```
[39]
✓ 0s
# Fill missing PED_ROLE values with 'UNKNOWN'
df_persons['PED_ROLE'] = df_persons['PED_ROLE'].fillna('UNKNOWN')

print("Missing PED_ROLE values have been imputed with 'UNKNOWN'.")

Missing PED_ROLE values have been imputed with 'UNKNOWN'.

[40]
✓ 0s
# Count missing values
missing_count = df_persons['PED_ROLE'].isnull().sum()

# Calculate percentage
missing_percent = (missing_count / len(df_persons)) * 100

print(f"Missing values in PED_ROLE: {missing_count} rows")
print(f"Percentage missing: {missing_percent:.2f}%")

Missing values in PED_ROLE: 0 rows
Percentage missing: 0.00%
```

Fill missing PED_ROLE values with "UNKNOWN" to avoid bias toward the dominant role. This preserves the true distribution of categories. It also keeps imputed records traceable for analysis.

```
[42]
✓ 1s
# Fill missing PERSON_SEX values with 'U'
df_persons['PERSON_SEX'] = df_persons['PERSON_SEX'].fillna('U')

print("Missing PERSON_SEX values have been imputed with 'U'.")

Missing PERSON_SEX values have been imputed with 'U'.

[43]
✓ 0s
# Count missing values
missing_count = df_persons['PERSON_SEX'].isnull().sum()

# Calculate percentage
missing_percent = (missing_count / len(df_persons)) * 100

print(f"Missing values in PERSON_SEX: {missing_count} rows")
print(f"Percentage missing: {missing_percent:.2f}%")

Missing values in PERSON_SEX: 0 rows
Percentage missing: 0.00%
```

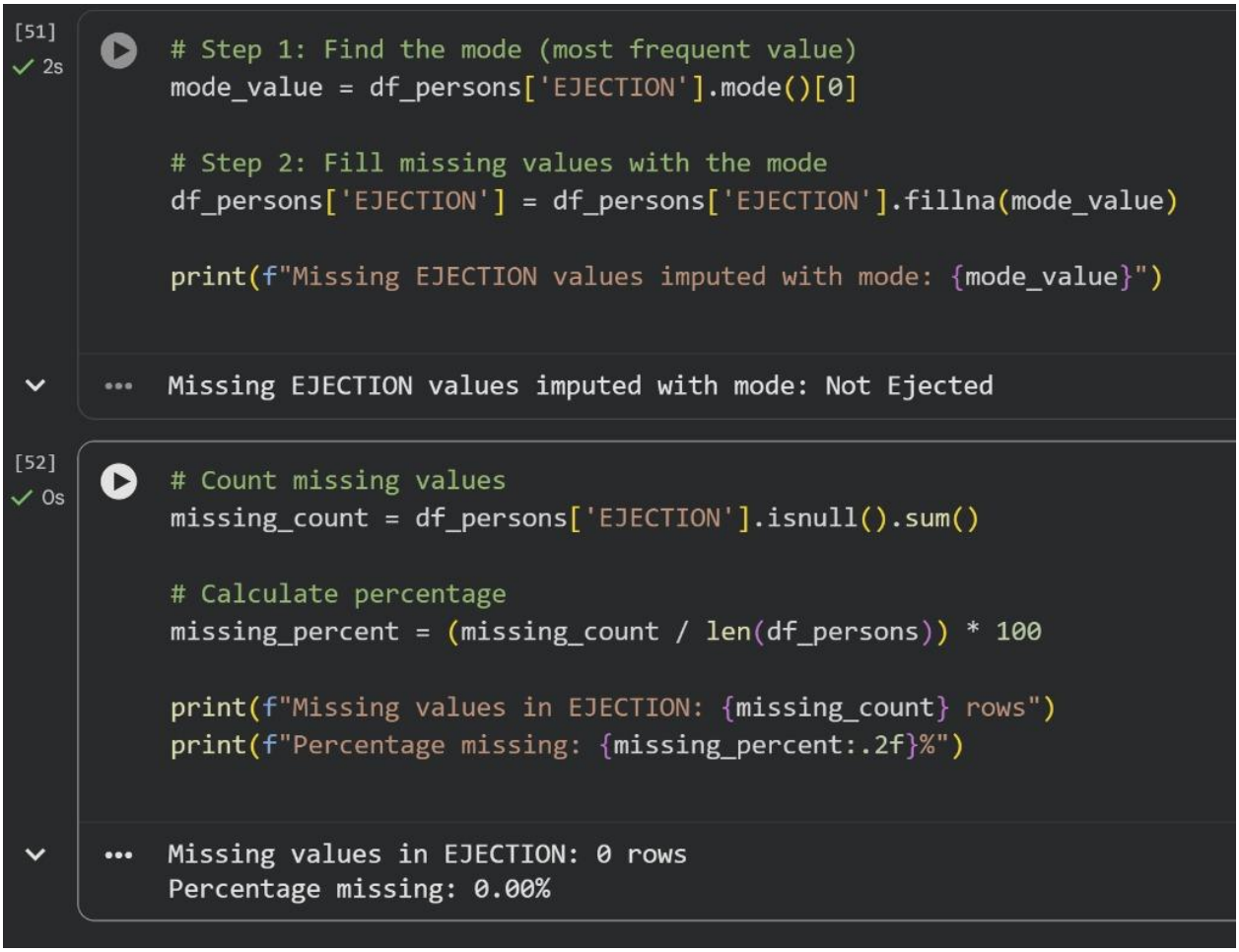
We filled missing PERSON_SEX values with "U" to match the dataset's existing coding scheme.
This avoids bias toward male or female categories during analysis.
It also keeps imputed records traceable and consistent for reproducibility.

```python
import numpy as np

# Replace invalid ages with -1
qt_persons['PERSON_AGE'] = qt_persons['PERSON_AGE'].apply(
    lambda x: x if 0 <= x <= 100 else -1
)

# Impute missing values with sex-specific median
def impute_by_sex(series):
    median_val = series[series != -1].median()
    return series.fillna(median_val).astype(int)

qt_persons['PERSON_AGE'] = qt_persons.groupby('PERSON_SEX')['PERSON_AGE'].transform(impute_by_sex)

print("Missing ages imputed with sex-specific medians; -1 kept for invalid ages.")
```

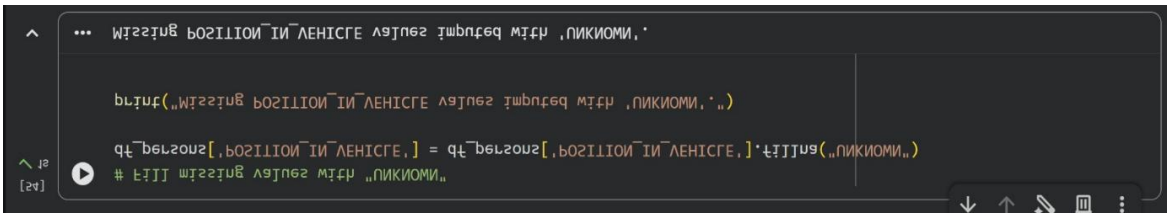
```
Missing ages imputed with sex-specific medians; -1 kept for invalid ages.
```

Imputing missing ages with the median by PERSON_SEX uses a logical demographic link. It avoids bias from unrealistic values while keeping distributions realistic. This ensures consistency and transparency in the dataset.

Unrealistic error outliers (ages <0 or >100) were replaced with -1 as a sentinel marker. This ensures missing values are filled logically while invalid entries remain traceable for analysis.



imputed missing EJECTION values using the mode because it is a categorical variable. The mode represents the most common in this case 97% and therefore most likely category, making it a logical default. This ensures consistency in the dataset while avoiding arbitrary or unrealistic replacements.



We imputed missing POSITION_IN_VEHICLE values with "UNKNOWN" because the variable is categorical. Using "UNKNOWN" avoids introducing bias from mode imputation when the true position cannot be determined. It provides a clear, explicit placeholder that keeps the dataset consistent and traceable for later analysis.