Cancer - Admission

52103889

01/11/2021

Background

Mitosis is a fundamental process for life closely controlled by genes, but when body cell starts multiplying without control, cancer cells are created in the body. In today's world, technological advancement has led us to fathom unimaginable depths of science such as exploring furthest parts of the Universe, sequencing the human genome and eradicating smallpox to name a few. Still, after billions of dollars of research by some of the most amazing minds on Earth, we are still unable to achieve success over understanding and eradicating Cancer, a disease that effects more than 14 million people in the world and their near and dear ones at any given time. It is still considered as a dreaded word, mere mention of which brings despair and pain to those affected.

The aim of this exploratory data analysis is to visualize the trend of cancer hospital registration of Grampian Health Board as compared with rest of health boards of Scotland from 2003 to 2018.Also comparing Grampian with Highland Health board to see difference in its statistics from 2003 to 2018.

The cancer registration data has been acquired through "Public health information for Scotland". It is an open data source which can be used for various exploratory analysis. The Analysis of cancer registration over the period of almost two decades, will give us a glimpse of the trend and also emphasize on its criticality. This will help the healthcare system and various charities such as Cancer Research UK and Macmillan Cancer Support of Scotland in channeling their resources towards a common goal of improving the healthcare system in the worst affected health boards where significant surge is noted. This will also help such health boards to spread awareness on the lifestyle necessary for avoiding cancer and provide the necessary support to the common public which can play an important role in moderating the number of cancer registrations and help those diagnosed with cancer to lead a more normal life. For the purpose of this report, cancer registrations for all the available geographies within Scotland was taken from www.ScotPHO.com and the Grampian region health board was chosen for analyzing and showcasing the trend in cancer registration compared to rest of the health boards. Further, a comparison was carried out on the trend between the Grampian and Highland health boards to showcase how it has changed over a period of 15 years from 2003 to 2018

Data Acquisition

Public health information for Scotland profiles presents a range of indicators to give an overview of health and its wider determinants at a local level. The profiles give a snapshot of health for local areas and their variation with the help of varied visualisation. To explore the trend of cancer registration for all the health boards of Scotland, the data set has been taken from the "data" tab of **Public health information for Scotland** for all the available geographies and selecting the "Cancer Registrations" as the indicator. Geography codes and label provides data about geographical codes and its corresponding labels.

Load the package

For this exploratory data analysis object-oriented programming R has been used which is open and is widely used for data analysis. Few common libraries have been used which are as follows:

tidyverse – it is a collection of R packages designed for data analysis. It makes data analysis easier faster and more fun

viridis – This package provides a series of colour maps that are designed to improve graph readability for readers

patchwork – This package expands the API to allow for arbitrarily complex composition of plots

plotly – This package is used for creating interactive and quality graphs

```
library(tidyverse)
library(viridis)   # nice colour scheme
library(patchwork) # great for combining plots
library(plotly)    # interactive visualisations
```

Read in the data

read_csv() function has been used to read the cancer registration data in comma-separated values into data frame "cancer_data" and Geography data has been read into "HSC_data_zone"

```
## Cancer registration info

cancer_data <- read_csv("scotpho_data_extract - Cancer.csv", col_types = cols())

## Intermediate_data_zone info

HSC_data_zone <- read_csv("iz2011_codes_and_labels_21042020.csv", col_types = cols())
```

Prepare the data

"scotpho_data_extract - Cancer.csv" file provides information of 1312 rows of data for cancer registrations. The rows provide information about the mean cancer registration for a given year under the column "measure" with its confidence interval under "Lower confidence interval" and "Upper confidence interval" which has been standardized for age - sex rate per 100,000. Public Health Scotland is responsible the collection of Scottish cancer registry called "SMR06". Information is collected on every new diagnosis of cancer occurring in scottish residents . These information are categorized under different area type under " area_type" and its corresponding codes under "codes" and area name under "area_name. The indicator column simply reminds that below information is about cancer registrations.

```
glimpse(cancer_data)
```

```
## Rows: 3,312
## Columns: 12
## $ indicator                      <chr> "Cancer registrations", "Cancer registration~
## $ area_name                      <chr> "Scotland", "NHS Ayrshire & Arran", "NHS Bor~
## $ area_code                      <chr> "S00000001", "S08000015", "S08000016", "S080~
## $ area_type                      <chr> "Scotland", "Health board", "Health board", ~
## $ year                           <dbl> 2003, 2003, 2003, 2003, 2003, 2003, 2003, 20~
## $ period                    <chr> "2002 to 2004 calendar years; 3-year aggrega~
## $ numerator                 <dbl> 27247.00, 2123.00, 650.00, 916.00, 1503.00, ~
## $ measure                   <dbl> 643.95, 649.79, 614.58, 620.53, 651.27, 617.~
## $ lower_confidence_interval <dbl> 635.97, 621.25, 566.64, 579.71, 617.17, 593.~
## $ upper_confidence_interval <dbl> 652.01, 679.26, 665.39, 663.40, 686.70, 642.~
## $ definition                <chr> "Age-sex standardised rate per 100,000", "Ag~
## $ data_source               <chr> "Public Health Scotland (SMR06)", "Public He~
```

For this exploratory analysis information of cancer registration is needed for every “HSC partnership” hence all the HSC partnership has been filtered from area type. For the this analysis, visualization will be done under different health board of Scotland. Health Board information is provided under “area_name” hence only “area name, area code, measure and its corresponding year has been selected with the help of filter function on cancer data. After filtering 496 row of information is produced. The glimpse()has been used to have a glimpse of the filtered data frame

```r
iz_data <- cancer_data %>%
  filter(area_type == "HSC partnership") %>%            # Filtering "HSC Partnership"
  select(area_name,area_code,year,measure)

glimpse(iz_data)
```

```
## Rows: 496
## Columns: 4
## $ area_name <chr> "Aberdeen City", "Aberdeenshire", "Angus", "Argyll & Bute", ~
## $ area_code <chr> "S37000001", "S37000002", "S37000003", "S37000004", "S370000~
## $ year      <dbl> 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, ~
## $ measure   <dbl> 657.36, 587.40, 574.47, 594.66, 625.07, 620.53, 656.66, 654.~
```

“iz2011_codes_and_labels_21042020.csv” file provides 1279 rows of geographical area codes for different area type like “intzone”,“CA”,“HSC partnership” and “Health Board” names. The glimpse()has been used to have a glimpse of the filtered data frame

```r
glimpse(HSC_data_zone)
```

```
## Rows: 1,279
## Columns: 9
## $ IntZone     <chr> "S02001236", "S02001237", "S02001238", "S02001239", "S0200~
## $ IntZoneName <chr> "Culter", "Cults, Bieldside and Milltimber West", "Cults, ~
## $ CA          <chr> "S12000033", "S12000033", "S12000033", "S12000033", "S1200~
## $ CAName      <chr> "Aberdeen City", "Aberdeen City", "Aberdeen City", "Aberde~
## $ HSCP        <chr> "S37000001", "S37000001", "S37000001", "S37000001", "S3700~
## $ HSCPName    <chr> "Aberdeen City", "Aberdeen City", "Aberdeen City", "Aberde~
## $ HB          <chr> "S08000020", "S08000020", "S08000020", "S08000020", "S0800~
## $ HBName      <chr> "NHS Grampian", "NHS Grampian", "NHS Grampian", "NHS Gramp~
## $ Country     <chr> "S92000003", "S92000003", "S92000003", "S92000003", "S9200~
```

Since our data is filtered for all HSC partnership and visualization needs to be done for every health board, HSCP code and Health Board name has been filtered from the geography data through filter function

```r
iz_info <- HSC_data_zone %>%
  select(HSCP,HBName)

glimpse(iz_info)
```

```
## Rows: 1,279
## Columns: 2
## $ HSCP   <chr> "S37000001", "S37000001", "S37000001", "S37000001", "S37000001"~
## $ HBName <chr> "NHS Grampian", "NHS Grampian", "NHS Grampian", "NHS Grampian",~
```

Data Cleaning

Checking any missing values

The acquired data is a huge data set. Finding errors by naked human eye is inconceivable. Therefore, before starting any type of analysis, it is critical to ensure that the data is clean and is ready for analysis. “Table ()” function of R has been used on area code (area_code) in the cancer registration data set (iz_data) and health board name (HBName) within the geography data (iz_info). It is a quick check to identify any missing values as the function provides frequency of all values of column “area_code” and “ HBName” and therefore any missing value and its frequency would appear in the output. The “summary ()” function has been used to check any outliers in the data set as this function provide minimum and the maximum value of numerical data in the data frame. The minimum and maximum value of the mean values are 494.9 and 780.3 respectively and the data has been taken between years 2003 and 2018. Both the data set which will be used for exploratory analysis has neither any missing data nor any outlier. The data set is in standard form as well, hence it can be used for Data analysis.

```
iz_data$area_code %>%table()               # any missing value in area code
```

```
## .
## S37000001 S37000002 S37000003 S37000004 S37000005 S37000006 S37000007 S37000008
##        16        16        16        16        16        16        16        16
## S37000009 S37000010 S37000011 S37000012 S37000013 S37000016 S37000017 S37000018
##        16        16        16        16        16        16        16        16
## S37000019 S37000020 S37000022 S37000024 S37000025 S37000026 S37000027 S37000028
##        16        16        16        16        16        16        16        16
## S37000029 S37000030 S37000031 S37000032 S37000033 S37000034 S37000035
##        16        16        16        16        16        16        16
```

```
iz_info$HBName %>% table()                 # any missing value in HB Name
```

```
## .
##        NHS Ayrshire and Arran                   NHS Borders
##                            93                            30
##     NHS Dumfries and Galloway                      NHS Fife
##                            40                           104
##              NHS Forth Valley                  NHS Grampian
##                            78                           132
## NHS Greater Glasgow and Clyde                  NHS Highland
##                           257                            79
##              NHS Lanarkshire                   NHS Lothian
##                           160                           192
##                    NHS Orkney                  NHS Shetland
##                             6                             7
##                   NHS Tayside             NHS Western Isles
##                            92                             9
```

```
iz_data %>% summary()
```

```
##   area_name          area_code               year         measure
##  Length:496         Length:496         Min.   :2003   Min.   :494.9
##  Class :character   Class :character   1st Qu.:2007   1st Qu.:611.2
##  Mode  :character   Mode  :character   Median :2010   Median :639.1
##                                        Mean   :2010   Mean   :639.5
##                                        3rd Qu.:2014   3rd Qu.:664.3
##                                        Max.   :2018   Max.   :780.3
```

Join the data set

The cancer registration data set (iz_data) that is being used in this analysis has area codes, area name, mean cancer registration of a year and that year. The geography data (iz_info) has area code under "HSCP" and its corresponding health board name. Combining the two data set will produce one data set that will have all the column from both the data sets. The Left_Join () has been used to join the two data sets. This function produces combined data frame (admission_data) which has all the values of cancer registration data set (iz_data) and the matching geography data set (iz_info). Matching has been done on "area code" and "HSCF". The column name has been changed from "cancer_admission" to "measure" and "HBName " to Health_Board" for clear indexing of the column and the prefix "NHS" has also been taken away for all the Health Board's name to make the variable name more presentable with the help of mutate ()function.

```
admission_data <- left_join(iz_data,iz_info,by = c("area_code" = "HSCP"))

glimpse(admission_data)
```

```
## Rows: 20,464
## Columns: 5
## $ area_name <chr> "Aberdeen City", "Aberdeen City", "Aberdeen City", "Aberdeen~
## $ area_code <chr> "S37000001", "S37000001", "S37000001", "S37000001", "S370000~
## $ year      <dbl> 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, ~
## $ measure   <dbl> 657.36, 657.36, 657.36, 657.36, 657.36, 657.36, 657.36, 657.~
## $ HBName    <chr> "NHS Grampian", "NHS Grampian", "NHS Grampian", "NHS Grampia~
```

```
admission_data <- admission_data %>%
  mutate(HBName = gsub("NHS ", "", HBName)) %>%
  rename(Health_Board = HBName,Cancer_admissions= measure)

glimpse(admission_data)
```

```
## Rows: 20,464
## Columns: 5
## $ area_name         <chr> "Aberdeen City", "Aberdeen City", "Aberdeen City", "~
## $ area_code         <chr> "S37000001", "S37000001", "S37000001", "S37000001", ~
## $ year              <dbl> 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003~
## $ Cancer_admissions <dbl> 657.36, 657.36, 657.36, 657.36, 657.36, 657.36, 657.~
## $ Health_Board      <chr> "Grampian", "Grampian", "Grampian", "Grampian", "Gra~
```

Save the data

The admission_data is now clean and has all the information that is required for this exploratory analysis.It is critical that this data set is stored in safe place and can be referenced anytime during data-analysis if required. This data set can be saved in CSV file using the write() function.

```
write_csv(admission_data, "Cancer related admission.csv")
```

Data Analysis

Comparison of the Grampian and Highland data

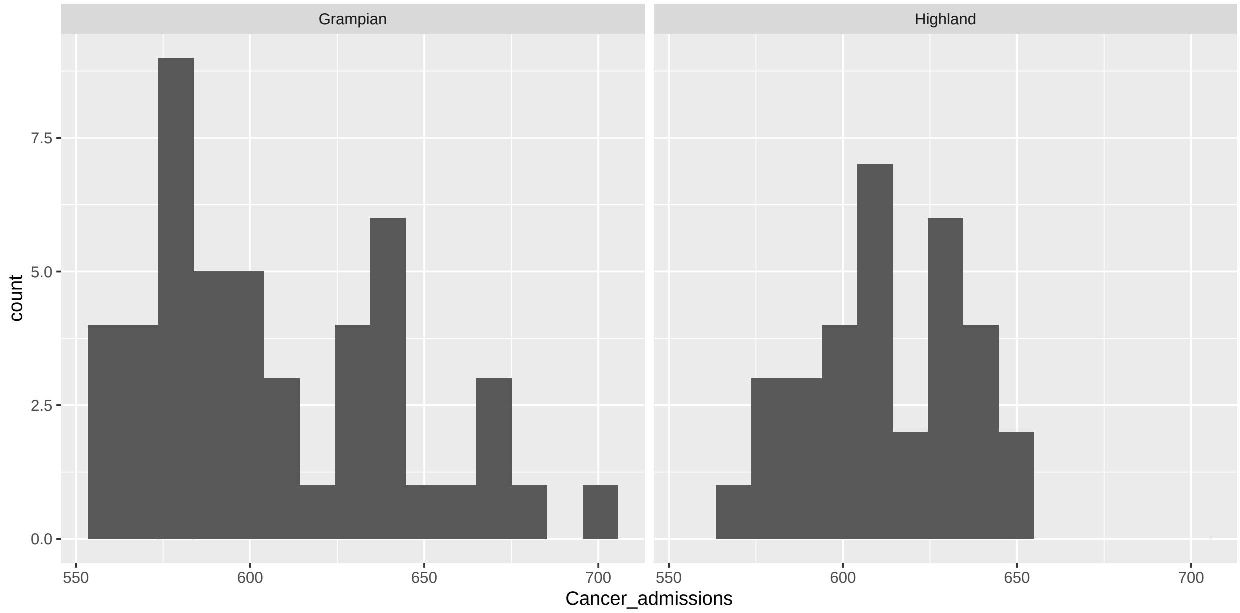
All the mean values of cancer registration for Grampians and Highland are filtered using filter function. ggplot() creates the x- axis and y axis for the plot, geom_histogram() creates the histogram and facet_wrap has been used to plot the histogram individually for Grampian and Highland. There were lots of repetitive data which has been corrected by using the unique () function. Histogram plot is showing the frequency distribution of mean cancer registration between the range of bins = 15.

The histogram for Grampian health board shows that the distribution of frequency of mean values is positively skewed and largest concentration of mean value is around 575 and few mean values are as high as 700. Similarly, the histogram of Highland health board shows that the distribution of frequency of mean values is not normally distributed either and highest concentration of mean values are around 600 which is slightly higher than the Grampians health board.There were lots of repetitive data which has been corrected by using the unique () function.

```
Grampian_Highland <- admission_data %>%
  filter(Health_Board %in% c("Grampian", "Highland"))%>%
  unique()


glimpse(Grampian_Highland)
```

```
## Rows: 80
## Columns: 5
## $ area_name         <chr> "Aberdeen City", "Aberdeenshire", "Argyll & Bute", "~
## $ area_code         <chr> "S37000001", "S37000002", "S37000004", "S37000016", ~
## $ year              <dbl> 2003, 2003, 2003, 2003, 2003, 2004, 2004, 2004, 2004~
## $ Cancer_admissions <dbl> 657.36, 587.40, 594.66, 606.50, 608.42, 638.40, 569.~
## $ Health_Board      <chr> "Grampian", "Grampian", "Highland", "Highland", "Gra~
```

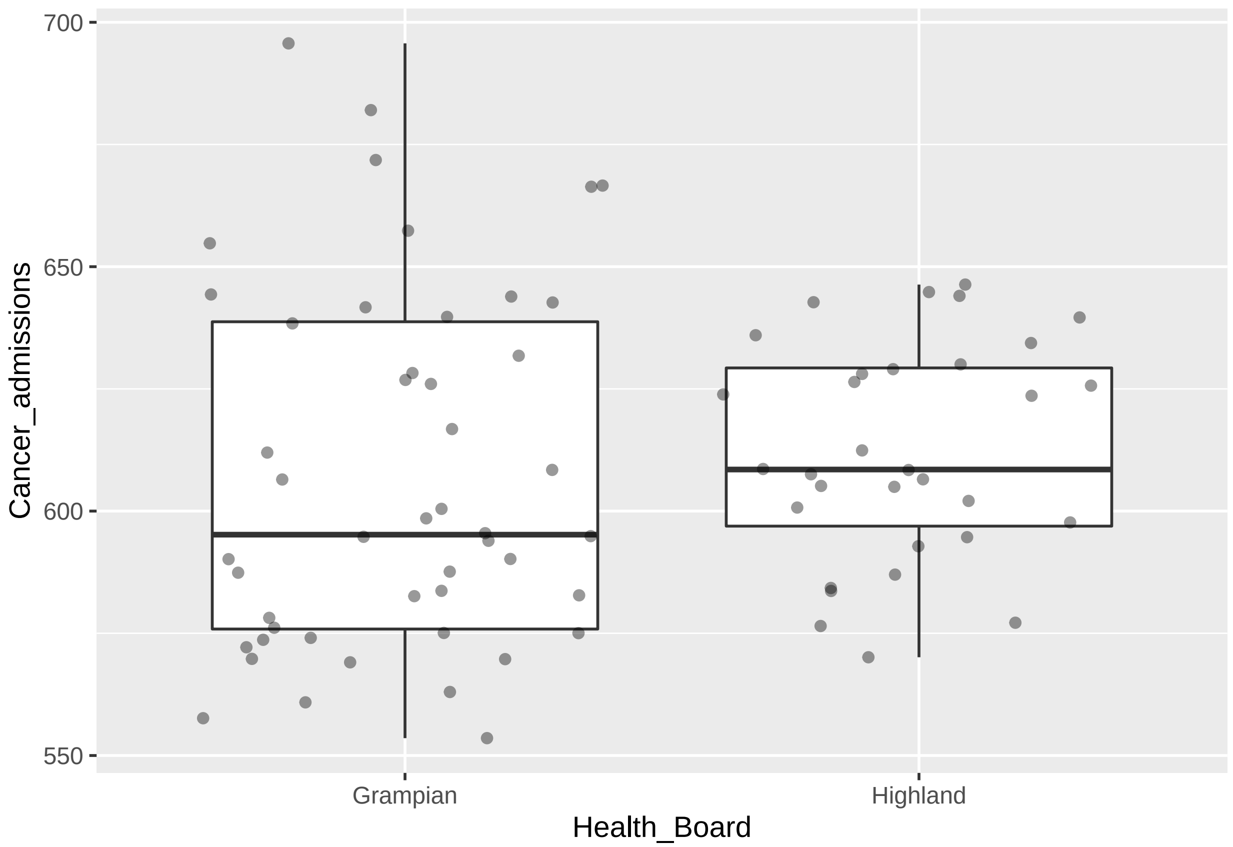
```
#histogram
Grampian_Highland %>%
  ggplot(aes(x = Cancer_admissions)) +
  geom_histogram(bins = 15)+
  facet_wrap(~Health_Board)
```



Box plot for Grampian and Highland

To compare the two health boards in an in-depth manner, a box plot has been plotted using ggplot() and geom_boxplot(). To visualize the individual median values, geom_jitter() function has been used. From the two box plot it is evident that the median of all the mean values for Grampian health board is just below 600 whereas the median of all the mean values for Highland is above 600. and the data is pretty much scattered for both the health boards. The spread of interquartile range for Grampians is little bit more than the spread of interquartile range of Highland Health board. The spread of minimum value and maximum value of mean values is bigger in Grampians as compared to Highland

```
# boxplot
Grampian_Highland %>%
  ggplot(aes(x = Health_Board,
             y = Cancer_admissions)) +
  geom_boxplot() +
  geom_jitter(alpha = 0.4) +              # add data points
  theme(legend.position = "none")         # remove legend
```



Communication

Median Grampian admissions 2003 to 2018

Since the data set for Grampians is not normally distributed, median would be appropriate to calculate the central tendency of the data. The median of all the mean values for a given year, filtered only for Grampian health board, has been calculated by Median () function within summarise () function.The data is then grouped by year using the function group_by (). The output is list of medians of mean values of cancer registrations only for Grampians health board for every year from 2003 to 2018.

Here we calculate the median admissions in Grampian for each year, using "group_by".

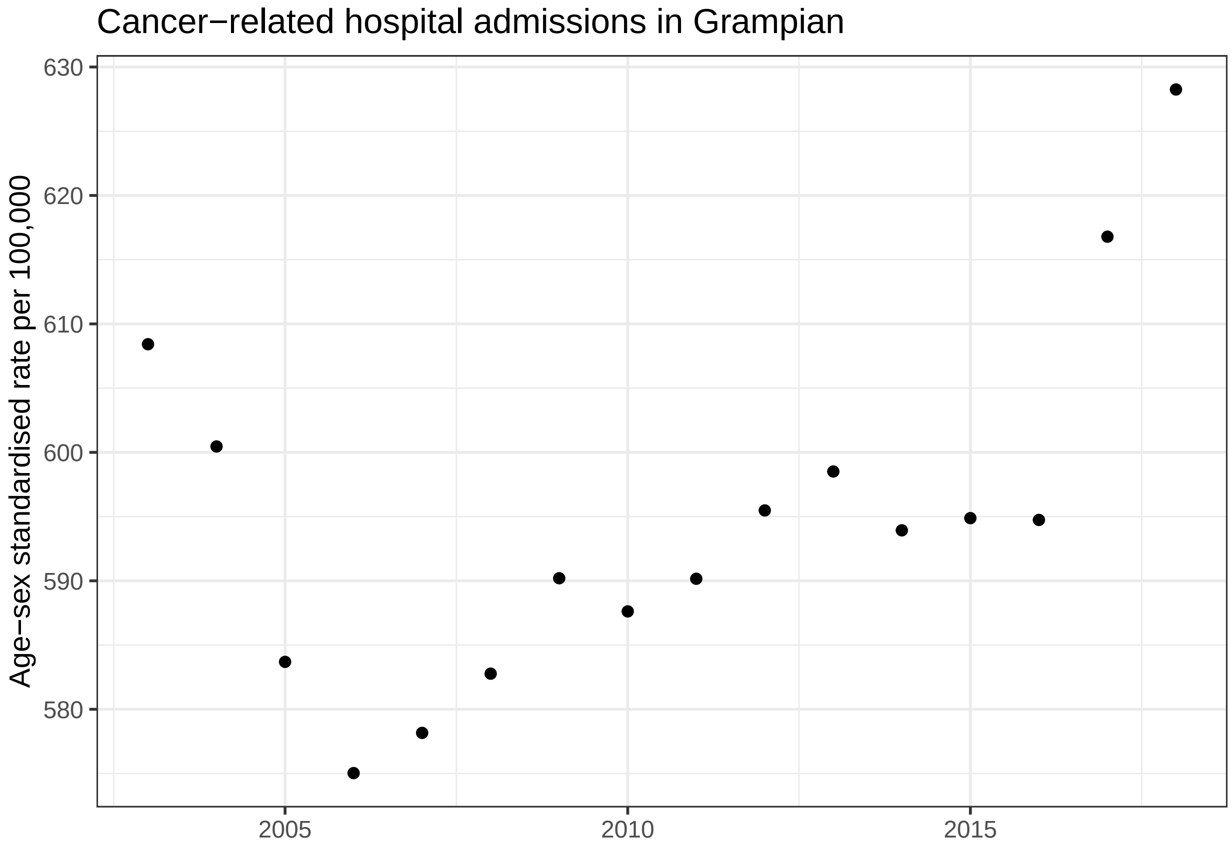
```
median_Grampian_admissions <- admission_data %>%
  filter(Health_Board == "Grampian") %>%
  group_by(year) %>%
  summarise(median_admissions = median(Cancer_admissions))

head(median_Grampian_admissions)
```

```
## # A tibble: 6 x 2
##    year median_admissions
##   <dbl>             <dbl>
## 1  2003              608.
## 2  2004              600.
## 3  2005               584.
## 4  2006               575.
## 5  2007               578.
## 6  2008               583.
```

Plotting a scatterplot using geom_point.

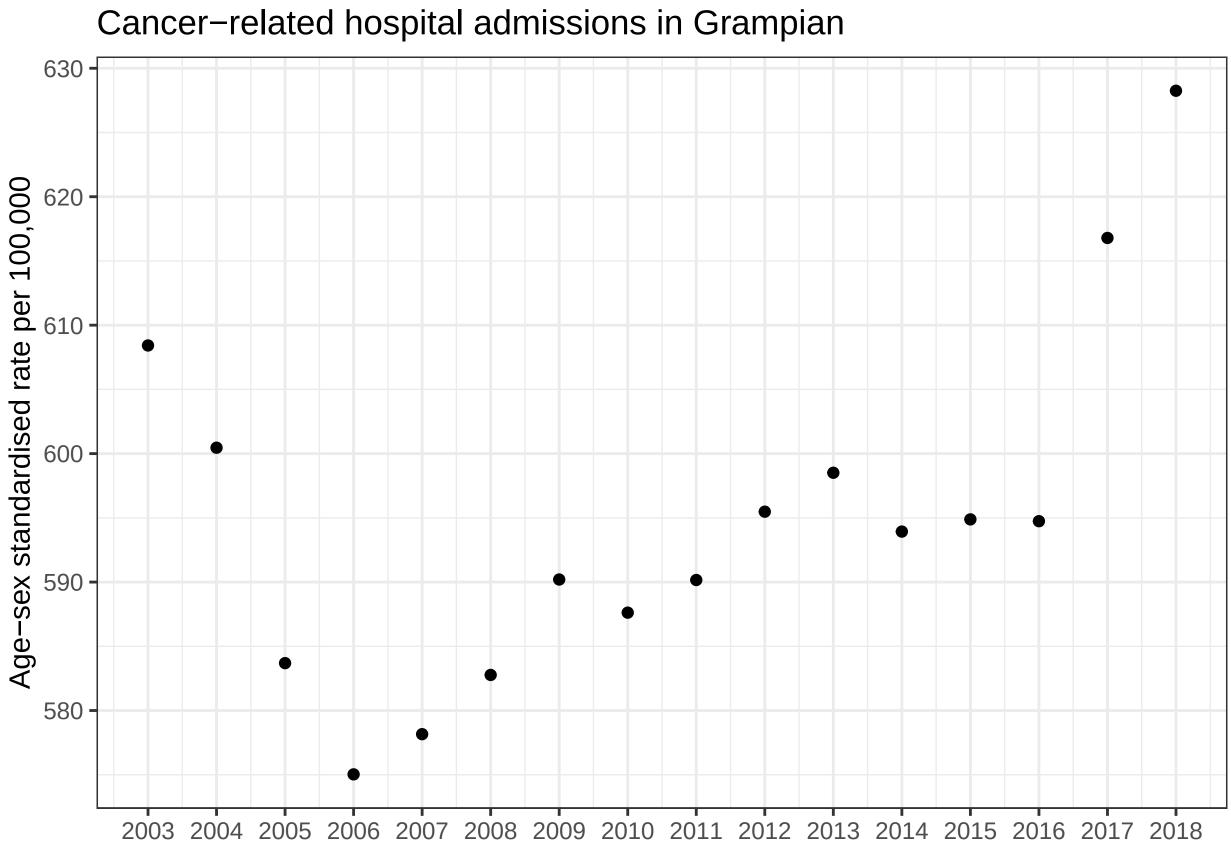
To see the trend in the median values of cancer admission against its respective year from 2003 to 2018 a scatter plot is a the best fit. Geom_point () has been used to plot a scatter plot where x axis is years and y axis is median cancer admission (median_admissions).

```
median_Grampian_admissions %>%
  ggplot(aes(x = year, y = median_admissions)) +
  geom_point()+
  xlab(NULL)+                                                # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +            # y-axis labeled
  ggtitle("Cancer-related hospital admissions in Grampian") +  # title added
  theme_bw()
```



Since the analysis needs to be done on year from 2003 to 2018,We can specify the ticks on the x axis using scale_x_continuous

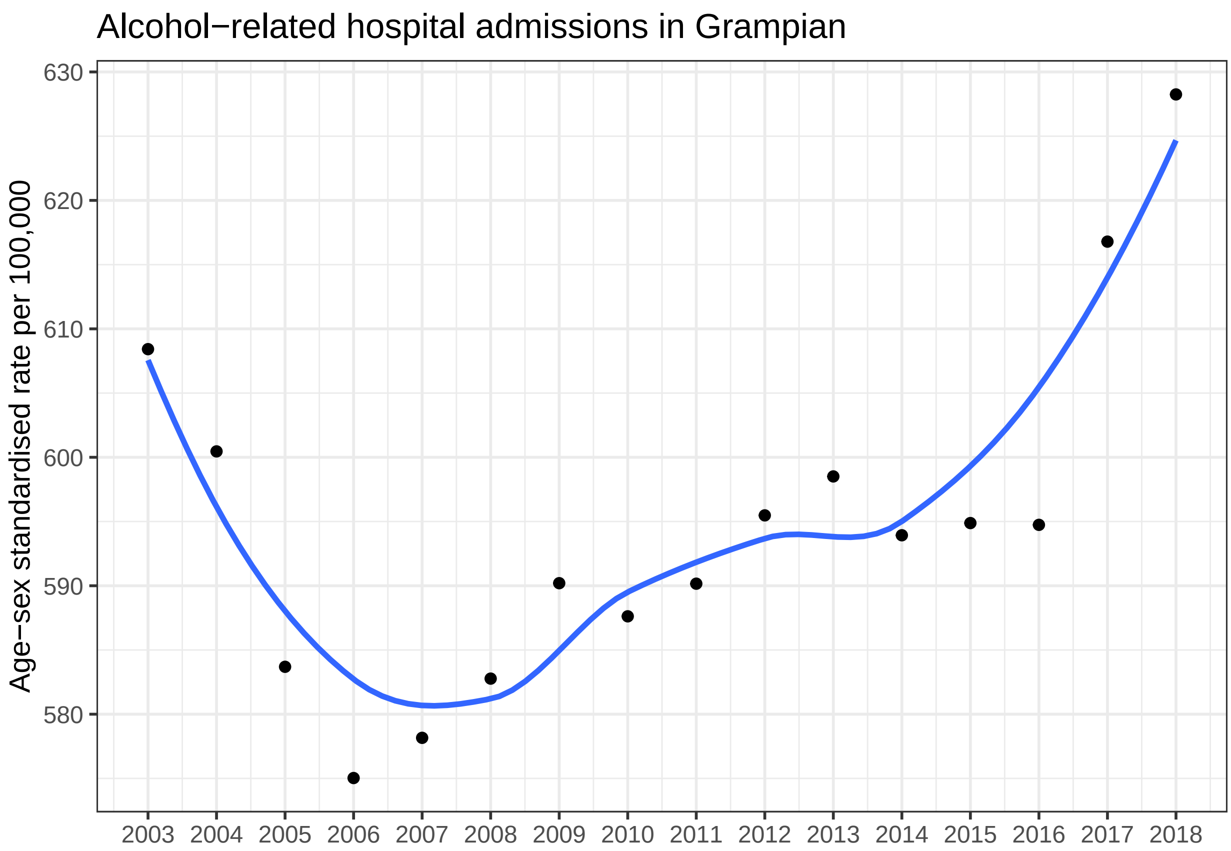
```
median_Grampian_admissions %>%
  ggplot(aes(x = year, y = median_admissions)) +
  geom_point()+
  scale_x_continuous(breaks = seq(2003, 2018, by=1) ) +
  xlab(NULL)+                                              # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +          # y-axis label
  ggtitle("Cancer-related hospital admissions in Grampian") +  # title added
  theme_bw()
```



The dots were joined to form a line to make the visualization even more clear by using geom_smooth () function.The function geom_smooth(), by default includes a confidence interval."se" within the function has been flagged FALSE to ignore the confidence interval.

```
median_Grampian_admissions %>%
  ggplot(aes(x = year, y = median_admissions)) +
  geom_point()+
  geom_smooth(se = FALSE) +
  scale_x_continuous(breaks = seq(2003, 2018, by=1) ) +
  xlab(NULL)+                                              # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +          # label y-axis
  ggtitle("Alcohol-related hospital admissions in Grampian") + # add title
  theme_bw()
```

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



R provides a fantastic graphical user interface called tooltip. When hovered over, it presents a brief informative message about that point. Tooltip has been applied by passing x axis and y axis values through tooltip in the ggplotly().

The scatter plot for Grampian conveys out few very interesting information about Grampian Health board.It shows the relationship between median cancer registration with its corresponding year and its strength. There is negative linear pattern between median cancer registration and its corresponding year between 2003 and 2006.Year 2006 has the minimum cancer registration median value between 2003 and 2018. Cancer registration has increased gradually between 2006 and 2018 then picked up exponentially from 2016 to 2018 showing a positive liner pattern. Tooltip provides exact value when hovered over these area. For example in 2006 the median cancer registration value was 575.03 and in 2018 the median cancer registration value was 628.25.

```r
p <- median_Grampian_admissions %>%
  ggplot(aes(x = year, y = median_admissions)) +
  geom_point()+
  geom_smooth(se = FALSE) +
  scale_x_continuous(breaks = seq(2003, 2018, by=1) ) +
  xlab(NULL)+                                              # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +          # label y-axis
  ggtitle("Cancer-related hospital admissions in Grampian") +  # add title
  theme_bw()

ggplotly(p,tooltip = c("x","y"))
```

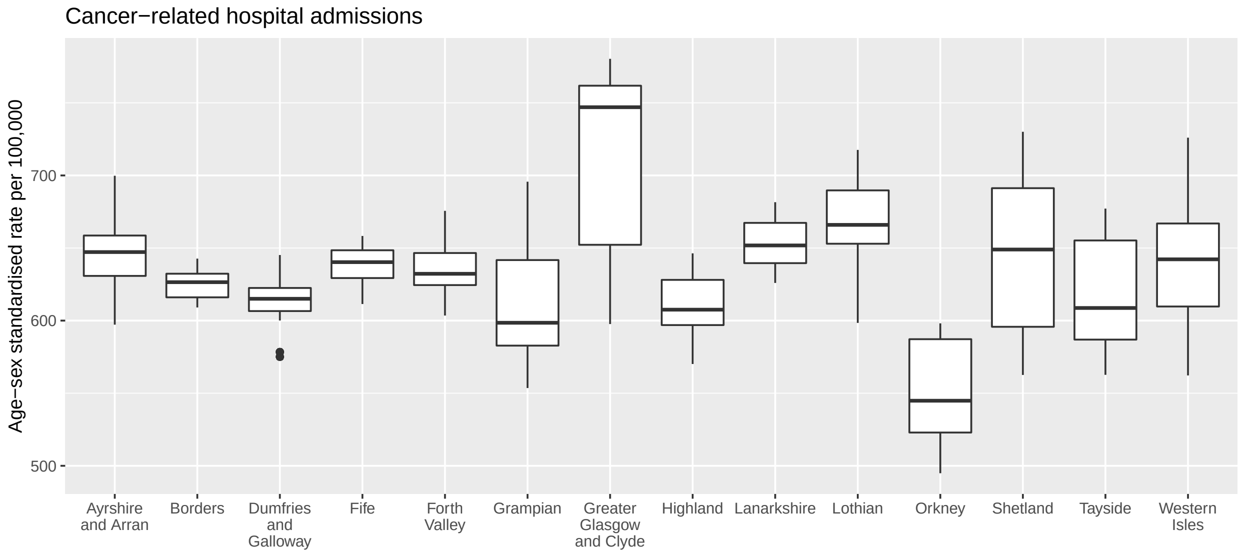
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

Boxplot for all the health boards from 2003 to 2018

A box plot for all the health board has been plotted taking the median cancer registration form 2003 to 2018 using the function geom_boxplot() and defining the width of the plot in the chunk header equals to 12. To make the labels at X-axis legible str_wrp() function has been used and values of X-axis have been passed through it.

Labels at X-axis and Y-axis have been customized in the plot. “Null” has been passed through xlab() which removes any label from X-axis as it is nor required and “ Age – sex standardised rate per 100,000” has been passed through Ylab() to prompt it on Y-axis of the plot. The title of the plot “has been passed through ggtitle() which is being reflected at the top of the plot

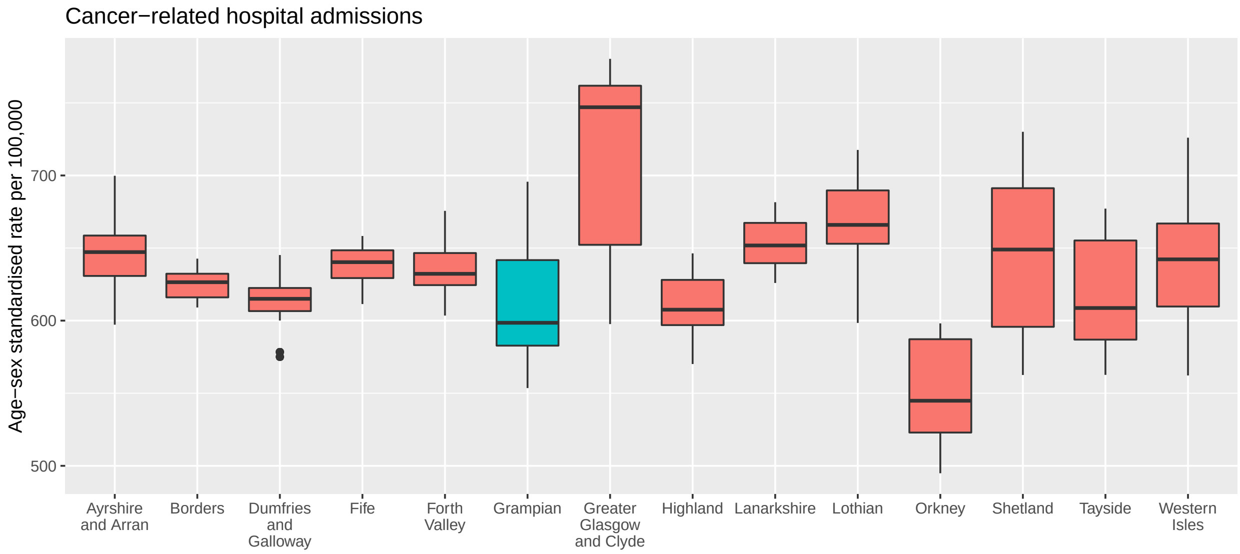
```
admission_data %>%
  ggplot(aes(x = str_wrap(Health_Board, 10),
             y = Cancer_admissions)) +
  geom_boxplot()+
  xlab(NULL) +                                           # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +        # label y-axis
  ggtitle("Cancer-related hospital admissions ")         # add title
```



Filling Colours in the plot

A new variable has been created called “isGrampian” with the help of mutate function. “cancer_admission” dataset has been bifurcated, where all rows that belongs to Grampian, the ifelse statement is “yes” for it and is stored in “isGrampian”, for rest of the data the ifelse statement is “no”. The new variable “IsGrampian” is then filled with different colour compared to rest of the health board to make it stand out and legend has been removed by theme() function.

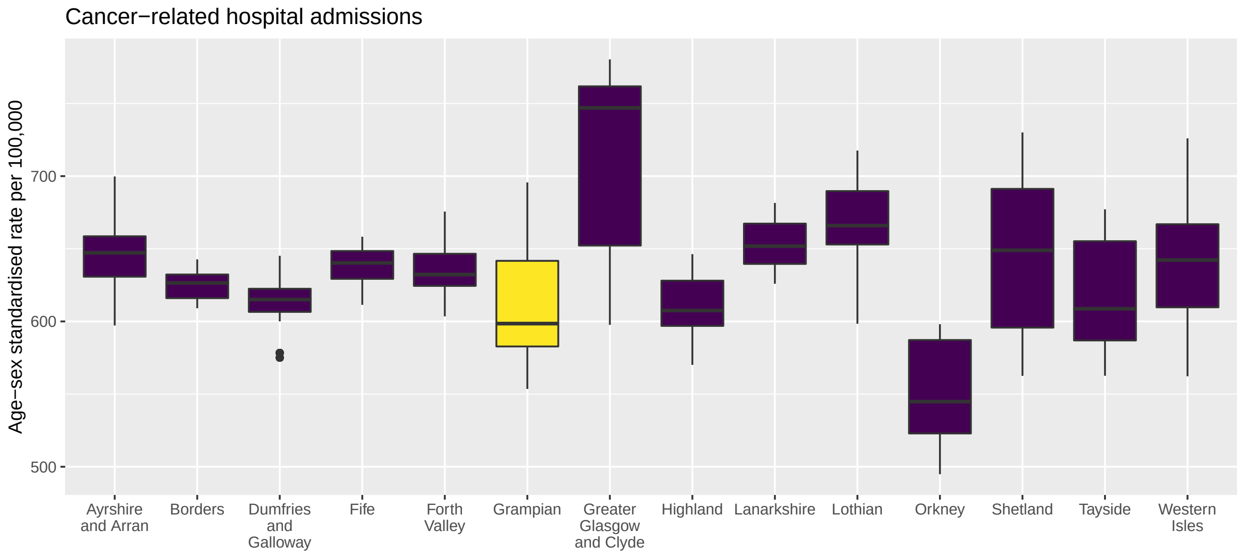
```
admission_data %>%
  mutate(isGrampian=ifelse(Health_Board=="Grampian","yes","no")) %>%
  ggplot(aes(x = str_wrap(Health_Board, 10),
           y = Cancer_admissions,
           fill=isGrampian)) +
  geom_boxplot() +
  xlab(NULL) +                                                # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +             # label y-axis
  ggtitle("Cancer-related hospital admissions ")+             # add title
  theme(legend.position = "none")
```



change the colour scheme to Viridis.

The fill function filled the boxplot with Red and Green. The most common form of colour blindness is known as “red/green colour blindness”. Colouring the plot with red and green might make this useless for the people with colour blindness. Viridis colour scheme has been used to colour the box plot which provide a wide perceptual range and rely more on blue - yellow contrast.

```
admission_data %>%
  mutate(isGrampian=ifelse(Health_Board=="Grampian","yes","no")) %>%
  ggplot(aes(x = str_wrap(Health_Board, 10),
             y = Cancer_admissions,
             fill=isGrampian)) +
  geom_boxplot() +
  xlab(NULL) +                                                  # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +               # label y-axis
  ggtitle("Cancer-related hospital admissions") +               # add title
  scale_fill_viridis(discrete=TRUE) +                           # use the viridis colour scheme
  theme(legend.position = "none")                               # remove the legend
```

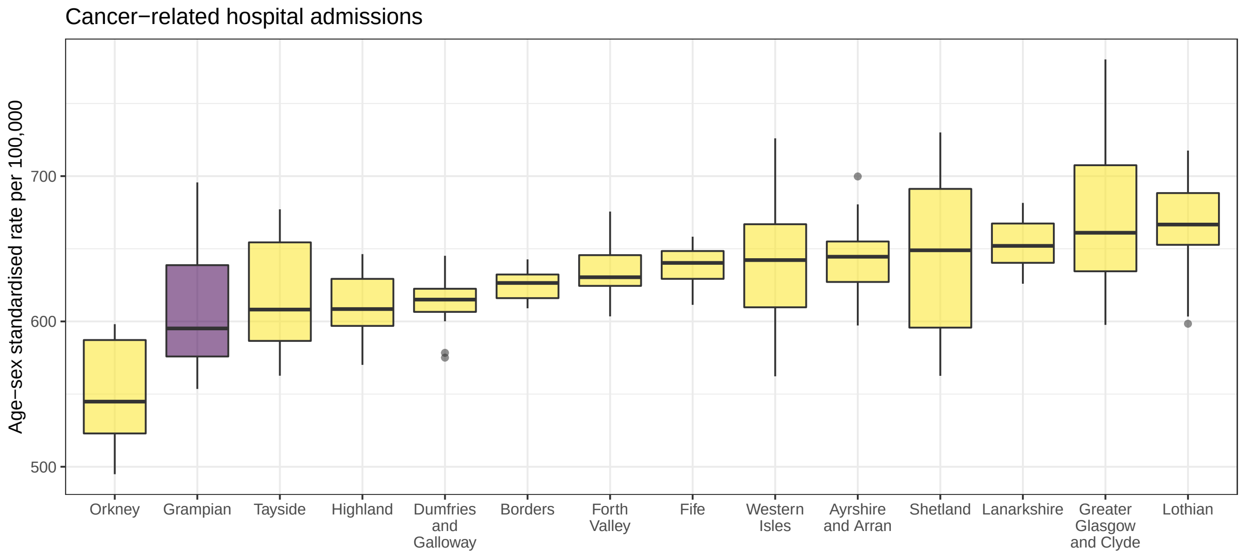


Graphs showing grampian

Colour plotted is a bit dark, making the median value illegible on the plot. To make it legible colours has been inverted by adding “direction = -1”, grey background has been removed by theme_bw() function and colours has been made transparent by passing “alpha = 0” through ggplot() function.

Below plot shows the median of the mean values of cancer registration of all the health boards from 2003 to 2018 ordered from minimum value of median to maximum value of median. “Orkeny” stands first in the row with a minimum value of median in the plot. Grampian stands on the second place whereas Lothian has the highest value of median in the given box plot. “Dumfries and Galloway”, “Ayrshire and Arran” and “Lothian” health boards have few outliers.

```
new_admission_data  <- admission_data %>%
unique()


new_admission_data %>%
  mutate(isGrampian=ifelse(Health_Board=="Grampian","yes","no")) %>%
  ggplot(aes(x = reorder(str_wrap(Health_Board,10),Cancer_admissions, FUN = median ),
             y = Cancer_admissions,
             fill=isGrampian,
             alpha=0)) +

  geom_boxplot()+
  xlab(NULL) +                                                # remove x-axis label
  ylab("Age-sex standardised rate per 100,000") +             # label y-axis
  ggtitle("Cancer-related hospital admissions ")+             # add title
  scale_fill_viridis(discrete=TRUE, direction = -1) +         # use the viridis colour scheme
  theme_bw() +
  theme(legend.position = "none")                             # remove the legend
```



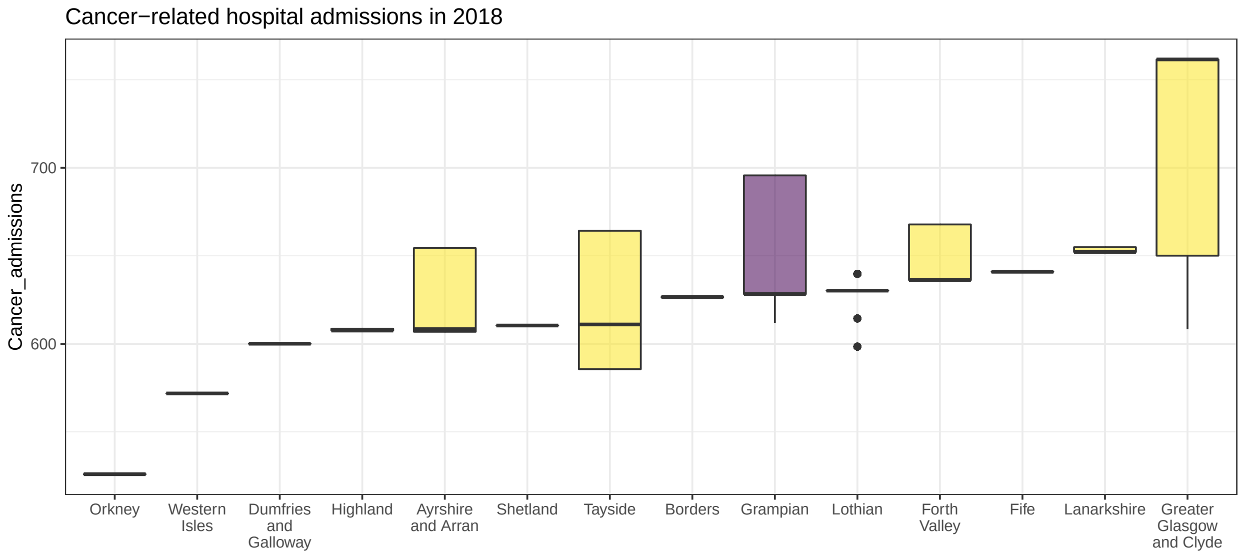
```
  geom_point(size = 1)
```

```
## geom_point: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

Comparing Grampian with other health board in 2018

To compare the cancer registration of Grampian with other health board in 2018, filter has been applied to the admission_data for 2018, which filters all the rows for the year 2018 only for all the health boards. The graph plotted shows very little variability for lots of health boards as the box plot is very slim and Lothian showing few outliers. For 2018 boxplot Grampian is in the top six health boards with high median values.

```
admission_data  %>%
filter(year == 2018) %>%
 mutate(isGrampian=ifelse(Health_Board=="Grampian","yes","no")) %>%
 ggplot(aes(x = reorder(str_wrap(Health_Board,10),Cancer_admissions, FUN = median ),
           y = Cancer_admissions,
           fill=isGrampian,
           alpha=0)) +

 geom_boxplot()+
 xlab(NULL) +                                             # remove x-axis label
 #ylab("Age-sex standardised rate per 100,000") +         # label y-axis
 ggtitle("Cancer-related hospital admissions in 2018")+   # add title
 scale_fill_viridis(discrete=TRUE, direction = -1) +      # use the viridis colour scheme
 theme_bw() +
 theme(legend.position = "none")                          # remove the legend
```



```
  geom_point(size = 1)
```

```
## geom_point: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

comparing Grampian in 2003,2010 and 2018

To compare Grampians for individual year in 2003, 2010 and 2018 with rest of the health boards gives an idea of trend of cancer registration of Grampian with rest of the Health boards. To see the trend clearly a box plot for all the health board for 2003, 2010 and 2018 has been plotted on top of each other by storing the code for each in three variable and positioning them as required. There is not much of change in the position of Grampians in 2003 and 2010.Infact the Grampians health board moved a position down in 2010 compared to 2003. 8 years down the lane the cancer registration for Grampians changed drastically and moved six positions up. While all the health boards moved up and down the ladder, two health board "Orkney" and "Greater Glasgow and Clyde" successfully managed to keep their position constant. Orkney managed to maintain the lowest value of median cancer registration and "Greater Glasgow and Clyde" maintained the highest value.

```
  P1 <- admission_data  %>%
 filter(year == 2003) %>%
  mutate(isGrampian=ifelse(Health_Board=="Grampian","yes","no")) %>%
  ggplot(aes(x = reorder(str_wrap(Health_Board,10),Cancer_admissions, FUN = median ),
            y = Cancer_admissions,
            fill=isGrampian,
            alpha=0)) +

  geom_boxplot()+
  xlab(NULL) +                                              # remove x-axis label
  #ylab("Age-sex standardised rate per 100,000") +           # label y-axis
  ggtitle("Cancer-related hospital admissions in 2003")+    # add title
  scale_fill_viridis(discrete=TRUE, direction = -1) +       # use the viridis colour scheme
  theme_bw() +
  theme(legend.position = "none")                            # remove the legend
  geom_point(size = 1)
```

```
## geom_point: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

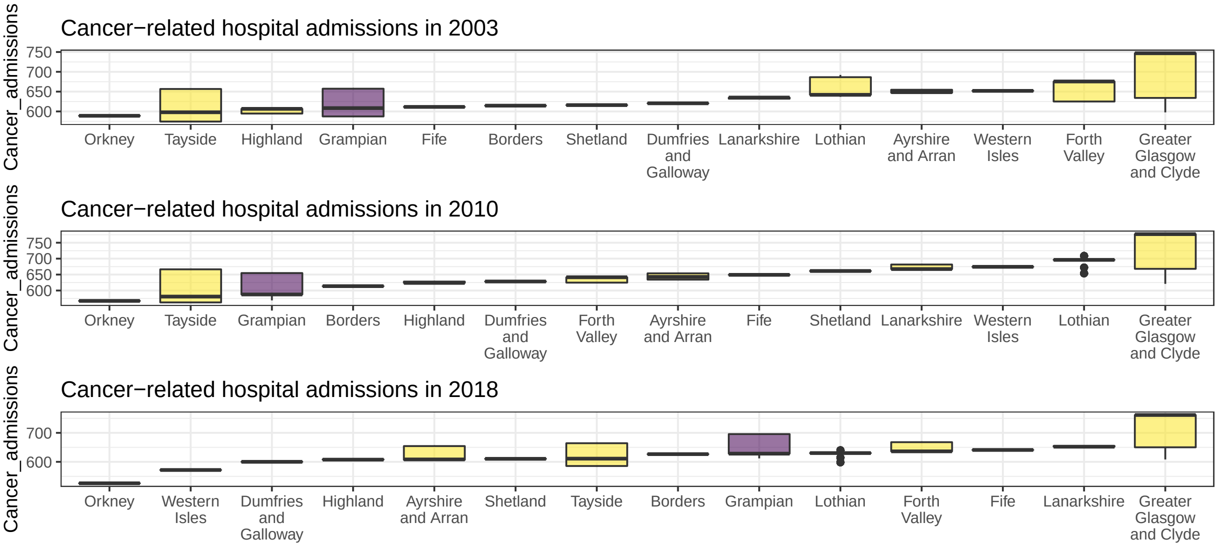
```
  P2 <- admission_data  %>%
 filter(year == 2010) %>%
  mutate(isGrampian=ifelse(Health_Board=="Grampian","yes","no")) %>%
  ggplot(aes(x = reorder(str_wrap(Health_Board,10),Cancer_admissions, FUN = median ),
             y = Cancer_admissions,
             fill=isGrampian,
             alpha=0)) +

  geom_boxplot()+
  xlab(NULL) +                                               # remove x-axis label
  #ylab("Age-sex standardised rate per 100,000") +            # label y-axis
  ggtitle("Cancer-related hospital admissions in 2010")+     # add title
  scale_fill_viridis(discrete=TRUE, direction = -1) +        # use the viridis colour scheme
  theme_bw() +
  theme(legend.position = "none")                            # remove the legend
  geom_point(size = 1)
```

```
## geom_point: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

```
  P3 <- admission_data  %>%
 filter(year == 2018) %>%
  mutate(isGrampian=ifelse(Health_Board=="Grampian","yes","no")) %>%
  ggplot(aes(x = reorder(str_wrap(Health_Board,10),Cancer_admissions, FUN = median ),
             y = Cancer_admissions,
             fill=isGrampian,
             alpha=0)) +

  geom_boxplot()+
  xlab(NULL) +                                               # remove x-axis label
  #ylab("Age-sex standardised rate per 100,000") +            # label y-axis
  ggtitle("Cancer-related hospital admissions in 2018")+     # add title
  scale_fill_viridis(discrete=TRUE, direction = -1) +        # use the viridis colour scheme
  theme_bw() +
  theme(legend.position = "none")                            # remove the legend
  geom_point(size = 1)
```

```
## geom_point: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

```
P1/P2/P3
```



Discussion

In this exploratory data analysis, the data used was from an open source. There were no missing values in the data set but there were very few data with lots of repetitions for single year for different Health boards, hence plotting a histogram for any health board for a particular year might not give a meaningful insight. While plotting box plots for all the health board for year 2018 it is evident that except few health boards the interquartile range was almost same as its median value. These few data had lots of repetitions, which was eluded by using the unique() function. Since there was lots of repetition of same data and very few data point, visualizing any health board in any particular year would not sketch correct interpretation. The choice of Grampian health board for detailed analysis and visualization is partly because it has considerable data within interquartile range which is missing in many other health boards like fife, Shetland, and Orkney. From the given data a scatter plot for these Health Board would not be very meaningful. Visualizing the trend of cancer registration for all the health board is very interesting and informative. Lots of information can be presented in a very simple way to wider audience in a very interactive way. Like the use of tooltip in scatter plot gives specific information of one date point when hovered over. The colour scheme provided by viridis package eliminates the colour blindness issue. The Data acquired is from 2003 to 2018 but it would be interesting to know how pandemic of 2020 has impacted the cancer registration in all the health board.