Details of Group & Members

Course – DAB501_23W

Section – 002

Group – 008

Members:

Harsimranjit Kaur - 0812147
Bhavitaben Bhatt - 0814912
Yashmeen Singh Chadha – 0804302

Statements of Integrity

I, Harsimranjit Kaur, confirm that my contribution to this project contains my own work. I have adhered to the St. Clair College's Integrity Policies and have acknowledged all the citations and references that I have used to complete this project.

I, Bhavitaben Bhatt, confirm that my contribution to this project contains my own work. I have adhered to the St. Clair College's Integrity Policies and have acknowledged all the citations and references that I have used to complete this project.

I, Yashmeen Singh Chadha, confirm that my contribution to this project contains my own work. I have adhered to the St. Clair College's Integrity Policies and have acknowledged all the citations and references that I have used to complete this project.

Software Version

Console - R version 4.2.2 (2022-10-31 ucrt)

About - RStudio 2022.12.0 Build 353
"Elsbeth Geranium" Release (7d165dcf, 2022-12-03) for Windows
Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) RStudio/2022.12.0+353 Chrome/102.0.5005.167 Electron/19.1.3 Safari/537.36

R Packages Used in this Project

dplyr – 1.1.0
ggplot2 – 3.4.0
tidyverse – 1.3.2
AER – 1.2.10
carData – 3.0.5
nlme – 3.1.162
dslabs – 0.7.4
openintro – 2.4.0
modeldata – 1.1.0

Documentation, Codes and Visualization of the Data Sets

Two Plots Displaying the distribution of a single continuous variable:

1. **Data Set** - Blackmore

Package – carData

Link - https://vincentarelbundock.github.io/Rdatasets/datasets.html

Attribution of the Owner/Creator of Data - Personal communication from Elizabeth Blackmore and Caroline Davis, York University.

Summary - This data frame studies 138 teenage girls hospitalized for eating disorders and 98 control subjects at an interval of 2 years to know about the amount of exercise they engaged in per week.

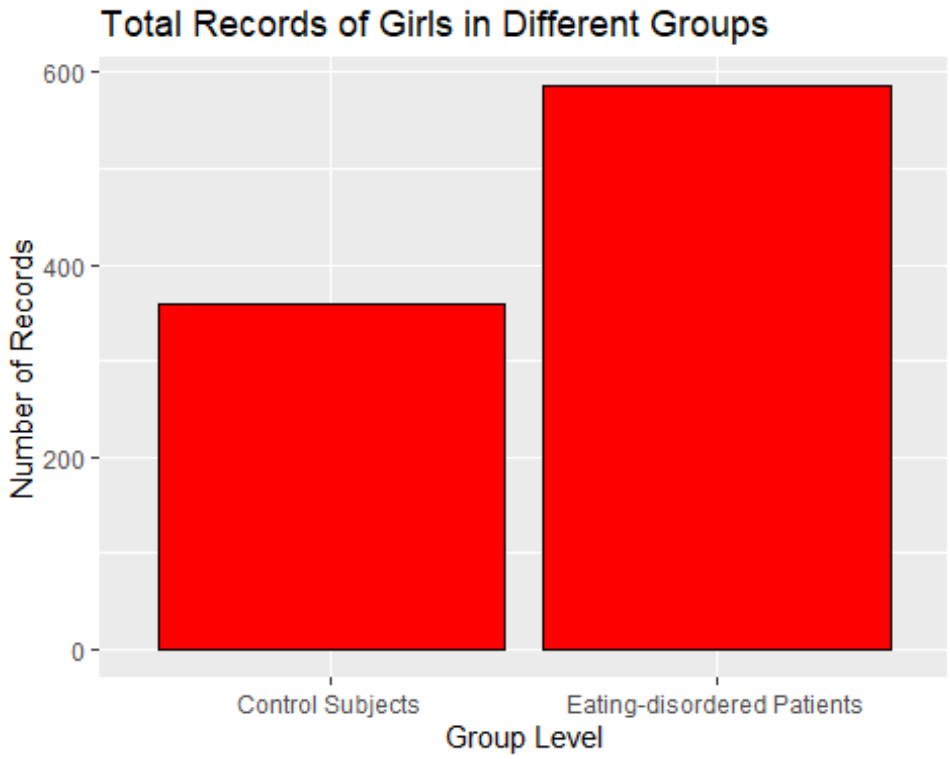
Below is the code and visualization for importing dataset, studying it, and creating the plot displaying distribution of a single continuous variable, i.e., Group Level.

```
#Loading the package required for dataset
library(carData)

#Loading the "Blackmore" Dataset
Blackmore

#Loading the dplyr library
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

#Details of Variables

#Getting a glimpse of the dataframe
glimpse(Blackmore)

## Rows: 945
## Columns: 4
## $ subject  <fct> 100, 100, 100, 100, 100, 101, 101, 101, 101, 101, 102,
102, 1…
## $ age      <dbl> 8.00, 10.00, 12.00, 14.00, 15.92, 8.00, 10.00, 12.00,
14.00, …
## $ exercise <dbl> 2.71, 1.94, 2.36, 1.54, 8.63, 0.14, 0.14, 0.00, 0.00,
```

```
5.08, 0…
## $ group    <fct> patient, patient, patient, patient, patient, patient,
patient…

#Getting the summary of the dataset
summary(Blackmore)

##     subject         age           exercise          group
##  100    :  5   Min.   : 8.00   Min.   : 0.000   control:359
##  101    :  5   1st Qu.:10.00   1st Qu.: 0.400   patient:586
##  105    :  5   Median :12.00   Median : 1.330
##  106    :  5   Mean   :11.44   Mean   : 2.531
##  107    :  5   3rd Qu.:14.00   3rd Qu.: 3.040
##  108    :  5   Max.   :17.92   Max.   :29.960
##  (Other):915

#Getting the information about the dataset to understand its features
?Blackmore

## starting httpd help server ...

##  done
```

Creating the bar plot to display single continuous variable.

```
#Loading the ggplot2 package
library(ggplot2)

#Creating a bar graph to show the distribution of observed teenage girls
among Eating-Disordered Patients and Control Subjects groups.

#ggplot defines the global features of the plot.
#goem_bar defines the attributes of the bar plot.
#ggtitle, xlab and ylab are used to rename the title and axis of the plot.
#scale_x_discrete is used to change/define the labels along x-axis.

ggplot(data = Blackmore, mapping = aes(x= group)) +
  geom_bar(color = "black",
           fill = "red") +
  ggtitle("Total Records of Girls in Different Groups") +
  xlab("Group Level") +
  ylab("Number of Records") +
  scale_x_discrete(labels=c('Control Subjects', 'Eating-disordered
Patients'))
```



The above Bar plot in R is created to represent the total number of recorded cases of girls that had been hospitalized under the two groups that were studied. Since the study was conducted 138 girls over the years, there are multiple records of same subject(girl) recorded at different ages.

The bin width of the bars is kept to the default value of 10, while red color is used to fill the bars and black color is used to outline them.

The plot can be used to understand the recorded number of cases of the same group of girls throughout their teenage years. Here, we can conclude that there were more reported cases of eating-disorder patients as compared to that of control subjects.

---

2. Data Set - Stroke Prediction Dataset

Link - https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Attribution of the Owner/Creator of Data - Kaggle contribution by fedesoriano (Owner) https://www.kaggle.com/fedesoriano

Summary - This data set contains information of various parameters related to stroke like gender, age, hypertension, heart disease, marriage status, work type and so on. It contains 5110 observations and 12 features.

Below is the code and visualization for importing dataset, studying it, and creating the plot displaying distribution of a single continuous variable, i.e., Group Level.

```r
# Loading the CSV file from  local machine

data_stroke = read.csv("C:\\Users\\91966\\Desktop\\Semester 1\\501 -  BASIC STATS & EXPL DATA ANALYS\\Labs\\Assignment\\healthcare-dataset-stroke-data.csv")
glimpse(data_stroke)
```

```
## Rows: 5,110
## Columns: 12
## $ id                <int> 9046, 51676, 31112, 60182, 1665, 56669, 53882, 10434…
## $ gender            <chr> "Male", "Female", "Male", "Female", "Female", "Male"…
## $ age               <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, …
## $ hypertension      <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1…
## $ heart_disease     <int> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0…
## $ ever_married      <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No…
## $ work_type         <chr> "Private", "Self-employed", "Private", "Private", "S…
## $ Residence_type    <chr> "Urban", "Rural", "Rural", "Urban", "Rural", "Urban"…
## $ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.0…
## $ bmi               <chr> "36.6", "N/A", "32.5", "34.4", "24", "29", "27.4", "…
## $ smoking_status    <chr> "formerly smoked", "never smoked", "never smoked", "…
## $ stroke            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
```

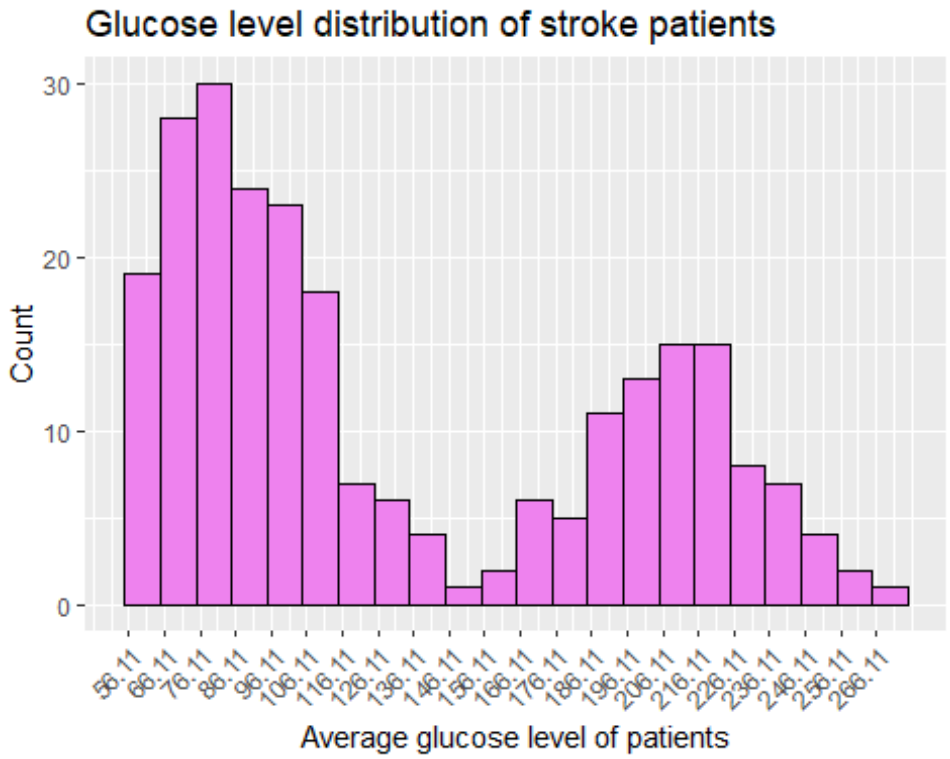
```r
summary(data_stroke)
```

```
##        id           gender               age         hypertension
##  Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Mode  :character   Median :45.00   Median :0.00000
##  Mean   :36518                      Mean   :43.23   Mean   :0.09746
##  3rd Qu.:54682                      3rd Qu.:61.00   3rd Qu.:0.00000
##  Max.   :72940                      Max.   :82.00   Max.   :1.00000
##  heart_disease     ever_married        work_type         Residence_type
##  Min.   :0.00000   Length:5110        Length:5110        Length:5110
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.05401
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

```
##  avg_glucose_level     bmi            smoking_status         stroke
##  Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000
##  1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000
##  Median : 91.89    Mode  :character   Mode  :character   Median :0.00000
##  Mean   :106.15                                          Mean   :0.04873
##  3rd Qu.:114.09                                          3rd Qu.:0.00000
##  Max.   :271.74                                          Max.   :1.00000

stroke_patient <- data_stroke %>% filter(stroke == 1)
require(ggplot2)
```

Create a histogram plot displaying the distribution of a single continuous variable.

```
# ggplot defines the global features of the plot.
# geom_histogram creates histogram with bin width,color and fill.
# lab provide title of chart and labels for both axis.
# scale_x_continuous is used for altering defaulting sequence, it generate
# sequence from min glucose level to maximum glucose level.
#theme function is used to rorate the x axis label and maintain the space of
it from graph.

ggplot( stroke_patient, aes(x= avg_glucose_level)) +
  geom_histogram(binwidth= 10, colour="black", fill="violet") +
  labs(title = "Glucose level distribution of stroke patients",
  x="Average glucose level of patients",
  y = "Count") +
  scale_x_continuous(breaks = seq(min(stroke_patient$avg_glucose_level),
max(stroke_patient$avg_glucose_level), 10)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Above graph represents average glucose level of patients who had stroke. To fetch the patients with the history of heart stroke we have used filter function on the integer column stroke. 1 means patient had stroke.

In this code chunk we are trying to figure out what was the range of average glucose level for stroke patients.

Since we have filtered patients, we have used stroke_patient as data for creating histogram where x-axis contains average glucose level and y axis contains count of it. To represent it elegantly bin width is set to 20, x axis contains range of interval starting from minimum glucose level to maximum glucose level, color of border is set as black in combination with violet as bin color.

The most common level of glucose (for 30 patients) in stroke patients is between 76.11 to 86.11, very few patients have average glucose level above 266.11.

Two plots displaying information about a single categorical variable:

1. **Dataset** - Spruce

Package - nlme

Link - https://vincentarelbundock.github.io/Rdatasets/datasets.html

Attribution of the Owner/Creator of Data - Pinheiro, J. C. and Bates, D. M. (2000), Mixed-Effects Models in S and S-PLUS, Springer, New York. (Appendix A.28)

Diggle, Peter J., Liang, Kung-Yee and Zeger, Scott L. (1994), Analysis of longitudinal data, Oxford University Press, Oxford.

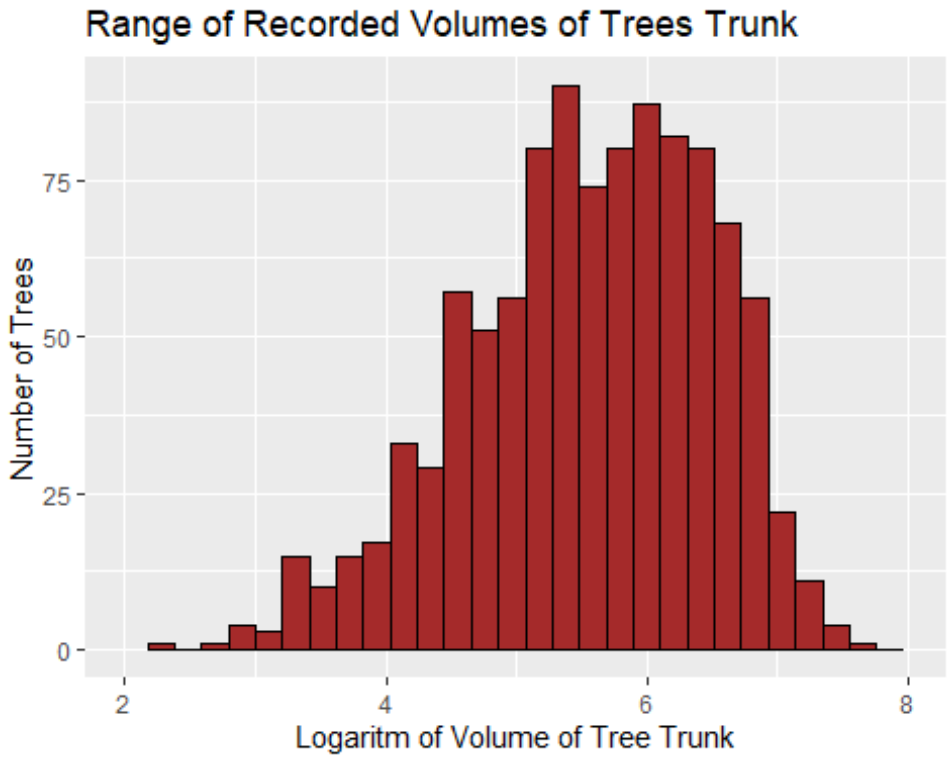
Summary - This data frame was collected in 1994 and it describes the information about the growth of spruce trees planted on different plots, that have either been exposed to a normal atmosphere or an ozone-rich atmosphere.

```
#Loading the package required for dataset
library(nlme)

#Loading the "Spruce" Dataset
Spruce

#Loading the dplyr library
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:nlme':
##
##     collapse

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

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

#Details of Variables

#Getting a glimpse of the dataframe
glimpse(Spruce)

## Rows: 1,027
## Columns: 4
## $ Tree    <ord> O1T01, O1T01, O1T01, O1T01, O1T01, O1T01, O1T01, O1T01,
O1T01,…
## $ days    <dbl> 152, 174, 201, 227, 258, 469, 496, 528, 556, 579, 613,
```

```
639, 67…
## $ logSize <dbl> 4.51, 4.98, 5.41, 5.90, 6.15, 6.16, 6.18, 6.48, 6.65,
6.87, 6.…
## $ plot    <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1,…

#Getting the information about the dataset to understand its features
?Spruce

## starting httpd help server ...

##  done

#Getting the summary of the dataset
summary(Spruce)

##       Tree           days           logSize      plot
##  O1T24  : 13   Min.   :152.0   Min.   :2.230   1:351
##  O1T18  : 13   1st Qu.:227.0   1st Qu.:4.945   2:351
##  O1T19  : 13   Median :496.0   Median :5.630   3:156
##  O1T15  : 13   Mean   :428.2   Mean   :5.548   4:169
##  O1T10  : 13   3rd Qu.:579.0   3rd Qu.:6.250
##  O1T26  : 13   Max.   :674.0   Max.   :7.560
##  (Other):949
```

Creating the histogram plot to display single categorical variable.

```
#Loading the ggplot2 package
library(ggplot2)

#Creating a histogram showing the range of logarithm of an estimate of the
volume of the tree trunk for all the spruce tree observations.

#ggplot defines the global features of the plot.
#geom_histogram defines the attributes of the histogram plot.
#xlim is used to define the scale range of the x-axis.
#ggtitle, xlab and ylab are used to rename the title and axis of the plot.

ggplot(data = Spruce) +
  geom_histogram(mapping = aes(x = logSize),
                 fill = 'brown',
                 colour = 'black') +
  xlim(2,8) +
  ggtitle("Range of Recorded Volumes of Trees Trunk") +
  xlab("Logaritm of Volume of Tree Trunk") +
  ylab("Number of Trees")

## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



The above Histogram plot is created in R to represent the logarithm value of the volume of the tree trunks across the trees the spruce trees that were studied.

The bin width of the bars is kept to the default value of 10, while brown color is used to fill the bars and black color is used to outline them.

The plot can be used to understand the logarithm of the volume of the tree trunks over the number of days for which the experiment was conducted. Here, we can conclude that the trunk volume of most of the trees lies roughly between 4.3 to 7.

---

2. Dataset - Stroke Prediction Dataset

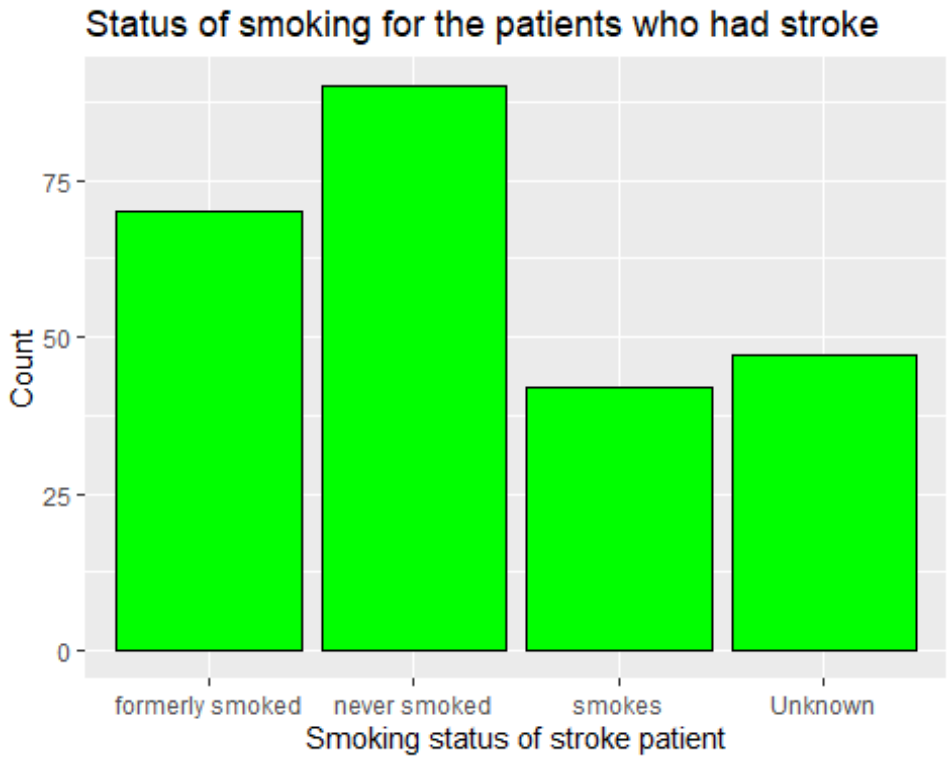
Link - https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Attribution of the Owner/Creator of Data - Kaggle contribution by fedesoriano (Owner) https://www.kaggle.com/fedesoriano

Summary - This data set contains information of various parameters related to stroke like gender, age, hypertension, heart disease, marriage status, work type and so on. It contains 5110 observations and 12 features.

```
# ggplot defines the global features of the plot.
# goem_bar defines the attributes of the bar plot.
# # lab provide title of chart and labels for both axis.

ggplot( stroke_patient, aes(x= smoking_status)) +
  geom_bar( fill="green", colour="black") +
```

```
labs(title = "Status of smoking for the patients who had stroke",
x="Smoking status of stroke patient",
y = "Count")
```



This graph represents various categories of smoking which are formerly smoked, never smoked, smokes and unknown. Surprisingly, majority patients who were not smoking got the stroke and those who smokes are the least who got stroke. The x axis contains status of smoking whereas Y axis contains the count of it. The bar graph is filled with green colour and black boarder. For categorical distribution smoking_status attribute is used.

The geom_bar function portrait graph with fill and colour parameter and labs function defines X-axis, Y-axis and graph label as Smoking status of stroke patient, Count and Status of smoking for the patients who had stroke.

One plot displaying information about both a continuous variable and a categorical variable:

Dataset - us_contagious_diseases

Package - dslabs

Link - https://vincentarelbundock.github.io/Rdatasets/datasets.html

Original data - Tycho Project (http://www.tycho.pitt.edu/)

Attribution of the Owner/Creator of Data - Willem G. van Panhuis, John Grefenstette, Su Yon Jung, Nian Shong Chok, Anne Cross, Heather Eng, Bruce Y Lee, Vladimir Zadorozhny, Shawn Brown, Derek Cummings, Donald S. Burke.

Summary - This data frame contains the yearly counts of cases reported from 1928 to 2011 for 7 diseases - Hepatitis A, Measles, Mumps, Pertussis, Polio, Rubella, and Smallpox across different states in US.

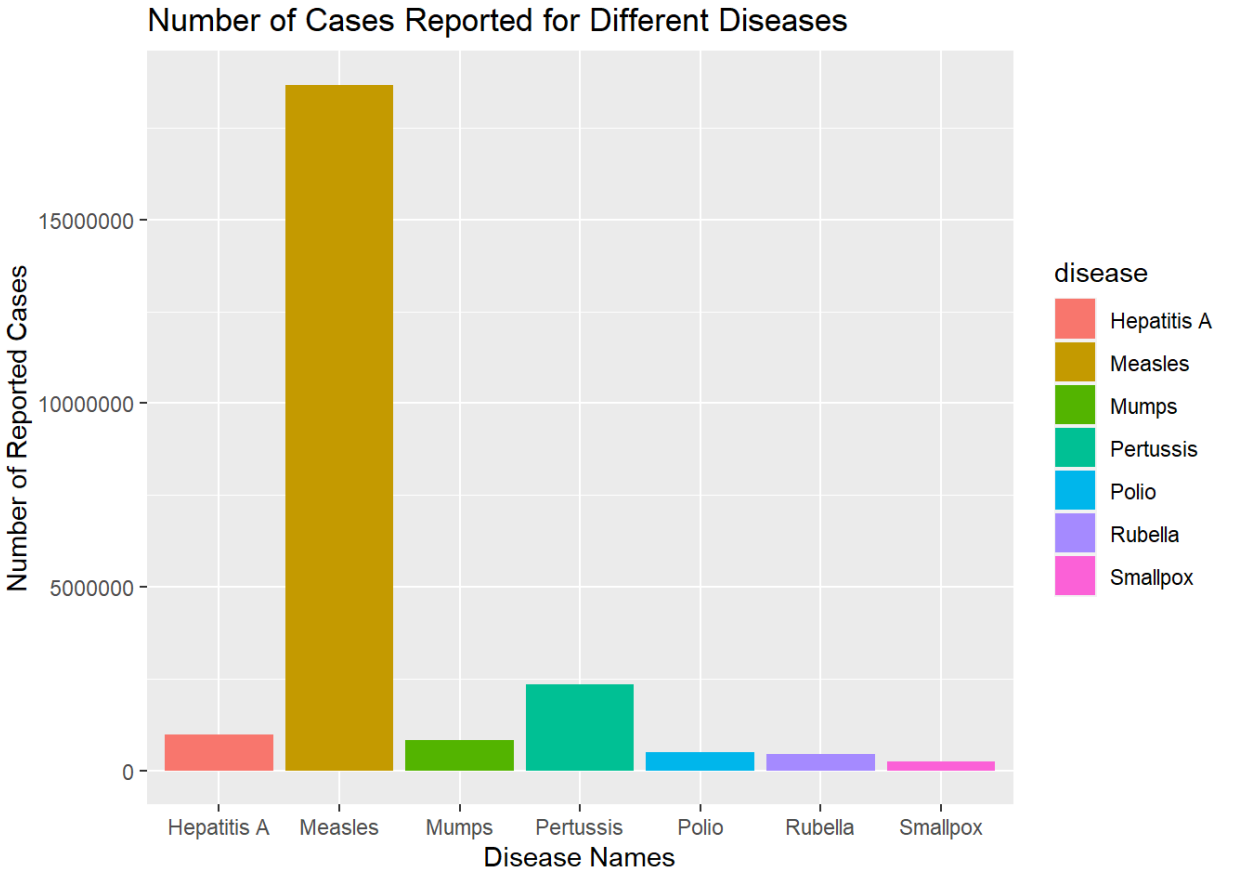
```
#Loading the package required for dataset
library(dslabs)

#Loading the "Us Contagious Diseases" Dataset
us_contagious_diseases

#Loading the dplyr library
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

#Details of Variables

#Getting a glimpse of the dataframe
glimpse(us_contagious_diseases)

## Rows: 16,065
## Columns: 6
## $ disease         <fct> Hepatitis A, Hepatitis A, Hepatitis A, Hepatitis
A, He…
## $ state           <fct> Alabama, Alabama, Alabama, Alabama, Alabama,
Alabama, …
## $ year            <dbl> 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973,
1974, …
```

```
## $ weeks_reporting <dbl> 50, 49, 52, 49, 51, 51, 45, 45, 45, 46, 50, 43,
41, 47…
## $ count           <dbl> 321, 291, 314, 380, 413, 378, 342, 467, 244, 286,
220,…
## $ population      <dbl> 3345787, 3364130, 3386068, 3412450, 3444165,
3481798, …

#Getting the information about the dataset to understand its features
?us_contagious_diseases

## starting httpd help server ...

##  done

#Getting the summary of datset
summary(us_contagious_diseases)

##         disease            state            year      weeks_reporting
##  Hepatitis A:2346   Alabama   :  315   Min.   :1928   Min.   : 0.00
##  Measles    :3825   Alaska    :  315   1st Qu.:1950   1st Qu.:31.00
##  Mumps      :1785   Arizona   :  315   Median :1975   Median :46.00
##  Pertussis  :2856   Arkansas  :  315   Mean   :1971   Mean   :37.38
##  Polio      :2091   California:  315   3rd Qu.:1990   3rd Qu.:50.00
##  Rubella    :1887   Colorado  :  315   Max.   :2011   Max.   :52.00
##  Smallpox   :1275   (Other)   :14175
##      count          population
##  Min.   :     0   Min.   :   86853
##  1st Qu.:     7   1st Qu.: 1018755
##  Median :    69   Median : 2749249
##  Mean   :  1492   Mean   : 4107584
##  3rd Qu.:   525   3rd Qu.: 4996229
##  Max.   :132342   Max.   :37607525
##                   NA's   :214
```

Creating the bar plot to display both a continuous variable and a categorical variable.

```
#Loading the ggplot2 package
library(ggplot2)

#Removing the scientific notations from the plot.
options(scipen=999)

#Creating a bar graph to show the number of cases that were reported over the
years in different states, for the contagious diseases that were being
studied.

#ggplot defines the global features of the plot.
#goem_bar defines the attributes of the bar plot.
#ggtitle, xlab and ylab are used to rename the title and axis of the plot.

ggplot(data = us_contagious_diseases,
```

```
      mapping = aes(x = disease, y = count, fill = disease)) +
  geom_bar(stat = "identity") +
  ggtitle("Number of Cases Reported for Different Diseases") +
  xlab("Disease Name") +
  ylab("Number of Reported Cases")
```



The above Bar plot is created in R to represent the total number of cases reported for the contagious diseases that were being studied over the years, across different US states.

The bin width of the bars is kept to the default value of 10, while the disease names, a feature that is displayed on the X axis is used to fill the bar colours too.

The plot can be used to understand the number of cases that were reported for various contagious diseases over the years. Here, we can clearly conclude that the most contagious disease during those years of study in US states has been Measles while the least cases has been that of Smallpox.

Two plots displaying the information that shows a relationship between two variables:

1. **Data set** - Affairs

Package - AER

Attribution of the Source of Data - Below graph uses R's data set named Affairs {AER}. Online complements to Greene (2003). Table F22.2.

Summary - Basically, this is infidelity data, known as Fair's Affairs. It is gathered by conducting survey by Psychology Today in 1969. It contains 601 observations and 9 attributes, emphasizing on number of affairs and various other correlated factor like years of marriage, no of children, religiousness and so on.

```
library(AER)
```

```
## Loading required package: car
```

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

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

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

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

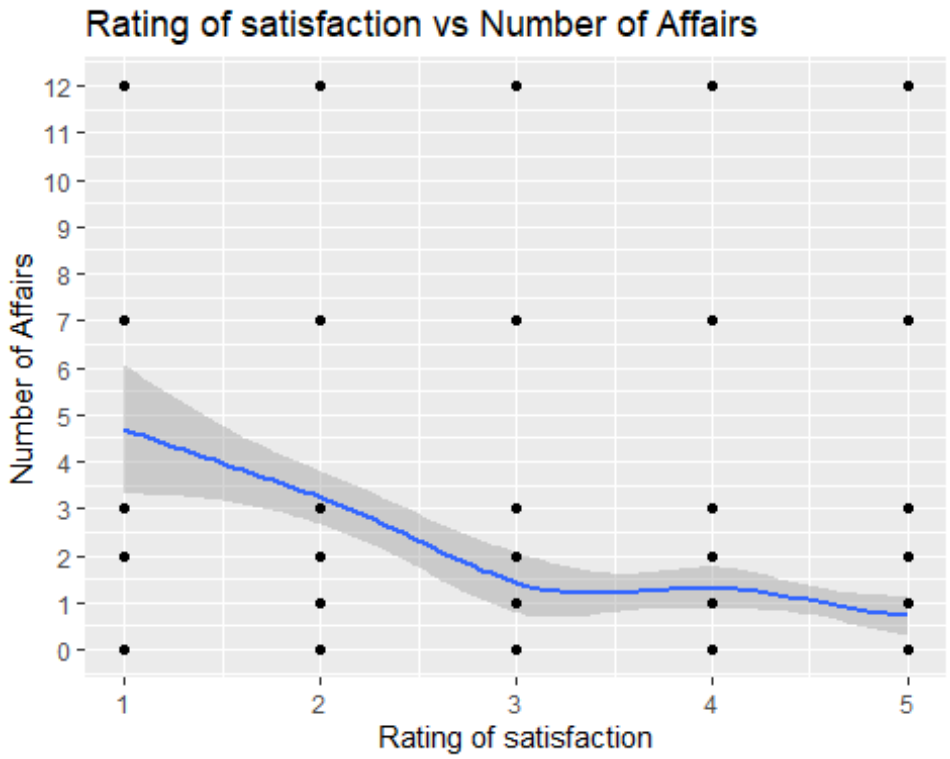
```
## Loading required package: survival
```

```
data(Affairs)
glimpse(Affairs)
```

```
## Rows: 601
## Columns: 9
## $ affairs       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

```
0, 0,…
## $ gender        <fct> male, female, female, male, male, female, female,
male, …
## $ age           <dbl> 37, 27, 32, 57, 22, 32, 22, 57, 32, 22, 37, 27, 47,
22, …
## $ yearsmarried  <dbl> 10.00, 4.00, 15.00, 15.00, 0.75, 1.50, 0.75, 15.00,
15.0…
## $ children      <fct> no, no, yes, yes, no, no, no, yes, yes, no, yes,
yes, ye…
## $ religiousness <int> 3, 4, 1, 5, 2, 2, 2, 2, 4, 4, 2, 4, 5, 2, 4, 1, 2,
3, 2,…
## $ education     <dbl> 18, 14, 12, 18, 17, 17, 12, 14, 16, 14, 20, 18, 17,
17, …
## $ occupation    <int> 7, 6, 1, 6, 6, 5, 1, 4, 1, 4, 7, 6, 6, 5, 5, 5, 4,
5, 5,…
## $ rating        <int> 4, 4, 4, 5, 3, 5, 3, 4, 2, 5, 2, 4, 4, 4, 4, 5, 3,
4, 5,…

summary(Affairs)

##     affairs          gender         age         yearsmarried    children
##  Min.   : 0.000   female:315   Min.   :17.50   Min.   : 0.125   no :171
##  1st Qu.: 0.000   male  :286   1st Qu.:27.00   1st Qu.: 4.000   yes:430
##  Median : 0.000                Median :32.00   Median : 7.000
##  Mean   : 1.456                Mean   :32.49   Mean   : 8.178
##  3rd Qu.: 0.000                3rd Qu.:37.00   3rd Qu.:15.000
##  Max.   :12.000                Max.   :57.00   Max.   :15.000
##  religiousness     education       occupation        rating
##  Min.   :1.000   Min.   : 9.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:14.00   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :16.00   Median :5.000   Median :4.000
##  Mean   :3.116   Mean   :16.17   Mean   :4.195   Mean   :3.932
##  3rd Qu.:4.000   3rd Qu.:18.00   3rd Qu.:6.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :20.00   Max.   :7.000   Max.   :5.000

# min provides least value from series of number which is column here.
# max provides highest value from series of number which is column here.
# geom_smooth is used here to observe trend line
# geom_point is used here to represent the data points.
# scale_y_continuous gives dynamic range to Y axis.

min_affair <- min(Affairs$affairs)
max_affair <- max(Affairs$affairs)
ggplot(Affairs, aes(x = rating, y = affairs)) +
  geom_smooth() +
  geom_point() +
  labs(title = "Rating of satisfaction vs Number of Affairs",
  x="Rating of satisfaction",
  y = "Number of Affairs") +
  scale_y_continuous(breaks = seq(min_affair, max_affair))
```

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

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 5.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 2.0116e-15

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 1

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used
at
## 5.02

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood
radius 2.02

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal
condition
## number 2.0116e-15

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other
near
## singularities as well. 1
```



In the above plot, on the X axis rating of marriage is presented and on y axis Number of Affairs. It shows that the higher satisfaction in marriage leads to a smaller number of affairs.

For Y-axis range scale_y_continuous is used which considers min and max values of affairs.

X-axis is labelled as Rating of satisfaction, Y-axis is labelled as Number of Affairs, graph title is set as Rating of satisfaction vs Number of Affairs and geom_smmoth() adds trend line which is in downward trend over here. Downward trend shows that if person is satisfied in marriage, then number of affairs will be significantly less compared to those who are unsatisfied in marriage.

---

2. **Dataset** - lending_club

Package – modeldata

Link - https://vincentarelbundock.github.io/Rdatasets/datasets.html

Attribution of the Owner/Creator of Data -
https://www.lendingclub.com/info/download-data.action

Summary - The Data Frame used in the above Plot is lending_club imported from the library "modeldata". The Data Set we are using is about the loan data providing us information about the funded loan amount, term, interest rate, verification status, annual income, employment length etc.

Below is the code and visualization for importing dataset, studying it, and creating the plot displaying two plots should display information that shows a relationship between two variables.

```
# Loading the Library and Packages
library (tidyverse)
```

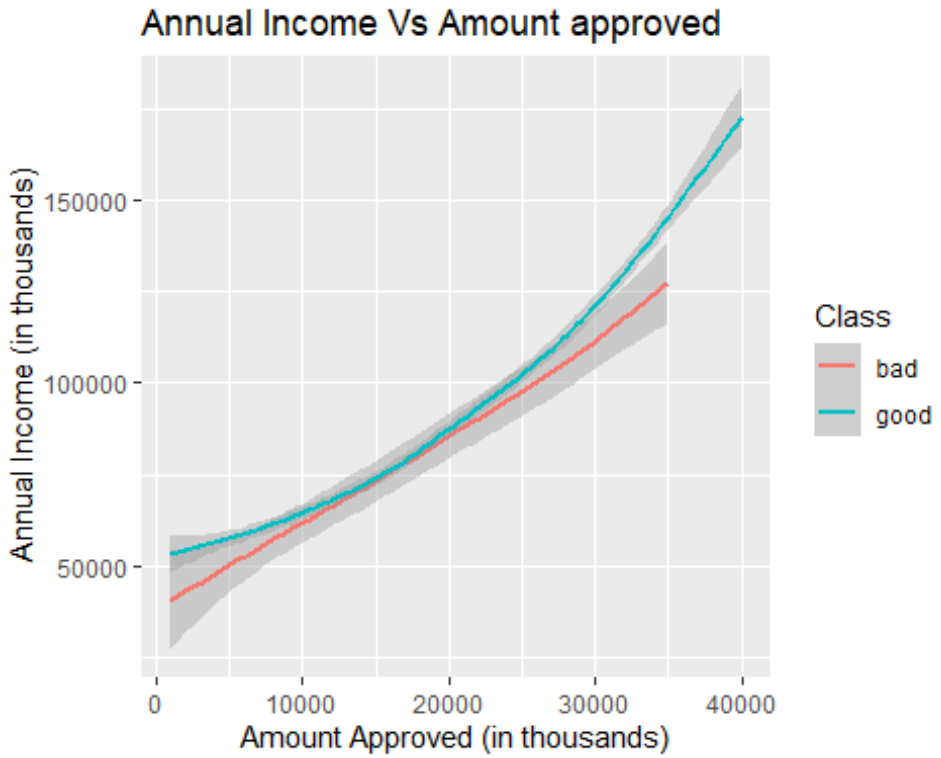
```
## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.4.0     ✓ purrr   1.0.1
## ✓ tibble  3.1.8     ✓ dplyr   1.1.0
## ✓ tidyr   1.2.1     ✓ stringr 1.5.0
## ✓ readr   2.1.4     ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library("modeldata")

# Loading the First few entries of the Data Frame

head(lending_club)
```

```
## # A tibble: 6 × 23
##   funded…¹ term  int_r…² sub_g…³ addr_…⁴ verif…⁵ annua…⁶ emp_l…⁷ delin…⁸
inq_l…⁹
##      <int> <fct>   <dbl> <fct>   <fct>   <fct>     <dbl> <fct>     <int>
<int>
## 1    16100 term…   14.0  C4      CT      Not_Ve…   35000 emp_5         0
0
## 2    32000 term…   12.0  C1      MN      Verifi…   72000 emp_ge…       0
0
## 3    10000 term…   16.3  D1      OH      Source…   72000 emp_ge…       0
2
## 4    16800 term…   13.7  C3      NV      Verifi…  101000 emp_lt…       0
0
## 5     3500 term…    7.39 A4      CA      Source…   50100 emp_unk       0
0
## 6    10000 term…   11.5  B5      TX      Source…   32000 emp_lt…       0
0
## # … with 13 more variables: revol_util <dbl>, acc_now_delinq <int>,
## #   open_il_6m <int>, open_il_12m <int>, open_il_24m <int>, total_bal_il
<int>,
## #   all_util <int>, inq_fi <int>, inq_last_12m <int>, delinq_amnt <int>,
## #   num_il_tl <int>, total_il_high_credit_limit <int>, Class <fct>, and
## #   abbreviated variable names ¹funded_amnt, ²int_rate, ³sub_grade,
## #   ⁴addr_state, ⁵verification_status, ⁶annual_inc, ⁷emp_length, ⁸
delinq_2yrs,
## #   ⁹inq_last_6mths
```

```r
# Creating the Plot
ggplot(lending_club, aes(x = funded_amnt, y = annual_inc)) +
    geom_smooth( aes(color = Class))+
  ggtitle("Annual Income Vs Amount approved") +
  xlab ("Amount Approved (in thousands)")+
  ylab (" Annual Income (in thousands)")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



The plot we have used is a Geom smooth plot comparing two attributes Annual Income and the Amount funded. According to the graph we could see the positive regression towards the amount funded as compared to the Annual Income. This displays that more the annual income of the loan applicant the larger is the Loan amount which is approved. But it is not the only deciding factor. Another factor which decides the Loan amount is the credit report whether it's good or bad. this plays an important part because no matter if the person has a higher income if he has bad credit history he might not be approved with relatively higher amount as compared to the person with higher income and good credit score.

One plot showing faceting and displaying information about 4 variables:

Dataset - lending_club

Package – modeldata

Link - https://vincentarelbundock.github.io/Rdatasets/datasets.html

Attribution of the Owner/Creator of Data - https://www.lendingclub.com/info/download-data.action

Summary - The Data Frame used in the above Plot is lending_club imported from the library "modeldata". The Data Set we are using is about the loan data providing us information about the funded loan amount, term, interest rate, verification status, annual income, employment length etc.

Below is the code and visualization displaying plot using faceting and display information about 4 variables. Here we used the same Data Set explained Above

```
ggplot(lending_club, aes(x = funded_amnt, y = annual_inc)) +
    geom_smooth( aes(color = Class))+
  facet_grid(term ~ verification_status)+
  ggtitle("Annual Income Vs Amount approved") +
  xlab ("Amount Approved (in thousands)")+
  ylab (" Annual Income (in thousands)")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

In the above plot we are comparing 4 variables that is annual Income, Amount approved , Credit History, verification status of the applicant and the Loan term. According to the Data there three types of verification status which are conducted on the applicant's application. One is verified through the normal Process. The second is Source verified where the applicant details are verified from the Source end from where the application is been Submitted. The above is a facet Grip plot comparing all the attributes Defined earlier. From this we can understand that there are only two terms for the loan amount being sanctioned which is 36 months and 60 months. This visually analyze the data and helps us the to give the result as to what category of applicants have the highest loan value and longest terms in comparison of their annual income and credit score.

Competition Plot: an opportunity to explore what's possible and get creative:

Dataset - ncbirths

Package – openintro

Link - https://vincentarelbundock.github.io/Rdatasets/datasets.html

Attribution of the Owner/Creator of Data - This is a random sample of 1,000 cases from a state collected data.

Summary – This data set is a random sample of 1,000 cases that have been taken from a study that was conducted by the state of North Carolina in 2014 containing the information about the birth records in the state. It has since been used by the medical researchers to understand the relationship between the birth of children and the practices as well as habits of the expectant mothers.

Below is the code and visualization for importing dataset, studying it, and creating the plot displaying competition plot.

This plot must use ggplot2 but additional packages may be used

for example:

- Interactive plots (ggplotly);
- different data sets in same plot;
- subplots; animation (gganimate); etc.

```
# Loading the Libraries
library(ggplot2)
library(plotly)

##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

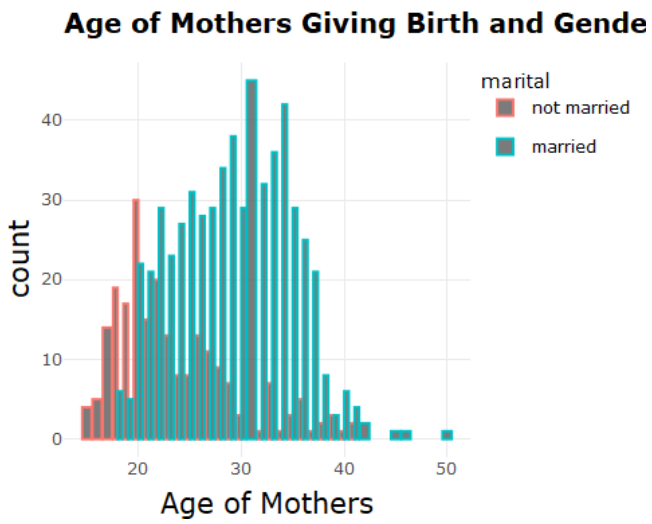
```
library("openintro")
```

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata

##
## Attaching package: 'openintro'

## The following object is masked from 'package:modeldata':
##
##     ames
```

```
# Reading the Data Frame

head(ncbirths)
```

```
## # A tibble: 6 × 13
##    fage  mage mature    weeks premie visits marital gained weight lowbi…¹
gender
##   <int> <int> <fct>     <int> <fct>   <int> <fct>    <int>  <dbl> <fct>
<fct>
## 1    NA    13 younger …    39 full …     10 not ma…     38   7.63 not low
male
## 2    NA    14 younger …    42 full …     15 not ma…     20   7.88 not low
male
## 3    19    15 younger …    37 full …     11 not ma…     38   6.63 not low
female
## 4    21    15 younger …    41 full …      6 not ma…     34   8    not low
male
## 5    NA    15 younger …    39 full …      9 not ma…     27   6.38 not low
female
## 6    NA    15 younger …    38 full …     19 not ma…     22   5.38 low
male
## # … with 2 more variables: habit <fct>, whitemom <fct>, and abbreviated
## #   variable name ¹lowbirthweight
```

```r
# Cleaning the Data and removing the Na values from the column Marital
ncbirths_clean <- ncbirths %>%
  na.omit(marital)

# Creating a Plot

plot <- ggplot(data = ncbirths_clean, mapping = aes(x = mage)) +
  geom_bar(aes(color = marital),
           position = "dodge",
           alpha = 0.8,
           show.legend = TRUE)+
  ggtitle("Age of Mothers Giving Birth and Gender of the Baby born") +
  xlab ("Age of Mothers")+
  theme_minimal() +
  theme(legend.position = "right",
        plot.title = element_text(size = 15, face = "bold"),
        axis.title = element_text(size=16))


# Creating a ggplotly

ggplotly(plot)
```



This code produces a bar plot showing the distribution of the age of mothers giving birth, separated by marital status, and with the color representing the gender of the baby born. Firstly,the na.omit() function is used to remove any rows with missing values in the marital variable from the original data set ncbirths. The resulting cleaned data set is stored in the ncbirths_clean object. Then, the ggplot() function is used to create a plot with ncbirths_clean as the data source. The geom_bar() function is used to create a bar plot of the age of mothers (mage variable) with the color aesthetic mapped to the marital variable. The position argument is set to “dodge” to separate the bars by marital status, and alpha is set to 0.8 to make the bars slightly transparent. The show.legend argument is set to TRUE to display the legend. The ggtitle() and xlab() functions are used to set the plot title and x-

axis label, respectively. The theme_minimal() function is used to set the plot theme, and the theme() function is used to adjust the legend position, plot title size and font weight, and axis title size.

Reference Links

https://www.statology.org/r-plot-change-axis-scale/

https://www.tutorialspoint.com/how-to-remove-scientific-notation-form-base-r-plot

https://statisticsglobe.com/change-colors-of-bars-in-ggplot2-barchart-in-r#example-1-drawing-ggplot2-barplot-with-default-colors

https://r-graph-gallery.com/ggplot2-package.html

https://www.tidyverse.org/#:~:text=The%20tidyverse%20is%20an%20opinionated,%2C%20grammar%2C%20and%20data%20structures.&text=See%20how%20the%20tidyverse%20makes,%E2%80%9CR%20for%20Data%20Science%E2%80%9D.

https://ggplot2.tidyverse.org/reference/scale_continuous.html

https://stackoverflow.com/questions/17216358/eliminating-nas-from-a-ggplot

---

Questions

In what ways do you think data visualization is important to understanding a data set?

Data Visualization is important to display the data in a format that is easy to read, attractive, and understandable for even those audiences who might not know anything about the data set or its history.

Sometimes, the data is too vast to understand and read quickly and data visualization helps in reading different features of the data set comprehensively. For instance, a scatter plot would help us understand the distribution of features and any disparity that might have occurred.

In what ways do you think data visualization is important to communicating important aspects of a data set?

There is a famous saying – "A picture is worth a thousand words and data visualization proves that to be correct. It is used to ensure effective communication of the data set.

When working in organizations, the data is usually complex and is a huge amount. Even though data scientists can find the patterns and desired results from the pool of data without visualization too, it is often a challenge for them to communicate the findings to their audience as they might not be familiar with the syntaxes and coding language used.

Data visualization helps data scientists in communicating their findings clearly and interactively so that it could aid the decision-making process.

What role does your integrity as an analyst play when creating a data visualization for communicating results to others?

Integrity is one of the most important traits to possess for an analyst. When creating data visualization, one must realize that the results that have been depicted in those plots and the information that is being communicated will form the basis for some of the major business decisions that an organization would take.

Now, if the analyst has not been honest in calculating the findings and presenting the results, it would result in misinterpreted results. Such a scenario would ultimately cause the management to make wrong decisions and that could be risky for the business.

How many variables do you think you can successfully represent in a visualization? What happens when you exceed this number?

The number of variables that can be represented in a visualization depends on the findings that an analyst is trying to communicate.

However, as a generic rule, it is advisable to not represent more than 4 variables on a single plot as that might result in too much information being displayed in a single frame and that could be difficult to interpret for some. After all, the entire purpose of visualization is to ensure that the data is easy and quick to read.

Members Contribution

Harsimranjit Kaur – 0812147

I am responsible for the creation of bar plots using the Blackmore and US Contagious Diseases data sets, and histogram plot using the Spruce data set. I have also contributed to answering the questions and created all the documentations for this project.

Bhavitaben Bhatt – 0814912

I am responsible for creation of histogram plot and bar plot using the Stroke Prediction data set. I have also created the point plot using Affairs dataset and have contributed to answering the questions.

Yashmeen Singh Chadha – 0804302

I am responsible for creation of the two smooth plots and using the Lending Club data set along with the bar graph using the NC Births data set. I have also contributed to answering the questions.