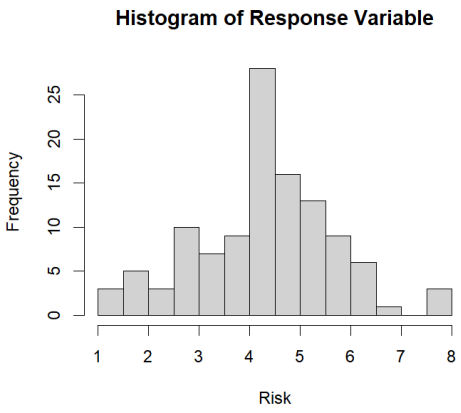
What affects the risk of picking up an Infection?

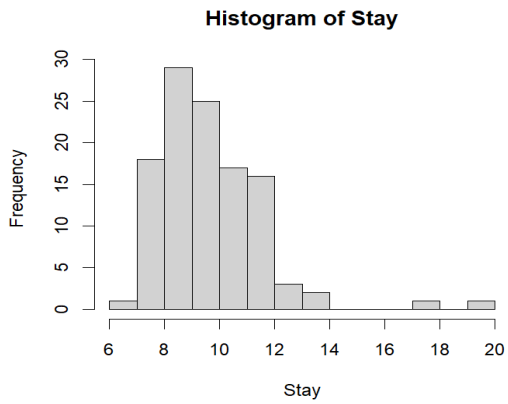
Checking data and finding correlations

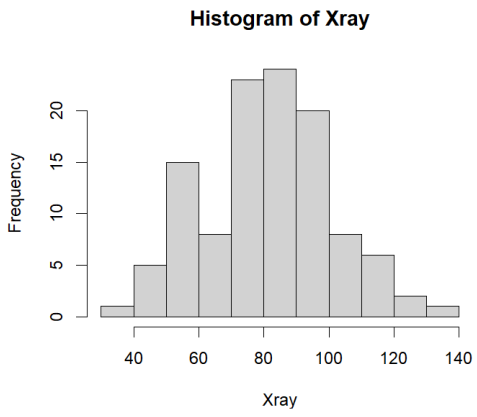
Inputting my data into R, I checked the dimensions to see if it had loaded correctly, the data came up as what I had expected with 113 rows and 4 columns. Response variable being risk and the explanatory variables being Stay, Xray, and Age I plotted the response variable as a histogram to check for normal distribution.



This distribution is showing a bell shaped normal distribution which indicates that the data for my response variable is accurate and that it's okay to fit a model to it.

I also plotted a histogram of Stay and Xray as I may need to use these as response variables later on when testing interactions / confounding variables.

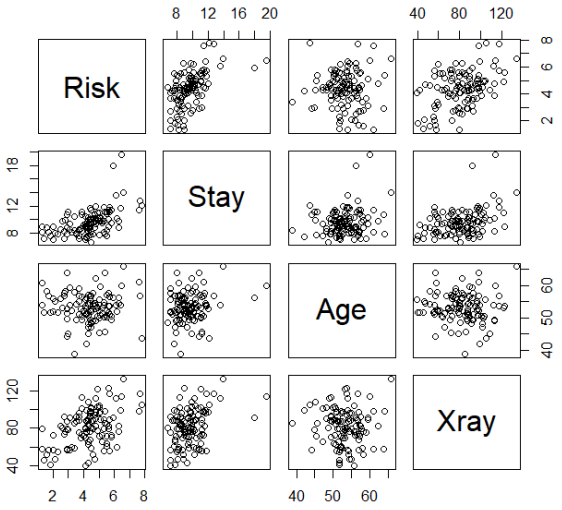




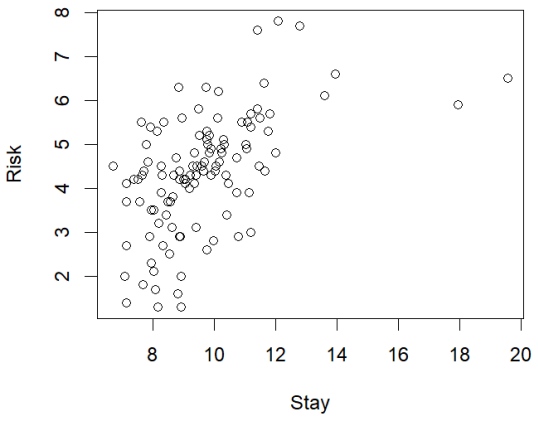
Histogram of Stay shows a slightly right skewed graph meaning that it isn't ideal to have as a response variable

Histogram of XRay shows a normally distributed graph meaning that it is ideal to have this as a response variable.

To try and predict any sort of relationships between my variables, I plotted a matrix with all variables and noted that there's possible correlations between Stay and Risk, Stay and Xray, Xray and Ris. I plotted these individual graphs against each other to get clarity as to whether or not my predictions are assumptions.



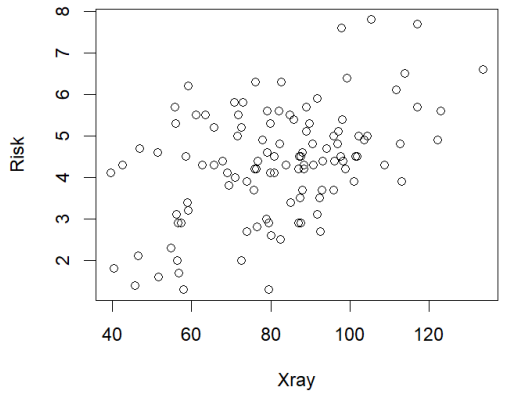
(Matrix of all variables plotted)



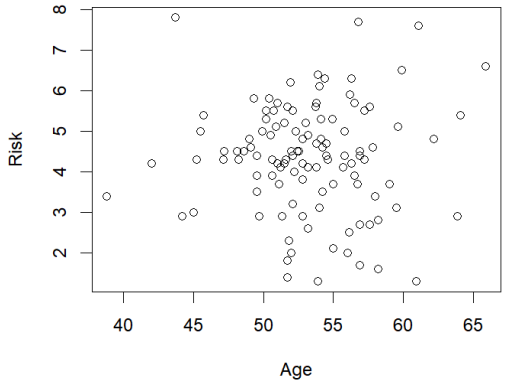

```
Call:
lm(formula = Risk ~ Stay, data = infection)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7823 -0.7039  0.1281  0.6767  2.5859

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.74430    0.55386   1.344    0.182
Stay         0.37422    0.05632   6.645 1.18e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.139 on 111 degrees of freedom
Multiple R-squared:  0.2846,    Adjusted R-squared:  0.2781
F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09
```

The data here seems to be following a positive linear relationship where the longer the patient stays in the hospital, the higher the risk of getting infected. I plotted a linear regression model against this and summarised the data to find that the model is significant and that there's a positive relationship between these two variables.

Next, I plotted XRay and Risk, observing a clear positive linear relationship. This is suggesting that as the number of XRays increases, a patient's average risk of infection also rises. The summary of my linear regression model confirmed this relationship as statistically significant. However, given the noticeable scatter in the graph additional variance testing may be necessary to conclude my prediction.




```
Call:
lm(formula = Risk ~ Xray, data = infection)

Residuals:
     Min       1Q   Median       3Q      Max
-2.98805 -0.85492 -0.04643  0.91416  2.73426

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.792024   0.491363   3.647 0.000406 ***
Xray        0.031396   0.005858   5.359 4.58e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.201 on 111 degrees of freedom
Multiple R-squared:  0.2056,    Adjusted R-squared:  0.1984
F-statistic: 28.72 on 1 and 111 DF,  p-value: 4.585e-07
```

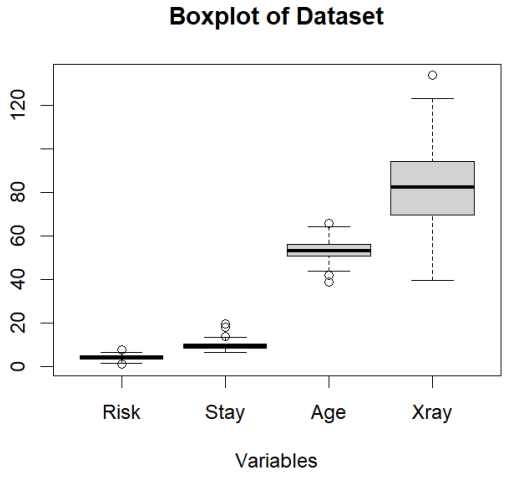


Following the trend, I also plotted Age and Risk and found no obvious relationship between the two variables, there’s too much scatter not following any sort of shape. To check my prediction here, again, I summarised a linear regression model for this data and found no significant relationship.

The summary coefficient is showing that age isn’t statistically significant here sitting at a p-value of 0.99 which is very high. Therefore allowing me to make the conclusion that there’s no linear relationship between these variables.

```
Call:
lm(formula = Risk ~ Age, data = infection)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0574 -0.6568  0.0455  0.8457  3.4483

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.3373782  1.5237868   2.846  0.00527 **
Age         0.0003285  0.0285263   0.012  0.99083
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.347 on 111 degrees of freedom
Multiple R-squared:  1.195e-06, Adjusted R-squared:  -0.009008
F-statistic: 0.0001326 on 1 and 111 DF,  p-value: 0.9908
```

I also plotted a boxplot to show the difference in my variables which allows me to clearly see any uncommon residuals that R has picked up. I can see residuals for all these variables. I will need to check them as they could be skewing my data by not occurring randomly.



Creating a linear regression model between Predictor Variables and Risk.

I used the given code from the assignment question to create a linear regression model with all 3 of my predictor variables and summarised the data with my response variable Risk. Below I was given the output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.001162   1.314724   0.761 0.448003
Stay         0.308181   0.059396   5.189 9.88e-07 ***
Age         -0.023005   0.023516  -0.978 0.330098
Xray         0.019661   0.005759   3.414 0.000899 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.085 on 109 degrees of freedom
Multiple R-squared:  0.363,     Adjusted R-squared:  0.3455
F-statistic:  20.7 on 3 and 109 DF,  p-value: 1.087e-10
```

With an F-Statistic p-value of 1.087e-10, I can confidently conclude that the overall model is statistically significant, meaning that at least one of the predictor variables is related to the Risk of Infection. The Multiple R-Squared value of 0.363 indicates that approximately 36.3% of the variation in the Risk of Infection is explained by the model. Although 36.3% is not very high, it

still suggests a moderate level of explanation, and there may be other factors influencing the risk that are not included in the model.

The intercept isn't statistically significant with a p-value indicating that it is not reliably different from zero. This suggests that when all predictor variables are zero, the predicted baseline risk may not be accurately estimated by this model. This is logical because in practical terms, if a patient does not attend the hospital, it makes sense that their risk of infection would be negligible.

Looking at individual predictors:

Age: The p-value of 0.33 indicates that age is not significantly associated with the Risk of Infection.

XRay and Stay: Both variables show a significant relationship with the Risk of Infection, which is consistent with their relationships observed in individual plots.

Based on the significance of these predictors, I can consider simplifying the model by focusing on the variables that significantly contribute to the risk, namely XRay and Stay.

Simplifying the model

Because the variable Age had no statistical significance in this model, I can simplify the model by removing Age from it. I realise that Age may contribute to my Xray and Stay variables which are significant in this model so I created models for these two variables summarising Age against both variables to see any statistical significance. **XRay given age** significance: 8.42, **Stay given Age**: 0.04. Since I had statistical significance between my Age and Stay variables, I need to look further into this as it may be a confounding variable affecting my model. The p-value of Stay given Age = 0.045 meaning it's borderline significant $p<0.05$, the R squared value is only 0.036 meaning that only 3.6% of the variation is accounted for because of Age. I created a new model without age and compared it using an ANOVA test with the model with Age. This tells me if there's improved statistical significance in which case I can remove the variable "Age". Here's what I found…

```
Analysis of Variance Table

Model 1: Risk ~ Stay + Age + Xray
Model 2: Risk ~ Stay + Xray
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    109 128.28
2    110 129.41 -1   -1.1263 0.957 0.3301
```

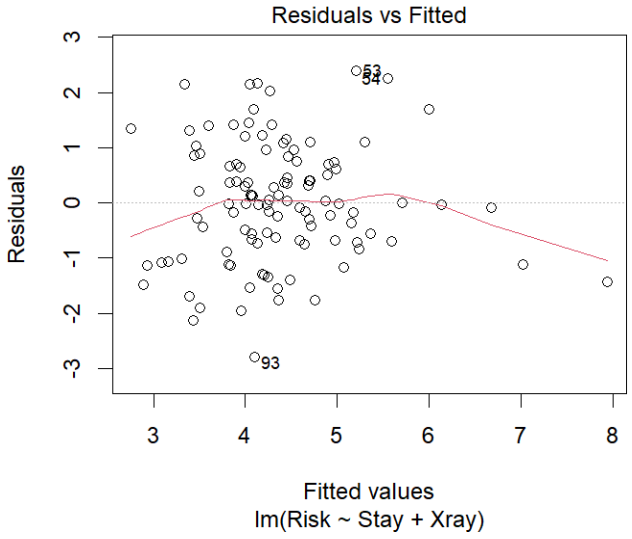
Comparing the 2 models in an ANOVA table gives a significance of 0.33 which is well over 0.05, this means that Age isn't playing a significant factor in the explanatory for the Risk of Infection therefore I'm able to remove it from the model and proceed with my simplified version Risk ~ Stay + Xray

```
Call:
lm(formula = Risk ~ Stay + Xray, data = infection)

Residuals:
     Min       1Q   Median       3Q      Max
-2.79635 -0.71592 -0.01511  0.73899  2.39483

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.150603   0.585036  -0.257 0.797331
Stay         0.295845   0.058030   5.098 1.44e-06 ***
Xray         0.020227   0.005728   3.531 0.000606 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.085 on 110 degrees of freedom
Multiple R-squared:  0.3574,    Adjusted R-squared:  0.3457
F-statistic: 30.59 on 2 and 110 DF,  p-value: 2.734e-11
```

After summarising my new linear regression model I'm already seeing that the significance in the F-statistic P-value and the F-statistic has improved dramatically (1.087e-10 - 2.734e-11) and (20.7 - 30.59). Both of my explanatory variables are highly significant indicating that this is a suitable model to interpret the relationship they have with the Risk of Infection.

Now that my model has been simplified and fitted, I will check the residual graphs to see whether or not my variance scatter is as it should be and if there are any outliers affecting my data.

Checking residuals for my new model :)

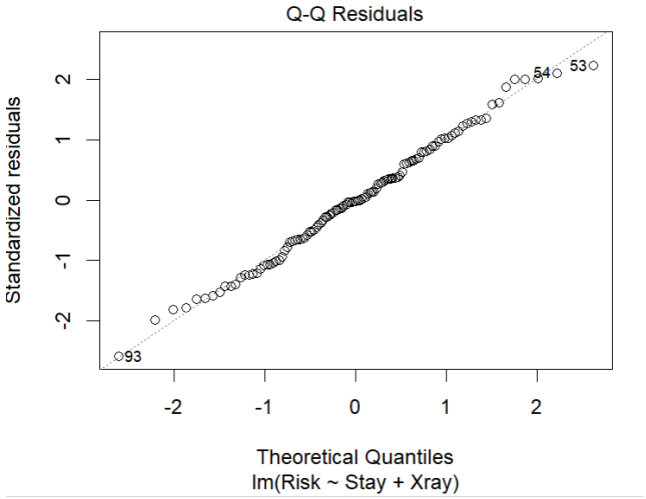


Which = 1

From this residual graph, I'm seeing a trend in scatter being around the 3-5 region, along with this bent line. I'm not sure if I can trust this model although it's acceptable. After seeing this graph I created a quadratic model and summarised it to see if it would better fit, these were the results.

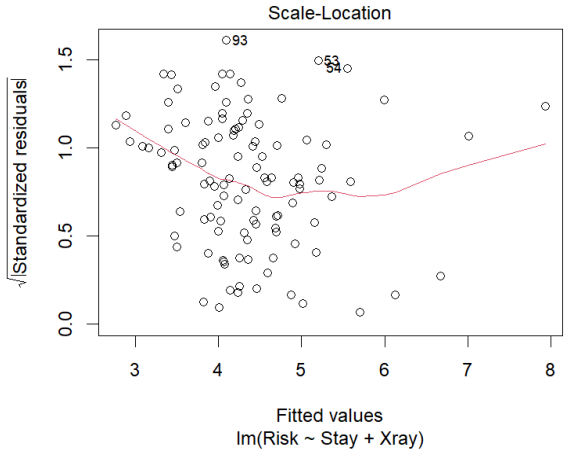
```
Call:
lm(formula = Risk ~ Stay + I(Stay^2) + Xray, data = infection)

Residuals:
     Min       1Q   Median       3Q      Max
-2.80182 -0.75364 -0.05399  0.64317  2.33773

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.018905   1.600887  -1.886 0.061988 .
Stay         0.828245   0.282974   2.927 0.004167 **
I(Stay^2)   -0.022846   0.011891  -1.921 0.057306 .
Xray         0.019504   0.005672   3.439 0.000829 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 109 degrees of freedom
Multiple R-squared:  0.3784,    Adjusted R-squared:  0.3613
F-statistic: 22.12 on 3 and 109 DF,  p-value: 2.907e-11
```

This quadratic model in comparison to my other model shows a similar level of significance (2.734e-11 - 2.907e-11) and shows that my variables aren't significant. I can't use this model so I will continue to use the model from earlier and look at the rest of the residual graphs as I suspect there may be outliers influencing my data.



Which = 2

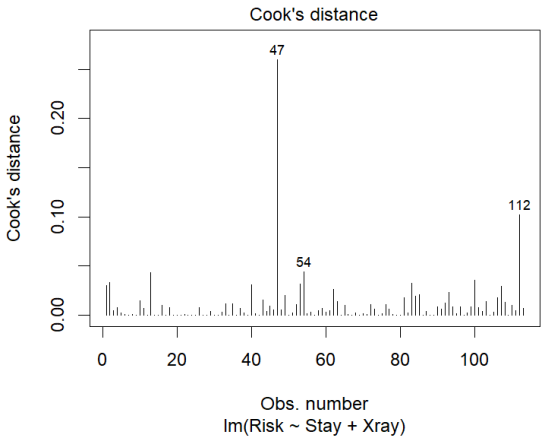
From this graph, I can see that my model residuals are normally distributed which ticks off an assumption, the residuals at the extremes are notable (93, 54, 53)



Which = 3

This graph shows consistent spread which allows me to assume constant variance although I'm seeing again, a bent red line which could indicate there's some non-linearity in my model that isn't being accounted for. This is okay because the line isn't too dramatic so I can still work with the linear model. Points (93, 53, 54) are still notable in this residual plot

Which = 4



The provided Cook's distance graph shows points passing 0.4 although (47) gets fairly close. After seeing this it's okay to say that the outliers aren't playing much of a factor in skewing my data.

Pulling results from the Simplified Model

The model has been refined and simplified down to Risk ~ Stay + Xray and the output summary is as below.

```
Call:
lm(formula = Risk ~ Stay + Xray, data = infection)

Residuals:
     Min       1Q   Median       3Q      Max
-2.79635 -0.71592 -0.01511  0.73899  2.39483

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.150603   0.585036  -0.257 0.797331
Stay         0.295845   0.058030   5.098 1.44e-06 ***
Xray         0.020227   0.005728   3.531 0.000606 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.085 on 110 degrees of freedom
Multiple R-squared:  0.3574,    Adjusted R-squared:  0.3457
F-statistic: 30.59 on 2 and 110 DF,  p-value: 2.734e-11
```

From these numbers because I don't have a statistically significant intercept, I can't accurately say that when a person doesn't stay at the hospital and has 0 XRays that on average the risk of an infection will be 0.8. What I can conclude is that there's a statistically significant relationship between Stay, XRay in accordance with the Risk of Infection.

Conclusion: For every day longer a patient stays in the hospital, on average, the risk of infection increases by 0.3 units. For every XRay a patient gets, on average, the risk of infection increases by 0.02 units. This is likely because a patient that stays longer has a higher chance of exposure to any infection from equipment, surfaces, or people around them. The XRay may have a correlation between the risk of infection because of the surfaces, human error from people operating the equipment, or different waiting rooms. The explanation from this model summary makes logical sense giving me more confidence in my conclusion.

Interaction model and the relationship with the response variable

```
Call:
lm(formula = Risk ~ Stay * Xray, data = infection)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8079 -0.8136 -0.0008  0.6488  2.3733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.092404   2.328880  -1.328   0.1870
Stay         0.606997   0.245391   2.474   0.0149 *
Xray         0.051957   0.024980   2.080   0.0399 *
Stay:Xray   -0.003306   0.002534  -1.305   0.1947
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.081 on 109 degrees of freedom
Multiple R-squared:  0.3673,    Adjusted R-squared:  0.3499
F-statistic: 21.09 on 3 and 109 DF,  p-value: 7.564e-11
```

In R to create my linear regression interaction model, I used the command `infection.lm3<-lm(Risk ~ Stay * Xray, data=infection)`
Where Stay * Xray is the interaction. Firstly in my F-statistic p-value of 7.564e-11 I see that this model is statistically significant. The interaction variable of Stay:XRay has a P-value of 0.19 indicating that there's no evidence of interaction between the amount of days a patient stays and the amount of XRays a patient receives on average. Based on this significance, the model Risk ~ Stay + XRay is still sufficient.

How many observations are there using degrees of freedom?

ANOVA table on my model gave me the output

```
Analysis of Variance Table

Response: Risk
           Df  Sum Sq Mean Sq F value    Pr(>F)
Stay        1  57.305  57.305  48.711 2.353e-10 ***
Xray        1  14.668  14.668  12.468  0.000606 ***
Residuals 110 129.407   1.176
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

To calculate the amount of observations, all I need to do is sum the degrees of freedom and add 1. This calculation is 1 + 1 + 110 + 1 = 113. To check this answer, I called the dimensions command and saw that my data has 113 observations therefore confirming my result.