Investigating Factors Affecting Airline Customer Satisfaction

A Study of Passenger Data from 2022 to 2024.

By: Aarshi Jain
Email: aarshi.jain@torontomu.ca
Student #: 501178324

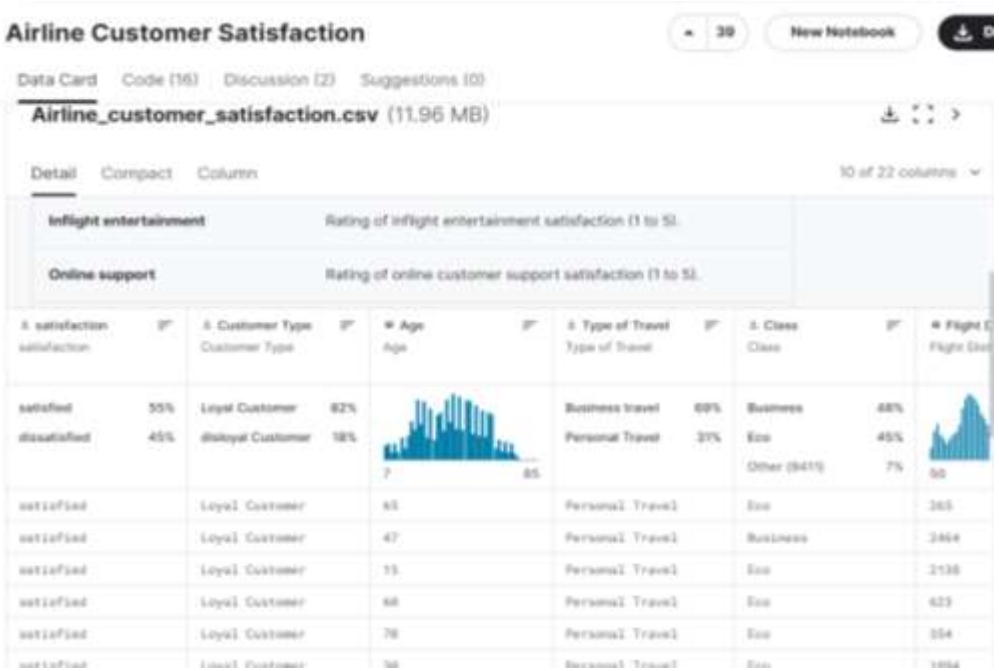
Ryerson University

Objective and Goals

- Objective: Find key factors influencing customer satisfaction in the airline industry from 2022-2024.
- Goals:
  1. Determine which factors are the most influential when predicting customer satisfaction.
  2. Analyze the influence of demographic variables.
  3. Evaluate service aspects that impact satisfaction.

Data and Methodology

- Dataset: Airline Customer Satisfaction (2022-2024).
- Features: 129,880 rows, 22 columns
- Techniques: Data cleaning, EDA, feature selection, model training, Apriori Algorithm




```
# Check for any missing values by finding the sum total of all null values in each variable.
print(cust_sat_df.isnull().sum())
```

```
# Replace the missing values with the median for each column.
for column in cust_sat_df.columns:
    if cust_sat_df[column].dtype in ['int64', 'float64']:      # Have to replace all numerical values only.
        median_value = cust_sat_df[column].median()
        cust_sat_df[column].fillna(median_value, inplace=True)
```

Ryerson University

SweetViz
2.3.1

Get updates, docs & report issues here

Created & maintained by Francois Bertrand
Graphic design by Jean-Francois Hains

DataFrame

NO COMPARISON TARGET

129880	ROWS
0	DUPLICATES
28.1 MB	RAM
27	FEATURES
23	CATEGORICAL
4	NUMERICAL
0	TEXT

ASSOCIATIONS

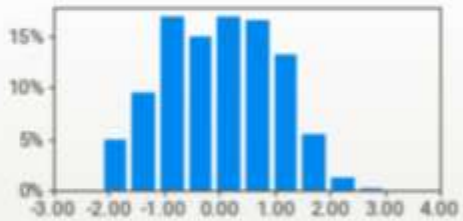
DataFrame

1 Age

VALUES	129,880 (100%)
MISSING	---
DISTINCT	75 (<1%)
ZEROES	---

MAX	3.01
95%	1.63
Q3	0.77
MEDIAN	0.04
AVG	-0.00
Q1	-0.82
5%	-1.62
MIN	-2.14

RANGE	5.16
IQR	1.59
STD	1.00
VAR	1.00
KURT	-0.719
SKEW	-0.004
SUM	-1.18e-11

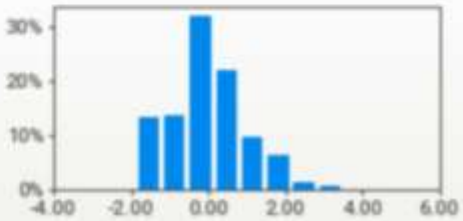


2 Flight Distance

VALUES	129,880 (100%)
MISSING	---
DISTINCT	5,398 (4%)
ZEROES	---

MAX	4.84
95%	1.80
Q3	0.55
AVG	0.00
MEDIAN	-0.05
Q1	-0.61
5%	-1.60
MIN	-1.88

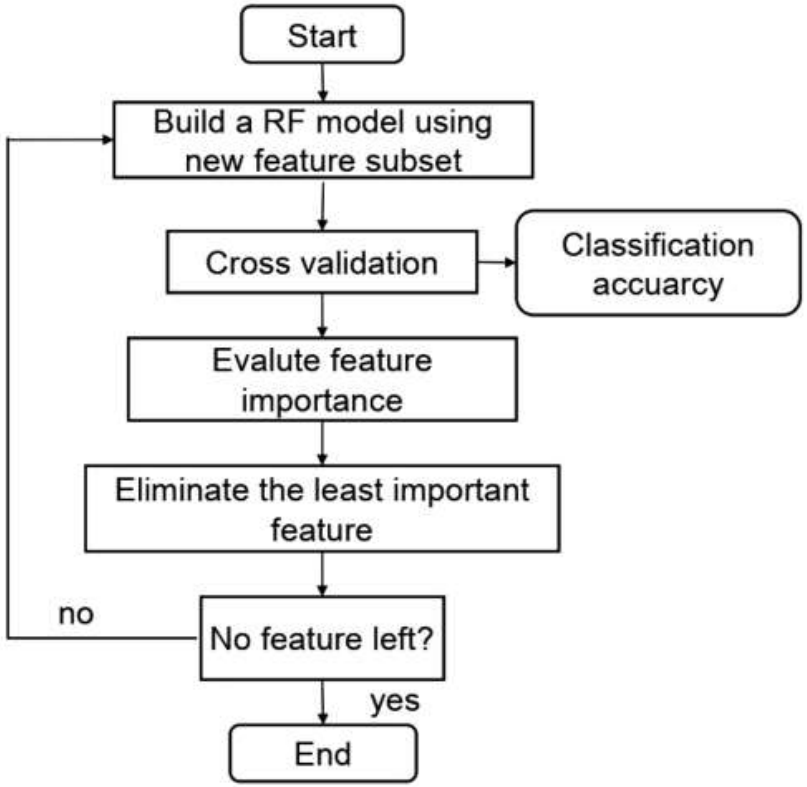
RANGE	6.72
IQR	1.15
STD	1.000
VAR	1.000
KURT	0.364
SKEW	0.467
SUM	1.48e-11



3 Seat comfort

Models Performance

- Models:
  1. Decision Trees
  2. Random Forest
  3. Logistic Regression
- Metrics
  1. Accuracy
  2. Precision
  3. Recall
  4. ROC-AUC



The RF-RFE flow.

$$\frac{1}{F} = \frac{1}{1+\beta^2} \cdot \left(\frac{1}{Precision} + \frac{\beta^2}{Recall} \right)$$

$$F = \frac{(1+\beta^2) \times Precision \times Recall}{(\beta^2 \times Precison) + Recall} \quad (23)$$

where β reflects the relative importance of precision to recall. When $\beta = 1$, that is, the precision is as important as the recall, the F value is the commonly used F1 value. Its equation is Eq. (24):

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (24)$$

The logic behind using Accuracy, Precision, and Recall.

Accuracy is the most commonly used performance measure in classification tasks. It refers to the proportion of the number of correctly classified samples to the total number of samples for a given test set. Its equation is Eq. (20):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

Precision refers to the proportion of samples whose real situation is positive in the samples determined as positive by the classifier for a given test set. Its equation is Eq. (21):

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

The recall rate refers to the proportion of samples determined as positive by the classifier in all positive samples for a given test set. The equation is Eq. (22):

$$Recall = \frac{TP}{TP + FN} \quad (22)$$

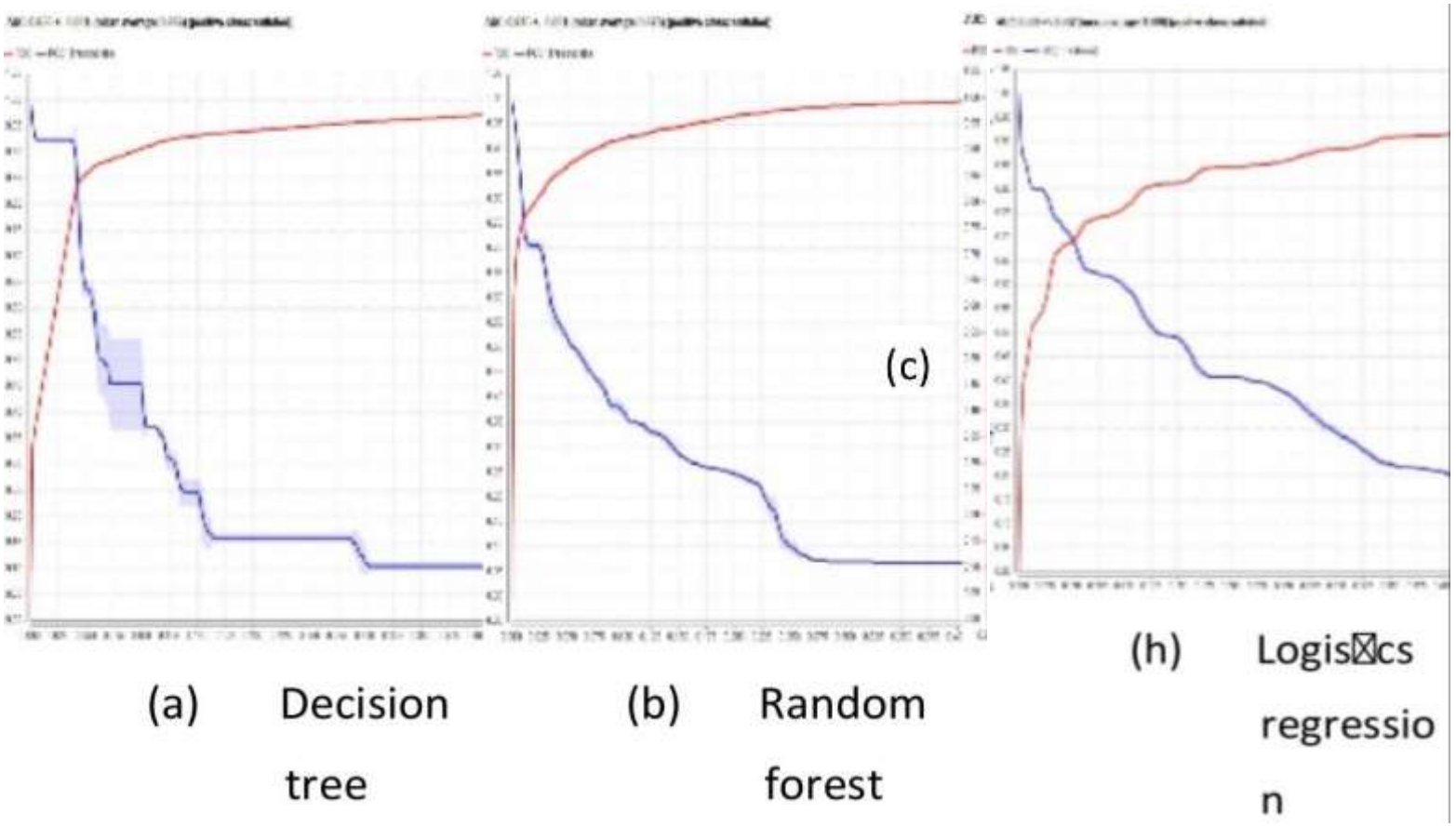
Comparing Models' Performances

```
Decision Tree Accuracy: 0.8669926085617493
Decision Tree Precision: 0.9168972939729397
Decision Tree Recall: 0.8339976225438781
Decision Tree F1 Score: 0.8734849317074957
Decision Tree ROC AUC: 0.9427457087725556
Random Forest Accuracy: 0.8968663381583
Random Forest Precision: 0.8912604363048747
Random Forest Recall: 0.9255996084189917
Random Forest F1 Score: 0.9081055122971906
Random Forest ROC AUC: 0.9560714296944239
Logistic Regression Accuracy: 0.8241453649522636
Logistic Regression Precision: 0.8396733440357367
Logistic Regression Recall: 0.841199916089784
Logistic Regression F1 Score: 0.8404359368450468
Logistic Regression ROC AUC: 0.8988780796838604
```

Overall, we see that Random Forest is the best model given

Ryerson University

ROC Curve of the 5 features chosen from feature selection.



Interpretability and Insights

- Key Findings:
  1. The importance of features such as inflight Wi-Fi service, seat comfort, and online boarding.
- Rationale:
  1. Customers fly better when they are provided with entertainment, they are comfortable, and they do not have to stress about their flight.

$$S_p = \frac{P(A \cap B)}{N} \quad (2)$$

Confidence (C_f) of a rule $A \rightarrow B$ can be defined as the ratio of the occurrence of A and B together with the total occurrence of A only in the dataset and can be evaluated as Eq. 3. The high confidence value close to 1 indicates a more strong rule.

$$C_f = \frac{P(A \cap B)}{P(A)} \quad (3)$$

Lift (L_f) is another important parameter to evaluate the effectiveness of a rule $A \rightarrow B$. It measure the occurrence of A and B together than expected. In other words, lift is the ratio of actual confidence value and expected confidence value. Actual confidence value evaluate the occurrence of A and B together with respect to the occurrence of A whereas expected confidence evaluate the occurrence of A and B with respect to the occurrence of B. The formula to evaluate the lift in given in Eq. 4. The expected values for lift may range from 0 *to* ∞. In general, the lift value greater than 1 is considered as good to select the rules to be evaluated on new data.

$$L_f = \frac{P(A \cap B)}{P(A) \times P(B)} \quad (4)$$

Finding the Support, Confidence, and Lift for the Apriori Algorithm.

The Apriori Algorithm was used to find hidden patterns and relationships in the dataset.

```
# Import the Apriori and Association Rules library.
from mlxtend.frequent_patterns import apriori, association_rules
```

```
x_selected_with_sat = X_selected.copy()
x_selected_with_sat['satisfaction_satisfied'] = cust_sat_df_std['satisfaction_satisfied']

# Now X_selected contains the satisfaction_satisfied column
print(x_selected_with_sat.tail())
```

```
        Seat comfort  Departure/Arrival time convenient  \
129875             5                                 5
129876             2                                 3
129877             3                                 0
129878             3                                 2
129879             3                                 4

        Inflight entertainment  Ease of Online booking  On-board service  \
129875                       5                       2                 3
129876                       1                       3                 2
129877                       2                       4                 4
129878                       2                       3                 3
129879                       3                       4                 5

        Leg room service  Checkin service  Customer Type_disloyal Customer  \
```

Then, a Feature Importance Graph is created.

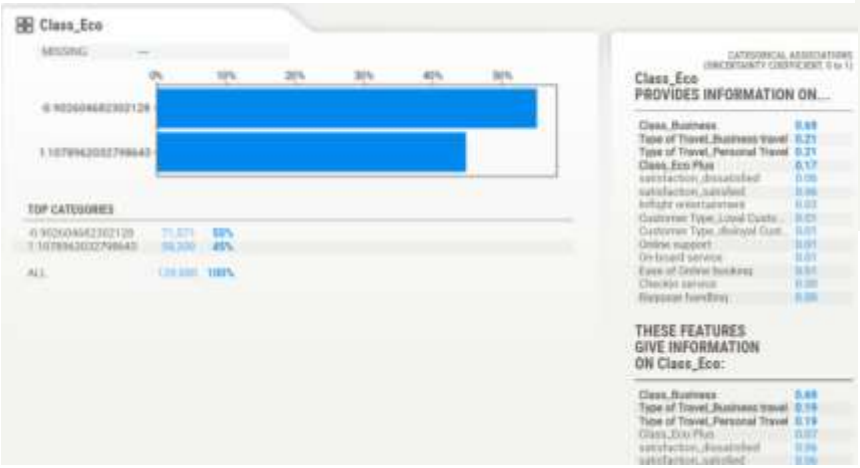
```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
%matplotlib inline

y = X_selected.pop('Survived')

model = RandomForestClassifier()
model.fit(X_selected, y)

(pd.Series(model.feature_importances_, index=X_selected.columns)
   .nlargest(4)
   .plot(kind='barh'))
```

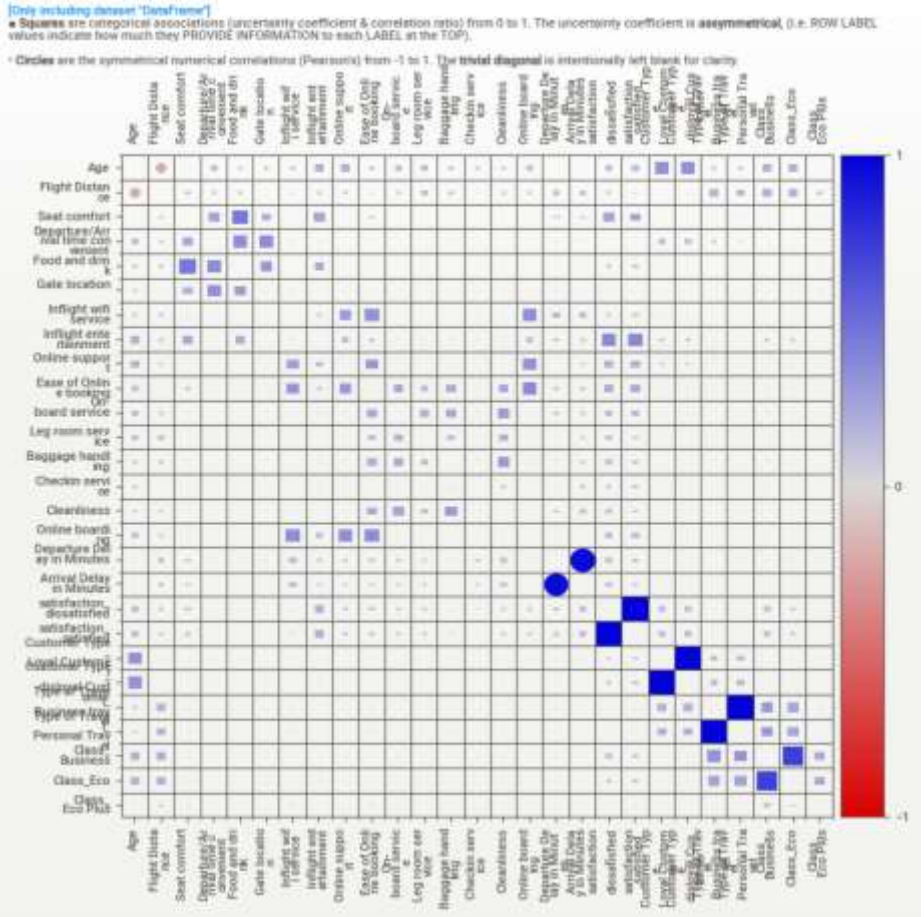
Finally, the results show the top three factors in predicting airline customer satisfaction to be:

1. Online Boarding;
2. Inflight Wifi Services, and;
3. Baggage Handling

Technical Insights

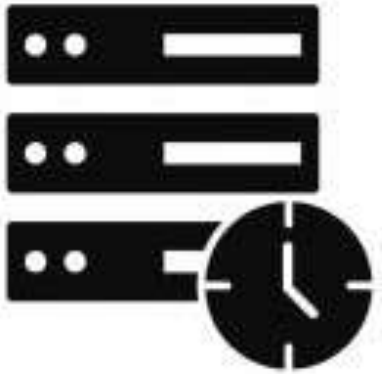
- Techniques used:
  1. EDA
  2. Feature Selection
  3. Model Training
  4. Evaluation



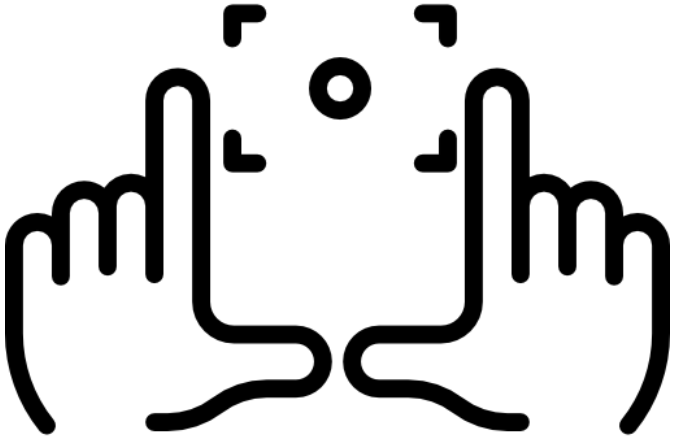


Limitations

- Scope: Limited to one unnamed airline and only specific years.
- Techniques: Did not use Lasso or Chi-Square due to complexity.



- **K-Best or Chi-Square**: To identify significant features, apple feature selection techniques.
    - Initial feature selection was found using K-Best and Chi-Square techniques (Bacani, 2022). This method reduces dimensionality and emphasizes the most important determinants which enhance the model's performance and interpretability.
- **Lasso Regression**: Refine the feature set by using Lasso Regression.
    - To remove collinearity, Sangah Kim used Lasso Regression to increase the robustness of the model (Kim, 2023).



Challenges

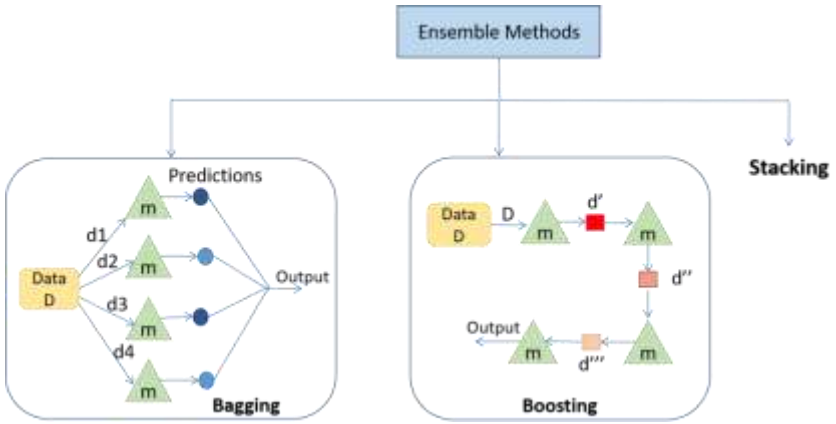
- Computational Issues such as tools crashing and processing time.
- Data volume: Handling large datasets.

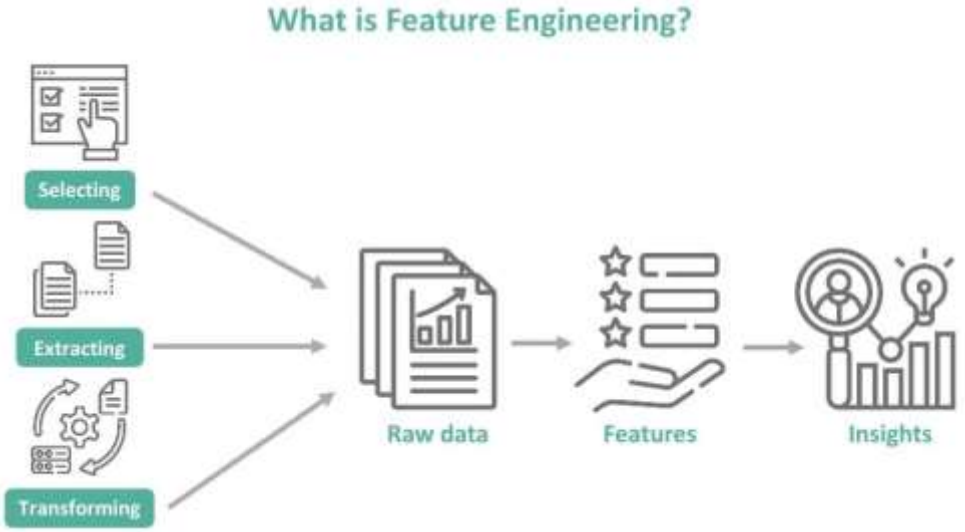
```
In [130…  # Save the report to an HTML file and save the findings as another file. Open the
          report.show_html('Customer_Airline_Satisfaction_EDA.html', open_browser = True)

          Report Customer_Airline_Satisfaction_EDA.html was generated! NOTEBOOK/COLAB USERS:
          book/colab files.

In [131…  # Import the IPython library to show the EDA report.
          import IPython

          # Display the EDA report within the study.
          IPython.display.HTML('Customer_Airline_Satisfaction_EDA.html')

Out[131]:
```

Ryerson University

Continuity and Next Steps

- Future Research will include a broader dataset and enhanced feature engineering.
- Applications: Applying techniques learned here to other industries.





Ryerson University

Ethical Considerations

- Factors such as Age, Class, and food and drink need to be interpreted ethically.
- Avoid discrimination or the exclusion of certain demographics based on the findings.

Conclusion

Thank you for your time!