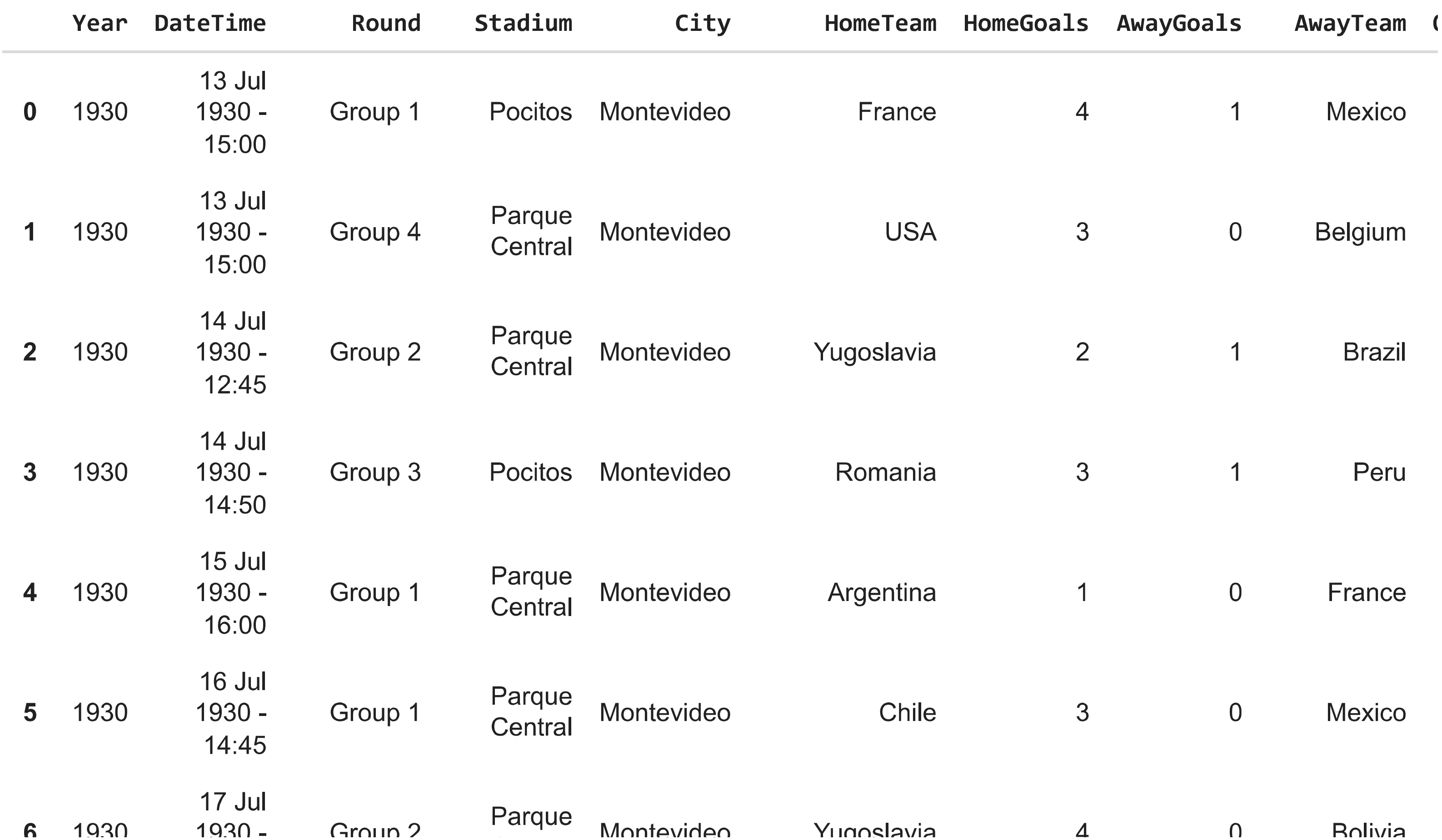
```
import pandas as pd
import numpy as np
import matplotlib as plt
import seaborn as sns


dataset = pd.read_csv("WorldCupMatches.csv")
dataset.head(30)
```



	Year	DateTime	Round	Stadium	City	HomeTeam	HomeGoals	AwayGoals	AwayTeam
0	1930	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4	1	Mexico
1	1930	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3	0	Belgium
2	1930	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2	1	Brazil
3	1930	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3	1	Peru
4	1930	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1	0	France
5	1930	16 Jul 1930 - 14:45	Group 1	Parque Central	Montevideo	Chile	3	0	Mexico
6	1930	17 Jul 1930 -	Group 2	Parque	Montevideo	Yugoslavia	4	0	Bolivia

```
dataset1= pd.read_csv("WorldCupMatches.csv",index_col=0)
dataset1
```

	DateTime	Round	Stadium	City	HomeTeam	HomeGoals	AwayGoals	AwayTeam	Observatio
Year									
1930	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4	1	Mexico	
1930	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3	0	Belgium	
1930	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2	1	Brazil	

```
dataset1=pd.read_csv("WorldCupMatches.csv",index_col=0,na_values=["??","***"])
dataset1
```

	DateTime	Round	Stadium	City	HomeTeam	HomeGoals	AwayGoals	AwayTeam	Observatio
Year									
	13 Jul								

```
dataset1.at[1930,"Stadium"]
```

```
Year
1930              Pocitos
1930       Parque Central
1930       Parque Central
1930              Pocitos
1930       Parque Central
1930       Parque Central
1930       Parque Central
1930       Parque Central
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
1930   Estadio Centenario
Name: Stadium, dtype: object
```

Stadiums which particpated in 1930

```
dataset.at[0,"Stadium"]
```

```
'Pocitos'
```

```
dataset.size
```

```
8520
```

```
dataset.shape
```

```
     (852, 10)
```

```
dataset.dtypes
```

```
     Year            int64
     DateTime       object
     Round          object
     Stadium        object
     City           object
     HomeTeam       object
     HomeGoals       int64
     AwayGoals       int64
     AwayTeam       object
     Observation    object
     dtype: object
```

```
dataset.columns
```

```
     Index(['Year', 'DateTime', 'Round', 'Stadium', 'City', 'HomeTeam', 'HomeGoals',
            'AwayGoals', 'AwayTeam', 'Observation'],
           dtype='object')
```

```
dataset.memory_usage()
```

```
     Index           128
     Year           6816
     DateTime       6816
     Round          6816
     Stadium        6816
     City           6816
     HomeTeam       6816
     HomeGoals      6816
     AwayGoals      6816
     AwayTeam       6816
```

```
Observation    6816
dtype: int64
```

```
dataset1.memory_usage()
```

```
Index          39880
DateTime        6816
Round           6816
Stadium         6816
City            6816
HomeTeam        6816
HomeGoals       6816
AwayGoals       6816
AwayTeam        6816
Observation     6816
dtype: int64
```

```
dataset1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 852 entries, 1930 to 2014
Data columns (total 9 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   DateTime     852 non-null    object
 1   Round        852 non-null    object
 2   Stadium      852 non-null    object
 3   City         852 non-null    object
 4   HomeTeam     847 non-null    object
 5   HomeGoals    852 non-null    int64 
 6   AwayGoals    852 non-null    int64 
 7   AwayTeam     844 non-null    object
 8   Observation  852 non-null    object
dtypes: int64(2), object(7)
memory usage: 98.9+ KB
```

```
dataset1.describe()
```

	HomeGoals	AwayGoals
count	852.000000	852.000000
mean	1.811033	1.022300
std	1.610255	1.087573
min	0.000000	0.000000
25%	1.000000	0.000000
50%	2.000000	1.000000
75%	3.000000	2.000000
max	10.000000	7.000000

```
dataset.describe()
```

	Year	HomeGoals	AwayGoals
count	852.000000	852.000000	852.000000
mean	1985.089202	1.811033	1.022300
std	22.448825	1.610255	1.087573
min	1930.000000	0.000000	0.000000
25%	1970.000000	1.000000	0.000000
50%	1990.000000	2.000000	1.000000
75%	2002.000000	3.000000	2.000000
max	2014.000000	10.000000	7.000000

```
#Slice the result for first 5 rows
```

```
print(dataset[0:5]['City'])

     0    Montevideo
     1    Montevideo
     2    Montevideo
     3    Montevideo
     4    Montevideo
     Name: City, dtype: object
```

```
dataset.loc[0]

     Year                           1930
     DateTime       13 Jul 1930 - 15:00 
     Round                       Group 1
     Stadium                     Pocitos
     City                     Montevideo 
     HomeTeam                     France
     HomeGoals                         4
     AwayGoals                         1
     AwayTeam                     Mexico
     Observation                        
     Name: 0, dtype: object
```

```
dataset.loc[1]

     Year                           1930
     DateTime       13 Jul 1930 - 15:00 
     Round                       Group 4
     Stadium              Parque Central
     City                     Montevideo 
     HomeTeam                        USA
     HomeGoals                         3
     AwayGoals                         0
     AwayTeam                    Belgium
     Observation                        
     Name: 1, dtype: object
```

```
dataset1.loc[2014]
```

	DateTime	Round	Stadium	City	HomeTeam	HomeGoals	AwayGoals	AwayTeam	Observation
Year									
2014	12 Jun 2014 - 17:00	Group A	Arena de Sao Paulo	Sao Paulo	Brazil	3	1	Croatia	
2014	13 Jun 2014 - 13:00	Group A	Estadio das Dunas	Natal	Mexico	1	0	Cameroon	
2014	13 Jun 2014 - 16:00	Group B	Arena Fonte Nova	Salvador	Spain	1	5	Netherlands	
2014	13 Jun 2014 - 18:00	Group B	Arena Pantanal	Cuiaba	Chile	3	1	Australia	
2014	14 Jun 2014 - 13:00	Group C	Estadio Mineirao	Belo Horizonte	Colombia	3	0	Greece	
...	...	...	...	...	...	...	...	...	...
	05 Jul		Arena						Netherlands

```
dataset.iloc[-2]
```

```
     Year                               2014
     DateTime            12 Jul 2014 - 17:00
     Round          Play-off for third place
     Stadium                Estadio Nacional
     City                          Brasilia 
     HomeTeam                         Brazil
     HomeGoals                             0
     AwayGoals                             3
     AwayTeam                    Netherlands
     Observation                            
     Name: 850, dtype: object
```

```
#Reading Specific Columns
print(dataset1.loc[0:6,['Stadium','City']])

     Empty DataFrame
     Columns: [Stadium, City]
     Index: []
```

Group By: Average home goals in each stadium in each year

```
print(dataset.groupby(["Year","Stadium"])['HomeGoals'].mean())

     Year  Stadium
     1930  Estadio Centenario     3.600000
           Parque Central         2.666667
           Pocitos                3.500000
     1934  Giorgio Ascarelli      3.500000
           Giovanni Berta         2.333333
                                    ...   
     2014  Estadio Castelao       1.625000
           Estadio Mineirao       1.250000
           Estadio Nacional       1.300000
           Estadio das Dunas      0.500000
           Estadio do Maracana    0.900000
     Name: HomeGoals, Length: 191, dtype: float64
```

```
#showing all null values
dataset.isnull()
```

	Year	DateTime	Round	Stadium	City	HomeTeam	HomeGoals	AwayGoals	AwayTeam	Observation
0	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...
847	False	False	False	False	False	False	False	False	False	False
848	False	False	False	False	False	False	False	False	False	False
849	False	False	False	False	False	False	False	False	False	False
850	False	False	False	False	False	False	False	False	False	False
851	False	False	False	False	False	False	False	False	False	False

852 rows × 10 columns

```
dataset.isnull().sum()
```

```
Year          0
DateTime      0
Round         0
Stadium       0
City          0
HomeTeam      5
HomeGoals     0
AwayGoals     0
AwayTeam      8
```

```
    Observation    0
    dtype: int64
```

```
dataset.isnull().values.any()
```

```
    True
```

```
print(dataset['HomeGoals'].isnull())
```

```
    0      False
    1      False
    2      False
    3      False
    4      False
           ...
    847    False
    848    False
    849    False
    850    False
    851    False
    Name: HomeGoals, Length: 852, dtype: bool
```

```
print(dataset['AwayGoals'].isnull())
```

```
    0      False
    1      False
    2      False
    3      False
    4      False
           ...
    847    False
    848    False
    849    False
    850    False
    851    False
    Name: AwayGoals, Length: 852, dtype: bool
```

```
dataset.fillna(0)
```

	Year	DateTime	Round	Stadium	City	HomeTeam	HomeGoals	AwayGoals	AwayTeam	Observ
0	1930	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4	1	Mexico	
1	1930	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3	0	Belgium	
2	1930	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2	1	Brazil	
3	1930	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3	1	Peru	
4	1930	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1	0	France	
...	...	...	...	...	...	...	...	...	...	
		05 Jul		Arena						Nethe

```
dataset.fillna(0,inplace=True)
dataset.shape
```

```
     (852, 10)
```

```
dataset.isnull().sum()
```

```
     Year           0
     DateTime       0
     Round          0
     Stadium        0
     City           0
     HomeTeam       0
     HomeGoals      0
     AwayGoals      0
     AwayTeam       0
     Observation    0
     dtype: int64
```

```
dataset.shape
```

```
     (852, 10)
```

```
dataset.dropna(inplace=True)
```

```
dataset.shape
```

```
     (852, 10)
```

```
dataset.head(30)
```

	Year	DateTime	Round	Stadium	City	HomeTeam	HomeGoals	AwayGoals	AwayTeam	(
0	1930	13 Jul 1930 - 15:00	Group 1	Pocitos	Montevideo	France	4	1	Mexico	
1	1930	13 Jul 1930 - 15:00	Group 4	Parque Central	Montevideo	USA	3	0	Belgium	
2	1930	14 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	2	1	Brazil	
3	1930	14 Jul 1930 - 14:50	Group 3	Pocitos	Montevideo	Romania	3	1	Peru	
4	1930	15 Jul 1930 - 16:00	Group 1	Parque Central	Montevideo	Argentina	1	0	France	
5	1930	16 Jul 1930 - 14:45	Group 1	Parque Central	Montevideo	Chile	3	0	Mexico	
6	1930	17 Jul 1930 - 12:45	Group 2	Parque Central	Montevideo	Yugoslavia	4	0	Bolivia	
7	1930	17 Jul 1930 - 14:45	Group 4	Parque Central	Montevideo	USA	3	0	Paraguay	
8	1930	18 Jul 1930 - 14:30	Group 3	Estadio Centenario	Montevideo	Uruguay	1	0	Peru	
9	1930	19 Jul 1930 - 12:50	Group 1	Estadio Centenario	Montevideo	Chile	1	0	France	
		19 Jul								

10	1930	19 Jul 1930 - 15:00	Group 1	Estadio Centenario	Montevideo	Argentina	6	3	Mexico
11	1930	20 Jul 1930 - 13:00	Group 2	Estadio Centenario	Montevideo	Brazil	4	0	Bolivia
12	1930	20 Jul 1930 - 15:00	Group 4	Estadio Centenario	Montevideo	Paraguay	1	0	Belgium
		21 Jul		Estadio					

```
from sklearn import preprocessing
```

		22 Jul		Estadio					

Data transformation categorical values

```
x = np.random.uniform(0.0,1.0,size=(10,2))
y = np.random.choice(('Male','Female'),size=(10))
x[0]
```

```
     array([0.24031914, 0.38565597])
```

```
x
```

```
     array([[0.24031914, 0.38565597],
            [0.99109438, 0.44363365],
            [0.48674251, 0.70725583],
            [0.82837732, 0.36564014],
            [0.17635625, 0.37069176],
            [0.11764371, 0.36422172],
            [0.10594998, 0.73399216],
            [0.87726682, 0.82292319],
            [0.72064142, 0.48927008],
            [0.65153335, 0.88297504]])
```

```
y
```

```
    array(['Male', 'Female', 'Female', 'Male', 'Male', 'Female', 'Male',
           'Female', 'Male', 'Male'], dtype='<U6')
```

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
yt = le.fit_transform(y)
print(yt)
```

```
    [1 0 0 1 1 0 1 0 1 1]
```

```
output = [0, 0, 0 ,1 ,0 ,0  ,1, 1, 0, 1]
```

```
decoded_output = [le.classes_[i] for i in output]
decoded_output
```

```
    ['Female',
     'Female',
     'Female',
     'Male',
     'Female',
     'Female',
     'Male',
     'Male',
     'Female',
     'Male']
```

```
census_names = ['age','workclass','fnlwgt','education','education_num','marital_status','occupation','relationship','race','sex','cap
```

```
df_census = pd.read_csv('adult.data',names=census_names)
df_census
```

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	rac
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	Whi
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	Whi
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	Whi
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Blac
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Blac
...	...	...	...	...	...	...	...	...	.
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	Whi
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	Whi
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	Whi
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	Whi
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	Whi

32561 rows × 15 columns

```
bool_series=df_census.duplicated(subset=None,keep=False)
bool_series
```

```
     0        False
     1        False
     2        False
     3        False
     4        False
              ...
     32556    False
     32557    False
     32558    False
     32559    False
     32560    False
     Length: 32561, dtype: bool
```

```
df_census_unique=df_census[-bool_series]
print("Before removing duplicates")
print(df_census.shape)
print("Before after duplicates")
print(df_census_unique.shape)
```

```
     Before removing duplicates
     (32561, 15)
     Before after duplicates
     (32514, 15)
```

```
train=pd.read_csv("/content/sample_data/california_housing_test.csv")
train
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	med
0	-122.05	37.37	27.0	3885.0	661.0	1537.0	606.0	
1	-118.30	34.26	43.0	1510.0	310.0	809.0	277.0	
2	-117.81	33.78	27.0	3589.0	507.0	1484.0	495.0	
3	-118.36	33.82	28.0	67.0	15.0	49.0	11.0	
4	-119.67	36.33	19.0	1241.0	244.0	850.0	237.0	
...	...	...	...	...	...	...	...	

```
d = preprocessing.normalize(train, axis=0)
scaled_df = pd.DataFrame(d)
scaled_df.head()
```

	0	1	2	3	4	5	6	7	8
0	-0.018631	0.019112	0.015670	0.021005	0.017919	0.016122	0.018104	0.028491	0.026795
1	-0.018058	0.017522	0.024956	0.008164	0.008404	0.008486	0.008275	0.015516	0.013720
2	-0.017983	0.017276	0.015670	0.019405	0.013745	0.015566	0.014788	0.024977	0.021027
3	-0.018067	0.017296	0.016250	0.000362	0.000407	0.000514	0.000329	0.026454	0.025652
4	-0.018267	0.018580	0.011027	0.006710	0.006615	0.008916	0.007080	0.012664	0.006351

```
dataset.corr()
```

	Year	HomeGoals	AwayGoals
Year	1.000000	-0.381332	0.075339
HomeGoals	-0.381332	1.000000	0.012474
AwayGoals	0.075339	0.012474	1.000000

Colab paid products - Cancel contracts here

● ✕