**Customer Lifetime Value Prediction**

**Group Name:** Group Sigma

**Group Members:**

|  |  |  |
| --- | --- | --- |
| First name | Last Name | Student number |
| Amrita Kaur | Badhan | C0928455 |
| Jaldhi | Shah | C0924359 |
| Mohamed | Gaafar | C0905127 |
| Mukul | Sharma | C0926138 |

**Submission date: 12 August 2024**

Contents

[Abstract 3](#_Toc4571289)

[Introduction 3](#_Toc4571290)

[Methods 4](#_Toc4571291)

[Results 47](#_Toc4571292)

[Conclusions and Future Work 48](#_Toc4571293)

[References 48](#_Toc4571294)

**Abstract**

**In this project we aim to predict the customer live time value (CLV) for 19,119 customers of a business from 121,317 transaction records. We acquired the business data by querying the Microsoft Adventure Works sample database using SQL, Microsoft SQL Management Server and Azure Data Studio. We performed data pre-processing to handle missing values, data formatting, data integrity, and detecting outliers. Then we processed the data by trimming the outliers, features engineering, and metrics calculation. The CLV metrics we calculated are Frequency, Recency, T, and Monetary Values. To predict the CLV we performed data preparation by splitting our data to training and test datasets and recalculating the metrics and the labels for both datasets. After preparing our data we predicted the CLV with 5 methods, which are: Historical, Probabilistic BG/NBD, Machine Learning Regression using XGBoost, Liner Deep Learning Regression, and Machine Learning Segmentation and Classification. We achieved the least error with Liner Deep Learning scoring an RMSE of 5.33 model with which is very good error margin, followed by ML XGBoost Regression with an RMSE of 294.28, Benchmark Historical CLV with an RMSE of 4406.50, and the BG-NBD Probabilistic with RMSE of 4464.26.**

# **Introduction**

**When running a business, the customer is the most valuable asset. Modern businesses generate numerous amounts of data about their customers but this data isn’t useful until transformed into meaningful information that helps the decision maker to act on and drive knowledge about their business. Business information is usually called business metrics of key performance indicators (KPIs). One of the most important and valuable business metrics is customer lifetime value (CLV). The CLV indicates how useful the individual customer is to the business, it takes into account many factors to calculate the value (mostly the revenue) generated to the business by the customer.**

**CLV is very helpful and informative for the business when running a marketing campiness, knowing the most valuable customers and working on keeping them alive for the business is a key success step. When knowing the CLV a business can utilize it to make customers person for the most valuable customers and focus on acquiring this segment of customers. On the other hand, the business can manage to classify the less valuable customers to avoid spending less money on acquiring them. Also, the CLV can be very informative when pivoting it with different business elements such as business offers, it can show the effective offers from the less effective ones.**

**Predicting the CLV makes it more robust for the business to manage its customer’s base and enhance its processes. There are many different approaches to calculate and predict the CLV, in this project we encountered 5 methods which are: Historical, Probabilistic BG/NBD, Machine Learning Regression using XGBoost, Liner Deep Learning Regression, and Machine Learning Segmentation and Classification. For the business to be able to utilize these methods it needs to perform data pre-processing and data preparation to make the data ready for modeling.**

**In the project we queried the data from a business database, pre-processed it, understand it throw EDA, processed and prepared it, then we modeled and evaluated it.**

**Methods**

**The following steps show the workflow to get the CLV prediction from the business row data.**

**Step 1: Data Accusation:**

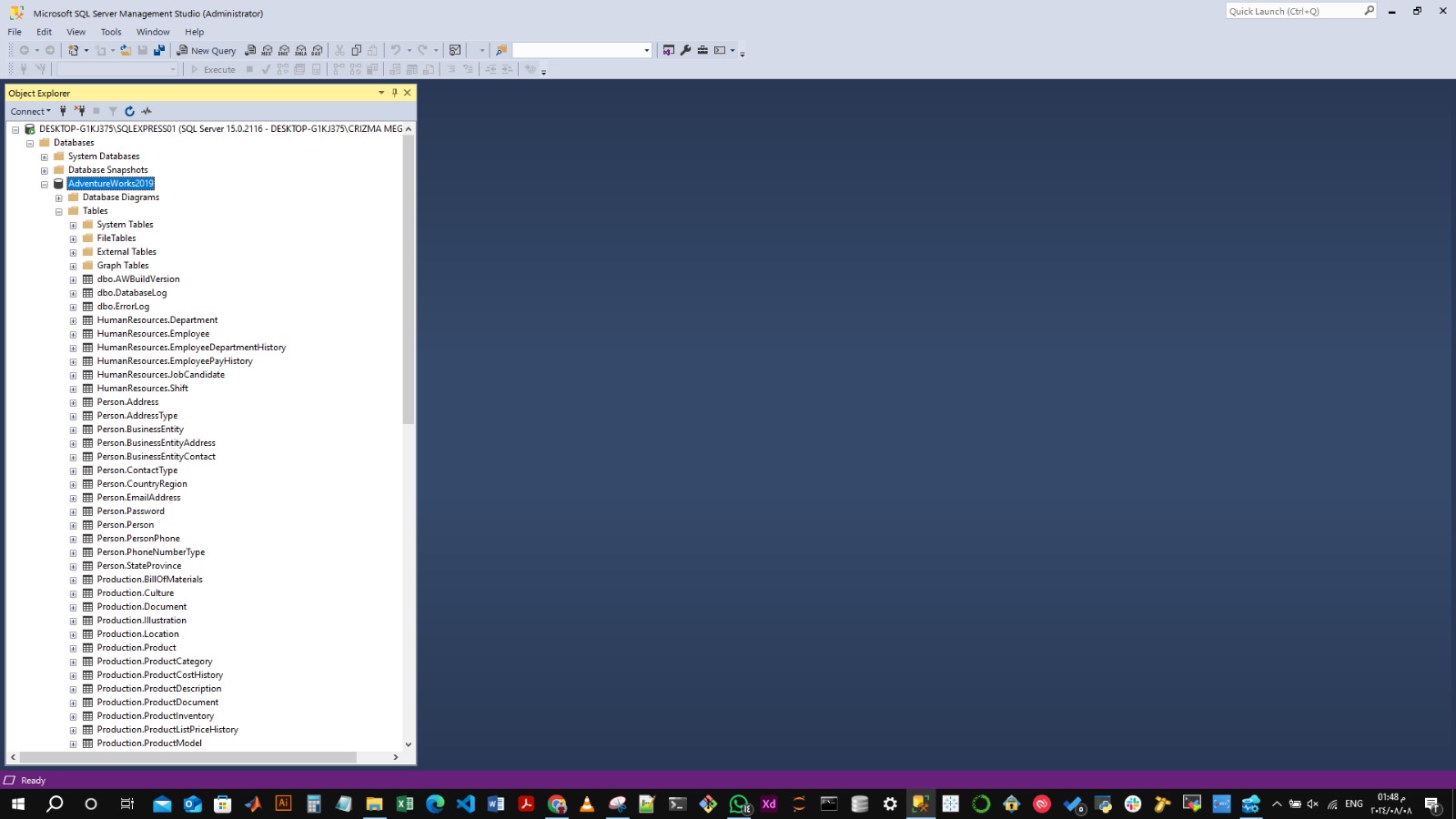
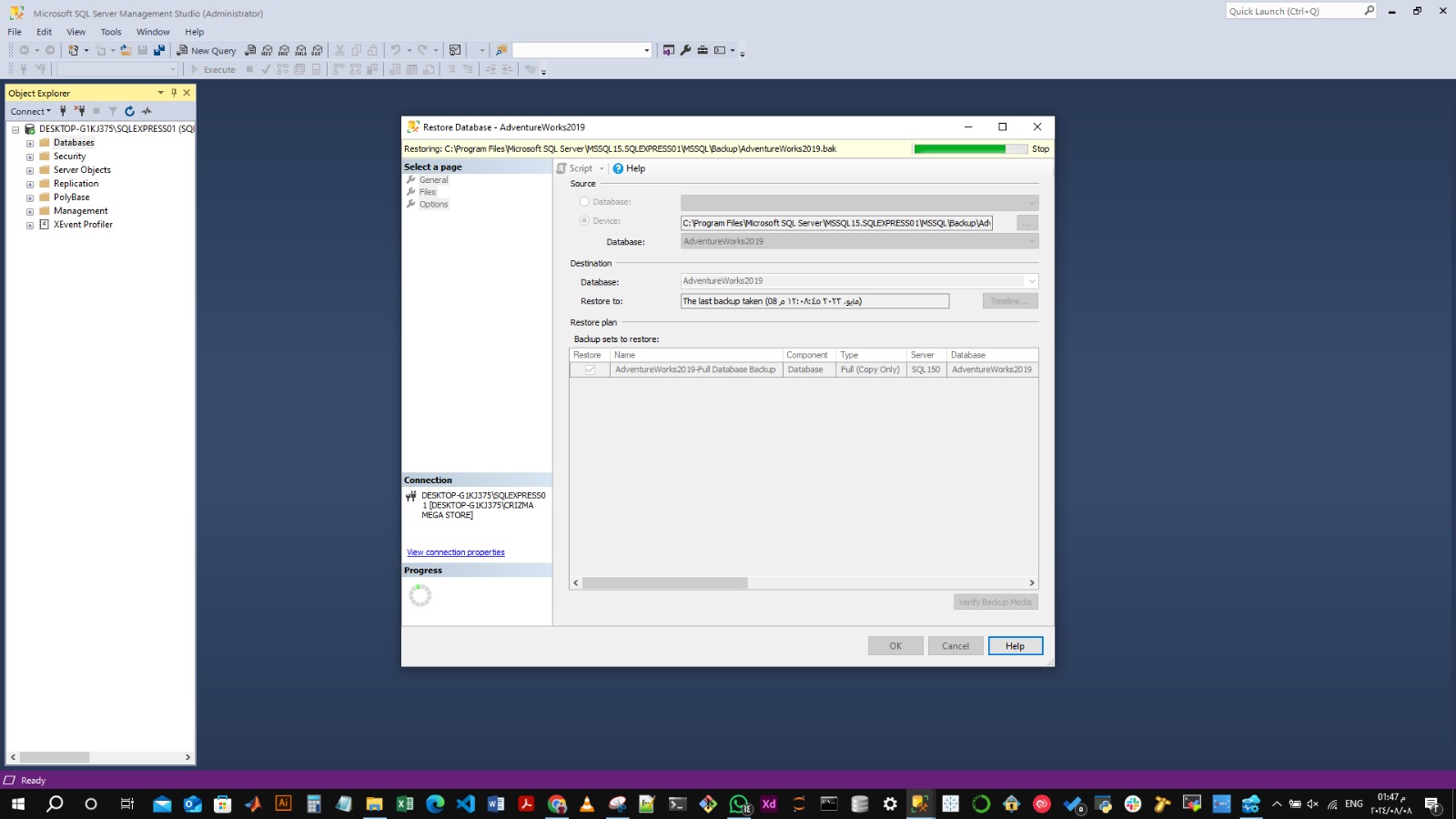
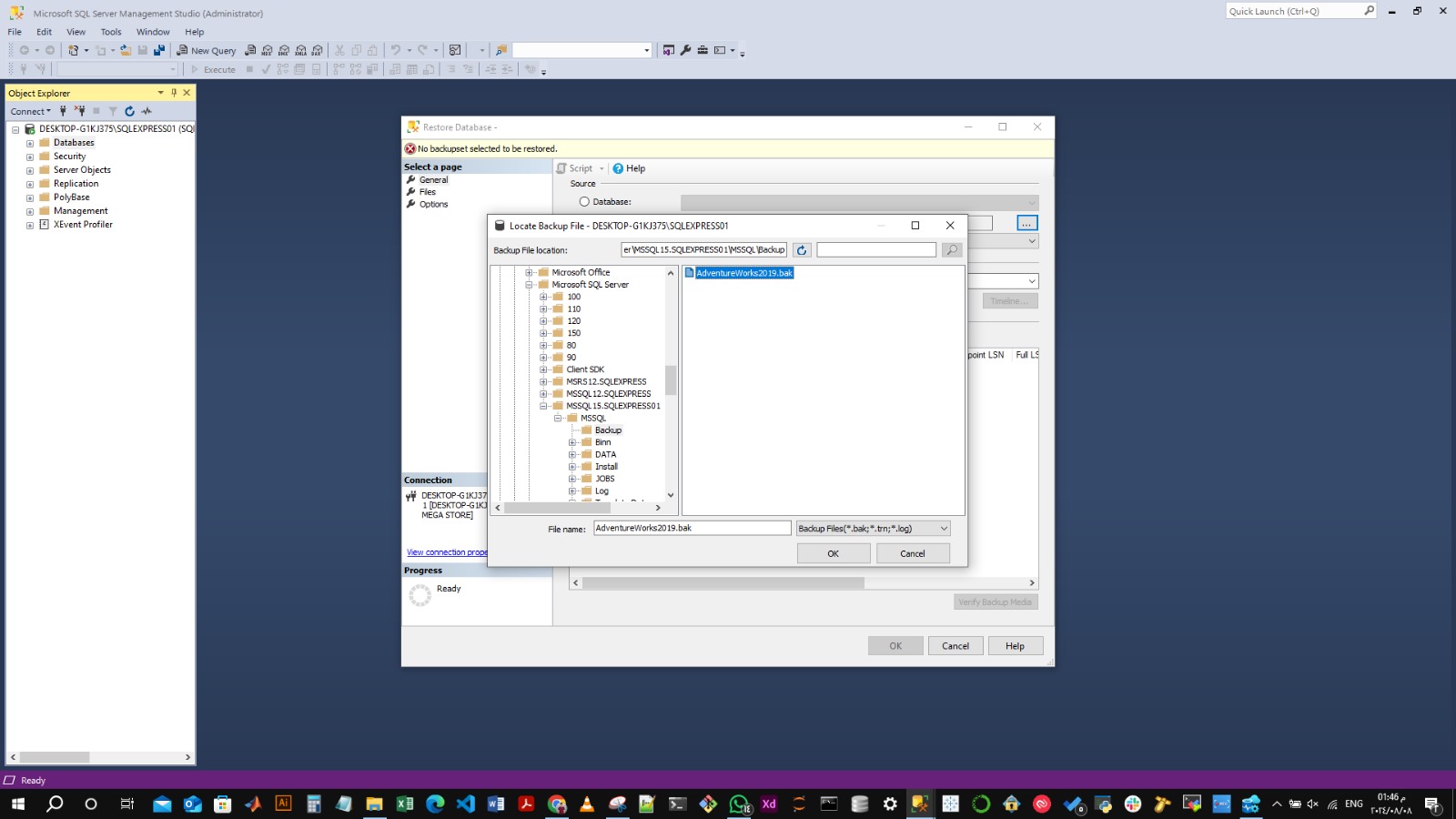
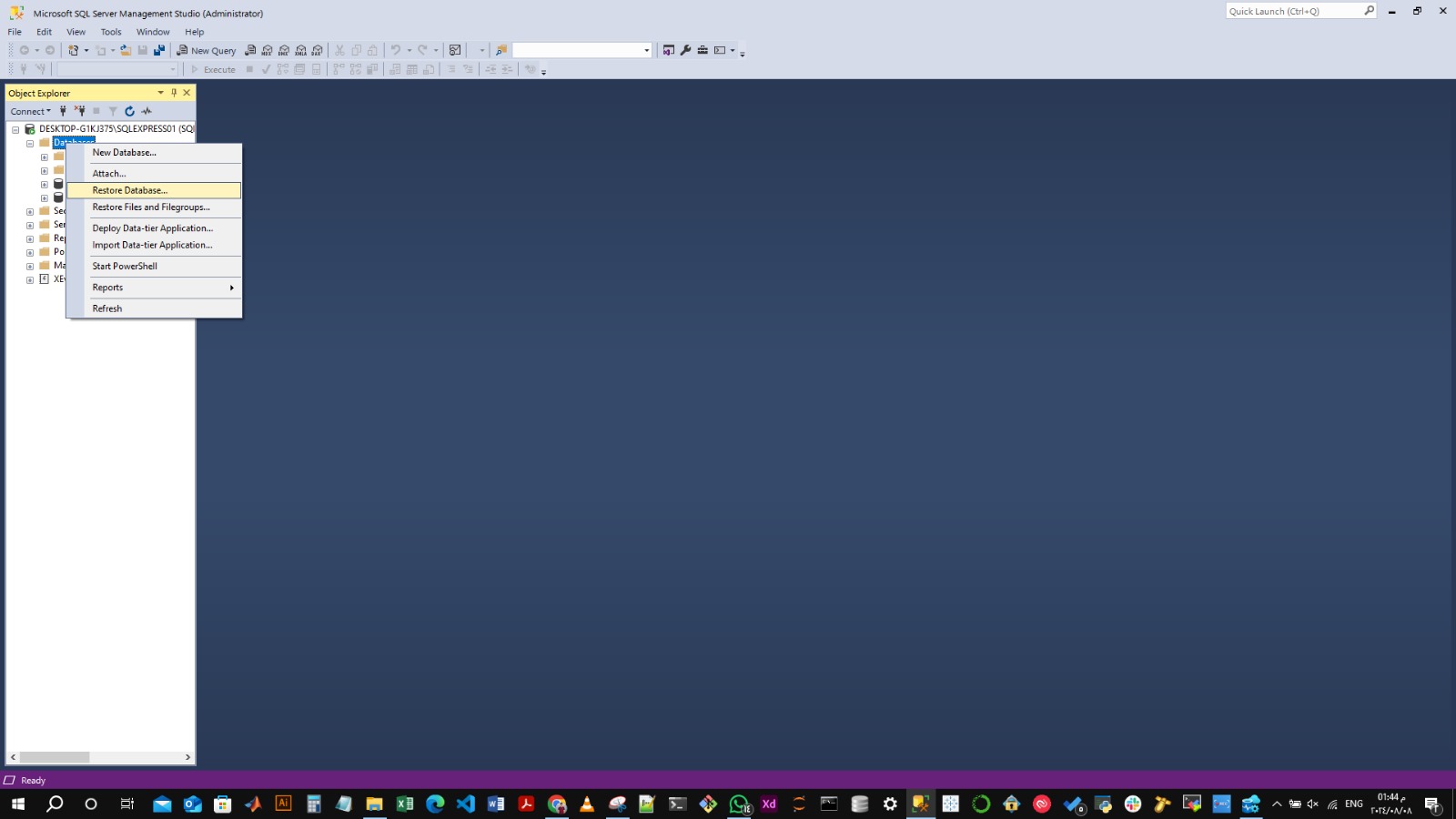
We acquired our data from Microsoft AdventureWorks sample databases. These are sample databases simulating real-world applications. The databases simulate real world scenarios such as Online Transaction Processing (OLTP), and Data Warehousing (DW).

We followed the following steps to acquire and extract the data we are interested in.

**1.1. We downloaded the database backup file from the Microsoft website.**

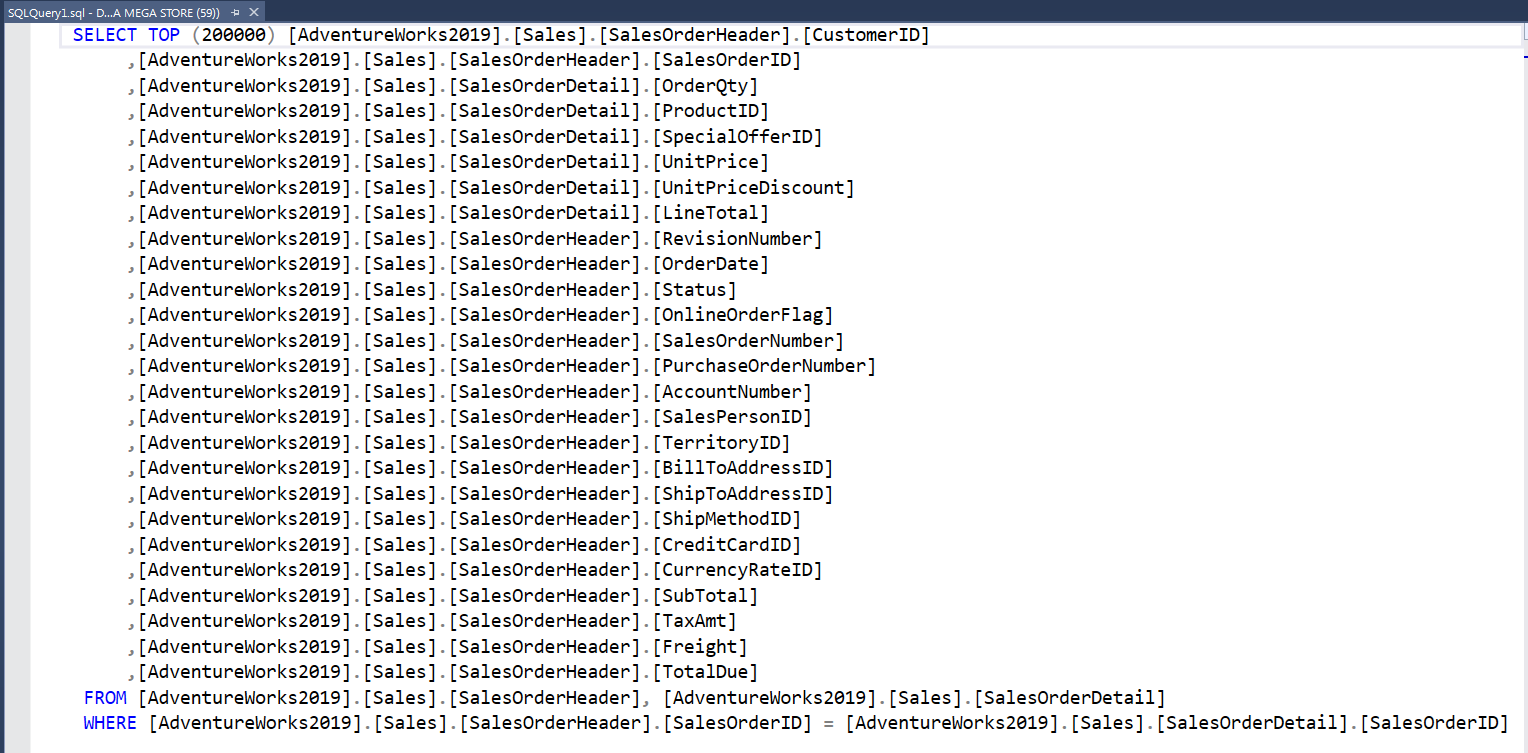
<https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms>

**1.2. Restoring the 'AdventureWorks2019.bak' file using SQL Server Management Studio.**

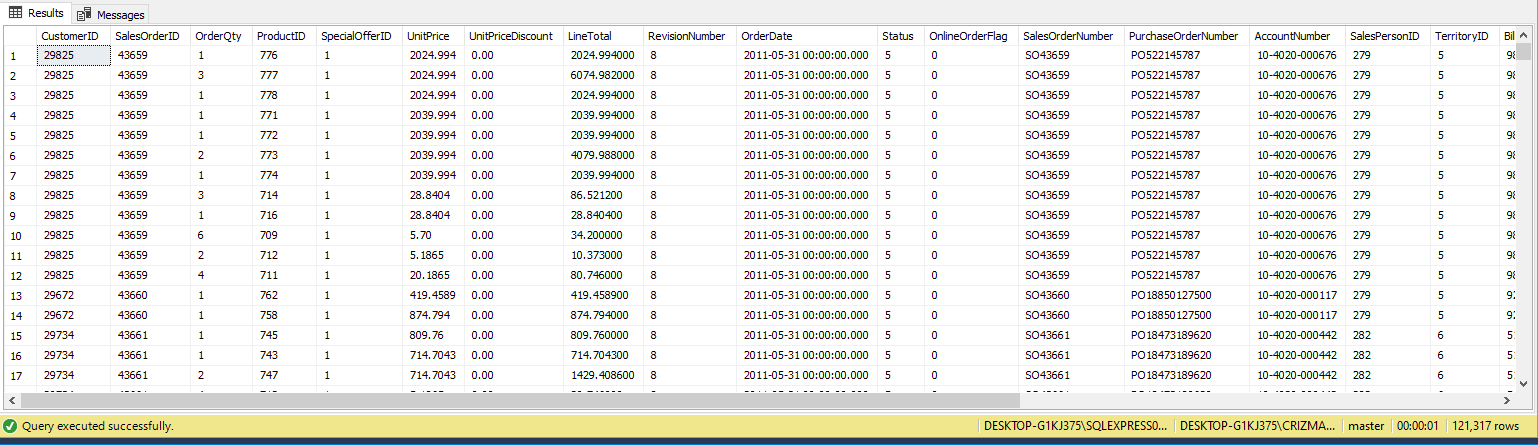


**1.3. Query the data using SQL Server Management or Azure Data Studio:**

After restoring the database we extracted the customers’ transactions data by running an SQL query against the database.



Executing the SQL query extracted 121,317 rows.

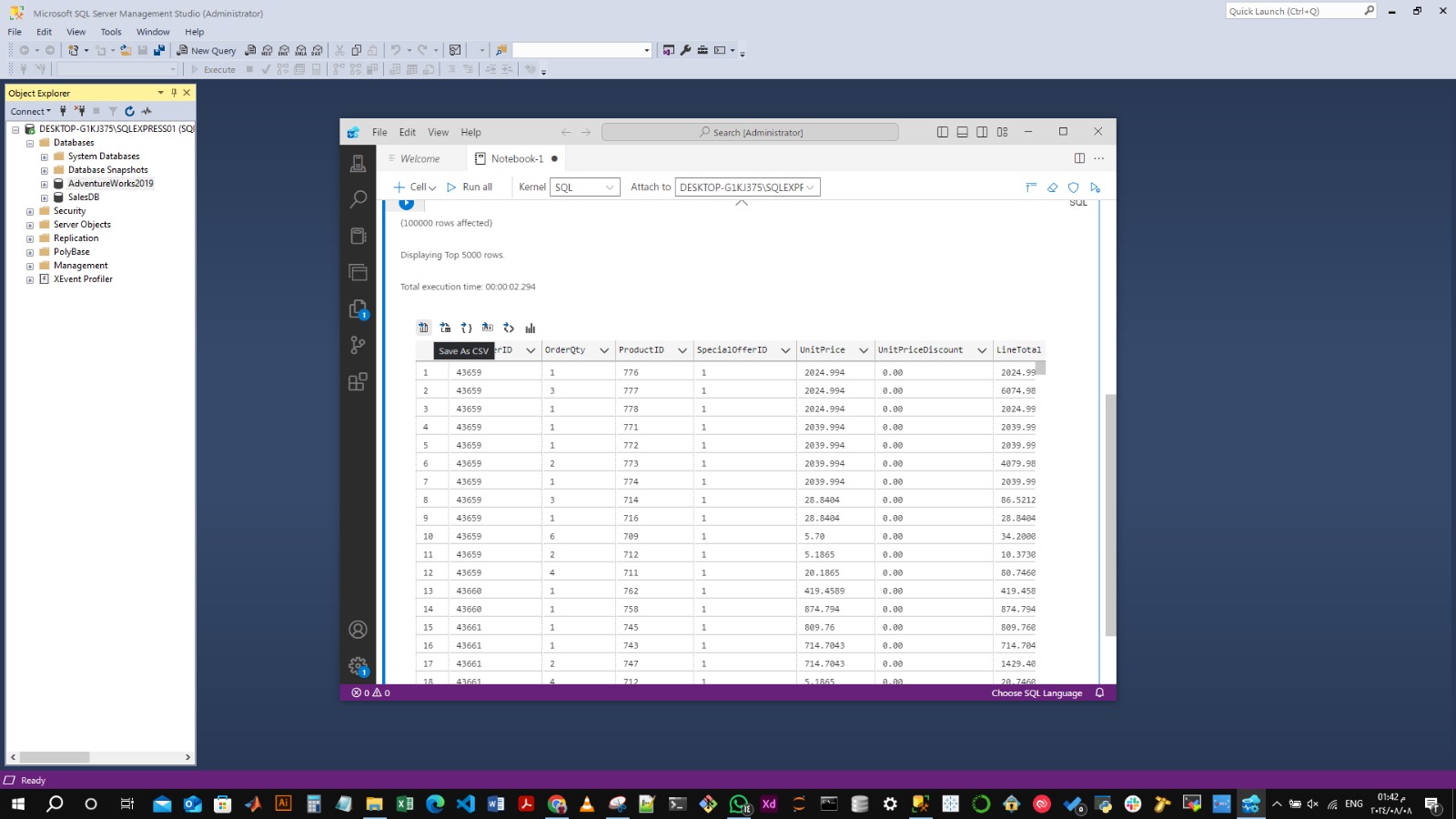
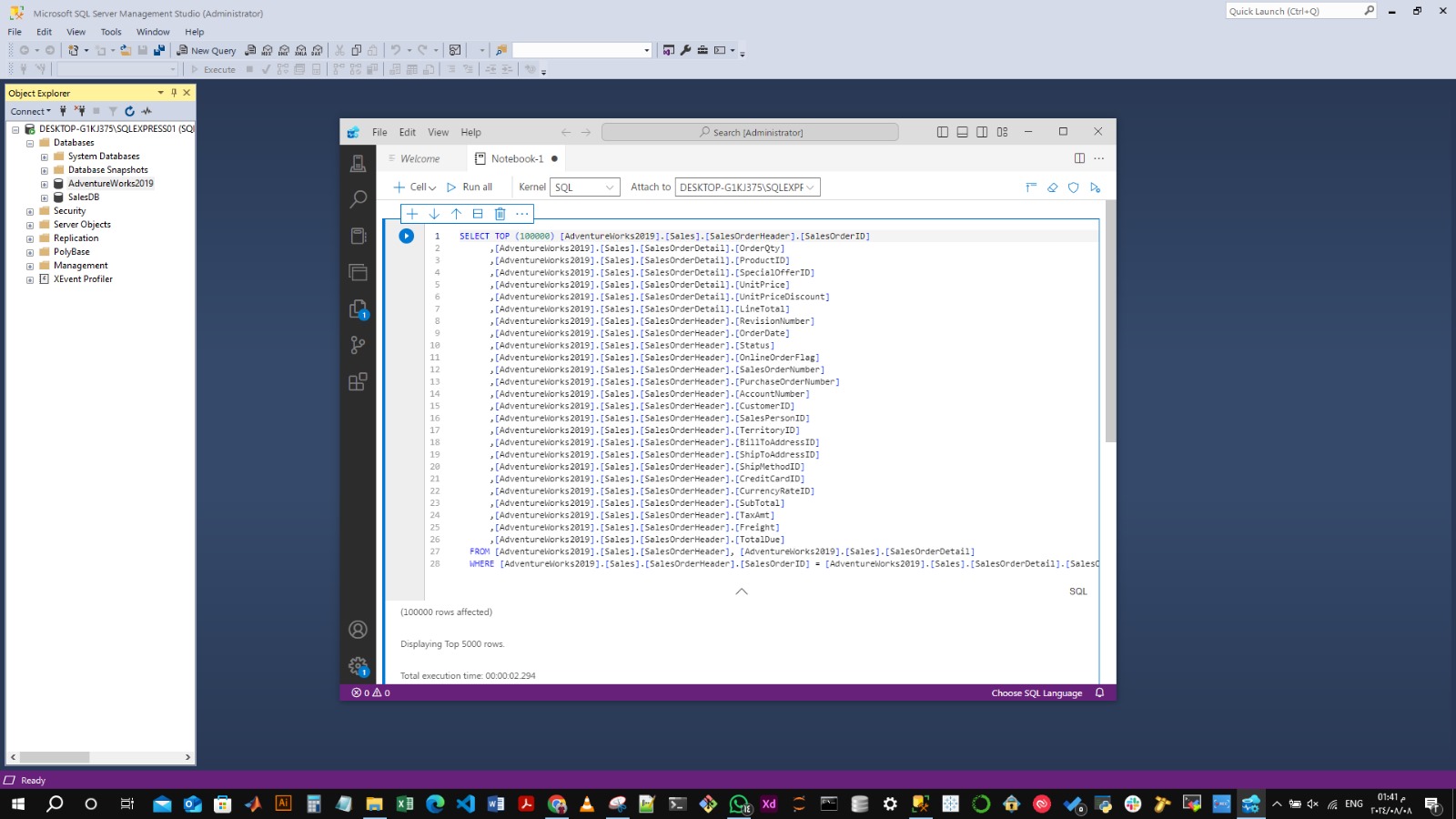
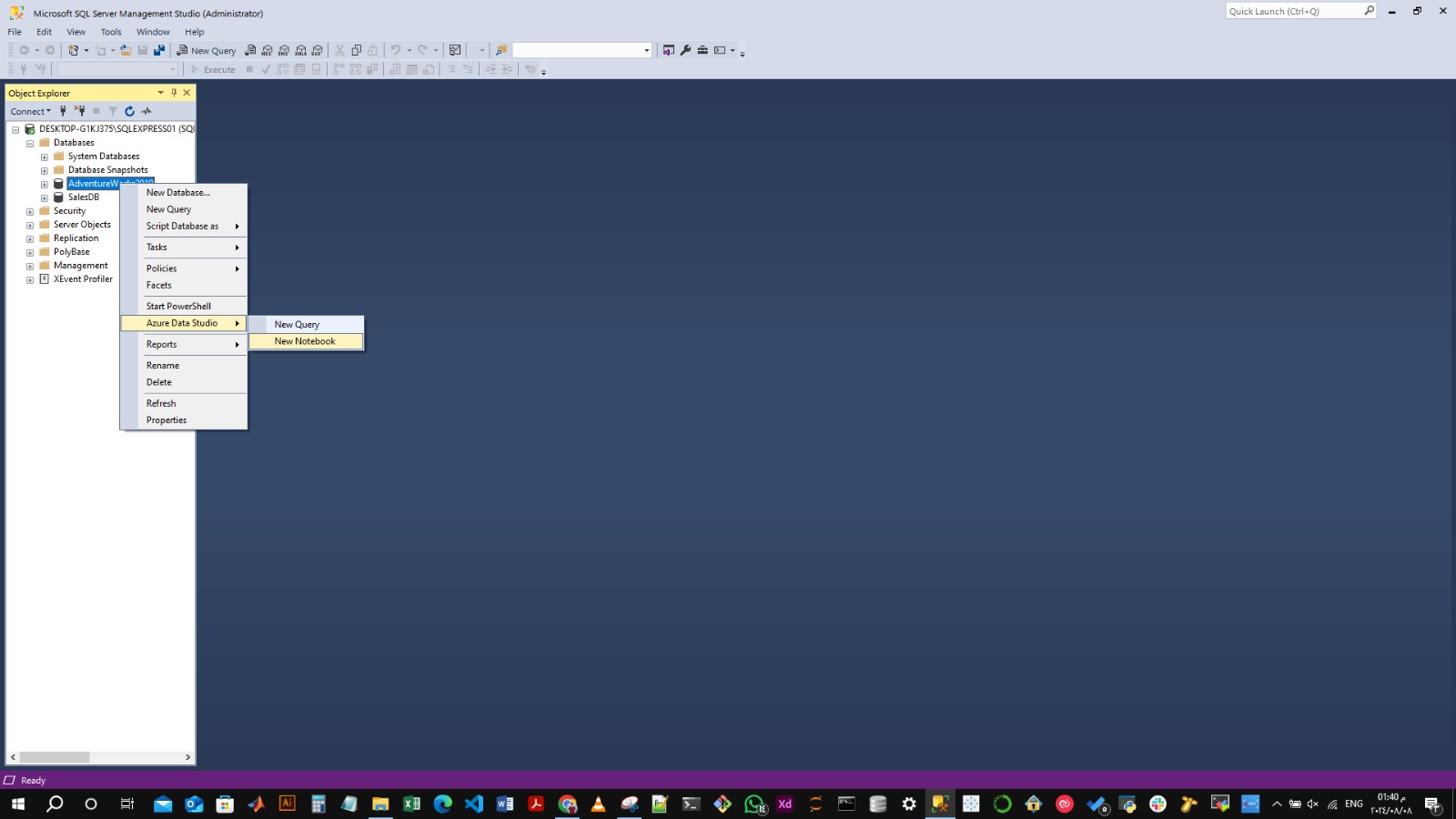


**1.4. Exporting the data to a .csv file:**

To work with the data and manipulate it in Python we needed to transform it to .CSV (or .XLSX) file.

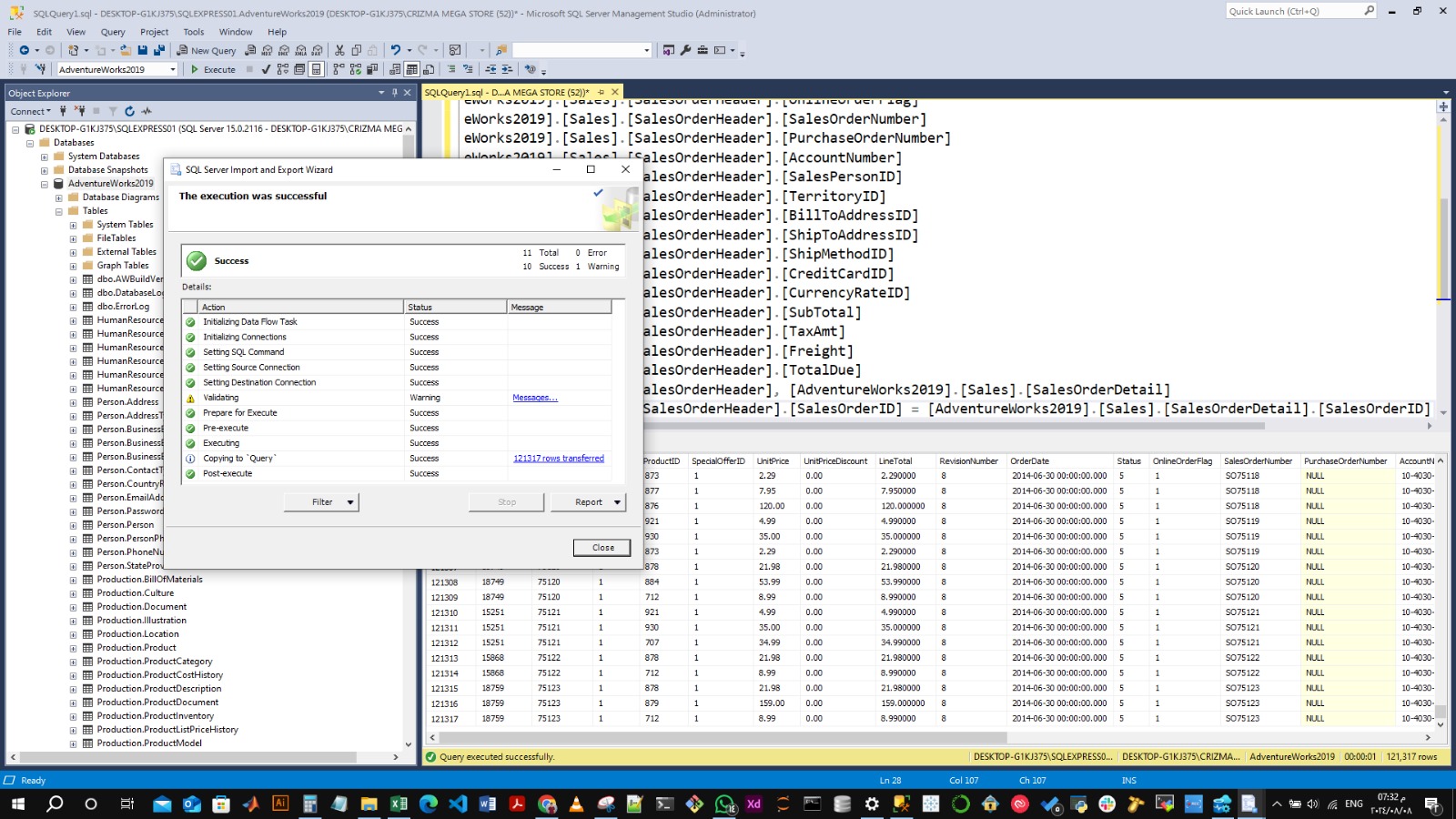
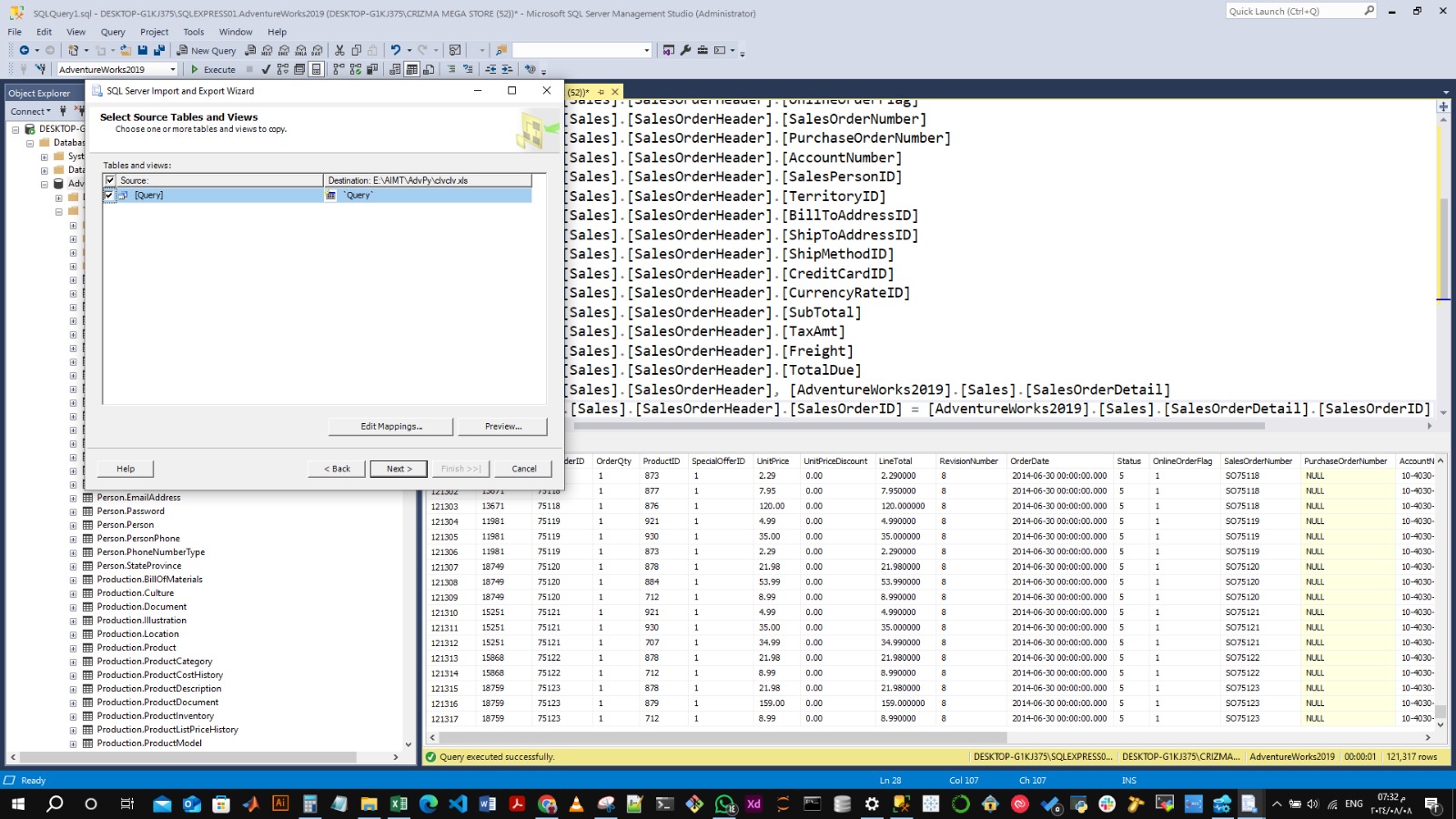
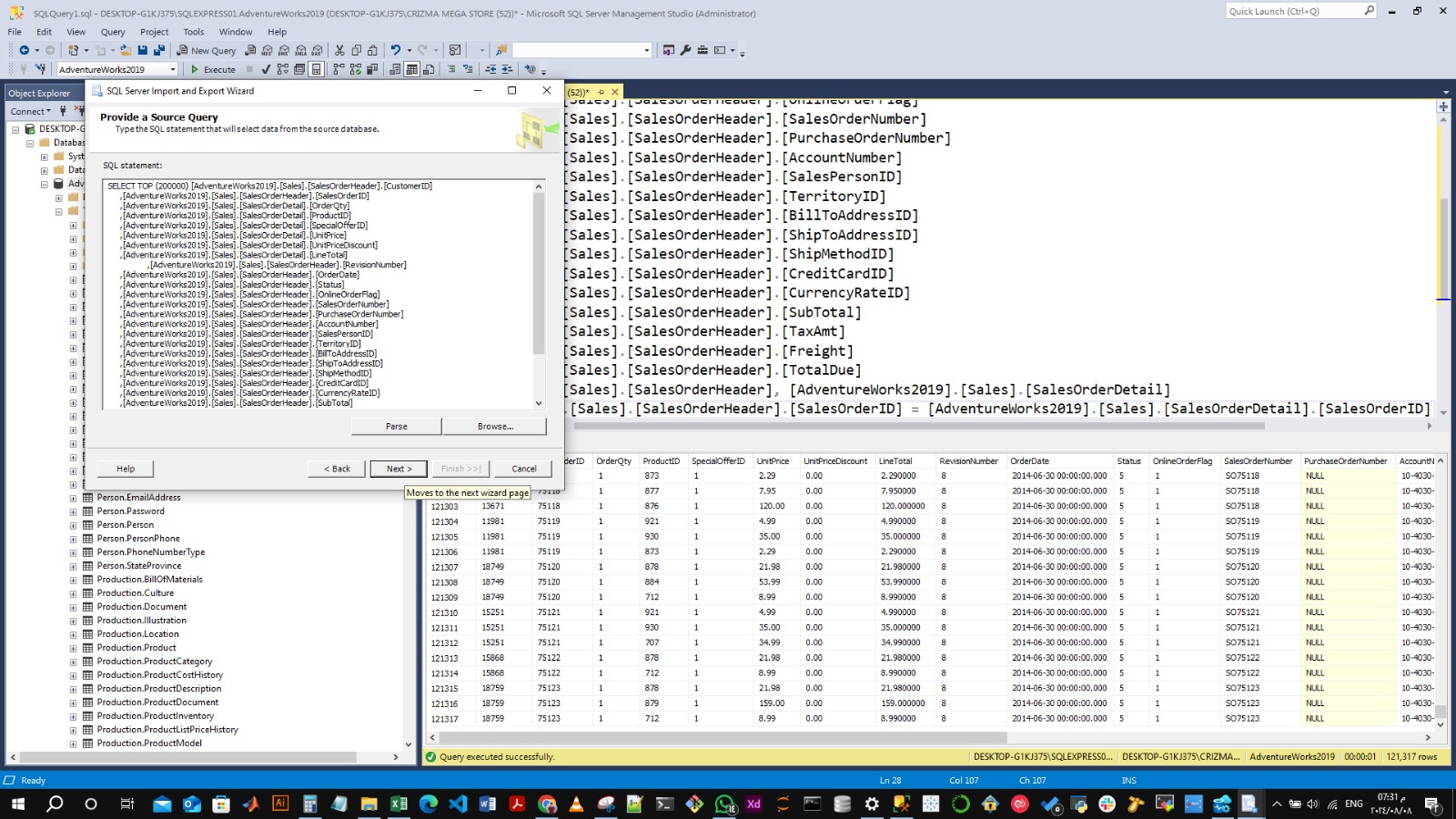
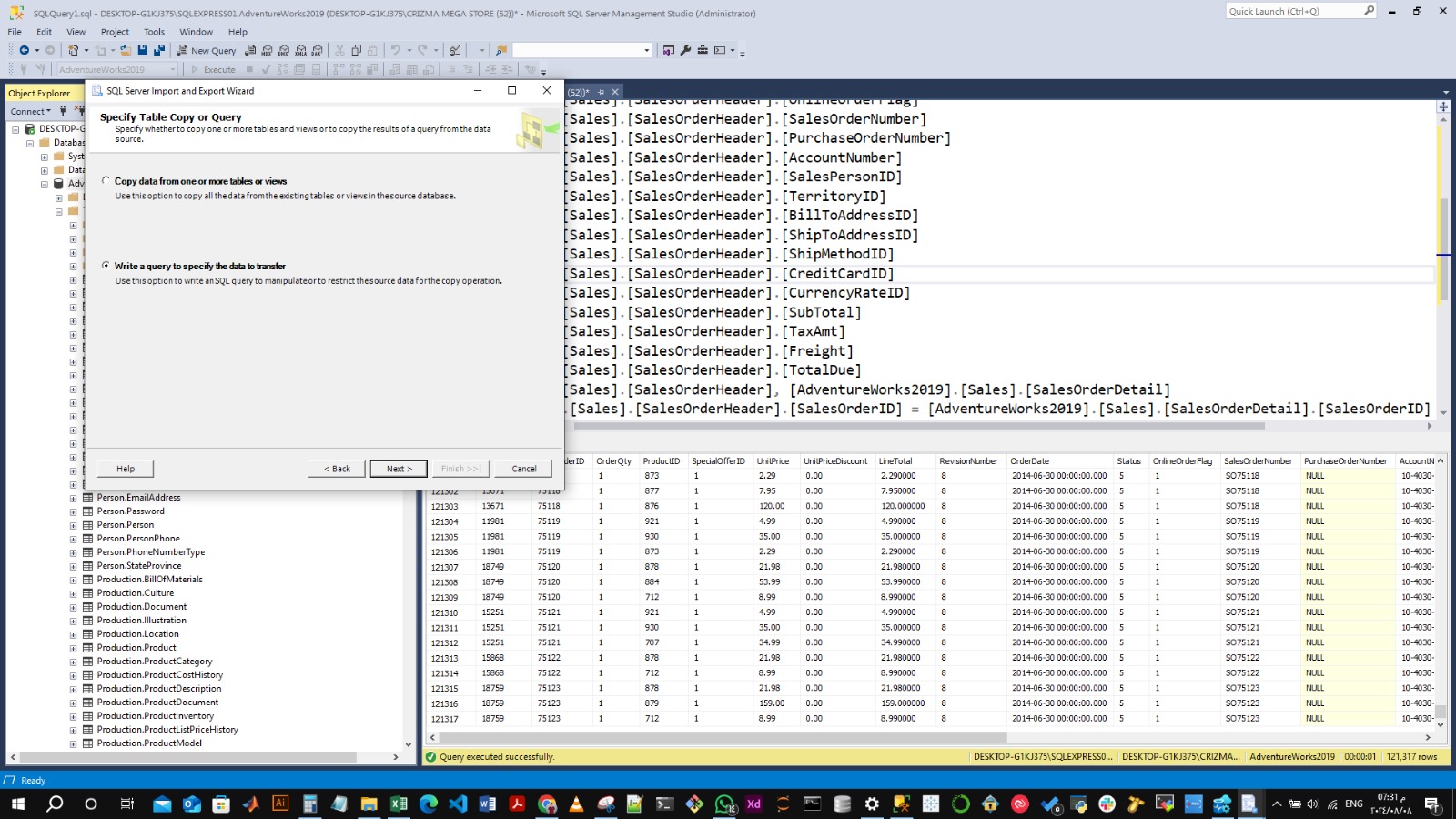
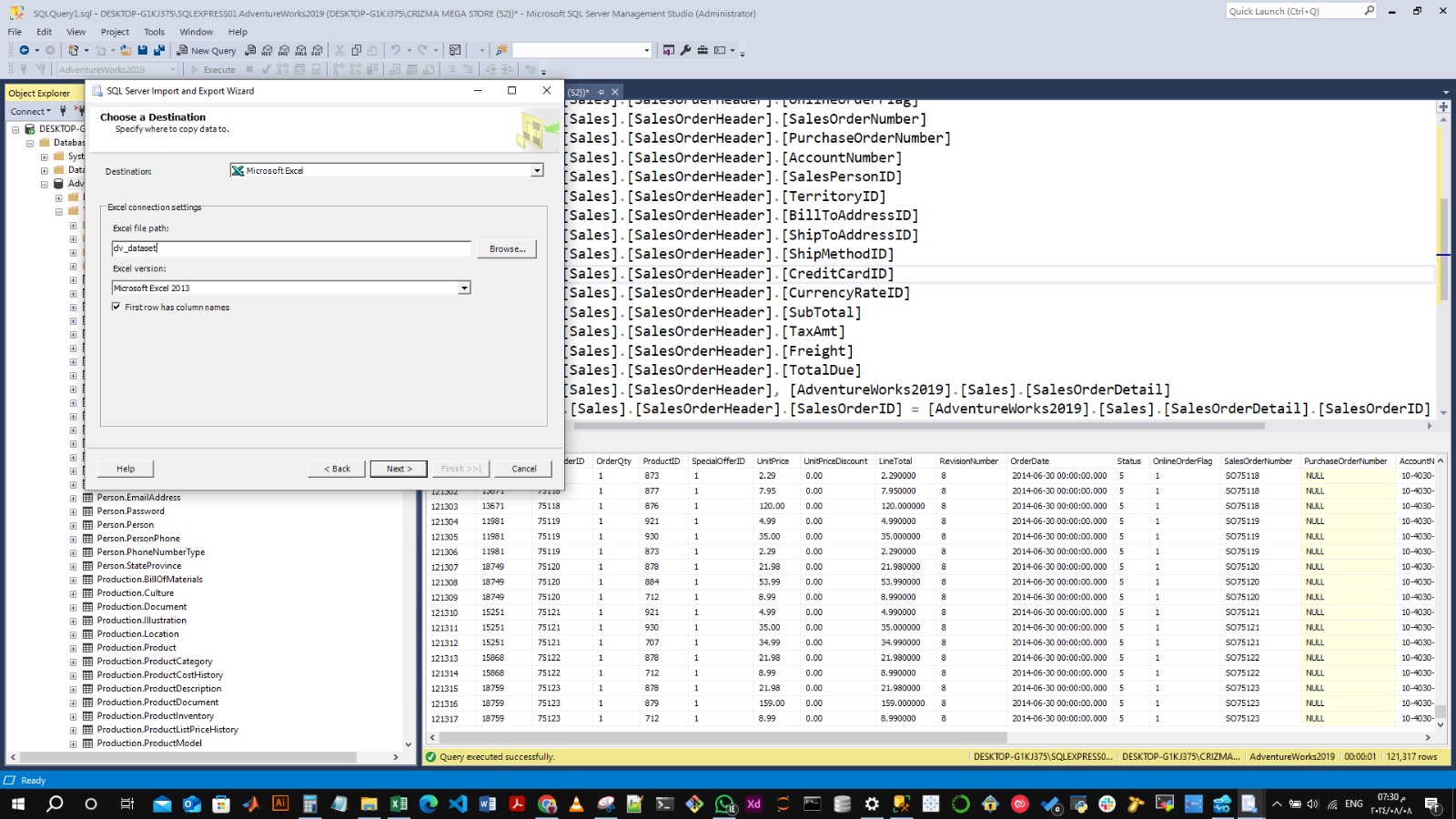
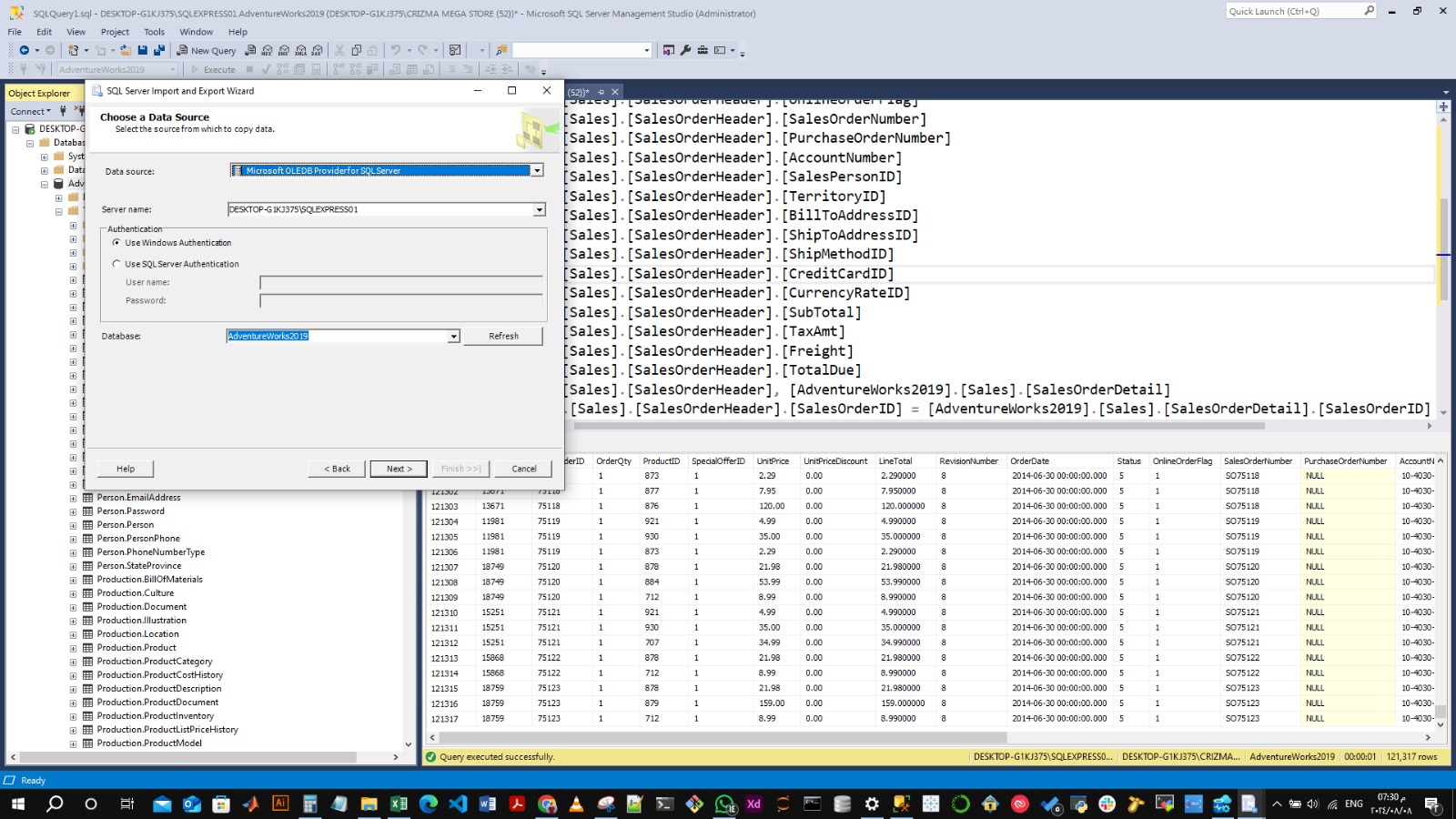
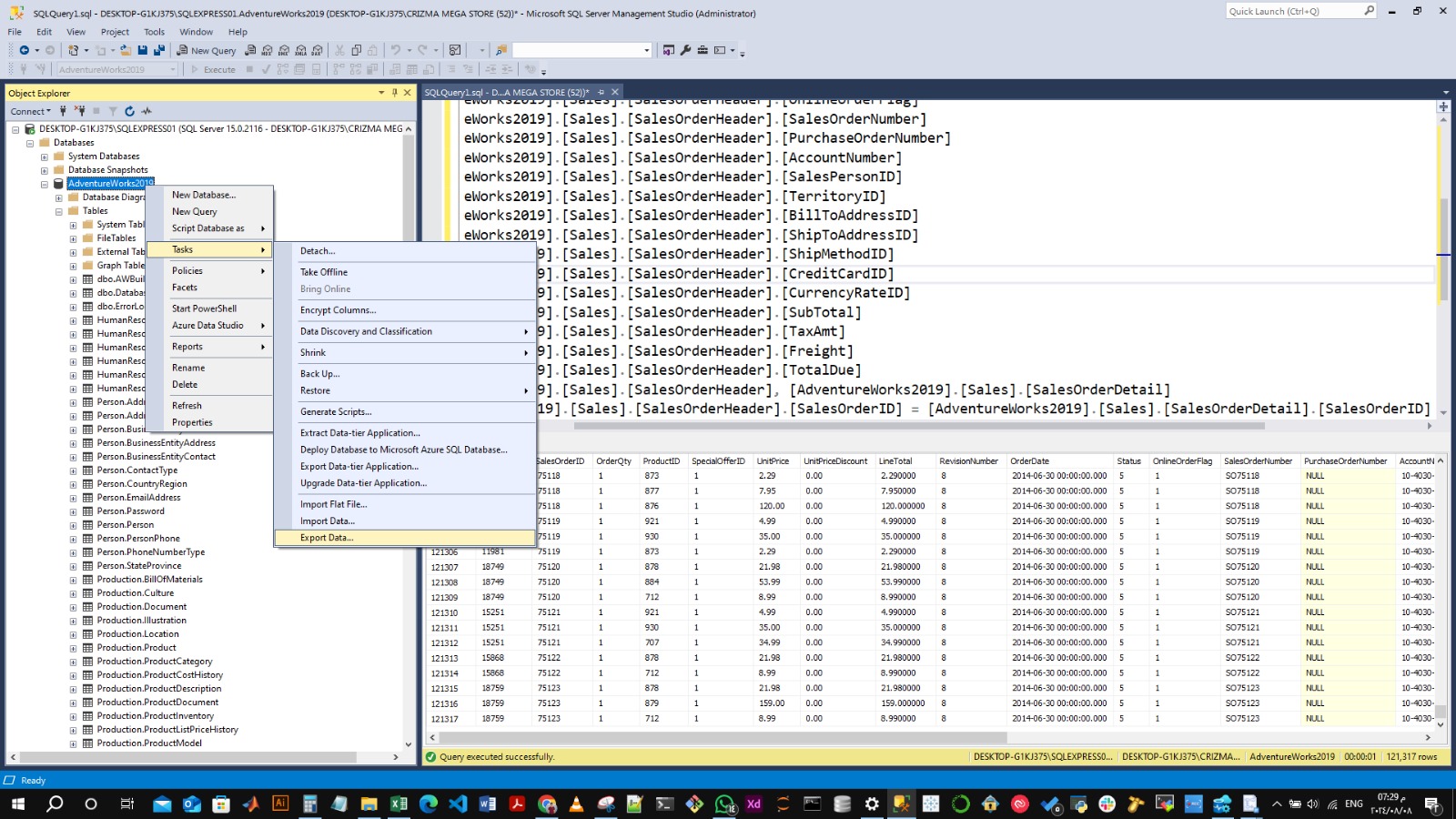
We used two methods to export our data.

**1.4.1. Azure Data Studio:**



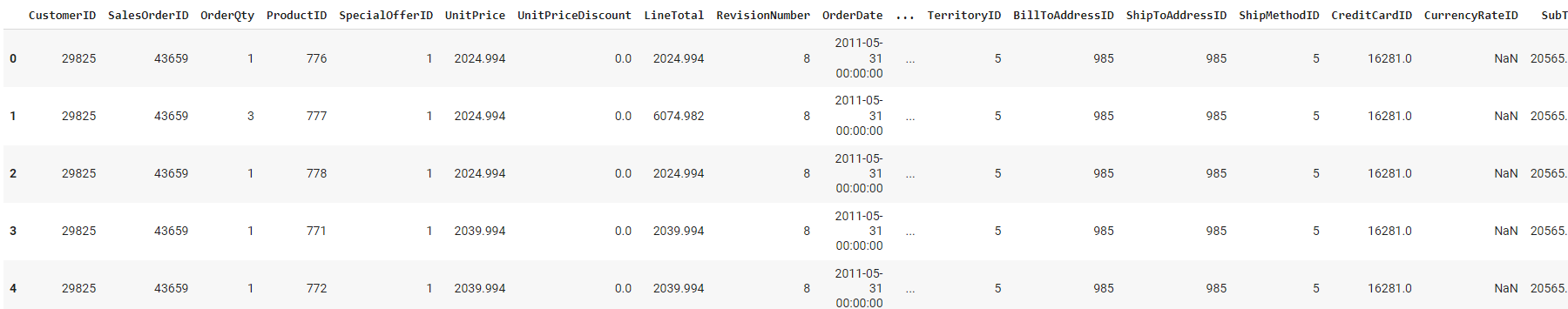
The issue with this method is its exporting limit to just export 5,000 rows of the data. We used different method to extract the full dataset.

**1.4.2. SQL Management Server:**



**1.5. Reading the dataset:**

We used the Pandas library to read the CSV file and store it in the Padas data frame.



Our dataset shape is 121,317 rows and 26 columns.

**Step 2: Data Pre-Processing:**

**In this step, we aim to clean our data, format it correctly, ensure its integrity, and detect the outliers.**

**2.1. Features Selection:**

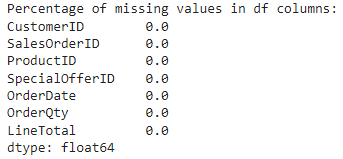
**Starting by selecting only the features we are interested in. For predicting the CLV we need our data to have the following columns 'CustomerID', 'SalesOrderID', 'OrderDate', 'OrderQty', 'LineTotal' to calculate the metrics values (Frequency, Recency, T, Monetary Values and Revenue).**

**For the EDA we need the columns 'ProductID', 'SpecialOfferID' to visualize useful findings and answer interesting questions about the business.**

**Our selected features data frame snapshot:**

**2.2. Data Cleaning: Missing Values**

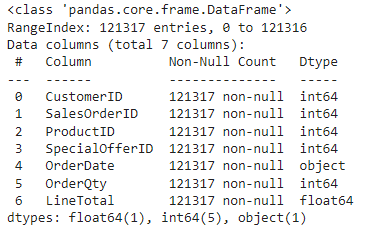
**Next step is to handle missing values in our data.**



**We found 0 missing values in our dataset.**

**2.3. Data Formatting and Structuring: Data Types**

**Ensuring our data are stored in the right format is a crucial pre-processing step.**



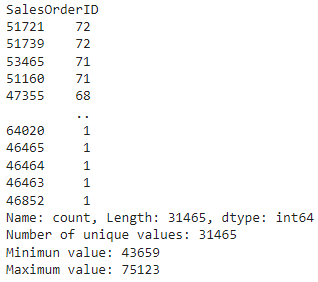
**We found the Orders Date values stored in the wrong format, so we need to cast it to datetime format instead of object format.**

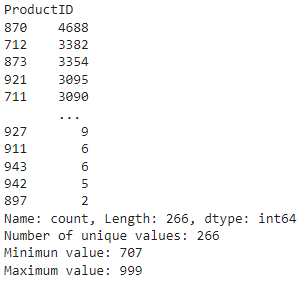


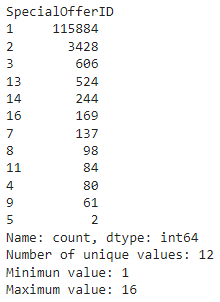
**2.4. Data Validation: Data Integrity**

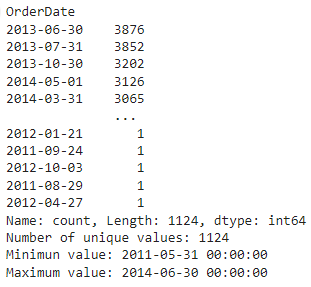
**We investigated the data inside our columns to ensure it integrity.**

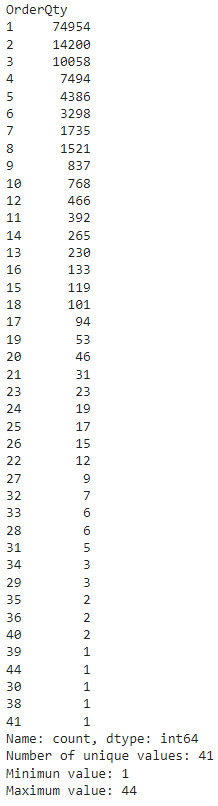


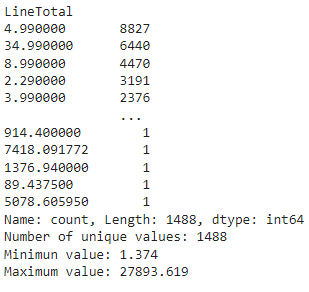








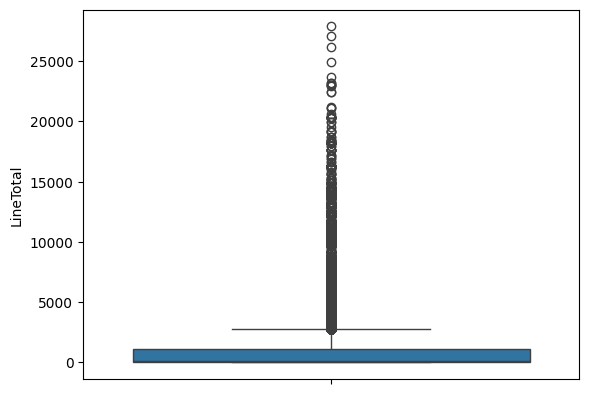
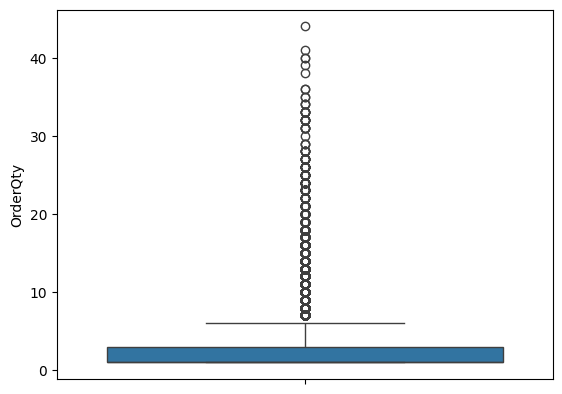
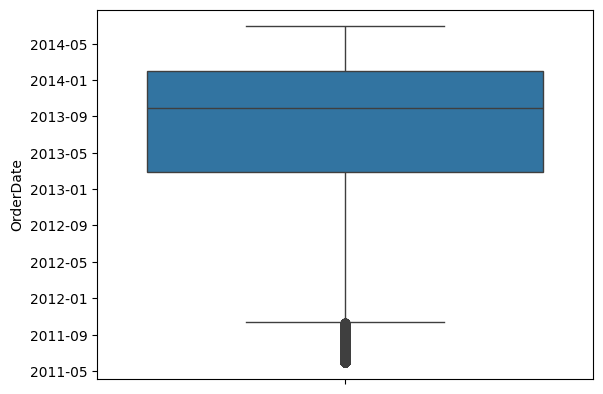
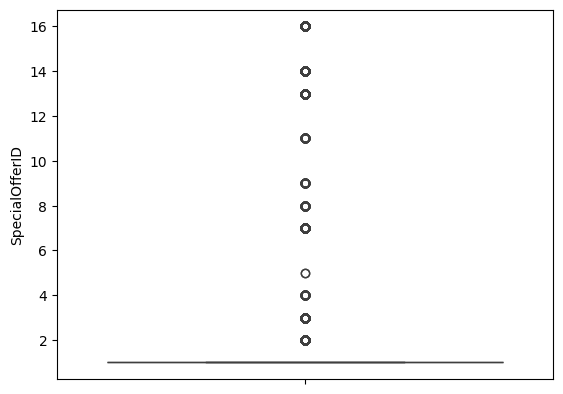
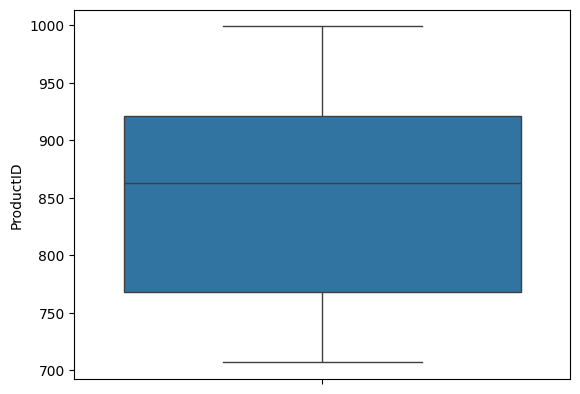
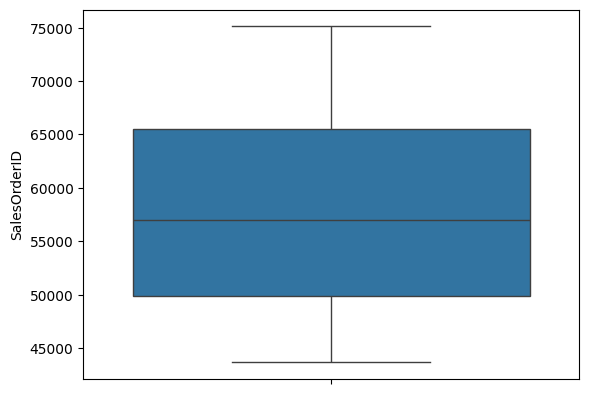
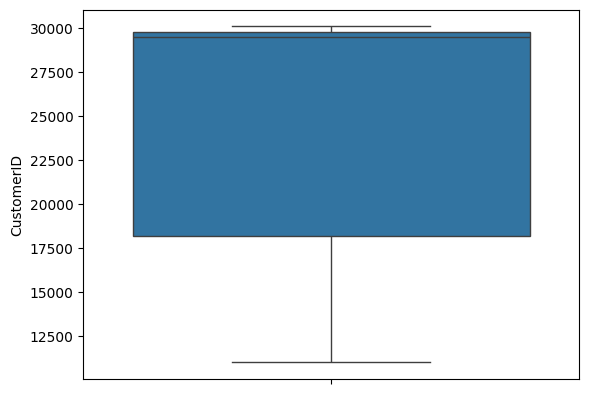




**There are no issues with the data stored in our columns. We conclude that our data are healthy.**

**2.5. Outliers Detection:**

**We will end our data pre-processing by defining the outliers in our dataset using boxplot, so we can handle it in the data processing step.**



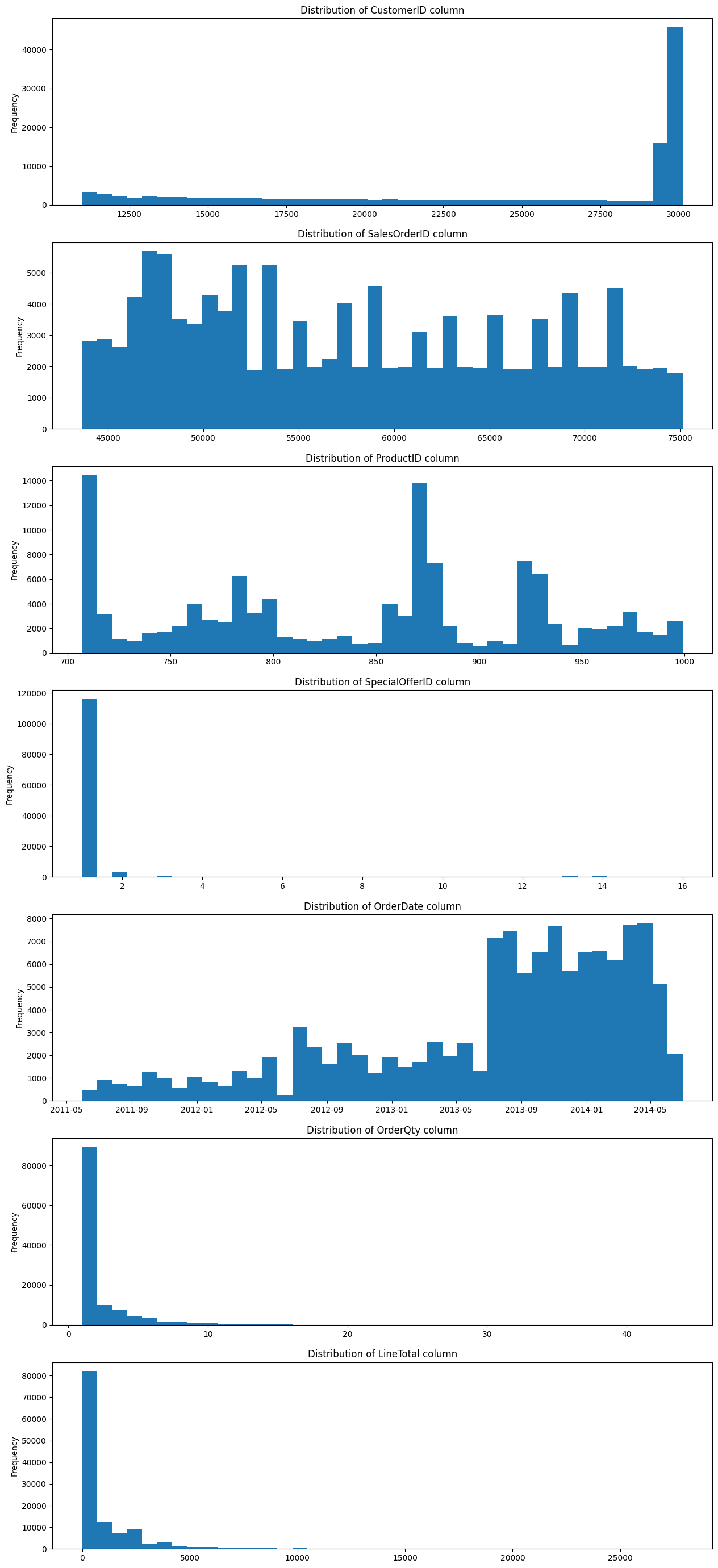
**We found outliers in columns 'SpecialOfferID', 'OrderDate', 'OrderQty', 'LineTotal’. We will address it the data processing step.**

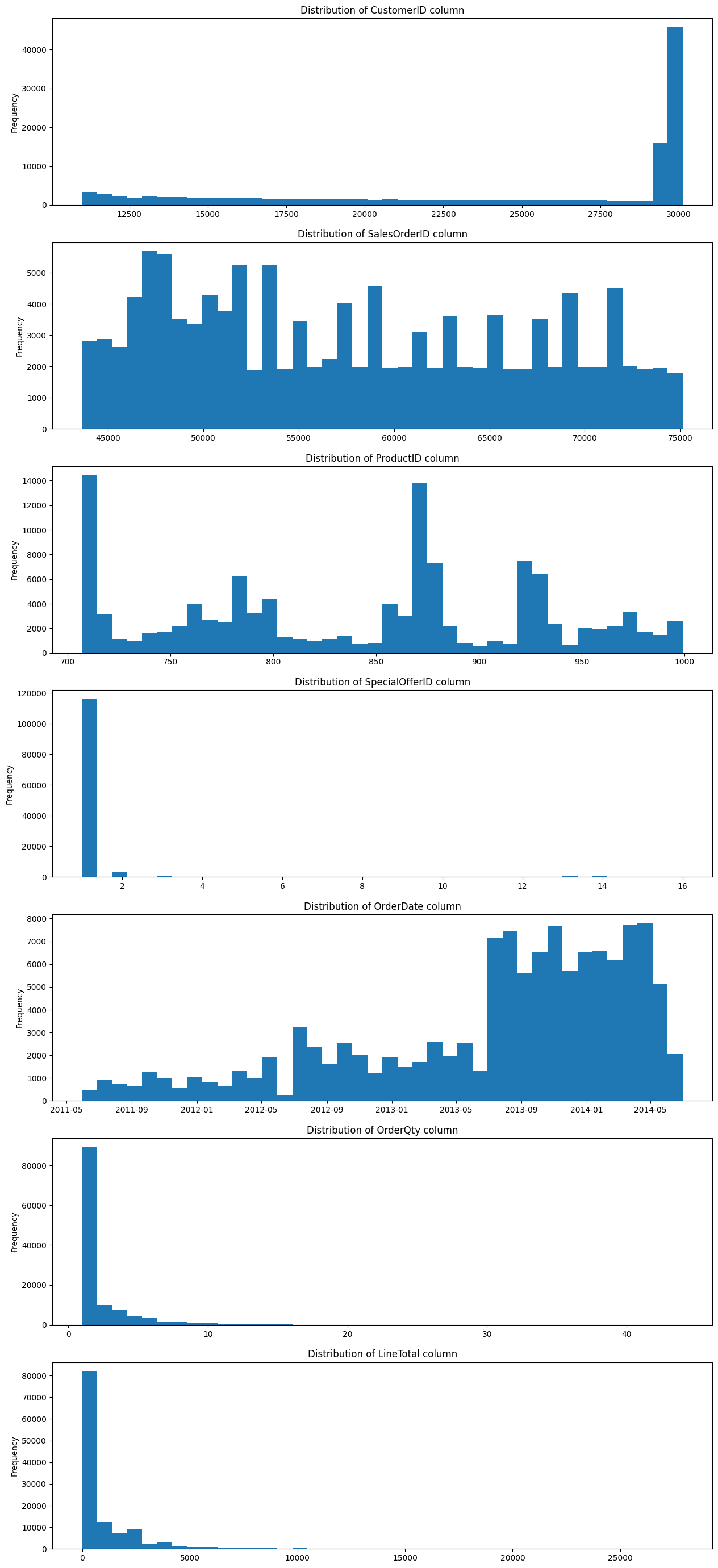
**Step 3: Data Analysis:**

**To understand our dataset better we will perform the following data analysis techniques.**

**3.1. Data Distribution:**

**Since all our features data are numerical features we will use the histogram to plot the distribution of our features data.**

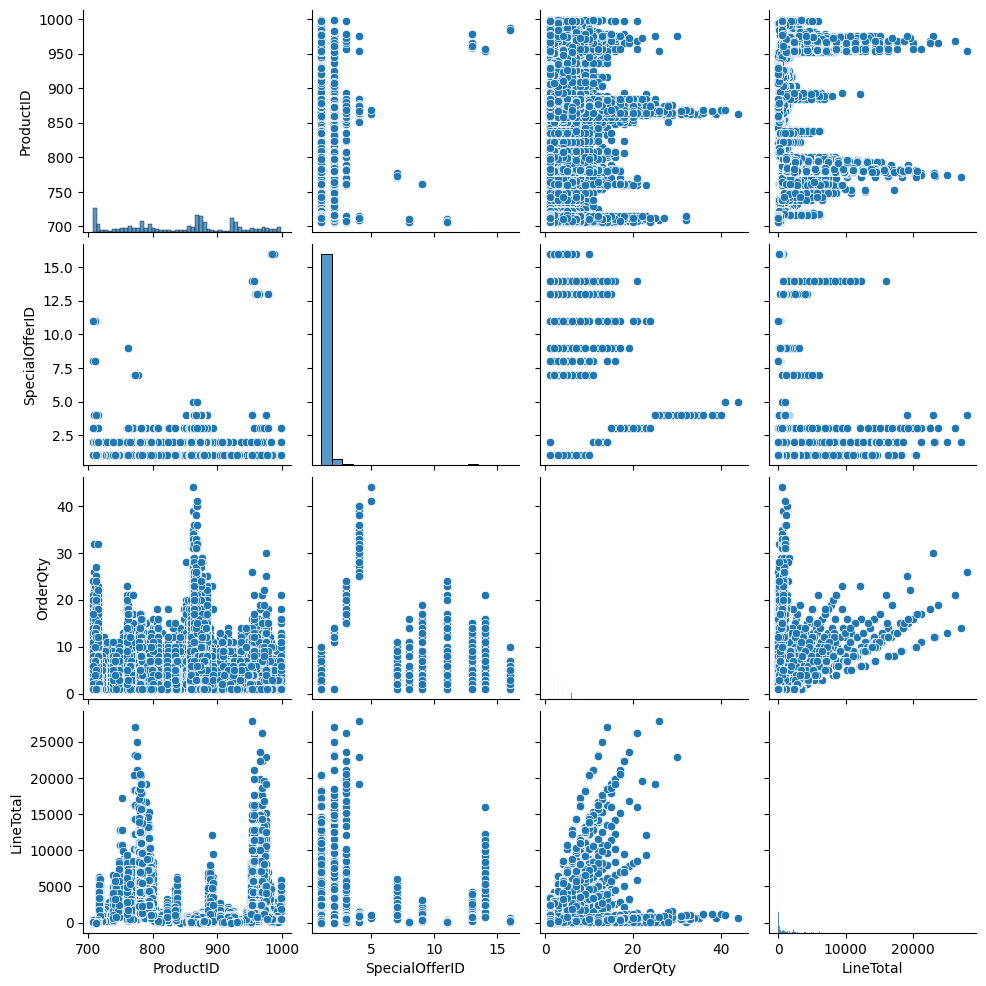


****

**We can interpret that the ‘OrderQty’ and the ‘LineTotal’ columns are right-skewed. We can also say that the ‘OrderDate’ is left skewed, and the rest of the columns don’t follow any distribution pattern.**

**3.2. Correlation Analysis: Pair plot (Scatter plot):**

**The next step in our analysis is to define the correlation between our columns. We used a scatter plot to visualize the correlation between our columns.**



**We can see there is no high correlation between our meaningful columns. But there is a sign of positive correlation between OrderQty and LineTotal which make sense.**

**3.3. Panda Profiling**

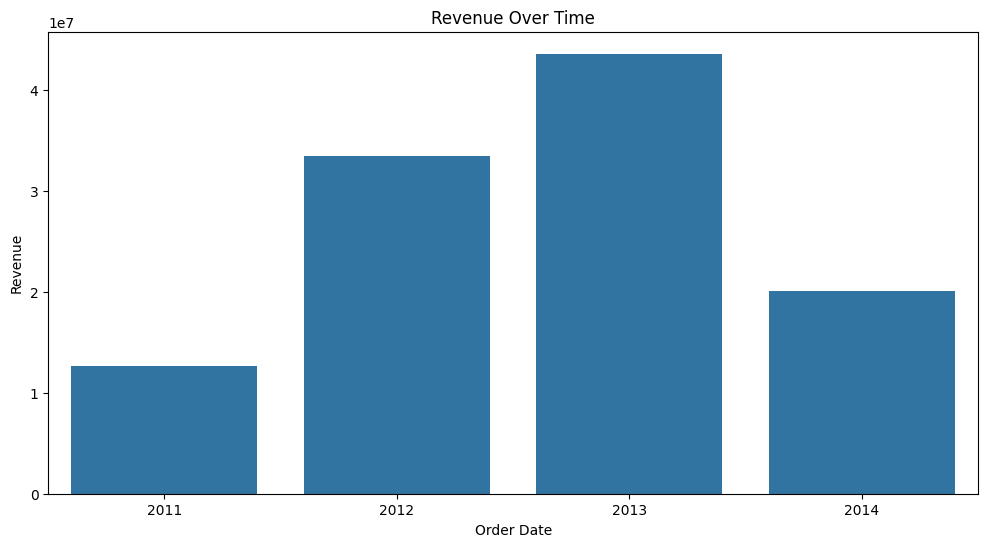
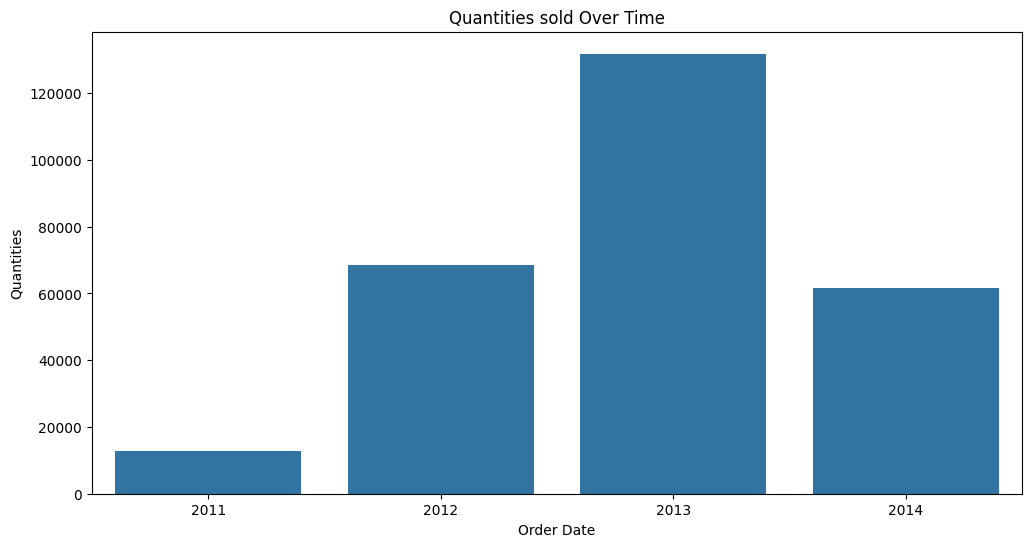
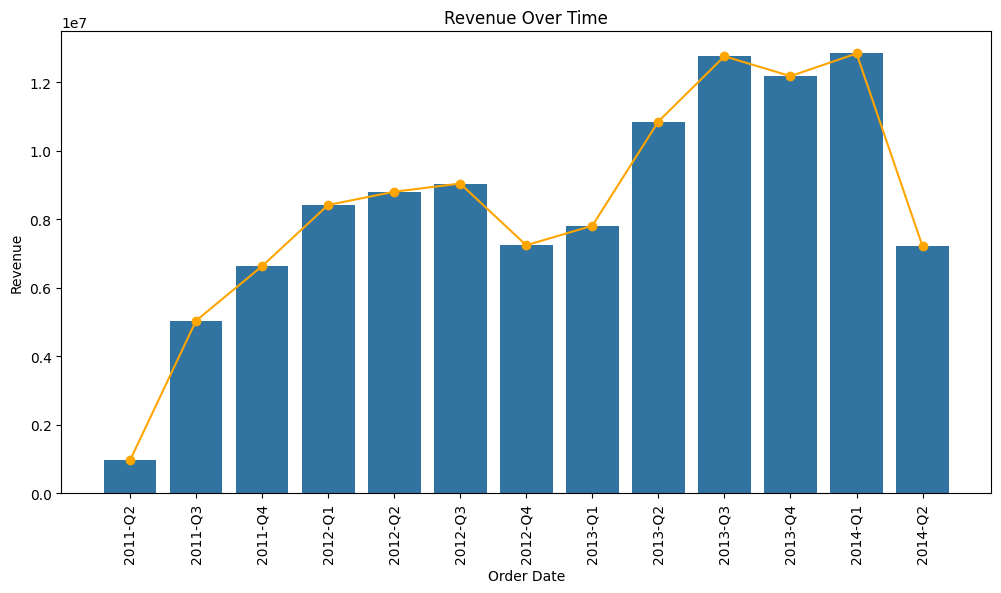
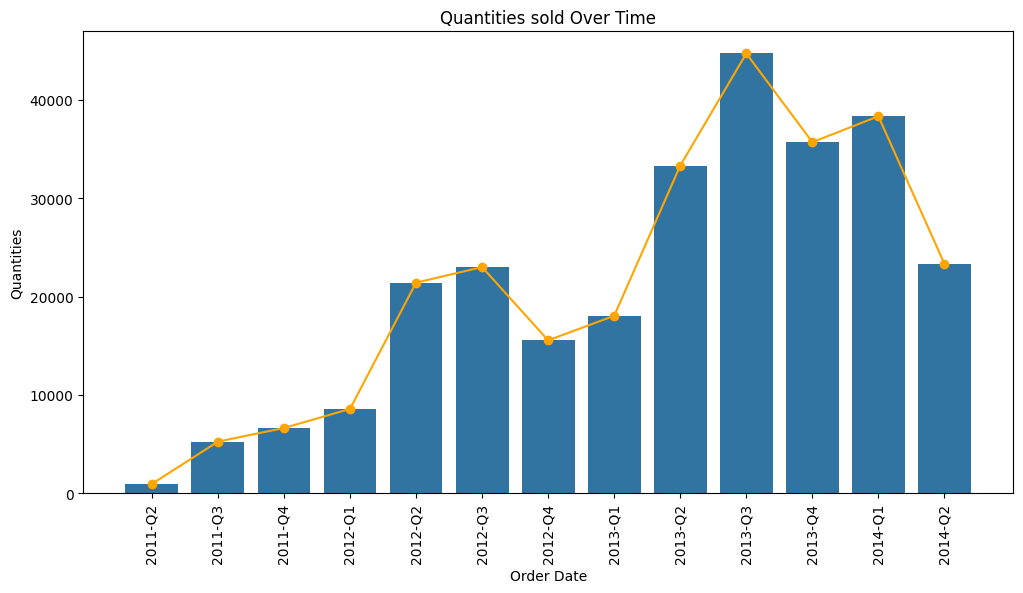
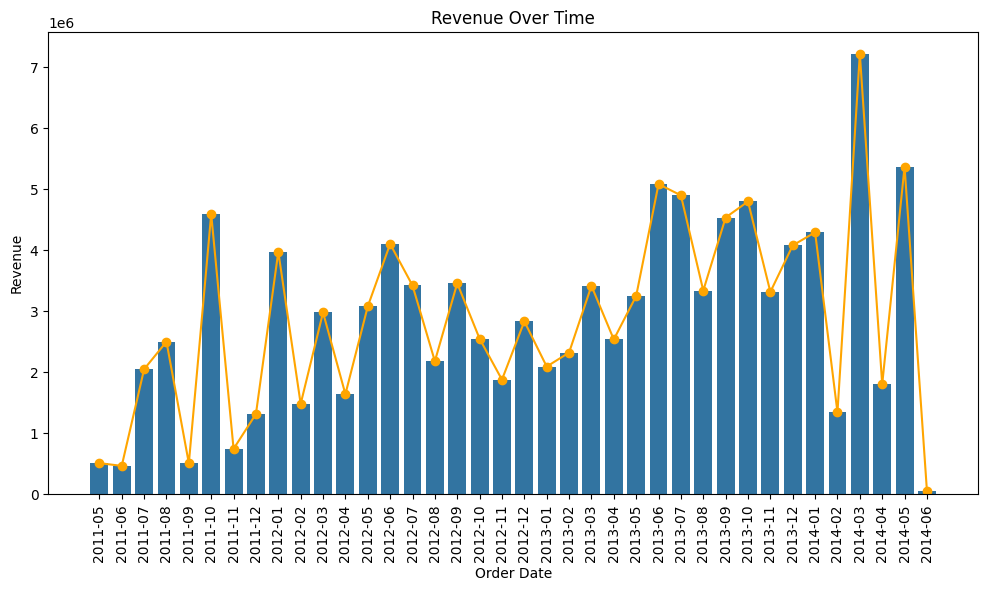
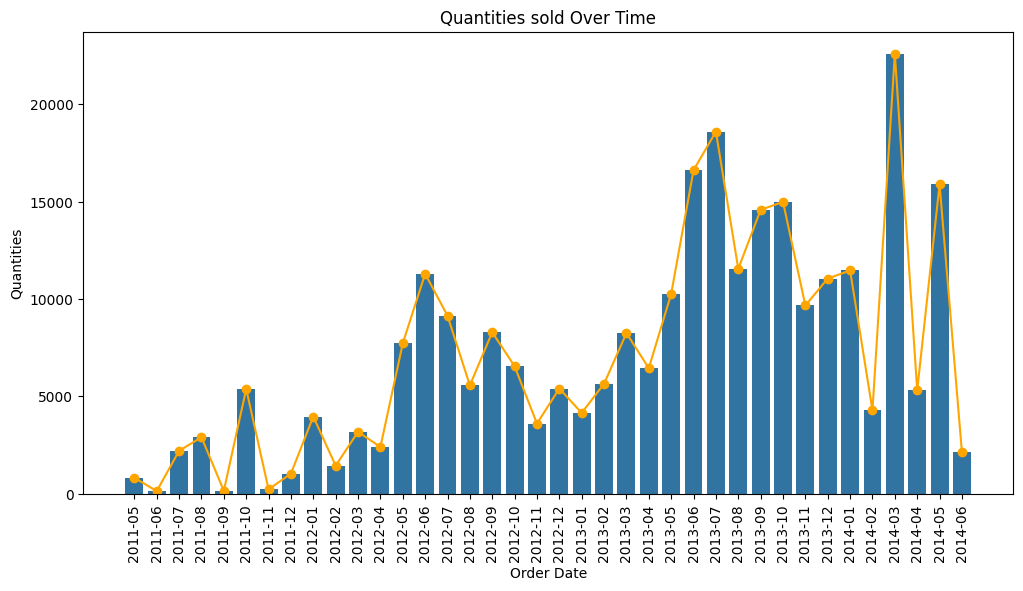
**To summaries our data profiling in one document we used Pandas profiling library.**

**The profiling report shows the same result we already done, with 2 alerts about the correlation between the CustomerID and OrderQty which may be result of the sampling method used to generate the data by Microsoft, and it’s not worth worry about since it doesn’t indicate any meaningful information.**

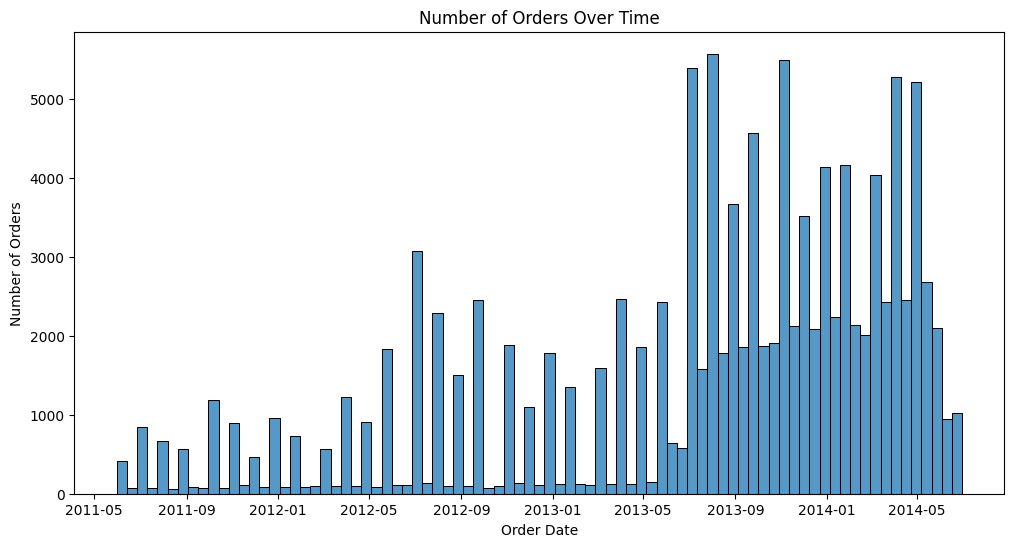
**3.4. EDA:**

**Pursuing with the data analysis we want to understand our dataset better, to do so we will explore our data by answering the following question:**

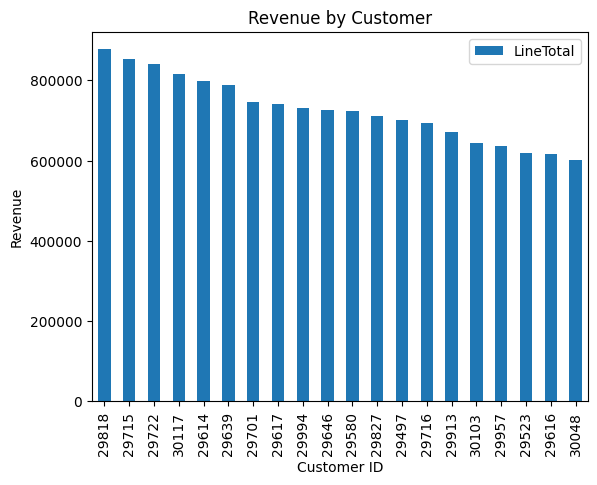
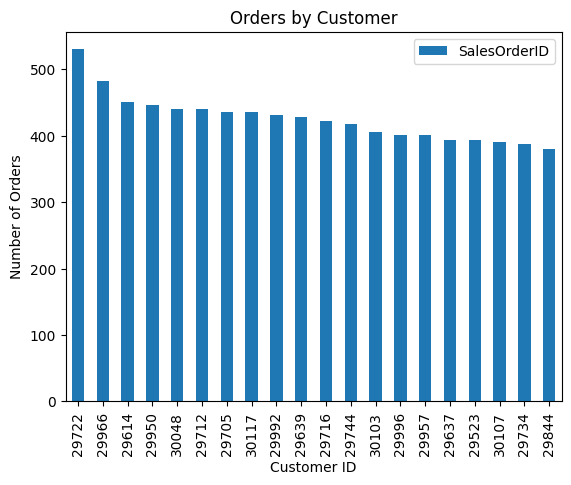
**3.4.1. What are the Sales Trends over Time?**

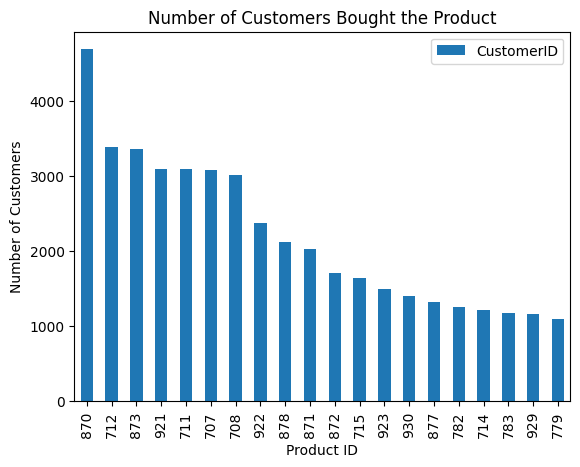


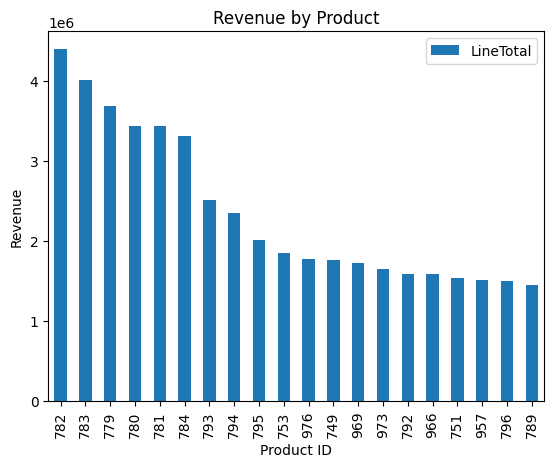
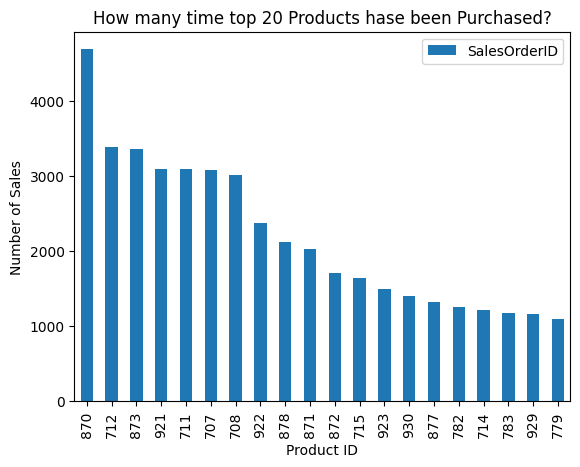
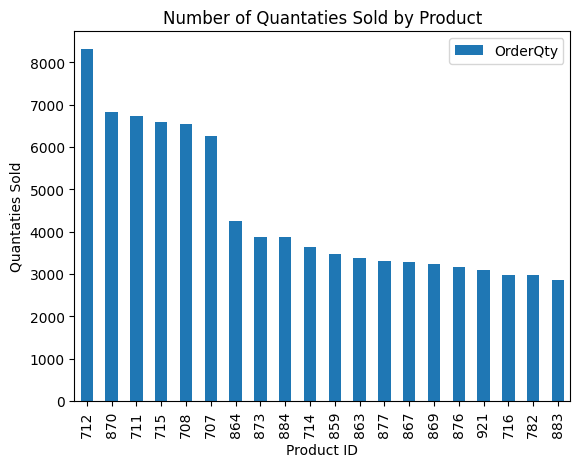
**3.4.2. What are the number of order over time?**



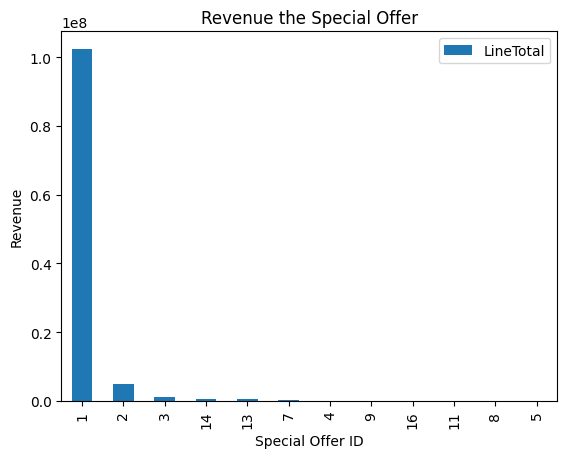
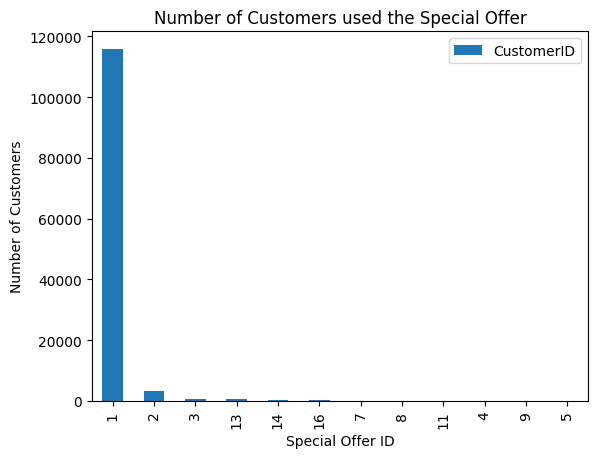
**3.4.3. What are the customers purchase behaviors?**



**3.4.4. What are the popularity for the products?**



**3.4.5. What are the most effective offers?**



**Step 4: Data Processing and Metrics Calculation:**

**4.1. Outliers Handling: Trimming**

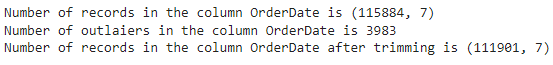
We will drop the observations greater than the upper whisker, or less than the lower whisker.











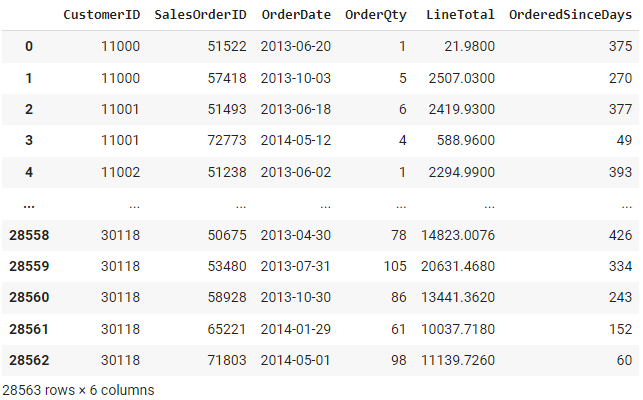




The shape of our data after trimming the outliers is 99,562 rows and 7 columns

**4.2. Features Engineering**

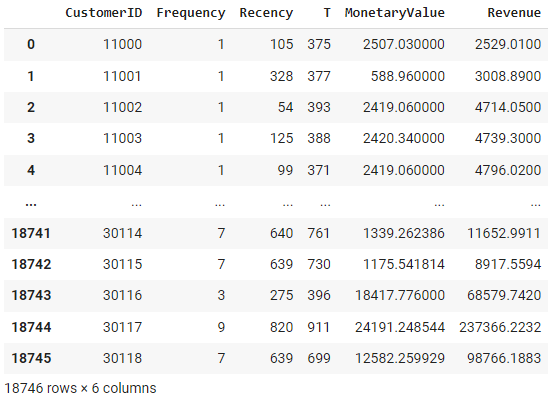
We engineered new feature to represent the days from the last order and the last observed date in our dataset for each customer. This features help us define the inactive customer and it can be used to indicate if the churned customers for further analysis.



**4.3. Metrics Calculation**

To calculate and then predict the CLV we need to transform our observation data to a set of metrics data represent a specific feature about the customer instead of dealing with non-informative data forms such as OrderDate, SalesOrderID, and OrderQty in their original format. We will calculate the following metrics.

* Frequency — the number of repeat purchases (more than 1 purchases)
* Recency — the time between the first and the last transaction
* T — the time between the first purchase and the end of the transaction period (last date of the time frame considered for the analysis)
* Monetary Value — it is the mean of a given customers sales value



We calculated the metrics for our 18,746 customers.

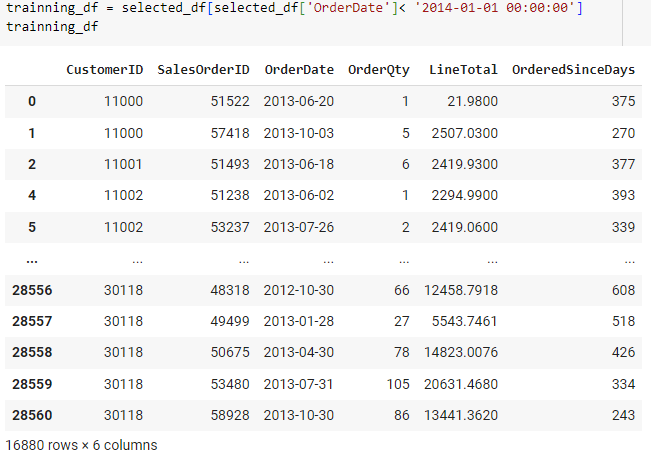
**Step 5: Data Modeling:**

**Next step is to use modeling methods to calculate or predict the CLV.**

**4.1. Data Preparation:**

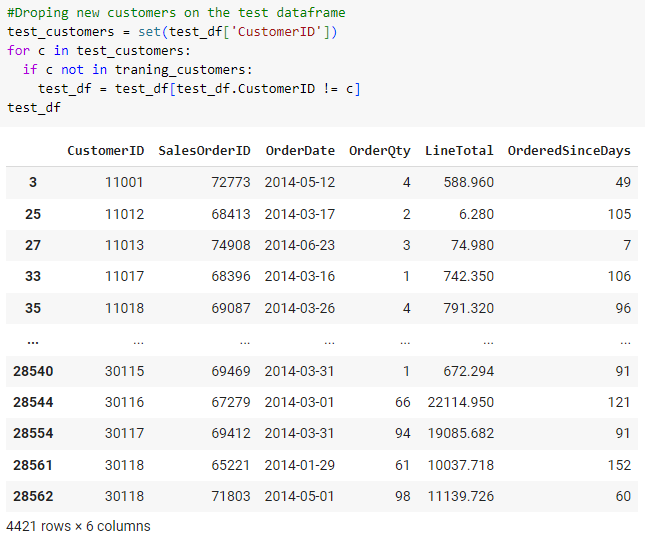
**Before we process/train our data with models we need to prepare it.**

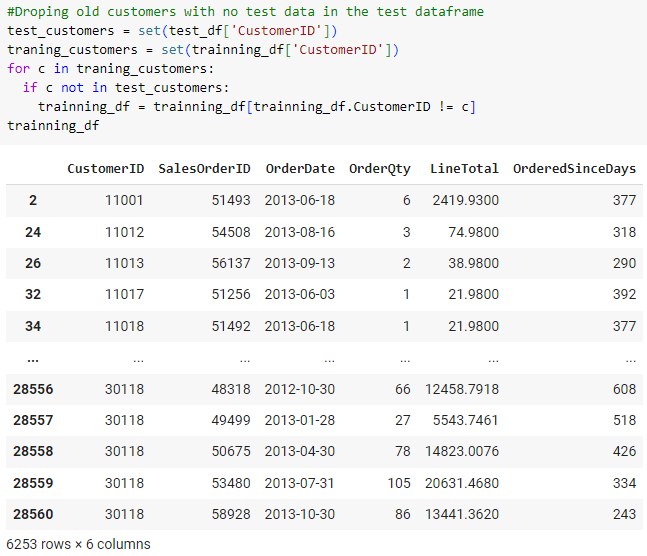
**We need to split our data in time-wise. We will spate the data of the last 6 months observed to be our target data which we will use to evaluate our predictive models.**



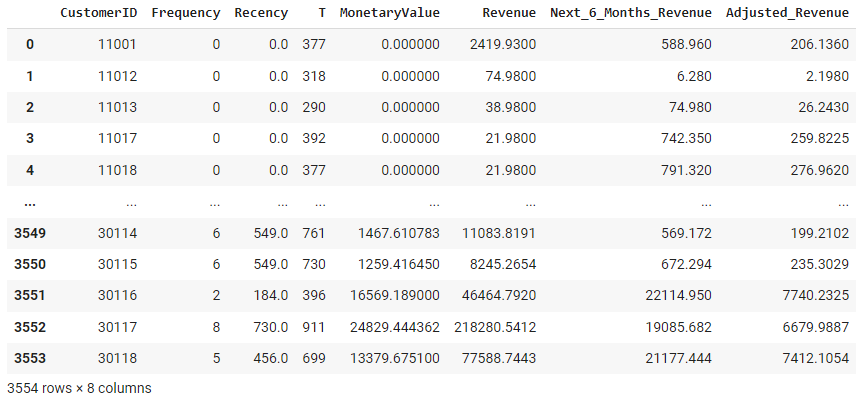


**Then we will drop the uncommon observations.**





**Now we are ready to calculate the metrics for the customers eligible for modeling.**



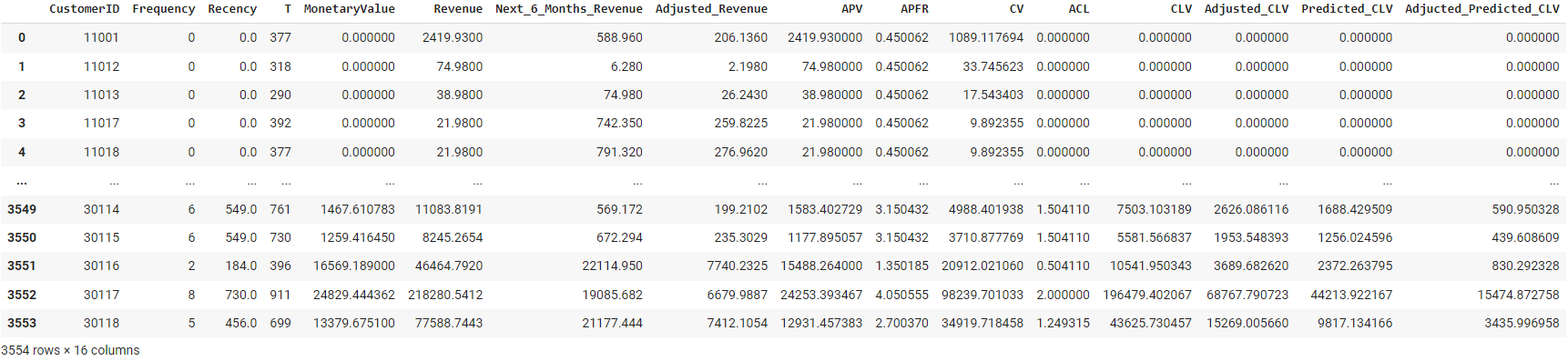
**We ended up with 3554 customers that we have enough data for modeling to predict and validate their CLV and model performance.**

**Method 1: [Benchmark] Historical CLV**

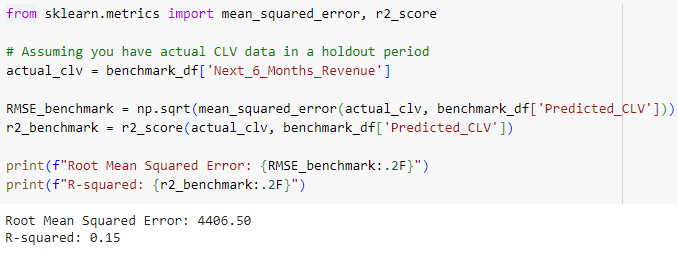
**We will use the simple historical method to predict the CLV as a benchmark to compare the models performance with.**

**To calculate the historical CLV and predict the CLV for the next 6 months based on historical CLV we will follow the following steps:**

1. **Calculate AVP = Total Revenue/ Number of Purchases**
2. **Calculate APFR = Number of Purchases/ Number of Unique Periods**
3. **Calculate CV = APV \* APFR**
4. **Calculate ACL = Last Purchase - First Purchase (in years)**
5. **Calculate CLV = CV \* ACL**
6. **Calculate Adjusted\_CLV = CLV\* Average gross margin**
7. **Calculate Predectied\_CLV for next 6 months = CLV \* (0.5 / Period in years)**
8. **Calculate Adjusted\_Predicted\_CLV = Predectied\_CLV\* Average gross margin**



**To measure the accuracy of the method results we will compare the predicted CLV to the known CLV for the last 6 months for each customer.**



**The RMSE for the benchmark historical method is 4406.50 from the actual value.**

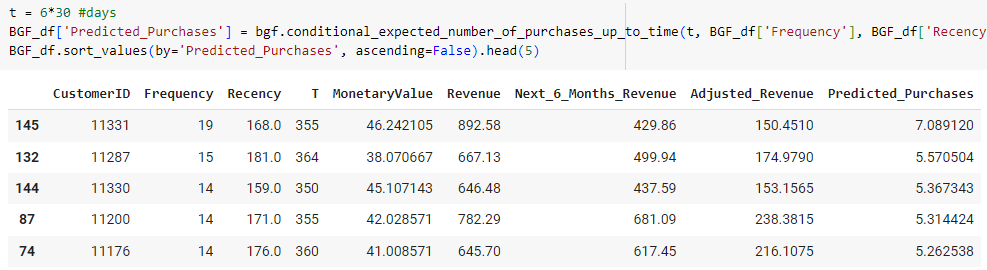
**Method 2: Probabilistic BG/NBD**

**The BG/NBD is a probabilistic model uses the customer buying behavior and customer churn rate to predict the number of next purchases the customer will place in specific amount of time, the customer churn state (if the customer will stay alive customer), and the CLV.**

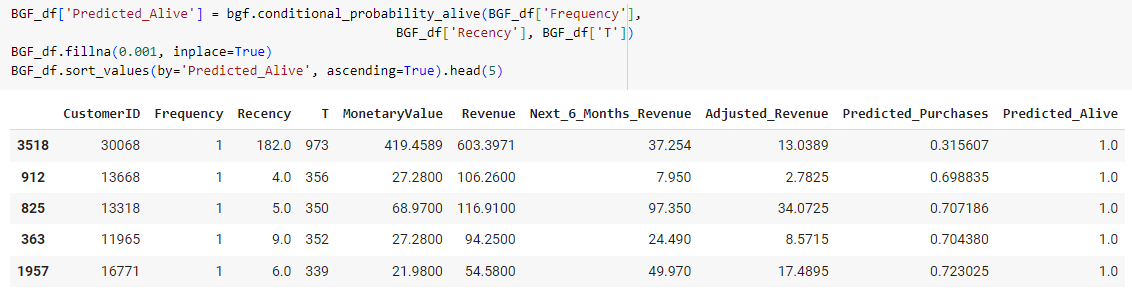
**The model uses the combination of Poisson and Gamma distribution which is called negative binomial distribution (NBD) to model the customer buying behavior.**

**It uses Beta distribution to predict the churn probability and the CLV of the customer.**

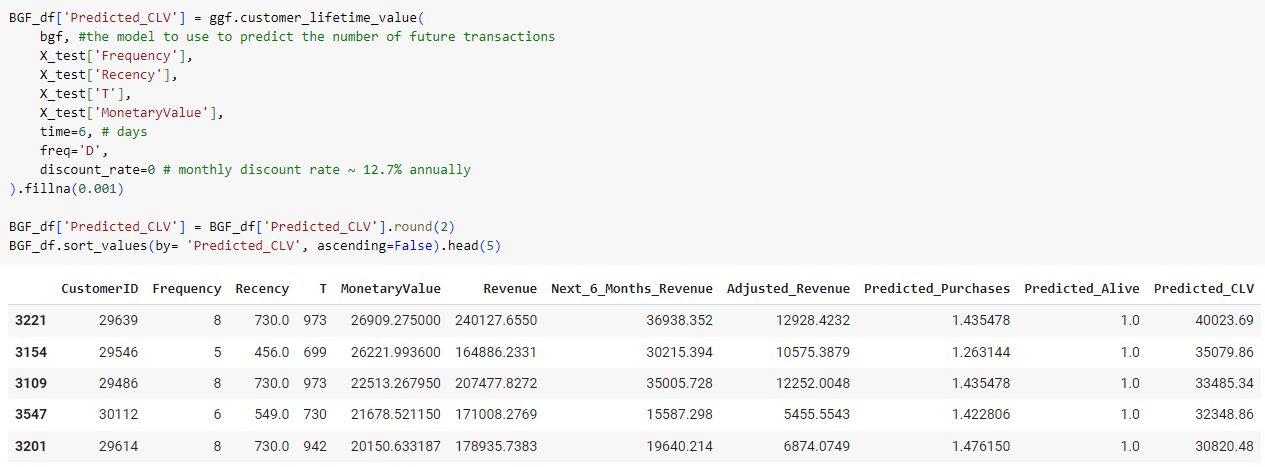
**To perform the model, we will start with predicting the number of purchases for each customer in the next 6 months.**



**Then we will predict the churn probability where 1 indicate the customer won’t churn in the specified period.**



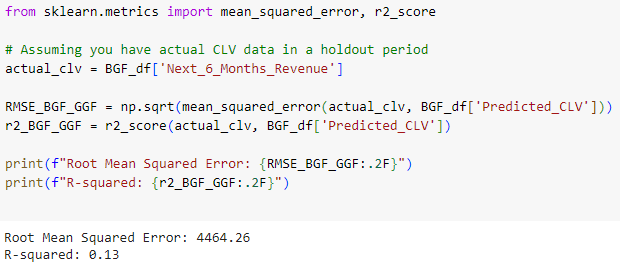
**Now, we will utilize the model to predict the CLV.**



**We notice a NaN results for some customers from the model, this could happen due to one or more of the following reasons:**

* **If Recency and T are very close.**
* **If the monetary values are too similar.**
* **If there is only one transaction (Frequency = 1), the model may not be able to generate a meaningful prediction, leading to NaN values.**

**Finally, we will evaluate our model.**

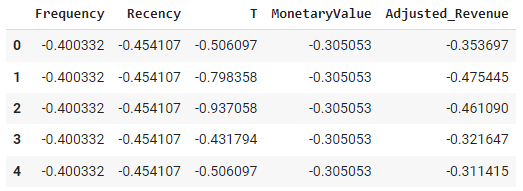


**The BG/NBG model predict the CLV with RMSE of 4464.26 which is higher than the benchmark model meaning either the data is not sufficient for the model or our implementation for the model is not perfect.**

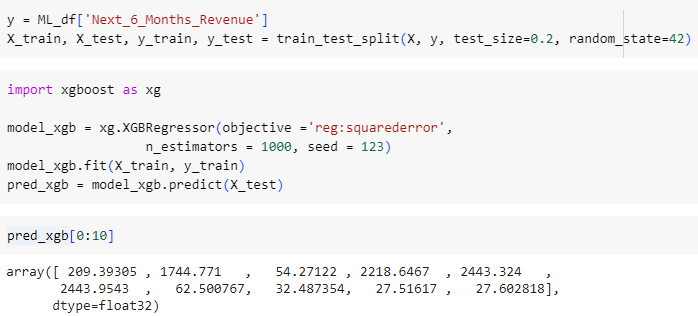
**Method 3: Machine Learning Regression**

**Next, we will use a supervised machine learning regression technique to predict the CLV for the next 6 months using XGBoost.**

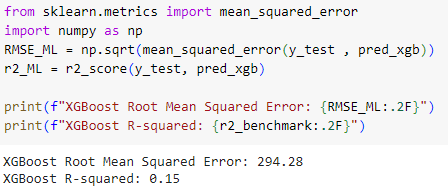
**Stating with standardizing our data to avoid bias.**



**Next, we will train our model**



**Then validating our model**

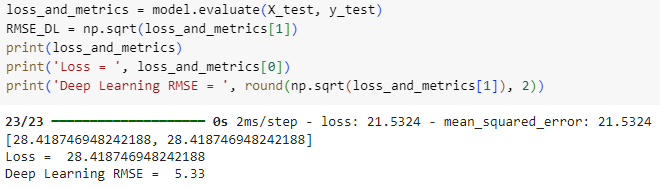


**The XGBoost model predicted the CLV with RMSE of 294.28 which represent a high accuracy for the model.**

**Using this trained model we ca predict the CLV for the next 6 months with expected error margin of 294.28 revenue unit.**

**Method 4: Deep Learning Linear**

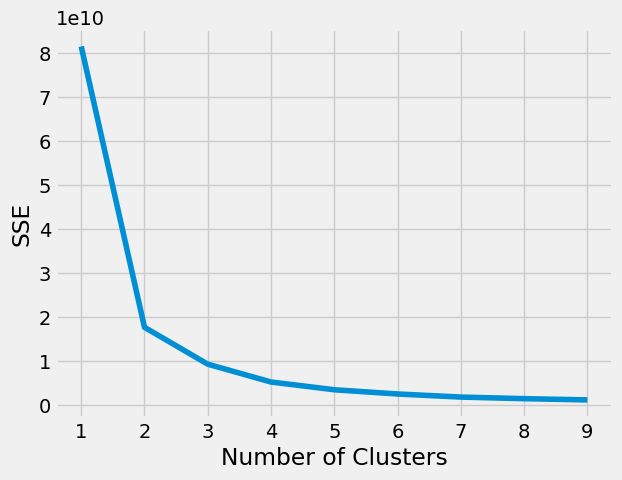
**Next, we will model an artificial neural network (ANN) to predict the CLV for the next 6 months using deep learning.**



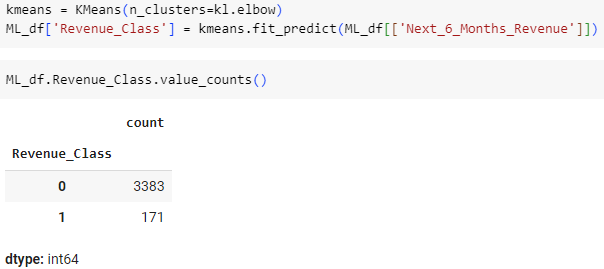
**The RMSE for the deep learning linear model is 5.33 that beats the XGBoost model and rank as the best predictive model for CLV.**

**Method 5: Machine Learning Segmentation and Classification**

**In this step we will try to classify and predict the customer Life Value classes (High LV, Low LV) using both segmentation with KMeans and classification with XGBoost.**



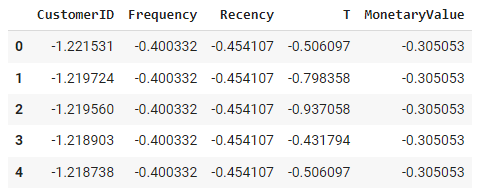
**The elbow method indicate there are 2 classes for our revenue (LV) column.**

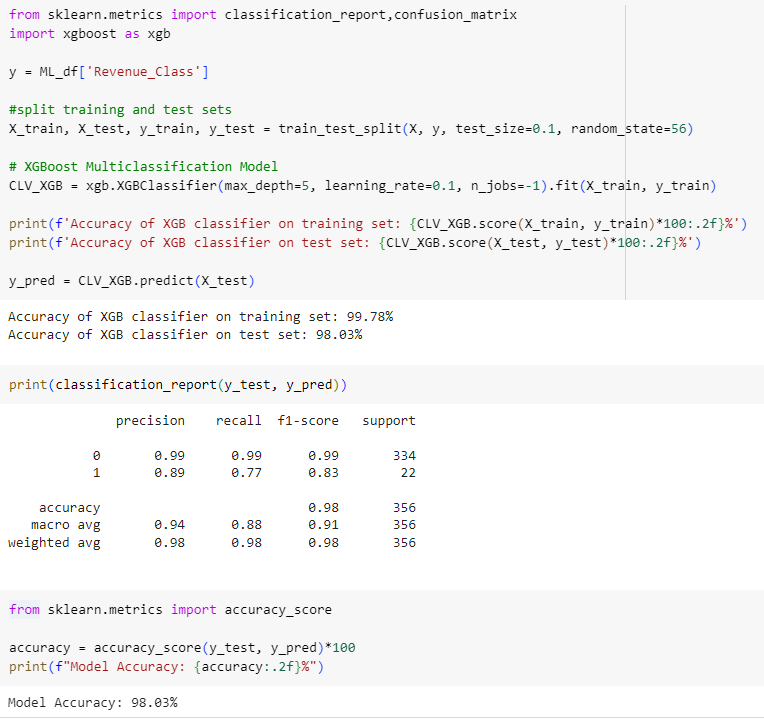


**Most of our customers are on the Low LV class**



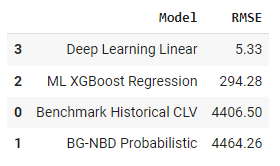
**Standardization to avoid bias.**

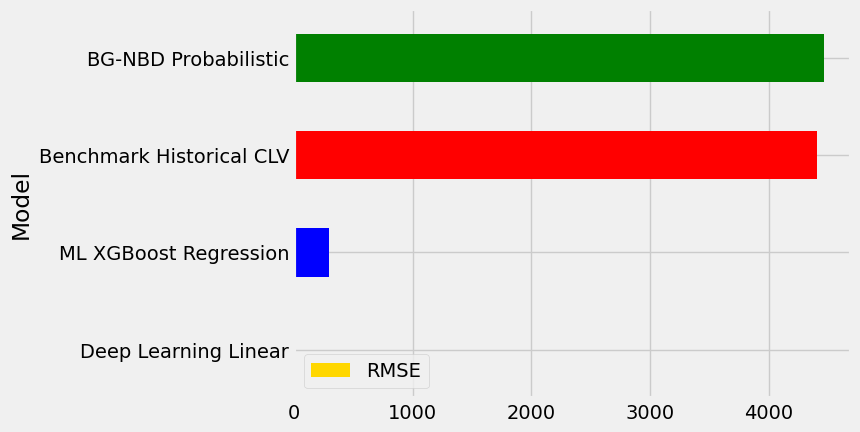




**Step 6: Models Evaluations**

**Summary and rank of the models performance:**





**The very low error for the linear deep learning model may indicate an overfitting, this can be validated/solved by having more data points.**

**The ML model error indicates a balanced model this also can be enhanced by providing more data to the model.**

**The BG-NBD errors indicate either the data is not sufficient for the model makes it under fitted or our implementation of the model is not perfect. Some potential errors with the data:**

* **If Recency and T are very close.**
* **If the monetary values are too similar.**
* **If there is only one transaction (Frequency = 1), the model may not be able to generate a meaningful prediction, leading to NaN values.**

**The benchmark error represents an average performance for a simple historical model.**

**The classification model is fitting very well for the dominant class (Low LV) and doing good for the non-rich data class (High LV) this resulted in loss in recall that can be increased with more data points for the class (High LV).**

**Results**

**We have successfully predicted the CLV for 3554 customers with 5 different methods, ranked by the RMSE as follows Liner Deep Learning 5.33 RMSE (a potential over fitted model), ML XGBoost Regression 294.28 RMSE (balanced model), Benchmark Historical CLV 4406.50 RMSE (simple model), and the BG-NBD Probabilistic 4464.26 RMSE (under fitted or wrong implementation), and The classification model is fitting very well for both classes with classification accuracy of 98%.**

**Through the EDA we found discovered that a single special offer is used 93% of the times the customers used a special offer. The sales trended to increase from the years 1-3 and dropped in the 4th years that may be resulting of only the first 6 months of the 4th year was recorded. The top customer by orders is different than the top customer by revenue with the first had over 500 orders and the second contributed over 800,000 revenue. We identified the most popular product have been ordered over 4000 times and the top selling product sold over 8000 quantities and the most revenue-generating product generated 4 million unit price.**

**To achieve this, we queried our data from database server with SQL resulting in acquiring 121,317 data records. We performed data pre-processing using Pandas resulting in cleaning the data from missing values, cast the data into the right format, ensured data integrity, and detect outliers. We analyzed the data by visualizing its distribution, and correlation resulting on define skewed features and highly correlated features. We summarized the data profile using Pandas profiling. We performed extensive data preparation, outliers trimming, data processing, modeling and evaluation.**

# **Conclusions and Future Work**

Predicting the CLV is valuable to the business due its association with each customer behavior and value contribution directly. By predicting the CLV correctly it opens the door to many advanced application and use cases to the predicted metrics. One of the potential future works is to build an automated pipeline to predict the CLV for the customers periodically and automatically, some of the technology stack that could be used is Airflow, Docker, and AutoML, the future application would increase the models performance by retraining them on larger and recent data, this app could utilize the distributed computing power of apache Hadoop for batch processing.

In other hand this project can be integrated to many CRM systems and business managements systems to add the predicted CLV into account for more advanced business decision.

# **References**

# **Scikit-learn Machine Learning in Python documentation:** <https://scikit-learn.org/stable/>

* Pandas documentation: <https://pandas.pydata.org/docs/>
* Pandas profiling: <https://pypi.org/project/pandas-profiling/>
* Customer Life Time Value Prediction: <https://www.analyticsvidhya.com/blog/2020/10/a-definitive-guide-for-predicting-customer-lifetime-value-clv/>