INFO304 Assignment 4 2022

Ethan Smith - 5652106

QUESTION ONE (30 MARKS) - Review of paper

PAPER: Regression tree construction by bootstrap: Model search for DRG-systems applied to Austrian health-data - Thomas Grubinger, Conrad Kobel and Karl-Peter Pfeiffer (2010)

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-9

In this paper Grubinger, Kobel and Pfeiffer aim to improve on existing methods to model length of stay using Austrian health data. Existing models to allocate hospital resources by a patient's length of stay based on conjunctive rules often do not align with how the hospitals operate, this means they often need to be adjusted through domain knowledge which decreases accuracy. The authors aim to improve on the CART models currently used with a different form of decision tree to increase the accuracy of LOS predictions and create more diverse trees to account for more scenarios. This also aims to create trees that require less readjustment to align with medical practices.

The Austrian health system uses conjunctive rules which allow Length of stay to be modelled as a decision tree. These rules mean that every condition must be met to be true. The Austrian-DRG data used is comprised of two main categories for codes. MEL codes are used for patients who are at the hospital for a pre-arranged procedure while general admission patients receive an HDG code. Both of these codes have subcodes called LDF referring to the specific procedure or diagnosis a patient receives. In this paper, the authors use eight different datasets. Four of these are MEL-LDF codes and the other four and HDG-LDF codes. The number of features is different for each dataset and no detailed list of variables is provided by the authors but they do mention the datasets include patient information such as diagnosis, procedures, age and sex.

Prior to modelling the authors do not mention any analysis of the features of the eight datasets nor do they mention any transformations or explicit feature selection. Due to trees being modelled feature selection will be provided by the model at each split. The only analysis provided prior to modelling is an explanation of how they believe the CART model can be improved. They mention that the CART model is a greedy algorithm, so their aim is to try and improve on this by modelling globally optimised trees. Possible approaches to this they considered were: evolutionary algorithms, bootstrapping and Monte-Carlo simulation methods. Bootstrapping was decided on as the method of choice as they aimed to maintain accuracy over a wide range of cases, bootstrapping allows this as many smaller datasets can be formed to build trees. This helps to optimise a final model as many potential scenarios have been modelled which can account for small changes in data having a significant effect on the resulting model. The bootstrapping method of choice was bumping, unlike ensemble methods bumping only results in a single tree which allows for a more interpretable model. By using B bootstrapped datasets to create B trees a wide range of candidate models are able to be assessed for their accuracy and suitability (B = 200). Each sample had a minimum of 30 observations.

To maintain accuracy in comparisons the authors only compared trees if they had the same size, trees were given a limit of 16 internal nodes and a max depth of 5. This resulted in a maximum of 17 possible groups patients could be filtered into based on the maximum of 5 tree rules. No formal explanation was given for these parameters but the range of values seems to imply that a wide range of complexities needed to be tested while trying to maintain consistency and interpretability in the models. As the authors are attempting to improve on a previous model their models use the mean squared error, for a direct comparison with the existing CART models.

In the first modelling stage, the authors wanted to show evidence that bumping can improve the CART models. To do this, trees were trained and tested on bootstrapped data using a single run of k-fold cross-validation with 10 folds on all datasets for all tree sizes between 0 and 20 internal nodes. No minimum split was set on the tree nodes to allow all branches to grow to the same depth for consistency. Figure 3 in the paper shows the reduction in MSE in comparison to the CART models from the best tree built by the bootstrap. Each tree appears to reduce MSE by 0.4%-1.0%.

After K-fold cross-validation a second evaluation was made for all tree sizes between 2 and 16 internal nodes by taking all trees at least as good as the CART model and selecting candidate models at each tree size. To be considered a candidate model a tree had to share the same split variables as the majority. From this group, the best-performing model was selected as a candidate model. This provided a good set of candidate models for length of stay that not only improved on the existing CART models but also provided more diversity from the varying bootstrap samples and tree complexities.

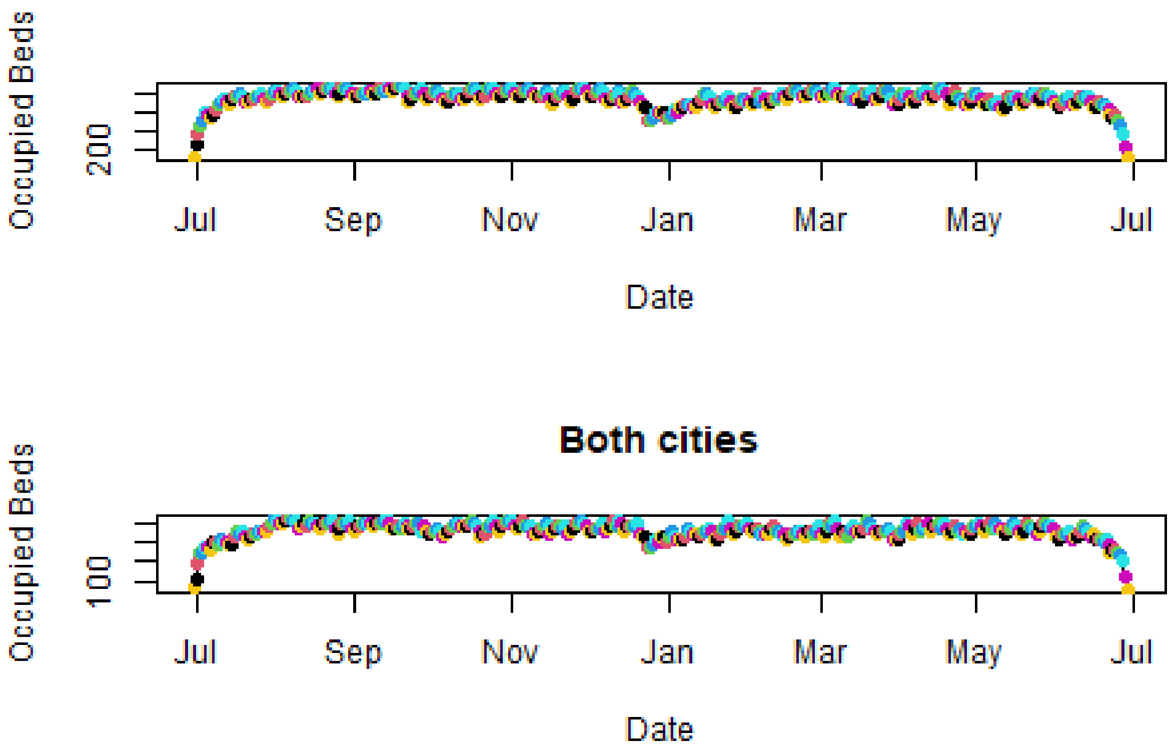
Along with creating a range of more diverse trees, bumping was also shown to improve accuracy against the CART trees on average between 1.06% to 4.90% on the eight datasets across a varying range of tree sizes. In some cases, Bumped trees were found that were no worse than the CART model while having a lower model complexity. As the authors have only provided the change in MSE between the CART and bumped models there is no indication on precisely how accurate either of these models are. Having more accurate and diverse models means that not only are the results more reliable than the CART models but can also account for more scenarios and give more options to medical professionals that may be more in line with hospital operation protocols.

QUESTION TWO (70 MARKS) - LOS modelling

1. Describe and visualise the datasets for both cities, focusing on the length of stay (LOS) and the relationship to the explanatories that are known at time of admission. Include a discussion that compares the cities in terms of health delivery, the population cohort, bed count over time, etc. (20 marks)

Based on the behaviour of bed count what other explanatory variable should be created? Ensure you create this variable and save the cityA and cityB data with the additional explanatory. What (if any) transformations are suggested by the data?

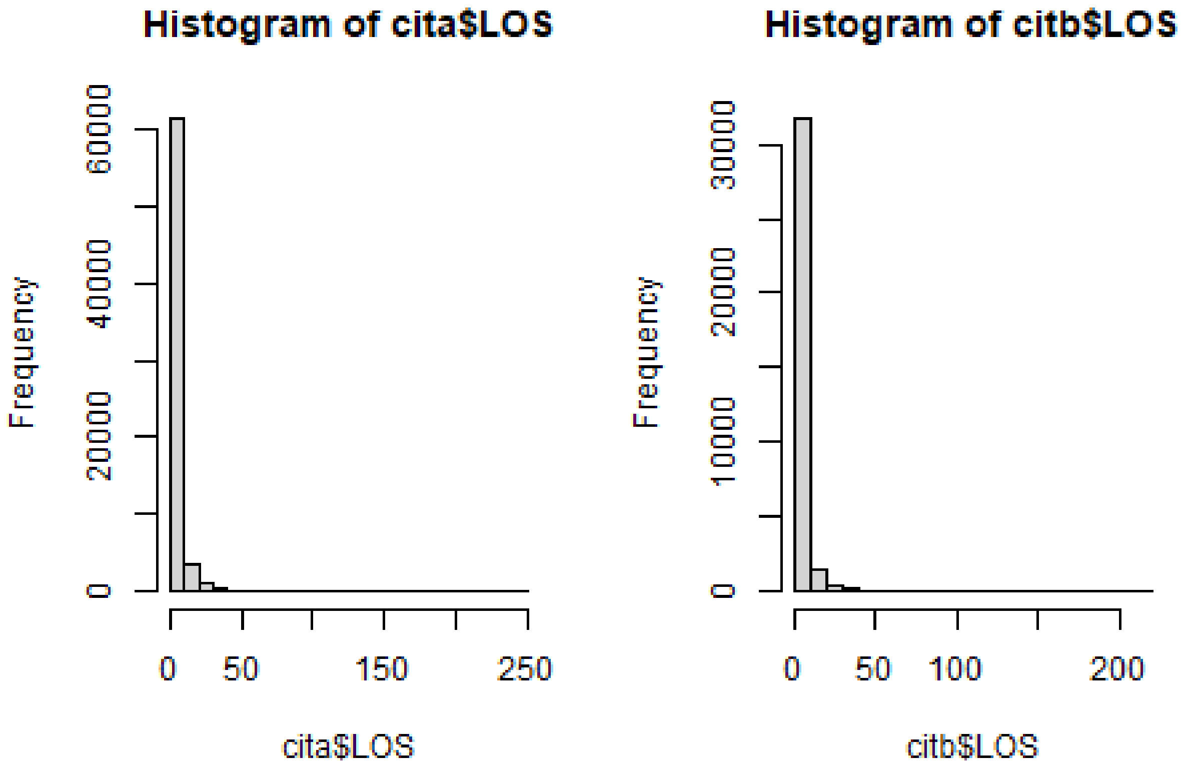
```r
cita <- read.csv("cityA.csv")
citb <- read.csv("cityB.csv")
#convert dates to date format
cita$EVSTDATE <- as.Date(cita$EVSTDATE, "%Y-%m-%d")
cita$EVENDATE <- as.Date(cita$EVENDATE, "%Y-%m-%d")
citb$EVSTDATE <- as.Date(citb$EVSTDATE, "%Y-%m-%d")
citb$EVENDATE <- as.Date(citb$EVENDATE, "%Y-%m-%d")

#create bedcounts
cita.bedcount <- bedcount(cita)
citb.bedcount <- bedcount(citb)
par(mfrow = c(2,1))
plot.bc(cita.bedcount)
plot.bc(citb.bedcount)
title("Both cities")
```



Looking at the overall bed counts for both cities they appear to have the same trends. The coloured dots refer to specific days of the week so it appears some day tend to have less patients in the hospital than others. For both cities it also appears that the number of beds occupied decreases around the Christmas / new years period. The tails on the end do not provide an accurate representation of the bed counts as they only exist due to cut off dates for admission within this one year sample. From the y axis scale it appears city A's hospital is twice as large as city B's hospital which appears to align with the size of the datasets as city A's dataset has twice as many patients.

NOTE: Due to the initial diagnosis variable having approximately 4000 factor levels, I will not be analyzing it here.

```
par(mfrow=c(1,2))
hist(cita$LOS)
hist(citb$LOS)
```



```
cita$LOS <- log(cita$LOS)
citb$LOS <- log(citb$LOS)
```

It is visible here that the response variable length of stay has an extreme right skew. I am going to apply a log transformation to this variable. Due to how extreme the skew is, this transformation will not normalize the length of stay variable but it will help to separate the patients more.

```
cita_clean <- cita
cita <- cita %>% arrange(cita$EVSTDATE)
cita_clean <- cita_clean[cita_clean$EVSTDATE >= "1984-08-01", ]
```

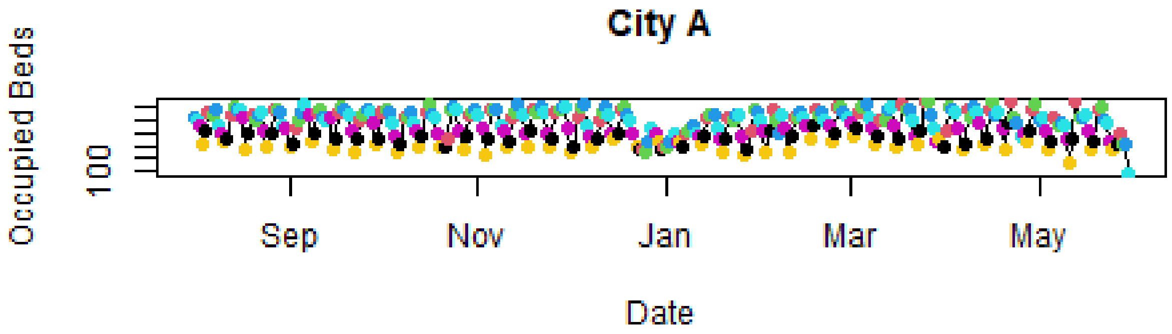
```r
cita_clean <- cita_clean[cita_clean$EVENDATE <= "1985-05-31", ]
cita_clean$GENDER[cita_clean$GENDER == "M"] <- 0
cita_clean$GENDER[cita_clean$GENDER == "F"] <- 1
cita_clean$GENDER <- as.numeric(cita_clean$GENDER)
cor(cita_clean[,c(1,2,6,8)])
```

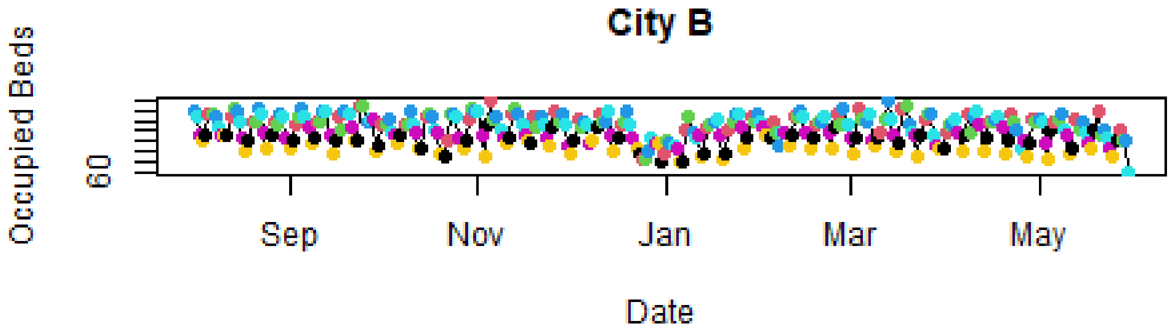
```
##             AGE_ADM       GENDER          LOS        Dep06
## AGE_ADM  1.00000000 -0.018749133  0.151103400 -0.084827644
## GENDER  -0.01874913  1.000000000 -0.003648690 -0.004451553
## LOS      0.15110340 -0.003648690  1.000000000 -0.009664902
## Dep06   -0.08482764 -0.004451553 -0.009664902  1.000000000
```

```r
citb_clean <- citb
citb <- citb %>% arrange(citb$EVSTDATE)
citb_clean <- citb_clean[citb_clean$EVSTDATE >= "1984-08-01", ]
citb_clean <- citb_clean[citb_clean$EVENDATE <= "1985-05-31", ]
citb_clean$GENDER[citb_clean$GENDER == "M"] <- 0
citb_clean$GENDER[citb_clean$GENDER == "F"] <- 1
citb_clean$GENDER <- as.numeric(citb_clean$GENDER)
cor(citb_clean[,c(1,2,6,8)])
```

```
##              AGE_ADM      GENDER         LOS        Dep06
## AGE_ADM  1.000000000 -0.02986826  0.13979457 -0.009268011
## GENDER  -0.029868260  1.00000000 -0.02911427 -0.028052189
## LOS      0.139794569 -0.02911427  1.00000000  0.013347501
## Dep06   -0.009268011 -0.02805219  0.01334750  1.000000000
```

```r
par(mfrow = c(2,1))
cita.bedcount <- bedcount(cita_clean)
citb.bedcount <- bedcount(citb_clean)
plot.bc(cita.bedcount)
title("City A")
plot.bc(citb.bedcount)
title("City B")
```

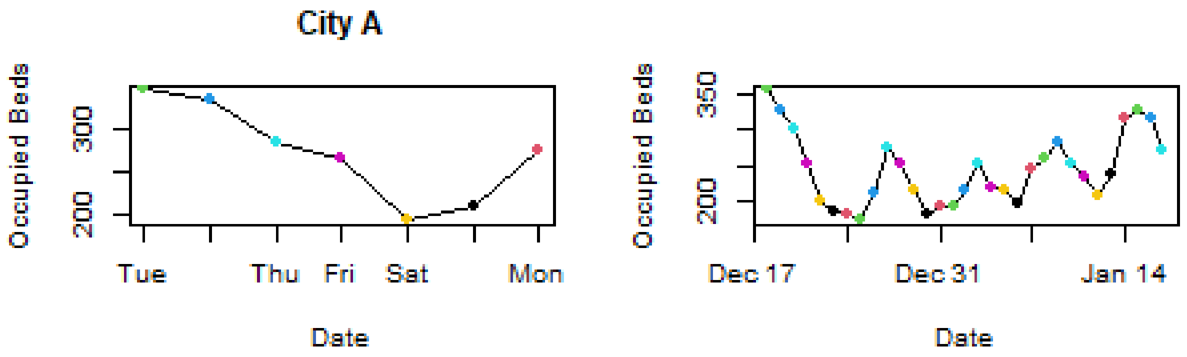

City A

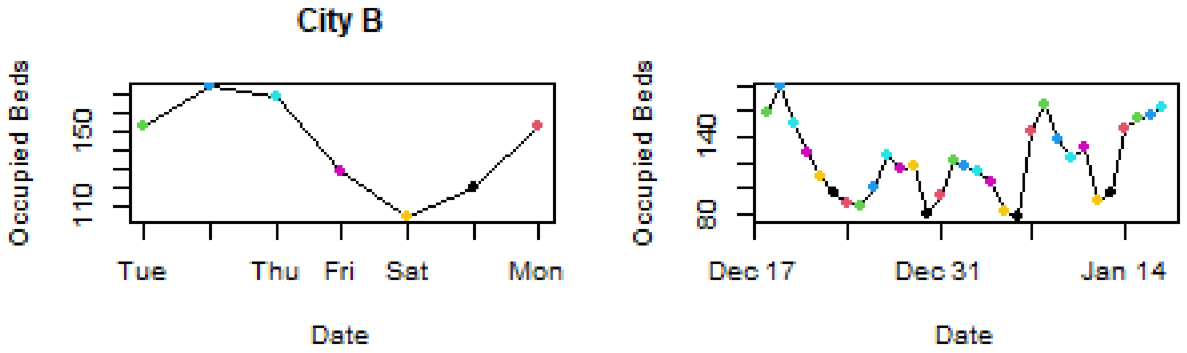

City B

Here I have performed some modifications on the data to try lessen the effect of the tails. The tails are still existent but by tidying the ends I have made the change in bed count less extreme as the data cuts off. I have also displayed some correlations between some of the variables. From here we can see a severe lack of relationship between the variables. This makes modelling length of stay much more difficult as there is a severe lack of variation in length of stay that can be provided by these predictors. The information we have is essentially noise.

The new bed count plots I have provided using the tided date ranges and the log transformation on LOS makes the affect of bed count by day much more visible and the drop in the holiday period is also much more noticeable. These plots shows that both of these are clear systematic trends based on hospital operating methods so they should be modelled.

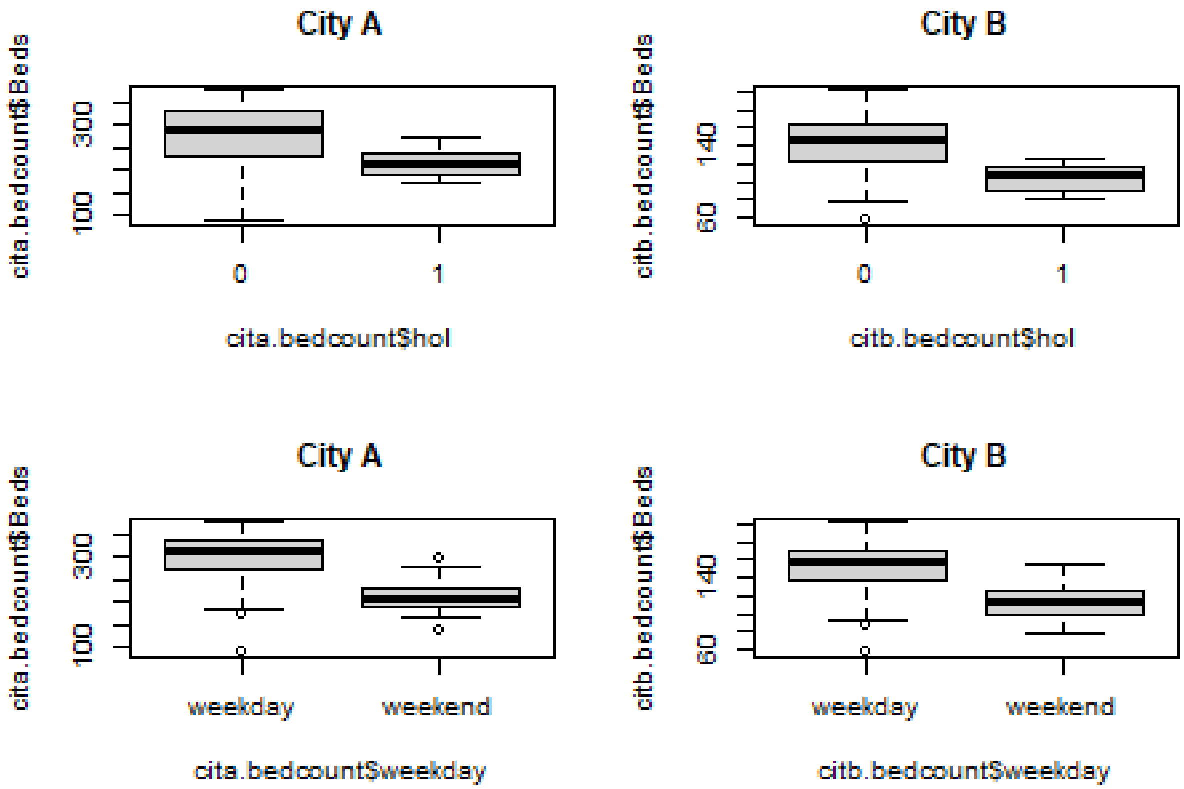
```r
par(mfrow = c(2,2))
plot.bc(cita.bedcount, r=28:34)
title("City A")
plot.bc(cita.bedcount, r=140:170)
plot.bc(citb.bedcount, r=28:34)
title("City B")
plot.bc(citb.bedcount, r=140:170)
```



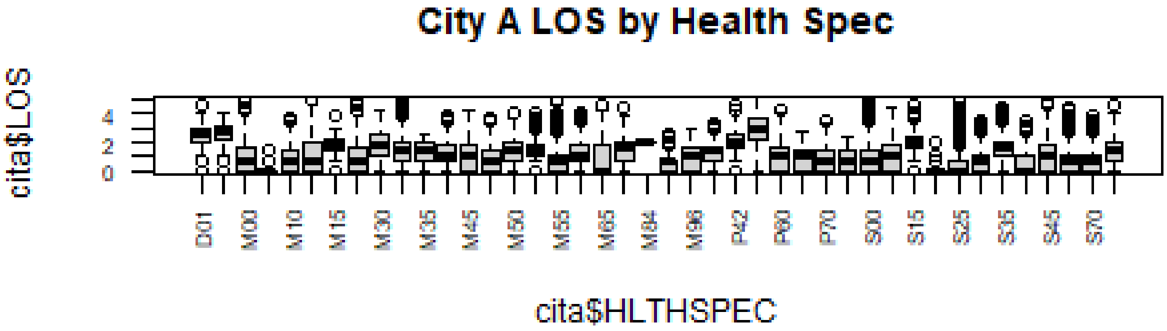


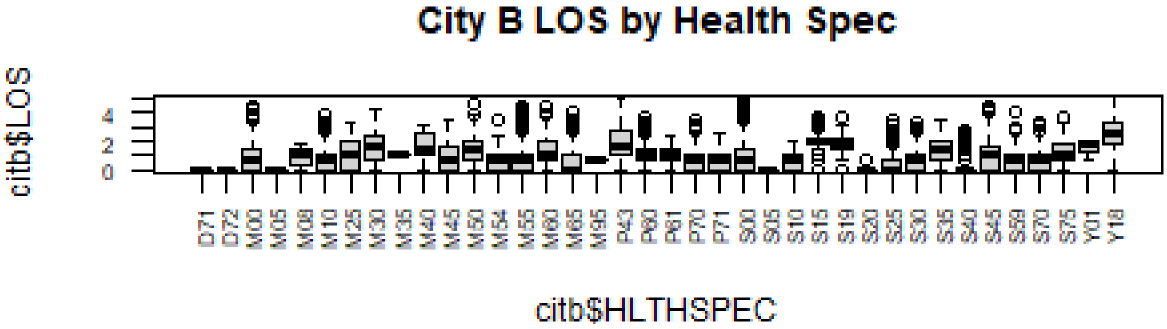
```
hol_a <- which(cita_clean$EVSTDATE >= "1984-12-21" & cita_clean$EVSTDATE <= "1985-01-05")
hol_b <- which(citb_clean$EVSTDATE >= "1984-12-21" & citb_clean$EVSTDATE <= "1985-01-05")

cita_clean$hol <- 0
citb_clean$hol <- 0
cita_clean$hol[hol_a] <- 1
citb_clean$hol[hol_b] <- 1

hol_a <- which(cita.bedcount$Day >= "1984-12-21" & cita.bedcount$Day <= "1985-01-05")
hol_b <- which(citb.bedcount$Day >= "1984-12-21" & citb.bedcount$Day <= "1985-01-05")

cita.bedcount$hol <- 0
citb.bedcount$hol <- 0
cita.bedcount$hol[hol_a] <- 1
citb.bedcount$hol[hol_b] <- 1


cita.bedcount$weekday <- "weekday"
for (i in 1:nrow(cita.bedcount)) {
  if (i %% 7 == 4 || i %% 7 == 5) {
    cita.bedcount[i,]$weekday <- "weekend"
  }
}

cita_clean$weekday <- "weekday"
for (i in 1:nrow(cita_clean)) {
```

```r
  val <- cita_clean$EVSTDATE[i]
  idx <- which(cita.bedcount$Day == val)
  new <- cita.bedcount$weekday[idx]
  cita_clean$weekday[i] <- new
}

citb.bedcount$weekday <- "weekday"
for (i in 1:nrow(citb.bedcount)) {
  if (i %% 7 == 4 || i %% 7 == 5) {
    citb.bedcount[i,]$weekday <- "weekend"
  }
}

citb_clean$weekday <- "weekday"
for (i in 1:nrow(citb_clean)) {
  val <- citb_clean$EVSTDATE[i]
  idx <- which(citb.bedcount$Day == val)
  new <- citb.bedcount$weekday[idx]
  citb_clean$weekday[i] <- new
}

par(mfrow = c(2,2))
boxplot(cita.bedcount$Beds ~ cita.bedcount$hol)
title("City A")
boxplot(citb.bedcount$Beds ~ citb.bedcount$hol)
title("City B")


boxplot(cita.bedcount$Beds ~ cita.bedcount$weekday)
title("City A")
boxplot(citb.bedcount$Beds ~ citb.bedcount$weekday)
title("City B")
```

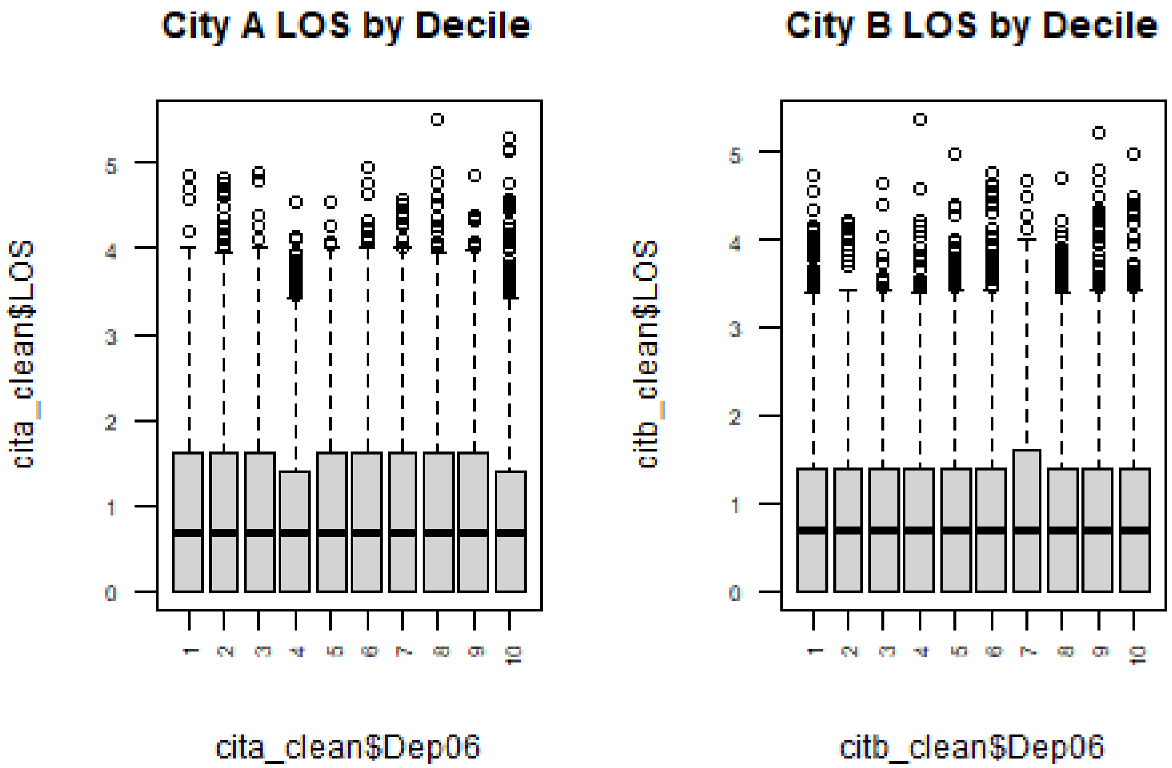


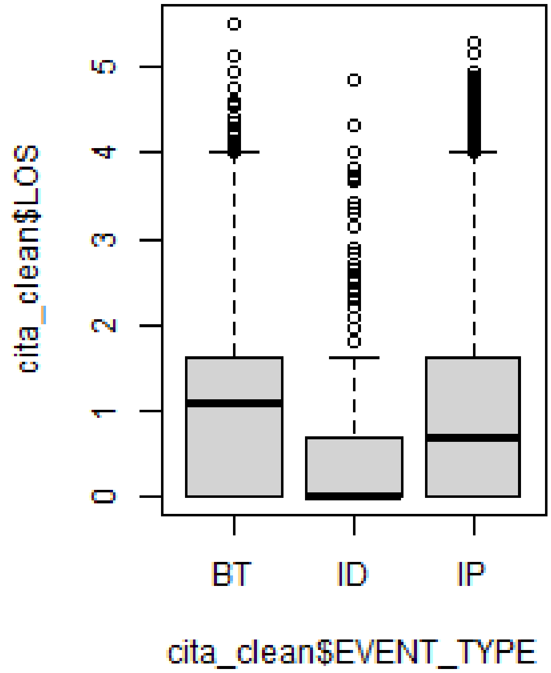
Here I have shown a closer view at the two visible trends. For the trends by day it can be seen that the hospital appears to try and reduce the number of patients hospitalized over the weekend before bringing their intake back up at the start of the working week. For the holiday period we can see that intake leading up to christmas consistently reduces and normal capacity does not return until early January. After implementing these as variables for both datasets and plotting them it appears that there is significant evidence that supports the reduction in bed count shown in previous figures for both the weekend and holiday period. The amount of variance in these periods also appears to significantly decrease which may make these time periods easier to model.

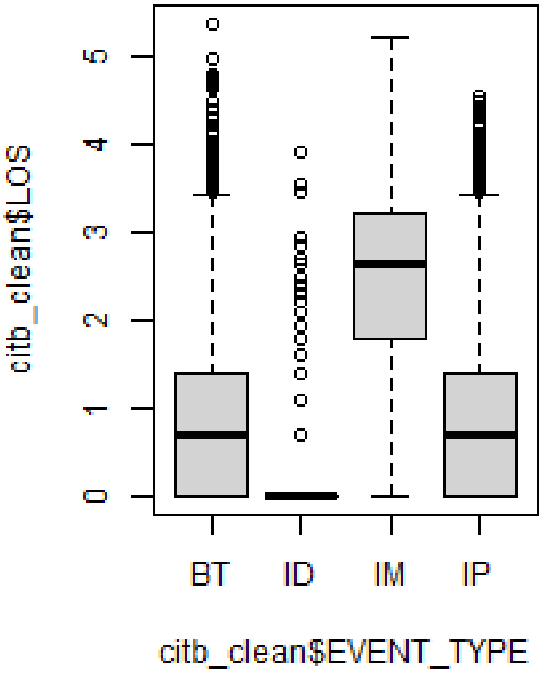




Comparing the two cities LOS by Health specification it appears that City A appears to deal with a much wider range of cases which is to be expected as the hospital is larger. Overall Length of stay appears to be similar between the two cities for matching cases. Both hospitals also appear to have unique cases they have dealt with (eg city A has no Y code patients but more types of cases overall).


City A LOS by Decile
City B LOS by Decile
cita_clean$LOS
citb_clean$LOS
cita_clean$Dep06
citb_clean$Dep06
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10





As a final comparison of the hospital operations for both hospitals it appears that patients who are only intended to stay for a day do stay for a significantly shorter time. It also appears that city B's hospital has a psychiatric ward (IM) which city A does not have. This once again highlights that the different hospitals do have different procedures and cases that need to be accounted for in a model which can add extra complexities.

Finally it also appears that a patients decile has no affect on Length of stay and just adds noise.

2. Build separate linear models (with additional explanatories and transformations) to predict LOS for cityA and cityB using all of the data – DO NOT USE the initial diagnosis class (diag01).

Discuss why LOS is difficult to model. How might you go about making the problem easier? Explore and discuss several options for improving LOS modelling and present some initial approaches. This can all be done just using linear regression, logistic regression or simple decision trees (rpart) depending on how you modify the data. If you want to assess accuracy use a 90% training-10% test split over 50 replicates using RRSE as the error measurement. (25 marks)

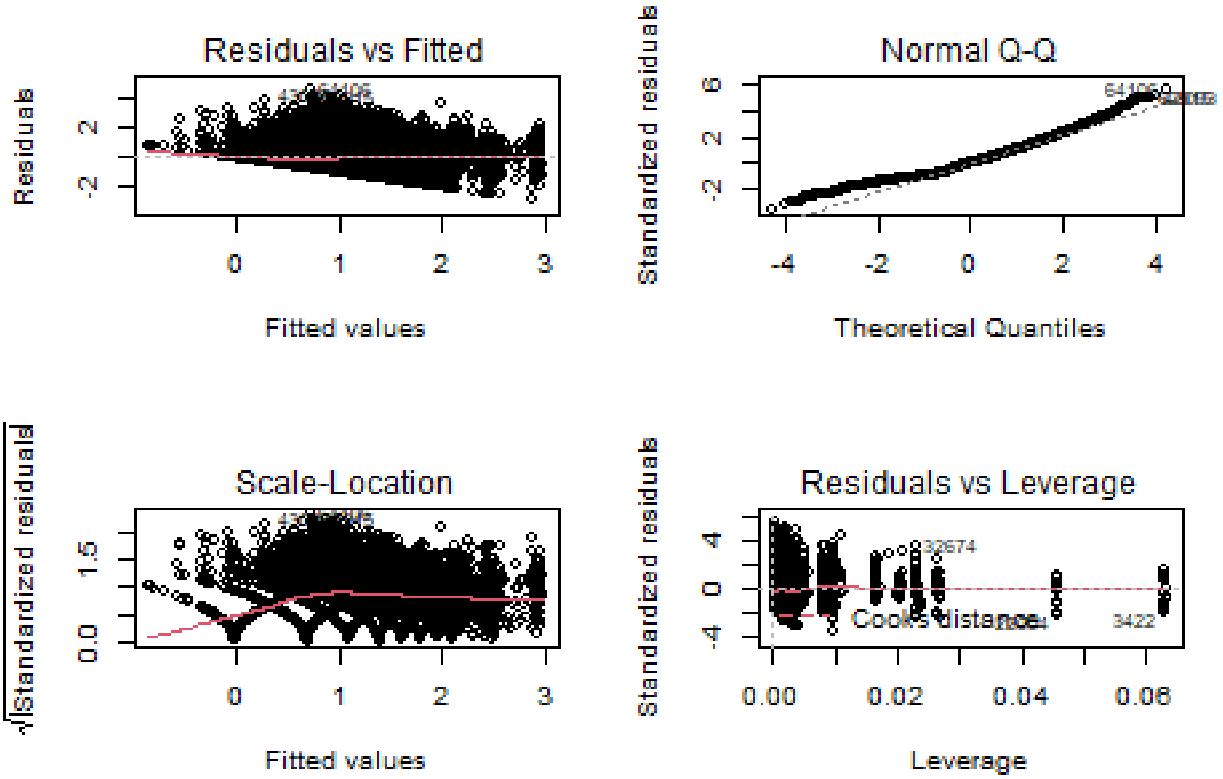
```
c <- c(2,3,7,8,9,10,11)
cita_clean[,c] <- lapply(cita_clean[c], factor)
citb_clean[,c] <- lapply(citb_clean[c], factor)

calm <- lm(LOS ~ AGE_ADM + GENDER + EVENT_TYPE + HLTHSPEC + Dep06 + hol + weekday, data = cita_clean)
summary(calm)
```

```
##
## Call:
## lm(formula = LOS ~ AGE_ADM + GENDER + EVENT_TYPE + HLTHSPEC +
##     Dep06 + hol + weekday, data = cita_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8724 -0.6821 -0.0402  0.5025  4.5524
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.9394869  0.0901106  32.621  < 2e-16 ***
## AGE_ADM        0.0080581  0.0001832  43.994  < 2e-16 ***
## GENDER1        0.0273843  0.0074793   3.661 0.000251 ***
## EVENT_TYPEID  -1.7580094  0.0795471 -22.100  < 2e-16 ***
## EVENT_TYPEIP  -1.1995047  0.0757823 -15.828  < 2e-16 ***
## HLTHSPECD40    0.4965485  0.1769415   2.806 0.005013 **
## HLTHSPECM00   -1.4476467  0.0458798 -31.553  < 2e-16 ***
## HLTHSPECM05   -2.0221674  0.0481201 -42.023  < 2e-16 ***
## HLTHSPECM10   -1.3760683  0.0483937 -28.435  < 2e-16 ***
## HLTHSPECM14   -0.7408242  0.0586522 -12.631  < 2e-16 ***
## HLTHSPECM15   -0.5279219  0.1378183  -3.831 0.000128 ***
## HLTHSPECM25   -1.0154747  0.0561467 -18.086  < 2e-16 ***
## HLTHSPECM30   -0.4383982  0.0560139  -7.827 5.10e-15 ***
## HLTHSPECM34   -0.3384709  0.0627894  -5.391 7.05e-08 ***
## HLTHSPECM35   -0.7299965  0.2058779  -3.546 0.000392 ***
## HLTHSPECM40   -0.8162317  0.0838679  -9.732  < 2e-16 ***
## HLTHSPECM45   -1.0930276  0.0531504 -20.565  < 2e-16 ***
## HLTHSPECM49   -0.9292236  0.0724846 -12.820  < 2e-16 ***
## HLTHSPECM50   -0.9360781  0.0500040 -18.720  < 2e-16 ***
## HLTHSPECM54   -0.4148637  0.0660389  -6.282 3.36e-10 ***
## HLTHSPECM55   -1.0674523  0.0508315 -21.000  < 2e-16 ***
## HLTHSPECM60   -1.0448593  0.0525501 -19.883  < 2e-16 ***
## HLTHSPECM65   -1.3285291  0.0488028 -27.222  < 2e-16 ***
## HLTHSPECM70   -0.4365294  0.1294324  -3.373 0.000745 ***
## HLTHSPECM84    0.0850565  0.8035878   0.106 0.915705
## HLTHSPECM95   -1.3342756  0.1125115 -11.859  < 2e-16 ***
## HLTHSPECM96    -0.8013668  0.0888698  -9.017  < 2e-16 ***
## HLTHSPECP41    -1.7813049  0.0911308 -19.547  < 2e-16 ***
## HLTHSPECP42    -0.8555010  0.0954119  -8.966  < 2e-16 ***
## HLTHSPECP43    -0.0656660  0.1177515  -0.558 0.577075
## HLTHSPECP60    -1.0099660  0.0481949 -20.956  < 2e-16 ***
## HLTHSPECP61    -2.0706693  0.0917049 -22.580  < 2e-16 ***
## HLTHSPECP70    -1.3271147  0.0489232 -27.126  < 2e-16 ***
## HLTHSPECP71    -2.2835848  0.0915066 -24.955  < 2e-16 ***
## HLTHSPECS00    -1.2398701  0.0466321 -26.588  < 2e-16 ***
## HLTHSPECS10    -1.1454888  0.0637259 -17.975  < 2e-16 ***
## HLTHSPECS15    -0.3850526  0.0531428  -7.246 4.36e-13 ***
## HLTHSPECS20    -1.8299915  0.1222955 -14.964  < 2e-16 ***
## HLTHSPECS25    -1.4945422  0.0490925 -30.443  < 2e-16 ***
## HLTHSPECS30    -1.4944628  0.0492033 -30.373  < 2e-16 ***
## HLTHSPECS35    -0.5966096  0.0522441 -11.420  < 2e-16 ***
## HLTHSPECS40    -1.4913997  0.0554318 -26.905  < 2e-16 ***
## HLTHSPECS45    -0.9454602  0.0471830 -20.038  < 2e-16 ***
## HLTHSPECS59    -1.1672173  0.0514058 -22.706  < 2e-16 ***
## HLTHSPECS70    -1.5250059  0.0488552 -31.215  < 2e-16 ***
## HLTHSPECS75    -0.8652815  0.0543360 -15.925  < 2e-16 ***
## Dep062          0.0135810  0.0171153   0.794 0.427489
## Dep063          0.0096819  0.0171458   0.565 0.572295
## Dep064         -0.0145665  0.0172742  -0.843 0.399089
## Dep065          0.0376391  0.0175159   2.149 0.031651 *
## Dep066         -0.0014283  0.0162972  -0.088 0.930162
## Dep067          0.0171255  0.0173152   0.989 0.322643
## Dep068          0.0377431  0.0165603   2.279 0.022663 *
## Dep069          0.0637463  0.0173238   3.680 0.000234 ***
## Dep0610         0.0241586  0.0160899   1.501 0.133238
## hol1            0.0033138  0.0174566   0.190 0.849440
## weekdayweekend  0.0095910  0.0086116   1.114 0.265401
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8021 on 54947 degrees of freedom
## Multiple R-squared:  0.2176, Adjusted R-squared:  0.2168
## F-statistic: 272.8 on 56 and 54947 DF,  p-value: < 2.2e-16
```

```
exp(sqrt(mean((cita_clean$LOS - predict.lm(calm, cita_clean)) ^ 2)))
```

```
## [1] 2.229351
```

```
par(mfrow=c(2,2))
plot(calm)
```



```
calm <- test.lm(citb_clean, LOS ~ AGE_ADM + GENDER + EVENT_TYPE + HLTHSPEC + hol + Dep06 + weekday)
calm
```

```
## [1] 0.8563222
```

```
cblm <- test.lm(citb_clean, LOS ~ AGE_ADM + GENDER + EVENT_TYPE + HLTHSPEC + hol + Dep06 + weekday)
cblm
```

```
## [1] 0.8638736
```

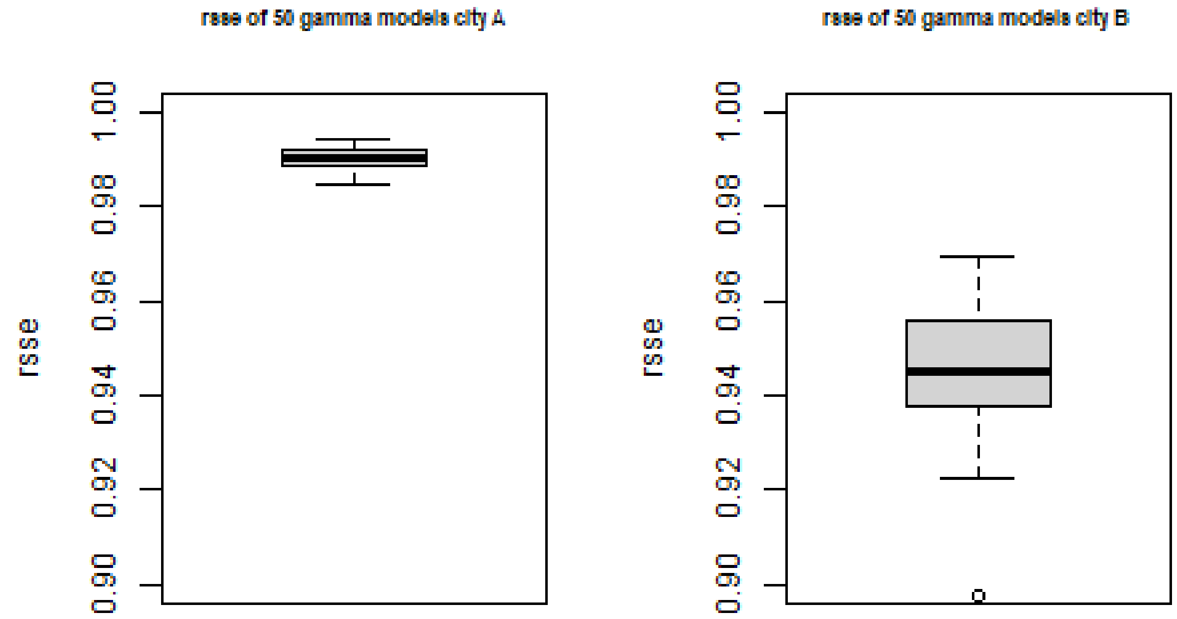
Length of stay is an incredibly hard problem to model due to a number of factors. The distribution of Length of stay itself is highly right skewed and non linear, this makes modelling much harder as many forms of models are affected by this as linearity is assumed (the linear model is one of these). Most predictors used to try and model length of stay struggle to explain any variation in the data. This is due to the distribution of length of stay but also due to how complex the data and hospital system itself is. Each hospital has it's own cases and procedures they are able to perform while each patient requires different forms of treatment from the many treatments available. This means that the overall data itself is incredibly diverse and very few observations in the dataset will be similar, this can make it very difficult to find a way to model length of stay without creating a model to a very specific group of patients.

By looking at the figures only the linear models produced to not appear to be awful, from a few tests I did the root relative squared error fell between 0.83-0.88 for both cities which shows the models predictions were better than just estimating the mean, for a problem of this complexity that seems fair. When breaking the linear models down further some major issues arise. I have displayed figures city A to show this. Although a large portion of the parameters appeared to explain something about length of stay the adjusted r-squared is still only 0.2168 which means the majority of variance in length of stay is still unexplained. The diagnostic plots also show major departures from the linear model assumptions regarding linearity and homoscedacity. On average it also appears that the model for city A is on average predicting length of stay wrong by two days.

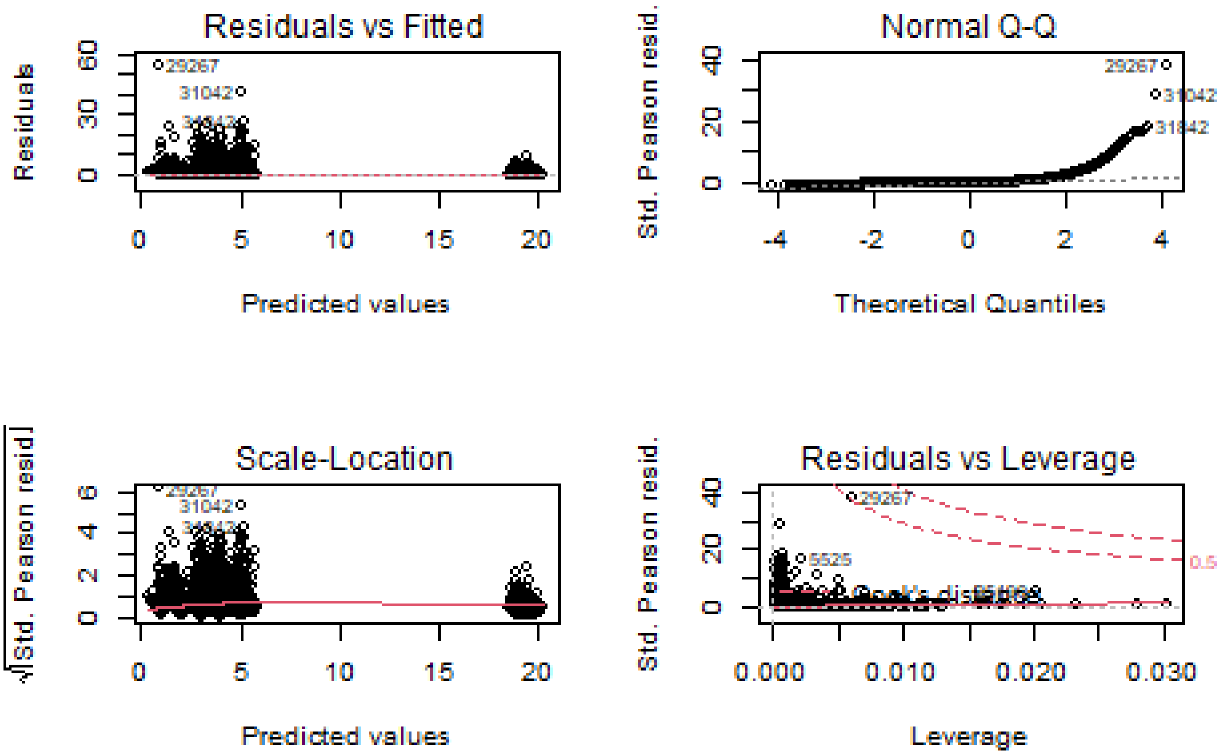
In order to model length of stay the data either needs to be separated to focus on building a model for a specific aspect or treatment or a more robust / suitable model needs to be used.

The first model I am interested in trialing is a generalized linear model with a gamma response distribution. I am interested in trying this as length of stay is right-skewed, continuous and non-negative which are the assumptions for the gamma family.

```
cita_g <- cita_clean
cita_g$LOS <- exp(cita_g$LOS)
citb_g <- citb_clean
citb_g$LOS <- exp(citb_g$LOS)
rssea <- c()
rsseb <- c()

for (i in 1:50) {
train.row <- sample(nrow(cita_g),0.9*nrow(cita_g),replace=FALSE)
train.data <- cita_g[train.row,]  # Create training and testing data
test.data <- cita_g[-train.row,]

cagam <- glm(LOS ~ AGE_ADM + GENDER + EVENT_TYPE + Dep06 + hol + weekday,
             data = train.data, family = Gamma(link = "identity"))

rssea[i] <- sqrt(sum((test.data$LOS - predict(cagam, test.data))^2,na.rm=T) / sum((test.data$LOS - mean

train.row <- sample(nrow(citb_g),0.9*nrow(citb_g),replace=FALSE)
train.data <- citb_g[train.row,]  # Create training and testing data
test.data <- citb_g[-train.row,]

cagam <- glm(LOS ~ AGE_ADM + GENDER + EVENT_TYPE + Dep06 + hol + weekday,
             data = train.data, family = Gamma(link = "identity"))

rsseb[i] <- sqrt(sum((test.data$LOS - predict(cagam, test.data))^2,na.rm=T) / sum((test.data$LOS - mean
}

par(mfrow = c(1,2))
boxplot(rssea, main = "rsse of 50 gamma models city A", ylab = "rsse",
        cex.main = 0.7, ylim = c(0.9,1))
boxplot(rsseb, main = "rsse of 50 gamma models city B", ylab = "rsse",
        cex.main = 0.7, ylim = c(0.9,1))
```



```r
par(mfrow=c(2,2))
plot(cagam)
```

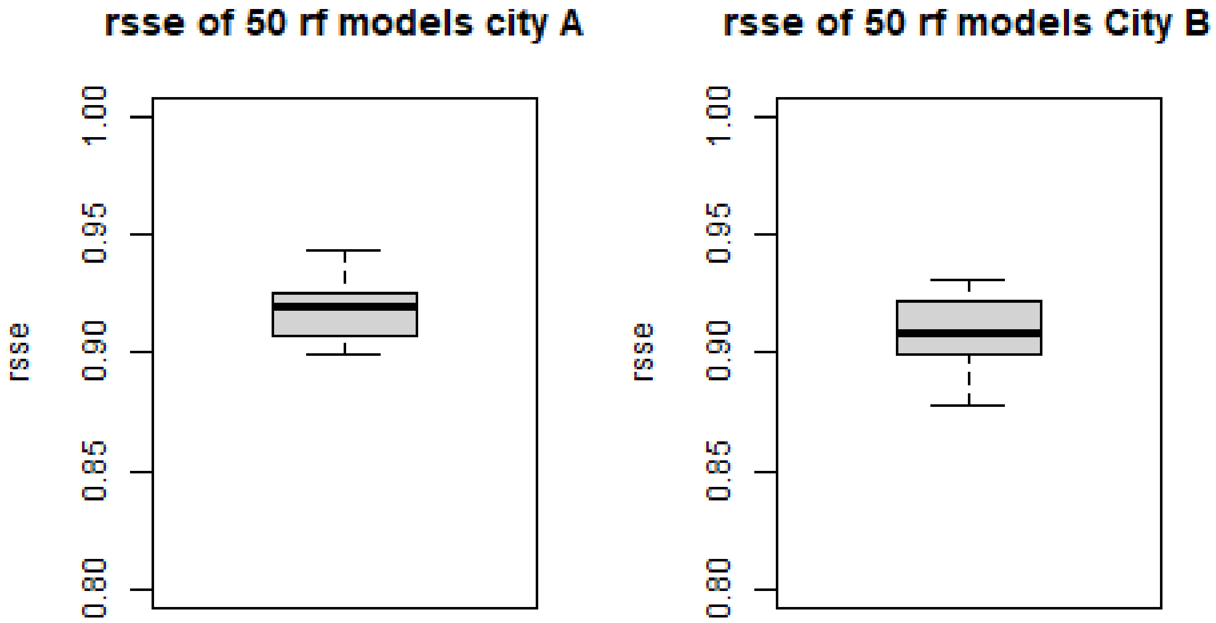


Surprisingly to me trying to account for the response distribution in the model has resulted in a worse performing model which has also not fixed the failures in model assumptions (rrse increase of 0.05-0.15). This makes me believe that the issues with modelling length of stay may be more associated with the quality of predictors than the highly skewed behavior of the response. This is also the same for a model built on all data with health codes included.

The second model I am going to trial is fitting a random forest to the data, I want to fit this model for two reasons. The first being that tree methods are incredibly stable and are not affected by non-linearity so the skew of length of stay will not have an adverse effect on modelling. The second reason for a random forest is that because we have many different aspects of the data (many different treatments ect) I believe an ensemble method may help to obtain more accurate predictions

```
library(randomForest)
rfa <- c()
rfb <- c()

for (i in 1:10) {
train.row <- sample(nrow(cita_g),0.9*nrow(cita_g),replace=FALSE)
train.data <- cita_g[train.row,]  # Create training and testing data
test.data <- cita_g[-train.row,]

mdl <- randomForest(x = train.data[-c(4,5,6,9)], y = train.data$LOS, ntree = 50)

rfa[i] <- sqrt(sum((test.data$LOS - predict(mdl, test.data))^2,na.rm=T) /
                 sum((test.data$LOS - mean(test.data$LOS,na.rm=T))^2))

train.row <- sample(nrow(citb_g),0.9*nrow(citb_g),replace=FALSE)
```

```r
train.data <- citb_g[train.row,]  # Create training and testing data
test.data <- citb_g[-train.row,]

mdl <- randomForest(x = train.data[-c(4,5,6,9)], y = train.data$LOS,
                    ntree= 50)

rfb[i] <- sqrt(sum((test.data$LOS - predict(mdl, test.data))^2,na.rm=T) /
                 sum((test.data$LOS - mean(test.data$LOS,na.rm=T))^2))
}
par(mfrow = c(1,2))
boxplot(rfa, main = "rsse of 50 rf models city A", ylab = "rsse",
        ylim = c(0.8,1))
boxplot(rfb, main = "rsse of 50 rf models City B", ylab = "rsse",
        ylim = c(0.8,1))
```



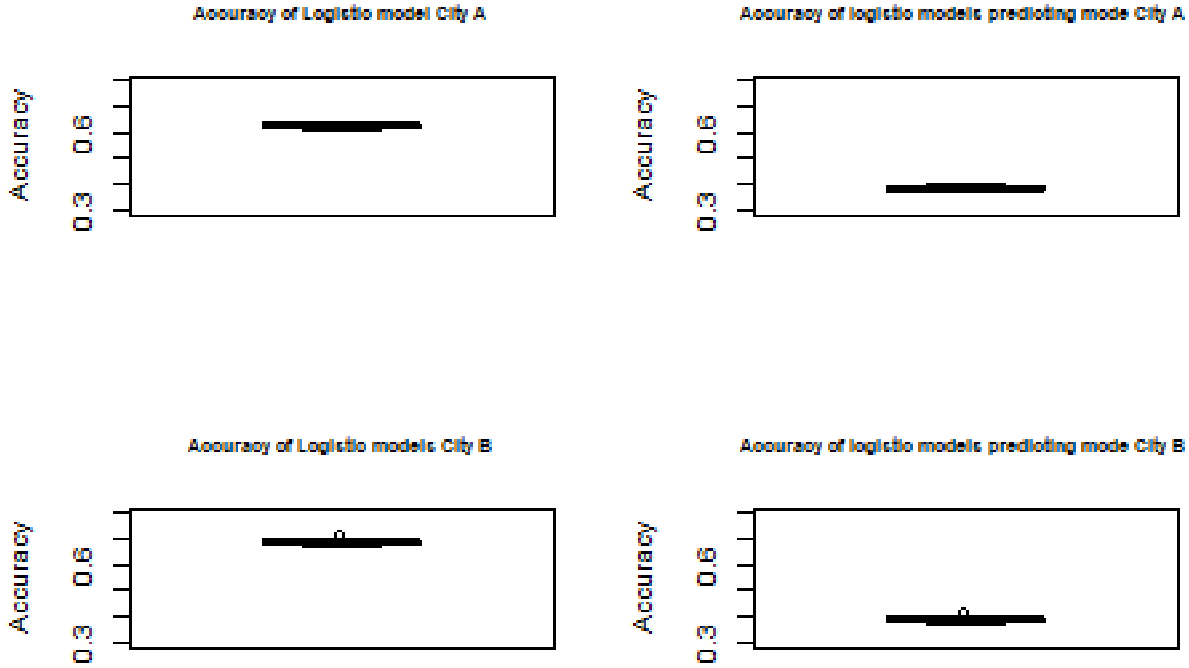
The use of a more stable random forest model does not seem to improve on the linear model. This suggests that the issues with modelling length of stay are due to the quality of the predictor variables. I also believe the issues with modelling length of stay are also due to how the complexity of the problem itself so methods to try and solve particular aspects of the problem should be using local models should be used over a global length of stay model in order to gain any true insight to the problem.

NOTE: Due to the complexity of the random forest model I have run this code as 10 runs for each city with 50 trees per run, this is a possible reason why the variance is lower in rrse and also a potential reason for a lack of change in rrse. I did trial a single model which used 500 trees which did appear to slightly reduce rrse but I have not considered this a conclusive result for consistency.

To do this I am going to fit a logistic regression model which aims to predict if a patient is going to stay in hospital for longer than a day. This seems like a fair problem to model as patients who only stayed in hospital for a single day make up over half the observations for both datasets.

To assess this problem I will compare the accuracy of the logistic regression model to a set of predictions where every observation is predicted as the mode of the dataset (patients only stay one day).

```
rssea <- c()
rssea_base <- c()
rsseb <- c()
rsseb_base <- c()

cita_clean$multiple <- 0
citb_clean$multiple <- 0

for (i in 1:nrow(cita_clean)) {
  if (cita_clean$LOS[i] > 0) {
    cita_clean$multiple[i] <- 1
  }
}

for (i in 1:nrow(citb_clean)) {
  if (citb_clean$LOS[i] > 0) {
    citb_clean$multiple[i] <- 1
  }
}

cita_clean$multiple <- as.factor(cita_clean$multiple)
citb_clean$multiple <- as.factor(citb_clean$multiple)

for (i in 1:50) {
train.row <- sample(nrow(cita_clean),0.9*nrow(cita_clean),replace=FALSE)
train.data <- cita_clean[train.row,]   # Create training and testing data
test.data <- cita_clean[-train.row,]



log.mod <- glm(multiple ~ AGE_ADM + GENDER + EVENT_TYPE + Dep06 + hol +
                 weekday, data=train.data, family="binomial")
train.pred <- predict(log.mod,newdata=train.data,type="response")

act <- as.numeric(train.data$multiple)-1
p.order <- order(train.pred)
probs <- train.pred[p.order]
act.order <- act[p.order]
act.cs <- cumsum(act.order)
x <- 1:length(act.cs)
threshold <- x - 2*act.cs
t <- as.numeric(probs[(which(threshold==max(threshold)))[1]])

test.pred <- predict(log.mod,newdata=test.data,type="response")
pred.vals <- as.factor(ifelse(test.pred >=t,1,0))
conf.mat <- caret::confusionMatrix(data = pred.vals,
                                     reference=test.data$multiple)
test.pred2 <- predict(log.mod,newdata=test.data,type="response")
pred.vals2 <- as.factor(ifelse(test.pred2 >=1,1,0))
conf.mat2 <- caret::confusionMatrix(data = pred.vals2,
                                    reference=test.data$multiple)
rssea[i] <- (conf.mat$table[1]+conf.mat$table[4])/nrow(test.data)
rssea_base[i] <- (conf.mat2$table[1]+conf.mat2$table[4])/nrow(test.data)

train.row <- sample(nrow(citb_clean),0.9*nrow(citb_clean),replace=FALSE)
train.data <- citb_clean[train.row,]  # Create training and testing data
test.data <- citb_clean[-train.row,]

log.mod <- glm(multiple ~ AGE_ADM + GENDER + EVENT_TYPE + Dep06 + hol +
                 weekday, data=train.data, family="binomial")
train.pred <- predict(log.mod,newdata=train.data,type="response")

act <- as.numeric(train.data$multiple)-1
p.order <- order(train.pred)
probs <- train.pred[p.order]
act.order <- act[p.order]
act.cs <- cumsum(act.order)
x <- 1:length(act.cs)
threshold <- x - 2*act.cs
t <- as.numeric(probs[(which(threshold==max(threshold)))[1]])

test.pred <- predict(log.mod,newdata=test.data,type="response")
pred.vals <- as.factor(ifelse(test.pred >=t,1,0))
conf.mat <- caret::confusionMatrix(data = pred.vals,
                                    reference=test.data$multiple)

test.pred2 <- predict(log.mod,newdata=test.data,type="response")
pred.vals2 <- as.factor(ifelse(test.pred2 >=1,1,0))
conf.mat2 <- caret::confusionMatrix(data = pred.vals2,
                                    reference=test.data$multiple)
rsseb[i] <- (conf.mat$table[1]+conf.mat$table[4])/nrow(test.data)
rsseb_base[i] <- (conf.mat2$table[1]+conf.mat2$table[4])/nrow(test.data)

}

par(mfrow = c(2,2))
boxplot(rssea, main = "Accuracy of Logistic model City A",
        ylab ="Accuracy", ylim = c(0.3,0.8), cex.main = 0.7)
boxplot(rssea_base,
        main = "Accuracy of logistic models predicting mode City A",
        ylab ="Accuracy",ylim = c(0.3,0.8), cex.main = 0.7)
boxplot(rsseb, main = "Accuracy of Logistic models City B",
        ylab ="Accuracy", ylim = c(0.3,0.8), cex.main = 0.7)
boxplot(rsseb_base,
        main = "Accuracy of logistic models predicting mode City B",
        ylab = "Accuracy",ylim = c(0.3,0.8), cex.main = 0.7)
```



From these plots it is visible that modelling a smaller problem based on length of stay is much more appropriate than modelling length of stay itself. For both cities the logistic regression model greatly improves in accuracy compared to just predicting the most common class (LOS = 1). For these models I have also used some old code to optimize the prediction threshold for each model.

From these models I believe there is significant evidence to suggest that any form of LOS modelling should be performed through local models or a simpler choice of response to model.

3.The diag01 variable defines the initial diagnosis for each patient on admission. Select a diagnosis that has a reasonable number of occurrences for both cities (i.e. at least 300), and build a linear model to predict LOS just for patients with this diagnosis. To assess the quality of the model use a 90% training-10% test split for 50 replicates and measure error using RRSE (so the result gives an measure effectively against the mean – see lab notes). Compare and discuss the results in relation to the type of diagnosis you have chosen, it's distribution of LOS, patient characteristics, etc. (15 marks)

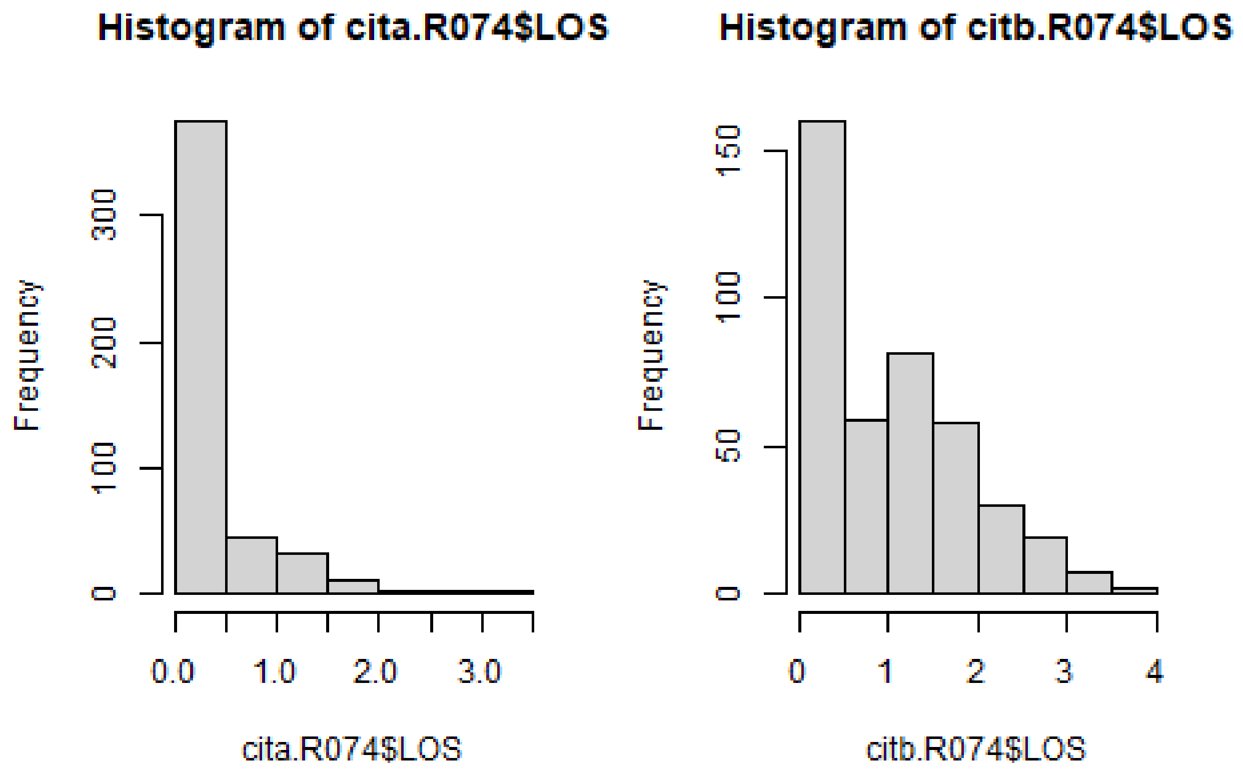
```r
head(sort(table(cita_clean$diag01), decreasing = TRUE), 10)
```

```
##
##  Z380 G4732  N390 L0311   I48  J189  O342  R103  R073  R074
##  1846   879   617   593   582   570   533   500   486   467
```

```r
head(sort(table(citb_clean$diag01), decreasing = TRUE), 10)
```

```
##
##  Z380  J189  R074  I500   R55  N390  O342 G4732  B349 L0311
##  1244   428   416   282   278   276   269   249   243   232
```

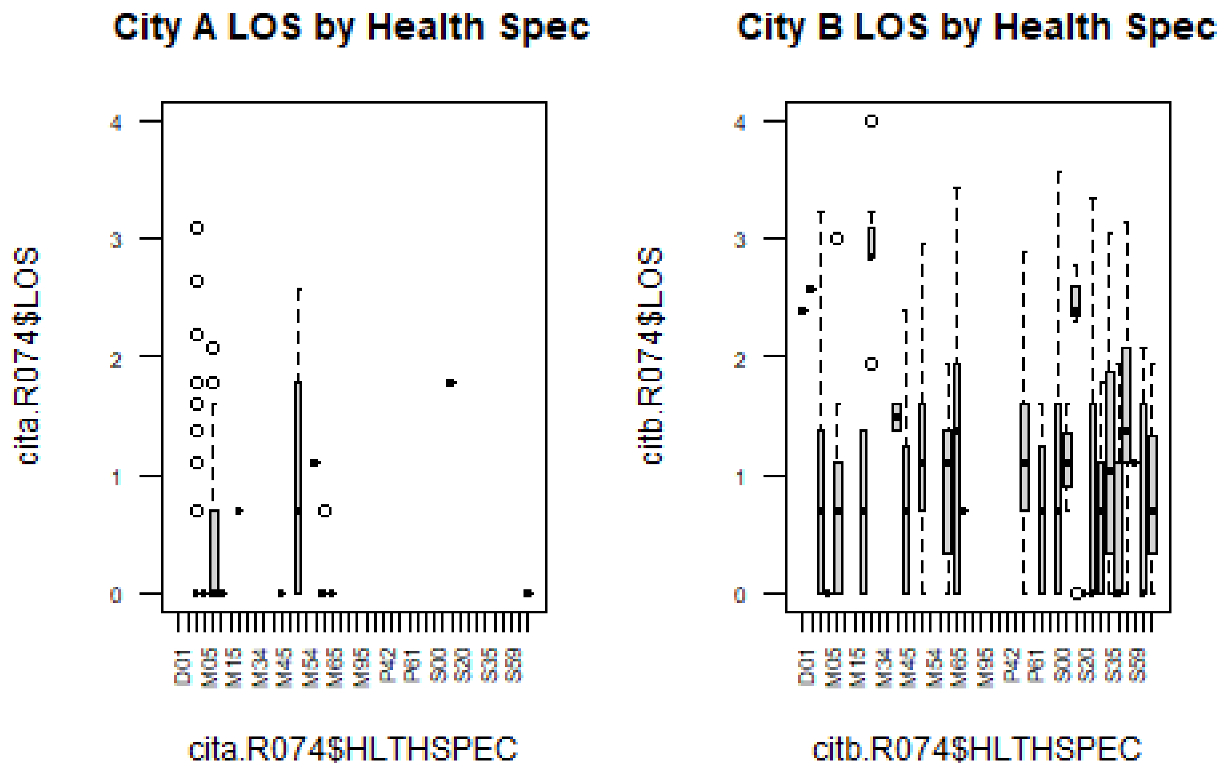
```r
cita.R074 <- cita_clean[which(cita_clean$diag01 == 'R074'),]
citb.R074 <- cita_clean[which(citb_clean$diag01 == 'R074'),]
par(mfrow=c(1,2))
hist(cita.R074$LOS)
hist(citb.R074$LOS)
```



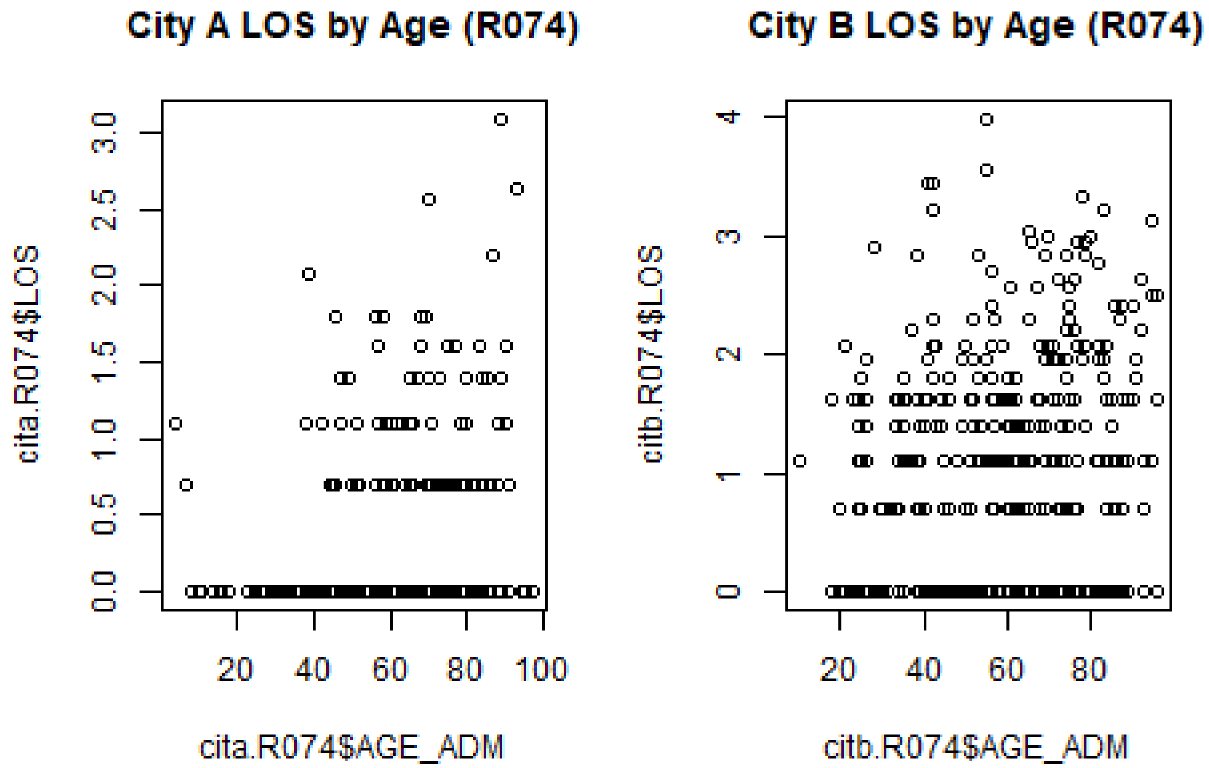
I chose the health code R074 as for both cities they have a similar number of observations. The health code is used for unspecified chest pain. I assume this mainly relates to patients who come with chest pain to the emergency ward. As this also relates to unspecified pain it is also plausible that this may contain a fair amount of variation in length of stay as it could relate to different diagnoses.

Looking at the distribution of length of stay across both cities for R074 patients it appears that city B tends to keep patients in for a longer period of time. This could potentially be due a smaller city being able to spend a longer period of time monitoring patients due to a smaller intake or could be due to a larger city having more patients come in which could have a much higher proportion of lower risk patients.

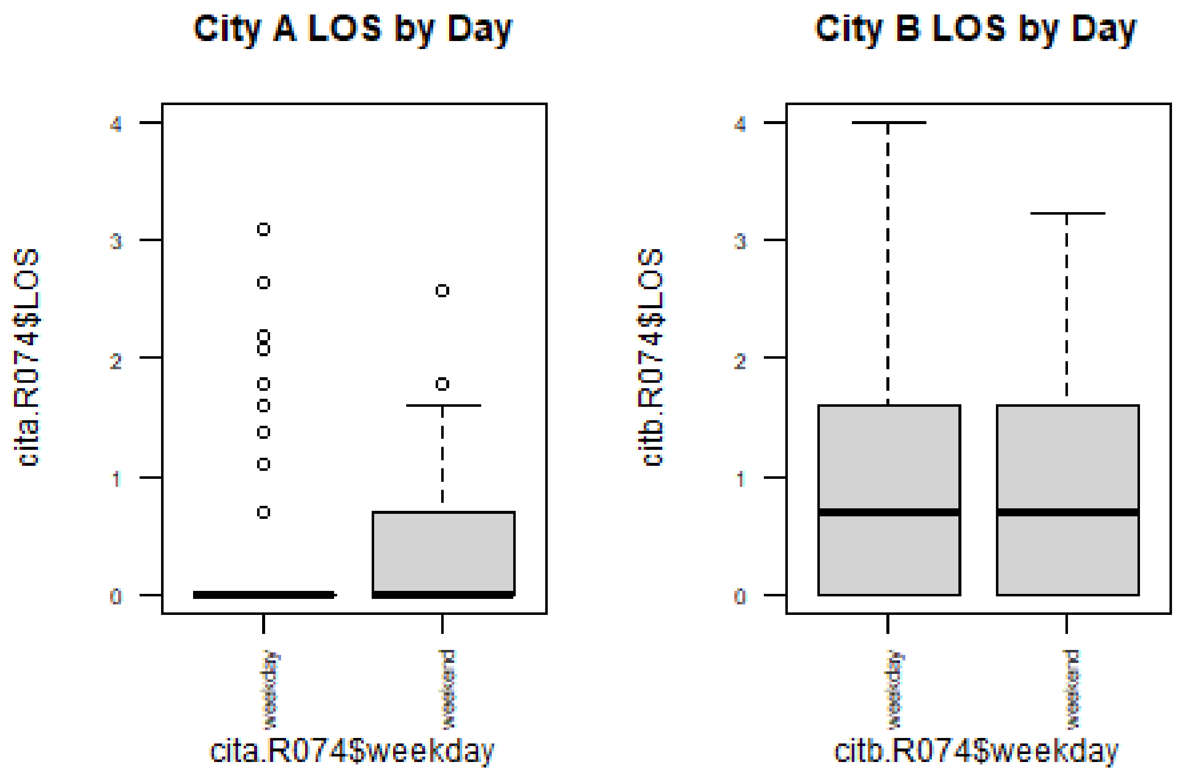
```
par(mfrow=c(1,2))
boxplot(cita.R074$LOS ~ cita.R074$HLTHSPEC, ylim = c(0,4), las = 2, cex.axis = 0.7,
        main = "City A LOS by Health Spec")
boxplot(citb.R074$LOS ~ citb.R074$HLTHSPEC, ylim = c(0,4), las = 2, cex.axis = 0.7,
        main = "City B LOS by Health Spec")
```



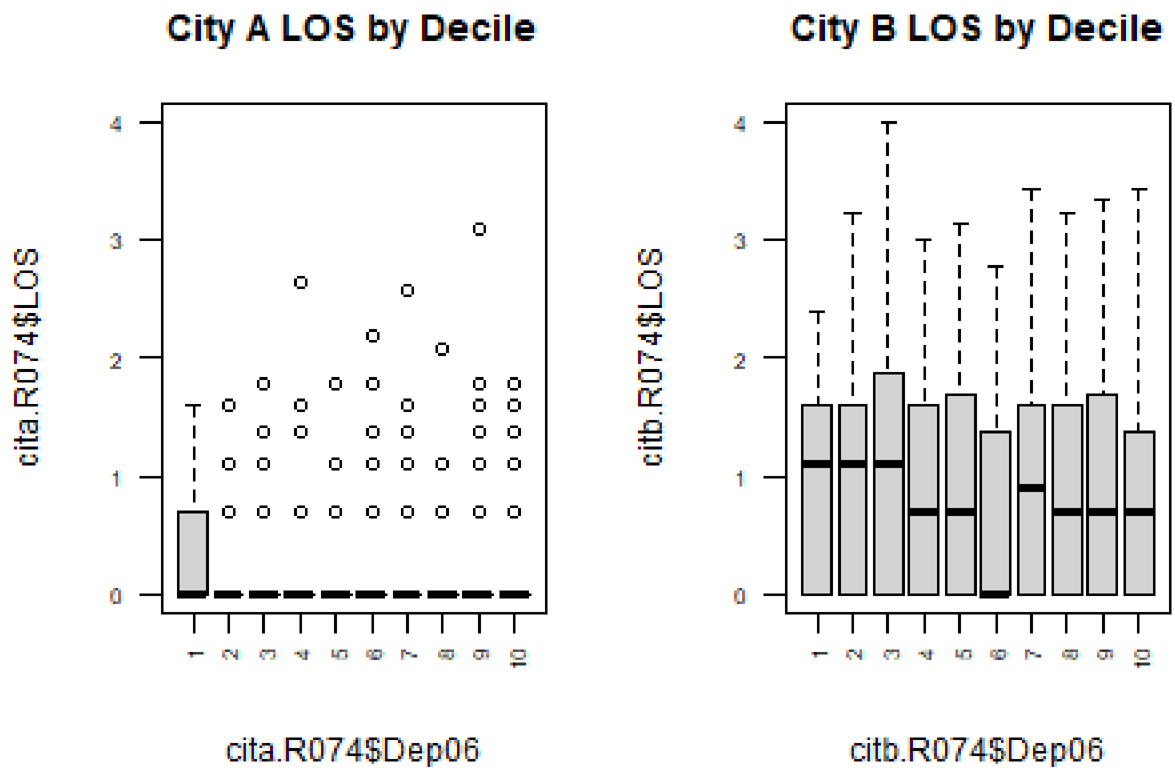
```
plot(cita.R074$AGE_ADM, cita.R074$LOS, main = "City A LOS by Age (R074)")
plot(citb.R074$AGE_ADM, citb.R074$LOS, main = "City B LOS by Age (R074)")
```



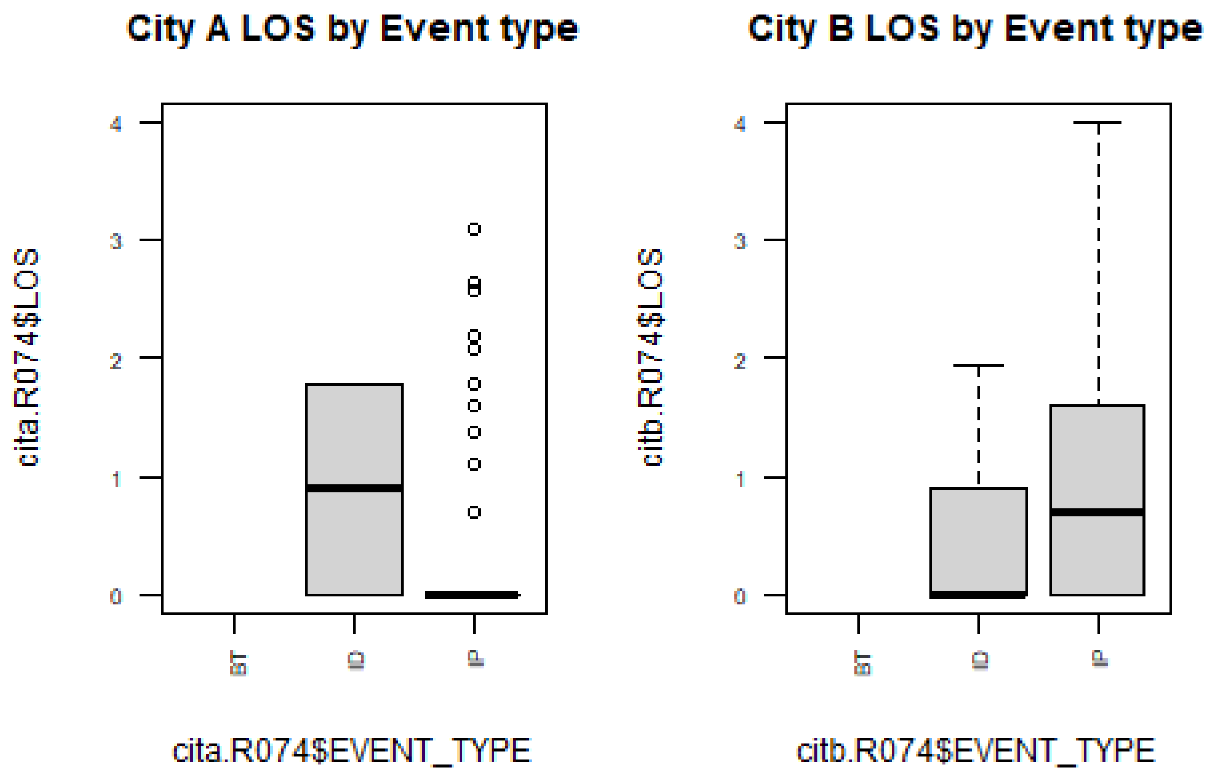
```
boxplot(cita.R074$LOS ~ cita.R074$weekday, ylim = c(0,4), las = 2, cex.axis = 0.7, main = "City A LOS by
boxplot(citb.R074$LOS ~ citb.R074$weekday, ylim = c(0,4), las = 2, cex.axis = 0.7, main = "City B LOS by
```



```
boxplot(cita.R074$LOS ~ cita.R074$Dep06, ylim = c(0,4), las = 2, cex.axis = 0.7, main = "City A LOS by
boxplot(citb.R074$LOS ~ citb.R074$Dep06, ylim = c(0,4), las = 2, cex.axis = 0.7, main = "City B LOS by
```



```
boxplot(cita.R074$LOS ~ cita.R074$EVENT_TYPE, ylim = c(0,4), las = 2, cex.axis = 0.7, main = "City A LO
boxplot(citb.R074$LOS ~ citb.R074$EVENT_TYPE, ylim = c(0,4), las = 2, cex.axis = 0.7, main = "City B LO
```



To reassess the explanatory based on the subset I have replotted some relationships between LOS and the predictors. First for patients who have a case where hospitalization is longer than a day there appears that the age of a patient has a relationship with their length of stay. This could be modeled with an interaction term between age and the response variable for multiple days I made for the logistic regression model.

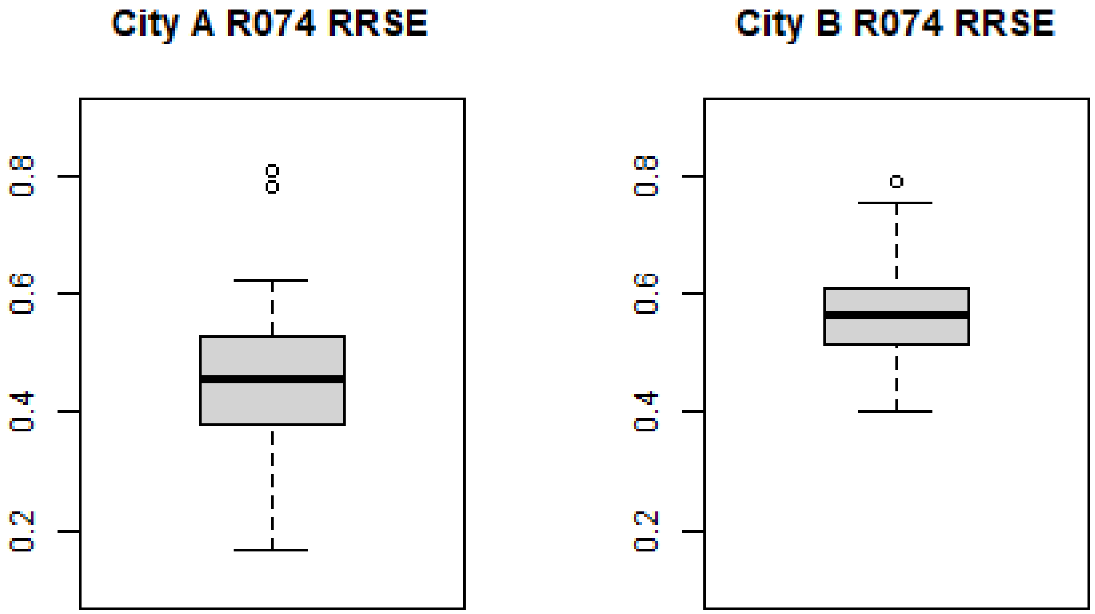
It also appears that city B deals with a wider range of health codes related to unspecified chest pain which may be a factor in the less skewed distribution of length of stay compared to city A. There is also a visual suggestion that event types may have different effects on Length of stay between the cities but this is not conlcusive evidence.

A patient's decile still has no effect on length of stay and the admission day is also no longer an effect. This makes sense as undiagnosed chest pain is likely a random event.

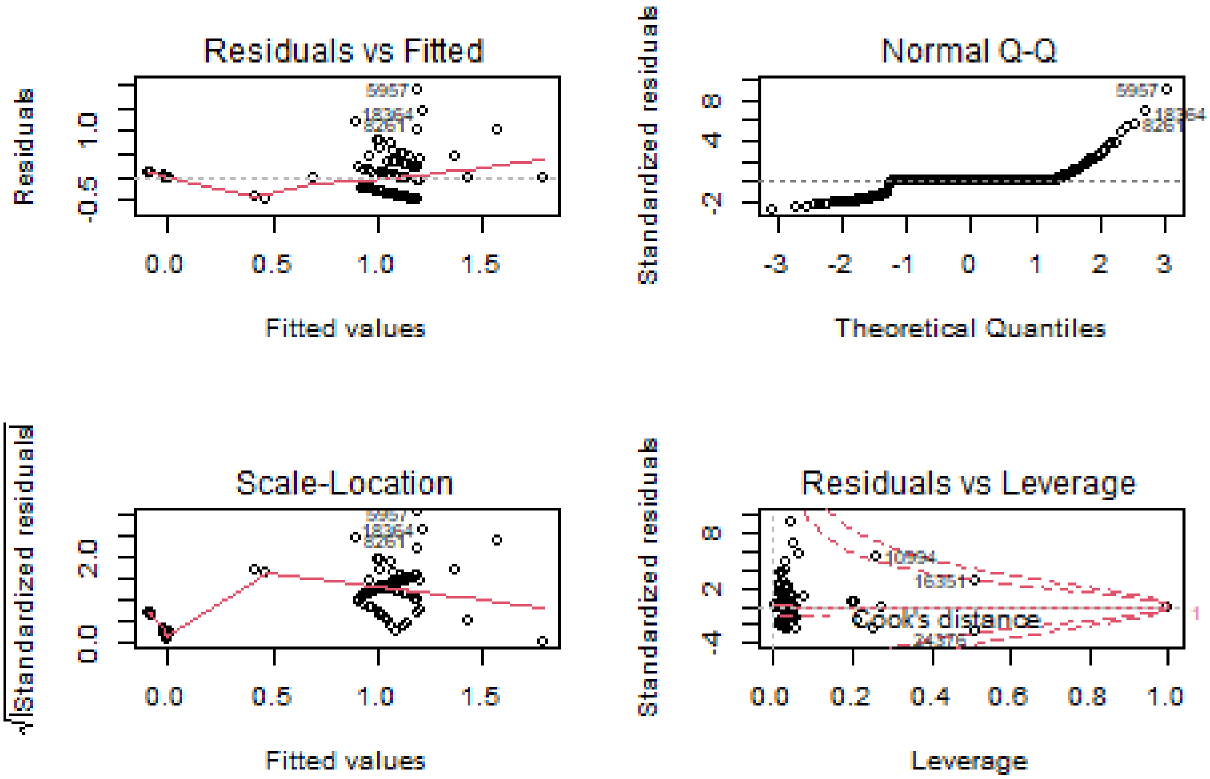
```
set.seed(3)
rrsea <- c()
rrseb <- c()

for (i in 1:100) {
  rrsea[i] <- test.lm(cita.R074, LOS ~ AGE_ADM + GENDER + EVENT_TYPE +
                      HLTHSPEC + AGE_ADM * multiple)

  rrseb[i] <- test.lm(citb.R074, LOS ~ AGE_ADM + GENDER + EVENT_TYPE +
                      HLTHSPEC + AGE_ADM * multiple)
}


R074ab <- lm(LOS ~ AGE_ADM + GENDER + EVENT_TYPE + HLTHSPEC +
               AGE_ADM * multiple, data = cita.R074)
summary(R074ab)
```

```
##
## Call:
## lm(formula = LOS ~ AGE_ADM + GENDER + EVENT_TYPE + HLTHSPEC +
##     AGE_ADM * multiple, data = cita.R074)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49952 -0.00353  0.00342  0.00700  1.90125
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.359e-02  1.909e-01  -0.385 0.700136
## AGE_ADM            9.674e-05  7.278e-04   0.133 0.894315
## GENDER1            8.495e-03  2.043e-02   0.416 0.677760
## EVENT_TYPEIP       6.262e-02  1.860e-01   0.337 0.736542
## HLTHSPECM05        3.952e-03  4.237e-02   0.093 0.925726
## HLTHSPECM10       -1.170e-02  2.727e-02  -0.429 0.668087
## HLTHSPECM14        9.416e-03  2.163e-01   0.044 0.965297
## HLTHSPECM25       -4.466e-02  2.402e-01  -0.186 0.852554
## HLTHSPECM45        6.998e-03  2.145e-01   0.033 0.973987
## HLTHSPECM50        4.793e-01  1.085e-01   4.418 1.25e-05 ***
## HLTHSPECM55        3.777e-01  2.424e-01   1.558 0.119990
## HLTHSPECM60       -7.686e-02  9.664e-02  -0.795 0.426830
## HLTHSPECM65        1.019e-02  2.172e-01   0.047 0.962603
## HLTHSPECS15        7.143e-01  2.156e-01   3.314 0.000996 ***
## HLTHSPECS75        2.451e-03  2.152e-01   0.011 0.990919
## multiple1          7.010e-01  1.276e-01   5.495 6.55e-08 ***
## AGE_ADM:multiple1  5.519e-03  1.833e-03   3.011 0.002752 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2135 on 450 degrees of freedom
## Multiple R-squared:  0.8133, Adjusted R-squared:  0.8067
## F-statistic: 122.5 on 16 and 450 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,2))
boxplot(rrsea, main = "City A R074 RRSE", ylim = c(0.1,0.9))
boxplot(rrseb, main = "City B R074 RRSE", ylim = c(0.1,0.9))
```



```
par(mfrow = c(2,2))
plot(R074ab)
```



From the plots it appears that a local model greatly improves on the model performance. It appears there is evidence to suggest that LOS for R074 patients in city A is easier to predict than city B. I believe that this is mainly down to the stronger relationship between age and LOS for patients who stayed longer than a day. My reasoning for this is that when I fitted a model for city A with and without the interaction term the adjusted R-squared went from 0.8067 to 0.1691, this indicates that the interaction for these patients is easily the most important predictor in the model.

Unfortunately from the diagnostic plots the model still fails the same assumptions of the linear model as the global model did. This suggests that in order to fit a linear model that meets the assumptions a specific range for length of stay may need to be modelled due to the two category nature of length of stay.

4. Discuss why predicting LOS is a difficult task and reflect on the way bed counts (i.e. admissions) are currently handled. Is there any way this might be improved? Discuss some approaches to break the problem down into more tractable aspects for managing admissions and predicting LOS. What does this suggest about building a model for LOS over the entire country? (10 marks)

From of all the models shown there is significant evidence that LOS is incredibly hard to predict due to the systematic skew of the data caused by operating procedures. It is clearly visible that each of wide range of treatments and procedures has unique characteristics that are very hard to assess on a consistent scale when many of the predictors provide a lack of information.

There also appears to be evidence that every hospital operates differently based on their size. To obtain any reliable results I believe there is no feasible way to create a single model that can predict length of stay for a patient. At a minimum I believe for each hospital every ward should have it's own local model. Although the R074 models failed on model assumptions I still believe they showed evidence that the problem has to be reduced to a smaller problem within each location to understand any true relationships in the data. The differences in the two hospitals alone show conclusive evidence that LOS is affected by different aspects in different locations so there is no possible way a global model for the entire country can provide reliable information.

Over time use of local modelling could allow more specific data to be collected which could further improve the models the effect of how each diagnosis and treatment given correlates to aspects of each individual patient so allocation of resources can further benefit both the efficiency of the hospital and the quality of treatment a patient receives.

I believe with some more work, a few of the approaches I trialed could be successful for managing aspects of hospital operations. As I mentioned above I believe a local model for each ward would help with this (something similar to how the R074 model approached the problem). I also believe the classification method I trialed earlier of predicting if a patient will stay for longer than a day could be a highly plausible and useful method due to the large amount of patients admitted only for a single day. Another interesting method I believe could be trialed for predicting bed counts is to try and predict the number of cases for each diagnosis that are likely over a certain time period and then use the length of stay from existing patients to try and predict the resources a hospital will need.

Overall while modelling LOS as a global problem is difficult there are many possible ways it can be broken down to model specific aspects of it or to use the information available to solve other existing problems within the hospitals that could provide more information to model LOS itself.