script.r

Jon Minton

Sun Feb 22 07:50:23 2015

# First script  
# 11/2/2015  
  
  
# To do :   
# 1) Send Poverty & Suburbs work to Ade  
# 2) reply to emails  
# 3) extract list of dzs as per Ellie's suggestion [ done]  
# 4) add packrat to see if this solves markdown issue [done - seems to ]  
# 5) add existing script and adapt where relevant  
  
rm(list=ls())  
  
require(rmarkdown)

## Loading required package: rmarkdown

require(repmis)

## Loading required package: repmis

require(plyr)

## Loading required package: plyr

require(dplyr)

## Loading required package: dplyr  
##   
## Attaching package: 'dplyr'  
##   
## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

require(tidyr)

## Loading required package: tidyr

require(ggplot2)

## Loading required package: ggplot2

require(ggtern)

## Loading required package: ggtern  
##   
## Attaching package: 'ggtern'  
##   
## The following objects are masked from 'package:ggplot2':  
##   
## %+%, %+replace%, aes, calc\_element, geom\_density2d,  
## geom\_segment, geom\_smooth, ggplot\_build, ggplot\_gtable,  
## ggsave, opts, stat\_density2d, stat\_smooth, theme, theme\_bw,  
## theme\_classic, theme\_get, theme\_gray, theme\_grey,  
## theme\_minimal, theme\_set, theme\_update

require(corrplot)

## Loading required package: corrplot

require(vegan)

## Loading required package: vegan  
## Loading required package: permute  
## Loading required package: lattice  
## This is vegan 2.2-1

# Linking chps to dzs -----------------------------------------------------  
  
  
# # find dzs referred to  
#   
# chps\_of\_interest <- read.csv("data/geographies/greater\_glasgow\_definitions\_simplified.csv") %>% tbl\_df()  
# chps\_of\_interest <- chps\_of\_interest %>% slice(1:12)  
#   
# chps\_to\_dzs <- read.csv("data/geographies/latestpcinfowithlinkpc.csv") %>% tbl\_df()  
# chps\_to\_dzs <- chps\_to\_dzs %>%   
# select(dz\_2001=Datazone, chp=CHP) %>%   
# distinct(dz\_2001)  
#   
# # n.b. need the 2001 not 2011 dz codes   
# chps\_of\_interest <- chps\_of\_interest %>% rename(chp=chcp\_code)   
#   
# dzs\_in\_greater\_glasgow <- chps\_to\_dzs %>%   
# inner\_join(chps\_of\_interest) %>%  
# select(-chcp\_name)  
#   
# write.csv(dzs\_in\_greater\_glasgow, file="data/geographies/dzs\_in\_greater\_glasgow.csv", row.names=F)  
# #   
  
# Main Analysis -----------------------------------------------------------  
  
greater\_glasgow\_dzs <- read.csv("data/geographies/dzs\_in\_greater\_glasgow.csv") %>% tbl\_df()  
  
#Tenure  
  
tenure\_households <- source\_DropboxData(  
 file="tenure\_households.csv",  
 key="6t6dss41g8fat1y"  
 ) %>% tbl\_df() %>% select(  
 dz\_2001=datazone, year,   
 all\_households=HO.allhouseholds,  
 council\_houses=HO.council,  
 rented\_from\_employer=HO.employ,  
 owned\_with\_mortgage=HO.ownmortloan,  
 owned\_outright=HO.ownoutright,  
 private\_rented=HO.privlet,  
 rented\_from\_relative=HO.relative,  
 shared\_ownership=HO.sharedown,  
 other\_social\_rented=HO.social  
) %>% mutate(  
 social=council\_houses + other\_social\_rented,  
 rented=rented\_from\_employer + private\_rented+ rented\_from\_relative,  
 owned=owned\_with\_mortgage + owned\_outright + shared\_ownership  
 ) %>%  
 mutate(  
 council\_houses=council\_houses/all\_households,  
 rented\_from\_employer=rented\_from\_employer/all\_households,  
 owned\_with\_mortgage=owned\_with\_mortgage/all\_households,  
 owned\_outright=owned\_outright/all\_households,  
 private\_rented=private\_rented/all\_households,  
 rented\_from\_relative=rented\_from\_relative/all\_households,  
 shared\_ownership=shared\_ownership/all\_households,  
 other\_social\_rented=other\_social\_rented/all\_households,  
 social = social/all\_households,  
 rented = rented/all\_households,  
 owned=owned/all\_households  
 )

## Downloading data from: https://dl.dropboxusercontent.com/s/6t6dss41g8fat1y/tenure\_households.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## 7cb8e90f51d1a3855e4ed4b9a7015a97d03282cd

# 6505 observations - whole of Scotland  
# left join to just Greater Glasgow  
tenure\_households <- greater\_glasgow\_dzs %>% left\_join(tenure\_households)

## Joining by: "dz\_2001"

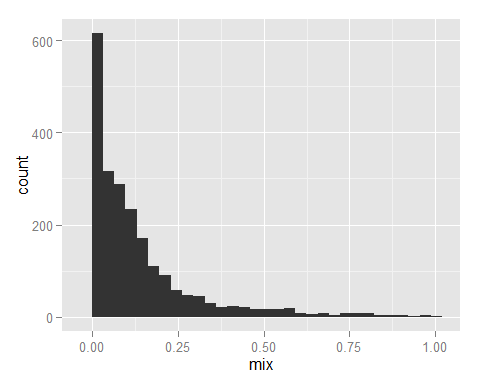
## Warning: joining character vector and factor, coercing into character  
## vector

# now 2200 observations  
# this is 34% of total, dzs are approx equal population  
# if Scot population is 5.3 million this implies   
# Pop of Greater Glasgow is about 1.8 Million -   
# does this seem reasonable?  
  
  
tenure\_households <- tenure\_households %>% mutate(  
 mix=(social \* rented \* owned) / (1/3)^3)  
 # this should be the maximum possible mix value  
  
tenure\_households %>% group\_by(year) %>% summarise()

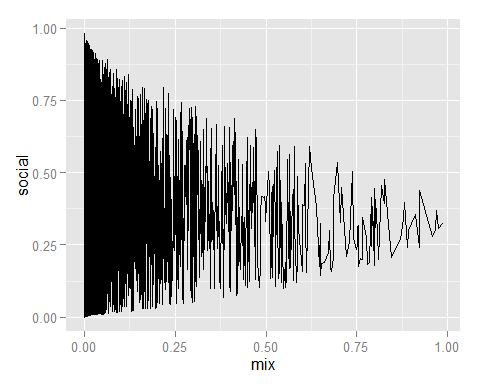
## Source: local data frame [1 x 1]  
##   
## year  
## 1 2001

# unfortunately this is only available for 2001  
  
#   
qplot(  
 x=mix, data=tenure\_households  
 )

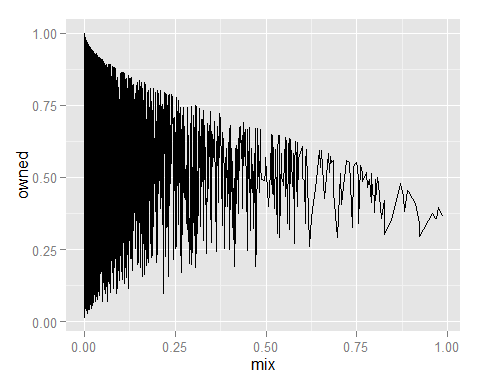
## stat\_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.



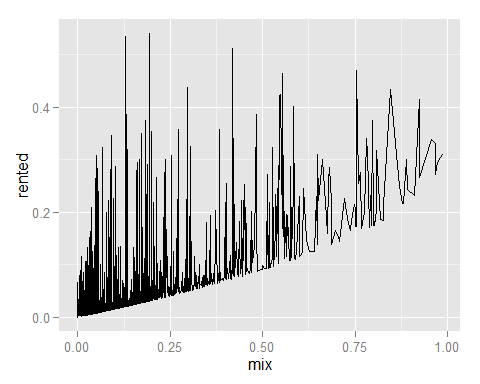
# a lot of excess 0s - no mix  
  
# arrange households by mix, then plot proportions of each tenure type along this linke  
  
tenure\_households <- tenure\_households %>% arrange(mix)  
  
ggplot(tenure\_households) + geom\_line(aes(x=mix, y=social))



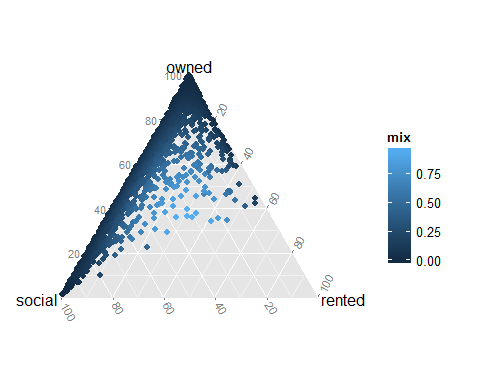
ggplot(tenure\_households) + geom\_line(aes(x=mix, y=owned))



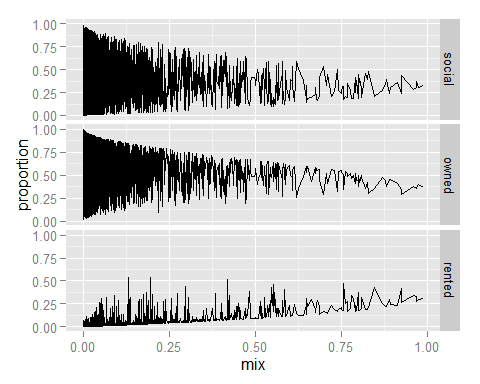
ggplot(tenure\_households) + geom\_line(aes(x=mix, y=rented))



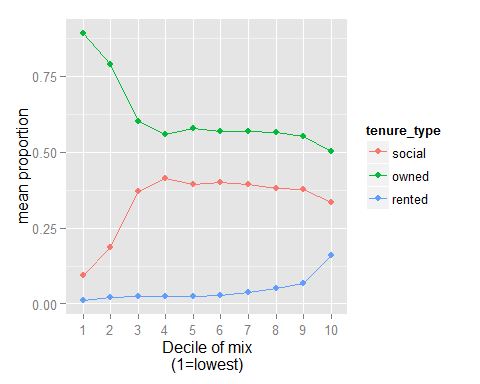
# ternery plot  
  
ggtern(data=tenure\_households, aes(x=social, y=owned, z=rented, colour=mix)) + geom\_point()



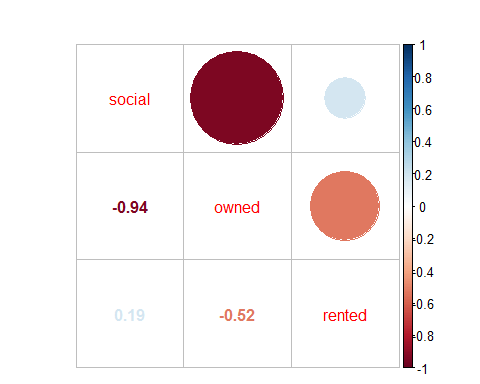
# want social, owned, and rented to be gathered   
  
tenure\_households %>%   
 select(dz\_2001, mix, social, owned, rented) %>%  
 gather(key = tenure\_type, value=tenure\_proportion, -dz\_2001, -mix) %>%  
 ggplot( aes(x=mix, y=tenure\_proportion)) +  
 geom\_line() +   
 facet\_grid(tenure\_type ~ . ) +   
 labs(y="proportion", x="mix")



# mix deciles   
  
tenure\_deciles <- tenure\_households %>%  
 select(dz\_2001, mix, social, owned, rented, all\_households) %>%  
 gather(key = tenure\_type, value=tenure\_proportion, -dz\_2001, -mix, -all\_households) %>%  
 mutate(m10=ntile(mix, 10)) %>%   
 group\_by(m10, tenure\_type) %>%   
 summarise(  
 tenure\_mean=mean(tenure\_proportion),  
 tenure\_sd=sd(tenure\_proportion),  
 n=sum(all\_households)  
 ) %>%  
 mutate(  
 ci = tenure\_sd / n^(1/2),  
 lower=tenure\_mean - 2 \* ci,  
 upper=tenure\_mean + 2 \* ci  
 )   
  
tenure\_deciles %>% ggplot(aes(  
 x=factor(m10), group=tenure\_type,   
 colour=tenure\_type, y=tenure\_mean  
 )) +  
 geom\_line(  
 ) +  
 geom\_pointrange(  
 aes(  
 ymax=upper,  
 ymin=lower  
 )  
 ) + labs(  
 y="mean proportion",  
 x="Decile of mix\n(1=lowest)"  
 )



# what's the correlation between the tenure types over the deciles?  
  
tenure\_deciles %>%   
 select(m10, tenure\_type, tenure\_mean) %>%  
 spread(key=tenure\_type, value=tenure\_mean) %>%   
 select(-m10) %>%   
 cor() %>%   
 corrplot.mixed()



# mix and dependancy ratios  
  
populations <- read.csv("data/derived/populations\_by\_age\_year\_sex.csv") %>%   
 tbl\_df()  
  
  
# reduce to just Greater Glasgow  
populations <- populations %>% rename(dz\_2001=datazone)  
populations <- populations %>% right\_join(greater\_glasgow\_dzs)

## Joining by: "dz\_2001"

## Warning: joining factors with different levels, coercing to character  
## vector

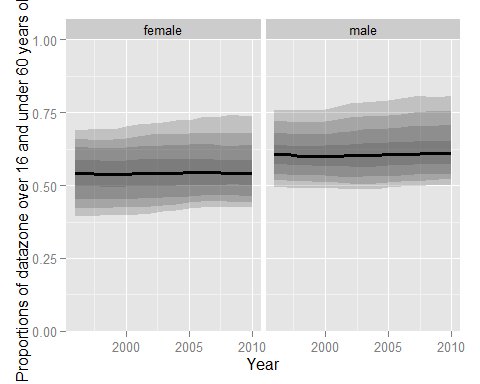
# now 1.21 million  
populations %>%   
 group\_by(year, age\_range, sex) %>%  
 summarise(  
 cells=n(),  
 sum=sum(count)  
 )

## Source: local data frame [590 x 5]  
## Groups: year, age\_range  
##   
## year age\_range sex cells sum  
## 1 1996 0\_0 female 2200 10218  
## 2 1996 0\_0 male 2200 10948  
## 3 1996 1\_4 female 2200 43974  
## 4 1996 1\_4 male 2200 46177  
## 5 1996 10\_12 female 2200 34126  
## 6 1996 10\_12 male 2200 35134  
## 7 1996 13\_14 female 2200 22109  
## 8 1996 13\_14 male 2200 22835  
## 9 1996 15\_15 female 2200 11504  
## 10 1996 15\_15 male 2200 11874  
## .. ... ... ... ... ...

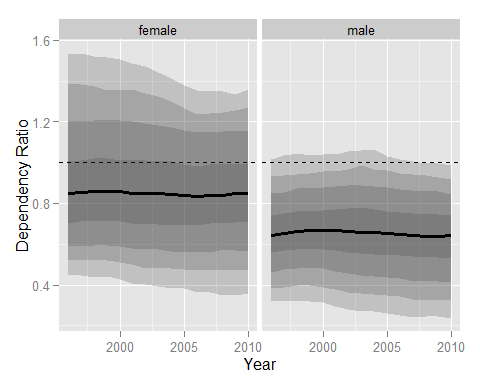
populations <- populations %>% arrange(year, sex, lower\_age)  
  
populations <- populations %>%   
 arrange(year,sex, lower\_age) %>%  
 mutate(  
 working\_age = ifelse(lower\_age > 16 & ((upper\_age < 60 & sex=="female") | (upper\_age < 65 & sex=="male")), 1, 0)  
 )   
  
populations %>%   
 group\_by(year, dz\_2001, sex, working\_age) %>%  
 summarise(count=sum(count))

## Source: local data frame [132,000 x 5]  
## Groups: year, dz\_2001, sex  
##   
## year dz\_2001 sex working\_age count  
## 1 1996 S01000758 female 0 222  
## 2 1996 S01000758 female 1 295  
## 3 1996 S01000758 male 0 183  
## 4 1996 S01000758 male 1 336  
## 5 1996 S01001423 female 0 148  
## 6 1996 S01001423 female 1 205  
## 7 1996 S01001423 male 0 109  
## 8 1996 S01001423 male 1 204  
## 9 1996 S01001424 female 0 247  
## 10 1996 S01001424 female 1 347  
## .. ... ... ... ... ...

prop\_working\_age <- populations %>%   
 group\_by(year, dz\_2001, sex) %>%  
 summarise(p\_wage=sum(count[working\_age==1])/sum(count))  
  
prop\_working\_age <- prop\_working\_age[!is.nan(prop\_working\_age$p\_wage),]   
prop\_working\_age <- prop\_working\_age %>% mutate(dep\_r = (1 - p\_wage) / p\_wage)  
  
prop\_working\_age %>% group\_by(year, sex) %>%  
 summarise(   
 q\_025=quantile(p\_wage, 0.025),   
 q\_050=quantile(p\_wage, 0.050),  
 q\_100=quantile(p\_wage, 0.100),  
 q\_250=quantile(p\_wage, 0.250),  
 q\_500=quantile(p\_wage, 0.500),  
 q\_750=quantile(p\_wage, 0.750),  
 q\_900=quantile(p\_wage, 0.900),  
 q\_950=quantile(p\_wage, 0.950),  
 q\_975=quantile(p\_wage, 0.975)  
 ) %>%  
 ggplot(aes(x=year, y=q\_500)) +  
 facet\_grid(. ~ sex) +  
 geom\_ribbon(aes(ymin=q\_025, ymax=q\_975), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_050, ymax=q\_950), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_100, ymax=q\_900), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_250, ymax=q\_750), alpha=0.2) +  
 geom\_line(size=1.1) +   
 coord\_cartesian(ylim=c(0,1)) +   
 labs(x="Year", y="Proportions of datazone over 16 and under 60 years old")



# equivalently, dependency ratio  
  
prop\_working\_age %>% group\_by(year, sex) %>%  
 summarise(   
 q\_025=quantile(dep\_r, 0.025),   
 q\_050=quantile(dep\_r, 0.050),  
 q\_100=quantile(dep\_r, 0.100),  
 q\_250=quantile(dep\_r, 0.250),  
 q\_500=quantile(dep\_r, 0.500),  
 q\_750=quantile(dep\_r, 0.750),  
 q\_900=quantile(dep\_r, 0.900),  
 q\_950=quantile(dep\_r, 0.950),  
 q\_975=quantile(dep\_r, 0.975)  
 ) %>%  
 ggplot(aes(x=year, y=q\_500)) +  
 facet\_grid(. ~ sex) +  
 geom\_ribbon(aes(ymin=q\_025, ymax=q\_975), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_050, ymax=q\_950), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_100, ymax=q\_900), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_250, ymax=q\_750), alpha=0.2) +  
 geom\_line(size=1.1) + geom\_hline(yintercept=1, linetype="dashed") +  
 labs(x="Year", y="Dependency Ratio")

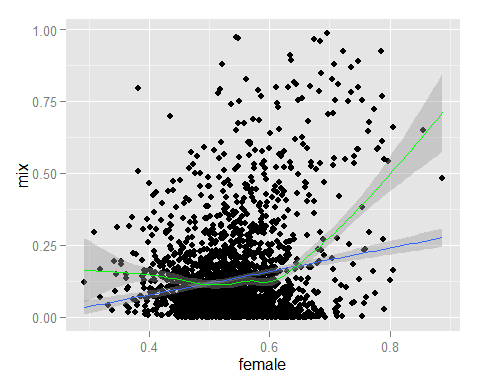


# Correlation between mix and proportion\_working\_age at datazone level  
  
tenure\_wage <- prop\_working\_age %>%   
 select(-dep\_r) %>%  
 spread(key=sex, value=p\_wage) %>%  
 inner\_join(tenure\_households)

## Joining by: c("year", "dz\_2001")

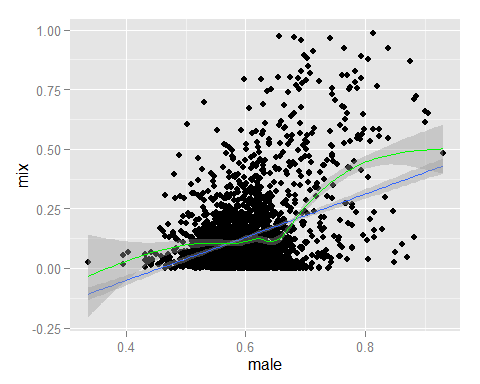
tenure\_wage %>%   
 ggplot(aes(x=female, y=mix)) +   
 geom\_point() +   
 stat\_smooth(method="lm") + stat\_smooth(colour="green")

## geom\_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.



tenure\_wage %>%   
 ggplot(aes(x=male, y=mix)) +   
 geom\_point() +   
 stat\_smooth(method="lm") + stat\_smooth(colour="green")

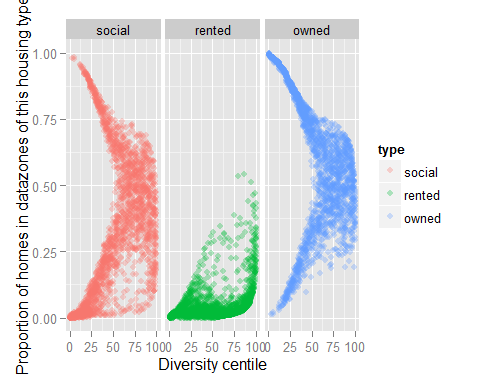
## geom\_smooth: method="auto" and size of largest group is >=1000, so using gam with formula: y ~ s(x, bs = "cs"). Use 'method = x' to change the smoothing method.



# Diversity Index ---------------------------------------------------------  
  
  
#########################################################################################  
# Instead of my own measure, how about using the diversity index?  
# Available in the vegan package  
# http://cc.oulu.fi/~jarioksa/softhelp/vegan/html/diversity.html  
  
# Trying out examples:  
# data(BCI)  
# H <- diversity(BCI)  
# simp <- diversity(BCI, "simpson")  
# invsimp <- diversity(BCI, "inv")  
# r.2 <- rarefy(BCI, 2)  
# alpha <- fisher.alpha(BCI)  
# pairs(cbind(H, simp, invsimp, r.2, alpha), pch="+", col="blue")  
# ## Species richness (S) and Pielou's evenness (J):  
# S <- specnumber(BCI) ## rowSums(BCI > 0) does the same...  
# J <- H/log(S)  
  
# So, the counts are needed  
  
  
tenure\_diversity <- tenure\_households %>%   
 filter(year==2001) %>%   
 select(all\_households, social, rented, owned) %>%   
 mutate(  
 social=social \* all\_households,  
 rented=rented \* all\_households,  
 owned=owned \* all\_households  
 ) %>%   
 select(-all\_households) %>%   
 diversity() %>%   
 as.data.frame() %>% tbl\_df()   
names(tenure\_diversity) = "diversity"  
  
tenure\_diversity <- tenure\_diversity %>%   
 mutate(dz\_2001 = tenure\_households$dz\_2001) %>%   
 tbl\_df() %>%   
 select(dz\_2001, diversity)  
  
tenure\_diversity <- tenure\_households %>%   
 filter(year==2001) %>%   
 select(dz\_2001, all\_households, social, rented, owned) %>%   
 mutate(  
 social=social ,  
 rented=rented ,  
 owned=owned   
 ) %>%   
 inner\_join(tenure\_diversity)

## Joining by: "dz\_2001"

# Create 100 bins based on diversity   
  
tenure\_diversity %>%   
 mutate(d\_bin=ntile(diversity, 100)) %>%  
 select(-all\_households) %>%   
 gather(key=type, value=proportion, social, rented, owned) %>%  
 arrange(d\_bin) %>%   
 ggplot(aes(x=d\_bin, y=proportion, fill=type, colour=type)) + geom\_point(alpha=0.3) +  
 facet\_wrap(~ type) +  
 labs(  
 x="Diversity centile",  
 y="Proportion of homes in datazones of this housing type")



# Other types of mix to look at:  
# social class mix  
#   
  
################################################################################################################  
  
  
# Now to look at other kinds of mix  
  
# mix types:  
  
# Dwelling type mix  
  
# dwellings, in by\_year  
  
require(plyr)  
require(stringr)

## Loading required package: stringr

require(tidyr)  
require(dplyr)  
require(ggplot2)  
require(repmis)  
  
greater\_glasgow\_dzs <- read.csv("data/geographies/dzs\_in\_greater\_glasgow.csv") %>%   
 tbl\_df %>%  
 rename(datazone=dz\_2001)  
  
  
dwellings <- source\_DropboxData(  
 file="dwellings.csv",   
 key="1oqfsgpfzotxji4"  
 ) %>% tbl\_df

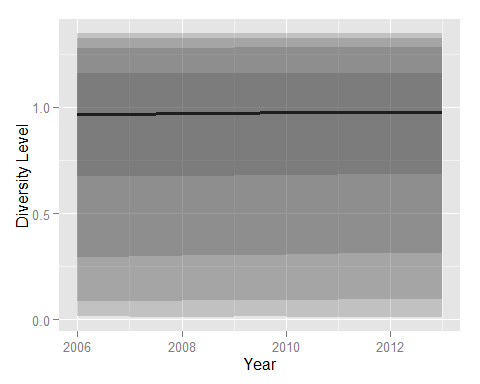
## Downloading data from: https://dl.dropboxusercontent.com/s/1oqfsgpfzotxji4/dwellings.csv   
##   
## SHA-1 hash of the downloaded data file is:  
## dfcd70415a2f07cf3e3460e32824e95875fc2ada

dwellings\_bands <- dwellings %>%  
 select(datazone, year, matches("\_[A-Z]\_")) %>%  
 inner\_join(greater\_glasgow\_dzs) %>%  
 select(-chp)

## Joining by: "datazone"

## Warning: joining character vector and factor, coercing into character  
## vector

names(dwellings\_bands) <- names(dwellings\_bands) %>%   
 str\_replace\_all("HO.Band\_", "") %>%   
 str\_replace\_all("\_dwellings", "")  
  
dwelling\_bands <- dwellings\_bands %>%  
 gather("band", "count", -datazone, -year)  
  
  
dwellings\_sizes <- dwellings %>%  
 select(datazone, year, matches("HO.Rooms[0123456789]{1,2}"), HO.Roomsge10)  
   
names(dwellings\_sizes) <- names(dwellings\_sizes) %>%   
 str\_replace\_all("HO.Rooms", "") %>%   
 str\_replace\_all("ge", "")  
  
dwellings\_sizes <- dwellings\_sizes %>%   
 gather(num\_of\_rooms, count, -datazone, -year) %>%  
 filter(!is.na(count))  
   
  
# Dwellings\_type  
dwellings\_types <- dwellings %>%  
 select(datazone, year, HO.Flat, HO.Terraced, HO.Semidetached, HO.Detached)  
  
names(dwellings\_types) <- names(dwellings\_types) %>% str\_replace\_all("HO.", "")   
  
dwellings\_types <- dwellings\_types %>%   
 gather(type, count, -datazone, -year) %>%  
 filter(!is.na(count))  
  
dwellings\_types$type <- dwellings\_types$type %>% tolower  
  
  
write.csv(dwellings\_bands, file="data/derived/dwellings\_by\_band.csv", row.names=FALSE)  
write.csv(dwellings\_sizes, file="data/derived/dwellings\_by\_size.csv", row.names=FALSE)  
write.csv(dwellings\_types, file="data/derived/dwellings\_by\_type.csv", row.names=FALSE)  
  
  
# Diversity by housing sizes?  
  
d\_dwelling\_types <- dwellings\_types %>%   
 spread(type, count) %>%   
 select(-datazone, -year) %>%   
 as.matrix %>%  
 diversity  
  
dwelling\_types\_wide <- dwellings\_types %>%  
 spread(type, count) %>%  
 mutate(diversity=d\_dwelling\_types)  
  
# What has been the change in diversity from year to year?  
  
div\_by\_year <- dwelling\_types\_wide %>%  
 group\_by(year) %>%  
 summarise(  
 q\_025=quantile(diversity, 0.025),  
 q\_050=quantile(diversity, 0.050),  
 q\_100=quantile(diversity, 0.100),  
 q\_250=quantile(diversity, 0.250),  
 q\_500=quantile(diversity, 0.500),  
 q\_750=quantile(diversity, 0.750),  
 q\_900=quantile(diversity, 0.900),  
 q\_950=quantile(diversity, 0.950),  
 q\_975=quantile(diversity, 0.975)  
 )  
  
  
div\_by\_year %>% ggplot(aes(x=year, y=q\_500)) +  
 geom\_line(size=1.1) +  
 geom\_ribbon(aes(ymin=q\_025, ymax=q\_975), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_050, ymax=q\_950), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_100, ymax=q\_900), alpha=0.2) +  
 geom\_ribbon(aes(ymin=q\_250, ymax=q\_750), alpha=0.2) +  
 labs(x="Year", y="Diversity Level")



#1) qualifications mix  
#   
#   
#   
# ES-degr-lev-all  
# # number of people with a degree  
# es-noquals   
# # number of people with no qualifications