IST 782 Portfolio Milestone

Name: Lu Guo

SUID: 292001282

Email: lguo15@syr.edu

Expected Graduation: May 2024

Program: M.S. of Applied Data Science

School of Information Studies

Syracuse University

1. Introduction

The Master's program in Applied Data Science is offered by the School of Information Studies at Syracuse University. It emphasizes the applications of data science to enterprise operations and processes, especially in the areas of data collection and management, data analysis, strategy and decisions, and implementation.

The program provides courses to develop students' proficiency in data science and teaches tools and methods in the data lifecycle:

- Data Collection and Processing:
  Python is taught in IST 652 (Scripting for Data Analysis) to collect and clean data. IST 718 (Big Data Analytics) teaches ETL and Pyspark to process data.
- Data Management:
  IST 659 (Data Administration Concepts and Database Management) is mainly about SQL and Azure to organize and query data.
- Data Analysis:
  IST 772 (Quantitative Reasoning for Data Science) teaches statistical analyzing methods. IST 687 (Introduction to Data Science) teaches traditional data analysis methods, e.g., sampling, linear model, decision tree, SVM, association rules mining et al. IST 707 (Applied Machine Learning), IST 691 (Deep Learning in Practice), and IST 736 (Text Mining) teach machine learning and deep learning to analyze more complex data like text and images.
- Data Visualization:

IST 719 (Information Visualization) teaches R and Adobe Illustrator to analyze and visualize data. SCM 651 (Business Analytics) teaches advanced Excel, PowerBI, and Tableau to visualize data.

- Decision Making:
  IST 687 (Introduction to Data Science), SCM 651 (Business Analytics), IST 719 (Information Visualization), and some other courses teach how to interpret analysis results and generate actionable insights.

I will introduce some of the courses I have taken to demonstrate the skills I have learned: IST 719 - Data Visualization, IST 659 - Database Administration & Database Management, and IST 736 - Text Mining.

2. IST 719 - Data Visualization

Under the direction of Professor Jeff Hemsley, I learned to use R code to analyze data and draw graphs to visualize the results. We also use Adobe Illustrator to modify graphs and create posters.

2.1 Project

(1) Project Introduction

This project's name is Data Scientist Job Information Analysis (Guo, 2023). As master's students in Applied Data Science, we are aiming for jobs as data scientists. Knowing more about these positions can give us valuable insights for job search. Master students in Applied Data Science or other students who

want to be a data analyst and data scientist may be interested.

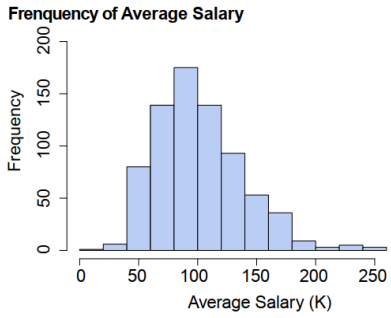
(2) Data

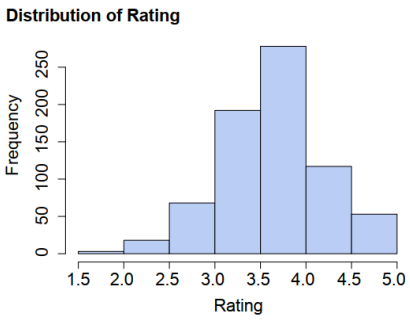
This data set is about the job information of data scientists and data analysts in 2021, with a total of 742 rows and 42 columns. The columns include job title, salary, rating, location, size, type of ownership, and skills. The data set is downloaded from Kaggle: https://www.kaggle.com/datasets/nikhilbhathi/data-scientist-salary-us-glassdoor.

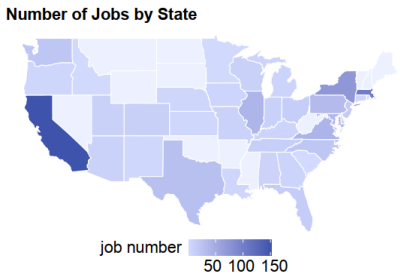
For the data cleaning, I used cosine similarity to divide the job titles into three types: data scientist, data engineer, and data analyst.

(3) Data Exploration

I use the R code to explore the distribution of the data set.



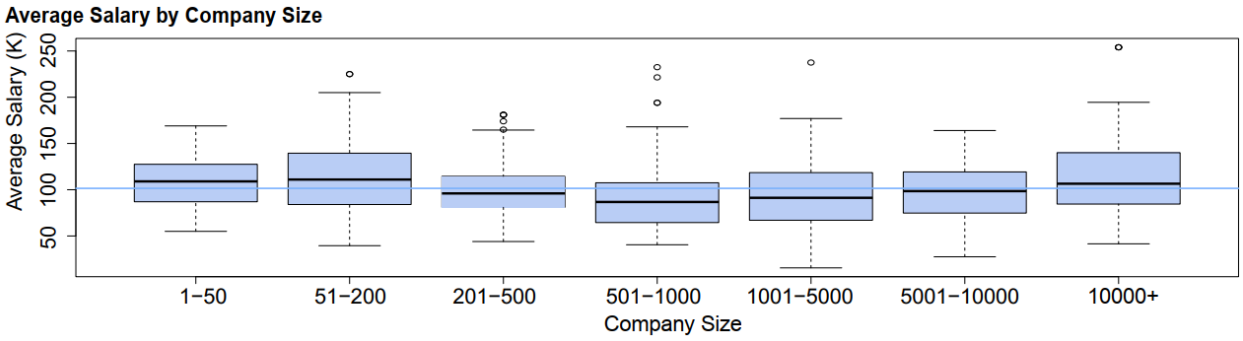




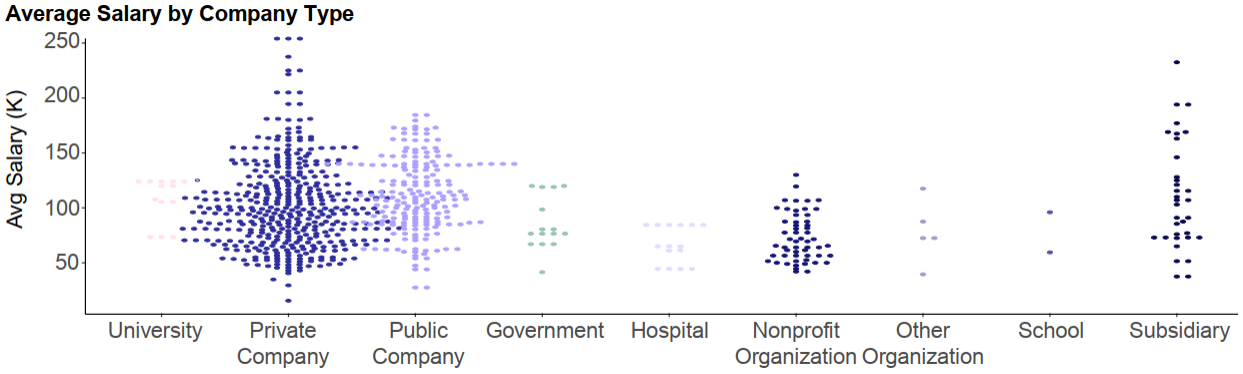
Most jobs are located in CA, MA, and NY states.

(4) Data Analysis

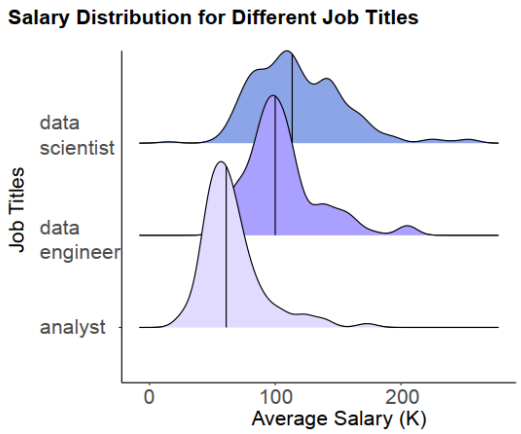
I try to answer two questions: What's the salary in different kinds of companies? What skills are required?

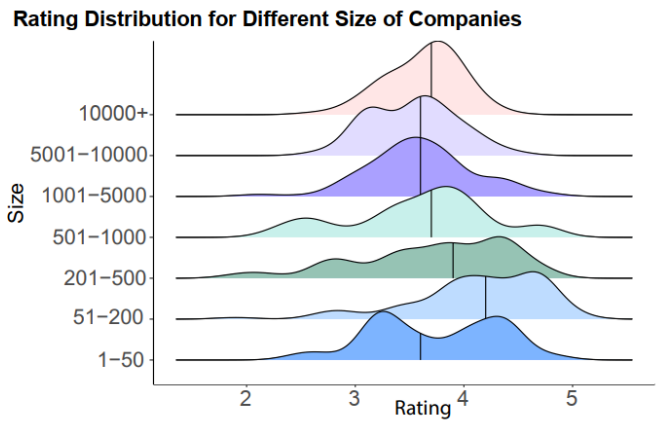


Companies with 51-200 employees offer top median salaries. Companies with 501-1000 employees offer the lowest median salaries.



Most positions are offered by private and public companies. The subsidiary companies offer higher average salaries.

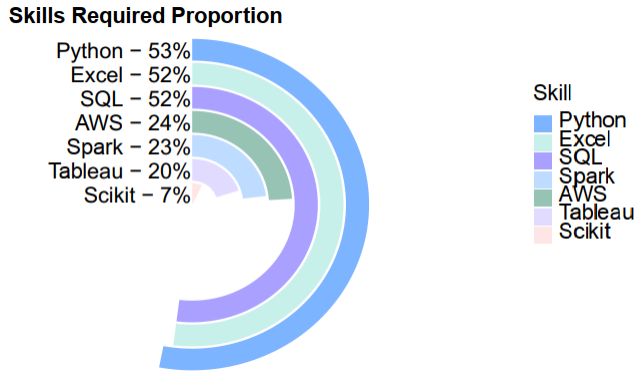




The average salary of data analysts, data engineers, and data scientists increases in order.

Companies with 51-200 employees receive the highest ratings.

Companies with 1-50 employees receive the lowest ratings.

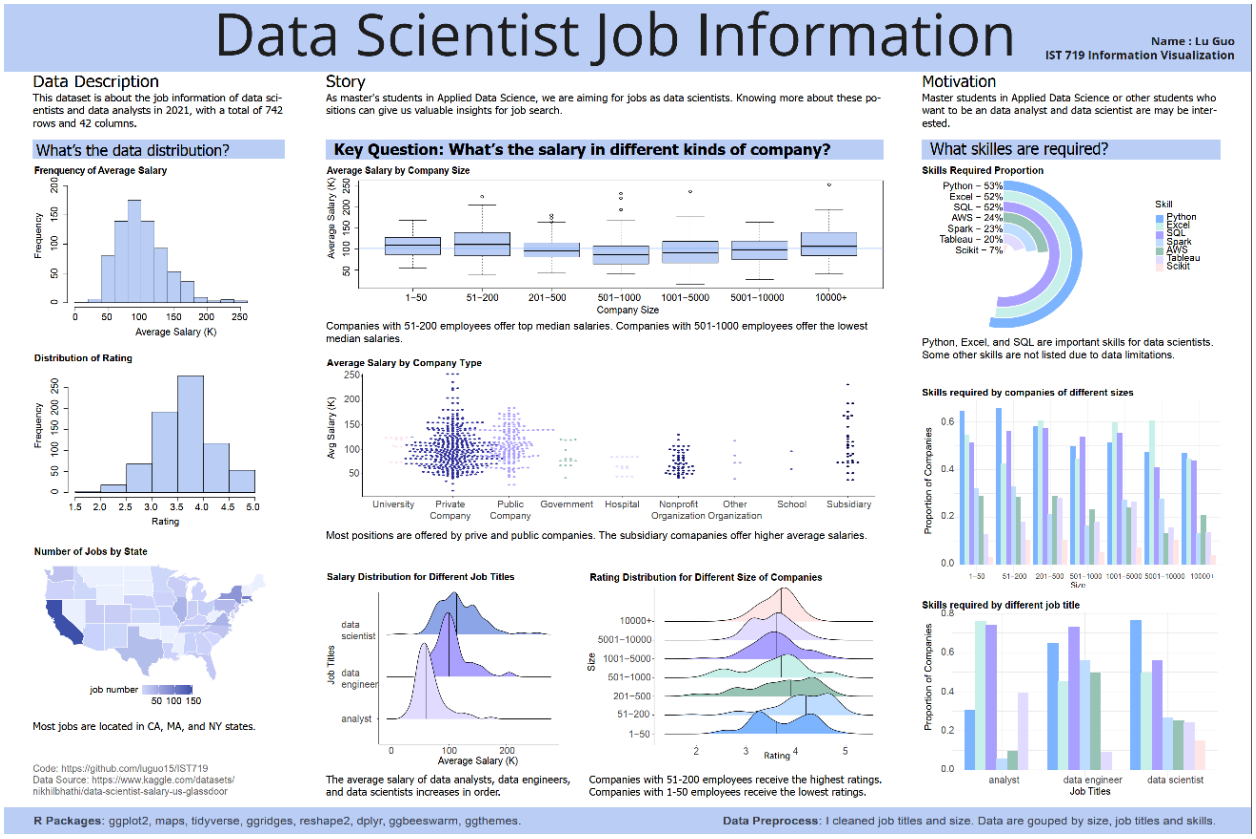


Python, Excel, and SQL are important skills for data scientists.

Some other skills are not listed due to data limitations.

(5) Data Visualization

I organized these results in a poster by Adobe Illustrator:



(6) Actionable Insights

CA, MA, and NY states have more data-related job opportunities. If we want to be a data scientist, data engineer, or data analyst, we can consider companies in these three states.

Companies with 51-200 employees offer top median salaries and receive the highest ratings. When applying for data-related positions, we can consider companies of this size.

If we want to be a data analyst, the most important skills to prepare are Excel and SQL. Python, Excel, SQL, Spark, and AWS are important for

data engineers. Python, Excel, and SQL are important for data scientists. Make sure you have these skills ready.

2.2 Summary

From this project, I learned how to use R code to clean data, analyze data (draw different graphs), and use Adobe Illustrator to create posters to visualize data.

3. IST 659 - Database Administration & Database Management

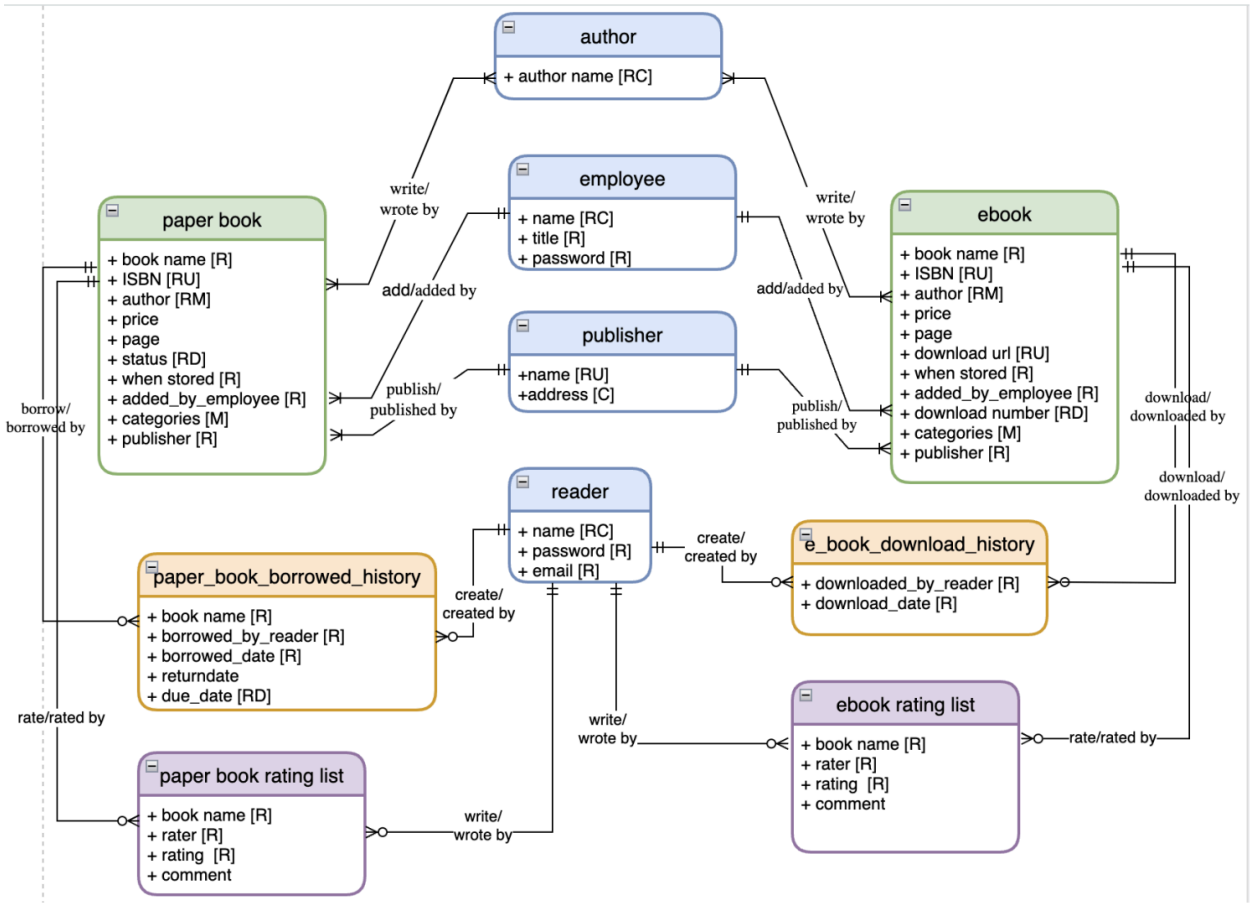
This course taught us to design databases, implement databases, and manage databases. The content contains advanced SQL, including Window Functions, Common Table Expressions (CTEs), Subqueries, Transactions and Concurrency Control, and Triggers.

3.1 Project

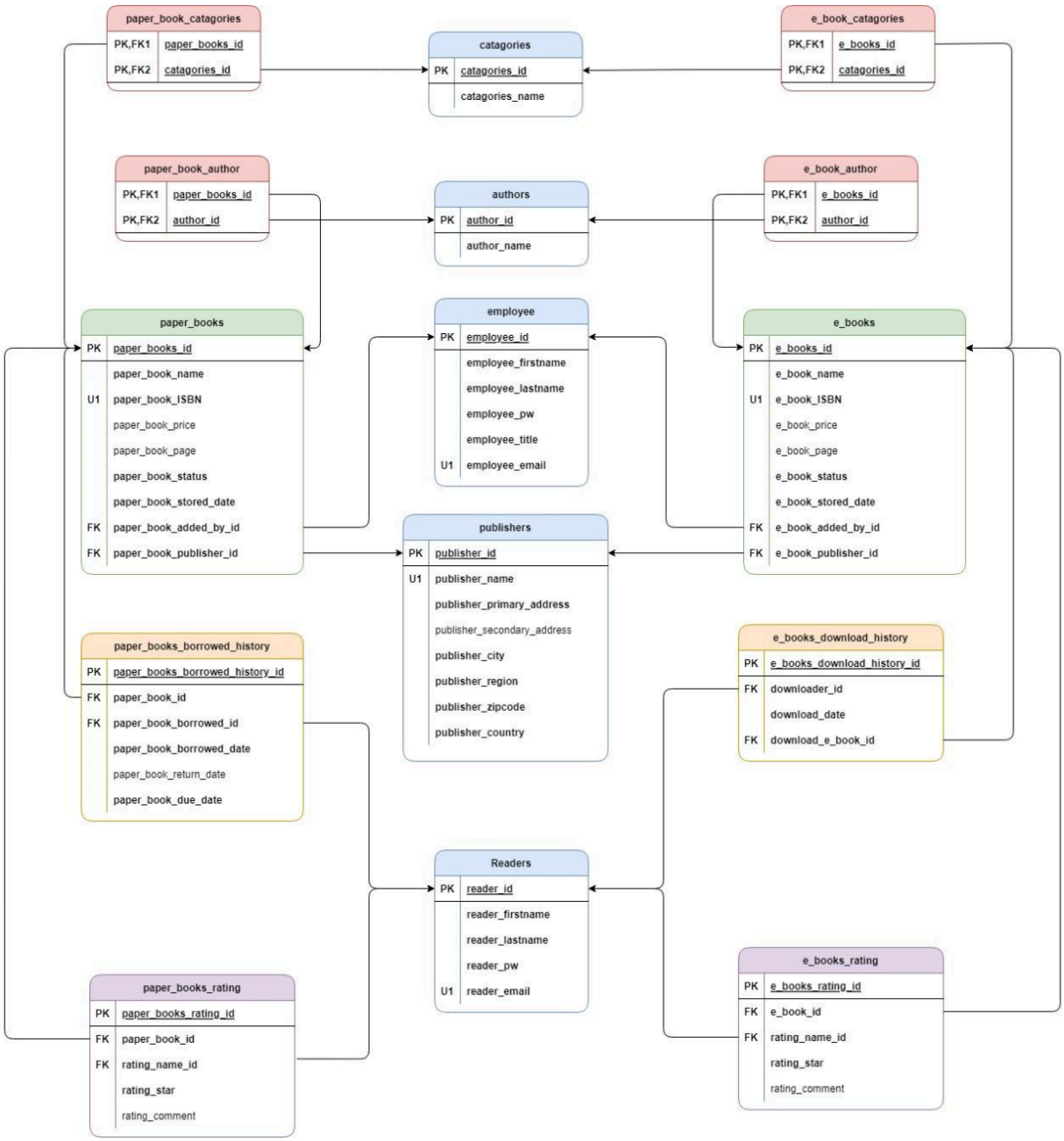
(1) Introduction

This is a team project. We established a new library management system, which encapsulates different books into data objects and provides various public access attributes by calling different data functions to meet the needs of users (Guo et al., 2022).

(2) Conceptual Data Model Diagram



(3) Logical Data Model Diagram



(4) Sample code to create tables

```
create table paper_books_rating(
    paper_book_rating_id int IDENTITY not null,
    paper_book_id int not null,
    rating_name_id int not null,
    rating_star int not null,
    rating_comment varchar(50),
    constraint pk_paper_books_rating_paper_book_rating_id primary key(paper_book_rating_id)

)
```

(5) Sample code to create procedures

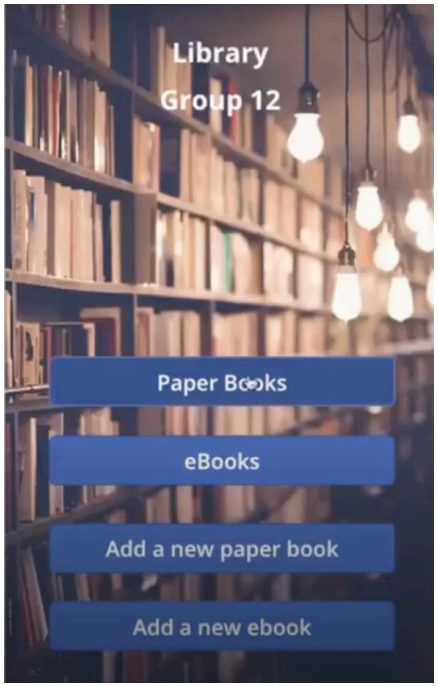
```
use library
GO

drop PROCEDURE if EXISTS dbo.p_insert_reader
drop PROCEDURE if EXISTS dbo.p_insert_employee
drop PROCEDURE if EXISTS dbo.p_insert_author
drop PROCEDURE if EXISTS dbo.p_insert_category
drop PROCEDURE if EXISTS dbo.p_insert_paper_book
drop PROCEDURE if EXISTS dbo.p_insert_e_book
drop PROCEDURE if EXISTS dbo.p_insert_paper_book_comment
drop PROCEDURE if EXISTS dbo.p_insert_e_book_comment
drop PROCEDURE if EXISTS dbo.p_borrowed_paper_book
drop PROCEDURE if EXISTS dbo.p_borrowed_e_book




go
create PROCEDURE dbo.p_insert_reader(
    @firstname varchar(50),
    @lastname varchar(50),
    @pw varchar(50),
    @email VARCHAR(50)

)
as BEGIN
    begin TRY
        begin TRANSACTION
        if exists(select * from readers where reader_email = @email)
        throw 50002, 'p-insert_reader: email exist',1

        else BEGIN

            insert into readers ( reader_firstname,reader_lastname,reader_pw,reader_email)
                VALUES (@firstname,@lastname,@pw,@email)
            if @@ROWCOUNT <> 1 throw 50002, 'p-insert_reader: Insert Erroe',1
        END
        COMMIT
    end try
    begin catch
         rollback;
         throw
    end CATCH
end
```

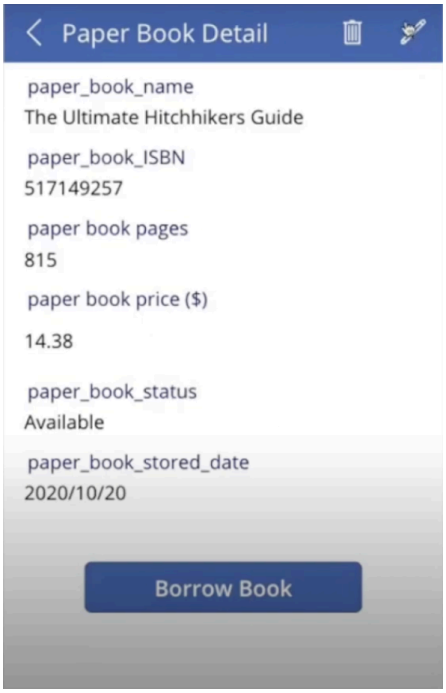
(6) Demo

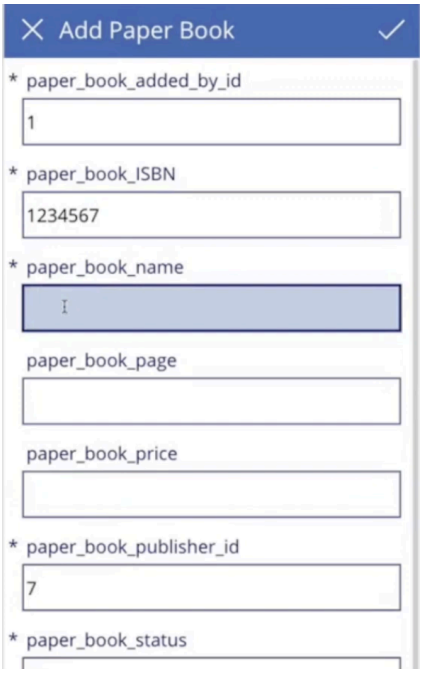
This is the link to the demo display:

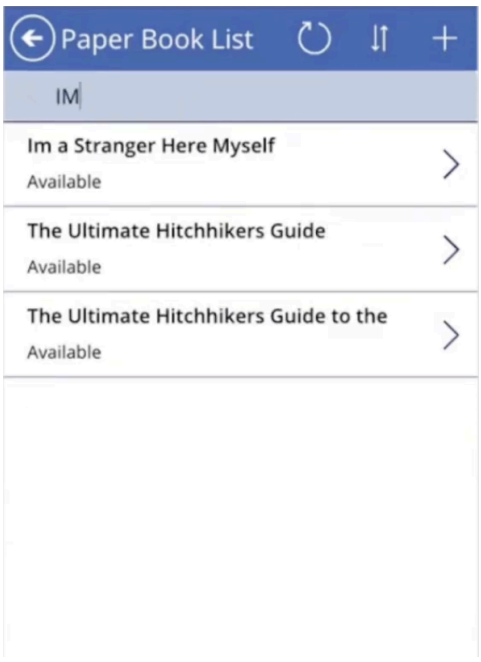
https://www.youtube.com/watch?v=SpxTjvtaCY4

We can search for book information, borrow books, and edit book information (manager only).









3.2 Summary

From this project, I learned how to design a database and implement a database. I also learned to query data from a database by SQL, the result can be used to do data analysis.

4. IST 736 - Text Mining

4.1 Project

(1) Introduction

Palliative care aims to improve seriously ill patients' and their family members' quality of life. The 4th edition of Clinical Practice Guidelines for Quality Palliative Care was released by The National Consensus Project for Quality Palliative Care in 2018. The updated guideline expands the offer of palliative care to all people with serious illnesses, regardless of diagnosis, prognosis, or age (Staff, n.d., 2023).

To understand palliative care from patients and their caregivers' perspectives, we analyze user-generated content from Reddit via natural language learning methods to investigate their knowledge, stance, and experiences of palliative care (Guo and Liu, 2023). Three research questions are proposed:

RQ1: What topics of palliative care have people posted on Reddit?

RQ2: What are users' attitudes/stances toward palliative care?

RQ3: Are they looking for and providing emotional support on online group forums?

(2) Data

We collect data from six cancer subreddits (testicularcancer, subredditcancer, cancer, breastcancer, braincancer, and thyroidcancer) until 2022-12-30, the data is available at https://atlantis.ischool.syr.edu/share/reddit/cancer/. Each subreddit includes submissions, comments, and related metadata. We used the keyword “palliative” to filter data, 596 submissions, and 2815 comments were left.

(3) Data Exploration

Image 1 is the data distribution trend, it shows that the discussion about palliative care significantly increased since 2018. The increase coincides with the release of the 4th edition of the palliative care guideline that extended the scope to all patients with serious illnesses (Staff, n.d., 2023).

Time trend analysis

```
[ ]:
import matplotlib.pyplot as plt

# Extract year from the datetime column in comments dataframe
merged_subset_2['year_month'] = merged_subset_2['created_utc'].dt.to_period('Y')

# Group by year and count the occurrences in comments dataframe
time_distribution_comments = merged_subset_2.groupby('year_month').size()

# Extract year from the datetime column in submissions dataframe
merged_subset_1['year_month'] = merged_subset_1['created_utc'].dt.to_period('Y')

# Group by year and count the occurrences in submissions dataframe
time_distribution_submissions = merged_subset_1.groupby('year_month').size()

# Plot the line chart for comments
time_distribution_comments.plot(kind='line', marker='o', figsize=(10, 6), label='Comments')

# Plot the line chart for submissions
time_distribution_submissions.plot(kind='line', marker='o', figsize=(10, 6), label='Submissions')

plt.title('Time Distribution of Limit Comments and Submissions')
plt.xlabel('Year')
plt.ylabel('Number')
plt.legend()
plt.grid(True)
plt.show()
```

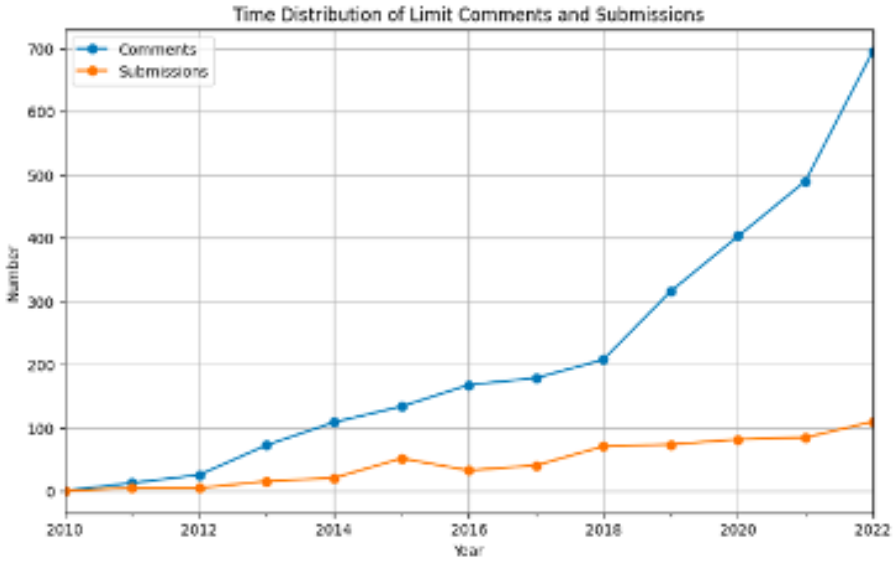


Image 1: Data distribution between 2010 to 2022.

(4) Data Analysis

To answer RQ1, we used Topic Modeling methods sentence_embedding+kmeans to cluster submissions and comments individually. The results show that seven topics are in the submissions: 1)

emotional distress after diagnosis, 2) challenges, 3) coping with a loved one with a terminal illness, 4) seeking alternative treatments during palliative care, 5) seeking advice of receiving palliative care or not, 6) seeking emotional support, and 7) seeking people with similar experiences. And five topics in the comments: 1) palliative care medications and pain management, 2) navigating palliative and hospice care options, 3) support and decision-making, 4) palliative care experiences, and 5) emotional support.

In summary, concerns related to palliative care mainly include mental health challenges, medical decision suggestion seeking, and emotional support seeking from Reddit data.

To answer RQ2, two authors separately annotated 20 submissions and 30 comments that were randomly selected and reached an agreement on different annotations after discussion. One irrelevant comment is removed. We use four labels: positive, negative, neutral, and cannot tell. For submissions, 20% (n=4) are positive, 65% (n=13) are neutral, 10%(n=2) cannot tell, and none (n=0) is negative. For comments, 62% (n=18) are positive, 3% (n=1) are negative, 24% (n=7) are neutral, and 10% (n=3) cannot tell (Table 1).

	submissions	comments
positive	4 (20%)	18 (62%)
negative	0	1 (3%)

neutral	13 (65%)	7 (24%)
cannot tell	2 (10%)	3 (10%)
Total	20	29

Table 1: Annotation results of submissions

We use ChatGPT to classify the stance. The prompt we used is "What's the stance towards palliative care in the below submission/comment from Reddit? Answer positive, negative, neutral, or cannot tell as accurately as possible. The submission/comment is: ###{}###". The macro-F1 score is 0.39, which is not very good. Previous research has shown that using the internal label of ChatGPT can improve the macro-F1 score (Kim et al., 2023). To get a higher macro-F1 score, we try to find labels from the internal of ChatGPT. We asked ChatGPT, "What labels are usually used in a stance classification task?" ChatGPT responded to four labels "Support, Oppose, Neutral, Unrelated." Then We revised our prompt to "What's the stance towards palliative care in the below submission/comment from Reddit? Answer Support, Oppose, Neutral, or Unrelated as accurately as possible. The submission/comment is: ###{}###". The macro F-1 score decreased to 0.33. The results show that stance classification does not work well for our data, we may try different models in the future.

To answer RQ3, one of the authors first manually annotated 121 submissions for training and evaluation purposes. Annotation results showed that 31% (n=37) of submissions sought emotional support and 69% (n=84) did not.

We then build SVM and BERT models to classify whether submissions seek or support emotional support. For SVM, we removed stop words and used a Boolean vectorizer. The macro F1 score of SVM is 0.61 and BERT is 0.38. The performance of the SVM model is better than the BERT model. The main error of the BERT model is that it predicts most sentences to have "no emotional support".

(5) Conclusion

Topic modeling results for the six cancer subreddits indicated that caregivers experienced severe stress during palliative care. The main sources of stress are the fear of the death of a loved one and the additional tasks that come with palliative care. Caregivers actively seek information, tools, and emotional support from others who have been through similar experiences. One specific use of these Reddit subreddits is to vent. Users vent their negative experiences and feelings in a group of others who have similar experiences, so they know they are understood and welcomed.

4.2 Summary

In this project, I used to analyze text data by NLP technologies. I used Python to clean text. I used topic modeling to analyze the content, and classification to decide the stance of the content. In the future, I can deal with text data by these methods.

5. Conclusion

From the courses in the Applied Data Science Project, I learned to scrawl data by R and Python programming, and then clean the data. The cleaned data can be stored and managed in a database. I learned to analyze and visualize data by R and Python. I also learned advanced methods including machine learning, deep learning, NLP, and LLMs to analyze complex data. I am equipped with the skills to become a data scientist in the future.

Thanks to all the professors, staff, advisors, and classmates for helping me.

6. References

Guo, L. (2023). IST 719: Data Visualization [GitHub repository]. Retrieved March 11, 2024, from https://github.com/luguo15/IST782-Portfolio/tree/main/Portfolio/IST%20718%20data%20visualization

Guo, L., Tsai, C., & Feng, P. (2022). IST 659: Database Administration & Database Management [GitHub repository]. Retrieved March 11, 2024, from https://github.com/luguo15/IST782-Portfolio/tree/main/IST%20659%20Database

Staff, A. I. W. (n.d.). Palliative care guidelines updated with increased focus on collaboration, communication. ACP Internist Weekly. Retrieved December 12, 2023, from https://acpinternist.org/weekly/archives/2018/11/06/4.htm

Guo, L., & Liu, X. (2023). IST 736: Text Mining [GitHub repository]. Retrieved March 11, 2024, from https://github.com/luguo15/IST782-Portfolio/tree/main/IST%20736%20Text%20Mining