DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING (INT 375)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sr.  No. | Registration NO. | Name of Student | Roll No | Total Marks | Marks Obtained | Signature |
| 1 | 12312809 | A.Manikanta | 48 |  |  |  |



Transforming Education Transforming India

Submitted to Ms. Maneet Kaur Lovely Professional University Jalandhar, Punjab, India

|  |  |
| --- | --- |
| Delivered by: | Received by: |
| Name of Student:   1. Manikanta,   Reg no:  12312809. | Name of Faculty: Ms. Maneet Kaur UID:  15709  Signature: |

DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING (INT 375)

# PROJECT REPORT

(Project Semester January-April 2025)

# (Exploratory Data Analysis on Real Time Air Quality Index)

Submitted by

A. Manikanta Eswar Reddy Registration No:12312809

Program and Section: - CSE K23GW Course Code INT375

Under the Guidance of

**(Ms. Maneet Kaur)**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Manikanta bearing Registration no. 12312809 has completed INT375 project titled, **“**Exploratory Data Analysis on Real Time Air Quality Index**”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

# Signature and Name of the Supervisor Designation of the Supervisor

**School of Computer Science and Engineering**

Lovely Professional University Phagwara, Punjab.

Date: 12th April 2025

**DECLARATION**

I, A. Manikanta, student of Computer Science and Engineering 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.

A close up of a paper

AI-generated content may be incorrect.

Date: 12th April 2025 Signature

Reg No: 12307855 Yaswanth

# Acknowledgement

I would like to express my sincere gratitude to all those who contributed to the successful completion of this project.

First and foremost, I am deeply thankful to **Maneet kaur**, my mentor and guide, for their constant support, valuable suggestions, and encouragement throughout the duration of this project. Their insights helped me explore the dataset more meaningfully and present my findings effectively.

I would also like to thank the Department of Computer Science and Engineering, **LOVELY PROFESSIONAL UNIVERSITY**, for providing me with the necessary resources and a learning environment that made this project possible. My heartfelt thanks to my friends and classmates who supported me during the various stages of this analysis, offering valuable feedback and motivation.

Lastly, I am grateful to the creators and maintainers of the dataset, as well as the open-source community for tools like Python, Pandas, Seaborn, and Matplotlib, which played an integral role in my data analysis journey.

This project has been a great learning experience, and I truly appreciate everyone who helped make it possible.

**Table of Contents**

1. **Introduction**
2. **Source of Dataset**
3. **Exploratory Data Analysis (EDA)**
   1. Overview of the Dataset
   2. Cleaning and Preprocessing
   3. Summary Statistics

# Detailed Analysis on Dataset

* 1. Hist Plot Analysis
  2. Pie Plot Analysis
  3. Bar Plot Analysis
  4. Count Chart Analysis
  5. Box Plot Analysis
  6. Hist Chart Analysis

4.7.Heat Map

1. Conclusion
2. Future Scope
3. References
4. Acknowledgement

# Exploratory Data Analysis on Amazon shipment details

**GitHub:** [**https://github.com/yaswanthreddyyvs/Amazon\_shipment**](https://github.com/yaswanthreddyyvs/Amazon_shipment)

# Introduction.

Exploratory Data Analysis (EDA) is a critical early step in the data science process. It helps understand the structure, trends, and patterns in the dataset before applying advanced techniques or models.

This report presents EDA on a dataset related to Amazon Shipment Details. It includes histogram, bar, box, pie and heatmap visualizations, as well as outlier detection.

# Source of Dataset

The dataset used in this project is sourced from the [Data.gov].

**Format**: CSV

**Size**: 11000 records

# Key Columns:

* Mode of Shipment
* Customer Rating
* Product Cost
* Discount
* Property Type
* On Time Delivery
* Product Importance

# EDA Process

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics, often using visual methods. This section describes the step-by-step EDA process followed in this project.

* 1. Importing Required Libraries

The analysis was performed using Python with the help of libraries such as:

* + - `pandas` – for data manipulation
    - `numpy` – for numerical operations
    - `matplotlib.pyplot` and `seaborn` – for visualization
  1. Data Loading

The dataset was loaded using `pandas` from a `.csv` file. An initial overview was done using:

Python

import pandas as pd import numpy as np

import matplotlib.pyplot as plt import seaborn as sns df=pd.read\_excel('File.xlsx')

df.info() # Gives data types and null values df.describe() # Statistical summary

# Data Cleaning

* Filling NA data
* Standardized column names (removed trailing spaces).
* Converted relevant columns to numeric data types.

# 3.5 Data Visualization Setup

* Subplots were used for comparative histogram and bar plot analysis.
* Visual styles were customized using Seaborn themes and custom color palettes.

# Analysis on Dataset

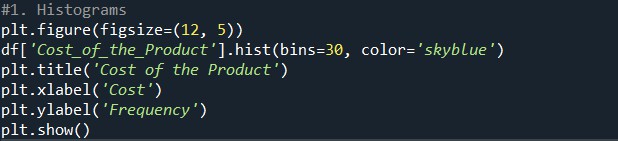
* + **Histogram:**

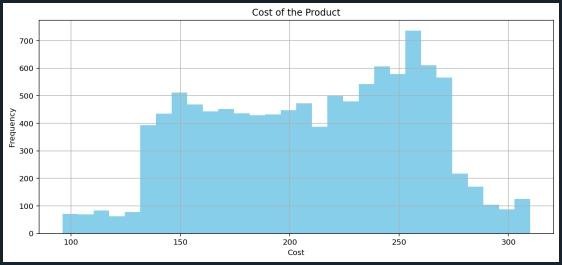
We use histograms to visualize the distribution of numerical data, showing how often values fall within specific intervals (bins). They help identify patterns, trends, or outliers, like whether data is skewed, normal, or multimodal, making it easier to understand the underlying structure of the dataset.

* + 1. **Purpose**: A histogram is ideal for visualizing the distribution of numerical data, such as product costs, to reveal patterns like the range, central tendency, spread, or skewness of the data.

**Insights**: For Product Cost, it shows how costs are distributed across different price ranges, helping identify common price points, outliers, or whether the costs are evenly spread or clustered.

**Decision-Making**: Understanding cost distribution can inform pricing strategies, inventory management, or customer targeting by highlighting which price ranges are most frequent.

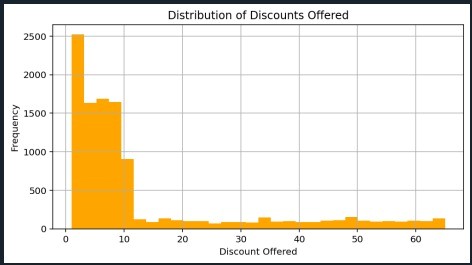
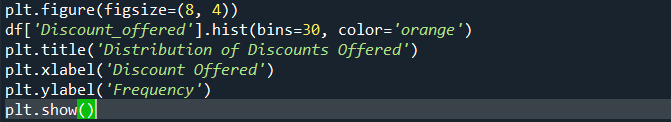




* + 1. **Purpose**: A histogram is used to visualize the distribution of numerical data, such as discounts offered, to understand how values are spread across different ranges. It reveals patterns like the most common discount levels, whether discounts are skewed, or if there are unusual outliers.

**Insights**: For Discount Offered, the histogram shows how frequently different discount amounts occur, helping identify typical discount ranges, rare high discounts, or clustering around specific values.

**Decision-Making**: Understanding discount distribution can guide marketing strategies, pricing adjustments, or promotional campaigns by highlighting which discount levels are most common or effective.



# Pie Chart

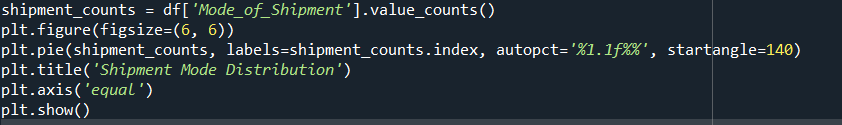
We use pie charts to visualize the proportion or percentage of categories within a dataset, showing how each category contributes to the whole.

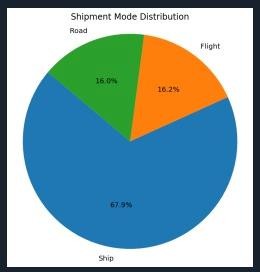
They are effective for displaying relative sizes of discrete categories, like market shares or budget allocations, when there are few categories, making it easy to compare parts to the total at a glance.

**Purpose**: A pie chart is ideal for showing the relative proportions or percentages of categorical data, such as different shipment modes (e.g., ship, truck, flight), to illustrate how each category contributes to the total.

It provides a clear, intuitive way to compare the share of each mode at a glance.

**Insights**: For Shipment Mode, the pie chart reveals which shipment modes are most or least common, highlighting the dominance or rarity of specific methods. For example, it can show if most shipments are by ship versus other modes.

**Decision-Making**: Understanding the distribution of shipment modes can inform logistics planning, cost optimization, or resource allocation by identifying which methods are used most frequently or may need adjustments.



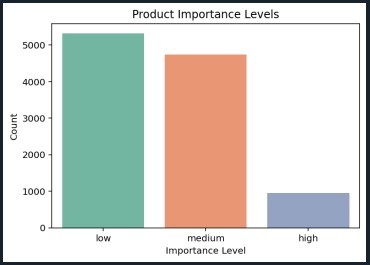
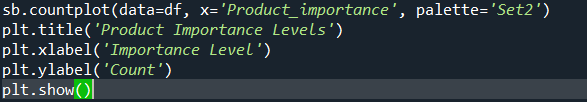
# Bar Plot

We use bar plots to compare the values of different categories or groups, displaying categorical data with rectangular bars where the length or height represents the value They are effective for showing differences across categories, like sales by region or counts of items, making it easy to compare magnitudes visually.

**Purpose:** A bar plot is ideal for visualizing and comparing the counts or values of categorical data, such as the levels of product importance .Each bar represents a category, and its height indicates the frequency or magnitude, making it easy to compare categories side by side.

**Insights:** For Product Importance, the bar plot shows how many products fall into each importance level, revealing which category is most or least common. For example, it can highlight if most products are of "low" importance or if "high" importance products are rare.

**Decision-Making:** Understanding the distribution of product importance can guide inventory management, marketing strategies, or resource allocation by identifying which types of products (based on importance) dominate the dataset or require attention.



# Count Plot

A **count plot** is a type of bar plot used to visualize the frequency of occurrences of each category in a categorical variable.

**Purpose**: A count plot is a type of bar plot that visualizes the frequency (count) of occurrences for each category in a categorical variable. It is ideal for showing how many times each category appears and can be extended with the hue parameter to compare counts across a second categorical variable.

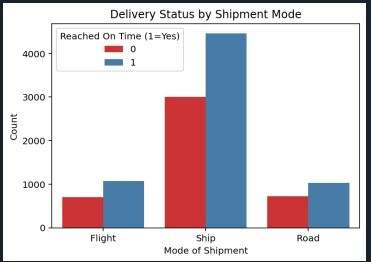
**Insights**: The plot reveals:

The popularity of each shipment mode (e.g., ship might have the most shipments).

The delivery performance for each mode (e.g., flights might have a higher on-time rate than trucks).

Patterns, such as modes with frequent delays, which could indicate logistical issues.





# Box Plot

This is also known as a box-and-whisker plot, is a graphical tool used to visualize the distribution of numerical data through its quartiles, highlighting central tendency, spread, and potential outliers. It is particularly useful for summarizing the distribution of a continuous variable and comparing distributions across categories.

**Purpose**: A box plot is used to summarize the distribution of a numerical variable through its quartiles, median, and outliers, making it ideal for comparing distributions across categories of a categorical variable. It shows the central tendency, spread, and potential outliers in the data.

# Insights:

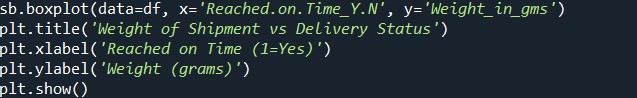
The median weight of on-time vs. late shipments Skewness

**Decision-Making:** Understanding the relationship between shipment weight and status can inform logistics strategies, such as:

Adjusting handling processes for heavier shipments if they’re prone to delays.

Optimizing transport modes for specific weight ranges to improve on- time rates.

Investigating outliers (e.g., very heavy shipments) that might require special handling.





# Correlation Heat Map

**Summarize Numerical Data**: It condenses a numerical variable into key statistics, showing how values are distributed without plotting every point.

**Compare Across Categories**: It’s great for comparing distributions across groups.

**Spot Outliers**: It clearly marks anomalies, which might indicate errors or special cases.

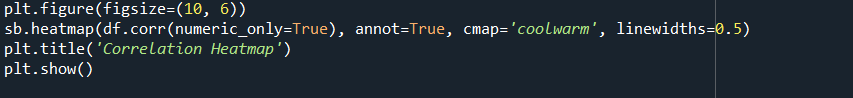
**Purpose:** A correlation heatmap displays a matrix of correlation coefficients between numerical variables, using colors to represent the strength and direction of relationships. It’s a powerful tool for exploratory data analysis to uncover patterns and dependencies.

**Decision-Making:** Understanding correlations can inform business decisions, such as:

Optimizing pricing if Cost\_of\_the\_Product and Discount\_offered are related.

Adjusting logistics if Weight\_in\_gms negatively correlates with Reached.on.Time\_Y.N

Avoiding redundant variables in predictive models (e.g., highly correlated variables might not add unique information).



# Conclusion

The visualizations collectively reveal that shipment weight and mode of transport are critical drivers of delivery performance, with heavier shipments potentially linked to delays and certain modes (e.g., ships) showing reliability issues. Pricing and discounts appear flexible but may influence delivery if not managed carefully, while product importance guides resource allocation. The correlation heatmap ties these factors together, confirming weight’s role in delays and highlighting pricing- weight relationships for optimization.

By addressing heavy shipment logistics, diversifying transport modes, refining pricing/discount strategies, and prioritizing high-importance products, the business can enhance on-time delivery rates, optimize costs, and improve customer satisfaction. These visualizations provide a data- driven foundation for targeted improvements, ensuring logistics efficiency and competitive pricing in the shipment process.

If you’d like a deeper focus on any specific aspect (e.g., predictive modeling, mode optimization, or outlier analysis), let me know, and I can expand further!

.

# Future Scope

The future scope for this shipment dataset is vast, building on the descriptive insights from your visualizations:

* + **Advanced Analytics**: Predictive models, clustering, and time-series analysis can forecast delays, segment shipments, and uncover trends, extending the **heatmap** and **box plot** findings.
  + **Operational Enhancements**: Optimizing mode allocation, weight handling, pricing, and product prioritization can directly address delays and inefficiencies, leveraging **pie chart**, **count plot**, and **histogram** insights.
  + **Technological Integration**: Real-time dashboards, automated systems, external data, and AI anomaly detection can scale insights into dynamic, proactive solutions.
  + **Business Expansion**: Improving customer experience, supply chain efficiency, and sustainability can broaden the impact beyond logistics, aligning with cost and weight patterns from **heatmap** and **box plot**.

# References

* 1. **Pandas Documentation** – Python Data Analysis Library <https://pandas.pydata.org/>
  2. **NumPy Documentation** – Scientific Computing with Python <https://numpy.org/>
  3. **Matplotlib Documentation** – Comprehensive 2D Plotting <https://matplotlib.org/>
  4. **Seaborn Documentation** – Statistical Data Visualization <https://seaborn.pydata.org/>
  5. **Scikit-learn Documentation** – Machine Learning in Python <https://scikit-learn.org/stable/>
  6. **Jupyter Project** – Interactive Python Notebook Environment <http://localhost:8888/notebooks/python.ipynb>