ECOLE INTERNATIONALE
DES SCIENCES DU
TRAITEMENT DE L'INFORMATION

EISTI

ADEOP1 ROJECT

Geographical statistics of engineering internships

Authors:
Gustavo FLEURY
Cylia BERKANE
Fagnuo CAI
INDURAJ PR
Quoc Viet PHAM
Yen Chu CHEN

Advisers:
Asma TALHI
Rachid CHELOUAH

Cergy - April 29, 2019

Contents

1 Introduction

EISTI - As an engineering school officially recognized by the State and made competent by the CTI (French Engineering accreditation institution) the EISTI's vocation trains future engineers in Mathematics and Computing. Having this in mind, our team choose to develop a system that could help the school gain more knowledge on the geographical statistics of the engineering internships chosen by student of both cergy and pau campus.

2 Background of the project

Every Educational institution have students going from the university to different parts of the world for internships to gain experience and explore career path. In EISTI's campus both Cergy and Pau, at least few hundred students relocate every semester to different geographical locations for the internships. But it is unclear as what's the choice of the students in choosing companies while applying for internships. It is thus crucial to analyze the internship preference of the students from the available data so as to get thorough insight on if the student preferred companies based on the distance between their home and campus or between their home and location of the company or between campus and internships location! The thorough insight on the student's preference and about the companies that hire the EISTIEN'S will help the university in many ways such as in adapting even more dynamic on-demand competitive edge curriculums, partnering with the companies for giving on-campus training sessions and seminars,etc. These insights can also be used to define a model which would predict as where the student is most likely to end up as intern!

3 General objectives

The general objective of the project is to build a tool by which we can visualize as what factor governs the most while the student decides to apply and go for the internship.

3.1 Specific Objective of the project

- Study and analyze the existing dataset
- Perform data cleansing and transformation
- Identify the basic functionality and non-functionalities of the system
- Design the front end (web based) and backend system
- Incorporating the front end and the back end with the database
- Implementing the designed system

4 Scope

To properly define the scope of this project we must understand the needs of the client, mainly the development of a desktop/web application that will allow users study and extract relevant information from a geographical database of EISTI students.

This application will have the following functunalities :

- Upload csv file and store it within a database.
- Automatically clean the data
- A menu to perform univariate and bivariate analysis and statistics that are defined in advance and can be configured.

- Distance Calculation between the different locations « home - campus », « home – internships location » and « campus – internships location ».
- User management (Login/Password /roles)

5 Technical solution

There are three major developing software that we use in our project, Python, Oracle SQL and Django. In the beginning of the research, the raw data is present in Excel file. Then, we import the data to python to do data cleaning and a general data profiling. After, we insert the data to the Oracle SQL. During the developing procedure, we use python again as data analysis software. At last, we choose Django do be our visualizing tool. With Django linking with the Oracle database, our final goal is to present the client dynamically analyzed relation model.

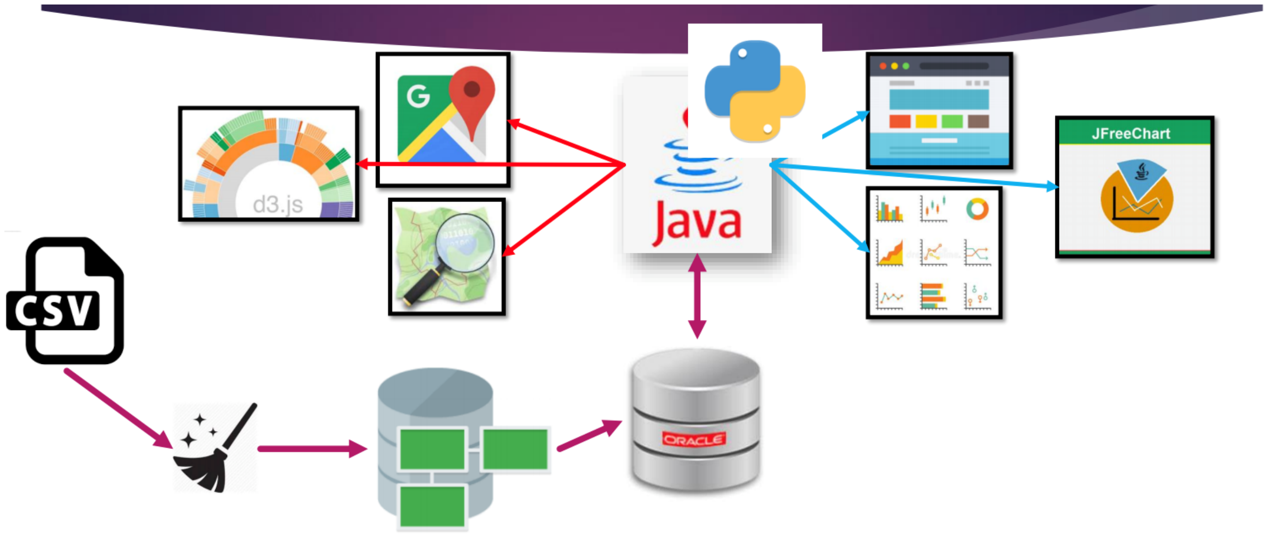


Figure 1: Technical Solution of the project.

6 Introduction of the developing software

6.1 Oracle

The oracle environment is a database that designated by the clients. The three major advantages with using oracle SQL developer is that it can manage not only oracle database, but also SQL server and MYSQL, and the second advantage is that it has the cross-platform execution capability with windows, Linux and Mac OX, and the last one is that it is an completely free resource.

6.2 Python

In this project, our team choose python over Java. For the first and the most important reason is that python is the mostly used software in the data science field. The essential of the python is that it simplified the complex coding grammar with fewer lines, and it is an open source that popular in business model. With more and more packages created, python has an comprehend function to analyze data. The the other reason for using python over Java is that python has a great flexibility that can easily connected with the web.

6.3 Django

7 Libraries and essential tools

7.1 NumPy

In scikit-learn, NumPy tables are the basic data structure. In fact, scikit-learn takes its data in the form of NumPy tables. All the data we use must be converted to a NumPy table. The central functionality of NumPy is the ndarray class, which is a multidimensional array (a n dimesions). All the elements of the table must be of the same type.

7.2 Pands

Pandas is a Python library for data manipulation and analysis. It is built around a data structure called DataFrame. Pandas provides a large number of methods to modify and process this table. Another important interest of pandas is that it is able to work with a large number of file formats and databases, such as SQL, Excel files or CSV. The most common data structures that we use in this project are pandas.Series (One-dimensional ndarray with axis labels) and pandas.DataFrame.

7.2.1 Attributes and methods in pandas.Series

Attributes/Methods	Description
.T	Return the transpose
.index	The index (axis labels) of the Series
.is_unique	Return boolean if values in the object are unique
.values	Return Series as ndarray or ndarray-like depending on the dtype
.str.upper()	Convert strings in the Series/Index to uppercase
.replace()	Replace values given in to_replace with value
.str.startswith()	Test if the start of each string element matches a pattern
.str.partition()	Split the string at the first occurrence of sep
.rstrip()	Remove trailing characters
.lstrip()	Remove leading characters
.isnull()	Detect missing values
.any()	Return whether any element is True, potentially over an axis

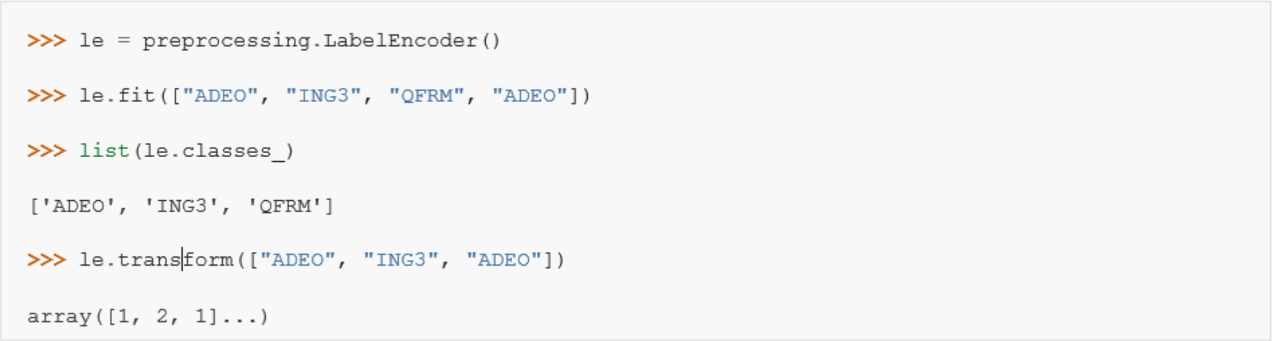
7.2.2 Attributes and methods in pandas.DataFrame

Attributes/Methods	Description
.loc	Access a group of rows and columns by label(s) or a boolean array
.iloc	Purely integer-location based indexing for selection by position.
.read_table	Read table

7.3 Sklearn.Preprocessing

The sklearn.preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.

The most common function we use is sklearn.preprocessing.LabelEncoder, which encodes labels with value between 0 and n_classes -1.



```
>>> le = preprocessing.LabelEncoder()

>>> le.fit(["ADEO", "ING3", "QFRM", "ADEO"])

>>> list(le.classes_)

['ADEO', 'ING3', 'QFRM']

>>> le.transform(["ADEO", "ING3", "ADEO"])

array([1, 2, 1]...)
```

Figure 2

7.4 sklearn.ensemble.RandomForestRegressor

A random forest regressor. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default). We use random forest regressor to fix missing values of attribute 'RENUMERATION' in table 'STUDENT_INTERNSHIP'

7.5 feature_selector. FeatureSelector

In this project we will walk-through using the FeatureSelector class for selecting features to remove from a dataset. This class has methods for finding features as:

Features	Description
Missing values	Find any columns with a missing fraction greater than a specified threshold
Single Unique value	Find any features that have only a single unique value
Collinear (highly correlated) Features	This method finds pairs of collinear features based on the Pearson correlation coefficient.

8 Front End - Functionalities for the User

9 Planning of the project

To organize the distribution of tasks and to meet deadlines of the project, we decided to use Scrum, an agile framework for managing the work. The idea of Scrum is to divide the project in a list of actions (backlog) that can be completed within time boxed iterations (sprints). Because of small time of all project, the sprints intervals are at maximum one week, and the daily scrums occurs before or after some normal class.

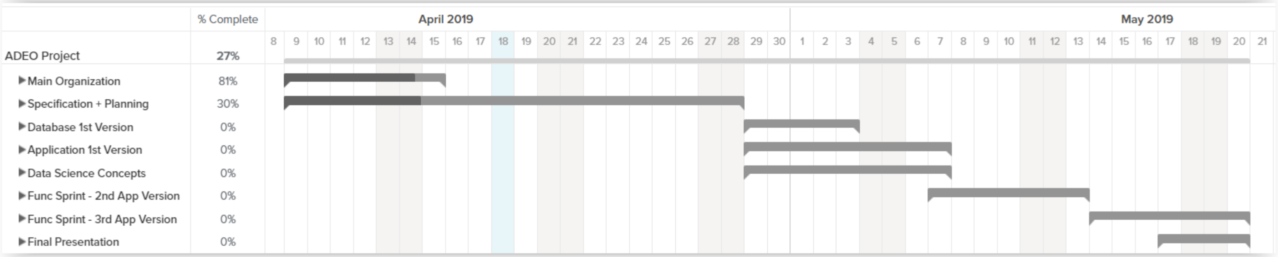
The tools chosen to assist in the execution of the scrum and in the follow-up of the project execution were Trello and TeamGantt. The Gantt graph is shown in the next topics.

To facilitate the exchange of information between the team, a group was created in chat tool. And the code is shared in github.

The project was divided in some groups, to evidence the deadlines of sub products and the application. Each group has some tasks, that could be linked to sub product. The following figures shows this groups, estimated time and deadline. The tasks will be chosen freely by team members, and dynamically allocated.

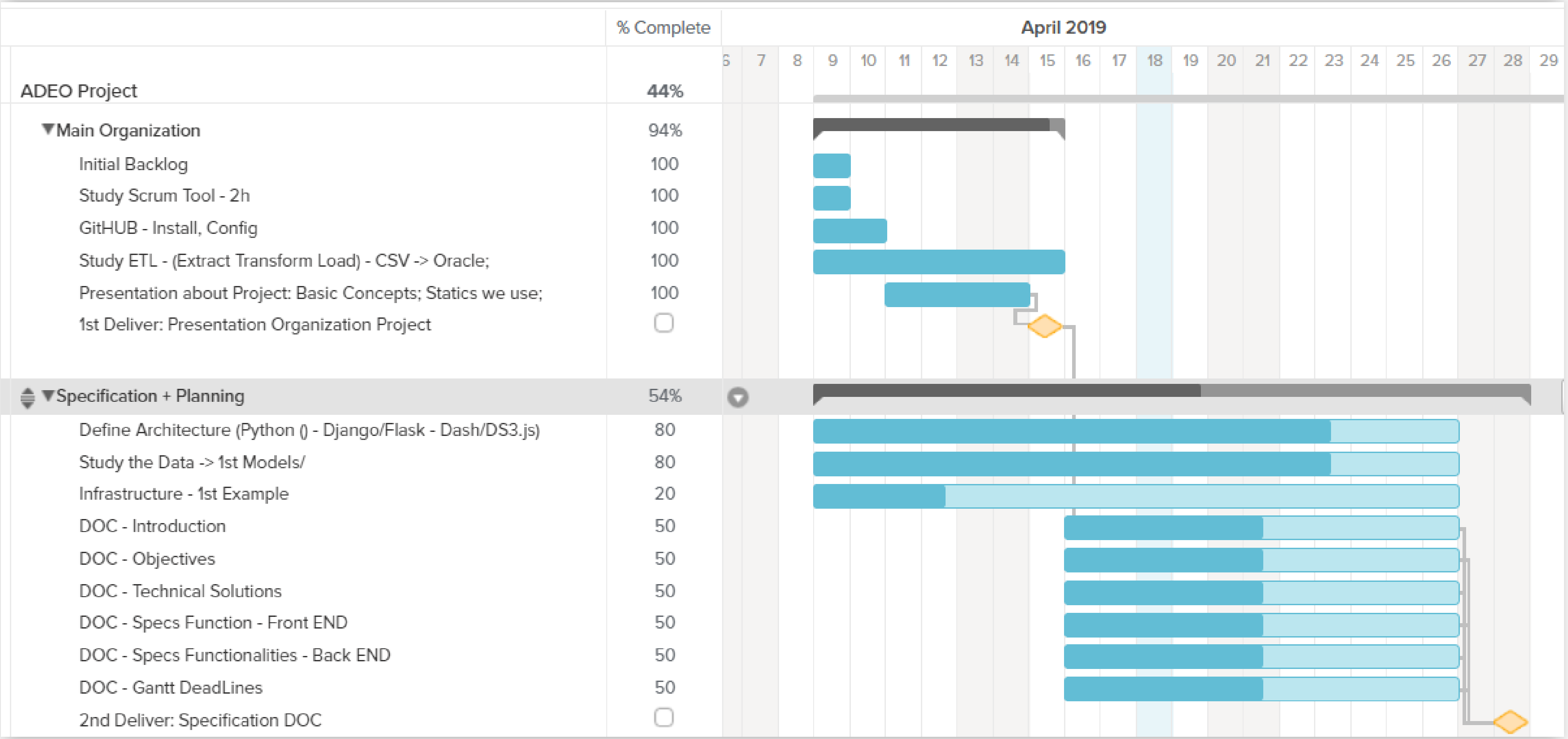
These initial tasks are just to estimate the functions we can add in the application, considering the deadline, number of team member and experience of team in the chosen technology. After define the additional functionalities, the backlog will be filled by the new tasks.

The second column in the Gantt figures shows the percentage of the execution of the product.



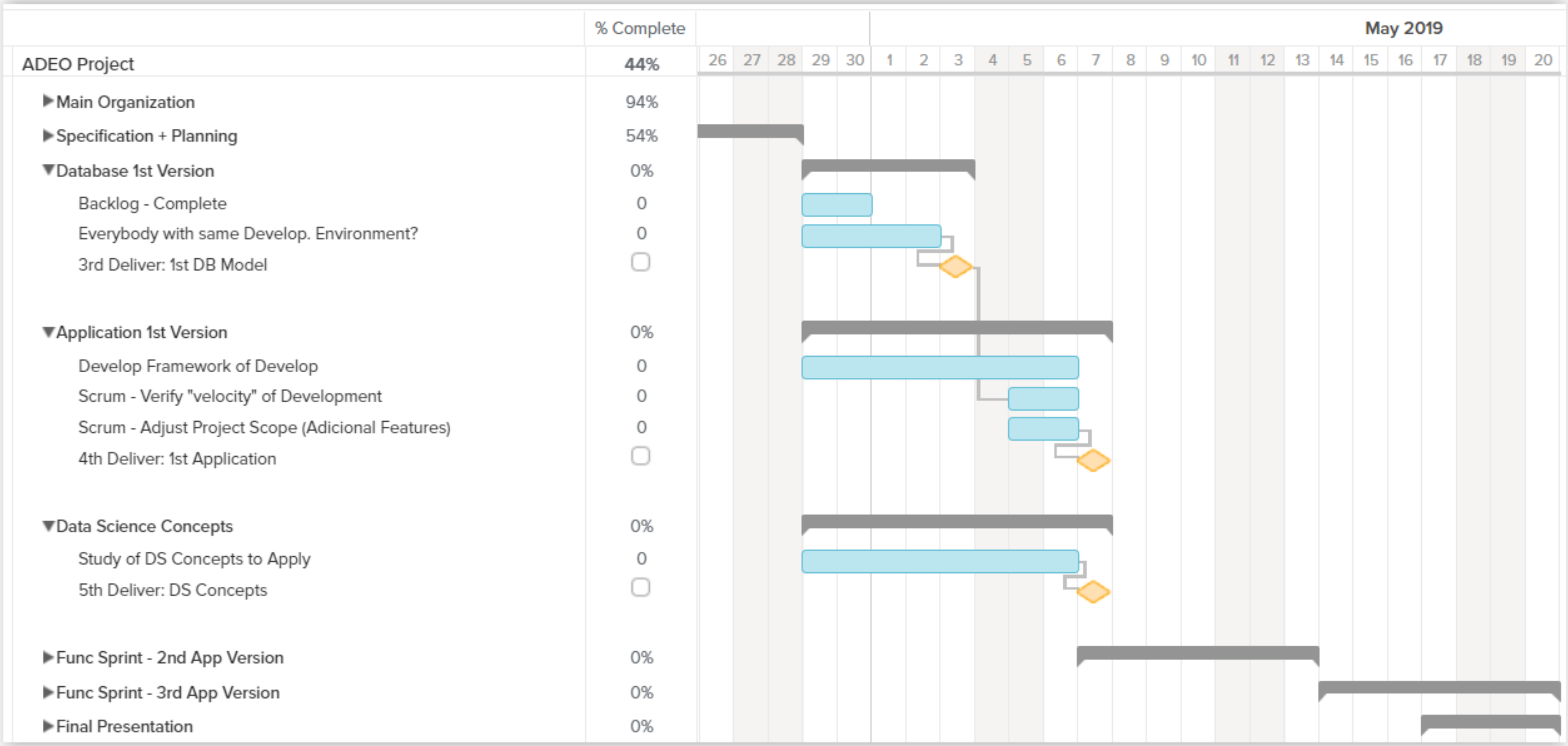
	% Complete
ADEO Project	**27%**
▶ Main Organization	81%
▶ Specification + Planning	30%
▶ Database 1st Version	0%
▶ Application 1st Version	0%
▶ Data Science Concepts	0%
▶ Func Sprint - 2nd App Version	0%
▶ Func Sprint - 3rd App Version	0%
▶ Final Presentation	0%

Figure 3: Schedule and current status of the final grade project.



	% Complete
ADEO Project	44%
▼Main Organization	94%
Initial Backlog	100
Study Scrum Tool - 2h	100
GitHUB - Install, Config	100
Study ETL - (Extract Transform Load) - CSV -> Oracle;	100
Presentation about Project: Basic Concepts; Statics we use;	100
1st Deliver: Presentation Organization Project	☐
▼Specification + Planning	54%
Define Architecture (Python () - Django/Flask - Dash/DS3.js)	80
Study the Data -> 1st Models/	80
Infrastructure - 1st Example	20
DOC - Introduction	50
DOC - Objectives	50
DOC - Technical Solutions	50
DOC - Specs Function - Front END	50
DOC - Specs Functionalities - Back END	50
DOC - Gantt DeadLines	50
2nd Deliver: Specification DOC	☐

Figure 4: Gantt for all project



	% Complete
ADEO Project	44%
▶Main Organization	94%
▶Specification + Planning	54%
▼Database 1st Version	0%
Backlog - Complete	0
Everybody with same Develop. Environment?	0
3rd Deliver: 1st DB Model	▢
▼Application 1st Version	0%
Develop Framework of Develop	0
Scrum - Verify "velocity" of Development	0
Scrum - Adjust Project Scope (Adicional Features)	0
4th Deliver: 1st Application	▢
▼Data Science Concepts	0%
Study of DS Concepts to Apply	0
5th Deliver: DS Concepts	▢
▶Func Sprint - 2nd App Version	0%
▶Func Sprint - 3rd App Version	0%
▶Final Presentation	0%

Figure 5: Tasks for "Database 1st Version", "Application 1st Version", "Data Science Concepts", "Functional Sprints" and "Final Presentation.