```
import pandas as pd
import numpy as np
import altair as alt
```

```
ufo_sightings = pd.read_csv('ufo_sightings.csv')
```

```
# select ufo sightings only in the US
ufo_sightings = ufo_sightings[ufo_sightings['country'] == 'us']
ufo_sightings = ufo_sightings.reset_index(drop = True)
```

```
# drop selected columns
ufo_sightings = ufo_sightings.drop(columns = ['described_encounter_length', 'description'])
```

```
# convert seconds to minutes
ufo_sightings.encounter_length = ufo_sightings.encounter_length / 60.0
```

```
# separate date_time into two distinct columns
ufo_sightings[['Date', 'Time']] = ufo_sightings.date_time.str.split(' ', expand = True)
# separate date(MMDDYY) into 3 different columns
ufo_sightings[['Month', 'Day', 'Year']] = ufo_sightings.Date.str.split('/', expand = True)
# separate time to hour and minutes
ufo_sightings[['Hour', 'Minute']] = ufo_sightings.Time.str.split(':', expand = True)

# drop the selected columns
ufo_sightings = ufo_sightings.drop(columns = ['date_time', 'Date', 'Day', 'Minute', 'Time'])
# reorder the columns
ufo_sightings = ufo_sightings.iloc[:, [9,8,10,0,1,2,3,4,5,6,7]]
```

```
# round all doubles to two decimal places
ufo_sightings.encounter_length = round(ufo_sightings.encounter_length, 2)
ufo_sightings.latitude = round(ufo_sightings.latitude, 2)
ufo_sightings.longitude = round(ufo_sightings.longitude, 2)
```



```
ufo_sightings.Month = ufo_sightings.Month.astype(int)
ufo_sightings.Year = ufo_sightings.Year.astype(int)
```

```
# rename columns
ufo_sightings = ufo_sightings.rename(columns = {'city_area': 'City', 'state': 'State', 'count
                                                'ufo_shape': 'UFO Shape', 'encounter_length':
                                                'date_documented': 'Date Documented', 'latitu
                                                'longitude': 'Longitude'})
```

Initial explorations:

```
# dimensions
ufo_sightings.shape
```

```
(65114, 11)
```

```
# missing values
ufo_sightings.isna().sum()
```

```
Year                   0
Month                  0
Hour                   0
City                   0
State                  0
Country                0
UFO Shape           1553
Encounter Length       2
Date Documented        0
Latitude               0
Longitude              0
dtype: int64
```

A possible reason for the missing values in the UFO Shape column might be that it was difficult to determine such qualities due to weather conditions, time of day, etc.

```
# number of unique values for each column
ufo_sightings.nunique()
```

```
Year                  83
Month                 12
Hour                  25
City               11364
State                 52
Country                1
UFO Shape             28
Encounter Length     435
Date Documented      315
Latitude            2255
Longitude           4485
dtype: int64
```

```
# mean, standard deviation, min, and max for columns of interest
ufo_sightings[['Year', 'Month', 'Hour', 'Encounter Length', 'Latitude', 'Longitude']].aggrega
    ['mean', 'std', 'min', 'max']
).transpose()
```

	mean	std	min	max
Year	2004.090057	10.385694	1910.00	2014.00
Month	6.861597	3.248243	1.00	12.00
Hour	15.835396	7.567445	0.00	24.00
Encounter Length	96.669800	6741.994933	0.00	1104600.00
Latitude	38.357840	5.570069	17.97	70.64
Longitude	-95.710903	18.108472	-176.66	-65.83

```
ufo_sightings
```

	Year	Month	Hour	City	State	Country	UFO Shape	Encounte
0	1949	10	20	san marcos	tx	us	cylinder	45.00
1	1956	10	21	edna	tx	us	circle	0.33
2	1960	10	20	kaneohe	hi	us	light	15.00
3	1961	10	19	bristol	tn	us	sphere	5.00
4	1965	10	23	norwalk	ct	us	disk	20.00
...	...	...	...	...	...	...	...	...
65109	2013	9	21	nashville	tn	us	light	10.00
65110	2013	9	22	boise	id	us	circle	20.00
65111	2013	9	22	napa	ca	us	other	20.00
65112	2013	9	22	vienna	va	us	circle	0.08
65113	2013	9	23	edmond	ok	us	cigar	17.00

65114 rows × 11 columns