***Google Data Analytics***

***Capstone Project***

**Analysis of Unemployment Rate in India during Pre and Post Covid Pandemic**

# Table of Contents

* [Introduction](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#Introduction)
* [Ask](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#ask)
  + [Guiding questions](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#ask-guiding)
  + [Key tasks](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#ask-key)
  + [Deliverable](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#ask-deliverable)
* [Prepare](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#prepare)
  + [Guiding questions](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#prepare-guiding)
  + [Key tasks](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#prepare-key)
  + [Deliverable](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#prepare-deliverable)
* [Process](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process)
  + [Code](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-code)
    - [Dependences](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-code-dependences)
    - [Concatenating](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-code-concatenating)
    - [Data cleaning](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-code-data)
    - [Manipulating the data](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-code-manipulating)
    - [Saving the result as a CSV](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-code-saving)
  + [Guiding questions](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-guiding)
  + [Key tasks](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-key)
  + [Deliverable](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#process-deliverable)
* [Analyze](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#analyze)
  + [Code](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#analyze-code)
    - [Data distribution](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook" \l "analyze-code-data" \t "_self)
    - [Other variables](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook" \l "analyze-code-other" \t "_self)
  + [Guiding questions](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#analyze-guiding)
  + [Key tasks](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#analyze-key)
  + [Deliverable](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#analyze-deliverable)
* [Share](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#share)
  + [Guiding questions](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#share-guiding)
  + [Key tasks](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#share-key)
  + [Deliverable](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#share-deliverable)
* [Act](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#act)
  + [Guiding questions](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#act-guiding)
  + [Key tasks](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#act-key)
  + [Deliverable](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#act-deliverable)
* [Conclusion](https://www.kaggle.com/code/jhelisonuchoa/google-data-analytics-capstone-case-study-1/notebook#conclusion)

# Introduction

This is my version of the Google Data Analytics Capstone - Case Study Track 2. The full document to the case study can be found in the [Google Data Analytics Capstone: Complete a Case Study](https://www.coursera.org/learn/google-data-analytics-capstone) course.

For this project this steps will be followed to ensure its completion:

* It will follow the 6 steps of the data analysis process:

1. **Ask**
2. **Prepare**
3. **Process**
4. **Analyse**
5. **Share**
6. **Act.**

* Each step will follow its own roadmap with:
  + Code, if needed on the step.
  + Guiding questions, with answers.
  + Key tasks, as a checklist.
  + Deliverable, as a checklist.

# Ask

For the **Ask** step, first we have to understand the context behind

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company’s future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes dierently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Characters and teams

Cyclistic: A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also oering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can’t use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use the assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

Lily Moreno: The director of marketing and your manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.

Cyclistic marketing analytics team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. You joined this team six months ago and have been busy learning about Cyclistic’s mission and business goals — as well as how you, as a junior data analyst, can help Cyclistic achieve them.

Cyclistic executive team: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

## **Guiding questions**

* **What is the problem you are trying to solve?**

The main objective is to understand the state of Unemployment in Indian States before Covid-19 Pandemic, understand how this pandemic affected Unemployment in India States and how much the conditions have improved post-pandemic.

* **How can your insights drive business decisions?**

The insights will help us to understand the effects of Pandemic and whether necessary steps have been taken to improve the situation.

## **Key tasks**

* Identify the task
* Consider key stakeholders

# Prepare

This project will use the data on [Unemployment Rate in India](https://unemploymentinindia.cmie.com/kommon/bin/sr.php?kall=wsttimeseries&index_code=050050000000&dtype=total) provided by the Centre for Monitoring Indian Economy Pvt. Ltd. ([CMIE](https://www.cmie.com/)).

## **Guiding questions**

* **Where is your data located?**

The data is located in CMIE website.

* **How is the data organized?**

The data is separated monthly from 2016 to 2022,

* **Are there issues with bias or credibility in this data? Does your data ROCCC?**

Bias isn't a problem,The full sample of 178,677 households is surveyed over a four month period. Therefore, the most comprehensive measure of unemployment can be produced at the end of execution of the survey over all these households. The sample size of 44,600 household per month is well distributed over the country to enable the estimation of a monthly unemployment rate. And have full credibility for the same reason. And finally, it's ROCCC because it's reliable, original, comprehensive, current and cited.

* **How are you addressing licensing, privacy, security, and accessibility?**

The company has their own licence over the dataset. Besides that, the dataset doesn't have any personal information.

* **How did you verify the data’s integrity?**

All the files have consistent columns and each column has the correct type of data.

* **How does it help you answer your question?**

It gives us an idea of how the unemployment rate varies every month in each Indian State.

* **Are there any problems with the data?**

This dataset doesn't include the unemployment rate from Nagaland and the Union Territories except for Jammu and Kashmir.

## **Key tasks**

* Download data and store it appropriately.
* Identify how it’s organized.
* Sort and filter the data.
* Determine the credibility of the data.

## **Deliverable**

* A description of all data sources used

The primary data used for this analysis is Unemployment Rate of India from Sept 2018 to Sept 2022 provided by CMIE.

# Process

This step will prepare the data for analysis. All the csv files will be merged into one file to improve workflow

## **Code**

### **Dependences**

The main dependencie for the project will be tidyverse.

In [1]:

library(tidyverse)

── **Attaching packages** ─────────────────────────────────────── tidyverse 1.3.0 ──

✔ ggplot2 3.3.3 ✔ purrr 0.3.4

✔ tibble 3.1.1 ✔ dplyr 1.0.5

✔ tidyr 1.1.3 ✔ stringr 1.4.0

✔ readr 1.4.0 ✔ forcats 0.5.0

── **Conflicts** ────────────────────────────────────────── tidyverse\_conflicts() ──

✖ dplyr::filter() masks stats::filter()

✖ dplyr::lag() masks stats::lag()

### **Concatenating**

All the csvs files will be concatenated into one dataframe.

In [2]:

csv\_files <- list.files(path = "../input", recursive = TRUE, full.names=TRUE)

cyclistic\_merged <- do.call(rbind, lapply(csv\_files, read.csv))

In [3]:

head(cyclistic\_merged)

| A data.frame: 6 × 13 | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ride\_id | rideable\_type | started\_at | ended\_at | start\_station\_name | start\_station\_id | end\_station\_name | end\_station\_id | start\_lat | start\_lng | end\_lat | end\_lng | member\_casual |
|  | <fct> | <fct> | <fct> | <fct> | <fct> | <chr> | <fct> | <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <fct> |
| 1 | A847FADBBC638E45 | docked\_bike | 2020-04-26 17:45:14 | 2020-04-26 18:12:03 | Eckhart Park | 86 | Lincoln Ave & Diversey Pkwy | 152 | 41.8964 | -87.6610 | 41.9322 | -87.6586 | member |
| 2 | 5405B80E996FF60D | docked\_bike | 2020-04-17 17:08:54 | 2020-04-17 17:17:03 | Drake Ave & Fullerton Ave | 503 | Kosciuszko Park | 499 | 41.9244 | -87.7154 | 41.9306 | -87.7238 | member |
| 3 | 5DD24A79A4E006F4 | docked\_bike | 2020-04-01 17:54:13 | 2020-04-01 18:08:36 | McClurg Ct & Erie St | 142 | Indiana Ave & Roosevelt Rd | 255 | 41.8945 | -87.6179 | 41.8679 | -87.6230 | member |
| 4 | 2A59BBDF5CDBA725 | docked\_bike | 2020-04-07 12:50:19 | 2020-04-07 13:02:31 | California Ave & Division St | 216 | Wood St & Augusta Blvd | 657 | 41.9030 | -87.6975 | 41.8992 | -87.6722 | member |
| 5 | 27AD306C119C6158 | docked\_bike | 2020-04-18 10:22:59 | 2020-04-18 11:15:54 | Rush St & Hubbard St | 125 | Sheridan Rd & Lawrence Ave | 323 | 41.8902 | -87.6262 | 41.9695 | -87.6547 | casual |
| 6 | 356216E875132F61 | docked\_bike | 2020-04-30 17:55:47 | 2020-04-30 18:01:11 | Mies van der Rohe Way & Chicago Ave | 173 | Streeter Dr & Grand Ave | 35 | 41.8969 | -87.6217 | 41.8923 | -87.6120 | member |

### **Data cleaning**

#### **Removing Duplicates**

In [4]:

cyclistic\_no\_dups <- cyclistic\_merged[!duplicated(cyclistic\_merged$ride\_id), ]

print(paste("Removed", nrow(cyclistic\_merged) - nrow(cyclistic\_no\_dups), "duplicated rows"))

[1] "Removed 209 duplicated rows"

#### **Parsing datetime columns**

In [5]:

cyclistic\_no\_dups$started\_at <- as.POSIXct(cyclistic\_no\_dups$started\_at, "%Y-%m-%d %H:%M:%S")

cyclistic\_no\_dups$ended\_at <- as.POSIXct(cyclistic\_no\_dups$ended\_at, "%Y-%m-%d %H:%M:%S")

### **Manipulating the data**

New columns will help improve calculation time in the future

#### **ride\_time\_m**

Represents the total time of a bike ride, in minutes

In [6]:

cyclistic\_no\_dups <- cyclistic\_no\_dups %>%

mutate(ride\_time\_m = as.numeric(cyclistic\_no\_dups$ended\_at - cyclistic\_no\_dups$started\_at) / 60)

summary(cyclistic\_no\_dups$ride\_time\_m)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-29049.97 7.88 14.52 26.48 26.63 58720.03

#### **year\_month**

Separate the year and the month into one column

In [7]:

cyclistic\_no\_dups <- cyclistic\_no\_dups %>%

mutate(year\_month = paste(strftime(cyclistic\_no\_dups$started\_at, "%Y"),

"-",

strftime(cyclistic\_no\_dups$started\_at, "%m"),

paste("(",strftime(cyclistic\_no\_dups$started\_at, "%b"), ")", sep="")))

unique(cyclistic\_no\_dups$year\_month)

1. '2020 - 04 (Apr)'
2. '2020 - 05 (May)'
3. '2020 - 06 (Jun)'
4. '2020 - 07 (Jul)'
5. '2020 - 08 (Aug)'
6. '2020 - 09 (Sep)'
7. '2020 - 10 (Oct)'
8. '2020 - 11 (Nov)'
9. '2020 - 12 (Dec)'
10. '2021 - 01 (Jan)'
11. '2021 - 02 (Feb)'
12. '2021 - 03 (Mar)'

#### **weekday**

The weekday will be useful to determine patterns of travels in the week

In [8]:

cyclistic\_no\_dups <- cyclistic\_no\_dups %>%

mutate(weekday = paste(strftime(cyclistic\_no\_dups$ended\_at, "%u"), "-", strftime(cyclistic\_no\_dups$ended\_at, "%a")))

unique(cyclistic\_no\_dups$weekday)

1. '7 - Sun'
2. '5 - Fri'
3. '3 - Wed'
4. '2 - Tue'
5. '6 - Sat'
6. '4 - Thu'
7. '1 - Mon'

#### **start\_hour**

Getting the hour of the day also may be useful for intra day analysis

In [9]:

cyclistic\_no\_dups <- cyclistic\_no\_dups %>%

mutate(start\_hour = strftime(cyclistic\_no\_dups$ended\_at, "%H"))

unique(cyclistic\_no\_dups$start\_hour)

1. '18'
2. '17'
3. '13'
4. '11'
5. '14'
6. '10'
7. '15'
8. '16'
9. '03'
10. '12'
11. '09'
12. '08'
13. '20'
14. '19'
15. '21'
16. '23'
17. '07'
18. '22'
19. '06'
20. '01'
21. '05'
22. '00'
23. '02'
24. '04'

### **Saving the result as a CSV**

In [10]:

cyclistic\_no\_dups %>%

write.csv("cyclistic\_clean.csv")

## **Guiding questions**

* **What tools are you choosing and why?**

I'm using R for this project, for two main reasons: Because of the large dataset and to gather experience with the language.

* **Have you ensured your data’s integrity?**

Yes, the data is consistent throughout the columns.

* **What steps have you taken to ensure that your data is clean?**

First the duplicated values where removed, then the columns where formatted to their correct format.

* **How can you verify that your data is clean and ready to analyze?**

It can be verified by this notebook.

* **Have you documented your cleaning process so you can review and share those results?**

Yes, it's all documented in this R notebook.

## **Key tasks**

* [x] Check the data for errors.
* [x] Choose your tools.
* [x] Transform the data so you can work with it eectively
* [x] Document the cleaning process.

## **Deliverable**

* [x] Documentation of any cleaning or manipulation of data

# Analyze

The data exploration will consist of building a profile for annual members and how they differ from casual riders.

Putting in a new variable with a simpler name will help reduce some typing in the future.

## **Code**

In [11]:

*# This function help to resize the plots*

fig <- function(width, heigth){options(repr.plot.width = width, repr.plot.height = heigth)}

In [12]:

cyclistic <- cyclistic\_no\_dups

head(cyclistic)

| A data.frame: 6 × 17 | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ride\_id | rideable\_type | started\_at | ended\_at | start\_station\_name | start\_station\_id | end\_station\_name | end\_station\_id | start\_lat | start\_lng | end\_lat | end\_lng | member\_casual | ride\_time\_m | year\_month | weekday | start\_hour |
|  | <fct> | <fct> | <dttm> | <dttm> | <fct> | <chr> | <fct> | <chr> | <dbl> | <dbl> | <dbl> | <dbl> | <fct> | <dbl> | <chr> | <chr> | <chr> |
| 1 | A847FADBBC638E45 | docked\_bike | 2020-04-26 17:45:14 | 2020-04-26 18:12:03 | Eckhart Park | 86 | Lincoln Ave & Diversey Pkwy | 152 | 41.8964 | -87.6610 | 41.9322 | -87.6586 | member | 26.81667 | 2020 - 04 (Apr) | 7 - Sun | 18 |
| 2 | 5405B80E996FF60D | docked\_bike | 2020-04-17 17:08:54 | 2020-04-17 17:17:03 | Drake Ave & Fullerton Ave | 503 | Kosciuszko Park | 499 | 41.9244 | -87.7154 | 41.9306 | -87.7238 | member | 8.15000 | 2020 - 04 (Apr) | 5 - Fri | 17 |
| 3 | 5DD24A79A4E006F4 | docked\_bike | 2020-04-01 17:54:13 | 2020-04-01 18:08:36 | McClurg Ct & Erie St | 142 | Indiana Ave & Roosevelt Rd | 255 | 41.8945 | -87.6179 | 41.8679 | -87.6230 | member | 14.38333 | 2020 - 04 (Apr) | 3 - Wed | 18 |
| 4 | 2A59BBDF5CDBA725 | docked\_bike | 2020-04-07 12:50:19 | 2020-04-07 13:02:31 | California Ave & Division St | 216 | Wood St & Augusta Blvd | 657 | 41.9030 | -87.6975 | 41.8992 | -87.6722 | member | 12.20000 | 2020 - 04 (Apr) | 2 - Tue | 13 |
| 5 | 27AD306C119C6158 | docked\_bike | 2020-04-18 10:22:59 | 2020-04-18 11:15:54 | Rush St & Hubbard St | 125 | Sheridan Rd & Lawrence Ave | 323 | 41.8902 | -87.6262 | 41.9695 | -87.6547 | casual | 52.91667 | 2020 - 04 (Apr) | 6 - Sat | 11 |
| 6 | 356216E875132F61 | docked\_bike | 2020-04-30 17:55:47 | 2020-04-30 18:01:11 | Mies van der Rohe Way & Chicago Ave | 173 | Streeter Dr & Grand Ave | 35 | 41.8969 | -87.6217 | 41.8923 | -87.6120 | member | 5.40000 | 2020 - 04 (Apr) | 4 - Thu | 18 |

To quick start, let's generate a summary of the dataset

In [13]:

summary(cyclistic)

ride\_id rideable\_type

00009D28BA1C43CB: 1 docked\_bike :2558260

000164C1A5EE8E34: 1 electric\_bike: 611406

00016D3A09FD9C54: 1 classic\_bike : 319873

00017CD018F07967: 1

0001C8191EAB7DA8: 1

0001ECEB9AF8C006: 1

(Other) :3489533

started\_at ended\_at

Min. :2020-04-01 00:00:30 Min. :2020-04-01 00:10:45

1st Qu.:2020-07-14 19:36:28 1st Qu.:2020-07-14 20:11:10

Median :2020-08-29 14:47:30 Median :2020-08-29 15:18:24

Mean :2020-09-10 01:13:26 Mean :2020-09-10 01:39:55

3rd Qu.:2020-10-20 18:07:35 3rd Qu.:2020-10-20 18:21:47

Max. :2021-03-31 23:59:08 Max. :2021-04-06 11:00:11

start\_station\_name start\_station\_id

: 122175 Length:3489539

Streeter Dr & Grand Ave : 35363 Class :character

Clark St & Elm St : 32471 Mode :character

Lake Shore Dr & Monroe St : 29918

Theater on the Lake : 29887

Lake Shore Dr & North Blvd: 27126

(Other) :3212599

end\_station\_name end\_station\_id start\_lat

: 143241 Length:3489539 Min. :41.64

Streeter Dr & Grand Ave : 37425 Class :character 1st Qu.:41.88

Clark St & Elm St : 32254 Mode :character Median :41.90

Theater on the Lake : 31281 Mean :41.90

Lake Shore Dr & Monroe St : 29280 3rd Qu.:41.93

Lake Shore Dr & North Blvd: 27611 Max. :42.08

(Other) :3188447

start\_lng end\_lat end\_lng member\_casual

Min. :-87.87 Min. :41.54 Min. :-88.07 casual:1430351

1st Qu.:-87.66 1st Qu.:41.88 1st Qu.:-87.66 member:2059188

Median :-87.64 Median :41.90 Median :-87.64

Mean :-87.64 Mean :41.90 Mean :-87.64

3rd Qu.:-87.63 3rd Qu.:41.93 3rd Qu.:-87.63

Max. :-87.52 Max. :42.16 Max. :-87.44

NA's :4737 NA's :4737

ride\_time\_m year\_month weekday start\_hour

Min. :-29049.97 Length:3489539 Length:3489539 Length:3489539

1st Qu.: 7.88 Class :character Class :character Class :character

Median : 14.52 Mode :character Mode :character Mode :character

Mean : 26.48

3rd Qu.: 26.63

Max. : 58720.03

One thing that immediately catches the attention is ride\_time\_m. This field has negative values, and the biggest value is 58720.03, which is 40 days and 46 hours. This field will be explored further in the document.

### **Data distribution**

Here we want to try to answer the most basic questions about how the data is distributed.

#### **Casuals vs members**

How much of the data is about members and how much is about casuals?

In [14]:

cyclistic %>%

group\_by(member\_casual) %>%

summarise(count = length(ride\_id),

'%' = (length(ride\_id) / nrow(cyclistic)) \* 100)

| A tibble: 2 × 3 | | |
| --- | --- | --- |
| member\_casual | count | % |
| <fct> | <int> | <dbl> |
| casual | 1430351 | 40.98968 |
| member | 2059188 | 59.01032 |

In [15]:

fig(16,8)

ggplot(cyclistic, aes(member\_casual, fill=member\_casual)) +

geom\_bar() +

labs(x="Casuals x Members", title="Chart 01 - Casuals x Members distribution")

As we can see on the member x casual table, members have a bigger proporcion of the dataset, composing ~59%, ~19% bigger than the count of casual riders.

#### **Month**

How much of the data is distributed by month?

In [16]:

cyclistic %>%

group\_by(year\_month) %>%

summarise(count = length(ride\_id),

'%' = (length(ride\_id) / nrow(cyclistic)) \* 100,

'members\_p' = (sum(member\_casual == "member") / length(ride\_id)) \* 100,

'casual\_p' = (sum(member\_casual == "casual") / length(ride\_id)) \* 100,

'Member x Casual Perc Difer' = members\_p - casual\_p)

| A tibble: 12 × 6 | | | | | |
| --- | --- | --- | --- | --- | --- |
| year\_month | count | % | members\_p | casual\_p | Member x Casual Perc Difer |
| <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| 2020 - 04 (Apr) | 84776 | 2.429433 | 72.12890 | 27.87110 | 44.257809 |
| 2020 - 05 (May) | 200274 | 5.739268 | 56.60495 | 43.39505 | 13.209902 |
| 2020 - 06 (Jun) | 343005 | 9.829522 | 54.89337 | 45.10663 | 9.786738 |
| 2020 - 07 (Jul) | 551480 | 15.803807 | 51.16849 | 48.83151 | 2.336984 |
| 2020 - 08 (Aug) | 622361 | 17.835049 | 53.45772 | 46.54228 | 6.915440 |
| 2020 - 09 (Sep) | 532958 | 15.273020 | 56.71479 | 43.28521 | 13.429576 |
| 2020 - 10 (Oct) | 388653 | 11.137660 | 62.68857 | 37.31143 | 25.377136 |
| 2020 - 11 (Nov) | 259716 | 7.442702 | 66.07872 | 33.92128 | 32.157434 |
| 2020 - 12 (Dec) | 131364 | 3.764509 | 77.12082 | 22.87918 | 54.241649 |
| 2021 - 01 (Jan) | 96834 | 2.774980 | 81.29066 | 18.70934 | 62.581325 |
| 2021 - 02 (Feb) | 49622 | 1.422022 | 79.58365 | 20.41635 | 59.167305 |
| 2021 - 03 (Mar) | 228496 | 6.548028 | 63.22343 | 36.77657 | 26.446852 |

In [17]:

cyclistic %>%

ggplot(aes(year\_month, fill=member\_casual)) +

geom\_bar() +

labs(x="Month", title="Chart 02 - Distribution by month") +

coord\_flip()

Some considerations can be taken by this chart:

* There's more data points at the last semester of 2020.
* The month with the biggest count of data points was August with ~18% of the dataset.
* In all months we have more members' rides than casual rides (Maybe because of returning members).
* The difference of proporcion of member x casual is smaller in the last semester of 2020.

The distribution looks cyclical. Let's compare it with climate data for Chicago. The data will be taken by [Climate of Chicago](https://en.wikipedia.org/wiki/Climate_of_Chicago) (Daily mean °C, 1991–2020).

In [18]:

chicago\_mean\_temp <- c(-3.2, -1.2, 4.4, 10.5, 16.6, 22.2, 24.8, 23.9, 19.9, 12.9, 5.8, -0.3)

month <- c("001 - Jan","002 - Feb","003 - Mar","004 - Apr","005 - May","006 - Jun","007 - Jul","008 - Aug","009 - Sep","010 - Oct","011 - Nov","012 - Dec")

data.frame(month, chicago\_mean\_temp) %>%

ggplot(aes(x=month, y=chicago\_mean\_temp)) +

labs(x="Month", y="Mean temperature", title="Chart 02.5 - Mean temperature for Chicago (1991-2020)") +

geom\_col()

The main takeaway is:

* Temperature heavily influence the volume of rides in the month.

#### **Weekday**

How much of the data is distributed by weekday?

In [19]:

cyclistic %>%

group\_by(weekday) %>%

summarise(count = length(ride\_id),

'%' = (length(ride\_id) / nrow(cyclistic)) \* 100,

'members\_p' = (sum(member\_casual == "member") / length(ride\_id)) \* 100,

'casual\_p' = (sum(member\_casual == "casual") / length(ride\_id)) \* 100,

'Member x Casual Perc Difer' = members\_p - casual\_p)

| A tibble: 7 × 6 | | | | | |
| --- | --- | --- | --- | --- | --- |
| weekday | count | % | members\_p | casual\_p | Member x Casual Perc Difer |
| <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 - Mon | 420613 | 12.05354 | 63.73911 | 36.26089 | 27.4782282 |
| 2 - Tue | 431131 | 12.35496 | 66.16898 | 33.83102 | 32.3379669 |
| 3 - Wed | 464879 | 13.32207 | 65.85563 | 34.14437 | 31.7112625 |
| 4 - Thu | 467450 | 13.39575 | 64.44368 | 35.55632 | 28.8873676 |
| 5 - Fri | 513585 | 14.71785 | 59.80140 | 40.19860 | 19.6027921 |
| 6 - Sat | 659637 | 18.90327 | 49.16628 | 50.83372 | -1.6674322 |
| 7 - Sun | 532244 | 15.25256 | 50.16064 | 49.83936 | 0.3212812 |

In [20]:

ggplot(cyclistic, aes(weekday, fill=member\_casual)) +

geom\_bar() +

labs(x="Weekdady", title="Chart 03 - Distribution by weekday") +

coord\_flip()

It's interesting to see:

* The biggest volume of data is on the weekend.
* Saturday has the biggest data points.
* Members may have the biggest volume of data, besides on saturday. On this weekday, casual take place as having most data points.
* Weekends have the biggest volume of casual, starting on friday, a ~20% increase.

#### **Hour of the day**

In [21]:

cyclistic %>%

group\_by(start\_hour) %>%

summarise(count = length(ride\_id),

'%' = (length(ride\_id) / nrow(cyclistic)) \* 100,

'members\_p' = (sum(member\_casual == "member") / length(ride\_id)) \* 100,

'casual\_p' = (sum(member\_casual == "casual") / length(ride\_id)) \* 100,

'member\_casual\_perc\_difer' = members\_p - casual\_p)

| A tibble: 24 × 6 | | | | | |
| --- | --- | --- | --- | --- | --- |
| start\_hour | count | % | members\_p | casual\_p | member\_casual\_perc\_difer |
| <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| 00 | 41924 | 1.2014194 | 33.39853 | 66.60147 | -33.202939 |
| 01 | 26372 | 0.7557445 | 29.46686 | 70.53314 | -41.066282 |
| 02 | 15386 | 0.4409178 | 27.46653 | 72.53347 | -45.066944 |
| 03 | 9038 | 0.2590027 | 27.63886 | 72.36114 | -44.722284 |
| 04 | 7391 | 0.2118045 | 41.25288 | 58.74712 | -17.494250 |
| 05 | 17987 | 0.5154549 | 74.99861 | 25.00139 | 49.997220 |
| 06 | 56915 | 1.6310177 | 81.50751 | 18.49249 | 63.015022 |
| 07 | 106045 | 3.0389401 | 81.58706 | 18.41294 | 63.174124 |
| 08 | 133253 | 3.8186419 | 78.41399 | 21.58601 | 56.827989 |
| 09 | 123699 | 3.5448522 | 72.02403 | 27.97597 | 44.048052 |
| 10 | 136067 | 3.8992830 | 65.53316 | 34.46684 | 31.066313 |
| 11 | 175790 | 5.0376282 | 61.92161 | 38.07839 | 23.843222 |
| 12 | 222341 | 6.3716439 | 60.23585 | 39.76415 | 20.471708 |
| 13 | 235068 | 6.7363626 | 57.68629 | 42.31371 | 15.372573 |
| 14 | 242837 | 6.9589995 | 55.76992 | 44.23008 | 11.539839 |
| 15 | 264763 | 7.5873346 | 55.58518 | 44.41482 | 11.170367 |
| 16 | 298020 | 8.5403831 | 57.83605 | 42.16395 | 15.672103 |
| 17 | 349360 | 10.0116376 | 60.23643 | 39.76357 | 20.472865 |
| 18 | 333746 | 9.5641860 | 59.53480 | 40.46520 | 19.069592 |
| 19 | 254651 | 7.2975542 | 56.86646 | 43.13354 | 13.732913 |
| 20 | 175658 | 5.0338454 | 53.05822 | 46.94178 | 6.116431 |
| 21 | 113455 | 3.2512891 | 48.69331 | 51.30669 | -2.613371 |
| 22 | 84132 | 2.4109775 | 43.85489 | 56.14511 | -12.290211 |
| 23 | 65641 | 1.8810794 | 39.39154 | 60.60846 | -21.216922 |

In [22]:

cyclistic %>%

ggplot(aes(start\_hour, fill=member\_casual)) +

labs(x="Hour of the day", title="Chart 04 - Distribution by hour of the day") +

geom\_bar()

From this chart, we can see:

* There's a bigger volume of bikers in the afternoon.
* We have more members during the morning, mainly in between 5am and 11am
* And more casuals between 11pm and 4am

This chart can be expanded ween seen it divided by day of the week.

In [23]:

cyclistic %>%

ggplot(aes(start\_hour, fill=member\_casual)) +

geom\_bar() +

labs(x="Hour of the day", title="Chart 05 - Distribution by hour of the day divided by weekday") +

facet\_wrap(~ weekday)

There's a clear diferrence between the midweek and weekends. Let's generate charts for this two configurations.

In [24]:

cyclistic %>%

mutate(type\_of\_weekday = ifelse(weekday == '6 - Sat' | weekday == '7 - Sun',

'weekend',

'midweek')) %>%

ggplot(aes(start\_hour, fill=member\_casual)) +

labs(x="Hour of the day", title="Chart 06 - Distribution by hour of the day in the midweek") +

geom\_bar() +

facet\_wrap(~ type\_of\_weekday)

The two plots differs on some key ways:

* While the weekends have a smooth flow of data points, the midweek have a more steep flow of data.
* The count of data points doesn't have much meaning knowing each plot represents a different amount of days.
* There's a big increase of data points in the midween between 6am to 8am. Then it fall a bit.
* Another big increase is from 5pm to 6pm.
* During the weekend we have a bigger flow of casuals between 11am to 6pm.

It's fundamental to question who are the riders who use the bikes during this time of day. We can assume some factors, one is that members may are people who use the bikes during they daily routine activities, like go to work (data points between 5am to 8am in midweek), go back from work (data points between 5pm to 6pm).

#### **Rideable type**

In [25]:

cyclistic %>%

group\_by(rideable\_type) %>%

summarise(count = length(ride\_id),

'%' = (length(ride\_id) / nrow(cyclistic)) \* 100,

'members\_p' = (sum(member\_casual == "member") / length(ride\_id)) \* 100,

'casual\_p' = (sum(member\_casual == "casual") / length(ride\_id)) \* 100,

'member\_casual\_perc\_difer' = members\_p - casual\_p)

| A tibble: 3 × 6 | | | | | |
| --- | --- | --- | --- | --- | --- |
| rideable\_type | count | % | members\_p | casual\_p | member\_casual\_perc\_difer |
| <fct> | <int> | <dbl> | <dbl> | <dbl> | <dbl> |
| docked\_bike | 2558260 | 73.312263 | 56.35479 | 43.64521 | 12.70958 |
| electric\_bike | 611406 | 17.521111 | 60.25685 | 39.74315 | 20.51370 |
| classic\_bike | 319873 | 9.166626 | 77.86590 | 22.13410 | 55.73181 |

In [26]:

ggplot(cyclistic, aes(rideable\_type, fill=member\_casual)) +

labs(x="Rideable type", title="Chart 07 - Distribution of types of bikes") +

geom\_bar() +

coord\_flip()

It's important to note that:

* Docked bikes have the biggest volume of rides, but this can be that the company may have more docked bikes.
* Members have a bigger preference for classic bikes, 56% more.
* Also for electric bikes.

### **Other variables**

Now let's get a look at some variables of the dataset.

#### **ride\_time\_m**

First get some summarized statistic from the dataset

In [27]:

summary(cyclistic$ride\_time\_m)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-29049.97 7.88 14.52 26.48 26.63 58720.03

The min and the max may be a problem to plot some charts. How the ride time of some bikes is a negative value? Maybe there's some malfunction stations return bad dates. Checking the start and end stations doesn't appear to have a problem.

In [28]:

ventiles = quantile(cyclistic$ride\_time\_m, seq(0, 1, by=0.05))

ventiles

**0%**

-29049.9666666667

**5%**

3.1

**10%**

4.51666666666667

**15%**

5.68333333333333

**20%**

6.78333333333333

**25%**

7.88333333333333

**30%**

9.03333333333333

**35%**

10.25

**40%**

11.5333333333333

**45%**

12.95

**50%**

14.5166666666667

**55%**

16.2833333333333

**60%**

18.3

**65%**

20.65

**70%**

23.3933333333349

**75%**

26.6333333333333

**80%**

30.5833333333333

**85%**

36.4

**90%**

46.1

**95%**

73.05

**100%**

58720.0333333333

We can see that:

* The difference between 0% and 100% is 87770.0 minutes.
* The difference between 5% and 95% is 69.95 minutes. Because of that, in the analysis of this variable we are going to use a subset of the dataset without outliners. The subset will contain 95% of the dataset.

In [29]:

cyclistic\_without\_outliners <- cyclistic %>%

filter(ride\_time\_m > as.numeric(ventiles['5%'])) %>%

filter(ride\_time\_m < as.numeric(ventiles['95%']))

print(paste("Removed", nrow(cyclistic) - nrow(cyclistic\_without\_outliners), "rows as outliners" ))

[1] "Removed 350103 rows as outliners"

##### *ride\_time\_m multivariable exploration*

One of the first interactions between the columns and ride\_length is a box plot, with subplots based on the casual\_members column. Also the summarized data.

In [30]:

cyclistic\_without\_outliners %>%

group\_by(member\_casual) %>%

summarise(mean = mean(ride\_time\_m),

'first\_quarter' = as.numeric(quantile(ride\_time\_m, .25)),

'median' = median(ride\_time\_m),

'third\_quarter' = as.numeric(quantile(ride\_time\_m, .75)),

'IR' = third\_quarter - first\_quarter)

| A tibble: 2 × 6 | | | | | |
| --- | --- | --- | --- | --- | --- |
| member\_casual | mean | first\_quarter | median | third\_quarter | IR |
| <fct> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| casual | 23.67794 | 11.3 | 19.16667 | 31.36667 | 20.06667 |
| member | 15.48325 | 7.3 | 12.18333 | 20.66667 | 13.36667 |

In [31]:

ggplot(cyclistic\_without\_outliners, aes(x=member\_casual, y=ride\_time\_m, fill=member\_casual)) +

labs(x="Member x Casual", y="Riding time", title="Chart 08 - Distribution of Riding time for Casual x Member") +

geom\_boxplot()

It's important to note that:

* Casual have more riding time thant members.
* Mean and IQR is also bigger for casual.

Let's see if we can extract more informations when ploting with weekday.

In [32]:

ggplot(cyclistic\_without\_outliners, aes(x=weekday, y=ride\_time\_m, fill=member\_casual)) +

geom\_boxplot() +

facet\_wrap(~ member\_casual) +

labs(x="Weekday", y="Riding time", title="Chart 09 - Distribution of Riding time for day of the week") +

coord\_flip()

* Riding time for members keeps unchanged during the midweek, increasing during weekends
* Casuals follow a more curve distribution, peaking on sundays and valleying on wednesday/thursday.

Lastly, let's do rideable\_type.

In [33]:

ggplot(cyclistic\_without\_outliners, aes(x=rideable\_type, y=ride\_time\_m, fill=member\_casual)) +

geom\_boxplot() +

facet\_wrap(~ member\_casual) +

labs(x="Rideable type", y="Riding time", title="Chart 10 - Distribution of Riding time for rideeable type") +

coord\_flip()

* Electric bikes have less riding time than other bikes, for both members and casuals.
* Docked bikes have more riding time. And for docked bikes, members have more riding time than casuals.

## **Guiding questions**

* **How should you organize your data to perform analysis on it?**

The data has been organized into a single CSV concatenating all the files from the dataset.

* **Has your data been properly formatted?**

Yes, all the columns have their correct data type.

* **What surprises did you discover in the data?**

One of the main surprises is how members differ from casuals when analysed from weekdays. Also that members have less riding time than casual.

* **What trends or relationships did you find in the data?**
  + There are more members than casuals in the dataset.
  + There are more data points in the last semester of 2020.
  + There are more of a difference between the flow of members/casual from midweek to weekends.
  + Members use bikes on schedules that differs from casual.
  + Members have less riding time.
  + Members tend to prefer docked bikes.
* **How will these insights help answer your business questions?**

This insights helps to build a profile for members.

## **Key tasks**

* [x] Aggregate your data so it’s useful and accessible.
* [x] Organize and format your data.
* [x] Perform calculations.
* [x] Identify trends and relationships.

## **Deliverable**

* [x] A summary of your analysis

# Share

The share phase is usually done by building a presentation. But for kaggle, the best representation of the analysis and conclusions is it's own notebook.

Let's go through the main finds and try to arrive at a conclusion.

What we know about the dataset:

* Members have the biggest proportion of the dataset, ~19% bigger thand casuals.
* There's more data points at the last semester of 2020.
* The month with the biggest count of data points was August with ~18% of the dataset.
* In all months we have more members' rides than casual rides.
* The difference of proporcion of member x casual is smaller in the last semester of 2020.
* Temperature heavily influences the volume of rides in the month.
* The biggest volume of data is on the the weekend.
* There's a bigger volume of bikers in the afternoon.

It's possible to notice that the distribution of rides by month is cyclical through years, as seen on chart 02 and it's influenced by the temperature. The remaining question is: Why are there more members than casual? One plausible answer is that members have a bigger need for the bikes than casuals, as can be seen on how there are more members than casuals on cold months.

Besides that, we have more bike rides on the weekends. Maybe because on those days the bikes were utilized for more recreational ways. This even more plausible when knowing that There's a bigger volume of bikers in the afternoon.

Now for how members differs from casuals:

* Members may have the biggest volume of data, besides on saturday. On this weekday, casuals take place as having the most data points.
* Weekends have the biggest volume of casuals, starting on friday, a ~20% increase.
* We have more members during the morning, mainly between 5am and 11am. And more casuals between 11pm and 4am.
* There's a big increase of data points in the midweek between 6am to 8am for members. Then it fell a bit. Another big increase is from 5pm to 6pm.
* During the weekend we have a bigger flow of casuals between 11am to 6pm.
* Members have a bigger preference for classic bikes, 56% more.
* Casuals have more riding time than members.
* Riding time for members keeps unchanged during the midweek, increasing during weekends.
* Casuals follow a more curve distribution, peaking on sundays and valleying on wednesday/thursday.

What we can take from this information is that members have a more fixed use for bikes besides casuals. Their uses is for more routine activities, like:

* Go to work.
* Use it as an exercise.

This can be proven we state that we have more members in between 6am to 8am and at 5pm to 6pm. Also, members may have set routes when using the bikes, as proven by riding time for members keeps unchanged during the midweek, increasing during weekends. The bikes is also heavily used for recreation on the weekends, when riding time increases and casuals take place.

Members also have a bigger preference for classic bikes, so they can exercise when going to work.

Concluding:

* Members use the bikes for fixed activities, one of those is going to work.
* Bikes are used for recreation on the weekends.
* Rides are influenced by temperature.

## **Guiding questions**

* **Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?**

Yes. The data points to several differences between casuals and members.

* **What story does your data tell?**

The main story the data tells is that members have set schedules, as seen on chart 06 on key timestamps. Those timestamps point out that members use the bikes for routine activities, like going to work. Charts like 08 also point out that they have less riding time, because they have a set route to take.

* **How do your findings relate to your original question?**

The findings build a profile for members, relating to "Find the keys differences between casuals and annual riders", also knowing whey they use the bikes helps to find "How digital media could influence them".

* **Who is your audience? What is the best way to communicate with them?**

The main target audience is my cyclistic marketing analytics team and Lily Moreno. The best way to communicate is through a slide presentation of the findings.

* **Can data visualization help you share your findings?**

Yes, the main core of the finds is through data visualization.

* **Is your presentation accessible to your audience?**

Yes, the plots were made using vibrant colors, and corresponding labels.

## **Key tasks**

* [x] Determine the best way to share your findings.
* [x] Create effective data visualizations.
* [x] Present your findings.
* [x] Ensure your work is accessible.

## **Deliverable**

* [x] Supporting visualizations and key findings

# Act

The act phase would be done by the marketing team of the company. The main takeaway will be the top three recommendations for the marketing.

## **Guiding questions**

* **What is your final conclusion based on your analysis?**

Members and casual have different habits when using the bikes. The conclusion is further stated on the share phase.

* **How could your team and business apply your insights?**

The insights could be implemented when preparing a marketing campaign for turning casual into members. The marketing can have a focus on workers as a green way to get to work.

* **What next steps would you or your stakeholders take based on your findings?**

Further analysis could be done to improve the findings, besides that, the marketing team can take the main information to build a marketing campaign.

* **Is there additional data you could use to expand on your findings?**
  + Mobility data.
  + Improved climate data.
  + More information members.

## **Key tasks**

* [x] Create your portfolio.
* [x] Add your case study.
* [x] Practice presenting your case study to a friend or family member.

## **Deliverable**

* Your top three recommendations based on your analysis

1. Build a marketing campaign focusing on show how bikes help people to get to work, while maintaining the planet green and avoid traffic. The ads could be show on professional social networks.
2. Increase benefits for riding during cold months. Coupons and discounts could be handed out.
3. As the bikes are also used for recreations on the weekends, ads campaigns could also be made showing people using the bikes for exercise during the weeks. The ads could focus on how practical and **consistent** the bikes can be.

# Conclusion

The [Google Analytics Professional Certificate](https://www.coursera.org/professional-certificates/google-data-analytics?utm_source=gg&utm_medium=sem&utm_campaign=15-GoogleDataAnalytics-LATAM&utm_content=15-GoogleDataAnalytics-LATAM&campaignid=12686019520&adgroupid=120140812253&device=c&keyword=google%20analytics%20course&matchtype=b&network=g&devicemodel=&adpostion=&creativeid=512414119178&hide_mobile_promo&gclid=EAIaIQobChMIwPGyn7f68AIVAgyRCh2nCwfVEAAYAyAAEgKjpvD_BwE) teached me a lot and the R language is really useful for analysing data (although I prefer preffer pandas). This took me more time than I expected, but it was fun.

Sorry for my bad english during the notebook. And thanks for anyone who is reading this.

In [ ]: