**DATA SCIENCE TOOLBOX PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

[Operations Survey In Business finance](https://www.stats.govt.nz/assets/Uploads/Business-operations-survey/Business-operations-survey-2022/Download-data/business-operations-survey-2022-business-finance.csv)

**Submitted by**

**NAME:** S.Surendranath Reddy

**REGISTRATION NO:** 12303218

**PROGRAMME AND SECTION:** K23GN05

**COURSE CODE:** INT375

Under the Guidance of

**MRS. AASHIMA (UID: 28968)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **S.Surendranath Reddy** bearing Registration no. **12303218** has completed **INT375** project titled, **“Mrs.Aashima”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, S.Surendranath Reddy student of CSE (Program name) under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature

Registration No: 12303218 Name of the student: S.Surendranath Reddy

|  |  |
| --- | --- |
| Title | **Pg .no** |
| **1. Introduction** | **5** |
| **2.** **Source of Dataset** | **5** |
| **3. EDA Process** | **6** |
| **4. Analysis on Dataset** | **6** |
| **4.1 Introduction to Dataset** | **6** |
| **4.2 General Description** | **7** |
| **4.3 Specific Requirements, Functions, and Formulas** | **7** |
| **4.4 Analysis Results** | **7** |
| **4.5 Visualization** | **8** |
| **5. Business Decision Support** | **13** |
| **6. Conclusion** | **14** |
| **7. References** | **15** |
| **8. Source Code** | **15** |

**1. Introduction**

Data analysis is a critical tool in modern business decision-making. By systematically analyzing datasets, businesses can gain valuable insights into customer behavior, financial trends, and operational performance. This enables organizations to make data-driven decisions that improve efficiency, profitability, and overall competitiveness.

In today’s fast-paced business environment, data is generated at an unprecedented rate, and being able to extract actionable insights from this data is crucial. Effective data analysis not only helps businesses understand past and current performance but also allows for future predictions, identification of opportunities, and risk mitigation.

This project aims to demonstrate how data analysis can support business decisions using a general business dataset. The dataset includes various transaction details and business characteristics, such as industry, business size, transaction value, and others, which provide a comprehensive view of the business landscape. The analysis leverages Python tools like Pandas, NumPy, Matplotlib, Seaborn, and SciPy to clean, explore, and visualize the data.

**Objective:** The main goal of this project is to clean, explore, analyze, and visualize the dataset to provide meaningful insights that can support business decision-making. By analyzing revenue trends, customer segmentation, and identifying areas for improvement, the project aims to offer data-driven recommendations to enhance business operations.

**2. Source of Dataset**

The dataset used for this analysis was sourced from a local file titled python\_project.csv, encoded in ISO-8859-1 format. It contains records related to various business transactions and their attributes. The key columns in the dataset include:

**line\_code:** Identifier for each specific business line.

**industry:** Type of industry the business belongs to.

**size:** The business size (Small, Medium, Large).

**level:** A numerical value representing the business or transaction tier.

**value:** The financial amount associated with the transaction.

**description:** A brief description of the transaction or business activity.

Each record represents a unique financial transaction, providing a broad view of business activity across different sectors.

The file was loaded into the environment using Pandas, a powerful Python library for data manipulation. Initial exploration of the dataset revealed some common data issues that needed to be addressed before deeper analysis

Source of data set: <https://www.stats.govt.nz/large-datasets/csv-files-for-download/>

**3. What is EDA**

Exploratory Data Analysis (EDA) is the process of analyzing a dataset to summarize its main characteristics, often using visualizations and statistical techniques. It is an essential first step in the data analysis pipeline, as it helps analysts understand the data’s distribution, identify patterns, and detect anomalies or outliers. EDA allows for a more intuitive understanding of the dataset before diving into complex statistical models or machine learning techniques.

The importance of EDA can be summarized in several key points:

**Identifying Trends and Outliers:** EDA helps to spot trends and outliers that could influence or distort statistical models, ensuring that they are properly handled before further analysis.

**Providing Early Insights:** Through EDA, analysts can gain valuable insights into the data early on without needing complex algorithms. This includes understanding data distribution and identifying key variables that could drive future analysis.

**Informs Data Cleaning:** During the EDA process, issues like missing values, inconsistencies, and formatting errors are identified, allowing for effective data cleaning and preparation.

**Supports Feature Engineering:** EDA helps uncover relationships between variables, which can guide the creation of new features that improve the performance of models.

**Visual Storytelling:** By using visualizations like histograms, scatter plots, and heatmaps, EDA makes complex data more accessible and understandable, aiding decision-makers in interpreting the results.

To achieve these insights, EDA typically involves steps such as:

**Data Cleaning:** Addressing missing values, correcting formats, and removing outliers to ensure data quality.

**Descriptive Statistics:** Summarizing key statistics like mean, median, and standard deviation to understand the data’s central tendencies and spread.

**Visualization:** Using various charts and plots to identify trends and relationships in the data.

**Grouping and Aggregation:** Grouping data by categories like industry or size to reveal patterns and key in sights.

**4. Analysis on Dataset**

**4.1 Introduction to Dataset**

The dataset captures transaction-level information from various businesses. With a focus on financial values and business characteristics, the dataset includes attributes that allow classification and revenue tracking across different categories.

**4.2 General Description**

The dataset includes a mix of numerical and categorical variables. Here's a quick overview:

**Numerical Columns:**

* **level:** Represents a numerical grade or tier of a business or transaction.
* **value:** Financial amount associated with the transaction or record.

**Categorical Columns:**

* **line\_code:** Code representing a specific line of business.
* **industry:** Type of industry the business belongs to.
* **size:** Size of the business (e.g., Small, Medium, Large).
* **description:** Brief narrative or label.

**4.3 Specific Requirements, Functions, and Formulas**

* **Data Cleaning:**
* Replaced encoding issues (en dash in size column).
* Filled missing values using appropriate strategies:
* Mode for categorical fields (e.g., industry, line\_code, size, description).
* Mean for numerical fields (level, value).
* Removed duplicates using df.drop\_duplicates().

**Descriptive Statistics and Aggregation:**

* Used df.describe() to explore distributions.
* Aggregated value by industry, size, and line\_code.

**4.4 Analysis Results**

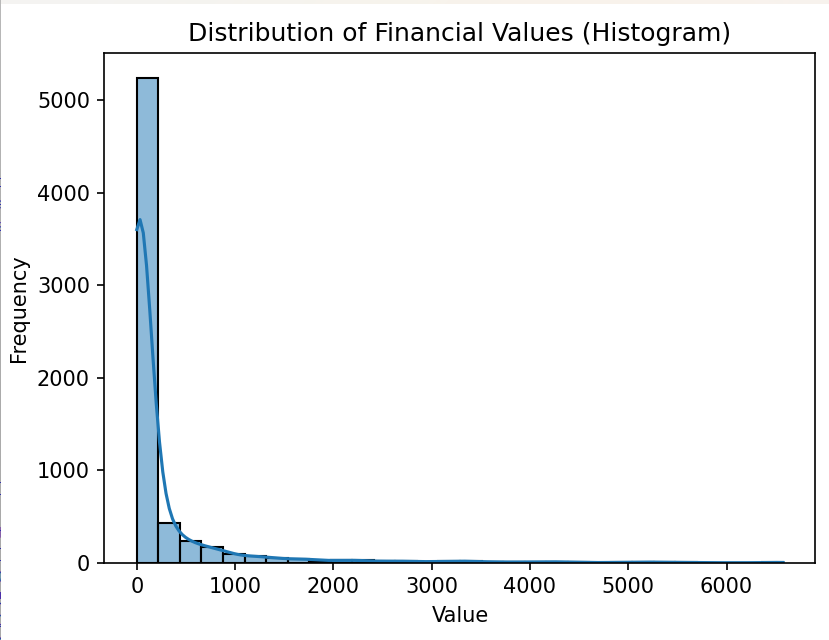
Post-cleaning, the dataset displayed improved consistency and reliability. Key insights included:

* Most industries reported a similar spread of values, although some dominated the revenue share.
* Business size played a role in determining the overall financial value.
* The value column exhibited positive skewness, indicating the presence of some high-value transactions.

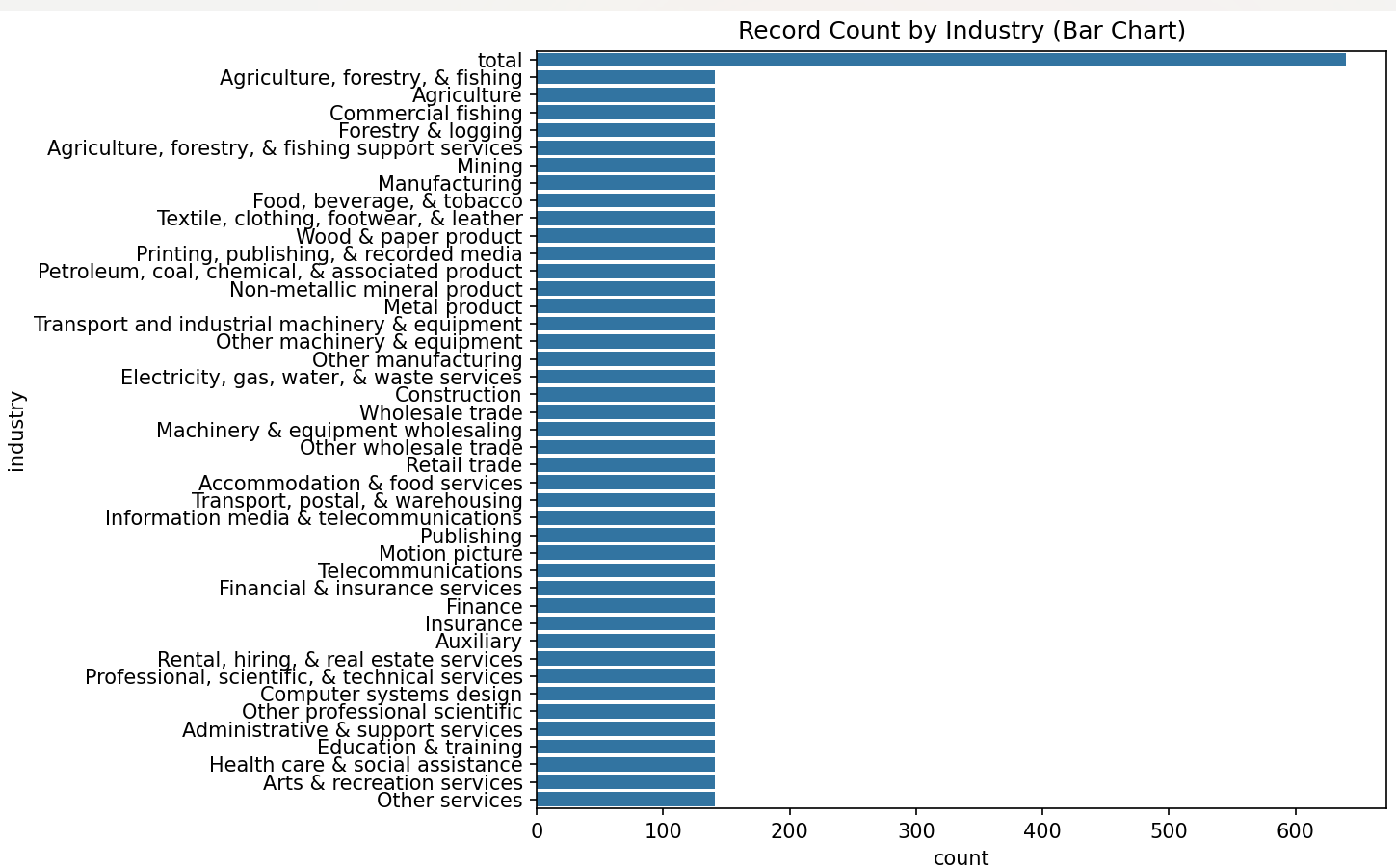
**4.5 Visualization**

You can insert your graphs in the following sections to illustrate the findings:

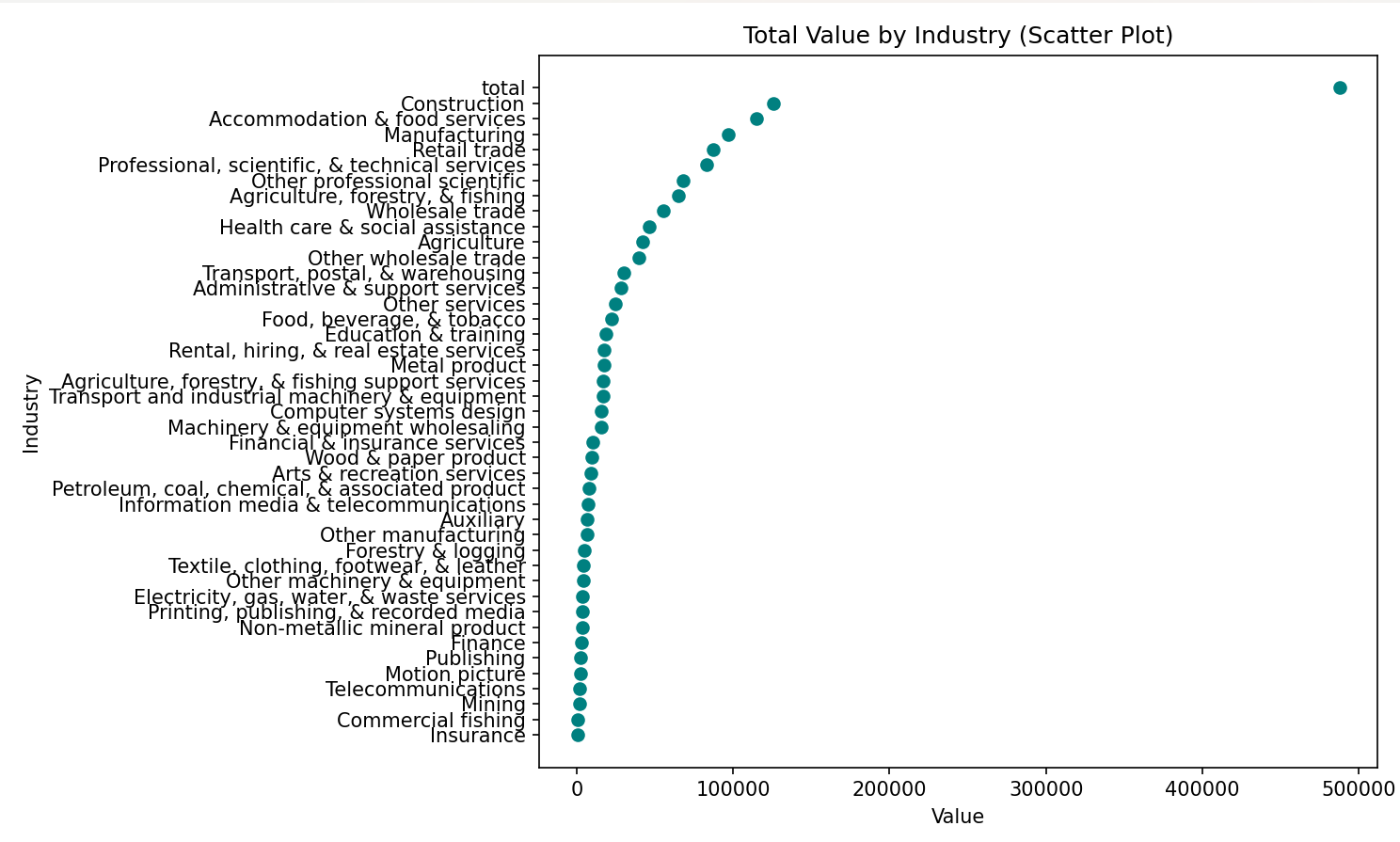
* **Histogram**: Place this graph to show the positively skewed distribution of financial values. This would go here:



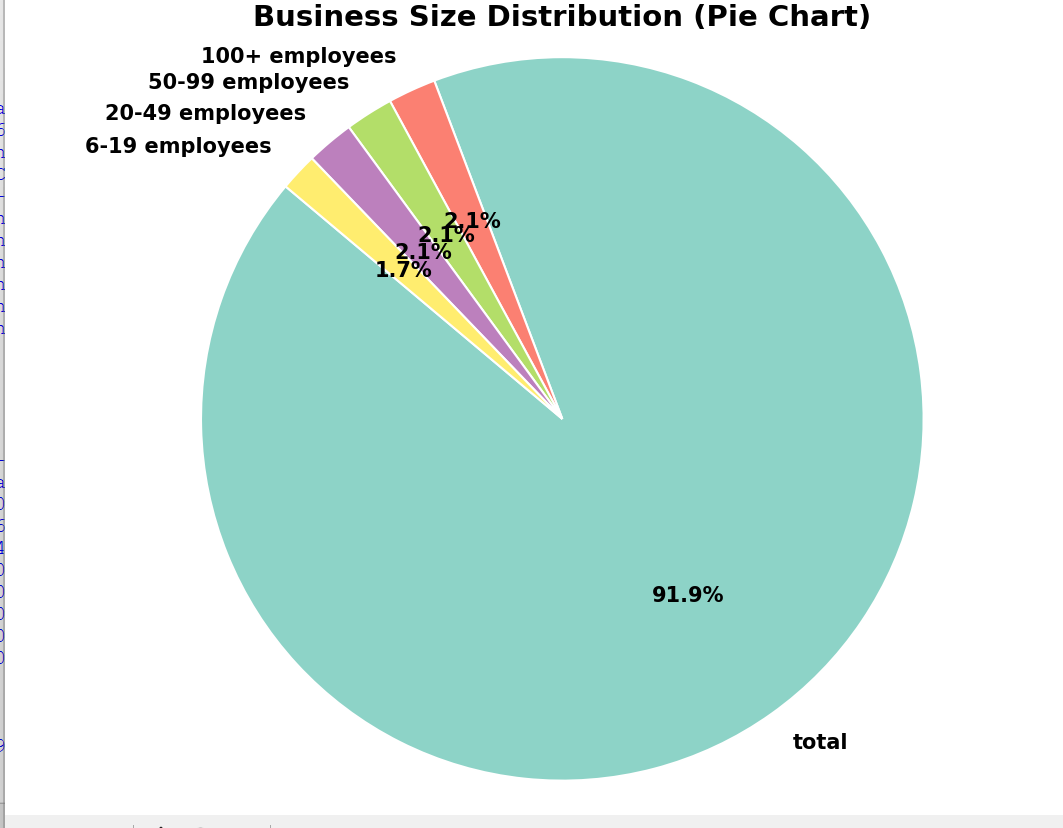
* **Bar Chart:** Demonstrates which industries have the highest number of records, helping identify high-activity sectors. This can be placed here:



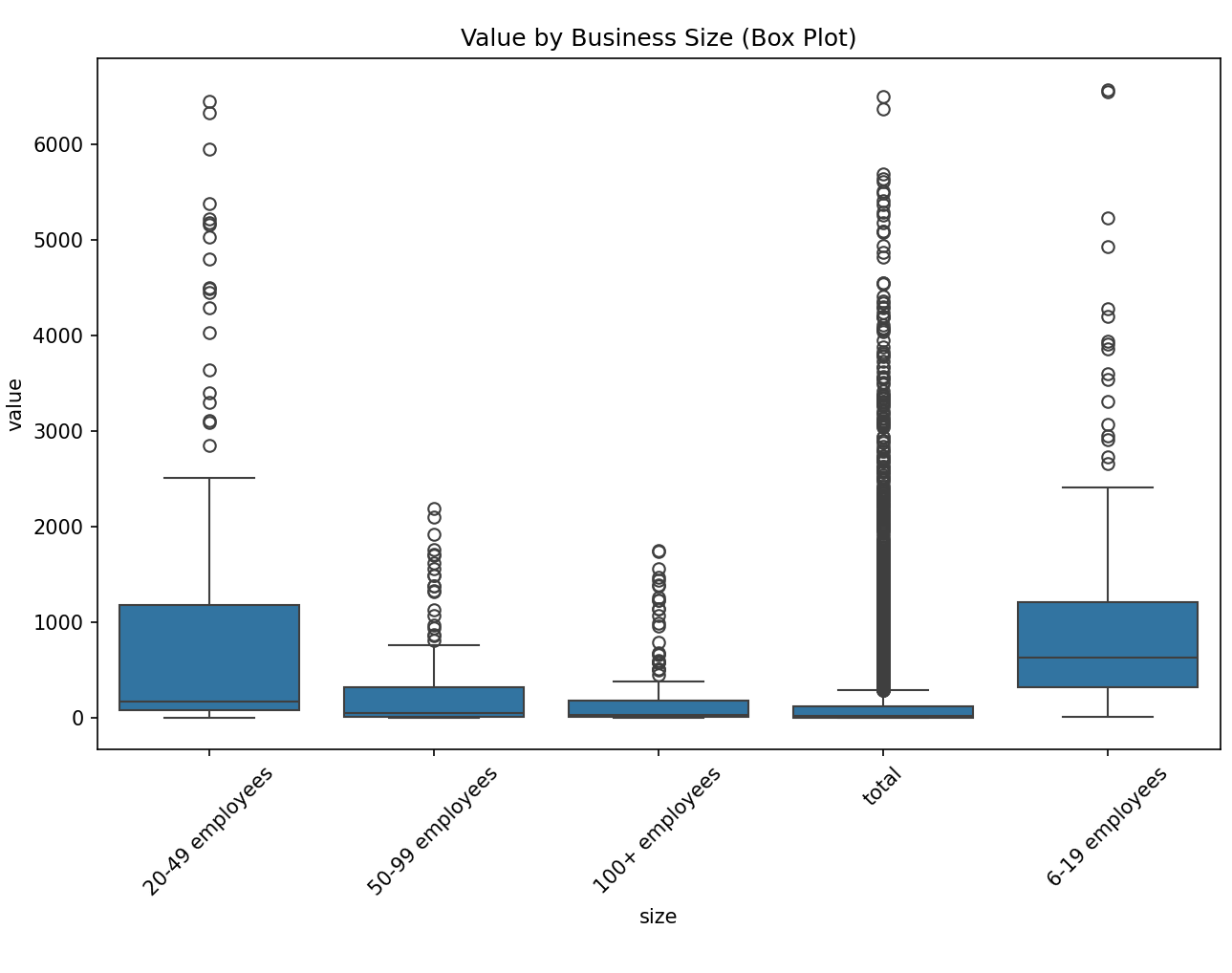
* **Scatter Plot:** Plot industry-wise aggregated values to visually separate high-performing industries. You can add it here:



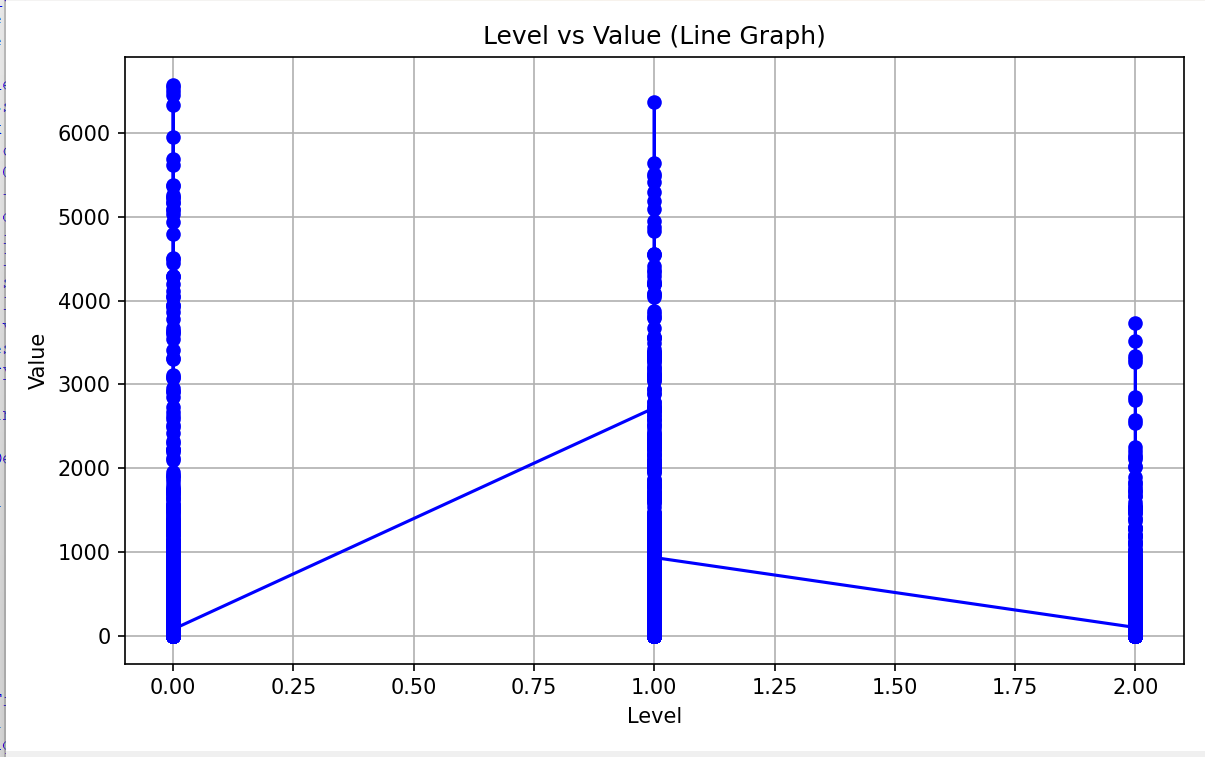
* **Pie Chart:** Reveals the proportion of business sizes, useful for segmenting strategy. This would fit here:



* **Box Plot:** Displays the interquartile range and outliers per business size. Insert this graph here:



* **Line Graph:** Offers a trend view of how value changes with level. This can be inserted here:



**5. Business Decision Support**

In this section, we leverage the insights obtained from the dataset to aid in decision-making and strategic planning for the business. By identifying key trends and patterns, we can recommend actions to enhance profitability, resource allocation, and business growth.

**Top Industries by Revenue**

* We can identify the industries that generate the most revenue using the following aggregation:
* df.groupby('industry')['value'].sum().sort\_values(ascending=False).head()
* **Actionable Insights:**
* **Prioritize marketing and outreach campaigns**: Focus on industries with the highest revenue generation, as they represent key target markets.
* **Tailor products and services:** Customize offerings to cater to the needs of high-revenue industries, ensuring the business meets their specific demands and challenges.

**Top Line Codes by Revenue**

* To identify which business lines are the most profitable, the dataset can be analyzed by line code:
* df.groupby('line\_code')['value'].sum().sort\_values(ascending=False).head(10)
* **Actionable Insights:**
* **Identify strong-performing business lines for expansion**: Focus on expanding and strengthening the lines of business with the highest revenue contribution.
* **Strategic decisions for growth:** By understanding which line codes are the most profitable, decisions on where to invest in new products or services can be made with more confidence.

**6. Conclusion**

The business dataset provides valuable insights that can significantly improve strategic planning. Through Exploratory Data Analysis (EDA), we were able to clean, process, and visualize the data to extract key trends that inform decision-making.

**Key Findings:**

* **Industries like [INSERT TOP INDUSTRY] generated the highest revenue:** Certain industries consistently contributed more to overall revenue, helping identify critical focus areas for marketing and business expansion.
* **Mid to large-size businesses were dominant in contribution:** Larger businesses generally generated higher financial values, which suggests they should be prioritized for tailored services and resources.
* **Specific line codes are repeat high-performers:** Certain business lines consistently outperformed others, revealing opportunities for expansion and increased focus on these areas.

**Implications:**

* Businesses can use these insights to optimize marketing efforts, adjust product strategies, and allocate resources more effectively.
* Focusing on high-performing industries and business sizes, and identifying profitable line codes, will drive future growth and operational efficiency.
* In conclusion, EDA has provided a structured way to transform raw data into actionable business insights, offering critical guidance for operational and strategic decision-making.

**7. References**

* pandas Documentation: <https://pandas.pydata.org/>
* matplotlib Documentation: <https://matplotlib.org/>
* seaborn Documentation: <https://seaborn.pydata.org/>
* Python Official: <https://www.python.org/>

**8. Source Code**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from scipy import stats

df = pd.read\_csv("C:\\Users\\singa\\Documents\\OneDrive\\Desktop\\python project.csv", encoding='ISO-8859-1')

# 1:Data Cleaning & Preprocessing

print(df.info())

print("\nMissing Values:")

print(df.isnull().sum())

df['size'] = df['size'].str.replace('\x96', '-', regex=True)

df['line\_code'] = df['line\_code'].fillna(df['line\_code'].mode()[0])

df['industry'] = df['industry'].fillna(df['industry'].mode()[0])

df['size'] = df['size'].fillna(df['size'].mode()[0])

df['level'] = df['level'].fillna(df['level'].mean())

df['description'] = df['description'].fillna(df['description'].mode()[0])

df['value'] = df['value'].fillna(df['value'].mean())

df.drop\_duplicates(inplace=True)

z = np.abs(stats.zscore(df[['level', 'value']]))

df = df[(z < 3).all(axis=1)]

print("\nCleaned Data Info:")

print(df.info())

print("Remaining null values:", df.isnull().sum().sum())

# 2:Exploratory Data Analysis (EDA)

print("\n--- Descriptive Statistics ---")

print(df.describe())

# Value distribution (Histogram)

plt.figure()

sns.histplot(df['value'], bins=30, kde=True)

plt.title("Distribution of Financial Values (Histogram)")

plt.xlabel("Value")

plt.ylabel("Frequency")

plt.tight\_layout()

plt.show()

# Count by industry (Bar Chart)

plt.figure(figsize=(10, 8))

sns.countplot(y='industry', data=df, order=df['industry'].value\_counts().index)

plt.title("Record Count by Industry (Bar Chart)")

plt.tight\_layout()

plt.show()

# 3:Financial Insights Extraction

total = df['value'].sum()

average = df['value'].mean()

maximum = df['value'].max()

print("\n--- Financial Insights ---")

print(f"Total Value: {total}")

print(f"Average Value: {average}")

print(f"Maximum Value: {maximum}")

# 4:Data Visualization

industry\_values = df.groupby('industry')['value'].sum().sort\_values()

# Scatter plot

plt.figure(figsize=(10, 6))

plt.scatter(industry\_values.values, industry\_values.index, color='teal') # x: values, y: industries

plt.title("Total Value by Industry (Scatter Plot)")

plt.xlabel("Value")

plt.ylabel("Industry")

plt.tight\_layout()

plt.show()

##pie chart

size\_counts = df['size'].value\_counts()

colors = plt.cm.Set3(np.linspace(0, 1, len(size\_counts)))

plt.figure(figsize=(8, 8))

plt.pie(

size\_counts,

labels=size\_counts.index,

autopct='%1.1f%%',

startangle=140,

colors=colors,

textprops={'fontsize': 10, 'fontweight': 'bold', 'color': 'black'},

wedgeprops={'edgecolor': 'white', 'linewidth': 1}

)

plt.title("Business Size Distribution (Pie Chart)", fontsize=14, fontweight='bold')

plt.axis('equal')

plt.tight\_layout()

plt.show()

##Box Plot

plt.figure(figsize=(10, 6))

sns.boxplot(x='size', y='value', data=df)

plt.title("Value by Business Size (Box Plot)")

plt.xticks(rotation=45)

plt.tight\_layout()

plt.show()

# Line Graph

df\_sorted = df.sort\_values(by='level')

plt.figure(figsize=(8, 5))

plt.plot(df\_sorted['level'], df\_sorted['value'], marker='o', color='blue')

plt.title("Level vs Value (Line Graph)")

plt.xlabel("Level")

plt.ylabel("Value")

plt.grid(True)

plt.tight\_layout()

plt.show()

# 5:Business Decision Support

print("\n--- Business Decision Support ---")

print("\nTop Industries by Revenue:")

print(df.groupby('industry')['value'].sum().sort\_values(ascending=False).head())

print("\nTop Business Sizes by Revenue:")

print(df.groupby('size')['value'].sum().sort\_values(ascending=False))

print("\nTop Line Codes by Revenue:")

print(df.groupby('line\_code')['value'].sum().sort\_values(ascending=False).head(10))