Other Data Sources & Report Automation

Ted Kwartler

HARVARD
UNIVERSITY

Agenda

Start	End	Item
		New Data Sources – APIs (JSON & XML)
		New Data Sources -Web Scraping
		New Data Sources – Microsoft Office
		Reporting

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

What is an API?

"Application Program Interface"
Clearly defined methods of communication
between various software components.
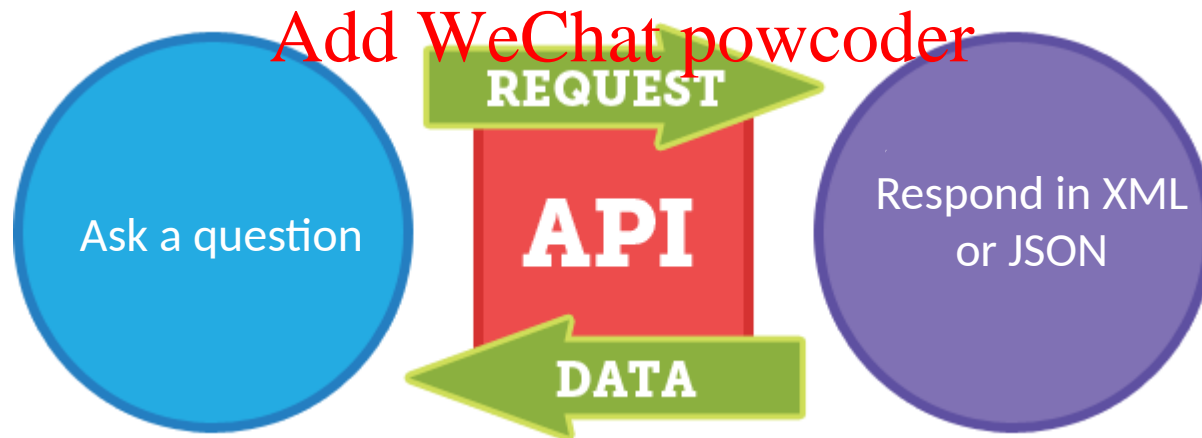
Kwartler CSCI 96

HARVARD
UNIVERSITY
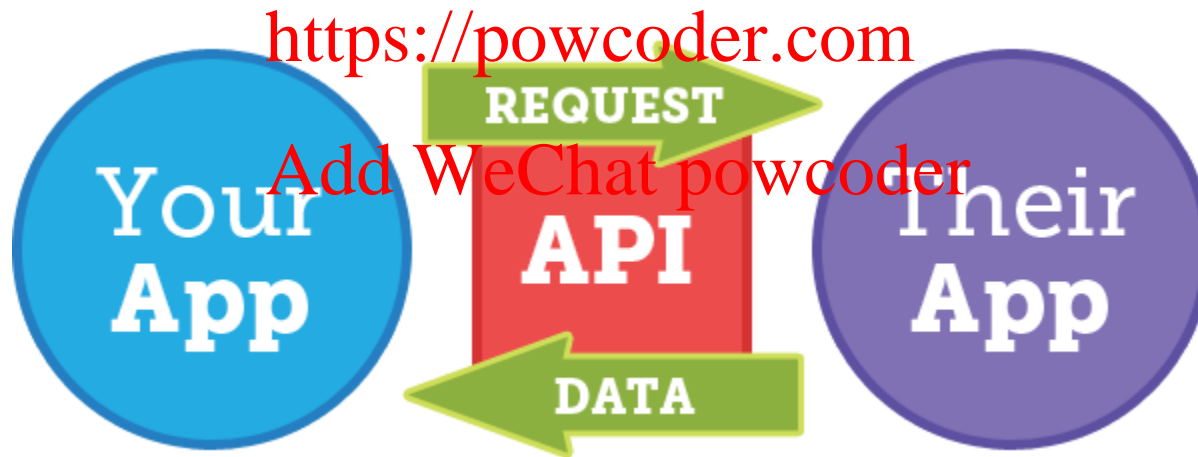
Two formats

- XML – Extensible Markup Language
  - Ever wonder why Excel files went from xls to xlsx? The data is stored as XML.

- JSON- Javascript Object Notation
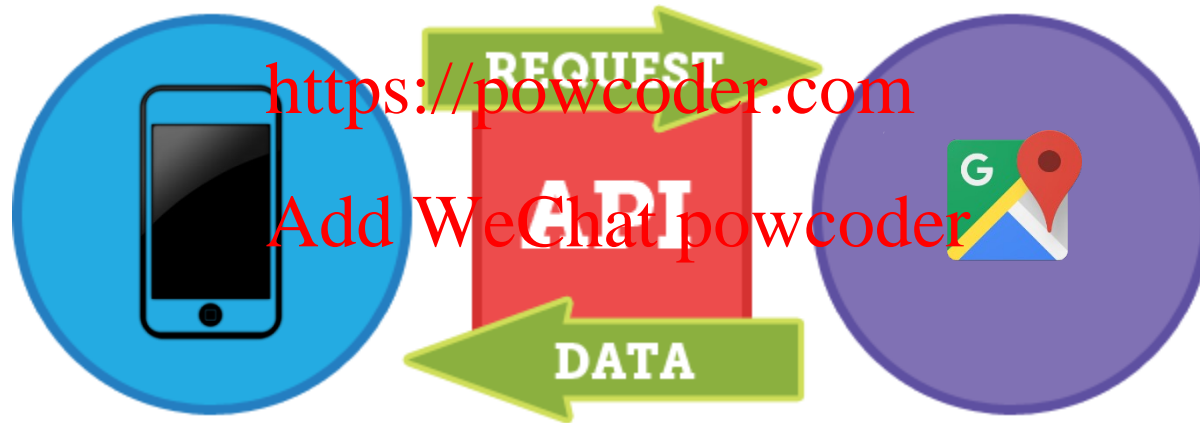 - Similar to R but used to make interactive objects in web browsers.

REQUEST

API

DATA

Ask a question

Respond in XML or JSON

HARVARD
UNIVERSITY

Your phone doesn't have every map in the world.

"Your phone requests information from the maps API services to render the information.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

REQUEST

API

DATA

Your phone doesn't have every map in the world.

www.google.com/maps/place/Cleveland,+OH/@41.4951143,-81.8462865,11z

Google Maps API Services include
- Geocoding
- Base map "tiles"
- Basic Geographic Info

Kwartler CSCI 96

HARVARD
UNIVERSITY
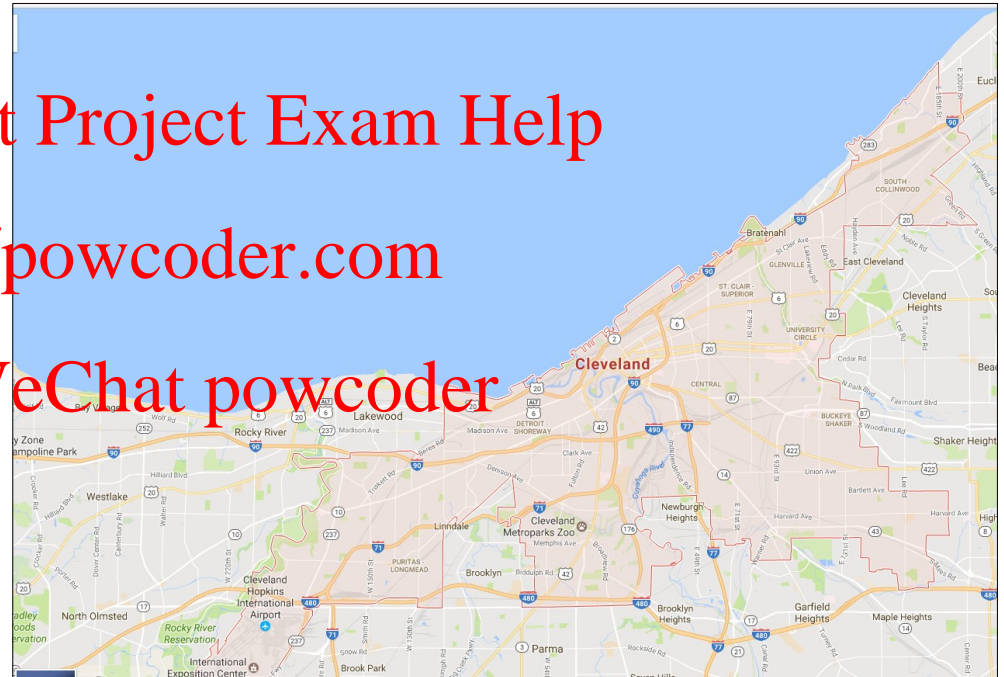
APIs are behind many of the sites you use everyday.

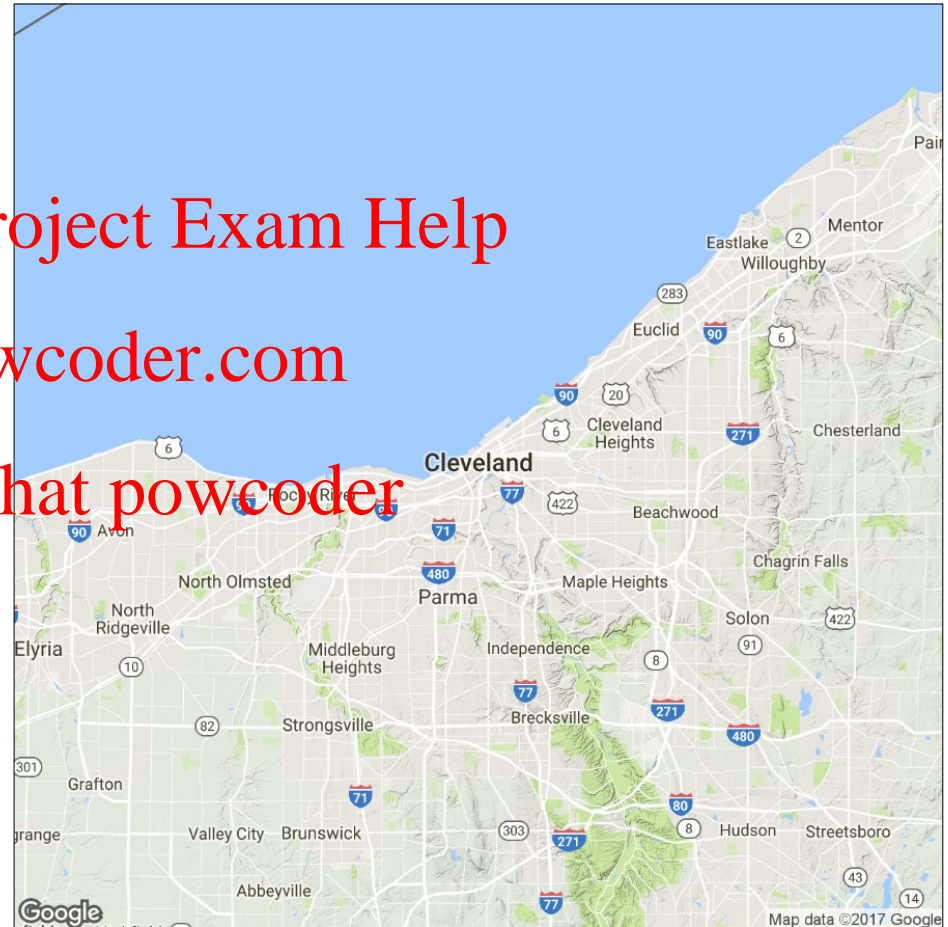[Static Map API](#)

Static images are loaded as files
underneath allowing you to scroll
and navigate in your browser.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

APIs are behind many of the sites you use everyday.

JSON Information API

Assignment Project Exam Help

https://powcoder.com

On top of the tiles, more information
is needed including the geo-political
information and coordinates for the
"bounding box"

Add WeChat powcoder

```
{
    "results" : [
        {
            "address_components" : [
                {
                    "long_name" : "Cleveland",
                    "short_name" : "Cleveland",
                    "types" : [ "locality", "political" ]
                },
                {
                    "long_name" : "Cuyahoga County",
                    "short_name" : "Cuyahoga County",
                    "types" : [ "administrative_area_level_2", "political" ]
                },
                {
                    "long_name" : "Ohio",
                    "short_name" : "OH",
                    "types" : [ "administrative_area_level_1", "political" ]
                },
                {
                    "long_name" : "United States",
                    "short_name" : "US",
                    "types" : [ "country", "political" ]
                }
            ],
            "formatted_address" : "Cleveland, OH, USA",
            "geometry" : {
                "bounds" : {
                    "northeast" : {
                        "lat" : 41.604436,
                        "lng" : -81.5327439
                    },
                    "southwest" : {
                        "lat" : 41.390628,
                        "lng" : -81.87897599999999
                    }
                },
                "location" : {
                    "lat" : 41.49932,
                    "lng" : -81.6943605
```

HARVARD
UNIVERSITY

Comparing XML to JSON…same info

This XML file does not appear to have any style information assoc

```
▼<GeocodeResponse>
   <status>OK</status>
 ▼<result>
    <type>locality</type>
    <type>political</type>
    <formatted_address>Boston, MA, USA</formatted_address>
  ▼<address_component>
     <long_name>Boston</long_name>
     <short_name>Boston</short_name>
     <type>locality</type>
     <type>political</type>
    </address_component>
  ▼<address_component>
     <long_name>Suffolk County</long_name>
     <short_name>Suffolk County</short_name>
     <type>administrative_area_level_2</type>
     <type>political</type>
    </address_component>
  ▼<address_component>
     <long_name>Massachusetts</long_name>
     <short_name>MA</short_name>
     <type>administrative_area_level_1</type>
     <type>political</type>
    </address_component>
  ▼<address_component>
     <long_name>United States</long_name>
     <short_name>US</short_name>
     <type>country</type>
     <type>political</type>
    </address_component>
  ▼<geometry>
   ▼<location>
      <lat>42.3600825</lat>
      <lng>-71.0588801</lng>
     </location>
     <location_type>APPROXIMATE</location_type>
   ▼<viewport>
```

```
{
   "results" : [
      {
         "address_components" : [
            {
               "long_name" : "Boston",
               "short_name" : "Boston",
               "types" : [ "locality", "political" ]
            },
            {
               "long_name" : "Suffolk County",
               "short_name" : "Suffolk County",
               "types" : [ "administrative_area_level_2", "political" ]
            },
            {
               "long_name" : "Massachusetts",
               "short_name" : "MA",
               "types" : [ "administrative_area_level_1", "political" ]
            },
            {
               "long_name" : "United States",
               "short_name" : "US",
               "types" : [ "country", "political" ]
            }
         ],
         "formatted_address" : "Boston, MA, USA",
         "geometry" : {
            "bounds" : {
               "northeast" : {
                  "lat" : 42.40081989999999,
                  "lng" : -70.749455
               },
               "southwest" : {
                  "lat" : 42.22788,
                  "lng" : -71.191113
               }
            },
            "location" : {
               "lat" : 42.3600825,
               "lng" : -71.0588801
```

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

https://maps.googleapis.com/maps/api/geocode/xml
?address=boston&sensor=false&key=AIzaSyCg5BhicmNdpk2Hg1dr0m-H3XPWjd0BtfU

https://maps.googleapis.com/maps/api/geocode/json
?address=boston&sensor=false&key=AIzaSyCg5BhicmNdpk2Hg1dr0m-H3XPWjd0BtfU

HARVARD UNIVERSITY

If you know where to look, you can access APIs for data!

 If the video has closed captioning, let's grab the text by clicking "cc"
https://www.youtube.com/watch?v=NYL-wPVzL64

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

WHERE NBC'S STEVE PATTERSON IS TONIGHT. STEVE, GOOD EVENING.

STEVE PATTERSON
PARADISE, CALIFORNIA

HARVARD
UNIVERSITY

In chrome access the developer console.

While the video is playing, press F12 & reload the page.

Kwartler CSCI 96

HARVARD
UNIVERSITY
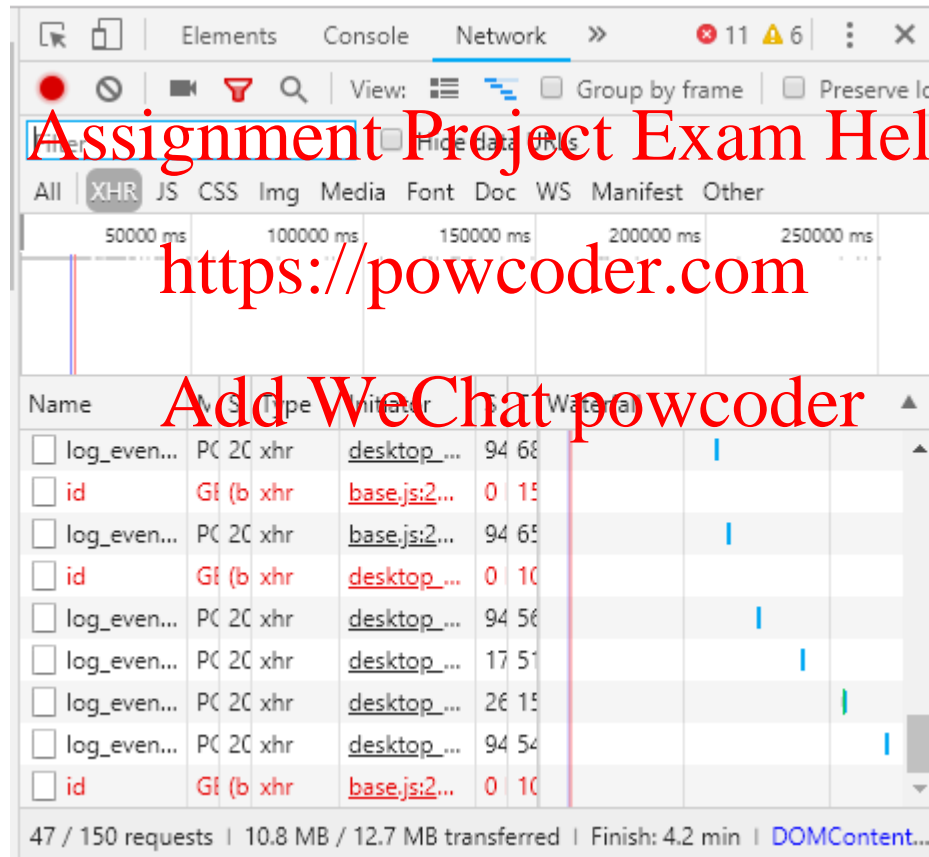
Click XHR and search for "timed" as in timedtext.
Right click on the request name and select "open in new tab"

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY
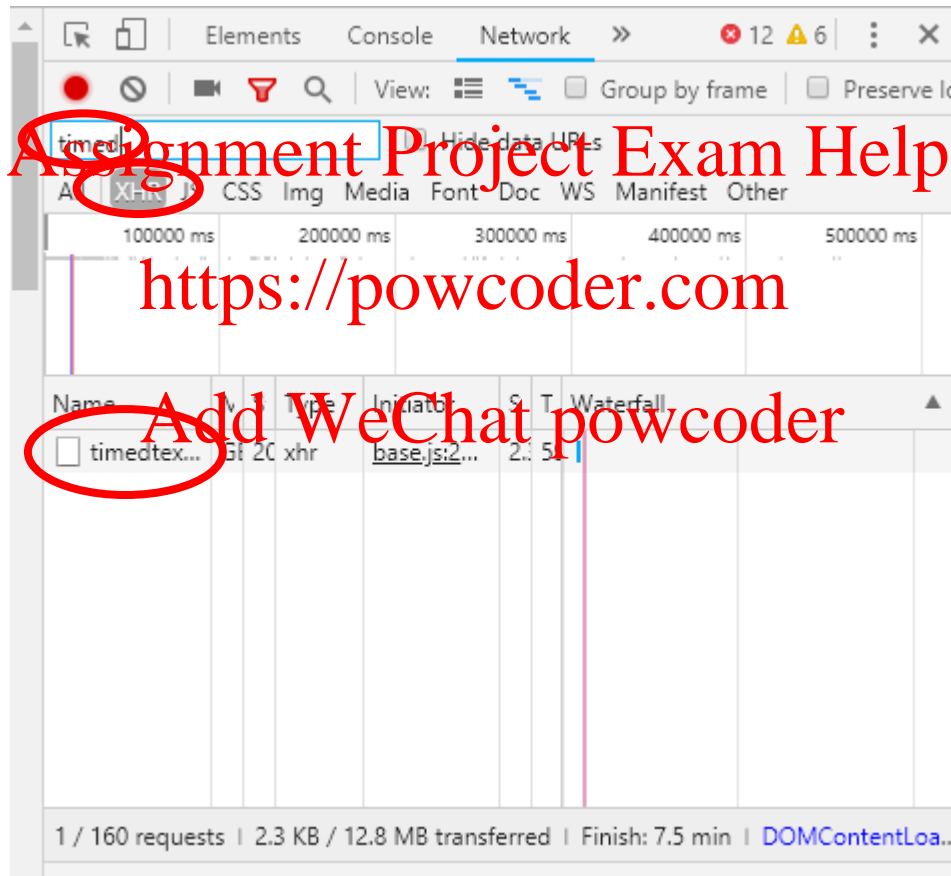
The closed caption XML Response

The response page is active for ~10min.
We can use R to read it and organize it.

This XML file does not appear to have any style information associated with it. The document tree is sho

```
▼<timedtext format="3">
  ▼<head>
     <w id="1" ju="0">
     <wp id="0" ap="7" ah="20" av="95" rc="11" cc="32">
     </head>
  ▼<body>
     ▼<p t="1202" d="1100" wp="1" ws="1">
        BEEN ALL BUT WIPED OFF THE MAP, AND THAT'S WHERE NBC'S STEVE
       </p>
     ▼<p t="2336" d="1168" wp="1" ws="1">
        THE MAP, AND THAT'S WHERE NBC'S STEVE PATTERSON IS TONIGHT.
       </p>
     ▼<p t="3504" d="1067" wp="1" ws="1">
        WHERE NBC'S STEVE PATTERSON IS TONIGHT. STEVE, GOOD EVENING.
       </p>
     ▼<p t="4572" d="733" wp="1" ws="1">
        PATTERSON IS TONIGHT. STEVE, GOOD EVENING. >> Reporter: KATE
       </p>
     ▼<p t="5306" d="766" wp="1" ws="1">
        STEVE, GOOD EVENING. >> Reporter: KATE THIS, COMMUNITY NOW
       </p>
     ▼<p t="6073" d="733" wp="1" ws="1">
        >> Reporter: KATE THIS, COMMUNITY NOW FILLED WITH GRIM
       </p>
     ▼<p t="6807" d="1501" wp="1" ws="1">
        THIS, COMMUNITY NOW FILLED WITH GRIM SCENES LIKE THIS, CARS
       </p>
     ▼<p t="8342" d="1668" wp="1" ws="1">
        FILLED WITH GRIM SCENES LIKE THIS, CARS GUTTED BY WALLS OF
       </p>
     <p t="10011" d="499" wp="1" ws="1">SCENES LIKE THIS, CARS GUTTED BY WALLS OF FIRE.</p>
     <p t="10512" d="4069" wp="1" ws="1">GUTTED BY WALLS OF FIRE. IN SOME CASES THEY</p>
     <p t="14615" d="2301" wp="1" ws="1">FIRE. IN SOME CASES THEY BECAME TOMBS FOR</p>
     ▼<p t="16918" d="1100" wp="1" ws="1">
        IN SOME CASES THEY BECAME TOMBS FOR RESIDENTS AND AS WE
       </p>
```

New version is JSON

```
  "wWinId": 1,
  "segs": [ {
    "utf8": "border",
    "acAsrConf": 239
  }, {
    "utf8": " is",
    "tOffsetMs": 300,
    "acAsrConf": 218
  }, {
    "utf8": " going",
    "tOffsetMs": 360,
    "acAsrConf": 150
  }, {
    "utf8": " to",
    "tOffsetMs": 540,
    "acAsrConf": 239
  }, {
    "utf8": " be",
    "tOffsetMs": 600,
    "acAsrConf": 252
  }, {
    "utf8": " closed",
    "tOffsetMs": 659,
    "acAsrConf": 252
  }, {
    "utf8": " hundred",
    "tOffsetMs": 1070,
    "acAsrConf": 187
  } ]
}, {
  "tStartMs": 5930,
  "dDurationMs": 1840,
  "wWinId": 1,
  "aAppend": 1,
  "segs": [ {
    "utf8": "\n"
  } ]
}, {
  "tStartMs": 5940,
  "dDurationMs": 3720,
  "wWinId": 1,
```

Let's Practice…

A_youtubeAPI_example.R

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

A Simple JSON API

News API   Get started   Documentation   News sources   Pricing   ehk116@gmail.com

Assignment Project Exam Help

Search worldwide news with code

https://powcoder.com

Get breaking news headlines, and search for articles from over 30,000

Add WeChat powcoder

news sources and blogs with our news API

Get API key

HARVARD
UNIVERSITY

Some APIs require a "handshake" to authenticate

Account

Email

ehk116@gmail.com

Assignment Project Exam Help

Update email address

Password

https://powcoder.com

Change password

Add WeChat powcoder

API key

Copy this Code...

Regenerate API key

```
# https://newsapi.org/ Key
apiKey <- 'XXXXXXXXXXXXXXXXXXXXXXXX'
```

HARVARD
UNIVERSITY

Now JSON

B_newsAPI_example.R

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

JSON Statham

New JSON object

HARVARD
UNIVERSITY
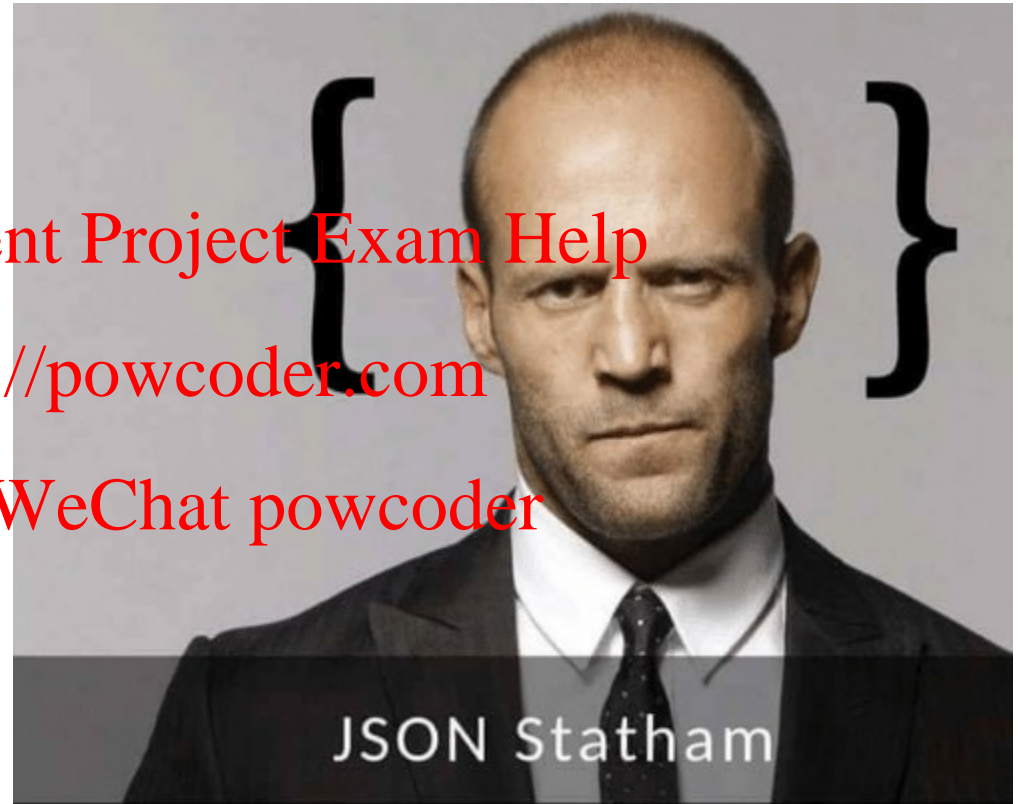
Let's Practice…

A_youtubeAPI_example.R

B_newsAPI_example.R

C_presidentialApproval.R – *not just text*

Kwartler CSCI 96

Other Straightforward APIs to Explore

- If you work with social media data (requires signing up)
- https://pipl.com/dev/demo

Assignment Project Exam Help

- Sources of APIs to explore
- https://www.programmableweb.com/apis/directory https://powcoder.com
- https://github.com/toddmotto/public-apis

Add WeChat powcoder

- Challenge find the APIs for the chart data here (hint: JSON using F12):
- https://projects.fivethirtyeight.com/trump-approval-ratings

HARVARD
UNIVERSITY

Agenda

Start	End	Item
		New Data Sources – APIs (JSON & XML)
		New Data Sources -Web Scraping
		New Data Sources – Microsoft Office
		Report Automation – library(officer)

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

Web Scraping

R and other languages let you extract information from websites. This is called "web scraping".

HARVARD
UNIVERSITY
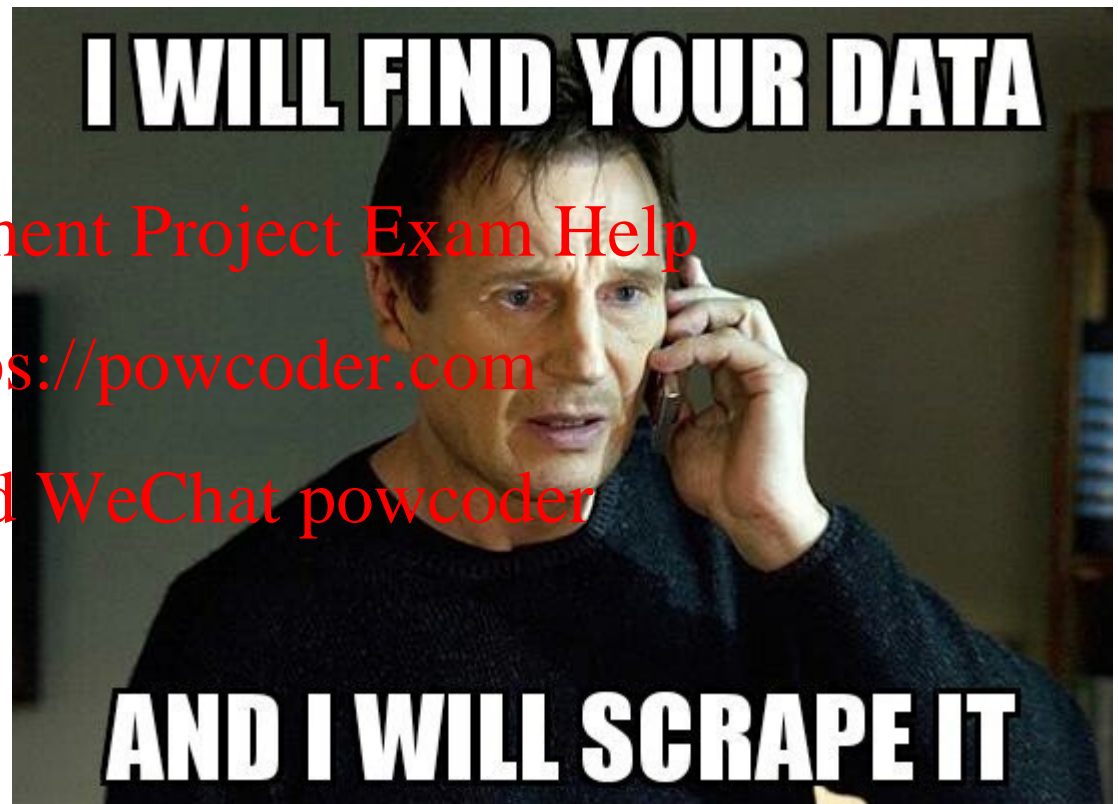
Questionable Practice

Websites don't like you taking information for various reasons. Always respect terms of service and any blocking software.

Web-scraping

- Everything in a webpage is "client" side (interpreted by **your** computer), therefore you can get anything you see.
- Each scrape is literally *custom* software
- Webpages have different styles, object names, and complexity...*always check for an API first*.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

Web Scraping Example

IMDB.com

- Internet Movie Database
- Owned by Amazon.com
- Realistically messy but doesn't stop scraping

HARVARD
UNIVERSITY

Getting some data

D_webscraping.R

Kwartler CSCI 96

HARVARD
UNIVERSITY

Let's get the dirt on GSERM's profs.

1. Open a "headless" browser to the instructor page. Programatically find all appropriate URLS which is represented by each picture.

Let's get the dirt on our profs.

1. Open a "headless" browser to the instructor page. Find all appropriate URLS which is represented by each picture.
2. Loop through all ID'ed URLs to get the bio of each professor.

HARVARD
UNIVERSITY

Let's get the dirt on our profs.

E_webScraping.R

1. Open a "headless" browser to the instructor page. Find all appropriate URLS which is represented by each picture.
2. Loop through all ID'ed URLs to get the bio of each professor.
3. Extract the bio text from each page and organize

TED KWARTLER

Ted Kwartler is an adjunct professor at Harvard University Extension School, author of the book "Text Mining in Practice with R" and has two NLP courses on DataCamp.com. He has worked at amazon.com and was an early employee at DataRobot. At both he had the opportunity to apply data science to help customers, and improve business decision making. Ted holds an MBA from the University of Notre Dame (USA) with a concentration in Marketing Analytics.
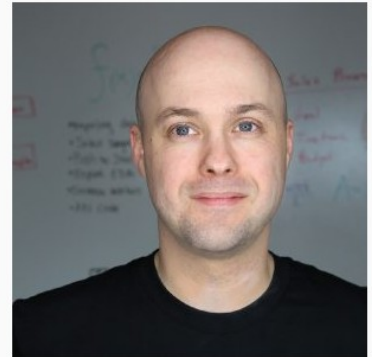
Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

Agenda

Start	End	Item
		New Data Sources – APIs (JSON & XML)
		New Data Sources -Web Scraping
		New Data Sources – Microsoft Office
		Reporting

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

Reading in other file formats

So far we have read in CSV (comma separated values) files.

Often you will be asked to use other data sources.
- .txt (plain text)
- .doc & .docx (Microsoft Word)
- .xls & .xlsx (Microsoft Excel)

R can even interpret PDF, image and other less popular files (ie SPSS).

Kwartler CSCI 96

HARVARD
UNIVERSITY

Let's Practice

F_readDOCX.R

- Get tables
- Get simple text

This is an example Word document for parsing in R. It is for the class CSCI 96

Here is a table

Extracting Text

ID	Some_Value	Another_Value
1	16	A
2	18	B
3	20	B
4	22	A
5	24	B
6	26	A
7	28	A

Parse the Table

HARVARD UNIVERSITY

Let's Practice

G_readEXCEL.R

- List all worksheets
- Read in individual sheets

	A	B	C	D	E	F	G	H	I	J	K
1	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
2	21	6	160	110	3.9	2.62	16.46	0	1	4	4
3	21	6	160	110	3.9	2.875	17.02	0	1	4	4
4	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
5	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
6	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
7	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
8	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
9	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
10	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
11	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
12	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
13	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
14	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
15	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
16	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
17	10.4	8	460	215	3	5.424	17.82	0	0	3	4
18	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
19	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
20	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
21	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
22	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
23	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
24	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2

mtcars | diamonds | profiles ⊕

Get the table data

Identify all Worksheets

HARVARD
UNIVERSITY

Agenda

Start	End	Item
		New Data Sources – APIs (JSON & XML)
		New Data Sources -Web Scraping
		New Data Sources – Microsoft Office
		Reporting

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

Automating Reports

Flexdashboard

Assignment Project Exam Help

- Use R Markdown to publish a group of related data visualizations as a dashboard.

https://powcoder.com

- Support for a wide variety of components including htmlwidgets; base, ggplot2

Add WeChat powcoder

- Flexible and easy to specify row and column-based layouts

You've been building dashboards all semester with markdown ☺

HARVARD
UNIVERSITY

Awesome. No more old cars!

No more bad requests, I work in a hospital now. Its all doctors so I am sure they are data fluent.

3902092

motionelements.com

Kwartler CSCI 96

HARVARD
UNIVERSITY

Let's make some routine reports

Dale! I am freaking out!!!
I need a report on my desk each
Monday that details our diabetic
patient results.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Use Case for Report Automation

cool.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

- Making routine reports is boring

- Routine reports waste analytical talents' time
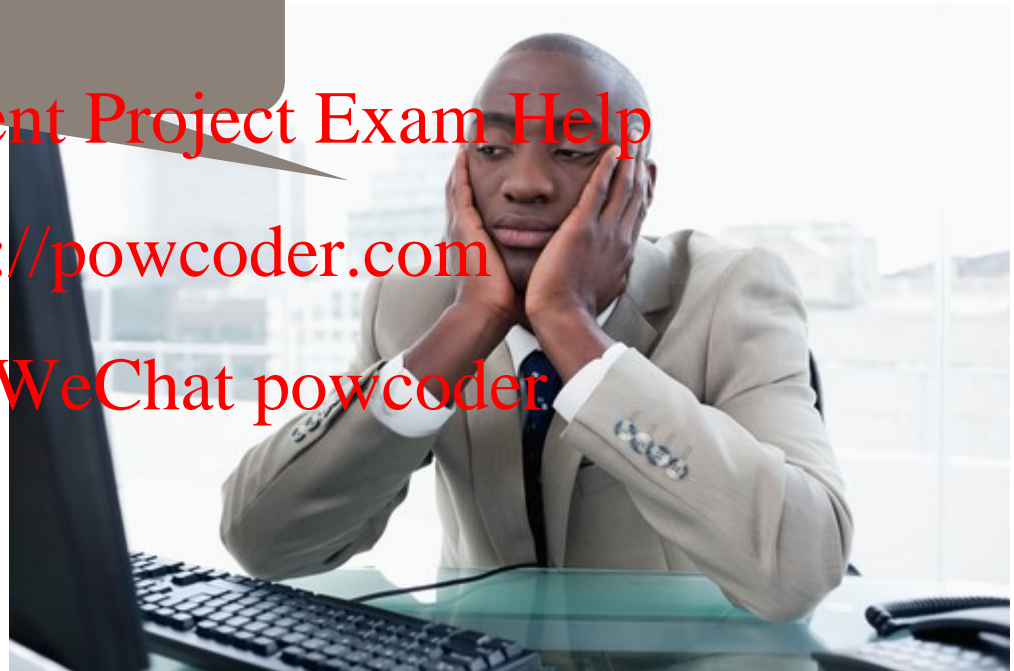
We need to automate Dale's misery!

HARVARD
UNIVERSITY

Multiple Methods

- R library(officer)
  - https://davidgohel.github.io/officer/
  - Docx and pptx
  - Flexible though clunky
  - Uses any static image ggplot etc
  - Easily allows for themes

- R Studio Presentation "PowerPoint"
  - https://support.rstudio.com/hc/en-us/articles/360004672913-Rendering-PowerPoint-Presentations-with-RStudio
  - Requires knowledge of markdown

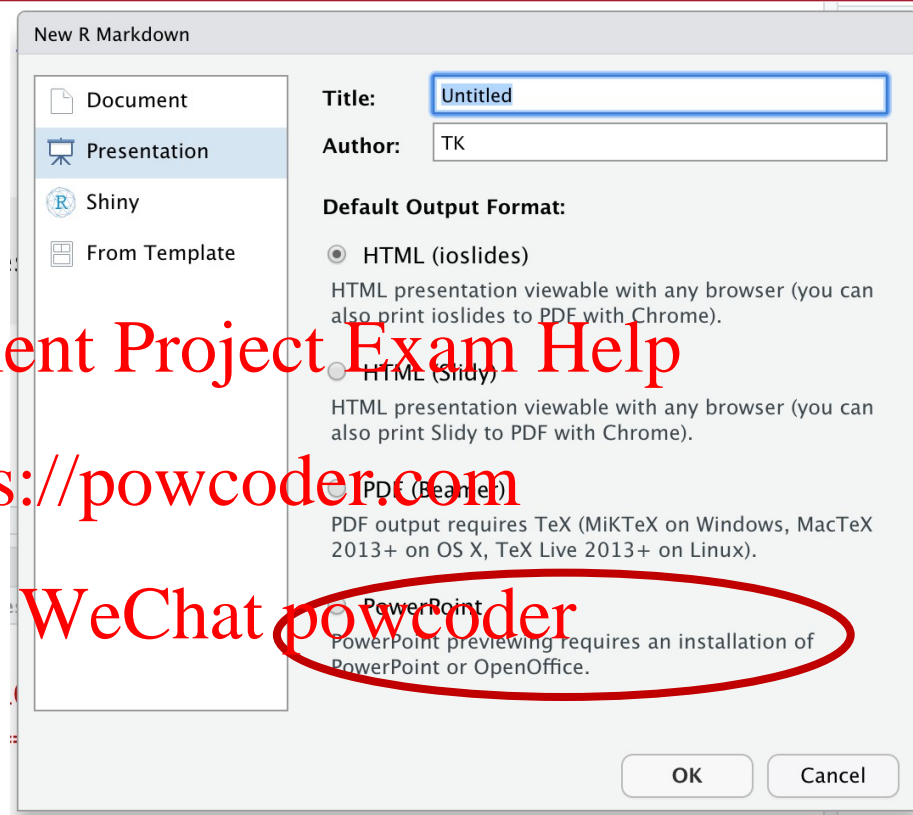Last semester we used library officer but this time the markdown version.

HARVARD
UNIVERSITY

Making powerpoints

File
New File
New Markdown
Select Presentation
Select PowerPoint

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Automating Reports w/FlexDashboard

But Dale, I need more interesting pictures. PowerPoint gives me a headache.

Assignment Project Exam Help

https://powcoder.com

- Use R Markdown to publish a group of related data visualizations as a dashboard.

- Support for a wide variety of components including htmlwidgets; base, ggplot2

Add WeChat powcoder

- Flexible and easy to specify row and column-based layouts.

You've been building dashboards all semester with markdown.

HARVARD
UNIVERSITY

fine.
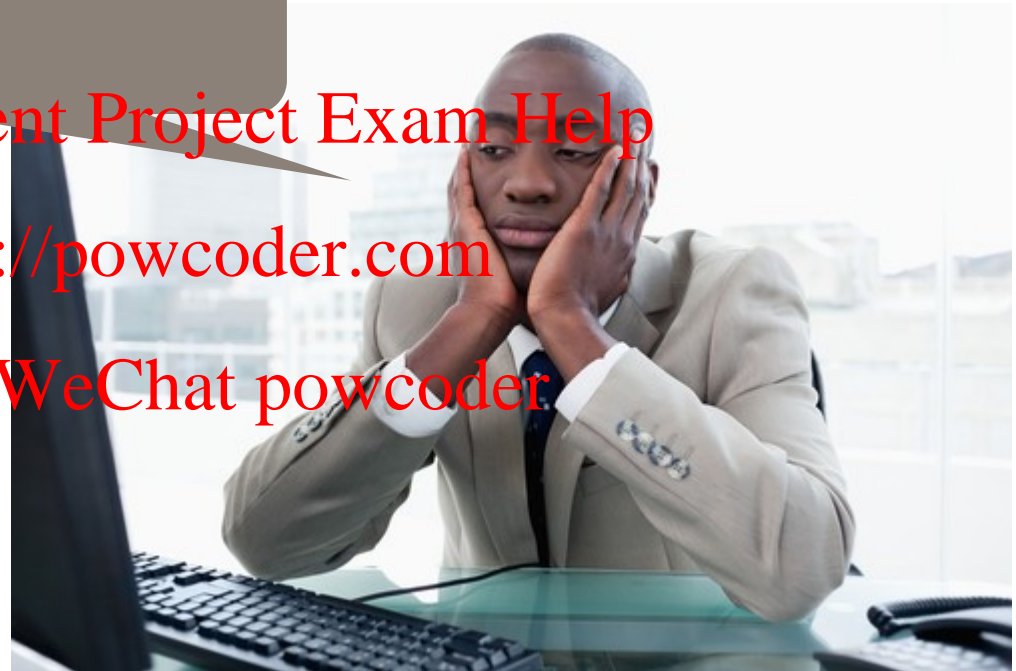
Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Let's add some dynamic reporting!

Let's practice!

H_flexdashboardHomework_REVISED.rmd

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

HARVARD
UNIVERSITY

Homework Clarification

- Make a markdown powerpoint using week 3 data
  - Create 3 or more slides
  - 1 ggplot barplot or histogram
  - 1 ggplot of your choosing
  - 1 with summary information for a variable, can be bullets or a table. No default information from the original markdown template.
  - Turn in the pptx.
- Build a flexdashboard using week 4 data, again change the charts and tables info.
  - Use at least 1 "widget" or dynamic plot
  - Turn in the html.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Kwartler CSCI 96

HARVARD
UNIVERSITY