IS 3005 – Statistics in Practice I
Industry Guest Lecture Series
Take Home Assignment - Building Alternative data.

S15055

01) a) Possible ways of collecting alternative data.

- Web Scraping
- Alternative data providers
- Online surveys
- Public records

b) Advantages of picking web-scraping to gather alternative data.

- It is a customizable method, because it can be customized to extract specific data points and information from websites.
- It is a cost-effective alternative to traditional data collection methods such as surveys or interviews.
- It is a method that we can collect data directly from the source without human intervention or bias.
- Web Scraping can extract large amount of data from multiple sources, and it provide real time data.
- Web scraping can be automated.

c) Limitations/challenges of web-scraping

- Web scraping can raise legal issues related to copyright, intellectual property, and terms of use violations.
- The quality of the data collected through web scraping can vary widely depending on the source website and the accuracy of the data presented on it.
- Data collected through web scraping may not be structured or formatted in a way that is easy to analyze, requiring additional processing or cleaning before it can be used effectively.
- Web scraping can raise ethical concerns related to privacy and data protection, particularly when it comes to personal or sensitive data.

d) Major python libraries and the use of them.

Pandas –

- Pandas provides functions that can read data directly from web pages or extract data from HTML tables. As well as after extracting data from web pages, we can use pandas for Data cleaning, Data analysis and pandas provides functions to write data to a variety of formats, including CSV, Excel, and SQL databases.

Request library with beautiful Soup –

- The combination of the Requests library and Beautiful Soup is a popular choice for web scraping projects in Python. First, we use the Requests library to make an HTTP request to the URL of the web page we want to scrape. Then use Beautiful Soup to parse the HTML content and create a parse tree. Beautiful Soup provides a range of methods for navigating and searching the parse tree, such as find(), find_all(), and select().

Selenium –

- Selenium is a powerful library that allows us to programmatically control a web browser and interact with web pages just as a human user would. As well as extract some dynamic content that is generated by JavaScript, Ajax, or other client-side technologies which may not be available in the HTML source code of the web page is difficult with using traditional web scraping techniques. So, selenium allows us to simulate user interactions with the web page, such as clicking buttons or filling out forms, and retrieve the resulting dynamic content.