## **Categorical Data Value Check Report**

This report compares the actual unique values found in the dataset's categorical columns against the expected values listed in the Description of Variables.docx document.

**1. First Language:**

* **Found:** 1.0, 2.0, 3.0, NaN
* **Expected:** 1 (English), 2 (French), 3 (Other)
* **Status:** **OK**. All found codes match the expected codes. NaN indicates missing values.

**2. Funding:**

* **Found:** 1, 2, 4, 5, 8, 9
* **Expected:** 1 through 9 (various funding types)
* **Status:** **OK**. All found codes are within the expected range defined in the description. (Codes 3, 6, 7 were expected but not found in the data).

**3. School:**

* **Found:** Only 6
* **Expected:** 1 through 7 (Advancement, Business, Communications, Community/Health, Hospitality, Engineering, Transportation)
* **Status:** **DISCREPANCY**. The dataset *only* contains code 6 (Engineering). All other expected school codes (1-5, 7) are missing. This suggests the dataset might be filtered for only Engineering students, or there's a data issue.

**4. Fast Track:**

* **Found:** 1, 2
* **Expected:** 1 (Y), 2 (N)
* **Status:** **OK**.

**5. Coop:**

* **Found:** 1, 2
* **Expected:** 1 (Y), 2 (N)
* **Status:** **OK**.

**6. Residency:**

* **Found:** 1, 2
* **Expected:** 1 (Domestic), 2 (International)
* **Status:** **OK**.

**7. Gender:**

* **Found:** 1, 2, 3
* **Expected:** 1 (Female), 2 (Male), 3 (Neutral)
* **Status:** **OK**.

**8. Prev Education:**

* **Found:** 0.0, 1.0, 2.0, NaN
* **Expected:** 1 (HighSchool), 2 (PostSecondary)
* **Status:** **DISCREPANCY**. Found an unexpected code 0.0 (88 times) which is not defined in the description. NaN indicates missing values. Recommend treating 0.0 as missing/unknown unless its meaning can be clarified.

**9. Age Group:**

* **Found:** 1.0 through 8.0, NaN
* **Expected:** 1 through 10 (various age ranges)
* **Status:** **OK**. All found codes are within the expected range. (Codes 9 and 10 were expected but not found in the data). NaN indicates missing values.

**10. English Grade:**

* **Found:** 1.0, 2.0, 3.0, 4.0, 7.0, 8.0, 9.0, 10.0, NaN
* **Expected:** 1 through 11 (various levels)
* **Status:** **OK**. All found codes are within the expected range. (Codes 5, 6, 11 were expected but not found in the data). NaN indicates missing values.

**Summary of Discrepancies:**

* **School:** Only code 6 (Engineering) is present. This is highly unusual and needs investigation.
* **Prev Education:** Code 0.0 is present but not defined in the documentation. Recommend treating as missing/unknown.