

LinuxMeeting

VisiData e Pandas



Ing. Aldo Maria Bracco

Il mondo dei dati



Dato → Informazione → Conoscenza

Operazioni preliminari



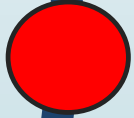
Gestione **missing**



Gestione **duplicati**



Operazioni **normalizzazione** e **standardizzazione**



Operazioni di **aggregazione**



Controllo di **qualità**



Scelta dello strumento

**Dammi sei ore per abbattere un
albero e spenderò le prime quattro
ore per affilare l'ascia.**

Abraham Lincoln

VisiData

- **VisiData** è un tool interattivo per operare su dati. Combina la **chiarezza** di un foglio di calcolo, l'**efficienza** del terminale e la **potenza** di Python in uno strumento che può gestire facilmente milioni di righe.

- Principali vantaggi sono:



rapidità d'utilizzo



free e **open source**



velocizza operazioni di ricerca, filtraggio, ordinamento...



utilizzo da **terminale**



scritto in **python3**

Installazione VisiData



- È possibile installare VisiData tramite l'installer di Python:
pip3 install visidata
- È possibile installare la versione di sviluppo tramite il comando:
pip3 install git+https://github.com/saulpw/visidata.git@develop
- È possibile personalizzare alcune caratteristiche modificando il file **.visidatarc**
- Se il file **.visidatarc** non è presente nella cartella **home** dell'utente è necessario crearlo

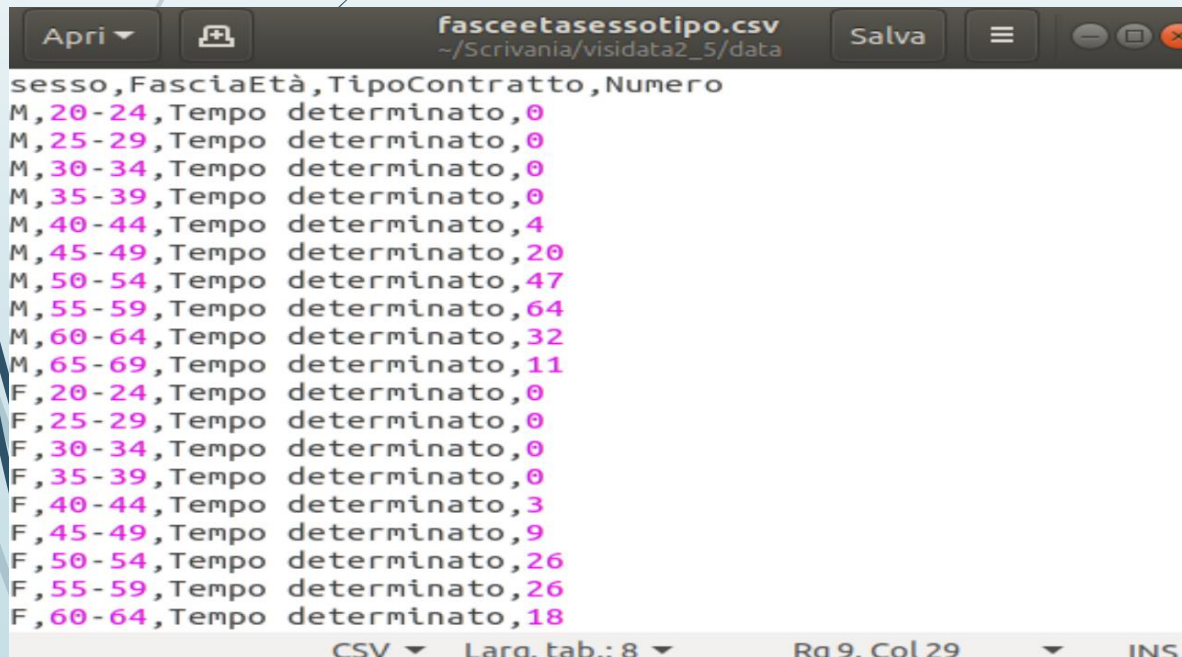
Aprire un file con Visidata

- VisiData supporta svariati **formati** di file:
 - CSV, TSV, sqlite, postgres, xlsx, json, vd...
- Per aprire un file con VisiData è sufficiente digitare il comando:
vd nomeFile
- Tramite VisiData è possibile aprire tramite l'**URL pubblico**:
vd 'https://...'

Esempio Visidata

- Tramite il link '<https://dati.regione.sicilia.it/dataset/personale-fasce-di-eta-per-sesso/resource/cac23b2a-c466-43ca-bb93-0ce28b3ae693>' è possibile effettuare il download dei dati relativi al personale regione Sicilia per fasce di età.
- Aprire il file tramite il comando: **vd fasceerasessotipo.csv**

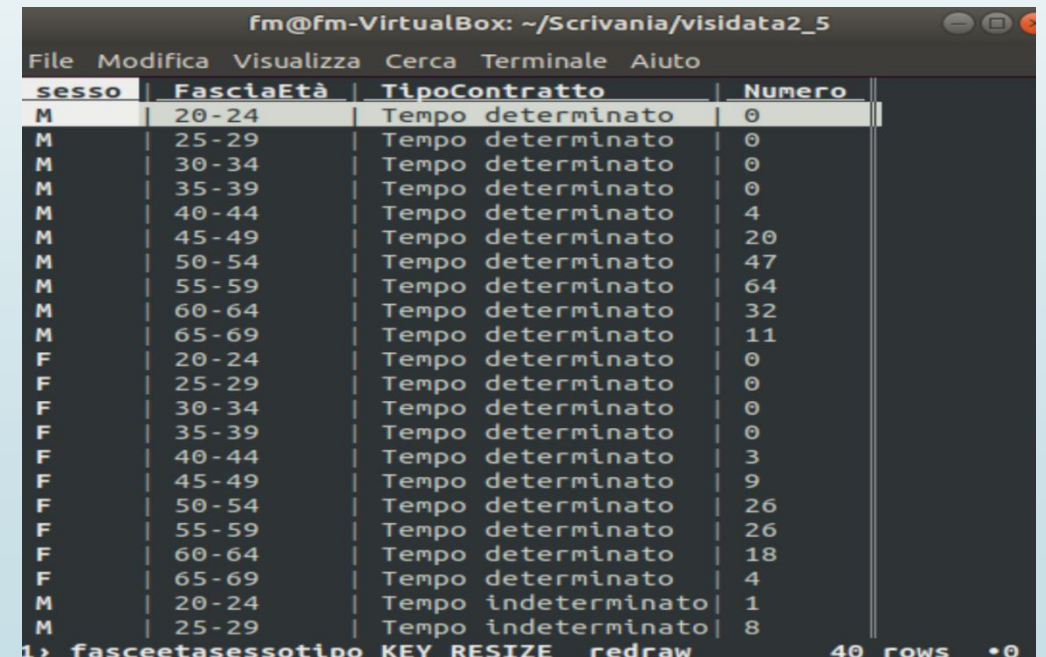
Editor di testo



The screenshot shows a text editor window titled 'fasceetasessotipo.csv' with the following content:

sezzo	FasciaEtà	TipoContratto	Numero
M	20-24	Tempo determinato	0
M	25-29	Tempo determinato	0
M	30-34	Tempo determinato	0
M	35-39	Tempo determinato	0
M	40-44	Tempo determinato	4
M	45-49	Tempo determinato	20
M	50-54	Tempo determinato	47
M	55-59	Tempo determinato	64
M	60-64	Tempo determinato	32
M	65-69	Tempo determinato	11
F	20-24	Tempo determinato	0
F	25-29	Tempo determinato	0
F	30-34	Tempo determinato	0
F	35-39	Tempo determinato	0
F	40-44	Tempo determinato	3
F	45-49	Tempo determinato	9
F	50-54	Tempo determinato	26
F	55-59	Tempo determinato	26
F	60-64	Tempo determinato	18

VisiData



The screenshot shows the VisiData application window titled 'fm@fm-VirtualBox: ~/Scrivania/visidata2_5'. The data is displayed in a table with the following content:

sezzo	FasciaEtà	TipoContratto	Numero
M	20-24	Tempo determinato	0
M	25-29	Tempo determinato	0
M	30-34	Tempo determinato	0
M	35-39	Tempo determinato	0
M	40-44	Tempo determinato	4
M	45-49	Tempo determinato	20
M	50-54	Tempo determinato	47
M	55-59	Tempo determinato	64
M	60-64	Tempo determinato	32
M	65-69	Tempo determinato	11
F	20-24	Tempo determinato	0
F	25-29	Tempo determinato	0
F	30-34	Tempo determinato	0
F	35-39	Tempo determinato	0
F	40-44	Tempo determinato	3
F	45-49	Tempo determinato	9
F	50-54	Tempo determinato	26
F	55-59	Tempo determinato	26
F	60-64	Tempo determinato	18
F	65-69	Tempo determinato	4
M	20-24	Tempo indeterminato	1
M	25-29	Tempo indeterminato	8

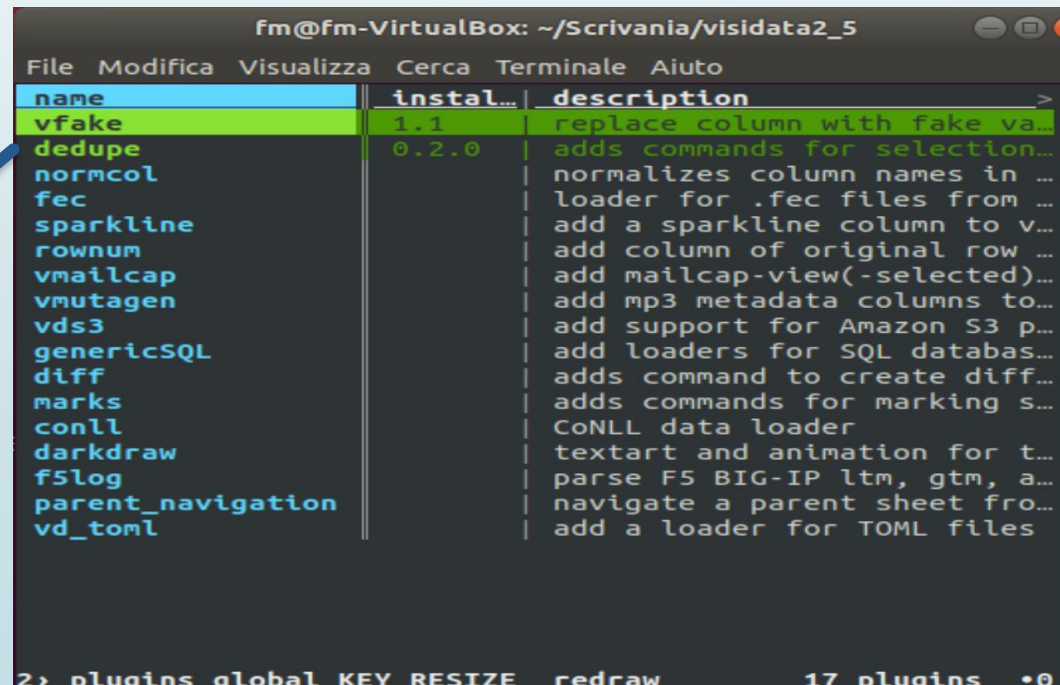
Importare moduli Python Visidata

- È possibile ampliare le funzionalità del tool importando al suo interno moduli Python. Per esempio, è possibile importare il modulo Python **re** relativo alle regular expression:

import re

- Oltre ai moduli Python, è possibile attivare/disattivare i **pluggings** già integrati oppure aggiungerne di nuovi.

Tasto 'a' per
attivare il plugin



```
fm@fm-VirtualBox: ~/Scrivania/visidata2_5
File Modifica Visualizza Cerca Terminale Aiuto
name instal... description >
vfake 1.1 replace column with fake va...
dedupe 0.2.0 adds commands for selection...
normcol normalizes column names in ...
fec loader for .fec files from ...
sparkline add a sparkline column to v...
rownum add column of original row ...
vmailcap add mailcap-view(-selected)...
vmutagen add mp3 metadata columns to...
vds3 add support for Amazon S3 p...
genericSQL add loaders for SQL databas...
diff adds command to create diff...
marks adds commands for marking s...
conll CoNLL data loader
darkdraw textart and animation for t...
f5log parse F5 BIG-IP ltm, gtm, a...
parent_navigation navigate a parent sheet fro...
vd_toml add a loader for TOML files

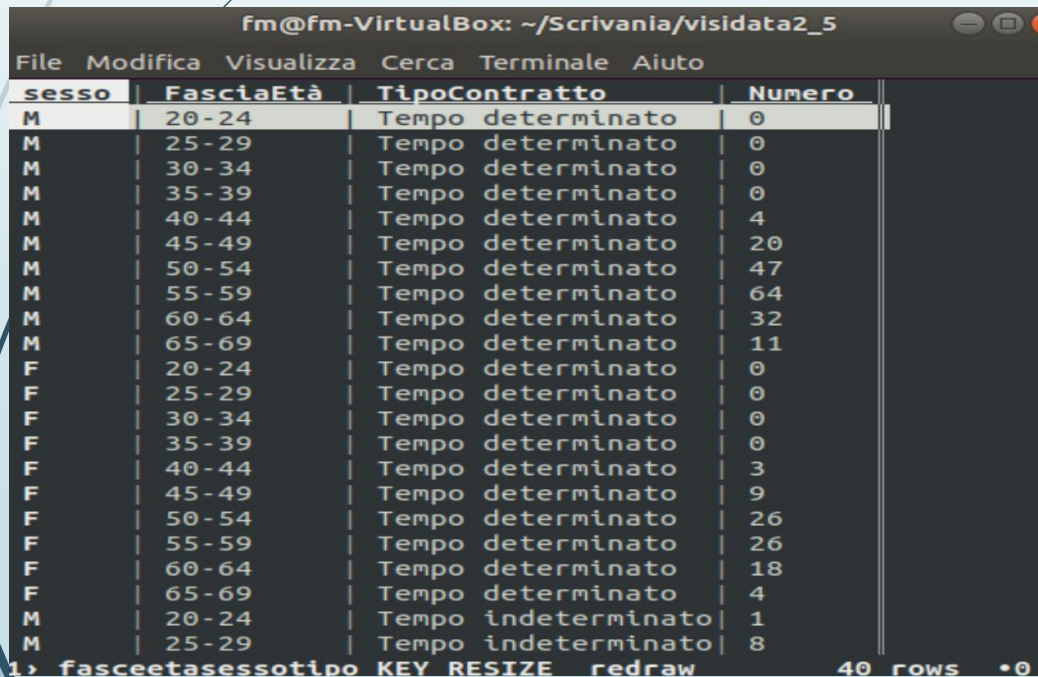
2> plugins global KEY RESIZE redraw 17 plugins •0
```

Tasto spazio +
open-plugins

VisiData

- Con la combinazione dei tasti **shift + f** è possibile calcolare la frequenza delle occorrenze di una colonna.

Posizionarsi sulla colonna di interesse

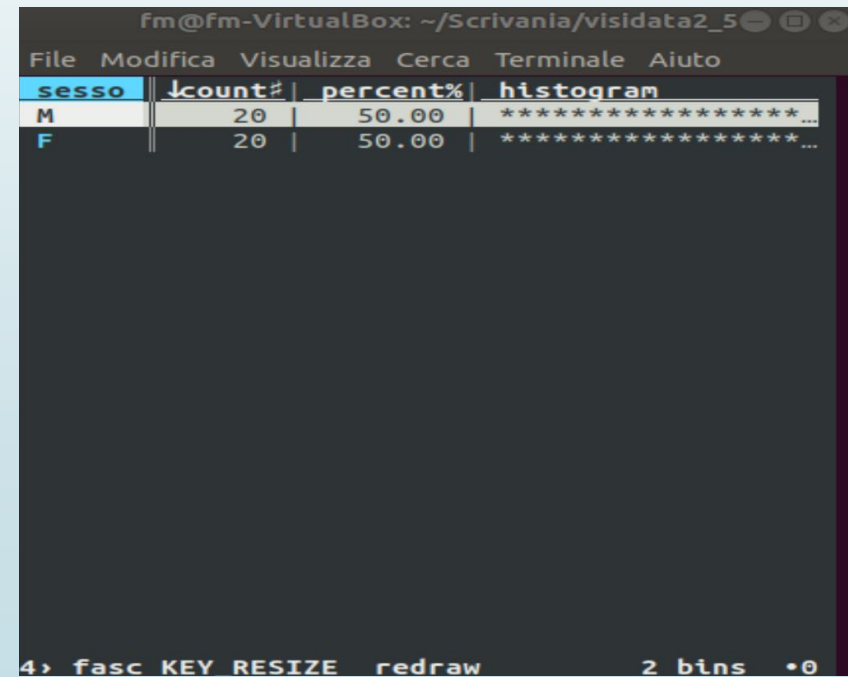


fm@fm-VirtualBox: ~/Scrivania/visidata2_5

sesso	FasciaEtà	TipoContratto	Numero
M	20-24	Tempo determinato	0
M	25-29	Tempo determinato	0
M	30-34	Tempo determinato	0
M	35-39	Tempo determinato	0
M	40-44	Tempo determinato	4
M	45-49	Tempo determinato	20
M	50-54	Tempo determinato	47
M	55-59	Tempo determinato	64
M	60-64	Tempo determinato	32
M	65-69	Tempo determinato	11
F	20-24	Tempo determinato	0
F	25-29	Tempo determinato	0
F	30-34	Tempo determinato	0
F	35-39	Tempo determinato	0
F	40-44	Tempo determinato	3
F	45-49	Tempo determinato	9
F	50-54	Tempo determinato	26
F	55-59	Tempo determinato	26
F	60-64	Tempo determinato	18
F	65-69	Tempo determinato	4
M	20-24	Tempo indeterminato	1
M	25-29	Tempo indeterminato	8

1> fasceetasessotipo KEY RESIZE redraw 40 rows •0

Risultato combinazione tasti **shift + f**



fm@fm-VirtualBox: ~/Scrivania/visidata2_5

sesso	count#	percent%	histogram
M	20	50.00	*****
F	20	50.00	*****

4> fasc KEY RESIZE redraw 2 bins •0

Selezione dati VisiData

- È possibile **selezionare/deselezionare** in modo **parziale** o **totale** le righe sulle quali svolgere operazioni. I tasti principali sono:
 - **s** : seleziona la riga corrente
 - **u** : deseleziona la riga corrente
 - **t** : inverte la selezione della riga corrente
 - **gs** : seleziona tutte le righe
 - **gu** : deseleziona tutte le righe
 - **gt** : inverte tutte di tutte le righe

In **arancione** le
righe selezionate

sesso	FasciaEtà	TipoContratto	Numero
M	20-24	Tempo determinato	0
M	25-29	Tempo determinato	0
M	30-34	Tempo determinato	0
M	35-39	Tempo determinato	0
M	40-44	Tempo determinato	4
M	45-49	Tempo determinato	20
M	50-54	Tempo determinato	47
M	55-59	Tempo determinato	64
M	60-64	Tempo determinato	32
M	65-69	Tempo determinato	11
F	20-24	Tempo determinato	0
F	25-29	Tempo determinato	0
F	30-34	Tempo determinato	0
F	35-39	Tempo determinato	0
F	40-44	Tempo determinato	3
F	45-49	Tempo determinato	9
F	50-54	Tempo determinato	26
F	55-59	Tempo determinato	26
F	60-64	Tempo determinato	18
F	65-69	Tempo determinato	4
M	20-24	Tempo indeterminato	1
M	25-29	Tempo indeterminato	8

1) fasceetasessotipo | KEY UP go-up 40 rows •5

Righe selezionate **5**

Selezione dati VisiData

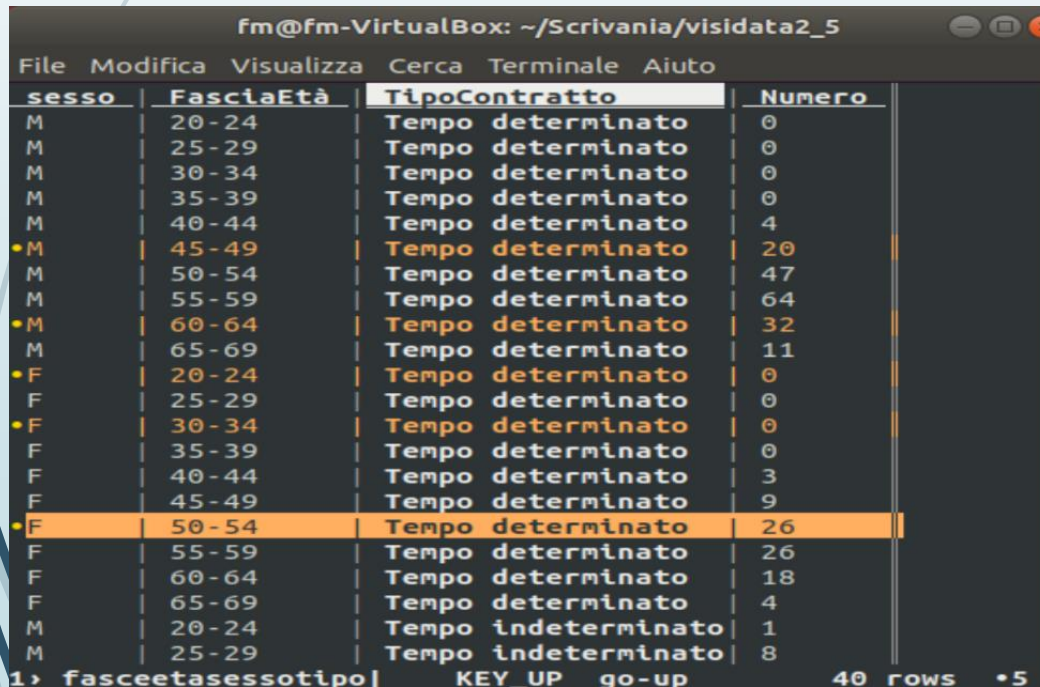
- È possibile **selezionare/deselezionare** tramite espressioni regolari:
 - **| + termine da ricercare**: seleziona tutte le righe nelle quali è verificata la corrispondenza per la colonna corrente
 - **\ + termine da ricercare**: deseleziona tutte le righe nelle quali è verificata la corrispondenza per la colonna corrente
- È possibile **selezionare/deselezionare** tramite un'espressione Python:
 - **z|**: seleziona tutte le righe nelle quali è valida l'espressione
 - **z**: deseleziona tutte le righe nelle quali è valida l'espressione
- Dopo aver importato il modulo Python **re** è possibile selezionare o deselezionare tramite espressioni regolari tramite il metodo **search**:
 - **re.search('F', sesso) or re.search(26, Numero)**

Selezione dati VisiData

- Una volta selezionate le righe di interesse è possibile creare un nuovo foglio tramite la combinazione dei tasti **shift + 2**

In **arancione** le
righe selezionate

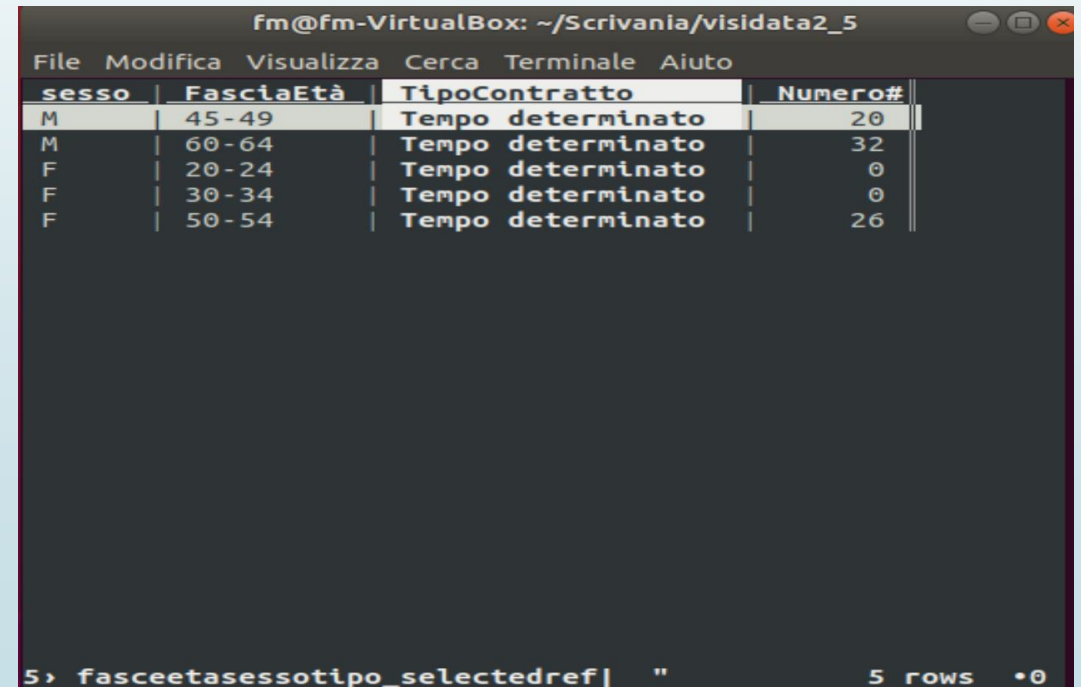
Creazione e apertura
nuovo foglio



fm@fm-VirtualBox: ~/Scrivania/visidata2_5

sesso	FasciaEtà	TipoContratto	Numero
M	20-24	Tempo determinato	0
M	25-29	Tempo determinato	0
M	30-34	Tempo determinato	0
M	35-39	Tempo determinato	0
M	40-44	Tempo determinato	4
M	45-49	Tempo determinato	20
M	50-54	Tempo determinato	47
M	55-59	Tempo determinato	64
M	60-64	Tempo determinato	32
M	65-69	Tempo determinato	11
F	20-24	Tempo determinato	0
F	25-29	Tempo determinato	0
F	30-34	Tempo determinato	0
F	35-39	Tempo determinato	0
F	40-44	Tempo determinato	3
F	45-49	Tempo determinato	9
F	50-54	Tempo determinato	26
F	55-59	Tempo determinato	26
F	60-64	Tempo determinato	18
F	65-69	Tempo determinato	4
M	20-24	Tempo indeterminato	1
M	25-29	Tempo indeterminato	8

1> fasceetasessotipo| KEY_UP go-up 40 rows •5



fm@fm-VirtualBox: ~/Scrivania/visidata2_5

sesso	FasciaEtà	TipoContratto	Numero#
M	45-49	Tempo determinato	20
M	60-64	Tempo determinato	32
F	20-24	Tempo determinato	0
F	30-34	Tempo determinato	0
F	50-54	Tempo determinato	26

5> fasceetasessotipo_selectedref| " 5 rows •0

Tipo di dato in VisiData

- Di default VisiData considera tutti i dati come **stringa** o testo
- È possibile scegliere il **tipo** di campo tra 5 tipi possibili:
 - **#** : definisce il campo della colonna corrente come **intero**
 - **%** : definisce il campo della colonna corrente come **float** o decimale
 - **\$** : definisce il campo della colonna corrente come **valuta** o moneta
 - **@** : definisce il campo della colonna corrente come **date**
 - **~** : definisce il campo della colonna corrente come **testo**

fm@fm-VirtualBox: ~/Scrivania/visidata2_5

File	Modifica	Visualizza	Cerca	Terminale	Aiuto
Sesso	FasciaEtà	TipoContratto	Numero#		
M	45-49	Tempo determinato	20		
M	60-64	Tempo determinato	32		
F	20-24	Tempo determinato	0		
F	30-34	Tempo determinato	0		
F	50-54	Tempo determinato	26		

La colonna **Numero** è impostata come **intero** (come indicato dal #)

Tipo di dato in VisiData

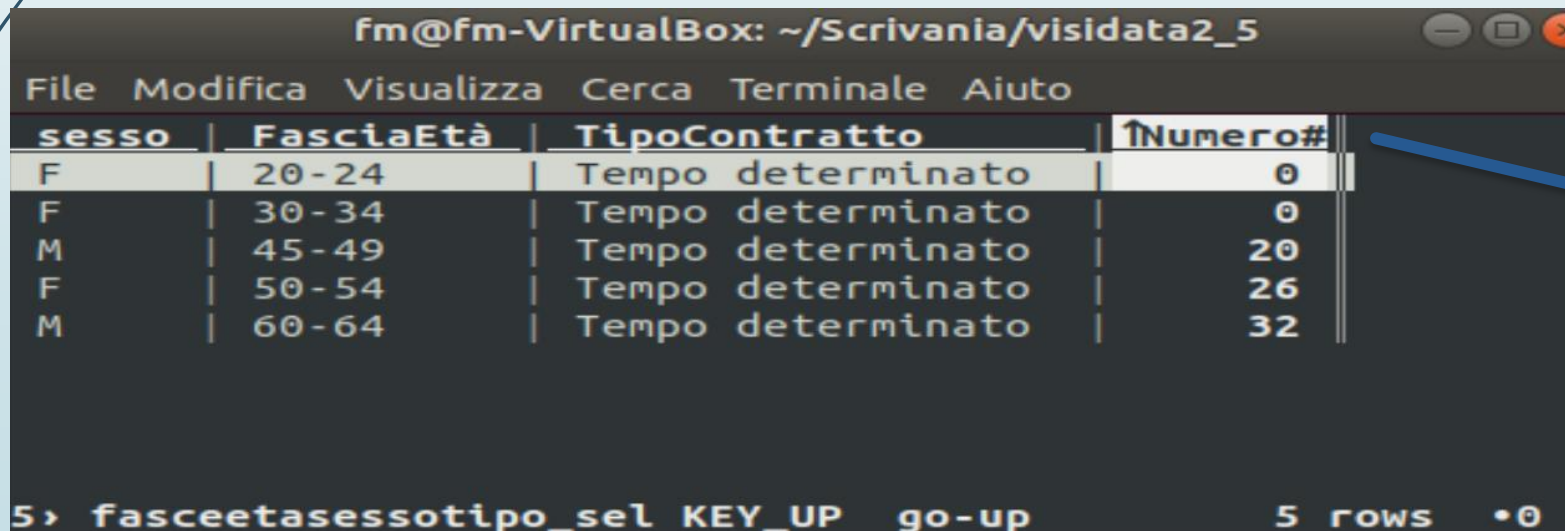
- Impostare il tipo di dato corretto è fondamentale se si vuole operare sui dati e svolgere operazioni su di essi
- '1' + '2' se trattato come testo non restituisce il valore aritmeticamente atteso, ovvero 3. Viceversa, se impostato come **tipo numero**, restituirà il valore aritmeticamente atteso, ovvero 3
- Se interpretato come testo il valore '9' è maggiore del valore '100.000.000.000'

Impostare correttamente il tipo di dato consente di svolgere operazioni fondamentali su di essi e avere risultati coerenti!

- Per il tipo **date** è possibile specificare il formato data personalizzato:
 - 1) combinazioni dei tasti **z** + **@**
 - 2) inserire il formato date desiderato
 - 3) premere **invio**

Ordinare le righe in VisiData

- È possibile ordinare i dati a seconda dei valori di una colonna in modo crescente o decrescente:
 - [: ordinare i dati in modo **crescente**
 -] : ordinare i dati in modo **decrescente**



sesso	FasciaEtà	TipoContratto	Numero#
F	20-24	Tempo determinato	0
F	30-34	Tempo determinato	0
M	45-49	Tempo determinato	20
F	50-54	Tempo determinato	26
M	60-64	Tempo determinato	32

Righe ordinate in modo **crescente**

Operazioni sui dati in VisiData

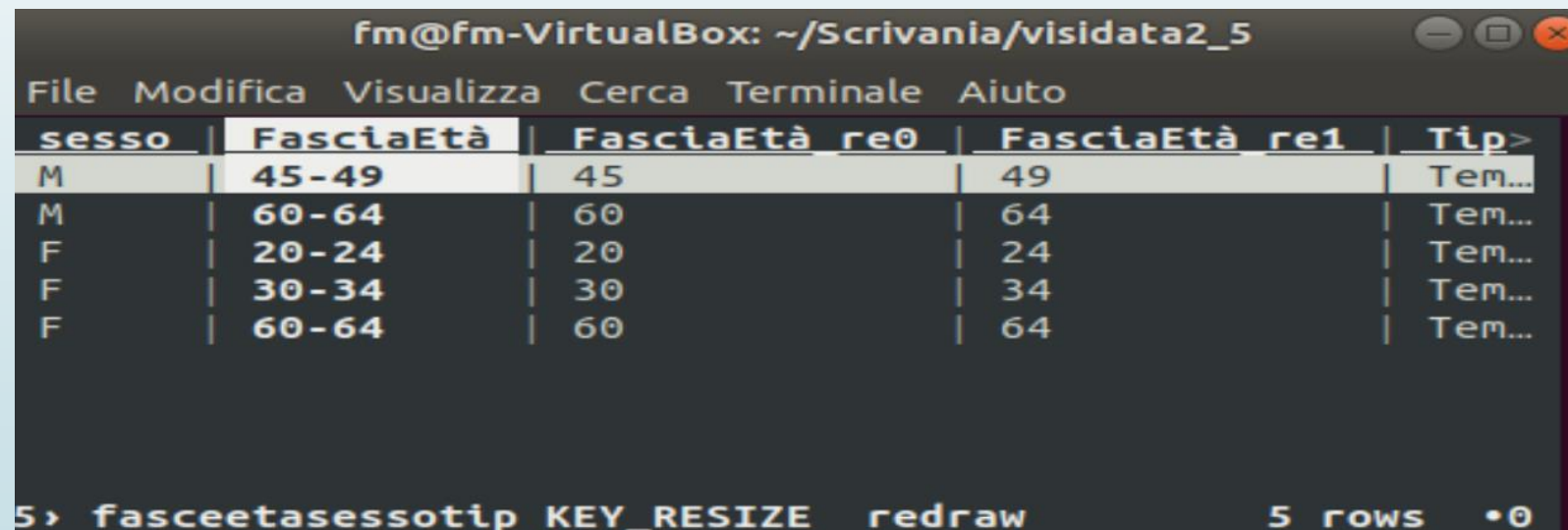
- È possibile svolgere dei calcoli su una colonna
- In totale vi sono **16 possibili operazioni**, tra cui somma, media, minimo, massimo....
- Per eseguire operazioni sui dati è necessario:
 - posizionarsi sulla colonna di interesse
 - Premere combinazione dei tasti **z + +**
 - Premere combinazione dei **CTRL + x** per aprire l'elenco delle operazioni possibili
 - Selezione dell'operazione desiderata

key	desc
min	minimum value
max	maximum value
avg	arithmetic mean of values
mean	arithmetic mean of values
median	median of values
mode	mode of values
sum	sum of values

Alcune possibili
operazioni

Creare colonne in VisiData

- Risulta spesso necessario **aggiungere** una nuova colonna alla tabella dei dati sulla quale si lavora.
- È possibile svolgere questa operazione principalmente in **3 modalità**:
 - Come risultato di un'operazione di **split**



sesso	FasciaEtà	FasciaEtà_re0	FasciaEtà_re1	Tip
M	45-49	45	49	Tem...
M	60-64	60	64	Tem...
F	20-24	20	24	Tem...
F	30-34	30	34	Tem...
F	60-64	60	64	Tem...

5> fasceetasessotip KEY_RESIZE redraw 5 rows *0

Creare colonne in VisiData

- Risulta spesso necessario **aggiungere** una nuova colonna alla tabella dei dati sulla quale si lavora.
- È possibile svolgere questa operazione principalmente in **3 modalità**:
 - Come risultato di una **espressione Python**
 - Come risultato di una **espressione regolare**

fm@fm-VirtualBox: ~/Scrivania/visidata2_5

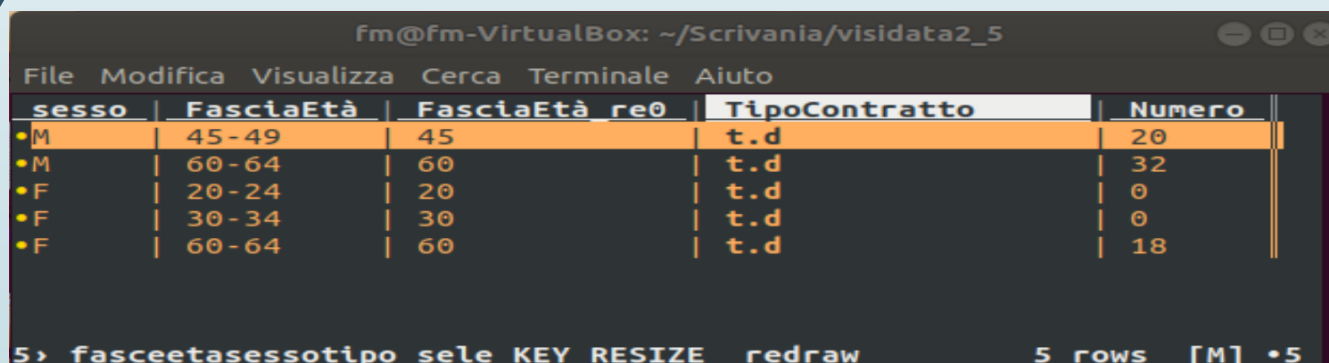
File	Modifica	Visualizza	Cerca	Terminale	Aiuto
Sesso	FasciaEtà	FasciaEtà re0	TipoContratto	Numero	
M	45-49	45	Tempo determinato	20	
M	60-64	60	Tempo determinato	32	
F	20-24	20	Tempo determinato	0	
F	30-34	30	Tempo determinato	0	
F	60-64	60	Tempo determinato	18	

5> fasceetasessotipo_selectedref| ; capture-col 5 rows •0

Premere ;
Inserire `^([0-9]{2})`
Viene creata la
terza colonna

Trova e sostituisci in VisiData

- È possibile effettuare l'operazione **trova e sostituisci** sui dati:
 - Selezionare tutte le righe (**gs**)
 - Digitare **gz***
 - Inserire **TrovaQuestaStringa / SostituisciConQuestaStringa**
 - Se non è inserito il carattere / l'operazione non andrà a buon fine



fm@fm-VirtualBox: ~/Scrivania/visidata2_5

sesso	FasciaEtà	FasciaEtà re0	TipoContratto	Numero
•M	45-49	45	t.d	20
•M	60-64	60	t.d	32
•F	20-24	20	t.d	0
•F	30-34	30	t.d	0
•F	60-64	60	t.d	18

5> fasceetasessotipo sele KEY RESIZE redraw 5 rows [M] •5

Digitare **gs**

Digitare **gz***

Digitare **Tempo determinato/t.d**

JOIN in VisiData

- Il **JOIN** è una clausola (da sql) che consente di combinare dati tra due o più tabelle in base a relazioni logiche tra le tabelle stesse.
- Per determinare quale colonna deve essere impostata come chiave di JOIN si utilizza il carattere !
- È possibile eseguire un JOIN tra due tabelle nel seguente modo:
 - Apro **prima tabella** e seleziono la **chiave di join** (premo ! sulla colonna)
 - Apro **seconda tabella** e seleziono la **chiave di join** (premo ! sulla colonna)
 - Digito il carattere &
 - Digito la combinazione di tasti **CTRL + x** la quale aprirà un menù dal quale è possibile selezionare il Join di interesse
 - Digito **invio**
- È possibile selezionare **una o più chiavi** di Join!

JOIN in VisiData

- In totale è possibile scegliere tra **7 Join** differenti

key	desc
inner	only rows which match keys on all sheets
outer	all rows from first selected sheet
full	all rows from all sheets (union)
diff	only rows NOT in all sheets
append	columns all sheets; extend with rows from all sheets
> choices	
KEY_UP go-up 7 choices •0	

Salva in VisiData

- Per salvare i dati di un foglio è necessario digitare **CTRL + s** e inserire il nome del file seguito dall'estensione
- È possibile salvare anche un **flusso di lavoro**, ovvero l'insieme ordinato delle operazioni eseguite sulla/e tabella/e così da poterle replicare successivamente.
- Per salvare un flusso di lavoro è necessario digitare la **CTRL + d** e inserire il nome del file seguito dall'estensione **.vd**
- Per eseguire un flusso di lavoro salvato è necessario digitare:
vd -p nomeFile.vd

Pandas

- Pandas è un pacchetto Python che consenti di gestire dati in modo **veloce** e **flessibile**
- Il nome **pandas** deriva dall'econometria, in particolare dalla combinazione dei termini **panel data**
- Ha l'obiettivo di diventare lo strumento **open source** principale e più potente per svolgere le operazioni di analisi e/o manipolazione dei dati



Installare pandas

- È possibile installare pandas tramite **Anaconda** o **Miniconda** poiché fa parte di questa distribuzione:
conda install pandas
- È possibile installare pandas tramite **pip** da **PyPI**:
pip install pandas

Tipi di dati in pandas

- ▶ Tramite pandas è possibile gestire due tipi di dati:
 - ▶ **Series**: rappresenta dati 1D, come le serie temporali. I dati sono memorizzati all'interno di un vettore **monodimensionale**.
 - ▶ **DataFrame**: rappresenta i dati come una **tabella** di oggetti eterogenei.

Series

- Gli elementi di una Series sono etichettati tramite un **index**
- **Creazione** di una **Series**:
 - **s = pd.Series({"a": 10, "z": 100, "g" : 45, "t" : 0})** passando un dizionario (chiave valore)
 - **s = pd.Series([4,3,7,8],index=[1,5,45,333])** Specificando sia dati che indici
 - **s = pd.Series([4,3,7,8])** Passando solo i dati (gli indici saranno sequenziali a partire da 0)

Operazioni sulle Series

- Le operazioni aritmetiche su una Series si applicano a tutti i suoi elementi (tale operazione viene detta broadcasting): **$s *= 2$**
- È possibile eseguire un'operazione su un solo elemento riferendosi ad esso per posizione o indice: **$s['c'] += 1$**
- i test logici si applicano a tutti gli elementi

Statistiche sulle Series

- Somma degli elementi di una Series: **series.sum()**
- Prodotto degli elementi di una Series: **product.sum()**
- Massimo (o minimo) tra gli elementi, ed indice corrispondente:
series.max()
series.argmax()

DataFrame

- Un DataFrame è una tabella di oggetti eterogenei. In pratica è l'equivalente bi-dimensionale di una Series.
- Un DataFrame ha indici sia per le righe che per le colonne:
 - index rappresenta le etichette delle righe
 - columns rappresenta le etichette delle colonne
- L'attributo shape descrive le dimensioni della tabella
- Ogni colonna di un DataFrame è una Series

Creazione di DataFrame

- Ci sono molti modi per creare un DataFrame. I più semplici sono:
 - Da un dizionario di liste

```
d = {'a' : [5,4,2,1,3], 'b' : [6,-1,1,9,1.8]}  
df = pd.DataFrame(d, index=['g1','g2','g3','g4','g5'])
```
 - Da un file in formato tabellare (e.g. CSV)

```
df = pd.read_csv('breast_cancer.txt', delimiter='\t')
```

Estrazione di righe e colonne

- Estrazione di colonne tramite indici di colonna

```
Col1 = df['colonna1']
```

```
Col = df['colonna1 ', 'colonna2']
```

- Una volta estratte le righe o le colonne è possibile operare su di esse tramite i metodi descritti in precedenza per le **Series**

Operazioni su DataFrame

- In generale, le operazioni sulle Series si applicano in modo analogo ai DataFrame
- Le statistiche si possono applicare a singole colonne o all'intera tabella
 - **df.mean()** : media sulle colonne del DataFrame
 - **df.std()** : deviazione standard sulle colonne del DataFrame

Raggruppare righe

- **GroupBy** di Pandas è una funzione potente e versatile. Consente di dividere i dati in gruppi separati sui quali eseguire calcoli per una migliore analisi
- GroupBy restituisce un oggetto **DataFrameGroupBy**
- Applicazione di funzioni ai DataFrameGroupBy
 - Aggregazione
 - Trasformazione
 - Filtraggio
 - Applicare la nostra funzione

Confronto

File 10.000.000 di righe (1.2 GB)	VisiData	Pandas	R	Excel
Tempo apertura	61s	30.5s	71s	FAIL
RAM occupata	14.4 Gb	7.1 Gb	6.2 Gb	FAIL



Grazie per l'attenzione