OneModel Academy

One AI

Data Exploration

Machine Learning

Austin Hambrick

Machine Learning Engineer

07 Jun 2024

Topics Covered

- Introduction to data exploration & its importance
- Data statistics, including descriptive statistics, distribution analysis, measures of dispersion, & correlation analysis
- Visual exploration, including charts in the EDA report such as histograms, bar charts, & violin plots
- Data quality checks to ensure the dataset is suitable for machine learning modeling

Learning Outcomes

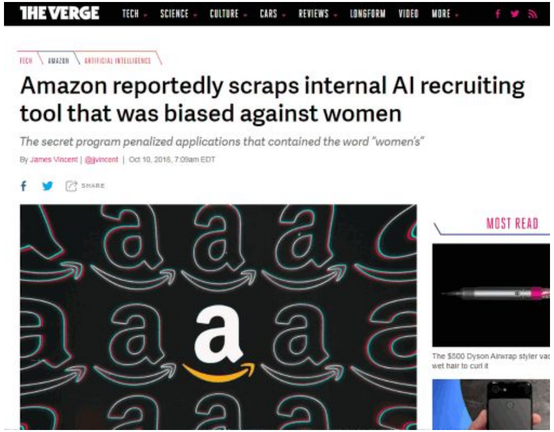
- You will understand what data exploration is & its critical role in the machine learning process, serving as a foundational step before model building
- You will respect the role of descriptive, distribution, & correlation statistics in uncovering data patterns & relationships
- You will become familiar with the diverse array of visualization tools, including histograms, bar charts, & violin plots, available in the Exploratory Data Analysis (EDA) report
- You will understand the importance of data quality checks around missing values, anomalies, & unreliable features to ensure clear & consistent data for modeling

Introduction to Data Exploration

Introduction to Data Exploration

Data exploration is an important early step in the ML process

- Involves analyzing the model dataset to understand its structure, relationships, & patterns
- Identifying data quality issues, informing preprocessing strategies, & guiding feature selection
- Should be done prior to model building to prevent errors & reduce biases **and** after the model is created & run with the **EDA report**
- Helps explain model behavior & the features that drive predictions
    - If we can't understand what the model is doing or the features it's using, it should not be used
- Involves data statistics, visual exploration, & data quality checks



Data Statistics

Data Statistics

- **Data statistics** are numerical summaries that reveal important characteristics of your dataset
    - Reveal averages, variations, & relationships between features
- We will walk through the following subsets of data statistics - some categories overlap; not everything fits neatly into one type
    - Descriptive statistics
    - Distribution analysis
    - Measures of dispersion
    - Correlation analysis

Descriptive Statistics

Descriptive Statistics

- Offer a quick summary of the dataset, describing key characteristics by input feature:
    - **Variable type** indicates the type of variable, either numeric or non numeric
    - **Non-null count** helps assess data quality & decide whether to drop or fill missing values.
    - **Unique count** shows how many distinct values exist in a feature, which is useful helping decide whether to treat a feature as binary or update the category size threshold
    - **Most frequent** identifies the most common value in a feature, which highlights dominant classes in categorical features & possible data imbalances
    - **Mean** is the average value of a numerical feature, providing central tendency
    - **Standard deviation** measures the spread of data around the mean. A high standard deviation means the data points are more spread out, while a low standard deviation indicates they are closely clustered around the mean. This is useful in identifying outliers.

One AI Supports Descriptive Statistics

- Generate data statistics from the One AI Query Builder validation step

Generate Data Statistics

Success - There are no issues with your data model that would cause the query to fail.

Column	Type	Count	Non-null Count	Unique Count	Most Frequent	Most Frequent Count	Min	Max	Mean	Standard Deviation
Time Periods	Non-numeric	1248	1248	1	2022-12-31	1248	-	-	-	-
Headcount (EOP)	Numeric	1248	1248	1	-	-	1.00	1.00	1.00	0.00

- View quantile statistics & descriptive statistics for each input variable from the EDA report variable analysis section

Histogram ▾ Statistics ▾ Common values ▾ Extreme values ▾ Violin Plot ▾ Overlay Histogram

Quantile statistics

Minimum	0
5-th percentile	7000036
Q1	209000014
Median	404000089
Q3	803000013
95-th percentile	1002000051
Maximum	1010000016
Range	1010000016
Interquartile range	593999999

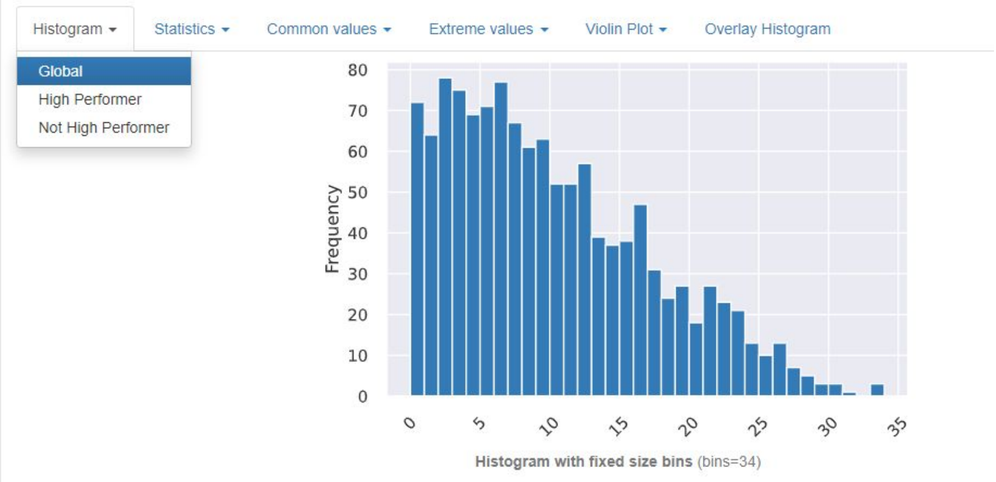
Descriptive statistics

Standard deviation	338643609.9
Coef of variation	0.6958616201
Kurtosis	-1.425210263
Mean	486653668.1
MAD	303947453.5
Skewness	0.07741788748
Sum	4.973600488e+11
Variance	1.146794945e+17
Memory size	48.3 KiB

Distribution Analysis

Distribution Analysis

- Helps you understand the spread & shape of features, revealing skewness, unusual patterns in the data, or outliers
- One AI supports this analysis with EDA report allowing you to view distribution globally or by class label



Measures of Dispersion

Measures of Dispersion

- Use statistics to help understand the data's variability
  - Variance & standard deviation measure how far data points are from the mean
  - Range is the difference between the minimum & maximum values
  - IQR is the range between the 25th & 75th percentile
- This data can be found in the One AI EDA report in the variable analysis section under statistics

Histogram ▾ | Statistics ▾ | Common values ▾ | Extreme values ▾ | Violin Plot ▾ | Overlay Histogram

Quantile statistics

Minimum	0
5-th percentile	7000036
Q1	209000014
Median	404000089
Q3	803000013
95-th percentile	1002000051
Maximum	1010000016
Range	1010000016
Interquartile range	593999999

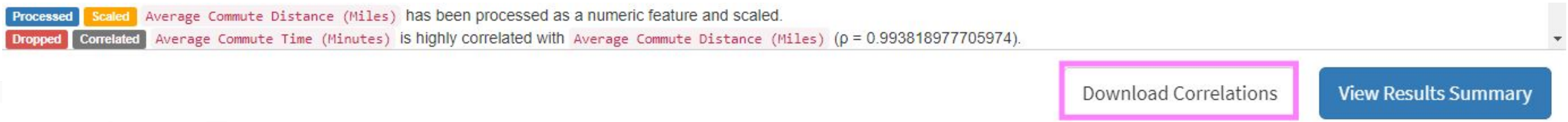
Descriptive statistics

Standard deviation	338643609.9
Coef of variation	0.6958616201
Kurtosis	-1.425210263
Mean	486653668.1
MAD	303947453.5
Skewness	0.07741788748
Sum	4.973600488e+11
Variance	1.146794945e+17
Memory size	48.3 KiB

Correlation Analysis

Correlation Analysis

- Identifies relationships between different features or variables
- One AI's EDA report provides downloadable correlation information, showing how each input variable is correlated with others & the target variable



	one.employee.annual_salary_range
Performance Review Occurrences	0.044063108
one.employee.annual_salary_range	1
one.employee.city	0.064557989
one.employee.corrective_action	0.054666633
one.employee.diverse	0.038390636

Visual Exploration

Data Quality Checks

Data Quality Checks

- Should occur throughout data exploration to ensure the model is trained on reliable, good data
- As you identify issues, make plans to address them before model building
    - Handle missing values
        - One AI automatically drops overly null columns & offers null-filling strategies
    - Identify anomalies
        - Remove anomalies, transform the data, or use an algorithm designed to handle them
    - Drop unreliable features
        - One AI automatically drops highly correlated & constant variables as well as those with no predictive power

OneModel Academy

Thanks for watching!

OneAI