\ s_t &= \gamma(y_t - l_t) + (1 - \gamma)s_{t-L} \\ \hat{y}_{t+m} &= l_t + mb_t + s_{t-L+1} + (m-1)\text{mod } L \end{aligned}$$

2.2.1 Exponential Smoothing

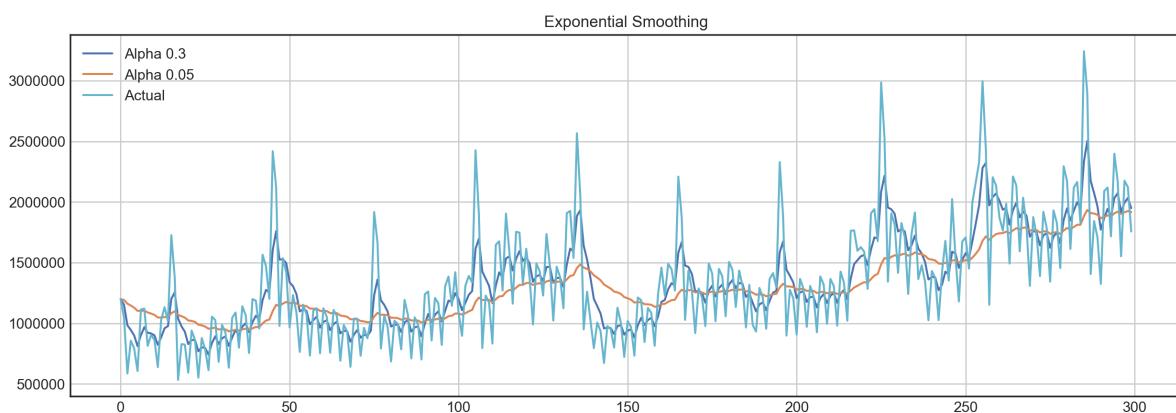
由下图可看见， α 越大，表示当前真实观测值的影响力越大，而前一个模型预测值的影响力越大，最终得到的时间序列将会越平滑（类似于moving average内的n越大则趋势图越平滑），由alpha的数值以及趋势图来看，两个数据的时序变化确实较为激烈

其中手游玩家每小时观看的广告量有上升趋势，因此可以考虑Double Exponential Smoothing或Triple Exponential Smoothing(Holt-Winters Model)

2.2.1.1 手游戏玩家每小时观看的广告量



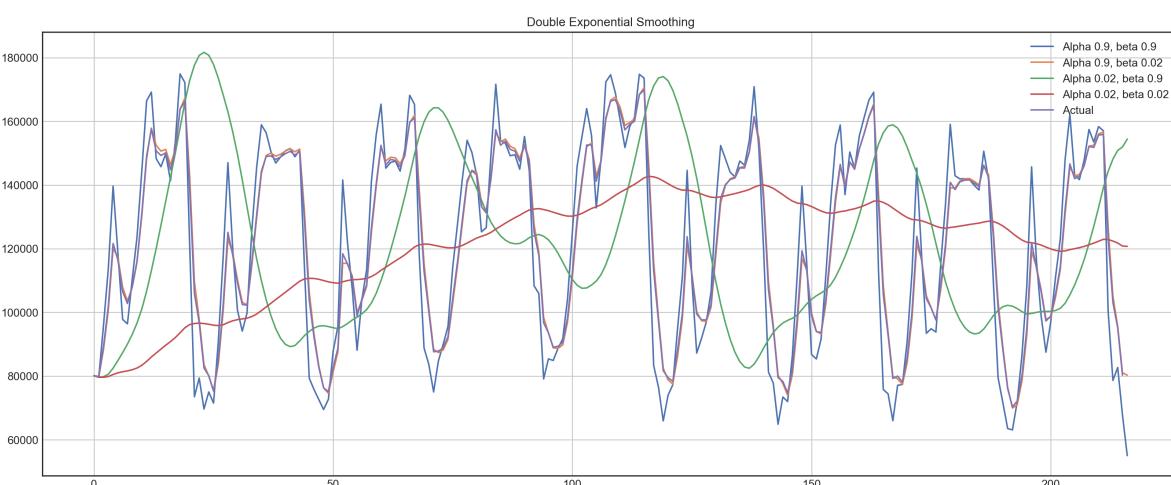
2.2.1.2 手游玩家每天的游戏币消费



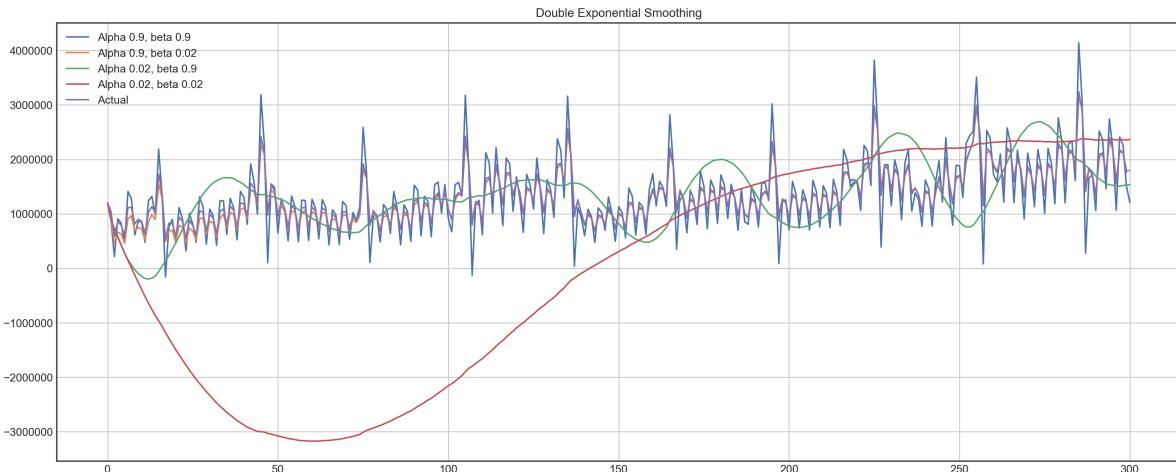
2.2.2 Double Exponential Smoothing

α 决定时间序列数据自身变化趋势的平滑程度， β 决定趋势的平滑程度，下图为不同 α & β 的组合，与Exponential Smoothing最明显不同之处在于手游玩家每小时观看的广告量的上升趋势被预测地明显许多

2.2.2.1 手游戏玩家每小时观看的广告量



2.2.2.2 手游玩家每天的游戏币消费



2.2.3 Triple Exponential Smoothing(Holt-Winters Model)

相较上述两个方式，此方法考量季节性变量

下面将应用Holt-Winters Model于两个数据，会使用到滚动交叉验证，以找到模型最佳的参数（对于时间序列数据，数据间存在时间的依赖性，我们就不能再随机划分数据集，否则会导致数据中的时间结构被破坏了，因此使用滚动交叉验证）

并且在Holt-Winters 模型中，对平滑参数的大小有一个限制，每个参数都需在0到1之间。因此必须选择支持模型参数约束的最优化算法，此数使用Truncated Newton conjugate gradient (截断牛顿共轭梯度法)

2.2.3.1 手游玩家每小时观看的广告量

所谓的季节性变化不一定得是真正的“季节”，这只是一个说法，真正意思比较像是周期变化，如这个数据就可以以一日为季节（也就是24小时），故可以设定变化周期为24小时

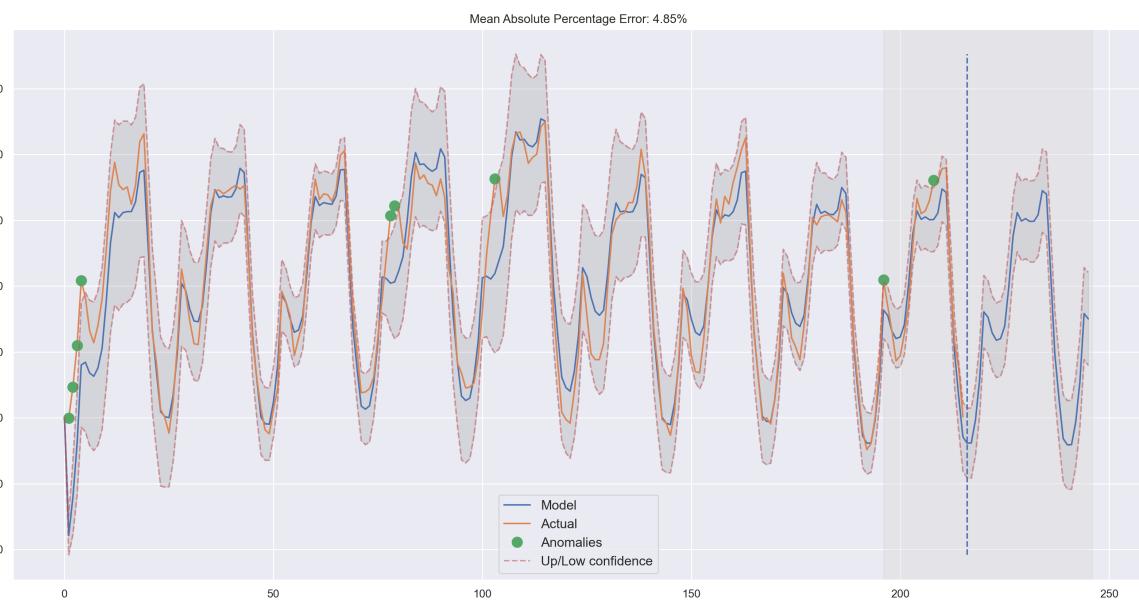
此为训练后得到的最合适参数组合，也就是可以使得Loss Function最小：

```
alpha_final: 0.11676236693712227 beta_final: 0.0026881337430822994
gamma_final: 0.055312622299154346
CPU times: user 1.11 s, sys: 24.1 ms, total: 1.13 s
Wall time: 1.12 s
```

下图为拟和出来的模型，以及实际值：

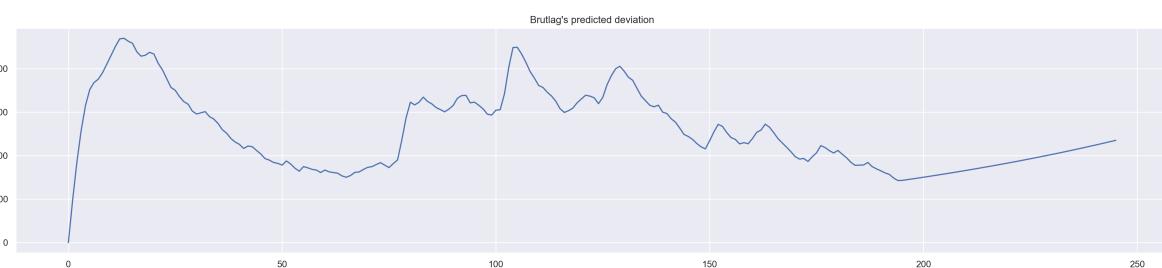


也能从信赖区间捕捉异常值，会发现异常值还是存在的



此为模型偏差图，可以清楚地看到模型对序列结构的变化反应非常强烈，但是很快就会把偏差恢复到正常值，“遗忘”过去

该模型的这一特性能快速构建异常检测系统，即使对于非常嘈杂的系列，也不需要花费太多的时间和金钱来准备数据和训练模型



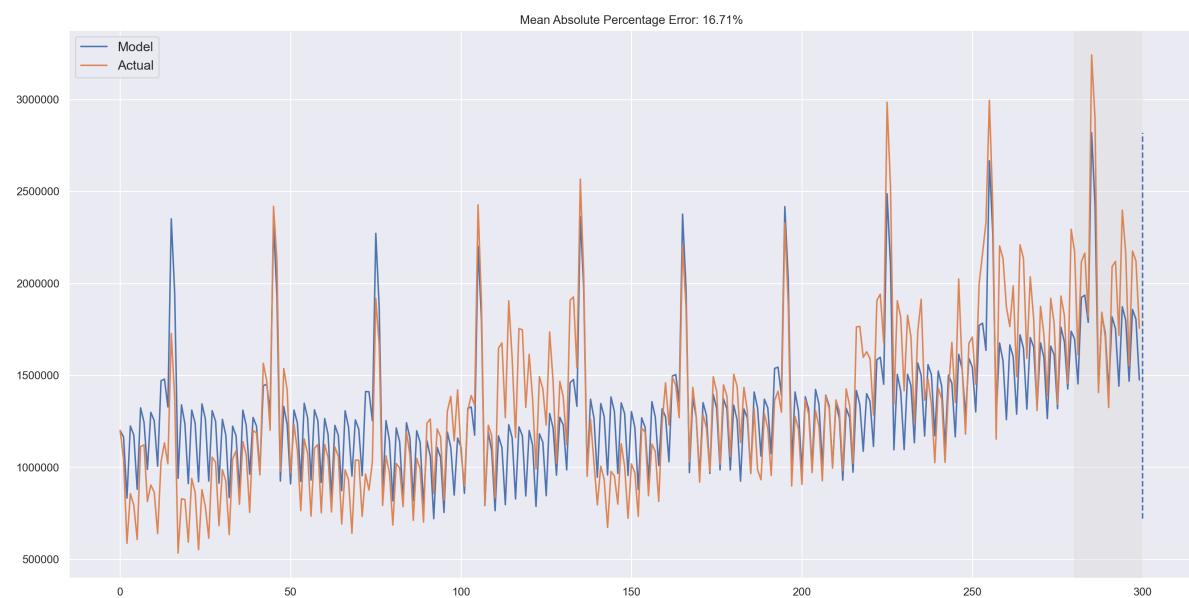
2.2.3.2 手游玩家每天的游戏币消费

设定变化周期30天（一个月）

此为训练后得到的最合适参数组合，也就是可以使得Loss Function最小：

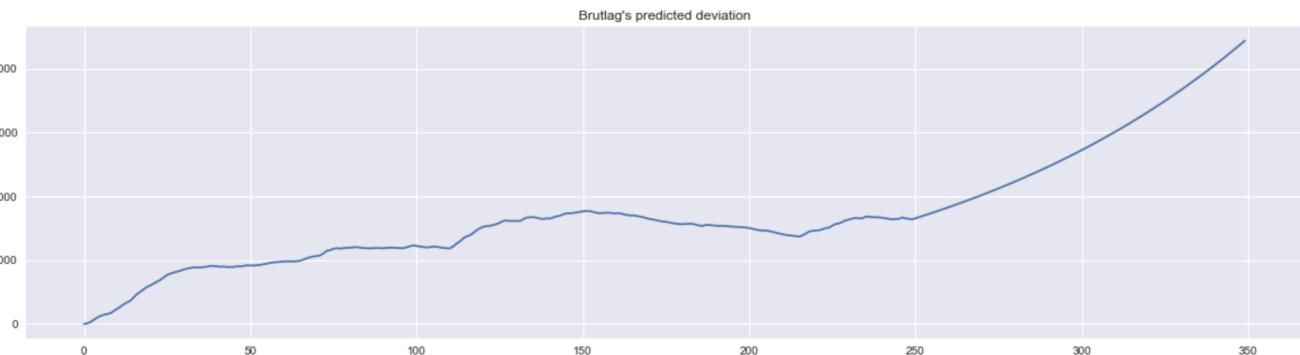
```
alpha_final: 0.013190344846993662 beta_final: 0.047616267647338284  
gamma_final: 0.0  
CPU times: user 1.78 s, sys: 21.6 ms, total: 1.8 s  
Wall time: 1.78 s
```

下图为拟和出来的模型，以及实际值：





此为模型偏差图，可以清楚地看到模型对序列结构的变化反应越趋稳定



2.3 小结

每个数据适合的方法都不同

- Exponential Smoothing: 因为不考虑整体上升或下降趋势，以此只适用于手游玩家每小时观看的广告量
- Double Exponential Smoothing: 因为不考虑整体上升或下降趋势，适用于手游玩家每天的游戏币消费
- Triple Exponential Smoothing(Holt-Winters Model): 手游戏玩家每小时观看的广告量和手游玩家每天的游戏币消费都为季节性时序数据，故都适用，且模型能够成功地逼近初始时间序列，捕捉到季节性、整体降趋势甚至一些异常

3 ARIMA

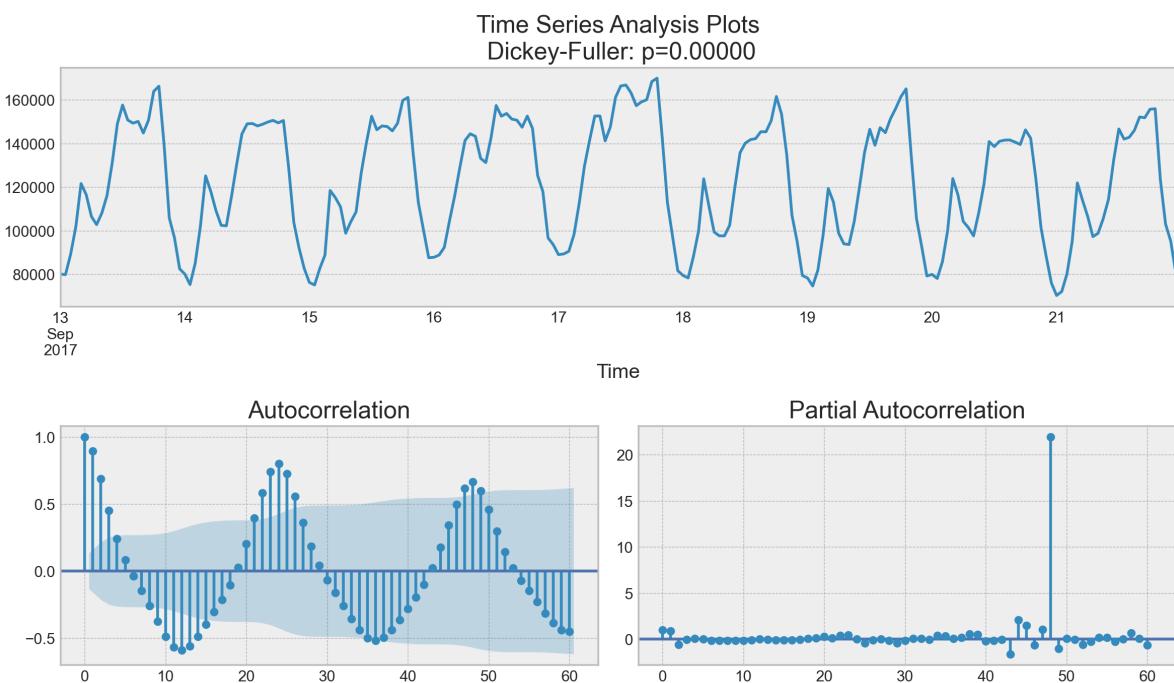
3.1 平稳性检验

以时间序列图、ACF、PACF、Dickey-Fuller来检验

3.1.1 手游戏玩家每小时观看的广告量

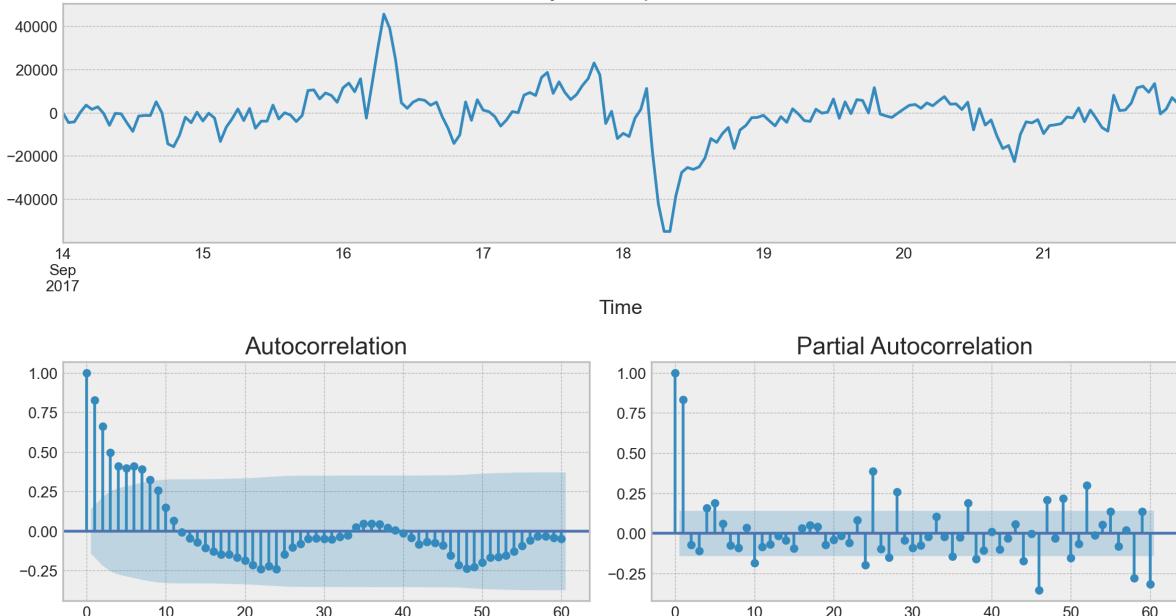
初始序列是平稳的，Dickey-Fuller拒绝了单位根存在的 H_0

而且也能从图形本身就可以看出—没有明显的趋势，故均值是恒定的，整个序列的方差也相对比较稳定 \Rightarrow 在建模之前只需处理季节性，为此让采用“季节差分”，也就是对序列进行简单的减法操作，时差等于季节周期



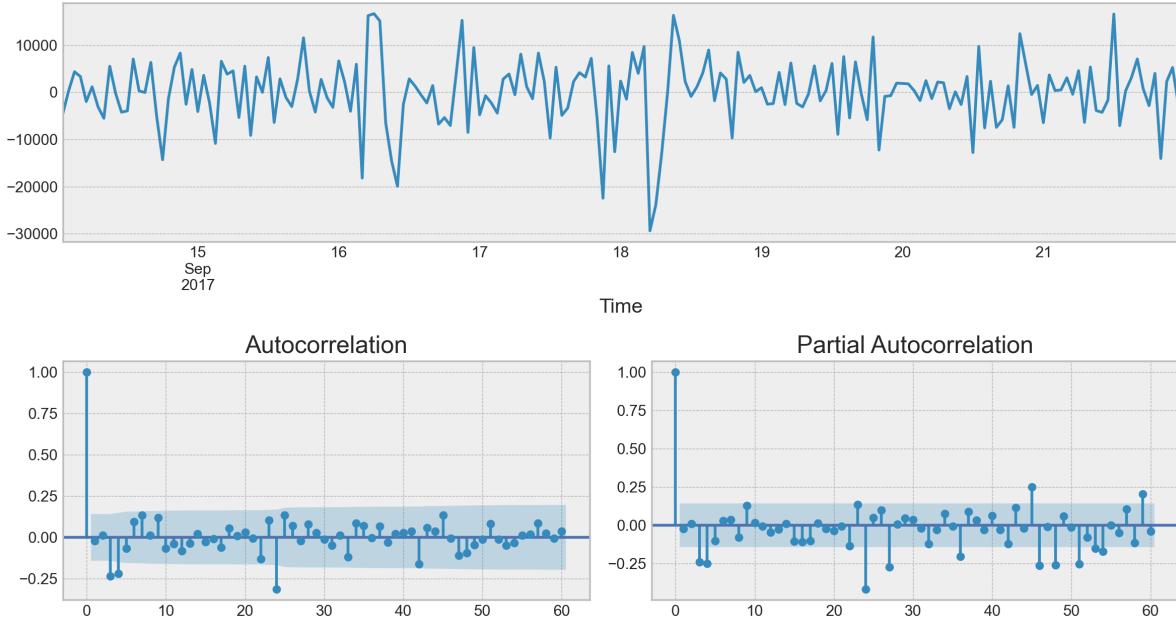
采用“季节差分”后，明显可以看出季节性消失，但是ACF仍然有太多的明显滞后(阴影外的点为滞后点)为了移除它们 \Rightarrow 取一阶差分：从序列中减去自身 (时差为1)

Time Series Analysis Plots Dickey-Fuller: p=0.04742



采用季节性差分和一阶差分后，序列可在零周围振荡，且Dickey-Fuller显示此序列平稳，又ACF中的滞后点变少，ACF的明显尖峰也不见了 ⇒ 可以开始建模

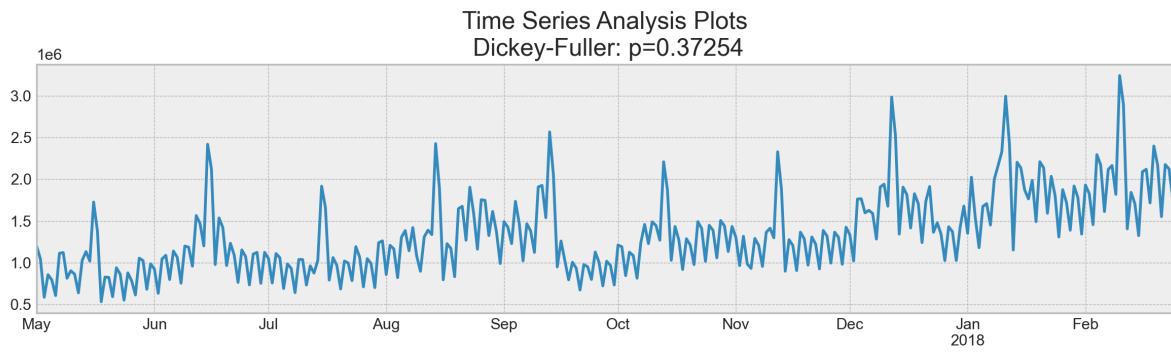
Time Series Analysis Plots Dickey-Fuller: p=0.00000



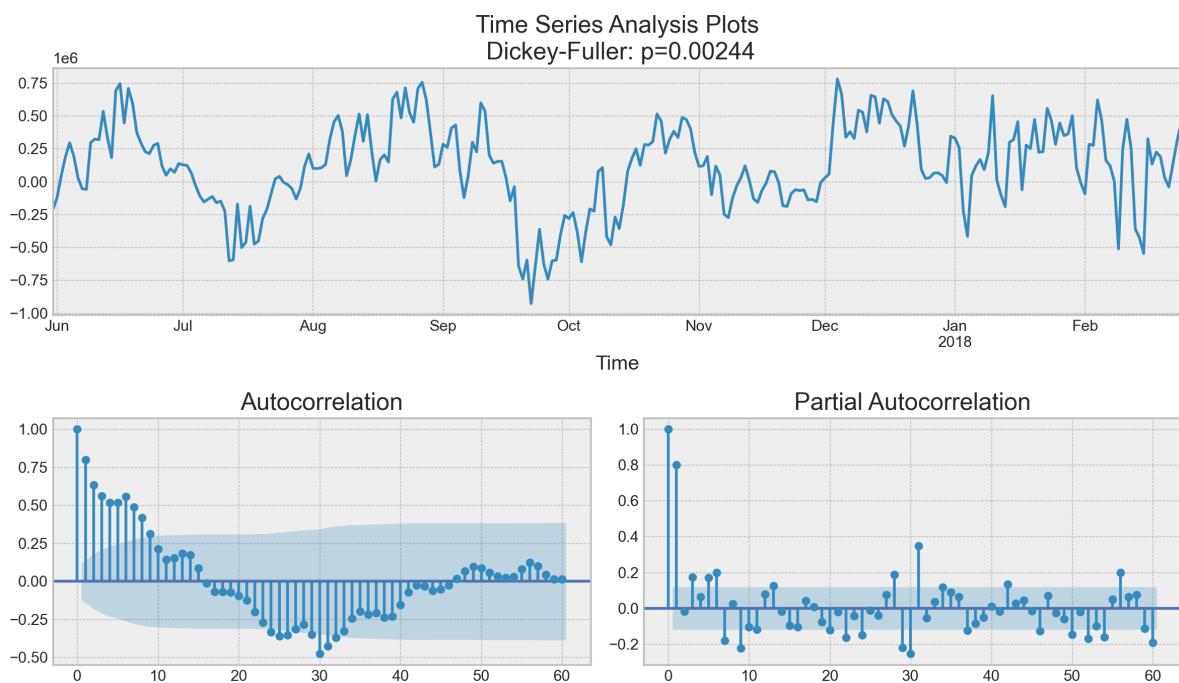
3.1.2 手游戏玩家每天的游戏币消费

初始序列是不平稳的，Dickey-Fuller不拒绝单位根存在的 H_0

而且也能从图形本身就可以看出有明显上升的趋势，故均值是不恒定的，且也有季节性趋势 ⇒ 在建模之前只需处理季节性，为此让采用“季节差分”，也就是对序列进行简单的减法操作，时差等于季节周期

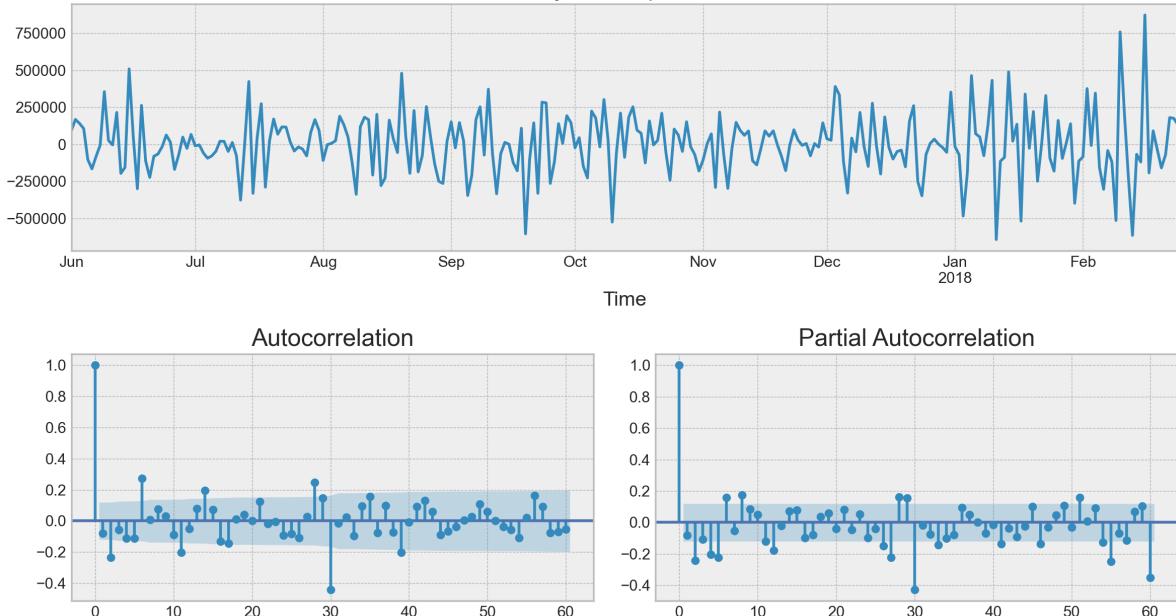


采用“季节差分”后，明显可以看出季节性大都消失，且Dickey-Fuller的p-value值下降许多，但是ACF仍然有太多的明显滞后(阴影外的点为滞后点)为了移除它们 ⇒ 取一阶差分：从序列中减去自身（时差为1）



采用季节性差分和一阶差分后，序列可在零周围振荡，且Dickey-Fuller显示此序列平稳，又ACF中的滞后点变少，ACF的明显尖峰也不见了 ⇒ 可以开始建模

Time Series Analysis Plots Dickey-Fuller: p=0.00006



3.2 Seasonal Autoregression Moving Average model

Seasonal Autoregression Moving Average model(季节自回归移动平均模型):

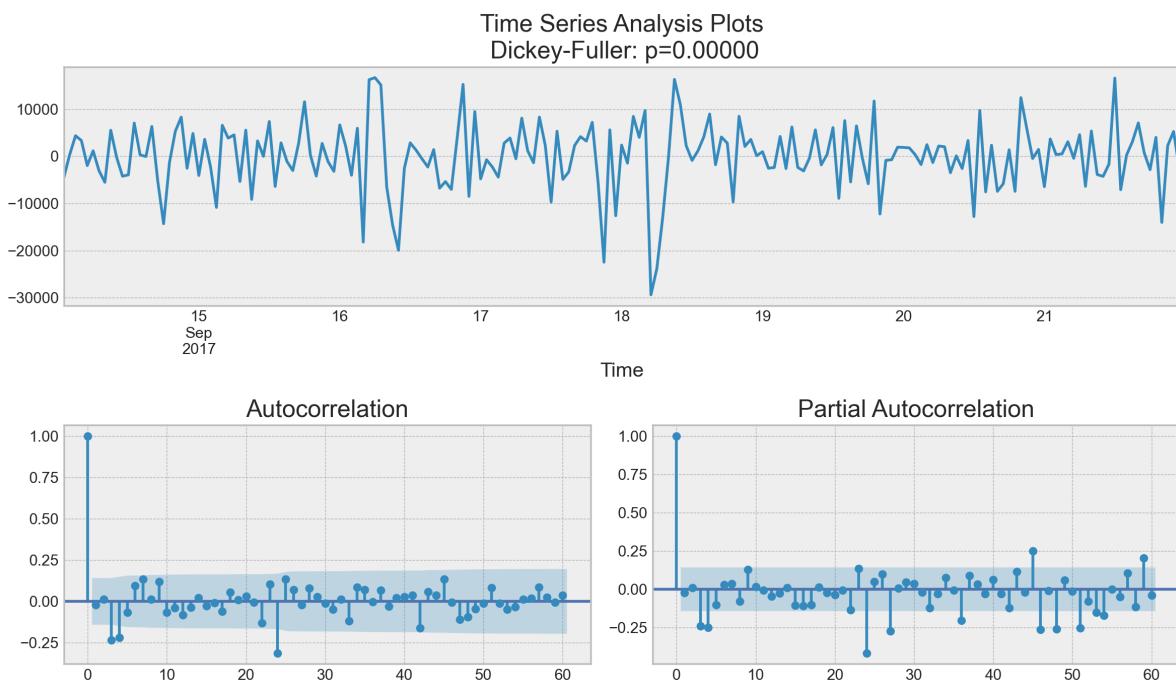
$ARIMA(p, d, q)(P, D, Q, s)$

- $AR(p)$: autoregression model(自回归模型)，即时间序列对自身的回归
基本假设是当前的序列值取决于它之前的值，并且存在一定的滞后 \Rightarrow 模型最大的滞后值称为p，要在PACF中确定初始的p值
- $MA(q)$: moving average model(移动平均值模型)，基于当前的误差依赖于先前的误差，有一定的滞后性 (滞后值记为q)
和上面一样，可以在ACF中找到滞后值q的初始值
 $\Rightarrow AR(p) + MA(q) = ARMA(p, q)$ (自回归移动平均模型)，如果初始序列是平稳的，可以通过这个模型逼近这一序列
- $I(d)$: d 阶，表示使得序列平稳所需的非季节性差别的数量
- $S(s)$: 表示序列是季节性的， $S(s)$ 等于这个季节周期的长度值
- P : 自回归模型的季节性分量的阶数，可以看PACF中显着滞后的数量为季节周期长度的几倍P，P为季节周期长度的倍数
- Q : 移动平均模型中季节分量的阶数，初始值的确定和P同理，不过使用的是ACF
- D : 季节性整合阶数，取值等于1或0，表示是否应用季节差分

3.2.1 手游玩家每小时观看的广告量

由3.1.1的平稳性检定和下图得出：

- p:最有可能为4
从PACF中，可以观测到的最大滞后值
- d:等于1
因为使用的是一阶差分
- q:最有可能为3
从ACF中可以看出
- P:最有可能为1
因为PACF中第24个是比较明显的滞后点
- D:等于1
表示季节差分处理
- Q:最有可能为1
因为ACF中，第24个滞后是比较明显的，第48个滞后不太明显



i. 设置参数搜索区间

```
parameters_list: 54
```

ii. 寻找ARIMA模型的最佳参数组合

iii. 设置 SARIMA 模型最佳参数，查看模型输出结果

- p:
 - 猜测为4
 - 配适为2
- d:等于1
 - 猜测为1
 - 配适为2
- q:
 - 猜测为3
 - 配适为3
- P:
 - 猜测为1
 - 配适为1
- D:
 - 猜测为1
 - 配适为1
- Q:
 - 猜测为1
 - 配适为1

SARIMAX Results

```
=====
=====
Dep. Variable:                      Ads      No. Observat
ions:                            216
Model: SARIMAX(2, 1, 3)x(1, 1, [1], 24)   Log Likeliho
od          -1936.321
Date:                    Fri, 14 May 2021      AIC
3888.642
Time:                           23:12:27      BIC
3914.660
Sample:                         09-13-2017      HQIC
3899.181
                           - 09-21-2017
Covariance Type:                  opg
=====
```

	coef	std err	z	P> z	[0.025
0.975]					
ar.L1	0.7913	0.270	2.928	0.003	0.262
1.321					
ar.L2	-0.5503	0.306	-1.799	0.072	-1.150
0.049					
ma.L1	-0.7316	0.262	-2.793	0.005	-1.245
-0.218					
ma.L2	0.5651	0.282	2.005	0.045	0.013
1.118					
ma.L3	-0.1811	0.092	-1.964	0.049	-0.362
-0.000					
ar.S.L24	0.3312	0.076	4.351	0.000	0.182
0.480					
ma.S.L24	-0.7635	0.104	-7.361	0.000	-0.967
-0.560					
sigma2	4.574e+07	5.61e-09	8.15e+15	0.000	4.57e+07
4.57e+07					

=====

Ljung-Box (Q):	43.70	Jarque-Bera (JB):
10.56		
Prob(Q):	0.32	Prob(JB):
0.01		
Heteroskedasticity (H):	0.65	Skew:
-0.28		
Prob(H) (two-sided):	0.09	Kurtosis:
4.00		

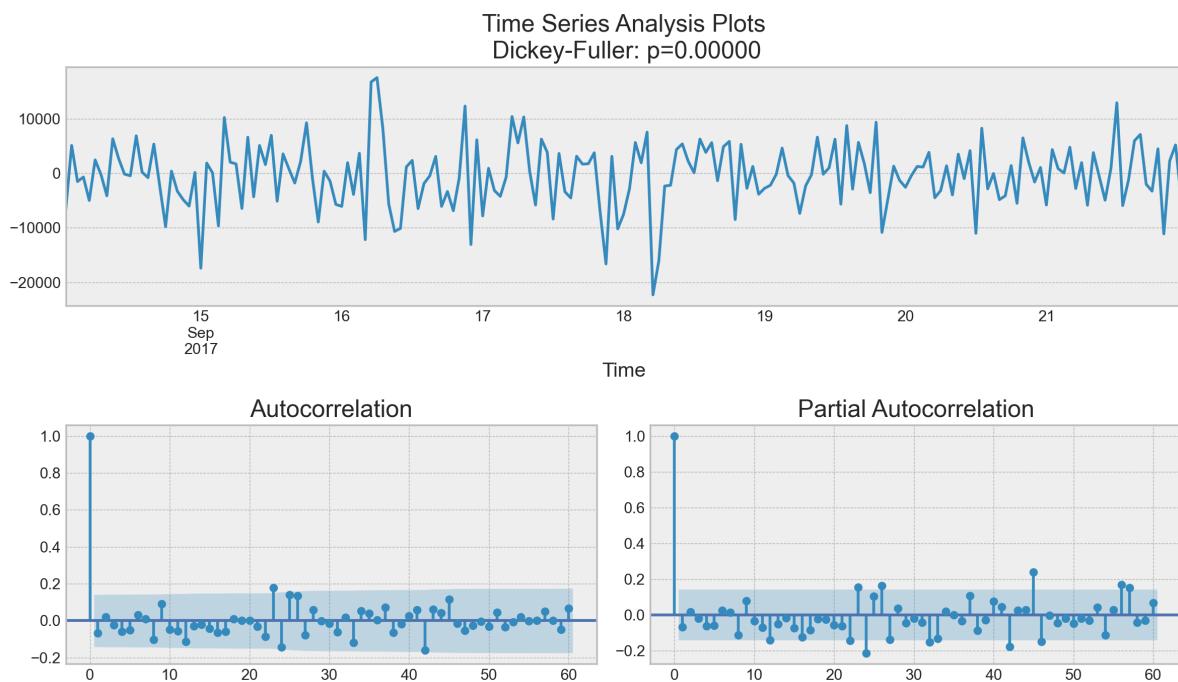
=====

Warnings:

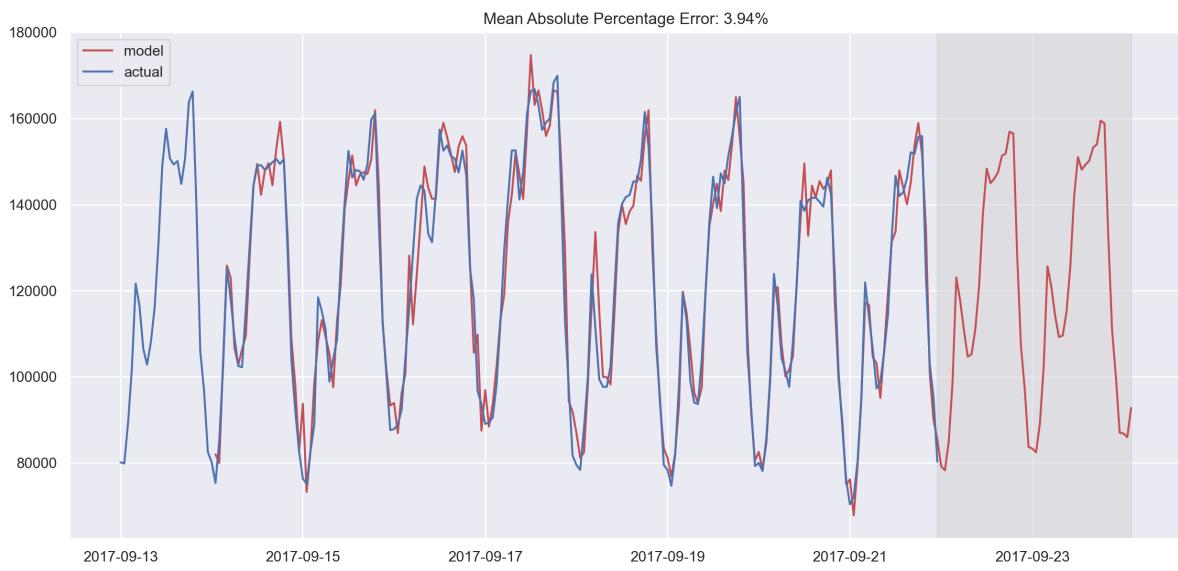
- [1] Covariance matrix calculated using the outer product of gradients (complex-step).
- [2] Covariance matrix is singular or near-singular, with condition number 6.86e+31. Standard errors may be unstable.

iv. 模型的残差分布情况

序列在零周围振荡，且Dickey-Fuller显示此序列平稳，又ACF、PACF中的滞后点几乎消失



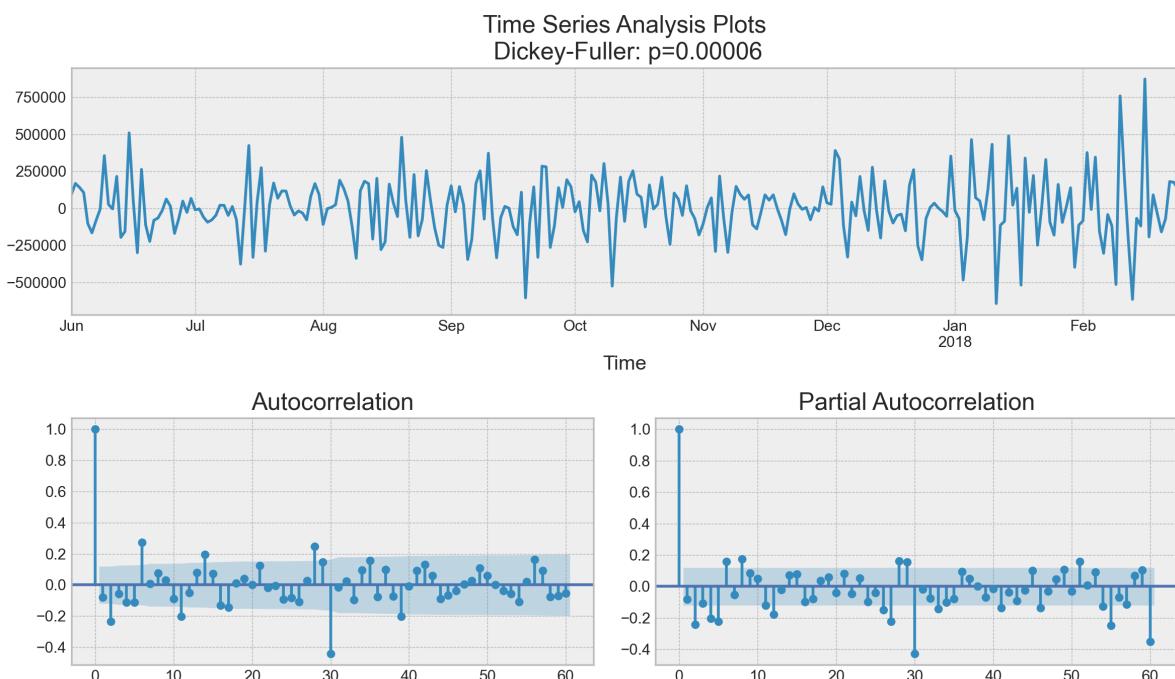
v. 模型预测 Mean Absolute Percentage Error=3.94%



3.2.2 手游玩家每天的游戏币消费

由3.1.2的平稳性检定和下图得出：

- p:最有可能为2
从PACF中，可以观测到的最大滞后值
- d:等于1
因为使用的是一阶差分
- q:最有可能为2
从ACF中可以看出
- P:最有可能为2
因为PACF中第30&60个是比较明显的滞后点
- D:等于1
表示季节差分处理
- Q:最有可能为1
因为ACF中，第30个滞后是比较明显的，第60个滞后不太明显



i. 设置参数搜索区间

```
parameters_list: 72
```

ii. 寻找ARIMA模型的最佳参数组合

iii. 设置 SARIMA 模型最佳参数，查看模型输出结果

- p:
 - 猜测为2
 - 配适为5
- d:等于1
 - 猜测为1
 - 配适为1
- q:
 - 猜测为2
 - 配适为2
- P:
 - 猜测为2
 - 配适为2
- D:
 - 猜测为1
 - 配适为1
- Q:
 - 猜测为1
 - 配适为1

```
SARIMAX Results
=====
Dep. Variable: GEMS_GEMS_SPENT    No. Observati
ons: 300
Model: SARIMAX(5, 1, 2)x(2, 1, [], 30) Log Likelihood
d -3605.888
Date: Sat, 15 May 2021      AIC
7231.777
Time: 00:33:25                BIC
7267.724
Sample: 05-01-2017            HQIC
7246.213
- 02-24-2018
Covariance Type: opg
=====

              coef      std err          z      P>|z|      [ 0.025
0.975 ]
-----
ar.L1      -0.9503      0.176     -5.402      0.000     -1.295
-0.605
ar.L2      -0.8488      0.203     -4.173      0.000     -1.247
-0.450
ar.L3      -0.4173      0.114     -3.651      0.000     -0.641
-0.193
ar.L4      -0.3876      0.111     -3.496      0.000     -0.605
-0.170
ar.L5      -0.3729      0.086     -4.359      0.000     -0.541
```

```

-0.205
ma.L1          0.7965    0.195     4.082      0.000     0.414
1.179
ma.L2          0.4331    0.194     2.238      0.025     0.054
0.812
ar.S.L30       -0.6555   0.073    -8.989      0.000    -0.798
-0.513
ar.S.L60       -0.4285   0.067    -6.382      0.000    -0.560
-0.297
sigma2        3.563e+10  6.4e-12   5.56e+21     0.000   3.56e+10
3.56e+10
=====
=====
```

Ljung-Box (Q): 43.42 Jarque-Bera (JB):

48.57

Prob(Q): 0.33

0.00

Heteroskedasticity (H): 1.57 Skew:

0.54

Prob(H) (two-sided): 0.03 Kurtosis:

4.78

```
=====
=====
```

Warnings:

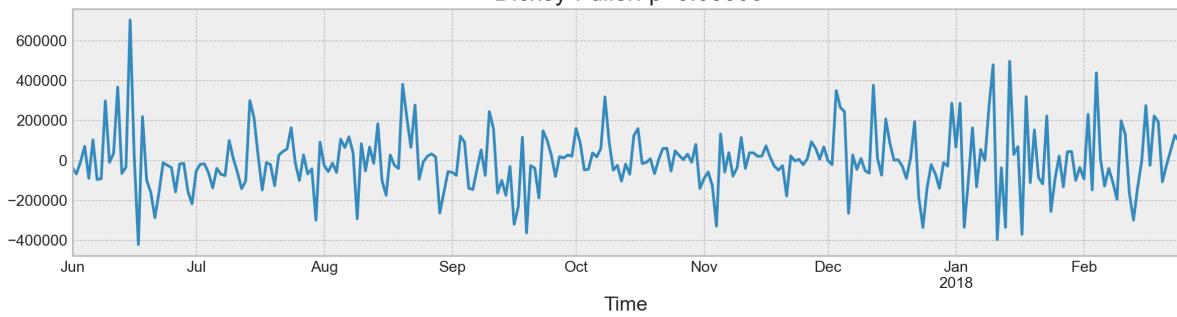
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

[2] Covariance matrix is singular or near-singular, with condition number 7.51e+38. Standard errors may be unstable.

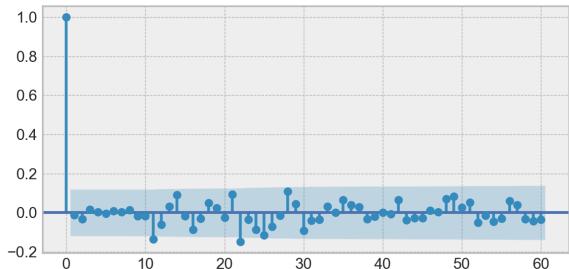
iv. 模型的残差分布情况

序列在零周围振荡，且Dickey-Fuller显示此序列平稳，又ACF、PACF中的滞后点几乎消失

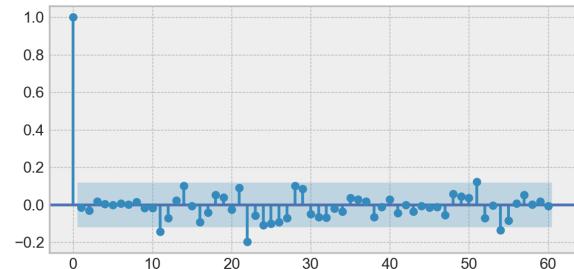
Time Series Analysis Plots
Dickey-Fuller: p=0.00000



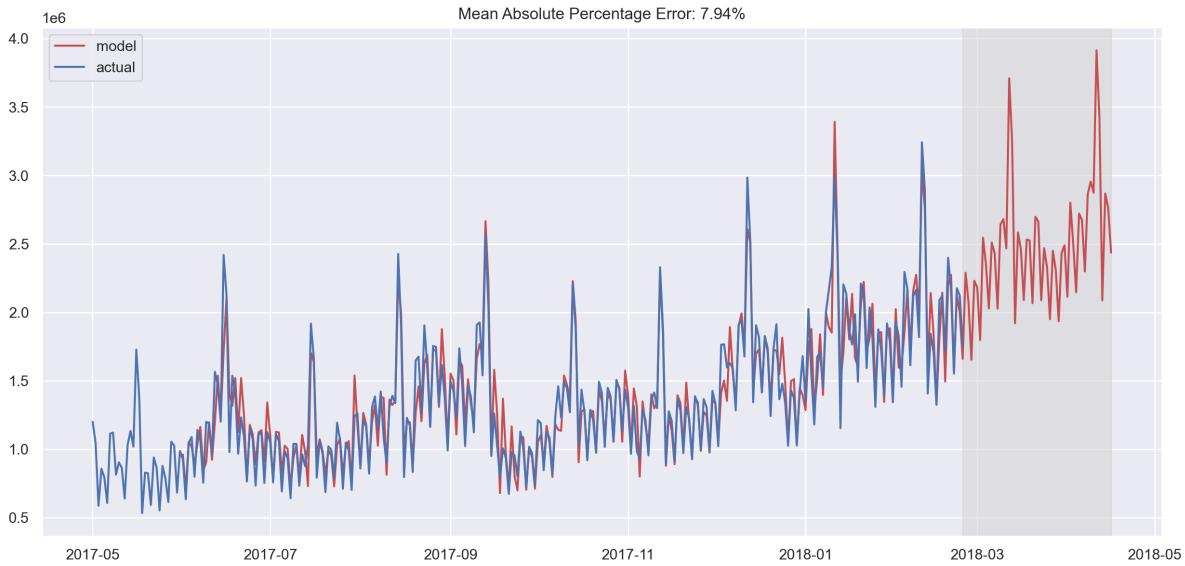
Autocorrelation



Partial Autocorrelation



v. 模型预测 Mean Absolute Percentage Error=7.94%



4. 结论

由于此报告内的两个数据都是季节性时间序列，因此在使用的方法中(Moving Average Method 、 Smoothing Method 、 Exponential Smoothing 、 Double Exponential Smoothing 、 Triple Exponential Smoothing(Holt-Winters Model) 、 Seasonal Autoregression Moving Average model)，考量季节性的Triple Exponential Smoothing(Holt-Winters Model)和Seasonal Autoregression Moving Average model应该最为合适

而这两个方法的Mean Absolute Percentage Error为：

- Triple Exponential Smoothing(Holt-Winters Model)
 - 手游戏玩家每小时观看的广告量: 4.85
 - 手游戏玩家每天的游戏币消费情况: 16.71
- Seasonal Autoregression Moving Average model
 - 手游戏玩家每小时观看的广告量: 3.94
 - 手游戏玩家每天的游戏币消费情况: 7.94

根据Mean Absolute Percentage Error和季节性因素，最适合这两个数据的是Seasonal Autoregression Moving Average model，且会发现比较特别的是，短时间序列(游戏玩家每小时观看的广告量)在两个模型间的Mean Absolute Percentage Error远小于中长时间序列(游戏玩家每天的游戏币消费情况)，可见若非短时间季节性时间序列则最适用Seasonal Autoregression Moving Average model