Assignment1

Haoluan Chen

Q1

a)

Assume $E[\theta] = 1$ and $Var(\theta) = \sigma^2$

By the Law of total expectation and $Y|\theta \sim \text{Poisson}(\mu\theta)$:

$E(Y) = E[E(Y|\theta)] = E[\mu\theta] = \mu E[\theta] = \mu$

By the Law of total variance and $Y|\theta \sim \text{Poisson}(\mu\theta)$:

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

b)

Assume $\theta \sim Gamma(\alpha, \beta)0$

$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

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

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

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

$$= \frac{\mu^y}{\beta^\alpha y!\Gamma(\alpha)} \frac{\Gamma(y+\alpha)}{(\mu+1/\beta)^{y+\alpha}}$$

$$= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)} \frac{\mu^y}{\beta^\alpha} (\frac{\mu\beta+1}{\beta})^{-y-\alpha}$$

$$= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)} \frac{\mu^y}{\beta^\alpha} \frac{\beta^{y+\alpha}}{(\mu\beta+1)^{y+\alpha}}$$

$$= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)} (\frac{\mu\beta}{\mu\beta+1})^y (\frac{1}{\mu\beta+1})^\alpha \sim NB(\alpha, \frac{\mu\beta}{\mu\beta+1})$$

c)

Since

$E(Y) = \mu = \frac{\alpha(1-\frac{\mu\beta}{\mu\beta+1})}{\frac{\mu\beta}{\mu\beta+1}} = \alpha\mu\beta => \alpha\beta = 1$

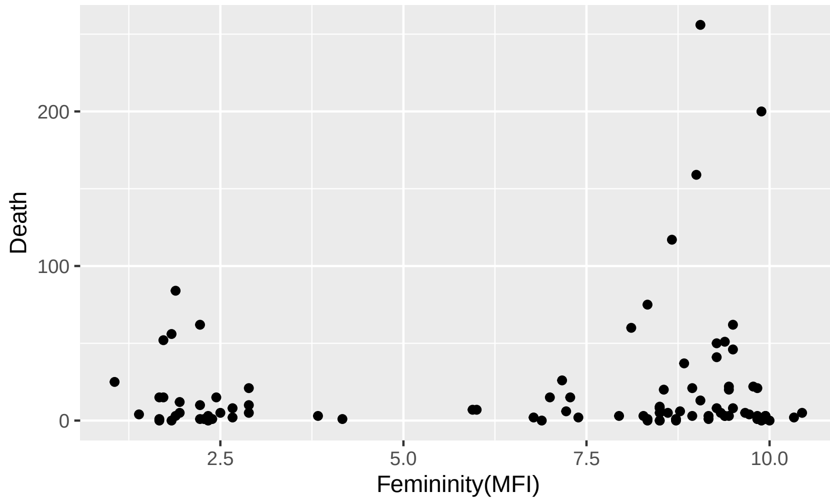
$Var(Y) = \mu(1+\mu\sigma^2) = \frac{\alpha(1-\frac{\mu\beta}{\mu\beta+1})}{(\frac{\mu\beta}{\mu\beta+1})^2} = \alpha\mu\beta + \alpha\mu^2\beta^2 => \alpha\beta^2 = \sigma^2$

Then

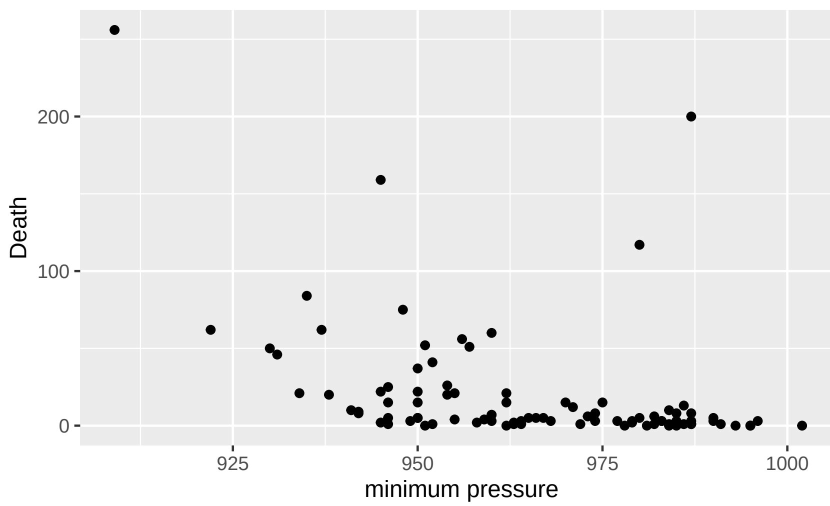
$\alpha = 1/\sigma^2, \beta = \sigma^2$

Q2

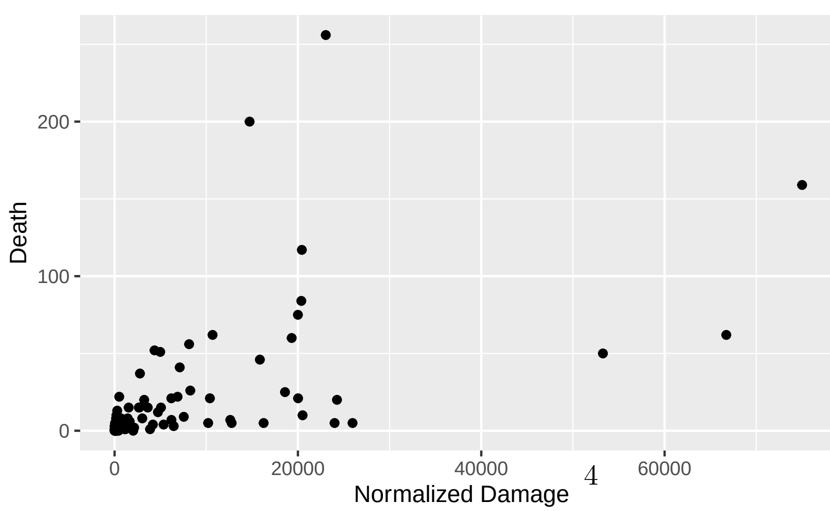
a)



Death
0
100
200
2.5
5.0
7.5
10.0
Femininity(MFI)



Death
0
100
200
925
950
975
1000
minimum pressure



Death
0
100
200
0
20000
40000
60000
Normalized Damage

From the death by femininity scatter plot, it looks like there is two cluster. A group centered around femininity value of 2 and a group centered around femininity value of 8.5. Higher femininity value has higher variability on the number of deaths. One extreme value of over 200 deaths. For minimum pressure, there is a slightly increasing trend as the minimum pressure goes below 950. Lastly, we see an increasing in deaths as normalized damage increase, the variation also increase.

b)

Fitting Poisson model(Estimates are exponentiated) :

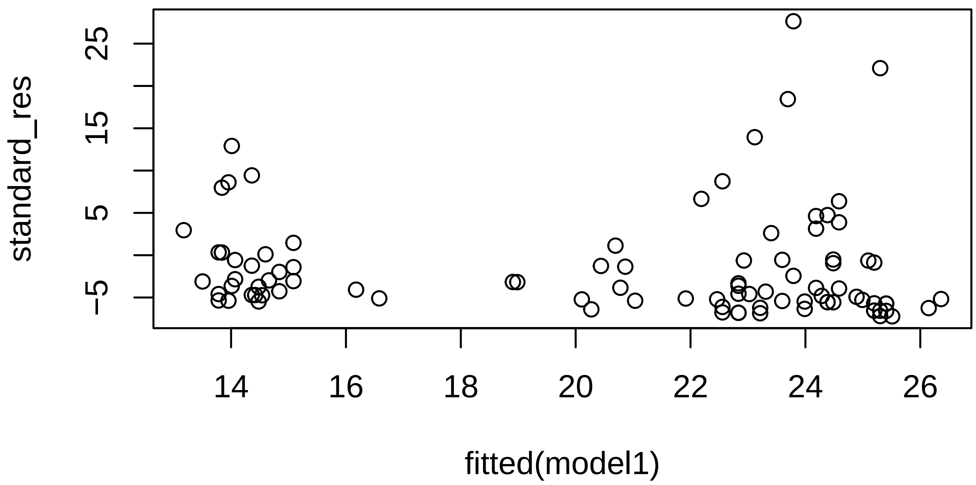
```
model1 <- glm(alldeaths~MasFem, family = poisson, data = q2)
est <- data.frame(summary(model1)$coefficients) %>%
  mutate(Estimate = exp(Estimate))
kable(round(est, 4))
```

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	12.1870	0.0633	39.5021	0
MasFem	1.0767	0.0079	9.3620	0

The poisson model suggested that as the MFI increase by one unit, the death count increase by a factor of 1.0767

Checking for overdispersion:

```
standard_res <- rstandard(model1)
plot(fitted(model1), standard_res)
```




```
n = 92
k = 2
sum(standard_res^2)/(n-k)
```

```
[1] 44.6563
```

```
1-pchisq(sum(standard_res^2), n-k)
```

```
[1] 0
```

There is an overdispersion!

Fitting quasi-poisson model(Estimates are exponentiated):

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	12.1870	0.5437	4.5987	0.0000
MasFem	1.0767	0.0678	1.0899	0.2787

Assuming the significant level to be 0.05. The quasi-poisson suggest that the MFI does not affect on the death count.

c)

Model 4(Estimates are exponentiated):

```
cmodel<-glm.nb(alldeaths ~ ZMasFem*ZMinPressure_A + ZMasFem*ZNDAM , data=q2)
kable(round(summary(cmodel)$coefficients, 4))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.4756	0.1222	20.2605	0.0000
ZMasFem	0.1723	0.1238	1.3918	0.1640
ZMinPressure_A	-0.5521	0.1503	-3.6734	0.0002
ZNDAM	0.8635	0.1445	5.9764	0.0000
ZMasFem:ZMinPressure_A	0.3948	0.1521	2.5952	0.0095
ZMasFem:ZNDAM	0.7051	0.1501	4.6988	0.0000

```
exp(0.1723)
```

```
[1] 1.188034
```

Assuming a hurricane with median pressure and damage ratings, the estimated effect of one unit increase in MFI on death count is 18.8%

d)

```
d <- q2%>% filter(Name == "Sandy")
d
```

```
# A tibble: 1 x 14
   Year Name  MasFem MinPressure~1 Minpr~2 Gende~3 Categ~4 allde~5  NDAM Elaps~6
  <dbl> <chr>  <dbl>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>
1  2012 Sandy      9           945     942       1       2     159 75000       1
# ... with 4 more variables: Source <chr>, ZMasFem <dbl>, ZMinPressure_A <dbl>,
#   ZNDAM <dbl>, and abbreviated variable names 1: MinPressure_before,
#   2: `Minpressure_Updated 2014`, 3: Gender_MF, 4: Category, 5: alldeaths,
#   6: `Elapsed Yrs`
```

```r
sandy <- d[12:14]
predict(cmodel, sandy, type="response")
```

```
       1
20806.74
```

The predicted death count for Sandy is 20807. However, the actual death count is only 159. The predicted death count is so high because Sandy has highest damage.

e)

weakness:

1. Only 9 independent coder were include in determine the MFI, which may be biased. More coder can be included.
2.

strength:

1. Recognizing the confounding variable: effect of gendered names on protective action, not simply conclude that Feminine-named hurricanes cause significantly more deaths.
2. Many experiment were carried out to test difference aspect about the perceived risk of the hurricanes, predicted intensity and evacuation intention. This wide range of experiment helps convince reader that gendered hurricanes names will affect how people feel and act.
3. Data set are available for reproducible

f)

I think I'm convinced by the results,

Q3

Loading and combining two datasets

```r
q3 <- read_csv("data/q3.csv")
```

```
Rows: 3283 Columns: 80
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (6): Date, FIPS, Recip_County, Recip_State, SVI_CTGY, Metro_status
dbl (47): MMWR_week, Completeness_pct, Administered_Dose1_Pop_Pct, Administe...
num (27): Administered_Dose1_Recip, Administered_Dose1_Recip_5Plus, Administ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
q3 <- dplyr::select(q3, FIPS, starts_with("Series_Complete"), )
acs <- read_csv("data/acs.csv")
```

```
Rows: 37704 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): fips, county_name, variable
dbl (1): value

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
acs <- acs %>% pivot_wider(names_from = variable, values_from=value)
acs <- acs %>% rename(FIPS = fips)
combined <- inner_join(q3, acs, by='FIPS')
```

a)

```
skim(acs)
```

Table 4: Data summary

Name	acs
Number of rows	3142
Number of columns	14
Column type frequency:	
character	2

Table 4: Data summary

numeric	12
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
FIPS	0	1	5	5	0	3142	0
county_name	0	1	16	42	0	3142	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
total_pop_18plus	0	1	79970.85	255904.46	65.00	8438.50	20154.50	52646.50	7866810.00	
prop_white	0	1	0.83	0.17	0.04	0.76	0.90	0.95	1.00	
prop_foreign_born	0	1	0.06	0.07	0.00	0.02	0.03	0.07	0.64	
median_age	0	1	41.43	5.42	22.30	38.20	41.30	44.50	67.40	
median_income	0	1	53475.91	14192.53	21504.00	44155.00	51757.50	59867.25	142299.00	
median_rent	4	1	774.54	226.89	313.00	633.00	716.00	851.00	2316.00	
prop_less_than_hs	0	1	0.13	0.06	0.01	0.08	0.12	0.17	0.74	
prop_bachelor_above	0	1	0.22	0.10	0.00	0.15	0.20	0.26	0.78	
prop_unemployed	0	1	0.03	0.01	0.00	0.02	0.03	0.04	0.16	
prop_nilf	0	1	0.42	0.08	0.17	0.36	0.41	0.47	0.85	
prop_health_insurance	0	1	0.90	0.05	0.54	0.88	0.91	0.94	1.00	
prop_low_ratio_ip	0	1	0.16	0.07	0.00	0.11	0.15	0.19	0.58	

```r
skim(q3)
```

Table 7: Data summary

Name	q3
Number of rows	3283
Number of columns	25
Column type frequency:	
character	1

Table 7: Data summary

numeric	24
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
FIPS	0	1	3	5	0	3225	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Series_Complete_Yes	17	0.99	69561.55	242131.74	32.0	5401.50	13727.5	39787.75	7502592	
Series_Complete_Pop_Pct	78	0.98	53.81	13.48	11.3	44.20	52.2	61.60	95	
Series_Complete_5Plus	17	0.99	69263.45	240854.47	32.0	5398.75	13714.5	39738.25	7465445	
Series_Complete_5PlusPop_Pct	78	0.98	56.90	13.97	12.1	46.90	55.3	65.30	95	
Series_Complete_5to17	17	0.99	7539.79	29730.44	2.0	328.25	972.5	3337.00	929501	
Series_Complete_5to17Pop_Pct	78	0.98	28.17	17.48	1.4	16.00	23.3	34.90	95	
Series_Complete_12Plus	17	0.99	66440.82	229753.61	32.0	5306.25	13431.5	38572.75	7132323	
Series_Complete_12PlusPop_Pct	78	0.98	60.74	13.92	13.2	50.90	59.6	69.30	95	
Series_Complete_18Plus	17	0.99	61723.62	211637.23	30.0	5056.50	12685.5	36343.00	6535944	
Series_Complete_18PlusPop_Pct	78	0.98	62.81	13.52	13.9	53.40	62.0	71.30	95	
Series_Complete_65Plus	17	0.99	15670.78	44814.17	9.0	1755.25	4177.5	11113.50	1253400	
Series_Complete_65PlusPop_Pct	78	0.98	80.99	11.95	17.6	73.80	83.0	90.70	95	
Series_Complete_Pop_Pct_SVI	79	0.98	7.81	4.47	1.0	4.75	9.0	13.00	16	
Series_Complete_5PlusPop_Pct_SVI	79	0.98	7.99	4.48	1.0	4.75	9.0	13.00	16	
Series_Complete_5to17Pop_Pct_SVI	79	0.98	7.64	4.54	1.0	4.75	9.0	13.00	16	
Series_Complete_12PlusPop_Pct_SVI	79	0.98	8.24	4.50	1.0	4.75	9.0	13.00	16	
Series_Complete_18PlusPop_Pct_SVI	79	0.98	8.37	4.49	1.0	4.75	9.0	13.00	16	
Series_Complete_65PlusPop_Pct_SVI	79	0.98	9.47	4.46	1.0	4.75	9.0	13.00	16	
Series_Complete_Pop_Pct_UR_Equity	80	0.98	4.27	1.87	1.0	2.00	5.0	6.00	8	
Series_Complete_5PlusPop_Pct_UR_Equity	80	0.98	4.45	1.87	1.0	3.00	5.0	6.00	8	
Series_Complete_5to17Pop_Pct_UR_Equity	80	0.98	4.10	1.85	1.0	2.00	5.0	5.00	8	
Series_Complete_12PlusPop_Pct_UR_Equity	80	0.98	4.69	1.87	1.0	3.00	5.0	6.00	8	
Series_Complete_18PlusPop_Pct_UR_Equity	80	0.98	4.83	1.88	1.0	3.00	5.0	6.00	8	
Series_Complete_65PlusPop_Pct_UR_Equity	80	0.98	5.93	1.91	1.0	4.00	7.0	8.00	8	

```r
skim(combined)
```

Table 10: Data summary

Name	combined
Number of rows	3142
Number of columns	38
Column type frequency:	
character	2
numeric	36
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
FIPS	0	1	5	5	0	3142	0
county_name	0	1	16	42	0	3142	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Series_Complete_Yes	16	0.99	68682.92	244993.43	12.00	5198.75	12971.00	38067.00	7502592.00	
Series_Complete_Pop_Pct	16	0.99	53.13	12.90	11.30	43.92	51.80	60.70	95.00	
Series_Complete_5Plus	16	0.99	68384.92	243713.73	73.00	5194.00	12934.50	38032.25	7465445.00	
Series_Complete_5PlusPop_Pct	16	0.99	56.22	13.44	12.10	46.70	54.80	64.20	95.00	
Series_Complete_5to17	16	0.99	7457.05	29904.48	2.00	313.25	906.00	3052.75	929501.00	
Series_Complete_5to17Pop_Pct	16	0.99	26.92	15.74	1.40	15.90	22.90	33.80	95.00	
Series_Complete_12Plus	16	0.99	65630.38	232586.67	12.00	5101.50	12693.50	37089.50	7132323.00	
Series_Complete_12PlusPop_Pct	16	0.99	60.14	13.52	13.20	50.70	59.10	68.40	95.00	
Series_Complete_18Plus	16	0.99	60927.86	214192.70	30.00	4859.00	11959.50	34746.50	6535944.00	
Series_Complete_18PlusPop_Pct	16	0.99	62.26	13.20	13.90	53.10	61.55	70.40	95.00	
Series_Complete_65Plus	16	0.99	15536.12	45319.52	9.00	1697.50	4064.50	10949.00	1253400.00	
Series_Complete_65PlusPop_Pct	16	0.99	80.88	12.04	17.60	73.60	82.90	90.80	95.00	
Series_Complete_Pop_Pct_SVI	16	0.99	7.76	4.47	1.00	5.00	9.00	13.00	16.00	
Series_Complete_5PlusPop_Pct_SVI	16	0.99	7.95	4.47	1.00	5.00	9.00	13.00	16.00	
Series_Complete_5to17Pop_Pct_SVI	16	0.99	7.58	4.52	1.00	5.00	9.00	13.00	16.00	
Series_Complete_12PlusPop_Pct_SVI	16	0.99	8.20	4.49	1.00	5.00	9.00	13.00	16.00	

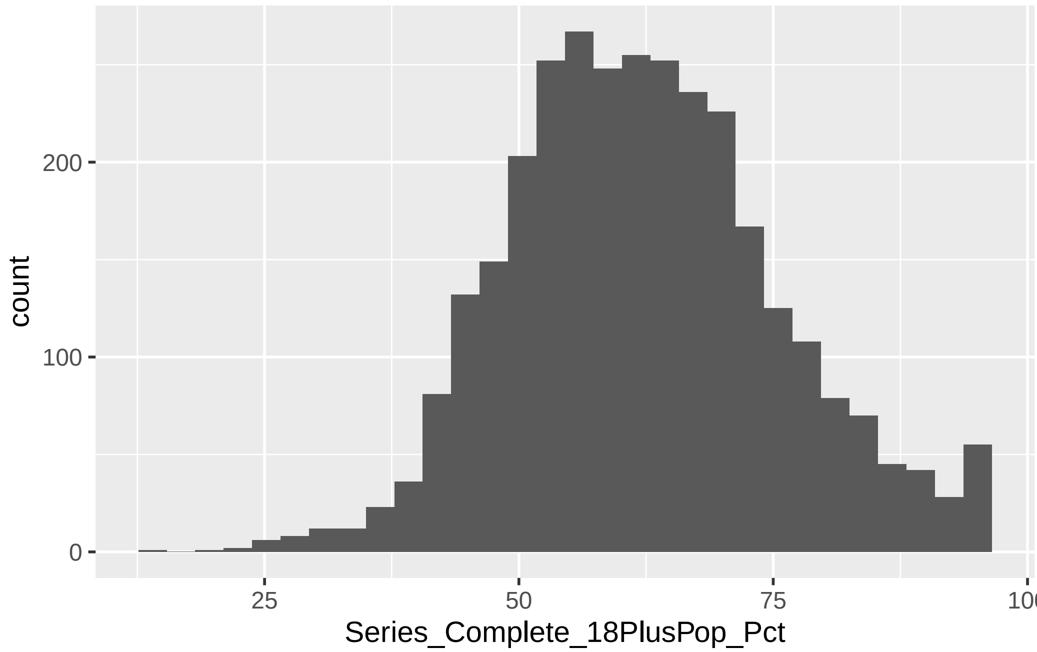
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Series_Complete_18PlusPop_Pct_SVI	16	0.99	8.34	4.48	1.00	5.00	9.00	13.00	16.00	
Series_Complete_65PlusPop_Pct_SVI	16	0.99	9.47	4.46	1.00	5.00	9.00	13.00	16.00	
Series_Complete_Pop_Pct_UR_Equity	17	0.99	4.27	1.88	1.00	2.00	5.00	6.00	8.00	
Series_Complete_5PlusPop_Pct_UR_Equity	17	0.99	4.46	1.88	1.00	3.00	5.00	6.00	8.00	
Series_Complete_5to17Pop_Pct_UR_Equity	17	0.99	4.09	1.86	1.00	2.00	5.00	5.00	8.00	
Series_Complete_12PlusPop_Pct_UR_Equity	17	0.99	4.71	1.88	1.00	3.00	5.00	6.00	8.00	
Series_Complete_18PlusPop_Pct_UR_Equity	17	0.99	4.85	1.89	1.00	3.00	5.00	6.00	8.00	
Series_Complete_65PlusPop_Pct_UR_Equity	17	0.99	5.97	1.91	1.00	4.00	7.00	8.00	8.00	
total_pop_18plus	0	1.00	79970.82	255904.46	65.00	8438.50	20154.50	52646.50	7866810.00	
prop_white	0	1.00	0.83	0.17	0.04	0.76	0.90	0.95	1.00	
prop_foreign_born	0	1.00	0.06	0.07	0.00	0.02	0.03	0.07	0.64	
median_age	0	1.00	41.43	5.42	22.30	38.20	41.30	44.50	67.40	
median_income	0	1.00	53475.91	14192.53	31504.00	44155.00	51757.50	59867.25	142299.00	
median_rent	4	1.00	774.54	226.89	313.00	633.00	716.00	851.00	2316.00	
prop_less_than_hs	0	1.00	0.13	0.06	0.01	0.08	0.12	0.17	0.74	
prop_bachelor_above	0	1.00	0.22	0.10	0.00	0.15	0.20	0.26	0.78	
prop_unemployed	0	1.00	0.03	0.01	0.00	0.02	0.03	0.04	0.16	
prop_nilf	0	1.00	0.42	0.08	0.17	0.36	0.41	0.47	0.85	
prop_health_insurance	0	1.00	0.90	0.05	0.54	0.88	0.91	0.94	1.00	
prop_low_ratio_ip	0	1.00	0.16	0.07	0.00	0.11	0.15	0.19	0.58	

```r
combined <- na.omit(combined)
```

Check distribution of the count

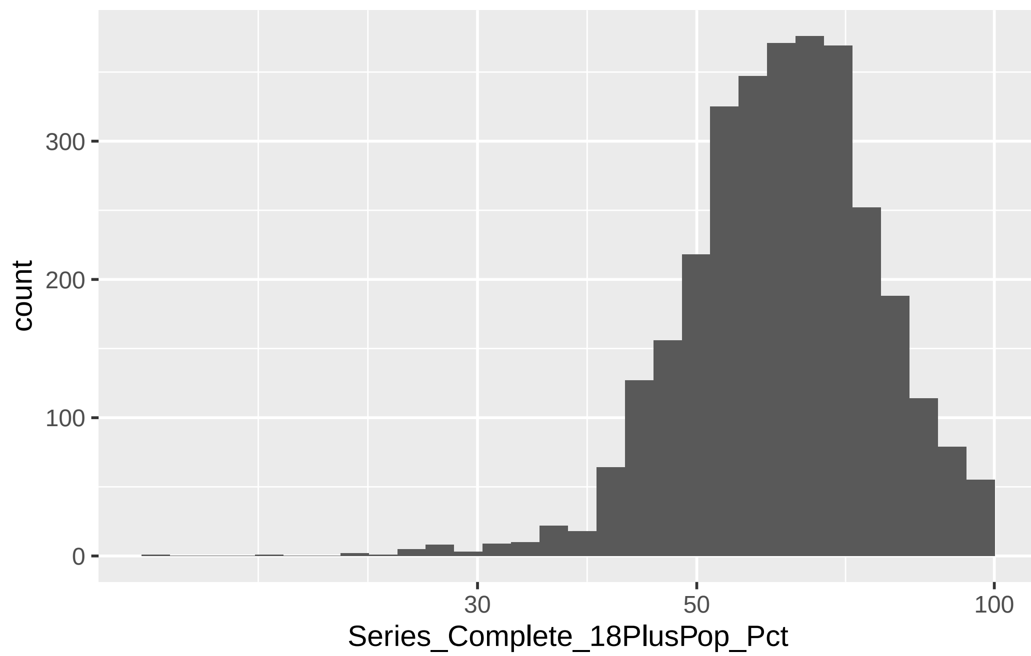
```r
combined %>% ggplot(aes(Series_Complete_18PlusPop_Pct))+ geom_histogram()
```

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




```
combined %>% ggplot(aes(Series_Complete_18PlusPop_Pct))+ geom_histogram() + scale_x_log10(
```

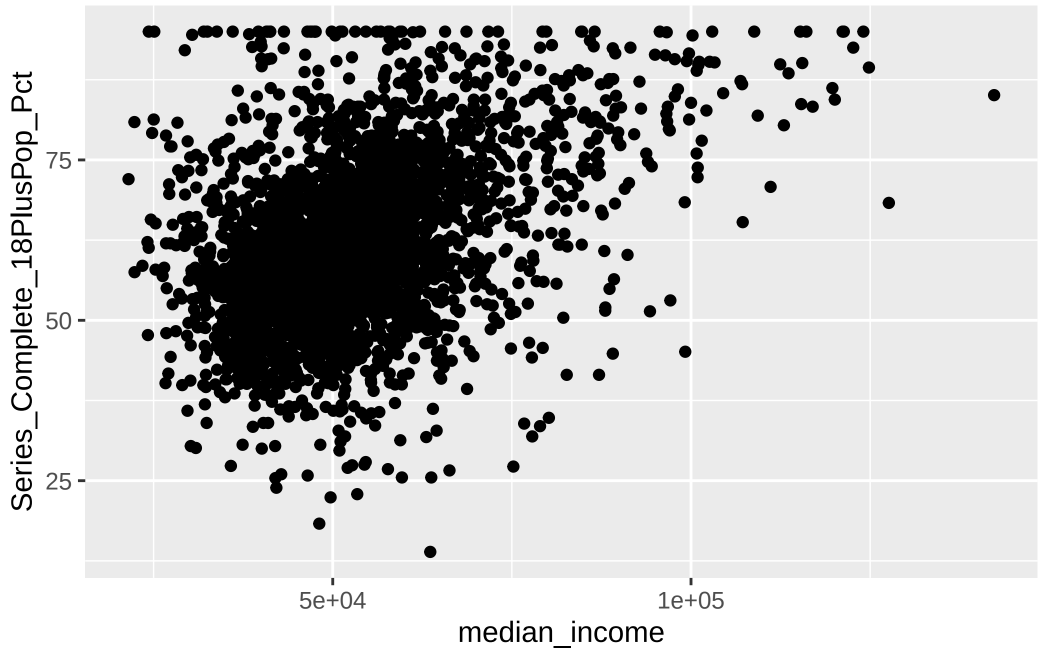
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



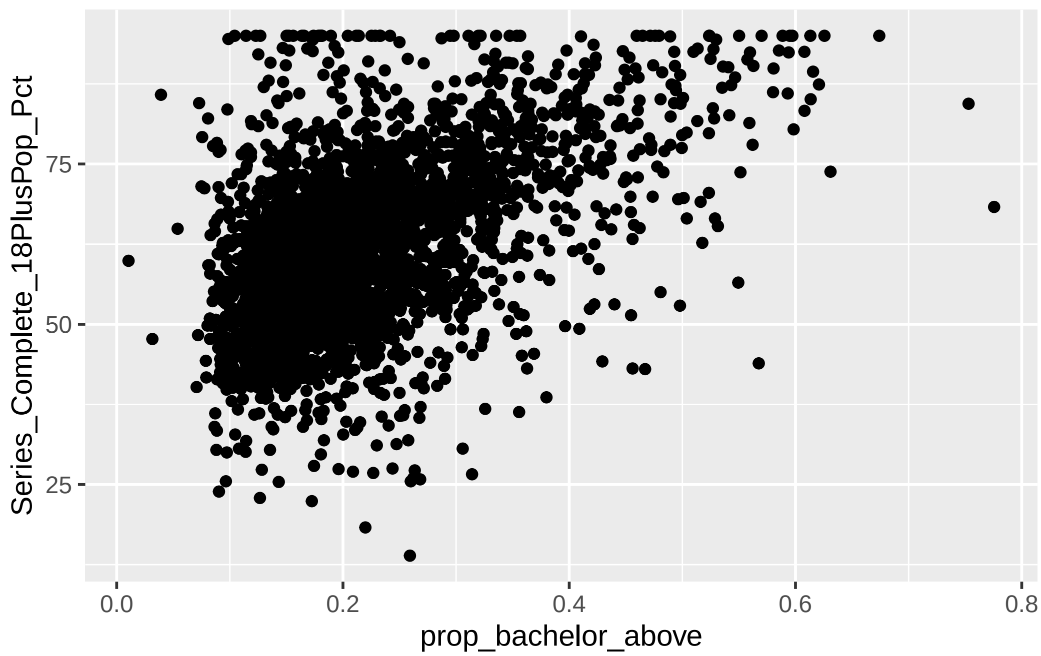
log transformation makes the count close to normal distribution, maybe I should consider log-linear model?

Complete count vs prop_white

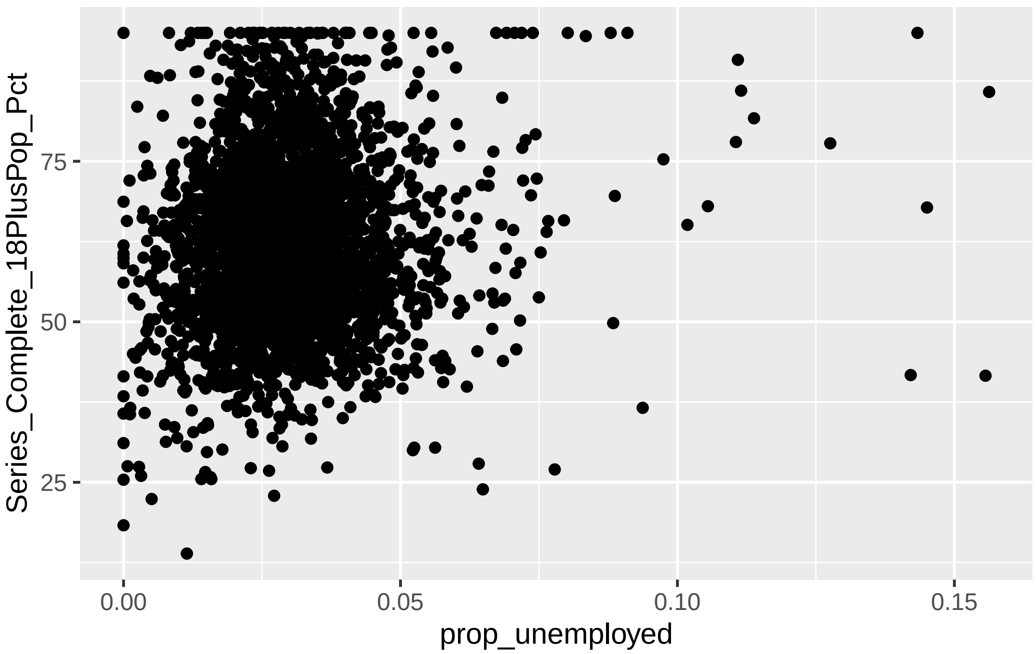
```
combined %>% ggplot(aes(x=median_income, y=Series_Complete_18PlusPop_Pct)) + geom_point()
```



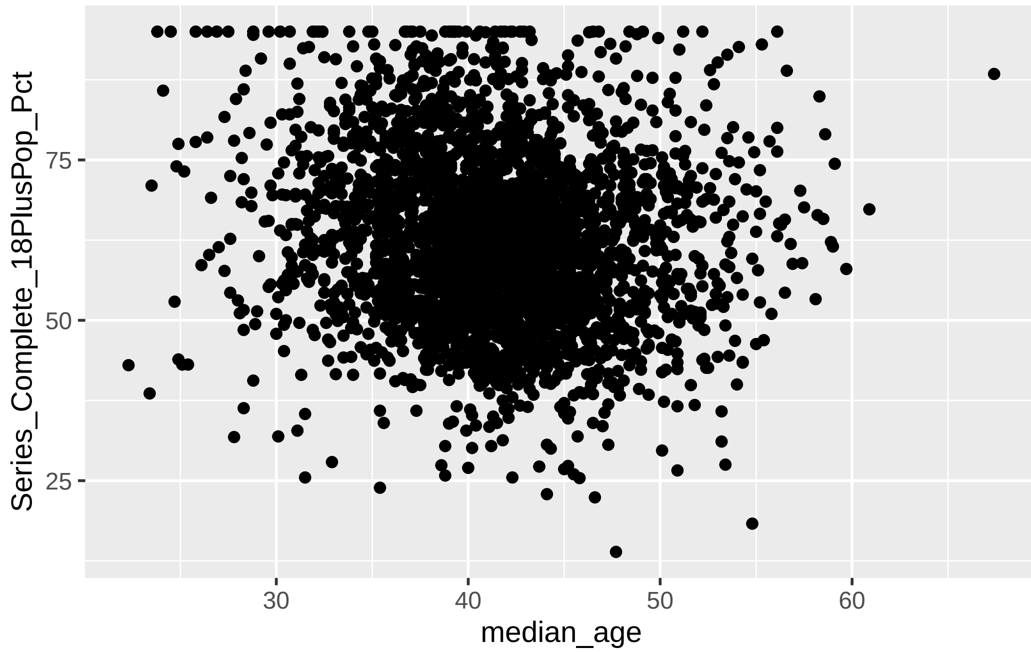

```
combined %>% ggplot(aes(x=prop_bachelor_above, y=Series_Complete_18PlusPop_Pct)) + geom_po
```



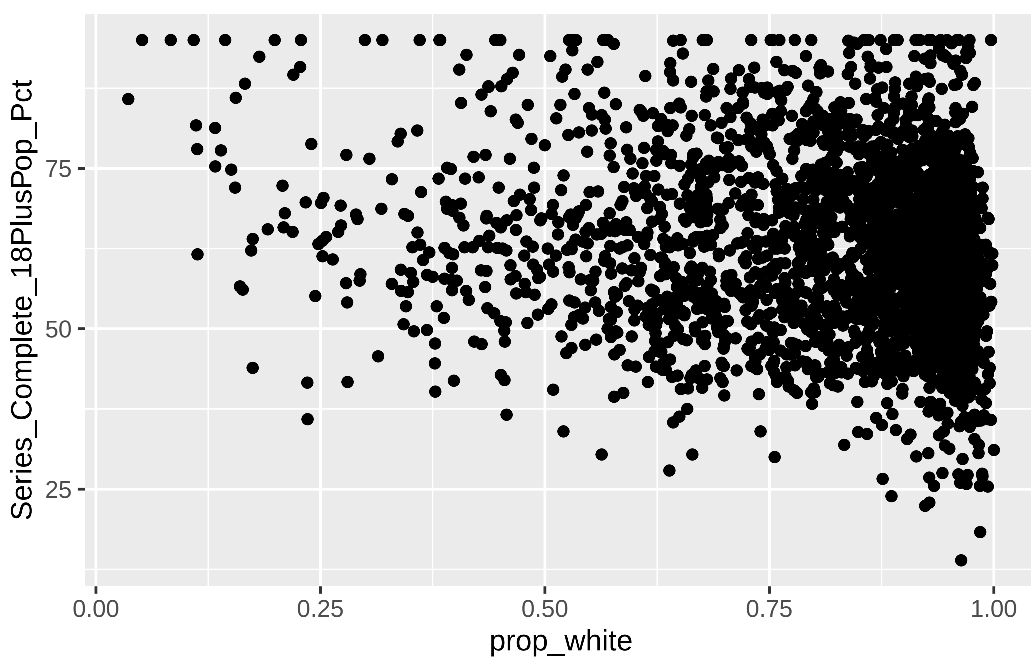
```
combined %>% ggplot(aes(x=prop_unemployed, y=Series_Complete_18PlusPop_Pct)) + geom_point(
```



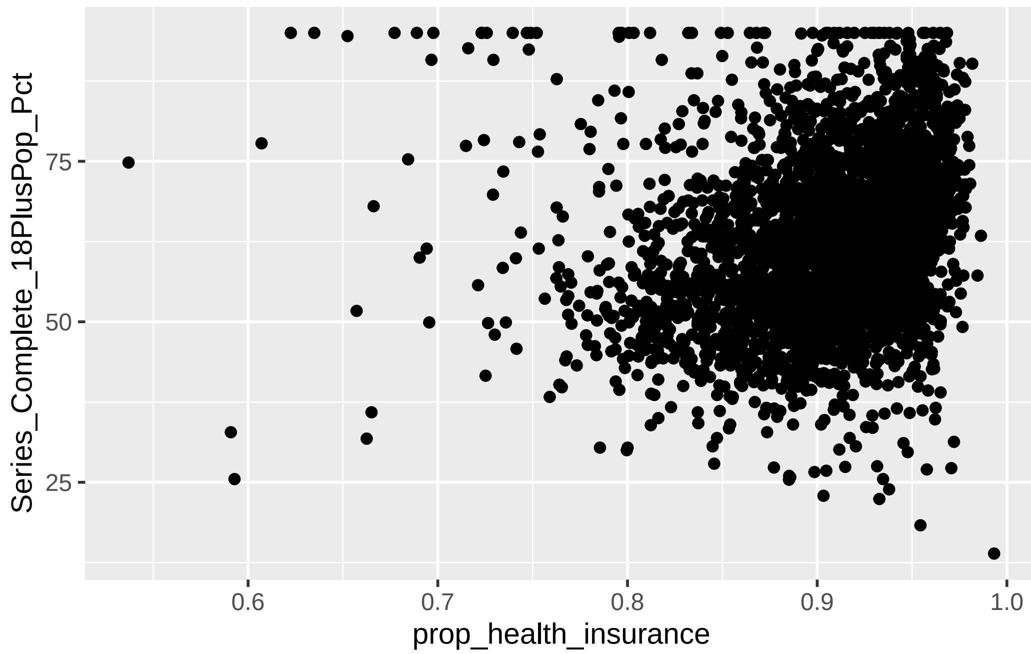

```
combined %>% ggplot(aes(x=median_age, y=Series_Complete_18PlusPop_Pct)) + geom_point()
```



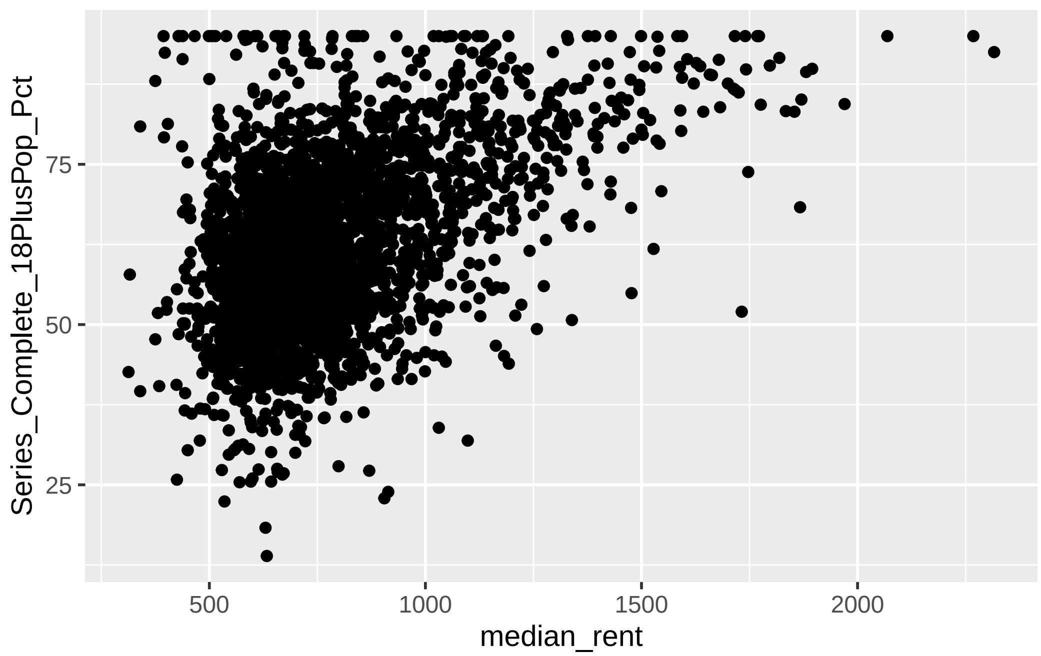

```
combined %>% ggplot(aes(x=prop_white, y=Series_Complete_18PlusPop_Pct)) + geom_point()
```



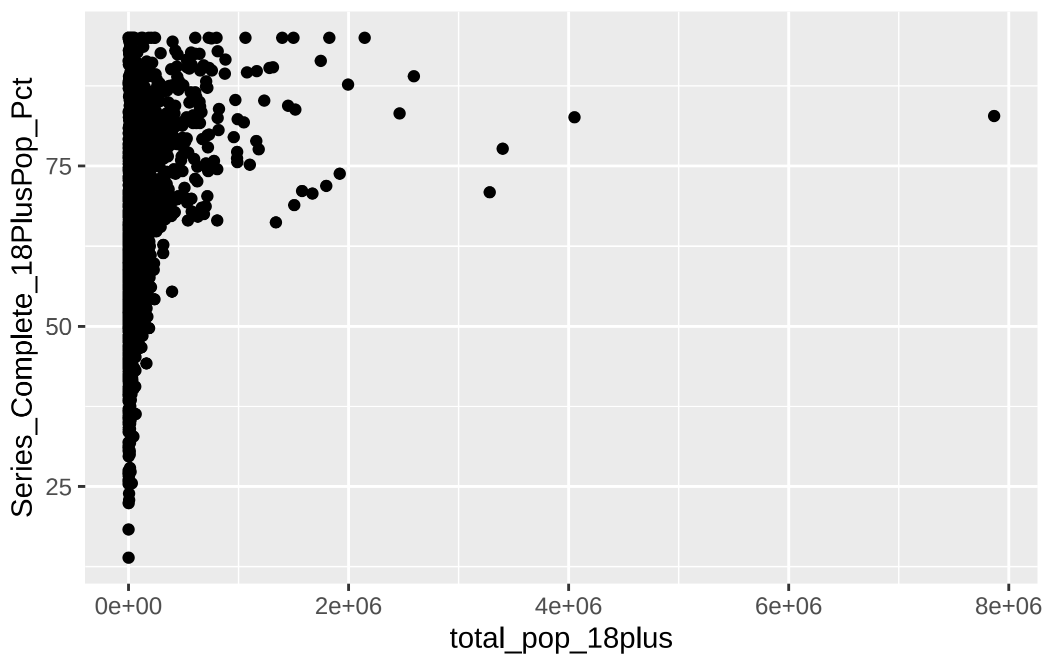
```
combined %>% ggplot(aes(x=prop_health_insurance, y=Series_Complete_18PlusPop_Pct)) + geom_
```



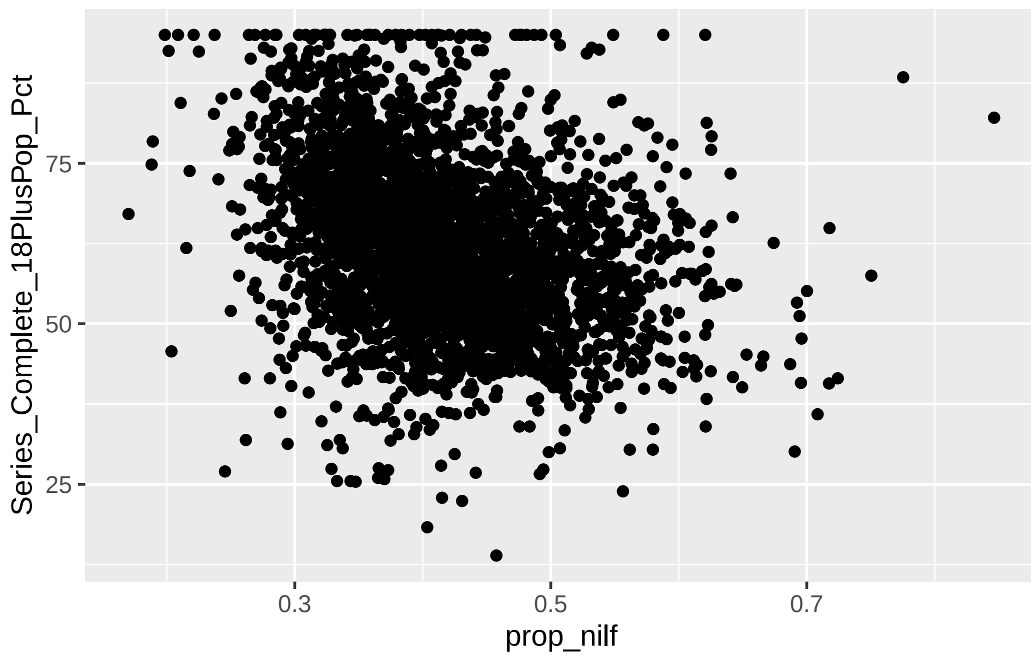

```
combined %>% ggplot(aes(x=median_rent, y=Series_Complete_18PlusPop_Pct)) + geom_point()
```



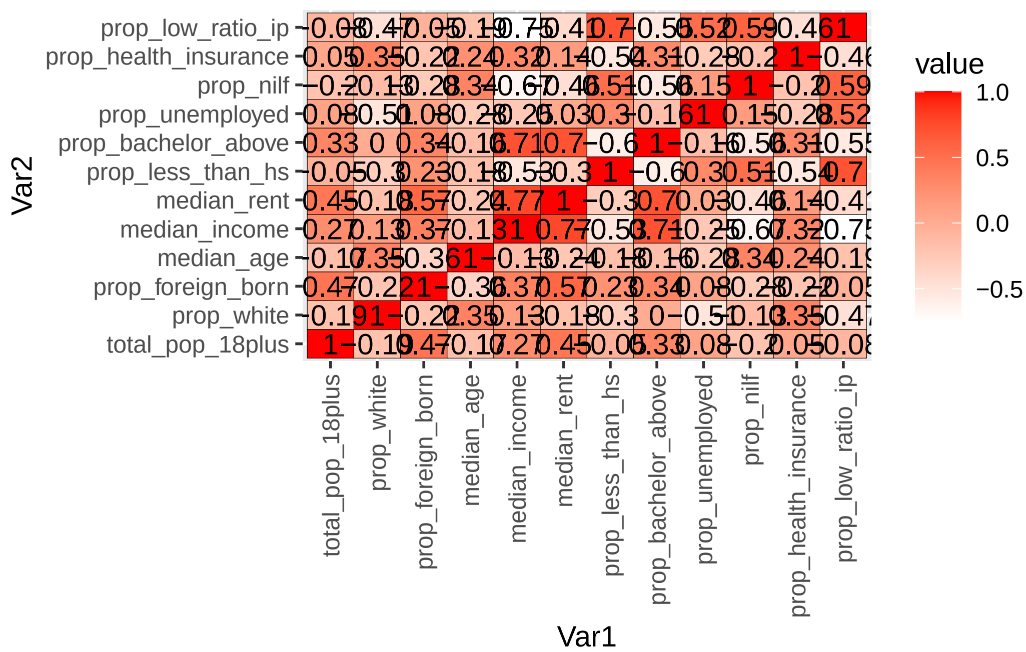

```
combined %>% ggplot(aes(x=total_pop_18plus, y=Series_Complete_18PlusPop_Pct)) + geom_point
```



```
combined %>% ggplot(aes(x=prop_nilf, y=Series_Complete_18PlusPop_Pct)) + geom_point()
```




```
correlationacs <- dplyr::select(acs, -FIPS, -county_name)
correlationacs<- na.omit(correlationacs)
# creating correlation matrix
corr_mat <- round(cor(correlationacs),2)
# reduce the size of correlation matrix
melted_corr_mat <- melt(corr_mat)
# head(melted_corr_mat)
# plotting the correlation heatmap
ggplot(data = melted_corr_mat, aes(x=Var1, y=Var2,
                                    fill=value)) +
  geom_tile(color = "black") +
  scale_fill_gradient(low = "white", high = "red") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  geom_text(aes(label = value), color = "black", size = 4)
```



b)

```
q3
```

```
# A tibble: 3,283 x 25
   FIPS  Serie~1 Serie~2 Serie~3 Serie~4 Serie~5 Serie~6 Serie~7 Serie~8 Serie~9
   <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
 1 38079   11316    79.8   11283    87.1    2594    73     10057    91.8    8689
 2 48391    3750    54      3749    57.2     323    28.4    3675    61.9    3426
 3 53025   57360    58.7   57325    63.6    6465    30.6   55453    70.5   50860
 4 19183   12872    58.6   12822    62.5    1184    30.2   12399    67.4   11638
 5 48313    5981    41.9    5981    44.4     276    12.4    5910    48.1    5705
 6 20105    1330    44.9    1329    47.1      66    13.2    1309    51      1263
 7 29103    1629    41.1    1627    44        94    14      1606    47.9    1533
 8 55035   66817    63.9   66290    67.1    7574    49.6   62925    69.4   58716
 9 51097    3683    52.4    3680    54.9     233    25.7    3593    57.7    3447
10 48029 1452251    72.5 1450278    77.8  186695    50.8 1389964    83.4 1263583
# ... with 3,273 more rows, 15 more variables:
#   Series_Complete_18PlusPop_Pct <dbl>, Series_Complete_65Plus <dbl>,
#   Series_Complete_65PlusPop_Pct <dbl>, Series_Complete_Pop_Pct_SVI <dbl>,
#   Series_Complete_5PlusPop_Pct_SVI <dbl>,
```

```
#   Series_Complete_5to17Pop_Pct_SVI <dbl>,
#   Series_Complete_12PlusPop_Pct_SVI <dbl>,
#   Series_Complete_18PlusPop_Pct_SVI <dbl>, ...
```

```
modeldata3b <- combined %>% mutate(Series_Complete_18PlusPop_Pct_model = Series_Complete_1
model3b1 <- glm(Series_Complete_18PlusPop_Pct_model ~ prop_white + prop_foreign_born+
                 median_income +prop_unemployed + prop_nilf +
                 prop_health_insurance + prop_low_ratio_ip,
               family = binomial, data = modeldata3b)
```

```
Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(model3b1)
```

```

Call:
glm(formula = Series_Complete_18PlusPop_Pct_model ~ prop_white +
    prop_foreign_born + median_income + prop_unemployed + prop_nilf +
    prop_health_insurance + prop_low_ratio_ip, family = binomial,
    data = modeldata3b)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.03921  -0.13545  -0.00147   0.13379   1.12401

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.637e+00  9.802e-01  -2.690 0.007144 **
prop_white            -6.614e-01  2.908e-01  -2.274 0.022964 *
prop_foreign_born      2.374e+00  7.013e-01   3.386 0.000710 ***
median_income          1.145e-05  5.507e-06   2.080 0.037552 *
prop_unemployed        5.345e+00  3.493e+00   1.530 0.125943
prop_nilf             -4.858e-01  6.571e-01  -0.739 0.459705
prop_health_insurance  3.269e+00  8.815e-01   3.709 0.000208 ***
prop_low_ratio_ip      2.877e-01  1.120e+00   0.257 0.797294
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 247.53  on 3120  degrees of freedom
Residual deviance: 172.04  on 3113  degrees of freedom
AIC: 3330.1

Number of Fisher Scoring iterations: 4
```

```r
model3b2 <- glm(Series_Complete_18PlusPop_Pct_model ~ prop_white + prop_foreign_born+
                 median_income +prop_unemployed +
                 prop_health_insurance,
               family = binomial, data = modeldata3b)
```

```
Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
summary(model3b2)
```

```
Call:
glm(formula = Series_Complete_18PlusPop_Pct_model ~ prop_white +
    prop_foreign_born + median_income + prop_unemployed + prop_health_insurance,
    family = binomial, data = modeldata3b)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0532  -0.1355  -0.0020   0.1342   1.1398

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.818e+00  7.610e-01  -3.704 0.000212 ***
prop_white            -6.705e-01  2.745e-01  -2.442 0.014590 *
prop_foreign_born      2.439e+00  6.868e-01   3.551 0.000384 ***
median_income          1.239e-05  3.447e-06   3.595 0.000324 ***
prop_unemployed        5.794e+00  3.269e+00   1.772 0.076340 .
prop_health_insurance  3.231e+00  8.718e-01   3.706 0.000211 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 247.53  on 3120  degrees of freedom
Residual deviance: 172.59  on 3115  degrees of freedom
```

```
AIC: 3329.5

Number of Fisher Scoring iterations: 4
```

c)

```r
Ada <- modeldata3b %>% filter(county_name == "Ada County, Idaho")
Ada
```

```
# A tibble: 1 x 39
  FIPS   Series~1 Serie~2 Serie~3 Serie~4 Serie~5 Serie~6 Serie~7 Serie~8 Serie~9
  <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 16001    323707    67.2  322173    70.9   37693    44.6  307890      75  284480
# ... with 29 more variables: Series_Complete_18PlusPop_Pct <dbl>,
#   Series_Complete_65Plus <dbl>, Series_Complete_65PlusPop_Pct <dbl>,
#   Series_Complete_Pop_Pct_SVI <dbl>, Series_Complete_5PlusPop_Pct_SVI <dbl>,
#   Series_Complete_5to17Pop_Pct_SVI <dbl>,
#   Series_Complete_12PlusPop_Pct_SVI <dbl>,
#   Series_Complete_18PlusPop_Pct_SVI <dbl>,
#   Series_Complete_65PlusPop_Pct_SVI <dbl>, ...
```

```r
Ada <- dplyr::select(Ada, prop_white, prop_foreign_born,
                 median_income,prop_unemployed, prop_nilf,
                 prop_health_insurance, prop_low_ratio_ip, total_pop_18plus)
Ada
```

```
# A tibble: 1 x 8
  prop_white prop_foreign_born median_~1 prop_~2 prop_~3 prop_~4 prop_~5 total~6
       <dbl>             <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1      0.905            0.0698     66293  0.0257   0.336   0.917  0.0902  347052
# ... with abbreviated variable names 1: median_income, 2: prop_unemployed,
#   3: prop_nilf, 4: prop_health_insurance, 5: prop_low_ratio_ip,
#   6: total_pop_18plus
```

```r
dplyr::select(modeldata3b, county_name, Series_Complete_18PlusPop_Pct) %>%
  filter(county_name == "Ada County, Idaho")
```

```
# A tibble: 1 x 2
  county_name        Series_Complete_18PlusPop_Pct
  <chr>                                      <dbl>
1 Ada County, Idaho                           76.9
```

```
predict(model3b2, Ada, type = "response")
```

```
        1
0.6635257
```

The prediction is about 10% off. I guess it is pretty good considering the variability in the data

d)

e)