Task 1

I have written everything (challenges faced, thinking process, solution etc) in the notebook, so reading that is preferred /very very recommended. I have made this document by copy pasting stuff from there and the notebook is very structured and readable

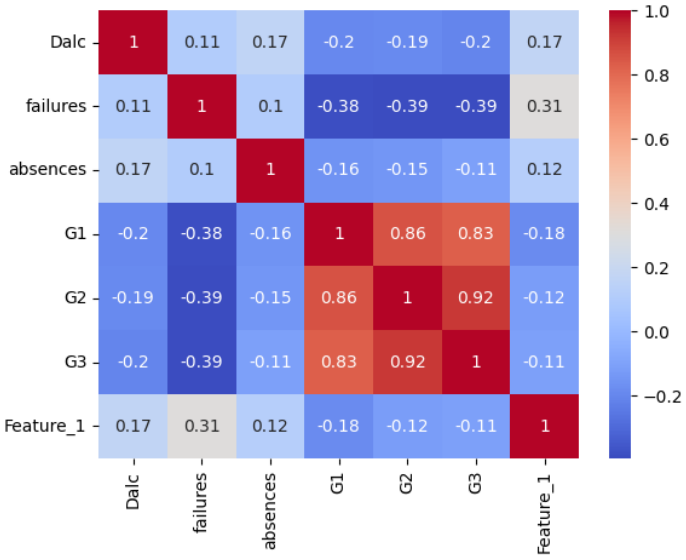
- Before starting, I first read all the resources and some extra videos for analysing data as the first task was related to that

- As our final task is to predict whether a student is in a relationship or not, I'll not consider relationship in the previous levels

Level 1: Variable Identification Protocol

- Starting with identifying **Feature 1**, it ranges from 15 to 22, used value_counts for that and the frequency decreases in value as the value of feature 1 increases. So obviously, considering that it's a high school, my first guess would be that it's age.

Testing my hypothesis:



- out of the following metrics, I am choosing the metrics that would corelate with age to support my hypothesis: famsup, traveltime, failures, higher, goout, Dalc, absences, G1, G2, G3, Fedu, Medu, Pstatus. I'll only choose metrics which should have high correlation with age:

- positive correlation {failures, Dalc, absences}

- negative correlation {grades}

- So, I plotted the correlation matrix, and voila, there is a correlation. in fact there is a pretty strong correlation between feature 1 and failure, and some weak correlations for absences and Dalc. Also there is a **negative** correlation between f1 and grades. So I think we can confidently conclude that feature 1 is age

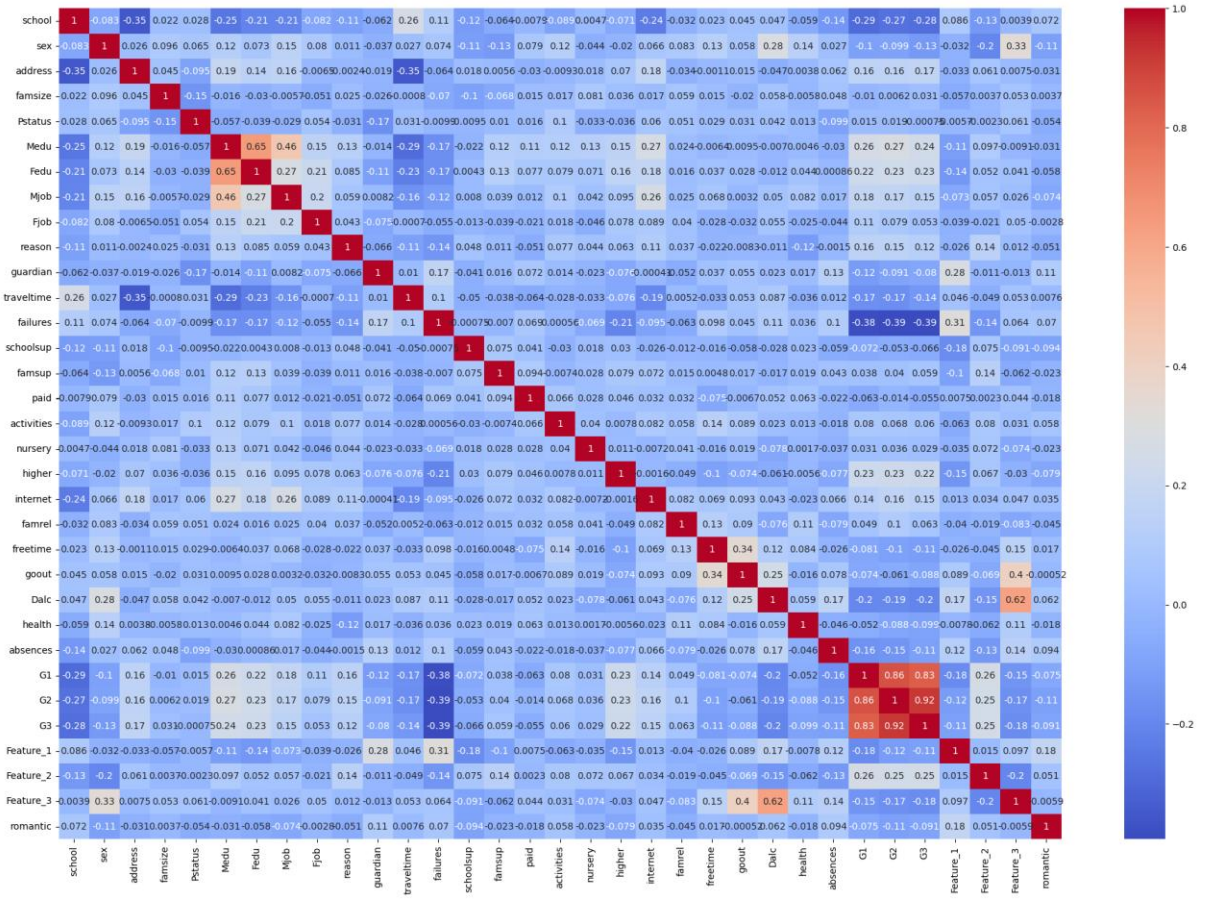
Answer: Feature 1 is Age

Before finding out feature 2, Let's draw the heatmap of correlation matrix between all data, as that would be useful for level 1 and other levels too!

But before drawing that, I converted all values to numerical values, so that it will be able to plot all values

(Label encoder encodes alphabatically and it encodes NA values too, so that's a fault but it won't matter for finding feature 1,2,3. So after cleaning the data I'll rerun this block to get more accurate results)

ATTACHING THE CORRELATION HEATMAP HERE! USED THIS THE MOST IN TASK 1!!



Feature 2:

This one was tricky:

- as it only has a strongish/okayish/most correlation with grades and sex.

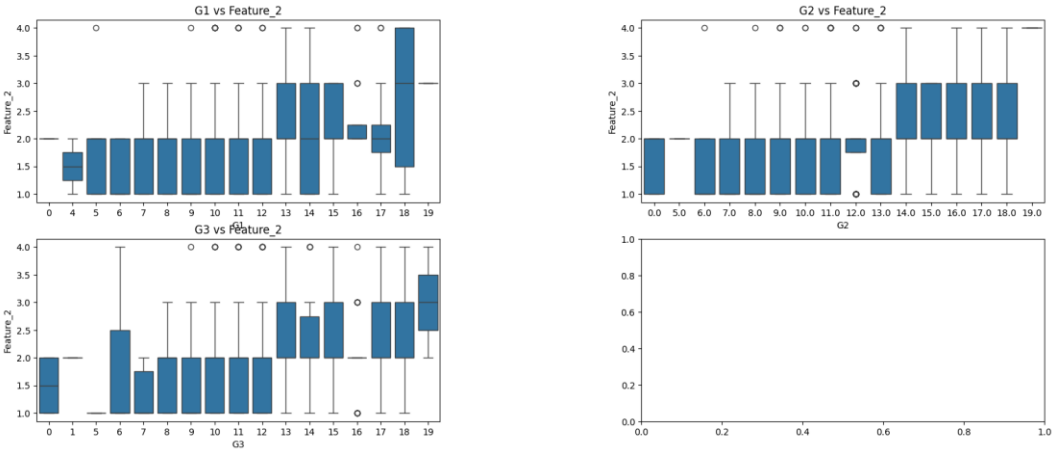
- And a positive correlation with grades, meaning it increases in size with grade and a negative correlation with sex, means it is higher for females.

- But I think that it has a correlation with sex cause sex has a correlation with grades, so I'll only consider grades

-Also, as 2 has the most frequency, 2 would be the average/mode values for students and 3 and 4 are pretty rare

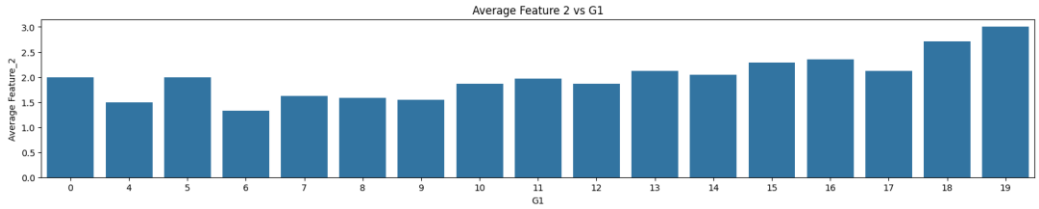
- Also noting down another point, these values are discreet and only contain {1,2,3,4}.

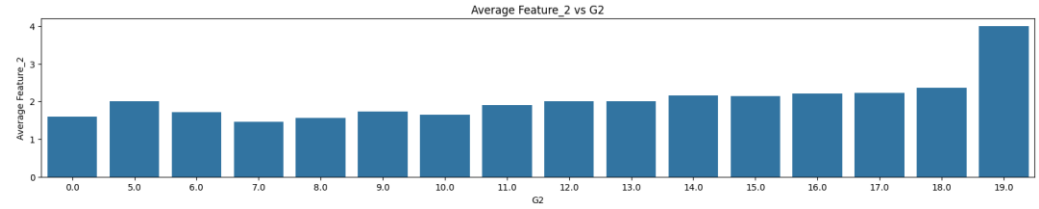
To analyse more, i want to see the distribution of students with each grade, so plotted a boxplot as this should show the distribution.



Though I see a upward trend, feature 2 is generally higher for people with higher greades for G2 but still not sure. So, there might be variance on time too, between diff terms, so let's ignore G3 as it's the final grade, so maybe the average of the two.

Tried drawing a scatter plot but it doesn't show anything. Now to get a clearer picture I want to know the average value of feature 2 for each value of each grade (yeah, a little confusing 😅, but what I mean is I want to see the relation between average value of feature 2 and grade)





Hmm, although I am not sure about F2 as i was about F1, but I think after seeing the graphs and distribution, I can make a credible guess:

- F2 is discreet values, so it might be certain grouping for ex, how health is

- It depends on G1 and G2 and has a more positive correlation with G2 means it also depends on terms/semester which means time

Ans for feature 2:

Feature 2: It is a method of grouping students by their academic ability, basically how coaching institute divide students, higher scoring students are placed in higher value of feature 2. Moreover, there is a reshuffling after term 1, that's why there is a more positive correltion with G2

Feature 3:

Feature 3 again has discreet values {1,2,3,4,5} and seeing the correlation heatmap, it has a high positive correlation with dalc, gout, sex(males, increases with males) and negative correlation with grades. Seeing this, I think it denotes:

- categorizing students based on behaviour (bad behaviour = high feature 3) or it can also be number of detentions (you get kicked out from school after 5 detentions)

Level 2: Data Integrity Audit

It doesn't make sense to remove the nan rows cause then we'll be wasting data(Around 7-8% of the values in some columns are missing)

1) ****Famsize****: It has high correlation with Pstatus, so I filled the nan's of famsize using Pstatus, when pstatus=a, famsize =LE3...similarly for pstatus=t (also considering that is parents are together then there is a high chance of fsize GE3, this is also evident from corr heatmap)

2) **Fedu**: Fedu and Medu has a very high correlation, and mostly the values of Fedu is same as Medu, using this I filled nans

3) **Traveltime**: It has a very strong corr with address, so filled nans like: If address = 'U', filled nans with mode of travel time under 'U' and similarly for 'R'

4) **Higher**: This is a discrete value, also it's distribution is skewed towards 'yes' with small number of outliers, so filling it with mode seemed appropriate

5) **Feature 1**: Same as 'higher'

6) **absences**: It doesn't correlate much and is a discrete value, so again mode seemed appropriate

7) The following data strongly correlate between each other, so filled nans of feature 2 and 3,freetime and G2 using KNN

Level 3: Exploratory Insight Report

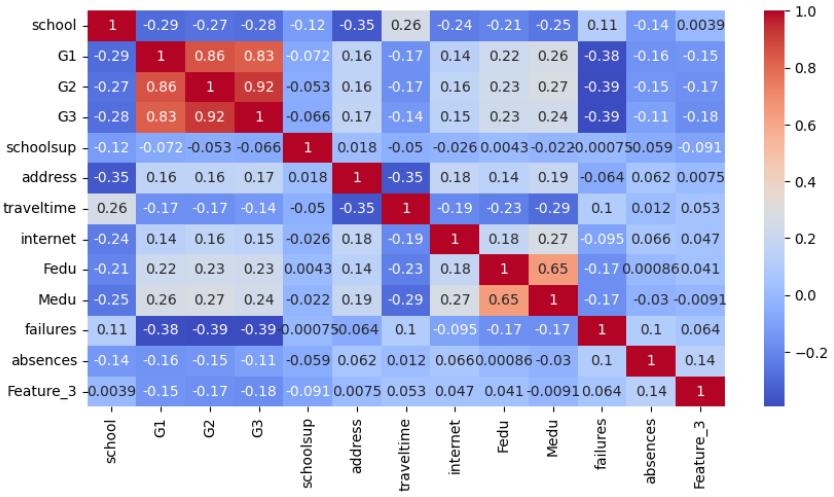
1) What's the difference between both the schools? ###

note that in the data_num GP is 0 and MS is 1 and strings are converted to numbers using alphabetical order

****I have used the corr heatmap again, decoded it using alphabetical order and saw these observations:****
The data reveals a huge gap between GP and MS.

****Observations on MS****:
- MS has lower grades and school support
- The students are from rural area, having more travel time and low internet access, their parent's education level is also low
- The students also have more failures but they have lower absences
- No difference between both schools in feature 3 (behaviour)

****Conclusions drawn:****
While students studying in MS school have a disadvantage, their low absences and behaviour (on par with GP) indicate that the students are trying their best and the government and their parents should support them



2) Is there any disparity between male and female students?

****Observations:****

- Males have comparitvely bad behaviour (Feature 3)

- Females have higher grades

- Females get more family and school support

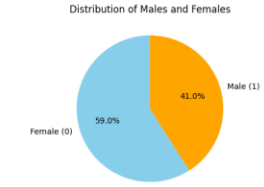
- There is strong correlation between mother's education and sex(male). There are also more females going to school. This might mean that boys are only going to school when mother's are more educated.

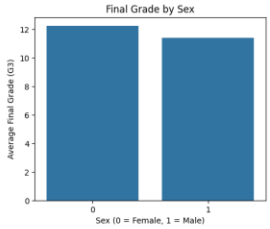
- Females are generally less healthy

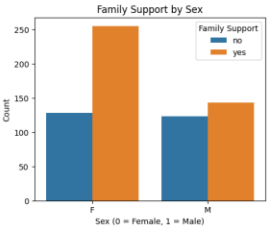
****Conclusions:****

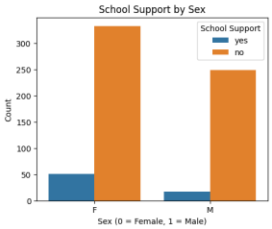
- Yes, there is a disparity. Male students are underrepresented, and with no school and family support. Although females' health score is low, so schools should look on that.

****#JusticeforMen ✊ ****









3) What all affects student's health? And how to improve it?

****Observations:****

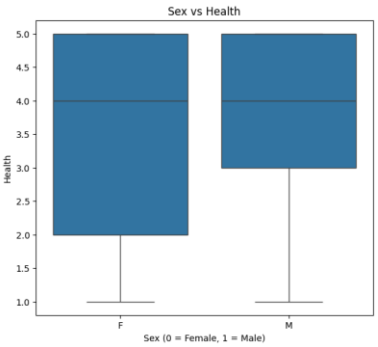
- Some of the females are more unhealthy (Health index=2 or 3), although an average woman is as healthy as an average man

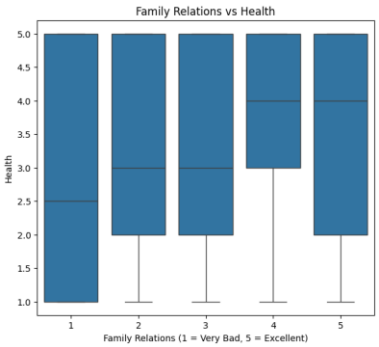
- Better family relations = better health

****Conclusions****

- Females in the lower health bracket should be checked upon

- Family relations affects health a lot, so a student should have good relations with his family members





4) How does pstatus affect student's health, academics and social life?

Although there isn't a strong correlation between health, academics and social life, the following observations can be made:

****Observations:****

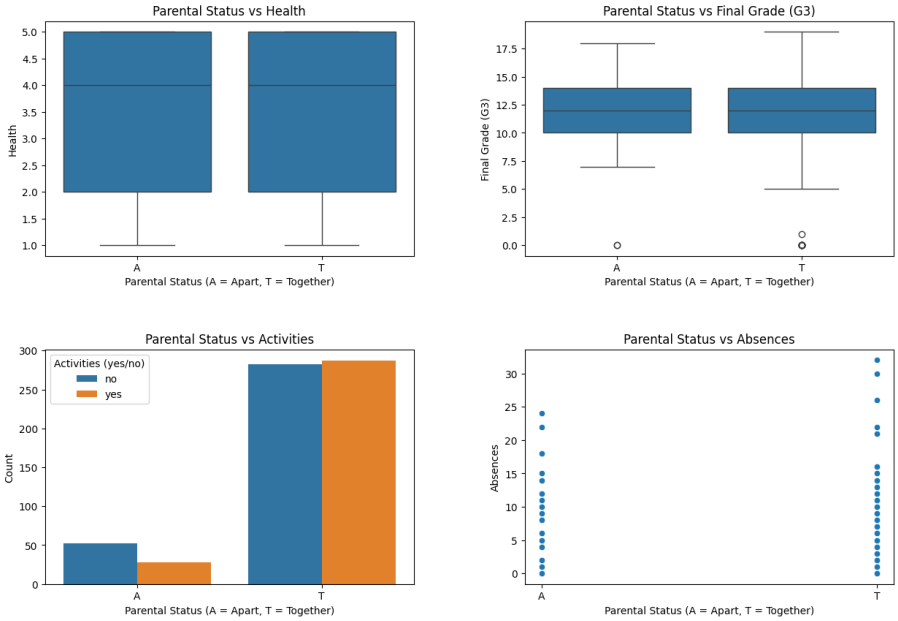
- Parental status doesn't determine health or academics

- But, there is a negative correlation between parental status and absences

- It also detrmines student's extra-curricular participation to some extent

****Conclusions****

- Parents should keep in mind their child's social life while seperating



5) Okay,now the big one : What all affects student's grades and then how to imporve school's grades?

I'll only be comparing G3 as other two are highly highly correlated with it

****Observations:****

- Absences affects the grades negatively (ignore high absences as they are relatively few and are outliers)

- Grades and health has a negative correlation

- Daily alcohol content also negatively impacts grades

- Intenet access also influences the grades

- Travel-time(already covered) and the guardian also affects the grade

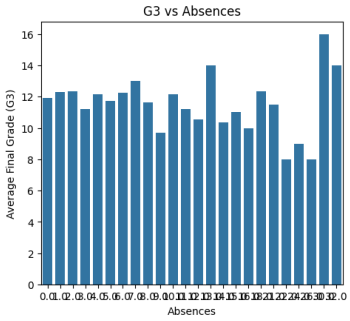
- Other factors include reason for joining school, mother's and father's education(corr to medu), address (already covered), sex (already covered) and school (already covered**)**

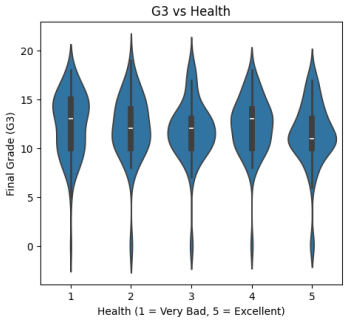
****Conclusions:****

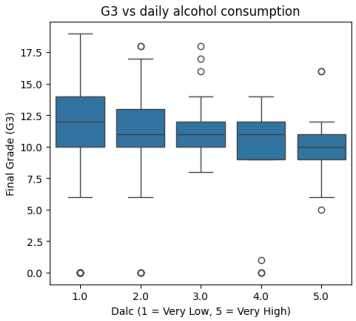
- If the mother or father is more educated, then the child gets better grades and thus the cycle continues, so parents and government should invest in education

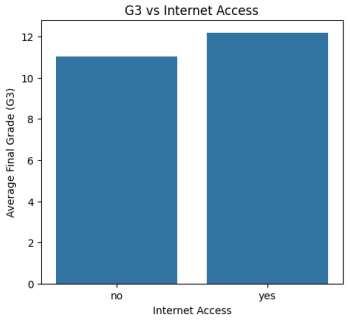
- Parent's should try to provide internet access to their children in order to get better grades

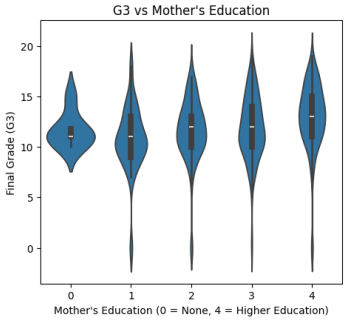
- And avoid alcohol, go to school regularly and be healthy to improve your grades :)

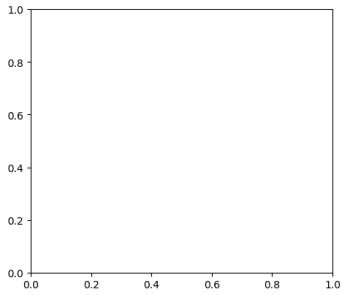












Level 4: Relationship Prediction Model

The factors on which it slightly depends:

- Feature_1, G2 (G1 and G3 too but they are **very** similar to G2, and G2 has the highest correlation, so not considering G1 and G3), absences, schoolsup, higher, guardian and sex

A logistic regression model should be good on this one as it have only binary values yes or no, but tried it and got only max 60% accuracy after trying out different combinations among them :(

I am a bit less on time, so wasn't able to read gridsearchcv :(

Yupp, so random forest gives the best (compared to others) accuracy which is 66.15%. But this isn't very effective. The input of random forest is grade, absences and sex:

- Sex is obvious as number of females are more in campus, so a female has a higher chance of being in a relationship

- But grades and absences might have happened ****due to**** relationship and not cause it. So if his grade is lower than some amount and absences is also greater than certain amount then one of the reasons causing it is that the student is in a relationship

So, this shows that if a student has absences > x and grades < y, then there is a good enough chance that relationship is behind it. So this tool might be helpful for teachers, if a female student is frequently absent and with low grades, then there is let's say around 70% chance that she is in a relationship and if we break up her from her boyfriend, then she might start scoring well. ****#BajrangDal****

TASK 2

I have written everything (challenges faced, thinking process, solution etc) in the notebook, so reading that is very very much preferred . Also the structure and readability of the notebook is also good

I recommend restarting the kernel before running each level

Level 1: Core Activation

What all I read before this task

I read the documentation, downloaded the API Keys and tried to understand what each block of code does(I don't exactly know how they work, but I have a pretty good idea what each block of code does and use for it)

I have also hard coded all the API keys (did ****a lot**** of requests so the free subscription might be on the verge of finishing, so use carefully XD)

And one thing I didn't like was that why tf are they changing libraries and locations of classes so fast, installed the latest versions but in 2-3 days, had to update some again! And the location of the classes is also not fixed!!

Challenges I faced for adding the tool

- I was trying to do like the Tavily search tool coded in the langgraph documentation. Was struggling a lot as I didn't know that the commented section under triple quotes is the description, I was removing that again and again from the documentation, finally got it right tho lol. I also read this - https://python.langchain.com/docs/concepts/tools/#tool-interface and referred material under Pre-built agents

I also encountered a problem in the last block which is mentioned in the notebook

Level 2: Senses of the World

So for this I used tavily, integrating it was simple as everything was mentioned in the document and for weather, I didn't initially see that it was given in the PS, so I used a youtube video for getting the API key for openweather and how to use it, then I put it in the form of a tool. I am attaching the link - https://www.youtube.com/watch?v=9P5MY_2i7K8 This cannot give the forecast or past weather tho :(

I have also commented out the part to test the code by invoking the tool

Level 3: Judgement and Memory

The code is mostly similar, just have added a few lines extra to save the memory and update it in the graph, I have put comments before those lines

Again I faced a problem in the last block which is covered in the notebook.

Level 4: The Architect's Trial – Multi-Agent Evolution

This one was the most challenging one yet. Took me a whole day for this level 👀. But made my understanding of graph builder strong. So here's my journey for this level:

- Read multi-agent subsection from the document but didn't understand it properly. So referred to a couple of videos but the best one was:

https://www.youtube.com/watch?v=JeyDrn1dSUQ It also had another repo in it's description (https://github.com/langchain-ai/langgraph-swarm-py) I first tried building a sample one from this and it worked!

- Now, the idea, I thought of building a wardrobe chatbot, but multiagents weren't needed for it, it could be done by using just a tool. So then I thought of building of travel packing assistant, included fashion agent, event agent and weather agent

- I was thinking of making it a swarm but midway thought that a supervisor with a packing agent as the supervisor would make more sense. But I had built everything for swarm so did a mixture of both.

- Do note that the event tool is for events within 1 month and weather tool is for weather forecast under 5 days.

-nailing the prompt of each agent was very very hard. I had to assign every agent it's exact role or it wouldn't work. For ex- If I didn't tell packing agent to approach sequentially, it would crash or if I didn't tell other agents their specific role then they would also try to do different things and fail. So I did use LLM here as I wanted to be perfectly clear. I gave it my demands, that I want it to do so and so things, write it in a consise and structured way. So after a few back and forths, got the final prompt which worked. If I had made a full supervisor structure instead of swarm then it would have worked well with a short prompt but I am a bit short on time and this is also working pretty good.

 # DONE #

Learned a lot here and it was the most practical and fun mini project ever!!!