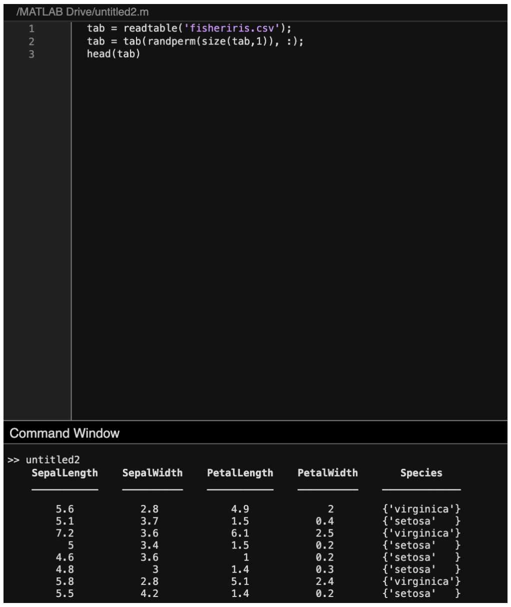
Trevor Moss, Alvaro Lorenzo, Ivan Marin Section C4

Deliverable C

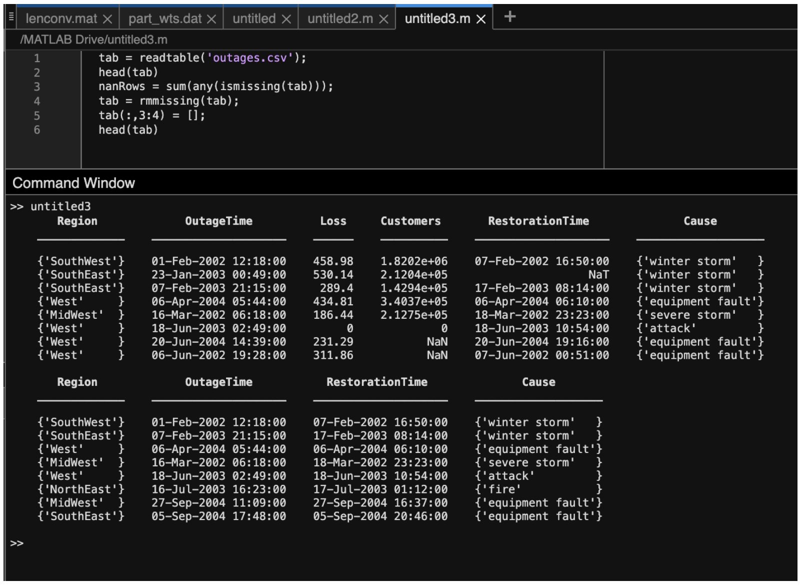
Steps 1 and 2

```
tab = readtable('fisheriris.csv');
tab = tab(randperm(size(tab,1)), :);
```



```
tab = readtable('fisheriris.csv');
tab = tab(randperm(size(tab,1)), :);
head(tab)
```

```
>> untitled2
    SepalLength    SepalWidth    PetalLength    PetalWidth      Species    
    ___________    __________    ___________    __________    _____________

        5.6           2.8            4.9             2        {'virginica'}
        5.1           3.7            1.5           0.4        {'setosa'   }
        7.2           3.6            6.1           2.5        {'virginica'}
          5           3.4            1.5           0.2        {'setosa'   }
        4.6           3.6              1           0.2        {'setosa'   }
        4.8             3            1.4           0.3        {'setosa'   }
        5.8           2.8            5.1           2.4        {'virginica'}
        5.5           4.2            1.4           0.2        {'setosa'   }
```

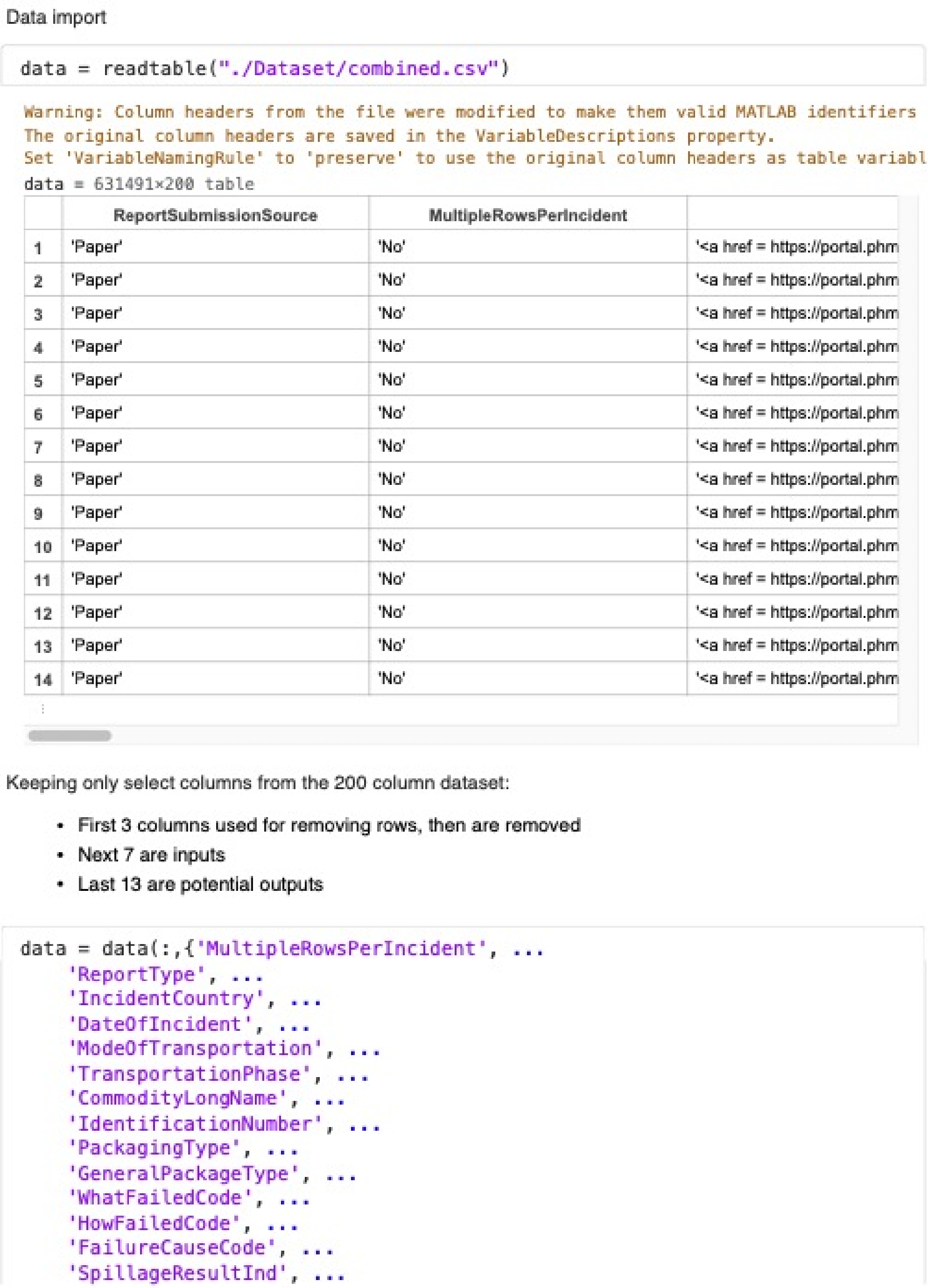


```
tab = readtable('outages.csv');
head(tab)
nanRows = sum(any(ismissing(tab)));
tab = rmmissing(tab);
tab(:,3:4) = [];
head(tab)
```

```
>> untitled3
      Region             OutageTime          Loss     Customers       RestorationTime             Cause       
    _____________    ____________________    ______    __________    ____________________    _________________

    {'SouthWest'}    01-Feb-2002 12:18:00    458.98    1.8202e+06    07-Feb-2002 16:50:00    {'winter storm'   }
    {'SouthEast'}    23-Jan-2003 00:49:00    530.14    2.1204e+05                     NaT    {'winter storm'   }
    {'SouthEast'}    07-Feb-2003 21:15:00     289.4    1.4294e+05    17-Feb-2003 08:14:00    {'winter storm'   }
    {'West'     }    06-Apr-2004 05:44:00    434.81    3.4037e+05    06-Apr-2004 06:10:00    {'equipment fault'}
    {'MidWest'  }    16-Mar-2002 06:18:00    186.44    2.1275e+05    18-Mar-2002 23:23:00    {'severe storm'   }
    {'West'     }    18-Jun-2003 02:49:00         0             0    18-Jun-2003 10:54:00    {'attack'         }
    {'West'     }    20-Jun-2004 14:39:00    231.29           NaN    20-Jun-2004 19:16:00    {'equipment fault'}
    {'West'     }    06-Jun-2002 19:28:00    311.86           NaN    07-Jun-2002 00:51:00    {'equipment fault'}

      Region             OutageTime            RestorationTime             Cause       
    _____________    ____________________    ____________________    _________________

    {'SouthWest'}    01-Feb-2002 12:18:00    07-Feb-2002 16:50:00    {'winter storm'   }
    {'SouthEast'}    07-Feb-2003 21:15:00    17-Feb-2003 08:14:00    {'winter storm'   }
    {'West'     }    06-Apr-2004 05:44:00    06-Apr-2004 06:10:00    {'equipment fault'}
    {'MidWest'  }    16-Mar-2002 06:18:00    18-Mar-2002 23:23:00    {'severe storm'   }
    {'West'     }    18-Jun-2003 02:49:00    18-Jun-2003 10:54:00    {'attack'         }
    {'NorthEast'}    16-Jul-2003 16:23:00    17-Jul-2003 01:12:00    {'fire'           }
    {'MidWest'  }    27-Sep-2004 11:09:00    27-Sep-2004 16:37:00    {'equipment fault'}
    {'SouthEast'}    05-Sep-2004 17:48:00    05-Sep-2004 20:46:00    {'equipment fault'}

>>
```

We decided to remove the Loss and Customer columns from the table because we thought it was best to use this data to focus on only the cause/location of the outages and the time it took to fix them. Loss and number of customers were irrelevant to that goal so we removed them.

Step 3
Our initial table had 200 columns of data, and many were unnecessary. We cut this down to 20 columns. Additionally, all missing values were removed. At the end of the cleaning, the dataset size went from 641491 to 300254 rows, with 87% of the reduction due to missing values.



Data import

```
data = readtable("./Dataset/combined.csv")
```

```
Warning: Column headers from the file were modified to make them valid MATLAB identifiers
The original column headers are saved in the VariableDescriptions property.
Set 'VariableNamingRule' to 'preserve' to use the original column headers as table variabl
data = 631491×200 table
```

	ReportSubmissionSource	MultipleRowsPerIncident	
1	'Paper'	'No'	'<a href = https://portal.phm
2	'Paper'	'No'	'<a href = https://portal.phm
3	'Paper'	'No'	'<a href = https://portal.phm
4	'Paper'	'No'	'<a href = https://portal.phm
5	'Paper'	'No'	'<a href = https://portal.phm
6	'Paper'	'No'	'<a href = https://portal.phm
7	'Paper'	'No'	'<a href = https://portal.phm
8	'Paper'	'No'	'<a href = https://portal.phm
9	'Paper'	'No'	'<a href = https://portal.phm
10	'Paper'	'No'	'<a href = https://portal.phm
11	'Paper'	'No'	'<a href = https://portal.phm
12	'Paper'	'No'	'<a href = https://portal.phm
13	'Paper'	'No'	'<a href = https://portal.phm
14	'Paper'	'No'	'<a href = https://portal.phm
⋮			

Keeping only select columns from the 200 column dataset:

- First 3 columns used for removing rows, then are removed
- Next 7 are inputs
- Last 13 are potential outputs

```
data = data(:,{'MultipleRowsPerIncident', ...
    'ReportType', ...
    'IncidentCountry', ...
    'DateOfIncident', ...
    'ModeOfTransportation', ...
    'TransportationPhase', ...
    'CommodityLongName', ...
    'IdentificationNumber', ...
    'PackagingType', ...
    'GeneralPackageType', ...
    'WhatFailedCode', ...
    'HowFailedCode', ...
    'FailureCauseCode', ...
    'SpillageResultInd', ...
    'FireResultInd', ...
    'ExplosionResultInd', ...
    'WaterSewerResultInd', ...
    'GasDispersionResultInd', ...
    'EnvironmentalDamageResult', ...
    'OtherCleanupInd', ...
    'MaterialLoss', ...
    'RemediationCleanupCost', ...
    'SeriousIncidentInd'});
```

Removal of select rows due to row content:

- MultipleRowsPerIncident == 'Yes' (Removal of 3% of dataset, n=19390) [For simplicity and clarity of counting rows and not double-counting]
- IncidentCountry ~= 'US' (Removal of 0.179%, n=1130) [Negligible reduction in rows, allows for removal of IncidentCountry column for file size]
- ReportType ~= 'A hazardous material incident' (Removal of 2.936%, n=18538) [For uniformity, the removed 2.936% were nonhazardous reports]

Overall reduction from 631491 to 596921 rows, n=34570, 5.5%

```
data(ismember(data.MultipleRowsPerIncident,'Yes'),:)=[];
data(~ismember(data.IncidentCountry,'US'),:)=[];
data(~ismember(data.ReportType,'A hazardous material incident'),:)=[];
data(:,1:3) = [];
```

Removal of rows with missing values:

Reduction from 596921 to 300254 rows, n=296667, 50%

```
data = rmmissing(data);
data = data(randperm(size(data,1)),:)
```

```
data = 300254×20 table
```

	DateOfIncident	ModeOfTransportation	TransportationPhase	
1	17-Nov-2017	'Highway'	'In Transit'	'SODIU
2	11-Jul-2020	'Highway'	'Unloading'	'DISINI
3	10-Sep-1995	'Highway'	'Unloading'	'INK PI
4	17-Sep-2012	'Highway'	'Unloading'	'SODIU
5	26-May-2021	'Highway'	'Unloading'	'CORR
6	01-Apr-2022	'Highway'	'Loading'	'ISOPR
7	19-Jan-2007	'Highway'	'Unloading'	'POTAS
8	10-Aug-2017	'Highway'	'Unloading'	'HYDR
9	12-Jul-2019	'Highway'	'Loading'	'ISOPR
10	10-Nov-2005	'Highway'	'Loading'	'PAINT
11	27-May-2023	'Highway'	'Loading'	'SODIU
12	25-Aug-2023	'Highway'	'Unloading'	'HYDR
13	12-Oct-2023	'Highway'	'Unloading'	'CORR
14	15-Aug-2012	'Highway'	'Unloading'	'PAINT