Introduction To Capstone Project

You have recently been hired as a Data Analyst by a global IT and business consulting services firm that is known for their expertise in IT solutions and their team of highly experienced IT consultants. **In order to keep pace with changing technologies and remain competitive, your organization regularly analyzes data to help identify future skill requirements.**

**As a Data Analyst, you will be assisting with this initiative and have been tasked with collecting data from various sources and identifying trends for this year's report on emerging skills.**

Your first task is to collect the top programming skills that are most in demand from various sources including:

* Job postings
* Training portals
* Surveys

Once you have collected enough data, you will begin analyzing the data and identify insights and trends that may include the following:

* What are the top programming languages in demand?
* What are the top database skills in demand?
* What are the popular IDEs?

You will begin by scraping internet web sites and accessing APIs to collect data in various formats like .csv files, excel sheets, and databases.

Once this is completed, you will make that data ready for analysis using data wrangling techniques.

When the data is ready you will then want to apply statistical techniques to analyze the data. Then bring all of your information together by using IBM Cognos Analytics to create your dashboard. And finally, show off your storytelling skills by sharing your findings in a presentation.

You will be evaluated using quizzes in each module as well as the final project presentation.

**Syllabus**

**Module 1: Data Collection**

* Collecting Data Using APIs
* Collecting Data Using Webscraping
* Exploring Data

**Module 2: Data Wrangling**

* Finding Missing Values
* Determine Missing Values
* Finding Duplicates
* Removing Duplicates
* Normalizing Data

**Module 3: Exploratory Data Analysis**

* Distribution
* Outliers
* Correlation

**Module 4: Data Visualization**

* Visualizing Distribution of Data
* Relationship
* Composition
* Comparison

**Module 5: Dashboard Creation**

* Dashboards

**Module 6: Presentation of Findings**

* Final Presentation

How to Get Data

In the coming future data will increasingly be shared using Application Programming Interfaces (APIs).  
An API is hosted on the public internet or a private network. APIs provide anytime access to the latest data.

Your assignment is to get the number of job openings using the GitHub Jobs API for technologies like:

* Java
* Python
* MySQL
* C++
* C#

GitHub Jobs API allows anyone to query for the jobs based upon:

* Technology like Python, MySQL
* City like New York, Bangalore

Here are the technical details to access the API.

The GitHub Jobs API allows you to search, and view jobs with JSON over HTTP.

To get the JSON representation of any search result or job listing, append .json to the URL you'd use on the HTML GitHub Jobs site.

**For example, below url will search for Python jobs near New York:**

[https://jobs.github.com/positions?description=python\&location=new+york](https://jobs.github.com/positions?description=python%5C&location=new+york)

To get the JSON representation of those jobs, just use positions.json:

[https://jobs.github.com/positions.json?description=python\&location=new+york](https://jobs.github.com/positions.json?description=python%5C&location=new+york)

Pagination

The API also supports pagination. /positions.json, for example, will only return 50 positions at a time. You can paginate results by adding a page parameter to your queries.

Pagination starts by default at 0.

Examples

[https://jobs.github.com/positions.json?description=ruby\&page=1](https://jobs.github.com/positions.json?description=ruby%5C&page=1)  
[https://jobs.github.com/positions.json?page=1\&search=code](https://jobs.github.com/positions.json?page=1%5C&search=code)

GET /positions.json

Search for jobs by term, location, full time vs part time, or any combination of the three. All parameters are optional.

Parameters

* description — A search term, such as "ruby" or "java". This parameter is aliased to search.
* location — A city name, zip code, or other location search term.
* lat — A specific latitude. If used, you must also send long and must not send location.
* long — A specific longitude. If used, you must also send lat and must not send location.
* full\_time — If you want to limit results to full time positions set this parameter to 'true'.

Examples

[https://jobs.github.com/positions.json?description=python\&full\_time=true\&location=sf](https://jobs.github.com/positions.json?description=python%5C&full_time=true%5C&location=sf)  
<https://jobs.github.com/positions.json?search=node>  
[https://jobs.github.com/positions.json?lat=37.3229978\&long=-122.0321823](https://jobs.github.com/positions.json?lat=37.3229978%5C&long=-122.0321823)

GET /positions/ID.json

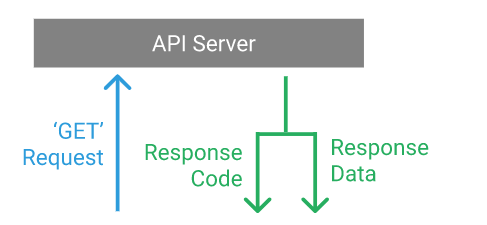
Retrieve the JSON representation of a single job posting.

Parameters

markdown — Set to 'true' to get the description and how\_to\_apply fields as Markdown.

Examples

<https://jobs.github.com/positions/21516.json>  
<https://jobs.github.com/positions/21516.json?markdown=true>



Assignment Overview

Data Wrangling or Munging is a process in which we clean up the data set and make it ready for data analysis. In this assignment you will perform the following tasks:

1. Identify duplicate rows in the data frame.
2. Remove duplicate rows from the data frame.
3. Find the number of missing values for all columns.
4. Find the value counts for the column "Employment".
5. Normalize the data using two existing columns.