Covid Analysis using the UK BioBank

The goal of this analysis is to see what a person's phenotypes, lifestyle based or genetic, affect your infection rate and death rate to Covid-19. To do this, logit models were used to compare the effect of these factors on infection and death separately.

All models used the sex, age, and whether the patient was white or another ethnicity as baseline phenotypes, and were compared to the other factors as seen below.

```
# load all data used in analysis
load(file = "/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/condensed_covid.rda")
load(file = "/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/globalData.rda")
load(file = "/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/infectionData.rda")
load(file = "/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/positiveTestDeathData.rda")
```

Infection Models:

The first model is to see the effect of exercise on one's infection rate.

```
# infection analysis

tmp <- infectionData

fitInfectionExercise <- glm(result~tmp$Sex+tmp$Age*tmp$WhiteBritish
                        +tmp$'Above Moderate Exercise'+tmp$'Above Moderate Walking Exercise'
                        ,family=binomial(),data=tmp)

summary(fitInfectionExercise)
```

```
##
## Call:
## glm(formula = result ~ tmp$Sex + tmp$Age * tmp$WhiteBritish +
##     tmp$"Above Moderate Exercise" + tmp$"Above Moderate Walking Exercise",
##     family = binomial(), data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9026  -0.6672  -0.6347  -0.5718   1.9865
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -0.227207   0.521807  -0.435 0.663255
## tmp$Sex                             0.243623   0.067915   3.587 0.000334 ***
## tmp$Age                            -0.019296   0.009453  -2.041 0.041224 *
## tmp$WhiteBritish                   -0.646590   0.571648  -1.131 0.258014
## tmp$"Above Moderate Exercise"       0.056513   0.080149   0.705 0.480746
## tmp$"Above Moderate Walking Exercise"  0.011421   0.099983   0.114 0.909058
## tmp$Age:tmp$WhiteBritish            0.005732   0.010340   0.554 0.579365
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

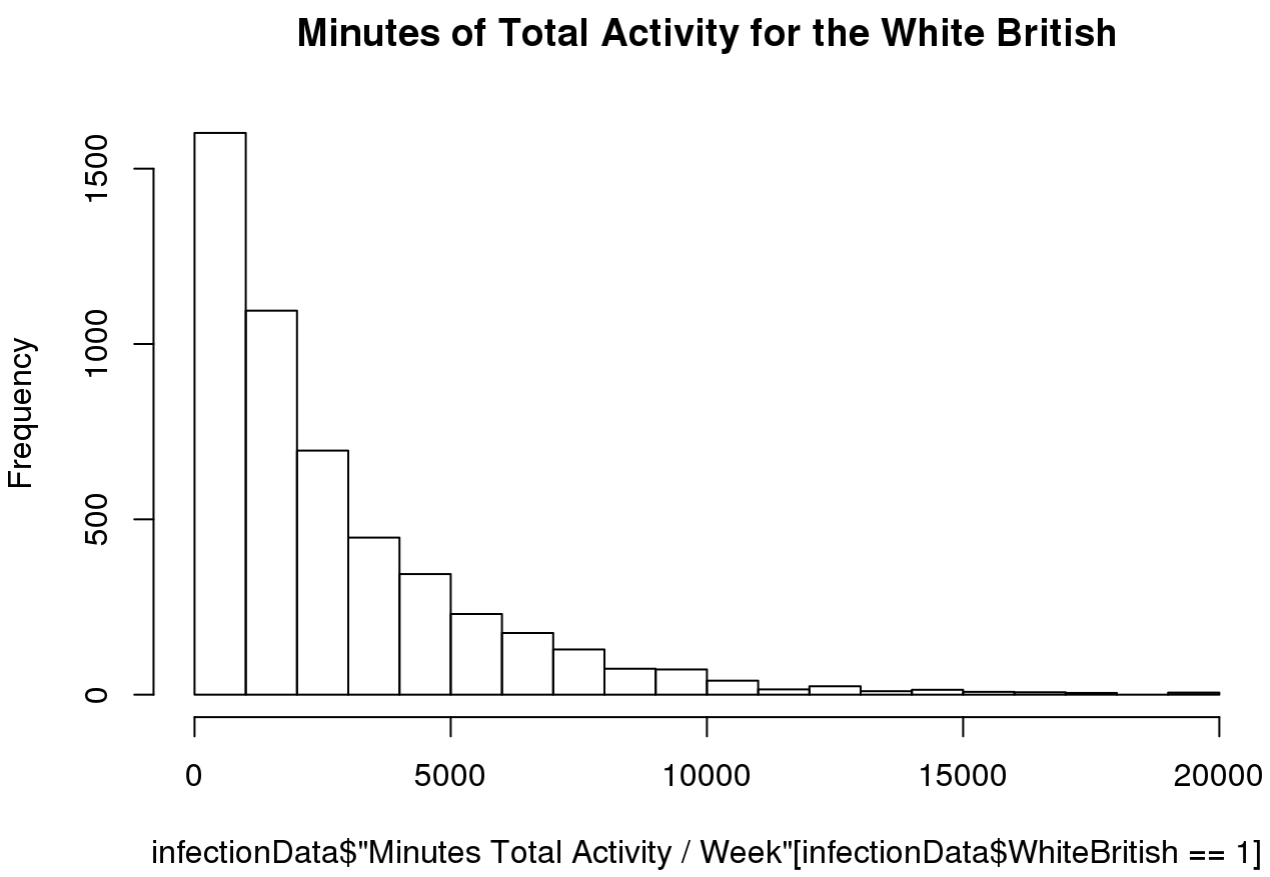
```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5684.7  on 5782  degrees of freedom
## Residual deviance: 5643.8  on 5776  degrees of freedom
##   (1591 observations deleted due to missingness)
## AIC: 5657.8
##
## Number of Fisher Scoring iterations: 4
```

Analysis:

- Being male increases your likelihood of being infection. This could be due to the risk taking nature of men, same reason our insurance is higher.
- You are less likely to be infected the older you are, this could be a result of the fact that the disease tends to be more dangerous for the elderly, and so they are taking more precautions to reducing infection than the young are.
- Ethnicity seems to have no effect on infection rates.
- Exercise and infection are not correlated.

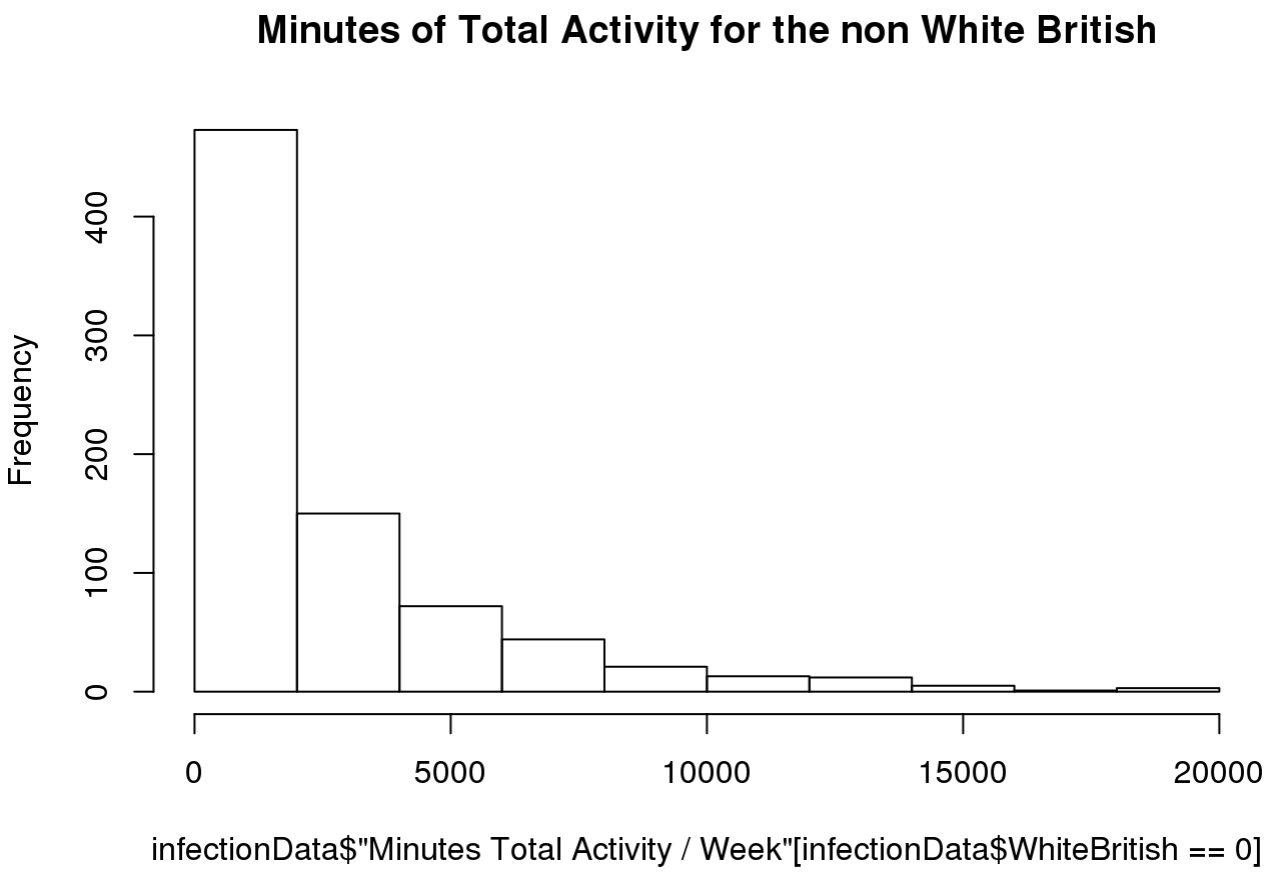
I try to understand the reason for ethnicity not playing a factor in this model, as in those to follow it does.

```
hist(infectionData$'Minutes Total Activity / Week'[infectionData$WhiteBritish==1],main="Minute
s of Total Activity for the White British")
```




```
hist(infectionData$'Minutes Total Activity / Week'[infectionData$WhiteBritish==0],main="Minute
```

```
s of Total Activity for the non White British")
```



It seems a person's exercise is irrelevant of their ethnicity, so this does not explain the model's disparity.

This model compares infection rates with sleep data measuring average hours of sleep per night, and whether a person is an insomniac.

```
fitInfectionSleep <- glm(result~tmp$Sex+tmp$Age
                      +tmp$WhiteBritish
                      +tmp$'Sleep Duration'+tmp$Insomnia
                      ,family=binomial(),data=tmp)

summary(fitInfectionSleep)
```

```
##
## Call:
## glm(formula = result ~ tmp$Sex + tmp$Age + tmp$WhiteBritish +
##     tmp$"Sleep Duration" + tmp$Insomnia, family = binomial(),
##     data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9179  -0.6826  -0.6361  -0.5782   1.9648
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)          -0.1353536  0.2563079  -0.528 0.597437
## tmp$Sex               0.2044952  0.0597661   3.422 0.000623 ***
## tmp$Age              -0.0160626  0.0034382  -4.672 2.99e-06 ***
## tmp$WhiteBritish     -0.3679570  0.0784543  -4.690 2.73e-06 ***
## tmp$"Sleep Duration" -0.0177426  0.0222060  -0.799 0.424290
## tmp$Insomnia         -0.0006758  0.0399354  -0.017 0.986500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7312.8  on 7302  degrees of freedom
## Residual deviance: 7253.9  on 7297  degrees of freedom
##   (71 observations deleted due to missingness)
## AIC: 7265.9
##
## Number of Fisher Scoring iterations: 4
```

Analysis:

- Sex and age results are consistent with the first model
- Not being white statistically significant in increasing infection in this model
- Sleep quality and duration has nothing to do with infection. This could be due to difficulties measuring the data accurately. Sleep will be re-evaluated later.

The following model seeks to analyze the specific effects of the number of different types of white blood cells in a person's immune system on infection rates.

```
cellprops <- cbind(tmp$NumBasophills,tmp$NumEosinophills,tmp$NumLymphocytes,
                   tmp$NumMonocytes,tmp$NumNeutrophills)
cellprops <- cellprops/apply(cellprops,1,sum)

fitInfectionImmuneSystem <- glm(result~tmp$Sex+tmp$Age+tmp$WhiteBritish
                            +tmp$NumWhiteBloodCells+cellprops[,1]+cellprops[,2] +
                              cellprops[,3]+cellprops[,4]+tmp$NumPlatelets+tmp$NumReticulocyte
s
                            +tmp$WhiteBritish
                            ,family=binomial(),data=tmp)

summary(fitInfectionImmuneSystem)
```

```
##
## Call:
## glm(formula = result ~ tmp$Sex + tmp$Age + tmp$WhiteBritish +
##     tmp$NumWhiteBloodCells + cellprops[, 1] + cellprops[, 2] +
##     cellprops[, 3] + cellprops[, 4] + tmp$NumPlatelets + tmp$NumReticulocytes +
##     tmp$WhiteBritish, family = binomial(), data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9678  -0.6825  -0.6360  -0.5787   2.0388
##
```

```
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.2601468  0.2999364  -0.867    0.386
## tmp$Sex                 0.1612942  0.0641769   2.513    0.012 *
## tmp$Age                -0.0154526  0.0035278  -4.380 1.19e-05 ***
## tmp$WhiteBritish       -0.3791160  0.0811362  -4.673 2.97e-06 ***
## tmp$NumWhiteBloodCells  0.0084978  0.0105930   0.802    0.422
## cellprops[, 1]         -9.7023426  5.1182616  -1.896    0.058 .
## cellprops[, 2]          0.8080292  1.3655417   0.592    0.554
## cellprops[, 3]          0.3425490  0.3764542   0.910    0.363
## cellprops[, 4]          0.5060004  1.0648041   0.475    0.635
## tmp$NumPlatelets       -0.0007890  0.0005117  -1.542    0.123
## tmp$NumReticulocytes    0.3954776  0.6819981   0.580    0.562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6995.6  on 6989  degrees of freedom
## Residual deviance: 6934.8  on 6979  degrees of freedom
##   (384 observations deleted due to missingness)
## AIC: 6956.8
##
## Number of Fisher Scoring iterations: 4
```

Analysis:

- Not being white is again relevant, suggesting the result of the first model on this aspect may be wrong
- Having less basophils may put people at a higher infection risk with a p value of .058. This would be interesting to keep examining. Basophils are part of the innate immune system, meaning they are circulating your blood and attack bodily invaders like parasites and viruses as they enter. This result suggests their importance tackling the initial covid-19 viral attack.

This last model aims to look at general life factors, including the Townsend Deprivation Index, to get a last look at a person's life experience, and its relation to infection rates.

```
# Number of basophills could put people at higher infection risk, p-val=.058. To look into

fitInfectionLife <- glm(result~tmp$'Townsend Deprivation Index'+tmp$Sex+tmp$Age+tmp$WhiteBriti
sh
                +tmp$'Sleep Duration'+Insomnia,data=tmp)
# All very corelated to infection except sleep and insomnia. Consistent with above results at
least

summary(fitInfectionLife)
```

```
##
## Call:
## glm(formula = result ~ tmp$"Townsend Deprivation Index" + tmp$Sex +
##     tmp$Age + tmp$WhiteBritish + tmp$"Sleep Duration" + Insomnia,
##     data = tmp)
##
```

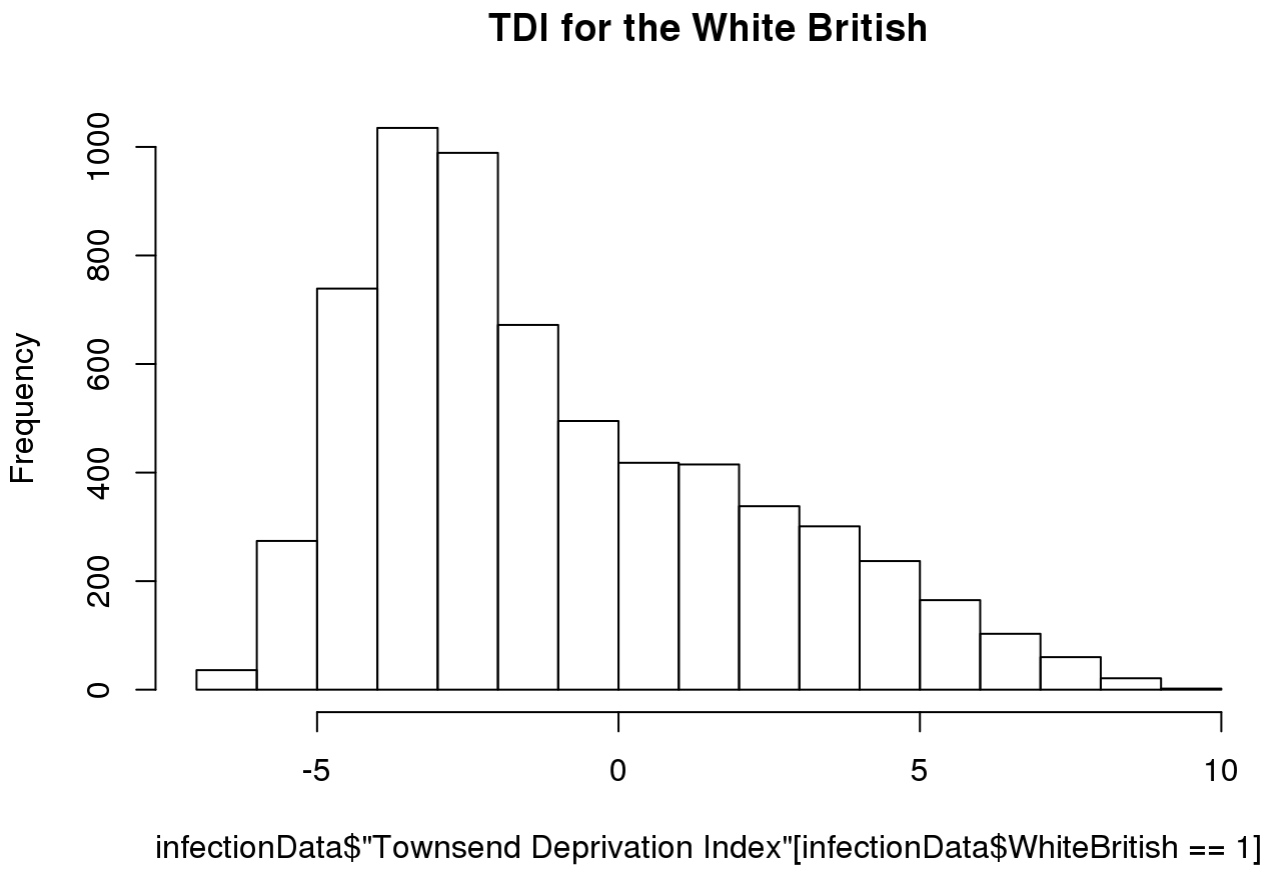
```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3557  -0.2134  -0.1814  -0.1438   0.8775
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.3977944  0.0421765   9.432  < 2e-16 ***
## tmp$"Townsend Deprivation Index"  0.0053222  0.0014411   3.693 0.000223 ***
## tmp$Sex                          0.0317055  0.0094874   3.342 0.000836 ***
## tmp$Age                         -0.0024906  0.0005543  -4.494 7.11e-06 ***
## tmp$WhiteBritish                -0.0538447  0.0138193  -3.896 9.85e-05 ***
## tmp$"Sleep Duration"            -0.0024666  0.0035952  -0.686 0.492688
## Insomnia                        -0.0012542  0.0064155  -0.195 0.845008
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1587078)
##
##     Null deviance: 1168.3  on 7291  degrees of freedom
## Residual deviance: 1156.2  on 7285  degrees of freedom
##   (82 observations deleted due to missingness)
## AIC: 7280.5
##
## Number of Fisher Scoring iterations: 2
```

Analysis:

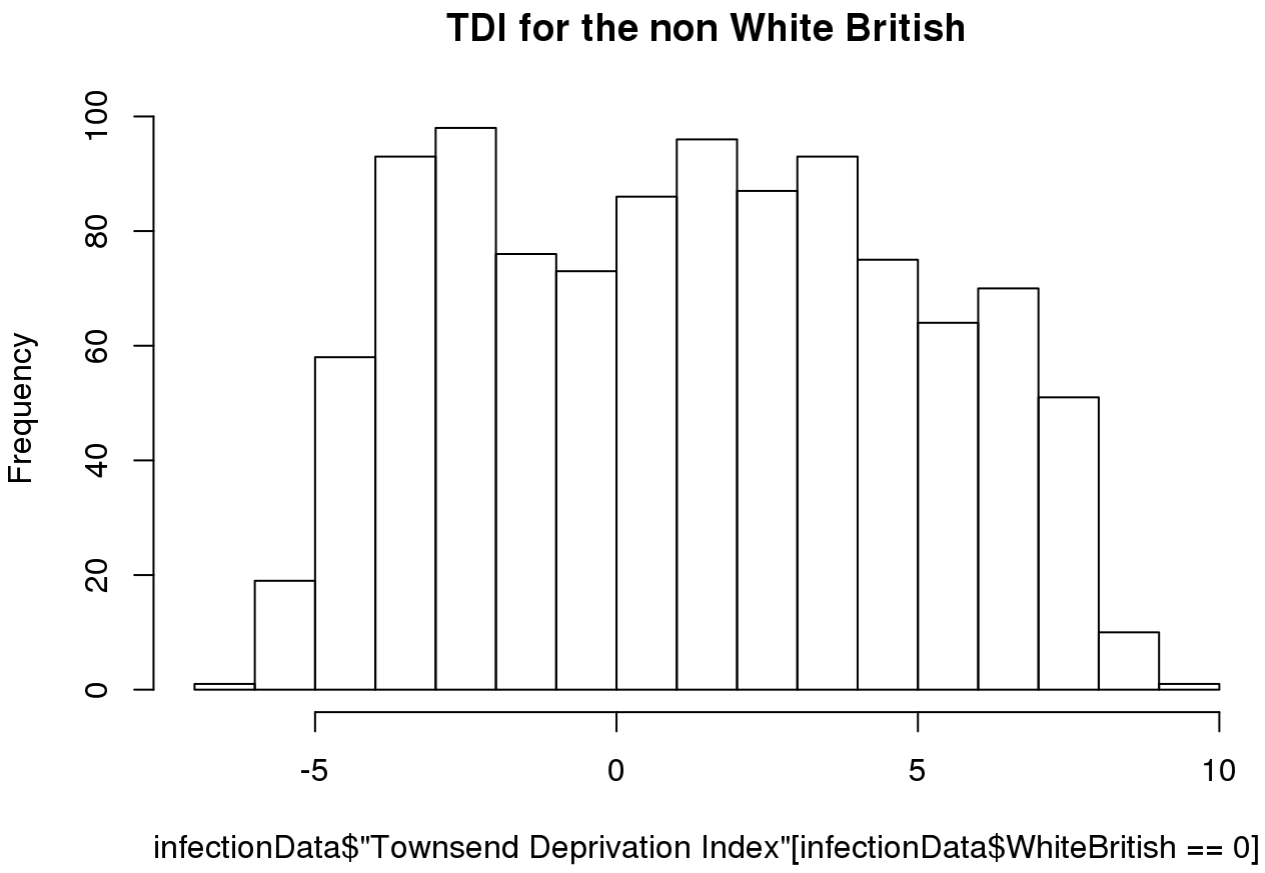
- The Townsend Deprivation Index seems to be a relevant factor for infection.
- All other results are consistent with above models.

Could Townsend Deprivation Index and being White and British be correlated? Let's find out.

```
hist(infectionData$'Townsend Deprivation Index'[infectionData$WhiteBritish==1],main="TDI for t
he White British")
```




```
hist(infectionData$'Townsend Deprivation Index'[infectionData$WhiteBritish==0],main="TDI for t
he non White British")
```



There is a clear distinction here and could explain the higher infection rates among the non White British.

*Death Models:**

The first model looks at a person's exericse habits in comparison to their Covid-19 death rates.

```
tmp <- positiveTestDeathData

fitDeathExercise <- glm(deathInd~tmp$Sex+tmp$Age
                +tmp$'Above Moderate Exercise'+tmp$'Above Moderate Walking Exercise'
                +tmp$'Minutes Walking / Week'+tmp$'Minutes Moderate Activity / Week'
                +tmp$'Minutes Vigorous Activity / Week'+tmp$'Minutes Total Activity / Week'
                +tmp$'Townsend Deprivation Index'
                +tmp$WhiteBritish
                ,family=binomial(),data=tmp)
summary(fitDeathExercise)
```

```
##
## Call:
## glm(formula = deathInd ~ tmp$Sex + tmp$Age + tmp$"Above Moderate Exercise" +
##     tmp$"Above Moderate Walking Exercise" + tmp$"Minutes Walking / Week" +
##     tmp$"Minutes Moderate Activity / Week" + tmp$"Minutes Vigorous Activity / Week" +
##     tmp$"Minutes Total Activity / Week" + tmp$"Townsend Deprivation Index" +
##     tmp$WhiteBritish, family = binomial(), data = tmp)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.1668  -0.6311  -0.3677  -0.2121   2.9118
##
## Coefficients: (1 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -7.981e+00  8.475e-01  -9.417  < 2e-16
## tmp$Sex                              5.226e-01  1.972e-01   2.650  0.00805
## tmp$Age                              1.098e-01  1.364e-02   8.049 8.34e-16
## tmp$"Above Moderate Exercise"       -3.222e-01  2.779e-01  -1.159  0.24642
## tmp$"Above Moderate Walking Exercise"  -4.258e-01  2.648e-01  -1.608  0.10781
## tmp$"Minutes Walking / Week"         4.212e-05  1.023e-04   0.412  0.68056
## tmp$"Minutes Moderate Activity / Week" -2.431e-05  1.054e-04  -0.231  0.81757
## tmp$"Minutes Vigorous Activity / Week"  2.660e-05  7.562e-05   0.352  0.72503
## tmp$"Minutes Total Activity / Week"            NA         NA      NA       NA
## tmp$"Townsend Deprivation Index"     2.332e-02  2.683e-02   0.869  0.38483
## tmp$WhiteBritish                    -3.034e-01  2.513e-01  -1.207  0.22734
##
## (Intercept)                         ***
## tmp$Sex                             **
## tmp$Age                             ***
## tmp$"Above Moderate Exercise"
## tmp$"Above Moderate Walking Exercise"
## tmp$"Minutes Walking / Week"
## tmp$"Minutes Moderate Activity / Week"
## tmp$"Minutes Vigorous Activity / Week"
## tmp$"Minutes Total Activity / Week"
## tmp$"Townsend Deprivation Index"
## tmp$WhiteBritish
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 900.56  on 1119  degrees of freedom
## Residual deviance: 785.45  on 1110  degrees of freedom
##   (365 observations deleted due to missingness)
## AIC: 805.45
##
## Number of Fisher Scoring iterations: 6
```

Analysis:

- Being male increases your chances of dying.
- As expected, the older you are, the higher your chances of dying.
- All exercise variables were irrelevant. Maybe an amalgamation of them could have an influence to get a greater picture on whether a person lives a healthy lifestyle or not. For now it is unclear.
- TDI does not affect a person's death rate.

The following death model inspects the influence of sleep habits.

```
fitDeathSleep <- glm(deathInd~tmp$Sex+tmp$Age
                       +tmp$WhiteBritish
```

```
                      +tmp$'Sleep Duration'+Insomnia
                         ,family=binomial(),data=tmp)

summary(fitDeathSleep)
```

```
##
## Call:
## glm(formula = deathInd ~ tmp$Sex + tmp$Age + tmp$WhiteBritish +
##     tmp$"Sleep Duration" + Insomnia, family = binomial(), data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0229  -0.6263  -0.3778  -0.2402   2.7714
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -8.30136    0.83266  -9.970  < 2e-16 ***
## tmp$Sex               0.57534    0.17260   3.333 0.000858 ***
## tmp$Age               0.09744    0.01164   8.373  < 2e-16 ***
## tmp$WhiteBritish     -0.22297    0.22002  -1.013 0.310851
## tmp$"Sleep Duration"  0.03126    0.06001   0.521 0.602484
## Insomnia              0.11555    0.11092   1.042 0.297549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1133.4  on 1461  degrees of freedom
## Residual deviance: 1016.1  on 1456  degrees of freedom
##   (23 observations deleted due to missingness)
## AIC: 1028.1
##
## Number of Fisher Scoring iterations: 5
```

Analysis:

- Whether you are a bad sleeper does not seem to affect your death likelihood.

The last model looks into the different immune system aspects.

```
cellprops <- cbind(tmp$NumBasophills,tmp$NumEosinophills,tmp$NumLymphocytes,
                         tmp$NumMonocytes,tmp$NumNeutrophills)
cellprops <- cellprops/apply(cellprops,1,sum)

fitDeathImmuneSystem <- glm(deathInd~tmp$Sex+tmp$Age
                            +tmp$NumWhiteBloodCells+cellprops[,1]+cellprops[,2] +
                              cellprops[,3]+cellprops[,4]+tmp$NumPlatelets+tmp$NumReticulocyte
s
                            ,family=binomial(),data=tmp)

summary(fitDeathImmuneSystem)
```

```
##
```

```
## Call:
## glm(formula = deathInd ~ tmp$Sex + tmp$Age + tmp$NumWhiteBloodCells +
##     cellprops[, 1] + cellprops[, 2] + cellprops[, 3] + cellprops[,
##     4] + tmp$NumPlatelets + tmp$NumReticulocytes, family = binomial(),
##     data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1604  -0.6202  -0.3796  -0.2331   2.7182
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -9.091720   0.994285  -9.144  < 2e-16 ***
## tmp$Sex                 0.660559   0.184111   3.588 0.000333 ***
## tmp$Age                 0.097827   0.011911   8.213  < 2e-16 ***
## tmp$NumWhiteBloodCells  0.012510   0.018633   0.671 0.501973
## cellprops[, 1]         18.002686  15.647330   1.151 0.249927
## cellprops[, 2]         -4.966619   5.054891  -0.983 0.325835
## cellprops[, 3]         -0.351798   1.011443  -0.348 0.727977
## cellprops[, 4]          2.604173   1.693057   1.538 0.124012
## tmp$NumPlatelets        0.002502   0.001343   1.862 0.062559 .
## tmp$NumReticulocytes    3.604307   2.536481   1.421 0.155320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1093.23  on 1399  degrees of freedom
## Residual deviance:  972.57  on 1390  degrees of freedom
##   (85 observations deleted due to missingness)
## AIC: 992.57
##
## Number of Fisher Scoring iterations: 5
```

Analysis:

- Only the number of platelets seems to almost be relevant to survival. With a positive correlation and a p value of 0.062, people with more platelets could be at a higher risk of death. Platelets functions are to form a sort of tape over injuries to prevent blood escaping out the injury. Not sure how this could affect the disease's attack on your body, maybe a deeper understanding of the body's response could explain this curiosity.

GWAS Polygenic Scores

To compare the sleep results with something that might be more accurate, I'll analyse a gwas on sleep duration and which genes affect it.

```
sleepGWAS <- read.csv(file = "/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/efotraits_7aug2020.csv",
                      as.is=TRUE)

gwasdata <- sleepGWAS
snpname <- rep(NA, nrow(gwasdata))
referencesnp <- rep(NA, nrow(gwasdata))
beta <- rep(NA, nrow(gwasdata))
```

```
for (i in 1:nrow(gwasdata)){
  snpnametemp <- strsplit(gwasdata[i,1],"-")

  string.split2 <- strsplit(gwasdata[i,1],">")

  allele <- strsplit(string.split2[[1]][2],"<")

  snpname[i] <- snpnametemp[[1]][1]

  referencesnp[i] <- allele[[1]][1]

  beta.split <- strsplit(gwasdata[i,6]," ")

  beta[i] <- beta.split[[1]][1]

}

setwd("/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID")
sleepgwas2plink <- data.frame(snpname=snpname,refallele=referencesnp,beta=as.numeric(beta),str
ingsAsFactors = FALSE)
```

```
## Warning in data.frame(snpname = snpname, refallele = referencesnp, beta =
## as.numeric(beta), : NAs introduced by coercion
```

```
sleepgwas2plink2 <- sleepgwas2plink[-which(is.na(sleepgwas2plink$beta)),]

# go back and get complete data
sleepgwas2plink3 <- sleepgwas2plink2[-which(sleepgwas2plink2$refallele=="?"),]

uniqueSnps <- unique(sleepgwas2plink3$snpname)
avgbeta.unique <-rep(NA, length(uniqueSnps))
refallel.unique<- rep(NA, length(uniqueSnps))

for (i in (1:length(uniqueSnps))){
  row <- sleepgwas2plink3[sleepgwas2plink3$snpname==uniqueSnps[i],]
  alleletable <- table(row$refallele)
  if (length(alleletable)==1) {
    avgbeta <- mean(row$beta,na.rm=TRUE)
    avgbeta.unique[i] <- avgbeta
    refallel.unique[i] <- row$refallele[1]
  }
  if (length(alleletable)>1){
    avgbeta <- (row$beta[1] - row$beta[2])/2
    avgbeta.unique[i] <- avgbeta
    refallel.unique[i] <- row$refallele[1]

  }
}

finalsleepgwas <- data.frame(uniqueSnps,refallel.unique,avgbeta.unique)
```

```
write.table(finalsleepgwas,file="/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/sleepgwas2plink.t
xt",
            col.names=FALSE,row.names=FALSE, quote=FALSE)
```

Calculate total gwas scores for each patient:

```
# Sleep gwas scores

path1 <- "/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/chr.profiles"
filelist <- list.files(path1)
file1 <- read.table(paste0(path1,"/chr1.txt.profile"),header=TRUE)
scores <- data.frame(id=file1[,1],scores=0)
for (j in 1:length(filelist)) {
  if(substr(filelist[j],nchar(filelist[j])-2,nchar(filelist[j]))=="ile") {
    filej <- read.table(paste0(path1,"/",filelist[j]),header=TRUE)
    scores[,2]<- scores[,2]+filej[,6]
  }
}

save(scores,file="/mnt/GREENWOOD_SCRATCH/liam.marengere/COVID/GWASscores")
```

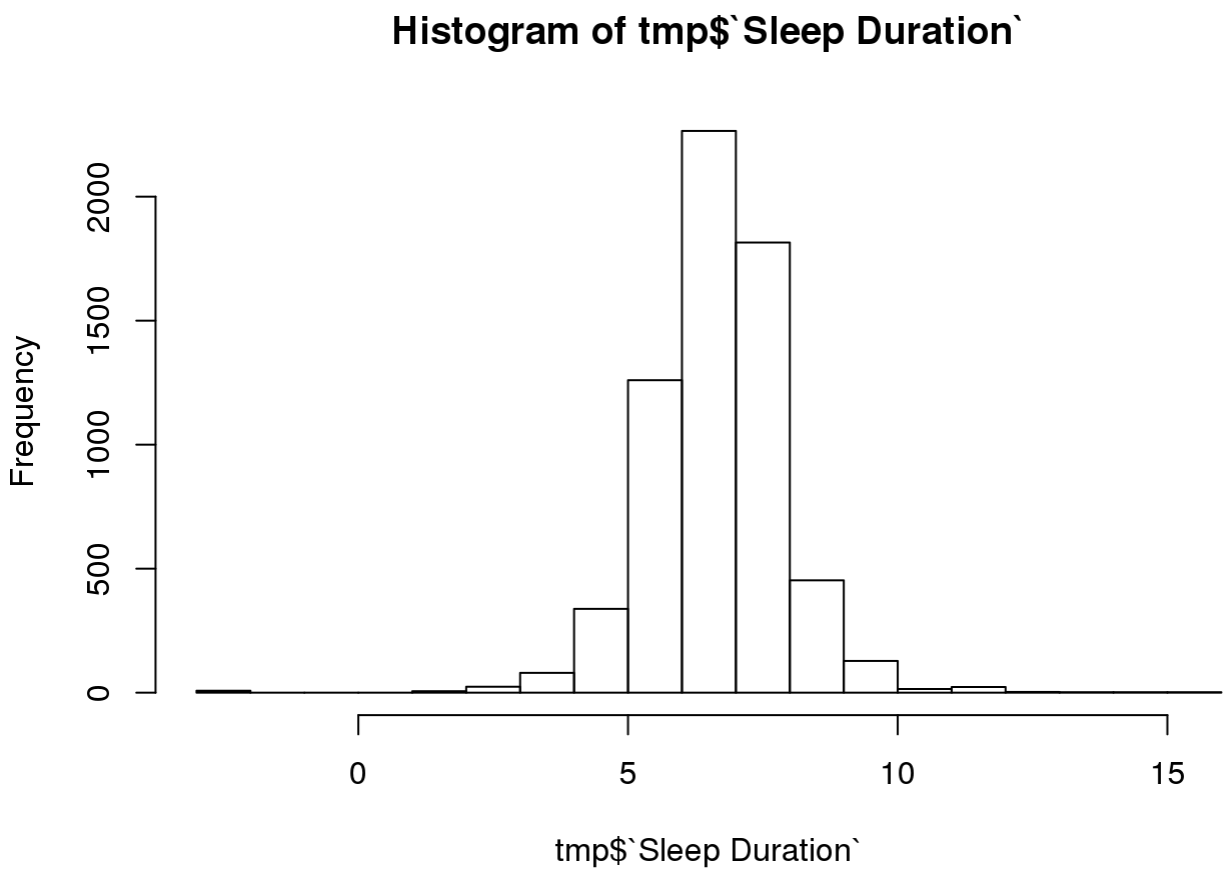
```
globalDataGwassed <- merge(globalData,scores,by.x=1,by.y=1)
covidDataGwassed <- merge(globalDataGwassed,condensed_covid,by.x=1,by.y=1)
infectionGwassed <- merge(infectionData, scores, by.x=1,by.y=1)
deathGwassed <- merge(positiveTestDeathData, scores, by.x=1, by.y=1)
```

Check if the sleep scores are associated:

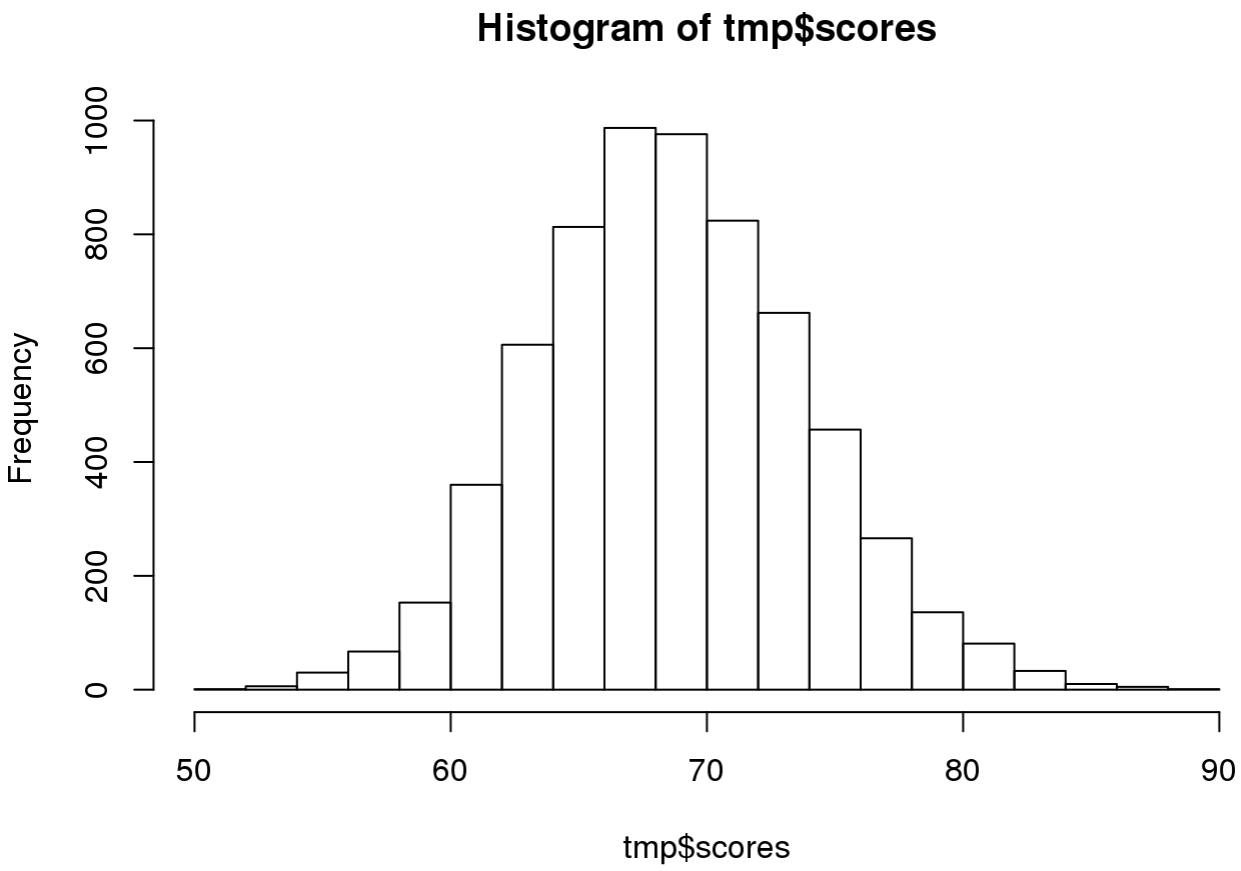
```
tmp <- infectionGwassed
fitSleepDurationVsScores <- lm(tmp$`Sleep Duration`~tmp$scores,data=tmp)

summary(fitSleepDurationVsScores)
```

```
##
## Call:
## lm(formula = tmp$`Sleep Duration` ~ tmp$scores, data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1980  -1.0772  -0.1252   0.8454   8.8097
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.448616   0.212090  30.405  < 2e-16 ***
## tmp$scores  0.010221   0.003081   3.317 0.000916 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.284 on 6422 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.00171,    Adjusted R-squared:  0.001555
## F-statistic:    11 on 1 and 6422 DF,  p-value: 0.0009157
```

```
hist(tmp$`Sleep Duration`)
```




```
hist(tmp$scores)
```



Analysis: - Sleep duration and the scores are highly correlated. - 1 unit of sleep scores corresponds to 6 minutes in sleep time. From the above histogram, we can see how big of a difference this can have on the individual's sleep patterns.

Modelling with the polygenic scores

Infection Model:

```
tmp <- infectionGwassed
fitInfectionGwasSleep <- glm(result~tmp$Sex+tmp$Age+tmp$WhiteBritish
                             +tmp$scores+tmp$`Townsend Deprivation Index`
                             ,family="binomial",data=tmp)

summary(fitInfectionGwasSleep)
```

```
##
## Call:
## glm(formula = result ~ tmp$Sex + tmp$Age + tmp$WhiteBritish +
##     tmp$scores + tmp$`Townsend Deprivation Index`, family = "binomial",
##     data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8771  -0.6762  -0.6305  -0.5758   2.0163
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                      -1.473816   0.482162  -3.057 0.002238 **
## tmp$Sex                           0.188617   0.063894   2.952 0.003157 **
## tmp$Age                          -0.013354   0.003698  -3.611 0.000305 ***
## tmp$WhiteBritish                 -0.113444   0.134508  -0.843 0.399007
## tmp$scores                        0.012412   0.006052   2.051 0.040290 *
## tmp$`Townsend Deprivation Index`  0.031770   0.009675   3.284 0.001024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6348.3  on 6462  degrees of freedom
## Residual deviance: 6311.5  on 6457  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 6323.5
##
## Number of Fisher Scoring iterations: 4
```

```
tmp <- infectionGwassed
fitInfectionGwasSleep <- glm(result~tmp$Sex+tmp$Age+tmp$WhiteBritish
                           +tmp$scores+tmp$NumBasophills+tmp$`Townsend Deprivation Index`
                           ,family="binomial",data=tmp)

summary(fitInfectionGwasSleep)
```

```
##
## Call:
## glm(formula = result ~ tmp$Sex + tmp$Age + tmp$WhiteBritish +
##     tmp$scores + tmp$NumBasophills + tmp$`Townsend Deprivation Index`,
##     family = "binomial", data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8729  -0.6744  -0.6305  -0.5704   2.0954
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.299464   0.492301  -2.640 0.008301 **
## tmp$Sex                           0.200776   0.065116   3.083 0.002047 **
## tmp$Age                          -0.012508   0.003774  -3.315 0.000918 ***
## tmp$WhiteBritish                 -0.139137   0.136553  -1.019 0.308240
## tmp$scores                        0.010300   0.006178   1.667 0.095444 .
## tmp$NumBasophills                -1.841859   0.756715  -2.434 0.014932 *
## tmp$`Townsend Deprivation Index`  0.031940   0.009861   3.239 0.001199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6132.1  on 6258  degrees of freedom
## Residual deviance: 6091.9  on 6252  degrees of freedom
##   (215 observations deleted due to missingness)
## AIC: 6105.9
```

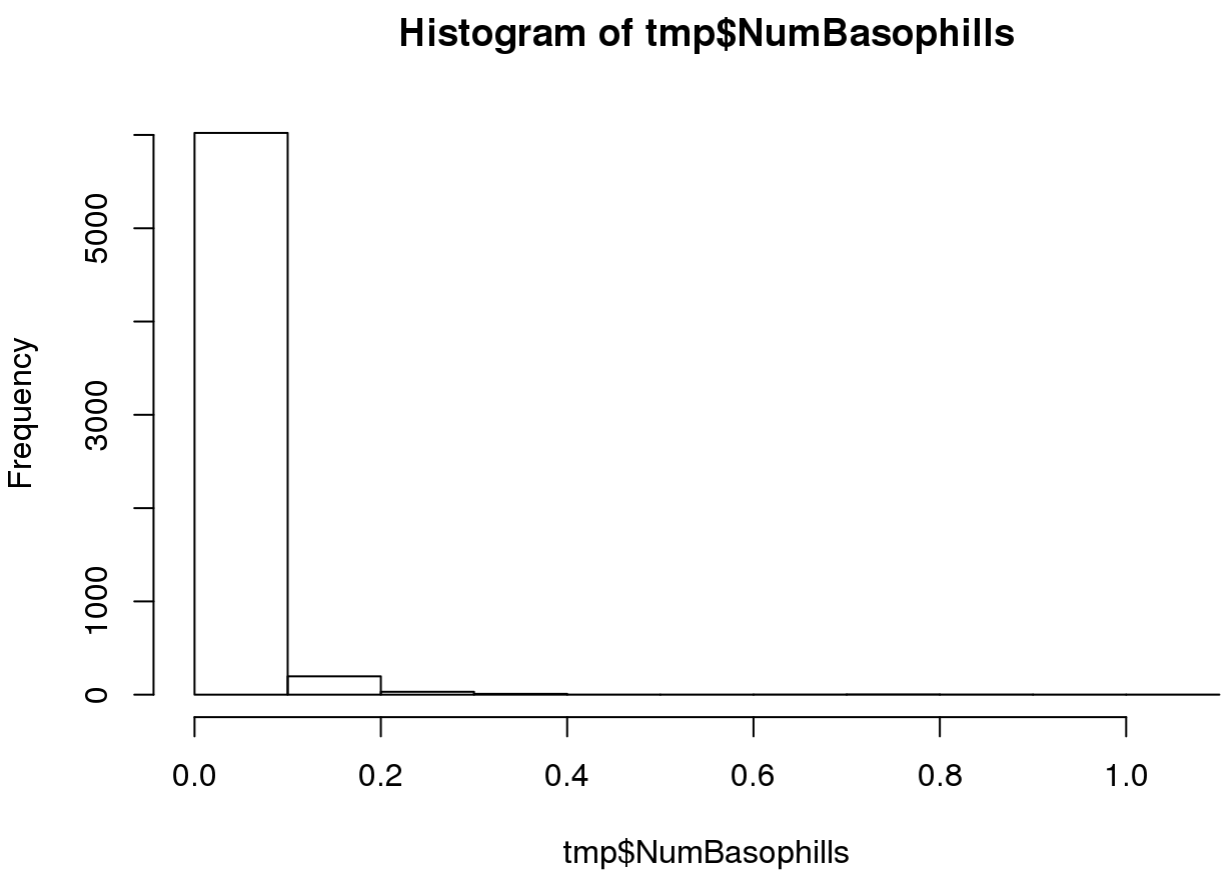
```
##
## Number of Fisher Scoring iterations: 4
```

Analysis:

- Adding the number of basophills into the model made the scores less relevant. Could this be because they are correlated? Basophills are more relevant to infection in this model, confirming earlier results.

Now to investigate the basophill to score correlation:

```
hist(tmp$NumBasophills)
```




```
tmp <- infectionGwassed

fitInfectionBasophils <- lm(tmp$NumBasophills~tmp$scores+tmp$WhiteBritish+tmp$Age+tmp$Sex+tmp$
`Townsend Deprivation Index`)

summary(fitInfectionBasophils)
```

```
##
## Call:
## lm(formula = tmp$NumBasophills ~ tmp$scores + tmp$WhiteBritish +
##     tmp$Age + tmp$Sex + tmp$`Townsend Deprivation Index`)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.05101 -0.02994 -0.01216  0.00983  1.00354
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.571e-02  9.994e-03   5.574  2.6e-08 ***
## tmp$scores                     -1.249e-04  1.251e-04  -0.998  0.31812
## tmp$WhiteBritish               -6.018e-03  2.879e-03  -2.090  0.03662 *
## tmp$Age                        -4.286e-05  7.715e-05  -0.556  0.57852
## tmp$Sex                        -2.603e-03  1.312e-03  -1.984  0.04731 *
## tmp$`Townsend Deprivation Index`  6.764e-04  2.034e-04   3.325  0.00089 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05139 on 6253 degrees of freedom
##   (215 observations deleted due to missingness)
## Multiple R-squared:  0.003582,   Adjusted R-squared:  0.002785
## F-statistic: 4.496 on 5 and 6253 DF,  p-value: 0.0004311
```

```
fitLogInfectionBasophils <- lm(log(tmp$NumBasophills+1)~tmp$scores+tmp$WhiteBritish+tmp$Age+tm
p$Sex+tmp$`Townsend Deprivation Index`)
summary(fitLogInfectionBasophils)
```

```
##
## Call:
## lm(formula = log(tmp$NumBasophills + 1) ~ tmp$scores + tmp$WhiteBritish +
##     tmp$Age + tmp$Sex + tmp$`Townsend Deprivation Index`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.04782 -0.02787 -0.01071  0.01076  0.67812
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.166e-02  8.530e-03   6.056 1.47e-09 ***
## tmp$scores                     -1.045e-04  1.068e-04  -0.978 0.328159
## tmp$WhiteBritish               -5.602e-03  2.457e-03  -2.280 0.022630 *
## tmp$Age                        -3.814e-05  6.585e-05  -0.579 0.562496
## tmp$Sex                        -2.235e-03  1.120e-03  -1.996 0.046019 *
## tmp$`Townsend Deprivation Index`  5.945e-04  1.736e-04   3.424 0.000621 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04386 on 6253 degrees of freedom
##   (215 observations deleted due to missingness)
## Multiple R-squared:  0.003854,   Adjusted R-squared:  0.003057
## F-statistic: 4.838 on 5 and 6253 DF,  p-value: 0.000203
```

Analysis:

- The skewed data does not seem like it influenced the results. However, this seems to confirm that the polygenic scores are mostly independent of the number of basophills.

The TDI is consistently relevant to the number of basophills.

Death Model:

```
tmp <- deathGwassed
fitDeathGwasSleep <- glm(tmp$deathInd~tmp$Sex+tmp$Age+tmp$WhiteBritish
                         +tmp$scores+tmp$NumPlatelets+tmp$`Townsend Deprivation Index`
                         ,family="binomial",data=tmp)

summary(fitDeathGwasSleep)
```

```
##
## Call:
## glm(formula = tmp$deathInd ~ tmp$Sex + tmp$Age + tmp$WhiteBritish +
##     tmp$scores + tmp$NumPlatelets + tmp$`Townsend Deprivation Index`,
##     family = "binomial", data = tmp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0697  -0.6197  -0.3945  -0.2301   2.7324
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.772489   1.525189  -5.096 3.47e-07 ***
## tmp$Sex                         0.637100   0.195626   3.257  0.00113 **
## tmp$Age                         0.101998   0.013279   7.681 1.58e-14 ***
## tmp$WhiteBritish                0.031584   0.381772   0.083  0.93407
## tmp$scores                     -0.020272   0.017272  -1.174  0.24052
## tmp$NumPlatelets                0.003046   0.001424   2.138  0.03249 *
## tmp$`Townsend Deprivation Index`  0.034308   0.026196   1.310  0.19031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 941.78  on 1210  degrees of freedom
## Residual deviance: 841.69  on 1204  degrees of freedom
##   (41 observations deleted due to missingness)
## AIC: 855.69
##
## Number of Fisher Scoring iterations: 6
```

Now to investigate connections to a person's number of platelets.

```
fitPlateletInfluences <- glm(tmp$NumPlatelets~tmp$Sex+tmp$Age+tmp$WhiteBritish+tmp$`Townsend D
eprivation Index`+tmp$scores)
summary(fitPlateletInfluences)
```

```
##
## Call:
## glm(formula = tmp$NumPlatelets ~ tmp$Sex + tmp$Age + tmp$WhiteBritish +
##     tmp$`Townsend Deprivation Index` + tmp$scores)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -200.52   -39.50    -7.04    30.49   349.75
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    300.94725   25.48598  11.808  < 2e-16 ***
## tmp$Sex                        -28.73578    3.50594  -8.196 6.28e-16 ***
## tmp$Age                         -0.58842    0.19240  -3.058  0.00228 **
## tmp$WhiteBritish                -0.80910    7.18198  -0.113  0.91032
## tmp$`Townsend Deprivation Index`  0.81650    0.52736   1.548  0.12182
## tmp$scores                      -0.01573    0.32804  -0.048  0.96176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 3575.339)
##
##     Null deviance: 4632671  on 1210  degrees of freedom
## Residual deviance: 4308283  on 1205  degrees of freedom
##   (41 observations deleted due to missingness)
## AIC: 13353
##
## Number of Fisher Scoring iterations: 2
```

Analysis:
- Being male and being older decrease your number of platelets. There must be another factor that increases platelets as all of these factors increase likelihood of death.

Conclusions and Further Questions
- What and how increases basophill count, or other parameters of the immune system?
- How does the TDI increase basophill count?
- If the scores do affect the infection rate independently from the basophill count, how? And are they truly independent?

- For death, the scores did not show any association. However an increased number of platelets did.
- If being male and older decrease your platelet count, what leads to them dying more frequently, and what does increase platelet count?