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

```
df = pd.read_csv('/content/202004-divvy-tripdata.csv')
```

```
df.head()
```

ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start
ADBBC638E45	docked_bike	2020-04-26 17:45:14	2020-04-26 18:12:03	Eckhart Park	86	Lincoln Ave & Diversey Pkwy	152.0	41.
B80E996FF60D	docked_bike	2020-04-17 17:08:54	2020-04-17 17:17:03	Drake Ave & Fullerton Ave	503	Kosciuszko Park	499.0	41.
4A79A4E006F4	docked_bike	2020-04-01 17:54:13	2020-04-01 18:08:36	McClurg Ct & Erie St	142	Indiana Ave & Roosevelt Rd	255.0	41.
BDF5CDBA725	docked_bike	2020-04-07 12:50:19	2020-04-07 13:02:31	California Ave & Division St	216	Wood St & Augusta Blvd	657.0	41.
306C119C6158	docked_bike	2020-04-18 10:22:59	2020-04-18 11:15:54	Rush St & Hubbard St	125	Sheridan Rd & Lawrence Ave	323.0	41.

Next steps: Generate code with df View recommended plots New interactive sheet

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84776 entries, 0 to 84775
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   ride_id             84776 non-null  object 
 1   rideable_type       84776 non-null  object 
 2   started_at          84776 non-null  object 
 3   ended_at            84776 non-null  object 
 4   start_station_name  84776 non-null  object 
 5   start_station_id    84776 non-null  int64  
 6   end_station_name    84677 non-null  object 
 7   end_station_id      84677 non-null  float64
 8   start_lat           84776 non-null  float64
 9   start_lng           84776 non-null  float64
 10  end_lat             84677 non-null  float64
 11  end_lng             84677 non-null  float64
 12  member_casual       84776 non-null  object 
dtypes: float64(5), int64(1), object(7)
memory usage: 8.4+ MB
```

```
# change Data types
df['started_at'] = pd.to_datetime(df['started_at'])
df['ended_at'] = pd.to_datetime(df['ended_at'])

df['start_station_id'] = df['start_station_id'].astype('int64')

df['start_station_name'] = df['start_station_name'].astype('string')
df['end_station_name'] = df['end_station_name'].astype('string')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84776 entries, 0 to 84775
Data columns (total 13 columns):
 #   Column              Non-Null Count  Dtype         
---  ------              --------------  -----         
 0   ride_id             84776 non-null  object        
 1   rideable_type       84776 non-null  object        
 2   started_at          84776 non-null  datetime64[ns]
 3   ended_at            84776 non-null  datetime64[ns]
 4   start_station_name  84776 non-null  string        
 5   start_station_id    84776 non-null  int64
 6   end_station_name    84677 non-null  string
 7   end_station_id      84677 non-null  float64
 8   start_lat           84776 non-null  float64
 9   start_lng           84776 non-null  float64
 10  end_lat             84677 non-null  float64
 11  end_lng             84677 non-null  float64
 12  member_casual       84776 non-null  object
dtypes: datetime64[ns](2), float64(5), int64(1), object(3), string(2)
memory usage: 8.4+ MB
```

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

```
                    0
ride_id             0
rideable_type       0
started_at          0
ended_at            0
start_station_name  0
start_station_id    0
end_station_name   99
end_station_id     99
start_lat           0
start_lng           0
end_lat            99
end_lng            99
member_casual       0

dtype: int64
```

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

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

```
                    0
ride_id             0
rideable_type       0
started_at          0
ended_at            0
start_station_name  0
start_station_id    0
end_station_name    0
end_station_id      0
start_lat           0
start_lng           0
end_lat             0
end_lng             0
member_casual       0

dtype: int64
```

```
print(df.duplicated().sum())
```

```
0
```

```
df['ride_length'] = (df['ended_at'] - df['started_at']).dt.total_seconds() / 60
```

```
df['day_of_week'] = df['started_at'].dt.day_name()
```

```
df.head(3)
```

pe	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start
ke	2020-04-26 17:45:14	2020-04-26 18:12:03	Eckhart Park	86	Lincoln Ave & Diversey Pkwy	152.0	41.
ke	2020-04-17 17:08:54	2020-04-17 17:17:03	Drake Ave & Fullerton Ave	503	Kosciuszko Park	499.0	41.
ke	2020-04-01 17:54:13	2020-04-01 18:08:36	McClurg Ct & Erie St	142	Indiana Ave & Roosevelt Rd	255.0	41.

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
# Average ride length
avg_length = df.groupby('member_casual')['ride_length'].mean()
avg_length
```

	ride_length
member_casual	
casual	72.435153
member	21.357266

dtype: float64

```
rides_per_day = df.groupby(['day_of_week', 'member_casual']).size().unstack()
rides_per_day
```

member_casual	casual	member
day_of_week		
Friday	2507	7454
Monday	2677	8055
Saturday	4065	8833
Sunday	6468	11429
Thursday	2426	9256
Tuesday	3651	9146
Wednesday	1794	6916

Next steps: Generate code with rides_per_day | View recommended plots | New interactive sheet

```
df.to_csv('cleaned_data.csv', index=False)
```