MATLAB EXPO

Cleaning and Preparing Time Series Data

Onomitra Ghosh, MathWorks

Heather Gorr, PhD, MathWorks

MathWorks

MathWorks
@MathWorks

Share the EXPO experience
#MATLABEXPO

Onomitra Ghosh

oghosh@mathworks.com

@linkedin.com/in/onomitra

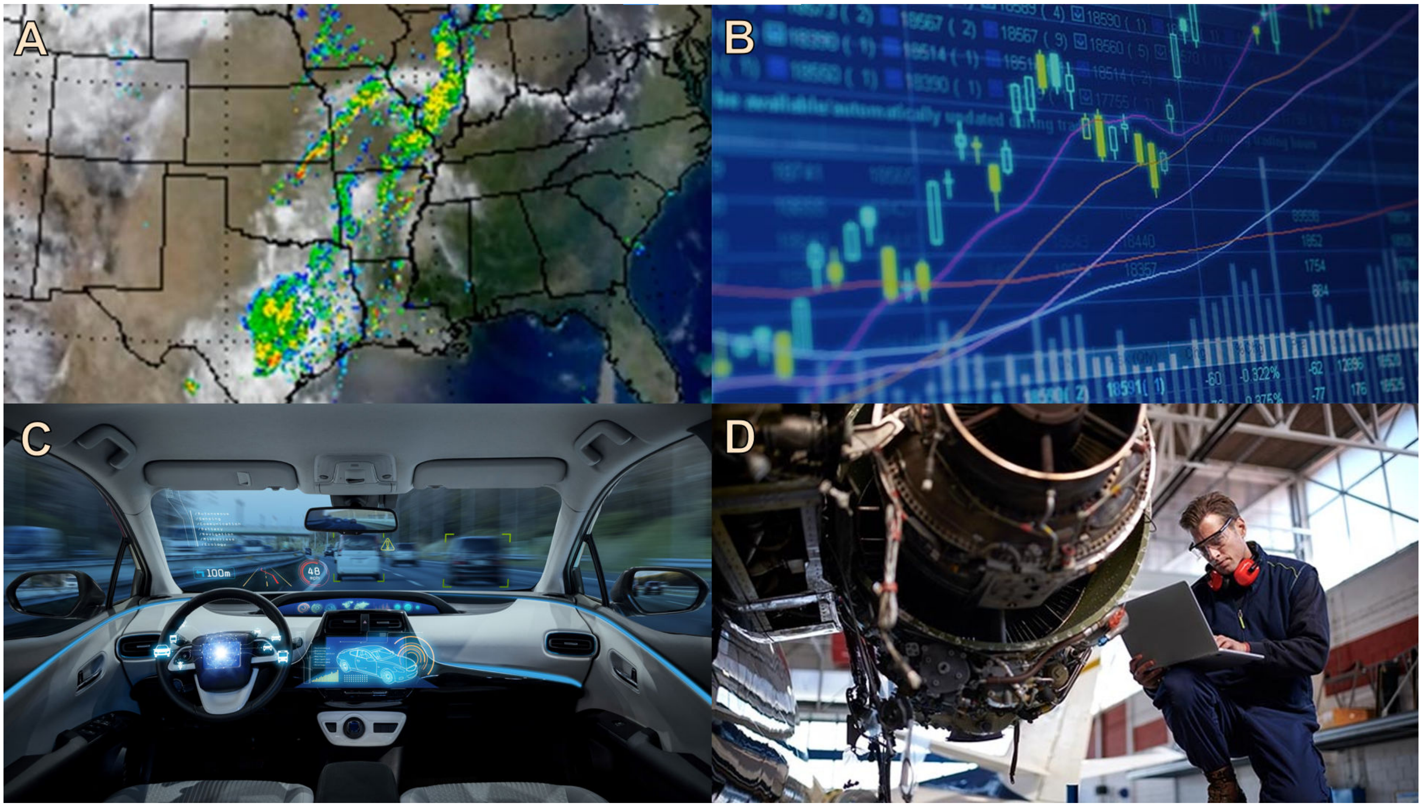
Heather Gorr, PhD

@HeatherGorr

@heather.codes

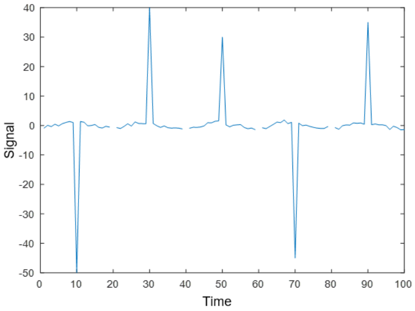
hgorr@mathworks.com

@linkedin.com/in/heather-gorr-phd

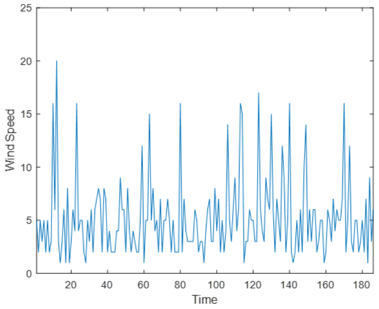


Which of these have time-series data?

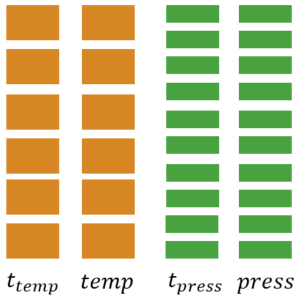
Real world time-series data is often messy



Missing data and outliers



Sensor noise



Measurements from different sensors

Background

Objective

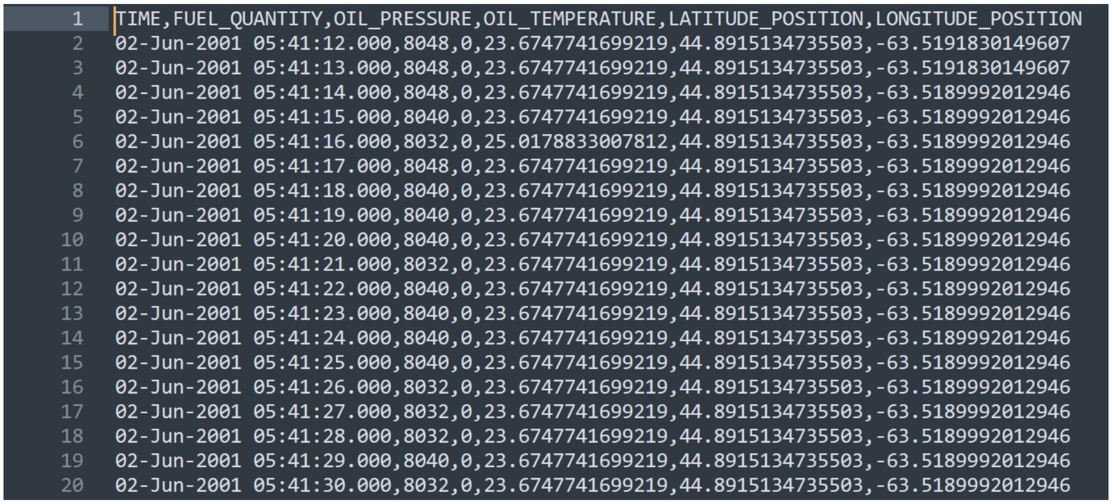
Predict True Air Speed (TAS) of flight based on other sensor measurements.

Dataset

The data set includes ~4600 different flights of plane with tail number 660. Each flight has measurement data from 2 different classes of sensors, one measured at 1 HZ and the other at 4 HZ.

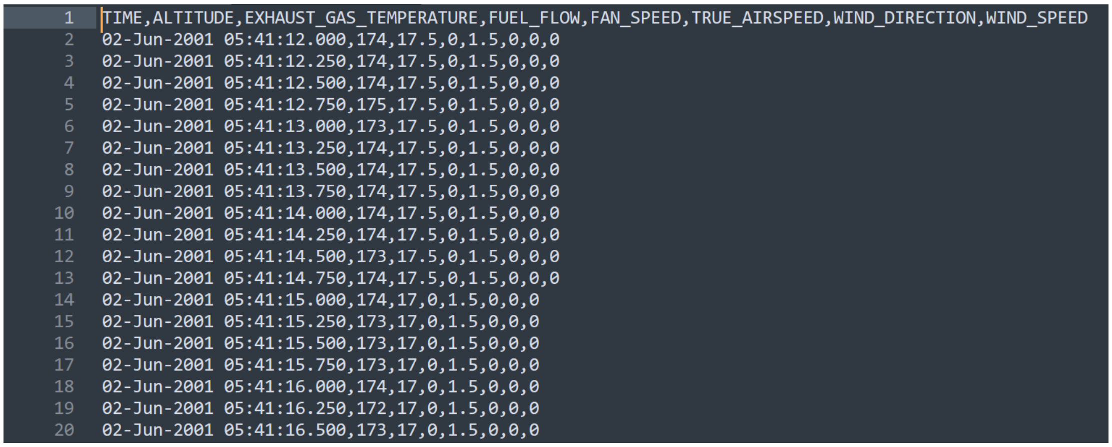
(This dataset is a small subset of the Flight Data from Dashlink which is modified for the purpose of this demo)

1 HZ



```
TIME,FUEL_QUANTITY,OIL_PRESSURE,OIL_TEMPERATURE,LATITUDE_POSITION,LONGITUDE_POSITION
02-Jun-2001 05:41:12.000,8048,0,23.6747741699219,44.8915134735503,-63.5191830149607
02-Jun-2001 05:41:13.000,8048,0,23.6747741699219,44.8915134735503,-63.5191830149607
02-Jun-2001 05:41:14.000,8048,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:15.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:16.000,8032,0,25.0178833007812,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:17.000,8048,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:18.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:19.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:20.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:21.000,8032,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:22.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:23.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:24.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:25.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:26.000,8032,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:27.000,8032,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:28.000,8032,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:29.000,8040,0,23.6747741699219,44.8915134735503,-63.5189992012946
02-Jun-2001 05:41:30.000,8032,0,23.6747741699219,44.8915134735503,-63.5189992012946
```

4 HZ



```
TIME,ALTITUDE,EXHAUST_GAS_TEMPERATURE,FUEL_FLOW,FAN_SPEED,TRUE_AIRSPEED,WIND_DIRECTION,WIND_SPEED
02-Jun-2001 05:41:12.000,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:12.250,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:12.500,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:12.750,175,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:13.000,173,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:13.250,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:13.500,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:13.750,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:14.000,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:14.250,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:14.500,173,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:14.750,174,17.5,0,1.5,0,0,0
02-Jun-2001 05:41:15.000,174,17,0,1.5,0,0,0
02-Jun-2001 05:41:15.250,173,17,0,1.5,0,0,0
02-Jun-2001 05:41:15.500,173,17,0,1.5,0,0,0
02-Jun-2001 05:41:15.750,173,17,0,1.5,0,0,0
02-Jun-2001 05:41:16.000,174,17,0,1.5,0,0,0
02-Jun-2001 05:41:16.250,172,17,0,1.5,0,0,0
02-Jun-2001 05:41:16.500,173,17,0,1.5,0,0,0
```

STEP 1. Import data into a timetable

Timetable is a special type of table for working with time series data.

```
t1HZ = readtimetable("flightData1HZ.csv")
```

```
t1HZ = 11828×5 timetable
```

...

	Time	FuelQuantity	OilPressure	OilTemperature
1	02-Jun-2001 05:41:12.000	8048	0	23.6748

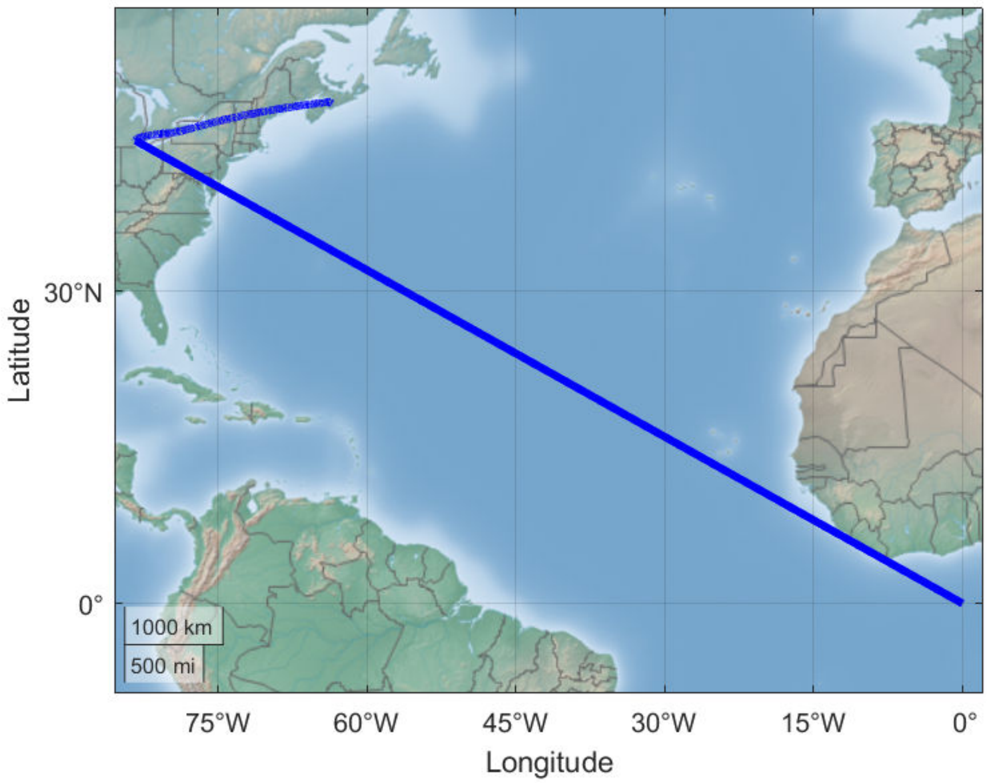
	Time	FuelQuantity	OilPressure	OilTemperature
2	02-Jun-2001 05:41:13.000	8048	0	23.6748
3	02-Jun-2001 05:41:14.000	8048	0	23.6748
4	02-Jun-2001 05:41:15.000	8040	0	23.6748
5	02-Jun-2001 05:41:16.000	8032	0	25.0179
6	02-Jun-2001 05:41:17.000	8048	0	23.6748
7	02-Jun-2001 05:41:18.000	8040	0	23.6748
8	02-Jun-2001 05:41:19.000	8040	0	23.6748
9	02-Jun-2001 05:41:20.000	8040	0	23.6748
10	02-Jun-2001 05:41:21.000	8032	0	23.6748
11	02-Jun-2001 05:41:22.000	8040	0	23.6748
12	02-Jun-2001 05:41:23.000	8040	0	23.6748
13	02-Jun-2001 05:41:24.000	8040	0	23.6748
14	02-Jun-2001 05:41:25.000	8040	0	23.6748

⋮

STEP 2. Visualize the flight path

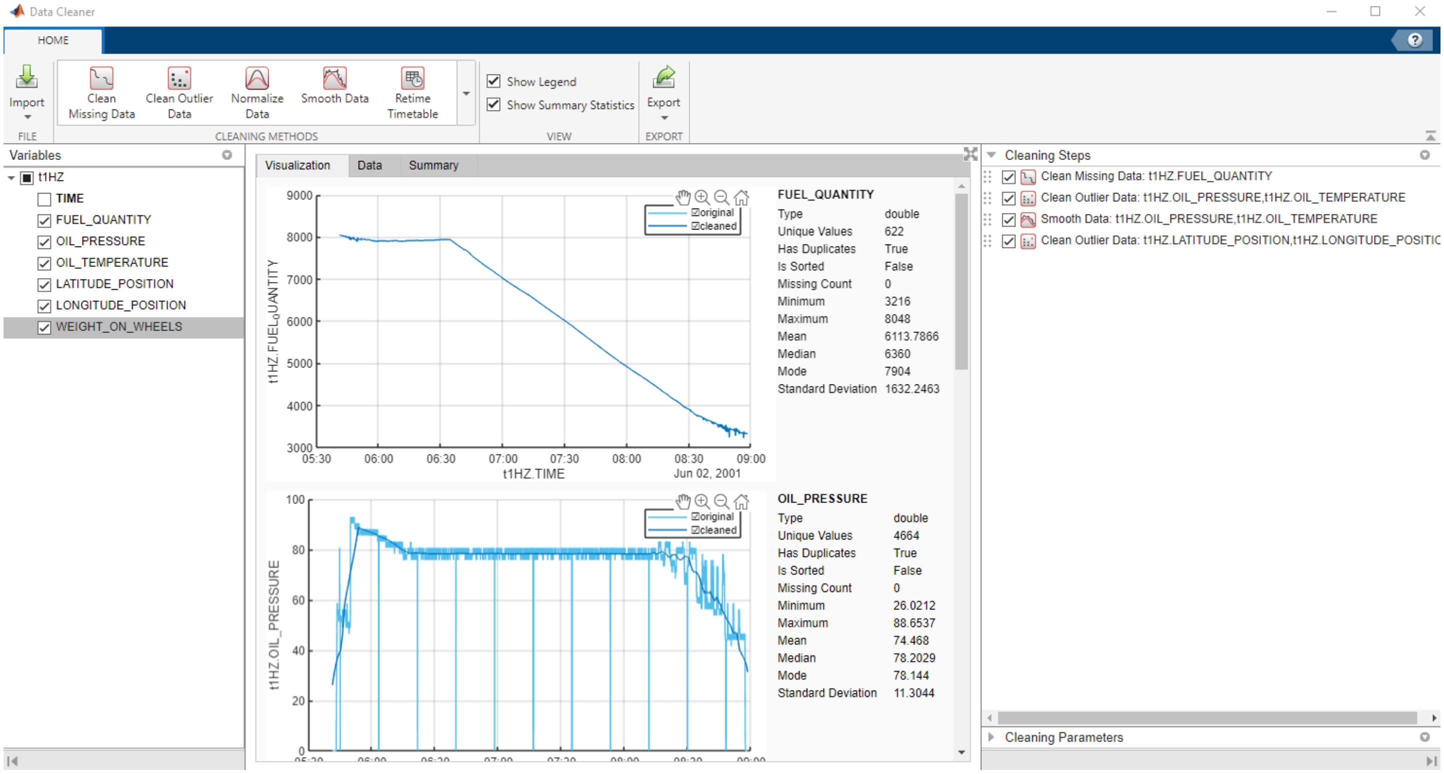
Let's check the starting and stopping locations.

```
geoplot(t1HZ.LatitudePosition,t1HZ.LongitudePosition,'Color','b','LineWidth',3);
geobasemap colorterrain
```



STEP 3. Clean data

Data Cleaner app (New in R2022a)



☑ Show Legend
☑ Show Summary Statistics

Variables

- ■ t1HZ
  - □ TIME
  - ☑ FUEL_QUANTITY
  - ☑ OIL_PRESSURE
  - ☑ OIL_TEMPERATURE
  - ☑ LATITUDE_POSITION
  - ☑ LONGITUDE_POSITION
  - ☑ WEIGHT_ON_WHEELS

Cleaning Steps

- ☑ Clean Missing Data: t1HZ.FUEL_QUANTITY
- ☑ Clean Outlier Data: t1HZ.OIL_PRESSURE,t1HZ.OIL_TEMPERATURE
- ☑ Smooth Data: t1HZ.OIL_PRESSURE,t1HZ.OIL_TEMPERATURE
- ☑ Clean Outlier Data: t1HZ.LATITUDE_POSITION,t1HZ.LONGITUDE_POSITIC

FUEL_QUANTITY

Type	double
Unique Values	622
Has Duplicates	True
Is Sorted	False
Missing Count	0
Minimum	3216
Maximum	8048
Mean	6113.7866
Median	6360
Mode	7904
Standard Deviation	1632.2463

OIL_PRESSURE

Type	double
Unique Values	4664
Has Duplicates	True
Is Sorted	False
Missing Count	0
Minimum	26.0212
Maximum	88.6537
Mean	74.468
Median	78.2029
Mode	78.144
Standard Deviation	11.3044

```
dataCleaner
```

Instead of just cleaning the current data, let's generate a MATLAB function that can be used for other 1HZ data files

```
t1HZ = clean1HZData(t1HZ)
```

t1HZ = 11828×5 timetable

...

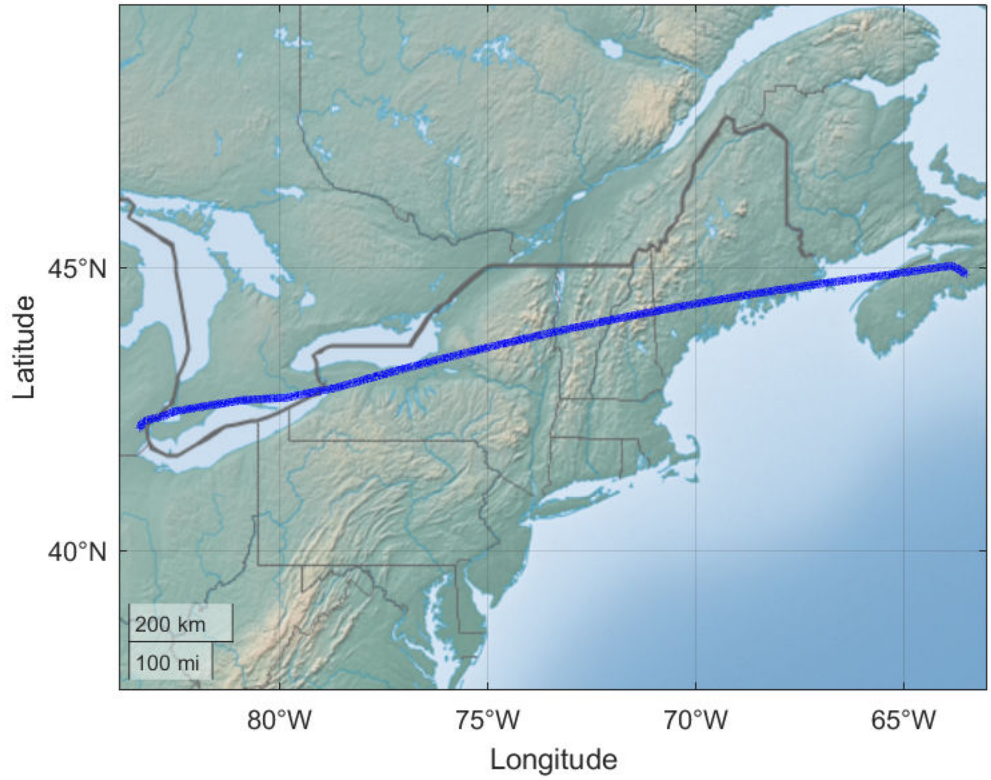
	Time	FuelQuantity	OilPressure	OilTemperature
1	02-Jun-2001 05:41:12.000	8048	0	23.6748
2	02-Jun-2001 05:41:13.000	8048	0	23.6748
3	02-Jun-2001 05:41:14.000	8048	0	23.6748
4	02-Jun-2001 05:41:15.000	8040	0	23.6748
5	02-Jun-2001 05:41:16.000	8032	0	23.6748
6	02-Jun-2001 05:41:17.000	8048	0	23.6748
7	02-Jun-2001 05:41:18.000	8040	0	23.6748
8	02-Jun-2001 05:41:19.000	8040	0	23.6748
9	02-Jun-2001 05:41:20.000	8040	0	23.6748
10	02-Jun-2001 05:41:21.000	8032	0	23.6748

	Time	FuelQuantity	OilPressure	OilTemperature
11	02-Jun-2001 05:41:22.000	8040	0	23.6748
12	02-Jun-2001 05:41:23.000	8040	0	23.6748
13	02-Jun-2001 05:41:24.000	8040	0	23.6748
14	02-Jun-2001 05:41:25.000	8040	0	23.6748

⋮

Verify flight path with cleaned data

```
geoplot(t1HZ.LatitudePosition,t1HZ.LongitudePosition,'Color','b','LineWidth',3)
geobasemap colorterrain
```



STEP 4. Load and clean the 4 HZ Data by going through the same steps

Perform the same importing and cleaning steps for the 4HZ data file

```
t4HZ = readtimetable("flightData4HZ.csv");
```

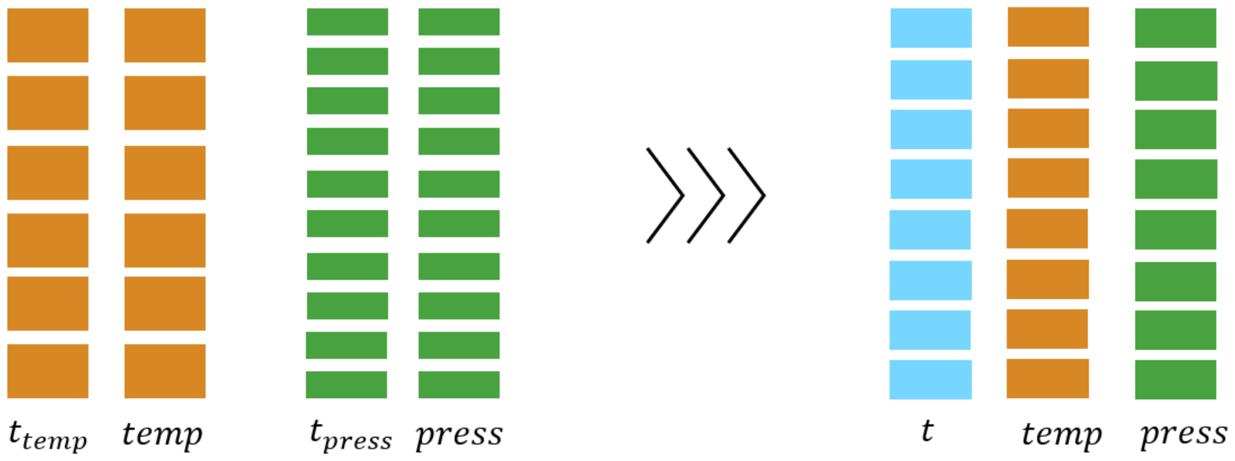
```
t4HZ = clean4HZData(t4HZ)
```

```
t4HZ = 47312×7 timetable
```

	Time	Altitude	ExhaustTemperature	FuelFlow	FanSpeed
1	02-Jun-2001 05:41:12.000	174	18.5736	0.0053	1.5000
2	02-Jun-2001 05:41:12.250	174	18.7119	0.0160	1.5000
3	02-Jun-2001 05:41:12.500	174	18.8555	0.0318	1.5000
4	02-Jun-2001 05:41:12.750	175	19.0046	0.0529	1.5000
5	02-Jun-2001 05:41:13.000	173	19.1589	0.0791	1.5000
6	02-Jun-2001 05:41:13.250	174	19.3186	0.1105	1.5000
7	02-Jun-2001 05:41:13.500	174	19.4834	0.1469	1.5000
8	02-Jun-2001 05:41:13.750	174	19.6535	0.1884	1.5000
9	02-Jun-2001 05:41:14.000	174	19.8287	0.2348	1.5000
10	02-Jun-2001 05:41:14.250	174	20.0090	0.2862	1.5000
11	02-Jun-2001 05:41:14.500	173	20.1944	0.3425	1.5000
12	02-Jun-2001 05:41:14.750	174	20.3848	0.4037	1.5000
13	02-Jun-2001 05:41:15.000	174	20.5802	0.4697	1.5000
14	02-Jun-2001 05:41:15.250	173	20.7805	0.5404	1.5000

⋮

STEP 5. Synchronize the two datasets



```
% Synchronize timetables
t = synchronize(t1HZ,t4HZ,"union","linear")
```

```
t = 47312×12 timetable
```

...

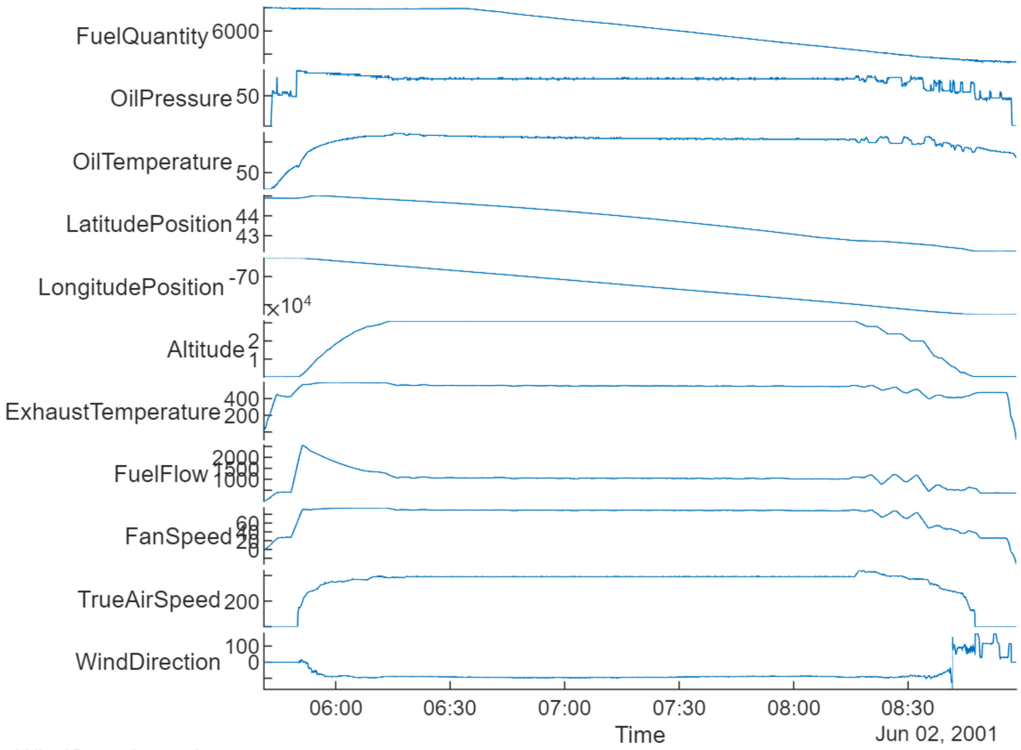
	Time	FuelQuantity	OilPressure	OilTemperature
1	02-Jun-2001 05:41:12.000	8048	0	23.6748
2	02-Jun-2001 05:41:12.250	8048	0	23.6748
3	02-Jun-2001 05:41:12.500	8048	0	23.6748
4	02-Jun-2001 05:41:12.750	8048	0	23.6748
5	02-Jun-2001 05:41:13.000	8048	0	23.6748
6	02-Jun-2001 05:41:13.250	8048	0	23.6748
7	02-Jun-2001 05:41:13.500	8048	0	23.6748
8	02-Jun-2001 05:41:13.750	8048	0	23.6748
9	02-Jun-2001 05:41:14.000	8048	0	23.6748
10	02-Jun-2001 05:41:14.250	8046	0	23.6748
11	02-Jun-2001 05:41:14.500	8044	0	23.6748
12	02-Jun-2001 05:41:14.750	8042	0	23.6748
13	02-Jun-2001 05:41:15.000	8040	0	23.6748
14	02-Jun-2001 05:41:15.250	8038	0	23.6748

⋮

STEP 6. Visualize timetables using stackedplot

stackedplot allows plotting multiple variables with common x-axis and exploring them in a synchronized way.

```
stackedplot(t)
```



WindSpeed not shown.

STEP 7. Finding and saving cruising data using ischange

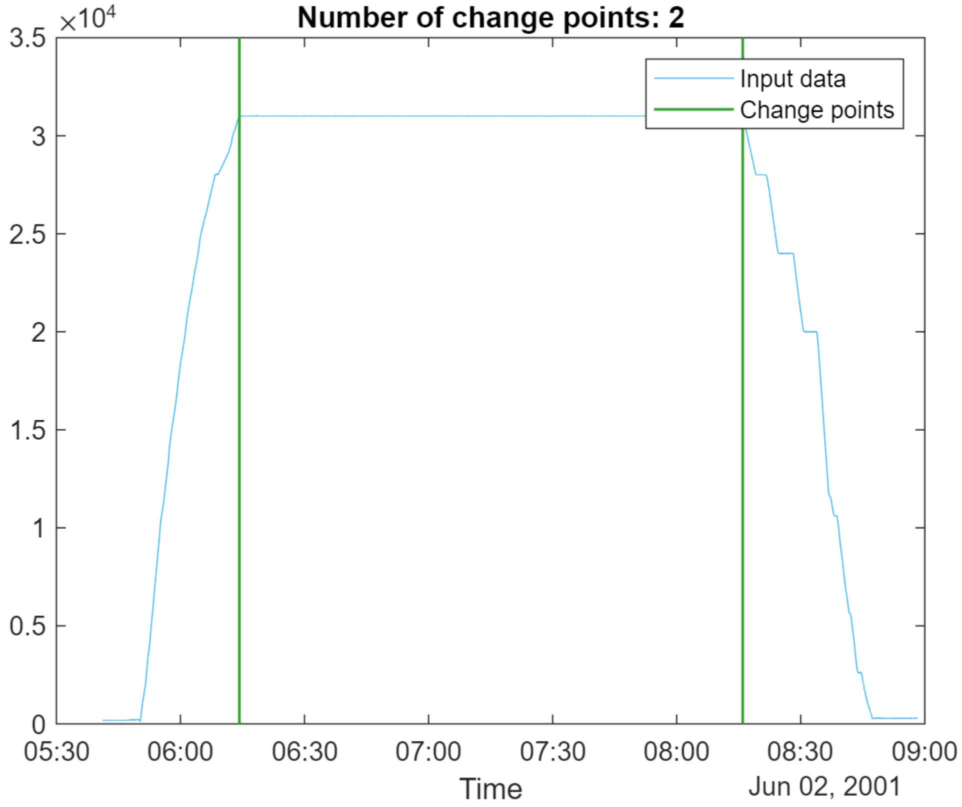
We can identify when the flight is cruising by checking when the altitude is not changing. So, we need to figure out when the cruise altitude starts and when it ends.

```
% Find change points
changeIndices = ischange(t.Altitude,"variance","MaxNumChanges",2,...
    "SamplePoints",t.Time);

% Display results
clf
plot(t.Time,t.Altitude,"Color",[77 190 238]/255,"DisplayName","Input data")
hold on

% Plot change points
x = repelem(t.Time(changeIndices),3);
y = repmat([ylim(gca) missing]',nnz(changeIndices),1);
plot(x,y,"Color",[51 160 44]/255,"LineWidth",1,"DisplayName","Change points")
title("Number of change points: " + nnz(changeIndices))

hold off
legend
xlabel("Time")
```



```
clear x y
```

Find the cruise start and end times

```
% Find the change points
changeIndices = find(changeIndices);
% Find cruise start and end times
cruiseStartTime = t.Time(changeIndices(1))
```

```
cruiseStartTime = datetime
   02-Jun-2001 06:14:15.250
```

```
cruiseEndTime = t.Time(changeIndices(2))
```

```
cruiseEndTime = datetime
   02-Jun-2001 08:15:58.750
```

Timetable allows slicing data based on a time range.

```
cruiseRange = timerange(cruiseStartTime,cruiseEndTime)
```

```
cruiseRange = 
    timetable timerange subscript:

        Select timetable rows with times in the half-open interval:
        [02-Jun-2001 06:14:15, 02-Jun-2001 08:15:58)
```

See Select Timetable Data by Row Time and Variable Type.

```
tcruise = t(cruiseRange,:)
```

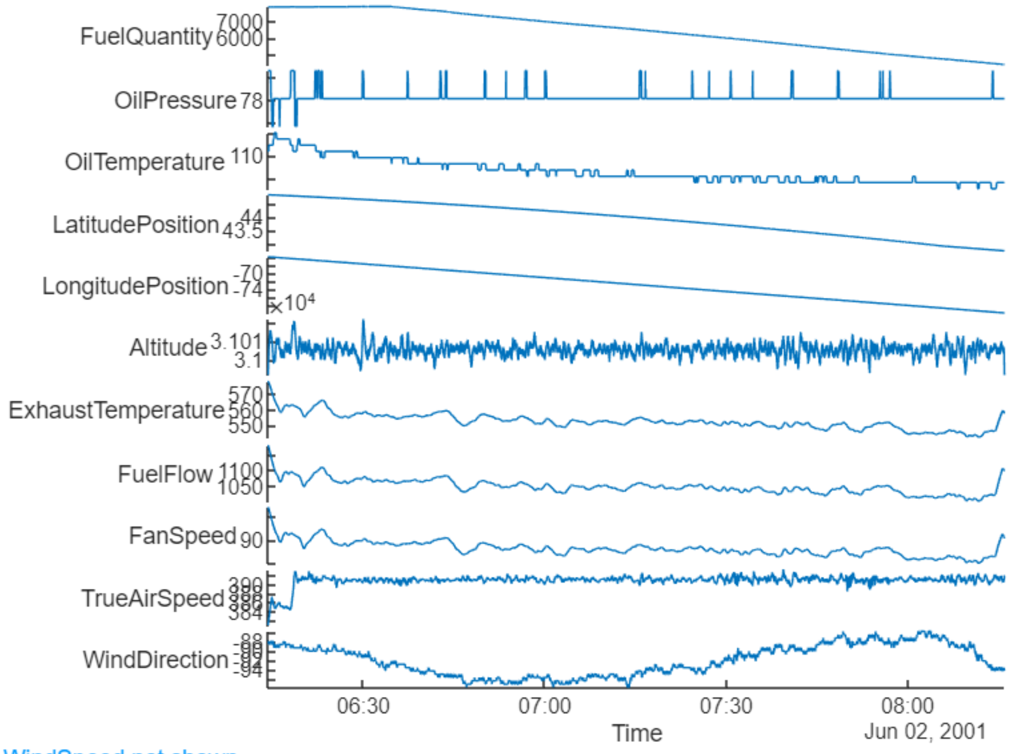
```
tcruise = 29214×12 timetable
```

...

	Time	FuelQuantity	OilPressure	OilTemperature
1	02-Jun-2001 06:14:15.250	7912	80.5860	110.9763
2	02-Jun-2001 06:14:15.500	7912	80.5860	110.9763
3	02-Jun-2001 06:14:15.750	7912	80.5860	110.9763
4	02-Jun-2001 06:14:16.000	7912	80.5860	110.9763
5	02-Jun-2001 06:14:16.250	7912	80.5860	110.9763
6	02-Jun-2001 06:14:16.500	7912	80.5860	110.9763
7	02-Jun-2001 06:14:16.750	7912	80.5860	110.9763
8	02-Jun-2001 06:14:17.000	7912	80.5860	110.9763
9	02-Jun-2001 06:14:17.250	7912	80.5860	110.9763
10	02-Jun-2001 06:14:17.500	7912	80.5860	110.9763
11	02-Jun-2001 06:14:17.750	7912	80.5860	110.9763
12	02-Jun-2001 06:14:18.000	7912	80.5860	110.9763
13	02-Jun-2001 06:14:18.250	7912	80.5860	110.9763
14	02-Jun-2001 06:14:18.500	7912	80.5860	110.9763

⋮

```
stackedplot(tcruise)
```

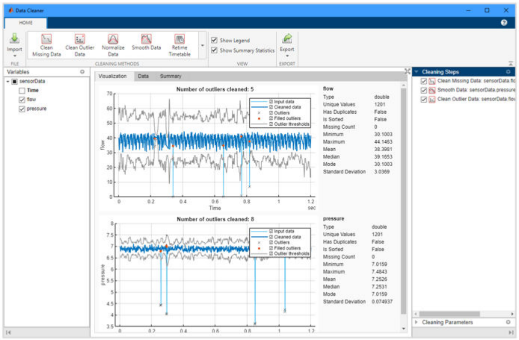


WindSpeed not shown.

Summary

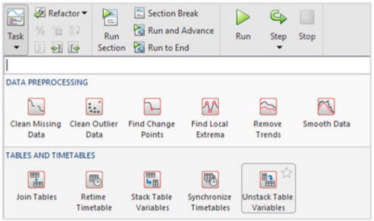
- MATLAB offers different low-code and code-based techniques for data cleaning and preparation
- Choose based on your familiarity with the data and the cleaning process

Low **Familiarity with data and cleaning process** **High**



flow	
Type	double
Unique Values	1201
Has Duplicates	False
Is Sorted	False
Missing Count	0
Minimum	30.1003
Maximum	44.1463
Mean	38.3981
Median	39.1653
Mode	30.1003
Standard Deviation	3.0369

pressure	
Type	double
Unique Values	1201
Has Duplicates	False
Is Sorted	False
Missing Count	0
Minimum	7.0159
Maximum	7.4843
Mean	7.2526
Median	7.2531
Mode	7.0159
Standard Deviation	0.074937



```
ismissing, rmmissing,
fillmissing,
standardizeMissing
isoutlier, rmoutliers,
filloutliers
smoothdata, movmean,
movmedian, movstd, movvar,
movsum
ischange, islocalmin,
islocalmax
join, innerjoin, outerjoin,
stack, unstack
retime, synchronize
normalize, rescale
```

Data Cleaner
R2022a

Live Editor Tasks

Command Line Functions

What kind of data cleaning challenges do you run into?