Final_Project

Elijah Russell

2025-12-11

Section 1: Introduction

The Data

To start this project I began by exploring what data set I wanted to work with. Of the initial data repositories that were suggested, the NHS Scotland Open Data immediately caught my eye. This repository particularly interested me because I have a passion for health and medical data and endeavor to have a career as a healthcare data analyst, as well as because my family emigrated from Scotland not too long ago. Next, it was a matter of finding a data set that was both interesting to me and met the specifications of the project: at least 50 observations, between 10 to 20 variables, and including categorical variables, discrete numerical variables, and continuous numerical variables. None of the first few data sets I checked had any continuous numerical variables, so I then started browsing for incidents rate data sets and shortly found the Annual Cancer Incidence at Scotland Level data set that I settled on using. Although the data set initially had more than 20 variables, after removing a few unnecessary or repetitive variables, such as variables just full of NAs, the data set met all of the specifications.

The annual publication linked to the data indicates that the data came from the Scottish Cancer Registry which “collects information on every cancer in Scotland and uses the data to inform cancer control”. The data dictionary identifies ‘Cancer Site’ as “the diagnosed cancer site”, ‘Year’ as “Year of diagnosis”, ‘IncidencesAge…’ as “Number of new cancer registrations in the age group …”, and ‘IncidenceRateAge…’ as “The incidence rate for the age group” where the rate is new cancer registrations per 100,000 people.

The Research Question

The next step in my process was to develop a research question. As I was looking through the data set I was struck with the idea to ask a question close to my heart. I recently lost my Scottish grandfather due to complications of cancer therapy, and felt that doing this project to analyze the relation between the demographic and time variables in the data set and the specific types of cancer that my grandfather had. Thus my research question is: How do age group, sex, and year each contribute to variation in cancer incidence rates in Scotland, and which factor has the strongest association with incidence rate?

Section 2: Data Analysis

I start the process of data analysis by loading libraries and importing the dataset.

I then restructure the data to best answer the research question, first by filtering to the types of cancer my grandfather had, bile duct cancer, liver cancer, and lymphoma, and then by making age group its own column and renaming the now age group variables to more readable names.

```r
types <- c("Liver and intrahepatic bile ducts", "Hodgkin lymphoma", "Non-Hodgkin lymphoma")
```

```r
RestructuredData <- opendata_inc9923_scotland %>%
  pivot_longer(
    cols = starts_with("IncidenceRateAge"),
    names_to = "Age_group",
    values_to = "Incidence_Rate")
```

```
RestructuredData <- RestructuredData %>%
  filter(CancerSite %in% types) %>%
    mutate(Age_group = case_when(
      Age_group == "IncidenceRateAgeUnder20" ~ "Under_20",
      Age_group == "IncidenceRateAge20-29" ~ "20_29",
      Age_group == "IncidenceRateAge30-39" ~ "30_39",
      Age_group == "IncidenceRateAge40-49" ~ "40_49",
      Age_group == "IncidenceRateAge50-59" ~ "50_59",
      Age_group == "IncidenceRateAge60-69" ~ "60_69",
      Age_group == "IncidenceRateAge70-79" ~ "70_79",
      Age_group == "IncidenceRateAge80AndOver" ~ "Over_80",
      TRUE ~ Age_group   # fallback so knitting doesn't choke
      ))
```

The next step of the data analysis is to construct a linear model predicting incidence rate by age group, sex, and year. In this model the outcome or response variable is the incidence rate, and the predictor or explanatory variables are age group, sex, and year.

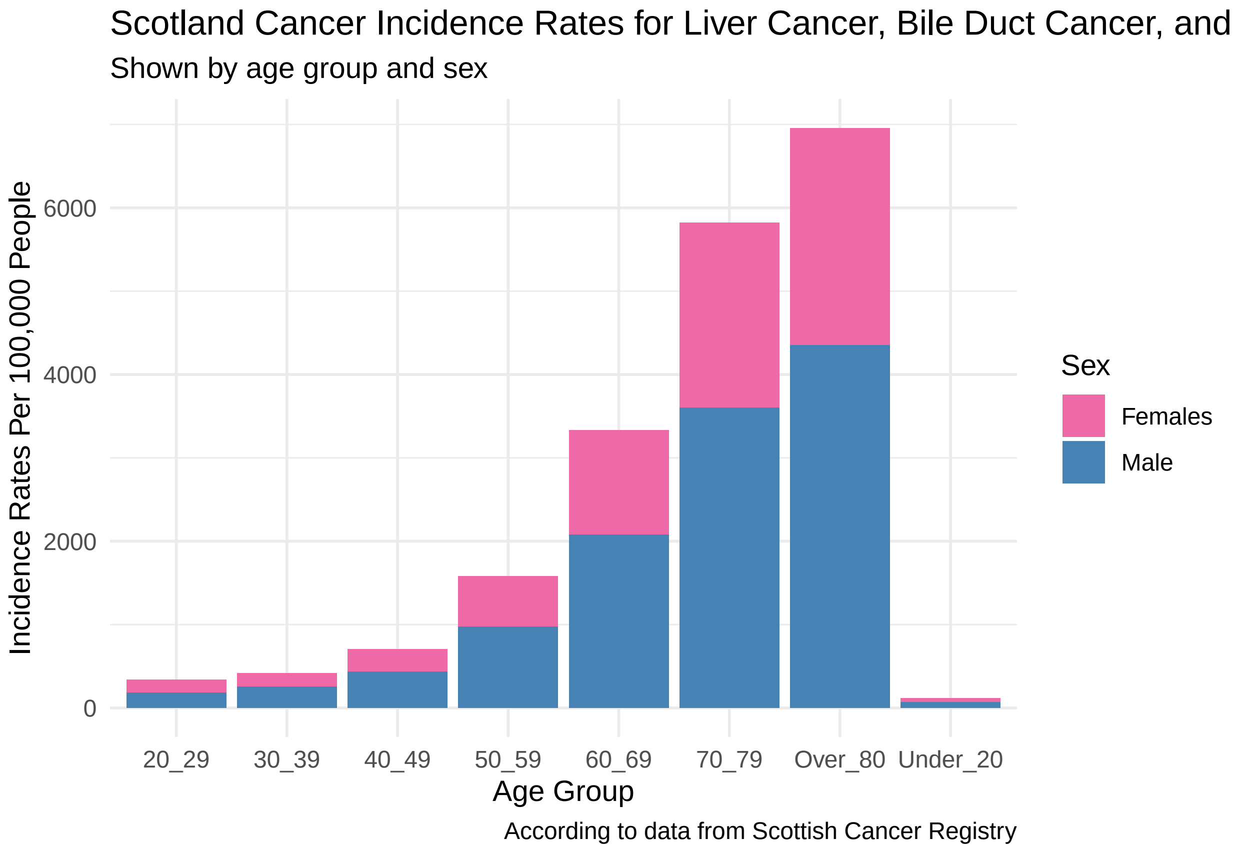
```
model <- lm(Incidence_Rate ~ Age_group + Sex + Year, data = RestructuredData)
summary(model)
```

```
## 
## Call:
## lm(formula = Incidence_Rate ~ Age_group + Sex + Year, data = RestructuredData)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -47.227  -5.396  -0.133   5.124  78.829 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -358.42918  115.66442  -3.099  0.00197 ** 
## Age_group30_39       0.51929    1.65890   0.313  0.75429    
## Age_group40_49       2.41247    1.65890   1.454  0.14605    
## Age_group50_59       8.26001    1.65890   4.979 7.00e-07 ***
## Age_group60_69      19.87951    1.65890  11.984  < 2e-16 ***
## Age_group70_79      36.20775    1.65890  21.826  < 2e-16 ***
## Age_groupOver_80    42.91070    1.65890  25.867  < 2e-16 ***
## Age_groupUnder_20   -1.47367    1.65890  -0.888  0.37448    
## SexFemales          -3.23595    1.01587  -3.185  0.00147 ** 
## SexMale              4.47201    1.01587   4.402 1.14e-05 ***
## Year                 0.17916    0.05751   3.115  0.00187 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 17.6 on 1789 degrees of freedom
## Multiple R-squared:  0.4764, Adjusted R-squared:  0.4735 
## F-statistic: 162.8 on 10 and 1789 DF,  p-value: < 2.2e-16
```

The output from the model tells us that the model explains about 47% of the variation in incidence rates. This is interesting as it is higher than I expected because although age is a big factor in cancer incidence, there was no data on the presence or lack thereof of risk factors. This goes to show that even disregarding risk factors, increasing age is a huge factor in the incidence of cancer. This observation is further supported by the analysis by age group. When compared to those in their 20s, those in their 50s, 60s, and 70's had rates of about 8, 20, 36, and 43 more cases per 100,000 people, at quite significant levels. Also notably the incidence rates for females were 3 cases per 100,000 people lower than the combined population average, while rates for males were 4 cases per 100,000 higher. There was also a small but significant increase in incidence rates of cancer year over year.

The next step for data analysis is to create a visualization to help communicate this.

```r
RestructuredData %>%
  filter(Sex %in% c("Male", "Females")) %>%
  ggplot(aes(x = Age_group, y = Incidence_Rate, fill = Sex)) +
  geom_col() +
  scale_fill_manual(values = c("Male" = "steelblue", "Females" = "hotpink2")) +
  theme_minimal() +
  labs(title = "Scotland Cancer Incidence Rates for Liver Cancer, Bile Duct Cancer, and Lymphoma", subt
```


Scotland Cancer Incidence Rates for Liver Cancer, Bile Duct Cancer, and
Shown by age group and sex
Incidence Rates Per 100,000 People
20_29
30_39
40_49
50_59
60_69
70_79
Over_80
Under_20
Age Group
Sex
Females
Male
According to data from Scottish Cancer Registry