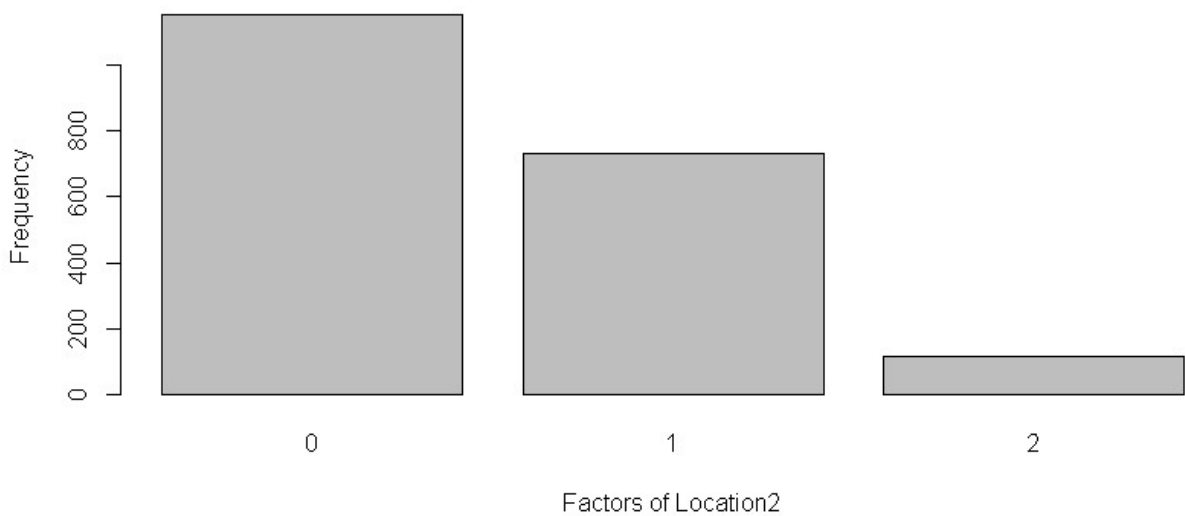
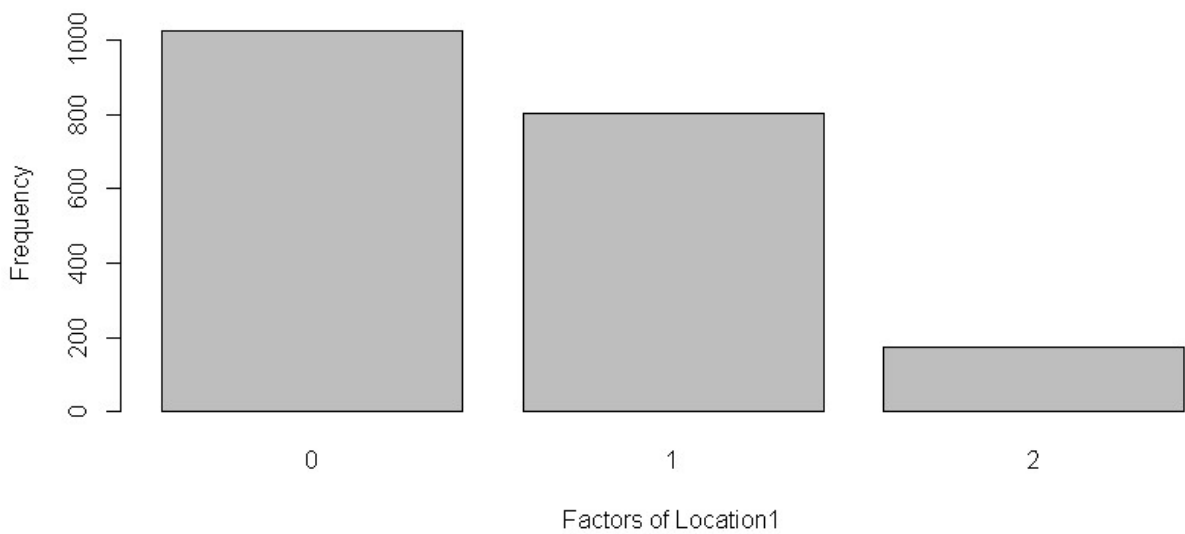


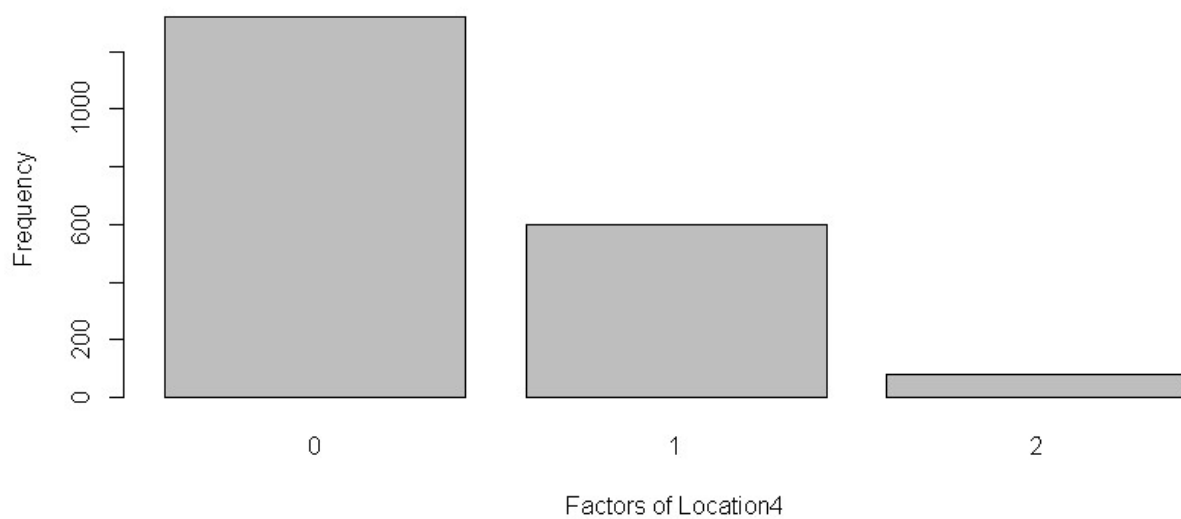
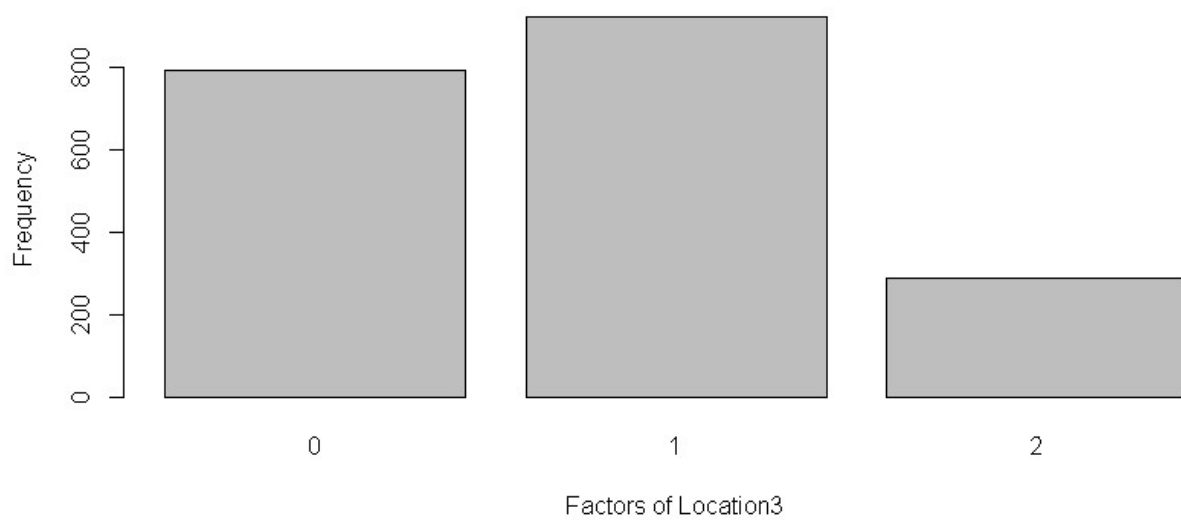
Initial Analysis

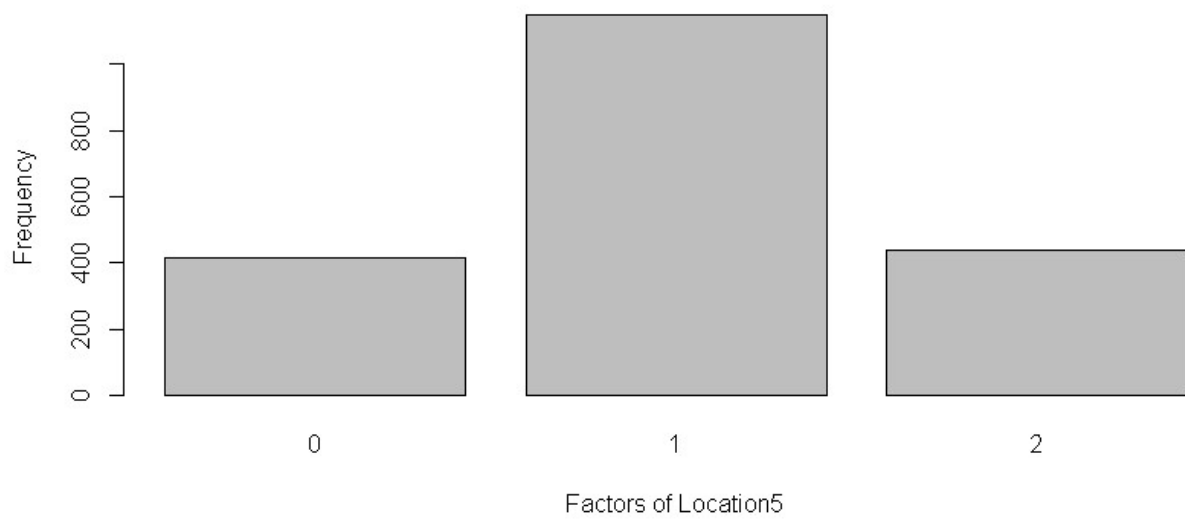
Summary of disease dataset;

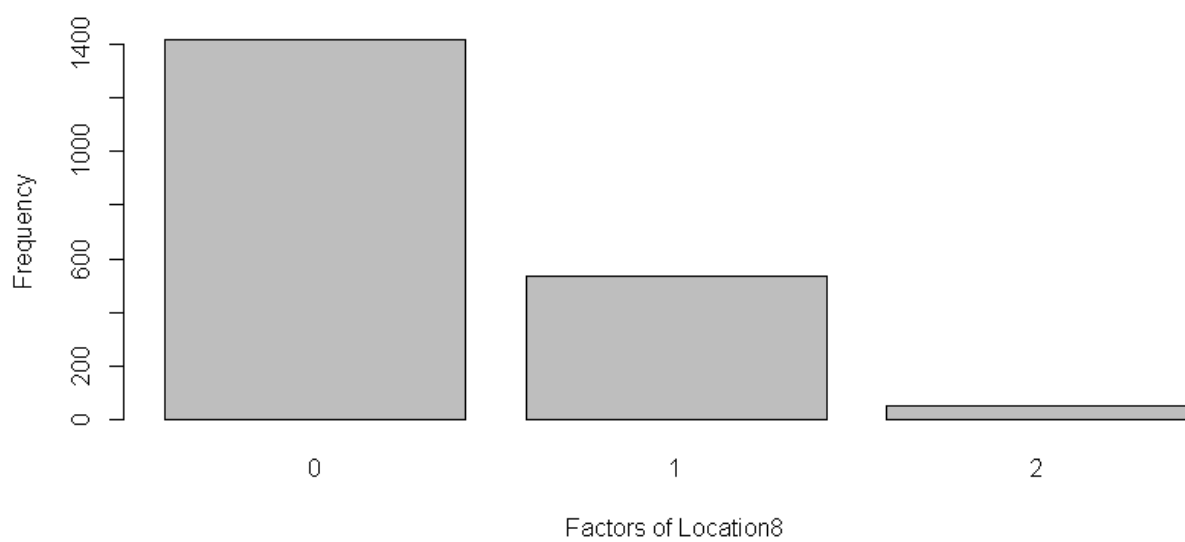
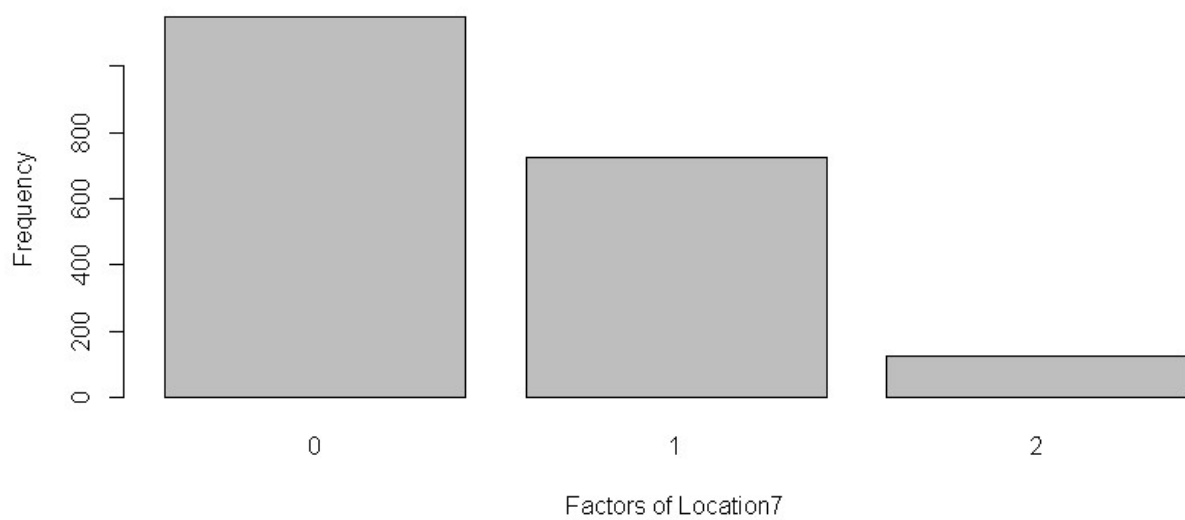
val	loc1	loc2	loc3	loc4	loc5	loc6	loc7	loc8	loc9	loc10	class
0	1025	1151	790	1319	415	868	1148	1416	916	436	950
1	803	732	920	602	1148	888	726	533	863	1166	1050
2	172	117	290	79	437	244	126	51	221	398	-

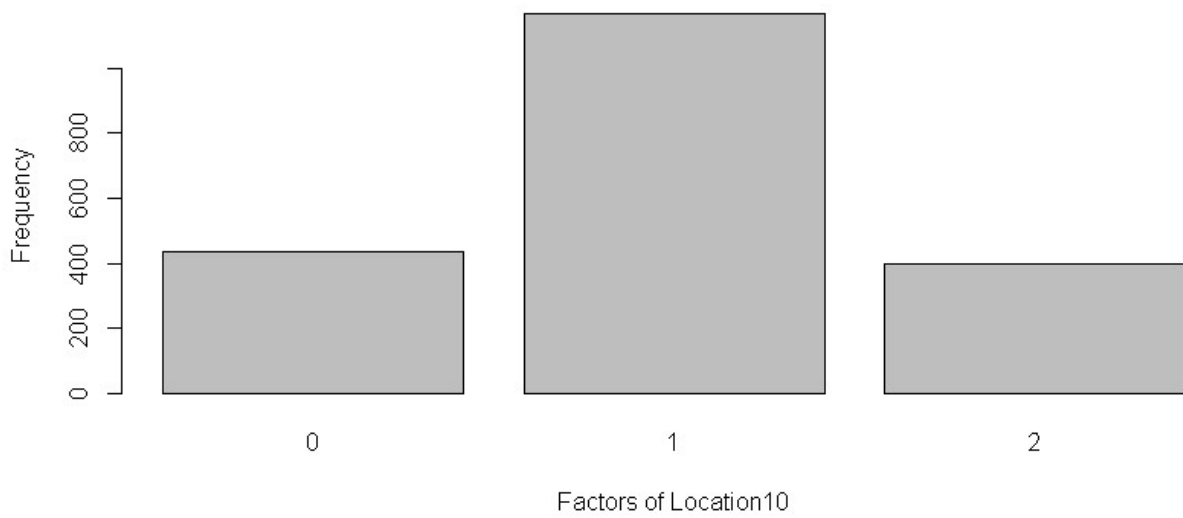
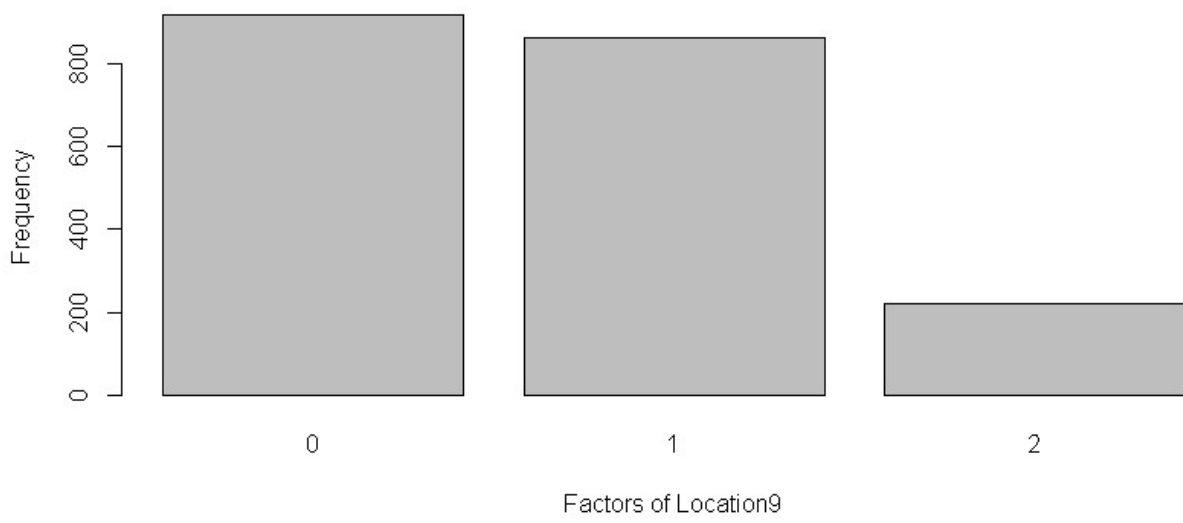
Histogram/Frequency Plots of factor values;



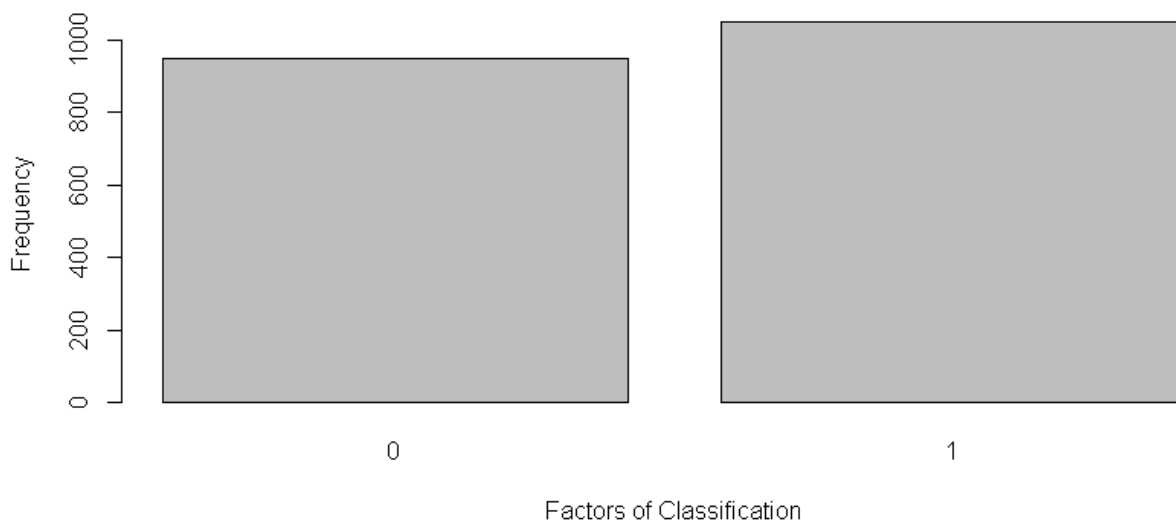








ACTUAL OBSERVATION OF CLASS;



Some Comparisons:

Model Name	AIC	Residual Deviance
All Main Effect (with Intercept)	2742.90	2700.90000
All Main Effect – Intercept	2742.90	2700.90000
Best Main [class~location1+location5+location10-1]	2722.60	2708.60000
All Interactions (Two way) – Intercept [class~.*.-1]	1832.50	1430.50000
Best Fit [class~location1+location5+location10+.*.-1]	1832.50	1430.50000
Full Model [class~.+.*.-1]	1832.50	1430.50000

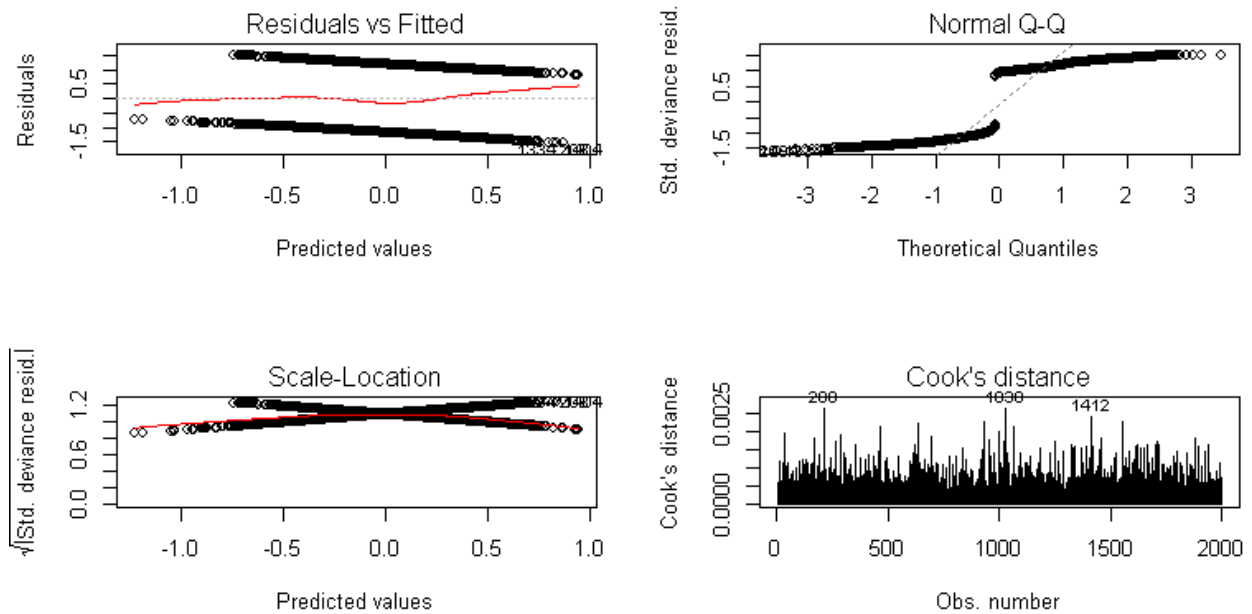
Significant Main Effects Predictors (Model 3) has following coefficients:

location10	location11	location12	location51	location52	location101	location102
0.5318079	0.4183030	0.4941323	1.6093885	1.7706004	1.8807271	1.4159356

Addendum:

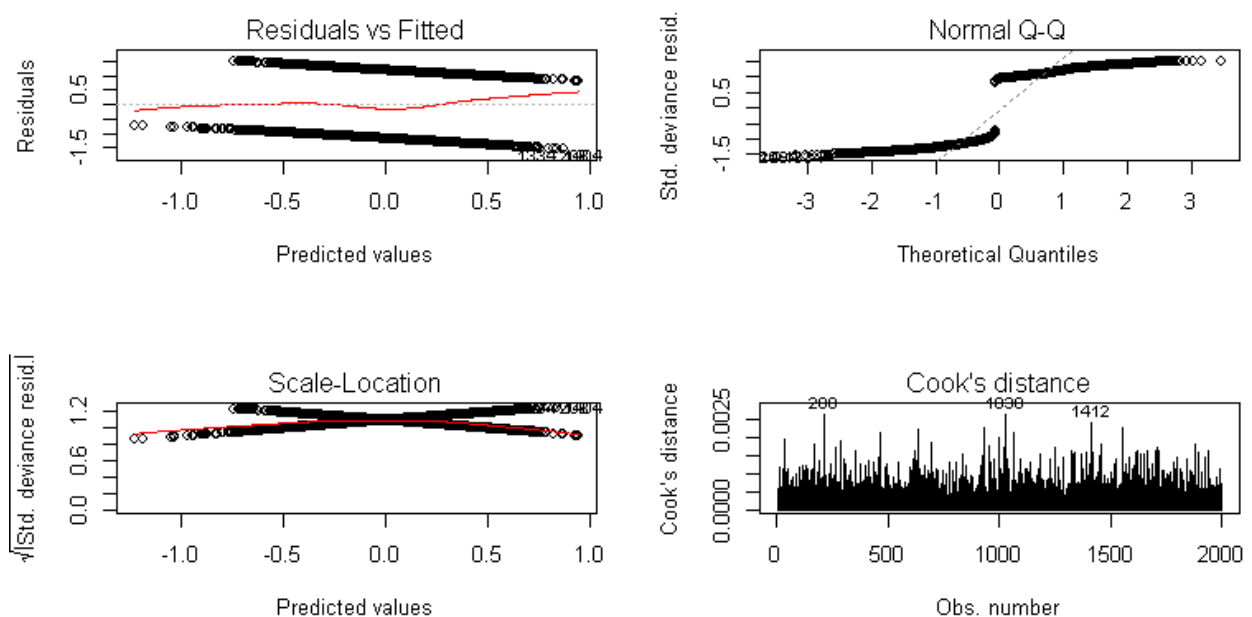
All Main Effect

Plots



All Main Effect – Intercept

Plots



Summary (All Main Effect with Intercept)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.690896	0.192695	-3.585	0.000337 ***
location11	-0.248695	0.096346	-2.581	0.009844 **
location12	-0.065724	0.168515	-0.390	0.696524

...

location42	0.009934	0.236499	0.042	0.966494
location51	0.485951	0.129270	3.759	0.000170 ***
location52	0.585198	0.153317	3.817	0.000135 ***
location61	0.022736	0.097615	0.233	0.815827
location62	0.198102	0.148934	1.330	0.183475
location71	0.033391	0.097068	0.344	0.730847
location72	0.071946	0.192655	0.373	0.708817
location81	0.077968	0.104523	0.746	0.455701
location82	-0.181098	0.292559	-0.619	0.535908
location91	-0.141754	0.097229	-1.458	0.144858
location92	0.070989	0.153832	0.461	0.644458
location101	0.624517	0.128318	4.867	1.13e-06 ***
location102	0.336888	0.154433	2.181	0.029150 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2767.6 on 1999 degrees of freedom

Residual deviance: 2700.9 on 1979 degrees of freedom

AIC: 2742.9

Number of Fisher Scoring iterations: 4

Summary (Without Intercept)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5601	-1.1938	0.9208	1.0755	1.5054

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
location10	-0.690896	0.192695	-3.585	0.000337 ***
location11	-0.939591	0.198204	-4.741	2.13e-06 ***
location12	-0.756620	0.242192	-3.124	0.001784 **
location21	0.039915	0.096852	0.412	0.680245

...

location42	0.009934	0.236499	0.042	0.966494
location51	0.485951	0.129270	3.759	0.000170 ***
location52	0.585198	0.153317	3.817	0.000135 ***
location61	0.022736	0.097615	0.233	0.815827

..

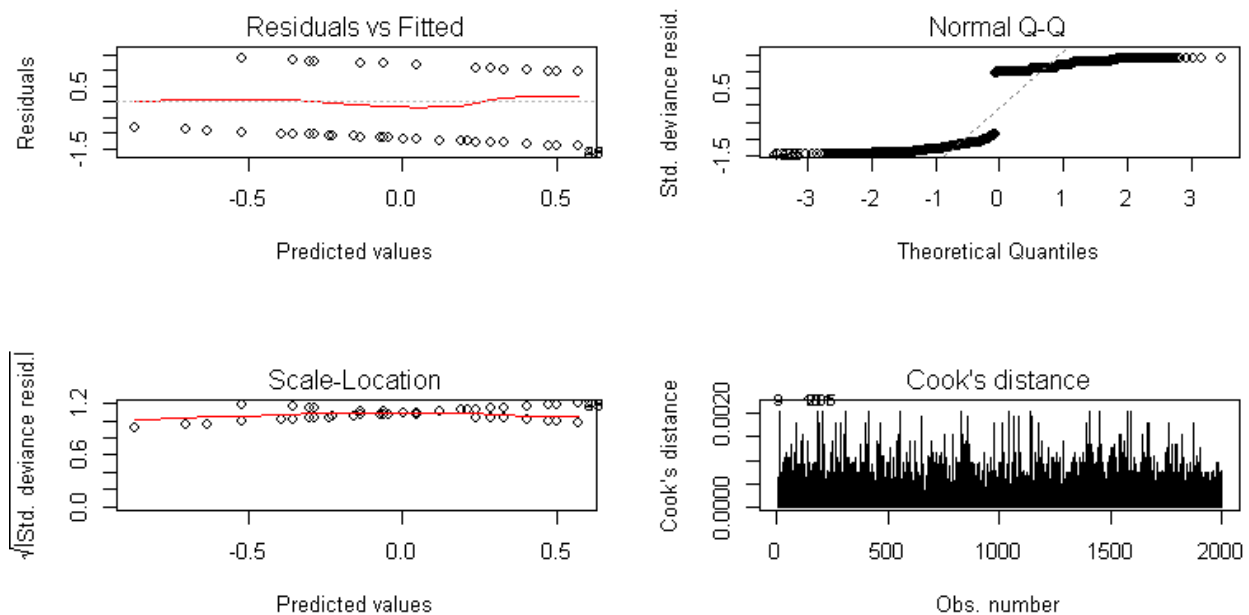
location92	0.070989	0.153832	0.461	0.644458
location101	0.624517	0.128318	4.867	1.13e-06 ***
location102	0.336888	0.154433	2.181	0.029150 *

Null deviance: 2772.6 on 2000 degrees of freedom

Residual deviance: 2700.9 on 1979 degrees of freedom, AIC: 2742.9, No. of Fisher Scoring it.: 4

Best Main Effect with Significant Predictors

Plots



Coefficients (in non log form.)

location10	location11	location12	location51	location52	location101	location102
0.5318079	0.4183030	0.4941323	1.6093885	1.7706004	1.8807271	1.4159356

Summary

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.428	-1.177	0.983	1.079	1.406

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
location10	-0.6315	0.1546	-4.085	4.40e-05 ***
location11	-0.8715	0.1609	-5.416	6.08e-08 ***
location12	-0.7050	0.2140	-3.295	0.000985 ***
location51	0.4759	0.1281	3.714	0.000204 ***
location52	0.5713	0.1524	3.749	0.000178 ***
location101	0.6317	0.1277	4.945	7.60e-07 ***
location102	0.3478	0.1536	2.265	0.023537 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2772.6 on 2000 degrees of freedom

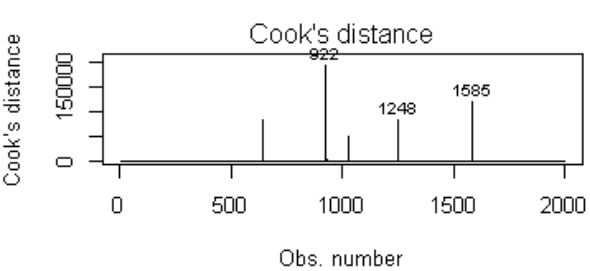
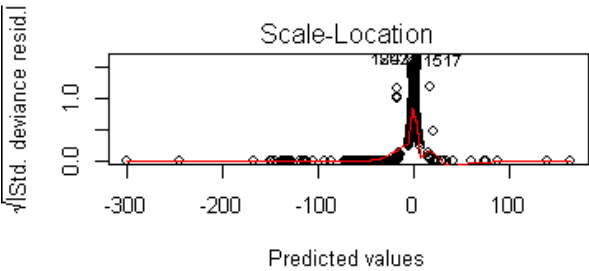
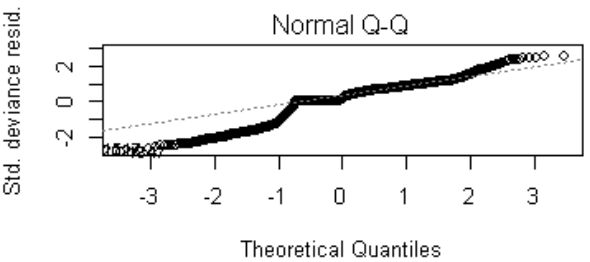
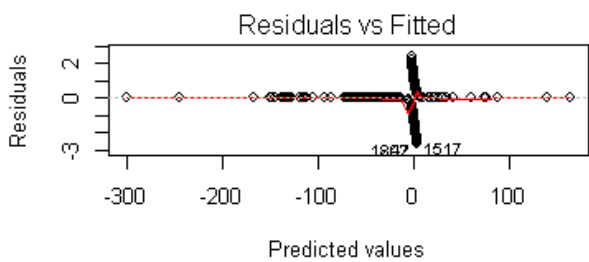
Residual deviance: 2708.6 on 1993 degrees of freedom

AIC: 2722.6

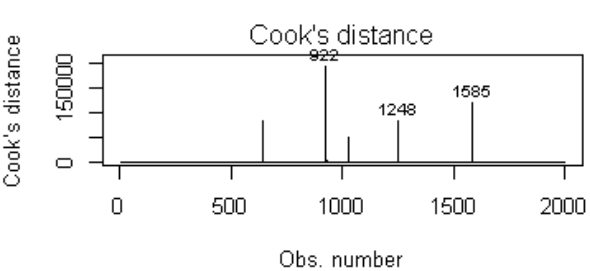
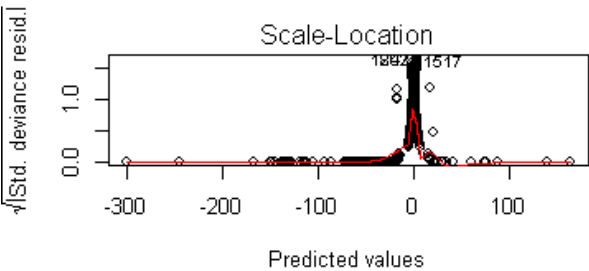
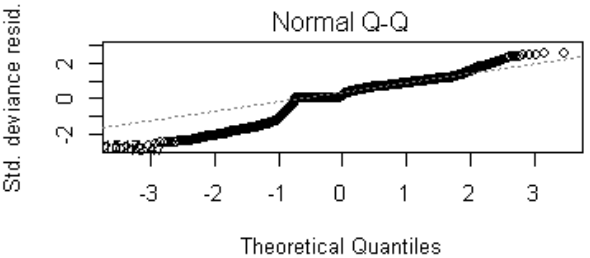
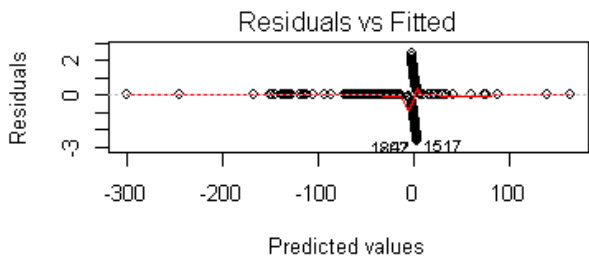
Number of Fisher Scoring iterations: 4

All Interactions Only Model

Plots



Full Model



R-Code

```
# setwd("C:/Users/Naseem Ashraf/Desktop 1/Fall 16/Bioinformatics/ML Assignment")
```

```
library("foreign")
raw <- read.arff("disease2a.arff")
head(raw)
summary(raw)
dim(raw)
```

```
plot(raw$location1, ylab = 'Frequency', xlab='Factors of Location1')
plot(raw$location2, ylab = "Frequency", xlab="Factors of Location2")
plot(raw$location3, ylab = "Frequency", xlab="Factors of Location3")
plot(raw$location4, ylab = "Frequency", xlab="Factors of Location4")
plot(raw$location5, ylab = "Frequency", xlab="Factors of Location5")
plot(raw$location6, ylab = "Frequency", xlab="Factors of Location6")
plot(raw$location7, ylab = "Frequency", xlab="Factors of Location7")
plot(raw$location8, ylab = "Frequency", xlab="Factors of Location8")
plot(raw$location9, ylab = "Frequency", xlab="Factors of Location9")
plot(raw$location10, ylab = "Frequency", xlab="Factors of Location10")
plot(raw$class, ylab = "Frequency", xlab="Factors of Classification")
```

```
## Modelling for all main systems
glm.fit=glm(class~., data=raw, family=binomial)
summary(glm.fit)
par(mfrow=c(2,2))
plot(glm.fit,1)
plot(glm.fit,2)
plot(glm.fit,3)
plot(glm.fit,4)
```

```
##Modelling Main Effects without the intercept
glm.fitno=glm(class~-1, data=raw, family=binomial)
summary(glm.fitno)
```

```
par(mfrow=c(2,2))
plot(glm.fitno,1)
plot(glm.fitno,2)
plot(glm.fitno,3)
plot(glm.fitno,4)
```

```
##Modelling Main Effects with statistically significant predictors without the intercept
glm.fit1=glm(class~location1+location5+location10-1, data=raw, family=binomial)
summary(glm.fit1)
exp(glm.fit1$coefficients)
```

```
par(mfrow=c(2,2))
plot(glm.fit1,1)
plot(glm.fit1,2)
plot(glm.fit1,3)
plot(glm.fit1,4)
```

```

# Coefficients:
# location10 location11 location12 location51 location52 location101 location102
# 0.5318079 0.4183030 0.4941323 1.6093885 1.7706004 1.8807271 1.4159356

## Modelling all two way interactions without the intercept
glm.fitinteractions=glm(class~ .*.-1, data=raw, family=binomial)
summary(glm.fitinteractions)

par(mfrow=c(2,2))
plot(glm.fitinteractions,1)
plot(glm.fitinteractions,2)
plot(glm.fitinteractions,3)
plot(glm.fitinteractions,4)

##Best of Both (trial)
glm.fitbest=glm(class~ location1+location5+location10+.*.-1, data=raw, family=binomial)
summary(glm.fitbest)

# Null deviance: 2772.6 on 2000 degrees of freedom
# Residual deviance: 1430.5 on 1799 degrees of freedom
# AIC: 1832.5
#
# Number of Fisher Scoring iterations: 20

##Full Model Analysis
glm.fitfull=glm(class~ .+.*.-1, data=raw, family=binomial)
summary(glm.fitfull)

par(mfrow=c(2,2))
plot(glm.fitfull,1)
plot(glm.fitfull,2)
plot(glm.fitfull,3)
plot(glm.fitfull,4)

# Deviance Residuals:
# Min 1Q Median 3Q Max
# -2.6654 0.0000 0.1625 0.6724 2.4397
#
# Coefficients:
# Estimate Std. Error z value Pr(>|z|)
# location21 -3.128e+00 1.478e+00 -2.116 0.034328 *
# location12:location21 -1.995e+00 6.499e-01 -3.070 0.002140 **
# location12:location32 -2.004e+00 8.465e-01 -2.367 0.017932 *
# location11:location42 -3.322e+00 1.388e+00 -2.392 0.016738 *
# location11:location72 -1.874e+00 7.297e-01 -2.569 0.010213 *
# location12:location81 -2.041e+00 7.407e-01 -2.756 0.005858 **
# location21:location61 -7.728e-01 3.170e-01 -2.438 0.014776 *
# location22:location61 -1.569e+00 6.955e-01 -2.256 0.024087 *
# location22:location62 -1.858e+00 9.244e-01 -2.010 0.044453 *
# location22:location101 3.716e+00 1.251e+00 2.971 0.002968 **
# location21:location102 3.290e+00 1.368e+00 2.405 0.016153 *
# location52:location72 4.656e+00 2.241e+00 2.078 0.037725 *

```

```
# location61:location71  6.896e-01  3.126e-01  2.206 0.027377 *
# location62:location72 -3.600e+00  1.071e+00 -3.363 0.000772 ***
# location62:location91 -1.444e+00  5.783e-01 -2.498 0.012502 *
# location71:location92  1.128e+00  5.735e-01  1.967 0.049187 *
# location81:location102 3.091e+00  1.462e+00  2.115 0.034463 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for binomial family taken to be 1)
#
# Null deviance: 2772.6  on 2000  degrees of freedom
# Residual deviance: 1430.5  on 1799  degrees of freedom
# AIC: 1832.5
#
# Number of Fisher Scoring iterations: 20
```