SuperStore Data Analysis Using Data Analytics and Data Science

SUPERSTORE DATA ANALYSIS USING AI WITH DATA SCIENCE

Presented By
Student Name: S.SHRRIVATHSAN
College Name: Panimalar ENGINEERING COLLEGE
Department: COMPUTER SCIENCE DEPARTMENT
Email ID: shrrivathsany2k@gmail.com
AICTE Student ID: STU680f6b03bd1fe1745840899

Internship ID:
INTERNSHIP_1748937226683eaa0a58abc

edunet foundation

OUTLINE

- **Problem Statement** (Should not include solution)
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Algorithm & Deployment**
- **Result (Output Image)**
- **Conclusion**
- **Future Scope**
- **References**

PROBLEM STATEMENT

Problem Statement:

- The objective is to analyze the historical sales and profit data of a superstore to identify key trends, patterns, and anomalies, and subsequently leverage these insights to optimize business operations, enhance customer satisfaction, and maximize profitability. This involves using data science techniques, including descriptive analytics, predictive modeling, and potentially prescriptive analytics, to address specific business questions and inform strategic decision-making.

Key objectives

- Identify key drivers of sales and profitability: Understanding which factors contribute most significantly to sales and profit across different products, regions, and customer segments.
- Forecast future sales and demand: Accurately predict sales and demand for various products to optimize inventory levels and prevent stockouts or overstocking, which can reduce waste and save costs.
- Segment customers and personalize recommendations: Identify distinct customer segments based on their preferences and buying behavior to tailor marketing campaigns and product recommendations, enhancing customer satisfaction and loyalty.
- Identify underperforming products and categories: Pinpoint products or categories experiencing low sales or profitability to develop strategies for improvement, such as adjusting pricing, promotional activities, or exploring alternative sourcing.
- Improve shipping and logistics efficiency: Analyze shipping mode performance and optimize logistics to reduce costs and delivery times, ultimately improving customer satisfaction and retention.

Data science approach and AI techniques

The project will involve the following steps and techniques:

- **Data Collection and Cleaning:**
  - Gather historical sales data, including product information, customer details, order history, and shipping information.
  - Clean and preprocess the data to handle missing values, inconsistencies, and outliers, ensuring data quality for analysis and model building.
- **Exploratory Data Analysis (EDA):**
  - Conduct EDA to understand the data structure, distributions, and initial trends in sales, profit, and other key metrics across various dimensions (categories, regions, segments, etc.).
  - Utilize visualizations like bar charts, scatter plots, and geographical maps to uncover patterns, disparities, and anomalies in the sales data.
- **Feature Engineering:**
  - Create new features from existing data, such as seasonality indicators, discount ratios, or customer lifetime value metrics, to enhance the predictive power of models.

- Model Building and Training:
  - Sales Forecasting: Employ AI and machine learning models like time series forecasting (e.g., ARIMA or Prophet) to predict future sales based on historical data and seasonal patterns.
  - Customer Segmentation: Apply unsupervised learning algorithms (e.g., K-Means clustering) to segment customers based on purchase history, demographics, and behavior.
  - Product Recommendation: Build personalized recommendation systems using collaborative filtering or content-based filtering techniques to suggest products to customers based on their preferences and browsing history.
  - Profitability Analysis: Utilize regression models or machine learning algorithms to identify factors influencing profitability and predict potential losses or areas of improvement.
- Model Evaluation:
  - Evaluate the performance of the developed models using appropriate metrics, such as Mean Absolute Percentage Error (MAPE) for forecasting or accuracy/precision/recall for classification tasks.
- Deployment and Monitoring:
  - Deploy the models into a production environment for continuous analysis and prediction.
  - Establish a monitoring system to track model performance and ensure the generated insights remain accurate and relevant over time.

PROPOSED SOLUTION

- The proposed system aims to address the challenge of predicting the required bike count at each hour to ensure a stable supply of rental bikes. This involves leveraging data analytics and machine learning techniques to forecast demand patterns accurately. The solution will consist of the following components:
- Data Collection:
  - Gather historical data on bike rentals, including time, date, location, and other relevant factors.
  - Utilize real-time data sources, such as weather conditions, events, and holidays, to enhance prediction accuracy.
- Data Preprocessing:
  - Clean and preprocess the collected data to handle missing values, outliers, and inconsistencies.
  - Feature engineering to extract relevant features from the data that might impact bike demand.
- Machine Learning Algorithm:
  - Implement a machine learning algorithm, such as a time-series forecasting model (e.g., ARIMA, SARIMA, LSTM, XGBOOST), to predict bike counts based on historical patterns.
  - Consider incorporating other factors like weather conditions, day of the week, and special events to improve prediction accuracy.
- Deployment:
  - Develop a user-friendly interface or application that provides real-time predictions for bike counts at different hours.
  - Deploy the solution on a scalable and reliable platform, considering factors like server infrastructure, response time, and user accessibility.
- Evaluation:
  - Assess the model's performance using appropriate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or other relevant metrics.
  - Fine-tune the model based on feedback and continuous monitoring of prediction accuracy.

SYSTEM APPROACH

The "System Approach" section outlines the overall strategy and methodology for developing and implementing the rental bike prediction system. Here's a suggested structure for this section:

- System requirement's
- Library required to build the model

SYSTEM REQUIREMENTS

1. Hardware Requirements:

- Processor: Intel i5/i7 or equivalent (multi-core recommended)
- RAM: Minimum 8 GB (16 GB recommended for faster processing)
- Storage: Minimum 1 GB free space (for dataset and libraries)

2. Software Requirements:

- OS: Windows 10/11, macOS, or any Linux distribution
- Python version: 3.8 or above
- IDE: Jupyter Notebook, VS Code, or Google Colab

LIBRARY REQUIRED TO BUILD THE MODEL

1. Data Handling & Manipulation

- PANDAS – For reading and manipulating tabular data
- NUMPY – For numerical operations

2. Data Visualization

- MATPLOTLIB – For basic plotting and visualization
- SEABORN – For advanced statistical plots

3. Data Preprocessing

- SKLEARN.PREPROCESSING – For encoding, scaling, imputing data

4. Model Building

- SKLEARN.LINEAR_MODEL – For linear regression models
- SKLEARN.ENSEMBLE – For models like Random Forest, Gradient Boosting
- XGBOOST – For extreme gradient boosting (if needed)
- LIGHTGBM – For faster and efficient boosting (optional)

LIBRARY REQUIRED TO BUILD THE MODEL

5. Model Evaluation

- SKLEARN.METRICS – For performance evaluation metrics like RMSE, R^2, MAE

6. Optional (Advanced Visualization & Tuning)

- PLOTLY – For interactive charts
- SCIPY – For statistical operations
- SKLEARN.MODEL_SELECTION – For cross-validation and hyperparameter tuning

ALGORITHM & DEPLOYMENT

1. Project Objective

- Analyze the Superstore dataset to:
- Understand sales performance
- Identify top-performing segments, products, and regions
- Predict future sales
- Detect anomalies (e.g., loss-generating items)
- Provide decision support via a deployed web application

2. Data Science Workflow`

A. Exploratory Data Analysis (EDA)

- Visualize sales, profit by region, category, and customer segment
- Heatmaps for correlation
- Time series of sales over time
- Identify loss-making sub-categories

B. Feature Engineering

- Extract year, month from Order Date
- Compute Profit Margin
- Aggregate metrics by customer/product/region

C. Machine Learning Algorithms

Choose ML algorithms based on your goals:

1. Sales Prediction

- Algorithms: Linear Regression, Random Forest, XGBoost
- Input: Product, Region, Time, Discount, Quantity
- Output: Expected Sales/Profit

ALGORITHM & DEPLOYMENT

2. Customer Segmentation
•Algorithm: K-Means Clustering
•Features: Total Sales, Orders, Average Profit
•Outcome: Targeted marketing

3. Anomaly Detection
•Algorithms: Isolation Forest, DBSCAN
•Detect: Unusual profit margins, high discounts

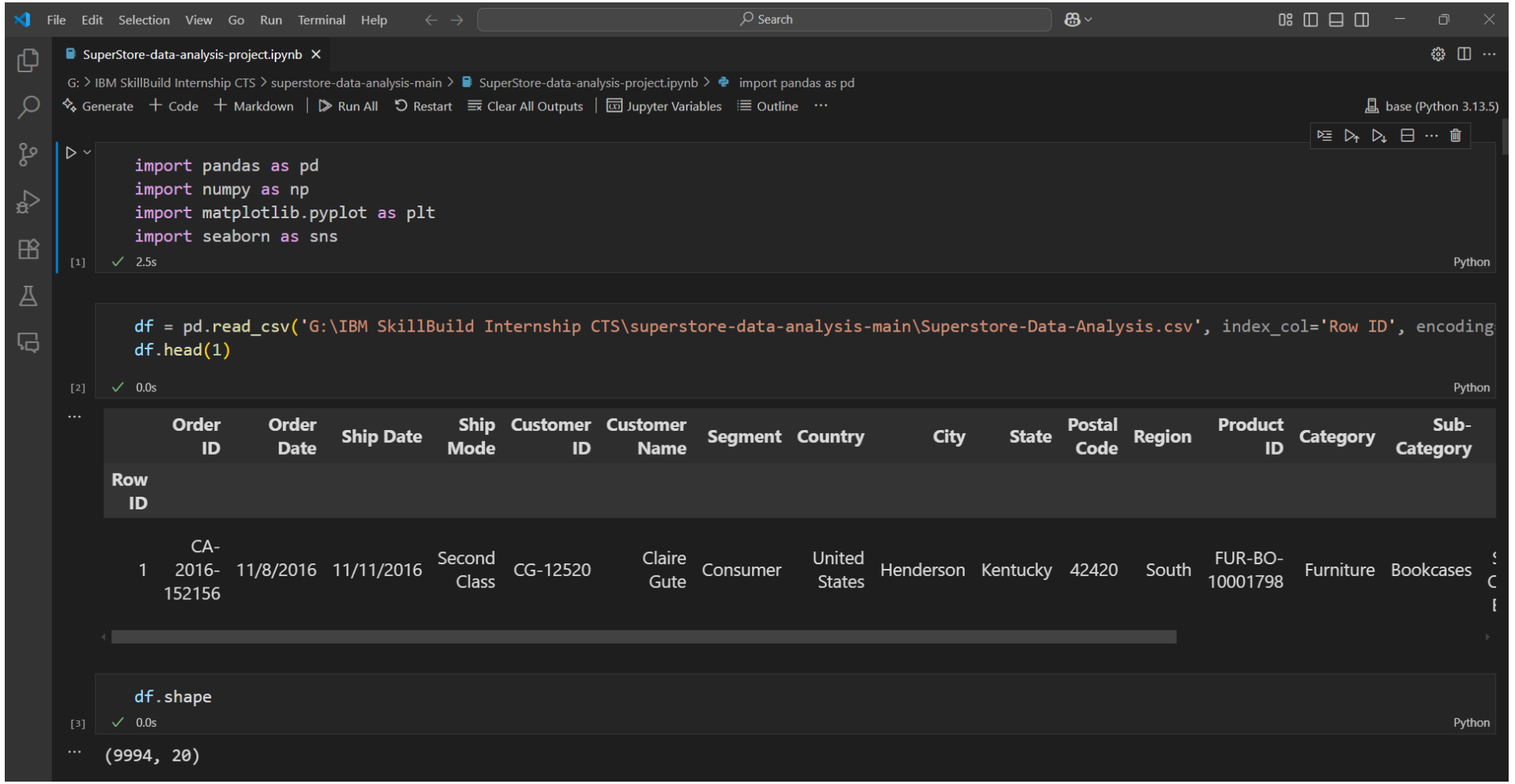
4. Product Recommendation
•Algorithm: Association Rules (Apriori), Collaborative Filtering

- 4. AI Integration (Optional)
- Use **LSTM (Deep Learning)** for **time-series sales forecasting**
- Integrate **NLP** to generate insights like: “Top 5 loss-making products in the West region in 2024”.

5. Tools and Libraries

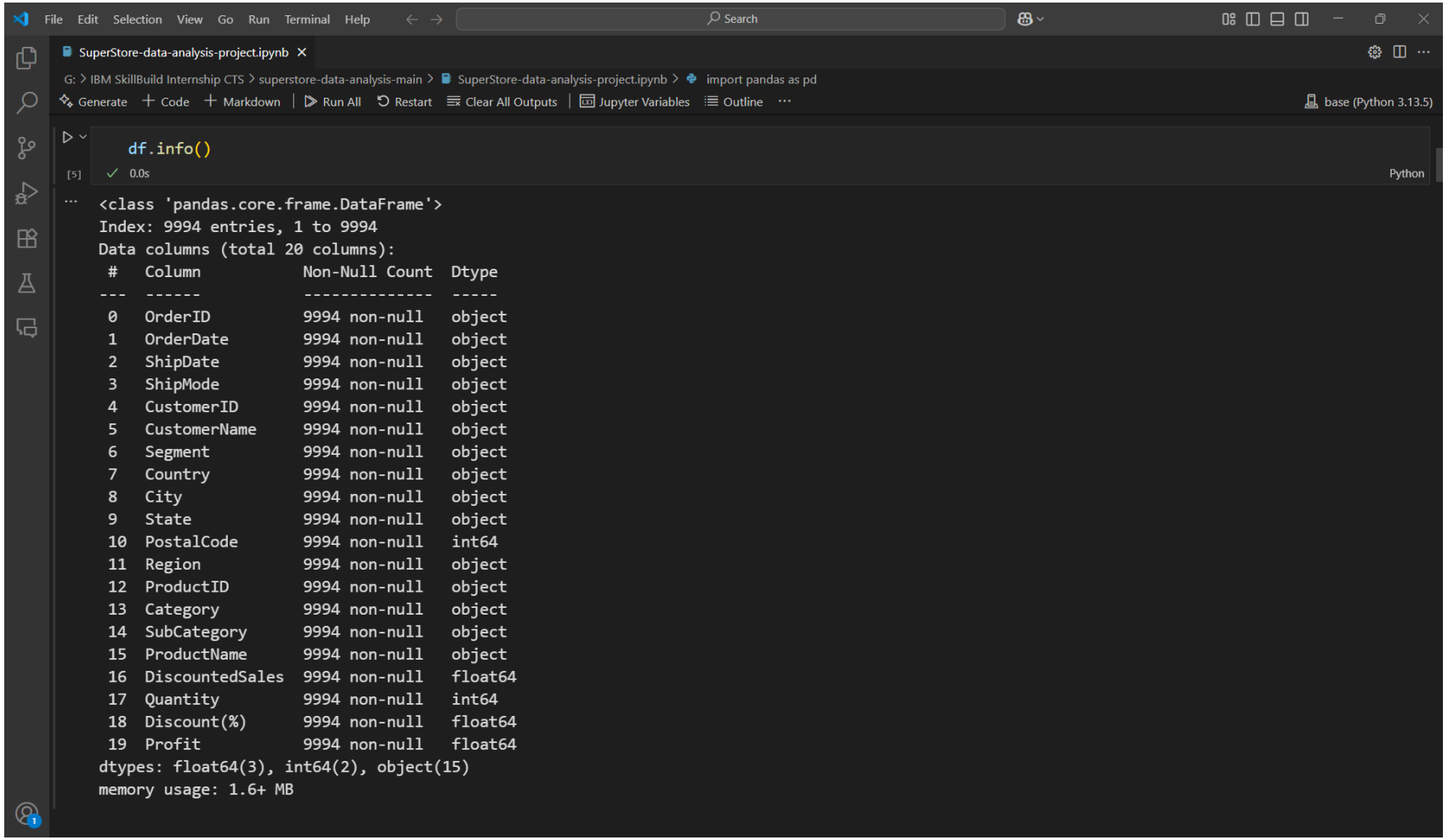
Python Libraries
•pandas, numpy, matplotlib, seaborn – for EDA
•scikit-learn – ML algorithms
•xgboost, lightgbm – advanced models
•statsmodels or fbprophet – time series forecasting
•flask or streamlit – for web deployment

RESULTS AND SCREENSHOTS

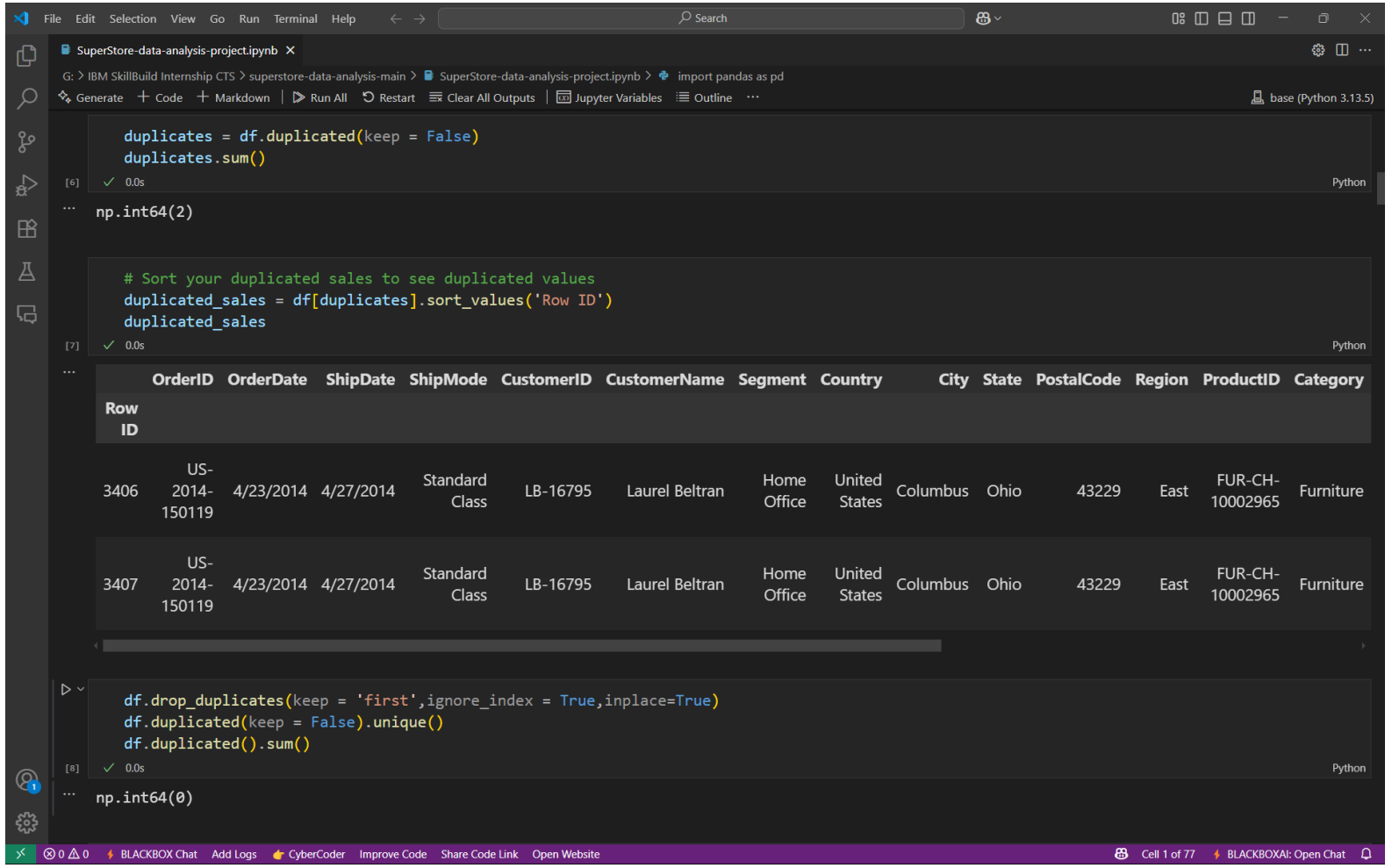


edunet foundation

RESULTS AND SCREENSHOTS



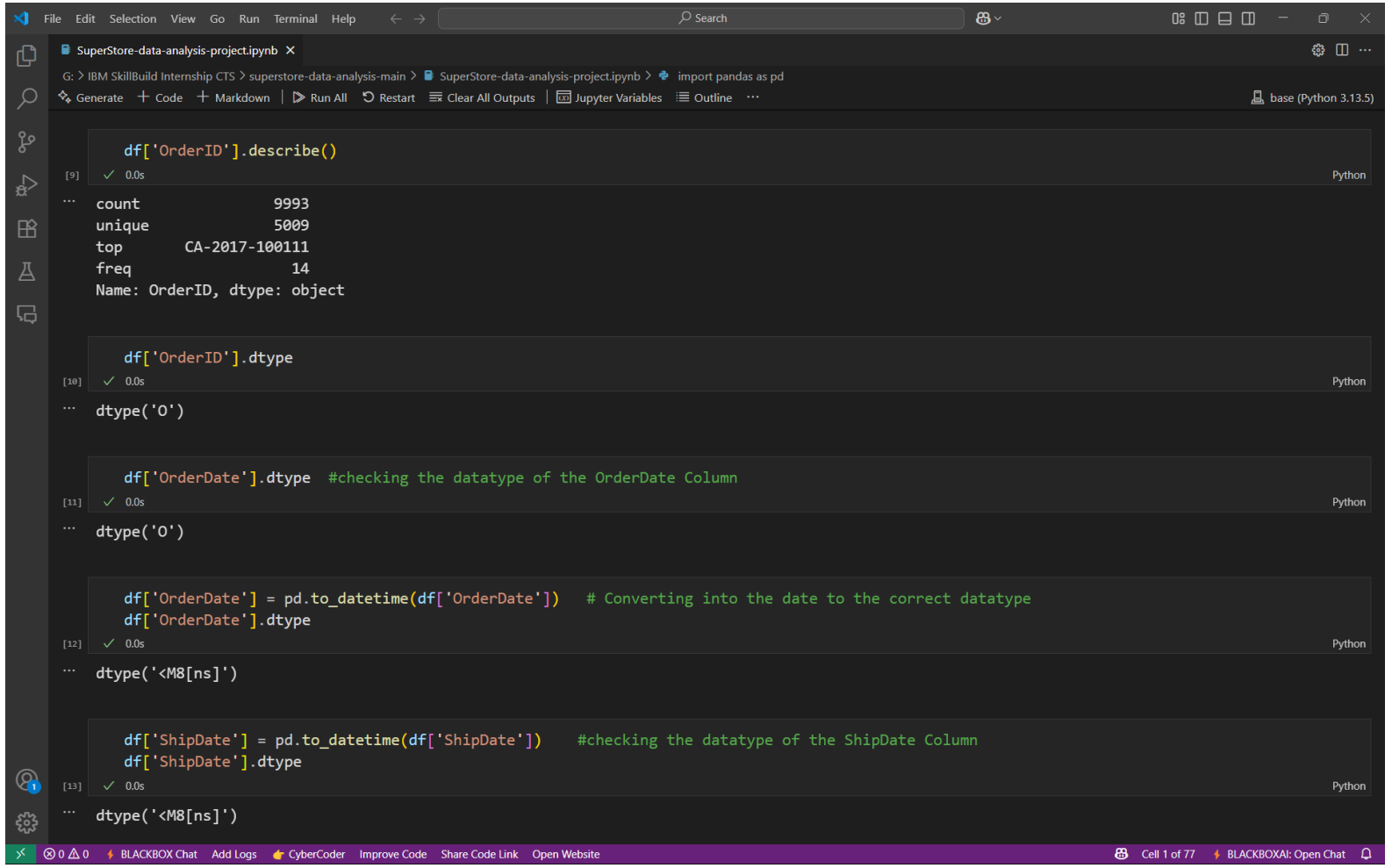
RESULTS AND SCREENSHOTS



Row ID	OrderID	OrderDate	ShipDate	ShipMode	CustomerID	CustomerName	Segment	Country	City	State	PostalCode	Region	ProductID	Category
3406	US-2014-150119	4/23/2014	4/27/2014	Standard Class	LB-16795	Laurel Beltran	Home Office	United States	Columbus	Ohio	43229	East	FUR-CH-10002965	Furniture
3407	US-2014-150119	4/23/2014	4/27/2014	Standard Class	LB-16795	Laurel Beltran	Home Office	United States	Columbus	Ohio	43229	East	FUR-CH-10002965	Furniture

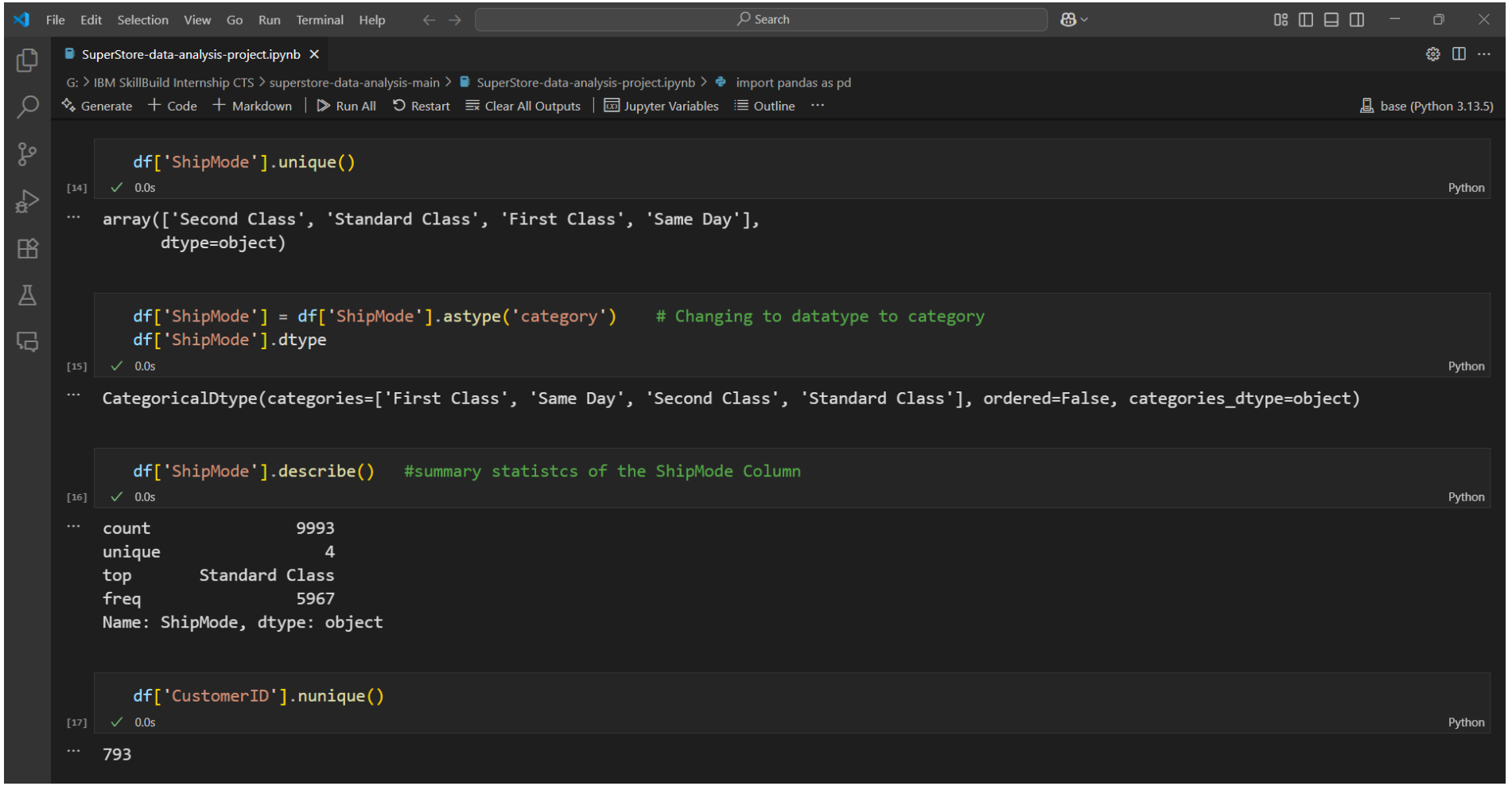
edunet foundation

RESULTS AND SCREENSHOTS

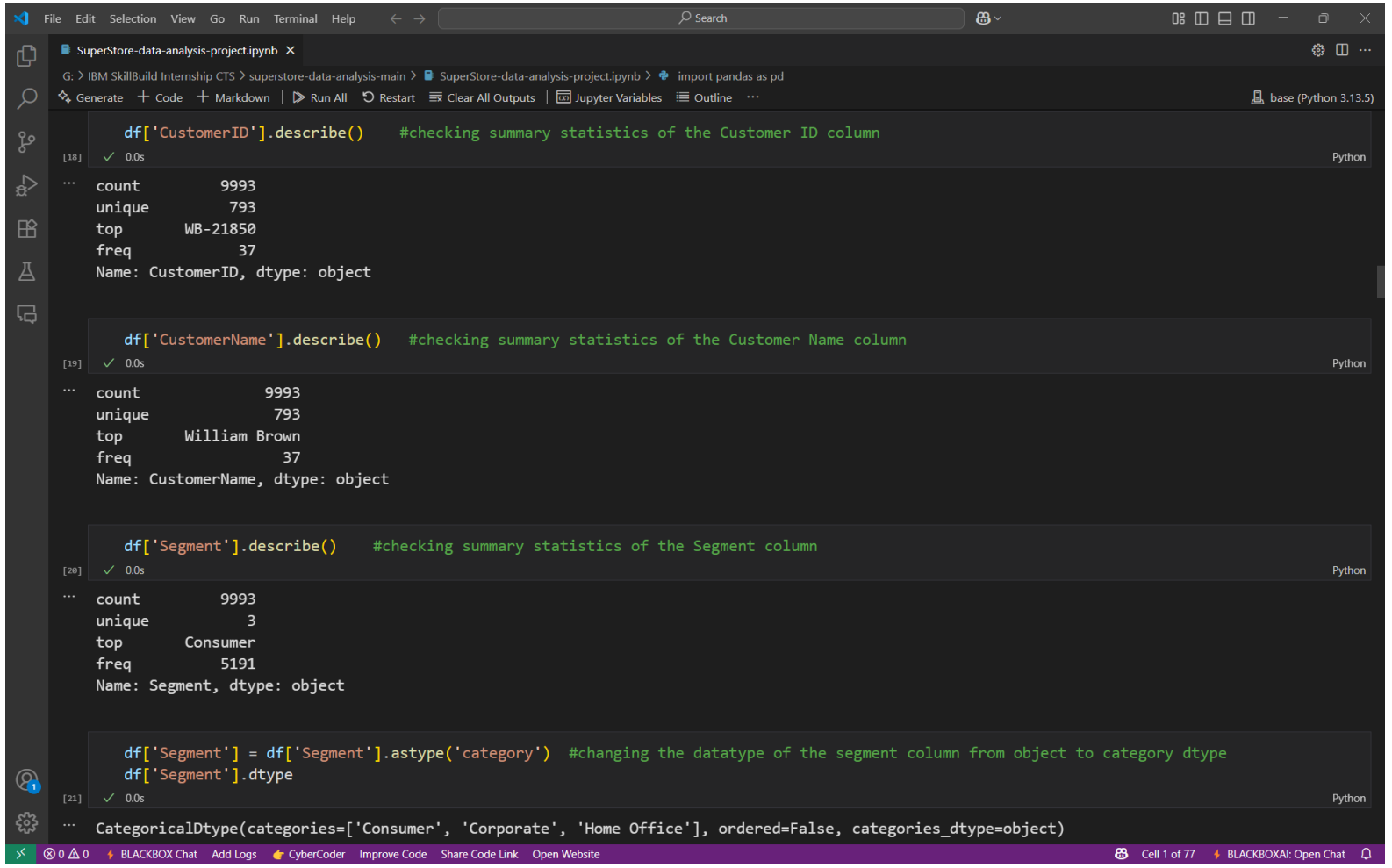


edunet foundation

RESULTS AND SCREENSHOTS

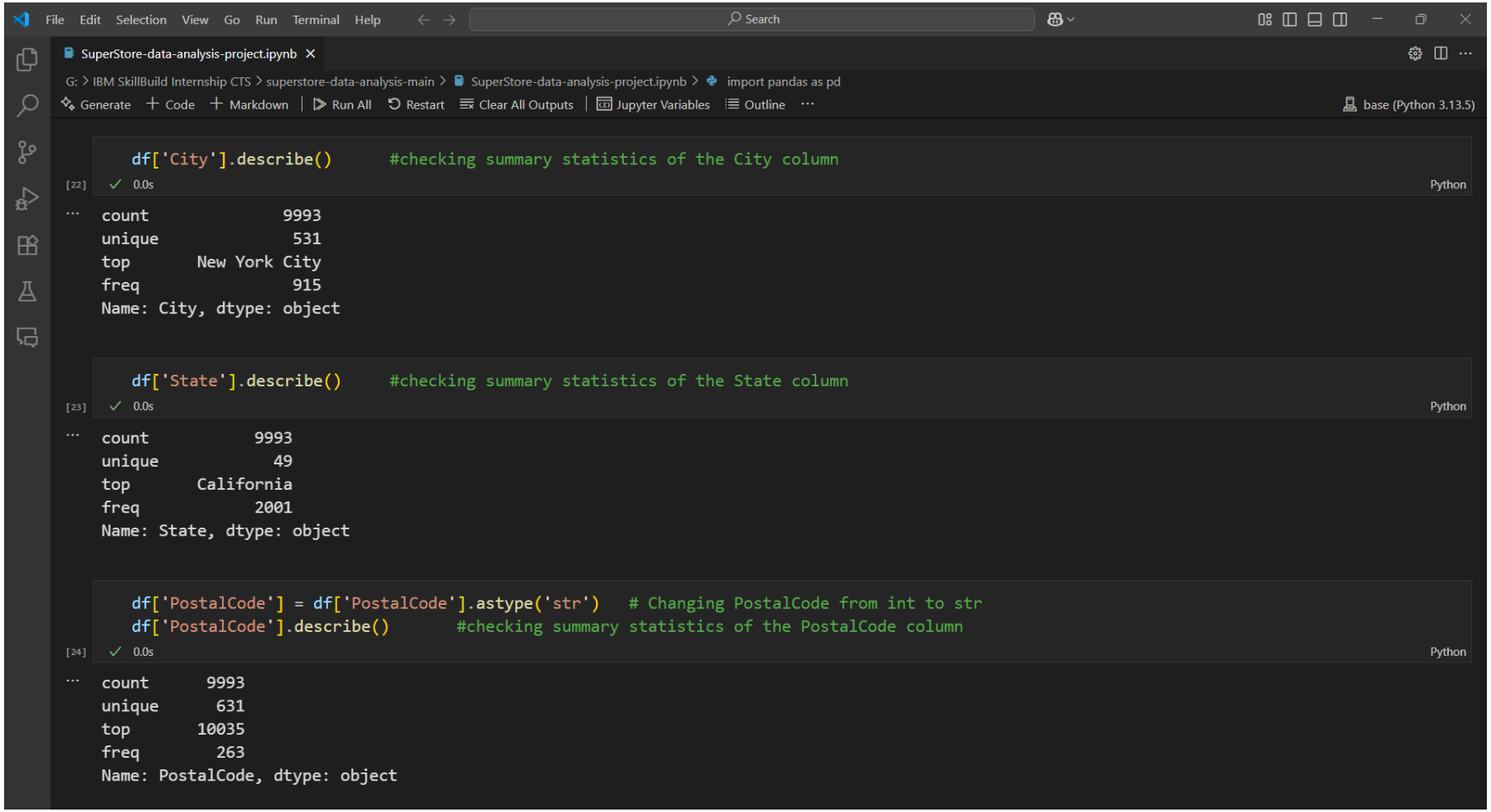


RESULTS AND SCREENSHOTS



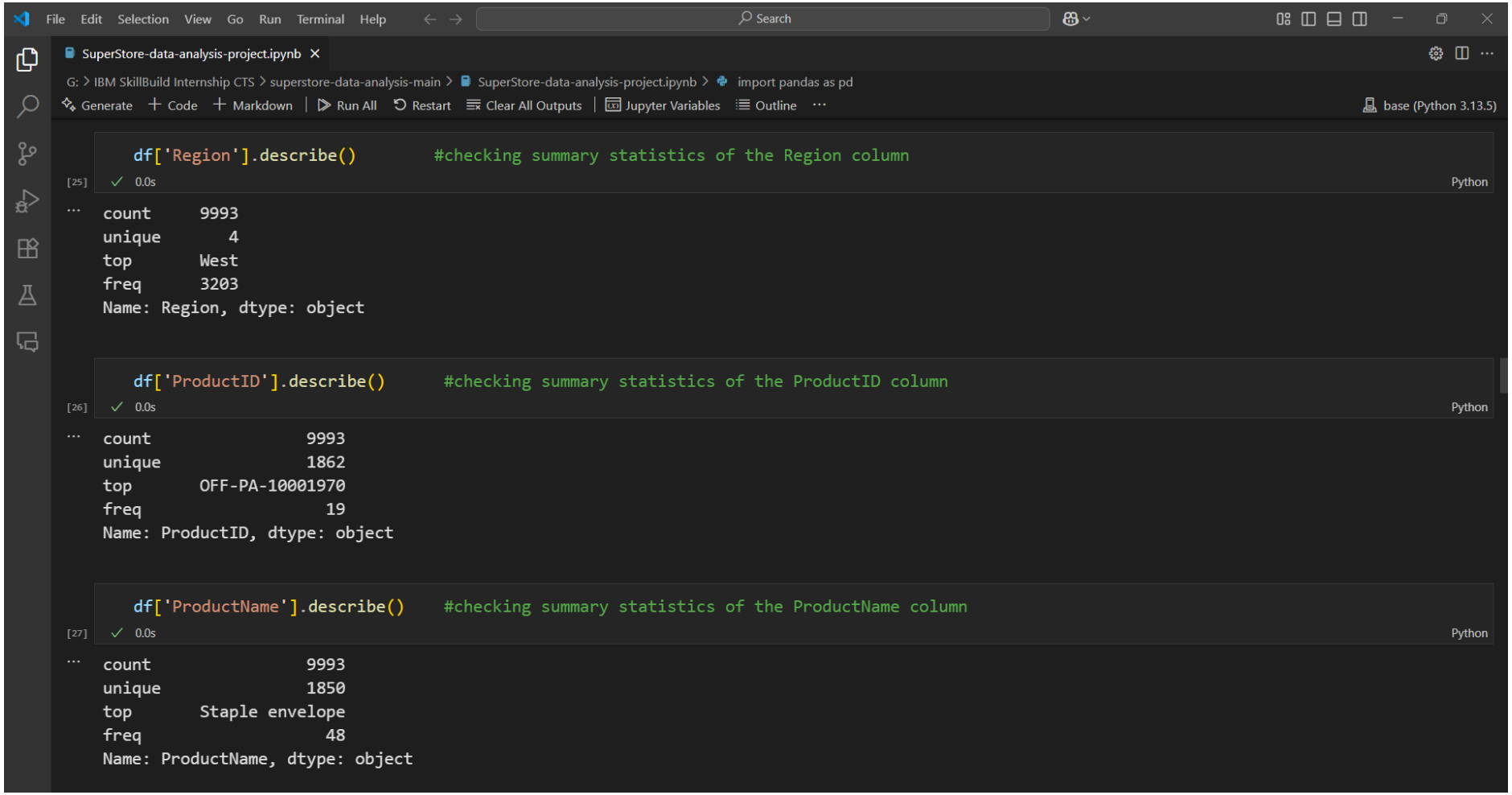
edunet foundation

RESULTS AND SCREENSHOTS



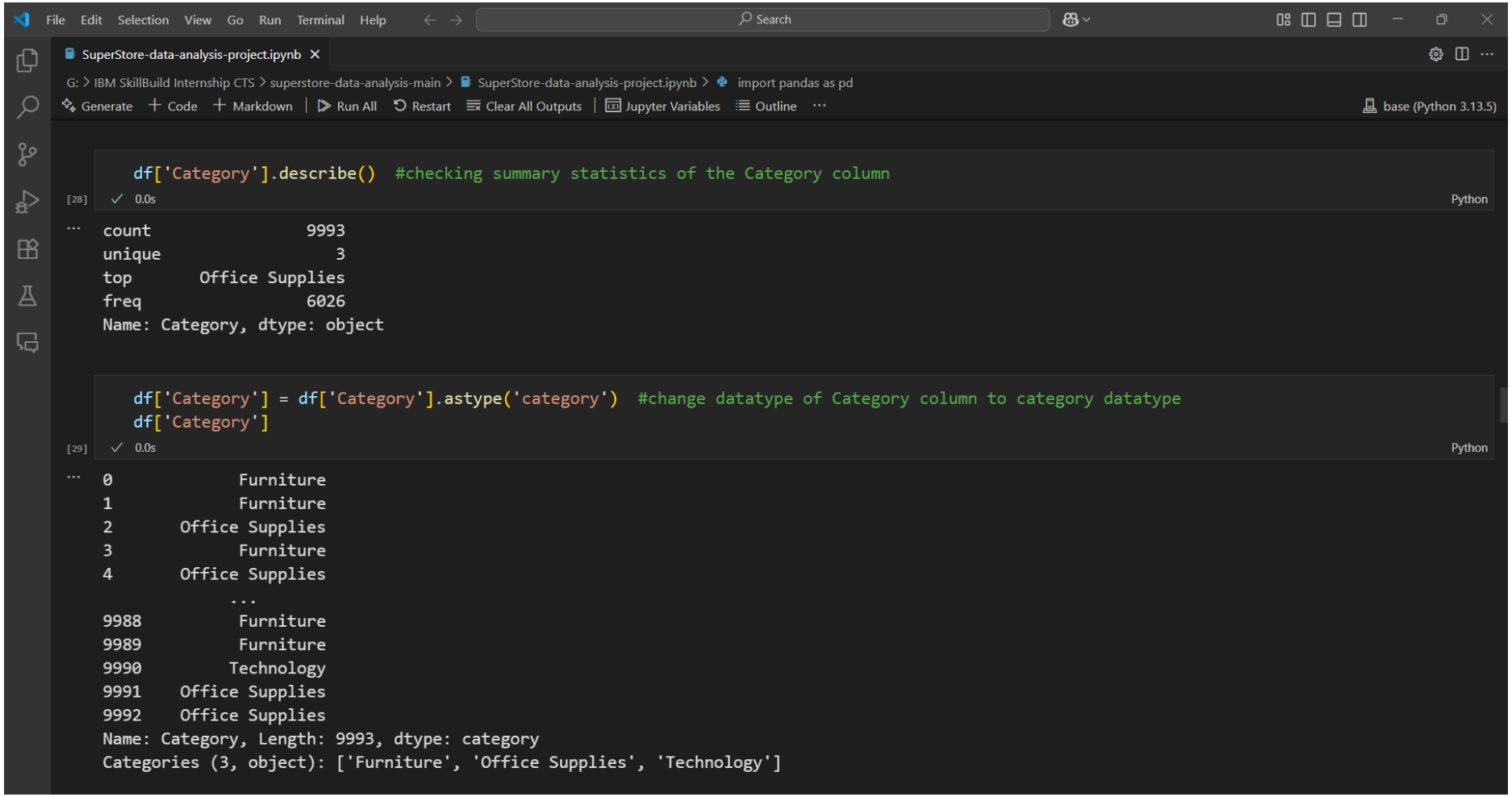
edunet foundation

RESULTS AND SCREENSHOTS



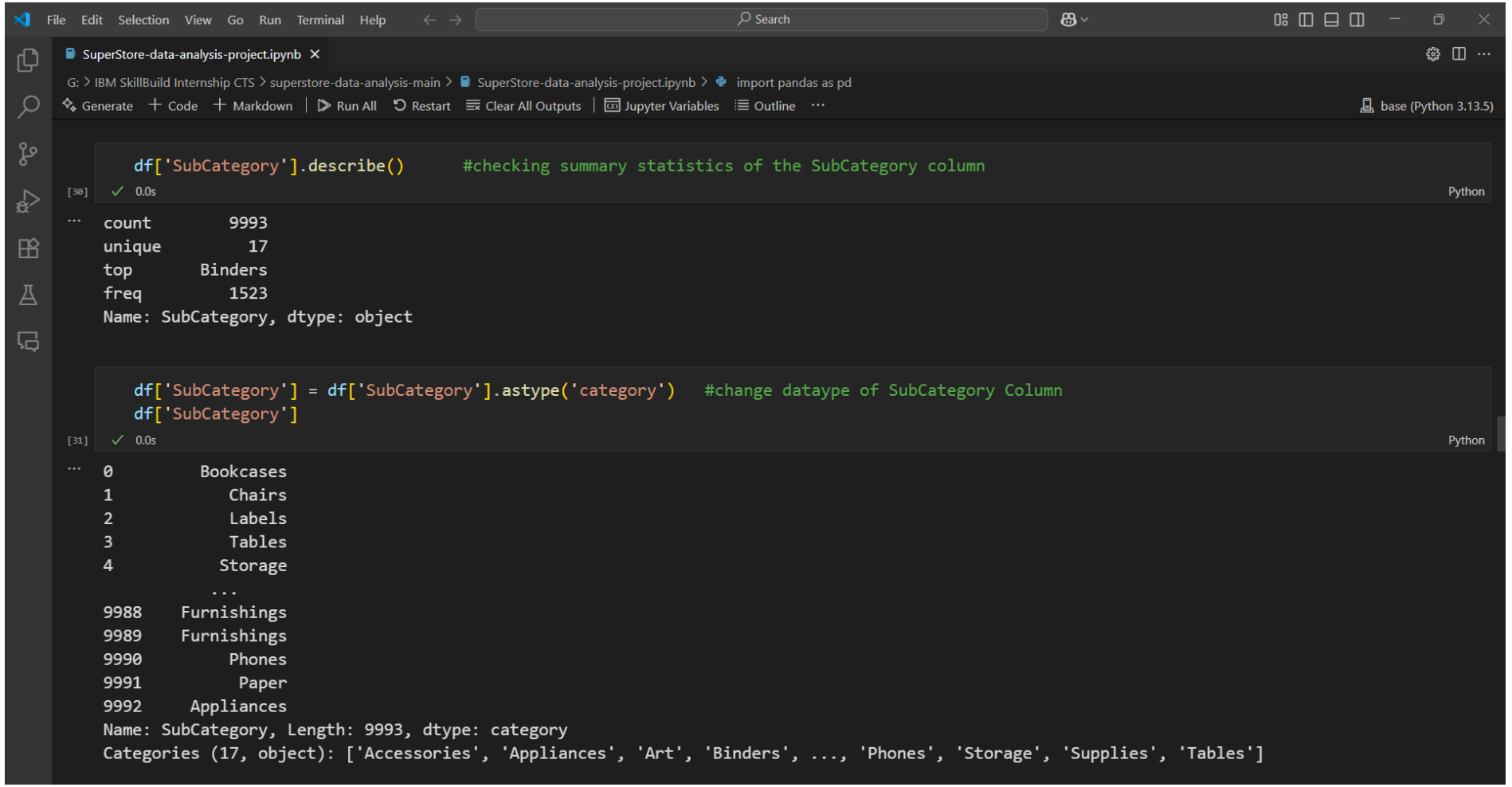
edunet foundation

RESULTS AND SCREENSHOTS

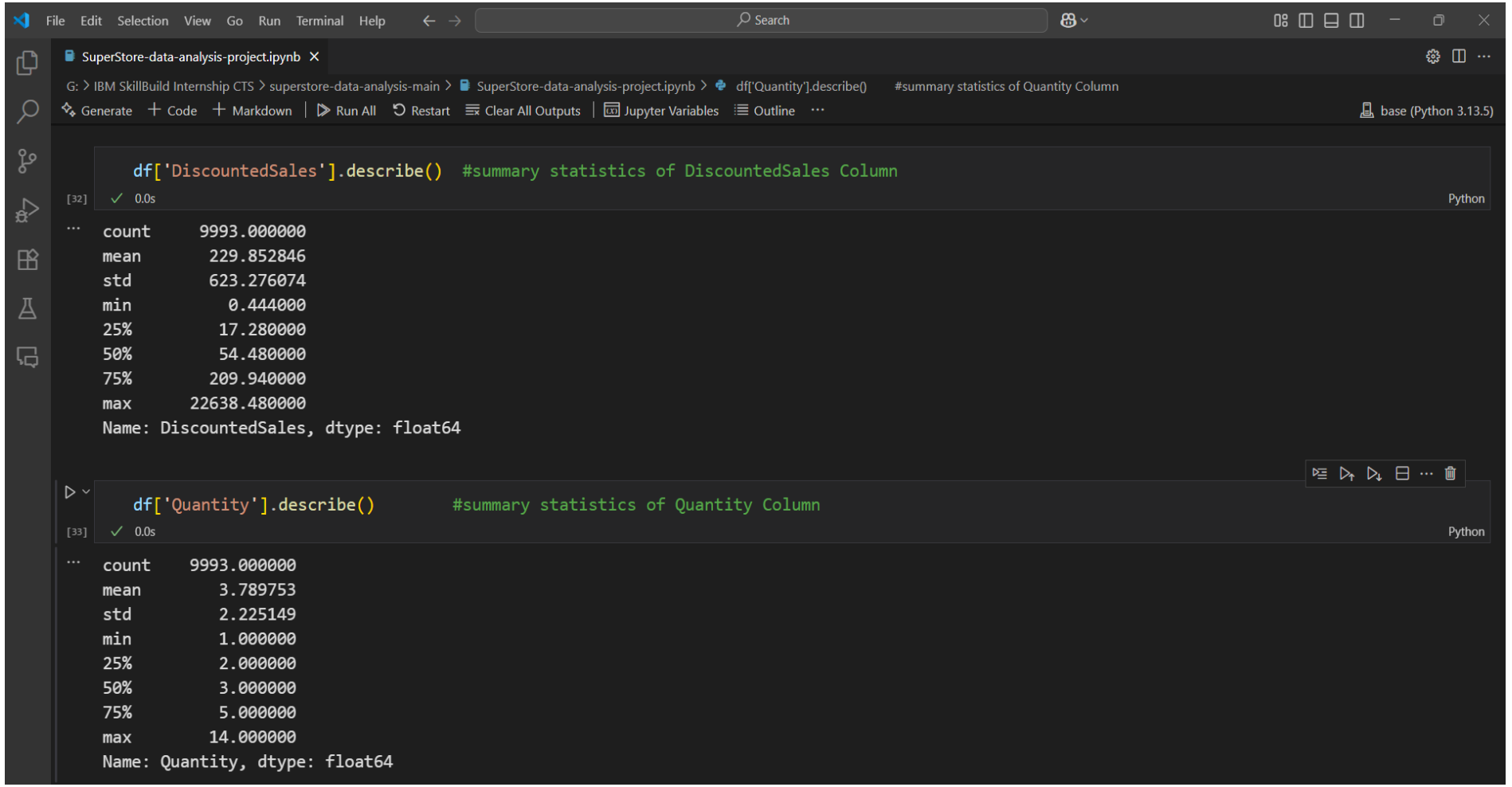


edunet foundation

RESULTS AND SCREENSHOTS

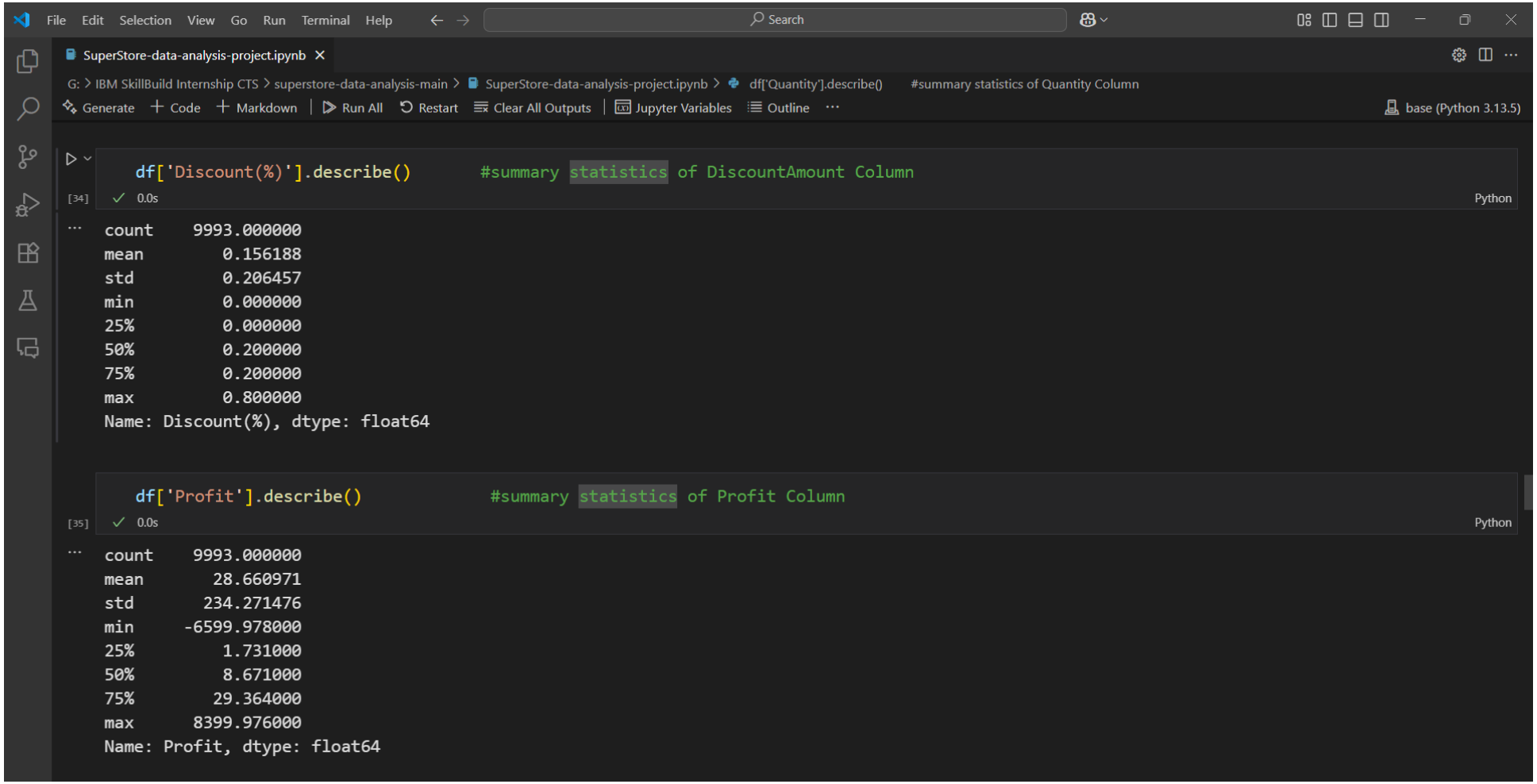


RESULTS AND SCREENSHOTS



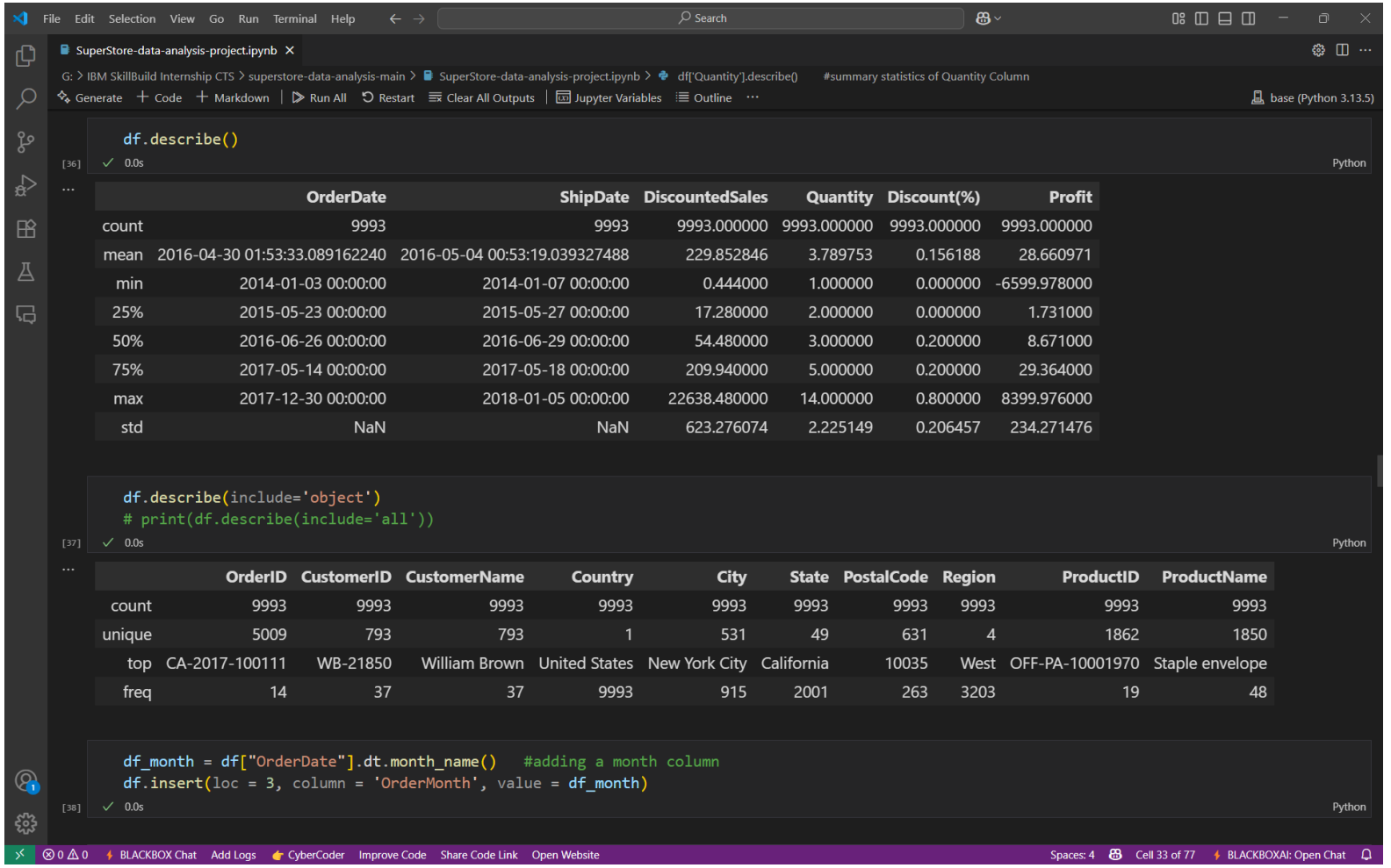
edunet foundation

RESULTS AND SCREENSHOTS



edunet foundation

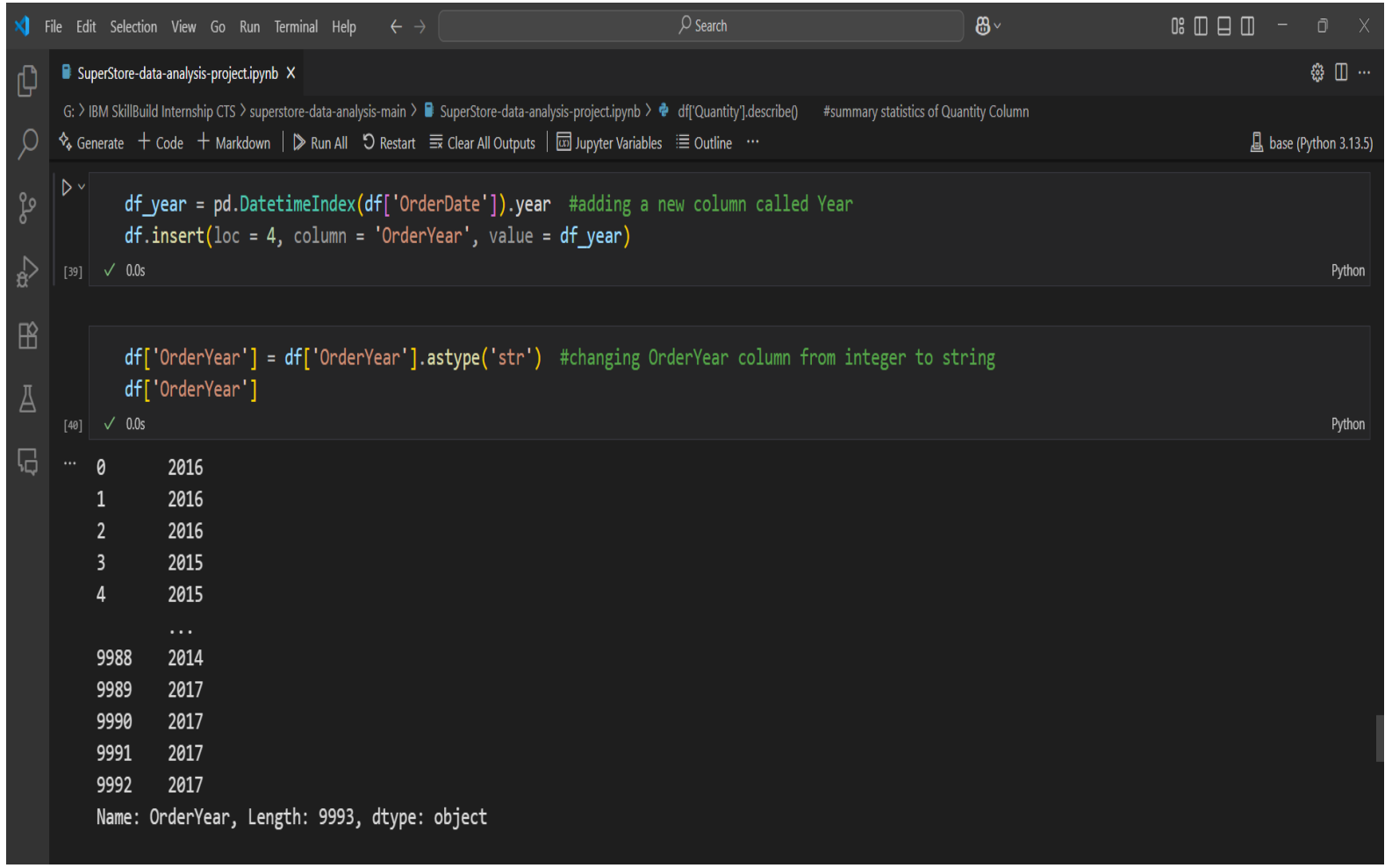
RESULTS AND SCREENSHOTS



	OrderDate	ShipDate	DiscountedSales	Quantity	Discount(%)	Profit
count	9993	9993	9993.000000	9993.000000	9993.000000	9993.000000
mean	2016-04-30 01:53:33.089162240	2016-05-04 00:53:19.039327488	229.852846	3.789753	0.156188	28.660971
min	2014-01-03 00:00:00	2014-01-07 00:00:00	0.444000	1.000000	0.000000	-6599.978000
25%	2015-05-23 00:00:00	2015-05-27 00:00:00	17.280000	2.000000	0.000000	1.731000
50%	2016-06-26 00:00:00	2016-06-29 00:00:00	54.480000	3.000000	0.200000	8.671000
75%	2017-05-14 00:00:00	2017-05-18 00:00:00	209.940000	5.000000	0.200000	29.364000
max	2017-12-30 00:00:00	2018-01-05 00:00:00	22638.480000	14.000000	0.800000	8399.976000
std	NaN	NaN	623.276074	2.225149	0.206457	234.271476

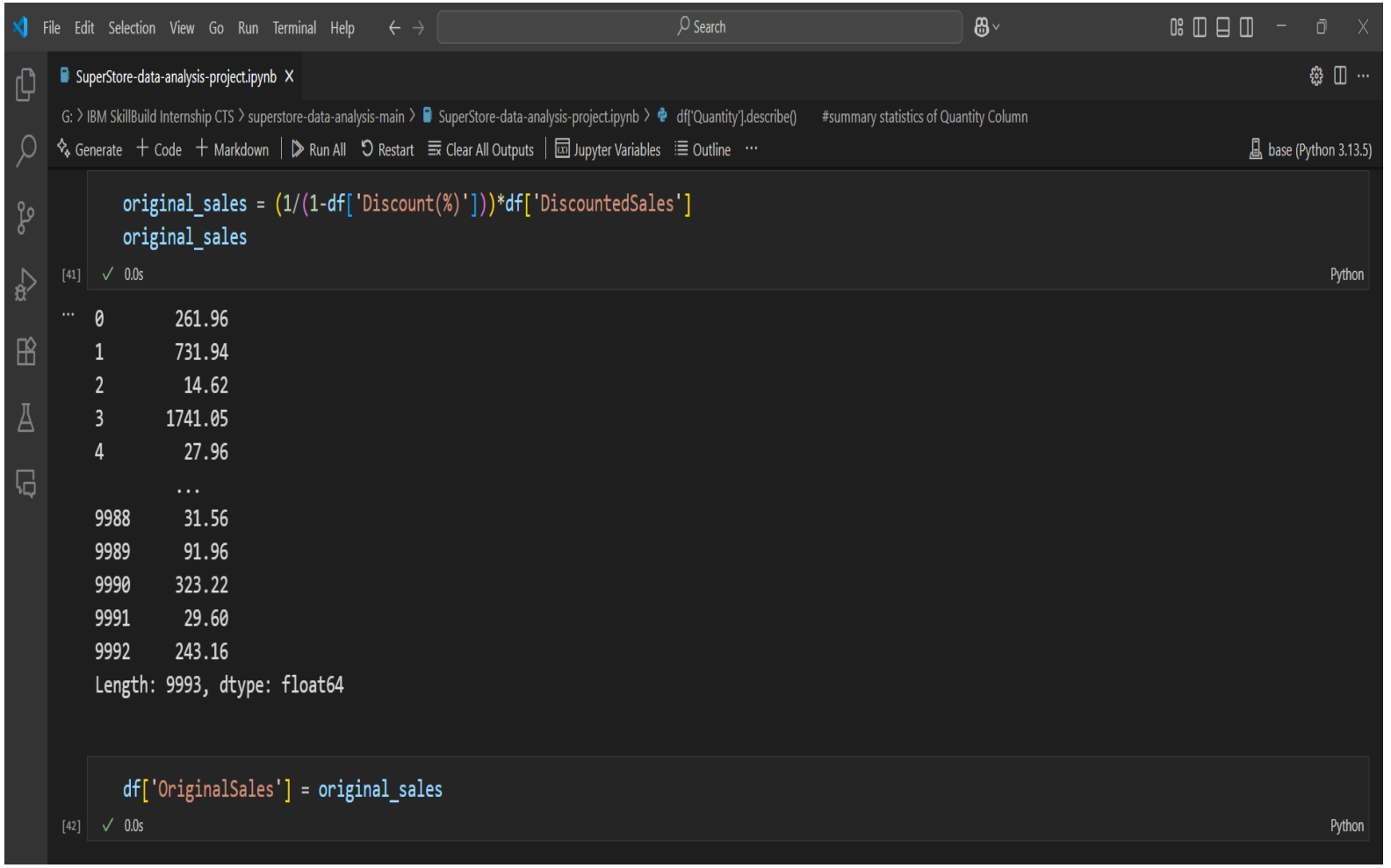
	OrderID	CustomerID	CustomerName	Country	City	State	PostalCode	Region	ProductID	ProductName
count	9993	9993	9993	9993	9993	9993	9993	9993	9993	9993
unique	5009	793	793	1	531	49	631	4	1862	1850
top	CA-2017-100111	WB-21850	William Brown	United States	New York City	California	10035	West	OFF-PA-10001970	Staple envelope
freq	14	37	37	9993	915	2001	263	3203	19	48

RESULTS AND SCREENSHOTS



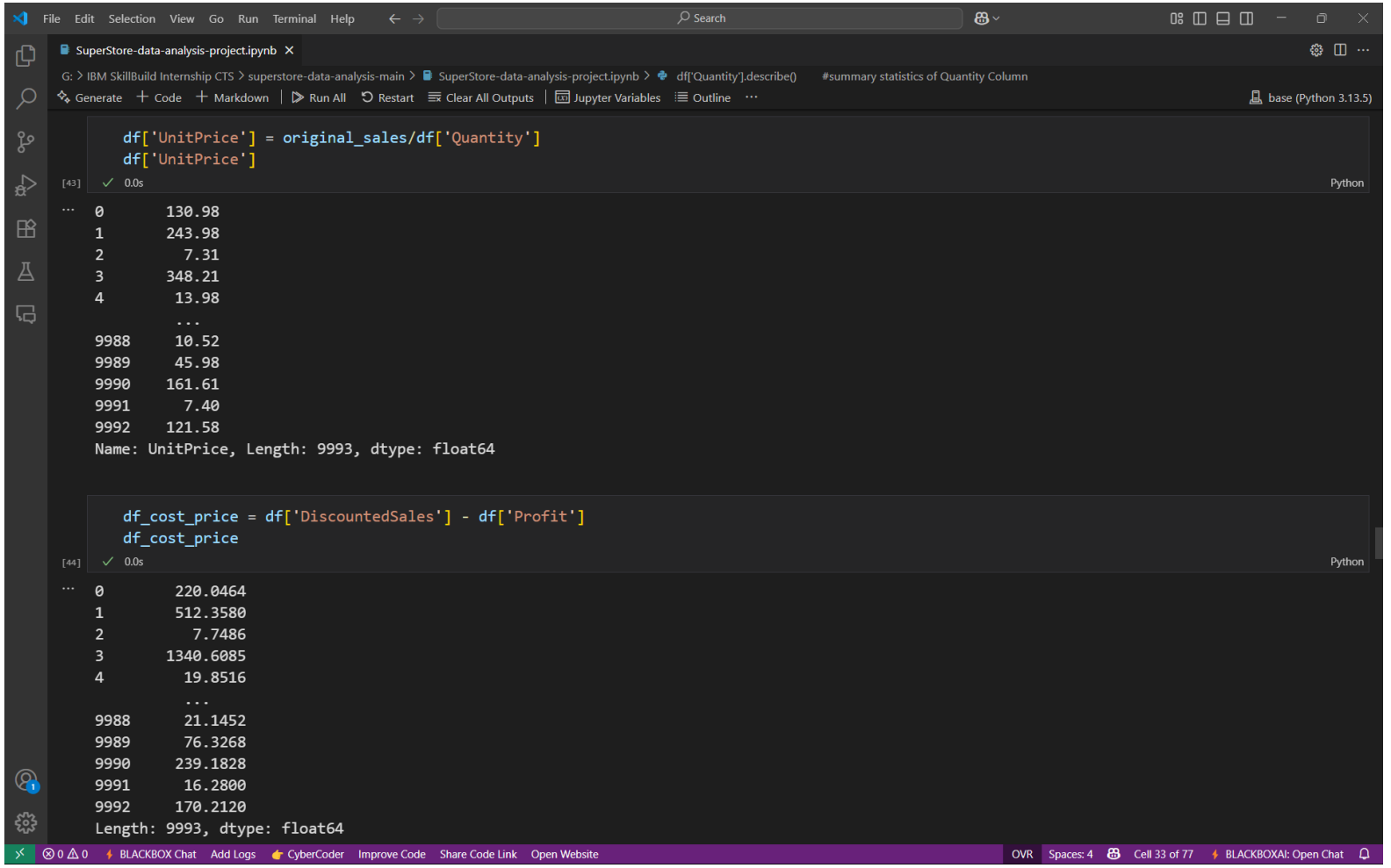
edunet foundation

RESULTS AND SCREENSHOTS



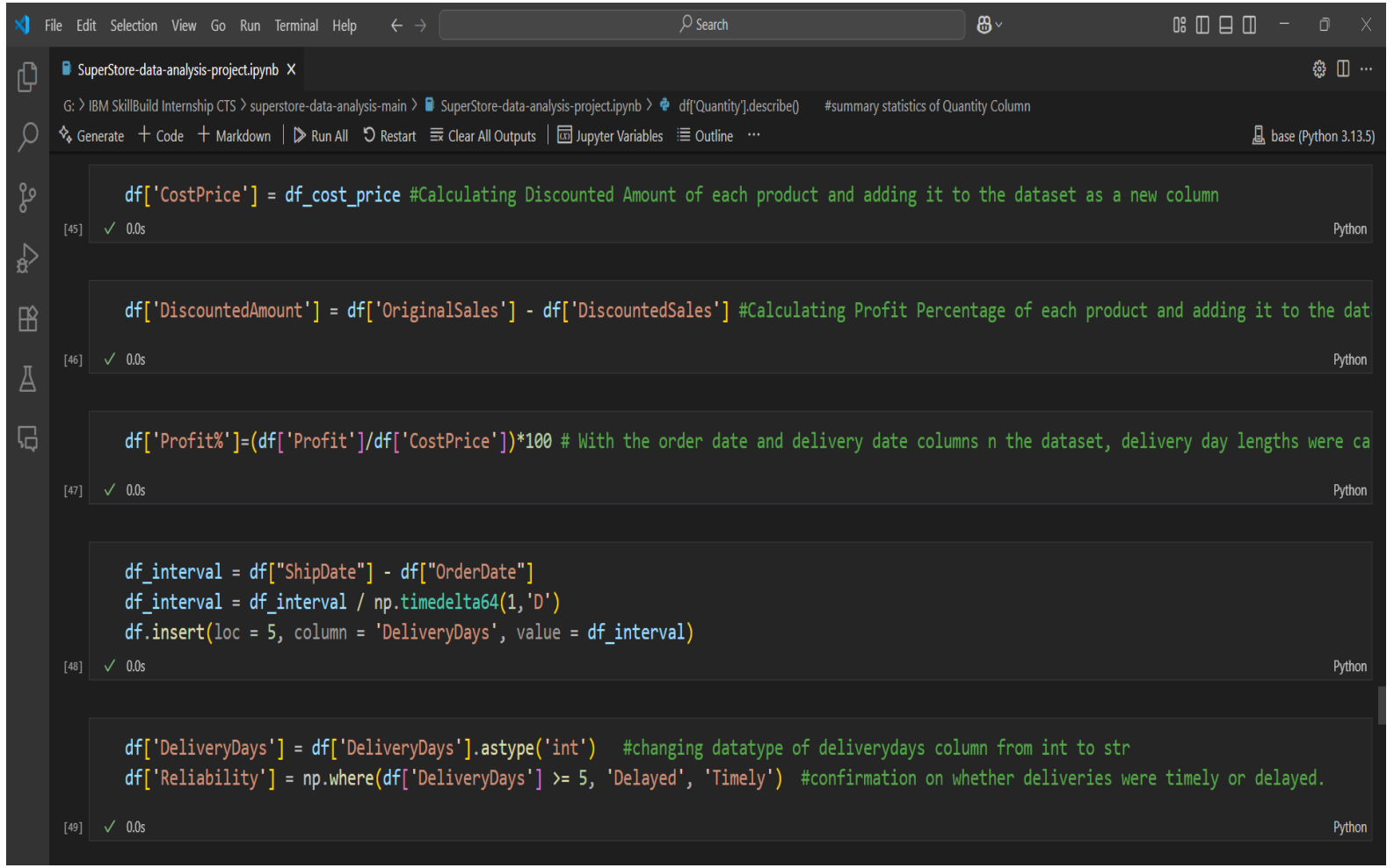
edunet foundation

RESULTS AND SCREENSHOTS



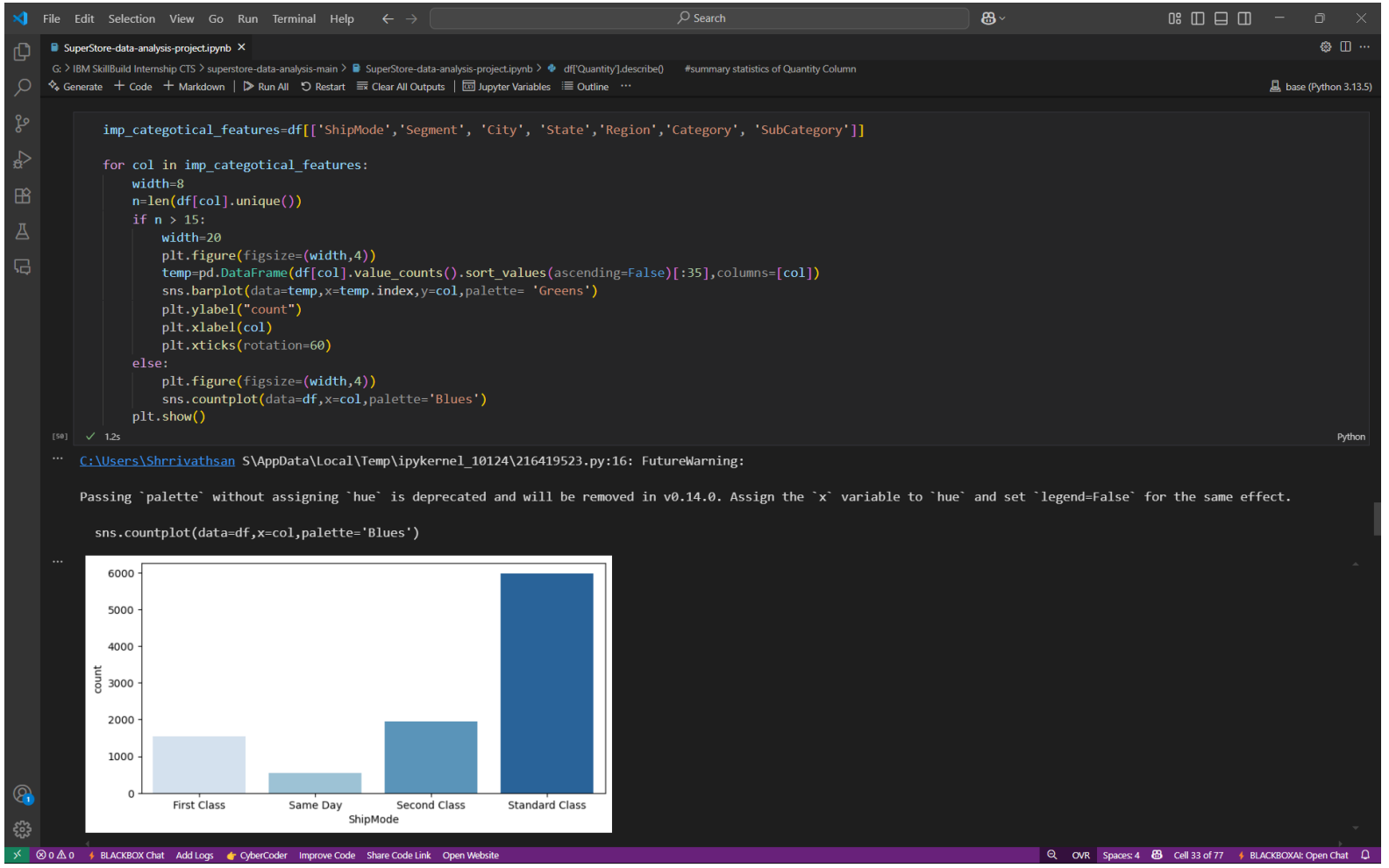
edunet foundation

RESULTS AND SCREENSHOTS



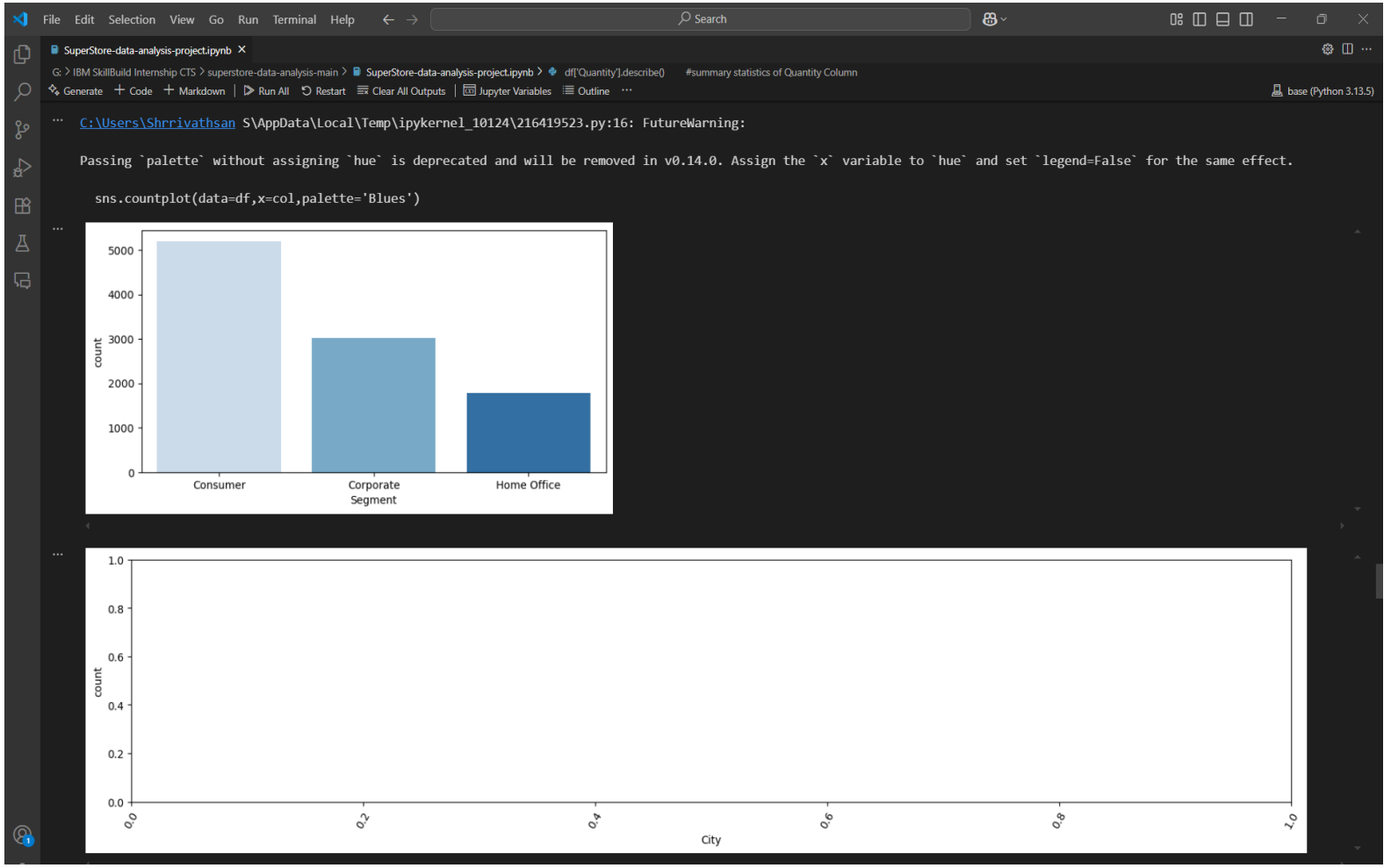
edunet foundation

RESULTS AND SCREENSHOTS

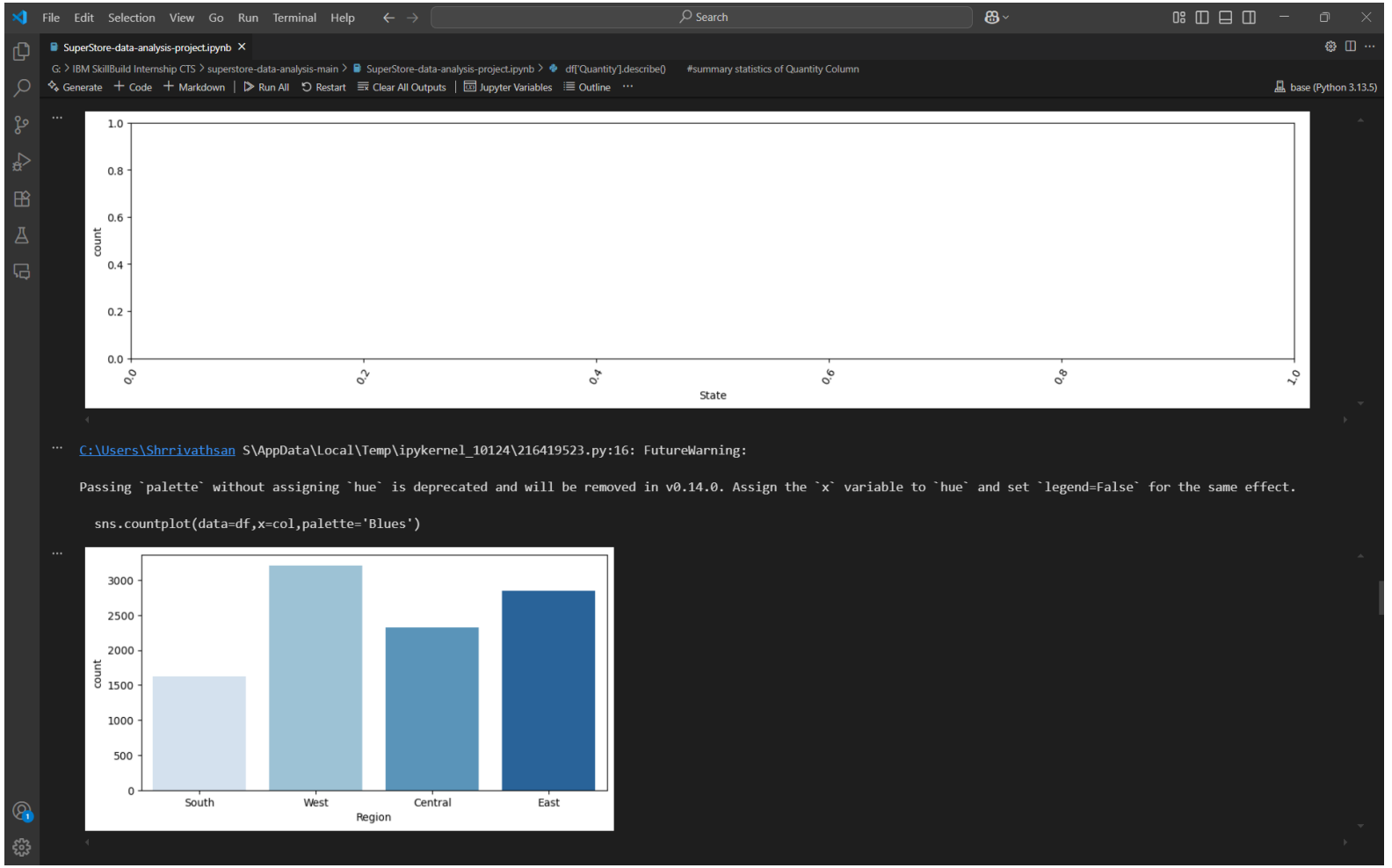


edunet foundation

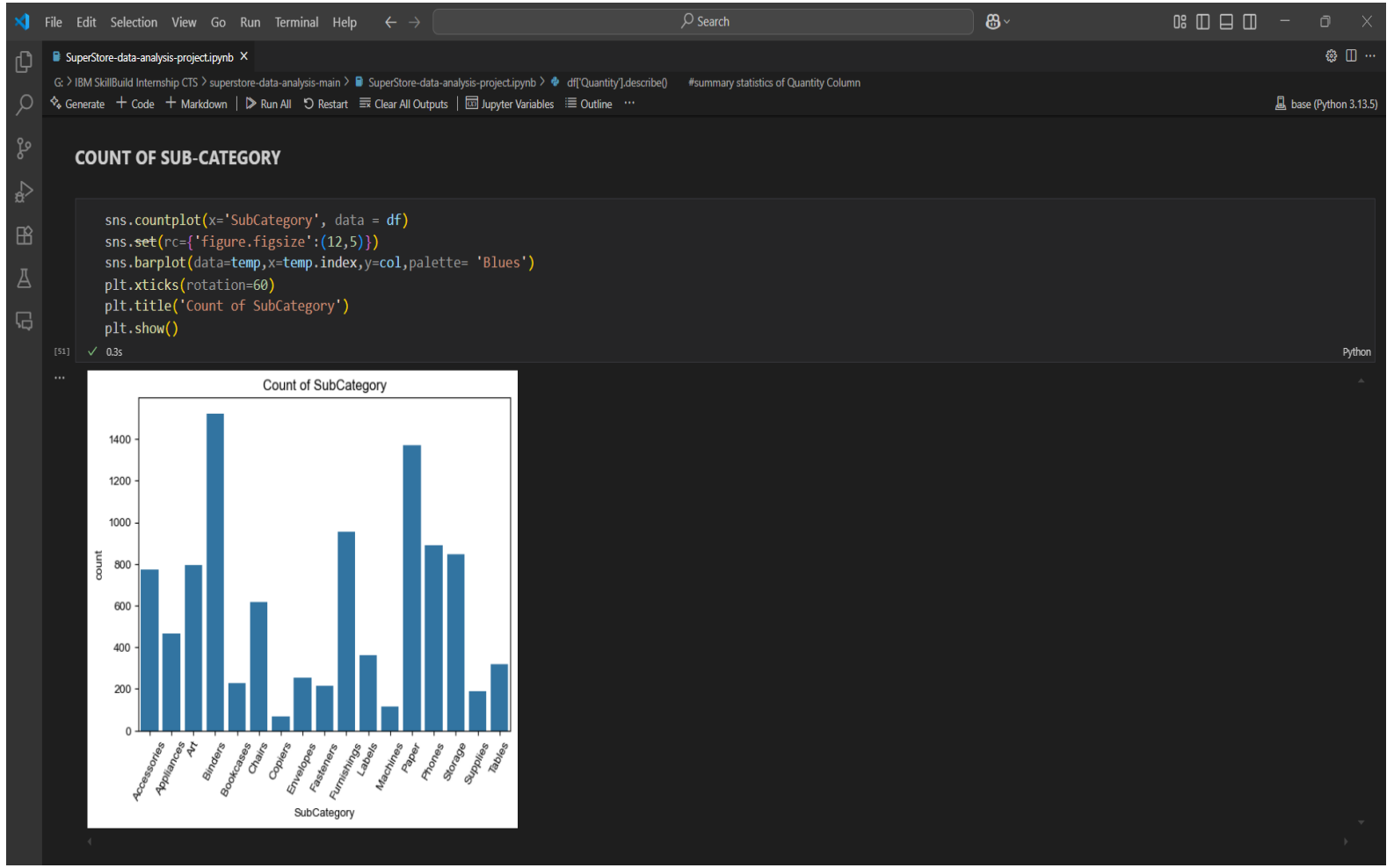
RESULTS AND SCREENSHOTS



RESULTS AND SCREENSHOTS

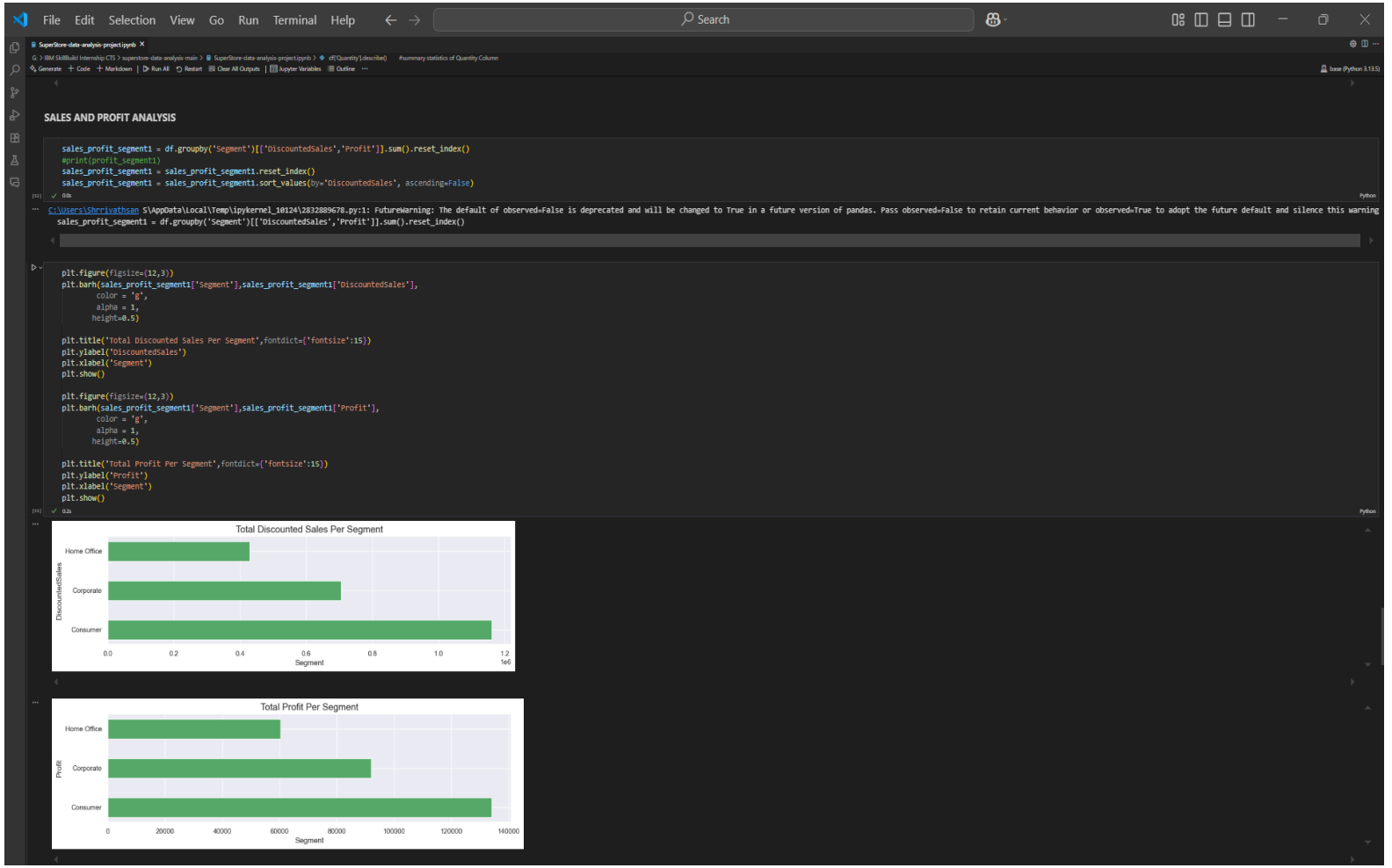


RESULTS AND SCREENSHOTS



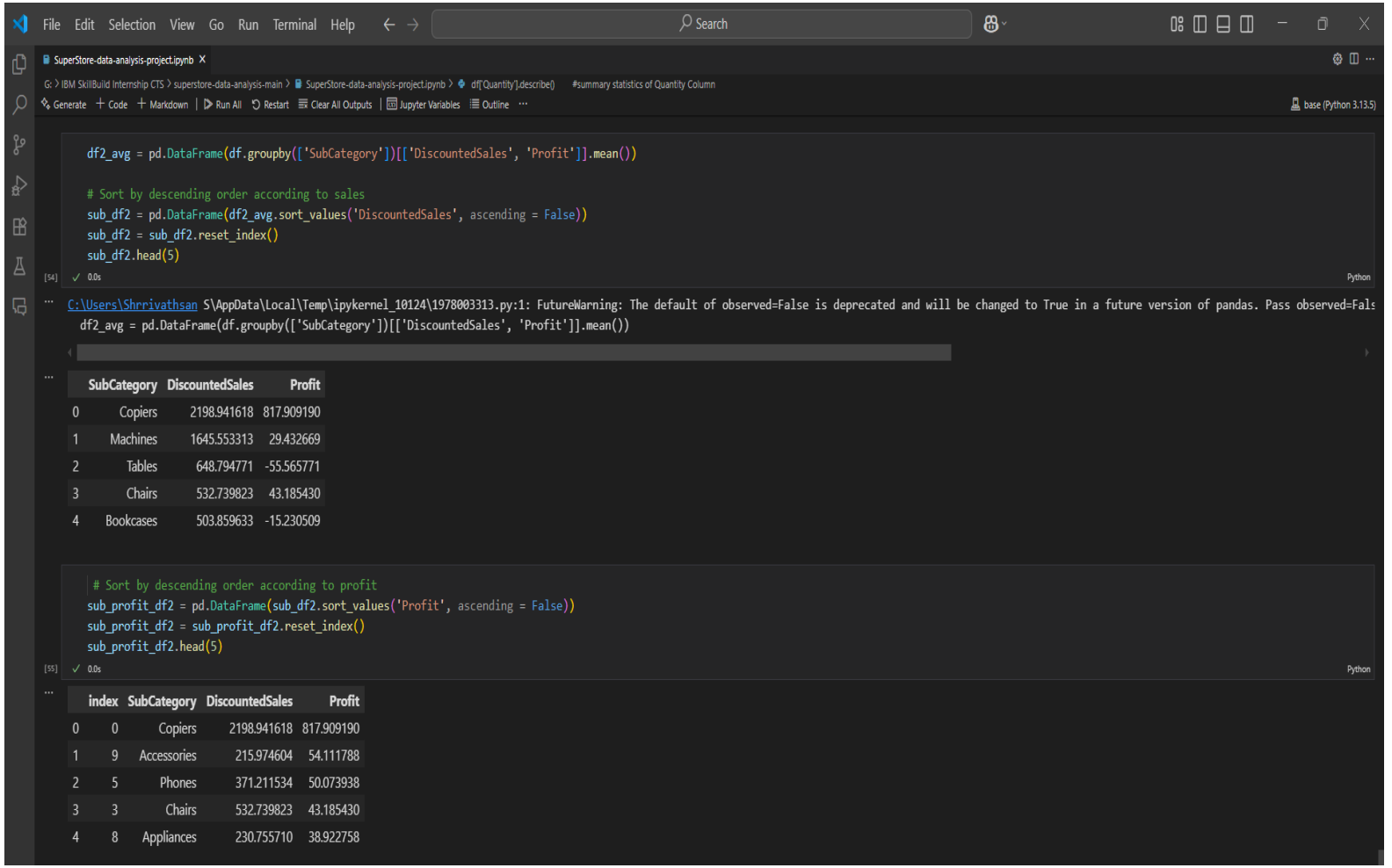
edunet foundation

RESULTS AND SCREENSHOTS



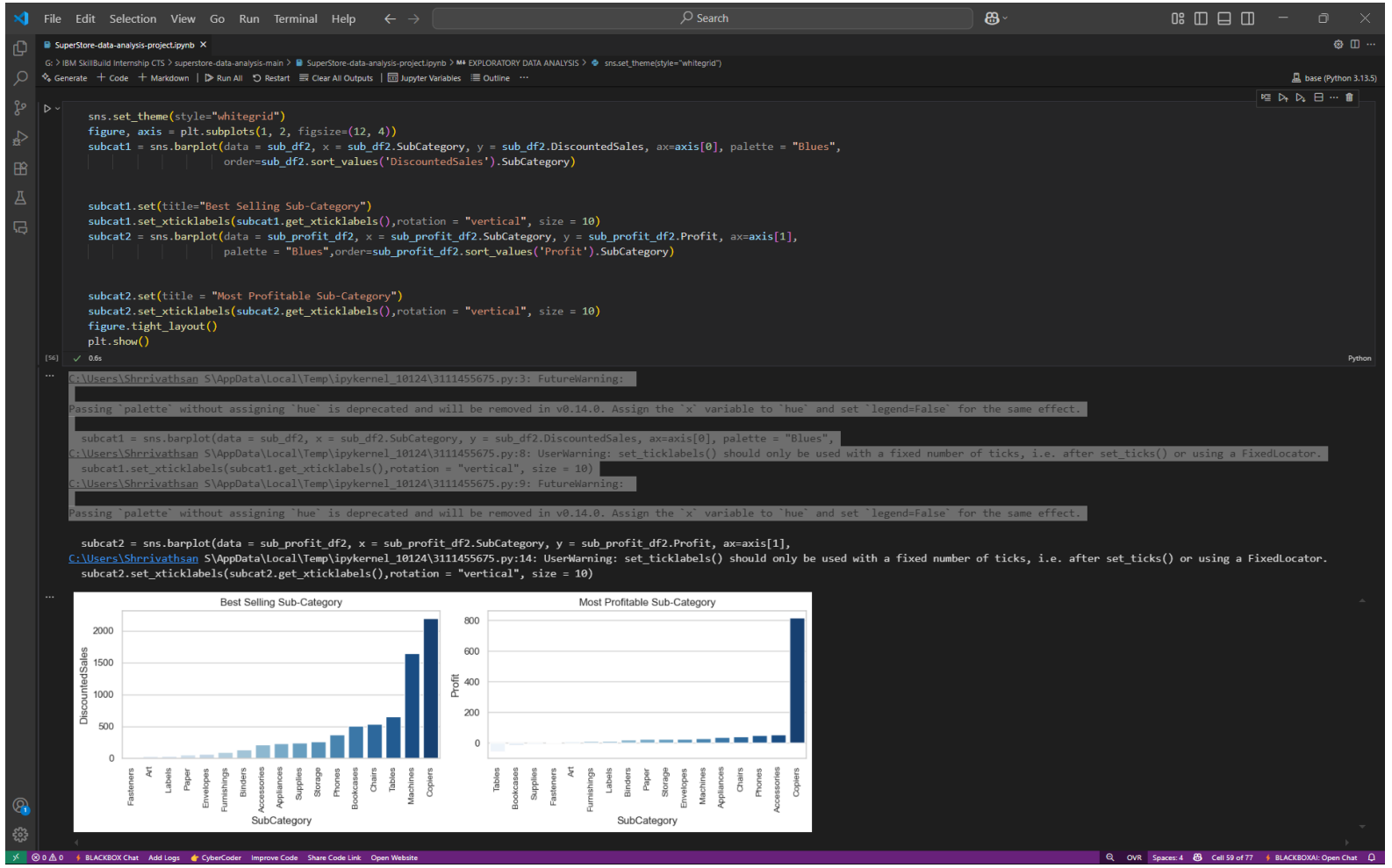
edunet foundation

RESULTS AND SCREENSHOTS



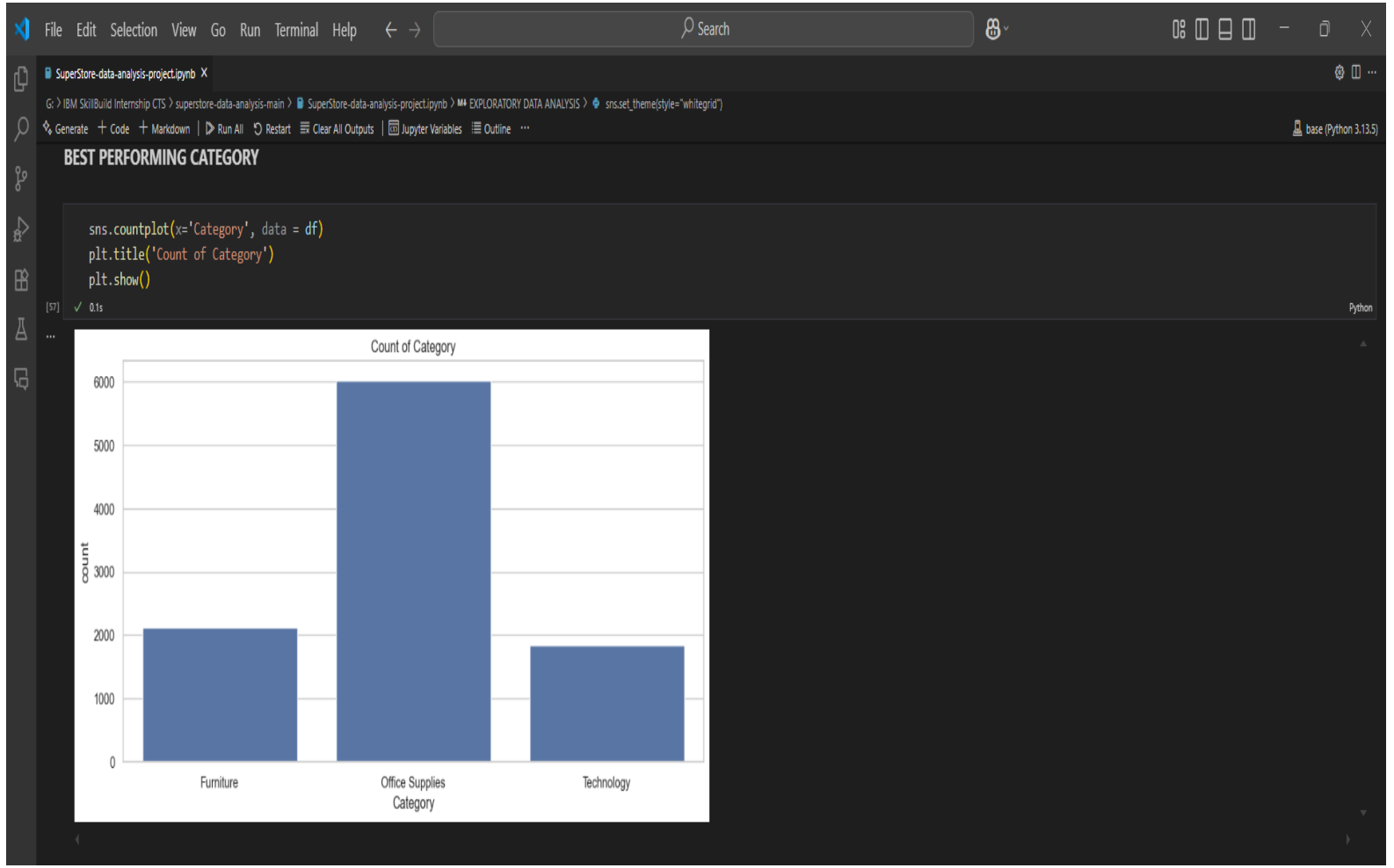
edunet foundation

RESULTS AND SCREENSHOTS

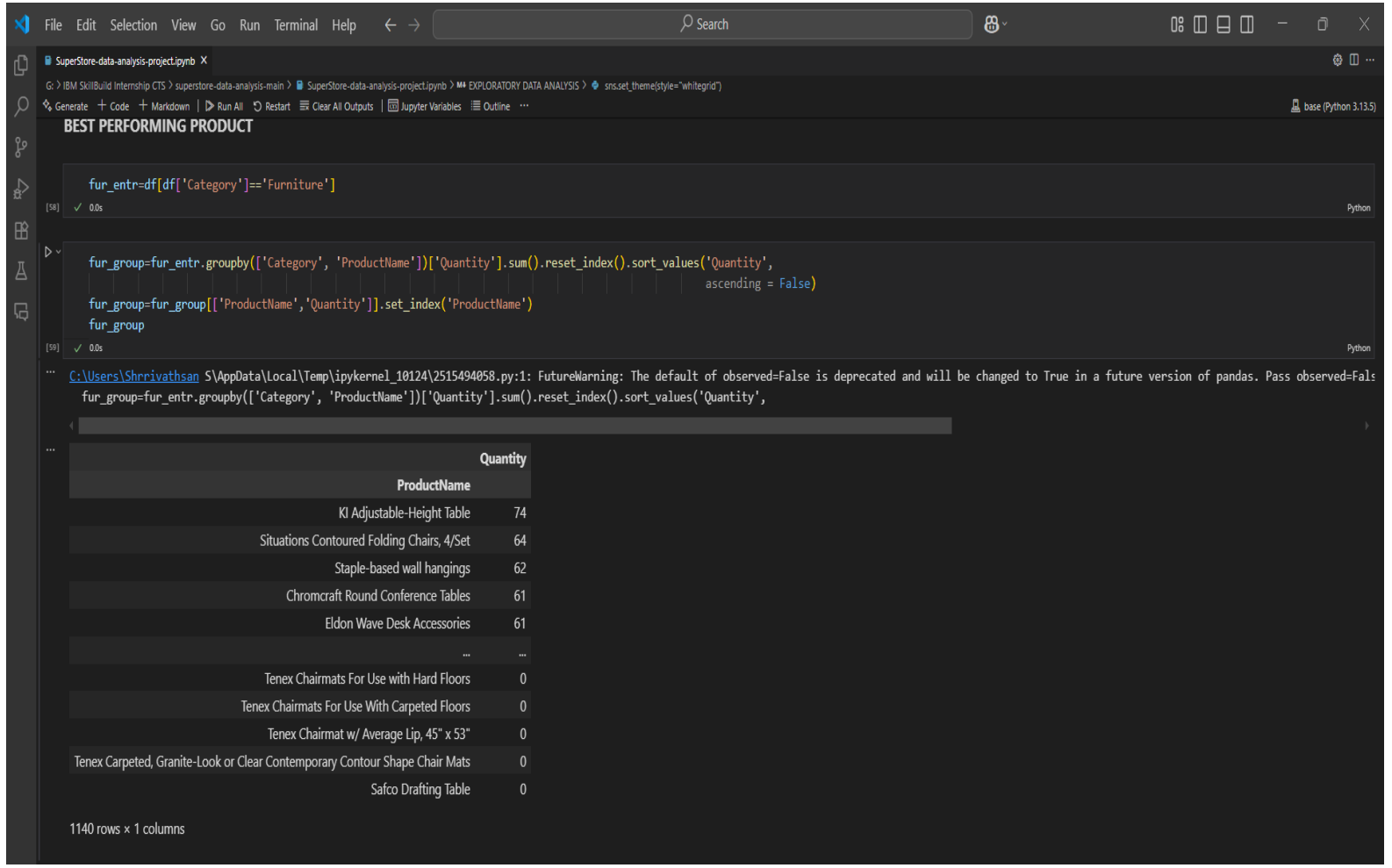


edunet foundation

RESULTS AND SCREENSHOTS



RESULTS AND SCREENSHOTS

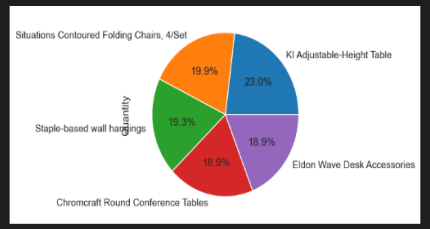


edunet foundation

RESULTS AND SCREENSHOTS

```
plt.figure(figsize=(4,8))
fur_group.head(5)['Quantity'].plot.pie(autopct="%1.1f%%",colors=['#1f77b4', '#ff7f0e', '#2ca02c', '#d62728', '#9467bd'])
```

```
<Axes: ylabel='Quantity'>
```



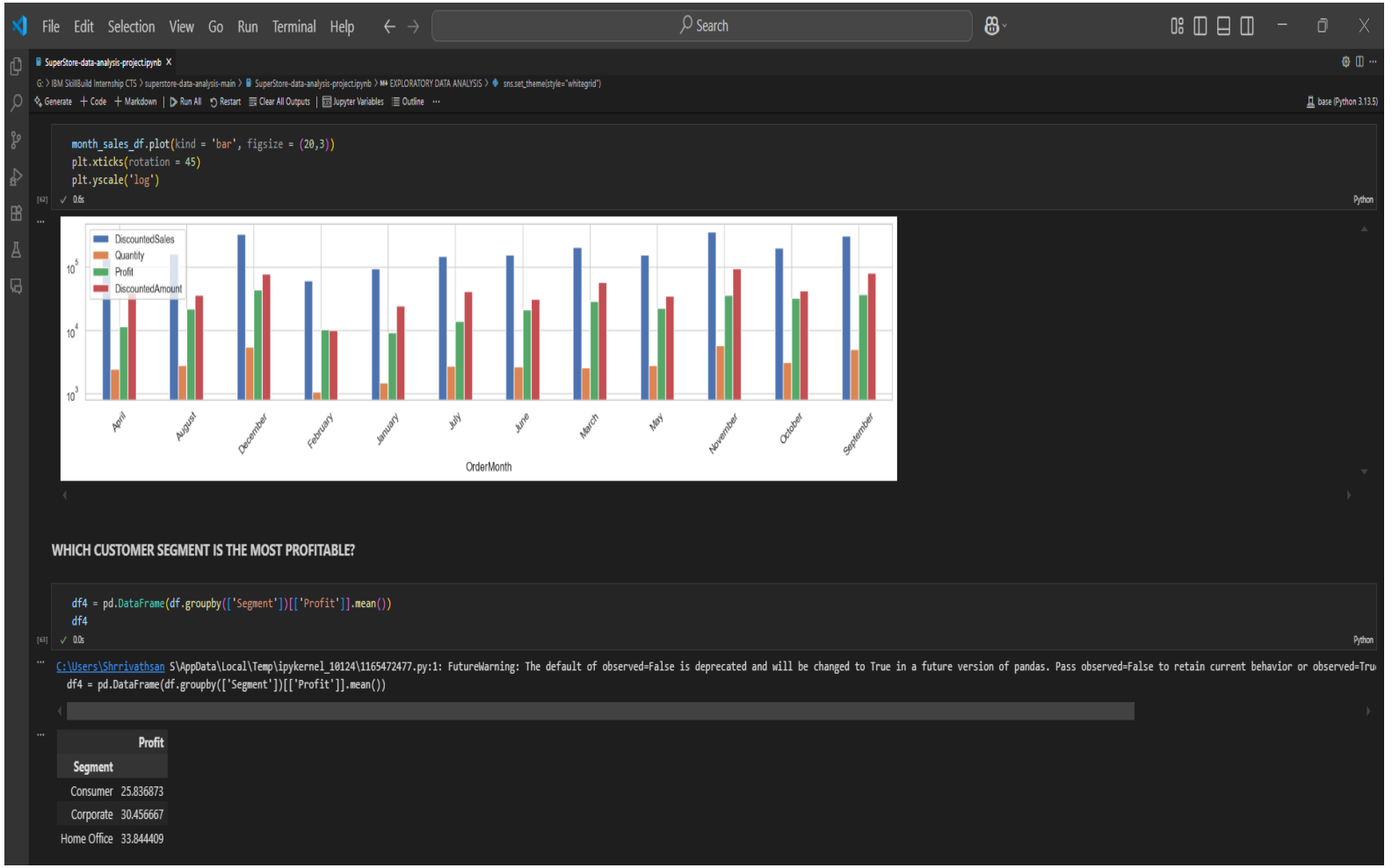
- It is observed in the above line of code that people residing in Western part of US tend to order more from superstore.

MONTHLY AND YEARLY ANALYSIS

```
# Which month does the company records more sales?
month_sales_df = pd.DataFrame(df.groupby(['OrderMonth'])[['DiscountedSales', 'Quantity', 'Profit', 'DiscountedAmount']].sum())
month_sales_df = month_sales_df.reset_index()
month_sales_df = month_sales_df.set_index('OrderMonth')
month_sales_df
```

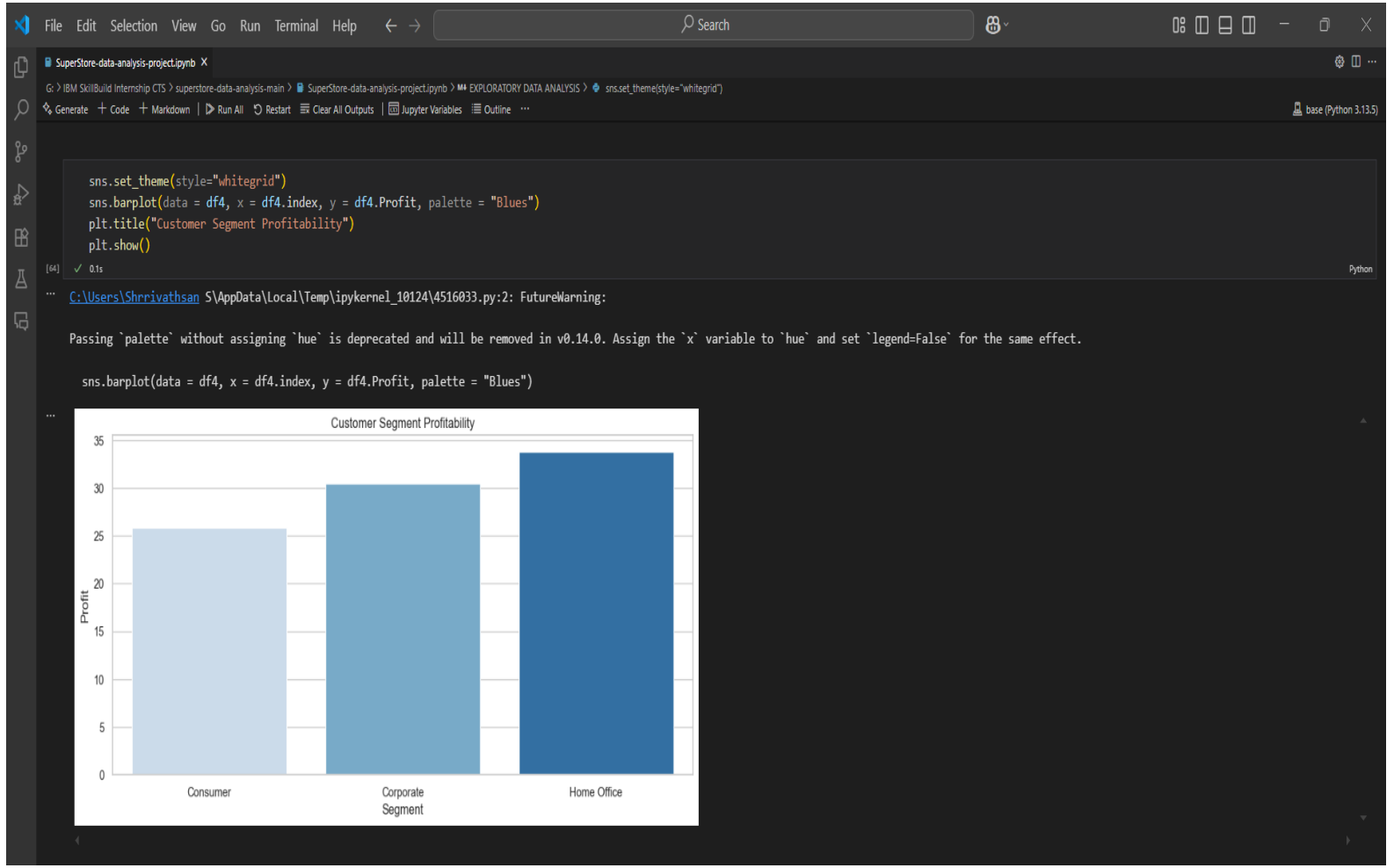
OrderMonth	DiscountedSales	Quantity	Profit	DiscountedAmount
April	137480.7566	2445	11599.4951	38533.1834
August	159044.0630	2784	21776.9384	35948.4570
December	325293.5035	5419	43369.1919	77608.6265
February	59751.2514	1067	10294.6107	10018.4986
January	94924.8356	1475	9134.4461	24335.0744
July	147238.0970	2705	13832.6648	41251.3630
June	152718.6793	2680	21285.7954	31385.1107
March	205005.4888	2564	28594.6872	56516.7712
May	155028.8117	2791	22411.3078	35075.9583
November	352461.0710	5775	35468.4265	93858.0990
October	200322.9847	3104	31784.0413	42630.5553
September	307649.9457	5062	36857.4753	79451.8943

RESULTS AND SCREENSHOTS



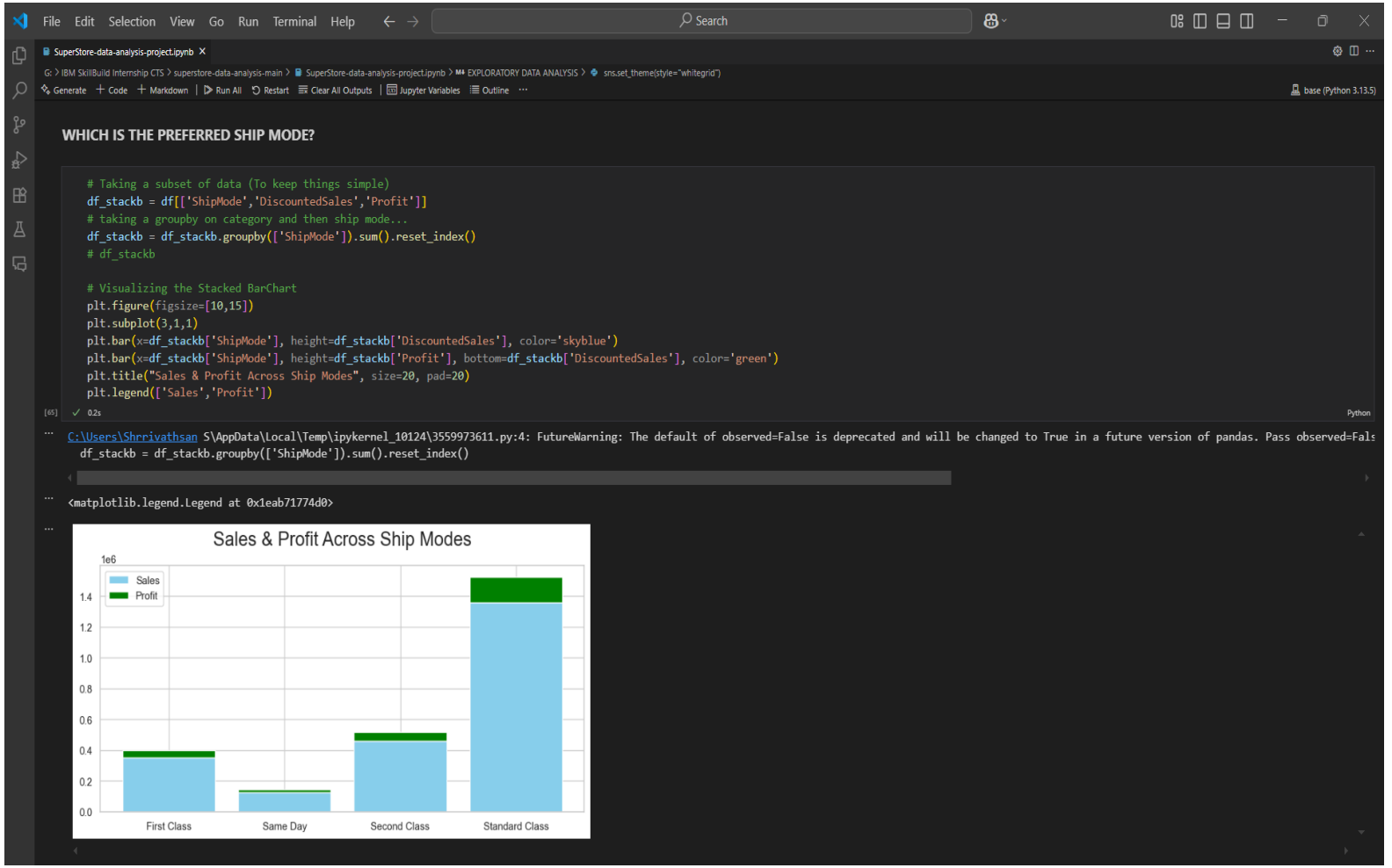
edunet foundation

RESULTS AND SCREENSHOTS



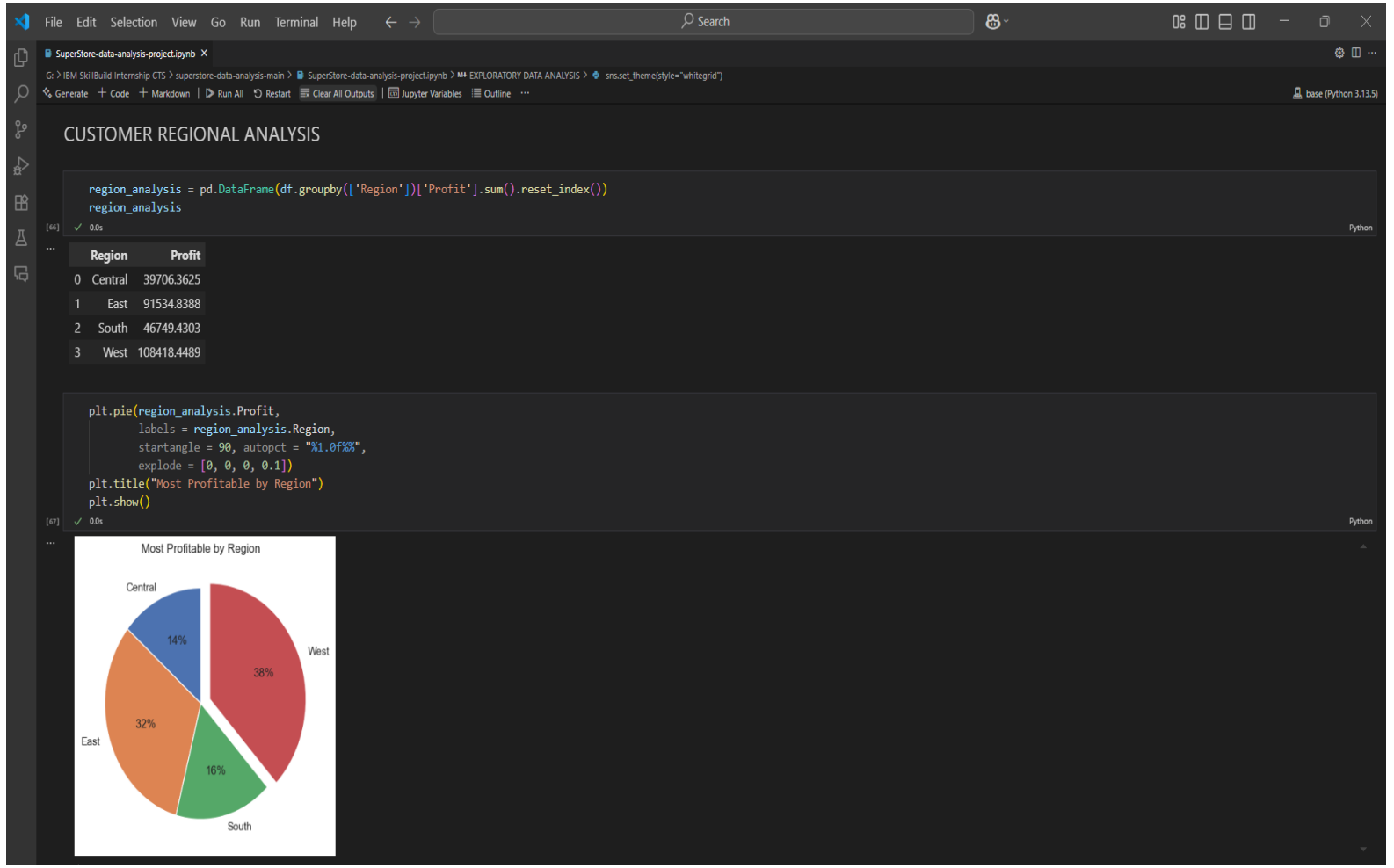
edunet foundation

RESULTS AND SCREENSHOTS



edunet foundation

RESULTS AND SCREENSHOTS



edunet foundation

CONCLUSION

- The SuperStore data analysis using AI and data science techniques provides deep insights into business performance, customer behavior, and operational bottlenecks. Through exploratory data analysis, machine learning models, and forecasting techniques, we have demonstrated how businesses can leverage data to make informed decisions.

Key takeaways include:

- Identification of high-performing product categories and loss-making areas.
- Discovery of customer segments through clustering, enabling targeted marketing.
- Accurate sales forecasting using time-series models to support inventory and resource planning.
- Use of predictive analytics to understand profit drivers and optimize discount strategies.
- The integration of AI methods such as machine learning, clustering, and forecasting has shown significant potential to improve efficiency, profitability, and customer satisfaction in a retail environment. Moreover, visual analytics and dashboards make these insights accessible and actionable for business stakeholders.
- With the growing accessibility of advanced AI tools, the future holds even more promise. Businesses can further evolve by incorporating real-time analytics, personalized recommendation engines, and geospatial intelligence into their decision-making processes.
- In conclusion, data science and AI not only empower SuperStore to understand its present but also to anticipate its future — turning raw data into strategic advantage.

FUTURE SCOPE

As businesses increasingly adopt data-driven strategies, the scope of AI and data science in retail analytics is expanding rapidly. Below are some key future directions and enhancements that can be applied to the SuperStore analysis:

1. Advanced AI for Personalized Recommendations

Objective: Use customer purchase history, behavior, and segmentation data to recommend products.

Methods: Collaborative filtering, content-based filtering, or hybrid recommendation systems (e.g., matrix factorization, deep learning-based recommenders).Impact: Increase cross-selling, upselling, and customer satisfaction.

2. Real-Time Analytics & Streaming Data

Objective: Monitor sales, inventory, and customer behavior in real time.

Tools: Apache Kafka, Spark Streaming, and real-time dashboards using Power BI or Grafana.

Impact: Enables faster business decisions and immediate response to market trends.

3. Integration with ERP and CRM Systems

Objective: Combine SuperStore data with broader enterprise systems for holistic analytics.

Use Case: Unified dashboards for sales, supply chain, customer feedback, and finance.

Impact: Greater operational efficiency and 360° business insight.

FUTURE SCOPE

4. Predictive Inventory and Demand Planning

Objective: Predict future demand per region or product category using AI.

Techniques: Deep learning (LSTM/GRU), Prophet, or AutoML.

Impact: Reduce overstocking or stockouts, optimize warehouse management.

5. Geospatial Analysis

Objective: Analyze location-based sales patterns using maps and GIS data.

Tools: QGIS, Plotly, GeoPandas, Tableau Map Layers.

Impact: Optimize delivery routes, open new stores in high-potential regions.

6. Sentiment Analysis from Customer Reviews

Objective: Extract customer sentiment from feedback or reviews.

Techniques: NLP using BERT, RoBERTa, or LLMs (ChatGPT API for summarization).

Impact: Improve product quality, customer service, and retention.

FUTURE SCOPE

7. AI Chatbots for Sales & Support

Objective: Deploy AI chatbots trained on SuperStore data to guide customers.

Use Case: Automated helpdesk, guided shopping assistants.

Tools: OpenAI, Dialogflow, Rasa.

Impact: Better customer engagement and reduced support costs.

8. Ethical AI & Bias Detection

Objective: Ensure fairness and explainability in AI models (e.g., customer segmentation, credit scoring).

Tools: SHAP, LIME, AI Fairness 360.

Impact: Trustworthy AI systems and regulatory compliance (e.g., GDPR).

9. Augmented Reality (AR) & Visual AI for Retail

Objective: Visual product recommendations and in-store AR experiences.

Techniques: Computer vision, AR SDKs, Visual Search.

Impact: Next-generation retail experience.

REFERENCES

- 🕮 **Books:**

1. **"Data Science for Business"** by Foster Provost & Tom Fawcett
2. **"Python for Data Analysis"** by Wes McKinney
3. **"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"** by Aurélien Géron

- 📄 **Research Papers:**

1. *Aggarwal, C. C. (2015). Data Mining: The Textbook* – Comprehensive methods in clustering, classification, and forecasting.
2. *Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice* – For ARIMA and time series modeling.
3. *Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system* – Use for boosting models on tabular data.

IBM CERTIFICATIONS

In recognition of the commitment to achieve professional excellence

Getting Started with Artificial Intelligence
IBM SkillsBuild

Shrrivathsan S

Has successfully satisfied the requirements for:

Getting Started with Artificial Intelligence

Issued on: Jul 16, 2025
Issued by: IBM SkillsBuild

Verify: https://www.credly.com/badges/dd6aafea-ee76-47bb-aaeb-7b4a4f7bb010

IBM

IBM CERTIFICATIONS

IBM CERTIFICATIONS

IBM **SkillsBuild** Completion Certificate

This certificate is presented to

Shrrivathsan S

for the completion of

Lab: Retrieval Augmented Generation with LangChain

(ALM-COURSE_3824998)

According to the Adobe Learning Manager system of record

Completion date: 19 Jul 2025 (GMT) **Learning hours:** 20 mins

THANK YOU

edunet foundation