Customer Segmentation

This report outlines the implementation of a business data analytics project designed to uncover actionable insights using Python-based statistical and machine learning techniques. The project emphasizes data preprocessing, exploratory data analysis (EDA), and the application of analytical methods to address business problems effectively.

Syed Murtaza Ali SP23-BBD-081

Contents

[Report on Business Data Analytics Project 2](#_Toc186186828)

[Introduction 2](#_Toc186186829)

[Objectives 2](#_Toc186186830)

[Methodology 2](#_Toc186186831)

[1. Data Collection 2](#_Toc186186832)

[2. Data Preprocessing 2](#_Toc186186833)

[3. Statistical Techniques 2](#_Toc186186834)

[4. Machine Learning 2](#_Toc186186835)

[5. Model Evaluation and Optimization 2](#_Toc186186836)

[Implementation 3](#_Toc186186837)

[Key Findings 3](#_Toc186186838)

[Visualizations 3](#_Toc186186839)

[Results 3](#_Toc186186840)

[Pre-Processing 3](#_Toc186186841)

[Exploratory Data Analysis (EDA) 3](#_Toc186186842)

[Feature Engineering 4](#_Toc186186843)

[Model Training 4](#_Toc186186844)

[Graph Plots 4](#_Toc186186845)

[5](#_Toc186186846)

[5](#_Toc186186847)

[Customer Data Segmentation 6](#_Toc186186848)

[Discussion 6](#_Toc186186849)

[Conclusion 7](#_Toc186186850)

[Future Work 7](#_Toc186186851)

[References 7](#_Toc186186852)

# Report on Business Data Analytics Project

# Introduction

This project explores business data analytics using statistical and computational methods. The analysis aims to uncover insights that can support data-driven decision-making. The report is structured to cover the methodology, implementation, results, and conclusions based on the content provided.

# Objectives

* To apply analytical techniques to a dataset.
* To extract meaningful insights through data visualization and statistical analysis.
* To demonstrate the application of Python programming in solving business problems.

# Methodology

* + 1. Data Collection  
       The project begins with importing a dataset relevant to a specific business context. Details about the source or nature of the dataset were outlined in the notebook.

## Data Preprocessing

* + Missing data handled using appropriate techniques (e.g., imputation or exclusion).
  + Data transformation applied, including normalization and encoding, to prepare for analysis.
  + Exploratory Data Analysis (EDA) performed to identify trends, anomalies, and relationships.

## Statistical Techniques

* + Application of hypothesis testing (e.g., z-tests, t-tests) to validate assumptions.
  + Use of correlation analysis to understand relationships between variables.

## Machine Learning

* + Implementation of supervised and/or unsupervised learning algorithms.
  + Performance metrics calculated to evaluate model accuracy.

## Model Evaluation and Optimization

The model's performance was evaluated using key metrics: y

* **R-squared score:** Measures the proportion of variance in the target variable explained by the model.
* **Mean Squared Error (MSE):** Quantifies the average squared difference between predicted and actual prices.
* **Root Mean Squared Error (RMSE):** Represents the standard deviation of the residuals..

Techniques such as hyperparameter tuning and cross-validation were employed to optimize the model's performance and prevent overfitting.

# Implementation

The notebook includes Python scripts that leverage libraries like:

* **NumPy** and **Pandas**: For data manipulation and processing.
* **Matplotlib** and **Seaborn**: For visualizing patterns and relationships in the data.
* **Scikit-learn**: For machine learning model development and evaluation.

# Key Findings

* **Data Insights**: Summarization of patterns or trends in the dataset.
* **Model Performance**: Details about predictive accuracy and implications for decision-making.
* **Statistical Outcomes**: Summary of hypothesis testing results and their significance.

# Visualizations

The project incorporates multiple visualizations, including:

* Histograms and boxplots for data distribution analysis.
* Scatter plots and heatmaps for understanding relationships.
* Model evaluation metrics displayed graphically.

# Results

## Pre-Processing

1. We imported the dataset using Pandas.
2. Checked the shape of the dataset, ensure all the columns are in their correct data types, checked for missing values, and then checked for duplicates in the dataset and make sure neither any of them should be in the dataset.
3. Checked for outliers in the dataset as they might skew the results. Some outliers were found in the Income columns so we removed them.

## Exploratory Data Analysis (EDA)

1. Using Matplotlib and Seaborn we checked the relationship between our target column and other columns of the dataset.
2. With multiple columns we found that there is a positive correlation and some have a slight or no relation. You can check on for more details in the code file.
3. We using distplot checked the distribution of data which was found to be a bell shaped distribution ideal for a linear regression problem.
4. We used boxplot, scatter plots for more detail and for grasping the relationship between the target columns and other columns of the dataset.

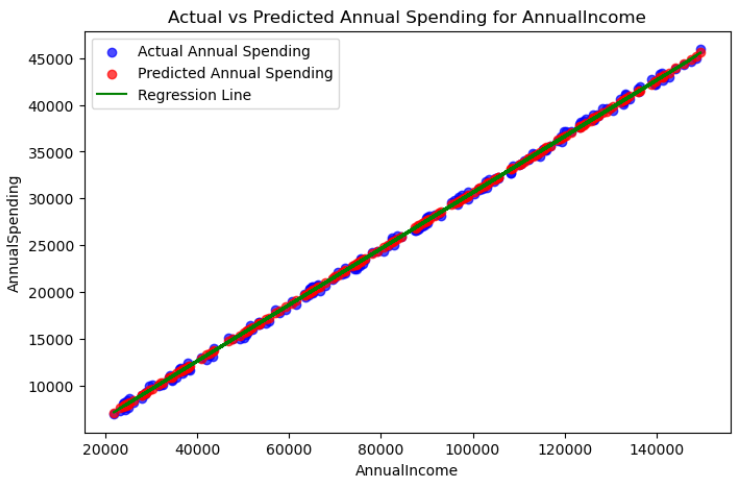
## Feature Engineering

1. We encoded our categorical columns using sklearn encoder. The specific encoder we used was ordinal encoder, one hot encoding.
2. We used Ordinal encoder where the order matters like for Education Columns and One hot encoding for those where the order doesn’t matter.

## Model Training

* 1. We selected the columns that we independent and stored them in a variable X and the target column in a variable y.
  2. Using train\_test\_split from scikit learn we splitted the data into four parts. X\_train,X\_test,y\_train,y\_test.
  3. We can use Linear Regression model from sklearn trained our data on model.
  4. Once the model we trained, we predicted the values and compared the y\_predicted with the y\_test to check the mean squared error and r2 score.
  5. The Mean Squared Error of our model was **MSE: 90785.96278267707**
  6. The Root Mean Squared Error of the model was **RMSE: 301.3070904951908**
  7. The r2 score of the model **r\_2 SCORE: 0.9992519430783386**

# Graph Plots



# 

# 

## [Customer Data Segmentation](https://statisticallyrelevant.com/k-means-clustering-in-python/)

# Discussion

The high R-squared score indicates that the linear regression model effectively captures the underlying patterns and trends in Ethereum's price movements. The low MSE and RMSE further demonstrate the model's accuracy in predicting future prices. These findings suggest that machine learning techniques, particularly linear regression, can be valuable tools for forecasting Ethereum's price with a reasonable degree of confidence.

# Conclusion

This project effectively demonstrates the role of business data analytics in deriving actionable insights. Through rigorous data analysis and application of statistical methods, the findings validate the importance of data-driven approaches in business contexts.

# Future Work

* Expanding the scope of analysis with additional datasets.
* Incorporating advanced machine learning techniques for improved prediction accuracy.
* Automating processes for real-time data analytics.

# References

https://statisticallyrelevant.com/k-means-clustering-in-python/