**Experiment 9**

**Aim:**

1. Case study on Power BI and Apache Spark.
2. Data Visualization using Power BI, Apache Spark.

**About Dataset:** This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being too toyish and too cumbersome. The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introduction dataset for teaching people about the basics of machine learning.

**Link:-** <https://www.kaggle.com/datasets/camnugent/california-housing-prices>

**Theory:**

**Power Bi:**

Microsoft Power BI is a business intelligence (BI) platform that provides nontechnical business users with tools for aggregating, analyzing, visualizing and sharing data. Power BI's user interface is fairly intuitive for users familiar with Excel, and its deep integration with other Microsoft products makes it a versatile self-service tool that requires little upfront training. Users can download an application for Windows 10, called Power BI Desktop, and native mobile apps for Windows, Android and iOS devices. There is also Power BI Report Server for companies that must maintain their data and reports on premises. That version of Power BI requires a special version of the desktop app -- aptly called Power BI Desktop for Power BI Report Server.

In the Power BI service, you can use the deployment pipeline tool to test your content before you release it to your users. The deployment pipeline tool can help you deploy reports, dashboards, datasets, and paginated reports. Read about how to get started with deployment pipelines in the Power BI service. Apart from these hundreds of data sources from files ( like Excel, PDF, SharePoint Folder, XML), databases (like SQL Server Database, Oracle Database, IBM databases, Amazon Redshift, Google BigQuery), other Power BI databases, Azure data connections, various online services (like Dynamics 365, Salesforce Reports, Google Analytics, Adobe Analytics, Facebook and others) can be connected into Power BI.

**Power BI components**

Microsoft Power BI works by connecting data sources and providing a dashboard of BI to the users. It can connect with just an Excel spreadsheet or bring together cloud-based and on-premises data warehouses. Data pulled from cloud-based sources, such as Salesforce CRM, is automatically refreshed. Included within Power BI are several components that help users create and share data reports. Those are the following:

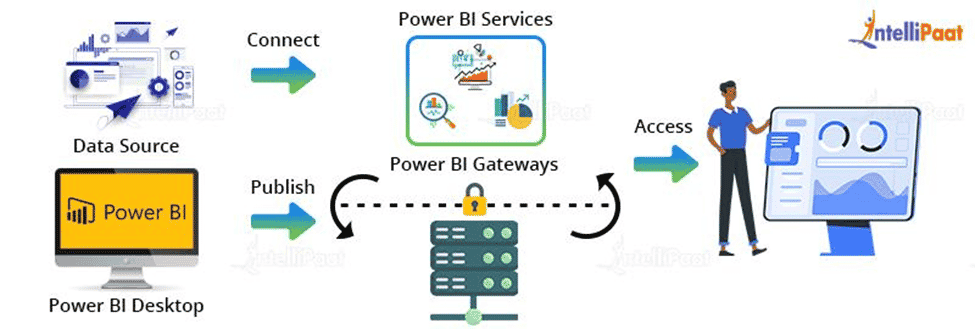
* Power Query: a data mashup and transformation tool
* Power Pivot: a memory tabular data modeling tool
* Power View: a data visualization tool
* Power Map: a 3D geospatial data visualization tool
* Power Q&A: a natural language question and answering engine

**Key features of Power BI**

Microsoft has added a number of data analytics features to Power BI since its inception, and continues to do so. Some of the most important features are the following:

* Artificial intelligence. Users can access image recognition and text analytics in Power BI, create machine learning models using automated ML capabilities and integrate with Azure Machine Learning.
* Hybrid deployment support. This feature provides built-in connectors that allow Power BI tools to connect with a number of different data sources from Microsoft, [Salesforce](https://www.techtarget.com/searchcustomerexperience/definition/Salesforcecom) and other vendors.
* Quick Insights. This feature allows users to create subsets of data and automatically apply analytics to that information.
* Common data model support. Power BI's support for the common data model allows the use of a standardized and extensible collection of data schemas (entities, attributes and relationships).
* Cortana integration. This feature, which is especially popular on mobile devices, allows users to verbally [query](https://www.techtarget.com/searchdatamanagement/definition/query) data using [natural language](https://www.techtarget.com/whatis/definition/natural-language) and access results using [Cortana](https://www.techtarget.com/searchenterprisedesktop/definition/Cortana), Microsoft's [digital assistant](https://www.techtarget.com/searchcustomerexperience/definition/virtual-assistant-AI-assistant).
* Customization. This feature allows developers to change the appearance of default visualization and reporting tools and import new tools into the platform.

**Power BI Architecture**



**Apache Spark:-**

Apache Spark is a lightning-fast, open source data-processing engine for machine learning and AI applications, backed by the largest open source community in big data. Apache Spark (Spark) is an open source data-processing engine for large data sets. It is designed to deliver the computational speed, scalability, and programmability required for Big Data—specifically for streaming data, graph data, machine learning, and artificial intelligence (AI) applications.

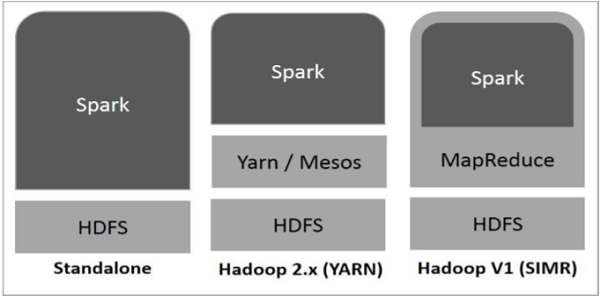
Spark's analytics engine processes data 10 to 100 times faster than alternatives. It scales by distributing processing work across large clusters of computers, with built-in parallelism and fault tolerance. It even includes APIs for programming languages that are popular among data analysts and data scientists, including Scala, Java, Python, and R. Spark is often compared to Apache Hadoop, and specifically to MapReduce, Hadoop’s native data-processing component. The chief difference between Spark and MapReduce is that Spark processes and keeps the data in memory for subsequent steps—without writing to or reading from disk—which results in dramatically faster processing speeds.

**Features of Apache Spark**

* **Fast -** It provides high performance for both batch and streaming data, using a state-of-the-art DAG scheduler, a query optimizer, and a physical execution engine.
* **Easy to Use -** It facilitates writing applications in Java, Scala, Python, R, and SQL. It also provides more than 80 high-level operators.
* **Generality -** It provides a collection of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming.
* **Lightweight -** It is a light unified analytics engine which is used for large scale data processing.
* **Runs Everywhere -** It can easily run on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud.

**Spark Built on Hadoop**

The following diagram shows three ways of how Spark can be built with Hadoop components.



There are three ways of Spark deployment as explained below.

* **Standalone −** Spark Standalone deployment means Spark occupies the place on top of HDFS(Hadoop Distributed File System) and space is allocated for HDFS, explicitly. Here, Spark and MapReduce will run side by side to cover all spark jobs on cluster.
* **Hadoop Yarn −** Hadoop Yarn deployment means, simply, spark runs on Yarn without any pre-installation or root access required. It helps to integrate Spark into Hadoop ecosystem or Hadoop stack. It allows other components to run on top of stack.
* **Spark in MapReduce (SIMR) −** Spark in MapReduce is used to launch spark job in addition to standalone deployment. With SIMR, user can start Spark and uses its shell without any administrative access.

**Usage of Spark**

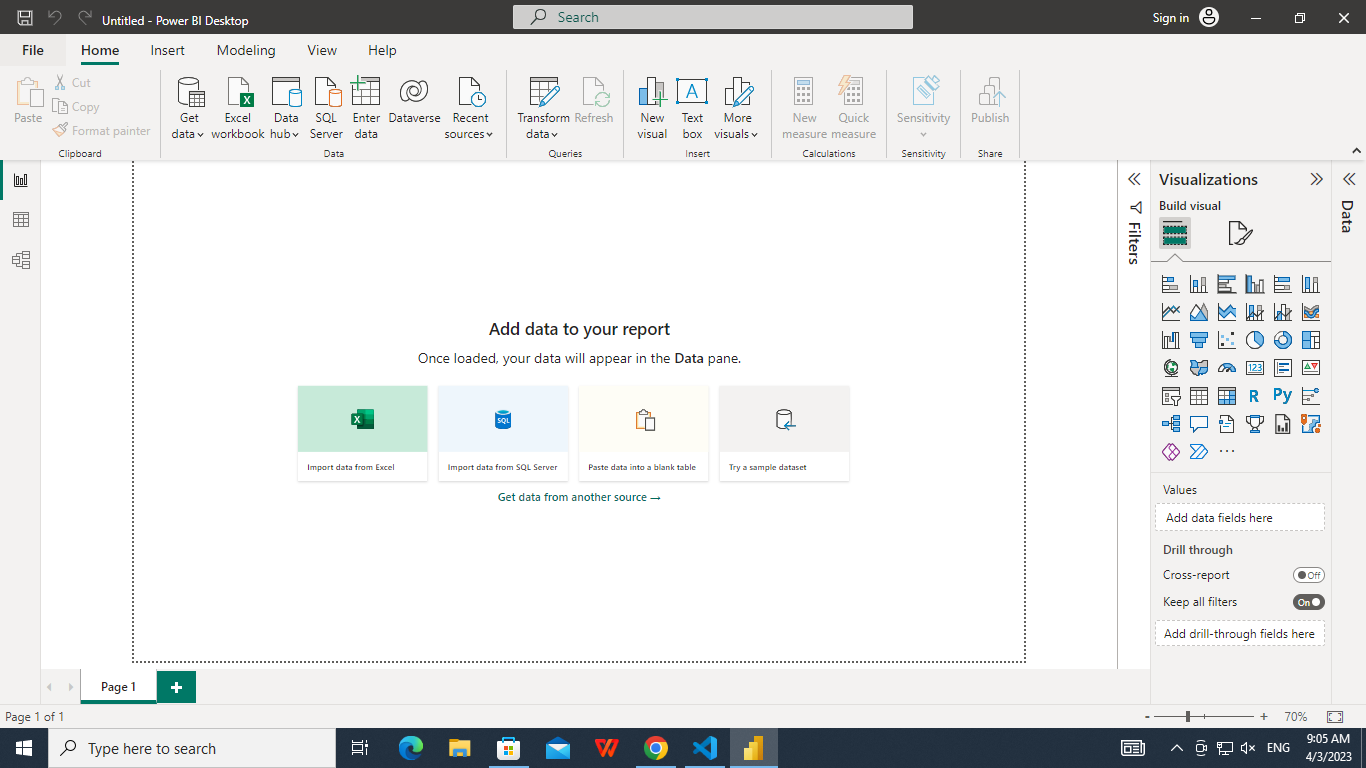
* **Data integration:** The data generated by systems are not consistent enough to combine for analysis. To fetch consistent data from systems we can use processes like Extract, transform, and load (ETL). Spark is used to reduce the cost and time required for this ETL process.
* **Stream processing:** It is always difficult to handle the real-time generated data such as log files. Spark is capable enough to operate streams of data and refuses potentially fraudulent operations.
* **Machine learning:** Machine learning approaches become more feasible and increasingly accurate due to enhancement in the volume of data. As spark is capable of storing data in memory and can run repeated queries quickly, it makes it easy to work on machine learning algorithms.
* **Interactive analytics:** Spark is able to generate responses rapidly. So, instead of running pre-defined queries, we can handle the data interactively.

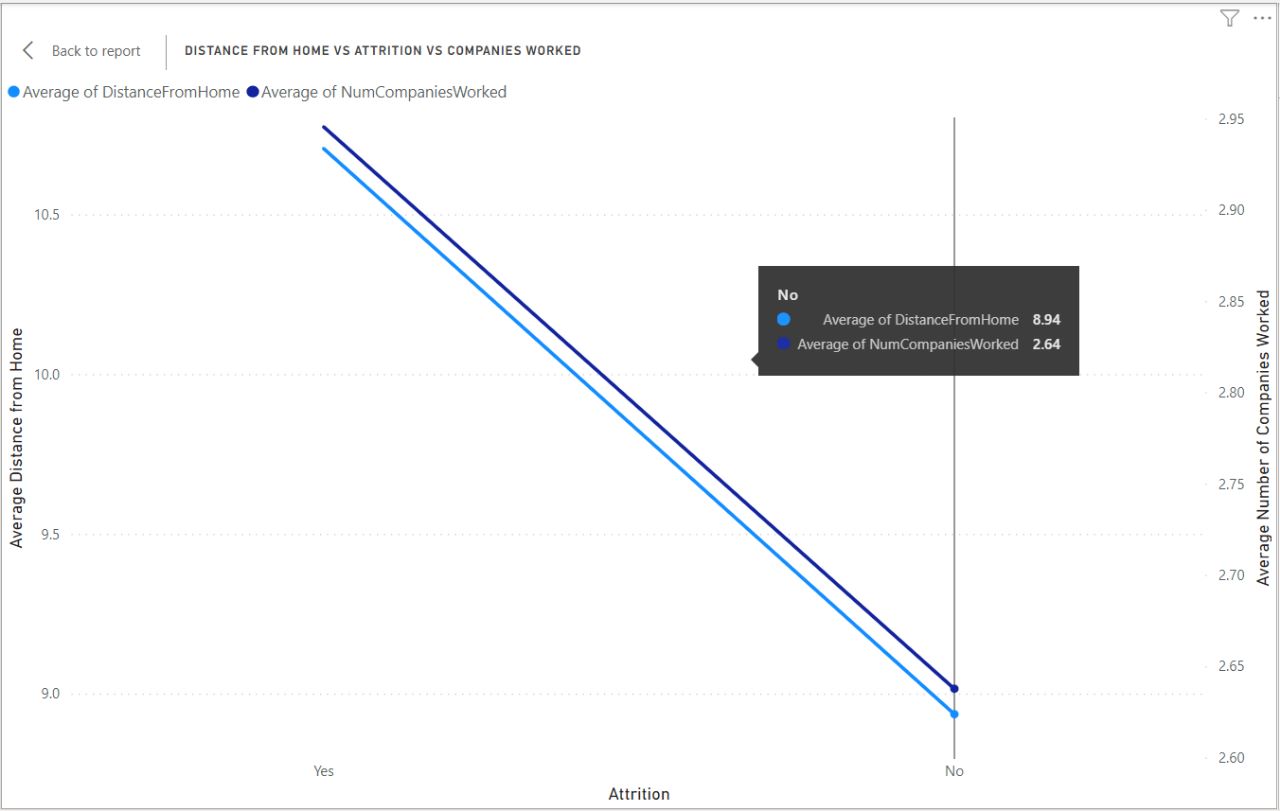
**What are the benefits of Apache Spark?**

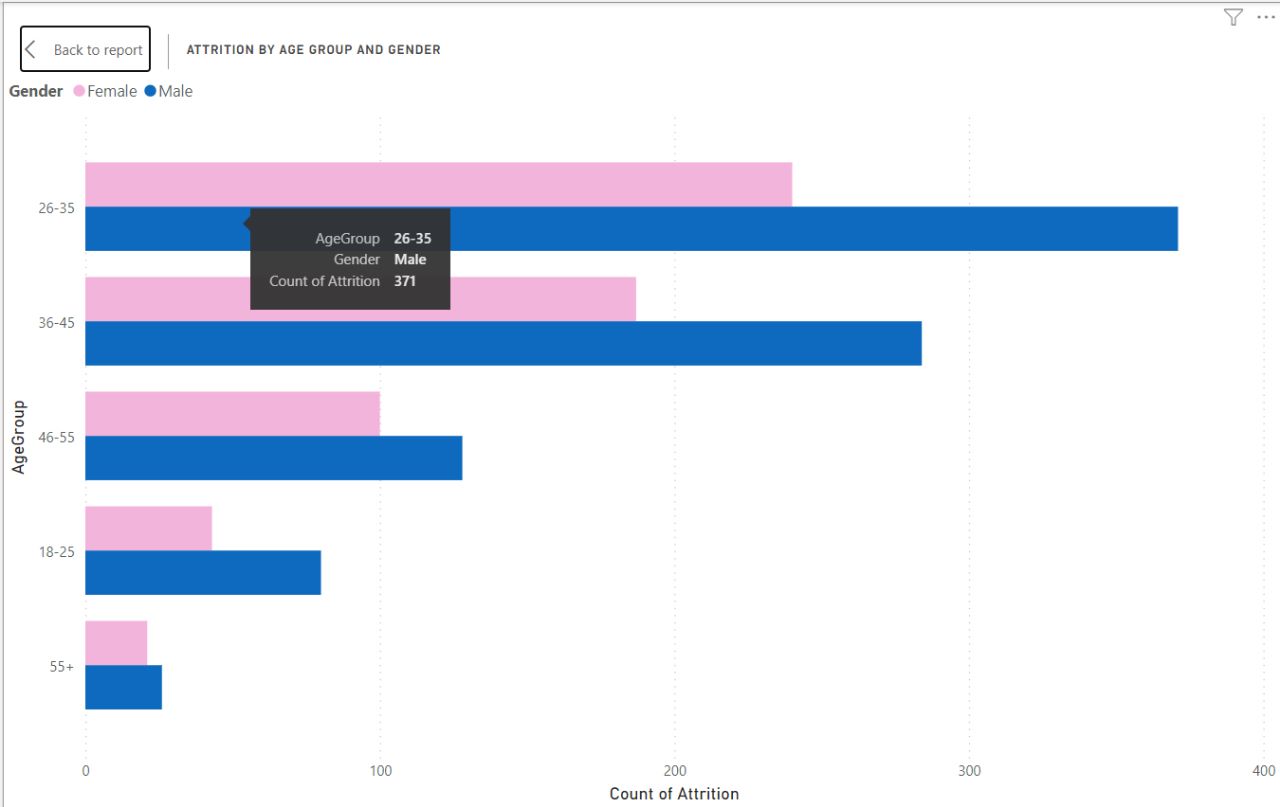
* **Fast:-** Through in-memory caching, and optimized query execution, Spark can run fast analytic queries against data of any size.
* **Developer friendly:-** Apache Spark natively supports Java, Scala, R, and Python, giving you a variety of languages for building your applications. These APIs make it easy for your developers, because they hide the complexity of distributed processing behind simple, high-level operators that dramatically lowers the amount of code required.
* **Multiple workloads:-** Apache Spark comes with the ability to run multiple workloads, including interactive queries, real-time analytics, machine learning, and graph processing. One application can combine multiple workloads seamlessly.

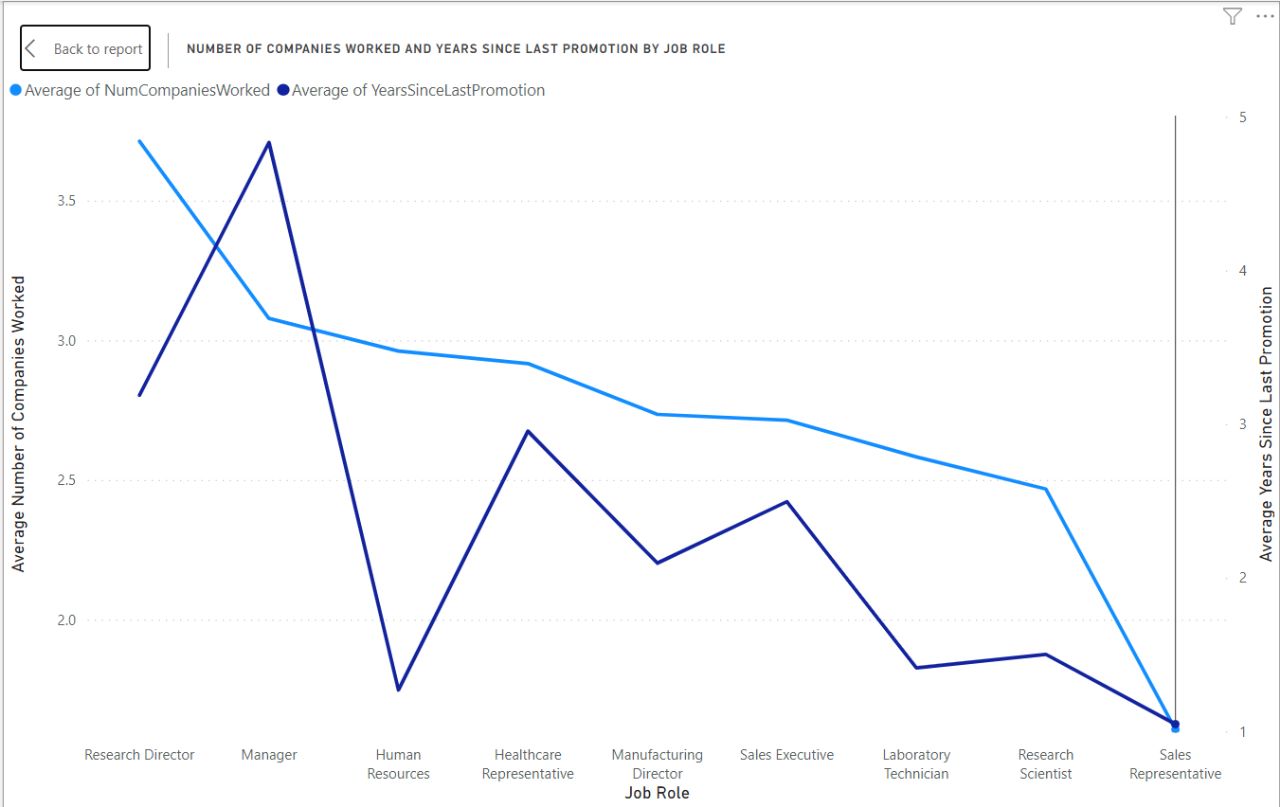
**Data Visualization using Power Bi:**

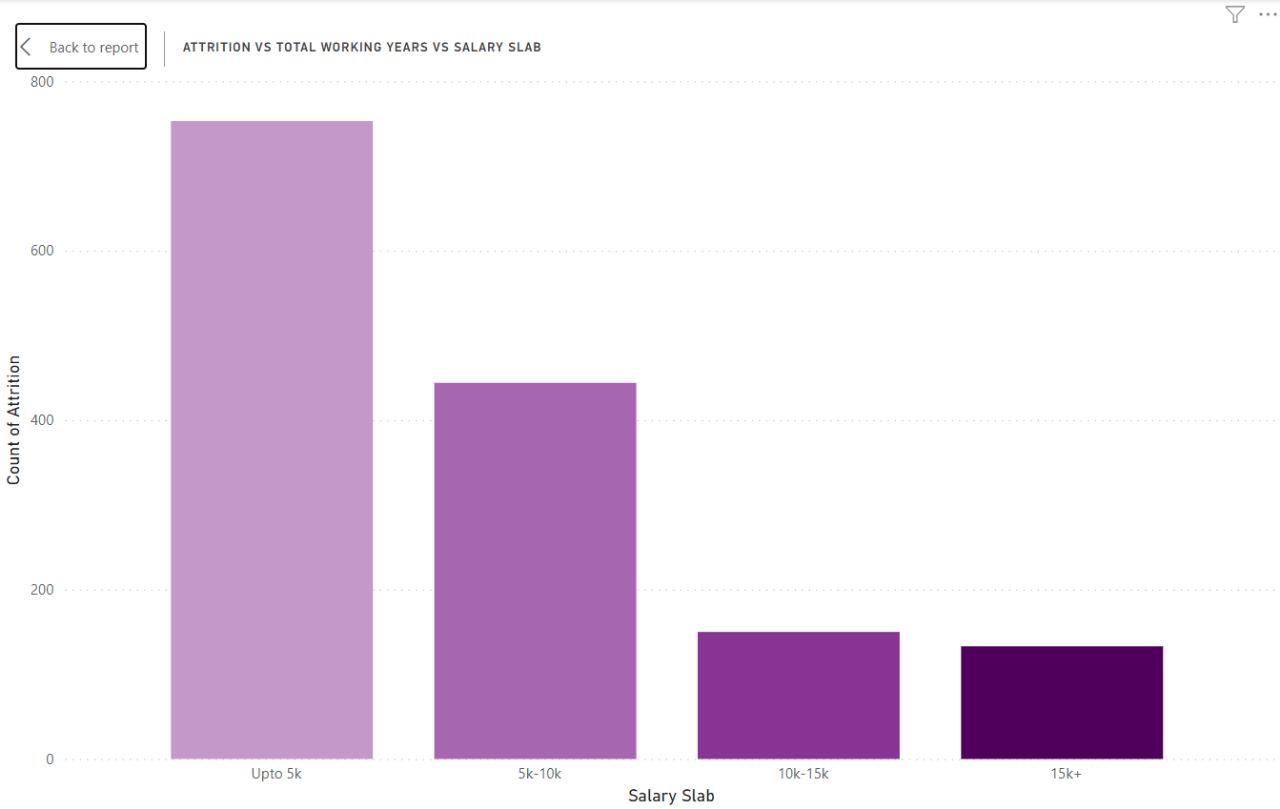
Importing dataset into Power Bi workspace.

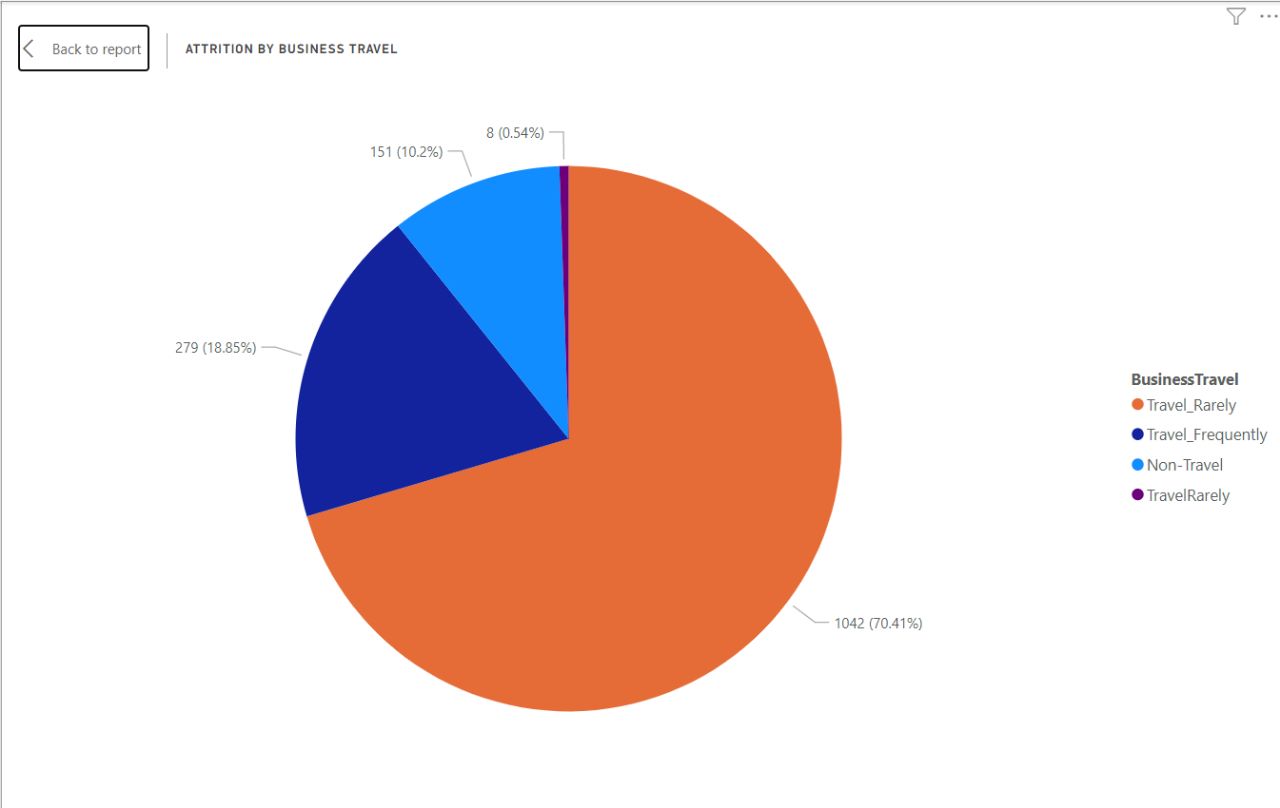
****

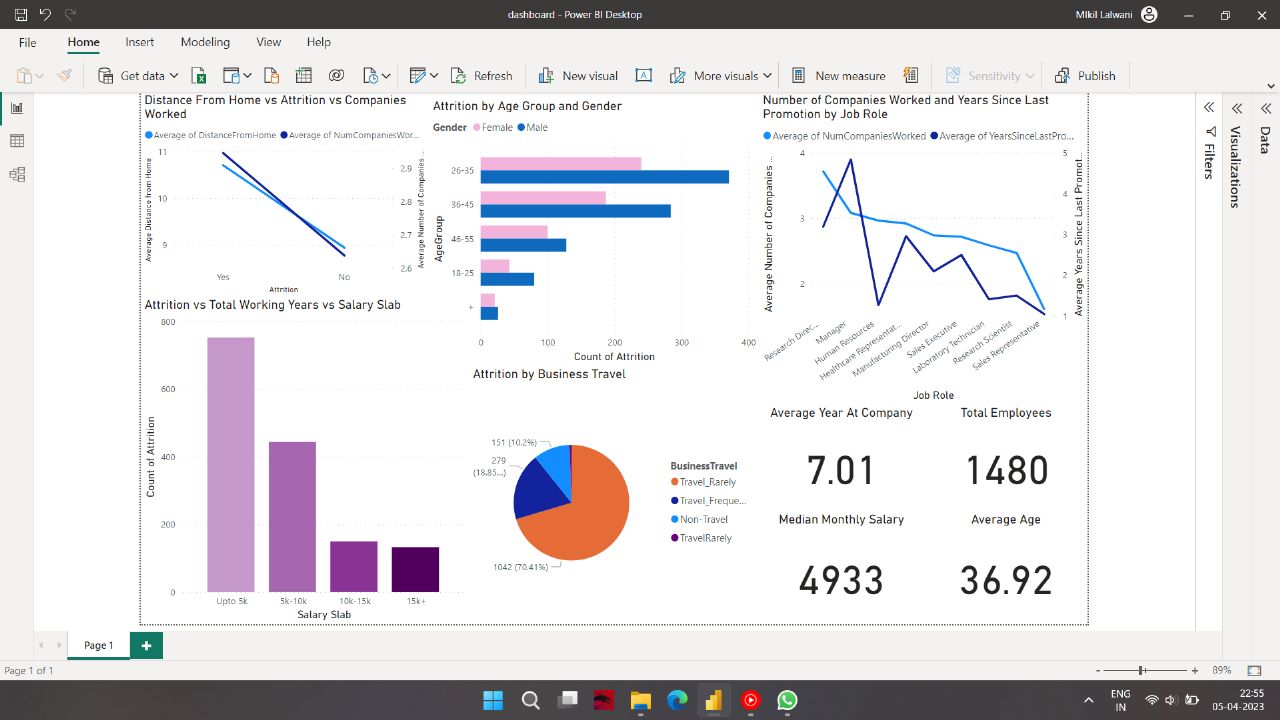
****

****

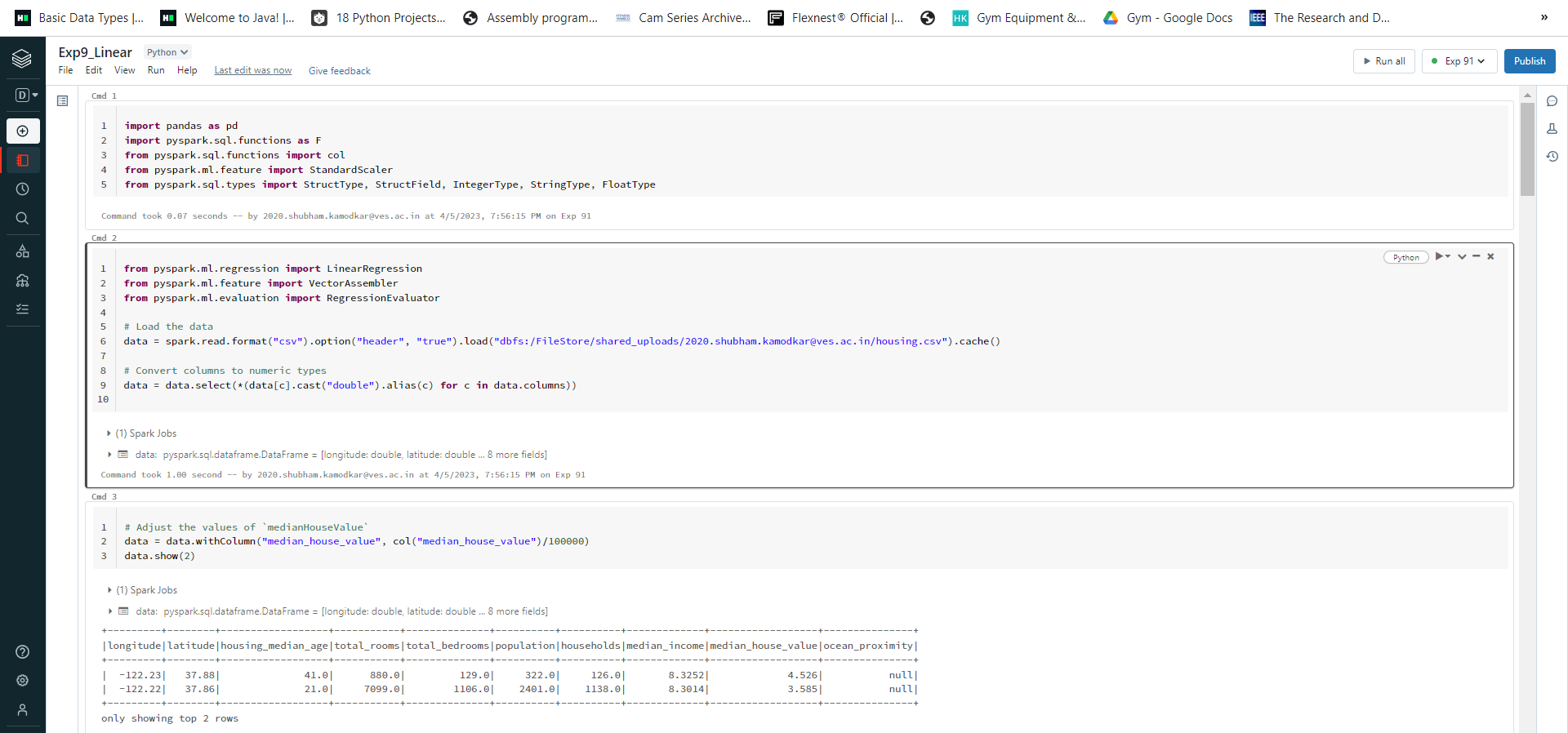
****

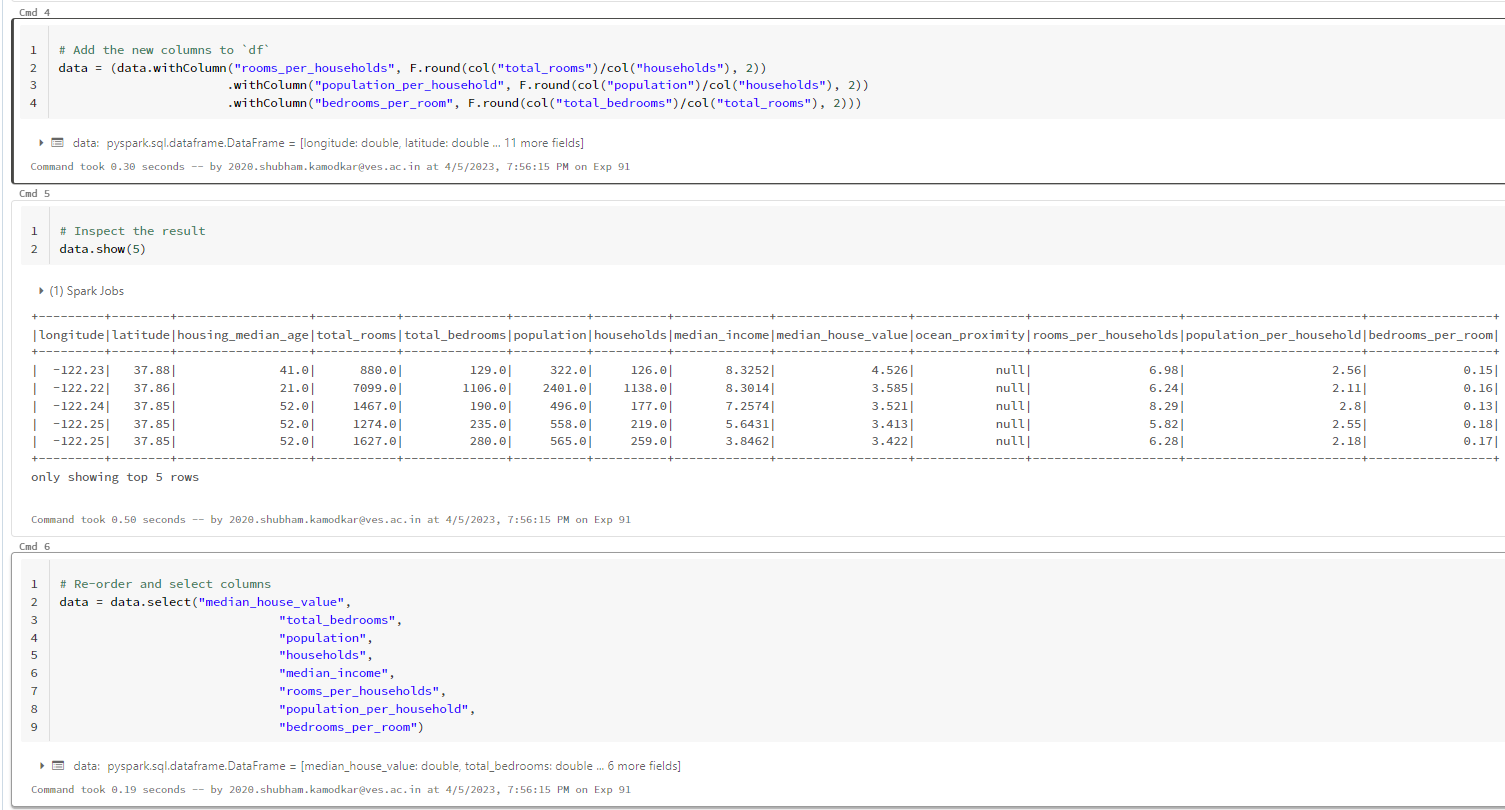
****

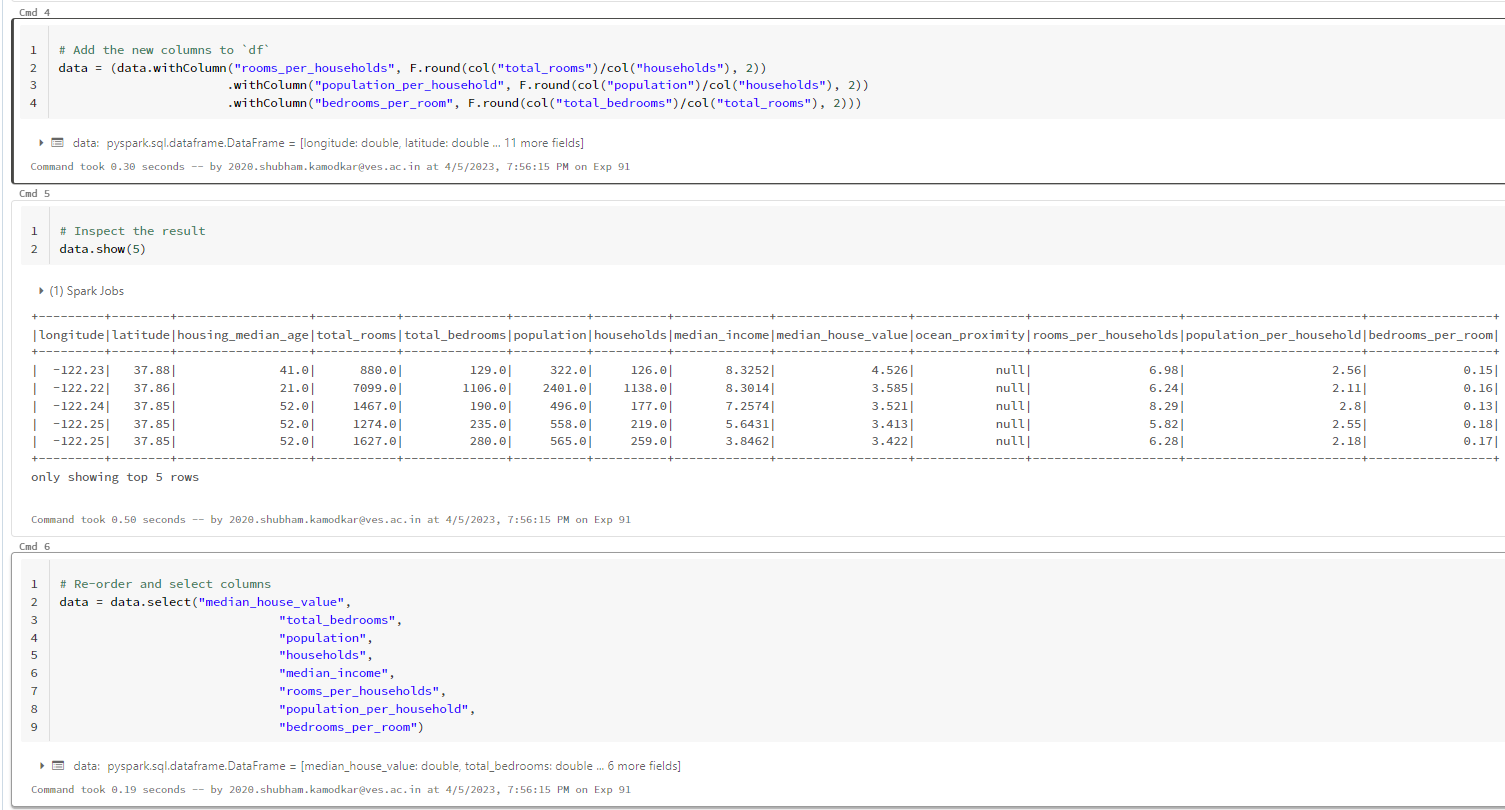
****

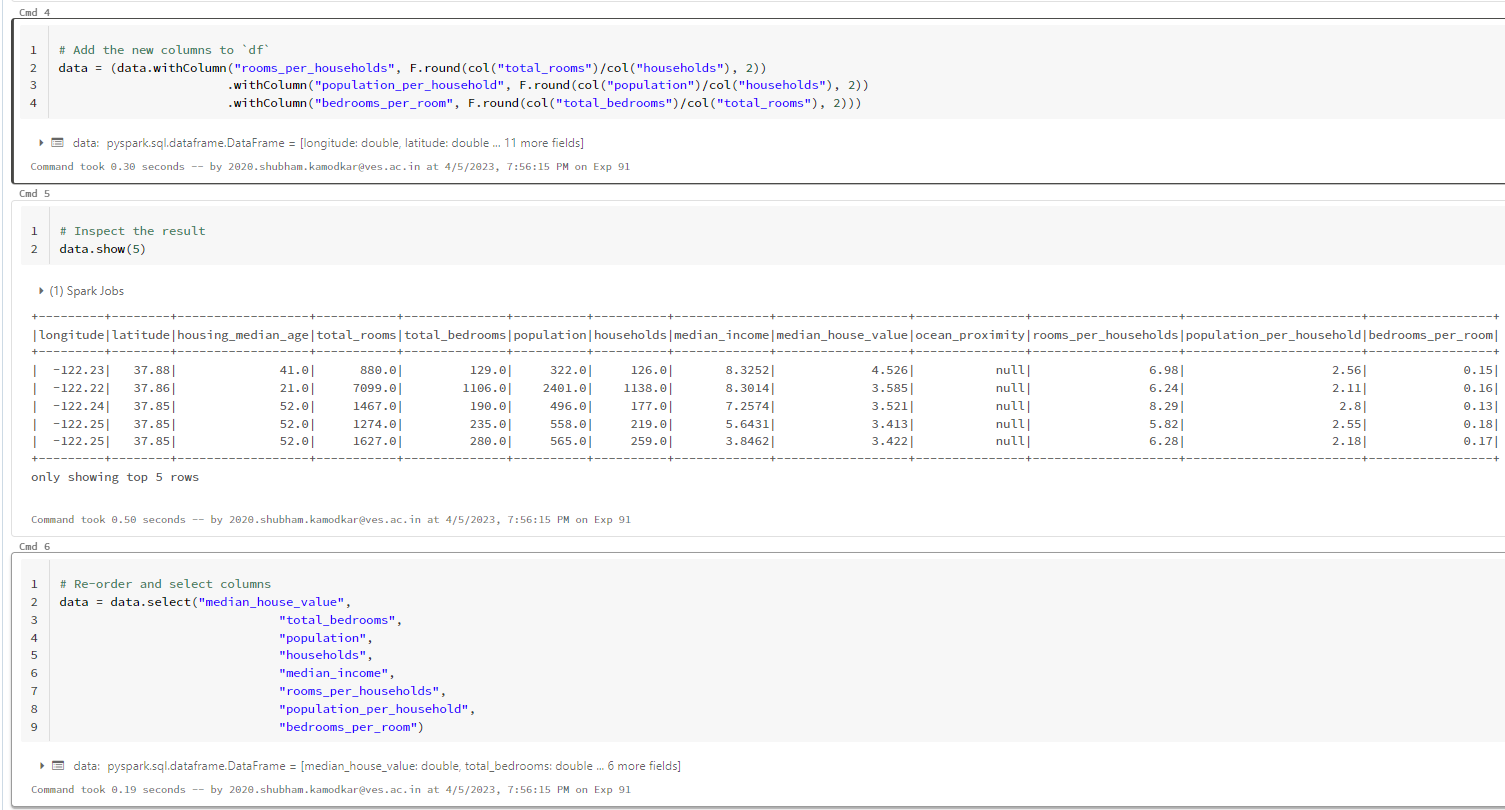
****

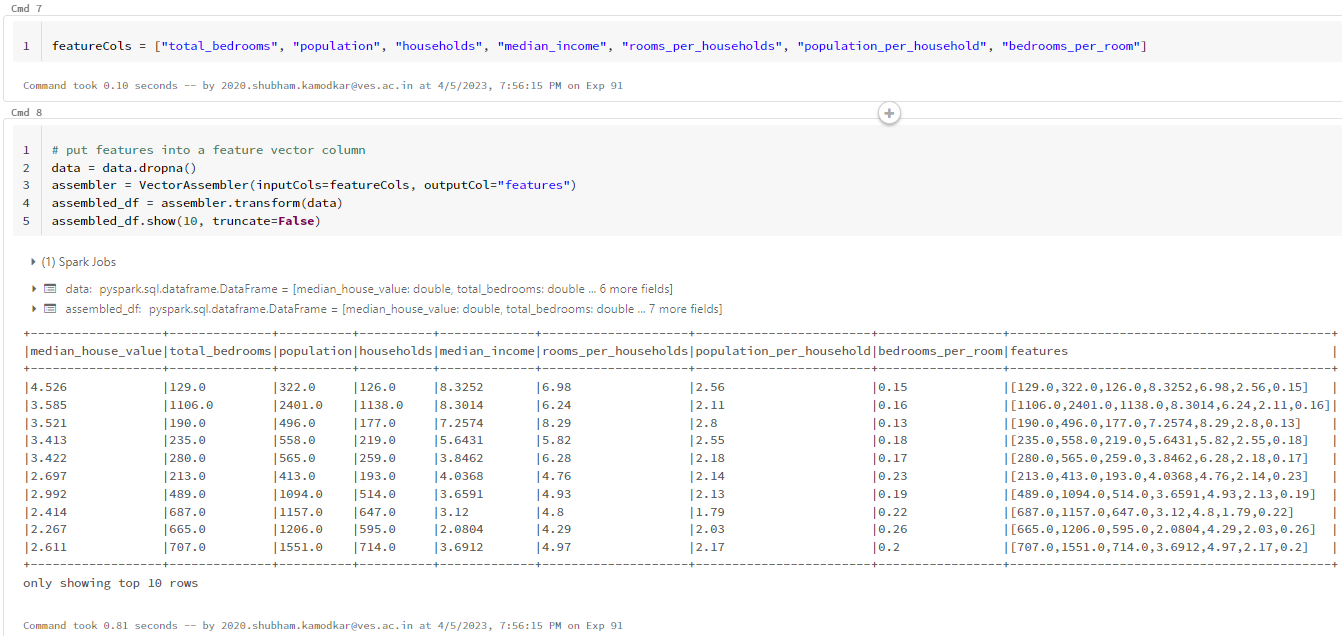
**Linear regression using Apache Pyspark:-**

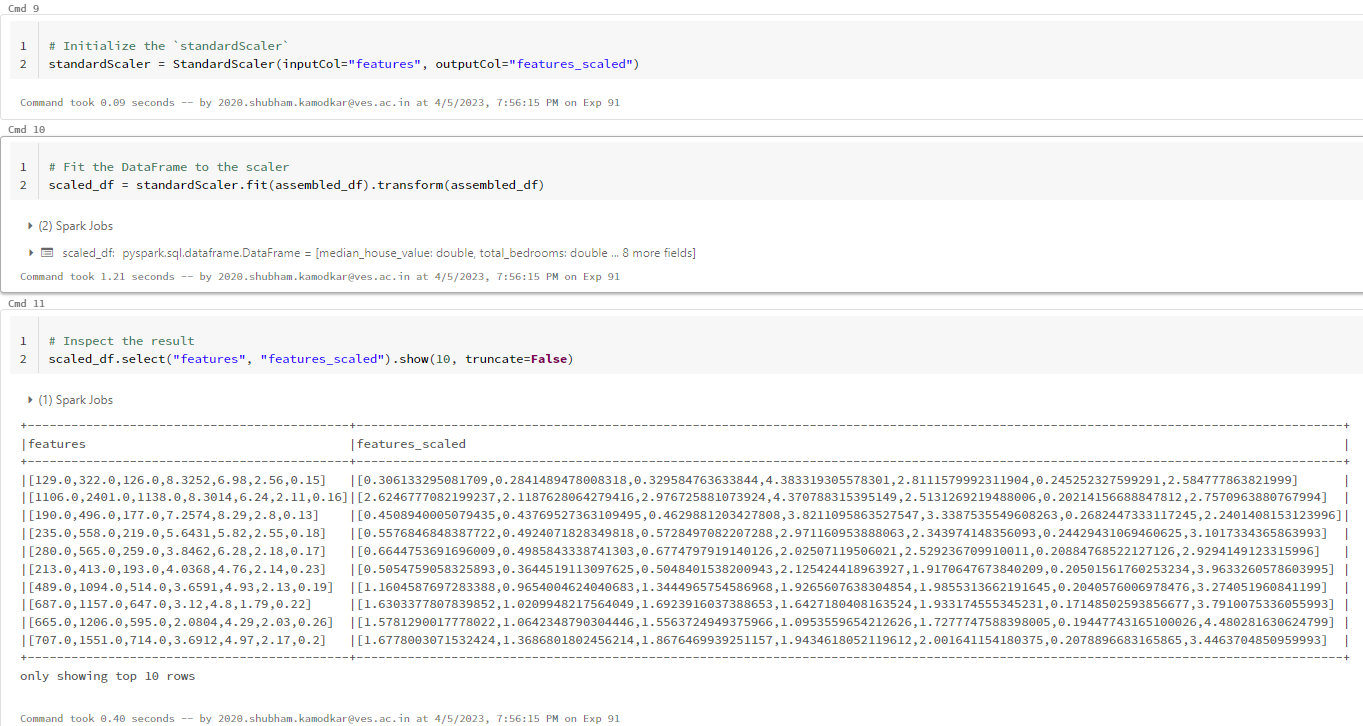
****

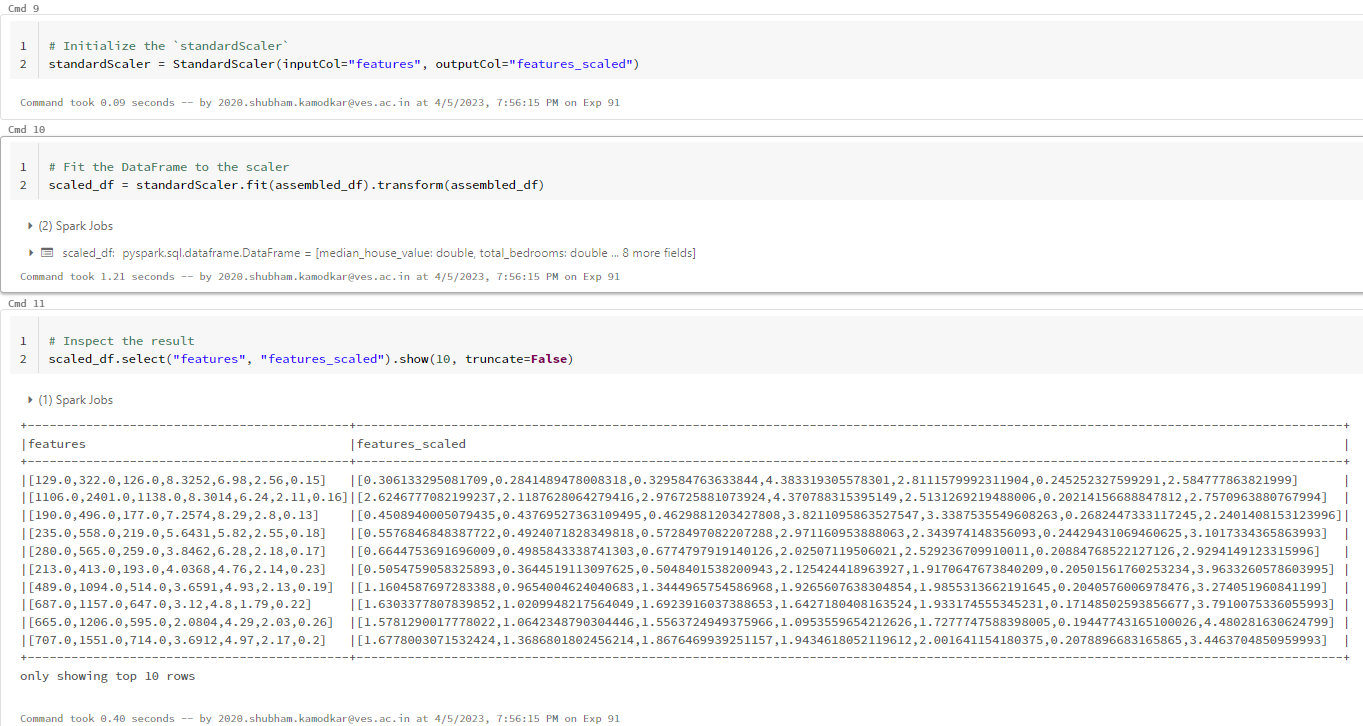
****

****

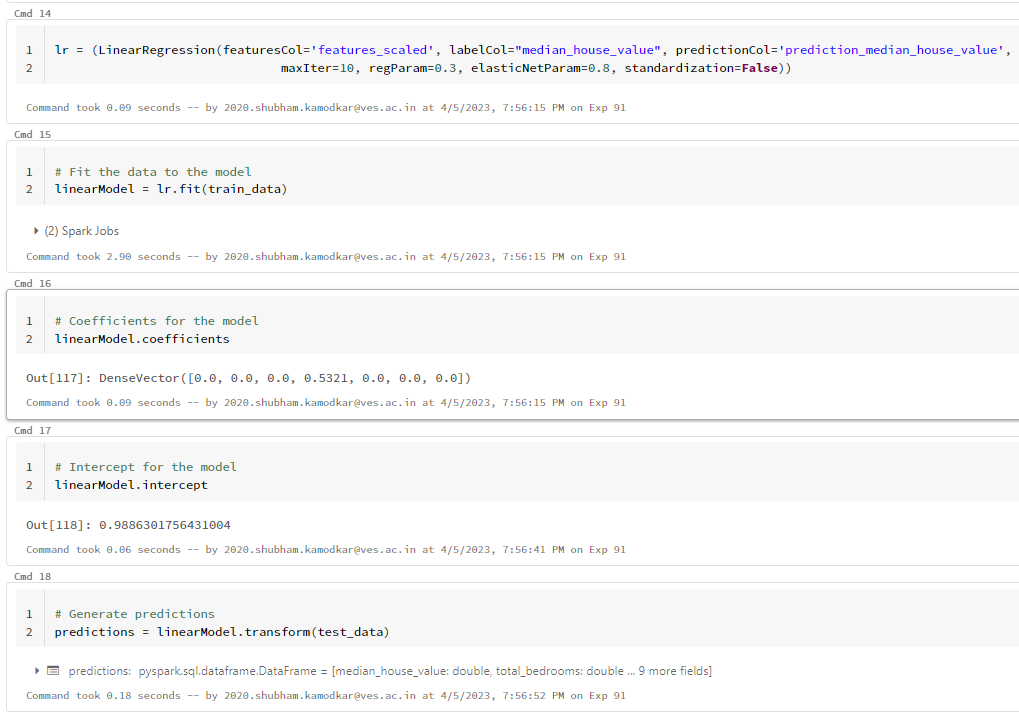
****

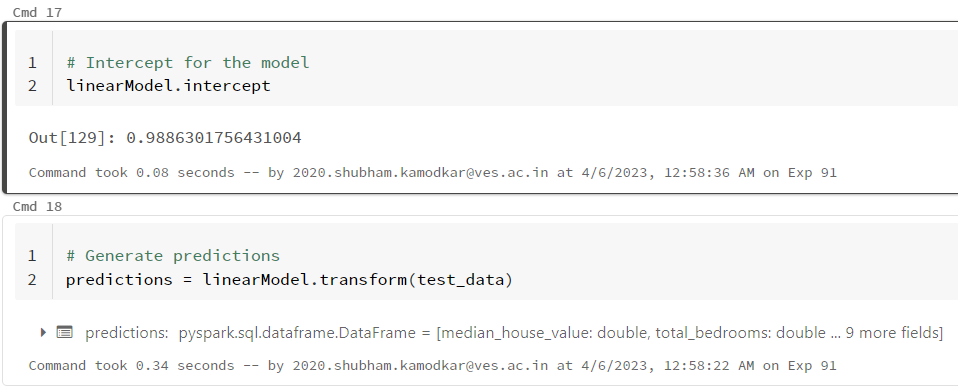
****

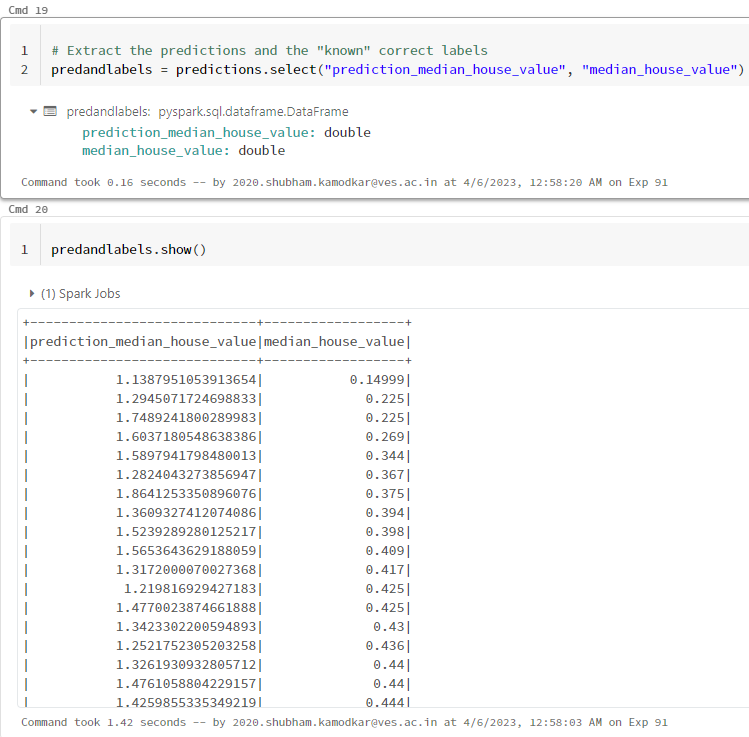
****

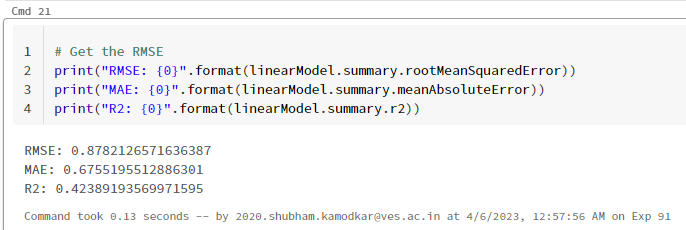
****

****

****

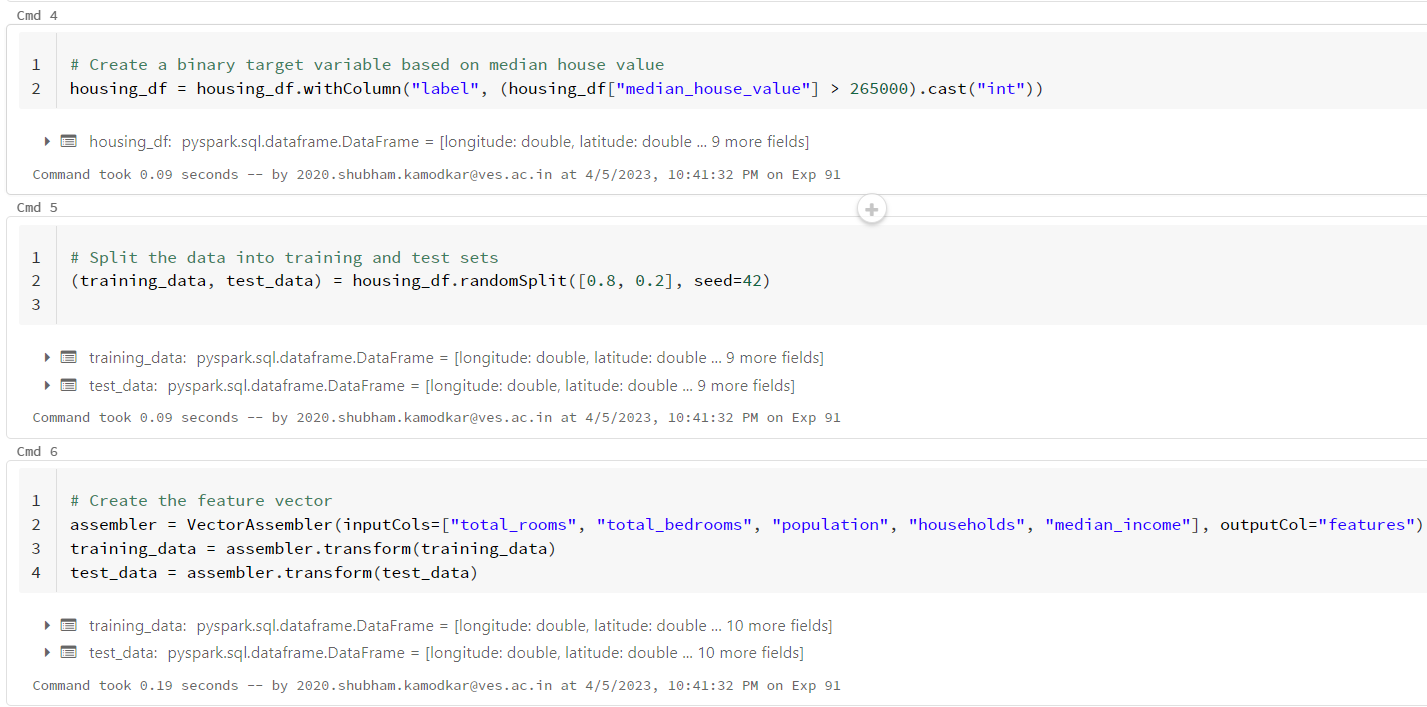
****

****

****

**Logistic regression using Apache Pyspark:-**











**Conclusion:**

Using Power Bi, we visualized data from the housing dataset. We learned about big data and also learned how to use Apache Spark and how to use its libraries for preprocessing and visualization.