EDA: Dr. Arbuthnot's Baptism Records

2022-07-19

loading the R Packages

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

Dataset 1: Dr. Arbuthnot's Baptism Records

The dataset is regarding the arbuthnot baptisum count of boys and girls from 1629 to 1710. The Arbuthnot data set refers to Dr. John Arbuthnot, an 18th century physician, writer, and mathematician. He was interested in the ratio of newborn boys to newborn girls, so he gathered the baptism records for children born in London for every year from 1629 to 1710.

```
data("arbuthnot")
head(arbuthnot)
```

```
##    year  boys girls
##   <int> <int> <int>
## 1  1629  5218  4683
## 2  1630  4858  4457
## 3  1631  4422  4102
## 4  1632  4994  4590
## 5  1633  5158  4839
## 6  1634  5035  4820
```

You can see the dimensions of this data frame by typing:

In R row comes first and column later, thus here its indicates there are 82 rows and 3 columns.

```
dim(arbuthnot)
```

```
## [1] 82  3
```

You can see the names of these columns (or variables) by typing:

```
names(arbuthnot)
```

```
## [1] "year"  "boys"  "girls"
```

What years are included in this dataset?

```
arbuthnot$year
```

```
##  [1] 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1640 1641 1642 1643
## [16] 1644 1645 1646 1647 1648 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658
## [31] 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1670 1671 1672 1673
```

```
## [46] 1674 1675 1676 1677 1678 1679 1680 1681 1682 1683 1684 1685 1686 1687 1688
## [61] 1689 1690 1691 1692 1693 1694 1695 1696 1697 1698 1699 1700 1701 1702 1703
## [76] 1704 1705 1706 1707 1708 1709 1710
```

Exploratory Data Analysis

number of boys baptized each year

$ –> Using this function we can access the data in a single column of a data frame separately.

```
arbuthnot$boys
```

```
##  [1] 5218 4858 4422 4994 5158 5035 5106 4917 4703 5359 5366 5518 5470 5460 4793
## [16] 4107 4047 3768 3796 3363 3079 2890 3231 3220 3196 3441 3655 3668 3396 3157
## [31] 3209 3724 4748 5216 5411 6041 5114 4678 5616 6073 6506 6278 6449 6443 6073
## [46] 6113 6058 6552 6423 6568 6247 6548 6822 6909 7577 7575 7484 7575 7737 7487
## [61] 7604 7909 7662 7602 7676 6985 7263 7632 8062 8426 7911 7578 8102 8031 7765
## [76] 6113 8366 7952 8379 8239 7840 7640
```

What command would you use to extract just the counts of girls born?

```
arbuthnot$girls
```

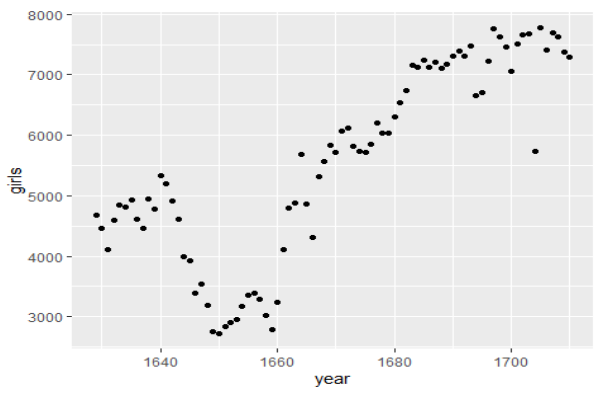
```
##  [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910 4617
## [16] 3997 3919 3395 3536 3181 2746 2722 2840 2908 2959 3179 3349 3382 3289 3013
## [31] 2781 3247 4107 4803 4881 5681 4858 4319 5322 5560 5829 5719 6061 6120 5822
## [46] 5738 5717 5847 6203 6033 6041 6299 6533 6744 7158 7127 7246 7119 7214 7101
## [61] 7167 7302 7392 7316 7483 6647 6713 7229 7767 7626 7452 7061 7514 7656 7683
## [76] 5738 7779 7417 7687 7623 7380 7288
```

A simple plot of the number of girls baptized per year with the command

Synthax explanation:

- The first argument is always the dataset.
- aes –> the x and the y axes.
- + -- > use for for another layer,

```
ggplot(data = arbuthnot, aes(x = year, y = girls)) +
  geom_point()
```



To seek help regarding the ggplot

```
?ggplot
```

Total number of baptisms

```
arbuthnot$boys + arbuthnot$girls
```

```
##  [1]  9901  9315  8524  9584  9997  9855 10034  9522  9160 10311 10150 10850
## [13] 10670 10370  9410  8104  7966  7163  7332  6544  5825  5612  6071  6128
## [25]  6155  6620  7004  7050  6685  6170  5990  6971  8855 10019 10292 11722
## [37]  9972  8997 10938 11633 12335 11997 12510 12563 11895 11851 11775 12399
## [49] 12626 12601 12288 12847 13355 13653 14735 14702 14730 14694 14951 14588
## [61] 14771 15211 15054 14918 15159 13632 13976 14861 15829 16052 15363 14639
## [73] 15616 15687 15448 11851 16145 15369 16066 15862 15220 14928
```

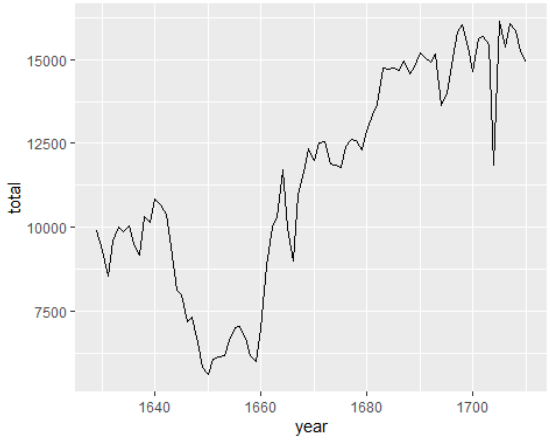
Adding a new variable to the data frame

The %>% operator is called the piping operator.Takes the output of the current line and pipes it into the following line of code.

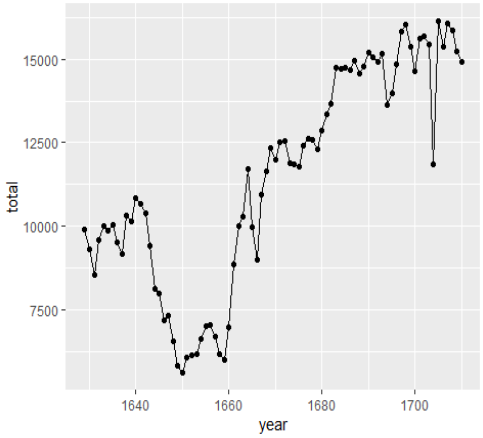
Take the arbuthnot dataset and pipe it into the mutate function. Using this mutate a new variable called total that is the sum of the variables called boys and girls. Then assign this new resulting dataset to the object called arbuthnot, i.e. overwrite the old arbuthnot dataset with the new one containing the new variable.

```
arbuthnot <- arbuthnot %>%
  mutate(total = boys + girls)

ggplot(data = arbuthnot, aes(x=year, y=total)) +
  geom_line()
```

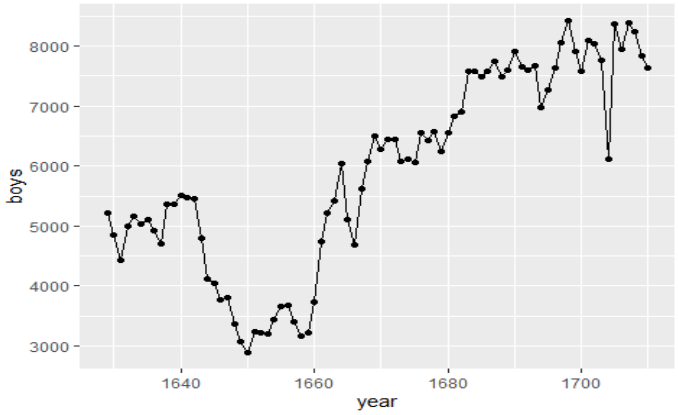


```
ggplot(data = arbuthnot, aes(x = year, y = total)) +
  geom_line() +
  geom_point()
```



Now, generate a plot of the proportion of boys born over time.

```
ggplot(data = arbuthnot, aes(x = year, y=boys)) +
  geom_line()+
  geom_point()
```



if boys outnumber girls in each year with the expression ?

```
arbuthnot <- arbuthnot %>%
  mutate(more_boys = boys > girls)
```