Module 3 - Data Manipulation

In [1]:
```
# Robert Fuss  06/08/2019
import pandas as pd
import numpy as np
```

Delete Duplicate Data

In [2]:
```
#create duplicate data
names = ['Jessica','John','Bob','Jessica','Mary','John','Mel','Mel']
grades = [95,78,76,95,77,78,99,100]
GradeList = list(zip(names,grades))
df = pd.DataFrame(data = GradeList, columns=['Names', 'Grades'])
df
df_shape = str(df.shape)
print (df_shape)
```

```
(8, 2)
```

In [3]:
```
dupe = df.duplicated() #creates list of True/False values
df[dupe] #shows rows where duplicated is True
```

Out[3]:

	Names	Grades
3	Jessica	95
5	John	78

In [4]:
```
#drops duplicate rows, default is to keep first observation
nodupedf = df.drop_duplicates()
nodupedf.count()
```

Out[4]:
```
Names     6
Grades    6
dtype: int64
```

In [8]:
```
#keep last observation
df.drop_duplicates(['Names'], keep='last')
```

Out[8]:

	Names	Grades
2	Bob	76
3	Jessica	95
4	Mary	77
5	John	78
7	Mel	100

In [9]:
```
# RLF Cmd
nodupedf
```

Out[9]:

	Names	Grades
0	Jessica	95
1	John	78
2	Bob	76
4	Mary	77
6	Mel	99
7	Mel	100

Select Rows by a Condition

In [10]:
```
#create fake data

names = ['Bob','Jessica','Mary','John','Mel']
grades = [76,-2,77,78,101]

GradeList = list(zip(names,grades))

df = pd.DataFrame(data = GradeList, columns=['Names', 'Grades'])
df
```

Out[10]:

	Names	Grades
0	Bob	76
1	Jessica	-2
2	Mary	77
3	John	78
4	Mel	101

In [11]:
```
#find rows equal to 101
df.loc[df['Grades'] == 101]
```

Out[11]:

	Names	Grades
4	Mel	101

In [12]:
```
#finds rows where the grade is less than or equal to 100
df.loc[df['Grades'] <= 100]
```

Out[12]:

	Names	Grades
0	Bob	76
1	Jessica	-2
2	Mary	77
3	John	78

In [13]:
```
#multiple conditions
df.loc[(df['Grades'] >= 60) & (df['Grades'] < 100)] #between 60 and 100
```

Out[13]:

	Names	Grades
0	Bob	76
2	Mary	77
3	John	78

In [14]:
```
#replacing values with new value
df.loc[df['Grades'] <= 77, 'Grades'] = 100
df
```

Out[14]:

	Names	Grades
0	Bob	100
1	Jessica	100
2	Mary	100
3	John	78
4	Mel	101

Missing Data

In [15]:
```
df = pd.read_csv("/home/robt/Projects/ITI340_rf/datasets//gradedatamissing.csv")
df.head()
```

Out[15]:

	fname	lname	gender	age	exercise	hours	grade	address
0	Marcia	Pugh	female	17.0	3.0	10.0	82.4	9253 Richardson Road, Matawan, NJ 07747
1	Kadeem	Morrison	male	18.0	4.0	4.0	78.2	33 Spring Dr., Taunton, MA 02780
2	Nash	Powell	male	18.0	5.0	9.0	79.3	41 Hill Avenue, Mentor, OH 44060
3	Noelani	Wagner	female	14.0	2.0	7.0	83.2	8839 Marshall St., Miami, FL 33125
4	Noelani	Cherry	female	18.0	4.0	15.0	87.4	8304 Charles Rd., Lewis Center, OH 43035

In [16]:
```
#Selecting rows with no missing age or gender
df[df['age'].notnull() & df['gender'].notnull()]
```

Out[16]:

	fname	lname	gender	age	exercise	hours	grade	address
0	Marcia	Pugh	female	17.0	3.0	10.0	82.4	9253 Richardson Road, Matawan, NJ 07747
1	Kadeem	Morrison	male	18.0	4.0	4.0	78.2	33 Spring Dr., Taunton, MA 02780
2	Nash	Powell	male	18.0	5.0	9.0	79.3	41 Hill Avenue, Mentor, OH 44060
3	Noelani	Wagner	female	14.0	2.0	7.0	83.2	8839 Marshall St., Miami, FL 33125
4	Noelani	Cherry	female	18.0	4.0	15.0	87.4	8304 Charles Rd., Lewis Center, OH 43035
5	Neil	Whitley	male	16.0	5.0	16.0	88.7	40 Washington Ave., Bloomfield, NJ 07003
6	Nelle	Golden	female	17.0	1.0	9.0	80.2	9768 Hanover Dr., Meadville, PA 16335
7	Armando	Hoffman	male	17.0	5.0	18.0	95.1	360 Manor Drive, Northville, MI 48167
8	Illiana	Rojas	female	15.0	5.0	9.0	76.5	9425 Studebaker Dr., Thibodaux, LA 70301
9	Victor	Richards	male	17.0	2.0	1.0	73.0	123 Main St., Warner Robbins, GA 22222
10	Neil	Wooten	male	15.0	3.0	15.0	89.7	400 Bridge Court, Soddy Daisy, TN 37379
11	Daquan	Alvarez	male	16.0	2.0	13.0	85.2	9028 Arnold Circle, Elizabeth, NJ 07202
12	Nola	Velazquez	female	15.0	2.0	10.0	75.3	72 Bradford Dr., Carlisle, PA 17013
13	Quinn	Warren	female	14.0	4.0	12.0	80.7	760 Smith Street, Appleton, WI 54911
14	Frances	Velasquez	female	15.0	2.0	15.0	84.2	57 Bridge St., Tupelo, MS 38801
15	Lareina	Poole	female	18.0	1.0	14.0	87.6	59 Court Dr., Waxhaw, NC 28173
16	Medge	Mccarthy	female	15.0	1.0	8.0	75.8	609 Warren Court, Prior Lake, MN 55372
17	Kibo	Gates	male	16.0	1.0	10.0	88.2	24 Vernon Street, Helena, MT 59601
18	Libby	Guzman	female	19.0	1.0	19.0	100.0	666 S. Pennington Rd., Dover, NH 03820
19	Shelly	Rosario	female	18.0	4.0	13.0	84.3	571 Miles Street, Flowery Branch, GA 30542
20	Lane	Tate	male	19.0	4.0	11.0	84.2	4 Old Westport St., Glen Burnie, MD 21060
21	Isadora	Case	female	18.0	3.0	11.0	79.1	44 Ocean Lane, Appleton, WI 54911
22	Maggy	Whitfield	female	15.0	1.0	15.0	90.5	2 Henry Ave., Palm Bay, FL 32907
23	Elton	Wagner	male	16.0	2.0	9.0	71.0	98 Indian Spring St., Athens, GA 30605
24	Lance	Benjamin	male	14.0	5.0	18.0	90.3	55 Creek Dr., Lorton, VA 22079
25	Kyle	Skinner	male	17.0	5.0	6.0	82.4	8593 East Branch St., Mooresville, NC 28115
26	Colin	Cohen	male	14.0	1.0	10.0	83.8	23 Lakewood Street, Lake Worth, FL 33460
27	Solomon	Mcpherson	male	15.0	5.0	18.0	94.5	7465 North Pearl St., Massapequa Park, NY 11762
28	Ulla	Warren	female	18.0	1.0	16.0	83.5	89 Fairview Avenue, Hopkins, MN 55343
29	Tyler	Collier	male	16.0	1.0	9.0	69.7	65 Lookout Street, Marshfield, WI 54449
...	...	...	...	...	...	...	...	...
1972	Armando	Mcclure	male	19.0	2.0	14.0	93.2	23 Hall Road, Hagerstown, MD 21740
1973	Haley	Mcgowan	female	17.0	3.0	16.0	90.4	4 Wellington Street, Saint Louis, MO 63109
1974	Fritz	Rojas	male	17.0	5.0	13.0	79.5	16 East Edgewood St., Ashtabula, OH 44004
1975	Allistair	Boyer	male	14.0	3.0	10.0	74.2	9373 Helen Drive, Leland, NC 28451
1976	Ella	Patterson	female	16.0	1.0	8.0	68.0	999 Nicolls Avenue, Oviedo, FL 32765
1977	Felix	Freeman	male	15.0	4.0	11.0	75.2	671 Division Ave., Vineland, NJ 08360
1978	Dean	Oneil	male	16.0	2.0	8.0	74.5	653 East Saxton Lane, Framingham, MA 01701

In [17]:
```
#drop rows with any missing data
df_no_missing = df.dropna()
df_no_missing
```

Out[17]:

	fname	lname	gender	age	exercise	hours	grade	address
0	Marcia	Pugh	female	17.0	3.0	10.0	82.4	9253 Richardson Road, Matawan, NJ 07747
1	Kadeem	Morrison	male	18.0	4.0	4.0	78.2	33 Spring Dr., Taunton, MA 02780
2	Nash	Powell	male	18.0	5.0	9.0	79.3	41 Hill Avenue, Mentor, OH 44060
3	Noelani	Wagner	female	14.0	2.0	7.0	83.2	8839 Marshall St., Miami, FL 33125
4	Noelani	Cherry	female	18.0	4.0	15.0	87.4	8304 Charles Rd., Lewis Center, OH 43035
5	Neil	Whitley	male	16.0	5.0	16.0	88.7	40 Washington Ave., Bloomfield, NJ 07003
6	Nelle	Golden	female	17.0	1.0	9.0	80.2	9768 Hanover Dr., Meadville, PA 16335
7	Armando	Hoffman	male	17.0	5.0	18.0	95.1	360 Manor Drive, Northville, MI 48167
8	Illiana	Rojas	female	15.0	5.0	9.0	76.5	9425 Studebaker Dr., Thibodaux, LA 70301
9	Victor	Richards	male	17.0	2.0	1.0	73.0	123 Main St., Warner Robbins, GA 22222
10	Neil	Wooten	male	15.0	3.0	15.0	89.7	400 Bridge Court, Soddy Daisy, TN 37379
11	Daquan	Alvarez	male	16.0	2.0	13.0	85.2	9028 Arnold Circle, Elizabeth, NJ 07202
12	Nola	Velazquez	female	15.0	2.0	10.0	75.3	72 Bradford Dr., Carlisle, PA 17013
13	Quinn	Warren	female	14.0	4.0	12.0	80.7	760 Smith Street, Appleton, WI 54911
14	Frances	Velasquez	female	15.0	2.0	15.0	84.2	57 Bridge St., Tupelo, MS 38801
15	Lareina	Poole	female	18.0	1.0	14.0	87.6	59 Court Dr., Waxhaw, NC 28173
16	Medge	Mccarthy	female	15.0	1.0	8.0	75.8	609 Warren Court, Prior Lake, MN 55372
17	Kibo	Gates	male	16.0	1.0	10.0	88.2	24 Vernon Street, Helena, MT 59601
18	Libby	Guzman	female	19.0	1.0	19.0	100.0	666 S. Pennington Rd., Dover, NH 03820
19	Shelly	Rosario	female	18.0	4.0	13.0	84.3	571 Miles Street, Flowery Branch, GA 30542
20	Lane	Tate	male	19.0	4.0	11.0	84.2	4 Old Westport St., Glen Burnie, MD 21060
21	Isadora	Case	female	18.0	3.0	11.0	79.1	44 Ocean Lane, Appleton, WI 54911
22	Maggy	Whitfield	female	15.0	1.0	15.0	90.5	2 Henry Ave., Palm Bay, FL 32907
23	Elton	Wagner	male	16.0	2.0	9.0	71.0	98 Indian Spring St., Athens, GA 30605
24	Lance	Benjamin	male	14.0	5.0	18.0	90.3	55 Creek Dr., Lorton, VA 22079
25	Kyle	Skinner	male	17.0	5.0	6.0	82.4	8593 East Branch St., Mooresville, NC 28115
26	Colin	Cohen	male	14.0	1.0	10.0	83.8	23 Lakewood Street, Lake Worth, FL 33460
27	Solomon	Mcpherson	male	15.0	5.0	18.0	94.5	7465 North Pearl St., Massapequa Park, NY 11762
28	Ulla	Warren	female	18.0	1.0	16.0	83.5	89 Fairview Avenue, Hopkins, MN 55343
29	Tyler	Collier	male	16.0	1.0	9.0	69.7	65 Lookout Street, Marshfield, WI 54449
...	...	...	...	...	...	...	...	...
1972	Armando	Mcclure	male	19.0	2.0	14.0	93.2	23 Hall Road, Hagerstown, MD 21740
1973	Haley	Mcgowan	female	17.0	3.0	16.0	90.4	4 Wellington Street, Saint Louis, MO 63109
1974	Fritz	Rojas	male	17.0	5.0	13.0	79.5	16 East Edgewood St., Ashtabula, OH 44004
1975	Allistair	Boyer	male	14.0	3.0	10.0	74.2	9373 Helen Drive, Leland, NC 28451
1976	Ella	Patterson	female	16.0	1.0	8.0	68.0	999 Nicolls Avenue, Oviedo, FL 32765
1977	Felix	Freeman	male	15.0	4.0	11.0	75.2	671 Division Ave., Vineland, NJ 08360
1978	Dean	Oneil	male	16.0	2.0	8.0	74.5	653 East Saxton Lane, Framingham, MA 01701

In [18]:
```
#replace empty cells with 0
df.fillna(0)
```

Out[18]:

	fname	lname	gender	age	exercise	hours	grade	address
0	Marcia	Pugh	female	17.0	3.0	10.0	82.4	9253 Richardson Road, Matawan, NJ 07747
1	Kadeem	Morrison	male	18.0	4.0	4.0	78.2	33 Spring Dr., Taunton, MA 02780
2	Nash	Powell	male	18.0	5.0	9.0	79.3	41 Hill Avenue, Mentor, OH 44060
3	Noelani	Wagner	female	14.0	2.0	7.0	83.2	8839 Marshall St., Miami, FL 33125
4	Noelani	Cherry	female	18.0	4.0	15.0	87.4	8304 Charles Rd., Lewis Center, OH 43035
5	Neil	Whitley	male	16.0	5.0	16.0	88.7	40 Washington Ave., Bloomfield, NJ 07003
6	Nelle	Golden	female	17.0	1.0	9.0	80.2	9768 Hanover Dr., Meadville, PA 16335
7	Armando	Hoffman	male	17.0	5.0	18.0	95.1	360 Manor Drive, Northville, MI 48167
8	Illiana	Rojas	female	15.0	5.0	9.0	76.5	9425 Studebaker Dr., Thibodaux, LA 70301
9	Victor	Richards	male	17.0	2.0	1.0	73.0	123 Main St., Warner Robbins, GA 22222
10	Neil	Wooten	male	15.0	3.0	15.0	89.7	400 Bridge Court, Soddy Daisy, TN 37379
11	Daquan	Alvarez	male	16.0	2.0	13.0	85.2	9028 Arnold Circle, Elizabeth, NJ 07202
12	Nola	Velazquez	female	15.0	2.0	10.0	75.3	72 Bradford Dr., Carlisle, PA 17013
13	Quinn	Warren	female	14.0	4.0	12.0	80.7	760 Smith Street, Appleton, WI 54911
14	Frances	Velasquez	female	15.0	2.0	15.0	84.2	57 Bridge St., Tupelo, MS 38801
15	Lareina	Poole	female	18.0	1.0	14.0	87.6	59 Court Dr., Waxhaw, NC 28173
16	Medge	Mccarthy	female	15.0	1.0	8.0	75.8	609 Warren Court, Prior Lake, MN 55372
17	Kibo	Gates	male	16.0	1.0	10.0	88.2	24 Vernon Street, Helena, MT 59601
18	Libby	Guzman	female	19.0	1.0	19.0	100.0	666 S. Pennington Rd., Dover, NH 03820
19	Shelly	Rosario	female	18.0	4.0	13.0	84.3	571 Miles Street, Flowery Branch, GA 30542
20	Lane	Tate	male	19.0	4.0	11.0	84.2	4 Old Westport St., Glen Burnie, MD 21060
21	Isadora	Case	female	18.0	3.0	11.0	79.1	44 Ocean Lane, Appleton, WI 54911
22	Maggy	Whitfield	female	15.0	1.0	15.0	90.5	2 Henry Ave., Palm Bay, FL 32907
23	Elton	Wagner	male	16.0	2.0	9.0	71.0	98 Indian Spring St., Athens, GA 30605
24	Lance	Benjamin	male	14.0	5.0	18.0	90.3	55 Creek Dr., Lorton, VA 22079
25	Kyle	Skinner	male	17.0	5.0	6.0	82.4	8593 East Branch St., Mooresville, NC 28115
26	Colin	Cohen	male	14.0	1.0	10.0	83.8	23 Lakewood Street, Lake Worth, FL 33460
27	Solomon	Mcpherson	male	15.0	5.0	18.0	94.5	7465 North Pearl St., Massapequa Park, NY 11762
28	Ulla	Warren	female	18.0	1.0	16.0	83.5	89 Fairview Avenue, Hopkins, MN 55343
29	Tyler	Collier	male	16.0	1.0	9.0	69.7	65 Lookout Street, Marshfield, WI 54449
...	...	...	...	...	...	...	...	...
1972	Armando	Mcclure	male	19.0	2.0	14.0	93.2	23 Hall Road, Hagerstown, MD 21740
1973	Haley	Mcgowan	female	17.0	3.0	16.0	90.4	4 Wellington Street, Saint Louis, MO 63109
1974	Fritz	Rojas	male	17.0	5.0	13.0	79.5	16 East Edgewood St., Ashtabula, OH 44004
1975	Allistair	Boyer	male	14.0	3.0	10.0	74.2	9373 Helen Drive, Leland, NC 28451
1976	Ella	Patterson	female	16.0	1.0	8.0	68.0	999 Nicolls Avenue, Oviedo, FL 32765
1977	Felix	Freeman	male	15.0	4.0	11.0	75.2	671 Division Ave., Vineland, NJ 08360
1978	Dean	Oneil	male	16.0	2.0	8.0	74.5	653 East Saxton Lane, Framingham, MA 01701

In [19]:
```
#replace empty cells with average of column
df["grade"].fillna(df["grade"].mean(), inplace=True)
```

Binning Data

In [20]:
```
filename = "/home/robt/Projects/ITI340_rf/datasets/gradedata.csv"
df = pd.read_csv(filename)

df.head()
```

Out[20]:

	fname	lname	gender	age	exercise	hours	grade	address
0	Marcia	Pugh	female	17	3	10	82.4	9253 Richardson Road, Matawan, NJ 07747
1	Kadeem	Morrison	male	18	4	4	78.2	33 Spring Dr., Taunton, MA 02780
2	Nash	Powell	male	18	5	9	79.3	41 Hill Avenue, Mentor, OH 44060
3	Noelani	Wagner	female	14	2	7	83.2	8839 Marshall St., Miami, FL 33125
4	Noelani	Cherry	female	18	4	15	87.4	8304 Charles Rd., Lewis Center, OH 43035

In [21]:
```
df['grade'].dtypes
```

Out[21]:
```
dtype('float64')
```

In [22]:
```
#Define bins as 0-59, 60-69, 70-79, 80-89, 90-100
bins = [0, 60, 70, 80, 90, 100]

# Create names for the four groups
group_names = ['F', 'D', 'C', 'B', 'A']

#make new column with letter grades
df['lettergrade'] = pd.cut(df['grade'], bins, labels=group_names)
df.head()
```

Out[22]:

	fname	lname	gender	age	exercise	hours	grade	address	lettergrade
0	Marcia	Pugh	female	17	3	10	82.4	9253 Richardson Road, Matawan, NJ 07747	B
1	Kadeem	Morrison	male	18	4	4	78.2	33 Spring Dr., Taunton, MA 02780	C
2	Nash	Powell	male	18	5	9	79.3	41 Hill Avenue, Mentor, OH 44060	C
3	Noelani	Wagner	female	14	2	7	83.2	8839 Marshall St., Miami, FL 33125	B
4	Noelani	Cherry	female	18	4	15	87.4	8304 Charles Rd., Lewis Center, OH 43035	B

In [23]:
```
pd.value_counts(df['lettergrade']) # number of observations per letter grade
```

Out[23]:
```
B    737
C    580
A    475
D    193
F     15
Name: lettergrade, dtype: int64
```

In [24]:
```
df.groupby('lettergrade')['hours'].mean() #mean of hours studied per letter grade
```

Out[24]:
```
lettergrade
F     3.933333
D     5.544041
C     8.381034
B    11.827680
A    15.305263
Name: hours, dtype: float64
```

In [25]:
```
#apply functions to data
df['grade'] = df['grade'].apply(lambda x: int(x)) #turns float type grades to int type
df.head()
```

Out[25]:

	fname	lname	gender	age	exercise	hours	grade	address	lettergrade
0	Marcia	Pugh	female	17	3	10	82	9253 Richardson Road, Matawan, NJ 07747	B
1	Kadeem	Morrison	male	18	4	4	78	33 Spring Dr., Taunton, MA 02780	C
2	Nash	Powell	male	18	5	9	79	41 Hill Avenue, Mentor, OH 44060	C
3	Noelani	Wagner	female	14	2	7	83	8839 Marshall St., Miami, FL 33125	B
4	Noelani	Cherry	female	18	4	15	87	8304 Charles Rd., Lewis Center, OH 43035	B

Add and Remove Columns

In [26]:
```
names = ['Bob','Jessica','Mary','John','Mel']
grades = [76,95,77,78,99]
bsdegress = [1,1,0,0,1]
msdegrees = [2,1,0,0,0]
phddegrees = [0,1,0,0,0]

GradeList = list(zip(names,grades,bsdegress,msdegrees,phddegrees))

df = pd.DataFrame(data = GradeList, columns=['Names','Grades','BS','MS','PhD'])
df
```

Out[26]:

	Names	Grades	BS	MS	PhD
0	Bob	76	1	2	0
1	Jessica	95	1	1	1
2	Mary	77	0	0	0
3	John	78	0	0	0
4	Mel	99	1	0	0

In [27]:
```
#make a new column with each value = 0
df['HighSchool']=0

#make new column with all NaN values
df['Preschool'] = np.nan

#manually add data for new column
d = ([0,1,0,1,0])
s = pd.Series(d, index= df.index)
df['DriversLicense'] = s

df
```

Out[27]:

	Names	Grades	BS	MS	PhD	HighSchool	Preschool	DriversLicense
0	Bob	76	1	2	0	0	NaN	0
1	Jessica	95	1	1	1	0	NaN	1
2	Mary	77	0	0	0	0	NaN	0
3	John	78	0	0	0	0	NaN	1
4	Mel	99	1	0	0	0	NaN	0

In [28]:
```
#drop a column
df.drop('PhD', axis=1)
```

Out[28]:

	Names	Grades	BS	MS	HighSchool	Preschool	DriversLicense
0	Bob	76	1	2	0	NaN	0
1	Jessica	95	1	1	0	NaN	1
2	Mary	77	0	0	0	NaN	0
3	John	78	0	0	0	NaN	1
4	Mel	99	1	0	0	NaN	0

In [29]:
```
#drop completely empty columns (all NaN/null values)
df.dropna(axis=1, how='all')
```

Out[29]:

	Names	Grades	BS	MS	PhD	HighSchool	DriversLicense
0	Bob	76	1	2	0	0	0
1	Jessica	95	1	1	1	0	1
2	Mary	77	0	0	0	0	0
3	John	78	0	0	0	0	1
4	Mel	99	1	0	0	0	0

Make new columns

In [30]:
```
#using functions
filename = "/home/robt/Projects/ITI340_rf/datasets/gradedata.csv"
df = pd.read_csv(filename)

def singlename(fn, ln):
    return fn + " " + ln

df['fullname'] = singlename(df['fname'], df['lname'])
df.head()
```

Out[30]:

	fname	lname	gender	age	exercise	hours	grade	address	fullname
0	Marcia	Pugh	female	17	3	10	82.4	9253 Richardson Road, Matawan, NJ 07747	Marcia Pugh
1	Kadeem	Morrison	male	18	4	4	78.2	33 Spring Dr., Taunton, MA 02780	Kadeem Morrison
2	Nash	Powell	male	18	5	9	79.3	41 Hill Avenue, Mentor, OH 44060	Nash Powell
3	Noelani	Wagner	female	14	2	7	83.2	8839 Marshall St., Miami, FL 33125	Noelani Wagner
4	Noelani	Cherry	female	18	4	15	87.4	8304 Charles Rd., Lewis Center, OH 43035	Noelani Cherry

In [31]:
```
#create column based on binary conditional
df['isFailing'] = np.where(df['grade']<70, 'yes', 'no')
df.tail()
```

Out[31]:

	fname	lname	gender	age	exercise	hours	grade	address	fullname	isFailing
1995	Cody	Shepherd	male	19	1	8	80.1	982 West Street, Alexandria, VA 22304	Cody Shepherd	no
1996	Geraldine	Peterson	female	16	4	18	100.0	78 Morris Street, East Northport, NY 11731	Geraldine Peterson	no
1997	Mercedes	Leon	female	18	3	14	84.9	30 Glenridge Rd., Bountiful, UT 84010	Mercedes Leon	no
1998	Lucius	Rowland	male	16	1	7	69.1	342 West Meadowbrook Lane, Helena, MT 59601	Lucius Rowland	yes
1999	Linus	Morris	male	19	4	10	79.6	81 Homestead Drive, Voorhees, NJ 08043	Linus Morris	no

In [32]:
```
#make numeric value column based on string value column

#create a function that will return 1 if female and 0 if male
def score_to_numeric(x):
    if x=='female':
        return 1
    if x=='male':
        return 0

df['gender_val'] = df['gender'].apply(score_to_numeric)
df.tail()
```

Out[32]:

	fname	lname	gender	age	exercise	hours	grade	address	fullname	isFailing	gender_va
1995	Cody	Shepherd	male	19	1	8	80.1	982 West Street, Alexandria, VA 22304	Cody Shepherd	no	0
1996	Geraldine	Peterson	female	16	4	18	100.0	78 Morris Street, East Northport, NY 11731	Geraldine Peterson	no	1
1997	Mercedes	Leon	female	18	3	14	84.9	30 Glenridge Rd., Bountiful, UT 84010	Mercedes Leon	no	1
1998	Lucius	Rowland	male	16	1	7	69.1	342 West Meadowbrook Lane, Helena, MT 59601	Lucius Rowland	yes	0
1999	Linus	Morris	male	19	4	10	79.6	81 Homestead Drive, Voorhees, NJ 08043	Linus Morris	no	0

Sort Values

In [33]:
```
#still using gradedata.csv
#df.head() if needed

df.sort_values('hours') #default is ascending=True
#does not change structure of df
```

Out[33]:

	fname	lname	gender	age	exercise	hours	grade	address	fullname	isFailing	gende
1611	Garrett	Everett	male	17	2	0	78.1	7963 South Thorne Drive, Sterling, VA 20164	Garrett Everett	no	
1213	Arthur	Nieves	male	14	3	0	76.8	377 Beech Street, Chesterfield, VA 23832	Arthur Nieves	no	
356	Luke	Santana	male	17	1	0	75.0	2 Sunbeam Dr., Zionsville, IN 46077	Luke Santana	no	
80	Hoyt	Lyons	male	16	5	1	71.3	720 Linden Street, Springfield Gardens, NY 11413	Hoyt Lyons	no	
1507	Ferris	Avila	male	18	5	1	74.2	85 Brown Drive, Pensacola, FL 32503	Ferris Avila	no	
1931	Luke	Morrison	male	19	5	2	70.1	471 Windfall Ave., Santa Monica, CA 90403	Luke Morrison	no	
1317	Kessie	Mcdonald	female	17	1	2	62.2	311 Pierce St., Eugene, OR 97402	Kessie Mcdonald	yes	
1359	Jin	Valdez	male	19	2	2	77.0	604 Big Rock Cove Drive, Gibsonia, PA 15044	Jin Valdez	no	
1110	Xerxes	Mooney	female	17	1	2	70.4	974 Theatre Dr., Arvada, CO 80003	Xerxes Mooney	no	
903	Lee	Barber	male	15	2	2	62.3	449 Pearl Street, Largo, FL 33771	Lee Barber	yes	
1916	Kevin	Townsend	male	14	5	2	60.1	9805 Walnutwood Dr., Panama City, FL 32404	Kevin Townsend	yes	
836	Abel	Stephens	male	18	4	2	68.2	7 Bow Ridge Ave., Lansing, MI 48910	Abel Stephens	yes	
1640	Nolan	Small	male	14	3	2	71.5	9205 Myers Drive, Ellicott City, MD 21042	Nolan Small	no	
700	Steven	Sherman	male	18	1	2	60.0	8029 Depot Street, Port Charlotte, FL 33952	Steven Sherman	yes	
410	Linda	Baldwin	female	16	5	2	59.0	970 SW. Second Ave., Cedar Falls, IA 50613	Linda Baldwin	yes	
								7986 Briarwood			

Calculate and Remove Outliers

In [34]:
```
#still using gradedata.csv
#check df.head() if needed

#Standard Deviation Method

meangrade = df['grade'].mean()
stdgrade = df['grade'].std()
toprange = meangrade + stdgrade * 1.96
botrange = meangrade - stdgrade * 1.96

newdf = df.copy() #to not mess up the original df
newdf = newdf.drop(newdf[newdf['grade'] > toprange].index)
newdf = newdf.drop(newdf[newdf['grade'] < botrange].index)

newdf.head()
```

Out[34]:

	fname	lname	gender	age	exercise	hours	grade	address	fullname	isFailing	gender_val
0	Marcia	Pugh	female	17	3	10	82.4	9253 Richardson Road, Matawan, NJ 07747	Marcia Pugh	no	1
1	Kadeem	Morrison	male	18	4	4	78.2	33 Spring Dr., Taunton, MA 02780	Kadeem Morrison	no	0
2	Nash	Powell	male	18	5	9	79.3	41 Hill Avenue, Mentor, OH 44060	Nash Powell	no	0
3	Noelani	Wagner	female	14	2	7	83.2	8839 Marshall St., Miami, FL 33125	Noelani Wagner	no	1
4	Noelani	Cherry	female	18	4	15	87.4	8304 Charles Rd., Lewis Center, OH 43035	Noelani Cherry	no	1

In [35]:
```
#Interquartile Range Method

q1 = df['grade'].quantile(.25)
q3 = df['grade'].quantile(.75)
iqr = q3-q1
toprange = q3 + iqr * 1.5
botrange = q1 - iqr * 1.5

newdf = df.copy()
newdf = newdf.drop(newdf[newdf['grade'] > toprange].index)
newdf = newdf.drop(newdf[newdf['grade'] < botrange].index)

newdf.head()
```

Out[35]:

	fname	lname	gender	age	exercise	hours	grade	address	fullname	isFailing	gender_val
0	Marcia	Pugh	female	17	3	10	82.4	9253 Richardson Road, Matawan, NJ 07747	Marcia Pugh	no	1
1	Kadeem	Morrison	male	18	4	4	78.2	33 Spring Dr., Taunton, MA 02780	Kadeem Morrison	no	0
2	Nash	Powell	male	18	5	9	79.3	41 Hill Avenue, Mentor, OH 44060	Nash Powell	no	0
3	Noelani	Wagner	female	14	2	7	83.2	8839 Marshall St., Miami, FL 33125	Noelani Wagner	no	1
4	Noelani	Cherry	female	18	4	15	87.4	8304 Charles Rd., Lewis Center, OH 43035	Noelani Cherry	no	1

In [38]:
```
q1
```

Out[38]: 75.57499999999999

In [39]:
```
q3
```

Out[39]: 89.7

In [40]:
```
iqr
```

Out[40]: 14.125000000000014

In [41]:
```
toprange
```

Out[41]: 110.88750000000002

In [42]:
```
botrange
```

Out[42]: 54.38749999999997

In [ ]:
```

```