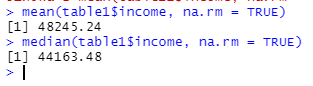
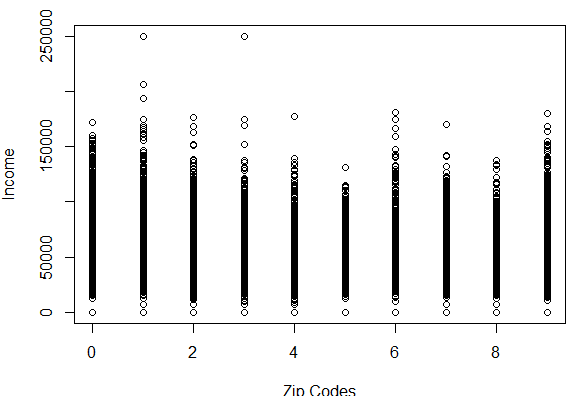
table1 <- read.table("D:/zipIncome.txt", header = TRUE, sep = "|") TRUE – первая строка читается как имя переменное

необходимо указывать, каким образом столбцы отделены друг от друга (аргумент sep, разделитель

а также учитывать, что представляет собой первая строка: наблюдение или шапку таблицы



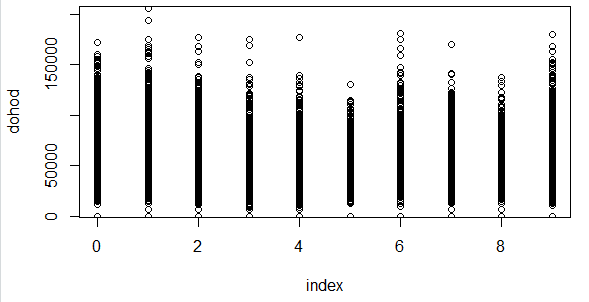
b. plot(table1$zipCode, table1$income, xlab = "Zip Codes", ylab = "Income")**In order to omit outliers, create a subset of the data so that:**

**$7,000 < income < $200,000**

res <- subset(table1, income > 7000, & income < 200000)

plot(table1$zipCode, table1$income, xlab = "index", ylab = "dohod", ylim = c(7000, 200000))

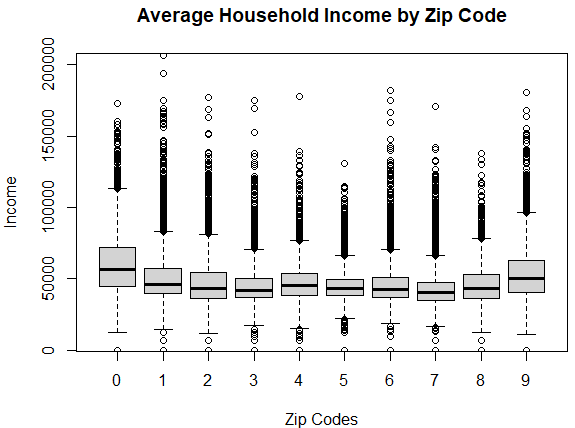
(Есть запасной вариант, если этот не правильный)

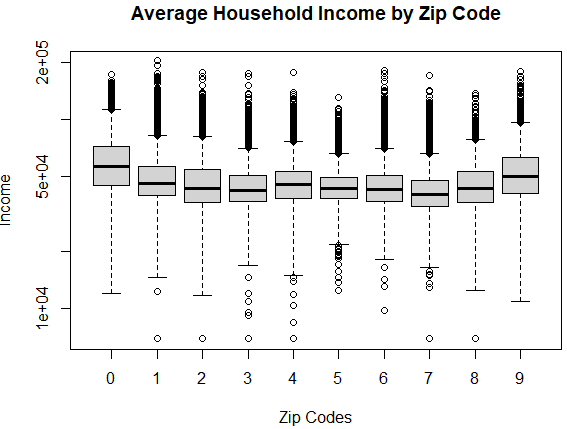




c. Выбросы – это максимальный доход, возможно. Так же можно считать, что выбросы, выделяющиеся из общего числа.

d. boxplot(income ~ zipCode, data = table1, main = "Average Household Income by Zip Code", xlab = "Zip Codes", ylab = "Income", ylim = c(7000, 200000))



e. boxplot(income ~ zipCode, data = table1, main = "Average Household Income by Zip Code", xlab = "Zip Codes", ylab = "Income", ylim = c(7000, 200000), log = "y")

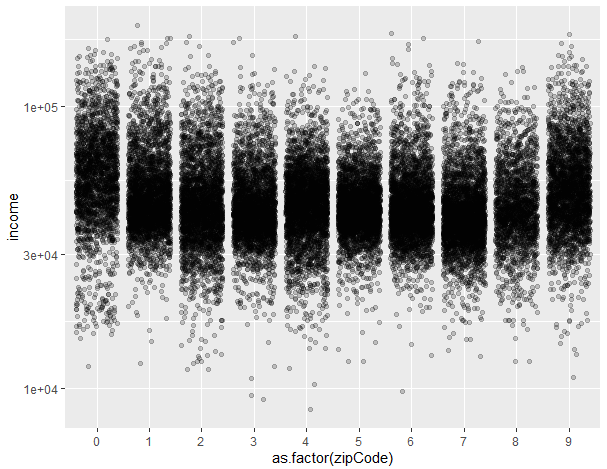
**Use the *ggplot* library in R, which enables you to create graphs with several different types of plots layered over each other.  Be sure to read the documentation for *ggplot* and load the library *ggplot2* (you may have to install this package into R).**

install.packages("ggplot2")

library(ggplot2)

ggplot2 – это библиотека для R, позволяющая более тонко работать с настройкой различных графиков. Графики jitter включают в себя специальные эффекты, с помощью которых могут быть изображены рассеянные графики. Jitter – это случайное значение, которое присваивается точкам для их разделения. Aes чтобы понять какие оси, factor чтобы понимать по чему группировать. Ggplot это конструктор для поля но котором будет все строиться. Point чтобы было отображение точек.

f. ggplot(data = res, aes(as.factor(zipCode), income)) + geom\_point(position = "jitter", alpha = 0.2) + scale\_y\_log10()

****

g. Jitter случайное значение чтобы точки не липли друг к друг , чтобы график нормально выглядел имея много точек на одном датасете. Alpa регулирует насыщенность, чтобы тоже было нормальное восприятие.

К графику, полученному выше, можно добавить слой блочной диаграммы и цветовое распределение. ggplot(res,aes(x=as.factor(zipCode),y=income))+geom\_point(aes(colour=factor(zipCode)),position = 'jitter',alpha=0.2)+ geom\_boxplot(alpha=0.1,outlier.size =0) + scale\_y\_log10()+labs(color="Region",x="Zip Code",y="Income",title="Average Income by Zip Code")

