Project Methodology:

The main goal of this project is to design and develop a web application which visualises and provides trends and data for international tourist arrivals to specific countries, regions and income categories. This project is mostly data science related but involves some software development if a web app is ultimately made so both types of methods were considered. Three methods were initially chosen to compare which were ‘CRISP-DM’, ‘Domino Life Cycle’ and ‘Scrum’ with the first two being data science methods and the last method being a software development method.

Five important criteria were defined for choosing the methods more efficiently:

1. Level of volatility/changeability in requirements: This should be high as there is little to no communication with the client and there are many initial unknowns as to their needs and wants. There needs to be high flexibility to make changes depending on new information gained.

2. Ease of learning the method/availability of documentation and support – it should be very easy to learn the method and documentation should be readily available, especially for users such as myself with no experience.

3. Should be able to handle large datasets – the dataset for international tourism used in this project is relatively large.

4. Timescale – the timescale is fixed and a relatively short timescale of around 4 months given.

Method: **Criterion:**	**1. CRISP-DM**	**3. TDSP**	**3. Scrum**
1. Level of volatility/changeability in requirements – should be high as there is low communication with the client and many initial unknowns	CRISP-DM is very flexible since it is mainly an agile method that allows freedom to move back and forth between different phases rather than following a strict linear sequence like other waterfall methods. This greatly aids the iterative process needed in a data science/software engineering project where, for example, moving between business and data understanding phases, where initially, there are many unknowns. So this flexibility allows each iteration and cycle to gain a deeper understanding of the data and problem for the method user.	TDSP is also an agile method like CRSIP-DM where there is high flexibility in moving between stages such as transferring between the data acquisition stage and deployment stage which again greatly aids the iterative process needed to work past unknowns and for the user to gain a deeper understanding of the data and problem after each iteration.	Scrum allows for the user of the method to quickly and effectively react and adapt to any changes. Proceeding each ‘sprint’, iterative changes can be made depending on the progress made in the sprint. Adapting to changes is also
2. Ease of learning the method/availability of documentation and support – should be very easy to learn and documentation should be readily available.	CRISP-DM provides a range of documentation available online with guidelines and experience documentation for users with little experience. All documents are readily available.	There is a wide range of documentation available by Microsoft that makes TDSP quick and easy to learn. However, some documents provided are inconsistent.	Scrum is relatively easy to learn, especially for teams as sprints can split tasks between members. The documentation is also available and ready to use.
3. Should be able to handle large datasets – the dataset for international tourism used in this project is relatively large	CRISP-DM’s data understanding stage is very flexible and effective for relatively large datasets, allowing documentation of any inconsistencies and visualising the data to then prepare it in the data preparation stage to clean construct and prepare data.	TDSP is very proficient with relatively large datasets as it provides a utility called “IDEAR” to help visualise any size dataset, particularly larger datasets and prepare data summary reports.	Scrum is very effective with large datasets again, due to the splitting of tasks into sprints. The sprint review process allows effective analysis of each task after it is done and can help to quickly analyse large datasets.

4. Timescale – fixed relatively short timescale of around 4 months given.	CRISP-DM can be relatively more time consuming than the other methods for understanding and preparing the data but overall, can match the relatively short timescale of this project and the flexibility outweighs the slight downside in time.	The TDSP method is reliable with shorter timescales particularly in the business understanding stage when success metrics are defined which must be "SMART". The relevant letter in this abbreviation is T which refers to Time-bound and the objectives defined are very time specific with goals of finishing each task within a specific timeframe.	Scrum relies on fast 'sprints' which are the basic unit of workflow for a team and are short iterations. (1 to 3 weeks) with specific tasks for each sprint ensuring effectiveness for a project with a short timescale such as this one. Particularly, the 'daily sprint' is very efficient for short timescales progress is evaluated each day with synchronising each activity and creating a plan for the next 24 hours.

Table 1: Comparison and analysis of 3 process methods

Conclusion:

The project methodology chosen was CRISP-DM. The main reason it was chosen is because of the first criterion. The flexibility this method allows with its high volatility allows it to adapt to changes at any stage of the process. For example, this is useful when new information to update the problem statement is found to help understand how to design the webapp. There are many initial unknowns, particularly for the client and the flexibility of moving back and forth between business understanding and data understanding phases allows for the method user to gain a deeper understanding of the problem and how it can be solved with the end product, in this case with the webapp in the context of this project. Also, there is a range of documentation available and it is very easy for someone with little to no experience to use which is the case with myself with no experience. There are some slight drawbacks with the long time taken in the data understanding stages but this extra time is ultimately beneficial since this is the most important initial part of a data science or software engineering project so it will allow a deeper understanding of the problem and data.

References:

CRISPDM criteria 1: https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview

CRISPDM criteria 2: https://www.datascience-pm.com/wp-content/uploads/2021/08/CRISP-DM-for-Data-Science.pdf

CRISPDM criteria 3: https://data-science-blog.com/blog/2021/01/06/crisp-dm-methodology-in-technical-view/

Crispdm criteria 4 (timescale): https://data-science-blog.com/blog/2021/01/06/crisp-dm-methodology-in-technical-view/

TDSP criteria 1: https://www.customerinsightleader.com/opinion/more-data-science-methodology-2/

TDSP criteria 2: https://github.com/Azure/Microsoft-TDSP/blob/master/Docs/README.md and https://www.datascience-pm.com/tdsp/

TDSP criteria 3: https://github.com/Azure/Microsoft-TDSP/blob/master/Docs/README.md

TDSP criteria 3: https://sagu94271.medium.com/tdsp-team-data-science-process-cf08f5b0d1dd

Scrum criteria 1: https://www.linkedin.com/pulse/achieving-flexibility-scrum-soumyaranjan-mukherjee-/

Scrum criteria 2: https://www.digite.com/agile/scrum-methodology/

Scrum criteria 3: https://www.digite.com/agile/scrum-methodology/

Scrum criteria 4: https://www.vthink.com.au/single-post/2017/11/05/how-to-apply-scrum-in-large-projects