Ask:

**Business problem:** The company wants more annual memberships to secure a steady profit for the company’s future success.

**Business objectives:**

Customer retention

More annual memberships

Converting daily customers to annual customers

Using historical data to aid media marketing

**Stakeholders:**

Cyclistic Executive board.

Cyclistic Marketing director.

Cyclistic analytics team.

**Key business questions:**

How do casual members' riding patterns differ from annual members?

When are casual riders most active vs members?

What trend in ride time or day can we observe for casual vs member?

**Business task(the deliverable):**

Analyze historical data to find out how do casual riders and annual members **differ** in how they use the Cyclistic bike-share system to provide insights that will help Cyclistic **convert** more casual riders into annual members.

Prepare:

You want to address these 5 key things (you can copy this as a checklist):

1. **What data is available?**

I have available data from May 2020 until June 2025

1. **Where did it come from?**

The data has been made available by Motivate International Inc. It represent historical data of Cyclistic which is a fictional company

1. **Is it complete and trustworthy?**

There are some missing station names in the data, no age or gender or any other specific data about the riders. Only rider behavior, no other data like previous marketing campains

1. **What’s in the data?** (columns, formats, etc.)

Format: .csv

1. **Is this data relevant to the business task?**

**Data sources used(the deliverable):**

The data used is from Cyclistic’s bike-share system during July 2024 until June 2025 covering a 12 month period. It contains relevant information such as the rider type and the rider behavior which will benefit our business task. But it’s missing some personal information about riders.

**Data structure:**

Each file contains the following:

| **Column Name** | **Description** |
| --- | --- |
| ride\_id | Unique ID for each ride |
| rideable\_type | Type of bike used |
| started\_at | Start date and time |
| ended\_at | End date and time |
| start\_station\_name / id | Location where the ride started |
| end\_station\_name / id | Location where the ride ended |
| member\_casual | Rider type (casual or member) |

**Data format:**

The data comes in .csv files, with each file containing rows in the range 300k-800k.

**Data credibility and integrity**:

Does this data ROCCC, is it reliable, original, comprehensive, current, cited?

The data structure provided is consistent throughout all months. The data has been made available by Motivate International Inc and it is original. It's easy to understand and follow and updated each month so it’s current. Since Cyclistic is an imaginary company and we know the source of the data, it’s cited.



Data relationships errors, Id’s and names?

How many missing values does the data have?

Can I fill those null values?

What about duplicates?

Do I have outliers?

Irrelevant data to the business task?

Spelling mistakes?

**Data relevance:**

The data includes the rider type and the rider behavior which is crucial to answering the question. How do annual members and casual riders use Cyclistic bikes differently?

**Importing the cleaned csv**:

Process:

Check the data for errors.

● Choose your tools.

● Transform the data so you can work with it effectively.

● Document the cleaning process.

# **Data Processing & Cleaning Checklist**

## **Tool Selection**

* Document which tools you're using (R, Python, Excel, etc.)
* Explain why you chose these tools for your analysis

## **Data Integrity & Error Checking**

* Document the data errors you found during exploration. **Detection** - finding the errors
* Record how you verified data integrity (duplicates, missing values, outliers).**Verification methodology** - how you systematically checked
* Note any data quality issues that could affect analysis. **Impact analysis** - understanding consequences for your analysis.

## **Data Transformation**

* Create new columns needed for analysis:
  + Calculate ride\_duration (from start/end times)(I applied a filter for ride duration: 1.5min and 2hours)
  + Extract day\_of\_week from date
  + Extract hour\_of\_day from start time
* Document any other transformations performed

## **Data Cleaning Verification**

* Test that calculated columns work correctly
* Verify final dataset has no critical errors
* Confirm data is ready for analysis phase
* Record before/after cleaning statistics

Analyze:

**Business problem:** The company wants more annual memberships to secure a steady profit for the company’s future success.

More annual memberships.

Do: how do casual riders and annual members **differ.**

I have two categories. I need to compare them.

Do I do this using a hypothesis? Or do I just analyze just like that?

**Business task(the deliverable):**

Analyze historical data to find out how do casual riders and annual members **differ** in how they use the Cyclistic bike-share system to provide insights that will help Cyclistic

**SQL Analysis “Grande” Checklist**

1. **Clarify Objectives & Metrics**
   * Decide exactly which KPIs you need (e.g. average ride duration, ride counts, frequency per user).
   * With what metrics can I compare both:

Ride duration: bar chart

Ride counts: bar chart

Frequency per user: bar chart

Time of day: chart

The change of time of day vs the year: chart

Day of week: chart

Month of year: chart

Stations used: chart

Bike type: KPI

Number of rides per year: KPI

Most start/end station used by both: KPI

Month with most rides for both: KPI.

Day of week with most rides.

Growth in rides

I am thinking about daily plot, then weekly, then monthly, then yearly?

* + Identify the dimensions you’ll compare (e.g. rider type, day-of-week, hour).

1. **Set Your Time Window**
   * Define the date range boundaries for analysis (start/end dates).  
     12 months.
   * Ensure consistency with business reporting periods.
2. **Filter for Data Quality**
   * Apply duration or value thresholds to exclude erroneous records.

Already did that in the cleaning part

* + Exclude test rides, maintenance trips, or incomplete records.  
    I already did that.

1. **Group & Aggregate**
   * Choose grouping keys (e.g. member vs. casual, weekday vs. weekend, hour bins).
   * Calculate your aggregates (means, counts, medians, percentiles).
2. **Validate Results**
   * Perform spot-checks against raw data to confirm correctness.
   * Look for outliers or unexpected zero/NULL aggregates.
3. **Optimize for Performance**
   * Leverage indexes on filtered or joined columns.
   * Consider materialized views or temporary tables for repeated queries.
4. **Standardize Column Names & Ordering**
   * Use clear, descriptive aliases for metrics and dimensions.
   * Order your grouping keys in the natural sequence you’ll plot (e.g. 0–6 for weekdays).
5. **Export or Persist**
   * Export the final result set (CSV, table export, or materialized view).
   * Ensure it’s stored in a location your visualization tool can access.
6. **Document Your Steps**
   * Note any filters, thresholds, or edge-case logic you applied.
   * Record query versions or view names for reproducibility.
7. **Handoff Prep**

* Confirm schema matches the plotting requirements.
* Tag your output with a timestamp or version for audit trail.

**KPIs (Big Picture)** Start with your headline metrics so everyone’s on the same page: total rides, conversion rate, top station, peak month, etc.

**Time Granularity** Drill into when activity happens: daily/weekly patterns, then zoom out to monthly and yearly trends. This shows seasonality and cyclical behavior.

**Stations / Bike-Type Breakdown** Finally, drill into location and equipment: which stations matter most for each group, and which bike types they prefer.

First Page:

**KPI’s:**

Total rides per member type

Average ride duration

Total yearly ride time

Peak month per member type

**Time granularity(trends):**

Daily trends in hour of day per member type

Weekly pattern per member type

Monthly pattern per member type

Seasonal pattern(winter, spring, summer, fall)

**Location and equipment:**

Top stations per member type

Prefered bike type per member type

Analyze:

What is the aim of my visualization: how do casual riders and annual members **differ** in how they use the Cyclistic bike-share system.

Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently? Yes, they use the bikes at a different concentration of hours per day.

What story does your data tell? It tells the story on how the members are workers while casual riders don’t need to commute, but probably a large section of them are commuters.

How do your findings relate to your original question?

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

Cyclistic Executive board.

Cyclistic Marketing director.

Cyclistic analytics team.

The best way is to make a dashboard that tells the story of the differences between the two rider types

Can data visualization help you share your findings? It’s the only way?

Is your presentation accessible to your audience? That I need to work on yes

My dashboard so far:



What I would like to edit in it:

Remove the total rider kpi and keep the bar chart for total riders but add the number to the bar chart so now I have a hybrid barchart/kpi

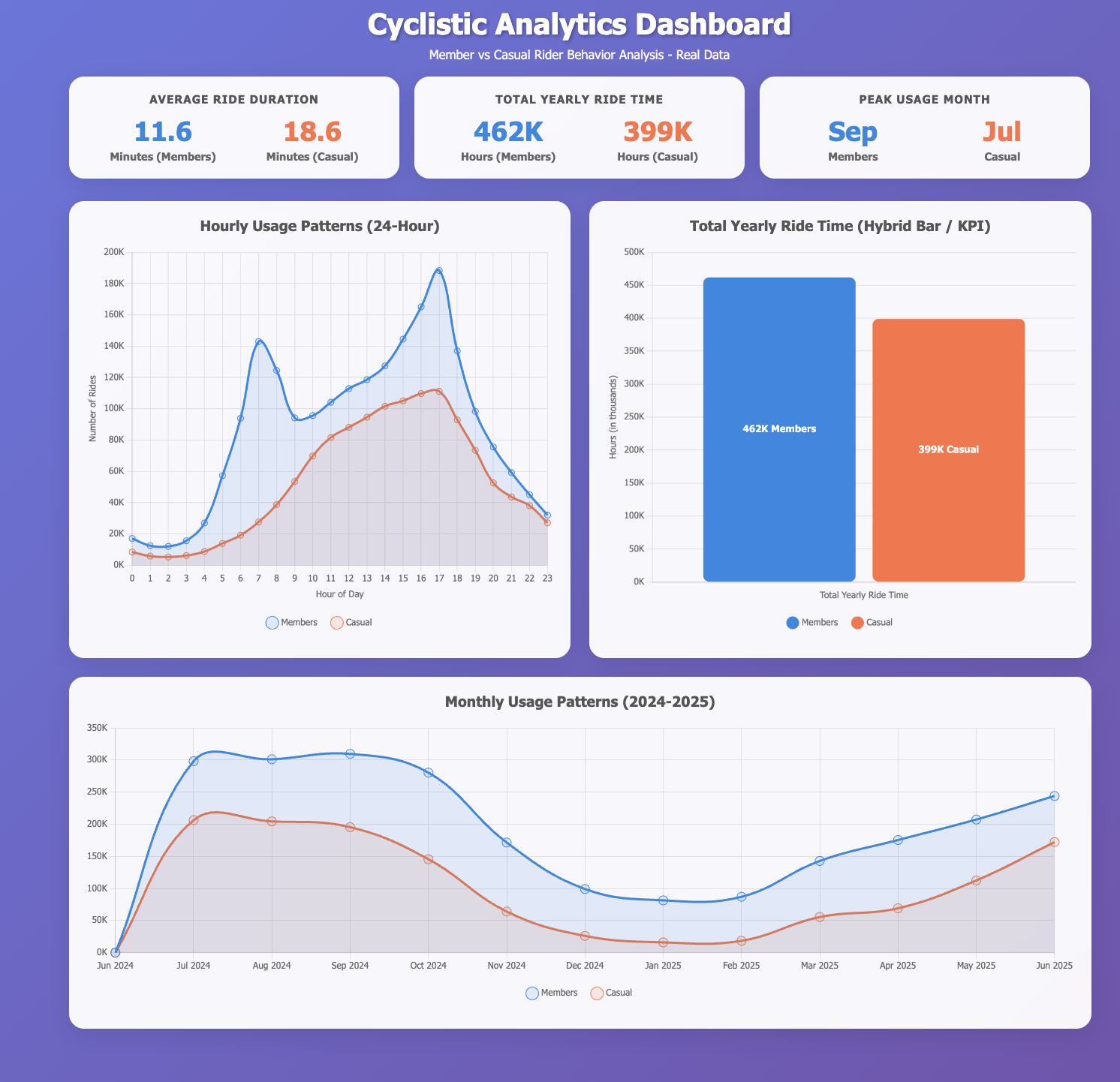
Remove the daily usage bar chart, replace that with the hourly usage pattern from the behavioral patterns page

Remove the behavioral patterns page

Organize the new dashboard in to one page, so we would have 3 KPI’s which are average ride duration, total yearly ride time, peak usage month and 3 charts, which are monthly usage patterns, hourly usage pattern and the hybrid barchart/kpi of total yearly ride time. Organize the new dashboard in a way that it had emphasis on the hybrid barchart/kpi and the hourly chart.  
It should be clear to anyone who’s seeing it on the first 5 seconds  
It should be accessible

Remove the cycler emoji from the title, keep it professional.

The edited dashboard



The hourly usage is good, the monthly is good.

Total yearly ride time is irrelevant because there’s a lot of dropped rows so max and min are not so relevant.

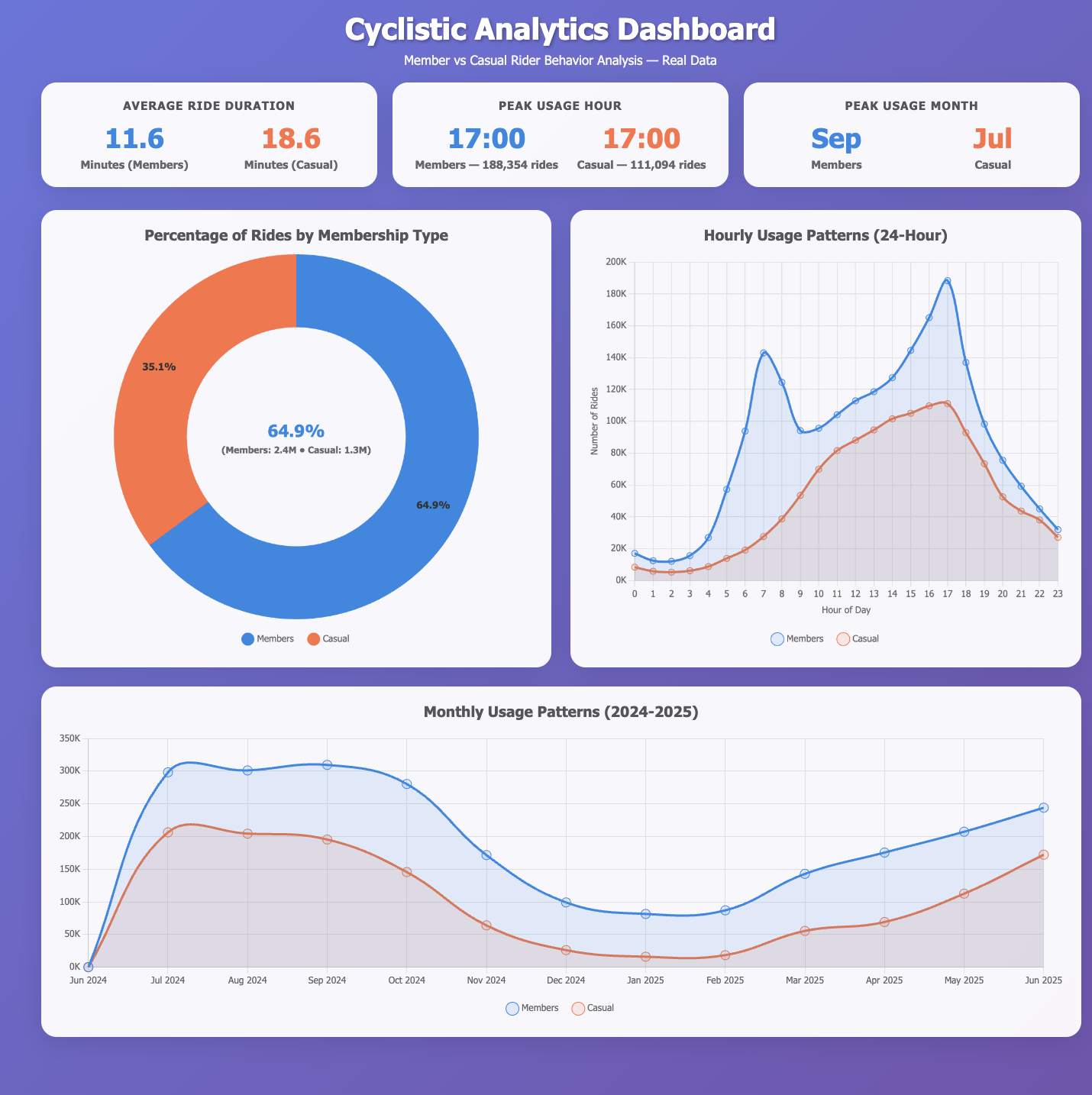
I need to drop total yearly ride time KPI and the chart as well

I need new measures:

**Percentage of rides by membership type rather than total yearly ride time chart**

**New hourly KPI, peak hour for both casual and members in place of total yearly ride time**

**The edited dashboard 02:**

****

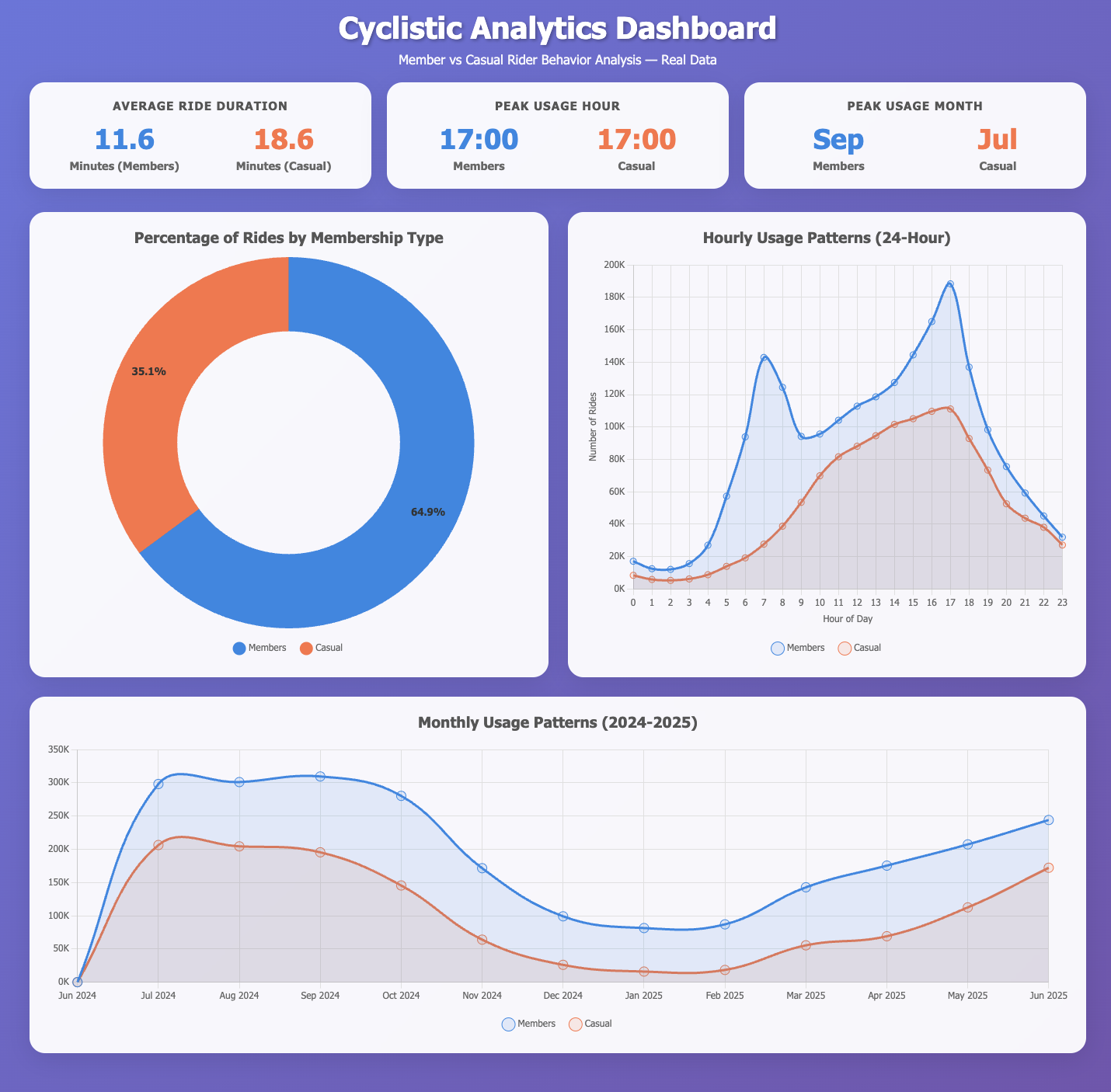
**In this I like the new donut chart but I don’t like that in the middle of it there’s a number written so that needs to go, i don’t need that count nor the percentage that’s written. I need to compare not get the max**

**The peak usage hour is actually relevant because it means there’s a lot of casuals who are commuting so i can keep that**

**The monthly clearly shows that the causal and the members follow the same pattern so that’s good**

**The casual riders have a higher average ride time despite having less rides so that’s also good.**

**The third edition:**

****

**I see a problem with the hourly usage and monthly usage patterns, they show number of rides on the y axis while i would want that to be a percentage**

**Actually, I’ll keep it that way.**

Act:

Professional

Deliverable: Your top three recommendations based on your analysis.

● What is your final conclusion based on your analysis?

That cyclistic already has more members than casual riders and that casual riders tend to have longer riders than members do, they are very similar in how they ride apart from the daily time they follow, the members seem to peak when commuting to work in the morning and in the evening. While the casual riders seem to peak during the evening rush hour.

● How could your team and business apply your insights?

My team could use these insights to decide on how to convert casual riders to member by targeting those who benefit from the evening rush hour as many of them are probably commuters who would be interested in a membership.

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

The next step would be to look at data that shows clearly the riders id so we know who would be interested in commuting and who would not be.

● Is there additional data you could use to expand on your findings?

Yes a dataset with riders ID, they’re type of work would be great to pinpoint who to target when coming up with a strategy to convert casuals to members.

Steps for me to finish this step:

1. Gather all the documents and files into github, create both readme files.
2. Prepare the outer readme file
3. Prepare the inner readme file( the deliverable)
4. Seal everything up and be done with the project

All the files I have worked on:

The raw data.

The kaggle notebooks.

The big query sql.

The html file for the dashboard.

The word file that I used to track my process.

The final deliverable.