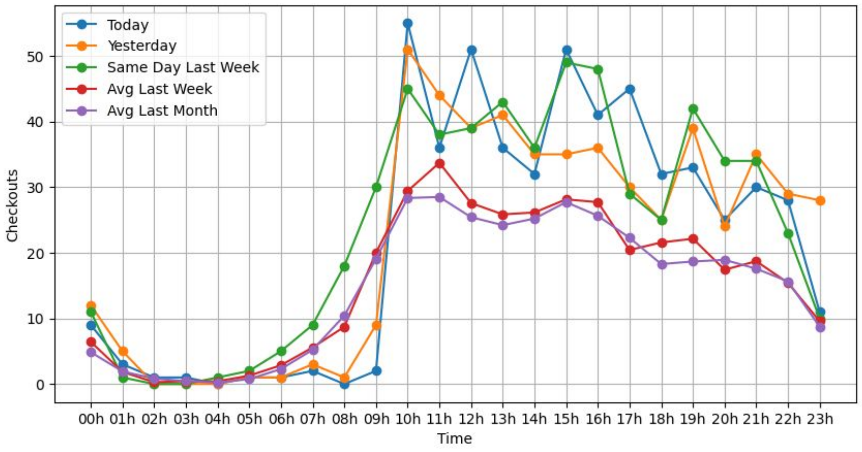
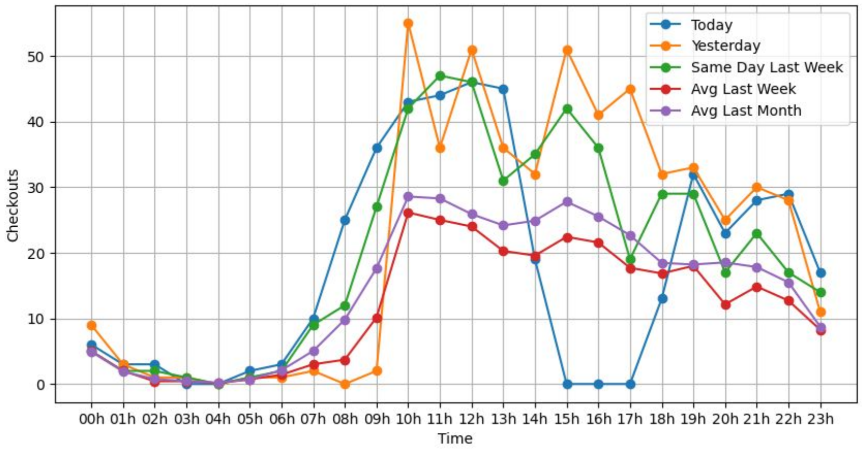
Anomalies on POS sales

Know the data

This graph enables a comparison of checkout patterns across various time frames, offering insights into customer behavior analysis. It clearly illustrates significant seasonal variations in the data. Notably, early morning hours show minimal checkout activity.

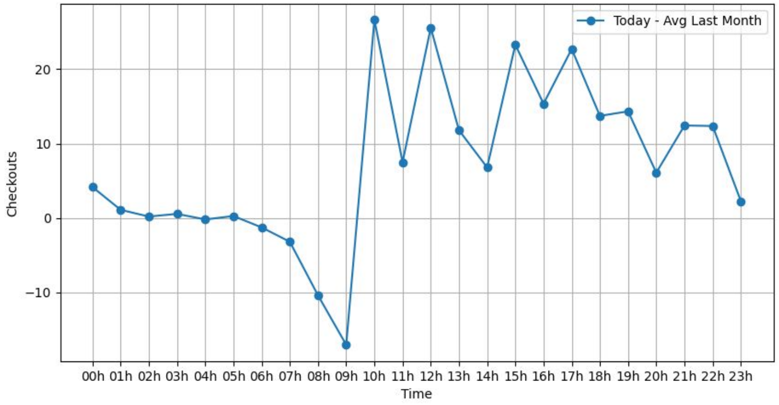


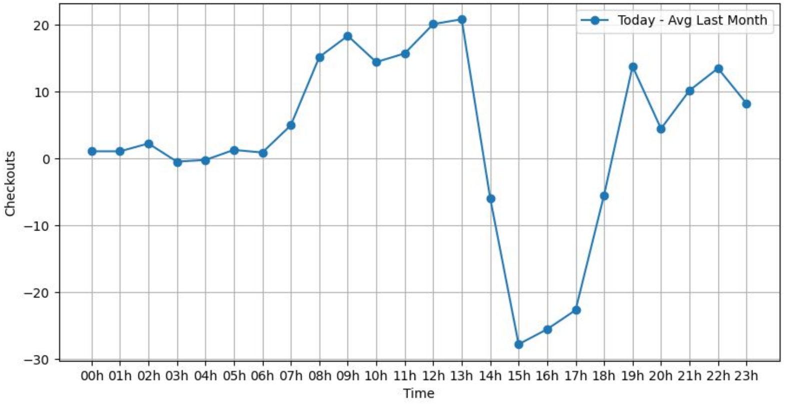
Day 1



Day 2

Detecting Anomalies in Sales Data

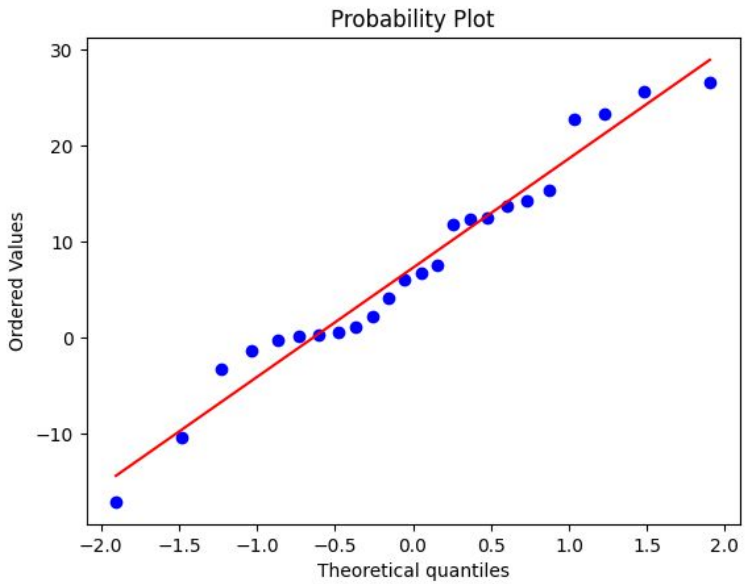




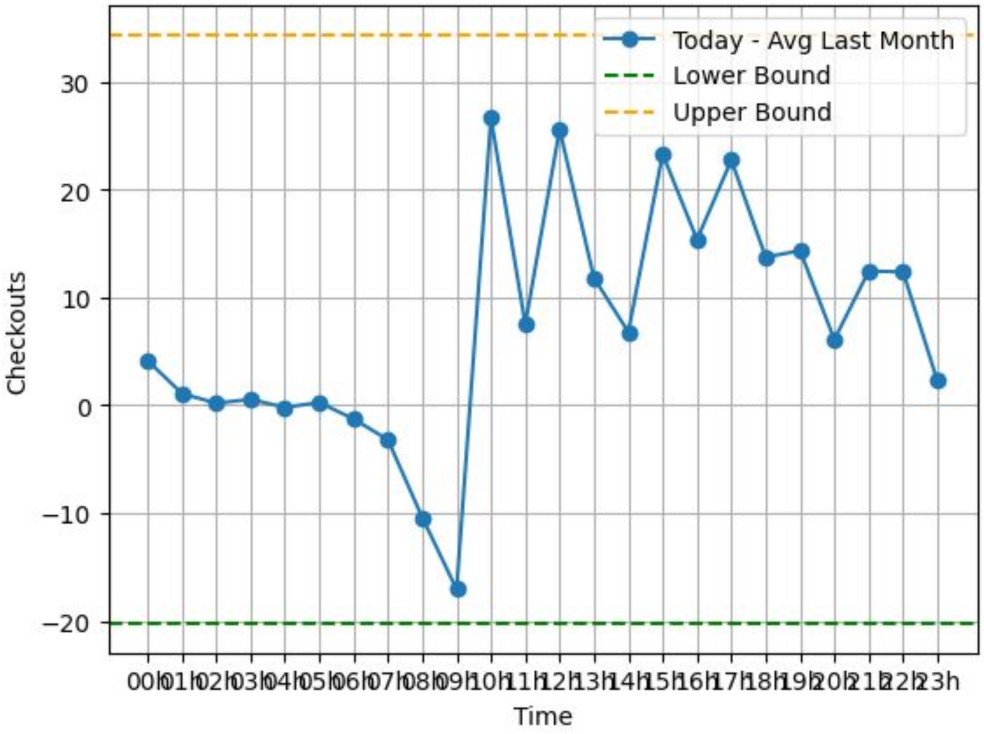
Given the **seasonal** nature of sales during different times of the day, we will compare sales values with their average **to establish a baseline for anomaly detection**. Since we have a small amount of data, we'll use a **statistical method** rather than a machine learning model. Choosing the residual between today's value and the monthly average, as it has the highest number of samples.

Q-Q Plot for Normality Check

In statistical analysis, a Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular distribution, such as a normal distribution. In this section, we'll create a Q-Q plot to check the normality of our sales data.



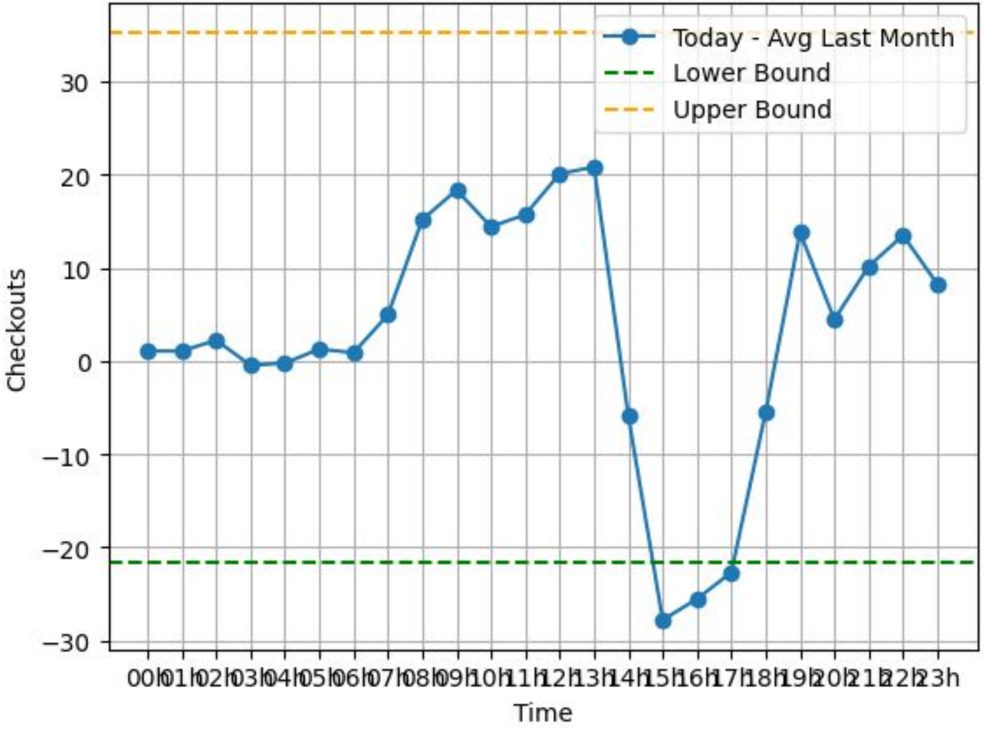
Non-Normality and Outlier Detection



Day 1

In our analysis, the data does not approximate a normal distribution. Therefore, we will not use the z-score method for outlier detection.

Non-Normality and Outlier Detection



Day 2

In our analysis, the data does not approximate a normal distribution. Therefore, we will not use the z-score method for outlier detection.

Conclusion

Based on our analysis, we have identified anomalies in the sales data. Specifically, only the values at 3 PM, 4 PM, and 5 PM on the second day are considered anomalies.

- **Second Day, 15:00:** 3 PM Sales value is anomalous.
- **Second Day, 16:00:** 4 PM Sales value is anomalous.
- **Second Day, 17:00:** 5 PM Sales value is anomalous.

Implementing a Monitoring System

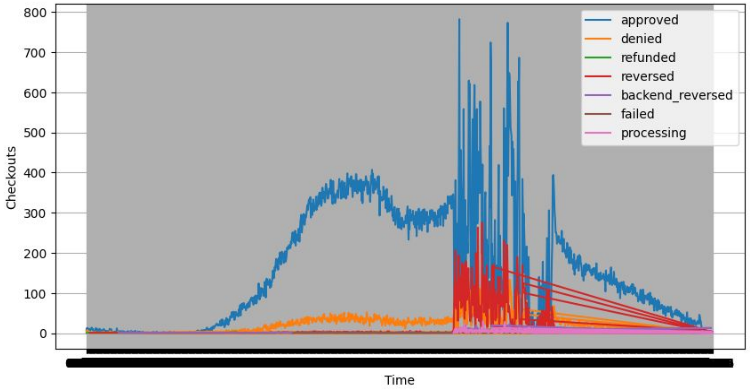
Know the data

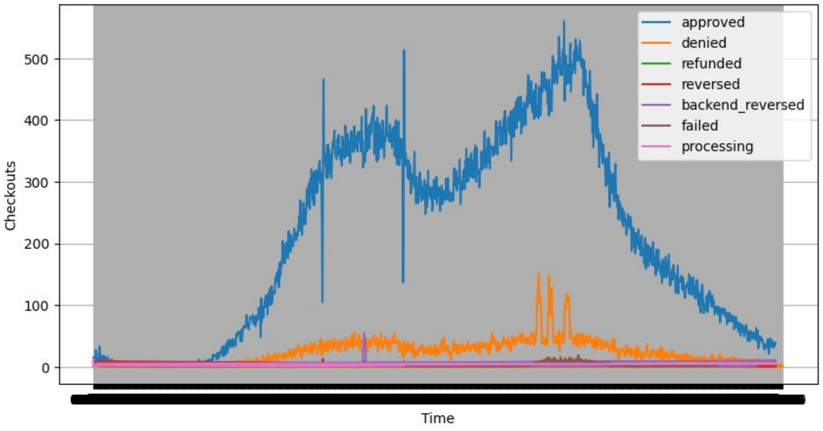
We can see the columns, the types, and that all values are non-null.

#	Column	Non-null Count	Dtype
0	time	8178	object
1	status	8178	object
2	count	8178	int64
3	day	8178	int64

Know the data

From this graphic we can start thinking about the anomalies detection.



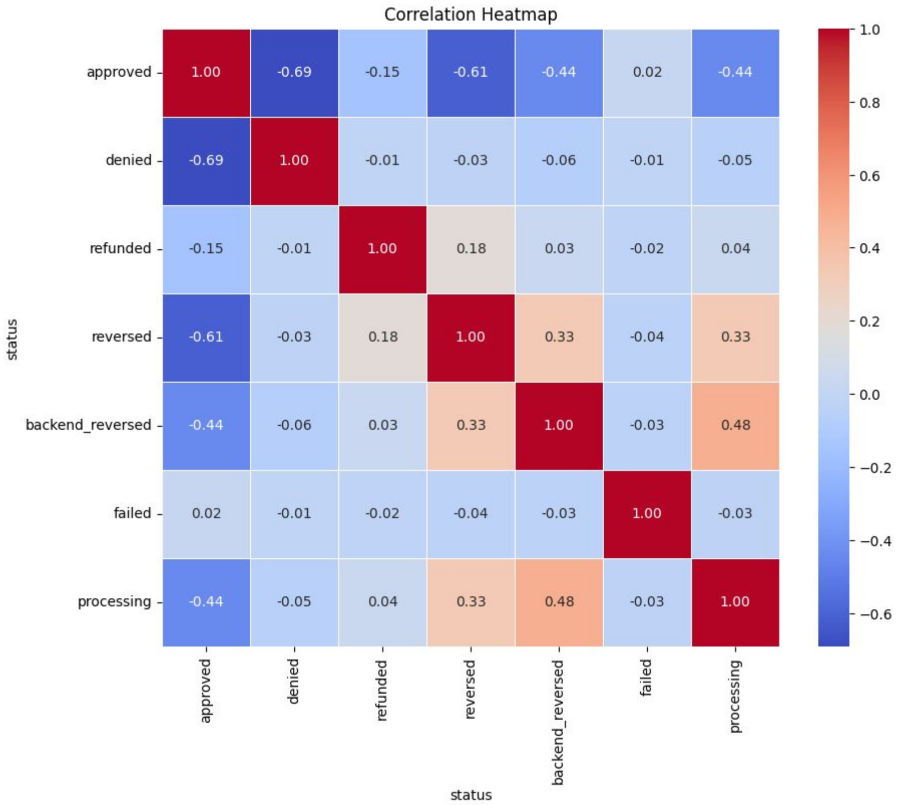


Data manipulation

To establish a baseline to compare the data, we are going to set the that to percentages of the status in that minute. For example, if there is 9 approved, 6 denied and 1 refunded the values are gone be:

approved	backend reversed	denied	failed	processing	refunded	reversed
56.25%	0.00%	37.50%	0.00%	0.00%	6.25%	0.00%

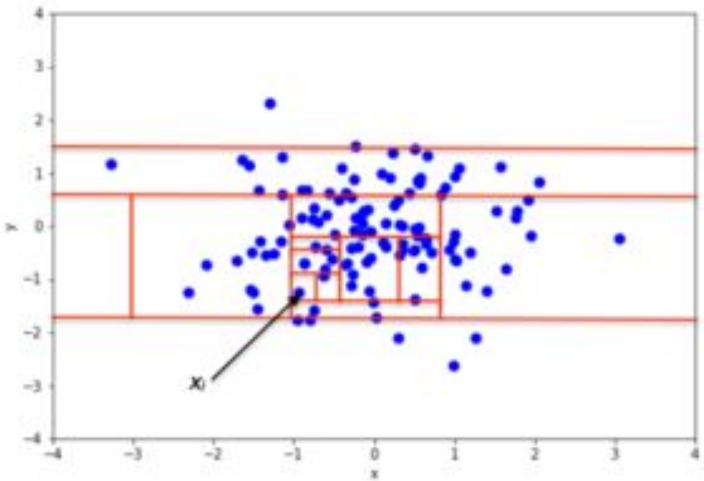
Correlation



Correlation analysis is important for selecting columns in machine learning because it helps identify relationships between variables, ensuring that redundant or highly correlated features are avoided to improve model interpretability and performance.

Isolation Forest

Isolation Forest, is an anomaly detection algorithm that isolates outliers by constructing random decision trees and measuring how quickly data points are isolated.



Server

I created a Flask server that will receive data for a specific minute in JSON format as follows:


```
[
  {
    "time": "18h 42",
    "status": "refunded",
    "count": 5
  },
  {
    "time": "18h 42",
    "status": "reversed",
    "count": 13
  }
]
```


and will return -1 if it's an anomaly and 1 if it's not. Based on the IsoForest model prediction.