Initial EDA Project

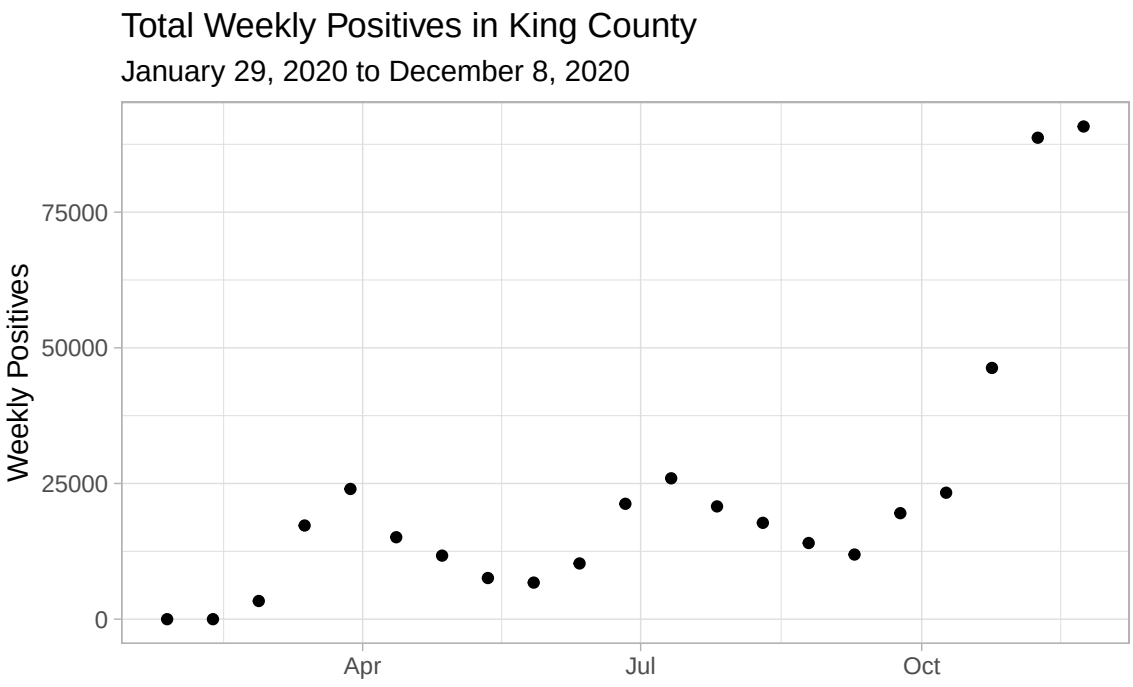
Ericka Smith

11/22/2020

```
covid_counts <- read_xlsx("../data/biweekly-counts-rates-by-geography-dec-8.xlsx",
                    col_types = c("text", "date", "date",
                                  "numeric", "text", "numeric", "text",
                                  "numeric", "text", "numeric", "text",
                                  "numeric", "numeric"))
covid_demos <- read_xlsx("../data/total-counts-by-date-city-demography-dec-8.xlsx")
```

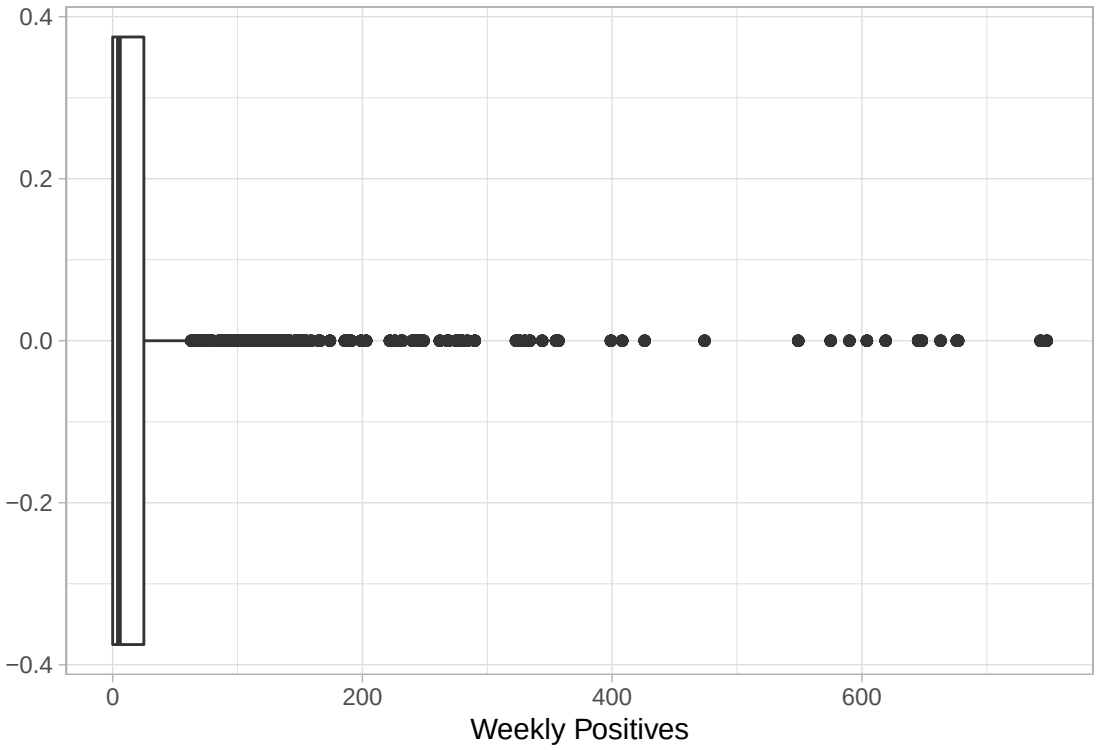
```
covid_counts <- covid_counts[, c("City", "Week_Start", "People_Tested_Rate", "Positives", "Deaths")]
covid_counts_clean <- covid_counts %>%
  mutate(City = factor(City),
         People_Tested_Rate = as.numeric(People_Tested_Rate)/100000)

covid_demos <- covid_demos[, c("Age_Group", "City", "Population")]
covid_demos_clean <- covid_demos %>%
  filter(City != "All King County") %>%
  mutate(City = factor(City),
         Age_Group = factor(Age_Group))

covid <- covid_counts_clean %>%
  left_join(covid_demos_clean, by = "City")
```

```
covid %>%
  group_by(Week_Start) %>%
  summarise(Total_Positives = sum(Positives)) %>%
ggplot() +
  geom_point(aes(Week_Start, Total_Positives)) +
  labs(title= "Total Weekly Positives in King County",
       subtitle = "January 29, 2020 to December 8, 2020",
       caption= "Source: Washington State Department of Health",
       x="",
       y="Weekly Positives")+
  theme_light()+
  theme(legend.position = "none")
```



Source: Washington State Department of Health

```
covid %>%
  filter(Positives <=750) %>%
ggplot() +
  geom_boxplot(aes(Positives)) +
  labs(y="",
       x="Weekly Positives")+
  theme_light()+
  theme(legend.position = "none")
```




```
# pois_mod <- glm(Positives~Week_Start+People_Tested_Rate+Deaths+Death_Rate+Positive_Rate, data=covid,
pois_mod <- glm(Positives~.-City, data=covid, family="poisson")

summary(pois_mod)
```

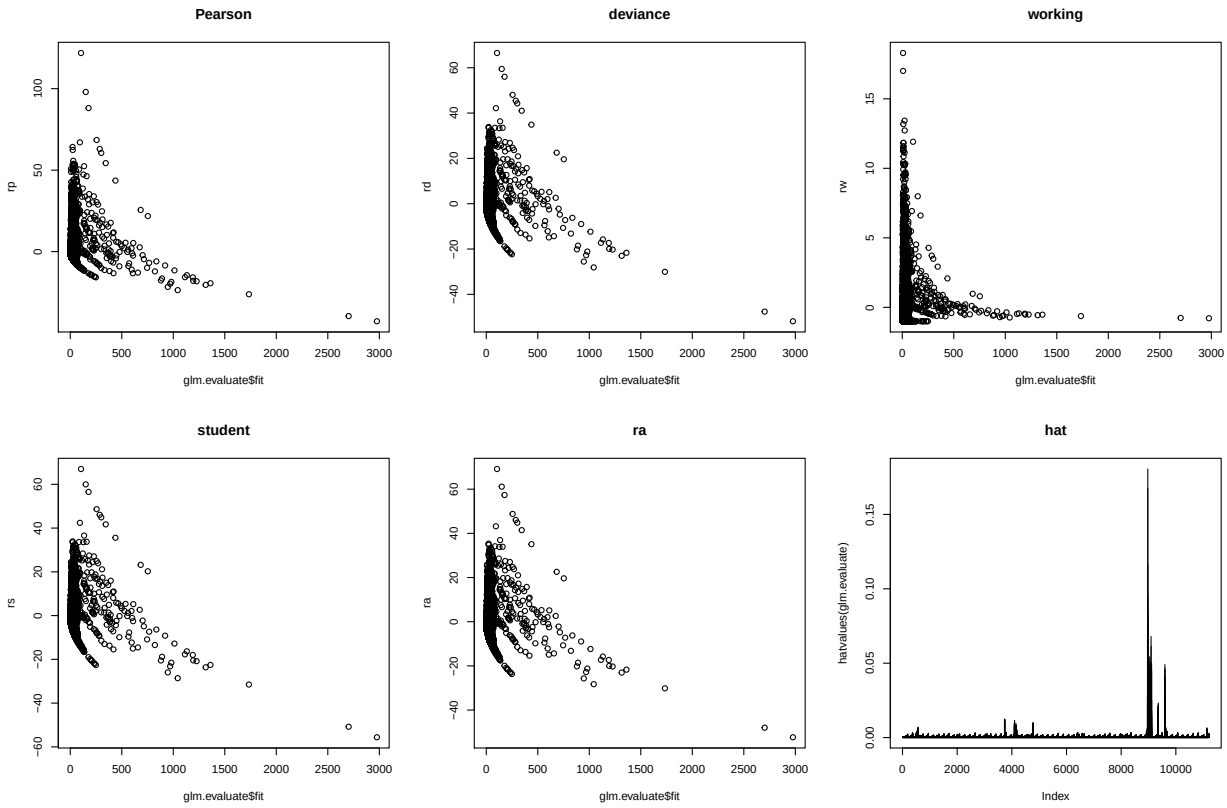
```
##
## Call:
## glm(formula = Positives ~ . - City, family = "poisson", data = covid)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -51.868   -4.563   -3.092    0.005   66.501
##
## Coefficients:
##                       Estimate Std. Error  z value Pr(>|z|)
## (Intercept)        -9.092e+01  6.070e-01 -149.781  < 2e-16 ***
## Week_Start          5.849e-08  3.814e-10  153.362  < 2e-16 ***
## People_Tested_Rate  4.443e+01  2.246e-01  197.812  < 2e-16 ***
## Deaths              1.071e-01  3.029e-04  353.554  < 2e-16 ***
## Age_Group10-19      3.211e-02  8.210e-03    3.911  9.2e-05 ***
## Age_Group20-29     -4.712e-01  8.754e-03  -53.826  < 2e-16 ***
## Age_Group30-39     -4.002e-01  8.582e-03  -46.630  < 2e-16 ***
## Age_Group40-49     -1.356e-01  8.245e-03  -16.445  < 2e-16 ***
## Age_Group50-59     -4.902e-02  8.213e-03   -5.968  2.4e-09 ***
## Age_Group60-69      4.482e-03  8.209e-03    0.546    0.585
## Age_Group70-79      1.271e-01  8.226e-03   15.452  < 2e-16 ***
## Age_Group80+        1.575e-01  8.233e-03   19.125  < 2e-16 ***
## Age_GroupUnknown    2.061e-01  8.246e-03   24.996  < 2e-16 ***
```

```
## Population          1.821e-05  6.380e-08  285.440  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 888944  on 11209  degrees of freedom
## Residual deviance: 442993  on 11196  degrees of freedom
##   (1180 observations deleted due to missingness)
## AIC: 477221
##
## Number of Fisher Scoring iterations: 7
```

```
Anova(pois_mod,
      type="II",
      test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Positives
##                     LR Chisq Df Pr(>Chisq)
## Week_Start            25314  1  < 2.2e-16 ***
## People_Tested_Rate    29481  1  < 2.2e-16 ***
## Deaths                95635  1  < 2.2e-16 ***
## Age_Group             10142  9  < 2.2e-16 ***
## Population            72909  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glm.evaluate=pois_mod

rp=resid(glm.evaluate, "pearson")
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```



Since residual deviance > df, this model is overdispersed. On to nb or quasipoisson

```
#nb_mod <- glm.nb(Positives~Week_Start+People_Tested_Rate, data=covid)
nb_mod <- glm.nb(Positives~.-City, data = covid, control = glm.control(maxit=75))
summary(nb_mod)
```

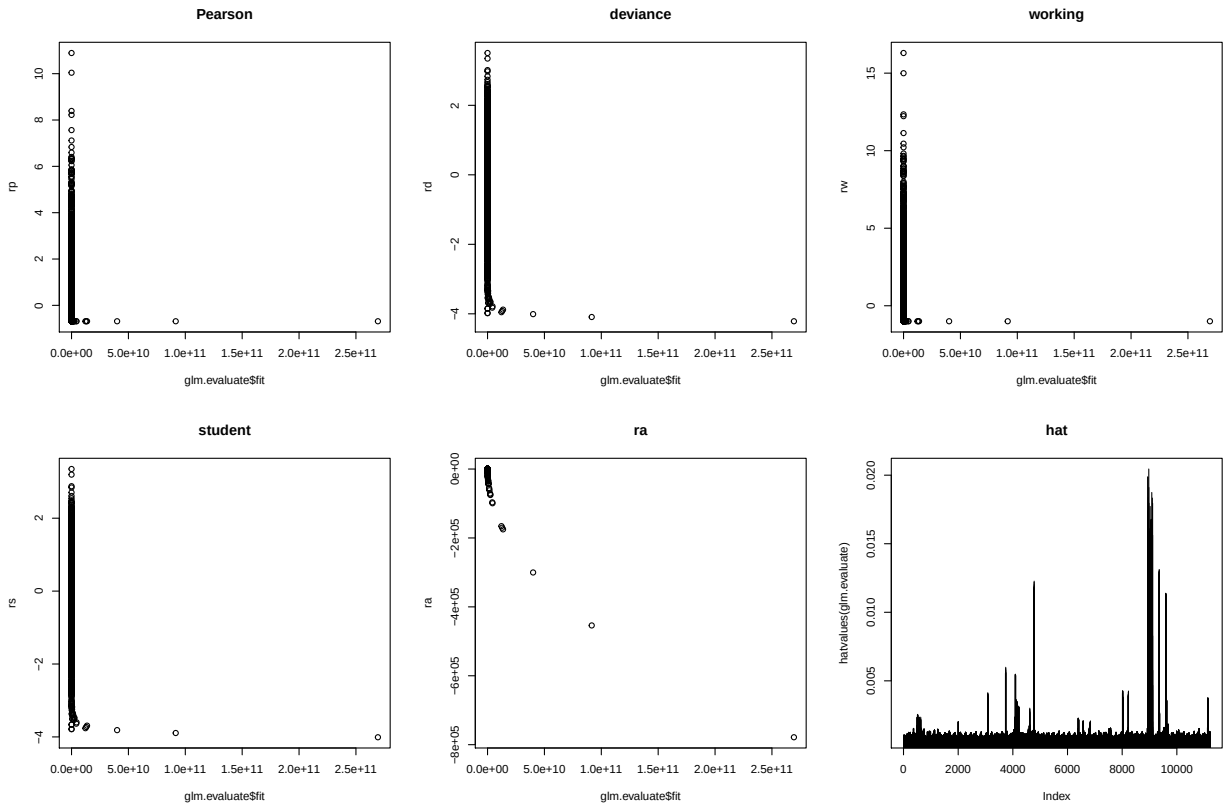
```
##
## Call:
## glm.nb(formula = Positives ~ . - City, data = covid, control = glm.control(maxit = 75),
##     init.theta = 0.4710586162, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2141  -1.2416  -0.5199   0.2161   3.5054
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.271e+01  4.825e+00  -8.852  < 2e-16 ***
## Week_Start          2.723e-08  3.043e-09   8.948  < 2e-16 ***
## People_Tested_Rate  9.067e+01  2.491e+00  36.399  < 2e-16 ***
## Deaths              2.781e-01  5.951e-03  46.736  < 2e-16 ***
## Age_Group10-19      6.841e-02  6.420e-02   1.066   0.2866
## Age_Group20-29      1.961e-02  6.430e-02   0.305   0.7604
## Age_Group30-39     -6.530e-02  6.437e-02  -1.015   0.3103
## Age_Group40-49     -7.655e-03  6.425e-02  -0.119   0.9052
## Age_Group50-59      1.414e-02  6.422e-02   0.220   0.8257
## Age_Group60-69      1.536e-01  6.413e-02   2.396   0.0166 *
```

```
## Age_Group70-79      3.942e-01  6.408e-02   6.152 7.67e-10 ***
## Age_Group80+        5.097e-01  6.407e-02   7.956 1.78e-15 ***
## Age_GroupUnknown    7.167e-01  6.410e-02  11.182  < 2e-16 ***
## Population          1.079e-04  1.400e-06  77.076  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.4711) family taken to be 1)
##
##     Null deviance: 21538  on 11209  degrees of freedom
## Residual deviance: 12392  on 11196  degrees of freedom
##   (1180 observations deleted due to missingness)
## AIC: 74905
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.47106
##           Std. Err.:  0.00697
##
##  2 x log-likelihood:  -74874.91100
```

```
Anova(nb_mod,
      type="II",
      test="LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Positives
##                     LR Chisq Df Pr(>Chisq)
## Week_Start            61.62  1   4.17e-15 ***
## People_Tested_Rate   693.17  1  < 2.2e-16 ***
## Deaths              1300.63  1  < 2.2e-16 ***
## Age_Group            258.72  9  < 2.2e-16 ***
## Population          1085.98  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glm.evaluate=nb_mod

rp=resid(glm.evaluate, "pearson")
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```




```
# qp_mod = glm(Positives~Week_Start+People_Tested_Rate, data=covid,
#                family="quasipoisson")
qp_mod = glm(Positives~.-City, data=covid,
             family=quasipoisson(link = "log"))
summary(qp_mod)
```

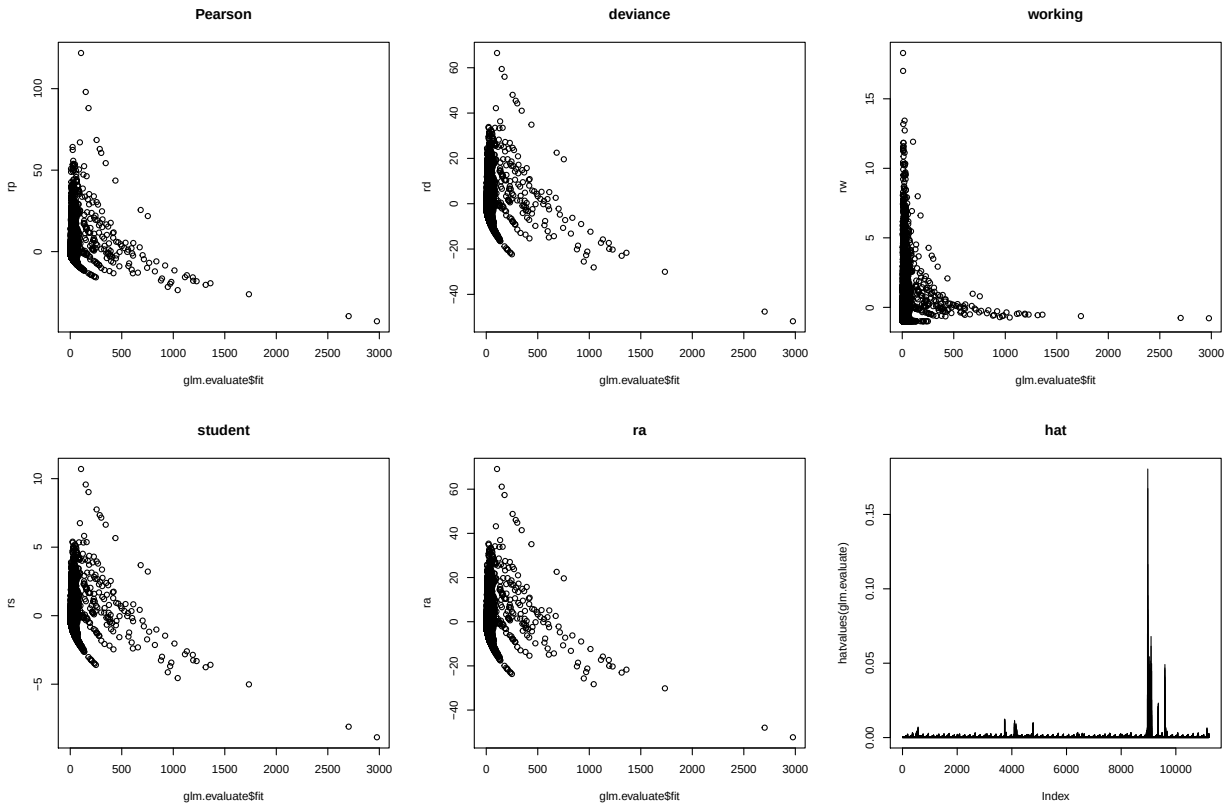
```
##
## Call:
## glm(formula = Positives ~ . - City, family = quasipoisson(link = "log"),
##     data = covid)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -51.868   -4.563   -3.092    0.005   66.501
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -9.092e+01  4.563e+00 -19.924  < 2e-16 ***
## Week_Start          5.849e-08  2.867e-09  20.401  < 2e-16 ***
## People_Tested_Rate  4.443e+01  1.688e+00  26.314  < 2e-16 ***
## Deaths              1.071e-01  2.277e-03  47.031  < 2e-16 ***
## Age_Group10-19      3.211e-02  6.172e-02   0.520 0.602917
## Age_Group20-29     -4.712e-01  6.581e-02  -7.160 8.57e-13 ***
## Age_Group30-39     -4.002e-01  6.451e-02  -6.203 5.74e-10 ***
## Age_Group40-49     -1.356e-01  6.198e-02  -2.188 0.028717 *
## Age_Group50-59     -4.902e-02  6.174e-02  -0.794 0.427264
```

```
## Age_Group60-69      4.482e-03  6.171e-02   0.073 0.942095
## Age_Group70-79      1.271e-01  6.184e-02   2.055 0.039857 *
## Age_Group80+        1.575e-01  6.189e-02   2.544 0.010972 *
## Age_GroupUnknown    2.061e-01  6.199e-02   3.325 0.000887 ***
## Population          1.821e-05  4.796e-07  37.970  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 56.51224)
##
##     Null deviance: 888944  on 11209  degrees of freedom
## Residual deviance: 442993  on 11196  degrees of freedom
##   (1180 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 7
```

```
Anova(qp_mod,
      type="II",
      test="LR")
```

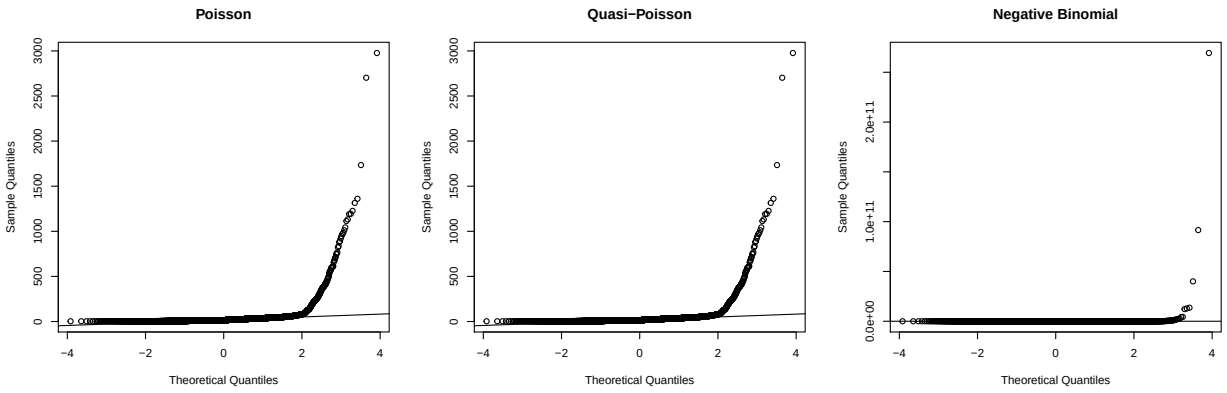
```
## Analysis of Deviance Table (Type II tests)
##
## Response: Positives
##                     LR Chisq Df Pr(>Chisq)
## Week_Start            447.93  1  < 2.2e-16 ***
## People_Tested_Rate    521.67  1  < 2.2e-16 ***
## Deaths               1692.29  1  < 2.2e-16 ***
## Age_Group             179.46  9  < 2.2e-16 ***
## Population           1290.14  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glm.evaluate=qp_mod

rp=resid(glm.evaluate, "pearson")
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```



Based on "How should we model overdispersed count data?" paper suggests that negative binomial may be better for my scenario because it gives smaller values more weight relative to quasi-Poisson and allows smaller values to have a greater effect on adjustments for negative binomial regression

```
par(mfrow=c(1,3))
qqnorm(pois_mod$fitted.values, main="Poisson")
qqline(pois_mod$fitted.values)

qqnorm(qp_mod$fitted.values, main="Quasi-Poisson")
qqline(qp_mod$fitted.values)

qqnorm(nb_mod$fitted.values, main="Negative Binomial")
qqline(nb_mod$fitted.values)
```



```r
outlierTest(pois_mod)
```

```
##        rstudent unadjusted p-value Bonferroni p
## 9913  -55.61178                  0            0
## 9914  -50.80727                  0            0
## 9980   42.39553                  0            0
## 10051  46.06647                  0            0
## 10052  48.63850                  0            0
## 10056  41.68576                  0            0
## 10057  44.90166                  0            0
## 10058  56.54814                  0            0
## 10059  59.96407                  0            0
## 10060  66.99511                  0            0
```

```r
outlierTest(qp_mod)
```

```
##        rstudent unadjusted p-value Bonferroni p
## 10060 10.703981         9.7497e-27   1.0929e-22
## 10059  9.570963         1.0591e-21   1.1873e-17
## 10058  9.021687         1.8521e-19   2.0763e-15
## 9913  -8.873518         7.0871e-19   7.9446e-15
## 9914  -8.101737         5.4180e-16   6.0736e-12
## 10052  7.752509         9.0094e-15   1.0100e-10
## 10051  7.340542         2.1273e-13   2.3847e-09
## 10057  7.154082         8.4235e-13   9.4427e-09
## 9980   6.753218         1.4460e-11   1.6210e-07
## 10056  6.639624         3.1449e-11   3.5254e-07
```

```r
outlierTest(nb_mod)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##       rstudent unadjusted p-value Bonferroni p
## 9913 -4.009378         6.1274e-05      0.68688
```