НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

КАФЕДРА ІНФОРМАТИКИ ТА ПРОГРАМНОЇ ІНЖЕНЕРІЇ

**КУРСОВА РОБОТА**

з дисципліни «Аналіз даних в інформаційних системах»

на тему: «Аналіз успіху фільмів за їх оцінюванням та оглядами»

Студента 2 курсу ІП-14 групи

Спеціальності: 121

«Інженерія програмного забезпечення»

Радзівіло В.А.

«ПРИЙНЯВ» з оцінкою

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

доц. Ліхоузова Т.А. / доц. Олійник Ю.О.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Підпис                    Дата

Київ - 2023 рік

Національний технічний університет України “КПІ ім. Ігоря Сікорського”

Кафедра інформатики та програмної інженерії

Дисципліна Аналіз даних в інформаційно-управляючих системах

Спеціальність 121 "Інженерія програмного забезпечення"

Курс 2 Група ІП-14 Семестр 4

**ЗАВДАННЯ**

**на курсову роботу студента**

|  |
| --- |
| Радзівіло Валерії Артемівни |

|  |  |
| --- | --- |
| 1.Тема роботи | Аналіз успіху фільмів за їх оцінюванням та оглядами |
|  | |
|  | |

|  |  |
| --- | --- |
| 2.Строк здачі студентом закінченої роботи | 19.06.2022 |

|  |  |
| --- | --- |
| 3. Вхідні дані до роботи | методичні вказівки до курсової робота, обрані дані з сайту |
| <https://www.kaggle.com/datasets/thedevastator/rotten-tomatoes-top-movies-ratings-and-technical> | |
| <https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney> | |
| <https://www.kaggle.com/datasets/atulanandjha/imdb-50k-movie-reviews-test-your-bert> | |
| <https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset> | |

4.Зміст розрахунково-пояснювальної записки (перелік питань, які підлягають розробці)

|  |
| --- |
| 1.Постановка задачі |
| 2.Аналіз предметної області |
| 3.Розробка сховища даних |
| 4.Інтелектуальний аналіз даних |

5.Перелік графічного матеріалу ( з точним зазначенням обов’язкових креслень )

|  |  |
| --- | --- |
|  | |
|  | |
|  | |
|  | |
| 6.Дата видачі завдання | 16.04.2022 |

**КАЛЕНДАРНИЙ ПЛАН**

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Назва етапів курсової роботи | Термін виконання етапів роботи | Підписи керівника, студента |
| 1. | Отримання теми курсової роботи | 16.04.2022 |  |
| 2. | Визначення зовнішніх джерел даних | … |  |
| 3. | Пошук та вивчення літератури з питань курсової роботи |  |  |
| 4. | Розробка моделі сховища даних |  |  |
| 4. | Розробка ETL процесів |  |  |
| 5. | Обґрунтування методів інтелектуального аналізу даних |  |  |
| 6. | Застосування та порівняння ефективності методів інтелектуального аналізу даних |  |  |
| 7. | Підготовка пояснювальної записки |  |  |
| 8. | Здача курсової роботи на перевірку |  |  |
| 9. | Захист курсової роботи | 19.06.2022 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Студент |  |  | ХХХ |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |

|  |  |  |  |
| --- | --- | --- | --- |
| Керівник |  |  | доц. Ліхоузова Т.А |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |
| Керівник |  |  | доц. Олійник Ю.О. |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |

"26" червня 2022 р.

**АНОТАЦІЯ**

Пояснювальна записка до курсової роботи: 85 сторінок, 52 рисунки, 10 таблиць, 11 посилань.

Об’єкт дослідження: інтелектуальний аналіз даних.

Предмет дослідження: створення програмного забезпечення, що проводить аналіз даних з подальшим прогнозуванням та графічним відображенням результатів.

Мета роботи: проектування та реалізація сховища даних та ETL процесів, а також реалізація програмного забезпечення для отримання даних зі сховища та їх подальшого аналізу та прогнозування.

Дана курсова робота включає в себе: опис проектування, створення та заповнення сховища даних за даною задачею за допомогою фізичної моделі бази даних, опис створення програмного забезпечення для інтелектуального аналізу даних, їх графічного відображення та прогнозування за допомогою різних моделей.

СХОВИЩЕ ДАННИХ, МОДЕЛЬ ПРОГНОЗУВАННЯ, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ФІЗИЧНА МОДЕЛЬ БАЗИ ДАНИХ, ELT ПРОЦЕСИ, МОДЕЛЬ SARIMAX, ЧАСОВИЙ ПРОМІЖОК, ДОВІРЧІ ІНТЕРВАЛИ, МОДЕЛЬ ARIMA.

**Зміст**

[вступ 5](#_Toc419641933)

[1.Постановка задачі 7](#_Toc419641934)

[2.Аналіз ПРЕДМЕТНОЇ ОБЛАСТІ 9](#_Toc419641935)

[3.розробка сховища даних 11](#_Toc419641936)

[3.1.РОЗРОБКА МОДЕЛІ СХОВИЩА ДАНИХ……... 11](#_Toc419641937)

[3.2.РОЗРОБКА ETL ПРОЦЕСІВ 19](#_Toc419641938)

[3.3.ЗАВАНТАЖЕННЯ ДАНИХ ЗА ДОПОМОГОЮ ETL ПРОЦЕСІВ…. 20](#_Toc419641938)

[4.ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ 37](#_Toc419641939)

[4.1.ОБҐРУНТУВАННЯ ВИБОРУ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ 37](#_Toc419641940)

[4.2. МАТЕМАТИЧНА МОДЕЛЬ ПРОГНОЗУВАННЯ ЧАСОВИХ ПРОМІЖКІВ 38](#_Toc419641940)

[4.3.ПРАКТИЧНЕ ЗАСТОСУВАННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ЧАСОВИХ ПРОМІЖКІВ 40](#_Toc419641940)

4.4.ПОРІВНЯННЯ ЕФЕКТИВНОСТІ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ 56

[Висновки 58](#_Toc419641949)

[Перелік посилань 60](#_Toc419641950)

ДОДАТок А ТЕКСТИ ПРОГРАМНОГО КОДУ 61

**Вступ**

Фільми мають важливе значення в житті людей. Вони є джерелом культурної спадщини, що допомагає зберегти та передати цінності й ідеї наступним поколінням. Фільми можуть надихати на нові ідеї, стимулювати наші почуття та емоції, допомагати розв'язувати проблеми та приймати важливі рішення. Вони також допомагають краще зрозуміти світ, у якому ми живемо, та почуватися зв'язаними з іншими людьми. Фільми навчають новим навичкам, збагачують нашу освіченість та можуть слугувати засобом розваги та відпочинку, знімаючи стрес і покращуючи настрій.

Аналіз даних в області фільмів може мати дуже значний вплив на розуміння споживачів та ефективність кінематографічної індустрії. Використання аналізу даних може допомогти кінокомпаніям визначити популярність різних жанрів та створити більш успішні фільми, залучаючи більше глядачів. Крім того, аналіз даних може допомогти зрозуміти, які елементи фільму найбільш сприймаються глядачами, такі як сюжет, головні герої, музика та багато іншого. Це може допомогти виробникам фільмів створювати більш привабливі продукти для своєї аудиторії. Крім того, аналіз даних може допомогти кінотеатрам і компаніям з продажу фільмів оптимізувати свої бізнес-процеси, збільшити прибуток та поліпшити якість обслуговування своїх клієнтів. Таким чином, використання аналізу даних в області фільмів може допомогти покращити якість продукту, збільшити прибуток та покращити взаємодію між кінематографічною індустрією та глядачами.

В рамках даної курсової роботи розроблено сховище даних відеоігор, на основі фізичної моделі бази даних, функціонал якої було розроблено за допомогою SQL скриптів, а саму роботу з базою даних представлено через реалізацію програмного забезпечення для імплементації ETL процесів та інтелектуального аналізу обраних даних.

В ролі системи керування сховищем даних для даної роботи буде виступати MySQL, а мова програмування для реалізації застосунку – Python3.

**1.Постановка задачі**

Під час виконання курсової роботи необхідно виконати наступні завдання:

Створення сховища даних типу «сніжинка». Структура курсової роботи узгоджена з керівником. Сховище даних повинне містити щонайменше 6 таблиць вимірів та таблицю фактів. Створення ETL процесів для завантаження даних до сховища, їх отримання зі самого сховища за допомогою запитів, а також оновлення та додавання даних до таблиць вимірів за допомогою SCD першого та другого типів.

Реалізувати спроектоване сховище даних з використанням PostgreSQL версії 12.1.

Створення застосунку, що отримує вибірку даних зі створеного сховища, графічно відображає отримані дані, проводить їх інтелектуальний аналіз для отримання передбачення за допомогою різних моделей прогнозування.

Проведення аналізу вибірки на сезонність, стаціонарність, шум за допомогою доповненого тесту Дікі-Фуллера, використовуючи декомпозицію даних.

Провести згладжування даних такими методами, як згладжування ковзного середнього, вейвлет-згладжування та швидке згладжування Фур’є.

Для аналізу даних використати такі моделі прогнозування, як Exponential Smoothing, ARIMA, SARIMAX, Holt. Для кожного прогнозу знайти середнє квадратичне відхилення.

Застосування найоптимальнішої моделі прогнозування на згладжувальних даних та порівняння отриманих результатів з прогнозуванням на початкових даних.

Провести автокореляцію та часткову автокореляцію для знаходження повторюваних ділянок даних, незалежно від шумів у даних.

Аналіз отриманих результатів, порівняння різних методів прогнозування на даній вибірці, отримання найоптимальнішого методу.

Знаходження довірчих інтервалів, використовуючи найоптимальнішу модель прогнозування даних.

Використати мову програмування Python 3 для реалізації застосунку.

Курсовий проект здати до дедлайну (початок сесії) та виконати у єдиному стилі написання коду (coding style).

**2.АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ**

Сучасна кінематографія є однією з найбільш динамічно-розвиваючих галузей світової економіки, що пов'язана з розробкою, виробництвом та дистрибуцією фільмів. Фільми є популярною формою розваги та масової культури, яка привертає все більше глядачів. Успіх фільму залежить від багатьох елементів, таких як сценарій, режисура, акторська гра, монтаж та багато іншого. Аналіз даних може допомогти кінокомпаніям визначити популярність різних жанрів та створити більш успішні фільми. Крім того, аналіз даних може допомогти зрозуміти, які елементи фільму найбільш сприймаються глядачами, що дозволить створювати більш привабливі продукти для своєї аудиторії. Таким чином, якісна розробка та вибір елементів має ключову роль в створенні будь-якого фільму.

У програмному забезпеченні буде реалізовано наступну функціональність, що включає в себе:

* проектування сховища даних;
* створення ETL процесів для завантаження і оновлення даних;
* створення вибірки даних з сховища;
* інтелектуальний аналіз даних;
* використання декількох моделей прогнозування даних;
* прогнозування перспектив продажів ігор;
* прогнозування факторів впливу на популярність ігор;
* графічне відображення отриманих результатів та їх аналіз.

**3.РОЗРОБКА СХОВИЩА ДАНИХ**

3.1 Розробка моделі сховища даних

Для виконання курсової роботи було обрано 4 джерела відкритих даних на сайті https://www.kaggle.com//, що включають в себе:

* IMDB Top 250 Movies Dataset: <https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset>
* Rotten Tomatoes Top Movies Ratings and Technical: <https://www.kaggle.com/datasets/thedevastator/rotten-tomatoes-top-movies-ratings-and-technical>
* Streaming Services for movies: <https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney>
* IMDB 50K Movie Reviews: <https://www.kaggle.com/datasets/atulanandjha/imdb-50k-movie-reviews-test-your-bert>

Предметною областю лабораторної роботи є фільми та їх оцінка. У моделі

сховища за типом сніжинка спроектовано чотири таблиці фактів та п’ять

таблиць вимірів.

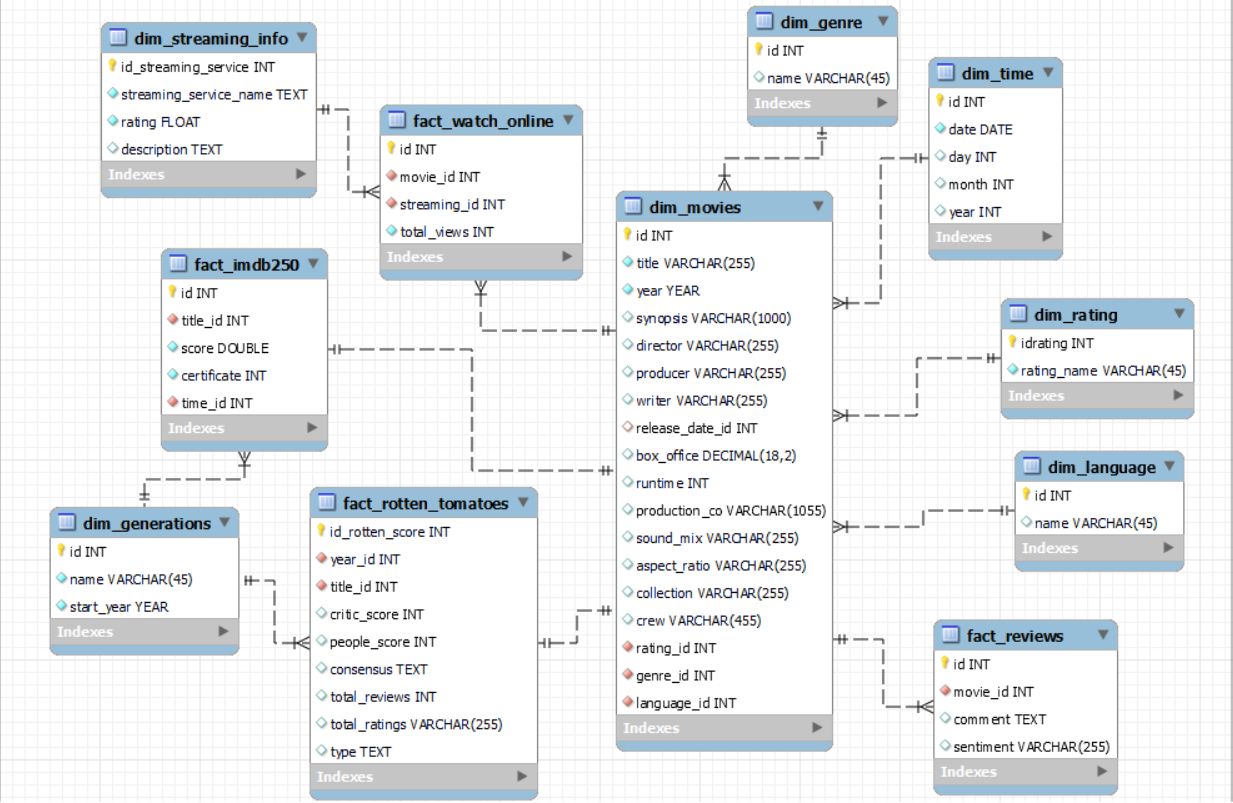


Рисунок 1.1. Проектування моделі сховища за типом сніжинка

Таблиця 1 – Таблиця виміру мов фільмів dim\_languages

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| idlanguages | Унікальний ідентифікатор мови |
| name | Назва мови |

Таблиця 2 – Таблиця виміру жанрів фільмів dim\_genres

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| idgenres | Унікальний ідентифікатор жанру |
| name | Назва жанру |

Таблиця 3 – Таблиця виміру мов фільмів dim\_rating

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| idrating | Унікальний ідентифікатор вікового рейтингу |
| rating\_name | Назва рейтингу |

Таблиця 4 – Таблиця фактів відгуків до фільмів fact\_reviews

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| id | Унікальний ідентифікатор відгуку |
| movie\_id | Ідентифікатор фільму |
| comment | Текст відгуку |
| sentiment | Текст емоційного відклику у коментатора |

Таблиця 5 – Таблиця виміру інформації про фільми dim\_movies

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| id | Унікальний ідентифікатор фільму |
| title | Назва фільму |
| year | Рік виходу фільму |
| synopsis | Короткий опис сюжету фільму |
| director | Режисер фільму |
| producer | Продюсер фільму |
| writer | Сценарист фільму |
| release\_date | Дата релізу |
| box\_office | Касові збори фільму |
| runtime | Тривалість фільму |
| production\_co | Компанія якій належать права на фільм |
| sound\_mix | Звукові ефекти в фільмі |
| aspect\_ratio | Формат фільму |
| collection | Колекція |
| crew | Список акторів фільму |
| rating\_id | Ідентифікатор вікового рейтингу |
| genre\_id | Ідентифікатор жанру |
| language\_id | Ідентифікатор мови |

Таблиця 6 – Таблиця фактів стрімінгових платформ для фільмів fact\_watch\_online

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| id\_streaming | Унікальний ідентифікатор стрімінгу |
| title\_id | Ідентифікатор фільмів |
| streaming\_service\_id | Ідентифікатор стрімінгового сервісу |

Таблиця 7 – Таблиця виміру стрімінгових сервісів для фільмів dim\_streaming\_info

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| id\_streaming\_service | Унікальний ідентифікатор стрімінгового сервісу. |
| streaming\_service\_name | Назва стрімінгового сервісу |
| rating | Рейтинг стрімінгового сервісу |
| description | Опис сервісу |

Таблиця 8 – Таблиця фактів рейтингу фільмів від imdb fact\_imdb250

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| rank\_imdb\_id | Унікальний ідентифікатор оцінки від imdb |
| title\_id | Ідентифікатор фільму |
| score | Оцінка від imdb |

Таблиця 9 – Таблиця фактів рейтингу фільмів від rotten\_tomatoes fact\_rotten\_tomatoes

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| id\_rotten\_score | Унікальний ідентифікатор оцінки від rotten tomatoes |
| title\_id | Ідентифікатор фільму |
| critic\_score | Оцінка від rotten tomatoes |
| people\_score | Оцінка від глядачів |
| consensus | Висновок |
| total\_reviews | Кількість відгуків |
| total\_ratings | Кількість оцінок |
| type | Визначений тип фільму |

Таблиця 10 – Таблиця виміру часу випуску фільмів dim\_generations

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| id | Унікальний ідентифікатор часу |
| name | Назва епохи цього часу |
| start\_year | Початковий рік часу |

Таблиця 10 – Таблиця виміру дати випуску фільмів dim\_time

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| id | Унікальний ідентифікатор часу |
| date | Повна дата |
| day | Виокремлений день |
| month | Виокремлений місяць |
| year | Виокремлений рік |

Створення таблиці виміру мов фільмів dim\_languages в сховищі даних:

CREATE TABLE `dim\_languages` (

`idlanguages` int NOT NULL AUTO\_INCREMENT,

`name` varchar(45) NOT NULL,

PRIMARY KEY (`idlanguages`),

UNIQUE KEY `name\_UNIQUE` (`name`)

)

Створення таблиці виміру жанрів фільмів dim\_genres в сховищі даних:

CREATE TABLE `dim\_genres` (

`idgenres` int NOT NULL AUTO\_INCREMENT,

`name` varchar(45) NOT NULL,

PRIMARY KEY (`idgenres`)

)

Створення таблиці виміру вікового рейтингу фільмів dim\_rating в сховищі даних:

CREATE TABLE `dim\_rating` (

`idrating` int NOT NULL AUTO\_INCREMENT,

`rating\_name` varchar(45) NOT NULL,

PRIMARY KEY (`idrating`),

UNIQUE KEY `rating\_name\_UNIQUE` (`rating\_name`)

)

Створення таблиці фактів відгуків до фільмів fact\_reviews в сховищі даних:

CREATE TABLE `fact\_reviews` (

`id` int NOT NULL AUTO\_INCREMENT,

`movie\_id` int NOT NULL,

`comment` text,

`sentiment` varchar(255) DEFAULT NULL,

PRIMARY KEY (`id`),

KEY `review\_movie\_idx` (`movie\_id`),

CONSTRAINT `review\_movie` FOREIGN KEY (`movie\_id`) REFERENCES `dim\_movies` (`id`)

)

Створення таблиці виміру інформації про фільми dim\_movies в сховищі даних:

CREATE TABLE `dim\_movies` (

`id` int NOT NULL AUTO\_INCREMENT,

`title` varchar(255) NOT NULL,

`year` int NOT NULL,

`synopsis` varchar(1000) DEFAULT NULL,

`director` varchar(255) DEFAULT NULL,

`producer` varchar(255) DEFAULT NULL,

`writer` varchar(255) DEFAULT NULL,

`release\_date` varchar(255) DEFAULT NULL,

`box\_office` varchar(255) DEFAULT NULL,

`runtime` varchar(255) DEFAULT NULL,

`production\_co` varchar(1055) DEFAULT NULL,

`sound\_mix` varchar(255) DEFAULT NULL,

`aspect\_ratio` varchar(255) DEFAULT NULL,

`collection` varchar(255) DEFAULT NULL,

`crew` varchar(455) DEFAULT NULL,

`rating\_id` int NOT NULL,

`genre\_id` int NOT NULL,

`language\_id` int NOT NULL,

PRIMARY KEY (`id`),

UNIQUE KEY `title\_UNIQUE` (`title`),

KEY `language\_idx` (`language\_id`),

KEY `rating\_idx` (`rating\_id`),

KEY `genre\_idx` (`genre\_id`),

CONSTRAINT `genre\_mov` FOREIGN KEY (`genre\_id`) REFERENCES `dim\_genres` (`idgenres`),

CONSTRAINT `language\_mov` FOREIGN KEY (`language\_id`) REFERENCES `dim\_languages` (`idlanguages`),

CONSTRAINT `rating\_mov` FOREIGN KEY (`rating\_id`) REFERENCES `dim\_rating` (`idrating`)

)

Створення таблиці фактів стрімінг фільмів fact\_watch\_online в сховищі даних:

CREATE TABLE `fact\_watch\_online` (

`id\_streaming` int NOT NULL AUTO\_INCREMENT,

`title\_id` int NOT NULL,

`streaming\_service\_id` int NOT NULL,

PRIMARY KEY (`id\_streaming`),

UNIQUE KEY `title\_id\_UNIQUE` (`title\_id`),

KEY `stream\_movie\_idx` (`title\_id`),

KEY `stream\_platform\_idx` (`streaming\_service\_id`),

CONSTRAINT `stream\_movie` FOREIGN KEY (`title\_id`) REFERENCES `dim\_movies` (`id`),

CONSTRAINT `stream\_platform` FOREIGN KEY (`streaming\_service\_id`) REFERENCES `dim\_streaming\_info` (`id\_streaming\_service`)

)

Створення таблиці виміру інформації про стрімінгові сервіси dim\_streaming\_info в сховищі даних:

CREATE TABLE `dim\_streaming\_info` (

`id\_streaming\_service` int NOT NULL AUTO\_INCREMENT,

`streaming\_service\_name` text NOT NULL,

`rating` float NOT NULL,

`description` text,

PRIMARY KEY (`id\_streaming\_service`)

)

Створення таблиці фактів imdb оцінки фільмів fact\_imdb250 в сховищі даних:

CREATE TABLE `fact\_imdb250` (

`id` int NOT NULL AUTO\_INCREMENT,

`title\_id` int NOT NULL,

`score` double NOT NULL,

`certificate` int NOT NULL,

`time\_id` int NOT NULL,

PRIMARY KEY (`id`),

KEY `imdb\_film\_idx` (`title\_id`),

KEY `imdb\_time` (`time\_id`),

CONSTRAINT `imdb\_film` FOREIGN KEY (`title\_id`) REFERENCES `dim\_movies` (`id`),

CONSTRAINT `imdb\_time` FOREIGN KEY (`time\_id`) REFERENCES `dim\_time` (`id`)

)

Створення таблиці фактів rotten tomatoes оцінки фільмів fact\_rotten\_tomatoes в сховищі даних:

CREATE TABLE `fact\_rotten\_tomatoes` (

`id\_rotten\_score` int NOT NULL AUTO\_INCREMENT,

`title\_id` int NOT NULL,

`critic\_score` int DEFAULT NULL,

`people\_score` int DEFAULT NULL,

`consensus` text,

`total\_reviews` int DEFAULT NULL,

`total\_ratings` text,

`type` text,

PRIMARY KEY (`id\_rotten\_score`),

UNIQUE KEY `title\_id\_UNIQUE` (`title\_id`),

KEY `rotten\_movie\_idx` (`title\_id`),

CONSTRAINT `rotten\_movie` FOREIGN KEY (`title\_id`) REFERENCES `dim\_movies` (`id`)

)

Створення таблиці виміру часу для рейтингу dim\_generations в сховищі даних:

CREATE TABLE `dim\_generations` (

`id` int NOT NULL AUTO\_INCREMENT,

`name` varchar(45) NOT NULL,

`start\_year` year NOT NULL,

PRIMARY KEY (`id`)

)

Створення таблиці виміру часу для виходу фільмів dim\_time в сховищі даних:

CREATE TABLE `dim\_time` (

`id` int NOT NULL AUTO\_INCREMENT,

`date` date NOT NULL,

`day` int DEFAULT NULL,

`month` int DEFAULT NULL,

`year` int DEFAULT NULL,

PRIMARY KEY (`id`),

UNIQUE KEY `date\_UNIQUE` (`date`)

)

3.2 Розробка ETL процесів

Для проектування та подальшої роботи з сховищем даних потрібно врахувати комплекс методів, які реалізовують процес переносу початкових даних в аналітичний формат, який буде підтримуватись в сховищі даних та не порушуватиме цілісність системи. Для цього опишемо основні функції ETL процесів, що включають:

**Процес завантаження.**

Цей етап передбачає зтягування даних довільної якості в ETL для подальшої обробки. Вхідні дані можуть бути в джерелах різних типів і форматів, бути створені в різних додатках, і, крім того, можуть використовувати різне кодування. Саме тому в подальшому для вирішення завдань аналізу дані повинні бути перетворені в єдиний універсальний формат, щоб запобігти вмісту різних факторів, які ускладнююють коректний аналіз даних.

**Процес валідації даних.**

На цьому етапі головною задачею є перевірка даних на коректність і повноту, після чого складається звіт про помилки для виправлення та узагальнення даних. Первинне очищення даних в процесі ETL носить переважно технічний характер, в якому основне завдання - підготувати дані до завантаження в сховище. Вторинна очистка в аналітичній системі є користувальницькою та спрямована на підготовку даних до вирішення конкретної аналітичної задачі. Завдяки чому обидва етапи очищення однаково важливі і необхідні.

**Перетворення даних.**

Даний етап ETL процесу потрібний для підготовки даних до розміщення в сховище даних і приведення їх до найбільш зручного виду для подальшого аналізу. При цьому повинна враховуватися вимога рівня якості даних. Тому в процесі перетворення можуть бути задіяні найрізноманітніші інструменти, які можуть бути як і найпростішими засобами ручного редагування даних так і реалізація досить складних методів обробки та очищення даних.

**Завантаження даних в сховище.**

Після того як дані отримано з різних джерел і виконано перетворення, агрегація і очищення даних, здійснюється останній етап ETL, тобто завантаження даних в сховище. Процес завантаження полягає в перенесенні даних з проміжних таблиць в структури сховища даних. Саме тому, щоб пришвидшити завантаження даних потрібно заздалегідь перевірити дані на повноту та коректність, а після спроектувати та обрати найоптимальніший процес заповнення сховища. При завантаженні даних в сховище переноситься не вся інформація з джерел, а тільки та, яка була змінена протягом проміжного часу, що пройшов з попереднього завантаження. ….

3.3 Завантаження даних за допомогою ETL процесів

Для заповнення сховища даних відеоігор було реалізовано наступне:

1. Було зведено дані до одного типу. А саме
   1. Для таблиць фактів fact\_rotten\_tomatoes, fact\_imdb250, fact\_watch\_online, fact\_reviews було зведено назву фільму до коду з таблиці dim\_movies.
   2. Для таблиці виміру dim\_movies мова була зведена до одного формату, мова гри була зведена до коду з таблиці dim\_language або 32 – інша мова, жанри були зведені до коду з таблиці dim\_genre або 24 – інший жанр, вікові рейтинги - аналогічно.
   3. Для таблиці fact\_imdb250 було зведено час до одного формату через збережену процедуру create\_fact\_imdb250
   4. Для таблиці fact\_watch\_online було зведено стрімінгову платформу до одного формату через збережену процедуру create\_fact\_watch\_online
   5. Для таблиці dim\_movies box\_office був зведений до одного числа -кількість мільйонів зароблених фільмом за допомоги функції extract\_million
   6. Для таблиці dim\_movies runtime було зведено до кількості хвилин, яку триває фільм за допомоги функції minutes\_from\_duration
   7. Для таблиці fact\_rotten\_tomatoes total\_ratings було зведено до мінімальної кількості оцінок за допомоги функції trim\_from\_plus та parse\_number
   8. Для таблиці dim\_movies жанр був зведений до першого з перелічених в сеті за допомоги функції trim\_first\_word
2. Було отримано дані з Stage зони для заповнення сховища даних. А саме:
   1. За допомоги збереженої процедури get\_movies було отримано повну інформацію про фільми з датасету rotten, на які було здійснено оцінку для таблиці dim\_movies.
   2. За допомоги збереженої процедури get\_languages було отримано усі мови з фільмів з датасету rotten, на які було здійснено оцінку для таблиці dim\_languages.
   3. За допомоги збереженої процедури get\_rating було отримано усі вікові обмеження з фільмів з датасету rotten, на які було здійснено оцінку для таблиці dim\_rating.
   4. За допомоги збереженої процедури get\_years було отримано усі роки оцінки фільмів з датасету imdb250, на основі яких було здійснено оцінку поколінь оглядів фільмів для таблиці dim\_time.
3. Було створено функції завантаження даних у сховище даних. А саме:
   1. За допомогою збереженої процедури create\_fact\_imdb250 було завантажено код фільму, оцінка, сертифікат та код часу для таблиці fact\_imdb.

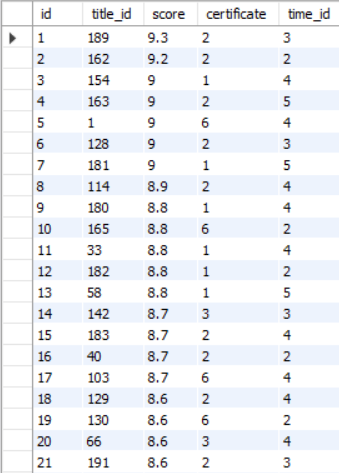


Рисунок 3.1. Завантаження даних у таблицю фактів fact\_imdb250

* 1. За допомогою збереженої процедури create\_fact\_watch\_online було завантажено код фільму, код стримінгової платформи, та кількість переглядів для таблиці fact\_watch\_online, код стрімінгової платформи визначається по цифрі 1 навпроти стрімінгу.

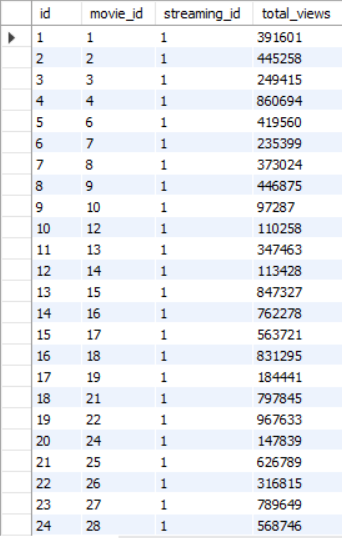


Рисунок 3.2. Завантаження даних у таблицю фактів fact\_watch\_online

* 1. За допомогою збереженої процедури create\_fact\_rotten було завантажено усю інформацію для таблиці fact\_rotten\_tomatoes з сету set\_rotten але замість назви фільму був код зазначений у таблиці dim\_movies.

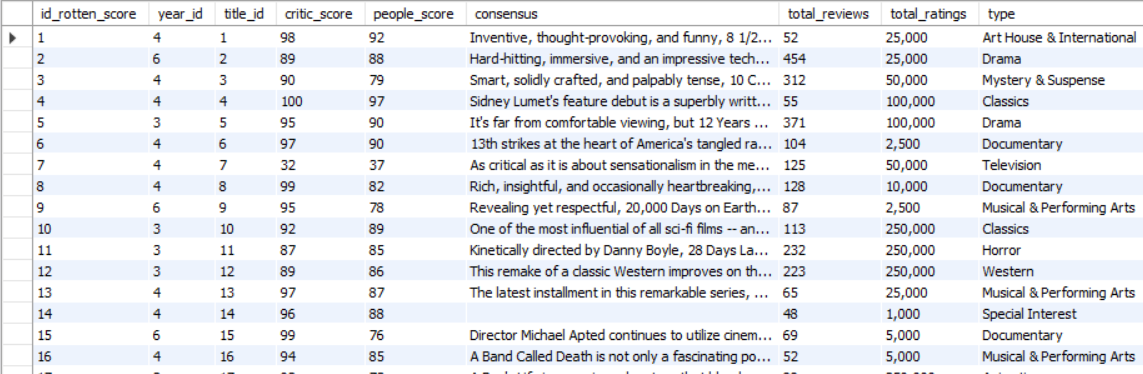


Рисунок 3.3 Завантаження даних у таблицю фактів fact\_rotten\_tomatoes

* 1. За допомогою збереженої процедури create\_dim\_movie було завантажено усю інформацію для таблиці dim\_movie з сету set\_rotten, адаптовано box\_office, duration, genre, language, rating.

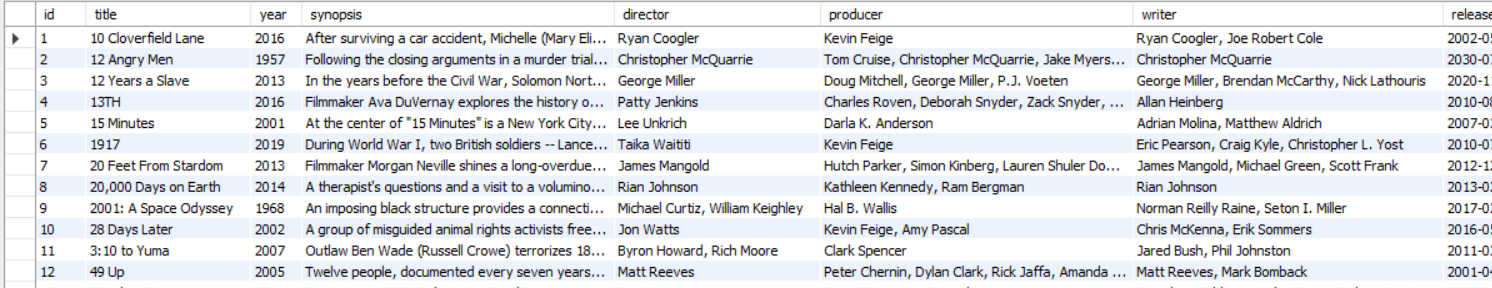


Рисунок 3.4 Завантаження даних у таблицю вимірів dim\_movie

* 1. За допомогою збереженої процедури create\_dim\_genres було завантажено усі жанри до таблиці dim\_genre з сету set\_rotten.

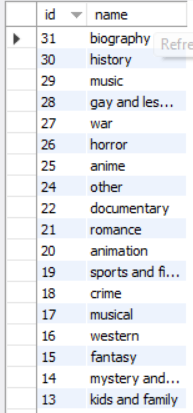


Рисунок 3.5 Завантаження даних у таблицю вимірів dim\_genre

1. Передбачено можливість завантаження змінених та додаткових даних. А саме:
   1. Передбачено додавання стімінгового сервісу до фільму через функцію add\_streaming\_service, після чого запустивши create\_fact\_watch\_online\_table, таблиця оновиться з новими, зміненими данними.

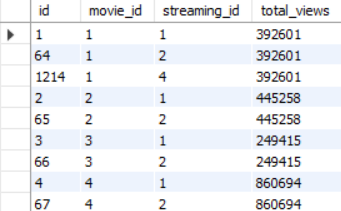


Рисунок 4.1.1 Таблиця до add\_streaming\_service

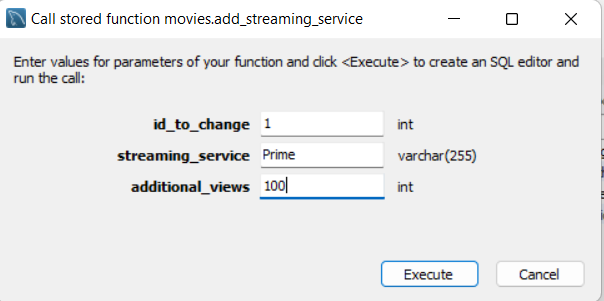


Рисунок 4.1.2 Запуск функції add\_streaming\_service

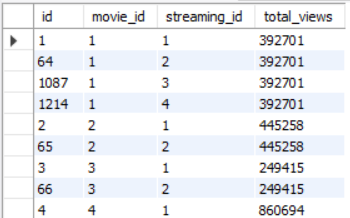


Рисунок 4.1.3 Результат функції add\_streaming\_service

* 1. Передбачено видалення фільму з set\_rotten, для подальшого не використання цього фільму в усіх таблицях та оцінюванні за допомоги функції delete\_movie\_from\_rotten.

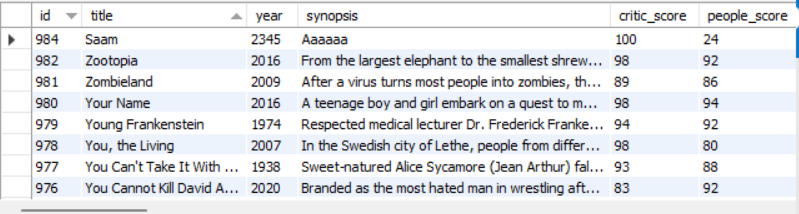


Рисунок 4.2.1 До функції delete\_movie\_from\_rotten.

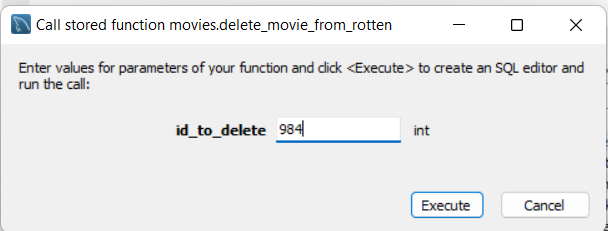


Рисунок 4.2.2 Запуск функції delete\_movie\_from\_rotten.

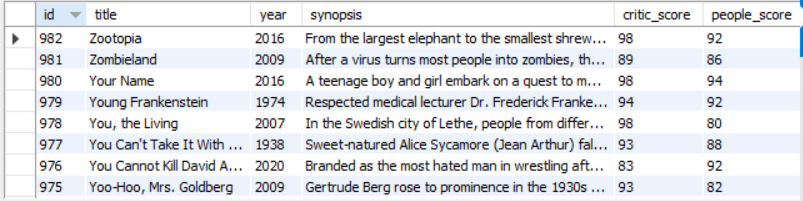


Рисунок 4.2.3 Результат виконання delete\_movie\_from\_rotten.

* 1. Передбачено оновлення даних в set\_imdb250 при зміненні оцінки фільму за допомогою функції update\_set\_imdb250.

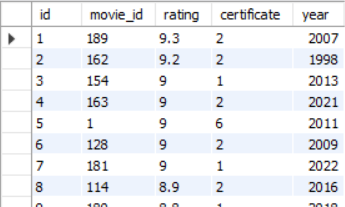


Рисунок 4.3.1 До виконання функції update\_set\_imdb250

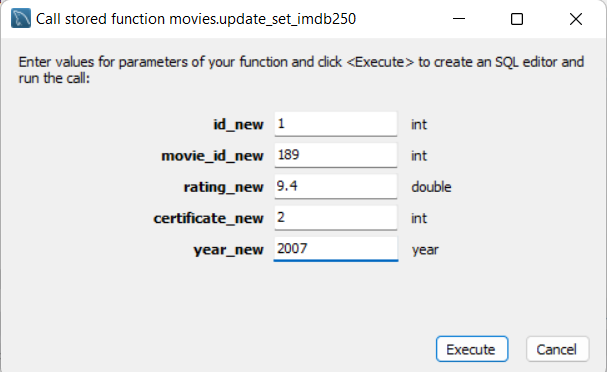


Рисунок 4.3.2 Запуск функції update\_set\_imdb250

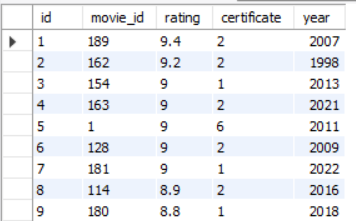


Рисунок 4.3.3 Виконання функції update\_set\_imdb250

**4.ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ**

4.1 Аналіз фільмів зазвичай вимагає використання інтелектуального аналізу даних для виявлення тенденцій та моделей. Це допомагає кінокомпаніям прогнозувати популярність та прибутковість фільмів, а також визначати ті елементи, які найбільше сприймаються глядачами. У процесі аналізу даних використовуються різні методи, включаючи Exponential Smoothing, ARIMA, SARIMAX та Holt. Вибір конкретної моделі залежить від джерела даних та потреб кінокомпанії, інформація з якої може бути використана для підтримки прийняття рішень. Для забезпечення ефективної роботи з даними, обрані моделі дають змогу аналітикам обробляти структурні розриви та зсуви даних, що змінюються з плином часу, а також оперативно змінювати параметри самої моделі. Подальший аналіз даних та порівняння їх ефективності допомагає компаніям створювати більш привабливі продукти для своєї аудиторії.

Вибір був зроблений на основі суттєвих факторів, що полегшують проведення прогнозування. Дані моделі дають змогу аналітикам обробляти структурні розриви та зсуви даних, що змінюються з плином часу, оперативно змінювати параметри самої моделі, що включає в себе коваріацію шумів, матриць переходів та спостереження. Обрані моделі для прогнозування дають високі показники ефективності у роботі з даними з різних джерел в одній моделі. Подальший аналіз даних на кожній з моделей та порівняння їх ефективності буде описано в наступних підрозділах[5].

4.2 Математична модель прогнозування часових проміжків

Для дослідження часових рядів головними чинниками є опис та моделювання їх структури. Побудована модель може використовуватися для екстраполяції або прогнозування часового ряду, ….

Модель ARIMA– це досить поширений клас статистичних моделей для аналізу та …[1].

Модель ARIMA для нестаціонарного часового ряду має вигляд:

де – стаціонарний часовий ряд, – параметри моделі, – оператор різниці часового ряду порядку k.

Модель SARIMAX …

Exponential Smoothing – це модель прогнозування, яка враховує як лінійний тренд у вибірці, так і сезонні варіації у даних за допомогою використання трьох додаткових параметрів. Дана модель описується чотирма формулами:

* експоненціально згладженний ряд
* оцінка тренду
* оцінка сезонності
* прогнозування на p-періодів вперед

де α, β, γ — константи згладжувания для рівня, тренду и сезонності, відповідно; s – тривалість періоду сезонного коливання[8].

Holt - модель прогнозування, … наступним чином:

(t,k)=(t)+k⋅(t),

де (t) – оцінка поточного (t-го) значення ряду; (t) – оцінка поточної зміни.

Далі визначається величина відхилення прогнозованого значення від реального e. При k=1 маємо:

e(t+1)=X(t+1)−(t,1),

(t)= (t−1)+ (t−1)+⋅e(t),

де - коефіцієнт згладжування (адаптації), що змінюється в межах від 0 до 1.

Параметри моделі обчислюються послідовно … на більш ранніх етапах[11].

4.3 Практичне застосування моделей прогнозування часових проміжків

Опираючись на теоретичний матеріал попереднього підпункту, опишемо практичне використання моделей прогнозування часових проміжків. Для початку роботи, оберемо потрібну вибірку даних, що включає в себе кількість гравців та кількість продажів певної гри з плином часу. В якості розробника обраних ігор буде виступати відома компанія … (рисунок 4.1).



Рисунок 4.1 – Вибірка даних для прогнозування

Відобразимо графічно зміну кількості гравців та кількості продажів гри у вигляді часового ряду (рисунки 4.2-4.3).

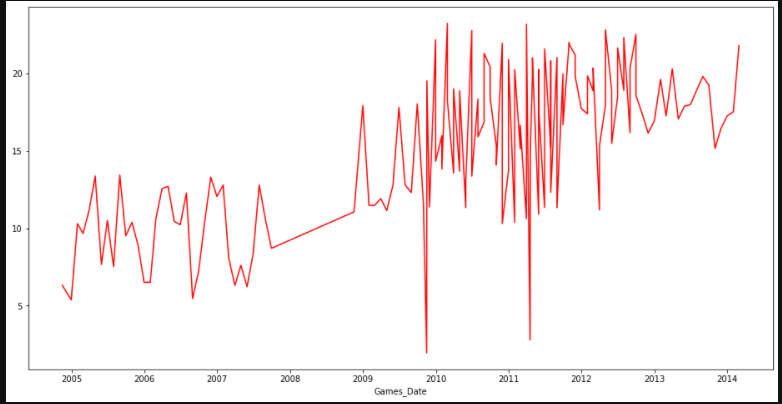


Рисунок 4.2 – Графічне відображення кількості продажів ігор у вигляді часового ряду

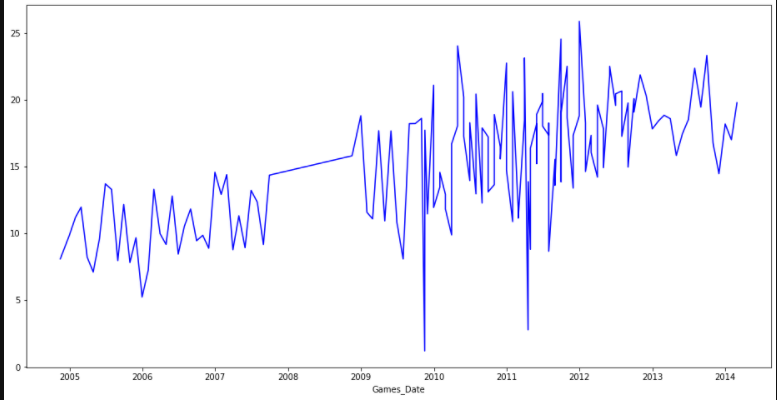


Рисунок 4.3 – Графічне відображення кількості гравців у вигляді часового ряду

Для перевірки на тренд, сезонність та шум застосуємо декомпозицію обраних даних (рисунки 4.4-4.5)

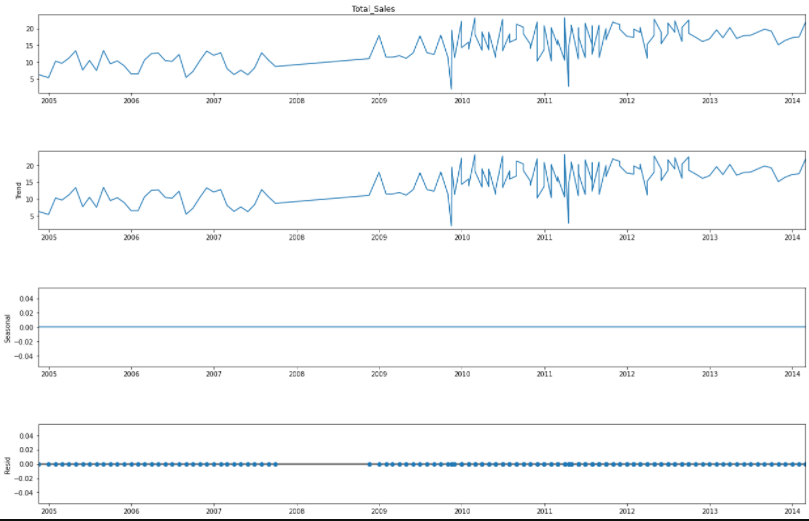


Рисунок 4.4 – Графічне зображення тренду, сезонності та залишкових атрибутів на прикладі даних кількості продажів ігор

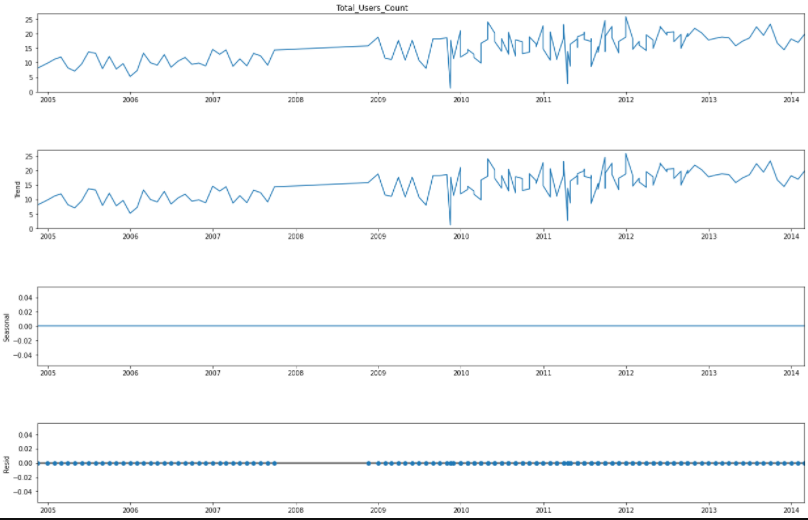


Рисунок 4.5 – Графічне зображення тренду, сезонності та залишкових атрибутів на прикладі даних кількості гравців

Дані графіки наочно демонструють те, що обрані дані є несезонними, а отже це дає змогу сказати, що дані є стаціонарними.

Для перевірки даних на сезонність було реалізовано розширений тест Дікі-Фуллера[10], який підтверджує результати отримані після декомпозиції даних. Отже, дані щодо кількості гравців та продажів ігор є стаціонарними (рисунки 4.6-4.7).

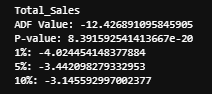


Рисунок 4.6 – Результат розширеного тесту Дікі-Фуллера для перевірки на сезонність кількості продажів ігор

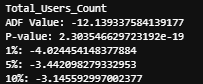


Рисунок 4.7 – Результат розширеного тесту Дікі-Фуллера для перевірки на сезонність кількості гравців

Отримані результати демонструють, що запропоновані у курсовій роботі моделі прогнозування підходитимуть під обрані дані для аналізу[9].

Для зменшення шуму …. В ролі вхідних даних в даній курсовій роботі виступають кількість гравців та кількість продажів ігор (рисунки 4.7 – 4.12).

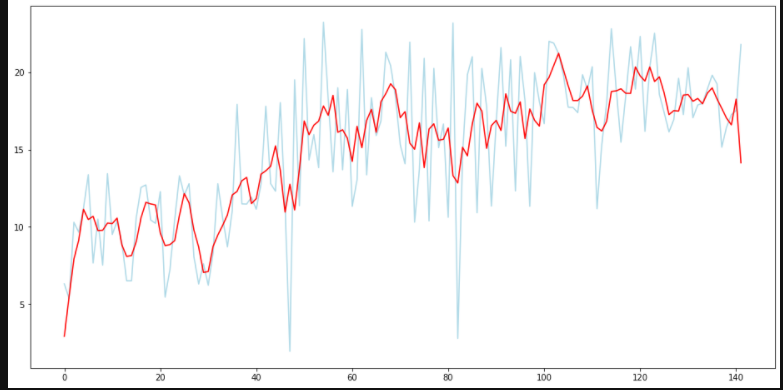


Рисунок 4.8 – Графічне відображення згладжування даних кількості продажів ігор на основі експоненціального згладжування

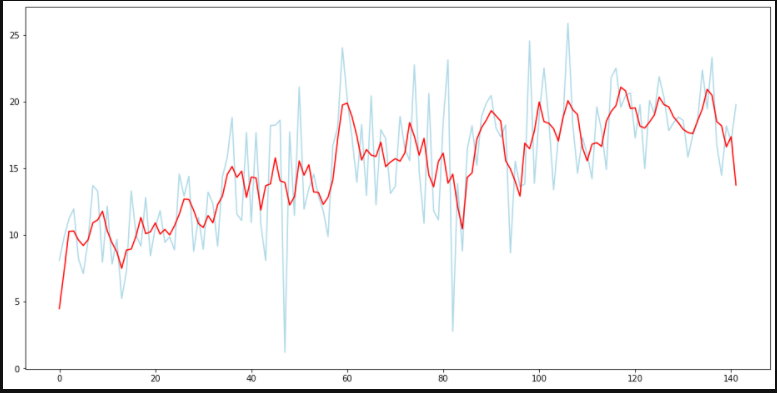


Рисунок 4.9 – Графічне відображення згладжування даних кількості гравців на основі експоненціального згладжування

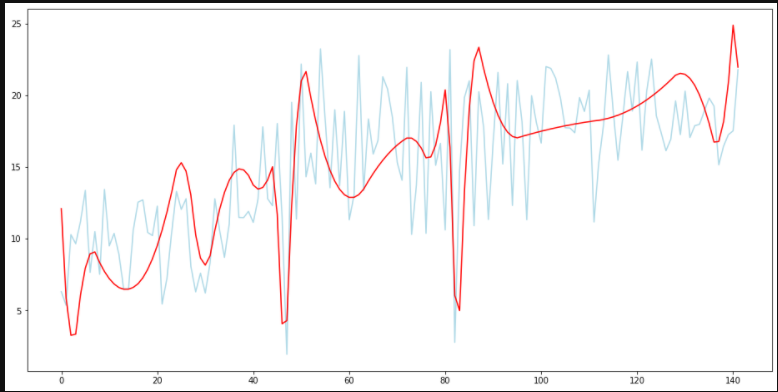


Рисунок 4.10 – Графічне відображення згладжування даних кількості продажів ігор на основі Вейвлет згладжування

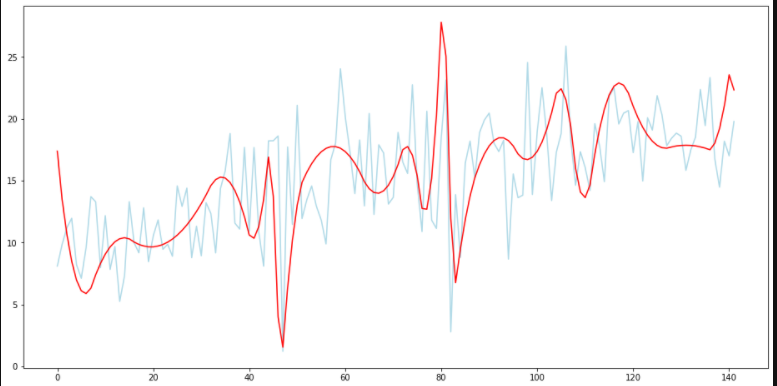


Рисунок 4.11 – Графічне відображення згладжування даних кількості гравців на основі Вейвлет згладжування

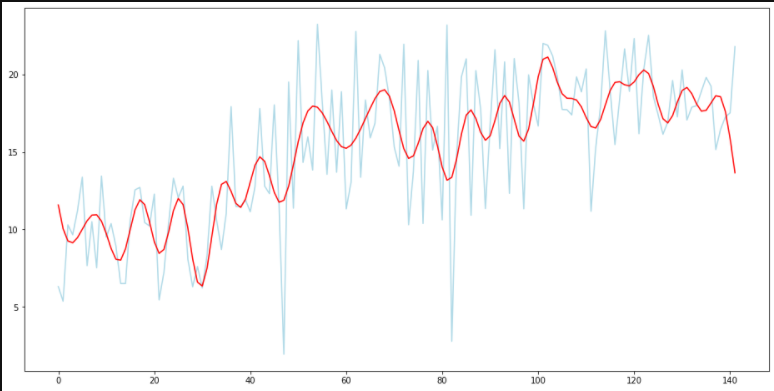


Рисунок 4.12 - Графічне відображення згладжування даних кількості продажів ігор на основі перетворення Фур'є

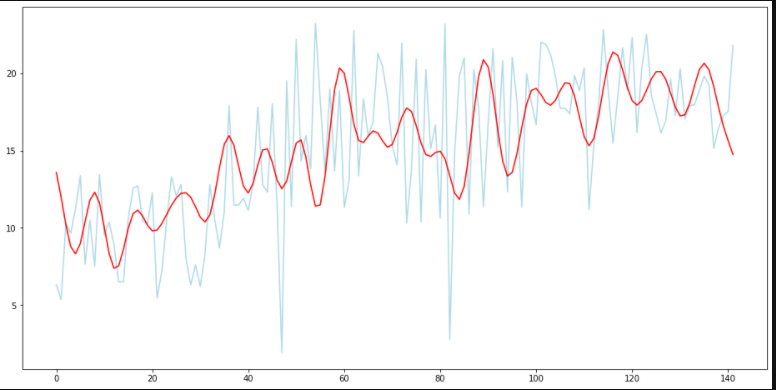


Рисунок 4.13 - Графічне відображення згладжування даних кількості гравців на основі перетворення Фур'є

З вище продемонстрованих графіків можна зробити висновок, що найкраще відтворює траєкторію кривої вхідних даних метод експоненціального згладжування, який і буде використовуватись для подальшого прогнозування даних відеоігор[8].

Перейдемо безпосередньо до графічного представлення реалізації прогнозування на обраних моделях такі як Exponential Smoothing, ARIMA, SARIMAX, Holt. На рисунках чорним кольором позначено вибірку тренувальних даних, а червоним та чорним – дані вибірки та прогнозовані дані відповідно.

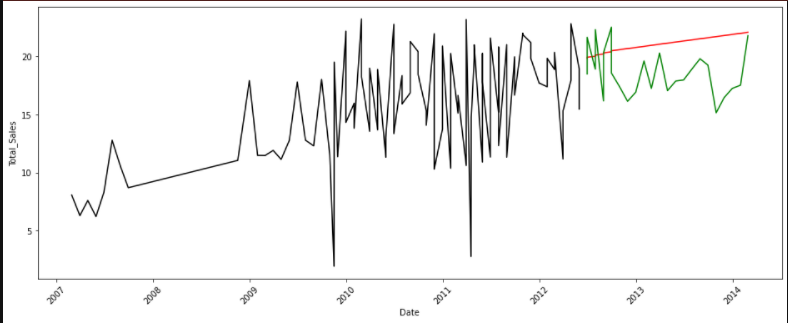


Рисунок 4.14 – Прогнозування вхідних даних кількості продажів ігор за моделлю Holt

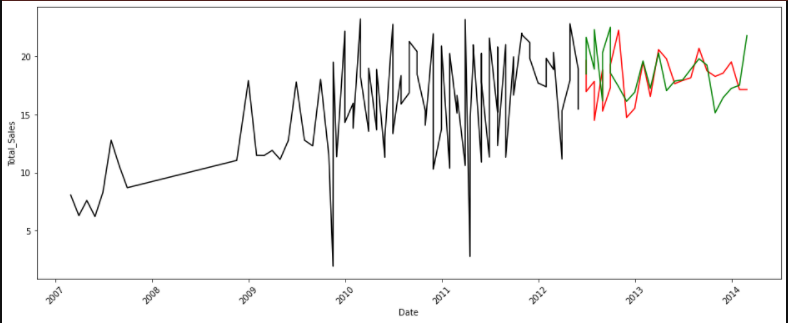


Рисунок 4.15 – Прогнозування вхідних даних кількості продажів ігор за моделлю Exponential Smoothing

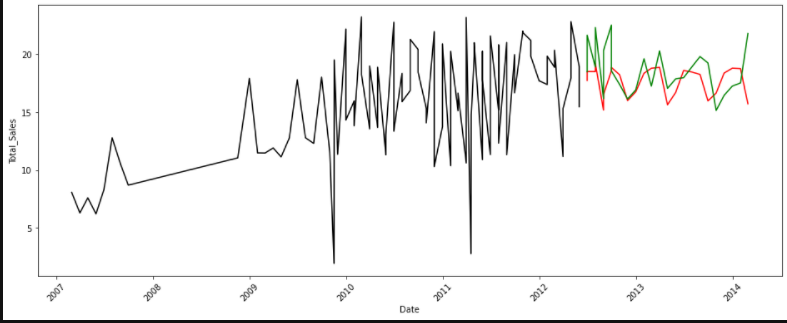


Рисунок 4.16 – Прогнозування вхідних даних кількості продажів ігор за моделлю SARIMAX

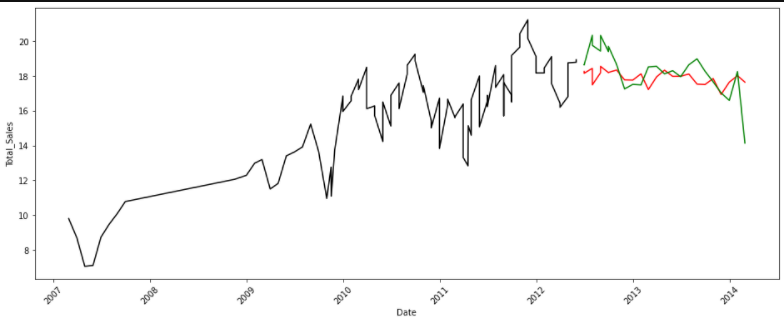


Рисунок 4.17 – Прогнозування згладжувальних даних кількості продажів ігор за моделлю SARIMAX

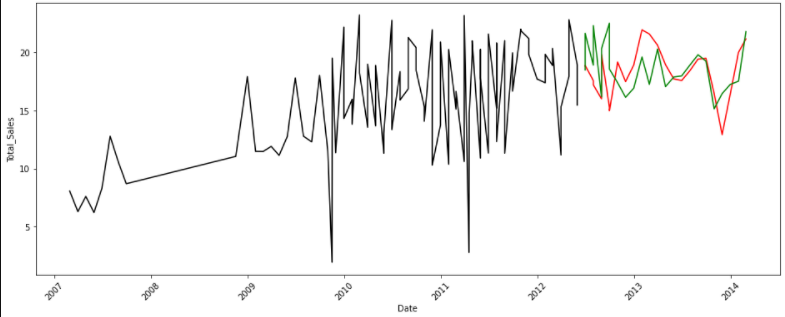


Рисунок 4.18 – Прогнозування вхідних даних кількості продажів ігор за моделлю ARIMA

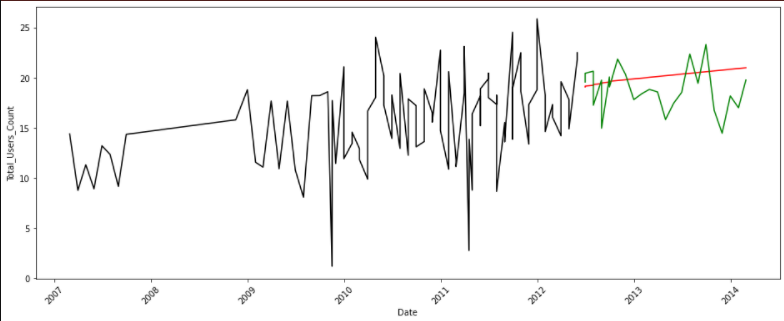


Рисунок 4.19 – Прогнозування вхідних даних кількості гравців за моделлю Holt

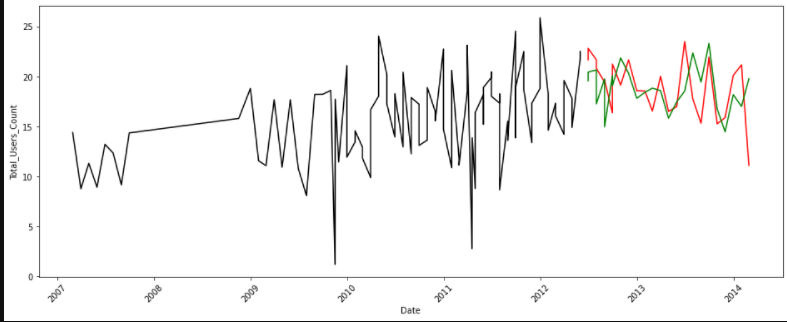


Рисунок 4.20 – Прогнозування вхідних даних кількості гравців за моделлю Exponential Smoothing

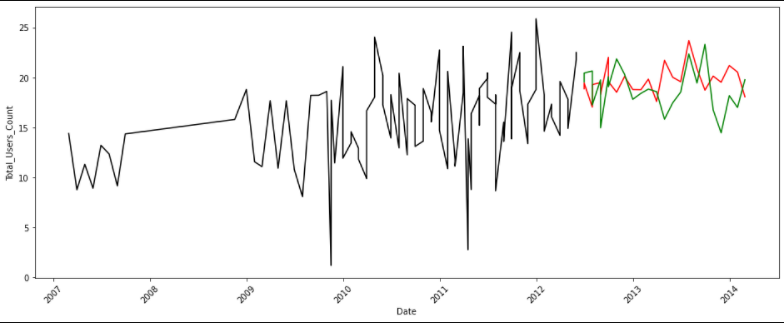


Рисунок 4.21 – Прогнозування вхідних даних кількості гравців за моделлю SARIMAX

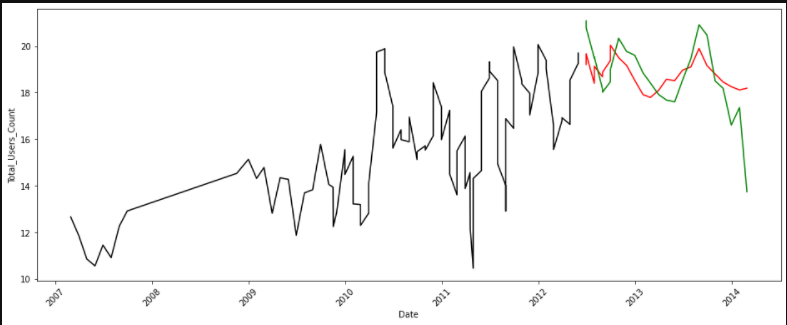


Рисунок 4.22 – Прогнозування згладжувальних даних кількості гравців за моделлю SARIMAX

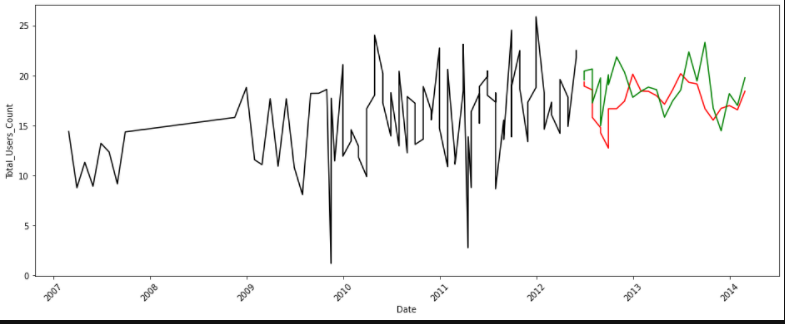


Рисунок 4.23 – Прогнозування вхідних даних кількості гравців за моделлю ARIMA

Для моделі SARIMAX виведемо діагностику для прогнозування кількості гравців та продажу ігор, що демонструє наступне (рисунки 4.24-4.25):

* Standardized residual for „T” … яка не пояснюється в моделі[7].

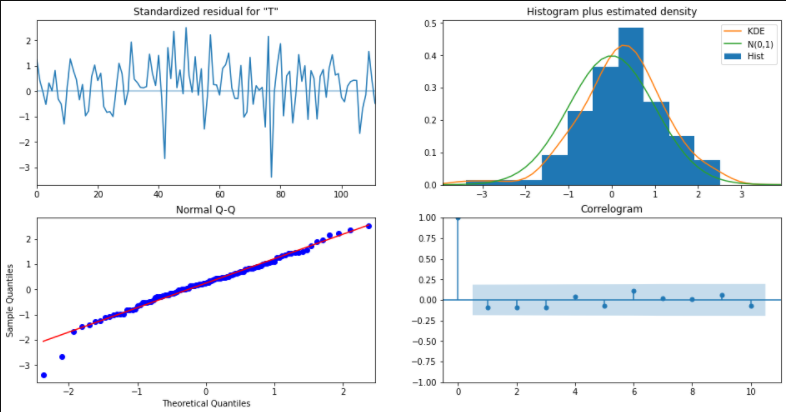


Рисунок 4.24 – Вивід діагностики для моделі SARIMAX на основі даних продажів ігор

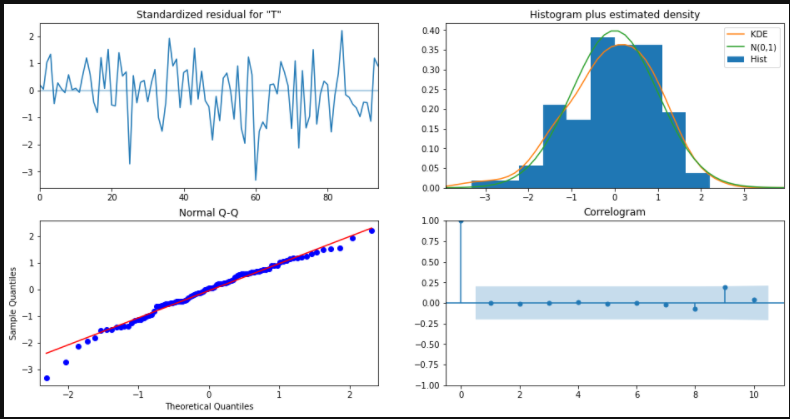


Рисунок 4.25 – Вивід діагностики для моделі SARIMAX на основі даних кількості гравців

Виведемо графіки автокореляції та часткової автокореляції для обраних у курсовій роботі даних для моделі SARIMAX(рисунки 4.26 - 4.27).

У цьому випадку автокореляція демонструє спостереження часового проміжку зі значенням того з цього ж проміжку в попередній інтервал часу. Часткова автокореляція дає змогу нам отримати взаємозв’язок між спостереженням в часовому проміжку зі спостереженнями на попередніх часових проміжках без урахування взаємозв’язків між проміжними спостереженнями[6].

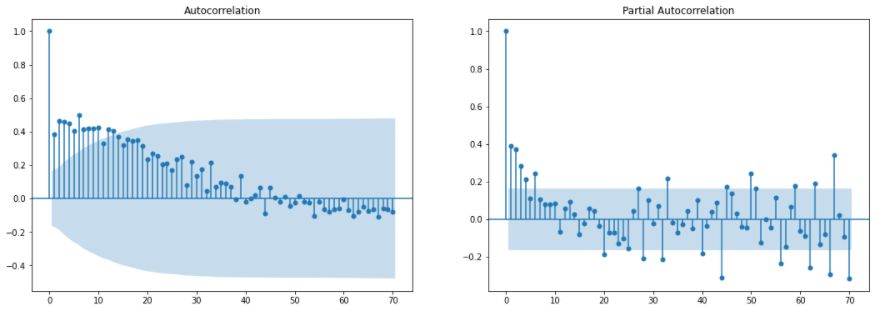


Рисунок 4.26 – Графіки автокореляції та часткової автокореляції на вибірці даних продажів ігор для моделі SARIMAX

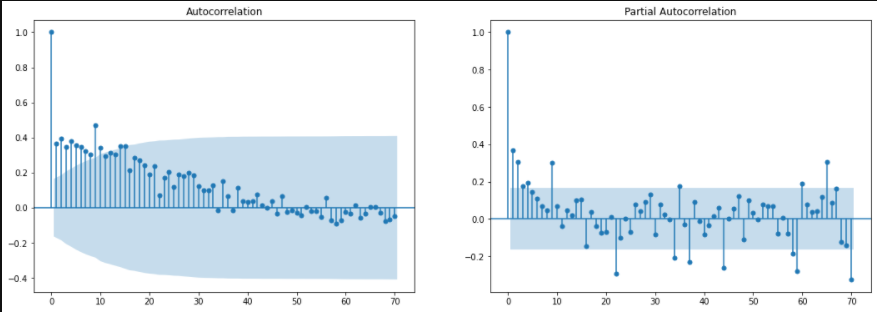


Рисунок 4.27 – Графіки автокореляції та часткової автокореляції на вибірці даних кількості гравців для моделі SARIMAX

…..

Враховуючи вищеописану реалізацію можна зробити висновки, проаналізувавши дані та вибрати найоптимальніший метод.

4.4 Порівняння ефективності методів інтелектуального аналізу

Аналізуючи прибутки з продажів ігор чи кількості гравців у певний момент часу важливо отримувати найбільш точний результат, щоб передбачати майбутні результати. Для аналітичного визначення точності результатів, отриманих на основі моделей для прогнозування, та вибору найоптимальнішого методу знайдемо середньоквадратичні помилки MSE для кожної з моделей вибірки даних кількості продажів ігор та кількості гравців.

Для кількості продажів ігор отримано наступні результати:

* Модель Exponential Smoothing – MSE = 8.952499933006333
* Модель ARIMA – MSE = 6.406898744038264
* Модель SARIMAX на вихідних даних – MSE = 4.864811488884299
* Модель SARIMAX на згладжувальних даних –
* …

З отриманих результатів середньоквадратичної помилки, а також враховуючи рисунки 4.14 - 4.23, можна зробити висновок, що найоптимальнішим методом для прогнозування що кількості продажів ігор, що кількості гравців є метод, реалізований на основі моделі …. Варто зазначити, що після проведення згладжування даних,….

**ВИСНОВКИ**

В результаті виконання курсової роботи було розроблено сховище даних відеоігор типу „сніжинка”, що містить дві таблиці фактів та вісім таблиць вимірів. Реалізовано ETL процеси для завантаження даних, …. Для реалізації поставленої задачі було використано PostgreSQL версії 12.1, мова програмування ….

На основі детального опису та проведеного аналізу предметної області інтелектуального аналізу даних для прогнозування кількості продажів ігор та факторів впливу на популярність відеоігор було отримано результати аналізу вибірки даних на сезонність, стаціонарність, шум за допомогою доповненого тесту …. Підтвердженням даних висновків є результати середньоквадратичних помилок, а для найкращого результату …. Результати досліджень показують, що …

Отже, поставлені задачі були виконані, а також планується розширення функціоналу для взаємодії з сховищем даних відеоігор з використанням більшої кількості факторів, які ….

**ПЕРЕЛІК ПОСИЛАНЬ**

1. Бокс Дж., Дженкинс Г.М. Анализ временных рядов. Прогноз и управление. –– Мир, 1974.
2. Scott Lundberg and Su-In Lee, “A Unified Approach to Interpreting Model Predictions”, 2017
3. Коровкин С. Д., Левенец И. А., Ратманова И. Д., Старых В. А., Щавелёв Л. В. Решение проблемы комплексного оперативного анализа информации хранилищ данных. // СУБД. – 1997. – № 5-6. – 47-51с.
4. Lloyd Shapley, “A value of n-person games. Contributions to the Theory of Games”, 1953
5. Tobias Lang and Matthias Rettenmeier, “Understanding Consumer Behavior with Recurrent Neural Networks”, 2017
6. Wikipedia. ARIMA // Википедия, свободная энциклопедия. –– 2014. –– URL: http://goo.gl/bNmcX8.
7. Wikipedia. Модель авторегрессии скользящего среднего // Википедия, свободная энциклопедия. –– 2014. –– URL: http://goo.gl/TSHFst.
8. Альперович М. Введение в OLAP и многомерные базы данных. – М.: Вильямс. 2005. – 267с.
9. Jerome Friedman, “Greedy function approximation: A gradient boosting machine”, 2001
10. NeuroProject. Методы прогнозирования. –– 2014. –– URL: http://goo.gl/9Jn0UV.
11. Tony Duan, Anand Avati, Daisy Ding, Khanh Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler, “NGBoost: Natural Gradient Boosting for Probabilistic Prediction”, 2020

**Додаток А Тексти програмного коду**

*студента групи ХХХ ІІ курсу*

*ХХХ*

(Обсяг програми (документа), арк., Кб)

*28 арк, 64 Кб*

(Вид носія даних)

*SSD*

(Найменування програми (документа))

*Тексти програмного коду* прогнозування доходів та факторів на популярність індустрії відеоігор

// code goes here