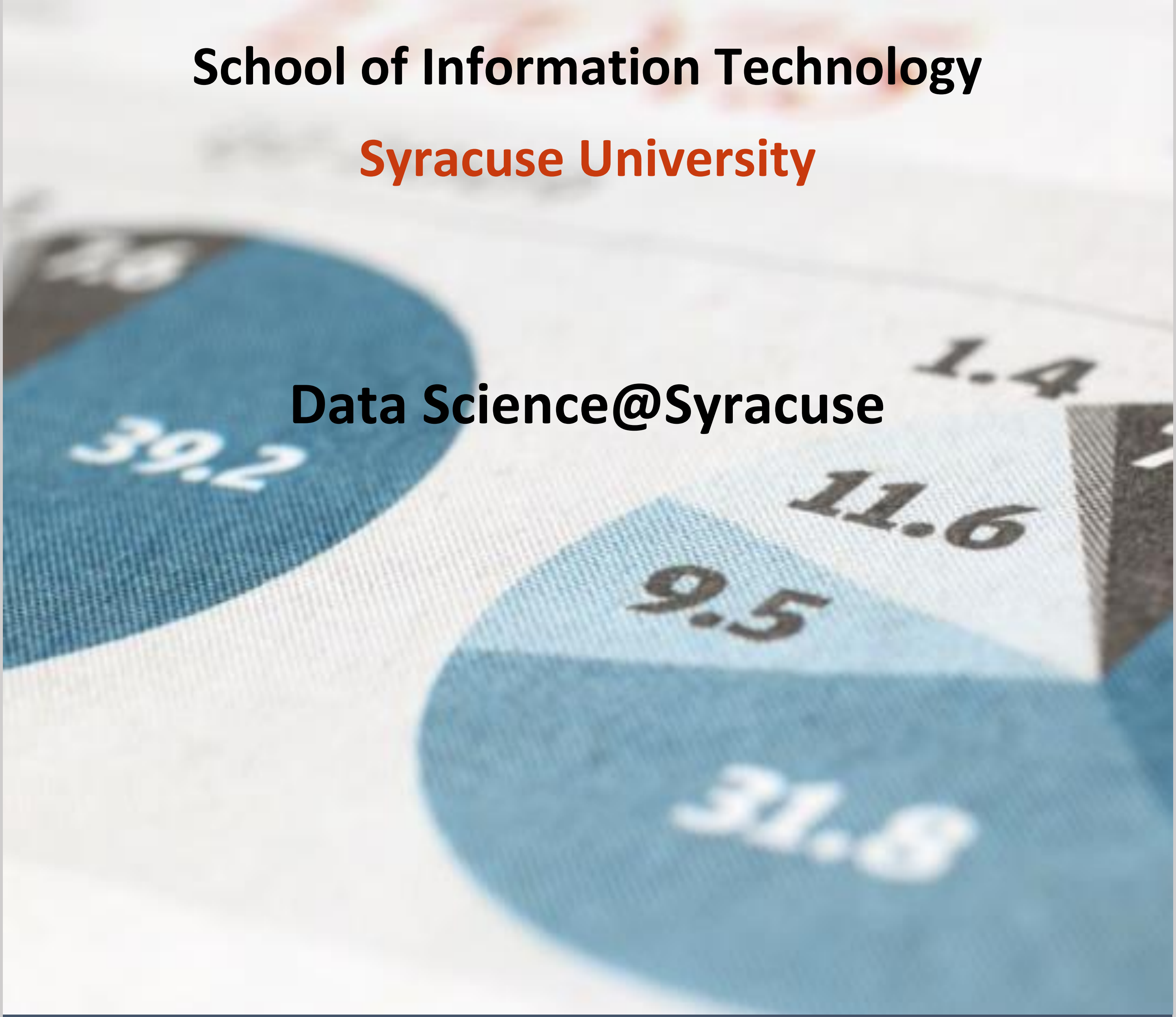
School of Information Technology

Syracuse University



Data Science@Syracuse

Project Portfolio Milestone Report

https://github.com/rdzidzornu/SU_ADS_PORTFOLIO_MILESTONE

Richard Dzidzornu 5/31/20 Milestone Portfolio

Table of Content

1.0 Introduction

The Applied Data Science Program at Syracuse University is designed to provide students with skillsets to collect, transform, manage, analyze, and develop in-depth understanding of data. It provides students with the practical analytical and technical skills to apply analytical concepts to gain insight from small and large datasets with no limitation to form or nature. Some of the underlying skills that were developed through assignments and projects in courses like Intro. to Applied Data Science(IST687), Data Visualization(IST719), Data Admin & Database Management(IST659), Data Analytics(IST707), Text Mining(IST736) include but not limited to scripting in SQL and using Python and R programing language for data analysis. The program as rigorous and challenging as it is will enable students to be more valuable and key contributors in their various organizations. While the overall program focuses on applications of data science to enterprise operations and processes, individual courses tackle data handling issues such as data capture, management, analysis, and communication for decision-making.

As noted by Prof. Jeffrey Saltz, "Applied Data Science is the holistic view of how organizations turn data into actionable insight in domains ranging from sales and marketing to customer insight and supply chain management. It includes the full life cycle of the data science process ranging from how to obtain, clean, and store data to leveraging machine learning and visualization to gain actionable insight of that data."

The program is focused on developing and equipping individuals the ability to demonstrate the following seven objectives:

1. Describe a broad overview of the major practice areas in data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analysis.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians and other relevant professionals in their organization.
7. Synthesize the ethical demonstration of data science practice(e.g., privacy)

In order to demonstrate the seven learning objectives of the Applied Data Science Program, reports and presentations were created in various courses which exemplify the knowledge and skills developed over the period in the program. These skills and knowledge will be demonstrated via this report by focusing on the following four courses.

2.0 IST719 Information Visualization

2.1 Project Details

The information Visualization course through the instructions and guidance for Prof. Gary Krudys provided various skills and knowledge when it comes to data visualization and storytelling. This was translated into the design and development of a poster on Housing Costs in California.

The project, required using different tools including Adobe Illustrator to create and design a poster that is very captivating and very informative to the reader or viewer. In line with that, the California housing dataset drawn from the 1990 U.S Census provides the basis of determining the housing prices in the various districts in California. The data collected information on the variables using all the block groups in California from the 1990 Census. A block group on average includes 1425.5 individuals living in a geographically compact area simply referred to as a district. This data has metrics such as the population, median income, median housing price, longitude, latitude, house median age, median income, total rooms and total bedrooms so on for each block group in California. A district typically has a population of 600 to 3,000 people.

Data cleaning, exploration, and analysis were done using R programming language. Data exploration of the California housing data covered areas of the housing costs in relation to ocean proximity, the population distribution of California from ocean proximity to inland localities (***Fig. 1*****).** The relationship between median house prices and median Income was determined and how impacts the house prices in various regions in California(***Fig. 2***).

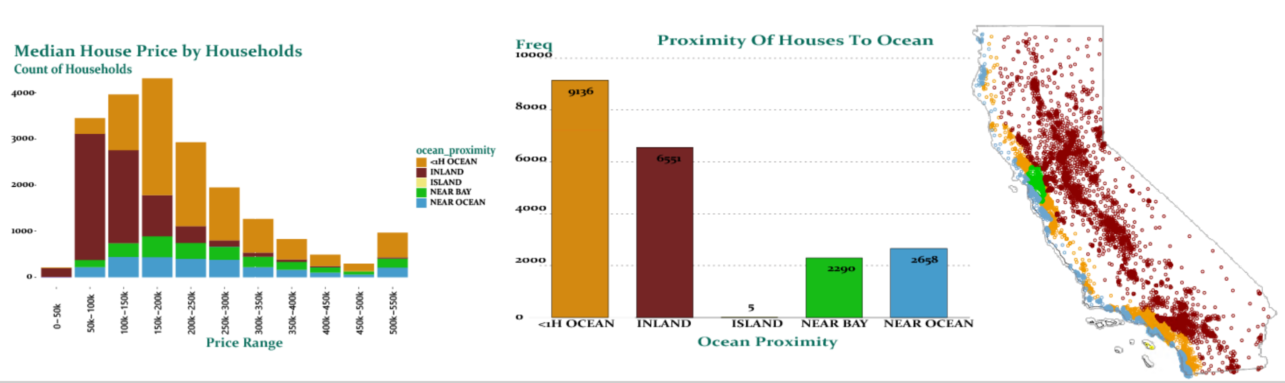


Figure 1: *Housing Costs in relation to Ocean Proximity in California*

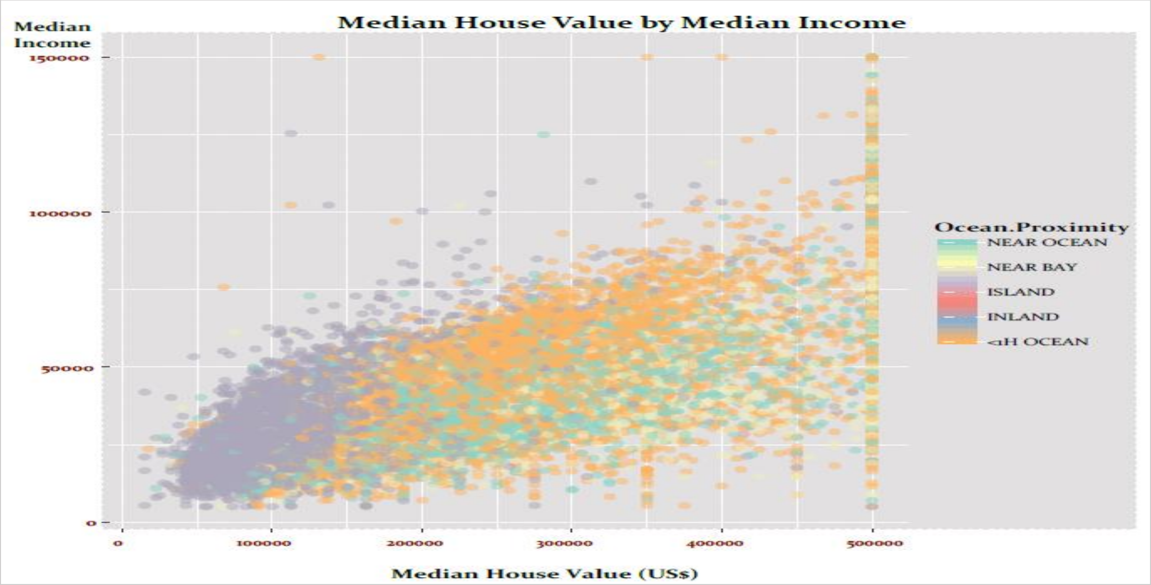


Figure 2: *Relationship between Median House Value and Median Income in (US$)*

The idea of the poster is to draw attention and interest of the reviewer into the intriguing facts and analysis projected by the poster. Humans by nature are drawn to colorful and well-designed visuals hence, the idea of a poster to transform raw data/numbers in a manner that draws audience to read and attain meaningful information from it. Techniques and technologies covered include but not limited to the use of R programming language, Adobe Illustrator, data mining, descriptive analytics, grouping and aggregation, plotting, data transformation, illustration, color theory.

2.2 Reflections and Learning Objectives

Every dataset has specific presentation or visualization needs and the reason that a set of data is used has just as much of an impact on the those needs as the data itself. The overall exercise of creating the poster using tools like R programming language and Adobe Illustrator showcased how significant it is to create a visually captivating image of the raw data to the intended audience. Its reveal how raw data is meaningless and of no significant if not projected in a manner that can be consumed and understood by the intended audience. It provided an insight into the importance of communicating results in visuals. Another important aspect noted in this exercise is the understanding and ability to identify the questions that need to be answered. Rather than thinking about how a set of data is collected, one came to the understanding of the need to think about how it would be used hence the need to work backwards to the data collected.

This project contributed to the successful application of the learning objectives which include the use of R to do basic data cleaning and preparation on any kind of datasets, identifying stories in datasets through data exploration using R to create rough plots to identify distributions and relationships, and the use of basic design principles to enhance viewer receptivity and convey meaningful.

3.0 IST659 Data Admin. Concepts $ Database Management

3.1 Project Details

The Database Administration and Database Management under the guidance and instructions of Dr. Gregory Block, provided skills and knowledge when it comes developing and managing different relational databases. The project, The Church Ministry Database, covered the design and creation of a database to track the activities and programs attendance of a local church from scratch using the Software Development Lifecycle (SDLC). The objective was to create a one stop data source for the church, which in the past, had no organized data on its ministries to monitor the progress and growth of the church. The Church Ministry Database will also as intended be used to create intuitive visualization for decision making.

Due to the time constraints, the project was limited to creating at most six entities in the database. The conceptual and physical models were developed to create and determine the relationship between the entities; Church, Person, Ministry, Program, Classroom, and Role. Microsoft Visio, a data modelling tool, was used to create the entity relationship in the conceptual model as well as the normalized logical model (Fig. 3) in building the database.

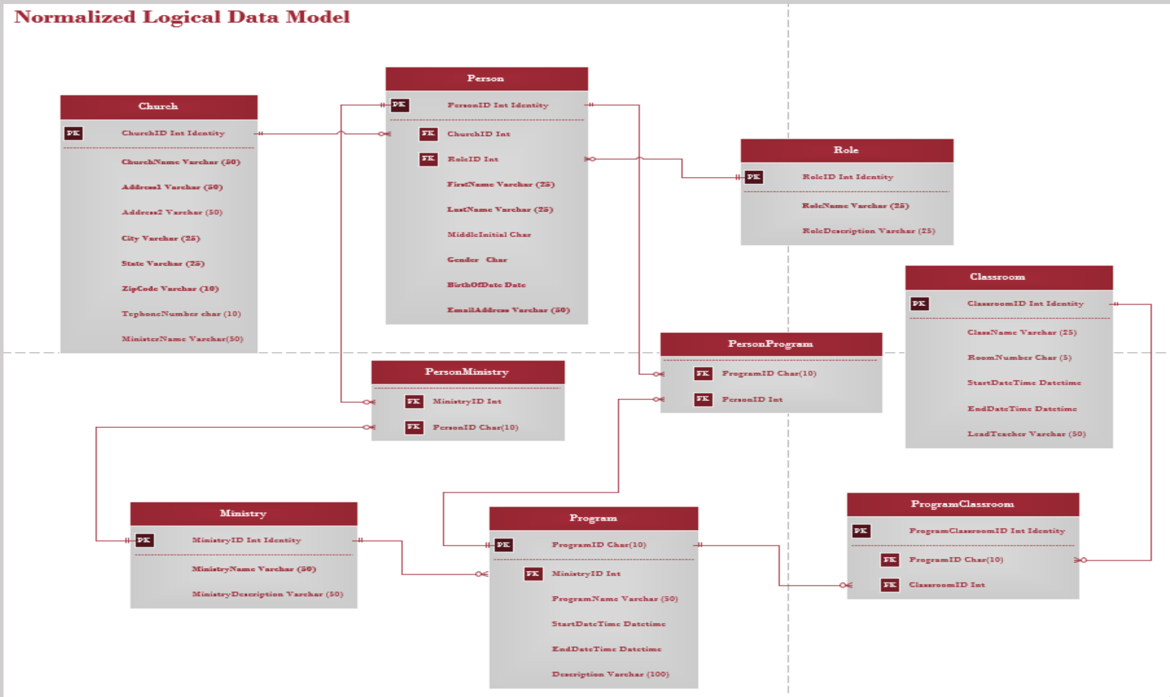
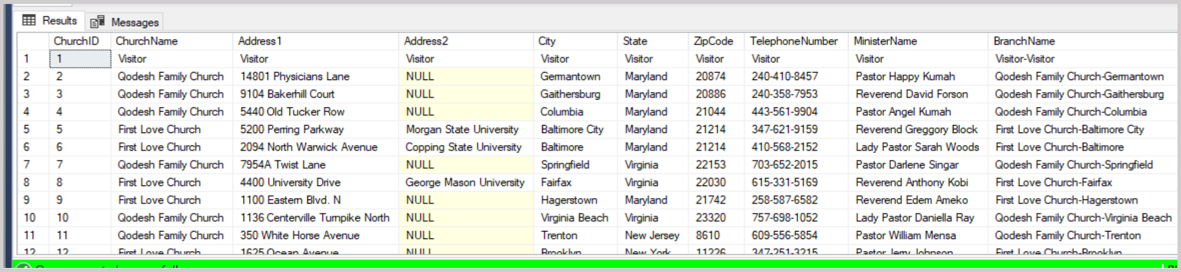


Figure 3: *Normalized logical Data Model depicting entity relationships*

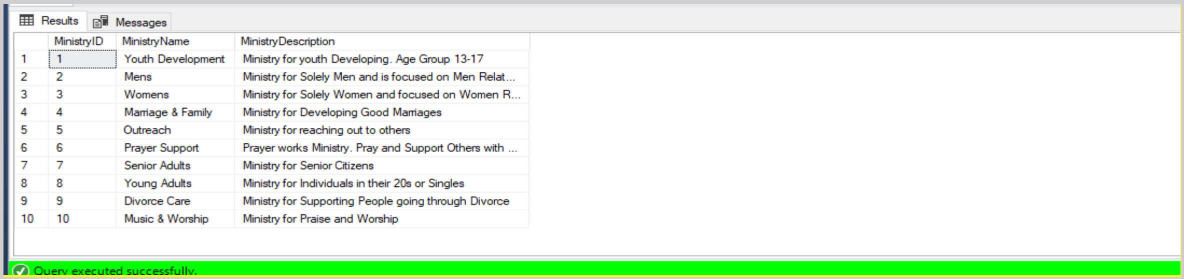
The database for the Church Ministry was created using Microsoft SQL Server application to develop all the database objects (tables, functions, stored procedures, and views) associated to the church ministry database project. The application aided in creating and inserting all the necessary records to develop the required database(Fig 4).

Project.Church Table



	ChurchID	ChurchName	Address1	Address2	City	State	ZipCode	TelephoneNumber	MinisterName	BranchName
1	1	Visitor	Visitor	Visitor	Visitor	Visitor	Visitor	Visitor	Visitor	Visitor-Visitor
2	2	Qodesh Family Church	14801 Physicians Lane	NULL	Germantown	Maryland	20874	240-410-8457	Pastor Happy Kumah	Qodesh Family Church-Germantown
3	3	Qodesh Family Church	9104 Bakerhill Court	NULL	Gaithersburg	Maryland	20886	240-358-7953	Reverend David Forson	Qodesh Family Church-Gaithersburg
4	4	Qodesh Family Church	5440 Old Tucker Row	NULL	Columbia	Maryland	21044	443-561-9904	Pastor Angel Kumah	Qodesh Family Church-Columbia
5	5	First Love Church	5200 Perring Parkway	Morgan State University	Baltimore City	Maryland	21214	347-621-9159	Reverend Greggory Block	First Love Church-Baltimore City
6	6	First Love Church	2094 North Warwick Avenue	Copping State University	Baltimore	Maryland	21214	410-568-2152	Lady Pastor Sarah Woods	First Love Church-Baltimore
7	7	Qodesh Family Church	7954A Twist Lane	NULL	Springfield	Virginia	22153	703-652-2015	Pastor Darlene Singar	Qodesh Family Church-Springfield
8	8	First Love Church	4400 University Drive	George Mason University	Fairfax	Virginia	22030	615-331-5169	Reverend Anthony Kobi	First Love Church-Fairfax
9	9	First Love Church	1100 Eastern Blvd. N	NULL	Hagerstown	Maryland	21742	258-587-6582	Reverend Edem Ameko	First Love Church-Hagerstown
10	10	Qodesh Family Church	1136 Centerville Turnpike North	NULL	Virginia Beach	Virginia	23320	757-698-1052	Lady Pastor Daniella Ray	Qodesh Family Church-Virginia Beach
11	11	Qodesh Family Church	350 White Horse Avenue	NULL	Trenton	New Jersey	8610	609-556-5854	Pastor William Mensa	Qodesh Family Church-Trenton
12	12	First Love Church	1625 Ocean Avenue	NULL	Brooklyn	New York	11226	347-251-3215	Pastor Jerry Johnson	First Love Church-Brooklyn

Project.Ministry Table

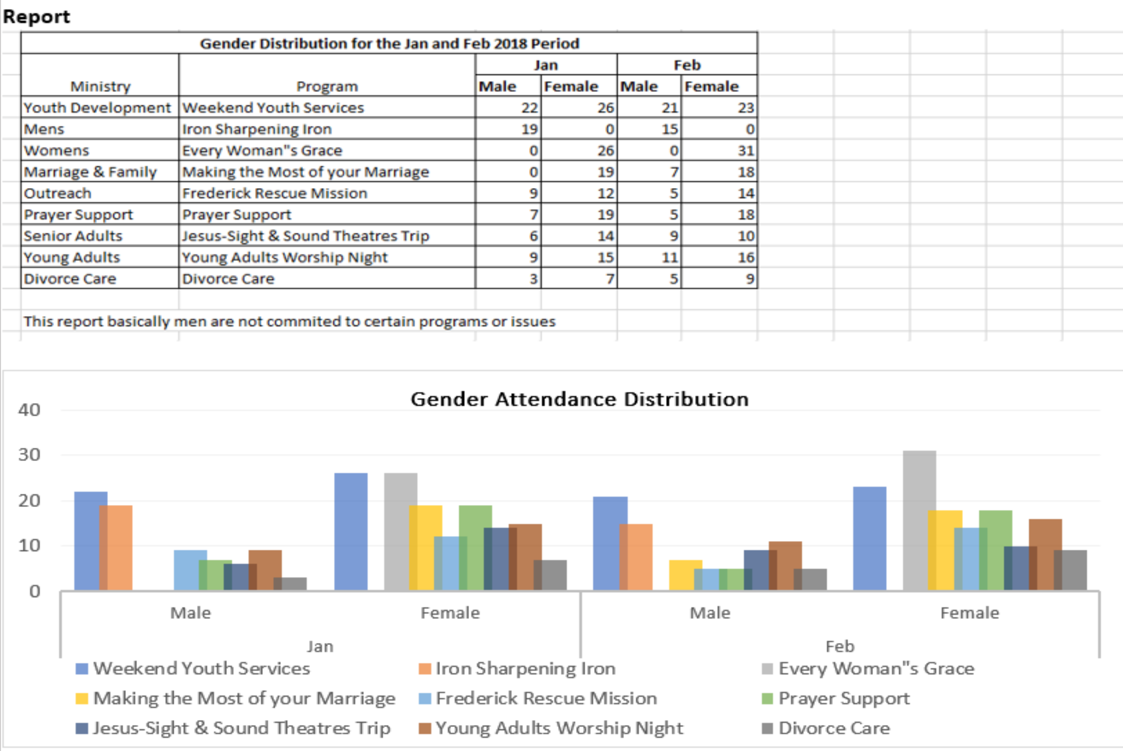


	MinistryID	MinistryName	MinistryDescription
1	1	Youth Development	Ministry for youth Developing. Age Group 13-17
2	2	Mens	Ministry for Solely Men and is focused on Men Relat...
3	3	Womens	Ministry for Solely Women and focused on Women R...
4	4	Marriage & Family	Ministry for Developing Good Marriages
5	5	Outreach	Ministry for reaching out to others
6	6	Prayer Support	Prayer works Ministry. Pray and Support Others with ...
7	7	Senior Adults	Ministry for Senior Citizens
8	8	Young Adults	Ministry for Individuals in their 20s or Singles
9	9	Divorce Care	Ministry for Supporting People going through Divorce
10	10	Music & Worship	Ministry for Praise and Worship

Query executed successfully.

Figure 4: *Sample tables with data extracted from MSSQL Server*

User interface was developed with Microsoft Access to give users and stakeholders the ability to query the data from a user perspective. Finally, interactive reports were created using Microsoft excel to answer most business questions like which ministry recorded the highest number of program attendances for a particular month, and what is the attendance for both males and females (Fig. 5).



Report

Gender Distribution for the Jan and Feb 2018 Period

		Jan		Feb	
Ministry	Program	Male	Female	Male	Female
Youth Development	Weekend Youth Services	22	26	21	23
Mens	Iron Sharpening Iron	19	0	15	0
Womens	Every Woman"s Grace	0	26	0	31
Marriage & Family	Making the Most of your Marriage	0	19	7	18
Outreach	Frederick Rescue Mission	9	12	5	14
Prayer Support	Prayer Support	7	19	5	18
Senior Adults	Jesus-Sight & Sound Theatres Trip	6	14	9	10
Young Adults	Young Adults Worship Night	9	15	11	16
Divorce Care	Divorce Care	3	7	5	9

This report basically men are not commited to certain programs or issues

Figure 5: *Interactive report created using excel with data from the Church Ministry Database Project*

3.2 Reflections and Learning Objectives

The design of a database as noted from this exercise is very demanding and time consuming. Understanding the business and processes of the user(client) is the foundation of every database development project. As noted from the project, the structure of the business process as well as the events that make up these processes determine how the data is stored. A well-structured and organized database provides a good source of information to decision makers. This project provided the skills and knowledge about the above-mentioned attributes of a database. The competence to draw an insight and develop a holistic approach to developing a relational database system has been an experience in my current position.

The completion of this project resulted in the successful application of the learning objectives such as ability to create databases and database objects using MSQL Server as a database management tool. It provided the chance to understand and solve problems by constructing database queries using Structured Query Language (SQL) which was leveraged in the Church Ministry Database. The ability to design databases using data modeling and data normalization techniques. Develop insights into future data management tools, techniques and trends to meet the needs of the user.

4.0 IST707 Data Analytics

4.1 Project Details

The Data Analytics under the guidance and instructions of Dr. Amy Gates, introduced skills and knowledge in data mining methods for extracting facts from data using open-source software packages in R programming language. Key concepts in the area of data mining, data preparation, association rule mining, classification, clustering, evaluation and analysis were implemented in the analysis of Patients Reviews on Specific Drugs project. The project aimed at analyzing the most common conditions for which people take drugs and the sentiments projected in these reviews

The Patients Reviews on Specific Drugs project encompassed the data retrieval, cleaning and pre-processing dataset from the Drugs.com. The dataset provided patient reviews on specific drugs along with related conditions and a 10-star patient rating reflecting overall patient satisfaction. The intention was to perform sentiment analysis of drug experiences over multiple facets. That is, sentiments on drugs aspects such as effectiveness and side effects. The project also used the dataset in developing different models and determining correlation between variables such as ratings, conditions, drugs and usefulness of drugs. Some aspects of the analysis were determining the most common conditions in the drug reviews dataset(*Fig. 6*), and common drugs in the sentiment reviews analysis(*Fig. 7*).

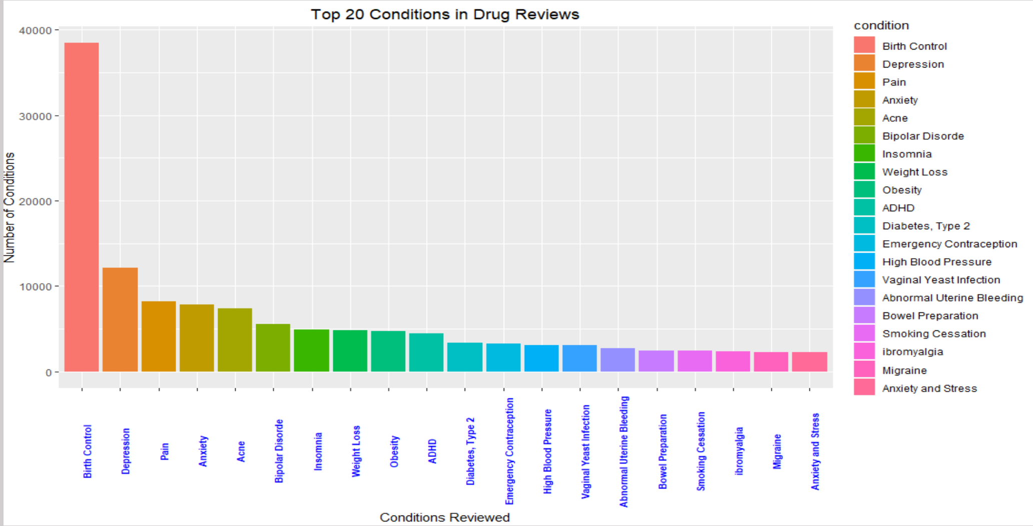


Figure 6*: Most Common Conditions in the Drug Reviews Dataset*

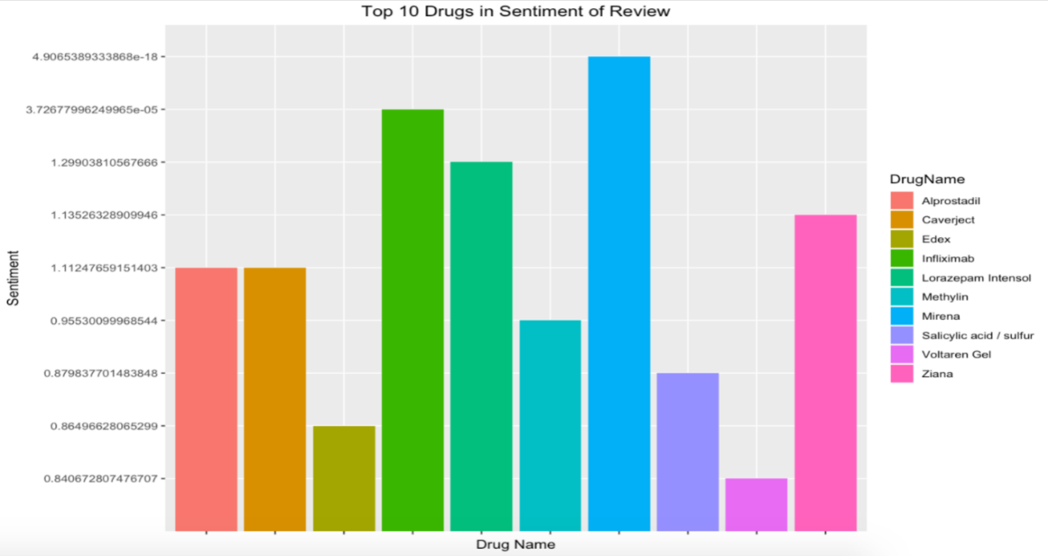


Figure 7: *Common Drugs in the Sentiment Analysis*

The project also looked at specific condition in relation to sentiment. A breakdown of how each sentiment was dispersed among different drugs that are treating depression(*Fig. 8*).

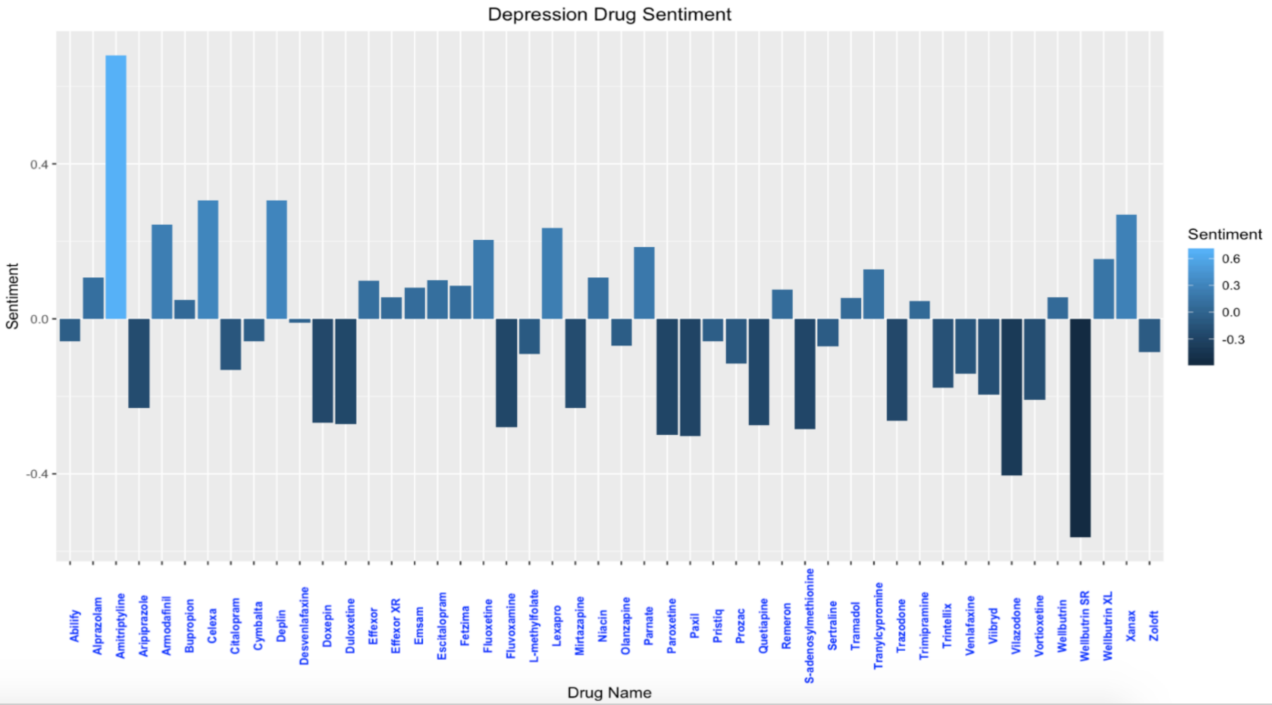


Figure 8: *Depression Drug Sentiment*

Interestingly as noted form the graph above, a couple are considered outliers with Wellbutrin SR which has an average sentiment of review of around –0.45.

4.2 Reflections and Learning Objectives

Understanding the data mining concepts, algorithms and evaluation methods to real-world problem is pivotal to finding useful patterns in every dataset and how to depict these patterns is very necessary in drawing meaningful conclusions as well. This project gave understanding and knowledge when it comes to creating models to determine patterns surrounding a given set of data.

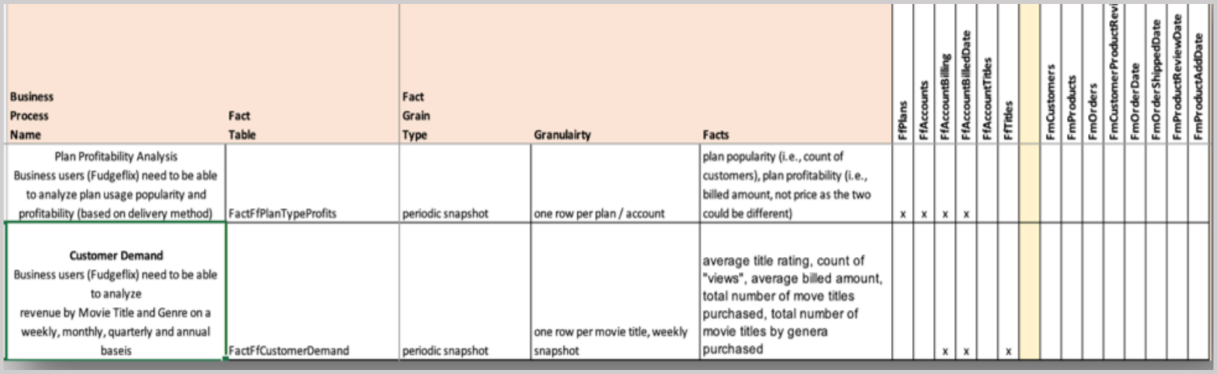
The project contributed to the successful application of the learning goals through the development of alternative strategies based on the data, and the communication of observations which translate to actionable insights. Data mining was also used in conjunction with visualization to identify patterns in the data for use in the sentiment analysis tasks.

5.0 IST722 Data Warehousing

5.1 Project Details

The Data Warehousing course, under the guidance and directions of Dr. Gregory Block also provided skills and knowledge in transforming business ideas and processes into creating and transforming data in transactional database system into a more intuitive way that can provide basis of performing analytical processes by all users. The course covered concepts, principles, and tools for designing, implementing a data warehouse. The project work under this course required the institution of a data warehouse for Fudgemart Inc. which holds two subsidiaries to store and deliver the organization's data assets which in otherwise was lacking of a centralized data system.

Using the Kimball methodology, the functional processes or business requirements of Fudgemart Inc, were gathered. Requirements were gathered or collected to determine the key factors impacting the Fudgemart operations by focusing on what business users do (or want to do in the future), rather than asking "what do you want in the data warehouse?". Based on these requirements gathered, a high-level dimensional (Fig. 9) and detailed-level dimensional model worksheets (Fig. 10) were created. These documents break down the business processes which would later be used to develop the enterprise data architecture.



Business Process Name	Fact Table	Fact Grain Type	Granulairty	Facts	FfPlans	FfAccounts	FfAccountBilling	FfAccountBilledDate	FfAccountTitles	FfTitles		FmCustomers	FmProducts	FmOrders	FmCustomerProductRevi	FmOrderDate	FmOrderShippedDate	FmProductReviewDate	FmProductAddDate
Plan Profitability Analysis Business users (Fudgeflix) need to be able to analyze plan usage popularity and profitability (based on delivery method)	FactFfPlanTypeProfits	periodic snapshot	one row per plan / account	plan popularity (i.e., count of customers), plan profitability (i.e., billed amount, not price as the two could be different)	x	x	x	x											
Customer Demand Business users (Fudgeflix) need to be able to analyze revenue by Movie Title and Genre on a weekly, monthly, quarterly and annual baseis	FactFfCustomerDemand	periodic snapshot	one row per movie title, weekly snapshot	average title rating, count of "views", average billed amount, total number of move titles purchased, total number of movie titles by genera purchased			x	x		x									

Figure 9: *High-Level Dimensional Model Worksheet*

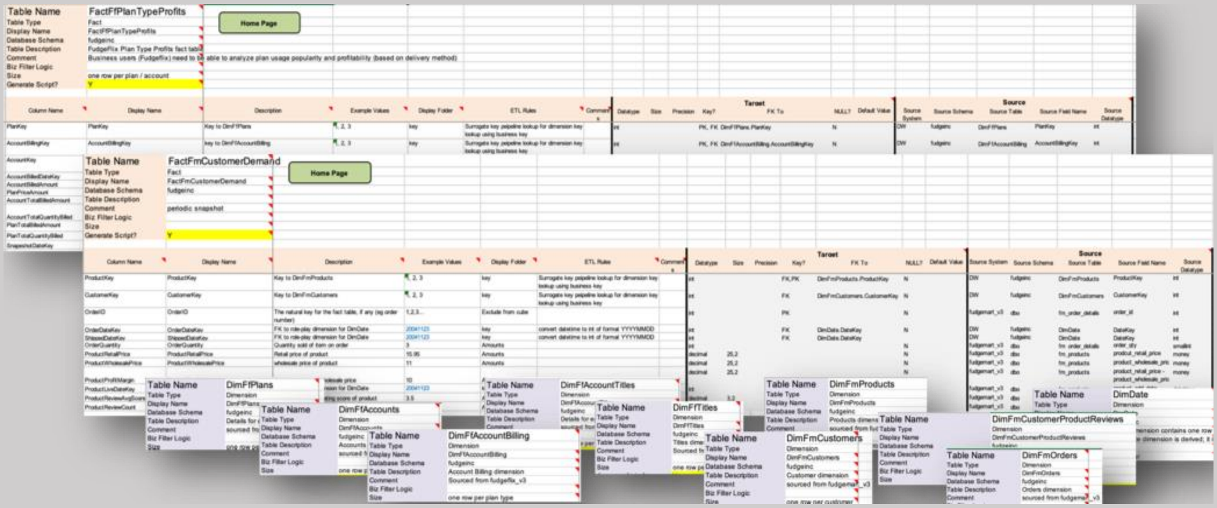
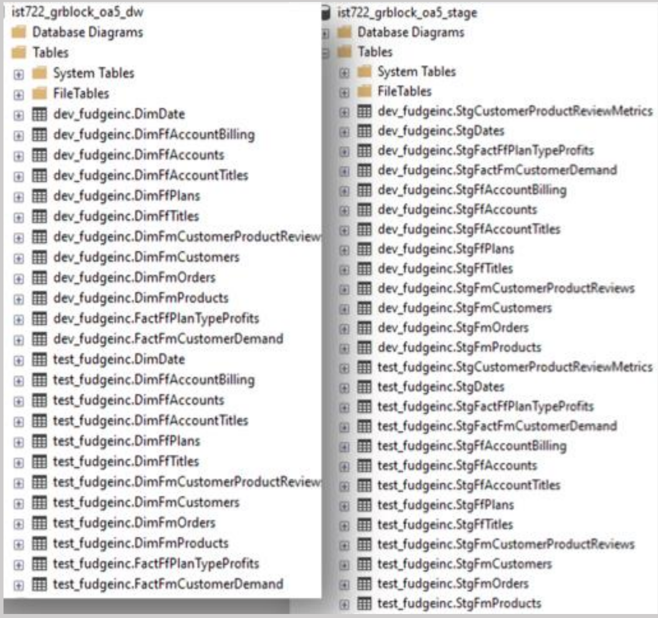


Figure 10: Detailed-Level Dimensional Model Worksheet

A portfolio of SQL Server tools that include SQL Server DBMS, SQL Server Integration Services (SSIS), SQL Server Reporting Services (SSRS) and SQL Server Analysis Service (SSAS) were leveraged to provide hands-on experience in implementing a reporting solution through projects.

The project implemented the dimensional model design (Kimball methodology) in a relational database as the approach for the enterprise bus architecture. Physical design (Fig. 11) included tables, keys, constraints, schemas, synonyms, and views. The final deliverable to the customer is relational online analytical process (ROLAP) star schema.



- ist722_grblock_oa5_dw
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - dev_fudgeinc.DimDate
    - dev_fudgeinc.DimFfAccountBilling
    - dev_fudgeinc.DimFfAccounts
    - dev_fudgeinc.DimFfAccountTitles
    - dev_fudgeinc.DimFfPlans
    - dev_fudgeinc.DimFfTitles
    - dev_fudgeinc.DimFmCustomerProductReview
    - dev_fudgeinc.DimFmCustomers
    - dev_fudgeinc.DimFmOrders
    - dev_fudgeinc.DimFmProducts
    - dev_fudgeinc.FactFfPlanTypeProfits
    - dev_fudgeinc.FactFmCustomerDemand
    - test_fudgeinc.DimDate
    - test_fudgeinc.DimFfAccountBilling
    - test_fudgeinc.DimFfAccounts
    - test_fudgeinc.DimFfAccountTitles
    - test_fudgeinc.DimFfPlans
    - test_fudgeinc.DimFfTitles
    - test_fudgeinc.DimFmCustomerProductReview
    - test_fudgeinc.DimFmCustomers
    - test_fudgeinc.DimFmOrders
    - test_fudgeinc.DimFmProducts
    - test_fudgeinc.FactFfPlanTypeProfits
    - test_fudgeinc.FactFmCustomerDemand

- ist722_grblock_oa5_stage
  - Database Diagrams
  - Tables
    - System Tables
    - FileTables
    - dev_fudgeinc.StgCustomerProductReviewMetrics
    - dev_fudgeinc.StgDates
    - dev_fudgeinc.StgFactFfPlanTypeProfits
    - dev_fudgeinc.StgFactFmCustomerDemand
    - dev_fudgeinc.StgFfAccountBilling
    - dev_fudgeinc.StgFfAccounts
    - dev_fudgeinc.StgFfAccountTitles
    - dev_fudgeinc.StgFfPlans
    - dev_fudgeinc.StgFfTitles
    - dev_fudgeinc.StgFmCustomerProductReviews
    - dev_fudgeinc.StgFmCustomers
    - dev_fudgeinc.StgFmOrders
    - dev_fudgeinc.StgFmProducts
    - test_fudgeinc.StgCustomerProductReviewMetrics
    - test_fudgeinc.StgDates
    - test_fudgeinc.StgFactFfPlanTypeProfits
    - test_fudgeinc.StgFactFmCustomerDemand
    - test_fudgeinc.StgFfAccountBilling
    - test_fudgeinc.StgFfAccounts
    - test_fudgeinc.StgFfAccountTitles
    - test_fudgeinc.StgFfPlans
    - test_fudgeinc.StgFfTitles
    - test_fudgeinc.StgFmCustomerProductReviews
    - test_fudgeinc.StgFmCustomers
    - test_fudgeinc.StgFmOrders
    - test_fudgeinc.StgFmProducts

Figure 11: Data Warehouse Fact and Dimensional Table

The project also implemented ETL packages in SQL Server DBMS, SQL Server Integration Services (SSIS) for source to target mappings (Fig. 12).

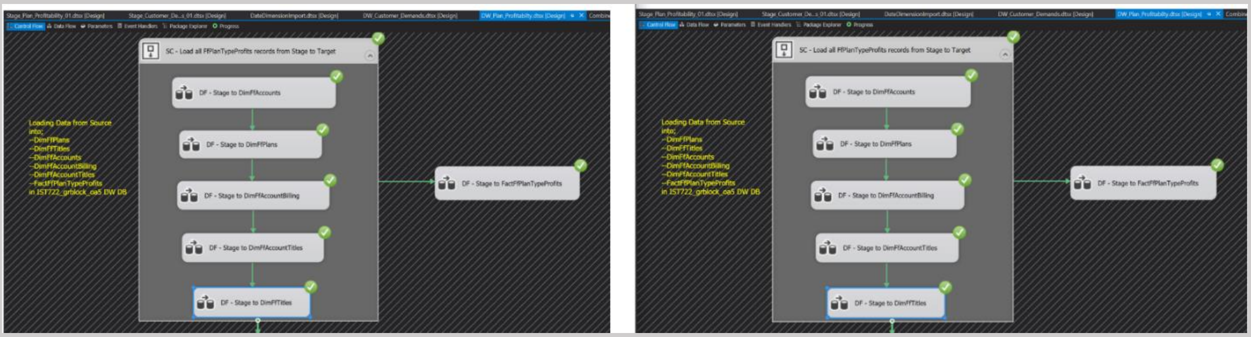

Stage_Plan_Profitability_01.dtsx [Design]
Stage_Customer_De...s_01.dtsx [Design]
DateDimensionImport.dtsx [Design]
DW_Customer_Demands.dtsx [Design]
DW_Plan_Profitabilty.dtsx [Design]
Combin
Control Flow
Data Flow
Parameters
Event Handlers
Package Explorer
Progress
SC - Load all FfPlanTypeProfits records from Stage to Target
DF - Stage to DimFfAccounts
DF - Stage to DimFfPlans
DF - Stage to DimFfAccountBilling
DF - Stage to DimFfAccountTitles
DF - Stage to DimFfTitles
DF - Stage to FactFfPlanTypeProfits
Loading Data from Source into;
--DimFfPlans
--DimFfTitles
--DimFfAccounts
--DimFfAccountBilling
--DimFfAccountTitles
--FactFfPlanTypeProfits
in IST722_grblock_oa5 DW DB

Figure 12: ETL Packages in SSIS

The data extraction, transformation and load process involved four major operations: extracting the data from the source, performing cleansing and conforming transformations, delivering the data to the presentation layer, and managing the backroom ETL processes and environment.

5.2 Reflections and Learning Objectives

It was well-known the design of a data warehouse and implementation of business intelligence tools that meets the needs of an organization from this exercise is very demanding, time consuming and cumbersome. Understanding the functional requirements as well as business process of all stakeholders is the foundation of every data warehouse and business intelligence implementation project. As noted from the project, the structure of the business process as well as the events that make up these processes determines how to holistically design and implement data warehouse project.

The completion of this project resulted in the successful application of the learning objectives such as ability to understand different approaches to creating and implementing data warehouse and business intelligence solutions. The application of Microsoft SQL Server Database Management System, SQL Server Integration Services (SSIS) in creating ETL solutions and the application of SQL Server Analysis Services (SSAS) for reporting solutions are just some of the high-level takeaways from the project.

6.0 Program Learning Objectives and Conclusion

Generally there are 5 types of analytics

- Retrospective: This is the traditional business intelligence/reporting: What happened? This concept of analytics relies heavily on Data Warehouse.
- Diagnostic: Analytic dashboards/Drill downs. Why did it happen?
- Descriptive: Real time dashboards. What's happening now.
- Predictive: Machine Learning and Forecasting. What is likely to happen.
- Prescriptive: Make a decision or take action. What should be done about it.

The last three is what data science represents. Data Science can be summed up as the combination of statistical analysis and programming skills to analyze high volume datasets and provide meaningful predictions and results. This requires the implementation of many skills like statistics, data mining, regression, classification, predictive modeling, and data visualization

An overview of the Applied Data Science program via the portfolio milestone indicated a successful completion and implementation of the learning objectives outlined above and major practice areas in data science. These encompass data collection which involves developing surveys and where necessary scraping the web, data cleaning and processing using programming languages to transform the data a form fit for exploration and analytical analysis and finally, data mining to derive patterns in a particular set of data. IST719 Data Visualization and IST707 Data Analytics used data from Kaggle and R programming language to manage the organize the data into the desired form.

Data visualizations concepts and tools were implemented in order to provide in-depth understanding of the data by applying statistical analysis and data mining skills and techniques to project patterns in worth knowing in the data. This was exhibited in all the projects. IST719 Data Visualization, in the Housing Costs in California project performed statistical analysis and data mining using R programming language, R packages and Adobe Illustrator for visualizations and recommended houses with low costs can be found in the Inlands area. IST722 Data Warehousing, the Fudgemart Inc. Data Warehouse used excel and Tableau desktop for both statistical analysis and visualization.

Various projects under-studied in the data science program and the outcomes can be complex and hard to explain. Presentation of approaches and findings to non-technical audience was a crucial part of all the underlying courses. Ability to interpret data, and tell the stories contain therein and in generate communicate and present these findings well were seen in all the above projects illustrated in this portfolio milestone. Communications of results were demonstrated in all PROJECTS. IST719 Data Visualization depicted this via the poster created and a recommendation that affordable houses could be are in the Inlands was made. Also attention was drawn to the fact that resources and social amenities should be routed to the Inlands since it contains the state's highest population.

Ethical issues in areas of unbiased analysis, data security, protection of personal individual information(PII) cannot be overlooked in data science. Throughout the program ethical issues were discussed and implemented in some of the projects. IST722 Data Warehousing, in developing the Fudgemart Inc. Data Warehouse dimension tables with PII were limited to individuals with privy to access such information. IST707 Data Analytics, Patients Review on Specific Drugs projects ensure privacy by removing details like name and age from data to ensure patients' privacy

7.0 References

- Yau, N. (2011). Visualize this: The Flowing Data guide to design, visualization, and statistics. Wiley Publishing. Retrieved from https://aimeeknight.files.wordpress.com/2016/07/visualize-this-nathan-yau.pdf

- Yau, N. (2013). Data points: Visualization that means something. Wiley Publishing. Retrieved from https://msucreativecomp.files.wordpress.com/2016/08/data_points.pdf

- Hoffer, J. A., Ramesh, V., & Topi, H. (2016). Modern database management (12th ed.). New York, NY: Pearson. ISBN13: 9780133544619

- Rainardi, V.(2008) Building a Data Warehouse with Examples in SQL Server. New York, NY: Apress. ISBN: 978-1590599310

- • W.H. Inmon, Imhoff, C., Sousa, R.(2001) Corporate Information Factory(2nd ed). Hoboken, New Jersey: Wiley Publishing. ISBN: 0-471-39961-2