**BIG DATA ANALYSIS WITH IBM CLOUD DATABASES**

Phase-3: Development Part 1

Big Data Analysis

**Problem Statement:**

* Dive into the world of big data analysis with IBM Cloud Databases.
* Uncover hidden insights from vast datasets, from climate trends to social patterns.
* Visualize your findings and derive valuable business intelligence.
* Embark on data-driven adventures, exploring the endless possibilities of big data.

**Problem Definition:**

* The project involves delving into big data analysis using IBM Cloud Databases.
* The objective is to extract valuable insights from extensive datasets, ranging from climate trends to social patterns.
* The project includes designing the analysis process, setting up IBM Cloud Databases, performing data analysis, and visualizing the results for business intelligence.

Phase 3

**INTRODUCTION:**

In this part we will begin building the project. Start building the big data analysis solution using shopping trending Dataset. In this project we will explore and analyze the selected dataset. Perform basic data cleaning and transformation as needed.

**PROCEDURE:**

**STEP 1:** Download the shopping trending dataset from the thesis paper.

**STEP 2:** Analyse the data from the given dataset.

**STEP 3:** If any duplicates, null value occurrence in the dataset then those errors are removed immediately from the dataset.

**STEP 4:** To remove those anomalies we are using processes called DATA CLEANING and DATA TRANFORMATION.

**DATA CLEANING:**

Data cleaning, also known as data cleansing or data scrubbing, is a fundamental step in the data preprocessing phase of a data science project. It involves identifying and correcting errors or inconsistencies in datasets to improve their quality and reliability for analysis. Data cleaning is essential because real-world data is often messy and may contain various issues that can lead to inaccurate or unreliable results if not addressed.

Common tasks in data cleaning include:

1. Handling Missing Values: Dealing with missing data by either removing the rows or columns with missing values, imputing missing values using statistical methods, or employing more advanced techniques like interpolation.
2. Correcting Inaccurate Data: Identifying and correcting inaccurate data points, such as outliers or values that fall outside of valid ranges. This may involve replacing or imputing incorrect values.
3. Standardizing Data: Ensuring that data follows a consistent format, including date formats, units of measurement, and encoding. Standardization makes it easier to work with the data and compare different data sources.
4. Deduplication: Identifying and removing duplicate records or entries in the dataset, which can skew analysis results and waste computational resources.
5. Data Transformation: Converting data into a suitable format for analysis, such as converting categorical data into numerical representations (e.g., one-hot encoding), scaling numeric features, or applying logarithmic transformations.
6. Handling Noisy Data: Addressing noisy data, which includes data points with errors or inconsistencies caused by factors such as measurement errors or data entry mistakes.
7. Dealing with Inconsistent Data: Ensuring that data from multiple sources align in terms of naming conventions, units, and other criteria.
8. Addressing Outliers: Identifying and either removing or transforming outliers that can have a significant impact on statistical analyses or machine learning models.
9. Addressing Data Integrity Issues: Checking for data integrity issues, such as referential integrity in databases, to ensure data relationships are maintained.
10. Data cleaning is a critical step in the data science workflow because the quality of the data directly influences the accuracy and reliability of any subsequent analyses, modeling, or machine learning tasks. Clean data leads to more meaningful insights and better decision-making. Data scientists often spend a significant portion of their time on data cleaning to ensure the data is fit for analysis.

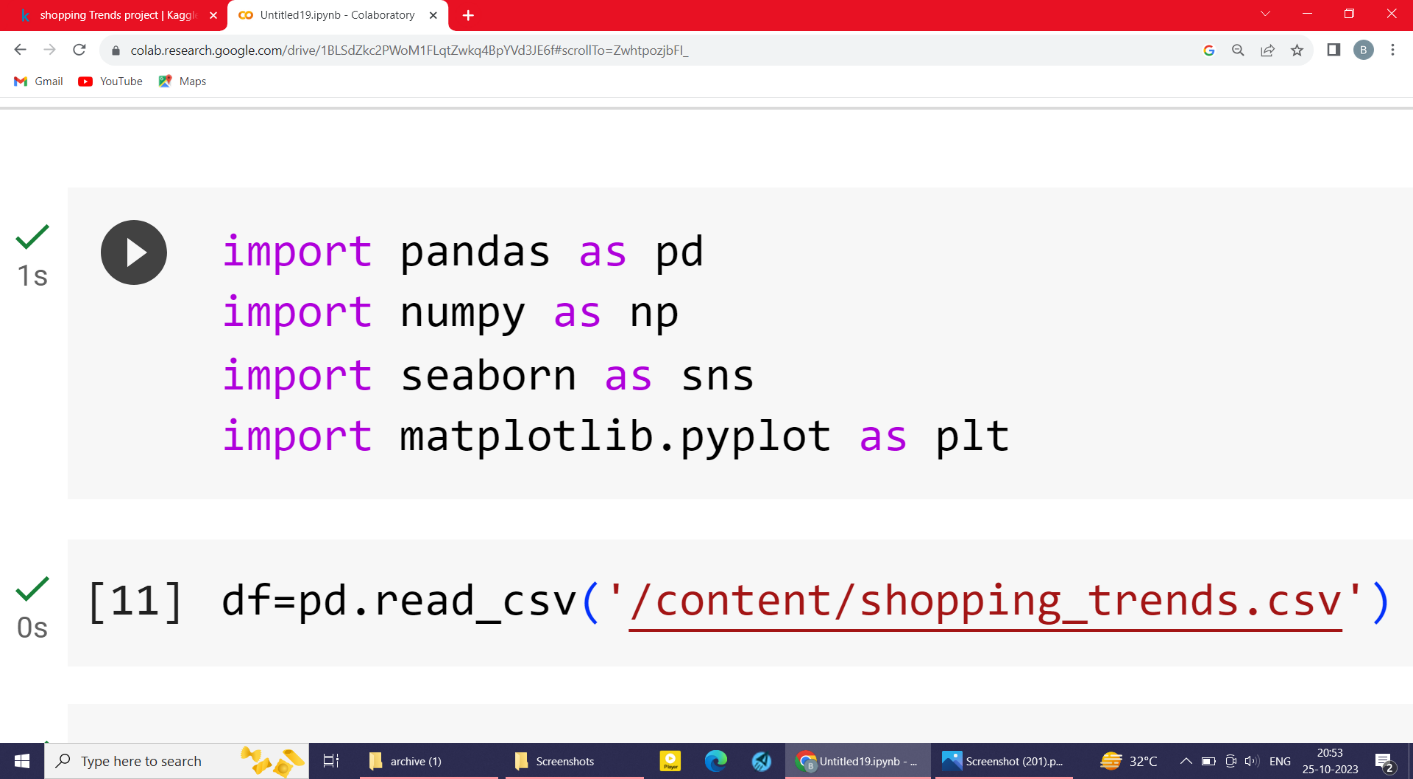
**DATA TRANSFORMATION:**

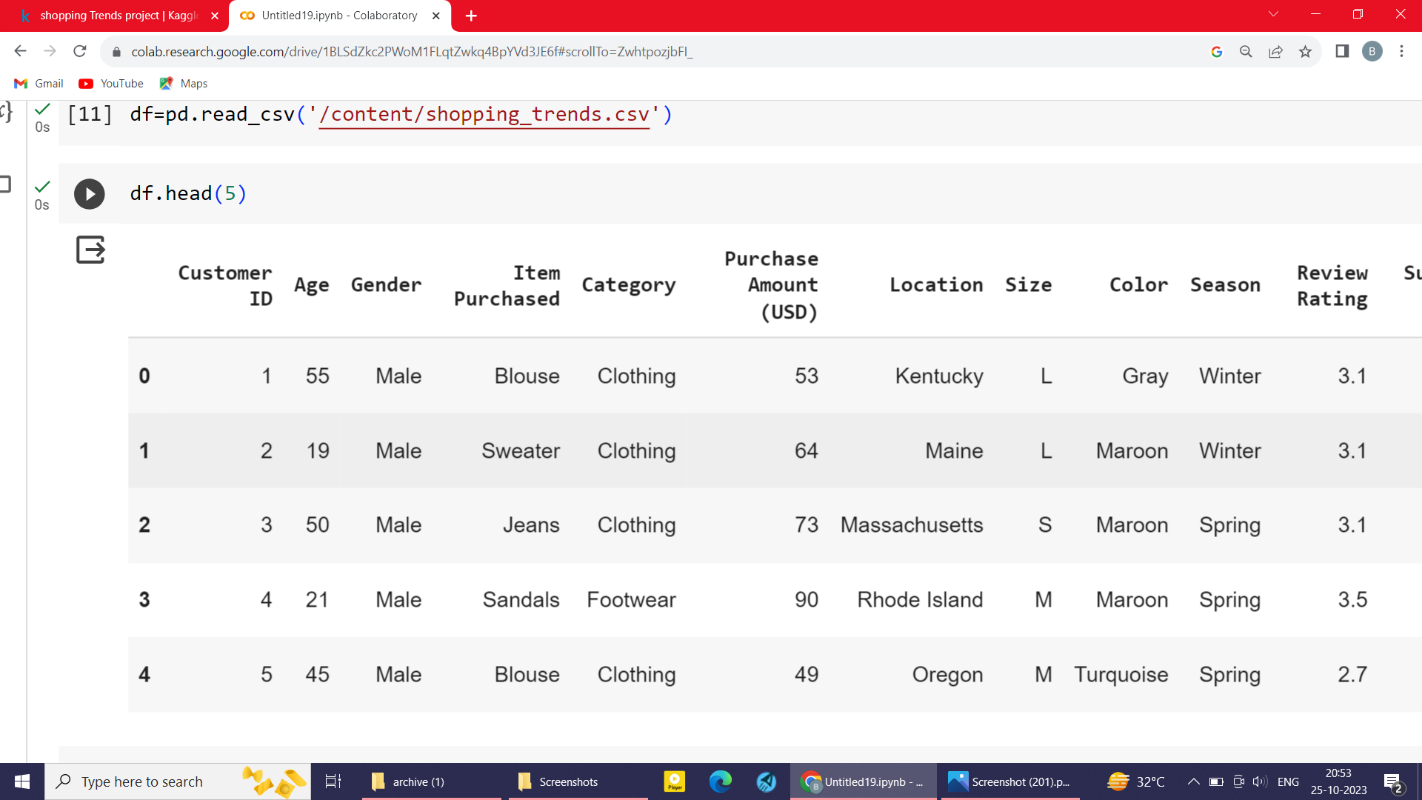
Data transformation in data science refers to the process of converting or modifying data from its original format into a different format or structure to make it more suitable for analysis, modeling, or other data-driven tasks. Data transformation is a crucial step in the data preprocessing phase and can involve various operations to enhance the quality and usability of the data.

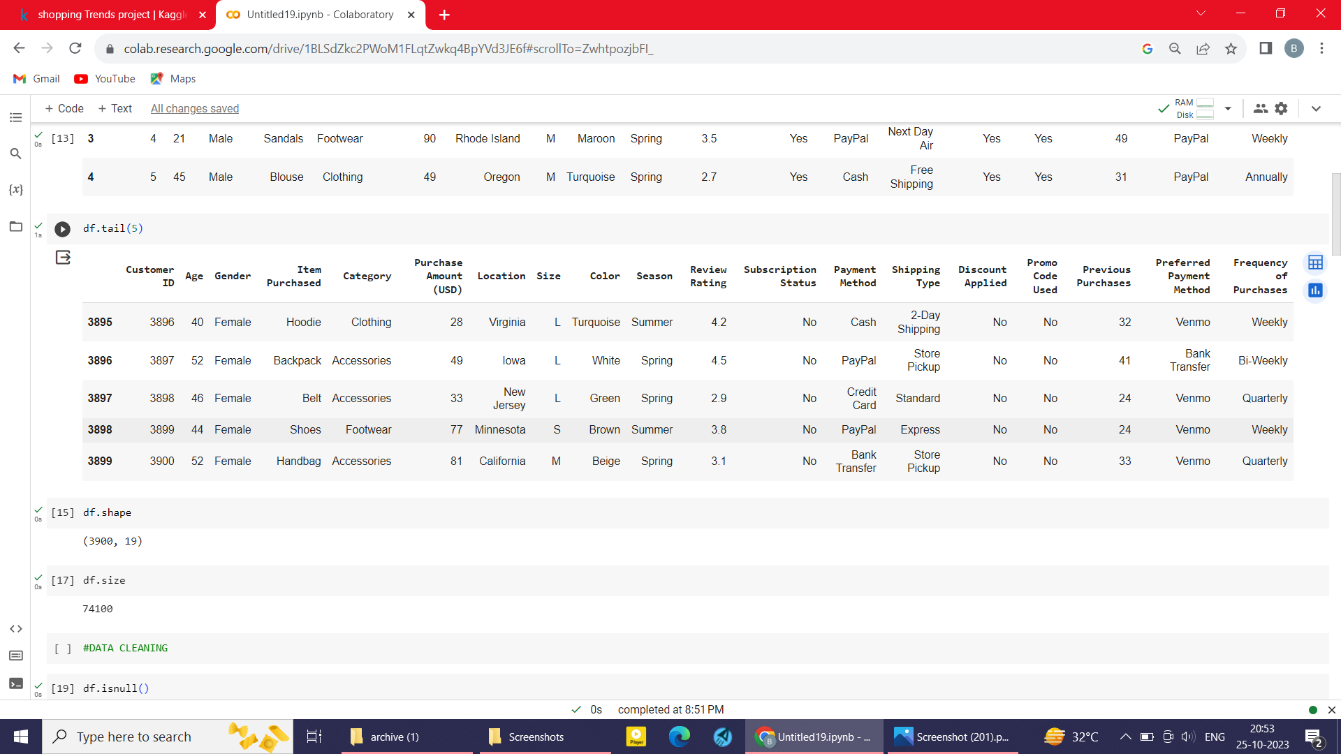
Some common aspects of data transformation include:

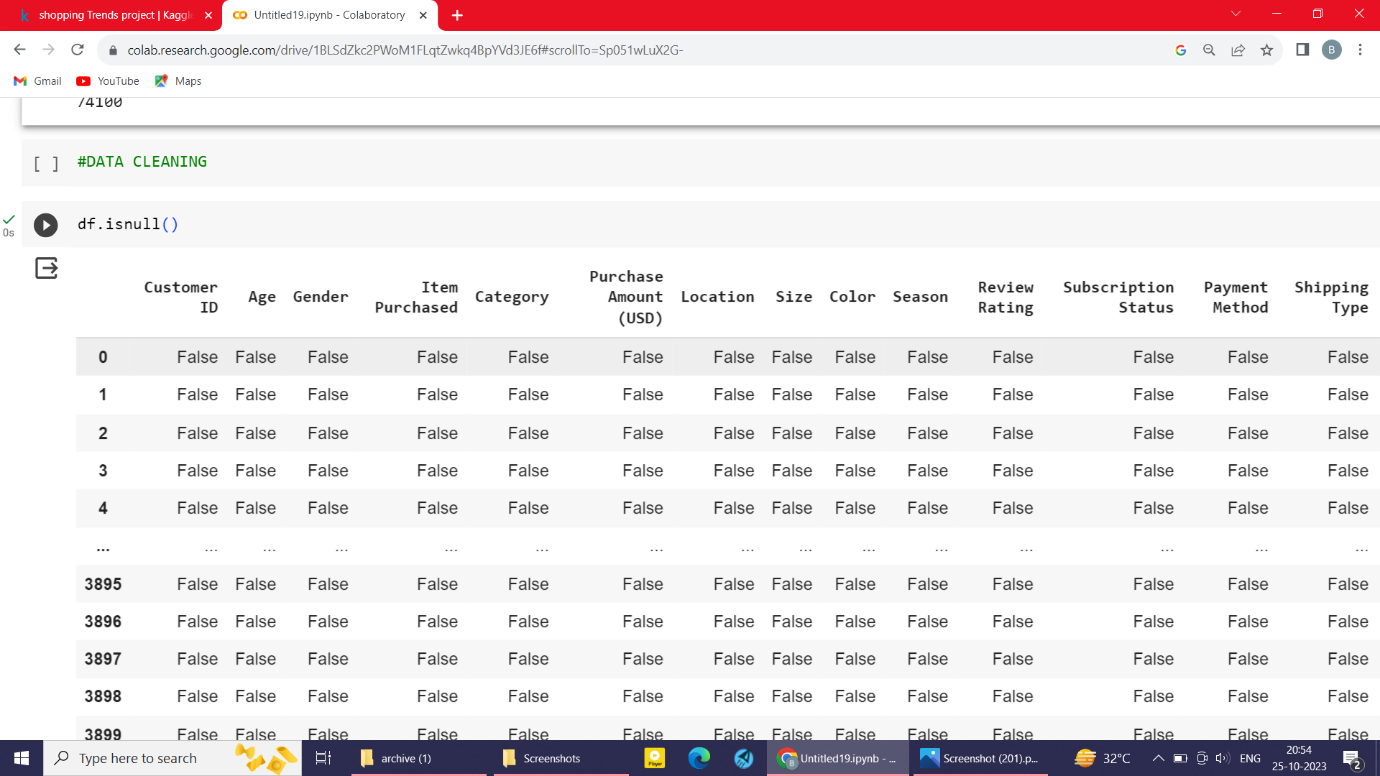
1. Feature Engineering: Feature engineering is a significant part of data transformation. It involves creating new features (variables) from existing data that may better represent the underlying patterns in the data. This can include mathematical transformations, aggregations, or the creation of interaction terms.
2. Scaling and Normalization: Scaling and normalization are techniques used to bring numerical features to a common scale. This is particularly important for machine learning algorithms that are sensitive to the magnitude of input features, such as distance-based algorithms (e.g., k-nearest neighbors) or gradient descent optimization. Common methods include min-max scaling, z-score normalization, and robust scaling.
3. Encoding Categorical Variables: Categorical variables, which represent categories or labels rather than numerical values, need to be encoded into numerical form for many machine learning algorithms. Common techniques include one-hot encoding, label encoding, or target encoding.
4. Data Aggregation: Data may need to be aggregated to a higher level of granularity. For example, you might aggregate daily sales data into monthly or yearly totals for trend analysis.
5. Binning or Discretization: Data can be transformed by grouping continuous values into discrete bins or categories. This can be useful when dealing with numerical data that has a wide range and you want to reduce its complexity.
6. Logarithmic or Exponential Transformations: These transformations are used to adjust the distribution of data, making it more symmetrical or linear. Logarithmic transformation can be useful for data with heavy tails or right-skewed distributions.
7. Datetime Parsing: When working with date and time data, you might need to parse datetime objects into different components (e.g., year, month, day, hour) for more granular analysis.
8. Text Processing: For natural language processing tasks, text data often requires transformations like tokenization (splitting text into words or tokens), stemming, or lemmatization (reducing words to their base form).
9. Handling Imbalanced Data: In classification tasks, where one class may be significantly underrepresented, you can apply techniques like oversampling or undersampling to balance the dataset.
10. Feature Scaling: Scaling numeric features to a similar range to prevent some features from dominating others in machine learning models.
11. Data transformation is highly context-dependent and driven by the specific goals of the analysis or modeling task. The goal is to create a dataset that is more informative, easier to work with, and more appropriate for the chosen data analysis or machine learning methods. These transformations often require domain knowledge and experimentation to determine the most suitable approach for a given dataset and problem.

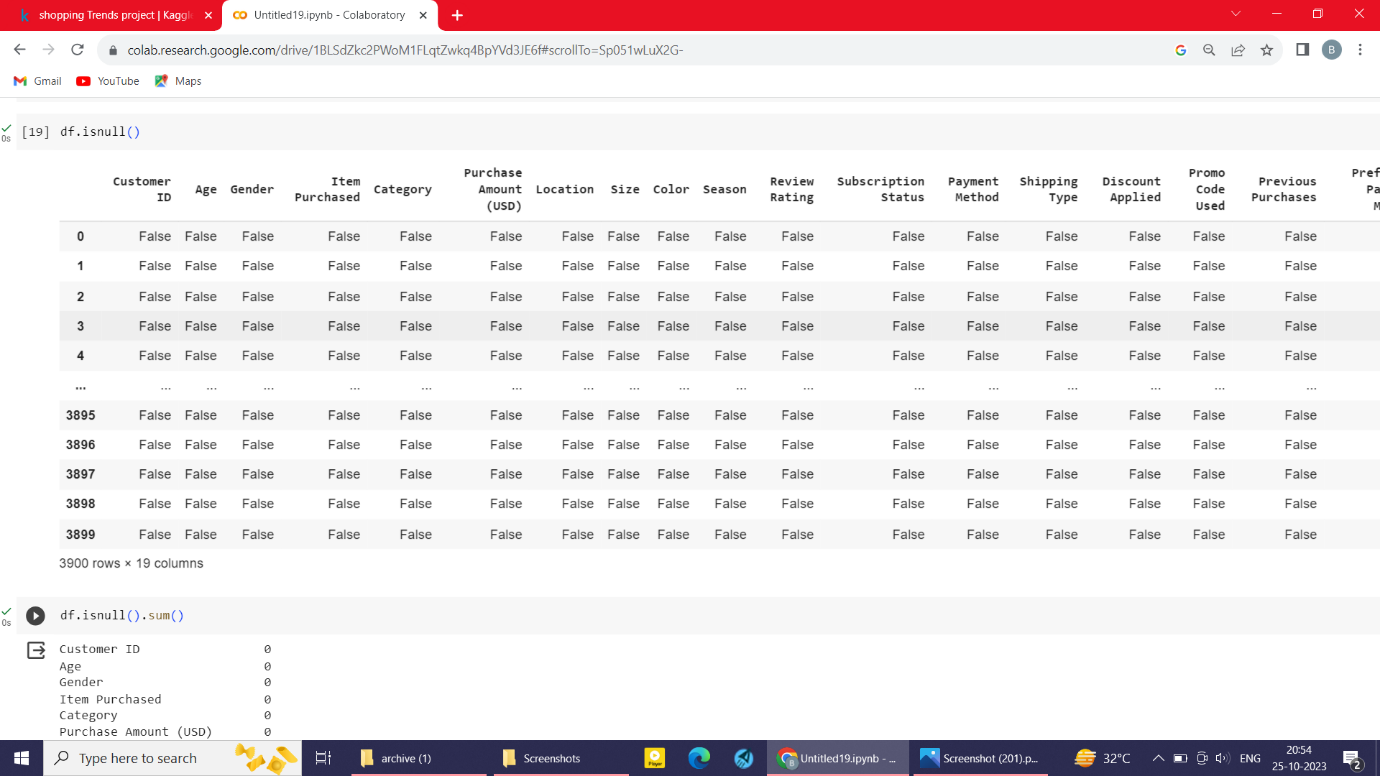
**PRACTICAL IMPLEMENTATIONS OF DATA CLEANING AND DATA TRANSFORMATION BY USING SHOPPING TREANDING DATASET.**

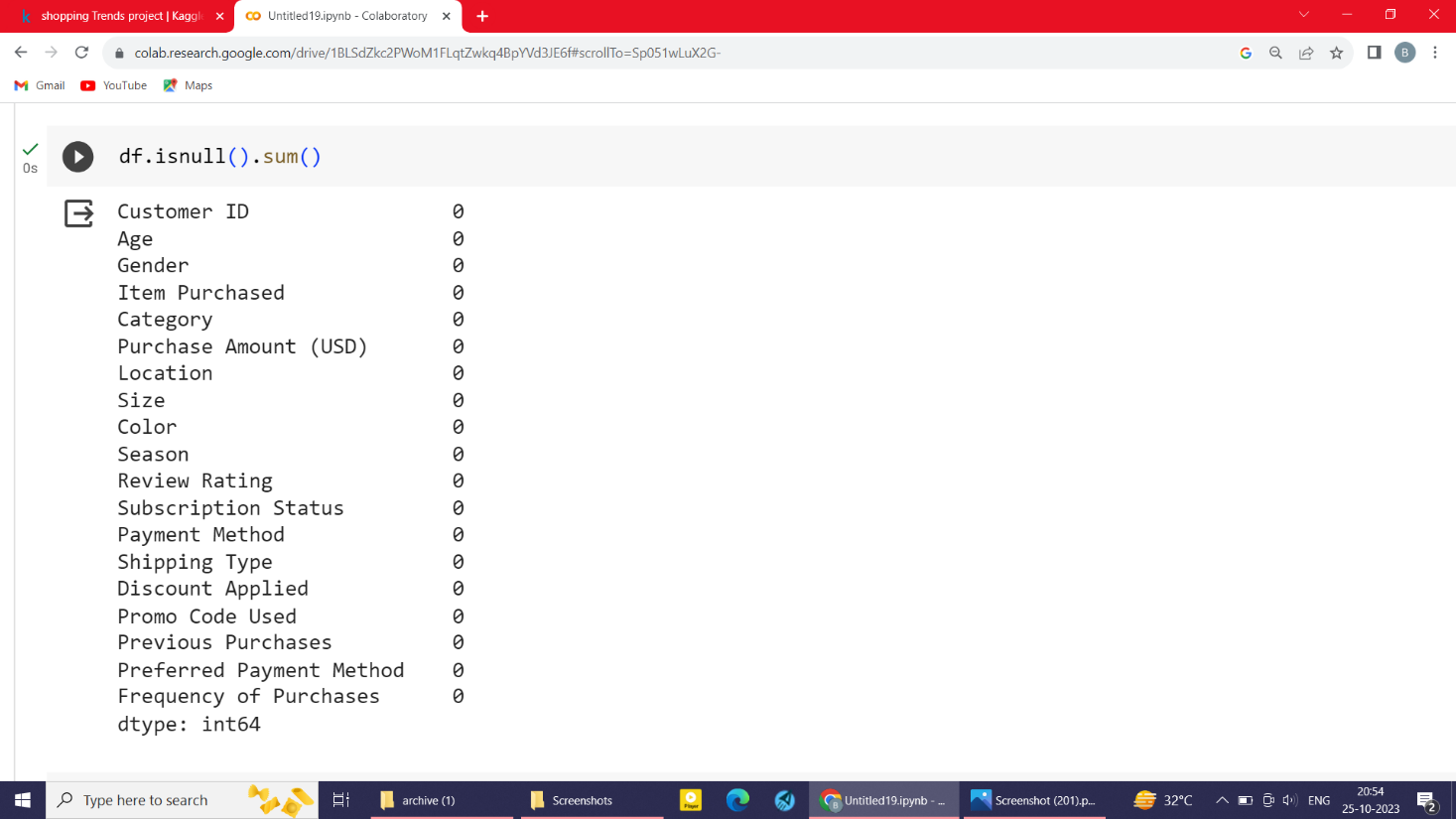




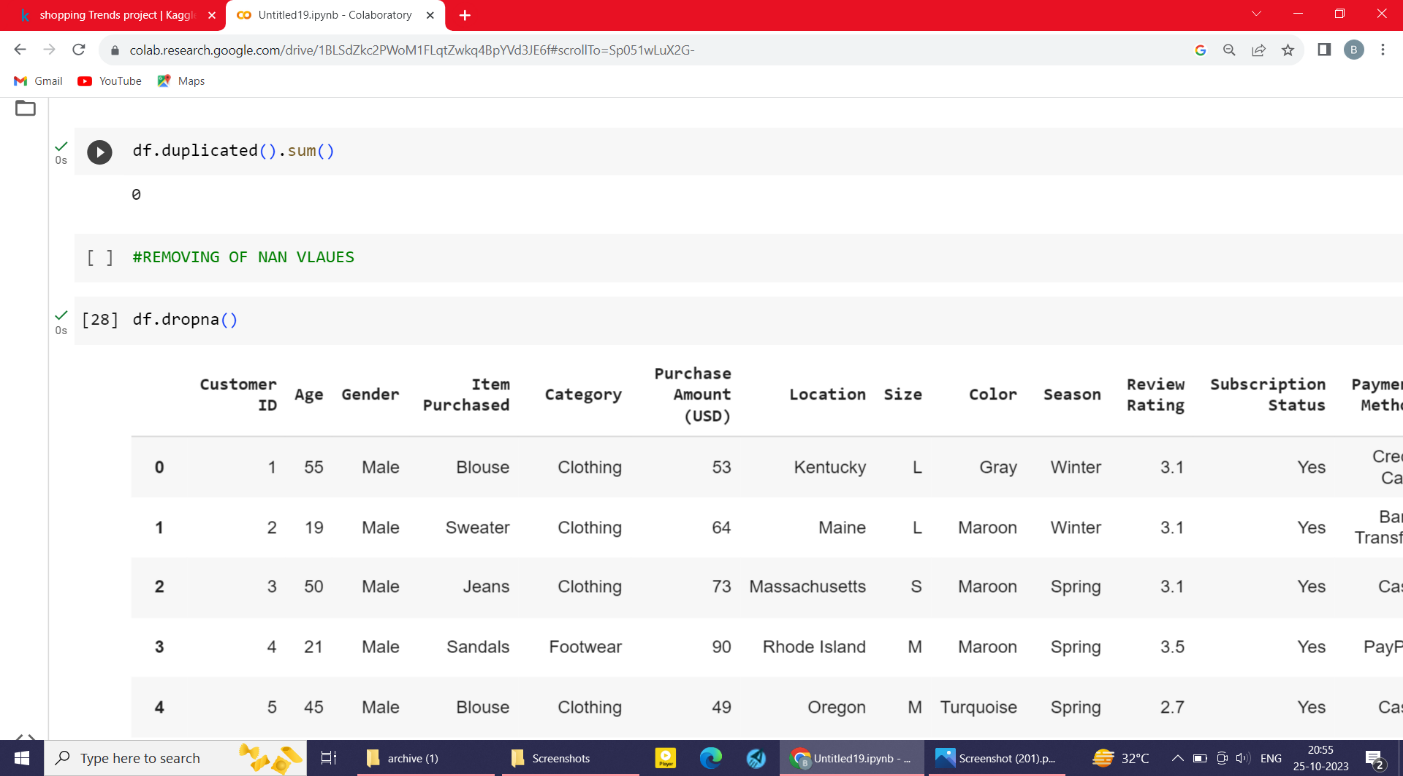


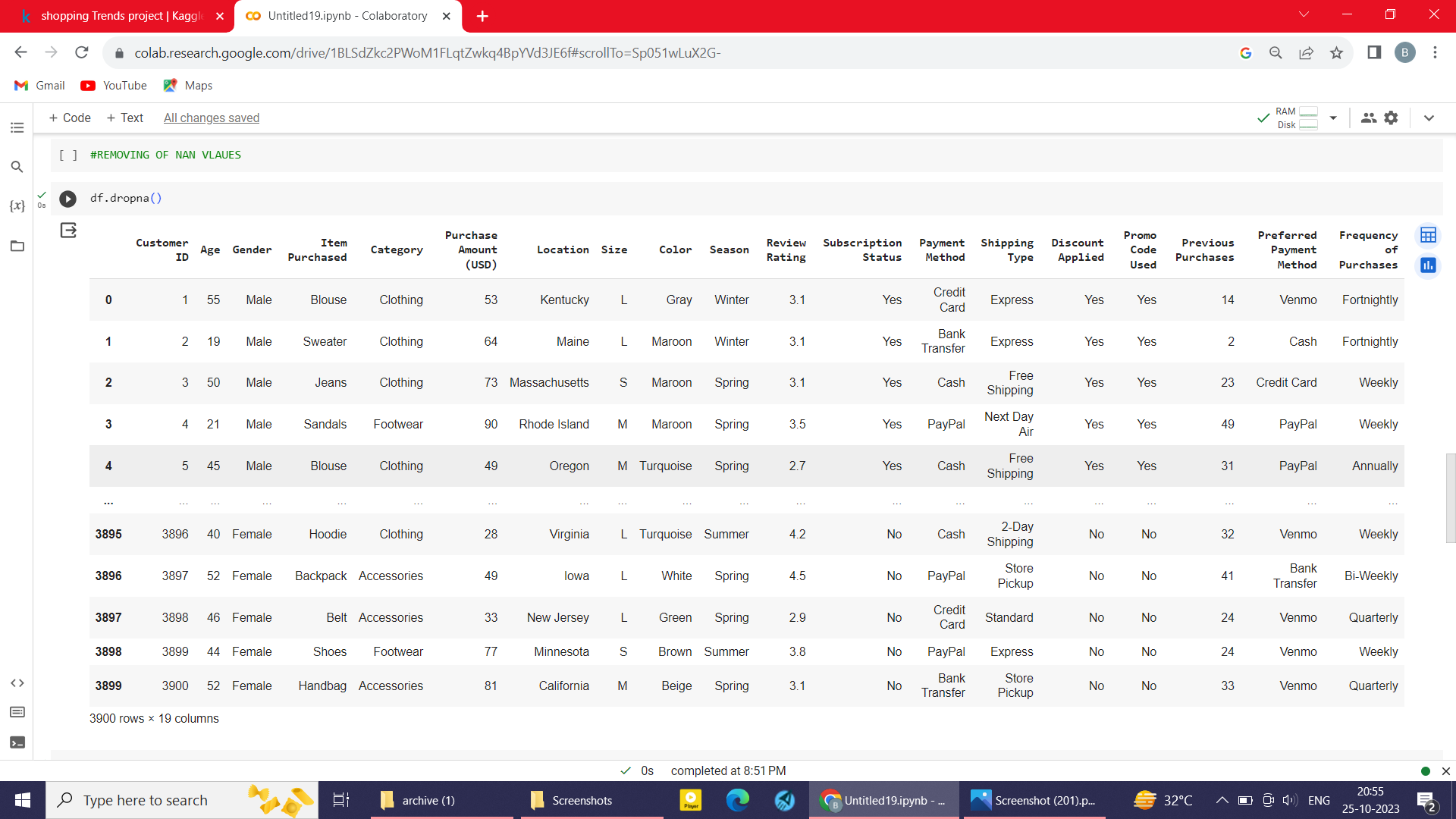


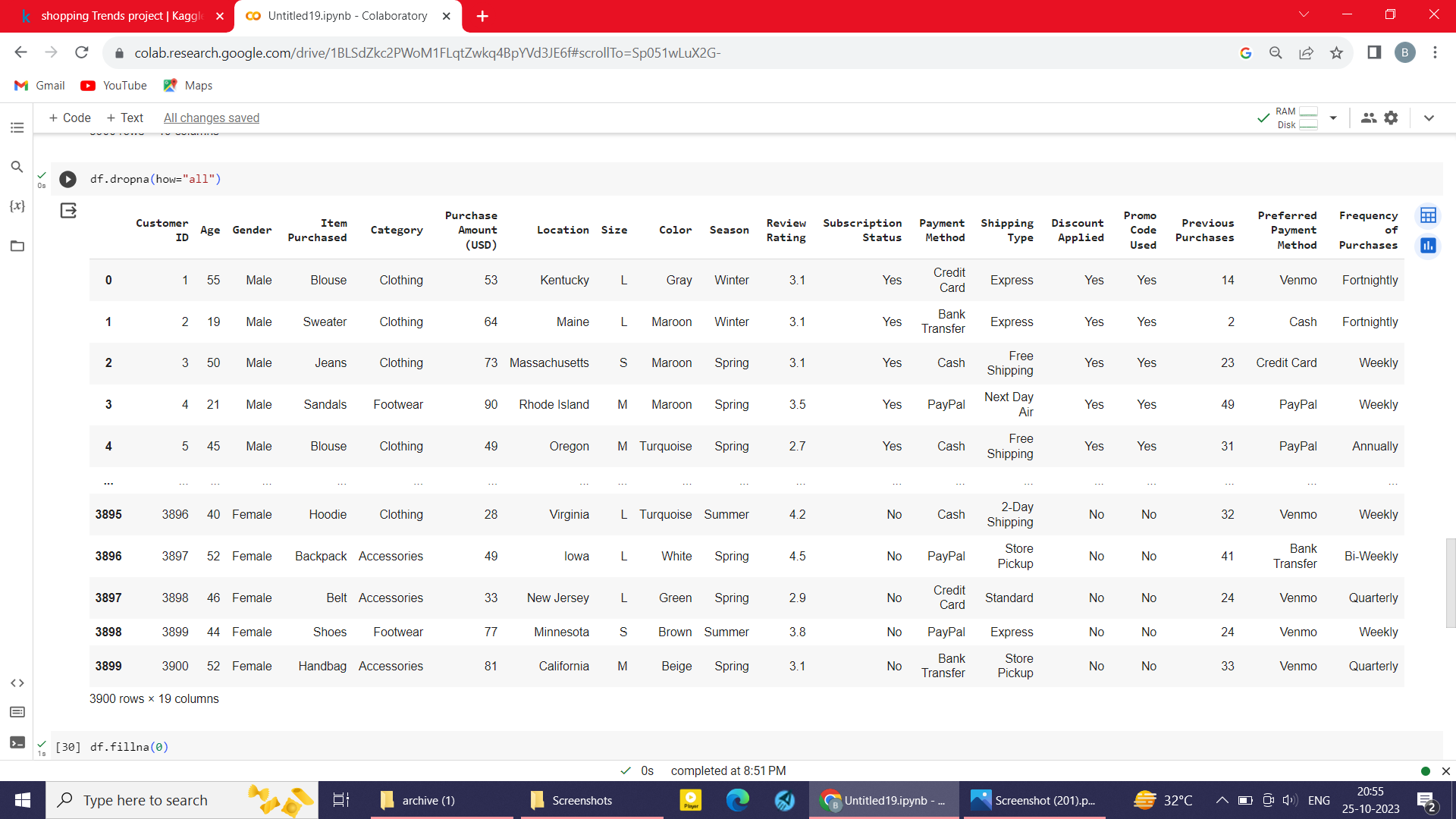


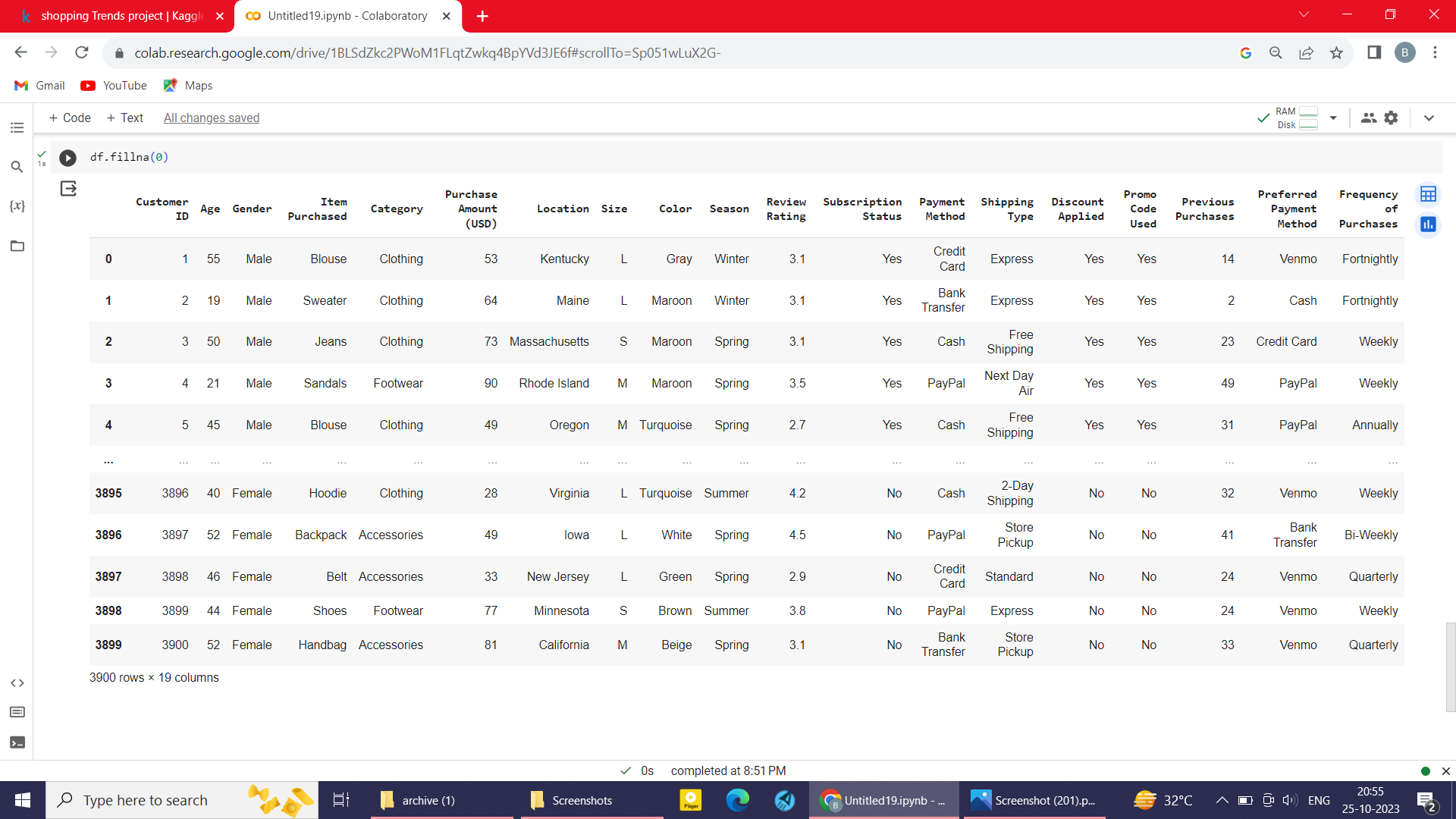


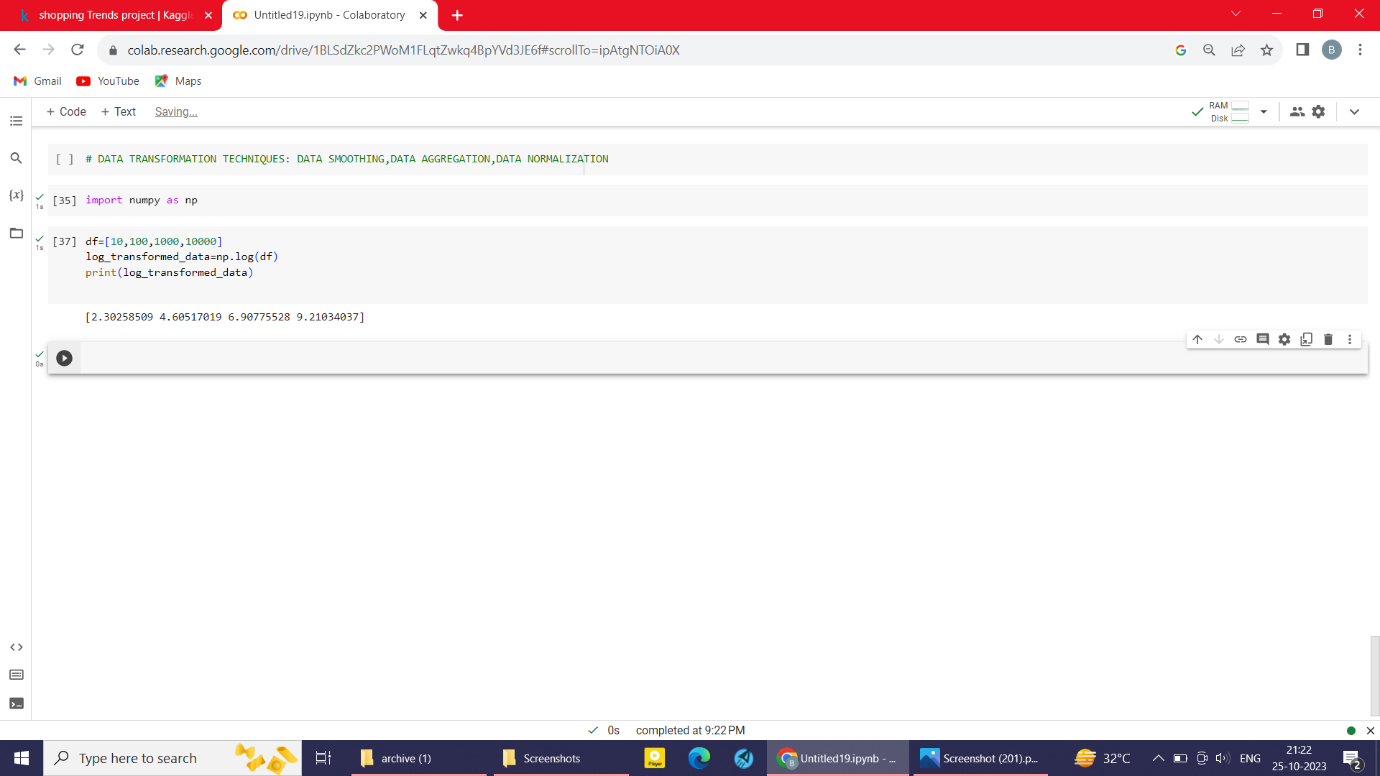












To begin building a big data analysis solution using IBM Cloud Databases, follow these steps:

Create an IBM Cloud account. You can create a free account to get started.

➢ Choose the appropriate database service. IBM Cloud offers a variety of database services, including Db2 and MongoDB. Choose the service that is best suited for your needs.

➢ Set up a database instance. Once you have chosen a database service, you need to set up a database instance. This will involve choosing a region and a plan.

➢ Develop queries or scripts to explore and analyze the selected dataset. Once you have set up a database instance, you can start to develop queries or scripts to explore and analyze the selected dataset. You can use the database console to develop and execute queries and scripts.

➢ Perform basic data cleaning and transformation as needed. Before you can analyze your data, you may need to perform some basic data cleaning and transformation. This may involve removing duplicate records, correcting errors, and transforming the data into a format that is compatible with your chosen analysis tools.

IBM DB2:

Description: IBM Db2 is a family of data management products, including database servers, developed by IBM.

Role in the Project: Used for storing structured data, providing a reliable and scalable database solution. To set up a database instance after choosing the database, you need to follow these steps:

• Create a database instance. This can be done using the database management tool that you are using. For example, to create a database instance in Db2, you would use the CREATE DB command.

• Configure the database instance. This includes setting things like the database name, the database user accounts, and the database parameters.

• Start the database instance. This can be done using the database management tool that you are using. For

example, to start a database instance in Db2, you would use the START DB command. Once the database instance is created, configured, and started, you can start using it to store and manage your data.