CDS501: PRINCIPLES & PRACTICES OF DATA SCIENCE & ANALYTICS

Data Exploration and Treatment

Introduction

- Attribute and Target identification
- Missing values detection and treatment
- Outlier detection and treatment
- Attribute transformation
- Data Analysis and Visualization

Attribute and Target Identification

- Identify the Predictors/Features/Attributes (input) and Target (output)
- Attributes are variables that are used to determine (predict) the Target
- Target is the variable that needs to be predicted
- Identify the data type of the attributes and target e.g. numerical, categorical etc.

Predictor and Target Identification

SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
35	0	33.6	0.627	50	1
29	0	26.6	0.351	31	0
0	0	23.3	0.672	32	1
23	94	28.1	0.167	21	0
35	168	43.1	2.288	33	1
0	0	25.6	0.201	30	0
32	88	31.0	0.248	26	1
0	0	35.3	0.134	29	0
45	543	30.5	0.158	53	1
0	0	0.0	0.232	54	1

Attributes (Input)

Target (Output)

Suppose the problem is to predict 'Outcome'

Data Type Identification

SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
35	0	33.6	0.627	50	1
29	0	26.6	0.351	31	0
0	0	23.3	0.672	32	1
23	94	28.1	0.167	21	0
35	168	43.1	2.288	33	1
0	0	25.6	0.201	30	0
32	88	31.0	0.248	26	1
0	0	35.3	0.134	29	0
45	543	30.5	0.158	53	1
0	0	0.0	0.232	54	1
Discrete Numerical	Discrete Numerical	Continuous Numerical	Continuous Numerical	Discrete Numerical	Nominal Categorical

Predictor and Target Identification

	miles_per_gallon	cylinders	weight	model_year	origin
105	12	8	4906	73	1
106	13	8	4654	73	1
107	12	8	4499	73	1
108	18	6	2789	73	1
109	20	4	2279	73	3
110	21	4	2401	73	1
111	22	4	2379	73	3
112	18	3	2124	73	3
113	19	4	2310	73	1
114	21	6	2472	73	1
115	26	4	2265	73	2
116	15	8	4082	73	1
117	16	8	4278	73	1
118	29	4	1867	73	2

Target (Output)

Attributes (Input)

Suppose the problem is to predict 'miles_per_gallon'

Data Type Identification

	miles_per_gallon	cylinders	weight	model_year	origin
105	12	8	4906	73	1
106	13	8	4654	73	1
107	12	8	4499	73	1
108	18	6	2789	73	1
109	20	4	2279	73	3
110	21	4	2401	73	1
111	22	4	2379	73	3
112	18	3	2124	73	3
113	19	4	2310	73	1
114	21	6	2472	73	1
115	26	4	2265	73	2
116	15	8	4082	73	1
117	16	8	4278	73	1
118	29	4	1867	73	2
	Continuous Numerical	Discrete Numerical	Continuous Numerical	Discrete Numerical	Nominal Categorical

Data Treatment

Missing Values

- Missing values may produce ineffective predictive models

kc_house_data_rev

	X	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above
1	0	7129300520	20141013T000000	221900	3	1.00	1180	5650	1.0	0	0	3	7	1180
2	1	6414100192	20141209T000000	538000	3	2.25	2570	7242	2.0	0	0	3	7	2170
3	2	5631500400	20150225T000000	180000	2	1.00	770	10000	1.0	0	0	3	6	770
4	3	2487200875	20141209T000000	604000	4	3.00	1960	5000	1.0	0	0	5	7	1050
5	4	1954400510	20150218T000000	510000	3	2.00	1680	8080	1.0	0	0	3	8	1680
6	5	7237550310	20140512T000000	1225000	4	4.50	5420	101930	1.0	0	0	3	11	NA
7	6	1321400060	20140627T000000	257500	3	2.25	1715	6819	2.0	0	0	3	7	1715
8	7	2008000270	20150115T000000	291850	3	1.50	1060	9711	1.0	0	0	3	7	1060
9	8	2414600126	20150415T000000	229500	3	1.00	1780	7470	1.0	0	0	3	7	1050
10	9	3793500160	20150312T000000	323000	3	2.50	1890	NA	2.0	0	0	3	7	1890
11	10	1736800520	20150403T000000	662500	NA	2.50	3560	9796	1.0	0	0	3	8	1860
12	11	9212900260	20140527T000000	468000	2	1.00	1160	6000	1.0	0	0	4	7	860
13	12	114101516	20140528T000000	310000	3	1.00	1430	19901	1.5	0	0	4	7	1430
14	13	6054650070	20141007T000000	400000	3	1.75	1370	9680	1.0	0	0	4	7	1370
15	14	1175000570	20150312T000000	530000	5	2.00	1810	4850	1.5	0	0	3	7	1810
16	15	9297300055	20150124T000000	650000	4	3.00	2950	5000	2.0	0	3	3	9	1980
17	16	1875500060	20140731T000000	395000	3	2.00	1890	14040	2.0	0	0	3	7	1890
18	17	6865200140	20140529T000000	485000	4	1.00	1600	4300	1.5	0	0	4	7	1600
19	18	16000397	20141205T000000	189000	2	1.00	1200	9850	1.0	0	0	4	7	1200
20	19	7983200060	20150424T000000	230000	3	1.00	1250	9774	1.0	0	0	4	7	1250

Missing Values

- Missing Completely at Random (MCAR)
- Not common
- Missing values on a given attribute is not associated with other attributes (no particular reason for the missing values)
- Missing values are randomly distributed across all observations
- A respondent decide to declare his/her income after tossing a coin
- Clinical samples might be damaged or contaminated in the lab

Missing Values

- Missing at Random (MAR)
- More common
- Missing values can be associated with other attributes
- Missing values are not randomly distributed across examples but are distributed within one or more sub-examples
- Men more likely to tell you their weight than women

Missing Values

- Missing Not at Random
- Missing value is related to the reason it's missing
- Weighing scale mechanism wear out over time
- Sensor node failure due to battery runs out of power

Handling Missing Values

- Listwise deletion - delete the rows (examples)
- Listwise deletion can be used if the number of samples is large
- Could introduce bias into the dataset

	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	st
582	829195	M	FALSE	3030	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	55.0000	O
583	~~829722~~	~~F~~	~~NA~~	~~NA~~	~~Widowed~~	~~TRUE~~	~~Homeowner free and clear~~	~~FALSE~~	~~0~~	~~78.0000~~	Te
584	830245	M	TRUE	120300	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	48.0000	O
585	830758	F	TRUE	4390	Never Married	FALSE	Rented	TRUE	1	24.0000	M
586	~~831497~~	~~M~~	~~TRUE~~	~~95000~~	~~Married~~	~~TRUE~~	~~Homeowner with mortgage/loan~~	~~FALSE~~	~~3~~	~~NA~~	Il
587	832435	M	TRUE	55001	Married	TRUE	Homeowner free and clear	FALSE	3	48.0000	N
588	~~832866~~	~~M~~	~~NA~~	~~6000~~	~~Divorced/Separated~~	~~FALSE~~	~~Homeowner with mortgage/loan~~	~~FALSE~~	~~2~~	~~NA~~	W
589	832949	M	TRUE	54000	Divorced/Separated	TRUE	Rented	FALSE	1	42.0000	K
590	833321	M	FALSE	30900	Never Married	FALSE	Homeowner with mortgage/loan	FALSE	2	51.0000	M
591	835830	F	TRUE	51000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	41.0000	W
592	~~837381~~	~~M~~	~~TRUE~~	~~NA~~	~~Never Married~~	~~TRUE~~	~~NA~~	~~NA~~	~~NA~~	~~20.0000~~	Vi
593	~~841342~~	~~M~~	~~TRUE~~	~~NA~~	~~Married~~	~~TRUE~~	~~Homeowner with mortgage/loan~~	~~FALSE~~	~~3~~	~~62.0000~~	Te
594	843811	F	TRUE	40000	Never Married	TRUE	Homeowner with mortgage/loan	FALSE	1	146.6802	In
595	~~844007~~	~~M~~	~~NA~~	~~0~~	~~Married~~	~~FALSE~~	~~NA~~	~~NA~~	~~NA~~	~~46.0000~~	C
596	845421	F	FALSE	4300	Married	TRUE	Rented	FALSE	0	27.0000	O
597	~~846713~~	~~M~~	~~NA~~	~~250~~	~~Never Married~~	~~TRUE~~	~~NA~~	~~NA~~	~~NA~~	~~45.0000~~	Fl
598	847156	M	TRUE	72000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	52.0000	C

Handling Missing Values

- Pairwise deletion - delete the cell (examples)
- Pairwise deletion needs to be used if the dataset is small
- Analysis will be based on sets with different number of samples (rows)

	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	st
582	829195	M	FALSE	3030	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	55.0000	O
583	829722	F	~~NA~~	~~NA~~	Widowed	TRUE	Homeowner free and clear	FALSE	0	78.0000	Te
584	830245	M	TRUE	120300	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	48.0000	O
585	830758	F	TRUE	4390	Never Married	FALSE	Rented	TRUE	1	24.0000	M
586	831497	M	TRUE	95000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	~~NA~~	Ill
587	832435	M	TRUE	55001	Married	TRUE	Homeowner free and clear	FALSE	3	48.0000	N
588	832866	M	~~NA~~	6000	Divorced/Separated	FALSE	Homeowner with mortgage/loan	FALSE	2	~~NA~~	W
589	832949	M	TRUE	54000	Divorced/Separated	TRUE	Rented	FALSE	1	42.0000	K
590	833321	M	FALSE	30900	Never Married	FALSE	Homeowner with mortgage/loan	FALSE	2	51.0000	M
591	835830	F	TRUE	51000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	41.0000	W
592	837381	M	TRUE	~~NA~~	Never Married	TRUE	~~NA~~	~~NA~~	~~NA~~	20.0000	Vi
593	841342	M	TRUE	~~NA~~	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	62.0000	Te
594	843811	F	TRUE	40000	Never Married	TRUE	Homeowner with mortgage/loan	FALSE	1	146.6802	In
595	844007	M	~~NA~~	0	Married	FALSE	~~NA~~	~~NA~~	~~NA~~	46.0000	C
596	845421	F	FALSE	4300	Married	TRUE	Rented	FALSE	0	27.0000	O
597	846713	M	~~NA~~	250	Never Married	TRUE	~~NA~~	~~NA~~	~~NA~~	45.0000	Fl
598	847156	M	TRUE	72000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	52.0000	C

Handling Missing Values

- Dropping Attributes
- Dropping attributes can be used if the attributes are not relevant or if the percentage is too high (subjective)

	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	st
582	829195	M	FALSE	3030	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	55.0000	O
583	829722	F	NA	NA	Widowed	TRUE	Homeowner free and clear	FALSE	0	78.0000	Te
584	830245	M	TRUE	120300	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	48.0000	O
585	830758	F	TRUE	4390	Never Married	FALSE	Rented	TRUE	1	24.0000	M
586	831497	M	TRUE	95000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	NA	Ill
587	832435	M	TRUE	55001	Married	TRUE	Homeowner free and clear	FALSE	3	48.0000	N
588	832866	M	NA	6000	Divorced/Separated	FALSE	Homeowner with mortgage/loan	FALSE	2	NA	W
589	832949	M	TRUE	54000	Divorced/Separated	TRUE	Rented	FALSE	1	42.0000	Ke
590	833321	M	FALSE	30900	Never Married	FALSE	Homeowner with mortgage/loan	FALSE	2	51.0000	M
591	835830	F	TRUE	51000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	41.0000	W
592	837381	M	TRUE	NA	Never Married	TRUE	NA	NA	NA	20.0000	Vi
593	841342	M	TRUE	NA	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	62.0000	Te
594	843811	F	TRUE	40000	Never Married	TRUE	Homeowner with mortgage/loan	FALSE	1	146.6802	In
595	844007	M	NA	0	Married	FALSE	NA	NA	NA	46.0000	C
596	845421	F	FALSE	4300	Married	TRUE	Rented	FALSE	0	27.0000	O
597	846713	M	NA	250	Never Married	TRUE	NA	NA	NA	45.0000	Fl
598	847156	M	TRUE	72000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	52.0000	C

Handling Missing Values

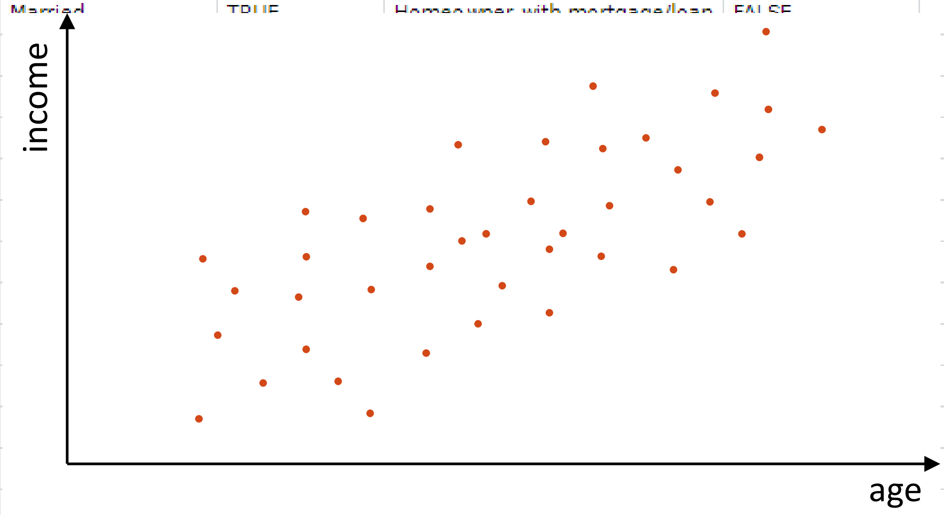
- Fill missing values with estimated values (mean/median imputation)
- Imputation needs to be used if the dataset is small
- Could reduce the variability of the data if there are many missing values

	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	st
582	829195	M	FALSE	3030	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	55.0000	O
583	829722	F	NA	mean/median NA	Widowed	TRUE	Homeowner free and clear	FALSE	0	78.0000	Te
584	830245	M	TRUE	120300	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	48.0000	O
585	830758	F	TRUE	4390	Never Married	FALSE	Rented	TRUE	1	24.0000	M
586	831497	M	TRUE	95000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	mean/median NA	I
587	832435	M	TRUE	55001	Married	TRUE	Homeowner free and clear	FALSE	3	48.0000	N
588	832866	M	NA	6000	Divorced/Separated	FALSE	Homeowner with mortgage/loan	FALSE	2	mean/median NA	W
589	832949	M	TRUE	54000	Divorced/Separated	TRUE	Rented	FALSE	1	42.0000	Ke
590	833321	M	FALSE	30900	Never Married	FALSE	Homeowner with mortgage/loan	FALSE	2	51.0000	M
591	835830	F	TRUE	51000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	41.0000	W
592	837381	M	TRUE	NA	Never Married	TRUE	Similar case e.g. Homeowner NA	NA	NA	20.0000	Vi
593	841342	M	TRUE	mean/median NA	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	62.0000	Te
594	843811	F	TRUE	40000	Never Married	TRUE	Homeowner with mortgage/loan	FALSE	1	146.6802	In
595	844007	M	Similar case e.g. False NA	0	Married	FALSE	NA	NA	NA	46.0000	C
596	845421	F	FALSE	4300	Married	TRUE	Rented	FALSE	0	27.0000	O
597	846713	M	NA	250	Never Married	TRUE	NA	NA	NA	45.0000	Fl
598	847156	M	TRUE	72000	Married	TRUE	Homeowner with mortgage/loan	FALSE	2	52.0000	C

Handling Missing Values

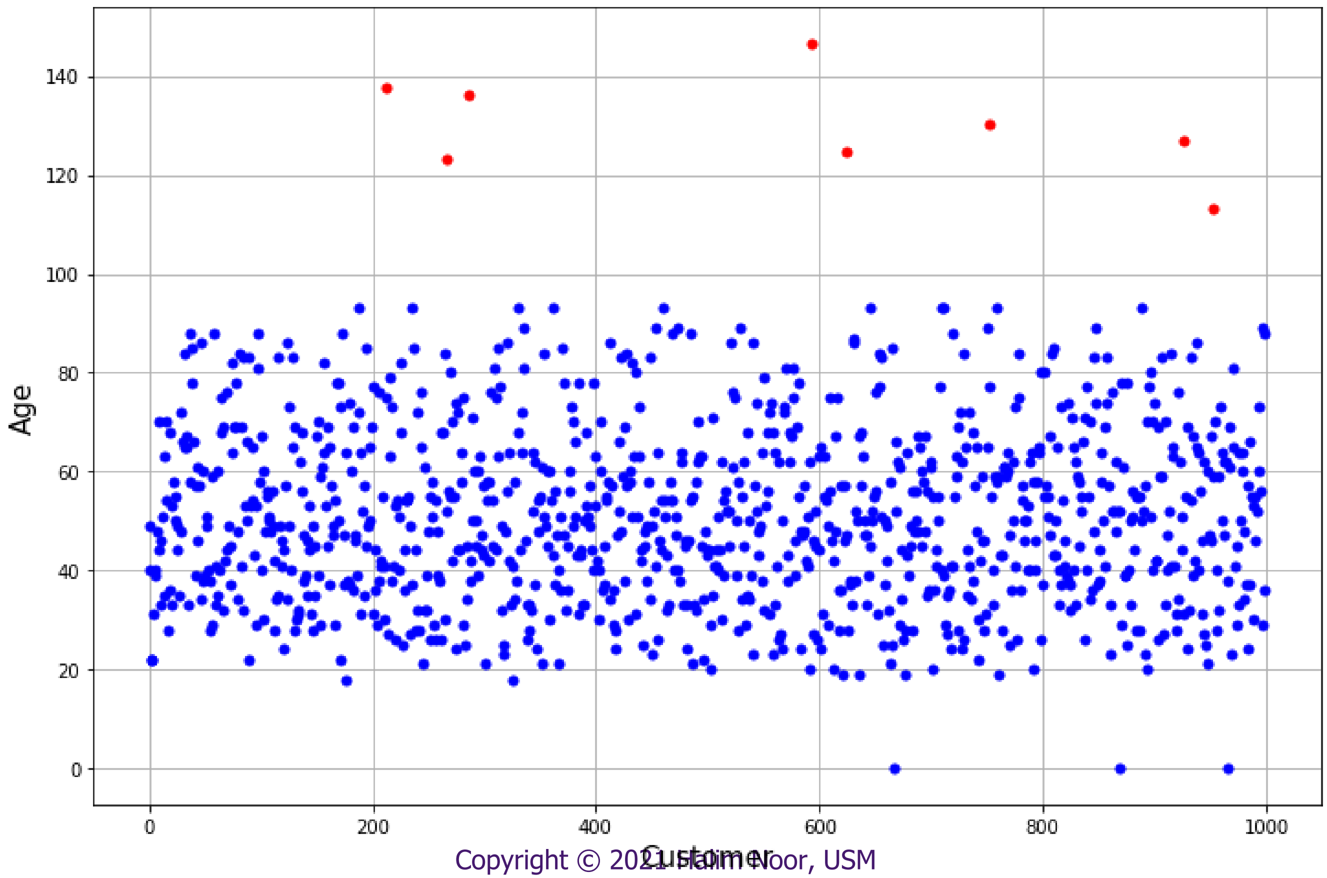
- Fill missing values with estimated values (regression imputation)

	custid	sex	is.employed	income	marital.stat	health.ins	housing.type	recent.move	num.vehicles	age	st
582	829195	M	FALSE	3030	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	55.0000	O
583	829722	F	NA	NA	Widowed	TRUE	Homeowner free and clear	FALSE	0	78.0000	Te
584	830245	M	TRUE	120300	Married	TRUE	Homeowner with mortgage/loan	FALSE	1	48.0000	O
585	830758	F	TRUE	4390	Never Married	FALSE	Rented	TRUE	1	24.0000	M
586	831497	M	TRUE	95000	Married	TRUE	Homeowner with mortgage/loan	FALSE	3	NA	Il
587	832435	M	TRUE	55001					3	48.0000	N
588	832866	M	NA	6000					2	NA	W
589	832949	M	TRUE	54000					1	42.0000	K
590	833321	M	FALSE	30900					2	51.0000	M
591	835830	F	TRUE	51000					3	41.0000	W
592	837381	M	TRUE	NA					NA	20.0000	Vi
593	841342	M	TRUE	NA					3	62.0000	Te
594	843811	F	TRUE	40000					1	146.6802	In
595	844007	M	NA	0					NA	46.0000	C
596	845421	F	FALSE	4300					0	27.0000	O
597	846713	M	NA	250					NA	45.0000	Fl
598	847156	M	TRUE	72000					2	52.0000	C



Outliers

- Observations that are appear far away and diverges from the overall pattern

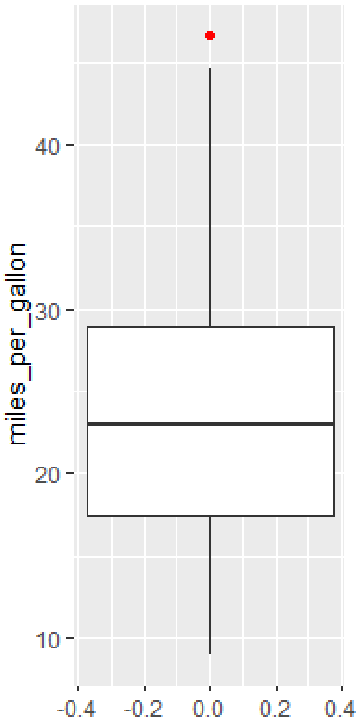


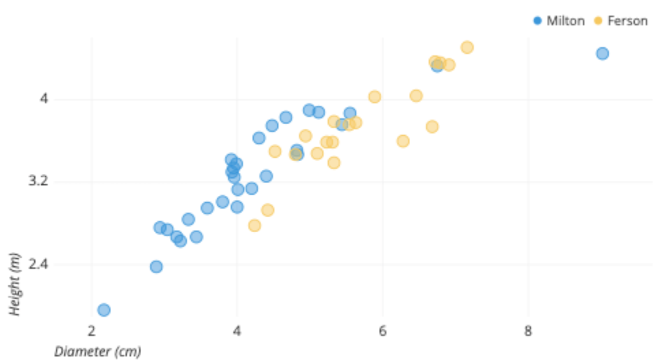
Outliers

- Artificial outliers (errors)
  - Human error – data collection, data entry
  - Measurement error – most common; faulty device, instrument
  - Sampling error – pick the wrong samples e.g. measuring the height of football players but include basketball players
  - Data processing error – errors occur during data manipulation and extraction e.g. wrong computation
  - Intentional outlier – involving sensitive data e.g. income, amount of assets
- Natural outliers
  - Not due to error
  - Valid data

Detecting Outliers

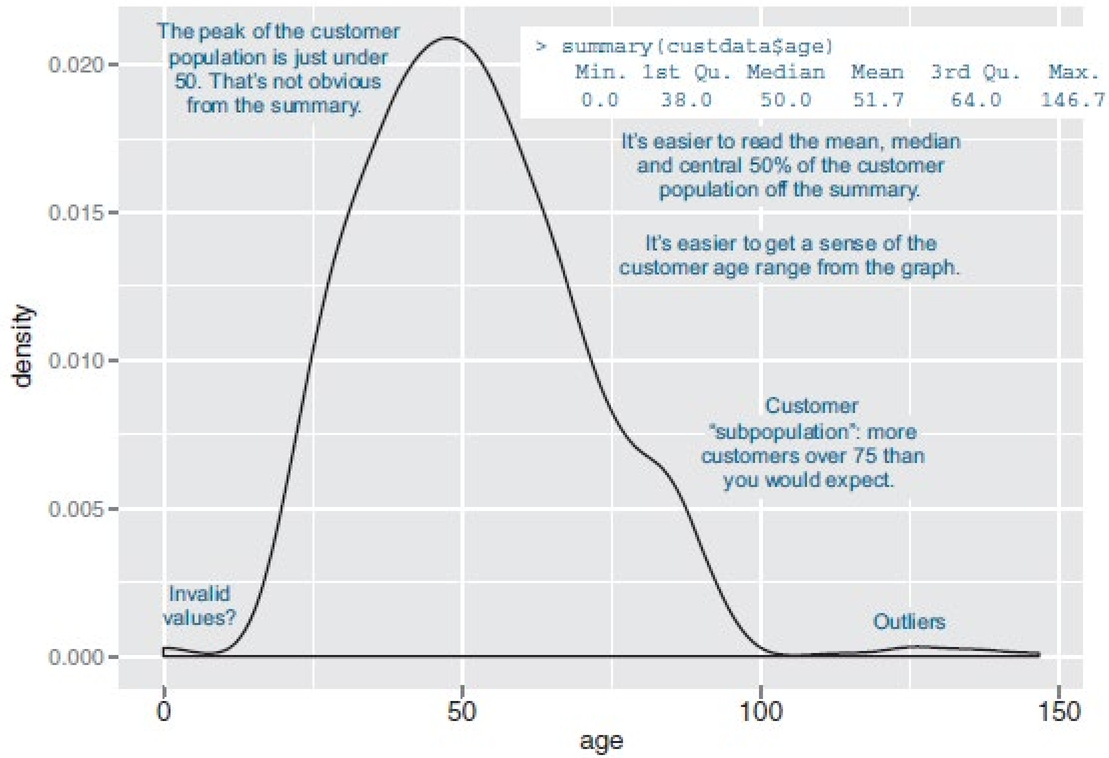
- Visualization – boxplot, scatter, density plot etc.





Detecting Outliers

- Visualization – histogram, scatter, density plot etc.

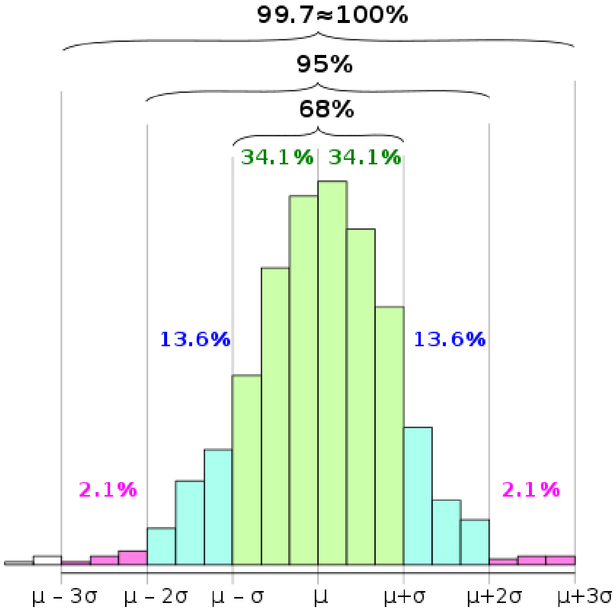


Detecting Outliers

- Outliers are values that fall out of the expected range
- Interquartile (IQR) Rule
  - Calculate the IQR for the data, IQR = Q3 – Q1
  - S = 1.5 × IQR
  - UB = Q3 + S
  - LB = Q1 – S
  - Valid range LB ≤ x ≤ UB

Detecting Outliers

- Empirical rule (68-95-99.7 rule) states that 99.7% of data observed following a normal distribution lies within 3 standard deviations of the mean
- Any observation more than 3 σ can be considered rare



Standard Score (Z Score)

- Measure of how far a data point is from the mean

$$z = \frac{(x - \mu)}{\sigma}$$

- Assume: range of valid data is -2σ < x < 2σ, $\mu = 150$ and $\sigma = 25$
- $x = 180, z = 1.2$ (within the range – x not outlier)
- $x = 210, z = 2.4$ (beyond the range – x outlier)

Handling Outliers

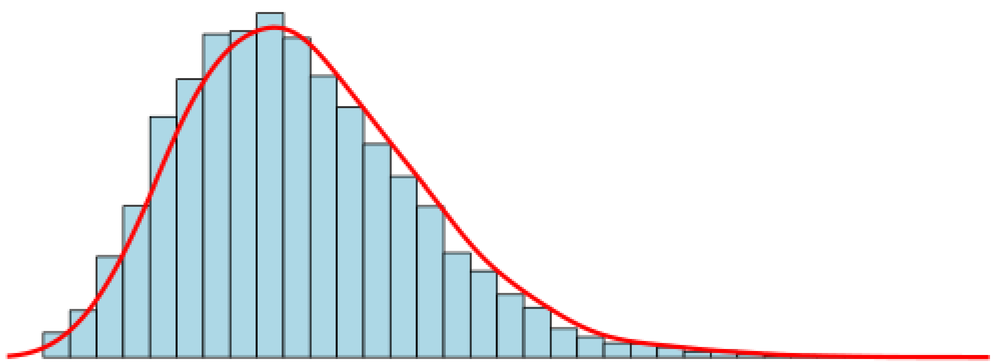
- Binning values – categorize the values into groups (numerical to categorical data)
- Trimming – remove the outliers
- Winsorization – replace with the upper or lower bound values
- Imputation with mean, median or mode

Attribute Transformation

- Normalization
  - Change the values of numeric columns to a common scale
  - Income: range from 0 – 50000
  - Age: range from 0 – 100
  - Scale: 0 to 1
  - $\acute{x}_i = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$
- Standardization
  - Change the data to have a mean of zero and standard deviation of 1
  - $\acute{x}_i = \frac{x_i - \mu}{\sigma}$
- Binning
  - Converting numerical data to categorical data

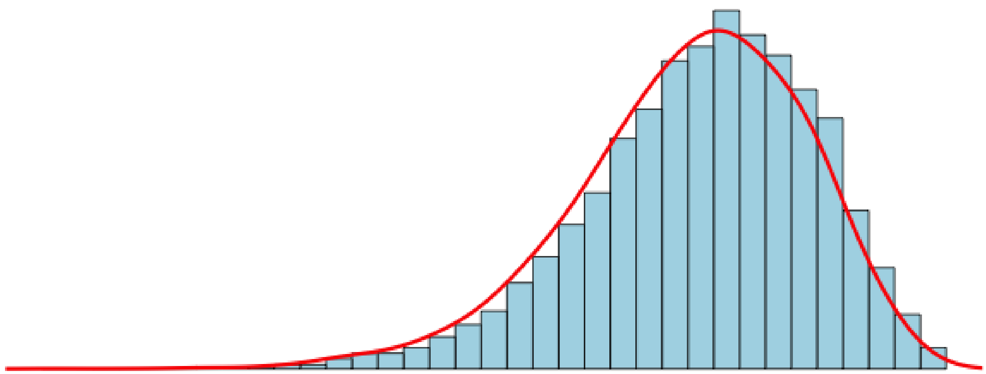
Attribute Transformation

- Reducing right skewness
- Logarithm
  - Cannot be applied to negative values
  - $\acute{x}_i = \log x_i$
- Square Root
  - Cannot be applied to negative values
  - $\acute{x}_i = \sqrt{x_i}$



Attribute Transformation

- Reducing left skewness
- Squares
  - $\acute{x}_i = x_i^2$
- Cubes
  - $\acute{x}_i = x_i^3$



Attribute Transformation

- Deriving new attribute(s) from existing attributes
- number_of_rooms_per_household = total_rooms/households
- population_per_household = population/households
- number_of_bedrooms_per_room = total_bedrooms/total_rooms

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY
5	-122.25	37.85	52.0	919.0	213.0	413.0	193.0	4.0368	269700.0	NEAR BAY
6	-122.25	37.84	52.0	2535.0	489.0	1094.0	514.0	3.6591	299200.0	NEAR BAY
7	-122.25	37.84	52.0	3104.0	687.0	1157.0	647.0	3.1200	241400.0	NEAR BAY
8	-122.26	37.84	42.0	2555.0	665.0	1206.0	595.0	2.0804	226700.0	NEAR BAY
9	-122.25	37.84	52.0	3549.0	707.0	1551.0	714.0	3.6912	261100.0	NEAR BAY

Data Analysis & Visualization

Univariate Analysis – Numerical

- Numerical data
- Central tendency – mean, median, mode

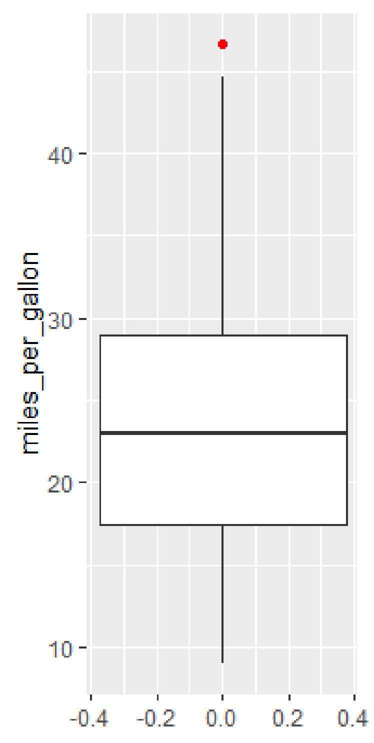
Univariate Analysis – Numerical

- Numerical data
- Central tendency – mean, median, mode
- Measure of dispersion – variance, standard deviation, range, quartile, interquartile range (IQR), skewness

* Cover more in next topics

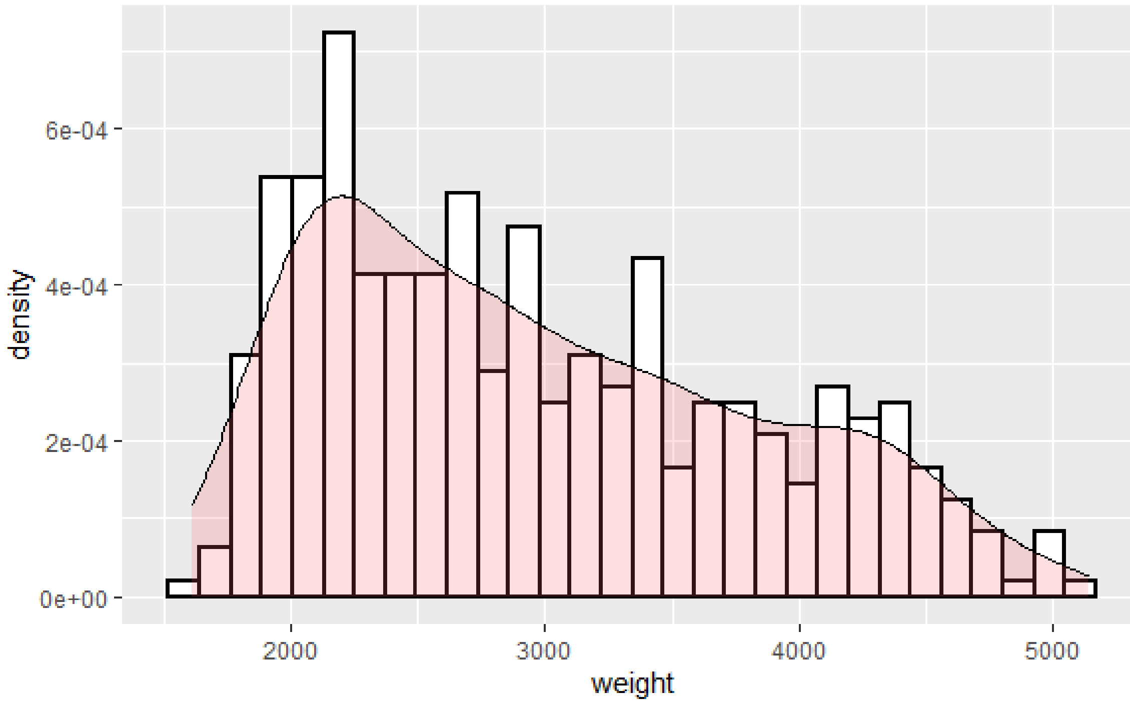
Univariate Analysis – Numerical

- Use visualization to analyze the data
  - Boxplot



Univariate Analysis – Numerical

- Use visualization to analyze the data
  - Boxplot
  - Histogram
  - Density



Univariate Analysis – Categorical

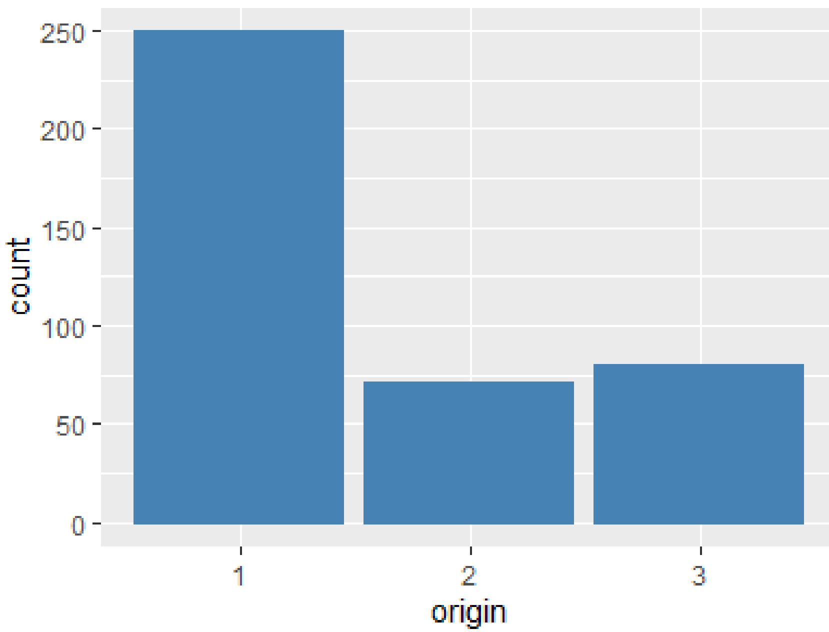
- Frequency table to count the number of times the value (observation) occurs

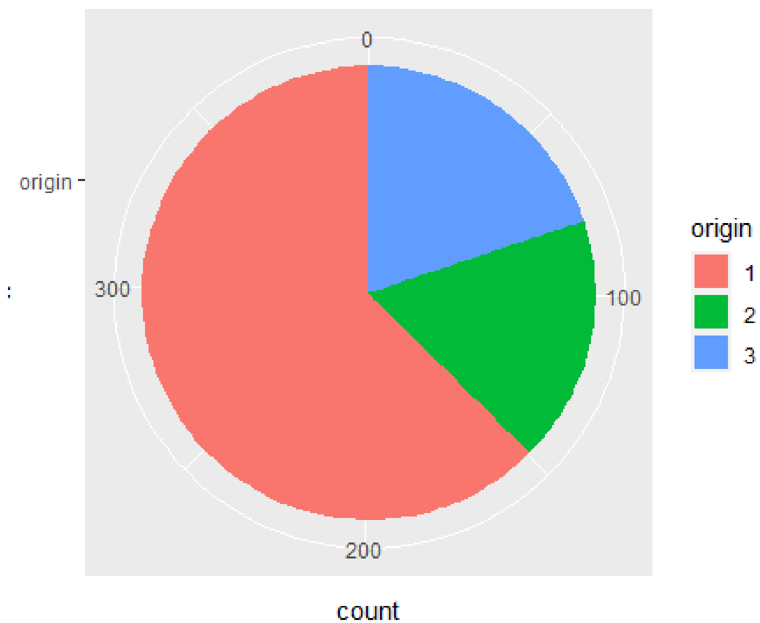
	miles_per_gallon	cylinders	weight	model_year	origin
105					1
106					1
107					1
108					1
109					3
110					1
111					3
112					3
113					1
114					1
115					2
116					1
117					1
118					2

Origin	Frequency
1	249
2	70
3	79

Univariate Analysis – Categorical

- Bar chart or pie chart to visualize data



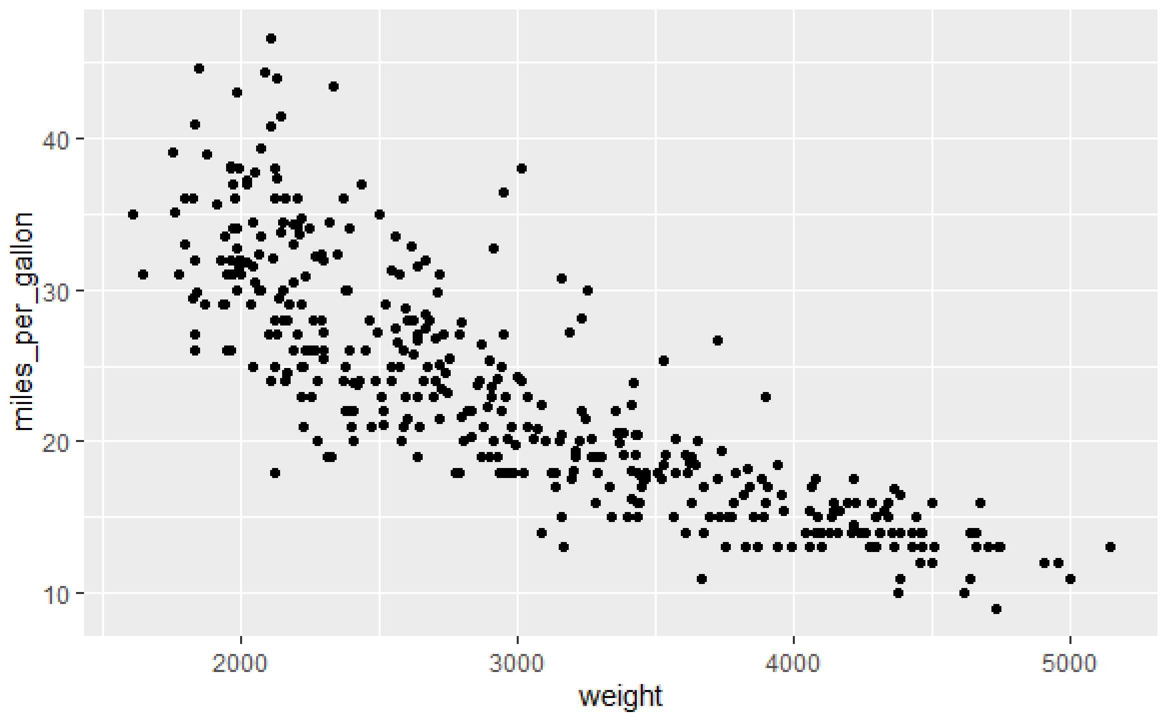


Bivariate Analysis

- Determine the relationship between two attributes
- Analysis between any combination of categorical and numerical
  - Numerical and numerical
  - Numerical and categorical
  - Categorical and numerical
  - Categorical and categorical

Bivariate Analysis

- Use visualization such as scatter plot



Bivariate Analysis

- To determine the strength of relationship
  - Covariance
  - Correlation
  - Statistical inference/test e.g. chi-square test, t test, ANOVA test

* Cover more in next topics

End