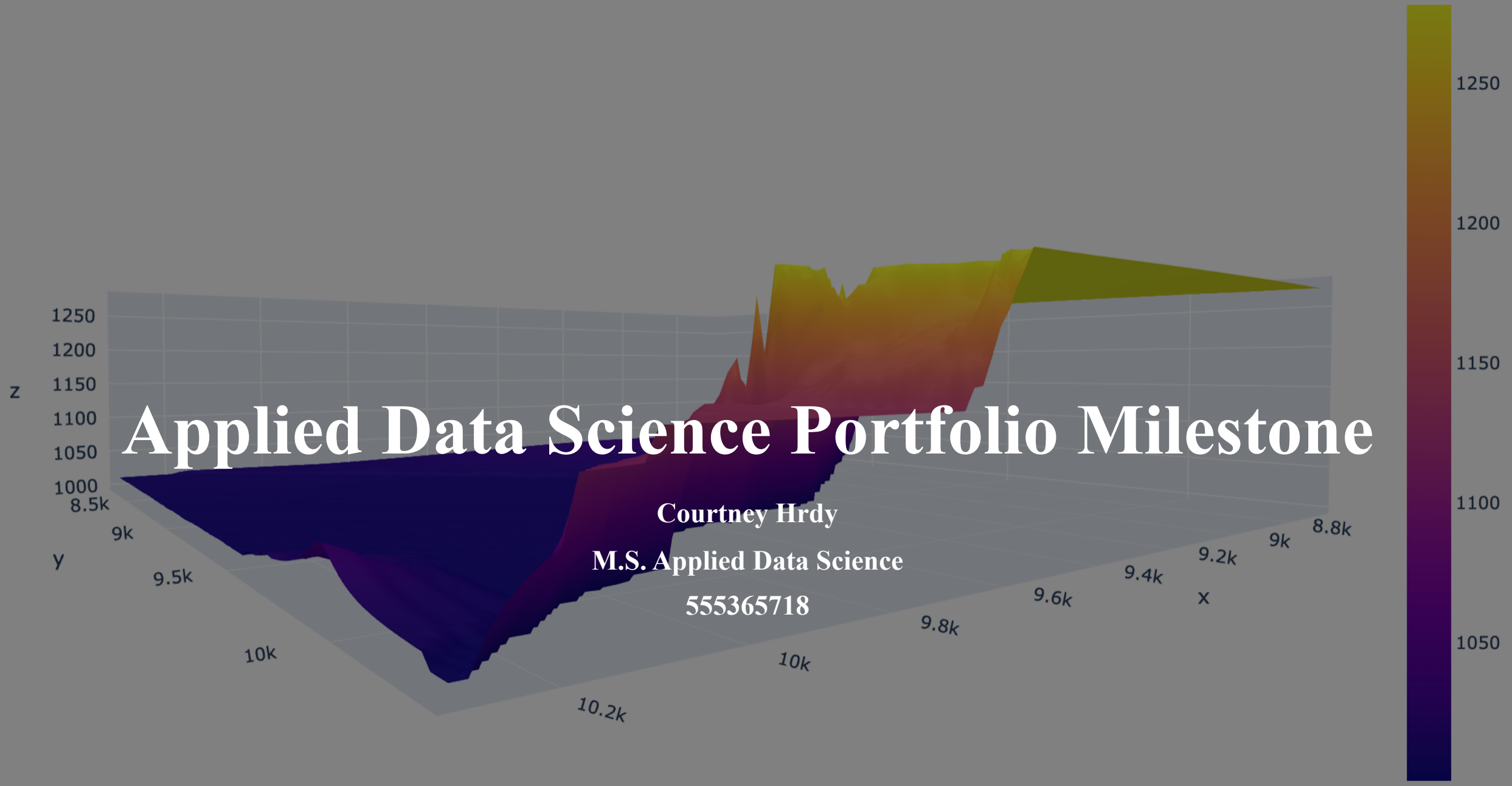


Applied Data Science Portfolio Milestone

Courtney Hrdy

M.S. Applied Data Science

555365718

Introduction

The MS in Applied Data Science is a practitioner's degree- while the curriculum is founded upon firm theoretical underpinnings, the project-based research reinforces professional underpinnings of data science.

- IST 707 Applied Machine Learning
- ACC 652 Accounting Analytics
- IST 718 Big Data Analytics
- IST 659 Database Administration and Concepts
- IST 974 Internship in Data Science

Achievement of Program Learning Goals

Collect, store, and access data by identifying and leveraging applicable technologies

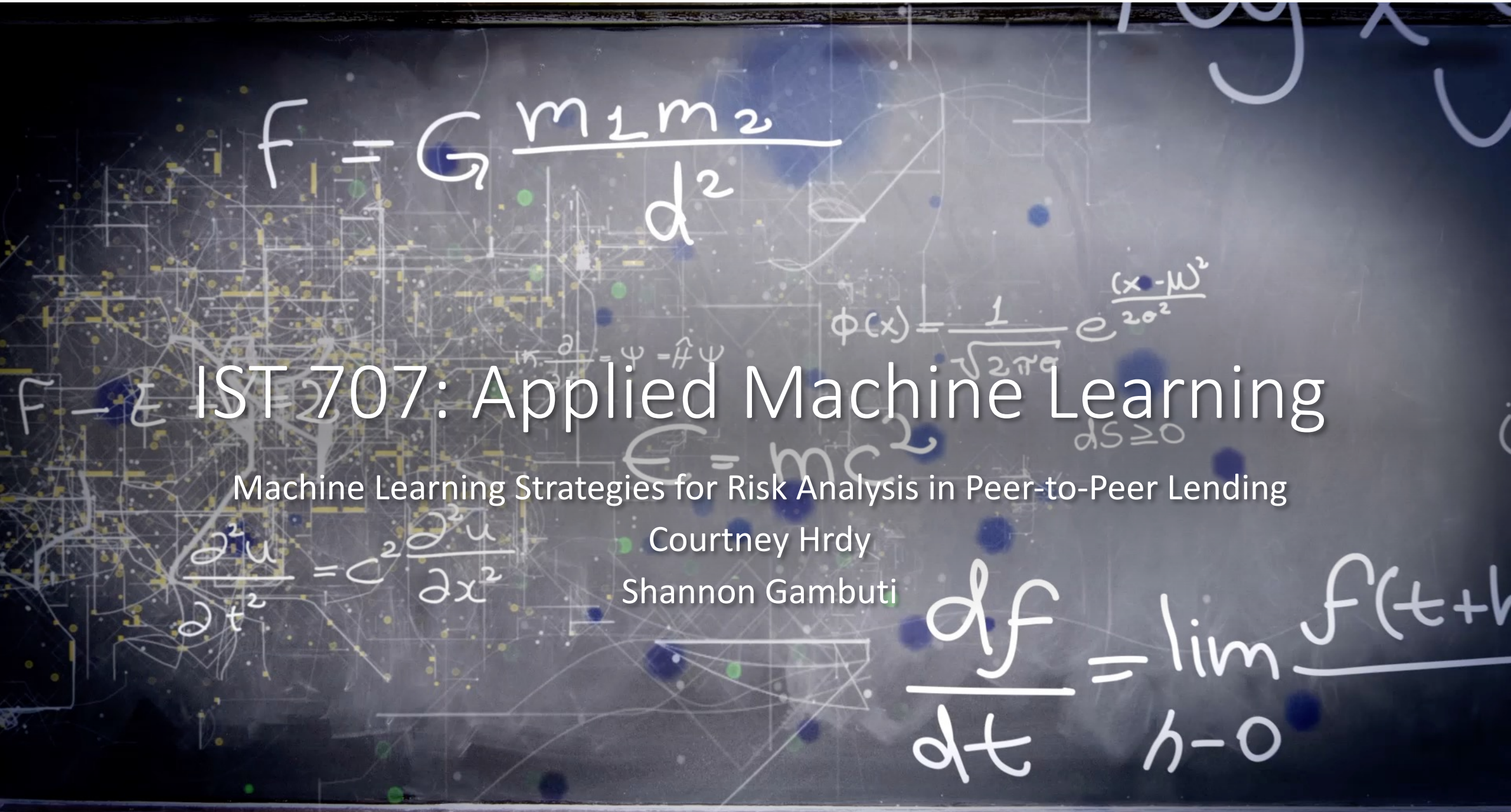
Create actionable insights across a range of contexts (accounting, aerospace, banking, financial, human resources)

Apply visualization and prediction models to help generate actionable insight

Use programming languages such as R, SQL, and Python to support the generation of actionable insight

Communicate insights gained via visualization and analytics to a broad range of audiences (project sponsors and team leads)

Apply ethics in the development, use, and evaluation of data and predictive models (fairness, bias, transparency, privacy)



IST 707: Applied Machine Learning

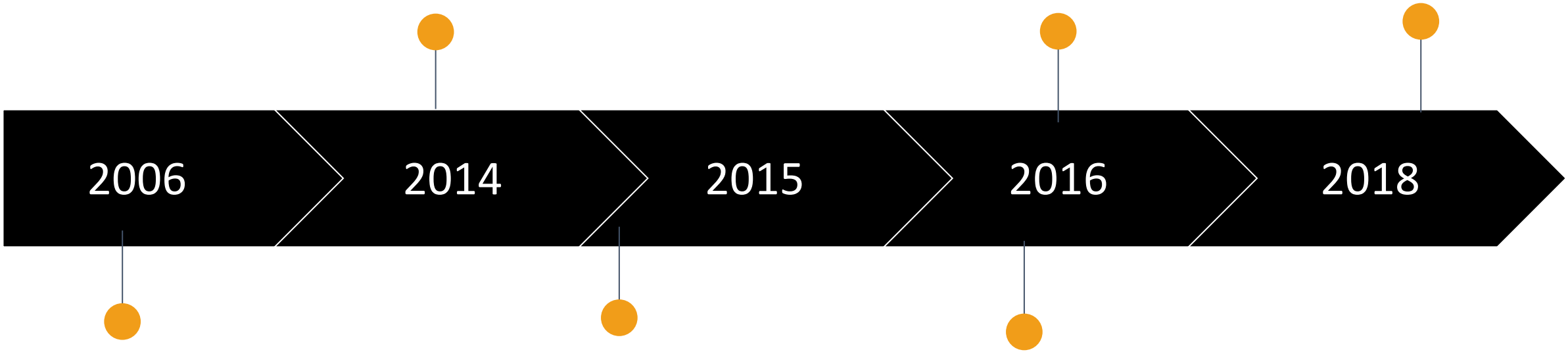
Machine Learning Strategies for Risk Analysis in Peer-to-Peer Lending

Courtney Hrdy

Shannon Gambuti

Overview

- LendingClub is a U.S.-based peer-to-peer lending company that provisions 3–5-year loans between **$1,000** and **$40,000** to accepted applicants based on a risk score calculation. The assigned risk score to all loan applicants determines whether a loan is accepted, and the interest rate the approved loans will receive.
- The risk score that has been used requires evaluation and improvement, given the increase in defaulted and charged off loans. LendingClub is looking for a new proposed model that would help potentially assign a risk score to loan applicants.



CEO Renaud Laplanche launches Lending-Club as a Facebook application

LendingClub goes public and is worth more than $10 billion.

Allen Griffin designs an algorithm to classify duplicate borrowers and starts data mining the Lending Club database. Realizes how bad things are

CEO is fired and IPO drops to $8.

LendingCub data scientist is found to have falsified $22 million of loans sold to Jeffries and Goldman Sachs

Defaulted loans surge, data released to the public for risk scoring.

The Data

Accepted Loans Dataset
253,629 records
107 attributes

Rejected Loans Dataset
16,384 records
9 attributes

7 common attributes:

Amount Requested, Month, Purpose, DBI Ratio, State, Years of Employment, class (reject/approve)

Target Variable: ***Charged off***

LendingClub has given up on being repaid according to the original terms of the loan

Data Cleaning and Preprocessing Steps:

1. Removing null values:
   Removed all rows with missing values – **39% of data**
2. Converted percentages in interest rate column to decimals
3. Removed '<','+','year', 'years' from employment length column
4. Converted issue date from m-yy format to m-yyyy
5. Created new DataFrame with the following columns:

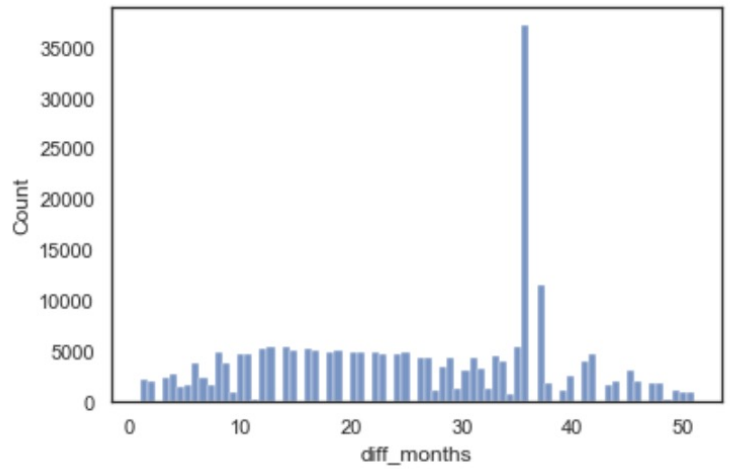
loan amount | interest rate | term | grade | employment length | issue date | month |loan status | payment plan | purpose | zip code | state | Debt-to-Income | delinquency | earliest credit line | open account | revolving balance | revolving utilization rate | total credit lines | application type

6. Merged new DataFrame with rejected dataset in Spark – joined on Class

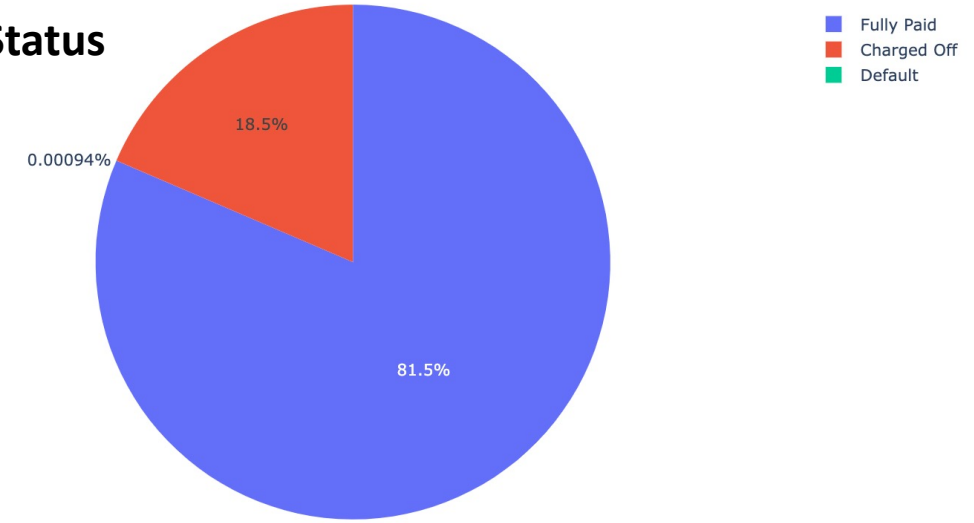
After preprocessing: **235,629** records

Exploratory Data Analysis

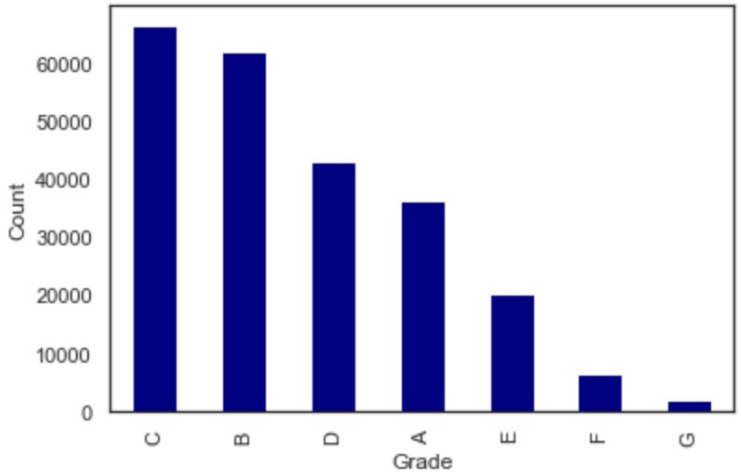
Frequency of Loan Duration



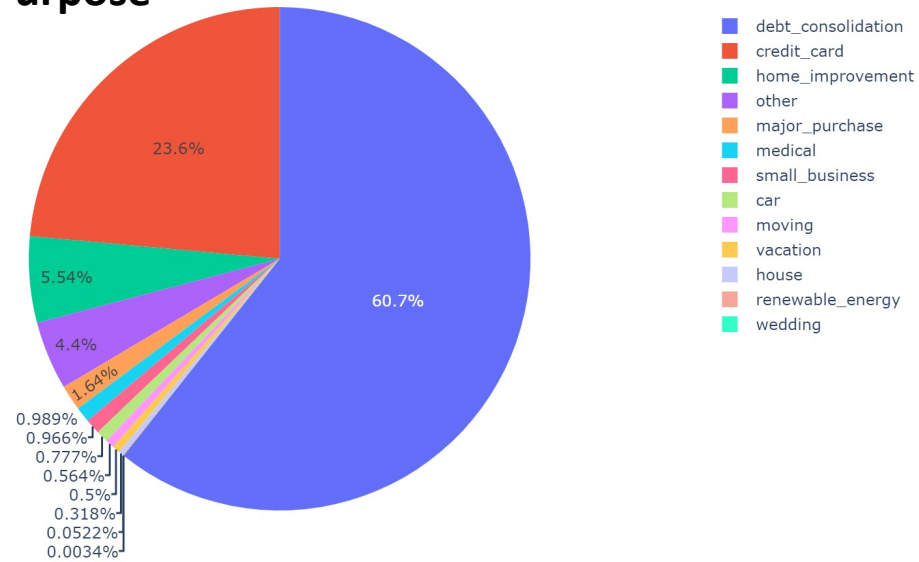
Loan Status



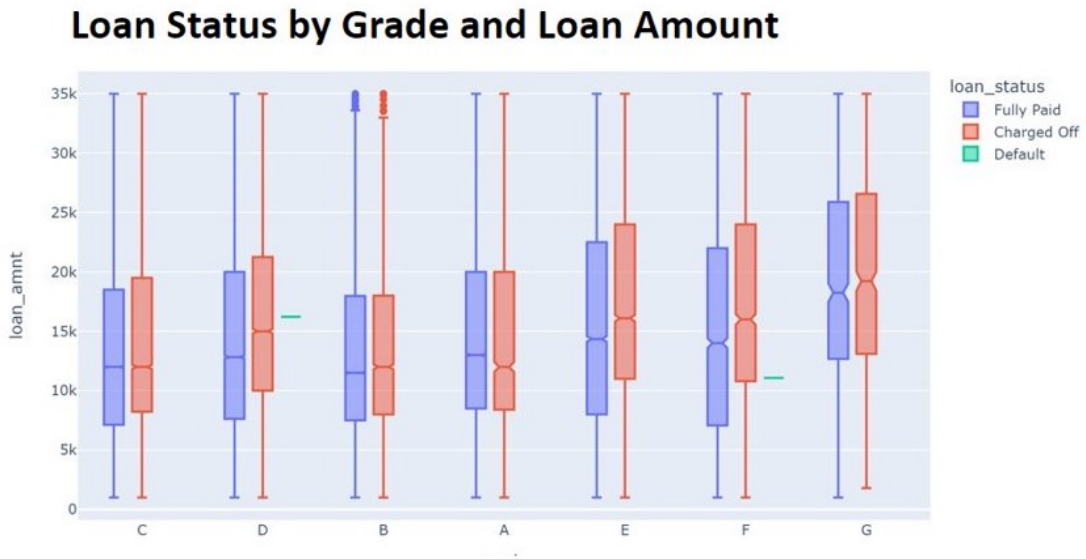
Distribution of Grades

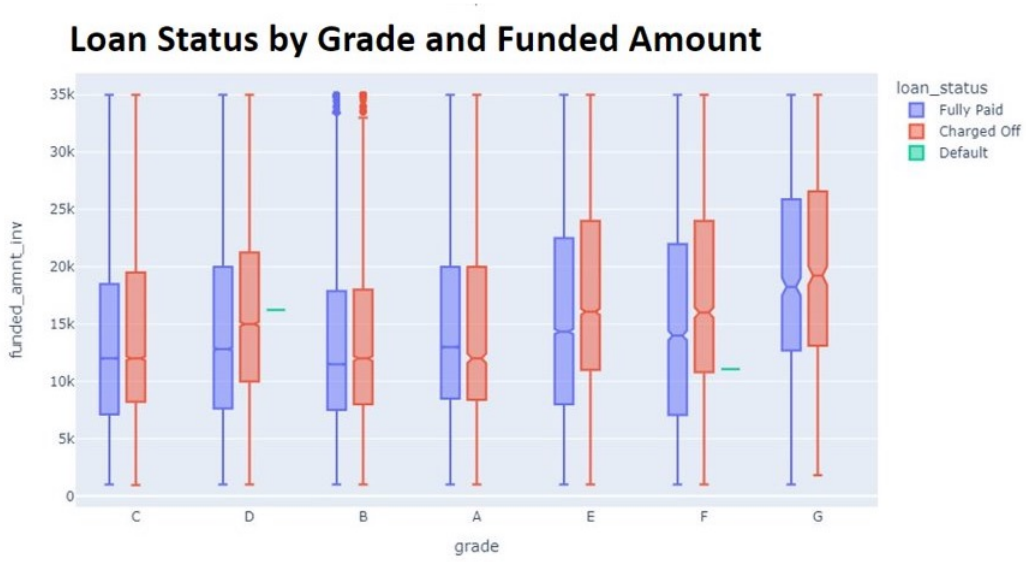


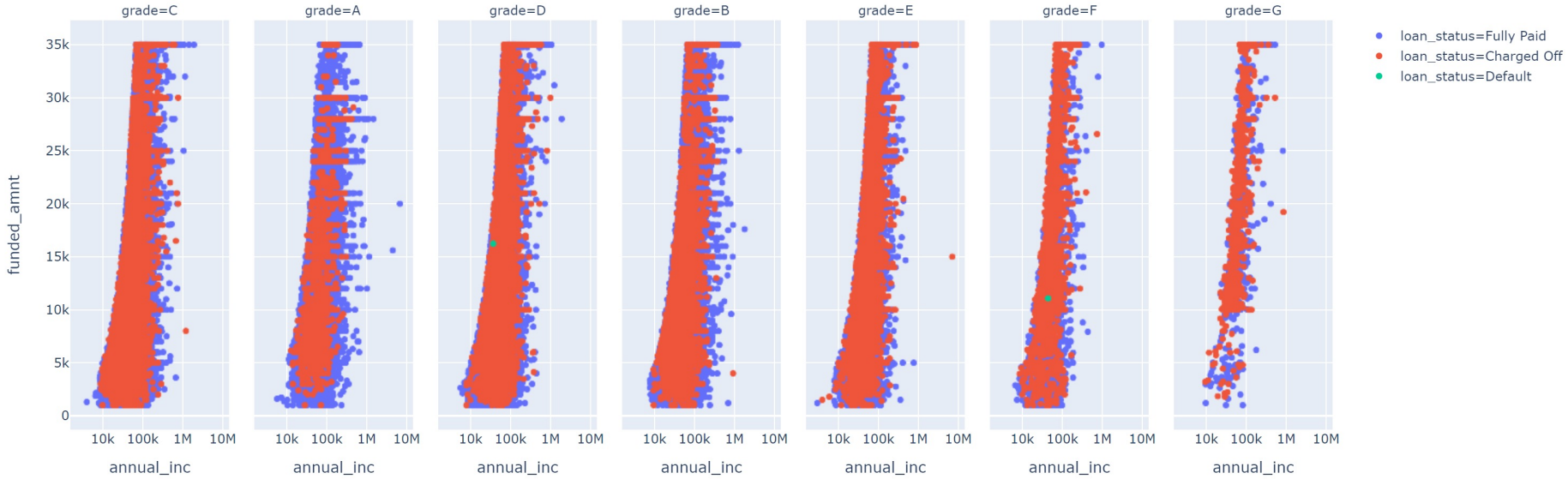
Loan Purpose



Exploratory Data Analysis



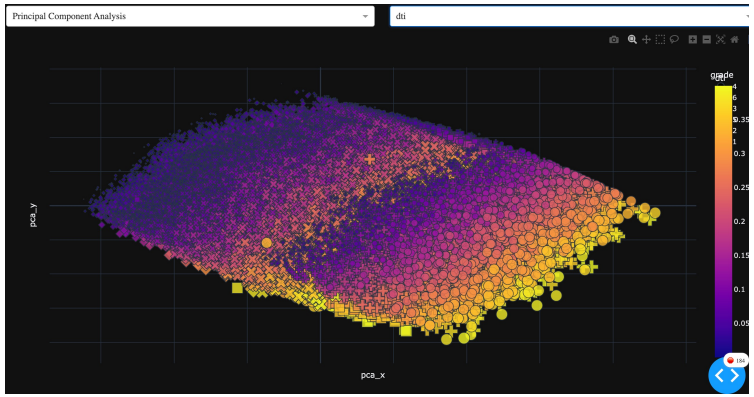


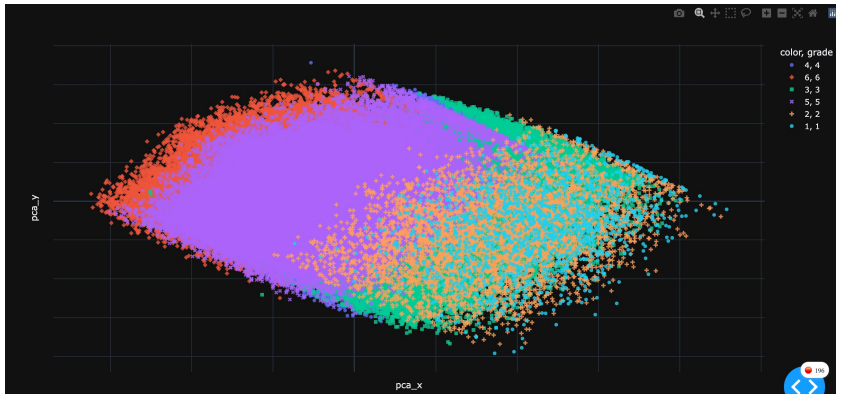


Unsupervised Learning Models

- Principal Component Analysis has practical implementation to risk analytics in finance
  - Good way for investors to evaluate their portfolio returns
  - Each principal component represents the percentage return on a statistical risk factor
  - Investors can optimize their portfolios by adjusting weights to match risk/return

PCA Visualization





Feature Selection

1. Spearman correlation matrix is used to identify variables with the strongest correlation to the target variable “charge off.”
2. Logistic regression and Random Forest use these features as independent variables.
3. All financial information is fit to Support Vector Machine, Gradient Boosting, KNN, and Logistic Regression.
4. Backwards regression is used to compare feature importance of financial and categorical information.
5. Linear regression to predict the Net Annualized Return based on features of a portfolio.

Random Forest

	feature	importance
6	open_acc	0.208423
2	int_rate	0.116139
4	dti	0.097576
1	loan_amnt	0.069827
5	delinq_2yrs	0.066586
3	emp_length	0.053145
0	annual_inc	0.017631
9	purpose	0.001621
10	region	0.000391
8	total_acc	0.000373
12	grade	0.000334
7	revol_bal	0.000000
11	term	0.000000

```
bce.evaluate(rf_model1
```

```
0.7027632111357343
```

Logistic Regression

	column	weight
6	open_acc	-0.673558
3	emp_length	-0.497545
11	term	-0.351590
9	purpose	-0.296352
7	revol_bal	-0.274637
12	grade	-0.234980
10	region	-0.219360
1	loan_amnt	-0.139913
8	total_acc	-0.128300
5	delinq_2yrs	-0.034754
0	annual_inc	0.094430
2	int_rate	0.144028
4	dti	0.154139

```
from pyspark.ml.evaluati
bce = BinaryClassificati
bce.evaluate(model.trans
```

```
0.7157492123585163
```

Model Evaluation

Model	Accuracy	F1	Precision	Recall
Logistic Regression	0.99	0.97	1.00	0.94
Random Forest	0.99	0.97	0.97	0.96
SVM -- Linear Kernel	0.87	0.53	1.00	0.36
SVM – Gaussian Kernel	0.87	0.53	1.00	0.36
SVM – Polynomial Kernel	0.84	0.31	0.99	0.18
Gradient Boosting	0.99	0.97	0.94	1.00
KNN	0.83	0.45	0.65	0.35

Challenges for AI Based Lending

- Bias
- Errors
- Time
- Ethics and privacy
- Regulation
  - “We know that a history of racism in financial services contributed to much of the economic inequality we see today. As tech, data, and AI plays an ever larger role in delivering quality financial services, this regulation is critical to ensuring these tools don’t lead to unintended outcomes. Instead of weakening these rules, this is the moment to reaffirm our shared commitment to a financial system that is designed to promote inclusion and opportunity for all Americans.”

ACC 652: Accounting Analytics

The DuPont Analysis of XBLR Financial Statements in Tableau

ACC 652: Project Overview

XBLR
- Access United Health and Cardinal Health financial statements using eXtensible Business Reporting Language (XBRL).

Return on Equity
- Compare the performance (return on equity) of United Health and Cardinal Health.

Dupont Analysis
- Calculate and visualize the Dupont Ratio from XBLR data using Tableau to generate actionable insight for investors.

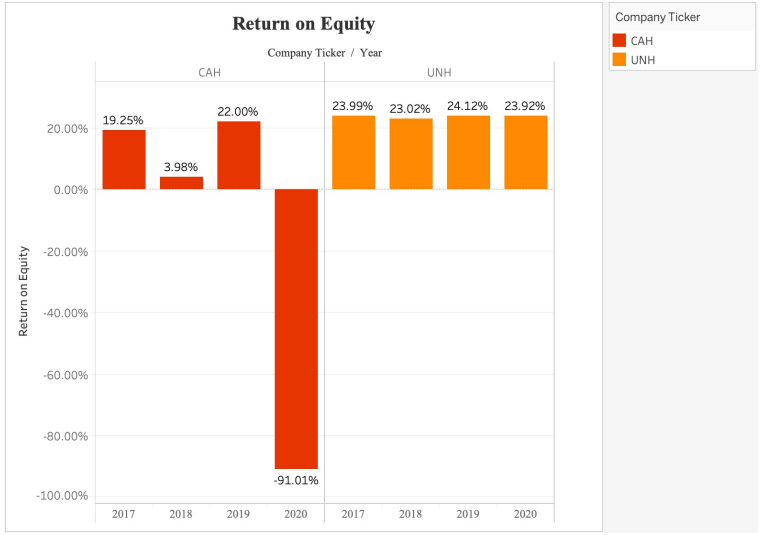
Financial Statement Analysis

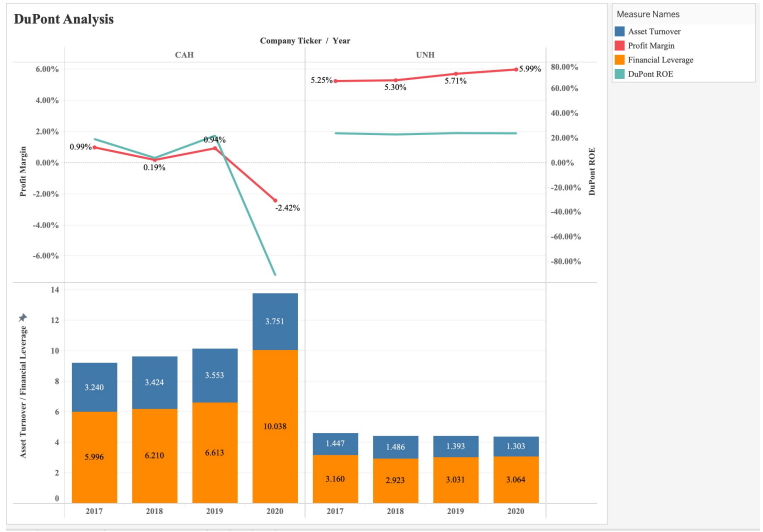
Main Company Ticker	UNH	(e.g. MSFT)	EntitySectorIndustryClassificationPrimary
Most Recent Year	2020	(e.g. 2014)	DocumentPeriodEndDate
Period	FY	(e.g. FY or Q2)	
Round to	1,000,000	(e.g. 100,000)	

Unitedhealth Group Inc

Financial Facts	2017	2018	2019	2020
CostOfGoodsAndServicesSold	$ 24,112	$ 26,998	$ 28,117	$ 30,745
OperatingIncomeLoss	$ 15,209	$ 17,344	$ 19,685	$ 22,405
ProfitLoss	$ 10,823	$ 12,382	$ 14,239	$ 15,769
Assets	$ 139,058	$ 152,221	$ 173,889	$ 197,289
StockholdersEquityIncludingPortionAttributableTo	$ 49,833	$ 54,319	$ 60,436	$ 68,328
NetIncomeLoss	$ 10,558	$ 11,986	$ 13,839	$ 15,403
Revenues	$ 201,159	$ 226,247	$ 242,155	$ 257,141

Analysis				
Sales Revenue - Cost of Goods Sold	$ 177,047	$ 199,249	$ 214,038	$ 226,396
Increase (Decrease) from previous year	0.00%	12.54%	7.42%	5.77%
ROA	7.78%	8.13%	8.19%	7.99%
Current Ratio	0.248	0.240	0.250	0.266
Profit Margin		5.30%	5.71%	5.99%
Change from previous		4.96%	3.36%	-2.05%
ROE (Profit Margin x ATO x FLEV)	21.72%	22.79%	23.56%	23.08%
Profit Margin (NI/Sales)	5.38%	5.47%	5.88%	6.13%
Asset Turnover (Sales/Total Assets)	1.447	1.486	1.393	1.303
Financial Leverage (Total Assets/Total SE)	2.790	2.802	2.877	2.887





Data Collection: XBLR

- United Health and Cardinal Health financial statements are collected using XBLR
- XBLR accesses data from the Securities and Exchange Commission Electronic Data Gathering, Analysis and Retrieval (EDGAR) system.

Financial Statement Analysis

Main Company Ticker	UNH	(e.g. MSFT)	EntitySectorIndustryClassificationPrimary
Most Recent Year	2020	(e.g. 2014)	DocumentPeriodEndDate
Period	FY	(e.g. FY or Q2)	
Round to	1,000,000	(e.g. 100,000)	

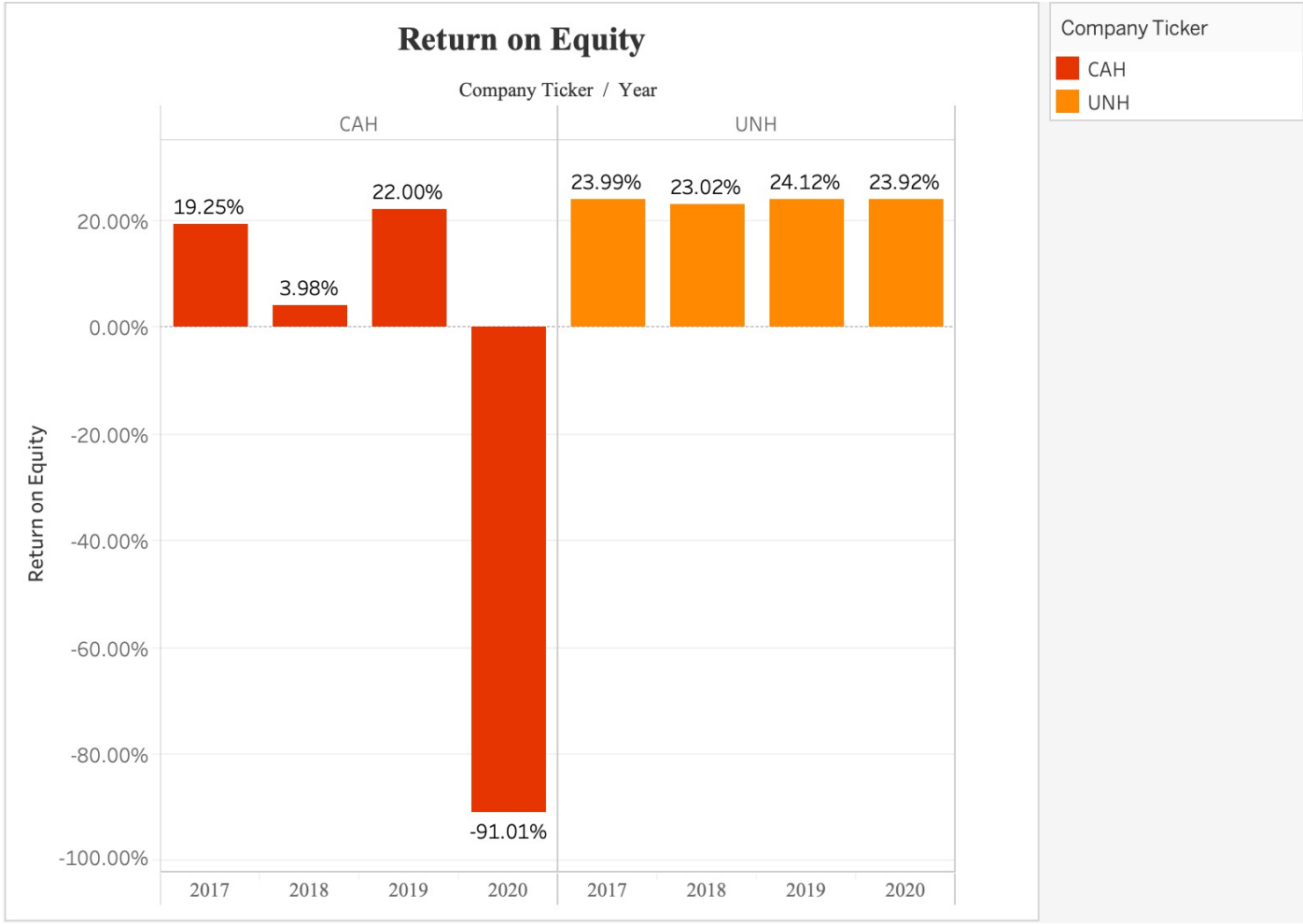
Unitedhealth Group Inc

Financial Facts	2017	2018	2019	2020
CostOfGoodsAndServicesSold	$ 24,112	$ 26,998	$ 28,117	$ 30,745
OperatingIncomeLoss	$ 15,209	$ 17,344	$ 19,685	$ 22,405
ProfitLoss	$ 10,823	$ 12,382	$ 14,239	$ 15,769
Assets	$ 139,058	$ 152,221	$ 173,889	$ 197,289
StockholdersEquityIncludingPortionAttributableTc	$ 49,833	$ 54,319	$ 60,436	$ 68,328
NetIncomeLoss	$ 10,558	$ 11,986	$ 13,839	$ 15,403
Revenues	$ 201,159	$ 226,247	$ 242,155	$ 257,141

Analysis				
Sales Revenue - Cost of Goods Sold	$ 177,047	$ 199,249	$ 214,038	$ 226,396
Increase (Decrease) from previous year	0.00%	12.54%	7.42%	5.77%
ROA	7.78%	8.13%	8.19%	7.99%
Current Ratio	0.248	0.240	0.250	0.266
Profit Margin		5.30%	5.71%	5.99%
Change from previous		4.96%	3.36%	-2.05%
ROE (Profit Margin x ATO x FLEV)	21.72%	22.79%	23.56%	23.08%
Profit Margin (NI/Sales)	5.38%	5.47%	5.88%	6.13%
Asset Turnover (Sales/Total Assets)	1.447	1.486	1.393	1.303
Financial Leverage (Total Assets/Total SE)	2.790	2.802	2.877	2.887

Data Understanding: Return on Equity

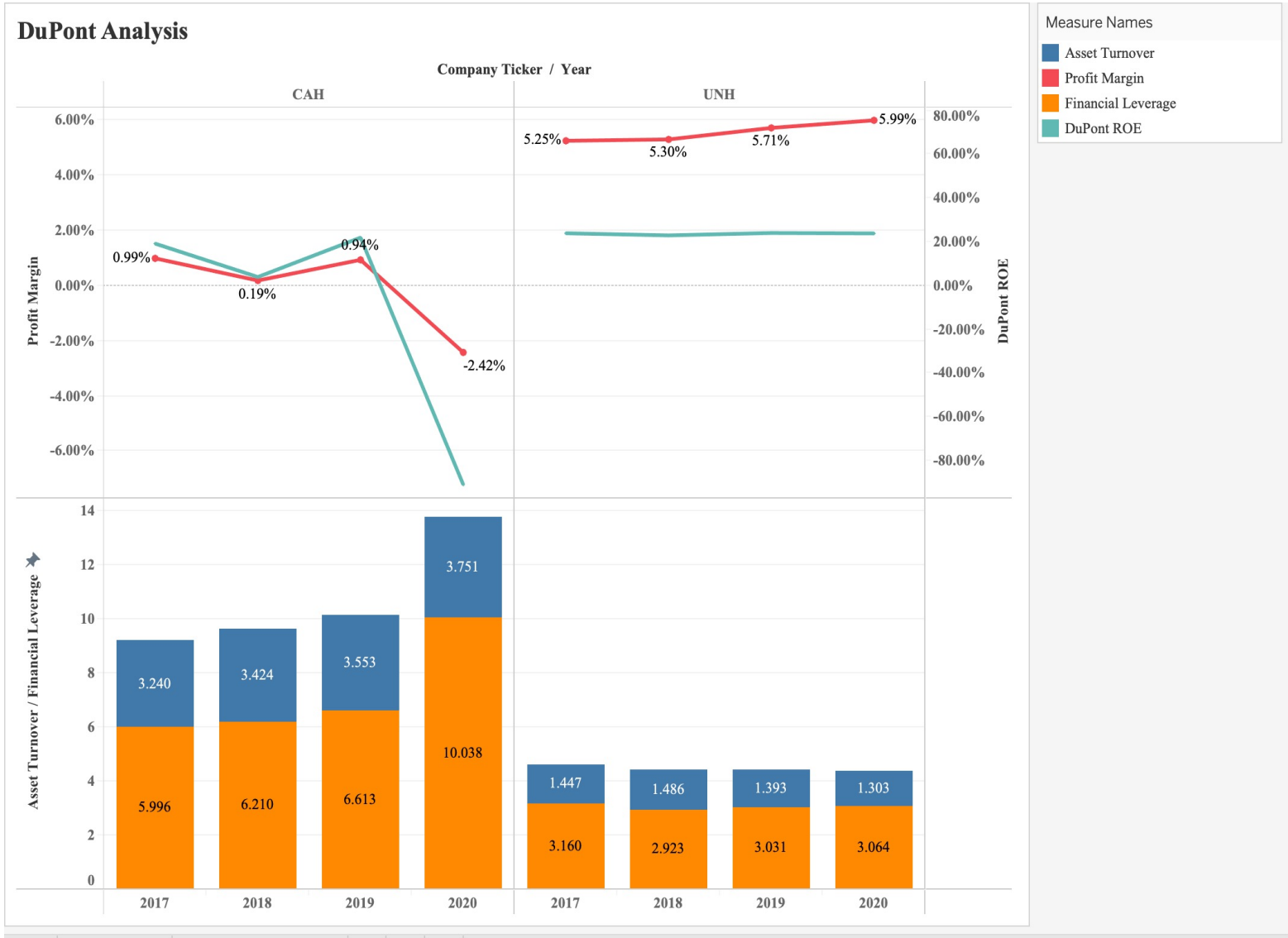
- The DuPont Analysis is used to better understand UnitedHealth's unscathed ROE during the pandemic and further analyze Cardinal Health's erratic decline.



The DuPont Analysis

$$ROE = \frac{Net\ Income}{Revenue} \times \frac{Revenue}{Total\ Assets} \times \frac{Total\ Assets}{Shareholder's\ Equity}$$

ROE = Profit Margin x Asset Turnover x Financial Leverage



ACC 652: Learning Outcomes

Create	Apply	Communicate	Apply
Create actionable insights across a range of contexts (auditing, managerial accounting, investment banking, insurance).	Apply ethics in the development, use, and evaluation of data in auditing and segregation of duties (fairness, bias, transparency, privacy).	Communicate insights gained via visualization and analytics to a broad range of audiences.	Apply Tableau visualizations to accounting models to help generate actionable insight.

IST 718: Big Data Analytics Optimizing Workplace Productivity

Dorothy Fang, Courtney Hrdy, Shannon Gambuti

Our Goal: offer a holistic, people-centric workplace evaluation

PRODUCTIVITY

Productivity is the main driver of profitability.

Example:

How can overall happiness or employee satisfaction link to increased productivity?

PROJECT OBJECTIVE

Evaluate and identify key areas of a workplace, **beyond financial incentive,** that can increase employee productivity.

Predict productivity through employee communication style.

DATASET

Survey from **Public Sector Commision of Western Australia** (2016)
Survey was completed by employees from 11 public sector organizations, with a total of **3883** valid responses and **109** questions (attributes)

Question Key

Question ID	Question
A1a	Please indicate your level of satisfaction with: My job overall
A1b	Please indicate your level of satisfaction with: My agency as an employer
A2a	My job allows me to utilise my skills, knowledge and abilities
A2b	I am clear what my duties and responsibilities are
A2c	I understand how my work contributes to my agency's objectives
A2d	I have the authority (e.g. the necessary delegations, autonomy, level of responsibility) to do my job effectively
A2e	I am sufficiently challenged by my work
A2f	I am recognised for the contribution I make
A2g	I am satisfied with the opportunities available to me for career progression in my current agency
A2h	I am proud to work in the Western Australian public sector
A3a	I feel that my agency on the whole is well managed
A3b	Change is managed well in my agency
A3c	My agency's senior leaders provide effective leadership
A3d	My agency uses technological advances to improve service design and delivery to customers/clients
A3e	Recruitment and promotion decisions in my agency are fair
A3f	My workplace culture supports people to achieve a suitable work/life balance
A3g	You are able to access and use flexible work arrangements to assist in your work/life balance
A3h	My agency is committed to health and wellbeing within the workplace
A3i	I feel a strong personal attachment to my agency
A3j	My agency motivates me to help it achieve its objectives
A3k	My agency inspires me to do the best in my job
A3l	I am proud to tell others I work for my agency
A3m	I would recommend my agency as a great place to work

Question ID	Value	Label
A1a	1	Very satisfied
	2	Moderately satisfied
	3	Mildly satisfied
	4	Neither satisfied nor dissatisfied
	5	Mildly dissatisfied
	6	Moderately dissatisfied
	7	Very dissatisfied
A1b	1	Very satisfied
	2	Moderately satisfied
	3	Mildly satisfied
	4	Neither satisfied nor dissatisfied
	5	Mildly dissatisfied
	6	Moderately dissatisfied
	7	Very dissatisfied
A2a	1	Strongly agree
	2	Moderately agree
	3	Mildly agree
	4	Neither agree nor disagree
	5	Mildly disagree
	6	Moderately disagree
	7	Strongly disagree
	8	Do not know or does not apply
A2b	1	Strongly agree
	2	Moderately agree
	3	Mildly agree
	4	Neither agree nor disagree
	5	Mildly disagree
	6	Moderately disagree
	7	Strongly disagree
	8	Do not know or does not apply

Data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA
A1a	A1b	A2a	A2b	A2c	A2d	A2e	A2f	A2g	A2h	A3a	A3b	A3c	A3d	A3e	A3f	A3g	A3h	A3i	A3j	A3k	A3l	A3m	A4	A4ai	A4aii	A4aiii
2	5	1	1	1	2	1	1	3	1	6	7	6	5	4	3	3	4	1	3	3	3	4	3			
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	3			
6	6	3	3	3	4	2	4	4	4	7	5	7	4	5	4	2	5	7	7	6	6	7	3			
7	7	7	7	4	7	1	7	7		7	7	7	6	7	7	7	7	7	7	7	7	7	3			
7	7	5	7	5	7	1	5	7	6	6	6	6	4	4	4	3	4	6	5	6	7	7	1		1	
1	3	2	1	2	5	2	3	2	1	3	2	2	2	4	2	3	1	1	3	3	2	2	3			
1	1	2	2	1	1	1	1	1	1	2	2	1	2	2	1	1	2	1	1	1	1	1				
4	4	3	2	2	3	1	4	5	4	3	5	3	3	6	3	3	4	4	4	4	4	5	3			
1	1	1	1	1	2	2	1	2	1	1	3	1	3		2	8	3	1	1	2	1	1	3			
2	2	2	2	2	1	1	2	3	2	3	3	4	3	4	2	2	3	4	3	3	3	3	2	1	1	
1		2	1	1	1	1	1	5	1	2	2	4	5	5	3	4	3	3	2	2	3	4	3			
2	3	2	2	2	3	3	6	4	2	2	3	4	4	4	5	5	5	4	4	4	3	3	3			
7	7	3	6	6	3	5	5	6	4	6	6	5	5	8	7	7	7	6	5	4	4	6	1			

DATASET - Descriptive Statistics

105005 null values

Question Types:

Very Satisfied-Very dissatisfied

Strongly Agree - Strongly Disagree

Never - Very Frequently

Yes/No

Agency Size:

Employees at agencies with over 1000 employees make up **63%** of the responses

Variance:

Lowest Variance

Question C3: Are you familiar with your agency code of conduct?

0.0374

Highest Variance

Question C1g: Purchasing decisions in my workplace are not influenced by gifts or incentives

6.1458

Target Variable

Question B3b: My work group achieves a high level of productivity

Variance: 2.396

B3b	Count
1.0	1540
2.0	1184
3.0	520
4.0	237
5.0	145
6.0	94
7.0	80
8.0	45
NaN	36

Correlation Matrix

```
array([[ 1.        , -0.48917882, -0.25711912, ...,  0.00116056,
         0.05838622, -0.05579996],
       [-0.48917882,  1.        , -0.36891848, ..., -0.01911292,
        -0.03894502,  0.04695736],
       [-0.25711912, -0.36891848,  1.        , ...,  0.04238243,
        -0.00620979, -0.0166348 ],
       ...,
       [ 0.00116056, -0.01911292,  0.04238243, ...,  1.        ,
        -0.17459612, -0.36595502],
       [ 0.05838622, -0.03894502, -0.00620979, ..., -0.17459612,
         1.        , -0.85244382],
       [-0.05579996,  0.04695736, -0.0166348 , ..., -0.36595502,
        -0.85244382,  1.        ]])
```

Highest correlation value: .44

AgencySize	Count
1	239
2	1170
3	2472

DATA CLEANING

Target question - B3b: My work group achieves a high level of productivity.

1 = Strongly Agree, 2 = Moderately Agree, 3 = Mildly Agree, 4 = Neither agree nor disagree, 5 = Mildly disagree, 6 = Moderately disagree, 7 = Strongly disagree, 8 = Do not know or does not apply

DATA CLEANING/PREPARATION:

Dummy variables created for each question:

```
[13] questions.show()

+-----------+--------------------+
|Question ID|            Question |
+-----------+--------------------+
|        A1a|Please indicate y...|
|        A1b|Please indicate y...|
|        A2a|My job allows me ...|
|        A2b|I am clear what m...|
|        A2c|I understand how ...|
|        A2d|I have the author...|
|        A2e|I am sufficiently...|
|        A2f|I am recognised f...|
|        A2g|I am satisfied wi...|
|        A2h|I am proud to wor...|
|        A3a|I feel that my ag...|
|        A3b|Change is managed...|
|        A3c|My agency's senio...|
|        A3d|My agency uses  t...|
|        A3e|Recruitment and p...|
|        A3f|My workplace cult...|
|        A3g|You are able to a...|
|        A3h|My agency is comm...|
|        A3i|I feel a strong p...|
|        A3j|My agency motivat...|
+-----------+--------------------+
only showing top 20 rows
```

```
[18] data.show()

+-------+-------+-------+-------+-------+-------+-------+--
|A1a_1.0|A1a_2.0|A1a_3.0|A1a_4.0|A1a_5.0|A1a_6.0|A1a_7.0|A1
+-------+-------+-------+-------+-------+-------+-------+--
|      0|      1|      0|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|
|      0|      0|      0|      0|      0|      1|      0|
|      0|      0|      0|      0|      0|      0|      1|
|      0|      0|      0|      0|      0|      0|      1|
|      1|      0|      0|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|
|      0|      0|      0|      1|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|
|      0|      1|      0|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|
|      0|      1|      0|      0|      0|      0|      0|
|      0|      0|      0|      0|      0|      0|      1|
|      0|      1|      0|      0|      0|      0|      0|
|      1|      0|      0|      0|      0|      0|      0|
|      0|      1|      0|      0|      0|      0|      0|
|      0|      0|      0|      0|      0|      0|      1|
|      0|      0|      0|      1|      0|      0|      0|
|      0|      0|      1|      0|      0|      0|      0|
|      0|      1|      0|      0|      0|      0|      0|
+-------+-------+-------+-------+-------+-------+-------+--
only showing top 20 rows
```

```
pipeline.transform(questionkey).select('Question ID','Question ','words', 'filtered','tf').show()

+-----------+--------------------+--------------------+--------------------+--------------------+
|Question ID|            Question |               words|            filtered|                  tf|
+-----------+--------------------+--------------------+--------------------+--------------------+
|        A1a|Please indicate y...|[please, indicate...|[indicate, level,...|(379,[27,35,76,11...|
|        A1b|Please indicate y...|[please, indicate...|[indicate, level,...|(379,[0,35,76,119...|
|        A2a|My job allows me ...|[my, job, allows,...|[job, allows, uti...|(379,[27,42,77,25...|
|        A2b|I am clear what m...|[am, clear, what,...|[clear, duties, r...|(379,[82,96,239],...|
|        A2c|I understand how ...|[understand, how,...|[understand, work...|(379,[0,5,78,273,...|
|        A2d|I have the author...|[have, the, autho...|[authority, neces...|(379,[27,35,48,10...|
|        A2e|I am sufficiently...|[am, sufficiently...|[sufficiently, ch...|(379,[5,288,351],...|
|        A2f|I am recognised f...|[am, recognised, ...|[recognised, cont...|(379,[63,163,333]...|
|        A2g|I am satisfied wi...|[am, satisfied, w...|[satisfied, oppor...|(379,[0,45,46,60,...|
|        A2h|I am proud to wor...|[am, proud, to, w...|[proud, work, wes...|(379,[5,14,15,97,...|
|        A3a|I feel that my ag...|[feel, that, my, ...|[feel, agency, ma...|(379,[0,59,112],[...|
|        A3b|Change is managed...|[change, is, mana...|[change, managed,...|(379,[0,59,258],[...|
|        A3c|My agency�s senio...|[my, agency, seni...|[agency, senior, ...|(379,[0,43,57,134...|
|        A3d|My agency uses  t...|[my, agency, uses...|[agency, uses, te...|(379,[0,88,161,17...|
|        A3e|Recruitment and p...|[recruitment, and...|[recruitment, pro...|(379,[0,54,164,22...|
|        A3f|My workplace cult...|[my, workplace, c...|[workplace, cultu...|(379,[2,5,34,70,8...|
|        A3g|You are able to a...|[you, are, able, ...|[able, access, us...|(379,[5,20,58,83,...|
|        A3h|My agency is comm...|[my, agency, is, ...|[agency, committe...|(379,[0,2,53,217,...|
|        A3i|I feel a strong p...|[feel, strong, pe...|[feel, strong, pe...|(379,[0,98,112,16...|
|        A3j|My agency motivat...|[my, agency, moti...|[agency, motivate...|(379,[0,67,70,78,...|
+-----------+--------------------+--------------------+--------------------+--------------------+
```

Model: LDA

Purpose: LDA was performed to gage the number of topics within the question list.

Evaluation Metric: Lacks golden standard

```
array([['behaviour', 'inappropriate', 'workplace', 'information',
        'unethical', 'witnessed', 'types', 'disclosure', 'months', 'use'],
       ['performance', 'recognised', 'contribution', 'tax', 'helped',
        'months', 'gross', 'total', 'make', 'enables'],
       ['decisions', 'repeated', 'subjected', 'past', 'recruitment',
        'adequately', 'fair', 'promotion', 'input', 'directly'],
       ['level', 'honesty', 'satisfaction', 'employer', 'indicate',
        'integrity', 'immediate', 'workers', 'highest', 'formal'],
       ['leave', 'describes', 'experienced', 'nature', 'bullying', 'yes',
        'intend', 'deliberately', 'changing', 'rumours'],
       ['public', 'sector', 'decision', 'standard', 'believe', 'did',
        'committed', 'code', 'past', 'ethics'],
       ['headcount', 'size', 'topics', 'discussed', 'meetings',
        'documented', 'managed', 'formal', 'activities', 'help'],
       ['achieve', 'suitable', 'life', 'supports', 'balance', 'personal',
        'culture', 'people', 'work', 'strong'],
       ['manager', 'witnessed', 'diversity', 'career', 'learning',
        'detriment', 'receive', 'people', 'pranks', 'rumours'],
       ['senior', 'effective', 'managers', 'provide', 'leadership',
        'leaders', 'area', 'work', 'communication', 'employees']],
      dtype='<U16')
```

Topic 1: Ethical Behaviors of employee/senior management
Topic 2: Recognition of achievement
Topic 3: Hiring decisions
Topic 4: Integrity and Satisfaction of employee
Topic 5: Bullying at the workplace
Topic 6: Code of Ethics
Topic 7:
Topic 8: Work-life balance support
Topic 9: Diversity in management
Topic 10: Senior management leadership and communication.

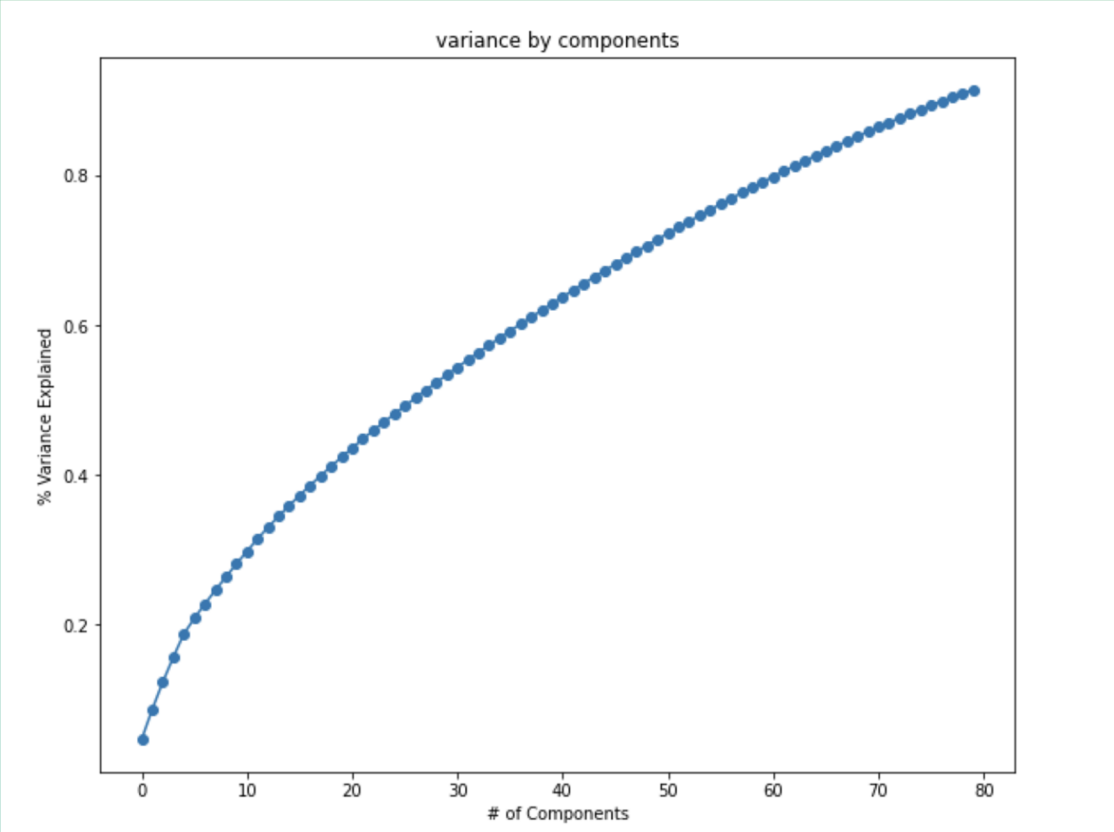
```
In [42]: lda.transform(questionkey).where(fn.col('Question ID') == 'B3b').first()['lda_feat']
         #Topic 4

Out[42]: DenseVector([0.032, 0.0287, 0.0279, 0.7381, 0.031, 0.0289, 0.0287, 0.0281, 0.0274, 0.0291])
```

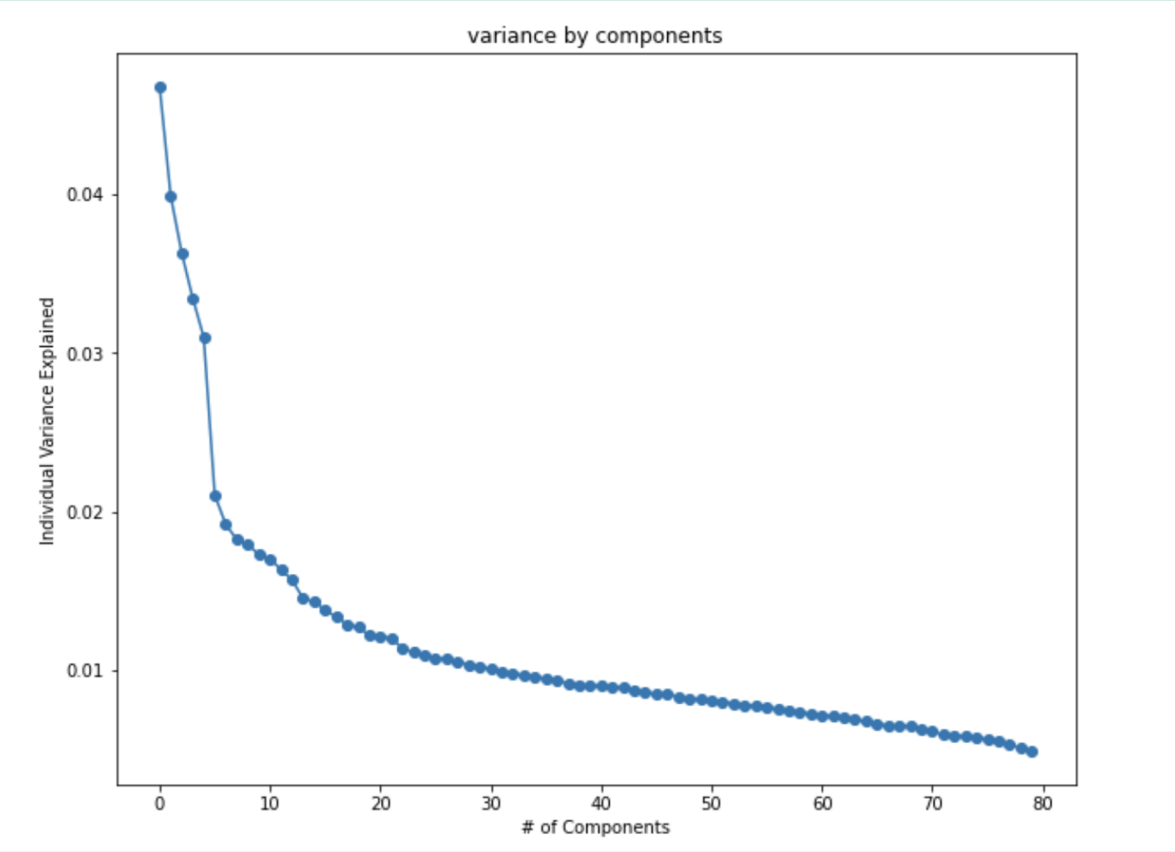
Question B3b belonging to topic 4

Model: PCA

Purpose: PCA produces a low dimensional representation of the original data with maximum variance.

Evaluation Metric: Variance Explained ratio shows how the variance is explained by each principal component.

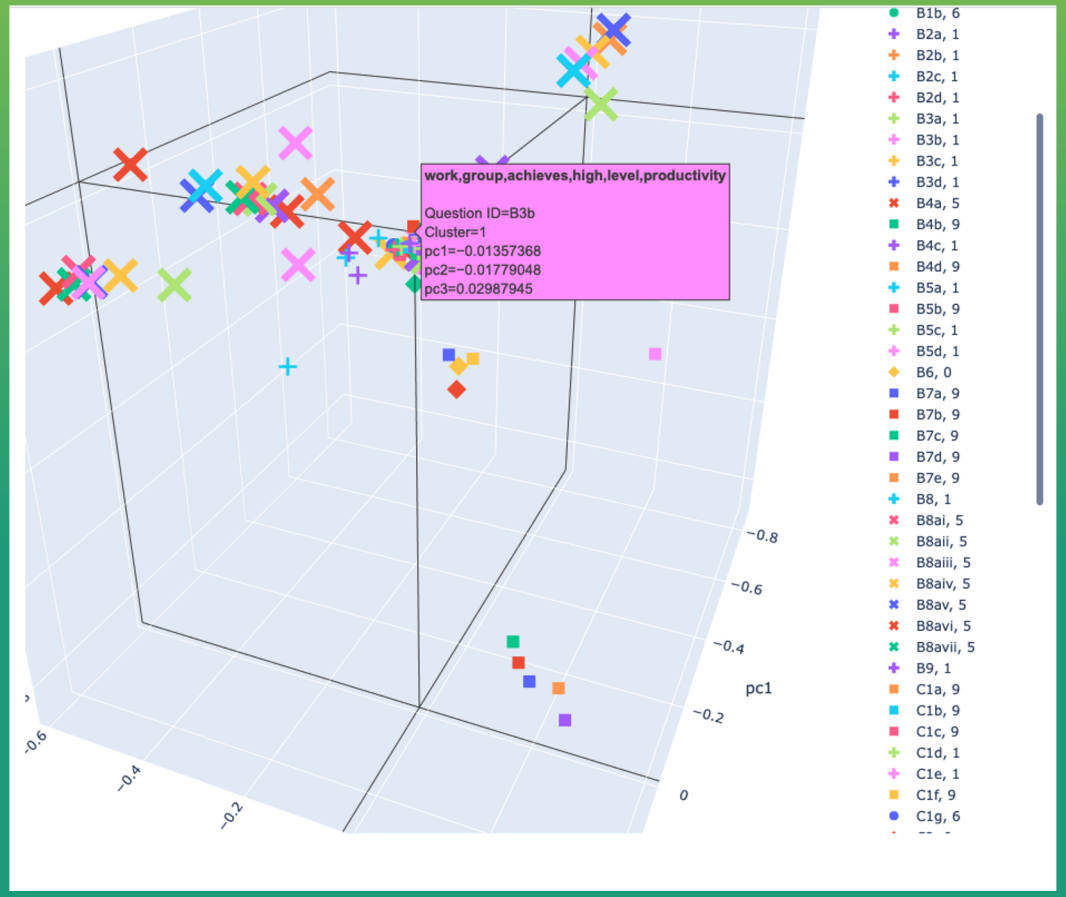




Model: K-MEANS

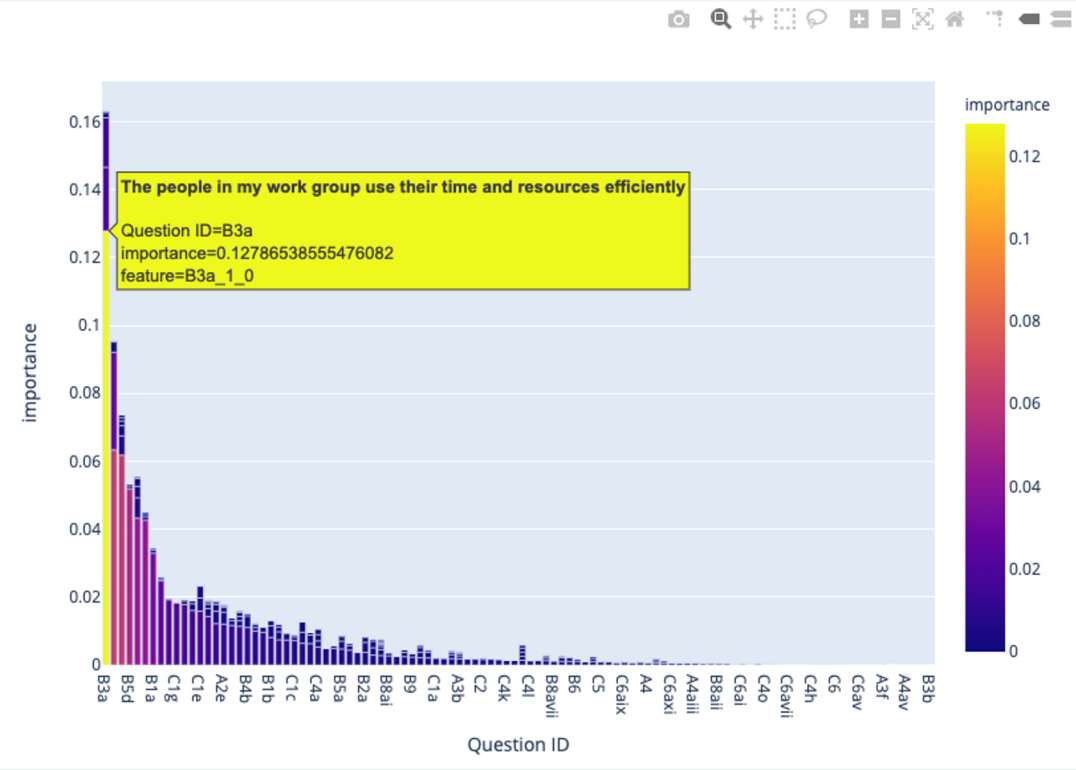
Using three principal components to plot the clusters maximizes variances separating the data.

Question ID	Question
A2c	I understand how my work contributes to my agency's objectives
A2e	I am sufficiently challenged by my work
A3a	I feel that my agency on the whole is well managed
A3b	Change is managed well in my agency
A3d	My agency uses technological advances to improve service design and delivery to customers/clients
A3e	Recruitment and promotion decisions in my agency are fair
A3g	You are able to access and use flexible work arrangements to assist in your work/life balance
A3i	I feel a strong personal attachment to my agency
A3j	My agency motivates me to help it achieve its objectives
A3l	I am proud to tell others I work for my agency
A3m	I would recommend my agency as a great place to work



https://chart-studio.plotly.com/~courtneyhrdy/1/#/

Model: RANDOM FOREST



AUC	Precision	Recall	F1-Score
0.81	0.88	0.67	0.76

https://chart-studio.plotly.com/~courtneyhrdy/4/#/

The feature importances are based on the averaged decrease in impurity over trees.

Question ID	Question
B3a	The people in my work group use their time and resources efficiently
B3c	In the last 12 months, my work group has implemented innovative processes or policies
B3d	The people in my work group are committed to providing excellent customer service and making a positive difference to the community
C1f	Confidential information in my workplace is only disclosed to appropriate people

Model: LOGISTIC REGRESSION

All Features:

Intercept:

-0.46390297661691365

AUC:

0.8147028381543826

	column	weight
155	A3j_8_0	443.586261
493	C4n_4_0	409.736187
163	A3k_8_0	244.907480
245	B3a_1_0	241.454263
338	B7b_1_0	225.325021
...	...	...
478	C4k_4_0	-353.476195
484	C4l_5_0	-363.917043
35	A2c_8_0	-389.365314
99	A3c_8_0	-468.207263
51	A2e_8_0	-979.869672

Top 5 features determined by Random Forest:

A3l_1_0: I am proud to tell others I work for my agency
A2a_1_0: My job allows me to use my skills, knowledge and abilities
A3a_1_0: I feel my agency is well managed
A3k_1_0: My agency inspires me to do my best
A1b_1_0: Indicate level of satisfaction: agency as an employer

Intercept:

-2.735958523998674

AUC:

0.9230811134717003

	column	weight
0	A1a_1_0	2.948996
1	A1a_2_0	1.577564
2	A1a_3_0	1.468410
4	A1a_5_0	0.532123
3	A1a_4_0	0.359440

All features on target variable B3b_7 (Strongly Disagree):

Intercept:

-4.096793222030586

AUC:

0.47137331808029626

	column	weight
459	C4g_5_0	19.017544
268	B3d_7_0	17.251142
458	C4g_4_0	16.445949
499	C4o_5_0	15.614347
251	B3a_7_0	12.329670
...	...	...
454	C4f_5_0	-7.035899
483	C4l_4_0	-7.472870
75	A2h_8_0	-8.371114
439	C4c_5_0	-10.695234
51	A2e_8_0	-17.039285

552 rows × 2 columns

CONCLUSION

The questions identified in the unsupervised learning algorithms are consistent with the feature importance we got from supervised learning

Employee's perception of their workplace is most predictive of productivity. Work where you love!

Highest Performing Models:
Random Forest,
Logistic Regression on 5 features

IST 659: Database Administration and Concepts

Credit Card Portfolio Analysis for Community Banks
PowerApp

IST 659: Credit Card Portfolio Analysis

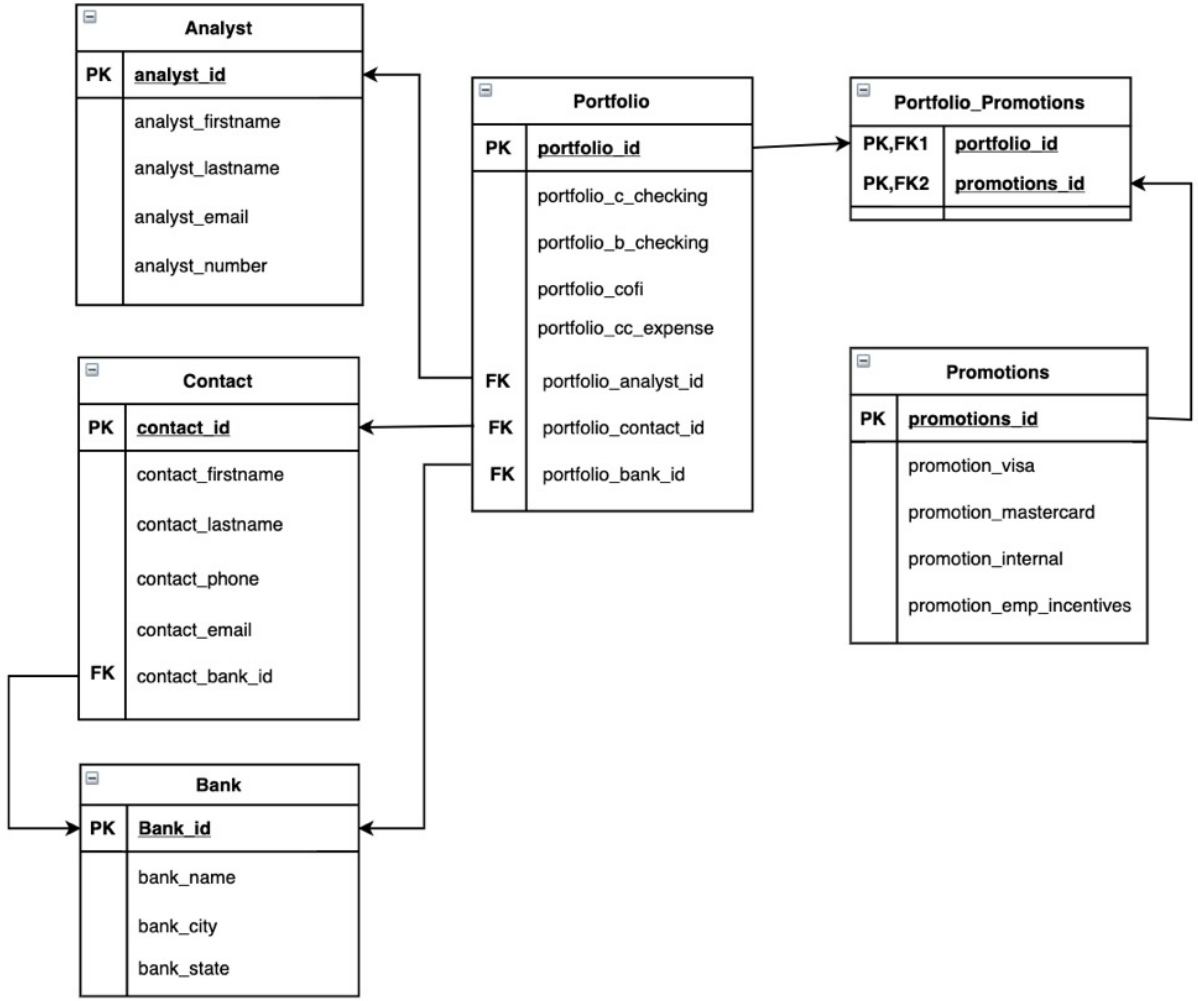
This project objective is to transform the Independent Community of Bankers Association (ICBA) Automated Credit Expert (ACE) form to enhance customer service and improve data accessibility for analysis.

PowerApps was used to improve accessibility to assessments and enhance dataflow to staff experts. This form would be used by the staff experts in initial meetings with banks seeking membership.

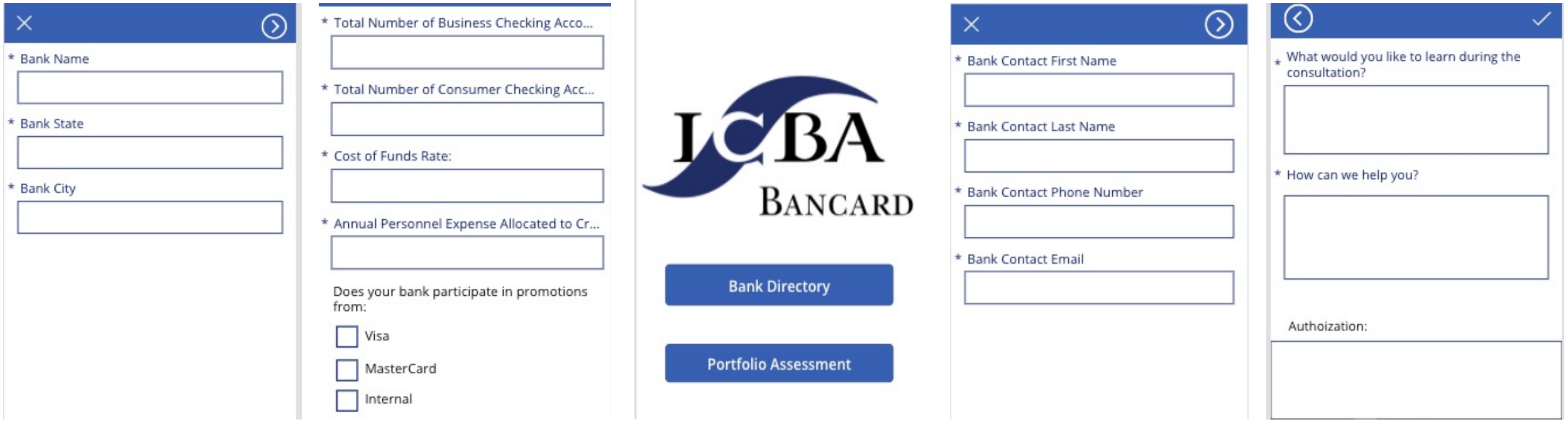
The Data Requirements

Attribute	Description
Bank Name	Community bank name
City	City of branch location
State	State of branch location
Contact Person	First and last name of community bank representative
Phone Number	Phone number of community bank representative
Email	Email of community bank representative
Total Number of Consumer Checking Accounts	Number of consumer checking accounts with the community bank
Total Number of Business Checking Accounts	Number of business checking accounts with the community bank
Cost of Funds Rate	The weighted average of interest rates that banks pay on savings accounts held by their customers and money borrowed from other institutions. That percentage is then divided by 12 and multiplied by the outstandings to show an annualized amount. (Cost of Funds % / 12) x Outstandings = Annualized Cost of Funds in $
Annual Personal Expense Related to Allocated Credit Cards	The expenses entered by your bank are allocated equally by the number of card plans in your portfolio, except for the cost of funds figure.
Promotion participation	Does your bank participate in promotions from Visa, MasterCard, Internal or Employee Incentives

The Logical Data Model



The PowerApp



```
    from icba_bank_portfolios p
    join icba_banks b on p.portfolio_bank_id = b.bank_id

GO
select * from p_icba_portfolio_bank

GO
create table icba_promotions
(
    promotion_id int not null,
    promotion_option_visa varchar(20) null,
    promotion_option_mastercard varchar(20) null,
    promotion_option_employee_benefits varchar(20) null,
    promotion_option_internal varchar(20) null,
    constraint pk_icba_promotions_promotion_id primary key (promotion_id)
)
GO

create table portfolio_promotions
(
    portfolio_id int not null,
    promotion_id int not null,
    constraint pk_portfolios_portfolio_promotion_id primary key (portfolio_id, promotion_id)

)
GO
```

```
GO
create table icba_analysts
(
    analyst_id int not null,
    analyst_firstname varchar(50) not null,
    analyst_lastname varchar(50) not null,
    analyst_email varchar(50) not null,
    analyst_number varchar(10) not null
    constraint pk_icba_analysts_analyst_id primary key (analyst_id)
)
GO
create table icba_bank_portfolios
(
    portfolio_id int identity not null,
    portfolio_bank_id int not null,
    portfolio_assigned_bank_id int null,
    portfolio_total_number_consumer_checking_acct int not null,
    portfolio_total_number_business_checking_acct int not null,
    portfolio_cofi int not null,
    portfolio_cc_expense int not null,
    portfolio_q1 varchar(max) not null,
    portfolio_q2 varchar(max) not null,
    promotion_option_visa bit,
    promotion_option_mastercard bit,
    promotion_option_employee_benefits bit,
    promotion_option_internal bit,
    constraint pk_icba_portfolios_portfolio_id primary key (portfolio_i
    constraint fk_icba_portfolios_portfolio_bank_id foreign key (portfo
    CONSTRAINT fk_icba_portfolios_portfolio_assigned_bank_id foreign ke
)
GO
```

```
    [bank_state] [varchar](2) NOT NULL,
    [bank_contact_firstname] [varchar](50) NULL,
    [bank_contact_lastname] [varchar](50) NULL,
    [bank_contact_email] [varchar](50) NULL,
    [bank_contact_number] [varchar](50) NULL,
 CONSTRAINT [pk_icba_banks_bank_id] PRIMARY KEY ([bank_id] )
)
GO
create table icba_analysts
(
    analyst_id int not null,
    analyst_firstname varchar(50) not null,
    analyst_lastname varchar(50) not null,
    analyst_email varchar(50) not null,
    analyst_number varchar(10) not null
    constraint pk_icba_analysts_analyst_id primary key (analyst_id)
)
GO
create table icba_bank_portfolios
(
```

```
(
    portfolio_id int not null,
    promotion_id int not null,
    constraint pk_portfolios_portfolio_promotion_id primary key (portfolio_id, promotion_id)

)
GO

alter table portfolio_promotions
    add constraint fk_portfolios_portfolio_id foreign key (portfolio_id)
    references icba_bank_portfolios(portfolio_id)

alter table portfolio_promotions
    add constraint fk_promotions_promotion_id foreign key (promotion_id)
    references icba_promotions (promotion_id)


INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (1, 'Union Bank', 'Washington', 'DC')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (2, 'Georgetown Bank', 'Washington', 'DC')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (3, 'Spring Valley Bank', 'Bethesda', 'MD')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (4, 'National Capital Bank', 'Rockville', 'MD')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (5, 'NOVA Bank', 'Arlington', 'VA')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (6, 'Great Falls Bank', 'McLean', 'VA')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (7, 'Anacostia Bank', 'Washington', 'DC')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (8, 'Capitol Hill Bank', 'Washington', 'DC')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (9, 'Montgomery Bank', 'Bethesda', 'MD')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (10, 'Wheaton Bank', 'Rockville', 'MD')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (11, 'Pentagon City Bank', 'Arlington', 'VA')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (12, 'Langley Bank', 'McLean', 'VA')
INSERT [dbo].[icba_banks] ([bank_id], [bank_name], [bank_city], [bank_state]) VALUES (13, 'CCS Bank', 'McLean', 'VA')
```

The Database

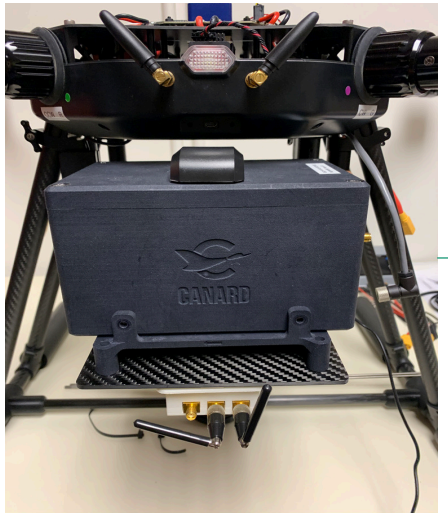
IST 974: Internship in Data Science

Syracuse University Center of Advanced Systems and Engineering (CASE)

Unmanned Aerial System (UAS) Instrument Landing System (ILS) Calibration

- Instrument Landing System (ILS) is a radio navigation system that guides the landing of planes.
- ILS inspection has always been done by manned aircraft, which is difficult for pilots and expensive for airports.
- Thales is developing a UAS-based ILS inspection solution to complement manned flight checks. This is the first in the United States.

The Data



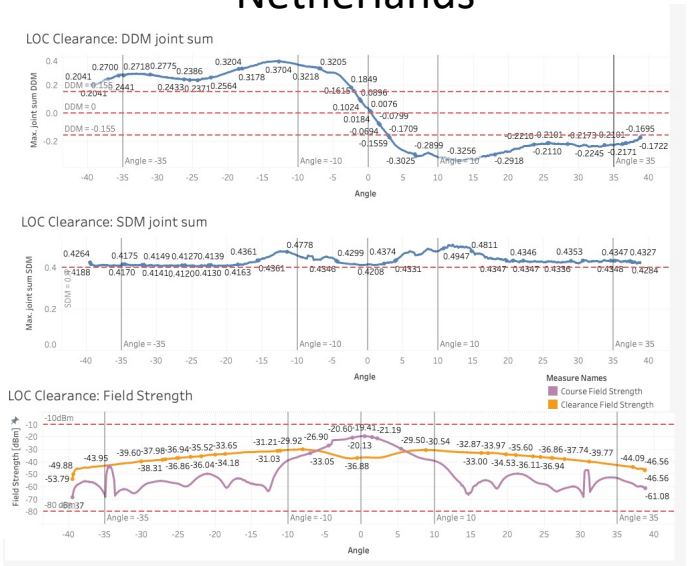
hmsl	233.68	233.68	233.68	233.68	233.68	233.68	233.68
latitude	43.2146	43.2146	43.2146	43.2146	43.2146	43.2146	43.2146
longitude	-75.3799	-75.3799	-75.3799	-75.3799	-75.3799	-75.3799	-75.3799
posFlag	1	1	1	1	1	1	1
sat_count	6	6	6	6	6	6	6
time	0	0	0	0	0	0	0
battery	0	0	0	0	0	0	0
clearance_DDM	-0.0348777	-0.0368207	-0.0399121	-0.040922	-0.0390534	-0.0394584	-0.04547
clearance_DDM_ua	-33.7527	-35.633	-38.6247	-39.6019	-37.7936	-38.1855	-44.0032
clearance_SDM	0.653301	0.653962	0.6568	0.656595	0.653988	0.654712	0.659889
clearance_f150	149.578	149.578	149.578	149.578	149.578	149.578	149.578
clearance_f90	81.0135	81.0135	81.0135	81.0135	81.0135	81.0135	81.0135
clearance_field_strength	-81.392	-81.3811	-81.394	-81.3893	-81.3777	-81.3809	-81.3965
clearance_freq_offset	-4043.95	-4043.95	-4043.95	-4043.95	-4043.95	-4043.95	-4043.95
clearance_m150	0.344089	0.345391	0.348356	0.348758	0.346521	0.347085	0.35268
clearance_m90	0.309212	0.30857	0.308444	0.307836	0.307467	0.307627	0.30721
course_DDM	-0.000566142	-0.000936882	-0.00058847	-0.000436939	-0.000594615	-0.000235735	0.000477056
course_DDM_ua	-0.54788	-0.90666	-0.569487	-0.422844	-0.575434	-0.228131	0.461667
course_SDM	0.409976	0.40918	0.407889	0.407094	0.406621	0.406283	0.406225
course_f150	149.979	149.979	149.979	149.979	149.979	149.979	149.979
course_f90	90.5085	90.5085	90.5085	90.5085	90.5085	90.5085	90.5085
course_field_strength	-62.1839	-62.1824	-62.191	-62.1992	-62.2015	-62.2036	-62.2086
course_freq_offset	3957.03	3957.03	3957.03	3957.03	3957.03	3957.03	3957.03
course_m150	0.205271	0.205058	0.204239	0.203766	0.203608	0.203259	0.202874
course_m90	0.204705	0.204121	0.20365	0.203329	0.203013	0.203023	0.203351
dual_frequency	1	1	1	1	1	1	1
joint_freq_separation	8000.98	8000.98	8000.98	8000.98	8000.98	8000.98	8000.98
joint_sum_DDM	-0.000772019	-0.00115266	-0.000824698	-0.000680862	-0.000827069	-0.000472749	0.000200083

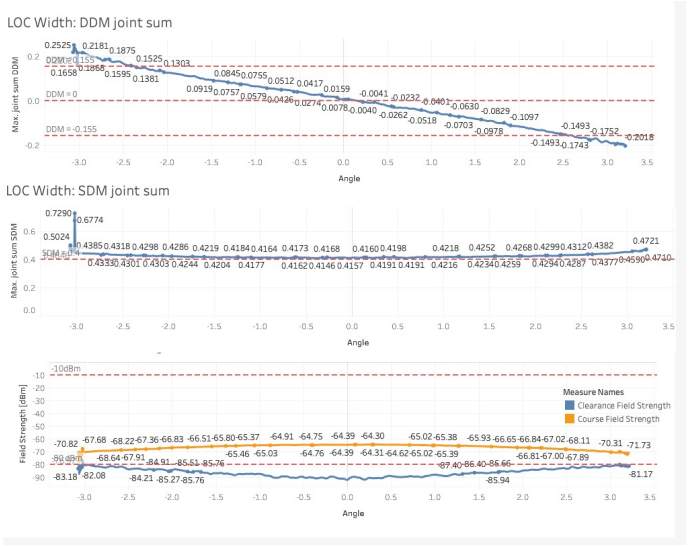
ILS Receiver	Description
date	date as transmitted to the receiver (from the flight controller)
hmsl	Height above mean sea level (from the flight controller)
lat	Latitude of the drone transmitted to the receiver (from the flight controller)
long	Longitude of the drone transmitted to the receiver (from the flight controller)
posFlag	Indicator of the GPS mode according to NMEA protocol
sat_count	Number of satellites (from the flight controller)
time	time as transmitted to the receiver (from the flight controller)
***_DDM	Difference in Depth of Modulation for each signal
***_DDM_ua	Difference in Depth of Modulation for each signal in microamperes
***_f150	Frequency of the 150Hz component for each signal
***_f90	Frequency of the 90Hz component for each signal
***_m150	Modulation of the 150Hz component for each signal
***_m90	Modulation of the 90Hz component for each signal
***_field_strength	Field strength for each signal in dBm
***_freq_offset	Frequency offset of each signal with respect to the center of the channel (or channel frequency)
***_SDM	Sum of Depth of Modulation for each signal
angle (LOC)	horizontal angle of the drone relative to the localizer signal
angle (GP)	vertical angle of the drone relative to the ground
distance	distance calculated using the threshold coordinates and the coordinates of the flight
timestamp	The time recorded in Unix
datetime	unix converted to hh.mm.ss.000

ILS Calibration: ICAO vs. FAA

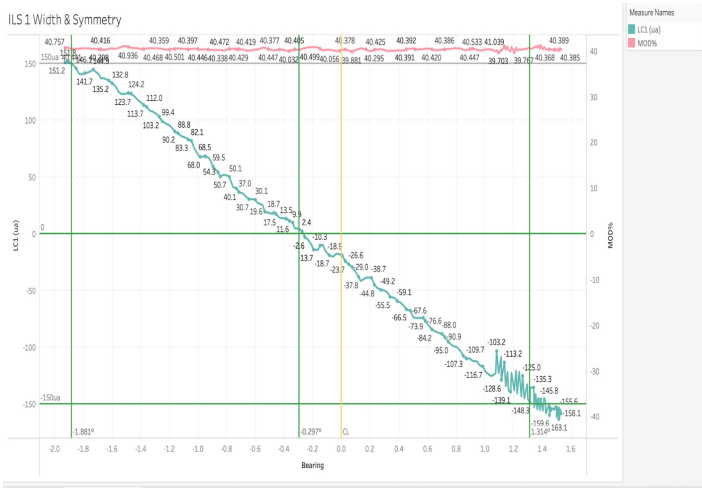
- Data modeled in the European Aviation Safety Agency (EASA) format using data from Groningen Airport Eelde, Netherlands.
- Data analysis in a format compatible to FAA ILS calibration operations using data collected from Griffiss International Airport, New York.

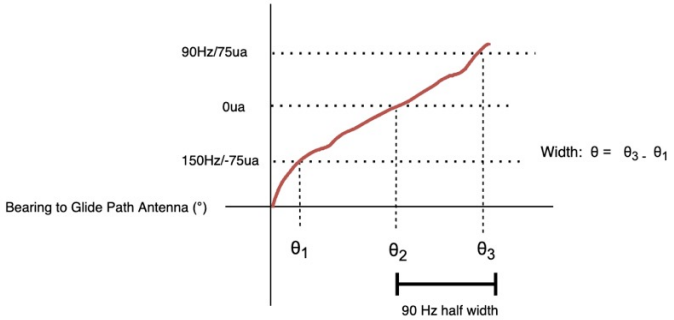
Groningen Airport Eelde, Netherlands



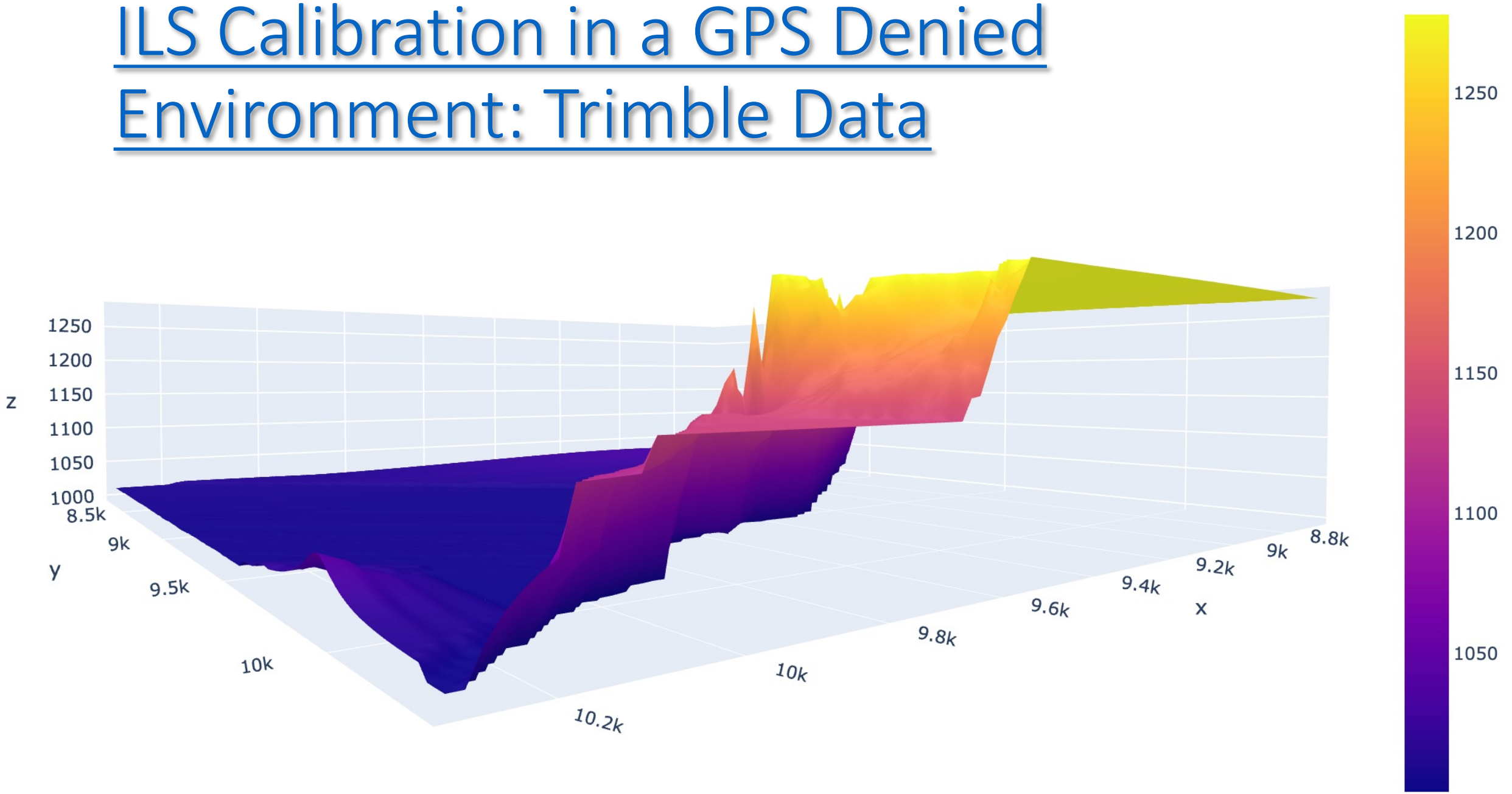


Griffiss UAS Testing Site, New York





ILS Calibration in a GPS Denied Environment: Trimble Data



My Role as a Data Science Intern

- Conduct geospatial analysis for UAS ILS calibration in a GPS-denied environment.
- Calculate pre-flight spatial requirements in Excel for flight tests at Griffiss Airport.
- Collect, store, and process ILS sensor data directly from the UAS.
- Validate flight path coordinates in Python.
- Present visualizations made in Tableau for meetings with the FAA and CANARD Drones.
- Model, validate, and deploy data in a format compatible to the UAS ILS calibrations done previously in Europe.
- Collect, analyze, and visualize data in a format compatible to FAA ILS calibration operations.