

Q1:

1. Consider the following contingency table

Treatment	Controls	Cases	
<i>A</i>	Y_{11}	Y_{12}	N_1
<i>B</i>	Y_{21}	Y_{22}	N_2

In this question, we show how you could use Poisson regression to test for an effect of treatment on the proportion of cases and controls.

Normally we could use logisitic regression, and so we will show that these are equivalent.

Poisson model

Let

$$Y_{ij} \sim Po(\lambda_{ij})$$

and

$$\log(\lambda_{ij}) = \gamma_0 + \gamma_1 x_i + \gamma_2 x_j + \gamma_3 x_i x_j,$$

where

$$x_i = \begin{cases} 0 & \text{if treatment is } A, \\ 1 & \text{if treatment is } B \end{cases}$$

and

$$x_j = \begin{cases} 0 & \text{if controls,} \\ 1 & \text{if cases} \end{cases}$$

a)

(a) Write in terms of the γ s, equations for

- i. $\log(\lambda_{11})$,
- ii. $\log(\lambda_{12})$,
- iii. $\log(\lambda_{21})$, and
- iv. $\log(\lambda_{22})$.

$$y_{ij} \sim \text{pois}(\lambda_{ij})$$

$$\log(\lambda_{ij}) = \gamma_0 + \gamma_1 x_i + \gamma_2 x_j + \gamma_3 x_i x_j$$

$$\log(\lambda_{11}) = \gamma_0 + \gamma_1 \overset{0}{x_1} + \gamma_2 \overset{0}{x_j} + \underbrace{\gamma_3 \overset{0}{x_1} \overset{0}{x_j}}_{\text{interaction term}}$$

$$= \gamma_0 + \gamma_1$$

$$\begin{aligned}\log(\lambda_{12}) &= r_0 + r_1 \cdot 0 + r_2 \cdot 1 + r_3 \cdot 0 \cdot 1 \\ &= r_0 + r_2\end{aligned}$$

$$\log(\lambda_{21}) = r_0 + r_1$$

$$\begin{aligned}\log(\lambda_{22}) &= r_0 + r_1 \cdot 1 + r_2 \cdot 1 + r_3 \cdot 1 \cdot 1 \\ &= r_0 + r_1 + r_2 + r_3\end{aligned}$$

(b) Show that the conditional distribution of

$$Y_{i2} | Y_{i1} + Y_{i2} = n_i$$

is binomial $B(n_i, \pi_i)$ and find an expression for π_i in terms of λ_{i1} and λ_{i2} .

show $Y_{i2} | Y_{i1} + Y_{i2} = n_i$ is
Binomial (n_i, π_i)

show $P(Y_{i2} = k | Y_{i1} + Y_{i2} = n_i)$
for some binomial probability.

$$= \frac{P(Y_{i2} = k \cap Y_{i1} + Y_{i2} = n_i)}{P(Y_{i1} + Y_{i2} = n_i)}$$

$$= \frac{P(Y_{i2} = k) \cdot P(Y_{i1} = n - k)}{e^{-(\lambda_{i1} + \lambda_{i2})} (\lambda_{i1} + \lambda_{i2})^{n_i}} \cdot \frac{n_i!}{k!(n-k)!}$$

$$= \frac{\frac{e^{-\lambda_{i2}} \lambda_{i2}^k}{k!} \cdot \frac{e^{-\lambda_{i1}} \lambda_{i1}^{(n-k)}}{(n-k)!}}{\frac{e^{-(\lambda_{i1} + \lambda_{i2})} (\lambda_{i1} + \lambda_{i2})}{n!}}$$

$$= \frac{\frac{\cancel{e^{-\lambda_{i2}}} \lambda_{i2}^k}{k!} \cdot \frac{\cancel{e^{-\lambda_{i1}}} \lambda_{i1}^{(n-k)}}{(n-k)!}}{\frac{\cancel{e^{-(\lambda_{i1} + \lambda_{i2})}} (\lambda_{i1} + \lambda_{i2})}{n!}}$$

$$= \frac{n!}{(n-k)! k!} \cdot \left(\frac{\lambda_{i2}}{\lambda_{i1} + \lambda_{i2}} \right)^k \left(\frac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}} \right)^{n-k}$$

$$\Rightarrow \pi_i = \frac{\lambda_{i2}}{\lambda_{i1} + \lambda_{i2}}$$

(c) Let

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i$$

c)

where

$$x_i = \begin{cases} 0 & \text{if treatment is A,} \\ 1 & \text{if treatment is B} \end{cases}$$

Hence, or otherwise, write an expression for β_1 in terms of γ s.

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i$$

$$\frac{\pi_i}{1 - \pi_i} = \frac{\lambda_{i2}}{\lambda_{i1} + \lambda_{i2}}$$
$$1 - \pi_i = \frac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}}$$

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i = \underbrace{\log(\lambda_{i2})}_{\beta_0 + \beta_1 x_i} - \underbrace{\log(\lambda_{i1})}_{\beta_0 + \beta_1 x_i}$$

$$= r_0 + r_1 x_i + r_2 \cdot 1 + r_3 \cdot x_i \cdot 1$$
$$- (r_0 + r_1 x_i + 0 + 0)$$

$$= r_2 + r_3 x_i$$

$$\beta_0 + \beta_1 x_i$$

$$\Rightarrow \beta_1 = r_3$$

d)

- (d) Hence, or otherwise, show that testing that treatment has no effect on the probability of being a case, is equivalent for test for no interaction in the Poisson model.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_0 + \theta_1 x_i$$

↑ treatment

$$H_0 : \theta_1 = 0$$

$$H_a : \theta_1 \neq 0$$

Test for interaction in the Poisson model.

$$x_i x_j = \gamma_3$$

$$H_0 : \gamma_3 = 0$$

$$H_a : \gamma_3 \neq 0$$

$$\boxed{\theta_1 = \gamma_3} \text{ from } \subset$$

Project 3

Mustaafa

17/05/2021

```
pacman::p_load(tidyverse, ggglm)
```

The data set lung cancer, collected by Anderson (1977)¹, contains the population size and number of cases of lung cancer in four Danish cities, stratified by age.

The purpose of the analysis is to fit a Poisson regression model to predict the number of cases of lung cancer.

(a)

Read the data lung_cancer.csv into R.

```
lung_Cancer<- read_csv(here::here("lung_cancer.csv"))
```

```
##
## — Column specification —————
## cols(
##   city = col_character(),
##   age = col_character(),
##   pop = col_double(),
##   cases = col_double()
## )
```

```
lung_Cancer
```

```
## # A tibble: 24 x 4
##   city      age      pop cases
##   <chr>    <chr> <dbl> <dbl>
## 1 Fredericia 40-54  3059    11
## 2 Horsens    40-54  2879    13
## 3 Kolding    40-54  3142     4
## 4 Vejle      40-54  2520     5
## 5 Fredericia 55-59   800    11
## 6 Horsens    55-59  1083     6
## 7 Kolding    55-59  1050     8
## 8 Vejle      55-59   878     7
## 9 Fredericia 60-64   710    11
## 10 Horsens   60-64   923    15
## # ... with 14 more rows
```

(b)

Perform an EDA of the data. In particular produce appropriate plots to look at the relationship between the number of cases and age and city. Also look at the relationship between the proportion of the population that was a cases and age and city.

Perform an EDA of the data

```
skimr::skim(lung_Cancer)
```



Data summary

Name	lung_Cancer
Number of rows	24
Number of columns	4
Column type frequency:	
character	2
numeric	2
Group variables	
None	

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
city	0	1	5	10	0	4	0
age	0	1	3	5	0	6	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
pop	0	1	1100.33	842.23	509	628	791	954.75	3142	
cases	0	1	9.33	3.16	2	7	10	11.00	15	

```
summary(lung_Cancer)
```

```
##      city      age      pop      cases
## Length:24   Length:24   Min.   : 509.0   Min.   : 2.000
## Class :character Class :character 1st Qu.: 628.0   1st Qu.: 7.000
## Mode  :character Mode  :character Median  : 791.0   Median :10.000
##                               Mean   :1100.3   Mean   : 9.333
##                               3rd Qu.: 954.8   3rd Qu.:11.000
##                               Max.   :3142.0   Max.   :15.000
```

- There are 4 variables, with their two data types being: 2 categorical and 2 numerical. Data types will inform our decision making in regards to plotting.

The following code will be converting the two variables with character data type, to factor, for ease of computation when counting.

```
lung_Cancer<-
  lung_Cancer %>%
  mutate(
    across(where(is.character), factor)
  )
lung_Cancer
```

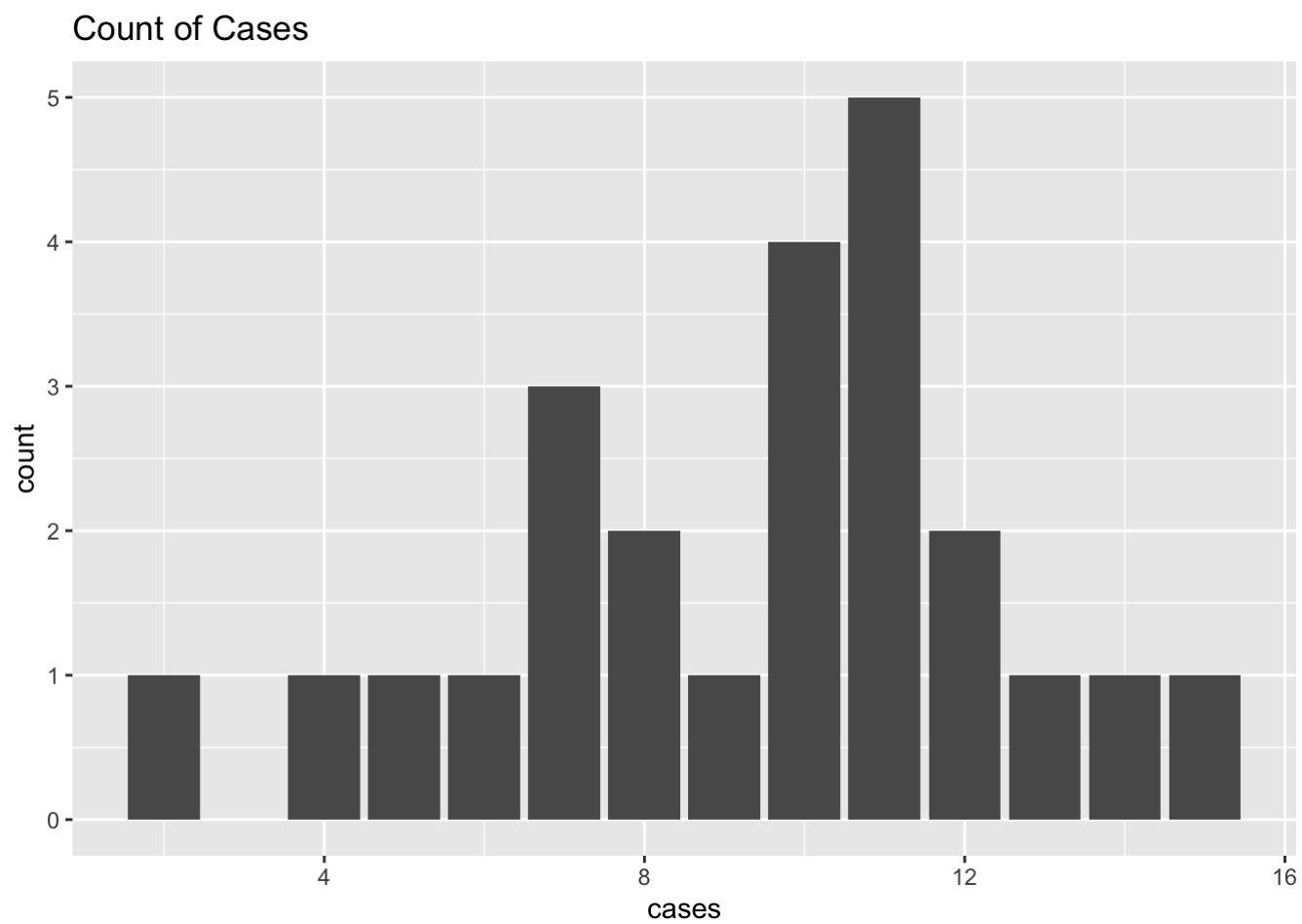
```
## # A tibble: 24 x 4
##   city      age      pop cases
##   <fct>    <fct> <dbl> <dbl>
## 1 Fredericia 40-54  3059    11
## 2 Horsens    40-54  2879    13
## 3 Kolding    40-54  3142     4
## 4 Vejle      40-54  2520     5
## 5 Fredericia 55-59   800    11
## 6 Horsens    55-59  1083     6
## 7 Kolding    55-59  1050     8
## 8 Vejle      55-59   878     7
## 9 Fredericia 60-64   710    11
## 10 Horsens   60-64   923    15
## # ... with 14 more rows
```

Produce appropriate plots to look at the relationship between the number of cases and age and city:

Analysing univariate plots for the variable Cases.

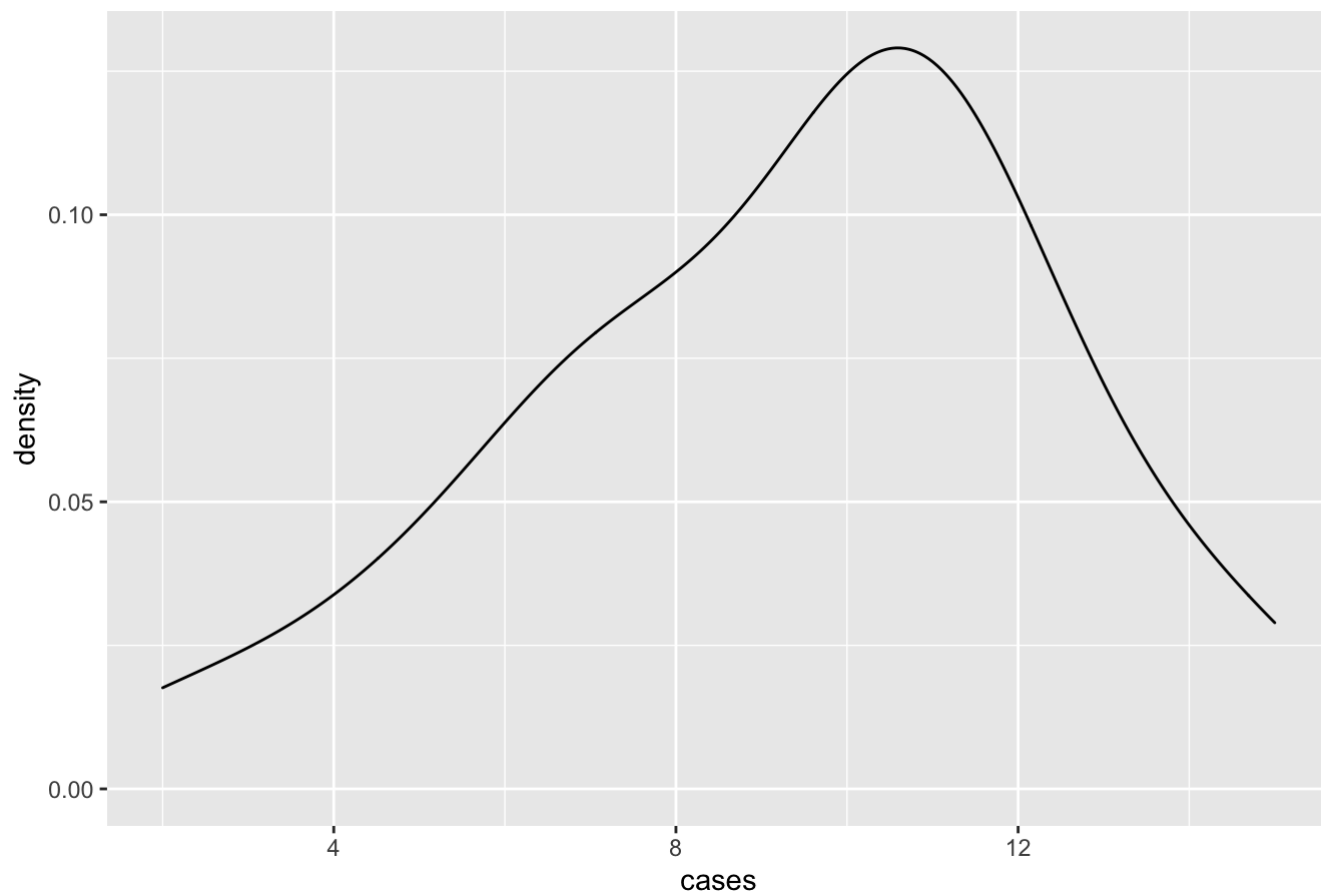
- This is univariate plot of case frequency.
- It has poisson distribution because it is counting of integers and non-negative lung cancer.

```
lung_Cancer %>%
  ggplot(aes(cases)) +
  geom_bar() +
  ggtitle("Count of Cases")
```



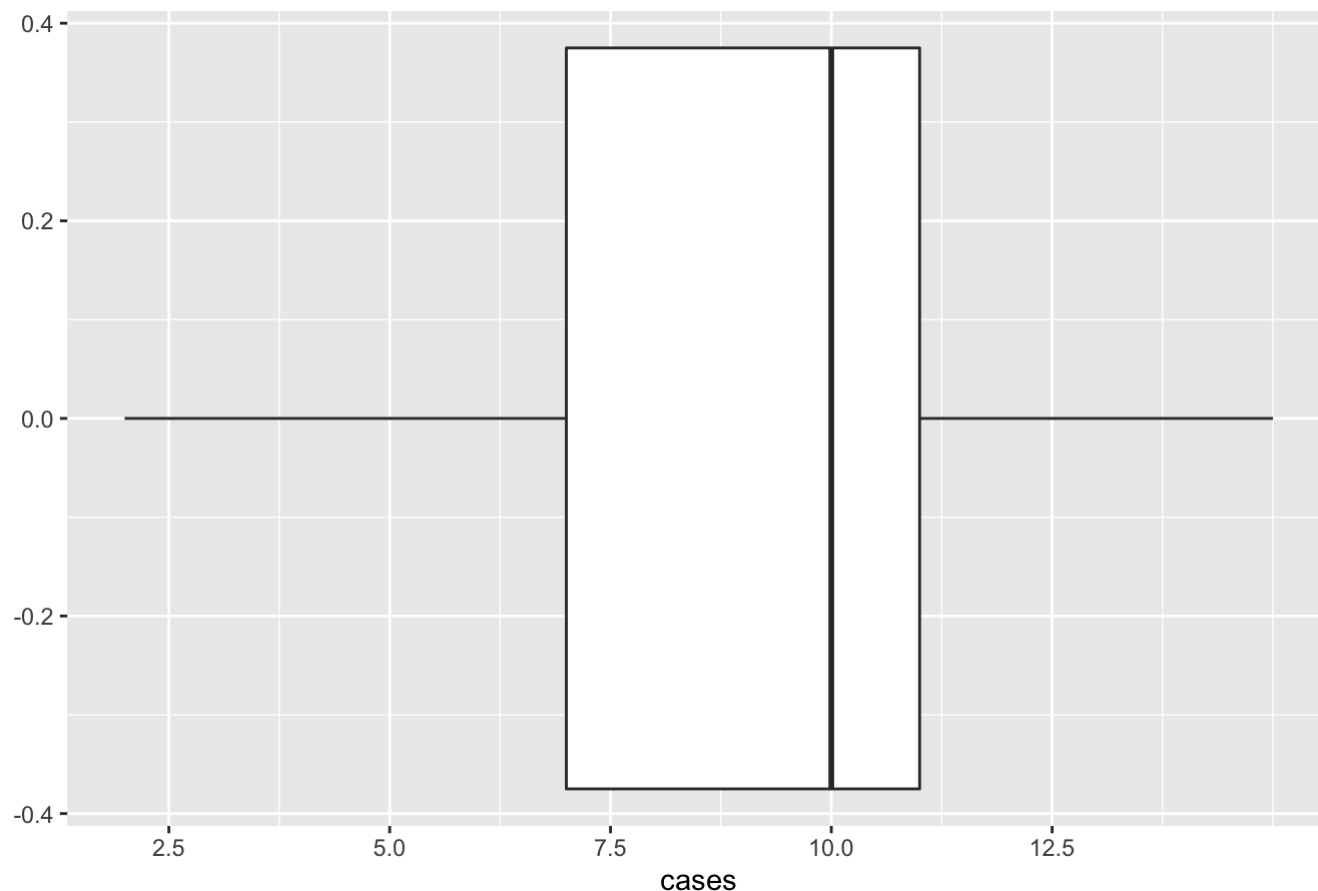
```
lung_Cancer %>%  
  ggplot(aes(cases)) +  
  geom_density() +  
  ggtitle("Density plot to confirm poisson distribution shape")
```

Density plot to confirm poisson distribution shape



```
lung_Cancer %>%  
  ggplot(aes(cases))+  
  geom_boxplot() +  
  ggtitle("Boxplot to exhibit the median number of lung cancer cases") +  
  scale_x_continuous(breaks=c(0, 2.5, 5, 7.5, 10,12.5))
```

Boxplot to exhibit the median number of lung cancer cases



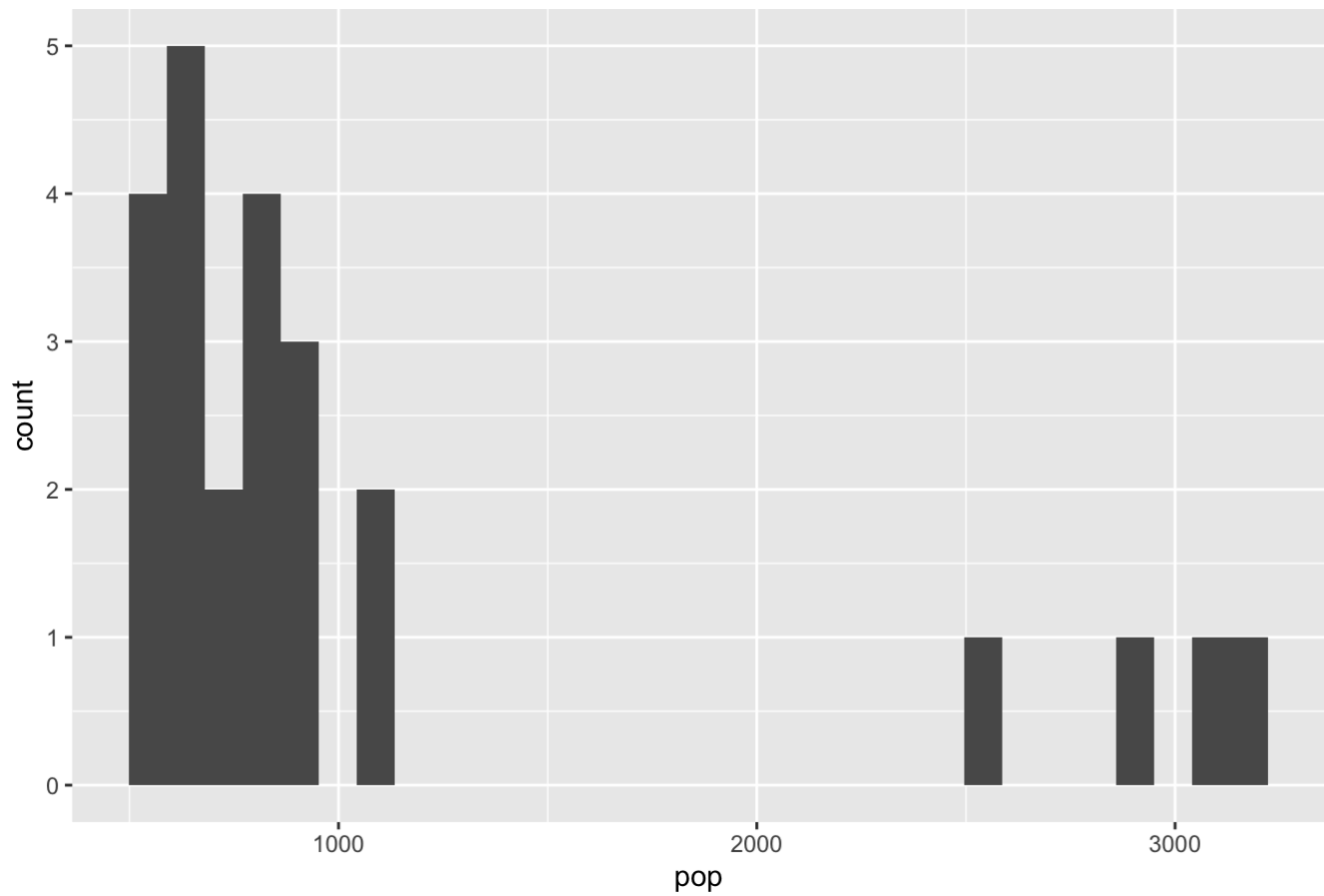
Analysing univariate plots for the variable Pop

- The population of the given cities are mostly less than one thousand, less in the middle, and few cities with higher than 2500 people as indicated by the boxplot below; further emphasized by the histogram, that shows clusters of small population at extreme. This insight can be meaningful for prediction of cases with respect to spatial dependencies, after checking the map of Danmark, there does not appear to be any spatial pattern.

```
lung_Cancer %>%  
  ggplot(aes(pop)) +  
  geom_histogram() +  
  ggtitle("The histogram of population distribution")
```

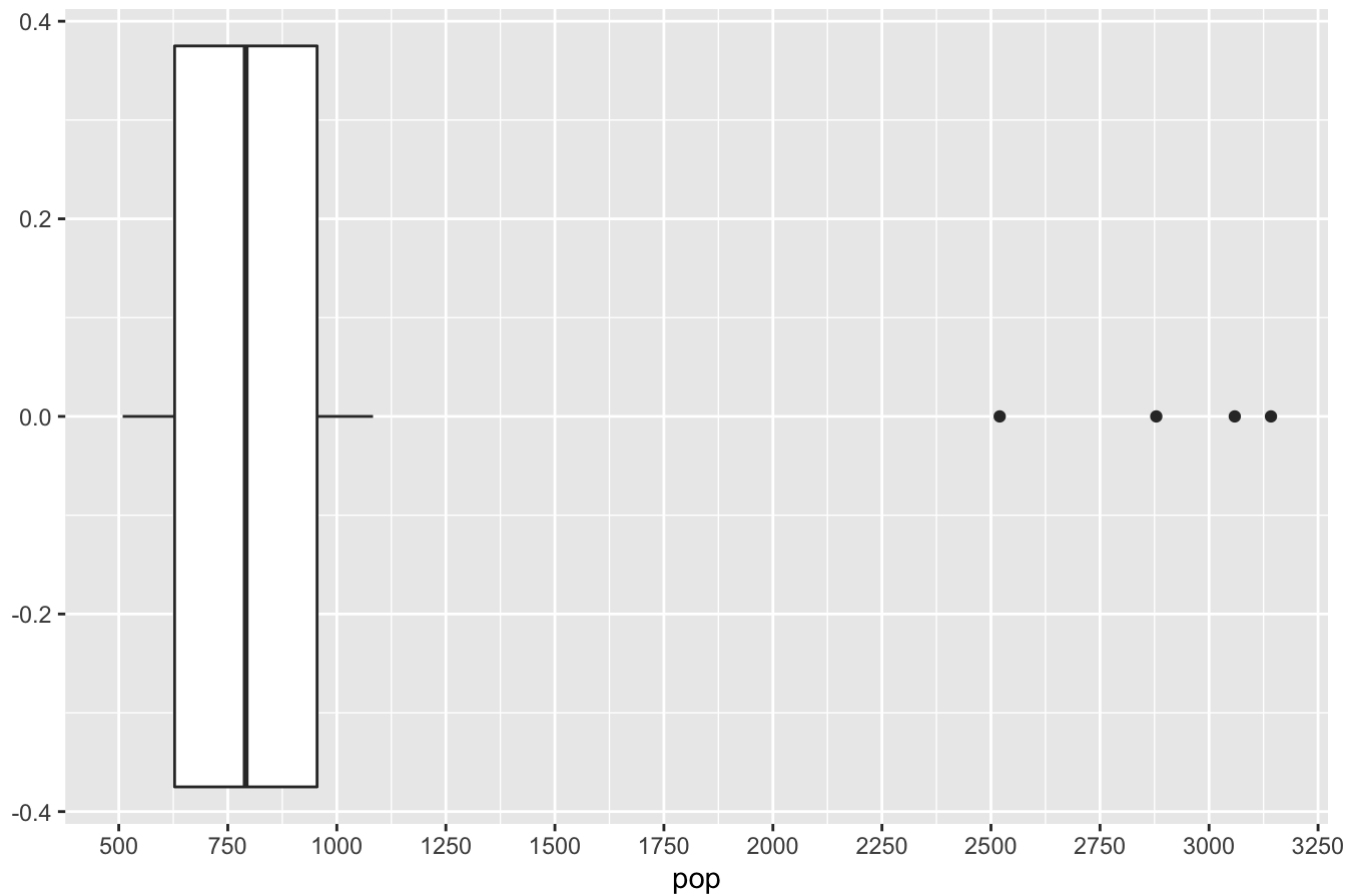
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

The histogram of population distribution



```
lung_Cancer %>%
  ggplot(aes(pop)) +
  geom_boxplot() +
  ggtitle("Boxplot to exhibit the median number of population in cities are approximately 800 people") +
  scale_x_continuous(breaks=c(250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, 3500))
```

Boxplot to exhibit the median number of population in cities are approximately 800

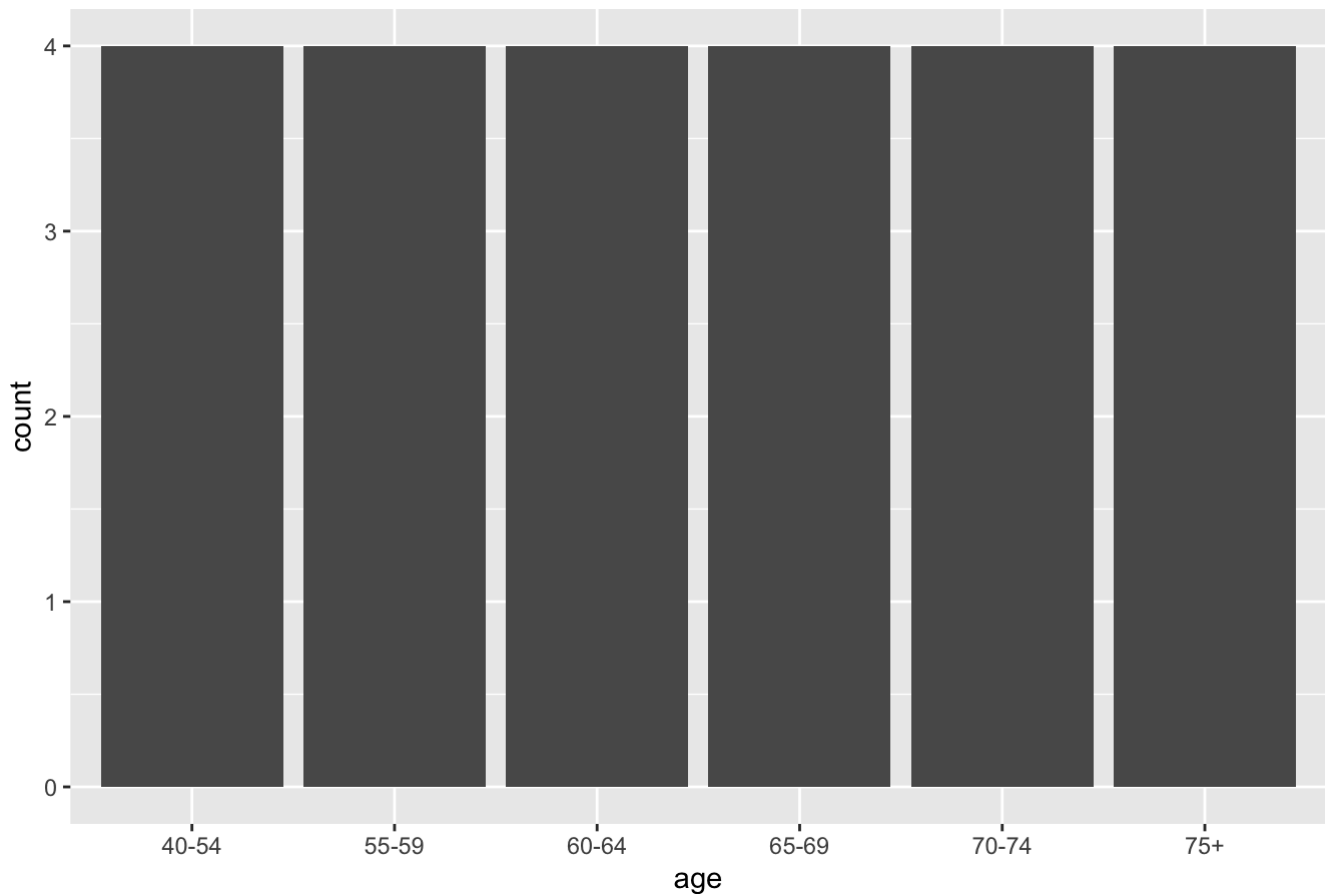


Analysing univariate plot for the variable Age

- The age distribution is uniform, which symbolises the the dataset does not represent the population, because there are missing data for the people less than 40, and above 75, going against the pyramid of human life span. Furthermore, number of cases divided by interval of age group shows that everybody in the age group interval has equal probability of getting lung cancer. For example, for the city of Fredericia, with population of 3059, 11 cases of lung cancer for the age group 40-54. Dividing 11 cases by the age 40 to 54. $(11/40) \times 100 = 27.5\%$, which makes each person between the age group of 40-54 to have equal probability of getting lung cancer. It is improbable for each person for age interval of 40-54 to have 27.5% equal chance of getting lung cancer; which is an indication the data does not represent the true population parameter. Since for each age group, there are four observations, which is not statistically adequate to represent a population.

```
lung_Cancer %>%
  ggplot(aes(age)) +
  geom_bar() +
  ggtitle("Barplot of Age distribution")
```

Barplot of Age distribution



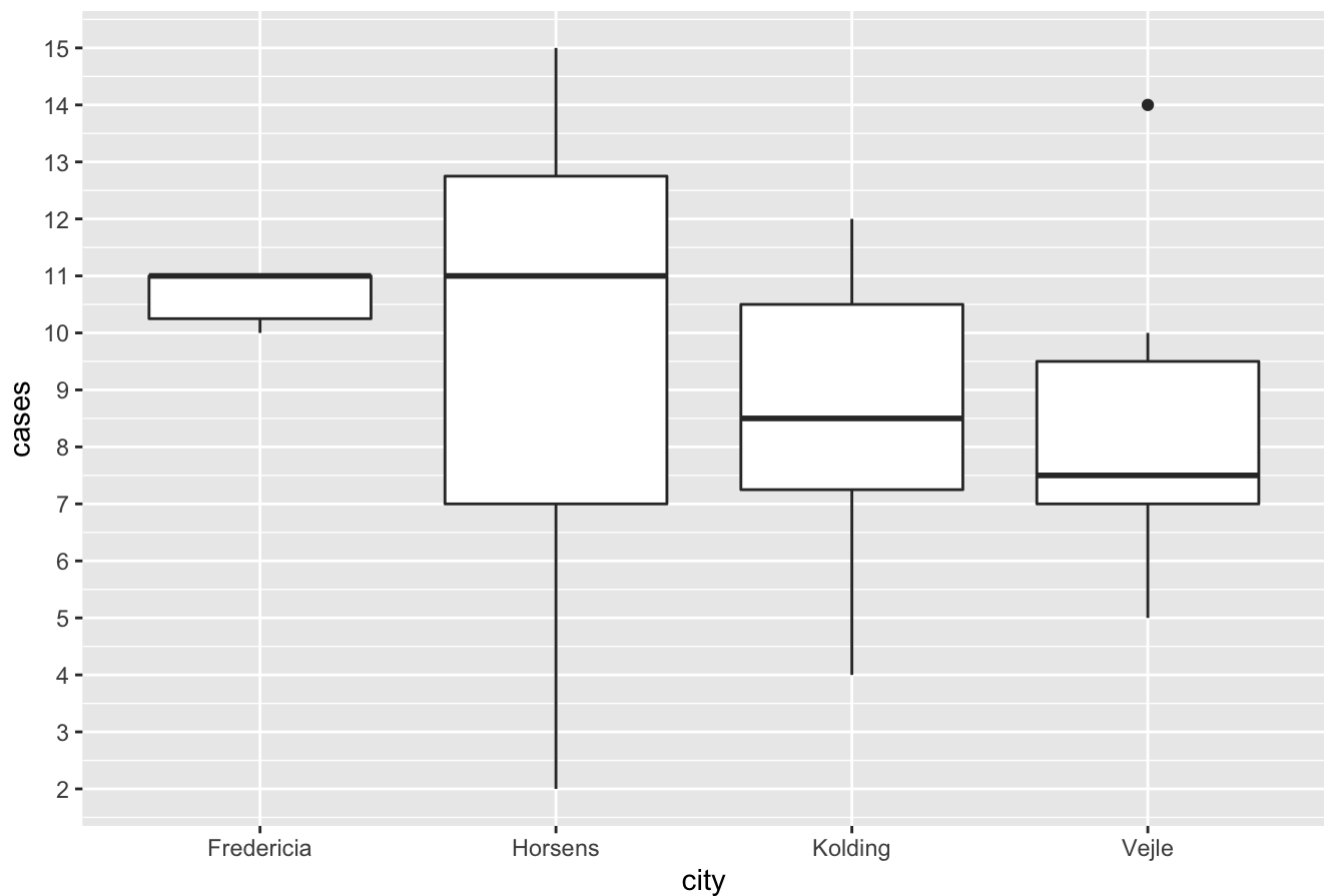
Analysing bivariate plots for each variable.

Analysing bivariate plots for the variable cases and city.

- Fredricia
 - plot shows the number of cases by city, the frequency of lung cancer cases is consistently 10 or 11, which is the lowest variance between cities.
- Horsens
 - plot shows the number of cases by city, the 6 points represent each of the age group. For Horsens the case of lung cancer varies a lot with extreme low and high among the four cities.
- Kolding
 - Kolding is similar to Horsens in distinctiveness and distribution of their cases. The variance of cases at Kolding is second high comparatively.
- Vije
 - is third highest variance comparatively similar to Horsens and Kolding, with an outlier 14 case, which sits well above the median. 75% of the cases is 10 and below.

```
lung_Cancer %>%
  ggplot(aes(city, cases))+
  geom_boxplot() +
  ggtitle("Boxplot of city and cases") +
  scale_y_continuous(breaks=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15))
```


Boxplot of city and cases



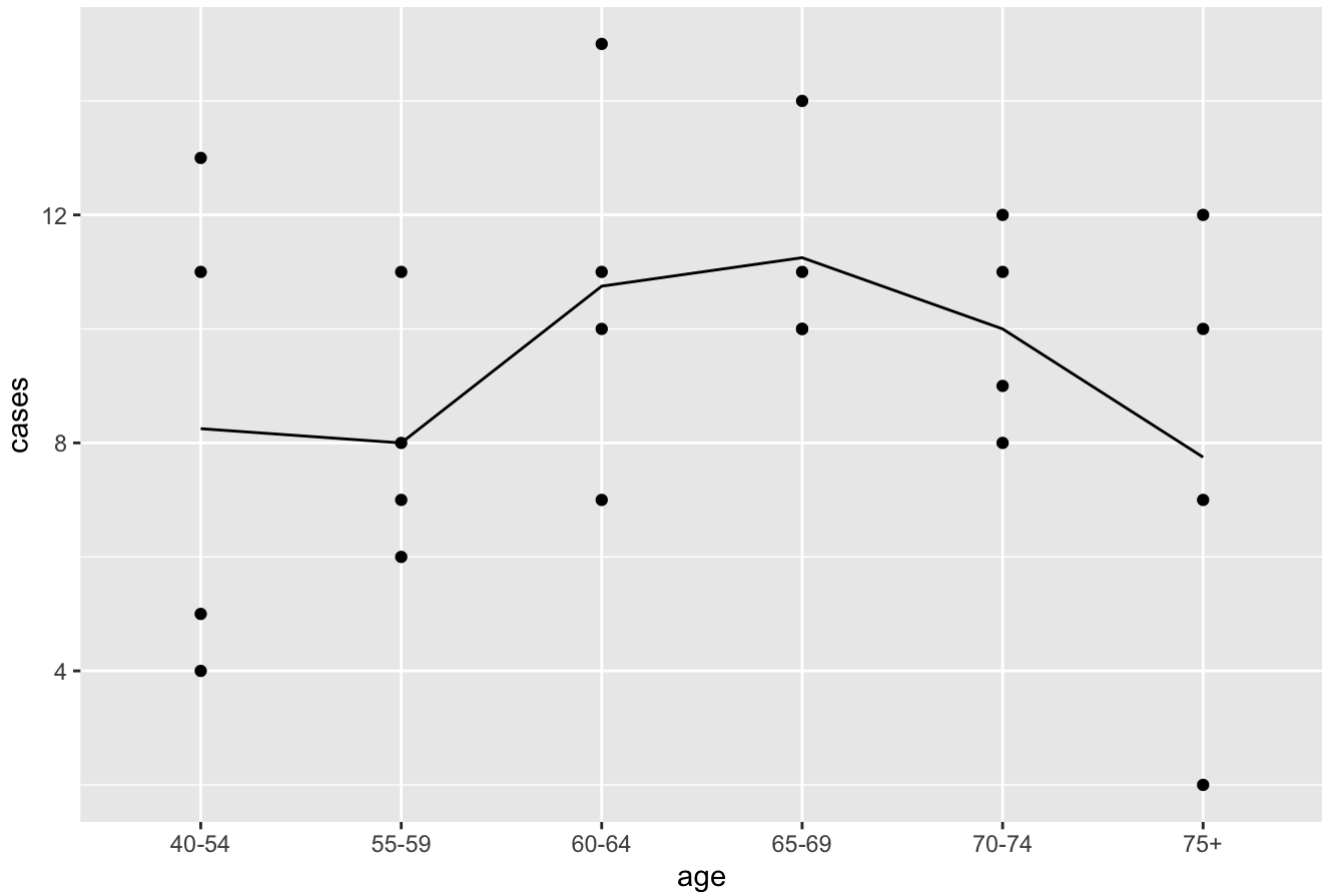
Analysing bivariate plots for the variable cases and age.

- The scatterplot of cases and age, shows a quadratic, or non linear trend, with a peak in number of lung cancer for age group 65-69. The down, up, and down, bell shaped trend does make sense as humans tend to become weak against disease as one gets older, and slowly die off.

```
lung_Cancer %>%
  ggplot(aes(age, cases))+
  geom_point()+
  stat_summary(fun.y = mean, geom = "line", aes(group = 1)) +
  ggtitle("Scattarplot of cases and age")
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

Scattarplot of cases and age



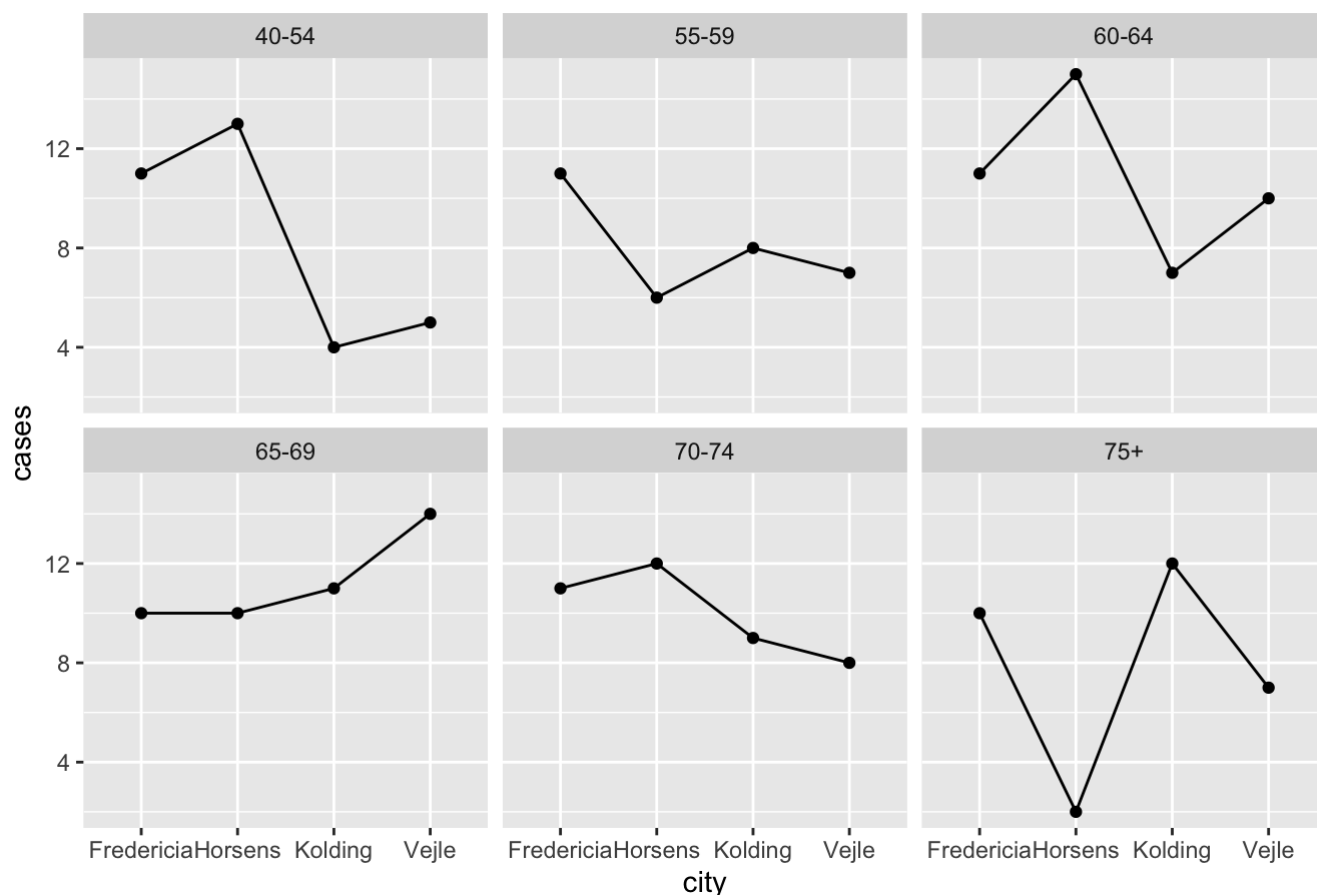
Analysing bivariate plots for the variable cases, age and city.

- The scatterplot below provides a unbrella explanation for the above plots. It shows a polynomial trend in age group 65-69, in increasing number of lung cancer as age increases. Which is an enforcement for the plots above. For example, the plot “Scattarplot of cases and age” shows that lung cancer peaks at the age 65-69. And peak is shown below which is at the city of Vejle. The city of Vejle was shown to have an outlier in the “Boxplot of city and cases”.

```
lung_Cancer %>%
  ggplot(aes(city, cases))+
  geom_point()+facet_wrap(~age) +
  stat_summary(fun.y = mean,geom = "line", aes(group = 1))+
  ggtitle("Scattarplot of cases and age and city")
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

Scattarplot of cases and age and city



```
lung_Cancer<-
  lung_Cancer %>%
  mutate(
    across(where(is.character), factor)
  )
lung_Cancer
```

```
## # A tibble: 24 x 4
##   city      age      pop cases
##   <fct>    <fct> <dbl> <dbl>
## 1 Fredericia 40-54  3059    11
## 2 Horsens    40-54  2879    13
## 3 Kolding    40-54  3142     4
## 4 Vejle      40-54  2520     5
## 5 Fredericia 55-59    800    11
## 6 Horsens    55-59  1083     6
## 7 Kolding    55-59  1050     8
## 8 Vejle      55-59   878     7
## 9 Fredericia 60-64   710    11
## 10 Horsens   60-64   923    15
## # ... with 14 more rows
```

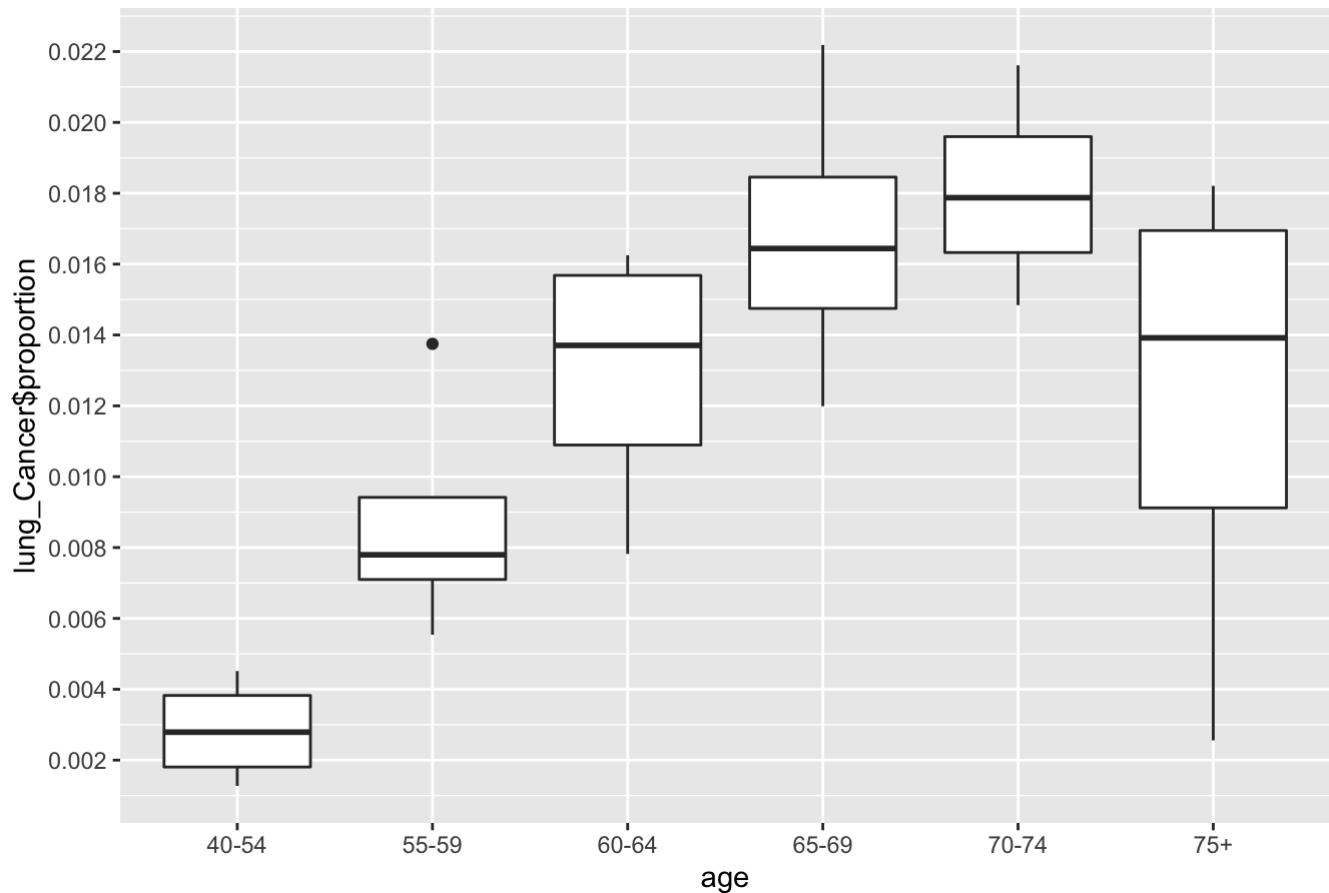
Looking at the relationship between the proportion of the population that was a cases and age and city.

- The histogram of population in the cities is proportionalised to be comparable with small cities. Otherwise, small cities will appear to have more cases compare to large cities.

```
lung_Cancer$proportion <- lung_Cancer$cases/lung_Cancer$pop # Creating Proportion
```

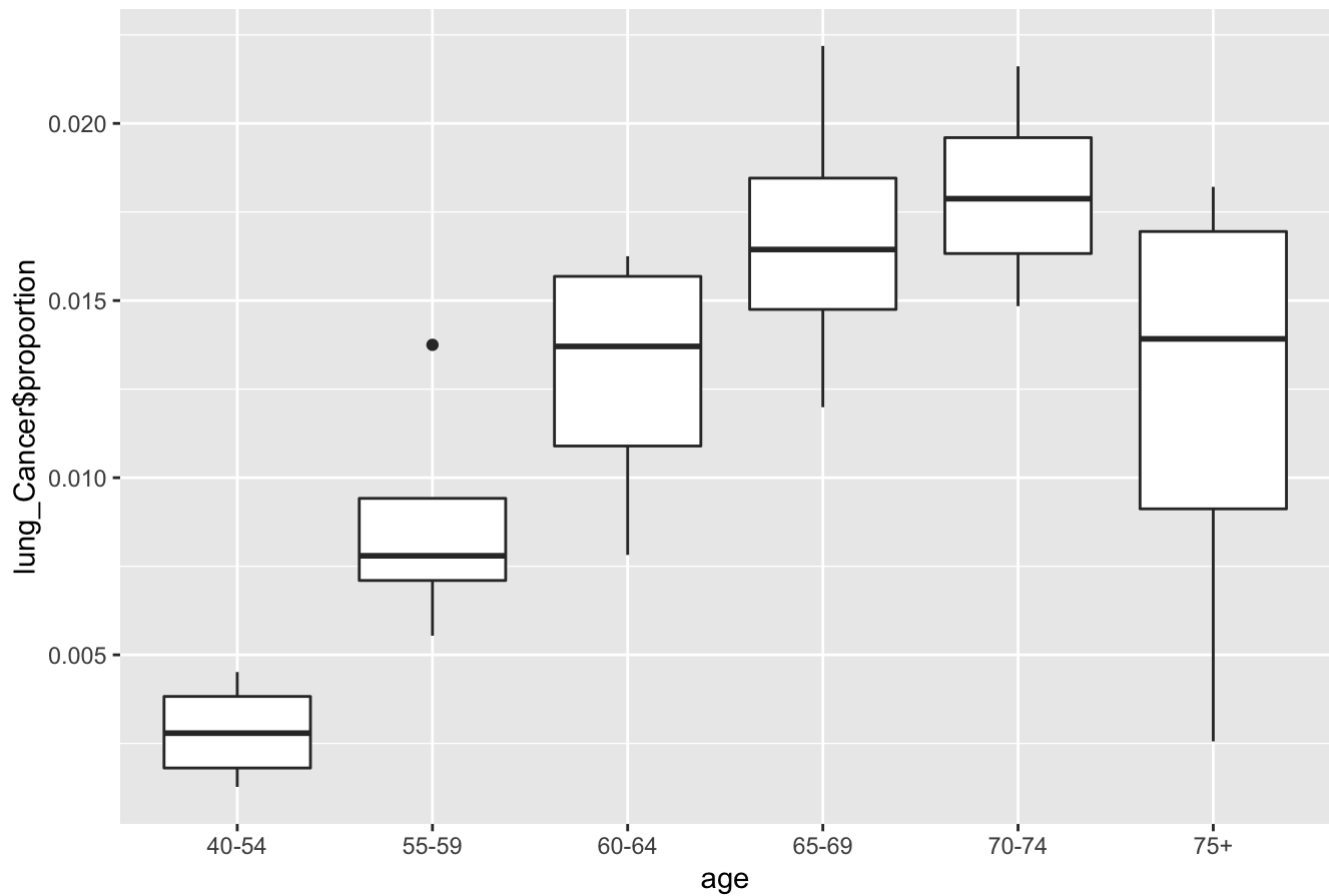
```
lung_Cancer %>%  
  ggplot(aes(age, lung_Cancer$proportion))+  
  geom_boxplot() +  
  ggtitle("BoxPlot of proportion for variable case and age")+  
  scale_y_continuous(breaks=c(0.002,0.004,0.006,0.008,0.010,0.012,0.014,0.016,0.018,  
0.020,0.022))
```

BoxPlot of proportion for variable case and age



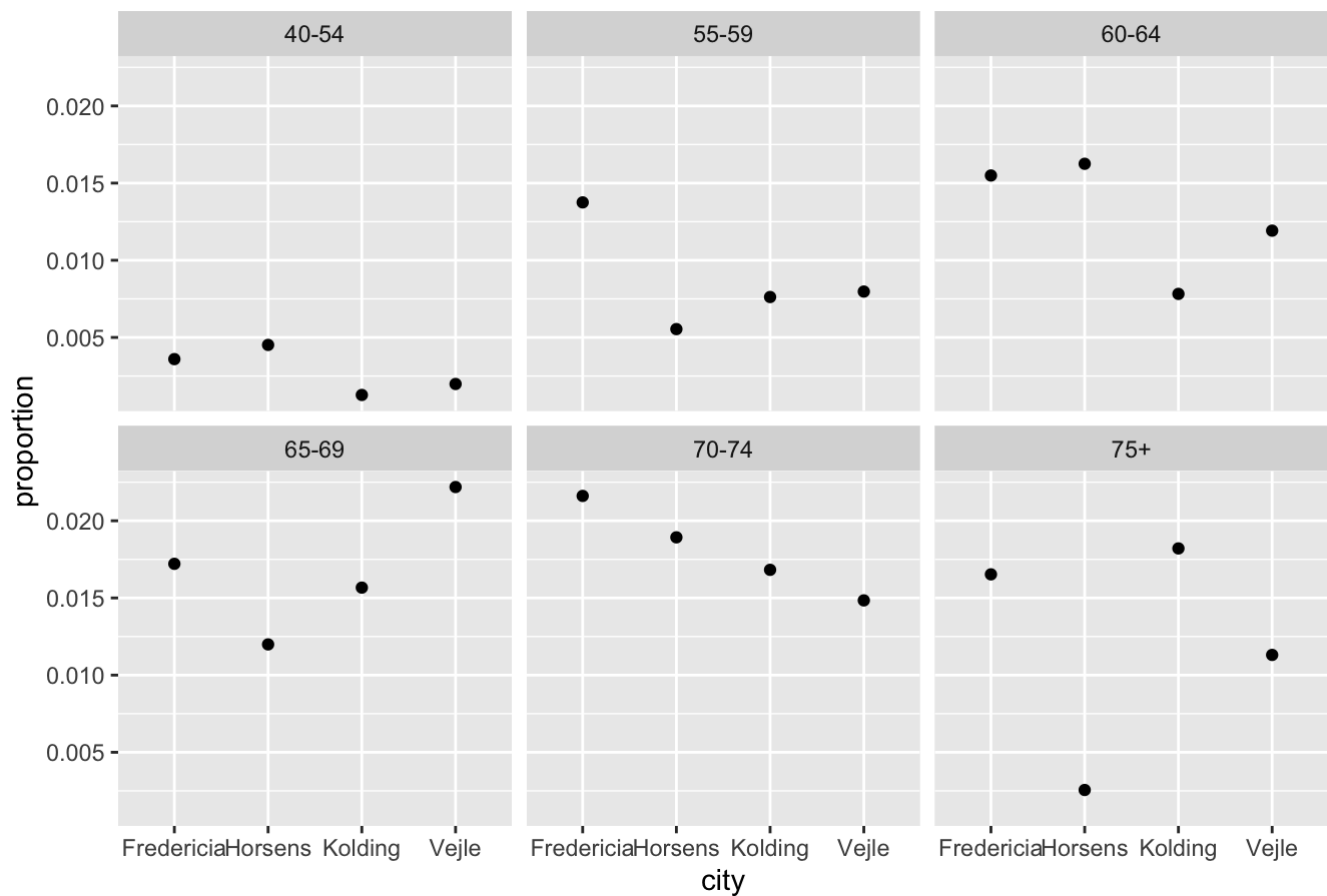
```
lung_Cancer %>%  
  ggplot(aes(age, lung_Cancer$proportion))+  
  geom_boxplot() +  
  ggtitle("BoxPlot of proportion for variable case and age")
```

BoxPlot of proportion for variable case and age



```
lung_Cancer %>%  
  ggplot(aes(city, proportion))+  
  geom_point()+facet_wrap(~age)+  
  ggtitle("Scatterplot of proportion for variable case and age and city")
```

Scatterplot of proportion for variable case and age and city



(c)

Fit a Poisson rate regression (denoted M1) with cases as the response variable, $\log(\text{pop})$ as an offset, and no predictors.

```
M1 <- glm(cases ~ 1, offset = log(pop), data = lung_Cancer, family = poisson())
summary(M1)
```

```
##
## Call:
## glm(formula = cases ~ 1, family = poisson(), data = lung_Cancer,
##      offset = log(pop))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -5.4891  -0.5126   1.2413   1.9248   3.1028
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.76978    0.06682  -71.39  <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: 129.91  on 23  degrees of freedom
## Residual deviance: 129.91  on 23  degrees of freedom
## AIC: 228.3
##
## Number of Fisher Scoring iterations: 5
```

(d)

Fit a Poisson rate regression (denoted M2) with cases as the response variable, $\log(\text{pop})$ as an offset, and age and city as the predictors.

```
M2 <- glm(
  cases ~ age+ city,
  family=poisson,
  offset = log(pop),
  data=lung_Cancer
)
summary(M2)
```

```
##
## Call:
## glm(formula = cases ~ age + city, family = poisson, data = lung_Cancer,
##      offset = log(pop))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63573  -0.67296  -0.03436   0.37258   1.85267
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.6321     0.2003  -28.125  < 2e-16 ***
## age55-59       1.1010     0.2483   4.434 9.23e-06 ***
## age60-64       1.5186     0.2316   6.556 5.53e-11 ***
## age65-69       1.7677     0.2294   7.704 1.31e-14 ***
## age70-74       1.8569     0.2353   7.891 3.00e-15 ***
## age75+         1.4197     0.2503   5.672 1.41e-08 ***
## cityHorsens    -0.3301     0.1815  -1.818  0.0690 .
## cityKolding   -0.3715     0.1878  -1.978  0.0479 *
## cityVejle     -0.2723     0.1879  -1.450  0.1472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 129.908  on 23  degrees of freedom
## Residual deviance:  23.447  on 15  degrees of freedom
## AIC: 137.84
##
## Number of Fisher Scoring iterations: 5
```

(e)

Fit a Poisson rate regression (denoted M3) with cases as the response variable, $\log(\text{pop})$ as an offset, and age, city and $\log(\text{pop})$ as the predictors.

```
M3 <- glm(
  cases ~ age + city + log(pop),
  family=poisson,
  offset = log(pop),
  data=lung_Cancer
)
summary(M3)
```



```
##
## Call:
## glm(formula = cases ~ age + city + log(pop), family = poisson,
##      data = lung_Cancer, offset = log(pop))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.44001  -0.64195  -0.04286   0.50052   1.51893
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   11.7496     8.8151   1.333  0.1826
## age55-59      -1.3842     1.2729  -1.087  0.2768
## age60-64      -1.2367     1.4049  -0.880  0.3787
## age65-69      -1.4378     1.6310  -0.882  0.3780
## age70-74      -1.8049     1.8608  -0.970  0.3321
## age75+        -1.8383     1.6588  -1.108  0.2678
## cityHorsens    0.1833     0.3193   0.574  0.5660
## cityKolding   -0.0483     0.2520  -0.192  0.8480
## cityVejle     -0.1679     0.1965  -0.855  0.3927
## log(pop)      -2.2096     1.1227  -1.968  0.0491 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 129.908  on 23  degrees of freedom
## Residual deviance:  19.498  on 14  degrees of freedom
## AIC: 135.89
##
## Number of Fisher Scoring iterations: 4
```

(f)

Use ANOVA to compare M1 and M2. Use this to decide if age and city are significant predictors?

- age and city are significant predictors. With p-value less than 0.05, therefore we reject the null hypothesis in support of the larger model, which is the M2, with more predictors.

```
anova(M1,M2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: cases ~ 1
## Model 2: cases ~ age + city
##      Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1          23      129.908
## 2          15       23.447   8    106.46 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(g)**Find the AIC for all three models.**

- M1, M2 has larger amount unexplained variance. And M3, has lower amount of unexplained variance; therefore, M3, is a better model, as it explains the data with more accuracy. However, since the difference in AIC between M2 and M3 is less than 1.9493, which is less than 2. Then we select M2 as model a preferred model. M2 is also not overly complicated.

```
AIC(M1, M2, M3)
```

```
##      df      AIC
## M1   1 228.2960
## M2   9 137.8355
## M3  10 135.8862
```

(h)**Produce a summary of the coefficients for M2.**

```
coef(M2)
```

```
## (Intercept)    age55-59    age60-64    age65-69    age70-74    age75+
## -5.6320645    1.1010140    1.5186123    1.7677062    1.8568633    1.4196534
## cityHorsens  cityKolding  cityVejle
## -0.3300600   -0.3715462   -0.2723177
```

```
summary(M2)
```

```
##
## Call:
## glm(formula = cases ~ age + city, family = poisson, data = lung_Cancer,
##      offset = log(pop))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.63573  -0.67296  -0.03436   0.37258   1.85267
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.6321     0.2003  -28.125  < 2e-16 ***
## age55-59       1.1010     0.2483   4.434 9.23e-06 ***
## age60-64       1.5186     0.2316   6.556 5.53e-11 ***
## age65-69       1.7677     0.2294   7.704 1.31e-14 ***
## age70-74       1.8569     0.2353   7.891 3.00e-15 ***
## age75+         1.4197     0.2503   5.672 1.41e-08 ***
## cityHorsens    -0.3301     0.1815  -1.818  0.0690 .
## cityKolding   -0.3715     0.1878  -1.978  0.0479 *
## cityVejle     -0.2723     0.1879  -1.450  0.1472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 129.908  on 23  degrees of freedom
## Residual deviance:  23.447  on 15  degrees of freedom
## AIC: 137.84
##
## Number of Fisher Scoring iterations: 5
```

(i)

Obtain the Pearson residuals for M2.

```
lung_Cancer <-
  lung_Cancer %>%
  add_column(
    M2_res = residuals(M2, type = "pearson"),
    M2_fit = fitted(M2)
  )

lung_Cancer$M2_res
```

```
##           1           2           3           4           5           6
## 0.013652695 2.052625514 -1.349807378 -0.714508940 0.812381642 -0.823478333
##           7           8           9          10          11          12
## 0.072106398 -0.075057830 -0.178845887 1.260080067 -0.973530842 -0.138695193
##           13          14          15          16          17          18
## -0.626542225 -0.726477239 0.264956939 1.234393293 -0.196875673 0.478838263
##           19          20          21          22          23          24
## 0.185144967 -0.460849280 0.347290318 -2.192356453 2.030745725 0.006703434
```

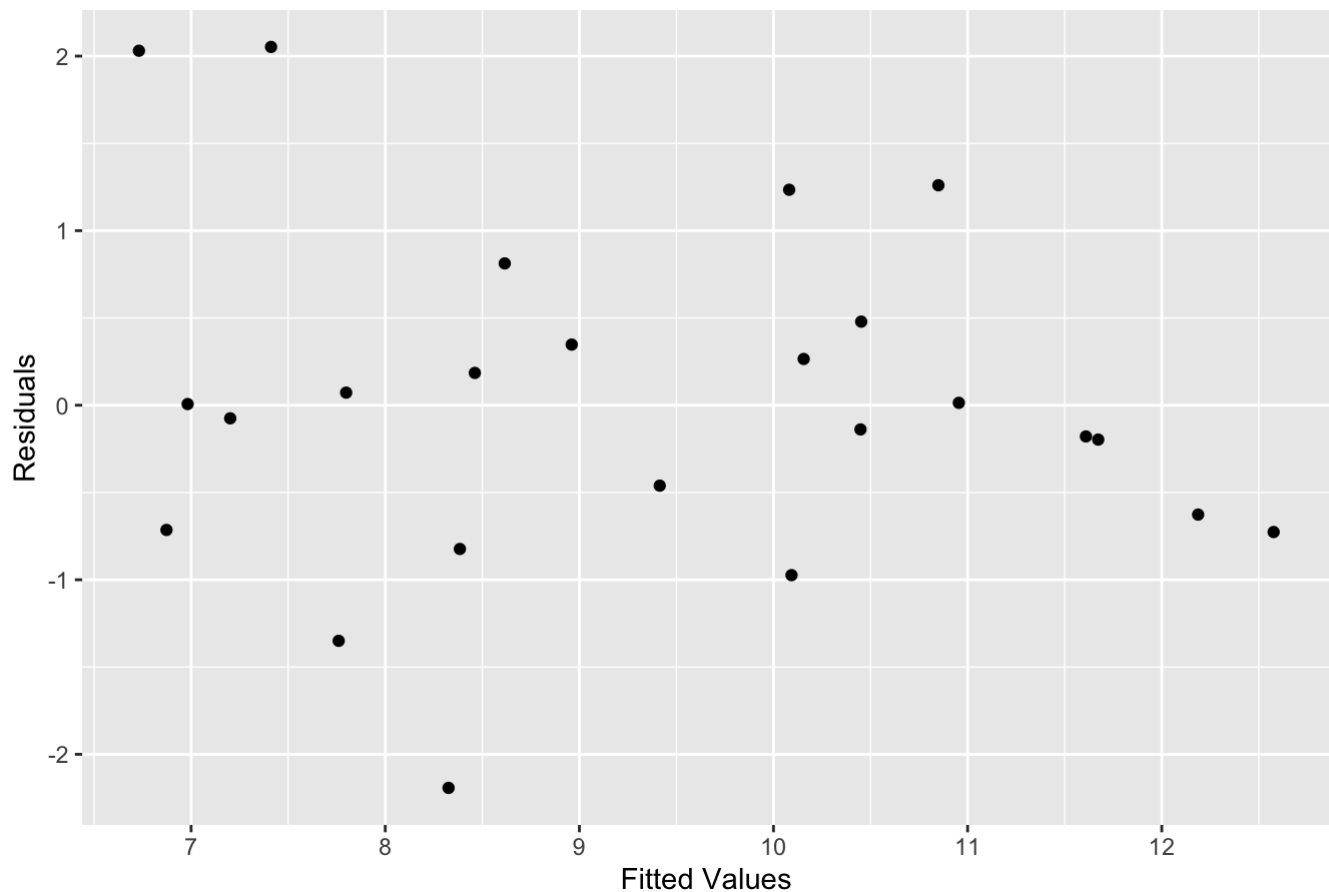
(j)

Plot the residuals versus

- fitted values,
- age, and
- city

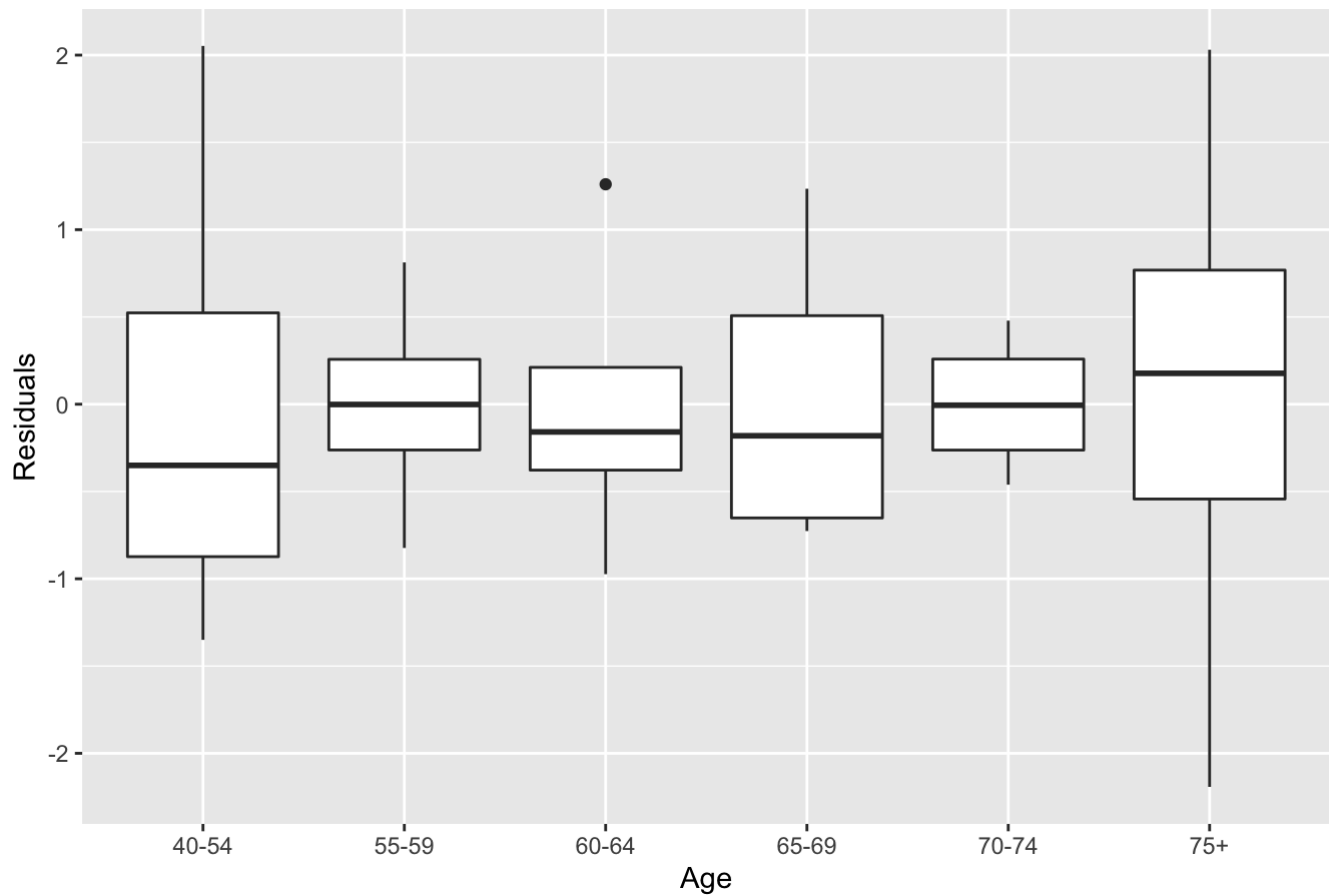
```
lung_Cancer %>%
  ggplot(aes(M2_fit, M2_res))+
  geom_point() +
  labs(title= "Scatter plot of Residuals Vs Fitted Values",y='Residuals', x = 'Fitted
Values')
```

Scatter plot of Residuals Vs Fitted Values



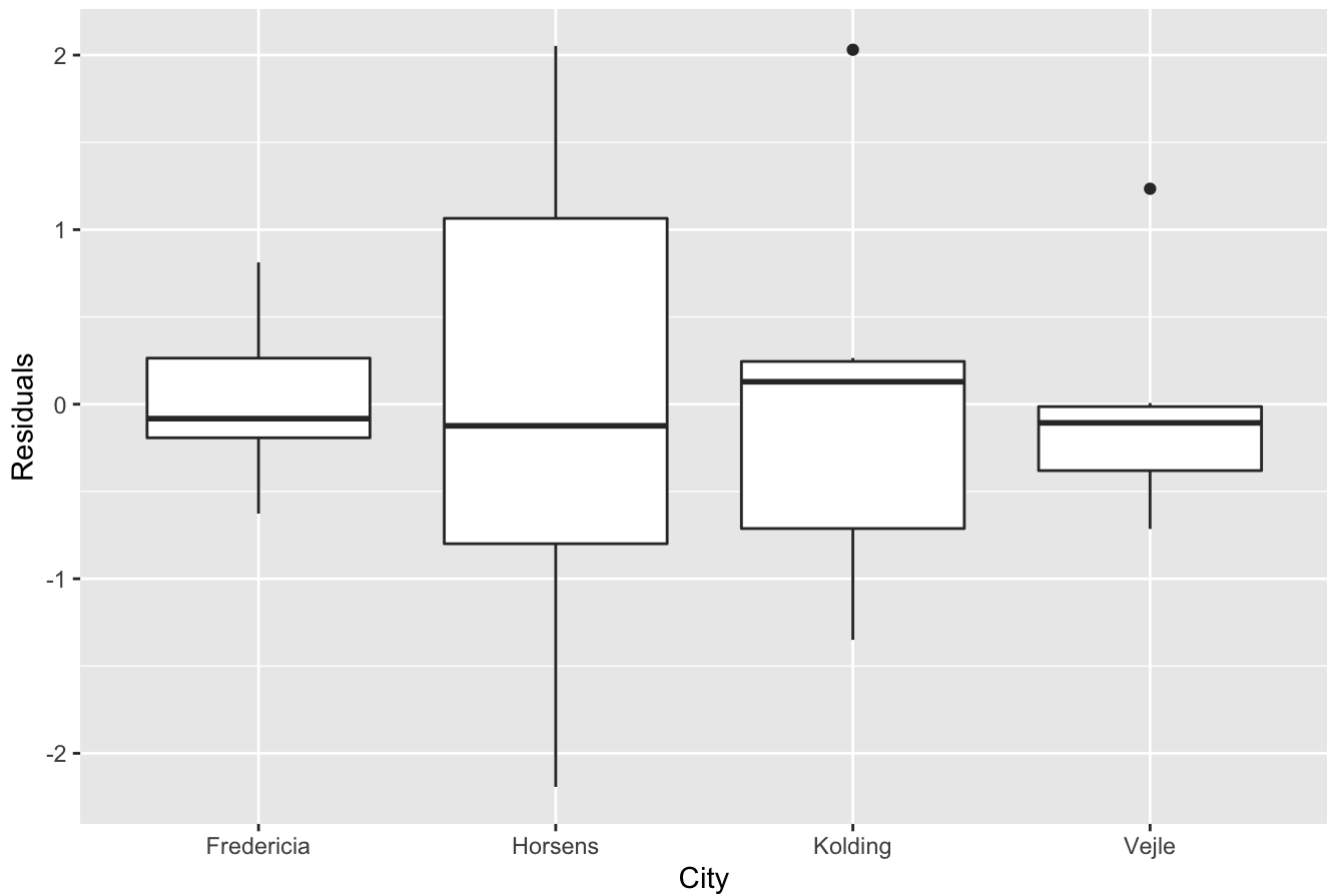
```
lung_Cancer %>%
  ggplot(aes(age, M2_res))+
  geom_boxplot()+
  labs(title= "Scatter plot of Residuals Vs Age",y='Residuals', x = 'Age')
```

Scatter plot of Residuals Vs Age



```
lung_Cancer %>%  
  ggplot(aes(city, M2_res))+  
  geom_boxplot() +  
  labs(title= "Scatter plot of Residuals Vs City",y='Residuals', x = 'City')
```

Scatter plot of Residuals Vs City

**(k)**

A health campaign was introduced in Fredericia. In 1980, the number of cases of lung cancer in the 40-54 age group (population now 4000) was 5. By calculating the probability of have five cases or fewer assuming the same rate as given by M2, decide if this is a significant decreases in the rate of lung cases.

- Yes, there's a significant decrease, because if we use the same data as M2, the chance of having 5 or fewer cases is very minimal, yet now we observed it. So the original model is not correct anymore.

```
lambda = exp(M2$coefficients[1] + log(4000))
sum(dpois(0:5, lambda = lambda))
```

```
## [1] 0.004440372
```