# Meeting notes

* Predict time series: 2026/ Christmas 2025
* Peak hour, off peak hour – Dang lam
* Weekday hay weekend, tháng nào, năm nào tap on, off nhiều nhất
* Public holiday
* Top 5 region đông nhất (top 3 – 4)

2023: Bthuong mới sau COVID

## Chia việc

Vinh: Pattern peak hour, weekday – weekend, holiday

* Time series (total patronage each month) 2020 – 2025
* Average Weekday vs Weekend Patronage
* Average patronage by month (2020 – 2025)
* Average peak vs off-peak patronage
* Average patronage by region (2020 – 2025)
* Time series daily patronage December 2020 – 2025
* Predict được tháng 7, 8 = predict được tháng 12
* Pattern tap-on, tap-off sẽ khác nhau

Anh Nhân: Map population density & job density

* Nối được số lượng business, chia theo industry => Tổng doanh thu business theo vùng => số job => ở đâu buổi sáng, tổng doanh thu service => ở đâu buổi tối
* Có được population từng vùng
* Giờ cần chuyển area sang postcode -> Nối với patronage

CAn: Tìm dataset có thể kết hợp được -> Point/ area

* Load xong data rainfall, max temp, min temp daily 2017 => 2020 – 2025
* Nhìn qua trend weather cả năm (monthly) + December
* Thử plot time series daily patronage December 2020 – 2025
* Ghép weather với patronage (chia ra theo khu vực). Chỉnh lại minTemp - maxTemp

Thời tiết -> restructure data để máy đọc được -> thử tìm xem có liên quan gì đến Opal patronage, thử chạy regression, decision tree

## Backup plan

Làm việc dựa trên thuần địa lý

1 ngày có 24h, có 8 region (ko kể Other)

Assumption: Mỗi region sẽ có distribution riêng

Loại ngày: Weekday/ Weekend/ Holiday

Assume seasonal factor ko tồn tại/ Làm typical day của 1 tháng

* Predict 24 số \* 3 loại ngày \* 8 region

3h Úc: Dừng làm analysis, vào họp

Muộn nhất 4h gộp hết ttin vào với nhau xong

Report, video: Chia việc sau

Vinh prefer làm report

Có thể làm ppt

Control variable (?):

Population density => Ảnh hưởng sluong tap on vào sáng, tap off vào tối

Job density (theo địa lý) => Ảnh hưởng sluong tap off vào sáng, tap on vào tối (Còn phụ thuộc vào business type – normal business, entertainment business làm việc về đêm…)

Development – sửa tàu, sửa ga => Thêm số tàu, số bus ở nơi đang có development

Thời tiết – Check những lúc thời tiết abnormal, tìm hiểu xem ảnh hưởng của nó là gì

Development & thời tiết => Phức tạp hơn, vì có thể ảnh hưởng đến tâm lý của người sử dụng ptien công cộng

Check xem distribution khi thời tiết ổn & thời tiết ko ổn liệu có như nhau ko

Occupancy rate của tàu – 2017 -> 2019 (ko có ích lắm vì pre-COVID, ko khớp khung tgian)

Research question:

* Pattern đi lại => Dự đoán demand (cố gắng cụ thể ở mức giờ, ga, vùng), biết bnhieu khách ở ga nào, đi chuyến nào => Cần xây thêm ga ở đâu

Dùng cả 2 cái cùng lúc (?)

Đã có:

* Time series từ 2020 (tổng tap on tap off theo tháng, tất cả phương tiện)

# Reminder

gitignore: .Rdata, .Rhistory

Đầu tiên: Cứ tính tổng tất cả các region, tất cả các transport mode, daily

2: Chia theo khu vực

3: Chia theo transport mode

Thử show Opal patronage & some weather data trên cùng 1 graph (biểu đồ kết hợp đường – cột)

Chỉnh lại maxTemp, minTemp về cùng 1 cách đo

maxTemp and minTemp measure 2 different things -> Might need to make some changes

The minimum temperature measured at that time is recorded against the day on which it was measured, while the maximum temperature is recorded against the previous day.

minTemp den 22/7, maxTemp den 21/7: maxTemp dung voi normal convention hon

Co the tinh trung binh maxTemp, minTemp but that’s not quite statistically meaningful => (maxTemp + minTemp) != averageTemp

Co the tinh maxTemp – minTemp nhưng cái đó ko liên quan gì đến tap on tap off lắm?

# Brainstorm

**Opal Patronage**: Only train, bus, ferry and light rail. Metro not included.

**Train station entries and exits**: Provided with data from August 2024. Can download data from July 2016. Have “typical day” data 2016 - 2023.

**Sydney Train Routes**: Seems like geospatial vector files => Don’t know how to open/ view yet

260 train stations, 30 metro stations?

There is only **entrance location** data for train stations. No such data (?) for bus, ferry, light rail, metro

Drop on Sep24, rise on Oct24. => Barangaroo, Tallawong,…

Really low on Jan25, really high on March25 => Lidcomb, Circular Quay, Domestic, Granville, Olympic Park, Bankstown, International, Miranda, Cronulla, Campsie (weird), Lakemba (weird), Wollongong (Jan ~ Feb), Marrickville (closure), Punchbowl (weird), Katoomba (peak on April), Dulwich Hill (super drop – closure?), Belmore (closure?), Hurlstone Park (closure?), Yagoona (weird), Canterbury (closure?), Wiley Park, East Hills, Birrong

Stations that don’t follow this pattern: Martin Place, Bankstown (station closure?), Olympic Park, Circular Quay, Gadigal, Crows Nest, Victoria Cross

Leppington Station is missing

Enquiry

Development: Later on 1st day/ Start of 2nd day

Evaluation: Later in 2nd day

If you’d get a different answer if you change the method, it is a problem

Refinement: Early Wednesday

Presentation

Report: a lot more ability to go into detail compared to the 3-minute video

# Q&A

Recommend to focus on a large question, instead of answering 3 separate

Make sense to have another questions following along

R Notebook/ Quarto Notebook/ Jupyter Notebook… -> Needs to have comments, documentation…

If the orginal data is in Excel and has to do some manual work -> Can document the steps in comments

Lunch time Tuesday -> Suggested structure for the report

Exec summary/ Motivation Statement, Posing the question, Analysis, Result, Conclusion

Try to pose an interesting/ impactful question

Focus on the question. Make sure that you don’t give excessive information, try to just address the question

Mark innovation, clarity, etc. across both the video & the report

Innovation: auxiliary information, drawing from other sources, really intriguing questions, creating a platform to examine the data

Can pitch the question on Ed

Maybe: At least 3 graphs in exploratory, at least 2 in answering the questions

Presentation & report weighted equally

3-minute thesis competition: take into account what kind of tactics they use to get engagement

Can use slides for the video

The audience is up to you for decide

# Research questions

* How did public transport usage shift during the COVID-19 pandemic, and how long did it take to recover?
* What impact did extreme weather events (e.g. storms, heatwaves) have on train travel?
* Are there observable trends in train station usage tied to changes in work-from-home policies or school schedules?
* How do travel patterns differ on public holidays or school breaks?
* Has the construction of new infrastructure, such as Metro lines or residential developments, shifted demand across the train network?
* Can you detect signs of inequity in transport access or usage across regions?
* How did Opal fare and other transportation methods’ costs change over time? Any substantial trend in transportation method? How did that impact public transport usage? Suggestions for future Opal fare policies
* Are there significant differences between public transport usage between the years?

# Weather data

Previous to Jun 2024:

* Maximum, minimum temperature (from 2017 Oct + from 1859 Jan) – Lots of locations
* Rainfall (from 2017 Oct + 1858 Jul) – Lots of locations
* Unavailable:
  + No recent data: Evaporation, sunshine
  + Not free: Temperature at exact time (dry bulb), wind at exact time (speed and direction), wind maximum gust

Region -> Station:

* Chatswood: **Observatory Hill** (9.33km) has all 3, North Ryde Golf Club (5.12km) only has rainfall
* Macquarie Park: **Sydney Olympic Park 066212** (8.33km) has all 3, North Ryde Golf Club 066213 (3.37km) only has rainfall
* Sydney CBD: **Observatory Hill 066214** (1.28km, 98%) has all 3, Sydney Botanic Gardens 066006 (0.73km, 90%) only has rainfall
* North Sydney: **Observatory Hill** (3.27km)
* Parramatta: Parramatta North 066124 (3.54km)
* Strathfield: **Sydney Olympic Park** (4.34km) has all 3, Concord Golf Club 066013 (2.11km, 90%) and Concord (Brays Rd) (3.05, 99%) only have rainfall
* ~~Newcastle: Newcastle Nobbys (2.14km)~~
* ~~Wollongong: Bellambi Aws (7.61km) has all 3, Port Kembla NTC Aws (5.19km) has temp til 2025 and rainfall til Jun 2024~~
* Sydney CBD & North Sydney: same weather data