Calories Burned During Exercise and Activities

igurrola

6/20/2021

1. Introduction

A fitness tracker is pretty likely that along with information on step counts and distance covered, your wearable is also dishing out data on estimated calories burned.The body burns calories through digestion, maintaining basic body functions and physical activity. In the case of most fitness trackers, they're offering a way of calculating energy expenditure from physical activities recorded with your device. Whether that's hitting a spin class, going for a run or just walking down to the shops. Along with following a sensible diet, monitoring this information can be really beneficial in the quest to lose weight. There are a variety of physical activities that produce a specific caloric consumption, depending on the age, gender and current weight will determine how fast the goal of getting fit will be accomplished.

2. Objective

Being an overweight man for a few years now, covid-19 survivor, I found this set of data interesting and motivating to understand the effects of different types of physical tasks in the search of getting fit. The information presented here will be approached with great responsibility and respect, seeking to create a comparative base that seeks to be useful for someone else in their need and affinity in the desire to start a physical activity.

I've been an enthusiast of sports and have practiced running and bike for sometime now, but a goal is to determine if there is a significant difference between the average number of calories burnt from cycling for an hour versus those burnt from running for an hour. I will verify in he data if running in a moderate way will have a positive impact either way.

A Two-Sample T Test was used to determine if there was a significant difference between the calories burnt during these two popular exercise activities.

3. Data preparation

Public dataset can be found in Kaggle as recommended in the Capstone Project. - Calories Burned During Exercise and Activities The author is Aadhav Vignesh, his dataset contains the amount of calories burned during several activities (version 2). It currently contains 248 activities and exercises ranging from running, cycling calisthenics, etc. Repository was compiled manually according with the author.

Data-mining includes 6 columns:

- Activity, Exercise or Sport (1 hour)
- 130 lb
- 155 lb
- 180 lb
- 205 lb

- Calories per lb

Some of the packages of the research were previously loaded but other were required to this specific study. To show the audience and run the program from scratch, packages are included to encourage the viewer to focus mainly in the structure of the code and motivate to extend the current document or try to build a similar or even bigger base with more variables.

```
Loading required package: tidyverse
```

```
Warning: package 'tidyverse' was built under R version 4.0.5
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
v ggplot2 3.3.3     v purrr   0.3.4
v tibble  3.1.2     v dplyr   1.0.6
v tidyr   1.1.3     v stringr 1.4.0
v readr   1.4.0     v forcats 0.5.1
```

```
Warning: package 'ggplot2' was built under R version 4.0.5
```

```
Warning: package 'tidyr' was built under R version 4.0.5
```

```
Warning: package 'dplyr' was built under R version 4.0.5
```

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

```
Loading required package: caret
```

```
Warning: package 'caret' was built under R version 4.0.5
```

```
Loading required package: lattice
```

```
Attaching package: 'caret'
```

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

    lift
```

```
Loading required package: 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

Loading required package: car

Warning: package 'car' was built under R version 4.0.5

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: magrittr

Attaching package: 'magrittr'

The following object is masked from 'package:purrr':

    set_names

The following object is masked from 'package:tidyr':

    extract

Loading required package: curl

Warning: package 'curl' was built under R version 4.0.5

Using libcurl 7.64.1 with Schannel

Attaching package: 'curl'

The following object is masked from 'package:readr':

    parse_date
```

3.1 Exercise database table

The introductory set of information contains as described initially, columns with weight calories for six different cycling activities.

```
  Activity..Exercise.or.Sport..1.hour. X130.lb X155.lb X180.lb X205.lb
1          Cycling, mountain bike, bmx     502     598     695     791
2  Cycling, <10 mph, leisure bicycling     236     281     327     372
3             Cycling, >20 mph, racing     944    1126    1308    1489
4          Cycling, 10-11.9 mph, light     354     422     490     558
5       Cycling, 12-13.9 mph, moderate     472     563     654     745
6       Cycling, 14-15.9 mph, vigorous     590     704     817     931
  Calories.per.kg
1       1.7507297
2       0.8232356
3       3.2949735
4       1.2348534
5       1.6478253
6       2.0594431
```

3.2 Adding Calories burned per category to Data Matrix

The calories burned per specific weight category can be calculated multiplying the calories by the weight section (in this case study categories are 130Lb, 155Lb, 180Lb and 205Lb). Four columns are created and added to the fixed matrix below. Added to the dataset, here are columns with calories burned per weight category.

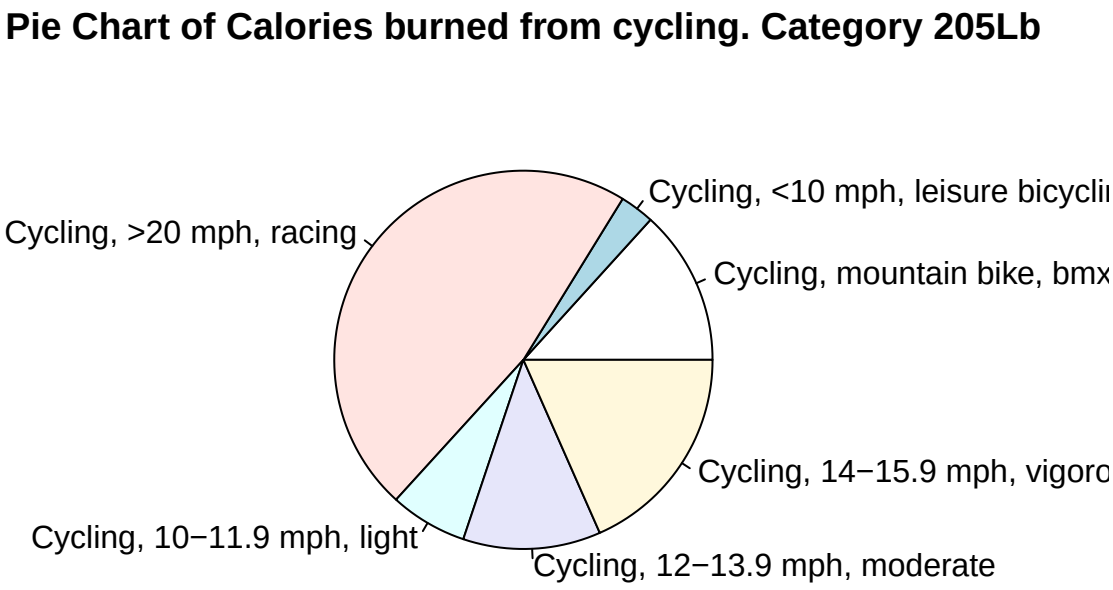
```
  Activity..Exercise.or.Sport..1.hour. X130.lb X155.lb X180.lb X205.lb
1          Cycling, mountain bike, bmx     502     598     695     791
2  Cycling, <10 mph, leisure bicycling     236     281     327     372
3             Cycling, >20 mph, racing     944    1126    1308    1489
4          Cycling, 10-11.9 mph, light     354     422     490     558
5       Cycling, 12-13.9 mph, moderate     472     563     654     745
6       Cycling, 14-15.9 mph, vigorous     590     704     817     931
  Calories.per.kg Cal_burned_130Lb Cal_burned_155Lb Cal_burned_180Lb
1       1.7507297         878.8663        1046.9364        1216.7572
2       0.8232356         194.2836         231.3292         269.1981
3       3.2949735        3110.4550        3710.1402        4309.8254
4       1.2348534         437.1381         521.1082         605.0782
5       1.6478253         777.7735         927.7256        1077.6777
6       2.0594431        1215.0714        1449.8479        1682.5650
  Cal_burned_205Lb
1        1384.8272
2         306.2437
3        4906.2156
4         689.0482
5        1227.6298
6        1917.3415
```

3.3 Compare each group's contribution

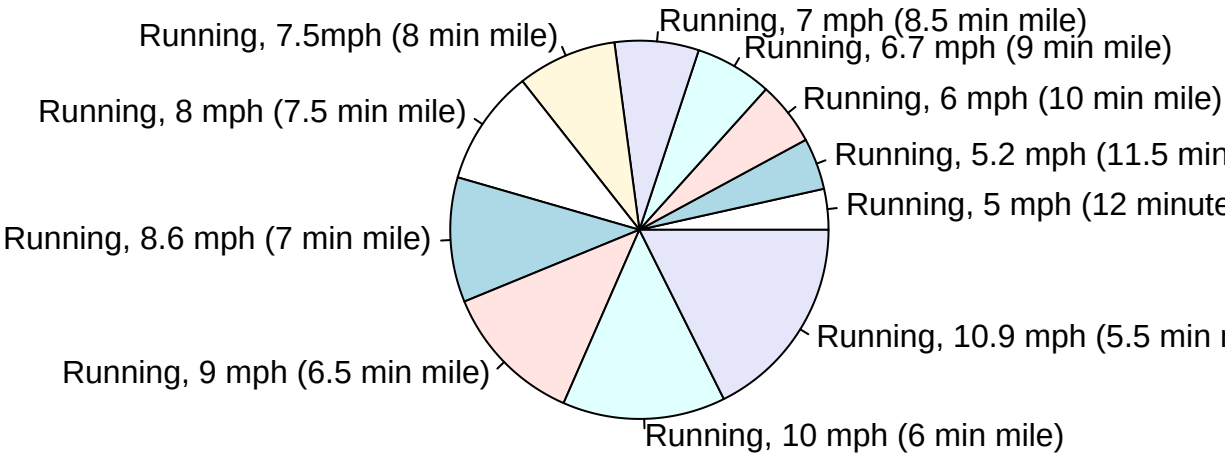
Pie charts are not recommended in the R documentation, and their features are somewhat limited. The authors recommend bar or dot plots over pie charts because people are able to judge length more accurately

than volume. Either way, I feel comfortable presenting the top activities from cycling and running.

The objective in the pie charts presented is to compare each group's contribution to the whole, as opposed to comparing groups to each other.



Since my weight is closer to 205Lb, the selection of data is based on that category. It is shown that Cycling, >20 mph have a higher effect on loosing calories that the other cycling activities. Either way it is better to start from less to more, even tough Cycling <10 mph have the lower effect it can be considered as a great strategy to start increasing until more vigorous speed can be implemented, always guided by a certified trainer and nutritionist.



Moving forward with running category, the chart is presenting the effects of burning calories depending of the running activity. As the previous cycling chart, it is suggested to start increasing until more vigorous speed can be implemented, always guided by a certified trainer and nutritionist. Sooner or later the goal will be accomplished.

3.4 Summarize the distribution of an univariate data set

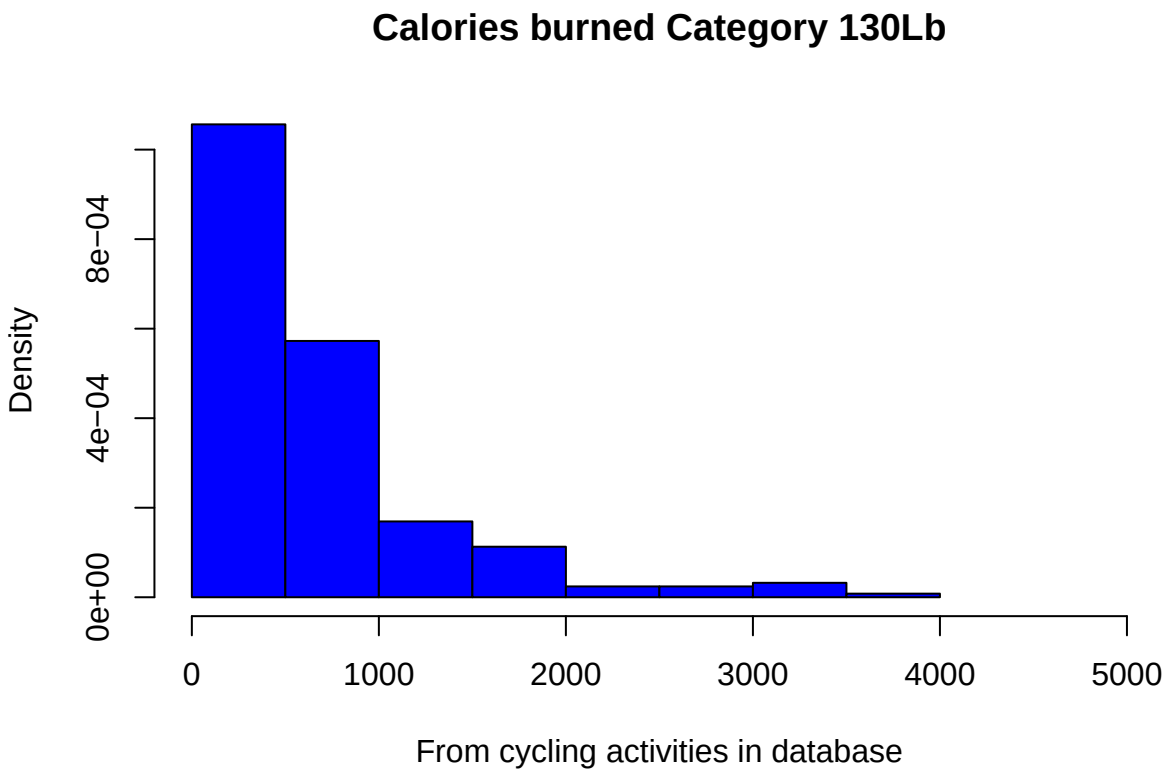
Histogram can be created using the hist() function in R programming language. This function takes in a vector of values for which the histogram is plotted.

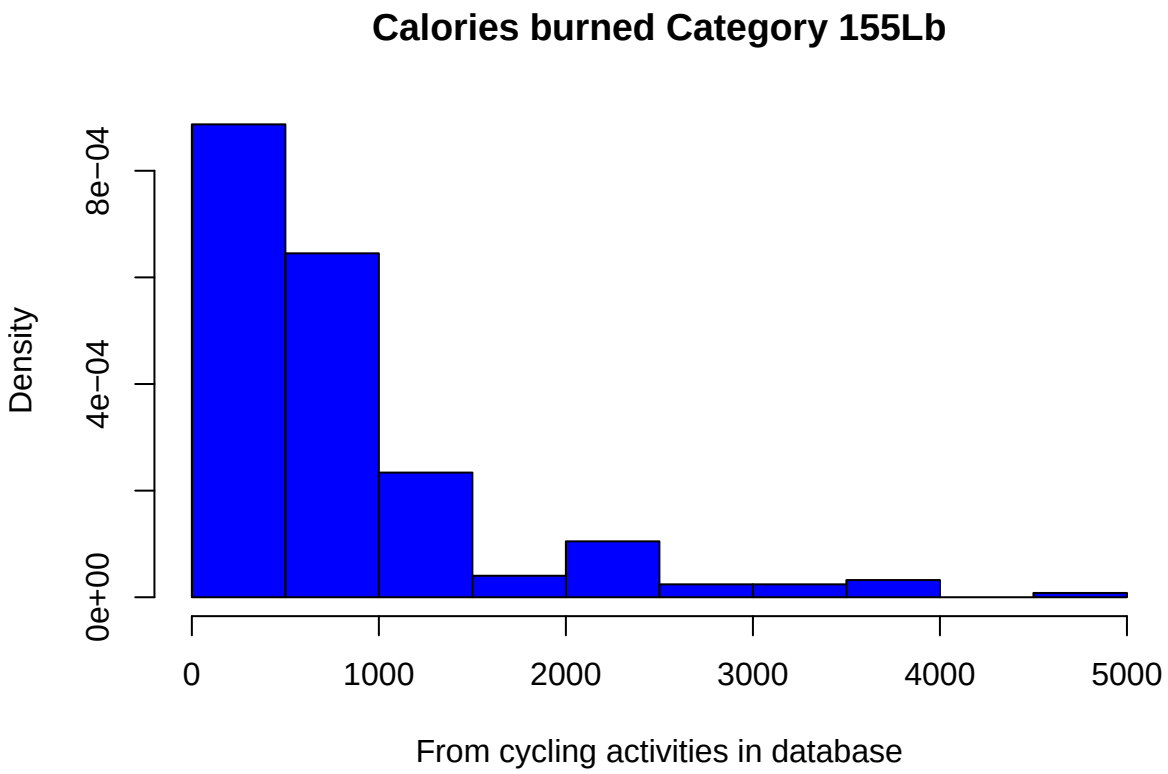
Four histograms are presented to have a better view of the result of each weight category.

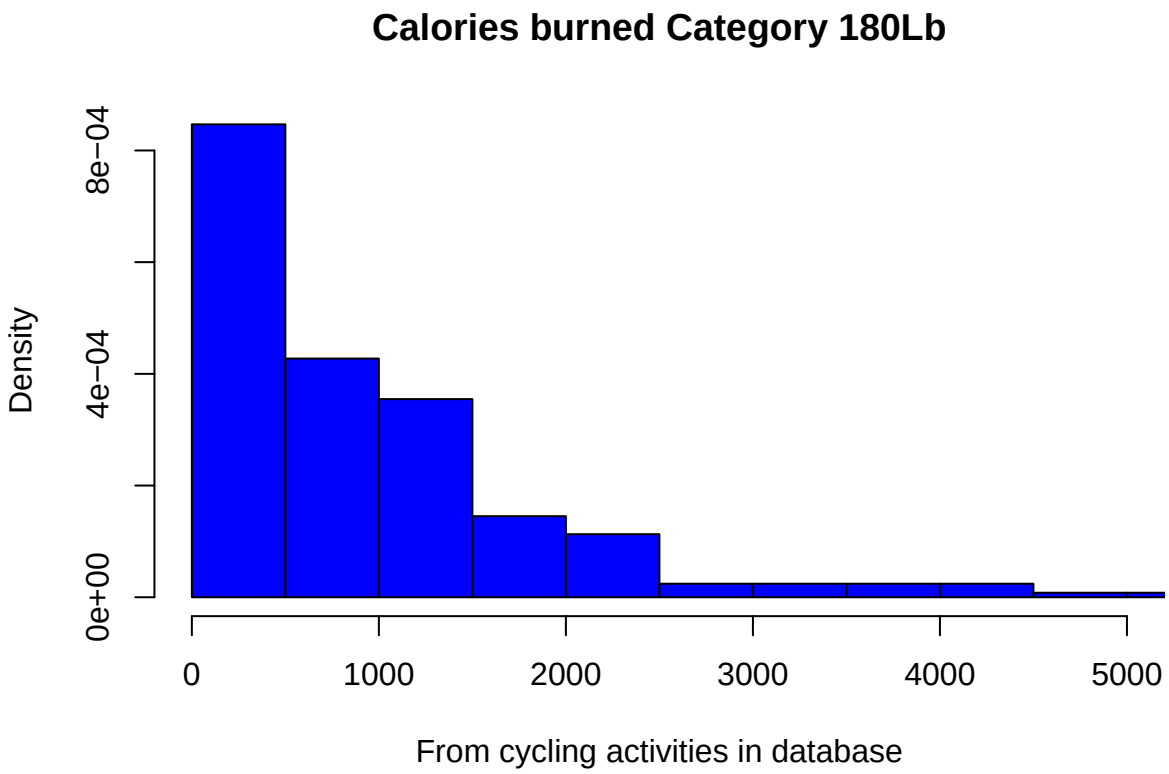
Note that the y axis is labeled density instead of frequency on each plot. In this case, the total area of the histogram is equal to 1.

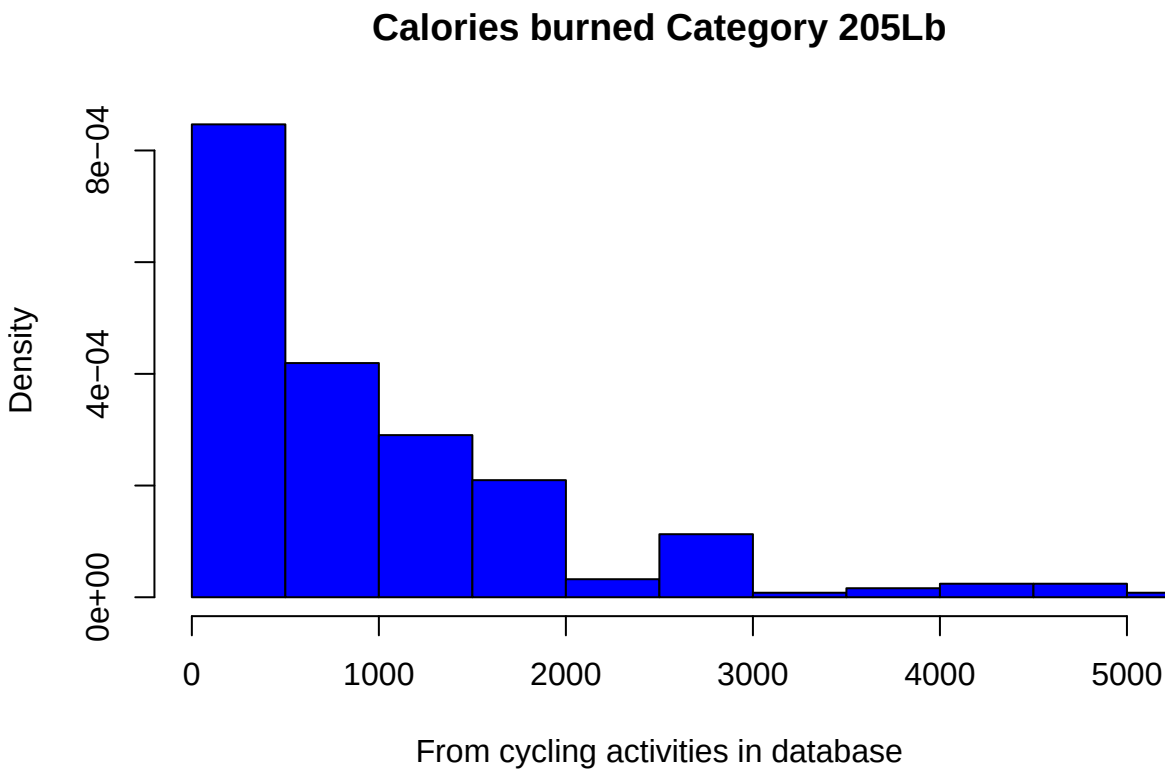
The purpose of a histogram is to graphically summarize the distribution of an univariate data set. The histogram graphically shows the following:

- Center (i.e., the location) of the data
- spread (i.e., the scale) of the data
- skewness of the data
- Presence of outliers
- Presence of multiple modes in the data


Calories burned Category 130Lb
Density
0e+00
4e−04
8e−04
0
1000
2000
3000
4000
5000
From cycling activities in database

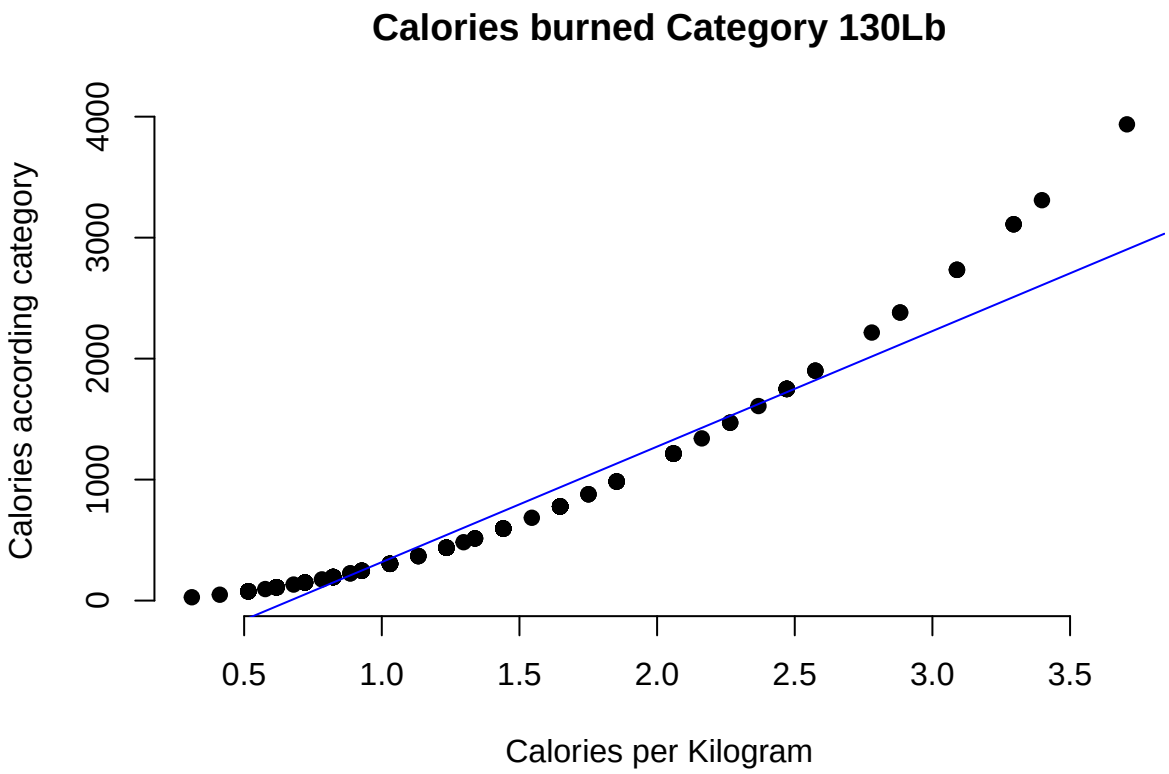

Calories burned Category 155Lb
Density
0e+00
4e−04
8e−04
0
1000
2000
3000
4000
5000
From cycling activities in database

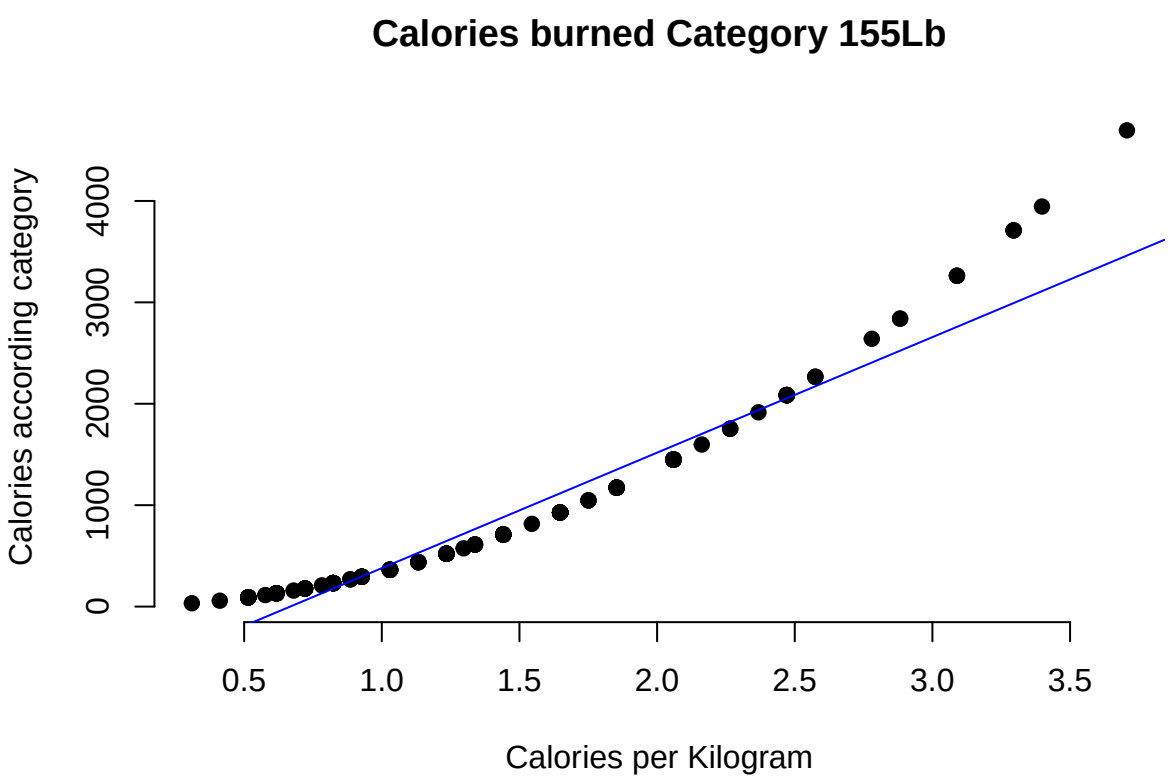

Calories burned Category 180Lb
Density
0e+00
4e−04
8e−04
0
1000
2000
3000
4000
5000
From cycling activities in database

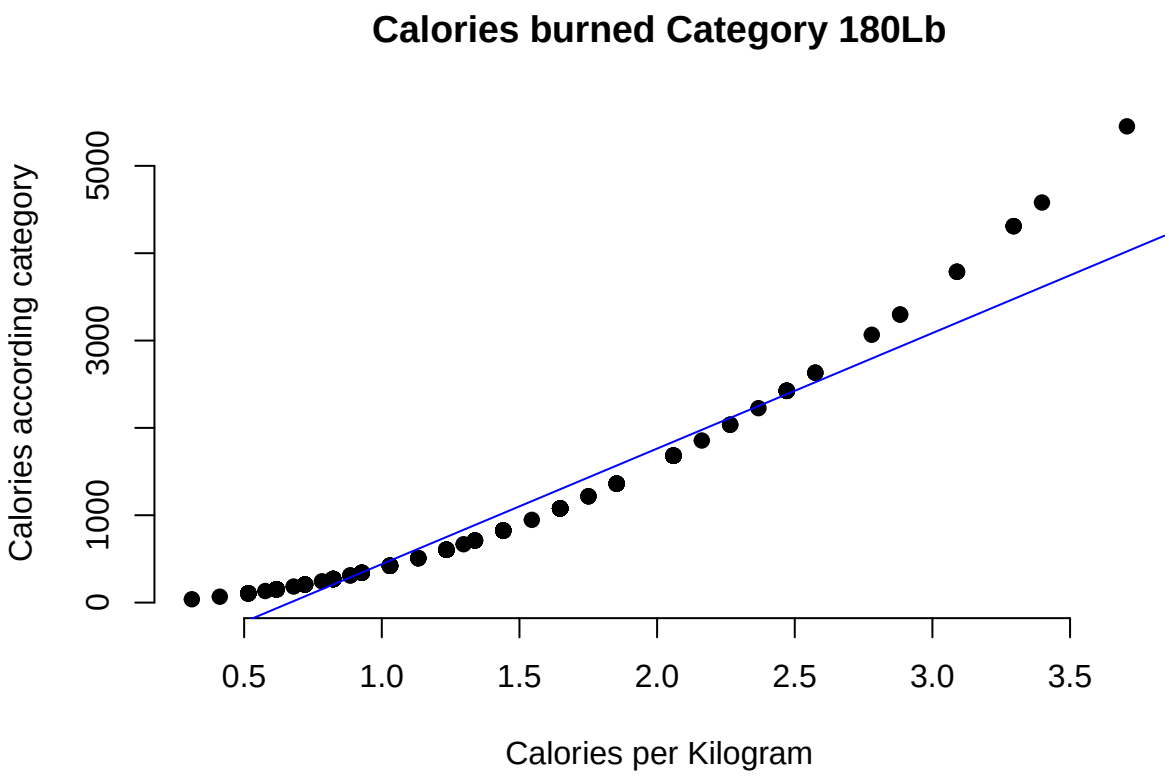


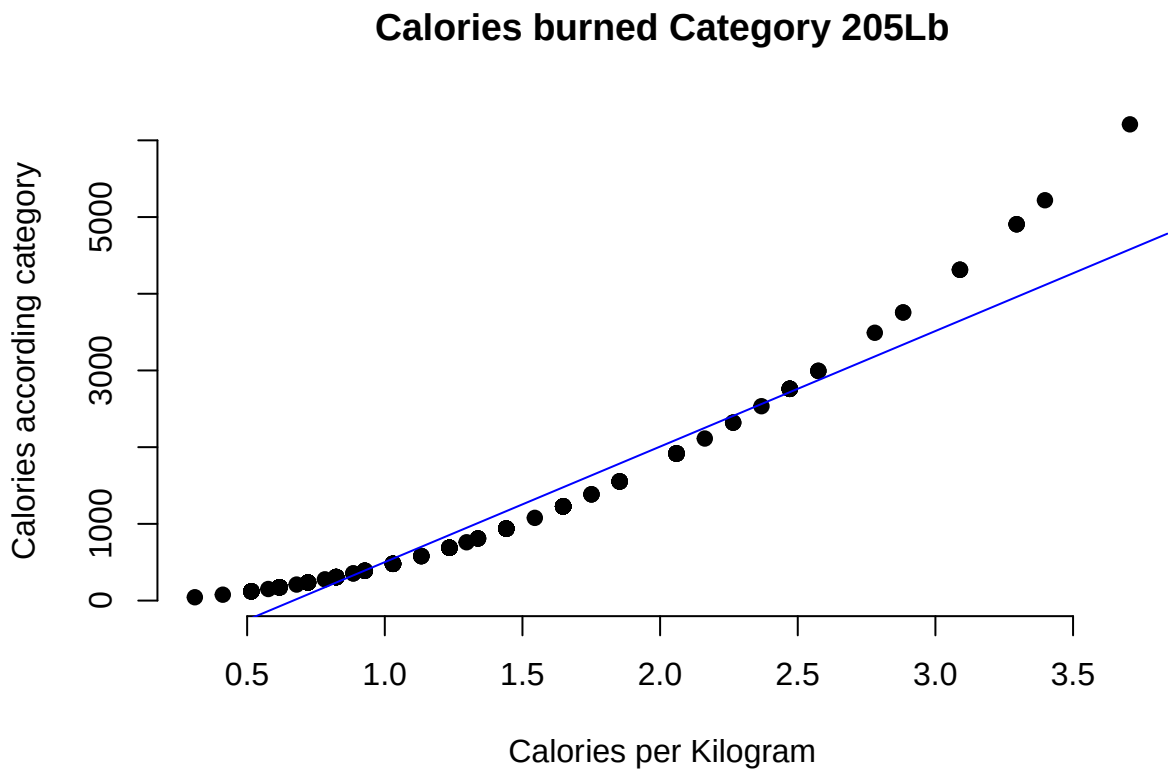
3.5 Scatter plot

Scatter plots help determine lines of best fit between plotted points that do not have perfect correlation with one another. Scatter plots can be used to determine regression equations by plugging in the values to a graphing calculator.


Calories burned Category 130Lb
Calories according category
0
1000
2000
3000
4000
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Calories per Kilogram


Calories burned Category 155Lb
Calories according category
0
1000
2000
3000
4000
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Calories per Kilogram


Calories burned Category 180Lb
Calories according category
0
1000
3000
5000
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Calories per Kilogram



4. Sorting to find the most intensive exercises of the database

Here you can see the most intensive exercises sorted on the basis of the calories burned per hour.

```
    Activity..Exercise.or.Sport..1.hour. X130.lb X155.lb X180.lb X205.lb
48      Running, 10.9 mph (5.5 min mile)    1062    1267    1471    1675
218         Cross country skiing, uphill     974    1161    1348    1536
3               Cycling, >20 mph, racing     944    1126    1308    1489
47          Running, 10 mph (6 min mile)     944    1126    1308    1489
189                    Skin diving, fast     944    1126    1308    1489
46         Running, 9 mph (6.5 min mile)     885    1056    1226    1396
    Calories.per.kg Cal_burned_130Lb Cal_burned_155Lb Cal_burned_180Lb
48         3.706591         3936.400         4696.251         5452.396
218        3.397878         3309.533         3944.936         4580.340
3          3.294974         3110.455         3710.140         4309.825
47         3.294974         3110.455         3710.140         4309.825
189        3.294974         3110.455         3710.140         4309.825
46         3.089165         2733.911         3262.158         3787.316
    Cal_burned_205Lb
48          6208.540
218         5219.141
3           4906.216
47          4906.216
189         4906.216
46          4312.474
```

According with the information provided, the data set explains that the top 5 most intensive exercise activities are:

- Running, 10.9 mph (5.5 min mile)
- Cross country skiing, uphill
- Cycling, >20 mph, racing
- Running, 10 mph (6 min mile)
- Skin diving, fast
- Running, 9 mph (6.5 min mile)

To incorporate all weight categories in the concluded samples, the data was converted into a tidy format using the gather function which combine weight categories into one variable.

```
  Activity..Exercise.or.Sport..1.hour. Calories.per.kg  weight Calories
1          Cycling, mountain bike, bmx       1.7507297 X130.1b      502
2  Cycling, <10 mph, leisure bicycling       0.8232356 X130.1b      236
3             Cycling, >20 mph, racing       3.2949735 X130.1b      944
4          Cycling, 10-11.9 mph, light       1.2348534 X130.1b      354
5       Cycling, 12-13.9 mph, moderate       1.6478253 X130.1b      472
6       Cycling, 14-15.9 mph, vigorous       2.0594431 X130.1b      590
```

Column names were reviewed and one was updated for coherence.

```
[1] "Activity..Exercise.or.Sport..1.hour."
[2] "Calories.per.kg"
[3] "weight"
[4] "Calories"
```

A variable was generated that include the two categories or sample groups that would be tested against each other cycling and running. The string detect function was used to identify any observations that included the words *cycling* or *running*. The function extracted these keywords so the variable included only two categories.

```
                            Activity Calories.per.kg  weight Calories Exercise
1         Cycling, mountain bike, bmx       1.7507297 X130.1b      502  cycling
2 Cycling, <10 mph, leisure bicycling       0.8232356 X130.1b      236  cycling
3            Cycling, >20 mph, racing       3.2949735 X130.1b      944  cycling
4         Cycling, 10-11.9 mph, light       1.2348534 X130.1b      354  cycling
5      Cycling, 12-13.9 mph, moderate       1.6478253 X130.1b      472  cycling
6      Cycling, 14-15.9 mph, vigorous       2.0594431 X130.1b      590  cycling
```

Data-mining was sectioned to contain only two variables, the exercise type and calories burnt. It was filtered to omit observations that didn't involve running or cycling.

```
  Calories Exercise
1      502  cycling
2      236  cycling
3      944  cycling
4      354  cycling
5      472  cycling
6      590  cycling
```

After pre-processing the data, the resulting data set include 2 variables:

Calories: The number of calories burnt per hour of exercise

Exercise: The type of exercise

Calories is an example of a discrete variable as calories are typically measured as counts and a decimal place is not meaningful. On the other hand, exercise is a categorical variable. To check that R assigned the correct data types to these variables, the data set structure was reviewed.

```
'data.frame':   116 obs. of  2 variables:
 $ Calories: int  502 236 944 354 472 590 708 295 177 325 ...
 $ Exercise: chr  "cycling" "cycling" "cycling" "cycling" ...
```

As anticipated, Calories is an integer and exercise has come up as a character so it was converted to a factor.

```
[1] TRUE
```

5. Descriptive Statistics and Data Visualisation

Data is quickly becoming a defining thing in the business world. It is the lifeblood of every company decision and thus, it defines what companies do. A company which doesn't pay attention to proper statistics can be at a serious disadvantage from companies who do, especially companies that use descriptive statistics and data visualization.

Data has to be good if a business wants to remain relevant and successful in the business world. The first step would be to collect the data, which is quite easy in many ways. Then the gathered information needs to be analyzed and understood.

Descriptive statistics describes data – it summarizes and organizes all of the collected data into something manageable and simple to understand. The descriptions can include the entire data set or just a part of the data set. One of the most important things to know about descriptive data analysis is that it focuses on the data instead of on the implication that can be far reaching and go beyond the represented data.

This is the main difference between inferential statistics and descriptive statistics. Inferential statistics uses complicated calculations to make predictions while descriptive statistics does not. This is just the basic information you need to know about descriptive statistics, but it's worth understanding the basics before we dive in any deeper.

Data Visualization means that you should take the data you have and that you should convert it to a visual form which is simpler to digest and understand. Instead of looking at numbers or spreadsheets, you can get a picture which shows you the information.

Descriptive statistics turn the data into something more understandable than raw data but data visualization goes further than that and creates a visual which quickly tells a story. (A. Halsey, 2019)

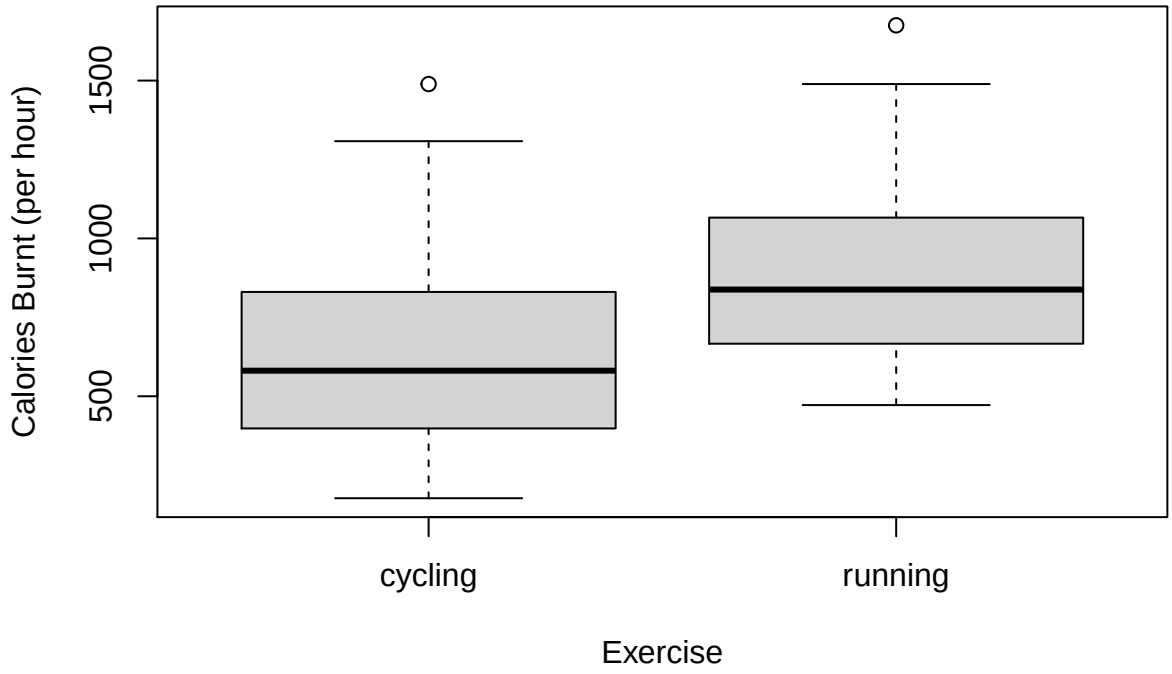
For example, a pie graph shows information much better than a bunch of numbers. And everyone has seen a pie chart many times already. Pie graphs are very simple but they are effective when used properly. But there are also different forms of data visualization like:

- Bar charts
- Line graphs
- Scatter plots
- Diagrams
- Spider charts

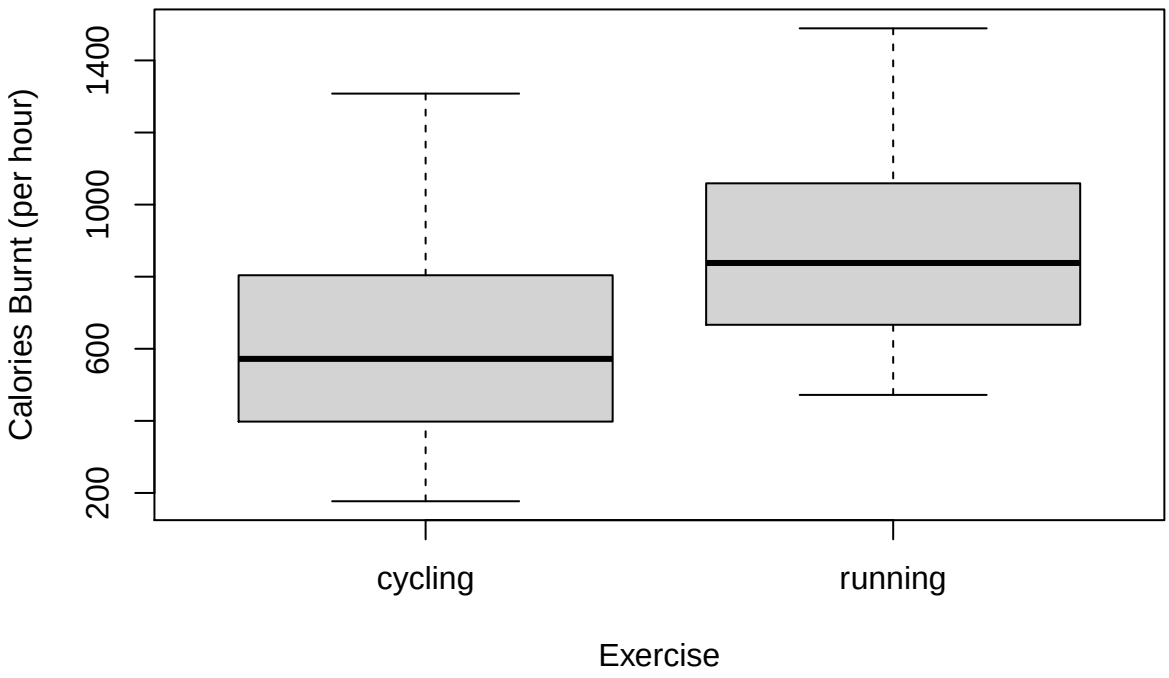
Summary statistics were taken to get a high level overview of the data. The output reveals that running has a larger sample size of 64 while cycling has 52 and there is quite a large different between the sample means **264**. Furthermore, neither samples have any missing values.

```
# A tibble: 2 x 10
  Exercise   Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <fct>    <int> <dbl>  <dbl> <dbl> <int> <dbl> <dbl> <int>   <int>
1 cycling    177  404.    581  824.  1489  632.  302.    52       0
2 running    472  673.    838 1064   1675  896.  282.    64       0
```

To visualize the information described in the summary statistics, box plots were generated. It seems running has a higher mean while cycling seems to have a larger spread. Also of note is that both samples have one outlier each.



Outliers were deleted to stop them from skewing the data and new box plots and summary statistics were created.




```
# A tibble: 2 x 10
  Exercise   Min    Q1 Median    Q3   Max  Mean    SD     n Missing
  <fct>    <int> <dbl>  <int> <dbl> <int> <dbl> <dbl> <int>   <int>
1 cycling    177  398     572   804  1308  615.  279.    51       0
2 running    472  666.    838  1059  1489  884.  267.    63       0
```

6. Hypothesis Testing

A hypothesis is an educated guess about something in the world around you. It should be testable, either by experiment or observation.

$Z = \frac{\hat{P}-p}{\sqrt{pq/n}}$

Hypothesis testing in statistics is a way to test the results of a survey or experiment to see if you have meaningful results. It is basically testing whether the results are valid by figuring out the odds that the results have happened by chance. If your results may have happened by chance, the experiment won't be repeatable and so has little use.

Hypothesis testing can be confusing, mostly because before you can even perform a test, you have to know what your null hypothesis is. Often, those tricky word problems that you are faced with can be difficult to decipher. All you need to do is:

- Figure out your null hypothesis
- State your null hypothesis
- Choose what kind of test you need to perform

- Either support or reject the null hypothesis

For the Independent Sample T Test the null hypothesis states that there is no difference between the average amount of calories burnt while cycling and running for an hour, that is, the difference is = 0. The alternate hypothesis states that there is a difference between the average amount of calories burnt while cycling and running for an hour.

$H_o : \mu_1 - \mu_2 = 0$ $H_A : \mu_1 - \mu_2 \neq 0$

Before applying the Two-Sample T Test the following assumptions must be met:

A. The variables are independent Cycling and running are independent because the distribution of one variable does not affect the other.

B. Data is normally distributed Both samples have n>30 so normal distribution has been assumed

C. Homogeneity of variance The Levene's test was used to compare the variances of cycling and running calories burnt

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   1  0.1357 0.7133
      112
```

The p-value for the Levene's test of equal variance between the two samples of running and cycling was p=0.71. Since p>.05, we fail to reject H0 and equal variance can be assumed.

The Two-sample T Test was then performed using the var.equal = TRUE and alternative = "two.sided" arguments since equal variance is assumed and it's a two-sided hypothesis test.

```
    Two Sample t-test

data:  Calories by Exercise
t = -5.2402, df = 112, p-value = 7.621e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -370.1807 -167.0481
sample estimates:
mean in group cycling mean in group running
             614.9412              883.5556
```

The T critical value was then calculated using $\alpha = 0.05$ and where degrees of freedom $\nu = nsample1 + nsample2 - 2$.

```
[1] -1.981372
```

As the $T - test$ test statistic, $t = -5.24$ was more extreme than -1.98, the critical value method has given a statistically significant result which means that H_o should be rejected.

Furthermore, the $p - value$ of the Two-Sample $T - test$ gives the probability of seeing a difference between the sample means of 269, or one more extreme, assuming that H_o is true. As the $p - value$ was $p < 0.001$ the decision should also be to reject the null hypothesis.

This means that there is evidence to support the alternate hypothesis - that there is a difference between calories burnt while running versus while cycling for an hour.

7. Future Work

The investigation compares average calories burnt over a range of cycling and running speeds and environments. A future work will be more useful if:

- Environmental factors such as humidity and temperature
- Consideration of similar quantity of men and women to avoid an unbiased comparison
- Adding columns for gender and age
- Speed control and contributing calorie-burning factors like incline
- Increasing the sample size so a wide range of weights are recorded, not only four weight categories as this dataset
- Combination of different sports for the training consider using combination of muscles for better results
- Receive feedback from coaches or experts in sports behavior and conduct

8. Conclusion

The results of this analysis suggest that the average number of calories burnt from an hour of running is significantly more than those burnt while cycling. While on a planned diet, it can be closer to a daily calories target, an hour jog will push you over the line faster than an hour of cycling.

For exercising, it is better to start from less to more, even tough Cycling <10 mph have the lower effect it can be considered as a great strategy to start increasing until more vigorous speed can be implemented, always guided by a certified trainer and nutritionist.

References

1. Rafael A. Irizarry (2019). Introduction to Data Science
2. Ander Fernandez Jauregui (2021). How to Code a recommendation System in R
3. Leah Wasser, NEON Data Skills. How to use R Markdown Code Chunks
4. R Markdown Syntax: Hyperlinks, Images & Tables
5. Yihui Xie (2005-2020).Chunk options and package options
6. Ashley Halsey (2019).Descriptive Statistics and Data Visualization