COVID-19 Cases in King County, Washington

Ericka Smith

Abstract: This report describes the investigation of factors that may help explain positive COVID-19 case count fluctuations from late January to early December of 2020 in King County, Washington. Three primary models were explored: Poisson, quasi-Poisson, and negative binomial. The negative binomial model was determined to be the best fit, though delving further into zero-inflated models and serial autocorrelation is suggested. City population size, rate of people tested, number of deaths, and week start date were all consistently significant (alpha=0.05) predictors in the models. Age group was more complex but the results affirm the idea that older age groups, particularly individuals 70 and above, may be more susceptible to the virus.

I. Introduction

This year has been defined by the novel Coronavirus Disease 2019 (COVID-19). The disease originated in Wuhan, China and the first confirmed case in the United States was in Washington State. Though Washington is no longer the center for the outbreak, the length of time that the disease has been present within the state makes it an interesting case study. The numbers of confirmed cases within the state are predominantly contained in four counties: King, Pierce, Snohomish (the site of the original case), and Yakima. The first three follow along the Interstate 5 corridor and are, respectively, the most populated in the state. The fourth has one of the largest proportions of essential workers in the state (63% as compared to 54% statewide). This analysis focuses on King County in part because it lies in the center of the three highly populated counties, has the most cases to date, and includes Seattle, the largest city in the state. More importantly though, King County was a prime candidate due to the plethora of freely available data on the progression of the virus.

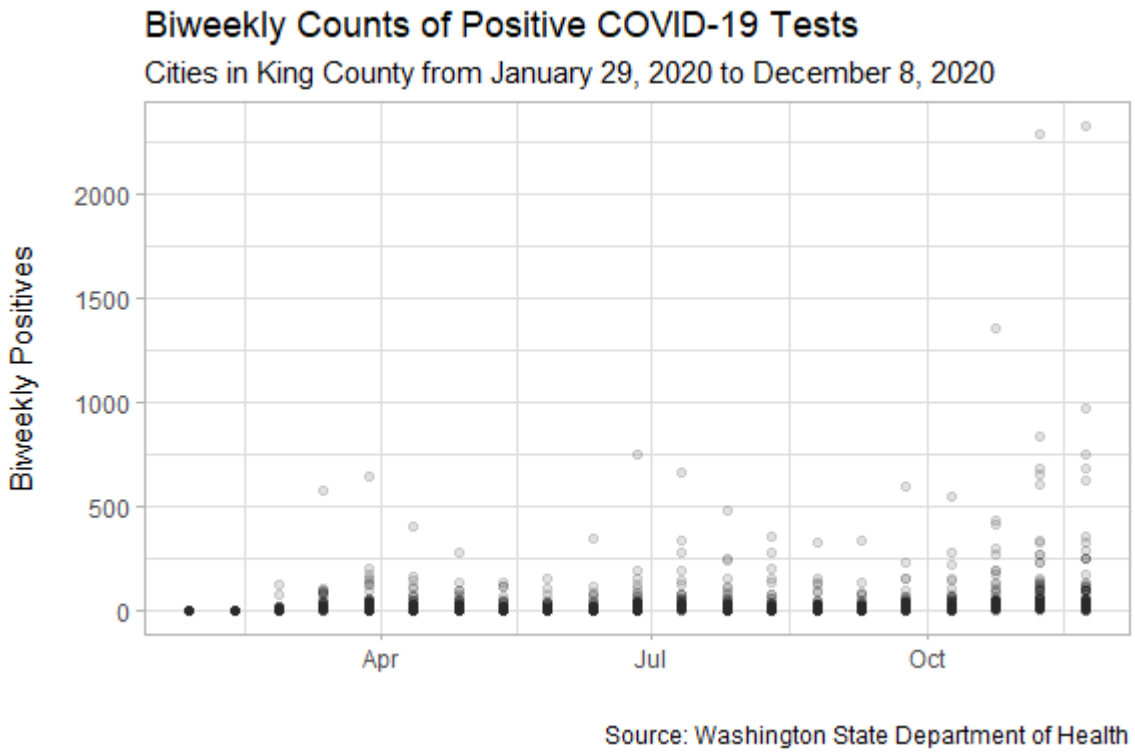


Figure 1. Case Counts.

The goal of this inquiry was to determine what demographic and testing factors may be useful in explaining the variation seen in positive test counts over time in King County, Washington. Therefore it focused on the biweekly count of unique King County residents who have had a confirmed positive COVID-19 PCR laboratory result

reported to the Washington State Department of Health as a response variable. The two central explanatory variables of interest were the date the week started and the rate of the population that was tested. Other variables were considered but ultimately deemed to be unsuitable for this analysis.

II. Analysis

❖ The Data

Two datasets available on the King County COVID-19 outbreak summary website were used. The first dataset includes our variable of interest, positive case count, along with the discrete time period over which these counts happen, biweekly. It also included the rate of people tested. The second dataset was added in order to get an idea of the effect of population and age demographics by city.

Originally the city was to be considered in the model, but as there are 59 unique cities within King County it was quickly determined to be both infeasible and of little practical interest. Instead population and age demographics were matched to the case counts by city and investigated. In addition, these data include deaths/death rate and hospitalizations/hospitalization rate, which were perceived to be responsive to the biweekly counts of positive tests, rather than explanatory, and therefore excluded. All test results/all test results rate are per test rather than per resident, which is in contrast to how positive cases were recorded. Due to the large number of retests these were also removed. Finally, end dates of the weeks and positive rate were excluded due to their high correlation, both practically and statistically, with the count of positive tests and the week start date.

❖ Models Fitted

Three primary regression models were fit in this analysis: Poisson, negative binomial, and quasipoisson. Though some other combinations of explanatory variables were fit, the best models were those in which the following were included: week start date, rate of people tested, count of deaths, age group, and population, so these will be the variables in the models discussed.

➢ Poisson

A Poisson regression model was fit initially. It was determined though that the data do not meet the assumptions of this model. The positive case events cannot be assumed to be independent. In addition, the variance is not equal to the mean. The data is in fact overdispersed, and this can be seen in the output from fitting the model in R (degrees of freedom = 11,786 < residual deviance=580,131). It is interesting to note though that in this model week start date, rate of people tested, number of deaths, population, and every age group except 60-69 were significant at alpha = 0.05 and AIC was 617,768.

➢ Quasi-Poisson

The quasi-poisson model does not assume the variance is equal to the mean, instead choosing it as a linear function of the mean. It also assumes that the variance parameter is the same across all of the events. This could be argued either way due to the aspects of testing that may have changed over time, but overall I believe that we can make this assumption. At alpha=0.05 week start date, rate of people tested, number of deaths, population were significant. Additionally in this model all unknown age groups, those from 20-49, and those 70 and over were significant at alpha=0.05. This model, of course, cannot have AIC, so it is not reported here. The null deviance and residual deviance are the same as that of the Poisson model.

➢ Negative Binomial

The negative binomial model also assumes that the variance parameter is the same across all of the events, but in contrast to the other two models the variance is a quadratic function of the mean. Considering the

overdispersion seen in this data that seems an apt determination. Note that this model required more iteration than is the default to converge but it did converge. Again, week start date, rate of people tested, number of deaths, population were significant. However in this model only unknown and age groups 60 and up were significant at alpha = 0.05. AIC for this model was 81,800. The null and residual deviance was much lower here than for the other two models (Figure 2).

Model	Null Deviance	Residual Deviance
Poisson	1,189,783 on 11,799 degrees of freedom	580,131 on 11,786 degrees of freedom
Quasi-Poisson	1,189,783 on 11,799 degrees of freedom	580,131 on 11,786 degrees of freedom
Negative Binomial	23,739 on 11,799 degrees of freedom	13,156 on 11,786 degrees of freedom

Figure 2. Null and Residual Deviance of the three models.

❖ Model Checking

The QQ-plots are displayed here (FIGURE XX) and the full residual plots are included in the appendix. Based on this graphical examination I argue that the negative binomial model is the best choice, though none of the models are a particularly great fit. This fits well with the model output discussed above about the negative binomial model as compared to the other two.

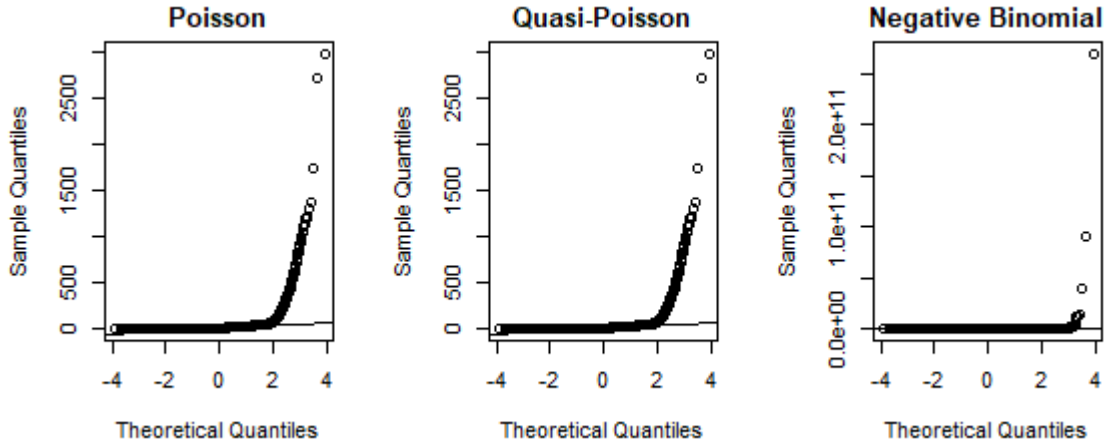


Figure 3. Quantile-Quantile Plots for the three models.

III. Discussion

Based on the analysis conducted the negative binomial model is the optimal one of those that were tested. Future analysis should investigate the suitability of a zero-inflated count model and further consider the issue of time series autocorrelation. Typically this might be handled by use of a lagged dependent variable, but since we have exponentiated coefficients the lagged variable becomes a growth rate rather than an autocorrelation coefficient. This very well may be appropriate since we expect some sort of trend in the case data, but is a limitation and is not within the scope of this analysis.

Though there are these limitations, there were certain consistencies across the board that give information

about the question of interest. Particularly, the population, rate of people tested, number of deaths, and week start date were all significant predictors in all three models. The first of these three make intuitive sense considering that our response variable is a count. The third is more interesting in that it may suggest that increased positive case counts within certain cities do not stay there, but are widespread throughout the county. Perhaps the cities mirror each other and there is more interaction occurring between individuals outside their home unit than is ideal.

On the other hand, there was not a very clear picture of the effect that age group has. Every model had some age groups as a significant predictor in summary but which age groups were significant varied. In contrast to its summary, for the ANOVA test (see appendix) for the quasi-Poisson model indicated that age group is not a significant predictor. The ANOVA tests for the Poisson and negative binomial tests were consistent with their summaries. Overall for the age groups, the primary pattern was that being in age groups 70-79 and 80+ were significant at alpha=0.05 for all models. In our model of choice, the negative binomial, the 95% confidence interval for 70-79 year olds is (0.35,0.47) and for 80+ year olds it is (0.47,0.59). Whether or not this is practically significant is up for debate, but I do believe this matches with the general message we're getting from health professionals about older individuals' susceptibility to the virus. There could be a wide variety of confounding factors here though.

Overall, this investigation makes some interesting conclusions about the timing of case counts and the impact age may have, but does not attempt to make any sweeping statements. The focus rather is on attempting to notice patterns in the COVID-19 data for King County, Washington, and any strong conclusions would require further analysis and testing.

IV. Appendix

Data Source

Data Extracts: https://www.kingcounty.gov/depts/health/covid-19/data/daily-summary/extracts.aspx

Site location of file downloads: https://www.kingcounty.gov/depts/health/covid-19/data/daily-summary.aspx

Download URL for Biweekly Counts and Rates by Geographic Levels:
https://www.kingcounty.gov/depts/health/covid-19/data/~/media/depts/health/communicable-diseases/documents/C19/data/biweekly-counts-rates-by-geography-dec-8.ashx

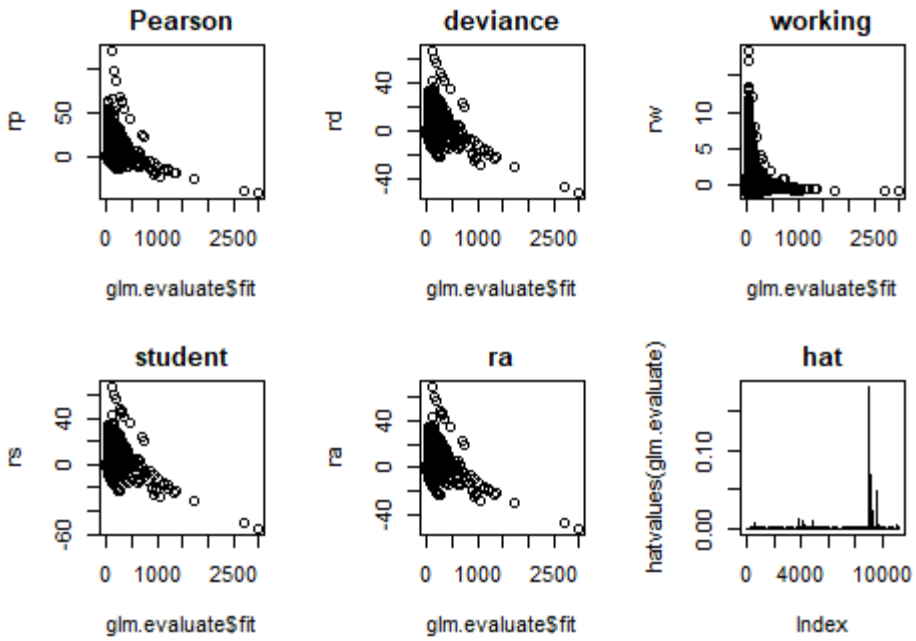
Download URL for Total Counts by City, Age, Sex at Birth, and Race/Ethnicity:
https://www.kingcounty.gov/depts/health/covid-19/data/~/media/depts/health/communicable-diseases/documents/C19/data/total-counts-by-date-city-demography-dec-8.ashx

References

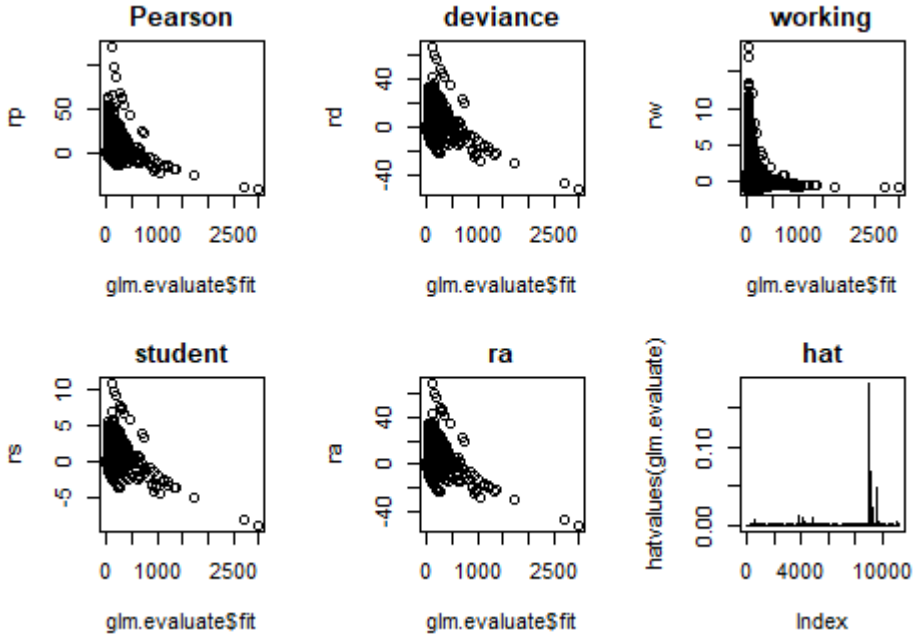
- Centers for Disease Control and Prevention. *United States COVID-19 Cases and Deaths by State.* Retrieved from https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days
- Public Health - Seattle & King County. *Coronavirus Disease 2019 (COVID-19).* Retrieved from https://www.kingcounty.gov/depts/health/covid-19.aspx
- Retka, Janelle. (2020, April 24). Yakima Herald-Republic. *Yakima County has highest rate of COVID-19 cases in Washington, double the state rate.* Retrieved from https://www.yakimaherald.com/special_projects/coronavirus/yakima-county-has-highest-rate-of-covid-19-cases-in-washington-double-the-state-rate/article_4daeb5ae-4646-5d2f-a5fb-78eed14fa050.html
- Schumakers, Erin. (2020, September 22). ABC News. *Timeline: How coronavirus got started.* Retrieved from https://abcnews.go.com/Health/timeline-coronavirus-started/story?id=69435165
- Ver Hoef, Jay M. and Boveng, Peter L., "QUASI-POISSON VS. NEGATIVE BINOMIAL REGRESSION: HOW SHOULD WE MODEL OVERDISPERSED COUNT DATA?" (2007). *Publications, Agencies and Staff of the U.S. Department of Commerce*. 142. https://digitalcommons.unl.edu/usdeptcommercepub/142
- Washington Demographics by Cubit. *Washington Counties by Population.* Retrieved from https://www.washington-demographics.com/counties_by_population
- Washington State Department of Health. *COVID-19 Data Dashboard.* Retrieved from https://www.doh.wa.gov/Emergencies/COVID19/DataDashboard

Additional Figures

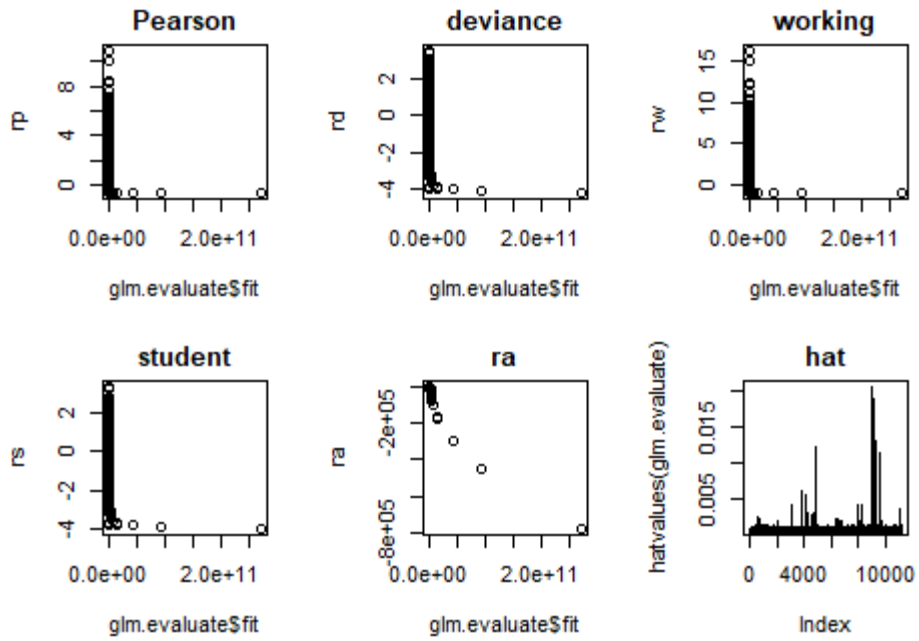
Poisson Residuals:



Quasi-Poisson Residuals:



Negative Binomial Residuals:



estigate the suitability of a zero-inflated count model (two-component mixture models combining a point mass at zero with a proper count dis

Poisson Model Analysis of Deviance Table (Type II Tests)			
Variable	**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

Quasi-Poisson Model Analysis of Deviance Table (Type I Tests)						
Variable	**df**	**Deviance**	**Residual df**	**Residual Deviance**	**F**	**Pr(>F)**
Week_Start	1	60002	11208	888942	1061.8	<2.2e-16

People_Tested_Rate	1	70106	11207	758835	1240.6	<2.2e-16
Deaths	1	242934	11206	515901	4298.8	<2.2e-16
Age_Group	9	0	11197	515901	0	0
Population	1	72909	11196	442993	1290.1	<2.2e-16

Negative Binomial Model Analysis of Deviance Table (Type II Tests)			
Variable	**LR Chisq**	**df**	**Pr(>Chisq)**
Week_Start	61.62	1	<2.2e-16
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

Poisson Coefficients Table:

Coefficient	Estimate	Standard Error	p value
Intercept	-1.48E+02	5.55E-01	< 2e-16
Week_Start	9.46E-08	3.48E-10	< 2e-16
People_Tested_Rate	3.45E+01	2.03E-01	< 2e-16
Deaths	1.15E-01	3.07E-04	< 2e-16
Age_Group10-19	3.29E-02	7.27E-03	6.13E-06
Age_Group20-29	-5.02E-01	7.78E-03	< 2e-16
Age_Group30-39	-4.24E-01	7.61E-03	< 2e-16
Age_Group40-49	-1.41E-01	7.30E-03	< 2e-16
Age_Group50-59	-5.01E-02	7.27E-03	5.63E-12
Age_Group60-69	5.64E-03	7.27E-03	0.438
Age_Group70-79	1.30E-01	7.28E-03	< 2e-16
Age_Group80+	1.60E-01	7.29E-03	< 2e-16
Age_GroupUnknown	2.08E-01	7.30E-03	< 2e-16
Population	1.96E-05	5.68E-08	< 2e-16

Quasi-Poisson Coefficients Table:

Coefficient	Estimate	Standard Error	p value
Intercept	-1.48E+02	4.75E+00	< 2e-16
Week_Start	9.46E-08	2.98E-09	< 2e-16
People_Tested_Rate	3.45E+01	1.73E+00	< 2e-16
Deaths	1.15E-01	2.63E-03	< 2e-16
Age_Group10-19	3.29E-02	6.22E-02	0.597289
Age_Group20-29	-5.02E-01	6.66E-02	4.88E-14
Age_Group30-39	-4.24E-01	6.52E-02	7.78E-11
Age_Group40-49	-1.41E-01	6.25E-02	0.024528
Age_Group50-59	-5.01E-02	6.22E-02	0.420941
Age_Group60-69	5.64E-03	6.22E-02	0.927756
Age_Group70-79	1.30E-01	6.23E-02	0.037325
Age_Group80+	1.60E-01	6.24E-02	0.010243
Age_GroupUnknown	2.08E-01	6.24E-02	0.000879
Population	1.96E-05	4.86E-07	< 2e-16

Negative Binomial Coefficients Table

Coefficient	Estimate	Standard Error	p value
Intercept	-1.48E+02	4.75E+00	< 2e-16
Week_Start	9.46E-08	2.98E-09	< 2e-16
People_Tested_Rate	3.45E+01	1.73E+00	< 2e-16
Deaths	1.15E-01	2.63E-03	< 2e-16
Age_Group10-19	3.29E-02	6.22E-02	0.597289
Age_Group20-29	-5.02E-01	6.66E-02	4.88E-14
Age_Group30-39	-4.24E-01	6.52E-02	7.78E-11
Age_Group40-49	-1.41E-01	6.25E-02	0.024528
Age_Group50-59	-5.01E-02	6.22E-02	0.420941
Age_Group60-69	5.64E-03	6.22E-02	0.927756
Age_Group70-79	1.30E-01	6.23E-02	0.037325
Age_Group80+	1.60E-01	6.24E-02	0.010243
Age_GroupUnknown	2.08E-01	6.24E-02	0.000879
Population	1.96E-05	4.86E-07	< 2e-16

Code

```
# Set up environment ------------------------------------
# Load Libraries
library(tidyverse)
library(here)
library(magrittr)
library(readxl)
library(MASS)
library(car)
library(robust)
library(GGally)
library(rcompanion)
library(hermite)
library(gee)


# Load Data - Updated 12-8!
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")

# Clean Data --------------------------------------------
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")

# Visualize Data ----------------------------------------
# Scatterplot
ggplot(covid) +
  geom_point(aes(Week_Start,
                 Positives),
             alpha=0.01) +
  labs(title= "Biweekly Counts of Positive COVID-19 Tests",
       subtitle = "Cities in King County from January 29, 2020 to December 8, 2020",
       caption= "Source: Washington State Department of Health",
       x="",
       y="Biweekly Positives\n")+
  theme_light()+
  theme(legend.position = "none")

# Histogram Code
ggplot(covid) +
  geom_boxplot(aes(Positives)) +
  labs(y="",
       x="Biweekly Positives")+
  theme_light()+
  theme(legend.position = "none")
```

```
# Poisson Model ----------------------------------------
pois_mod <- glm(Positives~.-City,
                data=covid,
                family="poisson")

summary(pois_mod)

# Poisson Model Residuals Analysis
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")


# Quasi-Poisson Model -----------------------------------
qp_mod = glm(Positives~.-City,
             data=covid,
             family=quasipoisson(link = "log"))
summary(qp_mod)

# Quasi-Poisson Residuals Analysis
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")


# Negative Binomial Model -------------------------------
nb_mod <- glm.nb(Positives~.-City,
                 data = covid,
                 control = glm.control(maxit=75))
summary(nb_mod)

# Negative Binomial Residuals Analysis
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")


# QQ plots ----------------------------------------------
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)

# Another check for outliers ----------------------------
outlierTest(pois_mod)
outlierTest(qp_mod)
outlierTest(nb_mod)

# Anova tables ------------------------------------------
Anova(pois_mod,
      type="II",
      test="LR")

anova(qp_mod,
      test="F")

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