No.: SD20036 NG JIE HAO

Final Exam BSD2333

Question 1

(a) i.) Missing Completely at Random (MCAR)
ii.) Missing at Random (MAR)
iii) Missing not at Random (MNAR)

(b) i) MCAR - means the data is missing due to completely random reasons. There is no specific structure as to why data might be missing.
ii) MAR - means the data is missing relative to the ~~obr~~ observed data.
iii) MNAR - means the data will be missing based on the column itself.

(c) Missing data is defined as the data value that is not stored for a variable in- the observation. The problem missing data is relatively common in almost all research.

(d) i) Imputation
ii) Remove data

(e) i) Imputation : Depending why the data are missing, imputation methods can deliver reasonably reliable result.

ii) Remove data : When dealing with data that is missing at random, related data can ~~be~~ be deleted to reduce bias.

No.: SD20036 NG JIE HAO

f) i) Find columns with missing data
ii) Get a list of colums of missing data
iii) Get the number total of missing data in DataFrame.
iv) Get the number of missing data per colums.
v) Remove columns that ~~containg~~ contains more than 50% of missing data.

Question 2

(a) i) Contradictory values
ii) Mislabeled values.
iii) Erroneous values
iv) Missing values
v) Don't care values

(b) i) pd. concat ([df1, df2])

	Name	Age	Height	Pace
0	L. Messi	34	170	85
1	R. Lewandowski	32	185	78
2	C. Ronaldo	36	187	87
3	Neymar Jr	29	175	91
4	K. Mbappe	22	182	97
5	H. Kane	27	188	70
6	M. Salah	29	175	90
7.	K. Benzema	33	185	76

Question 2 SD20036 NG JIE HAO

ii)

	Name	Age	Height	Pace	Weight	Position
0.	C. Ronaldo	36	187	87	83	ST
1.	Neymar. Jr	29	175	91	68	LW

iii) (left join) df1, df4

	Name	Age	Height	Pace	Position
0.	L. Messi	34	170	85	NaN
1.	R. Lewandowski	32	185	78	NaN
2.	C. Ronaldo	36	187	87	ST
3.	Neymar Jr	29	175	91	LW

iv) pd. merge ([df2, df3] how = 'outer')

	Name	Age	Height	Pace	Weight	Position
0	L. Messi	34	170	85	72	NaN
1	R. Lewandowski	32	185	78	81	NaN
2	C. Ronaldo	36	187	87	83	ST
3	Neymar Jr	29	185	91	68	LW
4	M. Salah	29	175	90	NaN	~~RW~~ RW
5	K. Benzama	33	185	76	NaN	CF

v) pd. merge ([df1, df4], how = 'inner')

	Name	Age	Height	Pace	~~Weight~~	Position
0	C. Ronaldo	36	187	87	~~83~~	ST
1.	Neymar Jr	29	175	91	~~68~~	LW

Question 3

(a) ~~Web sra Scrapping~~

i) Web scraping is extracting a large amount of specific data from online sources.

ii) Web Crawling is using tools to read, copy and store the content of the websites for indexing purposes.

(b) Web scraping

i) Comparing Prices

ii) Stock Market Analysis

iii) Market Research for new product.

Web crawling

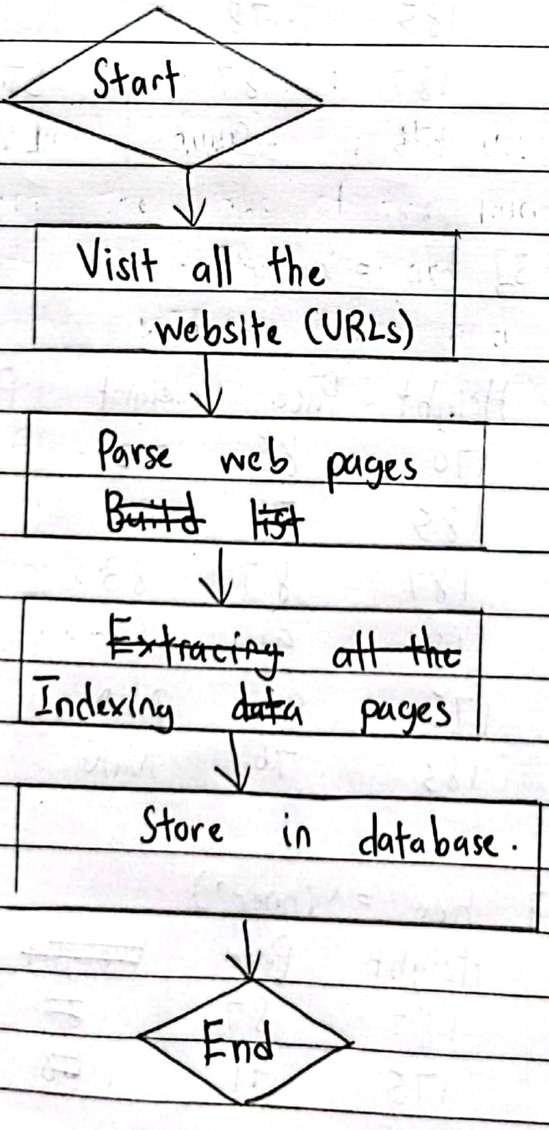
i) Performing website analysis

ii) Monitoring SEO analysis

iii) Generating search engine results.

Flowchart

(c)



Web Crawling

No.:

SD20036 NG JIE HAO

First, find out the target website and collect all the URLs. Then, parsing web page and indexing pages on the content. Lastly, stores in database.

~~(d)~~

(d) Differences between web scraping & web crawling.

Web ~~sta~~ scraping	Web crawling.
~~-Indexing pages based on the content~~.	
- Extracing information from the contents of the pages	- Indexing pages of the contents
- Scraper bots	- crawler bots.
-used by small and large business.	-Performs only by large corporations

Question 4

(a) The business of client was confronted with highly competitive market where their competitions frequently changed their prices and ~~amount~~ assortment, which was very difficult to track, considering the large scope of products to be monitored

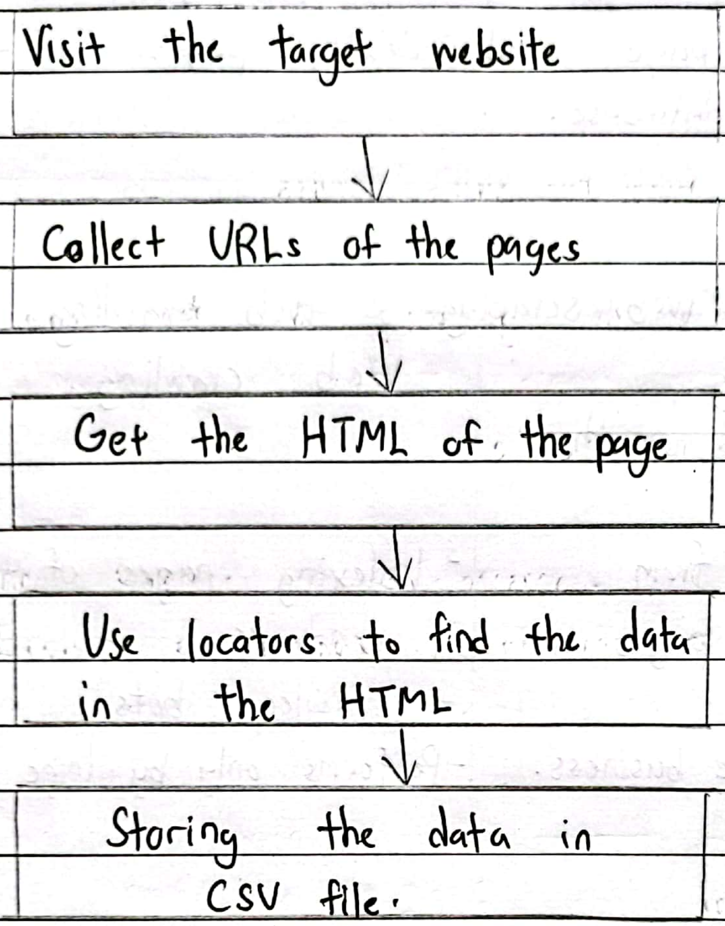
~~(b)~~ across very different product ~~ca~~ categories.

(b) The client needs to analyze their competitor's data to expand their own assortment and stay competitive. The client was in need of reliable information on ~~Competition~~ competitors' actions regarding how they run campaigns and promotions.

No.: SD20036 NG JIE HAO

Question 4

(c)



Visit the target website

Collect URLs of the pages

Get the HTML of the page

Use locators to find the data in the HTML

Storing the data in CSV file.

Web Scraping

First, we need find out the target website and visit it. Then, collect the URLs of the pages where you want to extract the data. Third, make a request to this URLs to get the HTML. pages. and use locators to find the data in the HTML. Lastly, storing the data in CSV file.

(d) i) Seo Crawler
ii) Parse Hub.

No.: SD20036 NG JIE HAO

Question 4

(e) i) JSON file

ii) XLSX format.

(f) i) Stay up to date with the price trends.

- The client are continually changing prices of product ~~iii~~ in an attempt to stay ahead of the competitors.

ii) Look at all the details.

- The client can check all the details like shipping costs, price, service ~~gura~~ guarantee and the popularity of the ~~retailer~~ ~~retailer~~ others competitors as well.

Question 5

(a) (i) ~~Pandas~~ Matplolib

(ii) Plotly Express

(iii) Seaborn

(iv) Altair

(v) Bokeh

(a) (ii) ~~(iii)~~ 1. Storytelling - Storytelling allows us to share our visualization and story with others.

2. ~~Inden~~ Identify emerging trends on community. These trends make more sense when they are graphically represented.

No.: SD20036 NG JIE HAO

Question 5

(b) i) To find out the effects of ~~New~~ Walmart stores number on the United States economy.

~~ii) 1) Consumers are looking for one~~

iii) 1) Consumers are looking for value.
- Walmart is one of the largest retailers in the world for a reasons it capitalizers on the consumer's desire for value.

2) The suceed in eCommerce.
- Ecommerce makes walmart to grow better, and increasing the walmart stores.

ii) This is because ~~the~~ ~~period~~ financial crisis happen on 1996 and 2006 and decrease the walmart store.

iv) i) Income
ii) Socioeconomic status
iii) Education