STA2201 ASSIGNMENT 1

Alice Huang

31/01/2023

QUESTION 1

Question 1a

Suppose $Y|\theta \sim Poisson(\mu\theta), E(\theta) = 1, Var(\theta) = \sigma^2$. Then $E(Y|\theta) = \mu\theta$ since $Y|\theta$ is Poisson.

By law of total expectation, $E(Y) = E_\theta(E_Y(Y|\theta)) = E_\theta(\mu\theta) = \mu E(\theta) = \mu(1) = \mu$.

Thus $E(Y) = \mu$.

By law of total variance, $Var(Y) = E(Var(Y|\theta)) + Var(E(Y|\theta))$.

Note that $Var(Y|\theta) = \mu\theta$. We get

$Var(Y) = E(\mu\theta) + Var(\mu\theta) = \mu E(\theta) + \mu^2 Var(\theta) = \mu(1) + \mu^2\sigma^2 = \mu(1 + \mu\sigma^2)$

Question 1b

Suppose $\theta \sim Gamma(\alpha, \beta)$ where α, β are respectively shape and scale parameters.

Then since θ is Gamma distributed $\pi(\theta) = \frac{\beta^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\theta}$ and since $x|\theta$ is Poisson distributed, $L(x|\theta) = \frac{(\mu\theta)^x}{x!} e^{-\mu\theta}$

We know that $f(y) = \int f(y|\theta)\pi(\theta)d\theta$ so

$f(y) = \int \frac{(\mu\theta)^x}{x!} e^{-\mu\theta} \frac{\beta^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\theta} d\theta = \int \frac{\mu^x \beta^\alpha}{x!\Gamma(\alpha)} \theta^{x+\alpha-1} e^{-(\mu+\beta)\theta} d\theta$

The Negative Binomial probability mass function is $\binom{k+r-1}{k}(1-p)^k p^r$

$E(\theta) = \alpha\beta, Var(\theta) = \alpha\beta^2, mgf = (1 - \beta t)^{-\alpha}, t < \frac{1}{\beta}$

By law of total expectation, $E(Y) = E_\theta(E_Y(Y|\theta)) = E_\theta(\mu\theta) = \mu E(\theta) = \mu\alpha\beta$.

Thus $E(Y) = \mu\alpha\beta$.

By law of total variance, $Var(Y) = E(Var(Y|\theta)) + Var(E(Y|\theta))$.

$Var(Y) = E(\mu\theta) + Var(\mu\theta) = \mu\alpha\beta + \mu^2\alpha\beta^2$

Question 1c

Suppose $\theta \sim Gamma(\alpha, \beta)$ where α, β are respectively shape and scale parameters.

$E(\theta) = \alpha\beta, Var(\theta) = \alpha\beta^2$

By law of total expectation, $E(Y) = E_\theta(E_Y(Y|\theta)) = E_\theta(\mu\theta) = \mu E(\theta) = \mu\alpha\beta$.

Thus $E(Y) = \mu\alpha\beta$. Compare this with $E(Y) = \mu$. We get that $\alpha\beta = 1$.

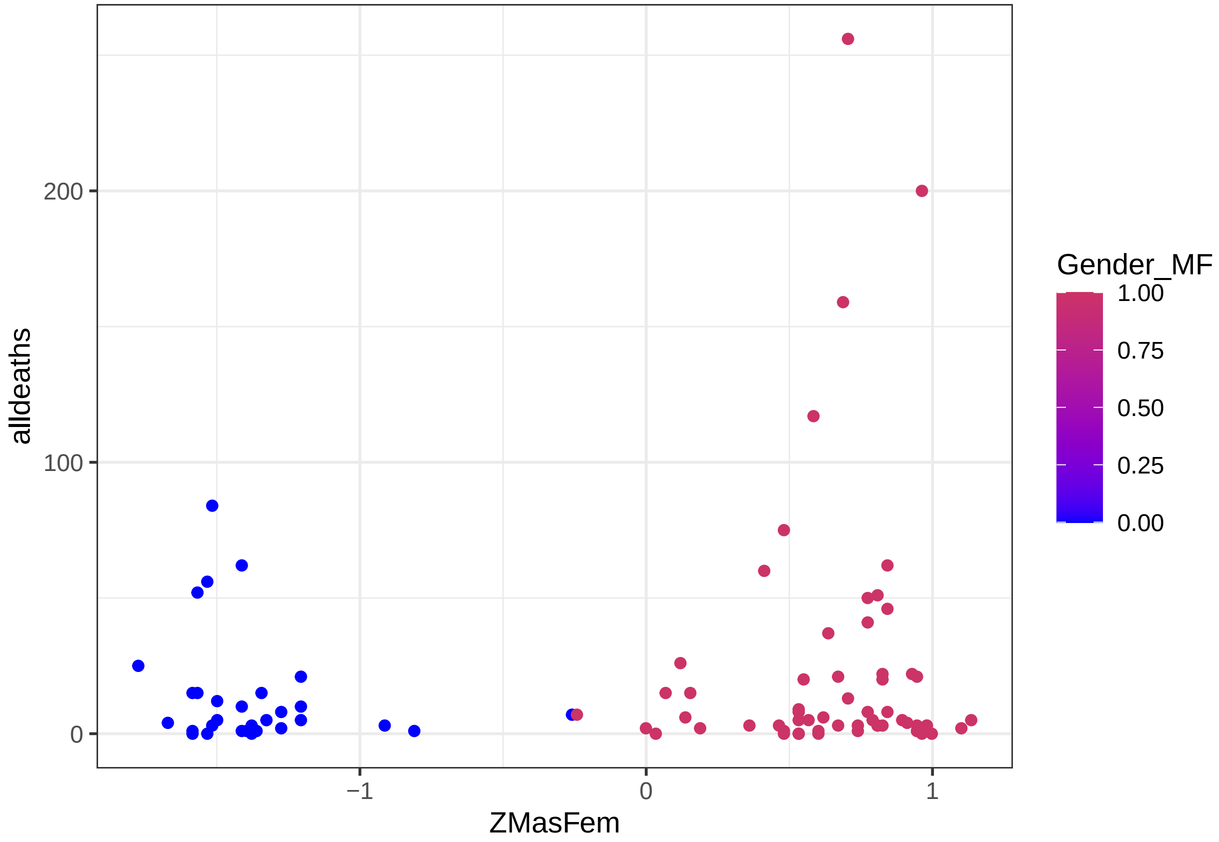
By law of total variance, $Var(Y) = E(Var(Y|\theta)) + Var(E(Y|\theta))$.

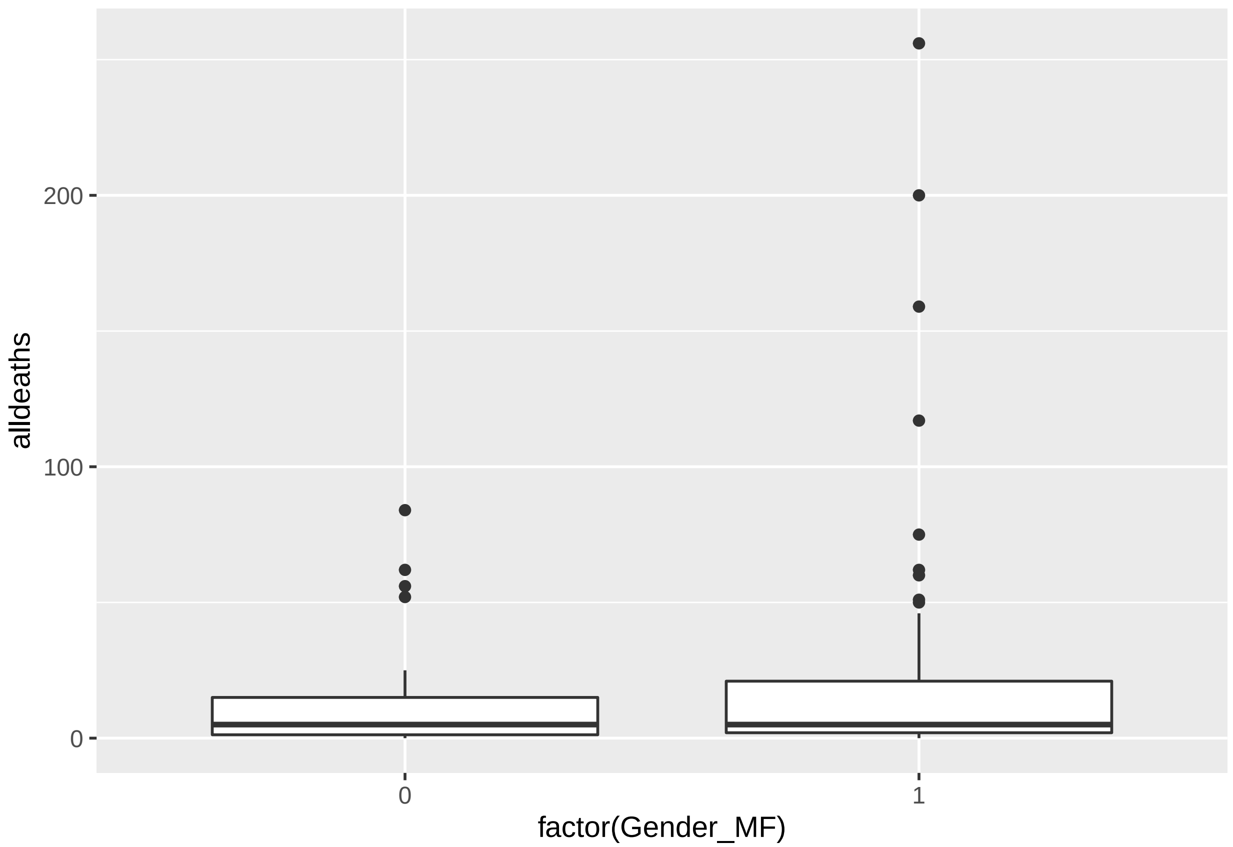
$Var(Y) = E(\mu\theta) + Var(\mu\theta) = \mu\alpha\beta + \mu^2\alpha\beta^2 = \mu(\alpha\beta + \mu\alpha\beta^2)$. Compare that with $Var(Y) = \mu(1 + \mu\sigma^2)$. We get that $\alpha\beta^2 = \sigma^2$. Since we must also have $\alpha\beta = 1$, this means $\alpha\beta^2 = (\alpha\beta)\beta = \beta = \sigma^2$.

We must have $\beta = \sigma^2$ and $\alpha = \frac{1}{\sigma^2}$.

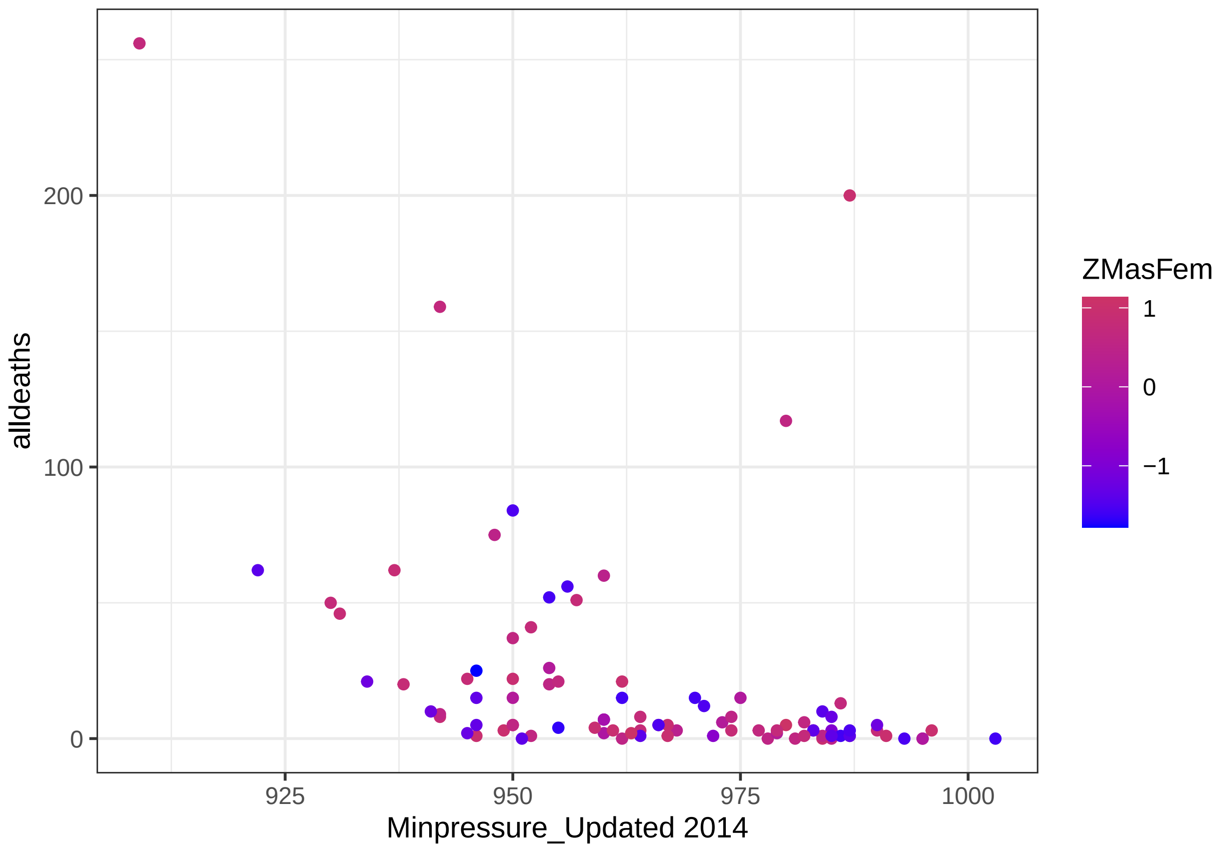
QUESTION 2

a) Create three graphs in ggplot that help to visualize patterns in deaths by femininity, minimum pressure, and damage. Discuss what you observe based on your visualizations.

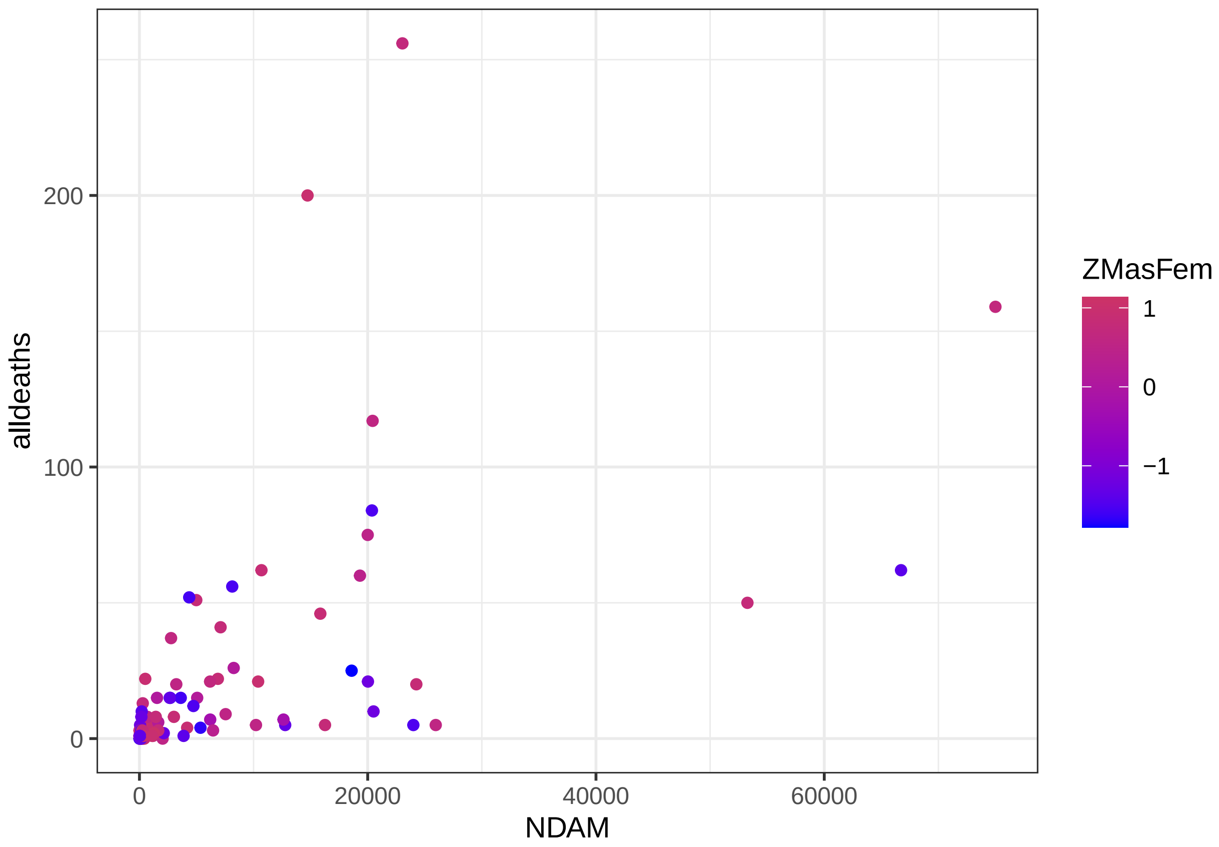




It appears that overall, there seem to be more feminine named hurricanes with very high death tolls (outliers). The distribution looks bimodal. There doesn't appear to be a clear trend through this scatterplot. If you compare the boxplots for death tolls of male named hurricanes and female named hurricanes, the median death tolls appear similar, but the feminine named hurricane group seems to have a higher standard error of death tolls and more outliers.



It seems that there is a weakly decreasing trend on the graph of minimum pressure against death toll.



It appears that on average, as the amount of damage increases, the death toll increases slightly. There are 3 hurricanes with significantly higher damage than the rest: Hurricanes Sandy, Andrew, and Donna.

b) Run a Poisson regression with `deaths` as the outcome and `femininity` as the explanatory variable. Interpret the resulting coefficient estimate. Check for overdispersion. If it is an issue, run a quasi-Poisson regression with the same variables. Interpret your results.

```
##
## Call:
## glm(formula = alldeaths ~ ZMasFem, family = "poisson", data = hdata1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.1429  -5.3716  -3.8288  -0.5364  27.4230
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.00128    0.02359 127.233   <2e-16 ***
## ZMasFem      0.23840    0.02546   9.362   <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: 4031.9  on 91  degrees of freedom
## Residual deviance: 3937.5  on 90  degrees of freedom
## AIC: 4266.4
##
## Number of Fisher Scoring iterations: 6
```

```
## [1] "Overdispersion factor of Poisson model"
```

```
## [1] 73.7846
```

```
## [1] "Probability that values greater than overdispersion factor are observed"
```

```
## [1] 0.8924091
```

Under a Poisson regression, for every unit increase in the standardized Masculinity-Femininity score, the death toll is expected to increase by exp(0.23840).

There seems to be overdispersion because the probability of observing values greater than the dispersion factor is 0.8924091 > 0.05. So the dispersion factor is not in the tails of the chi-squared distribution. So we consider a quasi-poisson model.

```
##
## Call:
## glm(formula = alldeaths ~ ZMasFem, family = "quasipoisson", data = hdata1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.1429  -5.3716  -3.8288  -0.5364  27.4230
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0013     0.2026   14.81   <2e-16 ***
## ZMasFem       0.2384     0.2187    1.09    0.279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 73.78496)
##
##     Null deviance: 4031.9  on 91  degrees of freedom
## Residual deviance: 3937.5  on 90  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Under a quasi-Poisson regression model, for every unit increase in the standardized Masculinity-Femininity score, the death toll is expected to increase by exp(0.23840).

c) Reproduce Model 4 (as described in the text and shown in Table S2).[^1] Report the estimated effect of femininity on deaths assuming a hurricane with median pressure and damage ratings.

```
##
## Call:
## MASS::glm.nb(formula = alldeaths ~ ZMinPressure_A + ZNDAM + ZMasFem +
##     ZMasFem * ZMinPressure_A + ZMasFem * ZNDAM, data = hdata1,
##     init.theta = 0.8112499791, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5088  -1.0527  -0.4759   0.2903   2.5741
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.4756     0.1222  20.261  < 2e-16 ***
## ZMinPressure_A          -0.5521     0.1503  -3.673 0.000239 ***
## ZNDAM                    0.8635     0.1445   5.976 2.28e-09 ***
## ZMasFem                  0.1723     0.1238   1.392 0.163988
## ZMinPressure_A:ZMasFem   0.3948     0.1521   2.595 0.009453 **
## ZNDAM:ZMasFem            0.7051     0.1501   4.699 2.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8112) family taken to be 1)
##
##     Null deviance: 184.86  on 91  degrees of freedom
## Residual deviance: 102.83  on 86  degrees of freedom
## AIC: 658.09
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  0.811
##          Std. Err.:  0.124
##
##  2 x log-likelihood:  -644.091
```

```
## [1] "Median Standardized Normalized Damage"
```

```
## [1] -0.434494
```

```
## [1] "Median Standardized Minimum Pressure"
```

```
## [1] -0.07239403
```

The regression equation of Model 4 was $alldeaths = 2.4756 - 0.5521'ZMinPressure'_A + 0.8635ZNDAM + 0.1723ZMasFem + 0.3948ZMasFem * `ZMinPressure_A` + 0.7051ZMasFem * ZNDAM$. The median standardized pressure is -0.07239403 and the median standardized damage is -0.434494. Plugging this in the equation above gives

$alldeaths = 2.4756 - 0.5521(-0.07239403) + 0.8635(-0.434494) + 0.1723ZMasFem + 0.3948(-0.07239403)ZMasFem + 0.7051(-0.434494)ZMasFem$

$alldeaths = 2.140383 - 0.1626429(ZMasFem)$

From this equation, it appears that as the ZMasFem index increases by 1 unit (ie the name is more feminine than masculine), the death toll decreases by 0.1626429.

d) Using Model 4, predict the number of deaths caused by Hurricane Sandy. Interpret your results.

```
##        1
## 20806.74
```

Model 4 predicts that Hurricane Sandy caused 20806.74 deaths. However, in reality, Hurricane Sandy only caused 159 deaths. So the prediction from Model 4 was too high. Perhaps this is due to the fact that Hurricane Sandy caused significantly more damage than other hurricanes. Perhaps Model 4 may be overfitted to the data, because the standard errors for ZNDAM and ZNDAM:ZMasFem coefficient estimates are small. Also, since there is a small number of variables in the model, large changes in ZNDAM yield disproportionately large changes in death tolls.

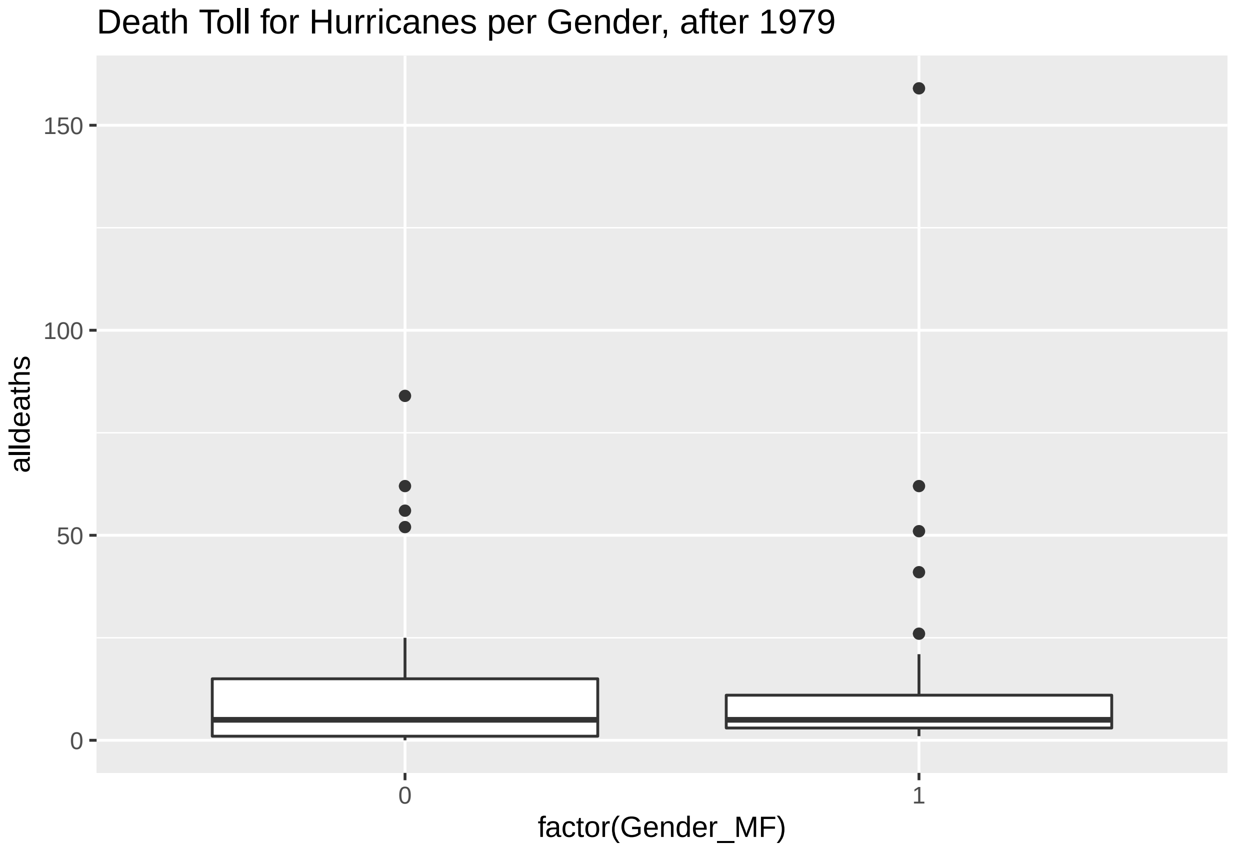
e) Describe at least two strengths and two weaknesses of this paper, focusing on the archival analysis. What was done well? What needed improvement?

One strength was that the reasoning behind choosing a negative binomial regression model was reasonable.

Another strength is that the raw dollar amounts of property damage caused by hurricanes were normalized to 2013 monetary values by adjusting them to inflation, wealth and population density.

One weakness is that the authors made a bold claim about female hurricanes having triple the death toll of male hurricanes. In general, the strength of the results seems to be overstated. In the Archival Study section, paragraph 2, the authors write "For example, a hurricane with a relatively masculine name (MFI = 3) is estimated to cause 15.15 deaths, whereas a hurricane with a relatively feminine name (MFI = 9) is estimated to cause 41.84 deaths. In other words, our model suggests that changing a severe hurricane's name from Charley (MFI = 2.889, 14.87 deaths) to Eloise (MFI = 8.944, 41.45 deaths) could nearly triple its death toll." I think the effect of name gender on hurricane death toll is overstated, especially in the latter sentence. The authors seemed to have obtained the 14.87 deaths number from the combined death tolls of two hurricanes named Charley, one from 1984 (5 deaths) and another from 2004 (10 deaths). There is only one Hurricane Eloise, which occurred in 1975. It seems strange to compare the death tolls of two hurricanes combined with the death toll of one hurricane. Hurricanes Charley and Eloise occurred in time periods with different technology, politics, and global climates. Furthermore, the death toll for Hurricane Eloise doesn't seem accurate, as the dataset suggests that Hurricane Eloise had a death toll of 21, not 41.45 as mentioned in the paper. It seems like specific hurricanes were cherry-picked to match the statistical results, the death toll of the feminine-named hurricane was incorrectly reported to be higher, and the differences were overstated.

Another issue is that the authors claim that gender of hurricane name has a significant effect on death toll despite the coefficient of the `ZMasFem` index not being statistically significant in Models 2 and 4. As previously seen from the boxplots comparing the death tolls for masculine-named and feminine-named hurricanes, the median death toll for male and female hurricanes is actually quite similar. This is still the case when you filter for the hurricanes after the male-female alternate naming rule was implemented.



In the Materials and Methods section, Additional Analyses subsection, the authors write "For hurricanes before 1979 (n = 38), a model in which normalized damage, minimum pressure, MFI, and two two-way interaction terms (MFI × normalized damage, MFI × minimum pressure) were entered generated similar but nonsignificant interactions (MFI × minimum pressure: β = 0.007, P = 0.408, SE = 0.008; MFI × normalized damage: β = 0.00003, P = 0.308, SE = 0.00003). For hurricanes after 1979 (n = 54), a model with normalized damage, minimum pressure, MFI, and two two-way interaction terms (MFI × normalized damage, MFI × minimum pressure) yielded a marginally significant interaction between MFI and normalized damage (β = 0.00001, P = 0.073, SE = 0.000004). The interaction between MFI and minimum pressure was nonsignificant (β = 0.003, P = 0.206, SE = 0.0028). In addition, using the gender of the hurricane name as a binary variable instead of MFI showed similar but nonsignificant interactions (gender of hurricane name × normalized damage: β = -0.00004, P = 0.128, SE = 0.00003; gender of hurricane name × minimum pressure: β = -0.019, P = 0.326, SE = 0.0197)." So if the researchers control for era, the two-way interaction effects of femininity of name and other variables seems to be insignificant, in addition to the main effect of femininity being insignificant in Models 2, and 4. I believe that the strength of the results is overstated in the paper, given that the femininity related coefficients and interactions are not significant when the models are controlled for era.

f) Are you convinced by the results? If you are, explain why. If you're not, describe what additional data and/or analyses you would like to see to further test the author's hypothesis.

I am not convinced by the results. The sample size of hurricanes after the male-female alternating rule was implemented seems rather small. There were only 30 male-named hurricanes. Furthermore, some names like "Able", "Ione" were rated as masculine despite being from the pre-1979 era when female names were much more prevalent in hurricane names. The gender of some old rare names like "Easy", "Inez" seem ambiguous to me. Coders from 2013 may not be familiar with names from the 1950s, and how people living during the 1950s would have responded to the names of the hurricanes. I think there should at least be an effort to recruit older participants who lived closer to those time periods, and may be able to give more accurate perceptions of the names' genders. Otherwise the ambiguous names may be removed, however I think this would worsen the issue of having a small sample set with low power, so I would avoid it if possible.

Only 9 coders determined the femininity of hurricane names. Since the researchers hypothesized that gender of hurricane name has an effect on response and death toll, I think more coders should have been recruited to determine the femininity of hurricane names for a more accurate response, especially given the variation in names from a span of over 60 years.

I also noticed that `ZMinPressure_A`, `ZNDAM` had a correlation of -0.5559824. I would like to see the summary of a negative binomial model with an interaction term of `ZMinPressure_A:ZNDAM`.

The death toll did not adjust for the population of the affected areas at the time. I would like to see the death tolls more accurately adjust for population density in the affected areas given census data from the year of the hurricane. I think that would give a more accurate perception of hurricane deadliness.

In Model 4, for a hurricane with median pressure and damage rating, femininity of hurricane name seems to be associated with a decline in death count. This does not seem consistent with the author's claims that feminine names predict higher hurricane death tolls. I would like to compare the coefficients of other models for hurricanes with median pressure and damage rating. It seems that there were 2 female-named hurricanes with significantly higher damage than the rest, so considering the median damage instead of the mean damage would probably give a more accurate measure of central tendency.

Vaccinations

This question relates to COVID-19 vaccination rates in the United States. We are interested in exploring factors that are associated with differences in vaccine coverage by US county.

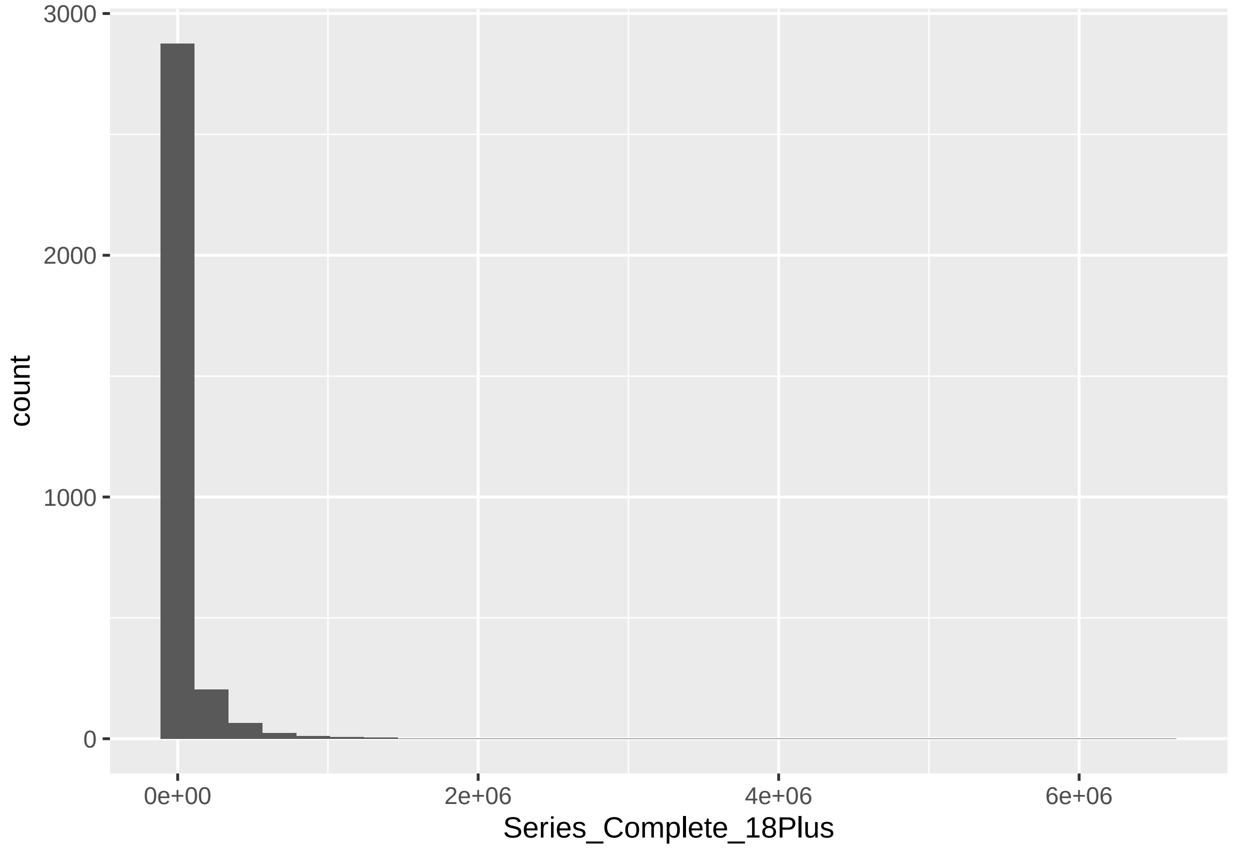
- You can download the latest data on vaccination coverage here: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/data. Note that this is updated most days so depending on when you download it, it might be slightly different from others (that's okay). For the purposes of the assignment, please consider data from the 11th of January 2023. Also note that on the same webpage you should be able to find a data dictionary. We will be interested in people who have completed a primary vaccine series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine), which refers to columns that have the `Series_Complete` prefix.
- The class repo has a dataset `acs` that contain a range of different demographic, socioeconomic, and health variables by county. These were obtained from the American Community Survey (ACS) via the R package `tidycensus`. For reference, the extraction code can be found in the repo (`acs.R`)

a) Perform some exploratory data analysis (EDA) using a dataset combining the vaccination and ACS data, and summarize your observations with the aid of 3-4 key tables or graphs.

```
## Warning: * You have not set a Census API key. Users without a key are limited to 500
## queries per day and may experience performance limitations.
## i For best results, get a Census API key at http://api.census.gov/data/
## key_signup.html and then supply the key to the ‘census_api_key()‘ function to
## use it throughout your tidycensus session.
## This warning is displayed once per session.
```

```
newvaxdata %>% ggplot(aes(Series_Complete_18Plus)) + geom_histogram()
```

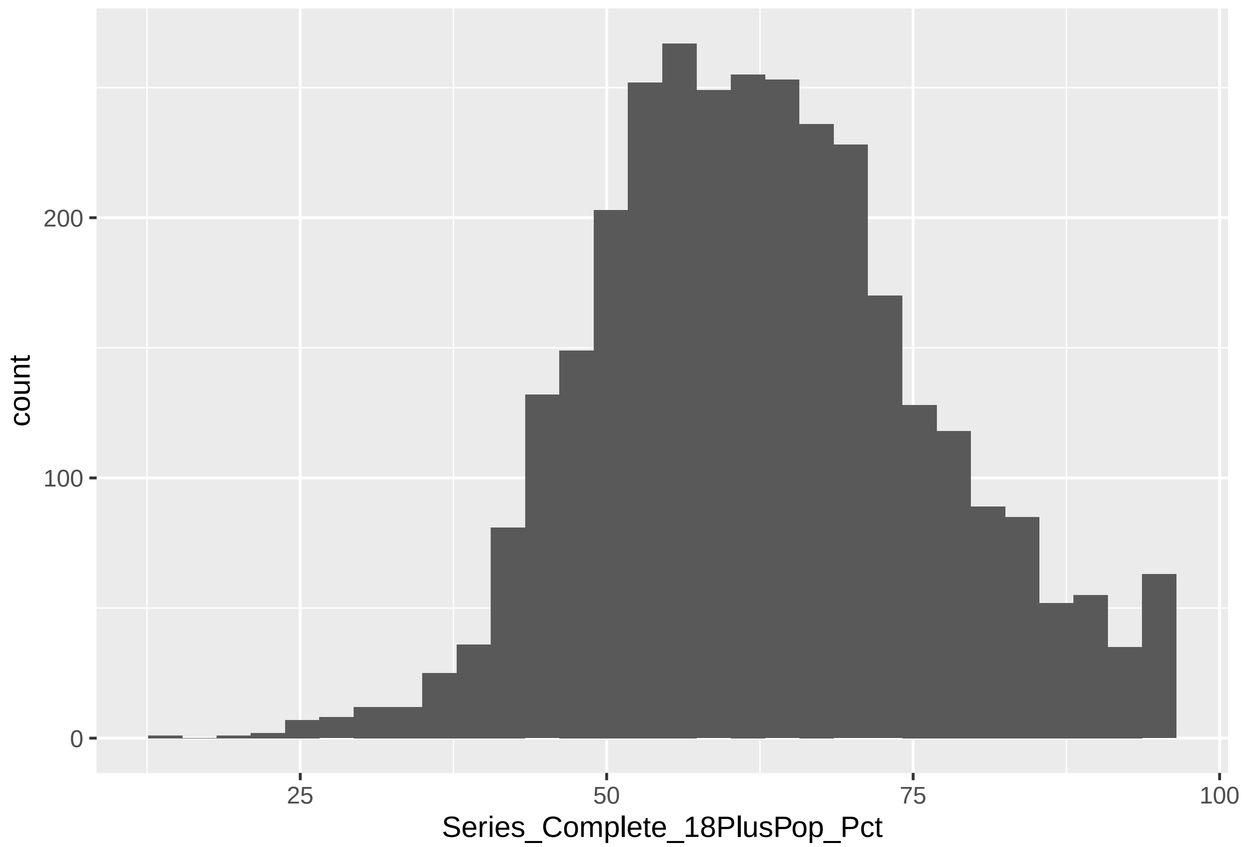
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## Warning: Removed 16 rows containing non-finite values (stat_bin).
```



```
newvaxdata %>% ggplot(aes(Series_Complete_18PlusPop_Pct)) + geom_histogram()
```

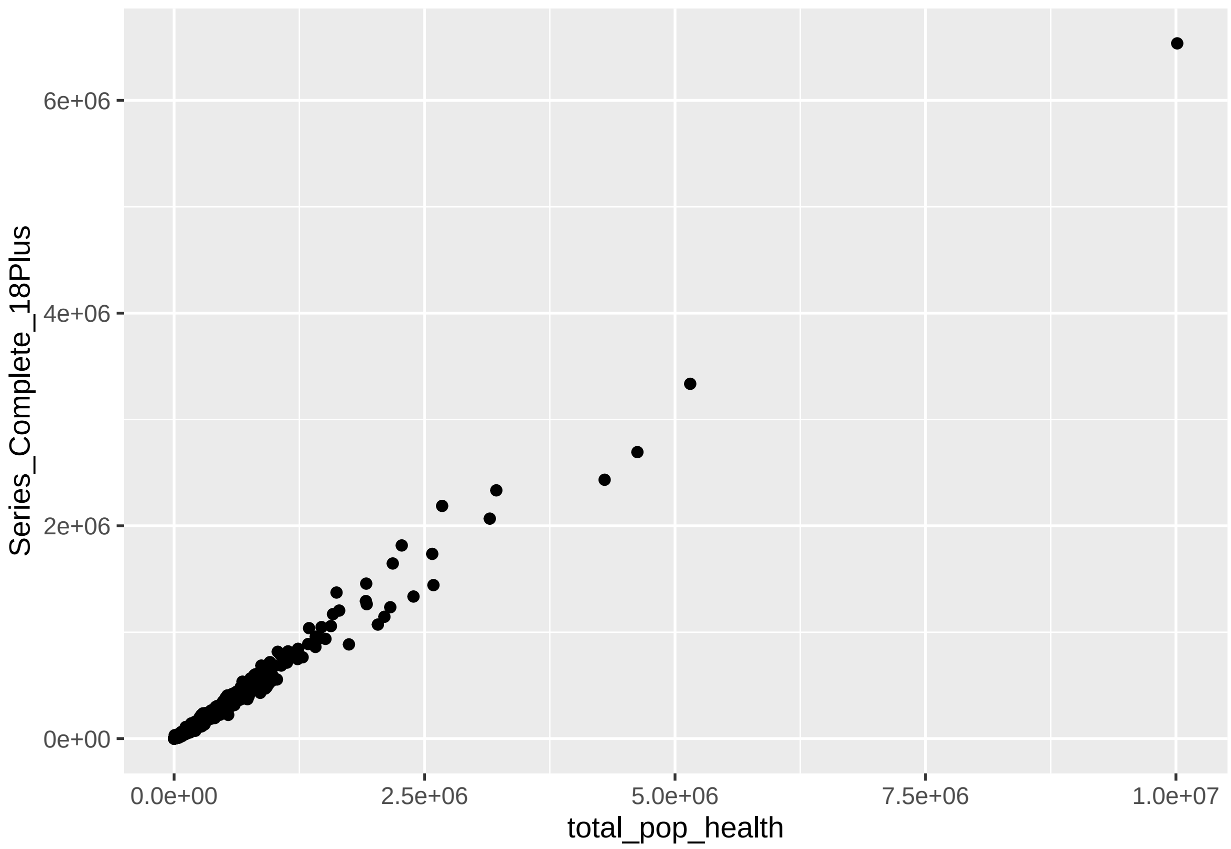
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
## Warning: Removed 16 rows containing non-finite values (stat_bin).
```



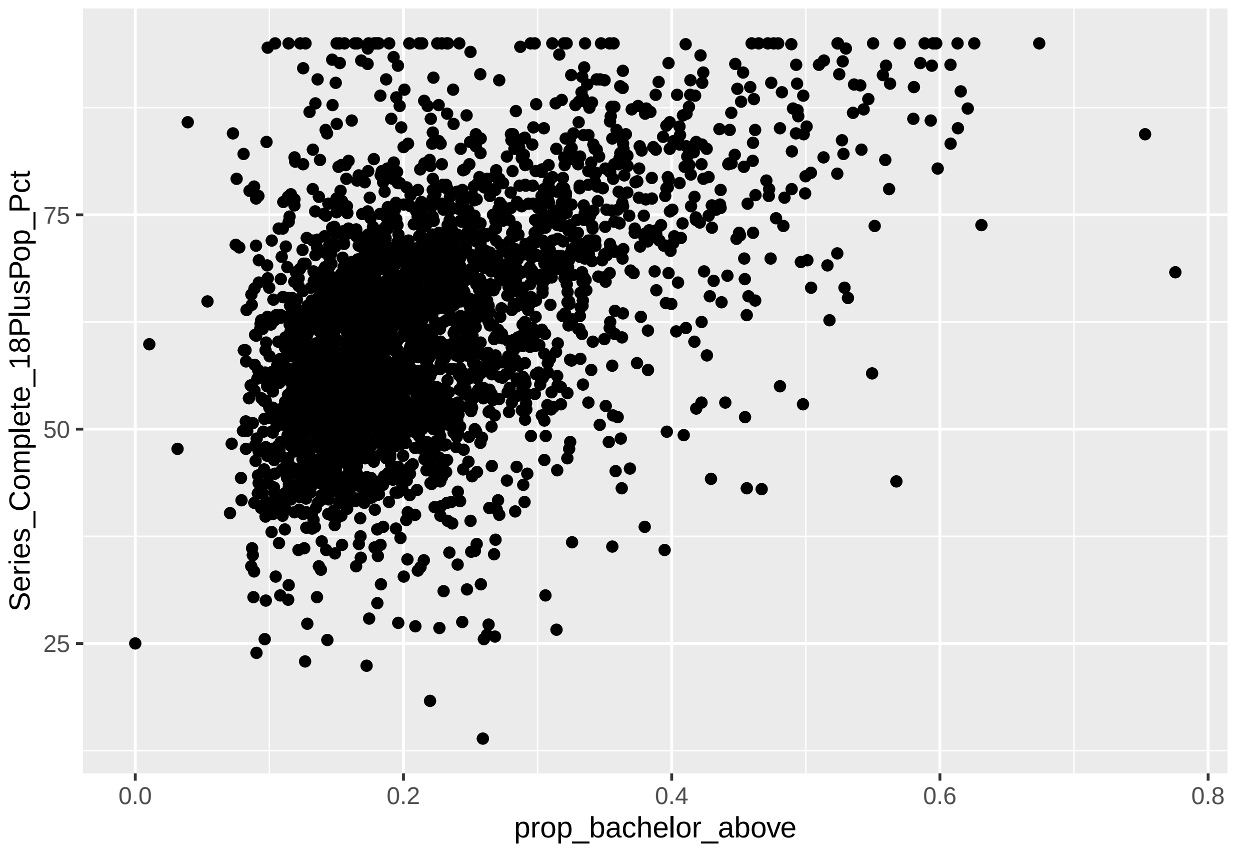
```
newvaxdata %>% ggplot(aes(x=total_pop_health, y=Series_Complete_18Plus)) + geom_point()
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```



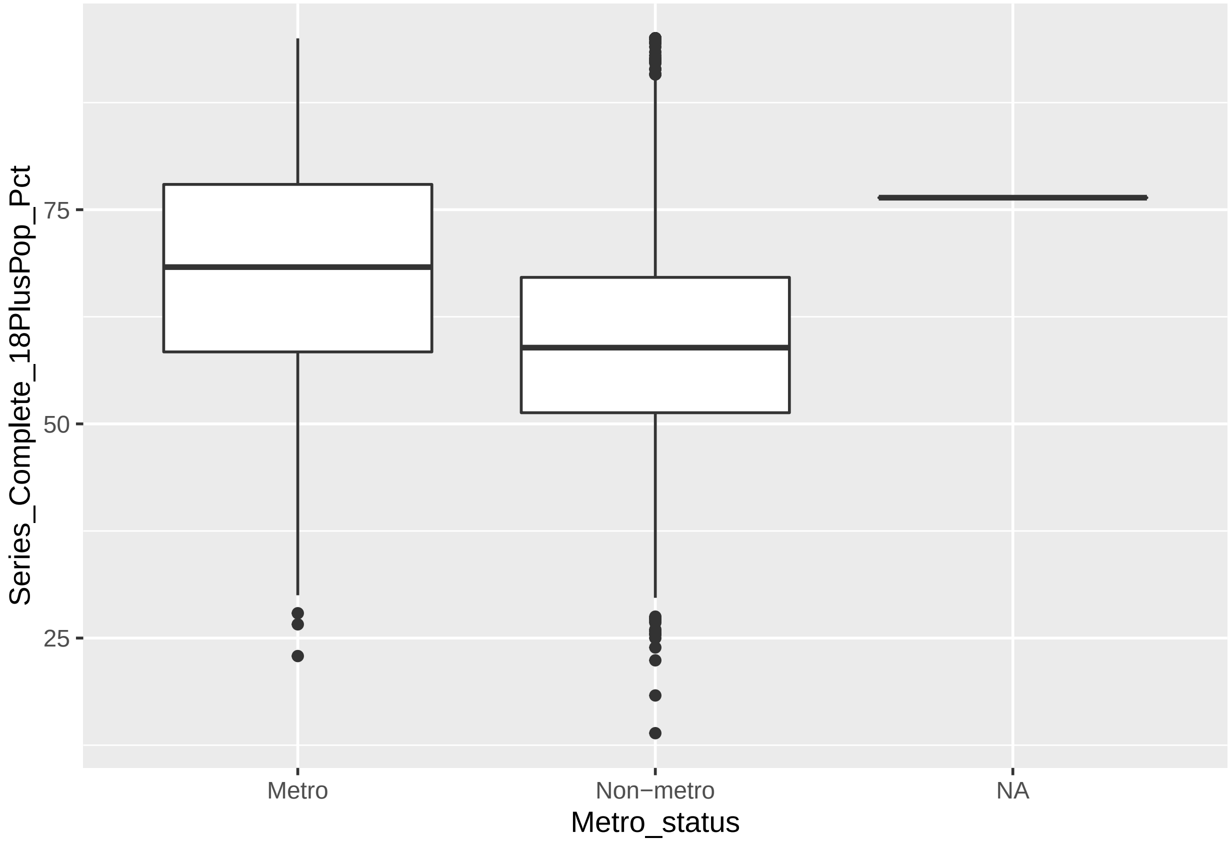
```
newvaxdata %>% ggplot(aes(x=prop_bachelor_above, y=Series_Complete_18PlusPop_Pct)) + geom_point()
```

```
## Warning: Removed 94 rows containing missing values (geom_point).
```



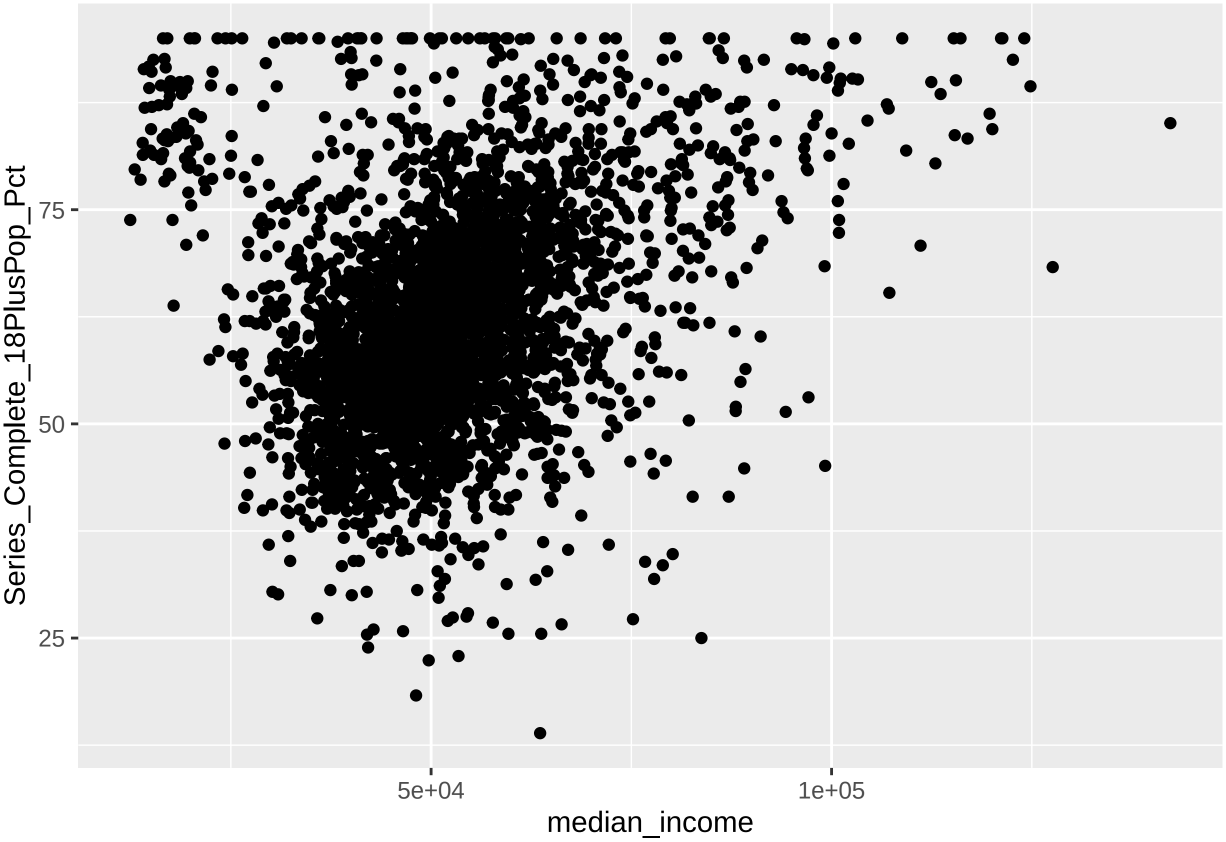
```
newvaxdata %>% ggplot(aes(x=Metro_status, y=Series_Complete_18PlusPop_Pct)) + geom_boxplot()
```

```
## Warning: Removed 16 rows containing non-finite values (stat_boxplot).
```



```
newvaxdata %>% ggplot(aes(x=median_income, y=Series_Complete_18PlusPop_Pct)) + geom_point()
```

```
## Warning: Removed 16 rows containing missing values (geom_point).
```

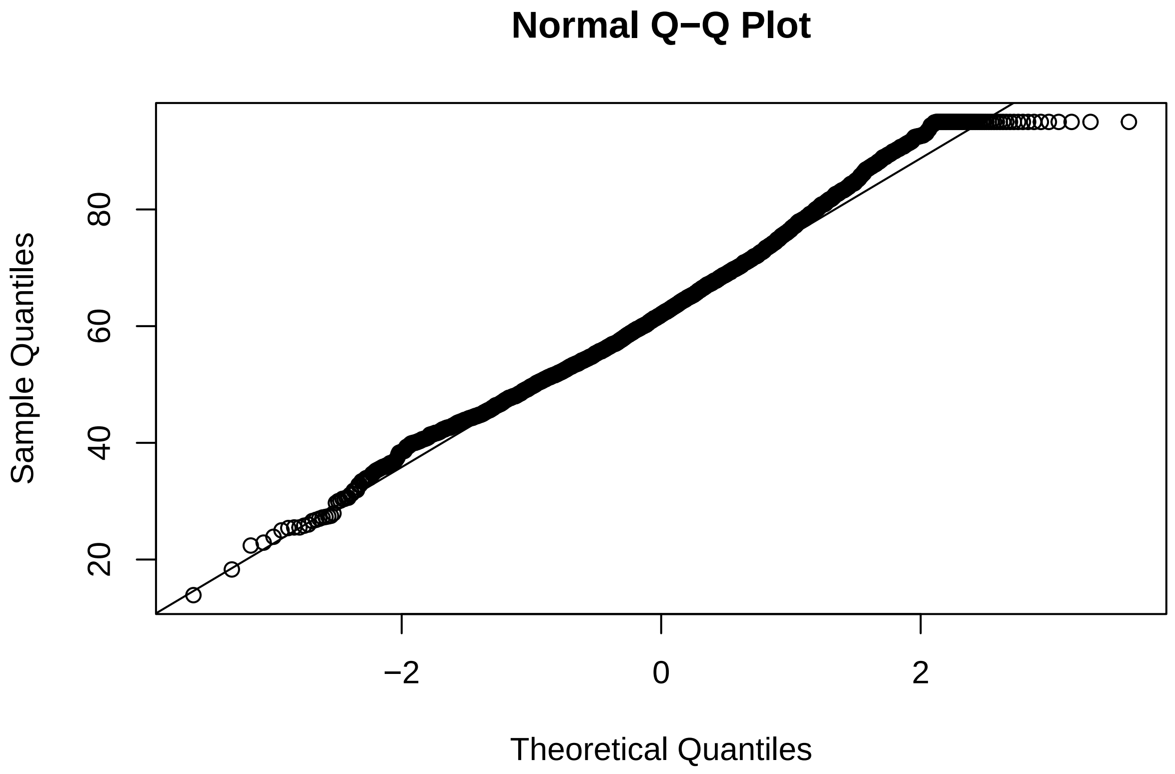


b) Build a regression model at the county level to help investigate patterns in the full vaccination rate for the population aged 18+ (that is, people aged 18+ who have completed a primary vaccine series). There is no one right answer here, but you should justify the outcome measure you are using (e.g. counts, proportions, rates, etc) and your distributional assumptions about the outcome measure (e.g. binary, poisson, normal, etc). You should also discuss briefly your model building strategy; what covariates you considered and why (motivated by your EDA)[1], and how the candidate model was chosen. Interpret your findings, including visualizations where appropriate.

To choose which variables I should include, I guessed which variables would probably have a significant effect on vaccination rates. For example, I guessed that counties with higher rates of post-secondary education would have higher vaccinations. I grouped variables together into clusters (education-related variables, employment-related variables, etc) and guessed which variables would be correlated with each other. For example, I guessed that `median_rent` and `Metro_status` would probably be correlated. I only picked one variable from each cluster to avoid correlation between similar variables.

First I considered modeling the outcome as proportions, instead of just counts because counts can be interpreted differently depending on the size of the county population. County populations can vary from 66-10081570. I considered making my response `Series_Complete_18PlusPop_Pct`, the percent of people ages 18+ who have completed a primary series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine) based on the jurisdiction and county where vaccine recipient lives. The histogram looked bell-shaped but unfortunately it did not appear to be normal, because the points on the quantile-quantile normal plot did not follow a straight line and it failed the Shapiro Wilk test (rejected null hypothesis is that data is normal). I also tried modelling the log of the proportion but the same issues with non-normality persisted.

[1] Note that the vaccines dataset also has a `Metro` variable which you are welcome to use in your analyses.



```
##
##  Shapiro-Wilk normality test
##
## data:  newvaxdata$Series_Complete_18PlusPop_Pct
## W = 0.99175, p-value = 1.236e-12
```

We also consider using a Poisson regression to model count data. To solve the issue of counts not showing significance, we offset by the population variable. Since the response is being modeled using counts, the explanatory variables should also be modeled using counts. Then we check for overdispersion.

```
##
## Call:
## glm(formula = Series_Complete_18Plus ~ Metro_status + median_income +
##     median_rent + total_pop_health + total_pop_employ + bachelor,
##     family = "poisson", data = newvaxdata, offset = log(Census2019_18PlusPop))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -184.10   -18.71    -4.78     7.74   320.86
##
## Coefficients:
##                          Estimate Std. Error  z value Pr(>|z|)
## (Intercept)           -5.804e-01  3.582e-04 -1620.25   <2e-16 ***
## Metro_statusNon-metro -1.027e-01  2.530e-04  -405.79   <2e-16 ***
## median_income          4.715e-07  7.931e-09    59.45   <2e-16 ***
## median_rent            2.408e-04  4.438e-07   542.66   <2e-16 ***
## total_pop_health      -2.094e-07  1.891e-09  -110.73   <2e-16 ***
## total_pop_employ       2.378e-07  2.554e-09    93.12   <2e-16 ***
## bachelor               1.532e-07  2.455e-09    62.41   <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: 5683153  on 3120  degrees of freedom
## Residual deviance: 3115507  on 3114  degrees of freedom
##   (99 observations deleted due to missingness)
## AIC: 3151023
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] "Overdispersion factor"
```

```
## [1] 1016.213
```

Unfortunately, there seemed to be a lot of overdispersion (1016.213) so a Poisson regression is not appropriate as the mean and variance would not be the same.

```
##
## Call:
## glm(formula = Series_Complete_18Plus ~ Metro_status + median_income +
##     median_rent + total_pop_health + total_pop_employ + total_pop_educ,
##     family = "quasipoisson", data = newvaxdata, offset = log(Census2019_18PlusPop))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -161.349   -18.600    -4.663     7.812   278.363
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -5.824e-01  1.005e-02 -57.974  < 2e-16 ***
## Metro_statusNon-metro -1.018e-01  7.981e-03 -12.759  < 2e-16 ***
## median_income          6.840e-07  2.329e-07   2.937  0.00334 **
## median_rent            2.276e-04  1.423e-05  15.989  < 2e-16 ***
## total_pop_health      -8.174e-08  6.272e-08  -1.303  0.19255
## total_pop_employ      -7.417e-07  1.945e-07  -3.814  0.00014 ***
## total_pop_educ         1.002e-06  1.711e-07   5.855 5.27e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasipoisson family taken to be 1007.812)
##
##     Null deviance: 5683153  on 3120  degrees of freedom
## Residual deviance: 3084855  on 3114  degrees of freedom
##   (99 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

The quasipoisson dispersion parameter was large. Unfortunately, the quasipoisson model seems to have a very high residual deviance given the number of residual degrees of freedom. So we consider a negative binomial model.

The negative binomial model is similar to the Poisson regression model but it does not require the mean and variance to be the same. We also include an offset term for the population in the negative binomial model.

```
##
## Call:
## MASS::glm.nb(formula = Series_Complete_18Plus ~ total_pop_educ *
##     median_income + total_pop_health * total_pop_employ + Metro_status +
##     median_income * median_rent + Metro_status * median_rent +
##     offset(log(Census2019_18PlusPop)), data = newvaxdata, init.theta = 25.48930145,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.6146  -0.6559  -0.0338   0.5378  11.9869
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -8.084e-01  6.534e-02 -12.371  < 2e-16 ***
## total_pop_educ                   -1.202e-06  8.655e-07  -1.389  0.16473
## median_income                     2.004e-06  8.429e-07   2.377  0.01745 *
## total_pop_health                 -1.287e-06  4.438e-07  -2.901  0.00372 **
## total_pop_employ                  3.053e-06  1.097e-06   2.783  0.00538 **
## Metro_statusNon-metro             7.267e-03  4.004e-02   0.181  0.85598
## median_rent                       2.796e-04  8.617e-05   3.245  0.00118 **
## total_pop_educ:median_income     -4.024e-12  1.471e-12  -2.735  0.00623 **
## total_pop_health:total_pop_employ -2.267e-14  4.131e-15  -5.488 4.06e-08 ***
## median_income:median_rent        -1.384e-10  9.603e-10  -0.144  0.88540
## Metro_statusNon-metro:median_rent -8.897e-06  5.074e-05  -0.175  0.86080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(25.4893) family taken to be 1)
##
##     Null deviance: 3985.2  on 3120  degrees of freedom
## Residual deviance: 3163.2  on 3110  degrees of freedom
##   (99 observations deleted due to missingness)
## AIC: 58387
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  25.489
##           Std. Err.:  0.649
##
##  2 x log-likelihood:  -58363.337
```

The negative binomial model on the counts seems to have a more reasonable Residual Deviance given the number of degrees of freedom, so we choose the negative binomial model.

c) Use your model from b) to predict the proportion of the population aged 18+ in Ada County, Idaho who are fully vaccinated. Briefly discuss how good you think this prediction is, and why.

The model predicts that 263117 people in Ada County will be vaccinated. We divide by the population of people aged 18+ to find the proportion of adults who are fully vaccinated.

```
##         1
## 0.7115759
```

```
##         1
## 0.7362406
```

```
## # A tibble: 1 x 1
##   Series_Complete_18PlusPop_Pct
##                           <dbl>
## 1                          76.9
```

The model predicted that the proportion of the population aged 18+ in Ada County, Idaho who are fully vaccinated, is around 71.1%, compared to the true value of 76.9%. I think this is pretty close.

For the sake of comparison, I also tried using the quasi-poisson model for prediction and that yielded 73.6%. However, I will stick with the negative binomial model because I think it will be a better model overall.

Of course, it is difficult to satisfy many theoretical assumptions of the model.

Given the limitations of the model, the prediction still looks reasonably close to the true value so I think the prediction still looks good.

d) Give a brief summary of your analysis. What other variables may be of interest to investigate in future?

e) Now consider the situation of analysing vaccination rates at the **state** level. Consider the three following options:

    1) Regression at the state level, outcome used is the total population 18+ fully vaccinated
    2) Regression at the state level, outcome used is the average of the county level full vaccination rates of 18+ population
    3) Regression at the county level, outcome used is the total population 18+ fully vaccinated, and include as a covariate a categorical variable (fixed effect) which indicates which state a county is in.

    Without performing these regressions, briefly discuss how you think these three approaches would differ in terms of the granularity of information used and the type of outcome measure. In your opinion which is the most appropriate analysis, or does it depend on the question being asked?

Here we assume that analysing vaccination rates at the **state** level refers to analysing the (number of vaccinated people in the state divided by the total number of people in the state). We assume that the goal is to compare and predict vaccination rates among different states.

1) It seems appropriate to do a regression at the state level when we are interested in analysing vaccination at the state level. If we consider the outcome of the total population 18+ who are fully vaccinated, we are considering a count outcome. However, if we are interested in a rate outcome, we would need to divide the outcome by the population. It doesn't seem appropriate to do a regression with a count outcome when we are interested in rates, unless we do further processing.

2) The regression at the state level seems to be the appropriate granularity. I think the outcome measure is appropriate as it is a rate and we are interested in vaccination rates.

3) Regression at the county level seems too granular. It may introduce too much noise if we are really interested in comparing vaccination rates among states. It is also measuring counts rather than rates, so if we are interested in rates, we would have to divide the counts by the appropriate population to get rates.

Overall, I believe that 2) is the most appropriate analysis.