Project Report

|  |  |
| --- | --- |
| **Project Title** | Fraud Detection for Financial Transactions |
| **Qualification Name** | Advanced Certificate in Data Science (E-Learning) |
| **Product Name** | Advanced Certificate in Data Science (E-Learning) |
| **Module Name** | WSQ Captone Project-data analytics (SF) |

|  |  |  |  |
| --- | --- | --- | --- |
| **Student name** | | **Assessor name** | |
| Tan Tiong Choon (Chen ZhongChun) | | Ryan Suryanto | |
| **Date issued** | **Completion date** | | **Submitted on** |
| 16 Jan 2024 | 23 February 2024 | | 23 February 2024 |

**Fraud Detection for Financial Transactions**

**Project title**

Date: 11 10 January 2024

Student signature:

I certify that the work submitted for this assignment is my own and research sources are fully acknowledged.

**Learner declaration**



23 February 2024

Table of Contents

1. [Project Background 2](#_TOC_250012)
2. [Project Objective 2](#_TOC_250011)
3. [Project Requirements Specifications 2](#_TOC_250010)
4. [Activity 1 2](#_TOC_250009)
5. [Activity 2 2](#_TOC_250008)
6. [Activity 3 2](#_TOC_250007)
7. [Activity 4 2](#_TOC_250006)
8. [Activity 5 2](#_TOC_250005)
9. [Activity 6 2](#_TOC_250004)
10. [Activity 7 2](#_TOC_250003)
11. [Activity 8 2](#_TOC_250002)
12. [Annexure (Excel Workbook) 3](#_TOC_250000)

# Project Background

## Financial institutions are being challenged of detecting fraudulent transactions to protect their

## customers and maintain the service integrity. Our client, a leading financial institution, is committed

## to enhance their fraud detection capabilities using data analytics and Azure Machine Learning.

## This Capstone Project aims to leverage Python and Azure Machine Learning to develop a robust

## system for real-time identification of suspicious financial transactions. By analyzing historical data,

## the goal is to minimize financial losses, enhance customer trust, and strengthen fraud prevention strategies.

# Project Objective

## Develop a highly accurate fraud detection model for real-time identification of suspicious financial transactions.

## Provide insights into the key features and patterns associated with fraudulent transactions to aid

## in fraud prevention and investigation.

## Implement the fraud detection model seamlessly into the client's transaction processing system

## using Azure Machine Learning, with goal to fortify security measures, minimize financial losses and

## bolster customer trust in the institution's services.

# Project Requirements Specifications

## Python is the programming language for preprocessing, feature engineering, exploratory data analysis:

## Data Collection: The project involves a diverse dataset of financial transactions details, ensuring data

## security and privacy compliance.

## Feature Engineering: To build relevant features such as transaction frequency, amounts, location and user behaviour patterns.

## Exploratory Data Analysis (EDA): Utilize visualization and statistical analysis to uncover patterns and

## anomalies in the transaction data to gain insights into fraudulent transaction characteristics.

## Azure Machine learning is used for model development and deployment:

## Model Selection: Choose appropriate machine learning algorithms for fraud detection, leveraging

## classification and anomaly detection techniques with Azure Machine Learning.

## Model Evaluation: Develop robust evaluation metrics such as accuracy, precision, recall, F1-score, and ROC AUC to provide actionable insights for real-time fraud detection.

## Model Deployment: Using Azure Machine Learning to integrate the fraud detection model into the client's transaction processing system, considering real-time scoring or batch processing.

## Interpretability: Ensure model's interpretability for fraud analysts to understand a flagged transaction,

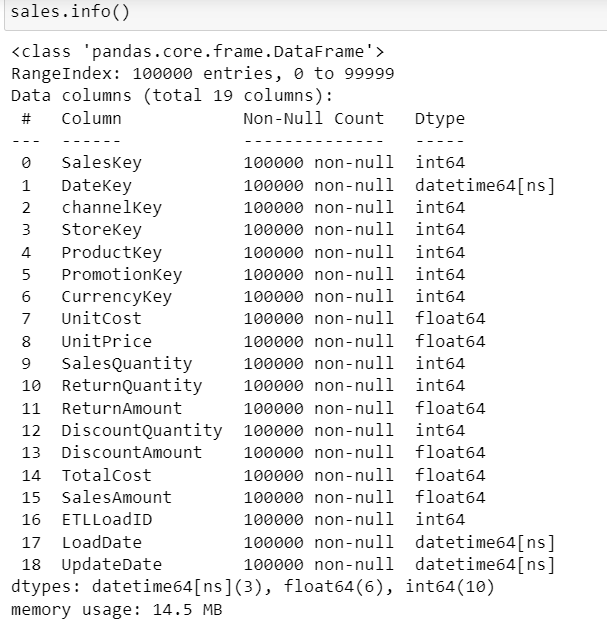
## aiding in investigation and decision-making.

# Activity 1 : Data Collection and Preprocessing

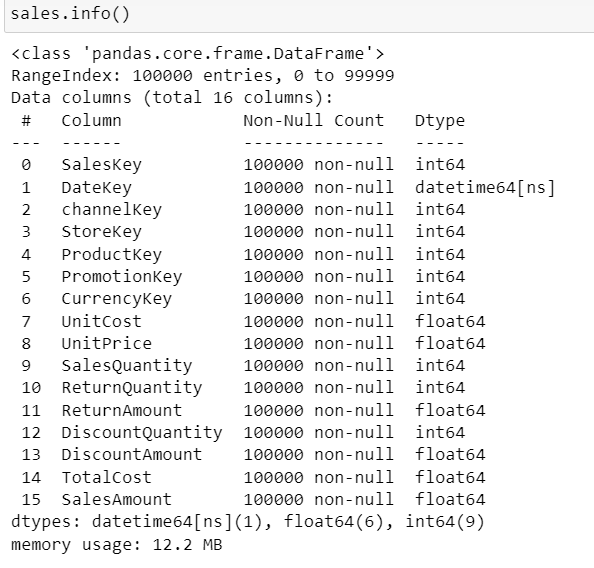
* Collect a diverse dataset of financial transactions, ensuring data security and compliance.
* Preprocess the data, handle missing values, and perform data cleansing.

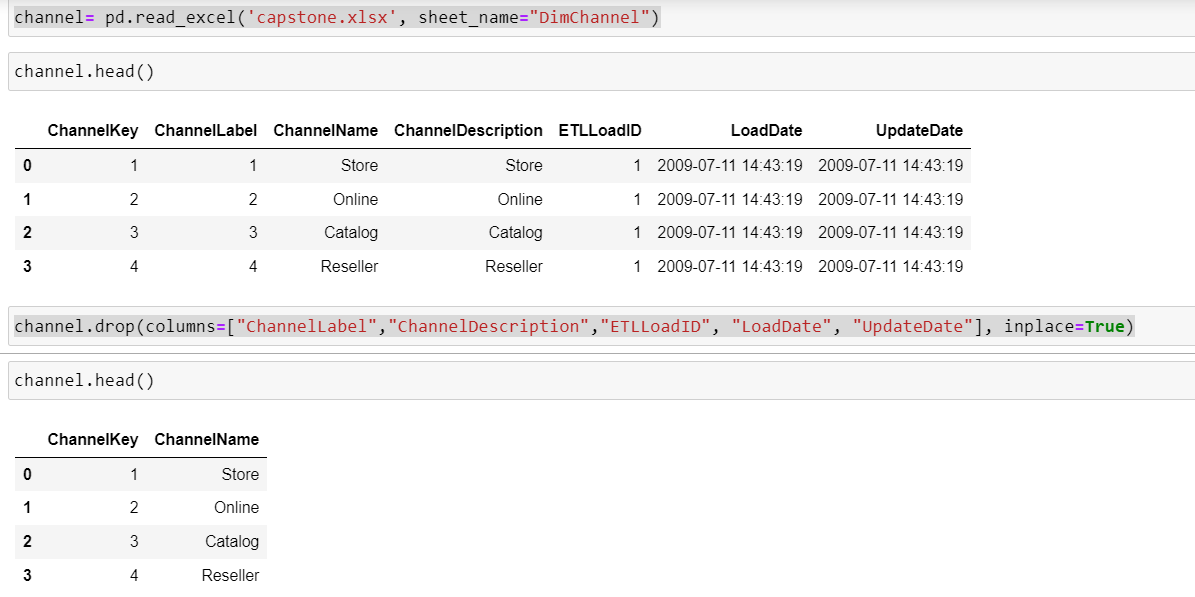


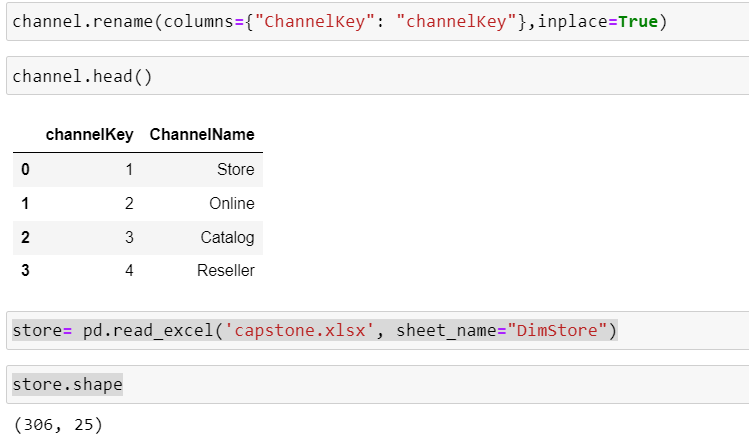


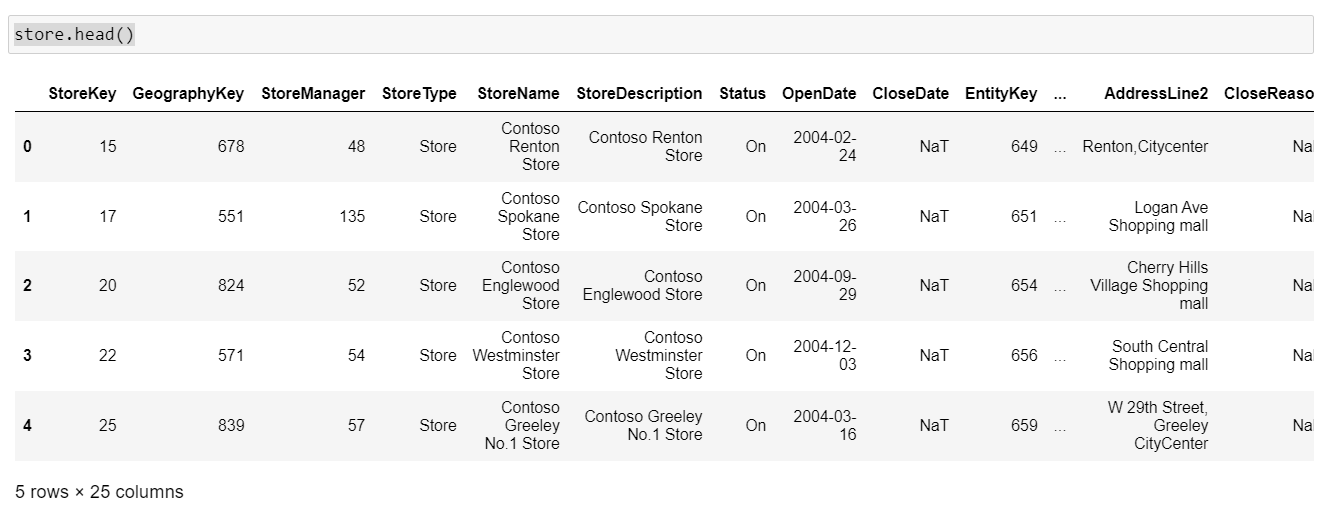


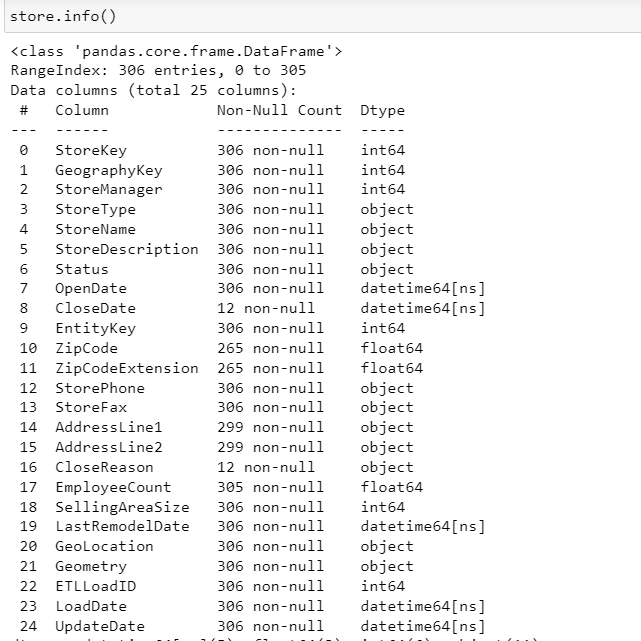


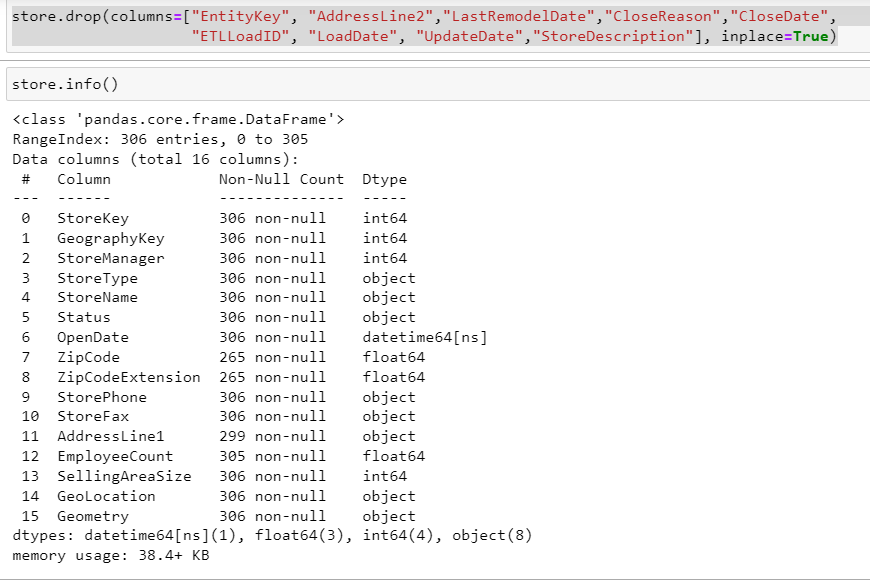


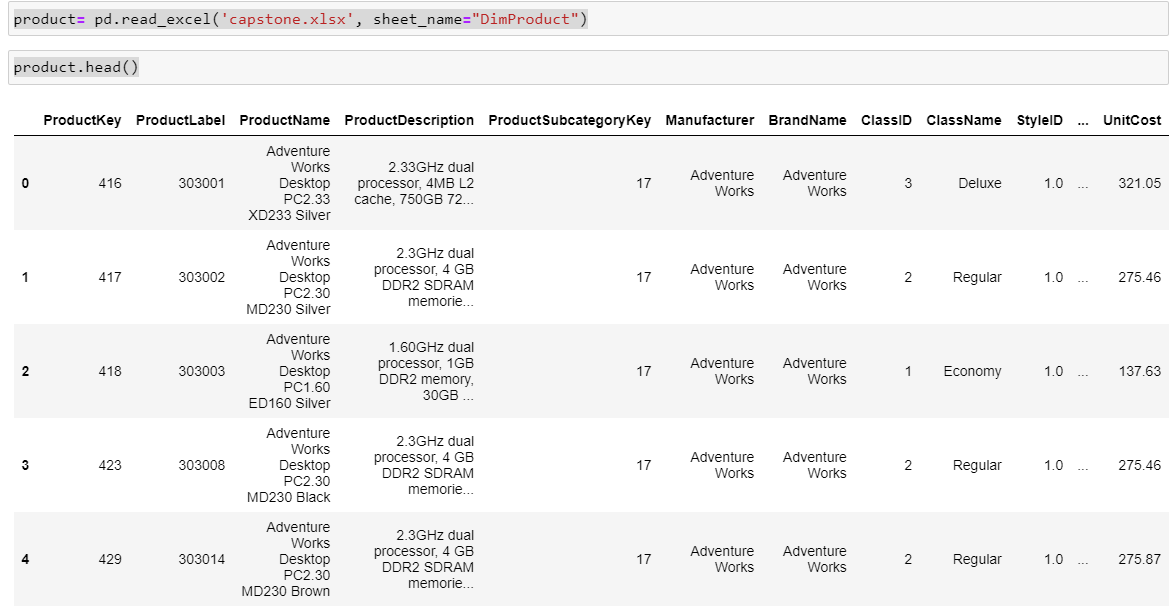


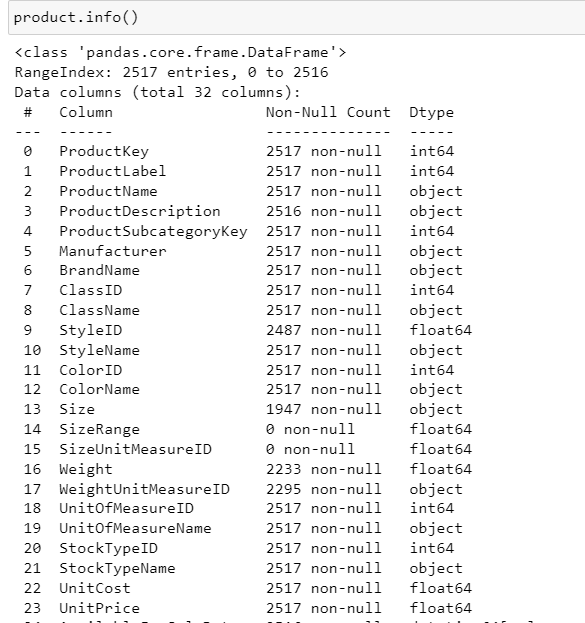


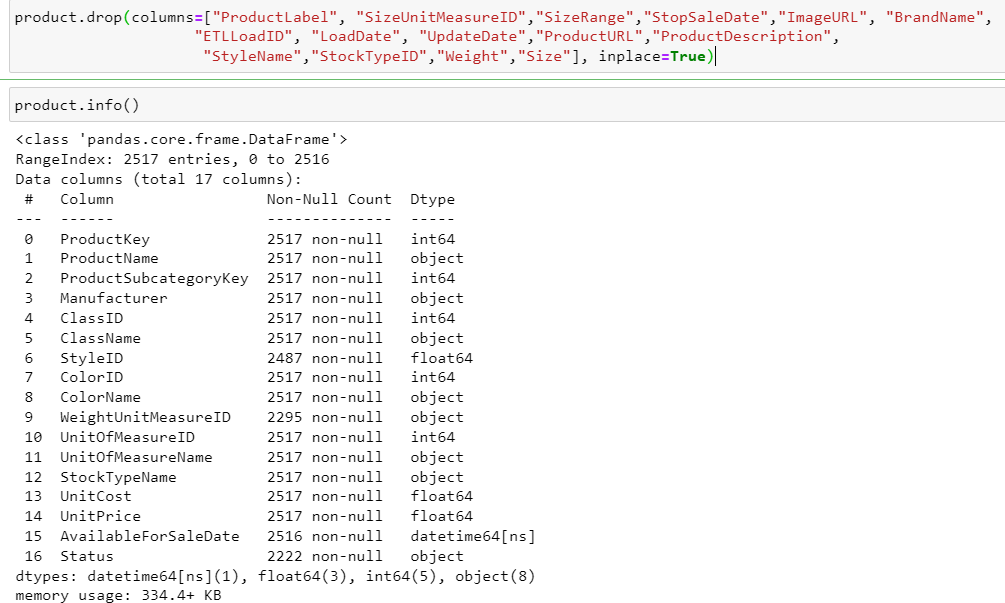


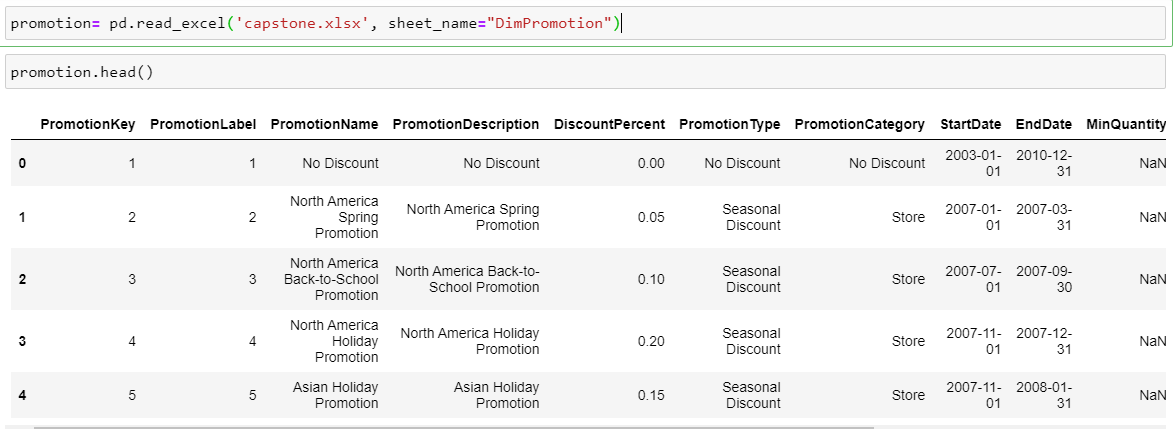


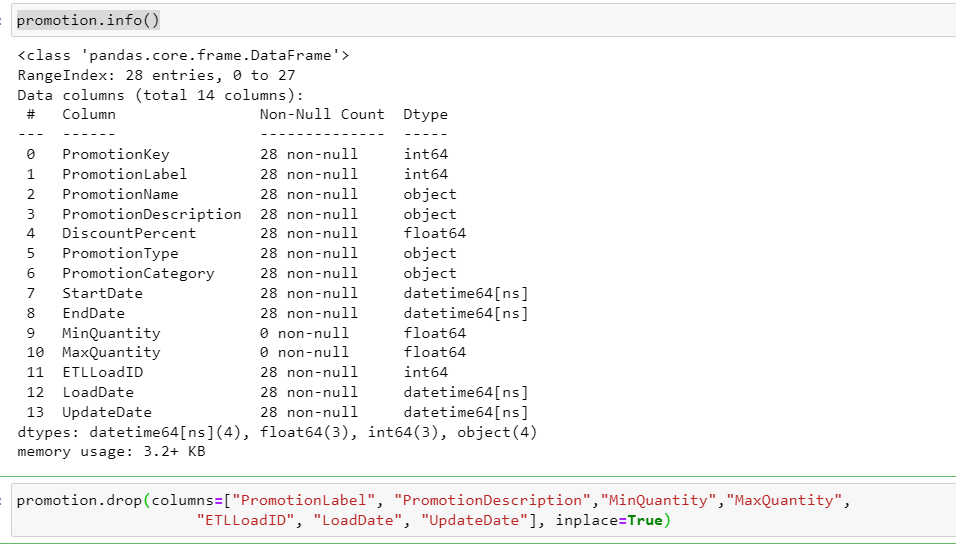


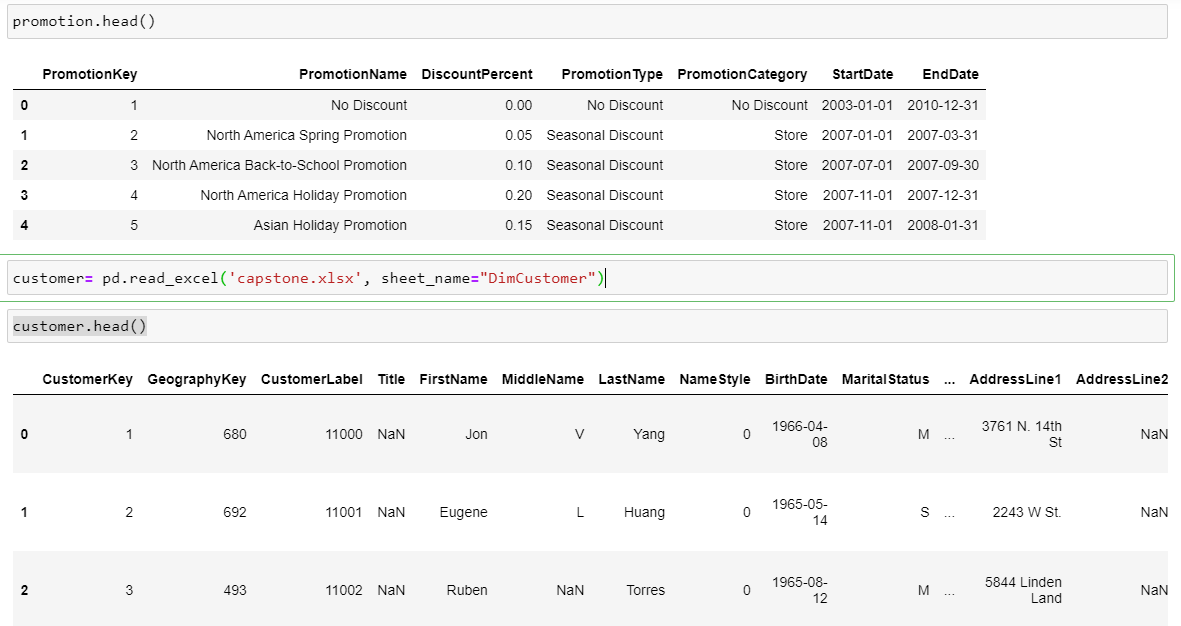


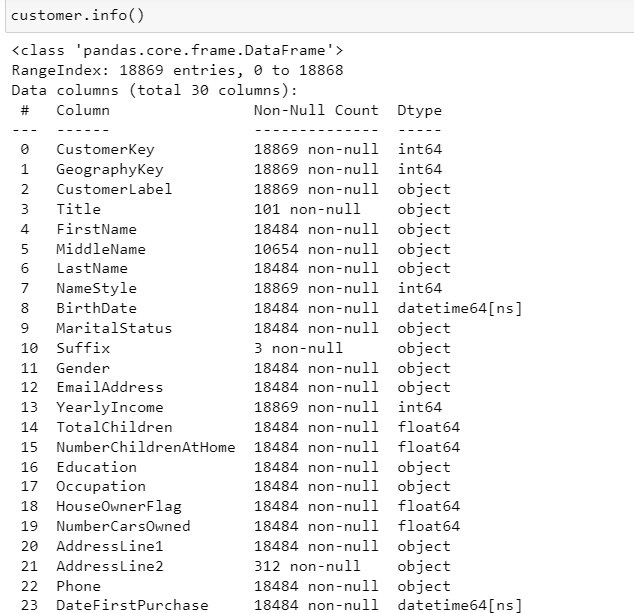


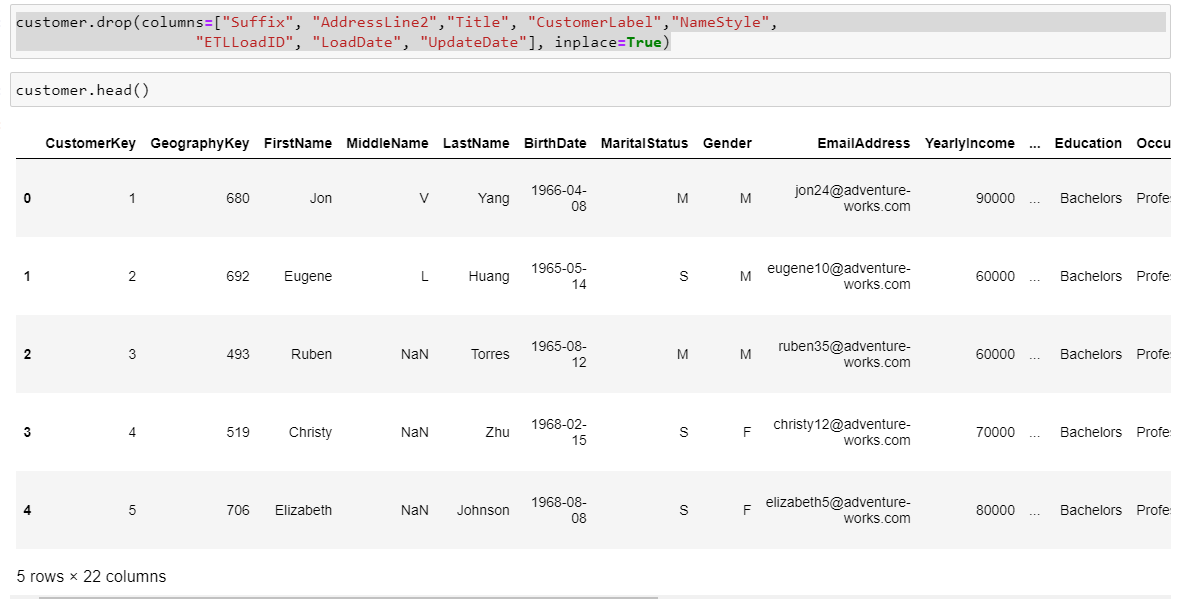


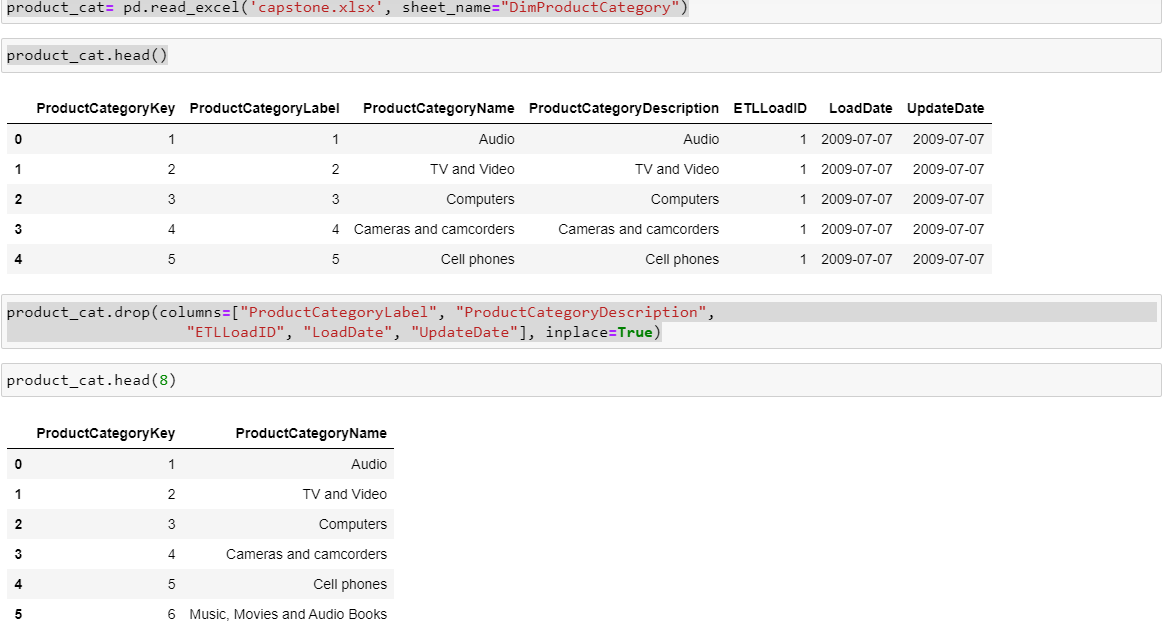


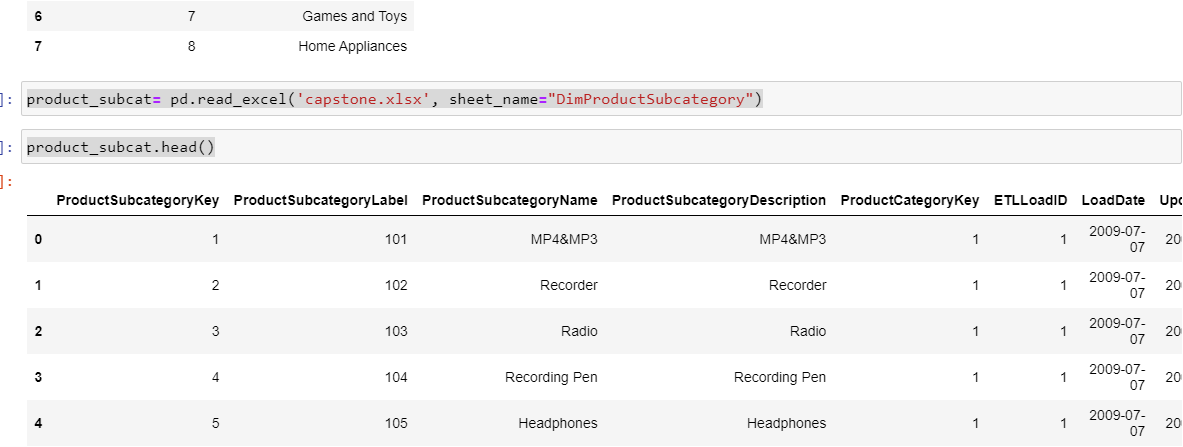


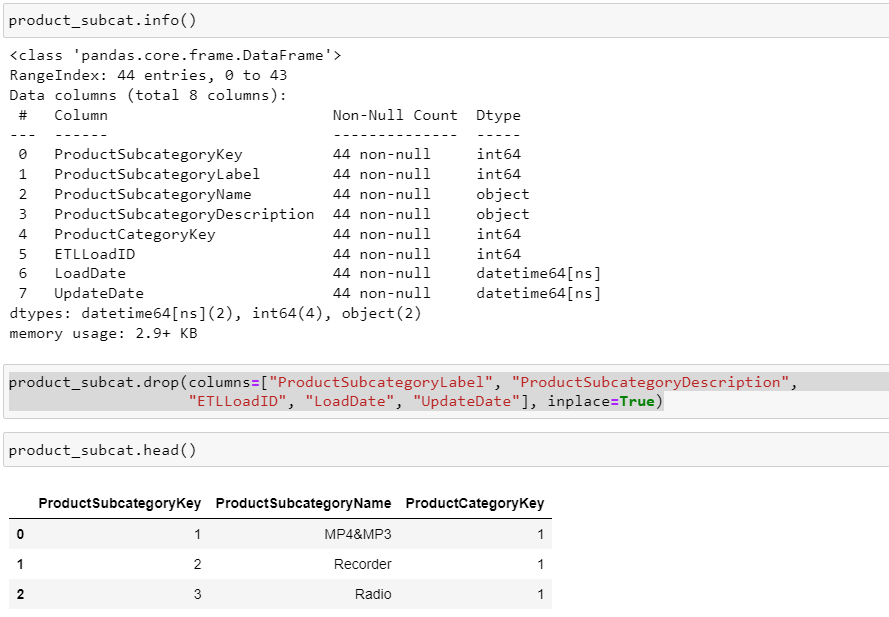


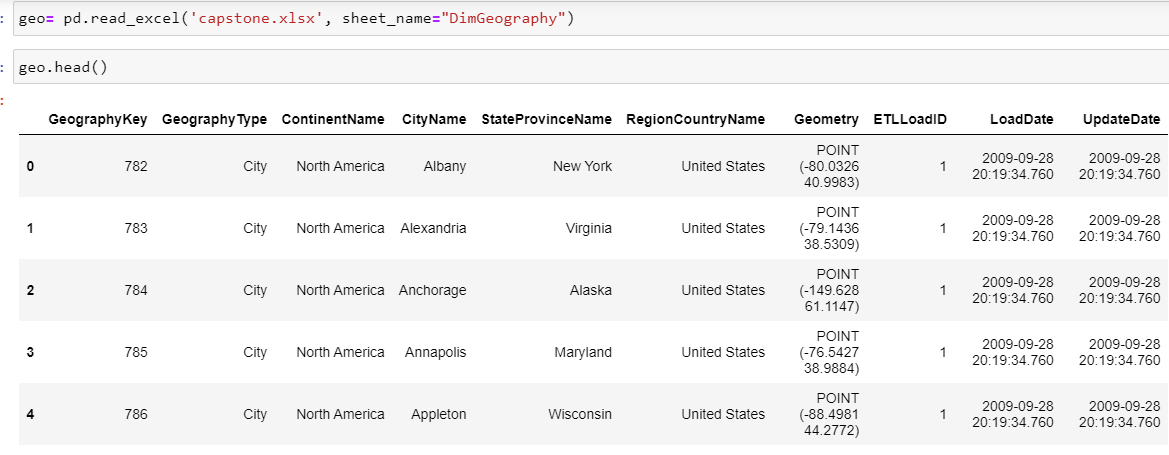




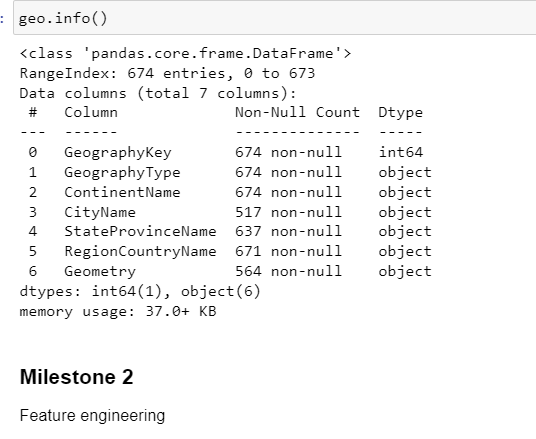






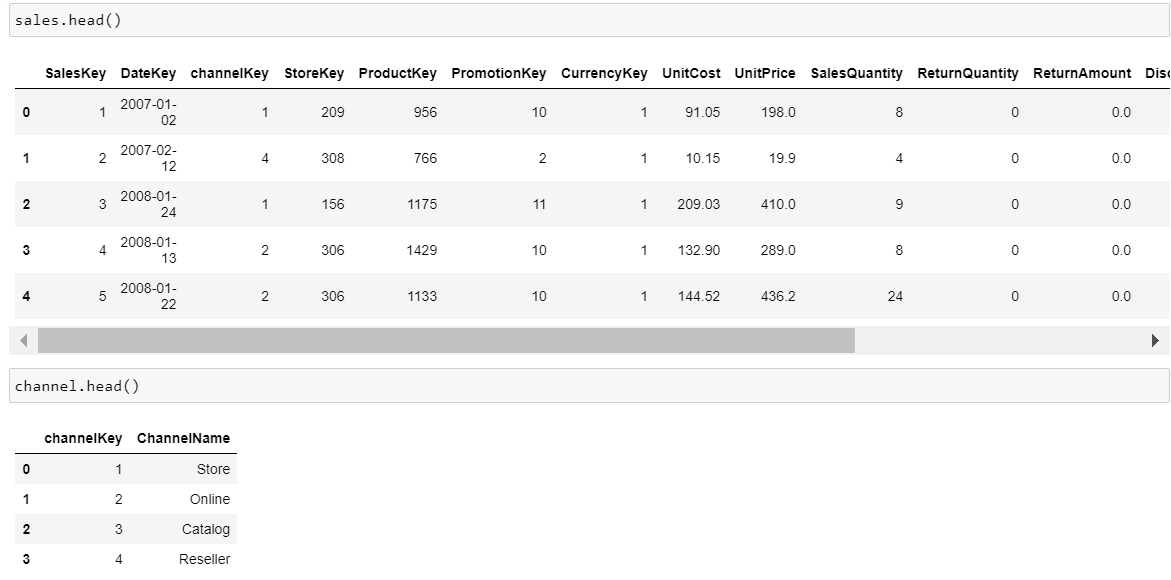


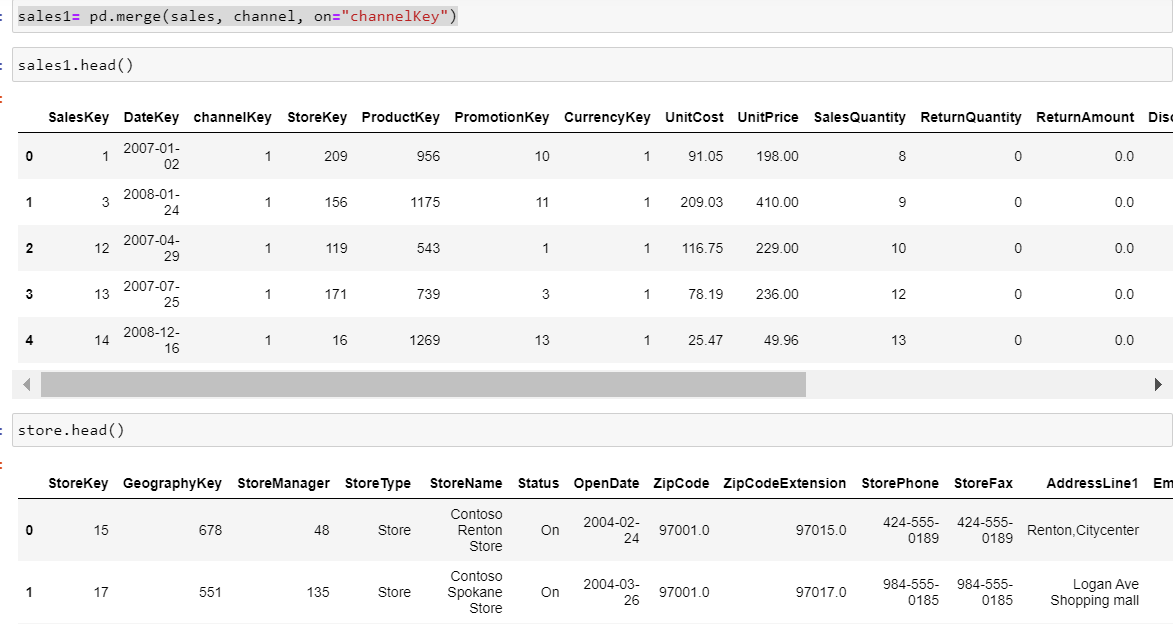


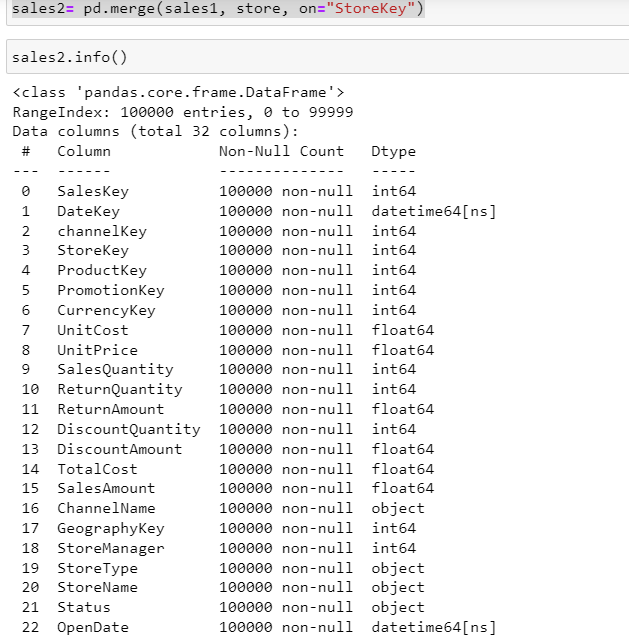


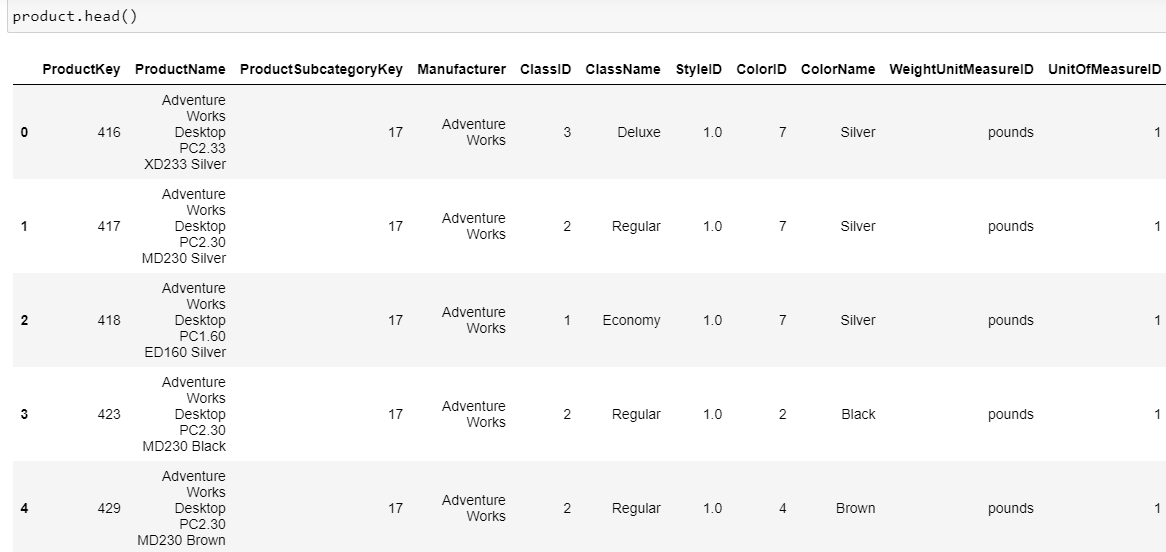
# Activity 2 : Feature Engineering

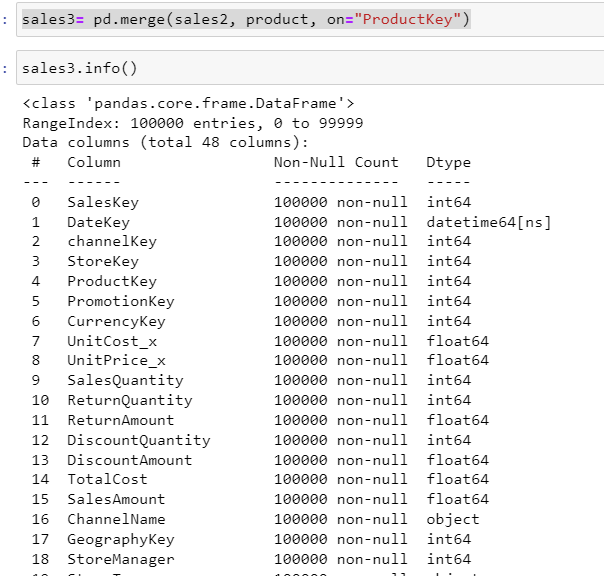
* Engineer relevant features from the transaction data, including transaction frequency, amounts, and user behavior patterns.

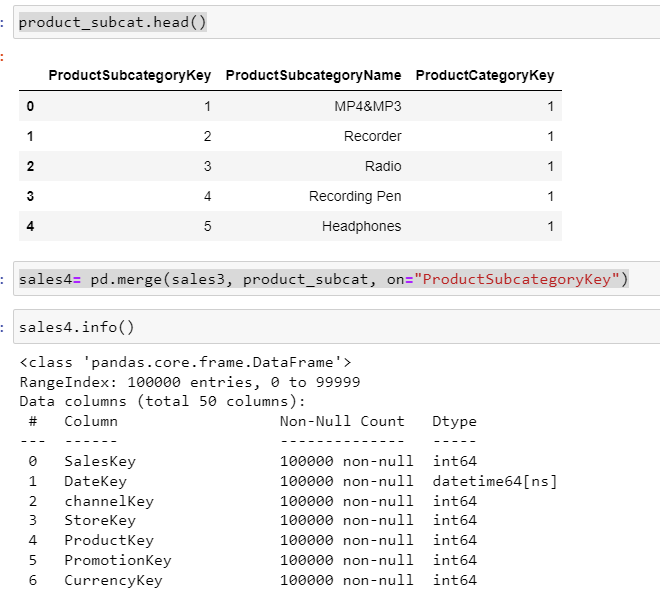


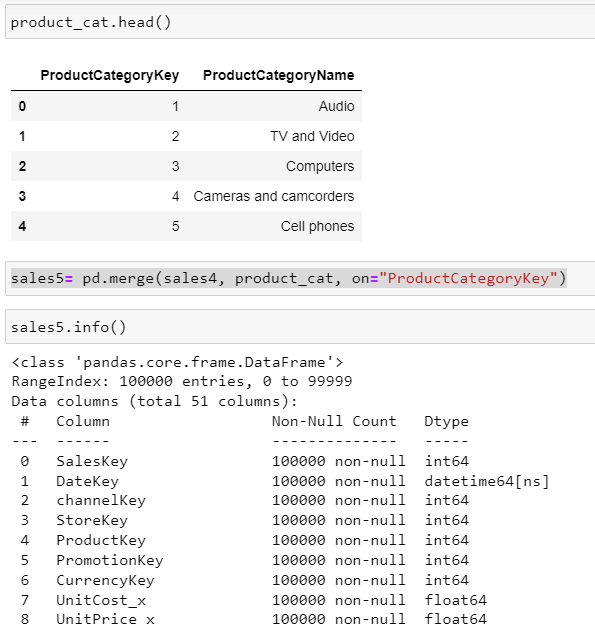


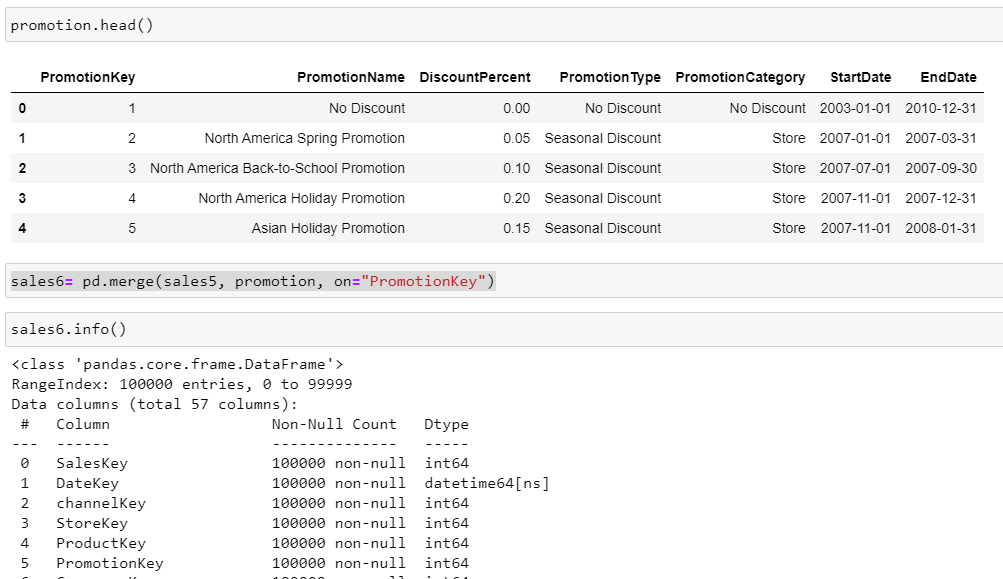




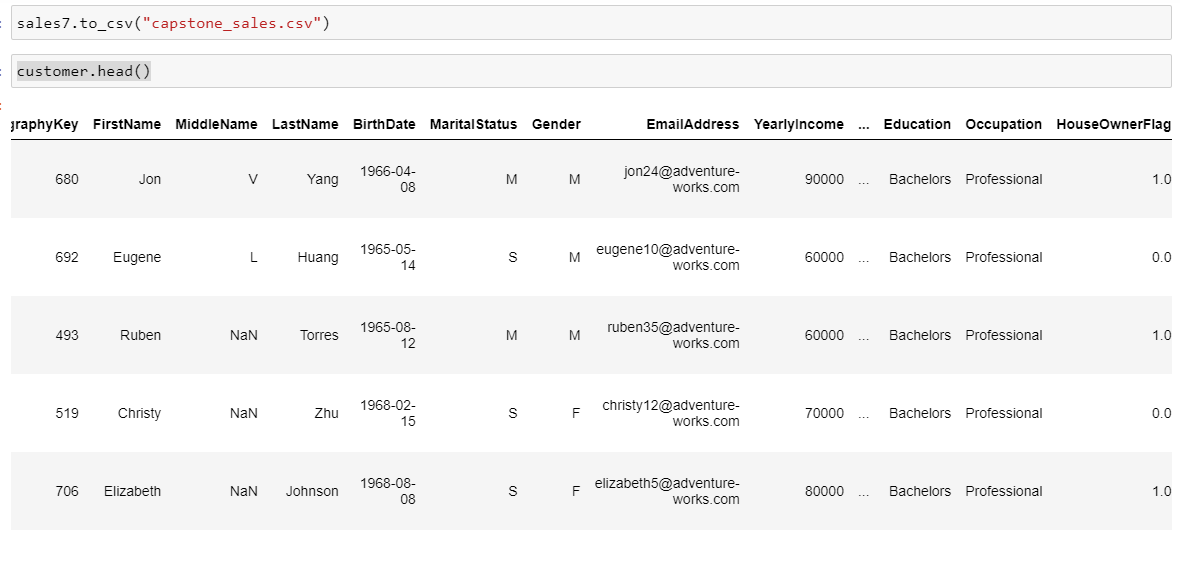


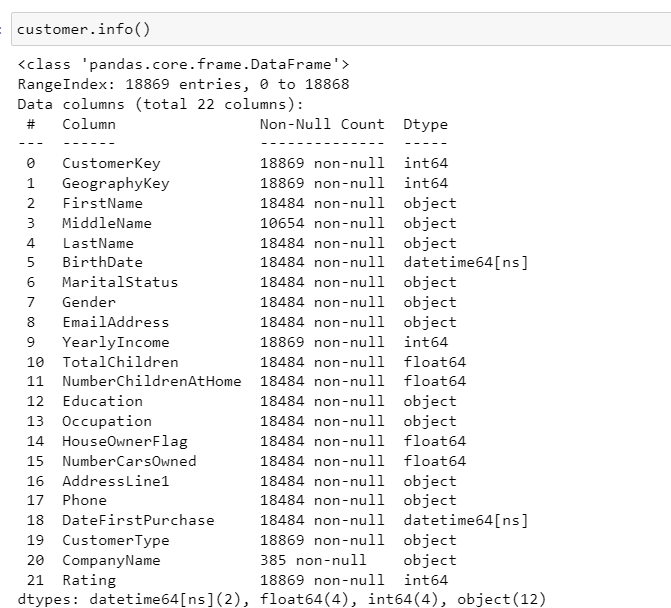


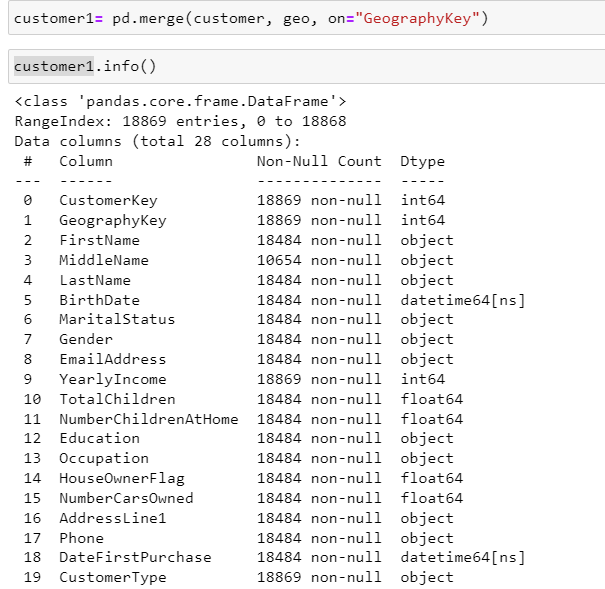


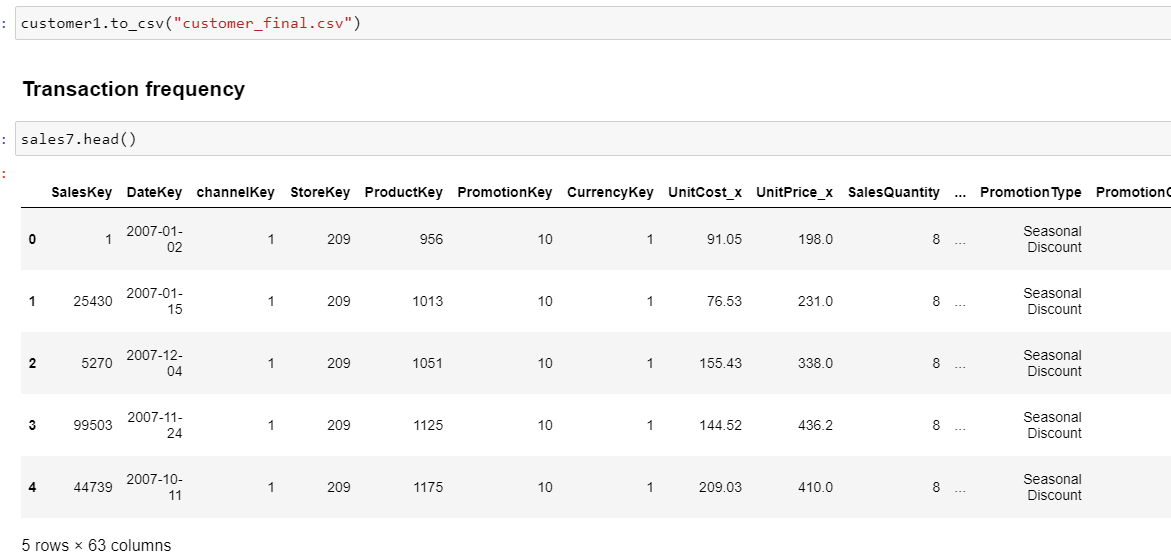


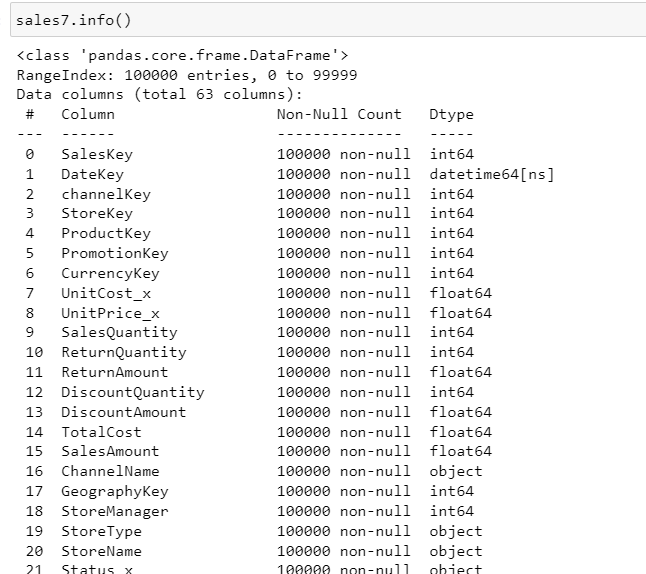


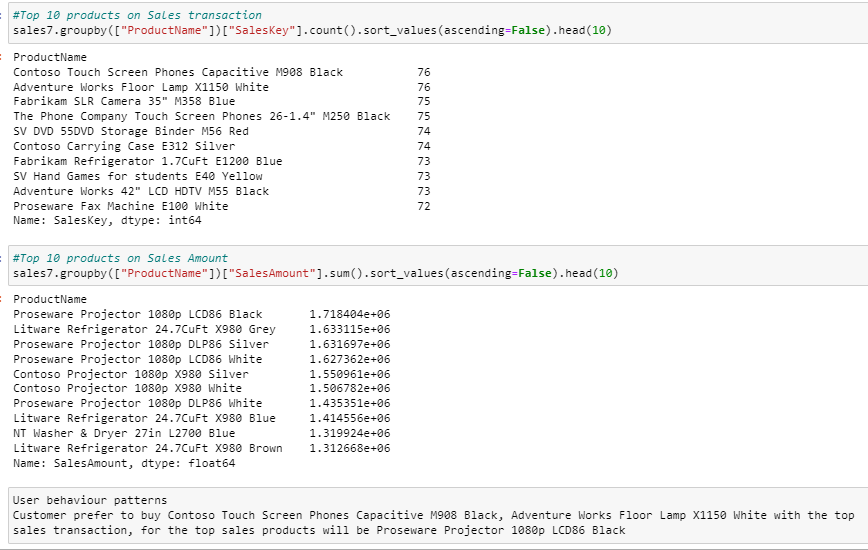






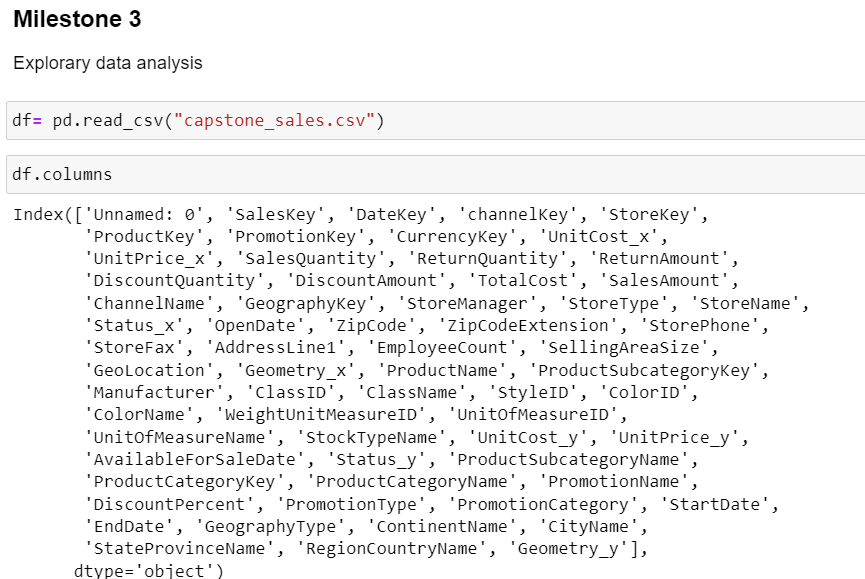


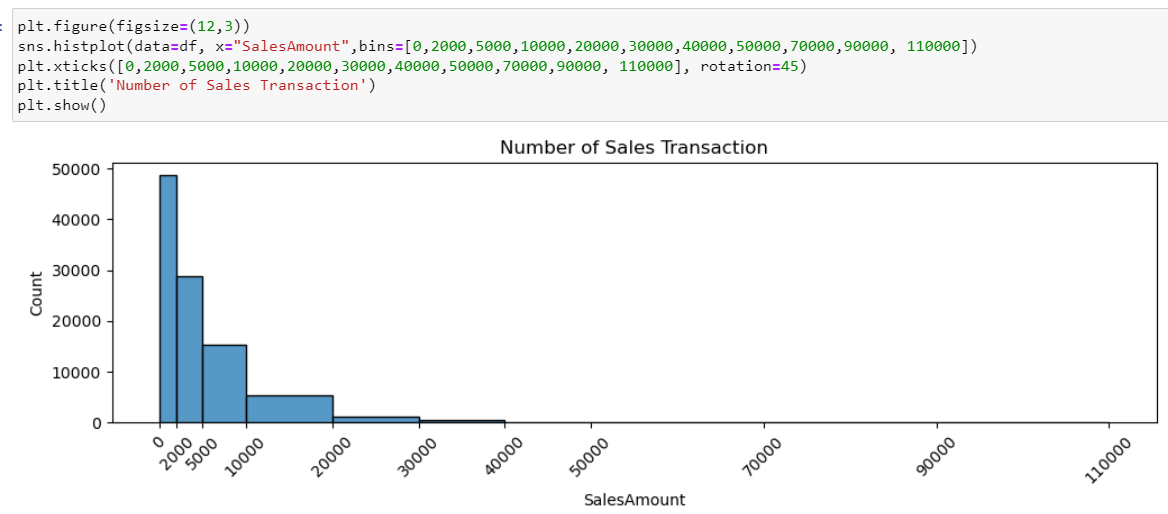


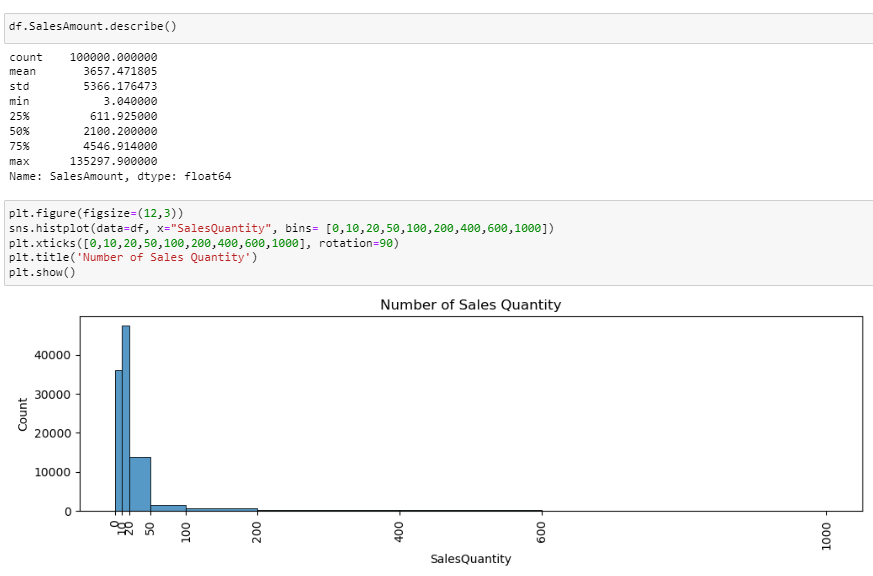


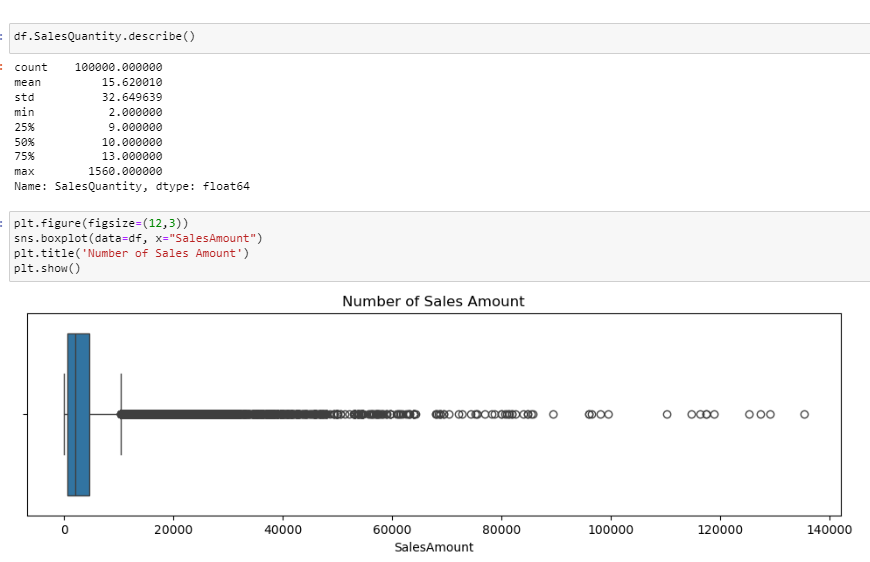
# Activity 3 : Exploratory Data Analysis

* Conduct comprehensive EDA to uncover patterns and anomalies in the transaction data.
* Visualize key insights and perform statistical analysis.

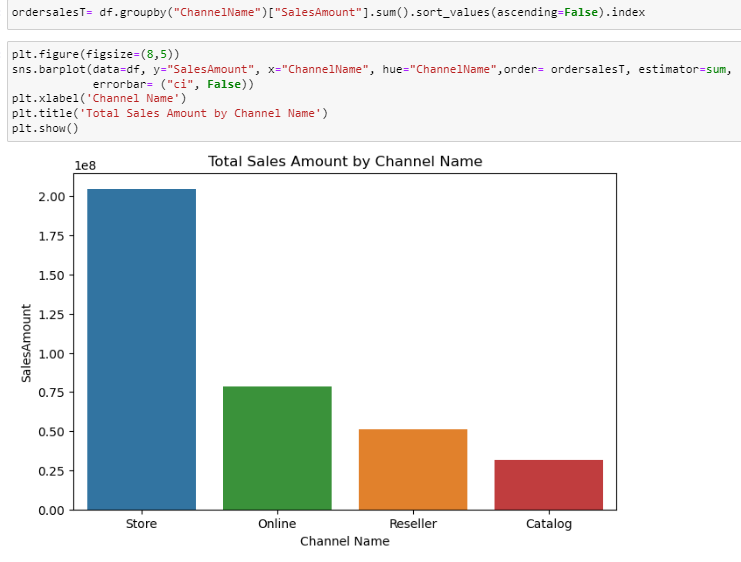


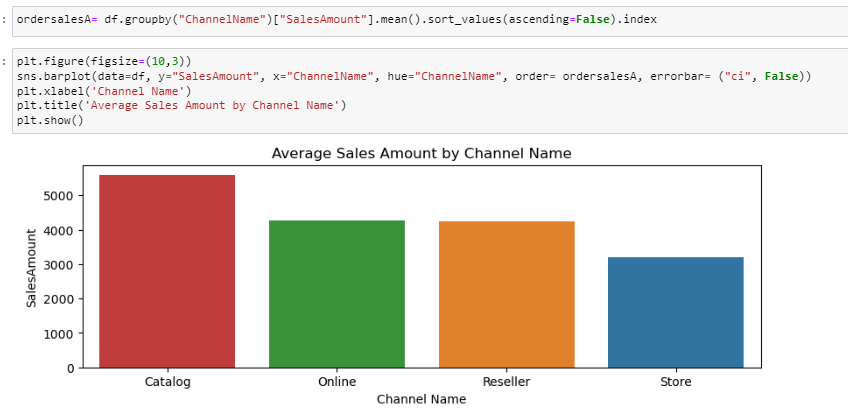




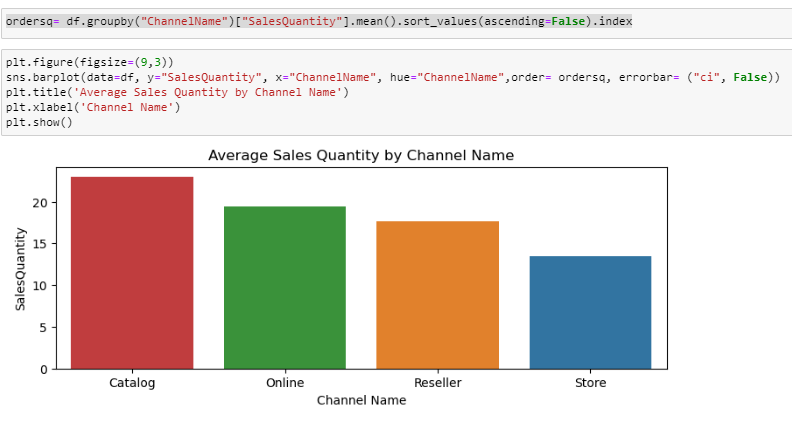


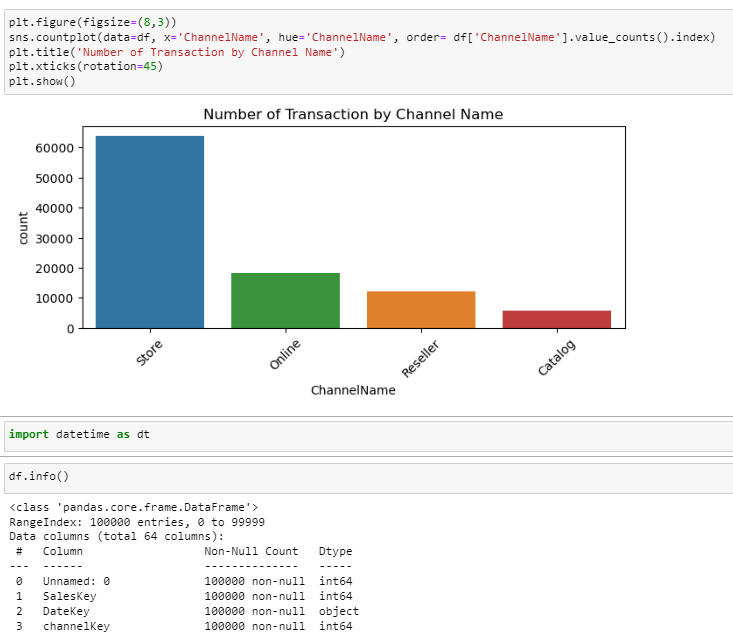




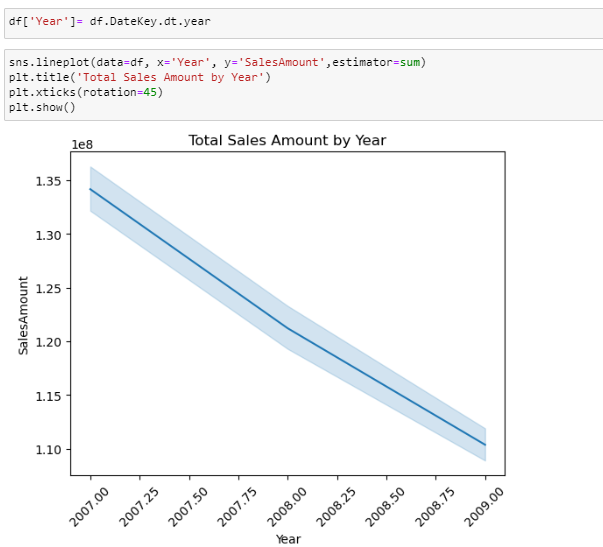


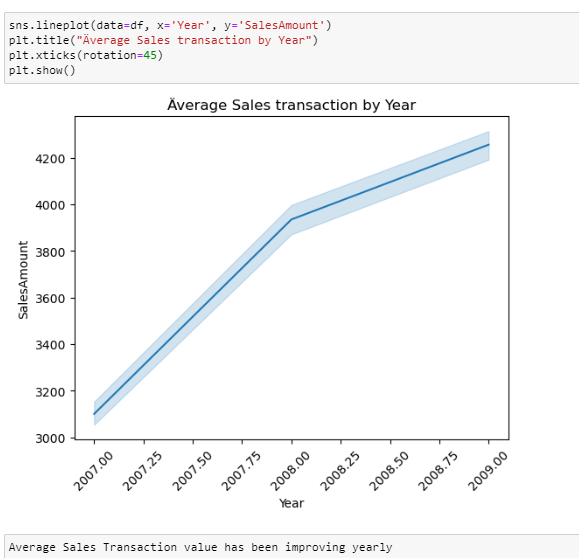


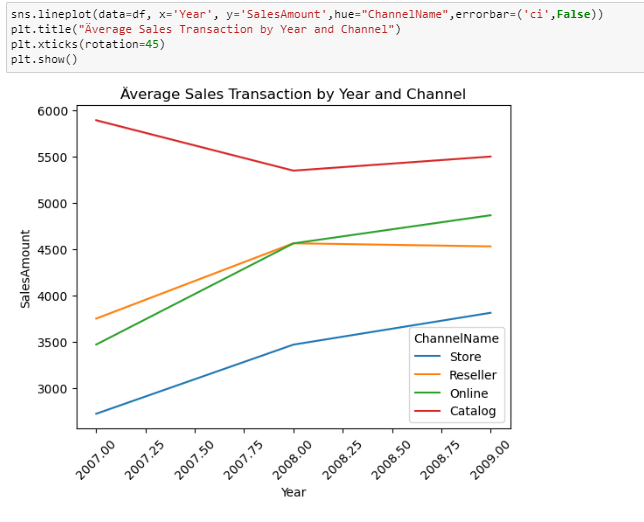




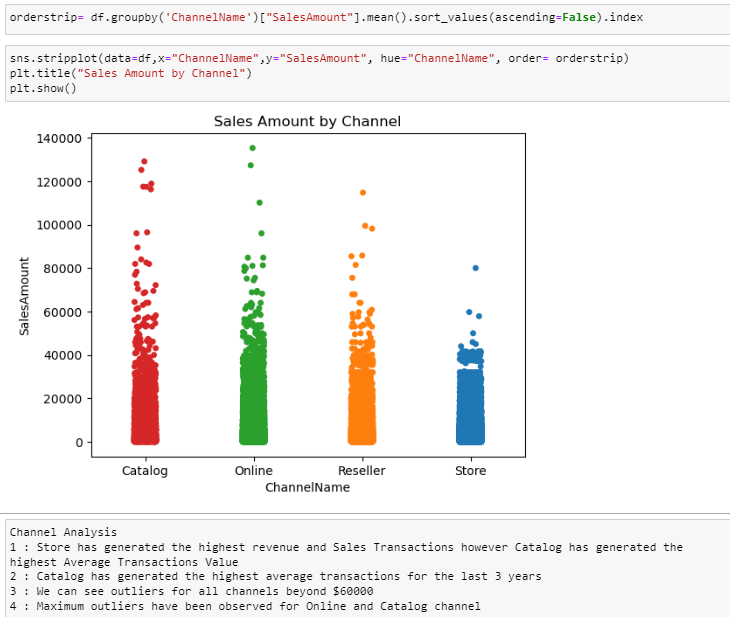


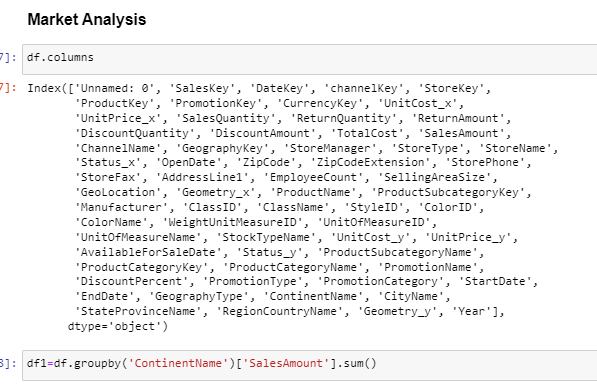


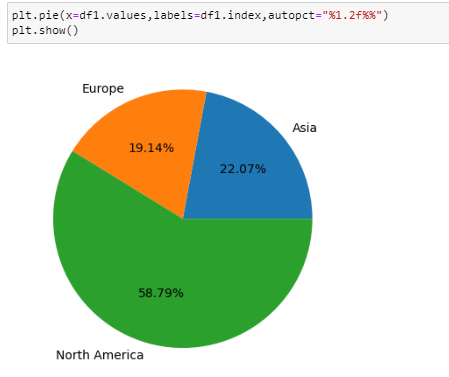


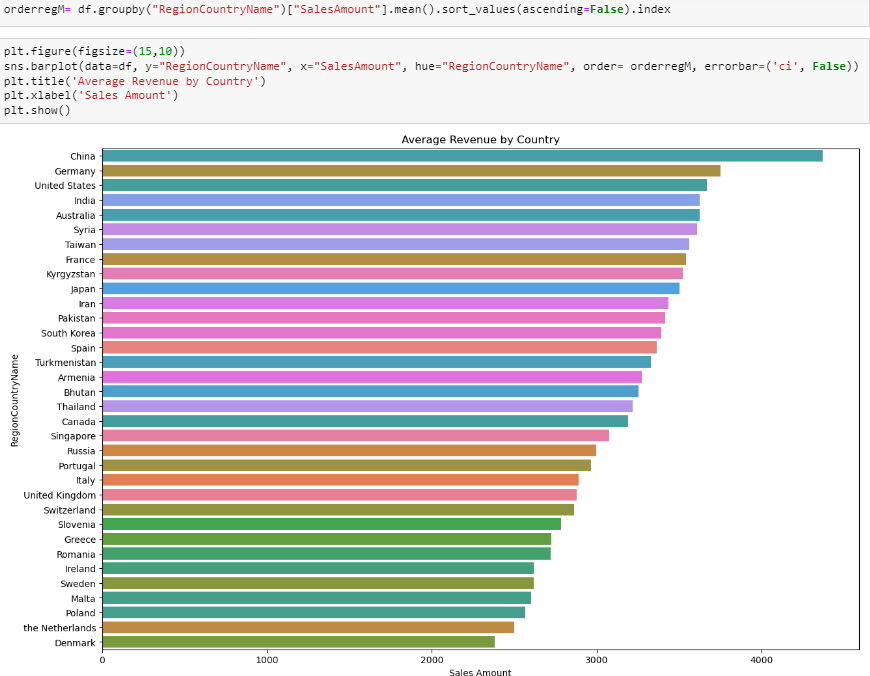


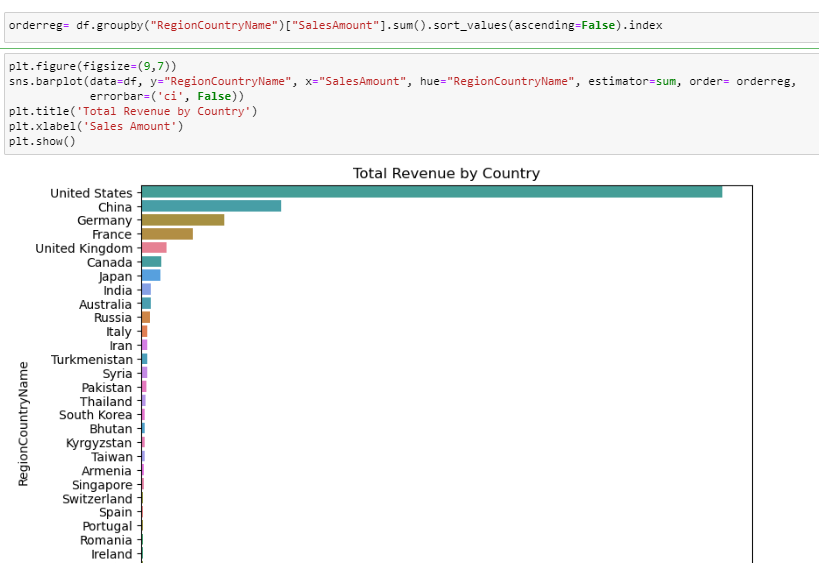


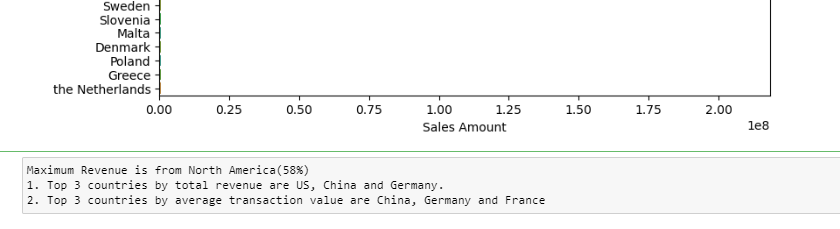


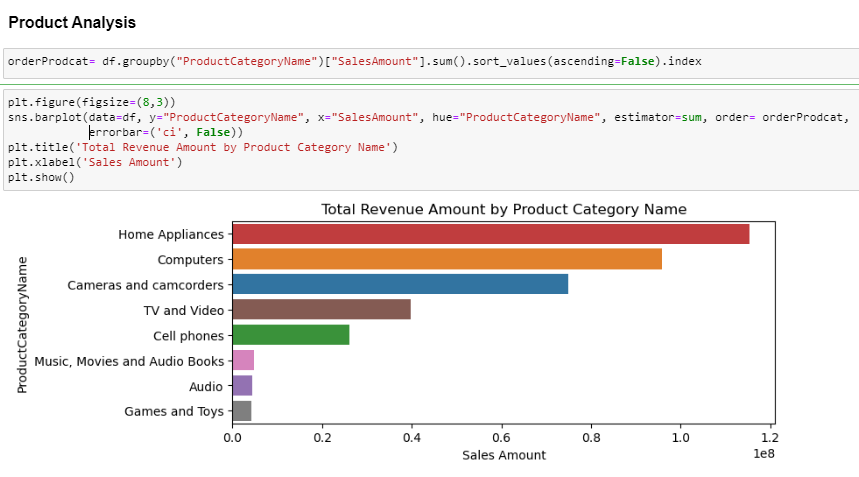


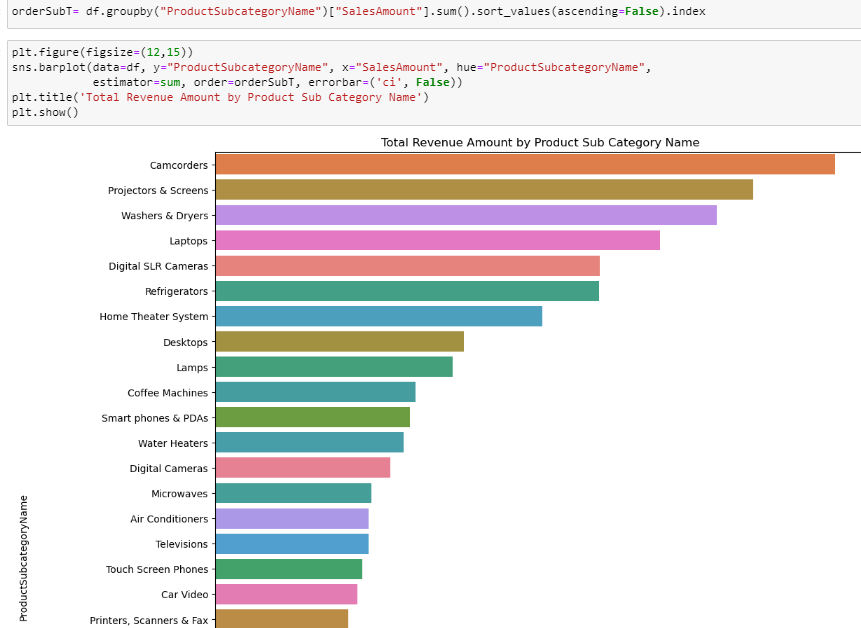


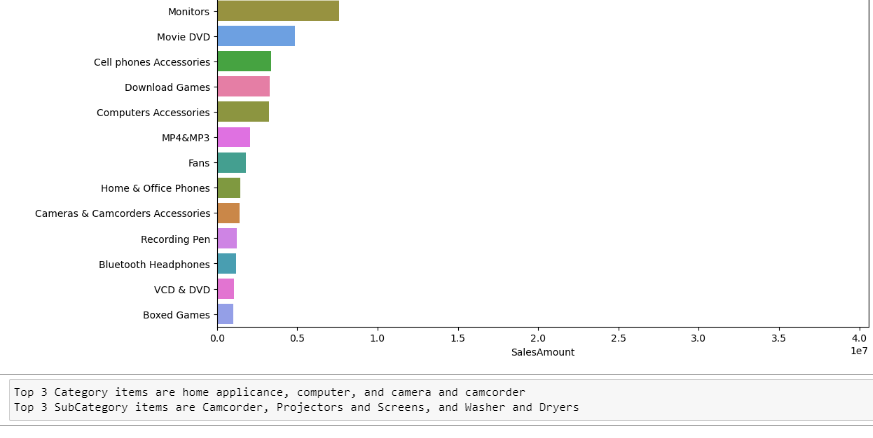


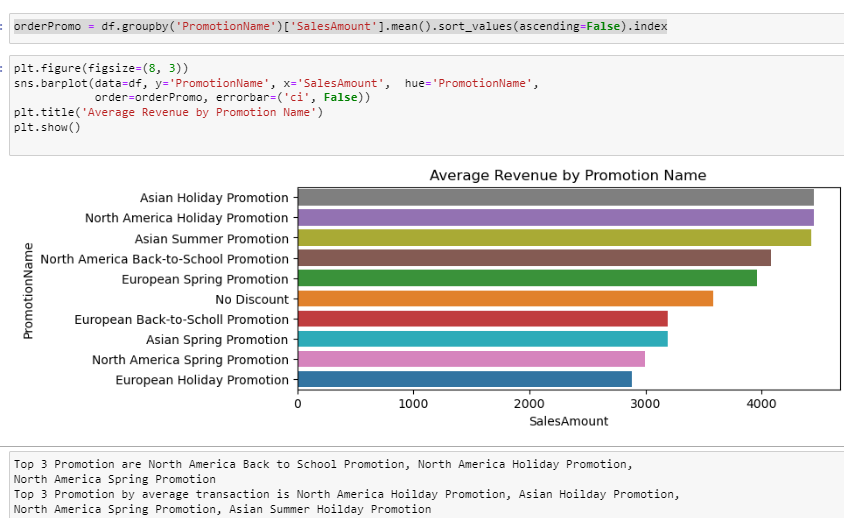
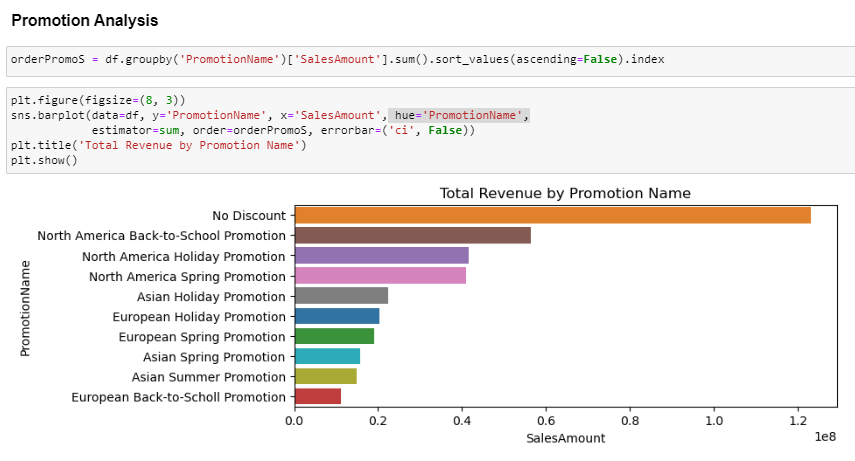


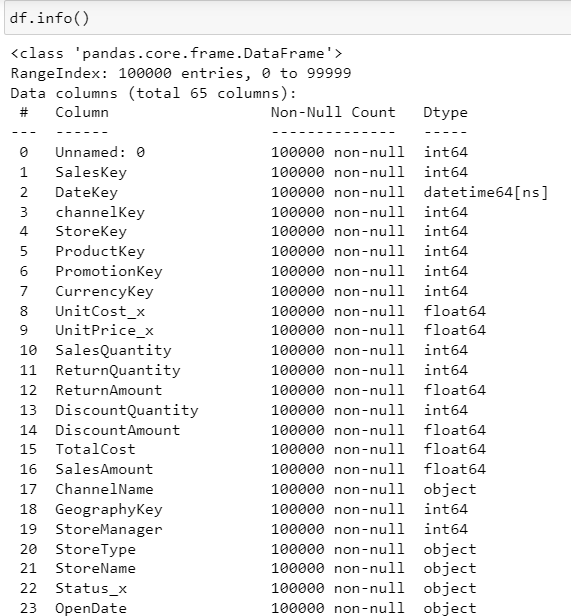


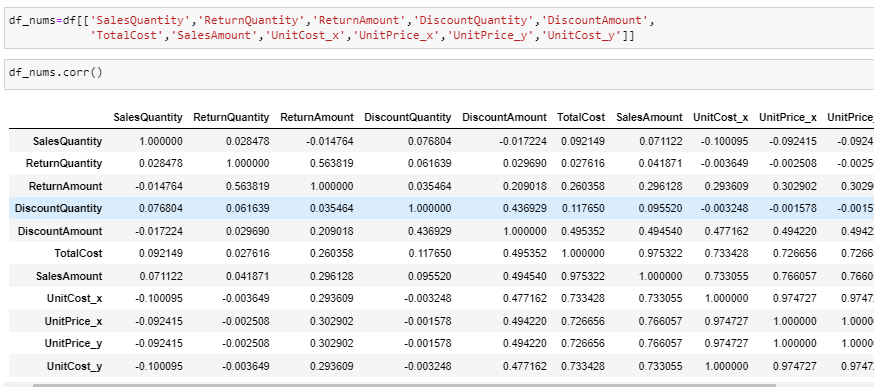


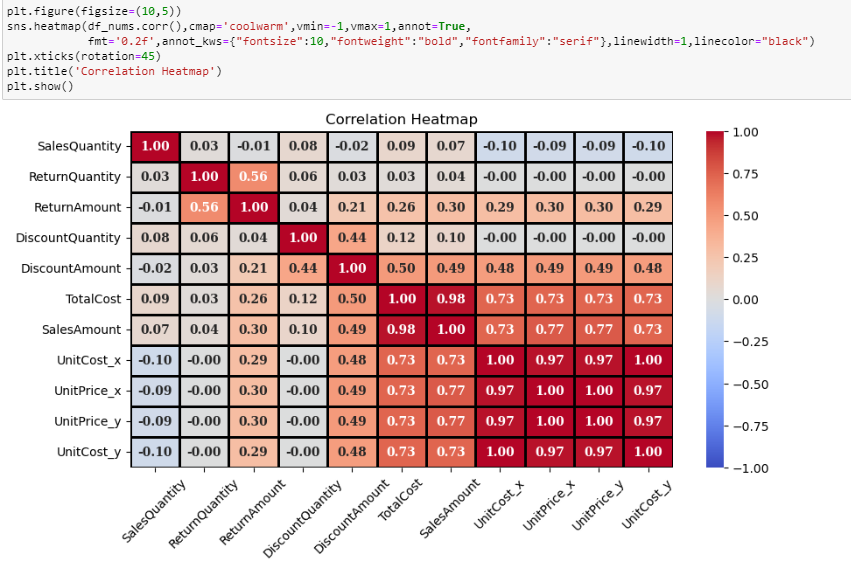


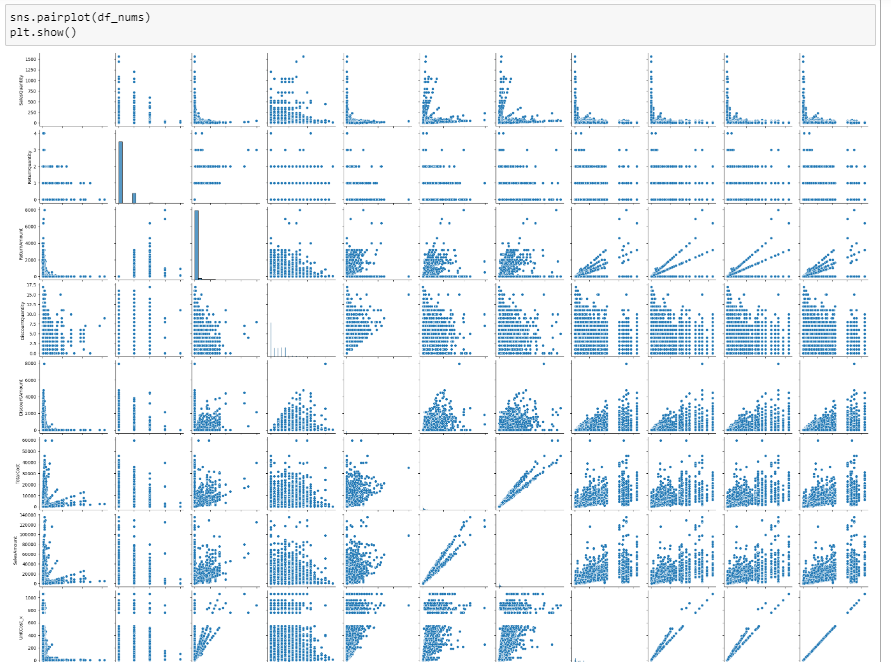


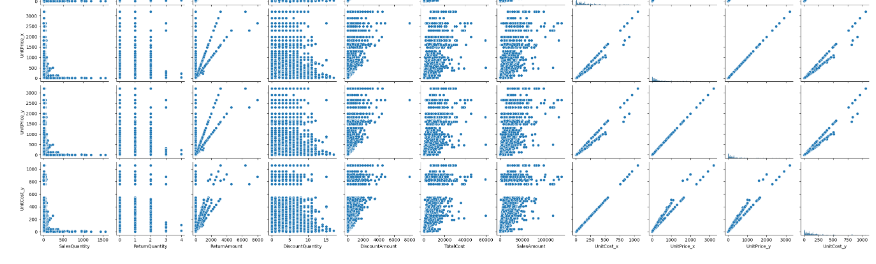


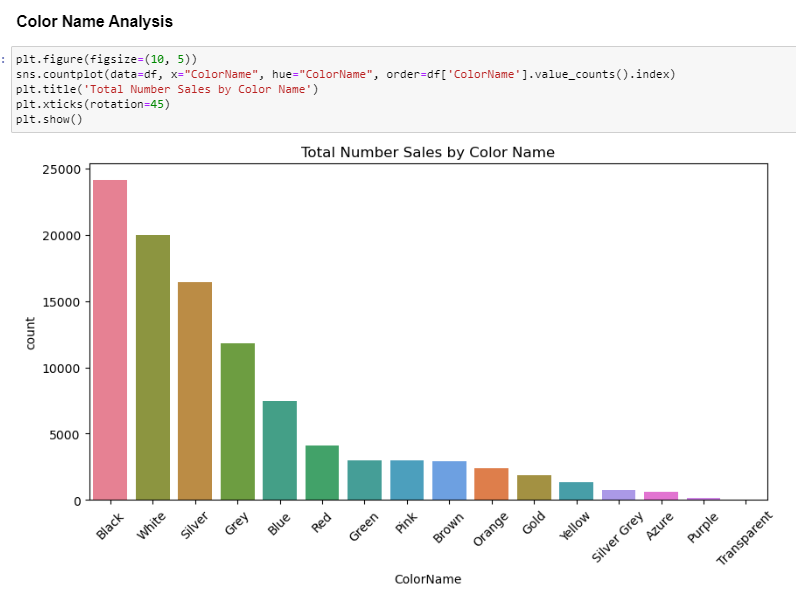


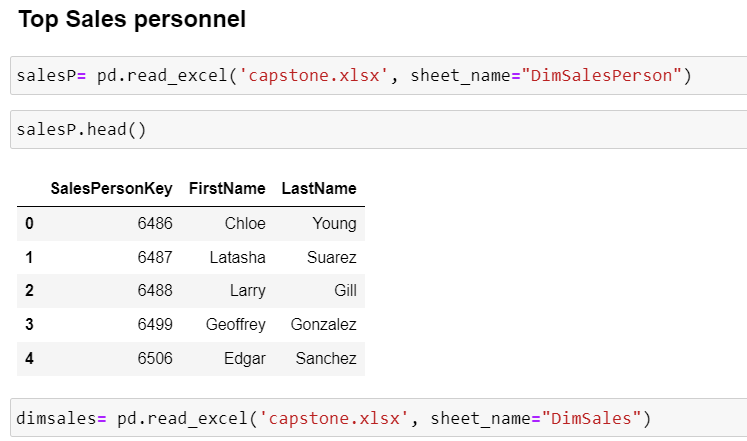


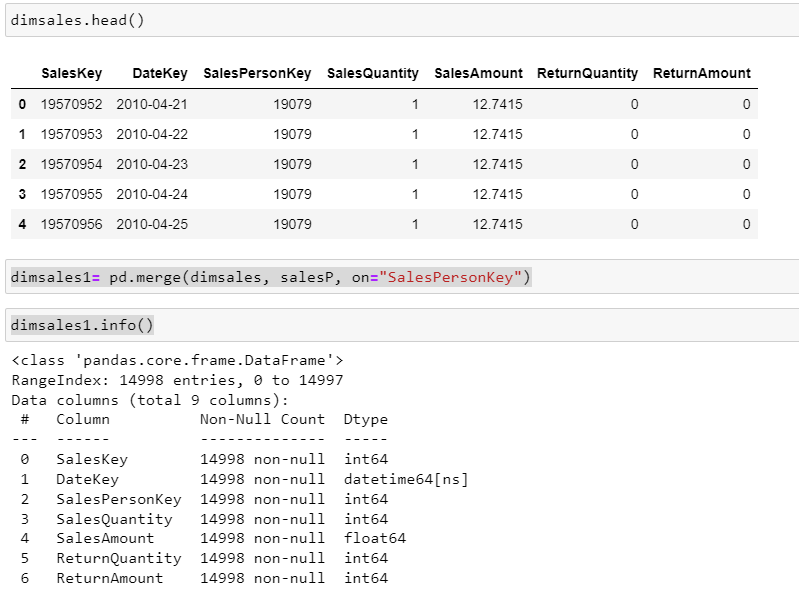


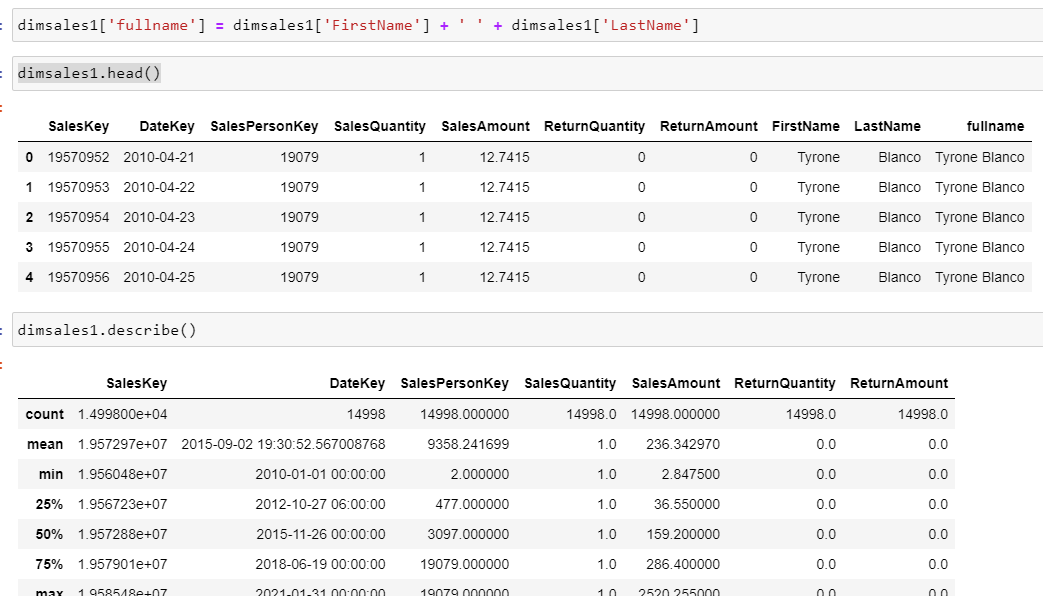


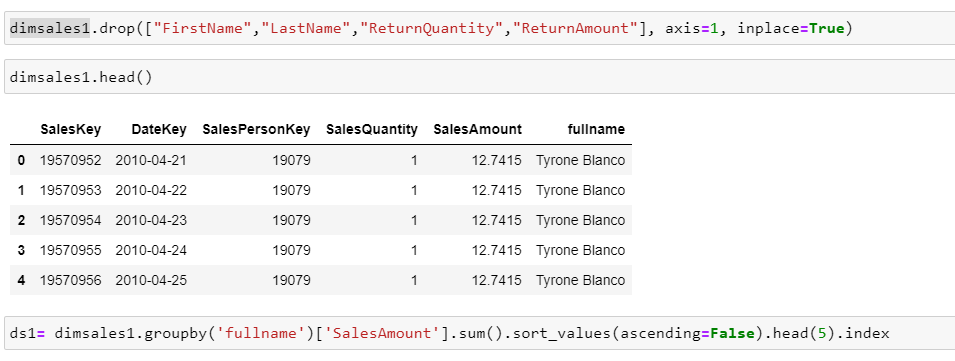


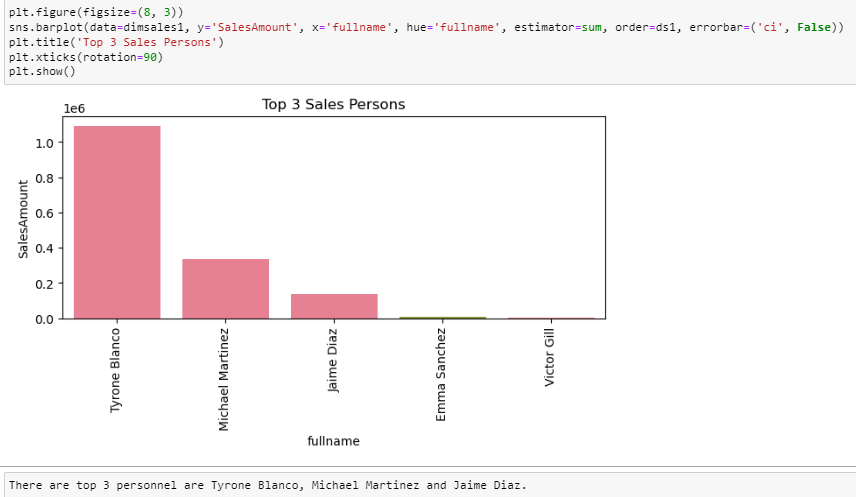


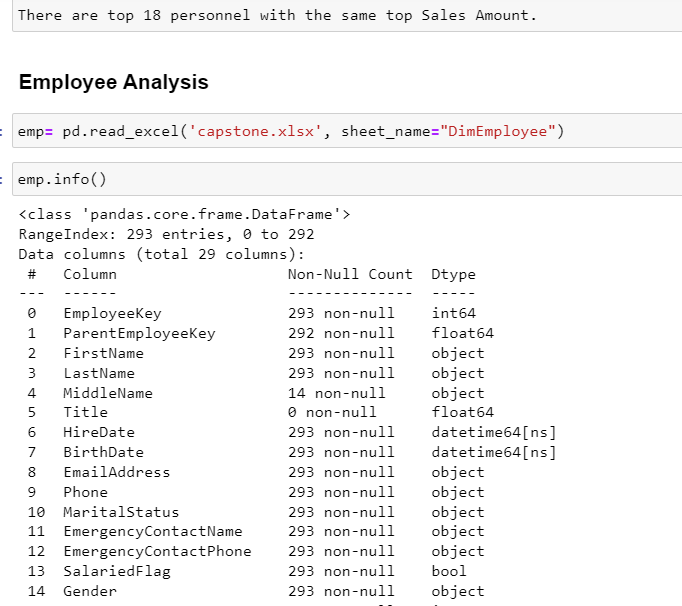




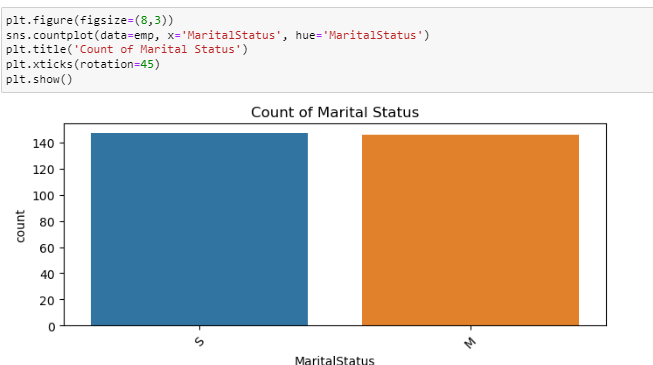


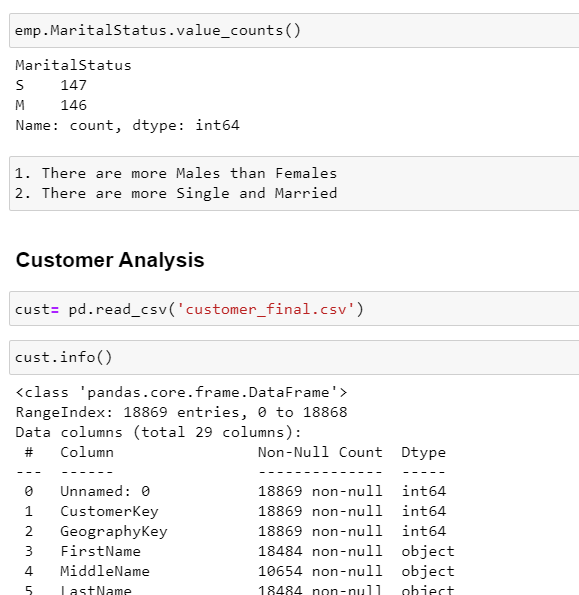


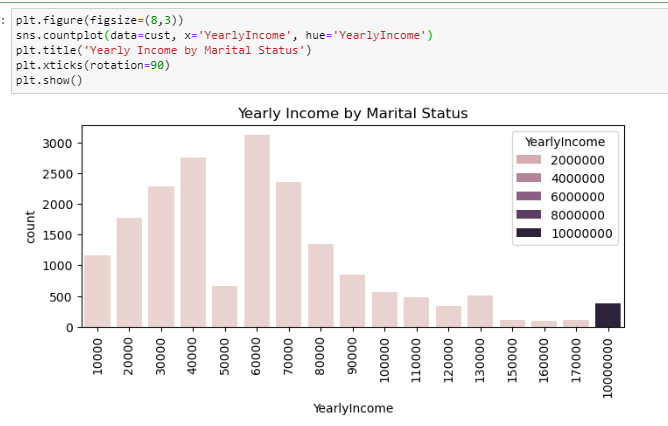


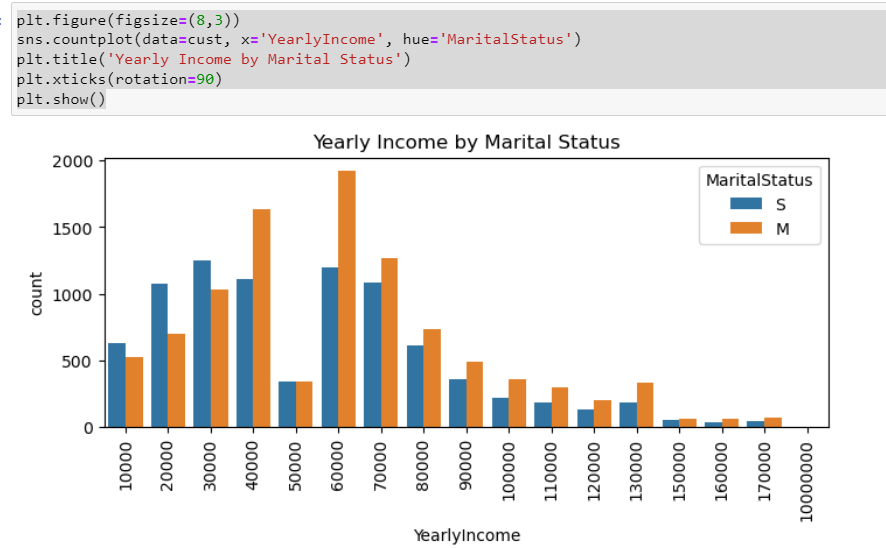


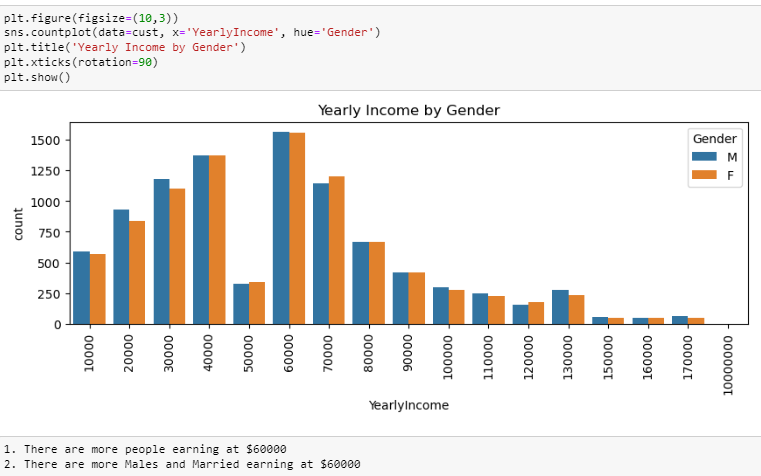






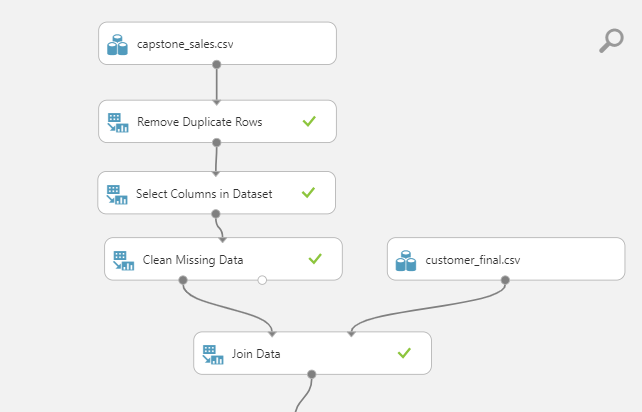


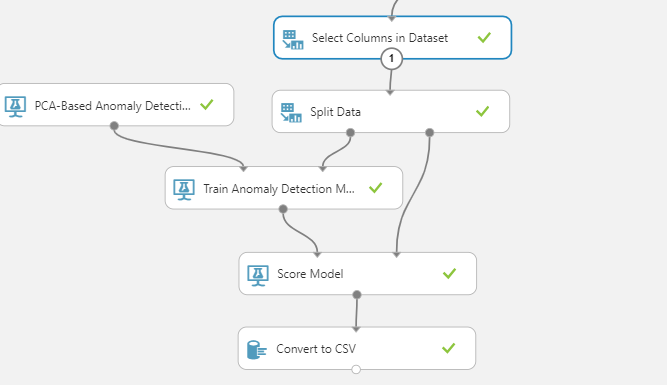


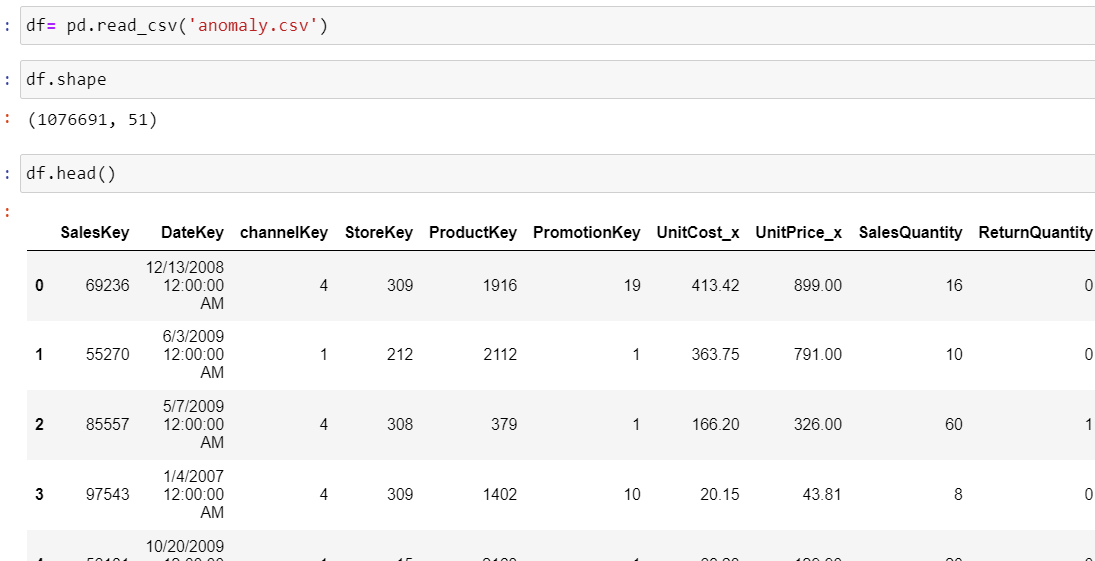


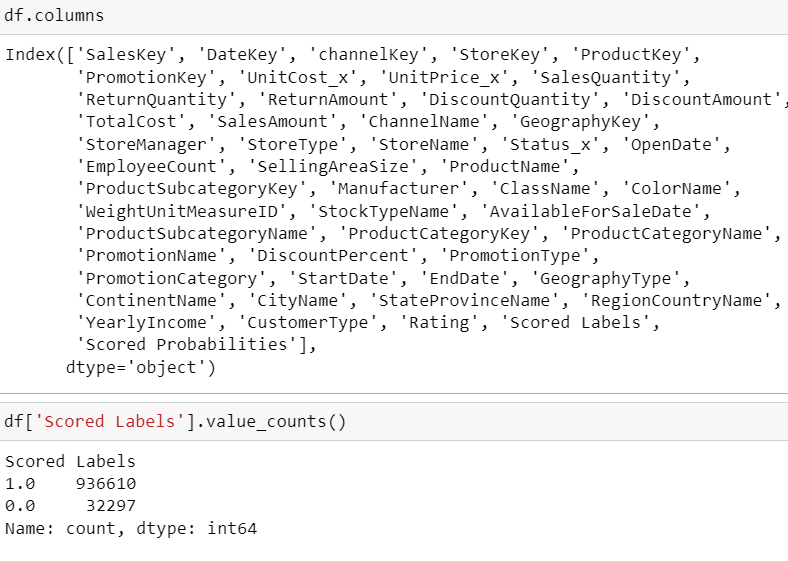
# Activity 4 : Model Selection

* Select appropriate machine learning algorithms for fraud detection, including classification and anomaly detection.
* Implement and fine-tune the chosen models.

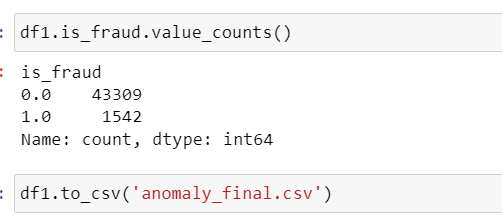






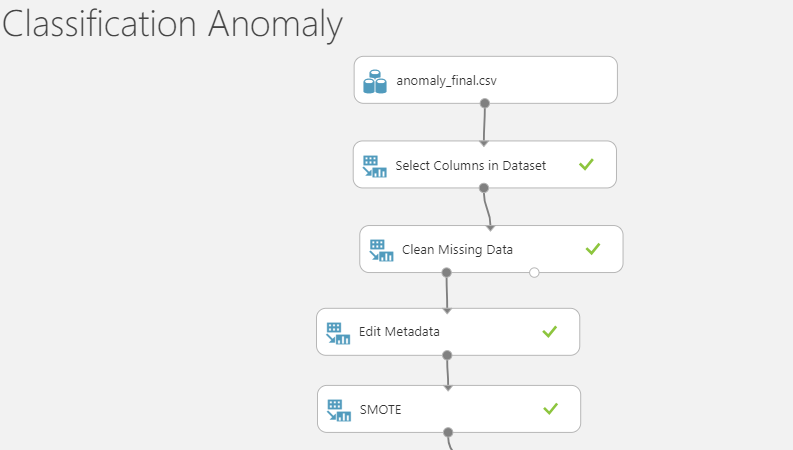


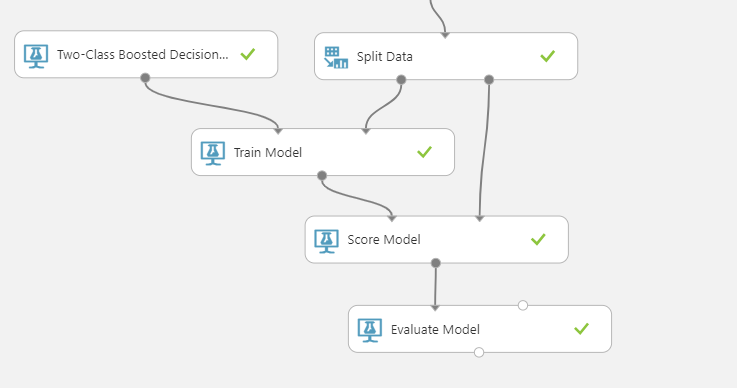


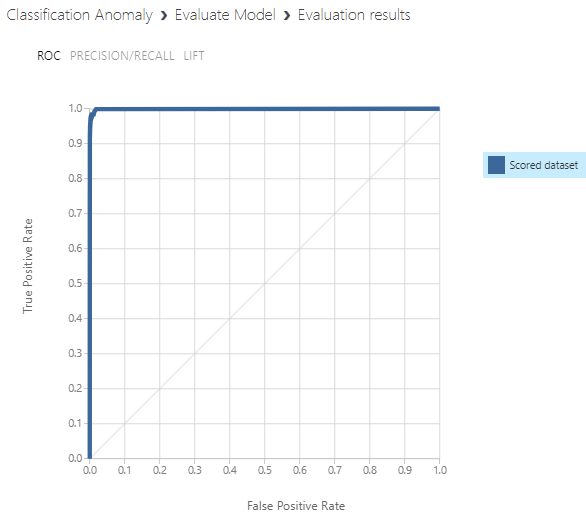


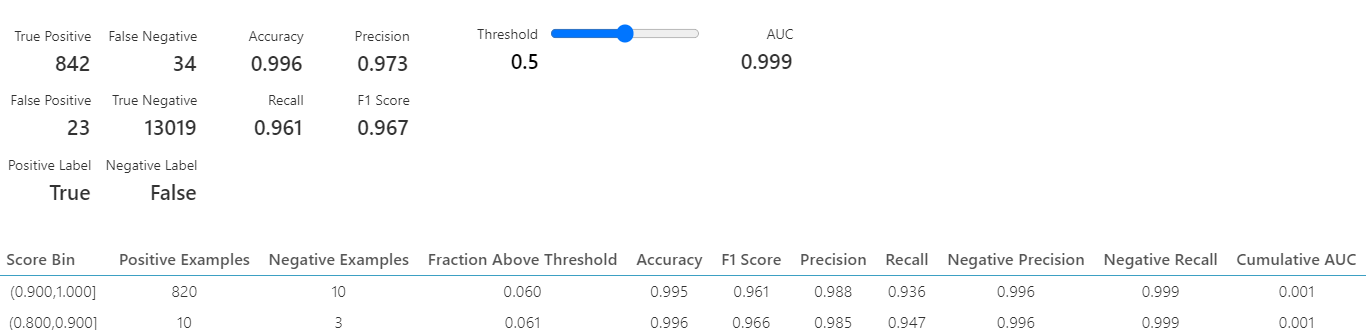
# Activity 5 : Model Evaluation

* Develop robust evaluation metrics to assess model performance, including accuracy, precision, recall, F1-score, and ROC AUC.
* Ensure the model provides actionable insights for real-time fraud detection.









**Precision is the ratio of correctly predicted positive observations to the total predicted positives so having a score of 0.973 means it had 97.3% accuracy of positive prediction.**

**Recall is the ratio of correctly predicted positive observations to the all observations in actual class so having a score of 0.961 mean it had 96.1% to capture all the relevant cases.**

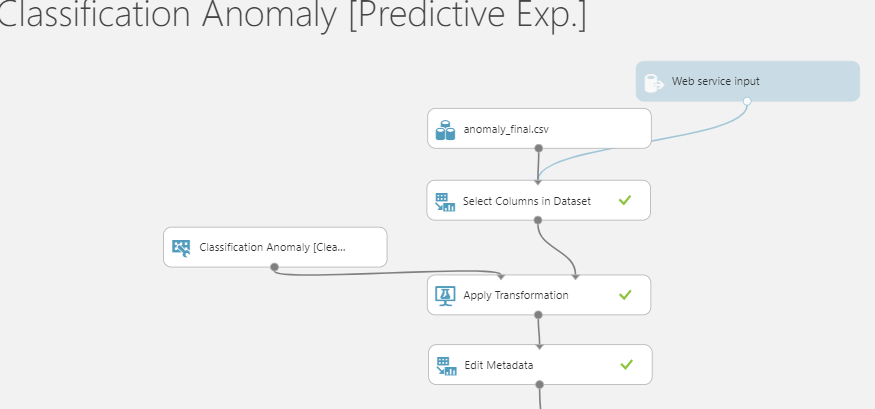
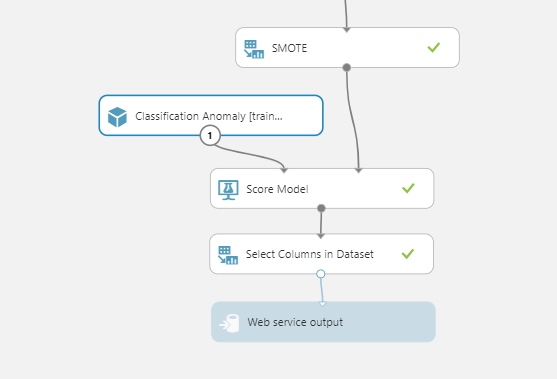
**Accuracy is the overall measure of how well the model is performing across all classes so for result of 0.996 mean that the model got 99.6% prediction right.**

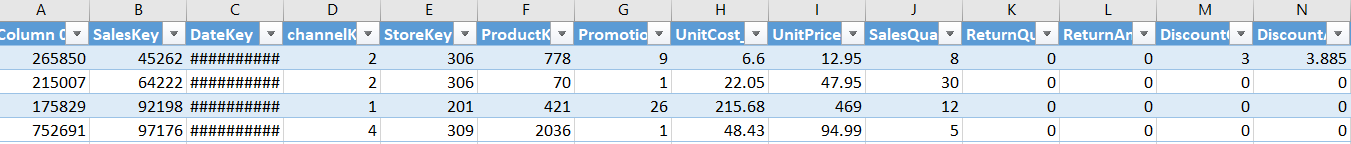
**F1 score evaluate the performance of a classification model so for a result of 0.967 means it provide 96.7% performance and strong balance between precision and recall.**

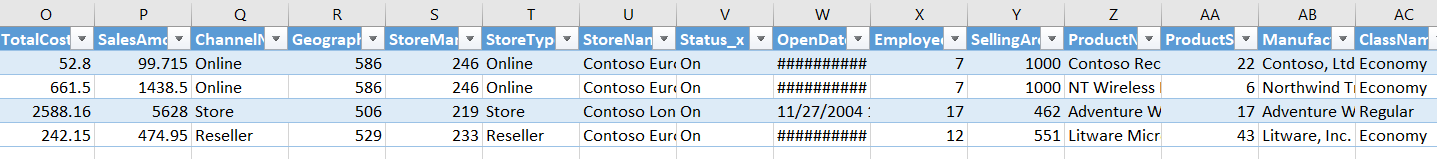
**AUC is a graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) across different thresholds a result of 0.999 mean it had 99.9% ability to discriminate between positive and negative instances.**

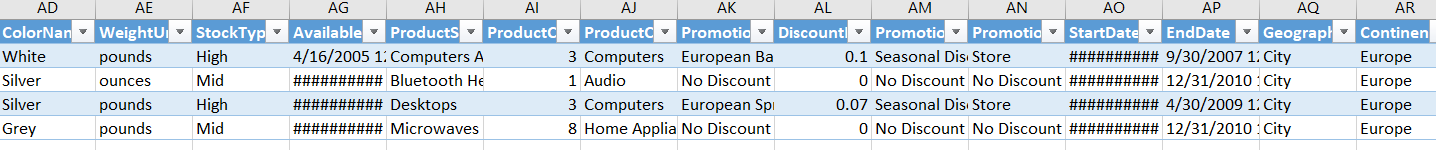
Activity 6 : Model Deployment

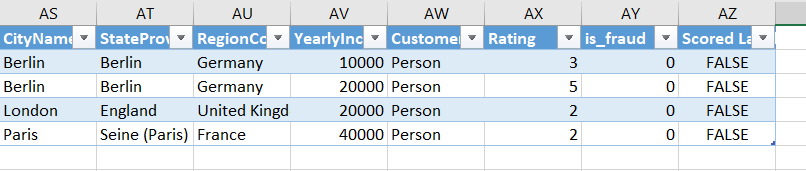
* Create a deployment strategy using Azure Machine Learning to integrate the fraud detection model into the client's transaction processing system.
* Implement real-time scoring or batch processing.

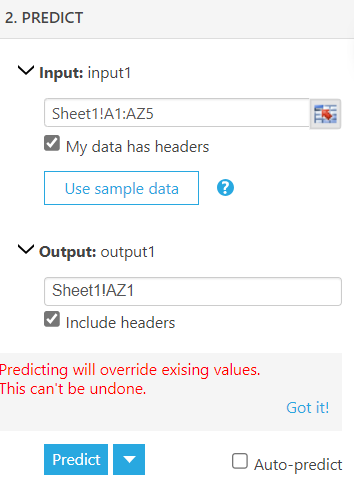












# Activity 7 : Interpretability

# Ensure the model's interpretability so that fraud analysts can understand why a particular transaction is flagged as suspicious.

# 

# 

# The above 5 features are most important contributing to the fraud detection.

# Develop visualization and explanation methods for model predictions.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# Activity 8 : Documentation and Presentation

* Document the entire project, including data sources, preprocessing steps, model selection, and deployment details.
* Prepare a comprehensive project report summarizing the project's objectives, methods, findings, and recommendations.

We are using Jupyter and Azure Machine Learning to develop a robust system for real-time identification of suspicious financial transactions.

Milestone 1:

* We read the necessary sheets and understand the data.
* We drop the unnecessary columns.
* We rename the columns to prepare for merging of datasets.

Milestone 2:

* We merge the sales and channel using channelKey, next merge on StoreKey, next merge on ProductKey

next merge on ProductSubcategoryKey, next merge on ProductCategoryKey, next merge on

PromotionKey, next merge on Geography Key.

* From the merged dataset we are able to understand the Top 10 product on Sales transaction and Sales amount and save the dataset as capstone\_sales.csv.
* We merge the customer and geo using GeographyKey, and save as customer.final.csv.

Milestone 3

* We read the capstone\_sales.csv file and visualize the Sales amount and Sales quantity understanding the outliers and the maximum sales transaction occurs between 0 to $2000 where sales quantity is

between 10 to 20

* We do Channel analysis using stripplot, lineplot, barplot, countplot, groupby and understand that

Store has generated the highest revenue and sales transactions, Catalog has generated the

highest average transactions for the last 3 years, we can see outliers for all channels beyond $60000, Maximum outliers have been observed for Online and Catalog channel.

* We do Market analysis using groupby, pie chart, barplot and understand that Maximum Revenue is from North America(58%), Top 3 countries by total revenue are US, China and Germany, Top 3 countries by average transaction value are China, Germany and France.
* We do Product analysis using barplot, groupby and understand that the Top 3 Category items are home appliance, computer, and camera and camcorder, Top 3 SubCategory items are Camcorder, Projectors

and Screens, and Washer and Dryers.

* We do Promotion analysis using groupby, barplot and understand that theTop 3 Promotion are North America Back to School Promotion, North America Holiday Promotion, North America Spring Promotion,

Top 3 Promotion by average transaction is Asian Holiday Promotion, North America Spring Promotion,

Asian Summer Holiday Promotion.

* We do corr, heatmap, pairplot on the numeric columns to visualize the relationship.
* We do Color Name analysis using countplot and understand that Top 3 Selling colors are Black, White

and Silver

* We merge DimSales and DimSalesPerson, drop the unnecessary columns, using groupby, barplot and

understand that the top 3 personnel are Tyrone Blanco, Michael Martinez and Jaime Diaz.

* We do Employee analysis using countplot and understand that there are more Males than Females,

there are more Single and Married

* We do Customer analysis using countplot and understand that there are more people earning at $60000,

there are more Males and Married earning at $60000

Milestone 4

* We use Azure Machine Learning to read capstone\_sales.csv, remove duplicated rows, select needed columns, clean missing data, join data with customer\_final.csv.
* Split the data into 70% train data and 30% test data.
* and using PCA-Based Anomaly detection and train the model
* Scored the model and Scored labels columns is created and convert to anomaly.csv.

Milestone 5

* We read the anomaly.csv convert the Scored Labels from 1 to 0, 0 to 1.
* Create a sample size of 50000 and save file as anomaly\_final.csv
* Using Azure, we read the anomaly.csv file, select the needed columns, cleaning missing data,

Edit Metadata, SMOTE to boost the minority data, split data in 70% train data and

30% test data.

* Train the model using Two-class Decision Tree and train the model, score the model.
* Evaluate the model produce F1 score, Precision, Accuracy, Recall and AUC.

Milestone 6

* We create the Predictive Experiment and select only Scored Labels and deploy the web service.
* Using Excel we use the Azure Machine Learning, use the sample data to input the data and predict the scored labels.

Milestone 7

* Using Permutation Feature Importance we are able to show the 5 features that are most important contributing to the fraud detection.
* We read the anomaly.csv using countplot, stripplot, barplot, lineplot, groupby, we understand that the most fraud cases occur with Below yearly income of 40000, Reseller, Contoso, White color, Regular

Class, , European Holiday Promotion, Lamps, Home appliances, Paris, France, more return amount, Sales amount is less than 20k, more discount quantity, more discount amount, average sales quantity about

12, average sales amount about 3700.

# Annexure (Excel Workbook)

* + 1. Attach your dataset (one excel file or multiple csv files)

## Submission Instructions

1. Report file – DSC-0423-<Your name>.docx or DSC-0423-<Your name>.pdf
2. Dataset
3. Azure ML file - DSC-0423-<Your name>.pbix
4. Compress item 1 to 3 => DSC-0423-<Your name>.zip

## File to submit - DSC-0423-<Your name>.zip