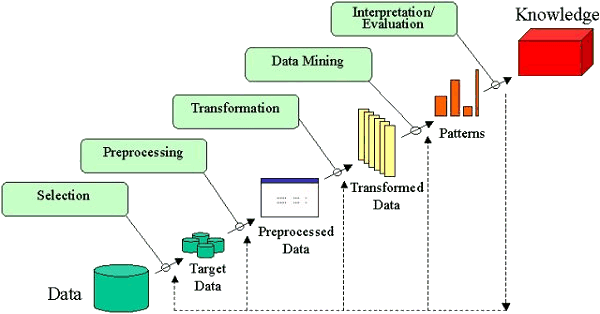
Annex 5: Nurse Salaries Project Plan

This is a checklist of all the items required for the project. The list has been organized in phases, and each phase has to be completed before the next.

Items highlighted in color are requirements of the project that will be evaluated.

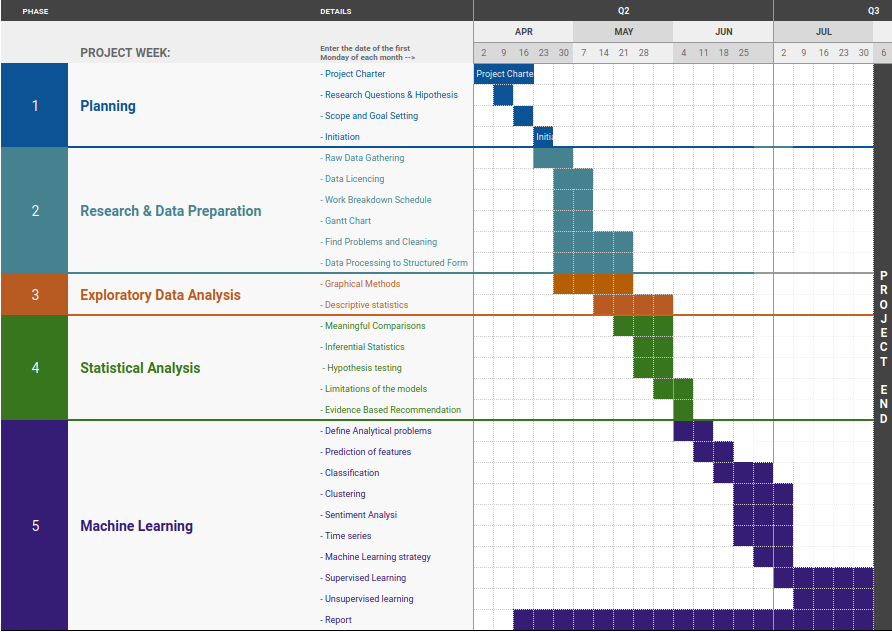
### Project Management

* PM framework selected: Classic Linear Cascade (Alternatives: Agile, Scrum, Adaptive, Extreme Programming)
  + Justification: Time and Human resources, but an Agile strategy will be better.
* Data Science framework selected: Knowledge Discovery in Databases or KDD (Alternatives: CRISP-DM, SEMMA)
  + Justification: Linear, Less reprocessing than other strategies, due to Tiem and Human resources limitation



Source: <http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/kdd.gif>

* Developing environment: Jupyter Local. (Alternatives: Google colab, Kaggle, Deepnote
  + Justification: Time and Human resources limitation. No need for collaboration. (Source: <https://noteable.io/blog/jupyter-notebook-alternatives/>)
* Git Versioning: Only uploading finished versions of datasets and notebooks.
  + Justification: Takes too much time and risk to learn to use Git efficiently
* Final Report 3000 words
  + Use a word counter <https://wordcounter.net/>
* Timeline (Gantt Chart)



# Work Breakdown Structure

## Module 1. Data Preparation

* 1. Planning (This Document)
     1. Research Questions
  2. Raw Data Gathering
     1. Data Licencing
     2. Load Datasets
  3. Find Problems and Cleaning
     1. Encoding and file format
     2. Consistency
     3. Relevance
     4. Data Augmentation and enrichment
  4. Data Processing to Structured Form
     1. Handling Outliers

### Programming Learning Outcomes

* Documentation
* Multiple Data Structures (CSV, SQL, API, JSON -> Series, dataframes)
* Aggregation techniques / methods (.agg .groupby .aggregate .filter .transform .apply)
* Test and Optimization strategy
* Compare, Contrast and choose processing libraries (Numpy, Pandas, NLTK, Sklearn, Scipy, seaborn)

## Module 2. Exploratory Data Analysis (EDA)

* 1. Graphical Methods
     1. Identify Issues
     2. Critique the Dataset
  2. Descriptive statistics
  3. Evaluate hypothesis
  4. Conclusions: Meaningful Comparisons between countries (Similarities, disparities)
     1. Biased comparisons and Causal relations
     2. Confounding variables
     3. Correlations

### Visualization Learning Outcomes

* Interactive Visualization solution (Dashboard) (Streamlit)

## Module 3. Statistical Analysis

* 1. Inferential Statistics and Hypothesis testing
     1. Point of view and Scope,
        1. Research Questions
        2. Research Hypothesis
        3. Target Variables and Independent Variables
     2. Applicability of tests (T-test, Anova, Wilkinson, Chi-Squared)
  2. Parametric Tests
  3. Non Parametric Tests
  4. Critique Limitations of the models
  5. Evidence Based Recommendation
     1. Employment
     2. Earnings

### Statistics Learning Outcomes

* Applicability of tests (T-test, Anova, Wilkinson, Chi-Squared)

## Module 4. Machine Learning

* 1. Machine Learning strategy
     1. Define Analytical problems
     2. Scoring Metrics
     3. Choose Appropriate Features
     4. Target Feature
  2. Supervised Learning
     1. Which data is labeled?
     2. Check instructable images
  3. Unsupervised learning
     1. Prediction of features
     2. Forecasting (prediction of future features)
     3. Classification
     4. Clustering
     5. Sentiment Analysis
     6. Time series
     7. Recommender Systems

### ML Learning Outcomes

* Hyperparameters tunning
* Test and compare 2 approaches
* Provide graphical comparison of accuracy and predictive power of ML algorithms
* GridSearchCV
* Cross Validation to check authenticity of model outcomes
* Dimensional reduction if needed

# 

# Notes During Execution

## Module 1.

The report has to look like a formal report, the idea is to find insights and write them, supported by charts, or formal tests. I will plot more charts in the morning.

**Backlog work and concerns:**

### Items to address :

I already copied the introduction and references to a new file: “Report”

* must fix it and complete it. Proofread it. **I did it yesterday and I will fix it properly later (done)**

it will look very professional if the source is well documented: have the data about the dataset:

* Fill the first sheet of [this](https://docs.google.com/spreadsheets/d/1sWYxUMNiccddqLqbPVHkeVHqe-pzgg9YgfSAfv4rRws/edit?usp=sharing) with the information about the source.

I figured out that the dataset is composed of 6 different measures (Units), all sorted by country, year, profession.

* Flags mark differences in methodology. Are we able to assume that the data with different flags is comparable?
  + What is “Break”
  + ~~The differences are from country to country. We can compare countries with confidence if they have the same methodology~~
  + ~~Make a plot with the flags to see visually where the data is having a different methodology~~
* Do we need a dataset of conversion of currencies to the dollar, for all years?
  + ~~No, Salaries have already been converted to the dollar. Some data is missing~~
* Is all the data yearly?
  + ~~It looks like all data is yearly. Some data is missing: I use Interpolation~~

I filled in the incomplete data using interpolation.

* + It works most of the time, but introduces bias. Try to fix the bias
* We need to understand the variable “UNIT” / “Measure”:
  + ~~NCU means National Currency Unit (dollar, pound, etc)~~
  + ~~NCU at 2015 GDP price level is Good to compare over time, but have to be converted to dollar to be comparable between countries~~
  + Income per average wage could potentially refer to the average income earned per unit of average wage

It compares how much outside wages people need to live?

Need more detail about how it was calculated

* + Income, /capita GDP is the income compared with the average income earned per person in the country

compares how better is the income compared to average person.  
can be used to compare countries over time

* + Income, US$ exchange rate is useful, but it is not comparable over time
  + Income, PPP for private consumption, in US$ refers to the Purchasing Power Parity conversion factor, the number of units of currency required to buy the same goods and services.  
    useful to compare real purchase power of salaries in country
  + Are some of the salaries standardized / normalized?
    - Yes, /capita GDP
    - Yes, income/average wage
    - Income averaged is different interpretation than income per capita
    - Is there any other way to normalize salaries that makes sense?
  + ~~Some salaries are so low, are they reported monthly, yearly, weekly, daily?~~ 
    - No, they are rates.
* We need to compare salaries using adjustment to present value to the last year available, There are many techniques: (prices inexpensive, inflation
  + ~~No, they are already NCU at 2015 GDP price level, assess if data is missing~~
  + ~~Data is adjusted, but not to the dollar, so they are only comparable within the country. Already create a new variable for this.~~
* Self employed is different than salaried
  + ~~There are two populations. Useful for inferences and comparisons~~
* Different specialities have to be compared separately
  + ~~Useful for comparison and test hypothesis~~
* In order to compare features, we need to cluster (ML) countries in regions… or we can use an ad-hoc classification to start
  + Cluster by differences in methodology
  + Cluster by GDP
  + Cluster by… [Human Development Index](https://en.wikipedia.org/wiki/Human_Development_Index) HDI or IHDI
  + Cluster by inequality, net wealth per capita…
* I create a new variable about employment type: salaried or self-employed.
  + ~~I also created a new variable that combines profession and type of employment~~
  + ~~I also rename “Country” variable values in order to include their ranking by 2022 GDP, so now the variable is more rich.~~
* I dropped the code columns, and the countries that had incomplete data.

**Final list of variables useful:**

Income, US$ exchange rate

Income at 2015 GDP price level, in US$

Income PPP for private consumption, in US$

Income / per capita GDP

Income per average wage

# I have finished notebook 1.

## Module 2. Items to address

Plots to make:

Here are some of the best plots to consider for each variable:

* Year (Categorical Variable):
  + Bar Plot: Use a bar plot to show the distribution of data across different years.
  + Line Plot: Display trends over time, especially if have sequential data points.
  + Histogram: If data is continuous, create a histogram to show the frequency distribution of salaries over the years.
* Salary (Continuous Variable):
  + Box Plot: Use a box plot to visualize the distribution of salaries, identify outliers, and compare salary ranges across different professions or countries.
  + Violin Plot: Similar to the box plot but also shows the kernel density estimation of the data, giving a better view of the salary distribution.
  + Scatter Plot: Compare salaries against other variables, like years, professions, or countries, to explore potential relationships or patterns.
* Profession (Categorical Variable):
  + Use as categories to compare the graphs listed above
* Country (Categorical Variable):
  + Use as categories to compare the graphs listed above

Additionally, consider using faceted or grouped plots to compare variables simultaneously. For example, create grouped bar plots to compare average salaries across different professions for each year, or a faceted scatter plot to compare salary distributions for various countries.

**I** finally cleaned and formatted the data. It took me a horrendous amount of time. I made 10 scatter plots and boxplots. They show the behavior of the 5 Main measures:

* Income converted to dollars
* Income adjusted for inflation
* Income adjusted for consumer prices
* Income vs average salary in the country
* Income vs average wage per person.

And I plotted these against 5 variables at the same time:

* Country
* GDP ranking of the country
* Profession
* Type of employment
* Time (Year)

These will bring enough information to extract many conclusions for the report. Without statistical testing it is evident the impact of profession, GDP, type of employment.

* Write conclusions about that… for example about the unfairness of the compensation, from the experience as a nurse.

# I have finished Notebook 2

## Module 3. Items to address

Applicability of tests (T-test, Anova, Wilkinson, Chi-Squared)

Rewrite the research questions with the things learned so far

Ireland is the base for the modeling

# I am in 3.1.1

## Module 4. Items to address

Ireland is the base…

Choose an easy problem.

* Fill missing data (predict a value that doesn't exist)
* Predict future behavior of the variables
* Cluster countries by similarity
  + Supervised
  + Unsupervised
* Find most similar country to Ireland