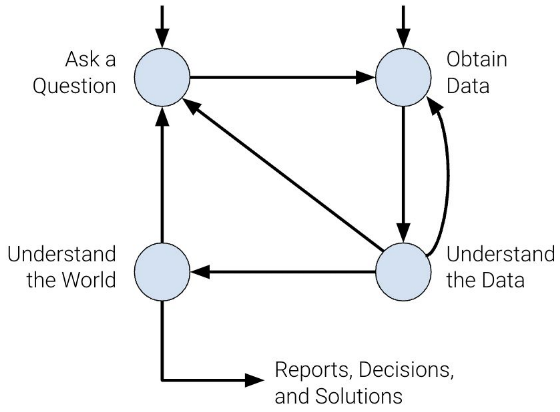
Data Cleaning and EDA

Introduction to Data Science
Spring 1403

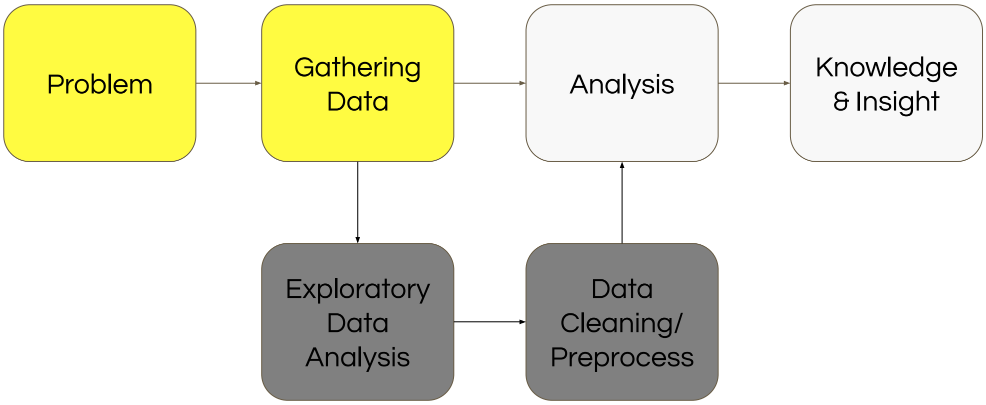
Yadollah Yaghoobzadeh

Most slides are from ADS course by Hesam Salavati.
Some from DS100 UC Berkeley.

Data Science Lifecycle



Data Analysis Pipeline



Problem

Gathering Data

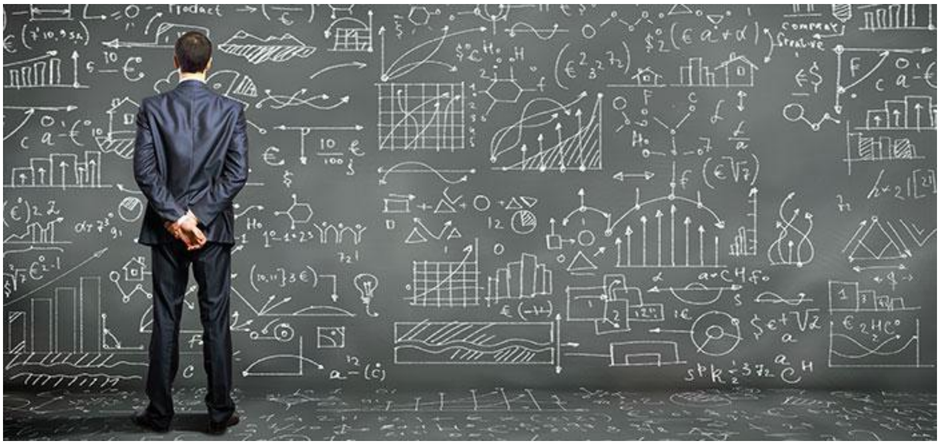
Analysis

Knowledge & Insight

Exploratory Data Analysis

Data Cleaning/ Preprocess

Exploratory Data Analysis



Credit https://www.linkedin.com/pulse/data-answer-needs-managed-colm-sherwin

Exploratory Data Analysis (EDA)

In essence, the purpose of EDA is to get to know your data and the problem better by:

- Summarizing it
- Visualizing it
- Looking for patterns
- Check for missing variables
- ...

Credit: http://querytreeapp.com/blog/make-sense-with-data-visualization/

Why EDA is Important

“If you skip this step then you might end up generating inaccurate models and choosing the insignificant variables in your model.”

“It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

John Tukey, developer of Exploratory Data Analysis

EDA Steps

1. Talk to the data owners, understand the context and the task
2. Check dataset head, shape and summary
3. Go over data columns, check their type, range, etc.
4. Check for missing data in columns
5. Get a glance of data distribution (visualization)
6. Find correlated features (columns)
7. Check for outliers

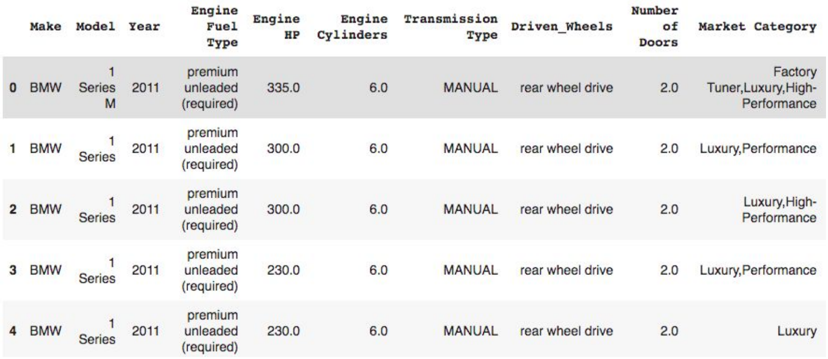
Example Dataset

A data set of around 12000 cars:

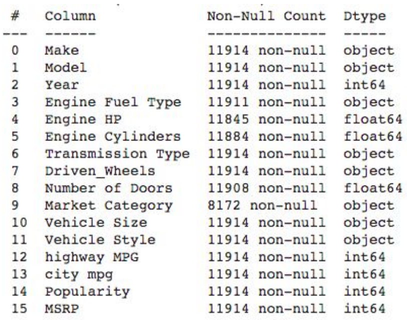
- Make
- Model
- Power
- Size
- MPG
- Year
- Price
- ….

Dataset Shape, Head, Summary

Shape:
(11914, 16)



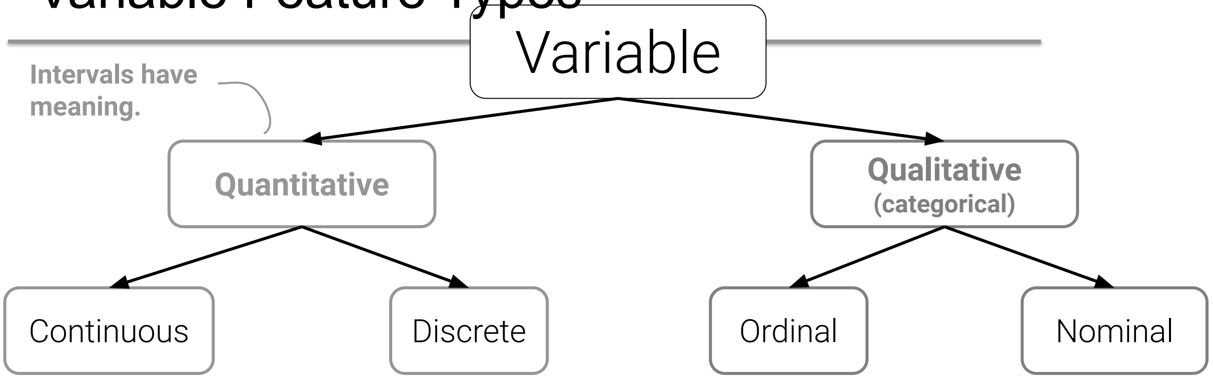
	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury



```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Make               11914 non-null  object
 1   Model              11914 non-null  object
 2   Year               11914 non-null  int64
 3   Engine Fuel Type   11911 non-null  object
 4   Engine HP          11845 non-null  float64
 5   Engine Cylinders   11884 non-null  float64
 6   Transmission Type  11914 non-null  object
 7   Driven_Wheels      11914 non-null  object
 8   Number of Doors    11908 non-null  float64
 9   Market Category    8172 non-null   object
 10  Vehicle Size       11914 non-null  object
 11  Vehicle Style      11914 non-null  object
 12  highway MPG        11914 non-null  int64
 13  city mpg           11914 non-null  int64
 14  Popularity         11914 non-null  int64
 15  MSRP               11914 non-null  int64
```

What Would You Do Next to Get More Insights from the Data?

Variable Feature Types



Continuous: Could be measured to arbitrary precision.

Examples:
- Price
- Temperature

Discrete: Finite possible values

Examples:
- Number of siblings
- Yrs of education

Ordinal: Categories w/ ordered levels; no consistent meaning to difference

Examples:
- Preferences
- Level of education

Nominal: Categories w/ no specific ordering.

Examples:
- Political Affiliation
- Cal ID number

Note that **qualitative variables** could have numeric levels; conversely, **quantitative variables** could be stored as strings!

Understand What Data Represents

❑ Granularity

❑ Scope

❑ Temporality

Granularity

- The granularity of a dataset is what a single row represents.
- To determine the data’s granularity, ask: what does each row in the dataset represent?
- For example
  - each record may represent one person.
  - each record may represent a group of people.

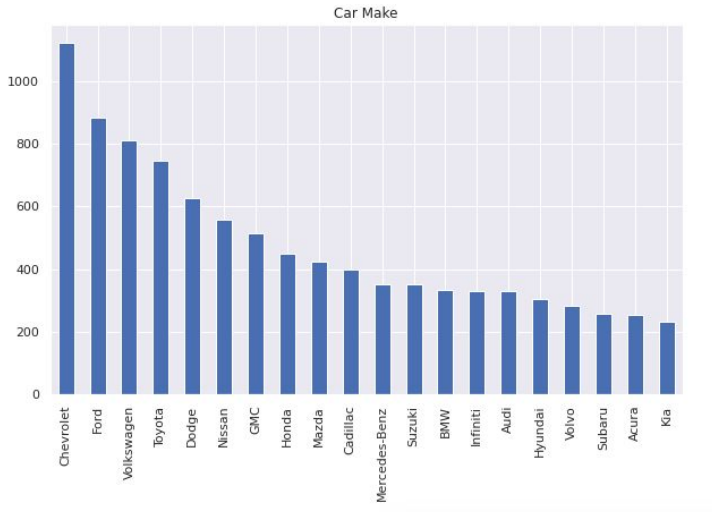
Scope

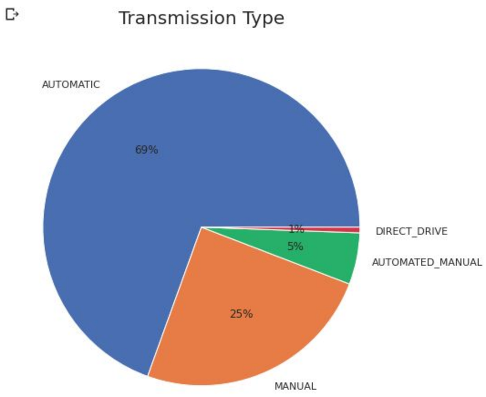
- The scope of a dataset is the subset of the population covered by the data.
- For example
  - students in our DS class
  - students in UT
  - Students in Tehran

Temporality

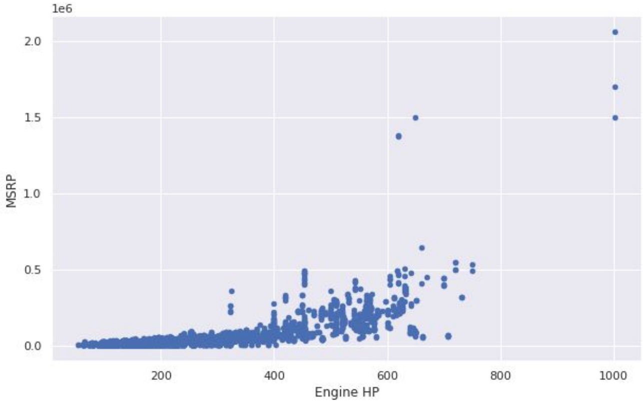
- The temporality of a dataset describes the periodicity over which the data was collected as well as when the data was most recently collected or updated.
- Time and date fields of a dataset could represent a few things:
  - when the “event” happened
  - when the data was collected, or when it was entered into the system
  - when the data was copied into the database

Getting Some Insight From the Data





Getting Some Insight From the Data



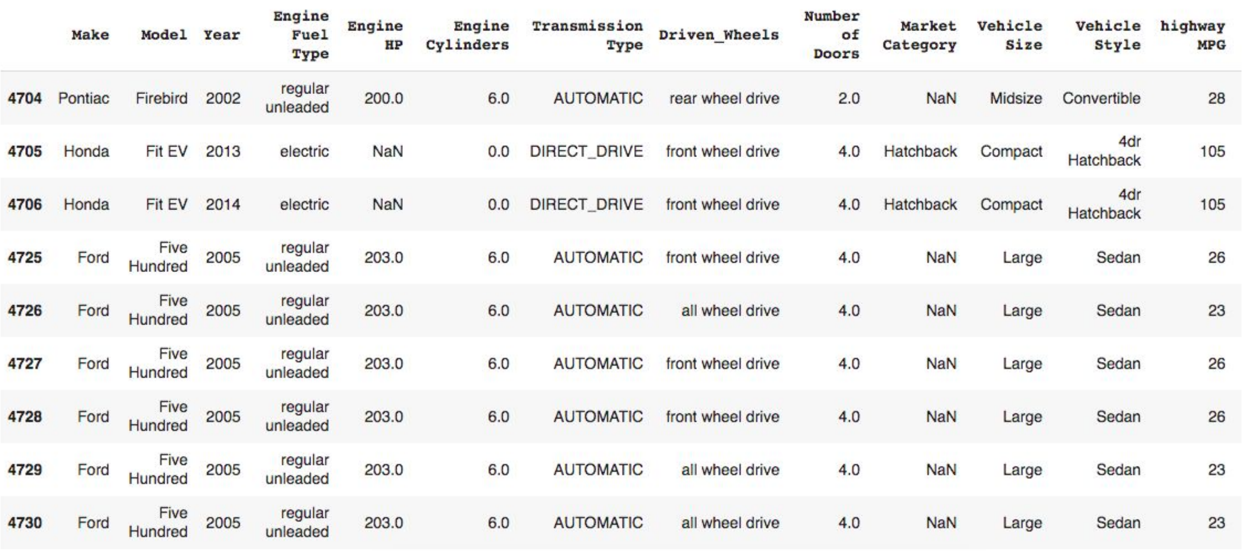
Data Cleaning

Origins of Dirty Data

- **Incomplete data:** fields left blank
- **Incorrect data:** invalid range, not-validated inputs, etc.
- **Inconsistent data:** different versions, different forms (VP/Vice President), etc.
- **Duplicate data**
- **Inaccurate data:** fake email address, etc.
- **Old data:** changed phone numbers, addresses, etc.

- Causes
  - Human error
  - Hardware/Software/etc failure

Incomplete/Missing Data



	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG
4704	Pontiac	Firebird	2002	regular unleaded	200.0	6.0	AUTOMATIC	rear wheel drive	2.0	NaN	Midsize	Convertible	28
4705	Honda	Fit EV	2013	electric	NaN	0.0	DIRECT_DRIVE	front wheel drive	4.0	Hatchback	Compact	4dr Hatchback	105
4706	Honda	Fit EV	2014	electric	NaN	0.0	DIRECT_DRIVE	front wheel drive	4.0	Hatchback	Compact	4dr Hatchback	105
4725	Ford	Five Hundred	2005	regular unleaded	203.0	6.0	AUTOMATIC	front wheel drive	4.0	NaN	Large	Sedan	26
4726	Ford	Five Hundred	2005	regular unleaded	203.0	6.0	AUTOMATIC	all wheel drive	4.0	NaN	Large	Sedan	23
4727	Ford	Five Hundred	2005	regular unleaded	203.0	6.0	AUTOMATIC	front wheel drive	4.0	NaN	Large	Sedan	26
4728	Ford	Five Hundred	2005	regular unleaded	203.0	6.0	AUTOMATIC	front wheel drive	4.0	NaN	Large	Sedan	26
4729	Ford	Five Hundred	2005	regular unleaded	203.0	6.0	AUTOMATIC	all wheel drive	4.0	NaN	Large	Sedan	23
4730	Ford	Five Hundred	2005	regular unleaded	203.0	6.0	AUTOMATIC	all wheel drive	4.0	NaN	Large	Sedan	23

How should we deal with missing data?

Handling Incomplete/Missing Data

Strategies to preserve data points:

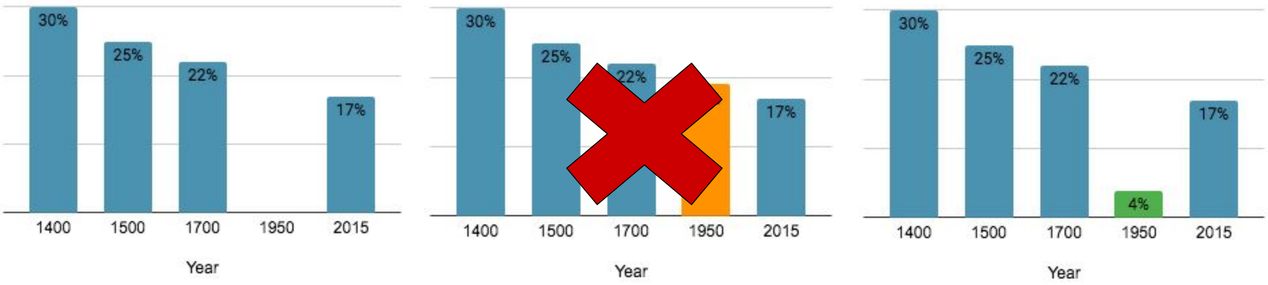
- Find accurate values for missing data
- Guess/Estimate missing values (Imputing)
- Just copy/pasting missing data from similar rows

Handling Incomplete/Missing Data

Ignoring the missing data

- Drop the column(s) with most missing values
- Drop the rows containing missing data

Handling Incomplete/Missing Data: Caution



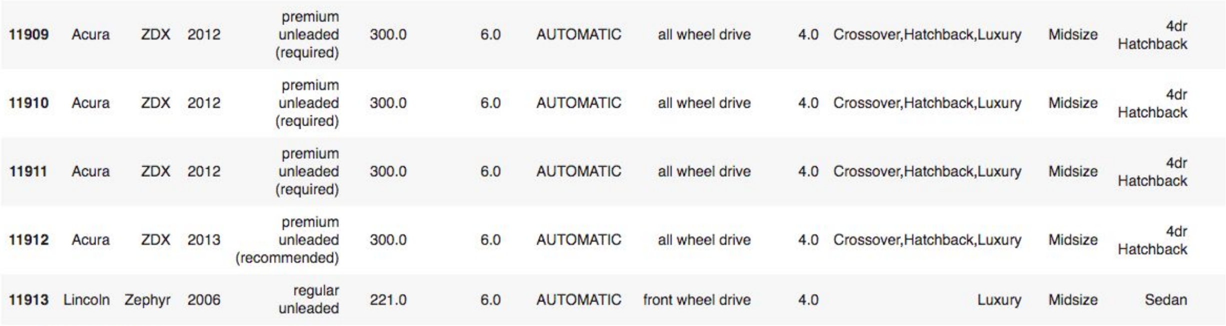
China's Share of Worldwide GDP

https://www.businessinsider.com/history-of-chinese-economy-1200-2017-2017-1

Incomplete/Missing data

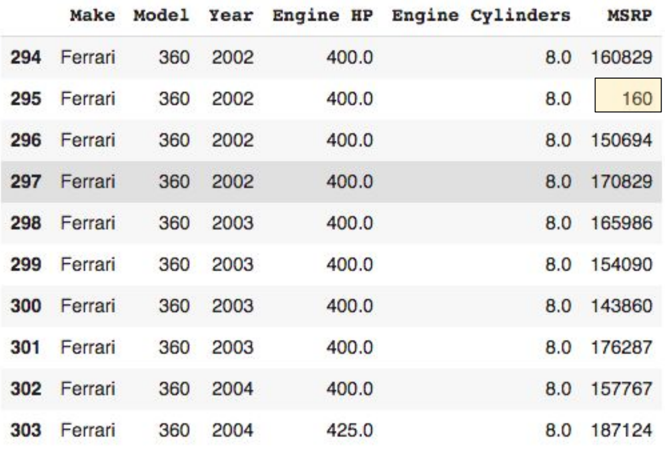
For our sample dataset:

- Find accurate values for missing data
- Drop a non-important column
- Drop the remaining rows



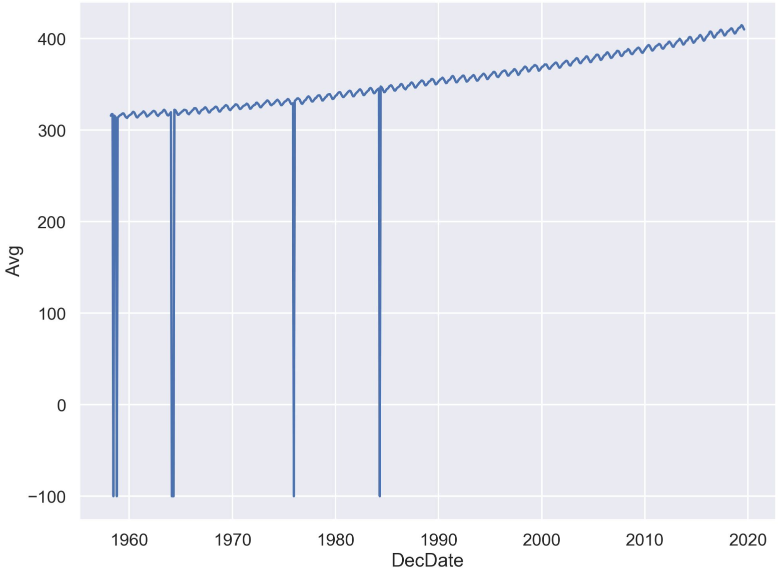
11909	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Crossover,Hatchback,Luxury	Midsize	4dr Hatchback
11910	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Crossover,Hatchback,Luxury	Midsize	4dr Hatchback
11911	Acura	ZDX	2012	premium unleaded (required)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Crossover,Hatchback,Luxury	Midsize	4dr Hatchback
11912	Acura	ZDX	2013	premium unleaded (recommended)	300.0	6.0	AUTOMATIC	all wheel drive	4.0	Crossover,Hatchback,Luxury	Midsize	4dr Hatchback
11913	Lincoln	Zephyr	2006	regular unleaded	221.0	6.0	AUTOMATIC	front wheel drive	4.0	Luxury	Midsize	Sedan

Data Cleaning: Invalid Data



	Make	Model	Year	Engine HP	Engine Cylinders	MSRP
294	Ferrari	360	2002	400.0	8.0	160829
295	Ferrari	360	2002	400.0	8.0	160
296	Ferrari	360	2002	400.0	8.0	150694
297	Ferrari	360	2002	400.0	8.0	170829
298	Ferrari	360	2003	400.0	8.0	165986
299	Ferrari	360	2003	400.0	8.0	154090
300	Ferrari	360	2003	400.0	8.0	143860
301	Ferrari	360	2003	400.0	8.0	176287
302	Ferrari	360	2004	400.0	8.0	157767
303	Ferrari	360	2004	425.0	8.0	187124

Identifying



How to Identify Invalid Data?

Invalid Data: What to Look For?

A great summary from Wikipedia:

- **Data-Type Constraints:** values in a particular column must be of a particular data type, e.g., Boolean, numeric, etc.
- **Range Constraints:** typically, numbers or dates should fall within a certain range, e.g. Age
- **Unique Constraints:** A field, or a combination of fields, must be unique across a dataset, e.g. National ID Number
- **Set-Membership constraints:** The values for a column come from a set of discrete values or codes, e.g. Gender
- **Regular expression patterns:** Occasionally, text fields will have to be validated this way, e.g. Mobile Phone Numbers
- **Cross-field validation:** Certain conditions that utilize multiple fields must hold, e.g. Date of Birth < Date of Death

Dealing with Invalid Data

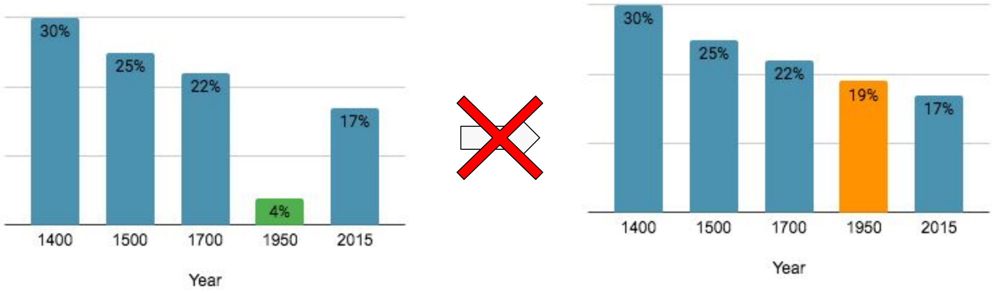
Strategies to preserve data points:

- Find accurate values for invalid data
- Guess/Estimate missing values

Ignoring the invalid data

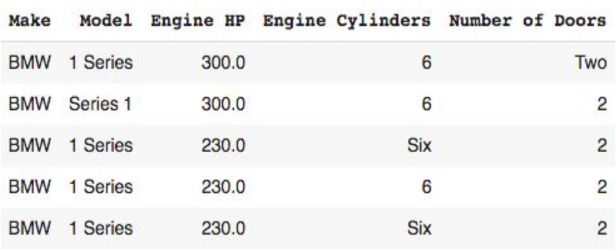
- Drop the rows/columns containing missing data

Dealing with Invalid Data: Caution!



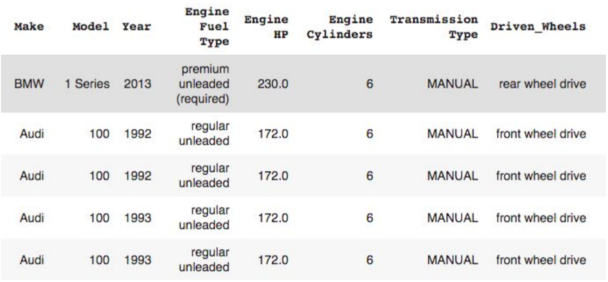
Inconsistent Data/Duplicate Data

Inconsistent Data:
Similar data in different formats



Make	Model	Engine HP	Engine Cylinders	Number of Doors
BMW	1 Series	300.0	6	Two
BMW	Series 1	300.0	6	2
BMW	1 Series	230.0	Six	2
BMW	1 Series	230.0	6	2
BMW	1 Series	230.0	Six	2

Duplicate Data:
Same data, repeated

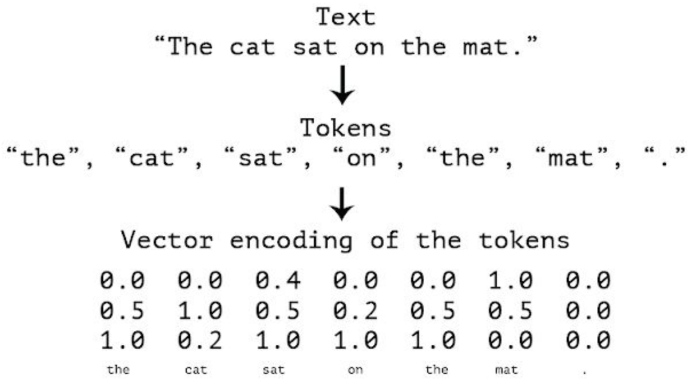


Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels
BMW	1 Series	2013	premium unleaded (required)	230.0	6	MANUAL	rear wheel drive
Audi	100	1992	regular unleaded	172.0	6	MANUAL	front wheel drive
Audi	100	1992	regular unleaded	172.0	6	MANUAL	front wheel drive
Audi	100	1993	regular unleaded	172.0	6	MANUAL	front wheel drive
Audi	100	1993	regular unleaded	172.0	6	MANUAL	front wheel drive

Data Preprocessing

Data Preprocessing: Convert to Numerical Values

- Convert “categorical” variables to numbers
- Transform Boolean values to 0/1
- Take care of “date/time” values
- Embedding text into vectors



Text
“The cat sat on the mat.”

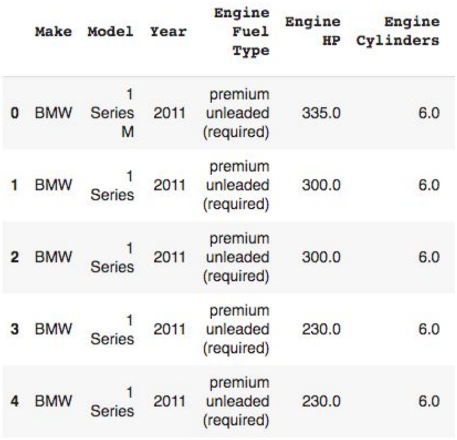
Tokens
“the”, “cat”, “sat”, “on”, “the”, “mat”, “.”

Vector encoding of the tokens

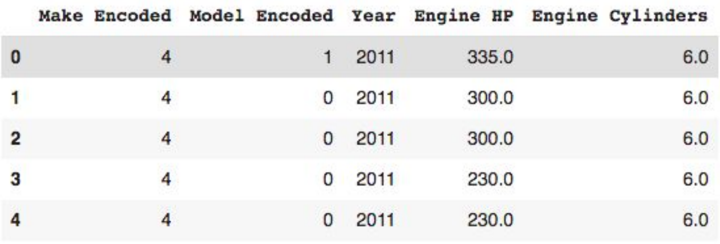
0.0	0.0	0.4	0.0	0.0	1.0	0.0
0.5	1.0	0.5	0.2	0.5	0.5	0.0
1.0	0.2	1.0	1.0	1.0	0.0	0.0
the	cat	sat	on	the	mat	.

Credit: https://plink.ir/7Y3pF

Data Preprocessing: Categorical Encoding

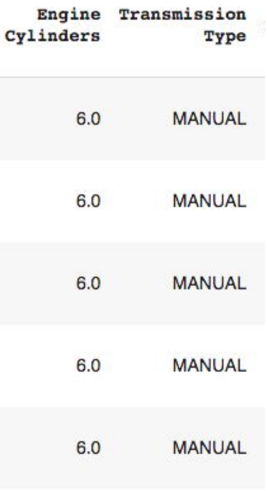


	Make	Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders
0	BMW	1 Series M	2011	premium unleaded (required)	335.0	6.0
1	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0
2	BMW	1 Series	2011	premium unleaded (required)	300.0	6.0
3	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0
4	BMW	1 Series	2011	premium unleaded (required)	230.0	6.0

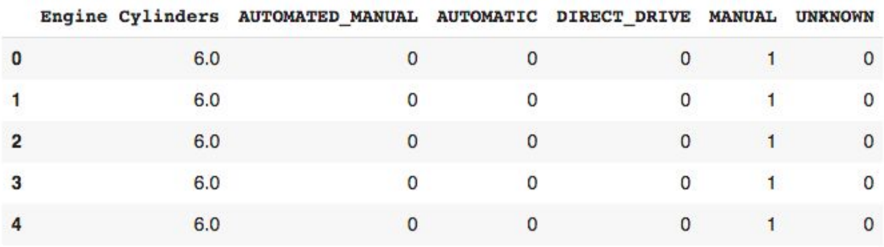


	Make Encoded	Model Encoded	Year	Engine HP	Engine Cylinders
0	4	1	2011	335.0	6.0
1	4	0	2011	300.0	6.0
2	4	0	2011	300.0	6.0
3	4	0	2011	230.0	6.0
4	4	0	2011	230.0	6.0

Data Preprocessing: One-Hot-Encoding



Engine Cylinders	Transmission Type
6.0	MANUAL
6.0	MANUAL
6.0	MANUAL
6.0	MANUAL
6.0	MANUAL

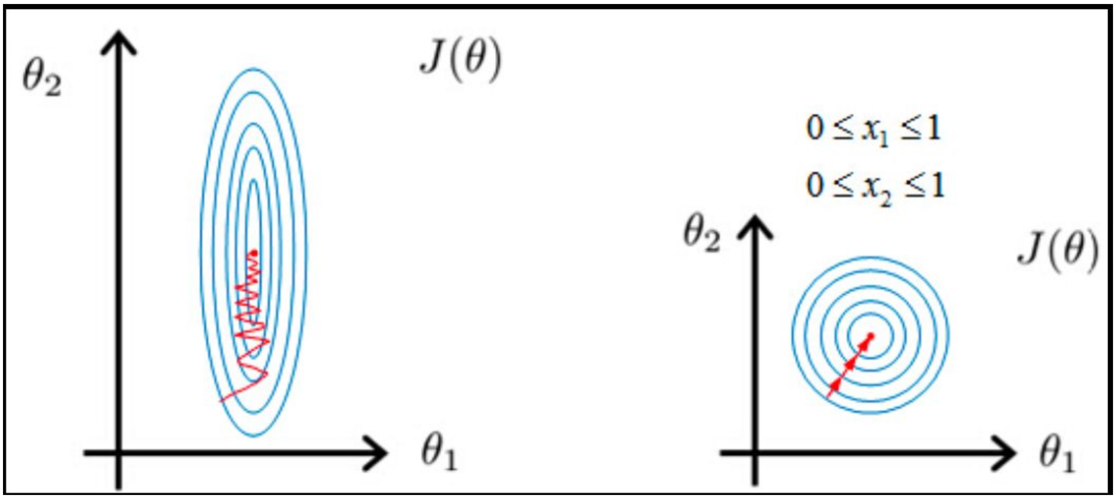


	Engine Cylinders	AUTOMATED_MANUAL	AUTOMATIC	DIRECT_DRIVE	MANUAL	UNKNOWN
0	6.0	0	0	0	1	0
1	6.0	0	0	0	1	0
2	6.0	0	0	0	1	0
3	6.0	0	0	0	1	0
4	6.0	0	0	0	1	0

Data Preprocessing: Normalization

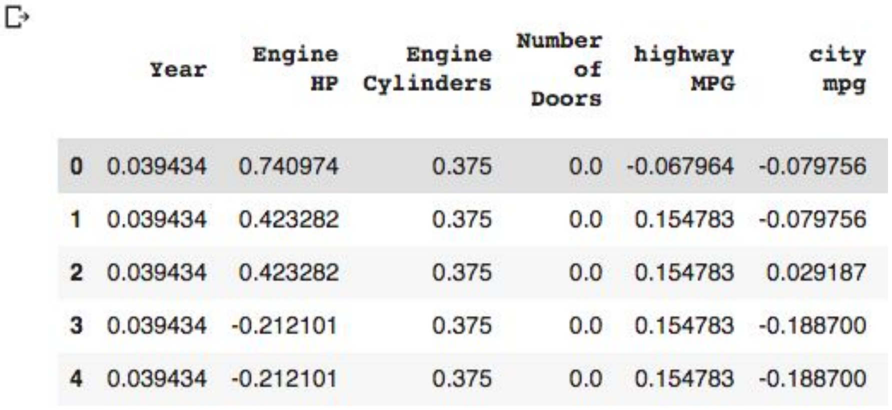
- Make all columns have a similar range
- For images it is equivalent to adjust certain properties like “brightness” across all channels.
- But why is it necessary?

Data Preprocessing: Normalization



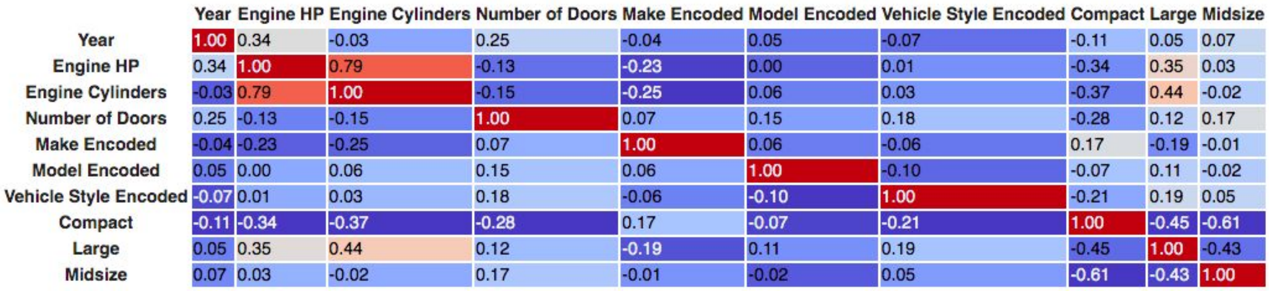
Credit: https://stackoverflow.com/a/46688787/14458418

Data Preprocessing: StanNormalization



	Year	Engine HP	Engine Cylinders	Number of Doors	highway MPG	city mpg
0	0.039434	0.740974	0.375	0.0	-0.067964	-0.079756
1	0.039434	0.423282	0.375	0.0	0.154783	-0.079756
2	0.039434	0.423282	0.375	0.0	0.154783	0.029187
3	0.039434	-0.212101	0.375	0.0	0.154783	-0.188700
4	0.039434	-0.212101	0.375	0.0	0.154783	-0.188700

Data Preprocessing: Correlation Analysis



	Year	Engine HP	Engine Cylinders	Number of Doors	Make Encoded	Model Encoded	Vehicle Style Encoded	Compact	Large	Midsize
Year	1.00	0.34	-0.03	0.25	-0.04	0.05	-0.07	-0.11	0.05	0.07
Engine HP	0.34	1.00	0.79	-0.13	-0.23	0.00	0.01	-0.34	0.35	0.03
Engine Cylinders	-0.03	0.79	1.00	-0.15	-0.25	0.06	0.03	-0.37	0.44	-0.02
Number of Doors	0.25	-0.13	-0.15	1.00	0.07	0.15	0.18	-0.28	0.12	0.17
Make Encoded	-0.04	-0.23	-0.25	0.07	1.00	0.06	-0.06	0.17	-0.19	-0.01
Model Encoded	0.05	0.00	0.06	0.15	0.06	1.00	-0.10	-0.07	0.11	-0.02
Vehicle Style Encoded	-0.07	0.01	0.03	0.18	-0.06	-0.10	1.00	-0.21	0.19	0.05
Compact	-0.11	-0.34	-0.37	-0.28	0.17	-0.07	-0.21	1.00	-0.45	-0.61
Large	0.05	0.35	0.44	0.12	-0.19	0.11	0.19	-0.45	1.00	-0.43
Midsize	0.07	0.03	-0.02	0.17	-0.01	-0.02	0.05	-0.61	-0.43	1.00

Summary

Know your data before diving in (EDA)

Data is almost always dirty, *carefully* clean it before starting the analysis

Finally, make your data machine-understandable

Signs that your data may not be faithful (and proposed solutions)

Truncated data
Early Microsoft Excel limits: 65536 Rows, 255 Columns

Spelling Errors

Apply corrections or drop records not in a dictionary

Time Zone Inconsistencies

Convert to a common timezone (e.g., UTC)

Duplicated Records or Fields

Identify and eliminate (use primary key).

Units not specified or consistent

Infer units, check values are in reasonable ranges for data

- Be aware of consequences in analysis when using data with inconsistencies.
- Understand the potential implications for how data were collected.

Missing Data???

Examples

```
"  “           1970, 1900
0, -1          NaN
999, 12345     Null
```

NaN: “Not a Number”