**INDEX**

1. **Software and Package requirements**
2. **Reading the Data set**
3. **Accessing Websites**
4. **Web scrapping**
5. **Dataset Update**
6. **Steps to execute the Jupyter notebook**
7. **Software and Package requirements:**

The following are the software that are required to run the model

* 1. Anaconda environment

The following packages and their dependent packages as mentioned by the anaconda are to be installed in anaconda environment to run the file successfully.

1. Below Python packages are used:

import pandas as pd

from urllib.request import urlopen, HTTPError

from datetime import datetime, timedelta

import urllib.request

from urllib.request import urlopen, Request

from bs4 import BeautifulSoup

from googlesearch import search

1. **Reading the data set**

Firstly reads the input dataset to retrieve the dunsName by using python’s pandas library.

1. **Accessing Websites**

Automated the search for each dunsName in google to access their official website by using Libraries. We are taking first search into consideration.

1. **Web scrapping**

Used Python Web scrapping technique to access the HTML page content of the website and search based on the keywords like women owned/minority owned/ diverse owned.

1. **Dataset Update**

Depending upon the results retrieved from the search, we will update the input Dataset with columns (isWomanowned/MinorityOwnedDesc). If the company is Womanowmned, then the flag will be set to 1.If the company is minorityowned, then the column “Minorityowned Desc” will be updated with diversity description like Asian, African American etc.

If the search retrieves nothing, “isWomanOwned” will be zero and “MinorityOwnedDesc” will be NA.

1. **Steps to execute the Jupyter notebook.**
   1. Make sure to have the input data set ('Hackathon\_Data\_MinorityWomenOwned\_2022 v1.xlsx') is in the same location as the Jupyter notebook.
   2. Execute the notebook ‘TheEngineers\_Hackathon2022.ipynb’.
   3. Results will be written into new file 'Hackathon\_Data\_MinorityWomenOwned\_2022 v1\_solution.xlsx'.