统计分析与建模期末项目报告

总述

本次期末项目由三个问题组成。问题一关于全国教育动态分析，多因素指向同一个或同一类指标；问题二关于泰坦尼克号存活率，聚焦于”是否存活“的分类问题；问题三关于股票分析，其数据具有明显的时序特征。

每个问题所处的实际语境不同、数据特点不同，不可一概而论，应当使用不同的模型加以分析。

具体分析

回归模型——教育动态分析问题

项目要求

该数据集提供了一个全球范围内的教育数据，包括29列，内容涵盖失学率、学业完成率、熟练程度、识字率、出生率以及小学和高等教育入学统计等信息。希望你能够通过数据分析对不同国家和地区的教育动态提供深刻的见解，希望你的工作能够给全球教育工作者一个评估、加强和重塑全球教育系统的机会。

确立模型

使用线性回归模型

解决过程

一、数据预处理

1.1 打开数据集

```
# 设置你的文件路径
file_path <-"C:/Users/Chen Yutong/Desktop/education.csv"

# 读入CSV文件
data <- read.csv(file_path)
```

1.2 数据预处理

1.删除缺失值

2.数据去重

3.数据过滤，保证各种rate在0~100之间

4.处理异常值（很多0）

将一行中0的个数大于15个的整行删掉

将0替换成每一列的中位数

二、建立模型

a. 建立高等教育总入学人数预测模型（Y：Gross_Tertiary_Education_Enrollment）

a.2 建模前的分析

a.2.1 初探基本关系

在进行线性回归之前，首先需要对数据进行查看基本关系，然后进行检验数据是否满足参与线性回归分析的基本条件。 基本关系包括数据的相关关系以及共线性的查看。

a.2.2 相关关系

在回归分析前一般需要做相关分析，因为有了相关关系，才可能有回归影响关系；如果没有相关关系，是不应该有回归影响关系的。 将每个变量与高等教育总入学人数进行两两相关分析，p值大于0.05，则说明没有相关关系，在进行回归分析时不放入！

因此，由皮尔逊（Pearson）相关分析得Gross_Primary_Education_Enrollment、Unemployment_Rate的p值大于0.05，说明没有相关关系，在进行回归分析时不放入。

a.2.3 共线性

共线性是指线性回归模型中的解释变量之间由于存在精确相关关系或高度相关关系（例如相关系数大于0.5）而使模型估计失真或难以估计准确。共线性的存在可能会降低估计的精准度，并且稳定性也会降低。无法判断单独变量的影响。回归方程的标准误差增大。变量显著性可能会失去意义等等。所以在分析前需要对共线性问题进行检查。一般VIF值大于10（严格来说大于5），存在共线性问题。如果存在共线性问题则不能使用线性回归，可以使用岭回归、Lasso回归等进行分析。

VIF值大于10的变量说明存在严重的多重共线性问题。 由于这些变量之间的共线性，可以考虑进行合并。可以看出大多数共线性是由于性别导致，因此采取取平均合并相关的性别变量。

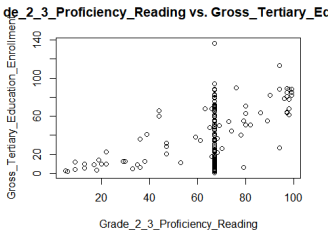
再做一次共线性分析，可以发现Completion_Rate_Lower_Secondary的vif>10，因此将这个变量删掉，再测试一下后发现没有共线性的情况了。

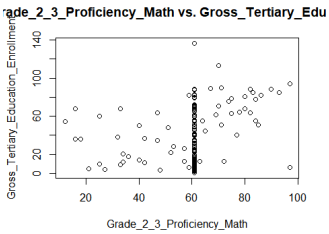
a.2.4 是否满足线性回归的几个前提条件

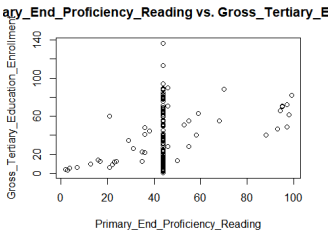
①因变量为连续性变量（√）

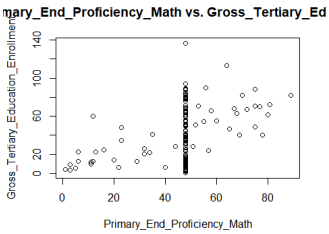
②因变量与自变量之间存在线性关系

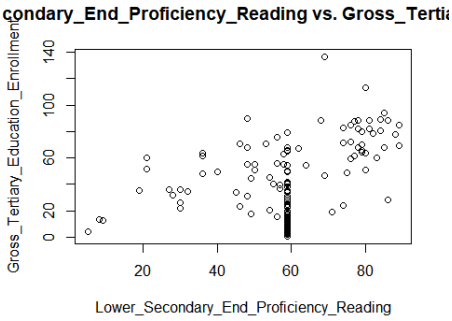
一般检验数据之间的线性关系，是为了考察因变量随自变量值变化的情况，可以做相关分析从侧面进行说明或者利用散点图进行说明，散点图更加直观，所以本次选择散点图进行描述。

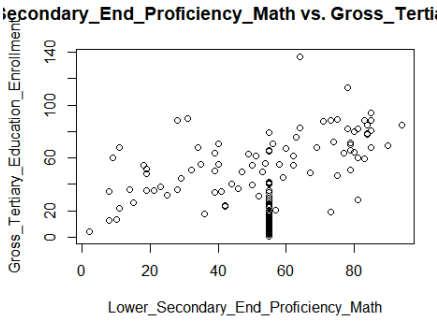


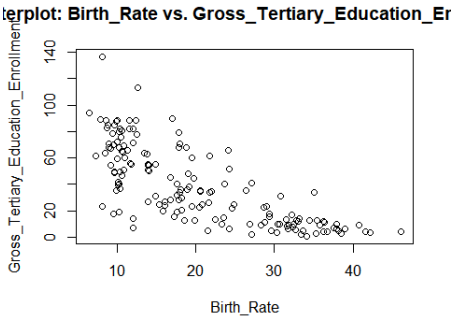


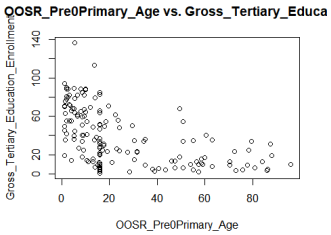


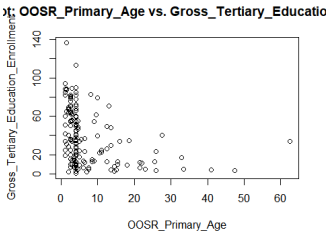


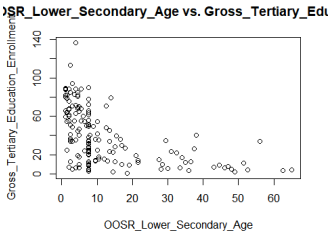


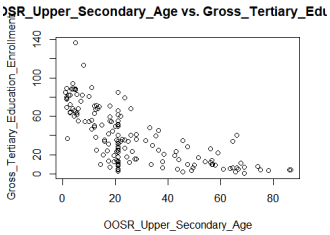


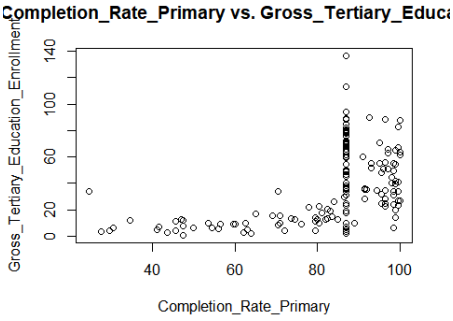


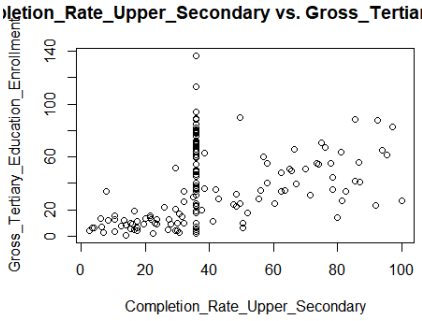


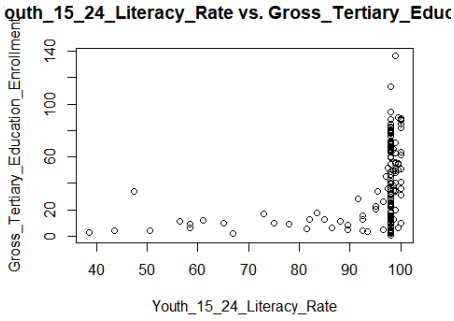












观察图后，我们认为存在较明显的线性关系。

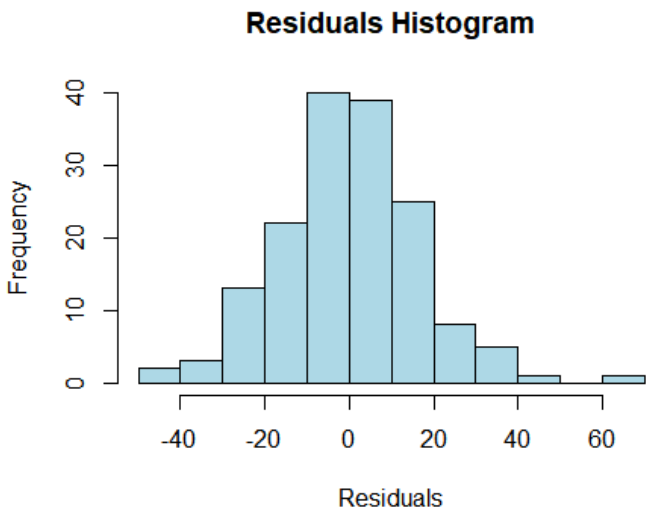
③样本个体间相互独立（由Durbin-Waston检验判断）

是指残差是独立的。可以查看DW值，一般在DW值在2附近（比如1.7-2.3之间），则说明没有自相关性，模型构建良好，反之若DW值明显偏离2，则说明具有自相关性，模型构建较差。

计算后得dw=2.193369（在1.7-2.3之间）说明没有自相关性，模型构建良好，残差相互独立。

④正态性：给定各个X值后，相应的Y值服从正态分布

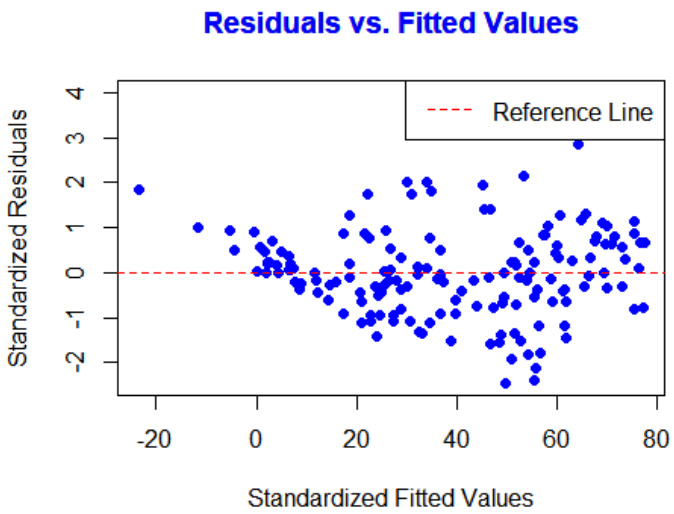
表示残差服从正态分布。其方差σ2 = var (ei)反映了回归模型的精度，一般 σ 越小，用所得到回归模型预测y的精确度越高。建立回归分析模型得到残差与预测值，利用残差绘制直方图查看残差是否满足正态分布，结果如下：



直方图呈现’中间高，两边低，左右基本对称的”钟形图”则基本服从正态分析。上图可以看出，数据呈现的分布对称，且近似’钟形’曲线，所以残差满足正态分布。，接下来验证方差齐性。

⑤方差齐性：各X值变动时，相应的Y有相同的变异度

方差齐性是指残差的大小不随所有变量取值水平的改变而改变，即方差齐性。首先对残差和预测值进行标准化，标准化残差为Y轴，标准化预测值为X轴绘制散点图，如果所有点均匀分布在直线Y=0的两侧，则可以认为是方差齐性，结果如下：



从散点图可以发现数据大致均匀分布在Y=0的两侧，所以可认为是方差齐性。

综上，数据满足回归分析的前提假设。可以进行线性回归。

a.3、数据建模及模型质量评估

分训练集和测试集，在训练集上进行建模

在该模型下， 残差标准差（均方误差RMSE）为17.82 拟合优度的判定系数（Multiple R-squared ）为0.6678，表示模型解释了目标变量方差的66.78%。值相对来说挺接近1的。 调整后的Adjusted R-squared 为0.6263 。调整的R-squared考虑了模型中使用的变量数量，防止了过多自变量引起的过拟

合。比Multiple R-squared更稳健。 最后观察到F-statistic (F统计量)的p-value: < 2.2e-16：表明模型作为整体是显著的，即至少有一个自变量对目标变量有显著影响。

a.4、模型优化

在逐步回归中，模型通过逐步添加或删除变量，以找到一个最优的拟合模型。AIC（Akaike Information Criterion）是一种常用的评价模型拟合优度和复杂度平衡的指标。AIC的目标是寻找一个对数据拟合较好但参数较少的模型。因此，可以通过比较AIC值来评估不同模型的相对优劣，选择AIC值最小的模型。

step()逐步回归后，得到AIC最小的（734.44），最优化的模型为：

```
mlr_better <- lm(formula = Gross_Tertiary_Education_Enrollment ~ Grade_2_3_Proficiency_Reading +
    Lower_Secondary_End_Proficiency_Reading + Birth_Rate + OOSR_Pre0Primary_Age +
    OOSR_Lower_Secondary_Age + OOSR_Upper_Secondary_Age, data = train_data)
```

根据该模型结果我们可以发现，观察p值，大部分回归系数在统计上是显著的。 在该模型下， 残差标准差（均方误差RMSE）为17.54 < 原模型17.82 调整后的Adjusted R-squared 为0.6379 > 原模型0.6263 。具有更好的解释性。 最后观察到F-statistic (F统计量)的p-value: < 2.2e-16：表明模型作为整体是显著的。

a.5、模型评估

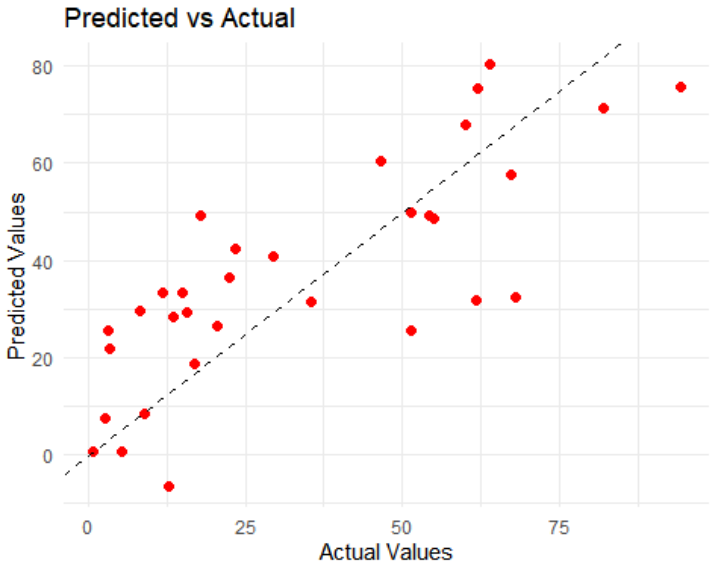
```
# 模型评估
predictions_tertiary <- predict(mlr_better, newdata = test_data)
rmse_tertiary <- sqrt(mean((predictions_tertiary - test_data$Gross_Tertiary_Education_Enrollment)^2))


rmse_tertiary
## [1] 16.6066
```

评估效果的解释：

rmse_tertiary: 是模型在测试集上的均方根误差。RMSE越小，表示模型在测试数据上的预测性能越好。

我们可以看到RMSE为16.6066，比较小，因此说明在测试数据上的表现较好

另外，我们可以可视化在测试集上的预测值与实际值的关系，画出对角线 由图我们可以看出大部分点还是集中在对角线附近，说明拟合效果较好。



a.6、模型解读：影响因素分析

在多个影响因素中，**初中教育结束时的阅读能力、出生率（反比）和高中的辍学率（反比）的入学总额**。 这些与高等教育的入学总额的线性相关有显著性。

因此根据以上数据分析工作得到的教育改进启示和建议是：

1.提高初中阅读水平： 通过制定和实施更有效的教育方案和教学方法，致力于提高初中生的阅读水平。可能的方法包括引入更富有趣和引人入胜的教材、提供额外的辅导和支持，以及使用现代技术手段提升学习体验。

2.增加更多高中教育资源： 根据出生率与高等教育的入学总额成反比，说明小孩越多，就会有越多的小孩入接受不了高等教育，这侧面反应了高等教育资源的不充分。因此着重解决可能影响高等教育的入学总额的社会经济问题，这样就可以让更多学生有机会接受高等教育。

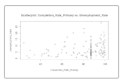
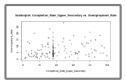
3.减少高中辍学率： 建立积极学校文化： 学校环境应该是积极、支持和鼓励学习的。建立积极的学校文化可以帮助学生感到更受欢迎和受到尊重，从而降低辍学的可能性。 个性化学习： 每个学生都有自己的学习风格和需求。通过个性化的学习计划，可以更好地满足学生的需求，提高他们的学术成绩和对学校的参与度。 提供支持服务： 学校可以提供学生支持服务，包括心理健康支持、学术指导和职业规划。有时候，学生辍学可能是因为他们面临的问题而不是学业本身。

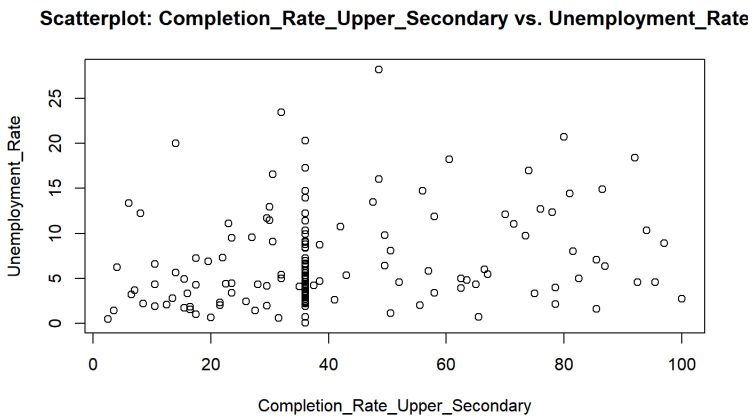
b. 建立失业率预测模型（Y：Unemployment_Rate）

步骤与a模型相同

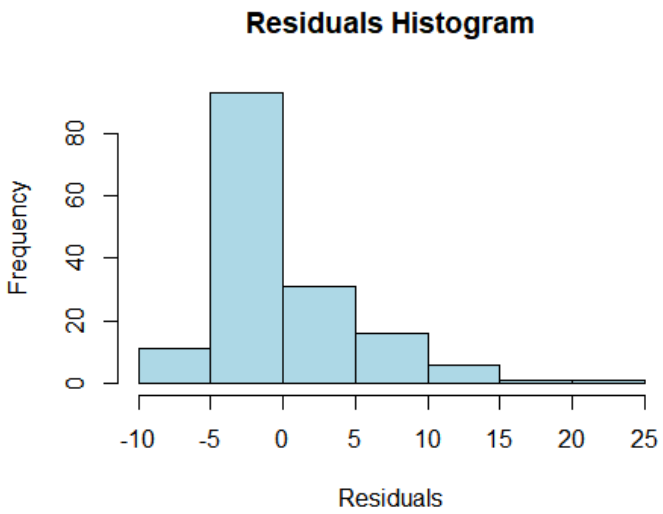
差异在于：

1.线性相关性不明显

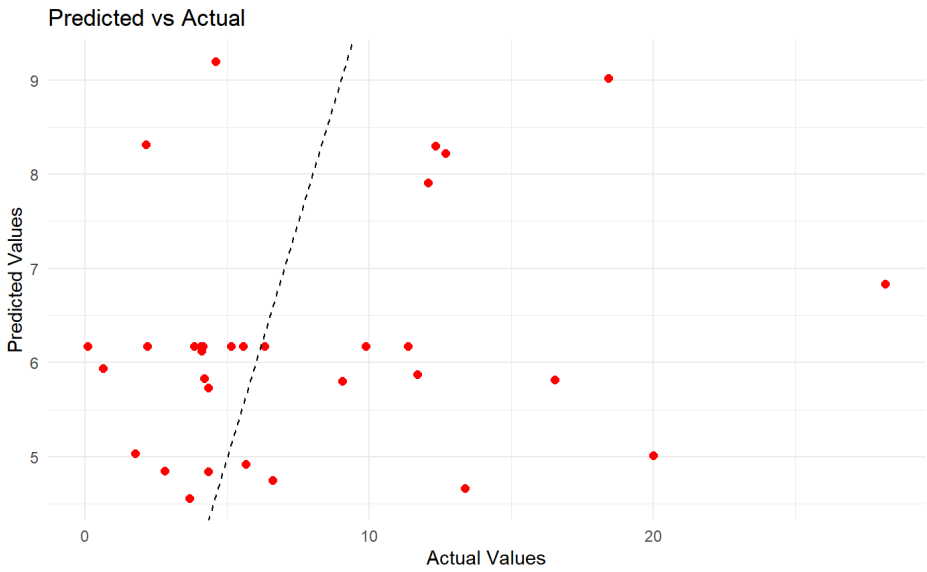


2.残差不满足正态分布



3.拟合效果不好

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.041092   2.230191   1.812   0.0724 .
Completion_Rate_Primary         0.003898   0.035031   0.111   0.9116
Completion_Rate_Upper_Secondary 0.049876   0.027773   1.796   0.0749 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.494 on 124 degrees of freedom
Multiple R-squared:  0.05939,   Adjusted R-squared:  0.04422
F-statistic: 3.915 on 2 and 124 DF,  p-value: 0.02246
```



结论

从模型结果来说，对于各因素和失业率之间的线性关系其实并不明显，难以建立起拟合效果好的线性回归模型。

如果非要从这个拟合一般的模型出发，那么初高中的完成率越高，失业率越高，显然是不符合常理的。

因此，对于在此数据集下的失业率模型，我们无法很好的解释其他因素与他的线性关系，无法得出准确有效的教育建议。

分类模型——泰坦尼克号存活率分析问题

项目要求

泰坦尼克号的沉没是历史上最著名的沉船事故之一。1912年4月15日，在处女航中，被认为“永不沉没”的皇家邮轮泰坦尼克号与冰山相撞后沉没。不幸的是，船上没有足够的救生艇容纳所有人，导致2224名乘客和船员中的1502人死亡。虽然生存有一些运气因素，但似乎有些群体比其他群体更有可能生存下来。请使用乘客数据（包括姓名、年龄、性别、社会经济阶层等）建立一个预测模型，尝试思考和回答以下问题：“什么样的人更有可能存活下来？”。

确立模型

对于这个问题，由于需要预测是否存活（Survived），是一个典型的二元分类问题。因此，可以使用各种二元分类模型来进行分析。在这里根据课程所学知识，可以使用**广义线性**回归中的逻辑回归模型。

解决过程

一、原始数据分析

在对数据进行预处理前，可以先对原数据进行一系列分析，包括数据概览和缺失值检查。

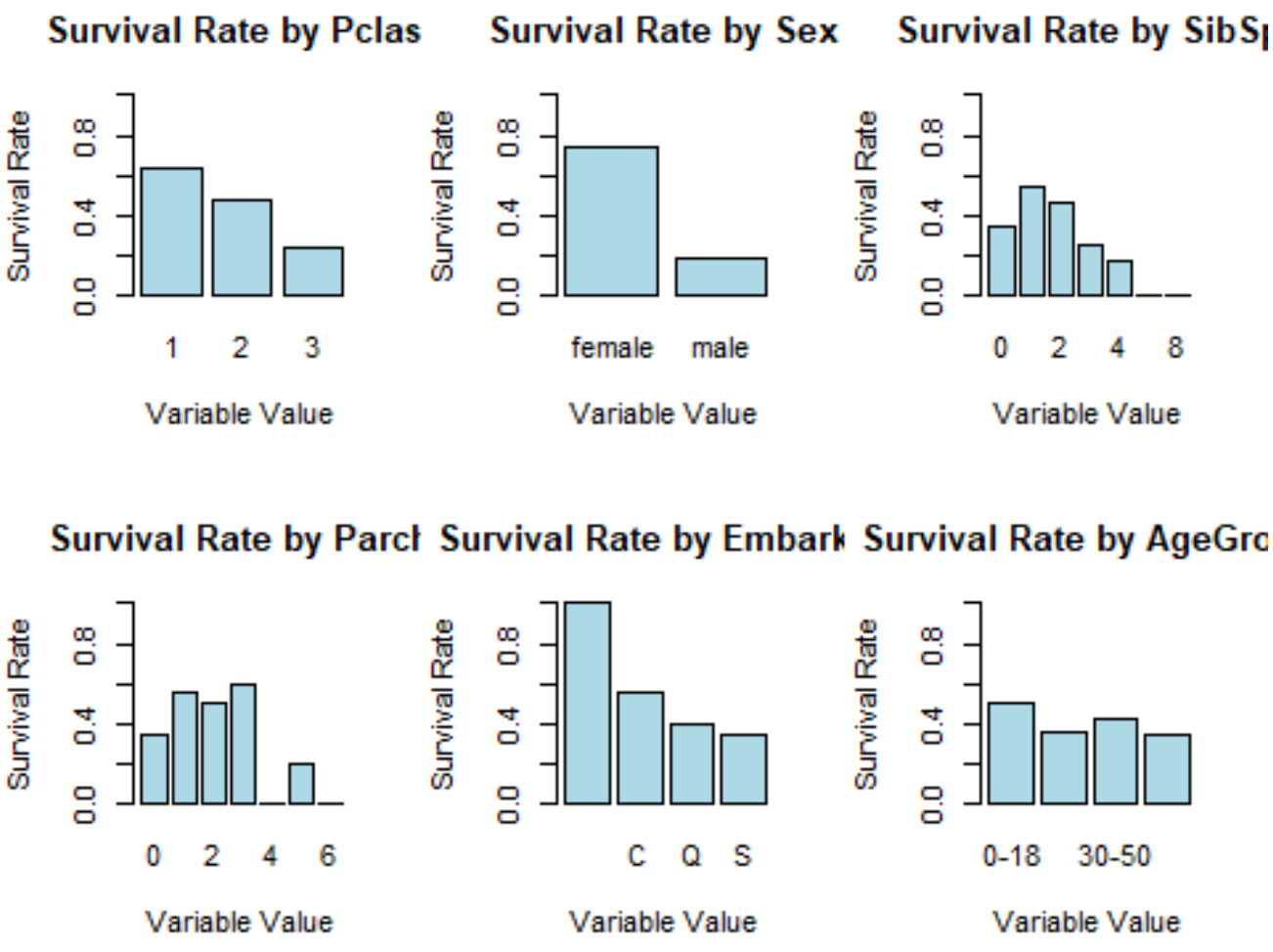
```r
# 将数据导入数据框
titanic_data <- read.csv("titanic.csv", stringsAsFactors = FALSE)
# 查看数据框的结构
str(titanic_data)
# 获取数据框的摘要统计信息
summary(titanic_data)
# 检查每列的缺失值，包括字符型变量
sapply(titanic_data, function(x) sum(is.na(x) | x == ""))
```

重点关注缺失值信息：

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         177
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0         687           2
```

可以看到年龄数据的缺失值较多。

另外可以计算整体生存率和不同子集的生存率作比较参考，并绘制如下的柱状图进行初步分析：



通过数据可视化直观感受：存活率可能与性别、船舱等级、同行人数有较大关系。

二、数据预处理

1.缺失值处理

根据原始数据分析的结果可知，Age属性有177个缺失值，Embarked属性有2个缺失值，Cabin属性有687个缺失值。对于这三种缺失值，分别采用不同的处理方式。

- 修补Embarked数据

对于Embarked这种缺失值极少的，适合用该属性的取值最多的值来填补缺失值。

- 修补Cabin数据

对于Cabin这种缺失值占比很大的，并且不为数字类型的，可以将缺失值作为新的一类填补进去。由于Cabin是表示客舱编号，客舱的位置很大可能影响人的生存情况，根据数据实际的情况来看，Cabin为空可能代表没有客舱可坐，所以将空值单独作为新值是有必要且有实际意义的。

- 修补Age数据

对于Age这种缺失值占比较大的，且为数字类型的，可以从特殊角度考虑数据填补的方法。注意到在Name属性一栏中有乘客身份相关的信息，例如Mr、Miss、Mrs等，考虑到乘客身份很有可能跟年龄有关系，所以先统计一下各种称呼的人数，统计的结果如下：

```
##
##          Capt
##             1
##           Col
##             2
##           Don
##             1
##            Dr
##             7
##      Jonkheer
##             1
##          Lady
##             1
##         Major
##             2
##        Master
##            40
##          Miss
##           182
##          Mlle
##             2
##           Mme
##             1
##            Mr
```

```
##                                                          517
##                                                          Mrs
##                                                          125
##                                                           Ms
##                                                            1
##                                                          Rev
##                                                            6
## Rothes, the Countess. of (Lucy Noel Martha Dyer-Edwards)
##                                                            1
##                                                          Sir
##                                                            1
```

可以发现最主要的身份包括Mr、Miss、Mrs、Master，其他身份都是少数，所以将乘客身份信息分为五类，每种身份分别计算年龄的平均值，同时也可以根据这五类扩充数据的身份信息作为新的属性。

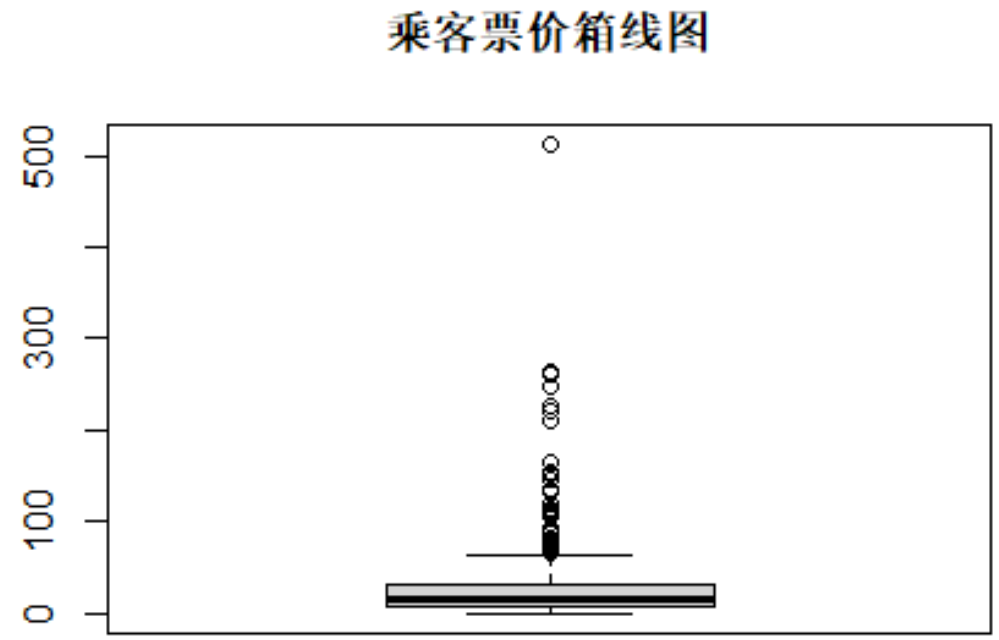
由于要计算平均值，所以先查找年龄数据是否有异常值，先做异常值处理。年龄的异常值需要根据实际情况来考虑，而不是根据数据分布情况，只需要考虑数据是否在正常年龄范围内（0~120岁）即可。根据以下代码：

```r
# 查找年龄中的异常值
age_outliers <- titanic_data$Age < 0 | titanic_data$Age > 120

# 输出异常值
print(titanic_data[age_outliers, c("PassengerId", "Age")])
```

可见没有异常值。 接下来分别计算不同身份的年龄平均数后四舍五入填入空缺的身份信息类即可。

2.异常值处理

年龄异常值前面已经寻找过，另一个可能出现异常值的属性为票价，根据原始数据分析中的票价箱线图中可以看出，有值明显偏离数据集中区域，下面进行分析：




```r
# 找出票价超过500的行
high_fare <- which(titanic_data$Fare > 500)

# 输出找到的行
print(titanic_data[high_fare, ])
```

```
##     PassengerId Survived Pclass                                Name    Sex
Age
## 259         259        1      1                    Ward, Miss. Anna female
35
## 680         680        1      1 Cardeza, Mr. Thomas Drake Martinez   male
36
## 738         738        1      1             Lesurer, Mr. Gustave J   male
35
##     SibSp Parch   Ticket     Fare       Cabin Embarked AgeGroup Title
## 259     0     0 PC 17755 512.3292          NC        C    31-50  Miss
## 680     0     1 PC 17755 512.3292 B51 B53 B55        C    31-50    Mr
## 738     0     0 PC 17755 512.3292        B101        C    31-50    Mr
##     TitleCategory
## 259          Miss
## 680            Mr
## 738            Mr
```

发现票价超过500的票价都为512.3292，且Pclass都为1，经过实际分析说明可能就是正常的高票价，为异常值的可能性很小，故不做处理。

3.数据去重

```r
# 检测是否有重复行
has_duplicates <- any(duplicated(titanic_data))

# 输出结果
print(has_duplicates)
```

```
## [1] FALSE
```

输出结果为FALSE，说明没有重复数据。

4.数据数值化及其他处理

由于数据分析通常需要对数值进行研究，所以有必要将一些字符型数据转变成数值型数据。由于Ticket数据类别过多，且实际情况中与存活率关系不大，可以考虑不处理这个属性。 另外由于需要编码的数据为没有明显顺序的类别，所以放弃标签编码，而使用独热编码。 分别对Embarked、TitleCategory、Cabin进行独热编码。

三、建立模型

划分训练集和测试集，按照4:1的比例划分，训练集有713份数据，测试集有178份数据。

在建立逻辑回归模型之前，要关注部分属性之间是否存在共线性问题，如果有则会导致p值输出异常。经分析可知属性 Sex、TitleCategoryMiss、TitleCategoryMrs、TitleCategoryMr有共线性问题，所以删除后三个后续扩充的属性。删除后共线性问题解决。

初次建模，使用R语言中的glm函数来进行逻辑回归预测，首先将所有的属性都填入分析：

```r
glm_model1 <- glm(Survived ~ Age + Sex + Pclass + SibSp + Parch + Fare +
Cabin_Code + EmbarkedC + EmbarkedQ + TitleCategoryMaster , family = binomial,
data = train_data)
summary(glm_model1)
predictions1 <- predict(glm_model1, newdata = test_data, type = "response")
predicted_classes <- ifelse(predictions1 > 0.5, 1, 0)
```

```r
accuracy <- sum(predicted_classes == test_data$Survived) /
length(test_data$Survived)
print(paste("模型1准确率:", round(accuracy, 4)))
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + SibSp + Parch +
##     Fare + Cabin_Code + EmbarkedC + EmbarkedQ + TitleCategoryMaster,
##     family = binomial, data = train_data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.972070   0.736053   6.755 1.43e-11 ***
## Age                  -0.017335   0.009572  -1.811 0.070127 .
## Sexmale              -3.286457   0.261139 -12.585  < 2e-16 ***
## Pclass               -1.086515   0.198984  -5.460 4.75e-08 ***
## SibSp                -0.479176   0.137688  -3.480 0.000501 ***
## Parch                -0.350050   0.154648  -2.264 0.023603 *
## Fare                 -0.001454   0.003526  -0.412 0.680016
## Cabin_Code           -0.002241   0.003843  -0.583 0.559863
## EmbarkedC             0.486603   0.271580   1.792 0.073174 .
## EmbarkedQ             0.005849   0.375150   0.016 0.987560
## TitleCategoryMaster   3.525086   0.604136   5.835 5.38e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 947.01  on 712  degrees of freedom
## Residual deviance: 588.09  on 702  degrees of freedom
## AIC: 610.09
##
## Number of Fisher Scoring iterations: 5
## [1] "模型1准确率: 0.8202"
```

可以看到p值较大属性包括 Fare、Cabin_Code、EmbarkedQ ，下一步建模时删除这些属性进行模型优化:

```r
glm_model2 <- glm(Survived ~ Age + Sex + Pclass + SibSp + Parch +
TitleCategoryMaster + EmbarkedC, family = binomial, data = train_data)
summary(glm_model2)
predictions2 <- predict(glm_model2, newdata = test_data, type = "response")
predicted_classes <- ifelse(predictions2 > 0.5, 1, 0)
```

```r
accuracy <- sum(predicted_classes == test_data$Survived) /
length(test_data$Survived)
print(paste("模型2准确率:", round(accuracy, 4)))
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + SibSp + Parch +
##     TitleCategoryMaster + EmbarkedC, family = binomial, data = train_data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.667116   0.569938   8.189 2.64e-16 ***
## Age                  -0.016704   0.009517  -1.755  0.07924 .
## Sexmale              -3.281486   0.256272 -12.805  < 2e-16 ***
## Pclass               -1.106263   0.143654  -7.701 1.35e-14 ***
## SibSp                -0.488048   0.135618  -3.599  0.00032 ***
## Parch                -0.359184   0.150161  -2.392  0.01676 *
## TitleCategoryMaster   3.545779   0.602113   5.889 3.89e-09 ***
## EmbarkedC             0.488537   0.262778   1.859  0.06301 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 947.01  on 712  degrees of freedom
## Residual deviance: 588.54  on 705  degrees of freedom
## AIC: 604.54
##
## Number of Fisher Scoring iterations: 5
## [1] "模型2准确率: 0.8315"
```

可以看到AIC值从610.09减少到604.54，明显减少，准确率也有所提升，说明模型得到了优化。

四、模型解读及结论

要解读模型，首先计算各变量的优势比：

```r
coefficients <- coef(glm_model2)
odds_ratios <- exp(coefficients)
print("各变量的优势比:")
print(odds_ratios)
```

```
## [1] "各变量的优势比:
##         (Intercept)                 Age             Sexmale
Pclass
##        106.39050413          0.98343498          0.03757237
0.33079277
##               SibSp               Parch TitleCategoryMaster
EmbarkedC
##          0.61382313          0.69824598         34.66666628
1.62993041
```

解读如下：

Age (年龄): 当年龄增加一个单位时，存活的优势相对于不存活的优势减少约1.65%。

Sex (性别): 女性相对于男性，存活的优势比约为0.0376，表示女性更有可能存活。

Pclass (船舱等级): 船舱等级增加一个单位时，存活的优势相对于不存活的优势减少约66.83%。

SibSp (同船兄弟姐妹或配偶的数量): 每增加一个同船兄弟姐妹或配偶的数量，存活的优势相对于不存活的优势减少约38.66%。

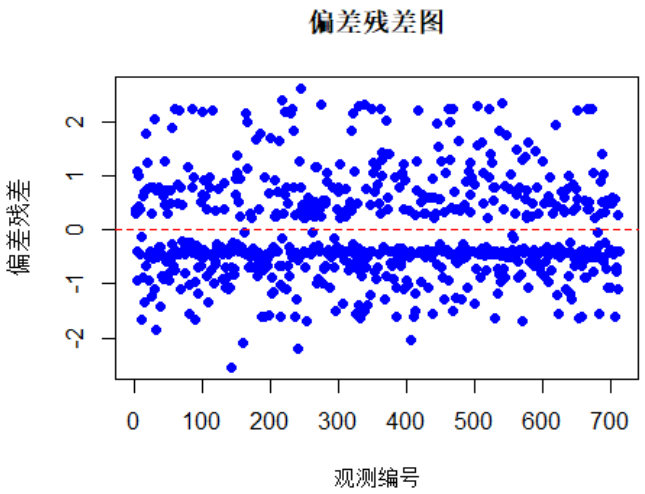
Parch (同船父母或子女的数量): 每增加一个同船父母或子女的数量，存活的优势相对于不存活的优势减少约30.12%。

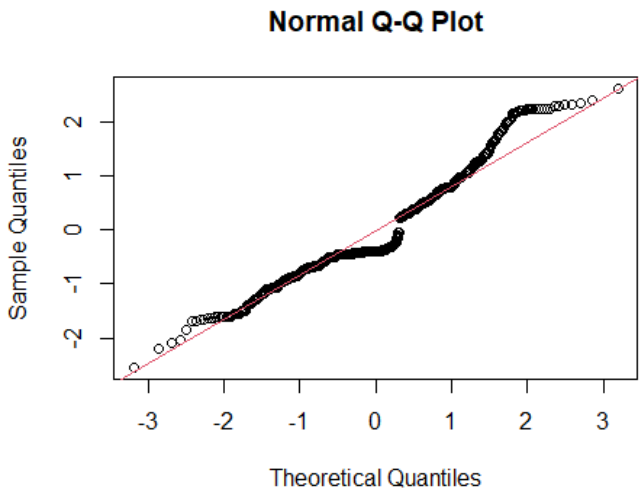
TitleCategoryMaster (头衔为Master的乘客): 相对于其他头衔，头衔为Master的乘客存活的优势比约为34.65，表示这一群体更有可能存活。

EmbarkedC (登船港口为C): 相对于其他港口，从C港口登船的乘客存活的优势比约为1.63，表示这一群体相对更有可能存活。

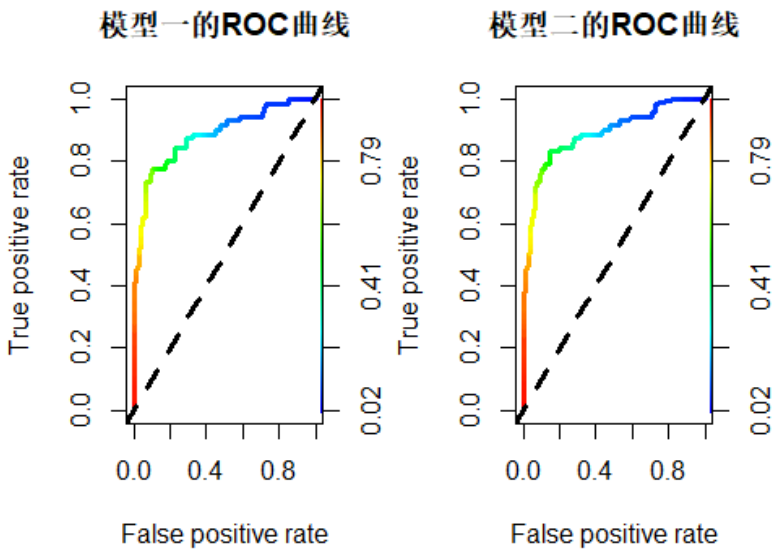
评估模型拟合效果，还可以绘制残差图和QQ图：

如图所示，残差图具有随机分布、无规律、零均值的特点，说明拟合程度较好。QQ图上的点也都拟合标准线较好。





计算最终模型对应的AUC值，同时计算灵敏度、特异度、精确率和F1值。

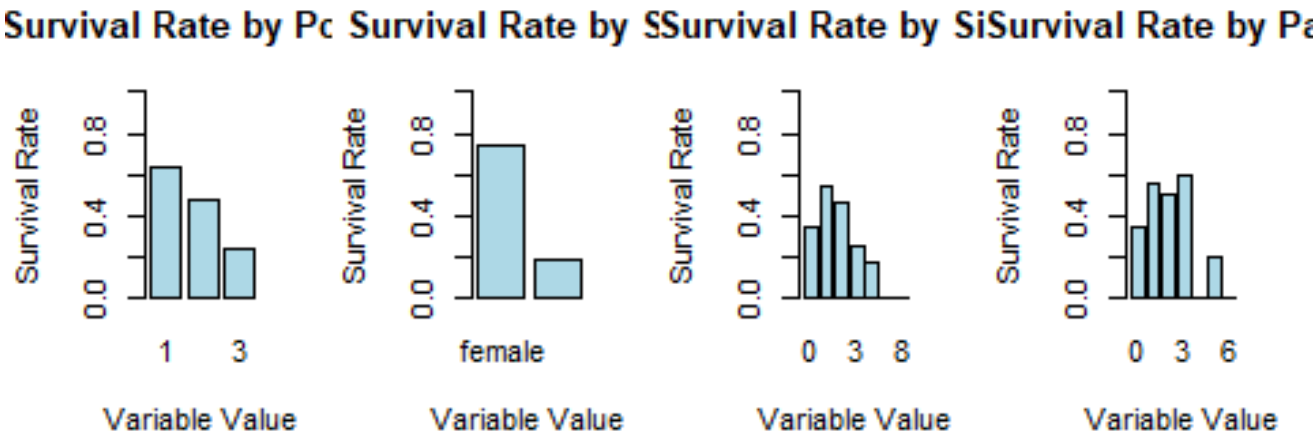


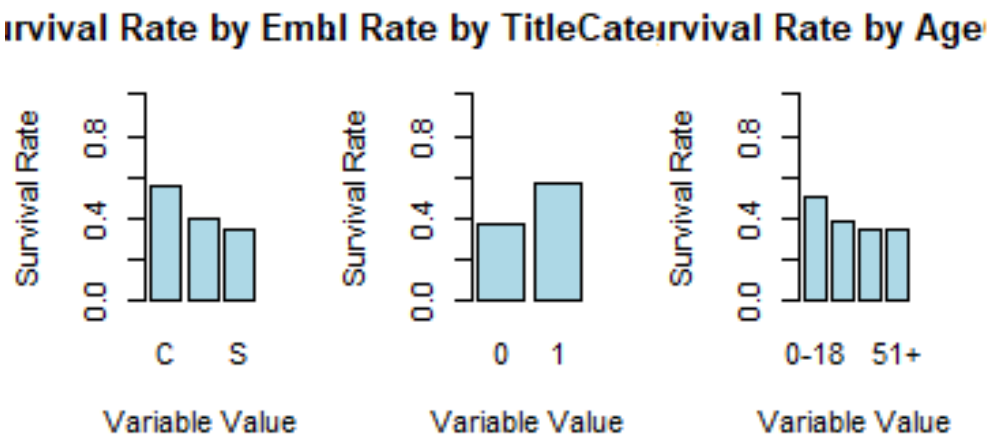

```
## 模型一的AUC值: 0.8844281
## 模型二的AUC值: 0.8889693
```

```
## 模型 1:
## 准确率 (Accuracy): 0.8202247
## 准确率 (Accuracy): 0.8202247
## 灵敏度 (Sensitivity): 0.7746479
## 特异度 (Specificity): 0.8504673
## 精准率 (Precision): 0.7746479
## F1值 (F1 Score): 0.7746479
```

```
## 模型 2:
## 准确率 (Accuracy): 0.8314607
## 灵敏度 (Sensitivity): 0.7887324
## 特异度 (Specificity): 0.8598131
## 精准率 (Precision): 0.7887324
## F1值 (F1 Score): 0.7887324
```

各项指标数据较好，说明模型拟合效果较好。





结论分析：经过如上图示并综合模型解读结果可以发现，生存率和性别、票级、兄弟姐妹/配偶数关系较大，女性、票级越高、有一到两个同行者存活率更高，同时也与年龄有较弱的关系，年龄更小的相对存活率高些，同时有一个特殊的发现，即头衔为Master的人存活率普遍高于其他人。

时序模型——股票分析问题

项目要求

现有从www.nasdaq.com通过网络抓取收集的数据，内容为上市公司的股价和交易量，如苹果、星巴克、微软、思科系统、高通、Meta、亚马逊、特斯拉、Advanced Micro Devices和Netflix，共包含25161行。请利用提供的数据集进行统计分析，并为投资者提出可靠并宝贵的建议。

解决思路

在粗略查看数据集之后，我们发现该数据集存在一些格式问题，因此我们决定首先对数据进行预处理，包括统一格式、进行数据类型的转换等。

由于股票价格和交易量等金融数据的明显时序性特征，我们决定采用时序模型进行建模。为了能够在建模时更好地确定时序模型需要的参数，我们首先对处理后的数据进行了初步分析，包括收盘价格图判断数据趋势、绘制季节图判断数据是否存在季节性、进行相关性分析以确定各公司股票价格间的潜在关系等。

时序模型我们选择的是ARIMA模型。由于使用该模型进行建模时，数据需要是平稳时间序列且为非白噪音序列，因此在建模之前，我们对所获取的数据进行了平稳序列检测和白噪音检测，并按照检测的结果对数据进行了差分处理，并利用差分后的序列进行建模。

由于公司数量较多，且相关性分析表明各公司股票相关性都很强，因此我们对AAPL这只股票的数据进行了详细的建模，对模型进行评估，使用模型进行价格预测并进行相应的分析。

最后，我们使用了ARIMA模型，结合RSI分析和布林带分析，确定了CSCO股票为最佳潜力股。

需要说明的是，通过查询资料，我们了解到许多投资者和分析师会首选使用收盘价进行分析，因为它代表了一天的最终交易价格，通常被视为当天的市场估值；因此从初步分析到建模，除K线图和柱形图外，我们均使用股票收盘价作为分析对象。

解决过程

一、数据预处理

通过summary检查数据类型：

```
   Company              Date             Close.Last          Volume         
 Length:25160       Length:25160       Length:25160       Min.   :1.144e+06  
 Class :character   Class :character   Class :character   1st Qu.:1.200e+07  
 Mode  :character   Mode  :character   Mode  :character   Median :2.672e+07  
                                                          Mean   :5.132e+07  
                                                          3rd Qu.:6.857e+07  
                                                          Max.   :1.065e+09  
     Open               High               Low          
 Length:25160       Length:25160       Length:25160      
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  
```

summary显示，在数据集中，除交易量外，所有数据类型均为字符型。因此，我们首先对处理数据类型，将公司名称转换为因子型、将日期转换为Date类型、将股票价格转换为数值型。

1. **“公司名称”数据类型转换**
2. 从character类型转为factor类型，以便按公司分类预测
3. **“日期”数据类型转换**
4. 把日期这一列的“-”都替换成“/”，使得日期格式统一化为形如07/17/2023的样式，然后转为Date类型
5. **对4列股价的数据类型转换**
6. 发现它之所以是character类型，是因为表格里的数据都包含" ”符，所以我们可以先删除” "符，再转为number类型
7. **整理数据，按日期升序排列**

```
1 data<-data[order(data$Date),]
```

8. 再次检查数据类型以及检查是否有空值

通过summary，发现没有空值，没有负数等异常值，各个数据类型也正确。

```
   Company          Date             Close.Last        Volume              Open        
AAPL    : 2516   Min.   :2013-07-18   Min.   :  1.62   Min.   :1.144e+06   Min.   :  1.62  
AMD     : 2516   1st Qu.:2016-01-14   1st Qu.: 36.57   1st Qu.:1.200e+07   1st Qu.: 36.51  
AMZN    : 2516   Median :2018-07-16   Median : 65.68   Median :2.672e+07   Median : 65.65  
CSCO    : 2516   Mean   :2018-07-15   Mean   :102.46   Mean   :5.132e+07   Mean   :102.43  
META    : 2516   3rd Qu.:2021-01-13   3rd Qu.:134.24   3rd Qu.:6.857e+07   3rd Qu.:134.32  
MSFT    : 2516   Max.   :2023-07-17   Max.   :691.69   Max.   :1.065e+09   Max.   :692.35  
(Other):10064                                                                             
      High              Low        
Min.   :  1.69   Min.   :  1.61  
1st Qu.: 36.89   1st Qu.: 36.13  
Median : 66.49   Median : 64.92  
Mean   :103.83   Mean   :101.01  
3rd Qu.:136.23   3rd Qu.:132.66  
Max.   :700.99   Max.   :686.09  
```

我们发现有的日期是缺失的，所以我们统计了每一年有哪些日期是有数据的：

2014	2015	2016	2017	2018	2019	2020	2021	2022
252	252	252	251	251	252	253	252	251

经过观察，我们发现这些数字与相应年份的证券交易日的数量完全吻合（一年里并不是每一天都是证券交易日，通常周末的时候证券市场不开放），我们可以确定我们的数据集是完整的，无需进行任何额外的补爬取等操作。

9. 转为ts类型

为了便于进行时间序列分析，我们将获取的数据转换为时间序列（ts）类型。

数据集中有每日股票的开盘价，收盘价，最高价，最低价。我们查找资料后认为收盘价更能反映股票变化趋势，所以选择了收盘价为研究重点。

10. 线性插值

由于我们创建的时间序列是从2013年7月18日到2023年7月17日之间的每一天的，而数据集只提供了合法交易日的数据，为了避免数据中缺失某些交易日的情况导致错误和障碍，我们采用线性插值方法来填充非交易日的数据，并将其添加到时间序列中。这样做可以确保我们不会因为数据中有某一天是空的而报错、而无法进行季节性分析等，这样就能更顺利地进后续分析。

```
ts_list<-list()#用list来存每个公司的时序数据
for(i in 1:length(company_data)){
   temp_ts<- ts(data = NA, start = start_date, end = end_date, frequency = 1)
   date_indices <- match(company_data[[i]]$Date, all_dates)
```

```
    temp_ts[date_indices] <- company_data[[i]]$Close.Last # 注意这里只把收盘价填入了ts
    temp_ts<-na.approx(temp_ts) # 线性插值
    ts_list[[length(ts_list) + 1]]<-temp_ts#放进列表
}
```

二、数据可视化

2.1 初步分析

2.1.1 趋势分析

K线图是股票技术分析中常用的图形，能够有效反映股票的走势。我们使用plot_ly函数画出了每个公司的走势，并且用红色标出这一天收盘价高于开盘价，用绿色标出这一天收盘价低于开盘价，并且通过这张图我们可以具体查看每一天的最高价、最低价等。

以TSLA公司为例，在图中可以直观看出它最近几个月中几乎每一天股价都有上升，说明该公司近期股价走势很好。

2.1.2 波动性分析

对每个公司的收盘价，取对数后再进行差分，即diff(log(target$Close.Last))，可以得到对数收益率。对数收益率可以消除价格级别的影响，更能反映出股票价格的变化。

标准差可以衡量数据分布的离散度，它越大表示公司的对数收益率数据的波动性越大。

因此，通过计算对数收益率的标准差，我们可以得到该股票在给定时间段内的波动性大小。

代码：

```r
for(i in seq_along(company_data)) {
  target <- company_data[[i]]
  returns <- diff(log(target$Close.Last))
  volatility <- sd(returns)
}
```

结果：

```
[1] "AAPL Volatility: 0.0180141051660119"
[1] "AMD Volatility: 0.0360936893240302"
[1] "AMZN Volatility: 0.0207401621388138"
[1] "CSCO Volatility: 0.0159020022335713"
[1] "META Volatility: 0.0246159153088786"
[1] "MSFT Volatility: 0.0172859692703764"
[1] "NFLX Volatility: 0.0285152513718961"
[1] "QCOM Volatility: 0.0220122529590671"
[1] "SBUX Volatility: 0.0164570921656759"
[1] "TSLA Volatility: 0.0353856602049915"
```

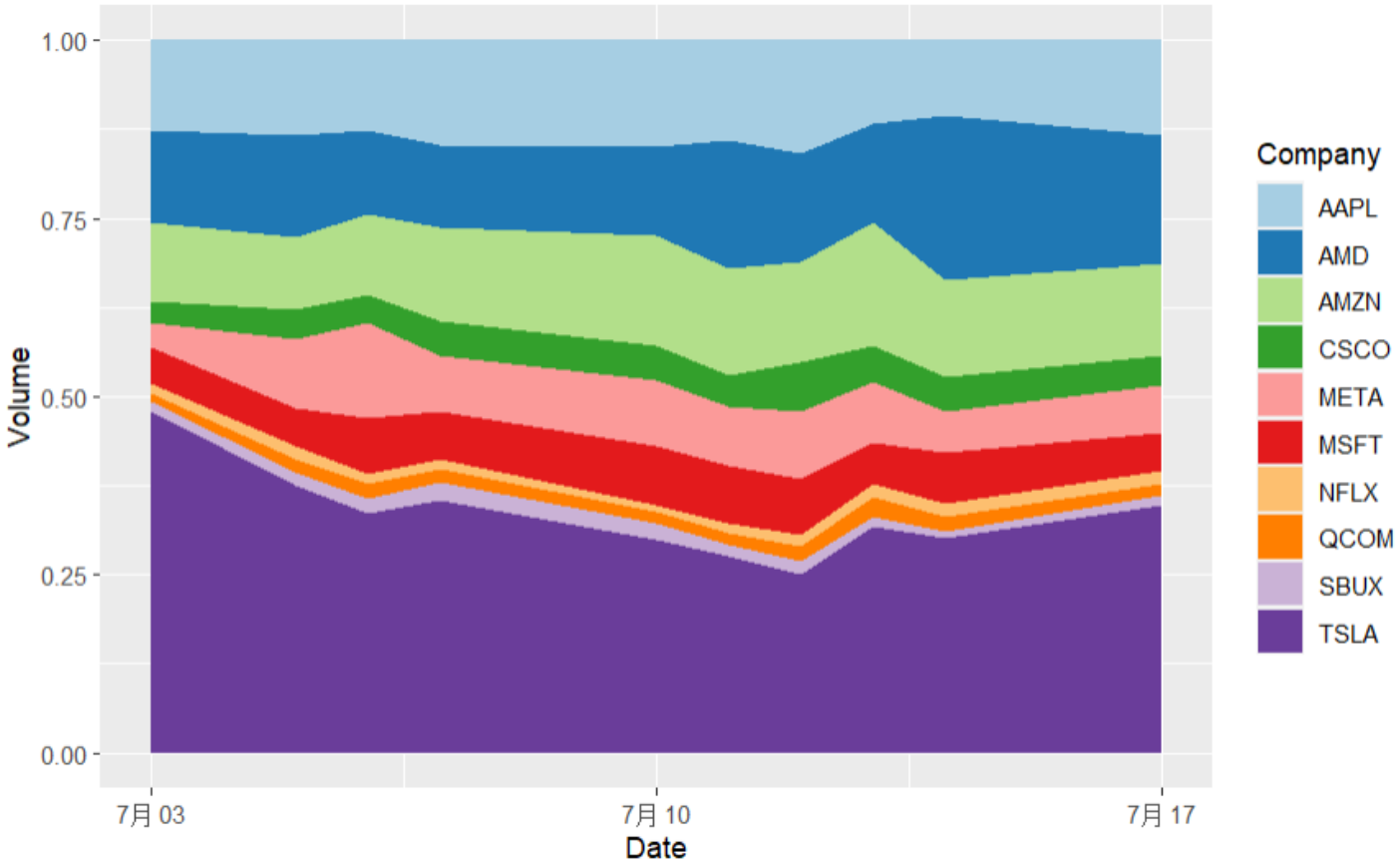
分析：

各个公司的波动性都比较低，这表明这些公司的股票价格在过去十年内的波动幅度较小，市场风险相对较低。

其中，TSLA、AMD公司的波动性明显高于其他公司。特斯拉是电动汽车行业的领导者，而AMD是在半导体市场上与英特尔竞争的重要参与者。这两家公司在新技术、产品创新和市场份额方面都取得了显著成就，这可能会引发投资者对它们未来表现的高度关注。

2.1.3 短期交易量对比

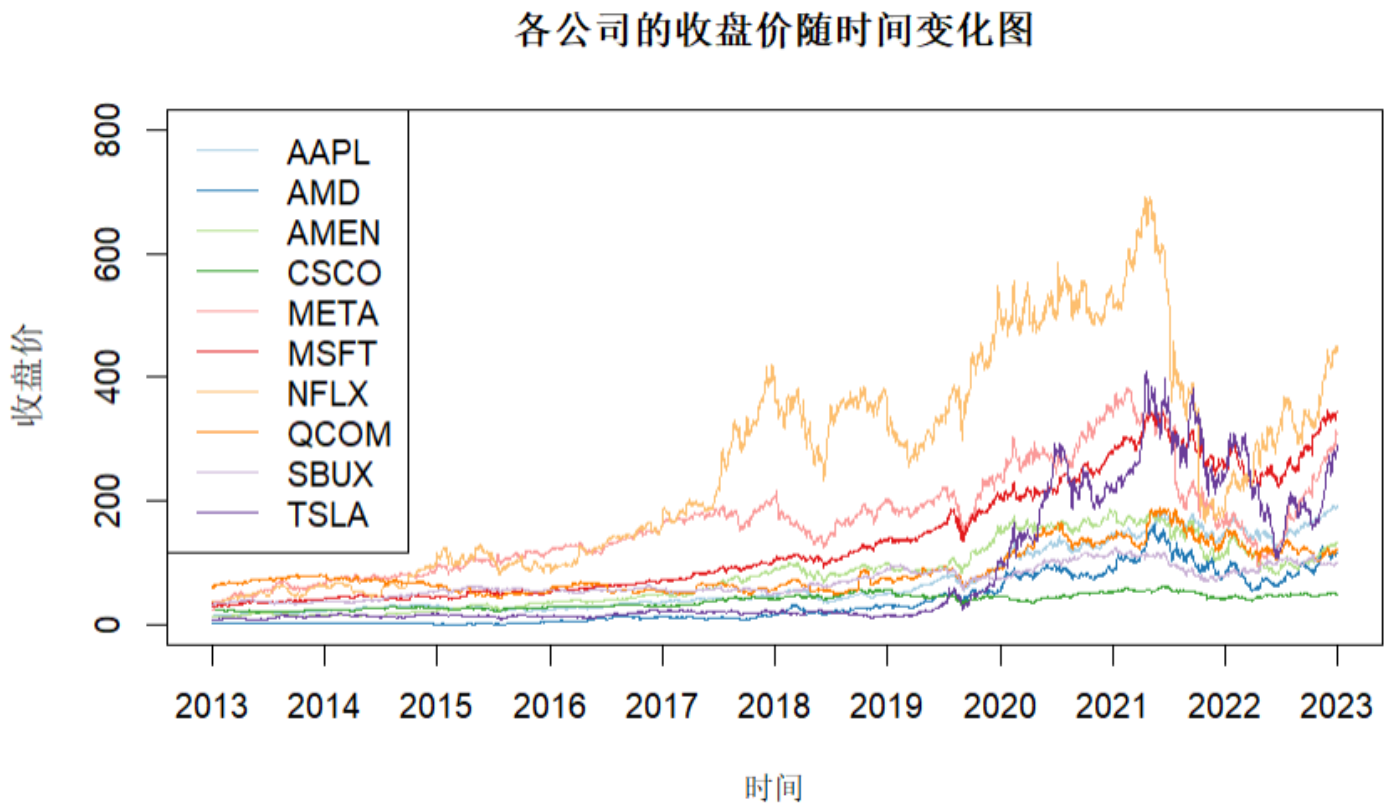
我们通过堆积图比较了各公司最近的交易量。



从交易量的对比图中，我们可以看出当前市场的划分状况。特斯拉、苹果、AMD、亚马逊在市场份额上有非常大的优势。

2.1.4 时序曲线对比

由于我们的研究重点是收盘价，因此针对收盘价绘制了时序曲线。通过时序曲线可以直接看出每个公司的股价随时间的变化趋势。



从图上能直观看出的是：

- 大多数时候QCOM（高通公司）股价显著高于其他公司。高通作为5G技术的主要供应商之一，受益于5G的广泛部署，是投资者青睐的对象，从而推高了股价。
- 收盘价数据波动较为显著，各公司可能都具有上升的趋势；
- 收盘价涨跌并**未表现出明显周期性**；
- 各家公司收盘价具体价格差异较大，但整体走势较为一致，可能具有一定相关性；
- 多家公司在2018-2023年间的平均股票收盘价明显高于2013-2018年，这可能是因为近年来云计算、人工智能和新兴技术的广泛应用。因此，十年前和现在的股价可能表现出不一样的趋势，因此后续分析时可能需要重新划定训练集，仅截取一段时间内的股票收盘价进行分析。

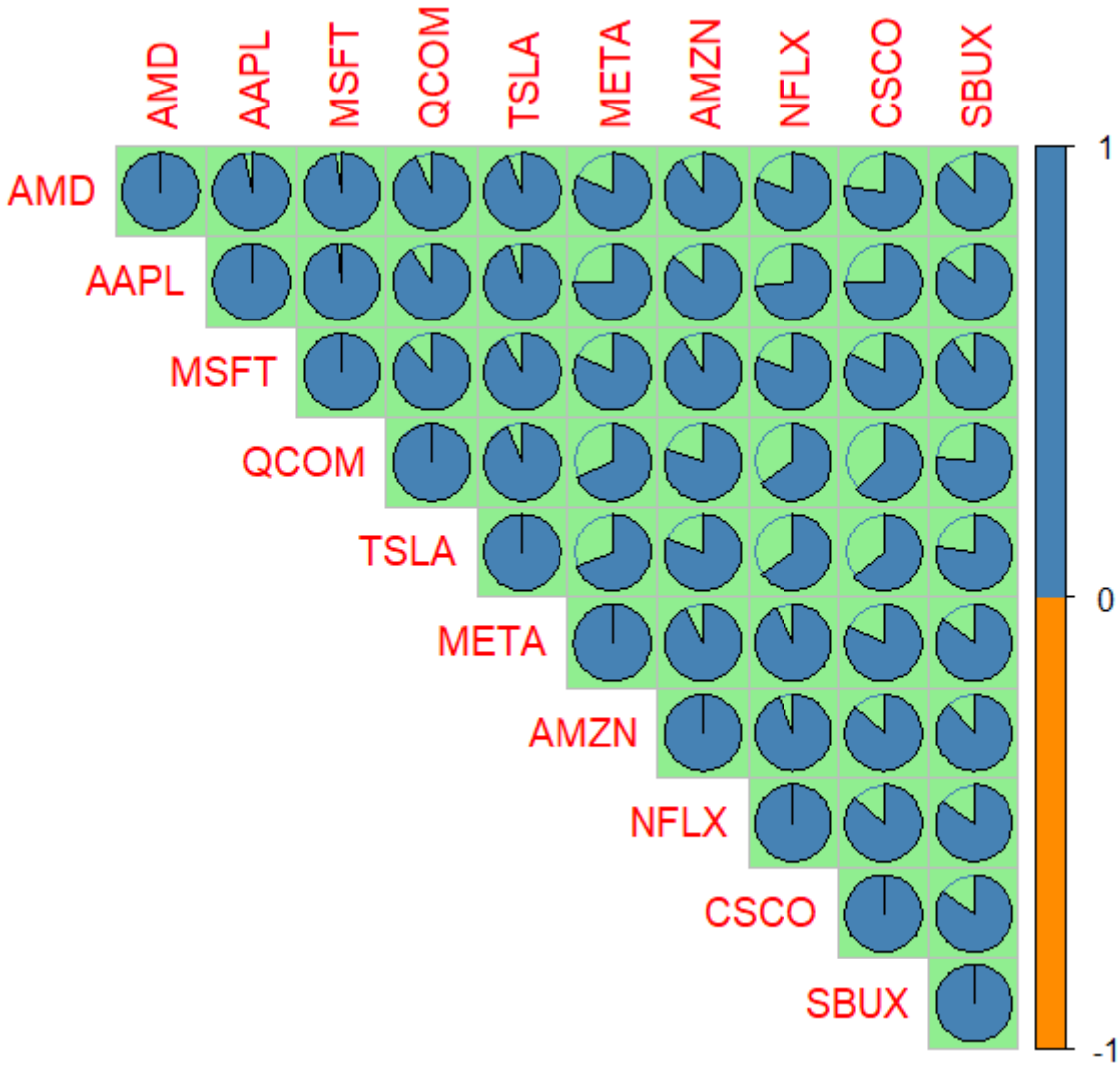
为了之后能够更好地进行建模，我们继续分析了各家公司股票收盘价的波动性、相关性、季节性。

2.2 相关性分析

2.2.1 公司之间的股价相关性分析

从所有数据中，我们按公司进行了分类，然后提取各个公司的收盘价，分析它们之间的相关性。

1. **相关性图**



2. **查看显著性水平**

```
cor_matrix <-
  rcorr(as.matrix(all_company_close.last[, -1]))
cor_matrix$P
```

代码运行结果：

	AAPL	AMD	AMZN	CSCO	META	MSFT	NFLX	QCOM	SBUX	TSLA
AAPL	NA	0	0	0	0	0	0	0	0	0
AMD	0	NA	0	0	0	0	0	0	0	0
AMZN	0	0	NA	0	0	0	0	0	0	0
CSCO	0	0	0	NA	0	0	0	0	0	0
META	0	0	0	0	NA	0	0	0	0	0
MSFT	0	0	0	0	0	NA	0	0	0	0
NFLX	0	0	0	0	0	0	NA	0	0	0
QCOM	0	0	0	0	0	0	0	NA	0	0
SBUX	0	0	0	0	0	0	0	0	NA	0
TSLA	0	0	0	0	0	0	0	0	0	NA

P值表示观察到的相关系数是否显著不同于零。我们得到的P值都为0，说明各个公司的收盘价存在非常强的线性相关性。

结合相关性矩阵和相关性图，我们可以得知各家公司的股票收盘价之间均存在强相关性。我们分析这是由于该数据集的股票均为科技市场股票，同市场下的股票股价变化规律类似。

在这样的结论下，我们决定在之后建模的过程中，聚焦于分析针对TSLA股票建立的模型，并将结论类推到其他股票的拟合模型中。

对于TSLA股票，我们也发现它和QCOM、AAPL等公司的相关性很强，因此在后续预测的时候可以考虑加入QCOM公司的股价作为相关因素。同时，注意到QCOM和AAPL公司的相关性系数达到了0.9083，说明它们之间的相关性也很强，为了避免模型出现共线性，我们最好不要把QCOM和AAPL公司同时加入模型。

2.2.2 股价和股票体量之间的相关性分析

我们找出了每个公司的所有数据，然后用cor函数分析了各个公司的股票体量和自身收盘价的关系。

结果如下：

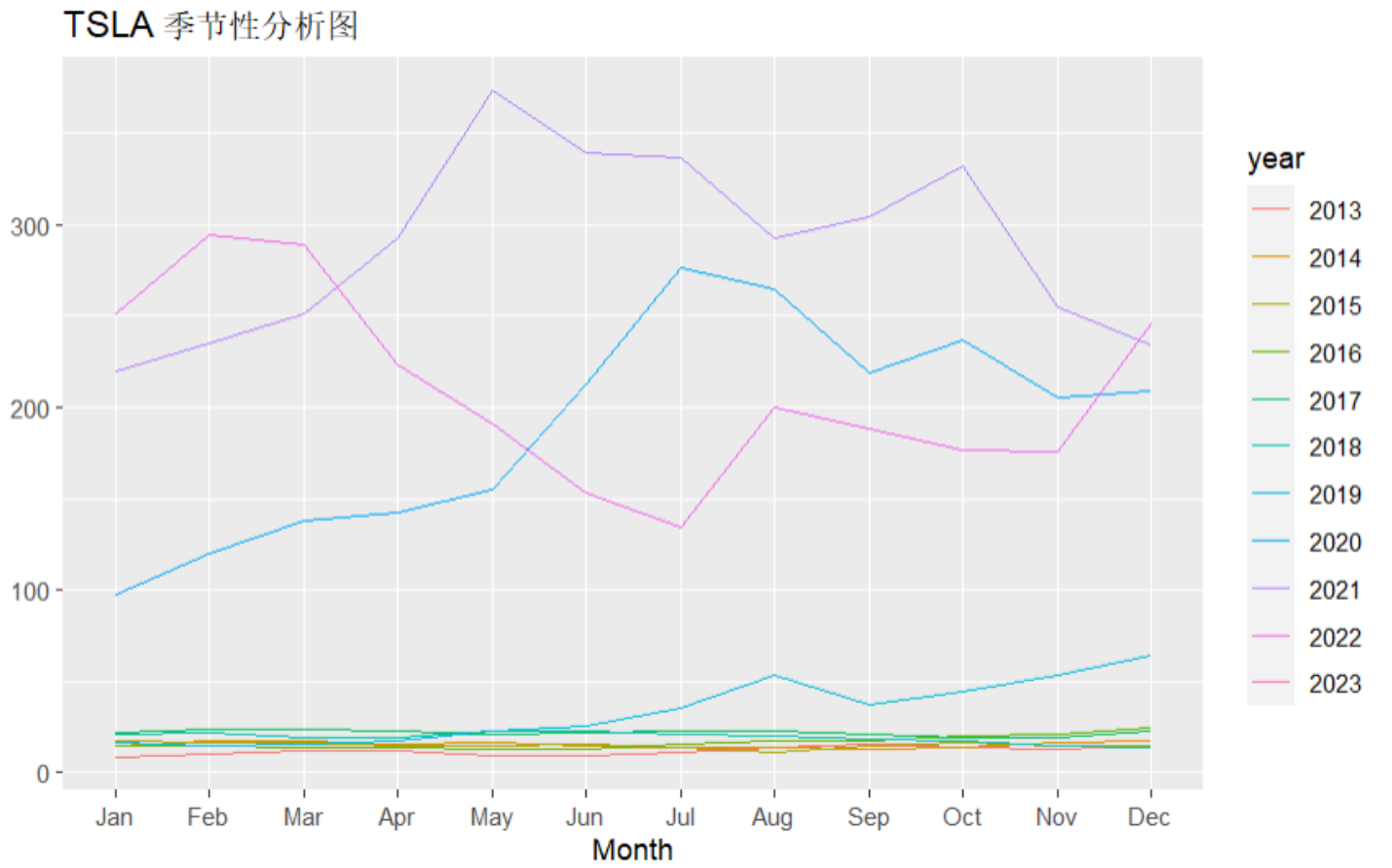
```
[1] "AAPL 的收盘价和自身体量的相关性系数为:  -0.472460812454884"
[1] "AMD 的收盘价和自身体量的相关性系数为:  0.357577906559595"
[1] "AMZN 的收盘价和自身体量的相关性系数为:  0.0102551776030346"
[1] "CSCO 的收盘价和自身体量的相关性系数为:  -0.341452191419137"
[1] "META 的收盘价和自身体量的相关性系数为:  -0.383190025950612"
[1] "MSFT 的收盘价和自身体量的相关性系数为:  -0.107740594084771"
[1] "NFLX 的收盘价和自身体量的相关性系数为:  -0.438112843682131"
[1] "QCOM 的收盘价和自身体量的相关性系数为:  -0.150571335899325"
[1] "SBUX 的收盘价和自身体量的相关性系数为:  -0.259468618397629"
[1] "TSLA 的收盘价和自身体量的相关性系数为:  -0.0722565999175452"
```

发现各家公司的收盘价和自己当天的股票体量基本上没有显著相关性，TSLA的股价和它的体量相关性绝对值只有0.07，因此预测时可以不把体量考虑进去。

2.2.3 季节性分析

我们将每家公司的数据以“月”为单位进行划分，使用该月每日股票收盘价的平均值作为该月的数据，进行季节性分析。

以TSLA公司为例：



可以看出它的收盘价和月份没有明显的关系，例如，对于同一个月份，股价的变化可能是不一致的。举个例子，2021年7月股价上升了，但是2020年和2022年的7月股价都下降了。因此，我们不能得出哪个月份的股价普遍会上升的结论。

我们查看了每个公司的季节性分析图，最后确定每家公司的股价都不存在季节性。

这也和公司性质是相符的，我们研究的10家公司都是科技公司，科技公司的股价通常受到更多的因素影响，如市场需求、技术创新和行业竞争等，而不太受季节因素的影响。通常，往往是旅游公司、暖气公司等容易受到季节影响。我们的研究结果也表明，这些科技公司的股价在不同的季节没有明显的规律性变化。这对于投资者来说是一个重要的信息，他们可以更加专注于分析其他与公司业务相关的因素，以做出更准确的投资决策。

三、模型建立

使用ARIMA模型进行建模。

它要求模型为**平稳序列**且为**非白噪音序列**，否则需要首先对数据进行差分等处理，然后再开始建模。

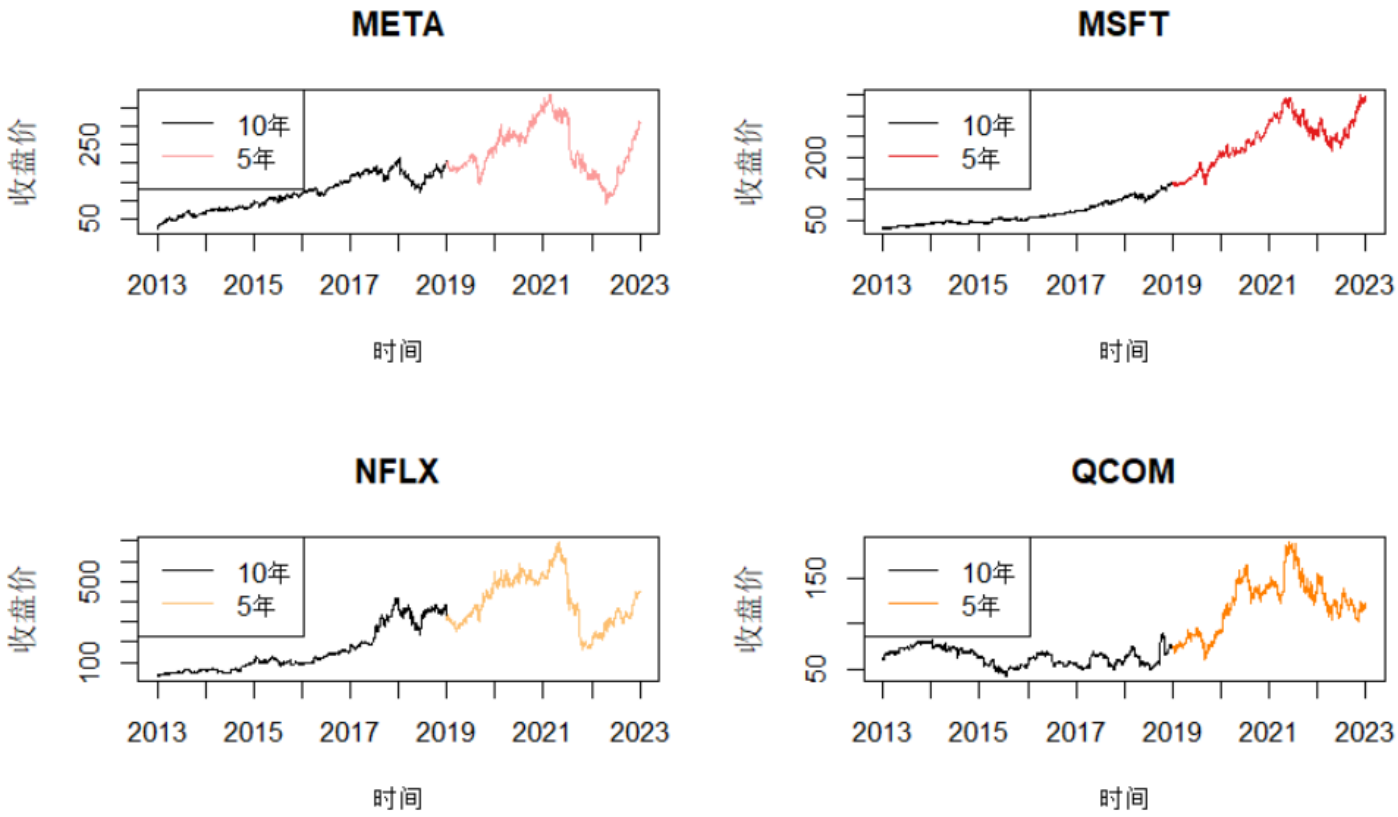
3.1 数据选取

我们一共有10年的数据，注意到在这十年中，后5年股价的规模明显大于前五年。因此，我们决定对数据进行两个不同的分析：一部分是针对整个10年的数据，另一部分是专注于最近5年的数据。

在正式进行分析之前，我们预测，10年的数据更能反映股票的长期变化趋势，而5年的数据更贴近当前情况。

5年的数据涵盖了股票出现巨幅增长的时期，因此可以更好地捕捉到目前市场上的走势和变化。通过综合分析这两组数据，我们可以得到更全面的股票状况评估，从而做出更准确的决策。

下图大致展示了5年数据和十年数据的不同，各个图的标题是公司名字：



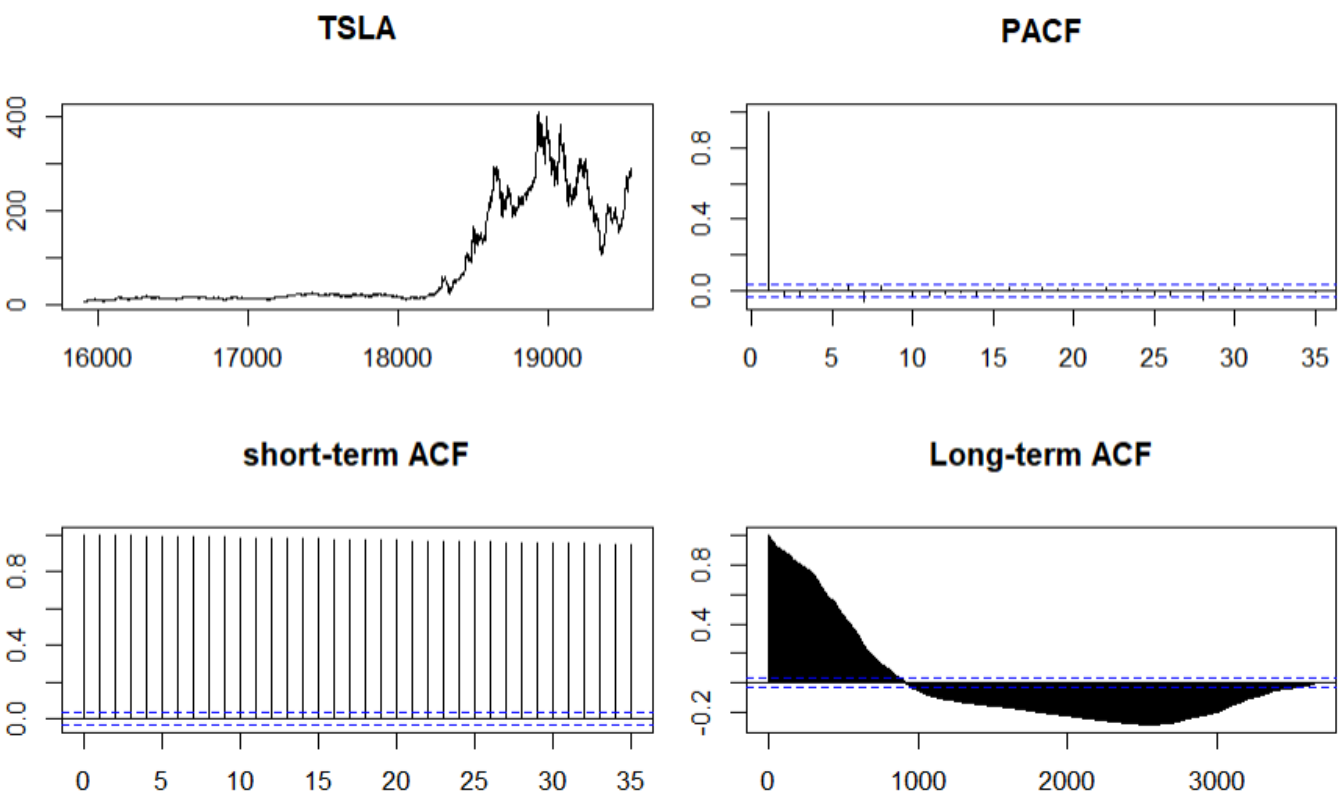
也就是说，我们取了两种时间跨度：十年和五年，对各家公司2013年7月18日-2023年7月17日的数据和2019年7月18日-2023年7月17日的数据都进行了分析，以更好地确定训练集。

3.2 建模条件分析

3.2.1 初步的平稳性检验

我们需要对每个公司画出时间序列图，PACF图，短期的ACF图和长期的ACF图（所有天数的ACF图）

1. **十年数据**
    - **画图分析**
2. 我们画出了每个公司的4种图，下面以TSLA公司为例进行分析



短期ACF图几乎看不出衰减，长期ACF图可见ACF系数随滞后期的增加而缓慢衰减，最终渐趋趋于0，因此数据具有长期相关性。初步判断十年的股价序列为非平稳序列。

类似地我们分析了每个公司，发现所有公司这十年的股价序列都不是平稳序列。

- **ADF检测**

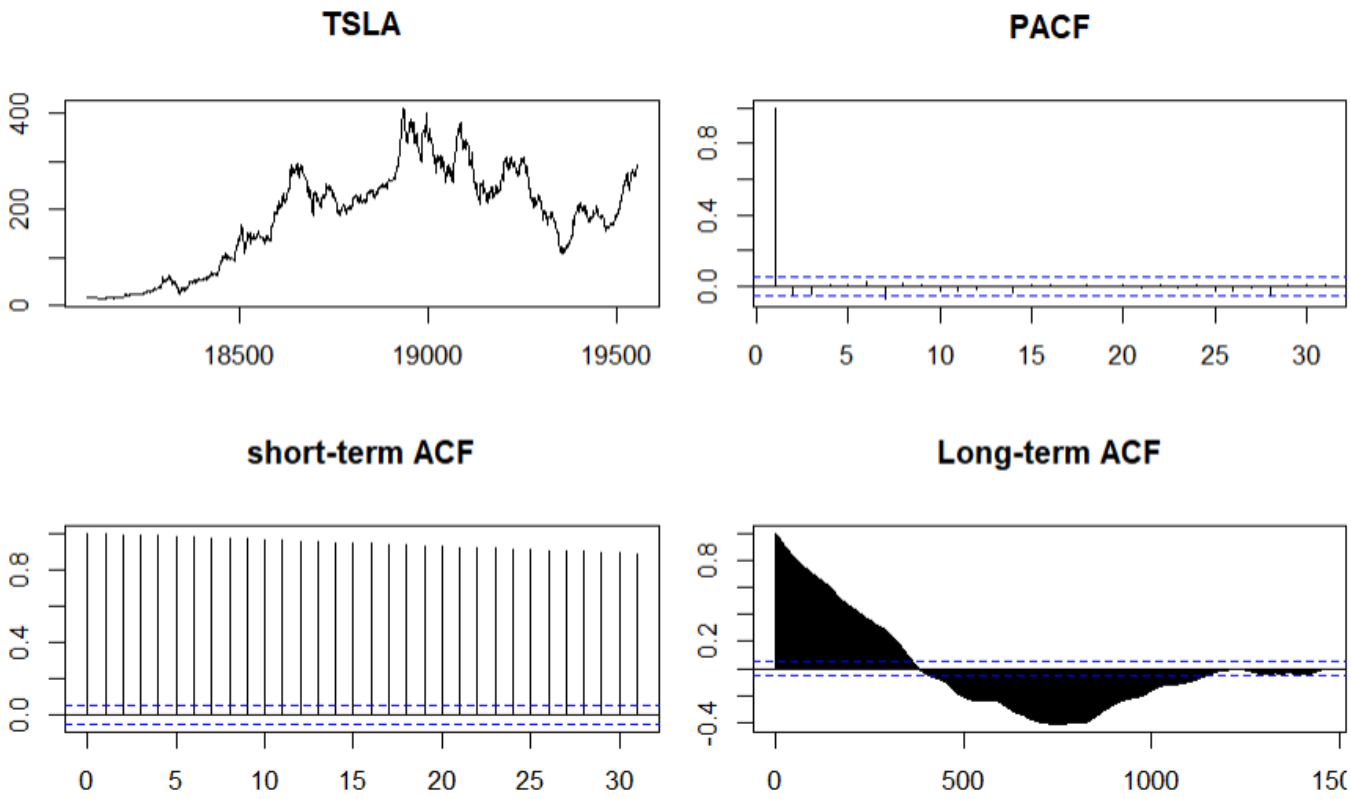
1. 为了更准确的判断数据平稳性，下面进行ADF检测，结果如下。

```
ADF Test for NFLX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -1.937, Lag order = 15, p-value = 0.605
alternative hypothesis: stationary

ADF Test for QCOM :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -2.2575, Lag order = 15, p-value = 0.4693
alternative hypothesis: stationary

ADF Test for SBUX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -2.5369, Lag order = 15, p-value = 0.351
alternative hypothesis: stationary

ADF Test for TSLA :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -2.1887, Lag order = 15, p-value = 0.4985
alternative hypothesis: stationary
```

ADF测试均显示**p值大于0.05**，因此接受**数据序列不平稳**的假设。

2. **五年数据**
    - **画图分析**

类似地可以进行画图和分析，发现各个公司近五年的股价也不是平稳的。



- **ADF检测**

同样地，对五年的数据进行了ADF检测，发现它们的P值均大于0.05 。说明近五年的股价确实不是平稳的。

```
ADF Test for NFLX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -1.3934, Lag order = 11  p-value = 0.8351
alternative hypothesis: stationary

ADF Test for QCOM :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -1.7066, Lag order = 11  p-value = 0.7025
alternative hypothesis: stationary

ADF Test for SBUX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -1.8477, Lag order = 11  p-value = 0.6428
alternative hypothesis: stationary

ADF Test for TSLA :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -1.9518, Lag order = 11  p-value = 0.5987
alternative hypothesis: stationary
```

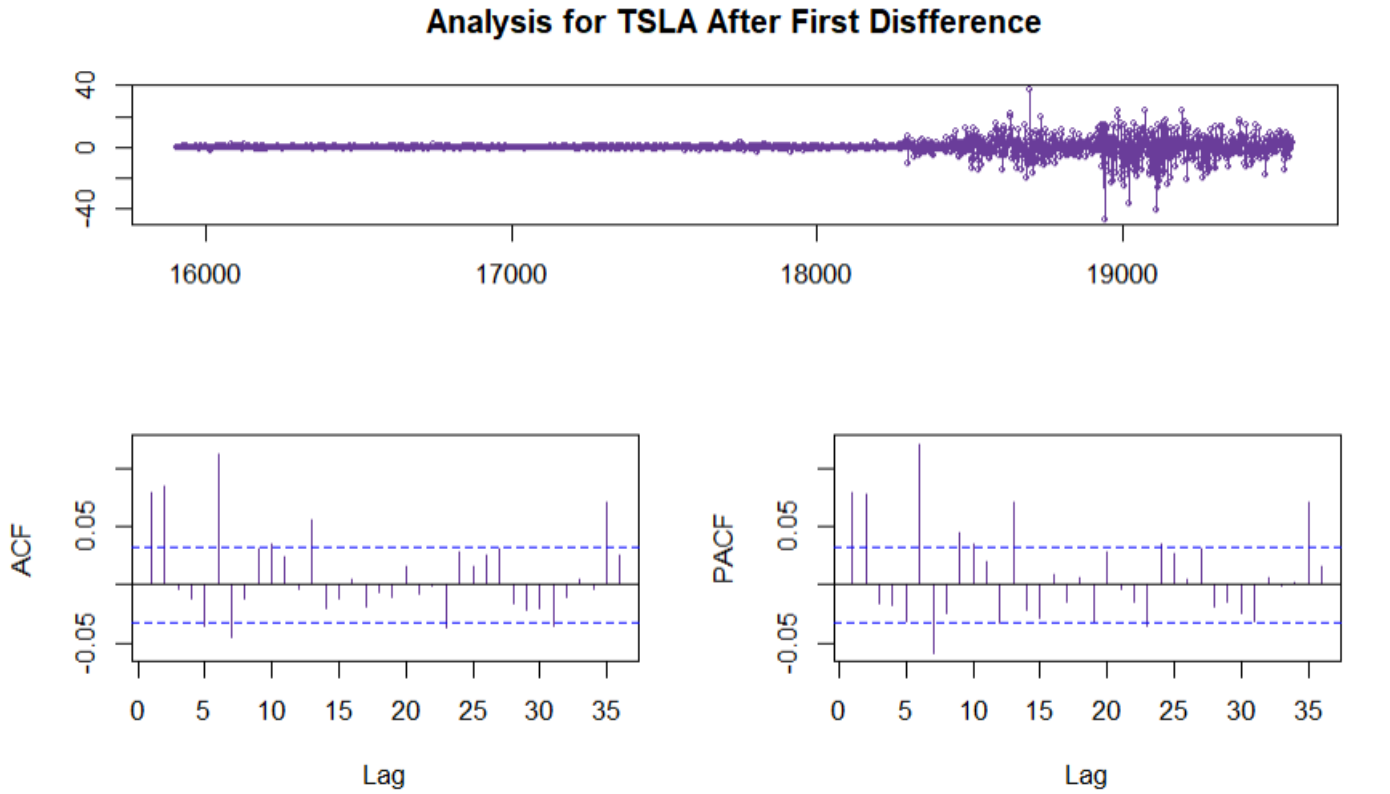
3.2.2 差分处理和进一步的平稳性检验

1. **画图分析**

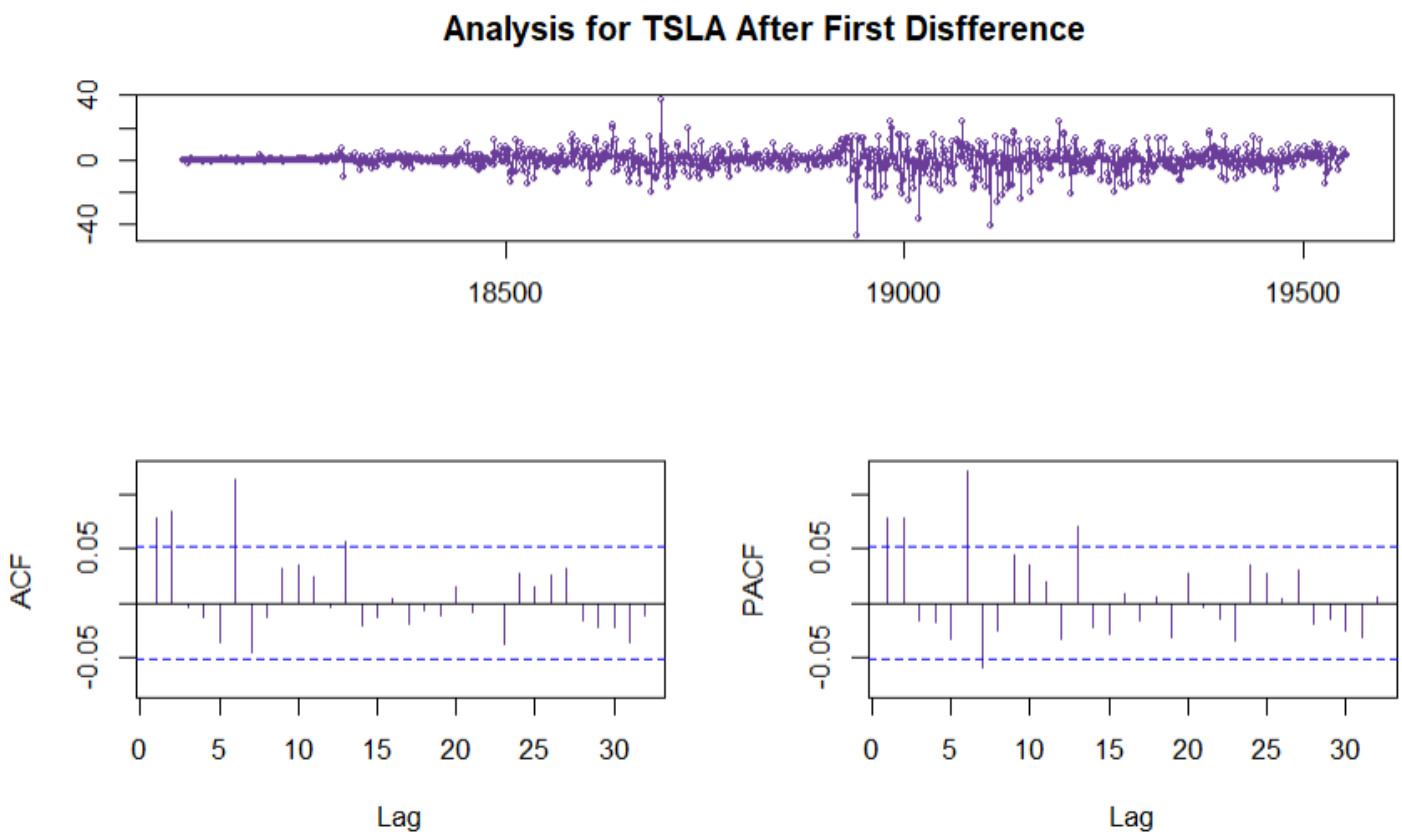
先采用diff函数进行差分，然后用tsdisplay函数，查看股价的时序图、ACF、PACF。

（1）十年数据

同样以TSLA公司为例，可以看出它已经基本平稳了，ACF和PACF都有收敛趋势。



- **（2）五年数据**



一阶差分后的序列，自相关性系数及偏自相关系数基本都在0值附近波动，偶尔有超过置信区间的时候，但是滞后阶数非常随机；ACF图中并未表现出周期性。初步判断一阶差分后的序列为平稳序列，无周期性。

此外，五年数据差分后的序列从数据分布的区间上看比十年数据更为均匀。

2. **ADF测试**
   - **（1）十年数据**

```
Warning: p-value smaller than printed p-valueADF Test for NFLX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -14.722, Lag order = 15, p-value = 0.01
alternative hypothesis: stationary

Warning: p-value smaller than printed p-valueADF Test for QCOM :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -14.457, Lag order = 15, p-value = 0.01
alternative hypothesis: stationary

Warning: p-value smaller than printed p-valueADF Test for SBUX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -14.906, Lag order = 15, p-value = 0.01
alternative hypothesis: stationary

Warning: p-value smaller than printed p-valueADF Test for TSLA :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -14.038, Lag order = 15, p-value = 0.01
alternative hypothesis: stationary
```

   - **（2）五年数据**

```
Warning: p-value smaller than printed p-valueADF Test for NFLX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -11.392, Lag order = 11, p-value = 0.01
alternative hypothesis: stationary

Warning: p-value smaller than printed p-valueADF Test for QCOM :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -10.857, Lag order = 11, p-value = 0.01
alternative hypothesis: stationary

Warning: p-value smaller than printed p-valueADF Test for SBUX :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -11.864, Lag order = 11, p-value = 0.01
alternative hypothesis: stationary

Warning: p-value smaller than printed p-valueADF Test for TSLA :

        Augmented Dickey-Fuller Test

data:  current.data
Dickey-Fuller = -10.305, Lag order = 11, p-value = 0.01
alternative hypothesis: stationary
```

采用ADF测试，发现无论是十年的数据还是五年的数据，p值均显著低于0.05，可以认定各个公司差分后序列为平稳序列。

3. **白噪音检验**

调用Box.test函数，通过获得的LB统计量和对应的p值来判断数据是否是纯随机的。结果如下。

- **（1）十年数据**

```
BOX Test for MSFT :

        Box-Pierce test

data:  current.data
X-squared = 1.6241, df = 1  p-value = 0.2025

BOX Test for NFLX :

        Box-Pierce test

data:  current.data
X-squared = 0.81448, df = 1, p-value = 0.3668

BOX Test for QCOM :

        Box-Pierce test

data:  current.data
X-squared = 0.28156, df = 1, p-value = 0.5957

BOX Test for SBUX :

        Box-Pierce test

data:  current.data
X-squared = 0.048579, df = 1, p-value = 0.8256

BOX Test for TSLA :

        Box-Pierce test

data:  current.data
X-squared = 22.908, df = 1  p-value = 1.699e-06
```

- **（2）五年数据**

```
BOX Test for MSFT :

        Box-Pierce test

data:  current.data
X-squared = 0.73042, df = 1  p-value = 0.3927

BOX Test for NFLX :

        Box-Pierce test

data:  current.data
X-squared = 0.034494, df =  , p-value = 0.8527

BOX Test for QCOM :

        Box-Pierce test

data:  current.data
X-squared = 1.679, df = 1,  -value = 0.1951

BOX Test for SBUX :

        Box-Pierce test

data:  current.data
X-squared = 0.43767, df = 1  p-value = 0.5083

BOX Test for TSLA :

        Box-Pierce test

data:  current.data
X-squared = 9.0607, df = 1, p-value = 0.002612
```

我们分为了十年数据和五年数据两种时间跨度，在两个时间跨度下，分别对所有公司的股价序列进行了白噪声检验。p值大于0.05说明接受原假设（该数据是白噪声的）。

Box-pierce检验表明，除了TSLA，其他公司的股票收盘价在进行一阶差分后是白噪音序列，无法使用。也就是说，只有TSLA公司的股价是可以使用ARIMA来建模预测的。

在这种情况下，根据上一步骤中队各家公司股票收盘价相关性的分析，我们决定**聚焦于对TSLA的建模**，并将该模型用于其他公司股票的预测。

通过上述分析，我们确定了建模时较为合适的**d值为1**（因为一阶差分后，数据变成了平稳的）。

另外，一阶差分后，TSLA的ACF图和PACF图整体上看都在13阶后截尾，但实际上只有1、2、6阶时的数据超出虚线较多，因此可尝试**p为1、2、6、13**及**q为1、2、6**的模型。

根据初步分析中的季节性分析以及差分后序列的ACF图，我们认为模型**无需添加季节差分参数**。

3.3 建立预测模型

3.3.1 测试集和训练集的数据划分

沿用之前划分的两种数据：十年数据和五年数据。

对于十年数据，我们将使用除了最后一个月之外的数据作为训练集，并将最后一个月的数据作为测试集。

对于五年数据，同样使用除了最后一个月之外的数据作为训练集，并将最后一个月的数据作为测试集。

3.3.2 ARIMA模型参数的选取

- **（1）十年数据**

通过观察TSLA公司的ACF图，我们选取了以下几组参数来构建ARIMA模型，并和aoto.arima得到的AIC进行比较：

```
arima(train_data_10y,order=c(1,1,2))
#aic = 19774.58
arima(train_data_10y,order=c(2,1,2))
#aic = 19737.59
arima(train_data_10y,order=c(6,1,2))
#aic = 19699.01
arima(train_data_10y,order=c(13,1,2))
#aic = 19682.22
auto.arima(train_data_10y)
#ARIMA(4,1,4) ,AIC=19668.04
```

使用auto.arima拟合时。模型参数同我们预先估计的相差较大，但从ACF图和PACF图上看，我们认为我们选择的参数更为合理，因此舍弃auto.arima提供的模型。

最后，对十年数据，我们选择arima(13,1,2)下的模型

- **（2）五年数据**

我们发现acf图2阶后截尾；pacf图整体上13阶后截尾，1、2、6、13超出虚线较多。因此选取了下列模型

```
arima(train_data_5y,order=c(1,1,2))
# aic = 4197.29
arima(train_data_5y,order=c(2,1,2))
# aic = 4192.57
arima(train_data_5y,order=c(6,1,2))
# aic = 4183.74
arima(train_data_5y,order=c(13,1,2))
```

```
# aic = 4187.41
auto.arima(train_data_5y)
#ARIMA(3,1,2), AIC = 4185.52
```

因此，对五年数据，我们选择AIC最小的arima(6,1,2)下的模型

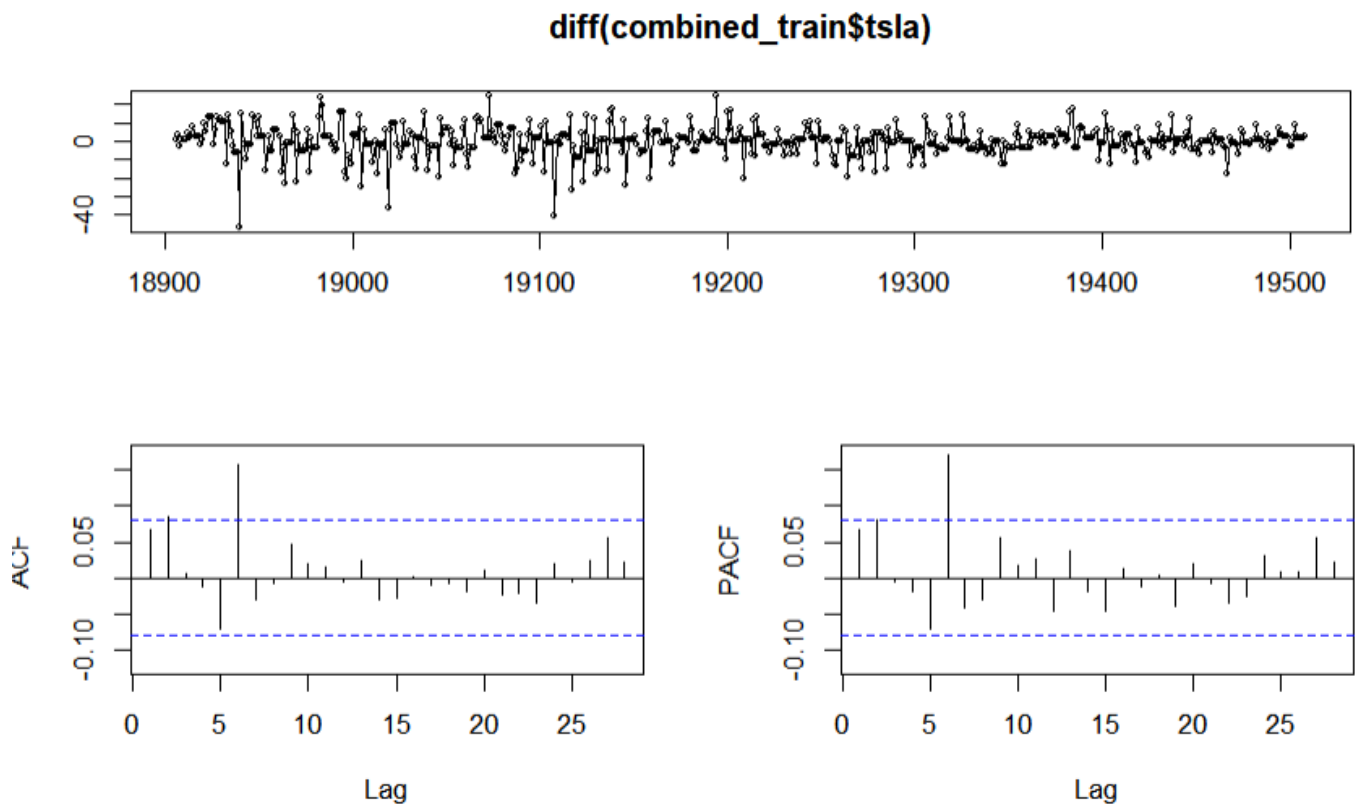
- **（3）基于五年数据的多元ARIMA**

基于前面的相关性分析，我们发现TSLA公司和AAPL公司有较强的相关性，和它自身的体量没有太大相关性。

因此，现在可以取出AAPL公司的相关数据，帮助预测。

- **AAPL公司和TSLA公司近五年的股价时序图比对：**

- 可以看出，它们有着较强的相似性。
  - **平稳性检验**



- 经过一阶差分之后，数据呈现出平稳的特征。这一点可以从自相关函数图和偏自相关函数图中得到佐证，图中前两个滞后阶的值较为显著。
  - **建模代码：**

```
fit_multi <- auto.arima(combined_train[,"tsla"], xreg =
combined_train[,"aapl"])
#指出，要预测的是tsla，参数是aapl
#通过auto.arima自动找到最优参数
```

  - **预测结果：**

```
Series: combined_train[, "tsla"]
Regression with ARIMA(2,1,2) errors

Coefficients:
          ar1      ar2     ma1     ma2    xreg
      -0.0364  -0.8875  0.0612  0.9623  1.9277
s.e.   0.0629   0.0690  0.0364  0.0529  0.1062

sigma^2 = 39.6:  log likelihood = -1962.64
AIC=3937.27   AICc=3937.41   BIC=3963.68
```

发现多元ARIMA的AIC为3937.27，而前面的单因素ARIMA找到的基于五年数据的最优预测模型的AIC为4183.74，说明加入AAPL公司的股价作为影响因子后，模型精度可能会得到一定提升。

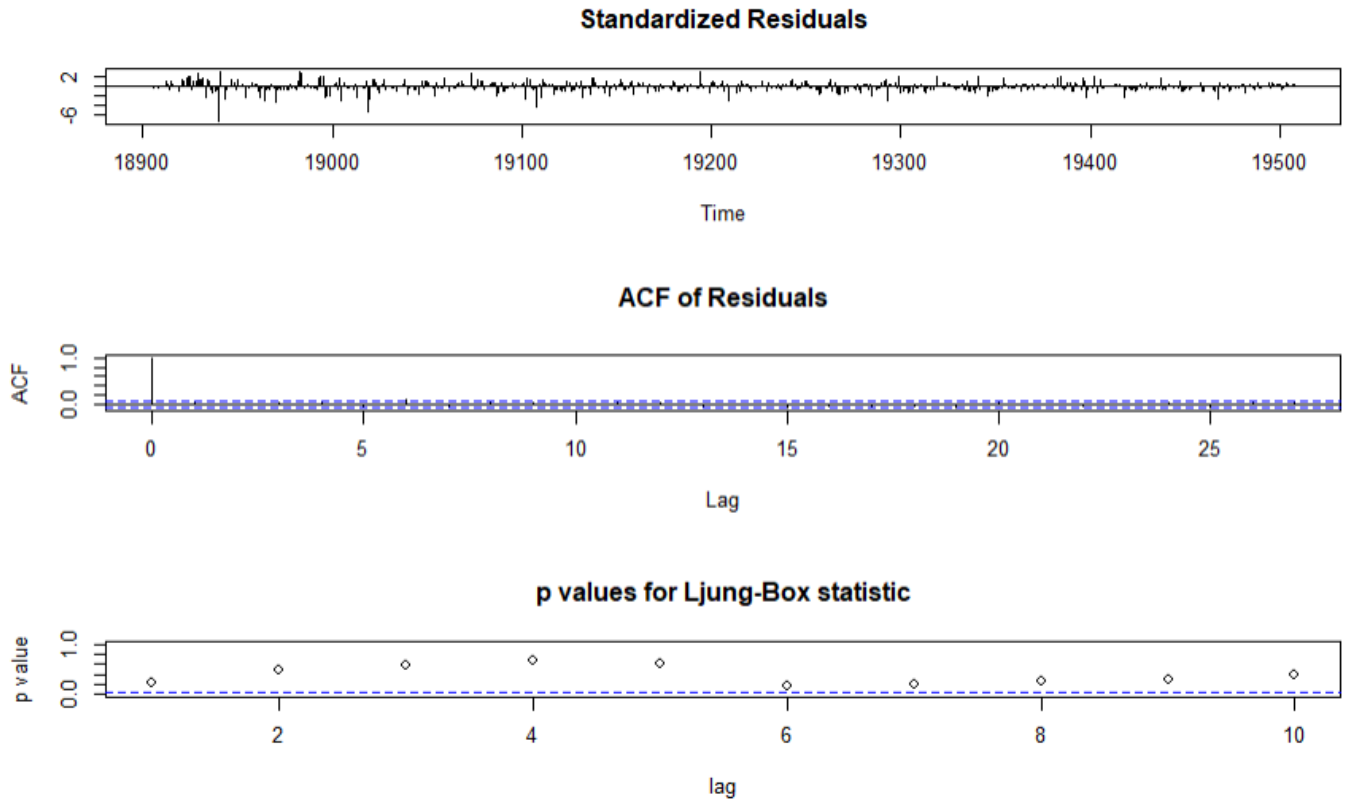
而且对于这个模型：ARIMA(2,1,2)也有着较好的可解释性

四、模型检验

使用R语言的tsdiag函数和accuracy函数对前面建立的三个模型：十年模型、五年模型、五年的多元ARIMA模型进行定量评估。

4.1 检查残差相关性

调用tsdiag来检查残差相关性，以多元预测模型为例进行分析



发现三种模型的残差自回归系数均为0，Ljung-Box检验的p值均在0.05之上，说明10年模型和5年模型以及多元arima的残差之间不具有相关性。

上面这些检测结果，说明我们已经选取了比较好的参数并较为成功地建立了预测模型。

4.2 准确度评估

我们之前已经取出了TSLA公司在数据集中最后47天的数据作为测试数据，因此可以令三个模型各自生成47天的预测数据，然后与真实值比较。

（1）使用accuracy函数

通过调用accuracy函数，我们可以得到模型在训练数据和测试数据上的平均绝对误差（MAE）、均方误差（MSE）、均方根误差（RMSE）、平均绝对百分比误差（MAPE）等，从而能衡量模型的预测能力。

- **在训练数据上比较准确度**

结果如下：

```
[1] "10 year"
                     ME     RMSE      MAE         MPE     MAPE      MASE          ACF1
Training set 0.04258063 3.695263 1.561282 0.04592887 1.696462 0.9813436 0.0005272948
[1] "5 year"
                      ME     RMSE      MAE          MPE     MAPE      MASE         ACF1
Training set -0.07277151 7.646159 5.256779 -0.06502664 2.110393 0.9839959 0.001020826
[1] "multi"
                     ME     RMSE      MAE         MPE     MAPE      MASE        ACF1
Training set -0.1969887 6.261856 4.228283 -0.1263541 1.712899 0.7914757 0.04789617
```

发现三种模型的性能都能较好地拟合训练数据，其中十年模型的均方差表现最好。

- **在测试数据上比较准确度**

由于三个模型均预测同一时间段的数据，因此这种方式有助于控制变量。

执行结果如下：

```
[1] "10 year"
                ME     RMSE      MAE      MPE     MAPE      ACF1 Theil's U
Test set 50.68296 54.74034 50.68296 19.10289 19.10289 0.8859979  10.20335
[1] "5 year"
                ME     RMSE      MAE      MPE     MAPE      ACF1 Theil's U
Test set 51.50093 55.69166 51.50093 19.40417 19.40417 0.8878321  10.38012
[1] "multi"
                ME     RMSE      MAE      MPE     MAPE      ACF1 Theil's U
Test set 34.72369 37.67395 34.84475 13.09344 13.15178 0.8637144  7.046834
```

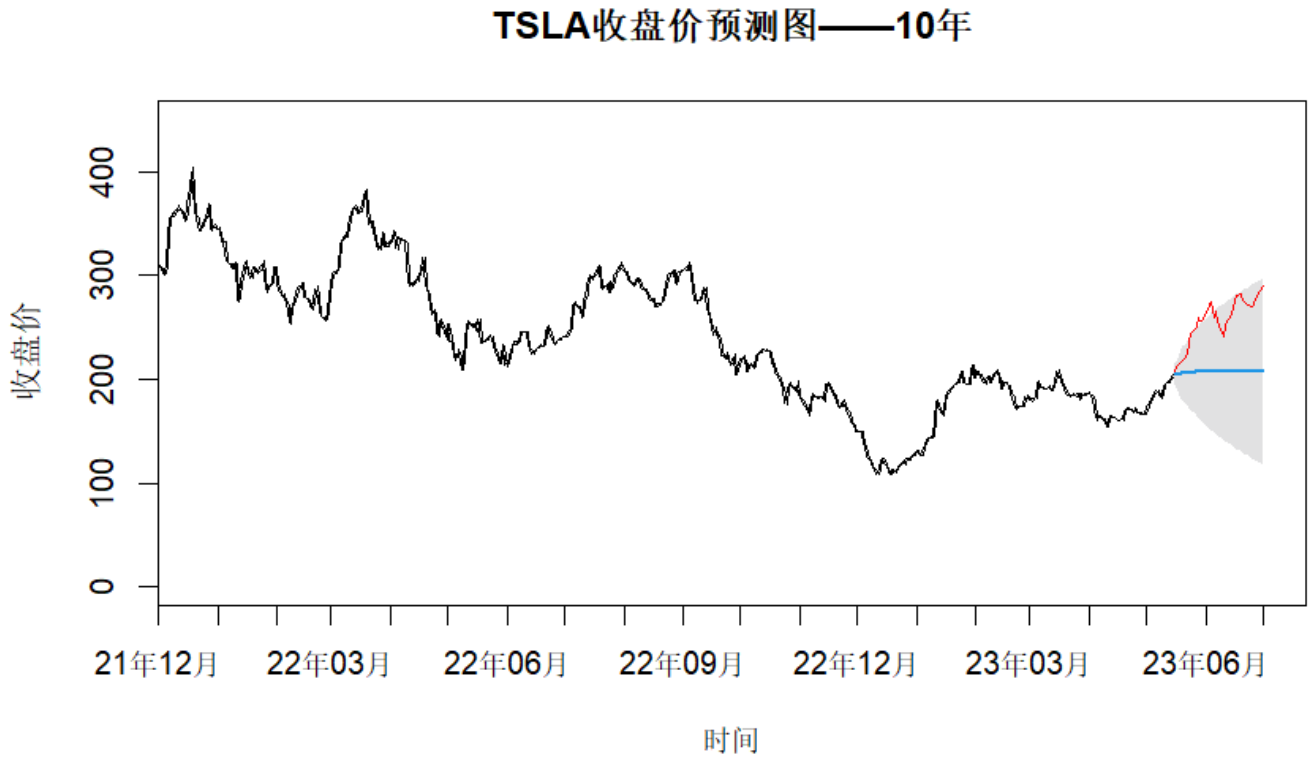
可以看到，多元预测模型有着最好的均方根误差，它的Theil's U指数也比较低，说明多元预测模型有着较好的预测能力。而五年模型和十年模型的性能则非常相近。

（2）作出预测曲线图

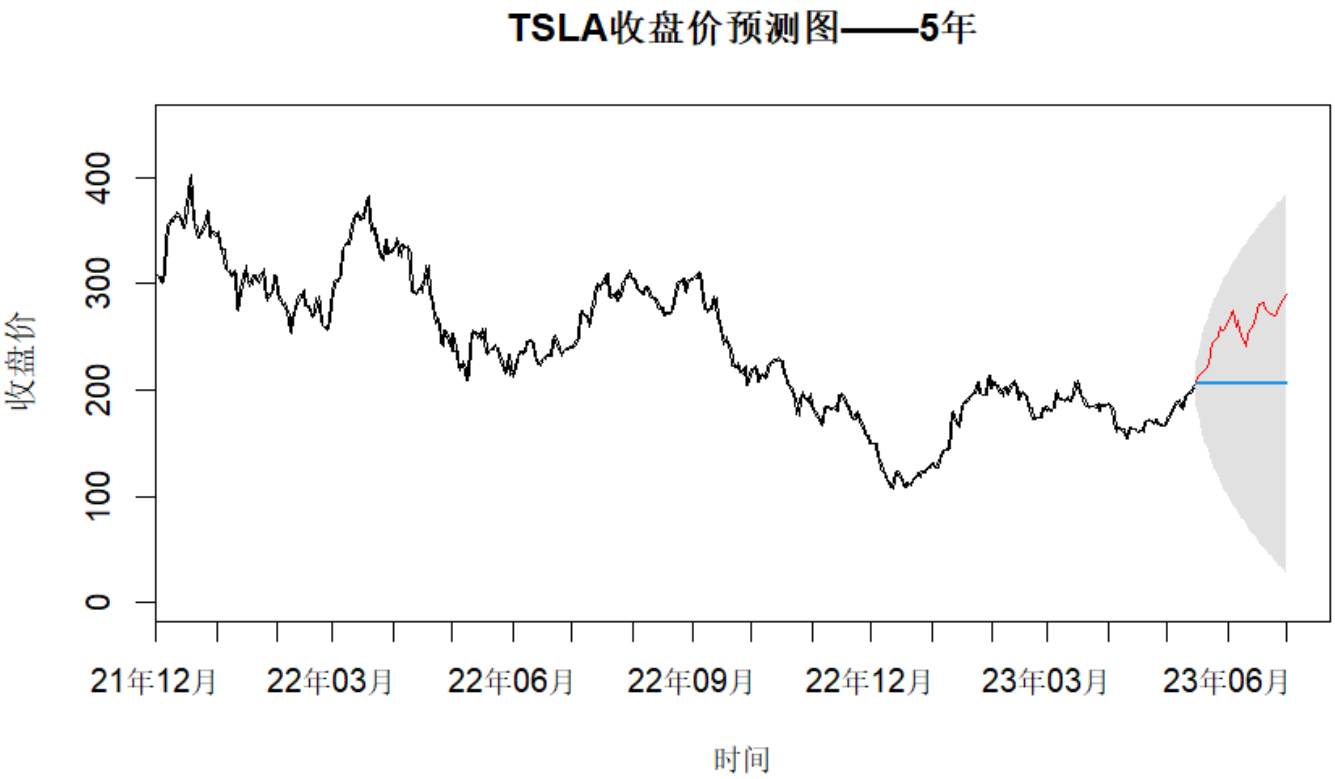
我们对三种预测模型进行预测，并将它们的预测结果与真实值绘制在同一张图上，以便直观地比较它们的表现。

为了控制变量，我们将三种图的横坐标和纵坐标范围设置为相同的范围。

- **十年预测图**

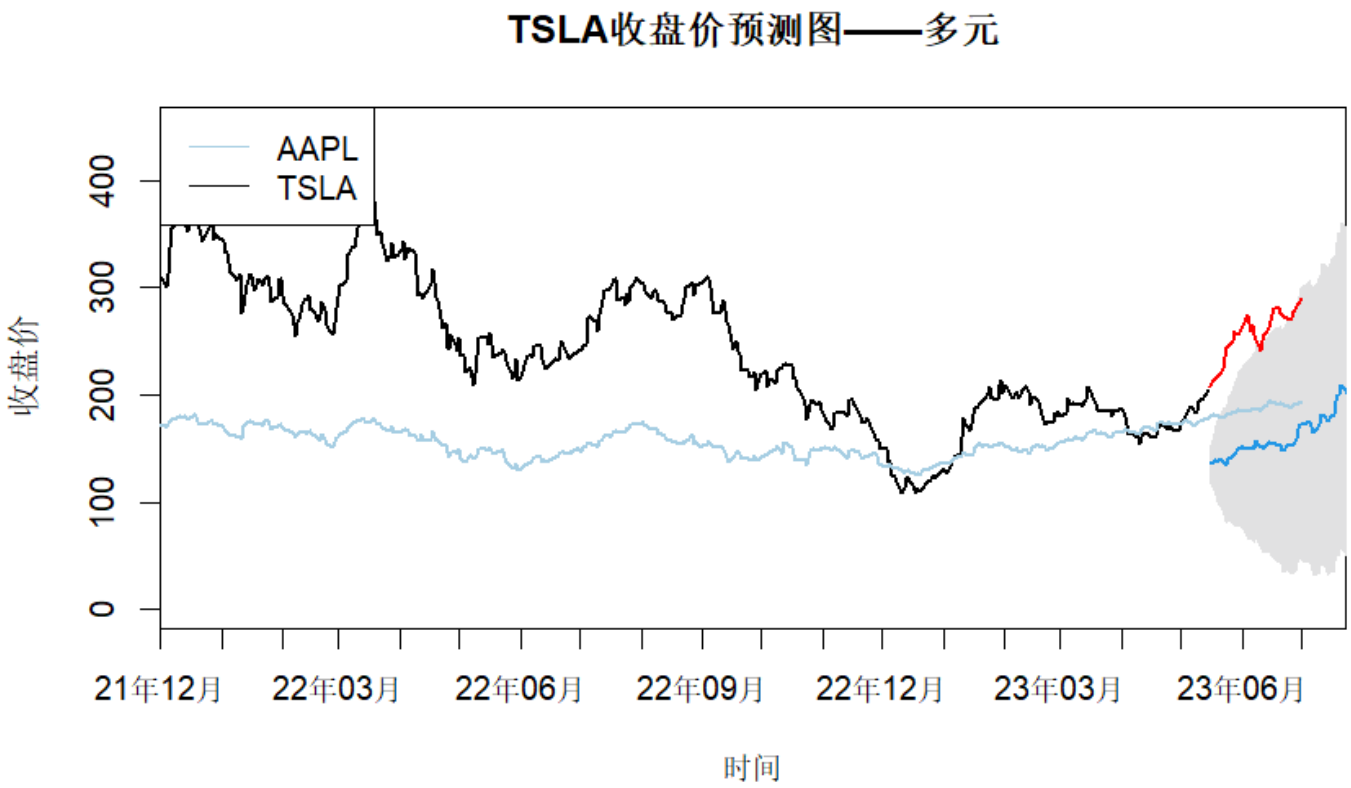


- **五年预测图**



- **多元预测模型的预测图**

由于单一因素下ARIMA建模预测效果欠佳，因此我们考虑建立多元预测模型。预分析中公司收盘价之间相关性的分析反映AAPL和TSLA的相关性最高，故我们将AAPL(公司因素)纳入建模中。



多元预测的模型效果不够好，在检查过后认为应该不是代码有问题，而是模型存在一定的过拟合，所以它虽然在训练数据上表现较好（有着较低的AIC），但是在测试数据上，真实值甚至没有落在预测的置信区间内。

无论是10年模型还是5年模型，预测数据均与真实数据相差很大；但从置信区间上看，股票真实值均落在置信区间内，尤其是5年模型，因此还是可以认为模型效果是不错的。基于十年的数据建立的模型得出的置信区间更小，这意味着该模型更为准确。这也符合我们的认知。在后续进行实际预测时，我们可以尝试使用置信区间的范围而非模型实际给出的值做出相关的分析。

五、挑选潜力股

由前所述，我们对五年和十年两个时间跨度进行了分析，发现除了TSLA公司以外的公司股票数据都是：（1）不差分时，是非平稳的（2）一阶差分后，是白噪声的。

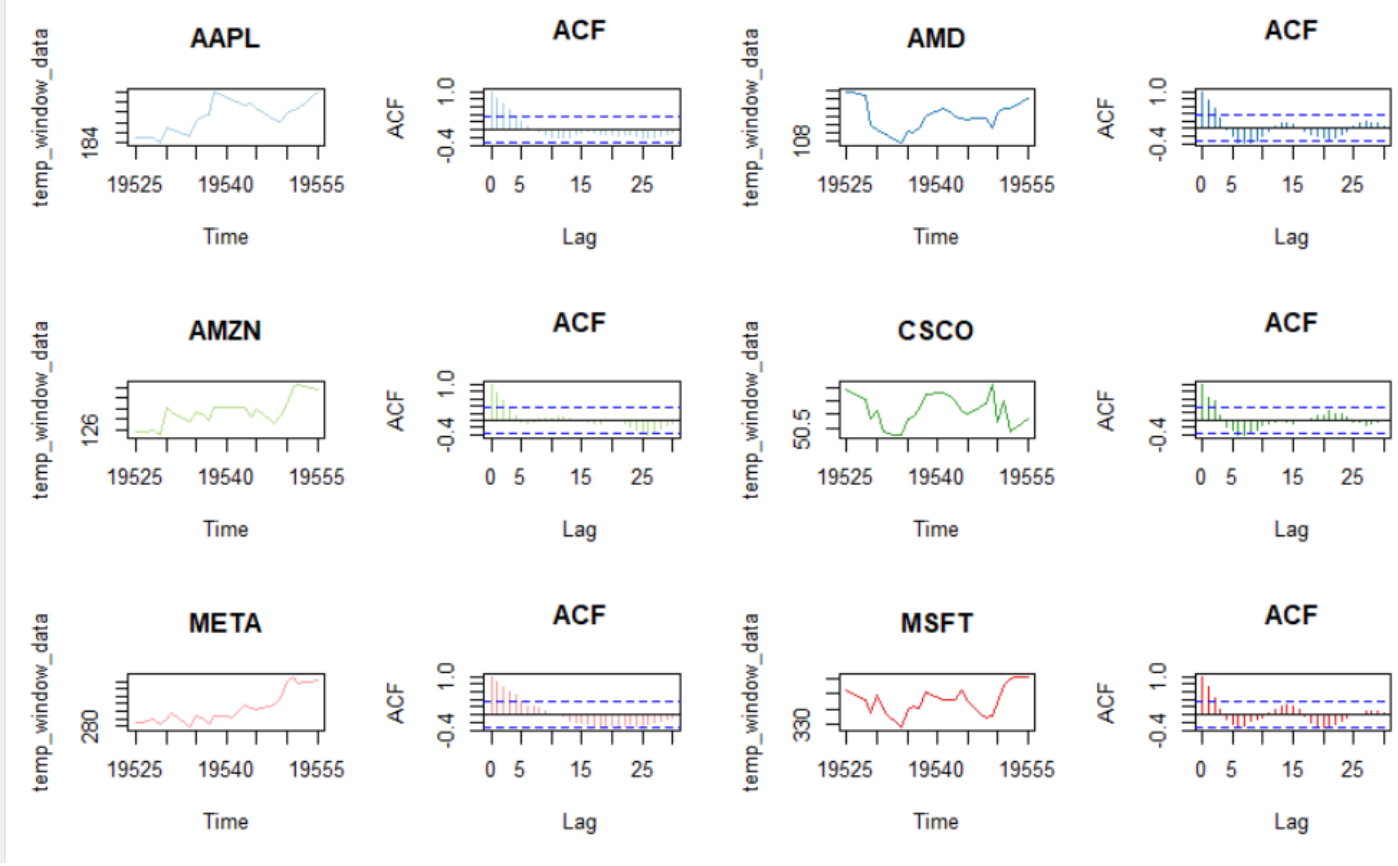
那么，如果把时间跨度取小一些，比如只取出一个月的数据，是否数据会变得平稳、容易预测呢？

在提出这个假设之后，我们进行了下面的观察和分析。

5.1 数据检验

1. **平稳性检验**

我们用tsdiag画出了每个公司最近一个月的时序图和ACF图，发现它们基本都是平稳的。



2. **白噪声检验**

调用Box.test函数，检查每个公司的LB统计量。

```
        Box-Pierce test

data:  ts_list_month[[i]]
X-squared = 53.891, df = 30  p-value = 0.004728

[1] "NFLX 的白噪声检验"

        Box-Pierce test

data:  ts_list_month[[i]]
X-squared = 91.361, df = 30  p-value = 4.078e-08

[1] "QCOM 的白噪声检验"

        Box-Pierce test

data:  ts_list_month[[i]]
X-squared = 73.479, df = 30  p-value = 1.636e-05

[1] "SBUX 的白噪声检验"

        Box-Pierce test

data:  ts_list_month[[i]]
X-squared = 77.814, df = 30  p-value = 4.054e-06
```

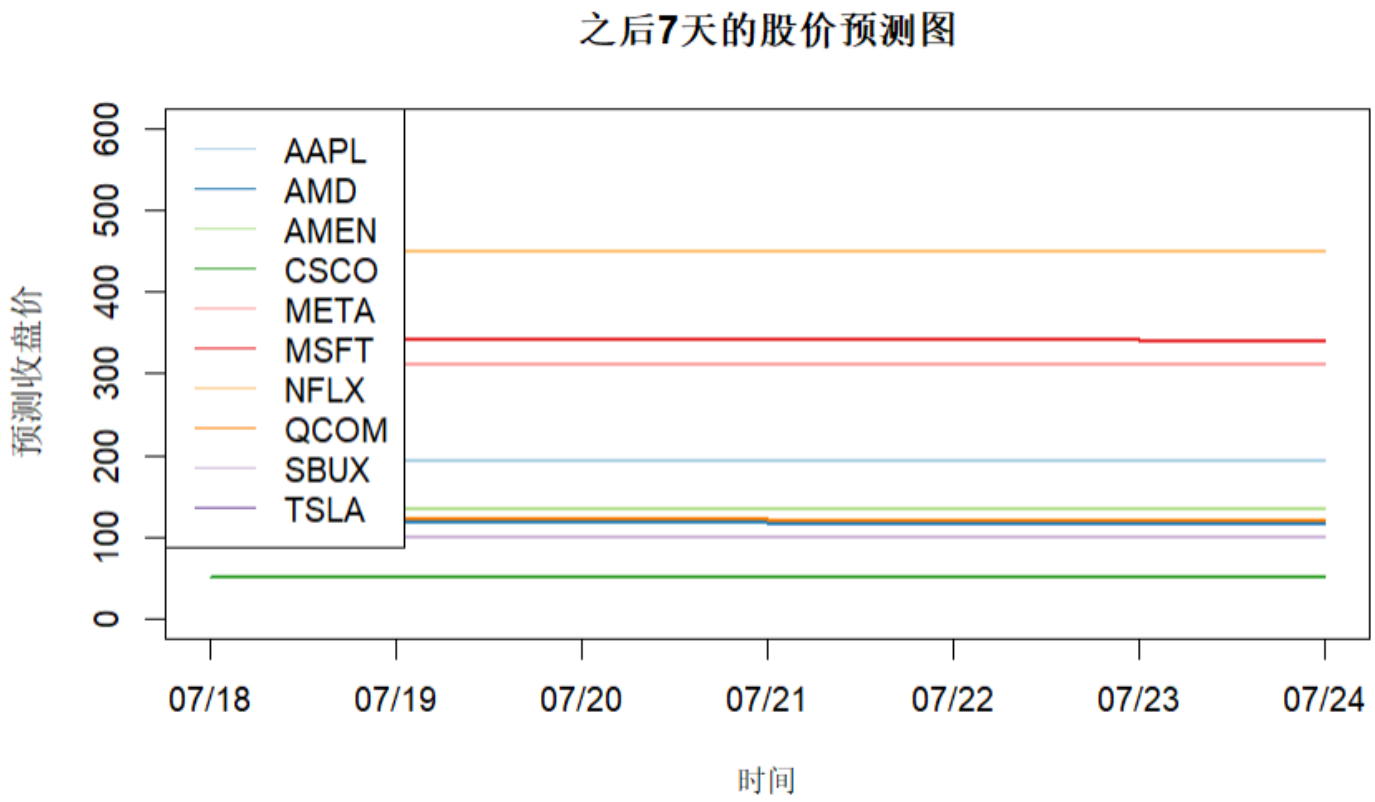
发现每个公司的p值都远小于0.05，说明显著拒绝原假设，说明各公司近一个月的股票数据不是白噪声的。

因此，可以尝试对每个公司使用arima方式建模

5.2 ARIMA预测

我们用auto.arima方法，用每家公司最近一个月的股票数据作为训练集，得到了每家公司的预测模型。

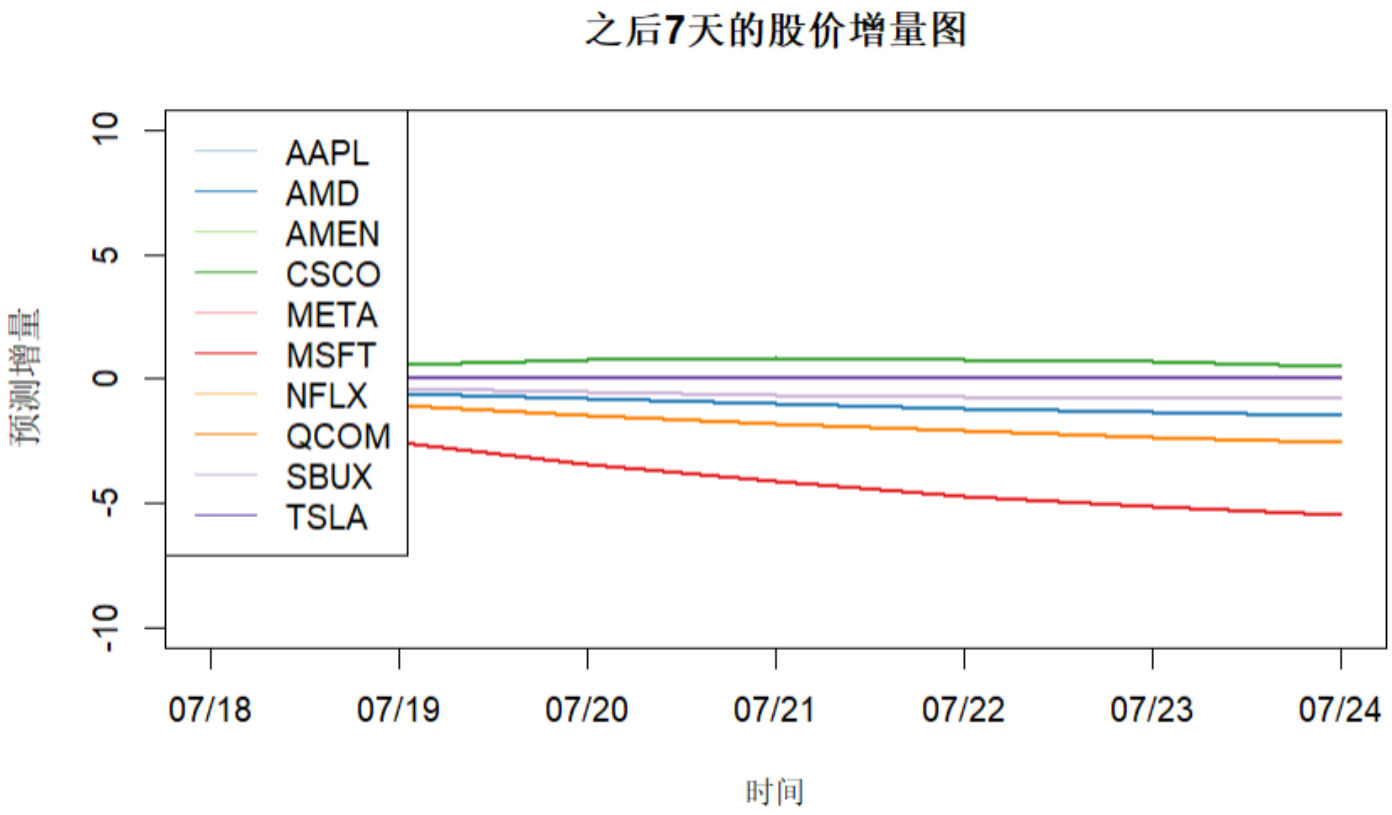
我们用得到的预测模型，预测出了各家公司之后7天的收盘价：



将这些数据减去各家公司在7月17日的股价，即可得到各家公司在之后7天的股价增量。

```
[1] "AAPL 的预测股价增量:  0"
[1] "AMD 的预测股价增量:  -1.47713416941812"
[1] "AMZN 的预测股价增量:  0"
[1] "CSCO 的预测股价增量:  0.49393794565249"
[1] "META 的预测股价增量:  0"
[1] "MSFT 的预测股价增量:  -5.471495994534"
[1] "NFLX 的预测股价增量:  0"
[1] "QCOM 的预测股价增量:  -2.54021447354899"
[1] "SBUX 的预测股价增量:  -0.811559243957134"
[1] "TSLA 的预测股价增量:  0"
```

我们也可以画出相应的预测图。



因此CSCO是用ARIMA模型预测出的最佳潜力股。

5.3 扩展：使用其他模型预测

经过前述分析，我们无法通过ARIMA模型进行预测，从而筛选出潜力股，因此我们改用股票分析中的常用的技术方法，通过分析各家公司股票的移动平均线(MA)、相对移动指数(RSI)、布林带(Bollinger Brands)。其中，由于布林带本身会包含移动平均线数据，故仅对**相对移动平均指数**、**布林带**进行分析。

此处我们在不进行线性插值的情况下取数据集中最近的15天数据，重新建立时序序列。

5.3.1 RSI分析

相对强弱指数（Relative Strength Index，简称RSI）是一种动量振荡器，用于衡量股票价格的最近涨跌动力，以评估股票或其他金融资产是过度买入（overbought）还是过度卖出（oversold）的状况。

RSI值超过70：通常被认为股票或资产处于过度买入状态。这可能意味着价格短期内可能会下跌，一些交易者可能会考虑卖出或减仓。

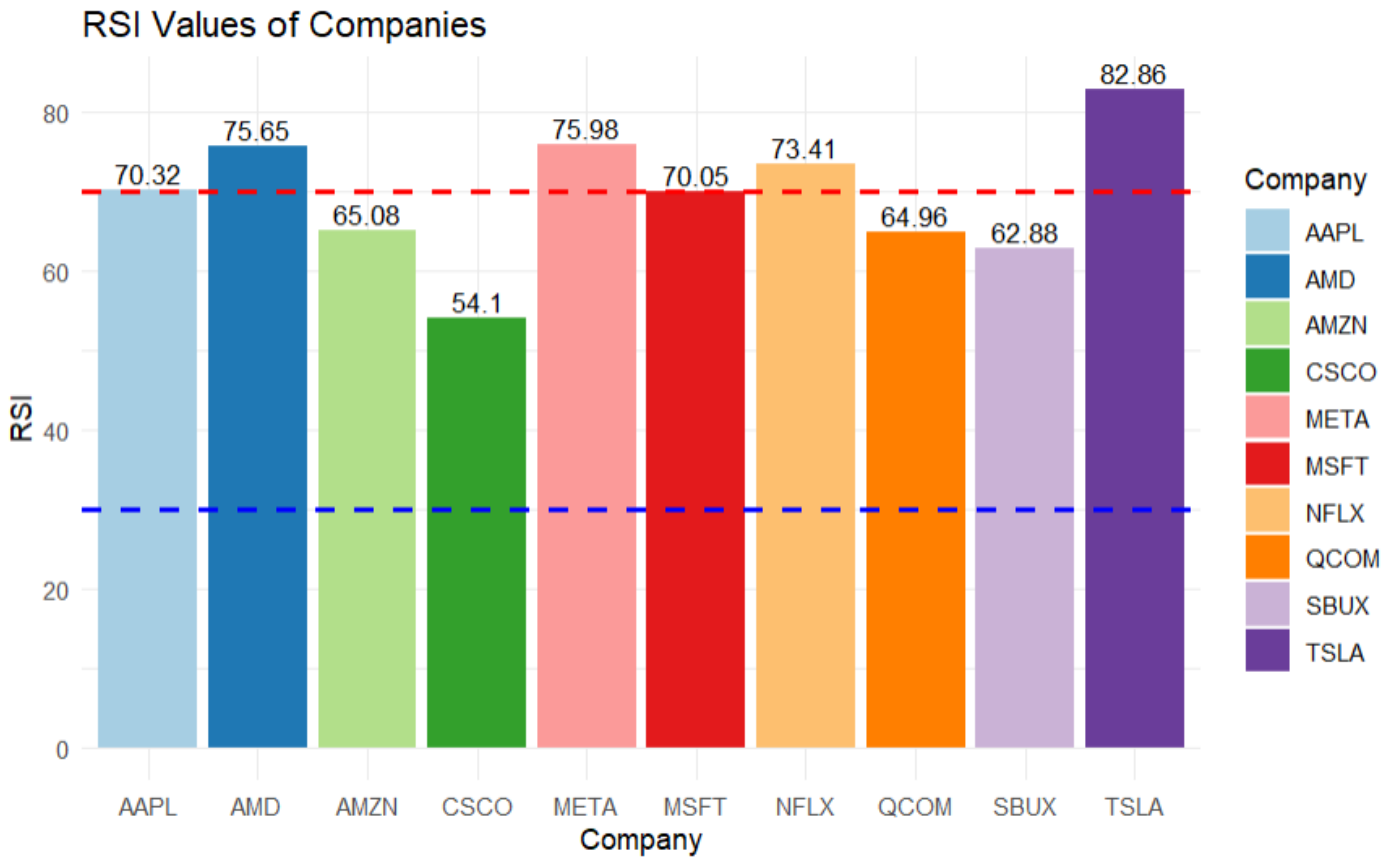
RSI值低于30：通常被认为股票或资产处于过度卖出状态。这可能意味着价格短期内可能会上涨，一些交易者可能会考虑买入或加仓。

1. **数据选取**

我们为每个公司创建一个xts对象，存放的是最近15个交易日的数据（即，不包括非交易日的数据，并且没有使用线性插值方法来填充空值）。

2. **计算RSI值**

接着我们对每个公司计算这十五天的RSI，得到各个公司的平均RSI。可以用下面这幅图直观看出各公司RSI的区别：



3. 结果分析

可以看到，特征最为明显的是TSLA和CSCO，短期内我们可认为TSLA股票非常受市场追捧，NFLX、Meta、AMD、AAPL也很受市场青睐，CSCO、QCOM、SBUX、AMAZN的情况较为稳定。

考虑到当前过度买入的股票可能出现的价格回落，我们认为短期内更有可能盈利的股票为情况更为稳定的股票。

5.3.2 布林带分析

布林带（Bollinger Band，简称BBands）是一种在金融技术分析中常用的工具，主要用于衡量市场的波动性和确定股票价格的相对高低位置。

其组成为：

中线（mavg）：通常是股票价格的简单移动平均线（SMA），用于表示价格的中期趋势。它是布林带的基准线。

上带（up）：表示中线上方的一个标准差范围。它是通过在中线的基础上加上两倍的股票价格的标准差来计算的。上带可以被视为价格的上限或阻力位。

下带（dn）：表示中线下方的一个标准差范围。它是通过在中线的基础上减去两倍的股票价格的标准差来计算的。下带可以被视为价格的下限或支撑位。

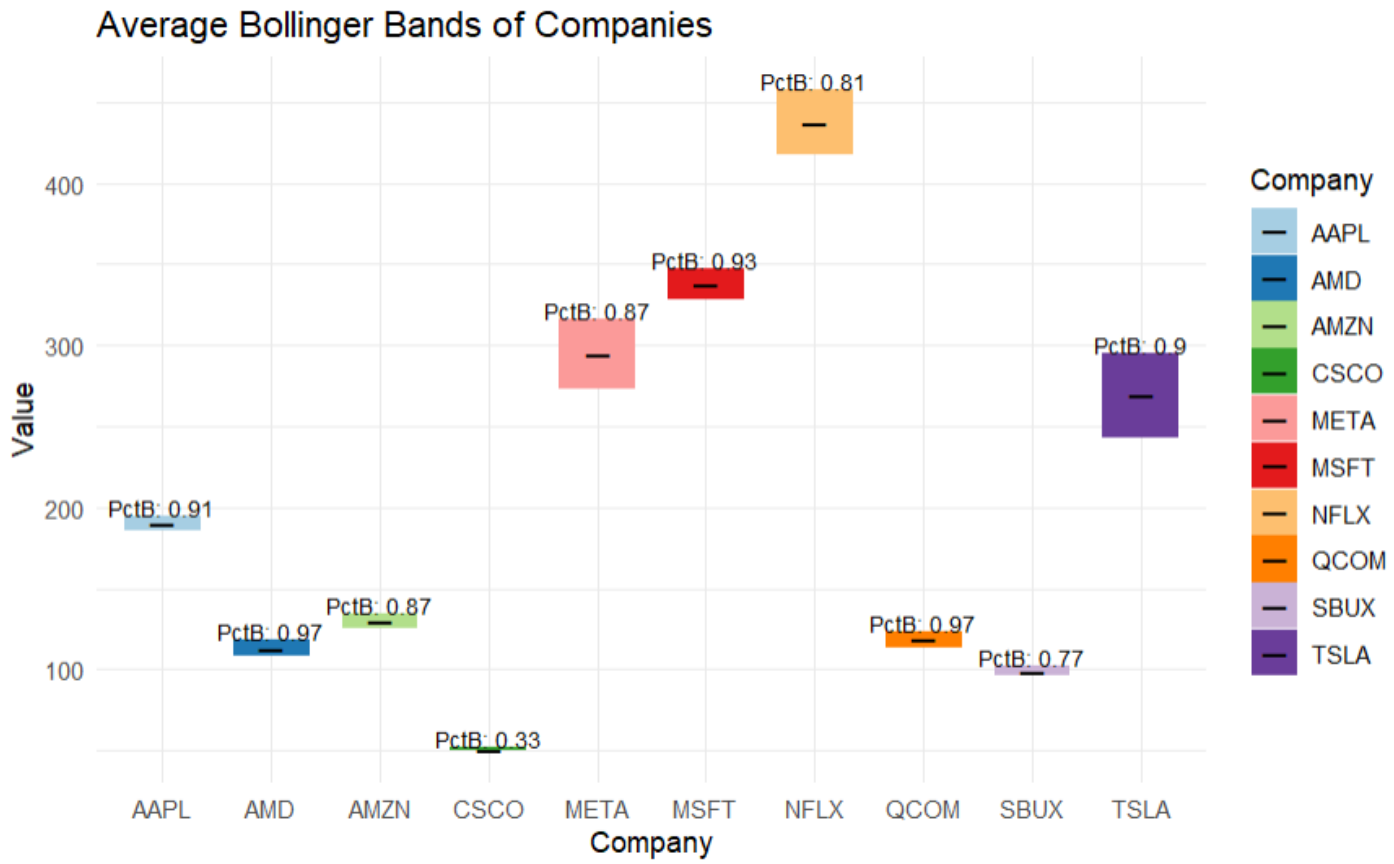
百分比带宽（pctB）：这是一个指标，用于显示价格相对于布林带的位置。其计算公式为：(价格 - 下带) / (上带 - 下带)。pctB的值范围在0到1之间，其中接近0表示价格接近下带，接近1表示价格接近上带。

1. **布林带指标计算**

使用BBands()函数计算布林带指标，并提取上轨、中轨、下轨和PctB的平均值，存进数据框。

2. **图表展示**

我们把前一步计算得到的各公司布林带指标在同一张图上展现了出来。



我们可以观察每支股票上带与下带的差值，从而看出公司的股价波动情况。可以看到，波动较大的几支股票为TSLA、NFLX、META、MSFT，波动较小的几支股票为CSCO、SBUX、AMD、AMAZON、QCOM、AAPL、SBUX。

我们可以通过图上标出的PctB值看出当前哪些公司的股价比较接近下带，一般来说，接近下带说明股价在短期内已经下跌到一个较低水平，可能会有反弹的机会。图中最接近下带的股票是CSCO，说明它未来是比较有潜力的。这也恰好与我们通过ARIMA预测的最佳潜力股是一致的。

两种数据分析时，被过度买入且PctB值较高的股票是当前发展势头非常好的股票，但在不久的将来可能会面临价格的回调；被过度卖出且PctB值较低的股票是当前发展势头没有很好的股票，但在不久的将来可能会面临价格的回升；RSI值在正常区间的股票是发展比较稳定的股票，尽管后续会有价格上的波动，但整体情况还是以稳定为主。

因此，基于这两种分析，在短期内我们认为TSLA、NFLX、META这几支股票的热度很高，但CSCO的潜力最大。

六、分析与建议

从模型的检测结果来看，我们最终建立的ARIMA模型的AIC值较低，残差检验结果非常优秀。这说明针对我们所选择的数据，模型呈现出的效果是非常好的；然而从测试集结果来看，该模型预测值与实际值偏差非常大，模型给出的测试结果并不完全可信。

但我们认为，投资者在物色投资对象时，了解一家公司是否有增长潜力比获得一个看似确切但实际并不确定的“股票价格”更为重要。因此，尽管股价预测的结果是不够准确的，其反映出的趋势却是十分有用的。

因此，我们给出如下的建议：

- 从趋势上看，CSCO公司具有更大的增长潜力，建议投资者着重考虑这家公司。
- 随着科技的不断发展，可以看出各个公司的股票长期呈现增长趋势，因此各位投资者可以多多关注科技市场。
- 根据收盘价来预测股票，虽然能大致看出趋势，但是建立出的模型和实际模型还是有很大出入。因此，单纯通过股票的价格走势来确定潜在投资对象也许不是很明智的选择。技术分析的结果对短期投资更有效，然后若想长期投资，还是应该对公司进行基本面分析，并结合该公司其他方面的表现（如公司领导人决策力、公司氛围登）予以综合考虑。

分工

任务	负责者
回归模型主要分析	陈雨彤
分类模型主要分析	沈加豪
时序模型主要分析	汤佳、叶洁颖
各模型改进	全组成员