Міністерство освіти і науки України

НТУУ «Київський політехнічний інститут»

Фізико-технічний інститут

СРП

Лабораторна робота №1

## Наука про дані: підготовчий етап

**Виконав:**

Студент 2 курсу ФТІ

групи ФІ-42

Корольчук М.С.

**Перевірив:**

Колотій А.В.

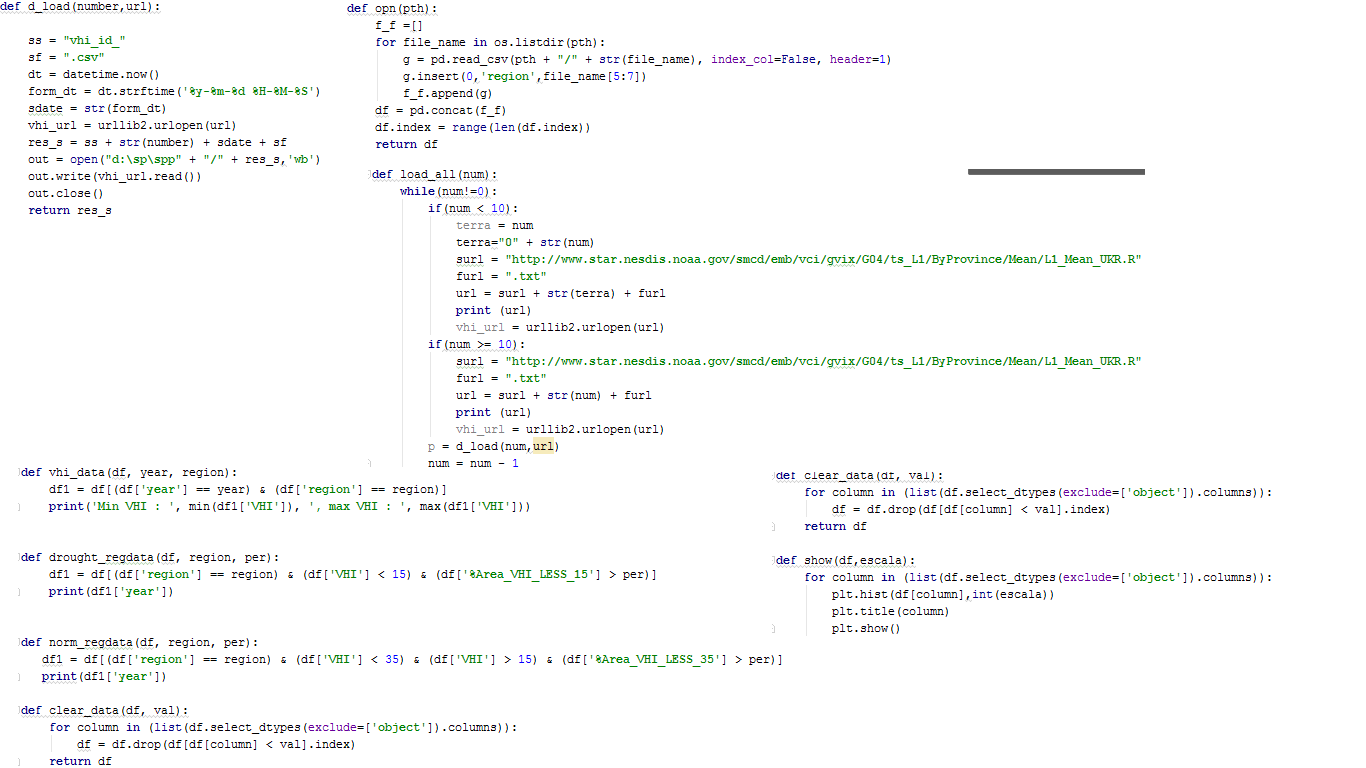
Київ 2016

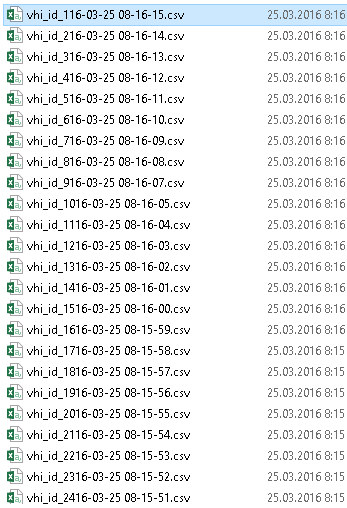
Мета роботи:

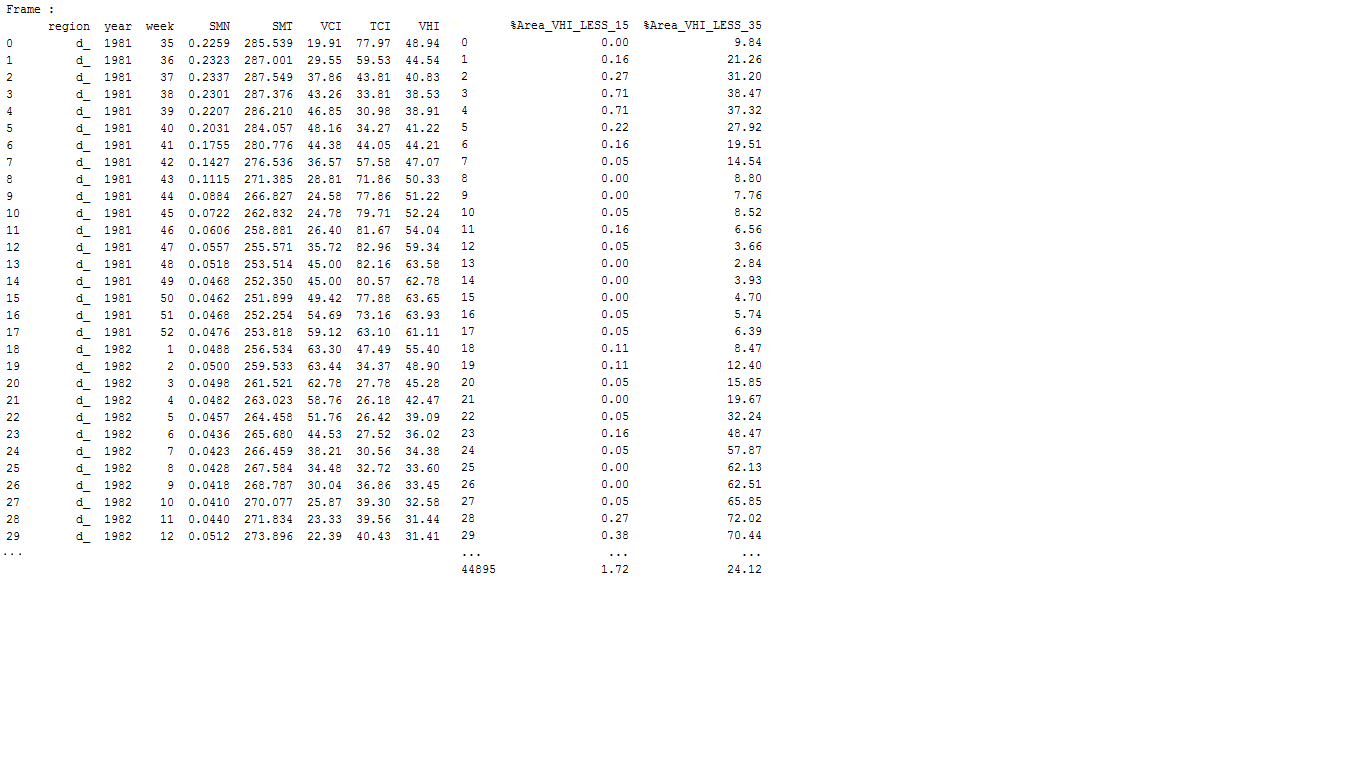
Навчитися здійснювати обробку даних для вирішення поставлених задач.

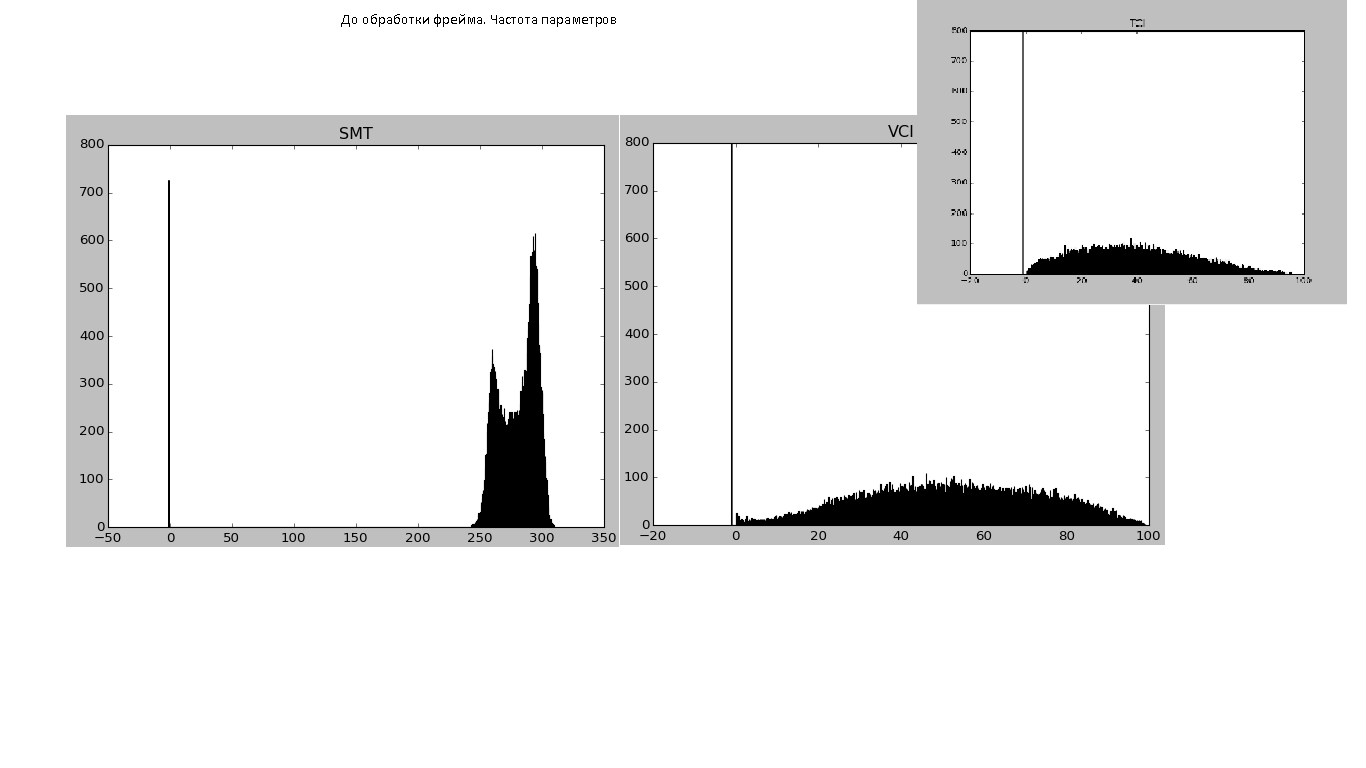
Хід роботи:

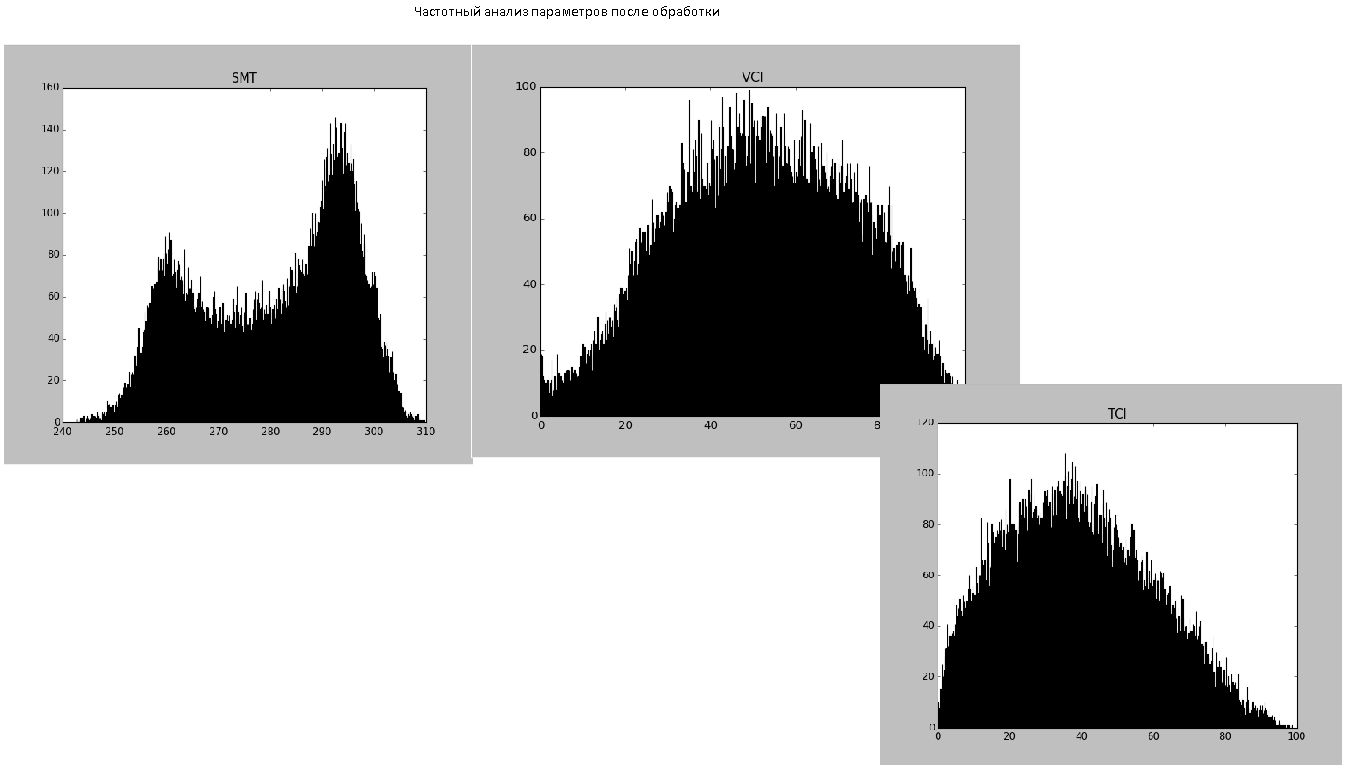
* Завантаження структурованих файлів зі вказаного джерела з вказанням дати завантаження.
* Зчитування даних у фрейм : 5e6 значень.
* Обробка фрейму таким чином, щоб спецсимволи, що були виявлені внаслідок графічного аналізу даних, та позначаючи місця відсутності даних не впливали на результати задач.
* Реалізація процедур, яких вимагає ряд поставлених задач.











Висновок:

Мною були виділені такі етапи аналізу даних: отримання даних, обробка, аналіз та інтерпретація результатів обробки.   
 Перевагами python перед іншими мовами програмування, на мою думку, це відносна нескладність цієї мови програмування та хороша документація сприяють ефективній роботі навіть новачкам в сфері DA .

Цитуючи експертів в сфері DA :

«*With python, getting the data will be trivial. But with R, it might be very hard in some cases.*

*Try to do some natural language processing with R packages ( e.g openNLP) and compare it with NLTK. Example : If you do part of speech tagging, NTLK with python gives you a nice tuple data structure to work with, whereas openNLP in R gives back an ugly looking string which you have to parse yourself.*

*So my advice would be  : Learn python. It's a beautiful language. I am using python primarily for data cleaning task (which takes 70% time in a project) and use R for building/testing the actual statistical models (which takes remaining 30% time).»*

Ідеальні дані – це набір повних даних. На мою думку, джерел даних досить багато : бази різних наукових інститутів, бази даних магазинів і т.д.

Data clean – це дані які не містять спецсимволів, які позначають відсутність даних(як в нашому випадку до очистки) , дані що не містять повторень(не повторень значень, а саме повторення тих самих даних, що буде негативно впливати на частотний аналіз), даних в неправильному форматі або неправильних даних.