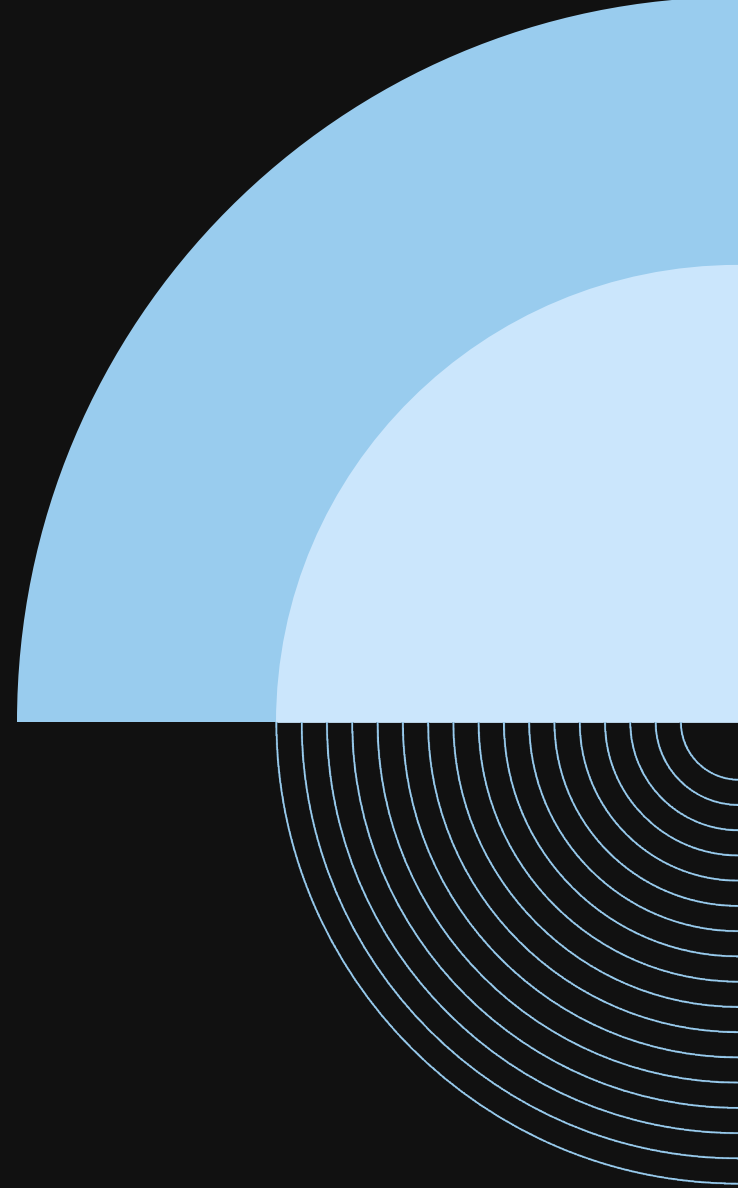


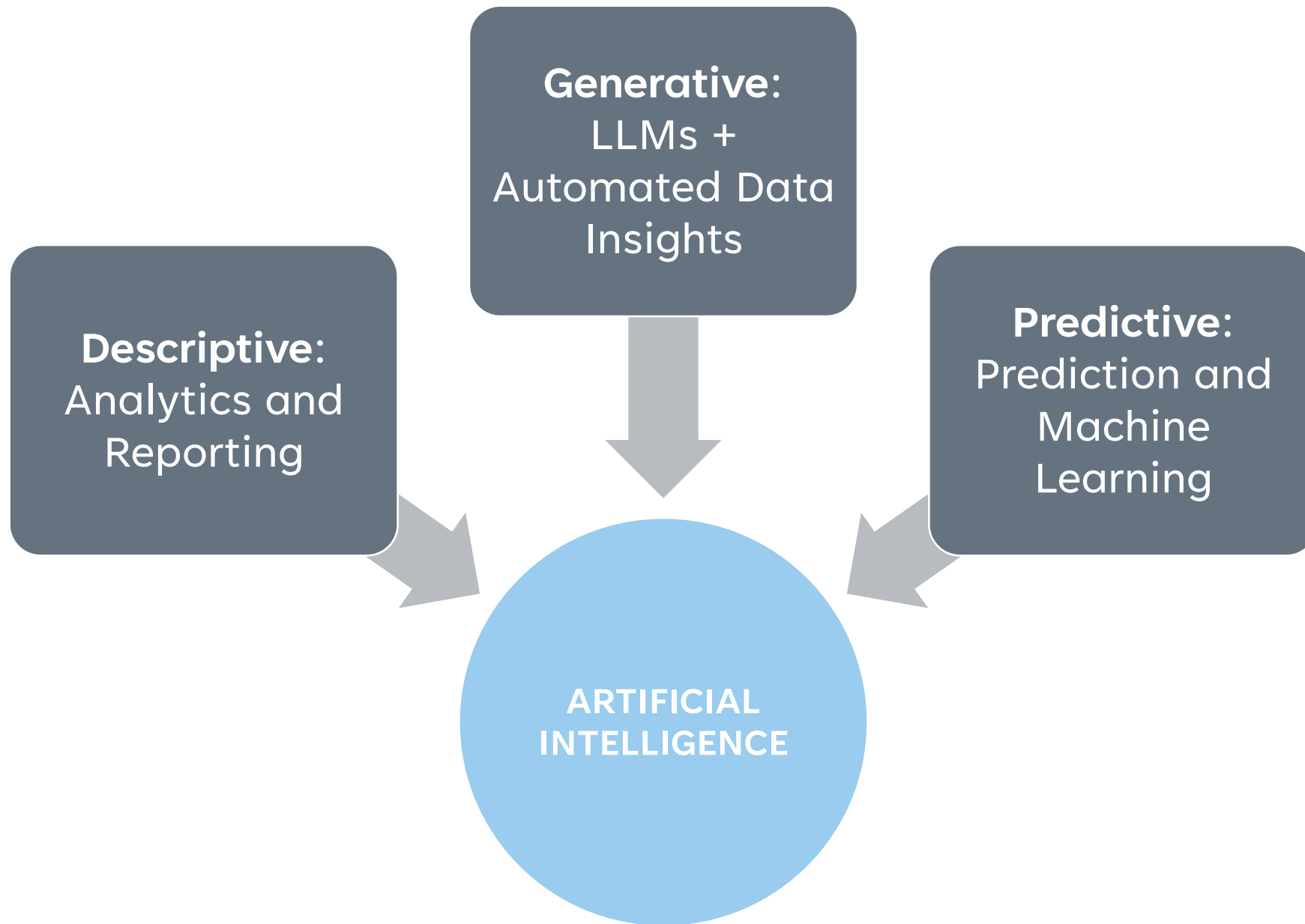
AI Ready: How do you avoid the pitfalls of AI?

Chris Willis: Chief Design Officer

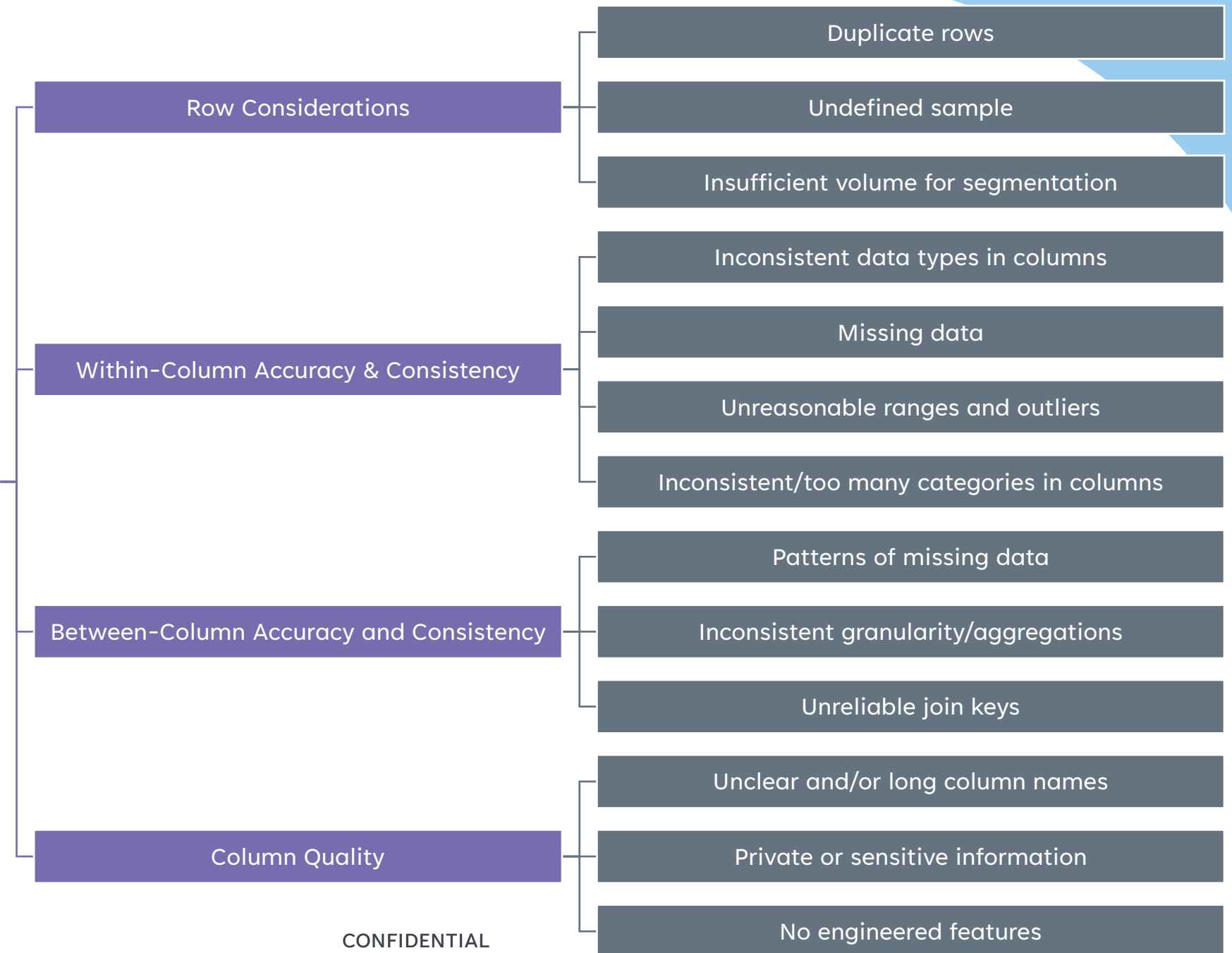
Florencia Silveira: Data Scientist



13 CARDINAL SINS OF AI-READY DATA



CARDINAL SINS FOR AI-READINESS



CONFIDENTIAL



- Duplicate rows
- Undefined sample
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

100,000,972	0	0	43
100,000,972	0	0	43



- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

C_000504	Yes
C_000507	Yes
C_000512	Yes
C_000274	Test
C_000360	Test
C_000408	Test
C_000509	Test
C_000001	No
C_000002	No
C_000004	No





Duplicate rows

Undefined sample of interest

Insufficient volume for segmentation

Inconsistent data types in columns

Missing data

Unreasonable ranges and outliers

Too many categories in columns

Patterns of missing data





Inconsistent granularity/aggregations

Unreliable join keys

Unclear and/or long column names

Private or sensitive information

No engineered features

Language Binary ⓘ	Genre 1 ⓘ	rating ⓘ	Count rating ⓘ
abc ⓘ	abc ⓘ	abc ⓘ	123 ⓘ COUNT ⓘ
			
2 unique values (approx) ⓘ + filter ⓘ	25 unique values (approx) ⓘ + filter ⓘ	8 unique values (approx) ⓘ + filter ⓘ	Loading distribution... ⓘ + filter ⓘ
Non-English	Adventure	PG-13	9
English	Romance	R	164
English	Comedy	PG	80
Non-English	Comedy	R	53
English	Horror	R	439
English	Drama	TV14	37
English	Mystery & thriller	R	366
English	Action	PG-13	200
Non-English	Drama	R	112
English	Comedy	PG-13	416
English	Comedy	R	757
English	Drama	PG-13	290
English	Documentary	PG-13	110
English	Musical	PG-13	21
English	Drama	R	658
Non-English	Horror	R	34
English	Adventure	PG-13	56
English	Fantasy	R	17
Non-English	Drama	PG-13	43
English	Romance	PG-13	173
English	Drama	PG	91
English	Western	R	28
English	Kids & family	PG	298



- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

invest_Total Amount	
abc	Text
<div></div>	
60 unique values (approx) + filter	
11867	
62979	
1357	
51630	
0 - No investment	
8315	
1072	
43742	
37698	
10231	
22468	
37025	
8315	
1072	





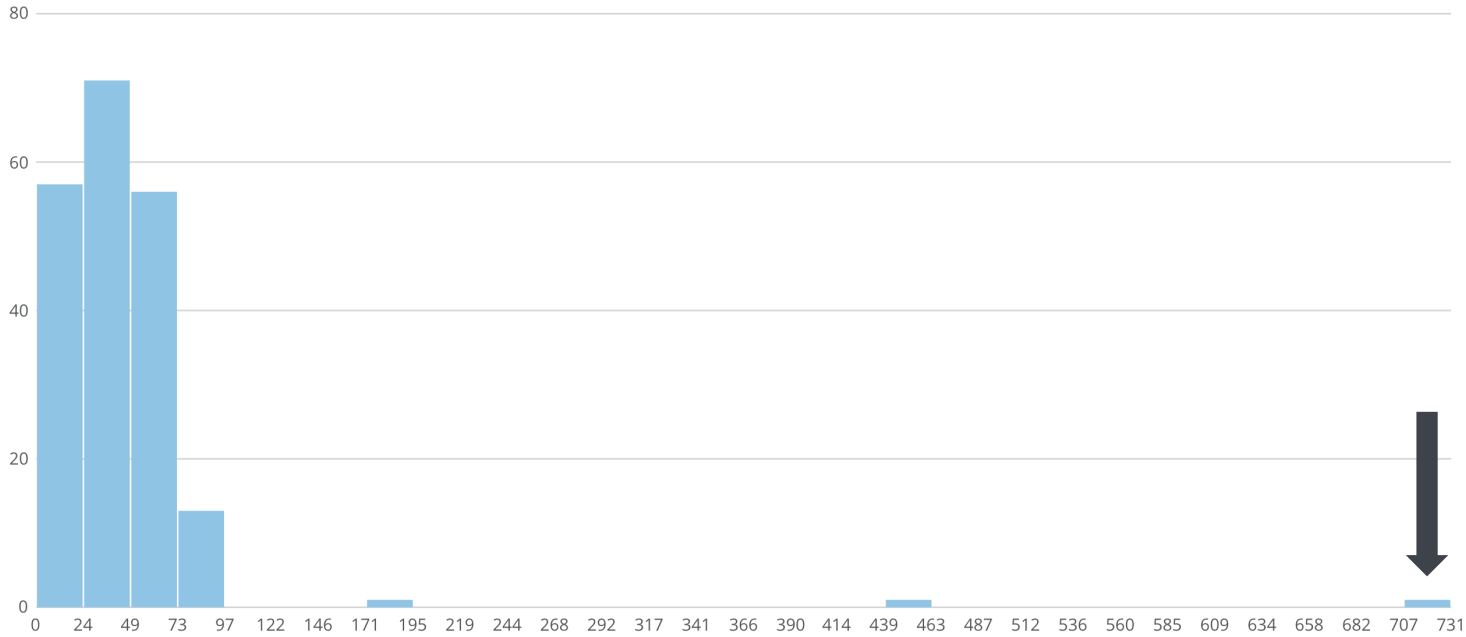
- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

Jun 30, 2021	32,760.31
Jul 31, 2021	33,414.70
Aug 31, 2021	34,982.10
Sep 30, 2021	
Oct 31, 2021	33,459.53
Nov 30, 2021	39,481.14
Dec 31, 2021	48,903.08
Jan 31, 2022	26,017.89
Feb 28, 2022	28,640.98
Mar 31, 2022	34,927.65
Apr 30, 2022	33,099.39
May 31, 2022	35,886.55
Jun 30, 2022	33,458.32
Aug 31, 2022	35,778.94
Sep 30, 2022	32,236.03
Oct 31, 2022	34,224.49



- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

46 Average of checking_Average Transactions per Month L12M ▾





- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

occupation	Count occupation
abc	123
Laborers	127,979
Sales staff	74,551
Core staff	65,179
Managers	50,164
Drivers	42,912
High skill tech staff	26,458
Accountants	22,715
Medicine staff	19,737
Security staff	15,480
Cooking staff	14,176
Cleaning staff	10,952
Private service staff	6,233
Low-skill Laborers	4,906
Secretaries	3,170
Waiters/barmen staff	3,147
Realty agents	1,700
HR staff	1,254
IT staff	1,079





Duplicate rows

Undefined sample of interest

Insufficient volume for segmentation

Inconsistent data types in columns

Missing data

Unreasonable ranges and outliers

Too many categories in columns

Patterns of missing data







Inconsistent granularity/aggregations

Unreliable join keys

Unclear and/or long column names

Private or sensitive information

No engineered features

	 DATE 		Metro Area 	Units 	
			abc	123	
1	Aug 1, 2017		Portland		
2	Aug 1, 2018		Portland		
3	Aug 1, 2019		Portland		
4	Aug 1, 2020		Portland		
5	Aug 1, 2021		Portland		
6	Aug 1, 2022		Portland		
7	Apr 1, 2017		Denver		
8	Apr 1, 2018		Denver		
9	Apr 1, 2019		Denver		
10	Apr 1, 2020		Denver		
11	Apr 1, 2021		Denver		
12	Apr 1, 2022		Denver		
13	Apr 1, 2023		Denver		



- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

applicant_income	co_applicant_income
1.23	1.23
1,600.00	19,200.00
4,166.00	49,992.00
4,895.00	58,740.00
9,703.00	116,436.00
5,166.00	61,992.00
3,254.00	39,048.00
5,923.00	71,076.00
11,500.00	138,000.00
4,333.00	51,996.00
3,708.00	44,496.00
3,029.00	36,348.00
6,080.00	72,960.00
7,200.00	86,400.00
4,566.00	54,792.00
4,917.00	59,004.00
1,836.00	22,032.00
5,800.00	69,600.00
3,676.00	44,112.00
2,957.00	35,484.00
6,783.00	81,396.00
3,676.00	44,112.00
8,334.00	100,008.00
3,667.00	44,004.00
8,666.00	103,992.00
3,547.00	42,564.00
3,036.00	36,432.00

- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

SKU_Master	
abc	
987654321098765	
890123456789012	
876543210987654	
765432109876543	
654321098765432	
567890123456789	
543210987654321	
456789012345678	
345678901234567	
321098765432109	
234567890123456	
123456789012345	
009856214789302	
007894562301478	
006789432105876	
003214567890189	
002134567890456	
001234567890123	
000987654321098	
000854762301987	
000678945612387	
000598742136985	
000432198765432	
000321654987032	



SKU ⓘ	
abc	
1234567890123	
123456789012345	
2134567890456	
234567890123456	
321098765432109	
3214567890189	
321654987032	
345678901234567	
432198765432	
456789012345678	
543210987654321	
567890123456789	
598742136985	
654321098765432	
6789432105876	
678945612387	
765432109876543	
7894562301478	
854762301987	
876543210987654	
890123456789012	
9856214789302	
987654321098	
987654321098765	





Duplicate rows	
Undefined sample of interest	
Insufficient volume for segmentation	
Inconsistent data types in columns	
Missing data	
Unreasonable ranges and outliers	
Too many categories in columns	
Patterns of missing data	
Inconsistent granularity/aggregations	
Unreliable join keys	
Unclear and/or long column names	
Private or sensitive information	
No engineered features	

4	Labor force participation rate, female (% of female population ages 15+) (national estimate)	
5	Proportion of seats held by women in national parliaments (%)	
6	Women Business and the Law Index Score (scale 1-100)	
7	Adolescents out of school, female (% of female lower secondary school age)	TR_14A
8	Children out of school, female (% of female primary school age)	
9	School enrollment, secondary, female (% gross)	TR_14B
10	GDP per capita (current US\$)	TR_14C
11	Adolescent fertility rate (births per 1,000 women ages 15-19)	
12	Birth rate, crude (per 1,000 people)	ITEM_DEPT49
13	Births attended by skilled health staff (% of total)	DEPT_LEV1
14	Current health expenditure (% of GDP)	
15	Current health expenditure per capita (current US\$)	DEPT_LEV2
16	Fertility rate, total (births per woman)	TOTAL
17	Hospital beds (per 1,000 people)	
18	Low-birthweight babies (% of births)	
19	Mortality rate, infant (per 1,000 live births)	
20	Mortality rate, neonatal (per 1,000 live births)	



Duplicate rows

Undefined sample of interest

Insufficient volume for segmentation

Inconsistent data types in columns

Missing data

Unreasonable ranges and outliers

Too many categories in columns

Patterns of missing data

Inconsistent granularity/aggregations

Unreliable join keys

Unclear and/or long column names

Private or sensitive information

No engineered features

Addressess Names
Email Addresses
Telephone Numbers
Dates of Birth
Credit Card Numbers
Social Security Numbers



- Duplicate rows
- Undefined sample of interest
- Insufficient volume for segmentation
- Inconsistent data types in columns
- Missing data
- Unreasonable ranges and outliers
- Too many categories in columns
- Patterns of missing data
- Inconsistent granularity/aggregations
- Unreliable join keys
- Unclear and/or long column names
- Private or sensitive information
- No engineered features

Start Date	Termination Date
Sep 12, 2011	Apr 29, 2022
Oct 13, 2019	Mar 28, 2023
Apr 24, 2005	
Jul 26, 2020	Jul 26, 2020
May 4, 2005	May 4, 2005
Nov 23, 2008	
Jun 6, 2000	Jun 15, 2023
Jul 24, 2003	
Dec 11, 2016	Jun 18, 2023
May 21, 2006	Sep 1, 2023
Sep 3, 2014	Jun 20, 2022
Aug 30, 2014	Feb 6, 2022
Mar 4, 2001	Dec 26, 2023
Jul 1, 2009	