Syracuse University

School of Information Studies

M.S. Applied Data Science

PORTFOLIO MILESTONE

Mark A. Roberts

SUID: 598273961

Github: https://github.com/dodekete/MSADS_Portfolio

MY INTRODUCTION TO DATA SCIENCE

A career in marketing/business development design and production professional in the commercial real estate industry served as my introduction to data science. I worked for a premier office and industrial real estate firms in the Greater Boston market. My firm specialized in biotech/life science but also high-end office and industrial leasing assignments for firms like TJX, State Street, Converse, Harvard Medical School/Boston Children's Hospital/Dana-Farber Cancer Institute, Corning, Boston Consulting Group, Shire Pharmaceuticals, TA Associates, Invesco, TIAA-Cref, Brookfield Properties and Normandy to name a few. Corporate real estate occupiers and owners had relied somewhat on power of relationships; working with a broker who was your college roommate whose insight was trusted implicitly. It was a 'Boys' Club'. After the Dot.com bust in 2001 and the banking crisis of 2008, that pattern rapidly became archaic. Suddenly, users needed "justification" for making significant moves, like moving a headquarters or leasing multiple stories of a downtown office. Enter data science.

Now, stakeholders were more diverse, not likely to be impressed by box seats at the big game, needed evidence to support significant transactions. Space utilization studies, employee commute analysis, labor supply dynamics, NER effects and lease up assumptions. Many of these items simply served an anecdotal and often purely visual representation that the move to a brand-new facility was not only justified, but necessary. There was a limit though: nobody wanted a regression equation, just a nicely color coded map to illustrate a(n already arrived at) decision point. I was keenly aware that we weren't "proving" anything but rather presenting a view of data that affirmed a conclusion we had already reached. Data was there to be there, not to be intrinsic to business practices. Still, it was better than nothing.

PRECONCEIVED NOTIONS

From watching too many procedural and fantasy/sci-fi shows, I was abused of the notion that data science was a magic trick. Pour in the right data, parameters and voila, an answer emerges as if from thin air. It can seem, from popular media, that data science can perform miracles. The truth is data science is more about guiding conversations toward truths rather than producing truths. Beyond a somewhat mythical view, I don't know if I knew that much about the actual practice of data science. Mean – Standard Deviation – Probability – Naïve-Bayes – F-Statistic: all of these terms were Greek to me. (And, I actually KNOW Ancient Greek). This innocence served me well as I entered the ADS program with an open mind.

APPLIED DATA SCIENCE AT SYRACUSE UNIVERSITY

The Applied Data Science program at Syracuse University's School of Information Studies equips students with the opportunity to develop analytic, technical and managerial skills in order to contribute measurable impacts in a highly competitive job market. The focus of the program is encapsulated in the word "Applied": students are taught to collect, manage, analyse and develop insights that can be communicated clearly to a broad audience of interested parties. Through courses such as Data Administration Concepts & Database Management (IST 659), Data Analytics (IST 707), Scripting for Data Analysis (IST 652) and Advanced Database Management (IST 769), I learned basic data structure (from byte to bit); how relational databases are built using SQL, Excel and Microsoft Access, meeting the business needs of organizations; how data mining techniques can be used to solve classification problems using R Studio; and, scripting in Python to explore a dataset to gain insights; using visualization to explore data and to communicate findings to a broad spectrum of audiences.

The Applied Data Science Program has seven learning objectives which were exemplified by the applications in this portfolio:

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 analyses.
6. Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.

PROJECT 1 – IST 659 DATA ADMINISTRATION CONCEPTS & DATABASE MANAGEMENT

Course Description

IST 659 is an introductory course to database management systems. This course examines data structures, file organizations, concepts, and principles of database management systems (DBMS) as well as data analysis, database design, data modeling, database management, and database implementation. More specifically, it introduces hierarchical, network, and relational data models; entity-relationship modeling; basics of Structured Query Language (SQL); data normalization; and database design. Using Microsoft's Access and SQL Server DBMSs as implementation vehicles, this course provides hands-on experience in database design and implementation through assignments, lab exercises, and course projects. This course also introduces advanced database concepts such as transaction management and concurrency control, distributed databases, multitier client/server architectures, web-based database applications, data warehousing, and NoSQL.

Learning Objectives:

After taking this course, the students will be able to:

- Describe fundamental data and database concepts
- Explain and use the database development lifecycle
- Create databases and database objects using popular database management system products
- Solve problems by constructing database queries using Structured Query Language (SQL)
- Design databases using data modeling and data normalization techniques
- Develop insights into future data management tool and technique trends
- Recommend and justify strategies for managing data security, privacy, audit/control, fraud detection, backup and recovery
- Critique the effectiveness of DBMS in computer information systems

Deliverable(s) – Custom Built SQL Data Base

Working under the tutelage of Dr. Gregory Block, I developed a database for a hypothetical real estate developer to manage: various communities; homes for sale within those communities; agents; buyers; all culminating in sale transactions. The goal of my project was to sort out some potentially complicated relationships (homes that are part of a community but can be sold individually, entities who could own or buy a house, and agents who could represent buyers or sellers).

With assumed client consultations, the project developed business rules which helped guide the first a conceptual model of the database, followed by a more refined logical/normalized model. Afterwards, I created tables and populated data using SQL Server Management Studio. Custom views and stored procedures were created as well as connection to Microsoft Access and Excel to produce reports to answer data questions, such as inventory of homes the developer/client has available, sales that have occurred and a detail of sale transactions.

Project Process & Development

Summary

A real estate developer has grown fast and needs help wrangling its far-flung projects into a cohesive data system, as opposed to a collection of spreadsheets, paper documents, accounting documents, tabbed folders and collections of emails, etc.

The developer builds "Communities". Communities are collections of "Homes" located in the suburbs. Communities may have the attributes "Normal" or "Adult Active". Each Home is located on a Street and is designated by a Number that is distinct for that Street; Street is unique but Number is not (there could be a 100 Maple Drive and a 100 Meadow Lane); the combination of the two is unique and also forms the Street Address of the Home.

Homes come in an array of types based on the Number of Bedrooms, ranging from 1 to 5 Bedrooms. Normal Communities have Homes in the full range from 1 to 5 – Families are welcome! Adult Active Communities are for quasi-retirees 55+, with a restriction of no young children. Therefore, Homes in Adult Active Communities may only have 1 or 2 Bedrooms (enough for 1 guest – No Ruffians!)

Homes feature amenities such as car parking, swimming pool, access to fitness center, nearest school. Some features depend on the Community (for instance, Adult Active Communities are not near any schools by design). Homes will feature different car parking configurations as a function of number of bedrooms. 5 bedroom homes may have more than one kitchen.

Each community features a dedicated Sales Team. A Team must have at least one member or "Sales Agent" and may have more than one; also, a Sales Agent may be on more than one Sales Team and function as a Sales Agent for more than one Community. A member of the Sales Team will act as "Owner" in a Sale Transaction (as a representative of the developer). Each Sale Transaction will have a Buyer. The Sale Transaction will have information associated with it such as Asking Price, Days on Market, Offer Price, and Final Sale Price.

Stakeholders/Entity/Attribute List & Glossary

Stakeholders	Entities	Attributes
DEVELOPER Creator of the Community/Homes, Original "Owner"	**COMMUNITIES** Collection of Homes developed in tandem with one another.	Type: • Adult Active • Normal
SALES TEAM Agents employed by the Developer to Sell the Homes and act as Seller Agent in the Sale Transaction	**HOMES** Individual units in Communities. Capable of being bought and sold.	Type • 1, 2, 3, 4 or 5 Bedroom Units • Street • Number • Community Type • Amenities • Sold Status
SALES Transactions Transfers of properties from Owner to Buyer via monetary purchase.	**Owner/Agent as Rep** **Buyer** **Home**	• Asking Price • Days on Market • Offer Price • Final Sale Price • Sale Date
OWNER Individual in possession of a Home.		Name Phone Address
BUYER Individual to whom the Home is sold.		Name Phone Address
AMENITIES Series of features attributed to houses based on Community and Home Type.	Amenity List	Car Parking/# Cars Nearest School Swimming Pool Additional Kitchen

Conceptual & Logical Data Models

ERD

E-ERD

Communities
PK CommunityID int identity
Type varchar(20)
Location varchar(20)
Number_Homes int
SalesTeam
PK SalesTeamID int identity
FK1 CommunityID int
FK2 SalesTeamMemberListID int
BedroomType
PK BedroomTypeID int identity
U1 NumberBedrooms int
Amenities
PK AmenityID int identity
NumberCars varchar (5)
NearestSchool varchar (20)
SwimmingPool boolean
AdditionalKitchen boolean
Homes
PK HomeID int identity
StreetNumber varchar(10)
StreetName varchar(20)
FK1 CommunityID int
FK2 BedroomTypeID int
FK3 AmenityTypeListID int
FK4 StatusListID int
SalesTeamMembersList
PK SalesTeamMemberListID int identity
FK1 SalesTeamID int
FK2 PartyID int
AmenityTypeList
PK AmenityTypeListID int identity
FK1 AmenityID
FK2 HomeID
Party
PK PartyID int identity
FirstName varchar(20)
LastName varchar(20)
U1 EmailAddress varchar(20)
PhoneNumber varchar(20)
StatusList
PK StatusListID int identity
Sold varchar(10)
Owners
PK OwnerID
FK1 PartyID int
SalesTransactions
PK SalesTransaction ID
AskingPrice float
OfferPrice float
FinalSalePrice float
DaysOnMarket int
SaleDate datetime
FK1 HomeID int
FK2 OwnerID int
FK3 BuyerID int
Buyers
PK BuyerID
FK1 PartyID int

Raw Data Sample

Raw Data Sample
Quaker Village
Seaview Cliff
Residential Place

SC
500 Mitchell Road — 3KR
└ Sold for $625K
└ Listed @ $450K
└ 2nd Level
No Pool

Sold by
W. Raleigh

Physical Database Design

```
-- Physical Database Design: Repeatable Script Drops/Creates Table According to Order of Dependency

-- Drop views
DROP VIEW IF EXISTS CommunityHomes;
GO
DROP VIEW IF EXISTS OwnersList;
GO
DROP VIEW IF EXISTS BuyersList;
GO
DROP VIEW IF EXISTS Sales;
GO
DROP VIEW IF EXISTS TotalAskingPriceByCommunity;
GO
DROP VIEW IF EXISTS HomesInventorybyBathrooms;
GO
DROP VIEW IF EXISTS TotalSales;
GO


-- Drop Table Statements

DROP TABLE IF EXISTS SaleTransaction;
GO
DROP TABLE IF EXISTS Buyers;
GO
DROP TABLE IF EXISTS Owners;
GO
DROP TABLE IF EXISTS AmenityTypeList;
GO
DROP TABLE IF EXISTS Homes;
GO
DROP TABLE IF EXISTS CommunitySalesTeam;
GO
DROP TABLE IF EXISTS Community;
GO
DROP TABLE IF EXISTS SalesTeam;
GO
DROP TABLE IF EXISTS BedroomType;
GO
DROP TABLE IF EXISTS Contact;
GO
DROP TABLE IF EXISTS AmenityProfiles;
GO

-- Create Table Statements, creating columns and constraints where indicated.
CREATE TABLE AmenityProfiles (
    AmenityProfilesID int identity,
    AmenityProfileName varchar(20) not null,
    NumberCars varchar(5) not null,
    SwimmingPool bit not null,
    AdditionalKitchen bit not null
    CONSTRAINT PK_AmenityProfiles PRIMARY KEY (AmenityProfilesID)
    );
    GO
```

```sql
CREATE TABLE Contact (
    ContactID int identity,
    FirstName varchar(20) not null,
    LastName varchar(20) not null,
    EmailAddress varchar(20) not null,
    PhoneNumber varchar(20)
    -- Contraints
    CONSTRAINT PK_Contact PRIMARY KEY (ContactID),
    CONSTRAINT U1_Contact UNIQUE (EmailAddress)
    );
    GO

CREATE TABLE BedroomType (
    BedroomTypeID int identity,
    NumberBedrooms varchar(10) not null
    CONSTRAINT PK_BedroomType PRIMARY KEY (BedroomTypeID),
    CONSTRAINT U1_BedroomType UNIQUE (NumberBedrooms)
    );
    GO

CREATE TABLE SalesTeam (
    SalesTeamID int identity,
    SalesTeamName varchar(50) not null,
    CommunityID int null,
    ContactID int null
    CONSTRAINT PK_SalesTeam PRIMARY KEY (SalesTeamID),
    CONSTRAINT FK1_SalesTeam FOREIGN KEY (ContactID) REFERENCES Contact(ContactID)
    );
    GO

CREATE TABLE Community (
    CommunityID int identity,
    CommunityName varchar(20) not null,
    CommunityType varchar(20) not null,
    CommunityLocation varchar(20) not null,
    NumberHomes int
    CONSTRAINT PK_Communities PRIMARY KEY (CommunityID)
    );
    GO

CREATE TABLE CommunitySalesTeam (
    CommunitySalesTeamID int identity,
    CommunityId int null,
    SalesTeamID int null,
    CONSTRAINT PK_CommunitySalesTeam PRIMARY KEY (CommunitySalesTeamID),
    CONSTRAINT FK1_CommunitySalesTeam FOREIGN KEY (CommunityID) REFERENCES
Community(CommunityID),
    CONSTRAINT FK2_CommunitySalesTeam FOREIGN KEY (SalesTeamID) REFERENCES
SalesTeam(SalesTeamID),
    );
    GO

CREATE TABLE Homes (
    HomesID int identity,
    StreetNumber varchar(10) not null,
    StreetName varchar(20) not null,
    CommunityID int null,
    BedroomTypeID int null,
```

```sql
        AmenityProfilesID int null,
        AskingPrice float not null
        CONSTRAINT PK_Homes PRIMARY KEY (HomesID),
        CONSTRAINT FK1_Homes FOREIGN KEY (CommunityID) REFERENCES Community(CommunityID),
        CONSTRAINT FK2_Homes FOREIGN KEY (BedroomTypeID) REFERENCES BedroomType(BedroomTypeID),
        CONSTRAINT FK3_Homes FOREIGN KEY (AmenityProfilesID) REFERENCES
AmenityProfiles(AmenityProfilesID)
        );
        GO

    CREATE TABLE AmenityTypeList (
        AmenityTypeListID int identity,
        AmenityProfilesID int not null,
        HomesID int not null
        CONSTRAINT PK_AmenityTypeList PRIMARY KEY (AmenityTypeListID),
        CONSTRAINT FK1_AmenityTypeList FOREIGN KEY (AmenityProfilesID) REFERENCES
AmenityProfiles(AmenityProfilesID),
        CONSTRAINT FK2_AmenityTypeList FOREIGN KEY (HomesID) REFERENCES Homes(HomesID)
        );
        GO

    CREATE TABLE Owners (
        OwnersID int identity,
        ContactID int not null
        CONSTRAINT PK_Owners PRIMARY KEY (OwnersID),
        CONSTRAINT FK1_Owners FOREIGN KEY (ContactID) REFERENCES Contact(ContactID)
        );
        GO

    CREATE TABLE Buyers (
        BuyersID int identity,
        ContactID int not null
        CONSTRAINT PK_Buyers PRIMARY KEY (BuyersID),
        CONSTRAINT FK1_Buyers FOREIGN KEY (ContactID) REFERENCES Contact(ContactID)
        );
        GO

    CREATE TABLE SaleTransaction (
        SaleTransactionID int identity,
        FinalPrice float,
        DaysOnMarket int,
        SaleDate datetime,
        HomesID int,
        OwnersID int,
        BuyersID int
        CONSTRAINT PK_SaleTransaction PRIMARY KEY (SaleTransactionID),
        CONSTRAINT FK1_SaleTransaction FOREIGN KEY (HomesID) REFERENCES Homes(HomesID),
        CONSTRAINT FK2_SaleTransaction FOREIGN KEY (OwnersID) REFERENCES Owners(OwnersID),
        CONSTRAINT FK3_SaleTransaction FOREIGN KEY (BuyersID) REFERENCES Buyers(BuyersID)
        );
        GO
```

Data Creation & Manipulation

```sql
    -- Insert statements
    -- Create Community records
```

```sql
INSERT INTO Community(CommunityName, CommunityLocation, NumberHomes, CommunityType)
VALUES ('Presidential Place', 'Quincy', 10, 'Gated')
, ('Seaview Cliff', 'Marblehead', 10, 'Adult Active')
, ('Quaker Village', 'Lincoln', 10, 'Retirement');
GO

SELECT * FROM Community

-- Create Bedroom Type records
INSERT INTO BedroomType(NumberBedrooms)
VALUES ('2 BR')
, ('3 BR')
, ('4 BR');
GO

SELECT * FROM BedroomType

-- Create Contact records
INSERT INTO Contact(FirstName, LastName, EmailAddress, PhoneNumber)
VALUES ('Abby', 'Adams', 'AAdams@domain.xyz', '(617) 555-1111')
, ('Dolly', 'Madison', 'DMadison@domain.xyz', '(617) 555-2222')
, ('Frank', 'Drake', 'FDrake@domain.xyz', '(617) 555-3333')
, ('Wally', 'Raleigh', 'WRaleigh@domain.xyz', '(617) 555-4444')
, ('Judy', 'Dench', 'JDench@domain.xyz', '(617) 555-5555')
, ('Bonny', 'Raitt', 'BRaitt@domain.xyz', '(617) 555-6666');
GO
SELECT * FROM Contact

-- Create AmenityProfiles records
INSERT INTO AmenityProfiles(AmenityProfileName,NumberCars,SwimmingPool,AdditionalKitchen)
VALUES ('Basic','2',0,0)
, ('Deluxe','3',0,1)
, ('Luxury','4',1,1);
GO

SELECT * FROM AmenityProfiles

-- Create Homes Records
INSERT INTO
Homes(StreetNumber,StreetName,Homes.BedroomTypeID,Homes.AmenityProfilesID,AskingPrice)
VALUES
('1','Washington Street',1,1,400000)
, ('3','Washington Street',1,1,400000)
, ('5','Washington Street',2,2,550000)
, ('2','Madison Boulevard',2,2,550000)
, ('4','Madison Boulevard',3,3,700000)
, ('6','Madison Boulevard',3,3,700000)
, ('100','Minot Road',1,1,500000)
, ('300','Minot Road',1,1,500000)
, ('500','Minot Road',2,2,650000)
, ('200','Quarterdeck Lane',2,2,650000)
, ('400','Quarterdeck Lane',3,3,725000)
, ('400','Quarterdeck Lane',3,3,725000)
, ('50','Friendship Drive',1,1,350000)
, ('60','Friendship Drive',1,1,350000)
, ('70','Friendship Drive',2,2,450000)
, ('10','Meetinghouse Road',2,2,450000)
, ('20','Meetinghouse Road',3,3,600000)
```

```sql
    , ('30','Meetinghouse Road',3,3,600000);
    GO

    SELECT * FROM Homes

    -- Create SalesTeam records
    INSERT INTO SalesTeam (SalesTeamName,CommunityID,ContactID)
    VALUES
        ('Presidential Place Team',(SELECT CommunityID FROM Community WHERE CommunityName =
'Presidential Place'),(SELECT ContactID FROM Contact WHERE LastName = 'Adams')),
        ('Presidential Place Team',(SELECT CommunityID FROM Community WHERE CommunityName =
'Presidential Place'),(SELECT ContactID FROM Contact WHERE LastName = 'Madison')),
        ('Seaview Cliff Team',(SELECT CommunityID FROM Community WHERE CommunityName = 'Seaview
Cliff'),(SELECT ContactID FROM Contact WHERE LastName = 'Drake')),
        ('Seaview Cliff Team',(SELECT CommunityID FROM Community WHERE CommunityName = 'Seaview
Cliff'),(SELECT ContactID FROM Contact WHERE LastName = 'Raleigh')),
        ('Quaker Village Team',(SELECT CommunityID FROM Community WHERE CommunityName =
'Presidential Place'),(SELECT ContactID FROM Contact WHERE LastName = 'Dench')),
        ('Quaker Village Team',(SELECT CommunityID FROM Community WHERE CommunityName = 'Seaview
Cliff'),(SELECT ContactID FROM Contact WHERE LastName = 'Raitt'));
        GO

    SELECT * FROM SalesTeam

    -- Create CommunitySalesTeam Records
    INSERT INTO
        CommunitySalesTeam(CommunityID,SalesTeamID)
        VALUES (1,1)
        , (2,2)
        , (3,3);
    GO

    SELECT * FROM CommunitySalesTeam

    -- Create Onwers Records
    INSERT INTO
        Owners(Owners.ContactID)
        VALUES
               ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'Adams')),
               ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'Madison')),
               ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'Drake')),
               ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'Dench'))
               GO

    SELECT * FROM Owners

    -- Update Homes CommunityID through FK reference to Community table
    UPDATE TGT
    SET CommunityID=SRC.CommunityID
    FROM Homes AS TGT
    INNER JOIN Community AS SRC ON SRC.CommunityName = 'Presidential Place'
    WHERE StreetName = 'Washington Street';
    GO

    UPDATE TGT
    SET CommunityID=SRC.CommunityID
    FROM Homes AS TGT
    INNER JOIN Community AS SRC ON SRC.CommunityName = 'Presidential Place'
```

```sql
WHERE StreetName = 'Madison Boulevard';
GO

UPDATE TGT
SET CommunityID=SRC.CommunityID
FROM Homes AS TGT
INNER JOIN Community AS SRC ON SRC.CommunityName = 'Seaview Cliff'
WHERE StreetName = 'Minot Road';
GO

UPDATE TGT
SET CommunityID=SRC.CommunityID
FROM Homes AS TGT
INNER JOIN Community AS SRC ON SRC.CommunityName = 'Seaview Cliff'
WHERE StreetName = 'Quarterdeck Lane';
GO

UPDATE TGT
SET CommunityID=SRC.CommunityID
FROM Homes AS TGT
INNER JOIN Community AS SRC ON SRC.CommunityName = 'Quaker Village'
WHERE StreetName = 'Friendship Drive';
GO

UPDATE TGT
SET CommunityID=SRC.CommunityID
FROM Homes AS TGT
INNER JOIN Community AS SRC ON SRC.CommunityName = 'Quaker Village'
WHERE StreetName = 'Meetinghouse Road';
GO

SELECT * FROM Community;
GO
SELECT * FROM Homes;
GO
SELECT * FROM BedroomType;
GO

UPDATE TGT
SET BedroomTypeID=BDR.BedroomTypeID
FROM Homes AS TGT
INNER JOIN BedroomType AS BDR ON BDR.NumberBedrooms = '2 BR'
WHERE StreetName = 'Washington Street' AND StreetNumber = '1';
GO

UPDATE TGT
SET BedroomTypeID=BDT.BedroomTypeID
FROM Homes AS TGT
INNER JOIN BedroomType AS BDT ON BDT.NumberBedrooms = 2
WHERE StreetName = 'Meetinghouse Road' AND StreetNumber = '3';
GO

UPDATE TGT
SET BedroomTypeID=BDT.BedroomTypeID
FROM Homes AS TGT
INNER JOIN BedroomType AS BDT ON BDT.NumberBedrooms = 3
WHERE StreetName = 'Meetinghouse Road' AND StreetNumber = '5';
GO
```

```sql
UPDATE TGT
SET BedroomTypeID=BDT.BedroomTypeID
FROM Homes AS TGT
INNER JOIN BedroomType AS BDT ON BDT.NumberBedrooms = 3
WHERE StreetName = 'Meetinghouse Road' AND StreetNumber = '5';
GO

SELECT * FROM Homes

-- Create Contact to populate Buyers table
INSERT INTO Contact(FirstName, LastName, EmailAddress, PhoneNumber)
VALUES ('Clark', 'Kent', 'CKent@domain.xyz', '(617) 555-7777')
, ('Diana', 'Prince', 'DPrince@domain.xyz', '(617) 555-8888')
, ('Bruce', 'Wayne', 'BWayne@domain.xyz', '(617) 555-9999')
, ('Wally', 'West', 'WWest@domain.xyz', '(617) 555-1010');
GO

-- Add Buyers Records
INSERT INTO
    Buyers(Buyers.ContactID)
    VALUES
           ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'Kent')),
           ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'Prince')),
           ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'Wayne')),
           ((SELECT Contact.ContactID FROM Contact WHERE LastName = 'West'))
           GO

-- Create SaleTransaction Records
INSERT INTO
SaleTransaction(FinalPrice,DaysOnMarket,SaleDate,SaleTransaction.HomesID,SaleTransaction.OwnersI
D,SaleTransaction.BuyersID)
VALUES
    (680000,64,GetDate(),5,1,1)
    , (380000,53,GetDate(),2,2,2)
    , (625000,53,GetDate(),9,3,3)
    , (705000,53,GetDate(),11,4,4)
GO
SELECT * FROM SaleTransaction

-- Create Views
-- Lists Homes showing each's community and bedroom type
CREATE VIEW dbo.CommunityHomes
AS
SELECT
    CommunityName,
    StreetNumber,
    StreetName,
    AskingPrice,
    BedroomType.NumberBedrooms
FROM Homes
JOIN Community ON Community.CommunityID = Homes.CommunityID
JOIN BedroomType ON BedroomType.BedroomTypeID = Homes.BedroomTypeID;
GO

-- Lists contacts designated as Owners
CREATE VIEW dbo.OwnersList
AS
```

```sql
SELECT
    FirstName AS OwnerFirstName,
    LastName AS OwnerLastName
FROM Owners
JOIN Contact ON Contact.ContactID = Owners.ContactID;
GO

-- Lists contacts designated as Buyers
CREATE VIEW dbo.BuyersList
AS
SELECT
    FirstName AS BuyerFirstName,
    LastName AS BuyerLastName
FROM Buyers
JOIN Contact ON Contact.ContactID = Buyers.ContactID;
GO

-- Lists sale transactions with Asking and Final Price
CREATE VIEW dbo.Sales
AS
SELECT
    StreetNumber,
    StreetName,
    AskingPrice,
    FinalPrice
FROM SaleTransaction
JOIN Homes ON Homes.HomesID = SaleTransaction.HomesID
GO

-- Lists total value of each community by summing asking price for each one's homes
CREATE VIEW dbo.TotalAskingPricebyCommunity
AS
SELECT
    CommunityName
    , SUM(AskingPrice) AS AskingPriceTotal
    , COUNT(HomesID) AS CountOfHomes
FROM
    Homes
RIGHT JOIN Community ON Community.CommunityID = Homes.CommunityID
GROUP BY CommunityName
GO

-- Lists number of bedroom type units per community
CREATE VIEW dbo.HomesInventorybyBathrooms
AS
SELECT
    CommunityName
    , NumberBedrooms
    , COUNT(Homes.BedroomTypeID) AS CountOfBedroomTypes
    , SUM(AskingPrice) AS AskingPriceTotal
FROM
    Homes
RIGHT JOIN BedroomType ON BedroomType.BedroomTypeID = Homes.BedroomTypeID
RIGHT JOIN Community ON Community.CommunityID = Homes.CommunityID
GROUP BY CommunityName
, NumberBedrooms;
GO
```

```sql
-- Lists sale transactions with Asking and Final Price
CREATE VIEW dbo.TotalSales
AS
SELECT
    StreetNumber,
    StreetName,
    CommunityName,
    NumberBedrooms,
    AskingPrice,
    FinalPrice
FROM
    SaleTransaction
INNER JOIN Homes ON Homes.HomesID = SaleTransaction.HomesID
INNER JOIN BedroomType ON BedroomType.BedroomTypeID = Homes.BedroomTypeID
INNER JOIN Community ON Community.CommunityID = Homes.CommunityID
GO
```

Answering Data Questions

1. Count of Bedroom Types by Community

Results | Messages

	CommunityName	NumberBedrooms	CountOfBedroomTypes	AskingPriceTotal
1	Presidential Place	2 BR	2	800000
2	Quaker Village	2 BR	2	700000
3	Seaview Cliff	2 BR	2	1000000
4	Presidential Place	3 BR	2	1100000
5	Quaker Village	3 BR	2	900000
6	Seaview Cliff	3 BR	2	1300000
7	Presidential Place	4 BR	2	1400000
8	Quaker Village	4 BR	2	1200000
9	Seaview Cliff	4 BR	2	1450000

2. Total Asking Price Value by Community

Results | Messages

	CommunityName	AskingPriceTotal	CountOfHomes
1	Presidential Place	3300000	6
2	Quaker Village	2800000	6
3	Seaview Cliff	3750000	6

3. Total Sales with Asking & Final Price

Results | Messages

	StreetNumber	StreetName	CommunityName	NumberBedrooms	AskingPrice	FinalPrice
1	4	Madison Boulevard	Presidential Place	4 BR	700000	680000
2	3	Washington Street	Presidential Place	2 BR	400000	380000
3	500	Minot Road	Seaview Cliff	3 BR	650000	625000
4	400	Quarterdeck Lane	Seaview Cliff	4 BR	725000	705000

4. Count of Team Members By Team

Results | Messages

	SalesTeamName	CountOfTeamMembers
1	Presidential Place Team	2
2	Quaker Village Team	2
3	Seaview Cliff Team	2

5. Count of Homes by Amenity Type

Results | Messages

	CommunityName	AmenityProfileName	CountOfAmenityTypes
1	Presidential Place	Basic	2
2	Quaker Village	Basic	2
3	Seaview Cliff	Basic	2
4	Presidential Place	Deluxe	2
5	Quaker Village	Deluxe	2
6	Seaview Cliff	Deluxe	2
7	Presidential Place	Luxury	2
8	Quaker Village	Luxury	2
9	Seaview Cliff	Luxury	2

Data Samples, Implementation, Reflection & Summary

Data Samples

```
SELECT * FROM Community;
```

Results | Messages

	CommunityID	CommunityName	CommunityType	CommunityLocation	NumberHomes
1	1	Presidential Place	Gated	Quincy	10
2	2	Seaview Cliff	Adult Active	Marblehead	10
3	3	Quaker Village	Retirement	Lincoln	10

```
SELECT * FROM AmenityProfiles;
```

Results | Messages

	AmenityProfilesID	AmenityProfileName	NumberCars	SwimmingPool	AdditionalKitchen
1	1	Basic	2	0	0
2	2	Deluxe	3	0	1
3	3	Luxury	4	1	1

```
SELECT * FROM Homes;
```

Results | Messages

	HomesID	StreetNumber	StreetName	CommunityID	BedroomTypeID	AmenityProfilesID	AskingPrice
1	1	1	Washington Street	1	1	1	400000
2	2	3	Washington Street	1	1	1	400000
3	3	5	Washington Street	1	2	2	550000
4	4	2	Madison Boulevard	1	2	2	550000
5	5	4	Madison Boulevard	1	3	3	700000
6	6	6	Madison Boulevard	1	3	3	700000
7	7	100	Minot Road	2	1	1	500000
8	8	300	Minot Road	2	1	1	500000
9	9	500	Minot Road	2	2	2	650000
10	10	200	Quarterdeck Lane	2	2	2	650000
11	11	400	Quarterdeck Lane	2	3	3	725000
12	12	400	Quarterdeck Lane	2	3	3	725000
13	13	50	Friendship Drive	3	1	1	350000
14	14	60	Friendship Drive	3	1	1	350000
15	15	70	Friendship Drive	3	2	2	450000
16	16	10	Meetinghouse Road	3	2	2	450000
17	17	20	Meetinghouse Road	3	3	3	600000
18	18	30	Meetinghouse Road	3	3	3	600000

Reports

Data Maintenance Form – Enter New Home

Enter New Home

Street Number: 1

Street Name: Washington Street

Asking Price: 400000

Community: Presidential Place

Bedrooms: 2 BR

Amenity Profile Type: Basic

Total Asking Price By Community

Community Name	Asking Price Total	Average Home Price
Presidential Place	$3,300,000	$550,000
Quaker Village	$2,800,000	$466,667
Seaview Cliff	$3,750,000	$625,000

Sales with Asking & Final Price with Sum

StreetNumber	StreetName	CommunityName	NumberBedrooms	AskingPrice	FinalPrice
4	Madison Boulevard	Presidential Place	4 BR	$700,000	$680,000
3	Washington Street	Presidential Place	2 BR	$400,000	$380,000
500	Minot Road	Seaview Cliff	3 BR	$650,000	$625,000
400	Quarterdeck Lane	Seaview Cliff	4 BR	$725,000	$705,000
				$2,475,000	$2,390,000

Inventory of Bathroom Units by Community

dbo_HomesInventorybyBathrooms

CommunityName	NumberBedrooms	CountOfBedroomTypes	AskingPriceTotal
Presidential Place	2 BR	2	800000
Presidential Place	3 BR	2	1100000
Presidential Place	4 BR	2	1400000
Quaker Village	2 BR	2	700000
Quaker Village	3 BR	2	900000
Quaker Village	4 BR	2	1200000
Seaview Cliff	2 BR	2	1000000
Seaview Cliff	3 BR	2	1300000
Seaview Cliff	4 BR	2	1450000

Inventory of Homes with Amenity Type by Community

Homes/Amenities

CommunityName	StreetNumber	StreetName	AmenityProfileName	NumberCars	SwimmingPool	AdditionalKitchen
Presidential Place	2	Madison Boulevard	Deluxe	3	False	True
Presidential Place	4	Madison Boulevard	Luxury	4	True	True
Presidential Place	6	Madison Boulevard	Luxury	4	True	True
Presidential Place	1	Washington Street	Basic	2	False	False
Presidential Place	3	Washington Street	Basic	2	False	False
Presidential Place	5	Washington Street	Deluxe	3	False	True
Quaker Village	50	Friendship Drive	Basic	2	False	False
Quaker Village	60	Friendship Drive	Basic	2	False	False
Quaker Village	70	Friendship Drive	Deluxe	3	False	True
Quaker Village	10	Meetinghouse Road	Deluxe	3	False	True
Quaker Village	20	Meetinghouse Road	Luxury	4	True	True
Quaker Village	30	Meetinghouse Road	Luxury	4	True	True
Seaview Cliff	100	Minot Road	Basic	2	False	False
Seaview Cliff	300	Minot Road	Basic	2	False	False
Seaview Cliff	500	Minot Road	Deluxe	3	False	True
Seaview Cliff	200	Quarterdeck Lane	Deluxe	3	False	True
Seaview Cliff	400	Quarterdeck Lane	Luxury	4	True	True
Seaview Cliff	400	Quarterdeck Lane	Luxury	4	True	True

List of Sales Team Members

Sales Team Members

SalesTeamName	LastName	FirstName	EmailAddress	PhoneNumber
Presidential Place Team	Adams	Abby	AAdams@domain.xyz	(617) 555-1111
Presidential Place Team	Madison	Dolly	DMadison@domain.x	(617) 555-2222
Quaker Village Team	Dench	Judy	JDench@domain.xyz	(617) 555-5555
Quaker Village Team	Raitt	Bonny	BRaitt@domain.xyz	(617) 555-6666
Seaview Cliff Team	Drake	Frank	FDrake@domain.xyz	(617) 555-3333
Seaview Cliff Team	Raleigh	Wally	WRaleigh@domain.xy	(617) 555-4444

Reflection

My biggest reflection is scale. I might have chosen a database type with fewer facets and intricacies. Given the limited time of the quarterly session, one ought to keep feasibility in mind. With that, I was only able to achieve what I consider to be rudimentary basics, with time being the essential missing ingredient. Scale is key.

Learning Outcomes & Reflection

●	●	○	●	●	●	○
Describe a broad overview of the major practice areas in data science.	Collect and organize data.	Identify patterns in data via visualization, statistical analysis, and data mining.	Develop alternative strategies based on the data.	Develop a plan of action to implement the business decisions derived from the analyses.	Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.	Synthesize the ethical dimensions of data science practice.

- **Describe a broad overview of the major practice areas in data science:**
  This project allowed me to consider how data organization and access impacts the needs of actual clients, and not a theoretical stance. A database must work for its intended user, or else, what's the point?

- **Collect and organize data:**
  This project gave me the opportunity to envision working with a client, developing business needs, and creating a plan to organize, create and populate data tables.

- **Develop alternative strategies based on the data.:**
  Initially, I imagined that I would keep the personal entities in separate data tables. It became more efficient to have a "People" table and to use connecting tables to identify roles that people can play in the data base (i.e., a person could own, buy, be an agent, or be a seller).

- **Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization:**
  From creation of stakeholders, data glossary and conceptual and, later, logical/normalized model, it is important for the data practitioner to be able to communicate clearly the direction undertaken. Later, data bases need to produce actionable insights, through aggregations, views and reports. A data base design must have utility as a core principle and communication as a guide to each facet of that utility.

The process of developing business rules, and each stage of the logical model to normalization, proved essential for creating a seamless data insertion process. I utilized Access and Excel ODBC connections to create my reports relying heavily on Views I created. A deep understanding of Excel/Access as well as the operations of a real estate firm helped me to complete these tasks.

Project 2– IST 652
Scripting for Data Analysis

Course Description

Scripting for the data science pipeline. Acquiring, accessing, and transforming data in the forms of structured, semistructured, and unstructured data. The goal of this class is to teach students the tools and skills of scripting needed to solve problems of accessing and preparing data in a variety of formats and situations, sometimes known as data wrangling. The scripting will provide the skills needed to form data science pipelines, from acquiring and cleaning data to accessing data and transforming data for analysis or visualization.

Learning Objectives:

Upon successful completion of this course, the student will be able to:

- Write scripts to access and amass data from fields in structured data, access fields in semistructured data, and define and find patterns of data in unstructured data; prepare and transform data to produce data summaries, lists, and networks;
- Analyze and solve data access problems for the three types of data and to find and deploy appropriate software packages that can be integrated into the problem solution; and
- Frame real-world data questions and show how they can be answered from data.

Deliverable(s) – Final Project Proposal, Project Code in Python,

Project Report & Presentation

Project Description

Working in groups, your final project you will demonstrate your ability to write Python scripts to access and amass data from fields in one or more of the three types of data studied in the course and to prepare and use data to produce data summaries, lists, and other structures. Pick a topic of investigation and the data that you will use, ideally from more than one source. The topic could focus on one main data set but also have supporting data. Your topic may focus on a single target topic or person, combinations of them, a comparison of more than one target topic and person, or comparisons over time. The data may come from any source: those that you have found online, collected from social media, or obtained through other means.

Pick several possible methods of analysis in order to give some initial idea of what analysis you will try. This analysis will be to answer the types of questions that you have worked on for the homework assignments. Since we are not focused on visualization, the results of your analysis can be reported as structured tables with a unit of analysis and collected, summarized, or computed values for those units.

Introduction

The Squirrel Census, an Atlanta-based organization with a mission to count and acknowledge the Eastern gray squirrel, set out to expand their mission to the Big Apple. In 2018, an 11-day squirrel census was conducted in Central Park in New York City. The organization engaged 323 volunteers who surveyed the 843-acre park for squirrels. The 323-volunteer squirrel sighters were assigned shifts in the morning and afternoon to count squirrels. Each was assigned 1 hectare (100-by-100-meter blocks) per shift to conduct their observations for 20-25 minutes. Each countable hectare in the park – 350 in total – was counted twice (in the morning and late afternoon). The census team assigned a random number to each hectare, which signified the order to survey each hectare. The numbers were randomly assigned to remove any selective bias.

The National Oceanic Atmospheric Administration's National Weather Service station near Belvedere Castle collected the weather data from hectare 17-D. The total time sighters spent counting squirrels was nearly 368.7 hours. This project presents the survey results.

About the Data

Two data sets are included for analysis.

Squirrels

The first is a .csv file containing the location, age, color, and other characteristics of 3,023 squirrels found in Central Park. The variables include:

Fields	Description	Type	Example
X	Longitude coordinate for squirrel sighting point	Number	-73.95613449
Y	Latitude coordinate for squirrel sighting point	Number	40.79408239
Unique Squirrel ID	Identification tag for each squirrel sightings. The tag is comprised of "Hectare ID" + "Shift" + "Date" + "Hectare Squirrel Number."	Plain Text	37F-PM-1014-03
Hectare	ID tag, which is derived from the hectare grid used to divide and count the park area. One axis that runs predominantly north-to-south is numerical (1-42), and the axis that runs predominantly east-to-west is Roman characters (A-I).	Plain Text	37F
Shift	Value is either "AM" or "PM," to communicate whether or not the sighting session occurred in the morning or late afternoon.	Plain Text	PM
Date	Concatenation of the sighting session day and month.	Plain Text	10142018
Hectare Squirrel Number	Number within the chronological sequence of squirrel sightings for a discrete sighting session.	Number	3
Age	Value is either "Adult" or "Juvenile."	Plain Text	Adult
Primary Fur Color	Value is either "Gray," "Cinnamon" or "Black."	Plain Text	Gray
Highlight Fur Color	Discrete value or string values comprised of "Gray," "Cinnamon" or "Black."	Plain Text	Black
Combination of Primary and Highlight Color	A combination of the previous two columns; this column gives the total permutations of primary and highlight colors observed.	Plain Text	+
Color notes	Sighters occasionally added commentary on the squirrel fur conditions. These notes are provided here.	Plain Text	Nothing selected as Primary. Gray selected as Highlights. Made executive adjustments.
Location	Value is either "Ground Plane" or "Above Ground." Sighters were instructed to indicate the location of where the squirrel was when first sighted.	Plain Text	Ground Plane
Above Ground Sighter Measurement	For squirrel sightings on the ground plane, fields were populated with a value of "FALSE."	Plain Text	10
Specific Location	Sighters occasionally added commentary on the squirrel location. These notes are provided here.	Plain Text	Branch

Fields	Description	Type	Example
Running	Squirrel was seen running.	Checkbox	TRUE
Chasing	Squirrel was seen chasing another squirrel.	Checkbox	TRUE
Climbing	Squirrel was seen climbing a tree or other environmental landmark.	Checkbox	FALSE
Eating	Squirrel was seen eating.	Checkbox	FALSE
Foraging	Squirrel was seen foraging for food.	Checkbox	FALSE
Other Activities		Plain Text	wrestling with mother
Kuks	Squirrel was heard kukking, a chirpy vocal communication used for a variety of reasons.	Checkbox	TRUE
Quaas	Squirrel was heard quaaing, an elongated vocal communication which can indicate the presence of a ground predator such as a dog.	Checkbox	FALSE
Moans	Squirrel was heard moaning, a high-pitched vocal communication which can indicate the presence of an air predator such as a hawk.	Checkbox	FALSE
Tail flags	Squirrel was seen flagging its tail. Flagging is a whipping motion used to exaggerate squirrel's size and confuse rivals or predators. Looks as if the squirrel is scribbling with tail into the air.	Checkbox	TRUE
Tail twitches	Squirrel was seen twitching its tail. Looks like a wave running through the tail, like a breakdancer doing the arm wave. Often used to communicate interest, curiosity.	Checkbox	FALSE
Approaches	Squirrel was seen approaching human, seeking food.	Checkbox	FALSE
Indifferent	Squirrel was indifferent to human presence.	Checkbox	TRUE
Runs from	Squirrel was seen running from humans, seeing them as a threat.	Checkbox	FALSE
Other Interactions	Sighter notes on other types of interactions between squirrels and humans.	Plain Text	ran from dog-walker
Lat/Long	Latitude and longitude	Point	POINT (-73.9561344937861 40.7940823884086)

Source: https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Squirrel-Data/vfnx-vebw

Hectares

The second data set includes location, weather, and environmental information on the 350 hectares of Central Park during the census:

Fields	Description	Type	Example
Hectare	ID tag, which is derived from the hectare grid used to divide and count the park area. One axis that runs predominantly north-to-south is numerical (1-42), and the axis that runs predominantly east-to-west	Plain Text	01A
Shift	Value is either "AM" or "PM," to communicate whether the sighting session occurred in the morning or late afternoon.	Plain Text	AM
Date	Concatenation of the sighting session month, day, and year (MMDDYYYY).	Plain Text	10072018
Anonymized Sighter	Sighter names were alphabetized then assigned a number (1-267) so that the individual abilities and characteristics of each sighter could be taken into account.	Number	110
Sighter Observed Weather Data	Each sighter was asked to describe the weather conditions before starting each sighting session. Weather values varied widely, and are in some cases inaccurate.	Plain Text	70¬∫ F, Foggy
Litter	Value can be "None", "Some" or "Abundant."	Plain Text	Some
Litter Notes	Sighters occasionally added commentary on the amount or quality of litter. These notes are provided here.	Plain Text	
Other Animal Sightings	Comma separated values indicating other animals sighted in the hectare. A list of other animals common to the park were provided; however, other animals were listed as well.	Plain Text	Humans, Pigeons
Hectare Conditions	Value is one of the following: “Calm”, “Moderate”, “Calm, busy” or “Busy.” The difference between “moderate” and “calm, busy” is as follows: “moderate” was either entered by the sighter, or translated from an entry that described a condition somewhere in-between busy and calm. “Calm, busy” was either entered by the sighter or translated from an entry that described discrete hectare sections as busy, but other hectare areas as separately calm.	Plain Text	Busy
Hectare Conditions Notes	Sighters occasionally added commentary on the hectare conditions. These notes are provided here.	Plain Text	Construction

Fields	Description	Type	Example
Number of sighters	The number of sighters that observed the hectare for the sighting session. Values range from "1" to "3."	Number	1
Number of Squirrels	The number of squirrels sighted in the hectare during the sighting session.	Number	4
Total Time of Sighting	The sighting session duration, in minutes.	Number	22

Source: https://data.cityofnewyork.us/Environment/2018-Central-Park-Squirrel-Census-Hectare-Data/ej9h-v6g2

Data Preprocessing

Squirrels

In reviewing the primary squirrels' data set, we observed several unnecessary columns and missing values. To start the cleaning process, the Unique Squirrel ID, Shift, Date, Hectare Squirrel Number, Combination of Primary and Highlight Color, Color notes, Location, Above Ground Sighter Measurements, Specific Location, Other Activities, Other Interactions, and Lat/Long columns were removed. Shift, Date, and Hectare Squirrel Number columns were duplicate columns from the hectare data frame. Combination of Primary and Highlight Color was a calculated field that combined Primary Fur Color and Highlight Fur Color with a "+". The original columns' format was ideal for our analysis, making the combination column no longer necessary. This is the same case for the Lat/Long column. The location column contained a significant number of missing values and was irrelevant for our analysis. The remaining notes columns included comments from surveyors. Most rows had missing values, and the values that were present were inconsistent in context.

Next, we addressed missing values in the remaining columns. For the age and primary fur color columns, if a value was missing, it was replaced with "Unknown." For the highlight fur color column, we exchanged blanks with "None." The reason for the difference is because some squirrels were only one color.

Hectares

The hectares data set also contained a few extraneous columns. Anonymized Sighter, which identifies who surveyed the hectare, was removed along with three columns pertaining to the litter and condition of the hectare. Since these columns were primarily blank, we decided to remove them. The date column was then correctly formatted.

Additionally, we observed two columns that needed to be normalized. The Sighter Observed Weather Data column contained the temperature and weather conditions (ex. 70°F, Foggy.) First, we separated the temperature from the condition, and then the condition was categorized into one of four main conditions: cloudy, sunny, rainy, and other. The other column requiring normalization was the Other Animal Sightings column. This column contained a list of animals in each row. Four more categories, Humans, Dogs, Birds, and Rodents, were created to classify each animal. Each category became their own column accepting a binary value.

Business Questions

Question 1

What are the primary visual characteristics of the squirrels? What color combinations were most popular? Were squirrels of similar markings found in a particular area of the park?

To evaluate this question, the Primary Fur Color and Highlight Fur Color columns in the squirrel dataset will be used to look for counts of each combination. Additionally, combining that information with the Hectare, X, and Y location columns, we can evaluate populated areas of the park by fur colors.

Question 2

Is there a section of the park that tends to have more squirrels? What are the characteristics of those hectares? Were other animals present?

Using the hectare dataset, the number of squirrels and hectare columns were used to determine the most populated part of the park. Then using the Other Animal Sighting column, those areas were evaluated for animal activity.

Question 3

Is there a relationship between the number of squirrels seen and the time of day? Does the weather at the time have any influence?

To summarize the squirrel sightings by time of day and weather, the hectare dataset was broken out by shift and weather condition. The temperature was then categorized in buckets of 10 degrees and summarized.

Question 4

What verbal noises were most popular? What behaviors are most common? How do these distribute around the park?

Using the squirrel dataset, we calculated the counts of types of sounds ("Kuks," "Quaas," and "Moans") as well as counts of squirrel behaviors (tail flagging/twitching, approaching/running from and indifference). The distribution of sounds is mapped on a Central Park map based on coordinate data.

Program Description

The program starts out by reading in the squirrel csv into a pandas dataframe and cleaning the dataframe via the outline in the cleaning section above. The same process then occurs for the hectare dataset. Some general descriptive statistics are performed to evaluate the distribution of the data and any correlations.

Question 1

The program creates a column to classify each squirrel as either in the lower, middle, or upper part of the park. A summary table is then created, grouping the squirrels by primary fur color. A plot of the distribution of the primary fur colors is then produced. The same summary table and bar chart is created for the highlight fur colors. Next, the program combines primary fur color and highlight fur colors into one summary table with 22 combinations displaying the count of squirrels for each combination and the percent of squirrels with each marking combination. A bar chart of this summary table is outputted. The squirrels are then mapped and colored by their primary fur color. A second map is created, removing the gray squirrels to focus on the clusters of the other primary fur colors. Next, a summary table of primary fur color by park area is created, along with an associated bar chart.

Question 2

Next, the program creates a summary table of the number of squirrels in each park area. A bar chart of this data is then outputted. The same park area column on the squirrel dataset is then added to the hectare dataset. Then a summary table was created, counting the number of human, dog, bird, and rodent sightings for each section of the park. The program then melts the table into a long table with a row for each area of the park, and then a bar chart was produced.

Question 3

The program then creates a dataframe summarizing the number of squirrels by time of day. A bar chart of this data is generated. Next, the program creates a dataframe summarizing the number of squirrels by weather condition and creates a bar chart. The program then combines the time of day and weather information by creating a dataframe to summarize the number of squirrels by weather condition and shift. Another bar chart is created. Next, five temperature buckets are created (40s, 50s, 60s, 70s, and 80s) and added to a temp_range dataframe. The program then adds this as a column to the hectare dataframe. Finally, a dataframe is created to summarize the number of squirrels by temperature range and generates a bar chart.

Question 4

Finally, the program creates a categorical column for the three "verbalizations" from the survey: "Kuks", "Quaas" and "Moans", as it were; subsequently, it provides a count and percentage of sounds by categorical name. Similarly, the program counts the number of each squirrel behavior observed by sighters, including "Tail flags", "Tail twitches," "Approaches," "Indifferent," and "Runs from." A further visualization shows the distribution of noises around the park, utilizing coordinate data points provided by the survey.

Analysis Results

Question 1

What are the primary visual characteristics of the squirrels? What color combinations were most popular? Were squirrels of similar markings found in a particular area of the park?

To evaluate the visual characteristics of the squirrels, we first summarized the squirrels by their primary fur color. There were three main colors, black, cinnamon, and gray, and some that had NAs. We categorized those NAs as unknown and created a separate category called unknown. Of those colors, 82% (2,473) of the central park squirrels were gray, 13% (392) were cinnamon, and only a small margin had primarily black fur.

Primary Fur Color	counts	percent
Black	103	0.03
Cinnamon	392	0.13
Gray	2473	0.82
Unknown	55	0.02

Next, we summarized the squirrels by their highlight fur color. There were four main colors, black, cinnamon, gray, and white. Since some had NAs, we categorized those NAs as none, creating a separate category. Of those colors, 36% (1,086) of the Central Park squirrels had no highlight color, 35% (1.076) had cinnamon markings, and 31% (954) had white markings. Black and/or gray highlights were only seen in 10% of squirrels.

	highlight_fur_color	counts	percent
0	Black	88.0	0.02
1	White	954.0	0.31
2	Gray	232.0	0.08
3	Cinnamon	1076.0	0.35
4	None	1086.0	0.36

To combine the primary and highlight fur markings, we created a summary table of all the combinations of colors. There were 22 color combinations: six black, seven cinnamon, eight gray, and one unknown. Most black squirrels had no highlight fur colors, most cinnamon squirrels had gray highlights, and most gray squirrels had no highlights. Gray squirrels were the only ones seen with up to three highlight colors.

	Primary Fur Color	Highlight Fur Color	counts	percent
0	Black	Cinnamon	15	0.005
1	Black	Cinnamon, White	3	0.001
2	Black	Gray	8	0.003
3	Black	Gray, White	1	0.000
4	Black	None	74	0.024
5	Black	White	2	0.001
6	Cinnamon	Black	10	0.003
7	Cinnamon	Black, White	3	0.001
8	Cinnamon	Gray	162	0.054
9	Cinnamon	Gray, Black	3	0.001
10	Cinnamon	Gray, White	58	0.019
11	Cinnamon	None	62	0.021
12	Cinnamon	White	94	0.031
13	Gray	Black	24	0.008
14	Gray	Black, Cinnamon	9	0.003
15	Gray	Black, Cinnamon, White	32	0.011
16	Gray	Black, White	7	0.002
17	Gray	Cinnamon	752	0.249
18	Gray	Cinnamon, White	265	0.088
19	Gray	None	895	0.296
20	Gray	White	489	0.162
21	Unknown	None	55	0.018

To evaluate if squirrels of similar markings were found in a particular part of the park, we first divided the hectares into lower, middle, and upper sections. This made it easier to summarize the data. For the most part, a majority of squirrels from each primary fur color were found in the lower part of the park.

	Park Area	Primary Fur Color	counts	percent
0	Lower	Black	42	0.014
1	Lower	Cinnamon	209	0.069
2	Lower	Gray	982	0.325
3	Lower	Unknown	20	0.007
4	Middle	Black	16	0.005
5	Middle	Cinnamon	106	0.035
6	Middle	Gray	681	0.225
7	Middle	Unknown	11	0.004
8	Upper	Black	45	0.015
9	Upper	Cinnamon	77	0.025
10	Upper	Gray	810	0.268
11	Upper	Unknown	24	0.008

We then mapped all the squirrels to visually see where the squirrels were found. The gray squirrels were fairly distributed across the entire park, but the black squirrels were clustered in the park's upper and lower areas. Additionally, we can see that the water reservoir in the center of Central Park contained no squirrels, which may have skewed the data in this section of the analysis.

Question 2

Is there a section of the park that tends to have more squirrels? What are the characteristics of those hectares? How busy were those areas? Were other animals present?

Here we summarized the number of squirrels in each area of the park. Each section had the same number of hectares. The middle area of the park was largely covered by water, potentially being the cause for the lower number of squirrel sightings. The upper and lower sections of the part had comparable amounts of water coverage but had a 10% difference in the number of squirrels.

	park area	counts	percent
0	Lower	1253	0.414
1	Middle	814	0.269
2	Upper	956	0.316

After looking at the area, we summarized the animal sightings by area of the park. Overall, the animal sightings were not linked to a particular region of the part except for rodents. Most of the hectares with rodents were seen in the upper portion of the park.

	park_area	hectare_counts	num_squirrels	animals	value
0	Lower	234	1252	humans	223
3	Lower	234	1252	birds	128
6	Lower	234	1252	dogs	179
9	Lower	234	1252	rodents	12
1	Middle	232	814	humans	206
4	Middle	232	814	birds	125
7	Middle	232	814	dogs	175
10	Middle	232	814	rodents	11
2	Upper	234	956	humans	207
5	Upper	234	956	birds	126
8	Upper	234	956	dogs	171
11	Upper	234	956	rodents	43

Question 3

Is there a relationship between the number of squirrels seen and the time of day? Does the weather at the time have any influence?

First, we summarized the number of squirrels observed by shift. The sightings occurred either in the morning (AM) or afternoon/evening (PM). As shown in the table and bar plot below, nearly 22% more squirrels were observed in the PM hours than the AM hours.

	shift	num_squirrels
0	AM	1347
1	PM	1675

Next, we summarized the squirrel sightings by weather condition. The four weather condition categories used were: cloudy, other, rainy, or sunny. The majority of squirrels were observed in cloudy conditions or other weather conditions. Squirrels were moderately observed in sunny weather and very rarely observed in rainy weather.

	weather_condition	num_squirrels
0	cloudy	1143
1	other	904
2	rainy	304
3	sunny	671

We also broke down the weather conditions by shift to see if certain weather conditions had more sightings in the AM or PM. As shown in the table and bar plot below, most squirrels were observed in the PM across all weather conditions except in the rain.

	weather_condition	shift	num_squirrels
0	cloudy	AM	420
1	cloudy	PM	723
2	other	AM	418
3	other	PM	486
4	rainy	AM	222
5	rainy	PM	82
6	sunny	AM	287
7	sunny	PM	384

After diving into weather conditions, we explored the number of squirrels observed across different temperature ranges. As shown in the table and bar plot below, most squirrels were observed in temperatures ranging in the 50s and 60s. There were fewer squirrel sightings on colder days in the 40s and hotter days in the 70s & 80s.

	temp_range	num_squirrels
0	40s	217
1	50s	953
2	60s	1132
3	70s	570
4	80s	150

Question 4

What verbal noises were most popular? What behaviors are most common? How do these distribute around the park?

With 87 observations, "Kuks" represent 63% of the total vocalizations, and nearly one-third of vocalizations are "Quaas". "Moan" barely registers on the tally with only three observations, at just 2% of occurrences. Whatever meaning is expressed by "Moans" is undoubtedly a squirrel event that may have eluded the sighters.

	Sound	counts	percent
0	Kuk	87	0.63
1	Moan	3	0.02
2	Quaa	49	0.35

Most verbal squirrels were observed kuking

Behaviors show a broader distribution, with "Indifferent" squirrels on top at 43, while "Runs from" is a close second at 39. "Tail flags" and "Tail twitches are at 18 and 17, respectively, with "Approaches" at the bottom with four counts.

	Behaviors	Counts
0	Tail flags	18
1	Tail twitches	17
2	Approaches	4
3	Indifferent	43
4	Runs from	39

Most squirrels were indifferent or ran from humans

There does not appear to be a pattern in the distribution of verbalization types throughout the park. Whatever the cause of one sound or another does not appear to be locally specific.

	Sound	counts	percent
0	Kuk	87	0.63
1	Moan	3	0.02
2	Quaa	49	0.35

Verbal squirrels were evenly distributed around the park

Conclusions

Squirrels

Overall, there were 3,023 squirrels sighted in New York City's Central Park during the 2018 squirrel census. They were of three primary fur colors, and four highlighted fur colors. There were 22 combinations of colors. Most of the squirrels sighted were gray and either had no highlights or cinnamon highlights. Central Park squirrels made two possible noises and had varying behavior towards humans. They are most likely to "Kuk" while being indifferent or perhaps running away from humans. Also, squirrel vocalizations occurred throughout the park in no discernable pattern. We didn't observe any upper or lower patois groupings.

Hectares

The squirrels were observed at two different times of the day over 11 days. Most squirrels were recorded in the PM, during cloudy weather conditions, and temperatures in the 50s-60s. There were less squirrel sightings in the AM, rainy conditions, and in extreme cold/warm. Additionally, most squirrels were sighted in the lower area of the park, which may have been due to less observable areas in the middle and upper regions because of a large water reservoir bordering those areas. Humans were the most observed "animal" in the park and were evenly sighted throughout the park.

Groupwork

The team worked together on the overall structure of the program and report. Each team member took the responsibilities laid out below:

- Squirrels Data Cleaning – Mark Roberts
- Hectare Data Cleaning – Sandy Spicer
- Business Question 1 – Katie Haugh
- Business Question 2 – Katie Haugh
- Business Question 3 – Sandy Spicer
- Business Question 4 – Mark Roberts

Learning Outcomes & Reflection

○	●	●	○	●	●	○
Describe a broad overview of the major practice areas in data science.	Collect and organize data.	Identify patterns in data via visualization, statistical analysis, and data mining.	Develop alternative strategies based on the data.	Develop a plan of action to implement the business decisions derived from the analyses.	Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.	Synthesize the ethical dimensions of data science practice.

Even a low stakes project like Squirrels could be a candidate for robust data scientific exploration and analysis. This project was an opportunity to learn data processing techniques in Pandas, make decisions on how to handle sticky data problems, and organize our data to be able to create visualizations and a compelling story. My team desired materials that could have just as slick an approach as if we were pitching a Class A office tower leasing assignment in New York City…just, it's about squirrels! Visual storytelling is a crucial ability that a Data Scientist can offer.

- **Collect and organize data:**
  As part of the project description, two datasets needed to be used. The squirrel set gave us access to data on their behavior and location, providing a challenge of organization.

- **Identify patterns in data via visualization, statistical analysis, and data mining:**
  From early experimentation, we discerned a few patterns regarding 1. Places squirrels were sighted, 2. Weather conditions, 3. Physical description of squirrels and 4. Behavior/Vocalizations. We centered our data analysis around these four insights gained from initial data mining.

- **Develop a plan of action to implement the business decisions derived from the analyses:**
  We sought to utilize scripting learned in the course materials to utilize our data to answer our business questions, providing a detailed analysis of our findings.

- **Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization:**
  Our project was centered around Business Questions, i.e. what would our 'Client" like to know about squirrels. Our analysis focused on being able to address each of our four chosen business questions and communicate all the necessary steps to answer them.

Project 3– IST 707
Data Analytics

Course Description

Introduction to data mining techniques, familiarity with particular real-world applications, challenges involved in these applications, and future directions of the field. This course will introduce popular data mining methods for extracting knowledge from data. The principles and theories of data mining methods will be discussed and will be related to the issues in applying data mining to problems. Students will also acquire hands-on experience using state-of-the-art software to develop data mining solutions to scientific and business problems. The focus of this course is on understanding of data and how to formulate data mining tasks to solve problems using the data.

The topics of the course will include the key tasks of data mining, including data preparation, concept description, association rule mining, classification, clustering, evaluation and analysis. Through the exploration of the concepts and techniques of data mining and practical exercises, students will develop skills that can be applied to business, science or other organizational problems.

Learning Objectives:

After taking this course, the students will be able to:

- Document, analyze, and translate data mining needs into technical designs and solutions.
- Apply data mining concepts, algorithms, and evaluation methods to real-world problems.
- Employ data storytelling and dive into the data, find useful patterns, and articulate what patterns have been found, how they are found, and why they are valuable and trustworthy.

Deliverable(s) – Final Project Classification Project Code & Report

The objective of the project is to use the main skills taught in this class to solve a real data mining problem. Students can choose to work individually or pair up with another student.

- Checkpoint 1: project idea proposal and presentation: Your idea proposal should include an overview of the data mining problem, the data set you will use and its availability, and your proposed data mining approach.
- Checkpoint 2: project progress presentation: Show preliminary results and major challenges.
- Checkpoint 3: Final project report: The final project report should describe the data mining problem, its significance and broader impact, the data mining approaches, results, and interpretation of the discovered patterns.

2019 Boston Marathon Dataset

Under the direction of Dr. Block, I focused my project on the 2019 Boston Marathon Results dataset. I pulled the data from a CSV file posted on Boston Athletic Association (which manages the event) website www.baa.org ; I wanted to pull raw data from a primary source as opposed to from GitHub, to have the full experience of a data set which might require extensive cleaning. After some experimentation with clustering that did not yield results (discussed with the professor at Checkpoint 1), I decided to make use of the commonplace classifiers inherent in running data (Gender, Age, Country). In researching methods of analyzing the data, I found some data scientists had created "Splits" data by arithmetic means. Dividing the finish time by 26, produces an average AKA Pace. Pace could then be used to calculate 5K, 10K and Half-Marathon splits. One problem: that data is not actually provided in the data set. Dr. Block and I decided that the results gained by such a method are problematic. The stakes of the results of a race are low, but transferred to another dataset, it's the pathway that a data scientist can allow bias to take root in analysis. We decided that the data scientists role is only to seek insights available from the data itself. "Noodling" is not allowed.

That said, I sought an separate BAA source (Qualification Times by Age/Gender Group) and applied a process to create a new variable "Achieved Qualifying Time" with respect to Gender and Age group. In other words, to qualify for the race a runner would have needed to reach a particular time, indexed by their particular grouping. I created my classification models to predict whether a runner could actually achieve their Qualifying Time. Using factors such as Age Quartile, Gender, Region, I used Decision Tree ("DT"), Support Vector Machine ("SVM") and Random Forest ("RF") and achieved generally 80% Accuracy rates with moderately good Kappa. An initial test with Naïve-Bayes yielded inferior results and was not included. Retuning my parameters yielded slightly better results.

Nearly 30,000 people run the Boston Marathon, normally held yearly on Patriots Day. Patriots Day is a local holiday celebrated in Massachusetts and Maine (which was formerly a part of Massachusetts), on the third Monday in April – the holiday is essentially Veterans Day but for pre-Independence veterans of the Revolutionary War. In Boston, the day is celebrated with near-religious fervor, an unofficial beginning of Spring. Men, women, younger and older adults come from all over the world, but many come from our own communities. Most Bostonians know a group of people 'training for the big day.' As the oldest (first event held in 1897) and one of the most prestigious (it is one of the six World Marathon Majors1), the Boston Marathon attracts a broad spectrum of runners, from Olympic-level athletes to somewhat casual running enthusiasts. The organization which sponsors the race, the Boston Athletic Association ("BAA") pays respect to its non-profit/amateur founding mission by allowing runners to run for charity2; if a runner does not "qualify" for the race conventionally, the BAA maintains spots for runners to run for charity. This adds to the diversity of the population of runners. The questions this analysis will seek to explore is whether any insights can be gained from on the success of a runner when applied to the dimensions of age/generation, gender, and nationality.

Data Acquisition

About the Data

The data is made up of 26,657 observations (=runners) of 20 variables:

[1] See, https://www.worldmarathonmajors.com/six-star.
[2] See, https://www.boston.com/sports/boston-marathon/2019/03/15/donate-marathon-charity-teams/.

```
> str(BAA2019Results)
'data.frame':   26657 obs. of  20 variables:
 $ BibNumber         : chr  "2" "6" "7" "8" ...
 $ FullName          : chr  "Lawrence Cherono" "Lelisa Desisa" "Kenneth Kipkemoi" "Felix
andie" ...
 $ SortName          : chr  "Cherono, Lawrence" "Desisa, Lelisa" "Kipkemoi, Kenneth" "Kar
ie, Felix" ...
 $ AgeOnRaceDay      : chr  "30" "29" "34" "32" ...
 $ Gender            : chr  "M" "M" "M" "M" ...
 $ City              : chr  "Eldoret" "Ambo" "Eldoret" "Iten" ...
 $ StateAbbrev       : chr  NA NA NA NA ...
 $ StateName         : chr  NA NA NA NA ...
 $ Zip               : chr  NA NA NA NA ...
 $ CountryOfResAbbrev: chr  "KEN" "ETH" "KEN" "KEN" ...
 $ CountryOfResName  : chr  "Kenya" "Ethiopia" "Kenya" "Kenya" ...
 $ CountryOfCtzAbbrev: chr  "KEN" "ETH" "KEN" "KEN" ...
 $ CountryOfCtzName  : chr  "Kenya" "Ethiopia" "Kenya" "Kenya" ...
 $ OfficialTime      : chr  "2:07:57" "2:07:59" "2:08:07" "2:08:54" ...
 $ RankOverall       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ RankOverGender    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ RankOverDivision  : chr  "1" "2" "3" "4" ...
 $ EventGroup        : chr  "Runners" "Runners" "Runners" "Runners" ...
 $ SubGroupLabel     : chr  NA NA NA NA ...
 $ SubGroup          : chr  NA NA NA NA ...
```

- **BibNumber:** Identification number used by BAA as a unique identifier for each runner. Variable not required for any data analytics modeling and will be dropped.
- **FullName:** Loosely defined as complete name combination of "First", "Middle" and "Last" names. As an international event, this system could be complicated given different cultures' nomenclature customs. Variable is not needed for data analytics modeling and will be dropped.
- **SortName:** Similar to FullName, with "Last" name first. Again, not needed for modeling and will be dropped.
- **AgeOnRaceDay:** Age of the runner on day of race. Runners may be 17 when registering but must be at least 18 on race day. There is no upper age limit. As a variable, it is numeric and continuous, but could also be ordinal and used to discretize. BAA offers 11 age group classes: 18-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+. 11 is too many classifiers; thus, this analysis will simply discretize to four bins based on quartiles.
- **City, StateAbbrev, StateName:** City, State designations. As many runners are international, this observation contains a fair amount of NA's. Fortunately, it will not be used in any analysis and will be dropped.
- **CountryOfRes/CtzAbbrev & CountryOfRes/CtzName:** Observations for the full country name and its attendant abbreviation. The dichotomous Res vs. Ctz relationship allows BAA to differentiate Citizenship designation from Residence. In most, cases, particularly for casual runners these values will likely be the same. For international "elite" or professionals, a runner may be a Citizen of say Kenya, but may 'train' in the US or UK. This analysis will explore the records where there is a difference but will decide on "CountryOfCtz" as the focus of analysis of runners' provenance. The variable will be used to discretize runners into regional classifiers.
- **OfficialTime:** The official finish time of each runner entered in "hh:mm:ss" format. This is the one true continuous variable in the data set and is the key target for any linear regression models. The observation will require transformation into seconds.
- **RankOverall, RankOverGender, RankOverDivision:** Rank number for each runner based on overall, within gender class and within one of the 11 age classes. This analysis will only consider RankOverall whereby it will discretize the set into quadrants "Top 25%", "Upper Middle 25%", "Lower Middle 25%", and "Bottom 25%". As this variable sorts the data set in order, steps will be taken to

"shuffle" the data set so that results are not skewed by position in the data set, but this marker will allow a somewhat general performance distinction for every runner, without a precise rank designation.

- **EventGroup, SubGroupLabel, SubGroup:** Participation in the Boston Marathon is not limited to runners; there are groupings for wheelchair, handcycles, vision impaired, etc. This data set and analysis only covers bipedal runners. Thus, this variable will be deleted.

```
> Keeps <- c("AgeOnRaceDay","Gender","CountryOfResName","CountryOfCtzName","OfficialTime","RankOverall")
> BAA2019Results <- BAA2019Results[Keeps]
> str(BAA2019Results)
'data.frame':   26657 obs. of  6 variables:
 $ AgeOnRaceDay    : chr  "30" "29" "34" "32" ...
 $ Gender          : chr  "M" "M" "M" "M" ...
 $ CountryOfResName: chr  "Kenya" "Ethiopia" "Kenya" "Kenya" ...
 $ CountryOfCtzName: chr  "Kenya" "Ethiopia" "Kenya" "Kenya" ...
 $ OfficialTime    : chr  "2:07:57" "2:07:59" "2:08:07" "2:08:54" ...
 $ RankOverall     : int  1 2 3 4 5 6 7 8 9 10 ...
```

Experimental Design

Cleaning Transformation

Checking for NA's

After removing unnecessary and problematic columns which contain all of the NA's, the data set conveniently has no NA's which need to be dealt with.

```
> sum(is.na(BAA2019Results))
[1] 0
```

However, in checking for unique values for classification, a problem in "Gender" is discovered. It would appear that one record was entered incorrectly either through a typo or some other human/data entry error.

```
> unique(BAA2019Results$Gender)
[1] "M"                "F"                "Leighton Buzzard"
```

"Leighton Buzzard" is a town. The runner in question has Bib Number 28913.

```
> BAA2019Results[BAA2019Results$Gender=="Leighton Buzzard",]
      BibNumber AgeOnRaceDay           Gender CountryOfResName CountryOfCtzName OfficialTime RankOverall
26143     28913            F Leighton Buzzard              GBR          5:46:30        26143       11695
```

From a look up at the BAA results website, the correct data was found to be and was corrected in an atomic fashion, one observation updated at at time:

Your search returned 1 result(s).
Criteria: Bib Number 28913, Race Year 2019, sorted by Last Name.

YEAR	BIB	NAME	AGE	M/F	CITY	STATE	COUNTRY
2019	28913	O"Connell, Pamela	62	F	Leighton Buzzard		GBR

Overall	Gender	Division	Official Time	Net Time
26143 / 26657	11695 / 11982	414 / 430	6:04:34	5:46:30

```
> BAA2019Results$Gender[26143]="F"
> BAA2019Results$OfficialTime[26143]="6:04:34"
> BAA2019Results$RankOverall[26143]="26143"
> BAA2019Results$CountryOfCtzName[26143]="United Kingdom"
> BAA2019Results$AgeOnRaceDay[26143]="62"
> BAA2019Results[BAA2019Results$BibNumber=='28913',]
      BibNumber AgeOnRaceDay Gender CountryOfResName CountryOfCtzName OfficialTime RankOverall
26143     28913           62      F              GBR   United Kingdom      6:04:34       26143
```

Once achieving the proper binomial Gender variable, look at the count shows a slightly greater number of Male runners vs Female runners, but not an extreme skew either way:

```
> table(BAA2019Results$Gender)

     F      M
 11982  14675
```

The distribution of completion times by gender shows a density of male population of finishers whereby the median roughly equates to the fastest lower quantile of the female runners, but generally there is overlap between the two distributions. There appears to be a population of outliers beginning ~20,000 seconds for both genders (for finishing time transformation, see below).

Transforming Official Time into Seconds & Creating "Pace" variable

The data set has precious few continuous variables. The key one and probably the most interesting possible dependent variable (were linear regression a model this analysis had planned) is finish time. Unfortunately, result time is currently in "time" format "HH:MM:SS"; using lubridate package, the analysis will transform time into seconds, a numerical value that will allow for regression.

```
> summary(BAA2019Results$OTSeconds)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7677   12026   13534   13982   15540   27828
```

The distribution of finish times appears to be in a bell-shaped curve, approaching a normal distribution. There's a slight right skew with median value to the left of the mean value. The population size is nearly 30,000 which is somewhat large. If the analysis were to consider more years than just one, the number of observations (=population) would increase and we'd likely see a more normal distribution shape.

Location Mismatches

There are 93 unique CountryOfResName values. There are 109 unique CountryOfCtzName values. This analysis will designate "CountryOfCtzName" for all runner Regional analytics. For point of interest, the accompanying R code provides an analysis of the number of instances in which CountryOrResName and CountryOfCtzName do not match. Since CountryOfResName will be dropped, this analysis will make no further investigation of the mismatches.

```
> length(unique(BAA2019Results$CountryOfResName))
[1] 93
> length(unique(BAA2019Results$CountryOfCtzName))
[1] 109
```

Since it's not a critical distinction for this analysis, a detailed sorted list of countries that "mismatches" and the most frequently occurring instances is only provided in the accompanying RMD file, for reference. However, this discrepancy provides a potential clue that some runners have a "professional" status whereby they may be FROM country X (Ctz) but RESIDE (Res, perhaps for training) in another.

Classification

The dataset comes handily equipped with several ready-made classification distinctions: Gender (M/F), Age/Group, Country (Regional). Below describes the process of discretizing the data into classes.

Region Classifier

Using the countrycode package, a new variable "Region" is created, taking CountryOfCtzName and applying the appropriate "Region."

```
> (table(BAA2019Results$Region))

      East Asia & Pacific      Europe & Central Asia  Latin America & Caribbean Middle East & North Africa
                     1734                       2823                       1233                         44
            North America                  South Asia         Sub-Saharan Africa
                    20663                         93                         67
```

The overwhelming majority of runners are from North America (20,663). Europe/Central Asia (2,823), East Asia & Pacific (1,734) and Latin America/Caribbean (1,233) each has a considerable number of runners, although if combined amount to just over a 25% of North America's representation. Regions with smaller representation are South Asia (93), Sub-Saharan Africa (67) and Middle East/North Africa (44). As the top runners in Marathons will likely be from Kenya or Ethiopia[3], it will be interesting to see this dynamic play out with a Sub-Saharan Africa Region bucket amounts to just 0.25% of the entire number of runners (26,657).

[3] https://www.runnersworld.com/advanced/a20842528/the-ethiopia-kenya-running-phenomenon/

Age Group Classifier

After transforming AgeOnRaceDay to a numeric value, into 4 quartile variables: "Youngest" (7,003 runners, aged 18-34), "Second Youngest" (12,725 runners, aged 35-50), "Older" (6,440 runners, aged 51-66) and "Oldest" (489 runners, aged 67-83).

```
> summary(BAA2019Results$AgeOnRaceDay[(BAA2019Results$AgeQuartiles=="Youngest")])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   18.0    26.0    29.0    28.4    31.0    34.0
> summary(BAA2019Results$AgeOnRaceDay[(BAA2019Results$AgeQuartiles=="Second Youngest")])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35.00   39.00   43.00   42.62   46.00   50.00
> summary(BAA2019Results$AgeOnRaceDay[(BAA2019Results$AgeQuartiles=="Older")])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  51.00   53.00   56.00   56.75   60.00   66.00
> summary(BAA2019Results$AgeOnRaceDay[(BAA2019Results$AgeQuartiles=="Oldest")])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  67.00   68.00   69.00   69.98   71.00   83.00
```

```
> table(BAA2019Results$AgeQuartiles)

      Youngest Second Youngest           Older          Oldest
          7003           12725            6440             489
```

Obviously, the "Oldest" category has the smallest group at 489. Oldest's distribution of Finish Times amounts shows a somewhat uniform distribution. Youngest, Second Youngest and Older have more normal distributions, all skewing right, with Second Youngest showing the greatest frequency around the median, naturally as the largest group. One might expect the youngest group to have the largest number of runners, but instead it's the Second Youngest group. Some factors that might explain this might be age limit (must be 18) and/or economic circumstances. A younger runner may not have the resources to pay for the race or accommodate a training regimen. Also, running a marathon is something a runner "works up to". It may be an interesting question to see how many runners are first time runners, what age those runners are and what had been their running resume before the Marathon (5K/10K/Half-Marathon). Unfortunately, that insight can't be inferred from the dataset as it is.

Rank Classifier

RankOverall was used to discretize four groups by 25% cuts: "Top 25%" (6,665 runners), "Upper Middle 25%" (6,664 runners), "Lower Middle 25%" (6,664 runners), and "Bottom 25%" (6,664 runners).

Histograms are provided in the R file which show with a smaller population size, each distribution is approaching a uniform shape. Using a runner's exact position in the race would not be very productive and could lead to some overfitting prediction issues, this analysis wanted to include a distinction for each runner in terms of general position in the race rank order.

Achieved Qualifying Time Classifier

Every few years, BAA puts out a list of qualifying times ("QT"). It's broken up by 11 age groups and then divided into Male and Female runners, thus creating 22 qualification categories. To outright qualify for the Boston Marathon, a runner must have run a race in the proximate years previously that satisfies the QT according to the runner's Gender and Age Group.

This analysis created a new data frame reflecting the QT for 2019:

	MenQT	WomenQT	MenQTSeconds	WomenQTSeconds
18–34	3:00:00	3:30:00	10800	12600
35–39	3:05:00	3:35:00	11100	12900
40–44	3:10:00	3:40:00	11400	13200
45–49	3:20:00	3:50:00	12000	13800
50–54	3:25:00	3:55:00	12300	14100
55–59	3:35:00	4:05:00	12900	14700
60–64	3:50:00	4:20:00	13800	15600
65–69	4:05:00	4:35:00	14700	16500
70–74	4:20:00	4:50:00	15600	17400
75–79	4:35:00	5:05:00	16500	18300
80 and Over	4:50:00	5:50:00	17400	21000

Once transformed into a numerical "seconds" variable, each observation was created as its own variable:

```r
#Qualifyingg Times by Gender/Age Group Each As Variable
M1834QT <- BAA2019QT[1,3]
M3539QT <- BAA2019QT[2,3]
M4044QT <- BAA2019QT[3,3]
M4549QT <- BAA2019QT[4,3]
M5054QT <- BAA2019QT[5,3]
M5559QT <- BAA2019QT[6,3]
M6064QT <- BAA2019QT[7,3]
M6569QT <- BAA2019QT[8,3]
M7074QT <- BAA2019QT[9,3]
M7579QT <- BAA2019QT[10,3]
M80PLUSQT <- BAA2019QT[11,3]
F1834QT <- BAA2019QT[1,4]
F3539QT <- BAA2019QT[2,4]
F4044QT <- BAA2019QT[3,4]
F4549QT <- BAA2019QT[4,4]
F5054QT <- BAA2019QT[5,4]
F5559QT <- BAA2019QT[6,4]
F6064QT <- BAA2019QT[7,4]
F6569QT <- BAA2019QT[8,4]
F7074QT <- BAA2019QT[9,4]
F7579QT <- BAA2019QT[10,4]
F80PLUSQT <- BAA2019QT[11,4]
```

Then a two-part process created a "QTClass" variable reflecting the appropriate grouping. Then, each observation that equated a particular QTClass AND was less than/equal to the matching QT seconds variable as assigned to a Boolean 1, or else assigned to 0.

```r
#Create Classification Variable of Each QT Group by Gender/Age
BAA2019Results$QTClass[(BAA2019Results$Gender=='M' & BAA2019Results$AgeOnRaceDay>=18 & BAA2019Results$AgeOnRaceDay<=34)] <- "M1834"
BAA2019Results$QTClass[(BAA2019Results$Gender=='M' & BAA2019Results$AgeOnRaceDay>=35 & BAA2019Results$AgeOnRaceDay<=39)] <- "M3539"
BAA2019Results$QTClass[(BAA2019Results$Gender=='M' & BAA2019Results$AgeOnRaceDay>=40 & BAA2019Results$AgeOnRaceDay<=44)] <- "M4044"
BAA2019Results$QTClass[(BAA2019Results$Gender=='M' & BAA2019Results$AgeOnRaceDay>=45 & BAA2019Results$AgeOnRaceDay<=49)] <- "M4549"
BAA2019Results$QTClass[(BAA2019Results$Gender=='M' & BAA2019Results$AgeOnRaceDay>=50 & BAA2019Results$AgeOnRaceDay<=54)] <- "M5054"
BAA2019Results$QTClass[(BAA2019Results$Gender=='M' & BAA2019Results$AgeOnRaceDay>=55 & BAA2019Results$AgeOnRaceDay<=59)] <- "M5559"
```

```
> table(BAA2019Results$QTClass)

  F1834   F3539   F4044   F4549   F5054   F5559   F6064   F6569   F7074   F7579 F80PLUS   M1834   M3539   M4044   M4549   M5054
   3867    1793    1840    1835    1200     829     430     141      37       9       1    3136    1956    1989    2447    1858
  M5559   M6064   M6569   M7074   M7579 M80PLUS
   1552    1086     469     132      39      11
```

```
BAA2019Results$AchievedQT[(BAA2019Results$QTClass=="M1834"&BAA2019Results$OTSeconds>=M1834QT)] <- 1
BAA2019Results$AchievedQT[(BAA2019Results$QTClass=="M3539"&BAA2019Results$OTSeconds>=M3539QT)] <- 1
BAA2019Results$AchievedQT[(BAA2019Results$QTClass=="M4044"&BAA2019Results$OTSeconds>=M4044QT)] <- 1
BAA2019Results$AchievedQT[(BAA2019Results$QTClass=="M4549"&BAA2019Results$OTSeconds>=M4549QT)] <- 1
BAA2019Results$AchievedQT[(BAA2019Results$QTClass=="M5054"&BAA2019Results$OTSeconds>=M5054QT)] <- 1
```

```
> table(BAA2019Results$AchievedQT)

    0     1
10122 16535
```

```
> table(BAA2019ModelsDF$AchievedQT)

   No   Yes
10122 16535
> (YesProb <- 16535/sum(table(BAA2019ModelsDF$AchievedQT)))
[1] 0.6202874
> (NoProb <- 10122/sum(table(BAA2019ModelsDF$AchievedQT)))
[1] 0.3797126
```

This variable will serve as the object classifier in the models I will be running. There is a slight skew in the data: 62% Yes/AchievedQT and 38% No/AchievedQT. The difference is not so stark that the analysis may not proceed. Exploration of balancing the data set may be an option to be revisited at a later stage.

Model Building

Planned Models

Next, the analysis will analyze the data set to determine whether a runner achieved qualifying time based on gender, rank, age group and region. The classification models being used are:

- **Decision Tree**
- **Support Vector Machines (SVM)**
- **Random Forest (RF)**

Model Preparation

After creating a data frame of only the variables needed for the analysis and removing any that may not be independent (RankOrder/OTSeconds, AgeQuartile/AgeOnRaceDay), the analysis randomizes the rows (which hitherto have been in rank order and therefore need to be 'shuffled'). Then using a 70/30 split, a "train" and "test" set are created.

```
set.seed(123)
ind <- sample(nrow(BAA2019ModelsDF), replace = T)
BAA2019_1 <- BAA2019ModelsDF[ind,]

dt = sort(sample(nrow(BAA2019_1), nrow(BAA2019_1)*.7))
train<-BAA2019_1[dt,]
test<-BAA2019_1[-dt,]
```

An initial set of models are created using CV with K-Fold of 10 as a conventionally recognized starting point. Each model is then run with a tuneLength definision of 10.

```
# Creating a control with cross validation starting at 10
control <- trainControl(method ='cv',number = 10)
metric <-  "Accuracy" # Metric for comparison will be accuracy for this project

## Train models to predict whether Runner Achieved QT

tree.model1 <- train(AchievedQT ~ ., data = train, method="rpart", metric=metric, trControl=
                     tuneLength = 10) # Decision Tree

svm.model1 <- train(AchievedQT ~ ., data = train, method="svmRadial",metric=metric,trControl
                    tuneLength = 10) # Support Vector Machine (SVM)

rf.model1 <- train(AchievedQT ~ ., data = train, method="rf", metric=metric, trControl=contr
                   tuneLength = 10) # Random Forest
```

Model Evaluation

A look at all three models' accuracy shows them all at about 80%, with a tighter confidence interval around Decision Tree. All three have a Cohen's Kappa score between 55-60%, which has a "moderate" significance. It's so-so. Better than a coinflip but not by much. Decision Tree has the highest Kappa at 58%.

Decision Tree's accuracy quality degrades as the model complexity increases, with a gradual decline beginning around 2-3, falling starkly at around 7-8. It may be worth trying to limit the k-fold number and adjusting the tuneLength parameter.

SVM's accuracy peaks early and drops precipitously just after 10. There's a second bump up around 15 and then a flatline. Again, an attempt will be made to adjust parameters to improve accuracy.

As with the other two models, the accuracy performance of Random Forest jumps at about 3 but then promptly levels off. As with the other, models parameter tuning will be attempted.

Model Prediction

Decision Tree Conf.Matrix	SVM Conf.Matrix	RF Conf.Matrix

`> (DTCM <- confusionMatrix(test$AchievedQT,predic`
`Confusion Matrix and Statistics`
 	
`Reference`
`Prediction No Yes`
`No 2283 799`
`Yes 717 4199`
 	
`Accuracy : 0.8105`
`95% CI : (0.8017, 0.819)`
`No Information Rate : 0.6249`
`P-Value [Acc > NIR] : < 2e-16`
 	
`Kappa : 0.5979`
 	
`Mcnemar's Test P-Value : 0.03749`
 	
`Sensitivity : 0.7610`
`Specificity : 0.8401`
`Pos Pred Value : 0.7408`
`Neg Pred Value : 0.8541`
`Prevalence : 0.3751`
`Detection Rate : 0.2854`
`Detection Prevalence : 0.3853`
`Balanced Accuracy : 0.8006`
 	
`'Positive' Class : No`
	`> (SVMCM <- confusionMatrix(test$AchievedQT,p`
	`Confusion Matrix and Statistics`
	
	`Reference`
	`Prediction No Yes`
	`No 2144 938`
	`Yes 609 4307`
	
	`Accuracy : 0.8066`
	`95% CI : (0.7977, 0.8152)`
	`No Information Rate : 0.6558`
	`P-Value [Acc > NIR] : < 2.2e-16`
	
	`Kappa : 0.5834`
	
	`Mcnemar's Test P-Value : < 2.2e-16`
	
	`Sensitivity : 0.7788`
	`Specificity : 0.8212`
	`Pos Pred Value : 0.6957`
	`Neg Pred Value : 0.8761`
	`Prevalence : 0.3442`
	`Detection Rate : 0.2681`
	`Detection Prevalence : 0.3853`
	`Balanced Accuracy : 0.8000`
	
	`'Positive' Class : No`
		`> (RFCM <- confusionMatrix(test$AchievedQT,p`
		`Confusion Matrix and Statistics`
		
		`Reference`
		`Prediction No Yes`
		`No 2082 1000`
		`Yes 564 4352`
		
		`Accuracy : 0.8045`
		`95% CI : (0.7956, 0.8131)`
		`No Information Rate : 0.6692`
		`P-Value [Acc > NIR] : < 2.2e-16`
		
		`Kappa : 0.576`
		
		`Mcnemar's Test P-Value : < 2.2e-16`
		
		`Sensitivity : 0.7868`
		`Specificity : 0.8132`
		`Pos Pred Value : 0.6755`
		`Neg Pred Value : 0.8853`
		`Prevalence : 0.3308`
		`Detection Rate : 0.2603`
		`Detection Prevalence : 0.3853`
		`Balanced Accuracy : 0.8000`
		
		`'Positive' Class : No`

All three models have an accuracy level of approximately 80% with Decision Tree slightly in the lead with 81.05% on prediction. Decision Tree also has the best Kappa value at .5979, just edging out SVM at .5834. All three have a low P-Value in McNemar's Test, all under 0.05 alpha level but Decision Tree is just barely under at 0.03749.

Random Forest has a slightly higher Sensitivity and No-Information rate. However, it's difficult to pick a clear winner. All three "get it right" and "get it wrong" but not all at the same time, as evidenced below:

```
> head(predictionTest,10)
    dt svm random_f test.AchievedQT
1   No Yes      Yes            Yes
2   No  No       No             No
3   No  No       No             No
4  Yes Yes      Yes            Yes
5   No Yes      Yes            Yes
6  Yes Yes      Yes             No
7  Yes Yes      Yes            Yes
8   No  No       No             No
9   No  No       No             No
10 Yes Yes      Yes            Yes
```

Model Tuning

After finding a cost in higher CV K-Fold values, a new set of models are created using CV with K-Fold of 3, since each model showed a drop-off in accuracy beyond this number. Each model is then run with a tuneLength definition of 5.

Tuning Model Evaluation

A look at all three models' accuracy shows them all three models still around 80, but Decision Tree has improved somewhat. Of note, is the better Kappa values with Decision Tree's confidence interval running to 60+, with brings the model out of "moderate" significance to "good." Albeit, modest, there has been an improvement.

Decision Tree Accuracy Plot

SVM Accuracy Plot

RF Accuracy Plot

Decision Tree shows a flattening at 2 and Random Forest at 4. It's unlikely that additional tuning will be able to improve the accuracy levels of the models in any significant way.

Model Prediction

Decision Tree Conf.Matrix

```
> (DT2CM <- confusionMatrix(test$AchievedQT,p
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2042 1040
       Yes  549 4367

               Accuracy : 0.8013
                 95% CI : (0.7924, 0.81)
    No Information Rate : 0.676
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5678

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7881
            Specificity : 0.8077
         Pos Pred Value : 0.6626
         Neg Pred Value : 0.8883
             Prevalence : 0.3240
         Detection Rate : 0.2553
   Detection Prevalence : 0.3853
      Balanced Accuracy : 0.7979

       'Positive' Class : No
```

SVM Conf.Matrix

```
> (SVM2CM <- confusionMatrix(test$AchievedQ
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2119  963
       Yes  587 4329

               Accuracy : 0.8062
                 95% CI : (0.7974, 0.8148)
    No Information Rate : 0.6617
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5814

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7831
            Specificity : 0.8180
         Pos Pred Value : 0.6875
         Neg Pred Value : 0.8806
             Prevalence : 0.3383
         Detection Rate : 0.2649
   Detection Prevalence : 0.3853
      Balanced Accuracy : 0.8006

       'Positive' Class : No
```

RF Conf.Matrix

```
> (RF2CM <- confusionMatrix(test$AchievedQ
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  2146  936
       Yes  609 4307

               Accuracy : 0.8068
                 95% CI : (0.798, 0.8154)
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.584

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7789
            Specificity : 0.8215
         Pos Pred Value : 0.6963
         Neg Pred Value : 0.8761
             Prevalence : 0.3445
         Detection Rate : 0.2683
   Detection Prevalence : 0.3853
      Balanced Accuracy : 0.8002

       'Positive' Class : No
```

All three models have an accuracy level of approximately 80% this time Random Forest slightly in the lead with 80.68% on prediction, although overall not an improvement over the initial models. Kappa values went down overall, the highest being Random Forest at .584. Of note, all retain a low P-Value in McNemar's Test, all under 0.05 alpha level this time including Decision Tree where the P-Value improved somewhat.

Decision Tree this time has the slightly higher Sensitivity and No-Information rate. However, it's difficult to pick a clear winner. All three still "get it right" and "get it wrong" but not all at the same time, as evidenced below:

```
> head(predictionTest2, 10)
   dt2 svm2 random_f2 test.AchievedQT
1  Yes  Yes       Yes             Yes
2   No   No        No              No
3   No   No        No              No
4  Yes  Yes       Yes             Yes
5  Yes  Yes       Yes             Yes
6  Yes  Yes       Yes              No
7  Yes  Yes       Yes             Yes
8   No   No        No              No
9   No   No        No              No
10 Yes  Yes       Yes             Yes
```

Conclusion

The Boston Marathon brings together runners from all over the globe. As such, the population is diverse. The accuracy and performance of the models are "so-so". 80% accuracy with ~60 Kappa is a decent result but leaves room for improvement. Things that cannot be known from the data set are very likely to impact the outcome. These include a runner's physique (height/weight), how first marathon or number of marathons previous, first time to the Boston Marathon; what is the training regimen; are you professional/sponsored. It is this analysis' contention that knowing these factors could improve the predictive quality of the models presented. Also, this analysis arrived classification based on convenient categories already provided by the dataset and by the author's domain knowledge of the event type. Early experiments with clustering did not prove fruitful in terms of designating previously undiscovered classes. Further, exploration may find useful insights into the complexity of the population.

Learning Outcomes & Reflection

○	●	●	○	●	●	●
Describe a broad overview of the major practice areas in data science.	Collect and organize data.	Identify patterns in data via visualization, statistical analysis, and data mining.	Develop alternative strategies based on the data.	Develop a plan of action to implement the business decisions derived from the analyses.	Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.	Synthesize the ethical dimensions of data science practice.

This exercise gave me a wealth of hands-on R keyboard experience. From solving the problem of accessing data, to translating a key continuous variable (finish time) from HH:MM:SS to a discrete numeral (number of seconds) to deciding to utilize domain knowledge that I have over competitive races, my data science skills grew leaps and bounds. Understanding the path from data collection to exploration to modeling to find an answer will often raise as many questions as are answered. A data scientist must be ready to adapt to the challenges that a particular dataset gives us. For instance, we rely on a binary variable such as "Gender", which in this case, is still received as a binary. There will come a day where gender spectrum will need to be respected and a binary variable may not be what we can count on. Data scientists must be prepared for this.

- **Collect and organize data:**
  Choosing my data set of Boston Marathon results provided and interesting challenge – the data has several built-in categories (age, gender, location), but it's ranked. Care had to be taken to randomize the data set so that classification models would not get stuck with skewed results, or poorly performing predictions.

- **Identify patterns in data via visualization, statistical analysis, and data mining:**
  Like I said, the data came with some ready and domain-standard categories: gender, age, location. Using these features, I was able to mine for additional categories (generation, region). Also, an auxiliary data source (BAA qualifying times by age/gender group), allowed me to create what would become my main dependent categorical value and question I was answering: how likely is it that a runner will achieve the qualifying time.

- **Develop a plan of action to implement the business decisions derived from the analyses:**
  Collecting the data, performing a thorough exploratory data analysis ('EDA'), plus my own domain knowledge around qualifying times for the Marathon, allowed me to construct the plan, develop my data set and test three classification models with an attempt to adjust tuning parameters to achieve greater accuracy.

- **Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization:**
  My project description entails a step-by-step narration of my process, steps taken to clean data, where needed, insights derived from EDA, and a full evaluation of model performance with suggestions on additional data which could improve the model. Consumers of the product data scientists should not have any doubt that the data has been capably explored and findings are clear.

Project 3– IST 769
Advanced Database Management

Course Description

An analysis of relational and nonrelational databases and their corresponding database management system architectures. Learn to build complex database objects to support a variety of needs from both the big data and traditional perspectives. Data systems performance, scalability, and security.

This course provides tour of relational, document, key-value, columnar, and streaming database systems through the lens of the CAP theorem. We will explore the strengths and weaknesses of various database systems in the relational, Hadoop, and noSQL spaces. Where possible you will experience these systems first-hand as to gain an understanding of how they can be used to address complex, big data challenges.

Learning Objectives:

After taking this course, the students will be able to:

- Understand advanced issues with the relational database model, such as transactions, performance, and security, as to understand the need for other database models.
- Explain the CAP theorem and describe how any given database system's architecture fits within the CAP context.
- Compare different database models such as document, key-value, column-family, streaming, and relational.
- Identify the most suitable database systems for a specific application's data storage requirements.
- Evaluate relational, Hadoop, and noSQL database tooling as to understand their underlying similarities and necessary differences.

Deliverable(s) – Weekly Coding Lab Assignments

Lab Homework assignments are technical activities which enforce asynchronous concepts through practice and are based on the demos in the course videos. Homework must be completed before the week's live session, where students must be prepared to discuss the outcomes of the assignment.

Learning Outcomes & Reflection

○	●	○	●	●	●	●
Describe a broad overview of the major practice areas in data science.	Collect and organize data.	Identify patterns in data via visualization, statistical analysis, and data mining.	Develop alternative strategies based on the data.	Develop a plan of action to implement the business decisions derived from the analyses.	Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.	Synthesize the ethical dimensions of data science practice.

The lectures (both live and asynchronous) and particularly the lab work provided me the opportunity for me to deepen my understanding of SQL, going into advanced topics not covered in IST659, such as transactions, views, stored procedures, temporal tables and security measures. The coverage of this material reinforced that the data scientist must be a steward of an organization's data, its most precious resource, and become familiar with ways to ensure its security. Digital threats are more the norm now than ever before. Also, organizations need to be concerned with scale and useability. SQL may not always be the solution. Exploration of the CAP theorem, and exploration of ETL/NoSQL programs like Hadoop, MongoDB and others provided an invaluable view into ways to structure data to meet data needs that may not be satisfied by RDBMS.