

# CaseStudies Live 2024-11-18

November 18, 2024

## 1 CaseStudies Live 2024-11-18 Q6

```
[1]: import pandas as pd
import matplotlib.pyplot as plt
```

### 1.1 Q6a

```
[52]: X = pd.read_csv("Data03.csv")
```

```
[5]: X.describe()
```

```
[5]:
```

	f1	f2
count	3840.000000	3840.000000
mean	-4.861466	-4.854412
std	1.604436	1.564089
min	-8.740101	-8.392300
25%	-5.704431	-5.649969
50%	-4.999363	-5.013370
75%	-4.317683	-4.317863
max	7.865602	7.463494

```
[6]: X.head()
```

```
[6]:
```

	timestamp	f1	f2
0	2023-07-20 06:00:00	-5.143580	-5.382025
1	2023-07-20 06:01:00	-4.605052	-5.503890
2	2023-07-20 06:02:00	-5.690262	-5.285184
3	2023-07-20 06:03:00	-3.917527	-6.951911
4	2023-07-20 06:04:00	-4.788862	-5.260651

```
[7]: X.timestamp
```

```
[7]:
```

0	2023-07-20 06:00:00
1	2023-07-20 06:01:00
2	2023-07-20 06:02:00
3	2023-07-20 06:03:00
4	2023-07-20 06:04:00

```

...
3835    2023-07-23 21:55:00
3836    2023-07-23 21:56:00
3837    2023-07-23 21:57:00
3838    2023-07-23 21:58:00
3839    2023-07-23 21:59:00
Name: timestamp, Length: 3840, dtype: object

```

```
[8]: X.timestamp.mean()
```

```

-----
TypeError                                Traceback (most recent call last)
Cell In[8], line 1
----> 1 X.timestamp.mean()

File /opt/conda/lib/python3.10/site-packages/pandas/core/series.py:6226, in
Series.mean(self, axis, skipna, numeric_only, **kwargs)
    6218 @doc(make_doc("mean", ndim=1))
    6219 def mean(
    6220     self,
    6221     (...),
    6222     **kwargs,
    6223 ):
-> 6226     return NDFrame.mean(self, axis, skipna, numeric_only, **kwargs)

File /opt/conda/lib/python3.10/site-packages/pandas/core/generic.py:11969, in
NDFrame.mean(self, axis, skipna, numeric_only, **kwargs)
    11962 def mean(
    11963     self,
    11964     axis: Axis | None = 0,
    11965     (...),
    11966     **kwargs,
    11967 ) -> Series | float:
> 11969     return self._stat_function(
    11970         "mean", nanops.nanmean, axis, skipna, numeric_only, **kwargs
    11971     )

File /opt/conda/lib/python3.10/site-packages/pandas/core/generic.py:11926, in
NDFrame._stat_function(self, name, func, axis, skipna, numeric_only, **kwargs)
    11922 nv.validate_func(name, (), kwargs)
    11924 validate_bool_kwarg(skipna, "skipna", none_allowed=False)
> 11926 return self._reduce(
    11927     func, name=name, axis=axis, skipna=skipna, numeric_only=numeric_onl
    11928 )

File /opt/conda/lib/python3.10/site-packages/pandas/core/series.py:6134, in
Series._reduce(self, op, name, axis, skipna, numeric_only, filter_type, **kwds)

```

```

6129     # GH#47500 - change to TypeError to match other methods
6130     raise TypeError(
6131         f"Series.{name} does not allow {kwd_name}={numeric_only} "
6132         "with non-numeric dtypes."
6133     )
-> 6134 return op(delegate, skipna=skipna, **kwds)

File /opt/conda/lib/python3.10/site-packages/pandas/core/nanops.py:147, in
↳ bottleneck_switch.__call__.<locals>.f(values, axis, skipna, **kwds)
    145         result = alt(values, axis=axis, skipna=skipna, **kwds)
    146     else:
-> 147         result = alt(values, axis=axis, skipna=skipna, **kwds)
    149 return result

File /opt/conda/lib/python3.10/site-packages/pandas/core/nanops.py:404, in
↳ _datetimelike_compat.<locals>.new_func(values, axis, skipna, mask, **kwargs)
    401 if datetimelike and mask is None:
    402     mask = isna(values)
-> 404 result = func(values, axis=axis, skipna=skipna, mask=mask, **kwargs)
    406 if datetimelike:
    407     result = _wrap_results(result, orig_values.dtype, fill_value=iNaT)

File /opt/conda/lib/python3.10/site-packages/pandas/core/nanops.py:720, in
↳ nanmean(values, axis, skipna, mask)
    718 count = _get_counts(values.shape, mask, axis, dtype=dtype_count)
    719 the_sum = values.sum(axis, dtype=dtype_sum)
-> 720 the_sum = _ensure_numeric(the_sum)
    722 if axis is not None and getattr(the_sum, "ndim", False):
    723     count = cast(np.ndarray, count)

File /opt/conda/lib/python3.10/site-packages/pandas/core/nanops.py:1693, in
↳ _ensure_numeric(x)
    1690 elif not (is_float(x) or is_integer(x) or is_complex(x)):
    1691     if isinstance(x, str):
    1692         # GH#44008, GH#36703 avoid casting e.g. strings to numeric
-> 1693     raise TypeError(f"Could not convert string '{x}' to numeric")
    1694     try:
    1695         x = float(x)

```



:-(( timestamp was not loaded as a timestamp!!!

Try to load it properly

```
[11]: X = pd.read_csv("Data03.csv", parse_dates=[0])
```

```
[12]: X.describe()
```

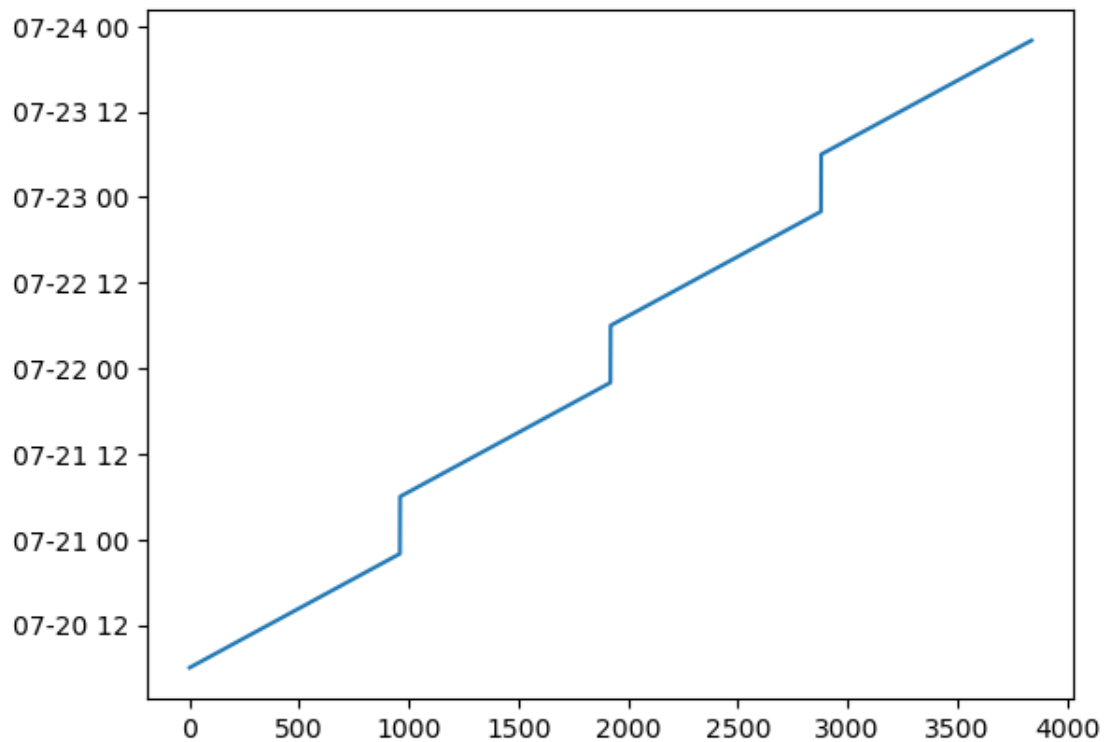
```
[12]:
```

	timestamp	f1	f2
count	3840	3840.000000	3840.000000
mean	2023-07-22 01:59:30	-4.861466	-4.854412
min	2023-07-20 06:00:00	-8.740101	-8.392300
25%	2023-07-21 03:59:45	-5.704431	-5.649969
50%	2023-07-22 01:59:30	-4.999363	-5.013370
75%	2023-07-22 23:59:15	-4.317683	-4.317863
max	2023-07-23 21:59:00	7.865602	7.463494
std	NaN	1.604436	1.564089

The timestamps are in the range 2023-07-20 06:00:00 to 2023-07-23 21:59:00. This is the timespan covered by the data.

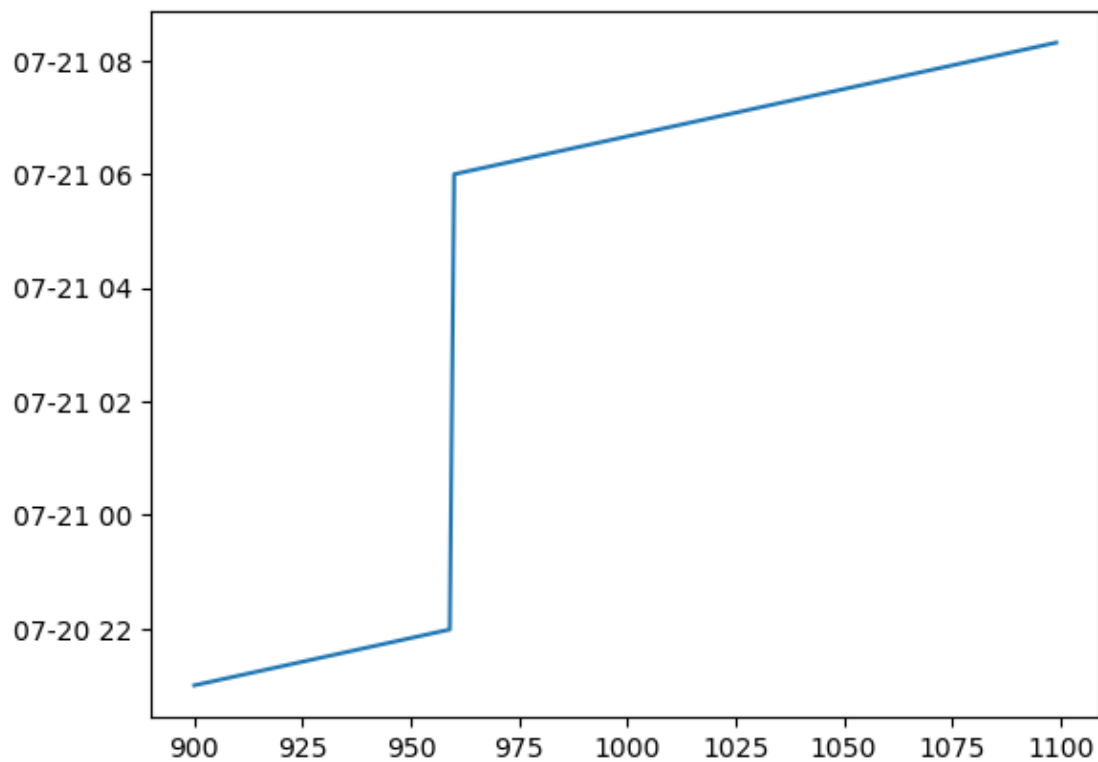
```
[13]: plt.plot(X.timestamp)
```

```
[13]: [<matplotlib.lines.Line2D at 0x7f81139916f0>]
```



```
[14]: plt.plot(X.timestamp[900:1100])
```

```
[14]: [<matplotlib.lines.Line2D at 0x7f8113996560>]
```



```
[16]: X.timestamp[950:970]
```

```
[16]: 950    2023-07-20 21:50:00
      951    2023-07-20 21:51:00
      952    2023-07-20 21:52:00
      953    2023-07-20 21:53:00
      954    2023-07-20 21:54:00
      955    2023-07-20 21:55:00
      956    2023-07-20 21:56:00
      957    2023-07-20 21:57:00
      958    2023-07-20 21:58:00
      959    2023-07-20 21:59:00
      960    2023-07-21 06:00:00
      961    2023-07-21 06:01:00
      962    2023-07-21 06:02:00
      963    2023-07-21 06:03:00
      964    2023-07-21 06:04:00
      965    2023-07-21 06:05:00
```

```

966    2023-07-21 06:06:00
967    2023-07-21 06:07:00
968    2023-07-21 06:08:00
969    2023-07-21 06:09:00
Name: timestamp, dtype: datetime64[ns]

```

```
[18]: X.timestamp[950:970].diff().max()
```

```
[18]: Timedelta('0 days 08:01:00')
```

```
[26]: X.timestamp.diff()[X.timestamp.diff()/pd.Timedelta(days=1) > 0.5/24]/pd.
      ↪Timedelta(days=1)
```

```

[26]: 960      0.334028
      1920      0.334028
      2880      0.334028
Name: timestamp, dtype: float64

```

Yes, we have three breaks of 8 hours each -> See plot.

```
[27]: X.timestamp[X.timestamp.diff()/pd.Timedelta(days=1) > 0.5/24]
```

```

[27]: 960      2023-07-21 06:00:00
      1920      2023-07-22 06:00:00
      2880      2023-07-23 06:00:00
Name: timestamp, dtype: datetime64[ns]

```

```
[31]: X.timestamp[X.timestamp.index[X.timestamp.diff()/pd.Timedelta(days=1) > 0.5/24]
      ↪- 1]
```

```

[31]: 959      2023-07-20 21:59:00
      1919      2023-07-21 21:59:00
      2879      2023-07-22 21:59:00
Name: timestamp, dtype: datetime64[ns]

```

## 1.2 Q6b

```
[32]: from sklearn.cluster import KMeans
```

```
[41]: clusterer = KMeans(n_clusters = 2)
```

```
[42]: X = X.drop(['timestamp'],axis = 1)
      X.head()
```

```
-----
KeyError
```

```
Cell In[42], line 1
```

```
Traceback (most recent call last)
```

```

----> 1 X = X.drop(['timestamp'],axis = 1)
      2 X.head()

```

File /opt/conda/lib/python3.10/site-packages/pandas/core/frame.py:5347, in `DataFrame.drop(self, labels, axis, index, columns, level, inplace, errors)`

```

5199 def drop(
5200     self,
5201     labels: IndexLabel | None = None,
5202     (...)
5203     errors: IgnoreRaise = "raise",
5204 ) -> DataFrame | None:
5210     """
5211     Drop specified labels from rows or columns.
5212     (...)
5345         weight 1.0      0.8
5346     """
-> 5347     return super().drop(
5348         labels=labels,
5349         axis=axis,
5350         index=index,
5351         columns=columns,
5352         level=level,
5353         inplace=inplace,
5354         errors=errors,
5355     )

```

File /opt/conda/lib/python3.10/site-packages/pandas/core/generic.py:4711, in `NDFrame.drop(self, labels, axis, index, columns, level, inplace, errors)`

```

4709 for axis, labels in axes.items():
4710     if labels is not None:
-> 4711         obj = obj._drop_axis(labels, axis, level=level, errors=errors)
4713 if inplace:
4714     self._update_inplace(obj)

```

File /opt/conda/lib/python3.10/site-packages/pandas/core/generic.py:4753, in `NDFrame._drop_axis(self, labels, axis, level, errors, only_slice)`

```

4751     new_axis = axis.drop(labels, level=level, errors=errors)
4752     else:
-> 4753     new_axis = axis.drop(labels, errors=errors)
4754     indexer = axis.get_indexer(new_axis)
4756 # Case for non-unique axis
4757 else:

```

File /opt/conda/lib/python3.10/site-packages/pandas/core/indexes/base.py:6992, in `Index.drop(self, labels, errors)`

```

6990 if mask.any():
6991     if errors != "ignore":

```



```
-> 6992         raise KeyError(f"{labels[mask].tolist()} not found in axis")
    6993     indexer = indexer[~mask]
    6994     return self.delete(indexer)
```

```
KeyError: "['timestamp'] not found in axis"
```

```
[43]: clusterer.fit(X)
      y = clusterer.predict(X)
```

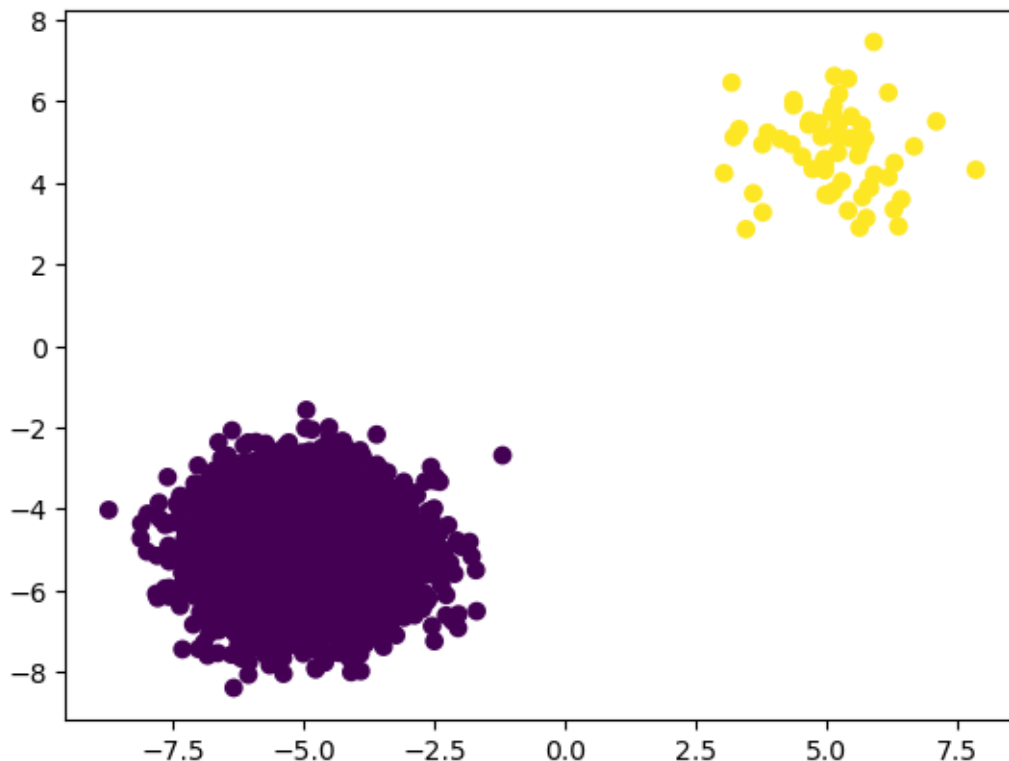
```
/opt/conda/lib/python3.10/site-packages/sklearn/cluster/_kmeans.py:1412:
FutureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  super()._check_params_vs_input(X, default_n_init=10)
```

```
[44]: y
```

```
[44]: array([0, 0, 0, ..., 0, 0, 0], dtype=int32)
```

```
[45]: plt.scatter(X.f1,X.f2,c=y)
```

```
[45]: <matplotlib.collections.PathCollection at 0x7f8111658a90>
```



A scatter plot indicates two distinct clusters: one at the bottom left, one at the top right ==> we look for two clusters

```
[49]: y = pd.Series(y)
      y.value_counts()
```

```
[49]: 0    3780
      1     60
      Name: count, dtype: int64
```

```
[53]: X.timestamp[y == 1]
```

```
[53]: 193    2023-07-20 09:13:00
      198    2023-07-20 09:18:00
      246    2023-07-20 10:06:00
      327    2023-07-20 11:27:00
      332    2023-07-20 11:32:00
      415    2023-07-20 12:55:00
      646    2023-07-20 16:46:00
      794    2023-07-20 19:14:00
      807    2023-07-20 19:27:00
      821    2023-07-20 19:41:00
      928    2023-07-20 21:28:00
      962    2023-07-21 06:02:00
      968    2023-07-21 06:08:00
      981    2023-07-21 06:21:00
      1100   2023-07-21 08:20:00
      1144   2023-07-21 09:04:00
      1239   2023-07-21 10:39:00
      1338   2023-07-21 12:18:00
      1379   2023-07-21 12:59:00
      1385   2023-07-21 13:05:00
      1470   2023-07-21 14:30:00
      1518   2023-07-21 15:18:00
      1572   2023-07-21 16:12:00
      1716   2023-07-21 18:36:00
      1741   2023-07-21 19:01:00
      1872   2023-07-21 21:12:00
      1876   2023-07-21 21:16:00
      1913   2023-07-21 21:53:00
      2015   2023-07-22 07:35:00
      2019   2023-07-22 07:39:00
      2087   2023-07-22 08:47:00
      2320   2023-07-22 12:40:00
      2379   2023-07-22 13:39:00
      2382   2023-07-22 13:42:00
      2432   2023-07-22 14:32:00
      2523   2023-07-22 16:03:00
```

```
2644    2023-07-22 18:04:00
2715    2023-07-22 19:15:00
2716    2023-07-22 19:16:00
2863    2023-07-22 21:43:00
2911    2023-07-23 06:31:00
2975    2023-07-23 07:35:00
3002    2023-07-23 08:02:00
3045    2023-07-23 08:45:00
3060    2023-07-23 09:00:00
3207    2023-07-23 11:27:00
3224    2023-07-23 11:44:00
3250    2023-07-23 12:10:00
3287    2023-07-23 12:47:00
3431    2023-07-23 15:11:00
3509    2023-07-23 16:29:00
3539    2023-07-23 16:59:00
3553    2023-07-23 17:13:00
3559    2023-07-23 17:19:00
3622    2023-07-23 18:22:00
3625    2023-07-23 18:25:00
3728    2023-07-23 20:08:00
3759    2023-07-23 20:39:00
3761    2023-07-23 20:41:00
3794    2023-07-23 21:14:00
Name: timestamp, dtype: object
```

[ ]: