Econometrics Lab04

## Packages

library(data.table)  
library(ggplot2)  
require(stargazer)

## Loading required package: stargazer

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

## Set Working Directory.

setwd("C:/Users/22700/Desktop/PhD Econometrics/28.10.2020")

## Laod the Data

list.files()

## [1] "DirectMarketing.csv" "Econometrics-Lab04.html"  
## [3] "Econometrics-Lab04.Rmd" "Econometrics Lab 4.Rmd"   
## [5] "Econometrics Lab04.Rmd" "lecture03"   
## [7] "lecture04"

load("lecture04/affairs.RData")  
dt.affairs <- data.table(data)  
rm(data)

## Summary Statistic

stargazer(dt.affairs, type = "text")

##   
## ============================================================  
## Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max   
## ------------------------------------------------------------  
## id 601 1,059.722 914.905 4 528 1,453 9,029  
## male 601 0.476 0.500 0 0 1 1   
## age 601 32.488 9.289 18 27 37 57   
## yrsmarr 601 8.178 5.571 0 4 15 15   
## kids 601 0.715 0.452 0 0 1 1   
## relig 601 3.116 1.168 1 2 4 5   
## educ 601 16.166 2.403 9 14 18 20   
## occup 601 4.195 1.819 1 3 6 7   
## ratemarr 601 3.932 1.103 1 3 5 5   
## naffairs 601 1.456 3.299 0 0 0 12   
## affair 601 0.250 0.433 0 0 0 1   
## vryhap 601 0.386 0.487 0 0 1 1   
## hapavg 601 0.323 0.468 0 0 1 1   
## avgmarr 601 0.155 0.362 0 0 0 1   
## unhap 601 0.110 0.313 0 0 0 1   
## vryrel 601 0.116 0.321 0 0 0 1   
## smerel 601 0.316 0.465 0 0 1 1   
## slghtrel 601 0.215 0.411 0 0 0 1   
## notrel 601 0.273 0.446 0 0 1 1   
## ------------------------------------------------------------

## Hypothesis

## Two-sided hypothesis test

## H0 : μnon−religious − μreligious = 0

## H1 : μnon−religious − μreligious 6= 0

## If using ‘dt.affairs[, religious:= ifelse(relig>3,1,0)]’ will be 0~ false, 1~ true

dt.affairs[, religious:= relig>3]

## Check how many people are in each group

dt.affairs[, .N, by = religious]

## religious N  
## 1: FALSE 341  
## 2: TRUE 260

## Run t.test( affair ~ religious)

dt.affairs[, t.test(affair ~ religious)]

##   
## Welch Two Sample t-test  
##   
## data: affair by religious  
## t = 3.7191, df = 594.76, p-value = 0.0002189  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.06043572 0.19568880  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 0.3049853 0.1769231

## Run t.test( naffair ~ religious)

dt.affairs[, t.test(naffairs ~ religious)]

##   
## Welch Two Sample t-test  
##   
## data: naffairs by religious  
## t = 4.0676, df = 593.3, p-value = 5.393e-05  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.5382493 1.5432981  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 1.9061584 0.8653846

## One-sided hypothesis test

## H0 : μnon−religious − μreligious <= 0

## H01μnon−religious − μreligious > 0

dt.affairs[, t.test(affair ~ religious, alternative = c("greater"))]

##   
## Welch Two Sample t-test  
##   
## data: affair by religious  
## t = 3.7191, df = 594.76, p-value = 0.0001094  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 0.07133542 Inf  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 0.3049853 0.1769231

dt.affairs[, t.test(naffairs ~ religious, alternative = c("greater"))]

##   
## Welch Two Sample t-test  
##   
## data: naffairs by religious  
## t = 4.0676, df = 593.3, p-value = 2.696e-05  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 0.6192441 Inf  
## sample estimates:  
## mean in group FALSE mean in group TRUE   
## 1.9061584 0.8653846

## Multiple Regression

## Case: Direct Marketing

## Load Data

dt.mktg <- data.table(read.csv("DirectMarketing.csv"))  
dt.mktg <- setnames(dt.mktg, tolower(names(dt.mktg)))

## Get to know the data

nrow(dt.mktg)

## [1] 1000

colnames(dt.mktg)

## [1] "age" "gender" "ownhome" "married" "location"   
## [6] "salary" "children" "history" "catalogs" "amountspent"

head(dt.mktg)

## age gender ownhome married location salary children history catalogs  
## 1: Old Female Own Single Far 47500 0 High 6  
## 2: Middle Male Rent Single Close 63600 0 High 6  
## 3: Young Female Rent Single Close 13500 0 Low 18  
## 4: Middle Male Own Married Close 85600 1 High 18  
## 5: Middle Female Own Single Close 68400 0 High 12  
## 6: Young Male Own Married Close 30400 0 Low 6  
## amountspent  
## 1: 755  
## 2: 1318  
## 3: 296  
## 4: 2436  
## 5: 1304  
## 6: 495

summary(dt.mktg)

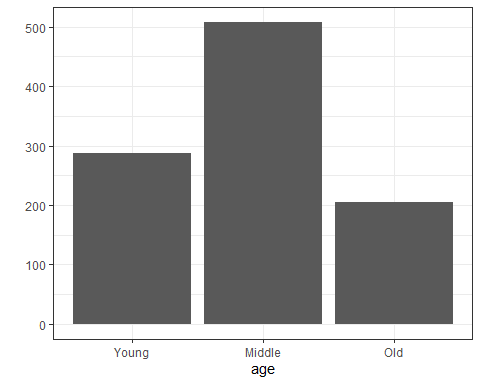
## age gender ownhome married   
## Length:1000 Length:1000 Length:1000 Length:1000   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## location salary children history   
## Length:1000 Min. : 10100 Min. :0.000 Length:1000   
## Class :character 1st Qu.: 29975 1st Qu.:0.000 Class :character   
## Mode :character Median : 53700 Median :1.000 Mode :character   
## Mean : 56104 Mean :0.934   
## 3rd Qu.: 77025 3rd Qu.:2.000   
## Max. :168800 Max. :3.000   
## catalogs amountspent   
## Min. : 6.00 Min. : 38.0   
## 1st Qu.: 6.00 1st Qu.: 488.2   
## Median :12.00 Median : 962.0   
## Mean :14.68 Mean :1216.8   
## 3rd Qu.:18.00 3rd Qu.:1688.5   
## Max. :24.00 Max. :6217.0

stargazer(dt.mktg, type = "text")

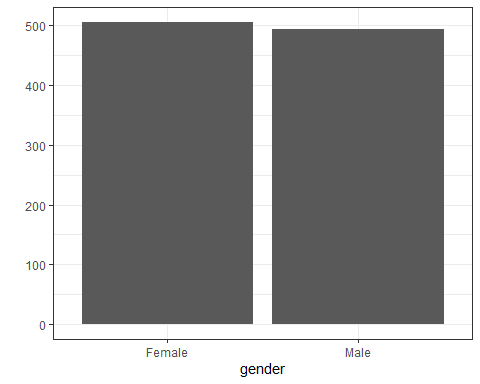
##   
## ========================================================================  
## Statistic N Mean St. Dev. Min Pctl(25) Pctl(75) Max   
## ------------------------------------------------------------------------  
## salary 1,000 56,103.900 30,616.310 10,100 29,975 77,025 168,800  
## children 1,000 0.934 1.051 0 0 2 3   
## catalogs 1,000 14.682 6.623 6 6 18 24   
## amountspent 1,000 1,216.770 961.069 38 488.2 1,688.5 6,217   
## ------------------------------------------------------------------------

## Explore the Data Graphically

qplot( data = dt.mktg  
, x = age  
, geom ="bar") + theme\_bw() + xlim("Young","Middle","Old")



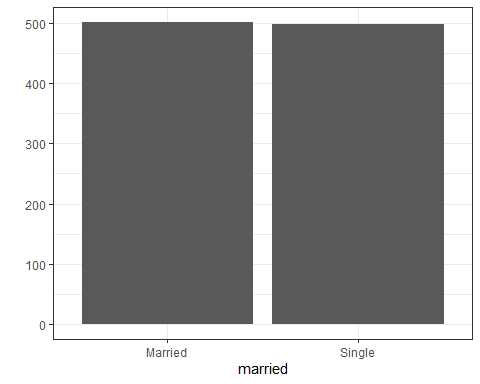
qplot( data = dt.mktg  
, x = gender  
, geom ="bar") + theme\_bw()



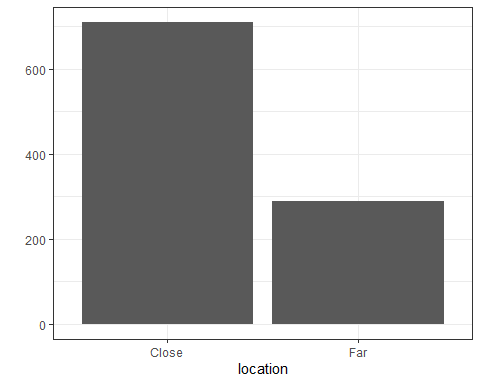
qplot( data = dt.mktg  
, x = ownhome  
, geom ="bar") + theme\_bw()



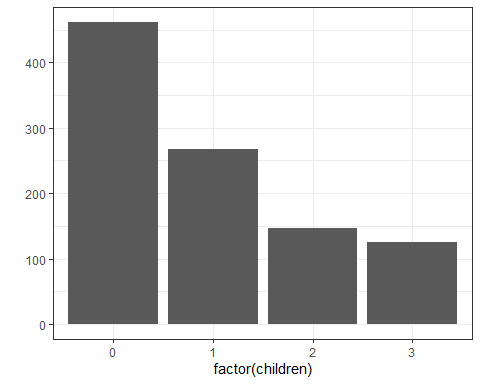
qplot( data = dt.mktg  
, x = married  
, geom ="bar") + theme\_bw()



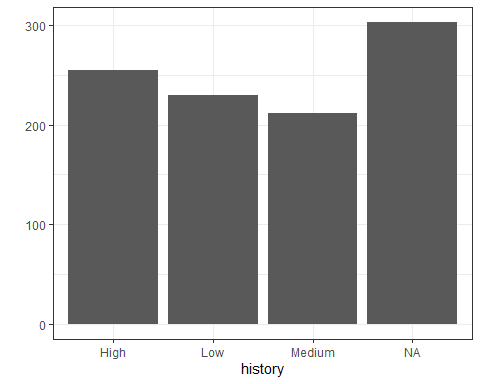
qplot( data = dt.mktg  
, x = location  
, geom ="bar") + theme\_bw()



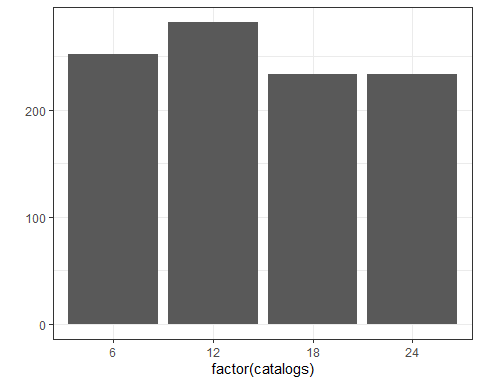
qplot( data = dt.mktg  
, x = factor(children)  
, geom ="bar") + theme\_bw()



qplot( data = dt.mktg  
, x = history  
, geom ="bar") + theme\_bw()

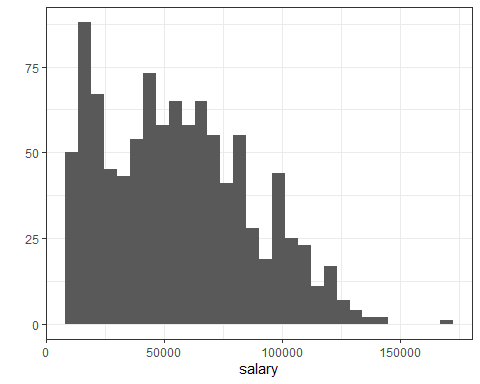


qplot( data = dt.mktg  
, x = factor(catalogs)  
, geom ="bar") + theme\_bw()

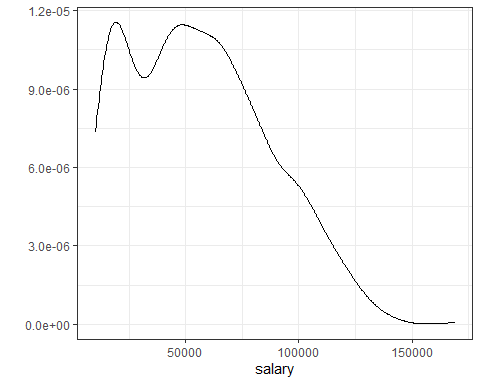


qplot( data = dt.mktg  
, x = salary  
, geom ="histogram") + theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

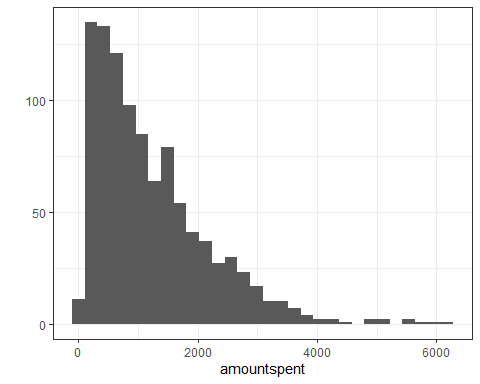


qplot( data = dt.mktg  
, x = salary  
, geom ="density") + theme\_bw()

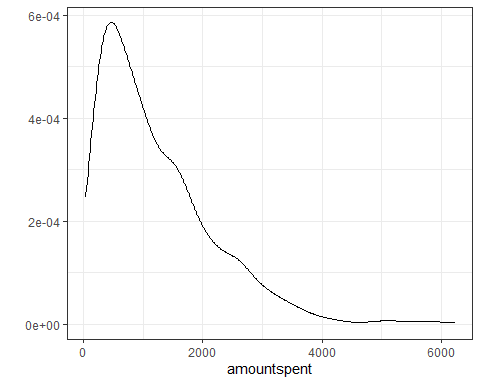


qplot( data = dt.mktg  
, x = amountspent  
, geom ="histogram") + theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

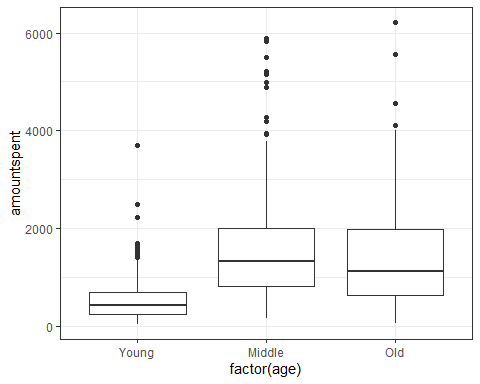


qplot( data = dt.mktg  
, x = amountspent  
, geom ="density") + theme\_bw()

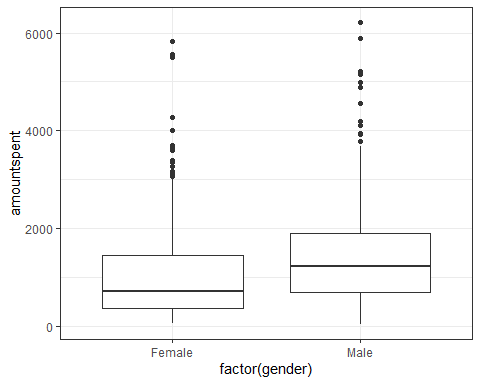


## Explore the Amountspent by the Different Customer Segments

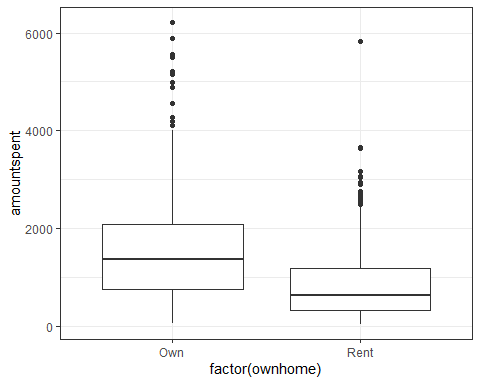
qplot( data = dt.mktg  
, x = factor(age)  
, y = amountspent  
, geom ="boxplot") + theme\_bw() + xlim("Young","Middle","Old")



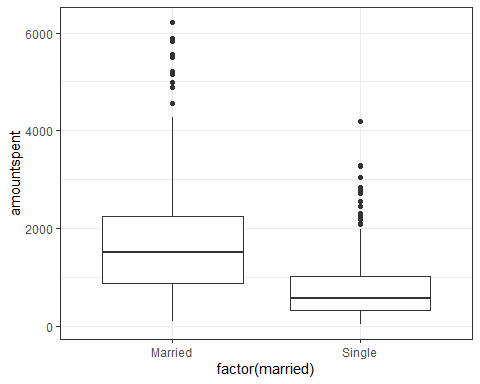
qplot( data = dt.mktg  
, x = factor(gender)  
, y = amountspent  
, geom ="boxplot") + theme\_bw()



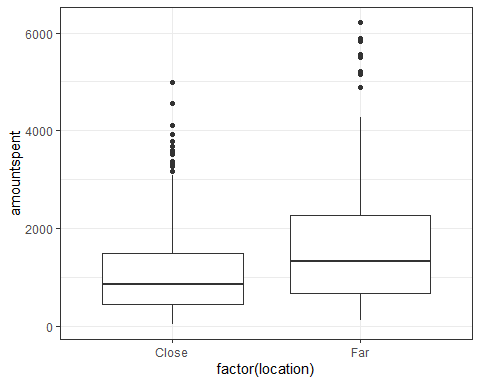
qplot( data = dt.mktg  
, x = factor(ownhome)  
, y = amountspent  
, geom ="boxplot") + theme\_bw()



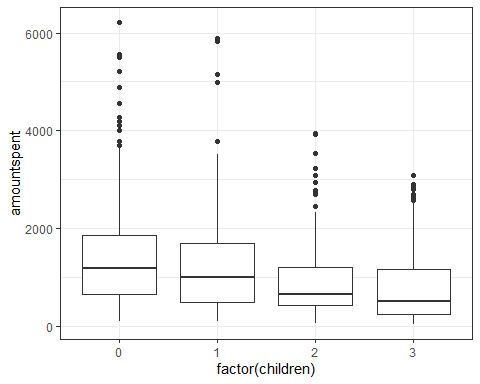
qplot( data = dt.mktg  
, x = factor(married)  
, y = amountspent  
, geom ="boxplot") + theme\_bw()



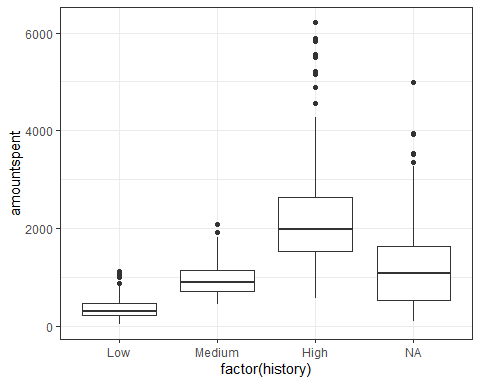
qplot( data = dt.mktg  
, x = factor(location)  
, y = amountspent  
, geom ="boxplot") + theme\_bw()



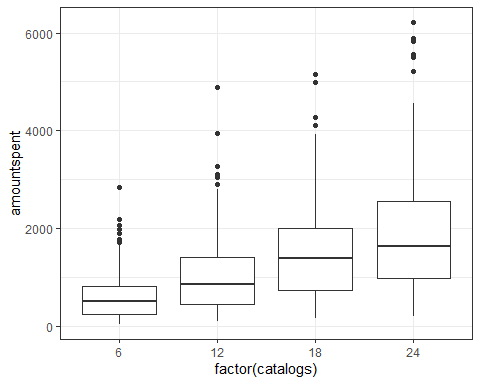
qplot( data = dt.mktg  
, x = factor(children)  
, y = amountspent  
, geom ="boxplot") + theme\_bw()



qplot( data = dt.mktg  
, x = factor(history)  
, y = amountspent  
, geom ="boxplot") + theme\_bw() + xlim("Low", "Medium", "High", NA)



qplot( data = dt.mktg  
, x = factor(catalogs)  
, y = amountspent  
, geom ="boxplot") + theme\_bw()



## Simple regression & Interpretation

lm1 <- lm(amountspent ~ salary, data = dt.mktg)  
stargazer(lm1, type = "text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## amountspent   
## -----------------------------------------------  
## salary 0.022\*\*\*   
## (0.001)   
##   
## Constant -15.318   
## (45.374)   
##   
## -----------------------------------------------  
## Observations 1,000   
## R2 0.489   
## Adjusted R2 0.489   
## Residual Std. Error 687.065 (df = 998)   
## F Statistic 956.694\*\*\* (df = 1; 998)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

lm2 <- lm(amountspent ~ location, data = dt.mktg)  
stargazer(lm2, type ="text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## amountspent   
## -----------------------------------------------  
## locationFar 534.773\*\*\*   
## (64.837)   
##   
## Constant 1,061.686\*\*\*   
## (34.916)   
##   
## -----------------------------------------------  
## Observations 1,000   
## R2 0.064   
## Adjusted R2 0.063   
## Residual Std. Error 930.364 (df = 998)   
## F Statistic 68.028\*\*\* (df = 1; 998)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

dt.mktg[location=="Close", mean(amountspent)]

## [1] 1061.686

dt.mktg[location=="Far", mean(amountspent)]

## [1] 1596.459

lm3 <- lm(amountspent ~ history, data = dt.mktg)  
stargazer(lm3 , type ="text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## amountspent   
## -----------------------------------------------  
## historyLow -1,829.050\*\*\*   
## (56.917)   
##   
## historyMedium -1,235.736\*\*\*   
## (58.174)   
##   
## Constant 2,186.137\*\*\*   
## (39.196)   
##   
## -----------------------------------------------  
## Observations 697   
## R2 0.610   
## Adjusted R2 0.608   
## Residual Std. Error 625.902 (df = 694)   
## F Statistic 541.884\*\*\* (df = 2; 694)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

dt.mktg[history=="High", mean(amountspent)]

## [1] 2186.137

dt.mktg[history=="Low", mean(amountspent)]

## [1] 357.087

dt.mktg[history=="Medium", mean(amountspent)]

## [1] 950.4009

## Multiple regression

lm.spend1 <- lm( amountspent ~ gender + location + salary + children + catalogs, data = dt.mktg)  
stargazer(lm.spend1 , type ="text")

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## amountspent   
## -----------------------------------------------  
## genderMale -42.309   
## (33.959)   
##   
## locationFar 508.129\*\*\*   
## (36.207)   
##   
## salary 0.021\*\*\*   
## (0.001)   
##   
## children -205.806\*\*\*   
## (15.731)   
##   
## catalogs 42.802\*\*\*   
## (2.544)   
##   
## Constant -528.143\*\*\*   
## (50.454)   
##   
## -----------------------------------------------  
## Observations 1,000   
## R2 0.715   
## Adjusted R2 0.714   
## Residual Std. Error 514.103 (df = 994)   
## F Statistic 499.438\*\*\* (df = 5; 994)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

lm.spend2 <- lm(amountspent ~ ., data = dt.mktg)  
stargazer(lm.spend1, lm.spend2 , type ="text")

##   
## ======================================================================  
## Dependent variable:   
## --------------------------------------------------  
## amountspent   
## (1) (2)   
## ----------------------------------------------------------------------  
## ageOld 41.385   
## (52.764)   
##   
## ageYoung 89.654   
## (58.741)   
##   
## genderMale -42.309 -53.701   
## (33.959) (38.016)   
##   
## ownhomeRent -18.288   
## (41.512)   
##   
## marriedSingle 19.503   
## (49.812)   
##   
## locationFar 508.129\*\*\* 608.992\*\*\*   
## (36.207) (43.985)   
##   
## salary 0.021\*\*\* 0.019\*\*\*   
## (0.001) (0.001)   
##   
## children -205.806\*\*\* -268.283\*\*\*   
## (15.731) (25.019)   
##   
## historyLow -267.514\*\*\*   
## (88.617)   
##   
## historyMedium -344.553\*\*\*   
## (59.964)   
##   
## catalogs 42.802\*\*\* 40.521\*\*\*   
## (2.544) (2.868)   
##   
## Constant -528.143\*\*\* -249.579\*   
## (50.454) (134.031)   
##   
## ----------------------------------------------------------------------  
## Observations 1,000 697   
## R2 0.715 0.789   
## Adjusted R2 0.714 0.785   
## Residual Std. Error 514.103 (df = 994) 463.457 (df = 685)   
## F Statistic 499.438\*\*\* (df = 5; 994) 232.493\*\*\* (df = 11; 685)  
## ======================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## Predict amount spent by new customer

new.client <- data.table( gender = "Male"  
, location = "Close"  
, salary = 53700  
, children = 1  
, catalogs = 12)  
new.client

## gender location salary children catalogs  
## 1: Male Close 53700 1 12

my.pred <- predict(lm.spend1, newdata = new.client)  
my.pred

## 1   
## 868.9695

my.pred <- predict(lm.spend1, newdata = new.client, interval="prediction", level = .95)  
my.pred

## fit lwr upr  
## 1 868.9695 -141.2554 1879.194