**DATA SCIENCE MASTER VIRTUAL INTERNSHIP**

Asummer internship report submitted in partial fulfilment of the requirements for the award of degree of

**BACHELOR OF TECHNOLOGY**

IN

**COMPUTER SCIENCE AND ENGINEERING**

BY

**PRASHANT KUMAR**

**HU21CSEN0101908**

Under the Guidance of

**Dr. Mula Malyadri**

Assistant professor



Department of Computer Science and Engineering

GITAM School of Technology

GITAM (Deemed to be University)

Hyderabad Campus - 502329

July 2024

**GANDHI INSTITUTE OF TECHNOLOGY AND MANAGEMENT**

**(GITAM)**

**(Declared as Deemed-to-be-University u/s 3 of UGC Act 1956)**

**HYDERABAD CAMPUS**



**DECLARATION**

I hereby declare that the summer internship report entitled “**DATA SCIENCE VIRTUAL INTERNSHIP**” to GITAM (Deemed to be university), Hyderabad in partial fulfilment of the requirements for the award of the degree of “**Bachelor of Technology**” in “**Computer Science and Engineering**”. I declare that it was carried out independently by me on the platform of Eduskills - Nation builds through skills.

This internship report had not been submitted to any other college or university for the award of any degree or diploma.

Place: Hyderabad Name:**PRASHANT KUMAR**

Date: 13-08-2024 Student Roll No.: **HU21CSEN0101908**

****

**CERTIFICATION OF COMPLETION**



**ACKNOWLEDGEMENT**

Apart from my effort, the success of this internship largely depends on the encouragement and guidance of many others. I take this opportunity to express my gratitude to the people who have helped me in the successful completion of this internship.

I would like to thank respected **D. Sambasiva Rao,** Pro Vice Chancellor,

GITAM Hyderabad and **N Seetharamaiah**, Principal , GITAM Hyderabad.I would like to thank respected **Shaik Mahaboob Basha**, Head of Computer Science and Engineering for giving me such a wonderful opportunity to expand my knowledge for my own branch and giving me guidelines to present an internship report. It helped me a lot to realize of what we study for.

I would like to thank the respected faculty **Dr. Mula Malyadri** who helped me to make this internship a successful accomplishment.

I would also like to thank the Eduskills and AICTE team who helped me to throughout the internship to make my work organised and well staked till the end.

Sincerely,

Prashant Kumar

HU21CSEN0101908

**ABSTRACT**

Data science involves analysing and interpreting complex data to extract insights and inform decision-making. It encompasses various techniques, including statistical analysis, machine learning, data visualisation, to uncover patterns and relationships within datasets. Compared to traditional methods, data science can handle larger and more diverse datasets, enabling more accurate predictions and deeper insights. Application of data science are vast, ranging from healthcare to finance, where it helps in optimizing operations, understanding customer behaviour, and driving innovation.

I completed the AICTE “Data science Virtual Internship” offered by Eduksills foundation and supported by Altair. This report is the reflection and also the journey of two months internship period along with the highlights of what I learned throughout this period through errors, work responsibilities and the importance of internship program. This program provided in-depth training in data science, covering essential topics such as data analysis, machine learning, statistical modelling. Throughout the course, I gained practical experience in handling and analysing large datasets, feature extraction, and developing predictive models. The internship emphasized the real world applications, enhancing my ability to solve complex problems using advance data science techniques. This experience has significantly strengthened my skills and prepared me for a career in the rapidly evolving field of data science.

**Table of contents**

**CHAPTER 1 : DATA SCIENCE**

1.1. Introduction -----------------------------------------------------------------------------1

1.2. Importance of Data Science -----------------------------------------------------------1

1.3. Uses of Data Science -------------------------------------------------------------------1

1.4. Key Concepts in Data Science --------------------------------------------------------2

1.4.1. Data Collection---------------------------------------------------------------2

1.4.2. Data Preparation--------------------------------------------------------------2

1.4.3. Data Exploration and Visualization----------------------------------------2

1.4.4. Data Modelling -------------------------------------------------------------2

1.4.5. Model Evaluation and Validation -----------------------------------------2

1.4.6. Deployment and Monitoring ----------------------------------------------2

1.5. Tools and Technologies in Data Science --------------------------------------------2

1.6. Challenges in Data Science -----------------------------------------------------------2

**CHAPTER 2: MACHINE LEARNING -------------------------------------------------------4**

2.1. Introduction ----------------------------------------------------------------------------4

2.2. Importance of Machine Learning ---------------------------------------------------4

2.3. Uses of Machine Learning -----------------------------------------------------------4

2.4. Types of Learning Algorithms -------------------------------------------------------5

2.4.1. Supervised Learning --------------------------------------------------------5

2.4.2. Unsupervised Learning -----------------------------------------------------5

2.4.3. Semi-Supervised Learning -------------------------------------------------6

2.5. Conclusion-------------------------------------------------------------------------------6

**CHAPTER 3: ALTAIR AI STUDIO (RAPIDMINER STUDIO)------------------------7**

3.1. Introduction ----------------------------------------------------------------------------7

3.2. Features of RapidMiner --------------------------------------------------------------7

3.3. Uses of RapidMiner -------------------------------------------------------------------7

3.4. Installation of RapidMiner -----------------------------------------------------------8

3.4.1. System Requirements ------------------------------------------------------8

3.4.2. Installation Steps------------------------------------------------------------8

3.5. Importance of RapidMiner Studio---------------------------------------------------9

3.6. Studio Glance --------------------------------------------------------------------------10

3.7. Conclusion -----------------------------------------------------------------------------11

**CHAPTER 4: BASICS OF AI, DATA SCIENCE, AND MACHINE LEARNING ---12**

4.1 Data Science Introduction and Types of Machine Learning ---------------------12

4.1.1 What is Data Science? ------------------------------------------------------12

4.1.2 What is Artificial Intelligence (AI), Machine Learning, and Deep Learning?----------------------------------------------------------------------------12

4.1.3 What Artificial Intelligence and Machine Learning CAN do – and what NOT ---------------------------------------------------------------------------12

4.1.4 Methodologies---------------------------------------------------------------13

4.2 From Descriptive to Predictive and Prescriptive ----------------------------------14

4.2.1 Introduction to Advanced Analytics---------------------------------------14

4.2.2 Back to the Future or "How to Get Predictive" -------------------------14

4.2.3 Prescriptive Analytics ------------------------------------------------------14

**CHAPTER 5: VISUALIZING DATA----------------------------------------------------------15**

5.1 Data visualization and graphical perception ---------------------------------------15

**CHAPTER 6: DATA VALIDATION ----------------------------------------------------------19**

6.1 Intro to Data Validation ---------------------------------------------------------------19

6.1.1 Importance of Data Validation ---------------------------------------------19

6.1.2 Types of Data Validation ---------------------------------------------------19

6.1.3 Data Validation Techniques ------------------------------------------------19

6.2 Cross Validation ------------------------------------------------------------------------19

6.2.1 Importance of Cross Validation--------------------------------------------19

6.2.2 Types of Cross Validation--------------------------------------------------19

6.2.3 Process of Cross Validation -----------------------------------------------19

6.3 Validating a Model --------------------------------------------------------------------20

6.4 How to Correctly Validate Machine Learning Models---------------------------20

**Chapter 7: PROJECT----------------------------------------------------------------------------21**

7.1 Introduction ----------------------------------------------------------------------------21

7.2 Tools and Technologies Used--------------------------------------------------------21

7.2.1 Altair AI Studio Overview ------------------------------------------------21

7.2.2 Key Features ----------------------------------------------------------------22

7.3 Project Dataset ------------------------------------------------------------------------22

7.3.1 Data Set Introduction ------------------------------------------------------22

7.3.2 Dataset Overview ----------------------------------------------------------22

7.4 Data Engineering ---------------------------------------------------------------------24

7.4.1 Importing the Data into AI Studio---------------------------------------24

7.4.2 Data Statistics --------------------------------------------------------------24

7.4.3 Data Visualization ---------------------------------------------------------25

7.5 Building and Running a Simple Process ------------------------------------------26

7.6 Correlation Matrix Process ---------------------------------------------------------28

7.6.1 Correlation Matrix Operator ---------------------------------------------28

7.6.2 Correlation Matrix Process Output -------------------------------------28

7.7 Data Aggregation Process ----------------------------------------------------------29

7.7.1 Aggregate Operator ------------------------------------------------------29

7.8 Linear Regression -------------------------------------------------------------------31

7.8.1 Linear Regression Process ----------------------------------------------32

7.9 Results --------------------------------------------------------------------------------36

7.9.1 ExampleSet (Apply Model)---------------------------------------------36

7.9.2 PerformanceVector (Performance)-------------------------------------39

7.9.3 Linear Regression (Linear Regression)--------------------------------40

**CONCLUSION----------------------------------------------------------------------------------42**

**REFERENCES----------------------------------------------------------------------------------43**

**CHAPTER 1: DATA SCIENCE**

**1.1. Introduction**

Data Science is a multidisciplinary field that combines statistics, computer science, domain expertise, and information technology to extract meaningful insights from structured and unstructured data. By employing scientific methods, processes, algorithms, and systems, data scientists can analyze complex data to inform decision-making, predict trends, and solve problems across various industries.

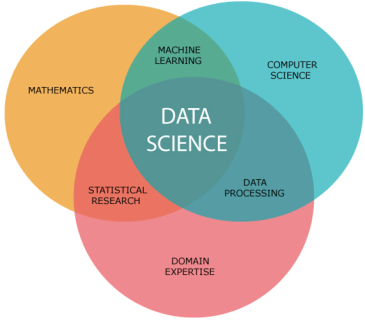


Fig.1.1

**1.2. Importance of Data Science**

The importance of Data Science cannot be overstated in today's data-driven world. It enables organizations to make informed decisions based on data analysis, optimize operations, improve customer experiences, and gain a competitive edge. Data Science drives innovation, supports strategic planning, and helps identify new business opportunities by uncovering patterns and trends hidden within large datasets.

**1.3. Uses of Data Science**

Data Science has a wide array of applications across different sectors:

* **Business**: Enhances decision-making through predictive analytics and business intelligence.
* **Healthcare**: Supports diagnostics, personalized medicine, and predictive modeling for patient outcomes.
* **Finance**: Improves risk management, fraud detection, and algorithmic trading.
* **Retail**: Optimizes inventory management, customer segmentation, and personalized marketing.
* **Manufacturing**: Enhances quality control, predictive maintenance, and supply chain optimization.

**1.4. Key Concepts in Data Science**

Data Science involves several key concepts that are essential for understanding and applying data-driven methods:

**1.4.1. Data Collection**

Data collection is the process of gathering raw data from various sources, including databases, APIs, sensors, and web scraping. This data forms the foundation for analysis and modeling.

**1.4.2. Data Preparation**

Data preparation involves cleaning, transforming, and organizing raw data to make it suitable for analysis. This step includes handling missing values, removing duplicates, normalizing data, and creating relevant features.

**1.4.3. Data Exploration and Visualization**

Data exploration and visualization help data scientists understand the underlying patterns and relationships within the data. Techniques such as descriptive statistics, data profiling, and visualizations (e.g., histograms, scatter plots) are used to summarize and interpret data.

**1.4.4. Data Modelling**

Data modelling involves creating mathematical models to represent the data and predict outcomes. Techniques such as regression analysis, classification, clustering, and neural networks are employed to build models that can learn from data and make predictions.

**1.4.5. Model Evaluation and Validation**

Model evaluation and validation are crucial steps to ensure the accuracy and reliability of predictive models. This involves using metrics such as accuracy, precision, recall, and F1-score, as well as techniques like cross-validation and split testing to assess model performance.

**1.4.6. Deployment and Monitoring**

Deploying data science models into production environments allows organizations to leverage insights and predictions in real-time applications. Continuous monitoring and maintenance of models ensure they remain effective and accurate over time.

**1.5. Tools and Technologies in Data Science**

Several tools and technologies are commonly used in Data Science to facilitate data analysis and modelling:

* **Programming Languages**: Python, R, SQL
* **Data Visualization Tools**: Tableau, Power BI, Matplotlib, Seaborn
* **Machine Learning Libraries**: Scikit-learn, TensorFlow, Keras, PyTorch
* **Data Processing Frameworks**: Apache Spark, Hadoop
* **Data Management Platforms**: SQL databases, NoSQL databases, cloud-based storage solutions

**1.6. Challenges in Data Science**

Data Science faces several challenges, including:

* **Data Quality**: Ensuring the accuracy, completeness, and reliability of data.
* **Data Privacy and Security**: Protecting sensitive information and complying with data protection regulations.
* **Scalability**: Managing and processing large volumes of data efficiently.
* **Interdisciplinary Collaboration**: Integrating knowledge and expertise from various domains to address complex problems.

In conclusion, Data Science is a powerful field that enables organizations to harness the potential of data for strategic advantage. By understanding its key concepts, tools, and applications, businesses can make data-driven decisions, optimize operations, and innovate in their respective industries.

**CHAPTER 2: MACHINE LEARNING**

**2.1. Introduction**

Machine Learning (ML) is a branch of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to perform tasks without explicit instructions. These models learn from data, identifying patterns and making decisions with minimal human intervention. As data becomes increasingly abundant, machine learning has emerged as a critical tool for extracting meaningful insights and automating complex processes.

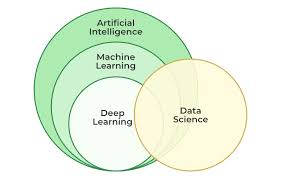


Fig.2.1

**2.2. Importance of Machine Learning**

Machine learning is vital in today's data-driven world because it allows for the efficient analysis of large datasets. It is instrumental in various fields, including healthcare, finance, marketing, and robotics, where it aids in making predictions, detecting anomalies, and optimizing processes. ML enhances decision-making by providing accurate, data-driven insights, ultimately leading to more informed and effective strategies across industries.

**2.3. Uses of Machine Learning**

Machine learning applications are vast and diverse. In healthcare, ML is used to predict patient outcomes, personalize treatments, and diagnose diseases. In finance, it assists in fraud detection, risk management, and algorithmic trading. Marketing leverages ML for customer segmentation, personalized advertising, and sentiment analysis. Furthermore, machine learning is crucial in developing autonomous systems, such as self-driving cars and intelligent robots.

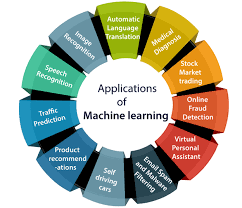


Fig.2.3

**2.4. Types of Learning Algorithms**

Machine learning algorithms are broadly categorized into three types based on how they learn from data: supervised learning, unsupervised learning, and semi-supervised learning.

**2.4.1. Supervised Learning**

Supervised learning algorithms are trained on labeled datasets, where the input data is paired with the correct output. The model learns to map inputs to the desired outputs by minimizing the error between predicted and actual results. Common applications include classification tasks like spam detection and regression tasks such as predicting housing prices.

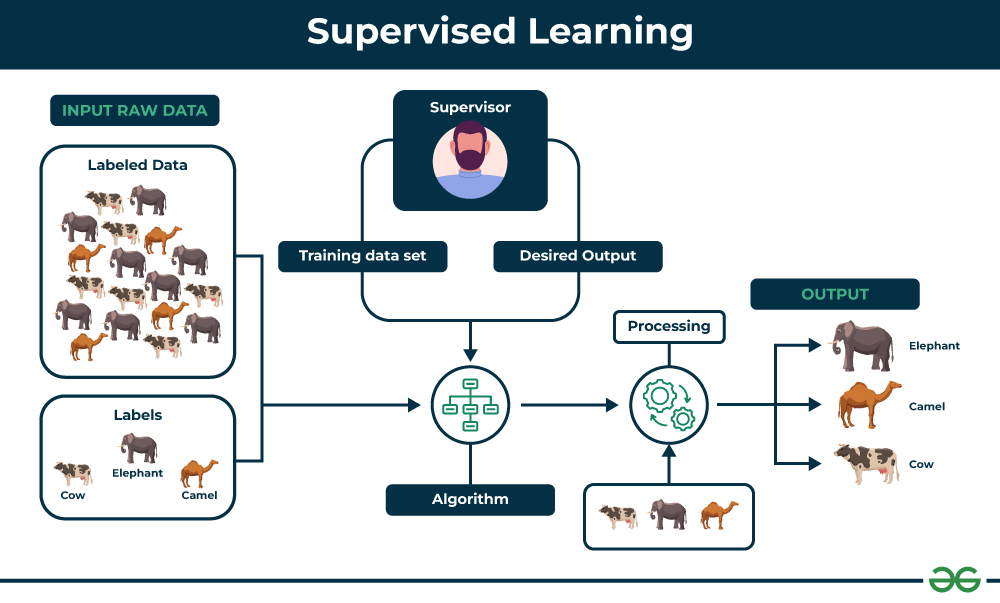


Fig.2.4.1

**2.4.2. Unsupervised Learning**

Unsupervised learning algorithms work with unlabeled data, identifying underlying patterns or structures without predefined labels. Clustering and association are common techniques in unsupervised learning. Examples include customer segmentation, where customers are grouped based on purchasing behavior, and anomaly detection in network security.

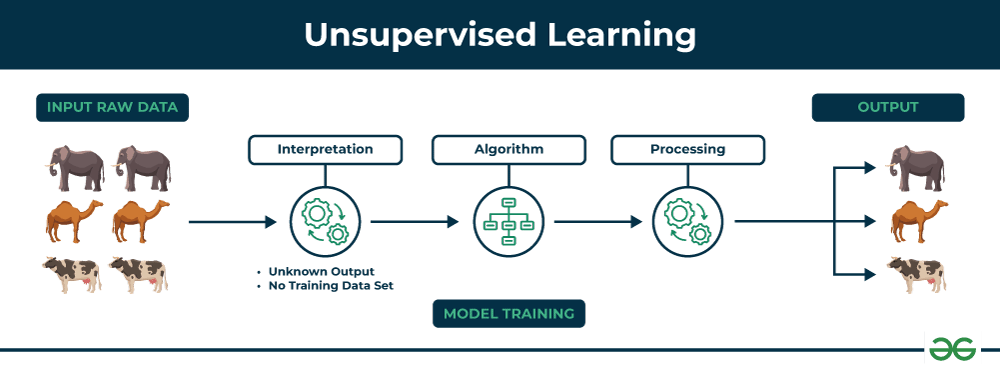


Fig.2.4.2

**2.4.3. Semi-Supervised Learning**

Semi-supervised learning is a hybrid approach that leverages both labeled and unlabeled data. It is particularly useful when labeled data is scarce or expensive to obtain, but a large amount of unlabeled data is available. This approach is often used in scenarios like image recognition, where labeling every image is impractical, but the model can still learn effectively from a combination of labeled and unlabeled data.

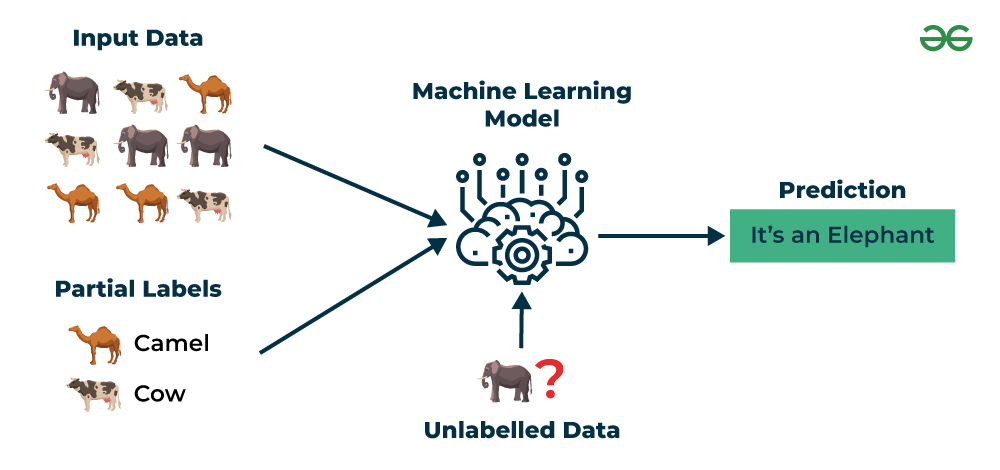


Fig.2.4.3

**2.5. Conclusion**

Machine learning is a transformative technology that is reshaping numerous industries by enabling more intelligent and automated decision-making. Understanding the different types of learning algorithms and their applications is essential for leveraging the full potential of machine learning in solving complex real-world problems. As the field continues to evolve, it will undoubtedly play an increasingly important role in driving innovation and efficiency across all sectors.

**CHAPTER 3: ALTAIR AI STUDIO (formerly known as RapidMiner Studio)**

**3.1. Introduction**

RapidMiner is a powerful, open-source data science platform that offers a comprehensive suite of tools for data preparation, machine learning, deep learning, text mining, and predictive analytics. It provides a user-friendly interface that allows both novice and experienced data scientists to build, deploy, and maintain predictive models quickly and efficiently. RapidMiner is widely used in academia and industry for its ability to simplify complex data science processes and its extensive support for various machine learning algorithms.



Fig.3.1

**3.2. Features of RapidMiner**

RapidMiner stands out for its versatility and ease of use. Key features include:

* **Drag-and-Drop Interface:** RapidMiner's visual workflow designer allows users to create complex models without writing code, making it accessible to non-programmers.
* **Extensive Algorithm Library:** The platform supports a wide range of machine learning algorithms, from basic to advanced, enabling users to tackle various data science challenges.
* **Data Integration:** RapidMiner can connect to multiple data sources, including databases, cloud storage, and flat files, allowing seamless data integration.
* **Automated Machine Learning (AutoML):** RapidMiner provides AutoML capabilities, which automatically select the best model and parameters for a given dataset, speeding up the model development process.
* **Scalability:** RapidMiner can scale from single-machine operations to distributed computing environments, handling large datasets efficiently.

**3.3. Uses of RapidMiner**

RapidMiner is used across various industries for different purposes:

* **Predictive Maintenance:** In manufacturing, RapidMiner helps predict equipment failures, reducing downtime and maintenance costs.
* **Customer Analytics:** Businesses use RapidMiner for customer segmentation, churn prediction, and personalized marketing strategies.
* **Fraud Detection:** Financial institutions leverage RapidMiner to detect and prevent fraudulent activities by analyzing transaction patterns.
* **Healthcare:** RapidMiner assists in predictive analytics for patient outcomes, resource allocation, and personalized treatment plans.

**3.4. Installation of RapidMiner**

**3.4.1. System Requirements**

Before installing RapidMiner, ensure that your system meets the following minimum requirements:

* **Operating System:** Windows 10 or later, macOS 10.13 or later, Linux (Ubuntu, CentOS)
* **RAM:** 8 GB minimum, 16 GB recommended
* **Disk Space:** 2 GB minimum, 4 GB recommended
* **Java Runtime Environment (JRE):** RapidMiner requires Java 8 or later.

**3.4.2. Installation Steps**

**Step 1: Download the Installer**

1. Visit the RapidMiner website and navigate to the "Downloads" section.
2. Select the appropriate version of RapidMiner for your operating system (Windows, macOS, or Linux).
3. Download the installer file to your computer.

**Step 2: Install RapidMiner**

1. **Windows:**
   * Run the downloaded .exe file.
     1. Follow the installation wizard instructions. You can choose the default settings or customize the installation path.
     2. After installation, launch RapidMiner Studio from the Start Menu or desktop shortcut.
2. **macOS:**
   * Open the downloaded .dmg file.
   * Drag the RapidMiner icon into the Applications folder.
   * Launch RapidMiner Studio from the Applications folder.
3. **Linux:**
   * Extract the downloaded .tar.gz file to your desired directory.
   * Navigate to the extracted folder and run the RapidMiner-Studio.sh script.
   * Follow any on-screen instructions to complete the installation.

For more info visit <https://docs.rapidminer.com/9.9/studio/installation/>

**Step 3: Launch RapidMiner and Set Up**

1. When you launch RapidMiner Studio for the first time, you will be prompted to log in or create a RapidMiner account.
2. After logging in, you can start a new project or explore sample datasets provided by RapidMiner.
3. Configure any necessary preferences or connect to your data sources through the interface.

**3.5 Importance of RapidMiner Studio**

RapidMiner Studio is an essential tool in the field of data science and analytics due to its comprehensive features and user-friendly design. Its importance can be understood through several key aspects:

1. **Ease of Use**: RapidMiner Studio is designed to be accessible to users with varying levels of technical expertise. Its intuitive drag-and-drop interface allows users to build complex analytical workflows without extensive programming knowledge. This democratizes data science, enabling business analysts, marketers, and other non-technical professionals to participate in data-driven decision-making processes.

2. **Comprehensive Functionality**: RapidMiner Studio offers a complete suite of tools for the entire data science lifecycle, from data preparation and cleaning to model building, validation, and deployment. It supports a wide range of data analysis tasks, including machine learning, deep learning, text mining, and time series analysis. This all-in-one platform reduces the need for multiple disparate tools, streamlining the analytical process and improving efficiency.

3. **Integration Capabilities**: The platform’s ability to connect to various data sources, both on-premises and in the cloud, is crucial for modern data analysis. It supports integration with databases, big data platforms, cloud services, and other applications, ensuring that users can easily access and analyse their data regardless of where it is stored. This flexibility is vital for organizations with diverse data ecosystems.

4. **Scalability and Collaboration**: RapidMiner Studio is designed to scale with the needs of the organization. It can handle projects of varying sizes and complexities, from small exploratory analyses to large-scale enterprise deployments. Additionally, it supports collaboration through project sharing and integration with RapidMiner Server and RapidMiner AI Hub, allowing teams to work together seamlessly on data science projects.

5. **Advanced Analytics**: The platform provides advanced analytics capabilities, including support for custom scripting in R and Python, which allows users to extend its functionality and tailor analyses to specific needs. This flexibility is crucial for organizations that require bespoke solutions or need to integrate cutting-edge research and techniques into their workflows.

6. **Community and Support**: RapidMiner has a strong user community and provides extensive resources, including forums, tutorials, documentation, and user-contributed extensions. This community support is invaluable for troubleshooting, learning new techniques, and staying updated with the latest developments. Professional support and consulting services are also available for organizations needing additional assistance.

7. **Operationalization of Data Science**: RapidMiner Studio facilitates the deployment of models into production environments, making it easier to operationalize data science. This means that insights and predictive models can be integrated into business processes in real-time, driving immediate impact. Features like automated scheduling and monitoring ensure that data workflows remain up-to-date and effective.

8. **Educational Value**: RapidMiner Studio is widely used in academic settings, providing students with practical experience in data science and analytics. Its user-friendly interface and comprehensive functionality make it an excellent teaching tool, helping to prepare the next generation of data scientists.

In summary, RapidMiner Studio’s importance lies in its ability to make data science accessible, efficient, and impactful. By providing a robust, scalable, and user-friendly platform, it enables organizations to leverage their data effectively, driving better decision-making and fostering a data-driven culture. Its integration capabilities, advanced analytics features, and strong community support further enhance its value, making it a critical tool in the modern data scientist’s toolkit.

**3.6 Studio Glance**

An introduction to RapidMiner Studio 9.3 covers the interface, data import, and process building. It explains using design and results views, creating subfolders for data organization, and generating a correlation matrix from a dataset.

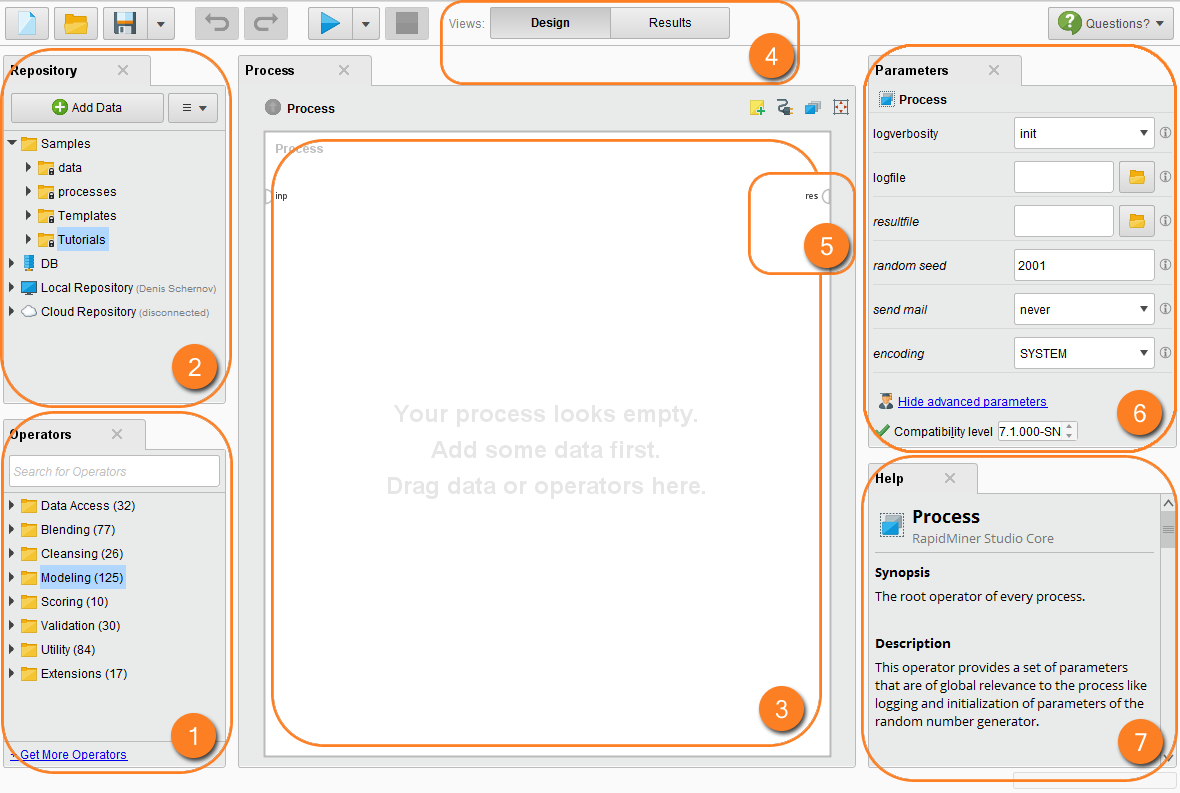


Fig. 3.6

RapidMiner Studio provides an intuitive interface for data analysis, focusing on the design and results views. Users can easily import datasets and build processes for various projects. The design view is essential for building models, while the results view displays outcomes. Users can switch easily between these two views for efficient workflow. The repository panel helps organize data and processes, with a local repository created by default upon installation. This aids in managing multiple projects effectively.

Operators are crucial for developing models, such as regression. Users can find helpful information about each operator in the dedicated help panel.

Users can customize the layout of Rapid Miner by expanding process panels and restoring default views whenever needed. Organizing data into subfolders like 'data' and 'processes' enhances workflow efficiency. Restoring default views in Rapid Miner is a straightforward process that helps users reset their workspace efficiently. This feature aids in maintaining a clean interface while working.

Creating structured subfolders within a repository allows for better organization of datasets and processes. This practice improves accessibility and management of various project components. When importing data, users can select specific worksheets and data ranges. This flexibility ensures that only relevant data is brought into Rapid Miner for analysis.

**3.7. Conclusion**

RapidMiner is a robust data science platform that simplifies the process of building and deploying machine learning models. Its intuitive interface, extensive features, and flexibility make it a valuable tool for both beginners and seasoned data scientists. With the ability to handle various data science tasks, from data preparation to model deployment, RapidMiner empowers users to extract insights and make data-driven decisions effectively. Whether you're looking to enhance your business processes, optimize operations, or explore new data science possibilities, RapidMiner offers the tools and capabilities to achieve your goals.

**CHAPTER 4: Basics of AI, Data Science, and Machine Learning**

**4.1 Data Science Introduction and Types of ML**

**4.1.1 What is Data Science?**

Data Science is an interdisciplinary field that involves extracting insights and knowledge from structured and unstructured data using scientific methods, processes, algorithms, and systems. It encompasses various techniques from statistics, machine learning, data mining, and big data analytics to analyze and interpret complex data. The primary goal of data science is to turn data into actionable insights that can guide decision-making in various domains such as business, healthcare, finance, and more.

**4.1.2 What is Artificial Intelligence (AI), Machine Learning, and Deep Learning?**

* Artificial Intelligence (AI): AI is a branch of computer science that aims to create machines or systems that can perform tasks that typically require human intelligence. These tasks include reasoning, learning, problem-solving, understanding natural language, and perception. AI can be categorized into two types: narrow AI, which is designed for specific tasks, and general AI, which would have the ability to perform any intellectual task that a human can do.
* Machine Learning (ML): ML is a subset of AI that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data. Unlike traditional programming, where explicit instructions are given, machine learning algorithms identify patterns in data and improve over time through experience.
* Deep Learning (DL): Deep Learning is a subfield of machine learning that involves neural networks with many layers (hence the term "deep"). These neural networks can automatically learn hierarchical representations of data, making them particularly effective for tasks like image and speech recognition.

**4.1.3 What Artificial Intelligence and Machine Learning CAN do – and what NOT**

* What AI and ML CAN do:
  + Automation: AI and ML can automate repetitive tasks, freeing up human workers for more complex activities.
  + Prediction: ML models can predict outcomes based on historical data, useful in fields like finance, healthcare, and marketing.
  + Personalization: AI systems can personalize user experiences, such as recommending products or content based on individual preferences.
  + Data Analysis: AI can process and analyze vast amounts of data faster and more accurately than humans.
* What AI and ML CANNOT do:
* General Intelligence: AI lacks the ability to think, understand, and reason like a human across a wide range of tasks.
* Creativity: While AI can generate content, it cannot truly create or innovate in the way humans can.
* Emotional Understanding: AI cannot genuinely understand or replicate human emotions and the nuances of social interactions.
* Ethical Decision-Making: AI systems can follow rules but struggle with ethical decision-making, especially in ambiguous situations.

**4.1.4 Methodologies**

In data science and machine learning, various methodologies guide the process of transforming raw data into actionable insights. These methodologies include:

* CRISP-DM (Cross-Industry Standard Process for Data Mining): A popular and widely used methodology that outlines six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. CRISP-DM is iterative, allowing for continuous refinement of models.

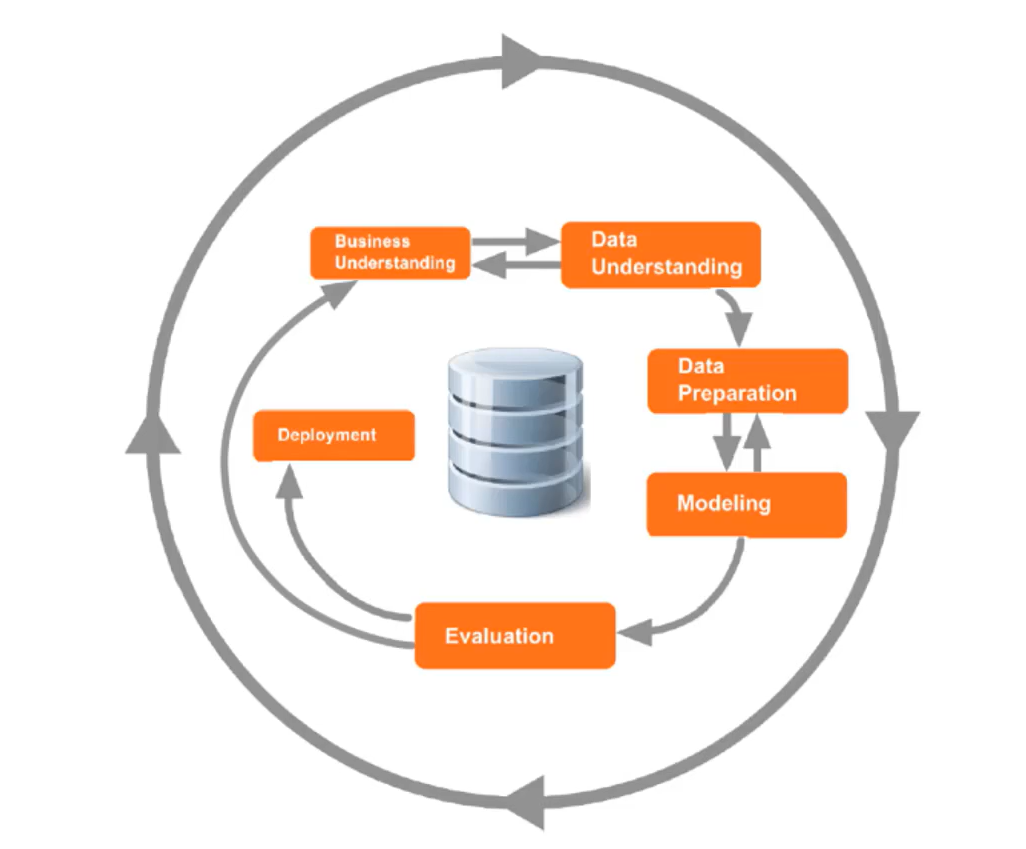


Fig.4.1.4

* RapidMiner: A data science platform that implements CRISP-DM and supports the entire data science lifecycle, from data preparation and modeling to deployment. RapidMiner offers a user-friendly interface for building predictive models and performing advanced analytics.

**4.2 From Descriptive to Predictive and Prescriptive**

**4.2.1 Introduction to Advanced Analytics**

Advanced analytics involves using sophisticated techniques and tools to analyze data beyond basic descriptive statistics. It includes predictive analytics, which forecasts future events based on historical data, and prescriptive analytics, which provides recommendations for decision-making. Advanced analytics helps organizations make data-driven decisions by uncovering patterns, trends, and relationships in data.

**4.2.2 Back to the Future or "How to Get Predictive"**

Predictive analytics uses statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. It involves creating predictive models that can forecast events such as customer behavior, equipment failure, or market trends. By leveraging predictive analytics, organizations can anticipate and respond to future challenges and opportunities.

**4.2.3 Prescriptive Analytics**

Prescriptive analytics goes beyond prediction by suggesting specific actions to achieve desired outcomes. It combines data, algorithms, and business rules to recommend optimal decisions. Prescriptive analytics helps organizations not only predict what will happen but also determine the best course of action to take in response. This approach is valuable in areas like supply chain management, pricing strategies, and personalized marketing.

**CHAPTER 5: VISUALIZING DATA**

**5.1 Data Visualization and Graphical Perception**

Data visualization helps the human brain interpret complex data, making it more accessible for understanding. Data science teams focus on cleaning data, building predictive models, and optimizing business outcomes. Effective communication of data patterns and trends to stakeholders relies on clear visualizations.

However, graphical perception—the ability to interpret visual information—is often overlooked in various phases of the data science process. For example, stakeholders may better understand bar charts or histograms over pie charts for frequency distribution analysis.

It's important to remember that while visual perception is powerful, it has its limits. This article introduces graphical perception concepts, best practices for creating effective charts, and how to choose appropriate visualizations based on these principles.

**Approaches to learning about data**

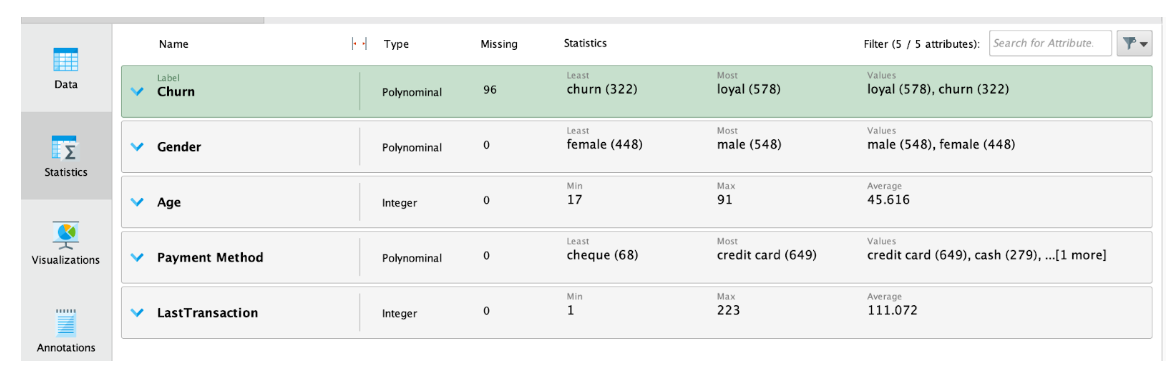
To better understand the approaches to learn about the data through data visualization, here we use the Customer Churn dataset in RapidMiner studio Training repository that consists of the following five attributes:

* Gender (Binominal - Male or Female)
* Age (Integer)
* Payment method (Polynominal - cheque, credit card, cash)
* Last transaction (Integer)
* Churn (Binominal - Churn, Loyal)

**Descriptive Analytics**

Descriptive analysis constructively summarizes a sample or the entire dataset in a meaningful way such that patterns might emerge in the data. The techniques often include calculating descriptive statistics - like mean, standard deviation, variance, or min/max - and univariate and multivariate analysis using different visualizations.

Within the RapidMiner studio, we will retrieve the customer churn data. Then, we will join the output to the results and run the process. In the Data section, you will see the entire dataset is displayed. In the Statistics section, we can see the descriptive statistics (like min, max, average) for each attribute and the number of missing values as well - in this case, Churn.

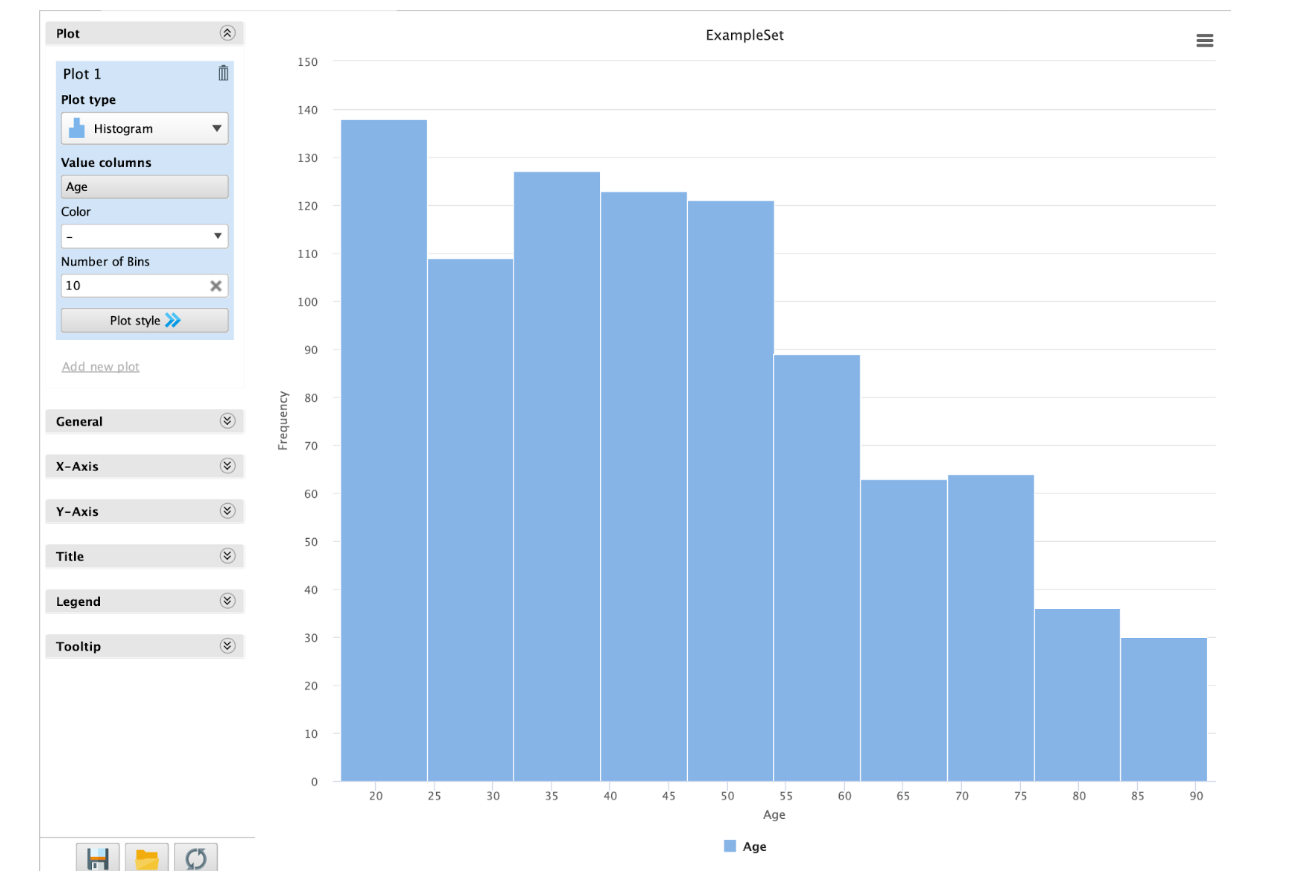


**Univariate Analysis**

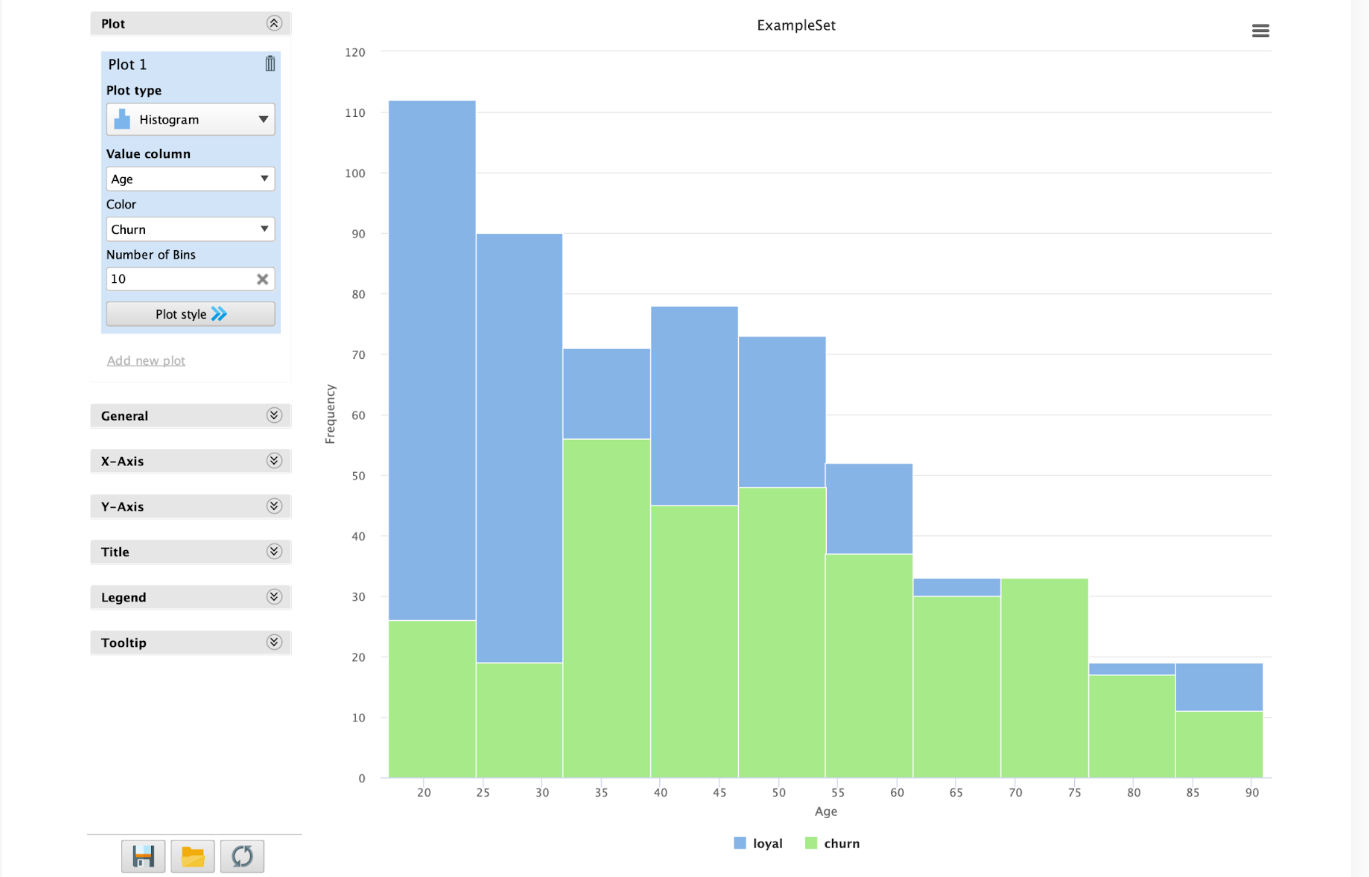
Univariate analysis means the analysis of one variable or one feature in a given data set. It tells us how data in each feature (or attribute) is distributed and also tells us about central tendencies like mean, median, and mode.

**Histograms**

One way to visualize the frequency distributions of an integer type feature is by using Histogram charts. In our customer churn dataset, below we can see the distribution of Age. Here it indicates the occurrences of each age bin -- in this case 10 bins.



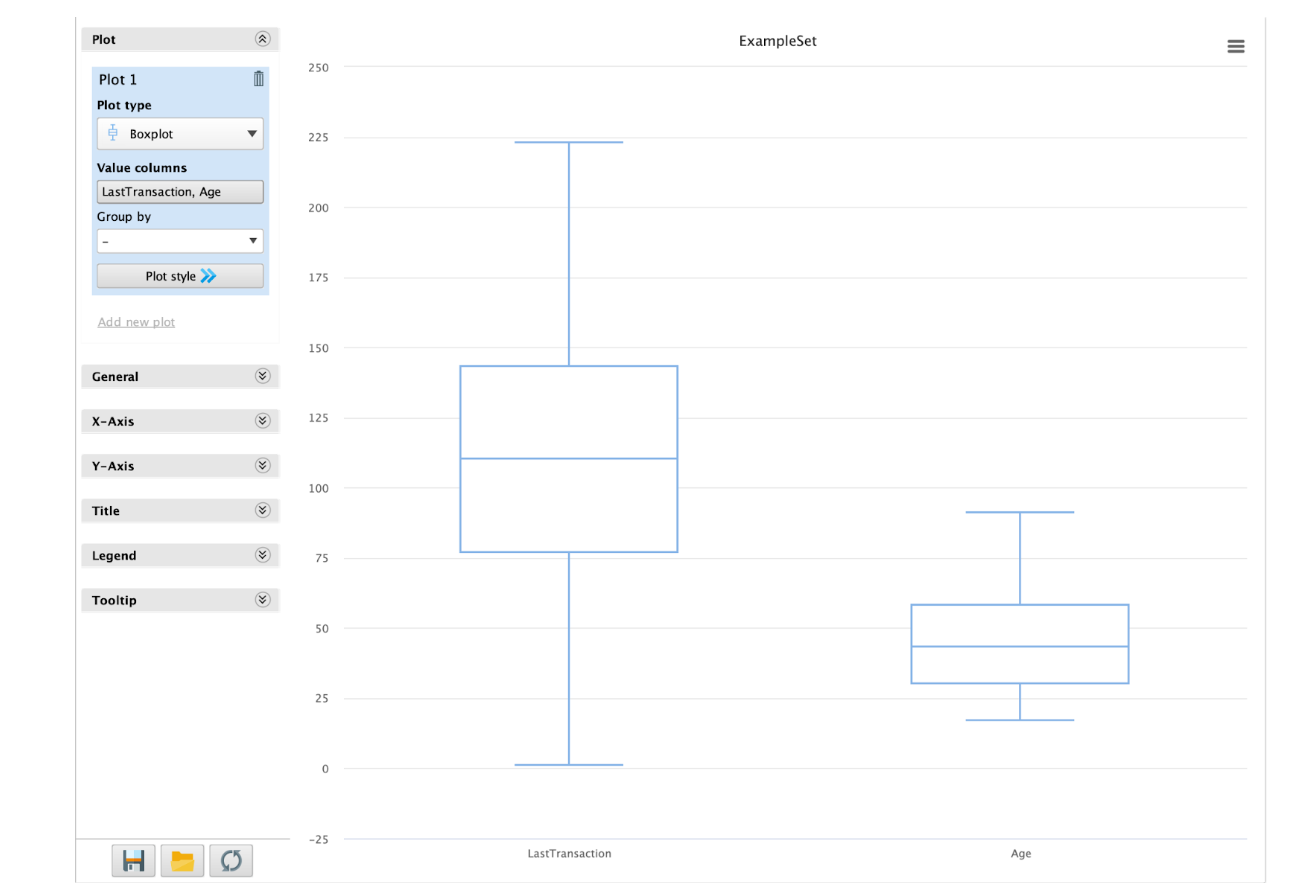
Now let’s say we want to visualize the above frequency distribution with the number of people that are loyal or churned in each age group/bin. In that case, we can set the “Color” setting as Churn. Below is the resulted visualization.



**Box plots**

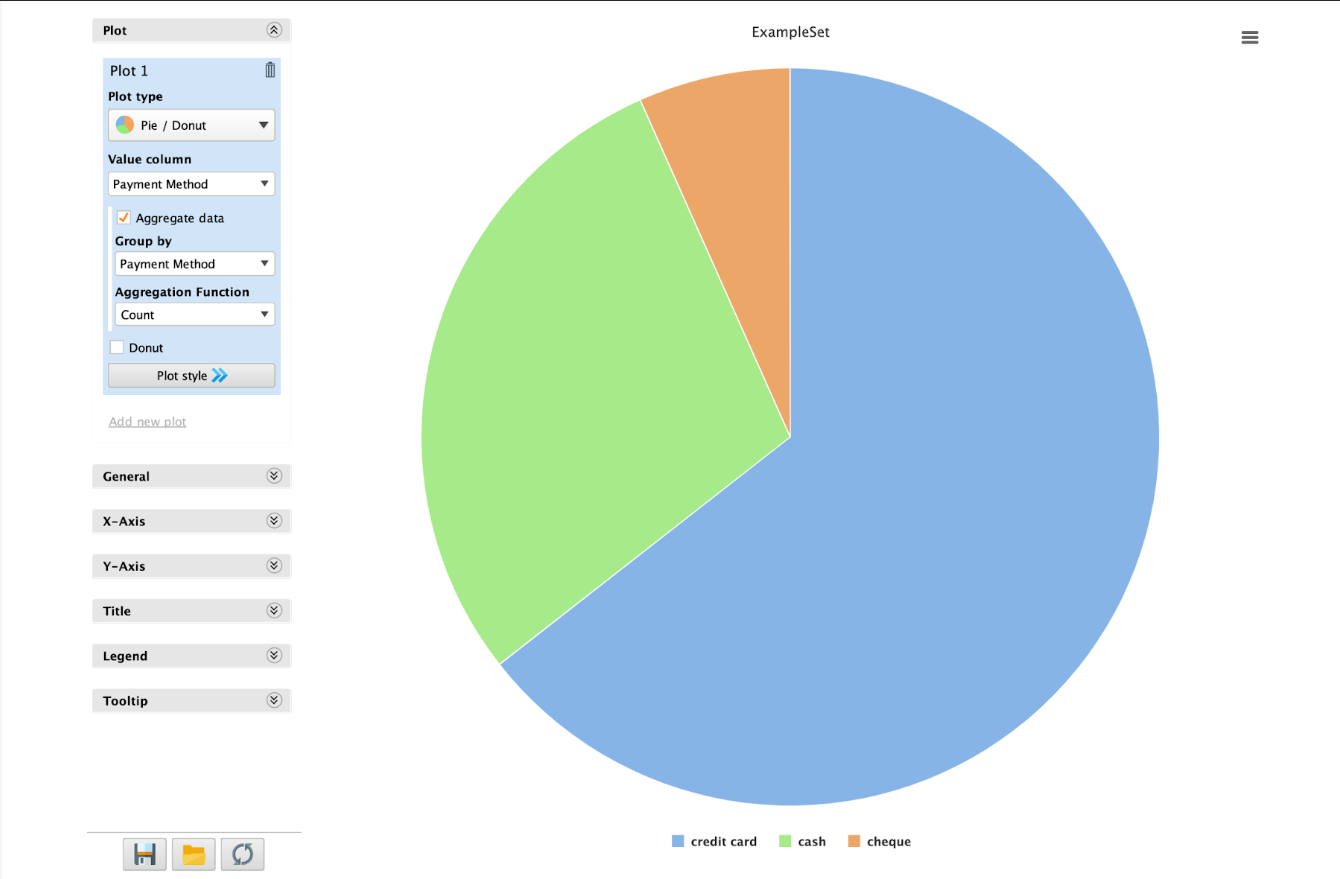
Histograms and box plots are quite similar as the purpose of both is to help visualize and describe numeric/integer data attributes. While histograms are better in determining the underlying distribution of any given attribute, box plots allow us to compare multiple attributes better than histograms as they are less detailed and take up less space on a common scale.

Below is a box plot visualization of the Last Transaction and Age attributes in the customer churn dataset.



**Pie charts**

Pie charts come to use when we want our viewers to have a general understanding of the part-of-whole relationship of a given attribute, and comparing the slices to each other is not so much important. As we mentioned previously, pie charts may not be the best visualization when we want to learn about the distribution of a dataset as compared to histograms or box plots, hence their rare applicability.



**For Continued Learning**

Apart from descriptive and univariate analysis, bivariate and multivariate analysis comes to use when we want to visualize the relationship between two or more attributes or features. There are many visualization techniques in the RapidMiner Studio to do such level analytics within a few clicks. Below are some types of charts to explore for continued future learning.

**Bivariate Analysis**

* Line chart for trends
* Correlation matrix
* Scatter plot and Scatter matrix (example: high correlation between gender and churn)
* Parallel coordinates

**Multivariate Analysis**

* Scatter plot - three dimensions (Example: Churn, Gender, Age)

**CHAPTER 6: DATA VALIDATION**

**6.1 Intro to Data Validation**

**6.1.1 Importance of Data Validation**

Data validation ensures the accuracy and quality of data used in analytics and machine learning models. It is crucial for preventing errors, ensuring reliability, and making informed decisions based on data. Proper validation helps avoid misleading results and ensures that models are built on accurate and relevant data.

**6.1.2 Types of Data Validation**

* **Syntax Validation:** Checks if data is in the correct format (e.g., date formats, numerical ranges).
* **Semantic Validation:** Ensures data values make sense in context (e.g., a negative age value).
* **Referential Integrity:** Verifies that data relationships are consistent (e.g., foreign key constraints in databases).
* **Business Rules Validation:** Ensures data adheres to predefined business rules (e.g., valid product codes or transaction limits).

**6.1.3 Data Validation Techniques**

* **Validation Rules:** Define specific criteria data must meet (e.g., range checks, format checks).
* **Data Profiling:** Analyzes data to understand its structure, content, and quality.
* **Outlier Detection:** Identifies data points that deviate significantly from the norm.
* **Data Cleaning:** Corrects or removes inaccurate or incomplete data.

**6.2 Cross Validation**

**6.2.1 Importance of Cross Validation**

Cross validation is a technique used to assess the performance and generalizability of a model. It helps ensure that the model performs well on unseen data and is not overfitted to the training data. Cross validation provides a more reliable estimate of model performance compared to a single train-test split.

**6.2.2 Types of Cross Validation**

* **K-Fold Cross Validation:** Splits the dataset into K subsets (folds), training the model K times, each time using a different fold as the test set and the remaining K-1 folds as the training set.
* **Leave-One-Out Cross Validation (LOOCV):** A special case of K-Fold where K equals the number of data points, with each point being used once as a test set.
* **Stratified Cross Validation:** Ensures that each fold maintains the same proportion of class labels as the original dataset, particularly useful for imbalanced datasets.

**6.2.3 Process of Cross Validation**

1. **Split Data:** Divide the dataset into K folds or subsets.
2. **Train and Test:** For each fold, train the model on K-1 folds and test it on the remaining fold.
3. **Evaluate Performance:** Collect performance metrics for each fold and average them to get the overall model performance.
4. **Select Model:** Choose the model with the best performance metrics based on cross-validation results.

**6.3 Validating a Model**

Validating a model involves assessing its performance using various metrics and techniques to ensure it meets the desired objectives and performs well on new data. Common metrics include accuracy, precision, recall, F1 score, and ROC-AUC for classification tasks, and mean squared error (MSE), root mean squared error (RMSE), and R-squared for regression tasks.

**6.4 How to Correctly Validate Machine Learning Models**

* **Use Appropriate Metrics:** Choose metrics that align with the specific goals of the model and the problem domain.
* **Perform Cross Validation:** Apply techniques like K-Fold or LOOCV to assess model performance across different data subsets.
* **Check for Overfitting:** Ensure the model generalizes well to unseen data by comparing training and validation performance.
* **Consider Data Leakage:** Ensure that information from the test set does not inadvertently influence the training process.
* **Assess Model Robustness:** Test the model’s performance under different conditions and scenarios to ensure its reliability.

**CHAPTER 7: PROJECT**

**7.1 Introduction**

This project aims to demonstrate the capabilities of Altair AI Studio (formerly known as RapidMiner Studio) in handling various data science tasks. The project involves importing a dataset, visualizing the data, building processes, and performing data analysis techniques such as correlation and linear regression. The objectives of this project are to explore the features of Altair AI Studio, apply data science methodologies, and interpret the results obtained from the analysis.

**7.2 Tools and Technologies Used**

**7.2.1 Altair AI Studio Overview**

* Formerly known as RapidMiner Studio.
* An approachable, end-to-end data science platform.
* Designed for users of all levels, from beginners to professionals.

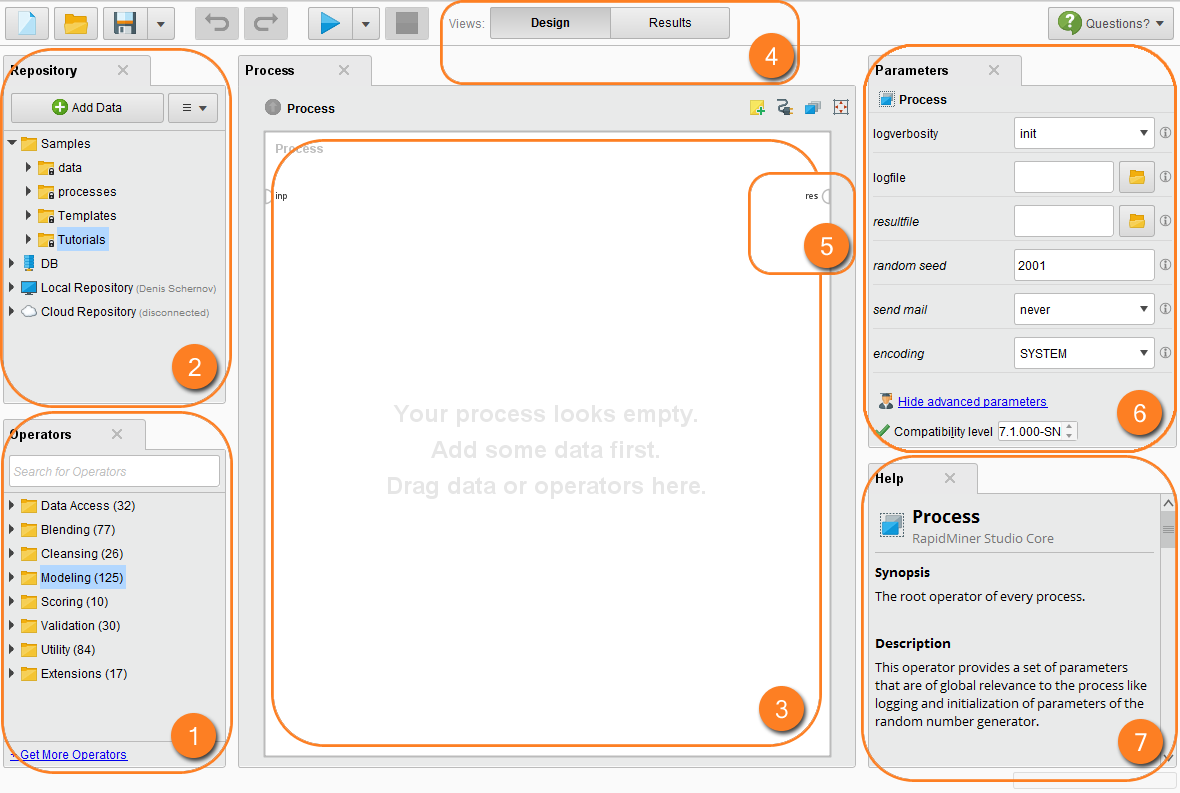


Fig. 7.2

1.operators 2.Repository 3.process panel 4. Views 5.Ports

6. Parameters 7.help

**7.2.2 Key features**

* User-Friendly Interface: Intuitive drag-and-drop functionality.
* Comprehensive Tools: Supports data preparation, modeling, and deployment.
* Robust Community Support: Extensive documentation, tutorials, and active user forums.

**7.3 Project Dataset**

**7.3.1 Data set introduction**

The Toyota Corolla dataset provides a comprehensive overview of 1,436 Toyota Corolla cars, encompassing a wide range of attributes related to the vehicles' specifications, features, and pricing. This dataset is valuable for analyses in areas such as car pricing, feature importance, and vehicle comparison.

**7.3.2 Dataset Overview**

The dataset contains 37 columns, each representing a different attribute of the cars. The attributes can be broadly categorized into:

**Identification and Model Information**:

**Id**: A unique identifier for each car.

**Model**: The specific model of the Toyota Corolla.

**Pricing and Age**:

**Price**: The selling price of the car (in euros).

**Age\_08\_04**: The age of the car in months as of August 2004.

**Manufacturing Details**:

**Mfg\_Month**: The manufacturing month.

**Mfg\_Year**: The manufacturing year.

**Performance and Specifications**:

**KM**: The total kilometers driven.

**HP**: Horsepower of the vehicle.

**cc**: Engine size in cubic centimeters.

**Doors**: Number of doors.

**Cylinders**: Number of engine cylinders.

**Gears**: Number of gears in the transmission.

**Weight**: The car's weight in kilograms.

**Fuel and Transmission**:

**Fuel\_Type**: Type of fuel used (Diesel, Petrol, etc.).

**Automatic**: Indicator of whether the car has an automatic transmission.

**Features and Accessories**:

**Met\_Color**: Presence of metallic color (1 for Yes, 0 for No).

**ABS**, **Airbags**, **Airco**: Safety and comfort features.

**CD\_Player**, **Radio**, **Boardcomputer**: Entertainment and convenience features.

**Power\_Steering**, **Powered\_Windows**: Functional features.

**Guarantees and Taxes**:

**Mfr\_Guarantee**, **BOVAG\_Guarantee**: Details on manufacturer and BOVAG guarantees.

**Quarterly\_Tax**: Quarterly road tax payable.

The Toyota Corolla dataset is rich in information and provides a detailed snapshot of various aspects of these vehicles. It serves as an excellent resource for analyzing the factors influencing car prices and understanding the distribution of features across different models. The dataset can be used for predictive modeling, correlation analysis, and other statistical evaluations to draw meaningful insights about the automotive market.

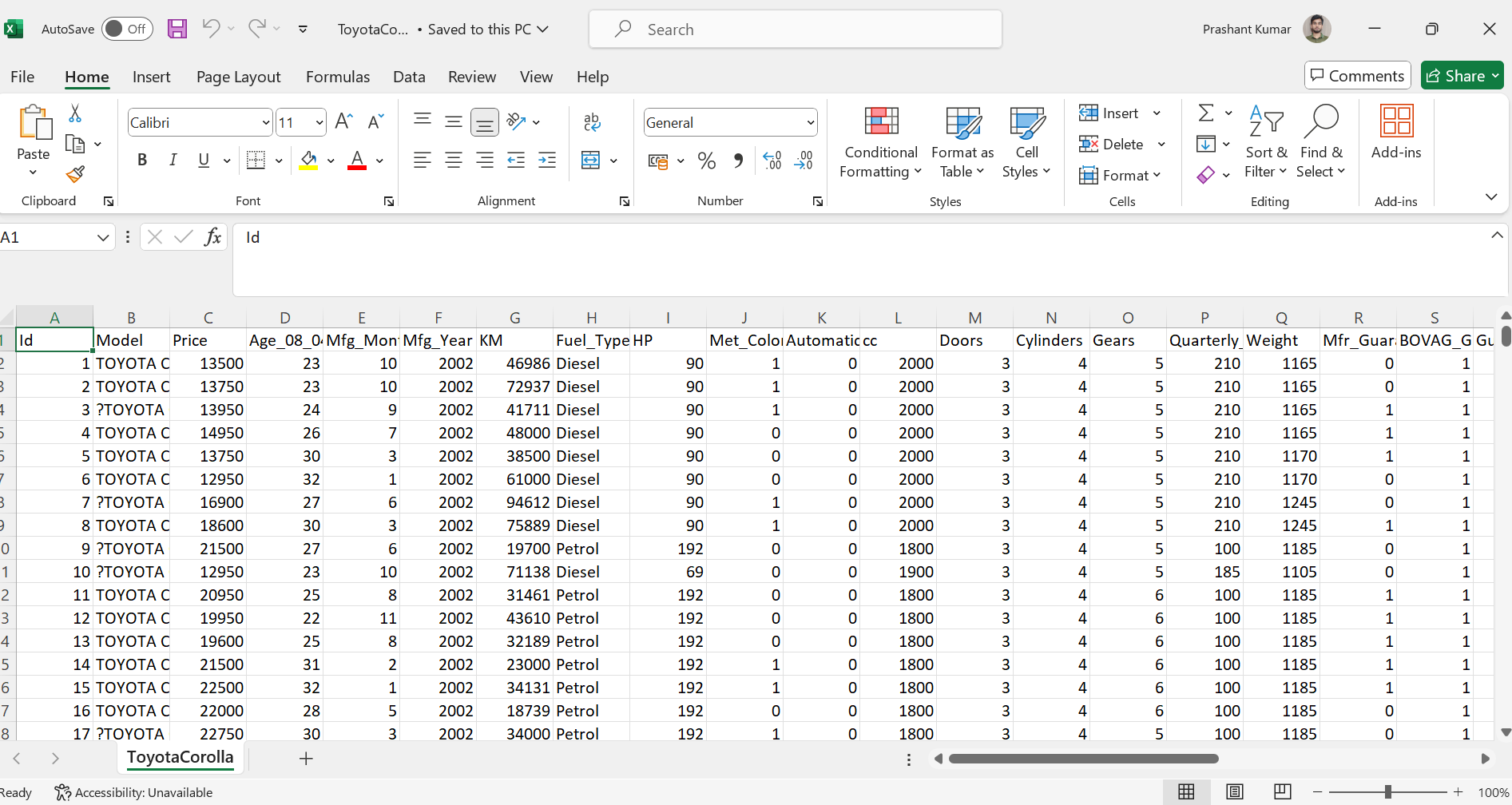


Fig. 7.3.2

**7.4 Data Engineering**

**7.4.1 Importing the data into AI studio**

Step1: open the Altair AI studio app

Step 2: on the left side of the in the repository section, click on “import data”

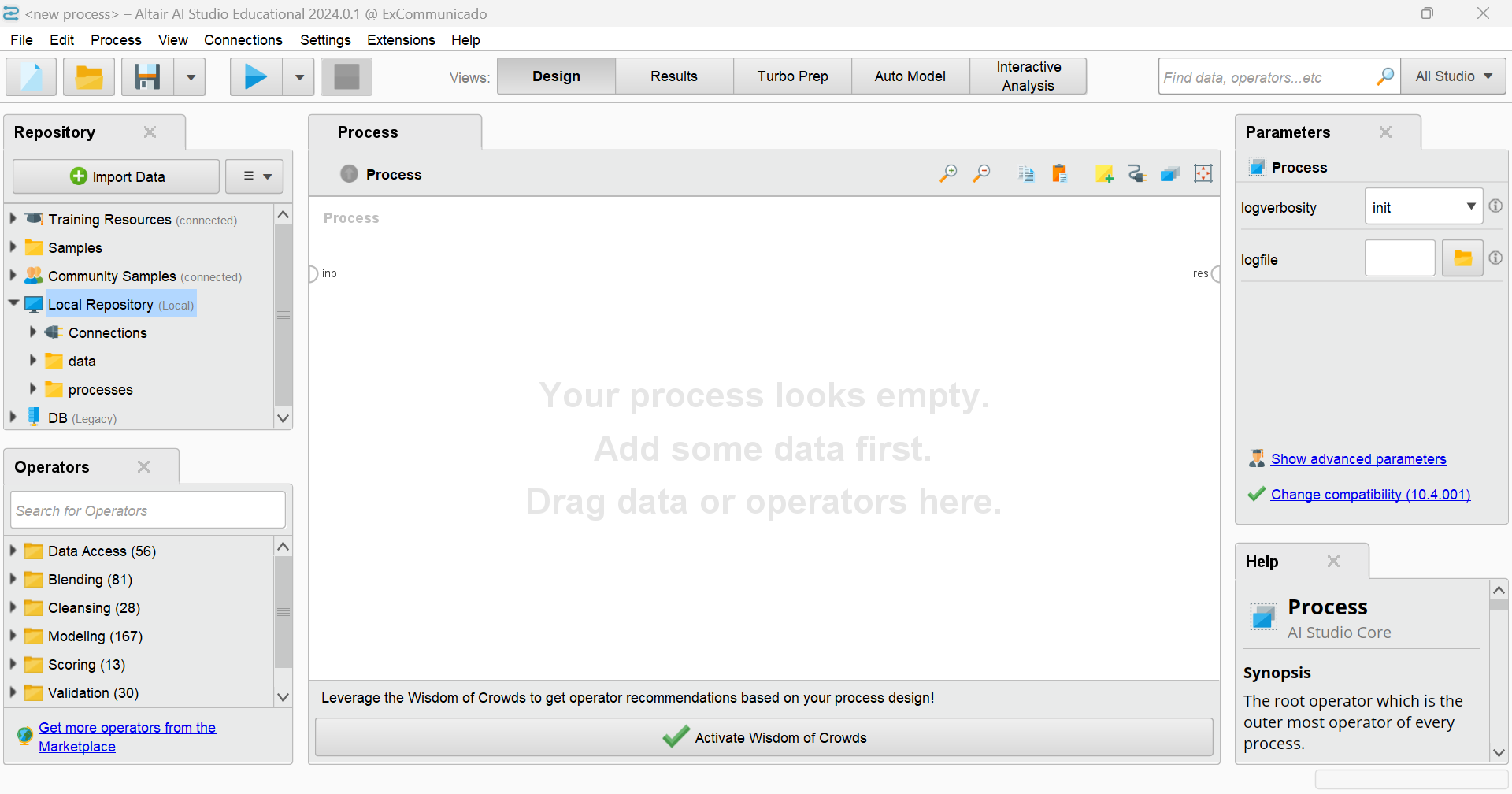
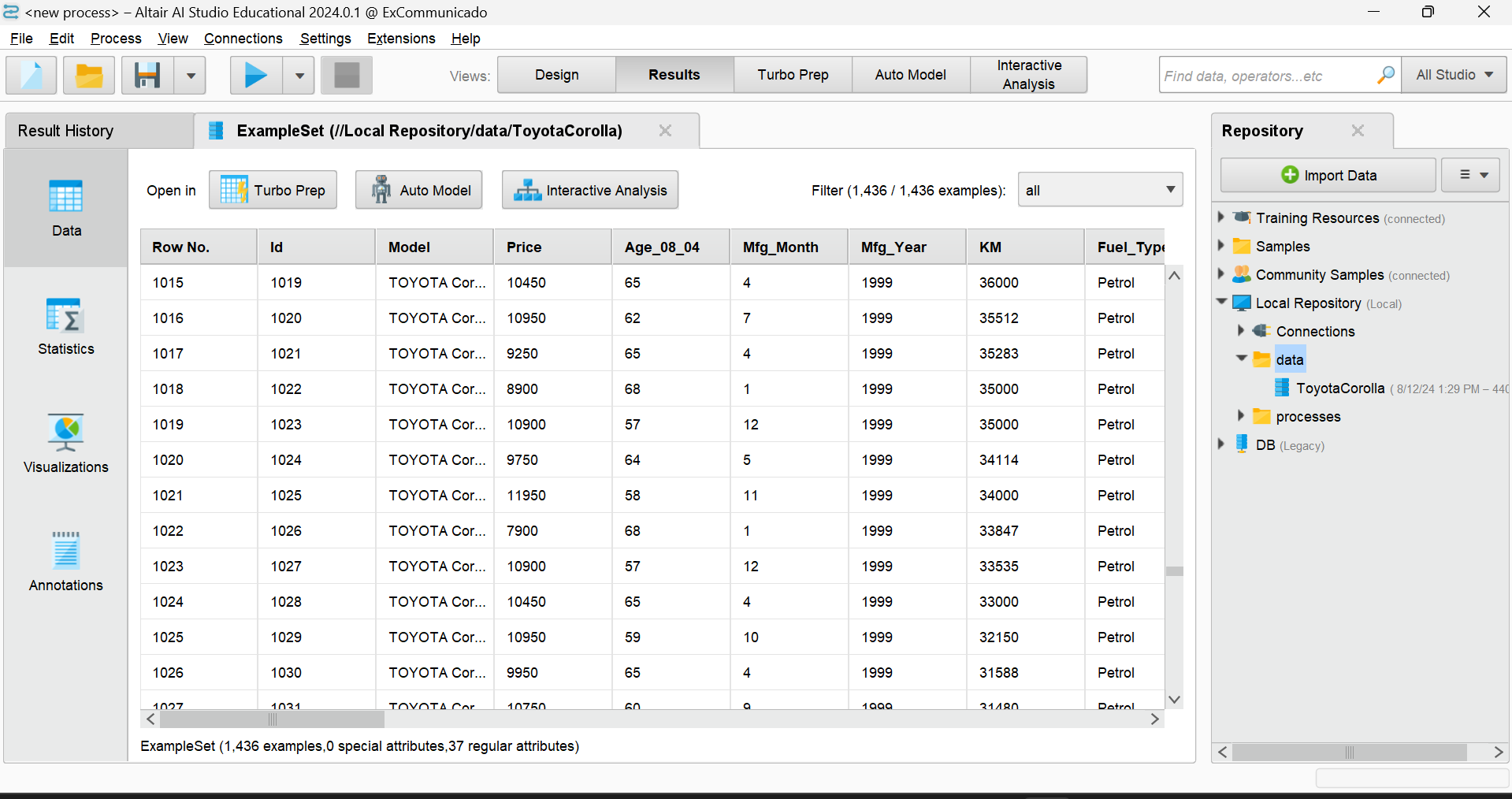


Fig. 7.4

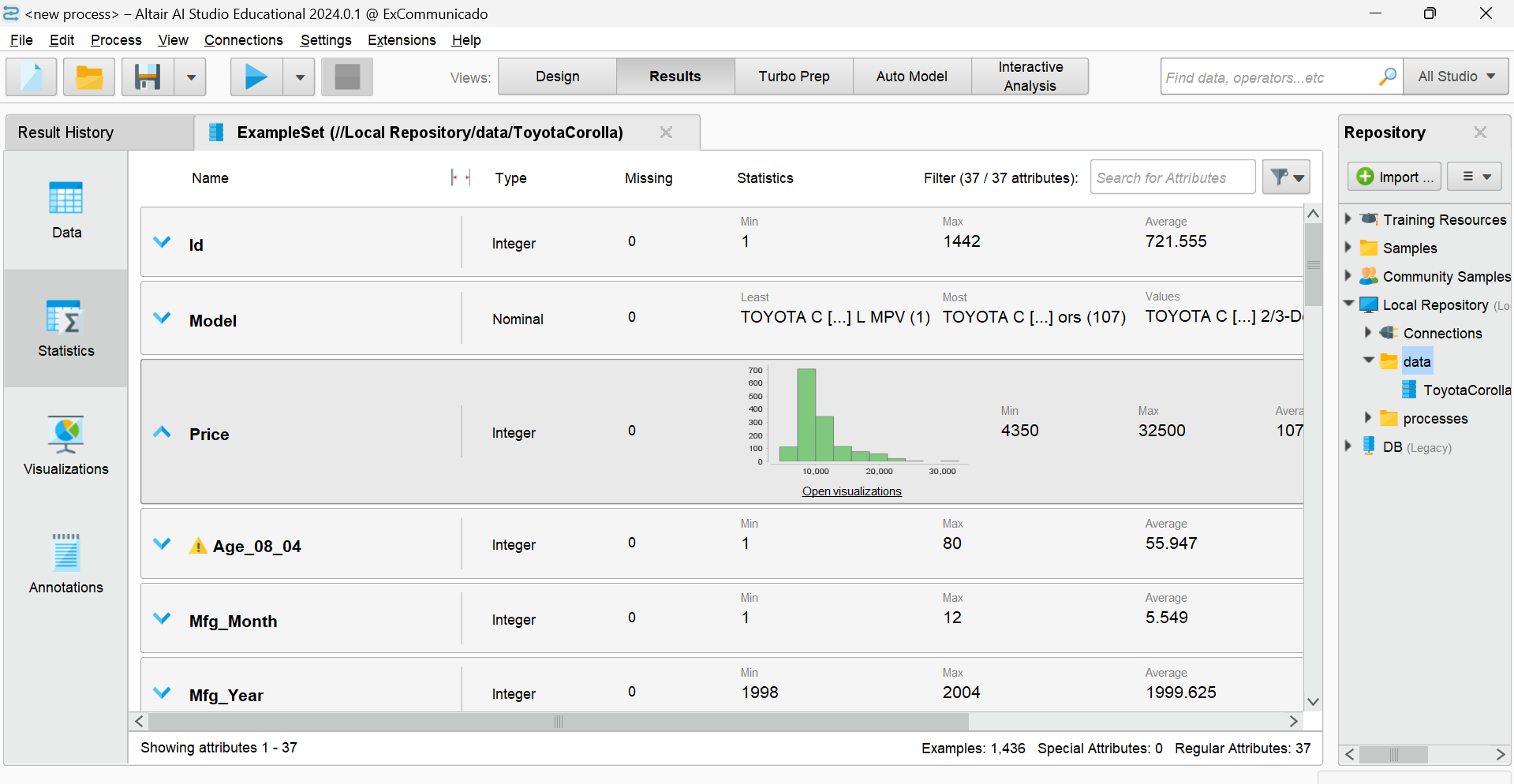
Step 3: select your datasets from the required location and continue..

Step 4: after importing, the screen will look like below

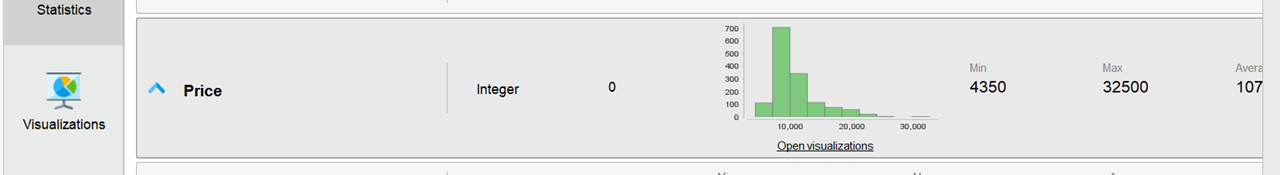


**7.4.2 Data statistics**

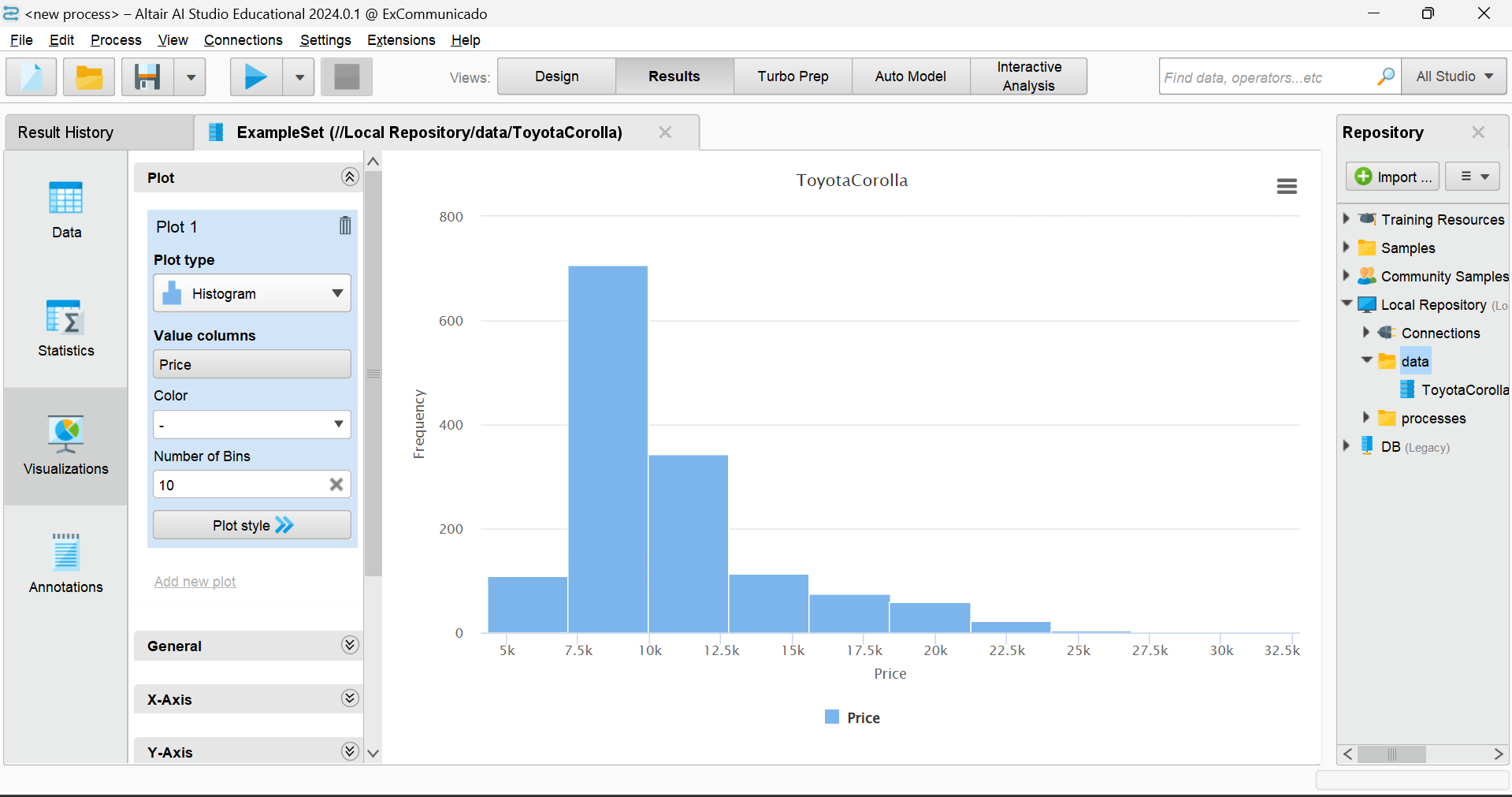
After clicking on the statistics section, we get something like this



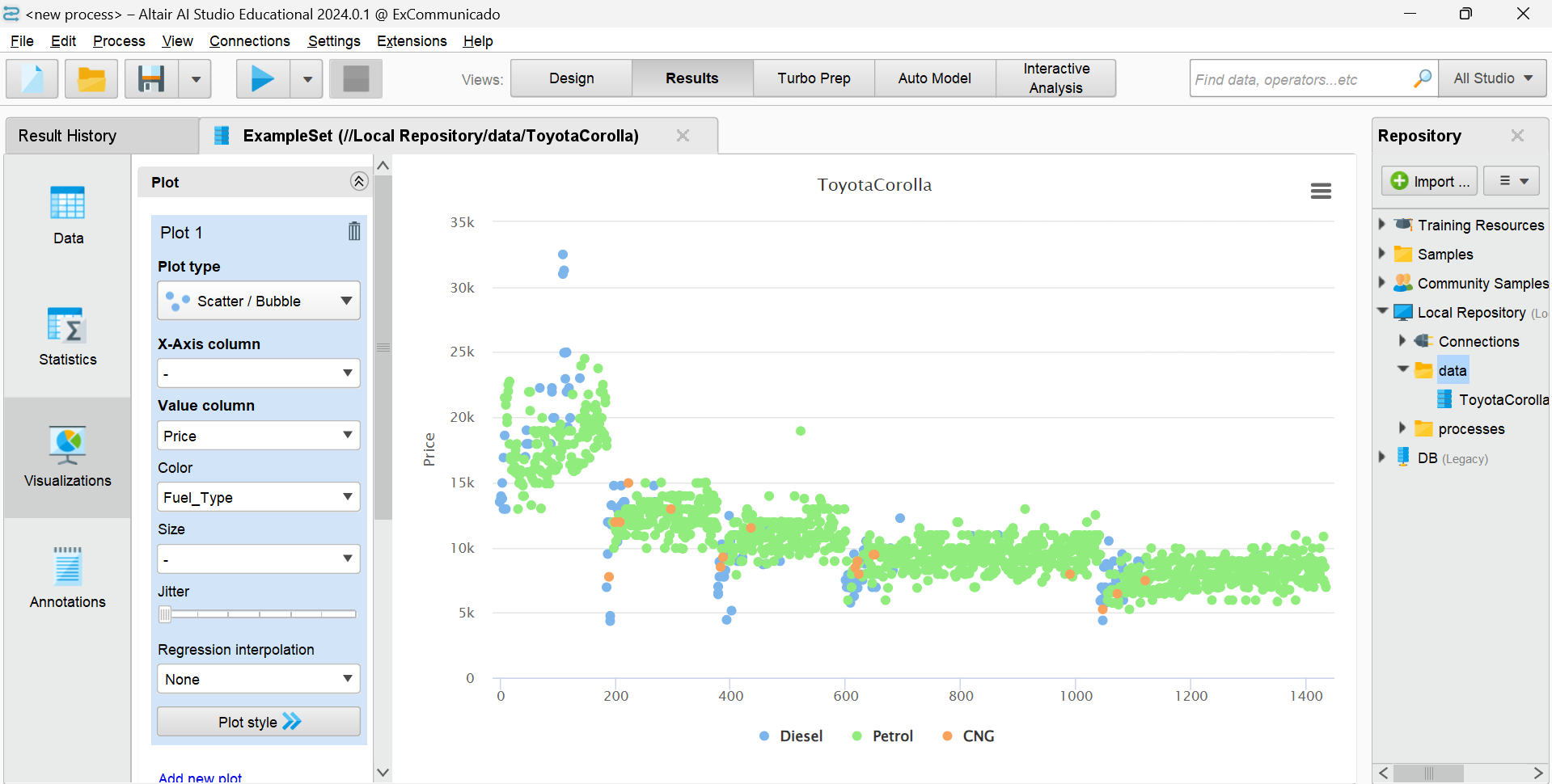
**7.4.3 Data Visualisation**



Click on open visualisation, it appears like below

****

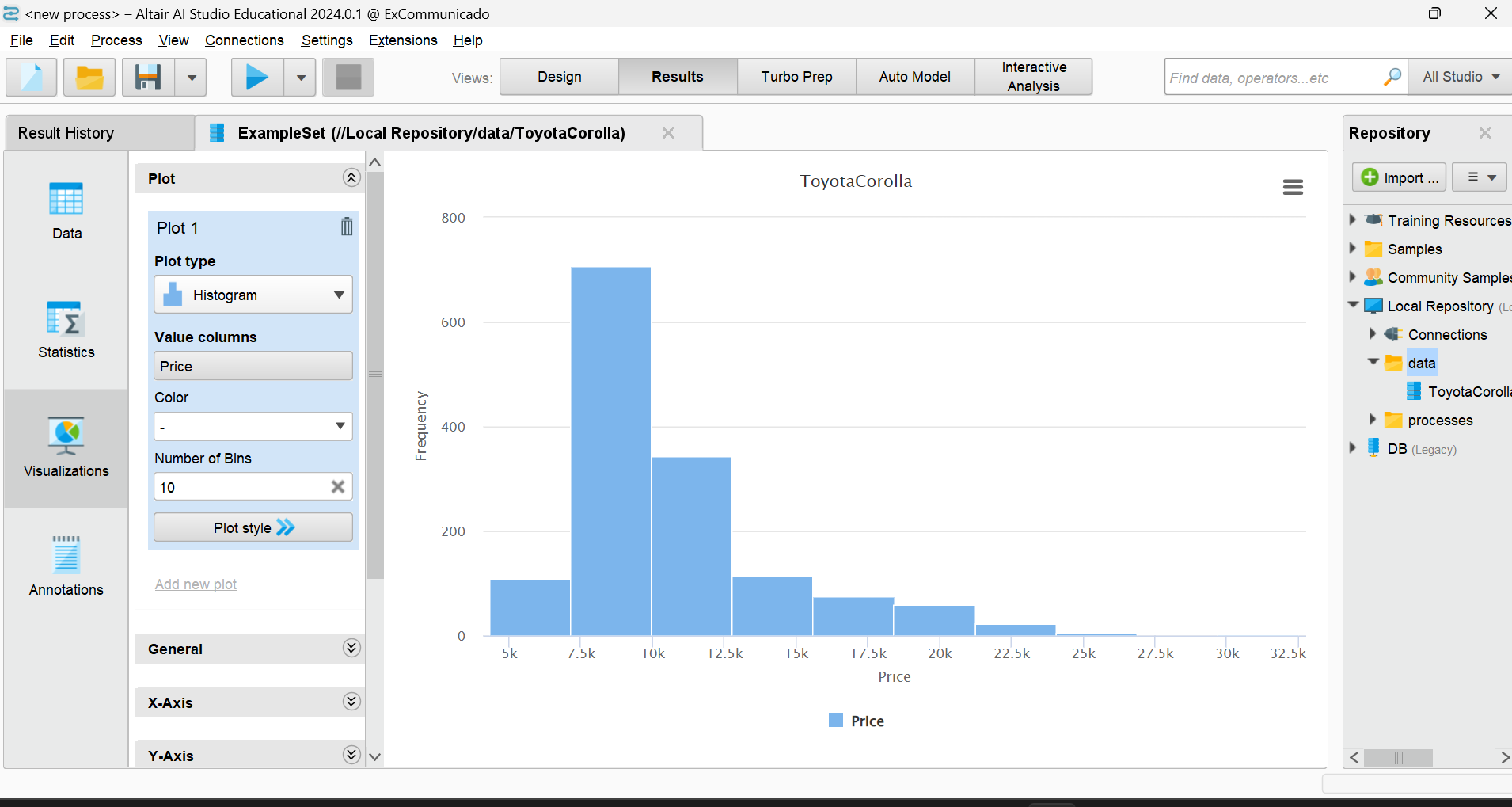
And one more example is given below

****

The above is the scatter plot which shows that we can use rapidminer, altair AI studio, for plotting also.

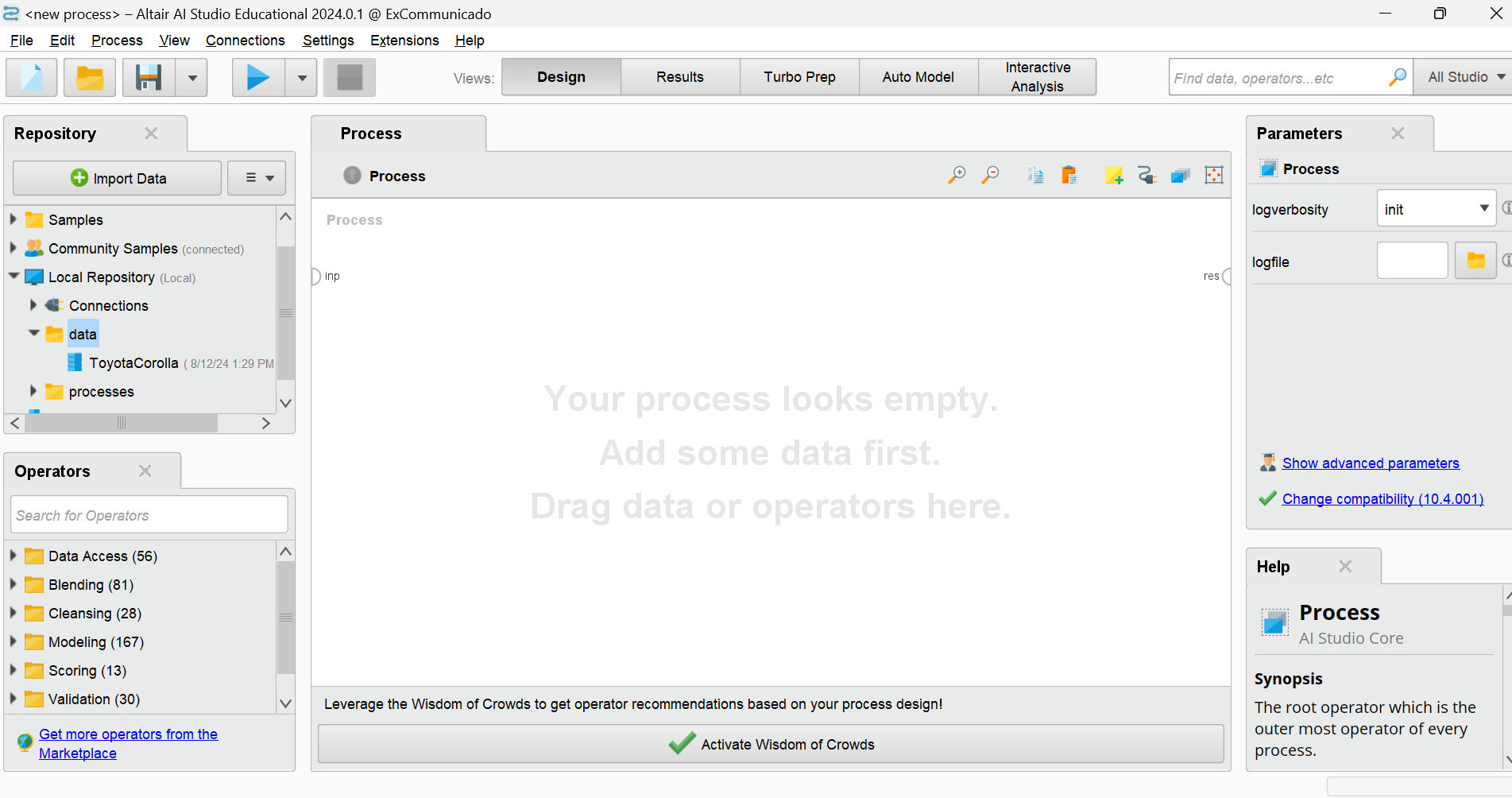
**7.5 Building and running a simple process**

In the Altair AI studio, click on design view



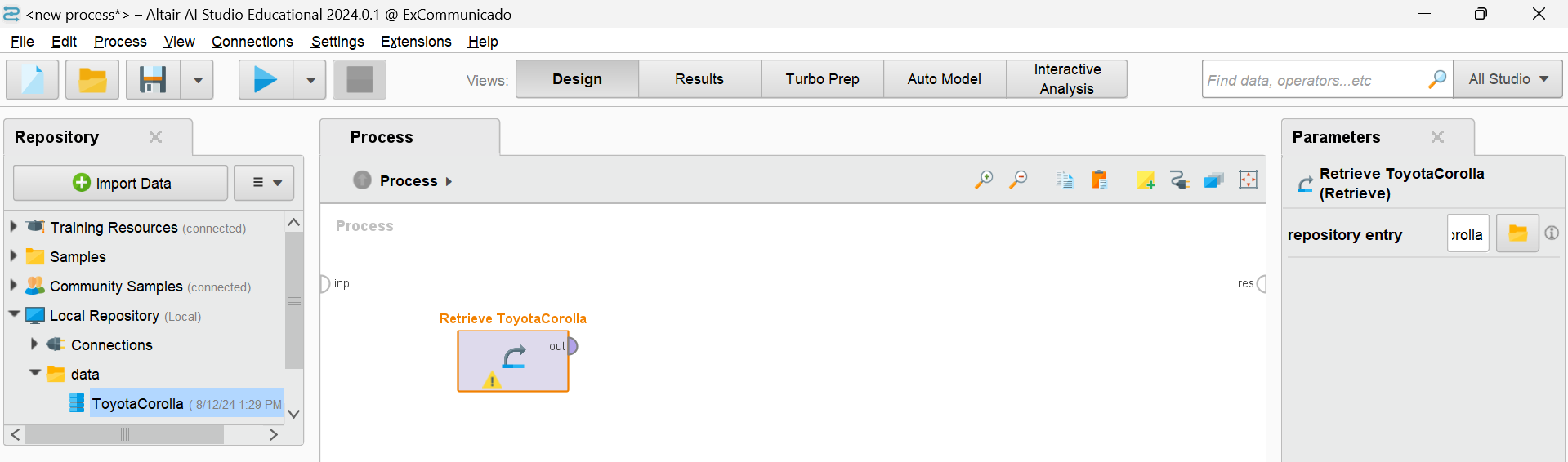
Click here

And then drag the file from repository to the process like below



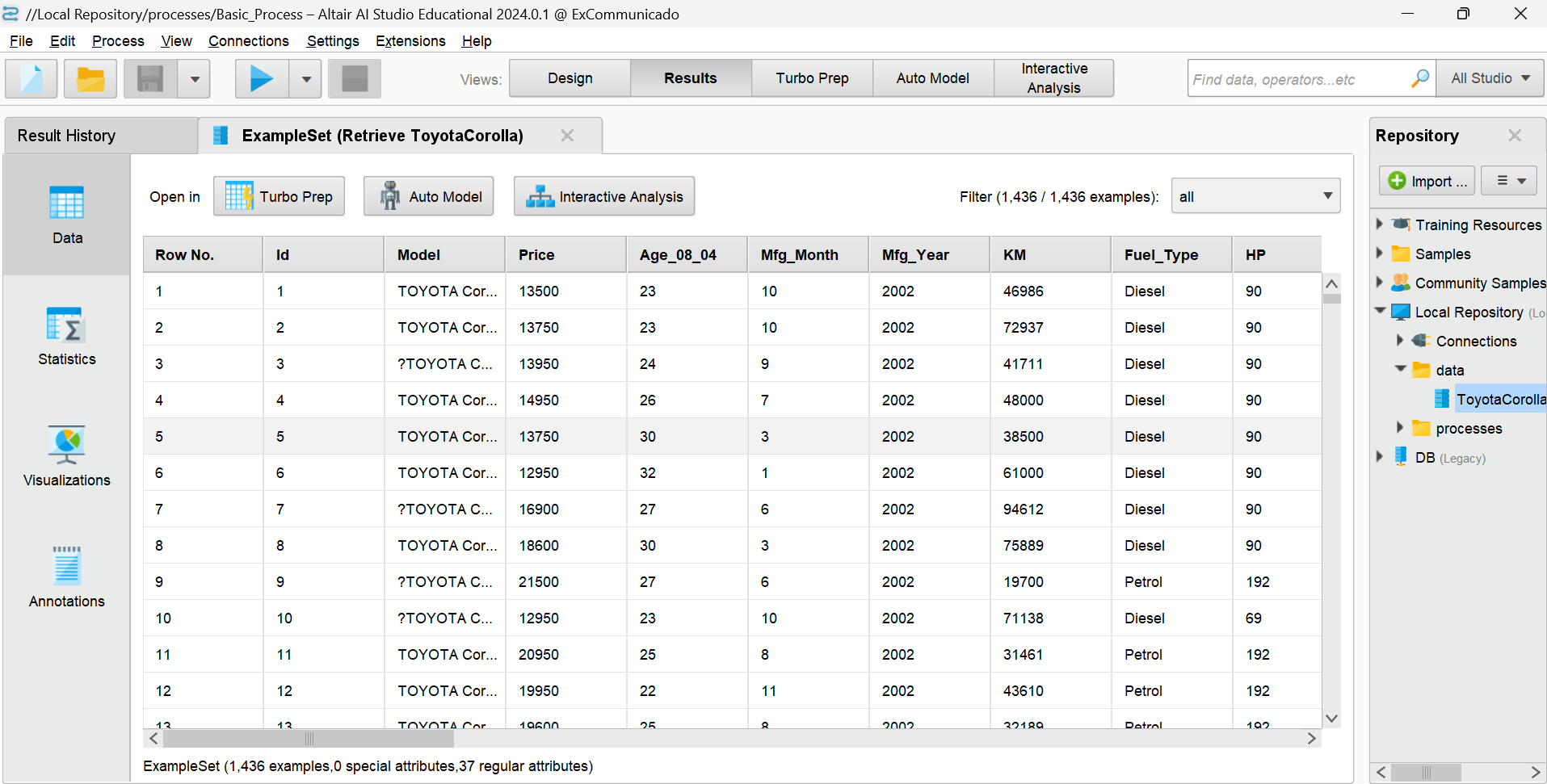
DRAG this file to here

After dragging below is how it looks like

You can see the out (output) and res (result), join that with by dragging after clicking “out” to the “res” and save it with any name and hit run on the top (this is what we call a running a process)

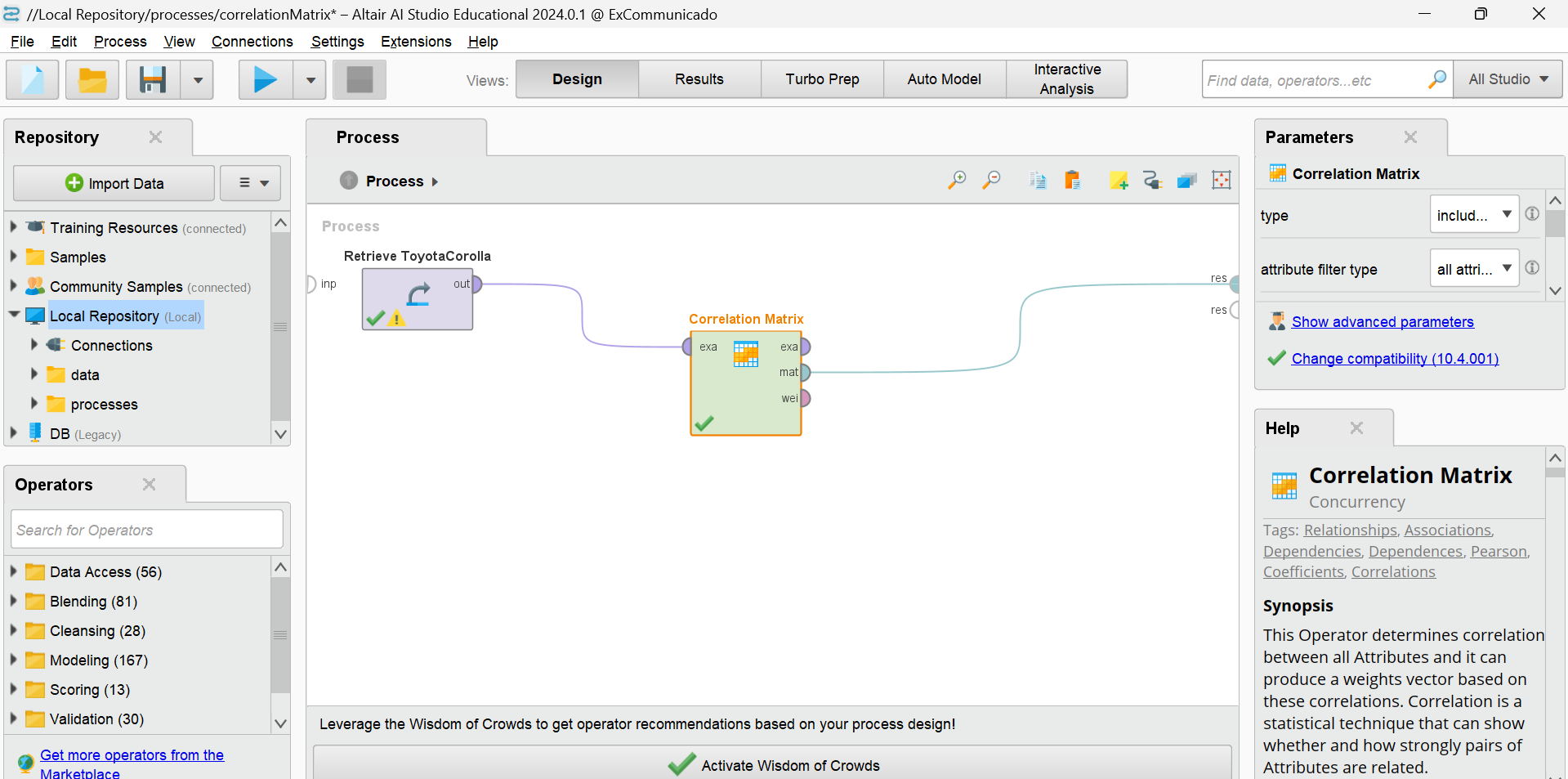
After running

The screen looks like



So this is how we build and run the process in ALTAIR AI studio.

**7.6 Correlation matrix process**

****

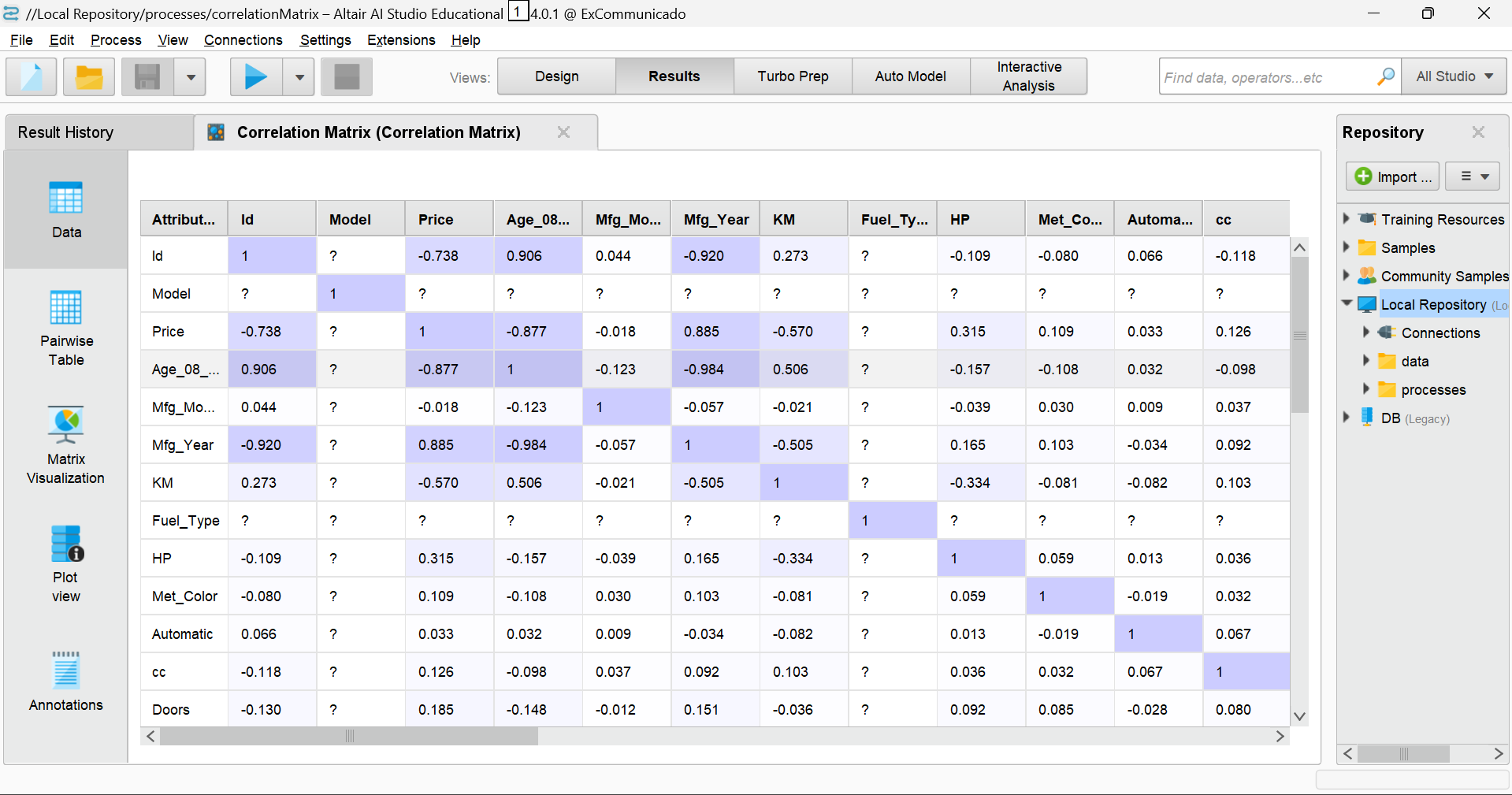
This is the process which I have created in Altair AI studio. Here I have used correlation

Matrix operator:

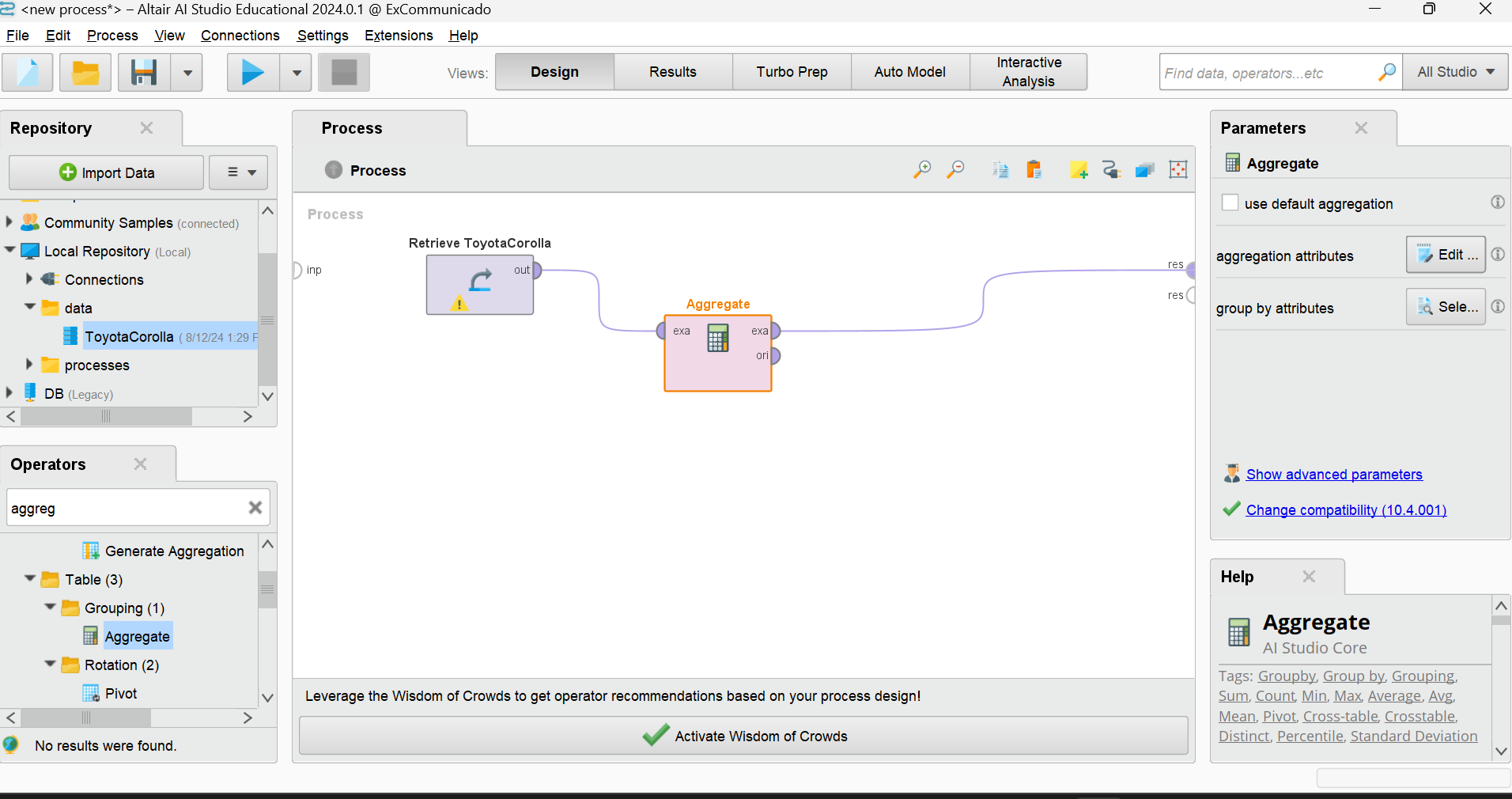
**7.6.1 correlation matrix operator**

* This Operator determines correlation between all Attributes
* It can produce a weights vector based on these correlations.
* Correlation is a statistical technique that can show whether and how strongly pairs of Attributes are related.
* A **Correlation Matrix** is a table showing correlation coefficients between variables.
* Each cell in the matrix shows the correlation between two variables
* The value ranges from
  + **-1 to 1**:
    - **+1**: Perfect positive correlation
    - **0**: No correlation
    - **-1**: Perfect negative correlation

**7.6.2 correlation matrix process output**

****

**7.7 Data Aggregation process**

****

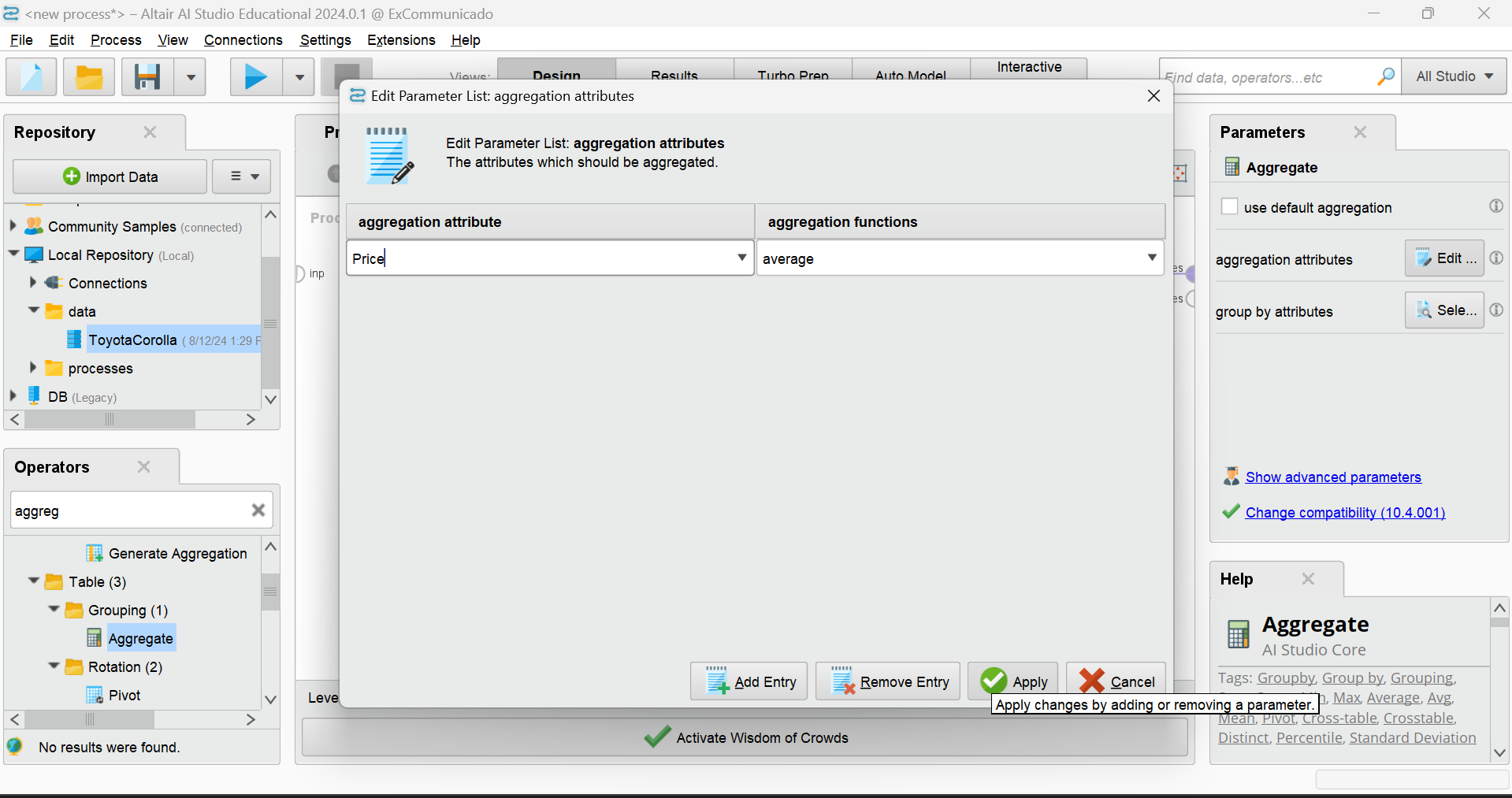
In the above figure, we can see “Aggregate” Operator . let’s talk about aggregate operator.

**7.7.1 Aggregate Operator**

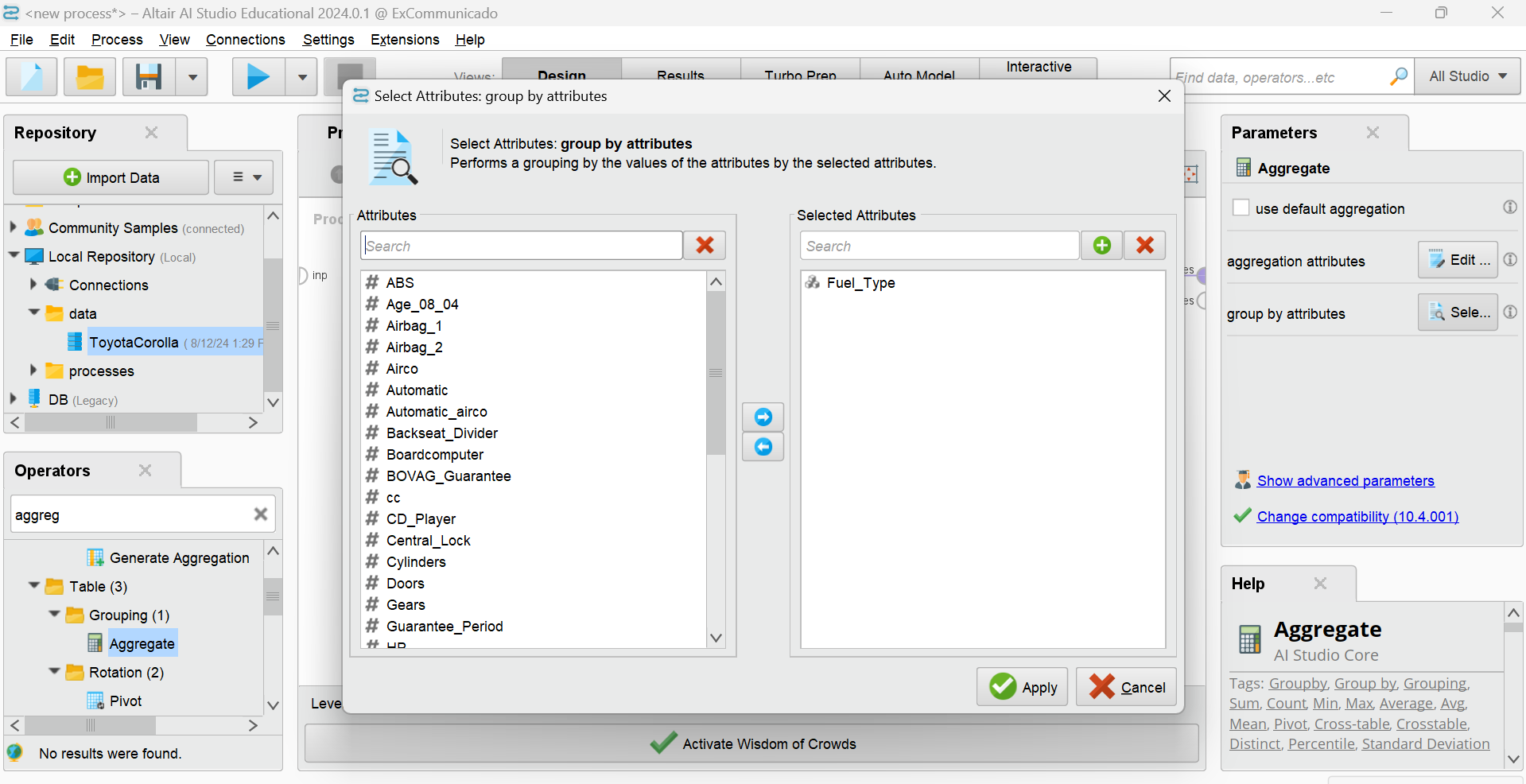
* This operator performs the aggregation functions known from SQL.
* This operator provides a lot of functionalities in the same format as provided by the SQL aggregation functions.
* SQL aggregation functions and GROUP BY and HAVING clauses can be imitated using this operator.
* We will use the aggregate operator for the data aggregation.
* Suppose we want to get the average price of the car group by fuel type
* How to do this?

The above process is made for the same.

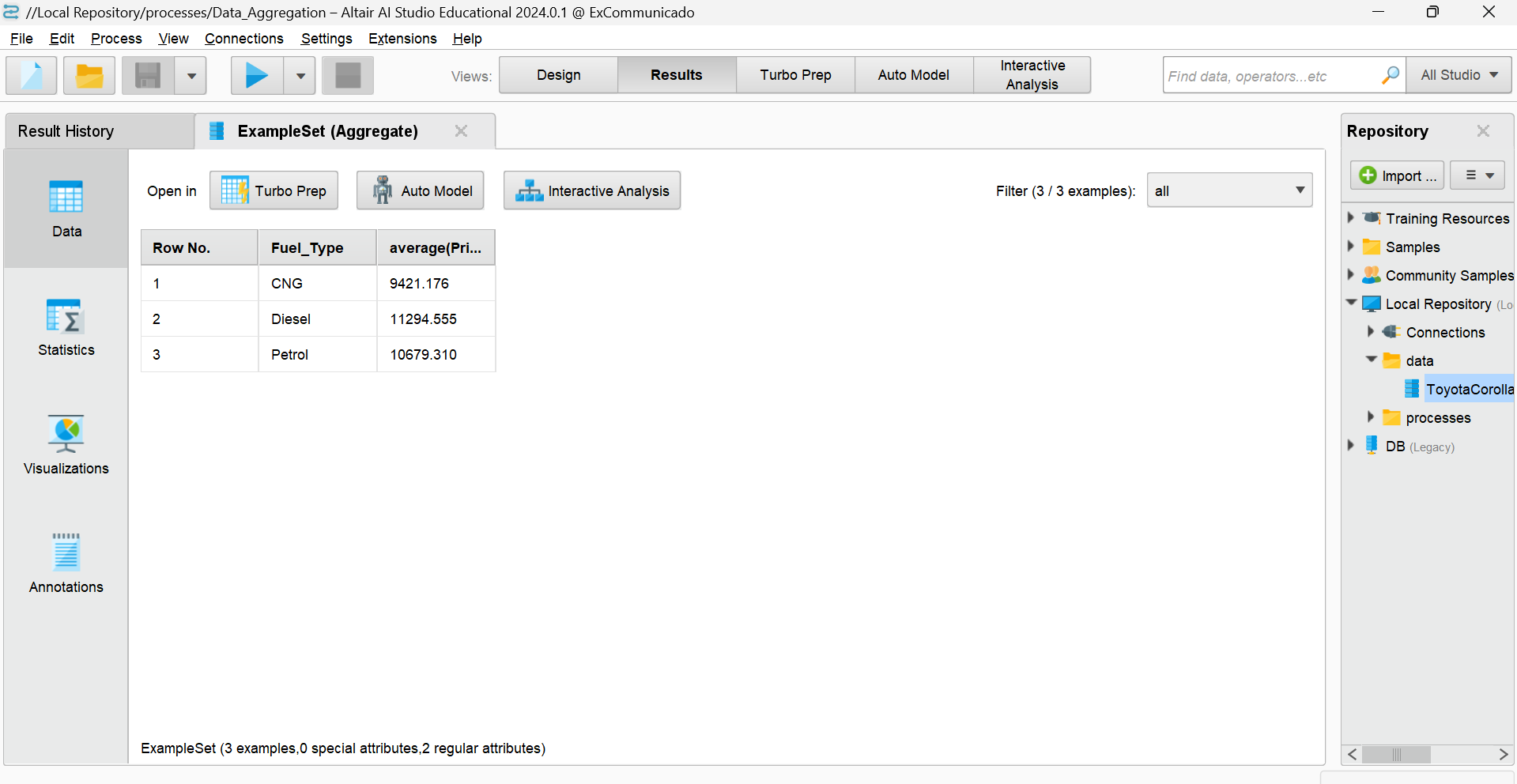
In the above figure, To the right of the process set the parameter aggregation attribute like below and click apply



And then Set the Group by to “fuel\_type” click on fuel\_type from the L.H.S and move it to the right using arrow button and click apply

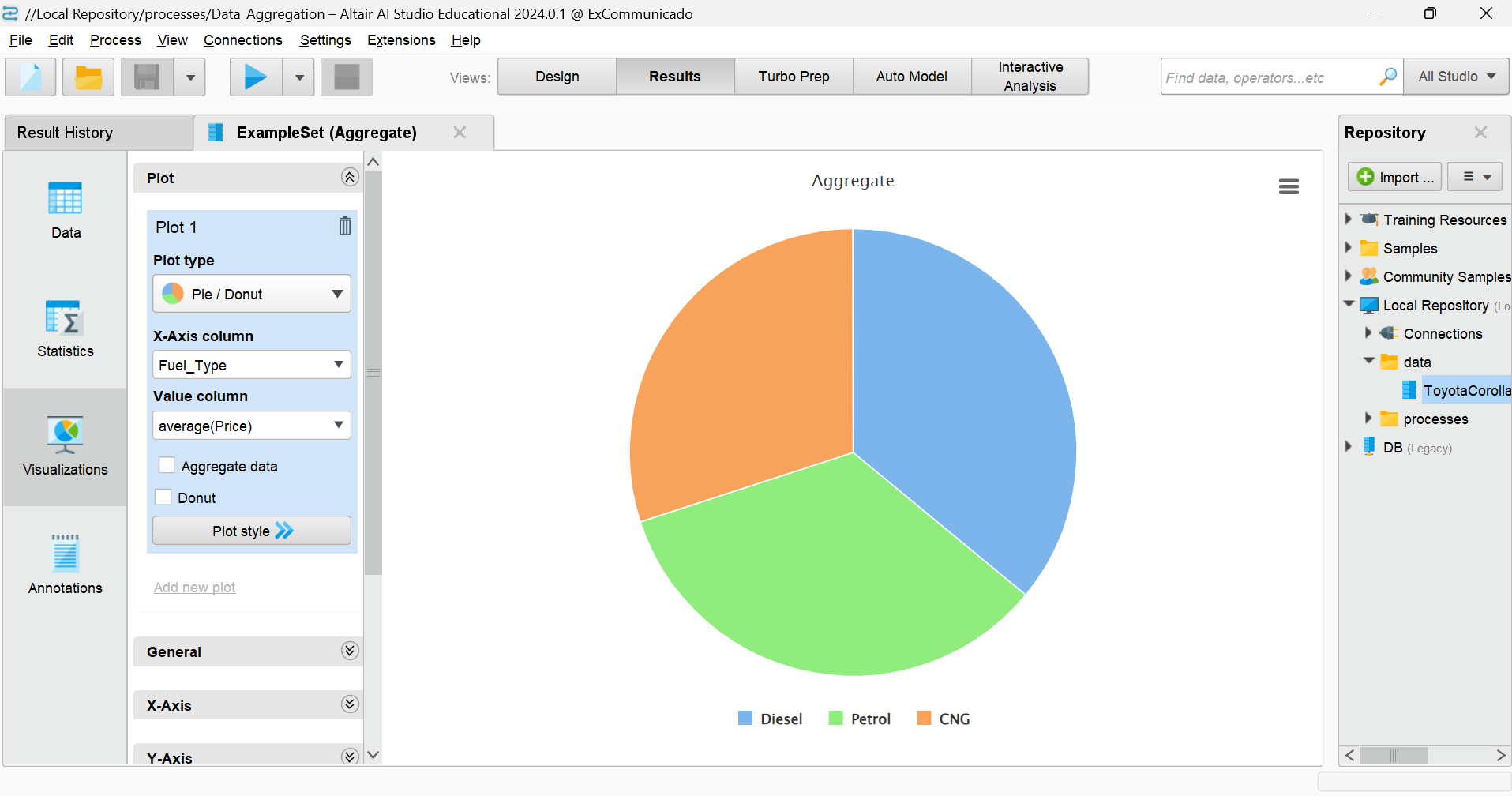


After all that, save and run…



We can see the fuel type and their average price.

Let’s represent it with the graph. On the left side, we have visualization, after clicking on it you will be able to see a plot for the same result ….(NEXT PAGE)

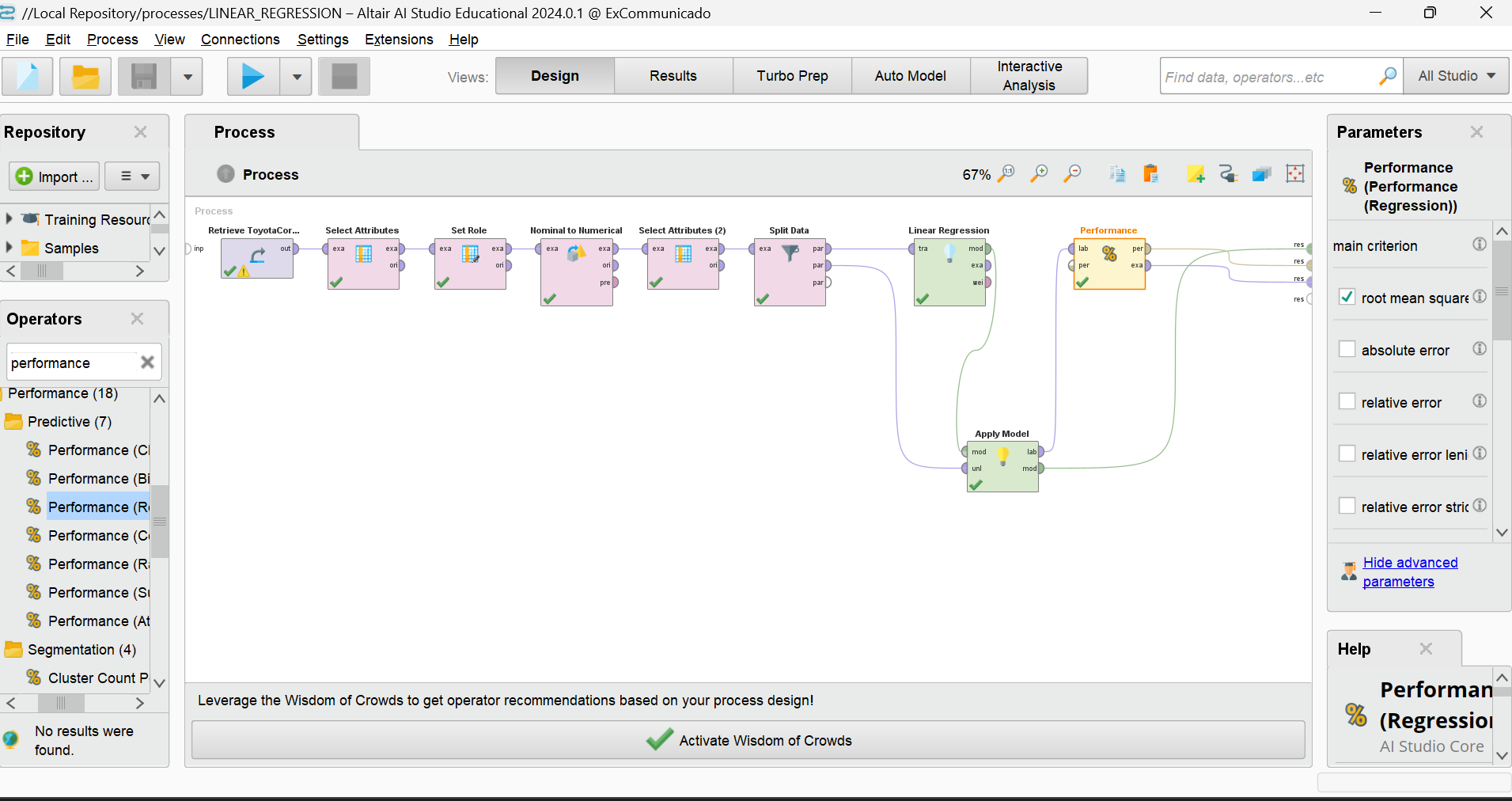


Visualisation defaultly it chose a plot type “pie chart” you can change the plot type to anything by clicking the plot type dropdown

**7.8 LINEAR REGRESSION**

* **Linear Regression** is a statistical method for modeling the relationship between a **dependent variable (Y)** and one or more **independent variables (X)**.
* It’s one of the simplest and most commonly used algorithms in **Data Science**.
* **Simple Linear Regression**: Models the relationship between two variables by fitting a linear equation to the observed data.
  + **Equation**: Y=b0+b1X+ϵ
  + b0​: Intercept
  + b1​: Slope (how much Y changes for a unit change in X)
  + ϵ: Error term (difference between observed and predicted Y)
* **Multiple Linear Regression**: Extends the model to include more than one independent variable.
  + **Equation**: Y=b0+b1X1+b2X2+...+bnXn+ϵ

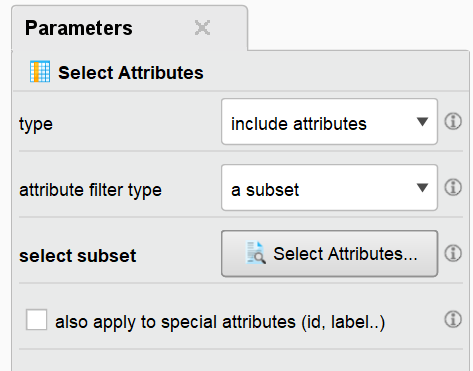
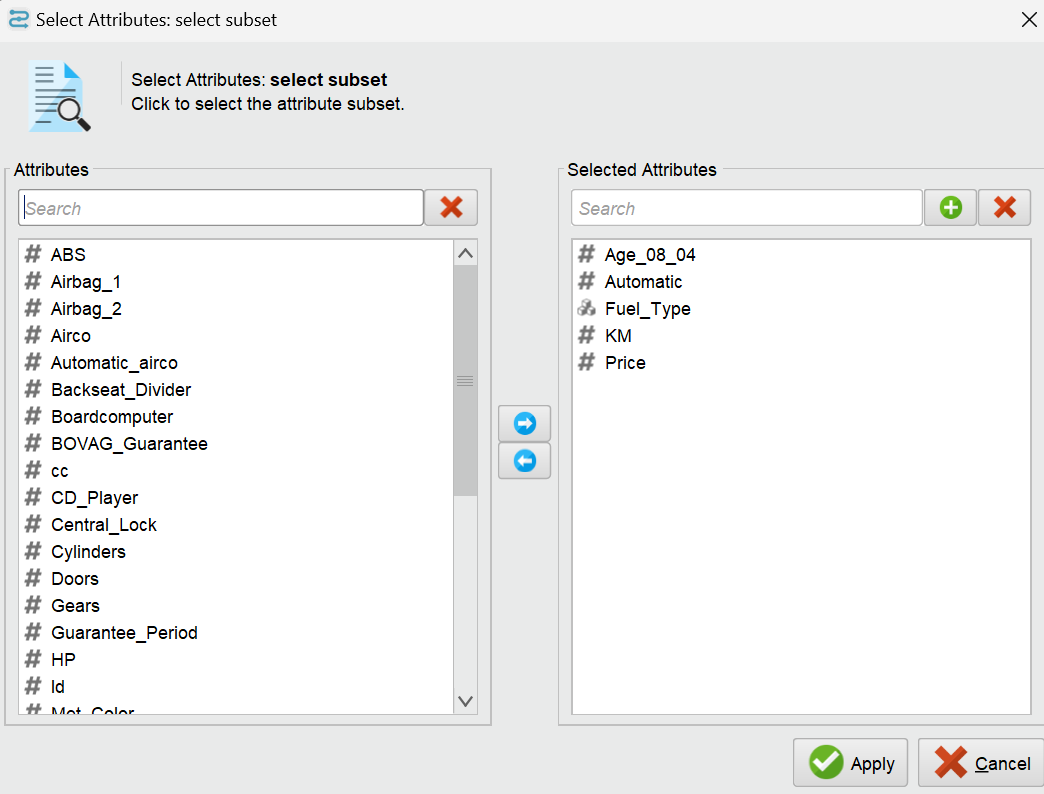
**7.8.1 Linear Regression Process**



This is what the process looks like. I have used so many operators here, names are given below:

1. **Select attribute**

This Operator selects a subset of Attributes of an ExampleSet and removes the other Attributes. In the linear regression process, make the following changes

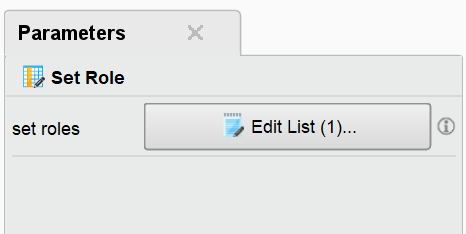
1. **Set Role**

This Operator is used to change the role of one or more Attributes.

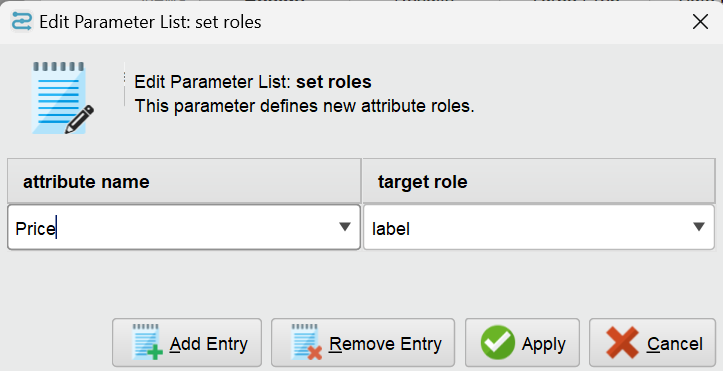
The role of an Attribute describes how other Operators handle this Attribute. The default role is regular, other roles are classified as special. The different types of roles are explained below in the parameter section.

An ExampleSet can have many special Attributes and you can assign special Attributes multiple times. This comes in handy, for example, if you want to feed the Attributes into a learner that accepts multiple labels. However, please note that some operators expect the special roles to be unique and they might not know how to handle duplicate special roles.

Made the two below changes in parameter like below



And other one is

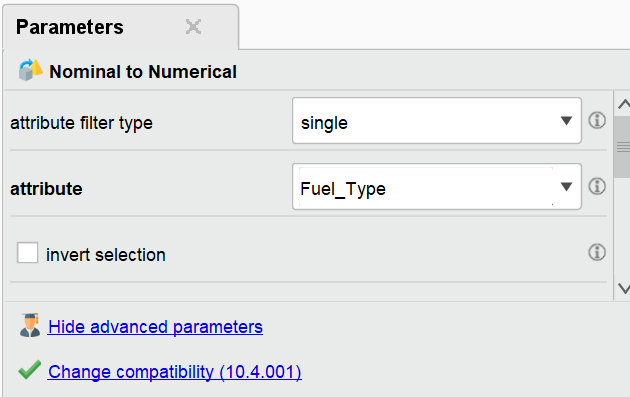


1. **Nominal to Numerical**

This operator changes the type of selected non-numeric attributes to a numeric type. It also maps all values of these attributes to numeric values

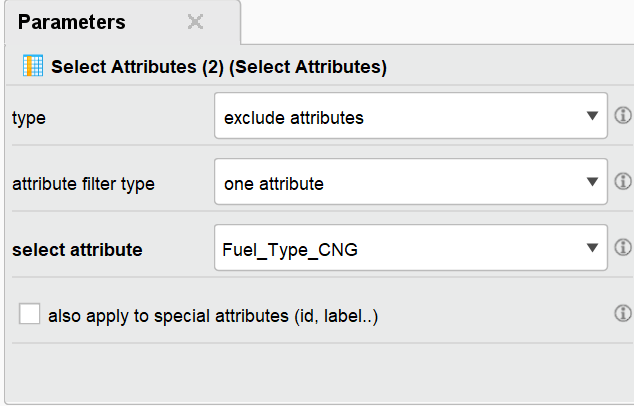
The Nominal to Numerical operator is used for changing the type of non-numeric attributes to a numeric type. This operator not only changes the type of selected attributes but it also maps all values of these attributes to numeric values. *Binary* attribute values are mapped to 0 and 1. Numeric attributes of input the ExampleSet remain unchanged. This operator provides three modes for conversion from nominal to numeric. This mode is selected by the *coding type* parameter. Explanation of these coding types is given in the parameters and they are also explained in the example process.

Make the following changes for the linear regression model



1. **Select Attribute**

Second time, we used the select attribute the parameter for that is



The Operator provides different filter types to make Attribute selection easy. Possibilities are for example: Direct selection of Attributes. Selection by a regular expression or selecting only Attributes without missing values. See parameter *attribute filter type* for a detailed description of the different filter types. The *type* parameter can be used to decide whether to include or exclude the selected Attributes. Special Attributes (Attributes with Roles, like id, label, weight) are by default ignored in the selection. They will always remain in the resulting output ExampleSet. The parameter *also apply to special attributes* changes this. Only the selected Attributes are delivered to the output port. The rest are removed from the ExampleSet.

1. **Split Data**

This operator produces the desired number of subsets of the given ExampleSet. The ExampleSet is partitioned into subsets according to the specified relative sizes.The Split Data operator takes an ExampleSet as its input and delivers the subsets of that ExampleSet through its output ports. The number of subsets (or partitions) and the relative size of each partition are specified through the *partitions* parameter. The *sampling type* parameter decides how the examples should be shuffled in the resultant partitions. For more information about this operator please study the parameters section of this description. This operator is different from other sampling and filtering operators in the sense that it is capable of delivering multiple partitions of the given ExampleSet.

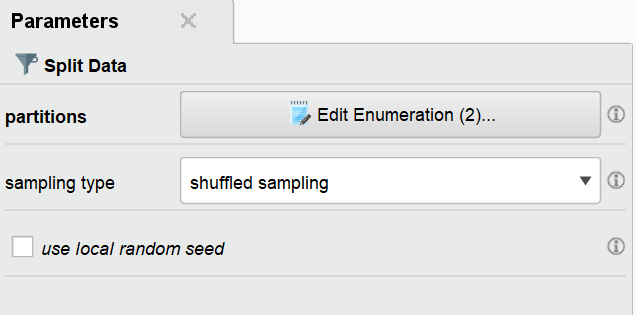
Input:

This input port expects an ExampleSet. It is the output of the Retrieve operator in the attached Example Process.

Output:

This operator can have multiple number of partition ports. The number of useful partition ports depends on the number of partitions (or subsets) this operator is configured to produce. The partitions parameter is used for specifying the desired number of partitions.

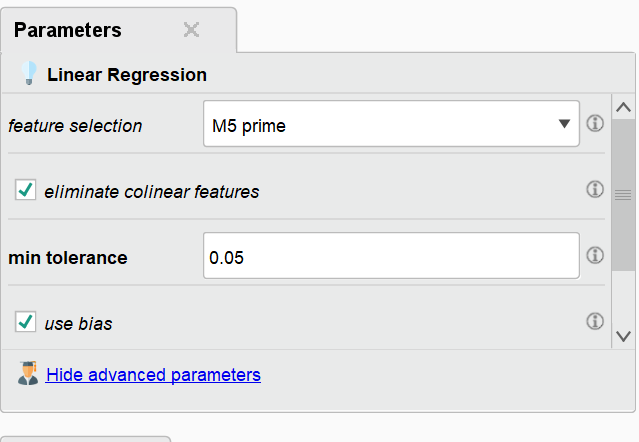
Make the below changes in the parameter for the split data operator



1. Linear regression

This operator calculates a linear regression model from the input ExampleSet.

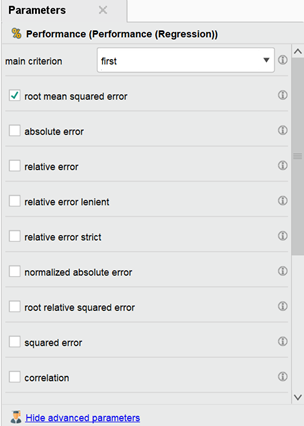
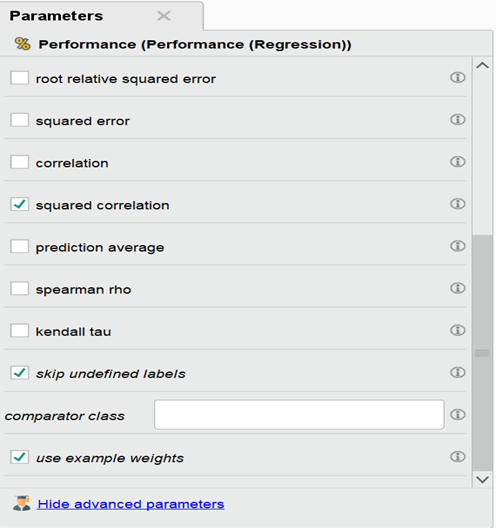
Didn’t make any changes in parameter. The default parameter which we used is



1. Apply Model

This Operator applies a model on an ExampleSet. A model is first trained on an ExampleSet by another Operator, which is often a learning algorithm. Afterwards, this model can be applied on another ExampleSet. Usually, the goal is to get a prediction on unseen data or to transform data by applying a preprocessing model. The ExampleSet upon which the model is applied, has to be compatible with the Attributes of the model. This means, that the ExampleSet has the same number, order, type and role of Attributes as the ExampleSet used to generate the model.

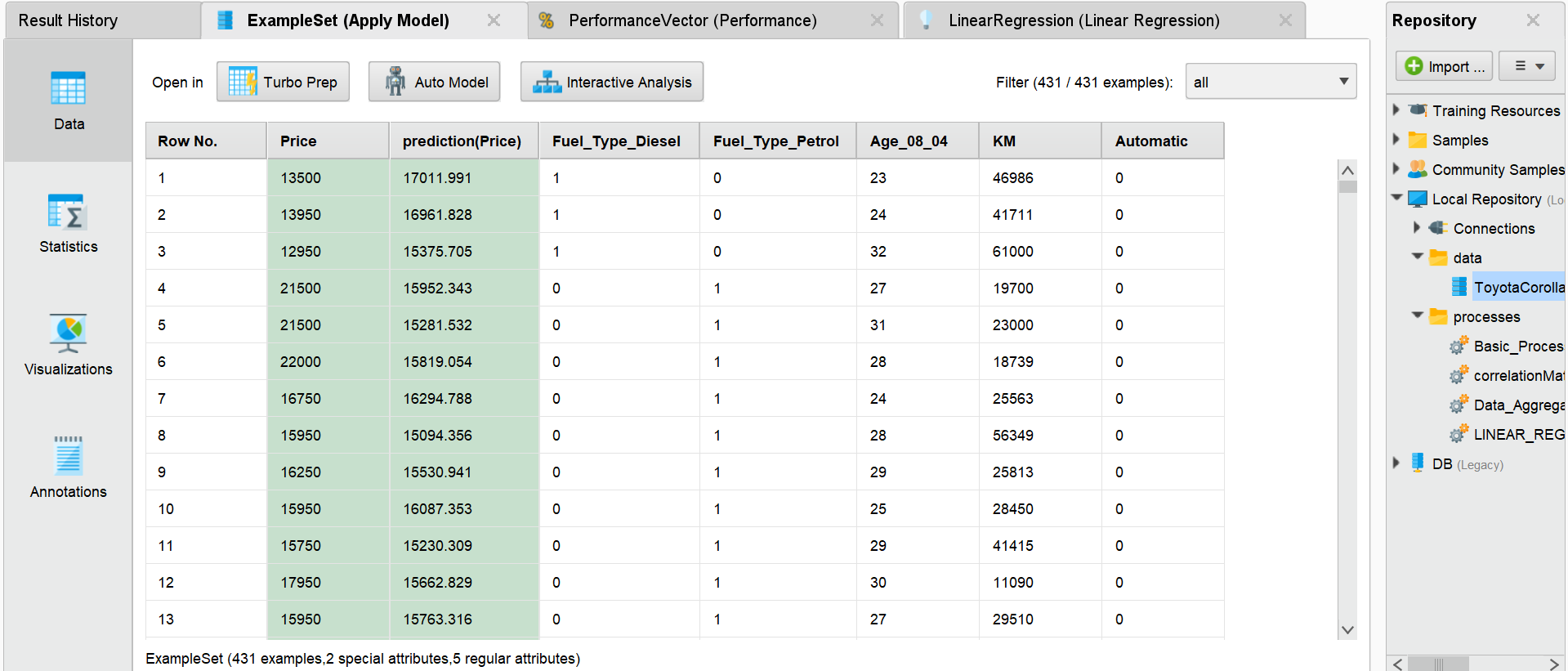
1. Performance regression

This operator is used for statistical performance evaluation of regression tasks and delivers a list of performance criteria values of the regression task. This operator should be used for performance evaluation of regression tasks only. This input port expects a labeled ExampleSet. The Apply Model operator is a good example of such operators that provide labeled data.Make sure that the ExampleSet has the label and prediction attribute. Make the changes in the parameter as given below  

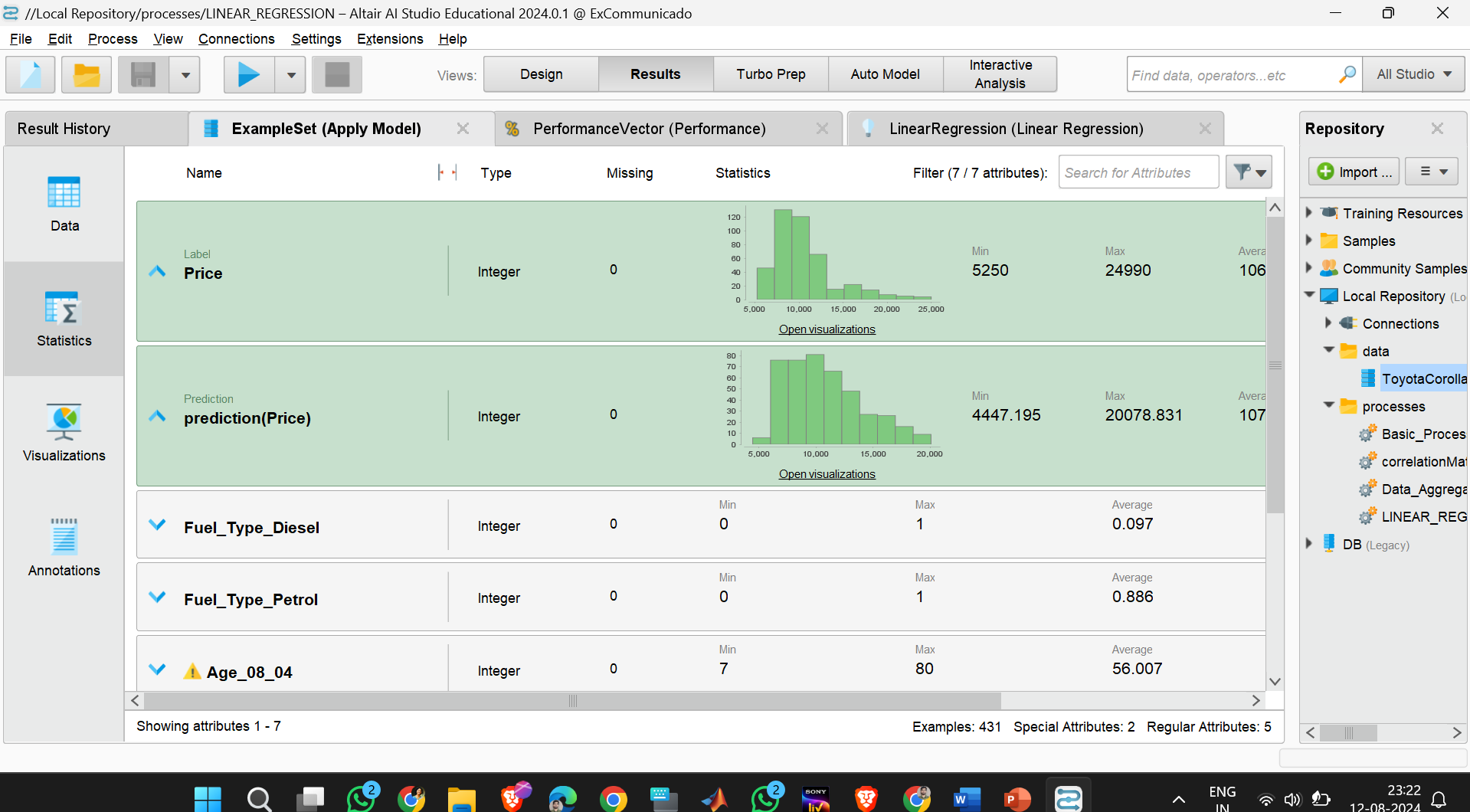
**7.9 Results**

We will get three results here

**7.9.1 ExampleSet (Apply Model)**

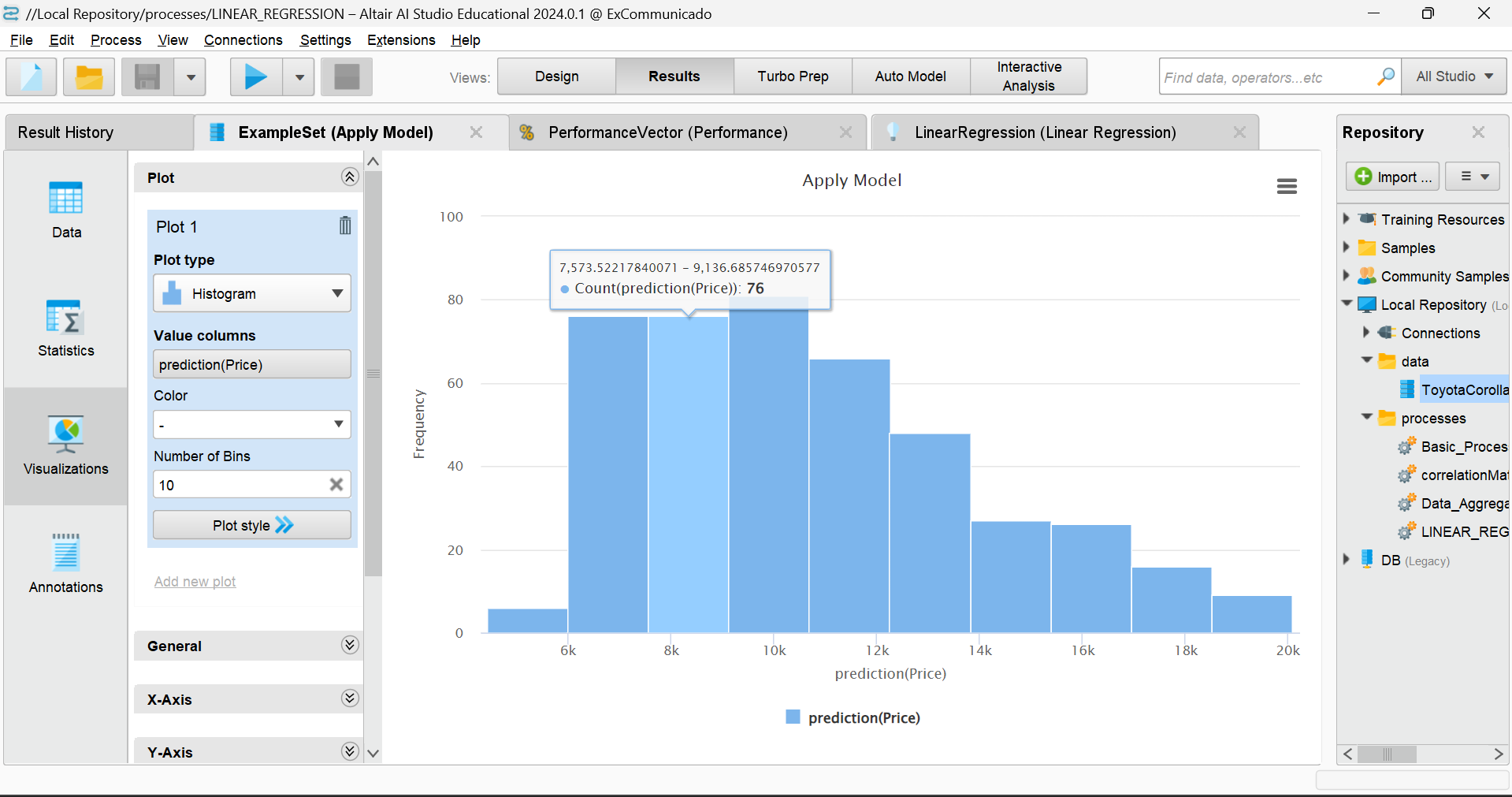
****

Statistics:

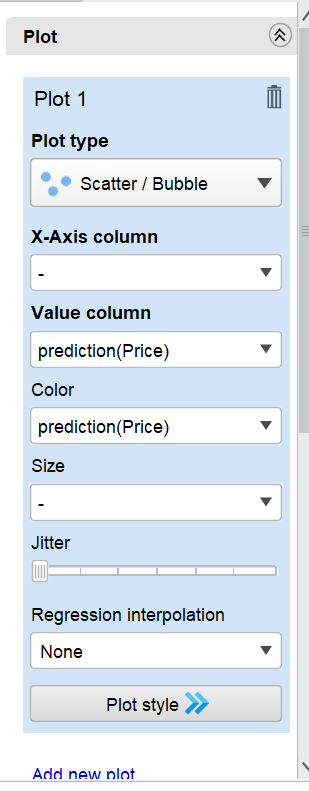


You can click on open visualisation to see the visualisation part

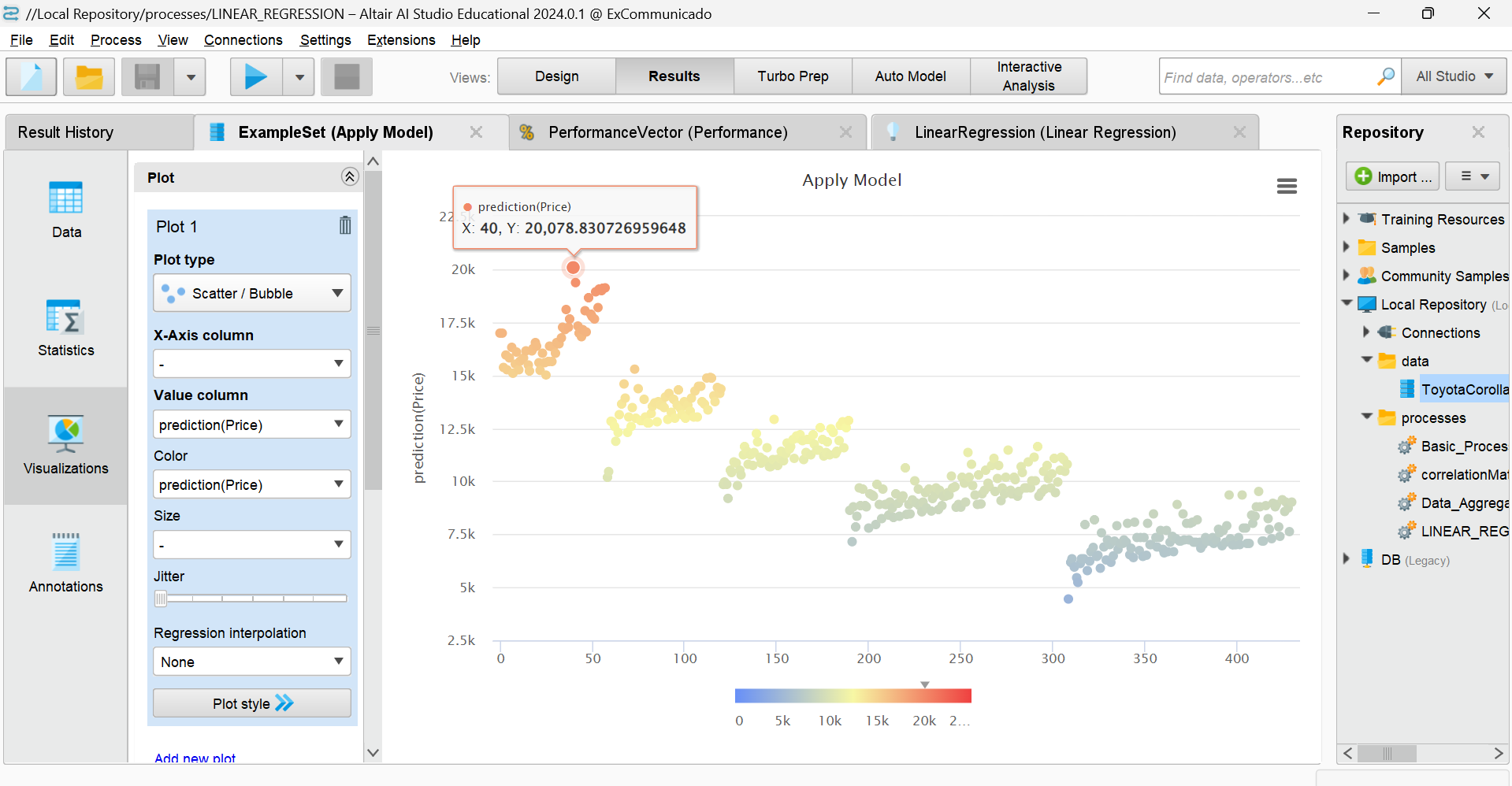
For example : for PREDICTION PRICE



Let’s make some changes like below:



As we have chosen plot type as scatter/bubble and color prediction (price), observe the below figure



**7.9.2 PerformanceVector(Performance)**

****

The Root Mean Squared Error (RMSE) is a commonly used metric to evaluate the accuracy of a regression model. It measures the average magnitude of the error between the predicted values and the actual values, where the errors are squared before being averaged. This metric provides a measure of how well the model's predictions match the actual data.



Where:

* ni is the number of data points.
* yi ​ is the actual value for the iiith data point.
* ​ yi cap is the predicted value for the ith data point.

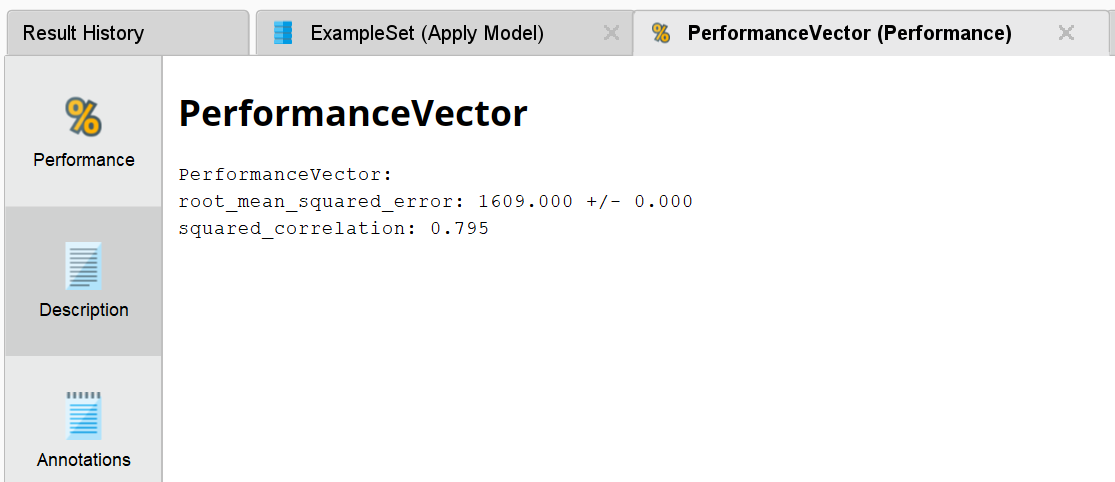
**Interpretation:**

* **Low RMSE:** Indicates that the model's predictions are close to the actual values, meaning the model is performing well.
* **High RMSE:** Suggests that the model's predictions deviate significantly from the actual values, indicating that the model may not be capturing the underlying data patterns effectively.



Squared correlation, often referred to as the coefficient of determination and denoted as R2, measures how well the variability of one variable is explained by another variable in a regression model. It provides an indication of the goodness-of-fit of the model.

As we move to the description section



**Root Mean Squared Error (RMSE):**

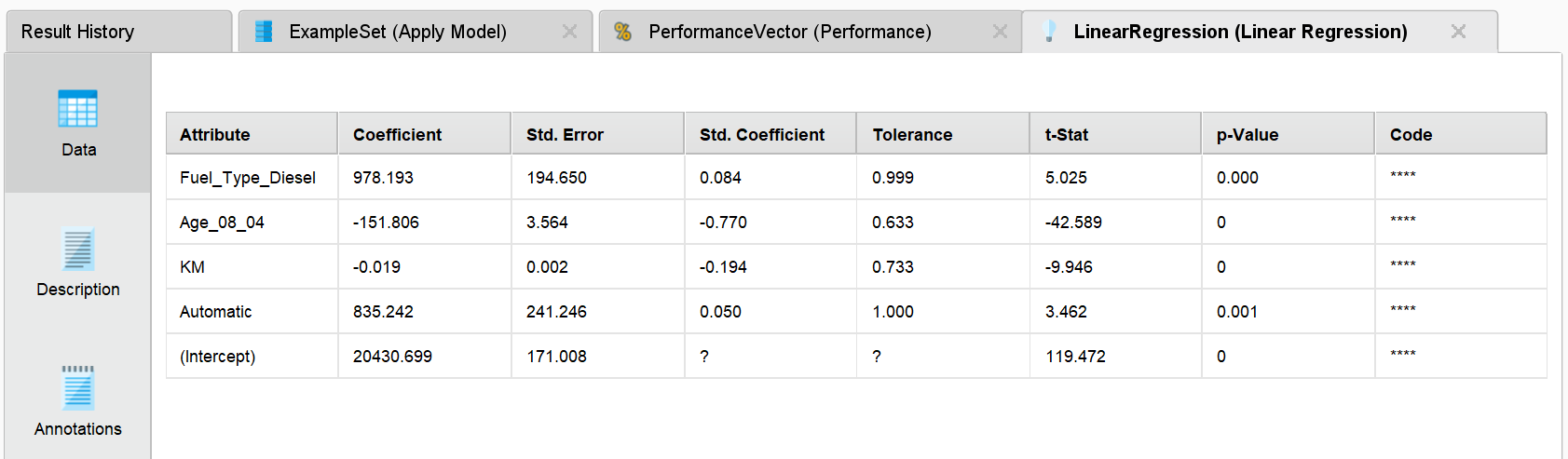
* **Value:** 1609.000
* **Interpretation:** RMSE measures the average magnitude of errors between the predicted and actual values. A value of 1609.000 suggests that, on average, the predictions deviate from the actual values by approximately 1609 units. The "+/- 0.000" indicates there is no variability in this error metric in the given context.

**Squared Correlation (R²):**

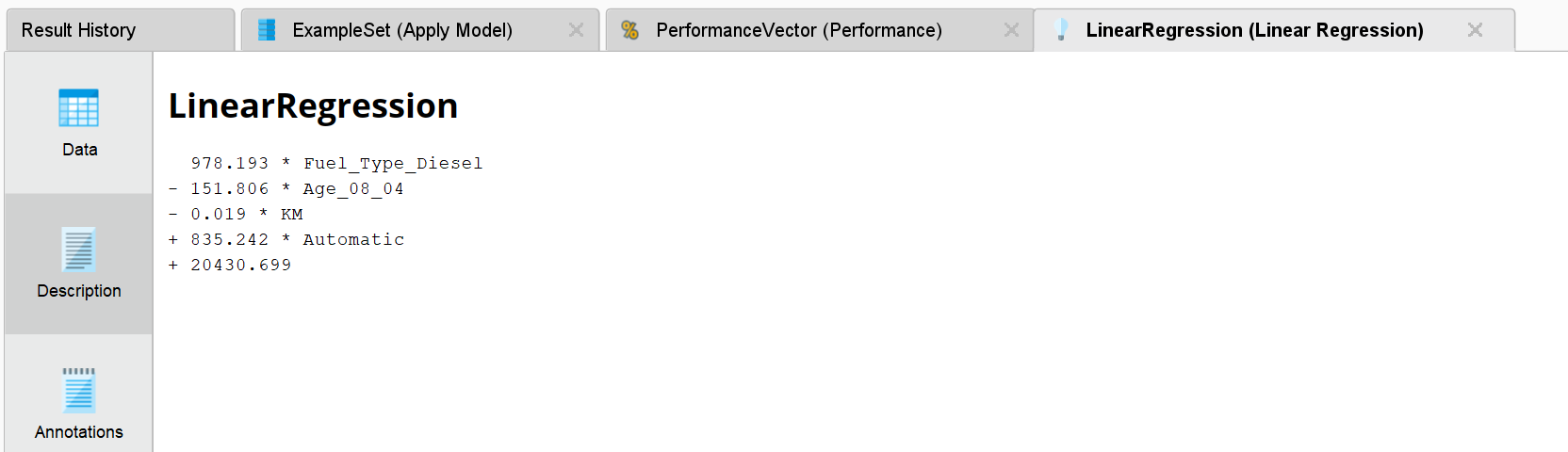
* **Value:** 0.79
* **Interpretation:** This squared correlation, or R2, of 0.79 means that 79% of the variance in the dependent variable is explained by the independent variable(s) in the model. This is a relatively high R2, suggesting that the model does a good job of explaining the variance in the data.

**7.9.3 LinearRegression (Linear Regression)**

Data section:



Description section



**978.193 \* Fuel\_Type\_Diesel:**

* This term indicates the impact of the Fuel\_Type\_Diesel variable on the dependent variable. The coefficient 978.193 suggests that, for a vehicle with diesel fuel, there is an additional value of 978.193 (in the unit of the dependent variable) compared to a reference fuel type (petrol).

**-151.806 \* Age\_08\_04:**

* This term shows the effect of the Age\_08\_04 variable on the dependent variable. The negative coefficient -151.806 means that for each unit increase in Age\_08\_04 (which is the age of the vehicle in months, for example), the value of the dependent variable decreases by 151.806.

**-0.019 \* KM:**

* This term represents the influence of the KM ( the number of kilometers driven) on the dependent variable. The coefficient -0.019 suggests that for each additional kilometer driven, the dependent variable decreases by 0.019.

**+835.242 \* Automatic:**

* This term reflects the impact of having an Automatic transmission on the dependent variable. The coefficient 835.242 indicates that vehicles with an automatic transmission have an additional value of 835.242 compared to vehicles with a manual transmission.

**+20430.699:**

* This is the intercept or constant term of the regression equation. It represents the baseline value of the dependent variable when all predictor variables are zero.

**CONCLUSION**

With the help of RapidMiner, data science tasks become more efficient and effective. The platform facilitates seamless data importing from various sources, offers powerful data visualization tools for uncovering insights, and simplifies process building with its intuitive drag-and-drop interface. It enables easy generation of correlation matrices to analyze variable relationships, efficient data aggregation for summarizing large datasets, and straightforward linear regression modeling for predictive analysis. Overall, RapidMiner's comprehensive features empower data scientists to streamline their workflows, enhance data analysis, and derive actionable insights with ease.

**REFERENCES**

* <https://en.wikipedia.org/wiki/Data_science>
* <https://en.wikipedia.org/wiki/Machine_learning>
* <https://www.kaggle.com/datasets/klkwak/toyotacorollacsv>
* <https://academy.rapidminer.com/learn/video/rapidminer-studio-gui-intro>
* <https://docs.rapidminer.com/latest/studio/installation/index.html>
* <https://docs.rapidminer.com/latest/studio/getting-started/index.html>
* <https://docs.rapidminer.com/latest/studio/guided/index.html>
* <https://www.geeksforgeeks.org/ml-linear-regression/>