Simplified Assessment Guide: Data Governance (DAG411)

What Are You Doing in This Assessment?

You are continuing your data journey, but now you're going beyond scraping. Youll be cleaning, protecting, and organizing the data you scrape. Youll look at real company data (like Shoprite), and ensure it's handled properly with laws and good practices in mind.

1. Best Practices for Data Governance (25 Marks)

- Explain how Shoprite (or any big company) should manage its data responsibly.
- What rules or plans should they follow?
- Use your web scraped data as an examplehow would you keep it clean, safe, and organized?

Write 3 short paragraphs explaining:

- What is data governance?
- Why is it important?
- How would you organize scraped data from Shoprite (e.g., using folders, labels, formats)?

2. Legislation (POPI Act) (25 Marks)

- Show that your scraped data respects South African law (POPIA).
- Talk about things like permission, privacy, and not collecting sensitive info.

Write a paragraph explaining:

- What POPIA is.
- How it protects peoples info.
- How you would apply this when scraping product info (e.g., avoid collecting names, emails, etc.).

3. Wrangling (25 Marks)

Simplified Assessment Guide: Data Governance (DAG411)

- Clean your scraped data using Python and Pandas.

- Focus on: Removing duplicates, fixing or removing missing data.

Example code:

```
import pandas as pd

df = pd.read_csv("shoprite_data.csv")

df.drop_duplicates(inplace=True)

df.dropna(inplace=True)

df.to_csv("cleaned_shoprite_data.csv", index=False)
```

Add screenshots and explain what your code does in simple words.

4. Data Structures (25 Marks)

- Explain how to organize large sets of data from Shoprite, e.g., using Spark DataFrames.

If you havent used Spark, just explain how big stores like Shoprite need powerful tools to handle big data.

Mention that Spark DataFrames are like Pandas but for bigger data.

Optional example:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("ShopriteData").getOrCreate()

df = spark.read.csv("shoprite_data.csv", header=True, inferSchema=True)

df.show()
```

Final Notes:

Simplified Assessment Guide: Data Governance (DAG411)

- Write 3 paragraphs per section (short and clear is okay!)

- Include screenshots of your code and cleaned data

- Use the Declaration of Authenticity page

- Submit all in one PDF no Word docs

- Harvard-style references for any websites or definitions you use.