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 гоночної серії FORMULA 1»

Студентів 2 курсу групи ІП-21

Спеціальності: 121

«Інженерія програмного забезпечення»

Загребельний О.А.

Голованьов Г.О.

«ПРИЙНЯВ» з оцінкою

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

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

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

Підпис                    Дата

Київ - 2024 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-21 Семестр 4

**ЗАВДАННЯ**

**на курсову роботу студента**

|  |
| --- |
| Загребельного О.А., Голованьова Г.О. |

|  |  |
| --- | --- |
| 1.Тема роботи | Аналіз гоночної серії FORMULA 1 |
|  | |
|  | |

|  |  |
| --- | --- |
| 2.Строк здачі студентом закінченої роботи | 29.05.2024 |

|  |  |
| --- | --- |
| 3. Вхідні дані до роботи | методичні вказівки до курсової роботи, обрані дані з сайтів |
| https://github.com/f1db/f1db | |
| https://docs.fastf1.dev/ | |

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 | 30.03.2024 |

**КАЛЕНДАРНИЙ ПЛАН**

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Назва етапів курсової роботи | Термін виконання етапів роботи | Підписи керівника, студента |
| 1. | Отримання теми курсової роботи | 30.03.2024 |  |
| 2. | Визначення зовнішніх джерел даних | 06.04.2024 |  |
| 3. | Пошук та вивчення літератури з питань курсової роботи | 13.04.2024 |  |
| 4. | Розробка моделі сховища даних | 20.04.2024 |  |
| 4. | Розробка ETL процесів | 27.04.2024 |  |
| 5. | Обґрунтування методів інтелектуального аналізу даних | 01.05.2024 |  |
| 6. | Застосування та порівняння ефективності методів інтелектуального аналізу даних | 15.05.2024 |  |
| 7. | Підготовка пояснювальної записки | 22.05.2024 |  |
| 8. | Здача курсової роботи на перевірку | 29.05.2024 |  |
| 9. | Захист курсової роботи | 03.06.2024 |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Студент |  |  | Загребельний О.А. |
|  | (підпис) |  | (прізвище, ім’я, по батькові) |
| Студент | |  |  | Голованьов Г.О. |
|  | | (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) |

"29" травня 2024 р.

**АНОТАЦІЯ**

Пояснювальна записка до курсової роботи: \_\_ сторінок, \_\_ рисунки, \_\_ таблиць, \_\_ 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**

У сучасному світі автоспорт, зокрема серія перегонів Formula 1, є одним із найпопулярніших та найпрестижніших видів спорту, що привертає увагу мільйонів глядачів по всьому світу. Формула 1 не лише захоплює своїми динамічними гонками та високотехнологічними автомобілями, але й приносить величезні прибутки та стимулює розвиток інновацій у різних галузях. Щорічно серія перегонів Formula 1 привертає нових вболівальників та встановлює нові рекорди за відвідуваністю та переглядами.

У рамках даної курсової роботи розроблено сховище даних, яке містить інформацію про результати перегонів Formula 1, базуючись на фізичній моделі бази даних. Основу сховища було створено за допомогою SQL скриптів, а робота з базою даних реалізована через програмне забезпечення для імплементації ETL процесів та інтелектуального аналізу даних.

Як система керування сховищем даних для даної роботи використовується PostgreSQL, а мова програмування для реалізації застосунку - Python3. Використовуючи ці інструменти, було розроблено програмне забезпечення, яке дозволяє отримувати дані зі сховища, проводити їх аналіз, прогнозувати результати перегонів та графічно відображати отримані результати.

Ця робота спрямована на дослідження та аналіз даних перегонів Formula 1, що дозволить не лише зрозуміти поточні тенденції, але й зробити прогнози щодо майбутніх перегонів, використовуючи різні моделі прогнозування, такі як K-nearest neighbours та Random Forest.

**1.Постановка задачі**

Під час виконання курсової роботи необхідно виконати комплекс завдань, спрямованих на створення ефективного інструменту для аналізу та прогнозування результатів перегонів Formula 1. Насамперед, потрібно створити сховище даних типу «сніжинка», яке міститиме щонайменше шість таблиць вимірів та одну таблицю фактів.

Наступним кроком є створення ETL процесів для завантаження даних до сховища, їх отримання зі сховища за допомогою запитів, а також оновлення та додавання даних до таблиць.

Реалізація спроектованого сховища даних буде здійснюватися з використанням PostgreSQL, що забезпечить високу продуктивність та надійність системи. Для подальшого аналізу даних необхідно створити програмне забезпечення, яке дозволить отримувати вибірки даних зі створеного сховища, графічно відображати їх, проводити інтелектуальний аналіз та здійснювати передбачення за допомогою різних моделей прогнозування.

Аналіз вибірки даних передбачає перевірку на сезонність, стаціонарність та шум за допомогою доповненого тесту Дікі-Фуллера, а також декомпозицію даних. Для покращення якості прогнозів необхідно провести згладжування даних методами ковзного середнього, вейвлет-згладжування та швидкого згладжування Фур'є.

У роботі будуть використовуватись моделі прогнозування, такі як k-nearest neighbours та random forest, для аналізу даних. Важливо застосувати найоптимальнішу модель прогнозування на згладжених даних та порівняти отримані результати з прогнозуванням на початкових даних.

Додатково, необхідно провести автокореляцію та часткову автокореляцію для виявлення повторюваних ділянок даних, незважаючи на наявність шумів. Аналіз отриманих результатів та порівняння різних методів прогнозування дозволить визначити найоптимальніший метод. Важливим аспектом є знаходження довірчих інтервалів з використанням найкращої моделі прогнозування.

Для реалізації застосунку буде використовуватись мова програмування Python 3. Весь проект повинен бути завершений до дедлайну, початку сесії, з дотриманням єдиного стилю написання коду (coding style), що забезпечить його читабельність та підтримуваність у майбутньому.

**2.АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ**

Сьогодні автоспорт, зокрема серія перегонів Formula 1, є одним із найпопулярніших видів спорту у світі, який постійно привертає увагу мільйонів глядачів. Formula 1, або "Королева автоспорту", не тільки захоплює своїми видовищними гонками та інноваційними технологіями, але й є значущою складовою світової економіки, генеруючи величезні доходи від трансляцій, спонсорства та продажу квитків. Гонки Formula 1 не лише пропонують захоплююче видовище, але й виступають платформою для впровадження новітніх технологій та інженерних рішень.

Аналіз результатів перегонів Formula 1 є складним завданням, яке включає багато факторів, таких як погодні умови, характеристики трас, технічний стан автомобілів та майстерність пілотів. Для команд та аналітиків важливо виявляти закономірності, що впливають на результати перегонів, аби оптимізувати стратегії, покращувати результати та передбачати майбутні успіхи. Вміння прогнозувати результати перегонів на основі історичних даних дозволяє командам краще готуватися до майбутніх змагань та приймати обґрунтовані рішення щодо розвитку техніки і тактики.

У рамках даної курсової роботи буде розроблено програмне забезпечення для аналізу та прогнозування результатів перегонів Formula 1. Функціональність цього програмного забезпечення включає:

* проектування сховища даних;
* створення ETL процесів для завантаження та оновлення даних;
* створення вибірки даних зі сховища;
* інтелектуальний аналіз даних;
* використання декількох моделей прогнозування даних;
* прогнозування результатів перегонів;
* визначення факторів, що впливають на результати перегонів;
* графічне відображення отриманих результатів та їх аналіз.

Сховище даних дозволить зберігати великі обсяги історичних даних про результати перегонів, що забезпечить основу для детального аналізу та прогнозування. ETL процеси забезпечать актуальність та повноту даних, необхідних для аналітичної роботи. Інтелектуальний аналіз даних, включаючи перевірку на сезонність, стаціонарність та шум, а також згладжування даних, дозволить виявити приховані тенденції та шаблони. Використання моделей прогнозування, таких як k-nearest neighbours та random forest, дозволить робити точні передбачення та оцінювати вплив різних факторів на результати перегонів. Графічне відображення результатів допоможе візуалізувати дані та зрозуміти основні висновки аналізу.

**3.РОЗРОБКА СХОВИЩА ДАНИХ**

3.1 Розробка моделі сховища даних

Для виконання курсової роботи було обрано 5 джерел відкритих даних на сайті https://www.kaggle.com//, що включають в себе:

* Загальна інформація про ігри: https://www.kaggle.com/nikdavis/steam-store-games
* Відгуки та популярність ігор: https://www.kaggle.com/floval/12-000-video-game-reviews-from-vandal
* Рейтинг ігор за регіонами: https://www.kaggle.com/juttugarakesh/video-game-data
* Платформи ігор: https://www.kaggle.com/skateddu/metacritic-all-time-games-stats
* Продаж ігор за регіонами: https://www.kaggle.com/gregorut/videogamesales

На основі детального опису та проведеного аналізу предметної області було розроблено наступну модель сховища даних відеоігор за типом „сніжинка” (рисунок 3.1).

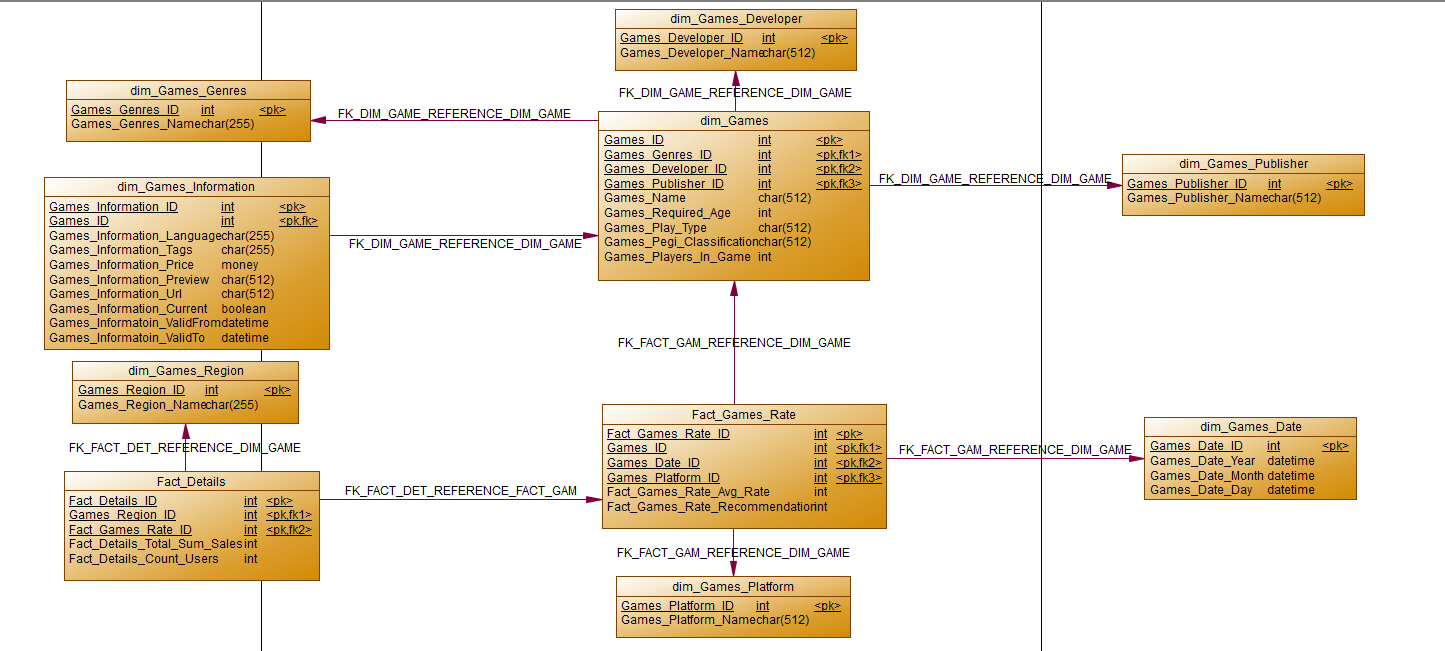


Рисунок 3.1 - Проектування моделі сховища за типом сніжинка

У моделі сховища за типом сніжинка спроектовано дві таблиці фактів та вісім таблиць вимірів. На даній моделі видно, що центральною таблицею вимірів є таблиця dim\_Games, яка відповідає за самі відеоігри. …. Нижче наведено опис полів кожної таблиці сховища даних відеоігор (таблиці 3.1-3.10).

Таблиця 3.1 – Таблиця виміру розробників ігор dim\_Games\_Developer

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| 1.Games\_Developer\_ID | Унікальний ідентифікатор запису розробника гри |
| 2.Games\_Developer\_Name | Назва розробника гри |

На основі спроектованої моделі було створено сховище даних відеоігор, що включає сім таблиць вимірів та дві таблиці фактів реалізованих за допомогою нижче наведених скриптів використовуючи PostgreSQL версії 12.1.

Створення таблиці виміру розробників ігор dim\_Games\_Developer в сховищі даних:

CREATE TABLE "dim\_Games\_Developer" (

"Games\_Developer\_ID" SERIAL PRIMARY KEY,

"Games\_Developer\_Name" VARCHAR(512)

);

….

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 таблиці виміру dim\_Games\_Date було розроблено функцію parseDate(x) розбиття дати випуску гри на окремі поля : рік, місяць та день випуску гри.
   2. …
2. Було отримано дані з Stage зони для заповнення сховища даних. А саме:
   1. За допомогою функції get\_dim\_games\_date(metacritic) було отримано дані року, місяця та дня випуску гри з датасету metacritic для таблиці дати випуску ігор dim\_Games\_Date.
   2. ….
3. Було створено функції завантаження даних у сховище даних. А саме:
   1. За допомогою функції initial\_loading() було завантажено дані року, місяця та дня випуску гри у створену таблицю виміру дати випуску ігор dim\_Games\_Date (рисунок 3.2).

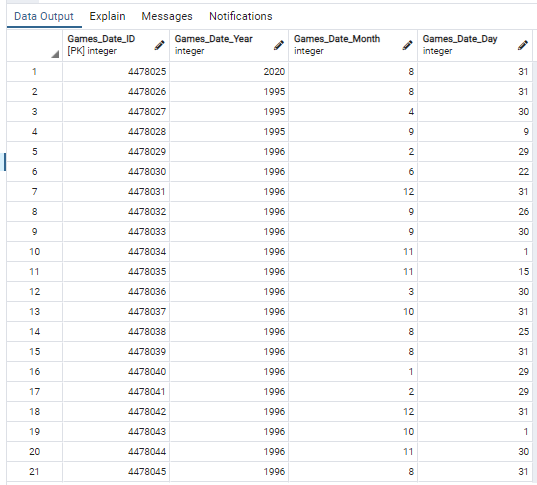


Рисунок 3.2 - Завантаження даних у таблицю вимірів dim\_Games\_Date

* 1. За допомогою функції initial\_loading() було завантажено дані назви видавців ігор …
  2. За допомогою функції upload\_one\_game(game\_name) було завантажено дані про мову випуску гри, теги, що стосуються гри, ціна на гру, короткий опис гри, посилання на гру, у створену таблицю виміру інформації про ігри dim\_Games\_Information (рисунок 3.9).

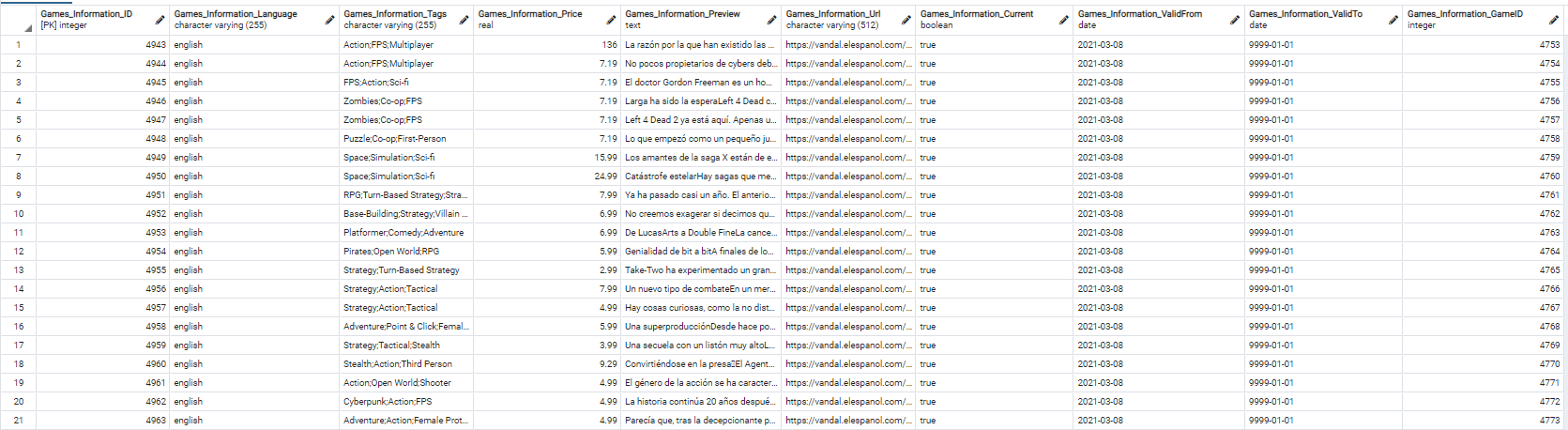


Рисунок 3.9 - Завантаження даних у таблицю вимірів dim\_Games\_Information

….

**4.ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ**

4.1 Обґрунтування вибору методів інтелектуального аналізу даних

Для обробки інформації та виявлення в ній моделей та тенденцій для прийняття рішень необхідно використовувати інтелектуальний аналіз даних. В даній курсовій роботі для передбачення аналізу популярності ігор та доходу компаній було обрано методи прогнозування даних. Спираючись на результати попередніх даних ігор за допомогою математичних методів, можна оцінити майбутні показники, що є актуальним для багатьох інтернет-магазинів відеоігор для виявлення закономірностей. В якості моделей для проведення прогнозування даних відеоігор було обрано Exponential Smoothing, ARIMA, SARIMAX, Holt.

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).

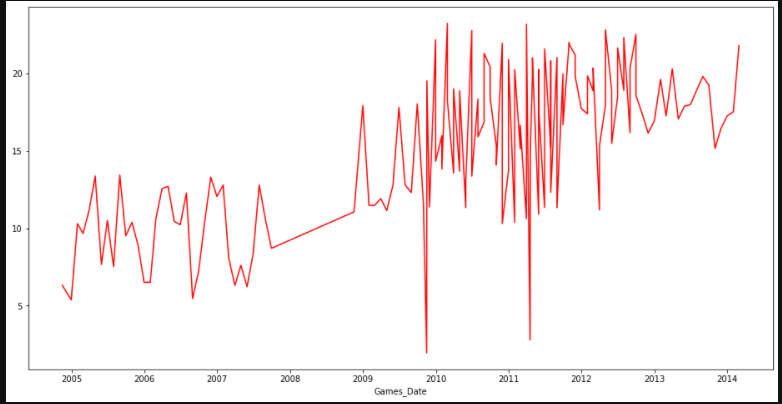


Рисунок 4.2 – Графічне відображення кількості продажів ігор у вигляді часового ряду

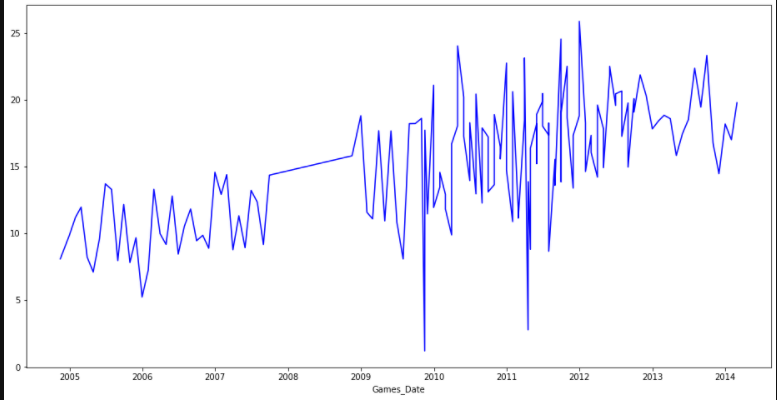


Рисунок 4.3 – Графічне відображення кількості гравців у вигляді часового ряду

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

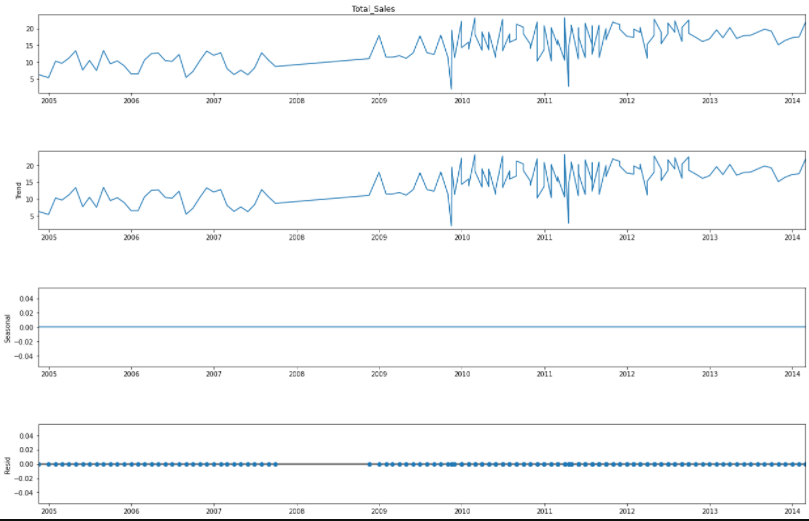


Рисунок 4.4 – Графічне зображення тренду, сезонності та залишкових атрибутів на прикладі даних кількості продажів ігор

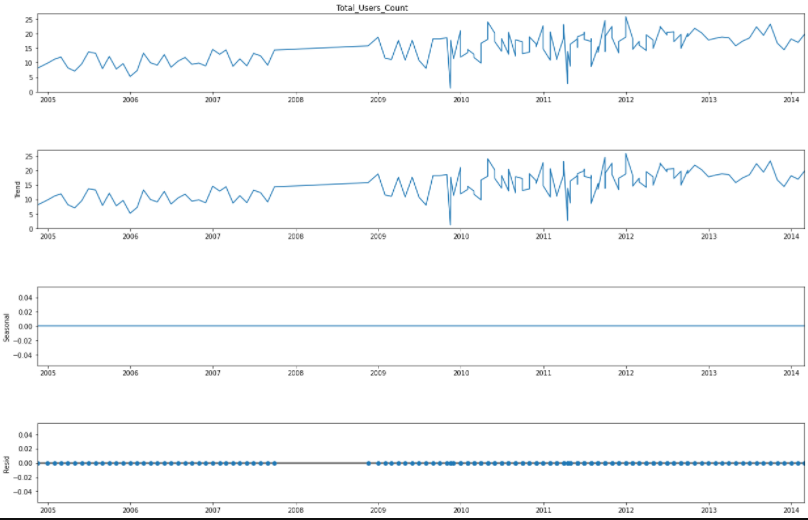


Рисунок 4.5 – Графічне зображення тренду, сезонності та залишкових атрибутів на прикладі даних кількості гравців

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).

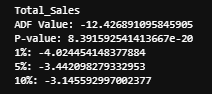


Рисунок 4.6 – Результат розширеного тесту Дікі-Фуллера для перевірки на сезонність кількості продажів ігор

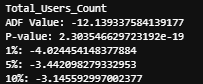


Рисунок 4.7 – Результат розширеного тесту Дікі-Фуллера для перевірки на сезонність кількості гравців

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).

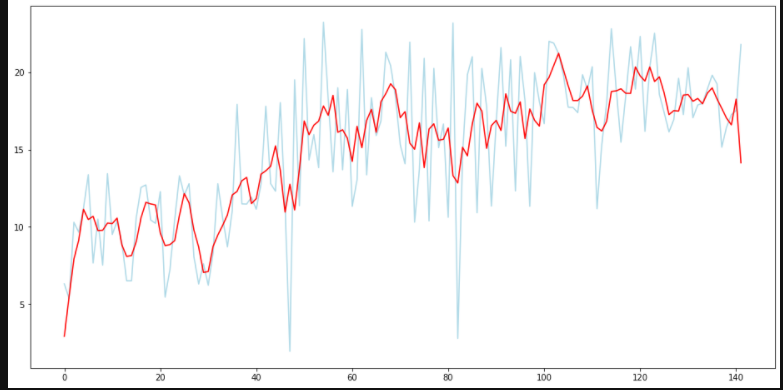


Рисунок 4.8 – Графічне відображення згладжування даних кількості продажів ігор на основі експоненціального згладжування

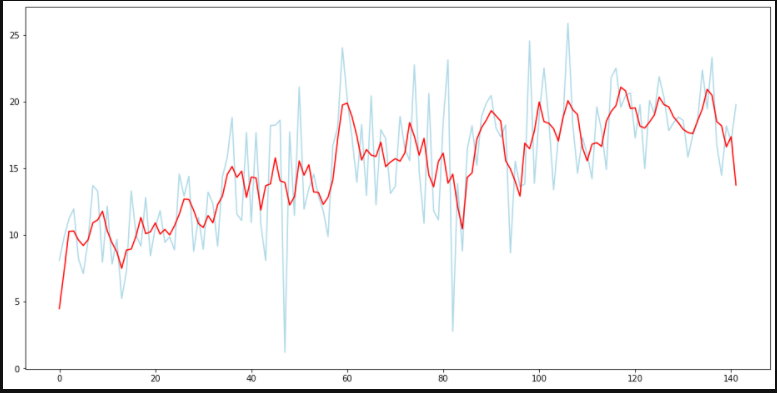


Рисунок 4.9 – Графічне відображення згладжування даних кількості гравців на основі експоненціального згладжування

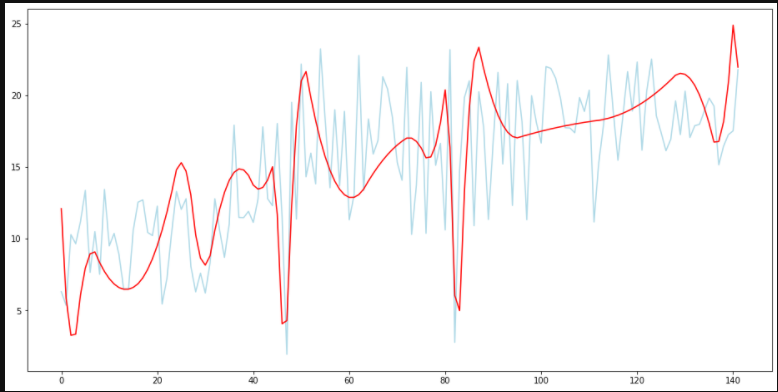


Рисунок 4.10 – Графічне відображення згладжування даних кількості продажів ігор на основі Вейвлет згладжування

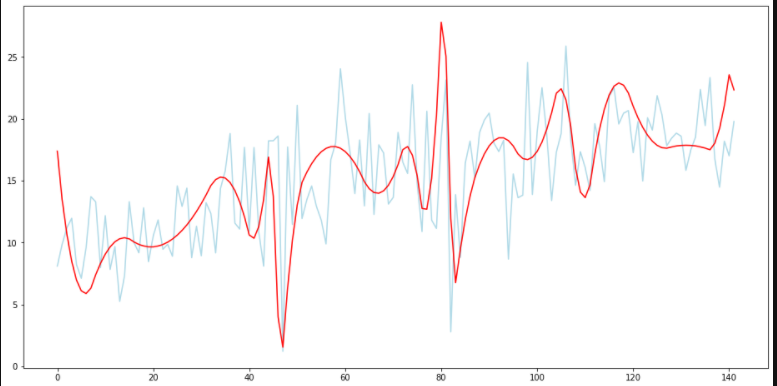


Рисунок 4.11 – Графічне відображення згладжування даних кількості гравців на основі Вейвлет згладжування

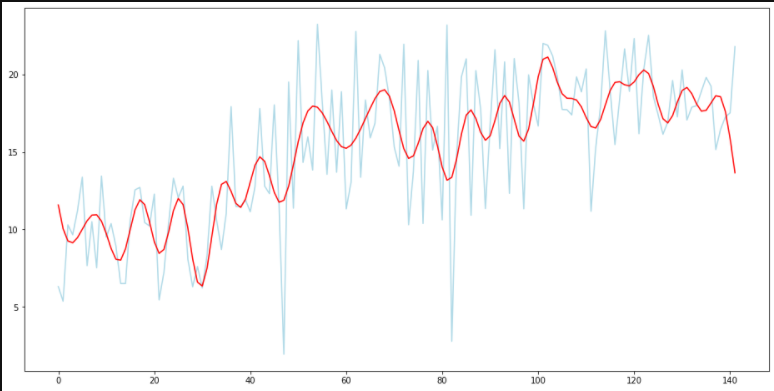


Рисунок 4.12 - Графічне відображення згладжування даних кількості продажів ігор на основі перетворення Фур'є

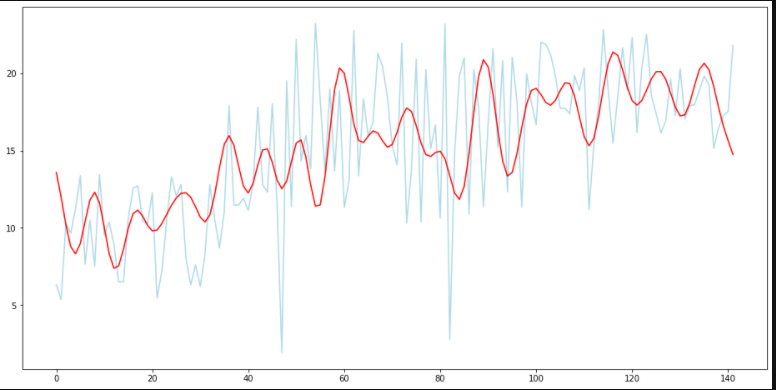


Рисунок 4.13 - Графічне відображення згладжування даних кількості гравців на основі перетворення Фур'є

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.

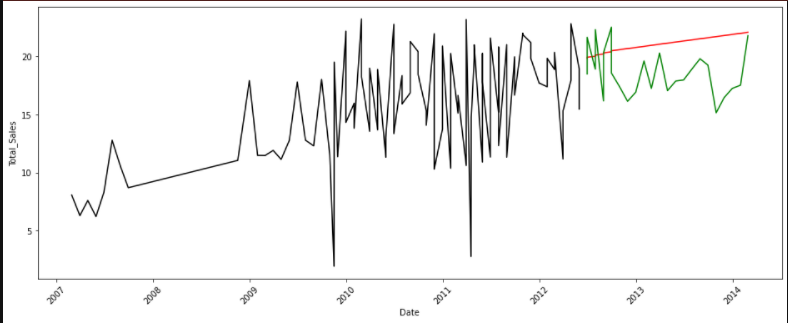


Рисунок 4.14 – Прогнозування вхідних даних кількості продажів ігор за моделлю Holt

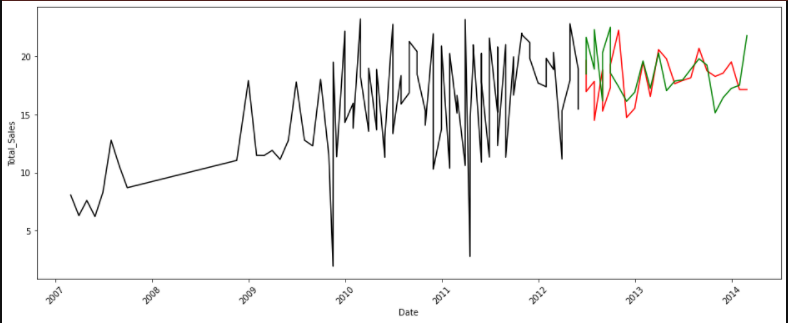


Рисунок 4.15 – Прогнозування вхідних даних кількості продажів ігор за моделлю Exponential Smoothing

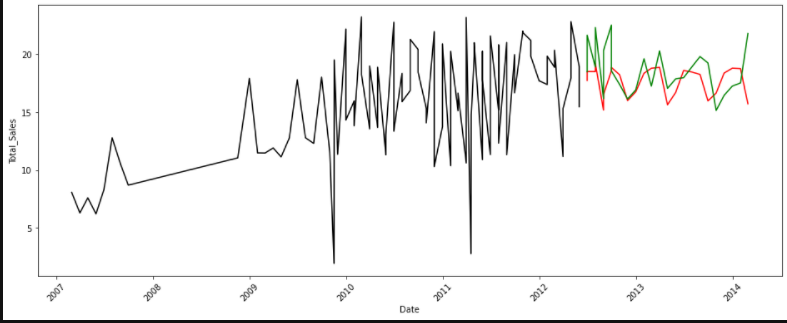


Рисунок 4.16 – Прогнозування вхідних даних кількості продажів ігор за моделлю SARIMAX

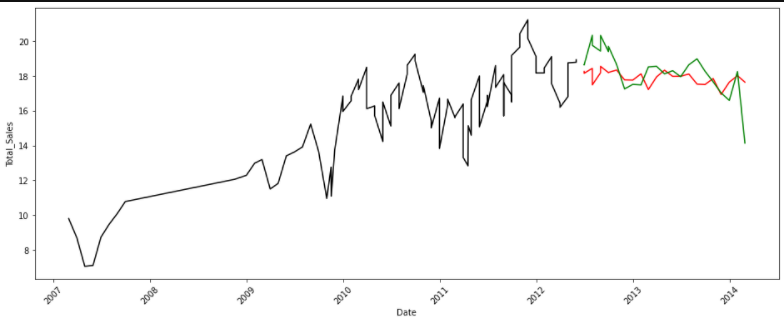


Рисунок 4.17 – Прогнозування згладжувальних даних кількості продажів ігор за моделлю SARIMAX

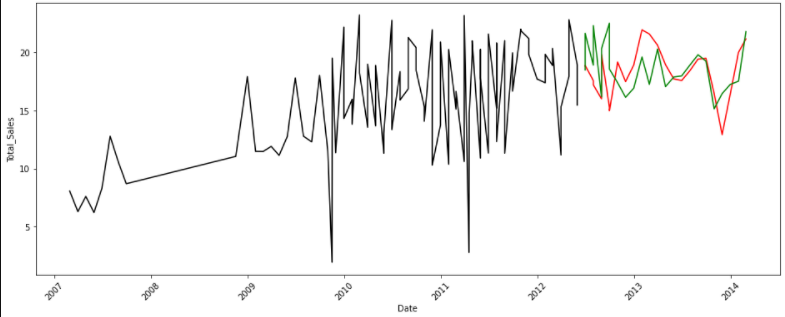


Рисунок 4.18 – Прогнозування вхідних даних кількості продажів ігор за моделлю ARIMA

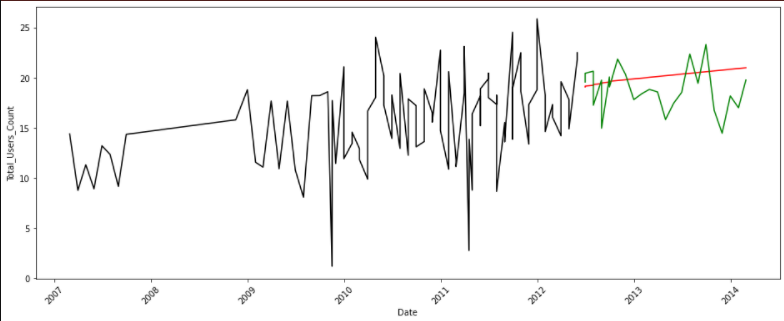


Рисунок 4.19 – Прогнозування вхідних даних кількості гравців за моделлю Holt

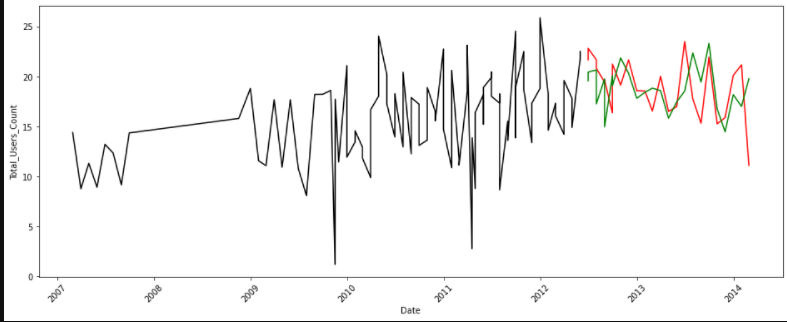


Рисунок 4.20 – Прогнозування вхідних даних кількості гравців за моделлю Exponential Smoothing

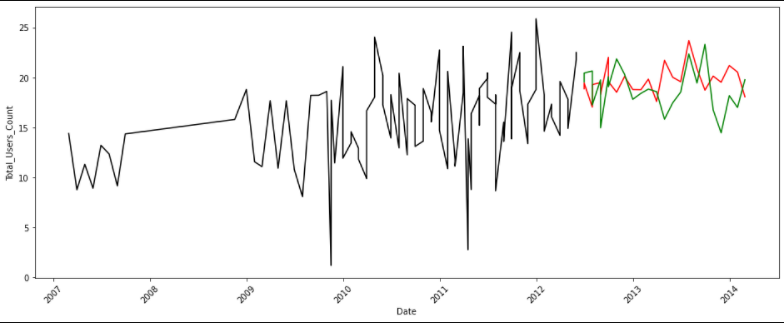


Рисунок 4.21 – Прогнозування вхідних даних кількості гравців за моделлю SARIMAX

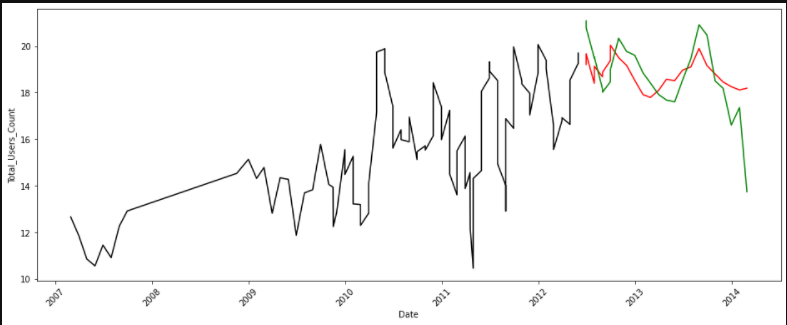


Рисунок 4.22 – Прогнозування згладжувальних даних кількості гравців за моделлю SARIMAX

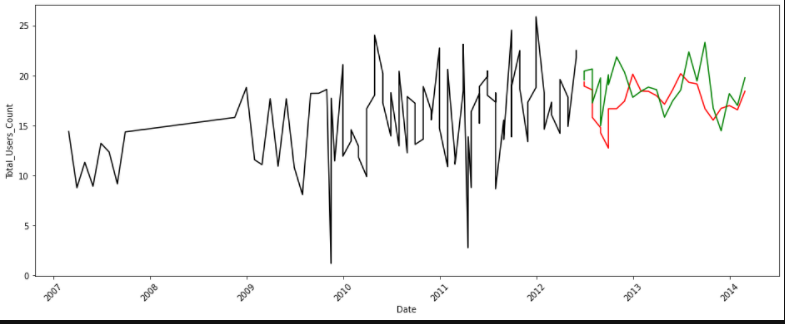


Рисунок 4.23 – Прогнозування вхідних даних кількості гравців за моделлю ARIMA

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].

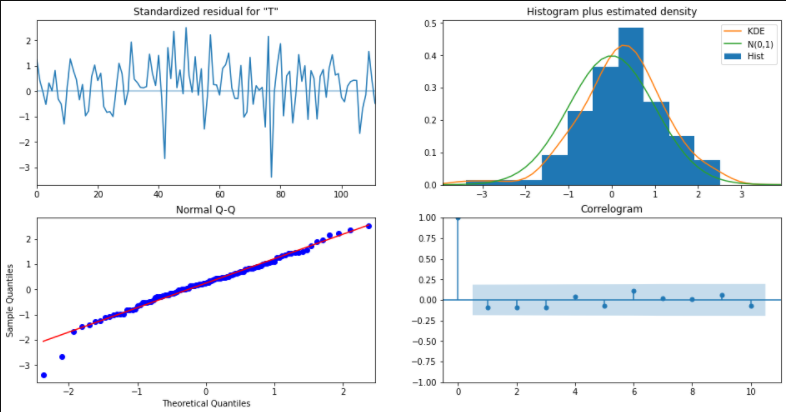


Рисунок 4.24 – Вивід діагностики для моделі SARIMAX на основі даних продажів ігор

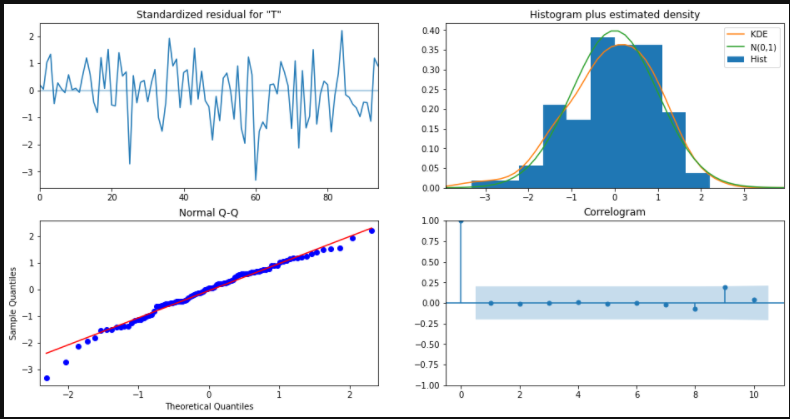


Рисунок 4.25 – Вивід діагностики для моделі SARIMAX на основі даних кількості гравців

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].

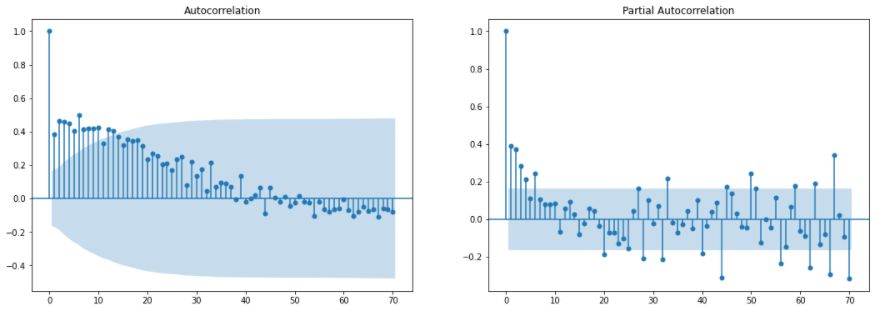


Рисунок 4.26 – Графіки автокореляції та часткової автокореляції на вибірці даних продажів ігор для моделі SARIMAX

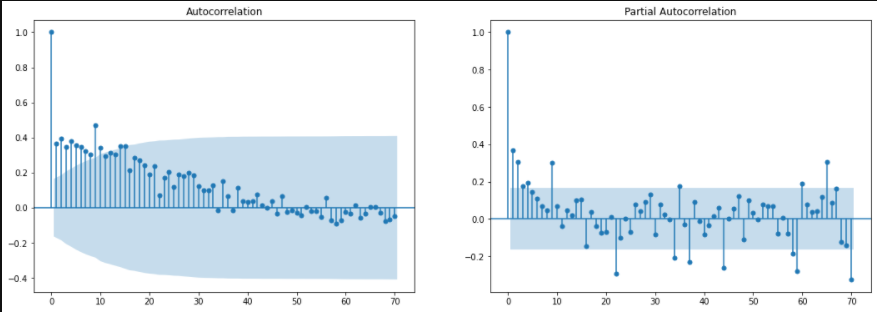


Рисунок 4.27 – Графіки автокореляції та часткової автокореляції на вибірці даних кількості гравців для моделі SARIMAX

…..

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

import pandas as pd

import numpy as np

import datetime

…