Product Requirements Document: "**Dark Sky Explorer NT**"

**1. Introduction**

This document outlines the product requirements for "Dark Sky Explorer NT," a data science project aimed at analyzing and exploring the potential for stargazing tourism in the Northern Territory (NT). The project will leverage public data and community insights to identify and rank optimal stargazing locations, providing actionable recommendations for local businesses and the government to promote NT as a **leading dark-sky destination**.

* Enhancing dark-sky experiences can attract niche tourism markets, extend visitor stays, and support regional development in remote NT communities.
* Our analysis also aims to estimate the potential economic value of improving dark-sky tourism.

**2. Background & Problem Statement**

The NT has a unique advantage due to its low light pollution and vast natural landscapes, making it ideal for stargazing tourism. However:

* There is no structured analysis of tourist feedback specific to dark sky experiences.
* Local businesses and policymakers lack data-driven insights to improve or market these destinations.

This project addresses these gaps by analysing existing tourist reviews, applying data science techniques, and generating practical recommendations.

**2. Objectives**

The primary objectives of this project are:

* **Data Collection & Preprocessing:** Gather and prepare datasets on weather, light pollution, and tourist reviews in the Northern Territory.

 We will collect and analyse publicly available reviews from platforms such as TripAdvisor, Google Maps, and travel blogs, specifically targeting dark-sky locations and stargazing experiences in the NT.

* Noted: Trying to find a way to collect data from travel blogs and make it useful for analysing

 These reviews provide authentic visitor feedback on the quality of the night sky, facilities, accessibility, and overall experience.

* **Exploratory Data Analysis (EDA):** Analyze the processed data using Python to identify trends in weather suitability, light pollution intensity, safety, services, and tourist sentiment.
* **Modeling:** Develop a simple model to rank potential stargazing locations by combining various factors into a single "Stargazing Score".

 By applying sentiment analysis and topic modelling, we can identify positive and negative themes (e.g., safety, accessibility, weather, atmosphere, tour quality).

 This enables us to quantify tourist satisfaction and highlight areas for improvement.

* **Recommendations:** Provide data-driven insights and recommendations to promote stargazing tourism in the NT.
  + Insights from the analysis will lead to practical recommendations for local businesses and government agencies, such as infrastructure upgrades, event organisation (e.g., astronomy festivals), or targeted marketing to promote the NT as a premier dark-sky tourism destination.
* **Presentation:** Present the findings and recommendations in a clear and concise 10-minute slide presentation.

**3. Data & Analysis**The project will involve the following data and analysis steps:

* **Data Sources:**
  + Weather data from sources like BOM and the OpenWeather API.
  + Light pollution data from VIIRS night-time satellite imagery.
  + Tourism review data from platforms such as TripAdvisor, Google Reviews, and other sample datasets.
* *Choose* ***4–6 NT dark-sky/stargazing spots*** *(e.g., national parks, tours, lookouts).  
  Define what “good” looks like:*
* *≥500 total reviews across all locations (target).*
* *At least 3 figures that tell a story (sentiment by site, top pain points, DSSI ranking).*
* *5–7 specific recommendations.*
  + - ***Checklist***
* *Location list & coordinates (rough).*
* *Short rationale for each site (remote, low light, famous view).*
  + - ***Fields to capture***
* *platform, location\_name, review\_id, review\_text, rating (1–5), review\_date, reviewer\_home (if shown), url*
* **Data Cleaning & Preprocessing:**
  + Standardize data formats for dates, units, and coordinates.
  + Sentiment analysis (positive, negative, neutral classification).
  + Topic modelling to group common themes (safety, weather, experience quality, facilities).
  + Handle missing values through dropping or imputation.
  + Create a consistent schema including   
    location, lat, lon, date, sky\_brightness, cloud\_cover, temperature, and sentiment\_score.
* **Exploratory Data Analysis (EDA):**
  + **Weather Suitability:** Analyze cloud cover trends by season.
  + **Light Pollution:** Map light pollution "hotspots" to identify areas with high intensity.
  + **Tourist Feedback:** Conduct sentiment analysis on tourist reviews to understand sentiment distribution.
  + **Visualizations:** Create figures using Python, such as histograms of cloud cover, line charts for seasonal variations, scatter maps of top spots, and word clouds of tourist reviews.
* **Modeling:**
  + **Index Creation:**
    - Weather index: A calculation based on cloud cover percentage and temperature comfort.
    - Darkness index: An inverse score of light pollution.
    - Sentiment index: The percentage of positive reviews.
  + **Stargazing Score:** Combine the above indices into a single "Stargazing Score" to rank locations.

**4. Recommendations**

Based on the analysis, the project will provide 2-3 strong recommendations, which may include:

* Promoting the highest-scoring stargazing sites.
* Organizing astronomy events, particularly during the dry season.
* Implementing measures to protect dark sky areas from urban light pollution.

**5. Deliverables**

The final deliverables for this project are:

* A clean dataset accompanied by a data dictionary.
* A Python Notebook containing the EDA, model, and visualizations.
* A slide deck of 8-10 slides for the presentation.
* A 10-minute live presentation with visual aids.

**6. Team Roles & Timeline**The project will follow a one-week timeline with defined roles for the team members:

* **Days 1-2:** Data collection and cleaning.
* **Days 3-4:** EDA and Python visualizations.
* **Day 5:** Model building for the Stargazing Score.
* **Day 6:** Drafting the presentation slides and assigning speaking roles.
* **Day 7:** Rehearsing the final presentation.

**PHASED PLAN**

## **Phase 1: Project Setup & Planning (Day 1 – Morning)**

**Tasks:**

* Confirm **scope and goals** (dark-sky tourism analysis, deliverables).
* Identify **success criteria**:
  + ≥500 reviews across all sites.
  + At least 3 clear figures.
  + 5–7 actionable recommendations.
* Finalize **team roles**:
  + Data Collection Lead
  + Data Cleaning/Preprocessing Lead
  + EDA & Modelling Lead
  + Visualization & Presentation Lead
* Select **4–6 dark-sky sites** (e.g., Uluru-Kata Tjuta, Alice Springs Desert Park, Devils Marbles, Kakadu NP lookouts, remote outback tour camps).
* Gather coordinates + short rationale for each site (e.g., “remote, low light, iconic landscape”).
* Set up tools: GitHub repo, Python Notebook structure, shared drive for raw data, Kanban board (Trello/Notion).

**Outputs:**

* Project charter (short doc).
* Finalized site shortlist with lat/lon and rationale.
* Shared workspace setup.

## **Phase 2: Data Collection (Day 1 – Afternoon & Day 2 – Morning)**

**Sub-Tasks:**

1. **Weather Data**
   1. Use **BOM historical data** + **OpenWeather API**.
   2. Collect: daily cloud cover %, temperature, rainfall.
   3. Export to CSV with fields: date, location, cloud\_cover, temperature, rain.
2. **Light Pollution Data**
   1. Download **VIIRS satellite night-time imagery**.
   2. Clip to NT boundaries (geospatial preprocessing in QGIS or Python).
   3. Extract **sky brightness index per site**.
3. **Tourist Reviews**
   1. Collect from **TripAdvisor, Google Maps, travel blogs**.
   2. Extract fields: platform, location, review\_id, review\_text, rating, review\_date, reviewer\_home, url.
   3. Use scraping tools or available CSV datasets.
   4. Ensure minimum target: 500 total reviews.

**Outputs:**

* 3 raw datasets: Weather, Light Pollution, Reviews.
* Metadata notes (source, collection method).

## **Phase 3: Data Cleaning & Preprocessing (Day 2 – Afternoon & Day 3 – Morning)**

**Steps:**

* Standardize formats:
  + Convert review dates → ISO format.
  + Convert coordinates → decimal degrees.
  + Normalize temperature units (°C).
* Handle missing values:
  + Drop empty reviews.
  + Impute missing weather values with seasonal averages.
* **Text Cleaning (Reviews):**
  + Tokenize words, remove stopwords/punctuation.
  + Lemmatization for consistency.
* **Sentiment Analysis:**
  + Use pre-trained Python library (TextBlob/VADER).
  + Assign sentiment score: Positive / Neutral / Negative.
* **Topic Modelling:**
  + Apply LDA (Latent Dirichlet Allocation).
  + Extract common themes (safety, accessibility, facilities, atmosphere).
* Create unified schema:  
   location, lat, lon, date, sky\_brightness, cloud\_cover, temperature, sentiment\_score.

**Outputs:**

* Clean dataset (CSV).
* Data dictionary describing fields & transformations.

## **Phase 4: Exploratory Data Analysis (Day 3 – Afternoon & Day 4 – Morning)**

**Analyses:**

1. **Weather Suitability**
   1. Seasonal cloud cover patterns (dry vs wet season).
   2. Histogram of cloud cover % by season.
   3. Line charts for seasonal variation.
2. **Light Pollution**
   1. Create scatter map of sites with VIIRS brightness overlay.
   2. Compare sky brightness index by site.
3. **Tourist Feedback**
   1. Sentiment distribution bar chart (per site).
   2. Word clouds of positive vs negative reviews.
   3. Identify top 5 pain points (e.g., safety, lack of toilets, access roads).

**Outputs:**

* Python Notebook with EDA visuals: histograms, line charts, maps, word clouds.
* Early insights summary: “Site A has best skies but poor accessibility,” etc.

## **Phase 5: Modeling & Index Creation (Day 4 – Afternoon & Day 5)**

**Steps:**

* **Weather Index:** Weighted score combining:
  + Avg. seasonal cloud cover (weight 0.6).
  + Avg. night-time comfort temperature (weight 0.4).
* **Darkness Index:**
  + Inverse of VIIRS brightness (lower light = higher score).
* **Sentiment Index:**
  + % of positive reviews per site.
* **Stargazing Score:**
  + Formula: 0.4\*WeatherIndex + 0.3\*DarknessIndex + 0.3\*SentimentIndex.
  + Normalize scores (0–100).
* Rank all sites.

**Outputs:**

* Ranked list of NT sites with Stargazing Score.
* Justification per rank.

## **Phase 6: Recommendations & Insights (Day 6 – Morning)**

**Tasks:**

* Translate findings into **5–7 recommendations**, e.g.:
  + Promote top 2–3 ranked sites in NT tourism campaigns.
  + Host **astronomy festivals** during May–September (dry season).
  + Improve site facilities (lighting, toilets, signage, safety).
  + Protect skies through **light pollution regulations** near Alice Springs/Darwin outskirts.
  + Partner with Indigenous rangers for **cultural astronomy tours**.
* Estimate potential **economic impact** (extra nights stayed × avg. visitor spend).

**Outputs:**

* Written recommendation brief (2–3 pages).
* Links to data insights.

## **Phase 7: Presentation Preparation (Day 6 – Afternoon & Day 7 – Morning)**

**Tasks:**

* Create **8–10 slide deck**, covering:
  + Introduction & Background.
  + Problem Statement.
  + Data Sources.
  + Cleaning & Preprocessing.
  + EDA Highlights.
  + Stargazing Score Results.
  + Recommendations.
  + Economic Value Estimate.
  + Conclusion.
* Visual design:
  + Include maps, charts, icons.
  + Keep text minimal, visuals strong.
* Assign **speaking roles**:
  + Intro & problem statement.
  + Data & analysis.
  + Results & recommendations.
* Practice for 10 minutes (timed run).

**Outputs:**

* Slide deck (PowerPoint/Google Slides).
* Speaker notes.

## **Phase 8: Final Delivery (Day 7 – Afternoon)**

**Steps:**

* Rehearse 2–3 full runs with Q&A.
* Deliver live 10-minute presentation.
* Submit:
  + Clean dataset + dictionary.
  + Python Notebook.
  + Slide deck.

**Outputs:**

* Completed project package delivered to stakeholders.