Module 3 Exercises - Data Manipulation

Exercise 1:

From the datasets folder, load in the "dupedata.csv" file as a dataframe. Drop the duplicates from the dataframe, keeping the first value (save the resulting dataframe to a new variable).

In [195]:

```
import pandas as pd
import numpy as np
```

In [196]:

```
import os
os.getcwd()
```

Out[196]:

```
'C:\\Users\\GBTC406001ur\\Downloads'
```

In [197]:

```
df = pd.read_csv("datasets/dupedata.csv")
df.head()
```

Out[197]:

	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 [198]:

```
df.count()
```

Out[198]:

```
fname       2038
lname       2038
gender      2038
age         2038
exercise    2038
hours       2038
grade       2038
address     2038
dtype: int64
```

In [199]:

```
df2 = df.drop_duplicates()
```

In [200]:
```
df2.head()
```

Out[200]:

	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 [201]:
```
df2.count()
```

Out[201]:
```
fname       2000
lname       2000
gender      2000
age         2000
exercise    2000
hours       2000
grade       2000
address     2000
dtype: int64
```

In [202]:
```
#keeping the last value
df3 = df.drop_duplicates(['fname'], keep='last')
```

In [203]:
```
df3.count()
```

Out[203]:
```
fname       958
lname       958
gender      958
age         958
exercise    958
hours       958
grade       958
address     958
dtype: int64
```

Exercise 2:

Using the dataframe in the previous exercise, select all the rows where students received a grade lower than 60 (they need a teacher conference on how to improve for the next test).

In [204]:

```
#finds rows where the grade is less than or equal to 100
df2.loc[df2['grade'] == 100, 'grade'] = 103
df2
```

Out[204]:

	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
5	Neil	Whitley	male	16	5	16	88.7	40 Washington Ave., Bloomfield, NJ 07003
6	Nelle	Golden	female	17	1	9	80.2	9768 Hanover Dr., Meadville, PA 16335
7	Armando	Hoffman	male	17	5	18	95.1	360 Manor Drive, Northville, MI 48167
8	Illiana	Rojas	female	15	5	9	76.5	9425 Studebaker Dr., Thibodaux, LA 70301
9	Neil	Wooten	male	15	3	15	89.7	400 Bridge Court, Soddy Daisy, TN 37379
10	Daquan	Alvarez	male	16	2	13	85.2	9028 Arnold Circle, Elizabeth, NJ 07202
11	Nola	Velazquez	female	15	2	10	75.3	72 Bradford Dr., Carlisle, PA 17013
12	Quinn	Warren	female	14	4	12	80.7	760 Smith Street, Appleton, WI 54911
13	Frances	Velasquez	female	15	2	15	84.2	57 Bridge St., Tupelo, MS 38801
14	Lareina	Poole	female	18	1	14	87.6	59 Court Dr., Waxhaw, NC 28173
15	Medge	Mccarthy	female	15	1	8	75.8	609 Warren Court, Prior Lake, MN 55372
16	Kibo	Gates	male	16	1	10	88.2	24 Vernon Street, Helena, MT 59601
17	Libby	Guzman	female	19	1	19	103.0	666 S. Pennington Rd., Dover, NH 03820
18	Shelly	Rosario	female	18	4	13	84.3	571 Miles Street, Flowery Branch, GA 30542
19	Lane	Tate	male	19	4	11	84.2	4 Old Westport St., Glen Burnie, MD 21060
20	Isadora	Case	female	18	3	11	79.1	44 Ocean Lane, Appleton, WI 54911
21	Maggy	Whitfield	female	15	1	15	90.5	2 Henry Ave., Palm Bay, FL 32907
22	Elton	Wagner	male	16	2	9	71.0	98 Indian Spring St., Athens, GA 30605

	fname	lname	gender	age	exercise	hours	grade	address
23	Lance	Benjamin	male	14	5	18	90.3	55 Creek Dr., Lorton, VA 22079
24	Kyle	Skinner	male	17	5	6	82.4	8593 East Branch St., Mooresville, NC 28115
25	Colin	Cohen	male	14	1	10	83.8	23 Lakewood Street, Lake Worth, FL 33460
26	Solomon	Mcpherson	male	15	5	18	94.5	7465 North Pearl St., Massapequa Park, NY 11762
27	Ulla	Warren	female	18	1	16	83.5	89 Fairview Avenue, Hopkins, MN 55343
28	Tyler	Collier	male	16	1	9	69.7	65 Lookout Street, Marshfield, WI 54449
29	Emma	Mccall	female	16	2	13	91.1	854 Sussex Street, Westford, MA 01886
...	...	...	...	...	...	...	...	...
1976	Armando	Mcclure	male	19	2	14	93.2	23 Hall Road, Hagerstown, MD 21740
1977	Haley	Mcgowan	female	17	3	16	90.4	4 Wellington Street, Saint Louis, MO 63109
1978	Fritz	Rojas	male	17	5	13	79.5	16 East Edgewood St., Ashtabula, OH 44004
1979	Allistair	Boyer	male	14	3	10	74.2	9373 Helen Drive, Leland, NC 28451
1980	Ella	Patterson	female	16	1	8	68.0	999 Nicolls Avenue, Oviedo, FL 32765
1981	Felix	Freeman	male	15	4	11	75.2	671 Division Ave., Vineland, NJ 08360
1982	Dean	Oneil	male	16	2	8	74.5	653 East Saxton Lane, Framingham, MA 01701
1983	Quinlan	Hawkins	male	14	3	18	85.6	9437 Longbranch Street, Rockville, MD 20850
1984	Kiara	Lott	female	17	2	12	82.7	7110 Ridge Road, Jacksonville, NC 28540
1985	Kai	Woodard	female	17	5	17	103.0	37 Addison St., Eastpointe, MI 48021
1986	Maxine	Raymond	female	17	4	19	91.9	739 Thomas Court, Memphis, TN 38106
1987	Leah	Lawrence	female	15	5	15	103.0	652 Pine Drive, Mountain View, CA 94043
1988	Kalia	Lewis	female	18	1	6	61.4	557 Theatre Lane, Green Cove Springs, FL 32043
1989	Anthony	Palmer	male	14	1	14	93.1	914 NW. Lawrence Street, Lawrenceville, GA 30043
1990	Joshua	Randolph	male	15	1	4	72.7	9632 NE. Lakeshore St., Riverside, NJ 08075
1991	Akeem	Luna	male	15	4	6	74.4	750 Adams Drive, Goldsboro, NC 27530

	fname	lname	gender	age	exercise	hours	grade	address
1992	Indigo	Mccoy	female	19	2	14	91.3	9652 Columbia Ave., Chattanooga, TN 37421
1993	Price	Wall	male	15	5	11	78.6	8672 S. 53rd Drive, Waterford, MI 48329
1994	Quinn	Patterson	male	15	1	14	78.0	634 Cedar Swamp Ave., Burbank, IL 60459
1995	John	Ford	male	14	2	14	91.2	64 Devonshire Street, Orange Park, FL 32065
1996	Adena	Battle	female	17	2	8	70.2	9272 Elizabeth Drive, Londonderry, NH 03053
1997	Craig	Obrien	male	16	3	7	64.9	524 Park Ave., Hollywood, FL 33020
1998	Isabelle	Barber	female	14	5	9	78.5	955 Glen Ridge Rd., Plattsburgh, NY 12901
1999	Risa	Watson	female	14	2	10	74.3	37 Augusta Lane, Montgomery Village, MD 20886
2000	Emerson	Gill	male	17	5	5	67.5	75 Wild Horse Street, Panama City, FL 32404
2001	Cody	Shepherd	male	19	1	8	80.1	982 West Street, Alexandria, VA 22304
2002	Geraldine	Peterson	female	16	4	18	103.0	78 Morris Street, East Northport, NY 11731
2003	Mercedes	Leon	female	18	3	14	84.9	30 Glenridge Rd., Bountiful, UT 84010
2004	Lucius	Rowland	male	16	1	7	69.1	342 West Meadowbrook Lane, Helena, MT 59601
2005	Linus	Morris	male	19	4	10	79.6	81 Homestead Drive, Voorhees, NJ 08043

2000 rows × 8 columns

```
In [205]: df2.loc[df2['grade'] == 103]
```

Out[205]:

	fname	lname	gender	age	exercise	hours	grade	address
17	Libby	Guzman	female	19	1	19	103.0	666 S. Pennington Rd., Dover, NH 03820
36	Ivor	Arnold	male	19	4	20	103.0	7027 Magnolia Dr., Catonsville, MD 21228
77	Quemby	Justice	female	14	2	17	103.0	346 Birchpond Court, Decatur, GA 30030
101	Sage	Cleveland	female	19	2	20	103.0	9721B Green Dr., Fairhope, AL 36532
109	Sophia	Gordon	female	19	4	17	103.0	7672 Smoky Hollow Street, Hillsboro, OR 97124
165	Ciaran	Johns	male	16	4	15	103.0	7350 Creek Avenue, Upper Marlboro, MD 20772
226	Uriah	Cummings	male	18	3	20	103.0	444 West Homestead Rd., Lebanon, PA 17042
249	Phyllis	Walters	female	18	4	19	103.0	33 Smith St., Hephzibah, GA 30815
252	Rose	Middleton	female	15	2	20	103.0	979 Andover Street, Cockeysville, MD 21030
303	Kamal	Walton	male	14	1	17	103.0	9125 Edgemont Lane, Attleboro, MA 02703
321	Lysandra	Copeland	female	19	5	19	103.0	20 Talbot Drive, Fort Lauderdale, FL 33308
325	Thor	Ramos	male	17	3	18	103.0	208 Plymouth St., Grove City, OH 43123
355	Kadeem	Marshall	male	17	5	18	103.0	56 North Glen Creek St., San Lorenzo, CA 94580
382	Jacob	Gray	male	15	3	17	103.0	8323 Alton Court, Clementon, NJ 08021
434	Kaden	Grant	female	17	5	14	103.0	73 SW. Leeton Ridge Road, Homestead, FL 33030
457	Myra	Parrish	female	19	5	15	103.0	854 3rd Ave., Nanuet, NY 10954
459	Inez	Stephenson	female	17	5	17	103.0	51 Academy St., Roselle, IL 60172
473	Declan	Manning	male	18	4	16	103.0	7993 E. Harvey Ave., Springfield, PA 19064
508	Dai	Osborne	female	19	5	20	103.0	2 Pilgrim Road, Alexandria, VA 22304
610	Kiara	Singleton	female	16	4	19	103.0	9800 Atlantic St., Whitestone, NY 11357
616	Ali	Franks	male	18	5	20	103.0	78 Essex Road, Stillwater, MN 55082
621	Gwendolyn	Vazquez	female	19	5	19	103.0	550 Selby St., Louisville, KY 40207

	fname	lname	gender	age	exercise	hours	grade	address
644	Drake	Winters	male	17	4	19	103.0	59 Greenrose St., Conyers, GA 30012
661	Camille	Barr	female	14	4	14	103.0	9665 S. Spring Rd., Kaukauna, WI 54130
690	Sylvester	Payne	male	17	5	20	103.0	242 Goldfield Ave., Encino, CA 91316
698	Ruth	Bowman	female	18	3	18	103.0	8621 Shub Farm Ave., Ocean Springs, MS 39564
701	Janna	Moran	female	17	5	17	103.0	4 Belmont St., Fremont, OH 43420
757	Erich	Stone	male	14	5	17	103.0	694 W. Heritage Road, Bedford, OH 44146
767	Lila	Wall	female	17	5	18	103.0	546 Fulton Lane, Warren, MI 48089
787	Josephine	Rivers	female	18	2	17	103.0	8867 Jackson Dr., Newark, NJ 07103
...	...	...	...	...	...	...	...	...
1398	Halla	Vang	female	16	3	17	103.0	9124 Fairway Road, Dawsonville, GA 30534
1421	Harriet	Long	female	17	3	17	103.0	797 Rockcrest Avenue, Lakewood, NJ 08701
1426	Shaine	Mcleod	female	16	2	18	103.0	8400 Brickell Drive, Clayton, NC 27520
1481	September	Norris	female	15	1	20	103.0	326 Birch Hill Street, Kent, OH 44240
1545	Asher	Guerrero	male	14	4	15	103.0	28 North Charles Road, Lithonia, GA 30038
1605	Clio	Glover	female	15	4	16	103.0	405 St Margarets Drive, Tampa, FL 33604
1629	Cassady	Ruiz	female	16	2	18	103.0	8233 Thomas Ave., North Miami Beach, FL 33160
1634	Blossom	Gonzalez	female	17	4	18	103.0	614 Locust Street, Winter Springs, FL 32708
1639	Colorado	Cash	male	16	5	18	103.0	48 King Circle, Griffin, GA 30223
1649	Igor	Conway	male	17	1	17	103.0	7859 Carriage Rd., Wilmington, MA 01887
1650	Denise	Sloan	female	18	4	20	103.0	36 Bowman Dr., Fort Dodge, IA 50501
1659	Honorato	Gutierrez	male	16	5	17	103.0	64 Mill Pond Street, Panama City, FL 32404
1665	Avram	Huber	male	18	3	20	103.0	36 S. Pacific Ave., Palm City, FL 34990
1680	Clayton	Yates	male	16	5	14	103.0	9528 Miller Drive, Klamath Falls, OR 97603
1719	Gay	Carlson	female	18	2	20	103.0	805 Lakeview Avenue, Villa Park, IL 60181

	fname	lname	gender	age	exercise	hours	grade	address
1724	Carol	Dillard	female	18	5	17	103.0	929 Pilgrim Road, Venice, FL 34293
1734	Naomi	Strong	female	15	2	20	103.0	507 E. Fifth Lane, Natick, MA 01760
1840	Phillip	Savage	male	19	2	19	103.0	579 High Ridge Rd., Matawan, NJ 07747
1841	Maxwell	Blake	male	15	2	17	103.0	30 Hawthorne Ave., Norfolk, VA 23503
1842	Louis	Joyner	male	19	2	19	103.0	7875 Tarkiln Hill Court, Windermere, FL 34786
1844	Vera	Russell	female	18	4	16	103.0	122 S. Pulaski St., Wausau, WI 54401
1850	Yael	Hatfield	female	19	2	18	103.0	9612 Sherman Avenue, Elk River, MN 55330
1859	Murphy	Michael	male	14	1	19	103.0	88 Garden Street, Union, NJ 07083
1899	Juliet	Good	female	14	5	15	103.0	69 Lake Forest Lane, Midlothian, VA 23112
1928	Phelan	Frye	male	16	5	19	103.0	9614 Winding Way St., Stone Mountain, GA 30083
1939	Adena	Robinson	female	14	3	19	103.0	575 Beech Street, Upper Marlboro, MD 20772
1951	Brianna	Holloway	female	18	4	18	103.0	8493 Locust Ave., Longwood, FL 32779
1985	Kai	Woodard	female	17	5	17	103.0	37 Addison St., Eastpointe, MI 48021
1987	Leah	Lawrence	female	15	5	15	103.0	652 Pine Drive, Mountain View, CA 94043
2002	Geraldine	Peterson	female	16	4	18	103.0	78 Morris Street, East Northport, NY 11731

87 rows × 8 columns

In [206]:

```
df3 = df2.loc[df2['fname'] == 'Noelani']
df3.head()
```

Out[206]:

	fname	lname	gender	age	exercise	hours	grade	address
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
527	Noelani	Villarreal	female	19	1	5	75.4	276 East Oxford Street, Lincolnton, NC 28092

In [207]:

```
df3 = df2.drop_duplicates(['fname'], keep='last')
```

In [208]:

```
df3.loc[df2['fname']=='Noelani']
```

Out[208]:

```
       fname      lname  gender  age  exercise  hours  grade                                     address
527  Noelani  Villarreal  female   19         1      5   75.4  276 East Oxford Street, Lincolnton, NC
                                                                                                 28092
```

Exercise 3:

Using the dataframe from Exercise 1, select all the rows where a student received a grade of 100 and change their grade to 103 (extra credit!).

In [209]:
```
df2.groupby(['grade']).max()
```

Out[209]:

	fname	lname	gender	age	exercise	hours	address
grade							
32.0	Alika	Poole	female	19	2	16	9282 Purple Finch Lane, Lexington, NC 27292
43.0	Keegan	Rasmussen	male	19	4	3	876 East Pilgrim Street, Chelmsford, MA 01824
55.9	Levi	Coleman	male	19	3	3	9453 Laurel Street, Jersey City, NJ 07302
56.1	Gail	Mcneil	female	17	2	3	8409A Spruce St., Fishers, IN 46037
56.3	Jenna	Wagner	female	16	1	3	8829 Shore Dr., Hopewell Junction, NY 12533
57.9	Lacey	Nieves	female	18	1	2	38 West Brickyard Avenue, Roslindale, MA 02131
58.9	Isaiah	Harrington	male	17	4	4	84 Rock Creek Lane, Durham, NC 27703
59.0	Linda	Baldwin	female	16	5	2	970 SW. Second Ave., Cedar Falls, IA 50613
59.2	Willa	Byers	female	14	2	4	9466 Wayne Lane, Torrington, CT 06790
59.3	Ciaran	Gay	male	19	4	3	157 Bridge Street, Corona, NY 11368
59.4	Selma	Stout	female	19	2	3	5 Pierce St., Chester, PA 19013
59.8	Xanthus	Mcneil	male	18	3	4	3 West Shipley Rd., Langhorne, PA 19047
60.0	Steven	Sherman	male	18	1	2	8029 Depot Street, Port Charlotte, FL 33952
60.1	Kevin	Vance	male	17	5	5	9805 Walnutwood Dr., Panama City, FL 32404
60.3	Dillon	Ochoa	male	19	2	4	75 Arrowhead Drive, Danvers, MA 01923
60.5	Tanek	Stephens	male	16	2	4	7994 Leatherwood St., Pittsfield, MA 01201
60.6	Olivia	Craig	female	14	4	3	555 San Pablo Court, Fond Du Lac, WI 54935
60.8	Yuri	Martinez	male	14	3	4	9 Military St., Springboro, OH 45066
60.9	Ryan	Webb	male	17	2	4	9961 State Ave., Union City, NJ 07087
61.1	Sade	Quinn	female	19	3	5	9223 Brookside Street, Wyoming, MI 49509
61.2	Porter	Rowland	male	19	2	5	8 Sleepy Hollow St., West Palm Beach, FL 33404
61.3	Cheyenne	Prince	female	15	2	4	7595 Fieldstone St., Lake Worth, FL 33460
61.4	Piper	Lowery	female	19	2	6	557 Theatre Lane, Green Cove Springs, FL 32043

	fname	lname	gender	age	exercise	hours	address
grade							
61.6	Anthony	Bishop	male	19	2	5	1 Walt Whitman Street, Hartselle, AL 35640
61.7	Linda	Moreno	female	15	3	3	736 Elmwood Ave., South Bend, IN 46614
61.8	Pamela	Holder	female	16	4	4	302 Sunnyslope St., Baldwinsville, NY 13027
62.0	Gabriel	Jordan	male	16	4	2	7986 Briarwood Road, Cranberry Twp, PA 16066
62.2	Lavinia	Mcdonald	female	17	2	4	993 Rockaway Road, Fleming Island, FL 32003
62.3	Lee	Barber	male	15	2	2	449 Pearl Street, Largo, FL 33771
62.5	Ciaran	Brady	male	19	3	3	650 Homestead Lane, Fond Du Lac, WI 54935
...	...	...	...	...	...	...	...
96.9	Arthur	Pacheco	male	15	3	14	704 E. Hill Field Dr., West Lafayette, IN 47906
97.1	Zelenia	Vasquez	male	18	5	16	9029 Marlborough Lane, Lacey, WA 98503
97.2	Regan	Ryan	female	17	3	15	8107 Country Street, Cedar Falls, IA 50613
97.3	Zephr	Vance	male	19	5	18	87 Carpenter Lane, Campbell, CA 95008
97.4	Quamar	Haynes	male	16	5	18	9119 Bridgeton Lane, Reisterstown, MD 21136
97.5	Sandra	Wilder	male	18	3	15	9092 Prairie Lane, Lebanon, PA 17042
97.6	Stuart	Rasmussen	male	14	3	14	9975 Lookout Court, Buffalo, NY 14215
97.7	Haley	Kramer	male	18	4	18	851 Lake Forest St., Ellicott City, MD 21042
97.8	Elton	Preston	male	17	5	20	83 Fremont Court, Manchester, NH 03102
97.9	Tanner	Velasquez	male	19	5	18	766 Pacific Dr., Rapid City, SD 57701
98.0	Vielka	Wilkins	male	19	5	20	8443 Hamilton St., Dundalk, MD 21222
98.1	Zephania	Webb	male	19	4	18	9033 NW. Pleasant St., Salisbury, MD 21801
98.2	Tiger	Ochoa	male	19	4	15	879 Central Drive, Cheshire, CT 06410
98.3	Lee	Gallagher	female	19	1	16	95 Applegate Drive, Lansdowne, PA 19050
98.5	Lance	Rush	male	17	5	15	97 Essex Drive, Windermere, FL 34786
98.6	Risa	Rush	male	18	5	18	92 Pennington St., Ephrata, PA 17522
98.7	Holmes	Lawson	male	18	5	17	9516 Airport Street, Staunton, VA 24401
98.8	Cedric	Acevedo	male	19	5	14	50 Vine Lane, Derby, KS 67037

	fname	lname	gender	age	exercise	hours	address
grade							
98.9	Thaddeus	Kirby	male	19	3	15	8360 Wakehurst Dr., Los Angeles, CA 90008
99.0	Georgia	Powell	female	17	5	20	84 New Saddle St., Revere, MA 02151
99.1	Summer	Parker	male	17	4	18	667 Cross St., Miami, FL 33125
99.2	Tasha	Wilkins	male	18	4	19	8933 Canal Dr., Tualatin, OR 97062
99.3	Graham	Morse	male	17	5	16	68 Birch Hill Road, Virginia Beach, VA 23451
99.4	Sydnee	Reynolds	female	18	3	15	64 Hill Field Ave., Kennesaw, GA 30144
99.5	Kendall	Sparks	female	15	5	16	990 Paris Hill Street, Romeoville, IL 60446
99.6	Mohammad	Fleming	male	18	5	18	8534 East Wild Rose Road, Teaneck, NJ 07666
99.7	Lee	Roberson	male	18	4	15	348 Hall Drive, Salem, MA 01970
99.8	Faith	Cotton	female	17	4	15	658 8th Street, Waterbury, CT 06705
99.9	Salvador	Perkins	male	17	5	17	59 Summerhouse Dr., Cartersville, GA 30120
103.0	Yael	Yates	male	19	5	20	9895 Beach Drive, Elizabeth City, NC 27909

385 rows × 7 columns

In [210]:

```
#finds rows where the grade is equal to 100
df2.loc[df2['grade'] == 100, 'grade'] = 103
```

In [211]:
```
df2.head(10)
```

Out[211]:

	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
5	Neil	Whitley	male	16	5	16	88.7	40 Washington Ave., Bloomfield, NJ 07003
6	Nelle	Golden	female	17	1	9	80.2	9768 Hanover Dr., Meadville, PA 16335
7	Armando	Hoffman	male	17	5	18	95.1	360 Manor Drive, Northville, MI 48167
8	Illiana	Rojas	female	15	5	9	76.5	9425 Studebaker Dr., Thibodaux, LA 70301
9	Neil	Wooten	male	15	3	15	89.7	400 Bridge Court, Soddy Daisy, TN 37379

Exercise 4:

Load in the "travel_times.csv" file as a dataframe. Drop the "Comments" column. Then remove rows from the dataframe that have missing values and assign the resulting dataframe as a new variable.

In [212]:
```
df = pd.read_csv("datasets/travel_times.csv")
df.head()
```

Out[212]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	F
0	1/6/2012	16:37	Friday	Home	51.29	127.4	78.3	84.8	
1	1/6/2012	08:20	Friday	GSK	51.63	130.3	81.8	88.9	
2	1/4/2012	16:17	Wednesday	Home	51.27	127.4	82.0	85.8	
3	1/4/2012	07:53	Wednesday	GSK	49.17	132.3	74.2	82.9	
4	1/3/2012	18:57	Tuesday	Home	51.15	136.2	83.4	88.1	

In [213]:

```
#drop column & assigned a ned dataframe name
df1 = df.drop('Comments', axis=1)
df1.head()
```

Out[213]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	F
0	1/6/2012	16:37	Friday	Home	51.29	127.4	78.3	84.8	
1	1/6/2012	08:20	Friday	GSK	51.63	130.3	81.8	88.9	
2	1/4/2012	16:17	Wednesday	Home	51.27	127.4	82.0	85.8	
3	1/4/2012	07:53	Wednesday	GSK	49.17	132.3	74.2	82.9	
4	1/3/2012	18:57	Tuesday	Home	51.15	136.2	83.4	88.1	

In [214]:

```
#drop Missing Value
df1.dropna(inplace=True)

df1.head()
```

Out[214]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed
6	1/2/2012	17:31	Monday	Home	51.37	123.2	82.9	87.3
7	1/2/2012	07:34	Monday	GSK	49.01	128.3	77.5	85.9
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.3
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.1
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4

In [215]:

```
#df.fillna()
```

Exercise 5:

Using the dataframe from the exercise above (w/ no missing values), create bins that will categorize the AvgSpeed column as "slow" or "fast", and make a new column called "Speed" to hold those new values. Values less than 75 are "slow" and everything above is "fast".

In [216]:

```
#average speed "slow = <75", "fast = >75"
binslist = [0, 75, 160]

speedlist = ['slow', 'fast']

df1['Speed'] = pd.cut(df1['AvgSpeed'], binslist, labels=speedlist)
```

In [217]:

```
df1.head()
```

Out[217]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed
6	1/2/2012	17:31	Monday	Home	51.37	123.2	82.9	87.3
7	1/2/2012	07:34	Monday	GSK	49.01	128.3	77.5	85.9
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.3
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.1
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4

In [218]:

```
# Pandas Dataframe.query() method

# You can refer to column names that contain spaces by
# surrounding them in backticks.


result = df1.query('Speed == "fast" and DayOfWeek in ("Monday","Tuesday") ')
result
```

Out[218]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpee
6	1/2/2012	17:31	Monday	Home	51.37	123.2	82.9	87.
7	1/2/2012	07:34	Monday	GSK	49.01	128.3	77.5	85.
12	12/20/2011	16:05	Tuesday	Home	51.45	130.1	75.2	82.
13	12/20/2011	06:04	Tuesday	GSK	49.01	119.0	77.4	82.
14	12/19/2011	16:18	Monday	Home	51.04	132.2	77.5	83.
15	12/19/2011	07:34	Monday	GSK	52.00	137.8	76.5	87.
31	12/6/2011	17:24	Tuesday	Home	51.25	123.5	77.3	81.
46	11/22/2011	16:15	Tuesday	Home	51.49	129.6	78.6	83.
47	11/22/2011	07:27	Tuesday	GSK	51.65	128.6	76.1	82.
62	11/8/2011	17:24	Tuesday	Home	50.75	131.3	89.5	93.
64	11/7/2011	16:05	Monday	Home	51.06	127.4	80.4	85.
73	10/31/2011	15:49	Monday	Home	51.06	125.0	76.4	85.
74	10/31/2011	06:21	Monday	GSK	50.58	125.0	104.4	106.
87	10/18/2011	08:14	Tuesday	GSK	51.74	130.8	80.8	85.
88	10/17/2011	16:58	Monday	Home	51.30	127.3	78.6	82.
89	10/17/2011	08:22	Monday	GSK	50.61	137.1	93.7	100.
94	10/11/2011	08:25	Tuesday	GSK	48.94	130.8	85.7	93.
101	10/4/2011	17:39	Tuesday	Home	51.15	128.8	76.0	85.
102	10/4/2011	07:42	Tuesday	GSK	50.67	127.3	94.9	97.
103	10/3/2011	17:31	Monday	Home	51.22	126.7	81.2	86.
104	10/3/2011	07:41	Monday	GSK	50.65	127.4	91.1	95.
110	9/27/2011	07:36	Tuesday	GSK	50.65	128.1	86.3	88.
111	9/26/2011	17:37	Monday	Home	50.69	132.3	97.2	103.
112	9/26/2011	08:02	Monday	GSK	50.65	129.4	88.2	91.
127	9/12/2011	17:04	Monday	Home	51.43	131.1	75.1	79.
133	9/6/2011	16:27	Tuesday	Home	52.88	131.6	95.4	98.
134	9/6/2011	07:50	Tuesday	GSK	54.36	132.5	95.1	98.
143	8/29/2011	17:11	Monday	Home	51.04	131.0	75.5	84.
151	8/23/2011	17:23	Tuesday	Home	51.22	129.7	79.7	84.

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpee
153	8/22/2011	16:44	Monday	Home	51.12	126.8	77.9	85.
161	8/16/2011	17:27	Tuesday	Home	51.14	133.4	82.4	87.
163	8/15/2011	17:38	Monday	Home	51.11	132.3	78.0	83.
173	8/8/2011	17:05	Monday	Home	52.35	127.5	76.9	84.
181	8/2/2011	17:22	Tuesday	Home	51.16	124.2	76.3	83.
198	7/19/2011	17:17	Tuesday	Home	51.16	126.7	92.2	102.
199	7/19/2011	08:11	Tuesday	GSK	50.96	124.3	82.3	96.

Exercise 6:

Using the dataframe in the previous exercise, make a new column called "Police" which is equal to all the values being "no" (they were never stopped by police for speeding while traveling).

In [219]:

```
df1['Police'] = 'no'
df1.head()
```

Out[219]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed
6	1/2/2012	17:31	Monday	Home	51.37	123.2	82.9	87.3
7	1/2/2012	07:34	Monday	GSK	49.01	128.3	77.5	85.9
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.3
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.1
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4

Exercise 7:

Using the dataframe from the previous exercise, pick a method (Standard Deviation or Interquartile Range) and remove the outliers from the "FuelEconomy" column.

In [220]:

```
df1.head()
```

Out[220]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed
6	1/2/2012	17:31	Monday	Home	51.37	123.2	82.9	87.3
7	1/2/2012	07:34	Monday	GSK	49.01	128.3	77.5	85.9
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.3
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.1
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4

In [221]:
```
df1.dtypes
```

Out[221]:
```
Date                 object
StartTime            object
DayOfWeek            object
GoingTo              object
Distance            float64
MaxSpeed            float64
AvgSpeed            float64
AvgMovingSpeed      float64
FuelEconomy          object
TotalTime           float64
MovingTime          float64
Take407All           object
Speed              category
Police               object
dtype: object
```

In [222]:
```
df1 = df1.dropna()
```

In [223]:

```
df1['FuelEconomy']
```

Out[223]:

```
6         -
7         -
8      8.89
9      8.89
10     8.89
11     8.89
12     8.89
13     8.89
14     8.89
15     8.89
16     9.08
17     9.08
18     9.08
19     9.08
20     9.08
21     9.08
22     9.08
23     9.08
24     9.76
25     9.76
26     9.76
27     9.76
28     9.76
29     9.76
30     9.76
31     9.16
32     9.16
33     9.16
42      9.3
43      9.3
       ... 
172    8.54
173    8.54
174    8.54
175    8.48
176    8.48
177    8.48
178    8.48
179    8.48
180    8.48
181    8.48
182    8.48
183    8.45
184    8.45
185    8.45
186    8.45
187    8.45
188    8.45
189    8.45
190    8.45
191    8.45
192    8.45
193    8.28
194    8.28
```

```
195    8.28
196    7.89
197    7.89
198    7.89
199    7.89
200    7.89
201    7.89
Name: FuelEconomy, Length: 188, dtype: object
```

In [227]:

```
df1.loc[df1['FuelEconomy']=='-', 'FuelEconomy'] = 0
df1
```

Out[227]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpee
6	1/2/2012	17:31	Monday	Home	51.37	123.2	82.9	87.
7	1/2/2012	07:34	Monday	GSK	49.01	128.3	77.5	85.
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.
11	12/21/2011	07:45	Wednesday	GSK	51.77	124.8	71.7	78.
12	12/20/2011	16:05	Tuesday	Home	51.45	130.1	75.2	82.
13	12/20/2011	06:04	Tuesday	GSK	49.01	119.0	77.4	82.
14	12/19/2011	16:18	Monday	Home	51.04	132.2	77.5	83.
15	12/19/2011	07:34	Monday	GSK	52.00	137.8	76.5	87.
16	12/16/2011	12:22	Friday	Home	51.05	128.4	86.9	90.
17	12/16/2011	07:21	Friday	GSK	49.04	124.6	71.1	80.
18	12/15/2011	16:14	Thursday	Home	51.06	126.9	80.5	84.
19	12/15/2011	07:19	Thursday	GSK	51.68	123.5	68.1	75.
20	12/14/2011	16:20	Wednesday	Home	51.04	123.4	75.1	79.
21	12/14/2011	07:23	Wednesday	GSK	51.67	123.5	76.6	82.
22	12/13/2011	17:43	Tuesday	Home	51.15	130.6	74.8	82.
23	12/13/2011	07:25	Tuesday	GSK	49.19	126.1	65.4	74.
24	12/12/2011	07:20	Monday	GSK	49.02	126.1	65.7	74.
25	12/9/2011	12:04	Friday	Home	51.14	126.8	87.3	90.
26	12/9/2011	07:22	Friday	GSK	51.69	128.4	74.0	77.
27	12/8/2011	17:41	Thursday	Home	51.07	125.0	74.6	81.
28	12/8/2011	07:14	Thursday	GSK	51.63	134.4	76.5	84.
29	12/7/2011	16:12	Wednesday	Home	51.10	126.5	79.9	85.
30	12/7/2011	07:18	Wednesday	GSK	51.64	124.6	73.6	82.
31	12/6/2011	17:24	Tuesday	Home	51.25	123.5	77.3	81.
32	12/6/2011	07:24	Tuesday	GSK	51.64	122.3	69.3	74.
33	12/5/2011	16:18	Monday	Home	50.18	124.0	71.0	79.
42	11/24/2011	16:15	Thursday	Home	51.49	126.6	74.0	82.
43	11/24/2011	07:23	Thursday	GSK	51.69	124.9	73.3	80.
...	...	...	...	...	...	...	...	.
172	8/9/2011	08:15	Tuesday	GSK	49.08	134.8	60.5	67.
173	8/8/2011	17:05	Monday	Home	52.35	127.5	76.9	84.

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpee
174	8/8/2011	08:07	Monday	GSK	49.25	126.3	68.5	78.
175	8/5/2011	17:00	Friday	Home	51.94	126.7	74.5	82.
176	8/5/2011	08:20	Friday	GSK	49.13	123.9	74.1	79.
177	8/4/2011	17:38	Thursday	Home	50.96	131.9	70.3	78.
178	8/4/2011	08:17	Thursday	GSK	49.12	122.4	71.5	77.
179	8/3/2011	17:14	Wednesday	Home	51.64	125.0	72.2	78.
180	8/3/2011	08:06	Wednesday	GSK	49.06	121.9	71.5	78.
181	8/2/2011	17:22	Tuesday	Home	51.16	124.2	76.3	83.
182	8/2/2011	07:38	Tuesday	GSK	53.48	124.9	68.8	78.
183	7/29/2011	20:31	Friday	Home	50.68	135.6	107.7	110.
184	7/29/2011	08:22	Friday	GSK	49.07	121.1	73.2	77.
185	7/28/2011	17:46	Thursday	Home	51.09	128.5	76.0	84.
186	7/28/2011	08:11	Thursday	GSK	49.11	120.1	69.1	73.
187	7/27/2011	17:24	Wednesday	Home	50.98	124.9	68.3	71.
188	7/27/2011	08:15	Wednesday	GSK	48.82	124.5	70.4	77.
189	7/26/2011	17:15	Tuesday	Home	51.28	122.1	43.7	51.
190	7/26/2011	08:11	Tuesday	GSK	49.16	122.6	71.9	76.
191	7/25/2011	16:59	Monday	Home	51.05	126.6	70.4	78.
192	7/25/2011	08:06	Monday	GSK	48.32	121.2	63.4	78.
193	7/22/2011	16:47	Friday	Home	51.24	126.3	75.8	81.
194	7/22/2011	08:28	Friday	GSK	51.05	123.3	88.9	96.
195	7/21/2011	07:59	Thursday	GSK	48.35	129.3	81.5	89.
196	7/20/2011	17:17	Wednesday	Home	53.47	124.0	58.6	71.
197	7/20/2011	08:24	Wednesday	GSK	48.50	125.8	75.7	87.
198	7/19/2011	17:17	Tuesday	Home	51.16	126.7	92.2	102.
199	7/19/2011	08:11	Tuesday	GSK	50.96	124.3	82.3	96.
200	7/18/2011	08:09	Monday	GSK	54.52	125.6	49.9	82.
201	7/14/2011	08:03	Thursday	GSK	50.90	123.7	76.2	95.

188 rows × 14 columns

```
In [228]: df1.dtypes
```

```
Out[228]: Date                object
          StartTime           object
          DayOfWeek           object
          GoingTo             object
          Distance           float64
          MaxSpeed           float64
          AvgSpeed           float64
          AvgMovingSpeed     float64
          FuelEconomy         object
          TotalTime          float64
          MovingTime         float64
          Take407All          object
          Speed             category
          Police              object
          dtype: object
```

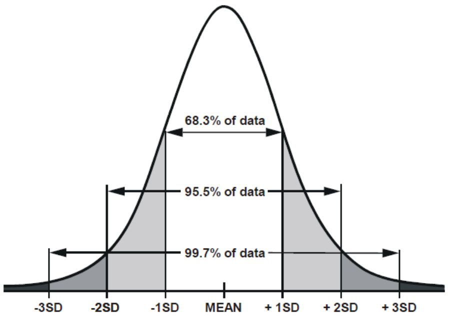
```
In [229]: df1['FuelEconomy'] = df1['FuelEconomy'].astype(float)
```

```
In [230]: df1.dtypes
```

```
Out[230]: Date                object
          StartTime           object
          DayOfWeek           object
          GoingTo             object
          Distance           float64
          MaxSpeed           float64
          AvgSpeed           float64
          AvgMovingSpeed     float64
          FuelEconomy        float64
          TotalTime          float64
          MovingTime         float64
          Take407All          object
          Speed             category
          Police              object
          dtype: object
```

Standary Deviation Method

Figure 3.9
Areas under the normal curve that lie between 1, 2, and 3
standard deviations on each side of the mean



In [231]:

```
#Standard Deviation Method

meangrade = df1['FuelEconomy'].mean()
stdgrade = df1['FuelEconomy'].std()
print("Remove outliers that are outside the 95% confidence interval")
toprange = meangrade + stdgrade * 1.96
botrange = meangrade - stdgrade * 1.96

print("Top range is %f" % toprange)
print("Bottom range is %f" % botrange)

copydf = df1.copy() #to not mess up the original df
copydf = copydf.drop(copydf[copydf['FuelEconomy'] > toprange].index)
copydf = copydf.drop(copydf[copydf['FuelEconomy'] < botrange].index)

copydf.head()
```

```
Remove outliers that are outside the 95% confidence interval
Top range is 10.607921
Bottom range is 6.588355
```

Out[231]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.3
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.1
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4
11	12/21/2011	07:45	Wednesday	GSK	51.77	124.8	71.7	78.9
12	12/20/2011	16:05	Tuesday	Home	51.45	130.1	75.2	82.7

In [232]:

```
#Interquartile Range Method

q1 = df1['FuelEconomy'].quantile(.25)
q3 = df1['FuelEconomy'].quantile(.75)
iqr = q3-q1
toprange = q3 + iqr * 1.5
botrange = q1 - iqr * 1.5

newdf = df1.copy()
newdf = newdf.drop(newdf[newdf['FuelEconomy'] > toprange].index)
newdf = newdf.drop(newdf[newdf['FuelEconomy'] < botrange].index)

newdf.head()
```

Out[232]:

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.3
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.1
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4
11	12/21/2011	07:45	Wednesday	GSK	51.77	124.8	71.7	78.9
12	12/20/2011	16:05	Tuesday	Home	51.45	130.1	75.2	82.7

In [ ]: