Below is a **comprehensive report** focusing exclusively on the three key RN-related variables in the PBJ Nurse dataset:

1. **Employee Hours (hrs\_rn\_emp)**
2. **Contract Hours (hrs\_rn\_ctr)**
3. **RN Temporary Staffing Ratio (rn\_temp\_ratio)**

We’ll summarize all the insights we’ve gathered so far and propose a few additional code snippets you could run to gain even deeper insights into these variables.

## **1. Overview & Key Observations**

1. **Distribution and Skewness**
   * **hrs\_rn\_emp** (Employee Hours): Ranges from 0 to ~888 hours, with a median of ~24 hours. About 1% of rows exceed ~148 hours.
   * **hrs\_rn\_ctr** (Contract Hours): Ranges from 0 to ~512 hours, with a median of 0 hours. About 1% of rows exceed ~43.6 hours.
   * **rn\_temp\_ratio** (Temporary Staffing Ratio): The ratio of contract to total (contract + employee) hours. Most values are 0, indicating no contract RN usage; about 1% are near 1, indicating full reliance on contract RNs.
2. **Visualizations**
   * **Histograms** showed that the data is heavily skewed:
     + Large clusters at 0 for contract hours and ratio.
     + A small tail of extremely high values for both hrs\_rn\_emp and hrs\_rn\_ctr.
   * **Scatter Plot** (hrs\_rn\_emp vs. rn\_temp\_ratio) revealed:
     + Most points cluster near (x>0, y=0), meaning some employee hours but no contract.
     + A small cluster near y=1, meaning a few facilities/days rely exclusively on contract RNs.
3. **Outliers**
   * The top ~1% of rows exhibit very high employee or contract hours, or an RN ratio near 1.
   * Domain context: Some facilities may legitimately have large aggregated hours or exclusively contract RNs on certain days.

## **2. Existing Code & Results**

We’ve already run the following analyses:

1. **Basic Summary Statistics** (count, mean, std, min, max, quartiles).
2. **Histograms** with various approaches (clipping, log scale, and fixed range).
3. **Percentile Analysis** at the 90th, 95th, and 99th percentiles.
4. **Scatter Plot** to observe the relationship between hrs\_rn\_emp and rn\_temp\_ratio.

From these, we concluded that outliers are real and likely represent legitimate facility-days. We also observed that the majority of rows have zero contract hours, leading to a ratio of zero, while a minority have near-total contract hours.

## **3. Additional Code Snippets for Deeper Insight**

Below are some optional code blocks you can run to further explore these three variables. You can pick and choose which are most relevant to your analysis goals.

### **3.1 Time-Based Analysis**

If your data includes daily entries with a date column (e.g., workdate), you might want to see how these variables evolve over time. The snippet below groups the data by month (or week) and computes the mean or median of each variable.

# Additional Analysis - Time-Based Trends

import pandas as pd

df\_nurse['year\_month'] = df\_nurse['workdate'].dt.to\_period('M') # Group by year-month

time\_trends = df\_nurse.groupby('year\_month').agg({

'hrs\_rn\_emp': 'mean',

'hrs\_rn\_ctr': 'mean',

'rn\_temp\_ratio': 'mean'

}).reset\_index()

print(time\_trends.head())

# Plot the monthly trends

time\_trends.plot(x='year\_month', y=['hrs\_rn\_emp', 'hrs\_rn\_ctr', 'rn\_temp\_ratio'], figsize=(10,5), marker='o')

plt.title("Monthly Average RN Staffing Metrics")

plt.xlabel("Year-Month")

plt.ylabel("Average Value")

plt.show()

**What This Does**:

* Groups by the year\_month period (derived from workdate) and calculates mean values.
* Plots how average employee hours, contract hours, and ratio change over time.

### **3.2 Facility-Level Analysis**

You might want to see if certain facilities are driving these outliers. The snippet below identifies the facilities with the highest average RN ratio or highest total hours.

# Additional Analysis - Facility-Level Outliers

# Group by 'provnum' to see which facilities have the highest average RN ratio

facility\_stats = df\_nurse.groupby('provnum').agg({

'hrs\_rn\_emp': 'sum',

'hrs\_rn\_ctr': 'sum',

'rn\_temp\_ratio': 'mean'

}).reset\_index()

# Sort by the average RN ratio descending

top\_ratio\_facilities = facility\_stats.sort\_values('rn\_temp\_ratio', ascending=False).head(10)

print("Top 10 Facilities by Average RN Temporary Staffing Ratio:")

print(top\_ratio\_facilities)

# Sort by total contract hours descending

top\_contract\_facilities = facility\_stats.sort\_values('hrs\_rn\_ctr', ascending=False).head(10)

print("\nTop 10 Facilities by Total RN Contract Hours:")

print(top\_contract\_facilities)

**What This Does**:

* Sums employee and contract hours at the facility level.
* Computes the average ratio per facility.
* Identifies which facilities rely most heavily on contract RNs.

### **3.3 Correlation with Other Nursing Variables**

If you’d like to see whether high RN contract hours correlate with other nursing roles (e.g., LPN, CNA), you can check a correlation matrix:

# Additional Analysis - Correlation with Other Nursing Variables

nursing\_cols = ['hrs\_rn\_emp', 'hrs\_rn\_ctr', 'rn\_temp\_ratio',

'hrs\_lpn\_emp', 'hrs\_lpn\_ctr', 'hrs\_cna\_emp', 'hrs\_cna\_ctr']

corr\_matrix = df\_nurse[nursing\_cols].corr()

print("Correlation Matrix for Nursing Hours:")

print(corr\_matrix)

**What This Does**:

* Computes pairwise correlations for RNs, LPNs, and CNAs (both contract and employee hours).
* Helps you see if, for instance, high RN contract usage tends to go hand in hand with high LPN contract usage or if they move inversely.

### **3.4 Advanced Outlier Profiling**

If you want to specifically profile the top 1% of outliers, you can do so with code like this:

# Additional Analysis - Profiling Top 1% Outliers

import numpy as np

threshold\_emp = np.percentile(df\_nurse['hrs\_rn\_emp'], 99)

threshold\_ctr = np.percentile(df\_nurse['hrs\_rn\_ctr'], 99)

threshold\_ratio = np.percentile(df\_nurse['rn\_temp\_ratio'], 99)

# Subset the dataset to only rows that exceed at least one threshold

outliers = df\_nurse[

(df\_nurse['hrs\_rn\_emp'] > threshold\_emp) |

(df\_nurse['hrs\_rn\_ctr'] > threshold\_ctr) |

(df\_nurse['rn\_temp\_ratio'] > threshold\_ratio)

]

print(f"Total outlier rows: {len(outliers)}")

# Explore how these outliers differ, e.g., average census, or which states/facilities

grouped\_outliers = outliers.groupby('state')['rn\_temp\_ratio'].agg(['count','mean']).sort\_values('count', ascending=False)

print(grouped\_outliers.head(10))

**What This Does**:

* Identifies any rows that exceed the 99th percentile for hrs\_rn\_emp, hrs\_rn\_ctr, or rn\_temp\_ratio.
* Groups them by state (or another relevant column) to see if certain regions/facilities drive these outliers.

## **4. Summary & Recommendations**

1. **Core Findings**
   * **hrs\_rn\_emp** has a long tail but a median of ~24 hours.
   * **hrs\_rn\_ctr** is zero for most rows, with a 99th percentile near 44 hours.
   * **rn\_temp\_ratio** is 0 for ~75% of rows and near 1 for ~1%—indicating a small fraction of entries rely solely on contract RNs.
2. **Domain Context Matters**
   * Large hour values may be legitimate for big facilities or aggregated shifts.
   * A ratio of 1 could be real if a facility depends entirely on contract RNs.
   * Validate outliers by checking if they align with known facility sizes or special circumstances.
3. **Further Analysis**
   * **Time-based** grouping can reveal trends across weeks or months.
   * **Facility-level** grouping identifies which facilities rely most on contract RNs.
   * **Correlation** with other variables (e.g., LPN, CNA, or external cost metrics) can shed light on broader staffing patterns and cost implications.
4. **Next Steps**
   * Decide whether to keep, exclude, or separately analyze the top 1% outliers.
   * Integrate external datasets (e.g., cost, penalties, ownership details) to see how these RN patterns correlate with financial or regulatory outcomes.
   * Consider a deeper domain review to confirm that extremely high hours or near-1.0 ratios are valid records.

## **Conclusion**

This report consolidates all insights specific to the **employee hours (hrs\_rn\_emp), contract hours (hrs\_rn\_ctr), and temporary staffing ratio (rn\_temp\_ratio)** for RNs in your PBJ Nurse dataset. We’ve provided additional code snippets for time-based, facility-level, and correlation analyses that can further illuminate how these variables behave and how outliers might impact your final conclusions.

If you need additional detail or would like to dive deeper into any particular angle—such as combining these variables with external quality measures or performing predictive modeling—just let me know!