Process Mining in the Coatings and Paints Industry: The Purchase Order Handling Process Business Process Intelligence Challenge 2019

Eda Eren

[eda.eren@studenti.unimi.it](mailto:eda.eren@studenti.unimi.it)

V09990

Business Information Systems

Prof. Paolo Ceravolo

2022- 2023

# **Abstract.** Process Mining is an extensively used field in process management to extract insights from event logs nowadays. The Business Process Intelligence Challenge of 2019 confers a problem and provides a real-life event log from a large multinational company operating in the area of coatings and paints from the Netherlands. This challenge focuses on the compliance analysis of the Purchase-to-pay process within this company. The process owner aims to gain business insights through addressing a variety of business questions. I tried to analyze the data using a variety of process mining and analytical tools.

# This report summarizes my understanding of the event log, my analytical approach, the different techniques and the steps undertaken to successfully answer business questions.

# **DESCRIPTION OF THE CASE STUDY**

# 

# This case study focuses on the gaining business insights through addressing a variety of business questions by focusing on compliance analysis of the purchase order handling process for a large multinational company operating in the area of coatings and paints in Netherland. The company aims to investigate the purchase order handling process of its 60 subsidiaries.

# **Application Areas**

# The application areas are the organizations and the businesses that manage purchase orders and engage in procurement. The case study gives us information about the difficulties and adversities in managing purchase orders, invoices and good receipts. The results and analysis are suitable for improving the financial control, uncover compliance issues and increase process efficiency.

# **Actors**

# *1-Batch Users*

The batch users are automated processes executed by different systems. While handling of the purchasing the orders, they have some certain tasks. Due to this their effects and actions can be examined to gain knowledge of the process’s automated elements.

# *2-Normal Users*

# The normal users refer to human actors in the process. They participate in the handling of purchase orders. [Task that are connected to invoices, good receipts and purchase orders.] The normal users responsibilities and activities are identified by the case study through the data. This situation allows us to examine their effects on the process.

# **Dataset And Peculiarities**

# In the data, each purchase order contains one more line items. For each line item, there are roughly four types of flows. These are:

# 3-way matching, invoice after goods receipt

# 3- way matching, invoice before goods receipt

# 2- way matching, no goods receipt needed

# Consignment.

There can be many goods receipt messages and corresponding invoices which are subsequently paid for each purchase item. So the complexity and adversity of the data goes further than just this division in four categories. For logistical services, there may even be hundreds of goods receipt messages for one line item.

Company, vendor, system, document names and IDs are anonymized in a consistent way throughout the log. The company has the key so any result can be translated by them to business insights about real customers and real purchase documents.

The event log is fully IEEE-XES compliant and is structured as follows. The case ID is a combination of the purchase document and the purchase item. There is a total of 76,349 purchase documents containing in total 251,734 items, so there are 251,734 cases. There are 1,595,923 events relating to 42 activities performed by 627 users (human and batch users).

For each purchase item or case the following attributes are recorded:

-Concept: Name: A combination of the purchase document ID and the item ID.

-Purchasing Document: The purchasing document ID

-Item: The item ID

-Item Type: The type of the item

-GR-Based Inv. Verif.: Flag indicating if GR-based invoicing is required

-Goods Receipt: Flag indicating if 3-way matching is required

-Source: The source system of this item

-Doc. Category Name: The name of the category from where the purchase originated

-Spend Classification Text: A text explaining the class of purchase item

-Spend Area Text: A text explaining the area for the purchase item

-Sub Spend Area Text: Another text explaining the area for the purchase item

-Vendor: The vendor to which the purchase document was sent

-Name: The name of the vendor

-Document Type: The document type

-Item Category: The category of the item based on the four flow types. (3-way with GR-based invoicing, 3-way without, 2-way, consignment)

**ORGANIZATIONAL GOALS**

# The main goal of the organizations and the companies is to improve their handling of purchase orders, improve their financial control, uncover compliance issues and increase process efficiency.

# *These might be the possible strategic goals that the organization is pursuing:*

# -Compliance is very important for a company. The company or the organization should be sure that the handling of purchase orders complies with the predetermined purchase rules, invoices and receipts. For example every line item should conform to its flow type.

# -The company has to be sure that if the receipts and invoices are matching or not.

# So they may want to increase the accuracy of the matching of that.

# -Company’s most wanted things are always the efficiency. So, they want to do the job fast with the best performance possible.

# Operational and tactical performance measures are components of strategic performance measures and they provide important insights to revise the strategic plans.

# *So, these might be the possible operational objectives of the organization:*

# -To increase compliance, the company can check the compliance at several points of the purchase order handling process.

# -To increase accuracy, companies can make a system, a software that automate the receipt-invoice matching control. Automated processes can accurately match receipts with corresponding invoices, reducing manual effort and minimizing errors.

# -To increase efficiency, companies can make a real time tracking system for receipt and invoice status provides visibility into the process. It allows to monitor the progress, identify bottlenecks and take prompt actions to resolve any discrepancies. Or the companies can develop an exception management system that helps identify and address discrepancies or exceptions that arise during the receipt-invoice matching process. This allows for timely resolution of issues, preventing delays and improving overall efficiency.

# *And lastly these might be the examples for the tactical objectives of the company:*

# -Reduce Processing Time by optimizing workflows, leveraging automation tools and eliminating bottlenecks.

# -Improve Accuracy by implementing validation checks, ensuring data integrity and providing training to the employees involved in the process.

# -Enhance Reporting and Analysis by including comprehensive dashboards, utilizing data analytics to gain insights for process optimization.

# Raw data map (directly follows graph visualization) by Disco

# **KNOWLEDGE UPLIFT TRAIL (KUT)**

# The goal is to examine the purchase order handling process for some of its 60 subsidiaries and address compliance questions raised by the process owner (organizations and the companies). To make sure that the purchase order handling process is compliant to the regulations, industry standards, financial operations and data privacy.. Procedures should outline the steps, roles and responsibilities involved in the process, ensuring adherence to compliance standards. After examining these situations and finding out the potential problems we can improve the system.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Input | Acquired Knowledge | | Output |
|  |  | Analytics/Models | Type |  |
| Step 1 | Purchase order records (event logs) | Filter the data | Prescriptive | 2019 data |
| Step 2 | Step 1 | Filter cases with zero duration | Prescriptive | Filtered Data |
| Step 3 | Step 2 | Filter cases to summarize | Prescriptive | Filtered Data |
| Step 4 | Step 3 | Filter cases according to item types | Prescriptive | Filtered Data |
| Step 5 | Step 4 | Filter cases with incomplete ending | Prescriptive | Filtered Data |
| Step 6 | Step 5 | Variant Analysis | Descriptive | Statistics about variants |
| Step 7 | Step 5 | Create heatmaps | Descriptive | Activity frequencies |
| Step 8 | Step 5 | Create directly flow graphs | Descriptive | Activity Flows |
| Step 9 | Step 5 | Petri nets and process trees (with different miners) | Prescriptive | Petri net/process tree |
| Step 10 | Step 5, Step 9 | Conformance checking | Descriptive | Compliant case statistics |
| Step 11 | Step 5, Step 10 | Comparative analysis | Prescriptive | If difference is significant |

# **Assumptions made while designing the trail**

# - The data is structured and it has accessible format with its mentioned attributes.

# - Even if we filter the data in the next steps, it will still going to has mostly relevant and reliable information which is sufficient enough to do process mining.

# - Necessary software tools to perform process mining are available.

# **Research questions that answered by the project**

# 

# Is the purchase handling process compliant with the company goals?

# Are there any discrepancies or non-compliance issues observed in the value matching and invoicing procedures?

# Does the process follow the appropriate value matching procedures for each type of purchase order?

# Are there any gaps in documentation, controls or communication that may contribute to compliance issues or inefficiencies?

# **PROJECT RESULTS**

# **VARIANT ANALYSIS**

# 

# In this case study there are 11973 variants, 1,595,923 events, 252,734 cases and 42 activities. Variants are the sets of similar cases, so they are important for understanding the diversity and differences between cases. By this analysis, it helps us to increase the efficiency, provide compliance and optimize or standardize systems for organizations or companies.

# Visualization of the case variants by Disco:

# 

# 

# 

# In the frequency distribution of variants within the purchase order handling process, a notable pattern emerges resembling the Pareto principle, where a small portion of variants represents the majority of cases. Unlike this situation, a significant number of variants which are not part of the common subset, contribute to the majority of cases and introduce diversity. These less common variants often play a crucial role in causing compliance issues and warrant closer examination.

# 

# Project Results

# 1% of the variants corresponds to 87% of the cases.

# When we analyze these frequencies with the variation filter, we can see that the most frequent variant follows this trace of activities (item category: 3-way matching, invoice after goods receipt (GR))

# 

# **FILTERING**

# 

# To conduct a focused analysis, it is necessary to filter the event logs and exclude irrelevant or false information. This process aims to remove unnecessary components from the event log, resulting in a more brief and sufficient data frame for analysis. By eliminating noise and irrelevant details, we can gather a more targeted and meaningful dataset for further examination.

# Some cases have zero duration. This type of cases might be considered as a noise. None of the cases can be completed in zero duration. In this case, we should filter this type of cases from the data. We can use Python for filtering the data:

# As you can see after the application of the filtering of the noise,

# 

# The related code is in the Google Colab File in “#5”th code.

# I used the Performance Filter and set the minimum duration time to 1 millisecond in disco and came through these statistics:

# As you can see, the noises (cases with zero durations) comprise 2% of the cases. The rest of them are the initial cases as filtered event log.

# Summarizing the data

# Frequency distribution of variants often conforms to a power-law distribution. This implies that a small proportion of variant allows us to capture the majority of cases. To optimize our analysis, I specifically concentrated on the top 40 variants for 3-way matching categories and consignment, while for 2-way matching, I expanded the scope to include 150 variants due to its relatively lower occurrence compared to the other categories within the event log. By selecting these specific variants, we can gather a more focused and representative subset of cases for further analysis.

# Through the process of filtering variants, we can narrow our focus to a more specific subset while retaining the majority of cases (approximately 80% of the total). This enables us to concentrate on a more targeted area within the variants without sacrificing the representation of the majority of cases.

# With python code we can see the number of events and number of cases and it’s graph as in the picture. You can access the related code in Google Colabs in code number #6

# 

1. In order to address the compliance questions pertaining to the item category of purchase orders, the data will be segmented into four distinct groups. The primary objective of this division is to enhance the efficiency of the process discovery and compliance checking stages. The four groups include 3-way matching {1}, 3-way matching {2}, 2- way matching, and consignment categories. By analyzing the data in these separate groups, we can effectively examine and evaluate the compliance- related aspects specific to each item category, enabling more streamlined and targeted insights into the compliance landscape of the company’s purchase order handling process. The details of this grouping procedure can be found in Google Colab in number #7



# Following the division of the data frame into four groups, a filtering process was implemented to eliminate incomplete cases. Specifically cases ending with “Delete Purchase Order Item” were identified and filtered out. These incomplete cases do not represent a complete purchase order flow and thus were deemed irrelevant for the process discovery. The filtering steps, corresponding to code in Google Colab in code number #8, #9, #10 and #11, effectively removed these incomplete cases from each respective group, ensuring that the subsequent analysis focuses on complete and representative purchase order flows.

# Incomplete cases 3 way-matching, invoice after goods receipts (GR) #8

# 

# 3-way matching, invoice before good receipts (GR) #9

# 

# 2-way matching #10

# 

# 

# Consignment #11

# 

# **PROCESS DISCOVERY**

# *Process discovery for the 3-way matching, invoice after good receipt*

# The Heatmap:

# 

# 

# According to the heatmap, “Create Purchase Order Item” and “Record Goods Receipts” are necessary activities. And the other activities may be a case or not.

# The Directly Follows Graph:

# 

# I compared the different process model visualizations with using the knowledge that we gained and the most fitting one seems to be that the Petri Net using ILP (Integer Linear Programming).

# 

# 

# All other process model visualization comparations are on Google Colab codes in the number #12.

# If we analyze them in terms of quality metrics,

# -Fitness (recall) is high.

# -Precision is medium.

# -Generalization is medium.

# -Simplicity is medium.

# For abstract (Fitness refers to how well the data fits the problem at the hand.

# Precision measures the accuracy and correctness of the data. Generalization assesses the ability of the data to be applied or generalized to different scenarios or contexts. Simplicity refers to the clarity and understandability of the data)

# *Process discovery for 3-way matching, invoice before good receipt*

# Heatmap:

# 

# According to the heatmap, “Create Purchase Order Item” is the necessary activity. And the other activities may be a case or not.

# Directly Follow Graph

# 

# I compared the different process model visualizations with using the knowledge that we gained and the most fitting one seems to be that BPMN net using inductive miner.

# 

You can find the other visualization techniques in Google Colab in code number #13

If we analyze this in terms of quality metrics, we might say that

-Fitness (recall) is high.

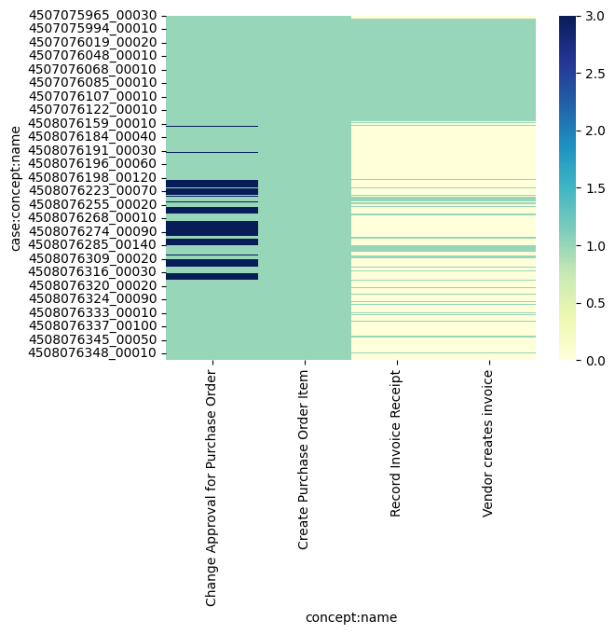
-Precision is low.

-Generalization is medium.

-Simplicity is medium.

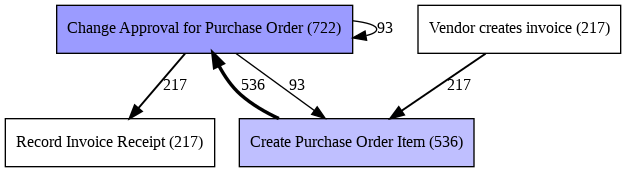
1. *Process discovery for 2-way matching*

The Heatmap

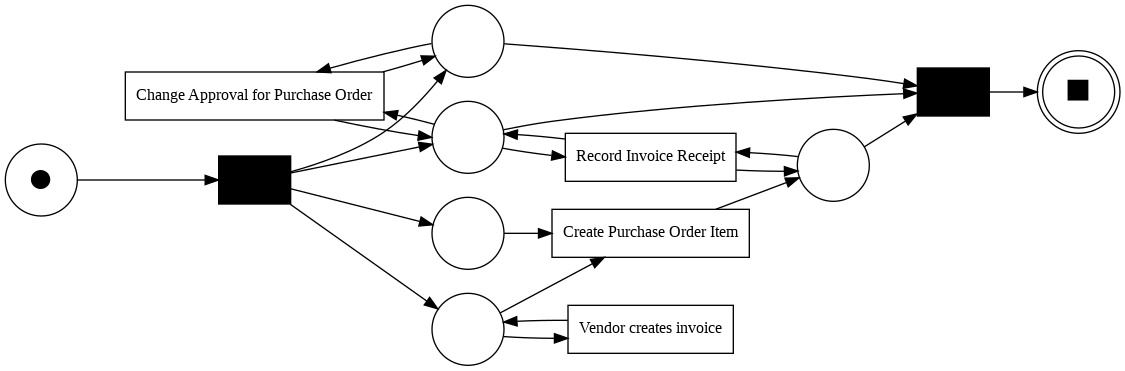


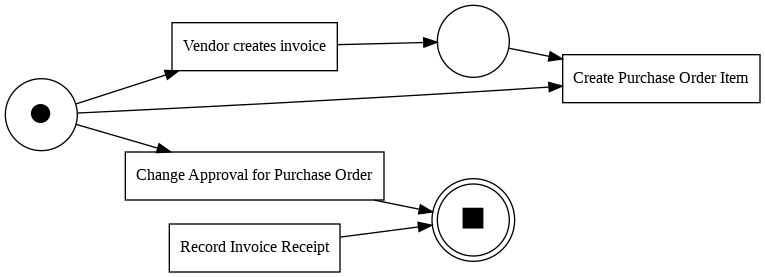
As we can from the heatmap there aren’t any necessary activities. But the “Change Approval For Purchase Order” is the most used one.

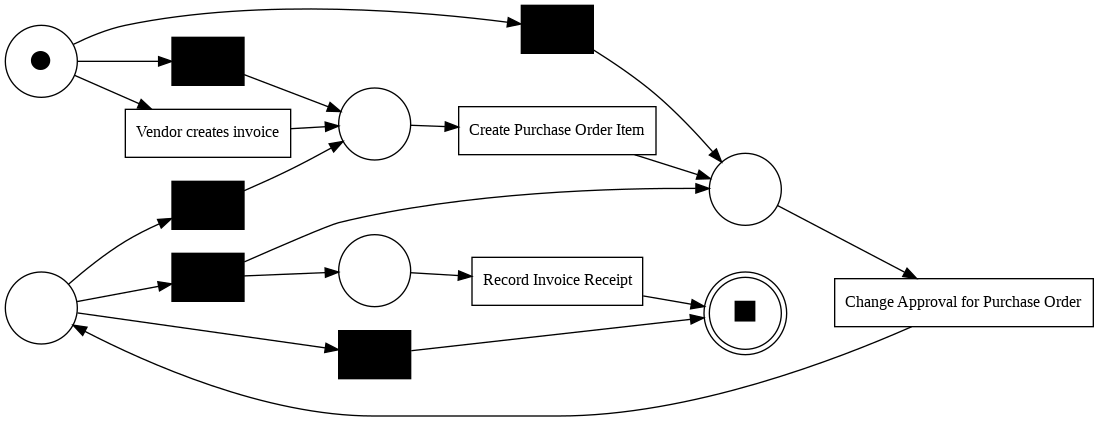
The Directly Follows Graph Of The Process

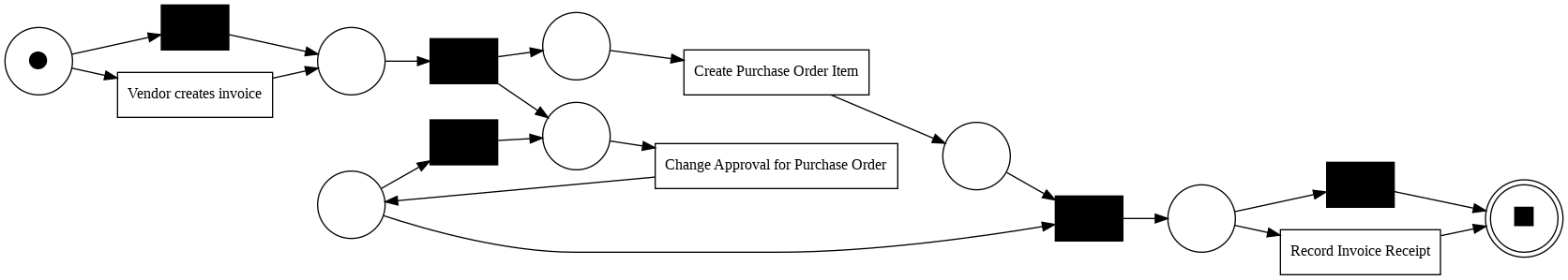


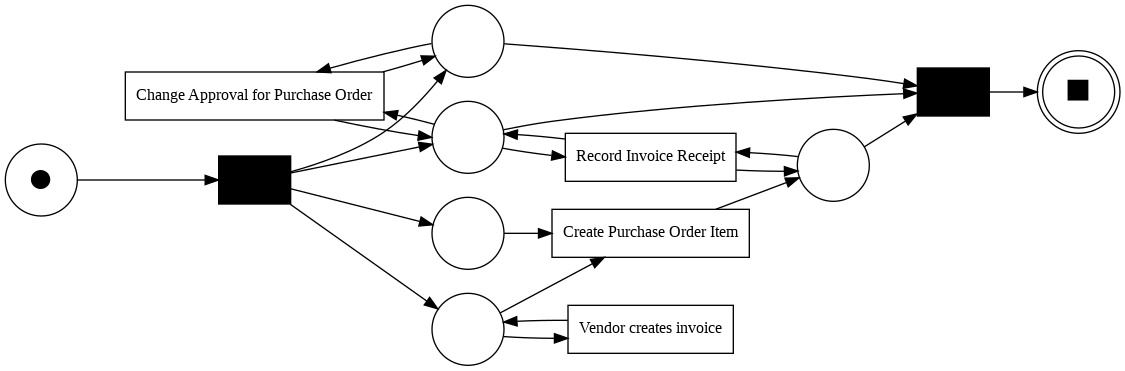
Petri Net Using ILP Miner



Alpha Miner Alghorithm

Petri Net Using Heuristic Miner

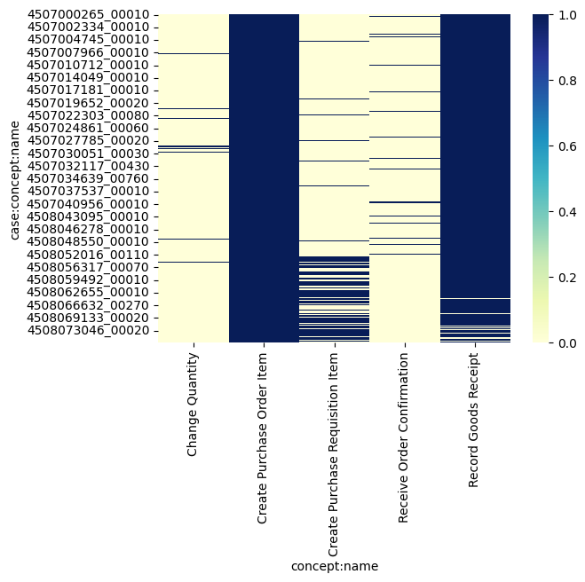
Petri Net Using Inductive Miner

Process Tree Using ILP Miner

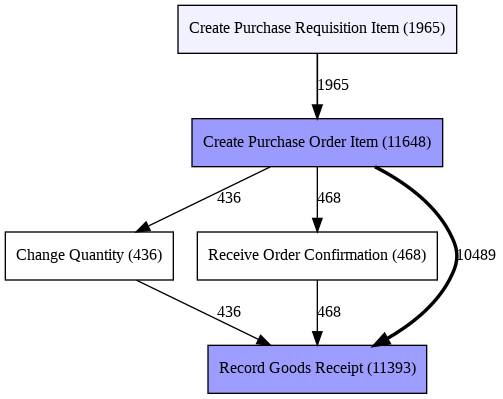
You can find more of the process discovery visualizations in the Google Colab in the code number #15.

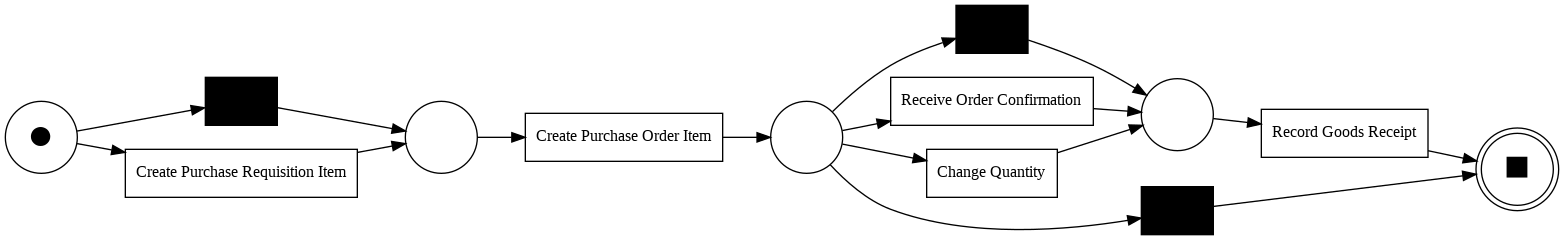
*4-Process Discovery For Consignment*

The HeatMap:



The Directly Follows Graph:



The Petri Net Using ILP Miner:

You can see more of the process discovery visualizations in Google Colab in the code number #16.

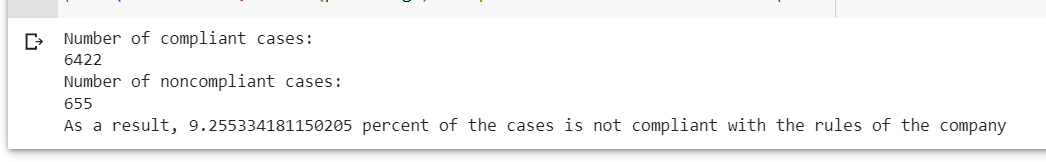
**CONFORMANCE CHECKING**

The company’s compliance checking for purchase order cases focuses on ensuring value matching according to item categories and their specified rules. Here are the expected procedures for value matching based on different item categories:

* *3-way (1):* The value of the good receipt message should be matched against both the value of the invoice receipt message and the value put during item creation.
* *3-way (2):* The value matches with the invoice and the value at creation of the item.
* *2-way:* The value of the invoice should match the value at creation.
* *Consignment:* There are no specific value matching criteria. Additionally, each activity within consignment cases is assigned a value of zero, eliminating the need for value checking.

These different value matching procedures enable the company to ensure compliance with the established rules and standards for purchase order cases across various item categories.

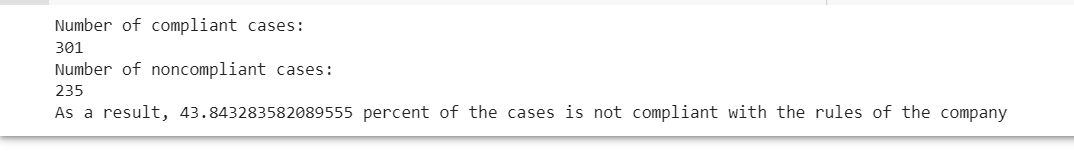
*Results of conformance checking for 3-way matching, invoice after GR, observing whether the values are equal or not:*



The code is in Google Colab, in code number #17.

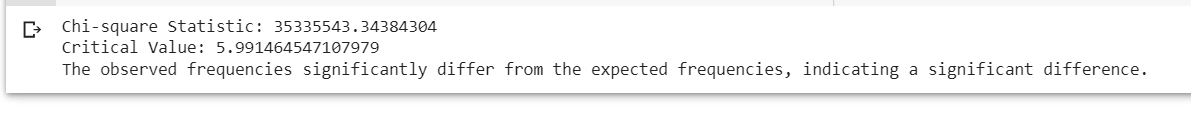
*Results of conformance checking for 3-way matching, invoice before GR, observing whether eur values are equal or not:*

You can find the related code in Google Colab, in code number #18.

*Results of conformance checking for 2-way matching, observing whether eur values are equal or not:*

You can find the related code in Google Colab, in code number #19.

**COMPARATIVE PROCESS MINING**

In order to assess the significance of the difference between the expected and observed frequencies of non-compliant cases, we need a test. This test helps evaluate whether it’s sticking to the compliance rules or if there are significant deviations from the expected frequencies.

You can find the related code in Google Colab, in code number #20.

**CONCLUSION**

In the light of the knowledge gained from the event log, the company has thoroughly investigated and addressed compliance concerns. The analysis reveals that the majority of purchase orders adhere to the company’s rules, with the exception of filtered cases (such as those with zero durations or deleted cases). Regarding value matching, the non-compliance rate for 3-way matching (1) and (2) is found to be unsignificant. However, the 2-way matching category exhibits a significant non-compliance ratio of 43.84% in terms of value matching. The comparative process mining analysis proves that non-compliant instances have an important impact on the overall process. Therefore, it is crucial for the company to prioritize the development of its process model, particularly for the 2-way matching item category, in order to optimize the system and decrease potential non-compliant cases effectively.

*Link for the code:*

<https://colab.research.google.com/drive/1osSeBdoZ1NVqjA2SvrHtRFEDS5sPma9Q?usp=sharing>

*File:*

**

<bis_project.py>

<BIS_Project.ipynb>

*Used dataset reference:*

van Dongen, B.F., Dataset BPI Challenge 2019. 4TU.Centre for Research Data.

https://doi.org/10.4121/uuid:d06aff4b-79f0-45e6-8ec8-e19730c248f1

*Links for the references that I took for the project:*

https://github.com/paoloceravolo/BIS2022/blob/main/Case3/case3clustering.py

https://icpmconference.org/2019/wp-content/uploads/sites/6/2019/07/BPI-Challenge-Student-Submission-6.pdf

<https://github.com/NicoRota-0/BIS-project>

<https://icpmconference.org/2019/wp-content/uploads/sites/6/2019/07/BPI-Challenge-Student-Submission-6.pdf>

<https://github.com/a-baran-orhan/Business_Information_Systems/blob/main/BIS_project.ipynb>

https://icpmconference.org/2019/icpm-2019/contests-challenges/bpi-challenge-2019/

*Links for the sites that helped me:*

<https://pm4py.fit.fraunhofer.de/>

https://pm4py-source.readthedocs.io/en/latest/pm4py.algo.filtering.log.variants.html