# Swiss Post Sorting Centers Package Sorting Performance Analysis and Prediction

# Objective

The goal of this project is to analyze the postal sorting center's performance by identifying the most influential factors contributing to sorting issues, including shipment attributes (e.g., dimensions, weight, coding stations, and timestamps) as well as chute congestion which we believe have a major impact on overall performance of the center and finally using a model to determine whether overburdened chutes or certain features create bottlenecks that reduce overall system efficiency.

1. **Determine Feature Importance**: Rank the shipment features by their importance and identify which shipment attributes (e.g., dimensions, weight, coding station) are most influential in causing sorting issues at postal centers. As well Correlation between the features and the impact on performance
2. **Determine chute congestion impact:** Determine whether chutes are handling disproportionately large volumes of packages can create bottlenecks that reduce overall system efficiency
3. **Predict Sorting Performance and Issues**: Develop a predictive model capable of forecasting sorting issues based on historical shipment data.
4. **Generate Actionable Insights**: Provide data-driven recommendations for improving chute utilization and enhance overall performance

# Sorting Process

A: Shipments **arrive** and are delivered to the sorting center, using designated units

B: Shipments are **scanned and transferred** to the conveyor belts, where automatic and manual sorting machines determine their route and send them to the appropriate chute based on the destination.

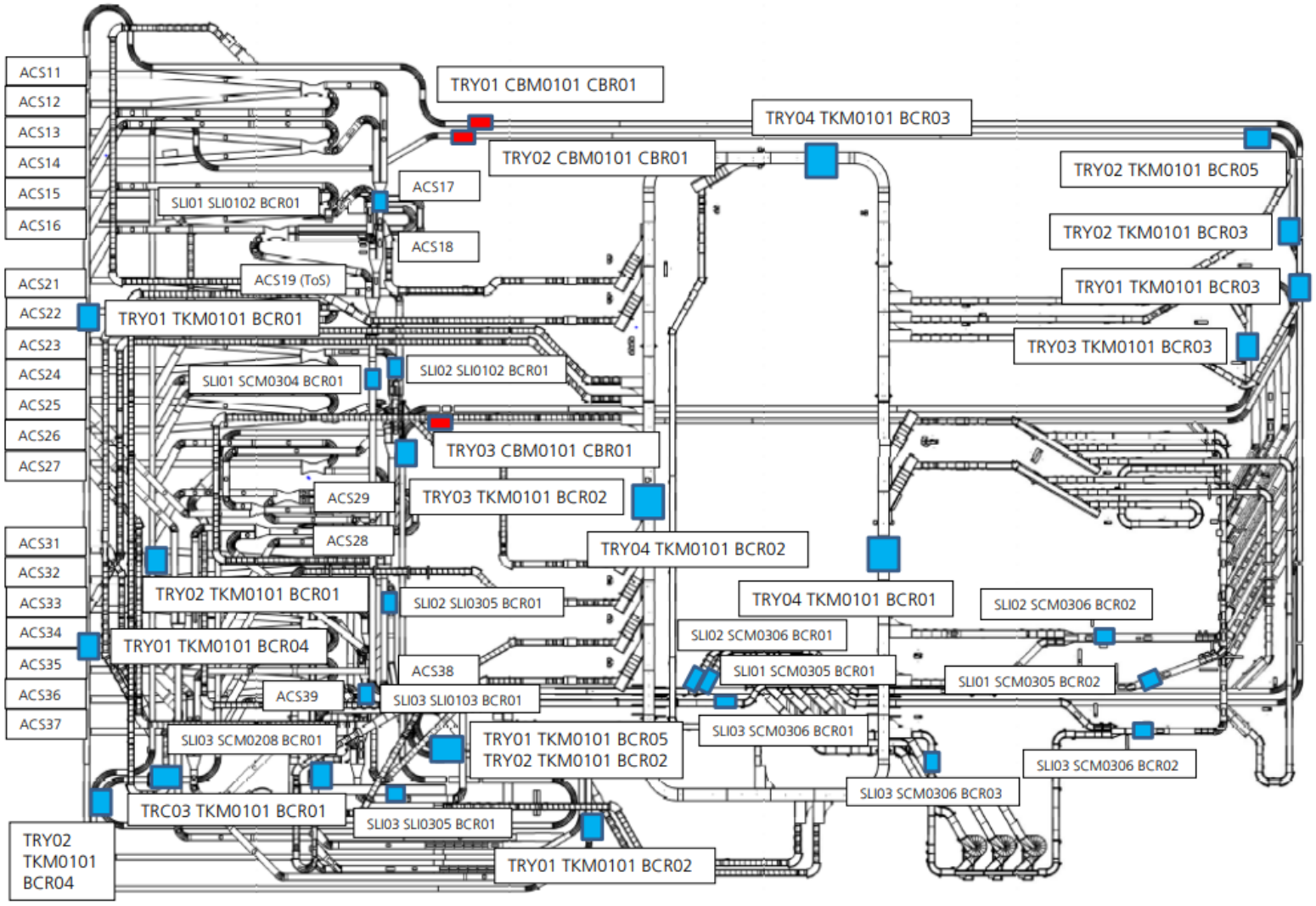
C: The **chute serves as the output** of the sorting machine, directing the parcels to different destinations depending on the ZIP code., one chute can serve several ZIP codes

The center's performance is measured by the number of parcels processed per time frame. Some centers have shown up to a 15% increase in processing efficiency compared to others with similar setups. Further investigation revealed that traffic bottlenecks at certain chutes, which handle significantly higher package volumes, cause an imbalance. This results in a non-normal distribution of packages across the chutes, leading to a noticeable reduction in overall performance.

Our goal is to achieve a balanced, normally distributed flow of packages across all available chutes, which would enhance the sorting rate and improve overall center performance.

# Sorting Process

Übersicht alle Lesesysteme PZ Frauenfeld Stand 2022



# Overview of provided Data

The data for this project comes from Swiss Post's shipment sorting system and includes:

- **Shipment number** (anonymized) SND_IDENTCODE
- **Shipment dimensions**: Length, width, and height (in millimeters) SND_CODS_DIM1, SND_CODS_DIM2, SND_CODS_DIM3
- **Shipment weight**: (n grams) SND_GEW
- **Scanning timestamps** when the item first scanned in the sorting center CODS_COD_DAT
- **Scanner station**: Sorting station identifier CODS_CO_STATION
- **Sorting center Number** CODS_ZENT_NR_x
- **leaving timestamps** when the item left the sorting center chute CODS_LERE_DAT
- **chute station** where the item is sent CODS_SD_RUTSCHE

The Engineered Features

- **Shipment dimensions SUM**: Length+width+height CODS_DIM_SUM as they always corelate together
- **Processing Time : `the processing time` processing_time_minutes = entry time – leaving time**
- **Sorting Performance : `the processing time if longer than 10 min its performance issue 1 else 0`**
- **Minute counter : `the time package leaving Center` as counter from 01.01.2024**

# Overview of provided Data

| Field Name | Description | Data Type | Example |
|---|---|---|---|
| **SND_IDENTCODE** | Unique identifier for each shipment (anonymized for privacy) | String | A12345 |
| **SND_CODS_DIM1** | Length of the shipment in millimeters | Integer | 300 |
| **SND_CODS_DIM2** | Width of the shipment in millimeters | Integer | 150 |
| **SND_CODS_DIM3** | Height of the shipment in millimeters | Integer | 50 |
| **SND_GEW** | Weight of the shipment in grams | Integer | 1000 |
| **CODS_COD_DAT** | Timestamp indicating when the shipment was scanned into the sorting center | Datetime | 15.01.2023 08:32 |
| **CODS_LERE_DAT** | Timestamp indicating when the shipment left the sorting center | Datetime | 15.01.2023 09:45 |
| **CODS_CO_STATION** | Station or scanner ID at which the shipment was processed | String | STATION01 |
| **CODS_SD_RUTSCHE** | Chute identifier where the package was routed for further processing | String | CHUTE10 |
| **processing_time_minutes** | Calculated field representing the time taken to process a shipment in minutes | Float | 73.5 |

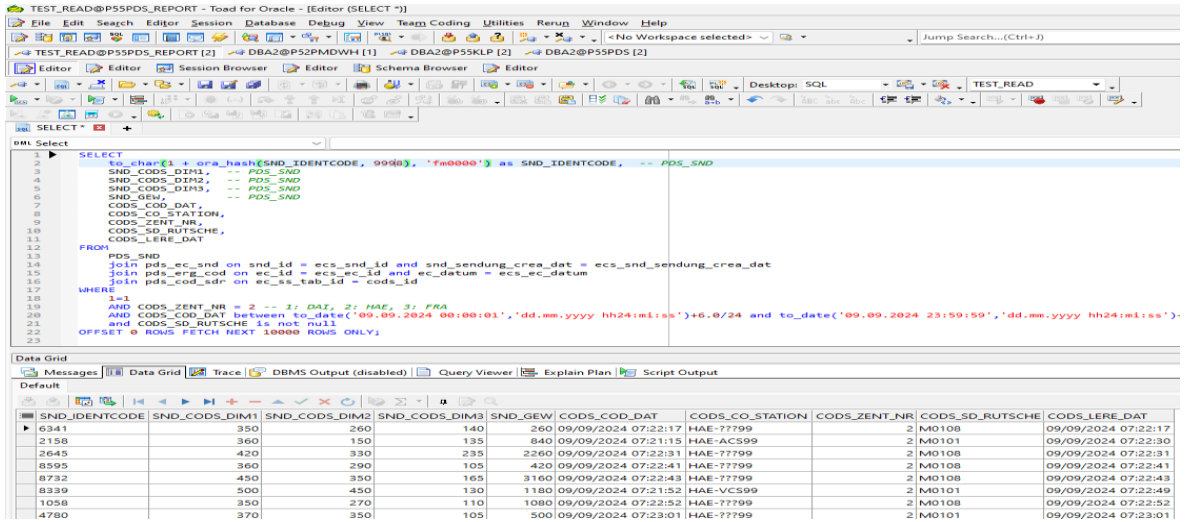
# Overview of provided Data

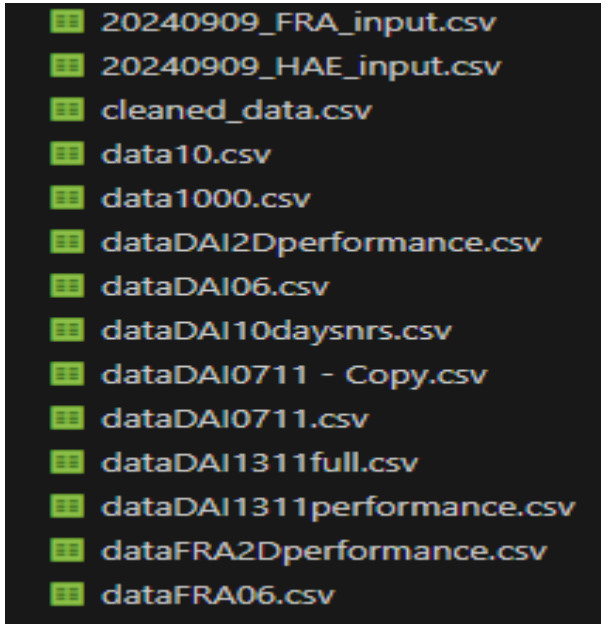
| | CODS_IDENTCODE_AN | CODS_DIM1 | CODS_DIM2 | CODS_DIM3 | CODS_GEW | CODS_ADR_PLZZZ_AN | CODS_CO_STATION | CODS_ZENT_NR | CODS_SD_RUTSCHE | DAY_15MIN | CODS_LERE_DAT | MINUTE_COUNTER | PROCESSING_TIME_MINUTES | SORTING_PERF_ISSUE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 242455661 | 700 | 454 | 54 | 1480 | 449561 | FRA-ACS18 | 3 | R0109 | 202411061015 | 06/11/2024 10:26:25 | 447026 | 1 | 0 |
| 1 | 482169492 | 704 | 454 | 48 | 1520 | 568172 | FRA-ACS18 | 3 | R0109 | 202411061015 | 06/11/2024 10:26:15 | 447026 | 2 | 0 |
| 2 | 407572193 | 400 | 380 | 235 | 4700 | 23223 | FRA-ACS21 | 3 | R2515 | 202411061015 | 06/11/2024 10:26:16 | 447026 | 4 | 0 |
| 3 | 999133381 | 470 | 450 | 120 | 1060 | 629595 | FRA-ACS21 | 3 | R0109 | 202411061015 | 06/11/2024 10:26:04 | 447026 | 7 | 0 |
| 4 | 748472889 | 360 | 290 | 110 | 1480 | 421619 | FRA-ACS21 | 3 | R2319 | 202411061015 | 06/11/2024 10:26:37 | 447027 | 15 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 170012 | 302468774 | 490 | 390 | 90 | 1280 | 712077 | DAI-ACS13 | 1 | R2419 | 202411061430 | 06/11/2024 14:38:26 | 447278 | 4 | 0 |
| 170013 | 225629123 | 170 | 110 | 55 | 120 | 869319 | DAI-ACS13 | 1 | R1223 | 202411061430 | 06/11/2024 14:38:22 | 447278 | 2 | 0 |
| 170014 | 822331195 | 410 | 360 | 55 | 820 | 49468 | DAI-ACS13 | 1 | R2704 | 202411061430 | 06/11/2024 14:38:25 | 447278 | 6 | 0 |
| 170015 | 804061373 | 400 | 310 | 110 | 1640 | 32705 | DAI-ACS13 | 1 | R1709 | 202411061430 | 06/11/2024 14:38:19 | 447278 | 6 | 0 |
| 170016 | 486028131 | 410 | 300 | 110 | 2900 | 608088 | DAI-ACS13 | 1 | R2207 | 202411061430 | 06/11/2024 14:38:32 | 447279 | 2 | 0 |

809022 rows × 14 columns

# Overview of provided Data

```
SELECT LPAD (TO_CHAR (ORA_HASH (cods_identcode, POWER (10, 9) - 1)), 9, '0')
          AS CODS_IDENTCODE_AN,                    --- to anonymize the codeid
       CODS_DIM1,
       CODS_DIM2,
       CODS_DIM3,
       CODS_GEW,
       LPAD (TO_CHAR (ORA_HASH (CODS_ADR_PLZZZ, POWER (10, 6) - 1)), 6, '0')
          AS CODS_ADR_PLZZZ_AN,                --- to anonymize the Postal code
       --CODS_CO_STATION,
       TRANSLATE (CODS_CO_STATION,
                  'ABCDEFGHIJKLMNOPQRSTUVWXYZ=.-?',
                  '12345678912345678912345678000')
          stationin,              -- to convert the coding station to a number
       CODS_ZENT_NR,                                  -- 1: DAI, 2: HAE, 3: FRA
       --CODS_SD_RUTSCHE,
       TRANSLATE (CODS_SD_RUTSCHE,
                  'ABCDEFGHIJKLMNOPQRSTUVWXYZ=.-?',
                  '12345678912345678912345678000')
          rutchout,                     --convert each rutch to unique number
         --CODS_COD_DAT,    CODS_LERE_DAT,
        TO_CHAR (CODS_COD_DAT, 'YYYYMMDDHH24')
       || LPAD (TRUNC (TO_NUMBER (TO_CHAR (CODS_COD_DAT, 'MI')) / 15) * 15,
                2,
                '0')
          AS day_15min,
       TRUNC ((CODS_LERE_DAT - CODS_COD_DAT) * 86400 / 60)
          processing_time_minutes
  FROM pds_cod_sdr
 WHERE     CODS_ZENT_NR = 2                         -- 1: DAI, 2: HAE, 3: FRA
       --and CODS_CO_STATION LIKE 'HAE-ACS%'  AND SUBSTR(CODS_CO_STATION, -2) BETWEEN '11' AND '66'
       -- AND CODS_COD_DAT > SYSDATE - 10
        AND CODS_COD_DAT BETWEEN TO_DATE('2024-11-06 00:00:00', 'YYYY-MM-DD HH24:MI:SS')
                             AND TO_DATE('2024-11-14 23:59:59', 'YYYY-MM-DD HH24:MI:SS')
       AND CODS_SD_RUTSCHE IS NOT NULL
       AND (    CODS_DIM1 IS NOT NULL
            AND CODS_DIM2 IS NOT NULL
            AND CODS_DIM3 IS NOT NULL)
       AND (CODS_LERE_DAT - CODS_COD_DAT) * 86400 / 60 > 30 --the min time taken by the package to be processed
       --  order by CODS_ADR_PLZZZ desc
       AND CODS_COD_DAT BETWEEN TO_TIMESTAMP ('2024-11-01 18:00',
                                              'YYYY-MM-DD HH24:MI')
                            AND TO_TIMESTAMP ('2024-11-17 20:30',
                                              'YYYY-MM-DD HH24:MI')
       AND CODS_ADR_PLZZZ IS NOT NULL
       AND CODS_DIM1 IS NOT NULL
       AND CODS_DIM2 IS NOT NULL
       AND CODS_DIM3 IS NOT NULL
       AND CODS_GEW IS NOT NULL
       AND CODS_CO_STATION IS NOT NULL
       AND CODS_SD_RUTSCHE IS NOT NULL
       AND CODS_ZENT_NR IS NOT NULL
--order by processing_time_minutes desc
```



# Descriptive Statistics

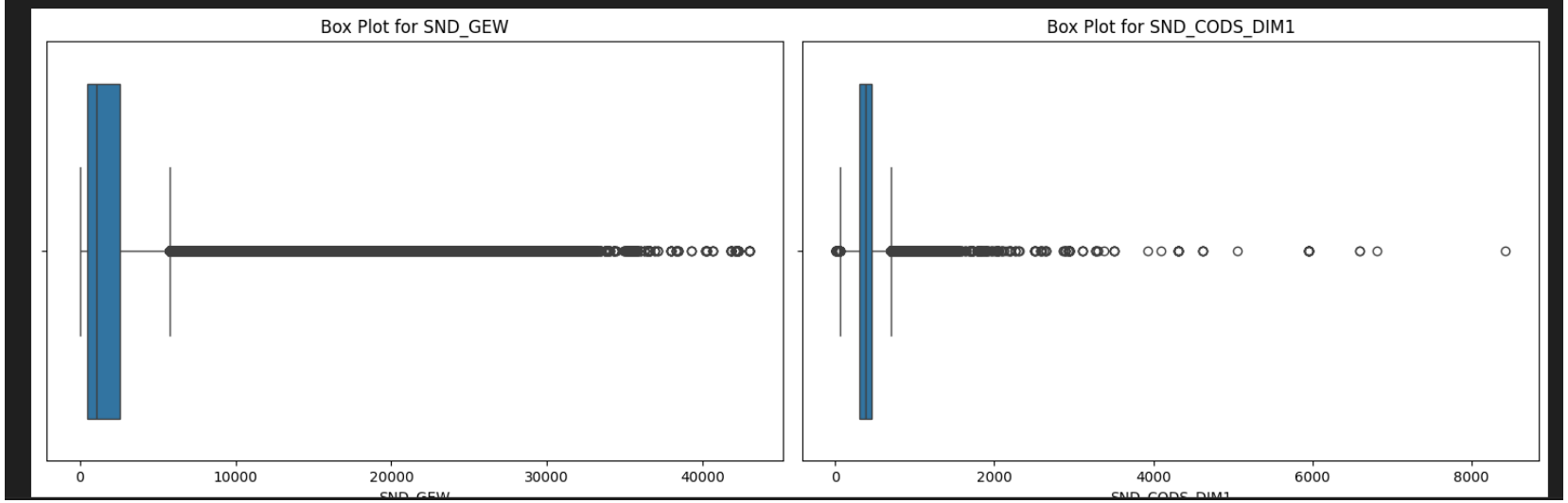
Summary of outliers in SND_GEW:

| | SND_IDENTCODE | SND_CODS_DIM1 | SND_CODS_DIM2 | SND_CODS_DIM3 | \ |
|---|---|---|---|---|---|
| count | 3.984580e+05 | 398385.000000 | 398385.000000 | 398385.000000 | |
| mean | 6.670258e+09 | 565.754303 | 407.914206 | 292.407407 | |
| std | 3.317503e+10 | 157.241636 | 95.041501 | 91.422043 | |
| min | 4.946300e+04 | 10.000000 | 10.000000 | 10.000000 | |
| 25% | 2.482289e+09 | 450.000000 | 340.000000 | 225.000000 | |
| 50% | 5.006381e+09 | 590.000000 | 400.000000 | 295.000000 | |
| 75% | 7.516051e+09 | 620.000000 | 460.000000 | 350.000000 | |
| max | 9.981578e+11 | 5960.000000 | 912.000000 | 740.000000 | |

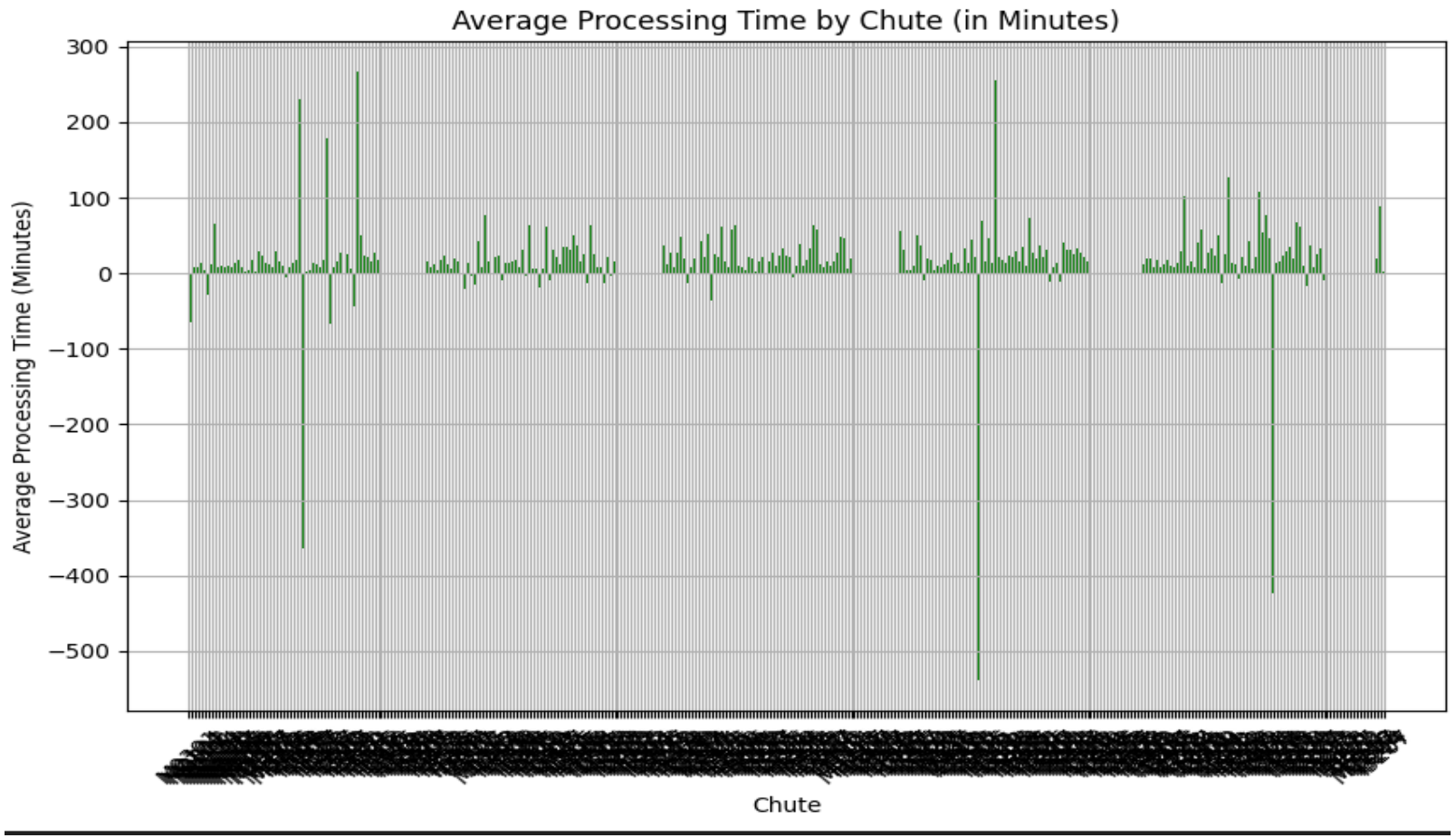
| | SND_GEW | CODS_ZENT_NR_x | CODS_ZENT_NR_y | processing_time_minutes |
|---|---|---|---|---|
| count | 398458.000000 | 398458.000000 | 398458.000000 | 398458.000000 |
| mean | 12179.256396 | 2.372365 | 2.372365 | 0.261907 |
| std | 5848.953016 | 0.483436 | 0.483436 | 348.457689 |
| min | 5785.000000 | 2.000000 | 2.000000 | -6413.600000 |
| 25% | 7560.000000 | 2.000000 | 2.000000 | 1.620000 |
| 50% | 10360.000000 | 2.000000 | 2.000000 | 3.170000 |
| 75% | 15160.000000 | 3.000000 | 3.000000 | 5.000000 |
| max | 42980.000000 | 3.000000 | 3.000000 | 6421.780000 |

Summary of outliers in SND_CODS_DIM1:

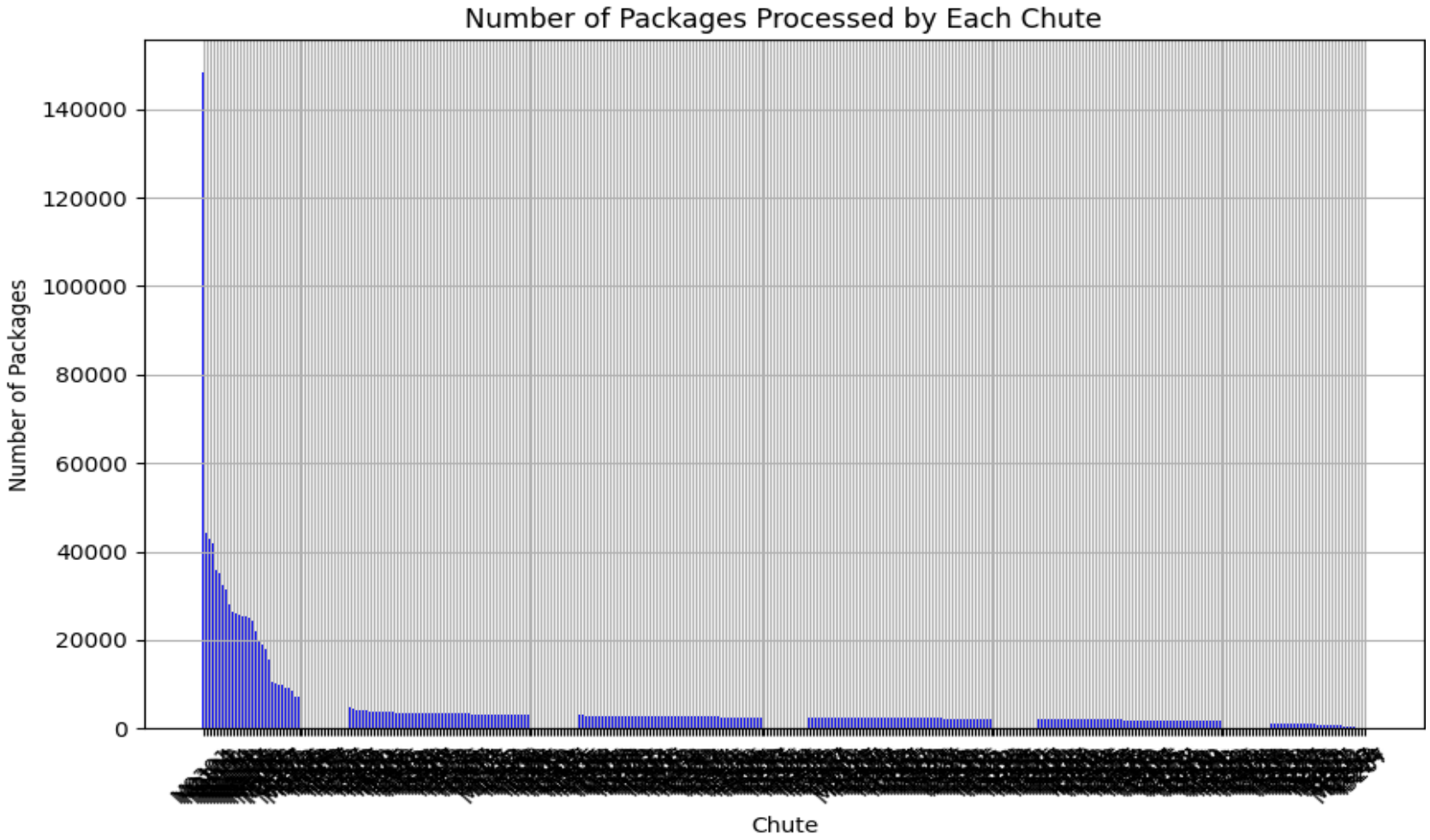
| | SND_IDENTCODE | SND_CODS_DIM1 | SND_CODS_DIM2 | SND_CODS_DIM3 | \ |
|---|---|---|---|---|---|
| count | 1.132030e+05 | 113203.000000 | 113203.000000 | 113203.000000 | |



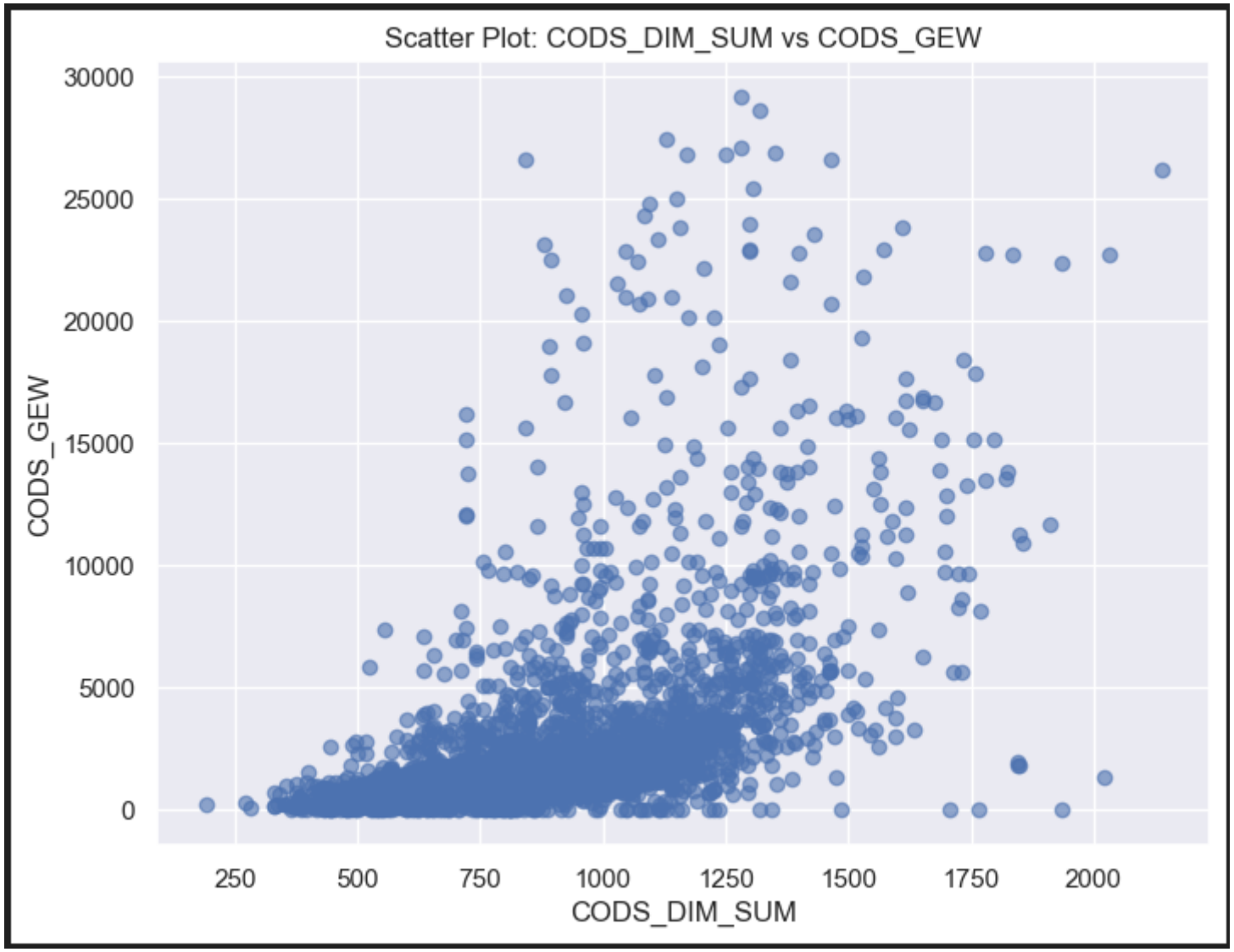
# Descriptive Statistics



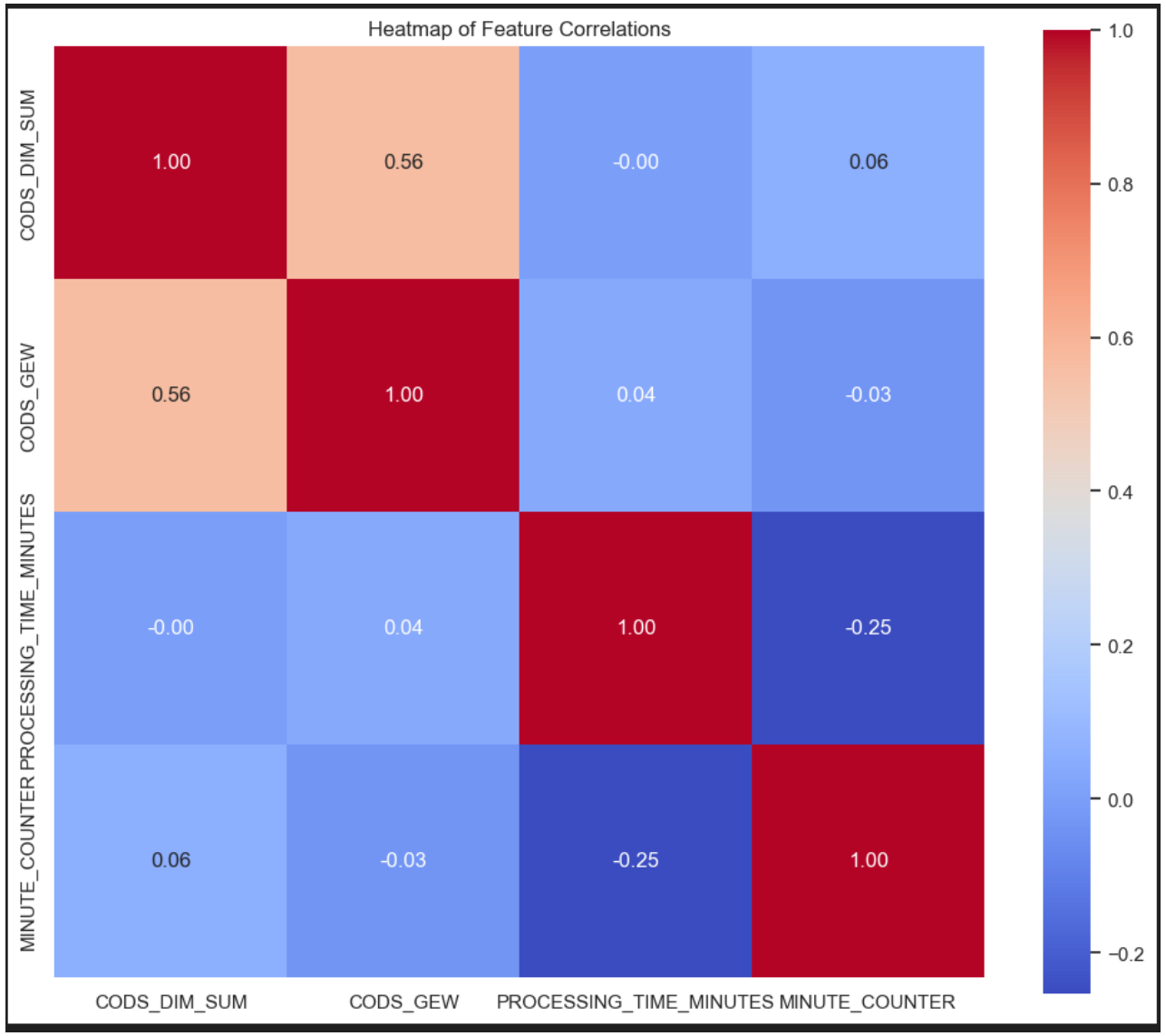
# Descriptive Statistics



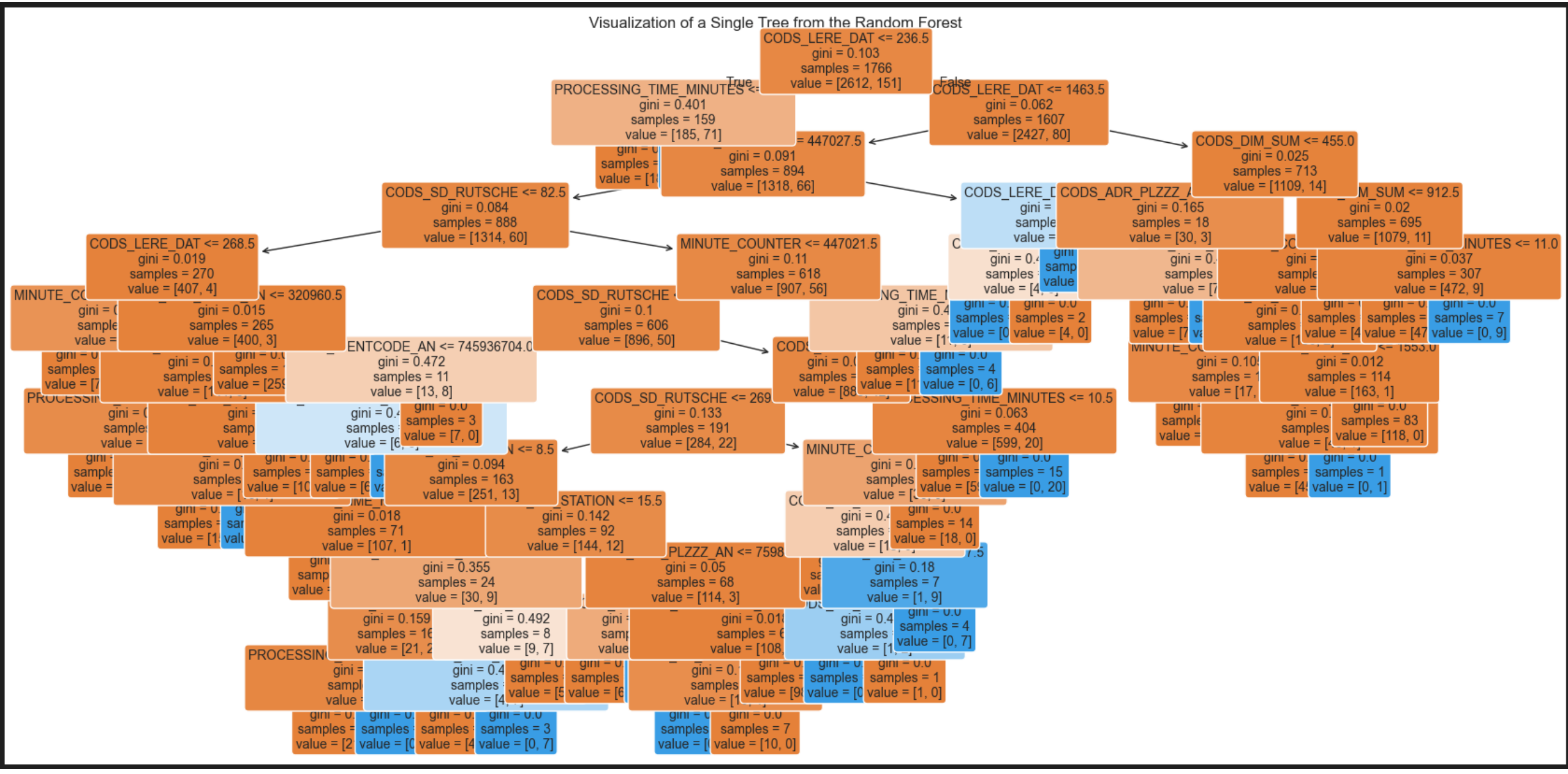
# Descriptive Statistics



# Heat Map


Heatmap of Feature Correlations
CODS_DIM_SUM
CODS_GEW
PROCESSING_TIME_MINUTES
MINUTE_COUNTER
1.00
0.56
-0.00
0.06
0.56
1.00
0.04
-0.03
-0.00
0.04
1.00
-0.25
0.06
-0.03
-0.25
1.00
CODS_DIM_SUM
CODS_GEW
PROCESSING_TIME_MINUTES
MINUTE_COUNTER
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

# Random Forest



# Result of Statistical Tests

Understanding which features most affect sorting issues provides insights that can be used to improve sorting operations at Swiss Post. Focusing on key factors like shipment size, weight, and station performance will help optimize sorting machine performance and reduce errors.

**Key Findings**

**Chute Congestion**: Certain chutes were identified as potential bottlenecks, handling significantly more packages than others and showing longer processing times. Managing chute congestion is critical to improving overall efficiency.

**Processing Time Variability**: There was substantial variability in processing times across shipments. Factors such as shipment dimensions, weight, and chute assignment contributed to this variability.

**Data Quality Issues**: Several data quality issues, such as missing or inconsistent timestamps, were identified. These issues were addressed to ensure accurate analysis, but continued data quality monitoring is recommended.

**Model Insights**

The Random Forest model provided insights into the factors most influencing sorting performance, with shipment weight and chute utilization being significant contributors. However, additional factors not captured in the dataset may also play a role in performance variations.

# Result of Statistical Tests

### Recommendations

**Chute Balancing**: Implement dynamic chute load balancing to distribute shipments more evenly across available chutes. This would reduce bottlenecks and improve throughput.

**Real-Time Monitoring**: Introduce real-time monitoring to detect and address chute congestion before it affects overall performance.

**Further Data Collection**: Collect additional data on shipment characteristics and operational factors to refine the performance models and improve accuracy.

### Next Steps

Continue refining the performance models with updated data and explore additional machine learning techniques to predict sorting center performance under different conditions.

Implement operational changes based on the findings and monitor their impact on sorting efficiency.

# Transitioning from Dimensions to ZIP and Chute-Based Modeling

**Rationale for Transitioning from Package Dimensions to ZIP Code and Chute-Based Modeling**

- **Package volume had minimal impact** on operational performance — the primary driver of chute congestion was the **number of packages**, not their size.
- ZIP code is the **decisive factor in routing**, directly determining chute assignment and downstream load.
- Tracking package dimensions introduced **unnecessary complexity** with **no measurable gain** in predictive accuracy.
- By focusing on **ZIP and chute-level volume**, we improved **model efficiency, reliability, and interpretability**.
- The new approach aligns better with **operational levers** — it's feasible to reallocate ZIP codes across chutes, but not to adjust package size.
- Result: a **leaner, faster, and more actionable model** that directly supports real-time decision-making.

# LSTM

**Why We Chose LSTM for This Project**

**1. Nature of the Problem: Sequential and Time-Dependent**

•Our data reflects **time-series behavior** — package volumes fluctuate over time, influenced by hour-of-day, ZIP routing patterns, and operational cycles.
•Chute congestion is not static; it depends on **past trends**, like package surges building up over time.
•We needed a model that could **learn from historical patterns** to predict upcoming performance issues.

**2. LSTM (Long Short-Term Memory) Is Built for Time-Series**

•LSTM is a specialized form of Recurrent Neural Network (RNN) designed to **retain memory over time**.
•Unlike standard models, LSTM can capture **long-term dependencies**, such as consistent overloads during specific hours or recurring ZIP surges.
•It excels at detecting trends and seasonality — **perfect for forecasting operational bottlenecks**.

**3. Handles Non-Linear and Complex Patterns**

•Chute performance is impacted by **non-linear combinations** of ZIP code volume, station load, and time.
•LSTM handles these **complex interactions** far better than traditional regression or rule-based models.

# LSTM Model Performance Comparison: All Chutes vs Top 20 Chutes

Original Model (All Chutes, Hourly Aggregation): Input Data: Included all chutes over 30 days, grouped by hour.

Problem: The vast majority of records had no performance issues (avg_processing_time_minutes < 10).

Consequence:
The model learned to predict near-zero processing time.
Overfitting to "normal" behavior → Poor at detecting or predicting real problems.
Forecasts were always flat and missed performance spikes.

Refined Model (Top 20 Chutes with Known Issues, 10-Minute Aggregation):

Input Data: Limited to top 20 chutes that had documented performance issues in the past 10 days.
Granularity: Switched to 10-minute intervals to better capture short-term load patterns.

Model now trains on meaningful variation and real issue patterns.
Reduced noise and class imbalance.
Much better fit for both historical prediction and 6-hour forecasting.
Early results show the model tracks spikes and processing slowdowns more accurately.

| Model Version | Data Scope | Granularity | Issue Presence | Prediction Quality |
|---|---|---|---|---|
| Original (Baseline) | All chutes | Hourly | Mostly 0s | Poor (overfitting) |
| Refined (Top 20) | Problematic only | 10-minutes | Real issues | Improved accuracy |

# What are we Predicting

**Primary Predictions**

**1.Average Processing Time per Chute per Hour** *(Regression Task)*

1. Helps us understand whether a chute is operating normally or slowing down due to volume buildup.
2. Used to detect early signs of potential performance degradation.

**2.Performance Degradation Risk** *(Classification Task)*

1. Binary output: Will there be a performance issue in the next time window?
2. Enables real-time **alerts and preventive actions** before delays occur.

**Current Prediction Scope**

•Predictions are made at the level of **ZIP code + chute + hour + station**.

•Inputs include:

- Historical package counts per ZIP/chute/hour.
- Time-based features (hour of day, day of week).
- Scanning station (to capture local load patterns).

•Output is both **continuous (processing time)** and **binary (issue/no issue)**.

**Business Value**

•Enables **proactive load balancing** by flagging upcoming pressure points.

•Supports **ZIP re-routing strategies** to reduce overload risk.

•Enhances operational visibility, helping floor managers **prioritize interventions**.