



# DATA EXPLORATION

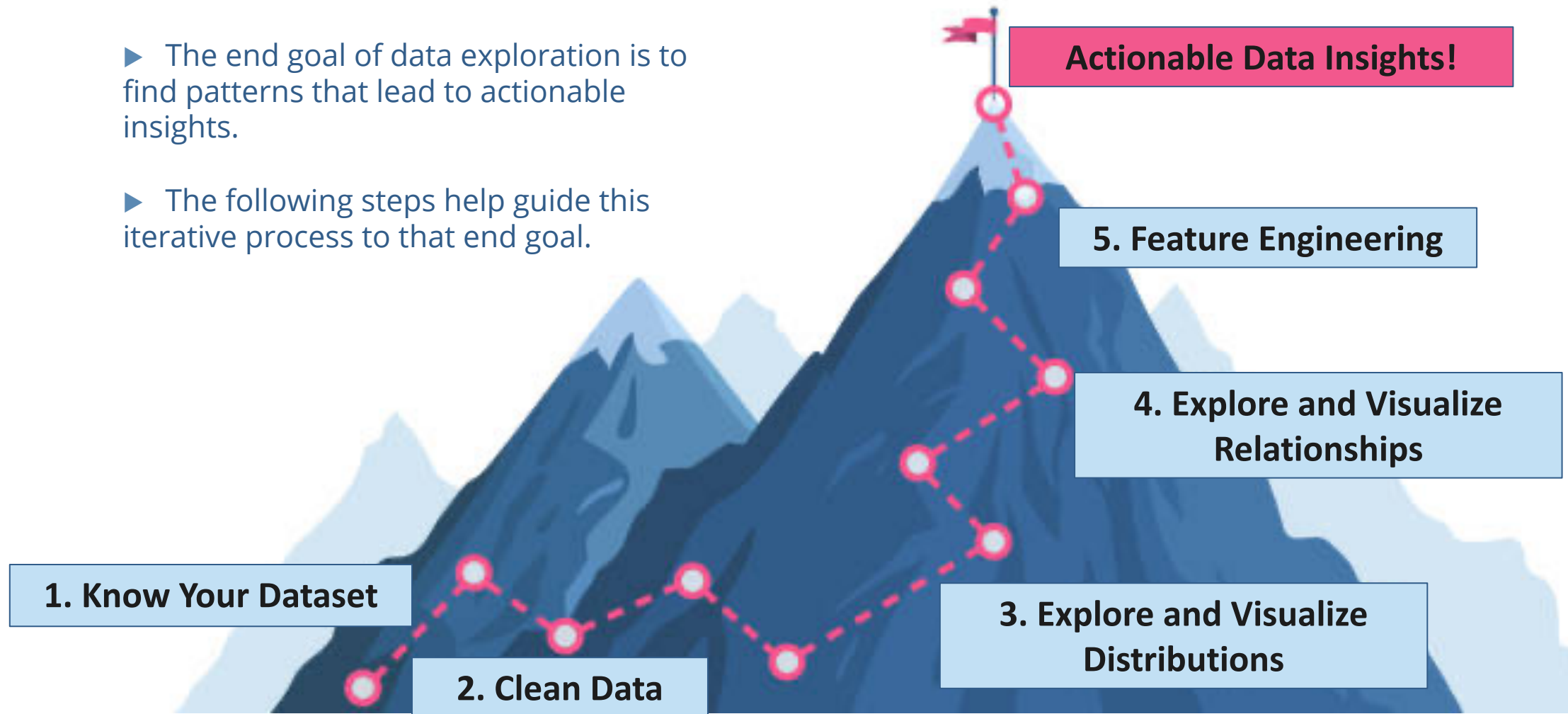
# Exploratory Data Analysis

- ▶ EDA is the initial analysis of a dataset to uncover patterns, relationships, and insights.
- ▶ EDA is an iterative process from which hypotheses are created and tested and questions are answered and derived through exploration of these patterns.
  - ▶ Effective EDA depends on a strong understanding of the business questions we wish to answer with the data and tools at our disposal.
- ▶ Before visualizing data to find patterns, the dataset must be understood and cleaned.
- ▶ Once the dataset is cleaned, various techniques and visualizations discussed in this section are used to uncover actionable insights and, if applicable, prepare the data for predictive modeling.

# Data Exploration Roadmap



- ▶ The end goal of data exploration is to find patterns that lead to actionable insights.
- ▶ The following steps help guide this iterative process to that end goal.



**1. Know Your Dataset**

**2. Clean Data**

**3. Explore and Visualize  
Distributions**

**4. Explore and Visualize  
Relationships**

**5. Feature Engineering**

**Actionable Data Insights!**

# KNOW YOUR DATASET

What equipment do we have for our exploratory journey?

# Know Your Dataset

- ▶ First step: Fully understand dataset.
- ▶ Important big picture components of a dataset include:
  - ▶ Dataset Structure
  - ▶ Unit of Analysis / Primary Key
  - ▶ Data Characteristics (Data type, dimensions, etc.)

# Dataset Structures

2 distinguishing features determine a dataset's structure:

## Time Component

A **cross-sectional dataset** is a dataset that holds individuals described by features **at a fixed point in time**.

A **time-series dataset** is a dataset that holds observations described by features **observed repeatedly over time**.

## Nesting/Clustering Component

A **simple dataset** is a dataset that **cannot be aggregated** into a larger group

A **clustered dataset** is a dataset that **can be aggregated** into a larger group

	Unit of analysis is NOT nested within larger units	Unit of analysis IS nested within larger units
Snapshot in Time	Simple Cross-Sectional	Clustered Cross-Sectional
Over Time	Simple Time Series	Clustered Time Series

# Cross-Sectional vs. Time-Series

**Simple Cross-Sectional Dataset:**  
Employee Directory for Marketing Team

Employee ID	Status	Employee Satisfaction Score	Manager Evaluation Score	Salary	Tenure	Gender
E46	Employed					
E29	Employed					
E121	Terminated					
E5	Employed					
E32	Terminated					
E24	Terminated					
E91	Terminated					
E74	Employed					
E2	Terminated					
E14	Employed					
E32	Employed					
E7	Terminated					
E44	Employed					
E85	Terminated					
E91	Employed					
E104	Employed					

**Simple Time-Series Dataset:**  
Quarterly Employee Directory for Marketing Team

Employee ID	Quarter	Employee Satisfaction Score	Manager Evaluation Score	Salary	Tenure	Gender
E46	Q1-2022					
E46	Q2-2022					
E46	Q3-2022					
E46	Q4-2022					
E32	Q1-2022					
E32	Q2-2022					
E32	Q3-2022					
E32	Q4-2022					
E6	Q1-2022					
E6	Q2-2022					
E6	Q3-2022					
E6	Q4-2022					
E44	Q1-2022					
E44	Q2-2022					
E44	Q3-2022					
E44	Q4-2022					

**Time Series Feature**

# Nested Dataset Examples

**Nested Cross-Sectional Dataset:**  
**Employee Satisfaction for All Departments**

Employee ID	Department	Employee Satisfaction Score	Manager Evaluation Score	Salary	Tenure	Gender
E46	Marketing					
E29	HR					
E121	Security					
E5	Management					
E32	Support					
E24	Marketing					
E91	Support					
E74	HR					
E2	Management					
E14	HR					
E32	Marketing					
E7	Marketing					
E44	Security					
E85	Support					
E91	Management					
E104	Support					

**Nesting Feature**

**Nested Time-Series Dataset:**  
**Quarterly Employee Satisfaction for All Departments**

Employee ID	Department	Quarter	Employee Satisfaction Score	Manager Evaluation Score	Salary	Tenure	Gender
E46	Marketing	Q1-2022					
E46	Marketing	Q2-2022					
E46	Marketing	Q3-2022					
E46	Marketing	Q4-2022					
E32	Support	Q1-2022					
E32	Support	Q2-2022					
E32	Support	Q3-2022					
E32	Support	Q4-2022					
E6	HR	Q1-2022					
E6	HR	Q2-2022					
E6	HR	Q3-2022					
E6	HR	Q4-2022					
E44	Management	Q1-2022					
E44	Management	Q2-2022					
E44	Management	Q3-2022					
E44	Management	Q4-2022					

**Time Series Feature**

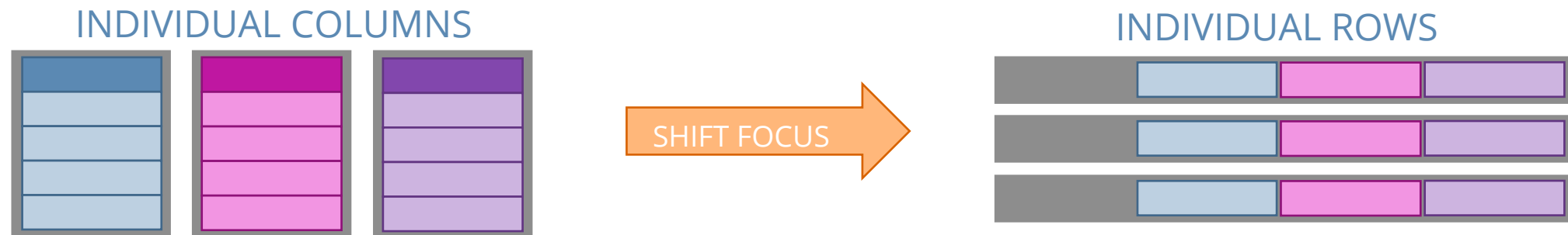


# Unit of Analysis

## What does an individual row represent in the data?

A clear Unit of Analysis contains the 3 S's:

- ▶ Singular
  - ▶ There should be no duplicate observations at the unit of analysis.
- ▶ Scheduled
  - ▶ If the dataset has a time-series component, the unit of analysis must describe it.
- ▶ Specific
  - ▶ If the dataset has a nesting component, the unit of analysis must describe it.



# Unit of Analysis

- ▶ Determined by:
  - ▶ The question to be addressed
    - ▶ Example: Are we predicting daily revenue or monthly revenue?
  - ▶ The availability and accuracy of data
- ▶ Informs Data Architecture:
  - ▶ More granular units are generally preferred
    - ▶ Example: daily rather than monthly
  - ▶ More granular units can be aggregated if necessary
    - ▶ Warning: Information is LOST in the aggregation process
  - ▶ Less granular units are usually difficult to disaggregate without compromising accuracy

# Granularity, Segmentation, and Aggregation

## Granularity

- ▶ The scale or level of detail present in a set of data
- ▶ The level at which the prediction for the outcome is to be made
- ▶ In time-series data, the granularity is the measurement of time intervals
- ▶ Examples
  - ▶ Daily vs. weekly vs. monthly
  - ▶ Region vs. division vs. group

## Segmentation

- ▶ The division of the data into logical (based on business logic) groupings.
- ▶ These groups have distinct characteristics and should therefore be treated separately
- ▶ Examples
  - ▶ Loan customers vs. account holders
  - ▶ Subscribed customers vs. not subscribed
  - ▶ U.S. vs. Canada homeowners

## Aggregation

- ▶ The process of summarizing raw data
- ▶ Generally utilizes measures of central tendency (mean, median, average) or sums/counts
- ▶ Very sensitive to outliers; lose variation of the measure
- ▶ Examples
  - ▶ Average monthly revenue
  - ▶ Total terminated employees last month
  - ▶ Average months of employment prior to termination

# Unit of Analysis Examples

	OPERATIONS	SALES	HR	MARKETING
BUSINESS QUESTION	How many page views should we expect tomorrow?	How long will it likely take for sales opportunities to close?	Which employees are likely to leave the company in the next month?	Based on purchase history, which customers should we target for each of our three different departments' advertising campaigns?
UNIT OF ANALYSIS	Daily sum of page views	Sales opportunities	Monthly observations of employees	Purchases by customer by department.
DATA ARCHITECTURE	One row per day	One row per sales opportunity	One row per employee per month of employment	One row per purchase per customer by department

# Dataset Characteristics

The final step in understanding your dataset is taking a bird's eye view of the features it contains. Data Type of features should be considered as different analysis approaches exist for each.

## CONTINUOUS

- ▶ A measure that takes on numeric values
  - ▶ Also referred to as interval/ratio
- ▶ Intervals between each possible value are equal
- ▶ Examples:
  - ▶ Customer age
  - ▶ Employee base pay
  - ▶ Percent increase in customer satisfaction

## CATEGORICAL

- ▶ A measure that takes on values that are described by non-numeric categories
- ▶ Categories could be names, numbers, or labels that have **no inherent numeric order**
- ▶ Examples:
  - ▶ Customer gender
  - ▶ Employee race/ethnicity
  - ▶ Salesperson assigned to an opportunity

## ORDINAL

- ▶ Feature w/values described by numeric categories. Can be ordered lowest to highest.
  - ▶ Can be treated as categorical or as numeric
- ▶ Categories could be numbers or names/labels that **have a clear order**.
- ▶ Examples:
  - ▶ Survey responses on a "Disagree" to "Agree" scale
  - ▶ Survey responses on a "1 to 10" scale
  - ▶ Employee performance ratings

## ID

- ▶ Unique identifiers. Numbers or combination of numbers and letters.
- ▶ Sometimes confused for a numeric column and included in numeric analysis.

## DATE

- ▶ Can be used as category to aggregate by day, month, year, etc.
- ▶ Can be turned into a continuous variable to calculate time since a date.

# DATA CLEANING

Does our equipment need mending before the exploratory journey?

# Data Cleaning

- ▶ Much like a hiker is sure to wear durable clothing and sturdy shoes with no holes, we must ensure that our data is prepared for exploration.
- ▶ Important big picture components of data cleaning include the identification and handling of:
  - ▶ Missing Data
  - ▶ Outliers
  - ▶ Duplicates

# Unusual Observations

- ▶ Outliers
  - ▶ Observations that are atypically distant from other observations (statistically unusual)
- ▶ Leverage points (unusual/unreasonable values)
  - ▶ Observations that are unusual/unreasonable for the given use case, regardless of their statistical properties
  - ▶ Observations that are significantly different from other observations in the same variable

Employee ID	Hourly Wage
5	18
2	19
9	19
13	19
1	20
6	20
7	20
8	20
11	21
3	22
10	22
12	23
14	23
4	24
15	110



# Why Do We Care About Outliers?

- ▶ Because outliers distort statistics about our data (e.g., means/averages, standard deviations, correlations, etc.)

Outlier Included?	Average Hourly Wage	Standard Deviation of Hourly Wage
Without Outlier	\$20.71	\$1.75
With Outlier	\$26.67	\$22.34

- ▶ As a result:
  - ▶ Hypothesis/significance tests (which are based on differences in means) between groups will be biased, which will subsequently affect model results
  - ▶ Larger standard deviation values reduces statistical power (which means hypothesis/significance tests will be less likely to detect an effect that exists), which will further affect model results

# How to Identify Outliers

- ▶ Business-driven caps
  - ▶ Limit data to appropriate ranges required by the business use-case
  - ▶ Example: If you are modeling car prices, you may consider removing cars from your analysis that are above a certain price. These high-end cars might not be true statistical outliers, but if your business-case is focused on modeling everyday automobiles, selecting an upper limit for price will be beneficial
- ▶ Common sense
  - ▶ Other thresholds may simply be the result of common-sense or "gut feelings"
  - ▶ This approach is often included in Data Quality Assurance Analysis
  - ▶ Example: If looking at the weights of mice, any negative value should be treated as an extreme value, regardless of whether these values were identified as outliers with statistical methods

# How to Handle Outliers

- ▶ **Keep them as is:** This is recommended if true anomalies exist and are important to model because they are important for informing both the results and the assumptions one would make about the data and interpretations of it
  - ▶ Example: A new tech company that wants to understand employee satisfaction, but only has a few employees who are older than 40
- ▶ **Keep them & flag them (or otherwise account for them):** Especially with time series analyses, outliers can be the result of expected peaks and valleys in the data that can be accounted for in a model
  - ▶ Example: A retail chain reports extremely high numbers of transactions on Black Friday
- ▶ **Correct them:** Outliers can be the result of a coding or record-keeping error that can be corrected
  - ▶ Example: A coding error resulted in the conversion of contracts over \$999,999 to negative numbers
- ▶ **Impute them:** Highly improbable outliers can be identified and imputed (similar to missing data)
  - ▶ Example: A survey respondent lists his birth year as 1783, which is most likely an error
- ▶ **Omit the offending cases:** If the outlier indicates that the entire case is invalid or problematic, it can be omitted
  - ▶ Example: A media company that counts homepage views experiences inflated counts as a result of a bot attack

# Missing Data

Employee ID	Status	Department	Months Employed	Most Recent Salary Bracket	Most Recent Satisfaction Score	Most Recent Evaluation Score
E1	Employed	Marketing	13	5	5	5
E2	Employed	Lending	1	4		5
E3	Employed	HR	85		4.8	3
E4	Employed	Management	32	4	3	4
E5	Employed		54	5		
T1	Terminated	Security	45			3
T2	Terminated	HR	2	5	1	2
T3	Terminated	IT	67	3	1.5	1
T4	Terminated	Marketing	34	4	2	1
T5	Terminated	HR	5	3		3

# What Is Missing Data?

- ▶ Cells that are not populated with meaningful data
- ▶ These may or may not be empty cells
- ▶ Why do they matter?
  - ▶ Missing data are problematic in data science
    - ▶ High percentages of missing data often diminish confidence in the data and in the results generated from them
    - ▶ Statistical software will often drop rows with missing data and exclude them completely from the analysis
    - ▶ Some types of missing data are easier to manage than others
    - ▶ Missing data often occurs in patterns across multiple variables, which can offer insights into the quality and integrity of certain cases or rows of data
    - ▶ Degree to which data are missing is important to assess
    - ▶ Missing values could be miscoded or not relevant to the analysis
      - ▶ Nulls act as place-holders for zeros or other values, or they represent nonresponse to “not applicable” questions or data fields, so null and missing values should be recoded appropriately

# How To Manage Missing Data

- ▶ No good way to deal with missing data
- ▶ Depending on the type of missing data, each strategy may distort the data and generate biased results
- ▶ Recoding
  - ▶ Missing data may indicate a “zero” value or some other value that can be managed through recoding
- ▶ Flagging or Categorizing
  - ▶ Missing data gets their own category for modeling
  - ▶ Missing cells are acknowledged but entire cases are not omitted because of a few missing cells
- ▶ Deleting
  - ▶ Delete rows (listwise deletion): If null values exists for any variable in the observation of a row, the entire row is removed; used if missing data is limited to a small number of observations
  - ▶ Deleting columns (dropping variables): If null values exist for any variable, the entire variable is removed; used if more than 60% of observations are missing and the variable is NOT significant
- ▶ Imputing
  - ▶ The process of replacing missing data with substituted values
    - ▶ Can use mean, median, mode, linear regression, multiple imputation, etc.

# EXPLORE AND VISUALIZE DISTRIBUTIONS

Let's start using the tools at our disposal.

# Feature Distributions

- ▶ After understanding and cleaning our data, we begin to examine individual features.
- ▶ Exploring distributions of individual features helps us understand the range of normal values and where an observation falls in that range.
  - ▶ E.g., does this checking account have a high or low balance relative to others?
- ▶ Several methods exist for exploring feature distributions. We will review:
  - ▶ Summary Statistics
    - ▶ Measures of Center, Standard Deviations, etc.
  - ▶ Box Plots
  - ▶ Histograms and Density Plots
  - ▶ Bar Charts

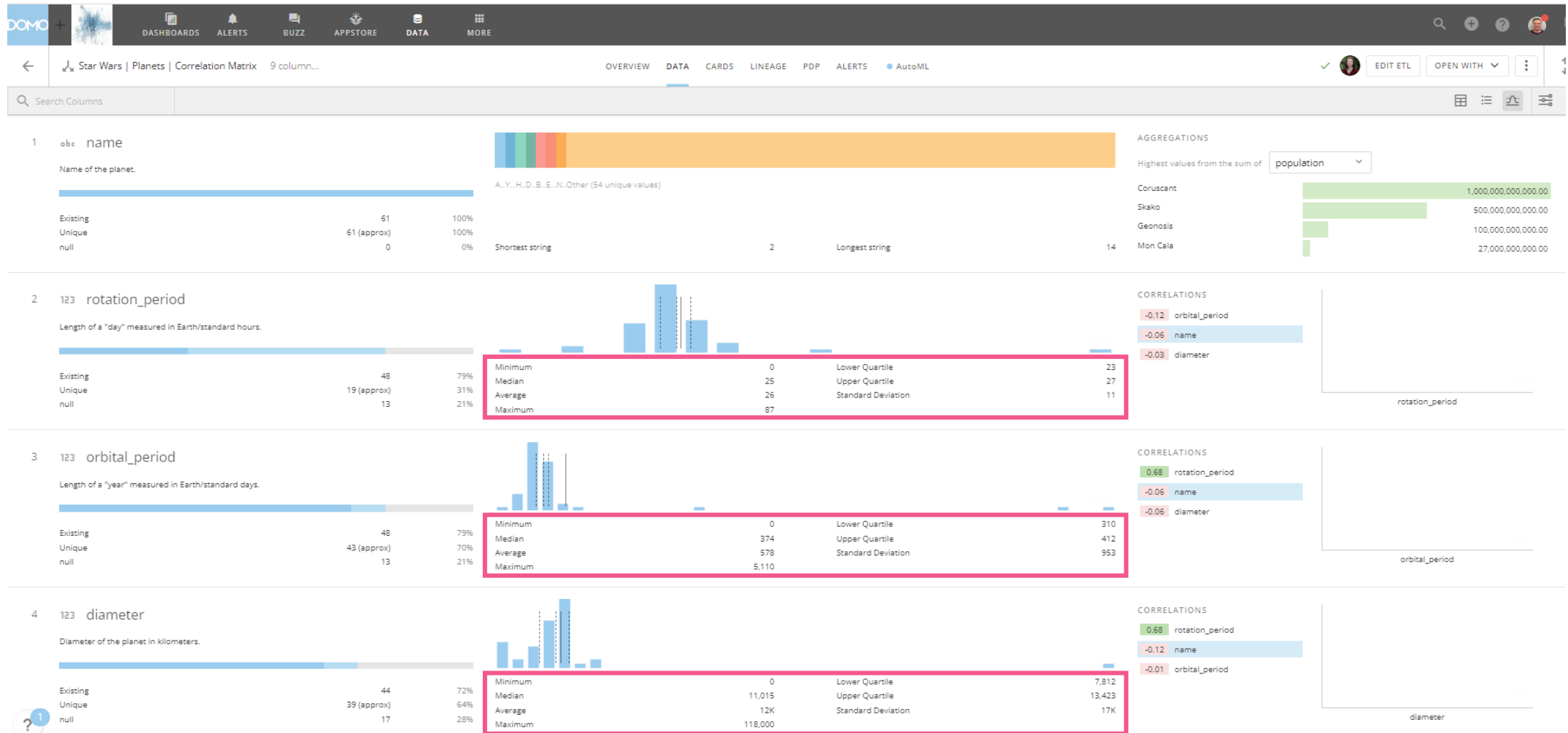


# Summary Statistics

► Summary statistics give us a numeric interpretation of a continuous feature's distribution and generally include:

- Minimum: Lowest value of the feature.
- Q1: First quartile value. Value under which 25% of the feature's observations fall when placed in ascending order.
- Median: Measure of Center. Value under which 50% of the feature's observations fall when placed in ascending order.
- Mean: Measure of Center. The feature's average value.
- Mode: Measure of center. Most common feature value.
- Q3: Third quartile value. Value under which 75% of the feature's observations fall when placed in ascending order.
- Maximum: Highest value of the feature.
- Standard Deviation: Measure of variation.
- Range: Difference between the Maximum and Minimum Values.

# Summary Statistics Example



# Summary Statistics Example

Summary Statistics



1,002 Total Rows



Column Name	min	25%	mean	50%	75%	max	count
Age	18.00	30.00	36.35	35.00	42.00	60.00	1,002.00
AttritionNumber	0.00	0.00	0.16	0.00	0.00	1.00	1,002.00
DailyRate	102.00	472.50	817.55	818.50	1,177.50	1,499.00	1,002.00
DistanceFromHome	1.00	2.00	9.47	7.00	15.00	29.00	1,002.00
Education	1.00	2.00	2.92	3.00	4.00	5.00	1,002.00
EmployeeNumber	2.00	483.25	1,041.40	1,064.00	1,587.50	2,068.00	1,002.00
EnvironmentSatisfaction	1.00	2.00	2.71	3.00	4.00	4.00	1,002.00
HourlyRate	30.00	48.00	65.82	65.00	83.00	100.00	1,002.00
JobInvolvement	1.00	2.00	2.71	3.00	3.00	4.00	1,002.00
JobLevel	1.00	1.00	1.89	2.00	2.00	4.00	1,002.00
JobSatisfaction	1.00	2.00	2.72	3.00	4.00	4.00	1,002.00
MonthlyIncome	1,009.00	2,853.75	5,754.29	4,762.00	7,101.50	17,426.00	1,002.00
MonthlyRate	2,094.00	8,054.50	14,421.79	14,575.50	20,872.00	26,999.00	1,002.00
NumCompaniesWorked	0.00	1.00	2.70	2.00	4.00	9.00	1,002.00
PerformanceRating	3.00	3.00	3.15	3.00	3.00	4.00	1,002.00
RelationshipSatisfaction	1.00	2.00	2.70	3.00	4.00	4.00	1,002.00
StandardHours	80.00	80.00	80.00	80.00	80.00	80.00	1,002.00
StockOptionLevel	0.00	0.00	0.80	1.00	1.00	3.00	1,002.00
TotalWorkingYears	0.00	6.00	10.52	9.00	14.00	40.00	1,002.00



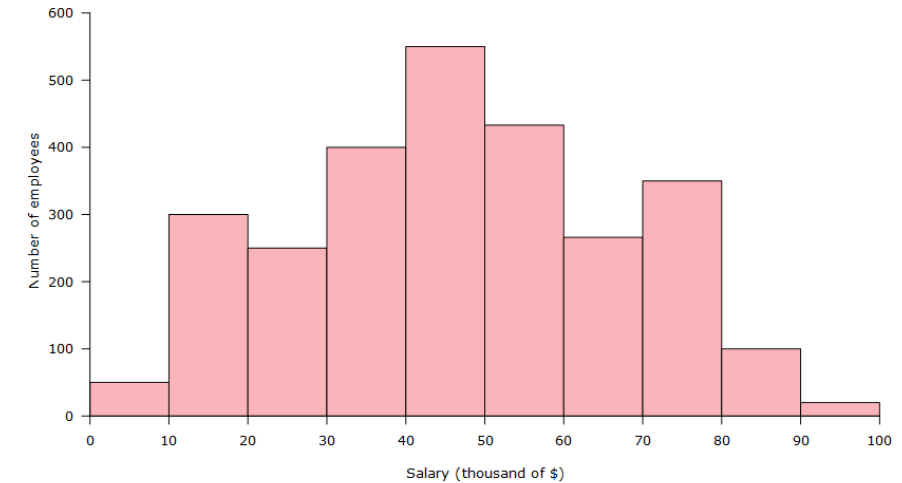
# Histograms

- ▶ A histogram is a visual representation of the shape and central tendency of a **numeric** variable's distribution.
- ▶ The x axis represents the range of the numeric variable split into several bins. The range within each of these bins should be equal.
- ▶ The y axis represents the number of observations that falls within each bin.
- ▶ Histograms answer several questions including:
  - ▶ Where is the center of the data?
  - ▶ What is the spread and shape of the data? Is it symmetric or skewed (influenced by outliers)?
- ▶ Appropriate bin selection is essential to creating a useful histogram.
  - ▶ Try several different bin ranges before settling on those you think best represent the data.
  - ▶ Overly large bins can make it difficult to see patterns that may be present on a smaller scale.

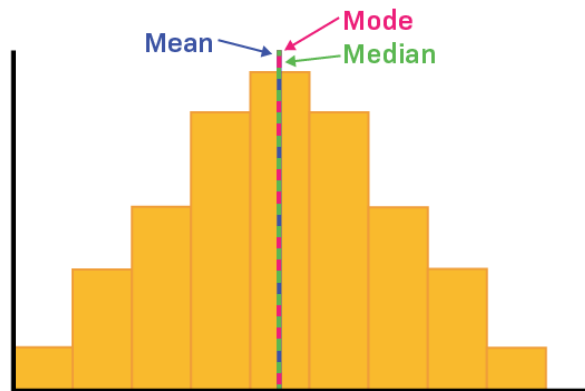
# Histogram Examples

- ▶ A skewed distribution may indicate that outliers are present.
- ▶ When a variable is skewed it is especially important to **use median as the measure of central tendency**. The mean is susceptible to influence from outliers.

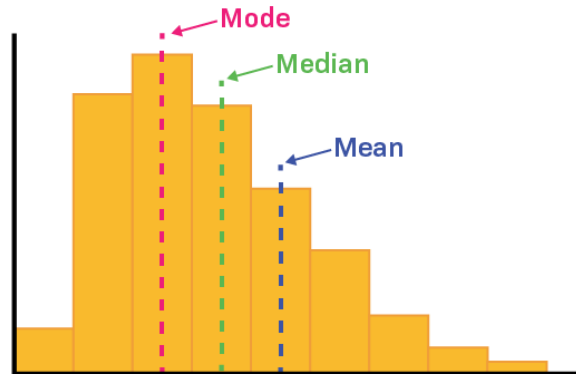
Distribution of salaries of the employees of ABC Corporation



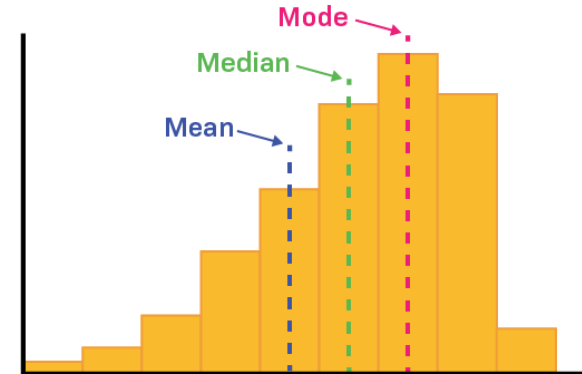
A. Symmetric



B. Right-skewed (or Positive-skewed)

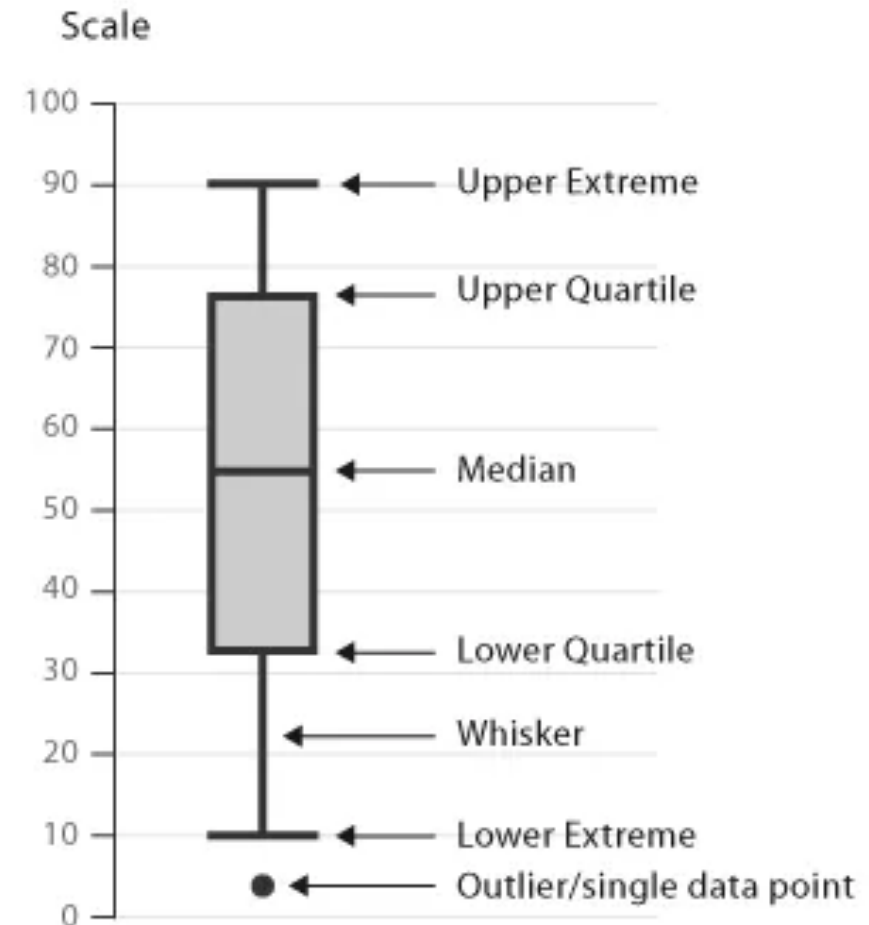


C. Left-skewed (or Negative-skewed)



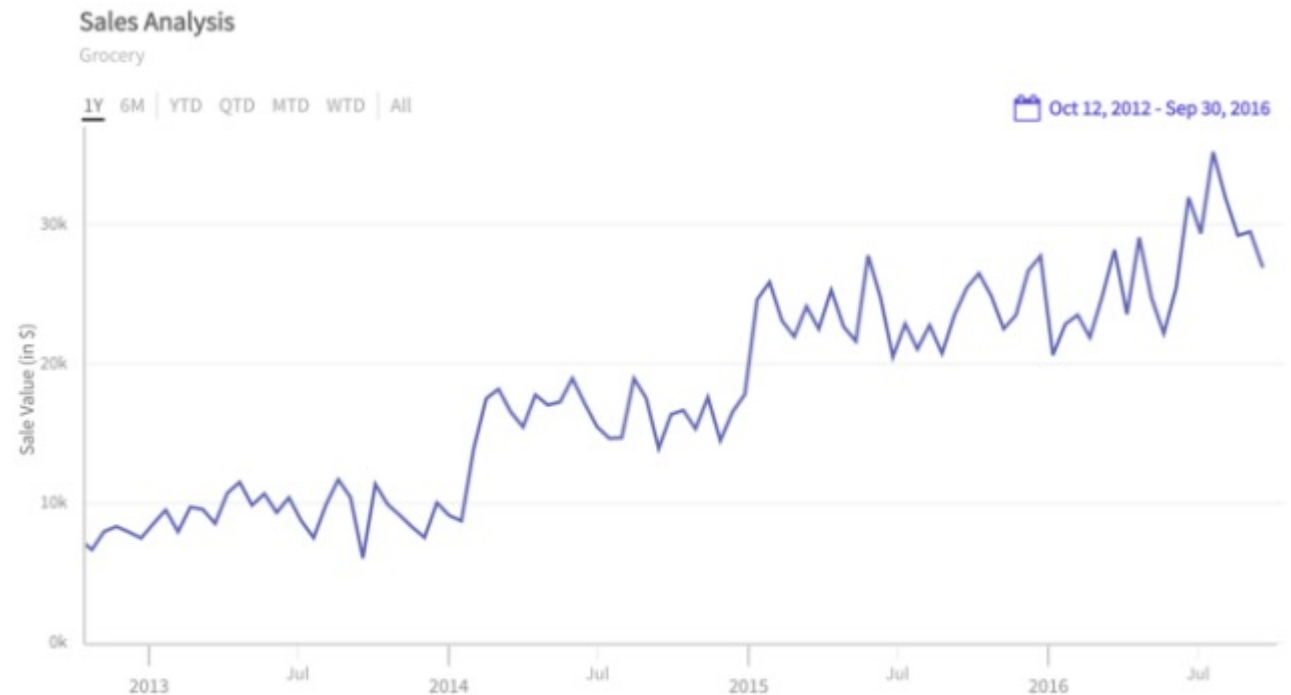
# Univariate Boxplot

- ▶ A boxplot is a visual representation of the feature's distribution. Specifically, it shows the Min, Q1, Median, Q3, and Maximum.
- ▶ They are a good tool to identify outliers.
  - ▶ If any outliers exist, the “whiskers” will reach out to  $1.5 * IQR$  ( $Q3 - Q1$ ) and outliers will be represented as dots beyond those.
- ▶ Bivariate Boxplots are especially good for showing differences between categorical groups (illustrated in relationships section.)



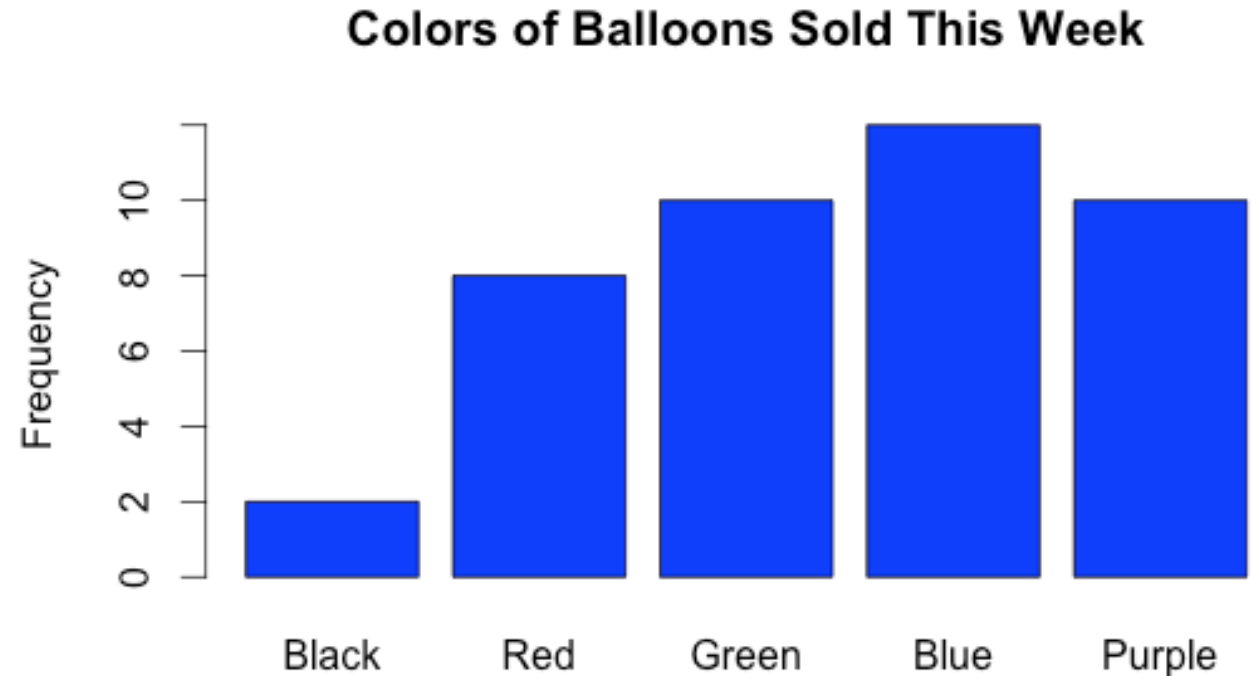
# Line Charts

- ▶ In the Data Structure and Unit of Analysis sections, we discussed time series data.
- ▶ For time-series measurement of a numeric variable, the best choice is a line chart.
- ▶ This chart has the date variable along the X-axis and the numeric variable on the Y-axis, showing the distribution of the numeric variable over time.
- ▶ Best practice dictates that the date variable is evenly distributed to clearly display the numeric distribution.



# Bar Charts

- ▶ The previous 4 visualization types are useful for examining numeric variable distributions.
- ▶ **For categorical variables**, a bar chart comparing count of observations by category is useful to understand the distribution.





# EXPLORE AND VISUALIZE RELATIONSHIPS

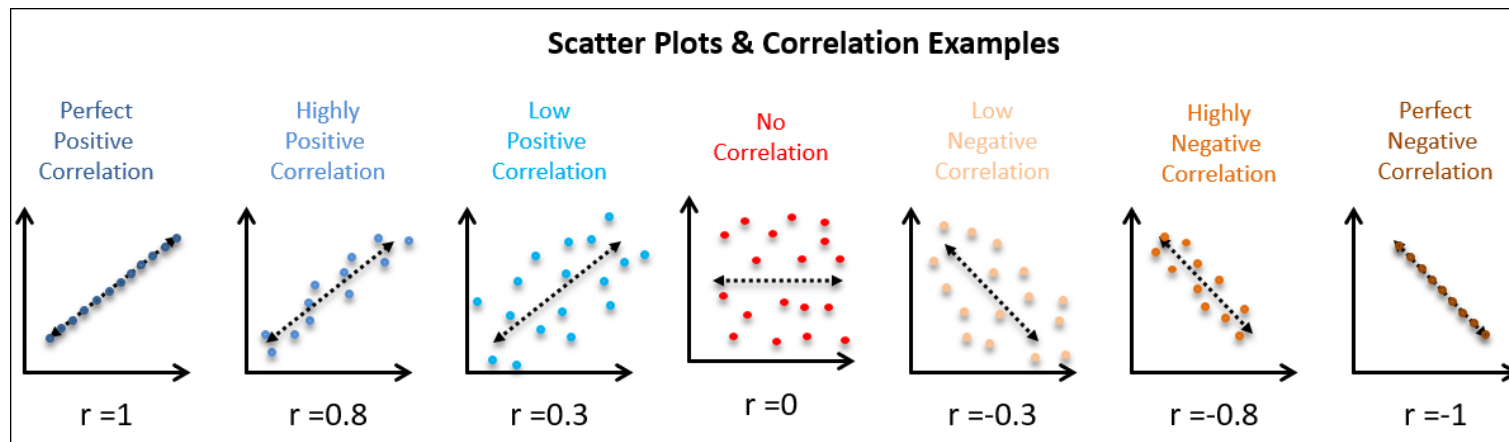
Leverage tools together to gain more power!

# Feature Relationships

- ▶ The basic building blocks of a predictive model are found in analyzing relationships between features.
- ▶ Finding features that are strongly related with each other can be a strong step towards pulling actionable insights from data exploration.
- ▶ Several methods exist for exploring these relationships. We will review:
  - ▶ Scatterplots
  - ▶ Correlation Matrices
  - ▶ Multi-variate Boxplots

# Scatterplots and Correlations

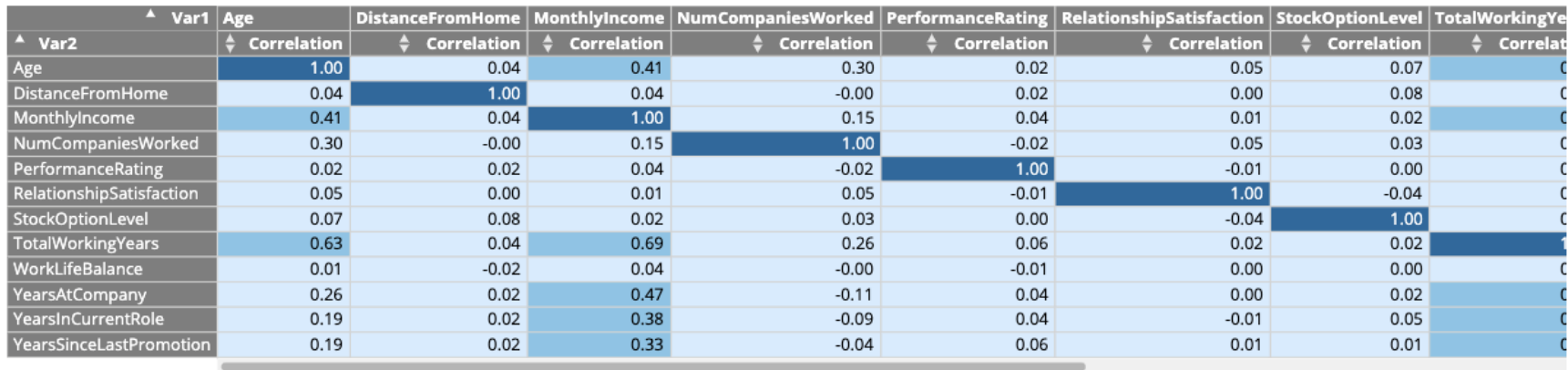
- ▶ Scatterplots are a visual representation of the relationship between 2 **numeric** variables.
- ▶ Correlation is a numeric measure of the strength of a relationship between 2 numeric variables.
  - ▶ It is represented by the lowercase  $r$ , measured on a scale of -1 to 1, and has 2 components:
    - ▶ Strength
      - ▶ The higher the absolute correlation value, the stronger the relationship between the 2 variables.
    - ▶ Direction
      - ▶ If 2 variables have a positive correlation value, this means that as the value of Variable 1 rises, the value of Variable 2 rises. The opposite is true of a negatively correlated relationship; as Variable 1 rises, Variable 2 will fall.



# Correlation Matrix

- ▶ A correlation matrix is a table that shows the strength and direction of relationship between all numeric variables in a dataset.
- ▶ A correlation matrix is a great starting point for understanding which variables in a dataset have relationships of interest.
- ▶ When 2 variables are highly correlated, there is a good chance that they will be predictive of each other.

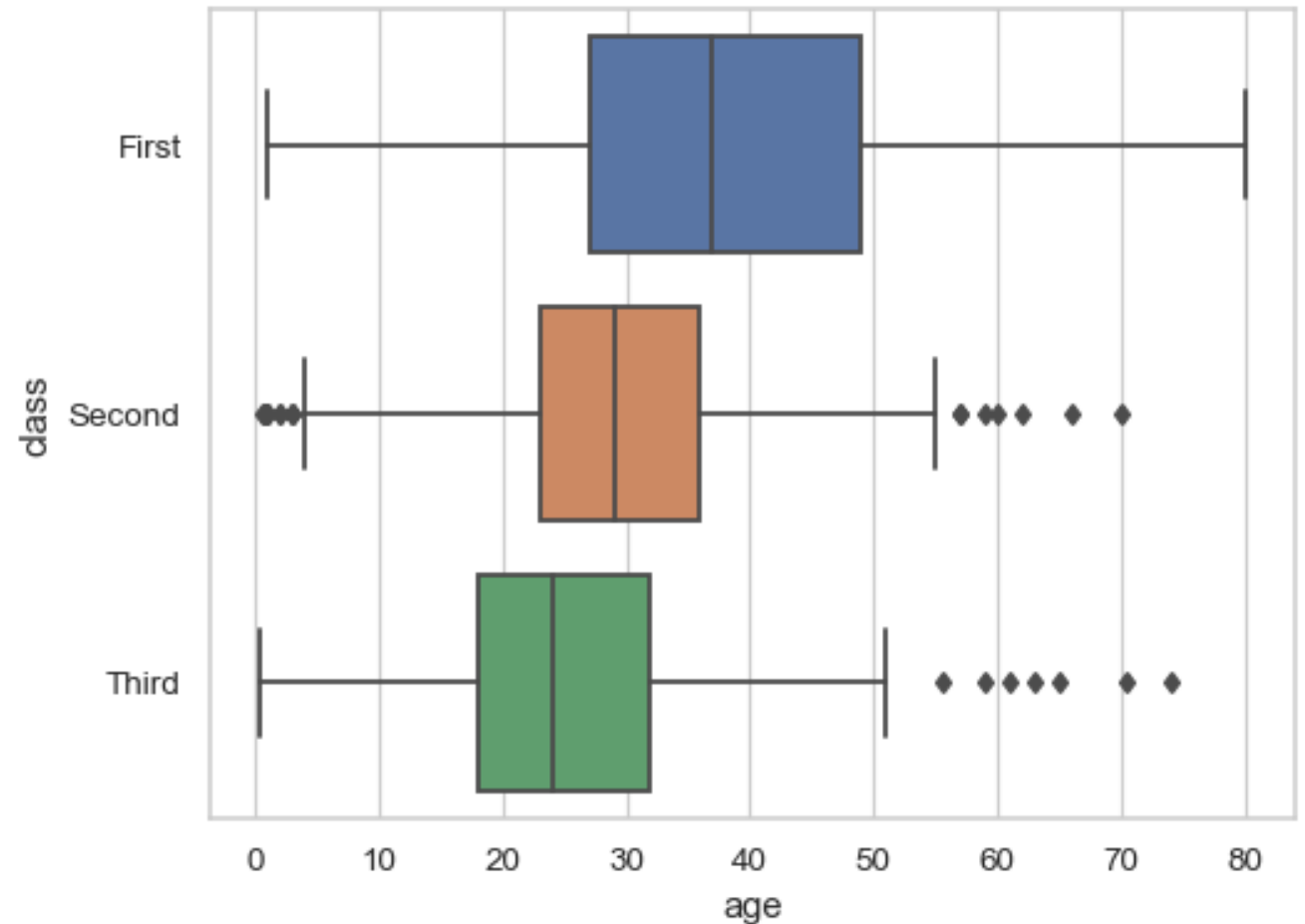
Correlation Matrix



Var1	Age	DistanceFromHome	MonthlyIncome	NumCompaniesWorked	PerformanceRating	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears
Age	1.00	0.04	0.41	0.30	0.02	0.05	0.07	0.63
DistanceFromHome	0.04	1.00	0.04	-0.00	0.02	0.00	0.08	0.04
MonthlyIncome	0.41	0.04	1.00	0.15	0.04	0.01	0.02	0.47
NumCompaniesWorked	0.30	-0.00	0.15	1.00	-0.02	0.05	0.03	-0.11
PerformanceRating	0.02	0.02	0.04	-0.02	1.00	-0.01	0.00	0.04
RelationshipSatisfaction	0.05	0.00	0.01	0.05	-0.01	1.00	-0.04	0.00
StockOptionLevel	0.07	0.08	0.02	0.03	0.00	-0.04	1.00	0.05
TotalWorkingYears	0.63	0.04	0.47	-0.11	0.04	0.00	0.05	1.00
WorkLifeBalance	0.01	-0.02	0.04	-0.00	-0.01	0.00	0.00	0.01
YearsAtCompany	0.26	0.02	0.38	-0.09	0.04	-0.01	0.01	0.33
YearsInCurrentRole	0.19	0.02	0.33	-0.04	0.06	0.01	0.01	0.33
YearsSinceLastPromotion	0.19	0.02	0.33	-0.04	0.06	0.01	0.01	0.33

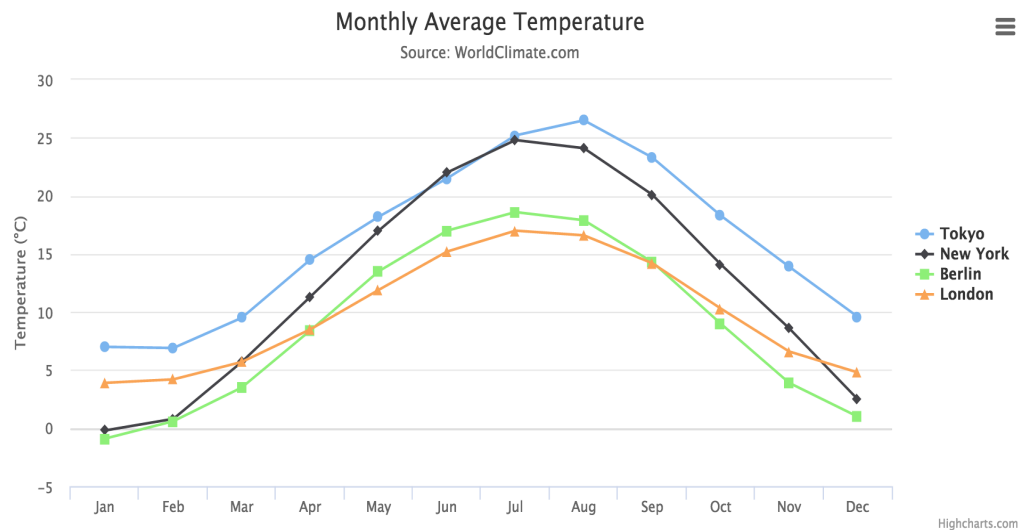
# Bivariate Boxplot

- ▶ As mentioned in previous slides, correlation and scatterplots show relationships between 2 numeric variables. **But what about categorical variables?**
- ▶ A bivariate boxplot is a great way to find relationships between **a categorical variable and a numeric variable**.
- ▶ This clearly answers the question, does a certain group or category behave differently than others?
- ▶ This chart type shows the distribution of the numeric variable for each category in the categorical variable.

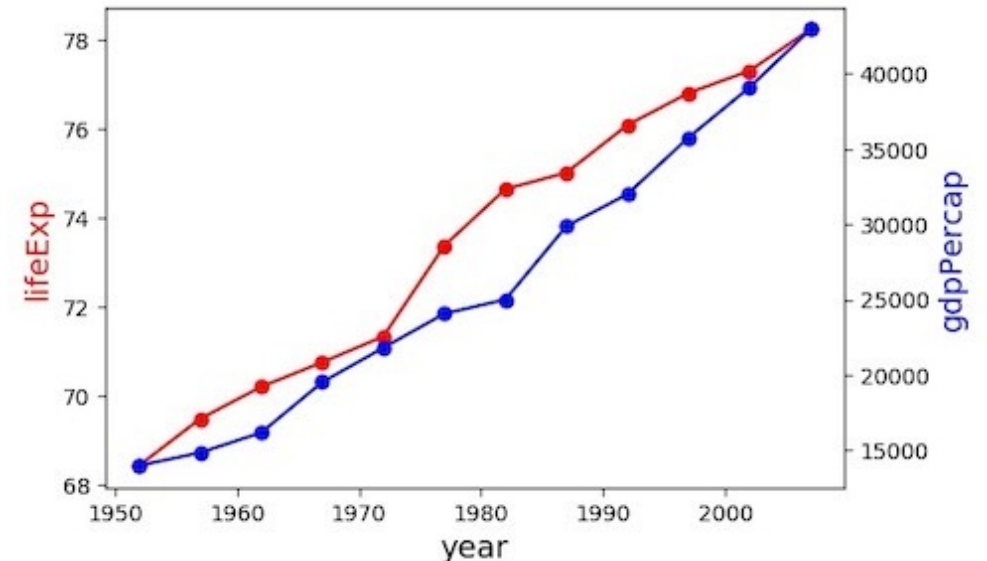


# Multi-Line Charts

- ▶ A multi-line chart is useful for 2 situations:
  - ▶ To show differences for 1 numeric variable over time across categories of a categorical variable.
  - ▶ To show multiple numeric variables over the same time period.
    - ▶ This should either be done for numeric features with similar scale. Or on 2 different scales, represented on the right and left of the Y-axis.



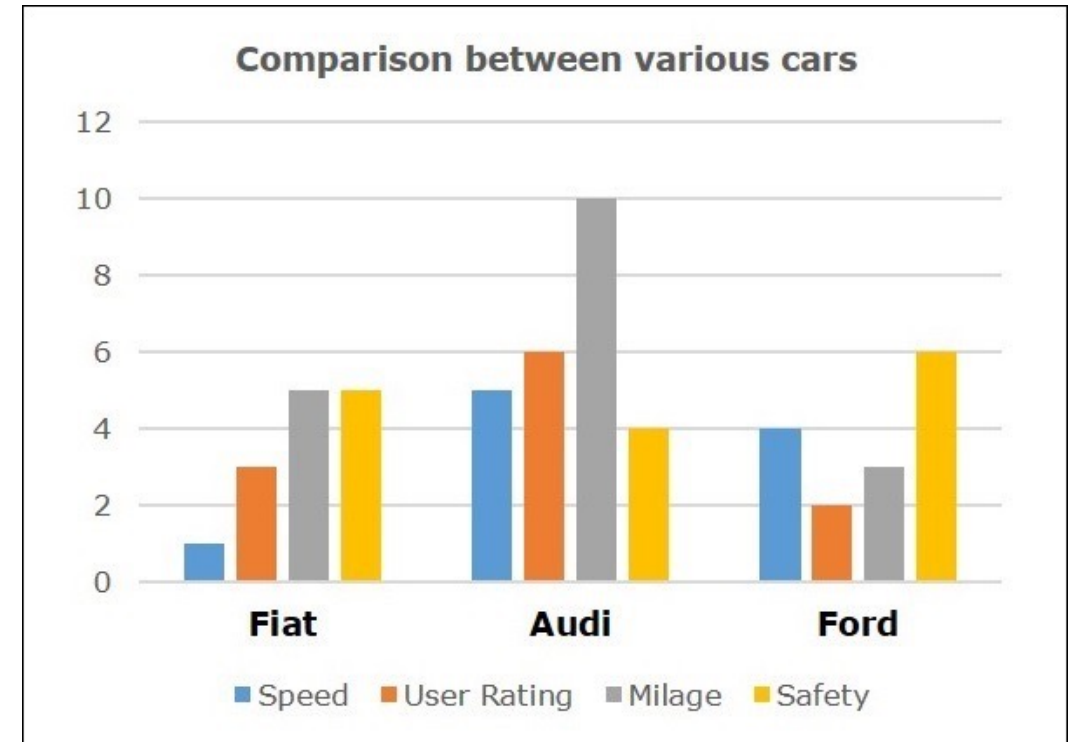
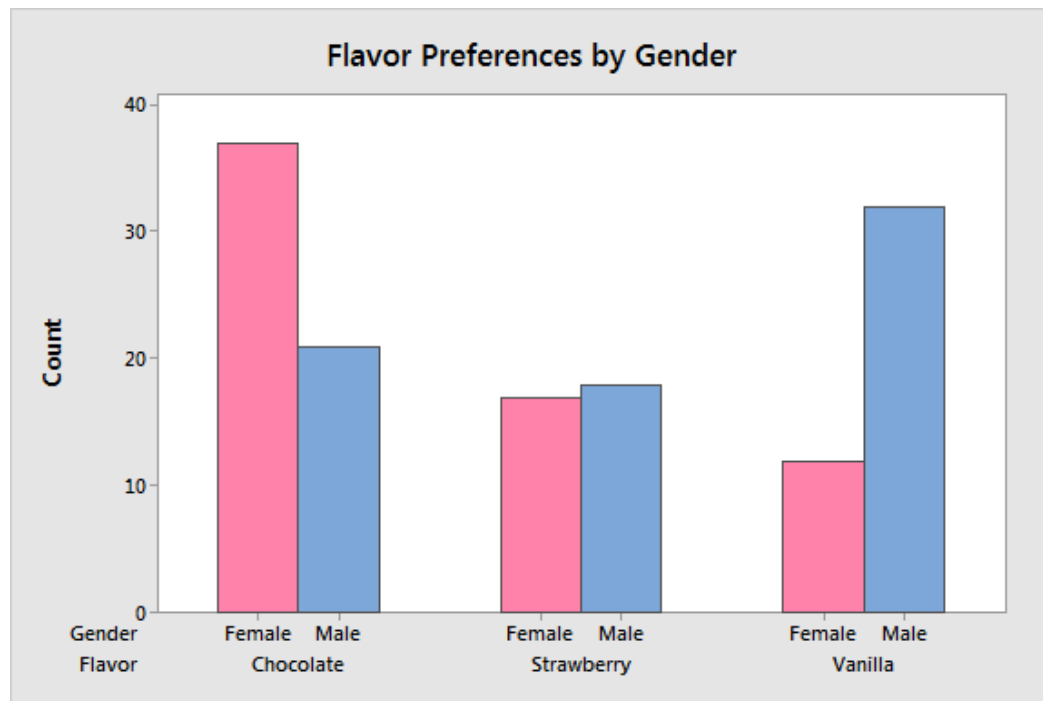
1 Numeric (Avg Temperature)  
1 Categorical (City)



2 Numeric (lifeExp & gdpPercap)  
2 Scales (Left & Right)

# Bar Chart

- ▶ Another useful visualization for categorical relationships are multi-variate bar charts.
- ▶ These can be used for comparing counts between multiple categorical variables or for comparing a numeric variable with a categorical variable.



# FEATURE ENGINEERING

Create new tools to fit our needs.



# Feature Engineering

- ▶ Often, further insights can be derived from the provided features by combining or transforming them.

# Transforming Variables for Analysis

## CATEGORICAL

- ▶ Categories often need to be represented numerically in many data science applications
- ▶ Use one hot encoding or create “dummy variables” for analysis (see next slide)

## ORDINAL

- ▶ Decide if measure should be treated as discrete or continuous in analysis
- ▶ This will depend on:
  - ▶ The modeling strategy
  - ▶ The distribution of the measure
  - ▶ The relationship between the measure and the dependent variable
- ▶ If measure will be treated as discrete:
  - ▶ Follow guidelines for Categorical measures
- ▶ If measure will be treated as continuous:
  - ▶ Follow guidelines for Interval/Ratio measures

## CONTINUOUS

- ▶ As a general rule, keep measure as is
- ▶ Determine if additional transformations are necessary
- ▶ This will depend on:
  - ▶ The modeling strategy
  - ▶ The distribution of the measure
  - ▶ The relationship between the measure and the dependent variable

# Dummy Variables

- ▶ Dummy variables/one hot encoding are data transformations that include adding a new column to the dataset for each category represented by a given categorical measure
- ▶ Each dummy variable takes on a value of 1 when the condition is true and 0 when the condition is false
- ▶ Some statistics packages will facilitate this transformation automatically
- ▶ Determine an appropriate reference group

Dummy Variables



ORIGINAL CATEGORICAL VARIABLE	Marketing_true	HR_true	Support_true	Sales_true	Management_true
Marketing	1	0	0	0	0
HR	0	1	0	0	0
Support	0	0	1	0	0
Marketing	1	0	0	0	0
Sales	0	0	0	1	0
HR	0	1	0	0	0
Support	0	0	1	0	0
Management	0	0	0	0	1
Sales	0	0	0	1	0
Sales	0	0	0	1	0
Management	0	0	0	0	1
HR	0	1	0	0	0

# Domo's Automated Data Exploration Tool: Data Profiler

# Domo Tool for Data Exploration: Data Profiler

- ▶ An easy-to-use pre-packaged Jupyter script
- ▶ Can be used for any dataset within the Domo instance
- ▶ Generates the following outputs in Domo dashboards:
  - ▶ Summary statistics table
  - ▶ Data profile
    - ▶ Histograms
    - ▶ Scatterplots
    - ▶ Boxplots
  - ▶ Correlation matrix

# Domo Tool for Data Exploration: Data Profiler

Summary Statistics

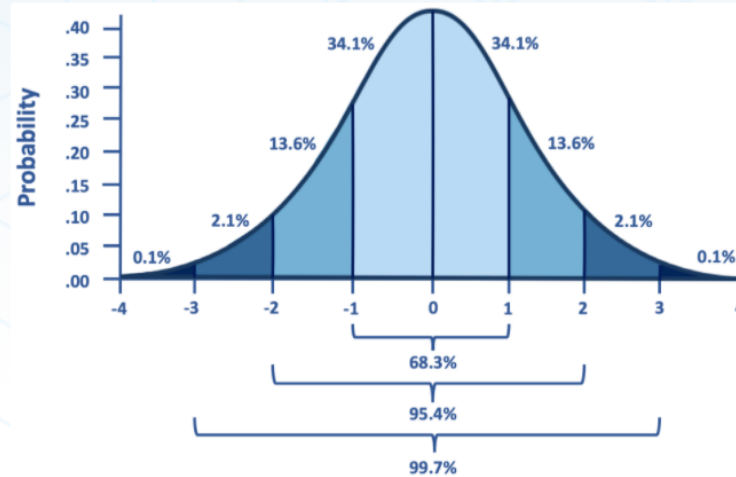
1,002 Total Rows



Column Name	min	25%	mean	50%	75%	max	count
Age	18.00	30.00	36.35	35.00	42.00	60.00	1,002.00
AttritionNumber	0.00	0.00	0.16	0.00	0.00	1.00	1,002.00
DailyRate	102.00	472.50	817.55	818.50	1,177.50	1,499.00	1,002.00
DistanceFromHome	1.00	2.00	9.47	7.00	15.00	29.00	1,002.00
Education	1.00	2.00	2.92	3.00	4.00	5.00	1,002.00
EmployeeNumber	2.00	483.25	1,041.40	1,064.00	1,587.50	2,068.00	1,002.00
EnvironmentSatisfaction	1.00	2.00	2.71	3.00	4.00	4.00	1,002.00
HourlyRate	30.00	48.00	65.82	65.00	83.00	100.00	1,002.00
JobInvolvement	1.00	2.00	2.71	3.00	3.00	4.00	1,002.00
JobLevel	1.00	1.00	1.89	2.00	2.00	4.00	1,002.00
JobSatisfaction	1.00	2.00	2.72	3.00	4.00	4.00	1,002.00
MonthlyIncome	1,009.00	2,853.75	5,754.29	4,762.00	7,101.50	17,426.00	1,002.00
MonthlyRate	2,094.00	8,054.50	14,421.79	14,575.50	20,872.00	26,999.00	1,002.00
NumCompaniesWorked	0.00	1.00	2.70	2.00	4.00	9.00	1,002.00
PerformanceRating	3.00	3.00	3.15	3.00	3.00	4.00	1,002.00
RelationshipSatisfaction	1.00	2.00	2.70	3.00	4.00	4.00	1,002.00
StandardHours	80.00	80.00	80.00	80.00	80.00	80.00	1,002.00
StockOptionLevel	0.00	0.00	0.80	1.00	1.00	3.00	1,002.00
TotalWorkingYears	0.00	6.00	10.52	9.00	14.00	40.00	1,002.00

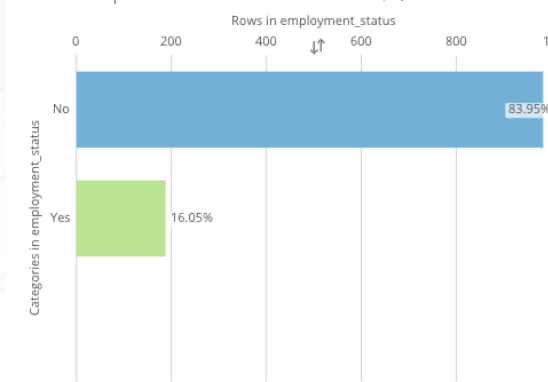
# Domo Tool for Data Exploration: Data Profiler

## Histogram Data Profile



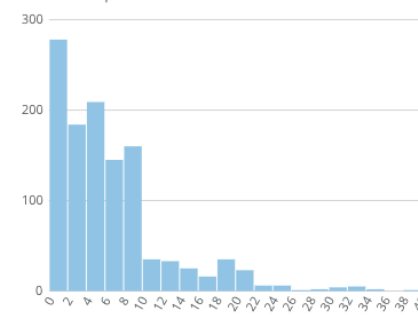
## employment\_status

1 Nulls | 0% of all rows NULL Count employment\_status



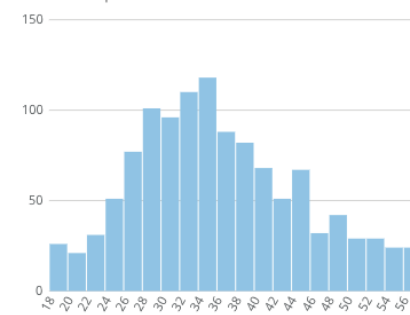
## YearsAtCompany Histogram

2 Nulls | 0% of all rows NULL Count YearsAtComp



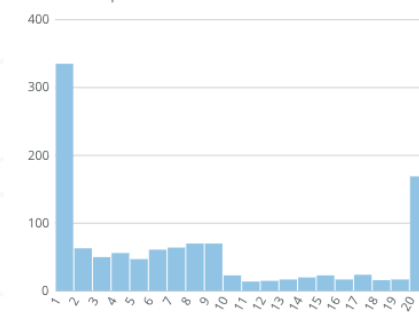
## Age Histogram

5 Nulls | 0% of all rows NULL Count Age



## DistanceFromHome Histogram

1 Nulls | 0% of all rows NULL Count DistanceFrom



# Domo Tool for Data Evaluation: Data Profiler

Correlation Matrix



^ Var1	Age	DistanceFromHome	MonthlyIncome	NumCompaniesWorked	PerformanceRating	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYe
^ Var2	Correlation	Correlation	Correlation	Correlation	Correlation	Correlation	Correlation	Correlat
Age	1.00	0.04	0.41	0.30	0.02	0.05	0.07	C
DistanceFromHome	0.04	1.00	0.04	-0.00	0.02	0.00	0.08	C
MonthlyIncome	0.41	0.04	1.00	0.15	0.04	0.01	0.02	C
NumCompaniesWorked	0.30	-0.00	0.15	1.00	-0.02	0.05	0.03	C
PerformanceRating	0.02	0.02	0.04	-0.02	1.00	-0.01	0.00	C
RelationshipSatisfaction	0.05	0.00	0.01	0.05	-0.01	1.00	-0.04	C
StockOptionLevel	0.07	0.08	0.02	0.03	0.00	-0.04	1.00	C
TotalWorkingYears	0.63	0.04	0.69	0.26	0.06	0.02	0.02	C
WorkLifeBalance	0.01	-0.02	0.04	-0.00	-0.01	0.00	0.00	C
YearsAtCompany	0.26	0.02	0.47	-0.11	0.04	0.00	0.02	C
YearsInCurrentRole	0.19	0.02	0.38	-0.09	0.04	-0.01	0.05	C
YearsSinceLastPromotion	0.19	0.02	0.33	-0.04	0.06	0.01	0.01	C