Import data

```
% all 12 months data is imported in a file datastore
raw_data= fileDatastore('D:\asus\15.Cody\Data Science\3. Predictive Modeling and Machine Learni
```

```
raw_data =
  FileDatastore with properties:

                      Files: {
                             ' ...\Taxi Data\Taxi Data\data\yellow_tripdata_2015-01.csv';
                             ' ...\Taxi Data\Taxi Data\data\yellow_tripdata_2015-02.csv';
                             ' ...\Taxi Data\Taxi Data\data\yellow_tripdata_2015-03.csv'
                              ... and 9 more
                             }
                    Folders: {
                             ' ...\3. Predictive Modeling and Machine Learning\Taxi Data\Taxi Data\data'
                             }
                UniformRead: 1
                   ReadMode: 'file'
                  BlockSize: Inf
                PreviewFcn: @importTaxiDataWithoutCleaning
      SupportedOutputFormats: ["txt"    "csv"    "xlsx"    "xls"    "parquet"    "parq"    "png"    "jpg"    "jpeg"
                   ReadFcn: @importTaxiDataWithoutCleaning
    AlternateFileSystemRoots: {}
```

```
data=readall(raw_data)
```

```
data = 2922266×19 table
```

...

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
1	2	2015-01-15 14:...	2015-01-15 14:...	1	3	-73.9643
2	2	2015-01-15 14:...	2015-01-15 14:...	1	0.6700	-73.9709
3	2	2015-01-07 14:...	2015-01-07 15:...	1	0.9800	-73.9487
4	2	2015-01-07 14:...	2015-01-07 15:...	3	4.3900	-73.9887
5	1	2015-01-20 23:...	2015-01-20 23:...	1	3.9000	-73.9750
6	2	2015-01-18 19:...	2015-01-18 20:...	6	4	-73.9710
7	2	2015-01-01 01:...	2015-01-01 01:...	1	5.7800	-74.0078
8	2	2015-01-01 01:...	2015-01-01 01:...	4	0.8800	-73.9642
9	1	2015-01-28 10:...	2015-01-28 10:...	1	0.6000	-73.9664
10	1	2015-01-23 16:...	2015-01-23 17:...	1	9.3000	-74.0067
11	1	2015-01-07 20:...	2015-01-07 20:...	1	6.9000	-73.9901
12	1	2015-01-10 19:...	2015-01-10 19:...	1	1	-73.9785
13	1	2015-01-10 19:...	2015-01-10 19:...	1	1.1000	-74.0016
14	2	2015-01-25 17:...	2015-01-25 17:...	1	0	-73.9757

⋮

The data contains 2,922,266 entries with 19 feature variables.

Exploratory Data Analysis (EDA)

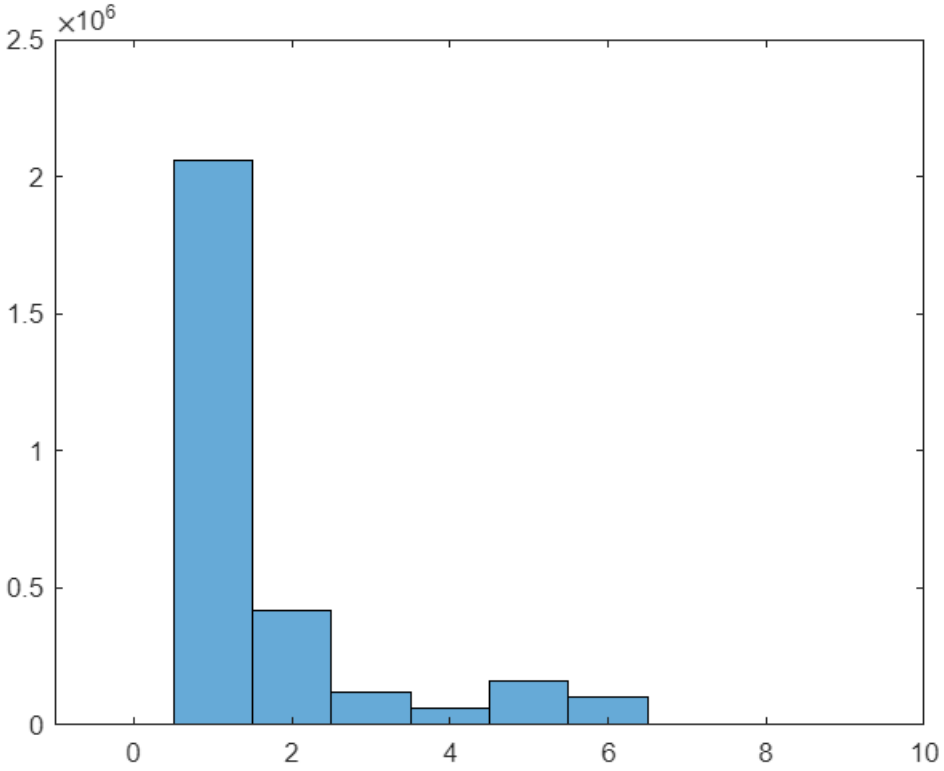
```
summary(data)
```

```
Variables:

    Vendor: 2922266×1 categorical

        Values:

            1    1.3876e+06
            2    1.5347e+06

    PickupTime: 2922266×1 datetime

        Values:

            Min       2015-01-01 00:00:43
            Median    2015-06-20 18:21:55
            Max       2015-12-31 23:59:59

    DropoffTime: 2922266×1 datetime

        Values:

            Min       2015-01-01 00:04:02
            Median    2015-06-20 18:35:14
            Max       2016-01-01 22:10:58

    Passengers: 2922266×1 double

        Values:

            Min          0
            Median       1
            Max          9

    Distance: 2922266×1 double

        Values:

            Min               0
            Median         1.71
            Max       1.468e+07

    PickupLon: 2922266×1 double

        Values:

            Min        -171.8
            Median    -73.982
            Max             0

    PickupLat: 2922266×1 double

        Values:

            Min             0
            Median     40.753
```

```
        Max        69.703

RateCode: 2922266×1 categorical

    Values:

        Standard      2.8453e+06
        JFK                61564
        Newark              5046
        Nassau              1051
        Negotiated          9110
        Group                 28
        99                   127

HeldFlag: 2922266×1 categorical

    Values:

        N    2.898e+06
        Y        24296

DropoffLon: 2922266×1 double

    Values:

        Min        -171.8
        Median     -73.98
        Max             0

DropoffLat: 2922266×1 double

    Values:

        Min             0
        Median     40.753
        Max        456.37

PayType: 2922266×1 categorical

    Values:

        Credit card    1.8323e+06
        Cash           1.0764e+06
        No charge           10130
        Dispute              3413
        Unknown                 3

Fare: 2922266×1 double

    Values:

        Min             -150
        Median           9.5
        Max       4.1027e+05

ExtraCharge: 2922266×1 double

    Values:

        Min         -45.2
        Median          0
        Max        579.72

Tax: 2922266×1 double
```

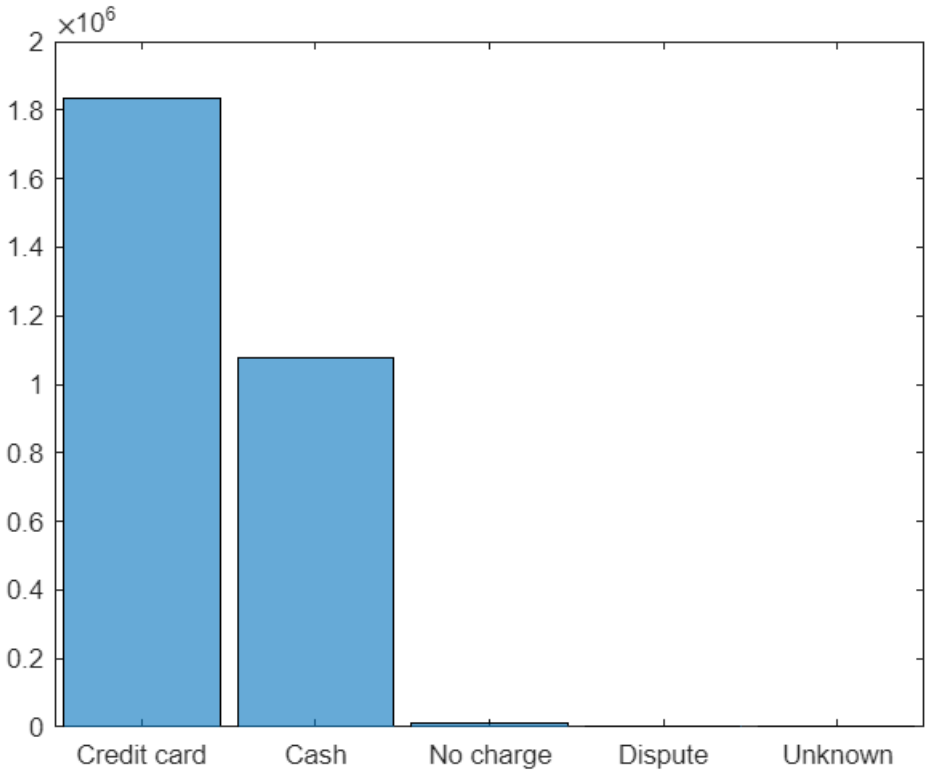
```
    Values:

        Min         -1.7
        Median       0.5
        Max        80.35

Tip: 2922266×1 double

    Values:

        Min         -2.7
        Median      1.16
        Max          650

Tolls: 2922266×1 double

    Values:

        Min           -15
        Median          0
        Max        911.08

ImpSurcharge: 2922266×1 double

    Values:

        Min         -0.3
        Median       0.3
        Max          0.3

TotalCharge: 2922266×1 double

    Values:

        Min          -150.8
        Median         11.8
        Max       4.1027e+05
```

```
histogram(data.Passengers)
```



Most of the rides were with single passengers.

```
histogram(data.PayType)
```



Adding taxi zones

```
data_2=addTaxiZones(data);
```

4 new feature variables are added using the shapefile information ("TaxiZones.shp" file).

Converting zones to taxis

Adding pick-up regions

```
% creating 2 new feature variables
data_2.pickup_region= data_2.PickupZone;
data_2.drop_region= data_2.DropoffZone;
```

```
% import the "Taxi Regions and Zones.csv" file as 'regions'
regions
```

```
regions = 19×6 table
```

...

	LowerManhattan	Midtown	UpperEastSide
1	"Alphabet City"	"Clinton East"	"Upper East Side North"

	LowerManhattan	Midtown	UpperEastSide
2	"Battery Park"	"Clinton West"	"Upper East Side South"
3	"Battery Park City"	"Midtown Center"	"Yorkville East"
4	"Chinatown"	"Midtown East"	"Yorkville West"
5	"East Village"	"Midtown North"	"Lenox Hill East"
6	"Financial District North"	"Midtown South"	"Lenox Hill West"
7	"Financial District South"	"Murray Hill"	"East Harlem South"
8	"Greenwich Village North"	"Penn Station/Madison Sq West"	""
9	"Greenwich Village South"	"Union Sq"	""
10	"Hudson Sq"	"UN/Turtle Bay South"	""
11	"Little Italy/NoLiTa"	"Times Sq/Theatre District"	""
12	"Lower East Side"	"Sutton Place/Turtle Bay North"	""
13	"Meatpacking/West Village ...	"Stuy Town/Peter Cooper Village"	""
14	"Seaport"	"West Chelsea/Hudson Yards"	""

⋮

```
% creating another copy
data_3=data_2;
```

```
% convert the variable to string
data_3.pickup_region=string(data_3.pickup_region);
```

```
% We want to group all these sub-regions into 6 overall region.
r1=replace(data_3.pickup_region, regions.LowerManhattan, "Lower Manhattan");
data_3.pickup_region=r1;

r2=replace(data_3.pickup_region,regions.Midtown, "Midtown");
data_3.pickup_region=r2;
```

```
r3=replace(data_3.pickup_region, regions.UpperEastSide(1:7), "Upper East Side");
data_3.pickup_region=r3;

r4=replace(data_3.pickup_region, regions.UpperWestSide(1:6), "Upper West Side");
data_3.pickup_region=r4;
```

```
data_3
```

```
data_3 = 2922266×25 table
```

...

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
1	2	2015-01-15 14:...	2015-01-15 14:...	1	3	-73.9643
2	2	2015-01-15 14:...	2015-01-15 14:...	1	0.6700	-73.9709
3	2	2015-01-07 14:...	2015-01-07 15:...	1	0.9800	-73.9487
4	2	2015-01-07 14:...	2015-01-07 15:...	3	4.3900	-73.9887
5	1	2015-01-20 23:...	2015-01-20 23:...	1	3.9000	-73.9750
6	2	2015-01-18 19:...	2015-01-18 20:...	6	4	-73.9710
7	2	2015-01-01 01:...	2015-01-01 01:...	1	5.7800	-74.0078
8	2	2015-01-01 01:...	2015-01-01 01:...	4	0.8800	-73.9642
9	1	2015-01-28 10:...	2015-01-28 10:...	1	0.6000	-73.9664
10	1	2015-01-23 16:...	2015-01-23 17:...	1	9.3000	-74.0067
11	1	2015-01-07 20:...	2015-01-07 20:...	1	6.9000	-73.9901
12	1	2015-01-10 19:...	2015-01-10 19:...	1	1	-73.9785
13	1	2015-01-10 19:...	2015-01-10 19:...	1	1.1000	-74.0016
14	2	2015-01-25 17:...	2015-01-25 17:...	1	0	-73.9757

⋮

```
data_4=data_3;
data_4.pickup_region=categorical(data_4.pickup_region);
```

```
summary(data_4.pickup_region)
```

```
Allerton/Pelham Gardens            13
Arrochar/Fort Wadsworth             2
Astoria                          6167
Astoria Park                       49
Auburndale                          9
Baisley Park                      414
Bath Beach                         18
Bay Ridge                         133
Bay Terrace/Fort Totten             7
Bayside                            19
Bedford                          1331
Bedford Park                       42
Bellerose                          12
Belmont                            36
Bensonhurst East                   31
Bensonhurst West                   54
Bloomfield/Emerson Hill             3
Boerum Hill                      3438
Borough Park                       60
Breezy Point/Fort Tilde…            1
Briarwood/Jamaica Hills           217
Brighton Beach                     18
Bronx Park                         11
Bronxdale                          14
Brooklyn Heights                 3509
```

Brooklyn Navy Yard	110
Brownsville	54
Bushwick North	886
Bushwick South	1689
Cambria Heights	14
Canarsie	37
Carroll Gardens	2040
Central Harlem	8758
Central Harlem North	4022
Central Park	37173
Charleston/Tottenville	1
City Island	4
Claremont/Bathgate	38
Clinton Hill	1716
Co-Op City	9
Cobble Hill	1856
College Point	18
Columbia Street	214
Coney Island	23
Corona	108
Country Club	2
Crotona Park	2
Crotona Park East	13
Crown Heights North	1226
Crown Heights South	225
Cypress Hills	17
DUMBO/Vinegar Hill	1307
Douglaston	4
Downtown Brooklyn/Metro…	4075
Dyker Heights	26
East Concourse/Concours…	132
East Elmhurst	487
East Flatbush/Farragut	53
East Flatbush/Remsen Vi…	50
East Flushing	5
East Harlem North	9290
East New York	95
East New York/Pennsylva…	21
East Tremont	27
East Williamsburg	2909
Eastchester	6
Elmhurst	1055
Elmhurst/Maspeth	485
Eltingville/Annadale/Pr…	1
Erasmus	105
Far Rockaway	7
Flatbush/Ditmas Park	385
Flatlands	47
Flushing	188
Flushing Meadows-Corona…	304
Fordham South	24
Forest Hills	583
Forest Park/Highland Park	7
Fort Greene	3103
Fresh Meadows	10
Glen Oaks	16
Glendale	35
Governor's Island/Ellis …	2
Gowanus	708
Gravesend	10
Great Kills	1
Green-Wood Cemetery	9
Greenpoint	2653
Hamilton Heights	3361

Hammels/Arverne	7
Heartland Village/Todt …	1
Highbridge	91
Highbridge Park	17
Hillcrest/Pomonok	33
Hollis	23
Homecrest	40
Howard Beach	20
Hunts Point	23
Inwood	297
Inwood Hill Park	13
JFK Airport	62178
Jackson Heights	1855
Jamaica	248
Jamaica Bay	2
Jamaica Estates	30
Kensington	175
Kew Gardens	169
Kew Gardens Hills	61
Kingsbridge Heights	39
LaGuardia Airport	70720
Laurelton	2
Long Island City/Hunter…	4303
Long Island City/Queens…	3242
Longwood	15
Lower Manhattan	554435
Lower Manhattan City	26156
Madison	21
Manhattan Beach	18
Manhattanville	2827
Marble Hill	21
Marine Park/Floyd Benne…	7
Marine Park/Mill Basin	23
Mariners Harbor	3
Maspeth	181
Melrose South	157
Middle Village	59
Midtown	1296074
Midtown-Queens	21
Midwood	66
Morningside Heights	14152
Morrisania/Melrose	72
Mott Haven/Port Morris	593
Mount Hope	55
New Dorp/Midland Beach	3
Newark Airport	160
North Corona	117
Norwood	33
Oakland Gardens	4
Ocean Hill	97
Ocean Parkway South	50
Old Astoria	2095
Ozone Park	27
Park Slope	4490
Parkchester	27
Pelham Bay	13
Pelham Parkway	20
Port Richmond	1
Prospect Heights	1224
Prospect Park	192
Prospect-Lefferts Gardens	487
Queens Village	32
Queensboro Hill	37
Queensbridge/Ravenswood	1076

Randalls Island	193
Red Hook	291
Rego Park	325
Richmond Hill	108
Ridgewood	126
Rikers Island	3
Riverdale/North Riverda…	29
Roosevelt Island	234
Rosedale	11
Saint Albans	9
Saint George/New Brighton	5
Saint Michaels Cemetery…	258
Schuylerville/Edgewater…	13
Sheepshead Bay	22
Soundview/Bruckner	31
Soundview/Castle Hill	41
South Beach/Dongan Hills	3
South Jamaica	174
South Ozone Park	178
South Williamsburg	375
Springfield Gardens North	18
Springfield Gardens South	149
Spuyten Duyvil/Kingsbri…	82
Starrett City	2
Steinway	2228
Stuyvesant Heights	609
Sunnyside	5435
Sunset Park East	34
Sunset Park West	426
University Heights/Morr…	79
Upper East Side	422235
Upper West Side	265523
Van Cortlandt Park	8
Van Cortlandt Village	41
Van Nest/Morris Park	30
Washington Heights North	653
Washington Heights South	2711
West Brighton	1
West Concourse	390
West Farms/Bronx River	24
Westchester Village/Uni…	35
Westerleigh	3
Whitestone	4
Willets Point	10
Williamsbridge/Olinville	20
Williamsburg (North Side)	6641
Williamsburg (South Side)	5352
Windsor Terrace	139
Woodhaven	60
Woodlawn/Wakefield	20
Woodside	1897
<undefined>	48799

So, there are many areas that did not fall into the 6 specified regions. We will take care of those later.

Adding Drop-off Regions

```
data_4.drop_region=string(data_4.drop_region);

r1=replace(data_4.drop_region, regions.LowerManhattan,"Lower Manhattan");
```

```
data_4.drop_region=r1;

r2=replace(data_4.drop_region, regions.Midtown, "Midtown");
data_4.drop_region=r2;

r3=replace(data_4.drop_region, regions.UpperEastSide(1:7),"Upper East Side");
data_4.drop_region=r3;

r4=replace(data_4.drop_region, regions.UpperWestSide(1:6),"Upper West Side");
data_4.drop_region=r4;

data_4.drop_region=categorical(data_4.drop_region)
```

data_4 = 2922266×25 table

...

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
1	2	2015-01-15 14:...	2015-01-15 14:...	1	3	-73.9643
2	2	2015-01-15 14:...	2015-01-15 14:...	1	0.6700	-73.9709
3	2	2015-01-07 14:...	2015-01-07 15:...	1	0.9800	-73.9487
4	2	2015-01-07 14:...	2015-01-07 15:...	3	4.3900	-73.9887
5	1	2015-01-20 23:...	2015-01-20 23:...	1	3.9000	-73.9750
6	2	2015-01-18 19:...	2015-01-18 20:...	6	4	-73.9710
7	2	2015-01-01 01:...	2015-01-01 01:...	1	5.7800	-74.0078
8	2	2015-01-01 01:...	2015-01-01 01:...	4	0.8800	-73.9642
9	1	2015-01-28 10:...	2015-01-28 10:...	1	0.6000	-73.9664
10	1	2015-01-23 16:...	2015-01-23 17:...	1	9.3000	-74.0067
11	1	2015-01-07 20:...	2015-01-07 20:...	1	6.9000	-73.9901
12	1	2015-01-10 19:...	2015-01-10 19:...	1	1	-73.9785
13	1	2015-01-10 19:...	2015-01-10 19:...	1	1.1000	-74.0016
14	2	2015-01-25 17:...	2015-01-25 17:...	1	0	-73.9757

⋮

```
summary(data_4.drop_region)
```

```
Allerton/Pelham Gardens          170
Arden Heights                     16
Arrochar/Fort Wadsworth           63
Astoria                        15209
Astoria Park                      77
Auburndale                       176
Baisley Park                     837
Bath Beach                       249
Bay Ridge                       2546
```

Bay Terrace/Fort Totten	217
Bayside	388
Bedford	6192
Bedford Park	468
Bellerose	143
Belmont	296
Bensonhurst East	397
Bensonhurst West	610
Bloomfield/Emerson Hill	69
Boerum Hill	5455
Borough Park	681
Breezy Point/Fort Tilde…	28
Briarwood/Jamaica Hills	803
Brighton Beach	278
Broad Channel	16
Bronx Park	128
Bronxdale	231
Brooklyn Heights	8264
Brooklyn Navy Yard	431
Brownsville	396
Bushwick North	3981
Bushwick South	6462
Cambria Heights	178
Canarsie	510
Carroll Gardens	3372
Central Harlem	17038
Central Harlem North	11821
Central Park	33415
Charleston/Tottenville	12
City Island	43
Claremont/Bathgate	266
Clinton Hill	6660
Co-Op City	210
Cobble Hill	2462
College Point	297
Columbia Street	972
Coney Island	234
Corona	1028
Country Club	62
Crotona Park	14
Crotona Park East	159
Crown Heights North	6454
Crown Heights South	1622
Cypress Hills	310
DUMBO/Vinegar Hill	4184
Douglaston	224
Downtown Brooklyn/Metro…	5395
Dyker Heights	493
East Concourse/Concours…	1112
East Elmhurst	1203
East Flatbush/Farragut	501
East Flatbush/Remsen Vi…	438
East Flushing	172
East Harlem North	20357
East New York	681
East New York/Pennsylva…	241
East Tremont	246
East Williamsburg	7665
Eastchester	124
Elmhurst	2853
Elmhurst/Maspeth	1462
Eltingville/Annadale/Pr…	24
Erasmus	485
Far Rockaway	117

Flatbush/Ditmas Park	2790
Flatlands	537
Flushing	1351
Flushing Meadows-Corona…	560
Fordham South	181
Forest Hills	3673
Forest Park/Highland Park	55
Fort Greene	5485
Fresh Meadows	316
Freshkills Park	2
Glen Oaks	153
Glendale	412
Governor's Island/Ellis …	1
Gowanus	1920
Gravesend	161
Great Kills	25
Green-Wood Cemetery	46
Greenpoint	9947
Grymes Hill/Clifton	24
Hamilton Heights	7957
Hammels/Arverne	107
Heartland Village/Todt …	59
Highbridge	562
Highbridge Park	149
Hillcrest/Pomonok	454
Hollis	121
Homecrest	380
Howard Beach	235
Hunts Point	310
Inwood	2237
Inwood Hill Park	190
JFK Airport	25608
Jackson Heights	5368
Jamaica	765
Jamaica Bay	3
Jamaica Estates	380
Kensington	968
Kew Gardens	703
Kew Gardens Hills	639
Kingsbridge Heights	343
LaGuardia Airport	35891
Laurelton	171
Long Island City/Hunter…	9892
Long Island City/Queens…	3846
Longwood	288
Lower Manhattan	496078
Lower Manhattan City	28694
Madison	345
Manhattan Beach	183
Manhattanville	4741
Marble Hill	152
Marine Park/Floyd Benne…	19
Marine Park/Mill Basin	339
Mariners Harbor	39
Maspeth	1135
Melrose South	850
Middle Village	961
Midtown	1205864
Midtown-Queens	394
Midwood	611
Morningside Heights	21721
Morrisania/Melrose	430
Mott Haven/Port Morris	2065
Mount Hope	560

New Dorp/Midland Beach	35
Newark Airport	4649
North Corona	866
Norwood	401
Oakland Gardens	208
Oakwood	13
Ocean Hill	693
Ocean Parkway South	279
Old Astoria	4417
Ozone Park	222
Park Slope	11263
Parkchester	372
Pelham Bay	163
Pelham Bay Park	32
Pelham Parkway	365
Port Richmond	13
Prospect Heights	3550
Prospect Park	697
Prospect-Lefferts Gardens	2373
Queens Village	283
Queensboro Hill	239
Queensbridge/Ravenswood	1737
Randalls Island	470
Red Hook	1295
Rego Park	1175
Richmond Hill	600
Ridgewood	1952
Rikers Island	3
Riverdale/North Riverda…	848
Rockaway Park	114
Roosevelt Island	1425
Rosedale	228
Rossville/Woodrow	18
Saint Albans	295
Saint George/New Brighton	66
Saint Michaels Cemetery…	267
Schuylerville/Edgewater…	337
Sheepshead Bay	355
Soundview/Bruckner	323
Soundview/Castle Hill	379
South Beach/Dongan Hills	55
South Jamaica	261
South Ozone Park	1059
South Williamsburg	1190
Springfield Gardens North	276
Springfield Gardens South	541
Spuyten Duyvil/Kingsbri…	1094
Stapleton	50
Starrett City	85
Steinway	6889
Stuyvesant Heights	3921
Sunnyside	8326
Sunset Park East	693
Sunset Park West	2163
University Heights/Morr…	525
Upper East Side	421602
Upper West Side	249292
Van Cortlandt Park	123
Van Cortlandt Village	510
Van Nest/Morris Park	301
Washington Heights North	5533
Washington Heights South	10589
West Brighton	35
West Concourse	1156

```
West Farms/Bronx River           301
Westchester Village/Uni…         255
Westerleigh                       45
Whitestone                       339
Willets Point                     28
Williamsbridge/Olinville         274
Williamsburg (North Side)      11766
Williamsburg (South Side)      10825
Windsor Terrace                 1601
Woodhaven                        552
Woodlawn/Wakefield               377
Woodside                        3860
<undefined>                    51915
```

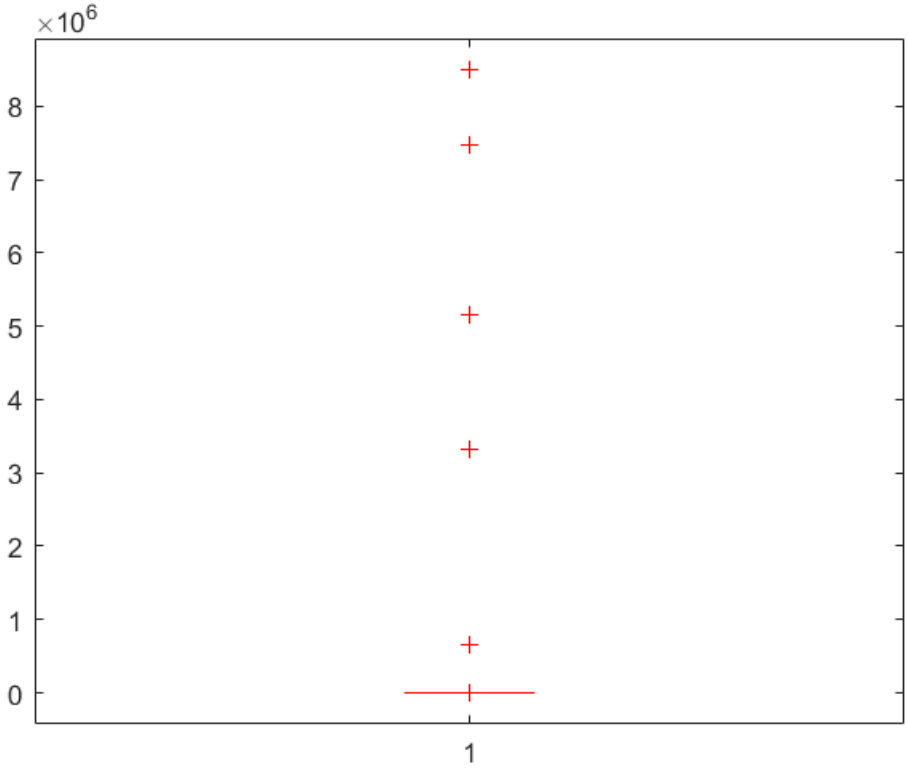
Data Preprocessing

```
dp = data_4;
```

```
%invalid ratecode
dp2 = dp(dp.RateCode ~= "99", :);
```

```
%invalid location
dp3 = standardizeMissing(dp2, 0, "DataVariables", ["PickupLat","PickupLon","DropoffLat","Dropo
dp3 = rmmissing(dp3, "DataVariables", ["PickupLat","PickupLon","DropoffLat","DropoffLon"]);
```

```
%passengers
dp4=dp3(dp3.Passengers>0,:);
boxplot(dp4.Distance)
```




```
prctile(dp4.Distance,[0,99])
```

```
ans = 1×2
         0   18.6000
```

```
dp5=dp4(dp4.Distance>0,:);
prctile(dp5.Distance,[0,99])
```

```
ans = 1×2
    0.0100   18.6000
```

```
prctile(dp5.Distance,[0,99.9])
```

```
ans = 1×2
    0.0100   24.5300
```

```
prctile(dp5.Distance,[0,99.99])
```

```
ans = 1×2
    0.0100   40.5000
```

```
dp6=rmoutliers(dp5,"percentiles",[0,99.99],"DataVariables","Distance")
```

```
dp6 = 2858743×25 table
```

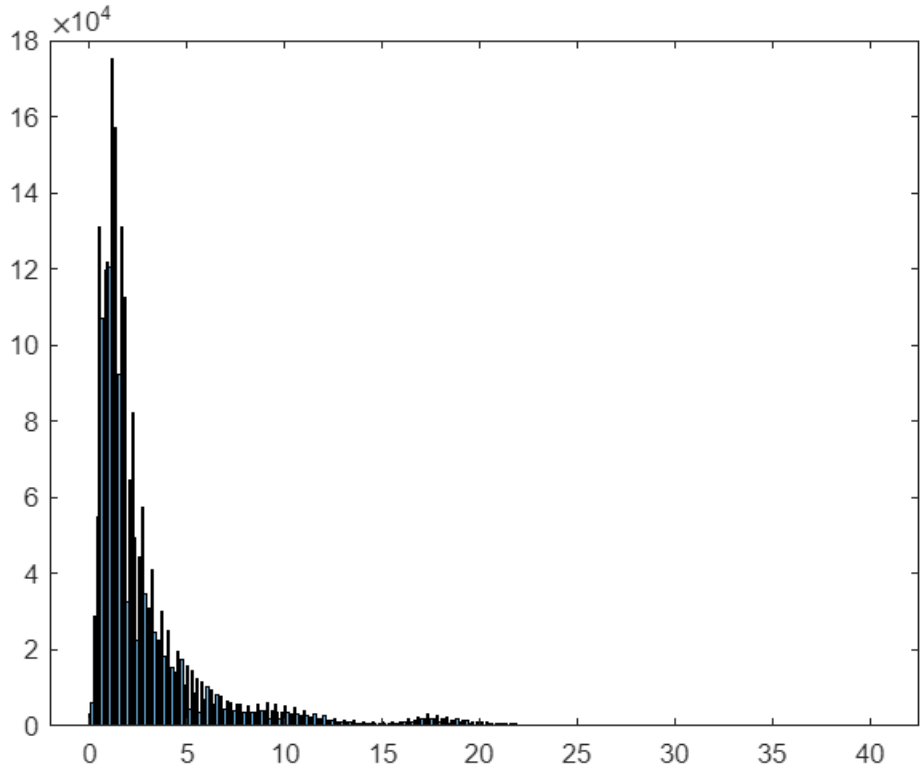
...

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
1	2	2015-01-15 14:...	2015-01-15 14:...	1	3	-73.9643

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
2	2	2015-01-15 14:...	2015-01-15 14:...	1	0.6700	-73.9709
3	2	2015-01-07 14:...	2015-01-07 15:...	1	0.9800	-73.9487
4	2	2015-01-07 14:...	2015-01-07 15:...	3	4.3900	-73.9887
5	1	2015-01-20 23:...	2015-01-20 23:...	1	3.9000	-73.9750
6	2	2015-01-18 19:...	2015-01-18 20:...	6	4	-73.9710
7	2	2015-01-01 01:...	2015-01-01 01:...	1	5.7800	-74.0078
8	2	2015-01-01 01:...	2015-01-01 01:...	4	0.8800	-73.9642
9	1	2015-01-28 10:...	2015-01-28 10:...	1	0.6000	-73.9664
10	1	2015-01-23 16:...	2015-01-23 17:...	1	9.3000	-74.0067
11	1	2015-01-07 20:...	2015-01-07 20:...	1	6.9000	-73.9901
12	1	2015-01-10 19:...	2015-01-10 19:...	1	1	-73.9785
13	1	2015-01-10 19:...	2015-01-10 19:...	1	1.1000	-74.0016
14	2	2015-01-23 00:...	2015-01-23 00:...	1	6.0300	-73.9852

⋮

```
histogram(dp6.Distance)
```




```
% fare
```

```
prctile(dp6.Fare,[0,99])
```

```
ans = 1×2
  -118    52
```

```
sum(dp6.Fare<=0)
```

```
ans = 1189
```

```
sum(dp6.Fare<=2.5)
```

```
ans = 5876
```

```
dp7=dp6(dp6.Fare>0,:);

prctile(dp7.Fare,[0,99.9])
```

```
ans = 1×2
    0.0100   76.0000
```

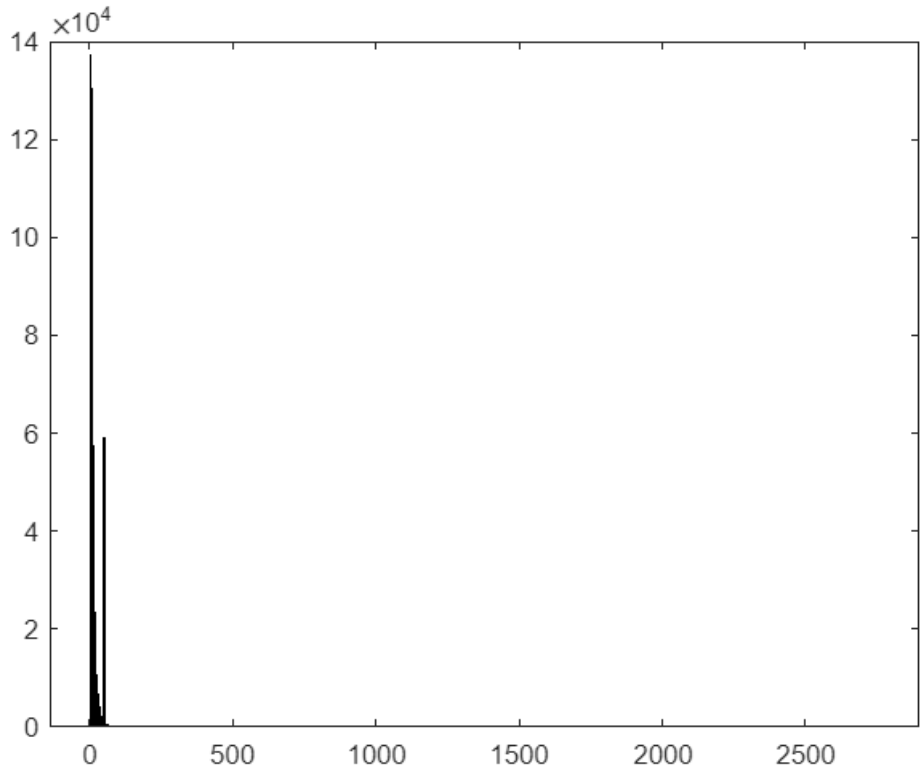
```
prctile(dp7.Fare,[0,99.99])
```

```
ans = 1×2
    0.0100  140.0000
```

```
sum(dp7.Fare>140)
```

```
ans = 270
```

```
histogram(dp7.Fare)
```




```
dp8=rmoutliers(dp7,"percentiles",[0,99.99],"DataVariables","Fare")
```

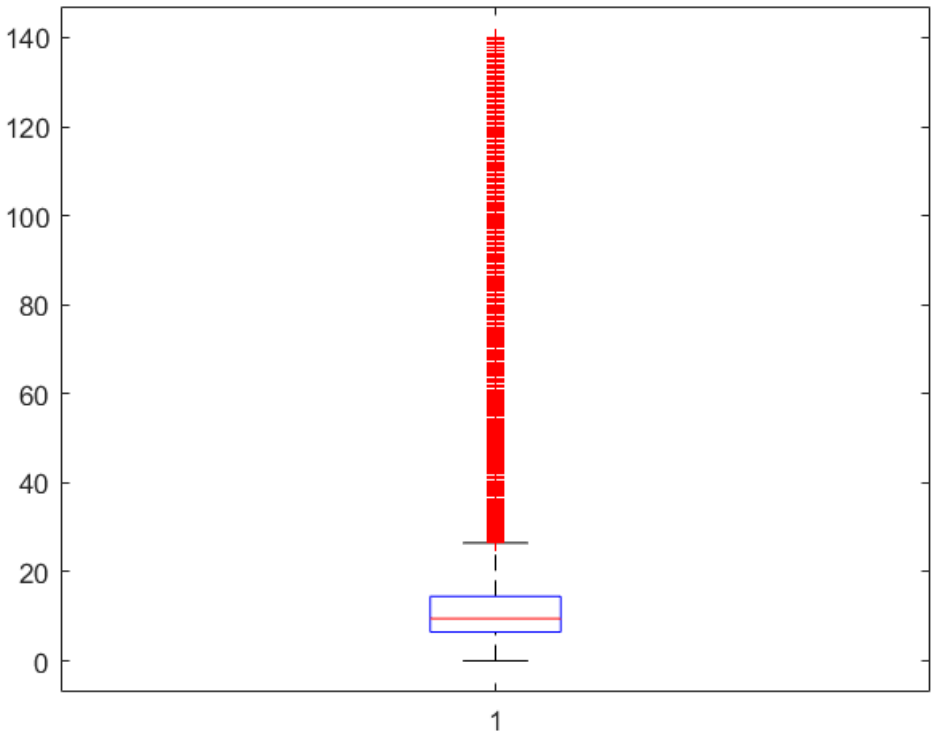
dp8 = 2857284×25 table

...

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
1	2	2015-01-15 14:...	2015-01-15 14:...	1	3	-73.9643
2	2	2015-01-15 14:...	2015-01-15 14:...	1	0.6700	-73.9709
3	2	2015-01-07 14:...	2015-01-07 15:...	1	0.9800	-73.9487
4	2	2015-01-07 14:...	2015-01-07 15:...	3	4.3900	-73.9887
5	1	2015-01-20 23:...	2015-01-20 23:...	1	3.9000	-73.9750
6	2	2015-01-18 19:...	2015-01-18 20:...	6	4	-73.9710
7	2	2015-01-01 01:...	2015-01-01 01:...	1	5.7800	-74.0078
8	2	2015-01-01 01:...	2015-01-01 01:...	4	0.8800	-73.9642
9	1	2015-01-28 10:...	2015-01-28 10:...	1	0.6000	-73.9664
10	1	2015-01-23 16:...	2015-01-23 17:...	1	9.3000	-74.0067
11	1	2015-01-07 20:...	2015-01-07 20:...	1	6.9000	-73.9901
12	1	2015-01-10 19:...	2015-01-10 19:...	1	1	-73.9785
13	1	2015-01-10 19:...	2015-01-10 19:...	1	1.1000	-74.0016
14	2	2015-01-23 00:...	2015-01-23 00:...	1	6.0300	-73.9852

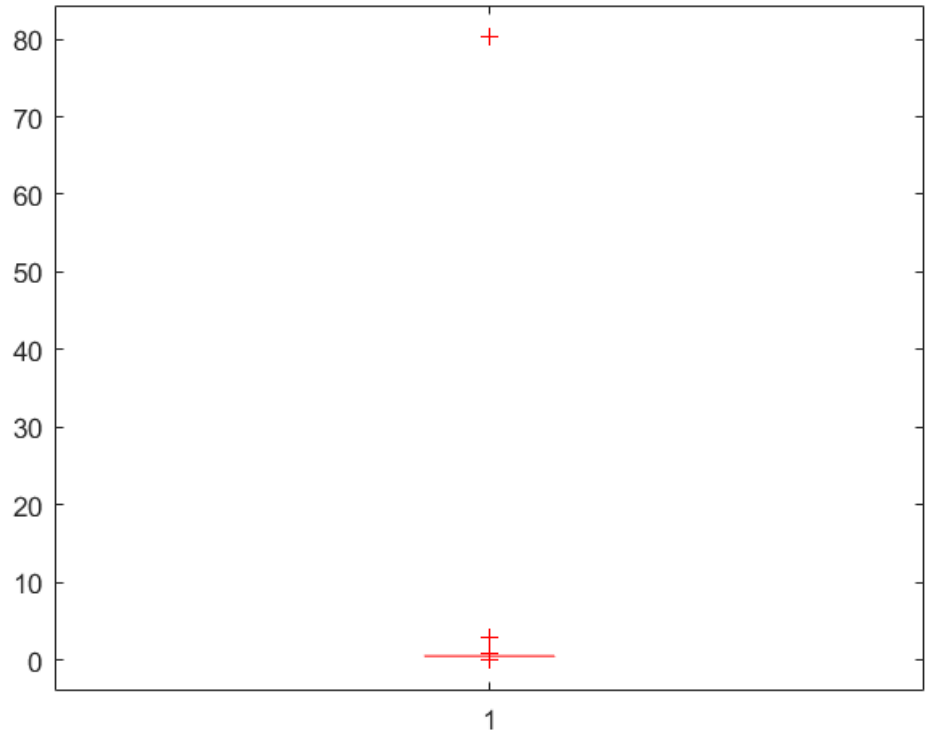
⋮

```
boxplot(dp8.Fare)
```

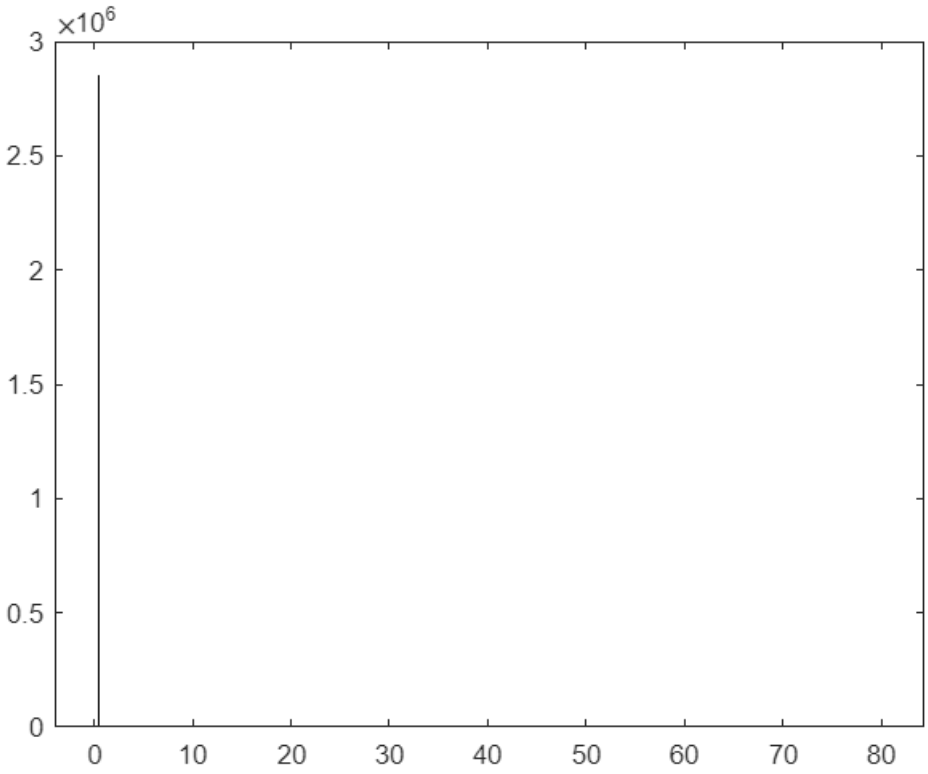



```
dp9=dp8(dp8.Fare>=2.5,:);
```

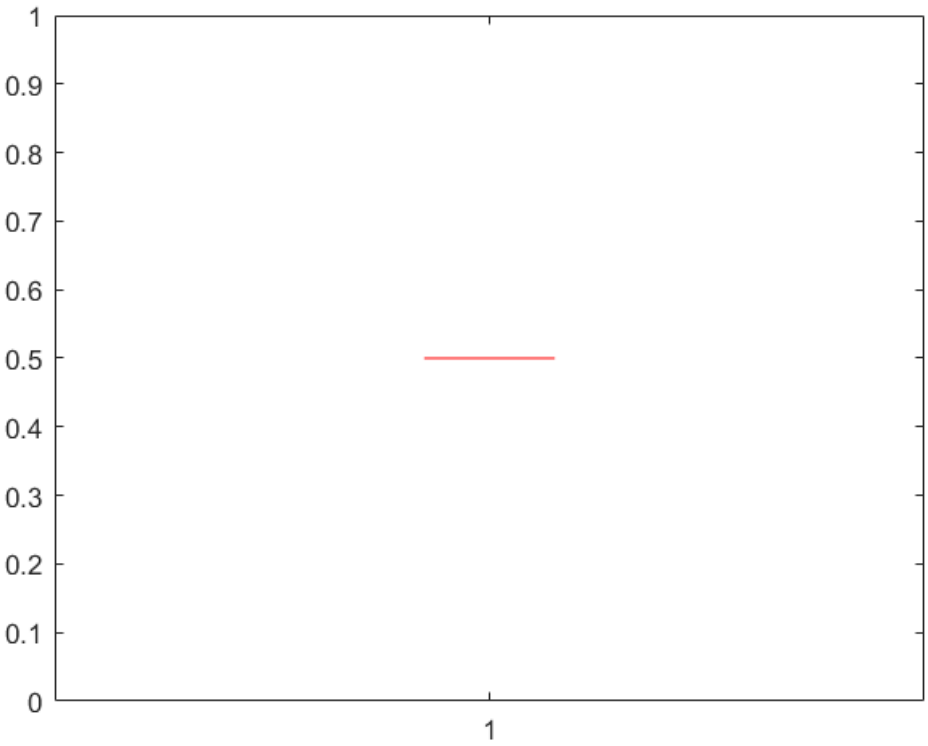
```
% extra charge
dp9=dp9(dp9.ExtraCharge>=0,:);
dp9=dp9(dp9.Tax>=0,:);
dp9=dp9(dp9.Tip>=0,:);
dp9=dp9(dp9.Tolls>=0,:);
dp9=dp9(dp9.ImpSurcharge>=0,:);
dp10=dp9(dp9.TotalCharge>=0,:);

boxplot(dp10.Tax)
```




```
histogram(dp10.Tax)
```



```
dp11 = dp10(abs(dp10.ImpSurcharge-0.3) < 0.01, :);
dp11 = dp11(abs(dp11.Tax-0.5) < 0.01, :);
dp11 = dp11(abs(dp11.Fare + dp11.ExtraCharge + dp11.Tax + dp11.Tip + dp11.Tolls + dp11.ImpSurch

boxplot(dp11.Tax)
```




```
% fare distance ratio
x=dp11.Fare./dp11.Distance;
prctile(x,[0,99])
```

```
ans = 1×2
    0.0712   15.1515
```

```
prctile(x,[0,99.99])
```

```
ans = 1×2
    0.0712  520.0000
```

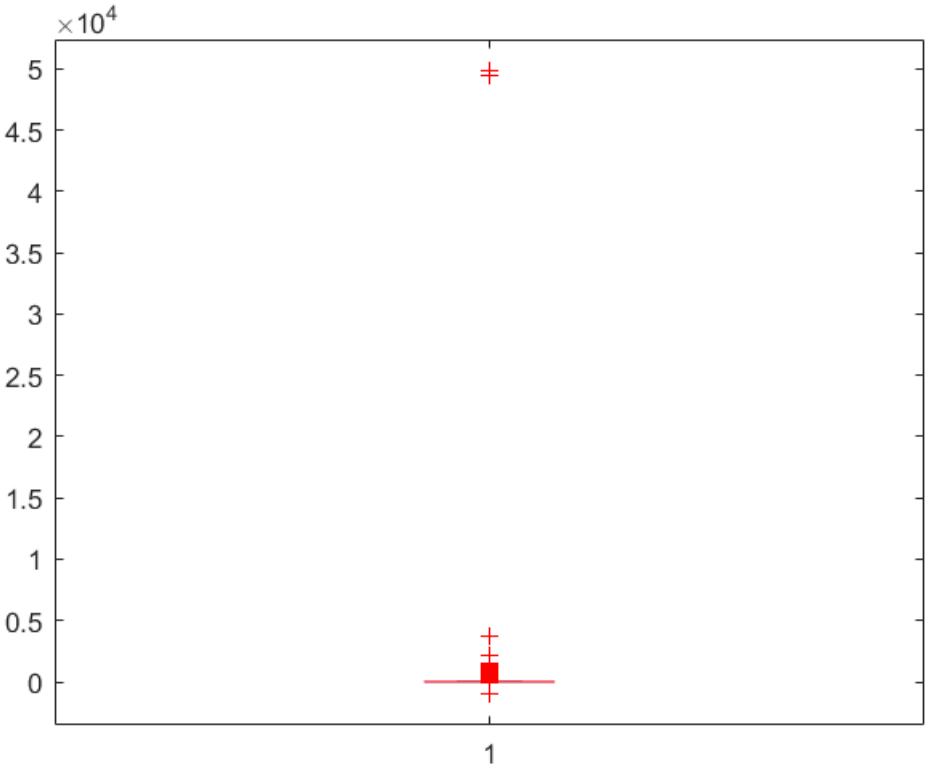
```
sum(x>520)
```

```
ans = 269
```

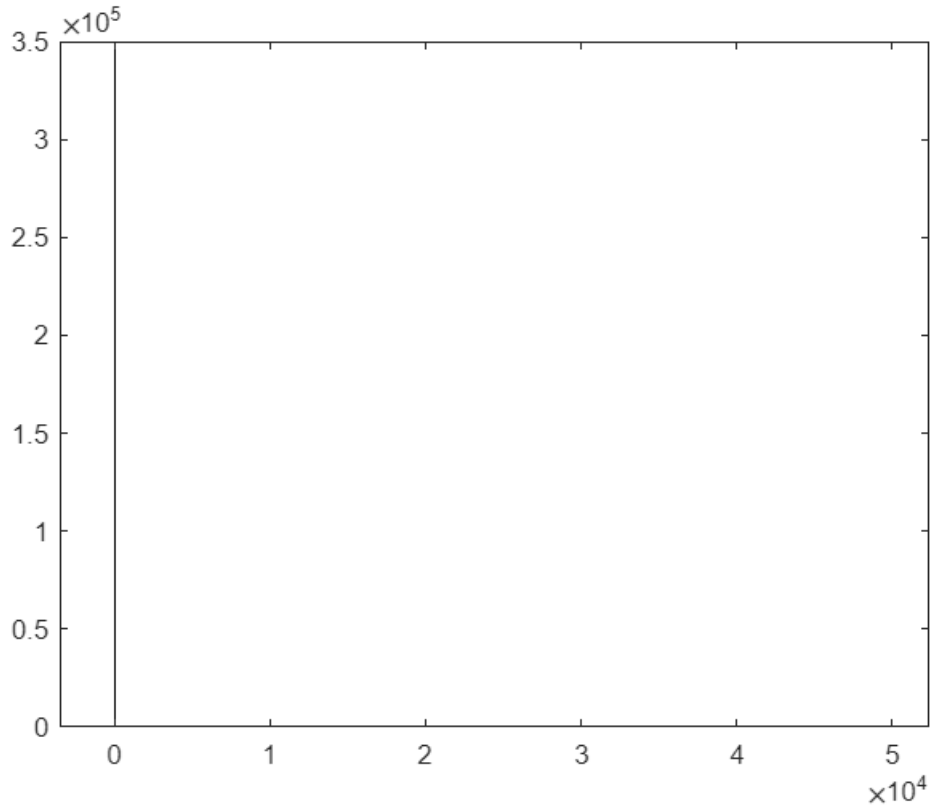
```
dp12=dp11(x<=520,:);
```

```
df = addDuration(dp12); % minutes
df = addAveSpeed(df); % mph
```

```
boxplot(df.Duration)
```




```
histogram(df.Duration)
```




```
prctile(df.Duration,[0,99])
```

```
ans = 1×2
 -985.5000   56.2667
```

```
prctile(df.Duration,[0,99.5])
```

```
ans = 1×2
 -985.5000   66.8667
```

```
sum(df.Duration<=0)
```

```
ans = 139
```

```
df2=df(df.Duration>=1 & df.Duration<120,:)
```

```
df2 = 2824737×27 table
```

...

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
1	2	2015-01-15 14:...	2015-01-15 14:...	1	3	-73.9643
2	2	2015-01-15 14:...	2015-01-15 14:...	1	0.6700	-73.9709
3	2	2015-01-07 14:...	2015-01-07 15:...	1	0.9800	-73.9487
4	2	2015-01-07 14:...	2015-01-07 15:...	3	4.3900	-73.9887
5	1	2015-01-20 23:...	2015-01-20 23:...	1	3.9000	-73.9750

	Vendor	PickupTime	DropoffTime	Passengers	Distance	PickupLon
6	2	2015-01-18 19:...	2015-01-18 20:...	6	4	-73.9710
7	2	2015-01-01 01:...	2015-01-01 01:...	1	5.7800	-74.0078
8	2	2015-01-01 01:...	2015-01-01 01:...	4	0.8800	-73.9642
9	1	2015-01-28 10:...	2015-01-28 10:...	1	0.6000	-73.9664
10	1	2015-01-23 16:...	2015-01-23 17:...	1	9.3000	-74.0067
11	1	2015-01-07 20:...	2015-01-07 20:...	1	6.9000	-73.9901
12	1	2015-01-10 19:...	2015-01-10 19:...	1	1	-73.9785
13	1	2015-01-10 19:...	2015-01-10 19:...	1	1.1000	-74.0016
14	2	2015-01-23 00:...	2015-01-23 00:...	1	6.0300	-73.9852

⋮

```
df2=df2(df2.AveSpeed>=0.1 & df2.AveSpeed<100,:);
df2=df2(df2.TotalCharge>=0.5 & df2.TotalCharge<=120,:);
df3=df2(df2.Tolls<=20,:);

timeofday(df3.PickupTime);

hour(df3.PickupTime(1:6));

%writetable(df3,'prepared_dataset_01.csv')
```

```
% Only keep trips that begin and end inside the region of interest.

lat = [40.5612 40.9637];
lon = [-74.1923 -73.5982];

inROI = inpolygon(df3.PickupLat,df3.PickupLon, lat([1 2 2 1]),lon([1 1 2 2])) ...
    & inpolygon(df3.DropoffLat,df3.DropoffLon, lat([1 2 2 1]),lon([1 1 2 2]));


df5 = df3(inROI,:);
```

Hourly data

```
df5.hourly_data= dateshift(df5.PickupTime,"start","hour");

df5.hourly_data_dropoff= dateshift(df5.DropoffTime,"start","hour");

df6=df5(:,["PickupTime","DropoffTime","Distance","Fare","ExtraCharge","Tax","Tip","Tolls","Imp
```

df6 = 2823681×16 table

...

	PickupTime	DropoffTime	Distance	Fare	ExtraCharge	Tax
1	2015-01-15 14:...	2015-01-15 14:...	3	12	0	0.5000
2	2015-01-15 14:...	2015-01-15 14:...	0.6700	5	0	0.5000
3	2015-01-07 14:...	2015-01-07 15:...	0.9800	7	0	0.5000
4	2015-01-07 14:...	2015-01-07 15:...	4.3900	15.5000	0	0.5000
5	2015-01-20 23:...	2015-01-20 23:...	3.9000	15	0.5000	0.5000
6	2015-01-18 19:...	2015-01-18 20:...	4	16	0	0.5000
7	2015-01-01 01:...	2015-01-01 01:...	5.7800	23	0.5000	0.5000
8	2015-01-01 01:...	2015-01-01 01:...	0.8800	6	0.5000	0.5000
9	2015-01-28 10:...	2015-01-28 10:...	0.6000	5.5000	0	0.5000
10	2015-01-23 16:...	2015-01-23 17:...	9.3000	39	1	0.5000
11	2015-01-07 20:...	2015-01-07 20:...	6.9000	23	0.5000	0.5000
12	2015-01-10 19:...	2015-01-10 19:...	1	6	0	0.5000
13	2015-01-10 19:...	2015-01-10 19:...	1.1000	8	0	0.5000
14	2015-01-23 00:...	2015-01-23 00:...	6.0300	20	0.5000	0.5000

⋮

```
%writetable(df6,'short_dataset.csv')
```

Removing Other Regions

```
ds=df6;

regions=["Lower Manhattan","Midtown","Upper East Side","Upper West Side","JFK Airport","LaGuard

ds2=ds((ds.pickup_region==regions(1) | ds.pickup_region==regions(2) | ds.pickup_region==regions
unique(ds2.pickup_region)
```

```
ans = 6×1 categorical
JFK Airport
LaGuardia Airport
Lower Manhattan
Midtown
Upper East Side
Upper West Side
```

```
ds3=ds2((ds2.drop_region==regions(1) | ds2.drop_region==regions(2) | ds2.drop_region==regions(3
```

```
ds3 = 2292216×16 table
```

...

	PickupTime	DropoffTime	Distance	Fare	ExtraCharge	Tax
1	2015-01-15 14:...	2015-01-15 14:...	0.6700	5	0	0.5000
2	2015-01-07 14:...	2015-01-07 15:...	0.9800	7	0	0.5000
3	2015-01-20 23:...	2015-01-20 23:...	3.9000	15	0.5000	0.5000
4	2015-01-18 19:...	2015-01-18 20:...	4	16	0	0.5000
5	2015-01-01 01:...	2015-01-01 01:...	0.8800	6	0.5000	0.5000
6	2015-01-28 10:...	2015-01-28 10:...	0.6000	5.5000	0	0.5000
7	2015-01-10 19:...	2015-01-10 19:...	1	6	0	0.5000
8	2015-01-10 19:...	2015-01-10 19:...	1.1000	8	0	0.5000
9	2015-01-23 17:...	2015-01-23 18:...	8.2000	34	1	0.5000
10	2015-01-17 19:...	2015-01-17 19:...	0.8900	5.5000	0	0.5000
11	2015-01-17 19:...	2015-01-17 19:...	2.5700	12.5000	0	0.5000
12	2015-01-17 23:...	2015-01-17 23:...	0.5000	5	0.5000	0.5000
13	2015-01-28 20:...	2015-01-28 20:...	0.8000	5.5000	0.5000	0.5000
14	2015-01-07 21:...	2015-01-07 21:...	0.5000	4.5000	0.5000	0.5000

⋮

```
unique(ds3.drop_region)
```

```
ans = 6×1 categorical
JFK Airport
LaGuardia Airport
Lower Manhattan
Midtown
Upper East Side
Upper West Side
```

Grouping

```
gp=groupsummary(ds3,["pickup_region","hourly_data"],"mean",["Duration","Distance","Fare"])
```

```
gp = 48734×6 table
```

...

	pickup_region	hourly_data	GroupCount	mean_Duration	mean_Distance
1	JFK Airport	2015-01-01 05:...	1	36	19.9300
2	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
3	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
4	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300
5	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
6	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550

	pickup_region	hourly_data	GroupCount	mean_Duration	mean_Distance
7	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
8	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
9	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
10	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050
11	JFK Airport	2015-01-01 18:...	2	30.4500	18.2300
12	JFK Airport	2015-01-01 19:...	1	31.3833	17.8200
13	JFK Airport	2015-01-01 20:...	3	30.4944	18.4100
14	JFK Airport	2015-01-01 21:...	3	32.2778	19.3700

⋮

```
gp.Properties.VariableNames(3)="pickup_count"
```

gp = 48734×6 table

...

	pickup_region	hourly_data	pickup_count	mean_Duration	mean_Distance
1	JFK Airport	2015-01-01 05:...	1	36	19.9300
2	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
3	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
4	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300
5	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
6	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550
7	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
8	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
9	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
10	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050
11	JFK Airport	2015-01-01 18:...	2	30.4500	18.2300
12	JFK Airport	2015-01-01 19:...	1	31.3833	17.8200
13	JFK Airport	2015-01-01 20:...	3	30.4944	18.4100
14	JFK Airport	2015-01-01 21:...	3	32.2778	19.3700

⋮

```
gd=groupsummary(ds3,["drop_region","hourly_data_dropoff"],"mean",["Duration","Distance","Fare"]
```

gd = 47468×6 table

...

	drop_region	hourly_data_dropoff	GroupCount	mean_Duration
1	JFK Airport	2015-01-01 04:00:00	1	31.2000

	drop_region	hourly_data_dropoff	GroupCount	mean_Duration
2	JFK Airport	2015-01-01 06:00:00	1	23.9500
3	JFK Airport	2015-01-01 07:00:00	2	22.1333
4	JFK Airport	2015-01-01 08:00:00	1	29.0833
5	JFK Airport	2015-01-01 09:00:00	3	25.1167
6	JFK Airport	2015-01-01 10:00:00	2	28.8500
7	JFK Airport	2015-01-01 11:00:00	5	29.3700
8	JFK Airport	2015-01-01 13:00:00	3	30.8333
9	JFK Airport	2015-01-01 15:00:00	2	32
10	JFK Airport	2015-01-01 16:00:00	1	30.1833
11	JFK Airport	2015-01-01 17:00:00	3	33.4278
12	JFK Airport	2015-01-01 18:00:00	1	22.8333
13	JFK Airport	2015-01-01 19:00:00	2	24.7917
14	JFK Airport	2015-01-01 20:00:00	1	33.5833

⋮

```
gd.Properties.VariableNames(3)="drop_count"
```

```
gd = 47468×6 table
```

...

	drop_region	hourly_data_dropoff	drop_count	mean_Duration
1	JFK Airport	2015-01-01 04:00:00	1	31.2000
2	JFK Airport	2015-01-01 06:00:00	1	23.9500
3	JFK Airport	2015-01-01 07:00:00	2	22.1333
4	JFK Airport	2015-01-01 08:00:00	1	29.0833
5	JFK Airport	2015-01-01 09:00:00	3	25.1167
6	JFK Airport	2015-01-01 10:00:00	2	28.8500
7	JFK Airport	2015-01-01 11:00:00	5	29.3700
8	JFK Airport	2015-01-01 13:00:00	3	30.8333
9	JFK Airport	2015-01-01 15:00:00	2	32
10	JFK Airport	2015-01-01 16:00:00	1	30.1833
11	JFK Airport	2015-01-01 17:00:00	3	33.4278
12	JFK Airport	2015-01-01 18:00:00	1	22.8333
13	JFK Airport	2015-01-01 19:00:00	2	24.7917
14	JFK Airport	2015-01-01 20:00:00	1	33.5833

⋮

```
sum(ismissing(gd))
```

```
ans = 1×6
     0     0     0     0     0     0
```

```
sum(ismissing(gp))
```

```
ans = 1×6
     0     0     0     0     0     0
```

```
gp.Properties.VariableNames(1)="region"
```

gp = 48734×6 table

...

	region	hourly_data	pickup_count	mean_Duration	mean_Distance
1	JFK Airport	2015-01-01 05:...	1	36	19.9300
2	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
3	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
4	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300
5	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
6	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550
7	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
8	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
9	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
10	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050
11	JFK Airport	2015-01-01 18:...	2	30.4500	18.2300
12	JFK Airport	2015-01-01 19:...	1	31.3833	17.8200
13	JFK Airport	2015-01-01 20:...	3	30.4944	18.4100
14	JFK Airport	2015-01-01 21:...	3	32.2778	19.3700

⋮

```
gd.Properties.VariableNames(1)="region"
```

gd = 47468×6 table

...

	region	hourly_data_dropoff	drop_count	mean_Duration	mean_Distance
1	JFK Airport	2015-01-01 04:00:00	1	31.2000	18.4100
2	JFK Airport	2015-01-01 06:00:00	1	23.9500	16.3800
3	JFK Airport	2015-01-01 07:00:00	2	22.1333	16.8150
4	JFK Airport	2015-01-01 08:00:00	1	29.0833	18.6500
5	JFK Airport	2015-01-01 09:00:00	3	25.1167	17.9633

	region	hourly_data_dropoff	drop_count	mean_Duration	mean_Distance
6	JFK Airport	2015-01-01 10:00:00	2	28.8500	17.8450
7	JFK Airport	2015-01-01 11:00:00	5	29.3700	18.2240
8	JFK Airport	2015-01-01 13:00:00	3	30.8333	19.0967
9	JFK Airport	2015-01-01 15:00:00	2	32	15.8350
10	JFK Airport	2015-01-01 16:00:00	1	30.1833	19.2500
11	JFK Airport	2015-01-01 17:00:00	3	33.4278	19.2667
12	JFK Airport	2015-01-01 18:00:00	1	22.8333	16.0600
13	JFK Airport	2015-01-01 19:00:00	2	24.7917	16.4550
14	JFK Airport	2015-01-01 20:00:00	1	33.5833	17

⋮

```
gd.Properties.VariableNames(2)="hourly_data"
```

gd = 47468×6 table

...

	region	hourly_data	drop_count	mean_Duration	mean_Distance
1	JFK Airport	2015-01-01 04:...	1	31.2000	18.4100
2	JFK Airport	2015-01-01 06:...	1	23.9500	16.3800
3	JFK Airport	2015-01-01 07:...	2	22.1333	16.8150
4	JFK Airport	2015-01-01 08:...	1	29.0833	18.6500
5	JFK Airport	2015-01-01 09:...	3	25.1167	17.9633
6	JFK Airport	2015-01-01 10:...	2	28.8500	17.8450
7	JFK Airport	2015-01-01 11:...	5	29.3700	18.2240
8	JFK Airport	2015-01-01 13:...	3	30.8333	19.0967
9	JFK Airport	2015-01-01 15:...	2	32	15.8350
10	JFK Airport	2015-01-01 16:...	1	30.1833	19.2500
11	JFK Airport	2015-01-01 17:...	3	33.4278	19.2667
12	JFK Airport	2015-01-01 18:...	1	22.8333	16.0600
13	JFK Airport	2015-01-01 19:...	2	24.7917	16.4550
14	JFK Airport	2015-01-01 20:...	1	33.5833	17

⋮

Joining

```
dj= outerjoin(gp,gd,"Keys",["region","hourly_data"]);
```

```
dj_2= outerjoin(gp,gd,"Keys",["region","hourly_data"],"MergeKeys",true);

%Data Missing
sum(ismissing(dj_2))
```

```
ans = 1×10
     0     0     1902     1902     1902     1902 ···
```

```
%dj_3=fillmissing(dj_2,"constant",0)

dj_3=fillmissing(dj_2,"constant",0,'DataVariables',@isnumeric);

%Combining

dj_3.netpickups= dj_3.pickup_count- dj_3.drop_count;

dj_3.avg_duration=(dj_3.mean_Duration_gd + dj_3.mean_Duration_gp)/2;

dj_3.avg_distance=(dj_3.mean_Distance_gd + dj_3.mean_Distance_gp)/2;

dj_3.avg_fare=(dj_3.mean_Fare_gd + dj_3.mean_Fare_gp)/2;

dj_3
```

dj_3 = 50636×14 table

...

	region	hourly_data	pickup_count	mean_Duration_gp	mean_Distance_gp
1	JFK Airport	2015-01-01 04:...	0	0	0
2	JFK Airport	2015-01-01 05:...	1	36	19.9300
3	JFK Airport	2015-01-01 06:...	0	0	0
4	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
5	JFK Airport	2015-01-01 08:...	0	0	0
6	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
7	JFK Airport	2015-01-01 10:...	0	0	0
8	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300
9	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
10	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550
11	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
12	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
13	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
14	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050

⋮

Train-test split

```
dj_4=dj_3;

rng(1)
partition=cvpartition(height(dj_4),"HoldOut",0.2)
```

```
partition =
Hold-out cross validation partition
   NumObservations: 50636
       NumTestSets: 1
         TrainSize: 40509
          TestSize: 10127
```

```
train_idx=training(partition);
test_idx=test(partition);

train_data= dj_4(train_idx,:);
test_data = dj_4 (test_idx,:);
```

Generating Response Variable

```
train_data.demand= discretize(train_data.netpickups,[-inf,0,15,inf],"categorical",["low","mediu
```

train_data = 40509×15 table

...

	region	hourly_data	pickup_count	mean_Duration_gp	mean_Distance_gp
1	JFK Airport	2015-01-01 06:...	0	0	0
2	JFK Airport	2015-01-01 08:...	0	0	0
3	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
4	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550
5	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
6	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
7	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
8	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050
9	JFK Airport	2015-01-01 18:...	2	30.4500	18.2300
10	JFK Airport	2015-01-01 19:...	1	31.3833	17.8200
11	JFK Airport	2015-01-01 20:...	3	30.4944	18.4100
12	JFK Airport	2015-01-01 21:...	3	32.2778	19.3700
13	JFK Airport	2015-01-01 22:...	5	29.8567	18.2060
14	JFK Airport	2015-01-02 00:...	2	25.1583	17.7750

⋮

```
test_data.demand= discretize(test_data.netpickups,[-inf,0,15,inf],"categorical",["low","medium"
```

test_data = 10127×15 table

	region	hourly_data	pickup_count	mean_Duration_gp	mean_Distance_gp
1	JFK Airport	2015-01-01 04:...	0	0	0
2	JFK Airport	2015-01-01 05:...	1	36	19.9300
3	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
4	JFK Airport	2015-01-01 10:...	0	0	0
5	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300
6	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
7	JFK Airport	2015-01-01 23:...	4	25.4833	18.1775
8	JFK Airport	2015-01-02 03:...	0	0	0
9	JFK Airport	2015-01-02 14:...	5	44.9400	18.8100
10	JFK Airport	2015-01-02 19:...	6	35.4861	18.8000
11	JFK Airport	2015-01-03 01:...	1	20.0667	18.9000
12	JFK Airport	2015-01-03 06:...	2	29.7833	17.4300
13	JFK Airport	2015-01-03 08:...	1	29.8667	19.2900
14	JFK Airport	2015-01-03 17:...	1	50.9000	19.0600

⋮

Summary Statistics

```
groupsummary(train_data,"demand")
```

ans = 3×2 table

	demand	GroupCount
1	low	18037
2	medium	19236
3	high	3236

```
groupsummary(test_data,"demand")
```

ans = 3×2 table

	demand	GroupCount
1	low	4414
2	medium	4916
3	high	797

```
groupsummary(train_data,["demand","region"])
```

ans = 17×3 table

	demand	region	GroupCount
1	low	JFK Airport	1795
2	low	LaGuardia Airport	2173
3	low	Lower Manhattan	3348
4	low	Midtown	3217
5	low	Upper East Side	3963
6	low	Upper West Side	3541
7	medium	JFK Airport	4536
8	medium	LaGuardia Airport	3944
9	medium	Lower Manhattan	3061
10	medium	Midtown	2476
11	medium	Upper East Side	2239
12	medium	Upper West Side	2980
13	high	LaGuardia Airport	79
14	high	Lower Manhattan	588

⋮

```
groupsummary(test_data,["demand","region"])
```

ans = 17×3 table

	demand	region	GroupCount
1	low	JFK Airport	487
2	low	LaGuardia Airport	499
3	low	Lower Manhattan	816
4	low	Midtown	792
5	low	Upper East Side	944
6	low	Upper West Side	876
7	medium	JFK Airport	1144
8	medium	LaGuardia Airport	965
9	medium	Lower Manhattan	802
10	medium	Midtown	665
11	medium	Upper East Side	597
12	medium	Upper West Side	743
13	high	LaGuardia Airport	12
14	high	Lower Manhattan	139

Feature Creation

```
df=dj_4;

[~,df.DayOfWeek] = weekday(df.hourly_data,"long")
```

df = 50636×15 table

	region	hourly_data	pickup_count	mean_Duration_gp	mean_Distance_gp
1	JFK Airport	2015-01-01 04:...	0	0	0
2	JFK Airport	2015-01-01 05:...	1	36	19.9300
3	JFK Airport	2015-01-01 06:...	0	0	0
4	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
5	JFK Airport	2015-01-01 08:...	0	0	0
6	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
7	JFK Airport	2015-01-01 10:...	0	0	0
8	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300
9	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
10	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550
11	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
12	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
13	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
14	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050

```
df.DayOfWeek = categorical(cellstr(df.DayOfWeek));

x=df.hourly_data(26)
```

x = *datetime*
2015-01-02 06:00:00

```
x2=datevec(x)
```

x2 = 1×6

2015	1	2	6	0	0

```
x3=datenum(x2(1:3))
```

x3 = 735966

```
day = x3 - datenum(x2(1), 1,0)
```

```
day = 2
```

```
df=adddayofyear(df)
```

```
df = 50636×16 table
```

...

	region	hourly_data	pickup_count	mean_Duration_gp	mean_Distance_gp
1	JFK Airport	2015-01-01 04:...	0	0	0
2	JFK Airport	2015-01-01 05:...	1	36	19.9300
3	JFK Airport	2015-01-01 06:...	0	0	0
4	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
5	JFK Airport	2015-01-01 08:...	0	0	0
6	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
7	JFK Airport	2015-01-01 10:...	0	0	0
8	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300
9	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
10	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550
11	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
12	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
13	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
14	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050

⋮

```
df.demand= discretize(df.netpickups,[-inf,0,15,inf],"categorical",["low","medium","high"])
```

```
df = 50636×17 table
```

...

	region	hourly_data	pickup_count	mean_Duration_gp	mean_Distance_gp
1	JFK Airport	2015-01-01 04:...	0	0	0
2	JFK Airport	2015-01-01 05:...	1	36	19.9300
3	JFK Airport	2015-01-01 06:...	0	0	0
4	JFK Airport	2015-01-01 07:...	1	23.7333	19.4700
5	JFK Airport	2015-01-01 08:...	0	0	0
6	JFK Airport	2015-01-01 09:...	1	23.7000	17.2700
7	JFK Airport	2015-01-01 10:...	0	0	0
8	JFK Airport	2015-01-01 11:...	1	36.3000	19.2300

	region	hourly_data	pickup_count	mean_Duration_gp	mean_Distance_gp
9	JFK Airport	2015-01-01 12:...	1	24.3000	16.9800
10	JFK Airport	2015-01-01 13:...	2	23.9750	18.1550
11	JFK Airport	2015-01-01 14:...	2	35.5750	20.1100
12	JFK Airport	2015-01-01 15:...	2	37.7667	19.1750
13	JFK Airport	2015-01-01 16:...	4	35.8792	18.2450
14	JFK Airport	2015-01-01 17:...	2	28.4833	19.2050

⋮

Feature Selection

```
%heatmap(df.demand,df.DayOfWeek)

crosstab(df.demand,df.DayOfWeek)
```

```
ans = 3×7
        3225        3157        3142        3453        3272        3087 ···
        3405        3516        3584        3305        3436        3462
         620         571         482         428         642         645
```

```
[a,chi2,p]=crosstab(df.demand,df.DayOfWeek)
```

```
a = 3×7
        3225        3157        3142        3453        3272        3087 ···
        3405        3516        3584        3305        3436        3462
         620         571         482         428         642         645
chi2 = 121.4323
p = 3.2012e-20
```

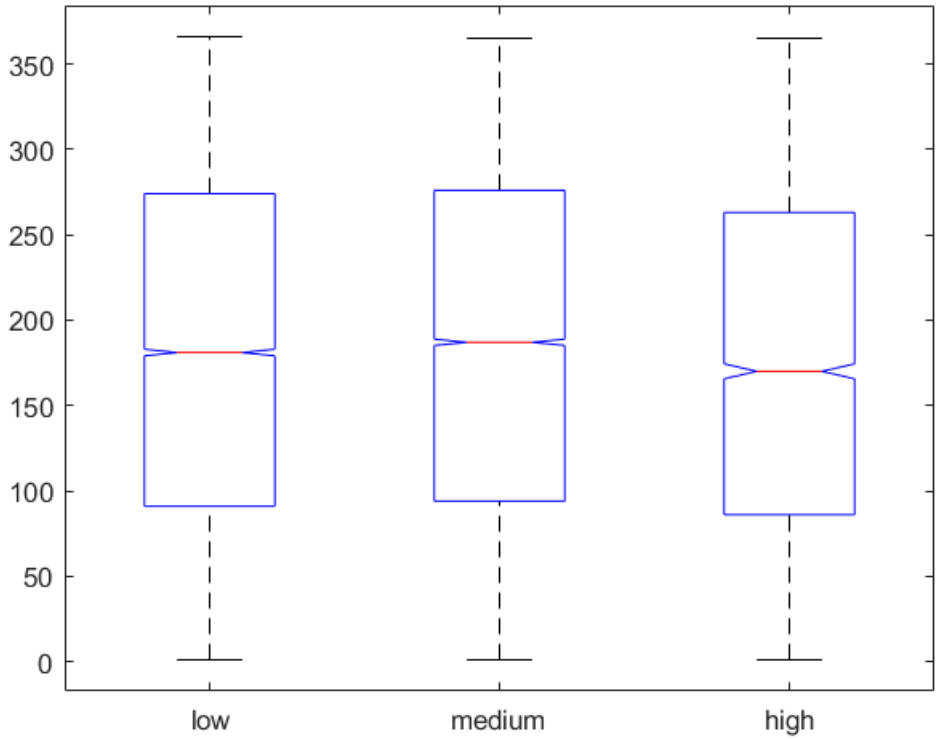
```
[a,chi2,p]=crosstab(df.demand,df.dayofyear)
```

```
a = 3×366
    61    61    63    66    62    59    58    60    62    59    64    51    57 ···
    74    74    69    70    67    69    67    62    65    66    69    75    67
     1     2     5     1    13    10    15    17    11    14     7    13    14
chi2 = 618.0252
p = 0.9990
```

```
[p,tbl]=anova1(df.dayofyear,df.demand)
```

ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	359427.9	2	179714	16.25	8.81705e-08
Error	559993961.7	50633	11059.9		
Total	560353389.6	50635			



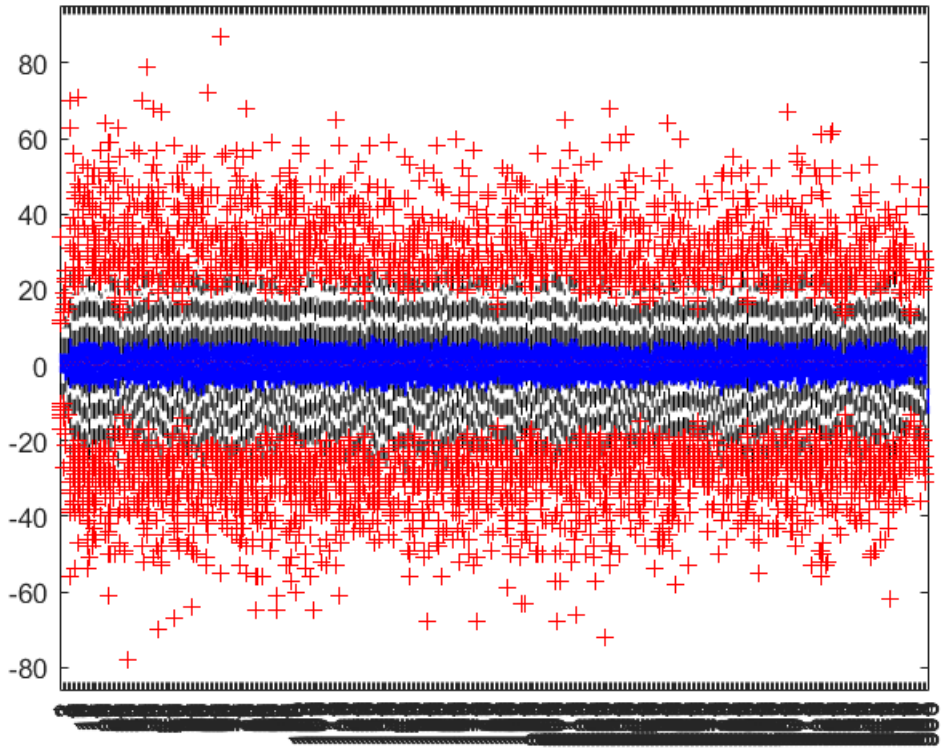

```
p = 8.8170e-08
tbl = 4×6 cell
```

	1	2	3	4	5	6
1	'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
2	'Groups'	3.5943e+05	2	1.7971e+05	16.2492	8.8170e-08
3	'Error'	5.5999e+08	50633	1.1060e+04	[ ]	[ ]
4	'Total'	5.6035e+08	50635	[ ]	[ ]	[ ]

```
[p,tbl]=anova1(df.netpickups,df.dayofyear)
```

ANOVA Table

```
Source        SS         df        MS        F      Prob>F
-----------------------------------------------------------
Groups      1702.44      365     4.664    0.03        1
Error    7794999.56    50270   155.063
Total    7796702       50635
```




```
p = 1
tbl = 4×6 cell
```

	1	2	3	4	5	6
1	'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
2	'Groups'	1.7024e+03	365	4.6642	0.0301	1
3	'Error'	7.7950e+06	50270	155.0627	[ ]	[ ]
4	'Total'	7796702	50635	[ ]	[ ]	[ ]

```
Warning: While saving an object of class 'matlab.graphics.primitive.Line':
Recursion limit exceeded when saving instance of class to a MAT-file. This is either because the instance
contains an extremely long chain of references, or the class definition contains an error.
Warning: While saving an object of class 'matlab.graphics.primitive.Line':
Recursion limit exceeded when saving instance of class to a MAT-file. This is either because the instance
contains an extremely long chain of references, or the class definition contains an error.
```

```
%[p,tbl]=anova1(string(df.demand),df.dayofyear)

%df_2= isholiday(df,holidays)

%[a,chi2,p]=crosstab(df_2.demand,df_2.isholiday)

%[a,chi2,p]=crosstab(df_2.demand,df_2.DayOfWeek)

%s=groupsummary(df_2,["DayOfWeek","region"],"mean","netpickups")

%gscatter(s.DayOfWeek,s.region,s.mean_netpickups)
```

```
%heatmap(s,"DayOfWeek","mean_netpickups")

%df_2.hourofday=hours(timeofday(df_2.hourly_data))

%corr(df_2.demand,df_2.avg_duration)

%df_3=df_2

%df_3.demand(df_2.demand=='low')=0

%df_3.demand=grp2idx(df_3.demand)

%summary(df_3)

%corr(df_3.demand,df_3.avg_duration)

%corr(df_3.demand,df_3.avg_distance)

%corr(df_3.demand,df_3.avg_fare)

%corr(df_3.demand,df_3.DayOfWeek)
```

Raw Model

```
y_pred=raw_model_bagged.predictFcn(test_data)

cMetrics(test_data.demand,y_pred)
```

Oversampling the Minority Class

```
x_train_v1= [x_train,y_train]
xhigh=x_train_v1(x_train_v1.demand=='high',:)
xothers=x_train_v1(x_train_v1.demand~='high',:)

histogram(x_train_v1.demand)
[a,b]=histcounts(x_train_v1.demand)
xhigh_os= datasample(xhigh,12000,"Replace",true)

combining

x_comb=[xhigh_os;xothers]

histogram(x_comb.demand)
```

Prediction on Test Data

```
y_pred=model_bag_unb_cost.predictFcn(test_data)
cMetrics(test_data.demand,y_pred)
confusionchart(test_data.demand,y_pred,"Normalization","row-normalized")

y_pred=model_bag_unb_2.predictFcn(test_data)
cMetrics(test_data.demand,y_pred)
confusionchart(test_data.demand,y_pred,"Normalization","row-normalized")
```

Classifier Function

```
function [trainedClassifier, validationAccuracy] = trainClassifier_01(trainingData)
% [trainedClassifier, validationAccuracy] = trainClassifier(trainingData)
% Returns a trained classifier and its accuracy. This code recreates the
% classification model trained in Classification Learner app. Use the
% generated code to automate training the same model with new data, or to
% learn how to programmatically train models.
%
%  Input:
%      trainingData: A table containing the same predictor and response
%       columns as those imported into the app.
%
%  Output:
%      trainedClassifier: A struct containing the trained classifier. The
%       struct contains various fields with information about the trained
%       classifier.
%
%      trainedClassifier.predictFcn: A function to make predictions on new
%       data.
%
%      validationAccuracy: A double containing the accuracy in percent. In
%       the app, the History list displays this overall accuracy score for
%       each model.
%
% Use the code to train the model with new data. To retrain your
% classifier, call the function from the command line with your original
% data or new data as the input argument trainingData.
%
% For example, to retrain a classifier trained with the original data set
% T, enter:
%   [trainedClassifier, validationAccuracy] = trainClassifier(T)
%
% To make predictions with the returned 'trainedClassifier' on new data T2,
% use
%   yfit = trainedClassifier.predictFcn(T2)
%
% T2 must be a table containing at least the same predictor columns as used
% during training. For details, enter:
%   trainedClassifier.HowToPredict

% Auto-generated by MATLAB on 09-Apr-2021 17:24:47


% Extract predictors and response
% This code processes the data into the right shape for training the
% model.
inputTable = trainingData;
predictorNames = {'region', 'avg_duration', 'avg_distance', 'avg_fare', 'DayOfWeek', 'dayofyear', 'isholiday',
predictors = inputTable(:, predictorNames);
response = inputTable.demand;
isCategoricalPredictor = [true, false, false, false, true, false, false, false];

% Train a classifier
```

```
% This code specifies all the classifier options and trains the classifier.
template = templateTree(...
    'MaxNumSplits', 40508);
classificationEnsemble = fitcensemble(...
    predictors, ...
    response, ...
    'Method', 'Bag', ...
    'NumLearningCycles', 30, ...
    'Learners', template, ...
    'Cost', [0 4 2; 10 0 6; 1 1 0], ...
    'ClassNames', categorical({'high'; 'low'; 'medium'}));

% Create the result struct with predict function
predictorExtractionFcn = @(t) t(:, predictorNames);
ensemblePredictFcn = @(x) predict(classificationEnsemble, x);
trainedClassifier.predictFcn = @(x) ensemblePredictFcn(predictorExtractionFcn(x));

% Add additional fields to the result struct
trainedClassifier.RequiredVariables = {'DayOfWeek', 'avg_distance', 'avg_duration', 'avg_fare', 'dayofyear', '
trainedClassifier.ClassificationEnsemble = classificationEnsemble;
trainedClassifier.About = 'This struct is a trained model exported from Classification Learner R2020a.';
trainedClassifier.HowToPredict = sprintf('To make predictions on a new table, T, use: \n  yfit = c.predictFcn(

% Extract predictors and response
% This code processes the data into the right shape for training the
% model.
inputTable = trainingData;
predictorNames = {'region', 'avg_duration', 'avg_distance', 'avg_fare', 'DayOfWeek', 'dayofyear', 'isholiday',
predictors = inputTable(:, predictorNames);
response = inputTable.demand;
isCategoricalPredictor = [true, false, false, false, true, false, false, false];

% Perform cross-validation
partitionedModel = crossval(trainedClassifier.ClassificationEnsemble, 'KFold', 5);

% Compute validation predictions
[validationPredictions, validationScores] = kfoldPredict(partitionedModel);

% Compute validation accuracy
validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
```

Loss analysis

```
p= raw_model.predictFcn(test_data)

idx_low_high= find(p=='high' & y_test.demand=='low')

raw_fare=x_test(idx_low_high,"avg_fare")

summary(raw_fare)

mean(raw_fare.avg_fare)

p= final_model.predictFcn(test_data)
idx_low_high= find(p=='high' & y_test.demand=='low')
raw_fare=x_test(idx_low_high,"avg_fare")
summary(raw_fare)
```

```
mean(raw_fare.avg_fare)

p= raw_model.predictFcn(test_data)
idx_low_high= find(p=='medium' & y_test.demand=='low')
raw_fare=x_test(idx_low_high,"avg_fare")
summary(raw_fare)
mean(raw_fare.avg_fare)

p= final_model.predictFcn(test_data)
idx_low_high= find(p=='medium' & y_test.demand=='low')
raw_fare=x_test(idx_low_high,"avg_fare")
summary(raw_fare)
mean(raw_fare.avg_fare)
```