**setwd("C:/STONY/Practice/R (No.7)/data\_1")**

**# 设定要用的包，前两个是过滤数据，最后一个是用于随机森林建模的包**

**library(dplyr)**

**library(plyr)**

**library(randomForest)**

**##########################################**

**# Q1&Q2 reputation and customer behavior #**

**##########################################**

**### read the data ###**

**# 读取msom\_seller\_data.csv 这张表**

**seller1 <- read.csv("C:/STONY/Practice/R (No.7)/data\_8/msom\_seller\_data.csv",header=FALSE)**

**# 给读取好的数据提供表头名字，也就是列名**

**colnames(seller) <- c("day","merchant\_id","subcategory\_id","pc\_pv","pc\_uv",**

**"app\_pv","app\_uv","avg\_logistic\_review\_score","avg\_order\_quality\_score",**

**"avg\_service\_quality\_score","if\_cainiao")**

**# 读取msom\_order\_data\_1.csv 这张表**

**odr1 <- read.csv("msom\_order\_data\_1.csv",header=FALSE)**

**# 给读取好的数据提供表头名字，也就是列名**

**colnames(odr1) <- c("day","order\_id","item\_det\_info","pay\_timestamp","buyer\_id",**

**"promise\_speed","if\_cainiao","merchant\_id","Logistics\_review\_score")**

**# 让if\_cainiao这个字段变成 “yes” 和”No” 这两个level**

**seller$cainiao <- ifelse(seller$if\_cainiao==0,"No","Yes")**

**seller$cainiao <- factor(seller$cainiao,levels=c("No","Yes"))**

**######## A. how rep system affects the purchase########**

**### subset the data ###**

**# 根据商家id以及天数来分类，算出每个商家每天的订单量**

**odr1\_num <- ddply(odr1,.(merchant\_id,day),summarize,odr\_number=length(order\_id))**

**#dim(odr1\_num)**

**# 根据商家id以及天数来分类，算出每个商家每天的平均的物流，服务，质量评分**

**seller\_na <- na.omit(seller)**

**seller1\_rep <- ddply(seller\_na,.(merchant\_id,day),summarize,**

**log\_score=sum(avg\_logistic\_review\_score)/length(avg\_logistic\_review\_score),**

**qual\_score=sum(avg\_order\_quality\_score)/length(avg\_order\_quality\_score),**

**serv\_score=sum(avg\_service\_quality\_score)/length(avg\_service\_quality\_score)**

**)**

**# 将以上两张表合并在一起，得出每个商家每天的订单量和物流，服务，质量评估分数**

**### merge the data into one table ###**

**cutomer\_behiv <- merge(odr1\_num, seller1\_rep, all=TRUE, sort=TRUE)**

**cutomer\_behiv <- na.omit(cutomer\_behiv)**

**# 算出每个商家一个月内的订单量和物流，服务，质量评估分数**

**### calculate the odr number of every merchant in one month###**

**customer\_behiv <- ddply(cutomer\_behiv,.(merchant\_id),summarize,**

**odr\_num=sum(odr\_number),**

**log\_score=sum(log\_score)/length(log\_score),**

**qual\_score=sum(qual\_score)/length(qual\_score),**

**serv\_score=sum(serv\_score)/length(serv\_score)**

**)**

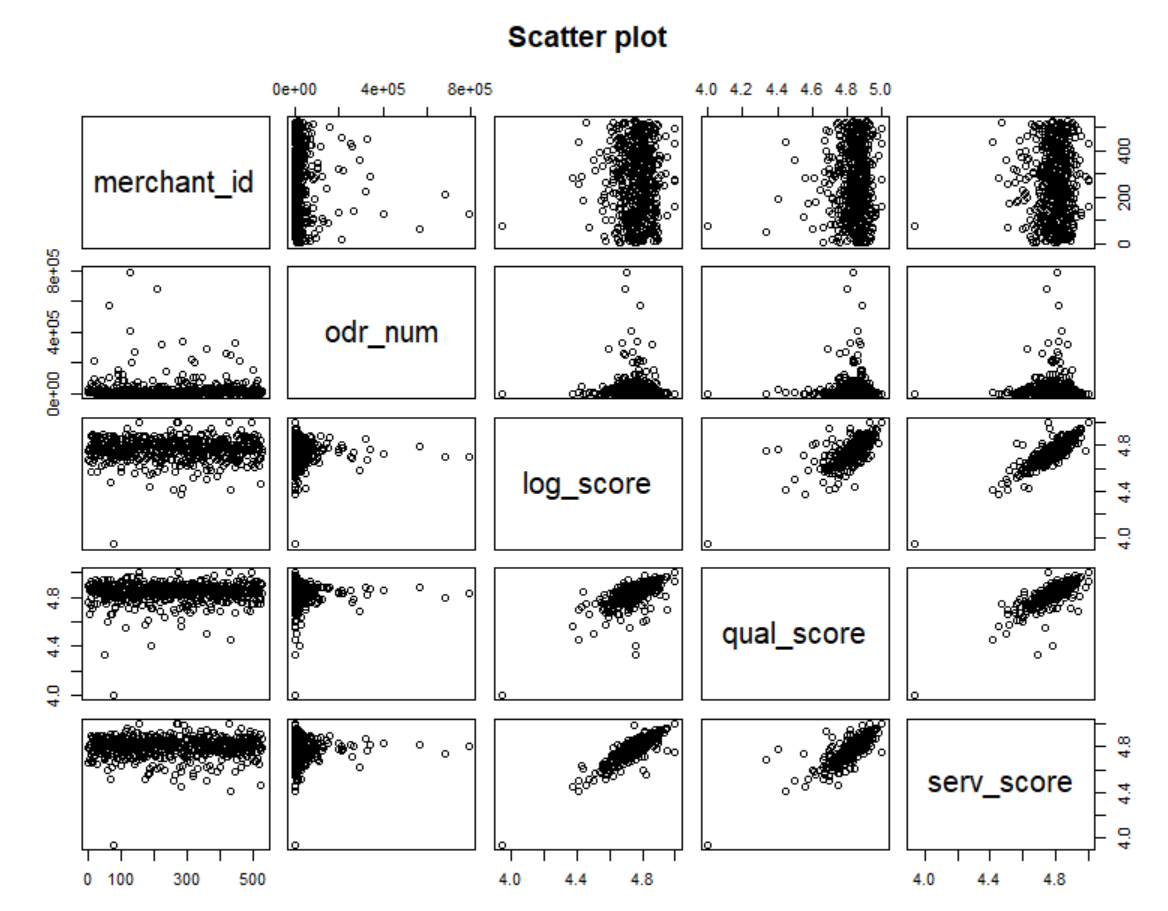
**# 建立线性回归模型，查看三种评分标准对于用户下单量的影响**

**### the relationship between rep\_score and purchase ###**

**# 先查看这些变量间的一个相关性，用scatter plot**

**par(mfrow=c(2,1))**

**plot(customer\_behiv, main="Scatter plot")**

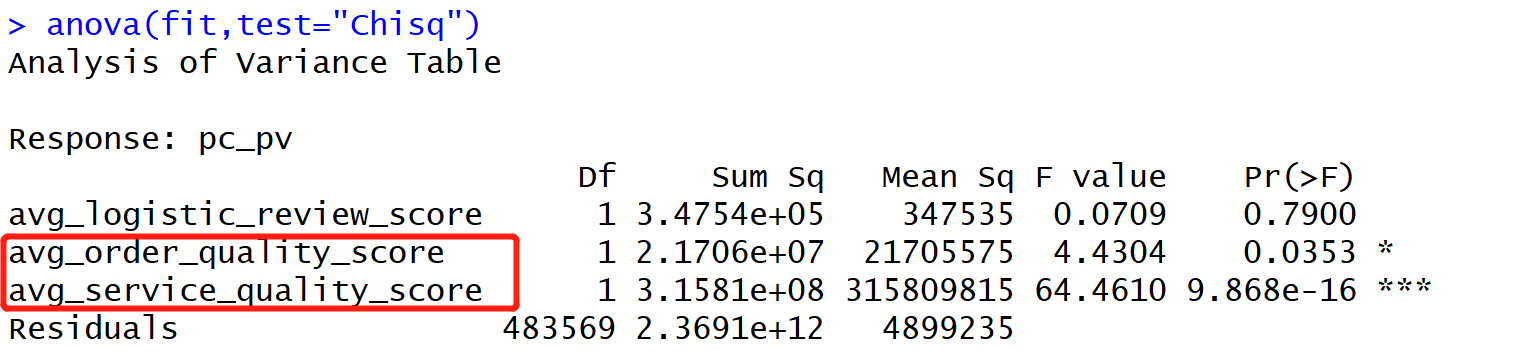
****

**model\_odr <- {odr\_num~log\_score+qual\_score+serv\_score}**

**fit <- lm(model\_odr,data=customer\_behiv)**

**summary(fit)**

**anova(fit,test="Chisq")**

****

**step(fit,direction="both")**

**# 利用随机森林进一步确定三种评分标准对于用户下单量的影响（可看出影响排序）**

**### important preditors ###**

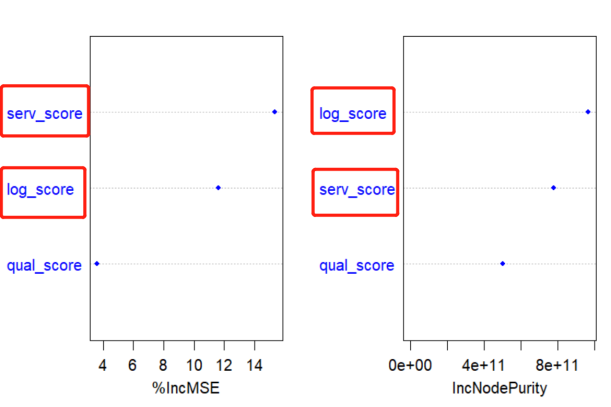
**set.seed(1234)**

**# 以下是随机森林公式**

**rf\_odr <- randomForest(model\_odr,data=customer\_behiv,mtry=3,ntree=1000,importance=TRUE)**

**# 以下是画出每个评分标准的重要程度，数值越大代表影响越大**

**varImpPlot(rf\_odr,color="blue",pch=20,cex=1.25,main="")**

****

**# 建立线性回归模型，查看三种评分标准对于用户在网页和在APP内浏览商品的行为影响**

**######## B. some indexes that affect the purchase########**

**# build the linear model**

**seller\_na <- na.omit(seller)**

**# 以下pc\_pv/uv 代表网页浏览次数和人数，APP\_pv/uv代表APP内浏览次数和人数**

**# 建立线性回归模型，查看每种浏览行为中哪些评分标准对其影响较大**

**# PC\_PV**

**model\_pcpv <- {pc\_pv~avg\_logistic\_review\_score+avg\_order\_quality\_score+avg\_service\_quality\_score+pc\_uv+**

**app\_uv+app\_pv+if\_cainiao}**

**fit <- lm(model\_pcpv,data=seller\_na)**

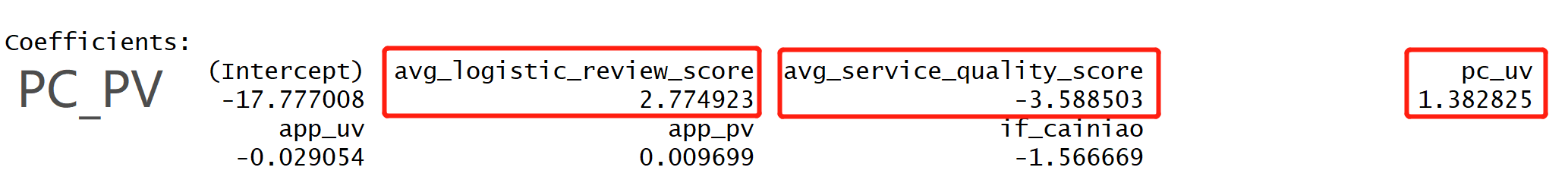
**summary(fit)**

**# 画出anova table, \*越多代表影响越大**

**anova(fit,test="Chisq")**

**# 以下是逐步回归stepwise的公式，将model进行前向以及后向迭代，得出最重要的变量以及变量的coeffients**

**step(fit,direction="both")**

****

**# PC\_uV**

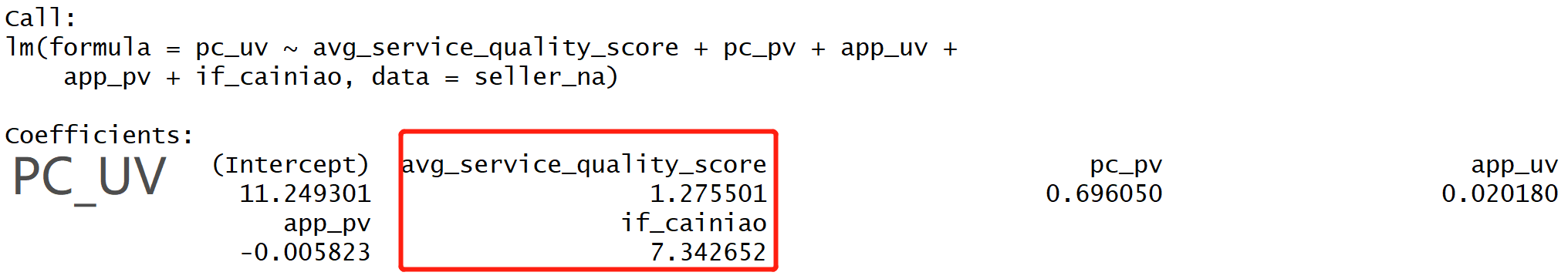
**model\_pcuv <- {pc\_uv~avg\_logistic\_review\_score+avg\_order\_quality\_score+avg\_service\_quality\_score+pc\_pv+**

**app\_uv+app\_pv+if\_cainiao}**

**fit <- lm(model\_pcuv,data=seller\_na)**

**anova(fit,test="Chisq")**

**step(fit,direction="both")**

****

**# APP\_PV**

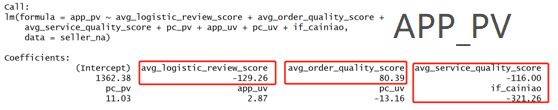
**model\_apppv <- {app\_pv~avg\_logistic\_review\_score+avg\_order\_quality\_score+avg\_service\_quality\_score+pc\_pv+**

**app\_uv+pc\_uv+if\_cainiao}**

**fit <- lm(model\_apppv ,data=seller\_na)**

**anova(fit,test="Chisq")**

**step(fit,direction="both")**

****

**# APP\_UV**

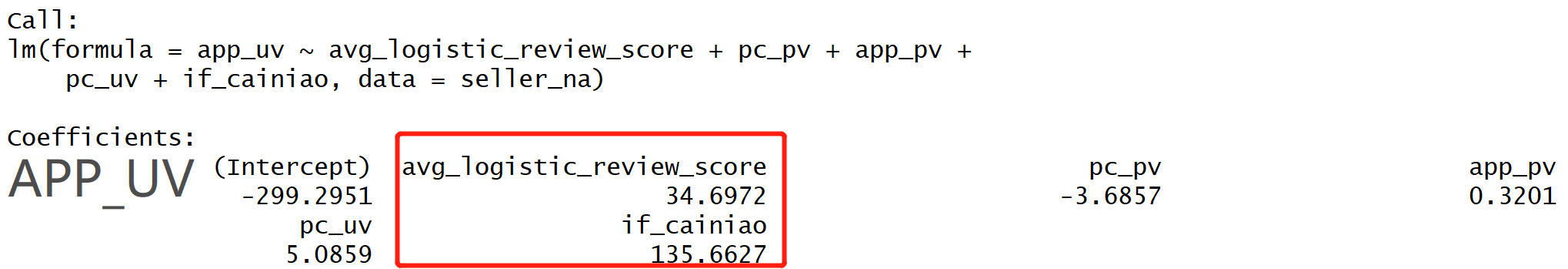
**model\_appuv <- {app\_uv~avg\_logistic\_review\_score+avg\_order\_quality\_score+avg\_service\_quality\_score+pc\_pv+**

**app\_pv+pc\_uv+if\_cainiao}**

**fit <- lm(model\_appuv ,data=seller\_na)**

**anova(fit,test="Chisq")**

**step(fit,direction="both")**

****

**# 查看使用以及不使用菜鸟是否会对商家带来订单数量以及评分上的影响**

**############################################**

**# Q3&4 analyze Cainiao services' influence #**

**############################################**

**# 选取**

**### the informatio about using cainiao or not**

**# 选取和不使用菜鸟的商家一个月内总的订单量**

**odr1\_sum <- ddply(odr1,.(merchant\_id,if\_cainiao),summarize,odr\_sum=length(order\_id))**

**# 选取和不使用菜鸟的商家一个月内总的评分**

**seller1\_cai <- ddply(seller\_na,.(merchant\_id,if\_cainiao),summarize,**

**log\_score=sum(avg\_logistic\_review\_score)/length(avg\_logistic\_review\_score),**

**qual\_score=sum(avg\_order\_quality\_score)/length(avg\_order\_quality\_score),**

**serv\_score=sum(avg\_service\_quality\_score)/length(avg\_service\_quality\_score)**

**)**

**cai\_evaluate<- merge(odr1\_sum, seller1\_cai,by=c("merchant\_id","if\_cainiao"))**

**cai\_scr\_odr <- ddply(cai\_evaluate,.(if\_cainiao),summarize,odr=sum(odr\_sum),**

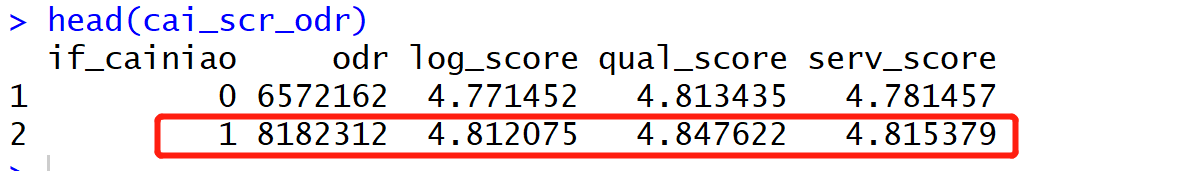
**log\_score=sum(log\_score)/length(log\_score),**

**qual\_score=sum(qual\_score)/length(qual\_score),**

**serv\_score=sum(serv\_score)/length(serv\_score)**

**)**

**head(cai\_scr\_odr)**

****

**####以上可以看出，使用了菜鸟包裹的商家的订单量和评分普遍比其他的高**

**### the percentage of using cainiao or not ###**

**# 选取和不使用菜鸟的全部商家一个月内总的订单量**

**odr1\_cai <- ddply(odr1,.(if\_cainiao),summarize,odr\_number=length(order\_id))**

**# 算出使用和不使用菜鸟的全部商家的市场份额百分比**

**cai\_percent <- ddply(odr1\_cai,**

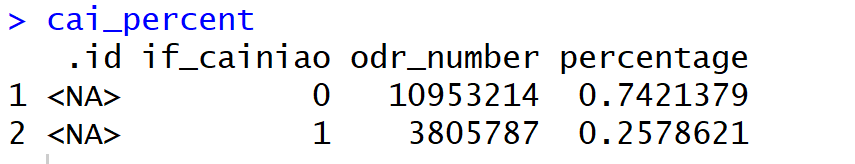
**.(),**

**.fun=function(x){**

**transform(x, percentage=with(x,ave(odr\_number,if\_cainiao)/sum(odr\_number)))**

**})**

**cai\_percent**

****

**####以上可以看出，菜鸟包裹的市场份额并不算很大，所以其对于商业订单量以及评分的提升效果，是可以纳入阿里巴巴后期和菜鸟包裹战略合作的重要因素**

**### build a regression reputation austen&customer behaviour & cainiao service influence**

**model\_3 <- {odr\_sum ~ if\_cainiao+log\_score+qual\_score+serv\_score}**

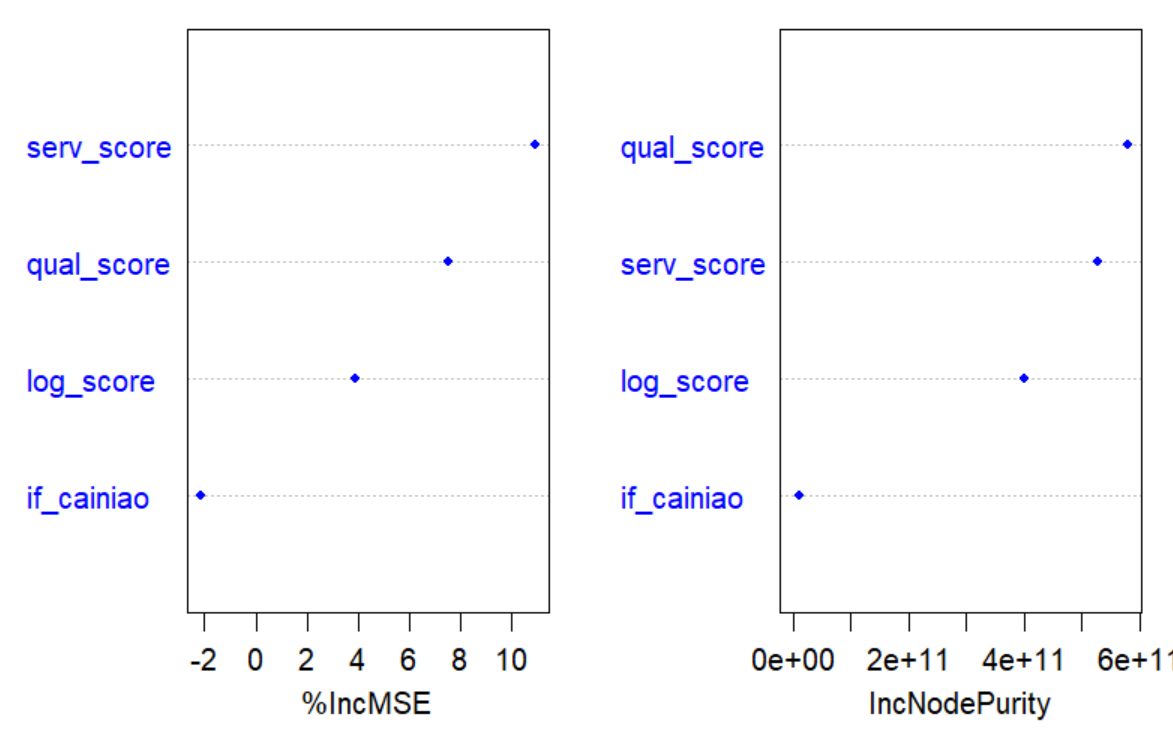
**fit\_3 <- lm(model\_3,data = cai\_evaluate)**

**### important preditors ###**

**set.seed(1234)**

**fac3 <- randomForest(model\_3,data=cai\_evaluate,mtry=3,ntree=1000,importance=TRUE)**

**varImpPlot(fac3,color="blue",pch=20,cex=1.25,main="")**

****

**### 可以看出荣誉系统对于订单量的影响更大，菜鸟的服务没有直接影响。**

**############################################**

**# Q5 analyze Cainiao accurate demand #**

**############################################**

**### 读取库房数据**

**inventory <- read.csv("C:/STONY/Practice/R (No.7)/data\_8/msom\_inventory\_data.csv",header=FALSE)**

**colnames(inventory) <- c("day","item\_id","warehouse\_id","city\_id","total\_begin\_qty",**

**"total\_end\_qty","replen\_in\_qty","transfer\_in\_qty","sale\_out\_qty",**

**"transfer\_out\_qty")**

**inventory <- na.omit(inventory)**

**library(plyr)**

**### 计算每天每个库房的货物需求量**

**## we can calculate the product demand of every city for warehouse**

**city\_qty <- ddply(inventory,.(city\_id),summarize,city\_qty=sum(sale\_out\_qty))**

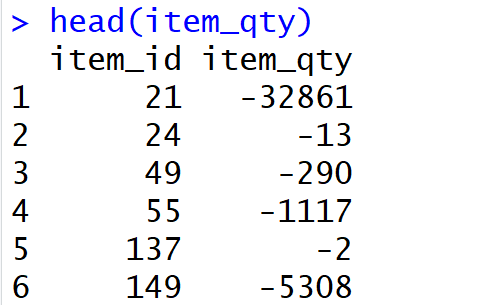
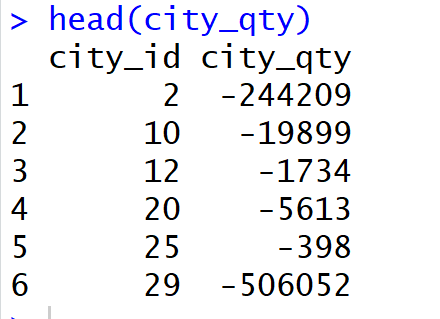
**head(city\_qty)**

**### 计算每天每个商品的货物需求量**

**## we can calculate the product demand of items for warehouse**

**item\_qty <- ddply(inventory,.(item\_id),summarize,item\_qty=sum(sale\_out\_qty))**

**head(item\_qty)**

****

**############################################**

**# Q7 Optimize the distirubution network #**

**############################################**

**### 计算每个货物在每座城市的货物需求量**

**## we can calculate the demands of every item in every city**

**itm\_city\_qty <- ddply(inventory,.(city\_id,item\_id),summarize,itm\_city\_qty=sum(sale\_out\_qty))**

**head(itm\_city\_qty)**

**### 计算每个订单在每座城市的快递所用市场，由于此数据为jason格式，需要解码并且文件庞大，个人认为也并不需要计算，第七八问只要是提供一个实践思路。**

**## we can figure out the time spent on the deliver of every item**

**logistic2 <- read.table("C:/STONY/Practice/R (No.7)/data\_2/msom\_logistic\_detail\_2.csv",nrows = 10000)**

**logis <- unlist(logistic2[,1])**

**logis <- data.frame(logis)**

**colnames(logistic2) <- c("order\_id","action","city\_id","timestamp")**

**### 这是后面几问的中文总结**

Sellers and buyers interact through the Alibaba and Cainiao platforms. Are there synergies that could be exploited between the two? How can Alibaba use Cainiao effectively to enhance its value proposition and its profits?

# 菜鸟包裹的好评率和订单都多，鼓励商家使用菜鸟包裹，可以使用减少阿里的抽成，从而提升商品的服务质量。2. 销量好的优先推荐使用菜鸟包裹 3. 加权平均的方法求出该商店的评分

Can Cainiao’s platform be used to develop more accurate demand forecasts?

# 可以的，我们拿到每个仓库每天卖出的货物数量，以及每个仓库每天所需的商品的数量，我们就可以通过今天的需求预测明天的需求，然后在明天开始前我们应该通过什么途径补货，调货，调货的，补货掉货的话，哪一个城市最快。

What is the optimal inventory allocation across products and warehouses?

# 根据货物在每个仓库的需求量来按比例分发，补货所需货物。然后在城市中间设置几个便捷的中转站，但凡出现货物售罄就从中转站及时补货和调货。而不同的商品会根据每个城市需求量的不同而按比例选择核心仓库，中转站和小型仓库。这样就能按货物需求来调整仓库和货物之间的流动性。

How to optimize the distribution network to achieve the best trade-off between cost efficiency and fast delivery?

# 首先，建立一个货物物流分发网络是一个具有可持续性发展的策略，所以我们需要每个仓库每种货物近几年的总销售量，下单城市的比重排名，按照就近原则，我们可以做一个聚类分析，根据货物出仓到到达目的地的时间， 每个仓库就近城市的货物需求量，每个仓库总的各类货物需求量。

How to monitor and incentivize third-party logistics firms to provide speedy service?

1. 时刻监控订单信息，如果超过承诺的发货以及到货日期，则给商家发送warning email,同时给顾客发送warning email。
2. 根据物流的速度进行true performance 打分，倘若在所有订单量中延期发货或到货的数量超过一定比例，将被扣分，分数可为1，2，3，4，5. 5为最高。 这一打分将计入商铺在首页的曝光及推荐系统中的考量，分数越高，被推荐的机会越大。促使效率低的商家选择优质的物流服务。