









Table Comparator

Problem	2
Rozwiązanie	2
Przykład – klucz prosty	3
Dane wejściowe	3
Wynik.....	3
Przykład – klucz złożony	5
Wynik.....	5
Obsługa wyjątków	7
Duplikaty	7
Niezgódna liczba kolumn	7
Niezgódne nazwy kolumn	7
Niezgódne typy kolumn.....	7
Sposób działania.....	8

Problem

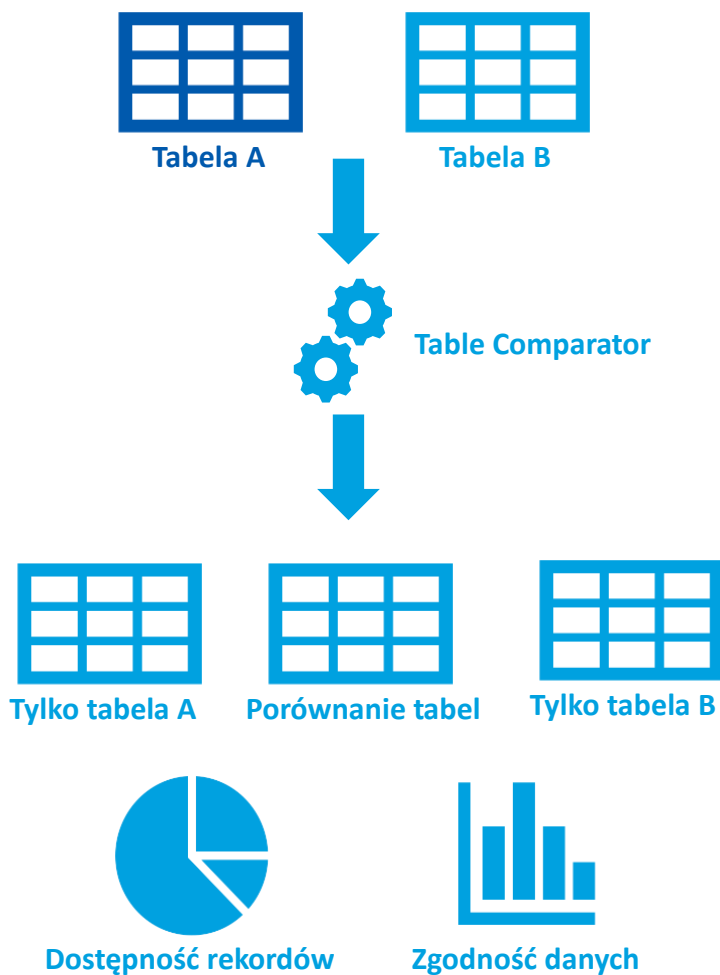
Mamy dwie tabele, które powinny być identyczne. W rzeczywistości występują jednak między nimi różnice.

Stan oczekiwany	Stan faktyczny
  =  	  ≠  

Chcemy te różnice wykryć, zwracając uwagę na to, czy:

- zawierają te same kolumny
- typy danych w kolumnach są takie same
- w obu tabelach występują te same rekordy
- atrybuty rekordów są takie same

Rozwiązanie



Przykład – klucz prosty

Dane wejściowe

Tabela A				Tabela B			
PESEL	Imię	Nazwisko	Zawód	PESEL	Imię	Nazwisko	Zawód
86485735888	Jan	Kowalski	informatyk	29432533772	Anna	Chmielewska	tancerka
79073971505	Tomasz	Nowak	przedsiębiorca	22340505862	Kacper	Karnowski	
29432533772	Anna	Chmielewska	tancerka	48395489710	Andrzej	Mostkowiak	mechanik
22340505862	Kacper	Karnowski		23594251784	Robert	Tomaszewski	nauczyciel
48395489710	Andrzej	Mostkowiak	mechanik	95097548444	Aneta	Morawska	sprzedawca
23594251784	Robert	Tomaszewski	nauczyciel	66488254309	Olek	Tomas	wykończeniowiec
95097548444	Aneta	Morawska	sprzedawczyni	64188385581	Barbara	Kurzewska	urzędniczka
66488254309	Aleksander	Tomas	wykończeniowiec	12345678987	Anna	Kołodziej	bezrobotna
64188385581	Barbara	Romczyk	urzędniczka	98765432123	Kacper	Stawarski	ogrodnik

Wynik

1. Wyłączenie tabela A:

PESEL	Imię	Nazwisko	Zawód
86485735888	Jan	Kowalski	informatyk
79073971505	Tomasz	Nowak	przedsiębiorca

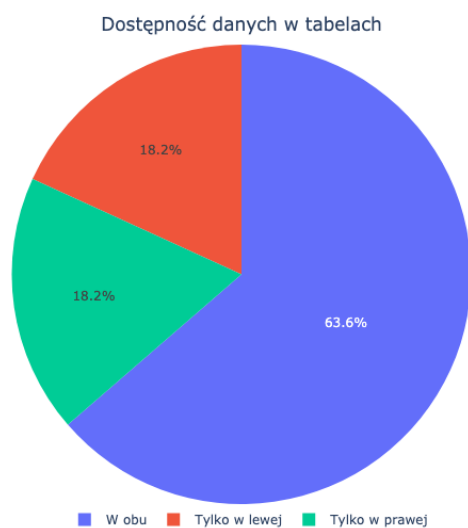
2. Wyłączenie tabela B:

PESEL	Imię	Nazwisko	Zawód
12345678987	Anna	Kołodziej	bezrobotna
98765432123	Kacper	Stawarski	ogrodnik

3. Tabela porównawcza:

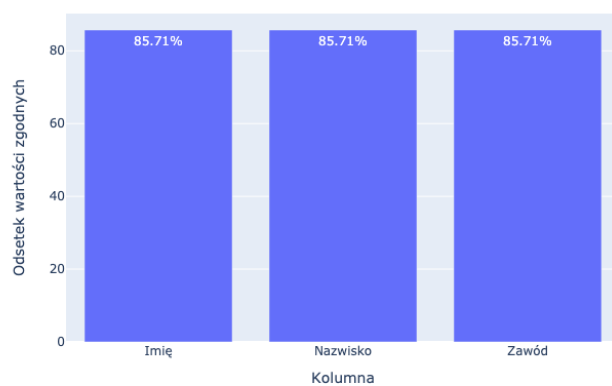
comparison	PESEL	Imię	Nazwisko	Zawód
left_table	29432533772	Anna	Chmielewska	tancerka
right_table	29432533772	Anna	Chmielewska	tancerka
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	22340505862	Kacper	Karnowski	
right_table	22340505862	Kacper	Karnowski	
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	48395489710	Andrzej	Mostkowiak	mechanik
right_table	48395489710	Andrzej	Mostkowiak	mechanik
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	23594251784	Robert	Tomaszewski	nauczyciel
right_table	23594251784	Robert	Tomaszewski	nauczyciel
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	95097548444	Aneta	Morawska	sprzedawczyni
right_table	95097548444	Aneta	Morawska	sprzedawca
comparison_result	TRUE	TRUE	TRUE	FALSE
left_table	66488254309	Aleksander	Tomas	wykończeniowiec
right_table	66488254309	Olek	Tomas	wykończeniowiec
comparison_result	TRUE	FALSE	TRUE	TRUE
left_table	64188385581	Barbara	Romczyk	urzędniczka
right_table	64188385581	Barbara	Kurzewska	urzędniczka
comparison_result	TRUE	TRUE	FALSE	TRUE

4. Dostępność danych w tabelach:



5. Zgodność wartości kolumn:

Zgodność wartości według kolumn (w %)



Przykład – klucz złożony

Tabela A				Tabela B			
Miasto	Numer	Marka	Model	Miasto	Numer	Marka	Model
SWD	AC5644	Honda	Civic	SWD	UBC45	BMW	M3
SWD	7JV7	Toyota	Corolla	SWD	AC5644	Honda	Civic
WA	53Y55	VW	Passat	SWD	7JV7	Toyota	Corolla
WA	AC5644	Renault	Clio	WA	53Y55	VW	Passat
WA	7JV7	Peugeot	3008	WA	AC5644	Renault	Clio
SWD	53Y55	Mercedes	CLA	WA	7JV7	Peugeot	3008
WA	UBC45	Opel	astra	SWD	53Y55	Audi	A4
WA	43GTA	Toyota	Avensis	WA	UBC45	Opel	Astra

Wynik

- Wyłącznie tabela A:

Miasto	Numer	Marka	Model
WA	43GTA	Toyota	Avensis

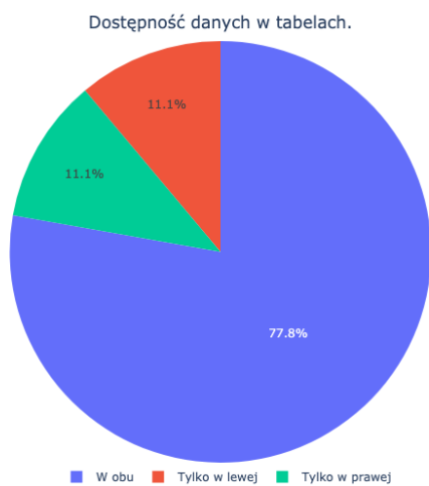
- Wyłącznie tabela B:

Miasto	Numer	Marka	Model
SWD	UBC45	BMW	M3

- Tabela porównawcza:

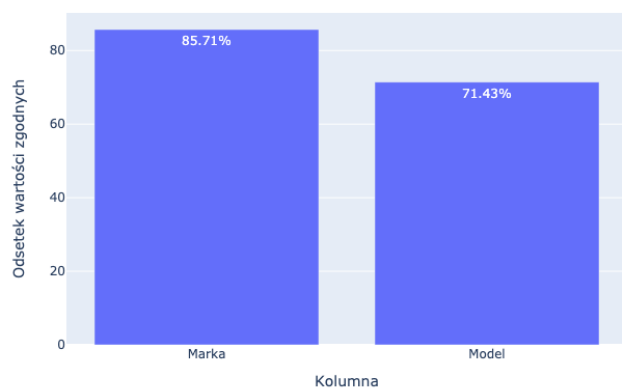
comparison	Miasto	Numer	Marka	Model
left_table	SWD	53Y55	Mercedes	CLA
right_table	SWD	53Y55	Audi	A4
comparison_result	TRUE	TRUE	FALSE	FALSE
left_table	SWD	AC5644	Honda	Civic
right_table	SWD	AC5644	Honda	Civic
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	SWD	7JV7	Toyota	Corolla
right_table	SWD	7JV7	Toyota	Corolla
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	WA	53Y55	VW	Passat
right_table	WA	53Y55	VW	Passat
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	WA	UBC45	Opel	astra
right_table	WA	UBC45	Opel	Astra
comparison_result	TRUE	TRUE	TRUE	FALSE
left_table	WA	AC5644	Renault	Clio
right_table	WA	AC5644	Renault	Clio
comparison_result	TRUE	TRUE	TRUE	TRUE
left_table	WA	7JV7	Peugeot	3008
right_table	WA	7JV7	Peugeot	3008
comparison_result	TRUE	TRUE	TRUE	TRUE

4. Dostępność danych w tabelach



5. Zgodność wartości kolumn:

Zgodność wartości według kolumn (w %)



Obsługa wyjątków

Duplikaty

PESEL	Imię	Nazwisko	Zawód
86485735888	Jan	Kowalski	informatyk
79073971505	Tomasz	Nowak	przedsiębiorca
29432533772	Anna	Chmielewska	tancerka
64188385581	Barbara	Romczyk	urzędniczka
48395489710	Andrzej	Mostkowiak	mechanik
23594251784	Robert	Tomaszewski	nauczyciel
95097548444	Aneta	Morawska	sprzedawczyni
66488254309	Aleksander	Tomas	wykończeniowiec
66488254309	Kacper	Stawarski	mechanik

Nie zgodna liczba kolumn

PESEL	Imię	Nazwisko	Zawód	Pies
86485735888	Jan	Kowalski	informatyk	Burek
79073971505	Tomasz	Nowak	przedsiębiorca	Reksio
29432533772	Anna	Chmielewska	tancerka	Wolf
64188385581	Barbara	Romczyk	urzędniczka	Scooby
48395489710	Andrzej	Mostkowiak	mechanik	Pluto
23594251784	Robert	Tomaszewski	nauczyciel	Kudłaty
95097548444	Aneta	Morawska	sprzedawczyni	Puszek
66488254309	Aleksander	Tomas	wykończeniowiec	Antek

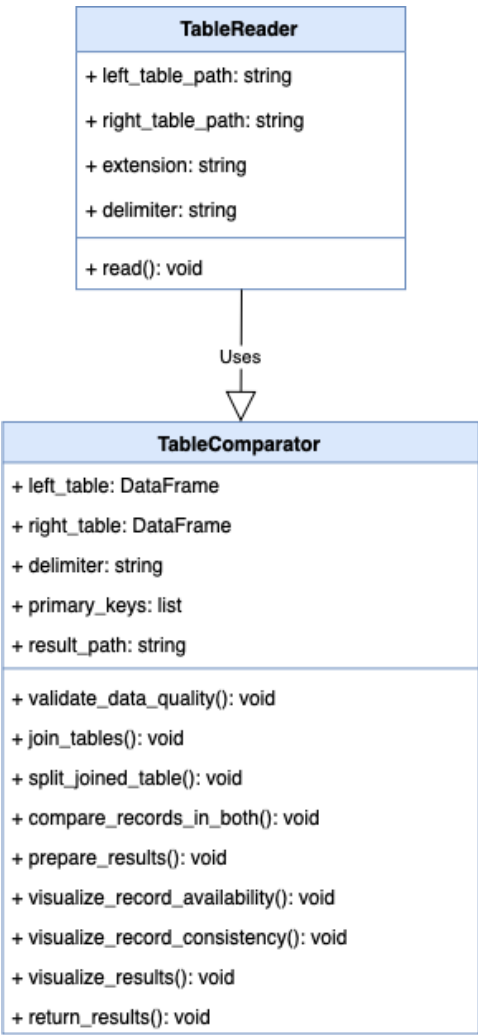
Nie zgodne nazwy kolumn

PESEL	Imię	Nazwisko	Profesja
86485735888	Jan	Kowalski	informatyk
79073971505	Tomasz	Nowak	przedsiębiorca
29432533772	Anna	Chmielewska	tancerka
64188385581	Barbara	Romczyk	urzędniczka
48395489710	Andrzej	Mostkowiak	mechanik
23594251784	Robert	Tomaszewski	nauczyciel
95097548444	Aneta	Morawska	sprzedawczyni
66488254309	Aleksander	Tomas	wykończeniowiec

Nie zgodne typy kolumn

PESEL	Imię	Nazwisko	Zawód
8648573588s	Jan	1	informatyk
79073971g05	Tomasz	2	przedsiębiorca
29432533772	Anna	3	tancerka
22340505862	Kacper	4	
48395489710	Andrzej	5	mechanik
23594251784	Robert	6	nauczyciel
95097548444	Aneta	7	sprzedawczyni
66488254309	Aleksander	8	wykończeniowiec
64188385581	Barbara	9	urzędniczka

Sposób działania

Diagram	Opis
 <pre> classDiagram class TableReader { +left_table_path: string +right_table_path: string +extension: string +delimiter: string +read(): void } class TableComparator { +left_table: DataFrame +right_table: DataFrame +delimiter: string +primary_keys: list +result_path: string +validate_data_quality(): void +join_tables(): void +split_joined_table(): void +compare_records_in_both(): void +prepare_results(): void +visualize_record_availability(): void +visualize_record_consistency(): void +visualize_results(): void +return_results(): void } TableReader --> TableComparator : Uses </pre>	<p>TableReader:</p> <ul style="list-style-type: none"> • odczytuje tabele • zapisuje tabele jako pd.DataFrame <p>Tabele te są wykorzystywane przez klasę TableComparator jako argumenty:</p> <ul style="list-style-type: none"> • left_table • right_table <p>TableComparator:</p> <ul style="list-style-type: none"> • dokonuje analizy danych • przygotowuje wizualizacje • zapisuje rezultaty <p>TableComparator zwraca następujące elementy:</p> <ul style="list-style-type: none"> • left_only: rekordy tylko w lewej tabeli • right_only: rekordy tylko w prawej tabeli • comparison_df: porównanie danych w obu tabelach • record_availability: wykres pokazujący dostępność danych • record_consistency: wykres pokazujący zgodność danych