Preprocessing steps:

Customers:

**ZIP Code Handling**:

* customer\_zip\_code\_prefix is cast to an integer (zip\_prefix\_int) for numeric joins.
* A zero-padded 5-character string version (zip\_prefix) is created using LPAD() to maintain formatting and enable string-based joins.

**Text Normalization**:

* customer\_city: Trimmed and converted to proper case using INITCAP(TRIM(...)).
* customer\_state: Trimmed and converted to uppercase using UPPER(TRIM(...)).

NOT NULL tests on:

* customer\_id
* customer\_unique\_id
* zip\_prefix\_int and zip\_prefix
* customer\_city and customer\_state

These tests ensure **referential integrity**, **completeness**, and **cleanliness** of dimension keys and attributes.

Geolocation:

**Preprocessing Steps**:

1. **ZIP Code Normalization**:
   * Similar to customers, two versions of ZIP code are created: numeric (zip\_prefix\_int) and padded string (zip\_prefix).
2. **Geospatial Cleaning**:
   * geolocation\_lat and geolocation\_lng retained as coordinates for spatial mapping.
3. **Text Cleaning**:
   * geolocation\_city: Trimmed and converted to proper case.
   * geolocation\_state: Trimmed and uppercased.

**Sanity Checks**​:

* NOT NULL tests on all columns:
  + zip\_prefix\_int, zip\_prefix, geolocation\_lat, geolocation\_lng
  + geolocation\_city, geolocation\_state
* These checks ensure all records are complete and usable for **location joins**, **spatial analysis**, or **regional aggregations**.

**stg\_order\_items**

**Preprocessing Steps**:

1. **Typed Fields**:
   * price and freight\_value are explicitly cast to numeric types for consistent cost calculations.
2. **Date Parsing**:
   * shipping\_limit\_dt is parsed into both timestamp and date-only formats:
     + shipping\_limit\_ts: precise timestamp for shipping cutoff.
     + shipping\_limit\_date: normalized date for daily aggregations and filters.

**Sanity Checks**​:

* NOT NULL tests on critical foreign keys:
  + order\_id, order\_item\_id, product\_id, seller\_id.
* shipping\_limit\_ts and shipping\_limit\_date are validated as not null to ensure reliable scheduling logic.
* These ensure **referential integrity**, **temporal completeness**, and **financial analysis readiness**.

**💳 stg\_order\_payments**

**Preprocessing Steps**:

* This table contains one or more payment records per order, often with varying types and installments.
* Preprocessing focuses on:
  + **Data typing** for payment\_value and payment\_installments.
  + **Sequential indexing** via payment\_sequential.

**Sanity Checks**​:

* NOT NULL for all core fields: order\_id, payment\_type, payment\_value, payment\_installments.
* **Uniqueness constraint** on (order\_id, payment\_sequential) ensures no duplicate payment records per order.
* These validations prevent **double-counting** and ensure **complete financial transaction records**.

**📝 stg\_order\_reviews**

**Preprocessing Steps**:

* Review data includes **qualitative and quantitative fields**.
* Derived field: response\_time\_days calculates how long it took to respond to a review (e.g., delivery to review timestamp).
* review\_score is extracted and validated.

**Sanity Checks**​:

* NOT NULL on:
  + review\_id, order\_id, and review\_score.
* **Value constraints**:
  + review\_score must be **between 1 and 5**, as defined by dbt\_expectations.
  + response\_time\_days must be **greater than or equal to 0** (no negative delays).
* Optional: order\_id may be linked to stg\_orders for referential integrity.

**📬 stg\_orders**

**Preprocessing Steps**:

* The table likely extracts and converts **timestamps** like order\_purchase\_ts to standard datetime formats.
* order\_status is cleaned and used for downstream business logic (e.g., fulfillment tracking, cancellations).

**Sanity Checks**​:

* NOT NULL + UNIQUE constraint on order\_id: Ensures **no duplicates or missing order records**.
* NOT NULL on customer\_id: Every order must belong to a valid customer.
* order\_status: Must be one of the accepted values:
  + 'created', 'approved', 'invoiced', 'processing', 'shipped', 'delivered', 'canceled', 'unavailable'
* order\_purchase\_ts: Required for time-based aggregation and analytics.

✅ These checks enforce **order traceability**, **data consistency**, and support advanced logic such as delivery delay calculation or status-based filtering.

**📦 stg\_products**

**Preprocessing Steps**:

* Likely includes casting of physical features like weight and dimensions.
* Calculation of volume\_cm3 is derived from height × width × length fields.
* category\_pt is standardized to represent product taxonomy in Portuguese.

**Sanity Checks**​:

* NOT NULL + UNIQUE on product\_id: Ensures product uniqueness.
* NOT NULL on category\_pt: Mandatory for category-level insights.
* weight\_g: Expected range between 50g and 50,000g.
* volume\_cm3: Must be between 10 cm³ and 500,000 cm³ (to prevent outliers).

✅ These validations guarantee **physical plausibility**, aid in **freight cost estimation**, and prevent outlier-driven skew in shipping or inventory models.

**🛒 stg\_sellers**

**Preprocessing Steps**:

* Standardization of seller\_city and seller\_state through trimming and uppercasing.
* ZIP prefix (zip\_prefix) may be padded similarly to the customers/geolocation table for join compatibility.

**Sanity Checks**​:

* NOT NULL + UNIQUE on seller\_id: Ensures reliable joins to sales and products.
* NOT NULL on zip\_prefix, seller\_city, seller\_state: Needed for regional logistics, mapping, and joins with geolocation.

✅ These constraints ensure **geographical completeness** and facilitate analysis like **seller density**, **delivery distances**, and **regional performance**.