**Statistical assignment 4**

[james hawkey] [add your candidate number here - mandatory]  
github: <https://github.com/jh994/assignment4>

**Read data**

In this assignment you will need to reproduce 5 ggplot graphs. I supply graphs as images; you need to write the ggplot2 code to reproduce them and knit and submit a Markdown document with the reproduced graphs (as well as your .Rmd file).

First we will need to open and recode the data. I supply the code for this; you only need to change the file paths.

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.4.4

## Warning: package 'ggplot2' was built under R version 3.4.4

## Warning: package 'tibble' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'readr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## Warning: package 'stringr' was built under R version 3.4.4

## Warning: package 'forcats' was built under R version 3.4.4

Data8 <- read\_tsv("C:/Users/User/Documents/datan3/data/UKDA-6614-tab/tab/ukhls\_w8/h\_indresp.tab")

Data8 <- Data8 %>%

select(pidp, h\_age\_dv, h\_payn\_dv, h\_gor\_dv)

Stable <- read\_tsv("c:/Users/User/Documents/datan3/data/UKDA-6614-tab/tab/ukhls\_wx/xwavedat.tab")

Stable <- Stable %>%

select(pidp, sex\_dv, ukborn, plbornc)

Data <- Data8 %>% left\_join(Stable, "pidp")

rm(Data8, Stable)

Data <- Data %>%

mutate(sex\_dv = ifelse(sex\_dv == 1, "male",

ifelse(sex\_dv == 2, "female", NA))) %>%

mutate(h\_payn\_dv = ifelse(h\_payn\_dv < 0, NA, h\_payn\_dv)) %>%

mutate(h\_gor\_dv = recode(h\_gor\_dv,

`-9` = NA\_character\_,

`1` = "North East",

`2` = "North West",

`3` = "Yorkshire",

`4` = "East Midlands",

`5` = "West Midlands",

`6` = "East of England",

`7` = "London",

`8` = "South East",

`9` = "South West",

`10` = "Wales",

`11` = "Scotland",

`12` = "Northern Ireland")) %>%

mutate(placeBorn = case\_when(

ukborn == -9 ~ NA\_character\_,

ukborn < 5 ~ "UK",

plbornc == 5 ~ "Ireland",

plbornc == 18 ~ "India",

plbornc == 19 ~ "Pakistan",

plbornc == 20 ~ "Bangladesh",

plbornc == 10 ~ "Poland",

plbornc == 27 ~ "Jamaica",

plbornc == 24 ~ "Nigeria",

TRUE ~ "other")

)

Reproduce the following graphs as close as you can. For each graph, write two sentences (not more!) describing its main message.

1. Univariate distribution (20 points).

#the graph shows the distroubtion of the net monthly pay of the respondents.

#the graph shows that the mode of the results is sligtly over the 1000 pounds mark with a steady decrease after.

#there is a slight increase just pasted the 5000 pound marker.

Data %>%

ggplot(aes(x=h\_payn\_dv)) +

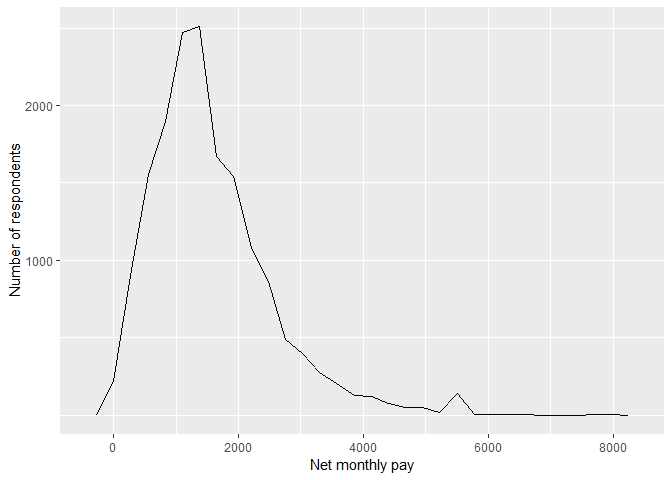
geom\_freqpoly(binwidth = 275)+

xlab("Net monthly pay") +

ylab("Number of respondents")+

scale\_y\_continuous(breaks = seq(1000,2000,1000))

## Warning: Removed 22638 rows containing non-finite values (stat\_bin).

 2. Line chart (20 points).

#the graph shows that both males and females have and increase in monthley earnings from the ages of 20 to 45.

# the male line shows that on a whole males have a higher monthley earings which also increase at a higher rate comparied to there female counterparts. this stead increase starts to reverse at the 45 age mark falling back to just over 1500 of monthly earings.

# the female line shows that while there is a simlair level of increase compared to the males up untill about the 25 age mark the increase after that becomes much less and progress much slower between the ages of 25 to 50 and then decrease as well.

DataSex <- Data[which(!is.na(Data$sex\_dv)),]

plotSmooth<- DataSex%>%

filter(h\_age\_dv<66) %>%

ggplot(aes(x=h\_age\_dv, y=h\_payn\_dv, group=sex\_dv, linetype=sex\_dv)) +

geom\_smooth(method = "loess", formula = y ~ x, color = "black") +

xlab("Age") +

ylab("Monthly earnings")

plotSmooth$labels$group <- "Sex"

plotSmooth$labels$linetype <- "Sex"

plotSmooth

## Warning: Removed 14403 rows containing non-finite values (stat\_smooth).



1. Faceted bar chart (20 points)

# these bar charts show the difference between the median monthly net pay between between males and female with in countrys in the data set. all of the charts show that males average high monthely incomes than females.

#there bar charts also highlight the differenct in median monthely net pay through out the defferent countrys.

#the different between male and females median monthely pay also differs between the countrys with banladesh haveing the closet level between the sexs while ireland and the uk have the largest.

Data\_medianPay <- Data %>%

group\_by(placeBorn, sex\_dv) %>%

summarise(medianPay = median(h\_payn\_dv,na.rm=TRUE))

Data\_medianPay <- Data\_medianPay[!is.na(Data\_medianPay$sex\_dv),]

Data\_medianPay <- Data\_medianPay[!is.na(Data\_medianPay$placeBorn),]

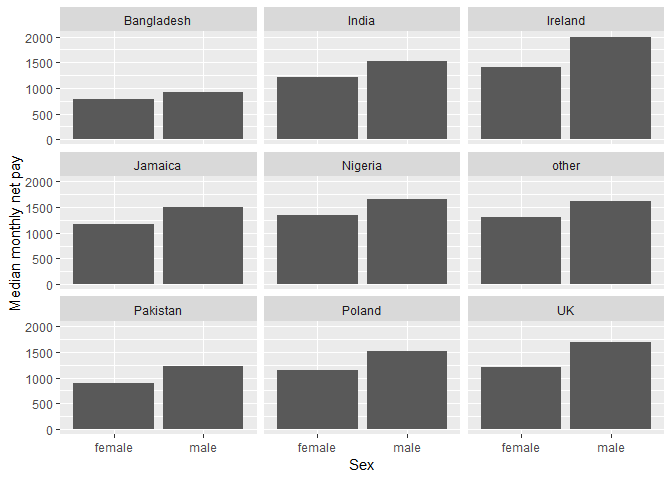
ggplot(Data\_medianPay, aes( x = sex\_dv, y = medianPay ) ) +

geom\_bar( stat = "identity" ) +

facet\_wrap( ~ placeBorn ) +

xlab("Sex") +

ylab("Median monthly net pay")



* 1. Heat map (20 points).

#this heat map shows the mean age of the different coutrys within the data set alnog with the different regions the darker the shade the youner the mean age nigeria as one of the youngest mean ages.

Data\_meanAge <- Data %>%

group\_by(h\_gor\_dv, placeBorn) %>%

summarise(meanAge = mean(h\_age\_dv))

Data\_meanAge <- Data\_meanAge[which(!is.na(Data\_meanAge$h\_gor\_dv)),]

Data\_meanAge <- Data\_meanAge[which(!is.na(Data\_meanAge$placeBorn)),]

Data\_meanAge %>%

ggplot(aes(h\_gor\_dv, placeBorn)) +

geom\_tile(aes(fill = meanAge)) +

xlab("Region") +

ylab("City of birth") +

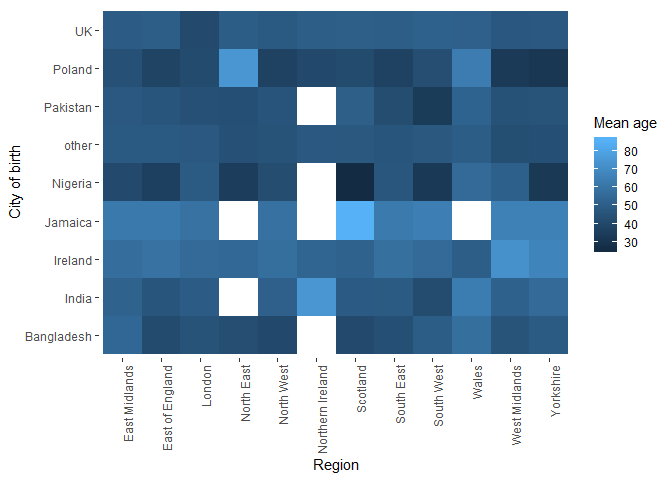
theme\_bw() +

theme(panel.border = element\_blank(), panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(), axis.line = element\_blank(),

axis.text.x = element\_text(angle = 90, hjust = 1)) +

labs(fill = "Mean age")



1. Population pyramid (20 points).

# this population pyramid allows us to veiw the age disturbtion through out the population in shows a developed coutrys distrubtion. the under 25 age marker there are peaks of 300 n but after this there is a small decrease in the percentage falling to just over 200 in regards to the male population and 250 for the female. after this decrease there is a increase of porotion up untill the age of 50 where both male and female reach there peaks. after this the population protion inregards to ages begins to decrease along with age.

Data\_nAge <- Data %>%

group\_by(h\_age\_dv, sex\_dv) %>%

summarise(nAge = n())

Data\_nAge$nAge <- ifelse(Data\_nAge$sex\_dv == "male", -1\*Data\_nAge$nAge, Data\_nAge$nAge)

ggplot(Data\_nAge, aes(x = h\_age\_dv, y= nAge, fill = sex\_dv)) +

geom\_bar(data = subset(Data\_nAge, sex\_dv == "female"), stat = "identity") +

geom\_bar(data = subset(Data\_nAge, sex\_dv == "male"), stat = "identity") +

scale\_fill\_manual("Sex", values = c("female" = "red", "male" = "blue")) +

xlab("Age") +

ylab("n") +

theme\_bw() +

coord\_flip()

