Report

1. **Executive Summary**

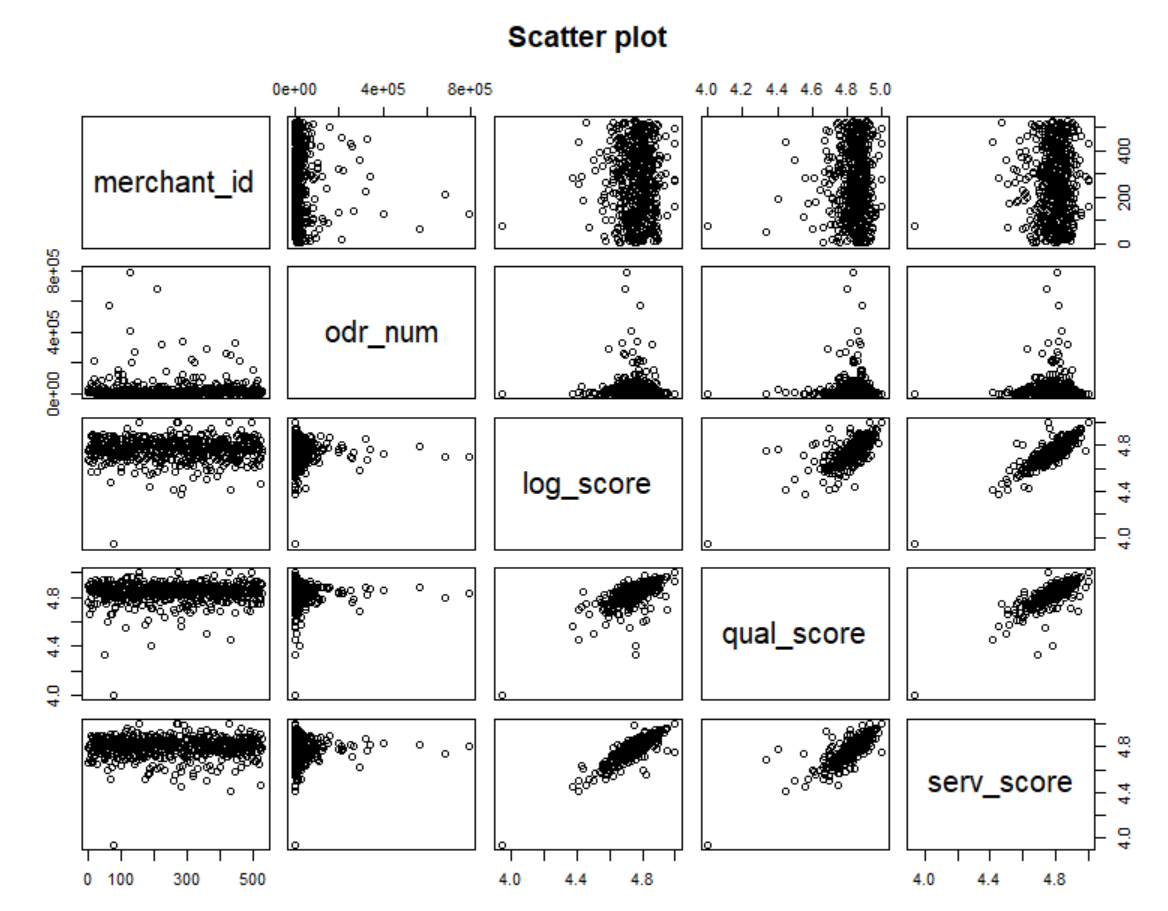
What does a deliver company care about most? Where do deliver company’s most fundamental interests come from? The answer, no doubt, is the customer. In this project, to analyze the relationship between the company and customers, and how to use this relationship to enhance the interests of the business, we used the data from Cainiao Network as a research. We have conducted statistical modeling to analyze and extract relevant information that has an impact on customer behaviors.

1. **Basic Description of Statistical Methodology**

In this project, because finding the relationship of the two is a statistical problem, we used linear regression model to eliminate relationships between factors(e.g., serve score, logistic score) and dependent variable(e.g., the order number). After getting the model, we need to use step() function to determine the formula. The stepwise function determined the variables that have the most significant predictive ability. In addition, we also use random forest method to ensure the accuracy of finding the important predictors.

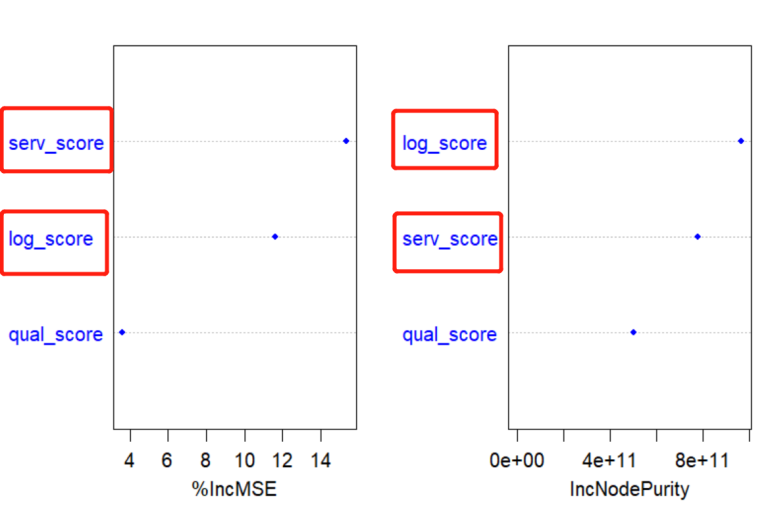
1. **Analysis of Data Findings**
2. **Data Clean**

Once we get the data, we are supposed to be clear about what the data look like and what the characteristics of the data are, which are critical for subsequent analysis. There are five tables included in our data set. In order to get the important information we want, we used ddply() and merge() function in R to achieve that. After that, we can check the scatter plot for the data we used. (*fig is below*)

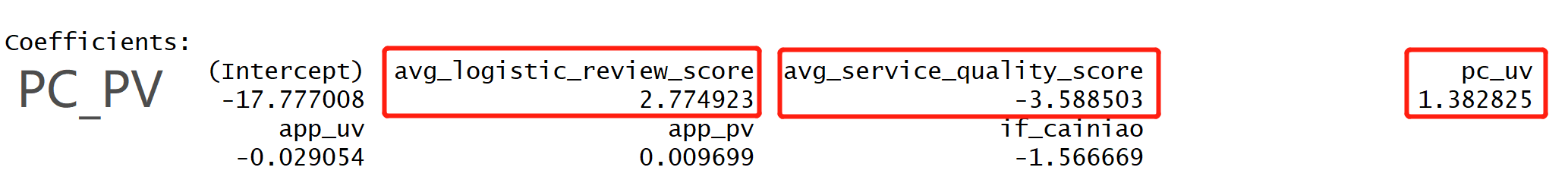


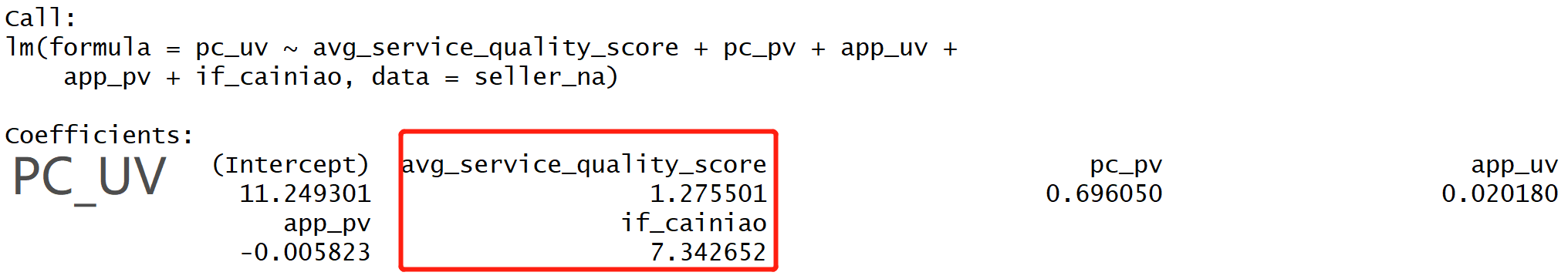
1. **Reputation System and Customer Behaviors**

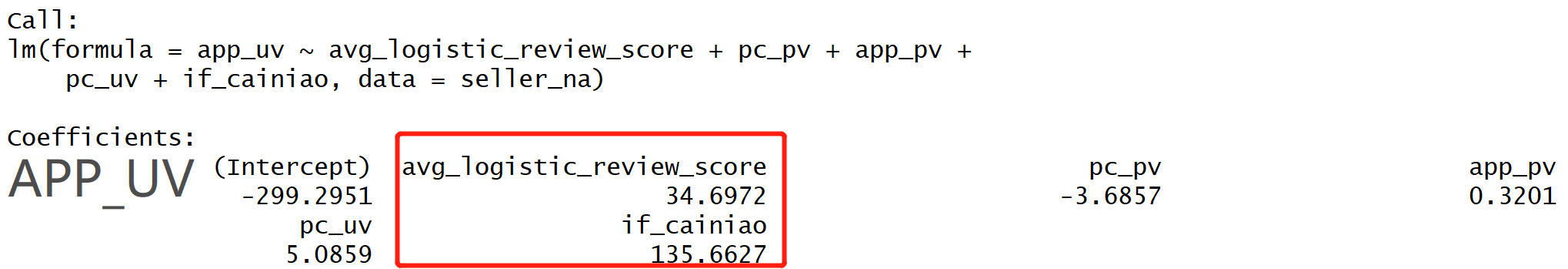
We put all the variables related to reputation system and purchase behaviors into the linear regression model. Then we use step() function and find out that the best model is involved with all of the variables. However, when it comes to the random forest, the result of varImpPlot() function shows that “serve\_score” and “logistic\_score” are more important among the reputation system. (*fig is below*)



In addition, we also defined the factors that have indirect impacts on the behavior of users ordering, like “pc\_pv”,“app\_uv” and etc. The result shows that for each indicator, the reputation system has a different impact on them. (*fig is below*)



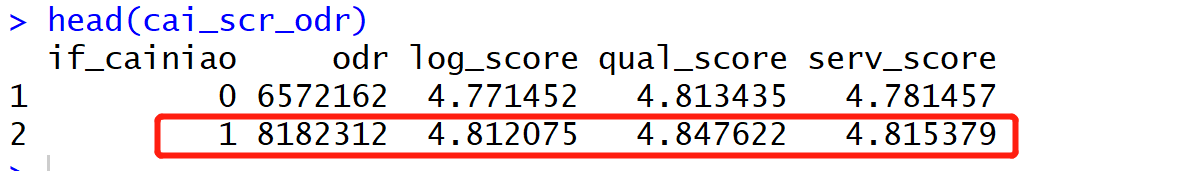


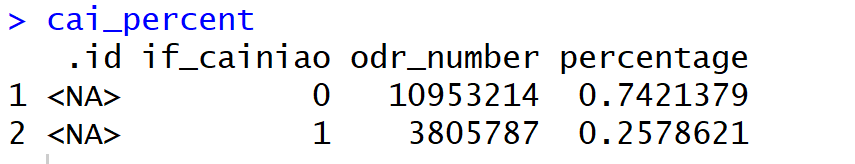


What we can see from the results above is that the coefficients of the reputation index are not the same or even similar. So we can say that behavioral biases do present in the reputation system. What we can do is that we can redesign the system to use the weighted average method to calculate the score of the merchants so as to put the bias into consideration.

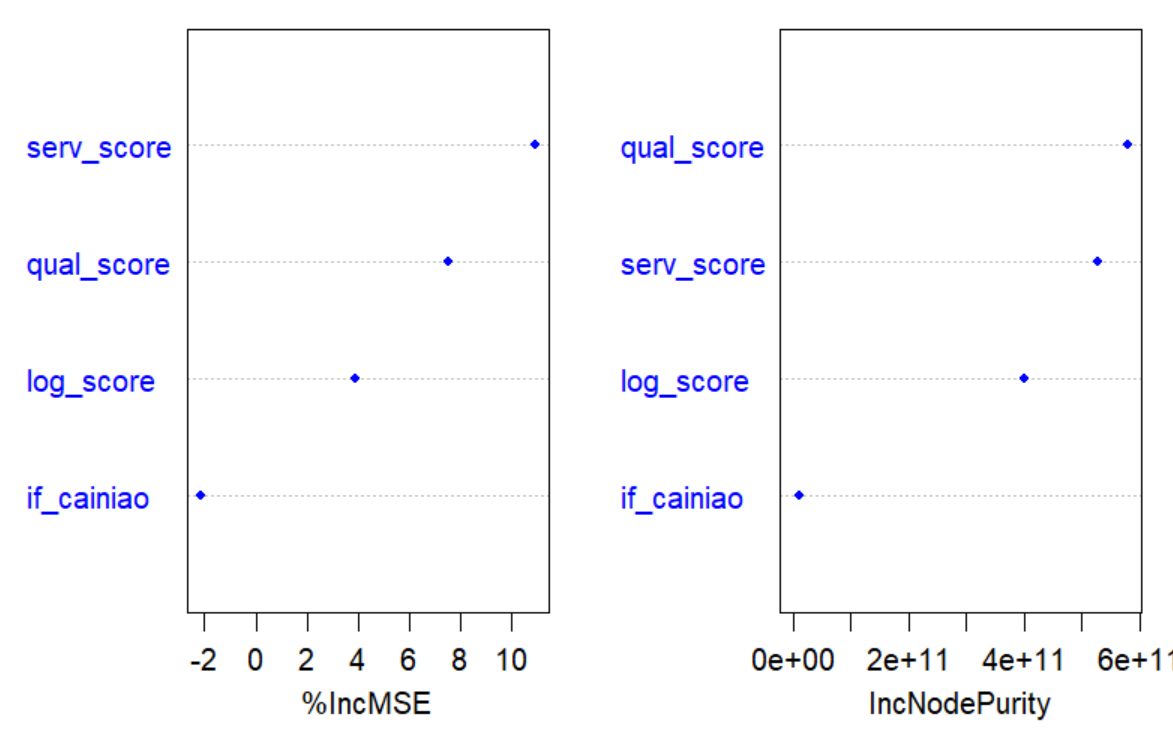
1. **Analyze Cainiao Services' Influence**

Researching only customer behaviors is not enough to build a program that upgrades its profits. To determine what other factors can make the company run better. We need to do more attempts. In this part, we divide the data into groups by the information about using cainiao or not. We try to figure out the market share of Cainiao and whether the companies that use Cainiao have more orders. The result proves that even though the market share of Cainiao is not very large, its service and logistic performance outperform than others. This is the economic value that it wraps in the market. (*fig is below*)





In addition, we also conducted a linear regression which regresses customer behavior on reputation system and Cainiao service. From the results we get, we can tell Cainiao has least impact on the order number, and reputation system has more important effects on it. (*fig is below*) So we can conclude that Cainiao service may have impacts on the reputation system, but it may not have large impact on order behavior directly.

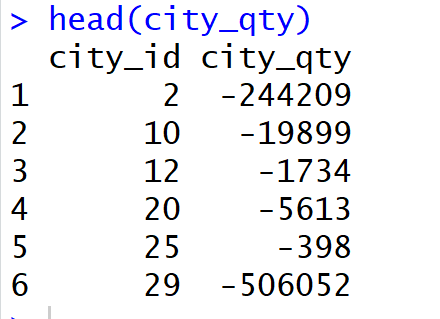
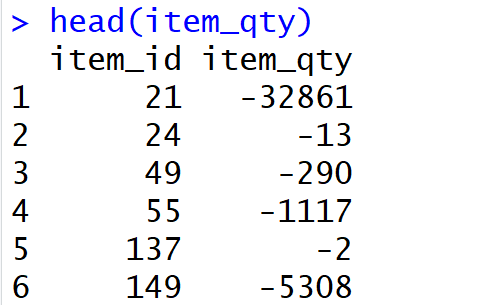


1. **Synergies between Alibaba and Cainiao**

Judging from the fact that the businesses using Cainiao has more favorable comments and more orders than other merchants, we should encourage merchants to use Cainiao through reducing Alibaba’s drawing rate, so as to improve the service quality of products. In addition, we can also recommend merchants whose products sell well to use Cainiao, by putting Cainiao into the Top1 deliver candidate list in their orders. So this could be a way to improve the profits both of Cainiao and Alibaba.

1. **Develop more accurate demand forecasts**

We can figure out how many goods each warehouse sells every day and how many goods each warehouse needs every day. We can also calculate the demand for every item in every city (*fig is below*)In this way, we can predict the demand of tomorrow through the demand of today, so that we can be clear about what kind of way we can use to replenish and adjust the goods before tomorrow starts. And we can also determine which city we should put into consideration when thinking about the deliver speed and the cost.

**** ****

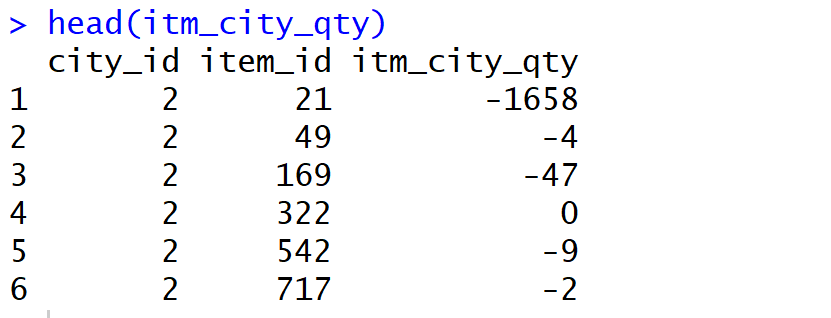
1. **Optimal inventory allocation across products and warehouses**

After getting the information between warehouse and items, we are supposed to distribute the goods proportionally according to the demand of the goods in each warehouse. Then, several convenient transfer stations will be set up in the middle among the cities. Whenever the goods are sold out, they will be replenishment or stuff will transfer goods in time from the stations.

For different goods, the core transfer stations and warehouses will be selected in proportion to the different demands of each city. In this way, the mobility between the warehouse and the goods could be adjusted according to the demand of goods.

1. **Optimize the distribution network**

In order to achieve the best trade-off between cost efficiency and fast delivery, establishing a distribution network of good logistics is a sustainable development strategy. Therefore, we need the total sales volume of each type of goods in each warehouse in recent years, which we can calculate though the data set. (*fig is below*) And the proportion ranking of the cities related with the demand of every item should be done. According to the principle of proximity, we can do a clustering analysis based on the time from the warehouse to the destination of the order, and the demand for goods in the cities close to each warehouse. After establishing this kind of clustering system, we can determine where to locate the warehouse for reducing the cost and improving the efficiency.

****

1. **How to monitor third-party logistics firms’ service**

For this part, what we can do first is to monitor order information at all times. If the merchant delays the delivery and arrival date promised to the customers, then Alibaba should send a warning email to the merchant as well as to the customers. What’s more, we are supposed to make a realistic score of the reputation system based on the speed of the logistics. Points will deducted for late shipment and arrival in excess of a certain percentage of all orders. The score can be divided into 1,2,3,4,5. The higher score the merchants get, the greater chance to enter the candidate list in the front page of the exposure machine recommendation they will have. This could promote inefficient merchants to choose high-quality logistics services.

1. **Conclusions**

After we reviewed the parts above we determined that it would be a good idea for a company to go deeper with the data to improve their business. The first major thing is that the data could provide the basic information, such as what kind of item has the highest number of order, what kind of merchant has good reputation and where can be improved between Alibaba and Cainiao to satisfy the market while still remaining cost effective. So coming up some brand-new strategies may lead to higher profits and more success for the long-term.

1. **Appendix**

**R code**

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

**library(dplyr)**

**library(plyr)**

**library(randomForest)**

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

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

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

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

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

**colnames(seller1) <- 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")**

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

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

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

**seller1$cainiao <- ifelse(seller1$if\_cainiao==0,"No","Yes")**

**seller1$cainiao <- factor(seller1$cainiao,levels=c("No","Yes"))**

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

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

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

**#dim(odr1\_num)**

**seller\_na <- na.omit(seller1)**

**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)**

**)**

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

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

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

**### 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)**

**)**

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

**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")**

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

**set.seed(1234)**

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

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

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

**# build the linear model**

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

**# 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)**

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

**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")**

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

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

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

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

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

**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)**

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

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

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

**.(),**

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

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

**})**

**cai\_percent**

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

**# Q5 analyze Cainiao accurate demand #**

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

**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)**

**## 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)**

**## 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 #**

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

**## 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)**

**## 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")**