1 read data ✓ 2 clean data ✓ 3 remove duplicated data ✓
4 outliers (Detecting and Filtering) ✓
5 Feature selection ✓ 6 Data normalization ✓ 1

Subject : Date:

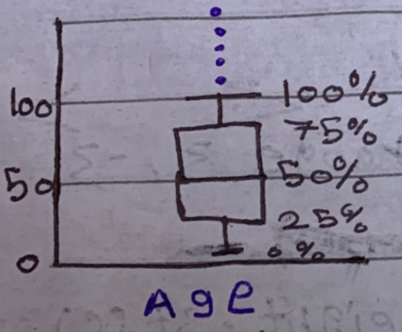
4 ≠ outliers detecting

مشكلة الـ outliers
وحلها

المعنى ⇐
إن بلاقي قيم خارج الـ Range إلى هي فيه بمعنى أنا عارف
إن متوسط الأعمار 65 سنة وأقصى عمر قد يصل إلى ١٠٠ عام
لكن ماذا عن وجدت أعمار تتجاوز الـ 500 والـ 700 هذا غير منطقي
فهذه القيم تسمى outliers

التعرف على هذه القيم بعدة طرق ⇐
١) عن طريق Domain Knowledge
أنا عارف إن الأعمار تصل إلى ١٠٠ فاهعمل شرط
age ≤ 105 إلى يتجاوز الشرط ده يبقى outliers
٢) عن طريق الـ Visualization

Box Plot , Bar Plot



precentile Max

Min

replace (outliers ممكن ـ , np.nan)   محمد عبد الله

Subject :

Date:

[1] Identify outliers by code   1- Domain Knowledge

1- Identify Max_Val, min_Val

معايا بيانات لطول الاشخاص المسموح بها للطول

[1] Max_Value = data['height'].quantile(0.95)

8.37 inch هنفترض هو وقف الطول لحد هنا مسموح بها

[2] Min_Value = data['height'].quantile(0.05)

2.3 أقل طول يسمح ~~بها~~ بة في الداتا

[3] Show The outliers [Identify outliers by domain]

data[data['height'] < min_value]

data[data['height'] > max_value]

name	highet
essra	0.9

name	height
amr	14.9
mostaFa	13.2
Ibrahim	15.3

[4] Remove outliers for simple data

dF[(dF['height'] < Max_Value) & (dF['height'] > min_Vale)]

الجملة دى هترجعلك الداتا من غير ال outliers

دى أهم شئ ال

Quntile