**General Notes on the Assignment in ICS 5110**

**Link to the assignment on Overleaf:** [**https://www.overleaf.com/project/671f5602b8967f23a9d7681e**](https://www.overleaf.com/project/671f5602b8967f23a9d7681e)

# Our Email Adresses

* [danou.nauck.24@um.edu.mt](mailto:danou.nauck.24@um.edu.mt)
* [owen.gauci.24@um.edu.mt](mailto:owen.gauci.24@um.edu.mt)
* [ivan.ciancio.01@um.edu.mt](mailto:ivan.ciancio.01@um.edu.mt)
* [paolo.fidicaro.24@um.edu.mt](mailto:paolo.fidicaro.24@um.edu.mt)

# Group Preparation and Setup:

If we should fill in this form to register in the VLE as a group: <https://docs.google.com/forms/d/e/1FAIpQLScedj6KJOor9PWrT4-BcVJ4EUD5tGCCOK_VmK6ata44Hht3Bw/viewform>

# Data collection and Links

Maybe useful CSV from kaggle.com

<https://www.kaggle.com/code/sovannvathanaksay/investigation-on-zodiac-compatibility>

Assignment templates are also in the Google Folder, but also here

<https://www.um.edu.mt/vle/mod/folder/view.php?id=1280102>

**I found a CSV Dataset with 5.000+ mexican people** and their marriages, duration of marriage, birthdates, education level, monthly income for each partner, amount of children etc. So basically quite a good list, only two downturns: It is mainly focused on mexico. And mexico only. And the evaluation looks at the time bracket from 2000 - 2015. So maybe too old?

Its here : <https://www.kaggle.com/code/kerneler/starter-divorce-marriage-dataset-with-3f6abe71-f/notebook>

They already evaluated the compatibility rate between different signs of zodiac, the results can be found here <https://www.kaggle.com/code/kerneler/starter-divorce-marriage-dataset-with-3f6abe71-f/input?select=Comp_matrix.csv>

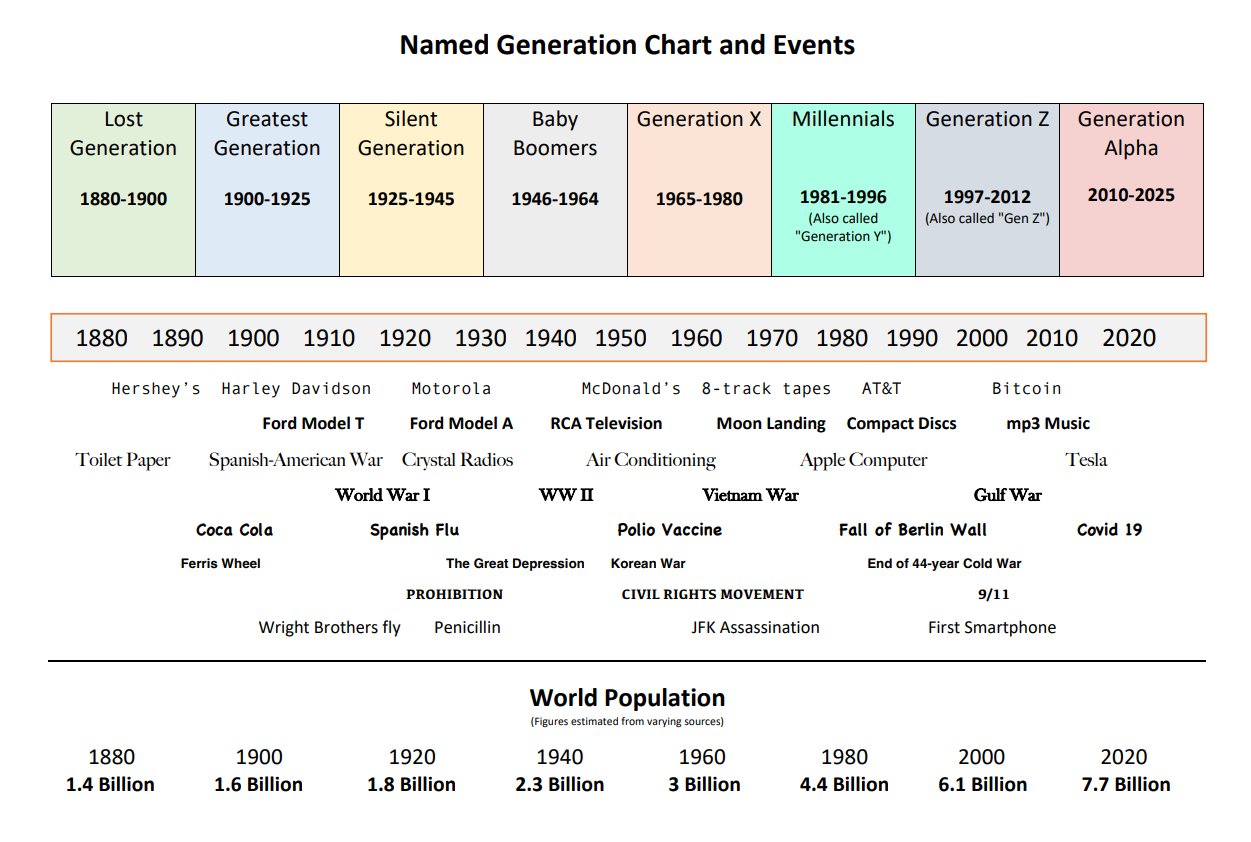
There are also all the necessary python scripts we could use, copy and extend if we want to extend further this data. The python scripts where itself cloned and extended by some other researchers. So basically we already have A LOT of working code to display graphically all sorts of graphs. Like male / female education, amount of kids / divorce, income difference to divorce ratio and all that stuph is all ready waiting for us. kaggle.com really rocks the house!

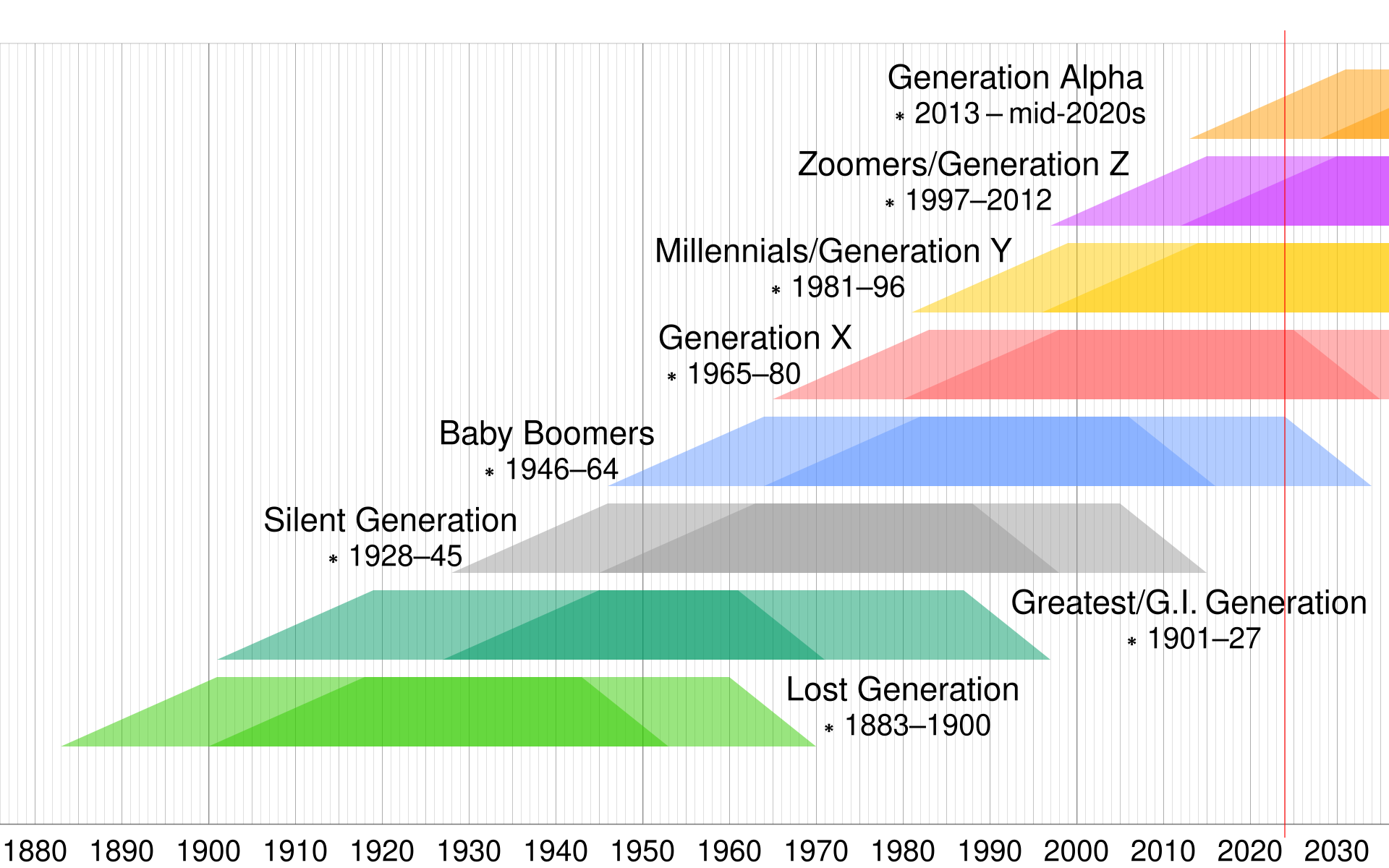
Anyways, i added the CSV Files also to the folder.

# Ideas for research topic

* **Marriage / Divorce status in comparison to birthdates / signs of zodiac.**
  + I suppose there will be hundrets of thousands of similar things already out there. Maybe we should be more creative and add some New / Extra knowledge? Therefore i want to extend this idea a little further: We go only one step further and **compare 2 different astrological believe systems** on their predictability of lasting marriages. What i mean?
  + The **Mayas** have the **Tzolkin calendar** <https://en.wikipedia.org/wiki/Tzolk%CA%BCin> It’s a *very easy mathematica*l translation from christian signs of zodiac to Mayan Sun signs and Moon signs (=“tones”). The big difference is that christians have only 12 signs, while Tzolkin calender knows 20 sun signs times 13 Moon tones. This goes to 260 different ones. Probably you know this symbols from some Indiana Jones Adventure Movies. ;-) 
  + Therefore my research proposal and question would be:  
      
    ***Is the Zodiac System more reliable in predicting long lasting marriages over the Tzolkin calendar?***
    - “Because Zodiac is more fuzzy with only 12 different signs” and is by far not so precise like a system with 260 different signs.
  + More resources: Find your own Maya Sign here: <https://www.mayansynchrony.com/tbuilder-layout-part/kin-calculator/>
  + In my case, i would be a **Red crystal earth**, that looks like this:

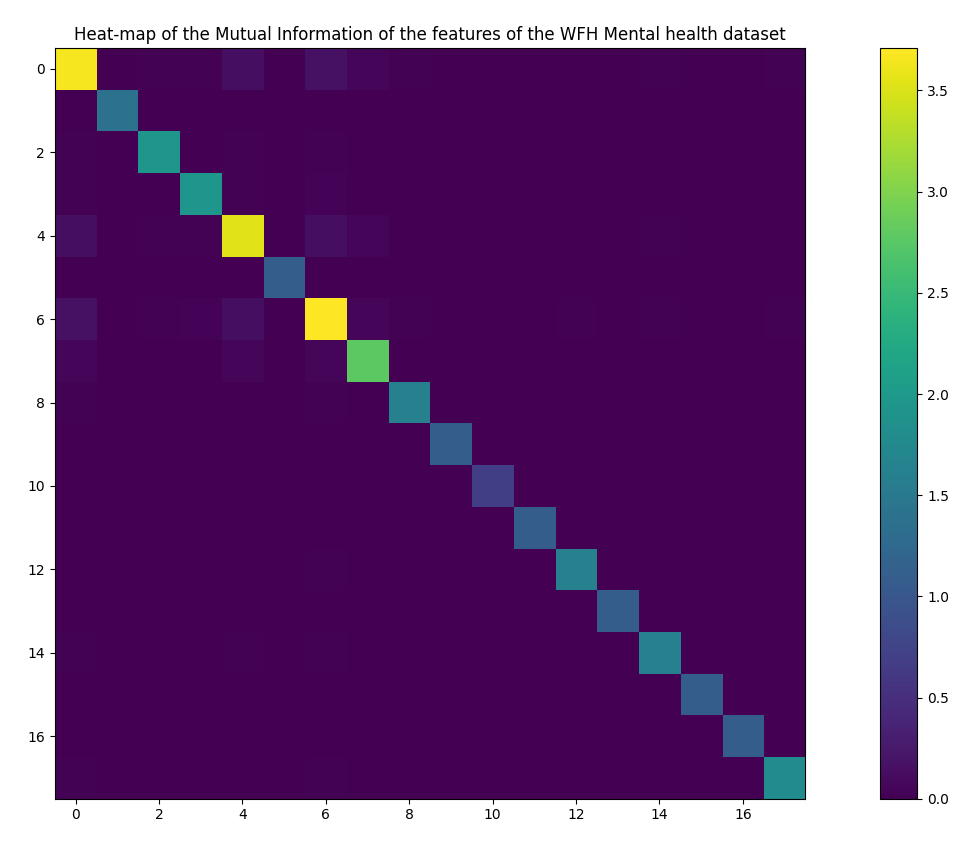
[PF] - Nice idea of using the Maya zodiac to make things more interesting. I have a few comments on the dataset and the problem definition. I don’t want to be negative guys, just trying to understand whether we have the right dataset at hand for the assignment:

* The initial idea was to compare the DOBs of married couples with the ones of divorced couples, to identify and then predict whether a combination of 2 zodiac signs will more probably lead toward a happy marriage or not. However, the Mexican dataset has only divorced couples, so it cannot be used for the initial idea.
* Based on what I have said above, it is good to change the problem statement, however in order to understand whether the Maya zodiac works better than the Western one to predict a happy marriage, we still need to have data regarding happy marriage and the dataset includes only divorced couples.
* We could change the problem statement in order to analyze only divorces, for example: what combination of zodiac signs is more likely to lead to a divorce. This is the analysis done by the author of the kaggle page (<https://www.kaggle.com/datasets/aagghh/divorcemarriage-dataset-with-birth-dates> , and, at the end of the day, requires only a statistical description of the data. I am not sure whether we would need to use a ML model to answer/predict this problem statement.
* [DN] Dear Paolo: I am happy with all these suggestions. Just to note, that your linked dataset is apparently exactly the same, as i stated above, isn’t it? I thought it’s just a clone / copy of the other data. → [PF] yes you are right Danou -> [DN] Hi, also online, right now? works well coop-mode. ;-) -> [PF] yep, let’s try to talk about it later on today, after the class.
* [OG] Just to leave in writing that with this data we might also be able to extrapolate the difference in relationships between generations and perhaps even cross-generation marriages/divorces, but we need more data to have a conclusive outcome i fear



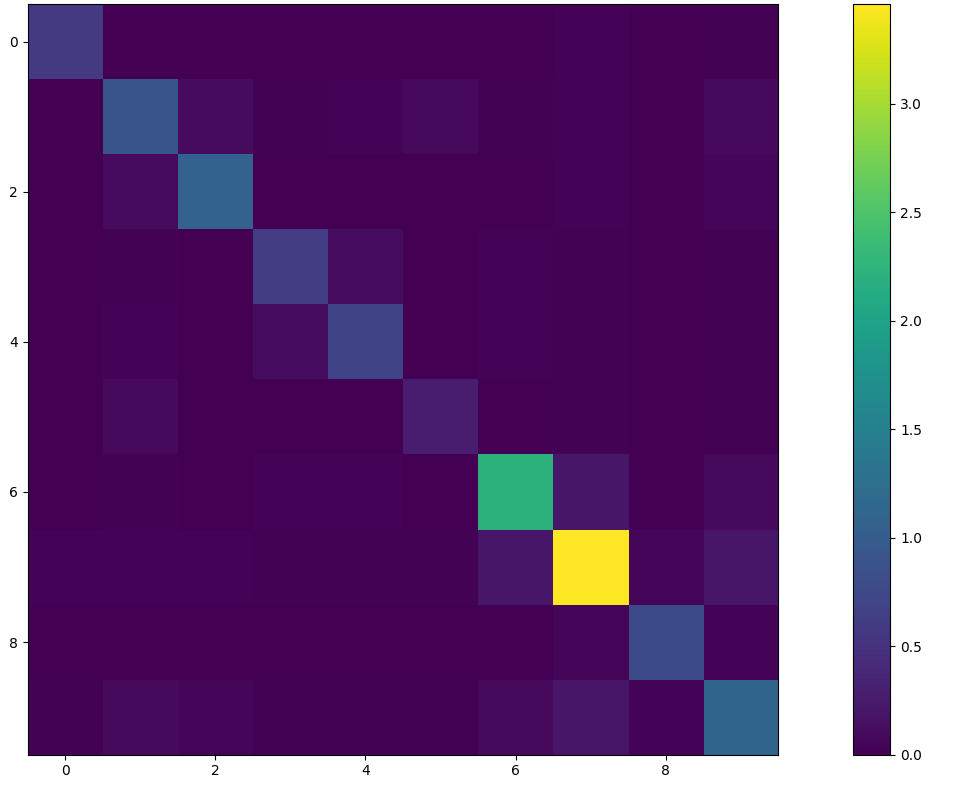
* 6/11/24 - Hi guys, this is the dataset I had talked you about, the one regarding mental health and working remotely: <https://www.kaggle.com/datasets/waqi786/remote-work-and-mental-health>

However, I have checked and this dataset doesn’t have any meta-data, which is not promising. So, I have performed the Mutual information analysis among all the features and the Mutual Information analysis between features and target variable (Mental helath), and I have found out that:

* the features are ALL NOT dependent between each other - which is good (see below heat map)
* however, none of the features has a relationship with the target variable, which is NOT good! This is the mutual information vector of the features against the target: [0.01 0. 0. 0. 0.01 0. 0.01 0.01 0. 0. 0. 0. 0. 0.

0. 0. 0. 0. ]

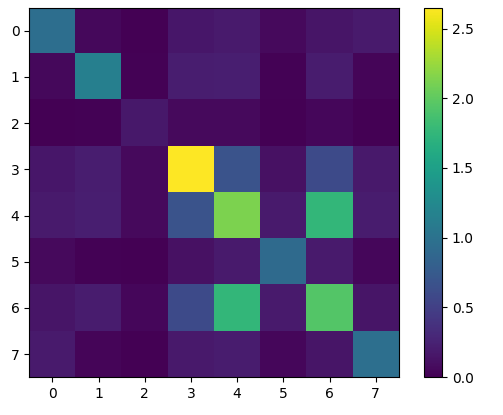
so I don’t think this dataset can be used for our purposes.

* [OG] Too bad about this, it was really an interesting study
* I have found another interesting dataset: <https://data.mendeley.com/datasets/mmnzx4w8cg/1> that is about mental health and IT jobs, however it is NOT about remote work. The dataset is supported by raw data from where a summary dataset is provided. The data goes from 2017 to 2021, so it includes the covid period too and it has meta-data… It seems very reliable. I have analyzed the features and even for this dataset the features are independent (good): 

and this time there are a few features that have a relationship with the target (Mental health). This is the mutual information vector of these features: [0. 0.02 0.01 0.05 **0.06** 0. **0.06 0.06** 0.01 **0.06**]. Please, let me know if this dataset seems good to you for our project.

* [OG] this seems very promising, but i’m still not sure what the end result or perhaps type of prediction we can achieve from it. It looks like this is more a categorization task sort of. Am I correct? or perhaps i am not understanding the idea behind this analysis. Is the end goal to show if an IT Job is mentally good or bad, or stressful or not stressful? Perhaps we can show the difference of the gender in the IT world?
* [PF] Hi Owen, the aim would be to build an ML module that learns from this data and can predict whether a person with specific characteristics (features) is more prone to have mental health issues.The more promising features for this prediction seems to be:
  + whether you work in an environment where it is welcome to talk with coworkers about mental issues
  + your age
  + the country

It doesn't seem that gender makes a difference, based on the above mutual information vector (0,01 of dependency), but maybe with further analysis we could find some differences there too.

* [PF] Hello guys, I am searching for other options for the datasets… they were right when they said that the selection of the dataset is important but really hard! Anyway, I have identified this dataset that seems interesting. It is about the Data Science job salaries. I have looked into it because it might be of interest for us. I have found the dataset here: <https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries> . There are about 600 entries, so not so many, and 8 relevant features (I have removed 2 features, salary in the employee currency and employee currency, because there is the feature salary in USD that can be used instead). I have analyzed the features and they are are independent (good), there is only a dependency between the employee country and company country, so maybe we could remove one of them:
* 

This is the mutual information vector of these features against the target/label, that is the salary in USD: [0.74 0.85 0.13 **2.15 1.93** 0.68 **1.79** 0.7 ]. The most relevant features in order to predict the salary are the ones in bold, which are respectively, the job title, the employee residence and the company residence. Strangely enough, the experience doesn’t seem to be so correlated with the final salary. Please, let me know if this dataset seems good to you for our project. T

<https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data>

data analysis

<https://www.kaggle.com/code/tawejssh/data-jobs-analysis-plotly-tutorial>

* [OG] Hi Fidipao and all, from the heatmap above all the features you highlighted seem useful so this is a potentially good datasource, I just wished it was a little bit more populated. 600 entries seems a bit too few for accuracy’s sake. But if we all like this dataset, we should use it and get on with it.  
  On another note, me and Ivan were talking and he seems to have some other good proposals as regards to dataset ideas.I will let him present them to you later today for your consideration.

# 2 more Datasets from Ivan

Olympia medals:

<https://www.kaggle.com/code/krantimohite/country-medals/input>

UNHCR

<https://www.unhcr.org/refugee-statistics/download?data-finder=on&data_finder%5BdataGroup%5D=displacement&data_finder%5Bdataset%5D=population&data_finder%5BdisplayType%5D=totals&data_finder%5BpopulationType%5D%5B%5D=REF&data_finder%5BpopulationType%5D%5B%5D=ASY&data_finder%5BpopulationType%5D%5B%5D=IDP&data_finder%5BpopulationType%5D%5B%5D=OIP&data_finder%5BpopulationType%5D%5B%5D=STA&data_finder%5BpopulationType%5D%5B%5D=HST&data_finder%5BpopulationType%5D%5B%5D=OOC&data_finder%5Byear__filterType%5D=range&data_finder%5Byear__rangeFrom%5D=2018&data_finder%5Byear__rangeTo%5D=2024&data_finder%5Bcoo__displayType%5D=doNotDisplay&data_finder%5Bcoa__displayType%5D=doNotDisplay&data_finder%5Byear__%5D=&data_finder%5Bcoo__%5D=&data_finder%5Bcoa__%5D=&data_finder%5Badvanced__%5D=&data_finder%5Bsubmit%5D=>

Covid

<https://ourworldindata.org/coronavirus>

## Millions new Datasets, high quality free to use from the EU

<https://data.europa.eu/data/datasets?locale=en>

# PROTOKOLL / Assignment Diary

This part is quite important, as we have to present 4x i think during mid term our current status-quo. Thus we should collect ideas here. Please feel free to add info to this document.

## Preparation Midpresentation 1

What happened so far?

* Group forming, Finding and discussion first thoughts
* Preparing infrastructure, WhatsApp Group, Google Drive Folder, LaTeX Template formatting and prefilling
* Evaluating and further adapting Research Ideas

## Preparation Midpresentation 2

## Preparation Midpresentation 3

## Preparation Midpresentation 4

# Formatting LaTeX, add ons:

\def\section{%

\@startsection{section}{1}{\z@}{+14\p@ \@plus 14\p@ \@minus -2\p@}% GM

{4\p@}{\baselineskip 14pt\secfnt\@ucheadtrue}%

}

\def\subsection{%

\@startsection{subsection}{2}{\z@}{+10\p@ \@plus 12\p@ \@minus -\p@}

{4\p@}{\secfnt}%

}

\def\subsubsection{%

\@startsection{subsubsection}{3}{\z@}{+8\p@ \@plus 10\p@ \@minus -\p@}%

{4\p@}{\subsecfnt}%

}

# Project implementation plan - what we have to do and how we can split and organize work

[Thanks man, this is a handsome Table, however made this!]

| ID | Task | Description | Assigned to | Status |
| --- | --- | --- | --- | --- |
| 1 | Data selection and extraction | After considering the proposed datasets we landed on the data science job salary dataset: [https://www.kaggle.com/datasets/ruchi798/data-science-job-salahttps://www.overleaf.com/project/671b2c1521d98e279ab53de5ries/data](https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data) | **Together** | Done |
| 2 | Data preparation | Conduct data cleaning, transformation, and feature engineering to prepare the data for ML. | Paolo |  |
| 3 | Implement and Evaluate Three ML techniques on the data | ML technique selection for each member | **Together** | Check this great article for the best python libraries:  [The Complete Guide to Python AI Libraries](https://blog.hubspot.com/website/python-ai-libraries#the-best-python-ai-libraries) |
| 3.1 |  | Implementation of the selected technique | each of us |  |
| 3.2 |  | Selection of the metrics to be used to measure the performance | **Together** |  |
| 3.3 |  | Evaluate each implemented ML performance | each of us |  |
| 3.3.1 |  | Include ethical consideration in each evaluation | each of us |  |
| 3.4 |  | GitHub repository creation and management | https://github.com/DocDanouBot/ICS5110 | Invitations Send, Struture established |
| 4 | Ethical review | Investigate and document potential biases, data privacy concerns, and other ethical considerations related to the data and its source. | E????? |  |
| 5 | Web tool | Develop a web-based interactive tool, using GitHub Pages, that allows users to interact with the model and see its predictions using tools such as Gradio | W????? |  |
| 5 | Documentation | 20 Pages (Maximum) Document in the IEEE Access format | ALL |  |
| 5.1 | Intro | Explain the properties of the chosen data set and plan of analysis b. Mention the machine-learning techniques that you will be using | Paolo |  |
| 5.2 | Background | Demonstrate and describe the mechanics of the selected machine learning techniques. b. Describe what rescaling and normalisation are and why they are important. c. Describe what cross-validation is and how (if applicable) it was used. d. Describe what dimensionality reduction and feature selection methods are e. Explain the quantitative measurements that you will be using to quantify the results; e.g. accuracy rate | Paolo |  |
| 5.3 | Data preparation | Describe the steps that you used to process the data set b. Discuss and explain the data cleaning, transformation, and feature engineering process to prepare the data for ML. | Paolo |  |
| 5.4 | Experiments | Describe the experiments that you carried out b. Describe the implementation of the ML techniques chosen. d. For each technique, assess individually how a modification in the parameters or cost function can affect the output with respect to a particular demographic. | each of us |  |
| 5.5 | Experiments comparison | c. Compare the results and insights from these models using the appropriate metrics | **Together** |  |
| 5.6 | Ethical Review | See point 4 above | E????? |  |
| 5.7 | Web Portal Usage Guide | Provide a brief user guide (maximum 2 pages) on how the data and insights can be explored from the web portal. (see point 5 above) | W????? |  |
| 5.8 | references and list of resources used |  | each of us |  |
| 6 | Generative AI Journal Guidelines | I will prepare a specific shared document with the sections required where each of us can input how he has been using Gen AI to support the implementation of the project : <https://docs.google.com/document/d/1OYasAGGiOdcpnwj0BQFHTR_HgXUwIvYkn92UqCV6fMk/edit?tab=t.0> | ALL |  |

**Data preparation**

The original dataset contains a total of 607 entries.

These are the fields of the original dataset.

| **Column** | **Description** |
| --- | --- |
| work\_year | The year the salary was paid:  - 2020  - 2021  - 2022 |
| experience\_level | The experience level in the job during the year with the following possible values: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director:  - EN  - MI  - SE  - EX |
| employment\_type | The type of employment for the role: PT Part-time, FT Full-time, CT Contract, FL Freelance: |
| job\_title | The role worked in during the year. |
| ~~salary~~ | ~~The total gross salary amount paid.~~ |
| ~~salary\_currency~~ | ~~The currency of the salary paid as an ISO 4217 currency code.~~ |
| salary\_in\_usd | The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com). |
| employee\_residence | Employee's primary country of residence in during the work year as an ISO 3166 country code. |
| remote\_ratio | The overall amount of work done remotely, possible values are as follows: 0 No remote work (less than 20%), 50 Partially remote, 100 Fully remote (more than 80%) |
| company\_location | The country of the employer's main office or contracting branch as an ISO 3166 country code. |
| company\_size | The average number of people that worked for the company during the year: S less than 50 employees (small), M 50 to 250 employees (medium), L more than 250 employees (large) |

**Cleaning**

Duplicates

- There are 42 duplicated entries. Duplicated data points might be:

1. Correct data that happen to be the same for more than one occurrence, in this case these true points should be taken in consideration since they are part of the overall phenomenon under analysis

2. Actual duplications of data points that have been incorrectly added to the dataset, in this case these data points must be dropped because they don’t add any information to the model, and they can skew the model reinforcing specific behaviours

It is NOT possible to determine which of above scenarios we are facing, therefore, better to keep a conservative approach, go for option 2) and remove duplications. The cleaned dataset has 565 unique entries.

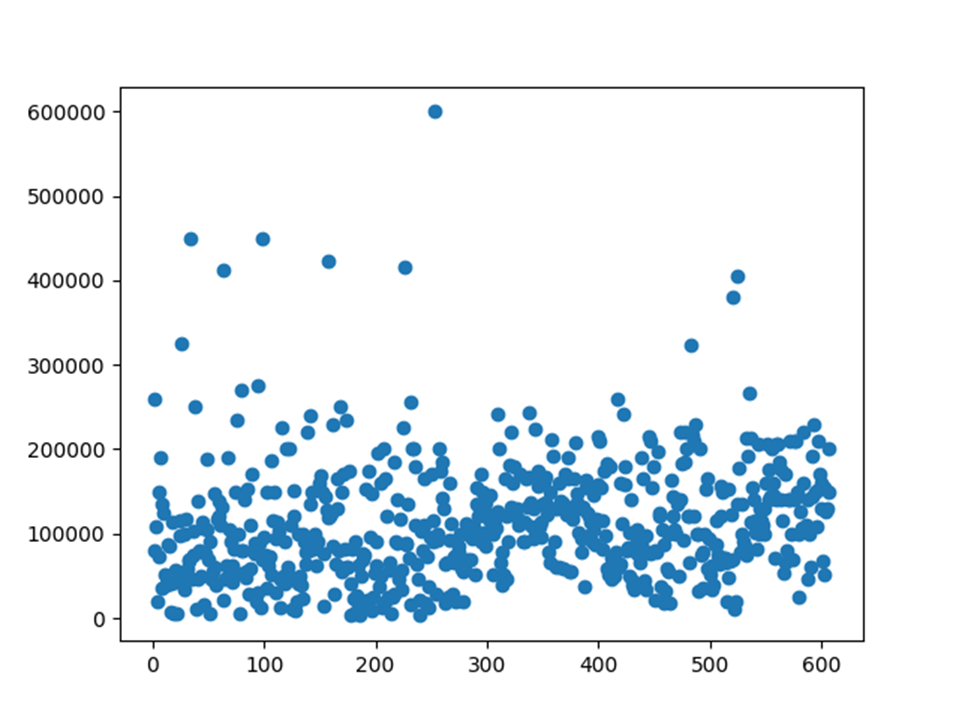
- *Salary* and *Salary\_currency* features have been removed because the *Salary\_in\_USD* feature already represents the same information and allows an homogeneous comparison among entries.

Omissions

The dataset doesn’t have any omission of features.

Balance

Usage



The dataset can be used for:

1. Predicting the annual salary in USD for jobs under the Data science umbrella – REGRESSION
2. Predicting the range of annual salary in USD jobs under the Data science umbrella – CLASSIFICATION
   1. A new feature (*Salary\_range*) will be introduced to enable this category of ML analysis
3. Finding the rules of classification and the number of classes of the dataset - CLUSTERING

Transformation

- Group in salary categories for classification problem

Feature engineering

- Features dependencies



- Mutual information vector between each feature and the target:

o work year = 0.74

o experience level = 0.85

o employment type = 0.13

o **job title = 2.15**

o **employee residence = 1.93**

o remote ratio = 0.68

o **company location = 1.79**

o company size = 0.7

- Grouping for classification

The dataset can be split in 3 categories of salaries, High, Medium and Low. The ranges of each category can be defined to have a balanced distribution:

· The 33th percentile of the distribution is 76,940 USD

· The 66th percentile of the distribution is 130,000 USD

The *salary\_group* feature is added to the dataset based on the following rules:

· Low salaries (L) are the ones below 76,940 USD

· Mid salaries (M) are the ones between 76,940 USD and 130,000 USD

· High salaries (H) are the ones above 130,000 USD

# Possible Research Questions:

1. What is the average data scientist salary?
2. Do startups pay data scientists more than large companies?
3. Does job title has a role to play in a data scientist salary?
4. Does your location change the numbers on your paycheck?
5. How country can impact the salary?
6. What is the most popular job title in data science?
7. Which country has better opportunities for data science jobs?
8. How much do remote & in-office employees get paid?
9. How many employees at each seniority level?
10. How many employees work remotely?
11. Which country pays the highest for an entry-level Data science professional?
12. Which option (job title) should an entry-level pick?
13. Which type of employment is best for an entry-level?
14. Which company size is right fit for an entry-level?
15. Should an entry-level try remote work or not?

[PF] [danou.nauck.24@um.edu.mt](mailto:danou.nauck.24@um.edu.mt) I like all of them, however, if we agree, some of them are descriptions of the dataset (1, 6, 7, 9, 10, 11) and we can put them in the Data collection and analysis section,that I would like to work on. The rest of the points could be written in the Research objectives sub-section section of the Introduction section, in this form:

“*Research Objectives*

1) To identify the features that directly influence the salary amount/category and reply to the following research questions:

a. Do small companies pay data scientists and machine learning engineers more than large companies?

b. Does the job title have a relevant role in determining the salary?

c. Are remote opportunities paid less than on-site/hybrid ones?

d. Which type of employment and which company size are better for an entry level?

2) To identify the optimal prediction model for predicting the salary amount or the salary category among selected participant segments

”

What do you think about it?