Introduction to Data Science

Data Science Essentials

NASHVILLE
SOFTWARE
SCHOOL

Goals for today

- **Review last session coding tasks**
- **Exploratory data analysis**

Review last session coding tasks

week2_review notebook

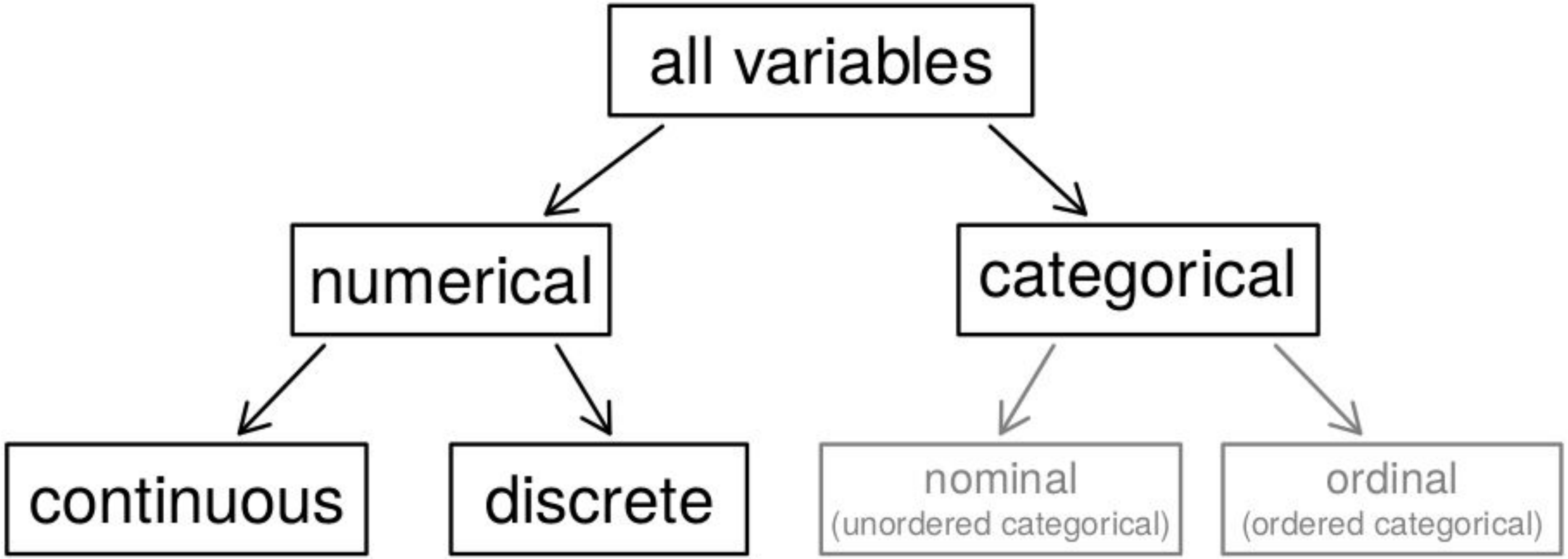
Data

Data is characteristics or information gathered through observation.

When conducting a study, we will be collecting data about **observations/cases/subjects**.

We will often collect a number of **variables** (height, weight, age, sex, annual income, favorite color, number of pets, etc.).

Types of Variables



Borrowed from the OpenIntro Stats Book:
https://leanpub.com/openintro-statistics

Types of Variables

Numerical Variables: Also called quantitative variables. A *measurement* with numerical meaning. Numerical variables can be measured using a scale or count.

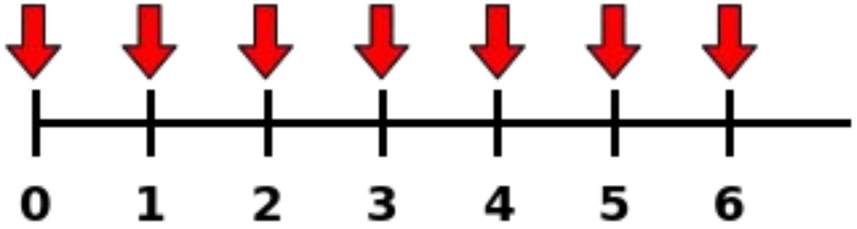
Divided into two types:

1. **Discrete Variables:** All possible outcomes can be listed
2. **Continuous Variables:** Can take on any value in a range

Types of Variables

Number of Pets:

Limited set of distinct values, meaning that it is a *discrete* variable. (You can't have 2.83 pets)



Temperature:

We can't list all possible temperatures, so it is *continuous* variable.

Types of Variables

Categorical Variables: Also called qualitative variables. Non-numeric data which falls into some number of **levels**.

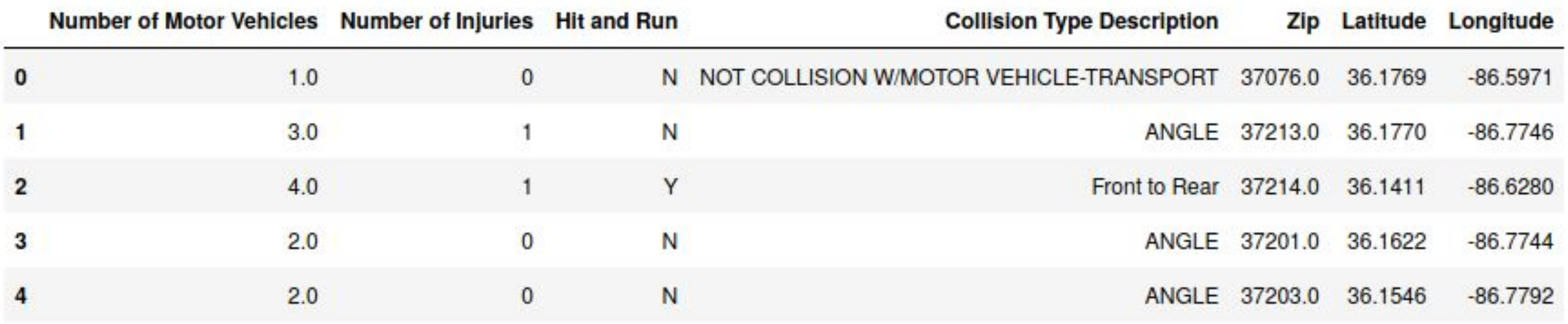
Caution: sometimes categorical variable are coded with a number

(eg. 0 = small, 1 = medium, 2 = large)

Divided into two types:

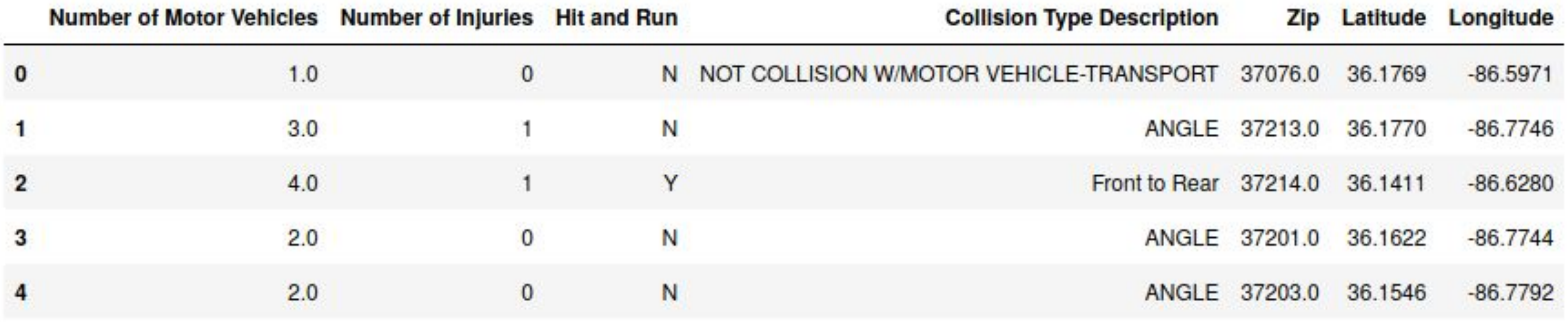
1. **Ordinal Variables:** Have a natural/intrinsic ordering (eg. grade level, small/medium/large)
2. **Nominal Variables:** No natural ordering (eg. gender, religious affiliation)

Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

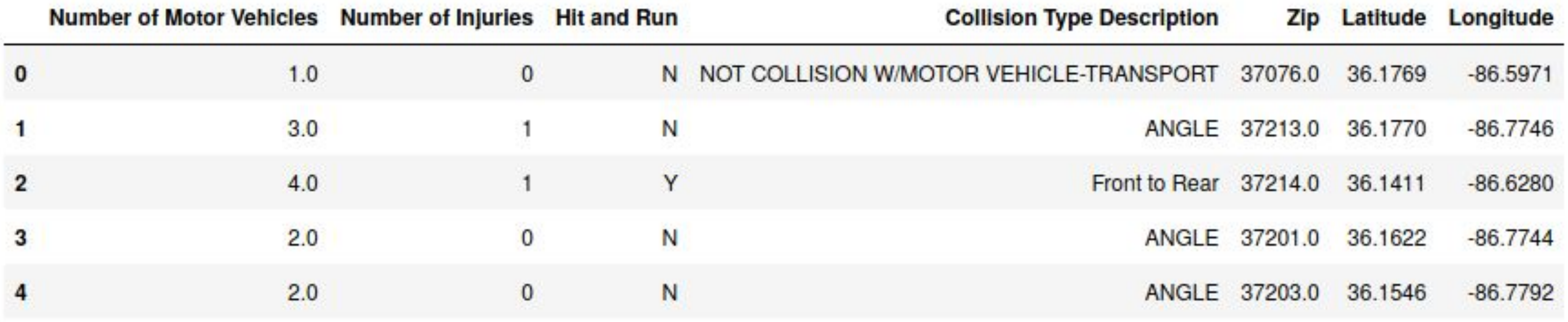
Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

Discrete Numeric

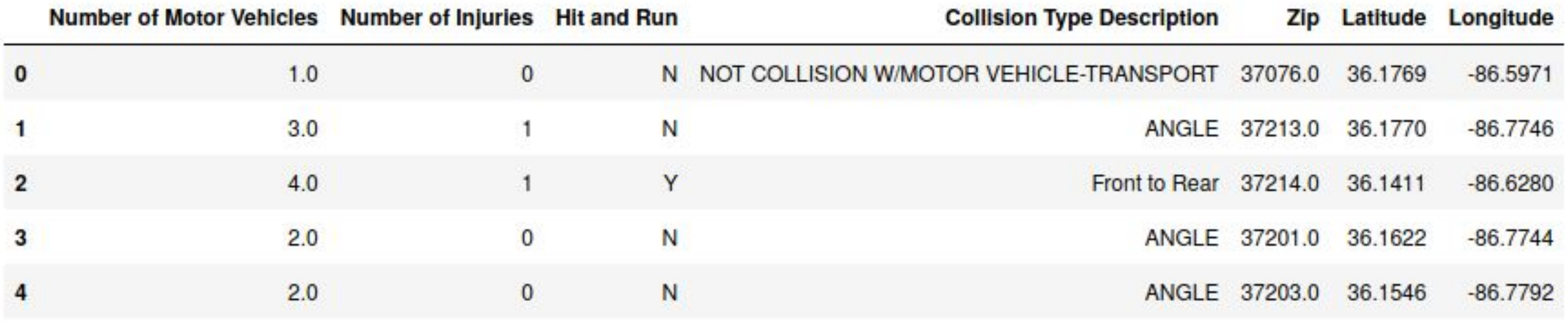
Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

Discrete Numeric

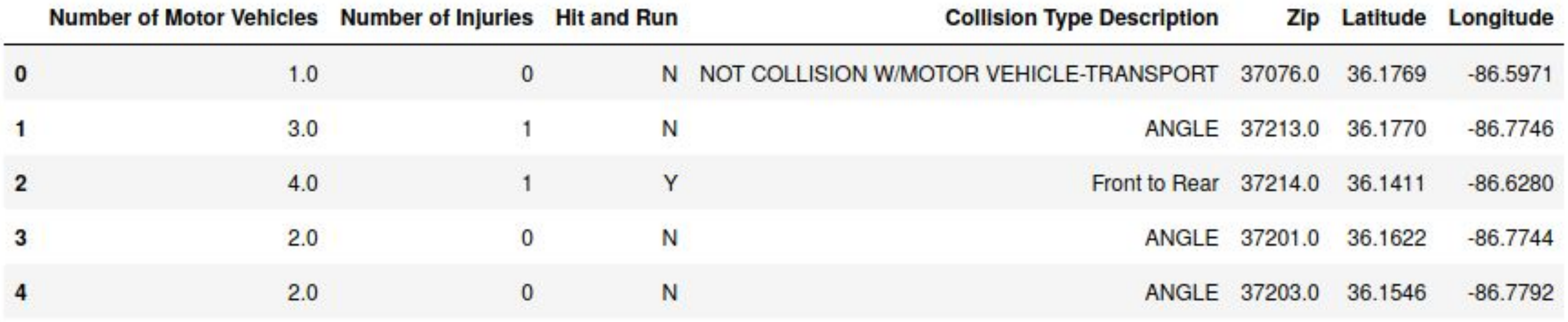
Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

Nominal Categorical

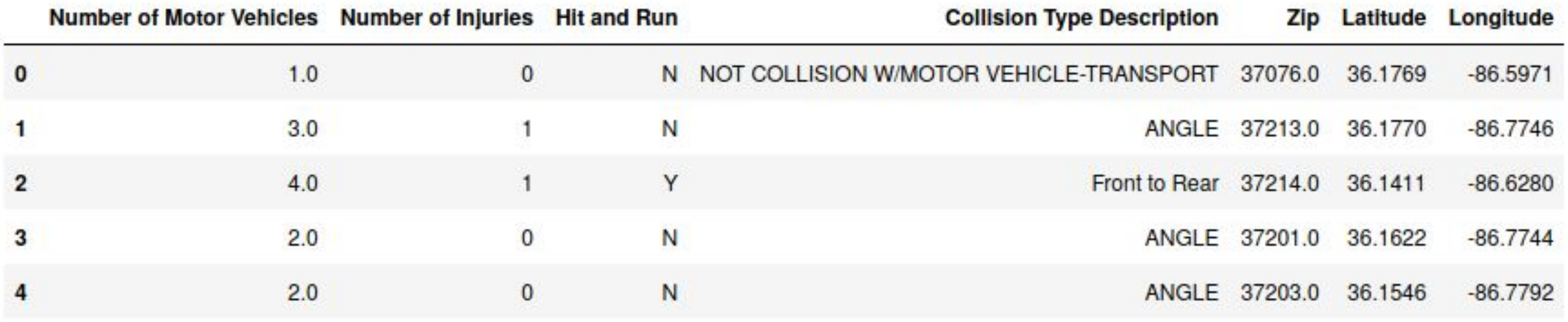
Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

Nominal Categorical

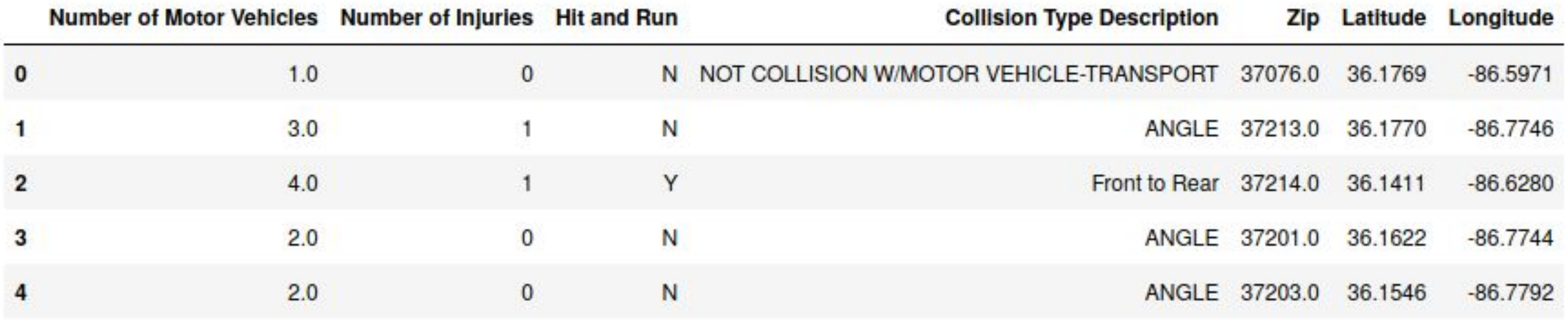
Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

Nominal Categorical

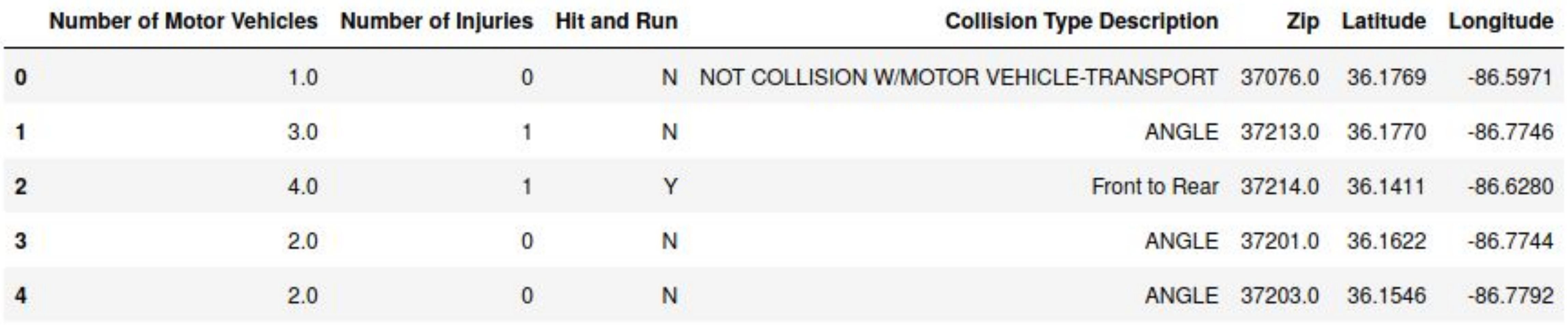
Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

Continuous Numeric

Types of Variables



	Number of Motor Vehicles	Number of Injuries	Hit and Run	Collision Type Description	Zip	Latitude	Longitude
0	1.0	0	N	NOT COLLISION W/MOTOR VEHICLE-TRANSPORT	37076.0	36.1769	-86.5971
1	3.0	1	N	ANGLE	37213.0	36.1770	-86.7746
2	4.0	1	Y	Front to Rear	37214.0	36.1411	-86.6280
3	2.0	0	N	ANGLE	37201.0	36.1622	-86.7744
4	2.0	0	N	ANGLE	37203.0	36.1546	-86.7792

Continuous Numeric

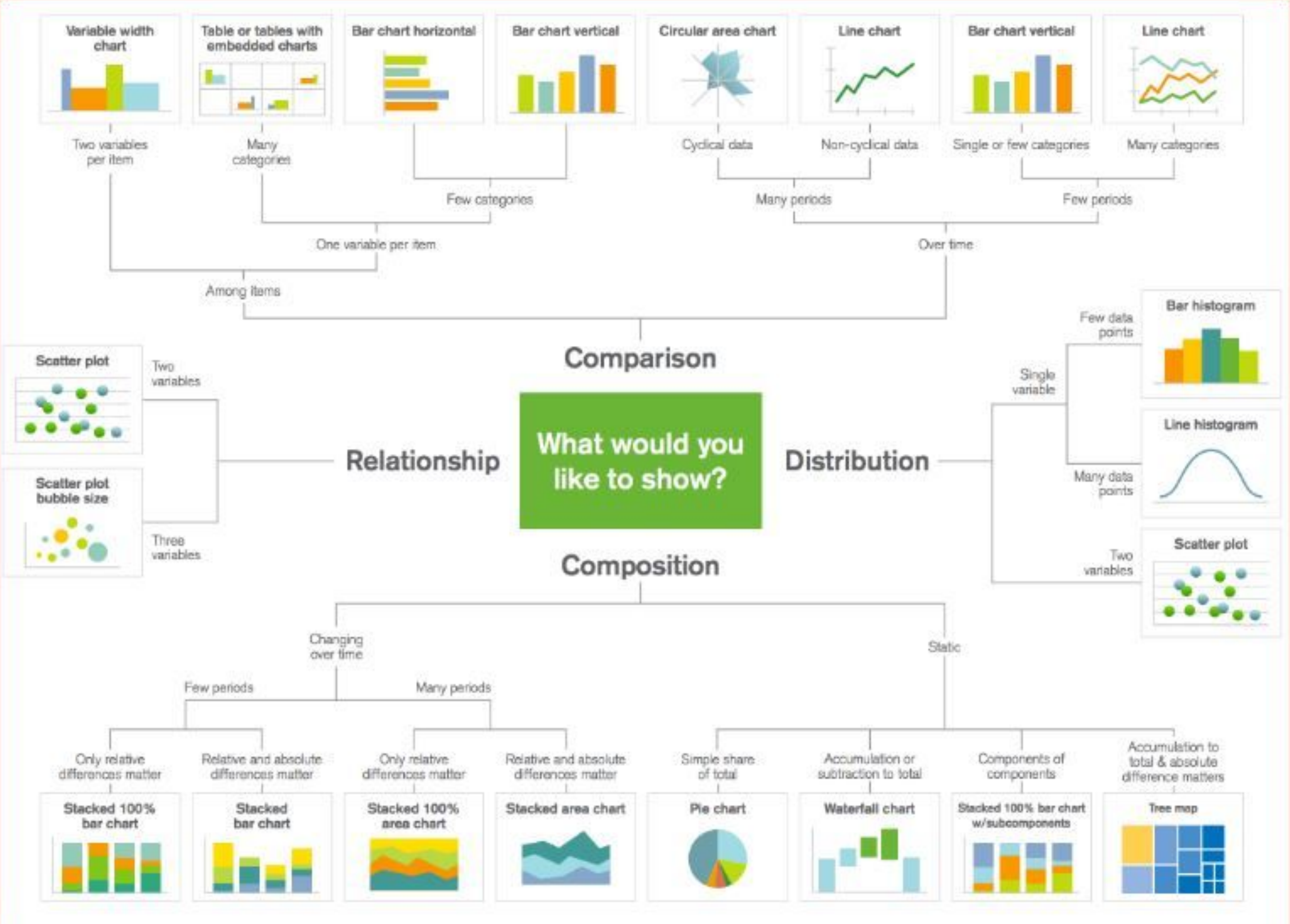
Exploratory Data Analysis

When we encounter a new dataset, we will often want to familiarize ourselves with the properties of that dataset.

It is very hard to understand characteristics of your dataset just by looking at it in tabular form, especially if you have more than a handful of observations (which you almost always will).

Exploratory Data Analysis (EDA) is the process of analyzing a dataset in order to summarize the main characteristics and generate potential ideas and hypotheses.

EDA can be done through plots or charts (visual EDA) or through numeric summaries (numeric EDA).



What would you like to show?

Comparison

Among items

Two variables per item

Variable width chart

One variable per item

Many categories

Table or tables with embedded charts

Few categories

Bar chart horizontal

Bar chart vertical

Over time

Many periods

Cyclical data

Circular area chart

Non-cyclical data

Line chart

Few periods

Single or few categories

Bar chart vertical

Many categories

Line chart

Relationship

Two variables

Scatter plot

Three variables

Scatter plot bubble size

Distribution

Single variable

Few data points

Bar histogram

Many data points

Line histogram

Two variables

Scatter plot

Composition

Changing over time

Few periods

Only relative differences matter

Stacked 100% bar chart

Relative and absolute differences matter

Stacked bar chart

Many periods

Only relative differences matter

Stacked 100% area chart

Relative and absolute differences matter

Stacked area chart

Static

Simple share of total

Pie chart

Accumulation or subtraction to total

Waterfall chart

Components of components

Stacked 100% bar chart w/subcomponents

Accumulation to total & absolute difference matters

Tree map

Source: ©A. Abela, 2010. www.ExtremePresentation.com

Questions?

NASHVILLE
SOFTWARE
SCHOOL