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**

We then Read in Files from years 2007-2015

#Data in year 2014/15  
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)  
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

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:

#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  
wgreen.2012 = sheet\_3\_201213[sheet\_3\_201213$`Test\_Centre`=="Wood Green (London)" ,]   
  
#2011/12  
wgreen.2011 = sheet\_4\_201112[sheet\_4\_201112$`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  
wgreen.2011 = sheet\_4\_201112[sheet\_4\_201112$`Test\_Centre`=="Wood Green (London)" ,]   
  
#2011/12 Data not available for Bury(Manchester) but it is available for Wood Green:  
wgreen.2011 = sheet\_4\_201112[sheet\_4\_201112$`Test\_Centre`=="Wood Green (London)" ,]   
  
#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,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$Year <- c(rep(2007,9),rep(2008,9), rep(2009,9),rep(2010,9), rep(2012,9), rep(2013,9), rep(2014,9))  
df\_wgreen$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:63 Length:63 Length:63   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Male\_Passes Male\_Pass\_Rate Female\_Tests\_Conducted  
## Length:63 Length:63 Length:63   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Female\_Passes Female\_Pass\_Rate Total\_Tests\_Conducted  
## Length:63 Length:63 Length:63   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
## Total\_Passes Pass\_Rate Year   
## Length:63 Length:63 Min. :2007   
## Class :character Class :character 1st Qu.:2008   
## Mode :character Mode :character Median :2010   
## Mean :2010   
## 3rd Qu.:2013   
## 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")]))   
  
summary(df\_bury)

## Test\_Centre Age Male\_Tests\_Conducted Male\_Passes   
## Length:63 Min. :17 Min. : 53.0 Min. : 25.0   
## Class :character 1st Qu.:19 1st Qu.: 87.5 1st Qu.: 45.0   
## Mode :character Median :21 Median :116.0 Median : 61.0   
## Mean :21 Mean :198.1 Mean :102.2   
## 3rd Qu.:23 3rd Qu.:251.0 3rd Qu.:132.5   
## Max. :25 Max. :784.0 Max. :451.0   
## 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.:47.05 1st Qu.:101.5 1st Qu.: 41.5 1st Qu.:39.91   
## Median :50.68 Median :149.0 Median : 58.0 Median :43.16   
## Mean :51.24 Mean :208.4 Mean : 94.0 Mean :43.28   
## 3rd Qu.:54.45 3rd Qu.:250.0 3rd Qu.:109.5 3rd Qu.:46.38   
## Max. :63.77 Max. :663.0 Max. :300.0 Max. :60.13   
## Total\_Tests\_Conducted Total\_Passes Pass\_Rate Year   
## Min. : 111.0 Min. : 48.0 Min. :37.21 Min. :2007   
## 1st Qu.: 194.5 1st Qu.: 88.0 1st Qu.:44.13 1st Qu.:2008   
## Median : 272.0 Median :123.0 Median :45.94 Median :2010   
## Mean : 406.6 Mean :196.2 Mean :46.82 Mean :2010   
## 3rd Qu.: 478.0 3rd Qu.:235.0 3rd Qu.:49.03 3rd Qu.:2013   
## Max. :1404.0 Max. :734.0 Max. :59.48 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

Expected(Mean) passing rate for 22 year old females at Bury (Manchester) and Wood Green (London) respectively:

mean(df\_bury$`Female\_Pass\_Rate`[df\_bury$Age==22])

## [1] 40.07298

mean(df\_wgreen$`Female\_Pass\_Rate`[df\_wgreen$Age])

## [1] 37.73707

Note that the above calculation was made on the strict assumption that Sarah belongs to the Age=22 category. If she wanted to generalise & find out what the mean pass rate was for her gender, then it would be:

mean(df\_bury$`Female\_Pass\_Rate`)

## [1] 43.2816

mean(df\_wgreen$`Female\_Pass\_Rate`)

## [1] 37.4059

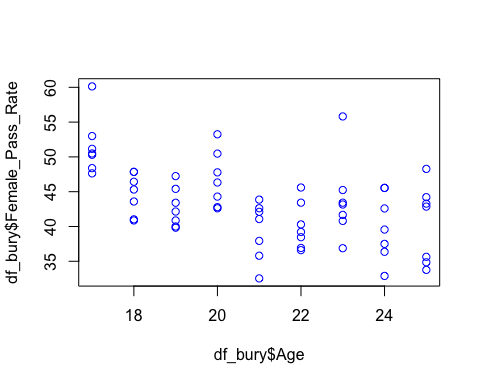
Thus, based on these findings, Sarah will have a better chance of passing the test given that she goes to the #Centre at Bury.

**Deciding where it would be best for Sarah to take the test**

Exploratory Data Analysis

I will now do some exploratory Data Analysis to get a sense of what the data shows(find any outliers or interesting correlations). The plot() function is used to display a scatter plot showing the relationship between two variables

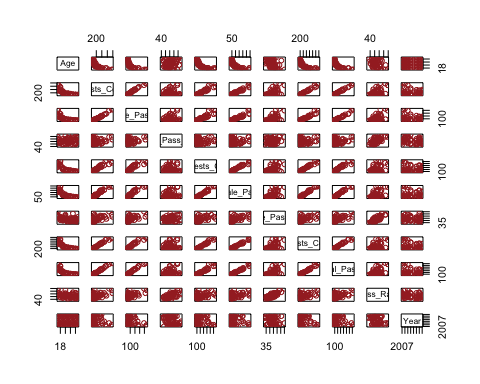
plot(df\_bury$Age, df\_bury$Female\_Pass\_Rate, col="blue")



This plot interestingly seems to suggest that as individuals get older(within the given range of 17-25), the probability of passing the test declines.Notice the two outliers represented by the circles; one states that a 17 year old individual has a passing rate of 60 % whereas the mean passing rate for the Age=17 category is only around 51%

The pairs() function can be used to create multiple scatter plots showing the relationship between  
independent continuous variables. How it works is that a variable on the horizontal axis is taken as an independent variable whereas one on the vertical axis (irrespective of whether it is positive/negative) is taken as the dependent/Response variable.

pairs(df\_bury[,-1],col="brown")



Correlations can only be performed for Numerix variables so we will remove the Test Centre and Age variables as they are categoical.

cor(df\_bury[-c(1,2)])

## Male\_Tests\_Conducted Male\_Passes Male\_Pass\_Rate  
## Male\_Tests\_Conducted 1.00000000 0.9919597 0.06878824  
## Male\_Passes 0.99195974 1.0000000 0.15990905  
## Male\_Pass\_Rate 0.06878824 0.1599091 1.00000000  
## Female\_Tests\_Conducted 0.95031462 0.9196238 0.01116382  
## Female\_Passes 0.97280982 0.9541803 0.05646824  
## Female\_Pass\_Rate 0.55382886 0.5711061 0.16053984  
## Total\_Tests\_Conducted 0.98938859 0.9706751 0.04275761  
## Total\_Passes 0.99481058 0.9910368 0.11556683  
## Pass\_Rate 0.47841785 0.5420329 0.65682619  
## Year -0.19598703 -0.2131284 -0.23588453  
## Female\_Tests\_Conducted Female\_Passes  
## Male\_Tests\_Conducted 0.95031462 0.97280982  
## Male\_Passes 0.91962380 0.95418034  
## Male\_Pass\_Rate 0.01116382 0.05646824  
## Female\_Tests\_Conducted 1.00000000 0.98697383  
## Female\_Passes 0.98697383 1.00000000  
## Female\_Pass\_Rate 0.46387324 0.57084803  
## Total\_Tests\_Conducted 0.98545903 0.99165777  
## Total\_Passes 0.96026024 0.98560194  
## Pass\_Rate 0.36930071 0.47388088  
## Year -0.16004753 -0.17205724  
## Female\_Pass\_Rate Total\_Tests\_Conducted Total\_Passes  
## Male\_Tests\_Conducted 0.5538289 0.98938859 0.9948106  
## Male\_Passes 0.5711061 0.97067515 0.9910368  
## Male\_Pass\_Rate 0.1605398 0.04275761 0.1155668  
## Female\_Tests\_Conducted 0.4638732 0.98545903 0.9602602  
## Female\_Passes 0.5708480 0.99165777 0.9856019  
## Female\_Pass\_Rate 1.0000000 0.51880912 0.5775544  
## Total\_Tests\_Conducted 0.5188091 1.00000000 0.9911981  
## Total\_Passes 0.5775544 0.99119810 1.0000000  
## Pass\_Rate 0.8419884 0.43350598 0.5178365  
## Year -0.1354530 -0.18167777 -0.1972420  
## Pass\_Rate Year  
## Male\_Tests\_Conducted 0.4784179 -0.1959870  
## Male\_Passes 0.5420329 -0.2131284  
## Male\_Pass\_Rate 0.6568262 -0.2358845  
## Female\_Tests\_Conducted 0.3693007 -0.1600475  
## Female\_Passes 0.4738809 -0.1720572  
## Female\_Pass\_Rate 0.8419884 -0.1354530  
## Total\_Tests\_Conducted 0.4335060 -0.1816778  
## Total\_Passes 0.5178365 -0.1972420  
## Pass\_Rate 1.0000000 -0.2375454  
## Year -0.2375454 1.0000000

cor(df\_wgreen[-c(1,2)])

## Male\_Tests\_Conducted Male\_Passes Male\_Pass\_Rate  
## Male\_Tests\_Conducted 1.00000000 0.9568870 -0.14876760  
## Male\_Passes 0.95688702 1.0000000 0.13649335  
## Male\_Pass\_Rate -0.14876760 0.1364934 1.00000000  
## Female\_Tests\_Conducted 0.50887920 0.4531091 -0.15024438  
## Female\_Passes 0.58966739 0.5713050 -0.01584812  
## Female\_Pass\_Rate 0.20198754 0.2971189 0.34122912  
## Total\_Tests\_Conducted 0.86563553 0.8086390 -0.17213176  
## Total\_Passes 0.88162615 0.8971686 0.07196338  
## Pass\_Rate 0.14631465 0.3451773 0.69580658  
## Year -0.06128154 -0.1080828 -0.17245788  
## Female\_Tests\_Conducted Female\_Passes  
## Male\_Tests\_Conducted 0.5088792 0.58966739  
## Male\_Passes 0.4531091 0.57130498  
## Male\_Pass\_Rate -0.1502444 -0.01584812  
## Female\_Tests\_Conducted 1.0000000 0.90110927  
## Female\_Passes 0.9011093 1.00000000  
## Female\_Pass\_Rate -0.2285856 0.20400622  
## Total\_Tests\_Conducted 0.8715037 0.86000970  
## Total\_Passes 0.7521462 0.87506713  
## Pass\_Rate -0.2534811 0.13207602  
## Year -0.1271441 -0.22540384  
## Female\_Pass\_Rate Total\_Tests\_Conducted Total\_Passes  
## Male\_Tests\_Conducted 0.20198754 0.86563553 0.88162615  
## Male\_Passes 0.29711887 0.80863897 0.89716862  
## Male\_Pass\_Rate 0.34122912 -0.17213176 0.07196338  
## Female\_Tests\_Conducted -0.22858562 0.87150367 0.75214620  
## Female\_Passes 0.20400622 0.86000970 0.87506713  
## Female\_Pass\_Rate 1.00000000 -0.01788323 0.28500347  
## Total\_Tests\_Conducted -0.01788323 1.00000000 0.93968942  
## Total\_Passes 0.28500347 0.93968942 1.00000000  
## Pass\_Rate 0.90218383 -0.06407767 0.27463429  
## Year -0.22081821 -0.10885859 -0.18504138  
## Pass\_Rate Year  
## Male\_Tests\_Conducted 0.14631465 -0.06128154  
## Male\_Passes 0.34517726 -0.10808280  
## Male\_Pass\_Rate 0.69580658 -0.17245788  
## Female\_Tests\_Conducted -0.25348113 -0.12714407  
## Female\_Passes 0.13207602 -0.22540384  
## Female\_Pass\_Rate 0.90218383 -0.22081821  
## Total\_Tests\_Conducted -0.06407767 -0.10885859  
## Total\_Passes 0.27463429 -0.18504138  
## Pass\_Rate 1.00000000 -0.21720516  
## Year -0.21720516 1.00000000

I am going to use the method of linear regression to decide which test Centre Sarah should do the test. In order to make a prediction on which centre is better, we will need to have a training set which will be used to train the model, and a testing set which (as its name implies) we can use to test the accuracy of the model.

The lm() function can be used to run a linear regression of the Y(dependent) variable on a set of predictors/ Independent variables(X). We will first decide which predictors are significant for the model in Bury and Wood Green and then use only the significant variables. If a variable is significant, this basically means that it has an effect on predicting the response and should therefore be included in the model.

The stars “\*" at the end of each row indicate how significant the variable is: the more stars a variable has, the more significant it is. If a variable has a dot associated with it , this means that it has significance as well but not as much as ones with the stars. We will now evaluate the models in each area and remove the insignificant variables (i.e. ones with no stars or dots) as they reduce model accuracy.

lm.bury <- lm(Female\_Pass\_Rate~ Age + Male\_Tests\_Conducted+ Male\_Passes + Male\_Pass\_Rate +Female\_Tests\_Conducted   
 +Female\_Passes + Pass\_Rate, data=df\_bury )  
summary(lm.bury)

##   
## Call:  
## lm(formula = Female\_Pass\_Rate ~ Age + Male\_Tests\_Conducted +   
## Male\_Passes + Male\_Pass\_Rate + Female\_Tests\_Conducted + Female\_Passes +   
## Pass\_Rate, data = df\_bury)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.77832 -0.37997 0.04408 0.37340 1.73984   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.118858 2.345115 1.330 0.189028   
## Age -0.041992 0.060213 -0.697 0.488502   
## Male\_Tests\_Conducted 0.010614 0.009257 1.147 0.256496   
## Male\_Passes -0.035873 0.015037 -2.386 0.020524 \*   
## Male\_Pass\_Rate -0.612289 0.037885 -16.162 < 2e-16 \*\*\*  
## Female\_Tests\_Conducted -0.028657 0.007264 -3.945 0.000228 \*\*\*  
## Female\_Passes 0.077431 0.015947 4.855 1.03e-05 \*\*\*  
## Pass\_Rate 1.552059 0.053109 29.224 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6589 on 55 degrees of freedom  
## Multiple R-squared: 0.9874, Adjusted R-squared: 0.9859   
## F-statistic: 618.2 on 7 and 55 DF, p-value: < 2.2e-16

Notice that the “Test \_Centre" variable is categorical and hence we will not include it in this calculation. Age , Male\_Tests\_Conducted and Year not significant

lm.wgreen <- lm(Female\_Pass\_Rate~ Age + Male\_Tests\_Conducted+ Male\_Passes + Male\_Pass\_Rate +Female\_Tests\_Conducted   
 +Female\_Passes + Pass\_Rate+Year, data=df\_wgreen )  
summary(lm.wgreen)

##   
## Call:  
## lm(formula = Female\_Pass\_Rate ~ Age + Male\_Tests\_Conducted +   
## Male\_Passes + Male\_Pass\_Rate + Female\_Tests\_Conducted + Female\_Passes +   
## Pass\_Rate + Year, data = df\_wgreen)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.9271 -0.2601 -0.1083 0.1695 2.5418   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 92.46630 62.38415 1.482 0.1433   
## Age -0.01583 0.04238 -0.374 0.7100   
## Male\_Tests\_Conducted 0.01488 0.01799 0.827 0.4113   
## Male\_Passes -0.05608 0.03996 -1.403 0.1654   
## Male\_Pass\_Rate -0.51591 0.09374 -5.503 7.27e-07 \*\*\*  
## Female\_Tests\_Conducted -0.02068 0.01273 -1.624 0.1094   
## Female\_Passes 0.06791 0.03076 2.208 0.0309 \*   
## Pass\_Rate 1.45305 0.12285 11.828 < 2e-16 \*\*\*  
## Year -0.04462 0.03101 -1.439 0.1551   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5691 on 63 degrees of freedom  
## Multiple R-squared: 0.9871, Adjusted R-squared: 0.9854   
## F-statistic: 600.7 on 8 and 63 DF, p-value: < 2.2e-16

Age , Male\_Tests\_Conducted, Male\_Passes and Year not significant

Remove insignificant variables in each models to improve predictive power:

lm.bury <- lm(Female\_Pass\_Rate~ Male\_Passes + Male\_Pass\_Rate +Female\_Tests\_Conducted   
 +Female\_Passes + Pass\_Rate, data=df\_bury )  
summary(lm.bury)

##   
## Call:  
## lm(formula = Female\_Pass\_Rate ~ Male\_Passes + Male\_Pass\_Rate +   
## Female\_Tests\_Conducted + Female\_Passes + Pass\_Rate, data = df\_bury)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.78577 -0.39680 0.05943 0.29227 1.77561   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.726029 1.530166 1.782 0.080154 .   
## Male\_Passes -0.019006 0.003546 -5.360 1.56e-06 \*\*\*  
## Male\_Pass\_Rate -0.647285 0.026875 -24.085 < 2e-16 \*\*\*  
## Female\_Tests\_Conducted -0.023261 0.006036 -3.854 0.000297 \*\*\*  
## Female\_Passes 0.071124 0.015218 4.674 1.85e-05 \*\*\*  
## Pass\_Rate 1.576667 0.049658 31.750 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6579 on 57 degrees of freedom  
## Multiple R-squared: 0.987, Adjusted R-squared: 0.9859   
## F-statistic: 867.5 on 5 and 57 DF, p-value: < 2.2e-16

lm.wgreen <- lm(Female\_Pass\_Rate~ Male\_Pass\_Rate +Female\_Tests\_Conducted   
 +Female\_Passes + Pass\_Rate , data=df\_wgreen )  
summary(lm.wgreen)

##   
## Call:  
## lm(formula = Female\_Pass\_Rate ~ Male\_Pass\_Rate + Female\_Tests\_Conducted +   
## Female\_Passes + Pass\_Rate, data = df\_wgreen)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.23585 -0.37674 -0.04022 0.25382 3.06129   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.73999 2.56245 4.972 4.87e-06 \*\*\*  
## Male\_Pass\_Rate -0.43097 0.05101 -8.448 3.75e-12 \*\*\*  
## Female\_Tests\_Conducted -0.05780 0.01175 -4.920 5.91e-06 \*\*\*  
## Female\_Passes 0.14786 0.02918 5.067 3.40e-06 \*\*\*  
## Pass\_Rate 1.09834 0.10422 10.539 7.32e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.674 on 67 degrees of freedom  
## Multiple R-squared: 0.9807, Adjusted R-squared: 0.9795   
## F-statistic: 851.1 on 4 and 67 DF, p-value: < 2.2e-16

Let the training set be a random sample of 75% of the observations in each model and the other 25% be the testing set.

set.seed(201316007) # To reproduce the exact set of random values each time  
split.bury <- sample(nrow(df\_bury), size = floor(0.75\*nrow(df\_bury)))  
train.bury = df\_bury[split.bury,]  
test.bury = df\_bury[-split.bury,]  
head(train.bury)

## Test\_Centre Age Male\_Tests\_Conducted Male\_Passes Male\_Pass\_Rate  
## 664 Bury (Manchester) 23 104 54 51.92308  
## 5891 Bury (Manchester) 21 88 46 52.27273  
## 570 Bury (Manchester) 20 169 88 52.07101  
## 666 Bury (Manchester) 18 346 169 48.84393  
## 661 Bury (Manchester) 20 150 68 45.33333  
## 668 Bury (Manchester) 20 156 86 55.12821  
## Female\_Tests\_Conducted Female\_Passes Female\_Pass\_Rate  
## 664 95 41 43.15789  
## 5891 126 41 32.53968  
## 570 176 75 42.61364  
## 666 420 195 46.42857  
## 661 187 80 42.78075  
## 668 180 86 47.77778  
## Total\_Tests\_Conducted Total\_Passes Pass\_Rate Year  
## 664 199 95 47.73869 2014  
## 5891 214 87 40.65421 2012  
## 570 345 163 47.24638 2008  
## 666 766 364 47.51958 2013  
## 661 337 148 43.91691 2014  
## 668 336 172 51.19048 2013

We do the same for Wood Green

set.seed(201316007) # To reproduce the exact set of random  
# values each time  
split.wgreen <- sample(nrow(df\_wgreen), size = floor(0.75\*nrow(df\_wgreen)))  
train.wgreen = df\_wgreen[split.wgreen,]  
test.wgreen = df\_wgreen[-split.wgreen,]  
head(train.wgreen)

## Test\_Centre Age Male\_Tests\_Conducted Male\_Passes  
## 3263 Wood Green (London) 22 250 96  
## 2878 Wood Green (London) 18 184 75  
## 2888 Wood Green (London) 22 205 90  
## 28851 Wood Green (London) 25 135 57  
## 3261 Wood Green (London) 20 247 108  
## 3341 Wood Green (London) 18 236 102  
## Male\_Pass\_Rate Female\_Tests\_Conducted Female\_Passes Female\_Pass\_Rate  
## 3263 38.40000 256 99 38.67188  
## 2878 40.76087 243 82 33.74486  
## 2888 43.90244 223 85 38.11659  
## 28851 42.22222 218 58 26.60550  
## 3261 43.72470 277 110 39.71119  
## 3341 43.22034 212 75 35.37736  
## Total\_Tests\_Conducted Total\_Passes Pass\_Rate Year  
## 3263 506 195 38.53755 2014  
## 2878 427 157 36.76815 2012  
## 2888 428 175 40.88785 2008  
## 28851 353 115 32.57790 2012  
## 3261 524 218 41.60305 2014  
## 3341 448 177 39.50893 2013

We now build the predictive models for Bury and Wood Green using only the significant variables as predictors and the training data

predictionModel.bury <- lm(Female\_Pass\_Rate~ Male\_Passes + Male\_Pass\_Rate +Female\_Tests\_Conducted   
 +Female\_Passes + Pass\_Rate, data=train.bury )  
summary(predictionModel.bury)

##   
## Call:  
## lm(formula = Female\_Pass\_Rate ~ Male\_Passes + Male\_Pass\_Rate +   
## Female\_Tests\_Conducted + Female\_Passes + Pass\_Rate, data = train.bury)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.55022 -0.43136 0.02434 0.34028 1.82432   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.313406 1.922188 0.683 0.498269   
## Male\_Passes -0.021189 0.004314 -4.911 1.49e-05 \*\*\*  
## Male\_Pass\_Rate -0.661423 0.033391 -19.808 < 2e-16 \*\*\*  
## Female\_Tests\_Conducted -0.022705 0.007057 -3.218 0.002527 \*\*   
## Female\_Passes 0.072294 0.018071 4.000 0.000258 \*\*\*  
## Pass\_Rate 1.622104 0.062211 26.074 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7153 on 41 degrees of freedom  
## Multiple R-squared: 0.985, Adjusted R-squared: 0.9832   
## F-statistic: 539.8 on 5 and 41 DF, p-value: < 2.2e-16

predictionModel.wgreen <-lm(Female\_Pass\_Rate~ Male\_Pass\_Rate +Female\_Tests\_Conducted   
 +Female\_Passes + Pass\_Rate, data=train.wgreen )  
summary(predictionModel.wgreen)

##   
## Call:  
## lm(formula = Female\_Pass\_Rate ~ Male\_Pass\_Rate + Female\_Tests\_Conducted +   
## Female\_Passes + Pass\_Rate, data = train.wgreen)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.18240 -0.36344 0.00182 0.24597 3.14893   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.97651 3.40678 3.809 0.000389 \*\*\*  
## Male\_Pass\_Rate -0.43831 0.06422 -6.825 1.24e-08 \*\*\*  
## Female\_Tests\_Conducted -0.05704 0.01487 -3.837 0.000357 \*\*\*  
## Female\_Passes 0.14681 0.03762 3.902 0.000291 \*\*\*  
## Pass\_Rate 1.09759 0.13693 8.015 1.81e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7159 on 49 degrees of freedom  
## Multiple R-squared: 0.9773, Adjusted R-squared: 0.9754   
## F-statistic: 526.5 on 4 and 49 DF, p-value: < 2.2e-16

Testing the predictive models:

**Bury**

prediction.bury <- predict(predictionModel.bury, newdata= test.bury)  
head(prediction.bury)

## 456 459 460 571 572 573   
## 50.97809 45.24736 45.84822 42.06016 36.07672 36.72546

head(test.bury$Female\_Pass\_Rate)

## [1] 50.47170 45.23810 45.53571 42.10526 36.91275 36.87943

**Wood Green**

prediction.wgreen <- predict(predictionModel.wgreen, newdata= test.wgreen)  
head(prediction.wgreen)

## 2395 2883 2884 2886 2889 2891   
## 39.95805 47.51873 41.60557 37.21748 37.69849 31.83785

head(test.wgreen$Female\_Pass\_Rate)

## [1] 39.53488 47.26027 41.19601 37.61755 38.03419 32.54902

CAlculate R-squared for the predictive model on the test set

**Bury**

RSS.bury <- sum((test.bury$Female\_Pass\_Rate -prediction.bury)^2)  
TSS.bury <- sum((test.bury$Female\_Pass\_Rate -mean(test.bury$Female\_Pass\_Rate))^2)  
ESS.bury <- 1-(RSS.bury/TSS.bury)  
ESS.bury

## [1] 0.9904062

**Wood Green**

RSS.wgreen <- sum((test.wgreen$Female\_Pass\_Rate -prediction.wgreen)^2)  
TSS.wgreen <- sum((test.wgreen$Female\_Pass\_Rate -mean(test.wgreen$Female\_Pass\_Rate))^2)  
ESS.wgreen <- 1-(RSS.wgreen/TSS.wgreen)  
ESS.wgreen

## [1] 0.9831382