STA440 Final Project

Sophia Li

2024-12-15

```
# Using readr for better performance
library(readr)
suicide_data <- read_csv("SuicideChina.csv")
```

```
## New names:
## Rows: 2571 Columns: 12
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (7): Hospitalised, Died, Urban, Sex, Education, Occupation, method dbl (5):
## ...1, Person_ID, Year, Month, Age
## 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.
## * `` -> `...1`
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
library(nnet)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
head(suicide_data)
```

```
## # A tibble: 6 x 12
##    ...1 Person_ID Hospitalised Died  Urban  Year Month Sex      Age Education
##   <dbl>     <dbl> <chr>        <chr> <chr> <dbl> <dbl> <chr>  <dbl> <chr>
## 1     1         1 yes          no    no     2010    12 female    39 Secondary
## 2     2         2 no           yes   no     2009     3 male      83 primary
## 3     3         3 no           yes   no     2010     2 male      60 primary
## 4     4         4 no           yes   no     2011     1 male      73 primary
## 5     5         5 yes          no    no     2009     8 male      51 Secondary
## 6     6         6 no           yes   no     2009    11 male      62 iliterate
## # i 2 more variables: Occupation <chr>, method <chr>
```

```r
str(suicide_data)
```

```
## spc_tbl_ [2,571 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1        : num [1:2571] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Person_ID   : num [1:2571] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Hospitalised: chr [1:2571] "yes" "no" "no" "no" ...
##  $ Died        : chr [1:2571] "no" "yes" "yes" "yes" ...
##  $ Urban       : chr [1:2571] "no" "no" "no" "no" ...
##  $ Year        : num [1:2571] 2010 2009 2010 2011 2009 ...
##  $ Month       : num [1:2571] 12 3 2 1 8 11 1 10 7 1 ...
##  $ Sex         : chr [1:2571] "female" "male" "male" "male" ...
##  $ Age         : num [1:2571] 39 83 60 73 51 62 90 54 66 30 ...
##  $ Education   : chr [1:2571] "Secondary" "primary" "primary" "primary" ...
##  $ Occupation  : chr [1:2571] "household" "farming" "farming" "farming" ...
##  $ method      : chr [1:2571] "Other poison" "Hanging" "Hanging" "Hanging" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_double(),
##   ..   Person_ID = col_double(),
##   ..   Hospitalised = col_character(),
##   ..   Died = col_character(),
##   ..   Urban = col_character(),
##   ..   Year = col_double(),
##   ..   Month = col_double(),
##   ..   Sex = col_character(),
##   ..   Age = col_double(),
```

```
##   ..   Education = col_character(),
##   ..   Occupation = col_character(),
##   ..   method = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
summary(suicide_data)
```

```
##       ...1          Person_ID      Hospitalised           Died
##  Min.   :   1.0   Min.   :   1.0   Length:2571        Length:2571
##  1st Qu.: 643.5   1st Qu.: 643.5   Class :character   Class :character
##  Median :1286.0   Median :1286.0   Mode  :character   Mode  :character
##  Mean   :1286.0   Mean   :1286.0
##  3rd Qu.:1928.5   3rd Qu.:1928.5
##  Max.   :2571.0   Max.   :2571.0
##     Urban                Year          Month            Sex
##  Length:2571        Min.   :2009   Min.   : 1.000   Length:2571
##  Class :character   1st Qu.:2009   1st Qu.: 4.000   Class :character
##  Mode  :character   Median :2010   Median : 6.000   Mode  :character
##                     Mean   :2010   Mean   : 6.298
##                     3rd Qu.:2011   3rd Qu.: 9.000
##                     Max.   :2011   Max.   :12.000
##       Age          Education          Occupation           method
##  Min.   : 12.00   Length:2571        Length:2571        Length:2571
##  1st Qu.: 37.00   Class :character   Class :character   Class :character
##  Median : 53.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 52.63
##  3rd Qu.: 69.00
##  Max.   :100.00
```

Let's do some EDA first

```
table(suicide_data$Died)
```

```
##
##   no  yes
## 1315 1256
```

```
table(suicide_data$Sex)
```

```
##
## female   male
##   1328   1243
```

```
table(suicide_data$Urban)
```

```
##
##      no unknown     yes
##    2213      81     277
```

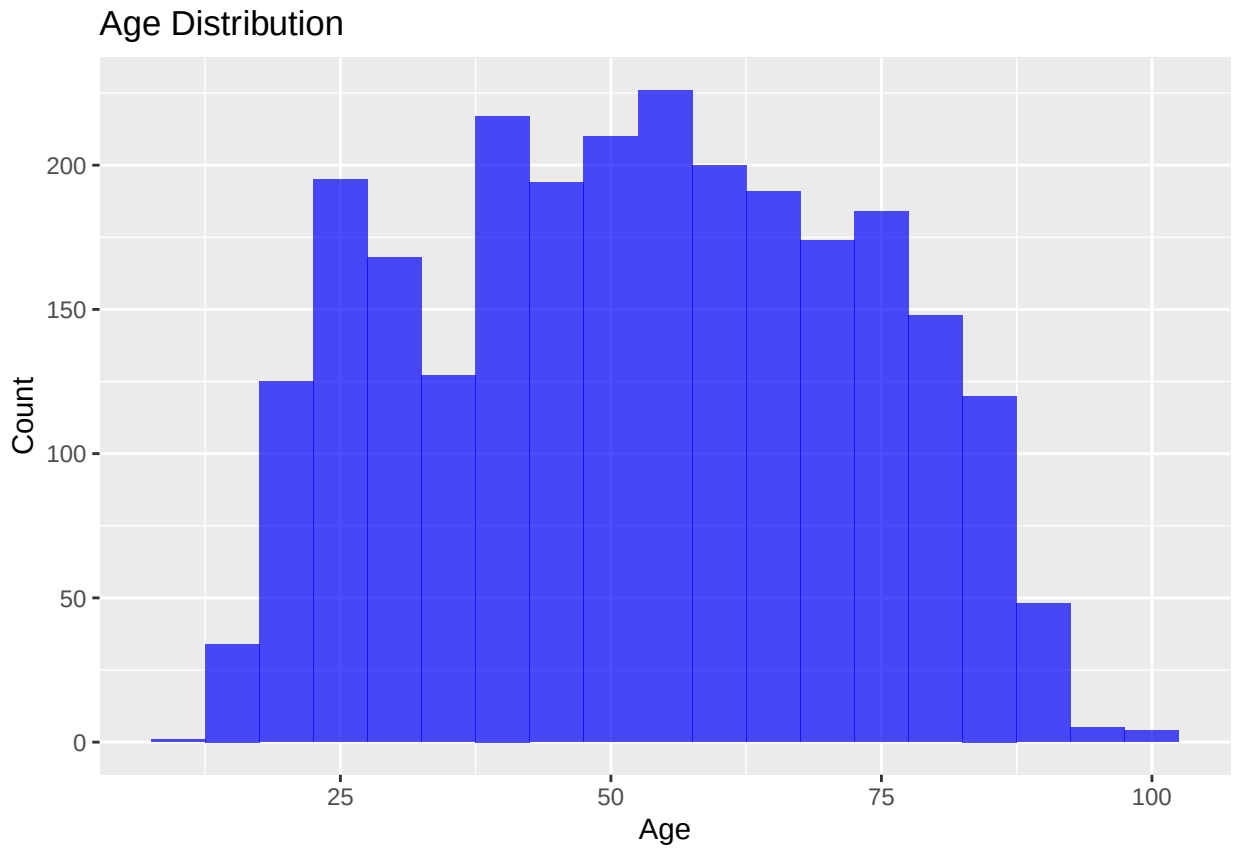
```r
table(suicide_data$method)
```

```
##
##       Cutting      Drowning       Hanging       Jumping  Other poison
##            29            26           431            15           146
##        Others     Pesticide Poison unspec   unspecified
##             1          1768           107            48
```

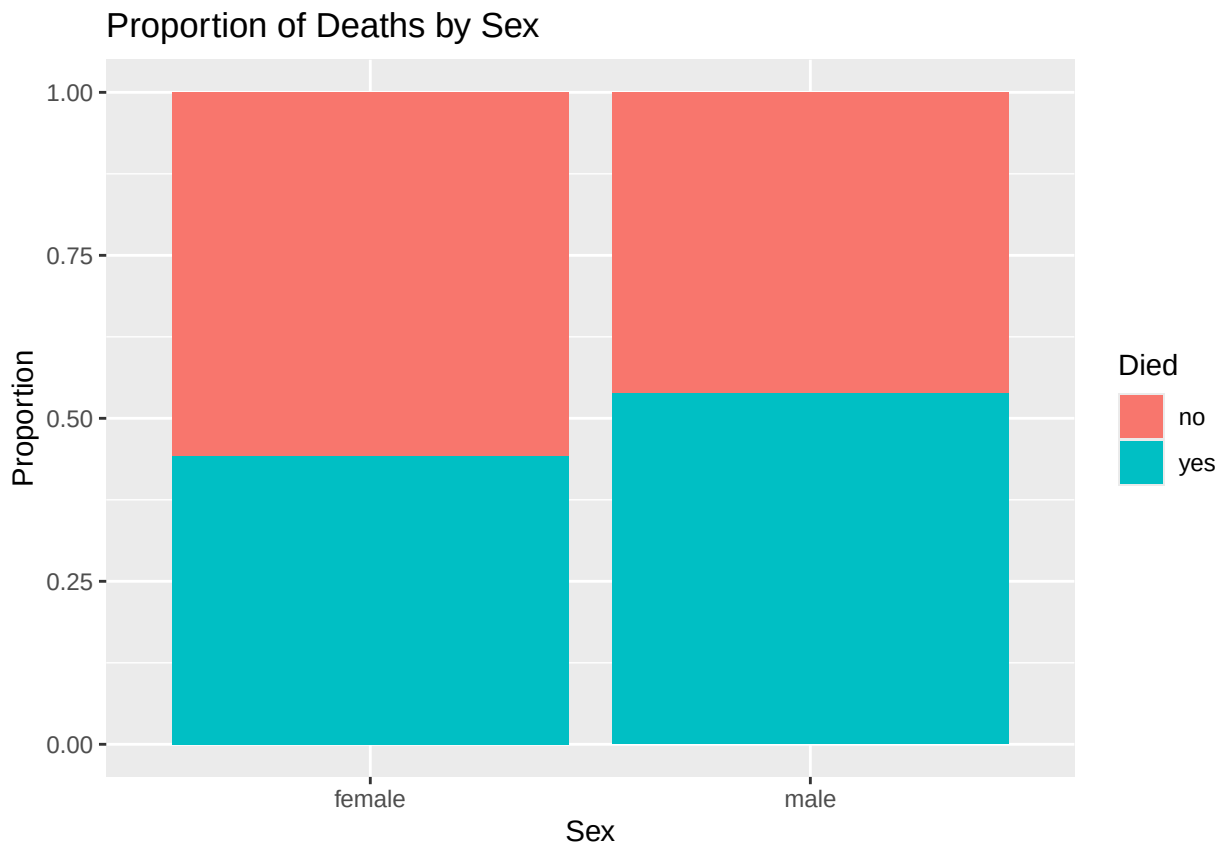
```r
summary(suicide_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   37.00   53.00   52.63   69.00  100.00
```

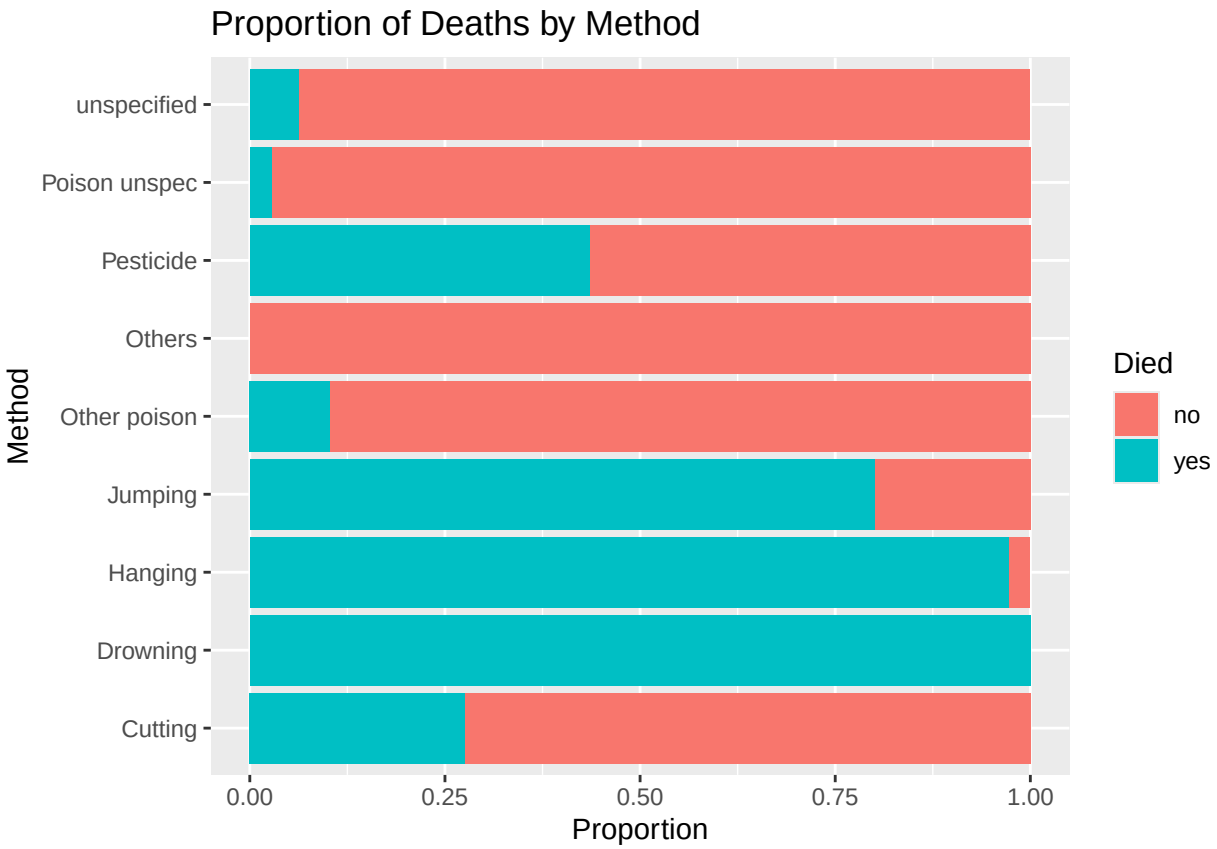
```r
ggplot(suicide_data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", alpha = 0.7) +
  labs(title = "Age Distribution", x = "Age", y = "Count")
```




```r
ggplot(suicide_data, aes(x = Sex, fill = Died)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Deaths by Sex", x = "Sex", y = "Proportion", fill = "Died")
```

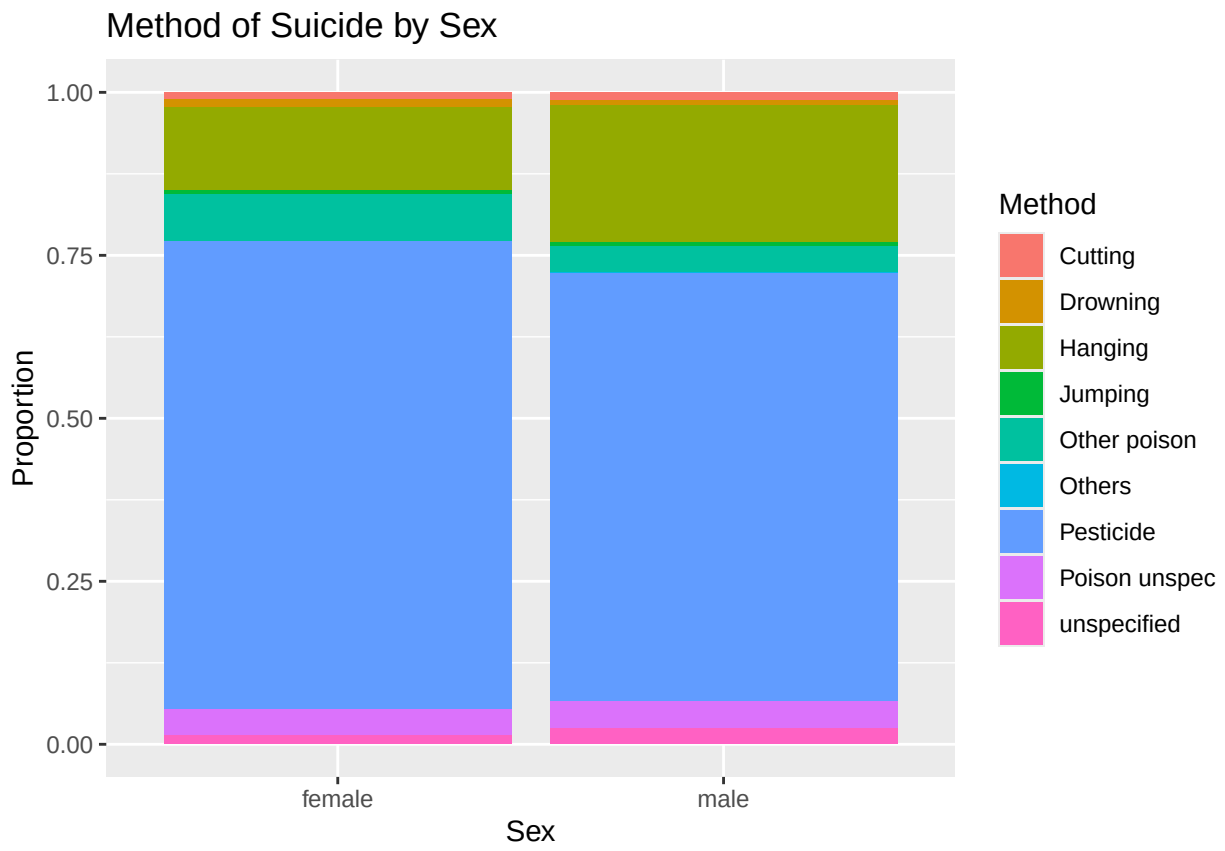



```r
ggplot(suicide_data, aes(x = method, fill = Died)) +
  geom_bar(position = "fill") +
  coord_flip() +
  labs(title = "Proportion of Deaths by Method", x = "Method", y = "Proportion", fill = "Died")
```

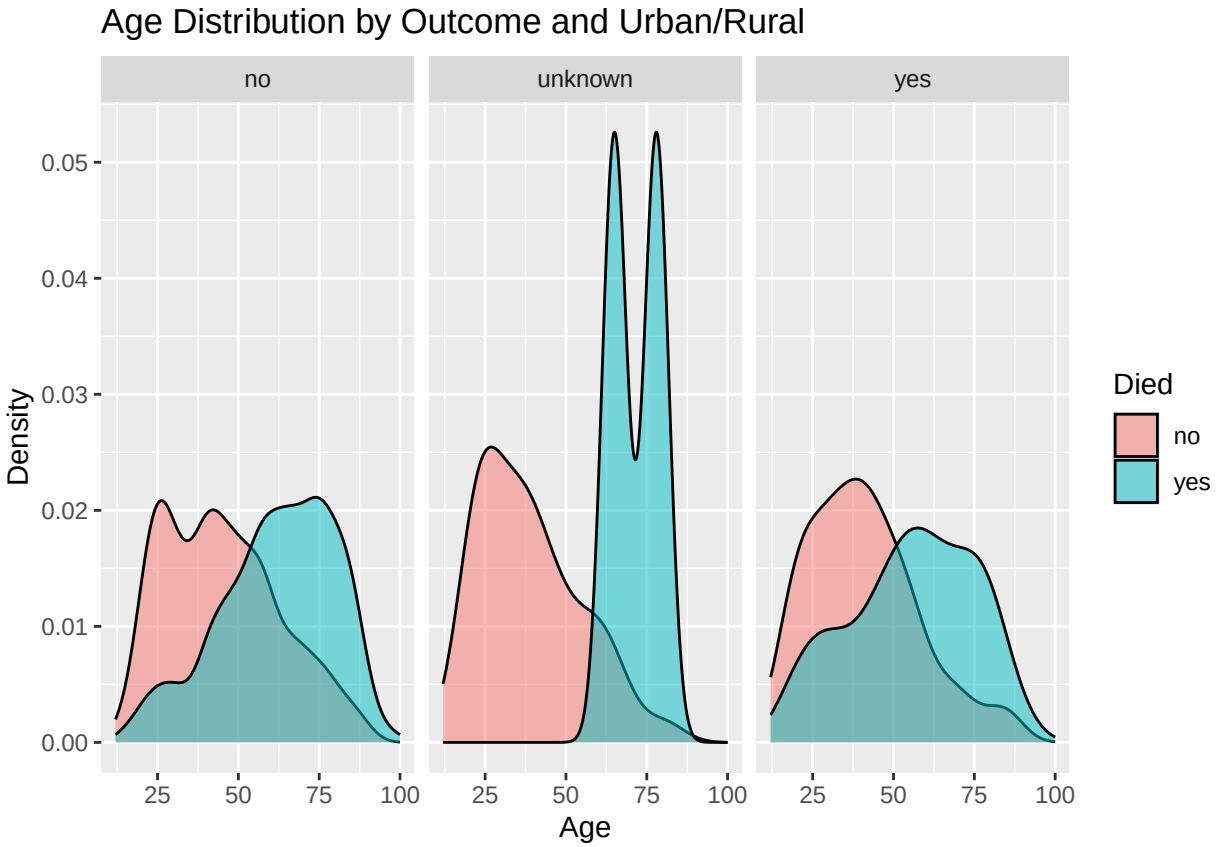



```
ggplot(suicide_data, aes(x = Age, fill = Died)) +
  geom_density(alpha = 0.5) +
  labs(title = "Age Distribution by Outcome", x = "Age", y = "Density", fill = "Died")
```

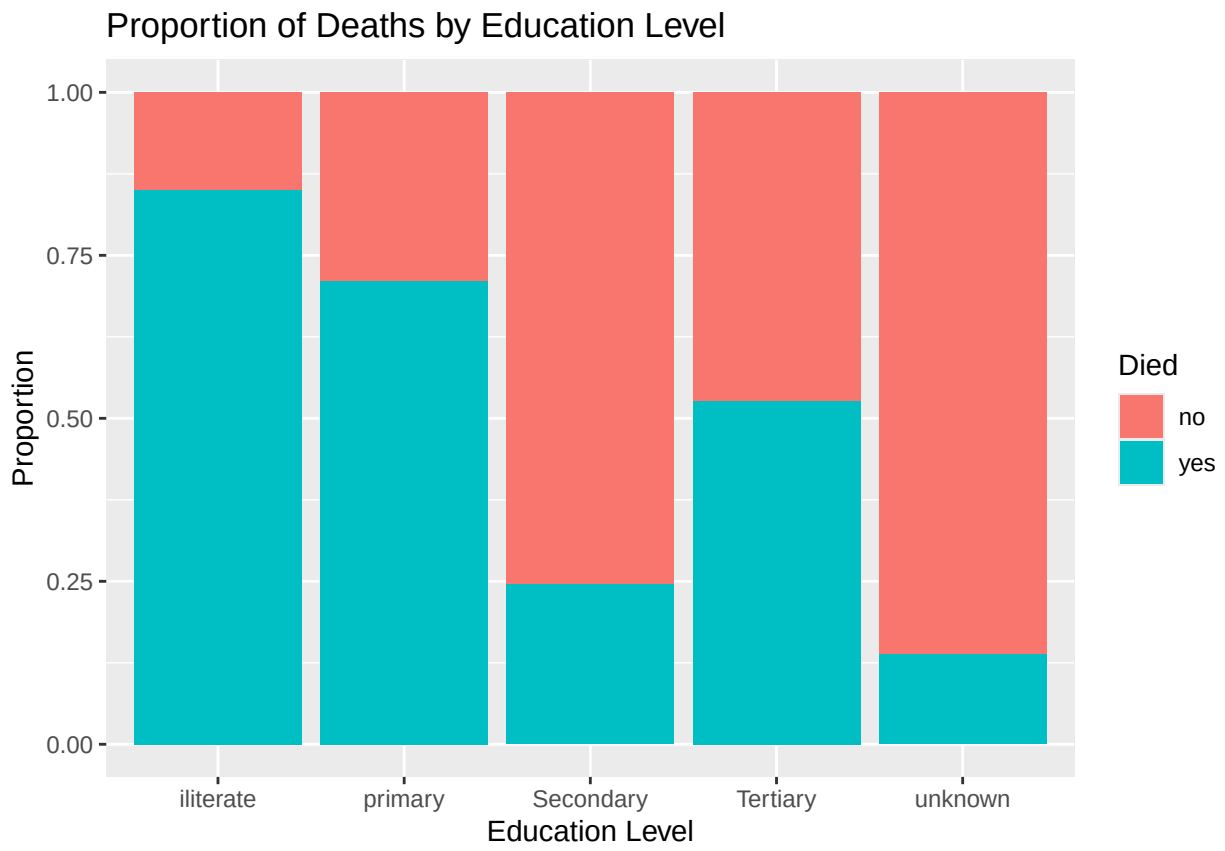

```
ggplot(suicide_data, aes(x = Sex, fill = method)) +
  geom_bar(position = "fill") +
  labs(title = "Method of Suicide by Sex", x = "Sex", y = "Proportion", fill = "Method")
```



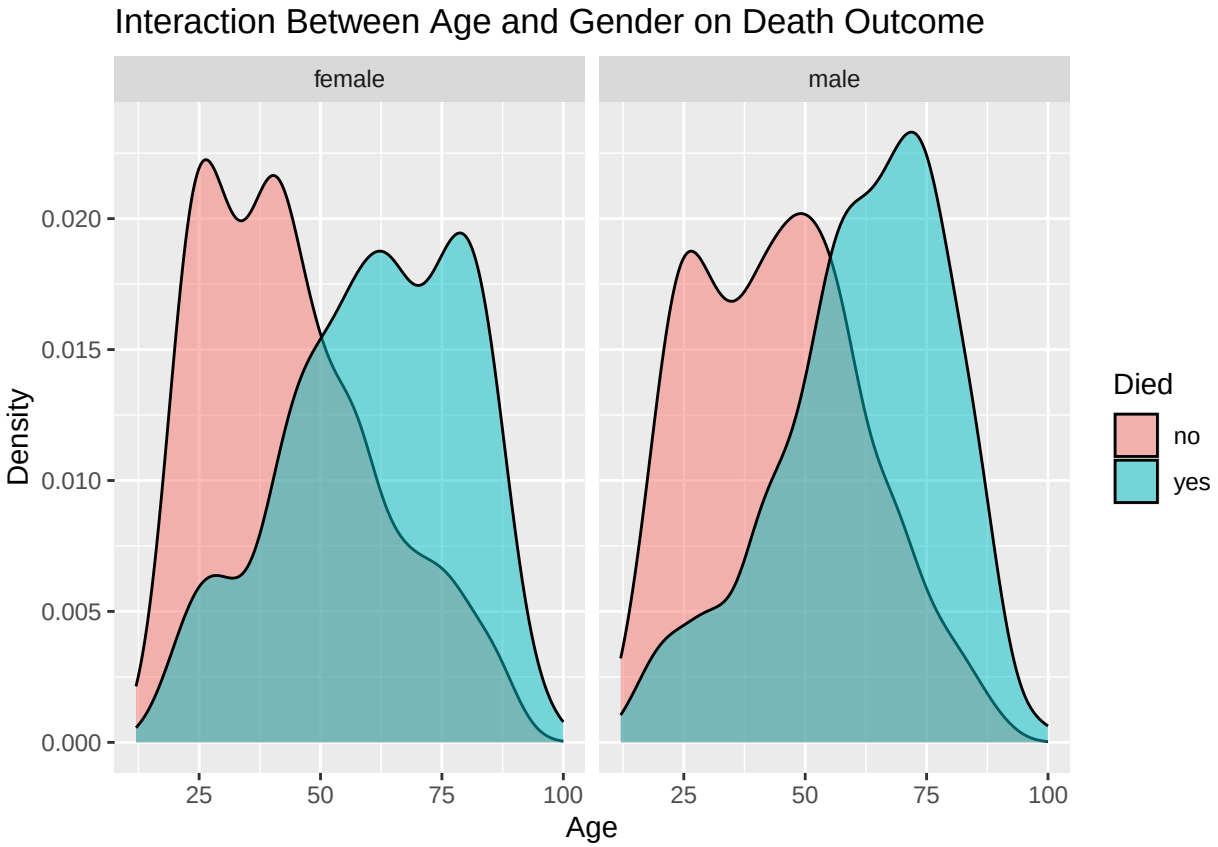

```
ggplot(suicide_data, aes(x = Age, fill = Died)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~Urban) +
  labs(title = "Age Distribution by Outcome and Urban/Rural", x = "Age", y = "Density", fill = "Died")
```



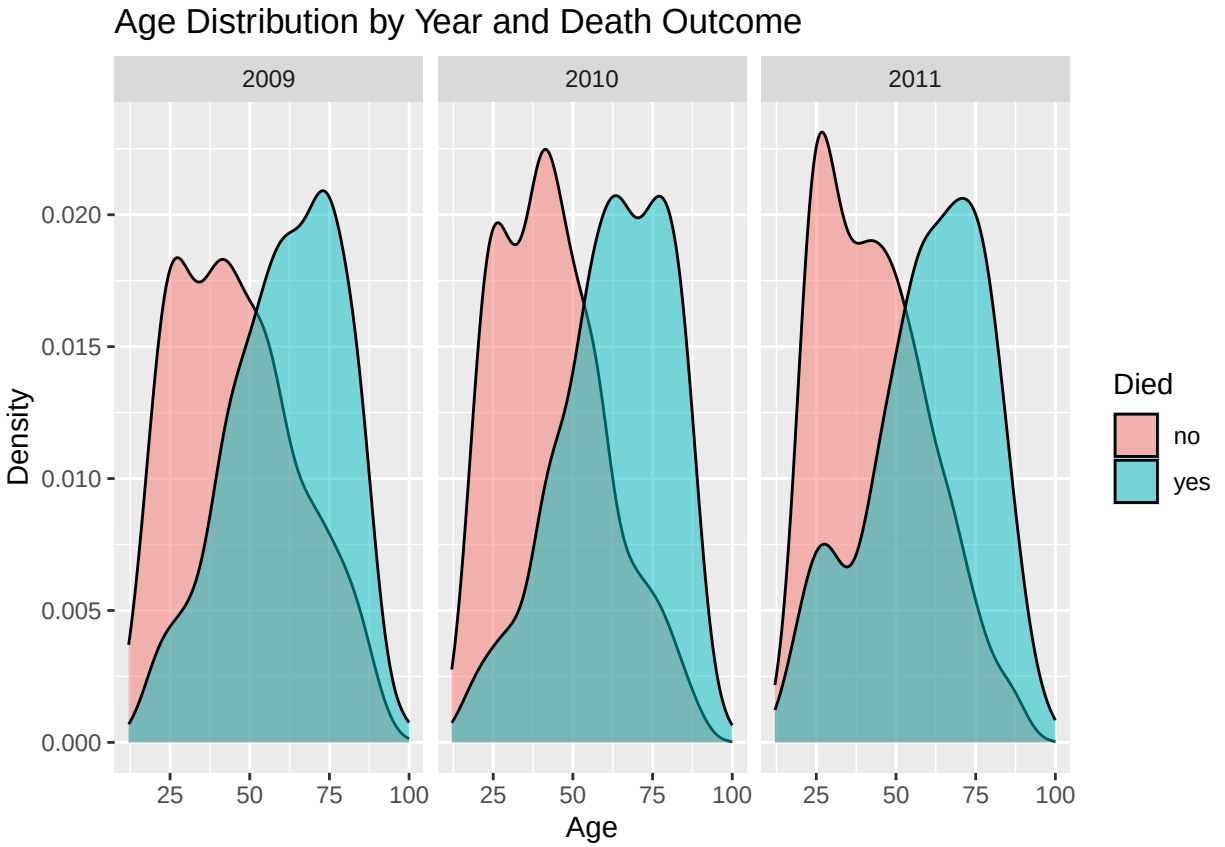

```
ggplot(suicide_data, aes(x = Education, fill = Died)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Deaths by Education Level", x = "Education Level", y = "Proportion", fill
```



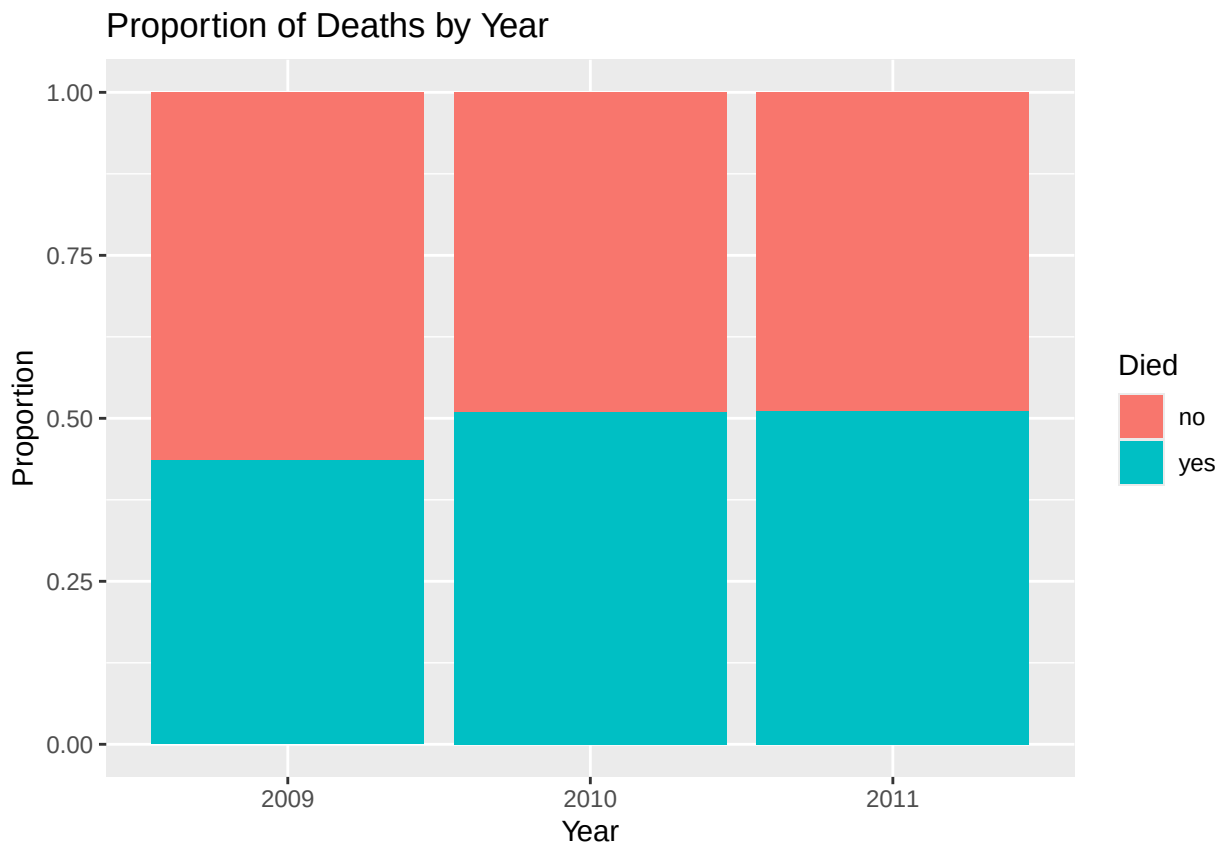

```r
# Interaction between Age and Gender on Death Outcome
ggplot(suicide_data, aes(x = Age, fill = Died)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~Sex) +
  labs(
    title = "Interaction Between Age and Gender on Death Outcome",
    x = "Age",
    y = "Density",
    fill = "Died"
  )
```



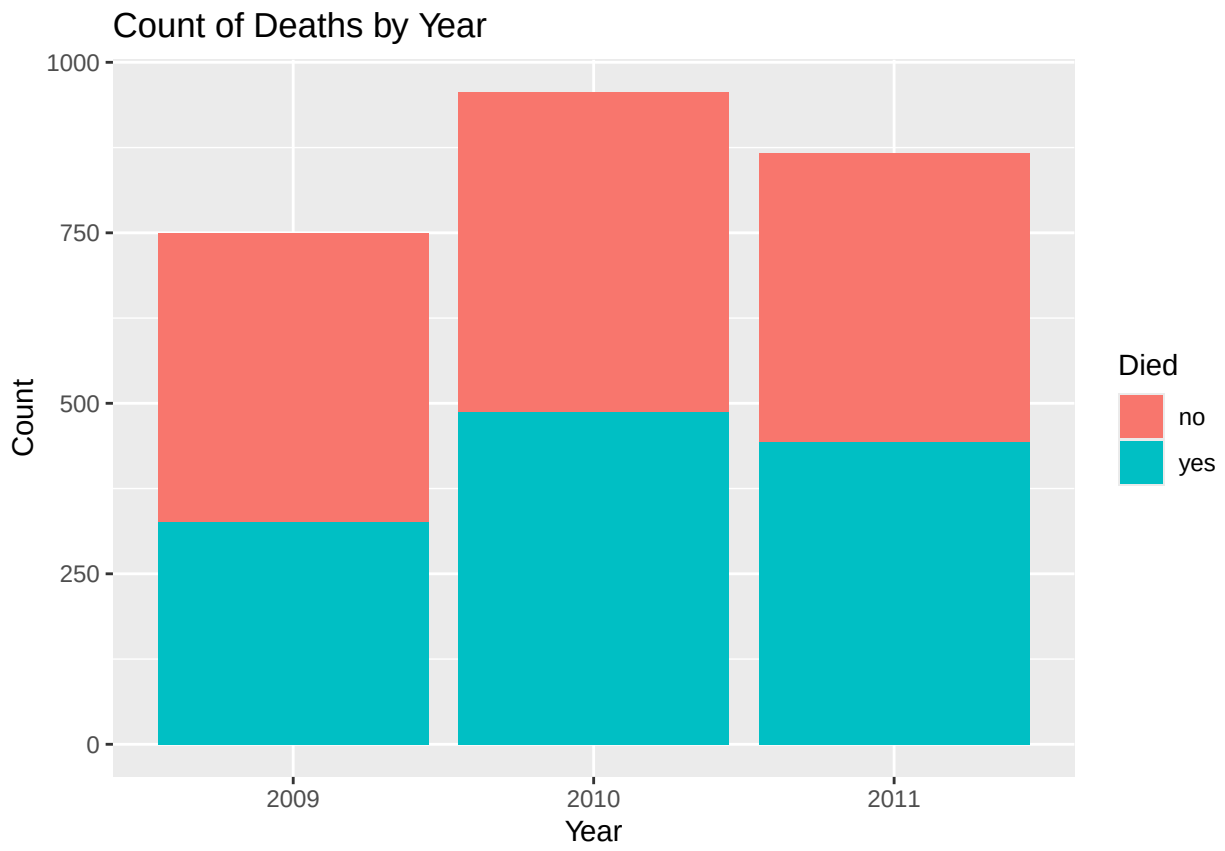

```r
ggplot(suicide_data, aes(x = Age, fill = Died)) +
  geom_density(alpha = 0.5) +
  facet_wrap(~Year) +
  labs(
    title = "Age Distribution by Year and Death Outcome",
    x = "Age",
    y = "Density",
    fill = "Died"
  )
```




```
ggplot(suicide_data, aes(x = as.factor(Year), fill = Died)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion of Deaths by Year",
    x = "Year",
    y = "Proportion",
    fill = "Died"
  )
```

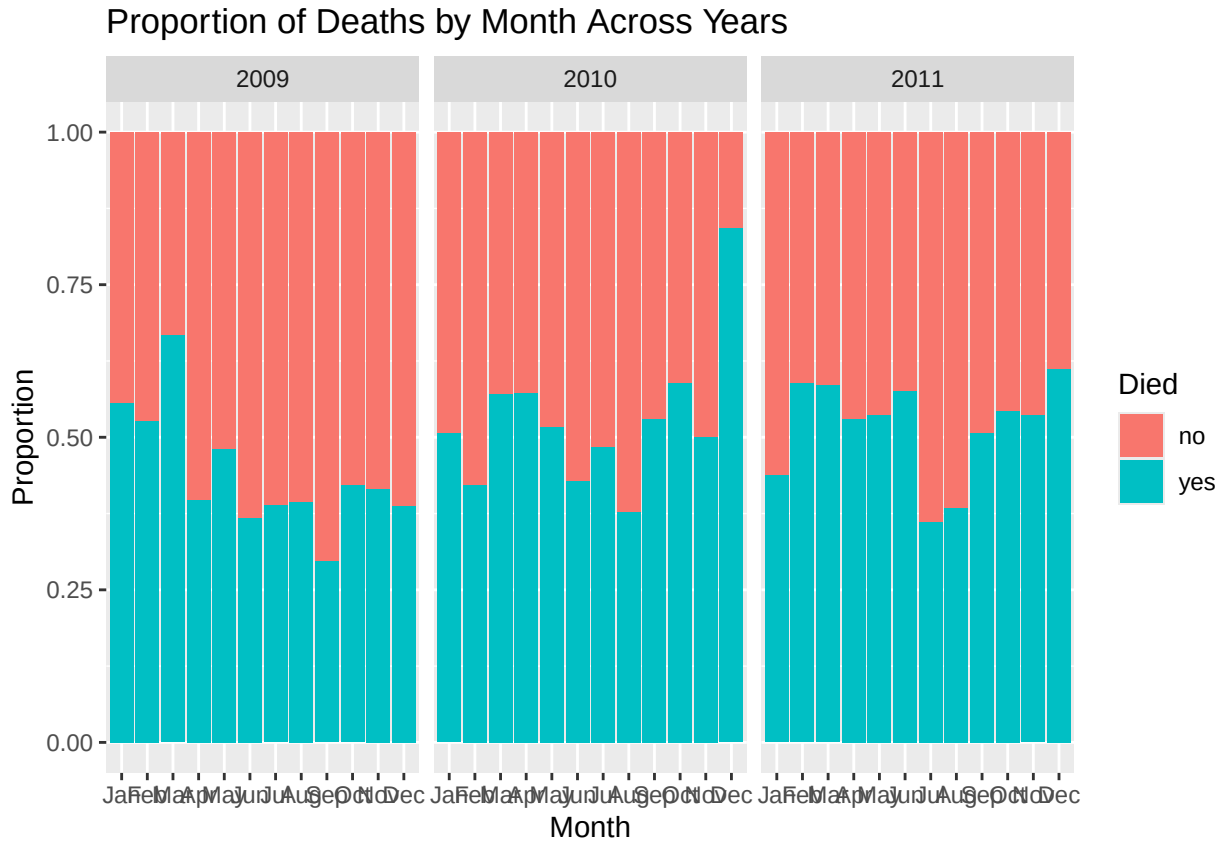



```
ggplot(suicide_data, aes(x = as.factor(Year), fill = Died)) +
  geom_bar() +
  labs(
    title = "Count of Deaths by Year",
    x = "Year",
    y = "Count",
    fill = "Died"
  )
```




```r
library(ggplot2)

ggplot(suicide_data, aes(x = as.factor(Month), fill = Died)) +
  geom_bar(position = "fill") +
  facet_wrap(~ Year) +
  labs(
    title = "Proportion of Deaths by Month Across Years",
    x = "Month",
    y = "Proportion",
    fill = "Died"
  ) +
  scale_x_discrete(labels = c("1" = "Jan", "2" = "Feb", "3" = "Mar", "4" = "Apr",
                              "5" = "May", "6" = "Jun", "7" = "Jul", "8" = "Aug",
                              "9" = "Sep", "10" = "Oct", "11" = "Nov", "12" = "Dec"))
```




```r
# Convert 'Died' to a factor variable
suicide_data$Died <- factor(suicide_data$Died, levels = c("no", "yes"))


logistic_model <- glm(Died ~ Age + Sex + Education + Occupation + method + Urban + Year + Month,
                      data = suicide_data,
                      family = binomial)

# Summary of the model
summary(logistic_model)
```

```
##
## Call:
## glm(formula = Died ~ Age + Sex + Education + Occupation + method +
##     Urban + Year + Month, family = binomial, data = suicide_data)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -7.440e+02  1.465e+02  -5.077 3.84e-07 ***
## Age                       1.922e-02  3.807e-03   5.050 4.43e-07 ***
## Sexmale                   2.091e-01  1.153e-01   1.814 0.069718 .
## Educationprimary         -7.085e-01  1.799e-01  -3.938 8.22e-05 ***
## EducationSecondary       -2.381e+00  1.881e-01 -12.658  < 2e-16 ***
## EducationTertiary        -1.429e+00  7.591e-01  -1.883 0.059736 .
## Educationunknown         -2.317e+00  6.461e-01  -3.586 0.000336 ***
```

```
## Occupationfarming        -7.934e-01  6.082e-01  -1.305 0.192040
## Occupationhousehold      -1.983e+00  6.385e-01  -3.106 0.001898 **
## Occupationothers          1.478e+01  1.164e+03   0.013 0.989867
## Occupationothers/unknown -2.106e+00  7.875e-01  -2.674 0.007493 **
## Occupationprofessional    8.011e-01  7.713e-01   1.039 0.298946
## Occupationretiree        -1.472e+01  1.250e+03  -0.012 0.990603
## Occupationstudent        -5.037e-01  7.648e-01  -0.659 0.510157
## Occupationunemployed     -1.184e+00  9.745e-01  -1.215 0.224225
## Occupationworker          1.552e+01  7.445e+02   0.021 0.983367
## methodDrowning            1.738e+01  4.064e+02   0.043 0.965889
## methodHanging             4.310e+00  6.010e-01   7.172 7.41e-13 ***
## methodJumping             4.227e+00  1.044e+00   4.051 5.11e-05 ***
## methodOther poison       -1.099e+00  5.947e-01  -1.848 0.064613 .
## methodOthers             -1.350e+01  2.400e+03  -0.006 0.995513
## methodPesticide           8.448e-01  5.079e-01   1.663 0.096219 .
## methodPoison unspec      -2.464e+00  7.991e-01  -3.083 0.002051 **
## methodunspecified        -2.064e+00  8.224e-01  -2.510 0.012089 *
## Urbanunknown             -4.064e+00  8.807e-01  -4.614 3.95e-06 ***
## Urbanyes                  1.367e-02  1.926e-01   0.071 0.943437
## Year                      3.703e-01  7.293e-02   5.078 3.81e-07 ***
## Month                    -1.905e-02  1.783e-02  -1.069 0.285205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3562.8  on 2570  degrees of freedom
## Residual deviance: 1988.0  on 2543  degrees of freedom
## AIC: 2044
##
## Number of Fisher Scoring iterations: 15
```

```r
# Odds Ratios
exp(coef(logistic_model))
```

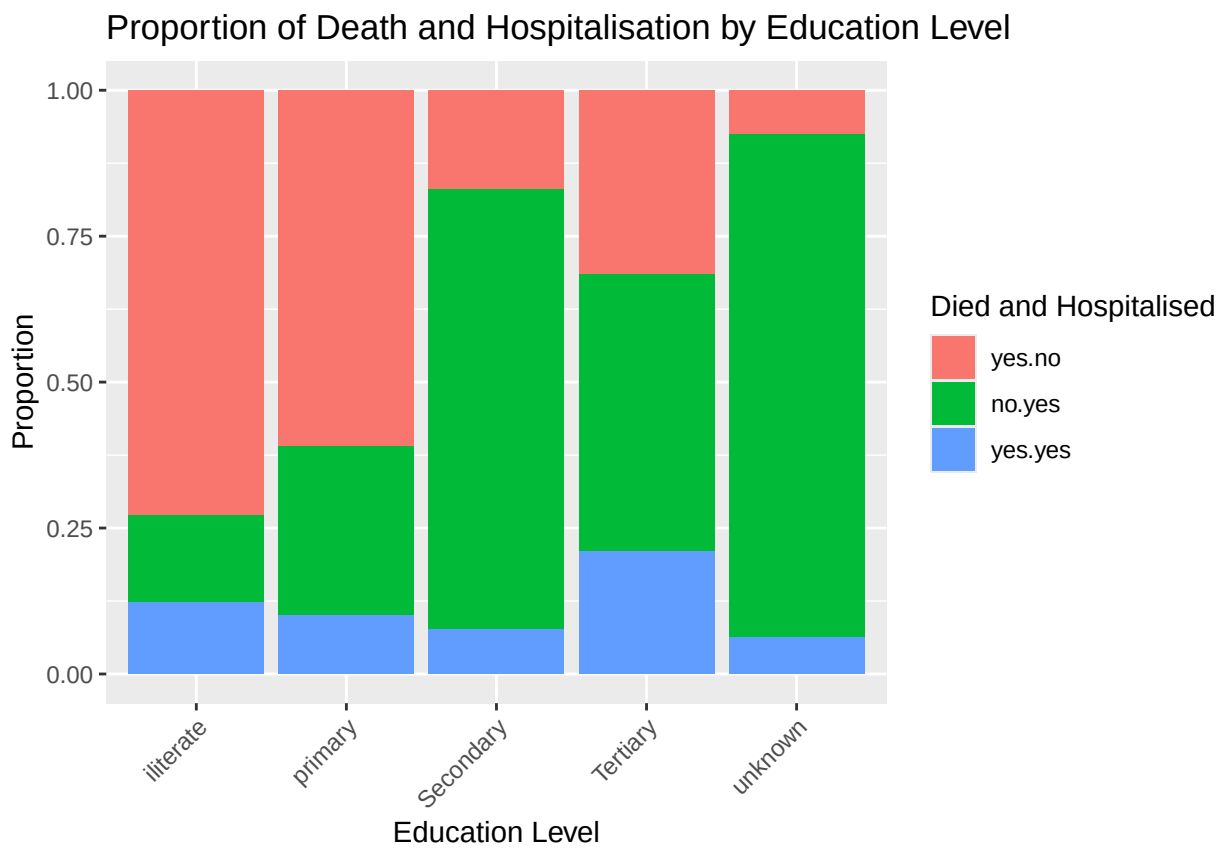
```
##              (Intercept)                      Age                  Sexmale
##            9.881313e-324             1.019411e+00             1.232597e+00
##         Educationprimary       EducationSecondary        EducationTertiary
##            4.923818e-01             9.245794e-02             2.395159e-01
##         Educationunknown        Occupationfarming      Occupationhousehold
##            9.860135e-02             4.522864e-01             1.376728e-01
##         Occupationothers Occupationothers/unknown   Occupationprofessional
##            2.619076e+06             1.217369e-01             2.228024e+00
##        Occupationretiree        Occupationstudent     Occupationunemployed
##            4.039341e-07             6.042897e-01             3.059355e-01
##         Occupationworker           methodDrowning            methodHanging
##            5.504952e+06             3.530801e+07             7.446246e+01
##            methodJumping       methodOther poison             methodOthers
##            6.852308e+01             3.331906e-01             1.376897e-06
##          methodPesticide      methodPoison unspec        methodunspecified
##            2.327507e+00             8.513527e-02             1.269561e-01
##             Urbanunknown                 Urbanyes                     Year
##            1.718620e-02             1.013761e+00             1.448229e+00
##                    Month
```

```
##             9.811273e-01
```

```r
# Count the frequency of each occupation
occupation_counts <- table(suicide_data$Occupation)
print(occupation_counts)
```

```
##
## business/service          farming        household           others
##               21             2032              248                3
##   others/unknown     professional          retiree          student
##              156               37                3               35
##       unemployed           worker
##               30                6
```

```r
# Create a stacked bar plot for Died and Hospitalised by Education
ggplot(suicide_data, aes(x = Education, fill = interaction(Died, Hospitalised))) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion of Death and Hospitalisation by Education Level",
    x = "Education Level",
    y = "Proportion",
    fill = "Died and Hospitalised"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
# Count the frequency of each occupation
educationlevels_counts <- table(suicide_data$Education)
print(educationlevels_counts)
```

```
##
## iliterate   primary Secondary  Tertiary   unknown
##       533       659      1280        19        80
```

```r
# Count the frequency of each occupation
month_counts <- table(suicide_data$Month)
print(month_counts)
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12
## 201 208 190 208 263 284 247 229 241 211 153 136
```

```r
# Count the frequency of each occupation
age_counts <- table(suicide_data$Age)
print(age_counts)
```

```
##
##  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31
##   1   3   5   5   9  12   9  21  29  32  34  39  40  42  40  34  37  29  38  31
##  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  33  29  23  29  18  28  32  37  58  45  45  52  28  34  38  42  53  36  47  35
##  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71
##  39  33  48  38  53  54  52  40  45  34  29  43  30  44  39  35  38  35  37  27
##  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91
##  37  30  39  45  37  33  29  31  30  30  28  27  29  24  25  15  23   8   9   2
##  92  94  95  96  97  98 100
##   6   1   2   1   1   3   1
```

```r
hos <- table(suicide_data$Hospitalised)
print(hos)
```

```
##
##   no  yes
## 1018 1553
```

```r
urban <- table(suicide_data$Urban)
print(urban)
```

```
##
##      no unknown     yes
##    2213      81     277
```

```r
# Filter dataset to include only farming and household occupations
suicide_data <- suicide_data %>%
  filter(Occupation %in% c("farming", "household"))
```

```r
# Filter for farming and household occupations and remove "unknown" values
cleaned_data <- suicide_data %>%
  filter(Occupation %in% c("farming", "household")) %>%   # Keep only farming and household
  filter(!Education %in% c("unknown"),                    # Remove "unknown" in Education
         !Died %in% c("unknown"),                         # Remove "unknown" in Died
         !Hospitalised %in% c("unknown"),                 # Remove "unknown" in Hospitalised
         !method %in% c("unknown"),                       # Remove "unknown" in method
         !Urban %in% c("unknown"))                        # Remove "unknown" in Urban

# Check the size of the cleaned dataset
nrow(cleaned_data)
```

```
## [1] 2211
```

```r
cleaned_data <- cleaned_data %>%
  mutate(method = case_when(
    method %in% c("Poison", "Poison unspec", "Other poison") ~ "Poison",
    TRUE ~ method  # Keep other categories unchanged
  ))

# Check the updated counts for the 'method' column
table(cleaned_data$method)
```

```
##
##     Cutting    Drowning     Hanging     Jumping   Pesticide      Poison
##          22          24         405           7        1576         139
## unspecified
##          38
```

```r
# Ensure 'Died' is a binary factor with levels "no" and "yes"
cleaned_data$Died <- factor(cleaned_data$Died, levels = c("no", "yes"))

# Build the logistic regression model
death_model <- glm(Died ~ Age + Sex + Education + method + Urban + Year + Month,
                   data = cleaned_data,
                   family = binomial)

# View model summary
summary(death_model)
```

```
##
## Call:
## glm(formula = Died ~ Age + Sex + Education + method + Urban +
##     Year + Month, family = binomial, data = cleaned_data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -6.424e+02  1.471e+02  -4.366 1.27e-05 ***
## Age                 1.830e-02  3.901e-03   4.690 2.73e-06 ***
## Sexmale             3.414e-01  1.162e-01   2.937  0.00331 **
## Educationprimary   -8.244e-01  1.795e-01  -4.593 4.37e-06 ***
## EducationSecondary -2.450e+00  1.905e-01 -12.861  < 2e-16 ***
```

```
## EducationTertiary  -8.442e-01  1.103e+00  -0.765  0.44408
## methodDrowning      1.756e+01  4.182e+02   0.042  0.96650
## methodHanging       4.108e+00  6.235e-01   6.589 4.44e-11 ***
## methodJumping       3.767e+00  1.232e+00   3.058  0.00223 **
## methodPesticide     9.481e-01  5.403e-01   1.755  0.07932 .
## methodPoison       -1.372e+00  6.093e-01  -2.252  0.02432 *
## methodunspecified  -1.799e+00  8.424e-01  -2.135  0.03275 *
## Urbanyes           -2.436e-02  2.111e-01  -0.115  0.90816
## Year                3.193e-01  7.322e-02   4.362 1.29e-05 ***
## Month              -1.981e-02  1.842e-02  -1.075  0.28216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3055.6  on 2210  degrees of freedom
## Residual deviance: 1879.4  on 2196  degrees of freedom
## AIC: 1909.4
##
## Number of Fisher Scoring iterations: 15
```

```r
# Calculate odds ratios for interpretation
exp(coef(death_model))
```

```
##        (Intercept)                Age            Sexmale   Educationprimary
##      1.009225e-279       1.018466e+00       1.406968e+00       4.384955e-01
## EducationSecondary  EducationTertiary     methodDrowning      methodHanging
##       8.632843e-02       4.298827e-01       4.243089e+07       6.084842e+01
##      methodJumping    methodPesticide       methodPoison  methodunspecified
##       4.323324e+01       2.580819e+00       2.535863e-01       1.655098e-01
##           Urbanyes               Year              Month
##       9.759392e-01       1.376226e+00       9.803862e-01
```

```r
# Create a new column 'Season' based on the month
cleaned_data <- cleaned_data %>%
  mutate(Season = case_when(
    Month %in% c(12, 1, 2) ~ "Winter",
    Month %in% c(3, 4, 5) ~ "Spring",
    Month %in% c(6, 7, 8) ~ "Summer",
    Month %in% c(9, 10, 11) ~ "Fall"
  ))

# Check the counts for each season
table(cleaned_data$Season)
```

```
##
##   Fall Spring Summer Winter
##    536    570    654    451
```

```r
# Update the logistic regression model with 'Season' instead of 'Month'
death_model_season <- glm(Died ~ Age + Sex + Education + method + Urban + Year + Season,
                          data = cleaned_data,
```

```
                          family = binomial)

# View the model summary
summary(death_model_season)
```

```
##
## Call:
## glm(formula = Died ~ Age + Sex + Education + method + Urban +
##     Year + Season, family = binomial, data = cleaned_data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -6.324e+02  1.476e+02  -4.286 1.82e-05 ***
## Age                 1.765e-02  3.915e-03   4.507 6.56e-06 ***
## Sexmale             3.186e-01  1.168e-01   2.727  0.00638 **
## Educationprimary   -8.200e-01  1.803e-01  -4.547 5.44e-06 ***
## EducationSecondary -2.479e+00  1.917e-01 -12.927  < 2e-16 ***
## EducationTertiary  -8.729e-01  1.094e+00  -0.798  0.42481
## methodDrowning      1.757e+01  4.169e+02   0.042  0.96637
## methodHanging       4.136e+00  6.241e-01   6.626 3.44e-11 ***
## methodJumping       3.736e+00  1.232e+00   3.033  0.00242 **
## methodPesticide     9.840e-01  5.403e-01   1.821  0.06861 .
## methodPoison       -1.396e+00  6.094e-01  -2.291  0.02199 *
## methodunspecified  -1.698e+00  8.394e-01  -2.023  0.04307 *
## Urbanyes           -6.548e-03  2.100e-01  -0.031  0.97512
## Year                3.142e-01  7.342e-02   4.280 1.87e-05 ***
## SeasonSpring        4.197e-01  1.642e-01   2.557  0.01057 *
## SeasonSummer       -5.829e-02  1.560e-01  -0.374  0.70868
## SeasonWinter        4.031e-01  1.749e-01   2.304  0.02121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3055.6  on 2210  degrees of freedom
## Residual deviance: 1865.8  on 2194  degrees of freedom
## AIC: 1899.8
##
## Number of Fisher Scoring iterations: 15
```

```
# Calculate odds ratios for interpretation
exp(coef(death_model_season))
```

```
##        (Intercept)                Age            Sexmale   Educationprimary
##      2.227514e-275       1.017804e+00       1.375154e+00       4.404200e-01
## EducationSecondary  EducationTertiary     methodDrowning      methodHanging
##       8.385456e-02       4.177344e-01       4.290448e+07       6.252607e+01
##      methodJumping    methodPesticide       methodPoison  methodunspecified
##       4.194650e+01       2.675025e+00       2.476527e-01       1.830092e-01
##           Urbanyes               Year       SeasonSpring       SeasonSummer
##       9.934738e-01       1.369212e+00       1.521546e+00       9.433714e-01
##       SeasonWinter
##       1.496428e+00
```

```r
# Convert Year into a categorical variable
cleaned_data$Year <- as.factor(cleaned_data$Year)

# Check the levels to confirm
levels(cleaned_data$Year)
```

```
## [1] "2009" "2010" "2011"
```

```r
# Logistic regression with Year as a categorical variable
death_model_categorical <- glm(Died ~ Age + Sex + Education + method + Urban + Season + Year,
                               data = cleaned_data,
                               family = binomial)

# View model summary
summary(death_model_categorical)
```

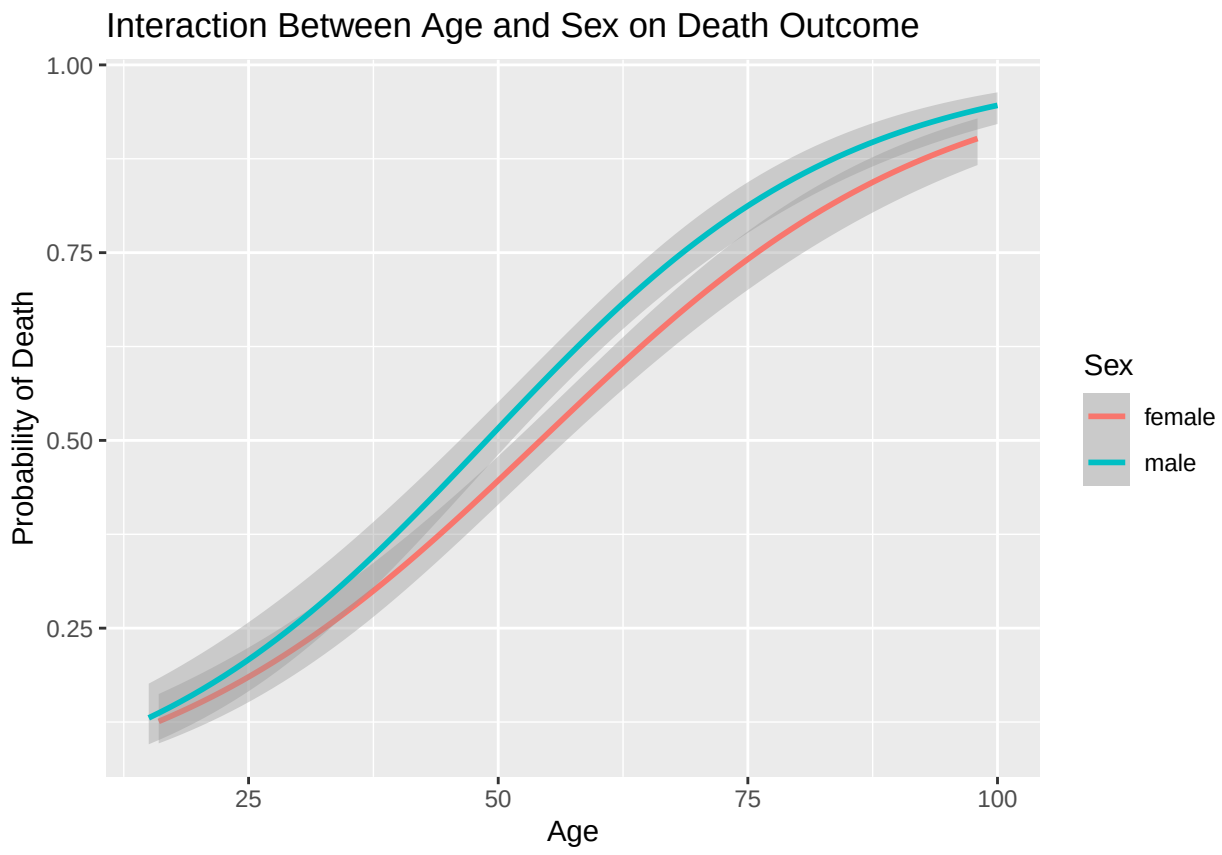
```
##
## Call:
## glm(formula = Died ~ Age + Sex + Education + method + Urban +
##     Season + Year, family = binomial, data = cleaned_data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.122755   0.607753  -1.847  0.06469 .
## Age                  0.017616   0.003917   4.497 6.88e-06 ***
## Sexmale              0.318369   0.116802   2.726  0.00642 **
## Educationprimary    -0.822828   0.180522  -4.558 5.16e-06 ***
## EducationSecondary  -2.480853   0.191859 -12.931  < 2e-16 ***
## EducationTertiary   -0.860457   1.093953  -0.787  0.43154
## methodDrowning      17.577865 416.561264   0.042  0.96634
## methodHanging        4.133303   0.624163   6.622 3.54e-11 ***
## methodJumping        3.740971   1.232428   3.035  0.00240 **
## methodPesticide      0.984014   0.540350   1.821  0.06860 .
## methodPoison        -1.389649   0.609688  -2.279  0.02265 *
## methodunspecified   -1.685206   0.840167  -2.006  0.04488 *
## Urbanyes            -0.007706   0.210086  -0.037  0.97074
## SeasonSpring         0.421035   0.164236   2.564  0.01036 *
## SeasonSummer        -0.058459   0.156024  -0.375  0.70790
## SeasonWinter         0.402815   0.174888   2.303  0.02126 *
## Year2010             0.356511   0.141806   2.514  0.01193 *
## Year2011             0.628978   0.146923   4.281 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3055.6  on 2210  degrees of freedom
## Residual deviance: 1865.7  on 2193  degrees of freedom
## AIC: 1901.7
##
## Number of Fisher Scoring iterations: 15
```

```r
# Calculate odds ratios
exp(coef(death_model_categorical))
```

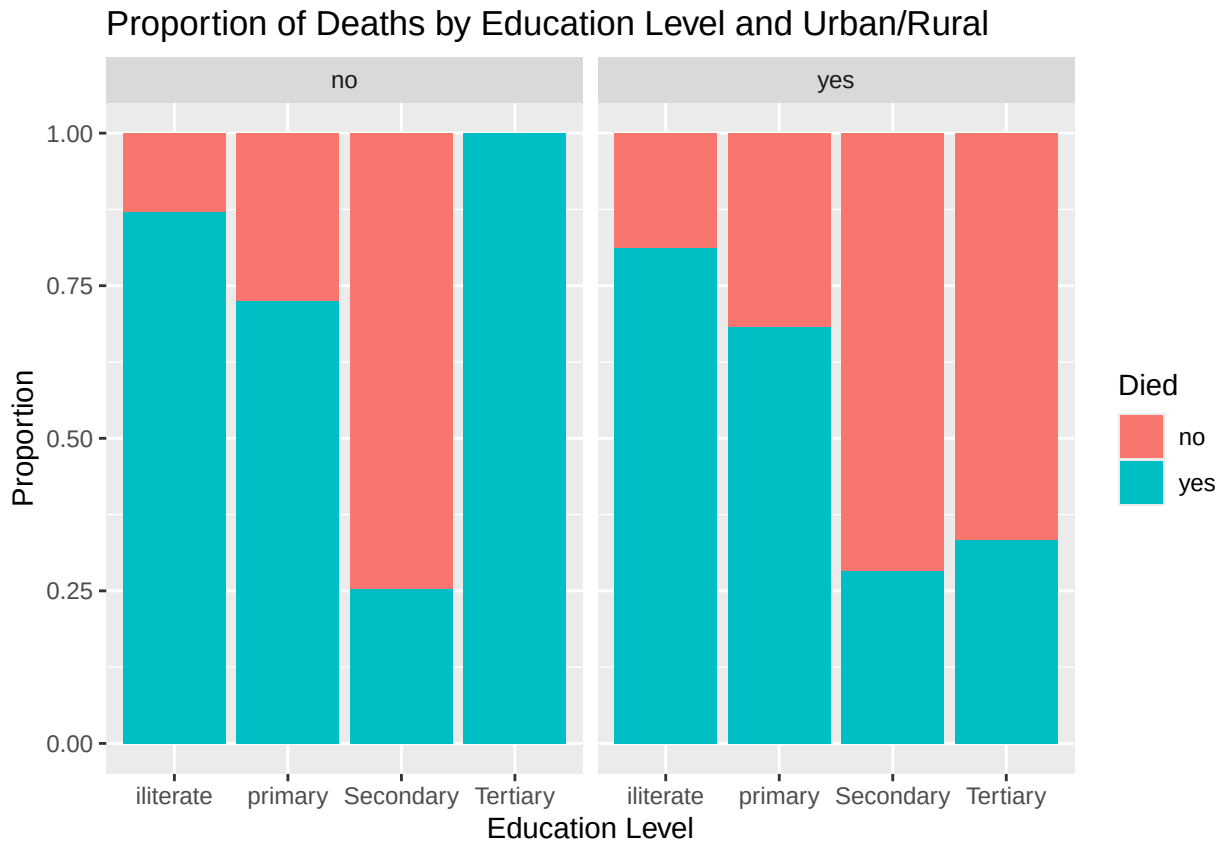
```
##         (Intercept)                 Age             Sexmale   Educationprimary
##        3.253822e-01        1.017772e+00        1.374883e+00        4.391877e-01
## EducationSecondary  EducationTertiary      methodDrowning      methodHanging
##        8.367179e-02        4.229688e-01        4.304966e+07        6.238363e+01
##      methodJumping    methodPesticide        methodPoison  methodunspecified
##        4.213887e+01        2.675174e+00        2.491627e-01        1.854062e-01
##            Urbanyes        SeasonSpring        SeasonSummer        SeasonWinter
##        9.923237e-01        1.523537e+00        9.432171e-01        1.496029e+00
##            Year2010            Year2011
##        1.428337e+00        1.875693e+00
```

```r
# Plot Age and Sex interaction
ggplot(cleaned_data, aes(x = Age, y = as.numeric(Died == "yes"), color = Sex)) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(
    title = "Interaction Between Age and Sex on Death Outcome",
    x = "Age",
    y = "Probability of Death"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
# Create a bar plot showing proportions
ggplot(cleaned_data, aes(x = Education, fill = Died)) +
  geom_bar(position = "fill") +
  facet_wrap(~ Urban) +
  labs(
    title = "Proportion of Deaths by Education Level and Urban/Rural",
    x = "Education Level",
    y = "Proportion",
    fill = "Died"
  )
```




```r
# Add Education and Urban interaction to the model
final_model <- glm(Died ~ Age + Education + method + Season + Year + Sex,
                   data = cleaned_data,
                   family = binomial)

# View the model summary
summary(final_model)
```

```
##
## Call:
## glm(formula = Died ~ Age + Education + method + Season + Year +
##     Sex, family = binomial, data = cleaned_data)
##
## Coefficients:
```

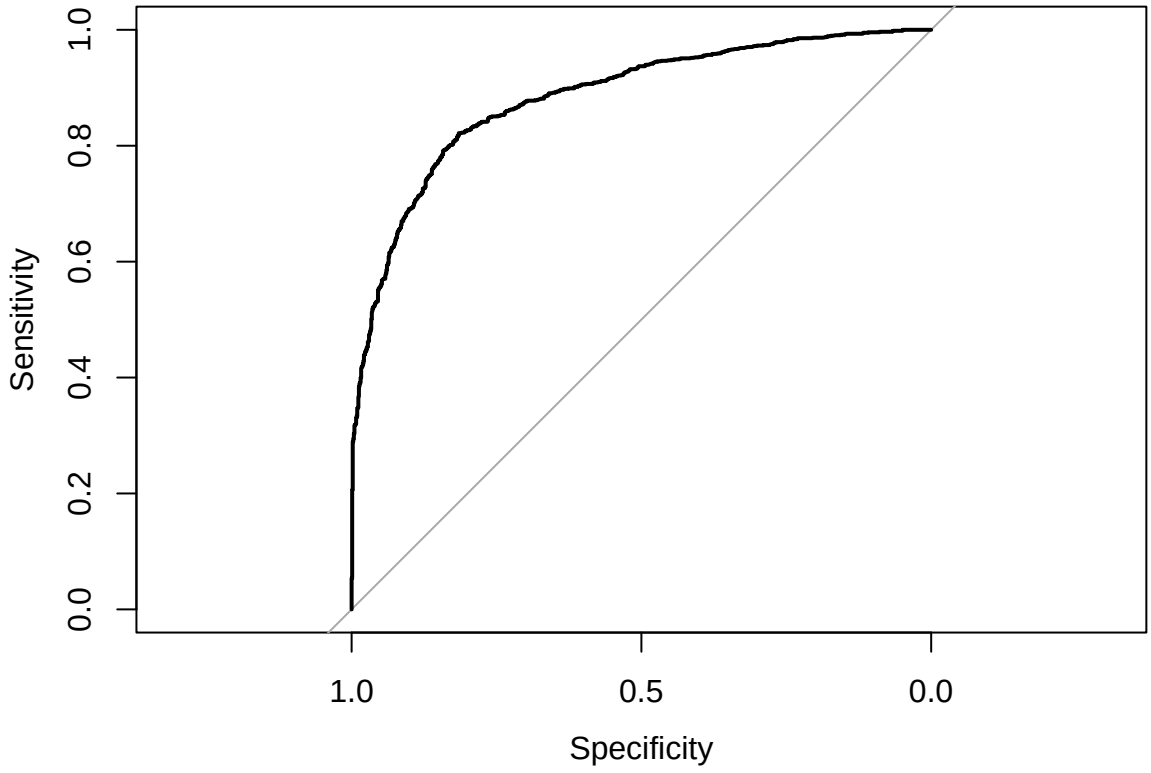
```
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.123417   0.607402  -1.850  0.06438 .
## Age                  0.017616   0.003917   4.498 6.87e-06 ***
## Educationprimary    -0.822793   0.180521  -4.558 5.17e-06 ***
## EducationSecondary  -2.481094   0.191751 -12.939  < 2e-16 ***
## EducationTertiary   -0.865654   1.084484  -0.798  0.42474
## methodDrowning      17.578842 416.613360   0.042  0.96634
## methodHanging        4.133257   0.624076   6.623 3.52e-11 ***
## methodJumping        3.737283   1.228425   3.042  0.00235 **
## methodPesticide      0.984085   0.540250   1.822  0.06853 .
## methodPoison        -1.389876   0.609575  -2.280  0.02260 *
## methodunspecified   -1.684696   0.839992  -2.006  0.04490 *
## SeasonSpring         0.421115   0.164219   2.564  0.01034 *
## SeasonSummer        -0.058410   0.156023  -0.374  0.70813
## SeasonWinter         0.402991   0.174818   2.305  0.02116 *
## Year2010             0.356449   0.141794   2.514  0.01194 *
## Year2011             0.628994   0.146920   4.281 1.86e-05 ***
## Sexmale              0.318367   0.116802   2.726  0.00642 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3055.6  on 2210  degrees of freedom
## Residual deviance: 1865.7  on 2194  degrees of freedom
## AIC: 1899.7
##
## Number of Fisher Scoring iterations: 15
```

```
roc_curve <- roc(cleaned_data$Died, predict(death_model_categorical, type = "response"))
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```

```
plot(roc_curve)
```




```
auc(roc_curve)
```

```
## Area under the curve: 0.884
```

```
final_model <- glm(Died ~ Hospitalised + method + Year + Education + Season + Sex + Age,
                   data = cleaned_data,
                   family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(final_model)
```

```
##
## Call:
## glm(formula = Died ~ Hospitalised + method + Year + Education +
##     Season + Sex + Age, family = binomial, data = cleaned_data)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.493e+01  7.568e+02   0.020 0.984265
## Hospitalisedyes    -3.243e+01  1.007e+03  -0.032 0.974313
## methodDrowning      3.339e+01  4.648e+03   0.007 0.994268
## methodHanging       1.700e+01  7.558e+02   0.022 0.982055
## methodJumping       1.792e+01  7.558e+02   0.024 0.981080
```

```
## methodPesticide     1.533e+01  7.558e+02   0.020 0.983815
## methodPoison        1.387e+01  7.558e+02   0.018 0.985363
## methodunspecified   1.394e+01  7.558e+02   0.018 0.985283
## Year2010            8.576e-01  2.313e-01   3.708 0.000209 ***
## Year2011            1.178e+00  2.327e-01   5.062 4.15e-07 ***
## Educationprimary   -7.934e-01  2.502e-01  -3.171 0.001517 **
## EducationSecondary -1.823e+00  2.624e-01  -6.949 3.68e-12 ***
## EducationTertiary  -1.580e+01  1.226e+03  -0.013 0.989713
## SeasonSpring        4.662e-01  2.424e-01   1.923 0.054449 .
## SeasonSummer        7.509e-02  2.322e-01   0.323 0.746447
## SeasonWinter        3.005e-01  2.632e-01   1.142 0.253499
## Sexmale             2.495e-01  1.714e-01   1.456 0.145501
## Age                 1.731e-02  5.712e-03   3.030 0.002445 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3055.58  on 2210  degrees of freedom
## Residual deviance:  917.82  on 2193  degrees of freedom
## AIC: 953.82
##
## Number of Fisher Scoring iterations: 20
```

```
model_no_hospital <- glm(Died ~ method + Year + Education + Season + Sex + Age,
                         data = cleaned_data, family = binomial)

# Fit model with "Hospitalised"
model_with_hospital <- glm(Died ~ Hospitalised + method + Year + Education + Season + Sex + Age,
                           data = cleaned_data, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
vif(model_with_hospital)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## Hospitalised 1.997443  1        1.413309
## method       1.988658  6        1.058961
## Year         1.069229  2        1.016875
## Education    1.815731  3        1.104524
## Season       1.043236  3        1.007079
## Sex          1.028566  1        1.014182
## Age          1.488270  1        1.219947
```

```
# To create a combined outcome variable
cleaned_data$Outcome <- case_when(
  cleaned_data$Died == "yes" ~ "Died",
  cleaned_data$Hospitalised == "yes" ~ "Hospitalised",
  TRUE ~ "Not_Hospitalised"
)
```

```r
cleaned_data$Outcome <- factor(cleaned_data$Outcome, levels = c("Died", "Hospitalised", "Not_Hospitalis

# Fit multinomial logistic regression
multinom_model <- multinom(Outcome ~ Age + Sex + Education + method + Urban + Year + Season,
                           data = cleaned_data)
```

```
## Warning in multinom(Outcome ~ Age + Sex + Education + method + Urban + Year + :
## group 'Not_Hospitalised' is empty
```

```
## # weights:  19 (18 variable)
## initial  value 1532.548416
## iter  10 value 1041.765204
## iter  20 value 934.689216
## iter  30 value 932.876210
## final  value 932.857566
## converged
```

```r
summary(multinom_model)
```

```
## Call:
## multinom(formula = Outcome ~ Age + Sex + Education + method +
##     Urban + Year + Season, data = cleaned_data)
##
## Coefficients:
##                           Values    Std. Err.
## (Intercept)          1.121185453 0.6076831017
## Age                 -0.017617059 0.0039168338
## Sexmale             -0.318360809 0.1168017984
## Educationprimary     0.822794401 0.1805220224
## EducationSecondary   2.480809455 0.1918588024
## EducationTertiary    0.860499851 1.0939475861
## methodDrowning     -13.576910086 0.0001168028
## methodHanging       -4.131878063 0.6241095774
## methodJumping       -3.737037597 1.2316706742
## methodPesticide     -0.982339257 0.5402683633
## methodPoison         1.391329935 0.6096188991
## methodunspecified    1.687101091 0.8401504430
## Urbanyes             0.007664267 0.2100843534
## Year2010            -0.356531399 0.1418062522
## Year2011            -0.629011036 0.1469231265
## SeasonSpring        -0.421014794 0.1642362140
## SeasonSummer         0.058490551 0.1560242625
## SeasonWinter        -0.402782230 0.1748877657
##
## Residual Deviance: 1865.715
## AIC: 1901.715
```

```r
exp(coef(multinom_model))
```

```
##        (Intercept)                 Age             Sexmale   Educationprimary
##       3.068490e+00       9.825372e-01       7.273403e-01       2.276853e+00
```

```
## EducationSecondary  EducationTertiary     methodDrowning      methodHanging
##       1.195093e+01       2.364342e+00       1.269471e-06       1.605270e-02
##      methodJumping    methodPesticide       methodPoison  methodunspecified
##       2.382458e-02       3.744342e-01       4.020193e+00       5.403793e+00
##           Urbanyes           Year2010           Year2011       SeasonSpring
##       1.007694e+00       7.001005e-01       5.331188e-01       6.563804e-01
##       SeasonSummer       SeasonWinter
##       1.060235e+00       6.684577e-01
```

```
# fixed - fit three level
cleaned_data$Outcome <- case_when(
  cleaned_data$Died == "yes" ~ "High_Fatality",                    # Dead, regardless of hospitalization
  cleaned_data$Died == "no" & cleaned_data$Hospitalised == "yes" ~ "Medium_Fatality",  # Not dead but h
  cleaned_data$Died == "no" & cleaned_data$Hospitalised == "no" ~ "Low_Fatality"      # Not dead and no
)

cleaned_data$Outcome <- factor(cleaned_data$Outcome,
                               levels = c("High_Fatality", "Medium_Fatality", "Low_Fatality"))


table(cleaned_data$Outcome)
```

```
##
##   High_Fatality Medium_Fatality    Low_Fatality
##            1178            1033               0
```

```
#only fit the two level - medium and high?
cleaned_data$Outcome <- case_when(
  cleaned_data$Died == "yes" ~ "High_Fatality",                    # Dead (high-fatality)
  cleaned_data$Died == "no" & cleaned_data$Hospitalised == "yes" ~ "Medium_Fatality" # Survived and hos
)


cleaned_data$Outcome <- factor(cleaned_data$Outcome,
                               levels = c("High_Fatality", "Medium_Fatality"))

table(cleaned_data$Outcome)
```

```
##
##   High_Fatality Medium_Fatality
##            1178            1033
```

```
# final model?
binary_model <- glm(Outcome ~ Age + Sex + Education + method + Urban + Year + Season,
                    data = cleaned_data,
                    family = binomial)

summary(binary_model)
```

```
##
## Call:
## glm(formula = Outcome ~ Age + Sex + Education + method + Urban +
```

```
##     Year + Season, family = binomial, data = cleaned_data)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.122755   0.607753   1.847  0.06469 .
## Age                  -0.017616   0.003917  -4.497 6.88e-06 ***
## Sexmale              -0.318369   0.116802  -2.726  0.00642 **
## Educationprimary      0.822828   0.180522   4.558 5.16e-06 ***
## EducationSecondary    2.480853   0.191859  12.931  < 2e-16 ***
## EducationTertiary     0.860457   1.093953   0.787  0.43154
## methodDrowning      -17.577865 416.561264  -0.042  0.96634
## methodHanging        -4.133303   0.624163  -6.622 3.54e-11 ***
## methodJumping        -3.740971   1.232428  -3.035  0.00240 **
## methodPesticide      -0.984014   0.540350  -1.821  0.06860 .
## methodPoison          1.389649   0.609688   2.279  0.02265 *
## methodunspecified     1.685206   0.840167   2.006  0.04488 *
## Urbanyes              0.007706   0.210086   0.037  0.97074
## Year2010             -0.356511   0.141806  -2.514  0.01193 *
## Year2011             -0.628978   0.146923  -4.281 1.86e-05 ***
## SeasonSpring         -0.421035   0.164236  -2.564  0.01036 *
## SeasonSummer          0.058459   0.156024   0.375  0.70790
## SeasonWinter         -0.402815   0.174888  -2.303  0.02126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3055.6  on 2210  degrees of freedom
## Residual deviance: 1865.7  on 2193  degrees of freedom
## AIC: 1901.7
##
## Number of Fisher Scoring iterations: 15
```

```r
# Convert coefficients to odds ratios
exp(coef(binary_model))
```

```
##        (Intercept)                Age            Sexmale   Educationprimary
##       3.073309e+00       9.825386e-01       7.273344e-01       2.276931e+00
## EducationSecondary  EducationTertiary     methodDrowning      methodHanging
##       1.195146e+01       2.364241e+00       2.322899e-08       1.602985e-02
##      methodJumping    methodPesticide       methodPoison  methodunspecified
##       2.373106e-02       3.738075e-01       4.013441e+00       5.393562e+00
##           Urbanyes           Year2010           Year2011       SeasonSpring
##       1.007736e+00       7.001148e-01       5.331363e-01       6.563672e-01
##       SeasonSummer       SeasonWinter
##       1.060201e+00       6.684360e-01
```