Homework #4

Instruction: Do all the following empirical exercises using R. Turn in your answer with tables and graphs, if any, (along with your program and output files appended at the end of document). Refer to the output file whenever appropriate when discussing your results.

Note that for all simulation exercises, set the seed number to 123456 to ensure the reproducibility of your results.

Question 1.

In `R`, type in the following code to simulate a dataset for this question.

```r
set.seed(123456)
y <- sample(c(1:5),1000, replace= TRUE, prob=c(.25,.25,.2,.2,.1))
x <- sample(c(0:1),1000, replace= TRUE, prob=c(.5,.5))
z1 <- sample(c(0:1),1000, replace= TRUE, prob=c(.25,.75))
z2 <- sample(c(0:1),1000, replace= TRUE, prob=c(.3,.6))
data <- as.data.frame(cbind(y,x,z1,z2))
attach(data)
```

Use the simulated data,

1. Calculate the conditional distribution $\Pr[Y|X=1]$ using two approaches illustrated in class.
   - Manually select a subsample and calculate the conditional distribution.
   - Generate dummy variables for each category of Y.
   - Use the `aggregate()` function to calculate the conditional distribution.
   - What is your prediction for Y when $X=1$?
2. Use whichever approach that you like to calculate the conditional distribution $\Pr[Y|X, Z1, Z2]$.
3. Using your results in 2), what is your prediction when $X=1, Z1=1, Z2=0$?

Question 2.

Use the `airbnb.csv` data on Canvas. This dataset contains (fake) historical data on various hosts' decisions for requests with different check-in gaps (as discussed in class). Answer the following questions:

1. If a customer requests a room with a check-in gap of 11 days, is it likely for this person to get his request accepted? Why?
2. Suppose that a customer submitted his request. Coincidentally, every host in your dataset has a check-in gap of 2 days for this particular request. Which host(s) should you recommend to this customer in order to maximize the chance that this recommended host will accept this request?

Question 3.

In `R`, type in the following code to simulate data

```r
library(MASS)
set.seed(123456)
y <- rnorm(1000)
z <- sample(1:3,1000, replace = TRUE)
```

Using the simulated data, answer the following questions

1.Calculate the conditional mean function with `aggregate()` command. And write it in one equation.

2. Similarly calculate the conditional median with two approaches.
   1. Subset approach (manually)
   2. Use `aggregate()`. You need to learn how to specify a different function in this case.
3. Suppose that this is some type of financial data for which I care about the lower-tail risks, and I would like to predict such value. z represents different markets. Suppose that I would like know about 5^{th} percentile of my returns to this stock in the market $z = 1$. Calculate this value.