BAN6800: Business Analytics Capstone

Nexford University

Milestone One Assignment

Title: Business Analytics Project-Ready Dataset for Behavioral Segmentation and Predictive Modeling of Purchasing Intent Among Takealot Online Shoppers

Name: Mubanga Nsofu
Learner ID: 149050
Date: 24th May 2025
Lecturer: Prof. Raphael Wanjiku

Introduction

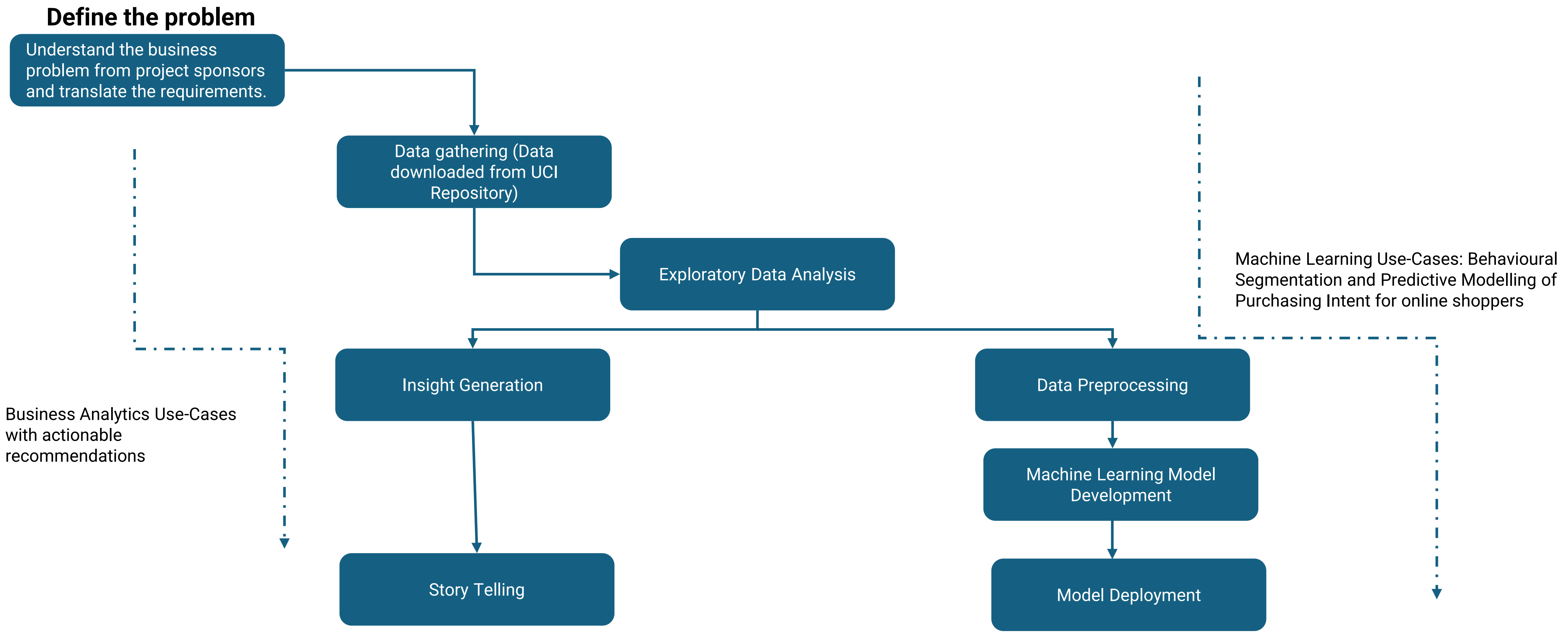
Takealot

Situation	Complication	Question	Answer
Dominanat e-commerce platform in South Africa for 14 years	Threat from e-commerce global giant Amazon a new entrant in the market	How can Takealot retain its competitive edge and market share?	Applied Machine learning using clustering and classification techniques

Takealot is a leading e-commerce player in South Africa (Wikipedia, 2025).

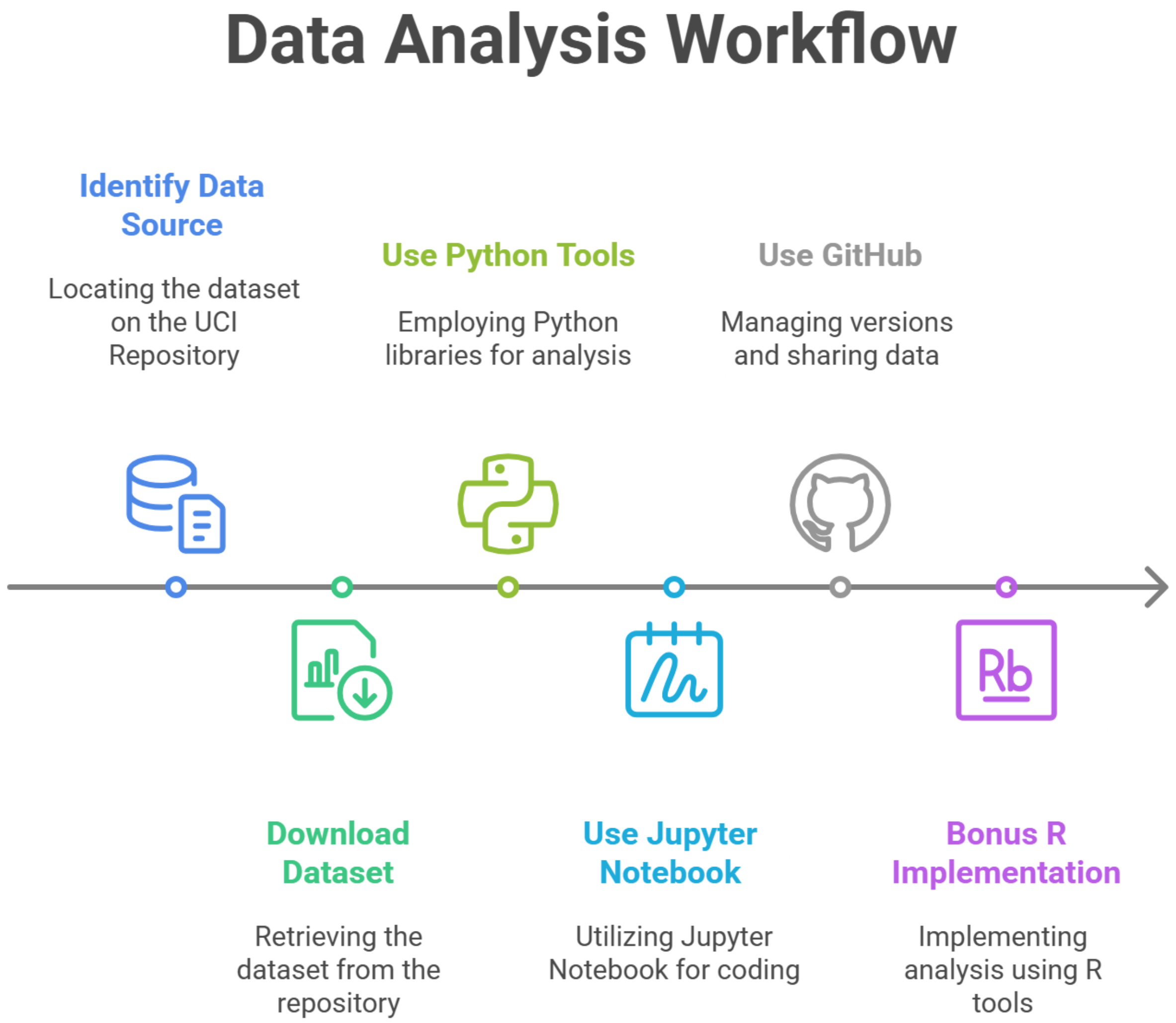
Nexford University

End-to-end workflow for the entire project



Note: Created by the Author (2024).

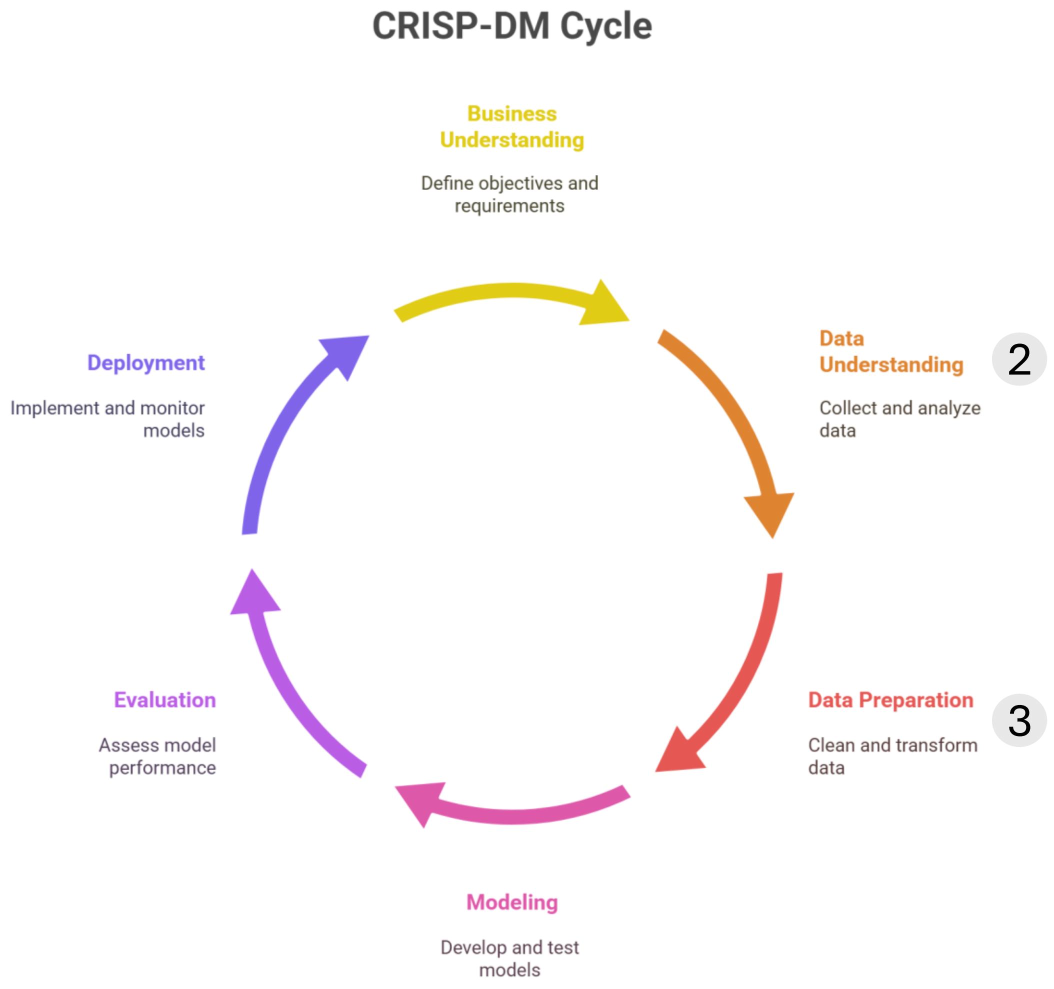
Relevant Data Sources, Methods and Tools for Collection



- The workflow identifies a relevant dataset for the project from the UCI repository
- Python libraries pandas, scikit-learn, matplotlib, seaborn and SweetViz are used for Exploratory Data Analysis (EDA) and data preprocessing
- The workflow is implemented in a Jupyter notebook and uploaded onto GitHub
- An R implementation of the entire process is also provided

Note: Created by the Author (2025).

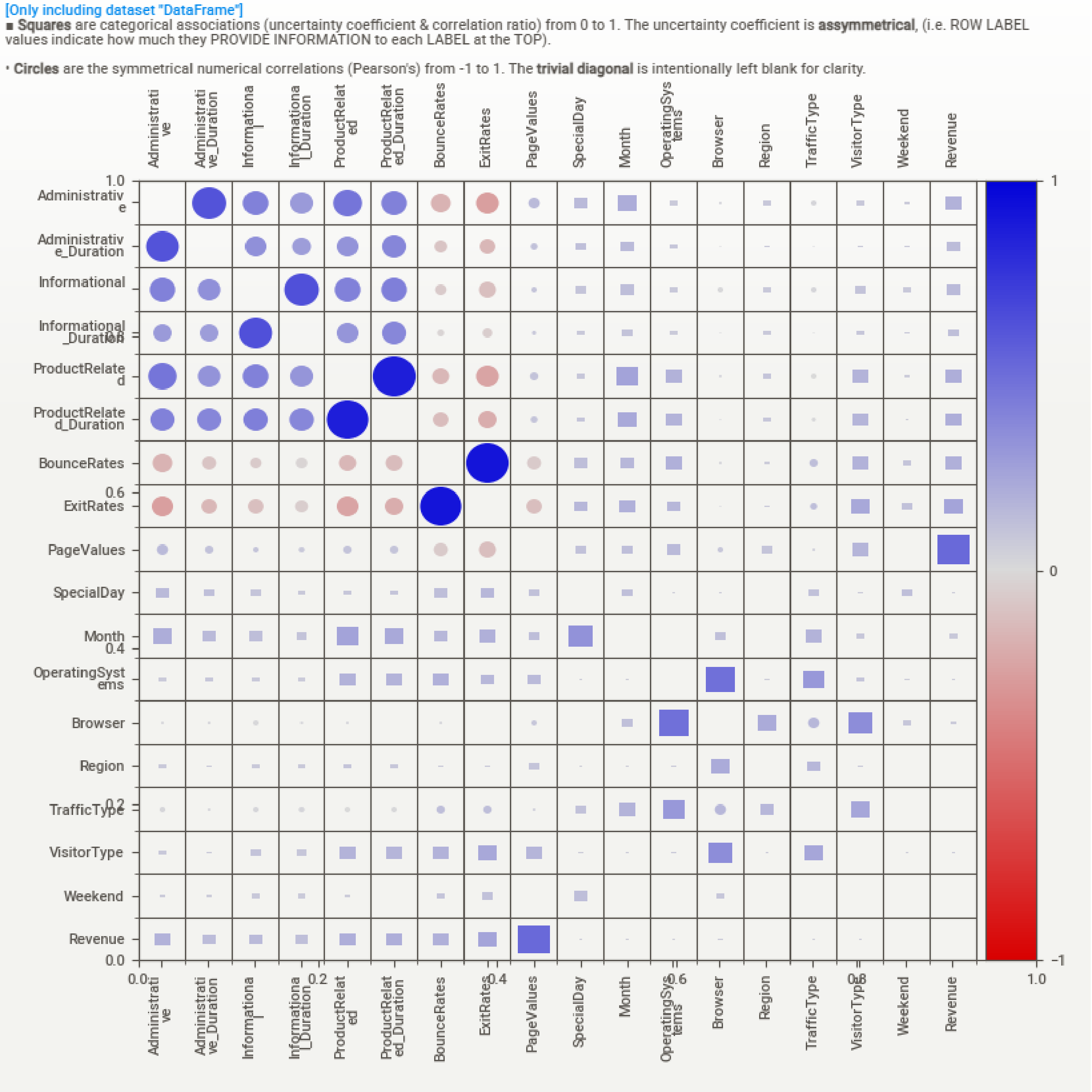
Steps 2 and 3 of CRISP-DM are applied to the dataset in Milestone One Assignment

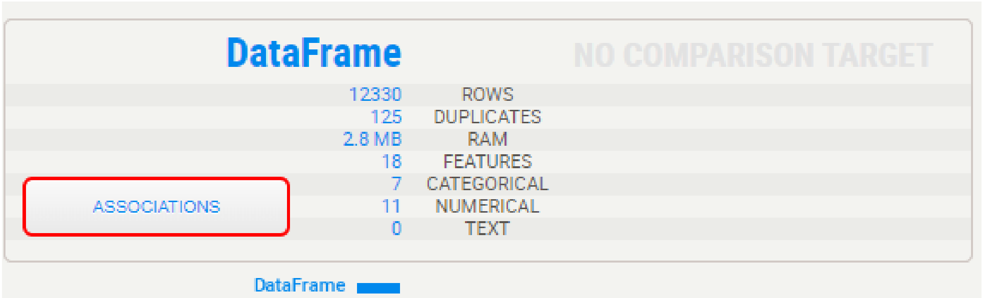


A clear understanding of the dataset before any modelling is important

Note: Created by the Author (2025).

Exploratory Data Analysis Using SweetViz





- The snippets show associations on the left and data frame details at the top: 12330 observations, 125 duplicates, and 18 features (7 categorical, 11 numerical).

- Automated EDA reveals correlations:
  - ProductRelated and ProductRelated_Duration are positively correlated, indicating that product browsing depth is crucial for engagement.
  - BounceRates and PageValues are negatively correlated, suggesting that high bounce rates lead to lower page values.

Nexford University

Exploratory Data Analysis using pandas [1/2]

```
# Using Pandas
data.info()
data.describe()

# Output
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
```

Using two functions info() and describe() from pandas. The output on the right is from data.info()

1.Data has 12330 entries and a total of 18 columns

2. There are no nulls/missing values

3. We have different data types including booleans

Exploratory Data Analysis using pandas and seaborn [2/2]


```python
# Code snippet for sns plot

# Create the pairplot
sns.pairplot(data[plot_features],
             x_vars=selected_features,
             y_vars=plot_features,
             hue="Revenue")
```

Key insights:

1. Users who made a purchase (red) generally:
   - Viewed more product pages
   - Spent more time on product-related content
   - Clear separation between True/False revenue labels at high values
2. Administrative / Informational Pages:
   - No strong visual separation between purchasers and non-purchasers.
3. Skweness in Data:
   - Most numerical features are right-skewed
   - Scaling or log transformation will be required before applying any machine learning algorithm

Data Cleaning and Feature Engineering Steps [1/2]

One Hot Encoding

Encoding categorical data into binary vectors (e.g. Visitor Type)

Boolean to Int Conversion

Converting boolean values to integers for analysis (e.g. Revenue & Weekend)

Ordinal Encoding

Encoding ordinal data into numerical order (e.g. Months)

Feature Engineering

Creating new features from existing data (e.g. Total Duration, product related duration, interaction intensity)

Note: Created by the Author (2025).

Data Cleaning and Feature Engineering Steps [2/2]

Snippet of One Hot Encoding from code

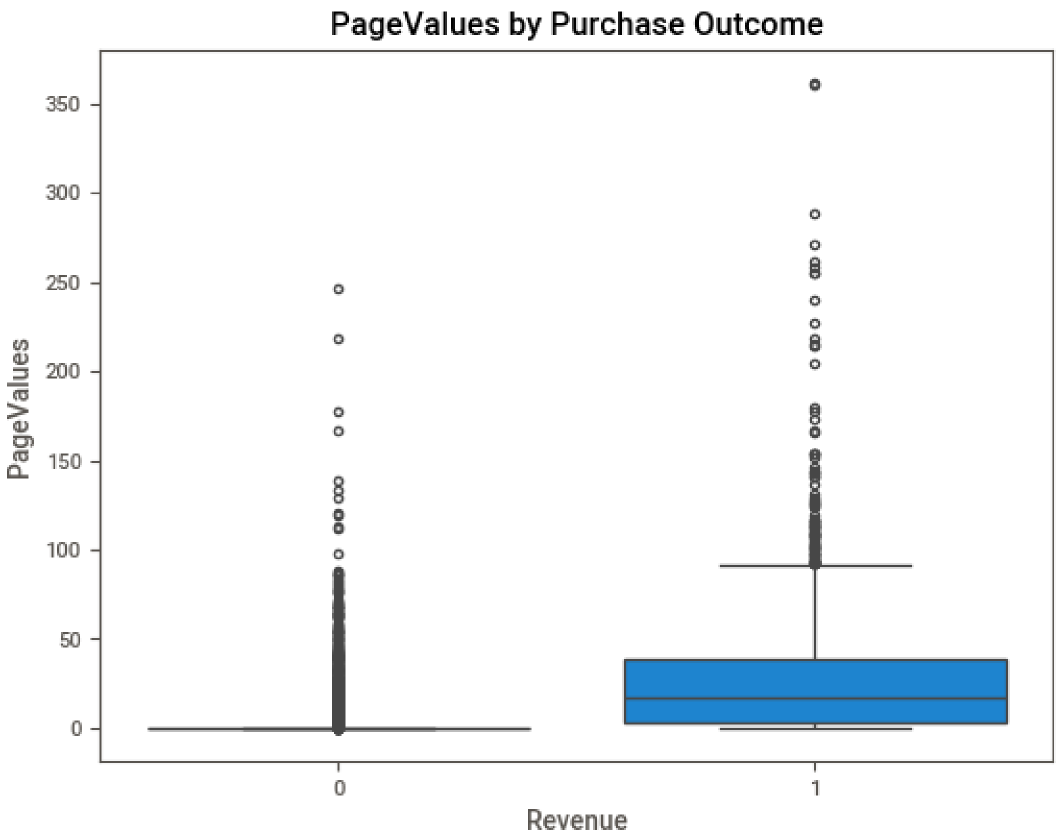
```
# Encode 'VisitorType' using one-hot encoding
(drop_first to avoid multicollinearity)
data = pd.get_dummies(data, columns=['VisitorType'],
drop_first=True)
```

Snippet of Feature Engineering from Code

```
# Feature Engineering
data['Total_Duration'] =
data['Administrative_Duration'] +
data['Informational_Duration'] +
data['ProductRelated_Duration']
data['Interaction_Intensity'] = data['PageValues'] /
(data['ProductRelated_Duration'] + 1e-5)  # avoid
division by zero
```

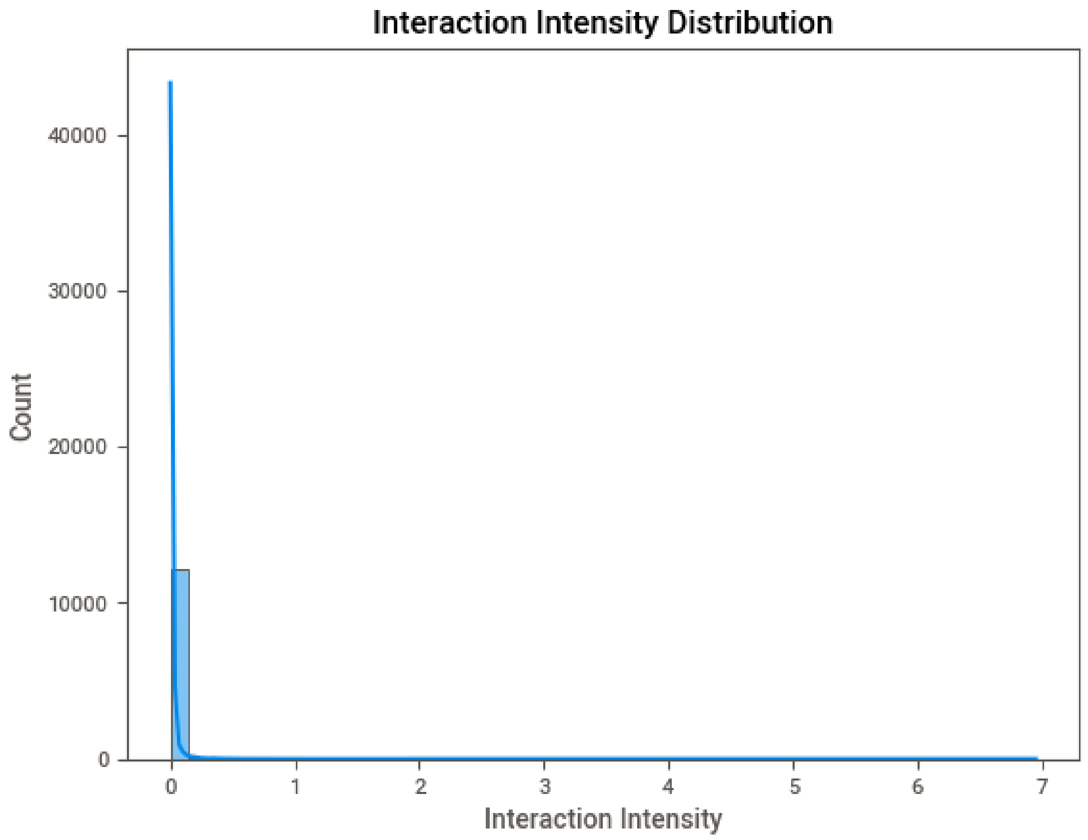
The entire data cleaning and feature engineering steps are included in the Jupyter notebook

Data visualisation of cleaned dataset with new features



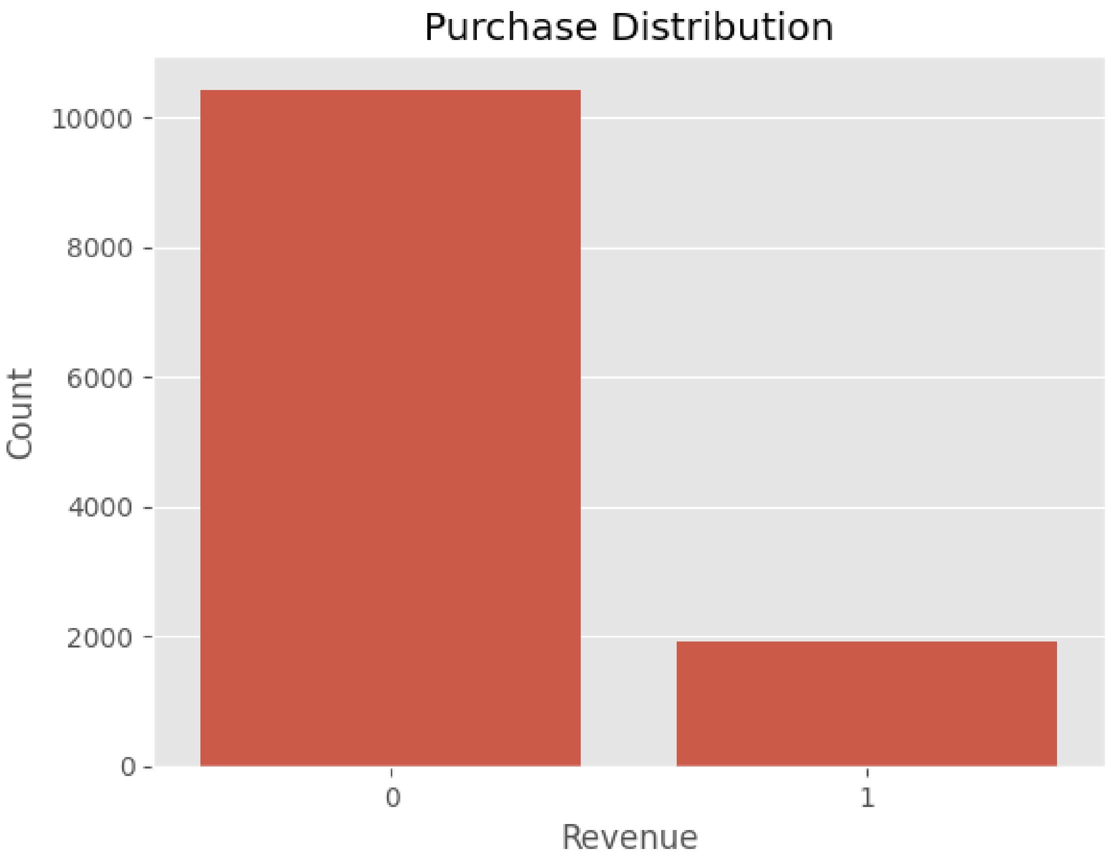
- Users who converted (Revenue = 1) had consistently higher PageValues.
- The median PageValue for Revenue = 1 is significantly higher than for Revenue = 0.
- Both categories have extreme outliers, but purchases have higher and more frequent outliers. This suggests that even among those who didn't purchase, a small number had high PageValues—possibly abandoned carts or last-minute dropouts.

Data visualisation of cleaned dataset with new features



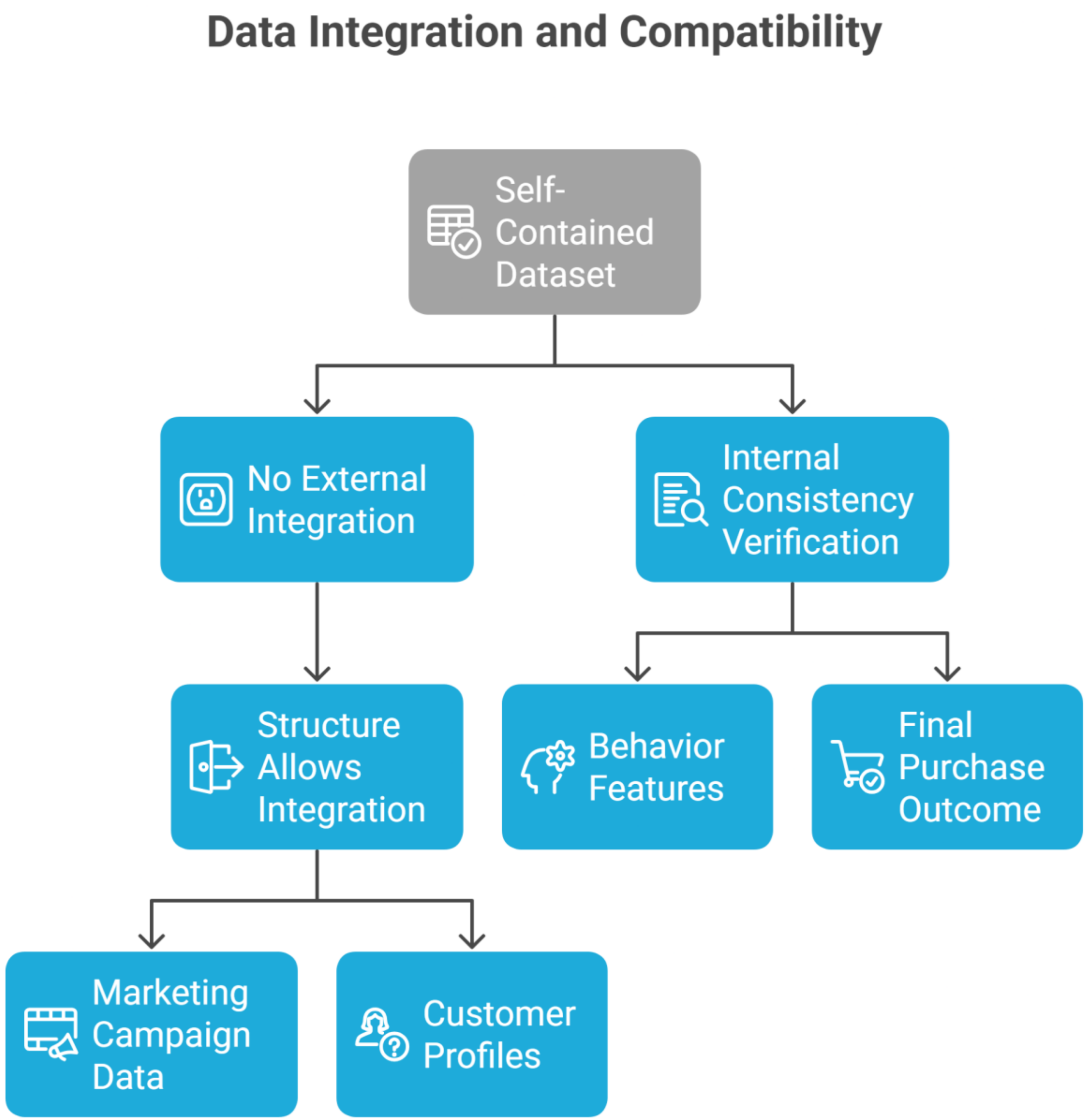
- The long tail suggests a small group of users are highly engaged (e.g., repeated or prolonged interaction with products).

Data visualisation of cleaned dataset with new features



- 85% of the visitors to Takealot's e-commerce website did not make any purchase
- This dataset has class imbalance.

Data Combination and Compatibility



Note: Created by the Author (2025).

- The dataset is self-contained and includes all behavioral signals needed for this project
- No external integration was necessary, but the data structure allows for integration with marketing campaign data and customer profiles
- Internal consistency of the dataset was verified between behavioural features and final purchase outcome

Data Quality Assessment (Bias and Ethics)

Data Quality Assessment Components

Quality Checks

Data completeness, distribution normality, and outlier analysis were performed. These checks ensure data reliability.

Bias

Target variable imbalance was detected, requiring mitigation strategies. Techniques like stratified sampling will be employed.

Ethical Use

The dataset contains no PII and is ethically safe. It is suitable for academic research.

- The Revenue variable shows class imbalance, and this will be addressed using techniques such as SMOTE during predictive modelling (Agular, n.d.).

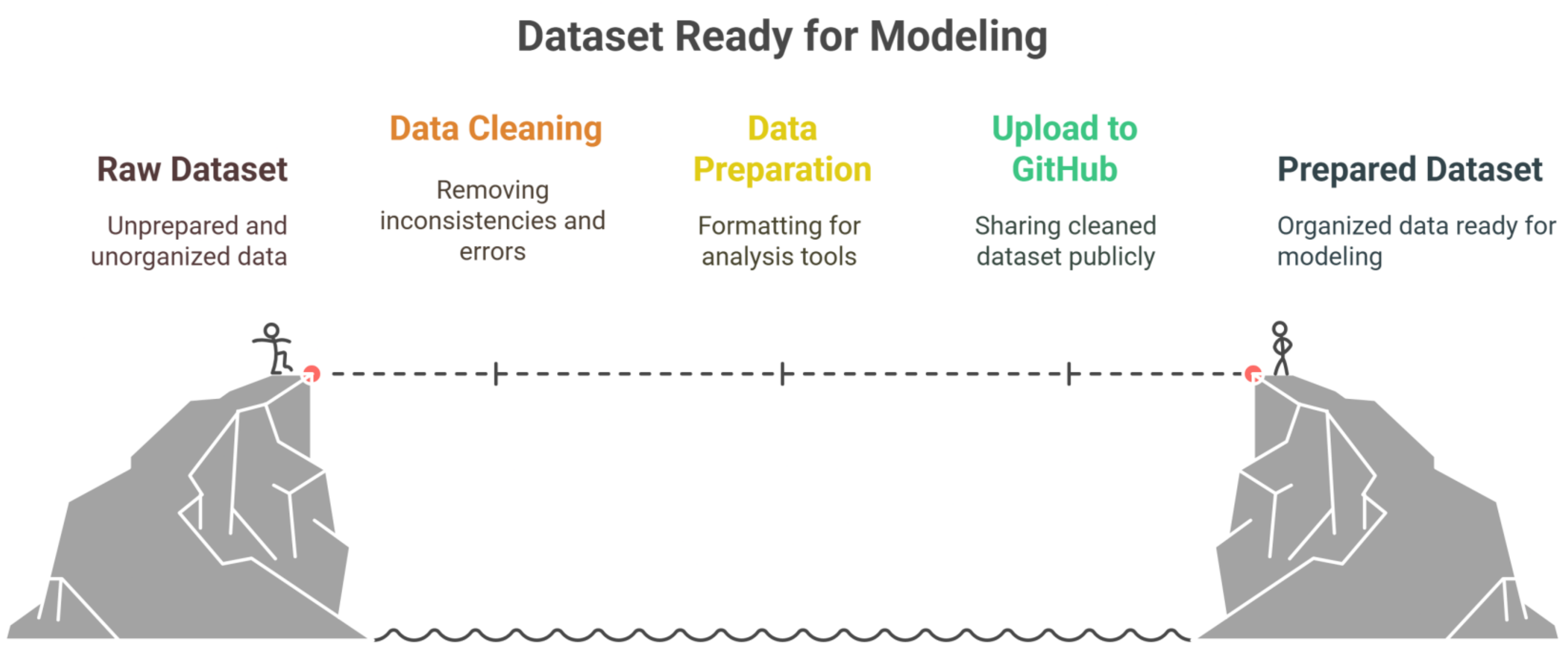
```
# Check for class imbalance in the Revenue
variable
data['Revenue'].value_counts()

# Output
Revenue
False    10422
True      1908
Name: count, dtype: int64
```

Note: Created by the Author (2025).

Note: SMOTE (Synthetic Minority Oversampling Technique)

Conclusion



- Link to prepared dataset: https://github.com/RProDigest/BAN6800/tree/main/Week-3

- Dataset is prepared for the next steps in the CRISP-DM process which is modelling
- Dataset will be used for both clustering and classification
- Cleaned dataset is uploaded to github at the provided link:
- https://github.com/RProDigest/BAN6800/tree/main/Week-3

Note: Created by the Author (2025).

References

- Wikipedia. (2025). *Takealot.com*. Wikipedia. https://en.wikipedia.org/wiki/Takealot.com
- Sakar, C. & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5F88Q.
- Roy, A. (2018). Chapter 1 - Introduction to CRISP DM Framework for Data Science and Machine Learning | LinkedIn. https://www.linkedin.com/pulse/chapter-1-introduction-crisp-dm-framework-data-science-anshul-roy/
- Agular, H. (n.d.). *What Is Imbalanced Data and How to Handle It? -*. TurinTech AI. Retrieved December 23, 2024, from https://www.turintech.ai/what-is-imbalanced-data-and-how-to-handle-it/

Nexford University

Thank you!