ST447 Data Analysis & Statistical Methods Final Project

Akeel Ahamed

31st January 2018



We first set the working directory to the folder used for this project, containing the data files

knitr::opts\_knit$set(root.dir = "~/Desktop/LSE\_Courses/ST447 /ST447 Final Project/")

We then Enter Student ID to get XYZ’s profile

ID = 201316007

Load the XYZ function to generate XYZ’s profile

source("XYZProfile.r")  
XYZprofile(201316007)

## The profile of XYZ:  
## - Age: 22  
## - Gender: Female  
## - Home address: Bury (Manchester)

From now on, throughtout this entire project, I am assuming that the name of my friend XYZ is **Sarah**

**DATA COLLECTION AND PREPROCESSING**

We then Read in Files from years 2007-2015, where the files have been downloaded & stored in the above directory.

sheet\_1\_201415 <- read.csv("sheet\_1.csv", stringsAsFactors = FALSE,header = T )  
  
#Data in year 2013/14  
sheet\_2\_201314 <- read.csv("sheet\_2.csv", stringsAsFactors = FALSE,header = T )  
  
#Data in year 2012/13  
sheet\_3\_201213 <- read.csv("sheet\_3.csv", stringsAsFactors = FALSE,header = T )  
  
#2011/12 Data not available for Bury(Manchester). Bolton used instead  
sheet\_4\_201112 <- read.csv("sheet\_4.csv", stringsAsFactors = FALSE, header = T)  
  
# Data in year 2010/11  
sheet\_5\_201011 <-read.csv("sheet\_5.csv", stringsAsFactors = FALSE,header = T )  
  
# Data in year 2009/10  
sheet\_6\_200910 <-read.csv("sheet\_6.csv", stringsAsFactors = FALSE,header = T )  
  
# Data in year 2008/09  
sheet\_7\_200809 <-read.csv("sheet\_7.csv", stringsAsFactors = FALSE,header = T )  
  
# Data in year 2007/08  
sheet\_8\_200708 <-read.csv("sheet\_8.csv", stringsAsFactors = FALSE,header = T )

Rename columns

#Rename columns  
colnames(sheet\_1\_201415) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
colnames(sheet\_2\_201314) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
colnames(sheet\_3\_201213) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
colnames(sheet\_4\_201112) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
colnames(sheet\_5\_201011) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
colnames(sheet\_6\_200910) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
colnames(sheet\_7\_200809) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
colnames(sheet\_8\_200708) <- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")

**Finding the mean Pass Rates at each Centre given that Sarah is a 22 year old Female**

Extract Data for Tests conducted in Bury (Manchester) and Wood Green (London) from years 2007 -2015:

# Extract Data for Tests conducted in Bury (Manchester) and Wood Green (London) from years 2007 -2015:  
#2014/15  
bury.2014 = sheet\_1\_201415[sheet\_1\_201415$`Test\_Centre`=="Bury (Manchester)" , ]   
wgreen.2014 = sheet\_1\_201415[sheet\_1\_201415$`Test\_Centre`=="Wood Green (London)" ,]   
  
#2013/14  
bury.2013 = sheet\_2\_201314[sheet\_2\_201314$`Test\_Centre`=="Bury (Manchester)",]   
wgreen.2013 = sheet\_2\_201314[sheet\_2\_201314$`Test\_Centre`=="Wood Green (London)" ,]   
  
#2012/13  
bury.2012 = sheet\_3\_201213[sheet\_3\_201213$`Test\_Centre`=="Bury (Manchester)" ,]   
wgreen.2012 = sheet\_3\_201213[sheet\_3\_201213$`Test\_Centre`=="Wood Green (London)" ,]   
  
#2011/12 Data not available for Bury(Manchester) but closest alternative at Bolton taken:  
wgreen.2011 = sheet\_4\_201112[sheet\_4\_201112$`Test\_Centre`=="Wood Green (London)" ,]   
bolton.2011 = sheet\_4\_201112[sheet\_4\_201112$Test\_Centre=="Bolton",]  
#2010/11  
bury.2010 = sheet\_5\_201011[sheet\_5\_201011$`Test\_Centre`=="Bury (Manchester)",]   
wgreen.2010 = sheet\_5\_201011[sheet\_5\_201011$`Test\_Centre`=="Wood Green (London)",]   
#This call does not seem to include the Age =25 row so we can manually add it using rbind(). Note that the values   
# have been typed in from the excel sheet.  
df<- data.frame( "Wood Green (London)",25, 163,   
 75, 46.0122699,195, 64,  
 32.8205128, 358, 139, 38.8268156)  
names(df)<- c("Test\_Centre","Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")  
wgreen.2010 = rbind(wgreen.2010,df)  
  
#2009/10  
bury.2009 = sheet\_6\_200910[sheet\_6\_200910$`Test\_Centre`=="Bury (Manchester)" ,]   
wgreen.2009 = sheet\_6\_200910[sheet\_6\_200910$`Test\_Centre`=="Wood Green (London)" ,]   
  
#2008/09  
bury.2008 = sheet\_7\_200809[sheet\_7\_200809$`Test\_Centre`=="Bury (Manchester)",]   
wgreen.2008 = sheet\_7\_200809[sheet\_7\_200809$`Test\_Centre`=="Wood Green (London)",]   
  
#2007/08  
bury.2007 = sheet\_8\_200708[sheet\_8\_200708$`Test\_Centre`=="Bury (Manchester)",]   
wgreen.2007 = sheet\_8\_200708[sheet\_8\_200708$`Test\_Centre`=="Wood Green (London)",]

I then merge all these data frames together to create one single data frame for each Area:

df\_bury<- rbind(bury.2007, bury.2008, bury.2009, bury.2010, bolton.2011, bury.2012, bury.2013, bury.2014)  
df\_wgreen<- rbind(wgreen.2007, wgreen.2008, wgreen.2009, wgreen.2010, wgreen.2011, wgreen.2012, wgreen.2013, wgreen.2014)

I then add the predictor “Year” using the $ sign:

df\_bury$Year <- c(rep(2007,9), rep(2008,9), rep(2009,9), rep(2010,9), rep(2011,9), rep(2012,9), rep(2013,9), rep(2014,9))  
df\_wgreen$Year <- c(rep(2007,9), rep(2008,9), rep(2009,9), rep(2010,9), rep(2011,9), rep(2012,9), rep(2013,9), rep(2014,9))

Upon further introspection of the dataframe we find that most of the numerical variables are of type “Character”:

summary(df\_bury)

## Test\_Centre Age Male\_Tests\_Conducted  
## Length:72 Length:72 Length:72   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Male\_Passes Male\_Pass\_Rate Female\_Tests\_Conducted  
## Length:72 Length:72 Length:72   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Female\_Passes Female\_Pass\_Rate Total\_Tests\_Conducted  
## Length:72 Length:72 Length:72   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Total\_Passes Pass\_Rate Year   
## Length:72 Length:72 Min. :2007   
## Class :character Class :character 1st Qu.:2009   
## Mode :character Mode :character Median :2010   
## Mean :2010   
## 3rd Qu.:2012   
## Max. :2014

summary(df\_wgreen)

## Test\_Centre Age Male\_Tests\_Conducted  
## Length:72 Length:72 Length:72   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Male\_Passes Male\_Pass\_Rate Female\_Tests\_Conducted  
## Length:72 Length:72 Length:72   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Female\_Passes Female\_Pass\_Rate Total\_Tests\_Conducted  
## Length:72 Length:72 Length:72   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Total\_Passes Pass\_Rate Year   
## Length:72 Length:72 Min. :2007   
## Class :character Class :character 1st Qu.:2009   
## Mode :character Mode :character Median :2010   
## Mean :2010   
## 3rd Qu.:2012   
## Max. :2014

Thus, we convert each numeric variable incorrectly classified as class “character” to class “numeric”:

df\_bury[,c("Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes",   
 "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")] <- as.numeric(   
 unlist(df\_bury[,c("Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted",   
 "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")]))   
  
df\_wgreen[,c("Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted", "Female\_Passes",   
 "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")] <- as.numeric(   
 unlist(df\_wgreen[,c("Age","Male\_Tests\_Conducted", "Male\_Passes", "Male\_Pass\_Rate","Female\_Tests\_Conducted",   
 "Female\_Passes", "Female\_Pass\_Rate", "Total\_Tests\_Conducted", "Total\_Passes", "Pass\_Rate")]))

Inspecting the summary of each data frame now would confirm that we indeed have each variable’s class/type as it should be(intuitively).

summary(df\_bury)

## Test\_Centre Age Male\_Tests\_Conducted Male\_Passes   
## Length:72 Min. :17 Min. : 53.00 Min. : 25.0   
## Class :character 1st Qu.:19 1st Qu.: 90.25 1st Qu.: 46.0   
## Mode :character Median :21 Median :143.00 Median : 70.5   
## Mean :21 Mean :212.01 Mean :111.3   
## 3rd Qu.:23 3rd Qu.:266.00 3rd Qu.:133.2   
## Max. :25 Max. :817.00 Max. :493.0   
## Male\_Pass\_Rate Female\_Tests\_Conducted Female\_Passes Female\_Pass\_Rate  
## Min. :39.68 Min. : 58.0 Min. : 23.0 Min. :32.54   
## 1st Qu.:47.80 1st Qu.:111.0 1st Qu.: 46.0 1st Qu.:40.67   
## Median :51.73 Median :158.0 Median : 70.5 Median :43.51   
## Mean :51.79 Mean :225.6 Mean :105.2 Mean :44.39   
## 3rd Qu.:55.62 3rd Qu.:292.5 3rd Qu.:137.2 3rd Qu.:47.85   
## Max. :63.77 Max. :737.0 Max. :434.0 Max. :60.13   
## Total\_Tests\_Conducted Total\_Passes Pass\_Rate Year   
## Min. : 111.0 Min. : 48.00 Min. :37.21 Min. :2007   
## 1st Qu.: 199.8 1st Qu.: 94.25 1st Qu.:44.49 1st Qu.:2009   
## Median : 293.0 Median :140.00 Median :47.14 Median :2010   
## Mean : 437.6 Mean :216.58 Mean :47.68 Mean :2010   
## 3rd Qu.: 537.8 3rd Qu.:261.50 3rd Qu.:50.93 3rd Qu.:2012   
## Max. :1554.0 Max. :927.00 Max. :59.65 Max. :2014

summary(df\_wgreen)

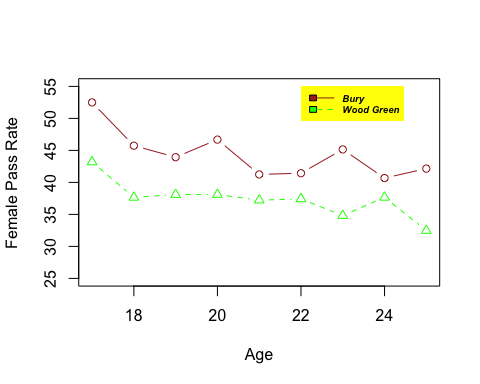
## Test\_Centre Age Male\_Tests\_Conducted Male\_Passes   
## Length:72 Min. :17 Min. :105.0 Min. : 43.00   
## Class :character 1st Qu.:19 1st Qu.:154.5 1st Qu.: 69.75   
## Mode :character Median :21 Median :185.5 Median : 81.50   
## Mean :21 Mean :199.0 Mean : 87.44   
## 3rd Qu.:23 3rd Qu.:236.0 3rd Qu.:102.25   
## Max. :25 Max. :363.0 Max. :164.00   
## Male\_Pass\_Rate Female\_Tests\_Conducted Female\_Passes Female\_Pass\_Rate  
## Min. :33.66 Min. : 69.0 Min. : 28.00 Min. :26.54   
## 1st Qu.:41.51 1st Qu.:203.0 1st Qu.: 68.50 1st Qu.:34.48   
## Median :43.68 Median :224.5 Median : 82.50 Median :38.22   
## Mean :44.11 Mean :227.8 Mean : 84.61 Mean :37.41   
## 3rd Qu.:45.47 3rd Qu.:263.0 3rd Qu.:101.00 3rd Qu.:40.05   
## Max. :58.21 Max. :372.0 Max. :137.00 Max. :47.76   
## Total\_Tests\_Conducted Total\_Passes Pass\_Rate Year   
## Min. :192.0 Min. : 84.0 Min. :32.58 Min. :2007   
## 1st Qu.:359.5 1st Qu.:142.0 1st Qu.:38.50 1st Qu.:2009   
## Median :407.0 Median :166.0 Median :40.21 Median :2010   
## Mean :426.8 Mean :172.1 Mean :40.37 Mean :2010   
## 3rd Qu.:492.5 3rd Qu.:199.2 3rd Qu.:41.94 3rd Qu.:2012   
## Max. :639.0 Max. :269.0 Max. :51.48 Max. :2014

#Inspecting names to look for explanatory variables  
names(df\_bury)

## [1] "Test\_Centre" "Age"   
## [3] "Male\_Tests\_Conducted" "Male\_Passes"   
## [5] "Male\_Pass\_Rate" "Female\_Tests\_Conducted"  
## [7] "Female\_Passes" "Female\_Pass\_Rate"   
## [9] "Total\_Tests\_Conducted" "Total\_Passes"   
## [11] "Pass\_Rate" "Year"

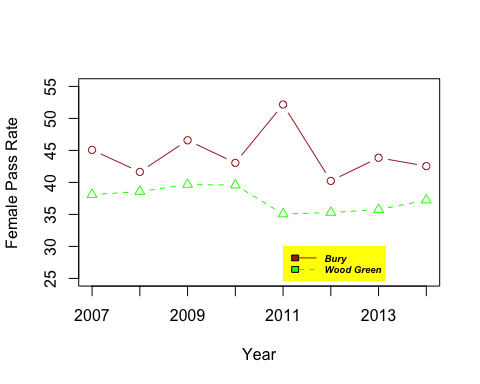
We will now do some exploratory visualisation & inspect Female Pass Rate according to Age in each Area

# Make a data frame consisting of a column for mean Female Pass Rate by Age and another for Age  
# Do this for each area  
df\_1\_bury <-data.frame(aggregate(df\_bury$Female\_Pass\_Rate, list(df\_bury$Age), mean))  
colnames(df\_1\_bury) = c("Age", "Female Pass Rate")  
df\_1\_wgreen <- aggregate(df\_wgreen$Female\_Pass\_Rate, list(df\_wgreen$Age), mean)  
colnames(df\_1\_wgreen) = c("Age", "Female Pass Rate")  
  
# Plot the data frame. Note that a function has been created to reduce replication of typing code  
makeplot <- function(x,y){  
 plot(x, col = "brown", type = "b", pch = 1, ylim = c(25,55), lty = 1)  
 lines(y, col = "Green", type = "b", pch = 2, lty = 2)  
}  
  
makeplot(df\_1\_bury, df\_1\_wgreen)  
legend(22, 55, legend = c("Bury", "Wood Green"), fill = c("brown", "green"), lty = 1:2,  
 col = c("brown", "green"), cex = 0.6, box.lty = 0, bg = "yellow", text.font = 4)



Do the same exploratory analysis as above this time inspecting female pass rate against year

df\_2\_bury <-data.frame(aggregate(df\_bury$Female\_Pass\_Rate, list(df\_bury$Year), mean))  
colnames(df\_2\_bury) = c("Year", "Female Pass Rate")  
df\_2\_wgreen <- aggregate(df\_wgreen$Female\_Pass\_Rate, list(df\_wgreen$Year), mean)  
colnames(df\_2\_wgreen) = c("Year", "Female Pass Rate")  
  
makeplot(df\_2\_bury, df\_2\_wgreen)  
legend(2011, 30, legend = c("Bury", "Wood Green"), fill = c("brown", "green"), lty = 1:2,  
 col = c("brown", "green"), cex = 0.6, box.lty = 0, bg = "yellow", text.font = 4)



Mean for all Females Aged 22 in each Area from Years 2007 - 2014:

Bury.Mean = mean(df\_bury$Female\_Pass\_Rate[df\_bury$Age == 22])  
WoodGreen.Mean = mean(df\_wgreen$Female\_Pass\_Rate[df\_wgreen$Age == 22])  
print(Bury.Mean)

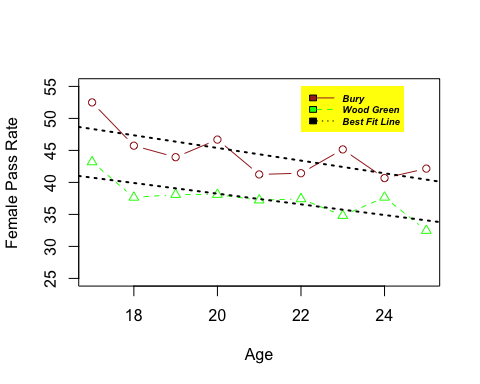
## [1] 41.43886

print(WoodGreen.Mean)

## [1] 37.42147

Fitting the line of best fit

plot(df\_1\_bury, col = "brown", type = "b", pch = 1, ylim = c(25,55), lty = 1)  
lm\_1\_bury <- lm(`Female Pass Rate` ~ Age, data = df\_1\_bury)  
abline(coef(lm\_1\_bury), lwd = 2, lty = 3)  
lines(df\_1\_wgreen, col = "green", type = "b", pch = 2, ylim = c(25,55), lty = 2)  
lm\_1\_wgreen <- lm(`Female Pass Rate` ~ Age, data = df\_1\_wgreen)  
abline(coef(lm\_1\_wgreen), lwd = 2, lty = 3)  
legend(22, 55, legend = c("Bury", "Wood Green", "Best Fit Line"), fill = c("brown", "green", "black"),  
 lty = 1:3,col = c("brown", "green", "black"), cex = 0.6, box.lty = 0, bg = "yellow", text.font = 4)



linear Regression Model for Female Pass Rate Against Age & Test Centre

#Make a new data frame combining Bury and Wood Green.   
df\_bury\_wgreen <- rbind(df\_bury, df\_wgreen)  
#Rename the Bolton Factor to Bury (Manchester) to be consistent with our initial assumption -   
#Since there is no data for Bury (Manchester) in 2011-12, I am taking the data for Bolton, which is the next   
#closest Test Centre to Bury of approximately 6 miles.   
df\_bury\_wgreen$Test\_Centre <- c(rep("Bury (Manchester)",72), rep("Wood Green (London)", 72))  
# Convert the test centre variable from character to factor in order to display contrasts (shown below)  
df\_bury\_wgreen$Test\_Centre <- as.factor(df\_bury\_wgreen$Test\_Centre)  
# Run a linear Regression model on Female Pass Rate against Test Centre and Age from the df\_bury\_wgreen dataset.  
# Note that the command to run linear regression in R is lm(y ~ x, data = ) where lm stands for linear model.  
lm\_bury\_wgreen <- lm(Female\_Pass\_Rate ~ Age + Test\_Centre , data = df\_bury\_wgreen)  
# Return the summary statistics of our linear model  
summary(lm\_bury\_wgreen)

##   
## Call:  
## lm(formula = Female\_Pass\_Rate ~ Age + Test\_Centre, data = df\_bury\_wgreen)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.8533 -3.4499 -0.4407 2.7066 13.6171   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.5433 3.4015 18.681 < 2e-16 \*\*\*  
## Age -0.9119 0.1596 -5.714 6.30e-08 \*\*\*  
## Test\_CentreWood Green (London) -6.9871 0.8241 -8.479 2.77e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.944 on 141 degrees of freedom  
## Multiple R-squared: 0.4258, Adjusted R-squared: 0.4176   
## F-statistic: 52.27 on 2 and 141 DF, p-value: < 2.2e-16

# Display the coding scheme used for the dummy variables  
contrasts(df\_bury\_wgreen$Test\_Centre)

## Wood Green (London)  
## Bury (Manchester) 0  
## Wood Green (London) 1

63.5433 - 0.9119\*(22) - 6.9871

## [1] 36.4944

63.5433 - 0.9119\*(22)

## [1] 43.4815