NCHS

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

NCHS\_New1 <- read.csv("NCHS.csv")  
str(NCHS\_New1)

## 'data.frame': 205920 obs. of 13 variables:  
## $ Year : int 2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...  
## $ Cause.of.Death : Factor w/ 5 levels "Cancer","Chronic Lower Respiratory Disease",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ State : Factor w/ 52 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ State.FIPS.Code : Factor w/ 52 levels "0","AK","AL",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ HHS.Region : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ Age.Range : Factor w/ 8 levels "0-49","0-54",..: 1 1 1 1 1 1 1 1 1 2 ...  
## $ Benchmark : Factor w/ 3 levels "2005 Fixed","2010 Fixed",..: 1 1 1 2 2 2 3 3 3 1 ...  
## $ Locality : Factor w/ 3 levels "All","Metropolitan",..: 1 2 3 1 2 3 1 2 3 1 ...  
## $ Observed.Deaths : int 756 556 200 756 556 200 756 556 200 1346 ...  
## $ Population : int 3148377 2379871 768506 3148377 2379871 768506 3148377 2379871 768506 3463216 ...  
## $ Expected.Deaths : int 451 341 111 421 318 103 451 341 111 784 ...  
## $ Potentially.Excess.Deaths : int 305 217 89 335 238 97 305 217 89 562 ...  
## $ Percent.Potentially.Excess.Deaths: Factor w/ 841 levels "","0%","0.10%",..: 355 332 397 395 380 437 355 332 397 370 ...

summary(NCHS\_New1)

## Year Cause.of.Death   
## Min. :2005 Cancer :41184   
## 1st Qu.:2007 Chronic Lower Respiratory Disease:41184   
## Median :2010 Heart Disease :41184   
## Mean :2010 Stroke :41184   
## 3rd Qu.:2013 Unintentional Injury :41184   
## Max. :2015   
##   
## State State.FIPS.Code HHS.Region Age.Range   
## Alabama : 3960 0 : 3960 Min. : 0.000 0-49 :25740   
## Alaska : 3960 AK : 3960 1st Qu.: 3.000 0-54 :25740   
## Arizona : 3960 AL : 3960 Median : 5.000 0-59 :25740   
## Arkansas : 3960 AR : 3960 Mean : 5.231 0-64 :25740   
## California: 3960 AZ : 3960 3rd Qu.: 8.000 0-69 :25740   
## Colorado : 3960 CA : 3960 Max. :10.000 0-74 :25740   
## (Other) :182160 (Other):182160 (Other):51480   
## Benchmark Locality Observed.Deaths   
## 2005 Fixed:68640 All :68640 Min. : 10   
## 2010 Fixed:68640 Metropolitan :68640 1st Qu.: 155   
## Floating :68640 Nonmetropolitan:68640 Median : 508   
## Mean : 2975   
## 3rd Qu.: 1586   
## Max. :493526   
## NA's :10212   
## Population Expected.Deaths Potentially.Excess.Deaths  
## Min. : 55536 Min. : 2 Min. : 0.0   
## 1st Qu.: 682826 1st Qu.: 92 1st Qu.: 41.0   
## Median : 1610936 Median : 298 Median : 159.0   
## Mean : 7008421 Mean : 2113 Mean : 875.1   
## 3rd Qu.: 4681585 3rd Qu.: 1019 3rd Qu.: 543.0   
## Max. :315131659 Max. :465126 Max. :175703.0   
## NA's :5280 NA's :10212 NA's :10212   
## Percent.Potentially.Excess.Deaths  
## : 10212   
## 0% : 2492   
## 50% : 1034   
## 33.30% : 874   
## 25% : 678   
## 40% : 671   
## (Other):189959

sum(is.na(NCHS\_New1))

## [1] 35916

aft\_mis\_NCHS\_New1<-na.omit(NCHS\_New1)  
sum(is.na(aft\_mis\_NCHS\_New1))

## [1] 0

str(aft\_mis\_NCHS\_New1)

## 'data.frame': 195708 obs. of 13 variables:  
## $ Year : int 2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...  
## $ Cause.of.Death : Factor w/ 5 levels "Cancer","Chronic Lower Respiratory Disease",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ State : Factor w/ 52 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ State.FIPS.Code : Factor w/ 52 levels "0","AK","AL",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ HHS.Region : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ Age.Range : Factor w/ 8 levels "0-49","0-54",..: 1 1 1 1 1 1 1 1 1 2 ...  
## $ Benchmark : Factor w/ 3 levels "2005 Fixed","2010 Fixed",..: 1 1 1 2 2 2 3 3 3 1 ...  
## $ Locality : Factor w/ 3 levels "All","Metropolitan",..: 1 2 3 1 2 3 1 2 3 1 ...  
## $ Observed.Deaths : int 756 556 200 756 556 200 756 556 200 1346 ...  
## $ Population : int 3148377 2379871 768506 3148377 2379871 768506 3148377 2379871 768506 3463216 ...  
## $ Expected.Deaths : int 451 341 111 421 318 103 451 341 111 784 ...  
## $ Potentially.Excess.Deaths : int 305 217 89 335 238 97 305 217 89 562 ...  
## $ Percent.Potentially.Excess.Deaths: Factor w/ 841 levels "","0%","0.10%",..: 355 332 397 395 380 437 355 332 397 370 ...  
## - attr(\*, "na.action")=Class 'omit' Named int [1:10212] 507 510 513 516 519 522 525 528 531 534 ...  
## .. ..- attr(\*, "names")= chr [1:10212] "507" "510" "513" "516" ...

#duplicated(NCHS\_New1)  
str(NCHS\_New1)

## 'data.frame': 205920 obs. of 13 variables:  
## $ Year : int 2005 2005 2005 2005 2005 2005 2005 2005 2005 2005 ...  
## $ Cause.of.Death : Factor w/ 5 levels "Cancer","Chronic Lower Respiratory Disease",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ State : Factor w/ 52 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ State.FIPS.Code : Factor w/ 52 levels "0","AK","AL",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ HHS.Region : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ Age.Range : Factor w/ 8 levels "0-49","0-54",..: 1 1 1 1 1 1 1 1 1 2 ...  
## $ Benchmark : Factor w/ 3 levels "2005 Fixed","2010 Fixed",..: 1 1 1 2 2 2 3 3 3 1 ...  
## $ Locality : Factor w/ 3 levels "All","Metropolitan",..: 1 2 3 1 2 3 1 2 3 1 ...  
## $ Observed.Deaths : int 756 556 200 756 556 200 756 556 200 1346 ...  
## $ Population : int 3148377 2379871 768506 3148377 2379871 768506 3148377 2379871 768506 3463216 ...  
## $ Expected.Deaths : int 451 341 111 421 318 103 451 341 111 784 ...  
## $ Potentially.Excess.Deaths : int 305 217 89 335 238 97 305 217 89 562 ...  
## $ Percent.Potentially.Excess.Deaths: Factor w/ 841 levels "","0%","0.10%",..: 355 332 397 395 380 437 355 332 397 370 ...

aft\_mis\_NCHS\_New2<- aft\_mis\_NCHS\_New1[!(aft\_mis\_NCHS\_New1$State == "United States"),]  
dat <- aft\_mis\_NCHS\_New2  
head(dat)

## Year Cause.of.Death State State.FIPS.Code HHS.Region Age.Range  
## 1 2005 Cancer Alabama AL 4 0-49  
## 2 2005 Cancer Alabama AL 4 0-49  
## 3 2005 Cancer Alabama AL 4 0-49  
## 4 2005 Cancer Alabama AL 4 0-49  
## 5 2005 Cancer Alabama AL 4 0-49  
## 6 2005 Cancer Alabama AL 4 0-49  
## Benchmark Locality Observed.Deaths Population Expected.Deaths  
## 1 2005 Fixed All 756 3148377 451  
## 2 2005 Fixed Metropolitan 556 2379871 341  
## 3 2005 Fixed Nonmetropolitan 200 768506 111  
## 4 2010 Fixed All 756 3148377 421  
## 5 2010 Fixed Metropolitan 556 2379871 318  
## 6 2010 Fixed Nonmetropolitan 200 768506 103  
## Potentially.Excess.Deaths Percent.Potentially.Excess.Deaths  
## 1 305 40.30%  
## 2 217 39%  
## 3 89 44.50%  
## 4 335 44.30%  
## 5 238 42.80%  
## 6 97 48.50%

library(noncensus)

## Warning: package 'noncensus' was built under R version 3.3.2

data("states")  
states

## state name region division  
## 1 AK Alaska West Pacific  
## 2 AL Alabama South East South Central  
## 3 AR Arkansas South West South Central  
## 4 AZ Arizona West Mountain  
## 5 CA California West Pacific  
## 6 CO Colorado West Mountain  
## 7 CT Connecticut Northeast New England  
## 8 DC District of Columbia South South Atlantic  
## 9 DE Delaware South South Atlantic  
## 10 FL Florida South South Atlantic  
## 11 GA Georgia South South Atlantic  
## 12 HI Hawaii West Pacific  
## 13 IA Iowa Midwest West North Central  
## 14 ID Idaho West Mountain  
## 15 IL Illinois Midwest East North Central  
## 16 IN Indiana Midwest East North Central  
## 17 KS Kansas Midwest West North Central  
## 18 KY Kentucky South East South Central  
## 19 LA Louisiana South West South Central  
## 20 MA Massachusetts Northeast New England  
## 21 MD Maryland South South Atlantic  
## 22 ME Maine Northeast New England  
## 23 MI Michigan Midwest East North Central  
## 24 MN Minnesota Midwest West North Central  
## 25 MO Missouri Midwest West North Central  
## 26 MS Mississippi South East South Central  
## 27 MT Montana West Mountain  
## 28 NC North Carolina South South Atlantic  
## 29 ND North Dakota Midwest West North Central  
## 30 NE Nebraska Midwest West North Central  
## 31 NH New Hampshire Northeast New England  
## 32 NJ New Jersey Northeast Mid-Atlantic  
## 33 NM New Mexico West Mountain  
## 34 NV Nevada West Mountain  
## 35 NY New York Northeast Mid-Atlantic  
## 36 OH Ohio Midwest East North Central  
## 37 OK Oklahoma South West South Central  
## 38 OR Oregon West Pacific  
## 39 PA Pennsylvania Northeast Mid-Atlantic  
## 40 RI Rhode Island Northeast New England  
## 41 SC South Carolina South South Atlantic  
## 42 SD South Dakota Midwest West North Central  
## 43 TN Tennessee South East South Central  
## 44 TX Texas South West South Central  
## 45 UT Utah West Mountain  
## 46 VA Virginia South South Atlantic  
## 47 VT Vermont Northeast New England  
## 48 WA Washington West Pacific  
## 49 WI Wisconsin Midwest East North Central  
## 50 WV West Virginia South South Atlantic  
## 51 WY Wyoming West Mountain  
## 52 AS American Samoa <NA> <NA>  
## 53 GU Guam <NA> <NA>  
## 54 MP Northern Mariana Islands <NA> <NA>  
## 55 PR Puerto Rico <NA> <NA>  
## 56 VI U.S. Virgin Islands <NA> <NA>  
## capital area population  
## 1 Juneau 589757 4779736  
## 2 Montgomery 51609 710231  
## 3 Little Rock 53104 6392017  
## 4 Phoenix 113909 2915918  
## 5 Sacramento 158693 37253956  
## 6 Denver 104247 5029196  
## 7 Hartford 5009 3574097  
## 8 <NA> 68.3 897934  
## 9 Dover 2057 601723  
## 10 Tallahassee 58560 18801310  
## 11 Atlanta 58876 9687653  
## 12 Honolulu 6450 1360301  
## 13 Des Moines 56290 1567582  
## 14 Boise 83557 12830632  
## 15 Springfield 56400 6483802  
## 16 Indianapolis 36291 3046355  
## 17 Topeka 82264 2853118  
## 18 Frankfort 40395 4339367  
## 19 Baton Rouge 48523 4533372  
## 20 Boston 8257 1328361  
## 21 Annapolis 10577 5773552  
## 22 Augusta 33215 6547629  
## 23 Lansing 58216 9883640  
## 24 Saint Paul 84068 5303925  
## 25 Jefferson City 69686 2967297  
## 26 Jackson 47716 5988927  
## 27 Helena 147138 989415  
## 28 Raleigh 52586 1826341  
## 29 Bismarck 70665 2700551  
## 30 Lincoln 77227 1316470  
## 31 Concord 9304 8791894  
## 32 Trenton 7836 2059179  
## 33 Santa Fe 121666 19378102  
## 34 Carson City 110540 9535483  
## 35 Albany 49576 672591  
## 36 Columbus 41222 11536504  
## 37 Oklahoma City 69919 3751351  
## 38 Salem 96981 3831074  
## 39 Harrisburg 45333 12702379  
## 40 Providence 1214 1052567  
## 41 Columbia 31055 4625364  
## 42 Pierre 77047 814180  
## 43 Nashville 42244 6346105  
## 44 Austin 267339 25145561  
## 45 Salt Lake City 84916 2763885  
## 46 Richmond 40815 625741  
## 47 Montpelier 9609 8001024  
## 48 Olympia 68192 6724540  
## 49 Madison 56154 1852994  
## 50 Charleston 24181 5686986  
## 51 Cheyenne 97914 563626  
## 52 Pago Pago 76.8 55519  
## 53 Hagåtña 212 159358  
## 54 Saipan 179.01 53833  
## 55 San Juan 3515 3725789  
## 56 Charlotte Amalie 133.73 106405

Reg <- states[,c(1,3)]  
Reg

## state region  
## 1 AK West  
## 2 AL South  
## 3 AR South  
## 4 AZ West  
## 5 CA West  
## 6 CO West  
## 7 CT Northeast  
## 8 DC South  
## 9 DE South  
## 10 FL South  
## 11 GA South  
## 12 HI West  
## 13 IA Midwest  
## 14 ID West  
## 15 IL Midwest  
## 16 IN Midwest  
## 17 KS Midwest  
## 18 KY South  
## 19 LA South  
## 20 MA Northeast  
## 21 MD South  
## 22 ME Northeast  
## 23 MI Midwest  
## 24 MN Midwest  
## 25 MO Midwest  
## 26 MS South  
## 27 MT West  
## 28 NC South  
## 29 ND Midwest  
## 30 NE Midwest  
## 31 NH Northeast  
## 32 NJ Northeast  
## 33 NM West  
## 34 NV West  
## 35 NY Northeast  
## 36 OH Midwest  
## 37 OK South  
## 38 OR West  
## 39 PA Northeast  
## 40 RI Northeast  
## 41 SC South  
## 42 SD Midwest  
## 43 TN South  
## 44 TX South  
## 45 UT West  
## 46 VA South  
## 47 VT Northeast  
## 48 WA West  
## 49 WI Midwest  
## 50 WV South  
## 51 WY West  
## 52 AS <NA>  
## 53 GU <NA>  
## 54 MP <NA>  
## 55 PR <NA>  
## 56 VI <NA>

Health\_data <- merge(dat,Reg,by.x = "State.FIPS.Code",by.y = "state")  
head(Health\_data)

## State.FIPS.Code Year Cause.of.Death State HHS.Region  
## 1 AK 2006 Heart Disease Alaska 10  
## 2 AK 2008 Stroke Alaska 10  
## 3 AK 2015 Cancer Alaska 10  
## 4 AK 2007 Heart Disease Alaska 10  
## 5 AK 2015 Cancer Alaska 10  
## 6 AK 2011 Chronic Lower Respiratory Disease Alaska 10  
## Age.Range Benchmark Locality Observed.Deaths Population  
## 1 0-49 2010 Fixed Nonmetropolitan 19 163571  
## 2 0-59 Floating Metropolitan 24 408596  
## 3 0-64 2005 Fixed All 390 665595  
## 4 0-59 2010 Fixed All 158 603065  
## 5 0-59 Floating Nonmetropolitan 88 189260  
## 6 0-64 2010 Fixed Metropolitan 24 449953  
## Expected.Deaths Potentially.Excess.Deaths  
## 1 15 4  
## 2 11 13  
## 3 353 43  
## 4 121 37  
## 5 63 27  
## 6 18 8  
## Percent.Potentially.Excess.Deaths region  
## 1 21.10% West  
## 2 54.20% West  
## 3 11% West  
## 4 23.40% West  
## 5 30.70% West  
## 6 33.30% West

# write.csv(Health\_data,"Health data.csv")  
  
# 1.trend of all the five leading cause of Observed deaths over time for each region  
  
Data\_all <- Health\_data[,c(2:3,9,14)]  
head(Data\_all)

## Year Cause.of.Death Observed.Deaths region  
## 1 2006 Heart Disease 19 West  
## 2 2008 Stroke 24 West  
## 3 2015 Cancer 390 West  
## 4 2007 Heart Disease 158 West  
## 5 2015 Cancer 88 West  
## 6 2011 Chronic Lower Respiratory Disease 24 West

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.3.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

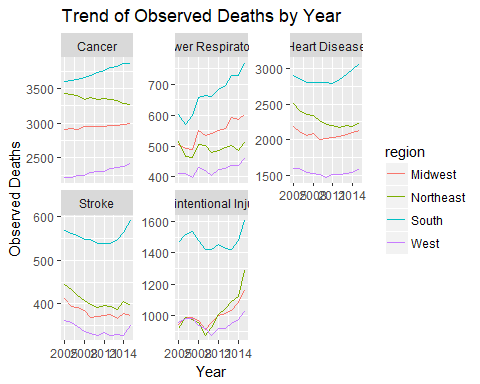
Data\_allNew <- Data\_all %>%  
 group\_by(region, Year, Cause.of.Death) %>%  
 summarise\_all(funs(mean))  
Data\_allNew

## # A tibble: 220 x 4  
## # Groups: region, Year [?]  
## region Year Cause.of.Death Observed.Deaths  
## <fctr> <int> <fctr> <dbl>  
## 1 Midwest 2005 Cancer 2908.1042  
## 2 Midwest 2005 Chronic Lower Respiratory Disease 503.8199  
## 3 Midwest 2005 Heart Disease 2185.6042  
## 4 Midwest 2005 Stroke 413.0691  
## 5 Midwest 2005 Unintentional Injury 943.3194  
## 6 Midwest 2006 Cancer 2925.4236  
## 7 Midwest 2006 Chronic Lower Respiratory Disease 493.5038  
## 8 Midwest 2006 Heart Disease 2121.4931  
## 9 Midwest 2006 Stroke 394.5018  
## 10 Midwest 2006 Unintentional Injury 986.9375  
## # ... with 210 more rows

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.3

ggplot(data = Data\_allNew, aes(x=Year, y = Observed.Deaths, colour = region, group = region)) + labs(title="Trend of Observed Deaths by Year", x="Year",y="Observed Deaths") +  
 geom\_line()+scale\_x\_continuous(breaks=seq(2005, 2015, 3)) + facet\_wrap(~Cause.of.Death,scales = "free\_y")



# 2. Max and Min value of all types of deaths in Each age group  
  
Data\_sec <- Health\_data[,c(3,6,9,11,12)]  
str(Data\_sec)

## 'data.frame': 191748 obs. of 5 variables:  
## $ Cause.of.Death : Factor w/ 5 levels "Cancer","Chronic Lower Respiratory Disease",..: 3 4 1 3 1 2 5 1 3 4 ...  
## $ Age.Range : Factor w/ 8 levels "0-49","0-54",..: 1 3 4 3 3 4 3 1 1 1 ...  
## $ Observed.Deaths : int 19 24 390 158 88 24 259 46 19 12 ...  
## $ Expected.Deaths : int 15 11 353 121 63 18 104 48 16 5 ...  
## $ Potentially.Excess.Deaths: int 4 13 43 37 27 8 155 1 3 7 ...

library(dplyr)  
Data\_sec\_New <- Data\_sec %>%  
 group\_by(Age.Range, Cause.of.Death) %>%  
 summarise\_all(funs(mean))  
Data\_sec\_New

## # A tibble: 40 x 5  
## # Groups: Age.Range [?]  
## Age.Range Cause.of.Death Observed.Deaths  
## <fctr> <fctr> <dbl>  
## 1 0-49 Cancer 485.26151  
## 2 0-49 Chronic Lower Respiratory Disease 52.27354  
## 3 0-49 Heart Disease 404.80074  
## 4 0-49 Stroke 86.90902  
## 5 0-49 Unintentional Injury 799.16290  
## 6 0-54 Cancer 903.57413  
## 7 0-54 Chronic Lower Respiratory Disease 90.56129  
## 8 0-54 Heart Disease 707.54162  
## 9 0-54 Stroke 132.63768  
## 10 0-54 Unintentional Injury 934.75290  
## # ... with 30 more rows, and 2 more variables: Expected.Deaths <dbl>,  
## # Potentially.Excess.Deaths <dbl>

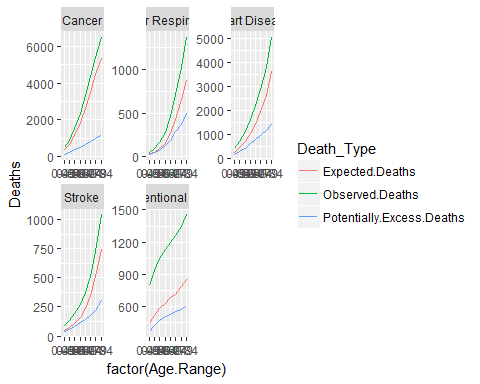
library(tidyr)

## Warning: package 'tidyr' was built under R version 3.3.3

long\_data\_sec\_New <- gather(Data\_sec\_New,"Death\_Type", "Deaths", Observed.Deaths, Expected.Deaths, Potentially.Excess.Deaths)  
long\_data\_sec\_New

## # A tibble: 120 x 4  
## # Groups: Age.Range [8]  
## Age.Range Cause.of.Death Death\_Type Deaths  
## <fctr> <fctr> <chr> <dbl>  
## 1 0-49 Cancer Observed.Deaths 485.26151  
## 2 0-49 Chronic Lower Respiratory Disease Observed.Deaths 52.27354  
## 3 0-49 Heart Disease Observed.Deaths 404.80074  
## 4 0-49 Stroke Observed.Deaths 86.90902  
## 5 0-49 Unintentional Injury Observed.Deaths 799.16290  
## 6 0-54 Cancer Observed.Deaths 903.57413  
## 7 0-54 Chronic Lower Respiratory Disease Observed.Deaths 90.56129  
## 8 0-54 Heart Disease Observed.Deaths 707.54162  
## 9 0-54 Stroke Observed.Deaths 132.63768  
## 10 0-54 Unintentional Injury Observed.Deaths 934.75290  
## # ... with 110 more rows

library(ggplot2)  
ggplot(data = long\_data\_sec\_New, aes(x=factor(Age.Range), y = Deaths, colour = Death\_Type, group = Death\_Type)) +   
 geom\_line() + facet\_wrap(~Cause.of.Death,scales = "free")



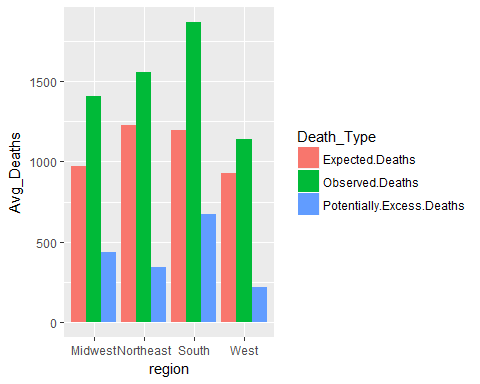
# 3. Max and Min Deaths by region  
library(dplyr)  
Data\_region <- Health\_data[,c(9,11,12,14)] %>%  
 group\_by(region) %>%  
 summarise\_all(funs(mean))  
Data\_region

## # A tibble: 4 x 4  
## region Observed.Deaths Expected.Deaths Potentially.Excess.Deaths  
## <fctr> <dbl> <dbl> <dbl>  
## 1 Midwest 1409.081 973.6223 439.8306  
## 2 Northeast 1556.521 1225.5080 340.7248  
## 3 South 1868.281 1195.5215 675.4204  
## 4 West 1140.032 930.2722 222.5200

library(tidyr)  
long\_data\_region <- gather(Data\_region,"Death\_Type", "Avg\_Deaths", Observed.Deaths, Expected.Deaths, Potentially.Excess.Deaths)  
long\_data\_region

## # A tibble: 12 x 3  
## region Death\_Type Avg\_Deaths  
## <fctr> <chr> <dbl>  
## 1 Midwest Observed.Deaths 1409.0814  
## 2 Northeast Observed.Deaths 1556.5210  
## 3 South Observed.Deaths 1868.2806  
## 4 West Observed.Deaths 1140.0323  
## 5 Midwest Expected.Deaths 973.6223  
## 6 Northeast Expected.Deaths 1225.5080  
## 7 South Expected.Deaths 1195.5215  
## 8 West Expected.Deaths 930.2722  
## 9 Midwest Potentially.Excess.Deaths 439.8306  
## 10 Northeast Potentially.Excess.Deaths 340.7248  
## 11 South Potentially.Excess.Deaths 675.4204  
## 12 West Potentially.Excess.Deaths 222.5200

library(ggplot2)  
ggplot(long\_data\_region,aes(x = region,y = Avg\_Deaths,fill = Death\_Type)) + geom\_bar(position = "dodge", stat = "identity")



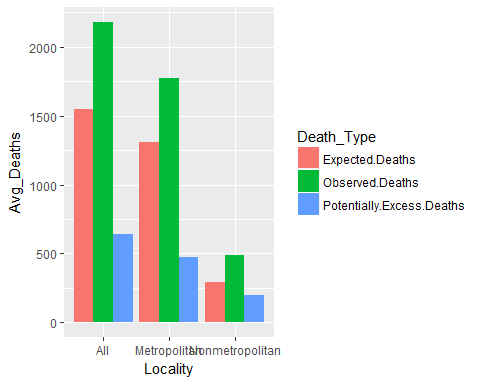
#4.  
library(dplyr)  
Data\_Locality <- Health\_data[,c(8,9,11,12)] %>%  
 group\_by(Locality) %>%  
 summarise\_all(funs(mean))  
Data\_Locality

## # A tibble: 3 x 4  
## Locality Observed.Deaths Expected.Deaths  
## <fctr> <dbl> <dbl>  
## 1 All 2181.6381 1549.6431  
## 2 Metropolitan 1771.5943 1308.0541  
## 3 Nonmetropolitan 487.3112 290.7998  
## # ... with 1 more variables: Potentially.Excess.Deaths <dbl>

library(tidyr)  
long\_Data\_Locality <- gather(Data\_Locality,"Death\_Type", "Avg\_Deaths", Observed.Deaths, Expected.Deaths, Potentially.Excess.Deaths)  
long\_Data\_Locality

## # A tibble: 9 x 3  
## Locality Death\_Type Avg\_Deaths  
## <fctr> <chr> <dbl>  
## 1 All Observed.Deaths 2181.6381  
## 2 Metropolitan Observed.Deaths 1771.5943  
## 3 Nonmetropolitan Observed.Deaths 487.3112  
## 4 All Expected.Deaths 1549.6431  
## 5 Metropolitan Expected.Deaths 1308.0541  
## 6 Nonmetropolitan Expected.Deaths 290.7998  
## 7 All Potentially.Excess.Deaths 640.8613  
## 8 Metropolitan Potentially.Excess.Deaths 473.3012  
## 9 Nonmetropolitan Potentially.Excess.Deaths 197.7998

library(ggplot2)  
ggplot(long\_Data\_Locality,aes(x = Locality,y = Avg\_Deaths,fill = Death\_Type)) + geom\_bar(position = "dodge", stat = "identity")



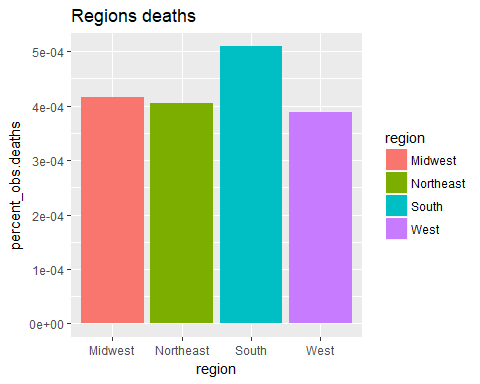
#5  
Health\_data$percent\_obs.deaths <- Health\_data$Observed.Deaths / Health\_data$Population  
head(Health\_data)

## State.FIPS.Code Year Cause.of.Death State HHS.Region  
## 1 AK 2006 Heart Disease Alaska 10  
## 2 AK 2008 Stroke Alaska 10  
## 3 AK 2015 Cancer Alaska 10  
## 4 AK 2007 Heart Disease Alaska 10  
## 5 AK 2015 Cancer Alaska 10  
## 6 AK 2011 Chronic Lower Respiratory Disease Alaska 10  
## Age.Range Benchmark Locality Observed.Deaths Population  
## 1 0-49 2010 Fixed Nonmetropolitan 19 163571  
## 2 0-59 Floating Metropolitan 24 408596  
## 3 0-64 2005 Fixed All 390 665595  
## 4 0-59 2010 Fixed All 158 603065  
## 5 0-59 Floating Nonmetropolitan 88 189260  
## 6 0-64 2010 Fixed Metropolitan 24 449953  
## Expected.Deaths Potentially.Excess.Deaths  
## 1 15 4  
## 2 11 13  
## 3 353 43  
## 4 121 37  
## 5 63 27  
## 6 18 8  
## Percent.Potentially.Excess.Deaths region percent\_obs.deaths  
## 1 21.10% West 1.161575e-04  
## 2 54.20% West 5.873773e-05  
## 3 11% West 5.859419e-04  
## 4 23.40% West 2.619950e-04  
## 5 30.70% West 4.649688e-04  
## 6 33.30% West 5.333890e-05

Obs.Deaths\_region <- Health\_data[,c(14:15)] %>%  
 group\_by(region) %>%  
 summarise\_all(funs(mean))  
Obs.Deaths\_region

## # A tibble: 4 x 2  
## region percent\_obs.deaths  
## <fctr> <dbl>  
## 1 Midwest 0.0004149184  
## 2 Northeast 0.0004044605  
## 3 South 0.0005085527  
## 4 West 0.0003889081

library(ggplot2)  
ggplot(Obs.Deaths\_region,aes(x = region,y = percent\_obs.deaths,fill = region)) + labs(title="Regions deaths") + geom\_bar(stat = "identity")



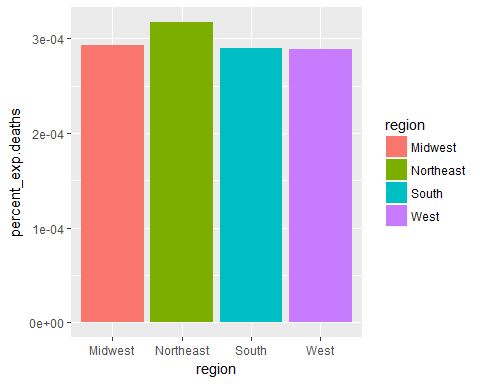
#6  
Health\_data$percent\_exp.deaths <- Health\_data$Expected.Deaths / Health\_data$Population  
head(Health\_data)

## State.FIPS.Code Year Cause.of.Death State HHS.Region  
## 1 AK 2006 Heart Disease Alaska 10  
## 2 AK 2008 Stroke Alaska 10  
## 3 AK 2015 Cancer Alaska 10  
## 4 AK 2007 Heart Disease Alaska 10  
## 5 AK 2015 Cancer Alaska 10  
## 6 AK 2011 Chronic Lower Respiratory Disease Alaska 10  
## Age.Range Benchmark Locality Observed.Deaths Population  
## 1 0-49 2010 Fixed Nonmetropolitan 19 163571  
## 2 0-59 Floating Metropolitan 24 408596  
## 3 0-64 2005 Fixed All 390 665595  
## 4 0-59 2010 Fixed All 158 603065  
## 5 0-59 Floating Nonmetropolitan 88 189260  
## 6 0-64 2010 Fixed Metropolitan 24 449953  
## Expected.Deaths Potentially.Excess.Deaths  
## 1 15 4  
## 2 11 13  
## 3 353 43  
## 4 121 37  
## 5 63 27  
## 6 18 8  
## Percent.Potentially.Excess.Deaths region percent\_obs.deaths  
## 1 21.10% West 1.161575e-04  
## 2 54.20% West 5.873773e-05  
## 3 11% West 5.859419e-04  
## 4 23.40% West 2.619950e-04  
## 5 30.70% West 4.649688e-04  
## 6 33.30% West 5.333890e-05  
## percent\_exp.deaths  
## 1 9.170330e-05  
## 2 2.692146e-05  
## 3 5.303525e-04  
## 4 2.006417e-04  
## 5 3.328754e-04  
## 6 4.000418e-05

exp.Deaths\_region <- Health\_data[,c(14,16)] %>%  
 group\_by(region) %>%  
 summarise\_all(funs(mean))  
exp.Deaths\_region

## # A tibble: 4 x 2  
## region percent\_exp.deaths  
## <fctr> <dbl>  
## 1 Midwest 0.0002925200  
## 2 Northeast 0.0003173100  
## 3 South 0.0002897919  
## 4 West 0.0002890023

library(ggplot2)  
ggplot(exp.Deaths\_region,aes(x = region,y = percent\_exp.deaths,fill = region)) + geom\_bar(stat = "identity")



#Standard diviation:  
  
sd(Health\_data$Potentially.Excess.Deaths)

## [1] 778.4071

sd(Health\_data$Observed.Deaths)

## [1] 3183.589

sd(Health\_data$Expected.Deaths)

## [1] 2600.873

#Anova table  
Anova.aov= aov(Health\_data$Expected.Deaths~Health\_data$Cause.of.Death)  
Anova.aov

## Call:  
## aov(formula = Health\_data$Expected.Deaths ~ Health\_data$Cause.of.Death)  
##   
## Terms:  
## Health\_data$Cause.of.Death Residuals  
## Sum of Squares 1.379290e+11 1.159151e+12  
## Deg. of Freedom 4 191743  
##   
## Residual standard error: 2458.727  
## Estimated effects may be unbalanced

summary(Anova.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Health\_data$Cause.of.Death 4 1.379e+11 3.448e+10 5704 <2e-16 \*\*\*  
## Residuals 191743 1.159e+12 6.045e+06   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Anova.aov= aov(Health\_data$Population~Health\_data$Observed.Death)  
Anova.aov

## Call:  
## aov(formula = Health\_data$Population ~ Health\_data$Observed.Death)  
##   
## Terms:  
## Health\_data$Observed.Death Residuals  
## Sum of Squares 2.081888e+18 3.194341e+18  
## Deg. of Freedom 1 191746  
##   
## Residual standard error: 4081572  
## Estimated effects may be unbalanced

summary(Anova.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Health\_data$Observed.Death 1 2.082e+18 2.082e+18 124969 <2e-16 \*\*\*  
## Residuals 191746 3.194e+18 1.666e+13   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Anova.aov= aov(Health\_data$Observed.Deaths~Health\_data$Cause.of.Death)  
Anova.aov

## Call:  
## aov(formula = Health\_data$Observed.Deaths ~ Health\_data$Cause.of.Death)  
##   
## Terms:  
## Health\_data$Cause.of.Death Residuals  
## Sum of Squares 2.023871e+11 1.741014e+12  
## Deg. of Freedom 4 191743  
##   
## Residual standard error: 3013.293  
## Estimated effects may be unbalanced

summary(Anova.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Health\_data$Cause.of.Death 4 2.024e+11 5.06e+10 5572 <2e-16 \*\*\*  
## Residuals 191743 1.741e+12 9.08e+06   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Linear Regression  
Regression<-lm(Health\_data$Expected.Deaths ~ Health\_data$Cause.of.Death + Health\_data$Population)  
summary(Regression)

##   
## Call:  
## lm(formula = Health\_data$Expected.Deaths ~ Health\_data$Cause.of.Death +   
## Health\_data$Population)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8189 -619 45 518 38526   
##   
## Coefficients:  
## Estimate  
## (Intercept) 1.421e+03  
## Health\_data$Cause.of.DeathChronic Lower Respiratory Disease -2.237e+03  
## Health\_data$Cause.of.DeathHeart Disease -9.892e+02  
## Health\_data$Cause.of.DeathStroke -2.272e+03  
## Health\_data$Cause.of.DeathUnintentional Injury -1.856e+03  
## Health\_data$Population 3.029e-04  
## Std. Error  
## (Intercept) 9.907e+00  
## Health\_data$Cause.of.DeathChronic Lower Respiratory Disease 1.363e+01  
## Health\_data$Cause.of.DeathHeart Disease 1.339e+01  
## Health\_data$Cause.of.DeathStroke 1.358e+01  
## Health\_data$Cause.of.DeathUnintentional Injury 1.338e+01  
## Health\_data$Population 8.171e-07  
## t value  
## (Intercept) 143.47  
## Health\_data$Cause.of.DeathChronic Lower Respiratory Disease -164.14  
## Health\_data$Cause.of.DeathHeart Disease -73.89  
## Health\_data$Cause.of.DeathStroke -167.35  
## Health\_data$Cause.of.DeathUnintentional Injury -138.64  
## Health\_data$Population 370.69  
## Pr(>|t|)   
## (Intercept) <2e-16 \*\*\*  
## Health\_data$Cause.of.DeathChronic Lower Respiratory Disease <2e-16 \*\*\*  
## Health\_data$Cause.of.DeathHeart Disease <2e-16 \*\*\*  
## Health\_data$Cause.of.DeathStroke <2e-16 \*\*\*  
## Health\_data$Cause.of.DeathUnintentional Injury <2e-16 \*\*\*  
## Health\_data$Population <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1877 on 191742 degrees of freedom  
## Multiple R-squared: 0.4794, Adjusted R-squared: 0.4794   
## F-statistic: 3.532e+04 on 5 and 191742 DF, p-value: < 2.2e-16

#Correlation  
cor.test(Health\_data$percent\_obs.deaths,Health\_data$Population)

##   
## Pearson's product-moment correlation  
##   
## data: Health\_data$percent\_obs.deaths and Health\_data$Population  
## t = -16.15, df = 191750, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.04132491 -0.03238521  
## sample estimates:  
## cor   
## -0.0368558

# partial correlations  
library(ggm)

## Warning: package 'ggm' was built under R version 3.3.3

## Loading required package: igraph

## Warning: package 'igraph' was built under R version 3.3.3

##   
## Attaching package: 'igraph'

## The following object is masked from 'package:tidyr':  
##   
## crossing

## The following objects are masked from 'package:dplyr':  
##   
## as\_data\_frame, groups, union

## The following objects are masked from 'package:stats':  
##   
## decompose, spectrum

## The following object is masked from 'package:base':  
##   
## union

##   
## Attaching package: 'ggm'

## The following object is masked from 'package:igraph':  
##   
## pa

pcor(c("Observed.Deaths", "percent\_obs.deaths"), var(Health\_data))

## Warning in var(Health\_data): NAs introduced by coercion

## [1] 0.4312337