

**Mata Kuliah Big Data & Data Analytics**

# **Laporan Praktikum Tugas 1: Retrive, Cleansing, and Manipulating**

**Dosen Pengampu: Agus Suhendar.S.T,.M.Eng**



**Disusun oleh:**

**Lathif Ramadhan (5231811022)**

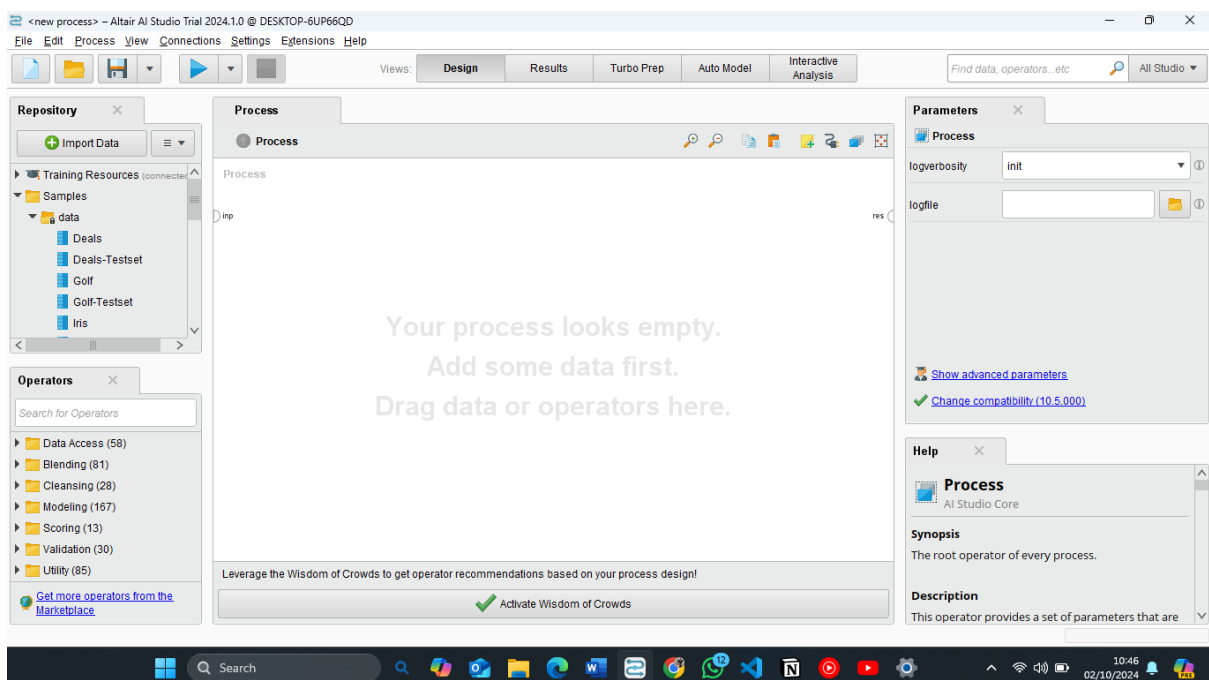
**PROGRAM STUDI SAINS DATA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS TEKNOLOGI YOGYAKARTA  
YOGYAKARTA**

**2024**

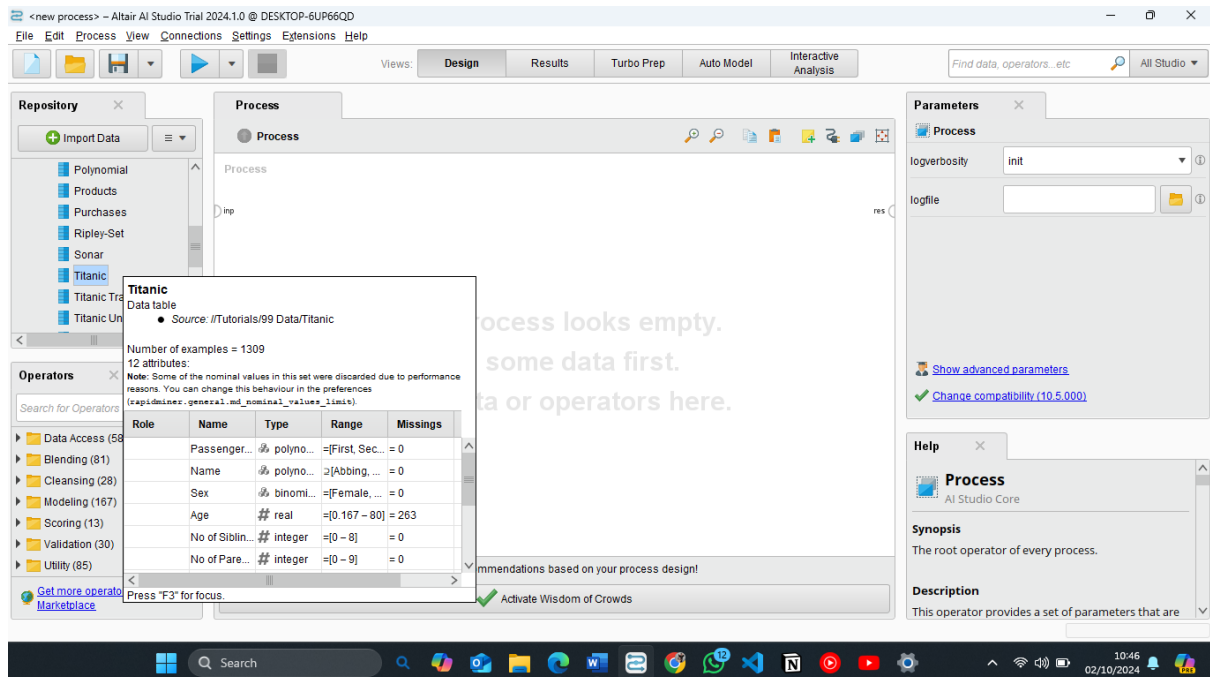
1. Pertama-tama, kita buka dulu aplikasi **Rapidminer** atau **Altair AI Studio**.



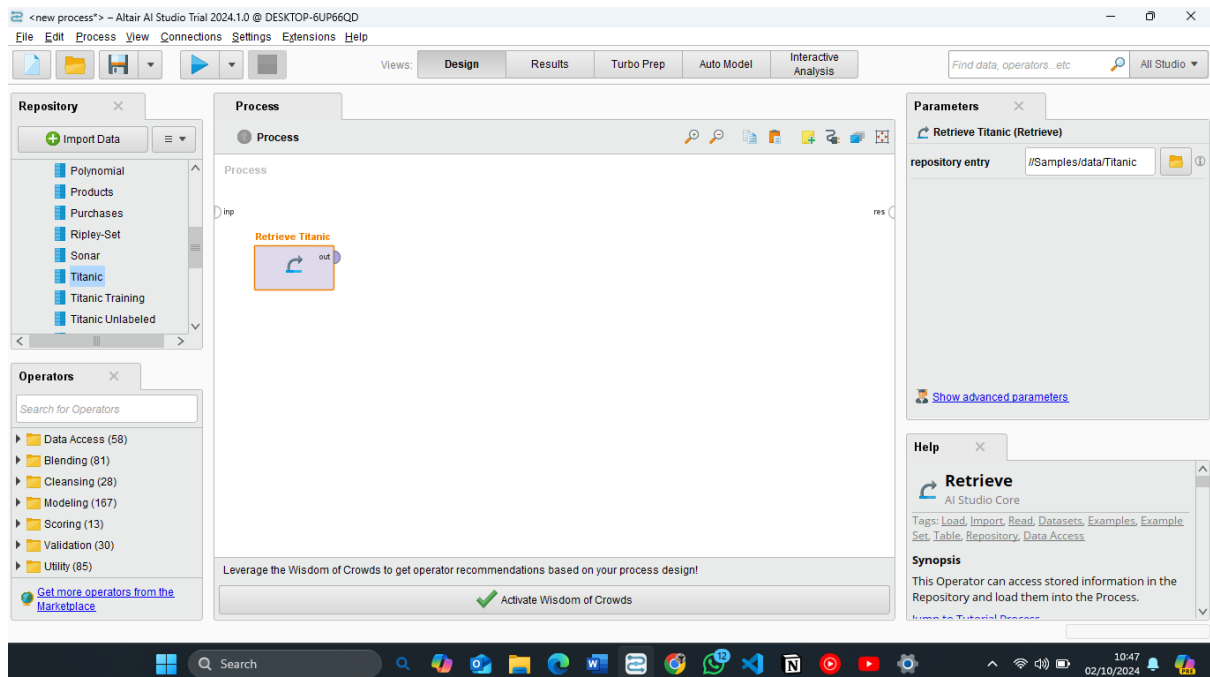
2. Setelah itu, kita buka pada **Repository > Samples > data > cari data Titanic!!**



3. Nah, kalau dataset ***Titanic*** telah kita temukan, maka kita ***drag*** dataset tersebut ke dalam ***Process***.



4. Ketika sudah di ***drag***, tarik ***out*** ke ***res***. Lalu, kita ***play/run***.



5. Maka, setelah di **play**, hasilnya akan seperti ini.

The screenshot shows the Altair AI Studio interface. The 'Results' tab is active, displaying a table titled 'ExampleSet (Retrieve Titanic)'. The table has 14 rows and 10 columns: Row No., Passenger..., Name, Sex, Age, No of Sibling..., No of Parent..., Ticket Num..., Passenger F..., and Cabin. The data includes passengers like Allen, Allison, Anderson, Andrews, Appleton, Artagaveyfia, Astor, Aubart, and Barber. The interface also shows a 'Repository' panel on the right with various datasets like Golf-TestSet, Iris, Labor-Negotiations, etc. The bottom status bar indicates 'ExampleSet (1,309 examples, 0 special attributes, 12 regular attributes)'.

Row No.	Passenger...	Name	Sex	Age	No of Sibling...	No of Parent...	Ticket Num...	Passenger F...	Cabin
1	First	Allen, Miss. E...	Female	29	0	0	24160	211.338	B5
2	First	Allison, Mast...	Male	0.917	1	2	113781	151.550	C22
3	First	Allison, Miss. ...	Female	2	1	2	113781	151.550	C22
4	First	Allison, Mr. H...	Male	30	1	2	113781	151.550	C22
5	First	Allison, Mrs. ...	Female	25	1	2	113781	151.550	C22
6	First	Anderson, Mr...	Male	48	0	0	19952	26.550	E12
7	First	Andrews, Mis...	Female	63	1	0	13502	77.958	D7
8	First	Andrews, Mr. ...	Male	39	0	0	112050	0	A36
9	First	Appleton, Mrs...	Female	53	2	0	11769	51.479	C10
10	First	Artagaveyfia, ...	Male	71	0	0	PC 17609	49.504	?
11	First	Astor, Col. Jo...	Male	47	1	0	PC 17757	227.525	C82
12	First	Astor, Mrs. Jo...	Female	18	1	0	PC 17757	227.525	C82
13	First	Aubart, Mme. ...	Female	24	0	0	PC 17477	69.300	B35
14	First	Barber, Miss. ...	Female	26	0	0	19877	78.850	?

6. Lalu, kita pergi ke bagian **Turbo Prep** untuk memulai **Manipulating Data** dan **Cleansing Data**. Kita fokus pada kolom-kolom yang diatas terdapat garis merahnya karena semakin panjang atau banyak garis merahnya, maka ada data yang kosong.

The screenshot shows the Altair AI Studio interface with the 'Turbo Prep' tab active. The 'Data Sets' panel on the left lists 'Titanic (3)', 'Titanic (2)', and 'Titanic'. The main area displays a table titled 'Titanic (3)' with 14 rows and 10 columns: Passenger Category, Name Category, Sex Category, Age Number, No of Sibling..., No of Parents..., Ticket Number Category, Passenger F..., Cabin Category, and Port of Category. The data includes passengers like Allen, Allison, Anderson, Andrews, Appleton, Artagaveyfia, Astor, Aubart, and Barber. The interface also shows a 'Repository' panel on the right with various datasets like Golf-TestSet, Iris, Labor-Negotiations, etc. The bottom status bar indicates '1,309 rows - 12 columns (8 nominal, 4 numerical)'.

Passenger CL...	Name	Sex	Age	No of Sibling...	No of Parents...	Ticket Number	Passenger F...	Cabin	Port of
Category	Category	Category	Number	Number	Number	Category	Number	Category	Category
First	Allen, Miss. Eli...	Female	29	0	0	24160	211.338	B5	Southa
First	Allison, Master...	Male	0.917	1	2	113781	151.550	C22 C26	Southa
First	Allison, Miss. H...	Female	2	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mr. Hud...	Male	30	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mrs. Hu...	Female	25	1	2	113781	151.550	C22 C26	Southa
First	Anderson, Mr. ...	Male	48	0	0	19952	26.550	E12	Southa
First	Andrews, Miss...	Female	63	1	0	13502	77.958	D7	Southa
First	Andrews, Mr. T...	Male	39	0	0	112050	0	A36	Southa
First	Appleton, Mrs. ...	Female	53	2	0	11769	51.479	C101	Southa

7. Untuk kolom pertama, kita pilih kolom **Cabin**. Maka, seluruh data beserta nama kolom **Cabin** akan bewarna orange seperti ini.

The screenshot shows the Altair AI Studio Turbo Prep interface. On the left, the 'Cleanse' sidebar is active, showing '1 column selected'. The main area displays a data table titled 'Titanic (3)'. The 'Cabin' column is highlighted in orange. The table has 10 columns: Passenger CL..., Name, Sex, Age, No of Sibling..., No of Parents..., Ticket Number, Passenger F..., Cabin, and Port of. The data rows show various passenger details, with the 'Cabin' column values like B5, C22 C26, E12, D7, A36, and C101.

Passenger CL...	Name	Sex	Age	No of Sibling...	No of Parents...	Ticket Number	Passenger F...	Cabin	Port of
First	Allen, Miss. Eli...	Female	29	0	0	24160	211.338	B5	Southa
First	Allison, Master...	Male	0.917	1	2	113781	151.550	C22 C26	Southa
First	Allison, Miss. H...	Female	2	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mr. Hud...	Male	30	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mrs. Hu...	Female	25	1	2	113781	151.550	C22 C26	Southa
First	Anderson, Mr. ...	Male	48	0	0	19952	26.550	E12	Southa
First	Andrews, Miss...	Female	63	1	0	13502	77.958	D7	Southa
First	Andrews, Mr. T...	Male	39	0	0	112050	0	A36	Southa
First	Appleton, Mrs. ...	Female	53	2	0	11769	51.479	C101	Southa

8. Kemudian, kita pergi ke **Cleanse > Replace Missing** > untuk **nominal missing**, pilih **specific value** > isi dengan kata “Tidak diketahui” > lalu klik **Apply**.

The screenshot shows the Altair AI Studio Turbo Prep interface. On the left, the 'Cleanse' sidebar is active, and the 'Replace Missing' option is selected. The 'Nominal missing' dropdown is set to 'specific value', and the text input field contains 'Tidak diketahui'. The main area displays the same data table as before, with the 'Cabin' column highlighted in orange.

Passenger CL...	Name	Sex	Age	No of Sibling...	No of Parents...	Ticket Number	Passenger F...	Cabin	Port of
First	Allen, Miss. Eli...	Female	29	0	0	24160	211.338	B5	Southa
First	Allison, Master...	Male	0.917	1	2	113781	151.550	C22 C26	Southa
First	Allison, Miss. H...	Female	2	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mr. Hud...	Male	30	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mrs. Hu...	Female	25	1	2	113781	151.550	C22 C26	Southa
First	Anderson, Mr. ...	Male	48	0	0	19952	26.550	E12	Southa
First	Andrews, Miss...	Female	63	1	0	13502	77.958	D7	Southa
First	Andrews, Mr. T...	Male	39	0	0	112050	0	A36	Southa
First	Appleton, Mrs. ...	Female	53	2	0	11769	51.479	C101	Southa

REMOVE CORRELATED

REPLACE MISSING

Nominal missings: specific value ▼

Tidak diketahui ✕

✓ APPLY

NORMALIZATION

Fi

Fi

Fi

Fi

Fi

Fi

Fi

Fi

Fi

9. Untuk menghindari data nama *double/duplicate* atau salah input. Kita klik kolom **Name** > dibagian **Cleanse** > klik **Remove Duplicate** > **Apply**.

<new process> - Altair AI Studio Trial 2024.1.0 @ DESKTOP-6UP66QD

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Interactive Analysis

Find data, operators, etc. All Studio

Cleanse

1 column selected

AUTO CLEANSING

REMOVE LOW QUALITY

REMOVE CORRELATED

REPLACE MISSING

NORMALIZATION

DISCRETIZATION

DUMMY ENCODING

PCA

Titanic (3)

Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make chang...

COMMIT CLEANSE

CANCEL

UNDO

SHOW HISTORY

Passenger Cl...  
Category

Name  
Category

Sex  
Category

Age  
Number

No of Sibling...  
Number

No of Parents...  
Number

Ticket Number  
Category

Passenger F...  
Number

Cabin  
Category

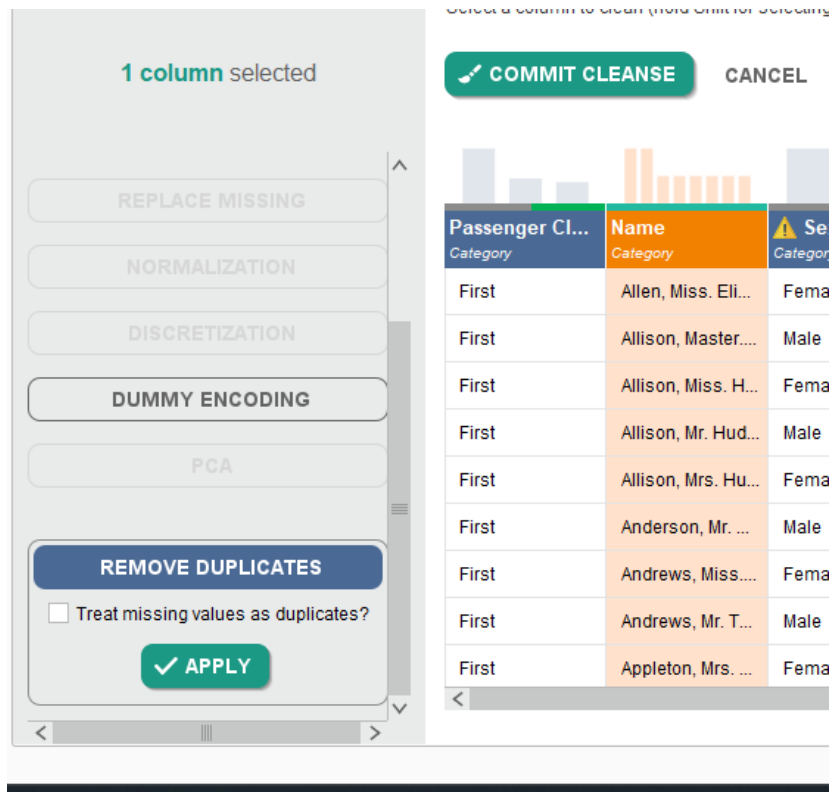
Port of  
Category

First	Allen, Miss. Eli...	Female	29	0	0	24150	211.338	B5	Southa
First	Allison, Master...	Male	0.917	1	2	113781	151.550	C22 C26	Southa
First	Allison, Miss. H...	Female	2	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mr. Hud...	Male	30	1	2	113781	151.550	C22 C26	Southa
First	Allison, Mrs. Hu...	Female	25	1	2	113781	151.550	C22 C26	Southa
First	Anderson, Mr. ...	Male	48	0	0	19952	26.550	E12	Southa
First	Andrews, Miss...	Female	63	1	0	13502	77.958	D7	Southa
First	Andrews, Mr. T...	Male	39	0	0	112050	0	A36	Southa
First	Appleton, Mrs. ...	Female	53	2	0	11769	51.479	C101	Southa

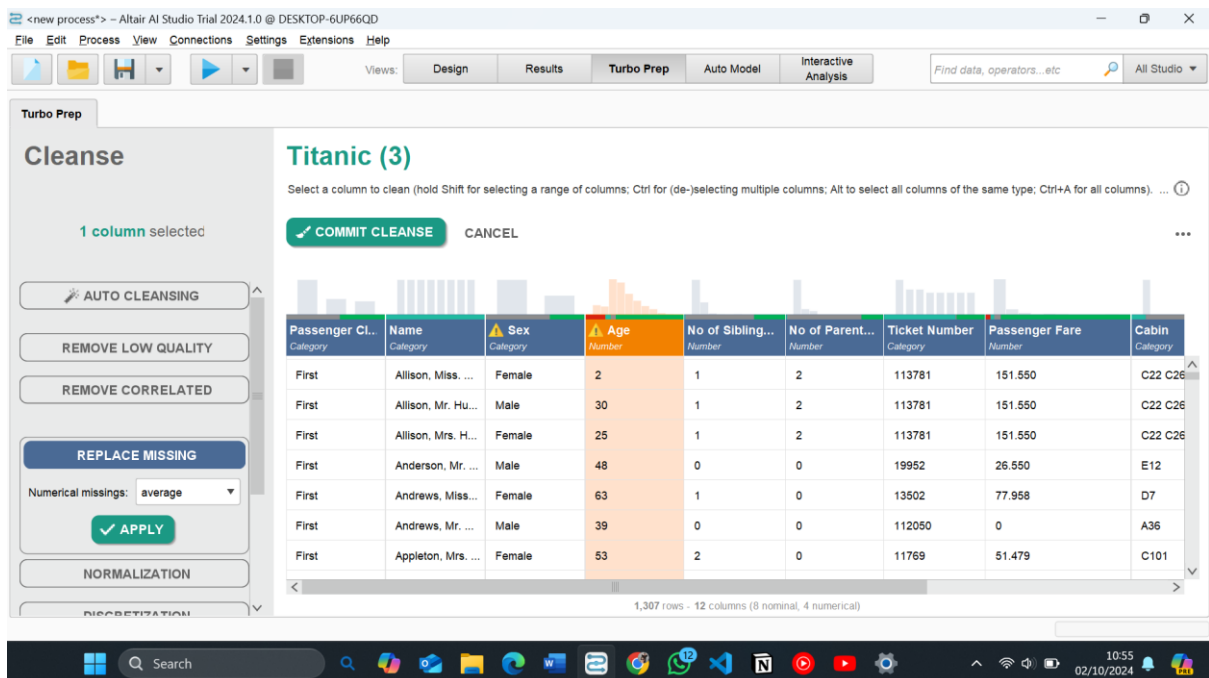
1,309 rows - 12 columns (8 nominal, 4 numerical)

Search

10:52 02/10/2024



10. Karena data di kolom **Age** juga terdapat data yang kosong. Maka, kita klik **Age > Replace Missing > Average**.





11. Selanjutnya kita olah data pada kolom **Life Boat** karena sebagian besar lebih banyak warna merah daripada hijau yang dimana artinya banyak data yang kosong/hilang. Maka, solusinya yaitu kita klik kolom **Life Boat > Replace Missing > Specific Value**.

The screenshot shows the Altair AI Studio Turbo Prep interface. On the left, the 'Cleanse' panel is active, showing '1 column selected'. The 'REPLACE MISSING' button is highlighted, and the 'specific value' dropdown is set to 'Tidak diketahui'. The 'TITANIC (3)' dataset is displayed in the center, with columns: No of Sibling..., No of Parent..., Ticket Number, Passenger Fare, Cabin, Port of Emb..., Life Boat, and Survived. The 'Life Boat' column is highlighted in orange, indicating it is the selected column for cleansing. The 'REPLACE MISSING' button is highlighted, and the 'specific value' dropdown is set to 'Tidak diketahui'. The 'TITANIC (3)' dataset is displayed in the center, with columns: No of Sibling..., No of Parent..., Ticket Number, Passenger Fare, Cabin, Port of Emb..., Life Boat, and Survived. The 'Life Boat' column is highlighted in orange, indicating it is the selected column for cleansing.

12. Setelah semua selesai, maka klik **Commit**.

The screenshot shows the Altair AI Studio Turbo Prep interface. On the left, the 'Cleanse' panel is active, showing '1 column selected'. The 'COMMIT CLEANSE' button is highlighted, indicating the cleansing process is complete. The 'TITANIC' dataset is displayed in the center, with columns: Passenger Cl..., Name, Sex, Age, No of Sibling..., No of Parent..., Ticket Number, Passenger Fa..., and Cabin. The 'COMMIT CLEANSE' button is highlighted, indicating the cleansing process is complete.



13. Setelah **Commit**, kita bisa klik **Export** untuk menyimpan data yang sudah kita kerjakan.

The screenshot shows the Altair AI Studio interface with the Titanic dataset loaded. The dataset table is visible with columns: Passenger Class, Name, Sex, Age, No of Siblings, No of Parents, Ticket Number, and Passenger Number. A context menu is open over the table, showing options: MODEL, CHARTS, CREATE PROCESS, HISTORY, EXPORT, COPY, and REMOVE. A tooltip over the 'EXPORT' option reads: 'Save the current data set in a repository or exports it to a file.'

Passenger Class	Name	Sex	Age	No of Siblings	No of Parents	Ticket Number	Passenger Number
First	Allen, Miss. Elis...	Female	29	0	0	24160	211.338
First	Allison, Master....	Male	0.917	1	2	113781	151.550
First	Allison, Miss. H...	Female	2	1	2	113781	151.550
First	Allison, Mr. Hu...	Male	30	1	2	113781	151.550
First	Allison, Mrs. Hu...	Female	25	1	2	113781	151.550
First	Anderson, Mr. ...	Male	48	0	0	19952	26.550
First	Andrews, Miss. ...	Female	63	1	0	13502	77.958

14. Nah, disini terdapat 4 pilihan format file yang akan digunakan untuk menyimpan data kita. Contohnya, kita memilih format **Excel** untuk file kita, maka klik **Excel** > lalu **Next**.

The screenshot shows the Altair AI Studio interface with the 'Export Data' dialog box open. The dialog box has four tabs: Select Format, Select Location, Writing, and Done. The 'Select Format' tab is active, showing four options: Repository, Qlik, Excel, and CSV. The 'Excel' option is selected, and the 'Next' button is highlighted. The background shows the Titanic dataset table.

Export Data 'Titanic'

Select Format | Select Location | Writing | Done

Repository: Store the data in a repository

Qlik: Export the data for Qlik (.qvx)

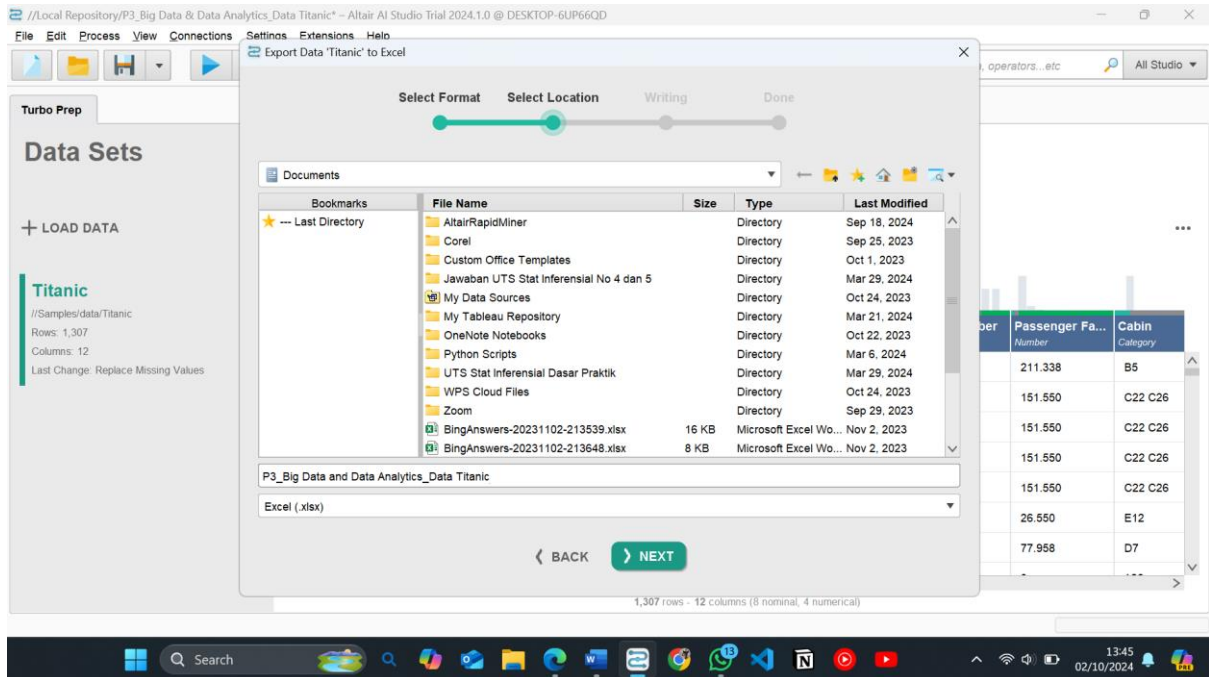
Excel: Export the data as Excel file (.xlsx)

CSV: Export the data as CSV file (.csv)

BACK | NEXT

1,307 rows - 12 columns (8 nominal, 4 numerical)

15. Lalu, kita beri nama file hasil pekerjaan kita dari data **Titanic** tadi. Kemudian pilih lokasi penyimpanan. Setelah itu klik **Next**.



16. Nah, sudah selesai. Kita bisa **Close**.

