**LAB REPORT**

**ON  
  
Get acquainted with data science tools and perform statistical analysis**

**FOUNDATION OF DATA SCIENCE**

**BY**

**AMAN KULUNG**

***PUR080BCT010***



**To   
ER. SUJAN KARKI**

**TRIBHUVAN UNIVERSITY  
INSTITUTE OF ENGINEERING**

**DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING**

**PURWANCHAL CAMPUS  
DHARAN, NEPAL**

**Lab 1**

**1. Get acquainted with data science tools and perform statistical analysis**

* 1. **Objective**
* To familiarize yourself with essential tools and libraries used in data science.
* To learn to describe data using descriptive statistics.  
  1. **Theory**

**Python**Python is a versatile and widely-used programming language for data science due to its simplicity and rich ecosystem of libraries. Its popular libraries include Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, etc. Python is used for data manipulation, visualization, machine learning, and automation.

**Anaconda**

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment.

**Jupyter Note Book**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

**NumPy**

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

NumPy, stands for Numerical Python, is used for the manipulation of elements of numerical array data.

**SciPy**

SciPy is a scientific computation library that uses NumPy underneath. SciPy stands for Scientific Python. It provides more utility functions for optimization, stats and signal processing. Like NumPy, SciPy is open source so we can use it freely. SciPy, stands for Scientific Python, is used for numerical computations in Python. Both these packages provide extended functionality to work with Python.

**Stats models**

Stats models is a popular library in Python that enables us to estimate and analyze various statistical models. It is built on numeric and scientific libraries like NumPy and SciPy. It includes various models of linear regression like ordinary least squares, generalized least squares, weighted least squares, etc

**Pandas**

Pandas are really powerful. They provide you with a huge set of important commands and features which are used to easily analyze your data. We can use Pandas to perform various tasks like filtering your data according to certain conditions, or segmenting and segregating the data according to preference, etc.

**Matplotlib**

Matplotlib is a Python library used for creating static, interactive, and animated visualizations. It provides a variety of plotting options, such as line plots, bar charts, scatter plots, and histograms. Matplotlib is highly customizable and serves as the foundation for other visualization libraries like Seaborn.

**Seaborn**

Seaborn is a Python library built on Matplotlib, designed to simplify the creation of visually appealing and informative statistical graphics. It includes features like heatmaps, violin plots, pair plots, and categorical plots. Seaborn integrates seamlessly with Pandas, making it ideal for exploratory data analysis.

**Scikit-learn**

Scikit-learn is a powerful machine-learning library in Python that provides tools for supervised and unsupervised learning, such as regression, classification, clustering, and dimensionality reduction. It also includes utilities for model selection, preprocessing, and evaluation. Scikit-learn is built on NumPy, SciPy, and Matplotlib.

**Google Colab**

Google Colab is a cloud-based platform for running Python code, offering free GPU and TPU resources. It supports popular data science libraries like TensorFlow, PyTorch, Pandas, and NumPy. Google Colab is commonly used for machine learning, data analysis, and collaborative projects.

**Descriptive statistics**

Descriptive statistics is a branch of statistics that focuses on summarizing and organizing data to uncover patterns, relationships, and trends. It provides a foundation for data analysis by offering methods to describe and present data meaningfully.

**Types of Descriptive statistics:**  
1. **Measures of Central Tendency:** Central tendency provides a single value that represents the center or typical value of a dataset.

* **Mean** (Average): The sum of all data points divided by the total number of points.
* **Median**: The middle value of an ordered dataset. It is less sensitive to outliers.
* **Mode**: The value that appears most frequently in a dataset.

1. **Measures of Dispersion (Spread):** Dispersion measures how spread out the data values are around the central tendency.

* **Range**: The difference between the maximum and minimum values.
* **Variance**: The average squared deviation from the mean.
* **Standard Deviation**: The square root of the variance, showing how data points deviate from the mean.
* **Interquartile Range** (**IQR**): The range of the middle 50% of the data, calculated as Q3 - Q1.

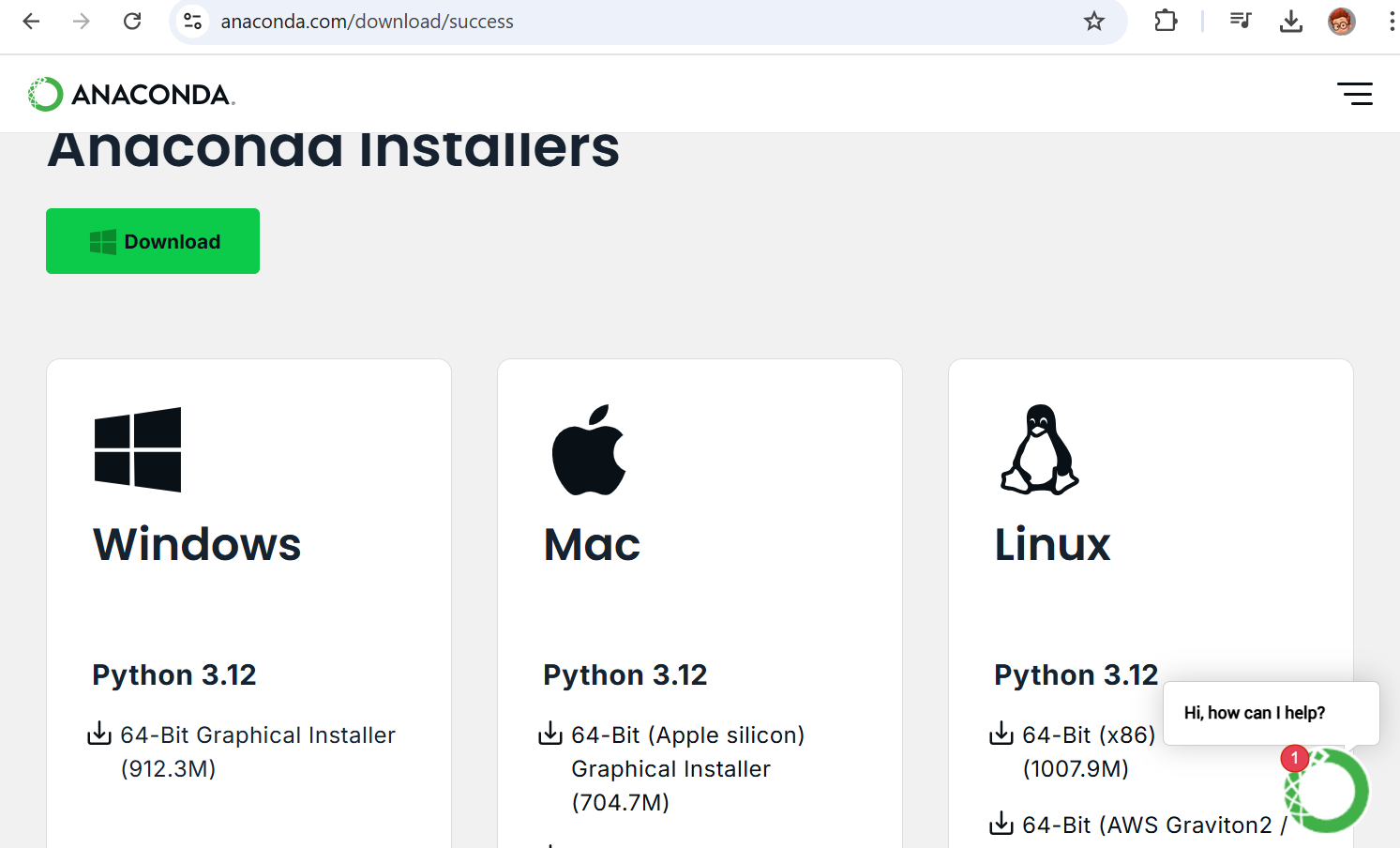
1. **Measures of Shape:** These metrics describe the distribution and symmetry of the data.

* Skewness: Measures asymmetry. Positive skew indicates a longer tail on the right; negative skew indicates a longer tail on the left.
* Kurtosis: Measures the "tailedness" of the distribution. High kurtosis means heavy tails; low kurtosis means light tails.
  1. **Installing Process**

**Download Anaconda:**

**Step1: Type “Anaconda Download” in Google Chrome.Visit the site “http://www.anaconda.com”.**

**Step2: Provide email to get a link for downloading or You can skip registration to download.**

**Step3: Scroll the anaconda products website below. Click the 64 bit installer as per your device.Let the download finish.**

**Step4: Click the downloaded.exe file and install Anaconda.**

**Step5: Agree terms and conditions and Click the Justme option.**

**Step6: Use default path to install in C drive.**

**Step7: Use default options…install it and after installation open the anaconda navigator.**

**Step8: Search for a Jupyter Notebook by scrolling.. If not installed , install it and Launch it.**

**Step9: Jupyter Notebook will be launched on your browser. Then Click on “New” and Python3.**

**Step10: Write your code and run it by shortcut… SHIFT+ENTER.**

**Installing Packages:**

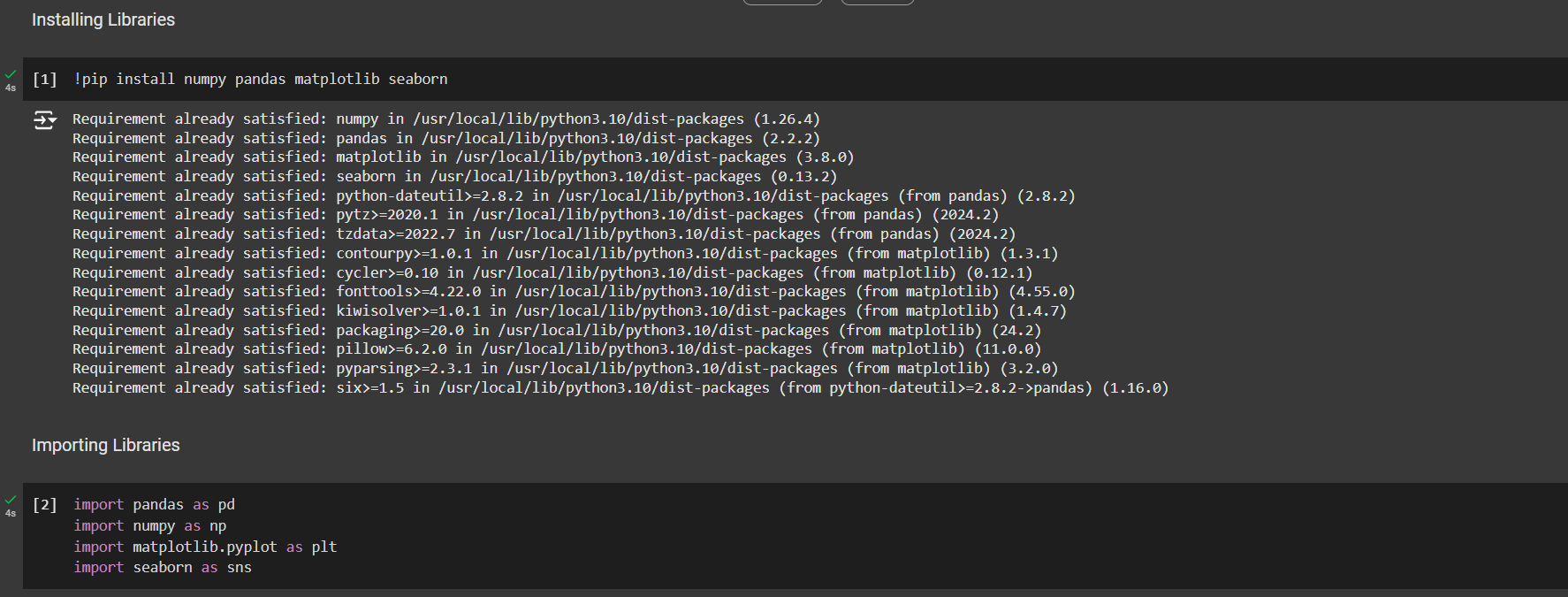
**Step 1: Go to the folder anaconda on your start menu and open “ Anaconda Prompt”.**

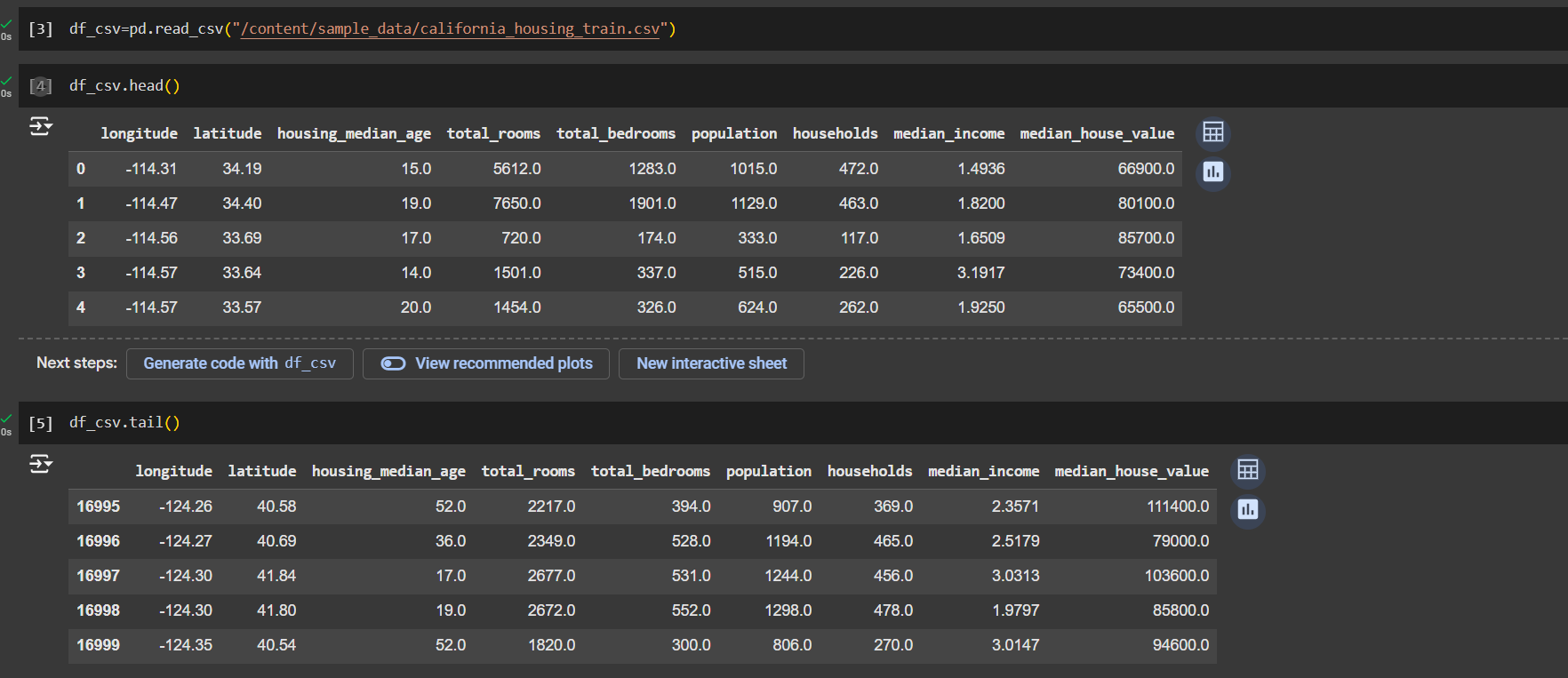
**Step 2: Installing**

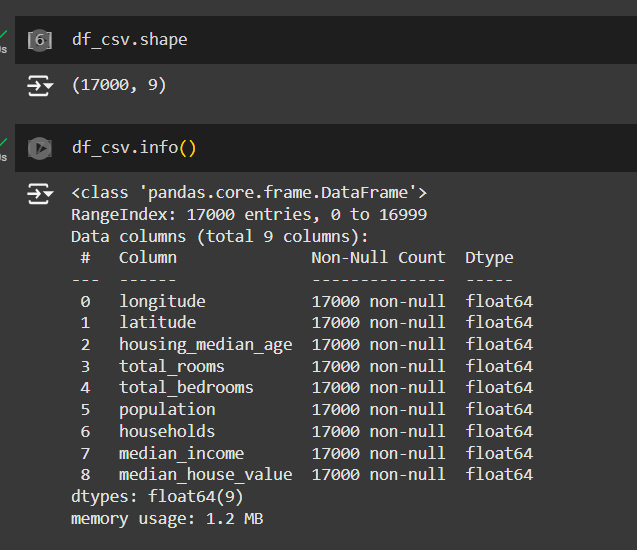
* **Confirm Python is installed correctly, by typing:  
  python -V**
* **Confirm conda is installed correctly, by typing:   
  conda -V**
* **You can install packages using pip or conda.  
  pip install <package\_name>  
  or,  
  conda install <package\_name>**

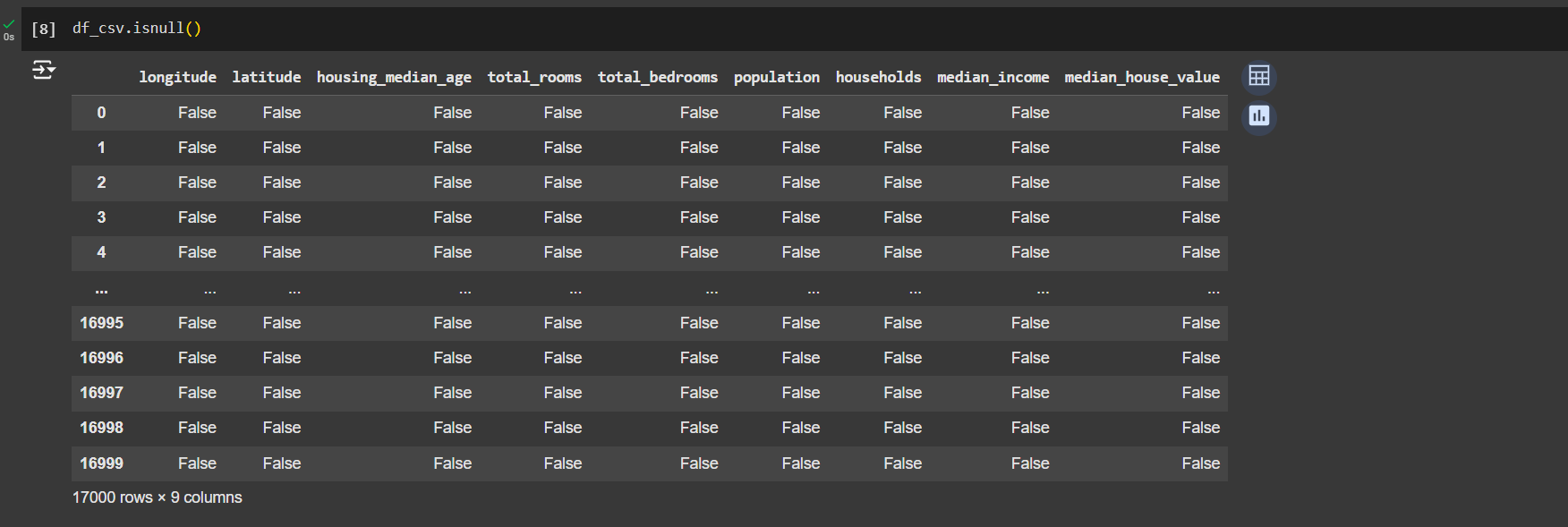
* **For installing a particular version of the package.  
  pip install <package\_name>=1.11.0**
* **For installing multiple packages all at once.  
  pip install <package\_name1> <package\_name2> <package\_name3>**
* **For updating package  
  pip install --upgrade <package\_name>**
* **For installing libraries for data-science  
  pip install numpy scipy statsmodels pandas seaborn matplotlib scikit-learn  
  OR,  
  conda install numpy scipy statsmodels pandas seaborn matplotlib scikit-learn**

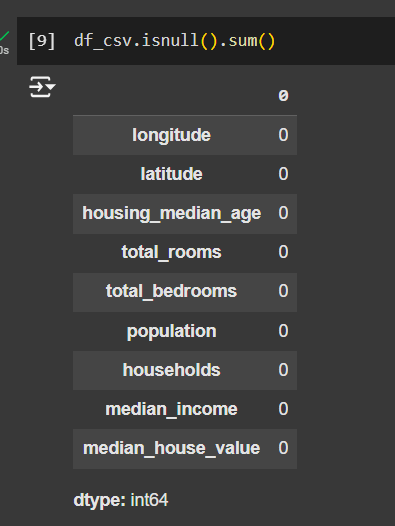
**CODE AND RESULT:**

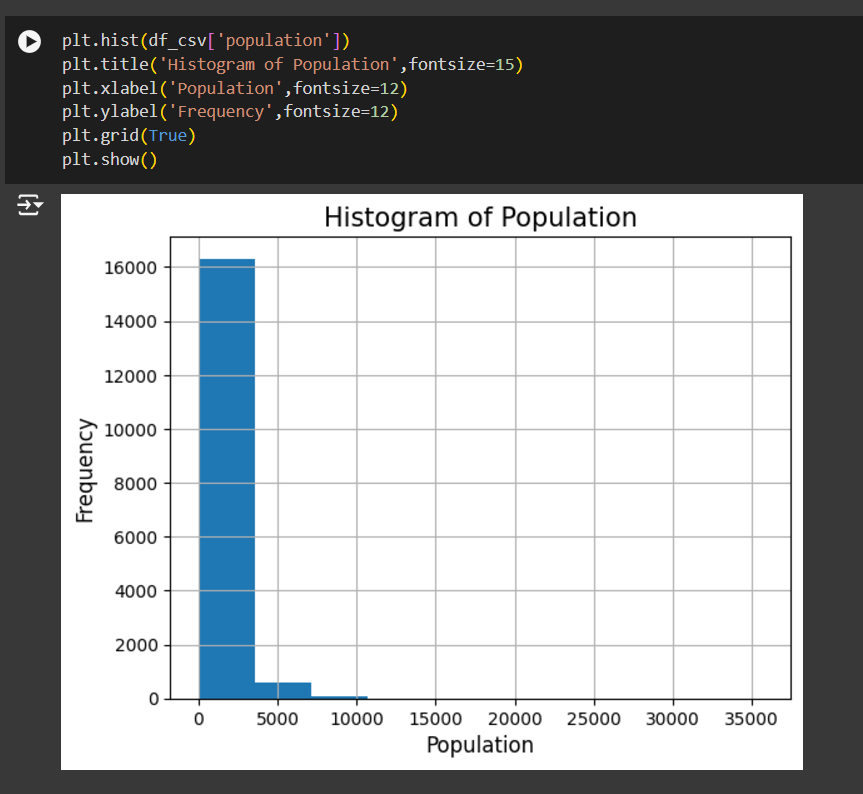
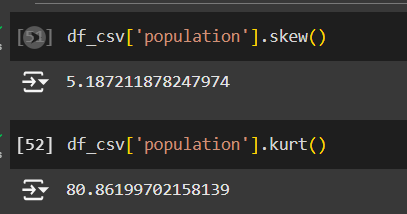
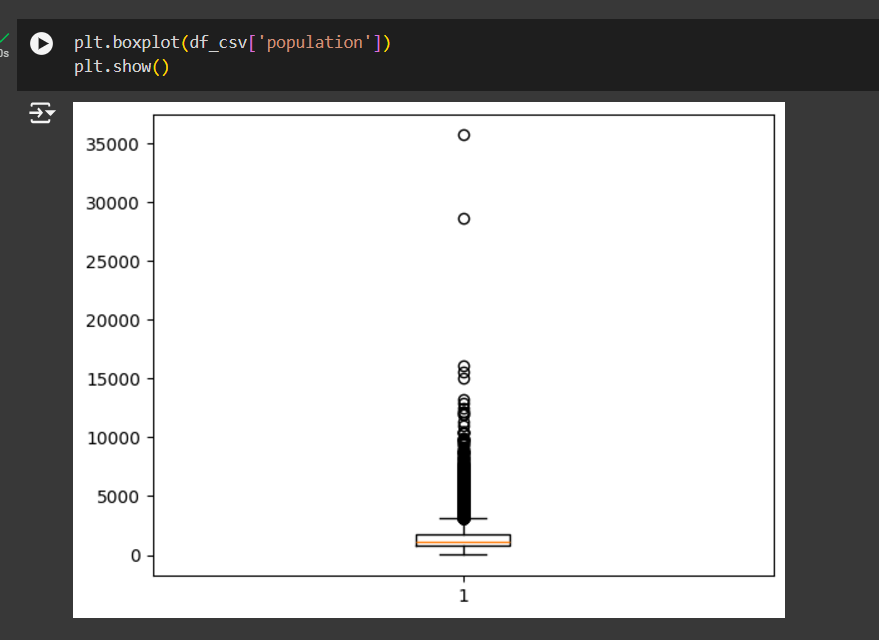
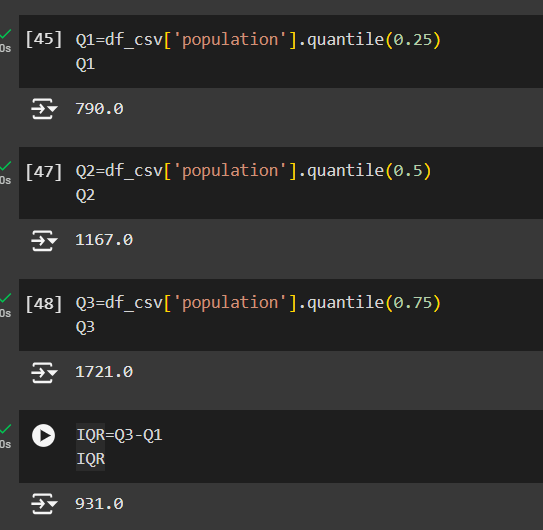
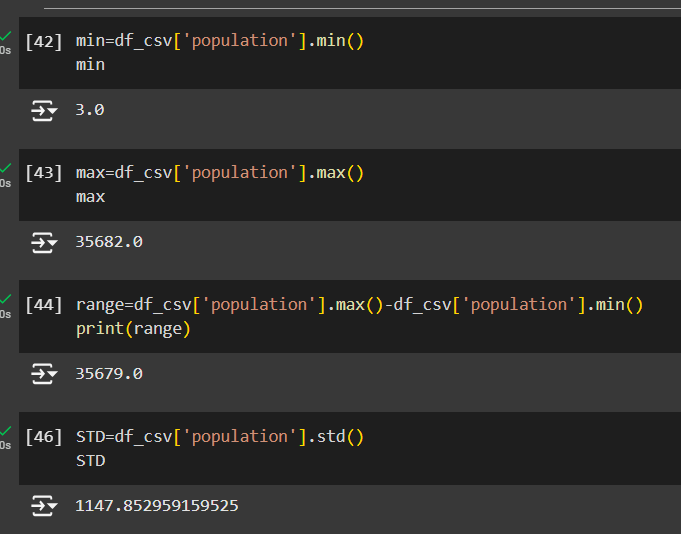
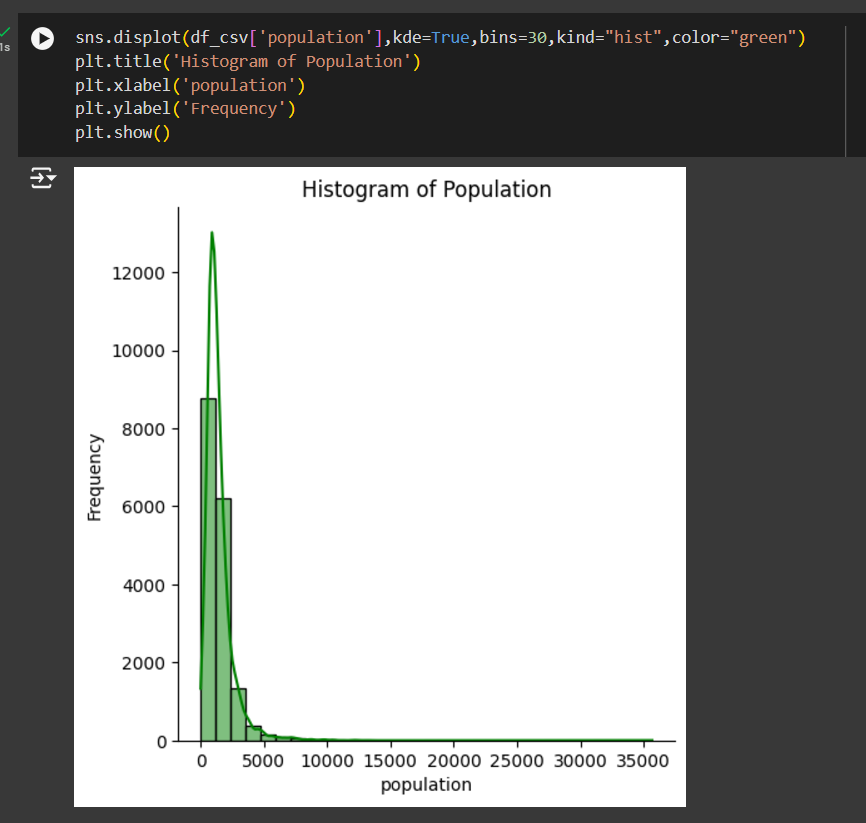
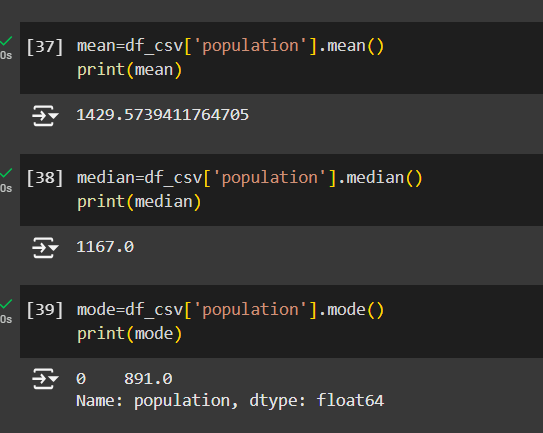
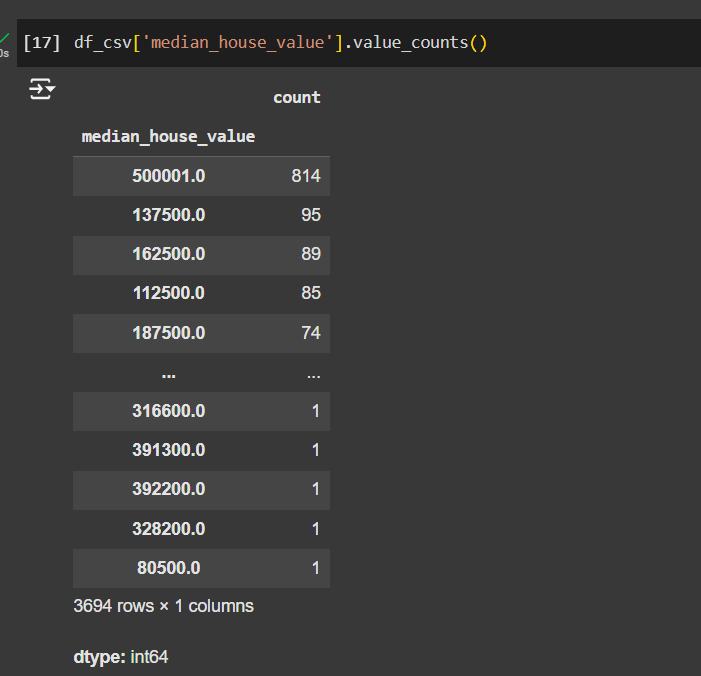
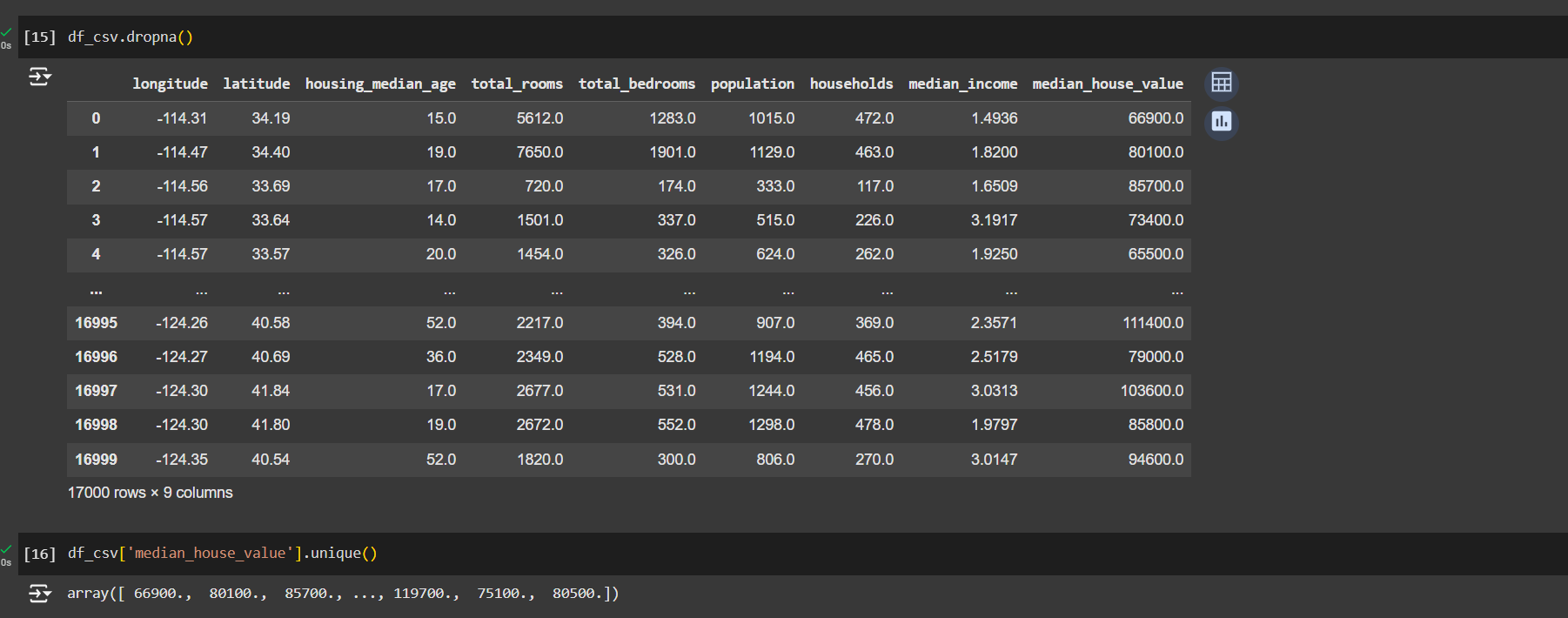
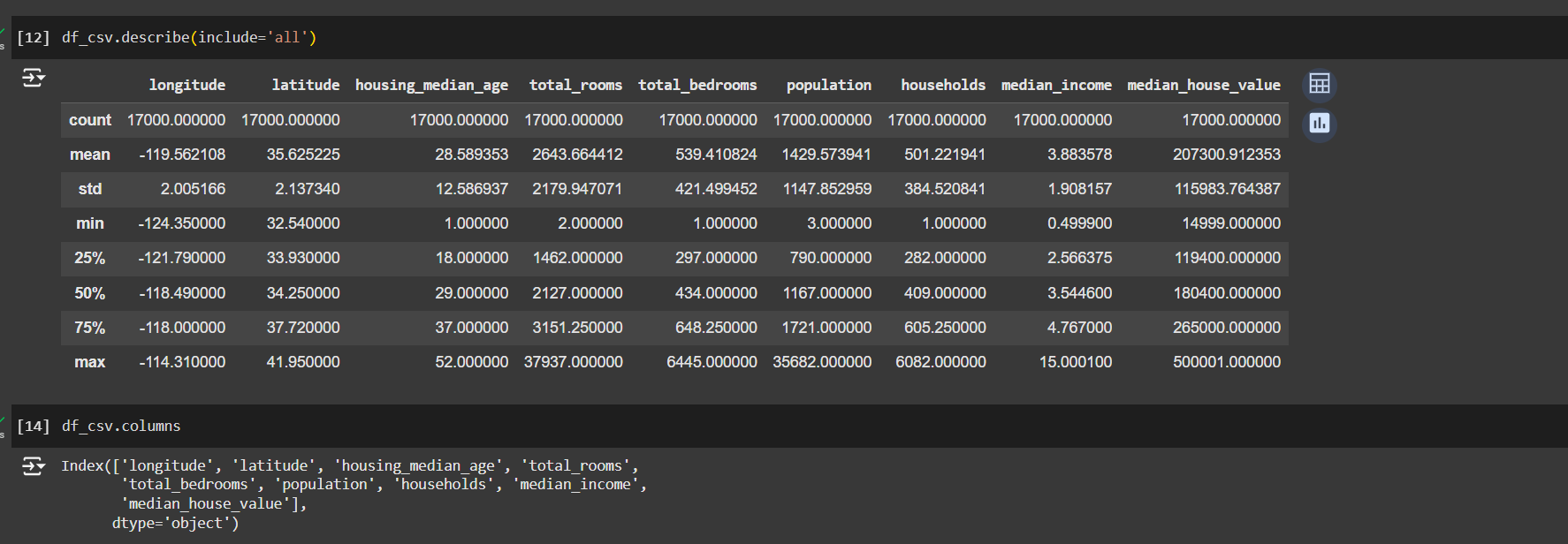
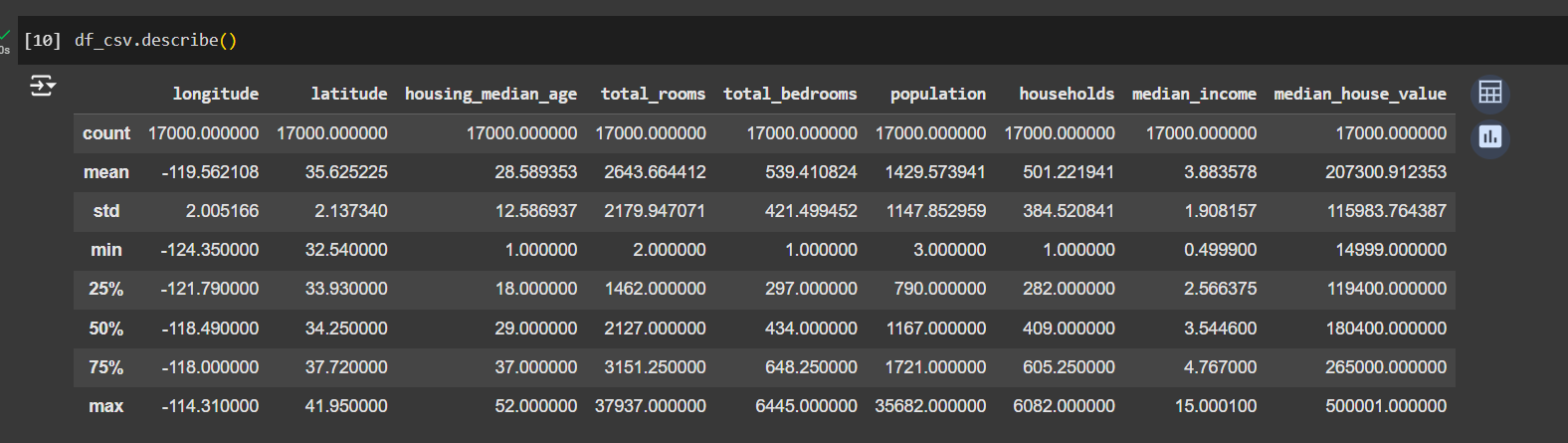
****

****

****

****

****

****

**1.5 Conclusion**

Hence, Anaconda Navigator and Jupyter Notebook, along with Python packages such as NumPy, Pandas, Matplotlib, and Seaborn, have been successfully downloaded and installed. Also, these data science packages were explored in Google Colab to perform descriptive statistical analysis.