Data Analysis assignment (20 points)

You have been given the dataset `travel-times` in a CSV format. This dataset comes from a driver that uses an app to track GPS coordinates as he drives to work and back each day. The app collects the location and elevation data. In total, data for about 200 trips are summarized in this data set.

Load the `travel-times` in a `df` variable using `pandas` and then perform the following:

- print the shape of the dataset **(1 points)**
- print the first 15 rows of the dataset **(1 points)**
- get information for the features (columns) with missing values **(1 points)**
- drop duplicate values (if any) by keeping only the last instance **(1 points)**
- calculate the total number of missing values (if any) on each column **(2 points)**
- create two copies of the dataframe, and then:
    - drop rows with missing values from the 1st copy **(1 points)**
    - drop columns with missing values from the 2nd copy **(1 points)**
- get summary statistics and see the correlation between the numerical columns **(1 points)**
- show rows 11 to 14 **(1 points)**
- create a subset with trips occurred on November 23, 2011 and January 6, 2012 **(2 points)**
- produce a scatterplot between `Distance` and `TotalTime` **(1 points)**
    - Use:

```
import matplotlib.pyplot as plt
plt.rcParams.update({'font.size': 20, 'figure.figsize': (10, 8)})
```

- produce boxplots for `AvgSpeed` and `AvgMovingSpeed` (use different cells for each) **(2 points)**

```
import matplotlib.pyplot as plt
import pandas as pd

plt.rcParams.update({'font.size': 20, 'figure.figsize': (10, 8)})

df = pd.read_csv("travel-times.csv")

df.shape
```

```
(205, 13)
```

```
df.head(15)
```

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed
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
5	1/3/2012	07:57	Tuesday	GSK	51.80	135.8	84.5	88.8
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
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
13	12/20/2011	06:04	Tuesday	GSK	49.01	119.0	77.4	82.0
14	12/19/2011	16:18	Monday	Home	51.04	132.2	77.5	83.5

Next steps: View recommended plots

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 205 entries, 0 to 204
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Date            205 non-null    object 
 1   StartTime       205 non-null    object 
 2   DayOfWeek       205 non-null    object 
 3   GoingTo         205 non-null    object 
 4   Distance        205 non-null    float64
 5   MaxSpeed        205 non-null    float64
 6   AvgSpeed        205 non-null    float64
 7   AvgMovingSpeed  205 non-null    float64
 8   FuelEconomy     188 non-null    object 
 9   TotalTime       205 non-null    float64
 10  MovingTime      205 non-null    float64
 11  Take407All      205 non-null    object 
 12  Comments        24 non-null     object 
dtypes: float64(6), object(7)
memory usage: 20.9+ KB
```

```
df.drop_duplicates(keep='last',inplace=True)
```

```
df.columns
```

```
Index(['Date', 'StartTime', 'DayOfWeek', 'GoingTo', 'Distance', 'MaxSpeed',
       'AvgSpeed', 'AvgMovingSpeed', 'FuelEconomy', 'TotalTime', 'MovingTime',
       'Take407All', 'Comments'],
      dtype='object')
```

```
df.isnull().sum()
```

```
Date               0
StartTime          0
DayOfWeek          0
GoingTo            0
Distance           0
MaxSpeed           0
```

```
AvgSpeed            0
AvgMovingSpeed      0
FuelEconomy        17
TotalTime           0
MovingTime          0
Take407All          0
Comments          181
dtype: int64
```

```
df1 = df.copy()
df2 = df.copy()
```

```
df1.dropna(inplace=True)
df1.shape
```

```
(23, 13)
```

```
df2.dropna(axis=1, inplace=True)
df2.shape
```

```
(205, 11)
```

```
df.describe()
```

	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	TotalTime	MovingTime
count	205.000000	205.000000	205.000000	205.000000	205.000000	205.000000
mean	50.981512	127.591707	74.477561	81.975610	41.904390	37.871707
std	1.321205	4.128450	11.409816	10.111544	6.849476	4.835072
min	48.320000	112.200000	38.100000	50.300000	28.200000	27.100000
25%	50.650000	124.900000	68.900000	76.600000	38.400000	35.700000
50%	51.140000	127.400000	73.600000	81.400000	41.300000	37.600000
75%	51.630000	129.800000	79.900000	86.000000	44.400000	39.900000
max	60.320000	140.900000	107.700000	112.100000	82.300000	62.400000

```
df.select_dtypes(exclude='object').corr()
```

	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	TotalTime	MovingTime
Distance	1.000000	0.145091	-0.006445	0.011874	0.197207	0.197044
MaxSpeed	0.145091	1.000000	0.253869	0.257823	-0.198775	-0.222574
AvgSpeed	-0.006445	0.253869	1.000000	0.872143	-0.877806	-0.835814
AvgMovingSpeed	0.011874	0.257823	0.872143	1.000000	-0.856986	-0.944433
TotalTime	0.197207	-0.198775	-0.877806	-0.856986	1.000000	0.920935
MovingTime	0.197044	-0.222574	-0.835814	-0.944433	0.920935	1.000000

```
df.iloc[10:14]
```

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	FuelEconomy	TotalTime	MovingTime
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4	8.89	39.8	36.2
11	12/21/2011	07:45	Wednesday	GSK	51.77	124.8	71.7	78.9	8.89	43.3	39.4
12	12/20/2011	16:05	Tuesday	Home	51.45	130.1	75.2	82.7	8.89	41.1	37.3
13	12/20/2011	06:04	Tuesday	GSK	49.01	119.0	77.4	82.0	8.89	38.0	35.9

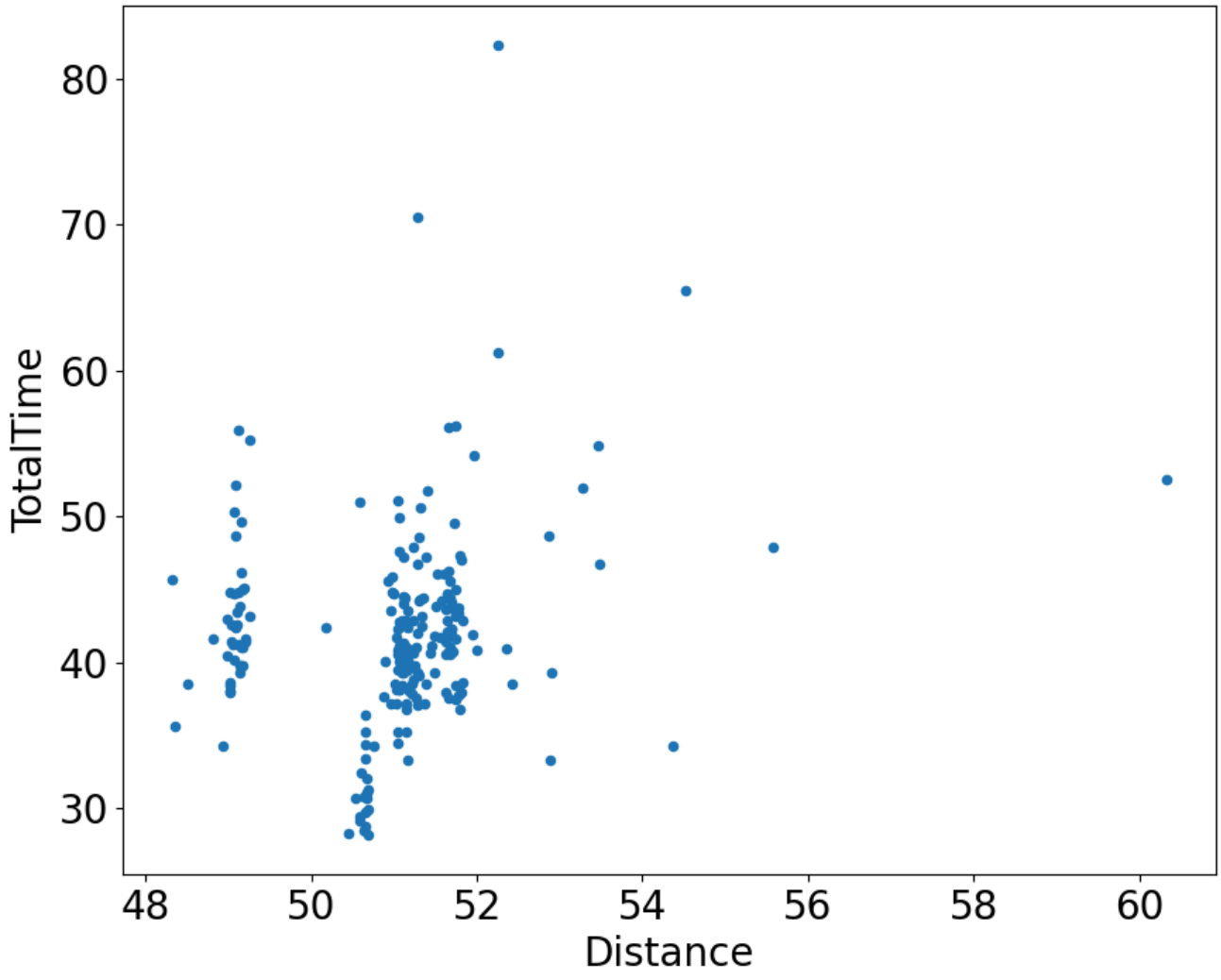
```
df[(df['Date']>= '11/13/2011') | (df['Date']<='01/26/2012')]
```

	Date	StartTime	DayOfWeek	GoingTo	Distance	MaxSpeed	AvgSpeed	AvgMovingSpeed	FuelEconomy	TotalTime	MovingTi
8	12/23/2011	08:01	Friday	GSK	52.91	130.3	80.9	88.3	8.89	39.3	36
9	12/22/2011	17:19	Thursday	Home	51.17	122.3	70.6	78.1	8.89	43.5	39
10	12/22/2011	08:16	Thursday	GSK	49.15	129.4	74.0	81.4	8.89	39.8	36
11	12/21/2011	07:45	Wednesday	GSK	51.77	124.8	71.7	78.9	8.89	43.3	39
12	12/20/2011	16:05	Tuesday	Home	51.45	130.1	75.2	82.7	8.89	41.1	37
...	...	...	...	...	...	...	...	...	...	...	
200	7/18/2011	08:09	Monday	GSK	54.52	125.6	49.9	82.4	7.89	65.5	39
201	7/14/2011	08:03	Thursday	GSK	50.90	123.7	76.2	95.1	7.89	40.1	32
202	7/13/2011	17:08	Wednesday	Home	51.96	132.6	57.5	76.7	NaN	54.2	40
203	7/12/2011	17:51	Tuesday	Home	53.28	125.8	61.6	87.6	NaN	51.9	36
204	7/11/2011	16:56	Monday	Home	51.73	125.0	62.8	92.5	NaN	49.5	33

162 rows × 13 columns

```
df.plot(kind='scatter', x='Distance', y='TotalTime')
```

```
<Axes: xlabel='Distance', ylabel='TotalTime'>
```




```
df['AvgMovingSpeed'].plot(kind='box')
```

<Axes: >