# 📄 Data Documentation – Adventure Works Analytics

Last Updated: 24/06/2025

## 🔹 Overview

This document provides full technical documentation of the data cleaning and transformation process for the Adventure Works analytics project. It outlines each dataset, the steps taken in Power Query and Excel, and key decisions regarding field normalization, deduplication, type formatting, and structural enhancements.

## 📦 Dataset Inventory

- Product  
- Sales  
- Reseller  
- SalesTerritory (Region)  
- Salesperson  
- SalespersonTerritory  
- SalespersonTarget  
- Dates (Date Dimension)

## 🧼 Data Cleaning Process Summary

- Removed nulls and duplicate rows where applicable  
- Standardized column names (e.g., 'Unit Price' → 'UnitPrice')  
- Converted data types (strings to numbers/dates)  
- Trimmed and lowercased textual values to prevent mismatches  
- Split or merged columns as needed  
- Parsed all date fields to actual Date format  
- Added calculated fields (e.g., Revenue = Quantity \* UnitPrice, Profit = Revenue - Cost)  
- Created surrogate keys where needed to ensure uniqueness

## 📆 Dates Table

Created a dedicated Dates table to support time-based analysis across the model.  
- Fields: FullDate (Date), Year, MonthNum, DayNum, YearWeekNum  
- Linked to Sales[OrderDate] and SalespersonTarget[TargetMonth]  
- Enables filtering, grouping, and trend analysis by consistent time grain

## 📐 Data Model Structure

The data model uses a star schema centered on the Sales fact table. It connects to multiple dimension tables:  
- Product, Reseller, SalesTerritory, Salesperson  
- The Date table bridges Sales and SalespersonTarget by FullDate  
- One-to-many relationships were enforced where possible, and cross-filtering is managed via Date

## 🛠 Key Transformations & Highlights

• Replaced column `sales` with calculated `Revenue = UnitPrice \* Quantity` using Power Query

• Added `Profit = Revenue - Cost` as custom column in Power Query

• Created `ProductNameKey` in Product table via grouping (by Name, Color, Subcategory, Category)

• Dropped unused columns: BackgroundColorFormat, FontColorFormat, etc.

• Cleaned BusinessType, ResellerName, City, State, Country for consistency

• Unified EmployeeKey and EmployeeID across tables to ensure referential integrity

• Renamed inconsistent columns like 'Reseller' → 'ResellerName', 'Title' → 'JobTitle'

## ✅ Final Note

The dataset is now fully cleaned, structured, and analysis-ready. It integrates well across all key entities and time dimensions, enabling rich and reliable BI reporting.