UNICORN VYSOKÁ ŠKOLA S.R.O.

MASTER'S THESIS

2024 Bc. Filip DITRICH

UNICORN VYSOKÁ ŠKOLA s.r.o.

Softwarový vývoj

MASTER'S THESIS

Cashless festival data analysis and analytical dashboard development

Author: Bc. Filip Ditrich

Supervisor: Mgr. Václav Alt

Prague 2024

Statutory Declaration

I hereby declare that I have written my Master's Thesis on the topic of *Cashless festival data analysis and analytical dashboard development* by myself, under the guidance of my thesis supervisor, using only the technical publications and other information sources which are all quoted in the thesis and listed in the bibliography.
I declare that artificial intelligence tools have been used only for support activities and in accordance with the principle of academic ethics.

As the author of this Master's Thesis, I also declare that in association with its writing I have not violated the copyright of any third party or parties and I am fully aware of the consequences of provisions of s. 11 et seq. of Act No. 121/2000 Coll., the Copyright Act.

Furthermore, I hereby declare that the submitted hard copy of this Master's Thesis is identical to the electronic version I have submitted.

In ..... on ....

Bc. Filip Ditrich

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Mgr. Václav Alt, for his guidance, support, and valuable advice throughout the process of writing this thesis. I would also like to thank my family and friends for their encouragement and understanding.

Cashless festival data analysis and analytical dashboard development

Analýza dat bezhotovostního festivalu a vývoj analytického dashboardu

UNICORN UNIVERSITY

Abstract

TODO

Keywords: TODO

Abstrakt

TODO

Klíčová slova: TODO

Contents

Introduction

Background and Motivation

Payments at festivals are a crucial part of the successful event management. The shift from cash to cashless payments has been a significant trend in the last decade that has brought many benefits to both festival organizers and attendees.

However, traditional cashless payment systems utilizing payment terminals are not only expensive to reliably implement at a venue, where the internet connection is often unreliable and requires expensive Base Transceiver Stations (BTS) set-ups or in-place wired/optical internet implemented which can cost up to several million CZK, but also do not provide any insights into the data generated by the transactions. In the best case scenario, the organizers are able to generate a report of processed transactions made by each terminal.

Which, frankly, is not enough to make any actionable decisions based on the data. Moreover, given the event organizers are not data scientists, they often lack the knowledge and tools to analyze the data and extract valuable insights from it.

That is where NFCtron comes in and offers a solution that not only provides a reliable cashless payment system, with credit-based NFC chip bracelets supporting offline mode or card terminal payment solutions, but also provides a comprehensive B2B platform that allows the organizers, the vendors and event the third-party partners to benefit from the data that the system operates with.

The system is a full-scope solution that provides from initial online ticket sales and online credit top-up, through attendee check-in, on-site credit top-up attendee access control, security monitoring, vendor sales and inventory management, most importantly the fast and reliable payment processing with the real-time data analytics and reporting, all the way to the post-event automatic settlement and

reporting.

It simply provides everything an event organizer needs to successfully and efficiently manage their event without the need to worry about any technicalities or staff management. Since NFCtron does not only provide the system as a service, but also provides experienced Event Managers, cashiers, check-in brigadiers and other staff to operate the system and the event itself ensuring that the organizer can focus on the event in terms of communication, marketing, line-ups and other important aspects of the event.

Put, NFCtron offers organizers a peace of mind and a guarantee that their event will be a success.

NFCtron Company

NFCtron is a Czech company that has been operating since around 2019. In its early beginning during a COVID-19 pandemic was on the verge of survival because of the event industry being paralyzed by the government restrictions. However, the company survived and even in these difficult times managed to turn the disadvantage into an advantage by focusing on the core system and product development.

Which years later resulted in a robust and reliable system. It is now used by many event organizers across the Czech Republic and Slovakia and is currently expanding to other countries in Central Europe such as Austria, Poland and Germany. In its primary market – the Czech Republic – NFCtron penetrated the market and is now the leading cashless payment system provider for festivals and other events.

In recent years, the company has also been focusing on expanding to the Payments market focusing both on Card Acquiring and Card Issuing. From the acquiring part, it is now actively developing its own SoftPOS solution that will allow vendors to accept payments via mobile phones. On the other hand, on the Issuing part, it is also working on Card Issuing; that will allow the company to issue its own NFCtron branded payment cards in cooperation with Mastercard.

A big part of the successful market penetration was the company's focus on the business-to-business (B2B) side of the business with event organizers. Giving the organizers amounts of data improving their decision-making and providing them

with insights allowing them to optimize their events. And most importantly, providing them with economic aspects and cashflow optimizations that allows many events and festivals to survive and continue to operate.

Personal Position and Motivation

I have been with the company from the COVID-19 times. My current position is a Chief Product Officer (CPO), and I am responsible for the product development and the product management of all the products and services that NFCtron offers. This allows me to have a deep insight into the system and most importantly, access to all the data that the system generates. As previously mentioned, the company's success is based on the B2B side of the business and that means services and products provided to the event organizers.

The main product, that organizers have access to is the platform called **NFCtron Hub**. My personal goal and personal motivation to work on this thesis is to discover new ways to improve the platform and provide even more valuable insights to the organizers.

Problem Statement

Even though the system provides a lot of data, it still has a lot of potential to provide even more valuable insights to the organizers. Currently, the previously mentioned B2B platform, **NFCtron Hub**, provides real time data analytics dashboard presenting the most important KPIs and metrics to the organizers.

These KPIs and metrics summarize:

- **Total sales**: the total amount spent on online ticket sales and on-site payments.
- **Total sales in time**: the total amount above split into time intervals.
- **Total refunds**: the total number of sale reversals or refunds made including refunds from online tickets, refunds of on-site payments and chip credit refunds.

- **Chip balances**: the current balance topped-up on the NFC chip bracelets on-site or pre-topped-up online.
- **Customer orders rating**: customers can rate their orders via a NFC-tron mobile application which provides the organizers with feedback on the vendor's performance and the quality of the products sold.

Moreover, it provides less clear data such as

- **List of vendors**: the list of vendors presents at the event with their sales and rating.
- **List of products**: the list of products sold at the event with their sales and rating.
- **List of points of sale**: the list of selling points at the event with their sales and rating.
- **List of top-up places**: the list of top-up places at the event with the amount of top-ups made.
- **List of customer chips**: the list of unique individual NFC chips issued to the customers with their balance, spending and security status.
- **List of customer ratings**: the list of customer ratings with their feedback from points of sale.

And finally, it also provides unstructured data in the form of data exports in tabular format that can be used for further analysis:

- **Product exports** – list of all products sold with summarized metrics.
- **Points of sale exports** – list of all selling points with summarized metrics.
- **Vendor exports** – list of all vendors with summarized metrics.
- **Deal exports** – list of summarized sales made under a deal[1] Between the organizer and the vendor.
- **Transaction exports** – a heavy export of all transactions made at the event.

[1] A deal is an arrangment between the organizer and the vendor that states the terms of the vendor's presence at the event, the products they are allowed to sell, the price of the products, the commission the vendor pays to the organizer and other terms of the deal.

- **Ticket redeems exports** – list of all tickets redeemed at the event.
- **And other exports regarding the online sales** – list of all online ticket sales, top-ups, receipts, customers and other data.

With the above giving some initial picture about the capabilities of the system, the platform (and thus the organizers) still face several problems or challenges that need to be addressed:

1. **Problem 1**: The main KPI metrics may provide some core insights, but the other less or unstructured data is not used to its potential.
2. **Problem 2**: The organizers are not data scientists and need a simple and clear way to understand the data.
3. **Problem 3**: Even with all this data, there is still a lot that can be done to dig deeper and provide more valuable insights.

Objectives of the Work

With the problems stated above, the main objective of this thesis is to analyze, answer and present results to important questions about the available data and the potential insights that can be extracted from it.

But to achieve this, it was a prerequisite to find a willing event organizer, that would provide the data and would be willing to cooperate on the project. Cooperate in terms of providing valuable insights into what they would like to know more about their event.

In Cooperation with the Event Organizer

For this purpose, I have chosen a not more undisclosed event organizer, that has been a close and helpful partner of NFCtron for many seasons. Together with the organizer in the first step, we have stated the following requirements to perform the analysis:

- **Requirement 1**: The event and organizer should be kept undisclosed.

- **Requirement 2**: The data should be anonymized to not leak any possible sensitive information about vendors or customers.

The next step was to choose an event from which the data will be used. As it cannot be disclosed any further, we will refer to the event as **The Event** and the organizer as **The Organizer**.

Now the important information about **The Event** for this study is the following:

- **The Event** is a music festival that has been organized for several years now.
- **The Event** takes place in the Czech Republic in the begging of July 2025.
- **The Event** is a 3-day event with multiple stages and multiple vendors.
- **The Event** uses NFCtron system for cashless payments and access control.
- **The Event** had around 7,000 attendees in 2024 and had a roughly 43% increase in 2025 to around 10,000 attendees.

Data Analysis Objectives

The final step was to define questions or data analysis objectives that should be answered or achieved by the end of the thesis.

Together in the cooperation with **The Organizer** and several internal colleagues in NFCtron, we have defined the following questions for the data analysis:

Cashflow and Revenue

- ***RQ1:*** *What was the total revenue of the organizer and what does it consist of?*
- ***RQ2:*** *How much and by what means was the balance topped up on the chips?*
- ***RQ3:*** *How much balance remained on all chips after the event and after refunds?*

- ***RQ4:*** *What was the total sales of the event, how much of it was the sales of the organizer and how many external vendors?*

These questions should provide currently unclear insights into the cashflow and revenue sources of the event. Possible answers to these questions could provide the organizer with valuable insights into the economic aspects of the event and could help optimize the cashflow and revenue sources for the next event.

Performance

- ***RQ5:*** *How many transactions were processed in total and what was the largest "peak" in the volume of transactions processed by the system (and when)?*
- ***RQ6:*** *What was the average transaction processing time during peak hours?*
- ***RQ7:*** *Were there any significant delays or downtimes in processing transactions?*
- ***RQ8:*** *What were the best-selling points?*
- ***RQ9:*** *What were the best top-up points?*
- ***RQ10:*** *Who were the best vendors?*
- ***RQ11:*** *What were the best products?*

The current platform already provides some performance metrics, but these questions should provide more detailed insights into the performance of the event.

Beverage Consumption

- ***RQ12:*** *How much was the total consumption of drinks/fluids?*
- ***RQ13:*** *How many returnable cups were issued, and how many were returned or not returned?*
- ***RQ14:*** *What was the most popular drink category?*

- ***RQ15:*** *What was the TOP beer brand, how much was consumed and sold?*

- ***RQ16:*** *What was the TOP brand of other alcoholic beverages, how much was consumed and sold?*
- ***RQ17:*** *What was the TOP brand of non-alcoholic beverages, how much was consumed and sold?*

Currently, a more in-depth product analysis is missing in the platform and the most important part of the product sales analysis at festivals is the beverage consumption. These questions should try to answer and give detailed insights into the beverage consumption, preferences and sales at the event.

Customers

- ***RQ18:*** *What was the total attendance at the event, how many active customers were at the event each day?*
- ***RQ29:*** *How many customers topped up credit in advance online?*
- ***RQ30:*** *What was the distribution of customers by their type (on-site, online, staff, guest, VIP)?*
- ***RQ31:*** *How many customers used the mobile application?*
- ***RQ19:*** *What is the distribution of bank cards used to refund credit?*
- ***RQ20:*** *What is the distribution of card schemes used to top up credit on-site and online?*
- ***RQ21:*** *What was the course of the event in terms of new visitors and when were the largest "peaks"?*
- ***RQ22:*** *What is the average time of a visitor from arrival in the first transaction?*
- ***RQ23:*** *What was the course of the event in terms of topping up credit on-site and when were the largest "peaks"?*
- ***RQ24:*** *How many customers topped up credit more than once and how much only once?*

- ***RQ25:*** *What was the difference between one-day, two-day and three-day visitors in terms of spending, topping up and refunded credit?*
- ***RQ26:*** *What was the difference between spending and topping up between different types of visitors (primarily online vs. on-site)?*
- ***RQ27:*** *What were the preferences for the type of drink throughout the day?*
- ***RQ28:*** *What were the most common combinations of products?*

The customer analysis is the most important part of the data analysis. It is crucial for the festival organizers to know their customer base and their behavior to optimize the event and make it more attractive for the customers.

Currently, no customer analysis, other than the customer ratings and list of customer chips, is available in the platform.

Answering these questions will possibly lead to the most valuable insights about the event's customer base and their behavior that no other platform or system currently provides.

Making these questions crucial and most valuable for the organizer and for the platform itself.

Technical Objectives

Answering the above questions will require a technical solution that will be able to process the data and provide the answers.

The scope of this study is not to implement a new system or a new platform and not even to implement any new changes to the existing NFCtron Hub platform. It is to find answers to the above questions and present them in a clear and understandable way in the form of a simple internal dashboard.

The technical goals of this study are:

- Prepare, process and analyze the data from **The Event**.
- Find answers to the above questions.

- Implement a simple internal dashboard that will present the answers to the questions.

Scope of the Study

To ensure the feasibility and focus of the study, certain boundaries have been defined in terms of what is included and excluded from the scope of the study.

Included in the Scope

- The study will focus on transactional, customer, and operational data from a specific event, referred to as **The Event**.
- Key areas of analysis include cashflow, revenue sources, performance indicators, beverage consumption, and customer segmentation and behavior.
- The data analyzed includes pre-event (e.g.,online top-ups), during-event (e.g.,chip transactions, sales), and post-event data (e.g.,credit refunds).
- A prototype dashboard will be developed using Python's Dash and Plotly libraries to present the key insights.
- The dashboard is intended for internal use and post-event analysis by the event organizer.

Excluded from the Scope

- **Real-Time Monitoring**: While the dashboard may be designed with real-time data potential, this study will focus solely on post-event analysis.
- **Multiple Event Comparisons**: This study is limited to the analysis of a single event (**The Event**) and does not involve comparative studies.
- **Data Collection**: The study does not involve the collection of new data and relies on data provided by **The Organizer** and the NFCtron system.
- **Implementation in NFCtron Hub**: The thesis focuses on analyzing data and developing a standalone prototype dashboard, not on direct integration into the NFCtron Hub platform.

Limitations

- **Anonymized Data**: To protect privacy, all customer and vendor data has been anonymized, which may limit certain the analysis in some ways.
- **Single Event Focus**: Insights and recommendations are based solely on data from **The Event**, which may limit broader generalizations.
- **Time Constraints**: Given the timeline of the thesis, certain advanced features (e.g.,predictive analytics) and technical implementations have been deprioritized but kept in mind for future work.

1 Data and Methodology

This chapter address the process and challenges of local environment setup, obtaining, preparing and anonymizing the data. Most importantly, this chapter describes and explains the data that was used for this research. It also briefly describes the tools, technologies and methods employed to answer the research questions.

1.1 Environment and local setup

To start off, we needed to set up some kind of environment where we would later work with the data. The data we would be working with was stored in a PostgreSQL database.

Having direct access to the production database to perform the analysis was not a secure and ethical way to go. Exporting only the necessary and raw data from the production database was an initial thought, but we initially did not know what data we would need, and by exporting we would lose all the relations between the tables.

Therefore, we decided to set up a local database with the same structure as the production database where we can query and analyze the data safely. The next step was to import the data from the production database to the local database. Importing or simple cloning the full database was also not an option, because only a small fraction of its subset was required.

So a deep internal analysis of the tables that were relevant to our study was performed. This resulted in a list of total 21 tables that held the necessary data for the study and were necessary to be imported.

1.1.1 Data Obtaining and Preparation

Almost every table was easily queried for the event and exported from the production database to a local CSV file. But some tables (for example and not surprisingly, the *transactions* table with over 140k rows) were too large to be exported in one piece, so we had to split them into smaller parts. Later, these parts were joined together to a single CSV file using a simple Python script.

Since no direct access to the production database was used for the export but rather a database management tool, the export was not as fast as it could be and took a significant amount of time. Moreover, the exported data, most importantly the timestamps, were in a different format than we needed. And also all numeric values were exported as a formatted strings with a comma as a decimal separator. So a data preprocessing Python script was written to convert such invalid columns to the correct format.

1.1.2 Local Database Setup

Then a step to set up the local PostgreSQL database was needed. Due to the nature of this study, we wanted to keep the setup as simple as possible, so we used the default PostgreSQL installation without using any special environment using Docker or similar. However, during this process I made a mistake and forgot that a PostgreSQL with PostGIS extension was needed. So it required to re-set up the database with the PostGIS extension.

The next step was to import the data from the CSV files to the local database. For further database handling, analysis and visualization, we used DataSpell, a Python IDE with a built-in database explorer and data visualization tools. DataSpell was then used for the local database import, which prior to it required some necessary database relations and constraints modifications, since the data was exported without them and was not relevant for the study.

This whole process resulted in approximately 387k rows of data in the local database that were ready to be queried and analyzed.

1.1.3 Local Database Modifications

Before any analysis was performed, some modifications to the local database were needed due to some known limitations and missing data.

Beverage Volumes

The first necessary limitation that the Beverage Consumption Analysis section heavily relied on was the missing information about beverage products volume in milliliters. This information was crucial for the analysis, so a new column was added to the relevant product information tables. However, the next step was to back-fill this information which was not easily automated.

The First approach was to write a Python script that would try to find the volume information from the product name. This worked for some products, but not for all since the naming convention was not consistent.

After several attempts to automate this process, it was decided to manually fill in the missing information since only 425 products were present in the database. Only 159 of them were of beverage type and thus eligible for the volume information.

Depositable Products

Since one of the research questions was to analyze the depositable cups and this information was not easily available in the database, a new column was added to the product information tables.

This was a simple binary column that indicated whether the product was depositable or not. Back-filling this information was also pretty straightforward since only one product was a depositable cup.

Venue Map Visualization

One of the initial ideas was to visualize the venue map with the locations of the selling places, top-up service points, stages and other important places.

This would be invaluable, the database was partially ready for this, but the data would be significantly time-consuming to back-fill and the later analysis and visualization would require more time.

Since these facts and the fast that this process of preparing the data took place before completing the list of data analysis questions, this idea was later abandoned.

Event Program

To present some time-related data and its correlation with the event program, it would require to have the event program in the database.

Again, the database was ready for this, but no event program was set up, since it was unnecessary for the event. Therefore, this required getting the event program from the festival website and manually insert it into the database.

This was manually a very time-consuming process, but it was necessary for the analysis. For some simplification of the process, an AI tool was used to extract the data from the program schedule screenshots and instructed to prepare an SQL script that would insert the data into the database.

This seemed like a good idea, but the AI tool was initially hallucinating and made up some incorrect data. But after several iterations, it successfully extracted the data and prepared the SQL script which was used and the event program was successfully inserted into the database.

In the end, I doubt that this process was faster than manual data entry, but it was a good exercise and a good example of how AI can be used to automate some processes.

1.2 Data Anonymization

The Data Anonymization process was necessary due to requirements initially set by the data provider and later by the ethical considerations. This step was performed for the already imported data in the local database. It required identifying the sensitive data and replacing them with anonymized values.

In this case, the most sensitive data were:

- **Vendor names**: Since it included the legal names of the vendors, it was necessary to anonymize them.
- **Selling places**: Some selling places were named after the vendors, so it was necessary to anonymize them as well.
- **Customer information**: Some tables included customer information like names, emails, phone numbers, etc.

The process could have been done various ways, but the fact that this study will not be exposing internal database structure, it was decided to perform the anonymization directly in the database.

However, if one-way anonymization were to be performed, it would permanently overwrite the original data, losing the possibility to switch from anonymized to original data. Therefore, a two-way anonymization process was chosen and performed.

This was particularly useful during the analysis phase, where the results would contain the original data for better understanding, fact-checking and for the internal presentation and consultations with the organizer.

It was done on the database level, where two new internal tables were introduced – *public.anonymization_config* and *public.original_values*.

Where the *public.anonymization_config* table held the configuration about which schema, table and column should be anonymized and how. For the usage, a simple SQL function was created to define the anonymization configuration in a simple JSON format, that looked like in the **Anonymization configuration example.**.

```
1  SELECT configure_anonymization('[
2    { "table": "schema.seller", "columns": ["legal_name", "name_int",
     ↪  "name_pub"] },
3    ...
4    { "table": "schema.user_account", "columns": ["email", "last_name",
     ↪  "first_name", "phone"] },
5    ]'::JSONB
6  );
```

Source code 1: Anonymization configuration example.

The *public.original_values* table was used to store the original values of the anonymized columns. Again, using a simple SQL function *anonymize_database()*, it would store the original values and anonymize the configured table columns.

One particular challenge was anonymizing the values smartly. It could have been easily done by replacing the values with random strings, hashes or encrypted values. But working with data, where a vendor is named *fa65165b923e9cc* is not very convenient.

Therefore, a simple SQL function was written to anonymize the value depending on the configuration. This allowed to configure the anonymization to:

- replace vendor names with values like *Vendor 1*,
- customer emails with *03b09592-d0eb-43a3-9941-30d38ade6bce@gmail.com* keeping the original email domain,
- selling places with values like *Place 1* where the original name contained sensitive information, but keep original values for places like *BAR L2*, etc.

In the end, it resulted in a database with anonymized values per stated configuration, that could have been used for the analysis and results presentation safely. However, the original values were still present in the database, and the database could have been anytime easily restored to the original state if needed and vice versa.

1.3 Data Structure

Without exposing the internal database structure, the abstract data structure this study was working with can be described as follows:

1.3.1 Financial Data

Transactions (approx. 300k rows): Transactional data, that holds the information about the type, amount, timestamps and links to products, places and other related entities. This analysis relies on and works with several transaction types,

including **top-up charge** and **refund** transactions [1], **order sales** and **refund** transactions [2] and **chip registrations**[3].

Credit Refunds (approx. 15k rows): Post-event credit refund requested by the customers with the information about the amount, timestamp, customer and the bank account to which the refund was sent. This data enables our analysis to correctly handle disposable credit balances, more information about the anonymous customers [4] and their behavior.

Why is it important?

These records also form the backbone of the analysis, enabling insights into revenue, sales trends, and customer behavior.

1.3.2 Customer Data

User Accounts (approx. 5k rows): Registered customer accounts with the information about the user and its potential online order history and other related information.

Tickets and Orders (approx. 30k rows): Information about the online sold tickets, its types, prices, timestamps, online order related information.

Why is it important?

With the above data, this analysis can work with more customer information supporting the customer behavior analysis, customer segmentation and other related analysis.

1.3.3 Event Data

Places (approx. 400 rows): Selling and top-up service points, zones for access control and its other relations.

Products (approx. 500 rows): Essentially a product catalog including the product name, price, category, volume, seller ownership and seller-organizer deal re-

[1] Top-up transactions mean funding or refunding chip credit balances.
[2] Order transactions mean spending the credit balance for products.
[3] Chip registrations mean records of when the chip bracelets were registered by the system.
[4] Anonymous customer essentially means a user without registered account.

lated links.

Important information about products is their supported categorization that will later be used for the sales analysis and can be seen in Table 1.1.

Category	Description
Nonalcoholic	Any non-alcoholic beverages (e.g., coffe, water, etc.)
Beer	Any kind of beer.
Wine	Any kind of wine.
Other Alcohol	Any other kind of alcoholic beverages (e.g., shots, cocktails, etc.).
Salty	Any salty snacks.
Sweet	Any sweet snacks.
Other	Any other products that do not fit into the above categories.

Table 1.1: Product categories.

Event Program (approx. 140 rows): Event program schedule with the information about the stages, performers, times and other related information.

Why is it important?

This data provides more context to the event when combined with the financial and customer data above. Enabling the analysis to work with the event program, its correlation with the sales, customer behavior and other related analysis.

1.3.4 Data Processing Views

To efficiently query the studied data during the analysis, several SQL views and functions were created to simplify and speed up the process.

Transaction Commission Calculation: A function that calculates the commission for each transaction based on the product and the seller-organizer deal. This was a crucial method required to calculate and analyze the commission from event order sales contributing to the organizer's revenue.

Transaction Enrichment: Since the transactional data consisted of several transaction types which were not easily distinguishable, a view was created to enrich the transaction data with the transaction type information. It also benefited from the transaction commission calculation function mentioned above which helped to easily calculate the commission for each transaction.

Chip Customers: Probably the most complex function that returns the customers at the event. Since the transactional data is architected using Event Sourcing, the customer information is not directly available and needs to be compiled from the transactional history. This function was constructed in a way, where it supports time-based filtering and provides extensive insights into the customers, which is shown in Table 1.2 below:

Column Name	Description
CHIP_ID	Unique chip identifier.
CHIP_TYPE	Type of chip (e.g., regular, VIP, online, staff).
REG_AT	Timestamp of chip registration.
FIRST_TRX	First transaction timestamp associated with the chip.
LAST_TRX	Last transaction timestamp associated with the chip.
LAST_BALANCE	Last known balance at the specified time frame.
ACTUAL_BALANCE	Balance at the specified time frame after credit refunds.
IS_BLOCKED	Indicates if the chip is blocked due to suspicious activity.
HOURS_ACTIVE	Total active hours of the chip (daily sum).
DAYS_ACTIVE	Total number of days the chip was active.
T_COUNT	Total number of transactions associated with the chip.
O_TOTAL_CNT	Total number of orders placed using the chip.
O_TOTAL_AMT	Total amount spent through orders.
O_MAX_AMT	Maximum amount spent in a single order.
O_AVG_AMT	Average amount spent per order.
O_MODE_AMT	Most common amount spent per order.
OS_AVG_AMT	Average amount spent on sales orders (excluding refunds).
OS_MODE_AMT	Most common sale amount (excluding refunds).
TU_TOTAL_CNT	Total number of top-ups made to the chip.
TU_TOTAL_AMT	Total amount credited to the chip via top-ups.
TU_MAX_AMT	Maximum amount credited in a single top-up.
TU_AVG_AMT	Average amount credited per top-up.
TU_MODE_AMT	Most common amount credited per top-up.
TU_CARD_BRAND	Used card brand for top-ups (e.g., Visa, Mastercard).
BR_AMT	Total amount refunded to the customer's bank account.
BR_EMAIL_DOMAIN	Domain of the refund request email (e.g., gmail.com).
BR_COUNTRY	Country associated with the bank account for refund.
BR_REQ_SOURCE	Source of the refund request (e.g., iOS, Android, Web).
BR_BANK_NAME	Name of the Czech bank used for the refund.
BR_CREATED	Timestamp when the refund request was created.
BR_APPROVED	Timestamp when the refund request was approved.
A_EMAIL_DOMAIN	Email domain of the account.
A_COUNTRY_NAME	Country associated with the account.
A_REQ_SOURCE	Source of the account creation (e.g., iOS, Android, Web).
EO_PAYMENT_METHOD	Payment method for orders (e.g., card, bank transfer).
EO_CARD_BRAND	Used card brand for online order (Visa, Mastercard)
EO_REQ_SOURCE	Source of the order request (e.g., iOS, Android, Web).

Table 1.2: Customer chips function return table.

1.4 Tools and Technologies

As mentioned earlier, this process utilized a variety of tools and technologies to handle the data and prepare for the analysis. These main tools and technologies included:

- **PostgreSQL**: An open-source relational database management system used to store and query the data.
- **PostGIS**: An extension of PostgreSQL that supports geospatial data, enabling spatial analysis and visualization.
- **DataSpell**: An integrated development environment (IDE) for data science and analytics, used for database exploration and data visualization.
- **Python**: Used for data preprocessing, querying, and analysis, along with libraries like Pandas for data manipulation.
- **Claude AI**: Utilized for extracting data from unstructured sources (e.g., program schedules) and automating repetitive tasks like data entry.

Using such tools during this process provided a convenient environment for data handling and processing, initial analysis and ensuring the accuracy and efficiency for further analysis and results presentation.

1.5 Conclusion

This chapter laid out the process of setting up the local environment, collecting and preparing the data, anonymizing sensitive information and introducing the data structure.

Some key challenges, such as missing beverage volume information, data anonymization, and event program integration, were addressed through a combination of manual and automated processes.

In the end, the use of SQL views for data enrichment and functions for further data processing ensured efficient querying and analysis. And hopefully laid the groundwork for answering the research questions and achieving the study's goals, as detailed in the later chapters.

For better understanding and visualization of this initial comprehensive Knowledge Data Discover process, a simplified diagram was created; that can be seen in Figure 1.1.

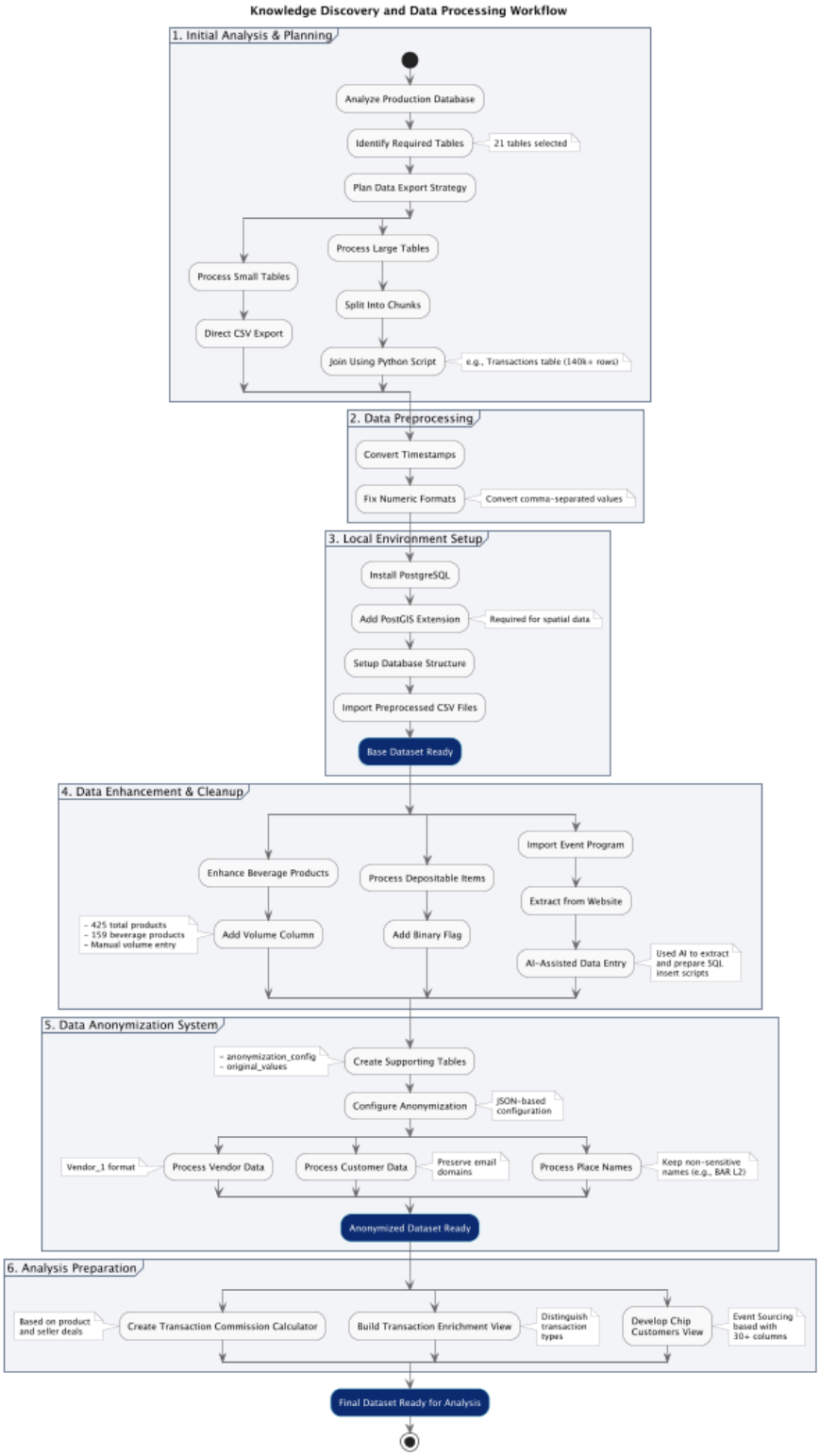


Figure 1.1: Knowledge Data Discovery workflow diagram.

2 Data Analysis and Results

This chapter presents the data analysis and results of the research. The goal is to address the research questions outlined earlier, with a focus on providing actionable insights for the event organizer.

The chapter is divided into several sections corresponding to key analytical areas:

1. **Cashflow and Revenue Sources Analysis**,
2. **Performance Indicators Analysis**,
3. **Beverage Consumption Analysis**,
4. and **Customer Analysis**

Each section focuses on a different aspect of the data analysis trying to answer the research questions, present quantitative results, visualizations, and interpretations.

2.1 Cashflow and Revenue Sources Analysis

This section provides a comprehensive view of the festival's financial performance and cash flows. It should answer critical questions about how finances were funded into the system, how were they processed, and what were the final outcomes.

For this analysis, four questions were previously formulated. However, they were reordered to better fit the narrative of the analysis and logical flow of the chapter:

1. ***RQ2:*** *How much and by what means was the balance topped up on the chips?*

2. ***RQ4:*** *What was the total sales of the event, how much of it was the sales of the organizer and how many external vendors?*

3. ***RQ3:*** *How much balance remained on all chips after the event and after refunds?*

4. ***RQ1:*** *What was the total revenue of the organizer and what does it consist of?*

In the end, this section should provide a clear picture of the financial flows during the event and easy understanding of the generated revenue from various sources.

2.1.1 Chip Top-Up Analysis

Research Question

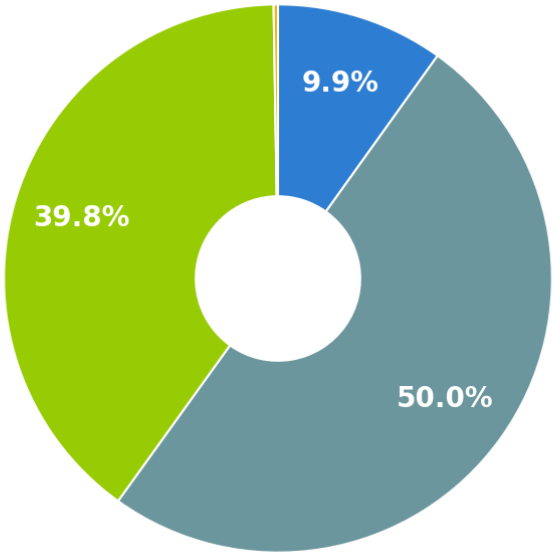
RQ2: *How much and by what means was the balance topped up on the chips?*

Attendees could top up their chip balances via online prepayments or on-site using cash or card. Additionally, the system allows to top-up "artificial" credit for VIP-issued chips which is also a mean of funding the system. However, these VIP credits are later not refundable, but this will be discussed in the next section.

This subsection quantifies these methods, highlighting their respective contributions to the overall top-up total.

To get the results, it was necessary to find all top-up transactions and their respective payment methods used. This resulted in **17,704** top-up transactions, with a total value of **14,520,973 CZK**.

When looking at the grouping by payment methods, the results in Figure 2.1 give a clear picture of the distribution.



Payment Method	Count	Total Value (CZK)
Card terminal	8,486	7,264,503 CZK
Cash	7,561	5,782,570 CZK
Online pre top-up	1,634	1,436,400 CZK
VIP issued	23	37,500 CZK

Figure 2.1: Top-Up Transactions by Payment Method

Thanks to the results, it is clear how many funds did the system receive and by what means.

Key Takeaways

- Total top-up amount was **14,520,973 CZK**.
- Most used payment method was card terminal at the event with 50% of all top-ups.
- Only around 10% of the top-ups were done online.

2.1.2 Sales Analysis

Research Question

RQ4: *What was the total sales of the event, how much of it was the sales of the organizer and how many external vendors?*

The sales analysis was crucial for understanding the overall sales behavior and served as a basis for further insights tightly connected to the revenue sources.

To answer the research question, it was necessary to find all sales transactions and their respective sellers and to divide them into two groups: the direct organizer's sales and external vendors' sales. And for better understanding, the sales were also grouped by the product categories (see Table 1.1 for the list of categories).

The results show that the total sales of the event were **11,711,807 CZK** with the organizer's sales being **8,240,264 CZK** and the external vendors' sales **3,471,543 CZK**.

The organizer, most importantly, sold all the beer beverages and most of the non-alcoholic and alcoholic (spirits) beverages. Whereas the external vendors sold mainly the food, wine beverages and other uncategorized products. This can be seen in Figure 2.2 below.

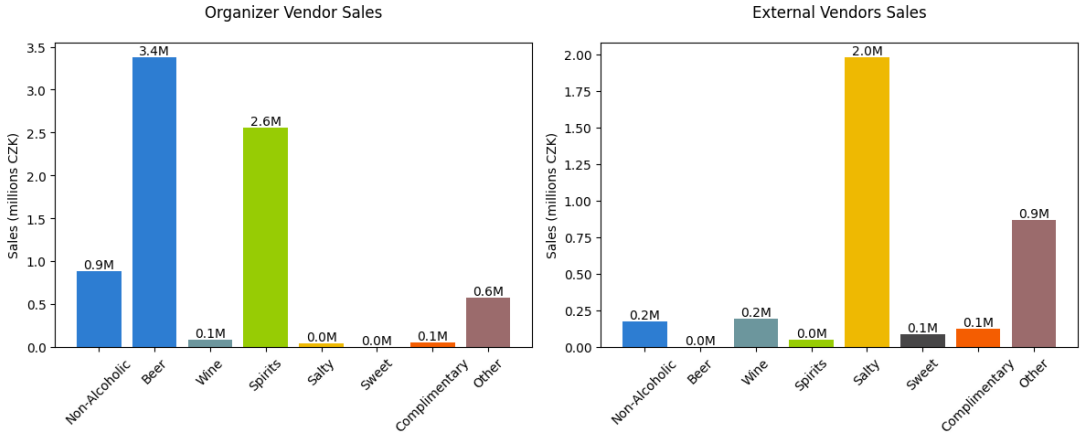


Figure 2.2: Sales of the Organizer vs. External Vendors

The organizer also sold not so little of uncategorized products, which after further investigation turned out to be ticket sales at the event amounting to **684,700 CZK**.

In total, the organizer direct sales were **70%** of the total sales, which is a significant portion, and thus the organizer itself has even bigger influence on the event's

financial performance.

Key Takeaways

- Total sales of the event were **11,711,807 CZK**, where organizer sales were **70%** of the total.
- The organizer sold all beer beverages and the majority of the non-alcoholic and alcoholic beverages.
- The organizer also sold tickets at the event amounting to **684,700 CZK**.
- External vendors sold mainly food, wine beverages, and other uncategorized products.

TODO: Better chart

2.1.3 Remaining Chip Balances

Research Question

RQ3: *How much balance remained on all chips after the event and after refunds?*

The remaining chip balances are crucial for the event organizer as they represent the potential revenue that can be still claimed. Any unclaimed balances after a given refund period, which is usually up to 14 days after the event will be considered as organizer's taxable revenue.

Out of the total top-up amount of **14,520,973 CZK**, the total spent credit amounted to **10,984,945 CZK**, which left a total of **3,536,028 CZK** on the chips before refunds. After refunds – done both at the event (**15,379 CZK**) and later via online bank refund requests (**3,163,567 CZK**) – the remaining balance was reduced to **357,082 CZK**.

However, this still included the artificially issued VIP credits with leftover balance of **12,405 CZK**. The system also reported integrity errors in the data, which resulted in a total of **10,246 CZK** due to fraudulent activities performed by some attendees which were automatically suspended by the system.

This left the total unclaimed balance at **334,431 CZK**, which has been claimed by the organizer as taxable revenue.

Since these numbers can be quite abstract, the results in a form of sankey diagram in Figure 2.3 below provide a clear picture of the flow of the funds.

TODO: Better chart

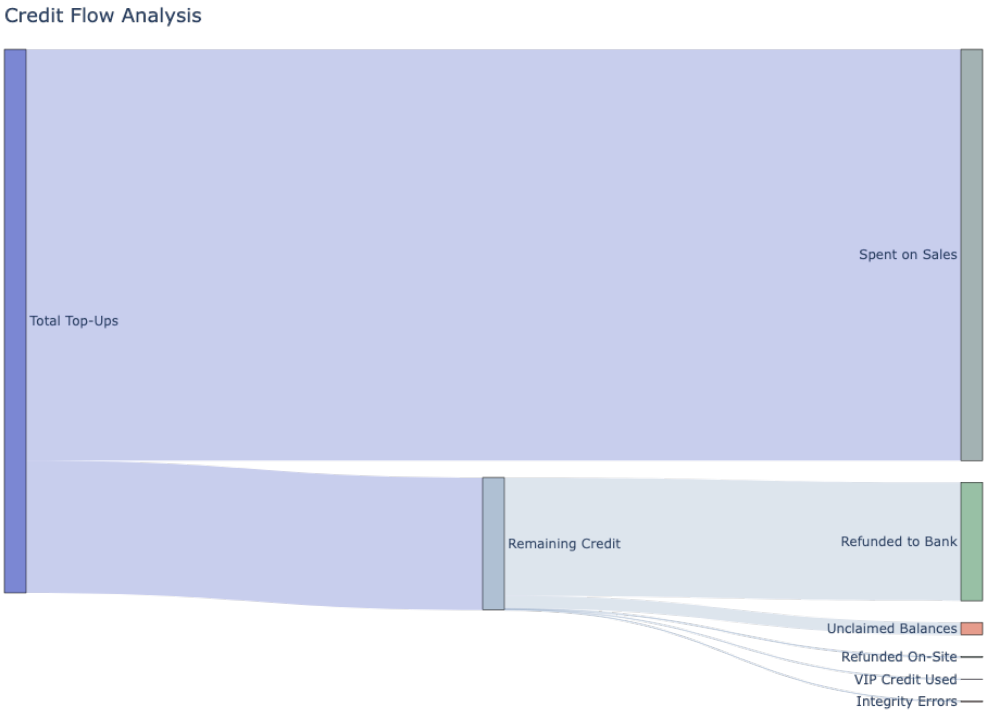


Figure 2.3: Remaining Chip Balances Sankey Diagram

Thanks to this breakdown, it is clear how the remaining balances were reduced and what was the final outcome. These results are important for the last part of this section, which is the total revenue of the organizer.

Key Takeaways

- Total unused credit was **3,536,028 CZK**.
- Credit refunded to customers was **3,178,946 CZK**.
- After VIP issued credits and system integrity error, the unclaimed balance was **334,431 CZK**.

2.1.4 Total Revenue of the Organizer

Research Question

RQ1: *What was the total revenue of the organizer and what does it consist of?*

The festival's financial model is based on a combination of revenue streams.

The most important stream is the **commission from the vendor sales**, which is arranged in advance between the organizer and the vendors. The commission is, in this case, a percentage (ranging from 15% to 30% depending on the deal) of the vendor sales amount without VAT.

Therefore, this required finding all sales transactions made at the external vendors' stands and calculating the commission based on the agreed percentage. However, this was not a straightforward task, since a transaction could contain multiple products even from different vendors.

This required a more complex calculation, for which was used the previously mentioned data processing views which were designed for this purpose. In the end, the total revenue from sales commissions was **820,712.79 CZK**.

Another source of revenue is the **unclaimed chip balances**, which, after a credit refund period, are considered as taxable revenue for the organizer. This, thanks to the previous subsection, was found to be **334,431 CZK**.

TODO: Better chart

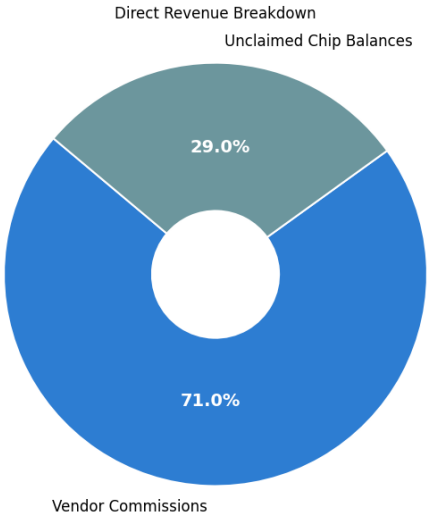


Figure 2.4: Breakdown of Direct Revenue Streams

Currently totalling **1,155,143.79 CZK** is the direct revenue of the organizer from the event and can be seen in Figure 2.4 above.

However, given the circumstances and setup of this event, there were also additional, but indirect revenue streams that were not included in the total revenue. These include **the online ticket sales**, which were sold by the organizer and **the direct sales of the organizer**. They were not included in the total direct revenue, as they may misinterpret the results since the analysis lacks expenses of the organizer.

If we were to include these, the total revenue would increase by **11,179,700 CZK** from the online ticket sales and **8,240,264 CZK** from the direct sales, which would result in a total revenue of **20,575,107.79 CZK**.

To better understand the revenue streams, the results are visualized in Table 2.1 and in Figure 2.5 below.

Revenue Stream	Amount (CZK)
Vendor Commissions	820,712.79 CZK
Unclaimed Chip Balances	334,431 CZK
Total Direct Revenue	**1,155,143.79 CZK**
Online Ticket Sales	11,179,700 CZK
Organizer Direct Sales	8,240,264 CZK
Total Revenue (All Streams)	**20,575,107.79 CZK**

Table 2.1: Revenue Summary Breakdown

TODO: Better chart

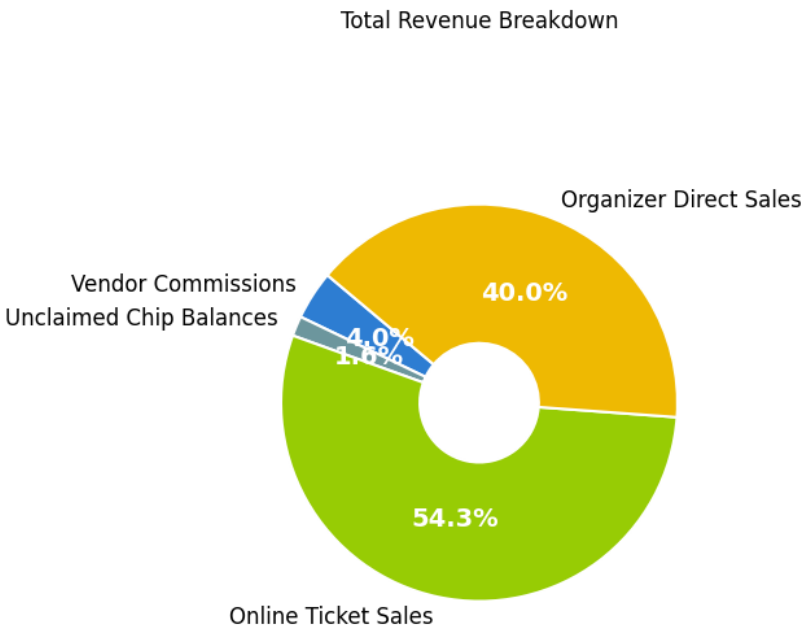


Figure 2.5: Breakdown of All Revenue Streams

Key Takeaways

- Total direct revenue of the organizer was **1,155,143.79 CZK**.
- Vendor sale commission contributed to approximately 71% of the total direct revenue.
- With other indirect revenue streams, the total revenue would be **20,575,107.79 CZK**.

2.1.5 Summary

This section provided a comprehensive view of the festival's financial performance and cash flows. The results covered the top-up transactions, sales analysis, remaining chip balances, and the total revenue of the organizer and contributed to a better understanding from the financial perspective of the festival.

Nevertheless, results covered in these subsections are only a part of the whole

picture and can be interpreted in various ways.

For this particular challenge, a summarized cash flow diagram of payments was created, containing thus only the direct revenue streams. This diagram can be seen in the Figure 2.6 below.

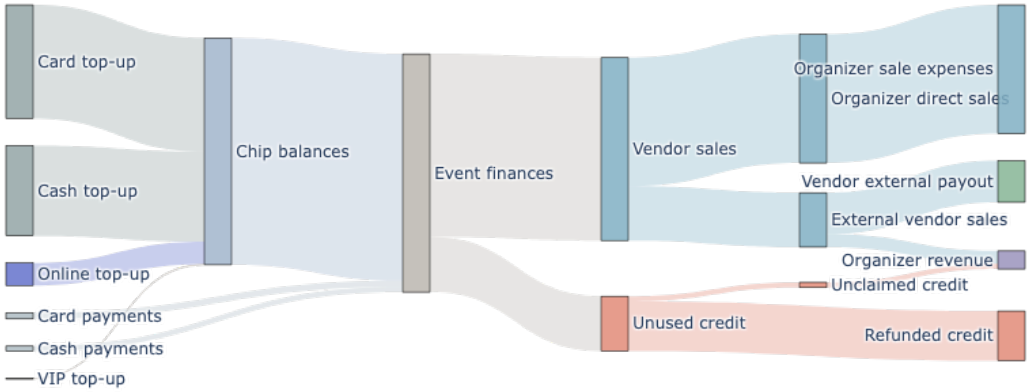


Figure 2.6: Overall Cash Flow Diagram

This diagram provides a clear overview of the financial flows during the festival and nicely summarizes the results of this analysis.

Key Takeaways

- Total incoming money flow was **14,520,973 CZK** from top-up transactions and **726,862 CZK** from non-chip sales.
- Total sales amounted to **11,711,807 CZK**.
- Which left a total of **3,536,028 CZK** in unused credit before refunds.
- After refunds and non-refundable chips, the remaining balance left was **334,431 CZK** claimed as taxable revenue.
- Commission from external vendor sales contributed to **820,712.79 CZK** of the total direct revenue.
- Together, the total direct revenue of the organizer was **1,155,143.79 CZK**.

2.2 Performance Indicators Analysis

This section focuses on the performance indicators of the event. The goal is to identify key metrics that can be used to further evaluate the event and its success. The potential of this analysis is to measure the "greatness" and the size of the event in terms of performance.

For this analysis, seven questions were previously formulated:

- ***RQ5:*** *How many transactions were processed in total and what was the largest "peak" in the volume of transactions processed by the system (and when)?*
- ***RQ6:*** *What was the average transaction processing time during peak hours?*
- ***RQ7:*** *Were there any significant delays or downtimes in processing transactions?*
- ***RQ8:*** *What were the best-selling points?*
- ***RQ9:*** *What were the best top-up points?*
- ***RQ10:*** *Who were the best vendors?*
- ***RQ11:*** *What were the best products?*

The results of this analysis should provide insights into the event's performance and help the organizer to understand the key metrics that can be used to evaluate the event's success.

To answer these questions, this section is divided into two parts:

1. **Transactions Processing Analysis**,
2. and **Best Sale & Top-Up Points, Vendors, and Products Analysis**.

2.2.1 Transactions Processing Analysis

This subsection will focus on the processing of transactions during the event in pursuit of answering the three first research questions of this section.

Research Question

RQ5: *How many transactions were processed in total and what was the largest "peak" in the volume of transactions processed by the system (and when)?*

This question actually consists of two sub-questions, which will be addressed separately.

The first part questions the total number of transactions processed during the event. Which was actually pretty straightforward to answer, as the system was designed to track all transactions and their respective types. The resulted total number of transactions was **141,381** consisting of **110,854** sales transactions, **17,726** top-up transactions and **12,801** chip register transactions.

The second part focuses on rather time-related metrics and asks about the processing peak times during the festival. For this part, it was necessary to spread out the above transactions over the time and find the peaks.

The results in Figure 2.7 below show the distribution of the processed transactions over time. It clearly identifies the peak on the last day of the festival at 18:00 amounting to **8,986** transactions.

TODO: Better chart

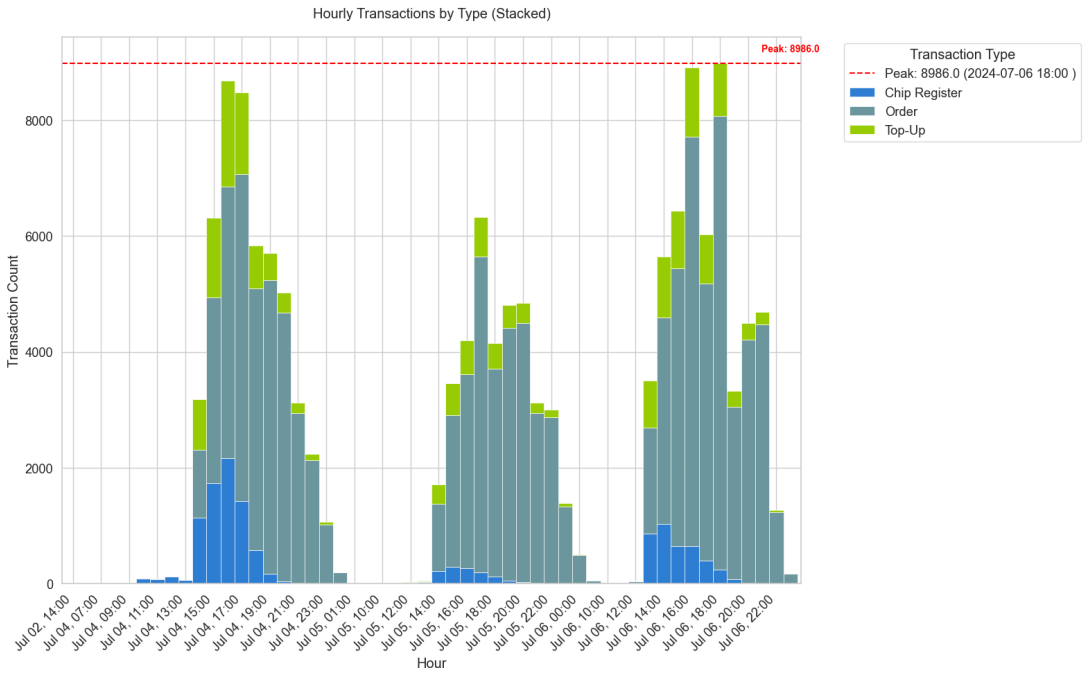


Figure 2.7: Transactions Processing Peaks

Key Takeaways

- Total number of transactions processed was **141,381** consisting mainly (78%) of order transactions.
- The peak of transactions was on the last day of the festival at 18:00 with **8,986** transactions processed at that hour.

The following two questions (RQ6 and RQ7), focus on processing times and potential delays during the event.

Research Question

RQ6: *What was the average transaction processing time during peak hours?*

The answer to this question is closely related to the previous one, as it requires the identification of the processing times during the peak times, which were already identified.

It required finding the average processing time, meaning difference between the transaction creation and its completion times.

What causes the processing time?

Time when the transaction is created is the time when the in-place offline-supported system created the transaction, and the processed time is later when the central system receives the transaction and processes it. The delays can be caused by various factors, such as network latency, offline mode active, system load, or even the transaction type.

The results show that the average processing time during the peak times was approximately **40** seconds.

When slightly changing the displayed data, we also get the answer to the RQ7 about the potential delays and downtimes during the event.

Research Question

RQ7: *Were there any significant delays or downtimes in processing transactions?*

The chart in the Figure 2.8 shows the distribution of the processing times over the time and identifies one high processing peak of approximately **13** minutes.

This is highly unusual and indicates a vendor misuse of the system or accidentally put the system into offline mode.

TODO: Better chart

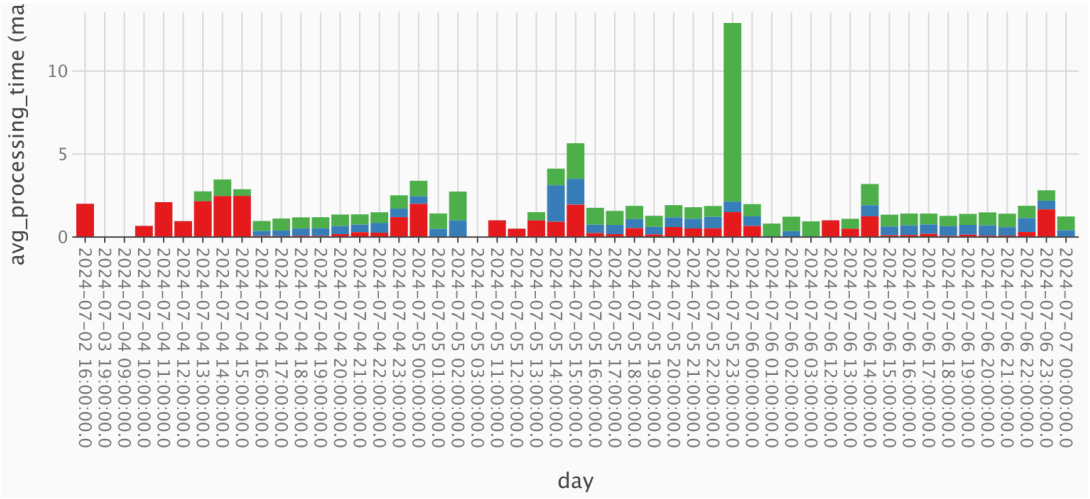


Figure 2.8: Transaction Processing Peaks

Other high peaks are visible on the second day in the afternoon with **5** minutes processing time, which was probably caused by the initial load on that day.

Key Takeaways

- The average processing time during the peak times was **40** seconds.
- The highest processing peak was approximately **13** minutes, indicating a potential misuse of the system.
- Other high peaks were visible on the second day in the afternoon with **5** minutes processing time.

These results provide insights into the system's performance during the event, its reliability, and potential bottlenecks. It also shows the festival's popularity and the system's ability to handle the load.

2.2.2 Best Sale & Top-Up Points, Vendors, and Products Analysis

In this subsection, the main goal will be to address the last four research questions of this section and provide insights into the best: selling points, top-up points, vendors, and products.

The problem with these question statements is that they are quite broad and can be interpreted in various ways. What does a "best" mean in this context?

It can be the most profitable, the best rated, the most visited, etc. But since we are exploring the performance indicators, the best should be understood as the "busiest". Which in terms of the system and this analysis should mean **the most transactions created** and the point's ability to handle the load.

Best Top-Up Points

The first focus will be on the best top-up points since unlike the selling points, vendors and products, the top-up points are not linked to any specific product or vendor.

Research Question

RQ9: *What were the best top-up points?*

To find these results, it required finding all top-up transactions, aggregate them in a bucket-like time frame and finally calculate their total counts, max peaks and averages over time.

This resulted in the following findings in the Table 2.2 below.

This indicates that the most busy top-up points were somehow evenly distributed with approximately around **1000 transactions** processed during the event with average peaks of around **100 transactions/hour**. The least busy top-up points were the specific ones, such as the Support tent, VIP and Accreditation points, which were not used so much for top-ups.

The overall distribution, shown in the Figure 2.9, also shows that Top-up points were more busy than Check-in points. That makes sense because the top-ups were done more frequently than the initial check-ins, but the check-ins were done

	Top-Up point	Customers	Transactions	Max trx./h
1	Pokladna 16	1,119	1,139	99
2	Pokladna 2	1,103	1,108	106
3	Pokladna 15	1,035	1,043	82
4	Pokladna 4	1,007	1,017	107
	...			
15	Pokladna 7	740	744	73
16	Pokladna 9	699	711	72
17	Pokladna 10	629	640	69
18	Odbavení	529	529	125
	...			
25	Odbavení 5	198	198	46
26	Odbavení 6	191	191	32
27	Support	64	85	11
28	VIP	23	23	5
29	Akreditace	17	17	8
30		0	0	0

Table 2.2: Best Top-Up Points

in a more concentrated time frame.

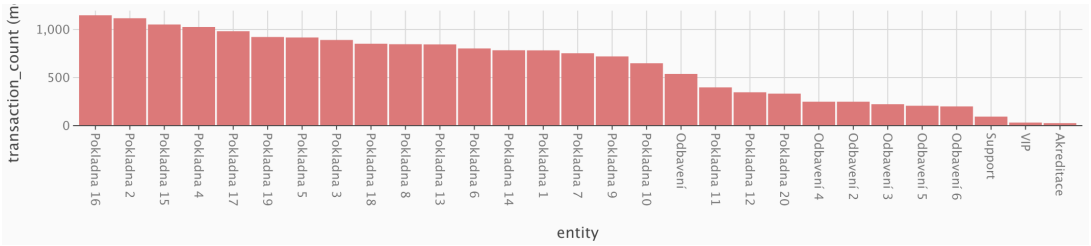


Figure 2.9: Best Top-Up Points

Especially **Odbavení** point processed only **529** transactions but peaked at **125** transactions per hour, which was much higher than the other check-in points and even higher than the best top-up points.

Key Takeaways

- The most busy top-up points processed around **1,000** transactions during the event.
- The average peak of the top-up points was around **100** transactions per hour.
- The least busy top-up points were the specific ones, such as the Support tent, VIP and Accreditation points.
- Check-in points were less busy than the top-up points, but the **Odbavení** point peaked at **125** transactions per hour.

Best Sale Points

The next focus will be on the best sale points, which are the points where the most orders were created.

Research Question

RQ8: *What were the best-selling points?*

The process of finding the best sale points was similar to the previous one, but this time it required finding all sales transactions and their respective points.

Out of total of **145** sale points, the best place was undeniable the **L20 PIVNÍ STAN 1** with total **10,114** orders processed and the maximum peak of **840** orders per hour.

Another interesting fact is the total unique users processed at the places. Where again the **L20 PIVNÍ STAN 1** processed **9,159** unique customers which, out of **10,009** active customers, is a significant portion (**91.50**%) of the total.

Based on this particular finding, we can assume that in the following analysis - the best vendors and products - the product preferences will be highly in favor of the beer beverages. And thus the best vendor will probably be the organizer, as they sold all the beer beverages at the festival.

	Sale Point	Customers	Orders	Max orders./h
1	L20 PIVNÍ STAN 1	9,159	10,114	840
2	Place 78	4,914	5,113	479
3	A19 PIVO 1	4,615	5,073	608
4	L16 FRISCO	3,420	3,646	235
5	Place 11	3,078	3,226	265
6	L22 BEEFEATER/ HAVANA 1	2,613	2,857	317
7	A19 VÝKUP KELÍMKŮ	2,683	2,743	316

Table 2.3: Best Sale Points

Key Takeaways

- The best sale point was the **L20 PIVNÍ STAN 1** with total **10,114** orders processed, which was **9.12**% of the total orders created.
- The best sale point peaked at **840** orders per hour.
- The **L20 PIVNÍ STAN 1** processed **9,159** unique customers, which was **91.50**% of all active customers.

In conclusion to these two questions, the results show clearly the busiest points of the festival and their ability to handle the load. However, the results can be visualized in a more interactive way, which would provide a better understanding of the data.

One especially interesting visualization of the best sale and top-up points would be a heatmap of the festival area with the points and their respective transaction counts. As this was initially intended to be part of the analysis, it was unfortunately not possible to create it due to the lack of the necessary data.

Best Vendors

TODO: -

Best Products

TODO: -

2.3 Beverage Consumption Analysis

TODO: -

2.4 Customer Analysis

TODO: -

3 Dashboard Implementation

TODO: This chapter describes the implementation of the analytical dashboard using Dash and Plotly.

4 Conclusion

4.1 Summary of Work

TODO: Write a summary of the work done in this thesis, recap the main objectives and how they were achieved. Note the key contributions of the research and practical side.

4.2 Reflections

TODO: Write reflections on the project, what was learned, how the project could change festival operations, etc.

4.3 Future Directions

TODO: Write about the future directions of the project, what could be done next, what could be improved, what could be added, etc.

List of Figures

Note: Unless stated otherwise, all figures are the author's own work.

List of Tables

List of Source Codes

List of Acronyms

List of Appendices