Extra Credit Homework **Submitted by:** Shivali Dalmia

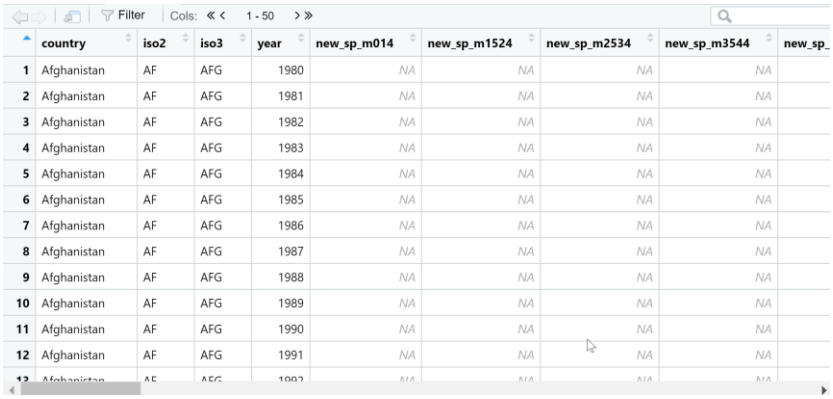
Step1.Loading the required libraries and dataset 'who' from tidyr package.

```
library(tidyverse)

library(tidyr)

library(ggplot2)

View(who)
```



Filter Cols: « < 1 - 50 > »

	country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp_
1	Afghanistan	AF	AFG	1980	NA	NA	NA	NA	
2	Afghanistan	AF	AFG	1981	NA	NA	NA	NA	
3	Afghanistan	AF	AFG	1982	NA	NA	NA	NA	
4	Afghanistan	AF	AFG	1983	NA	NA	NA	NA	
5	Afghanistan	AF	AFG	1984	NA	NA	NA	NA	
6	Afghanistan	AF	AFG	1985	NA	NA	NA	NA	
7	Afghanistan	AF	AFG	1986	NA	NA	NA	NA	
8	Afghanistan	AF	AFG	1987	NA	NA	NA	NA	
9	Afghanistan	AF	AFG	1988	NA	NA	NA	NA	
10	Afghanistan	AF	AFG	1989	NA	NA	NA	NA	
11	Afghanistan	AF	AFG	1990	NA	NA	NA	NA	
12	Afghanistan	AF	AFG	1991	NA	NA	NA	NA	
13	Afghanistan	AF	AFG	1992	NA	NA	NA	NA	

Dataset definition:

A subset of data from the World Health Organization Global Tuberculosis Report and accompanying global populations with 7,240 rows.

Column name	Description
country	Country Name
iso2	2-letter ISO country code
Iso3	3-letter ISO country code
Year	Year for which new cases were recorded
new_sp_m014 - new_rel_f65	Count of TB cases recorded by for different age groups for both males and females. e.g., new_sp_m014 new - stands for new cases sp - code for method of diagnosis • rel = relapse • sn = negative pulmonary smear • sp = positive pulmonary smear • ep = extrapulmonary m - code for gender • f = female • m = male 014 - code for age group • 014 = 0-14 years • 1524 = 15-24 years • 2534 = 25 to 34 years • 3544 = 35 to 44 years • 4554 = 45 to 54 years • 5564 = 55 to 64 years • 65 = 65 years or older

Step2.Data exploration.

This analysis is done using a subset for India country.

who_IN <- filter(who,who$iso2 == "IN")

View(who_IN) #Viewing new data frame

	country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp_m4554	new_sp_m5564	new_sp_m65	new_sp_f014	new_sp_f
1	India	IN	IND	1980	NA	NA	NA	NA	NA	NA	NA	NA	
2	India	IN	IND	1981	NA	NA	NA	NA	NA	NA	NA	NA	
3	India	IN	IND	1982	NA	NA	NA	NA	NA	NA	NA	NA	
4	India	IN	IND	1983	NA	NA	NA	NA	NA	NA	NA	NA	
5	India	IN	IND	1984	NA	NA	NA	NA	NA	NA	NA	NA	
6	India	IN	IND	1985	NA	NA	NA	NA	NA	NA	NA	NA	
7	India	IN	IND	1986	NA	NA	NA	NA	NA	NA	NA	NA	
8	India	IN	IND	1987	NA	NA	NA	NA	NA	NA	NA	NA	
9	India	IN	IND	1988	NA	NA	NA	NA	NA	NA	NA	NA	
10	India	IN	IND	1989	NA	NA	NA	NA	NA	NA	NA	NA	
11	India	IN	IND	1990	NA	NA	NA	NA	NA	NA	NA	NA	
12	India	IN	IND	1991	NA	NA	NA	NA	NA	NA	NA	NA	
13	India	IN	IND	1992	NA	NA	NA	NA	NA	NA	NA	NA	
14	India	IN	IND	1993	NA	NA	NA	NA	NA	NA	NA	NA	
15	India	IN	IND	1994	NA	NA	NA	NA	NA	NA	NA	NA	
16	India	IN	IND	1995	16	334	391	287	216	123	68	32	
17	India	IN	IND	1996	47	966	1143	934	666	424	213	79	
18	India	IN	IND	1997	50	1257	1351	1056	753	499	245	125	
19	India	IN	IND	1998	84	1773	2013	1851	1389	885	419	190	
20	India	IN	IND	1999	327	7058	8856	7900	6172	3864	1982	785	
21	India	IN	IND	2000	1588	20963	31090	30829	24230	15308	8534	2250	
22	India	IN	IND	2001	1063	22483	30007	29649	23961	14879	7779	2125	

There is total 34 rows for IN as country. Checking the summary statistics.

summary(who_IN)

```
  country              iso2               iso3                year     
Length:34          Length:34          Length:34          Min.   :1980  
Class :character   Class :character   Class :character   1st Qu.:1988  
Mode  :character   Mode  :character   Mode  :character   Median :1996  
                                                         Mean   :1996  
                                                         3rd Qu.:2005  
                                                         Max.   :2013  

 new_sp_m014     new_sp_m1524    new_sp_m2534    new_sp_m3544    new_sp_m4554   
Min.   :  16   Min.   :  334   Min.   :  391   Min.   :  287   Min.   :  216  
1st Qu.: 511   1st Qu.:10534   1st Qu.:14144   1st Qu.:13337   1st Qu.:10619  
Median :2784   Median :52230   Median :66945   Median :69019   Median :57519  
Mean   :2560   Mean   :43961   Mean   :50774   Mean   :53545   Mean   :46402  
3rd Qu.:4562   3rd Qu.:75113   3rd Qu.:81966   3rd Qu.:88095   3rd Qu.:78213  
Max.   :5001   Max.   :78278   Max.   :84003   Max.   :90830   Max.   :82921  
NA's   :16     NA's   :16      NA's   :16      NA's   :16      NA's   :16     
 new_sp_m5564     new_sp_m65     new_sp_f014     new_sp_f1524    new_sp_f2534   
Min.   :  123   Min.   :  68   Min.   :  32   Min.   :  179   Min.   :  169  
1st Qu.: 6618   1st Qu.: 3431   1st Qu.:1120   1st Qu.: 7746   1st Qu.: 7822  
Median :37254   Median :20203   Median :5302   Median :37764   Median :39563  
Mean   :32709   Mean   :19760   Mean   :4625   Mean   :30611   Mean   :30054  
3rd Qu.:56050   3rd Qu.:35040   3rd Qu.:8089   3rd Qu.:51186   3rd Qu.:48798  
Max.   :63814   Max.   :42443   Max.   :8576   Max.   :53975   Max.   :49887  
NA's   :16      NA's   :16      NA's   :16     NA's   :16      NA's   :16     
 new_sp_f3544    new_sp_f4554    new_sp_f5564     new_sp_f65       new_sn_m014   
Min.   :  80   Min.   :  49   Min.   :  30   Min.   :  11.0   Min.   : NA  
1st Qu.: 4606   1st Qu.: 2536   1st Qu.: 1512   1st Qu.:  708.5   1st Qu.: NA  
Median :25160   Median :15088   Median : 9014   Median : 4500.5   Median : NA  
```

From summary statistics we can see that there are 30 plus NA values for other diagnosis methods like 'sn','ep' etc except 'sp'. Hence, subsetting data for diagnosis method 'sp' for this analysis.

```
 new_sp_f3544    new_sp_f4554    new_sp_f5564     new_sp_f65       new_sn_m014   
Min.   :  80   Min.   :  49   Min.   :  30   Min.   :  11.0   Min.   : NA  
1st Qu.: 4606   1st Qu.: 2536   1st Qu.: 1512   1st Qu.:  708.5   1st Qu.: NA  
Median :25160   Median :15088   Median : 9014   Median : 4500.5   Median : NA  
Mean   :19966   Mean   :12696   Mean   : 8319   Mean   : 4731.8   Mean   :NaN  
3rd Qu.:33135   3rd Qu.:21194   3rd Qu.:14104   3rd Qu.: 8116.5   3rd Qu.: NA  
Max.   :34698   Max.   :23977   Max.   :17300   Max.   :10731.0   Max.   : NA  
NA's   :16      NA's   :16      NA's   :16      NA's   :16        NA's   :34   
 new_sn_m1524  new_sn_m2534  new_sn_m3544     new_sn_m4554  new_sn_m5564  
Min.   : NA   Min.   : NA   Min.   :250051   Min.   : NA   Min.   : NA  
1st Qu.: NA   1st Qu.: NA   1st Qu.:250051   1st Qu.: NA   1st Qu.: NA  
Median : NA   Median : NA   Median :250051   Median : NA   Median : NA  
Mean   :NaN   Mean   :NaN   Mean   :250051   Mean   :NaN   Mean   :NaN  
3rd Qu.: NA   3rd Qu.: NA   3rd Qu.:250051   3rd Qu.: NA   3rd Qu.: NA  
Max.   : NA   Max.   : NA   Max.   :250051   Max.   : NA   Max.   : NA  
NA's   :34    NA's   :34    NA's   :33       NA's   :34    NA's   :34   
  new_sn_m65   new_sn_f014   new_sn_f1524  new_sn_f2534  new_sn_f3544  
Min.   : NA   Min.   : NA   Min.   : NA   Min.   : NA   Min.   :148811  
1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.: NA   1st Qu.:148811  
Median : NA   Median : NA   Median : NA   Median : NA   Median :148811  
Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :NaN   Mean   :148811  
3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.: NA   3rd Qu.:148811  
Max.   : NA   Max.   : NA   Max.   : NA   Max.   : NA   Max.   :148811  
NA's   :34    NA's   :34    NA's   :34    NA's   :34    NA's   :33      
```

who_IN_sp <- select(who_IN,country,iso2,iso3,year,contains("sp"))

View(who_IN_sp)

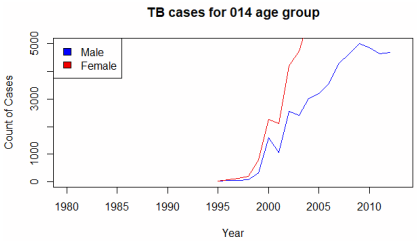
	country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp_m4554	new_sp_m5564	new_sp_m65	new_sp_f014	new_sp_f1524	ne
1	India	IN	IND	1980	NA	NA	NA	NA	NA	NA	NA	NA	NA	
2	India	IN	IND	1981	NA	NA	NA	NA	NA	NA	NA	NA	NA	
3	India	IN	IND	1982	NA	NA	NA	NA	NA	NA	NA	NA	NA	
4	India	IN	IND	1983	NA	NA	NA	NA	NA	NA	NA	NA	NA	
5	India	IN	IND	1984	NA	NA	NA	NA	NA	NA	NA	NA	NA	
6	India	IN	IND	1985	NA	NA	NA	NA	NA	NA	NA	NA	NA	
7	India	IN	IND	1986	NA	NA	NA	NA	NA	NA	NA	NA	NA	
8	India	IN	IND	1987	NA	NA	NA	NA	NA	NA	NA	NA	NA	
9	India	IN	IND	1988	NA	NA	NA	NA	NA	NA	NA	NA	NA	
10	India	IN	IND	1989	NA	NA	NA	NA	NA	NA	NA	NA	NA	
11	India	IN	IND	1990	NA	NA	NA	NA	NA	NA	NA	NA	NA	
12	India	IN	IND	1991	NA	NA	NA	NA	NA	NA	NA	NA	NA	
13	India	IN	IND	1992	NA	NA	NA	NA	NA	NA	NA	NA	NA	
14	India	IN	IND	1993	NA	NA	NA	NA	NA	NA	NA	NA	NA	
15	India	IN	IND	1994	NA	NA	NA	NA	NA	NA	NA	NA	NA	
16	India	IN	IND	1995	16	334	391	287	216	123	68	32	179	
17	India	IN	IND	1996	47	966	1143	934	666	424	213	79	618	
18	India	IN	IND	1997	50	1257	1351	1056	753	499	245	125	861	
19	India	IN	IND	1998	84	1773	2013	1651	1389	885	419	190	1375	
20	India	IN	IND	1999	327	7058	8856	7900	6172	3864	1982	785	5497	
21	India	IN	IND	2000	1588	20963	31090	30829	24230	15308	8534	2250	14495	
22	India	IN	IND	2001	1063	22483	30007	29649	23961	14879	7779	2125	15973	

Visualizing new TB cases for different age groups from 1980-2013.

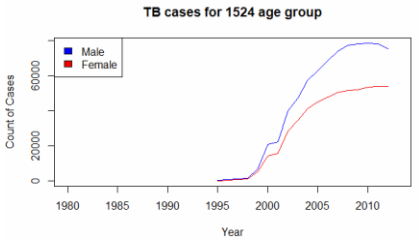
 # Function for plotting new cases for different age groups

```
plotNewTBCases <- function(ageGrp) {
  plot(who_IN_sp[[paste("new_sp_m",ageGrp,sep="")]]~who_IN_sp$year,
       xlab="Year",
       ylab="Count of Cases",
       main=paste("TB cases for",ageGrp,"age group"),
       type="l",
       col="blue")
  lines(who_IN_sp[[paste("new_sp_f",ageGrp,sep="")]]~who_IN_sp$year,col="red")
  legend("topleft",c("Male","Female"),fill=c("blue","red"))
}
```

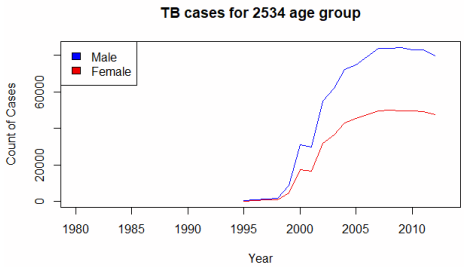
plotNewTBCases("014") #0-14



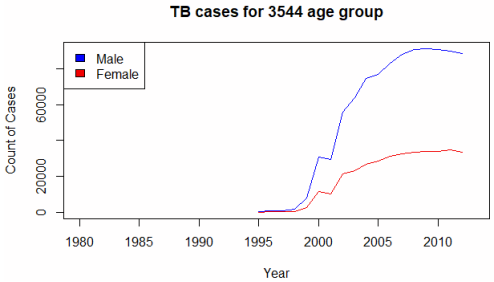
plotNewTBCases("1524") #15-24



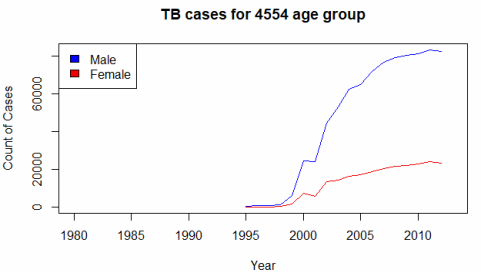
plotNewTBCases("2534") #25-34



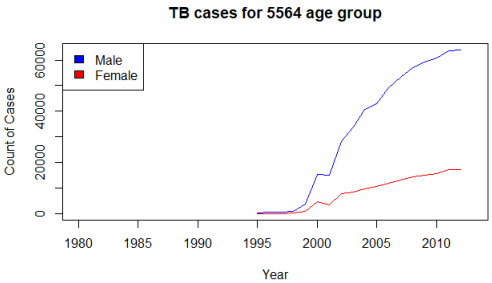
plotNewTBCases ("3544") #35-44



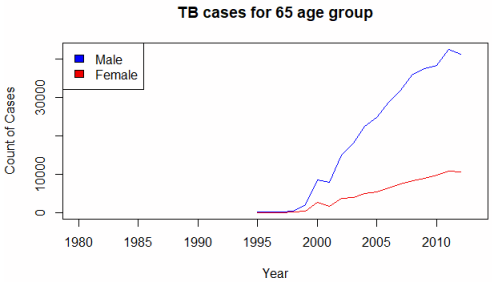
plotNewTBCases("4554") #45-54



plotNewTBCases("5564") #55-64



plotNewTBCases("65") #65 and above



Observations:

1.All the plots show visible trend from year 1995 to 2015 as from the year 1980 to 1994 the values are NA. Verifying this by calculating summary stats for years 1980-1994.

who_IN_SP_1980_1994 <- filter(who_IN_sp,between(year,1980,1994))

View(who_IN_SP_1980_1994)

	country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp_m4554	new_sp_m5564	new_sp_m65	new_sp_f014	new_sp_f1524	ne
1	India	IN	IND	1980	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	India	IN	IND	1981	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	India	IN	IND	1982	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	India	IN	IND	1983	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	India	IN	IND	1984	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	India	IN	IND	1985	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	India	IN	IND	1986	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	India	IN	IND	1987	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	India	IN	IND	1988	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	India	IN	IND	1989	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	India	IN	IND	1990	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	India	IN	IND	1991	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	India	IN	IND	1992	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	India	IN	IND	1993	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	India	IN	IND	1994	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

summary(who_IN_SP_1980_1994)

```
   country              iso2               iso3                year     
 Length:15          Length:15          Length:15          Min.   :1980  
 Class :character   Class :character   Class :character   1st Qu.:1984  
 Mode  :character   Mode  :character   Mode  :character   Median :1987  
                                                          Mean   :1987  
                                                          3rd Qu.:1990  
                                                          Max.   :1994  

  new_sp_m014     new_sp_m1524    new_sp_m2534    new_sp_m3544    new_sp_m4554   
 Min.   : NA     Min.   : NA     Min.   : NA     Min.   : NA     Min.   : NA    
 1st Qu.: NA     1st Qu.: NA     1st Qu.: NA     1st Qu.: NA     1st Qu.: NA    
 Median : NA     Median : NA     Median : NA     Median : NA     Median : NA    
 Mean   :NaN     Mean   :NaN     Mean   :NaN     Mean   :NaN     Mean   :NaN    
 3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA    
 Max.   : NA     Max.   : NA     Max.   : NA     Max.   : NA     Max.   : NA    
 NA's   :15      NA's   :15      NA's   :15      NA's   :15      NA's   :15     
  new_sp_m5564    new_sp_m65     new_sp_f014     new_sp_f1524    new_sp_f2534   
 Min.   : NA     Min.   : NA     Min.   : NA     Min.   : NA     Min.   : NA    
 1st Qu.: NA     1st Qu.: NA     1st Qu.: NA     1st Qu.: NA     1st Qu.: NA    
 Median : NA     Median : NA     Median : NA     Median : NA     Median : NA    
 Mean   :NaN     Mean   :NaN     Mean   :NaN     Mean   :NaN     Mean   :NaN    
 3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA    
 Max.   : NA     Max.   : NA     Max.   : NA     Max.   : NA     Max.   : NA    
 NA's   :15      NA's   :15      NA's   :15      NA's   :15      NA's   :15     
  new_sp_f3544    new_sp_f4554    new_sp_f5564    new_sp_f65    
 Min.   : NA     Min.   : NA     Min.   : NA     Min.   : NA    
 1st Qu.: NA     1st Qu.: NA     1st Qu.: NA     1st Qu.: NA    
 Median : NA     Median : NA     Median : NA     Median : NA    
 Mean   :NaN     Mean   :NaN     Mean   :NaN     Mean   :NaN    
 3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA     3rd Qu.: NA    
 Max.   : NA     Max.   : NA     Max.   : NA     Max.   : NA    
 NA's   :15      NA's   :15      NA's   :15      NA's   :15     
> 
```

From summary we can see that all columns from new_sp_m014 to new_sp_f65 have NA values. Hence we remove these rows before performing the further analysis.

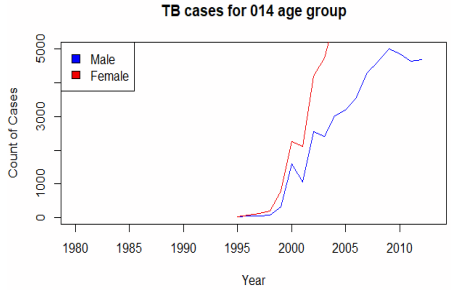
who_mod <- filter(who_IN_sp,between(year,1995,2012))

View(who_mod)

	country	iso2	iso3	year	new_sp_m014	new_sp_m1524	new_sp_m2534	new_sp_m3544	new_sp_m4554	new_sp_m5564	new_sp_m65
1	India	IN	IND	1995	16	334	391	287	216	123	
2	India	IN	IND	1996	47	966	1143	934	666	424	
3	India	IN	IND	1997	50	1257	1351	1056	753	499	
4	India	IN	IND	1998	84	1773	2013	1851	1389	885	
5	India	IN	IND	1999	327	7058	8856	7900	6172	3864	
6	India	IN	IND	2000	1588	20963	31090	30829	24230	15308	
7	India	IN	IND	2001	1063	22483	30007	29649	23961	14879	
8	India	IN	IND	2002	2551	39923	54719	55829	44532	28199	
9	India	IN	IND	2003	2411	47251	61758	63587	52865	33739	
10	India	IN	IND	2004	3018	57208	72132	74450	62173	40769	
11	India	IN	IND	2005	3185	62620	74678	76870	64843	43038	
12	India	IN	IND	2006	3566	68346	79037	82939	71621	49320	
13	India	IN	IND	2007	4305	73947	83850	88045	76408	53414	
14	India	IN	IND	2008	4648	77121	83798	90498	78815	56928	
15	India	IN	IND	2009	5001	78177	84003	90830	80097	59163	
16	India	IN	IND	2010	4871	78278	82757	90440	81210	60766	
17	India	IN	IND	2011	4649	78096	82762	89706	82921	63625	
18	India	IN	IND	2012	4697	75502	79594	88111	82356	63814	

2.From the line plots for all age groups in Step1, we can say that as from 1995 to 2005 count of new cases increases for both males and females and then slight dip is observed.

3. From the plot for 0-14 age group, it looks that the number of cases for females are higher than males over the years. Let's calculate the average number of new cases for both males and females.



colnames(who_mod)

summary(select(who_mod,new_sp_m014,new_sp_f014))

```
  new_sp_m014     new_sp_f014    
 Min.   :  16   Min.   :  32  
 1st Qu.: 511   1st Qu.:1120  
 Median :2784   Median :5302  
 Mean   :2560   Mean   :4625  
 3rd Qu.:4562   3rd Qu.:8089  
 Max.   :5001   Max.   :8576  
> 
```

From summary stats we can see that the average number of new cases for females is 4625 which are nearly twice of the number of cases for males i.e., 2560.

4.Plot the average count of cases(average of cases from 1994-2012) for males across different age groups.

#Creating a new data frame with average number of cases for both males and females across all age groups

mean_males <- c(mean(who_mod$new_sp_m014,na.rm = TRUE),mean(who_mod$new_sp_m1524,na.rm = TRUE),mean(who_mod$new_sp_m2534,na.rm = TRUE),mean(who_mod$new_sp_m3544,na.rm = TRUE),mean(who_mod$new_sp_m4554,na.rm = TRUE),mean(who_mod$new_sp_m5564,na.rm = TRUE),mean(who_mod$new_sp_m65,na.rm = TRUE))

mean_females <- c(mean(who_mod$new_sp_f014,na.rm = TRUE),mean(who_mod$new_sp_f1524,na.rm = TRUE),mean(who_mod$new_sp_f2534,na.rm = TRUE),mean(who_mod$new_sp_f3544,na.rm = TRUE),mean(who_mod$new_sp_f4554,na.rm = TRUE),mean(who_mod$new_sp_f5564,na.rm = TRUE),mean(who_mod$new_sp_f65,na.rm = TRUE))

sd_males <-c(sd(who_mod$new_sp_m014,na.rm = TRUE),sd(who_mod$new_sp_m1524,na.rm = TRUE),sd(who_mod$new_sp_m2534,na.rm = TRUE),sd(who_mod$new_sp_m3544,na.rm = TRUE),sd(who_mod$new_sp_m4554,na.rm = TRUE),sd(who_mod$new_sp_m5564,na.rm = TRUE),sd(who_mod$new_sp_m65,na.rm = TRUE))

sd_females <- c(sd(who_mod$new_sp_f014,na.rm = TRUE),sd(who_mod$new_sp_f1524,na.rm = TRUE),sd(who_mod$new_sp_f2534,na.rm = TRUE),sd(who_mod$new_sp_f3544,na.rm = TRUE),sd(who_mod$new_sp_f4554,na.rm = TRUE),sd(who_mod$new_sp_f5564,na.rm = TRUE),sd(who_mod$new_sp_f65,na.rm = TRUE))

n_males <- c(18,18,18,18,18,18,18)
n_females <- c(18,18,18,18,18,18,18)

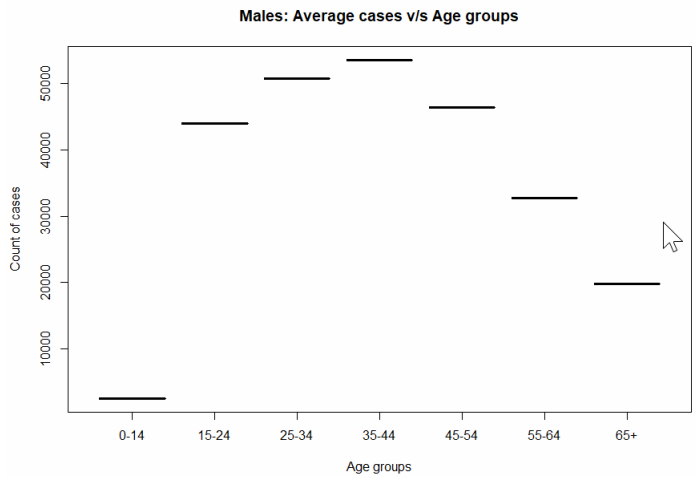
```
who_avg_mf <- data.frame("Age group" = c("0-14","15-24","25-34","35-44","45-54","55-64","65+"),
           "Mean_cases_males" = mean_males,
           "Mean_cases_females" = mean_females,
           "SD_cases_males" = sd_males,
           "SD_cases_females" = sd_females,
           "N_males" = n_males,
           "N_females" = n_females)
```

View(who_avg_mf)

	Age.group	Mean_cases_males	Mean_cases_females	SD_cases_males	SD_cases_females	N_males	N_females
1	0-14	2559.833	4625.333	1919.853	3417.812	18	18
2	15-24	43961.278	30611.333	31955.812	21949.135	18	18
3	25-34	50774.389	30054.111	34747.731	20846.947	18	18
4	35-44	53545.056	19965.722	37544.445	14117.674	18	18
5	45-54	46401.556	12695.722	33556.694	9269.348	18	18
6	55-64	32708.722	8319.333	24854.544	6407.439	18	18
7	65+	19759.722	4731.778	15909.239	3892.235	18	18

Visualizing average count of cases across all age groups for males.

```
boxplot(who_avg_mf$Mean_cases_males~who_avg_mf$Age.group,ylab="Count of cases",
xlab="Age groups",main="Males: Average cases v/s Age groups")
```

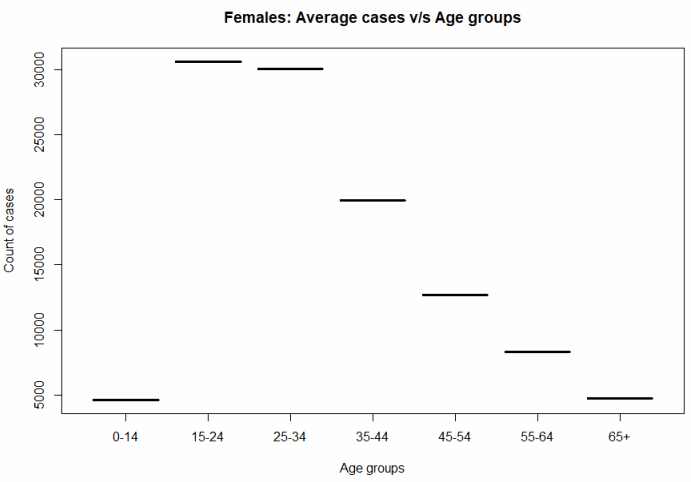


Conclusion: From the above plots we can say that average number of cases vary for males across different age groups.

We can test this using hypothesis testing.

Visualizing average count of cases across all age groups for females.

```
boxplot(who_avg_mf$Mean_cases_females~who_avg_mf$Age.group,ylab="Count of cases",
xlab="Age groups",main="Females: Average cases v/s Age groups")
```



Conclusion: From the above plots we can say that average number of cases vary for females as well across different age groups.

We can test this claim using hypothesis testing.

Step3.Questions to be tested using Hypothesis testing (Reasoning explained in Step3)

Question1: The average number of new TB cases are different for different age groups for males.

Question2: The average number of new TB cases are different for different age groups for females.

Test the claim using ANOVA.

Step4.Hypothesis testing using ANNOVA.

Testing claim: The average number of new TB cases are different for different age groups for males

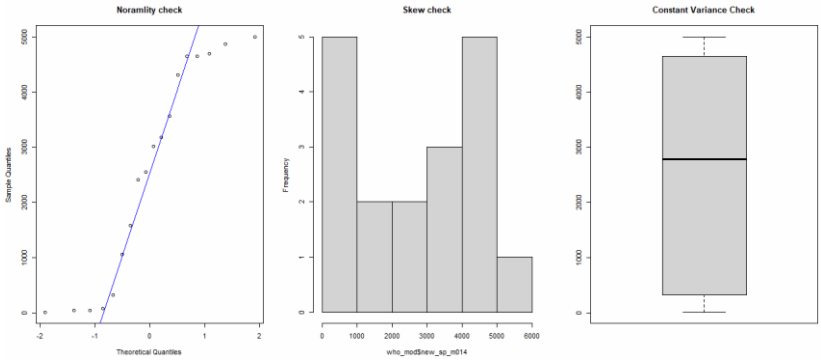
a. Question1: The average number of new TB cases are different for different age groups for males. Test using ANOVA test statistics.
b. Reason: As seen from the boxplot in step2 point 4 for males we see considerable difference in average number of new cases for different age groups. Hence, decided to confirm this claim using hypothesis testing.
c. Type of test: ANOVA analysis using F statistics.
d. Checking model assumptions

1.Independence: In this study the diagnosis of TB in single individual is independent of another individual. Hence observations are independent.

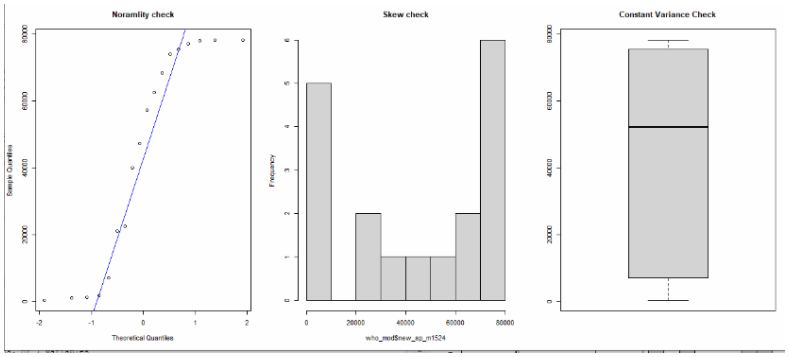
2.Normal approximation: By plotting the normal qqplot and histogram we can verify normality of data for all age groups.

3.Constant Variance: By plotting boxplots we can check for constant variance across all age groups.

#0-14
```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_m014,main="Noramlity check")
qqline(who_mod$new_sp_m014,col = "blue")
hist(who_mod$new_sp_m014,main="Skew check")
boxplot(who_mod$new_sp_m014,main="Constant Variance Check")
```

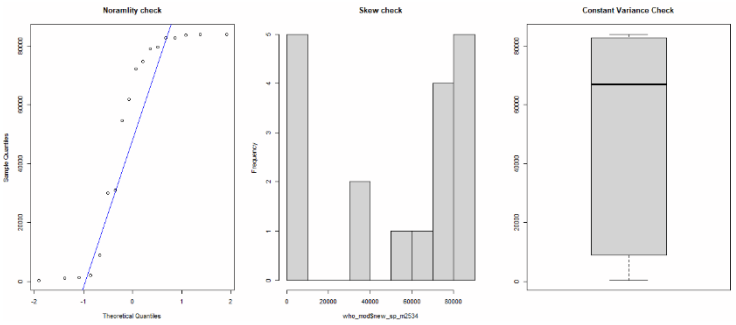


#15-24
```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_m1524,main="Noramlity check")
qqline(who_mod$new_sp_m1524,col="blue")
hist(who_mod$new_sp_m1524,main="Skew check")
boxplot(who_mod$new_sp_m1524,main="Constant Variance Check")
```



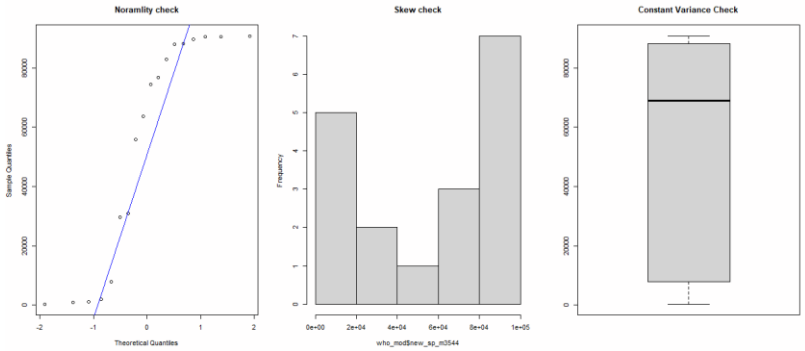
#25-34

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_m2534, main="Noramlity
check")
qqline(who_mod$new_sp_m2534, col = "blue")
hist(who_mod$new_sp_m2534,main="Skew check")
boxplot(who_mod$new_sp_m2534, main="Constant Variance
Check" )
```



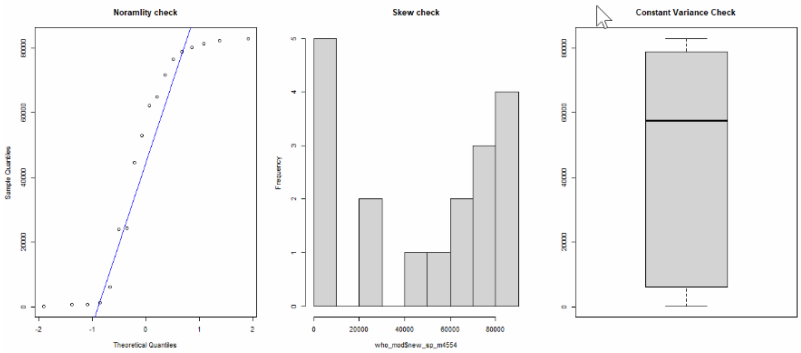
#35-44

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_m3544,main="Noramlity check")
qqline(who_mod$new_sp_m3544, col = "blue")
hist(who_mod$new_sp_m3544,main="Skew check")
boxplot(who_mod$new_sp_m3544,main="Constant Variance
Check")
```



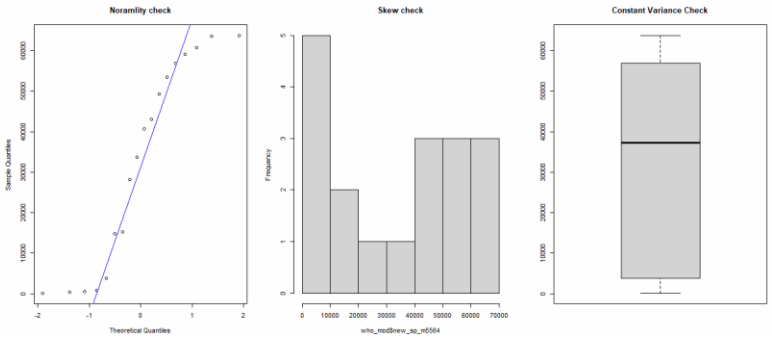
#45-54

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_m4554,main="Noramlity check")
qqline(who_mod$new_sp_m4554, col = "blue")
hist(who_mod$new_sp_m4554,main="Skew check")
boxplot(who_mod$new_sp_m4554,main="Constant Variance
Check")
```

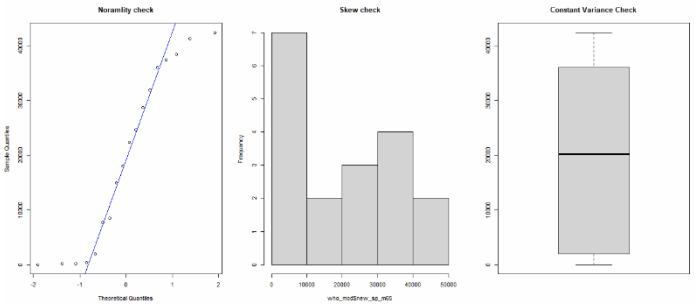


#55-64

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_m5564,main="Noramlity check")
qqline(who_mod$new_sp_m5564,col = "blue")
hist(who_mod$new_sp_m5564,main="Skew check")
boxplot(who_mod$new_sp_m5564,main="Constant Variance
Check")
```



```
#65+
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_m65,main="Noramlity check")
qqline(who_mod$new_sp_m65, col = "blue")
hist(who_mod$new_sp_m65,main="Skew check")
boxplot(who_mod$new_sp_m65,main="Constant Variance
Check")
```



Conclusion: From all normality plots we can see a normal trend, histogram doesn't show signs of strong skewness. Finally, from boxplots we can conclude constant variance.

e. Null and alternate hypothesis

H_0: $\mu_{m\_014} = \mu_{m\_1524} = \mu_{m\_2534} = \mu_{m\_3544} = \mu_{m\_4554} = \mu_{m\_5564} = \mu_{m\_65}$
H_A: At least average number of cases is different for one age group for males.

f. Calculating f statistics. (Using data frame 'who_avg_mf')
Note: I have not used R's inbuilt annova() to calculate values because dataset was not in the correct format for lm() model. So, calculated them using math formulas and then verified the result with online ANNOVA calculator (https://goodcalculators.com/one-way-anova-calculator/).

```
# Total sample size
(n<- sum(who_avg_mf$N_males))
>126

# Total groups
(k <- length(who_avg_mf$Age.group))
>7

# Degree of freedom for groups, error, and total.
(dfg <- k-1)
>6

(dft <- n-1)
>125

(dfe <- dft-dfg)
> 119

#Determine mean of mean number of cases for all groups
t_mean <- mean(who_avg_mf$Mean_cases_males)
> 35672.94

#Appending number of cases for all age groups for males
(no_of_sp_m <- (c(who_mod$new_sp_m014,who_mod$new_sp_m1524,who_mod$new_sp_m2534,
            who_mod$new_sp_m3544,who_mod$new_sp_m4554,who_mod$new_sp_m5564,
            who_mod$new_sp_m65)))

# Sum of squares total
(SST <- sum((no_of_sp_m-t_mean)^2))
> 133474512097

# Sum of squares between groups
(SSG <- sum(who_avg_mf$N_males * (who_avg_mf$Mean_cases_males - t_mean)^2))
> 37615696704
```

Sum of squares error
(SSE <- SST - SSG)
> 95858815393
Mean square error
(MSE <- SSE/dfe)
> 805536264

Mean square for groups
(MSG <- SSG/dfg)
> 6269282784

F-statistics
(F <- MSG/MSE)
> 7.782744

p-value
(round(pf(F,dfg,dfe,lower.tail = FALSE))) #Rounding off value 4.487155e-07 to 0
>0

g. Conclusion:

ANOVA Summary					
Source	Degrees of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-Value
Between Groups	6	37615696817.7622	6269282802.9604	7.7827	0
Within Groups	119	95858815348.1566	805536263.4299		
Total:	125	133474512165.9187			

(Table from online ANNOVA calculator: https://goodcalculators.com/one-way-anova-calculator/)

The p-value is 0 which is less than significance level of 0.05. Hence, we reject the null hypothesis.
We have enough evidence to say that average number of new cases for males are different for at least one of the age groups.

Testing claim: The average number of new TB cases are different for different age groups for females

a. Question1: The average number of new TB cases are different for different age groups for females. Test using ANOVA test statistics.
b. Reason: As seen from the boxplot in step2 point 4 for females we see considerable difference in average number of new cases for different age groups. Hence, decided to confirm this claim using hypothesis testing.
c. Type of test: ANOVA analysis using F statistics.
d. Checking model assumptions

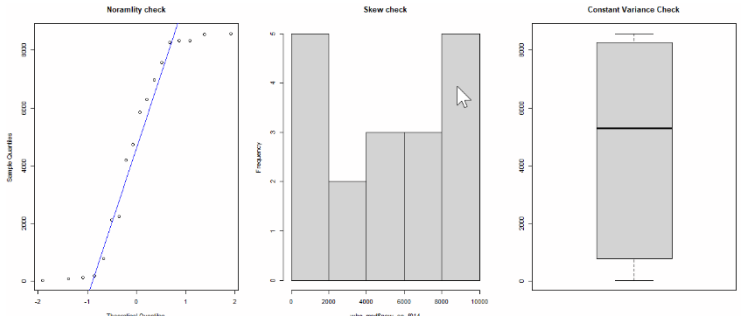
1.Independence: In this study the diagnosis of TB in single individual is independent of another individual. Hence observations are independent.

2.Normal approximation: By plotting the normal qqplot and histogram we can verify normality of data for all age groups.

3.Constant Variance: By plotting boxplots we can check for constant variance across all age groups.

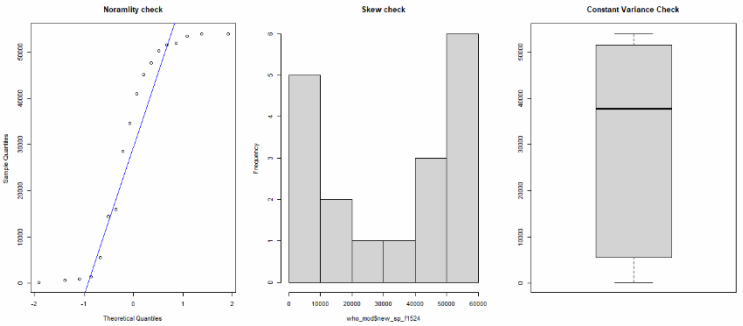
#0-14

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_f014,main="Noramlity check")
qqline(who_mod$new_sp_f014,col="blue")
hist(who_mod$new_sp_f014,main="Skew check")
boxplot(who_mod$new_sp_f014,main="Constant Variance Check")
```



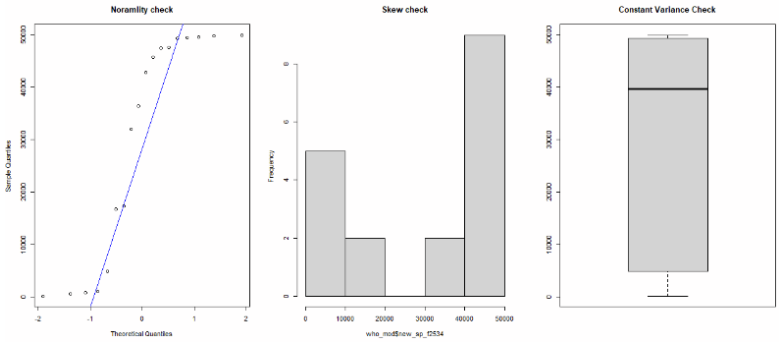
#15-24

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_f1524,main="Noramlity check")
qqline(who_mod$new_sp_f1524,col="blue")
hist(who_mod$new_sp_f1524,main="Skew check")
boxplot(who_mod$new_sp_f1524,main="Constant Variance Check")
```



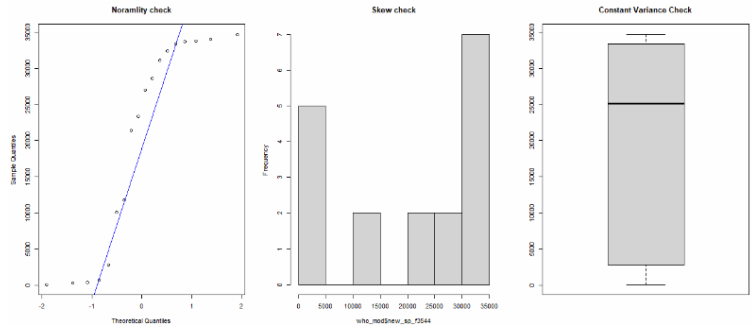
#25-34

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_f2534,main="Noramlity check")
qqline(who_mod$new_sp_f2534,col = "blue")
hist(who_mod$new_sp_f2534,main="Skew check")
boxplot(who_mod$new_sp_f2534,main="Constant Variance Check" )
```



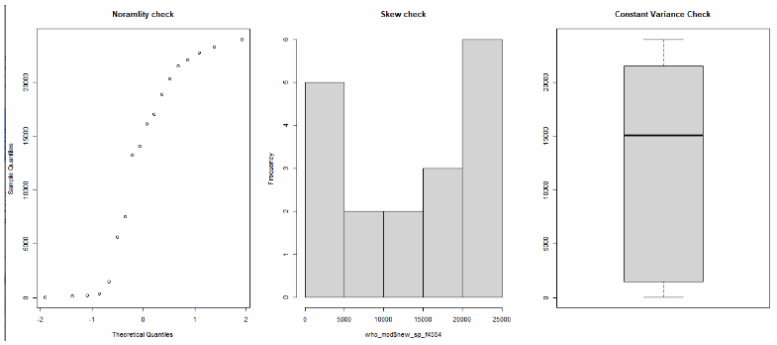
#35-44

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_f3544,main="Noramlity check")
qqline(who_mod$new_sp_f3544, col = "blue")
hist(who_mod$new_sp_f3544,main="Skew check")
boxplot(who_mod$new_sp_f3544,main="Constant Variance Check")
```



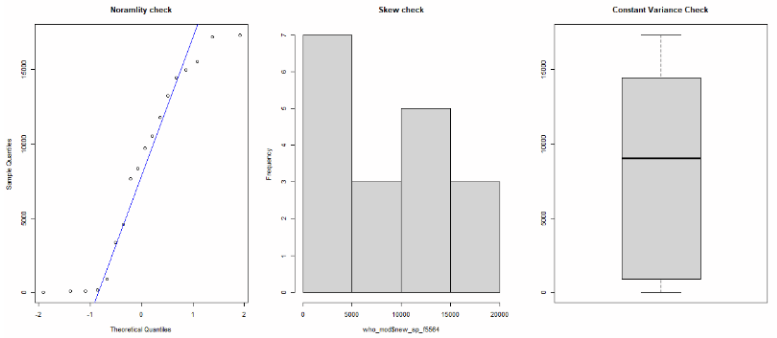
#45-54

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_f4554,main="Noramlity check")
qqline(who_mod$new_sp_4554, col = "blue")
hist(who_mod$new_sp_f4554,main="Skew check")
boxplot(who_mod$new_sp_f4554,main="Constant Variance Check")
```



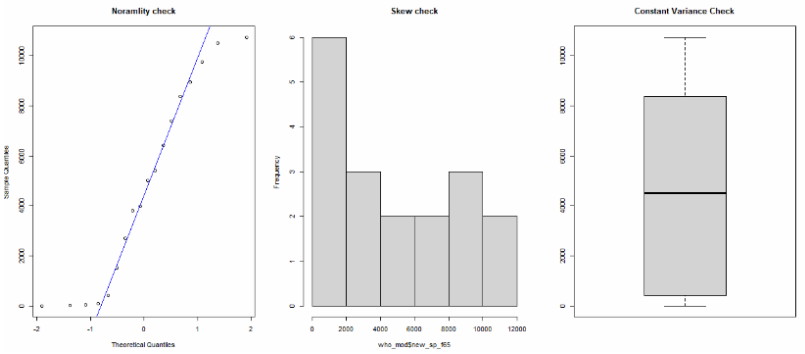
#55-64

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_f5564,main="Noramlity check")
qqline(who_mod$new_sp_f5564,col = "blue")
hist(who_mod$new_sp_f5564,main="Skew check")
boxplot(who_mod$new_sp_f5564,main="Constant Variance Check")
```



#65+

```
par(mfrow=c(1,3))
qqnorm(who_mod$new_sp_f65,main="Noramlity check")
qqline(who_mod$new_sp_f65, col = "blue")
hist(who_mod$new_sp_f65,main="Skew check")
boxplot(who_mod$new_sp_f65,main="Constant Variance Check")
```



Conclusion: From all normality plots we can see a normal trend, histogram doesn't show signs of strong skewness. Finally, from boxplots we can conclude constant variance.

e. Null and alternate hypothesis

H_0: $\mu_{f\_014} = \mu_{f\_1524} = \mu_{f\_2534} = \mu_{f\_3544} = \mu_{f\_4554} = \mu_{f\_5564} = \mu_{f\_65}$
H_A: At least average number of cases is different for one age group for females.

f. Calculating f statistics. (Using data frame 'who_avg_mf')
Note: I have not used R's inbuilt annova() to calculate values because dataset was not in the correct format for lm() model. So, calculated them using math formulas and then verified the result with online ANNOVA calculator (https://goodcalculators.com/one-way-anova-calculator/).

```
# Total sample size
(n<- sum(who_avg_mf$N_females))
>126

# Total groups
(k <- length(who_avg_mf$Age.group))
>7
```

```
# Degree of freedom for groups, error, and total.
(dfg <- k-1)
>6

(dft <- n-1)
>125

(dfe <- dft-dfg)
> 119

#Determine mean of mean number of cases for all groups
t_mean <- mean(who_avg_mf$Mean_cases_females)
> 15857.62

#Appending number of cases for all observations
(no_of_sp_f <- (c(who_mod$new_sp_f014,who_mod$new_sp_f1524,who_mod$new_sp_f2534,
            who_mod$new_sp_f3544,who_mod$new_sp_f4554,who_mod$new_sp_f5564,
            who_mod$new_sp_f65)))

# Sum of squares total
(SST <- sum((no_of_sp_f-t_mean)^2))
> 35132579340

# Sum of squares between groups
(SSG <- sum(who_avg_mf$N_females * (who_avg_mf$Mean_cases_females - t_mean)^2))
> 13551496438

# Sum of squares error
(SSE <- SST - SSG)
> 21581082902

# Mean square error
(MSE <- SSE/dfe)
> 181353638

# Mean square for groups
(MSG <- SSG/dfg)
> 2258582740

# F-statistics
(F <- MSG/MSE)
>  12.45403

# p-value
(round(pf(F,dfg,dfe,lower.tail = FALSE)))    #Rounding off value 7.457776e-11 to 0
>0
```

g. Conclusion:

ANOVA Summary					
Source	Degrees of Freedom DF	Sum of Squares SS	Mean Square MS	F-Stat	P-Value
Between Groups	6	13551496427.0872	2258582737.8479	12.454	0
Within Groups	119	21581082896.0756	181353637.7821		
Total:	125	35132579323.1628			

(Table from online ANNOVA calculator: https://goodcalculators.com/one-way-anova-calculator/)

The p-value is 0 which is less than significance level of 0.05. Hence, we reject the null hypothesis.
We have enough evidence to say that average number of new cases for females are different for at least one of the age groups.