alcohol-related-admissions

D Blana

17/09/2021

Load packages

We just need the tidyverse package. (also the here package)

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.2.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(here)
```

```
## here() starts at /home/myke/Desktop/Intro2HDS_R_WEEK5_MAIN
```

Read in data

Our data is saved in csv files, so we will use the read_csv function.

```
# alcohol-related hospital admissions
data <- read_csv(here("INPUTS/WEEK7/scotpho_data_extract.csv"))
```

```
## Rows: 15160 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): indicator, area_name, area_code, area_type, period, definition, dat...
## dbl (5): year, numerator, measure, lower_confidence_interval, upper_confiden...
##
## 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.
```

```
# intermediate data zone info
intermediate_zone_codes <- read_csv(here("INPUTS/WEEK7/iz2011_codes_and_labels_21042020.csv"))
```

```
## Rows: 1279 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): IntZone, IntZoneName, CA, CAName, HSCP, HSCPName, HB, HBName, Country
##
## 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.
```

Prepare data

The data in these open datasets is already clean, so we don't need to worry about missing or wrong values. It is also tidy. We will only select the data we want: data from intermediate zones, and a subset of the variables from each dataset.

```
glimpse(data)
```

```
## Rows: 15,160
## Columns: 12
## $ indicator                <chr> "Alcohol-related hospital admissions", "Alco~
## $ area_name                <chr> "Scotland", "Culter", "Cults, Bieldside & Mi~
## $ area_code                <chr> "S00000001", "S02001236", "S02001237", "S020~
## $ area_type                <chr> "Scotland", "Intermediate zone", "Intermedia~
## $ year                     <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 20~
## $ period                   <chr> "2010/11 financial year", "2010/11 financial~
## $ numerator                <dbl> 38976, 24, 24, 24, 60, 15, 15, 9, 30, 45, 12~
## $ measure                  <dbl> 759.79, 475.18, 732.32, 372.46, 1306.45, 359~
## $ lower_confidence_interval <dbl> 752.20, 299.39, 452.81, 233.27, 985.89, 203.~
## $ upper_confidence_interval <dbl> 767.44, 715.42, 1114.28, 562.68, 1695.69, 58~
## $ definition               <chr> "Age-sex standardised rate per 100,000", "Ag~
## $ data_source              <chr> "Public Health Scotland (SMR01)", "Public He~
```

```
iz_data <- data %>%
  filter(area_type == "Intermediate zone") %>%
  select(area_name,area_code,year,measure)

glimpse(intermediate_zone_codes)
```

```
## 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~
```

```
iz_info <- intermediate_zone_codes %>%
  select(IntZone,HBName)
```

Join the datasets

Let's join the two datasets, so we know which health board each data zone belongs to. We will also remove "NHS" from the health board name, and rename a couple of the columns.

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

glimpse(admission_data)
```

```
## Rows: 12,790
## Columns: 5
## $ area_name <chr> "Culter", "Cults, Bieldside & Milltimber West", "Cults, Biel~
## $ area_code <chr> "S02001236", "S02001237", "S02001238", "S02001239", "S020012~
## $ year      <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, ~
## $ measure   <dbl> 475.18, 732.32, 372.46, 1306.45, 359.21, 416.62, 233.21, 601~
## $ HBName    <chr> "NHS Grampian", "NHS Grampian", "NHS Grampian", "NHS Grampia~
```

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

glimpse(admission_data)
```

```
## Rows: 12,790
## Columns: 5
## $ area_name          <chr> "Culter", "Cults, Bieldside & Milltimber West", "Cu~
## $ area_code          <chr> "S02001236", "S02001237", "S02001238", "S02001239",~
## $ year               <dbl> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 201~
## $ alcohol_admissions <dbl> 475.18, 732.32, 372.46, 1306.45, 359.21, 416.62, 23~
## $ health_board       <chr> "Grampian", "Grampian", "Grampian", "Grampian", "Gr~
```

Save dataframe

To save the dataframe to a CSV file, we use the write_csv function.

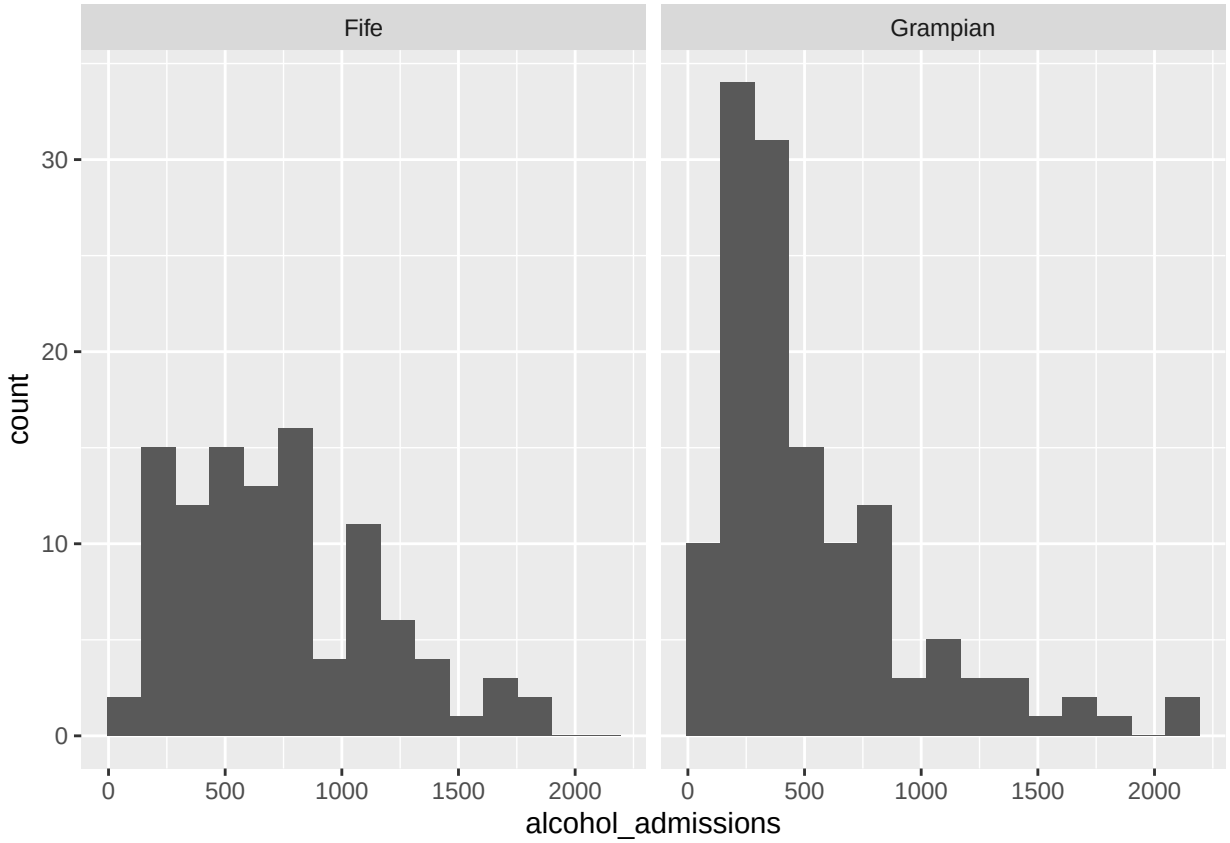
```
write_csv(admission_data, here("INPUTS/WEEK7/alcohol_related_admissions.csv"))
```

Let's compare Grampian and Fife data in 2019

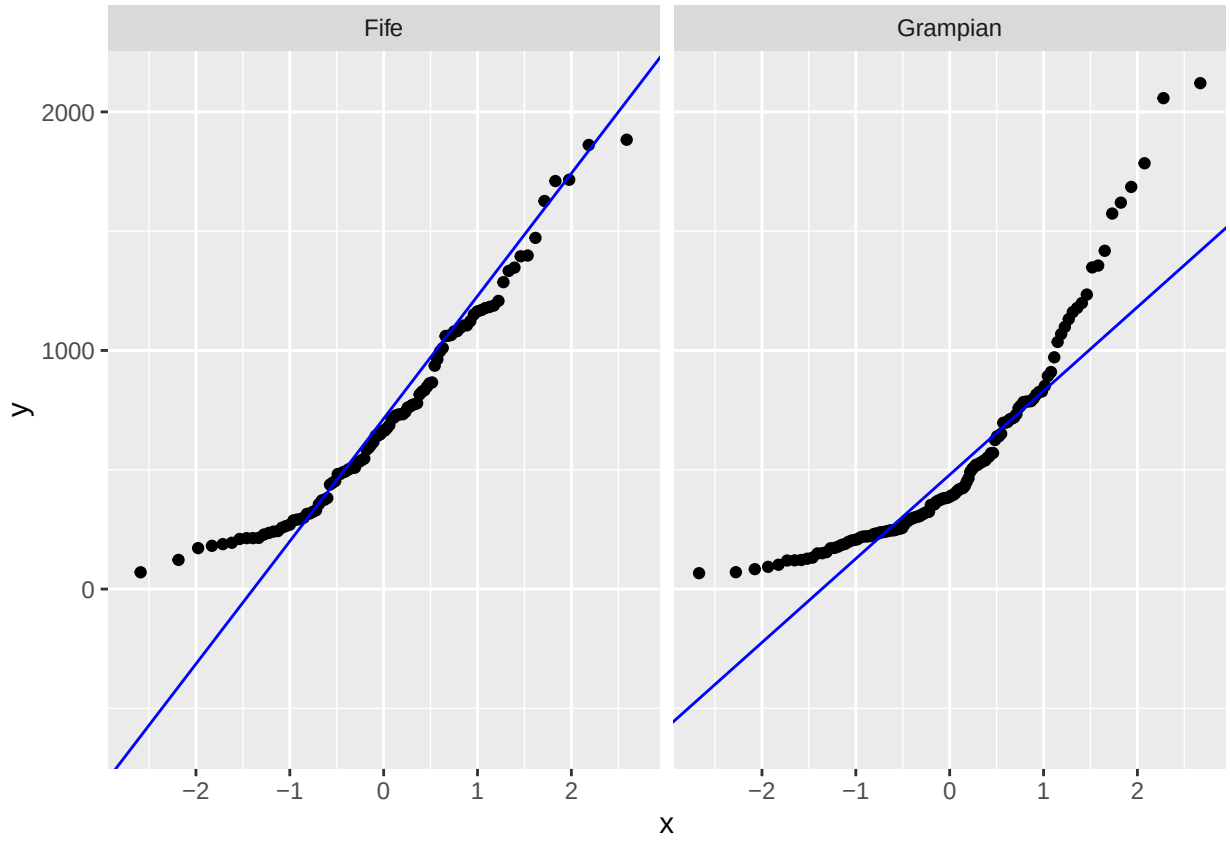
We will use histograms, Q-Q plots and boxplots, before doing a statistical comparison.

```
grampian_fife_data <- admission_data %>%
  filter(year == 2019,      # only 2019
         health_board %in% c("Grampian", "Fife"))  # only Grampian and Fife
```

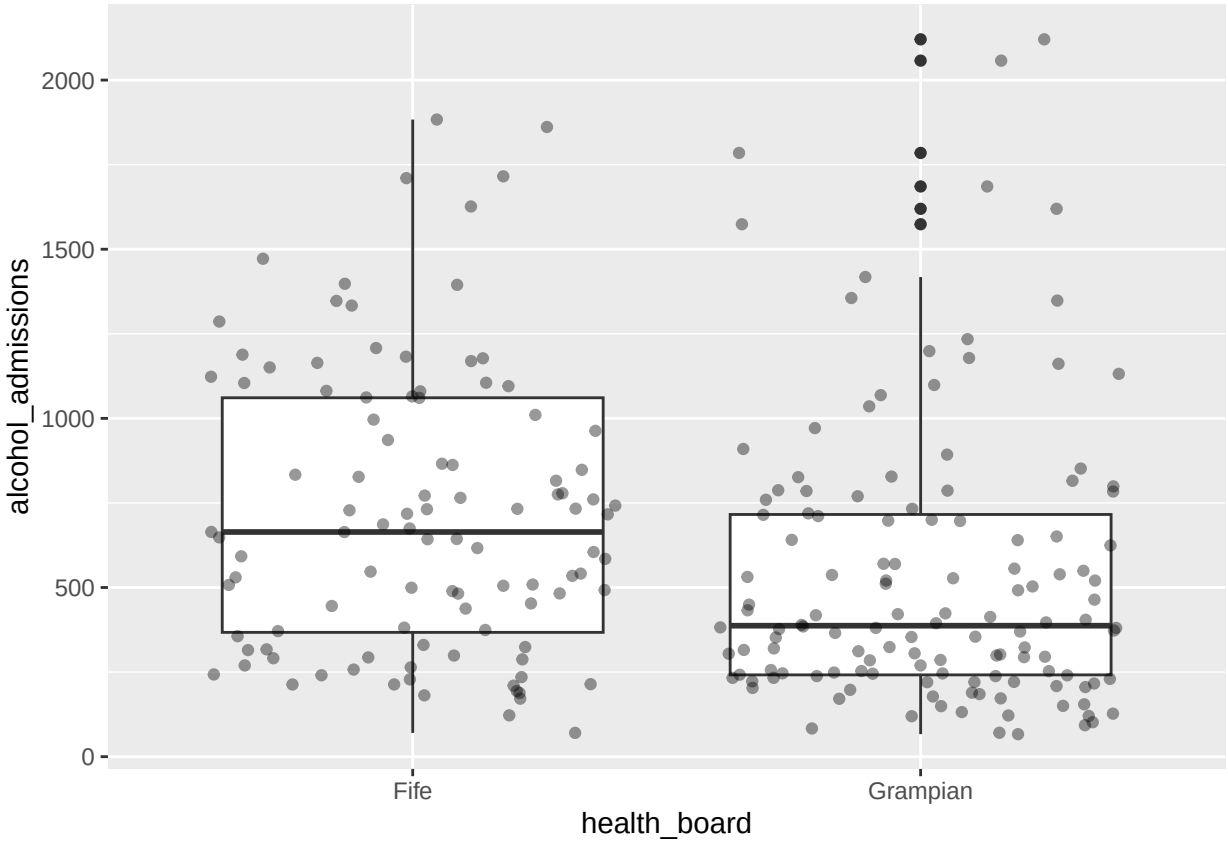
```r
#histogram
grampian_fife_data %>%
  ggplot(aes(x = alcohol_admissions)) +
  geom_histogram(bins = 15) +
  facet_wrap(~health_board)
```




```r
# Q-Q plot
grampian_fife_data %>%
  ggplot(aes(sample = alcohol_admissions)) +
  geom_qq() +
  geom_qq_line(colour = "blue") +
  facet_wrap(~health_board)
```




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



Statistical comparison

Non-parametric test

```
wilcox.test(alcohol_admissions ~ health_board, data = grampian_fife_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  alcohol_admissions by health_board
## W = 8871, p-value = 0.0001165
## alternative hypothesis: true location shift is not equal to 0
```

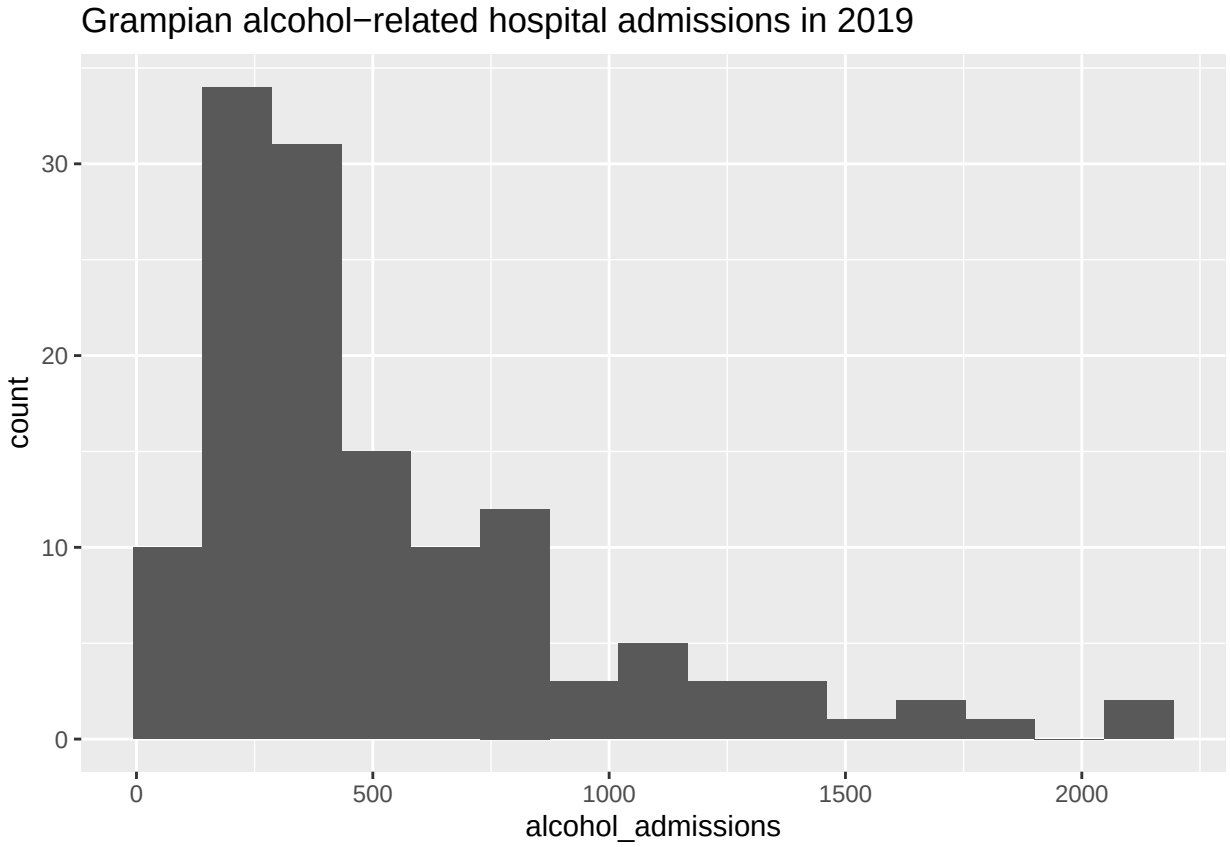
```
# non-parametric test for comparing more than two groups
admission_data %>%
  filter(year == 2019,
         health_board %in% c("Grampian", "Fife", "Lothian")) %>%
  kruskal.test(alcohol_admissions~health_board, data = .)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  alcohol_admissions by health_board
## Kruskal-Wallis chi-squared = 15.516, df = 2, p-value = 0.0004274
```

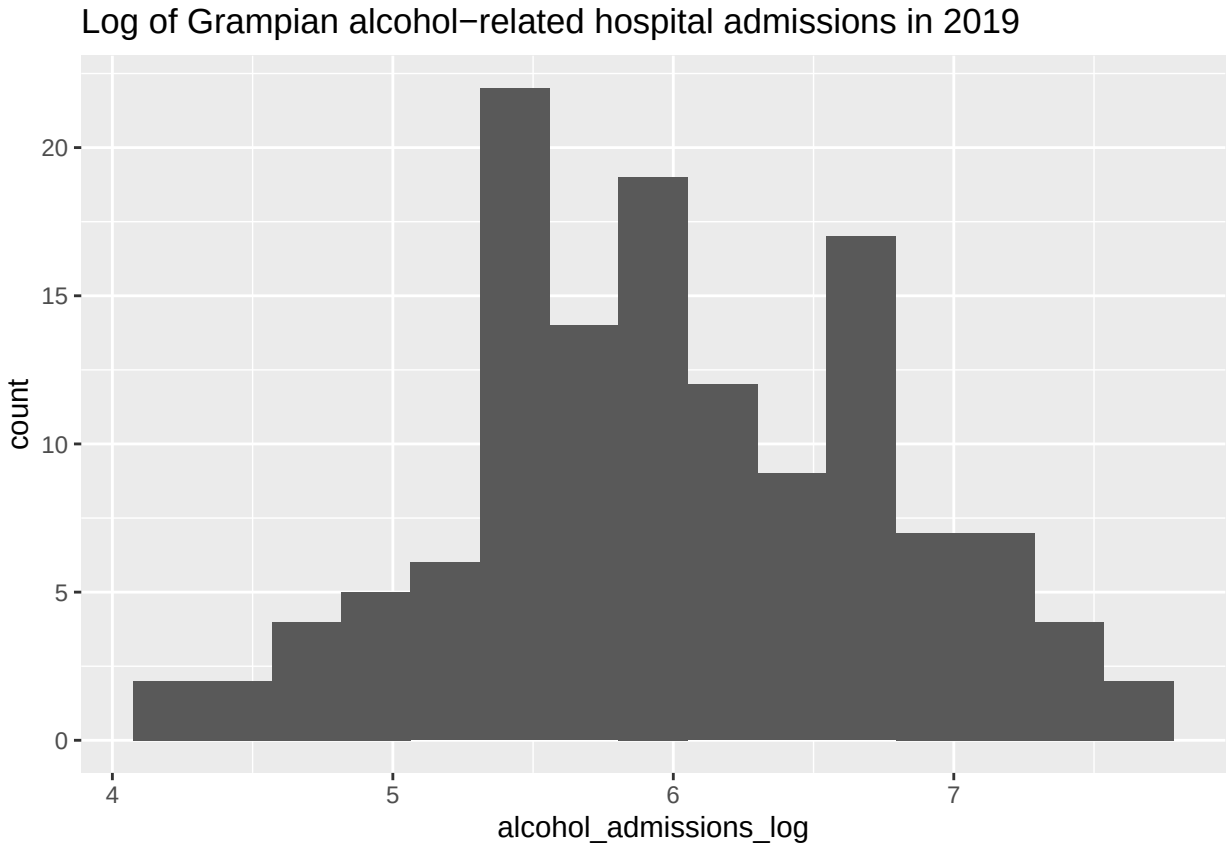
Transformation to normal

```
Grampian2019 <- admission_data %>%     # save as Grampian2019
  filter(year == 2019,             # only 2019
         health_board == "Grampian") %>%     # only Grampian
  mutate(alcohol_admissions_log = log(alcohol_admissions))         # log hospital admissions

Grampian2019 %>%
  ggplot(aes(x = alcohol_admissions)) +
  geom_histogram(bins = 15) +           # make histogram
  ggtitle("Grampian alcohol-related hospital admissions in 2019") # add title
```




```
Grampian2019 %>%
  ggplot(aes(x = alcohol_admissions_log)) +
  geom_histogram(bins = 15) +           # make histogram
  ggtitle("Log of Grampian alcohol-related hospital admissions in 2019") # add title
```


Log of Grampian alcohol−related hospital admissions in 2019
count
20
15
10
5
0
4
5
6
7
alcohol_admissions_log