Data Science and Business Analytics

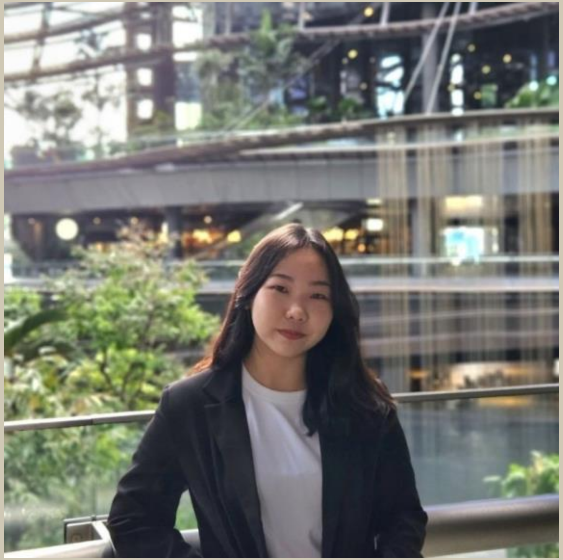
Portfolio

By Elizabeth Leonny Efendi

About Me

Elizabeth Leonny Efendi

Enthusiastic and driven Data Science and Business Analytics student from Indonesia, eager to apply analytical skills in real-world settings. Passionate about solving complex problems, contributing to dynamic teams, and continuously expanding my knowledge through hands-on experience.



Business Analysis: Canadian Market

2.1 Canada Market Overview

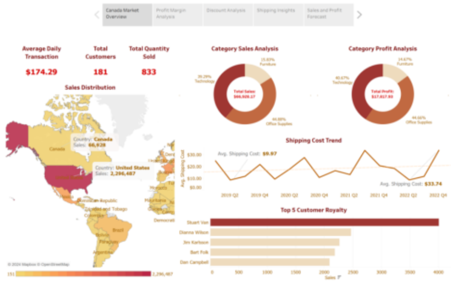


Figure 1: Canada Market Overview

2.2 Profit Margin Analysis

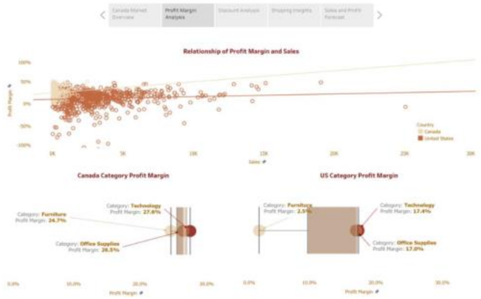

Relationship of Profit Margin and Sales
Canada Category Profit Margin
US Category Profit Margin

Figure 2: Canada and the US Profit Margin Analysis

2.3 Discount Analysis

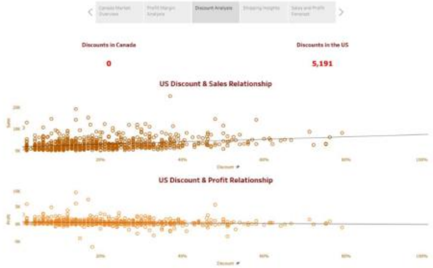

Discounts in Canada
0
Discounts in the US
5,191
US Discount & Sales Relationship
US Discount & Profit Relationship

Figure 3: Sales, Discount and Profit Relationship

2.5 Forecasts

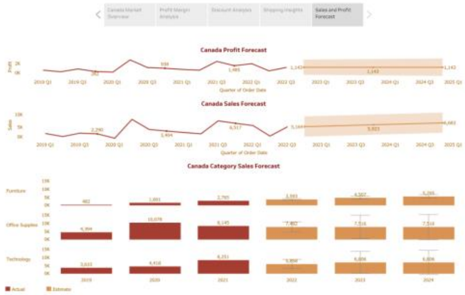


Figure 6: Sales and Profit Forecasts

2.4 Shipping Insights

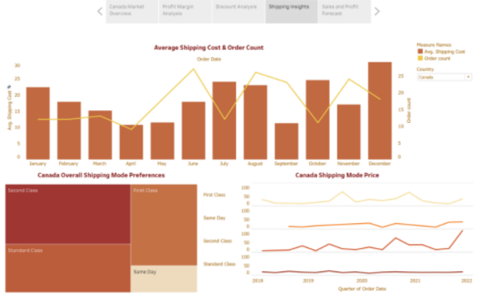

Average Shipping Cost & Order Count
Canada Overall Shipping Mode Preferences
Canada Shipping Mode Price
Second Class
First Class
Standard Class
Same Day

Figure 4: Shipping Insights 1 (Canada)

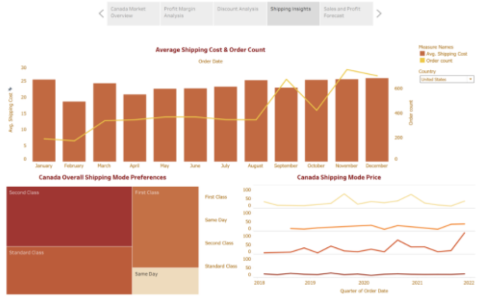

Average Shipping Cost & Order Count
Canada Overall Shipping Mode Preferences
Canada Shipping Mode Price
Second Class
First Class
Standard Class
Same Day

Figure 5: Shipping Insights 2 (US)

Full report is available at https://github.com/gezie1/Portfolio

Market Research: Marriott International

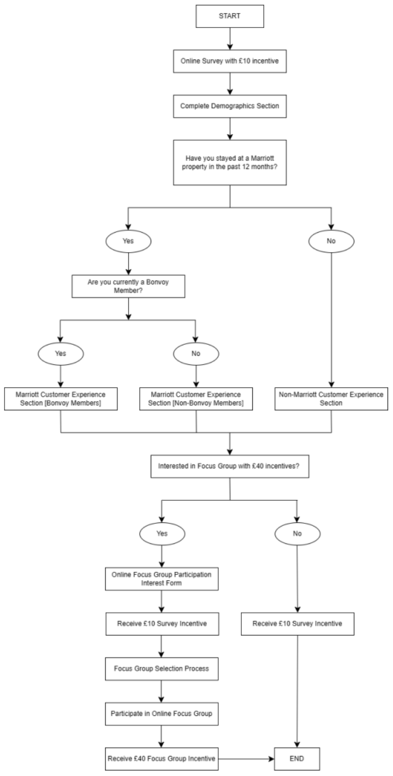
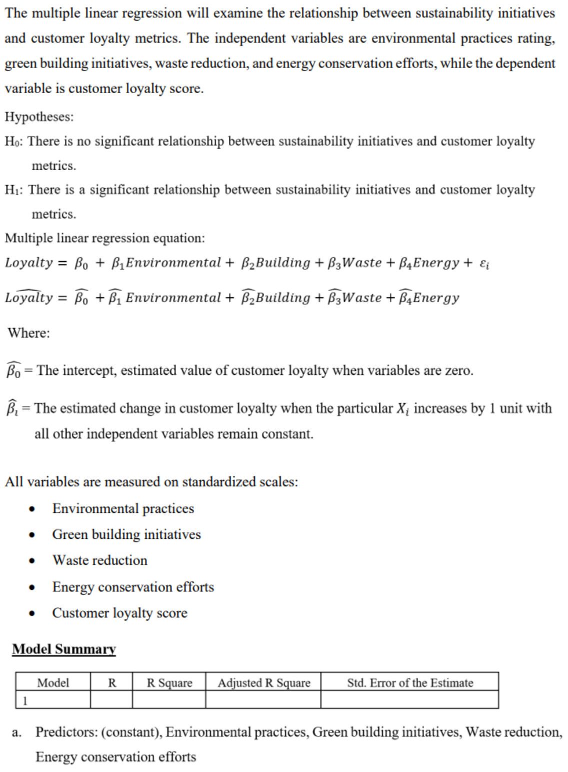

START
Online Survey with £10 incentive
Complete Demographics Section
Have you stayed at a Marriott property in the past 12 months?
Yes
No
Are you currently a Bonvoy Member?
Yes
No
Marriott Customer Experience Section [Bonvoy Members]
Marriott Customer Experience Section [Non-Bonvoy Members]
Non-Marriott Customer Experience Section
Interested in Focus Group with £40 incentives?
Yes
No
Online Focus Group Participation Interest Form
Receive £10 Survey Incentive
Focus Group Selection Process
Participate in Online Focus Group
Receive £40 Focus Group Incentive
Receive £10 Survey Incentive
END

Figure 1: Flow Chart



The multiple linear regression will examine the relationship between sustainability initiatives and customer loyalty metrics. The independent variables are environmental practices rating, green building initiatives, waste reduction, and energy conservation efforts, while the dependent variable is customer loyalty score.

Hypotheses:

H₀: There is no significant relationship between sustainability initiatives and customer loyalty metrics.

H₁: There is a significant relationship between sustainability initiatives and customer loyalty metrics.

Multiple linear regression equation:

$$Loyalty = \beta_0 + \beta_1 Environmental + \beta_2 Building + \beta_3 Waste + \beta_4 Energy + \varepsilon_i$$

$$\widehat{Loyalty} = \widehat{\beta_0} + \widehat{\beta_1}\, Environmental + \widehat{\beta_2} Building + \widehat{\beta_3} Waste + \widehat{\beta_4} Energy$$

Where:

$\widehat{\beta_0}$ = The intercept, estimated value of customer loyalty when variables are zero.

$\widehat{\beta_i}$ = The estimated change in customer loyalty when the particular X_i increases by 1 unit with all other independent variables remain constant.

All variables are measured on standardized scales:

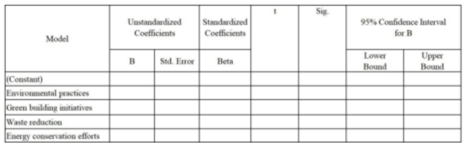
- Environmental practices
- Green building initiatives
- Waste reduction
- Energy conservation efforts
- Customer loyalty score

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1				

a. Predictors: (constant), Environmental practices, Green building initiatives, Waste reduction, Energy conservation efforts

Figure 8: Model Summary



Coefficients Table

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)							
Environmental practices							
Green building initiatives							
Waste reduction							
Energy conservation efforts							

Figure 9: Coefficients Table

We will employ t-tests and F-tests using SPSS to examine both the individual significance of variables and the overall significance of the model. This analysis will help us determine whether sustainability initiatives have a meaningful influence on customer loyalty. The adjusted R-squared value will be used to assess the strength of the relationship in the multiple linear regression model. A higher adjusted R-squared value would indicate a better model fit.

5. Timeline

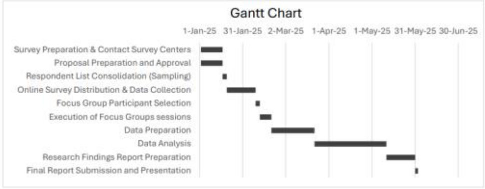


Figure 17: Gantt Chart

6. Budget

Description	Cost (£)
Incentives for Online Survey respondents	355,000
Incentives for Focus Group sessions + Moderator	8,000
Data Collection and Preparation	350,000
Data Analysis	200,000
Administrative and Operational Costs	100,000
Miscellaneous Fees	50,000
Total Cost	**1,063,000**

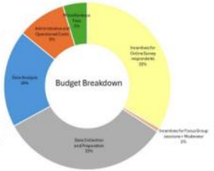
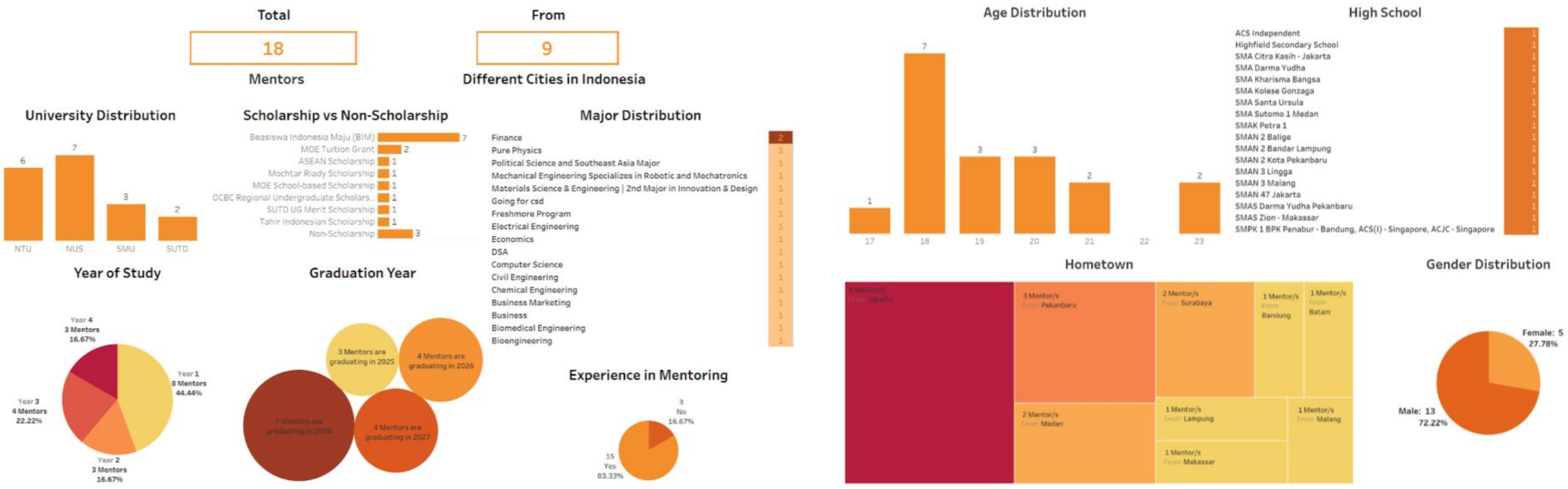
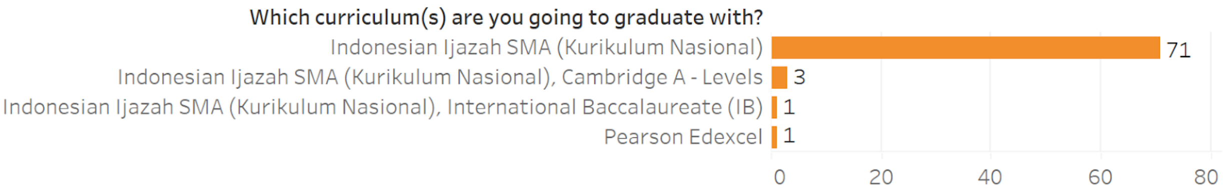


Figure 18: Budget Breakdown

Full report is available at https://github.com/gezie1/Portfolio

Analysis: Sentre Mentors and Mentees





Full report is available at https://github.com/gezie1/Portfolio

Stock Price Prediction: Meta

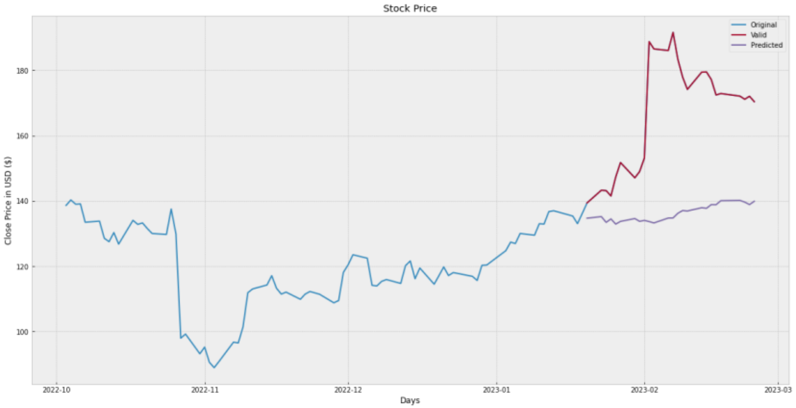
Language:

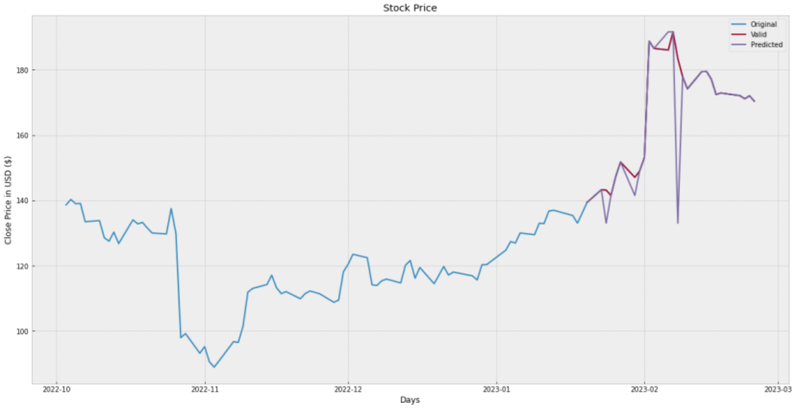
Data:

- <u>Sources: </u>Yahoo Finance
- <u>Features:</u> Moving averages, trading volume, financial ratios

Methodology:

- <u>Data Preprocessing:</u> Handled missing values, feature engineering
- <u>Models:</u> Linear Regression, Decision Tree





Flight Analysis

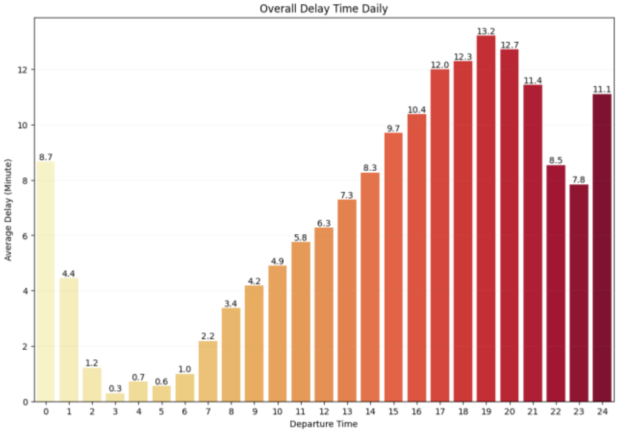
Language:

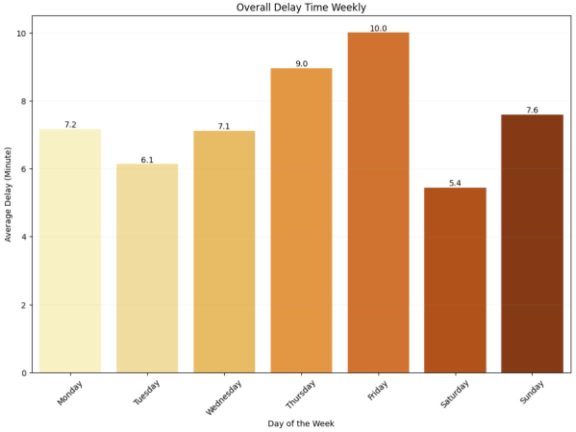
Data:

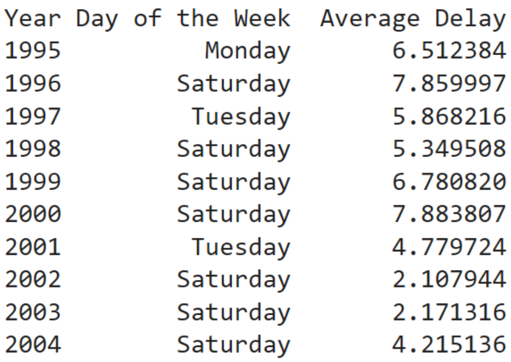
- Source: The 2009 ASA Statistical Computing and Graphics Data Expo

Methodology:

- Data Collection
- Data Cleaning
- Data Transformation
- Statistical Analysis
- Logistic Regression







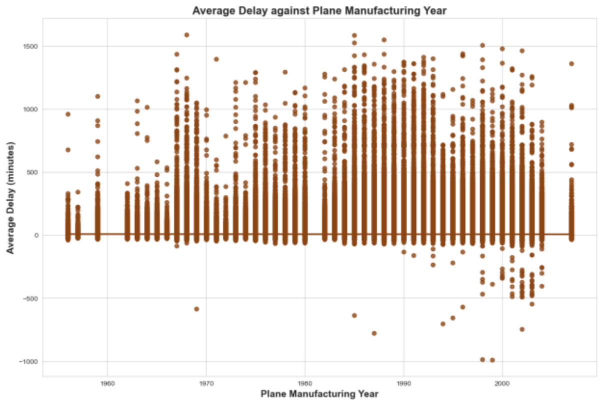
```
Year Day of the Week  Average Delay
1995          Monday       6.512384
1996        Saturday       7.859997
1997         Tuesday       5.868216
1998        Saturday       5.349508
1999        Saturday       6.780820
2000        Saturday       7.883807
2001         Tuesday       4.779724
2002        Saturday       2.107944
2003        Saturday       2.171316
2004        Saturday       4.215136
```

Flight Analysis

Language:

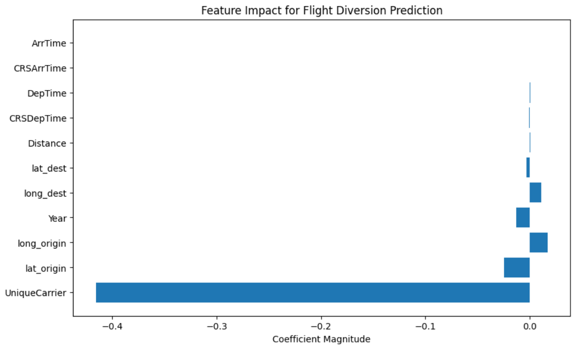
Data:

- Source: The 2009 ASA Statistical Computing and Graphics Data Expo

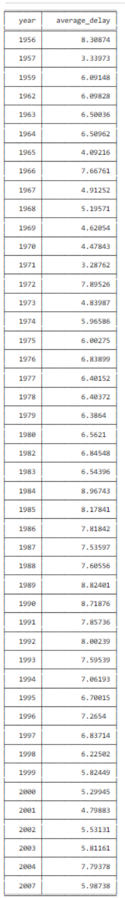


Methodology:

- Data Collection
- Data Cleaning
- Data Transformation
- Statistical Analysis
- Logistic Regression



```
Feature coefficients:
UniqueCarrier: -0.41592616010853733
lat_origin: -0.02472788594958321
long_origin: 0.01732631721702439
Year: -0.012916715774805342
long_dest: 0.011310413610952624
lat_dest: -0.003262644191880429
Distance: 0.0006178011191978469
CRSDepTime: -0.00047556415542441486
DepTime: 0.00044985926237693137
CRSArrTime: -8.565315048043337e-05
ArrTime: 5.088578261431095e-05
```



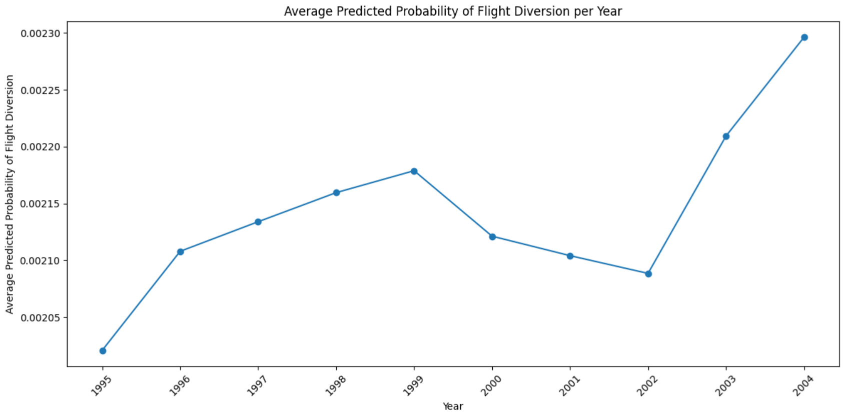
year	average_delay
1956	8.30874
1957	3.33973
1959	6.09148
1962	6.09828
1963	6.50036
1964	6.50962
1965	4.09216
1966	7.66761
1967	4.91252
1968	5.19571
1969	4.62054
1970	4.47843
1971	3.28762
1972	7.89526
1973	4.83987
1974	5.96586
1975	6.00275
1976	6.83899
1977	6.40152
1978	6.40372
1979	6.3864
1980	6.5621
1982	6.84548
1983	6.54396
1984	8.96743
1985	8.17841
1986	7.81842
1987	7.53597
1988	7.60556
1989	8.82401
1990	8.71876
1991	7.85736
1992	8.00239
1993	7.59539
1994	7.06193
1995	6.70015
1996	7.2654
1997	6.83714
1998	6.22502
1999	5.82449
2000	5.29945
2001	4.79883
2002	5.53131
2003	5.81161
2004	7.79378
2007	5.98738

Flight Analysis

Language:

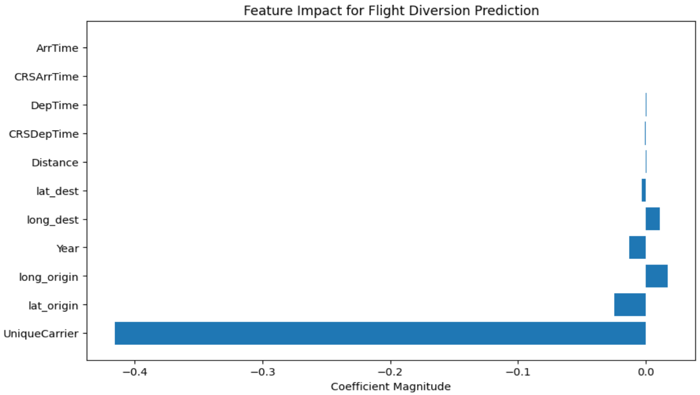
Data:

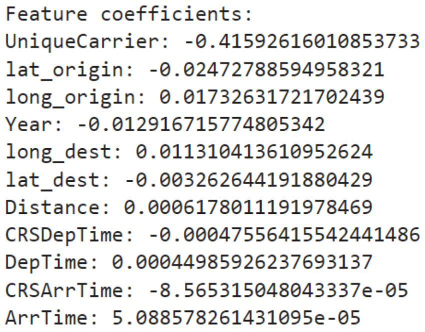
- Source: The 2009 ASA Statistical Computing and Graphics Data Expo


Average Predicted Probability of Flight Diversion per Year

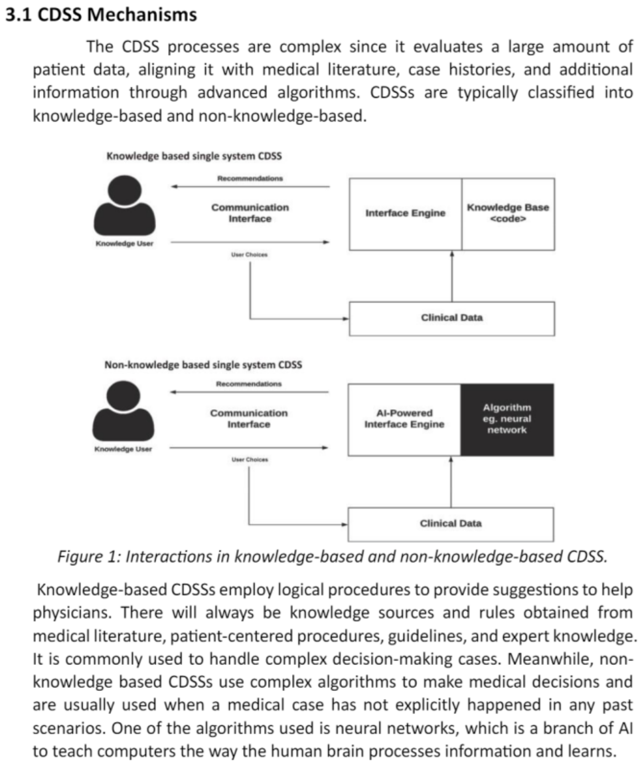
Methodology:

- Data Collection
- Data Cleaning
- Data Transformation
- Statistical Analysis
- Logistic Regression


Feature Impact for Flight Diversion Prediction


```
Feature coefficients:
UniqueCarrier: -0.41592616010853733
lat_origin: -0.02472788594958321
long_origin: 0.01732631721702439
Year: -0.012916715774805342
long_dest: 0.011310413610952624
lat_dest: -0.003262644191880429
Distance: 0.0006178011191978469
CRSDepTime: -0.00047556415542441486
DepTime: 0.00044985926237693137
CRSArrTime: -8.565315048043337e-05
ArrTime: 5.088578261431095e-05
```

Research Paper: Clinical Decision Support System



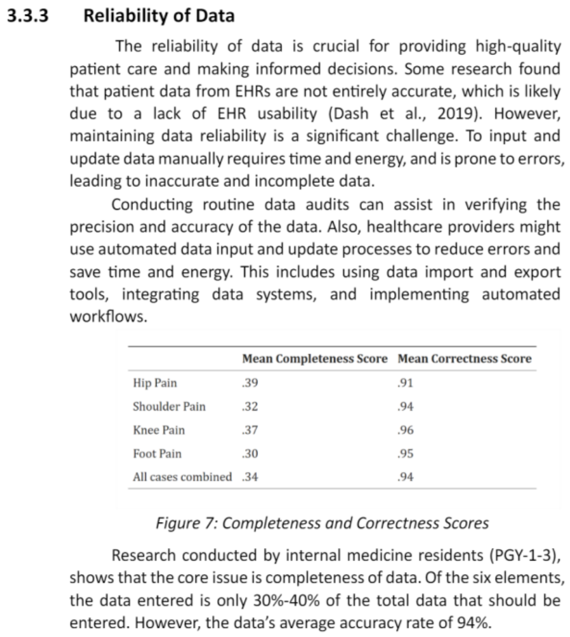
3.1 CDSS Mechanisms

The CDSS processes are complex since it evaluates a large amount of patient data, aligning it with medical literature, case histories, and additional information through advanced algorithms. CDSSs are typically classified into knowledge-based and non-knowledge-based.

Knowledge based single system CDSS
Recommendations
Communication Interface
Knowledge User
User Choices
Interface Engine
Knowledge Base <code>
Clinical Data
Non-knowledge based single system CDSS
Recommendations
Communication Interface
Knowledge User
User Choices
AI-Powered Interface Engine
Algorithm eg. neural network
Clinical Data

Figure 1: Interactions in knowledge-based and non-knowledge-based CDSS.

Knowledge-based CDSSs employ logical procedures to provide suggestions to help physicians. There will always be knowledge sources and rules obtained from medical literature, patient-centered procedures, guidelines, and expert knowledge. It is commonly used to handle complex decision-making cases. Meanwhile, non-knowledge based CDSSs use complex algorithms to make medical decisions and are usually used when a medical case has not explicitly happened in any past scenarios. One of the algorithms used is neural networks, which is a branch of AI to teach computers the way the human brain processes information and learns.



3.3.3 Reliability of Data

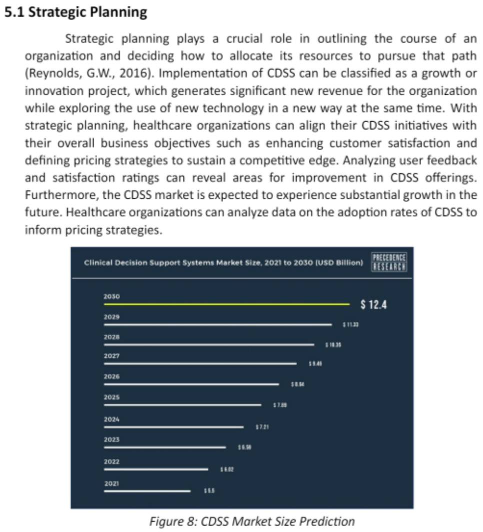
The reliability of data is crucial for providing high-quality patient care and making informed decisions. Some research found that patient data from EHRs are not entirely accurate, which is likely due to a lack of EHR usability (Dash et al., 2019). However, maintaining data reliability is a significant challenge. To input and update data manually requires time and energy, and is prone to errors, leading to inaccurate and incomplete data.

Conducting routine data audits can assist in verifying the precision and accuracy of the data. Also, healthcare providers might use automated data input and update processes to reduce errors and save time and energy. This includes using data import and export tools, integrating data systems, and implementing automated workflows.

	Mean Completeness Score	Mean Correctness Score
Hip Pain	.39	.91
Shoulder Pain	.32	.94
Knee Pain	.37	.96
Foot Pain	.30	.95
All cases combined	.34	.94

Figure 7: Completeness and Correctness Scores

Research conducted by internal medicine residents (PGY-1-3), shows that the core issue is completeness of data. Of the six elements, the data entered is only 30%-40% of the total data that should be entered. However, the data's average accuracy rate of 94%.



5.1 Strategic Planning

Strategic planning plays a crucial role in outlining the course of an organization and deciding how to allocate its resources to pursue that path (Reynolds, G.W., 2016). Implementation of CDSS can be classified as a growth or innovation project, which generates significant new revenue for the organization while exploring the use of new technology in a new way at the same time. With strategic planning, healthcare organizations can align their CDSS initiatives with their overall business objectives such as enhancing customer satisfaction and defining pricing strategies to sustain a competitive edge. Analyzing user feedback and satisfaction ratings can reveal areas for improvement in CDSS offerings. Furthermore, the CDSS market is expected to experience substantial growth in the future. Healthcare organizations can analyze data on the adoption rates of CDSS to inform pricing strategies.

Figure 8: CDSS Market Size Prediction

Full report is available at https://github.com/gezie1/Portfolio

Stay In Touch

elizabethleonnyefendi@gmail.com

https://github.com/gezie1/Portfolio