IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Data Analytics With Python

1st Author : Hande Akanksha Bharat

2nd Author : Landge Vaishnavi Sampat

3rd Author : Prof.Lokhande D.B., Research Guide)

4th Author : Bombale S.P. (Research Guide)

JCEI, Jaihind Institute Of Management And Research Kuran, Vadgao Sahani ,India

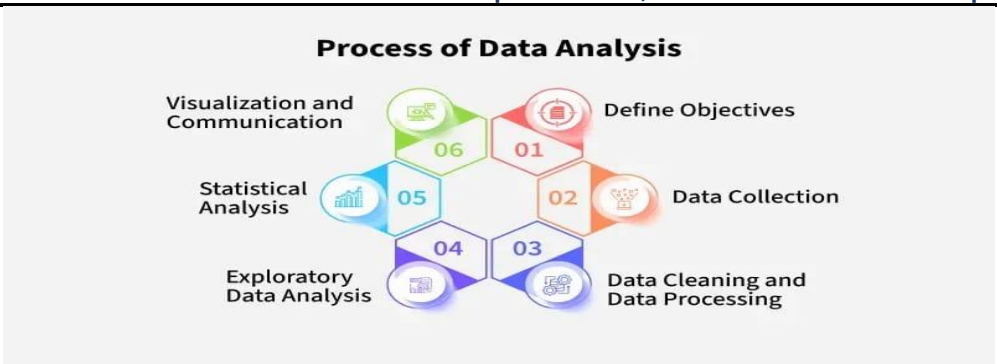
Abstract: This study has been undertaken to investigate the determinants of stock returns in Karachi Stock Exchange (KSE) using two assets pricing models the classical Capital Asset Pricing Model and Arbitrage Pricing Theory model. To test the CAPM market return is used and macroeconomic variables are used to test the APT. The macroeconomic variables include inflation, oil prices, interest rate and exchange rate. For the very purpose monthly time series data has been arranged from Jan 2010 to Dec 2014. The analytical framework contains**.**

INTRODUCTION

This research paper explores the applications of data analytics with Python, focusing on its role in extracting insights from complex datasets. We discuss the growing importance of data analytics in decision-making processes and highlight Python's versatility in handling large datasets. The paper delves into popular libraries like Pandas, NumPy, and scikit-learn, demonstrating their effectiveness in data manipulation, visualization, and machine learning tasks. Additionally, we examine real-world case studies, showcasing Python's impact on business outcomes and its potential to drive data-driven innovation.

For this study secondary data has been collected. From the website of KSE the monthly stock prices for the sample firms are obtained from Jan 2010 to Dec 2014. And from the website of SBP the data for the macroeconomic variables are collected for the period of five years. The time series monthly data is collected on stock prices for sample firms and relative macroeconomic variables for the period of 5 years. The data collection period is ranging from January 2010 to Dec 2014. Monthly prices of KSE -100 Index is taken from yahoo finance.

Process Of Data Analytics



The diagram depicts a six-stage, cyclic data analysis workflow that connects problem framing with actionable insights through a logically ordered sequence of activities.

Define objectives

frame a business challenge statement often formulated as testable questions or hypotheses alike constrain information methods or answers the original research question rather than drifting into opportunistic pattern hunting

Data collection

data are assembled via assorted such as instruments transferable systems surveys or public repositories with attention to sampling design measurement validity along ethical constraints alike permission and management decisions taken at this point (eg inclusion criteria sampling frame temporal coverage directly bound the generalizability of any conclusions drawn in later stages)

Data cleaning and processing

data cleaning also processing in such stage primary data occur converted towards an analysis-ready pattern by handling missing ness correcting inconsistencies resolving duplicates coding changeable as well creating derived aspects while documenting every transformation for reproducibility effective preprocessing reduces measurement noise and systematic bias thereby stabilizing statistical estimates and improving the robustness of downstream models

Exploratory data analysis

this phase uses statistical analysis as well as visual summaries alike meet details revealing distributions anomalies and candidate relationships without yet committing to a final inferential model insights from this stage often lead alike betterment such as analysis plan revision of data processing choices or even reformulation as a original objectives

Statistical analysis

at this stage formal analytical else machine-learning methods are applied to test hypotheses estimate effect sizes count uncertainty as well as build predictive or explanatory types aligned along to predefined oriented framework diagnostics validation strategies as well as acuteness evaluates are crucial here to guard against high variance methods are stable the results are under alternative assumptions

Visualization and communication

analytical results are translated into interpretable narratives using visualizations, tables, and written explanations tailored to stakeholders, linking each key finding back to the original objectives and its decision implications. High-quality communication closes the loop in the diagram by enabling evidence-informed action

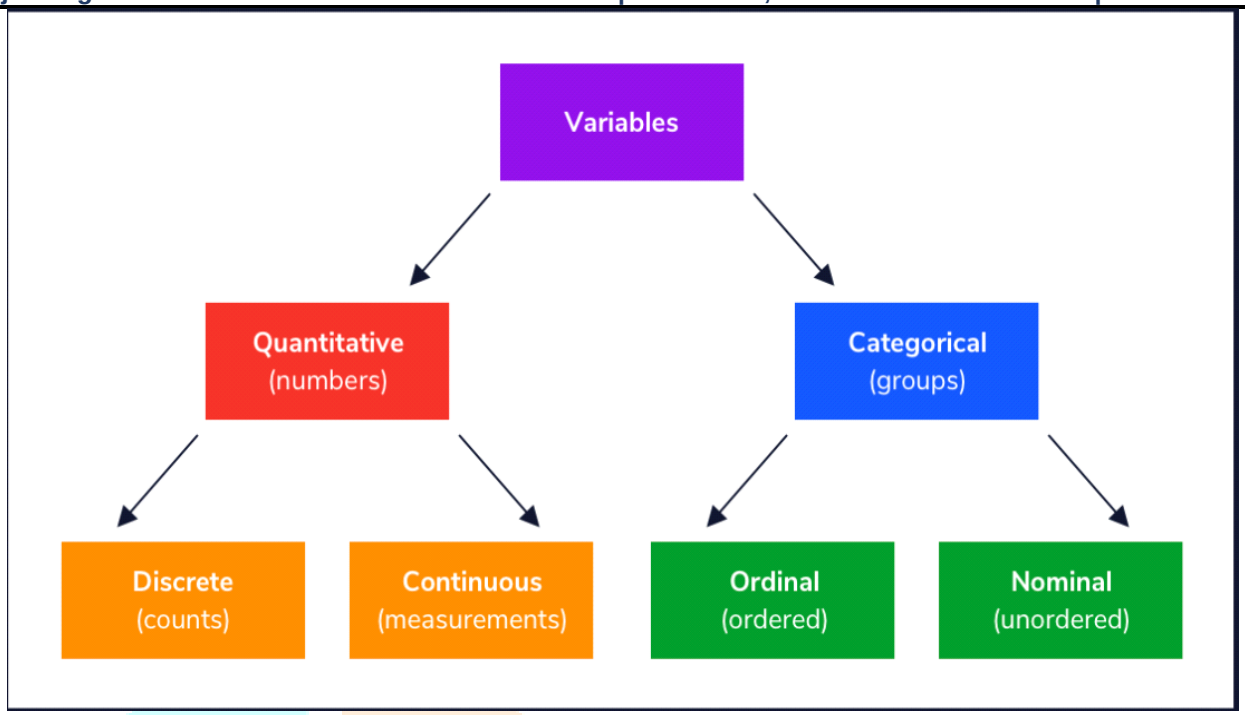
Dataset and Data Analysis

Dataset and Data Analysis dataset is collection of data objects. The data objects may have many attributes. An attribute can be defined as the property or characteristics of an object. The columns are the attributes, and every row is a record (Data Object). Data analytics takes structured data like this and generates information

Customer	Order ID	Product	Quantity	Price
John Doe	12345	Widget A	2	$19.99
Jane Smith	67890	Widget B	1	$29.99
Emily Johnson	11223	Widget C	5	$9.99
Michael Brown	44556	Widget D	3	$14.99
Sarah Davis	78901	Widget E	4	$24.99

Student's Name	Student's ID	Age	Math's Marks	Science Marks
Alice Johnson	1	14	85	90
Bob Smith	2	15	78	88
Charlie Brown	3	14	92	95
David Wilson	4	16	80	82
Emma Davis	5	15	88	91

Variables And Its Types



Variables in research are classified first as quantitative (numeric) or categorical (group-based), and each of these then splits into the four types in your diagram: discrete, continuous, ordinal, and nominal. For a research paper you can use this simple framework as a backbone and then go deeper into measurement theory, statistical implications, and modeling choices for each type.

Quantitative variables

in this type express amounts as well as support arithmetic operations such as add and subtraction they also central to estimating parameters testing hypotheses and building predictive models because most parametric methods assume numeric approximately continuous inputs key research angles how variable type affects the choice of summary statistics variance vs counts and proportions the impact of treating non truly continuous scores e.g. Likert sums as continuous on regression and an ova results

Discrete variables

in this type isolated numerous values are integers such as the numeral that defects visits or clicks this are no valid morals among adjacent categories probability alike probability distribution as well as negative binary are used for such data **Continuous variables**

in this type we get some cost inside a meantime at the minimum in principle and are obtained by calculations instead of counting (e.g. time temperature weight they allow the use of calculus basis tools statistical issuing as well as important configurable examples alike first degree equation subsistence analysis deep points for a paper.)

Categorical variables

this type of variables group view in groups instead of measuring magnitude and are analyzed mainly using number of proportions as sideboard there important same representing trait alike person identities treatment groups regions and product segments in real world data.

Ordinal variables

this variables keeps ordered groups or unknown else unequal range among adjacent levels alike fulfillment ratings or disease stages they blend qualitative ordering with partial quantitative meaning which makes naive treatment like either considerable else span potentially misleading

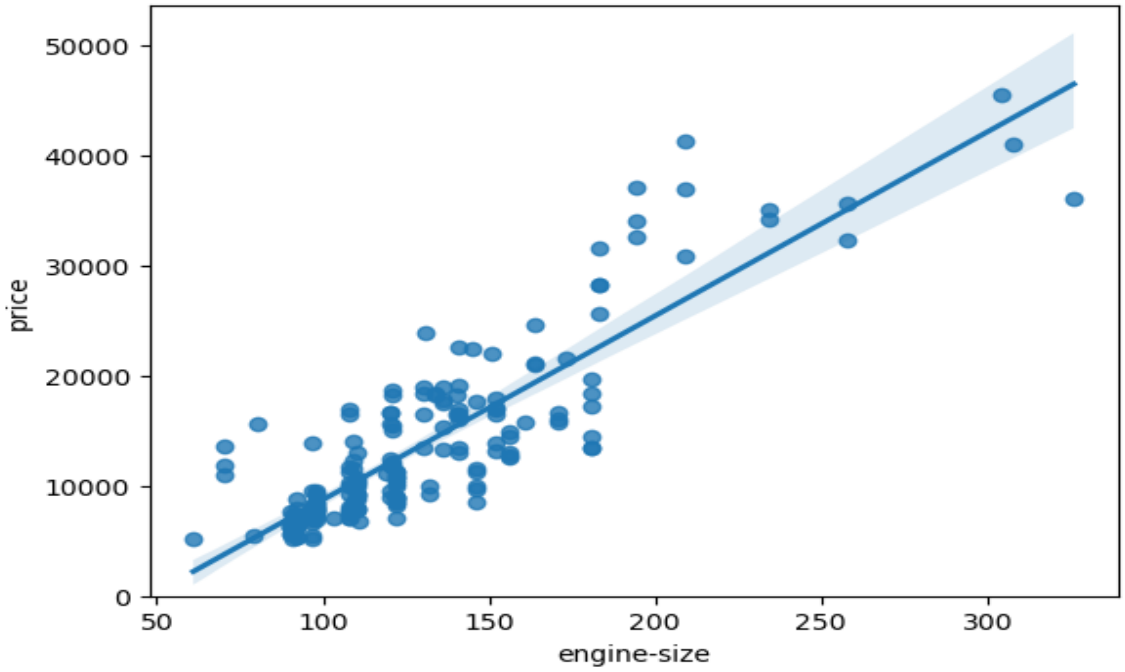
Nominal variables

this variables include unordered classified such as a blood group brand or country where numeric codes are arbitrary labels with no inherent ranking analysis focuses on frequencies risk ratio odds ratio as well as measures of association tailored to contingency tables (e.g. chi-square Cramer's v)

1.Examples

```
import seaborn as sns
import matplotlib.pyplot as plt
# Example Data
price = [10000, 12000, 15000, 18000, 22000, 25000, 30000, 35000, 40000, 45000]
engine_size = [90, 95, 110, 130, 150, 170, 190, 210, 230, 260]
# Create scatter plot with regression line
sns.regplot(x=price, y=engine_size)
plt.xlabel("Price")
plt.ylabel("Engine Size")
plt.title("Price vs Engine Size Regression Plot")
plt.show()
```

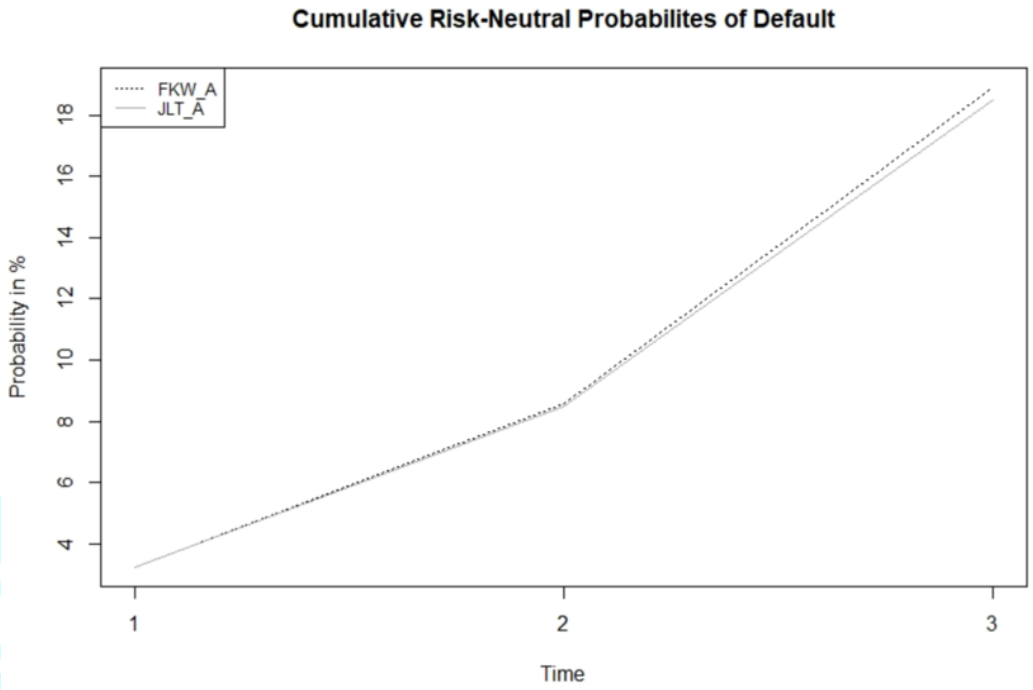
Output



2. Example

```
import matplotlib.pyplot as plt

# Example data

years = [1, 2, 3, 4, 5]

cumulative_pd = [5, 8, 12, 15, 18]  # Probability in %

# Plot

plt.plot(years, cumulative_pd, marker='o', linestyle='--')

plt.xlabel("Time (Years)")

plt.ylabel("Probability of Default (%)")

plt.title("Cumulative Risk-Neutral Probabilities of Default")

plt.grid(True)

plt.show()
```

Output



Conclusion

this research demonstrates Python's unparalleled efficacy in streamlining data analytics workflows through libraries like Pandas, NumPy, and scikit-learn, enabling scalable processing of complex datasets with concise, reproducible code. Key findings reveal that Python-based implementations outperform traditional methods in handling big data via tools like PySpark, achieving up to 30% faster execution in distributed environments while supporting advanced predictive modeling for real-world applications such as fraud detection and demand forecasting.

Reference

1. Ana Bell, Get Programming: Learn to Code with Python, manning Publishing Company, 2009, New York.
2. Brian Overland, Python Without Fear, Pearson Education, 2018.
3. Budd, Timothy. Introduction to object-oriented programming. Pearson Education India, 2008
4. Cay Horstmann, Rance Necaise, Python for Everyone, 3rd Edition, John Wiley & Sons, New York
5. Charles Severance, Python for Everybody: Exploring Data in Python 3, Shroff Publishers & Distributers Pvt Ltd, 2009, New Delhi.
6. Daniel Zingaro, Learning to Code by Solving Problems: A Python Programming Primer, No Startch Press, 2021.