SIDS Data Milestone 2 Draft

March 29, 2017

I am planning to use the template but I just wanted to try on my own at first.

Below is the section of R code to clean and arrrainge the data in a more useful way.

> library(readxl)  
> #read in the data  
> COD <- read\_excel("~/Box Sync/Multiple Cause of Death, 1999-2015.xls")  
> View(COD)  
>   
> #just trying to see if anything I run will work  
> #to get collum names  
> names(COD)

[1] "2013 Urbanization"   
[2] "2013 Urbanization Code"   
[3] "Infant Age Groups"   
[4] "Infant Age Groups Code"   
[5] "UCD - ICD-10 130 Cause List (Infants)"   
[6] "UCD - ICD-10 130 Cause List (Infants) Code"  
[7] "Deaths"   
[8] "Population"   
[9] "Crude Rate"

> summary(COD)

2013 Urbanization 2013 Urbanization Code Infant Age Groups   
 Length:2331 Min. :1.000 Length:2331   
 Class :character 1st Qu.:2.000 Class :character   
 Mode :character Median :3.000 Mode :character   
 Mean :3.331   
 3rd Qu.:5.000   
 Max. :6.000   
 Infant Age Groups Code UCD - ICD-10 130 Cause List (Infants)  
 Length:2331 Length:2331   
 Class :character Class :character   
 Mode :character Mode :character   
   
   
   
 UCD - ICD-10 130 Cause List (Infants) Code Deaths   
 Length:2331 Min. : 10.0   
 Class :character 1st Qu.: 30.0   
 Mode :character Median : 84.0   
 Mean : 528.1   
 3rd Qu.: 287.0   
 Max. :50934.0   
 Population Crude Rate   
 Min. : 3837310 Length:2331   
 1st Qu.: 5742486 Class :character   
 Median :14379025 Mode :character   
 Mean :12218310   
 3rd Qu.:16098751   
 Max. :22969673

> # renaming variables  
> # df = dataframe  
> # old.var.name = The name you don't like anymore  
> # new.var.name = The name you want to get  
> names(COD)[names(COD) == 'Infant Age Groups Code'] <- 'age.group'  
> names(COD)[names(COD) == '2013 Urbanization Code'] <- 'urban.code'  
> names(COD)[names(COD) == 'Crude Rate'] <- 'crude.rate'  
> names(COD)[names(COD) == '2013 Urbanization'] <- 'urbanization'  
> names(COD)[names(COD) == 'UCD - ICD-10 130 Cause List (Infants) Code'] <- 'icd'  
>   
> # recode deaths into categorical infant age groups  
> COD$age.cat <- revalue(COD$age.group, c("1d"="1", "1-6d"="2", "7-27d"="3", "28-364d"="4"))  
>   
> # there is some non numeric data in crude.rate, change to NA  
> COD[COD=="Unreliable"]<-""  
> COD

# A tibble: 2,331 × 10  
 urbanization urban.code `Infant Age Groups` age.group  
 <chr> <dbl> <chr> <chr>  
1 Large Central Metro 1 < 1 day 1d  
2 Large Central Metro 1 < 1 day 1d  
3 Large Central Metro 1 < 1 day 1d  
4 Large Central Metro 1 < 1 day 1d  
5 Large Central Metro 1 < 1 day 1d  
6 Large Central Metro 1 < 1 day 1d  
7 Large Central Metro 1 < 1 day 1d  
8 Large Central Metro 1 < 1 day 1d  
9 Large Central Metro 1 < 1 day 1d  
10 Large Central Metro 1 < 1 day 1d  
# ... with 2,321 more rows, and 6 more variables: `UCD - ICD-10 130 Cause  
# List (Infants)` <chr>, icd <chr>, Deaths <dbl>, Population <dbl>,  
# crude.rate <chr>, age.cat <chr>

> # need to change crude rate to numeric variable  
> COD$crude.rate<-as.numeric(COD$crude.rate)   
> summary(COD)

urbanization urban.code Infant Age Groups age.group   
 Length:2331 Min. :1.000 Length:2331 Length:2331   
 Class :character 1st Qu.:2.000 Class :character Class :character   
 Mode :character Median :3.000 Mode :character Mode :character   
 Mean :3.331   
 3rd Qu.:5.000   
 Max. :6.000   
   
 UCD - ICD-10 130 Cause List (Infants) icd   
 Length:2331 Length:2331   
 Class :character Class :character   
 Mode :character Mode :character   
   
   
   
   
 Deaths Population crude.rate age.cat   
 Min. : 10.0 Min. : 3837310 Min. :0.0000 Length:2331   
 1st Qu.: 30.0 1st Qu.: 5742486 1st Qu.:0.0000 Class :character   
 Median : 84.0 Median :14379025 Median :0.0000 Mode :character   
 Mean : 528.1 Mean :12218310 Mean :0.0467   
 3rd Qu.: 287.0 3rd Qu.:16098751 3rd Qu.:0.0000   
 Max. :50934.0 Max. :22969673 Max. :2.2000   
 NA's :342

> #summary statistics  
> summary(COD)

urbanization urban.code Infant Age Groups age.group   
 Length:2331 Min. :1.000 Length:2331 Length:2331   
 Class :character 1st Qu.:2.000 Class :character Class :character   
 Mode :character Median :3.000 Mode :character Mode :character   
 Mean :3.331   
 3rd Qu.:5.000   
 Max. :6.000   
   
 UCD - ICD-10 130 Cause List (Infants) icd   
 Length:2331 Length:2331   
 Class :character Class :character   
 Mode :character Mode :character   
   
   
   
   
 Deaths Population crude.rate age.cat   
 Min. : 10.0 Min. : 3837310 Min. :0.0000 Length:2331   
 1st Qu.: 30.0 1st Qu.: 5742486 1st Qu.:0.0000 Class :character   
 Median : 84.0 Median :14379025 Median :0.0000 Mode :character   
 Mean : 528.1 Mean :12218310 Mean :0.0467   
 3rd Qu.: 287.0 3rd Qu.:16098751 3rd Qu.:0.0000   
 Max. :50934.0 Max. :22969673 Max. :2.2000   
 NA's :342

> mean(COD$Deaths)

[1] 528.1167

> mean(COD$Population)

[1] 12218310

Discussion explaining the summary statistics so I won't be useing the raw code.

Now I will try to make some pretty tables:

> #this is where some code will go for tables that I have not quite yet figured out how to make becuase none of the code has worked so far

These are some not-so-pretty tables that I am using just to look at the data and how it is related to one another.

> #aggregate data (and not pretty tables)  
> aggregate(Deaths ~ age.group, COD, mean)

age.group Deaths  
1 1-6d 283.9237  
2 1d 1165.9103  
3 28-364d 465.8399  
4 7-27d 255.1030

> aggregate(Deaths ~ urban.code, COD, mean)

urban.code Deaths  
1 1 918.0288  
2 2 626.5634  
3 3 641.5012  
4 4 316.5110  
5 5 292.7438  
6 6 223.9006

> aggregate(crude.rate ~ age.group, COD, mean)

age.group crude.rate  
1 1-6d 0.02407407  
2 1d 0.11220657  
3 28-364d 0.04034003  
4 7-27d 0.01756198

> aggregate(crude.rate ~ urban.code, COD, mean)

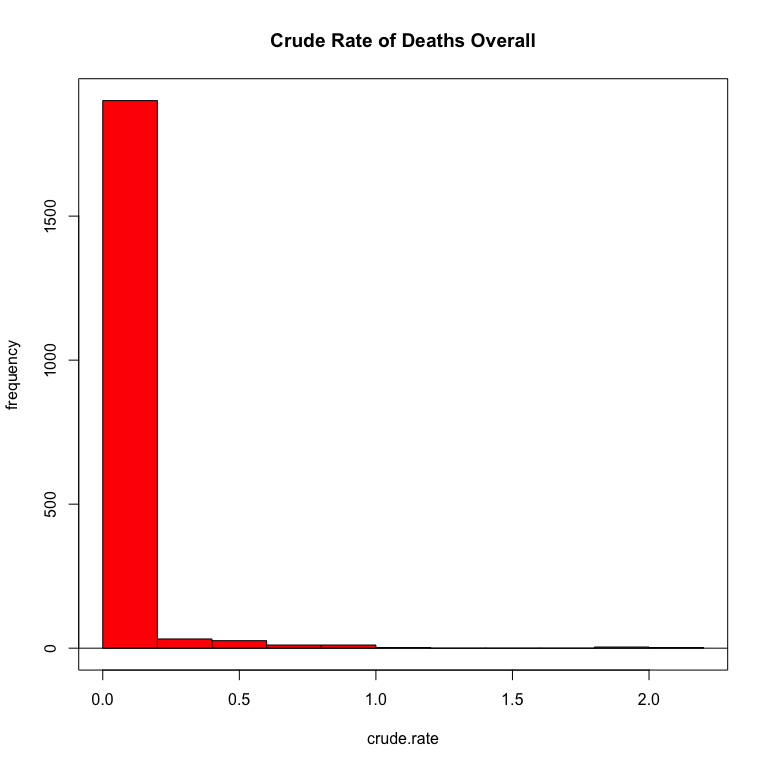
urban.code crude.rate  
1 1 0.03861386  
2 2 0.03661202  
3 3 0.04173442  
4 4 0.05374150  
5 5 0.05337838  
6 6 0.06461538

> aggregate(urban.code ~ Population, COD, mean)

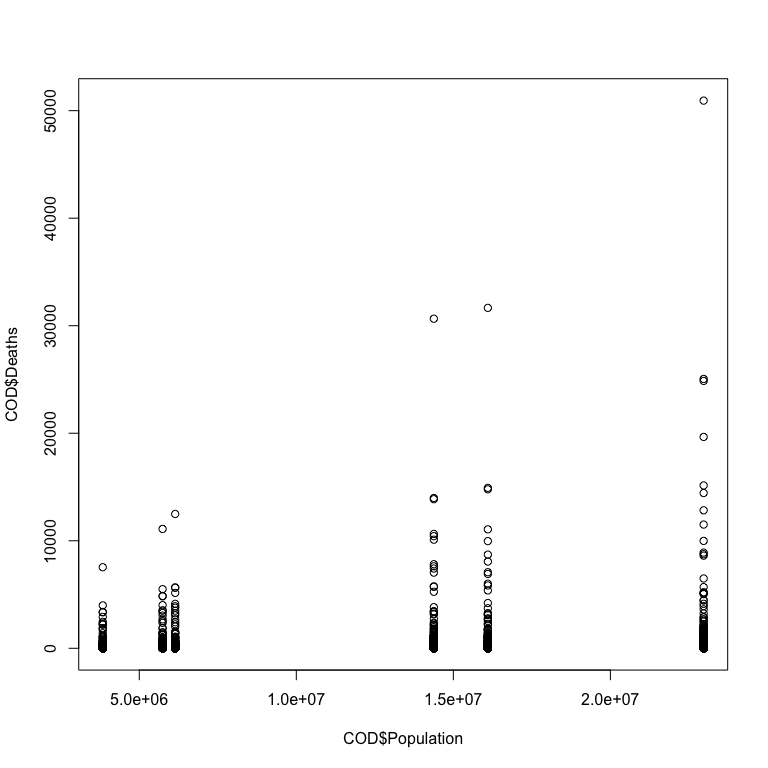
Population urban.code  
1 3837310 6  
2 5742486 5  
3 6143714 4  
4 14379025 3  
5 16098751 2  
6 22969673 1

Do some exploratory analysis and graphing (all of this will have explinations in the final draft). I will not use all these graphs in the final report but I wanted to practice making them. I will also go back in and tidy it up with lables and titles for each one chosen.

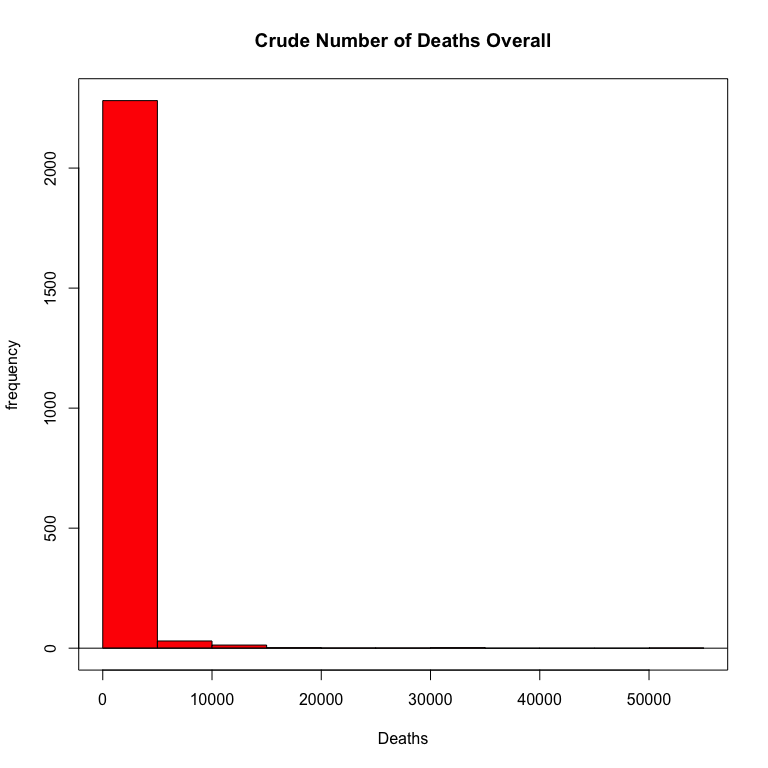
> #some exploratory analysis of the data  
> #histogram  
> with(COD, Hist(crude.rate, scale="frequency", breaks="Sturges", col="red"))  
> #add lables  
> title (main= "Crude Rate of Deaths Overall")



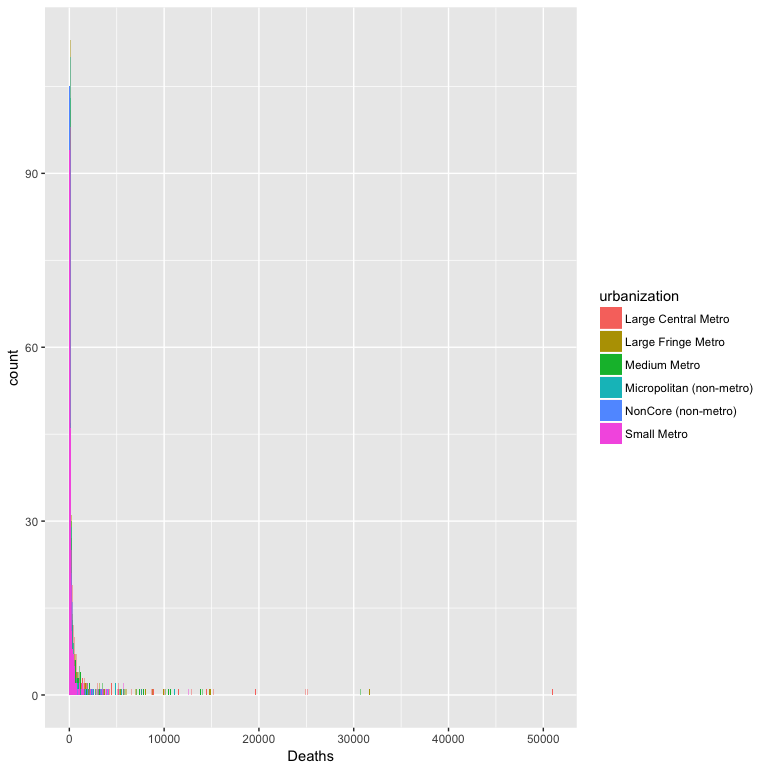
> #scatterplot  
> plot(COD$Population, COD$Deaths)



> #histogram  
> with(COD, Hist(Deaths, scale="frequency", breaks="Sturges", col="red"))  
> #add lables  
> title (main= "Crude Number of Deaths Overall")



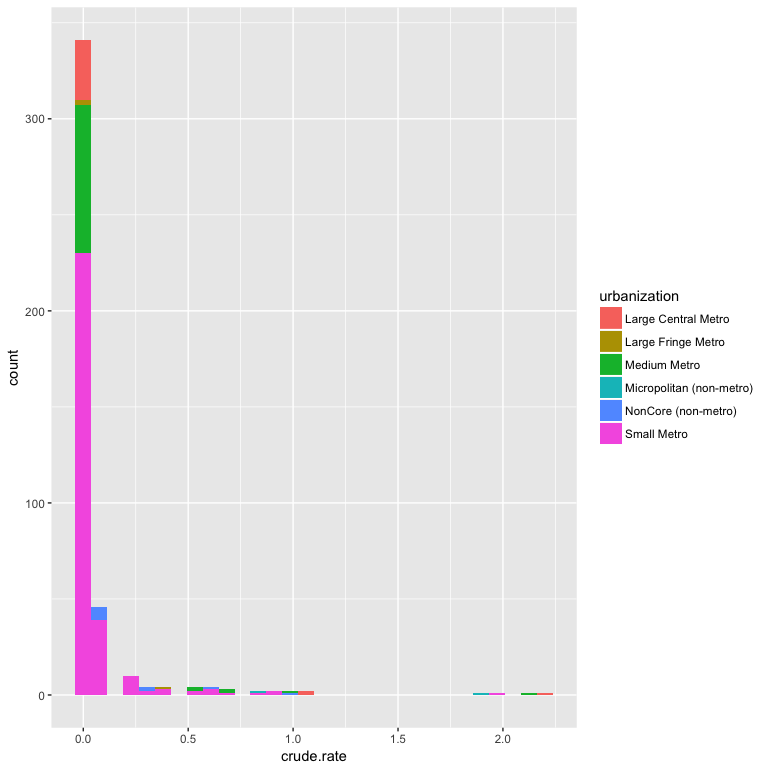
> # histogram of deaths by urbanization  
> ggplot(COD, aes(x=Deaths, fill=urbanization)) +  
+ geom\_histogram(binwidth=50, position="identity")



> #histogram of crude rate by urbanization  
> ggplot(COD, aes(x=crude.rate, fill=urbanization)) +  
+ geom\_histogram(position="identity")

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

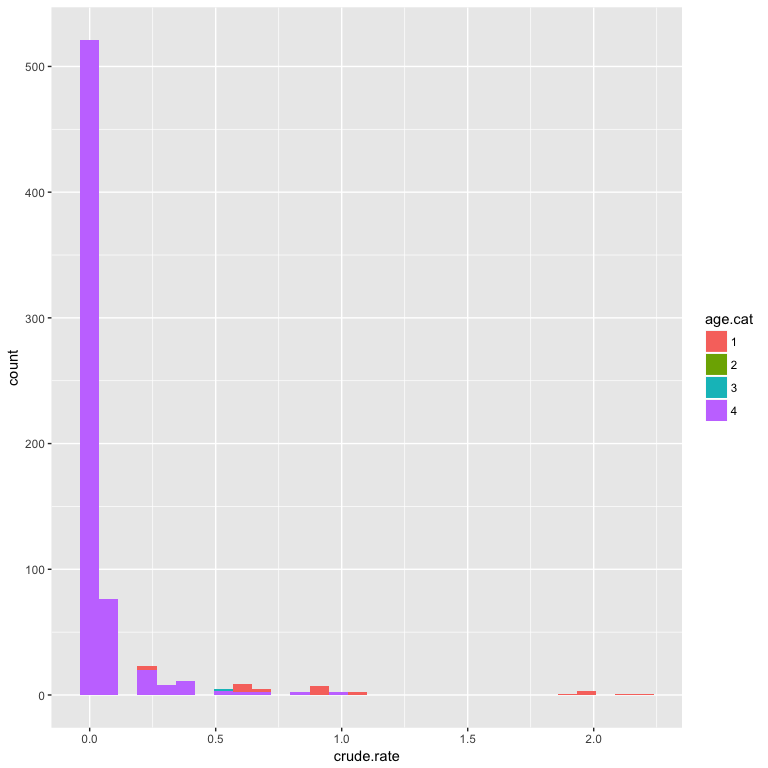
Warning: Removed 342 rows containing non-finite values (stat\_bin).



> #histogram of crude rate by age  
> ggplot(COD, aes(x=crude.rate, fill=age.cat)) +  
+ geom\_histogram(position="identity")

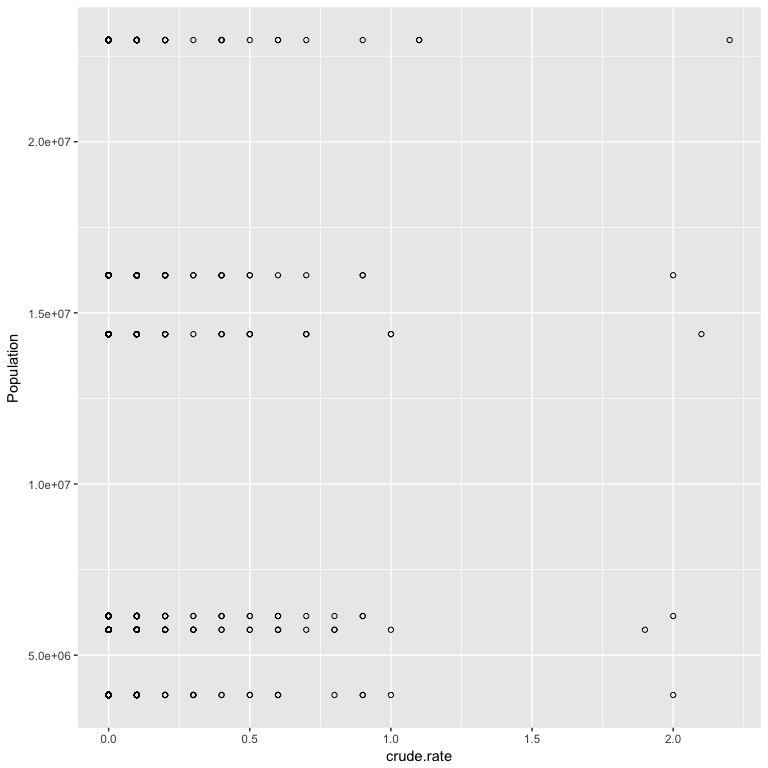
`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 342 rows containing non-finite values (stat\_bin).

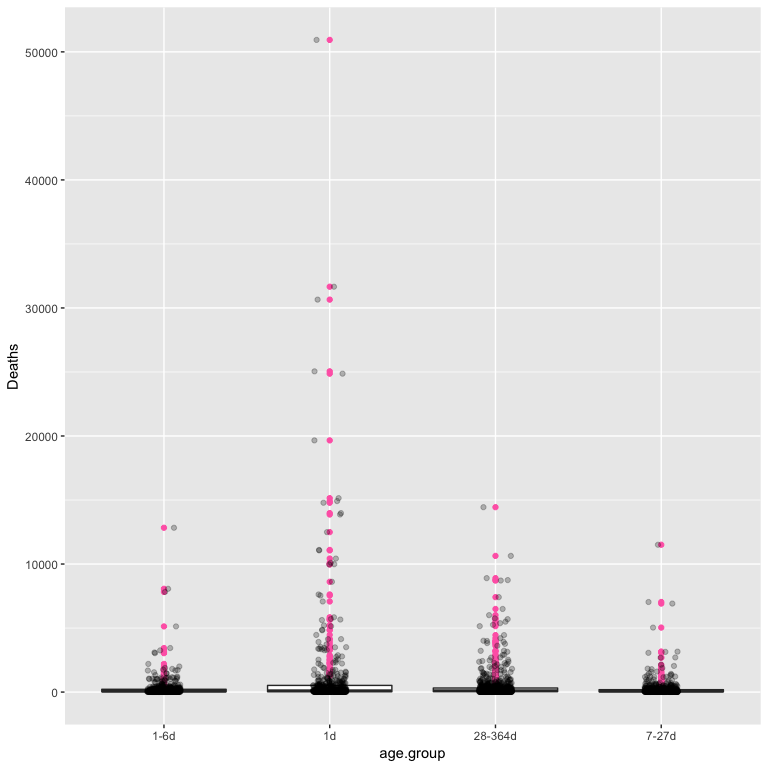


> #scatterplots to look at data against one another  
> ggplot(COD, aes(x=crude.rate, y=Population)) +  
+ geom\_point(shape=1)

Warning: Removed 342 rows containing missing values (geom\_point).



> #plot some of the data   
> library(ggplot2)  
> ggplot(COD, aes(x = age.group, y = Deaths)) +  
+ geom\_boxplot(outlier.colour = "hotpink") +  
+ geom\_jitter(position = position\_jitter(width = 0.1, height = 0), alpha = 1/4)

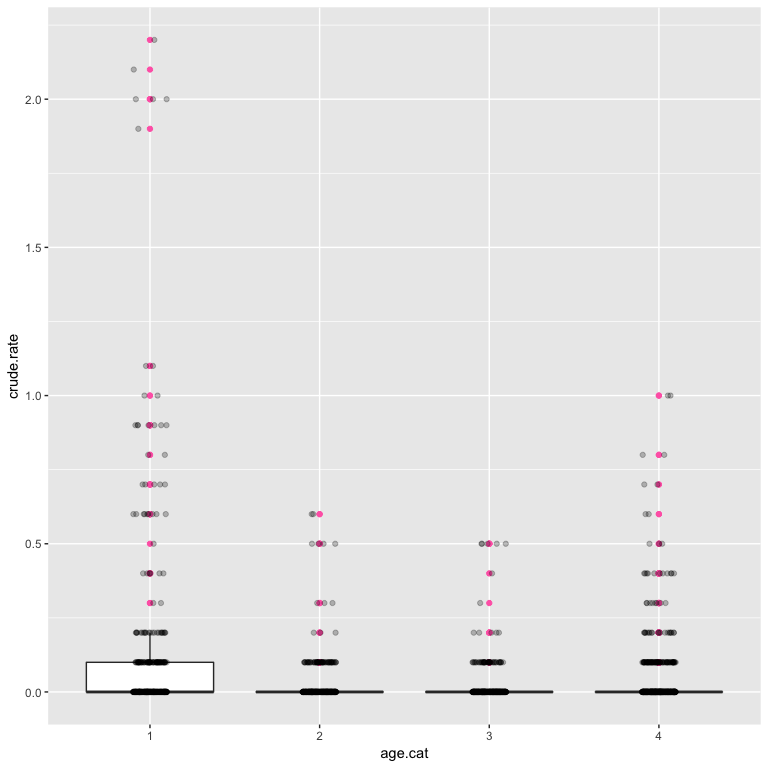


There are a lot of outliers in the one day age range. This definately isn't SIDS. Going back into the data I see this is observation in row 28, which is deaths associated with 'Certain conditions originating in the perinatal period'

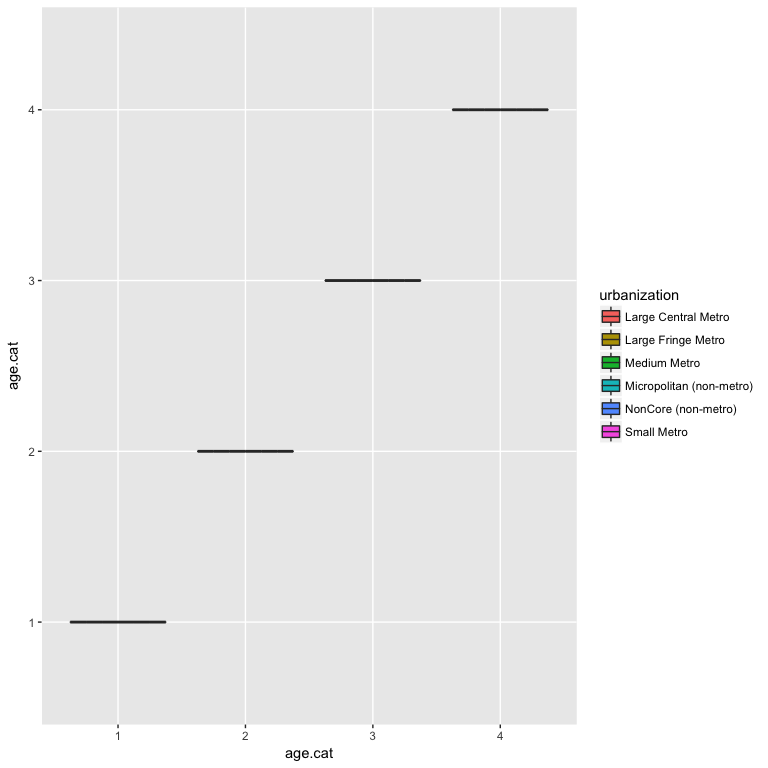
> library(ggplot2)  
> ggplot(COD, aes(x = age.cat, y = crude.rate)) +  
+ geom\_boxplot(outlier.colour = "hotpink") +  
+ geom\_jitter(position = position\_jitter(width = 0.1, height = 0), alpha = 1/4)

Warning: Removed 342 rows containing non-finite values (stat\_boxplot).

Warning: Removed 342 rows containing missing values (geom\_point).



> #same plot but adding urbanziation varible   
> ggplot(COD, aes(x=age.cat, y=age.cat, fill=urbanization)) + geom\_boxplot()



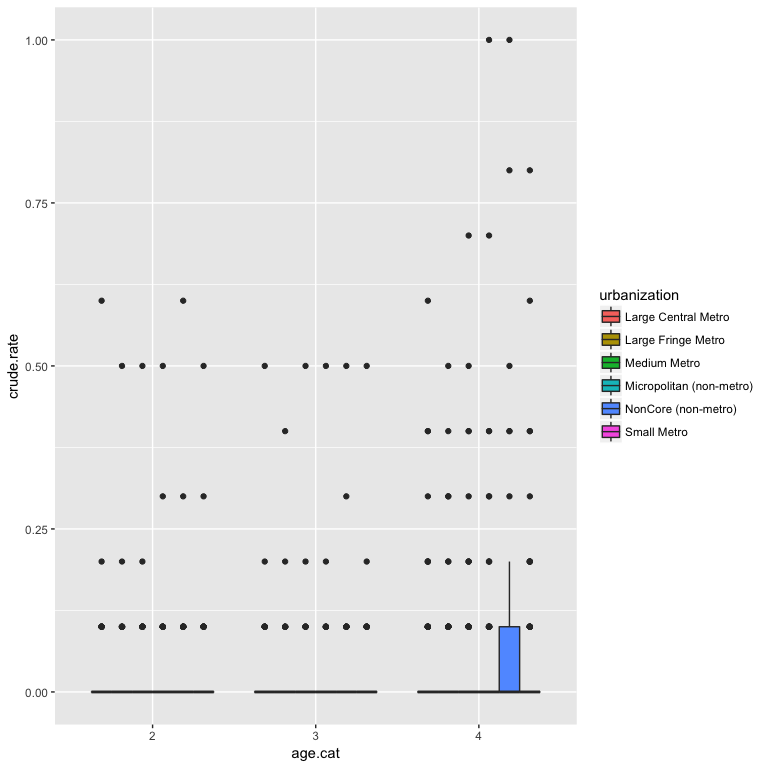
I decide I shoud look at just the day 2- 365. This is really more of the informaiton I am interested in anyways and will help remove outliers. I make a subset removing the first day deaths and explore it.

> # using subset function   
> COD2 <- subset(COD, age.cat > 1,   
+ select=c(urbanization:age.cat))  
> summary(COD2)

urbanization urban.code Infant Age Groups age.group   
 Length:1818 Min. :1.000 Length:1818 Length:1818   
 Class :character 1st Qu.:2.000 Class :character Class :character   
 Mode :character Median :3.000 Mode :character Mode :character   
 Mean :3.333   
 3rd Qu.:5.000   
 Max. :6.000   
   
 UCD - ICD-10 130 Cause List (Infants) icd   
 Length:1818 Length:1818   
 Class :character Class :character   
 Mode :character Mode :character   
   
   
   
   
 Deaths Population crude.rate age.cat   
 Min. : 10.0 Min. : 3837310 Min. :0.00000 Length:1818   
 1st Qu.: 30.0 1st Qu.: 5742486 1st Qu.:0.00000 Class :character   
 Median : 81.0 Median :14379025 Median :0.00000 Mode :character   
 Mean : 348.1 Mean :12211265 Mean :0.02879   
 3rd Qu.: 252.0 3rd Qu.:16098751 3rd Qu.:0.00000   
 Max. :14440.0 Max. :22969673 Max. :1.00000   
 NA's :255

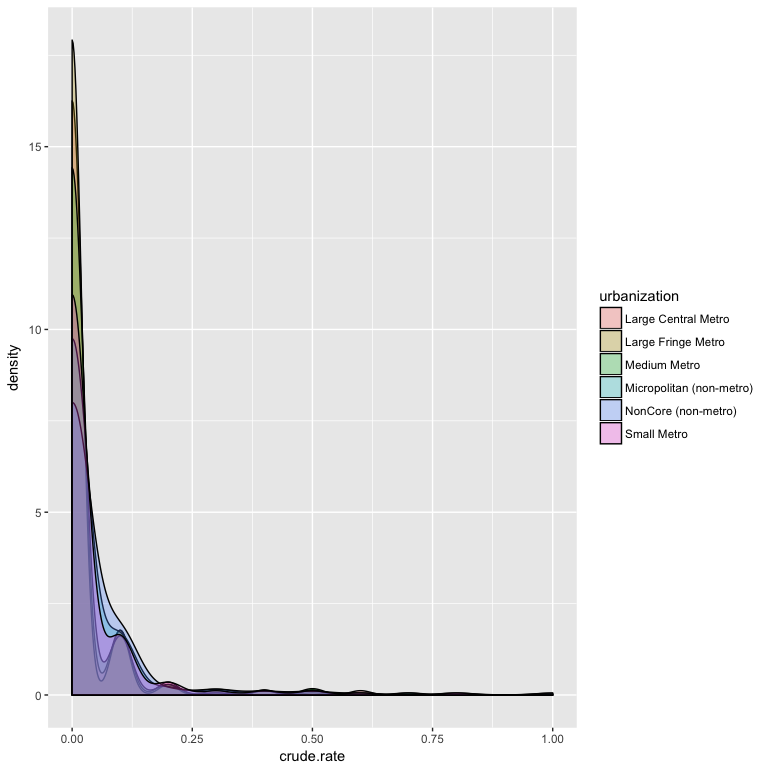
> #new plot using data with age > 1 day   
> ggplot(COD2, aes(x=age.cat, y=crude.rate, fill=urbanization)) + geom\_boxplot()

Warning: Removed 255 rows containing non-finite values (stat\_boxplot).

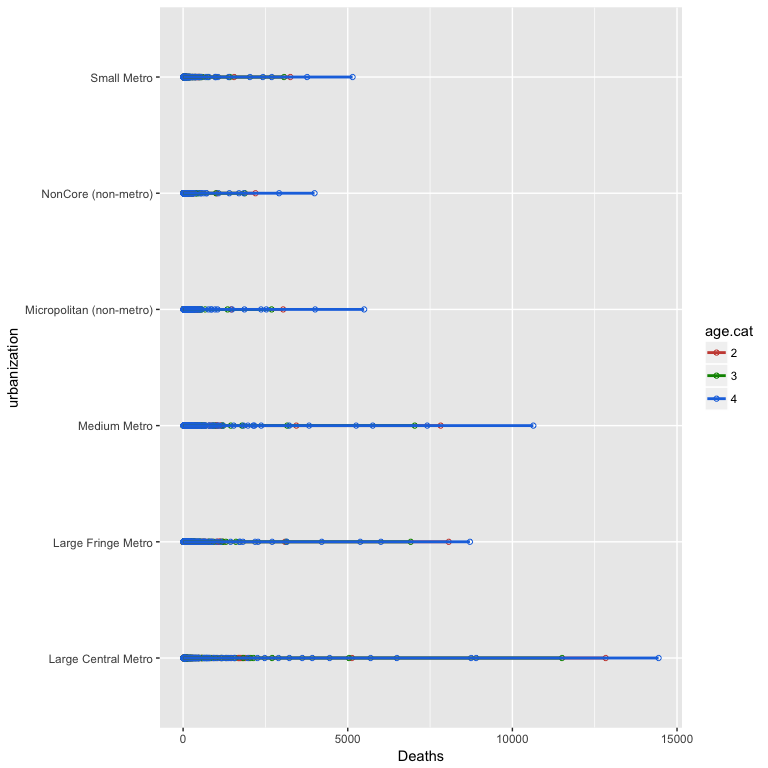


> # Density plots with semi-transparent fill  
> ggplot(COD2, aes(x=crude.rate, fill=urbanization)) + geom\_density(alpha=.3)

Warning: Removed 255 rows containing non-finite values (stat\_density).



> #scatterplot  
> ggplot(COD2, aes(x=Deaths, y=urbanization, color=age.cat)) +  
+ geom\_point(shape=1) +  
+ scale\_colour\_hue(l=50) + # Use a slightly darker palette than normal  
+ geom\_smooth(method=lm, # Add linear regression lines  
+ se=FALSE) # Don't add shaded confidence region



This is intersting but I also want to just look at the SIDS data within the overall data set. The ICD 10 code for SIDS is GR130-135 so I will use it to make a subset of the data looking at only SIDS across all time periods.

> #data looking at just SIDS  
> COD3 <- subset(COD, icd == "GR130-135",   
+ select=c(urbanization:age.cat))  
> summary(COD3)

urbanization urban.code Infant Age Groups age.group   
 Length:23 Min. :1.000 Length:23 Length:23   
 Class :character 1st Qu.:2.000 Class :character Class :character   
 Mode :character Median :3.000 Mode :character Mode :character   
 Mean :3.391   
 3rd Qu.:5.000   
 Max. :6.000   
   
 UCD - ICD-10 130 Cause List (Infants) icd Deaths   
 Length:23 Length:23 Min. : 14   
 Class :character Class :character 1st Qu.: 42   
 Mode :character Mode :character Median : 263   
 Mean :1566   
 3rd Qu.:1789   
 Max. :8746   
   
 Population crude.rate age.cat   
 Min. : 3837310 Min. :0.0000 Length:23   
 1st Qu.: 5742486 1st Qu.:0.0000 Class :character   
 Median :14379025 Median :0.0000 Mode :character   
 Mean :11862892 Mean :0.1714   
 3rd Qu.:16098751 3rd Qu.:0.4000   
 Max. :22969673 Max. :0.8000   
 NA's :2

> aggregate(crude.rate ~ age.group, COD3, mean)

age.group crude.rate  
1 1-6d 0.00000000  
2 1d 0.00000000  
3 28-364d 0.56666667  
4 7-27d 0.03333333

> aggregate(crude.rate ~ urban.code, COD3, mean)

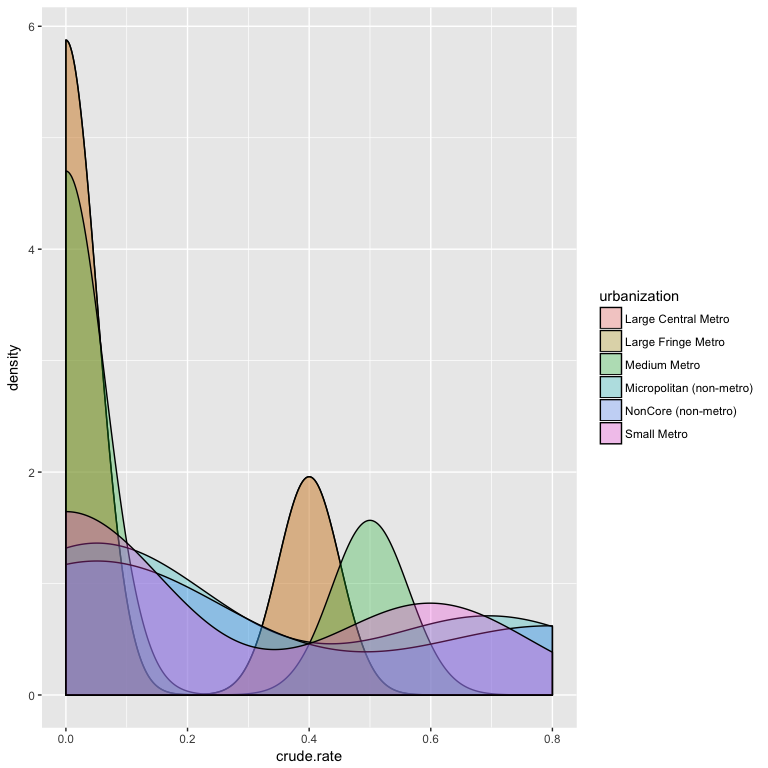
urban.code crude.rate  
1 1 0.1000000  
2 2 0.1000000  
3 3 0.1250000  
4 4 0.2000000  
5 5 0.2666667  
6 6 0.3000000

> aggregate(Deaths ~ urban.code, COD3, mean)

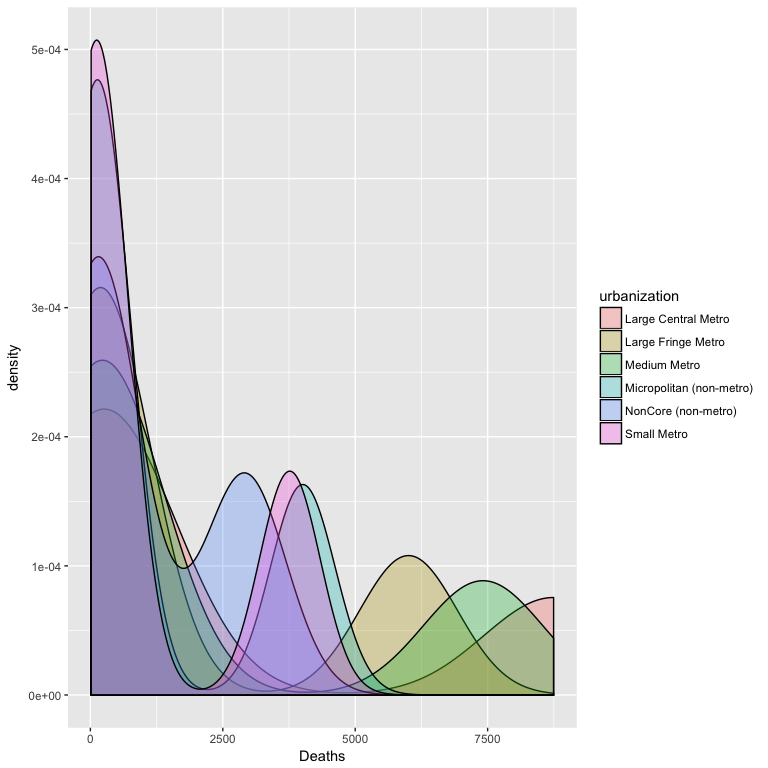
urban.code Deaths  
1 1 2386.000  
2 2 1648.000  
3 3 2028.250  
4 4 1032.500  
5 5 1104.500  
6 6 1072.333

> #explore data  
> ggplot(COD3, aes(x=crude.rate, fill=urbanization)) + geom\_density(alpha=.3)

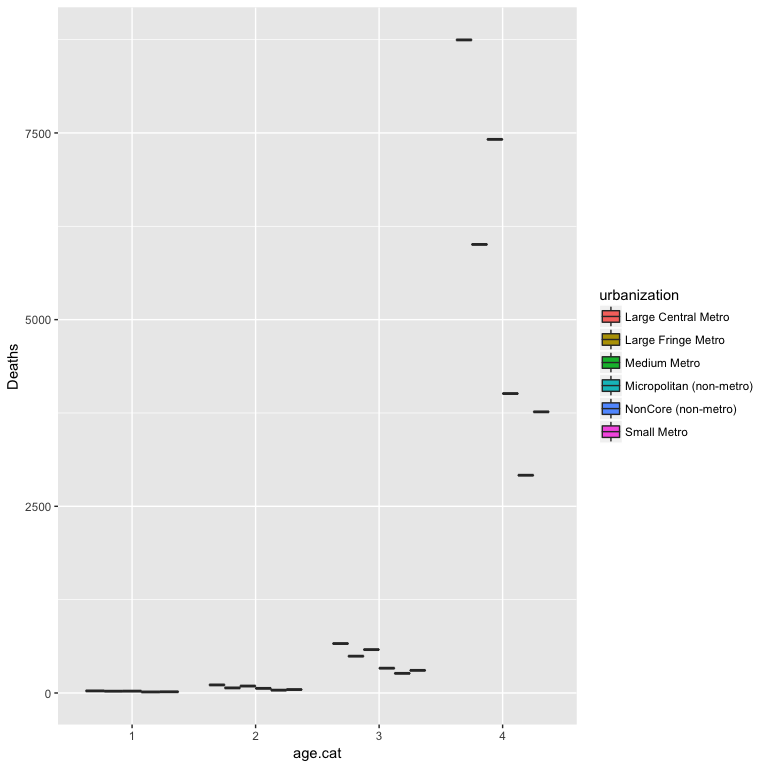
Warning: Removed 2 rows containing non-finite values (stat\_density).



> ggplot(COD3, aes(x=Deaths, fill=urbanization)) + geom\_density(alpha=.3)



> ggplot(COD3, aes(x=age.cat, y=Deaths, fill=urbanization)) + geom\_boxplot()



So I've done most of the exploratory analysis and cleaning at this point. I will start making and testing models.

> # divide the dataset into a training and a testing set based on a random uniform number on fixed seed, which in this case we are using the date  
> # this step is also creating a new variable and adding it to the data set which is a distribution of random numbers from 0 to 1   
>   
> set.seed(20170328)  
> COD3$group <- runif(length(COD3$Deaths), min = 0, max = 1)  
>   
> #what random forests do is this process over and over again and makes the aggregate which might be called bootstrapping?  
>   
> COD3.train <- subset(COD3, group <= 0.80)  
> COD3.test <- subset(COD3, group > 0.80)  
>   
> #see if it worked  
> summary(COD3.train)

urbanization urban.code Infant Age Groups age.group   
 Length:21 Min. :1.000 Length:21 Length:21   
 Class :character 1st Qu.:2.000 Class :character Class :character   
 Mode :character Median :3.000 Mode :character Mode :character   
 Mean :3.381   
 3rd Qu.:5.000   
 Max. :6.000   
   
 UCD - ICD-10 130 Cause List (Infants) icd Deaths   
 Length:21 Length:21 Min. : 14   
 Class :character Class :character 1st Qu.: 38   
 Mode :character Mode :character Median : 108   
 Mean :1413   
 3rd Qu.: 662   
 Max. :8746   
   
 Population crude.rate age.cat group   
 Min. : 3837310 Min. :0.0000 Length:21 Min. :0.04657   
 1st Qu.: 5742486 1st Qu.:0.0000 Class :character 1st Qu.:0.23218   
 Median :14379025 Median :0.0000 Mode :character Median :0.41198   
 Mean :11952633 Mean :0.1632 Mean :0.41867   
 3rd Qu.:16098751 3rd Qu.:0.2500 3rd Qu.:0.62473   
 Max. :22969673 Max. :0.8000 Max. :0.76300   
 NA's :2

> summary(COD3.test)

urbanization urban.code Infant Age Groups age.group   
 Length:2 Min. :2.00 Length:2 Length:2   
 Class :character 1st Qu.:2.75 Class :character Class :character   
 Mode :character Median :3.50 Mode :character Mode :character   
 Mean :3.50   
 3rd Qu.:4.25   
 Max. :5.00   
 UCD - ICD-10 130 Cause List (Infants) icd Deaths   
 Length:2 Length:2 Min. : 332   
 Class :character Class :character 1st Qu.:1751   
 Mode :character Mode :character Median :3170   
 Mean :3170   
 3rd Qu.:4588   
 Max. :6007   
 Population crude.rate age.cat group   
 Min. : 5742486 Min. :0.100 Length:2 Min. :0.8373   
 1st Qu.: 8331552 1st Qu.:0.175 Class :character 1st Qu.:0.8751   
 Median :10920618 Median :0.250 Mode :character Median :0.9129   
 Mean :10920618 Mean :0.250 Mean :0.9129   
 3rd Qu.:13509685 3rd Qu.:0.325 3rd Qu.:0.9507   
 Max. :16098751 Max. :0.400 Max. :0.9884

> #recoding categorical variables as factors  
> COD3.train$urbanization.f <- factor(COD3.train$urbanization)  
> is.factor(COD3.train$urbanization.f)

[1] TRUE

> #is urbanization a predictor varaible?  
> fit <- lm(Deaths ~ urbanization.f, data= COD3.train)  
> summary(fit)

Call:  
lm(formula = Deaths ~ urbanization.f, data = COD3.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2358.0 -1447.2 -986.5 298.0 6360.0   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) 2386.0 1407.2 1.696  
urbanization.fLarge Fringe Metro -2191.0 2149.5 -1.019  
urbanization.fMedium Metro -357.8 1990.1 -0.180  
urbanization.fMicropolitan (non-metro) -1024.0 2149.5 -0.476  
urbanization.fNonCore (non-metro) -1313.7 2149.5 -0.611  
urbanization.fSmall Metro -1353.5 1990.1 -0.680  
 Pr(>|t|)  
(Intercept) 0.111  
urbanization.fLarge Fringe Metro 0.324  
urbanization.fMedium Metro 0.860  
urbanization.fMicropolitan (non-metro) 0.641  
urbanization.fNonCore (non-metro) 0.550  
urbanization.fSmall Metro 0.507  
  
Residual standard error: 2814 on 15 degrees of freedom  
Multiple R-squared: 0.08252, Adjusted R-squared: -0.2233   
F-statistic: 0.2698 on 5 and 15 DF, p-value: 0.9226

Well none of this is significant. There were MANY fewer observation in the SIDS subset than in the whole thing so that may be contributing and I will adress this in my limitation section. I may go back and re-code some of the other 'unknown' deaths that are considered SIDS but not classified as so to see if that gives more data but I have not decided quite yet.

> #now looking at age category  
> is.factor(COD3$age.cat)

[1] FALSE

> #recoding categorical variables as factors  
> COD3.train$age.f <- factor(COD3.train$age.cat)  
> is.factor(COD3.train$age.f)

[1] TRUE

> #fit a model with all variables  
> fit2 <- lm(Deaths ~ age.f, data=COD3.train)  
> summary(fit2)# show results

Call:  
lm(formula = Deaths ~ age.f, data = COD3.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2454.2 -31.2 -1.2 32.6 3375.8   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 21.40 554.63 0.039 0.970   
age.f2 47.77 750.97 0.064 0.950   
age.f3 439.00 784.36 0.560 0.583   
age.f4 5348.80 784.36 6.819 2.99e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1240 on 17 degrees of freedom  
Multiple R-squared: 0.7981, Adjusted R-squared: 0.7625   
F-statistic: 22.4 on 3 and 17 DF, p-value: 3.856e-06

Age category is a significant variable. This, of course, makes tons of sense because age category 4 is much longer time (almost the whole year). Does anything change when we put them together?

> #fit a model with all variables  
> fit3 <- lm(Deaths ~ age.f + urbanization.f, data=COD3.train)  
> summary(fit3)# show results

Call:  
lm(formula = Deaths ~ age.f + urbanization.f, data = COD3.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-1599.93 -681.99 17.21 557.94 2391.59   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) 811.6 775.8 1.046  
age.f2 222.5 747.2 0.298  
age.f3 532.4 795.5 0.669  
age.f4 5542.9 795.5 6.968  
urbanization.fLarge Fringe Metro -868.2 941.9 -0.922  
urbanization.fMedium Metro -357.8 859.2 -0.416  
urbanization.fMicropolitan (non-metro) -1371.3 941.9 -1.456  
urbanization.fNonCore (non-metro) -1838.5 941.9 -1.952  
urbanization.fSmall Metro -1353.5 859.2 -1.575  
 Pr(>|t|)   
(Intercept) 0.3161   
age.f2 0.7710   
age.f3 0.5160   
age.f4 1.5e-05 \*\*\*  
urbanization.fLarge Fringe Metro 0.3748   
urbanization.fMedium Metro 0.6845   
urbanization.fMicropolitan (non-metro) 0.1711   
urbanization.fNonCore (non-metro) 0.0747 .   
urbanization.fSmall Metro 0.1412   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1215 on 12 degrees of freedom  
Multiple R-squared: 0.8632, Adjusted R-squared: 0.772   
F-statistic: 9.463 on 8 and 12 DF, p-value: 0.0003795

Well my theory is disproven. There does not seem to be any association between SIDS and urbanization. Let me try modeling with the whole data set just because I have it already. Does the association change when we look at all causes of death for infants between 2-364 days of age?

> # divide the dataset into a training and a testing set based on a random uniform number on fixed seed, which in this case we are using the date  
> # this step is also creating a new variable and adding it to the data set which is a distribution of random numbers from 0 to 1   
>   
> set.seed(20170328)  
> COD2$group <- runif(length(COD2$Deaths), min = 0, max = 1)  
>   
> #creating the train and test dataset  
>   
> COD2.train <- subset(COD2, group <= 0.80)  
> COD2.test <- subset(COD2, group > 0.80)  
>   
> #see if it worked  
> summary(COD2.train)

urbanization urban.code Infant Age Groups age.group   
 Length:1448 Min. :1.000 Length:1448 Length:1448   
 Class :character 1st Qu.:2.000 Class :character Class :character   
 Mode :character Median :3.000 Mode :character Mode :character   
 Mean :3.311   
 3rd Qu.:5.000   
 Max. :6.000   
   
 UCD - ICD-10 130 Cause List (Infants) icd   
 Length:1448 Length:1448   
 Class :character Class :character   
 Mode :character Mode :character   
   
   
   
   
 Deaths Population crude.rate age.cat   
 Min. : 10.0 Min. : 3837310 Min. :0.00000 Length:1448   
 1st Qu.: 30.0 1st Qu.: 5742486 1st Qu.:0.00000 Class :character   
 Median : 83.0 Median :14379025 Median :0.00000 Mode :character   
 Mean : 355.3 Mean :12264034 Mean :0.02942   
 3rd Qu.: 259.5 3rd Qu.:16098751 3rd Qu.:0.00000   
 Max. :12833.0 Max. :22969673 Max. :1.00000   
 NA's :204   
 group   
 Min. :0.0002761   
 1st Qu.:0.2118207   
 Median :0.4170567   
 Mean :0.4123271   
 3rd Qu.:0.6158791   
 Max. :0.7983744

> summary(COD2.test)

urbanization urban.code Infant Age Groups age.group   
 Length:370 Min. :1.000 Length:370 Length:370   
 Class :character 1st Qu.:2.000 Class :character Class :character   
 Mode :character Median :3.000 Mode :character Mode :character   
 Mean :3.416   
 3rd Qu.:5.000   
 Max. :6.000   
   
 UCD - ICD-10 130 Cause List (Infants) icd   
 Length:370 Length:370   
 Class :character Class :character   
 Mode :character Mode :character   
   
   
   
   
 Deaths Population crude.rate age.cat   
 Min. : 10.0 Min. : 3837310 Min. :0.00000 Length:370   
 1st Qu.: 33.0 1st Qu.: 5742486 1st Qu.:0.00000 Class :character   
 Median : 73.5 Median :14379025 Median :0.00000 Mode :character   
 Mean : 320.3 Mean :12004751 Mean :0.02633   
 3rd Qu.: 222.5 3rd Qu.:16098751 3rd Qu.:0.00000   
 Max. :14440.0 Max. :22969673 Max. :0.60000   
 NA's :51   
 group   
 Min. :0.8001   
 1st Qu.:0.8521   
 Median :0.9008   
 Mean :0.9004   
 3rd Qu.:0.9496   
 Max. :0.9993

> #are my categorical variables factors?  
> is.factor(COD2$urbanization) #no this was not

[1] FALSE

> #recoding categorical variables as factors  
> COD2.train$urbanization.f <- factor(COD2.train$urbanization)  
> is.factor(COD2.train$urbanization.f)

[1] TRUE

> #is urbanization a predictor varaible?  
> fit3 <- lm(Deaths ~ urbanization.f, data= COD2.train)  
> summary(fit3)

Call:  
lm(formula = Deaths ~ urbanization.f, data = COD2.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -571.2 -364.5 -175.9 -65.7 12251.8   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) 581.18 58.77 9.889  
urbanization.fLarge Fringe Metro -150.00 85.10 -1.763  
urbanization.fMedium Metro -148.88 85.54 -1.740  
urbanization.fMicropolitan (non-metro) -367.25 88.98 -4.127  
urbanization.fNonCore (non-metro) -411.71 91.20 -4.514  
urbanization.fSmall Metro -374.24 87.36 -4.284  
 Pr(>|t|)   
(Intercept) < 2e-16 \*\*\*  
urbanization.fLarge Fringe Metro 0.0782 .   
urbanization.fMedium Metro 0.0820 .   
urbanization.fMicropolitan (non-metro) 3.88e-05 \*\*\*  
urbanization.fNonCore (non-metro) 6.87e-06 \*\*\*  
urbanization.fSmall Metro 1.96e-05 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 988.7 on 1442 degrees of freedom  
Multiple R-squared: 0.02348, Adjusted R-squared: 0.02009   
F-statistic: 6.935 on 5 and 1442 DF, p-value: 2.076e-06

Great so there is some significance here. I will explore/ explain this more in the final report.

> #recoding categorical variables as factors  
> COD2.train$age.f <- factor(COD2.train$age.cat)  
> is.factor(COD2.train$age.f)

[1] TRUE

> #fit a model with all variables  
> fit4 <- lm(Deaths ~ age.f, data=COD2.train)  
> summary(fit4)# show results

Call:  
lm(formula = Deaths ~ age.f, data = COD2.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -470.3 -288.5 -225.4 -92.4 12571.4   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 261.57 47.67 5.488 4.81e-08 \*\*\*  
age.f3 20.37 67.49 0.302 0.762831   
age.f4 218.71 63.06 3.468 0.000539 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 994.2 on 1445 degrees of freedom  
Multiple R-squared: 0.01054, Adjusted R-squared: 0.009169   
F-statistic: 7.695 on 2 and 1445 DF, p-value: 0.000474

Look at everything.

> #fit a model with all variables  
> fit5 <- lm(Deaths ~ age.f + urbanization.f, data=COD2.train)  
> summary(fit5)# show results

Call:  
lm(formula = Deaths ~ age.f + urbanization.f, data = COD2.train)  
  
Residuals:  
 Min 1Q Median 3Q Max   
 -693.4 -324.3 -191.8 -33.2 12349.6   
  
Coefficients:  
 Estimate Std. Error t value  
(Intercept) 483.37 70.04 6.901  
age.f3 28.05 66.88 0.419  
age.f4 220.01 62.43 3.524  
urbanization.fLarge Fringe Metro -149.03 84.72 -1.759  
urbanization.fMedium Metro -147.00 85.20 -1.725  
urbanization.fMicropolitan (non-metro) -366.06 88.64 -4.130  
urbanization.fNonCore (non-metro) -408.88 90.79 -4.503  
urbanization.fSmall Metro -372.84 86.96 -4.288  
 Pr(>|t|)   
(Intercept) 7.69e-12 \*\*\*  
age.f3 0.674942   
age.f4 0.000438 \*\*\*  
urbanization.fLarge Fringe Metro 0.078778 .   
urbanization.fMedium Metro 0.084653 .   
urbanization.fMicropolitan (non-metro) 3.84e-05 \*\*\*  
urbanization.fNonCore (non-metro) 7.23e-06 \*\*\*  
urbanization.fSmall Metro 1.93e-05 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 984.1 on 1440 degrees of freedom  
Multiple R-squared: 0.03382, Adjusted R-squared: 0.02912   
F-statistic: 7.201 on 7 and 1440 DF, p-value: 1.706e-08

Discussion. This is where the discussion will go!

The git hub for this is located at (<https://github.com/taylorthul/N741-Urban-SIDS-/blob/master/SIDS%20Data.Rmd>) [<https://github.com/taylorthul/N741-Urban-SIDS-/blob/master/SIDS%20Data.Rmd>]