# Capstone Project Datasets Summary

This document provides a summary of the key datasets used in the Springboard Data Science Capstone Project focused on ski resorts. Each dataset is briefly described along with its typical use in the analysis pipeline.

|  |  |  |
| --- | --- | --- |
| Dataset Name | Filename (if applicable) | Description |
| ski\_data | ski\_data.csv | Core resort-level dataset with features like ticket prices, runs, vertical drop, snowmaking, etc. |
| state\_summary | Derived/cleaned version | Aggregated resort data at the state level, summarizing key metrics by state. |
| state\_summary\_scale | Derived from state\_summary | Standardized version of state\_summary used for PCA input. |
| usa\_states\_sub | usa\_states.csv | State-level population and area data used for contextual enrichment. |
| state\_price\_means | Derived via .groupby() | Average ticket prices by state, used to calculate quartiles. |
| pca\_df | Derived | Contains PCA results including components, quartiles, and state labels for visualization. |

Data Flow Summary:

1. ski\_data → cleaning/filtering → state\_summary  
2. state\_summary → standardization → state\_summary\_scale  
3. state\_summary\_scale → PCA → pca\_df  
4. usa\_states\_sub → merged with state\_summary or pca\_df  
5. state\_price\_means → used to calculate quartiles for visualization