Python Project

Gabriella Cerrato (gac2625)

In [12]:
```
#import dataset
flights = sns.load_dataset('flights')
flights.to_csv("flights.csv")
```

In [16]:
```
flights= pd.read_csv("flights.csv")
```

In [20]:
```
flights = flights.drop(flights.columns[[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
, 11, 12, 13]], axis=1)
```

In [21]:
```
flights.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 144 entries, 0 to 143
Data columns (total 3 columns):
year          144 non-null int64
month         144 non-null object
passengers    144 non-null int64
dtypes: int64(2), object(1)
memory usage: 3.5+ KB
```

The dataset has 3 columns and 144 rows. The dataset has information about number of passengers on flights during each month from the year 1949 to the year 1960. The variable 'month' is a categorical variable and the variable 'passengers' is a numeric variable that tells us the number of passengers that have flown during a certain month and year.

In [22]:
```
#view the first rows
flights.head()
```

Out[22]:

	year	month	passengers
0	1949	January	112
1	1949	February	118
2	1949	March	132
3	1949	April	129
4	1949	May	121

In [23]:
```
#perform EDA

flights.describe()
```

Out[23]:

	year	passengers
count	144.000000	144.000000
mean	1954.500000	280.298611
std	3.464102	119.966317
min	1949.000000	104.000000
25%	1951.750000	180.000000
50%	1954.500000	265.500000
75%	1957.250000	360.500000
max	1960.000000	622.000000

In [7]:
```
#filter
(flights.filter(['passengers', 'month'])
#group by sex
.groupby(['month'])
#compute mean, standard deviation, and counts
.agg(['mean','std','count']))
```

Out[7]:

	passengers		
	mean	**std**	**count**
month			
April	267.083333	107.374839	12
August	351.083333	155.783333	12
December	261.833333	103.093808	12
February	235.000000	89.619397	12
January	241.750000	101.032960	12
July	351.333333	156.827255	12
June	311.666667	134.219856	12
March	270.166667	100.559194	12
May	271.833333	114.739890	12
November	232.833333	95.185783	12
October	266.583333	110.744964	12
September	302.416667	123.954140	12

In [8]:
```
flights.passengers.mean()
```

Out[8]: 280.2986111111111

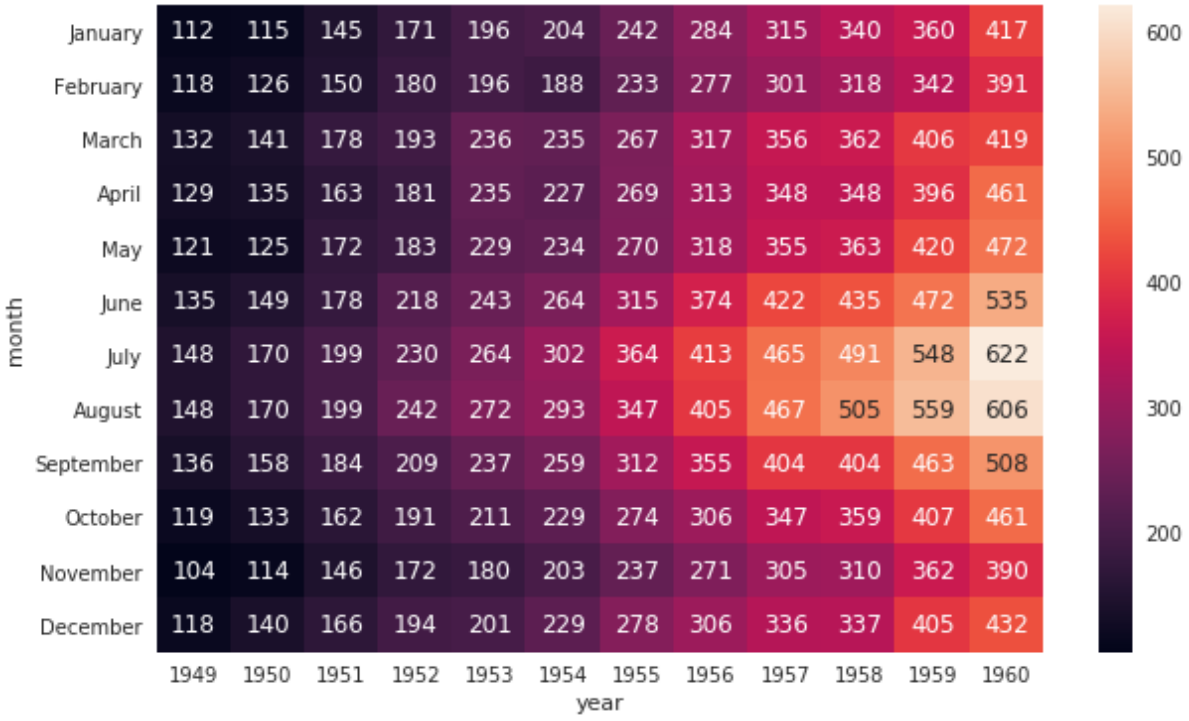
The average number of passengers in the dataset is 280.29 or about 280 people. The month with the highest average number of passengers is July. This is probably due to July being a summer month, which is when a lot of traveling takes place.

In [25]:
```
#correlation heat map
sns.set()

# Load the example flights dataset and convert to long-form
flights_long = sns.load_dataset("flights")
flights = flights_long.pivot("month", "year", "passengers")

# Draw a heatmap with the numeric values in each cell
ax, f = plt.subplots(figsize=(10, 6))
sns.heatmap(flights, annot=True, fmt="d")
```

Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x7fb0da753860>



*The correlation heat map shows that the month of July during the year 1960 had the highest correlation to number of passengers.

In [ ]: