DATA EXPLORATION : terms

→ part of data understanding & data preparing of CRISP-DM.

→ aims: ① understand characteristics of each feature in ABT

.... a features: - value types
- value ranges
- value distribution

② identify data quality issues
- missing values
- invalid values + outliers
- noise
- diff units of measurement

→ applications ① feature engineering
using domain knowledge to extract features from raw data to improve the results obtained from Modelling phase

② data quality plan
develop appropriate strategies to handle data quality issues

③ data transformation
make features more compatible with specific analytical models

→ can be broadly classified into

- descriptive stats : condensing key characs. of a data set into simple numeric metrics
e.g. mean, std. deviation, correlation

- data visualisation: projecting data into multi-dimensional space / abstract images
e.g. histograms, box plots, scatter plots.

→ Univariate exploration:
- one feature at a time
- distribution of feature values + distr. shape
- e.g. histograms, box plots, scatter plots

Multivariate exploration:
- 1+ feature simultaneously
- relationship btw. features. Useful for predicting target?
- scatterplots, heatmaps, collection of bar plots.

DATA EXPLORATION TECHNIQUES

CONTINUOUS	NUMERIC	true numeric values; arith. operations (price, age)
	INTERVAL	ordering + subtraction, but not arith. (date, time).
CATEGORIAL	BINARY	set 2 values (fraud vs not-fraud)
	ORDINAL	ordering, but not arith. (size S, M, L)
	CATEGORICAL	finite, x order, x arith (country, product code)
	TEXTUAL	free-form, usually short, text data (name, address)

MAKING A PREDICTIVE MODEL

→ GOAL: explain variation in target feature

→ given the target feature,

1st step: UNDERSTAND DISTRIBUTION OF TARGET FEAUTURE

provides: ① estimate of lower bound of expected performance

② whether target feature should be transformed

③ clues for creating features that may help predict target feature

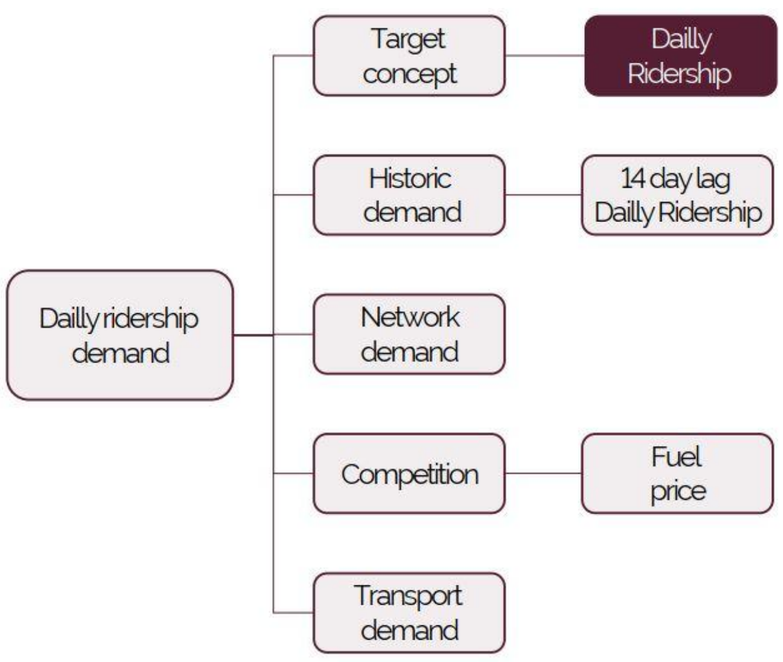
→ can use to understand: univariate visualisations (for contin. feature)

— box plots, violin plots, histograms

DOMAIN CONCEPTS

→ = topic within a field of knowledge

→ used to identify features



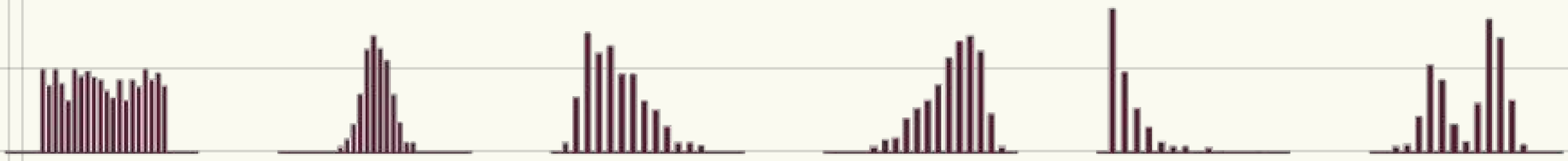
Example domain concepts for the ridership prediction problem

BOXPLOTS

→ quick way to assess distro of a feature

→ min, LQ, median, UQ, max

HISTOGRAM

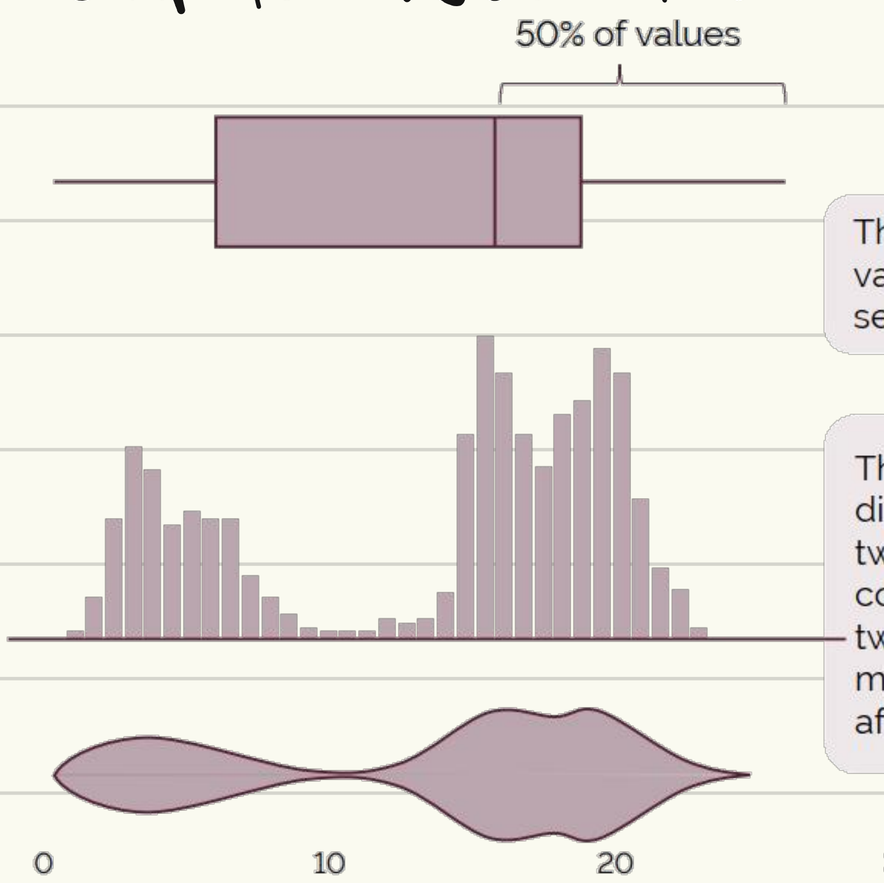


Uniform	Normal (unimodal)	Unimodal (skewed right)	A Unimodal (skewed left)	Exponential	Multimodal
A feature is ***equally likely*** to take a value in any of the ranges present e.g. ID	A strong tendency toward a ***central value*** and ***symmetrical variation*** to either side of this central tendency e.g. heights of randomly selected females	A tendency toward **very high** values e.g. salaries since only a small number of people are paid very large salaries	A tendency towards **very low** values e.g. life expectancy	The values of ***low values*** occurring are ***very high*** but ***diminish rapidly*** for higher values e.g. number of times a person has been married	***Two*** or ***more commonly occurring ranges*** of values that are separated e.g. height of randomly selected students

VIOLIN PLOTS

→ created by generating a density/distro of the data + its mirror image.

→ compact visualisation of the distro of data, + histo characteristics.



used to identify if distro has multiple peaks

ADDITIONAL DIMENSIONS:

→ can be added using colours, shapes, facets ↑ create same type of plot + split into diff. panels based on same feature

→ identify patterns + engineer new features

e.g. peaks ... ridership differs for week + weekend

∴ part of week must be added to ABT (0 or 1)

HEATMAPS

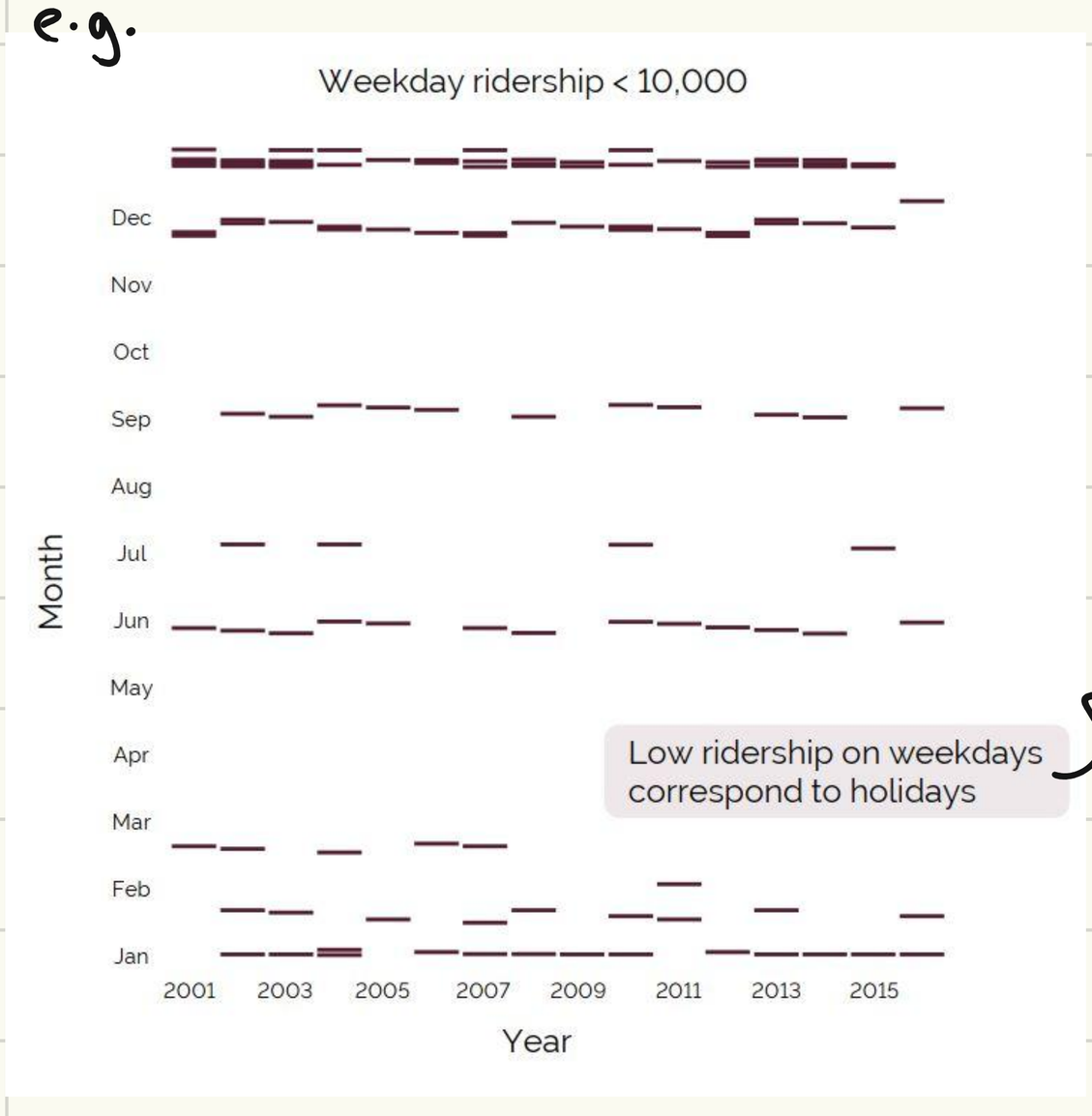
→ one feature on x-axis, one feature on y-axis

→ form a grid

→ filled with colour based on values of a third feature

↳ filling indicator can be categorical/continuous value.

e.g.



∴ add 'holiday' to ABT

Time association

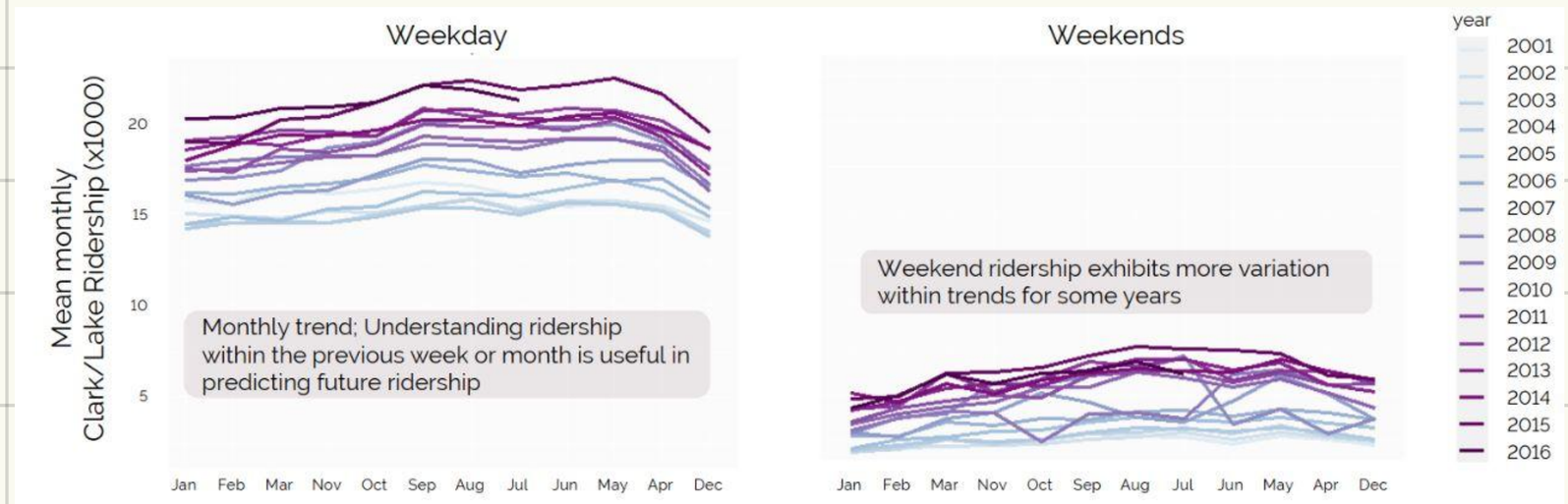
→ feature collected over time

likely to have trends/patterns

that are associated incrementally over time

∴ features current value more related to recent values

→ line plot : relationship btw time and value of a feature



Analytic Base Table

date	Clark/Lake rides 14-day lag	Weekday	Holiday	Clark/Lake rides
22 Jan 2001	7543	1	0	7601
23 Jan 2001	6967	1	0	7690
24 Jan 2001	7603	1	0	8050
25 Jan 2001	7820	1	0	7952
26 Jan 2001	6779	1	0	6996
27 Jan 2001	2738	0	0	4598
28 Jan 2001	2236	0	0	3258
...	...	...	...	...
12 Sep 2016	7615	1	0	7847

∴ to account for time-association, add 14-day lag feature to ABT

SCATTER PLOT

→ one feature on x-axis, one feature on y-axis] each pt

→ relationship btw features ... is feature useful?

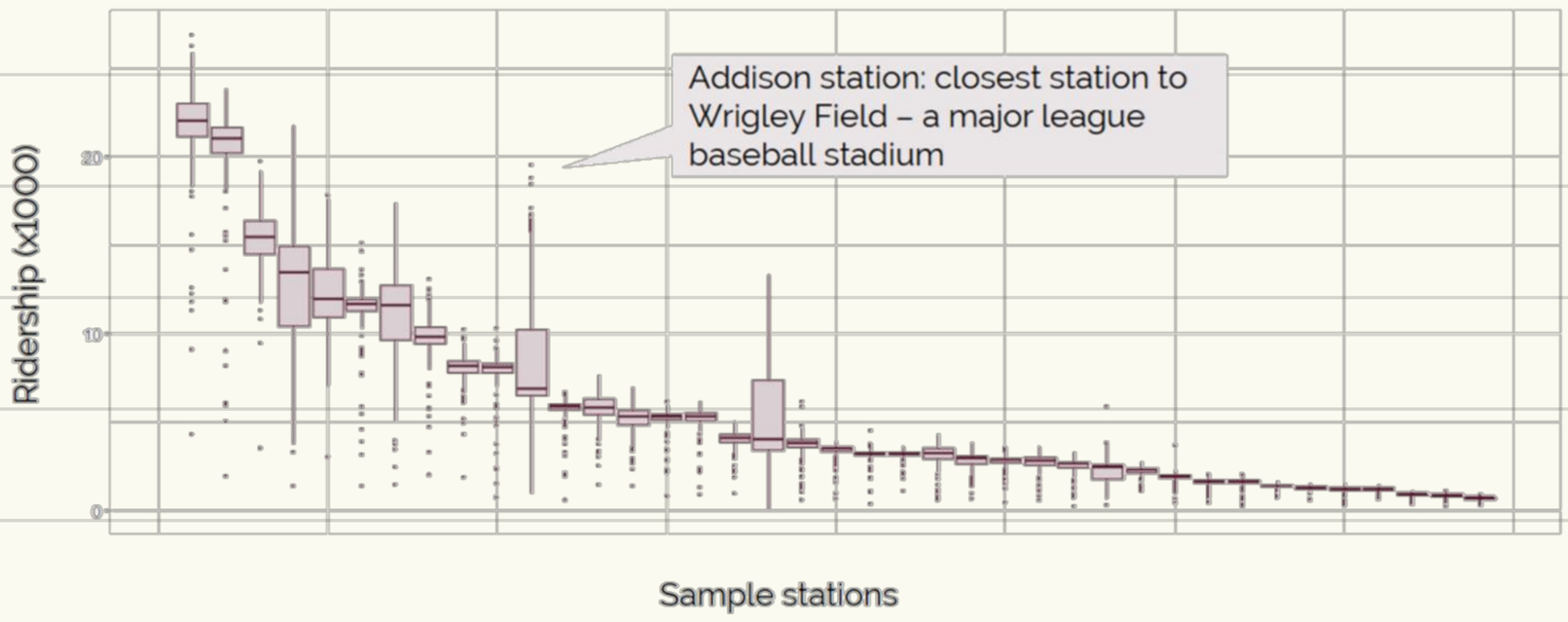
→ for prediction → past should be related to future case

SIDE-BY-SIDE

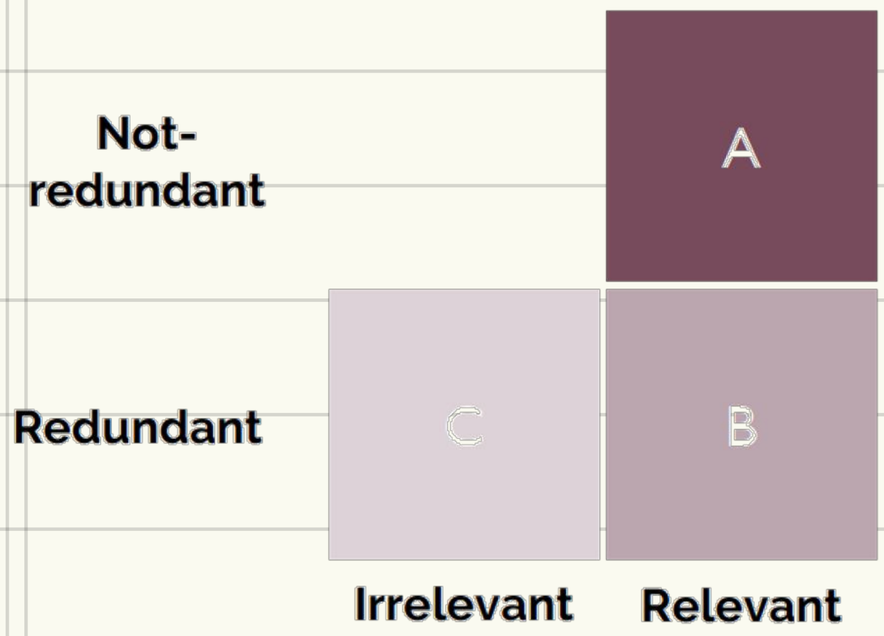
Side-by-side box or ***violin plots*** can be used to ***visualise multiple descriptive features***

When we have a ***moderate number*** of ***features*** and when the ***features*** are in the ***same order*** of ***magnitude***, we can ***visualise*** the ***distributions*** of the ***features side-by-side*** using a box or violin plot

- The variability and range of ridership across stations can be explored using a side-by-side boxplot
- A few stations have distinctly large variations and unusual low and high values i.e. possible outliers



FEATURE RELEVANCE (selection)



A feature can be classified as either:

A ***An informative feature***: Features that are correlated with the output target, but is not correlated with other features

B ***A redundant feature***: Features that are correlated with the output target, but is correlated with other features

C ***Or an irrelevant feature***: Features with no correlation to the output target

SCATTER PLOT MATRIX (SPLOM)

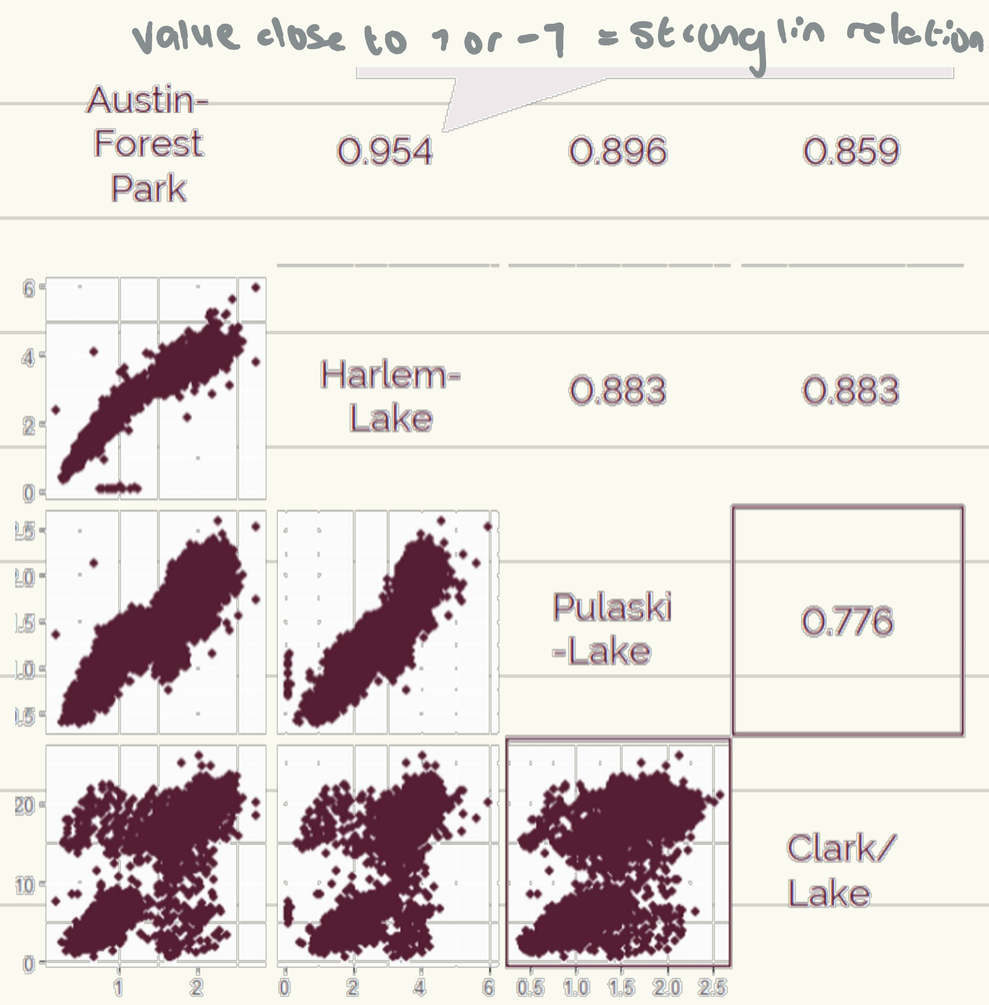
→ scatter plots arranged in a matrix

→ relationships btw groups of features (e.g. all cont feats. in the ABT)

→ effectiveness of SPLOM ↓ as no. of features > 8

→ relat. can also be measured using covariance + correlation

→ corr coeffs included above diag.



RELATIONS

COVARIANCE

For 2 features, a and b, in a dataset of n instances

sample covariance

values of ith instance

$$\text{cov}(a,b) = \frac{1}{n-1}\sum_{i=1}^{n}((a_i-\bar{a})\times(b_i-\bar{b}))$$

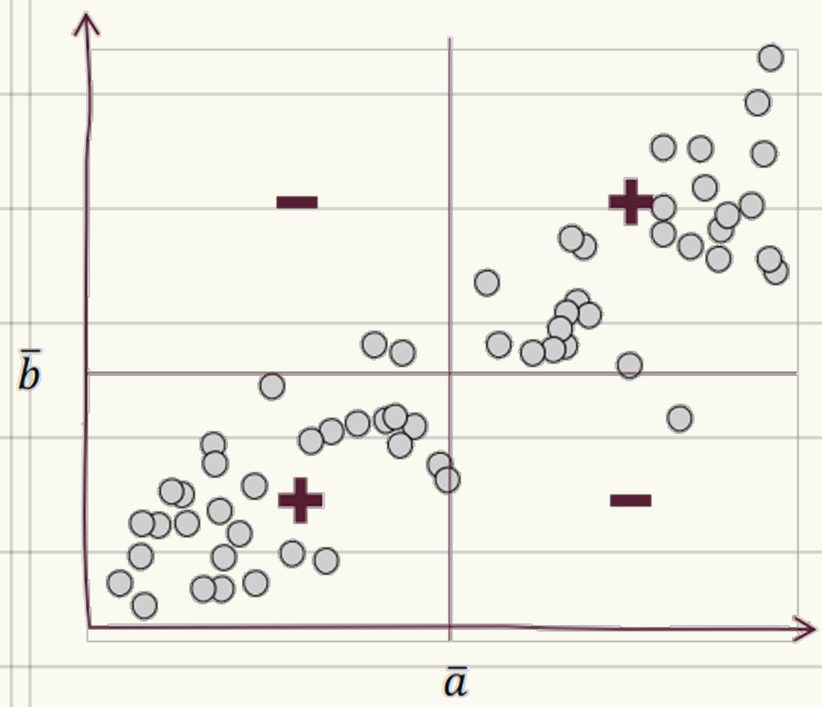
[all same units]

sample means

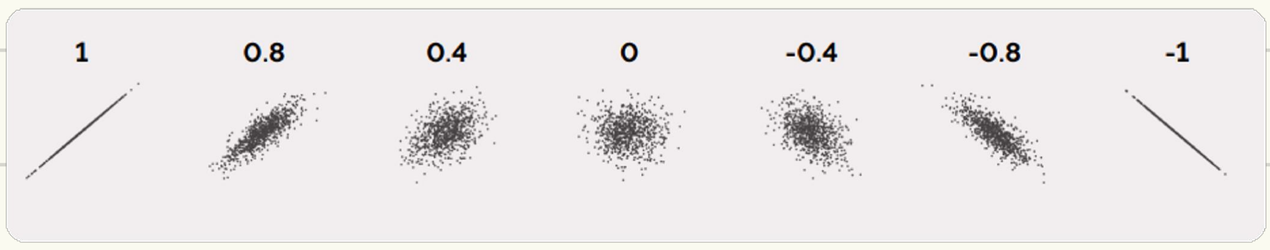
- $\ominus$: linear negative relationship
- $\oplus$: linear positive relationship.
- ≈ 0 : little / no relationship btw features

→ does not capture non-lin rel.

→ sensitive to outliers

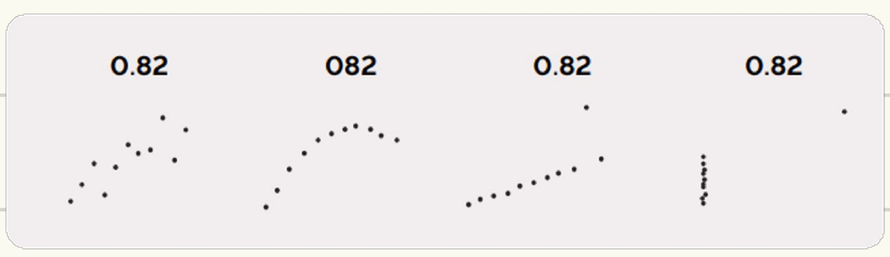


Correlation captures linear relationship between features



Clear relationship between features, but correlation is zero

Different shape; same correlation



CORRELATION → normalised measure of covariance

correlation

covariance

$$\text{corr}(a,b) = \frac{\text{cov}(a,b)}{\sigma_a \times \sigma_B}$$

→ std. deviations:

$$\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{n}}$$

≈ -1 : strong linear negative correlation

$\approx +1$: strong linear positive correlation

≈ 0 : no linear correlation

→ correlation ≠ causation

correlation heatmap:

relationship among several contin. features

$$\underset{\{a,b,\ldots z\}}{\text{correlation matrix}} = \begin{bmatrix} corr(a,a) & corr(a,b) & \ldots & corr(a,z) \\ corr(b,a) & corr(b,b) & \ldots & corr(b,z) \\ \ldots & \ldots & \ddots & \ldots \\ corr(z,a) & corr(z,b) & \ldots & corr(z,z) \end{bmatrix}$$

often displayed in a sPLOM

CORRELATION MATRIX = HEATMAP

↳ for a large no. of features (correlation is 3rd feature)

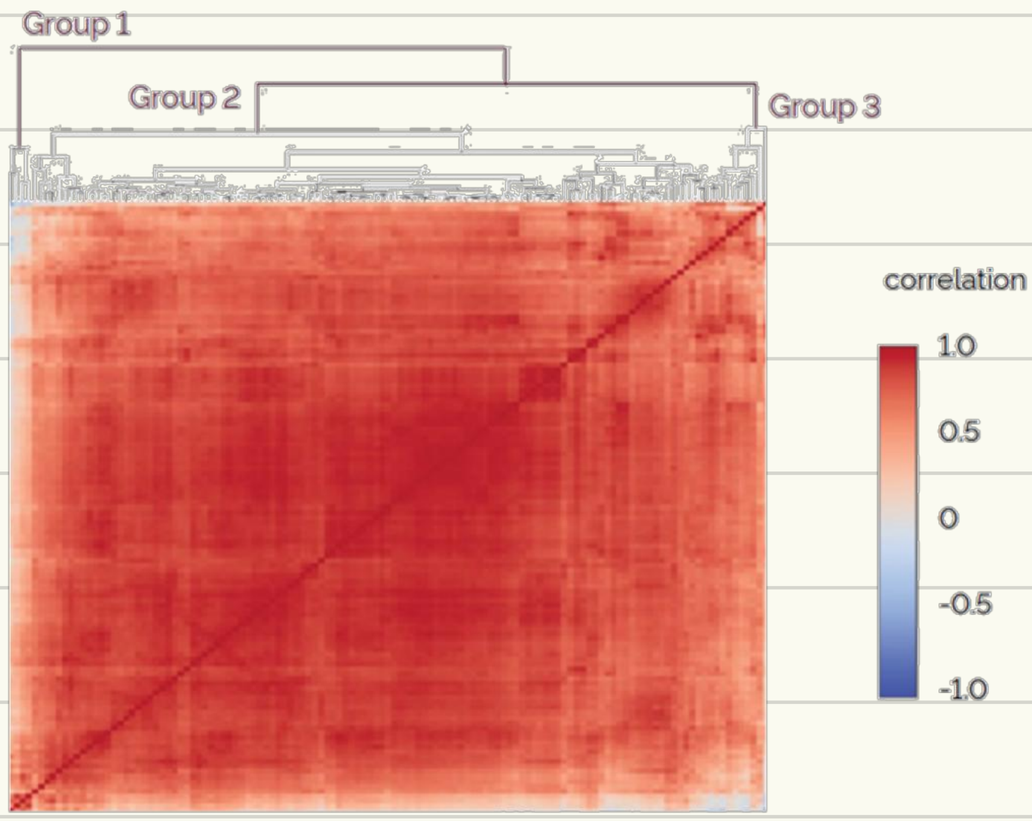
The ***high degree*** of ***correlation*** is a clear indicator that ***information*** present across the stations is ***redundant*** and could be ***eliminated*** or ***removed***

To identify possible features (stations) to remove ***hierarchical clustering*** can be applied to the correlation matrix

Hierarchical clustering groups all the ***stations*** based on how ***similar*** they are

The tree-like structure added to the heatmap is called a ***dendrogram*** and connects the ***sample*** based on their correlation vector ***proximity***

For example, the stations on the very left-hand side of the x-axis are grouped separately. This group contains stations connected to airports which have very different ridership patterns



Visualisation of the correlation matrix of the 14-day lag ridership predictions for non-holidays, weekdays in 2016.

BAR CHARTS → represent counts of categorical values

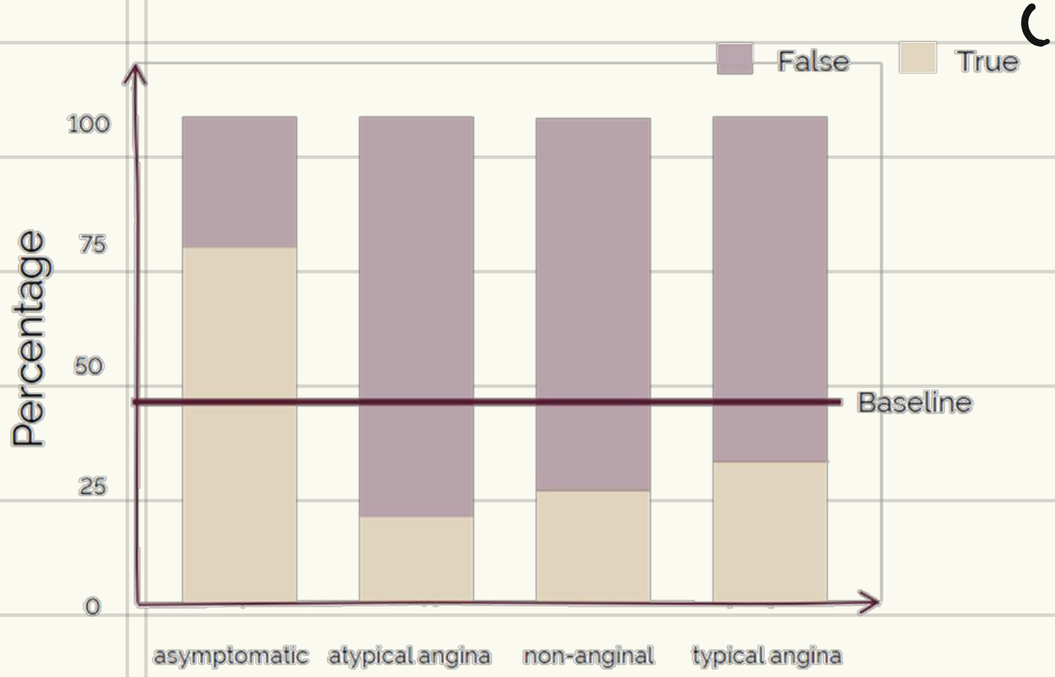
PROPORTIONS

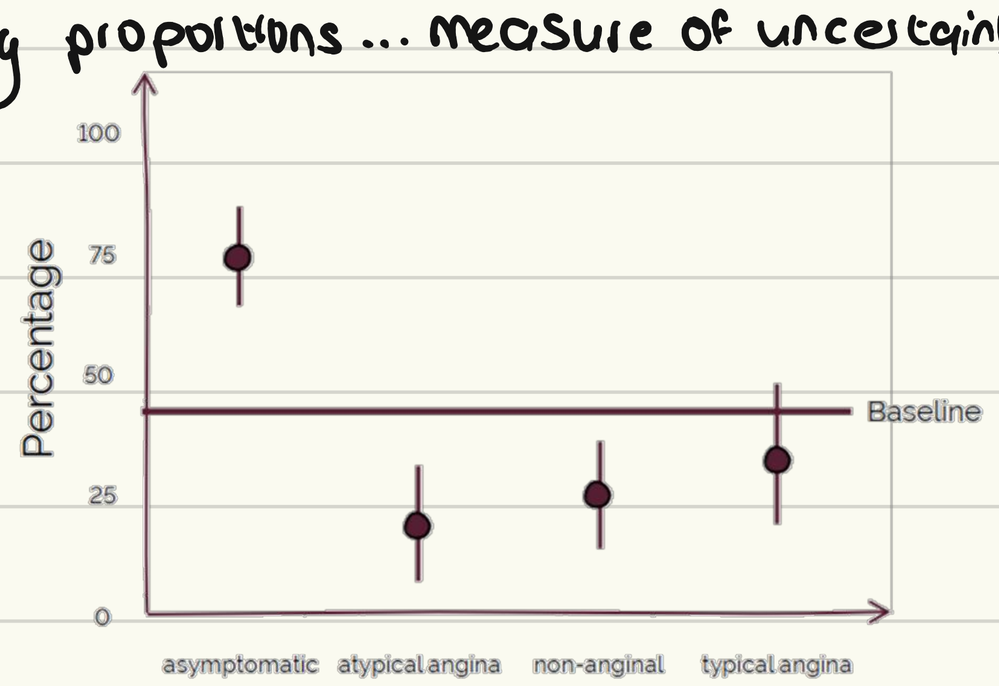
→ can better observe distn of each category relative to whole dataset

→ how much deviate from baseline, relative differences

CONFIDENCE INTERVALS ∴ helps establish if a proportion is due to chance.

(plotting proportions ... measure of uncertainty)





CHI-SQUARE

→ assess association btw 2 categorical vars.

→ ① compare observed vs expected frequencies in contingency table

② sum diffs

③ divide by expected freqs

④ sum values.

	observed			expected			chi	
	True	False		True	False		True	False
asymptomatic	105	39	**144**	66	78	**144**	23.0	19.5
atypical angina	9	41	**50**	23	27	**50**	8.5	7.2
non-angina	18	68	**86**	39	47	**86**	11.2	9.9
typical angina	7	16	**23**	10	13	**23**	1.2	1.0

THE DATA QUALITY REPORT

used to ① understand the data in an ABT

② identify any data quality issues

hmm seems familiar...

Data Quality Report

- ***Analytical base table***
  - ***Continuous features***
    - Tabular report
    - Histogram of each feature
  - ***Categorical features***
    - Tabular report
    - Bar plot of each feature

TABULAR REPORTS

describes characteristics of each feature using std. statistical measures of central tendency and variation

DATA VISUALISAIONS

distro of each feature

CONTINUOUS FEATURES

cardinality: number of distinct values

Data quality report – ***continuous features***

Feature	Count	% Missing	Card.	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.	Std. Dev.
Every continuous feature in the analytical base table	——	——	——	——	——	——	——	——	——	——
	——	——	——	——	——	——	——	——	——	——
	——	——	——	——	——	——	——	——	——	——

→ describes central tendency + variation of each feature using min, Q1 etc. + std dev.

→ histogram created for each continuous feature

EXCEPT: bar plot for card < 10

CATEGORICAL FEATURES

Data quality report – ***categorical features***

Feature	Count	% Missing	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
Every categorical feature in the analytical base table	——	——	——	——	——	——	——	——	——
	——	——	——	——	——	——	——	——	——
	——	——	——	——	——	——	——	——	——

Most frequent value | *2nd most frequent value*

→ 2 most frequent levels for each categorical feature

→ bar plot for each continuous feature.

ANALYSIS

The ***data quality report*** provides an ***in-depth*** picture of the ***data*** in an ***analytical base table*** and should be studied to ***get to know*** the ***data***

- For ***categorical*** features
  - Examine the mode, 2nd mode, mode %, and 2nd mode % as these indicate the most common levels within these features and will identify if any levels dominate the data set
- For ***continuous*** features
  - Examine the mean and standard division to understand the variation of the values within the data set
  - Examine the minimum and maximum values to understand the range that is possible for each feature

Possible ***data quality issues*** should be ***highlighted***, a corrective ***plan*** should be ***developed*** and ***implemented***, and the ***data quality plan*** should be ***updated***

Illustrative examp

Motor Insurance Fraud
Analytical base table

To illustrate how the ***data quality report*** can be ***constructed*** a ***claim data set*** where the task is to ***predict motor instance fraud*** will be used. An extract from the data is shown below

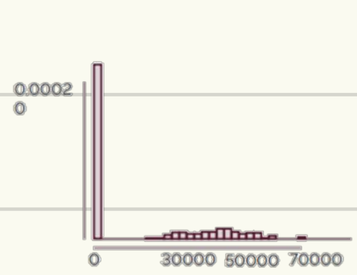
ID	Type	Income	Martial status	Num. of clmnts.	Injury type	Hospital stay	Claim amount	Total claimed	Num. of claims	Num. of soft tissue	% Soft Tissue	Claim amount received	Fraud flag
1	ci	0		2	soft tissue	no	1,625	3,250	2	2	1	0	1
2	ci	0		2	back	yes	15,028	60,112	1		0	15,028	0
3	ci	54,613	married	1	broken limb	no	-99,999	0	0	0	0	572	0
4	ci	0		4	broken limb	yes	5,097	11,661	1	1	1	7,864	0
5	ci	0		4	soft tissue	no	8,869	0	0	0	0	0	1
496	ci	0		1	soft tissue	no	2,118	0	0	0	0	0	1
497	ci	29,280	married	4	broken limb	yes	3,199	0	0	0	0	0	1
498	ci	0		1	broken limb	yes	32,469	0	0	0	0	16,763	0
499	ci	46,683	married	1	broken limb	no	179,448	0	0		0	179,448	0
500	ci	0		1	broken limb	no	8,259	0	0	0	0	0	1

Illustrative example

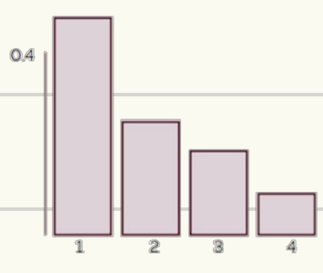
Data quality report: continuous features

Data quality report – ***continuous features (1/2)***

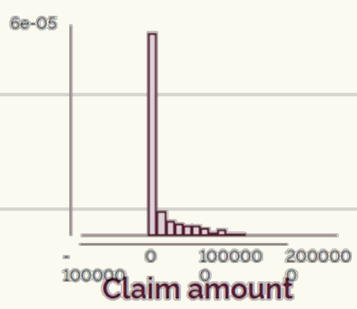
Feature	Count	% Missing	Card.	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.	Std. Dev.
Income	500	0.0	171	0	0	13,740	0	33,918	71,284	20,081
Num. of claimants	500	0.0	4	1	1	2	2	3	4	1
Claim amount	500	0.0	493	-99,999	3,233	16,373	5663	12,246	270,200	29,246
Total claimed	500	0.0	235	0	0	9,597	0	11,283	729,792	35,656



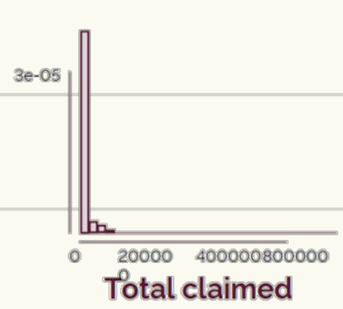
Income



Num. of claimants



Claim amount



Total claimed

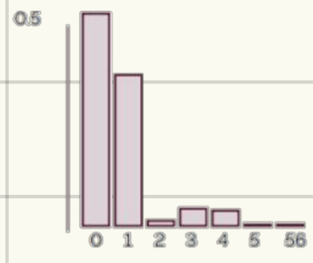
Income:
The income of several claimants is zero. Could indicate missing values

Claim amount:
One large negative claim amount[A]; could be a data entry mistake or a system default

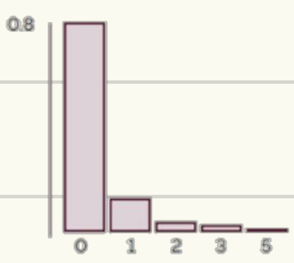
A Note the absence of multiple negative values in the corresponding histogram

Data quality report – ***continuous features (2/2)***

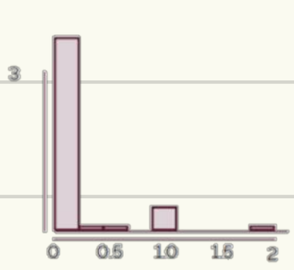
Feature	Count	% Missing	Card.	Min.	1st Quartile	Mean	Median	3rd Quartile	Max.	Std. Dev.
Num. claims	500	0.0	7	0	0	1	0	1	56	2.7
Num. soft tissue	500	2.0	6	0	0	0.2	0	0	6	0.6
% soft tissue	500	0.0	9	0	0	0.2	0	0	2	0.4
Amount received	500	0.0	329	0	0	13,051	3254	8,192	270,200	30,547
Fraud flag	500	0.0	2	0	0	0.3	0	1	1	0.5



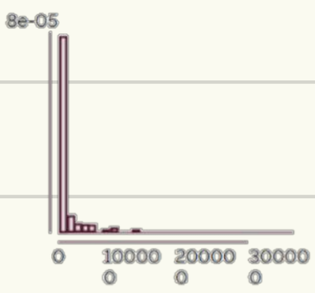
Num. claims



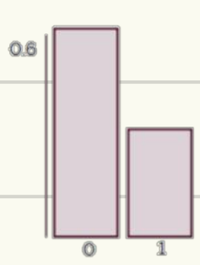
Num. soft tissue



% soft tissue



Amount received



Fraud flag

Num. claims & Amount received:
Contains irregular maximum values. Could represent a company policy

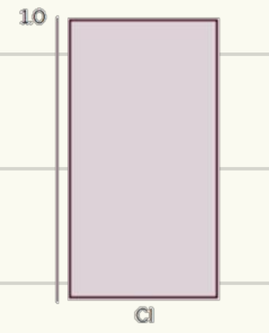
Num. soft tissue:
Contains missing values

Fraud:
Categorical feature encoded as a continuous feature; move to categorical quality report. Target is unbalanced

Motor Insurance Fraud
Data quality report: categorical features

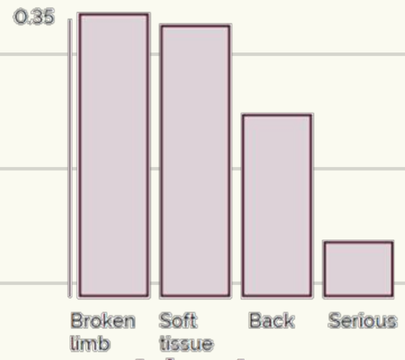
Data quality report – ***categorical features***

Feature	Count	% Missing	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
Insurance type	500	0.0	1	ci	500	100	-	-	-
Marital status	500	61.2	4	married	99	51	single	48	24.7
Injury type	500	0.0	4	broken limb	177	35.4	soft tissue	172	34.4
Hospital stay	500	0.0	2	no	354	70.8	yes	146	29.2



Insurance type

Marital status



Injury type

Hospital stay

Insurance type:
The feature has an irregular cardinality of 1. The feature can be removed contains no predictive power

Martial status:
Contains a high number of missing values. Several data analytical models cannot handle missing values