Final Project Report

EXTRACTING INSIGHTS FROM DENTAL CLINIC DATA THROUGH DATA SCIENCE TECHNIQUES

10.04.2024

—

Araiz Haider - 0757495

Sujan Bimali - 0776750

Sundeep Kumar - 0770178

Trent University
Peterborough, Ontario

Table of contents

Introduction

Dental practices at present are much different than they used to be in the past. Dental clinics are generating vast and enormous amounts of data, which opens up a series of unrealized possibilities and provides several chances to improve patient care and expedite administrative procedures. One possible way to advance evidence-based dentistry is to integrate data science approaches to extract insights from dental clinic data. This proposed project is to investigate the complex process of deriving significant insights from the vast datasets intrinsic to dental practice, with a focus on using data science as a focal point. By building on the seminal works of scholars like Jain and Wynne (2021) and Favaretto et al. (2021), who have investigated the repurposing of electronic patient data for dental clinical research and the ethical considerations entailed in the digitalization of dentistry, it hopes to contribute to the ongoing conversation surrounding data-driven approaches in dentistry. This study aims to provide insightful viewpoints to academics, policymakers, and dental practitioners alike by thoroughly examining the viability, effectiveness, and ethical implications of utilizing data science tools for the analysis of dental clinic data.

Research Objectives/ Purpose of Study:

1. To investigate the feasibility of employing data science techniques for finding insights from dental clinic data.
2. To explore the potential applications of data science in optimizing dental practice management and improving patient outcomes.

3. To assess the challenges associated with the implication of data science techniques in the context of dental clinic data analysis.
4. To enrich the understanding on how data science implications can revolutionize dental practice by leveraging data produced in dental clinics.
5. To facilitate the advancement of data-driven decision-making processes within the dental practices.
6. To provide realistic insights for integrating data science into dental clinic settings.

Significance of Study:

1. To address the growing significance of using data science in healthcare, especially dentistry, in order to enhance patient care and practice effectiveness.
2. To provide insightful information about the possible advantages and difficulties of implementing data science approaches in dental clinics for researchers, policymakers, and dental practitioners.
3. To contribute in the continuing discussion on the use of data analytics and technology in contemporary dentistry.

Problem of Study:

1. When dealing with large amounts of patient data, ethical implications are necessary. The main challenge with working with clinical data is working with personal identifiable information. A scientist has to work keeping in mind the privacy of patients along with ensuring suitable results.

2. Despite the tremendous volume of data generated in dental clinics, there is a lack of comprehensive understanding of how to effectively utilize data science techniques to extract meaningful insights.
3. Challenges related to data privacy, security, and interoperability are considered as the significant hurdles to the successful implementation of data science in dental clinic settings.
4. Lack of established protocols and guidelines for data collection, processing, and interpretation impedes the use of data science approaches in dentistry.

This research aims to provide light on the transformational potential of data science in extracting insights from dental clinic data by addressing these objectives, elucidating the purpose, highlighting the relevance, and outlining the study's issues.

Literature Review

Schwendicke and Krois (2021) explored the transformative impact of data on dental care and research in their article titled "Data Dentistry: How Data Are Changing Clinical Care and Research." The authors outlined the importance of data as a resource of information in present society and how it can improve the safety, fairness, cost, accessibility, and quality of healthcare. The researchers mentioned that dental research and treatment have changed throughout time, which has been reflected by the emergence of "data dentistry," which comprises three primary applications. First, they highlighted the use of deep learning techniques to explorative medical data analysis, which allowed for the mastery of a variety of data kinds, including audio, language, and photos, and therefore made them more useful in a range of dental scenarios. They further explored how advancements in data application, analysis, & gathering are changing dental care, including more individualized treatment plans and enhanced diagnostic tools. The study also revealed that in order to maximize dental treatments and improve patient outcomes, businesses must integrate several data sources, including genetic data, imaging technology, and electronic health records. Furthermore, the potential of artificial intelligence and big data analytics is explored in terms of pattern recognition, illness progression prognosis, and evidence-based dentistry procedures.

Uribe et al. (2021) studied the availability and quality of dental research data in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) principle. With an emphasis on open-access publications from Europe PubMed Central, the researchers collected articles from 2016 to 2021 that were published in dentistry journals with PubMed indexes. A random sample of 500 dentistry papers that were not open-access were also collected and only 1.5% of the 7,509

articles were examined in the research shared data. The average adherence to FAIR criteria was 32.6%, with findability, accessibility, interoperability, and reusability all showing below-average levels. The study showed that there hasn't been much progress made in terms of data sharing or quality over the years. It showed that dental researchers didn't share their data very often, which made their work harder to replicate and reduced the amount of data that machine learning algorithms could use. These findings emphasize the need for increased efforts to improve data sharing practices in dental research which could make the future analysis on medical dentistry much easier.

Dhopte and Badge (2023) explored the transformative potential of artificial intelligence (AI) in dentistry, emphasizing its ability to revolutionize clinical practice, enhance patient outcomes, and improve overall dental care efficiency. The study focused on the use of AI in a number of fields, including image analysis, patient management, diagnosis, treatment planning, and personalized care. The authors talked about how AI systems showed promise in automatically identifying and diagnosing dental disorders, facilitating early intervention & betterment of treatment outcomes. The study also examined AI-driven treatment planning systems that employ machine learning to assess large amounts of patient data while taking into account anatomical variances, medical histories, and treatment success rates. These technologies gave dentists important information and encouragement for making evidence-based treatment decisions, which resulted in more dependable and customized procedures and thus has increased the business acumen.

Surdilovic and Ille (2022) explored the evolving role of AI in dentistry, emphasizing its applications in diagnostic suggestions, therapeutic protocols, personalized medicine, patient monitoring, and epidemiological disease tracking. The authors emphasized on useful data that AI-driven clinical decision support systems may give, enhancing patient outcomes and those of

the general public. The research highlighted how software used in dental offices are always evolving, with artificial intelligence (AI) being a key factor in increasing business productivity. By boosting efficiency and guaranteeing the evidence-based paperwork required for insurance claims, artificial intelligence (AI) enables intelligent patient scheduling, ideal staffing, and financial advantages. The authors also highlighted the increasing difficulties in dental practice management software & present artificial intelligence (AI) as a driving force behind improvements in the dental industry.

Wynsor (2023) examined the evolving landscape of dental practice management systems, forecasting their future trajectory. The study explored the trends and technical developments that are influencing these systems, such as the incorporation of automation, artificial intelligence, and improved data analytics capabilities. The study highlighted how these developments might simplify administrative duties, enhance patient care, and maximize practice productivity. Dental offices that adopt these advances may expect increased efficiency, improved patient results, and a smoother user experience for both patients and staff. The study emphasized on how important CloudPlus could be used for maximizing a number of dental office management features and enhancing patient experiences by optimizing workflows during the patient's visit. Time-saving benefits include key features including pre-approved appointments, user-defined contemporaneous note templates, and easier charting and treatment planning.

Acharya et al. (2017) aimed to re-characterize these trends and factors, particularly focusing on the impact of the Health Information Technology for Economic and Clinical Health (HITECH) act on adoption rates through 2012. An innovative, statistically-modeled strategy based on early response rates was used to distribute a 39-question survey over the course of three months in order to get a predefined sample for the study. The results showed a 52% acceptance rate of EDR

for clinical assistance. It was noted that dentists in group practices, younger dentists, and dentists with less than 15 years of experience all had greater adoption rates. The results revealed that dental offices have made considerable strides in implementing EDRs and clinical computer systems, suggesting that the industry is moving more and more towards digitalization. The report does, however, also point out enduring difficulties such as issues with cost, compatibility, and usability.

Wanyonyi et al. (2021) explored the underexplored use of electronic patient records for research in dentistry. The study focused on the University of Portsmouth Dental Academy (UPDA) and utilized the R4/Clinical+ electronic patient management system widely used in UK dental practices. The study used a two-step procedure wherein a pilot study was conducted before the main data extraction. Systems Query Language (SQL) was utilized to produce data extracts including factors pertaining to the socioeconomic position, dental treatment, and demographics of patients. The study revealed that researchers can access a multitude of data by utilizing electronic records, which helps them perform more thorough studies and enhance oral treatment. This study emphasized on how crucial it is to guarantee the integrity and correctness of electronic dental data in order to support significant research findings and improve the provision of dental treatment.

AbuSalim et al. (2021) focused on a systematic literature review of deep learning techniques for dental informatics problems. The authors suggested developing thorough and understandable frameworks that might help the healthcare sector in addition to offering insights into the most recent research. The investigation looks at how deep learning is used in dentistry diagnostics, health informatics, and other relevant fields. Along with highlighting the shortcomings of current

methods, the researchers stress the necessity for improved development and offer fresh viewpoints on this fascinating advancement in the area.

Song et al. (2013) conducted a comprehensive review of the current status of reusing electronic patient data for dental clinical research. The purpose of the study was to evaluate how electronic patient data is used in dental research and how this can affect dentistry's use of evidence-based procedures. The results showed that although electronic patient data reuse in dentistry research is starting to pick up steam, it is still in its infancy. The capacity to analyze large samples and time savings were two of the benefits that the researchers noted. However, there were also some drawbacks, such as worries about the quality of the data and the inability to collect study-specific data. The examined studies notably underutilized electronic dental record (EDR) data from private practices. This study explored the advantages and drawbacks of this strategy, making it a thorough resource for scholars and professionals interested in dental research.

Favaretto et al. (2021) conducted a systematic review on the ethical issues surrounding the implementation of Big Data and digitalization in dentistry. The study found that there are several ethical issues when it comes to the use of ICT and big data in healthcare, especially in dental treatment. It was discovered by a thorough analysis of the literature from four databases (Web of Science, PubMed, Scopus, and Cinahl) that these difficulties are consistent with more general ethical concerns in healthcare, such as informed consent, privacy, anonymity, and security. In addition, questions about picture manipulation for insurance fraud and scientific misconduct were brought up, along with worries about online professionalism and business interests supported by digital platforms. This study sheds light on the complex ethical landscape that dentistry is experiencing as it embraces technological advancements.

Jain and Wynne (2021) revealed that digitalization has become essential in contemporary dentistry, and most treatments are likely to switch to digital methods in the near future. This change includes a number of activities that have adopted digital forms, such as taking impressions, documenting jaw motions, teaching new dentists, and encouraging patient participation in the practice. The quickening speed of technical development, however, poses a serious obstacle. Dental clinics and labs have a lot of room for technology and digitalization such as - digital radiography, virtual patient programmes, dental software, and CAD-CAM technologies. This research highlighted the current convergence of artificial intelligence and Big Data, emphasizing the digitalization of medical records and the capacity for cross-border data sharing and manipulation.

Data Set

The final queries that we used to extract data from the provided database are below:

Query 1: Patient Data Extraction Sample

This Query Fetches Patient Data From Table.

```
SELECT
fld_auto_intPatId as PatientId,
CONCAT(fld_strFName , ' ', fld_strLName) AS FullName,
fld_intFamId as FamilyId,
fld_strAddr1 AS patientAddress,
fld_strCity as city,
fld_strProv as province,
fld_strCountry as country,
fld_strPCode as postalCode,
fld_strFamDr as familyDoctor,
fld_strFamDrAddr as doctorAddress
FROM [ClearDent].[dbo].[tbl_PatInfo];
```

PatientId	FullName	FamilyId	patientAddress	city	province	country	postalCode	familyDoctor	doctorAddress
12345	John Doe	6789	123 Elm St.	Courtice	ON	CA	L1E0B7	Dr. Smith	321 Oak St.

Query 2: Patients Payments Made

This Query fetches Patients payments made.

```
SELECT
PTI.fld_auto_intPatId as PatientId,
CONCAT(PTI.fld_strFName , ' ', PTI.fld_strLName) AS FullName,
PTI.fld_intFamId as FamilyId,
PTI.fld_strAddr1 AS patientAddress, PTI.fld_strCity as city,
PTI.fld_strProv as province,
PTI.fld_strCountry as country, PTI.fld_strPCode as postalCode,
PTI.fld_strFamDr as familyDoctor, PTI.fld_strFamDrAddr as
doctorAddress, DS.Description as paymentMethod,
SUM(DS.InsClm) as insuranceClaim, SUM(DS.InsPmt) as insurancePayment,
SUM(DS.TotalFee) as patientBill,SUM(DS.PatPmt) as patientPayment
FROM
[ClearDent].[dbo].[tbl_PatInfo] PTI
INNER JOIN
[ClearDent].[dbo].[tbl_DaySummary] DS
ON DS.PatId = PTI.fld_auto_intPatId
GROUP BY
PTI.fld_auto_intPatId, PTI.fld_intFamId, PTI.fld_strAddr1,
PTI.fld_strCity, PTI.fld_strProv, PTI.fld_strCountry,
PTI.fld_strPCode, PTI.fld_strFamDr, PTI.fld_strFamDrAddr,
DS.Description
ORDER BY PTI.fld_auto_intPatId;
```

PatientId	FamilyId	patientAddress	city	province	country	postalCode	familyDoctor	doctorAddress	paymentMethod	insuranceClaim	insurancePayment	patientBill	patientPayment
12345	6789	123 Elm St.	Courtice	ON	CA	L1E0B7	Dr. Smith	321 Oak St.	Cash	200.00	150.00	350.00	200.00

Query 3: Individual Bank Deposit Details

This Query Contains Individual Bank Deposit Details By Patient on different days

```
SELECT
PTI.fld_auto_intPatId as PatientId,
CONCAT(PTI.fld_strFName , ' ', PTI.fld_strLName) AS FullName,
PTI.fld_intFamId as FamilyId, PTI.fld_strAddr1 AS patientAddress,
PTI.fld_strCity as city, PTI.fld_strProv as province,
PTI.fld_strCountry as country, PTI.fld_strPCode as postalCode,
PTI.fld_strFamDr as familyDoctor, PTI.fld_strFamDrAddr as
doctorAddress, BD.fld_fltDeposit AS paymentRequired,
BD.fld_fltInsPmt as paymentByInsurance, BD.fld_fltPatPmt AS
paymentByPatient,BD.fld_dtmDateTime AS paymentDate
FROM
[ClearDent].[dbo].[tbl_PatInfo] PTI
INNER JOIN
[ClearDent].[dbo].[tbl_Payment] PT
ON PT.fld_intPatId = PTI.fld_auto_intPatId
LEFT OUTER JOIN
[ClearDent].[dbo].[tbl_BankDeposit] BD
```

```
ON PT.fld_intBankDepositId = BD.fld_auto_intBankDepositId
GROUP BY
PTI.fld_auto_intPatId, PTI.fld_intFamId, PTI.fld_strFName,
PTI.fld_strLName, PTI.fld_strAddr1, PTI.fld_strCity, PTI.fld_strProv,
PTI.fld_strCountry, PTI.fld_strPCode, PTI.fld_strFamDr,
PTI.fld_strFamDrAddr, BD.fld_fltDeposit, BD.fld_fltInsPmt,
BD.fld_fltPatPmt, BD.fld_dtmDateTime
ORDER BY PTI.fld_auto_intPatId;
```

PatientId	FullName	FamilyId	patientAddress	city	province	country	postalCode	familyDoctor	doctorAddress	paymentRequired	paymentByInsurance	paymentByPatient	paymentDate
12345	John Doe	6789	123 Elm St.	Courtice	ON	CA	L1E0B7	Dr. Smith	321 Oak St.	500.00	150.00	350.00	2021-10-01

Query 4: Total Money Spent and Insurance Coverage Sample

This query fetches patient details combined with their total money spent and its division (How much insurance covered.

```
SELECT
PTI.fld_auto_intPatId as PatientId,
CONCAT(PTI.fld_strFName , ' ', PTI.fld_strLName) AS FullName,
PTI.fld_intFamId as FamilyId,
PTI.fld_strAddr1 AS patientAddress,
PTI.fld_strCity as city,
PTI.fld_strProv as province,
PTI.fld_strCountry as country,
PTI.fld_strPCode as postalCode,
```

```
PTI.fld_strFamDr as familyDoctor,
PTI.fld_strFamDrAddr as doctorAddress,
SUM(BD.fld_fltDeposit) AS totalPaymentRequired,
SUM(BD.fld_fltInsPmt) as totalPaymentByInsurance,
SUM(BD.fld_fltPatPmt) AS totalPaymentByPatient
FROM
[ClearDent].[dbo].[tbl_PatInfo] PTI
INNER JOIN
[ClearDent].[dbo].[tbl_Payment] PT
ON PT.fld_intPatId = PTI.fld_auto_intPatId
LEFT OUTER JOIN
[ClearDent].[dbo].[tbl_BankDeposit] BD
ON PT.fld_intBankDepositId = BD.fld_auto_intBankDepositId
GROUP BY
PTI.fld_auto_intPatId, PTI.fld_intFamId,
PTI.fld_strFName, PTI.fld_strLName,
PTI.fld_strAddr1, PTI.fld_strCity, PTI.fld_strProv,
PTI.fld_strCountry, PTI.fld_strPCode,
PTI.fld_strFamDr, PTI.fld_strFamDrAddr
ORDER BY PTI.fld_auto_intPatId;
```

PatientId	FullName	FamilyId	patientAddress	city	province	country	postalCode	familyDoctor	doctorAddress	totalPaymentRequired	totalPaymentByInsurance	totalPaymentByPatient
12345	John Doe	6789	123 Elm St.	Courtice	ON	CA	L1E0B7	Dr. Smith	321 Oak St.	1200.00	400.00	800.00

Query 5: Procedures Performed on Patients Sample

This query fetches patient details combined with the procedures performed on them.

```
SELECT PTI.fld_auto_intPatId as PatientId, CONCAT(PTI.fld_strFName , ' ',
PTI.fld_strLName) AS FullName, PTI.fld_intFamId as FamilyId, PTI.fld_strAddr1
AS patientAddress,  PTI.fld_strCity as city, PTI.fld_strProv as province,
PTI.fld_strCountry as country, PTI.fld_strPCode as postalCode,
PTI.fld_strFamDr as familyDoctor, PTI.fld_strFamDrAddr as doctorAddress,
fld_strDesc AS procedurePerformed, fld_strODesc AS procedureDetails,
fld_dtmDueDate AS procedureDate, fld_strCity AS procedureCity, fld_strProv AS
procedureProvince, fld_strCountry AS procedureCountry
FROM
tbl_ProcRecType PRT INNER JOIN tbl_ProcCat PC ON PRT.fld_bytRecTypeId =
PC.fld_auto_bytProcCatId INNER JOIN tbl_ProcInfo PCI
ON PRT.fld_strProcCode = PCI.fld_strProcCode INNER JOIN tbl_Rec RC
ON RC.fld_bytRecType = PRT.fld_bytRecTypeId INNER JOIN tbl_PatInfo PTI
ON RC.fld_intPatId = PTI.fld_auto_intPatId
GROUP BY  PTI.fld_auto_intPatId, PTI.fld_intFamId, PTI.fld_strFName,
PTI.fld_strLName, PTI.fld_strAddr1, PTI.fld_strCity, PTI.fld_strProv,
PTI.fld_strCountry, PTI.fld_strPCode, PTI.fld_strFamDr, PTI.fld_strFamDrAddr,
fld_strDesc, fld_strODesc, fld_dtmDueDate, fld_strCity, fld_strProv,
fld_strCountry
ORDER BY PTI.fld_auto_intPatId;
```

PatientId	FullName	FamilyId	patientAddress	city	province	country	postalCode	familyDoctor	doctorAddress	procedurePerformed	procedureDetails	procedureDate	procedureCity	procedureProvince	procedureCountry
12345	John Doe	6789	123 Elm St.	Courtice	ON	CA	L1E0B7	Dr. Smith	321 Oak St.	Teeth Cleaning	Full clean	2021-10-02	Courtice	ON	CA

Attributes:

- PatientId: A unique identifier for a patient within the database.
- FullName: The patient's full name, combining first and last names.
- FamilyId: A number linking a patient to their family group or unit.
- patientAddress, city, province, country, postalCode: The various components of the patient's address.
- familyDoctor: The name of the patient’s primary healthcare provider.
- doctorAddress: The address of the family doctor's practice.
- paymentMethod, insuranceClaim, insurancePayment, patientBill, patientPayment, paymentRequired, paymentByInsurance, paymentByPatient, paymentDate: These fields detail the financial transactions related to the patient's care, including how payments were made, amounts claimed and paid by insurance, patient bills, and out-of-pocket expenses.
- totalPaymentRequired, totalPaymentByInsurance, totalPaymentByPatient: Summarized financial details representing the total monetary aspects of the patient's care.
- procedurePerformed, procedureDetails, procedureDate, procedureCity, procedureProvince, procedureCountry: Information pertaining to medical procedures conducted, including descriptions, dates, and locations.

Each sample row provides a snapshot of the type of data

Methodology

In our project, we have used different methodologies to make sense of the data we have. Firstly, we extracted data from different SQL tables, we got a number of csv files that contained different attributes. Once we had that, we made use of each data file separately. This was needed because each file contained a different set of attributes, therefore, different techniques needed to be applied to each dataframe. We treated each data file separately, and analyzed the procedure that could be applied to the type of data we had. After that we performed specific procedures such as chi-square test, time series, and classification algorithm etc.

Generally, we made use of different data science techniques. Firstly, we performed exploratory data analysis on the data to make meaning from it. After that, we also plotted some boxplots to look for outliers in different variables. Later, we created visualizations and scatter plots to understand the relationship between different variables.

Secondly, we performed time series models on some variables. Assessed the performance of time series models and visualized the results. Furthermore, we also performed a chi-square test too. Before training models we performed necessary pre-processing steps which included sorting variables, checking outliers, and removing missing values etc. In short, our method consisted of extracting data, getting it in the right shape, understanding data, applying procedures based on type of data, visualizing results and evaluating performance of models.

The reason for choosing this method and dealing with each table separately is because of the nature of data we have. Each table contained a different set of variables, so it was not possible to treat each table in the same way and apply the same preprocessing steps or models to each table.

This way it would not have made much sense. The pros of using this method is, we have been able to see and analyze data through different viewpoints. Usage of statistical tests (chi-square), time series models (ARIMA and SARIMA), classification algorithms (KNN and Decision Tree) enabled us to extract value and gain helpful insights from the raw data. Similarly, it further proves that clinical data that we have can be used for a variety of purposes and it can answer many questions. The data has a great potential for data science techniques to be applied to it. However, there is still a large room for improvement in data collection and data handling practice.

Results

We made use of different techniques to extract value from the data. We performed a hypothesis test on the patient-procedures table to find the relation between postalCode and procedurePerformed by using chi-square test. Here is the null and alternative hypothesis:

Null Hypothesis: No association between postal code and procedure performed.

Alternative Hypothesis: There is association between postal code and procedure performed.

Here are the results of Chi-square test:

Chi-Squared Value : 658.7

P-Value: 2.36

Degrees of Freedom: 90

When we used full postal code, the chi-square test didn't give a reasonable value. It produced (p-value = 1) exactly, which meant there was something wrong with the variables being chosen. Therefore, we used only the first 3 characters of postal code to perform chi-square because the first 3 characters of postal code give us a more generalized demographic division. However, a full postal code might include only a few houses therefore the relationship doesn't make sense.

Chi-square test results in p-val = 2.36, which indicates we fail to reject the null hypothesis against significance level of alpha = 0.05. Therefore, there is no significant relationship between postal code and procedure performed.

On the other hand, we focused on the ChartPerio table to assess patients' dental health, utilizing four key columns—Classification, Hygiene, Calculus, and ExamType—for machine learning models to classify dental health conditions. We made use of Decision Tree and K-Nearest Neighbors (KNN) classifiers, achieving 70% and 66% accuracy, respectively. The analysis was limited by a reduced dataset from 607 to 177 samples due to missing values, highlighting the critical need for comprehensive data collection and suggesting further exploration with machine learning to enhance model performance. Additionally, the bar plot below visually demonstrates the daily financial transactions for medical services over a two to three-month period, indicating the variability in charges.

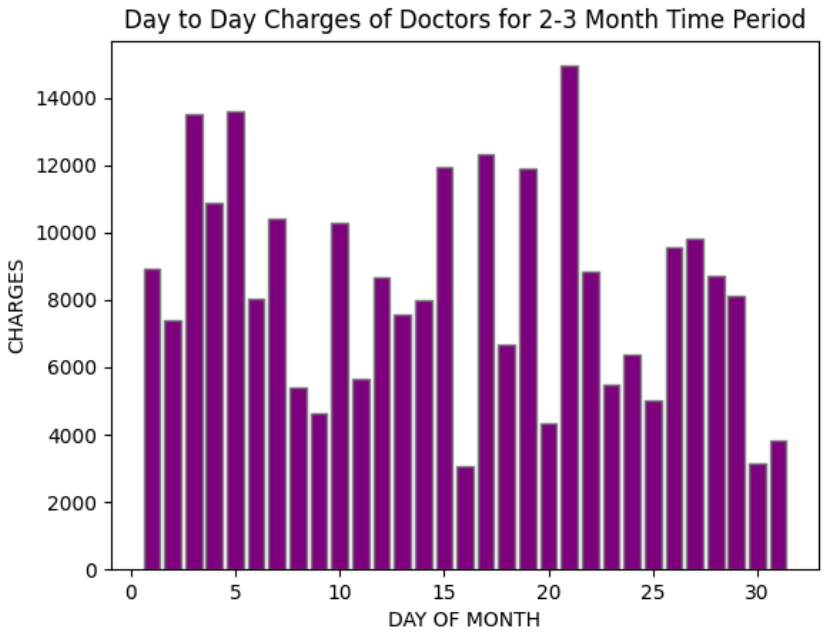


Figure 1

Figure 1 reveals significant day-to-day fluctuations in charges, reflecting variations in patient volume or service complexity, with notable peaks suggesting busier periods or costly procedures. Despite the difficulty in identifying clear monthly trends due to these fluctuations, the charges' wide range from below 2,000 to over 12,000 indicates unpredictable daily potential. This variability might hint at underlying patterns correlating with specific months, though such speculation requires more data for confirmation. The observations suggest potential uses for this data in optimizing resource allocation, scheduling, financial forecasting, and maintaining care quality. Essentially, this analysis underlines the importance of understanding charge variability for better administrative and strategic planning in medical practice management.

The bar chart below in Figure 2 illustrates the month-to-month revenue fluctuations of a business or practice as reflected by the deposited amounts. The vertical axis represents the revenue in Canadian dollars (CAD), and the horizontal axis represents each month of the year. It reveals significant seasonal revenue variations, with peaks in the winter months (November to February) attributed to year-end activities and a noticeable decline from May to July, possibly due to reduced customer demand or operational slowdowns. These trends highlight the importance of strategic planning in inventory management, staffing, and marketing to navigate seasonal impacts. Insights also suggest focusing on cash flow management to address revenue dips and considering customer engagement techniques and operational adjustments to maintain steady revenue throughout the year.

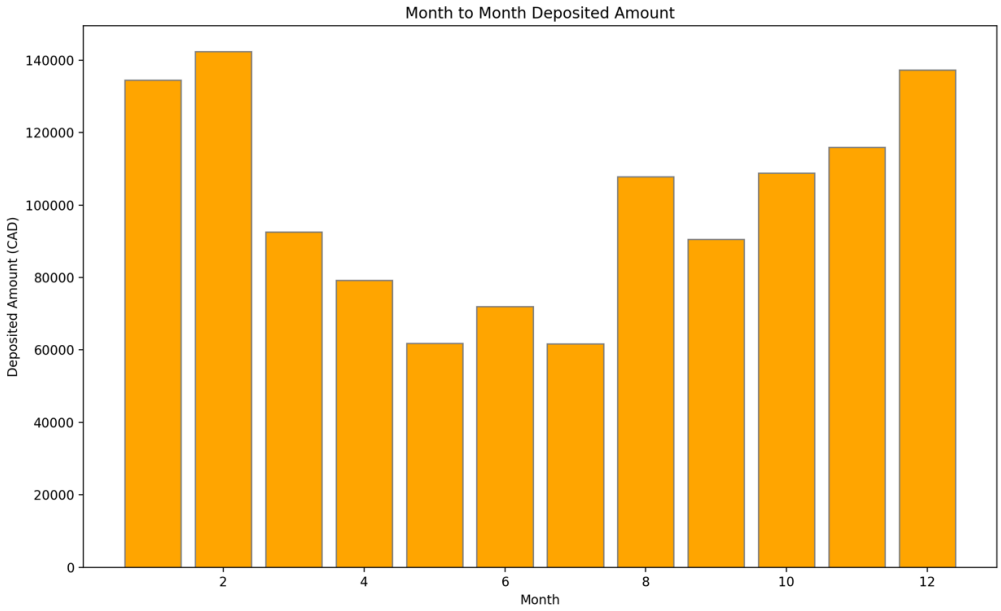


Figure 2

The bar chart below in Figure 3 now presents the revenue fluctuations on a day-to-day basis over a period of 2-3 years, as measured by the amounts deposited. The vertical axis quantifies the revenue in Canadian dollars (CAD), while the horizontal axis denotes the days of the month.

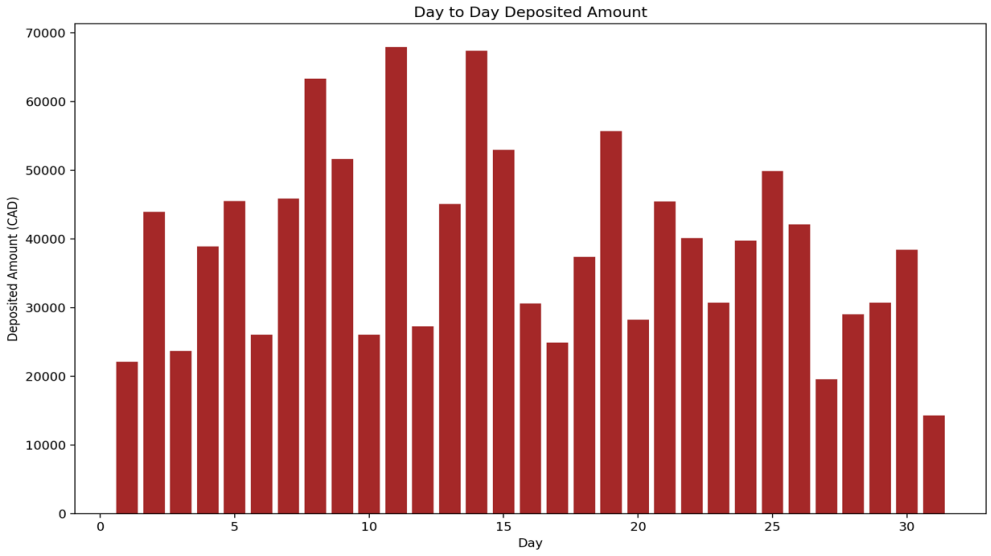


Figure 3

The bar chart analysis reveals distinct days throughout the month that stand out due to significantly higher revenue, suggesting a pattern of increased transactions or the impact of specific events. This pattern underscores the potential for strategic planning, pointing to a cycle of heightened customer engagement or business activity that could be instrumental for operational planning. The variability observed in the revenue underscores the dynamic nature of daily business operations, with fluctuations between days of high and low earnings highlighting the diverse performance of the business.

Insights garnered from the analysis of daily revenue patterns offer valuable information for refining business strategies. By identifying and understanding the factors behind high-revenue days, businesses can optimize resource allocation and staffing to accommodate the increased demand. Further investigation into these patterns, incorporating variables like customer traffic, promotional activities, or broader economic conditions, could enhance understanding and leverage these trends for business advantage. For instance, recognizing revenue spikes linked to specific promotions or events enables targeted strategy adjustments for future endeavors. Similarly, understanding customer spending behaviors on high-revenue days could inform more effective marketing strategies. Conversely, examining days with lower revenue might reveal opportunities for improving business efficiency or identifying strategies to boost customer interest and business intake.

The plot below in Figure 4 is a comprehensive representation of the types and counts of dental procedures carried out each month. The vertical axis quantifies the number of procedures, while the horizontal axis categorizes the months of the year. Different colors represent different types of procedures, as defined in the legend. The visualization of procedure counts by month provides

insights into patient preferences and behaviors, which can be leveraged to enhance service delivery, patient satisfaction, and practice efficiency. It would also be beneficial to overlay this data with financial performance figures to see how different types of procedures contribute to the practice's overall revenue.

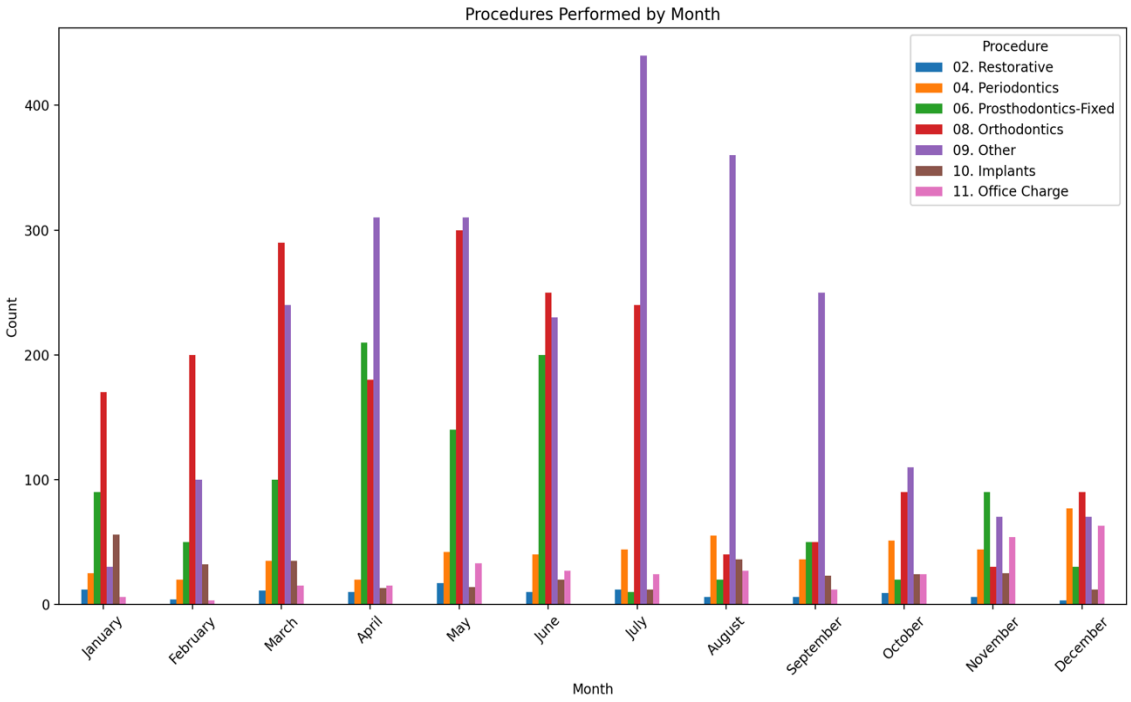


Figure 4

The analysis of the data highlights several key points, such as a significant surge in unspecified procedures in July, indicating a mid-year spike in demand for services that fall outside the usual categories. Orthodontic treatments also see an increase during March and May, likely aligning with periods favorable for initiating such treatments due to academic schedules or financial considerations. Conversely, there's a marked reduction in the volume of procedures from October to December, possibly as a result of the holiday season influencing patients to delay non-urgent dental work. These trends offer valuable insights for dental practices in several areas.

Firstly, they underline the importance of tailoring staffing and scheduling to accommodate fluctuating demand, ensuring operational efficiency. Additionally, awareness of when certain procedures are most sought-after can help administration to manage inventory, ensuring necessary supplies are readily available. Furthermore, strategically planned marketing initiatives during quieter months could help maintain a steady flow of procedures. Lastly, recognizing patterns in procedure frequency throughout the year can aid in more accurate financial forecasting, allowing practices to plan effectively for busier or slower periods.

This scatter plot aims to visualize any potential correlation between the number of available units and the charges for medical services rendered by doctors.

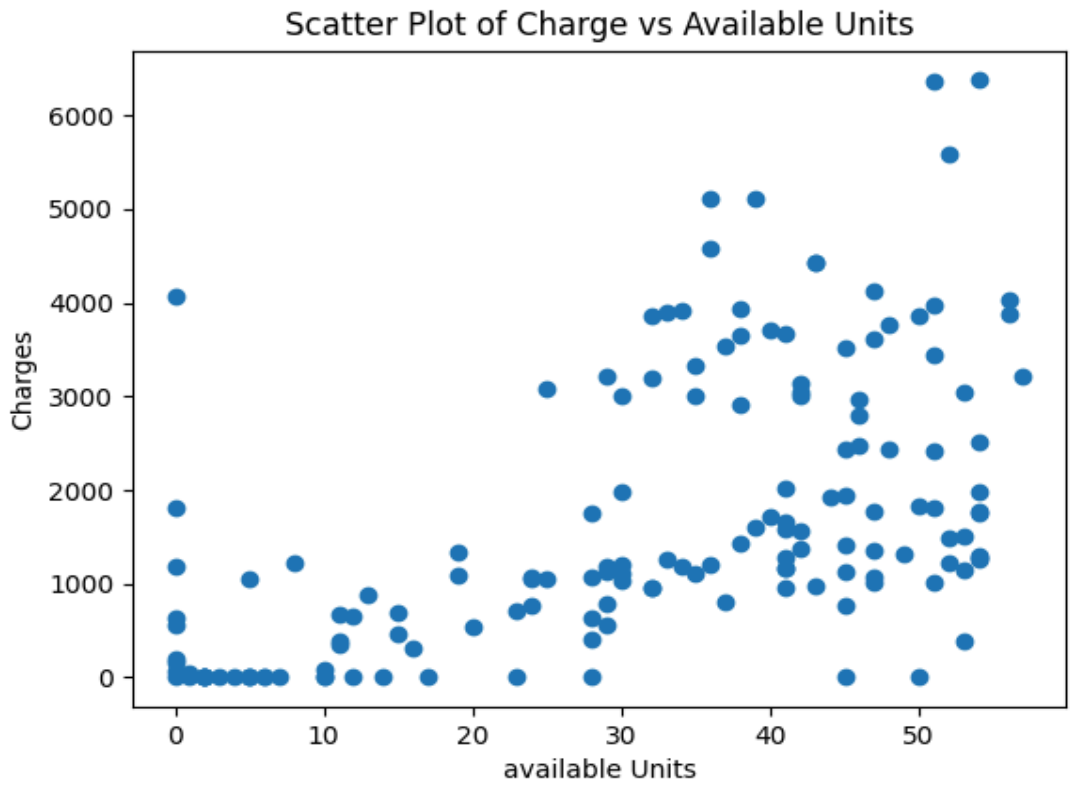


Figure 5

Figure 5's graph illustrates a broad range of fees across varying levels of service availability, with costs spanning from nearly zero to upwards of 5,000, regardless of how many slots are open. Notably, a higher density of data points is observed at lower availability levels, hinting that lower service availability tends to coincide with the recording of most fees. This could also mean the presence of outliers and mistakes committed by representatives while entering data. There's an absence of a clear trend linking service availability to the fees charged, suggesting a random distribution of data points and indicating that the fees levied for services aren't directly determined by a doctor's availability.

The analysis reveals fluctuations in charges across different availability levels, implying that factors beyond mere availability, such as service complexity, the nature of the cases, or the individual pricing strategies of doctors, might influence the cost. Interestingly, higher fees are not exclusively associated with greater availability, debunking the assumption that more available doctors necessarily command higher fees. The widespread use of charges also highlights the potential necessity of investigating outliers, especially where charges are significantly above average, to grasp the underlying reasons fully. From this analysis, several implications arise. Firstly, understanding the practice's pricing structure may require considering factors other than availability, such as the type of service or patient demographics. Secondly, merely increasing availability may not be financially advantageous for the practice; a more nuanced strategy might involve optimizing the schedule to align with peak demand periods or more lucrative services. Finally, a deeper dive into the data, possibly by segregating it based on service type or individual practitioner, could yield insights more conducive to making informed, data-driven decisions for the practice.

This scatter plot in Figure 6 visualizes the correlation between 'Productive Units'—likely a measure of how much service a doctor provides—and the associated charges for those services.

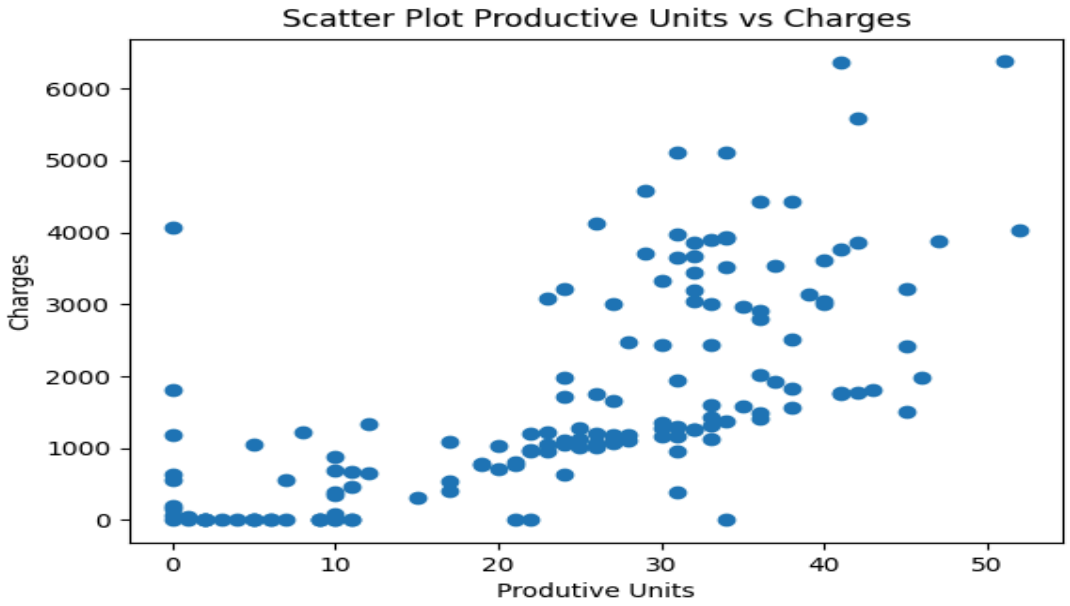


Figure 6

A positive relationship between the amount of services provided (productive units) and the corresponding charges has been observed, indicating that as services increase in volume or complexity, so do the charges. This pattern predominantly occurs within a specific range, with most services and charges clustering in the lower to middle spectrum. However, a notable variation in charges emerges beyond a certain productivity level, potentially due to diverse service complexities or differential pricing. The data points to a steady demand for services, reflecting an economic dynamic where increased doctor productivity could lead to higher charges. This scenario underscores the need for healthcare providers to strategically manage billing, evaluate the efficiency of service delivery, and allocate resources to maximize revenue, especially as the intensity and demand for services escalate.

This scatter plot in Figure 7 differentiates providers by color coding and illustrates the relationship between their productive units, presumably a measure of the time they work, and the charges they apply for their services.

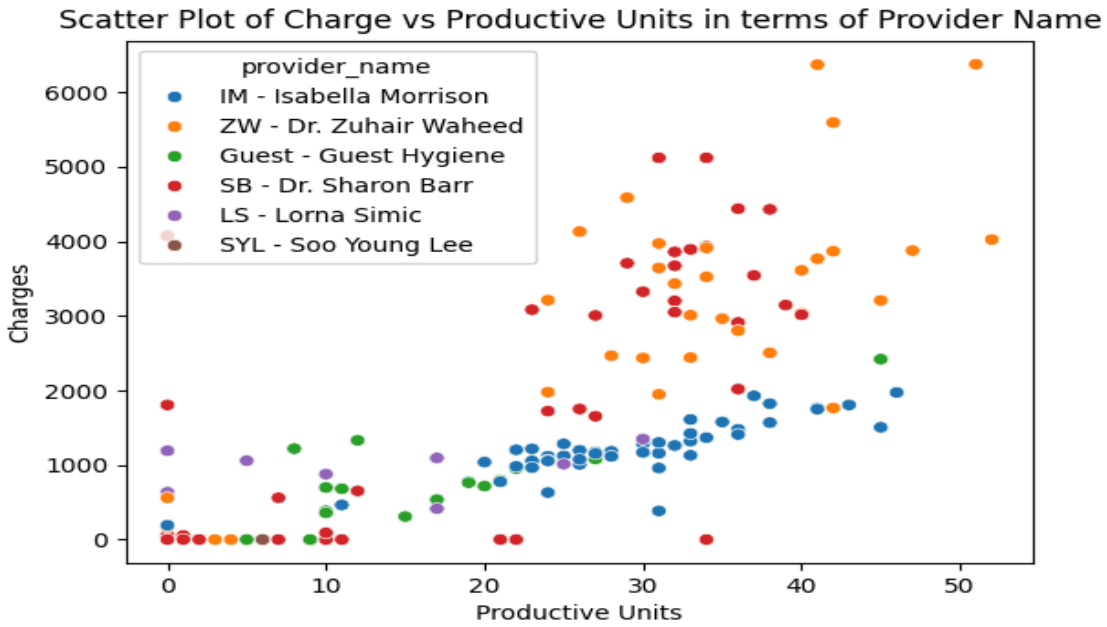


Figure 7

The plot facilitates a straightforward comparison among providers, revealing a variation in charges for comparable productive units, with some showing greater consistency than others. It indicates a general trend of increasing charges with more productive units, though this varies significantly across providers. Certain providers cluster at higher productive units with correspondingly higher charges, suggesting a link between service volume and pricing. This variability could mirror the efficiency of providers, their service complexity, or their strategic pricing based on specialization and reputation. The insights gained could inform efficiency and pricing strategies, providing a basis for evaluating provider performance and productivity,

identifying areas for operational enhancement, and aligning service offerings with market demand. This understanding could guide practices in optimizing schedules and services to meet demand effectively, potentially allowing for higher charges for sought-after services.

The next scatter plot offers a nuanced look at how different providers' availability correlates with the charges they issue. The plot's color coding distinguishes between providers, while the axes represent the available units and the corresponding charges.

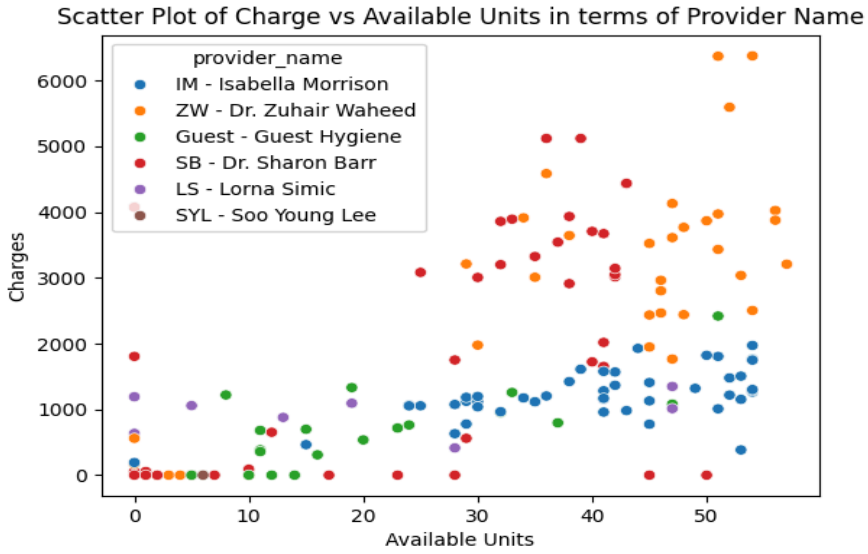


Figure 8

For the purpose of managing practice operations, such as establishing competitive prices and streamlining scheduling, it might be helpful to comprehend each provider's billing pattern in respect to their availability. It appears from the scatter that availability is not the only factor in fees. Rather, they probably represent the kind of service rendered, the level of experience of the practitioner, or the intricacy of the course of therapy. Certain suppliers could charge more when

they are less available, which could indicate increased productivity or a concentration on more profitable services.

Here is the scatter plot in Figure 9 that showcases the relationship between productive units and charges across two departments within a healthcare setting: hygiene and restorative (resto).

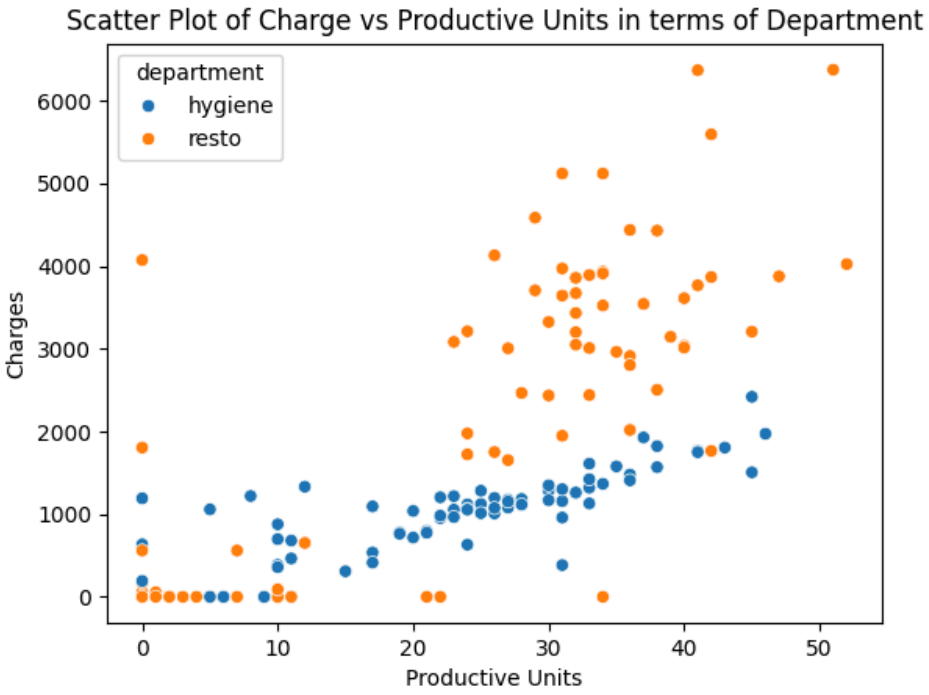


Figure 9

The graph clearly differentiates between hygiene and restorative departments using color codes, revealing a pattern where, on average, restorative services incur higher fees than hygiene ones. This is attributed to the relationship between the quantity of services provided and the resulting charges, with restorative services demonstrating a more pronounced increase in charges for every unit of service delivered. Charges for hygiene services are typically grouped at the lower end of the spectrum, while restorative services show a broader distribution, often commanding higher

fees. This disparity suggests that restorative services are priced higher, likely due to their complexity and the extensive resources they require. Additionally, the clustering of lower charges for hygiene services hints at a prevalence of quick, standard services, in contrast to the more involved and time-consuming restorative procedures. This analysis could guide strategic decisions regarding pricing, operational adjustments, and the distribution of resources, emphasizing support for the restorative department to optimize revenue.

This graph below in Figure 10 is a time series plot depicting deposited amounts over a period. Such plots are valuable for observing trends, patterns, and outliers in financial data over time. Let's break down the key elements of this graph.

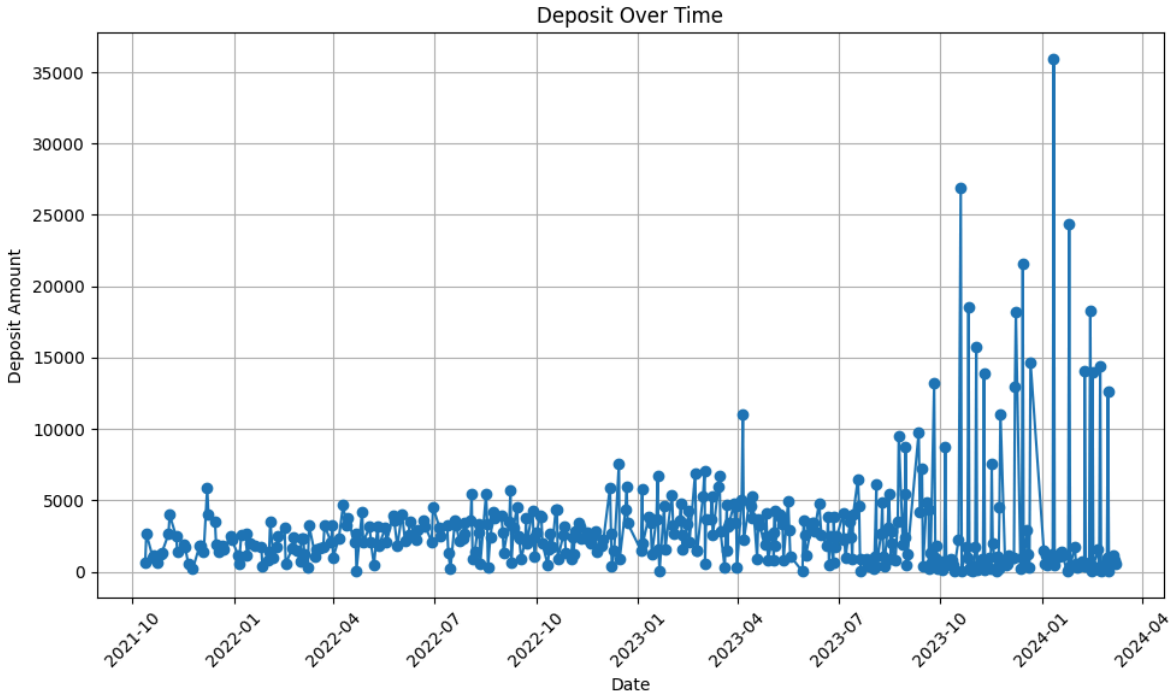


Figure 10

The plot spans several months, starting from October 2021 to April 2024, allowing for the observation of long-term trends in deposit amounts. There is a significant variability in deposit amounts over time, with certain days or periods showing much higher deposits than others. There appears to be no immediate, obvious upward or downward trend when considering the entire time frame. Instead, the deposit amounts fluctuate. There are conspicuous spikes in the data that could be outliers or could indicate days of unusually high deposit amounts, possibly due to large transactions or an aggregation of transactions on a specific day.

The analysis of the deposit trends reveals several key observations: The presence of peaks suggests that the deposits are influenced by seasonal trends or other initiatives that drive up transactions. Additionally, the clustering of data points around a baseline figure offers a glimpse into the consistent cash flow of the business, aside from exceptional events. While short-term variations are noticeable, determining a definitive long-term trend in deposits would necessitate a deeper statistical examination through methods like trend lines or moving averages. These insights carry significant implications for business operations. The observed deposit patterns could serve as a foundation for enhancing cash flow management strategies, ensuring the business has sufficient liquidity. Moreover, understanding the factors behind fluctuations in deposit amounts can aid in strategic planning, particularly in optimizing for periods of increased revenue. Finally, recognizing and leveraging consistent patterns in deposit amounts across years could prove invaluable for accurate financial forecasting and budgeting, allowing businesses to make informed decisions for future growth.

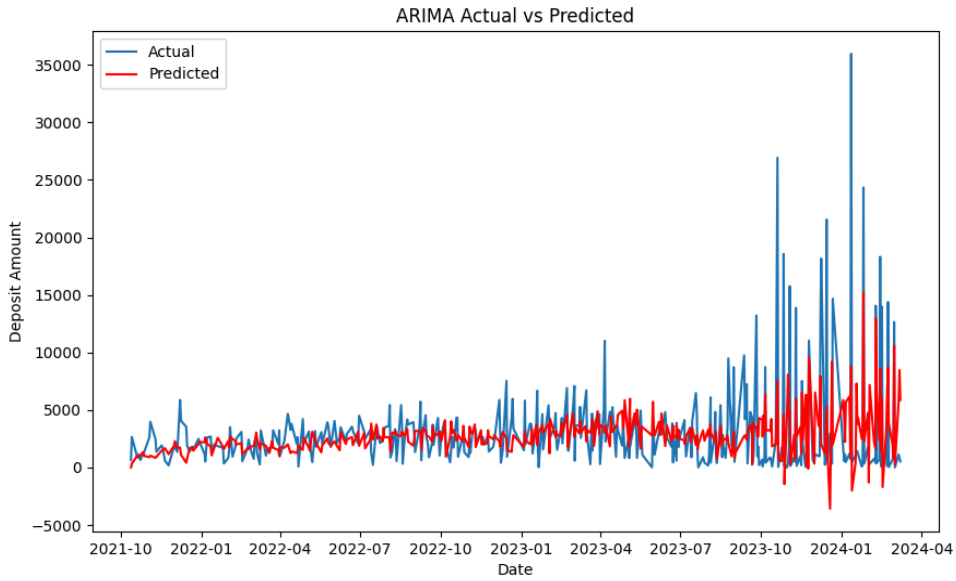


Figure 11

In figure 11, we have used the ARIMA (autoregressive integrated moving average) time series model. With time on the horizontal axis and deposit amounts on the vertical, the "ARIMA Actual vs. Predicted" graph compares actual and predicted deposit amounts from October 2021 to April 2024. The actual deposits, which are shown in blue, exhibit notable spikes and fluctuations that point to times when the deposits were significantly higher. The ARIMA model's red-colored predicted quantities are less volatile and generally follow the trend of the real data, however they appear smoother and underestimate the peaks and troughs, particularly in early 2024. The Root Mean Squared Error (RMSE) of 3374.905394588419 and the Mean Absolute Error (MAE) of 2006.6691566215247, which represent the average prediction errors and highlight the model's moderate predictive accuracy, are used to quantify the model's performance. The differences between actual and anticipated values, especially during periods of notable variation, point to the ARIMA model's limits in capturing all variables or seasonal

patterns, even while it captures the overall upward trend in deposits. These variations underline the need for care when applying these forecasts to financial choices and imply that further information or modifications may be required to enhance the model's dependability and accuracy for future planning.

The graph below titled "SARIMA Actual vs Predicted" exhibits the observed and forecasted deposit amounts over a specified timeframe, utilizing the SARIMA model framework. The SARIMA model is particularly adept at identifying and adjusting for patterns that recur at consistent intervals, known as seasonality, which might be observed across specific days, weeks, or months within a year.

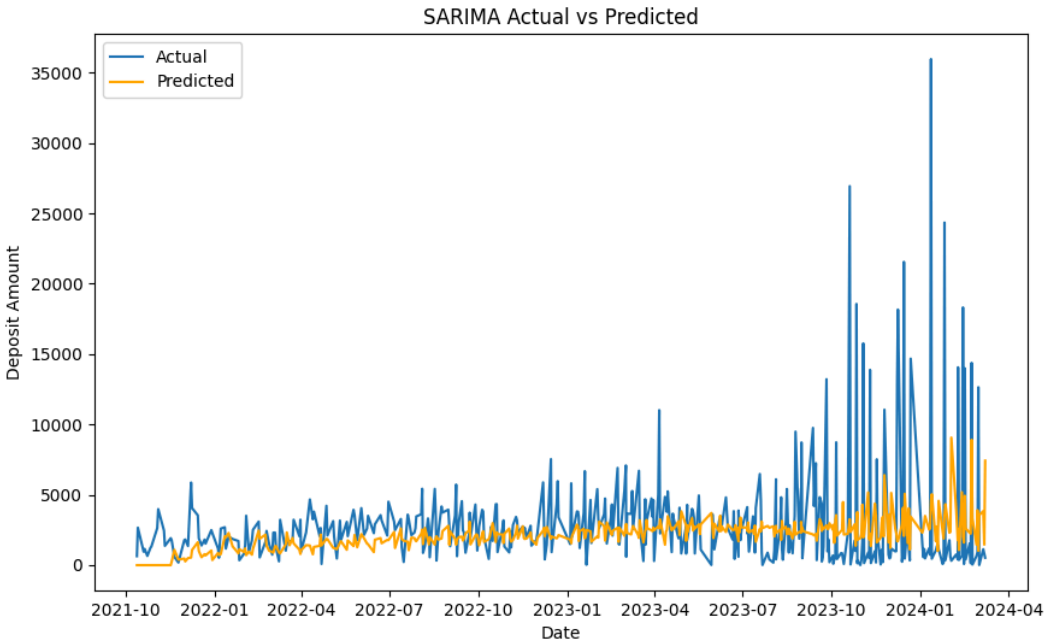


Figure 12

The model results are below:

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                Deposit   No. Observations:                  399
Model:                 ARIMA(5, 1, 0)   Log Likelihood               -3871.409
Date:                Fri, 08 Mar 2024   AIC                           7754.817
Time:                        22:26:14   BIC                           7778.736
Sample:                             0   HQIC                          7764.291
                                  - 399
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9493      0.044    -21.378      0.000      -1.036      -0.862
ar.L2         -0.8512      0.062    -13.673      0.000      -0.973      -0.729
ar.L3         -0.6549      0.062    -10.485      0.000      -0.777      -0.532
ar.L4         -0.3244      0.057     -5.644      0.000      -0.437      -0.212
ar.L5         -0.1521      0.038     -4.027      0.000      -0.226      -0.078
sigma2      1.663e+07    3.5e+05     47.563      0.000    1.59e+07    1.73e+07
===================================================================================
Ljung-Box (L1) (Q):                   0.45   Jarque-Bera (JB):              8403.38
Prob(Q):                              0.50   Prob(JB):                         0.00
Heteroskedasticity (H):              25.62   Skew:                             3.40
Prob(H) (two-sided):                  0.00   Kurtosis:                        24.46
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
Root Mean Squared Error (RMSE): 4048.973973655084
```

Figure 13

The SARIMAX analysis shows that, in spite of an inverse association shown by negative coefficients, historical observations have a significant impact on current projections when deposit amounts are predicted using an ARIMA(5, 1, 0) framework. Concerns about overfitting or overly complex models are raised by high AIC and BIC values. The correctness of the model is demonstrated by a high Root Mean Squared Error (RMSE) of about 4049 and a considerable residual variance, which both suggest that deposit variations are notably unpredictable. Diagnostic tests show heteroskedasticity and a non-normal distribution of residuals but not autocorrelation, indicating that more data and improved model refinement are required to improve predictive accuracy. Although analysis reveals a general increase in deposit amounts

over time, the SARIMA model has trouble correctly identifying data peaks. The model may have shortcomings in capturing seasonal patterns or irregular swings, as evidenced by its inability to closely match actual deposit trends, particularly during peak activity periods. Although the SARIMA model offers a basic forecast, this misalignment indicates that its precision at important moments is currently restricted, indicating the need for adjustment or the addition of more variables to increase its predictiveness.

Some other basic statistics we find are mentioned below:

Results of Payments Required with respect to Postal Code and City.

postalCode	city	
L1E0B7	Courtice	681797.37
L1E3J9	Courtice	635501.15
L1E2B5	Courtice	539948.25
L1E0H7	Courtice	511160.05
L1H7T4	Oshawa	490772.62
L1H8A3	Oshawa	478356.80
L1E0J2	Courtice	461145.14
L1E3E1	Courtice	456122.74
L1E0J6	Courtice	454402.69
L1E2B6	Courtice	448499.67
L1E0H5	Courtice	432357.32
L1H7X9	Oshawa	431479.81
L1E3H4	Courtice	427756.07
L1H7X6	Oshawa	422062.16
L1H8E5	Oshawa	416776.24
L1H8A4	Oshawa	390720.77
L1E3J7	Courtice	385690.64
L1H7W5	Oshawa	381047.42
L1E3H1	Courtice	380726.93
L1E2E6	Courtice	371655.60

Figure 14

Charges of each doctor along with their productive units and downtime.

Grouped by Doctors	provider_name	charge	productive_units	downtime_units
4	SYL – Soo Young Lee	0	6	0
2	LS – Lorna Simic	7632	104	55
0	Guest – Guest Hygiene	13506	324	86
1	IM – Isabella Morrison	55811	1336	498
3	SB – Dr. Sharon Barr	80917	862	210
5	ZW – Dr. Zuhair Waheed	99218	1033	285

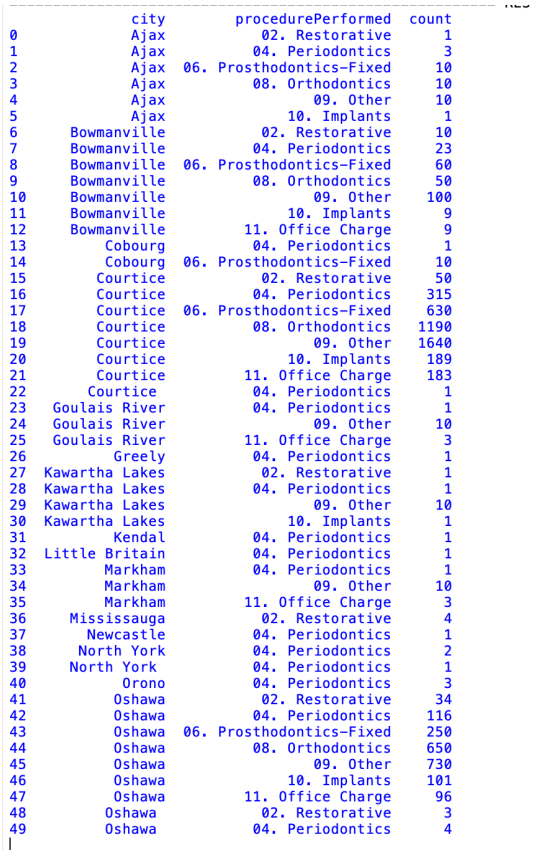
Figure 15

Charges with respect to the Department along with Productive units and downtime.

Grouped by Department:	department	charge	productive_units	downtime_units
0	hygiene	76949	1770	639
1	resto	180135	1895	495

Figure 16

Count of Medical Procedure Performed with respect to City.



	city	procedurePerformed	count
0	Ajax	02. Restorative	1
1	Ajax	04. Periodontics	3
2	Ajax	06. Prosthodontics-Fixed	10
3	Ajax	08. Orthodontics	10
4	Ajax	09. Other	10
5	Ajax	10. Implants	1
6	Bowmanville	02. Restorative	10
7	Bowmanville	04. Periodontics	23
8	Bowmanville	06. Prosthodontics-Fixed	60
9	Bowmanville	08. Orthodontics	50
10	Bowmanville	09. Other	100
11	Bowmanville	10. Implants	9
12	Bowmanville	11. Office Charge	9
13	Cobourg	04. Periodontics	1
14	Cobourg	06. Prosthodontics-Fixed	10
15	Courtice	02. Restorative	50
16	Courtice	04. Periodontics	315
17	Courtice	06. Prosthodontics-Fixed	630
18	Courtice	08. Orthodontics	1190
19	Courtice	09. Other	1640
20	Courtice	10. Implants	189
21	Courtice	11. Office Charge	183
22	Courtice	04. Periodontics	1
23	Goulais River	04. Periodontics	1
24	Goulais River	09. Other	10
25	Goulais River	11. Office Charge	3
26	Greely	04. Periodontics	1
27	Kawartha Lakes	02. Restorative	1
28	Kawartha Lakes	04. Periodontics	1
29	Kawartha Lakes	09. Other	10
30	Kawartha Lakes	10. Implants	1
31	Kendal	04. Periodontics	1
32	Little Britain	04. Periodontics	1
33	Markham	04. Periodontics	1
34	Markham	09. Other	10
35	Markham	11. Office Charge	3
36	Mississauga	02. Restorative	4
37	Newcastle	04. Periodontics	1
38	North York	04. Periodontics	2
39	North York	04. Periodontics	1
40	Orono	04. Periodontics	3
41	Oshawa	02. Restorative	34
42	Oshawa	04. Periodontics	116
43	Oshawa	06. Prosthodontics-Fixed	250
44	Oshawa	08. Orthodontics	650
45	Oshawa	09. Other	730
46	Oshawa	10. Implants	101
47	Oshawa	11. Office Charge	96
48	Oshawa	02. Restorative	3
49	Oshawa	04. Periodontics	4

Figure 17

Discussion

The research delved into integrating data science methods into dental offices to utilize the large amounts of data produced for improving patient care and administrative productivity. The goals were carefully established to evaluate the possibility, uses, obstacles, and the revolutionary effects of data science in dental environments. This project was based on a comprehensive research of existing literature which highlighted the possibilities of using data to transform dental care, the ethical considerations of converting it into digital format, and the potential benefits of utilizing artificial intelligence and big data to improve both clinical results and operational procedures.

One major obstacle faced in this project arose from the initial condition of the database and amount of the data. Exploring a vast collection of more than 200 database tables to determine the ones that are important for our research goals was a challenging task. As we started to unravel the complicated connections between these tables, the complexity increased, which was necessary for organizing our final datasets. Nevertheless, the triumph of conquering these obstacles was tinged with a sense of sadness. Following a thorough data cleaning procedure, we discovered that our dataset had drastically decreased in size. As we used our selected models, it became more clear that the limited data we had was not enough to produce meaningful results. This situation emphasized an important restriction of our project, overshadowing the ability to reach strong conclusions and emphasizing the key role of thorough, top-notch data in the success of machine learning efforts.

The findings provided valuable information, showing differences in costs and the factors impacting income, therefore helping understand how resources are allocated, appointments are scheduled, and finances are planned. Underlining the importance of comprehensive data collection was stressed to verify the precision of predictive models. Complex data analysis techniques like KMeans grouping.

Conclusion

In summary, this report provides important information on the possibility and effects of using data science methods in dental offices. By establishing a solid comprehension and recognizing the opportunities and challenges, it sets the stage for future research and the creation of effective plans for integrating data science into dental practices, ultimately enhancing patient care and improving administrative efficiency.

Although we have been able to extract meaningful patterns and helpful information from the dataset, yet, much of our work has been limited due to the database we have. There have been certain limitations which affected the performance of our models and analysis. For instance, as mentioned earlier, a large number of missing values caused severe problems to the accuracy of models. When we removed missing values, we were not left with sufficient data. Also, it was not possible to use other means of dealing with missing values such as replacing missing values by mean or median because some columns we were dealing with were categorical. Another problem we faced was, the database design and naming conventions of columns made it hard for us to join tables and make sense of some columns. One variable had different names across different tables so it took time to figure out the relations. In addition to it, database design needs

to be improved because a large number of tables are present in the database but most of them do not contain any data, and if some data is present in tables it is not useful. Also, representatives at dental clinics need to ensure that they enter complete and correct data into the database and do not leave any values missing. They will lead to better data which could further be used to make accurate analysis.

Studies in future can make use of this work to advance the research and further build on it. We believe that over time, the volume of data will increase which would be sufficient to build and train models on. Furthermore, making representatives at the clinic more responsible while entering data would help in ensuring that data entered into the database is suitable and valuable.

Bibliography

Schwendicke, F., & Krois, J. (2021). *Data Dentistry: How Data Are Changing Clinical Care and Research. Journal of Dental Research*, 101(1). https://doi.org/10.1177/00220345211020265

Uribe, S. E., Sofi-Mahmudi, A., Vilne, B., et al. (2021). *Dental Research Data Availability and Quality According to the FAIR Principles. Journal of Dental Research,* 101(11). https://doi.org/10.1177/002203452211013

Dhopte, A., & Bagde, H. (2023). *Smart Smile: Revolutionizing Dentistry with Artificial Intelligence. Cureus*, 15(6), e41227. https://doi.org/10.7759/cureus.41227

Surdilovic, D., & Ille, T. (2022). *Artificial Intelligence and Dental Practice Management. European Journal of Artificial Intelligence and Machine Learning*, 1. https://doi.org/10.24018/ejai.2022.1.3.8

Wynsor, G. (2023). *The future of dental practice management systems. British Dental Journal*, 235, 64. https://doi.org/10.1038/s41415-023-6102-4

Acharya, A., Schroeder, D., Schwei, K., & Chyou, P. H. (2017). *Update on Electronic Dental Record and Clinical Computing Adoption Among Dental Practices in the United States. Clinical Medicine & Research*, 15(3-4), 59–74. https://doi.org/10.3121/cmr.2017.1380

Wanyonyi, K. L., Radford, D. R., & Gallagher, J. E. (2021). *Electronic primary dental care records in research: A case study of validation and quality assurance strategies. International Journal of Medical Informatics*, 131, 104018. https://doi.org/10.1016/j.ijmedinf.2019.04.007

AbuSalim, S., Zakaria, N., Islam, M. R., Kumar, G., Mokhtar, N., & Abdulkadir, S. J. (2021). *Analysis of Deep Learning Techniques for Dental Informatics: A Systematic Literature Review. Healthcare*, 10(10), 1892. https://doi.org/10.3390/healthcare10101892

Song, M., Liu, K., Abromitis, R., & Schleyer, T. L. (2013). *Reusing electronic patient data for dental clinical research: a review of current status*. Journal of Dentistry, 41(12), 1148-1163. https://doi.org/10.1016/j.jdent.2013.04.006

Favaretto, M., Shaw, D., De Clercq, E., Joda, T., & Elger, B. S. (2020). *Big Data and Digitalization in Dentistry: A Systematic Review of the Ethical Issues*. International Journal of Environmental Research and Public Health, 17(7), 2495. https://doi.org/10.3390/ijerph17072495

Jain, P., Wynne, C. (2021). Artificial Intelligence and Big Data in Dentistry. In: Jain, P., Gupta, M. (eds) Digitization in Dentistry. Springer, Cham. https://doi.org/10.1007/978-3-030-65169-51

Appendix 1

```
## Importing necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns



## importing csv
df = pd.read_csv("data.csv")



df['date'] = pd.to_datetime(df['date'])

# Group the data by month and sum the charges for each month
monthly_data = df.groupby(df['date'].dt.day)['charge'].sum()


## plot for finding day to day charges of doctors
plt.bar(monthly_data.index, monthly_data.values, color='purple',
edgecolor='gray')
plt.xlabel("DAY OF MONTH")
plt.ylabel("CHARGES")
plt.title('Day to Day Charges of Doctors for 2-3 Month Time
Period')
plt.show()



## Total Charges, Productive Units, Downtime units grouped by
Department
g_d = df.groupby('department').agg({
    'charge': 'sum',
    'productive_units': 'sum',
    'downtime_units': 'sum'
}).reset_index()

print("Grouped by Department: ", g_d)
```

```
## Total Charges, Productive Units, Downtime units grouped by
Provider
g_d1 = df.groupby('provider_name').agg({
    'charge': 'sum',
    'productive_units': 'sum',
    'downtime_units': 'sum'
}).reset_index()

print("Grouped by Doctors", g_d1.sort_values(by=["charge"]))




df['date'] = pd.to_datetime(df['date'])
print(df["department"].value_counts())


## Boxplot for checking outliers
plt.boxplot(df['charge'])
plt.title("Boxplot of Charges")
plt.show()


## boxplot for available units to check outliers
plt.boxplot(df['available_units'])
plt.title("Boxplot of Charges")
plt.show()

## Making scatterplot for productive units and charges
plt.scatter(df['productive_units'],df['charge'])
plt.xlabel("Productive Units")
plt.ylabel("Charges")
plt.title("Scatter Plot Productive Units vs Charges")
plt.show()


## Scatterplot to show relation between available units and
charges
plt.scatter(df['available_units'],df['charge'])
plt.xlabel("available Units")
plt.ylabel("Charges")
plt.title('Scatter Plot of Charge vs Available Units')
plt.show()
```

```
## Using a plot using seasborn library. To show relation between
productive units and charges with respect to department
sns.scatterplot(data= df, x = "productive_units", y = "charge",
hue = "department")
plt.xlabel("Productive Units")
plt.ylabel("Charges")
plt.title('Scatter Plot of Charge vs Productive Units in terms
of Department')
plt.legend(title = "department")
plt.show()


## Using a plot using seasborn library. To show relation between
productive units and charges with respect to doctor name
sns.scatterplot(data= df, x = "productive_units", y = "charge",
hue = "provider_name")
plt.xlabel("Productive Units")
plt.ylabel("Charges")
plt.title('Scatter Plot of Charge vs Productive Units in terms
of Provider Name')
plt.legend(title = "provider_name")
plt.show()



## Using a plot using seasborn library. To show relation between
available  units and changes with respect to doctor name
sns.scatterplot(data= df, x = "available_units", y = "charge",
hue = "provider_name")
plt.xlabel("Available Units")
plt.ylabel("Charges")
plt.title('Scatter Plot of Charge vs Available Units in terms of
Provider Name')
plt.legend(title = "provider_name")
plt.show()
```

Appendix 2

```
## importing needed libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
from sklearn.metrics import mean_squared_error
import matplotlib.colors as mcolors
from statsmodels.tsa.statespace.sarimax import SARIMAX
import pmdarima as pmd


## import csv
data1 = pd.read_csv("BankDeposit.csv")




################### MONTH TO MONTH DEPOSIT BETWEEN 12/10/2021 - 07/03/2024
##################

## getting date in right format
data1['Date'] = pd.to_datetime(data1['Date'], format='%d/%m/%Y %H:%M')
### Sorting the data frame by date
data1.sort_values(by='Date', inplace=True)

# Grouping the data month by month and gettings sum of Deposit received
monthly = data1.groupby(data1['Date'].dt.month)['Deposit'].sum()

##  Bar plot for amounts earned month by month
plt.bar(monthly.index, monthly.values, color='orange', edgecolor='gray')
plt.xlabel("Month")
plt.ylabel("Deposited Amount (CAD)")
plt.title('Month to Month Deposited Amount')
plt.show()
##################



######## Day to Day DEPOSIT BETWEEN 12/10/2021 - 07/03/2024 ##################
daily = data1.groupby(data1['Date'].dt.day)['Deposit'].sum()
plt.bar(daily.index, daily.values, color = 'brown')
plt.xlabel("Day")
plt.ylabel("Deposited Amount (CAD)")
plt.title('Day to Day Deposited Amount')
plt.show()
###################################
```

```
###################################
###  Finding top paying Patients
cust = data1.sort_values(by= "Deposit", ascending = False)

cols = ["Date","Deposit", "LastPatient"]
print("Deposited Amounts : ",cust[cols].head(10))

###################################



## Normalizing data. I have used min-max normalization
min_deposit = data1["Deposit"].min()
max_deposit = data1["Deposit"].max()
data1["Normalized_Deposit"] = (data1["Deposit"] - min_deposit) / (max_deposit -
min_deposit)

## separated the column on which I need to apply time series models.
data = data1[['Date','Deposit']]



# Plotting Deposits received over time
plt.plot(data['Date'], data['Deposit'], marker='o', linestyle='-')
plt.title('Deposit Over Time')
plt.xlabel('Date')
plt.ylabel('Deposit Amount')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
#plt.show()


################ APPLYING ARIMA Time Series Model

## Defining Parameters
p = 10
d = 1
q = 0

## Defining model and assigning parameters
model = ARIMA(data["Deposit"], order =(p,d,q))
model_fit = model.fit()

print(model_fit.summary())

# Plotting residuals
resi = pd.DataFrame(model_fit.resid)
resi.plot(title='Residuals')
plt.show()


# Making predictions on fitted model
```

```
predictions = model_fit.predict(start=data.index[0], end=data.index[-1],
typ='levels')

# Plotting actual vs predicted values using ARIMA model
plt.plot(data['Date'], data['Deposit'], label='Actual')
plt.plot(data['Date'], predictions, label='Predicted', color='red')
plt.title('ARIMA Actual vs Predicted')
plt.xlabel('Date')
plt.ylabel('Deposit Amount')
plt.legend()
plt.show()



#
### Evaluating the model using Root Mean Squared error & Mean Squared Error
rmse = np.sqrt(mean_squared_error(data["Deposit"], predictions))
print("ARIMA RMSE:", rmse)

mae = np.mean(np.abs(data["Deposit"] - predictions))
print("ARIMA MAE:", mae)




############## APPLYING SARIMA (Seasonal Time series model)

## Setting Parameters for SARIMA Model
p1, d1,q1 = 0,0,0
p2,d2,q2,s2 = 5,0,2,12    ## Seasonal parameters

## Defining the model
sarima = SARIMAX(data["Deposit"], order = (p1,d1,q1), seasonal_order=
(p2,d2,q2,s2))
fit = sarima.fit()
pred = fit.predict()

## plotting actual vs predicted values on SARIMA model
plt.plot(data['Date'], data['Deposit'], label='Actual')
plt.plot(data['Date'], pred, label='Predicted', color='orange')
plt.title('SARIMA Actual vs Predicted')
plt.xlabel('Date')
plt.ylabel('Deposit Amount')
plt.legend()
plt.show()


### Evaluating the model using Root Mean Squared error & Mean Squared Error
s_rmse = np.sqrt(mean_squared_error(data["Deposit"], pred))
print("SARIMA RMSE:", s_rmse)


s_mae = np.mean(np.abs(data["Deposit"] - pred))
print("SARIMA MAE:", s_mae)
```

Appendix 3

```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency

## Here we are working with two different data tables

## dataframe 1
df = pd.read_csv("patient-procedures.csv")

## dataframe 2
df2= pd.read_csv("patient-billing.csv")
#print(df.head())


## PROCEDURES PERFORMED WITH RESPECT TO CITY
gp = df.groupby(['city', 'procedurePerformed'])
gp1 = gp.size().reset_index(name='count')
print(gp1.head(10))



## FINDING PAYMENT REQUIRED W.R.T CITY AND POSTAL CODE
res = df2.groupby(["postalCode",
"city"])["paymentRequired"].sum().sort_values(ascending = False)
print(res.head(20))




####### PERFORMING CHI-SQUARE BETWEEN POSTAL CODE AND PROCEDURE
PERFORMED
##

y = df['postalCode'].str[:2].dropna()  ## using only first 3
characters of postal code because complete postal code gives
unreasonable p-value (p=1)
x = df["procedurePerformed"].dropna()

## Creating frequency table
table = pd.crosstab(x,y)

## performing chi test
```

```
chi_val, p_val, degree, exp_freq =
chi2_contingency(table.values)
print("Chi-Squared Value :", chi_val)   ## chi-square value
print("P-Value: ", p_val)  ## p-value of the test
print("Degrees of Freedom", degree)  ## degrees of freedom


##########################################################

## FINDING PROCEDURES BY MONTH

df['procedureDate'] = pd.to_datetime(df['procedureDate'])

# Getting the month name from the procedureDate
df['Month'] = df['procedureDate'].dt.month_name()


months_list = ['January', 'February', 'March', 'April', 'May',
'June', 'July', 'August', 'September', 'October', 'November',
'December']

# grouping data by procedure performed, and month and counting
occurrence of each procedure
gb =
df.groupby(['Month','procedurePerformed',]).size().unstack(fill_
value=0)

## it was randomly ordering months in plot, so I have specify
order of month
gb = gb.reindex(months_list)

# Making bar graph
gb.plot(kind='bar')
plt.title('Procedures Performed by Month')
plt.xlabel('Month')
plt.ylabel('Count')
plt.legend(title='Procedure')
plt.xticks(rotation=45)
plt.show()

##########################################################
```

Appendix 4

```
import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score
import numpy as np
from sklearn.neighbors import KNeighborsClassifier


# importing csv
df = pd.read_csv("tbl_ChartPerio.csv")

# Removing Null values  in specific columns
data = df.dropna(subset=["fld_strClassification", "fld_strHygiene",
"fld_strCalculus", "fld_strExamType"])

# Specifying labels and features
my_label = data["fld_strClassification"]  ## label column for
classification algorithm
my_features = data[["fld_strHygiene", "fld_strCalculus",
"fld_strExamType"]]  ## features


## To check what each label is getting encoded to.
print("Before encoding", my_label[:5])

# Encoding labels
l= LabelEncoder()
my_label = l.fit_transform(my_label)

## To check what each label is getting encoded to.
print("After encoding", my_label[:5])

### encoding features
for i in my_features.columns:
    my_features[i] = LabelEncoder().fit_transform(my_features[i])


# Dividing data into training and testing. 20% for testing, 80% for
training

X_t, X_test, y_t, y_test = train_test_split(my_features, my_label,
test_size = 0.2, random_state = 34)
```

```
### Defining decision tree classification algorithm
decision_tree = DecisionTreeClassifier()

## Fitting classifier
decision_tree.fit(X_t, y_t)

# Making predictions
pred = decision_tree.predict(X_test)

# Evaluating performance of Model
acc = accuracy_score(y_test, pred)
print("Accuracy of Decision Tree:", acc)

##############

### APPLYING KNN CLASSIFIER

## Dividing data into testing and training. 85% for training and 15%
for testing. Model didn't perform good with 80% training data
X_tr, X_te, y_tr, y_te = train_test_split(my_features, my_label,
test_size=0.15, random_state=28)



# Defining Classifier with K =5
## I tried different k values and checked accuracy. K <= 5 gives
suitable accuracy. So I chose  K=5
knn = KNeighborsClassifier(n_neighbors=5)

# Fitting the model
knn.fit(X_tr, y_tr)

# Making predictions
predic = knn.predict(X_te)

# Finding accuracy of model
accuracy = accuracy_score(y_te, predic)
print("Accuracy of KNN: ", accuracy)

### Making prediction on my own chosen value

lst = [3,3,1]
ipt = np.array(lst).reshape(1,-1)
res = knn.predict(ipt)   ## predicting on KNN model
res2 = decision_tree.predict(ipt)  ## prediction on Decision Tree
model

print("my prediction", res, res2) ## Both give same results
```
