Portfolio Milestone
MS Applied Data Science

- Ryan Timbrook
- NetID: RTIMBROO

- Term: January/Winter, 2020

Portfolio - Course - Alignment
Overview

Course Projects included in Portfolio:

- IST 736 Text Mining
- IST 718 Big Data Analytics
- IST 707 Data Analytics
- IST 659 Database Administration Concepts and Database Management
- MBC 638 Data Analysis and Decision Making
- SCM 651 Business Analytics

Course Projects not included in Portfolio:

- ACC 652 Accounting Analytics
- MAR 653 Marketing Analytics
- IST 687 Introduction to Data Science
- IST 722 Data Warehouse
- IST 623 Intro to Information Security
- IST 600 Selected Topics – Transferred Credits – LING 570 Shallow Proc. for NLP
- CIS 668 Natural Language Processing – Transferred Credits – LING 571 Deep Proc. For NLP

Project - Course - Alignment Overview

- IST 736 Text Mining
  - Public Sentiment Toward NFL Teams, Coaches, and Players (Team Project)
- IST 718 Big Data Analytics
  - IEEE-CIS Fraud Detection (Team Project)
- IST 707 Data Analytics
  - Real Estate Property Investment (Individual Project)
- IST 659 Database Administration Concepts and Database Management
  - A4B KPI BizOps Organization Database (Individual Project)
- MBC 638 Data Analysis and Decision Making
  - Purchase Order Process Improvement (Individual Project)
- SCM 651 Business Analytics
  - Recruiting Advertising Strategy (Team Assignment)

Learning Objective - Project- Alignment Overview

- Collect and Organize Data
  - Public Sentiment Toward NFL Teams, Coaches, and Players
  - Real Estate Property Investment
  - Purchase Order Process Improvement

- Identify Patterns in data via visualization, statistical analysis, and data mining
  - Public Sentiment Toward NFL Teams, Coaches, and Players
  - IEEE-CIS Fraud Detection
  - Real Estate Property Investment

- Develop alternative strategies based on the data
  - Real Estate Property Investment
  - Recruiting Advertising Strategy

- Develop a plan of action to implement the business decisions derived from analysis
  - Purchase Order Process Improvement
  - A4B KPI BizOps Organization Schema

- Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals
  - Purchase Order Process Improvement
  - Public Sentiment Toward NFL Teams, Coaches, and Players
  - IEEE-CIS Fraud Detection

- Synthesize the ethical dimensions of data science practice
  - A4B KPI BizOps Organization Database
  - IEEE-CIS Fraud Detection

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

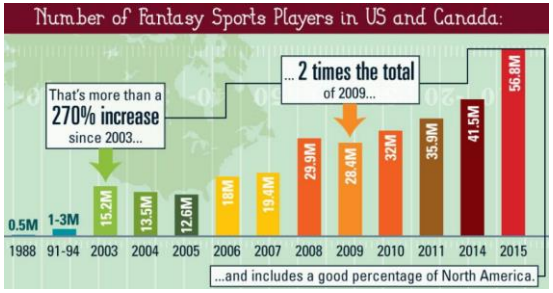
Question and Problem to Solve

Business Driver

- NFL revenue grew an estimated $900 million to $14 billion in 2017, in 2018 it generated about $15 billion.
- Fantasy football and the spread of legalized sports betting across the U.S. promises to lock in fans and keep them focused on the game.
- As a Fantasy football player, how can Data Science help me make the most intelligence selections when deciding my weekly roster? “How do I win more?”

Problem to Solve

- Identify if real-time public popular opinion on NFL teams, players, and coaches is a predictor of their weekly fantasy football stats.

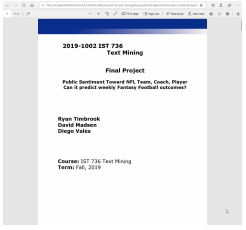


Fantasy Football is an $18.6 BILLION Market.

About the Data:

- Text data will be mined from three primary sources

1) Fantasy football websites like Yahoo Sports:
2) Twitter Social Media API live streams
3) Twitter Social Media Historical API search



***Link to Project Report Video | Link to Project Presentation Video

IST 718 Big Data Analytics
-IEEE-CIS Fraud Detection-
Question and Problem to Solve

Business Driver

- Improve the efficacy of fraudulent transaction alerts, helping hundreds of thousands of businesses reduce their fraud losses and increase their revenue; while securing consumer's peace of mind and wallets!

kaggle vesta

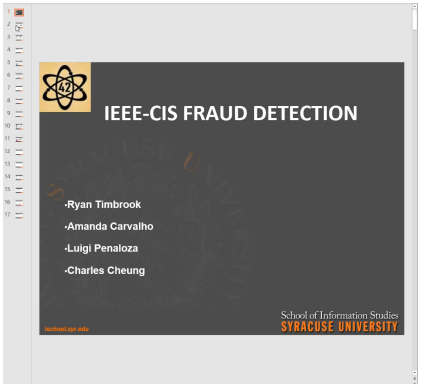
Problem to Solve

- Identify real-time fraudulent e-commerce transactions, using advanced Machine Learning algorithms, by automating alerts that block highly suspicious activities.

About the Data:

- The core data set for this project is provided by VESTA, the worlds leading payment service company, and is a kaggle competition being facilitated by the IEEE Computational Intelligence Society.
- The data is broken into two files **identity** and **transaction**, which are joined by **TransactionID**. *Not all transactions have corresponding identity information.*

***Link to Project Report Video



IST 707 Data Analytics
-Real Estate Property Investment-
Question and Problem to Solve

Business Driver

- Targeting low risk property investment opportunities for property management firms or individual investors who buy-rent-sell single family homes throughout the United States.

- Given a base set of investment criteria, provide a predicted N-best list of US geolocation regions by zip code that offer the best ROI.

Problem to Solve

- How to predict a low risk / high yield return on property investment in a volatile market.

- Buy low, rent fair, sell high…

- Where and when to buy and sell that maximizes investment profits.

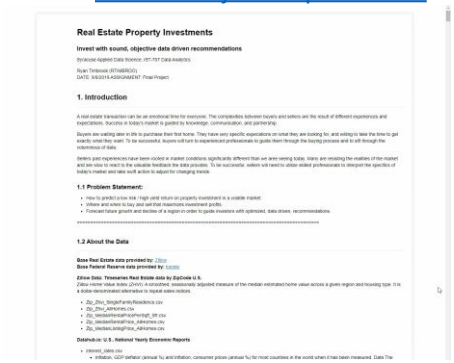
- Forecast future growth and decline of a region that yields Net Present Value (NPE) measurements significant enough to act on.

About the Data:

- Base datasets are provided by Zillow research data:
  - Home Values
  - Home Listings and Sales
  - Rental Values

- Additional data will be mined from:
  - Bureau of Labor Statistics and Census
  - Capital markets and economics

***Link to Project Report Video


Real Estate Property Investments

Invest with sound, objective data driven recommendations

1. Introduction

1.1 Problem Statement:

1.2 About the Data

IST 659 Database Administration & Management

-A4B KPI BizOps Organization Database-

Question and Problem to Solve

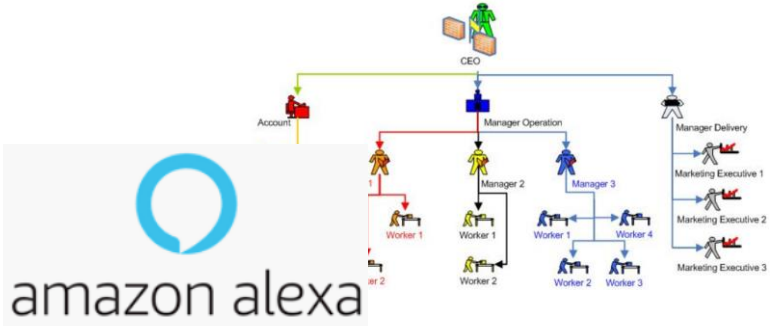
Business Driver

- Innovation of alternative Voice Channels for the Customer Service organization that are more cost effective than traditional telephony voice such as IVRs.

- Business and Operations KPIs are daily 'must know' factors that are challenging to get real-time updates on and take excessive resource time to pull reports together.

Problem to Solve

- No single source of truth exists for the Company's Organization Hierarchy that includes application and business operation domains along with their KPI mapping.



About the Data:

- Company Organization data is pulled from internal LDAP system.
- Application Ownership and KPI sources are gathered from internal audit and team interviews.

***Link to Project Report Video

MBC 638 Data Analysis & Decision Making
- Purchase Order Process Improvement-
Question and Problem to Solve

Business Driver

- 80% of vendor Purchase Orders (PO) for contingent staffing workforce are NOT being approved within the vendor offered 2% discount time period. On average, for each PO this is a missed opportunity of $5,911 in cost savings.
- Lengthy approval cycle-times and the hours software development team's management spend validating the accuracy of vendors PO Invoices leads to decreased product output hurting the business's competitive 'Time to Market' strategy.

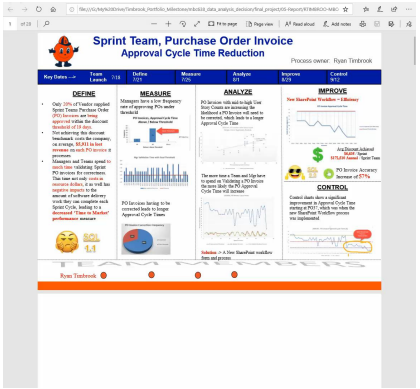
Problem to Solve

- Identify areas of waste within the PO approval workflow process.
- Devise a more efficient mechanism of collecting, organizing and reporting details of the work the vendor is submitting PO Invoices for.
- Implement a solution to be measured and establish control methods to ensure the process has improved the situation and is continually being enhanced.

About the Data:

- The data for this initiative comes from multiple disparate internal sources. It's generated from the Sprint Development Teams performing tasks associated with business deliverables. Systems like Rally and SharePoint capture the Sprint User Story tasks the Vendor submits PO Invoices for on a bi-weekly basis. PO Invoices are entered into SAP Ariba for internal processing including approvals for payment workflows.

***Link to Project Report Video



SCM 651 Business Analytics
-Recruiting Advertising Strategy-
Question and Problem to Solve

Business Driver

- Derive insights and knowledge from a Google ads advertisement recruitment campaign launched in 2011 for The Whitman School of Management in order to create a roadmap for 2020's new recruitment campaign.

Problem to Solve

- Identify time frames for each marketing campaign and the cost spent on each.
- Identify what the effectiveness was for each of the previous campaigns.
- Identify the key aspects of a United States campaign for 2020.
- Identify the costs for the advertising campaigns.
- Specify how to best measure performance of the decisions made after implementation.
- Specify other factors or considerations that are important.
- For 2020's campaign, stay within a $100,000 budget.

Analytics

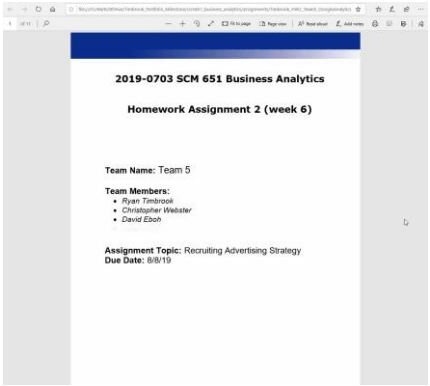
About the Data:

- Witman.syr.edu website user traffic analytics collected by google analytics from March 2011 to October 2012.

United States Campaign Measurements

Campaign	Start Date	End Date	Cost	Cost per Click	Clicks	Total Clicks	Total Sessions	Bounce Rate	Page per Session
whitman.syr.edu	2/15/2011	8/15/2011	$ 37,851.36	$ 3.93	133,732	98,379	192,134	48.08%	3.76
MBA Marketing – Full-time	Jan/15/2012	2/15/2012	$ 16,459.90	$ 4.64	1,347	1,348	1,468	97.07%	1.05
MBA Marketing – iMBA	2/1/2012	10/15/2012	$ 144,971.70	$ 7.46	2,753	2,748	3,079	88.96%	1.14
Delta	9/1/2013	11/1/2013	$ 10,000.00	$ -			284	28.17%	3.37

***Link to Project Report Video



2019-0703 SCM 651 Business Analytics

Homework Assignment 2 (week 6)

Team Name: Team 5

Team Members:
- Ryan Timbrook
- Christopher Webster
- David Eboh

Assignment Topic: Recruiting Advertising Strategy
Due Date: 8/8/19

Learning Outcome Objectives

- 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 analysis
- Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals
- Synthesize the ethical dimensions of data science practice

-Collect and Organize Data-
Overview

- Public Sentiment Toward NFL Teams, Coaches, and Players
- Real Estate Property Investment
- Purchase Order Process Improvement

Learning Objective
-Collect and Organize Data-

➢ Public Sentiment Toward NFL Teams, Coaches, and Players

The Bishop @BillBishopKHOU · Dec 4, 2019
Good morning. The NFL has named **Houston Texans** QB **Deshaun Watson** the AFC Offensive Player of the Week following his performance in the team's 28-22 victory over the New England Patriots in Week 13. #khou #HTownRush

DreamTeamApple @DreamTeamApple1 · Oct 28, 2019
Replying to @NFLFilms @deshaunwatson and 2 others
I've been disappointed with the **Texans** offensive line taking a step back with their pass protection since Tytus Howard's injury, but **Deshaun Watson** has been making sensational plays even when the pocket collapses! A lot like he did last year. **Deshaun Watson** is Michael Jordan!
1 · 15

AdvoSports @advosports · Dec 1, 2019
HOUSTON (AP) — **Deshaun Watson** threw three touchdown passes and had the first TD reception of his career, and the **Houston Texans** frustrated Tom Brady in a 28-22 victory over the New England Patriots on Sunday night.
victoriaadvocate.com/ap/sports/wats...
1 · 1

FleaFlickerFFB @FleaFlickerFFB · Nov 17, 2019
I'm taking **Deshaun Watson** and the **Houston Texans** to beat the Lamar Jackson led Baltimore Ravens.

Fun game in store between these two young QB's

LisaVoxNihili @LisaVoxNihili · Dec 22, 2019
Replying to @melissastark @HoustonTexans and @deshaunwatson
If you dislike **Deshaun Watson**, something is wrong with you. He is absolutely one of the nicest people on the planet. Just stating a fact, for those salty people in these comments.
5

Developer

Houston Texans
1st in AFC South

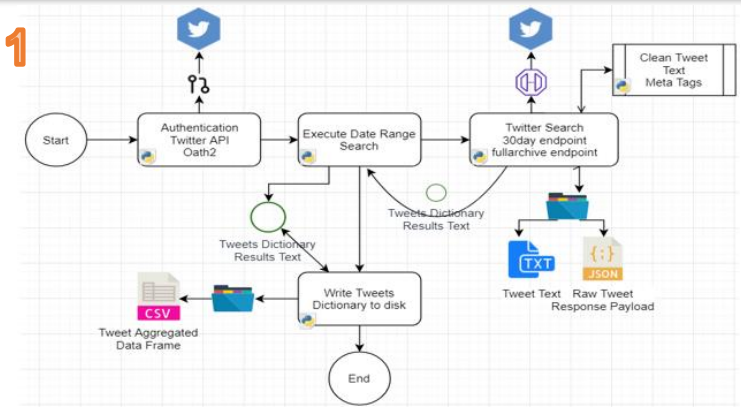
Collect and Organize Data
IST 736 Text Mining
-Public Sentiment Toward NFL Teams, Coaches, and Players-

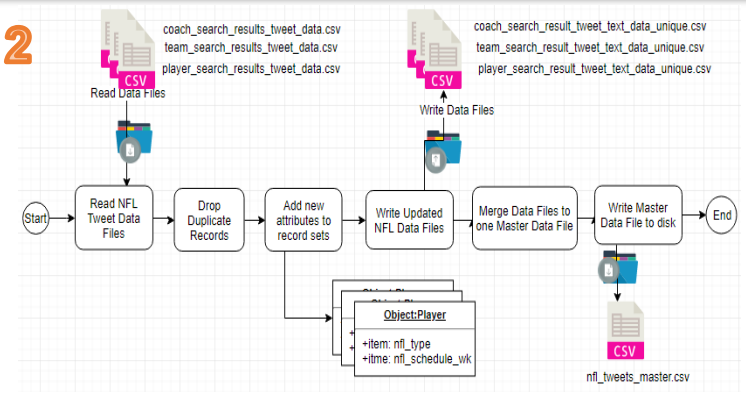
<u>Collection Methods:</u>

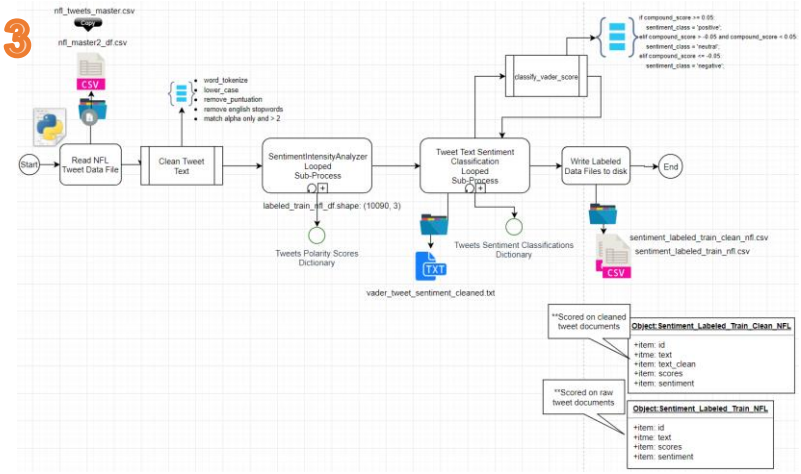
- Internet Web Scraping
- Twitter API Live Stream Capture
- Twitter API Historical Search Capture

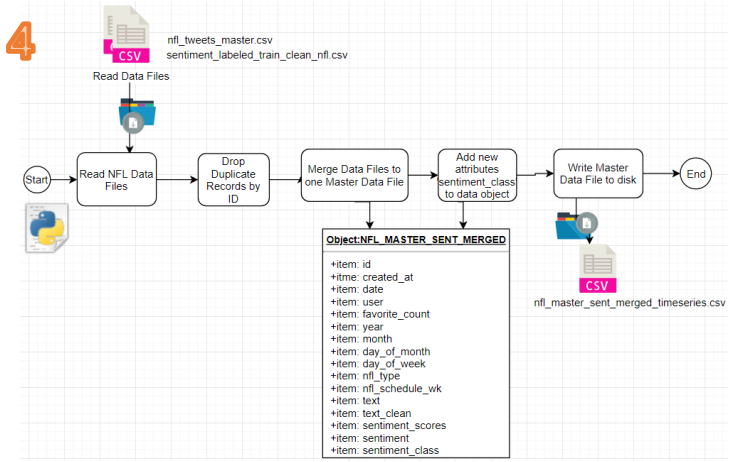
<u>Organization Methods:</u>

- NFL type object modeling
- Local file system, text corpus, storage and retrieval
- Data engineering pipeline







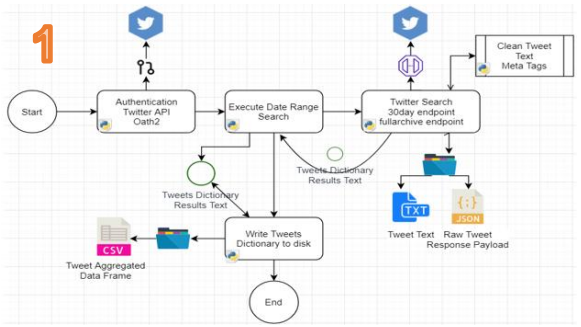


Web Scraping Code Overview | Twitter API Streaming Code Overview | Twitter API Search Code Overview | Datasets Merge Code Overview | Format Tweets JSON Code Overview | Sent Merge

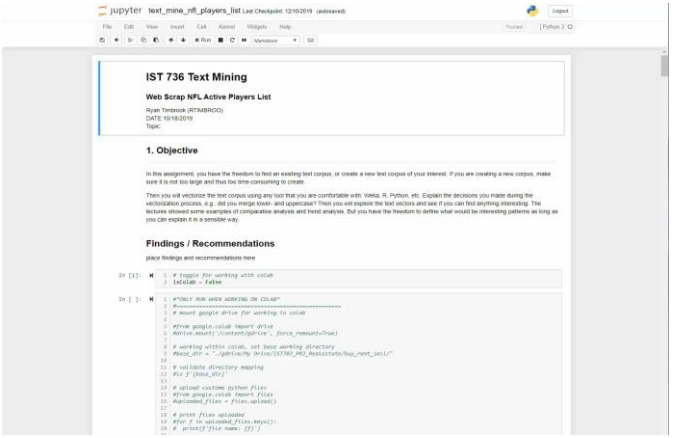
Collect and Organize Data
IST 736 Text Mining
-Public Sentiment Toward NFL Teams, Coaches, and Players-

Collection Methods:

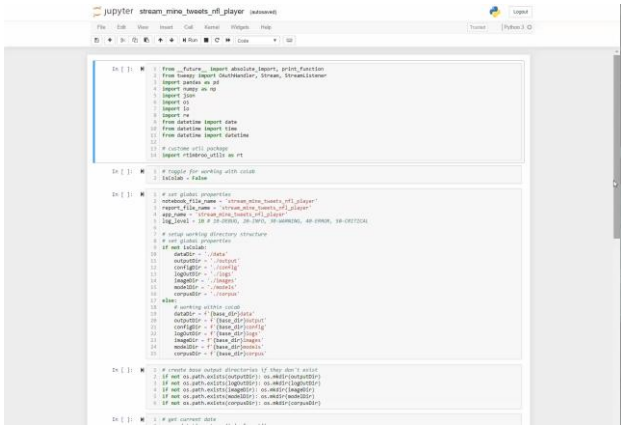
- Internet Web Scraping
- Twitter API Live Stream Capture
- Twitter API Historical Search Capture



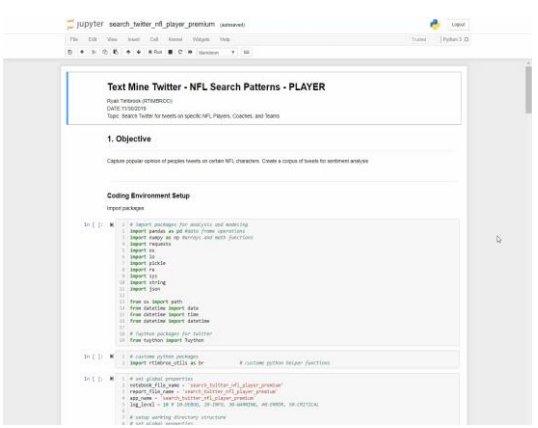
[Web Scraping Code Overview]



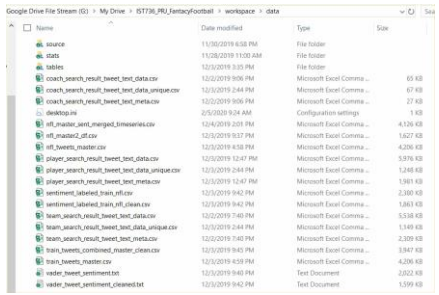
[Twitter API Streaming Code Overview]



[Twitter API Search Code Overview]



[Tweets Corpus Storage]

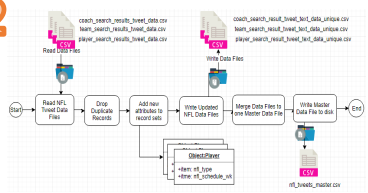


Collect and Organize Data
IST 736 Text Mining
-Public Sentiment Toward NFL Teams, Coaches, and Players-

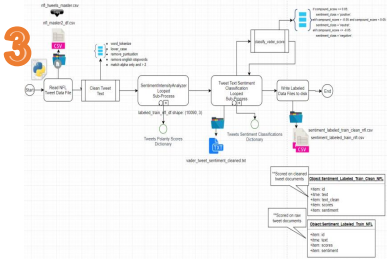
Organization Methods:

- NFL type object modeling
- Local file system, text corpus, storage and retrieval
- Data engineering pipeline

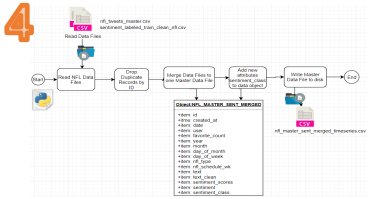
2



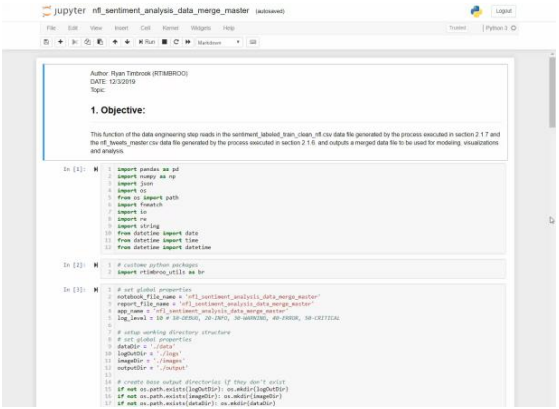
3



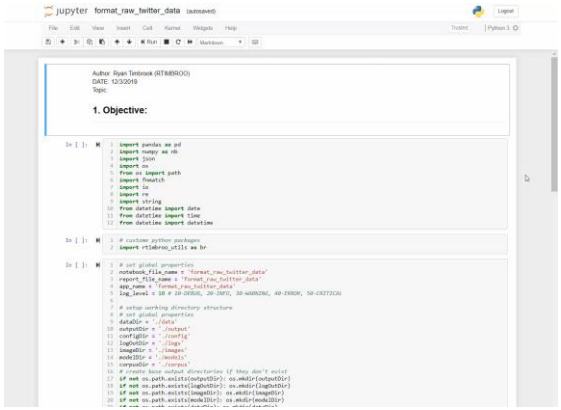
4



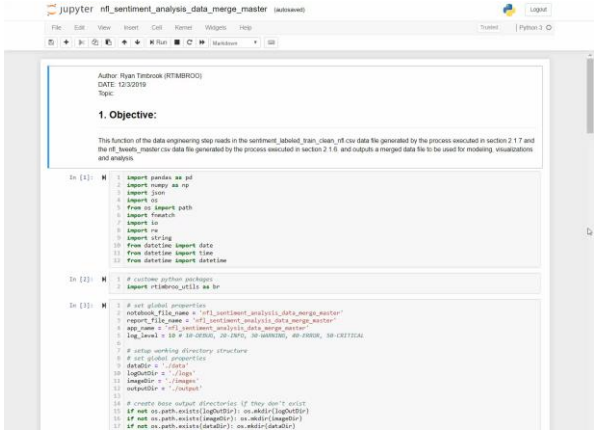
[Datasets Merge Code Overview]



[Format Tweets JSON Code Overview]



[Sentiment Analysis Merge]



Learning Objective
-Collect and Organize Data-

- Real Estate Property Investment

Zillow

Data Hub

Economic factors that affect Investment

Collect and Organize Data
IST 707 Data Analytics
-Real Estate Property Investment-

Collection Methods:

- Real Estate datasets pulled from Zillow Research web site.
- U.S. Economic indicators datasets pulled from Datahub.io

Organization Methods:

- Exploratory data analysis was performed on each of the datasets, five Zillow real estate and fourteen U.S. National Yearly Economic reports
- All datasets were reorganized and structured into timeseries format
- Data scrubbing, cleaning, transformation, and merging of disparate datasets was performed for analyses

** Merged all disparate Economic dataset features to one common source for analysis

Year	FF_Target_Rate_Avg	Inflation	GDP	GDP_Percent_Change	Education_Budget	Population	Investor_Flow_Avg	National_HPI_Avg	House_Hold_Income	Employed	Employed_Percent	Unemployed	Unemployed_percent	CPI_Index_Avg	Cash_Surp_Def	Bond_Yield_10y_Avg
1982	9.392857	6.203740	3345.0	8.8	15374.0	231664000.0	NaN	NaN	83918	99526	57.8	10678	9.7	96.500000	-2.735732	13.001667
1983	9.053125	3.948367	3638.1	11.1	15267.0	233792000.0	NaN	NaN	85290	100834	57.9	10717	9.6	99.600000	-4.819500	11.105000
1984	10.150000	3.548237	4040.7	7.6	15336.0	235825000.0	NaN	NaN	86789	105005	59.5	8539	7.5	103.883333	-3.676845	12.438333
1985	8.044643	3.199612	4346.7	5.6	18952.0	237924000.0	NaN	NaN	88458	107150	60.1	8312	7.2	107.566667	-3.927316	10.623333
1986	6.740132	2.017624	4590.2	6.1	17750.0	240133000.0	NaN	NaN	89479	109597	60.7	8237	7.0	109.608333	-4.248440	7.682500

Final Merged Dataset - Real Estate Combined with Economic Data Features

Time range - 1997 - 2017 (that was the cleanest that could be achieved at this time...
*Train classifiers on Feature 'Price_Point_Class'

- 0: means observation's price value is < 25% of the State Price Average
- 1: means observations fall within the normal (average) range of the State Price Average
- 2: means observations falls above the 75% range of the State Price Average

--Determin if classifiers can identify future home value classes based on prior date, location and economic features that have the most impact on both postive and negative price value swings...

- Dataset Shape: (88452, 16)

Date	ZipCode	log_Price	log_Price_Monthly_Avg	log_Price_diff	Price_Point_Class	CPI_Index_Avg_f	Interest_Rate_f	Housing_Price_Index_f	Bond_Yeild_10y_f	Inflation_f	GDP_f	Population_f	House_Hold_Income_f	Employment_f	Cash_Surp_Def_f
1997-01-01	98052	12.342486	11.804569	-0.537918	0	157.959370	4.814434	83.076214	6.055381	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-02-01	98052	12.352806	11.808970	-0.543837	0	158.404761	4.731693	83.392929	6.015761	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-03-01	98052	12.357526	11.811044	-0.546482	0	158.947674	4.633084	82.836534	6.073349	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-04-01	98052	12.367903	11.815443	-0.552460	0	159.387626	4.378458	83.893792	6.079290	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803
1997-05-01	98052	12.378131	11.819842	-0.558289	0	159.765440	4.652420	85.006970	6.126489	2.812443	8577.554463	272639888.8	11.537125	11.767443	-3.733803

Data Collection and Organization Code Overview

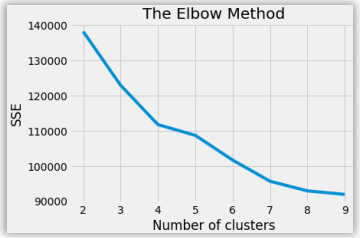
Collect and Organize Data
IST 707 Data Analytics
-Real Estate Property Investment-

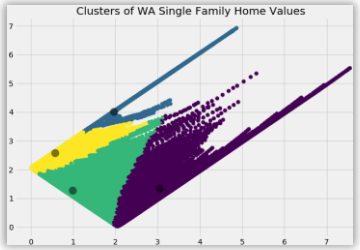
REAL ESTATE DATATSETS - ZILLOW

- Rename 'Region Name' Column to ZipCode
- Convert ZipCode field to string
- Remove columns of non-interest:
  - 'RegionID','SizeRank','City','Metro','CountyName'
  - '1996-04','1996-05','1996-06','1996-07','1996-08','1996-09','1996-10','1996-11','1996-12'
  - '2019-01','2019-02', '2019-03', '2019-04', '2019-05', '2019-06', '2019-07','2019-08','2019-09'
- Fill NaN with median value

Clean the forecast dataset for clustering

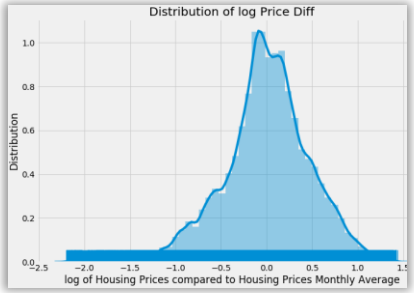
- limite the features for clustering - and the observations to just the predition time (5 years) + one year observed
- remove additive terms and multiplicative terms as well as the datetimestamp
- save series objects for later re joining

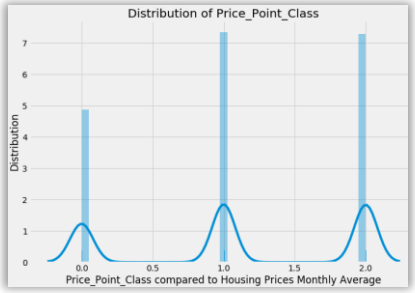




- Transformation of the data's necessary to merge the datasets together after processed through prophet.
- Look over the distribution of key features
- Set price thresholds for supervised learning classification
- Price_Point_Class is a generated feature for supervised classification. Details are shown below

Date	ZipCode	log_Price	log_Price_Monthly_Avg	log_Price_diff	Price_Point_Class
1997-01-01	98052	12.342486	11.804569	-0.537917	0
1997-02-01	98052	12.352806	11.808970	-0.543837	0
1997-03-01	98052	12.357526	11.811044	-0.546482	0
1997-04-01	98052	12.367903	11.815443	-0.552460	0
1997-05-01	98052	12.378131	11.819842	-0.558289	0





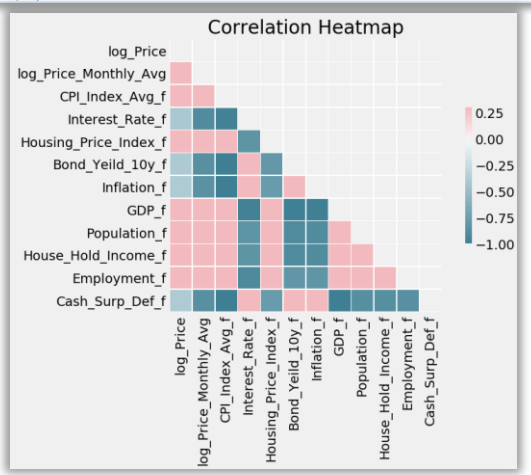
Final Merged Dataset - Real Estate Combined with Economic Data Features

Time range - 1997 - 2017 (that was the cleanest that could be achieved at this time...
*Train classifiers on Feature 'Price_Point_Class'

- 0: means observation's price value is < 25% of the State Price Average
- 1: means observations fall within the normal (average) range of the State Price Average
- 2: means observations falls above the 75% range of the State Price Average

--Determin if classifiers can identify future home value classes based on prior date, location and economic features that have the most impact on both postive and negative price value swings...

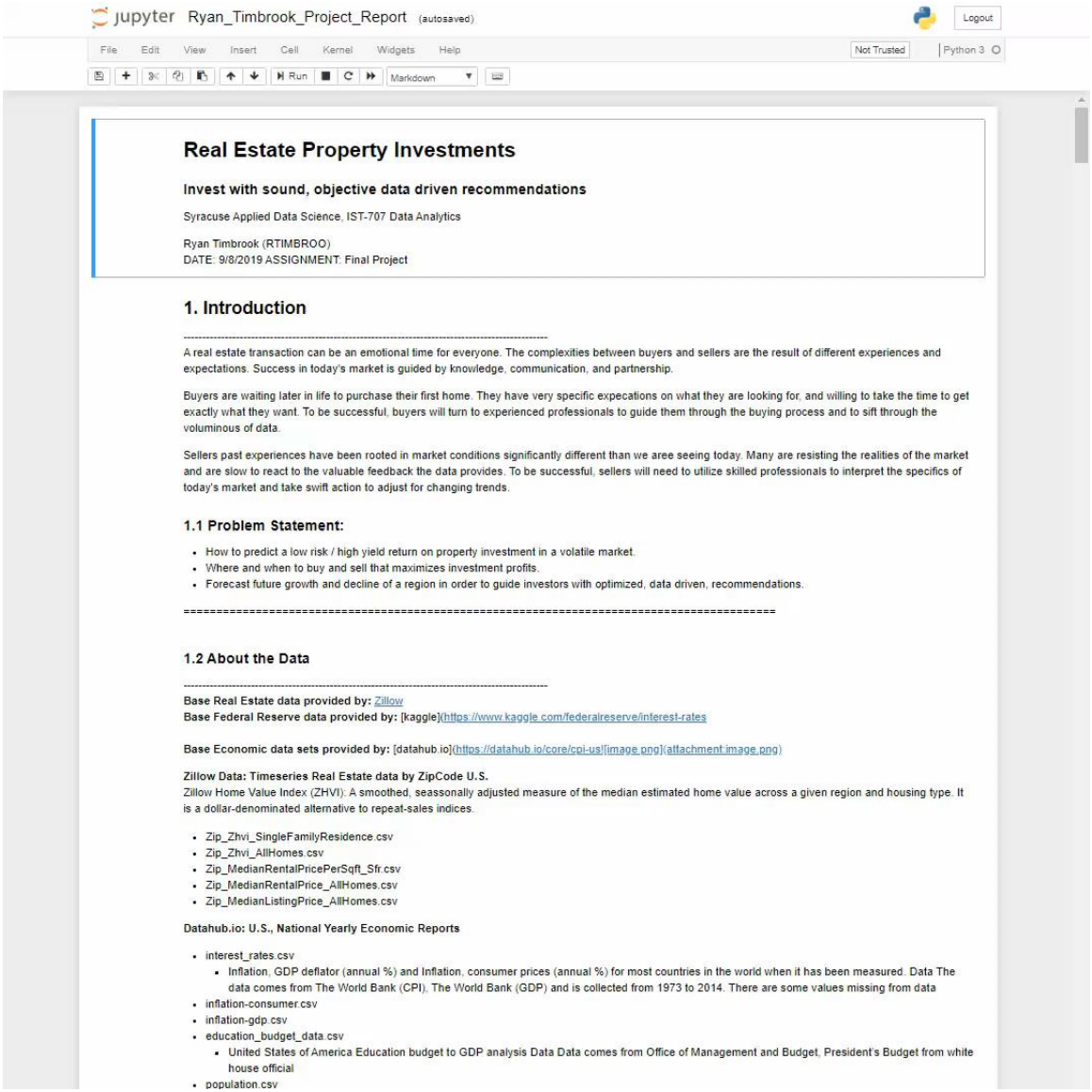
- Dataset Shape: (88452, 16)



[**Data Collection and Organization Code Overview**]

Collect and Organize Data
IST 707 Data Analytics
-Real Estate Property Investment-

[Data Collection and Organization Code Overview]



jupyter Ryan_Timbrook_Project_Report (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Markdown

Real Estate Property Investments

Invest with sound, objective data driven recommendations

Syracuse Applied Data Science, IST-707 Data Analytics

Ryan Timbrook (RTIMBROO)
DATE: 9/8/2019 ASSIGNMENT: Final Project

1. Introduction

A real estate transaction can be an emotional time for everyone. The complexities between buyers and sellers are the result of different experiences and expectations. Success in today's market is guided by knowledge, communication, and partnership.

Buyers are waiting later in life to purchase their first home. They have very specific expecations on what they are looking for, and willing to take the time to get exactly what they want. To be successful, buyers will turn to experienced professionals to guide them through the buying process and to sift through the voluminous of data.

Sellers past experiences have been rooted in market conditions significantly different than we aree seeing today. Many are resisting the realities of the market and are slow to react to the valuable feedback the data provides. To be successful, sellers will need to utilize skilled professionals to interpret the specifics of today's market and take swift action to adjust for changing trends.

1.1 Problem Statement:

- How to predict a low risk / high yield return on property investment in a volatile market.
- Where and when to buy and sell that maximizes investment profits.
- Forecast future growth and decline of a region in order to guide investors with optimized, data driven, recommendations.

==========

1.2 About the Data

Base Real Estate data provided by: Zillow
Base Federal Reserve data provided by: [kaggle](https://www.kaggle.com/federalreserve/interest-rates

Base Economic data sets provided by: [datahub.io](https://datahub.io/core/cpi-us![image.png](attachment:image.png)

Zillow Data: Timeseries Real Estate data by ZipCode U.S.
Zillow Home Value Index (ZHVI): A smoothed, seassonally adjusted measure of the median estimated home value across a given region and housing type. It is a dollar-denominated alternative to repeat-sales indices.

- Zip_Zhvi_SingleFamilyResidence.csv
- Zip_Zhvi_AllHomes.csv
- Zip_MedianRentalPricePerSqft_Sfr.csv
- Zip_MedianRentalPrice_AllHomes.csv
- Zip_MedianListingPrice_AllHomes.csv

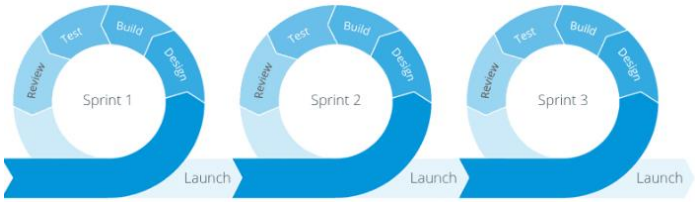
Datahub.io: U.S., National Yearly Economic Reports

- interest_rates.csv
  - Inflation, GDP deflator (annual %) and Inflation, consumer prices (annual %) for most countries in the world when it has been measured. Data The data comes from The World Bank (CPI), The World Bank (GDP) and is collected from 1973 to 2014. There are some values missing from data
- inflation-consumer.csv
- inflation-gdp.csv
- education_budget_data.csv
  - United States of America Education budget to GDP analysis Data Data comes from Office of Management and Budget, President's Budget from white house official
- population.csv

Learning Objective
-Collect and Organize Data-

- Purchase Order Process Improvement

Agile Sprint Delivery



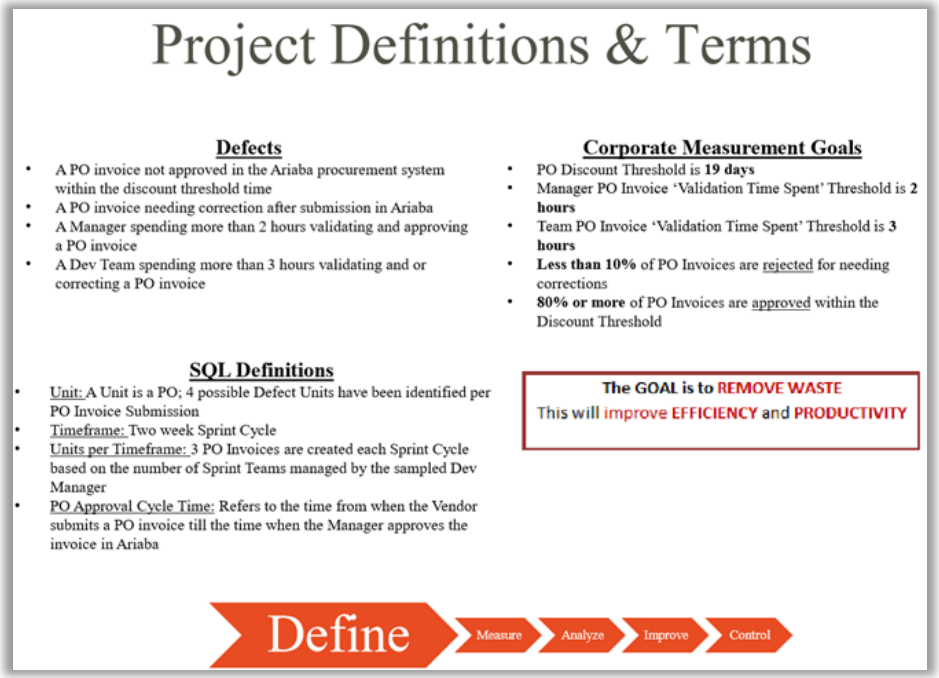
Collect and Organize Data
MBC 638 Data Analysis and Decision Making
-Purchase Order Process Improvement-

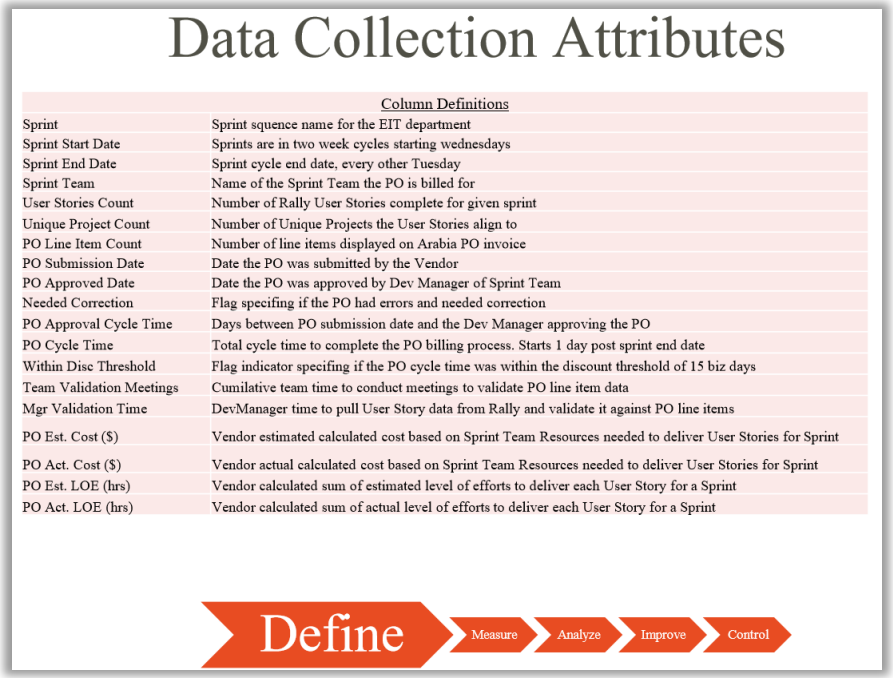
Collection Methods:

- Baseline data sources for this initiative are found in three primary internal business operations systems such as SAP Ariba, CA Rally, and Microsoft SharePoint. Each system was quired to extract the necessary data needed for analysis.

Organization Methods:

- After each of the collection cycles, the data is entered into a master spreadsheet for analyses.
- During the Define phase of the DMAIC lifecycle, Project Definitions & Terms along with Data Collection Attributes are defined.
- As a function of the process improvement workflow, data is entered into SharePoint forms that organizes and standardizes the data entry for consistency and automated workflow mechanisms.





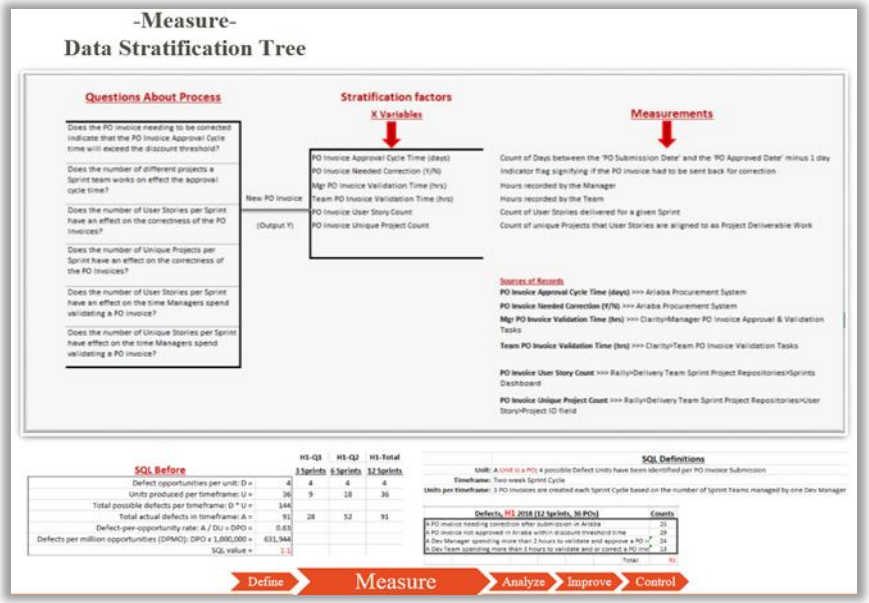
Data Collection and Organization Code Overview

Collect and Organize Data

MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-

-Measure-
Data Stratification Tree



SQL Before		H1-Q1 3 Sprints	H1-Q2 6 Sprints	H1-Total 12 Sprints
Defect opportunities per unit: D =	4	4	4	4
Units produced per timeframe: U =	36	9	18	36
Total possible defects per timeframe: D * U =	144			
Total actual defects in timeframe: A =	91	28	52	91
Defect-per-opportunity rate: A / DU = DPO =	0.63			
Defects per million opportunities (DPMO): DPO x 1,000,000 =	631,944			
SQL value =	3.1			

SQL Definitions

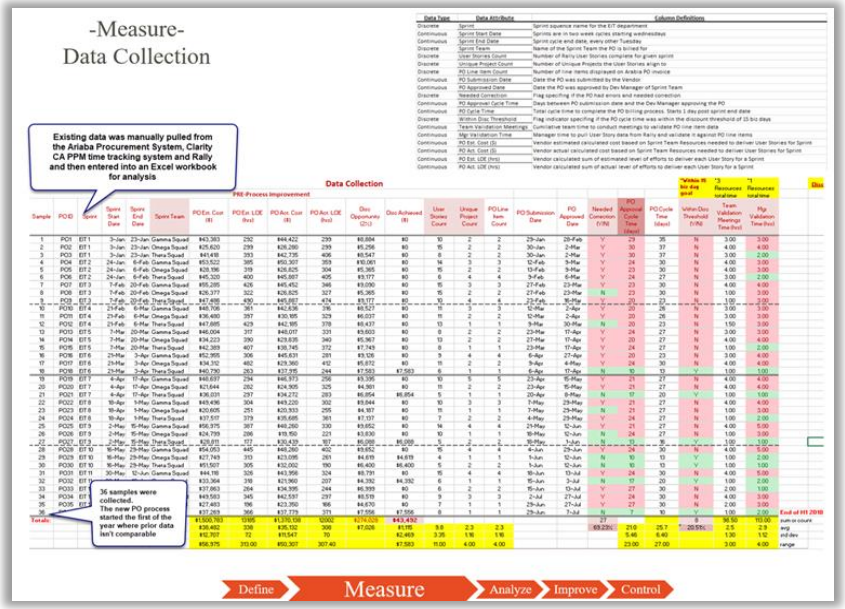
Unit: A Unit is a PO; 4 possible Defect Units have been identified per PO Invoice Submission
Timeframe: Two week Sprint Cycle
Units per timeframe: 3 PO Invoices are created each Sprint Cycle based on the number of Sprint Teams managed by one Dev Manager

Defects, H1 2018 (12 Sprints, 36 POs)	Counts
A PO invoice needing correction after submission in Ariaba	25
A PO invoice not approved in Ariaba within discount threshold time	29
A Dev Manager spending more than 2 hours to validate and approve a PO invoice	24
A Dev Team spending more than 3 hours to validate and or correct a PO invoice	13
Total:	91

Define | **Measure** | Analyze | Improve | Control

-Measure-
Data Collection

Existing data was manually pulled from the Ariaba Procurement System, Clarity CA PPM time tracking system and Rally and then entered into an Excel workbook for analysis

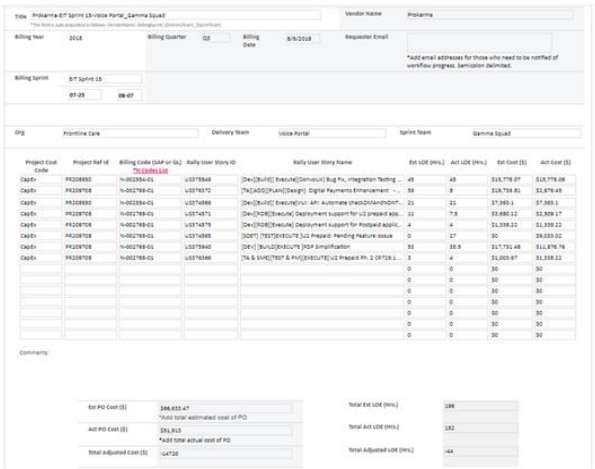


Define | **Measure** | Analyze | Improve | Control

-Improve-
New SharePoint Workflow and Data Input Form

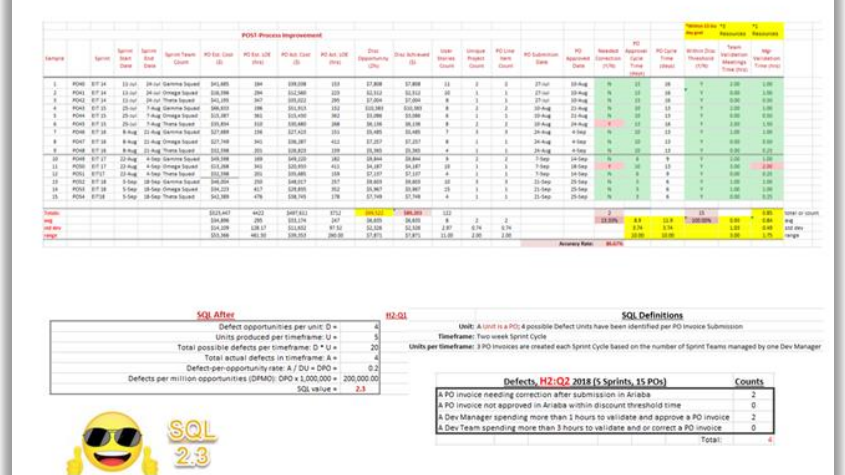
This form represents all of the attributes and data elements which need to be collected for a timely validation of Purchase Orders submitted by the Vendor.

Prior to this new process, all of this data had to be manually tracked down by the Dev Manager before approving a PO in the Ariba Procurement System.



Define | Measure | Analyze | **Improve** | Control

-Improve-
Data Collection



SQL After	H2-Q1
Defect opportunities per unit: D =	4
Units produced per timeframe: U =	5
Total possible defects per timeframe: D * U =	20
Total actual defects in timeframe: A =	4
Defect-per-opportunity rate: A / DU = DPO =	0.2
Defects per million opportunities (DPMO): DPO x 1,000,000 =	200,000.00
SQL value =	2.3

SQL Definitions

Unit: A Unit is a PO; 4 possible Defect Units have been identified per PO Invoice Submission
Timeframe: Two week Sprint Cycle
Units per timeframe: 3 PO Invoices are created each Sprint Cycle based on the number of Sprint Teams managed by one Dev Manager

Defects, H2:Q2 2018 (5 Sprints, 15 POs)	Counts
A PO invoice needing correction after submission in Ariaba	2
A PO invoice not approved in Ariaba within discount threshold time	0
A Dev Manager spending more than 2 hours to validate and approve a PO invoice	2
A Dev Team spending more than 3 hours to validate and or correct a PO invoice	0
Total:	4

Define | Measure | Analyze | **Improve** | Control

Collect and Organize Data
MBC 638 Data Analysis and Decision Making
-Purchase Order Process Improvement-

[Data Collection and Organization Code Overview]

file:///G:/My%20Drive/Timbrook_Portfolio_Milestone/mbc638_data_analysis_decision/final_project/05-Report/RTIMBROO-MBC638-ProssesImprovementProject.pdf

1 of 28 | Fit to page | Page view | Read aloud | Add notes

Sprint Team, Purchase Order Invoice
Approval Cycle Time Reduction

Process owner: Ryan Timbrook

Key Dates --->	Team Launch 7/18	Define 7/21	Measure 7/25	Analyze 8/1	Improve 8/29	Control 9/12

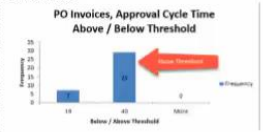
DEFINE

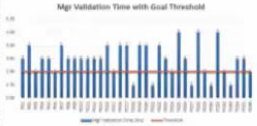
- Only 20% of Vendor supplied Sprint Teams Purchase Order (PO) Invoices are being approved within the discount threshold of 19 days.
- Not achieving this discount benchmark costs the company, on average, **$5,911 in lost revenue** on each PO invoice it processes
- Managers and Teams spend to much time validating Sprint PO invoices for correctness. This time not only costs in resource dollars, it as well has negative impacts to the amount of software delivery work they can complete each Sprint Cycle, leading to a **decreased 'Time to Market' performance** measure

SQL 1.1

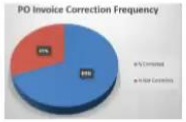
MEASURE

Managers have a low frequency rate of approving POs under threshold



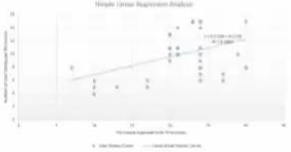


PO Invoices having to be corrected leads to longer Approval Cycle Times



ANALYZE

PO Invoices with mid-to-high User Story Counts are increasing the likelihood a PO Invoice will need to be corrected, which leads to a longer Approval Cycle Time



The more time a Team and Mgr have to spend on Validating a PO Invoice the more likely the PO Approval Cycle Time will increase

Solution -> A New SharePoint workflow form and process

IMPROVE

New SharePoint Workflow = Efficiency

Avg Discount Achieved
$6,635 / Sprint
$172,510 Annual / Sprint Team

SQL 2.3

PO Invoice Accuracy Increase of **57%**

CONTROL

Control charts show a significant improvement in Approval Cycle Time starting at PO37, which was when the new SharePoint Workflow process was implemented

TEAM MEMBERS

Ryan Timbrook

-Identify patterns in data via visualization, statistical analysis, and data mining-

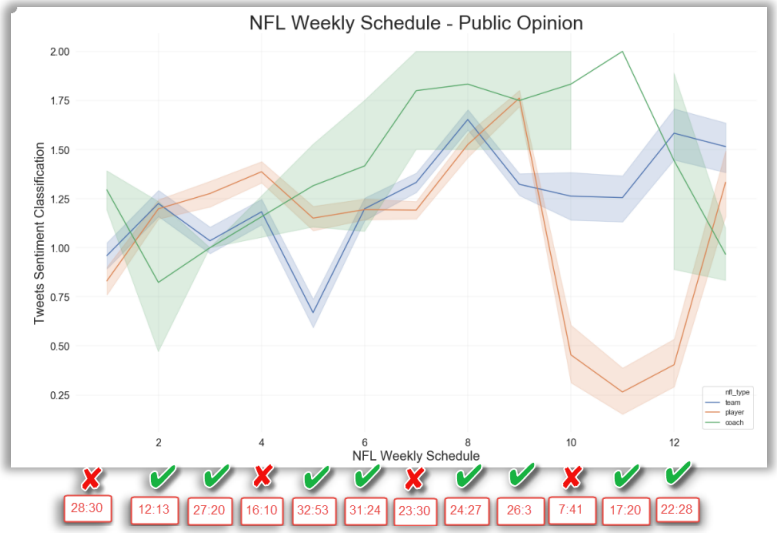
Overview

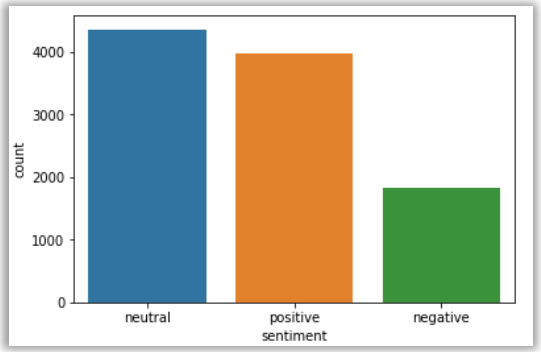
- Public Sentiment Toward NFL Teams, Coaches, and Players
- IEEE-CIS Fraud Detection
- Real Estate Property Investment

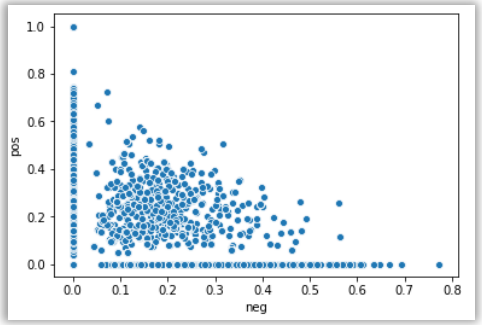
Learning Objective

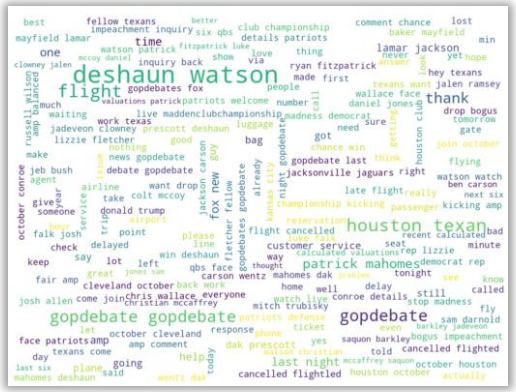
-Identify patterns in data via visualization, statistical analysis, and data mining-

➢ Public Sentiment Toward NFL Teams, Coaches, and Players









Identify patterns in data via visualization, statistical analysis, and data mining

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

Figure 2.1: Vedar Sentiment Scoring - NFL Trend over 13 Game Weeks

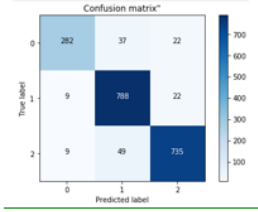
Most POSITIVE WORDS

Top and Bottome 10 of Most	POSITIVE	Learned Words
-1.3223	shit	1.9337 best
-1.2525	bad	1.9671 cheering
-1.1760	anymore	1.9906 winners
-1.1112	absolutely	2.0000 love
-1.1050	usually	2.0000 easy
-1.0301	kill	2.0000 easily
-1.0247	hate	2.0239 win
-1.0170	threat	2.0893 great
-1.0000	advertising	2.0930 better
-1.0000	crushed	2.1714 bold

Most NEGATIVE WORDS

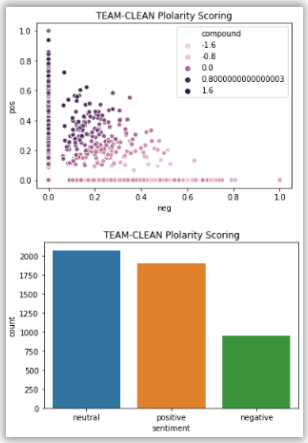
Top and Bottome 10 of Most	NEGATIVE	Learned Words
-1.6849	care	1.5858 axed
-1.4871	best	1.6238 offensive
-1.3031	wealth	1.6280 wrong
-1.2034	stay	1.6342 ill
-1.0856	improve	1.6606 worst
-1.0672	hope	1.6794 bad
-1.0495	fun	1.7560 problem
-1.0124	play	1.7898 ass
-1.0000	comparing	1.8333 hurt
-0.9472	receiver	1.9542 injury

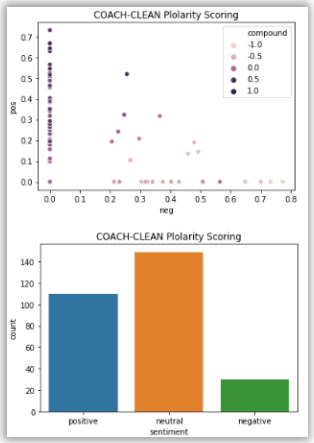
SVM Unigram Model Prediction Accuracy Results: 93%

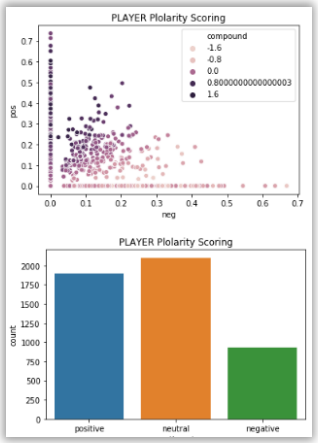


Classification Report:

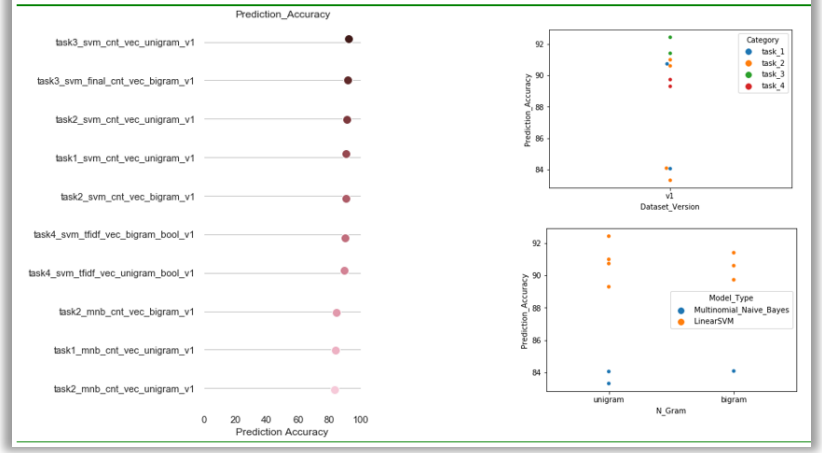
	precision	recall	f1-score	support
negative	0.94	0.83	0.88	341
neutral	0.90	0.96	0.93	819
positive	0.94	0.93	0.94	793
micro avg	0.92	0.92	0.92	1953
macro avg	0.93	0.91	0.92	1953
weighted avg	0.93	0.92	0.92	1953





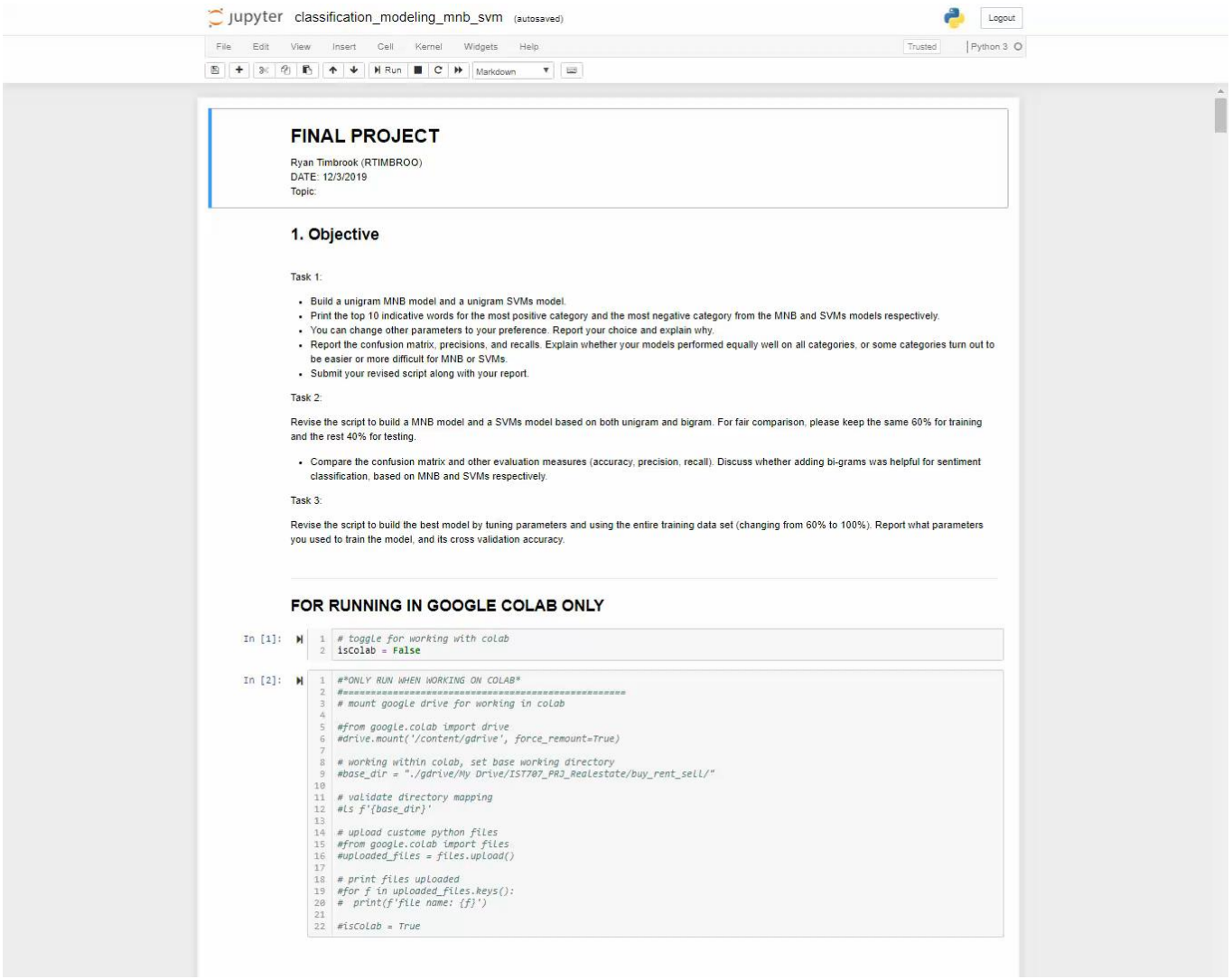


3.1.1.5.1 Model Accuracy Comparison Summary



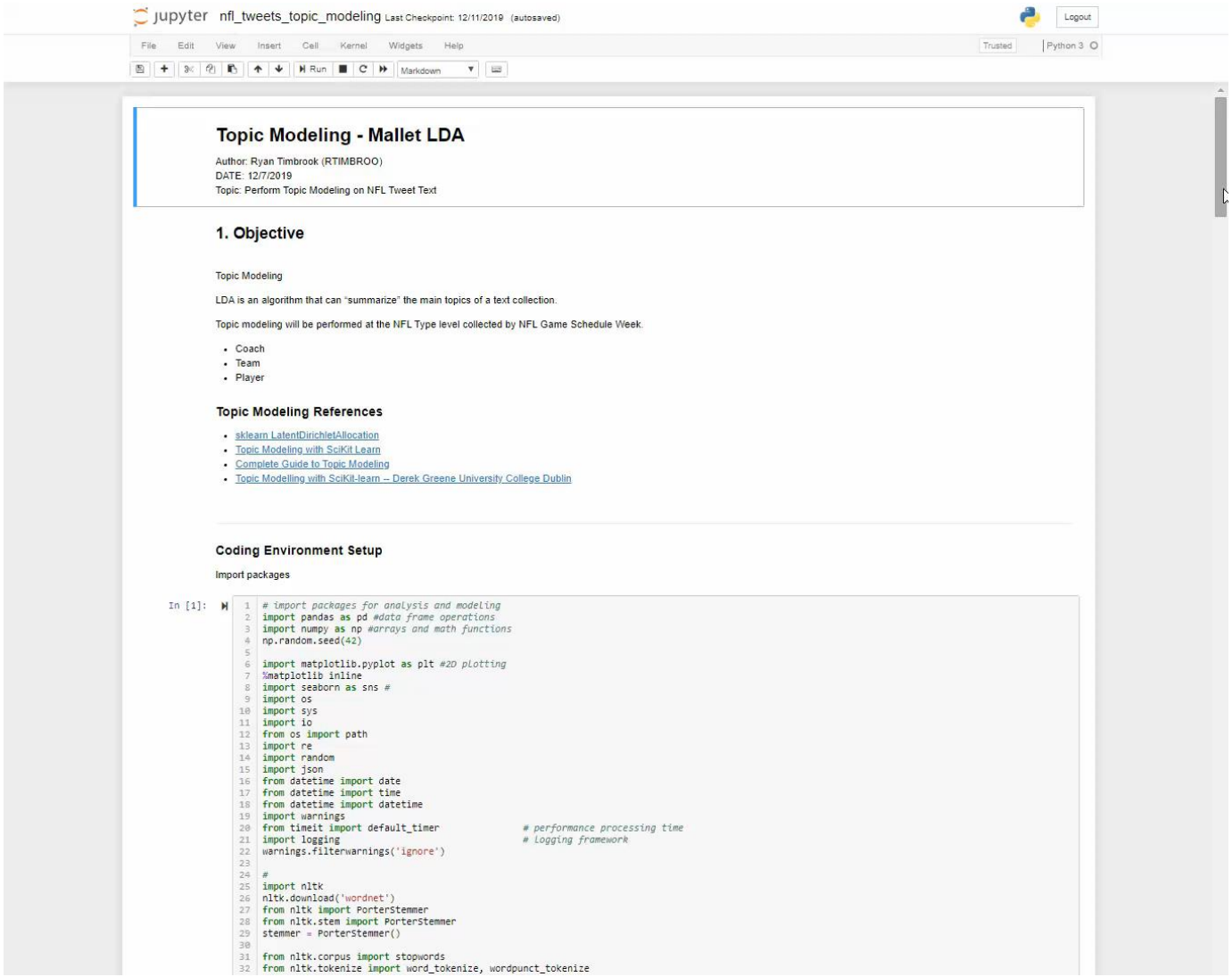
Identify patterns in data via visualization, statistical analysis, and data mining
IST 736 Text Mining
-Public Sentiment Toward NFL Teams, Coaches, and Players-

[Patterns in Classification Modeling]



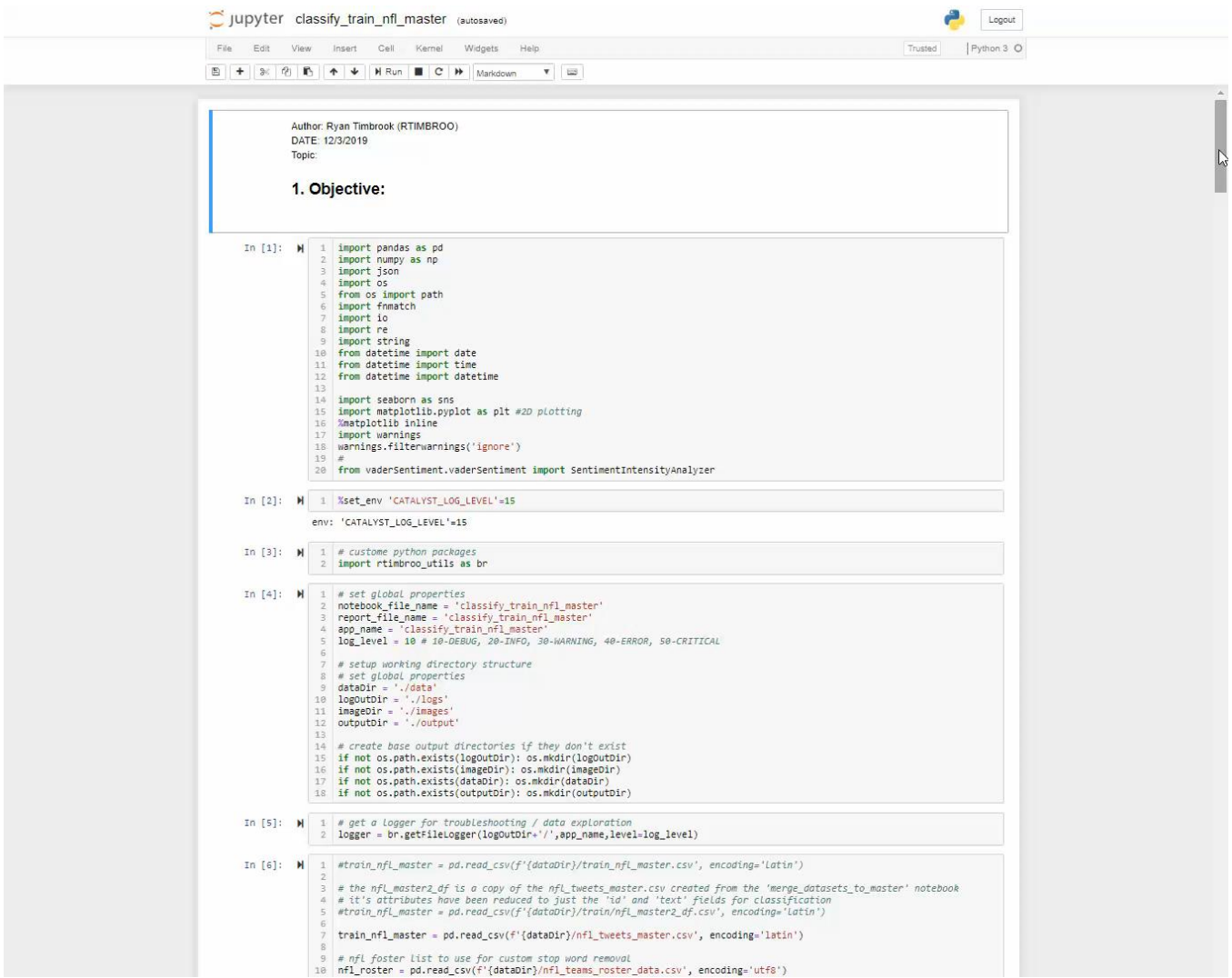
Identify patterns in data via visualization, statistical analysis, and data mining
IST 736 Text Mining
-Public Sentiment Toward NFL Teams, Coaches, and Players-

[Patterns in Topic Modeling]



Identify patterns in data via visualization, statistical analysis, and data mining
IST 736 Text Mining
-Public Sentiment Toward NFL Teams, Coaches, and Players-

[Patterns in Polarity Scoring]



jupyter classify_train_nfl_master (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Markdown

Author: Ryan Timbrook (RTIMBROO)
DATE: 12/3/2019
Topic:

1. Objective:

In [1]:
```
import pandas as pd
import numpy as np
import json
import os
from os import path
import fnmatch
import io
import re
import string
from datetime import date
from datetime import time
from datetime import datetime

import seaborn as sns
import matplotlib.pyplot as plt #2D plotting
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
#
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

In [2]:
```
%set_env 'CATALYST_LOG_LEVEL'=15
```
```
env: 'CATALYST_LOG_LEVEL'=15
```

In [3]:
```
# custome python packages
import rtimbroo_utils as br
```

In [4]:
```
# set global properties
notebook_file_name = 'classify_train_nfl_master'
report_file_name = 'classify_train_nfl_master'
app_name = 'classify_train_nfl_master'
log_level = 10 # 10-DEBUG, 20-INFO, 30-WARNING, 40-ERROR, 50-CRITICAL

# setup working directory structure
# set global properties
dataDir = './data'
logOutDir = './logs'
imageDir = './images'
outputDir = './output'

# create base output directories if they don't exist
if not os.path.exists(logOutDir): os.mkdir(logOutDir)
if not os.path.exists(imageDir): os.mkdir(imageDir)
if not os.path.exists(dataDir): os.mkdir(dataDir)
if not os.path.exists(outputDir): os.mkdir(outputDir)
```

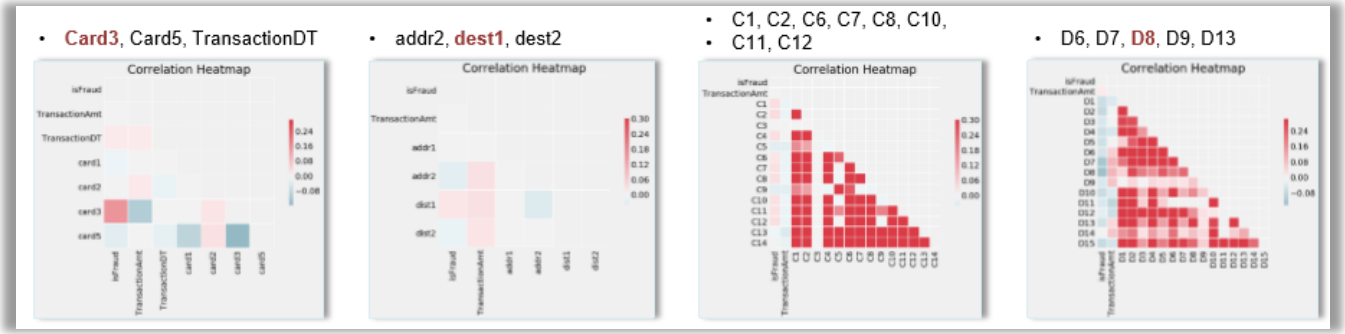
In [5]:
```
# get a logger for troubleshooting / data exploration
logger = br.getFileLogger(logOutDir+'/',app_name,level=log_level)
```

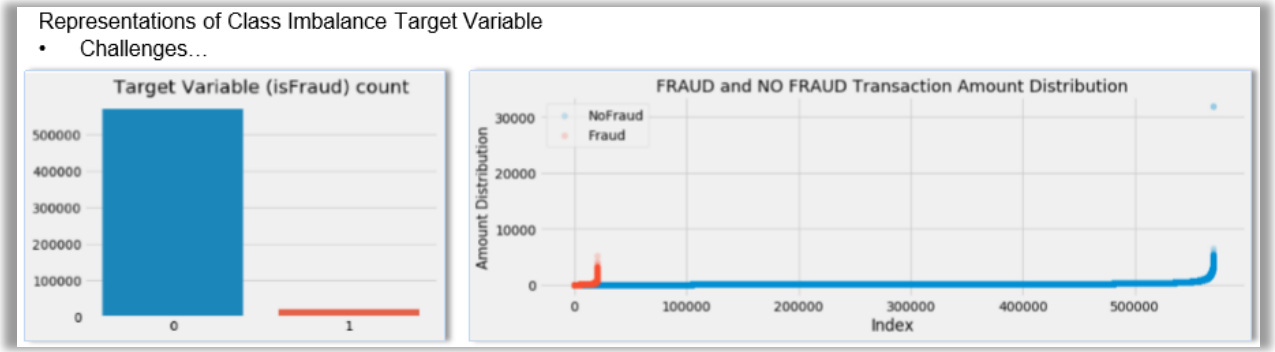
In [6]:
```
#train_nfl_master = pd.read_csv(f'{dataDir}/train_nfl_master.csv', encoding='latin')

# the nfl_master2_df is a copy of the nfl_tweets_master.csv created from the 'merge_datasets_to_master' notebook
# it's attributes have been reduced to just the 'id' and 'text' fields for classification
#train_nfl_master = pd.read_csv(f'{dataDir}/train/nfl_master2_df.csv', encoding='latin')

train_nfl_master = pd.read_csv(f'{dataDir}/nfl_tweets_master.csv', encoding='latin')

# nfl foster list to use for custom stop word removal
nfl_roster = pd.read_csv(f'{dataDir}/nfl_teams_roster_data.csv', encoding='utf8')
```

Learning Objective

-Identify patterns in data via visualization, statistical analysis, and data mining-

➢IEEE-CIS Fraud Detection





Identify patterns in data via visualization, statistical analysis, and data mining
IST 718 Big Data Analytics
-IEEE-CIS Fraud Detection-

Base Neural Network

Base Model Hyperparameters:

- Kernel_initializer: normal
- Dense Input Layer:
- Activation Function: **'relu'**
- Dense Output Layer:
- Activation Function: **'softmax'**
- Compiler:
- loss: **'categorical_crossentropy**
- optimizer: **'adam'**
- Built on:
- epochs: **300**
- batch_size: **1000**

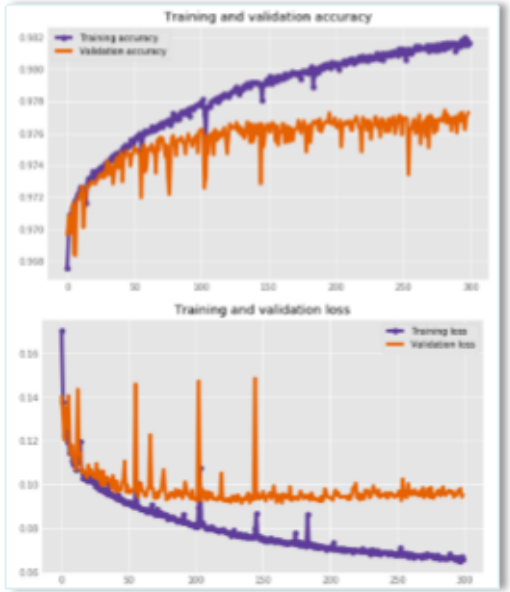
Model Summary:

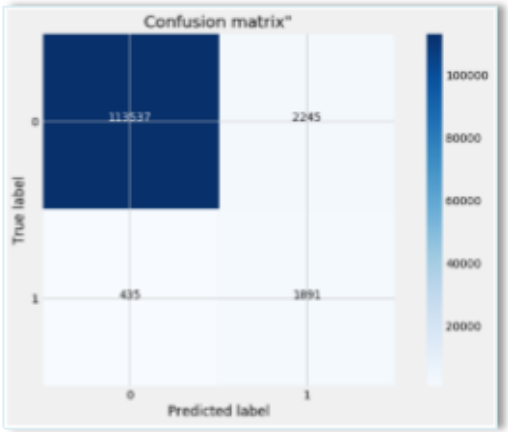
```
None
254520
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 503)               253512
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 1008
=================================================================
Total params: 254,520
Trainable params: 254,520
Non-trainable params: 0
_________________________________________________________________
```

```
Test loss: 0.0959736775085113
Test accuracy: 0.9773089037192809
Baseline Error: 2.2691096280719023
```

	precision	recall	f1-score	support
Class0	0.98	1.00	0.99	113972
Class1	0.81	0.46	0.59	4136
accuracy			0.98	118108
macro avg	0.90	0.73	0.79	118108
weighted avg	0.97	0.98	0.97	118108

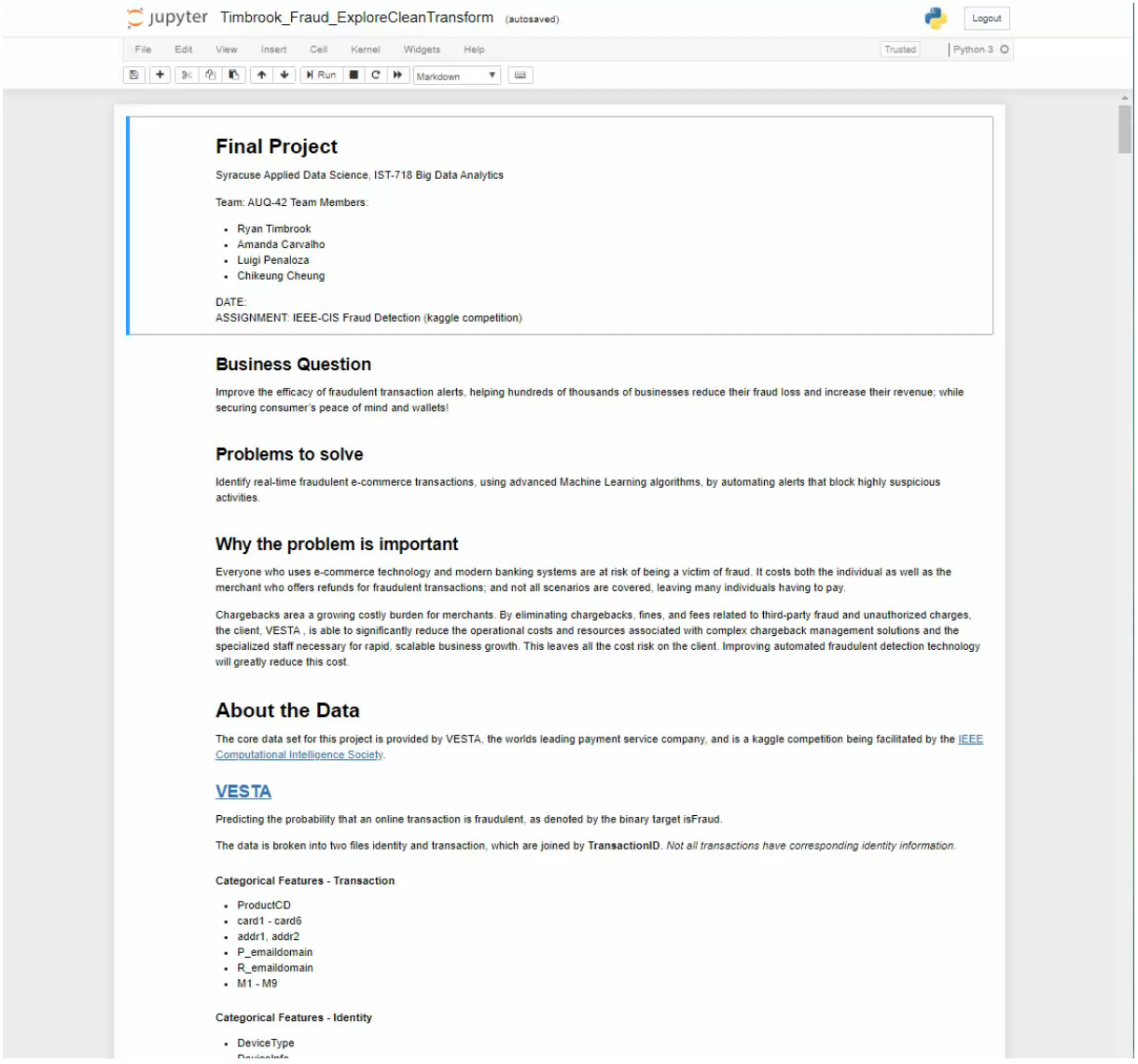
```
X_train shape: (330702, 205)
y_train shape: (330702, 2)
X_test shape: (118108, 205)
y_test shape: (118108, 2)
X_val shape: (141730, 205)
y_val shape: (141730, 2)
```





Identify patterns in data via visualization, statistical analysis, and data mining
IST 718 Big Data Analytics
-IEEE-CIS Fraud Detection-

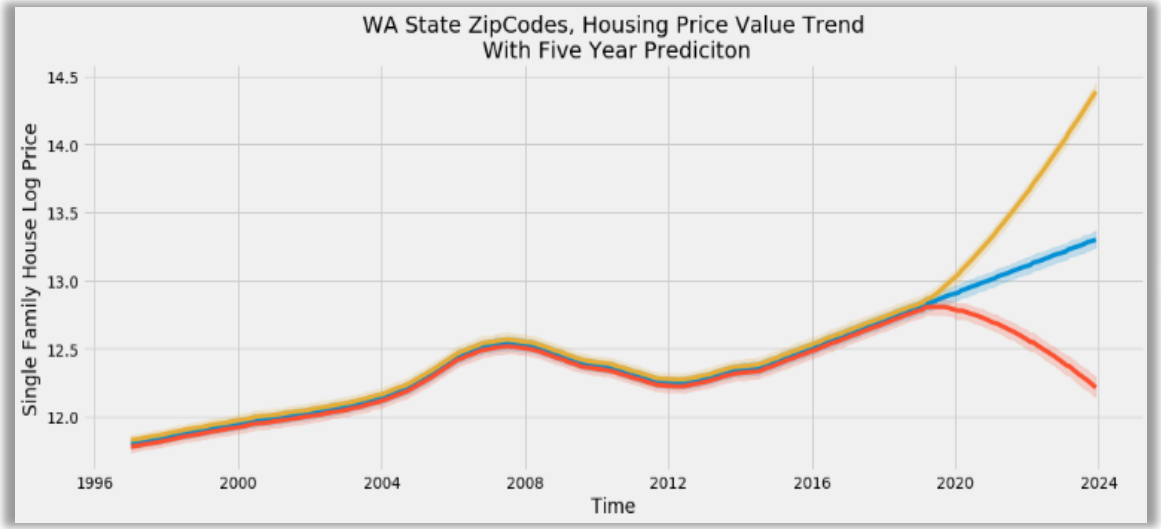
[Patterns in Exploratory Data Analysis and Neural Networks]

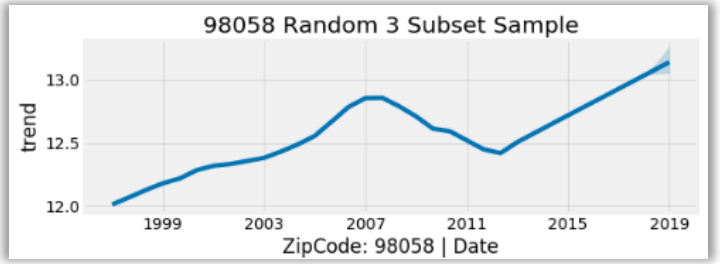


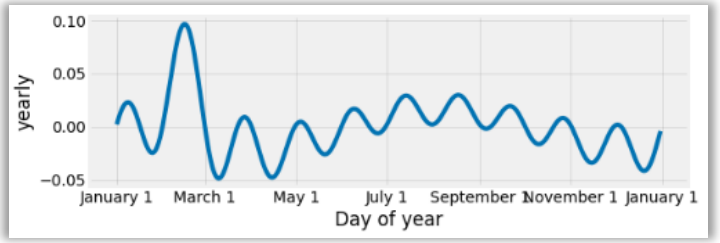
Learning Objective

-Identify patterns in data via visualization, statistical analysis, and data mining-

➢Real Estate Property Investment



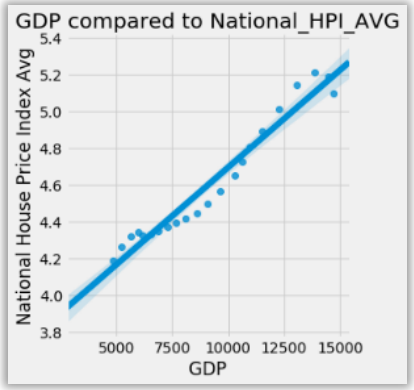


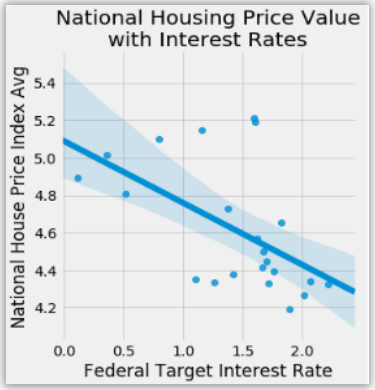


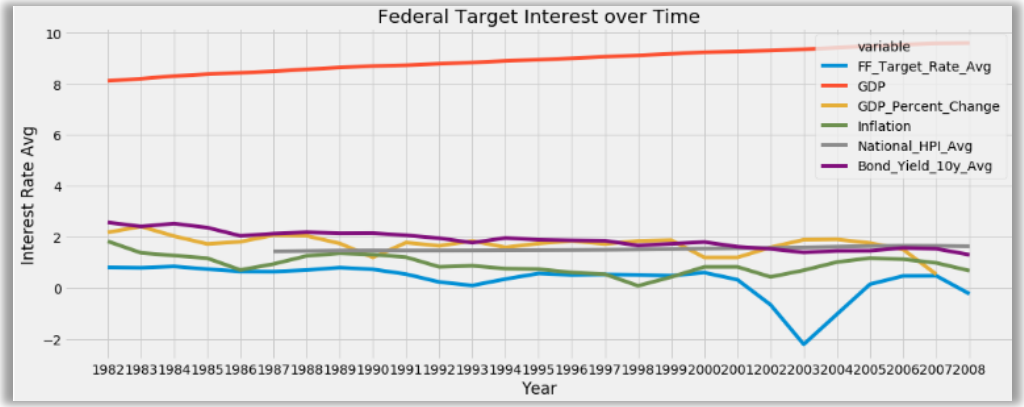
Identify patterns in data via visualization, statistical analysis, and data mining

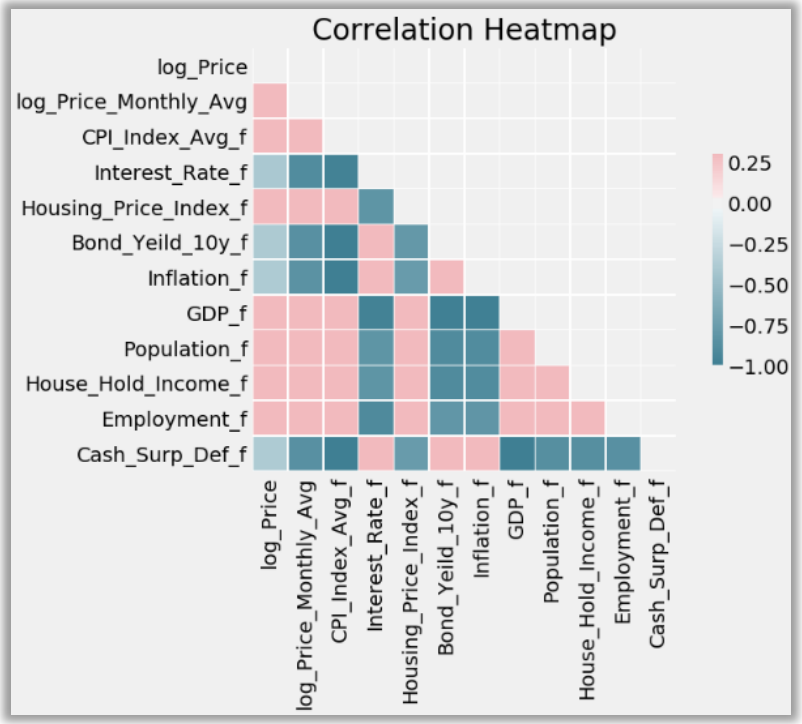
IST 707 Data Analytics

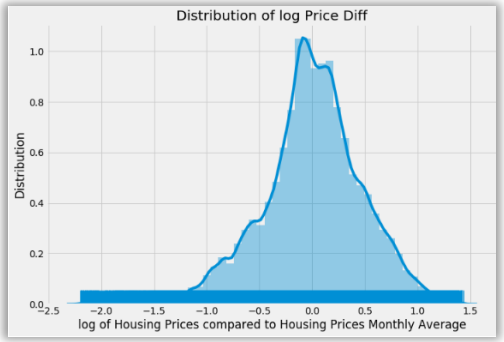
-Real Estate Property Investment-

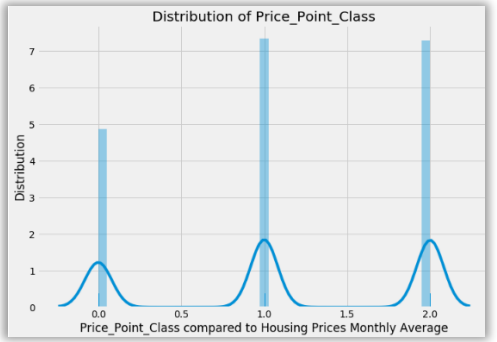


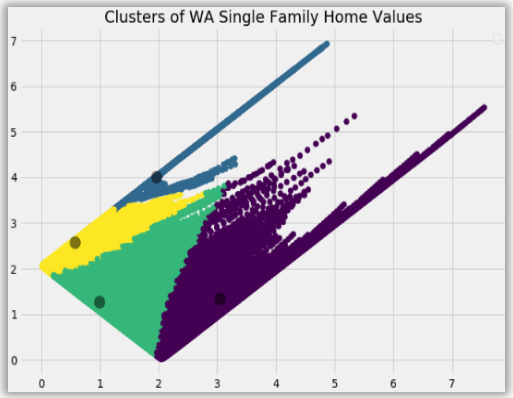












Identify patterns in data via visualization, statistical analysis, and data mining
IST 707 Data Analytics
-Real Estate Property Investment-

[Patterns in Exploratory Data Analysis]

jupyter Ryan_Timbrook_Project_Report (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

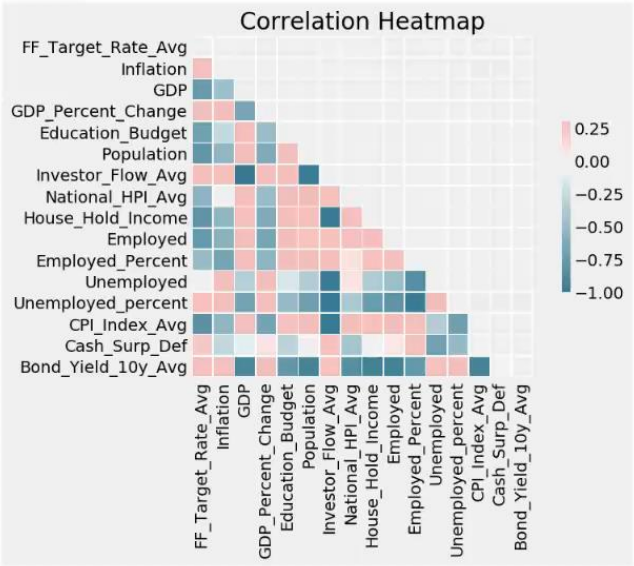
Run Markdown

```
    cpi_y,
    csd,
    by_10y_m,
    by_10y_y,
    datasets_to_merge_year,
    datasets_to_merge_month
except:
    pass
```

Correlation Heatmap of the new Economic Dataset's features

In [90]:

```
#plt.figure(figsize=(12,8))
rt.plot_corr_heatmap(economicDf_year,economicDf_year.drop(columns=['Year']).columns)
#plt.show()
```



In [91]:

```
p = economicDf_year
p['National_HPI_Avg'] = np.log(p['National_HPI_Avg'])
plt.figure(figsize=(16,6))
ax = sns.lineplot(x='Year',y='National_HPI_Avg',data=p)
plt.title('National House Price Index over Time')
plt.ylabel('Log National House Price Index')

plt.show()
del p
```

-Develop alternative strategies based on the data- Overview

- Real Estate Property Investment
- Recruiting Advertising Strategy

Learning Objective

- Develop alternative strategies based on the data-

- Real Estate Property Investment

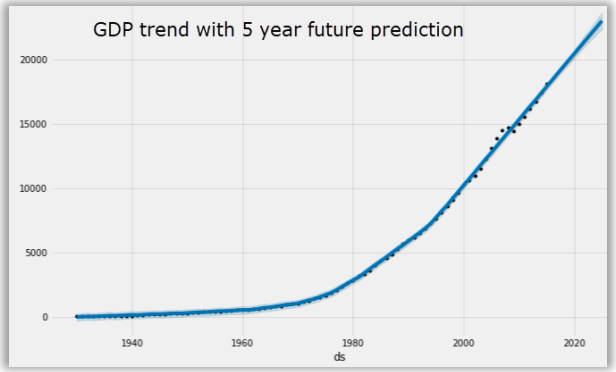
Develop alternative strategies based on the data
IST 707 Data Analytics
-Real Estate Property Investment-

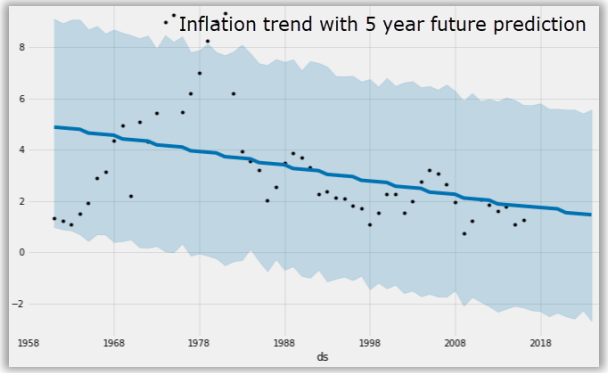
- Original **Economic Reports** dataset **incomplete**, unusable – **Alternative** found at **Datahub**.io

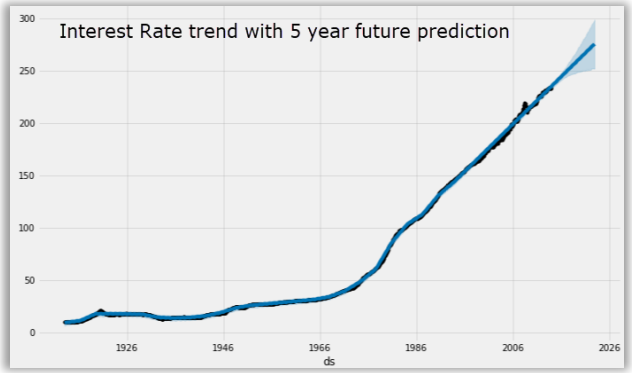
- Time series date ranges **short of target year** – Future **predictive** values created (Facebook Prophet)

    - **GDP** Year: Forecasted from 2016, 2017, 2018
    - **Inflation**: Forecasted for 2017, 2018
    - **Interest Rates**: Forecasted for 2016, 2017, 2018

PROPHET







Learning Objective

- Develop alternative strategies based on the data-

➢ Recruiting Advertising Strategy

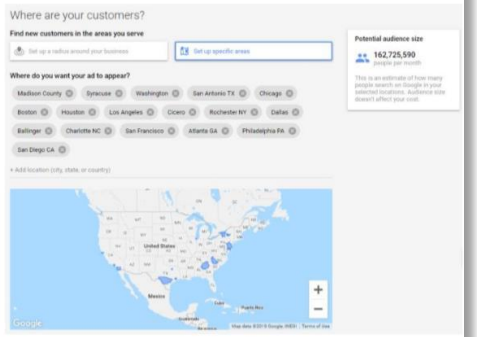
3.1 Recommendations

Advertising Regions:

- Taking the top ten US states by region for each prior campaign, it's recommended to target these regions for 2020.

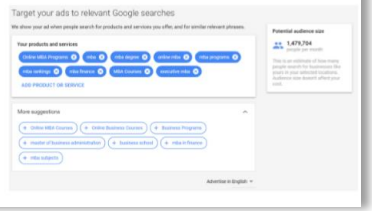
Google Ads estimates these US regions have a reach of ***162, 725, 590 people per month****.*

- Syracuse, NY
- New York, New York
- Washington, District of Columbia
- San Antonio, TX
- Chicago, IL
- Boston, MA
- Huston, TX
- Los Angeles, CA
- Cicero, NY
- Rochester, NY
- Ballinger, TX
- Charlotte, NC
- San Francisco, CA
- Atlanta, GA
- Philadelphia, PA
- San Diego, CA



Keywords to use:

- Online MBA Programs
- Mba
- Mba degree
- Online mba
- Mba programs
- Mba rankings
- Mba finance
- Mba Courses
- Execute mba



4.1.1 *Proposed 2020 Budget Strategy*

Campaign	2020 Total Duration	Ads Runtime Iterations	Provider	Budget
Whitman.syr.ed	9 months	1 month trial, then 4x2mo. iterations	Google Ads	$ 14,778.00
		1 month trial, then 4x2mo. iterations	Facebook	$ 11,666.67
		1 month trial, then 4x2mo. iterations	GMASS	$ 6,666.67
MBA Marketing - Full Time	9 months	1 month trial, then 4x2mo. iterations	Google Ads	$ 14,778.00
		1 month trial, then 4x2mo. iterations	Facebook	$ 11,666.67
		1 month trial, then 4x2mo. iterations	GMASS	$ 6,666.67
MBA Marketing - iMBA	9 months	1 month trial, then 4x2mo. iterations	Google Ads	$ 14,778.00
		1 month trial, then 4x2mo. iterations	Facebook	$ 11,666.67
		1 month trial, then 4x2mo. iterations	GMASS	$ 6,666.67
				$ 99,334.00

Develop alternative strategies based on the data
IST 707 Data Analytics
-Recruiting Advertising Strategy-

Based on the historical data and insights gained from analyzing the google analytics platform, using the following metrics: **Cost per Click**, **Users**, **New Users**, **Bounce Rate**, **Pages per session**, the **most effective** campaign identified was the **Whitman.syr.edu**.

United States Campaign Measurements

Campaign	Start Date	End Date	Cost	Cost per Click	Users	New Users	Total Sessions	Bounce Rate	Page per Session
whitman.syr.edu	2/15/2011	8/15/2011	$ 37,851.36	$ 3.93	133,732	98,379	192,134	48.08%	3.76
MBA Marketing – Full-time	Jan/15/2012	2/15/2012	$ 16,459.90	$ 4.64	1,347	1,348	1,468	97.07%	1.05
MBA Marketing – iMBA	2/1/2012	10/15/2012	$ 144,971.70	$ 7.46	2,753	2,748	3,079	88.96%	1.14
Delta	9/1/2013	11/1/2013	$ 10,000.00	$ -			284	28.17%	3.37

2020 Advertising Recruitment Recommendations:

- The 2020 Campaign recommendations include continued use of **Google Ad Marketing**, with the additional recommendations to incorporate **Facebook Ad Business** and **GMASS Targeting marketing**.

- The recommendations included in the report detail the **advertising regions** to focus their **ad campaigns** to, **keywords** to use in **search analytics**, the **best days of the week and time of the day** for the **advertising** to be **published**, the **expected advertising costs** which includes a **budget plan**, a post-implementation success measurement plan, and a **list** of **other factors** we found essential to the **success** of the upcoming **campaign.**

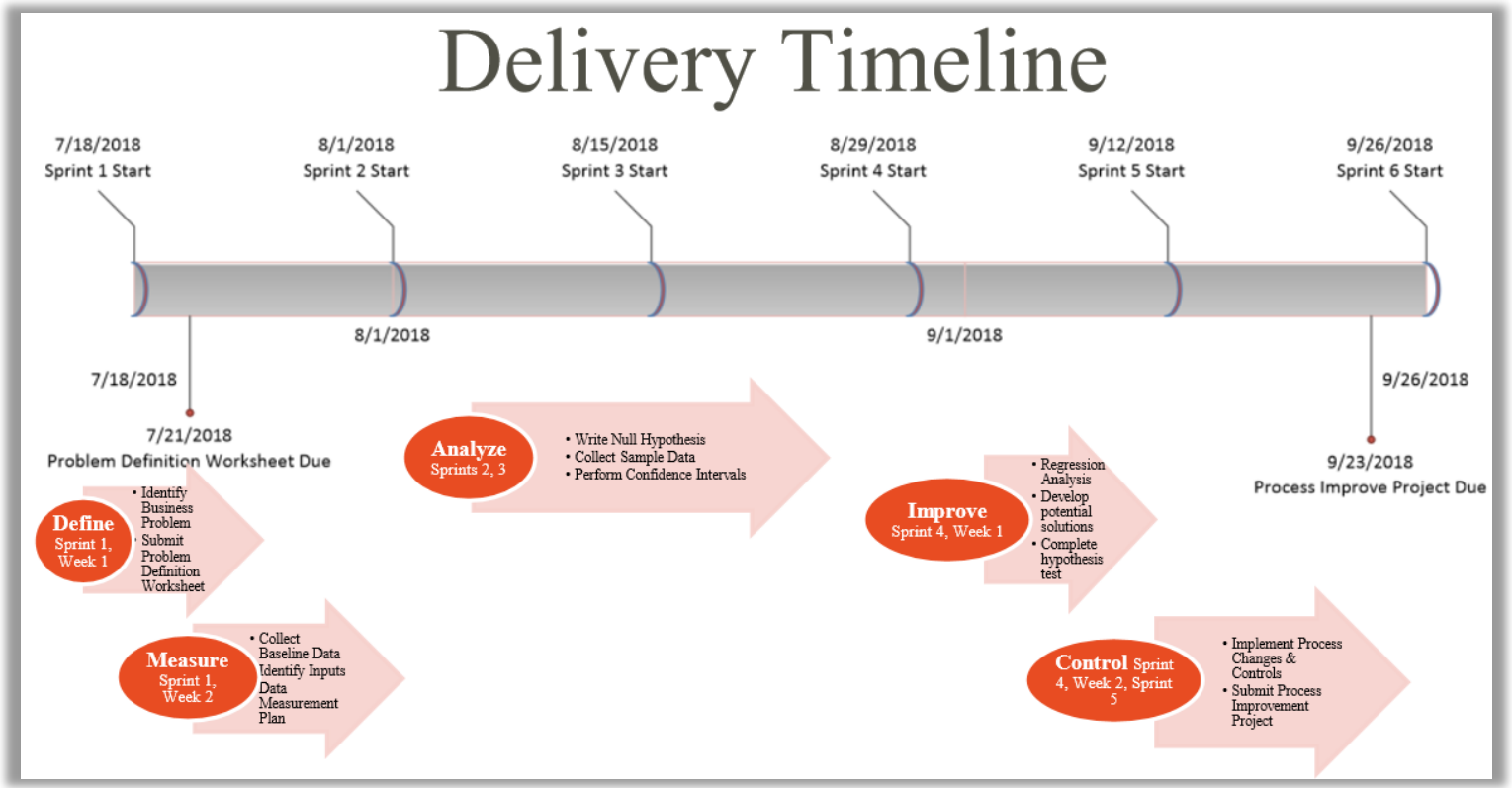
-Develop a plan of action to implement the business decisions derived from analysis-
Overview

- Purchase Order Process Improvement
- A4B KPI BizOps Organization Database

Learning Objective

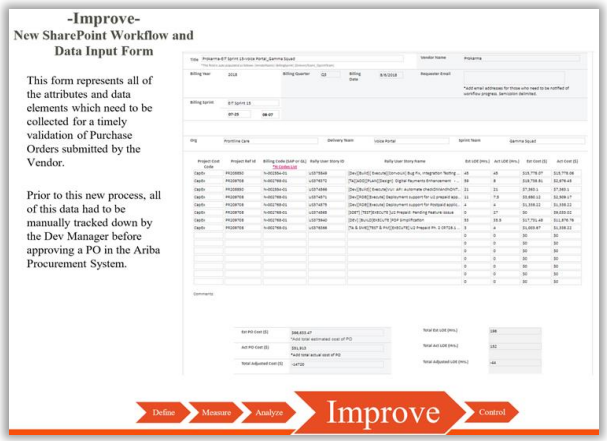
- Develop a plan of action to implement the business decisions derived from analysis-

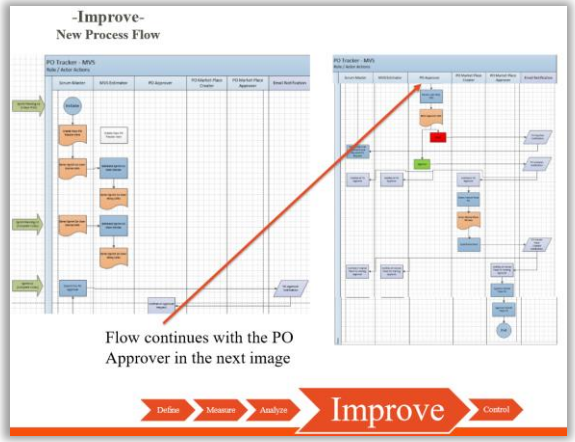
➢ Purchase Order Process Improvement

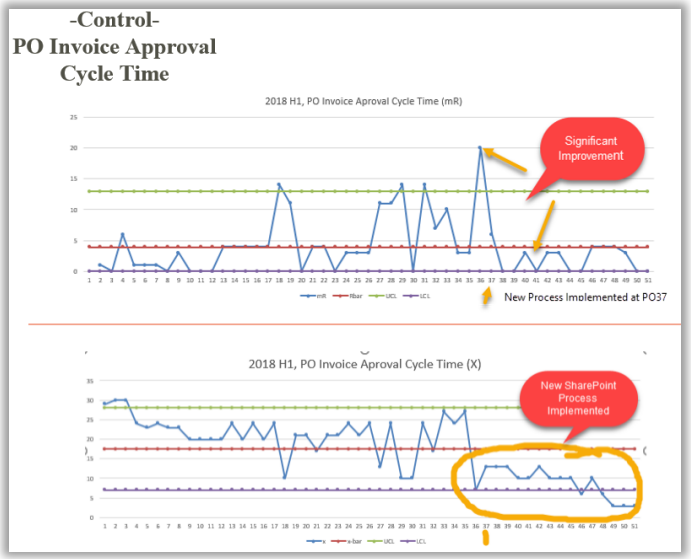


Develop a plan of action
MBC 638 Data Analysis and Decision Making
-Purchase Order Process Improvement-

In the **'Improve'** phase, a **plan** was **executed** that included the **design** and **pilot** of a **new process workflow** that incorporated automation through Microsoft SharePoint tooling. By designing a standard user interface for entering the necessary attributes needed for an accurate and timely PO validation and approval process along with a detailed cross-functional workflow diagram for role/resource training, an **improved cycle time** efficiency of **70%** was observed **during** the **pilot** of the **new process**.







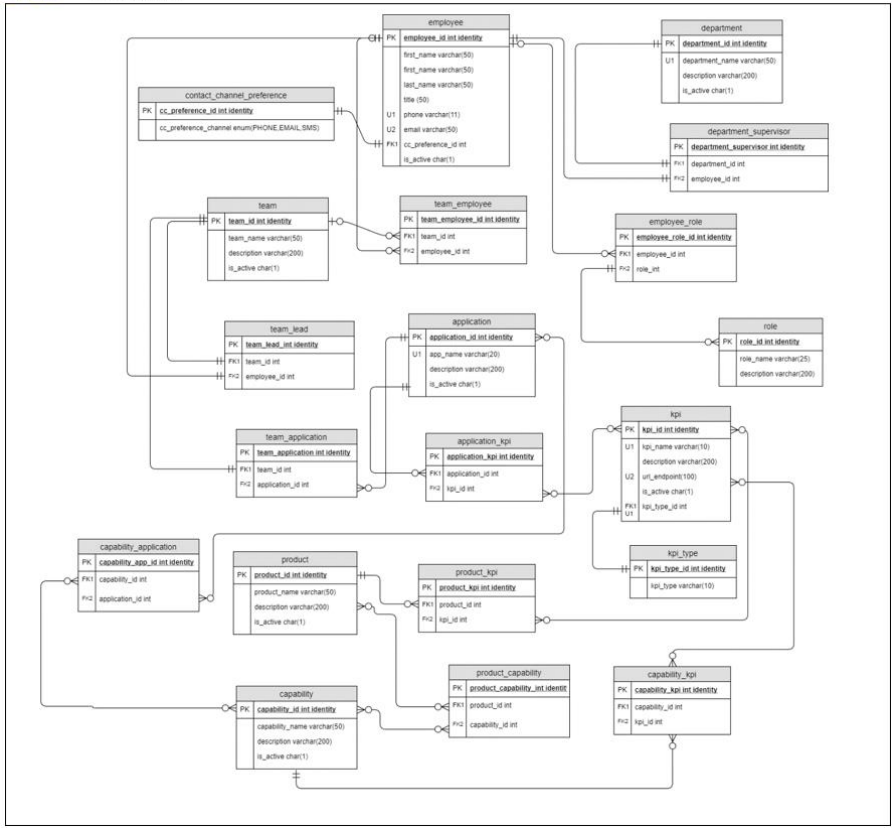
To **measure on-going success** and determine the business value of pushing this new process to other teams within the organization, **controls** were **implemented** that would **continue** to **measure** and **report cycle-times** along with POs being **approved** within the given discount thresholds.

Learning Objective

-Develop a plan of action to implement the business decisions derived from analysis-

➢ A4B KPI BizOps Organization Database

3.1.1 Diagram



Develop a plan of action
IST 659 Database Admin and Management
-A4B KPI BizOps Organization Database-

- **Project Initiative Name**: VXPD A4B 'KPI Buddy' Alexa Skill Pilot
- **Team**: VXPD DevOps
- **Project Type**: Agile Sprint, 2-Week Cycles
- **Project Duration**: Three Sprint Cycles (6 Weeks)
- **Team Roles**: Product Owner, Scrum Master, Voice Experience Designer, DBA, Voice Experience Developers, DevOps SRE

- Taking the design and implementation strategy laid out in this project, I ran an internal Pilot initiative with my DevOps team at work to prototype this database with new Alexa KPI Skills my team was building. Leveraging our AWS Cloud Environment, I set up an Amazon RDS for SQL Server instance and implemented this BizOps Organization Database on AWS.

- The Pilot was a success, offering our internal customers (Product and Technology Organization) a simple voice interface, customized to their level of KPI needs.

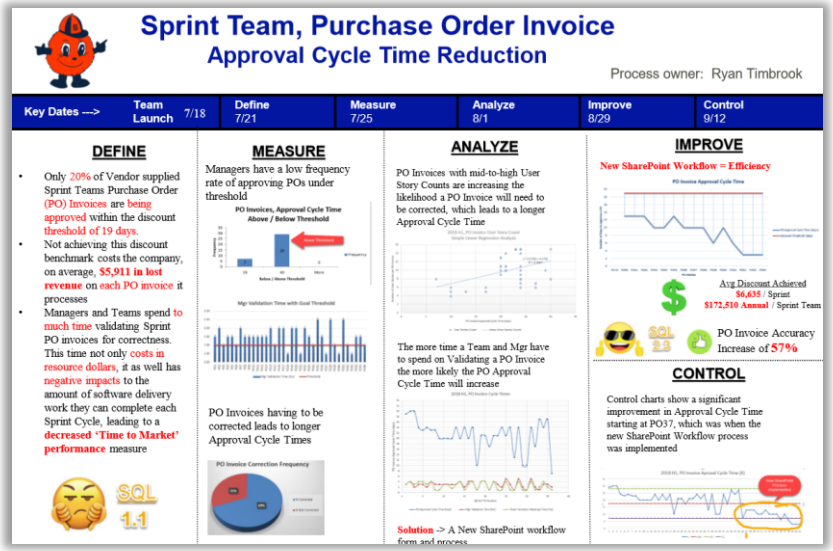
-Demonstrate communication skills regarding data and its analysis-
Overview

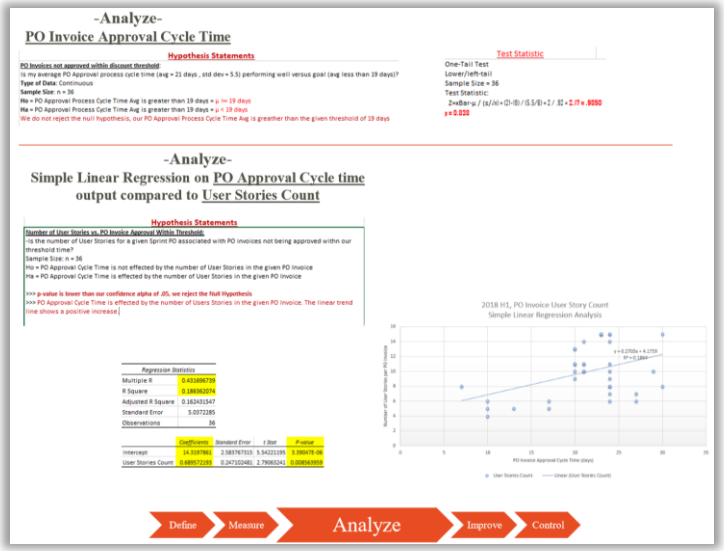
- Purchase Order Process Improvement
- Public Sentiment Toward NFL Teams, Coaches, and Players
- IEEE-CIS Fraud Detection

Learning Objective

-Demonstrate communication skills regarding data and its analysis-

➢ Purchase Order Process Improvement





Demonstrate communication skills regarding data and its analysis

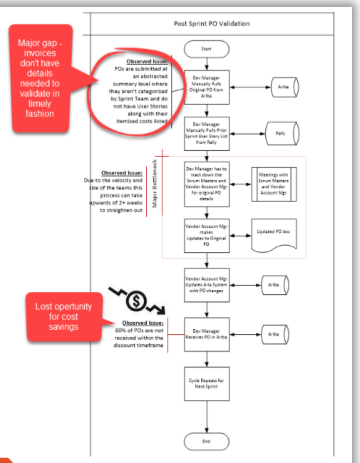
MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-

-Define-

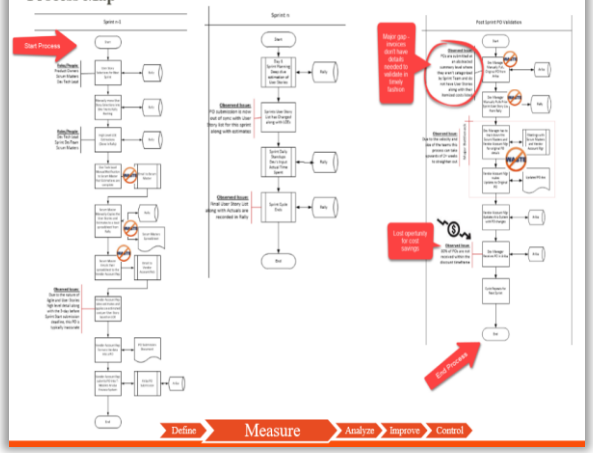
The Business Problem:

- Only 20% of Vendor supplied Sprint Teams Purchase Order (PO) Invoices are being approved within the discount threshold of 19 days.
- Not achieving this discount benchmark costs the company, on average, **$5,911 in lost revenue** on each PO invoice it processes
- Managers and Teams spend to much time validating Sprint PO invoices for correctness. This time not only costs in resource dollars, it as well has negative impacts to the amount of software delivery work they can complete each Sprint Cycle, leading to a **decreased 'Time to Market' performance** measure



Define | Measure | Analyze | Improve | Control

-Measure-
Process Map



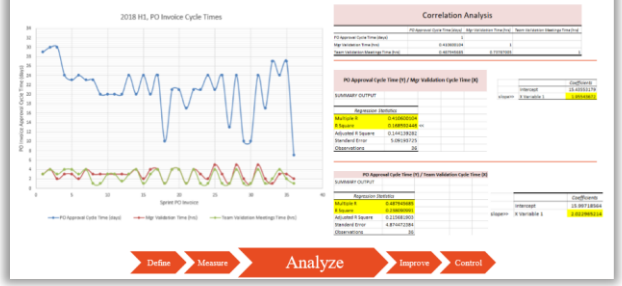
Define | Measure | Analyze | Improve | Control

-Measure-
PO Invoice Approval Cycle Time

Define | Measure | Analyze | Improve | Control

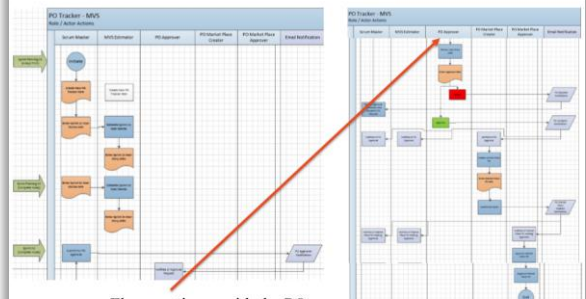
-Analyze-

Correlation Analysis of PO Approval Cycle Time with Mgr Validation Time and Team Validation Time



Define | Measure | Analyze | Improve | Control

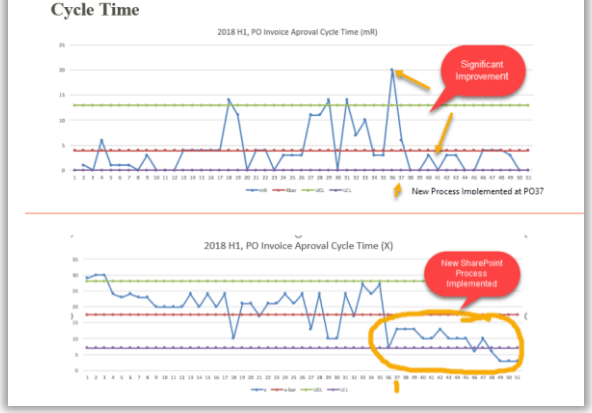
-Improve-
New Process Flow



Flow continues with the PO Approver in the next image

Define | Measure | Analyze | Improve | Control

-Control-
PO Invoice Approval Cycle Time

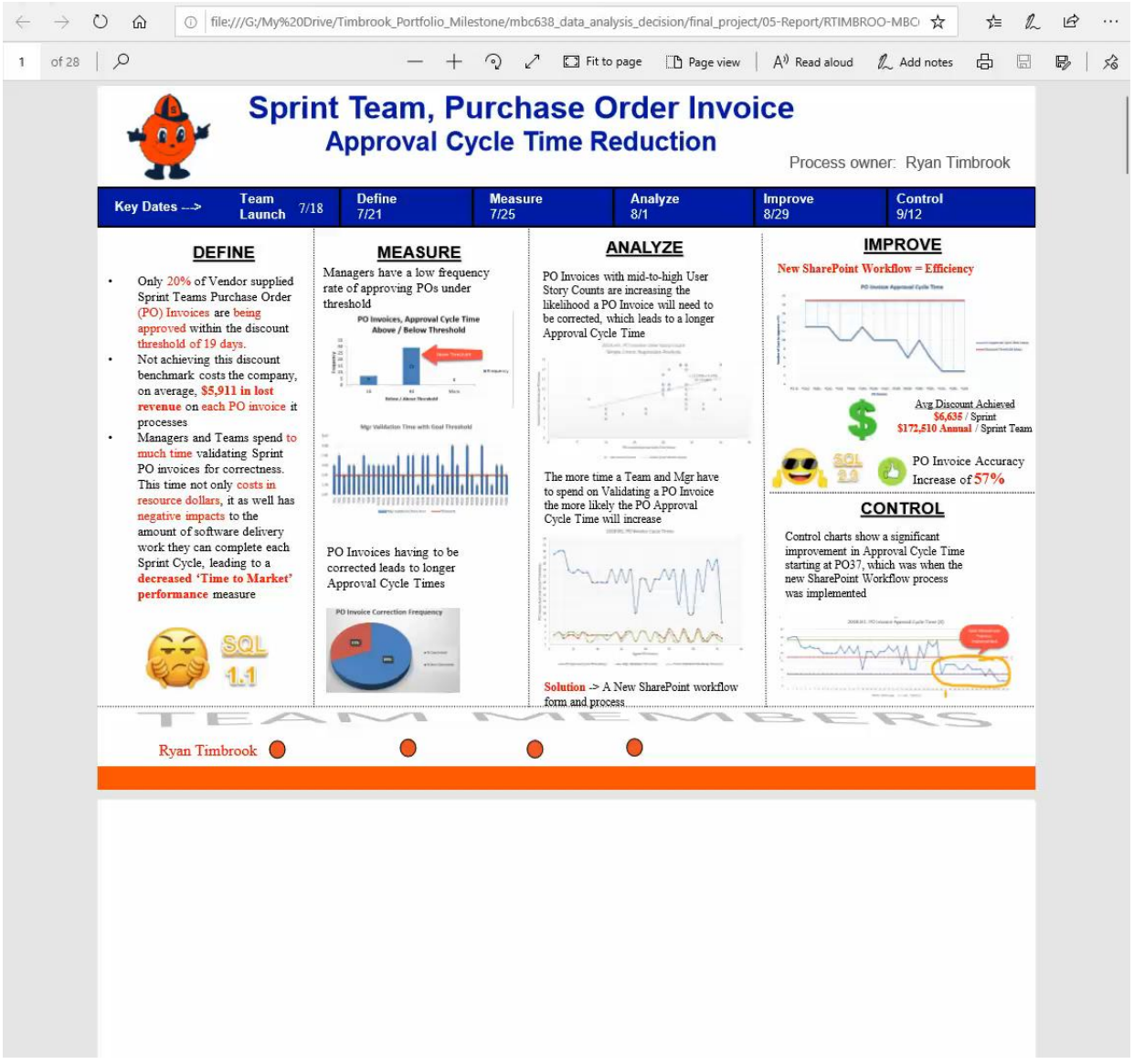


Demonstrate communication skills regarding data and its analysis

MBC 638 Data Analysis and Decision Making

-Purchase Order Process Improvement-

[Project Presentation Deck]

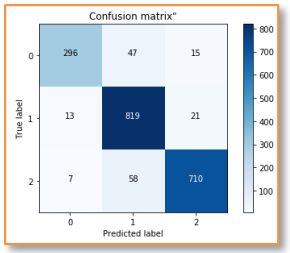


Learning Objective

- Demonstrate communication skills regarding data and its analysis-

➢ Public Sentiment Toward NFL Teams, Coaches, and Players

	precision	recall	f1-score	support
negative	0.94	0.83	0.88	358
neutral	0.89	0.96	0.92	853
positive	0.95	0.92	0.93	775
micro avg	0.92	0.92	0.92	1986
macro avg	0.92	0.90	0.91	1986
weighted avg	0.92	0.92	0.92	1986



Model_Type	Vectorize	N_Gram	Experiment_Model_Nam	Cross_Fold	Prediction_Accurac	Total_Prediction_Point	Test_Recall_Score_Av	Test_Precision_Score_Av	Train_Recall_Score_Av	Train_Precision_Score_Av	Total_Build_Time	Total_Predict_Tim	Confusion_Matri
LinearSVM	count	unigram	task3_svm_cnt_vec_unigram	10	92.42	1953	0.9133	0.9312	0.9877	0.9923	7.9983	0.001	[[282 37 22] [9 788 22] [9 49 735]]
LinearSVM	count	bigram	task3_svm_final_cnt_vec_big	10	91.4	1953	0.905	0.9218	0.9932	0.996	7.389	0.0017	[[295 40 22] [9 770 32] [11 54 720]]
LinearSVM	count	unigram	task2_svm_cnt_vec_unigram	10	90.99	3905	0.8963	0.9165	0.9951	0.9967	7.3672	0.0014	[[590 74 40] [29 1535 56] [27 126 1428]]
LinearSVM	count	unigram	task1_svm_cnt_vec_unigram	10	90.73	3905	0.9003	0.9189	0.9954	0.9972	5.5715	0.0008	[[578 77 56] [23 1522 60] [27 119 1443]]
LinearSVM	count	bigram	task2_svm_cnt_vec_bigram_	10	90.6	3905	0.8956	0.9141	0.9969	0.9979	9.7347	0.002	[[584 81 52] [26 1558 66] [22 120 1396]]
LinearSVM	tfidf	bigram	task4_svm_tfidf_vec_bigram	10	89.73	3905	0.8853	0.9138	0.9697	0.9795	5.126	0.0022	[[532 67 64] [20 1541 99] [21 130 1431]]
LinearSVM	tfidf	unigram	task4_svm_tfidf_vec_unigrar	10	89.3	3905	0.8755	0.9043	0.9588	0.9733	5.4828	0.0014	[[576 79 56] [30 1488 98] [13 142 1423]]
Multinomial_Naive_Bayes	count	bigram	task2_mnb_cnt_vec_bigram_	10	84.1	3905	0.8443	0.8478	0.9047	0.9099	0.1792	0.0015	[[562 68 87] [54 1410 200] [45 167 1312]]
Multinomial_Naive_Bayes	count	unigram	task1_mnb_cnt_vec_unigrar	10	84.07	3905	0.822	0.8274	0.8821	0.8866	0.1242	0.0024	[[530 66 84] [59 1410 201] [72 140 1343]]
Multinomial_Naive_Bayes	count	unigram	task2_mnb_cnt_vec_unigrar	10	83.33	3905	0.8226	0.8262	0.886	0.8919	0.0947	0.0011	[[544 86 80] [65 1361 233] [53 134 1349]]

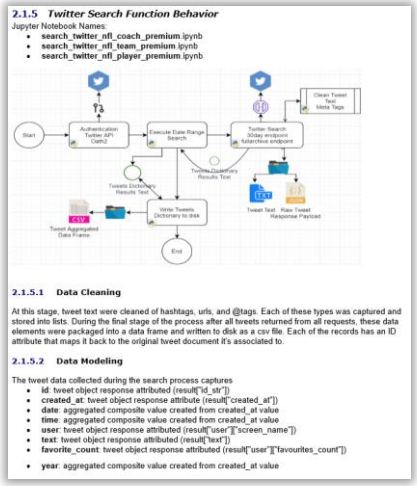
Demonstrate communication skills regarding data and its analysis

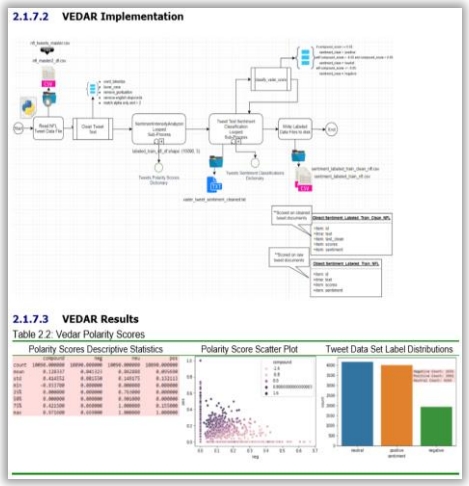
IST 736 Text Mining

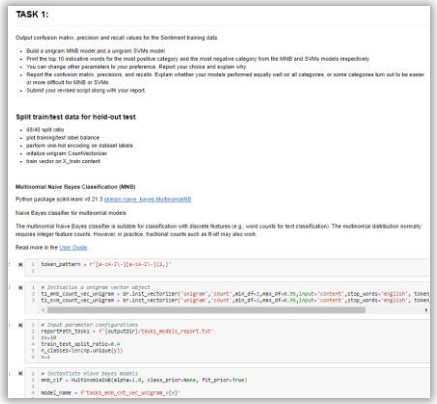
-Public Sentiment Toward NFL Teams, Coaches, and Players-

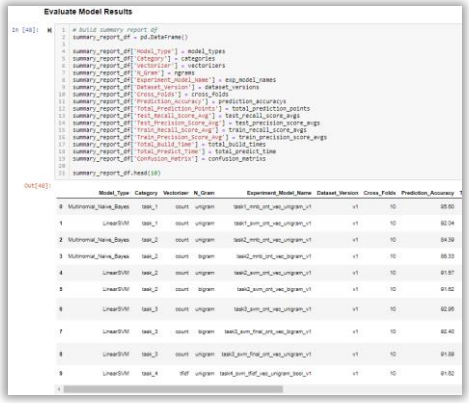
Types of Communication:

- Final Report (word document) – written as academic paper
- Final Presentation (power point deck) – live presentation summary of results and insights learned
- Autogenerated Model Performance Reports
- Source Code, Jupyter Notebooks Markup Language Comments









6 Conclusion

The most time-intensive, challenging, yet most critical aspect of this research is in mining the data. To be able to accurately apply machine learning algorithms to this type of unstructured, continuous-timeseries data for the purposes of trend analysis, mechanisms have to first be in place that can search social media sources like Twitter for relevant tweets on a daily and or historical bases. Setting up these programs should be well thought out and designed such that the data, in its various forms, is thought of as the final product. How the data is stored, cleaned and propagated through its pipeline will determine the quality of analysis and experimentation that can be done on it.

Technologies such as VADER (Hutto & Gilbert, 2014) for sentiment scoring, the Natural Language Toolkit (NLTK) (Bird, Loper, & Klein, 2009) for natural language processing and SciKit-Learn (Pedregosa, et al., 2011) for general machine learning have helped to commoditize these disciplines. The techniques made available to the average programmer today offer a wide range of options to attempt to extract value from data that is available either freely (albeit sometimes in limited quantities) or at a known subscription fee.

The National Football League (NFL) remains one of the most popular sports in America (Norman, 2018) and Fantasy Football has grown by leaps and bounds. In this age, popularity almost inevitably leads to an explosion of unstructured data across social networks such as Facebook and Twitter. This freely provided user opinion when correctly gathered and curated can provide a wealth of information about how those who follow both the NFL and fantasy sports feel about a particular team, coach, or player at any given week, especially during the active season. This data can be collected, and the sentiment of the fan base can be measured and tracked over time.

It is possible to look at this same user-provided data to determine the subjects about which changes to user sentiment are driven. The addition of topic modeling to this data may be a useful tool for the purposes of marketing. Not only could the NFL determine what the socially vocal fans are discussing, but it can determine the sentiment within those categories. The implications of this may hold well from a fantasy perspective as well, as these topics/sentiments can be used for assessing the cost of adding players to a fantasy team in games that are week to week.

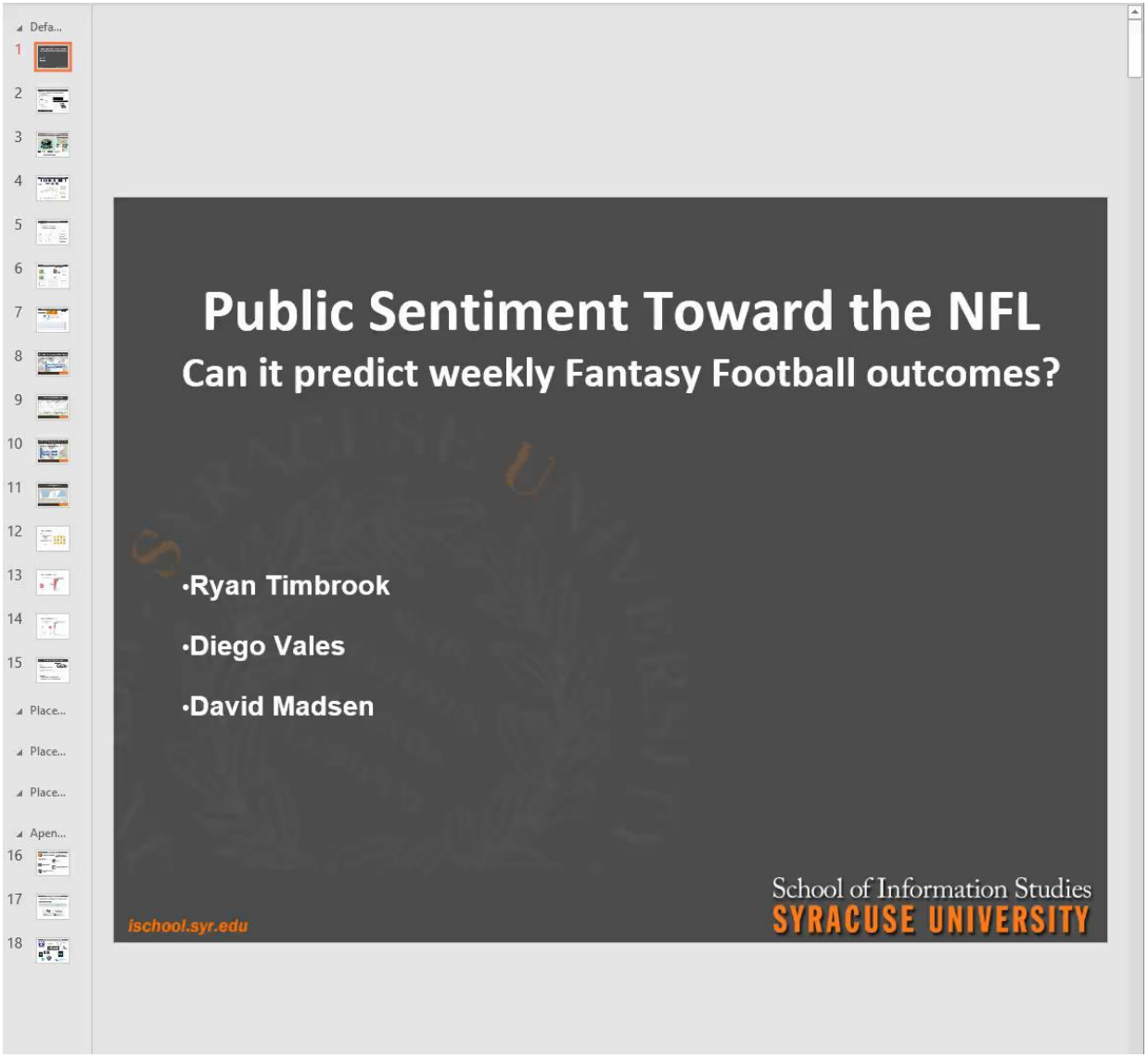
The validity of predicting player performance based on sentiment is, at the moment, inconclusive. While there is undoubtedly immediate benefit from analyzing social media sentiment for the NFL (it can be used to determine marketing strategy), nothing in the presented analysis can conclusively determine that a player's performance can be obtained from the consciousness of the 'Twittersphere.' Future efforts to determine the true validity of using sentiment in the prediction of a players' performance need to include data from each of the fantasy football service providers. Each provider uses different statistical methods and AI to forecast player performance. While this analysis did not show immediate promise based on the limited slice of data used, it is possible that the addition of new data sources and the altering of methods may lead sentiment to be a variable in predictive models.

Demonstrate communication skills regarding data and its analysis

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

[Project Presentation Deck]

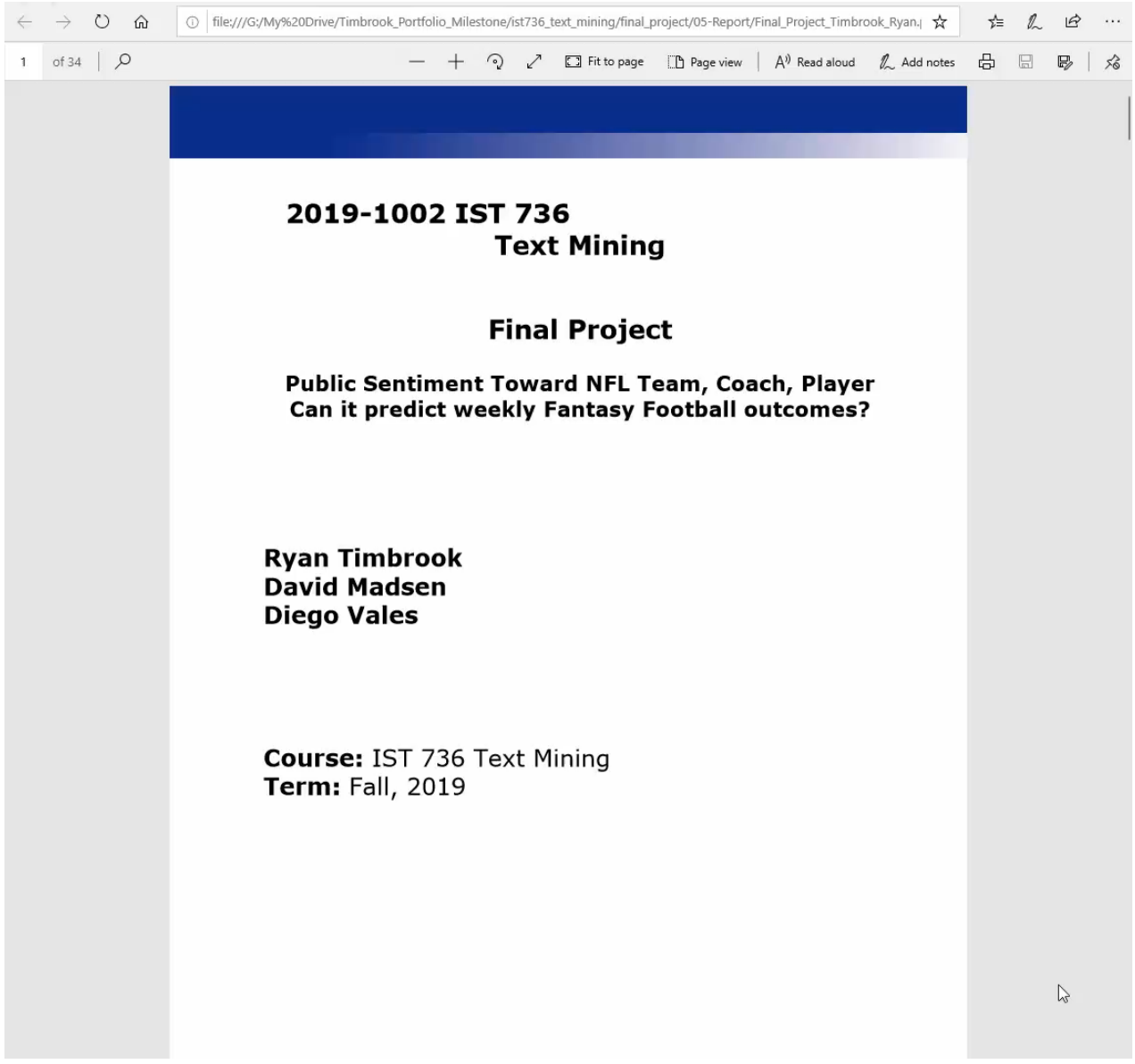


Demonstrate communication skills regarding data and its analysis

IST 736 Text Mining

-Public Sentiment Toward NFL Teams, Coaches, and Players-

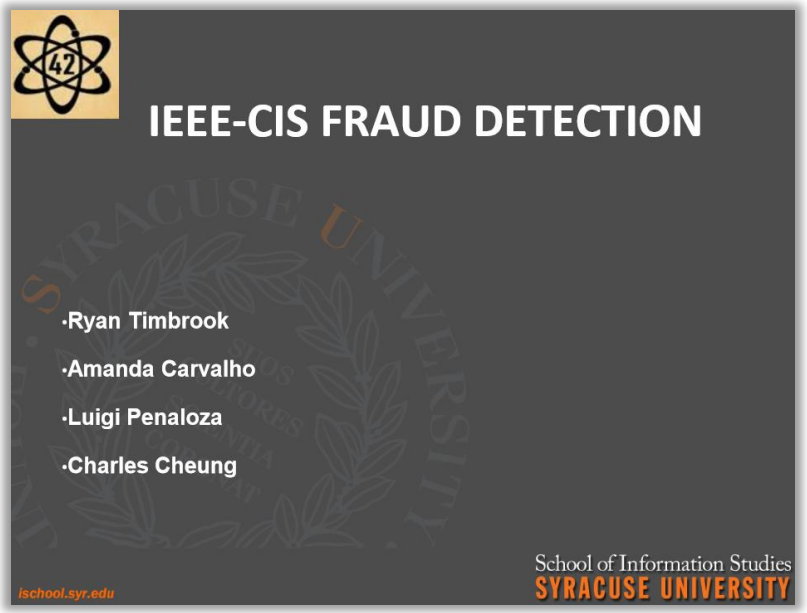
[Project Final Report]



Learning Objective

- Demonstrate communication skills regarding data and its analysis-

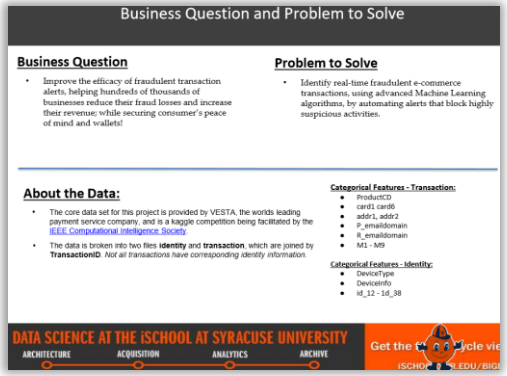
➢ IEEE-CIS Fraud Detection

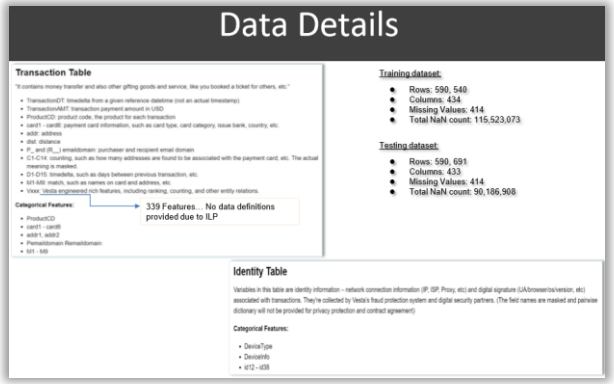


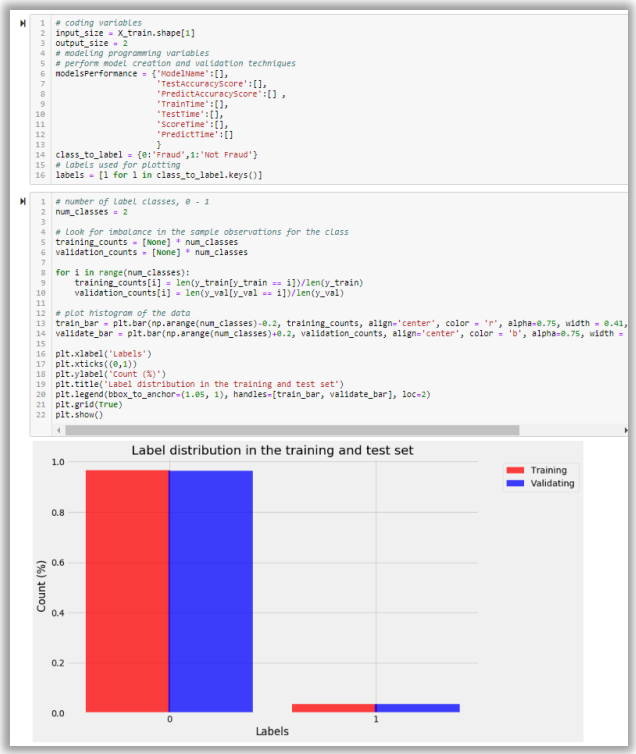
Demonstrate communication skills regarding data and its analysis
IST 718 Big Data Analytics
-IEEE-CIS Fraud Detection-

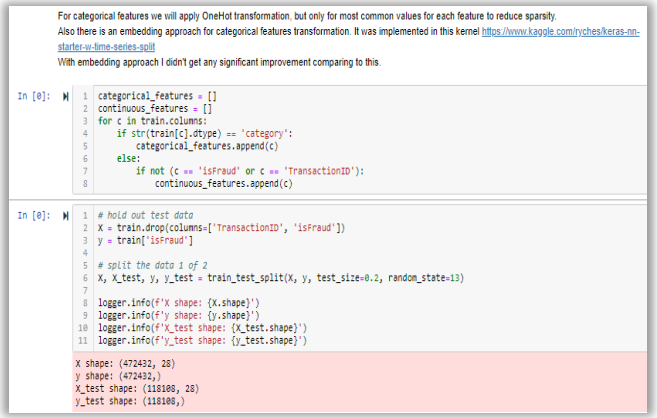
Types of Communication:

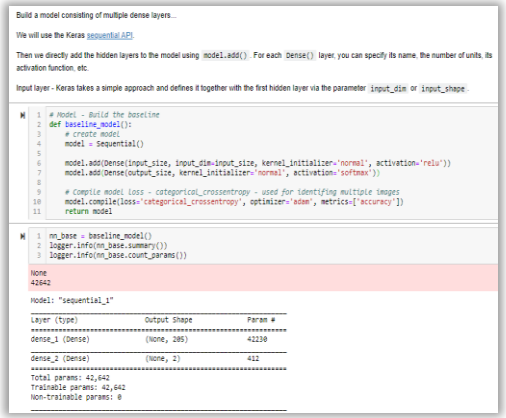
- Final Presentation (power point deck) – live presentation summary of results and insights learned
- Autogenerated Model Performance Reports
- Source Code, Jupyter Notebooks Markup Language Comments









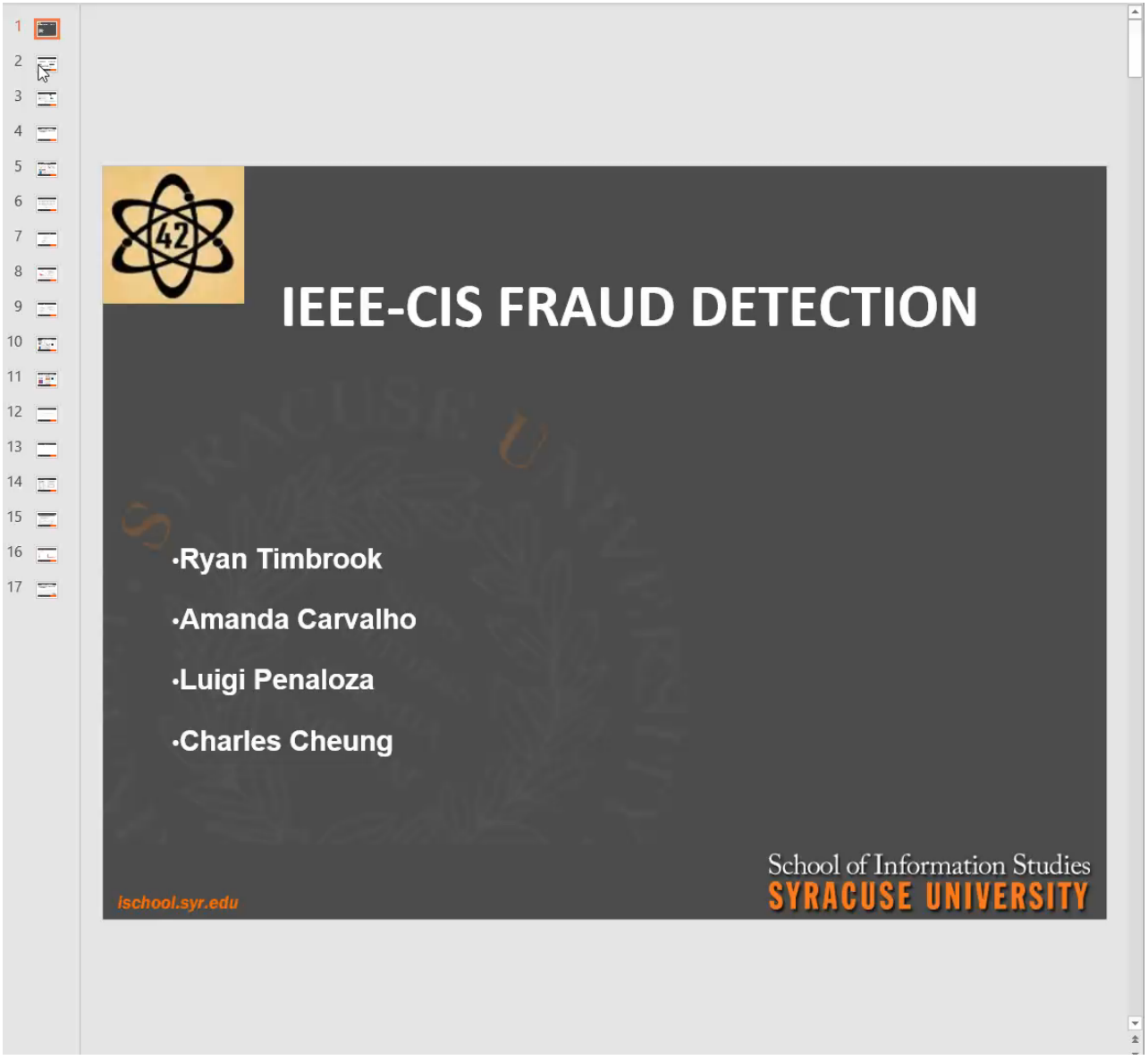


Demonstrate communication skills regarding data and its analysis

IST 718 Big Data Analytics

-IEEE-CIS Fraud Detection-

[Project Presentation Deck]



-Synthesize the ethical dimensions of data science practice- Overview

- A4B KPI BizOps Organization Database
- IEEE-CIS Fraud Detection

Learning Objective

-Synthesize the ethical dimensions of data science practice-

- A4B KPI BizOps Organization Database

Synthesize the ethical dimensions of data science practice
IST 659 Database Admin and Management
-A4B KPI BizOps Organization Database-

- IAM Rules – ‘least privilege’
- Hierarchical Relationship Model Design
- SQL Database Role-Based Permissions

Learning Objective

-Synthesize the ethical dimensions of data science practice-

- IEEE-CIS Fraud Detection

Synthesize the ethical dimensions of data science practice
IST 718 Big Data Analytics
-IEEE-CIS Fraud Detection-

- Machine Learning Algorithms to identify real-time fraudulent transactions
- Automatic alerting and blocking of highly suspicious activities
- Protect the people and the banking systems

- Just about **everyone** uses **e-commerce technology** and modern banking systems are at risk of being a victim of fraud.
- It cost both the individual as well as the merchant who offers refunds for fraudulent transactions; and not all scenarios are covered, leaving many individuals to pay.
- Company's who have had data security breaches put everyone who uses electronic forms of payment at risk.

A few of the larger breaches such as:

- **eBay in 2014 with 145 million** user **accounts compromised**
- **Heartland** Payment Systems in **2008** had **134 million** users **credit cards** stolen
- **Target** in **2013** had up to **110 million** customers **credit/debit card**
- **Yahoo** in **2014** with it's **1 billion** user **accounts** and **passwords** compromised

Thank you!

MS Applied Data Science program at Syracuse University School of Information Studies

Learning Outcome Objectives Covered:

- 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 analysis
- Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals
- Synthesize the ethical dimensions of data science practice

About Me:

Microsoft Word Document

RYAN TIMBROOK

Seattle, WA 98107
ryan.timbrook@gmail.com
206.516.9956
https://github.com/rtimbro185

SR. LEADER IN SOFTWARE ENGINEERING

Hands-on Technologist Leading High-velocity Product Development Teams That Exceed Quality Delivery Expectations

A top-performing Leader in Software Engineering with 21+ years of Product Development experience combining expert SDLC management skills with an extensive background in Enterprise Software Development, Speech/Text -- Natural Language Technologies, and Applied Data Science. Directs multiple projects/teams simultaneously; Hands-on technologist with a proven track record of leading multi-disciplinary, high-velocity, product development teams that exceed quality delivery expectations through Agile DevOps implementation methodologies. Relentless focus on solutions and continuous improvement opportunities. Display genuine expertise in business and technology, thus excelling in maintaining alignment of business and technology teams to achieve all business objectives — strong analytical, organizational and problem-solving skills, ability to identify operational deficiencies, and excellent relationship management skills. Recognized as a trusted advisor to executive management teams.

Areas of Expertise

- Applied Data Science
- Natural Language Technologies
- Speech Recognition, Voice Self-Service
- Enterprise Software Development
- Contact Center Technologies
- Agile DevOps, Continuos Product Delivery