# REPORTS MADE EASY

DIR-TERRANCE[1]

2021-02-08

[1] FOUNDER:RWILLS STATISTICAL COMPANY,EMAIL:consultancyrwillsstats@gmail.com

# Contents

**Prerequisites** **5**

0.1 Packages. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

**1 Introduction** **7**

1.1 THE CHALLLENGER. . . . . . . . . . . . . . . . . . . . . . . . . 7

1.2 DATA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

**2 EXPLANATORY DATA ANALYSIS** **11**

2.1 DATA WRANGLING . . . . . . . . . . . . . . . . . . . . . . . . 12

2.2 Visualization . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

**3 MODELS/METHODS** **15**

3.1 USES OF STATISTICAL MODELS{#statMods}. . . . . . . . . 15

**4 REMARKS** **19**

# Prerequisites

## 0.1 Packages.

Install the following packages to follow the book easily.

```r
#install.packages("bookdown") #to generate the book.
## or the development version
## devtools::install_github("rstudio/bookdown")
#
#install.packages("tidyverse")# a collection of r packages that makes working with data easier.

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
#

#install.packages("rticles")# journal like template

library(rticles)
#
#install.packages()
#
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

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

```
library(RColorBrewer)
```

### 0.1.1 Bibliography

automatically create a bib database for R packages

```
knitr::write_bib(c(
  .packages(), 'bookdown', 'knitr', 'rmarkdown', 'tidyverse'
), 'packages.bib')
```

# Chapter 1

# Introduction

## 1.1 THE CHALLLENGER.

If you have ever been in a data analytics class I bet you have ever heard of Edward Tufte. On his critisim on the engineers' failures to make sense of the concerns the had on lunching The Challenger on January 28, 1986.

> Being Right isn't enough, you have to be convincing.
>
> ~ Edward Tufte

This philosophy formed the basis of his arguments in his book *Visual Explanations and Quantities, Evidence and Narrative*

> An essential analytic task in making decisions based on evidence is to understand how things work—mechanism, trade-offs, process and dynamics, cause and effect. That is, intervention-thinking and policy-thinking demand causality-thinking

> Making decisions based on evidence requires the appropriate display of that evidence. Good displays of data help to reveal knowledge relevant to understanding mechanism, process and dynamics, cause and effect. That is, displays of statistical data should directly serve the analytic task at hand.

Hence we are going set our reports right even before we begin in whatever we are supposed to report on.

## 1.2 DATA

Lets take an hypothetical data where in Africa the distance students take to walk to get access to basic education affect their perfomance.

```
set.seed(123)
n <- 30
df <- tibble(
  score=sample(30:99, size = n, replace = T),
  dist=sample(1:8, size=n, replace = T)
)

head(df)#Checking the first six observations
```

```
## # A tibble: 6 x 2
##   score  dist
##   <int> <int>
## 1    60     8
## 2    80     6
## 3    43     1
## 4    96     2
## 5    71     5
## 6    79     5
```

```
### Any time you get data into your hands always check its health status ie, the first
###
### Height & Weight
nrow(df)#number of rows
```

```
## [1] 30
```

```
ncol(df)#number of columns
```

```
## [1] 2
```

```
### Pressure
str(df)#structure
```

```
## tibble [30 x 2] (S3: tbl_df/tbl/data.frame)
##  $ score: int [1:30] 60 80 43 96 71 79 72 43 54 98 ...
##  $ dist : int [1:30] 8 6 1 2 5 5 8 4 7 5 ...
```

```
### BMI
dim(df)
```

```
## [1] 30  2
```

```
colnames(df)
```

```
## [1] "score" "dist"
```

```
###
```

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 1. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

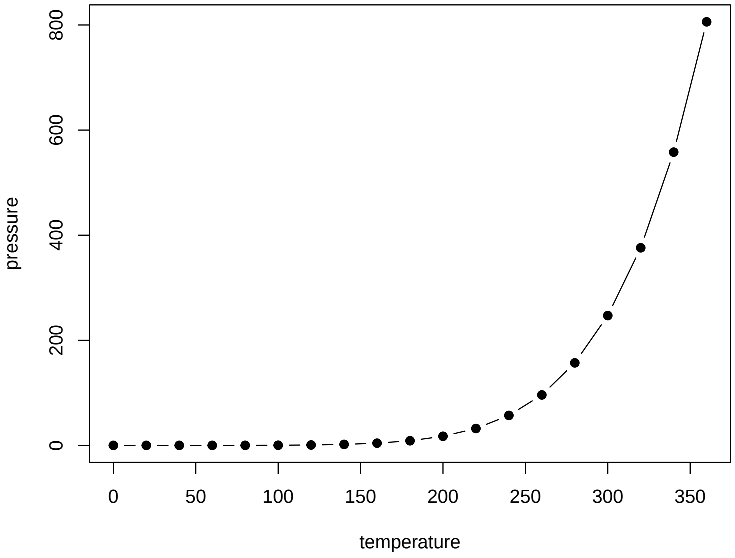

Figure 1.1: Here is a nice figure!

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 1.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 1.1.

Table 1.1: Here is a nice table!

| Sepal.Length | Sepal.Width | Petal.Length | Petal.Width | Species |
|---:|---:|---:|---:|:---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 5.0 | 3.6 | 1.4 | 0.2 | setosa |
| 5.4 | 3.9 | 1.7 | 0.4 | setosa |
| 4.6 | 3.4 | 1.4 | 0.3 | setosa |
| 5.0 | 3.4 | 1.5 | 0.2 | setosa |
| 4.4 | 2.9 | 1.4 | 0.2 | setosa |
| 4.9 | 3.1 | 1.5 | 0.1 | setosa |
| 5.4 | 3.7 | 1.5 | 0.2 | setosa |
| 4.8 | 3.4 | 1.6 | 0.2 | setosa |
| 4.8 | 3.0 | 1.4 | 0.1 | setosa |
| 4.3 | 3.0 | 1.1 | 0.1 | setosa |
| 5.8 | 4.0 | 1.2 | 0.2 | setosa |
| 5.7 | 4.4 | 1.5 | 0.4 | setosa |
| 5.4 | 3.9 | 1.3 | 0.4 | setosa |
| 5.1 | 3.5 | 1.4 | 0.3 | setosa |
| 5.7 | 3.8 | 1.7 | 0.3 | setosa |
| 5.1 | 3.8 | 1.5 | 0.3 | setosa |

```r
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2020) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

# Chapter 2

# EXPLANATORY DATA ANALYSIS

Here we form the foundation of our report. We listen to what the data has for us.This should be the entry step in a data project, where we start by knowing the correct data types and exploring distributions in numerical and categorical variables. Always getting your data into the right format will affect how you work with them hence inferences that they generate.

At this point we perfom dataset *'anatomy'* namely;

```
    *Getting metrics like total rows, columns, data types, zeros, and missing values
    *How each of the previous items impacts on different analysis
    *How to quickly filter and select variables.

    *Univariate analysis in categorical variable:

       +Frequency, percentage, cumulative value, and colorful plots
    *Univariate analysis with numerical variables:
       +Percentile, dispersion, standard deviation, mean, top and bottom values
      +Plotting distributions
```

Based on our 1.2 we can develop the following questions.

```
*Can we divide the perfomance into fail and passed?
*What distance community does one above categories fall into?
```

## 2.1 DATA WRANGLING

```
df$performance <- fifelse(df$score>=40, yes = "passed", no="failed") #answering the fi
df$dist_comm <- fifelse(df$dist<=1, yes = "near",
                        fifelse(df$dist>1 & df$dist<=4,
                                yes="average",
                                no="far")
                        )#answering the second question.
df$dist <- as.factor(df$dist)
df$dist_comm <- as.factor(df$dist_comm)
df$performance <- as.factor(df$performance)
```

## 2.2 Visualization

```
df %>%
  ggplot(aes(dist, score))+
  geom_boxplot()+
  labs(title = "BOXPLOT OF STUDENTS'\nSCORE BASED ON DISTANCE.")
```

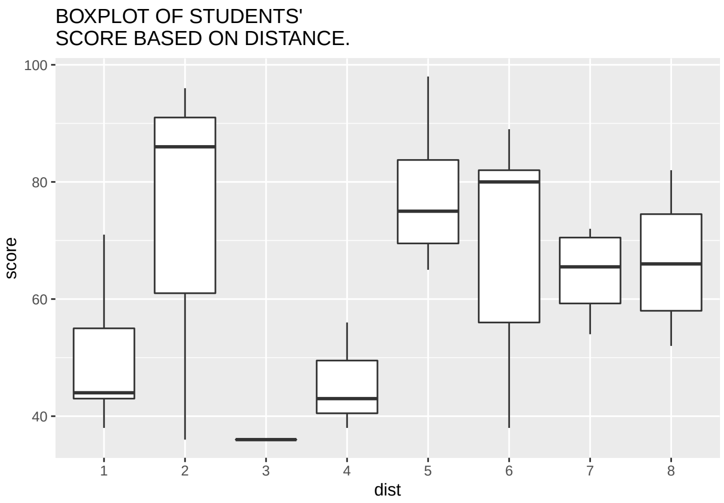

Figure 2.1: boxplot of score based pn distance students walk

Its results above is astonishing, for see students in coming from distance 3-4 perform even better than those who are close to school.

```
df %>%
  ggplot(aes(y=score, fill=performance))+
  geom_bar(position = "stack")+
  facet_wrap(~performance)+
  labs(title = "Bargraph OF STUDENTS'\nSCORE.")
```

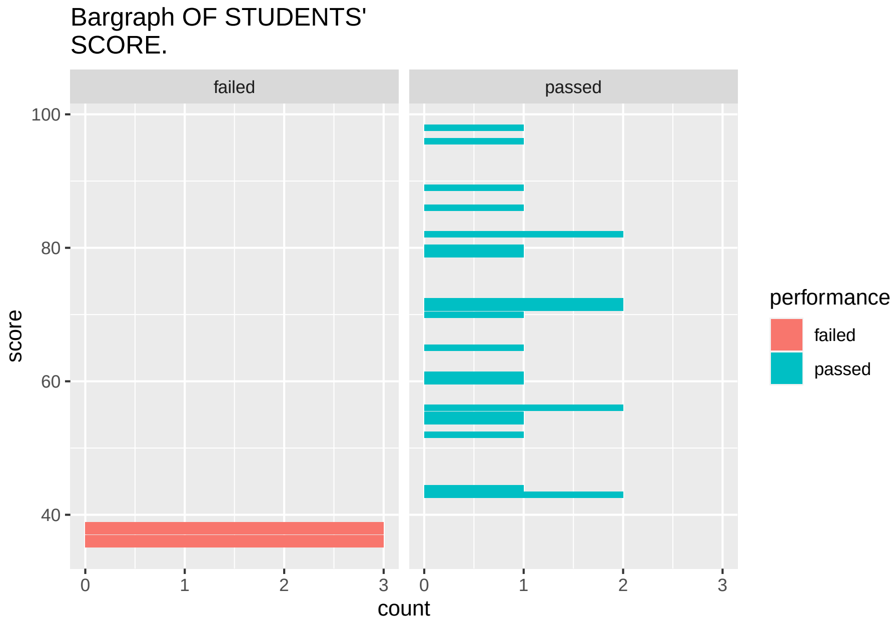

Figure 2.2: Bargraphs

```
df %>%
  ggplot(aes(y=score, fill=performance))+
  geom_bar(position = "identity")+
  facet_wrap(~dist_comm)+
  labs(title = "Bargraph OF STUDENTS'\nSCORE. IN DIFFERENT DISTANCE COMMUNITIES")
```

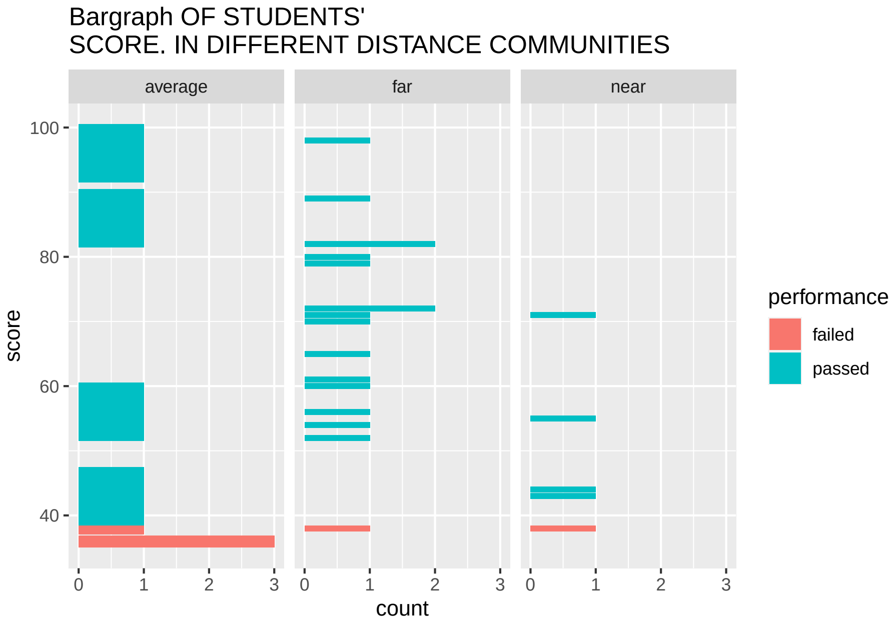

Figure 2.3: Bargraphs

# Chapter 3

# MODELS/METHODS

Now that we have seen some interesting patterns, lets try to get the finnier details of them. Modeling is important for they reveal subtle stories in a data.

> All models are wrong. Some models are useful.

`~Author(Statistical Modelling)`

> Art is a lie that tells the truth!

`~Pablo Picasso`

There are as many named lists of models as sand, yet they all have one thing in common;

> A model is a representative of a specific purpose.

Hence the appropriate form of the model depends on the task at hand and the expertise of the scientist.

## 3.1 USES OF STATISTICAL MODELS{#statMods}[1].

```
*Description models
*Classification or prediction models.
*Anticipating the consequences of interventions models.
```

[1] There are several classification of models eg. Mathematical and Statistical models, how you define a model changes its classification hence affecting its application

Its difficult to use observation to inform concrete knowledge becausse r/ships are complicated and involve multiple factors, do you still recall the mistakes made by the challengers' engineers?1.1.

> Mathematicians("Statisticians") do not study objects, butbthe relations among objetcs.

~Georges Braque

Lets consider our case data 2.1, we want to build a simple model that predicts a students performance based on the distance community(Sometimes variables will have collinearity hence be keen.)

$$Y = \beta_0 + \beta X + e$$

```
mod <- lm(formula =score~dist_comm, data = df)

anova(mod)
```

```
## Analysis of Variance Table
##
## Response: score
##           Df Sum Sq Mean Sq F value  Pr(>F)
## dist_comm  2 2240.1 1120.03  3.5759 0.04191 *
## Residuals 27 8456.9  313.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = score ~ dist_comm, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.471 -13.153  -1.971  10.279  42.625
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     53.375      6.257   8.530 3.82e-09 ***
## dist_commfar    16.096      7.588   2.121   0.0432 *
## dist_commnear   -3.175     10.089  -0.315   0.7554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.7 on 27 degrees of freedom
## Multiple R-squared:  0.2094, Adjusted R-squared:  0.1508
## F-statistic: 3.576 on 2 and 27 DF,  p-value: 0.04191
```

# Chapter 4

# REMARKS

Awesome unlike the engineers you will never have to face the Tuftes' criticisms.

# Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.21.