Homework2.R

student

Fri Feb 12 01:14:17 2016

##CS 480 Homework 2  
##Neeraj Asthana (nasthan2)  
  
library(biganalytics)

## Loading required package: bigmemory

## Loading required package: bigmemory.sri

## Loading required package: foreach

## Loading required package: biglm

## Loading required package: DBI

library(foreach)  
  
#Create database connection  
setwd("/home/student/container-data/RDataScience/AirlineDelays")  
  
##Q8  
#Which is the best day of the week to fly?  
#minimize average delay by day of the week  
y <- attach.big.matrix("air0708.desc")  
  
#Get mean arrival delays for each day  
totalDayDelays <- foreach(i = 1:7, .combine=c) %do% {  
 mean(y[y[,"DayOfWeek"]==i, "ArrDelay"], na.rm=TRUE)  
}  
totalDayDelays

## [1] 9.402934 7.880886 8.277983 10.608164 12.043961 5.819331 9.925622

##Q9  
#Which is the best day of the month to fly?  
#Get mean departure delays for each day  
totalMonthDayDelays <- foreach(i = 1:31, .combine=c) %do% {  
 mean(y[y[,"DayofMonth"]==i, "ArrDelay"], na.rm=TRUE)  
}  
totalMonthDayDelays

## [1] 10.038581 8.347670 7.157116 8.595337 8.361154 8.112159 8.348323  
## [8] 8.612245 8.623283 9.400009 8.951881 8.135688 7.894449 9.352992  
## [15] 11.843079 10.642763 9.013132 9.502615 11.331052 8.725341 10.841925  
## [22] 11.306045 9.115103 7.737640 9.192238 11.318314 11.172735 8.119991  
## [29] 7.181018 8.447295 10.614867

##Q14  
#How much do weather delays contribute to arrival delay?  
weatherDelays <- biglm.big.matrix( ArrDelay ~ WeatherDelay, data = y )  
summary(weatherDelays)

## Large data regression model: biglm(formula = formula, data = data, ...)  
## Sample size = 8800023   
## Coef (95% CI) SE p  
## (Intercept) 16.9910 16.9608 17.0212 0.0151 0  
## WeatherDelay 1.0862 1.0837 1.0887 0.0012 0

#I will calculate the percentage of each arrival delays that is caused by weather delays where an arrival delay exists  
arrDelayExists <- y[,"ArrDelay"] > 0  
weatherDelays <- y[arrDelayExists,"WeatherDelay"]  
arrDelays <- y[arrDelayExists,"ArrDelay"]  
percentWeatherContrib <- mean (weatherDelays/ arrDelays, na.rm=TRUE)  
percentWeatherContrib

## [1] 0.02588029

##Q15  
#Along with age, which other variables in the airline delay data set contribute to arrival delays?  
#TaxiIn, TaxiOut, Distance, Diverted, CarrierDelay, NASDelay, SecurityDelay, LateAircraftDelay  
taxiIn <- y[arrDelayExists,"TaxiIn"]  
taxiOut <- y[arrDelayExists,"TaxiOut"]  
Distance <- y[arrDelayExists,"Distance"]  
Diverted <- y[arrDelayExists,"Diverted"]  
CarrierDelay <- y[arrDelayExists,"CarrierDelay"]  
NASDelay <- y[arrDelayExists,"NASDelay"]  
SecurityDelay <- y[arrDelayExists,"SecurityDelay"]  
LateAircraftDelay <- y[arrDelayExists,"LateAircraftDelay"]  
  
blm <- biglm.big.matrix( ArrDelay ~ TaxiIn+TaxiOut+Distance+Diverted+CarrierDelay+NASDelay+SecurityDelay+LateAircraftDelay, data = y)  
summary(blm)

## Large data regression model: biglm(formula = formula, data = data, ...)  
## Sample size = 8800023   
## Coef (95% CI) SE p  
## (Intercept) -5.7810 -5.8048 -5.7572 0.0119 0  
## TaxiIn 0.2018 0.2000 0.2036 0.0009 0  
## TaxiOut 0.1879 0.1870 0.1887 0.0004 0  
## Distance -0.0017 -0.0017 -0.0017 0.0000 0  
## Diverted NA NA NA NA NA  
## CarrierDelay 1.0166 1.0162 1.0170 0.0002 0  
## NASDelay 0.9957 0.9952 0.9963 0.0003 0  
## SecurityDelay 1.0386 1.0309 1.0464 0.0039 0  
## LateAircraftDelay 1.0384 1.0380 1.0388 0.0002 0

percentCarrierContrib <- mean (CarrierDelay/ arrDelays, na.rm=TRUE)