

# DataViz1-ggplotBasics

Ellie Kuskie

## Setup

```
library(ggplot2)
library(here)
```

here() starts at C:/Users/conno/OneDrive/Documents/UCSC Undergrad/DataScience4EEB/DataViz1-ggplotBasics

```
# Read in the data and store as "ci_np" data object
ci_np <- read.csv("ci_np.csv")

# display the first 6 rows
head(ci_np)
```

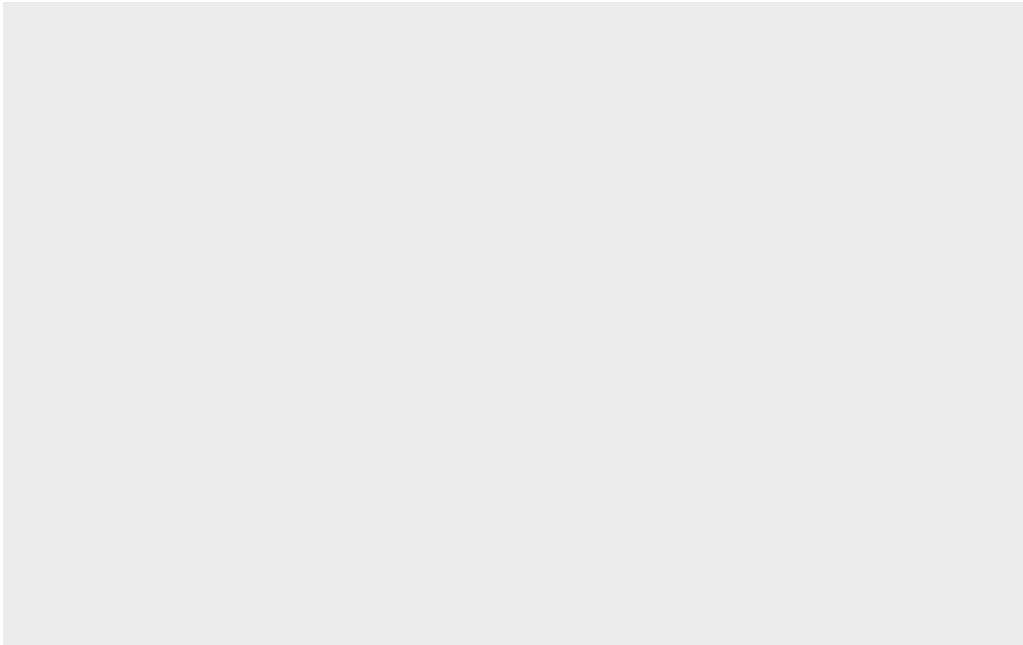
	region	state	code	park_name	type	visitors	year
1	PW	CA	CHIS	Channel Islands National Park	National Park	1200	1963
2	PW	CA	CHIS	Channel Islands National Park	National Park	1500	1964
3	PW	CA	CHIS	Channel Islands National Park	National Park	1600	1965
4	PW	CA	CHIS	Channel Islands National Park	National Park	300	1966
5	PW	CA	CHIS	Channel Islands National Park	National Park	15700	1967
6	PW	CA	CHIS	Channel Islands National Park	National Park	31000	1968

## Q1)

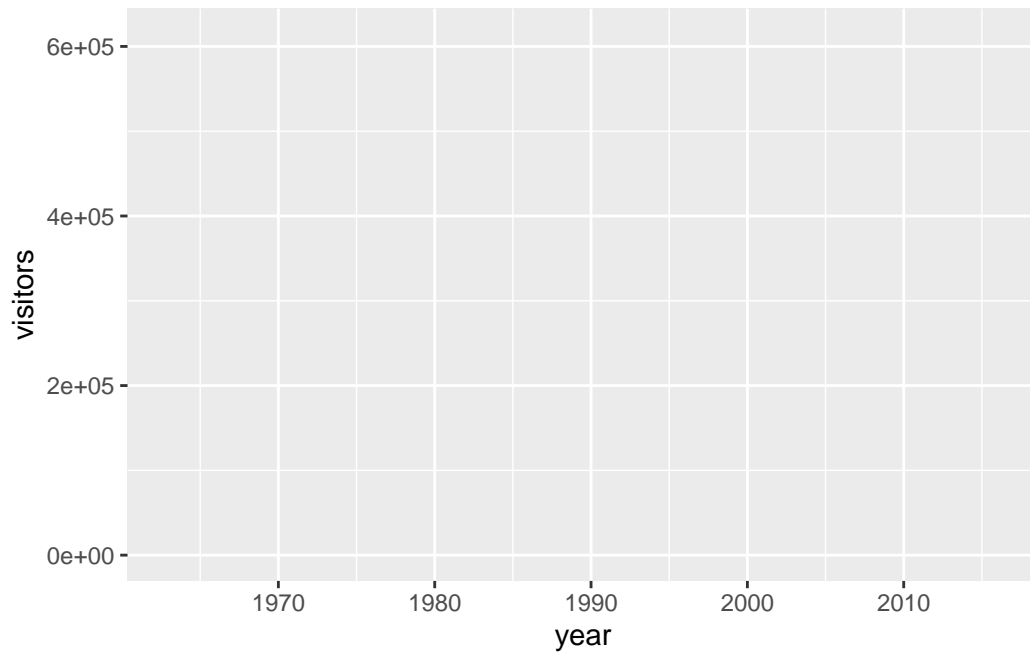
```
# got column names from ci_np dataset
colnames(ci_np)
```

```
[1] "region" "state" "code" "park_name" "type" "visitors"
[7] "year"
```

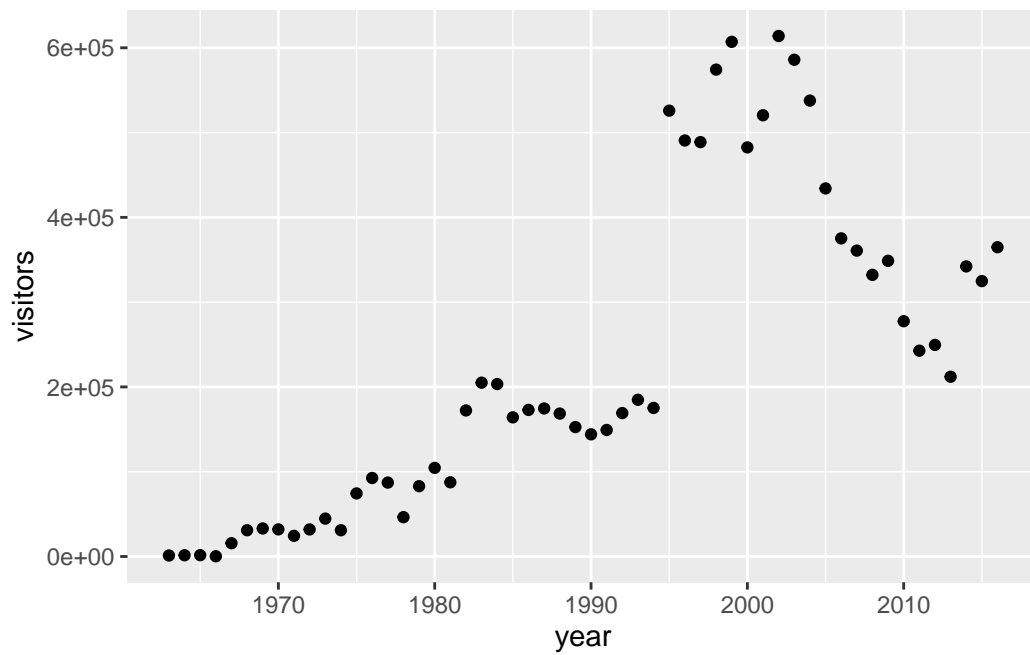
```
# created blank plot (no axes or points)
ggplot(data = ci_np)
```



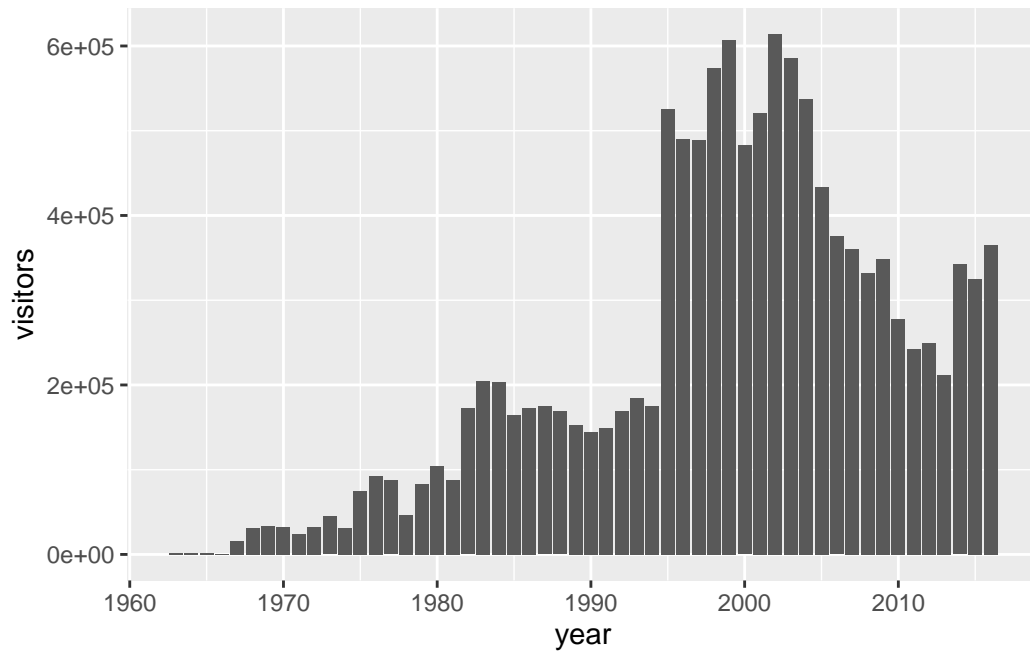
```
# made plot with axes but still no points
ggplot(data = ci_np, mapping = aes(x = year, y = visitors))
```



```
# added points to graph
ggplot(data = ci_np, mapping = aes(x = year, y = visitors)) +
  geom_point()
```

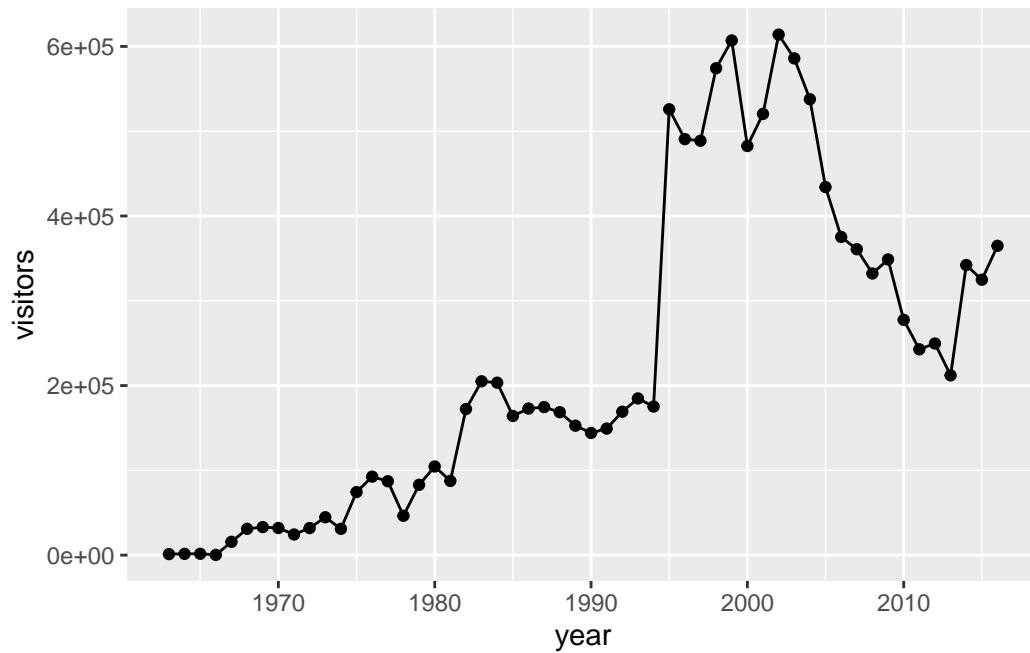


```
# changed graph to column chart
ggplot(data = ci_np, mapping = aes(x = year, y = visitors)) +
  geom_col()
```



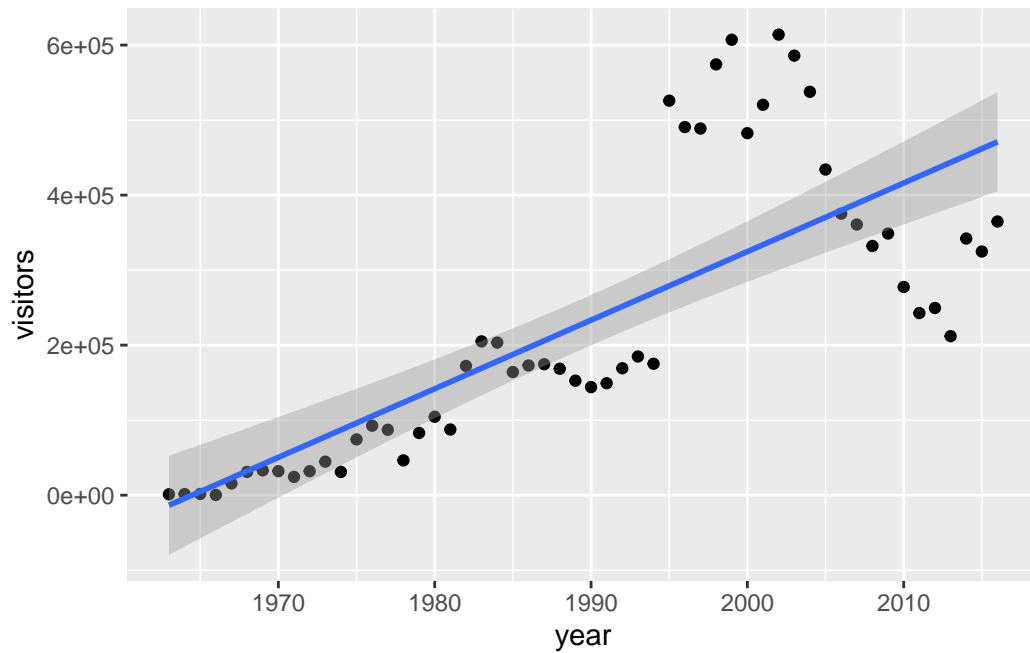
Q2)

```
# made into line graph
ggplot(data = ci_np, mapping = aes(x = year, y = visitors)) +
  geom_point() +
  geom_line()
```



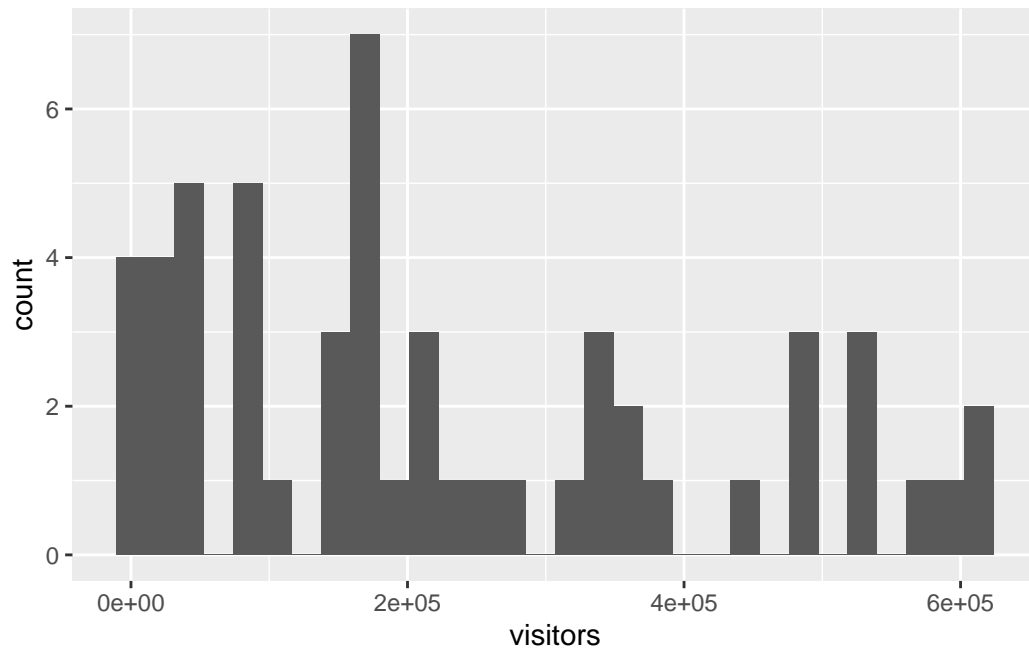
```
# made smooth trendline instead of just connecting dots
# then made it linear using method = lm
ggplot(data = ci_np, mapping = aes(x = year, y = visitors)) +
  geom_point() +
  geom_smooth(method = lm)
```

`geom\_smooth()` using formula = 'y ~ x'

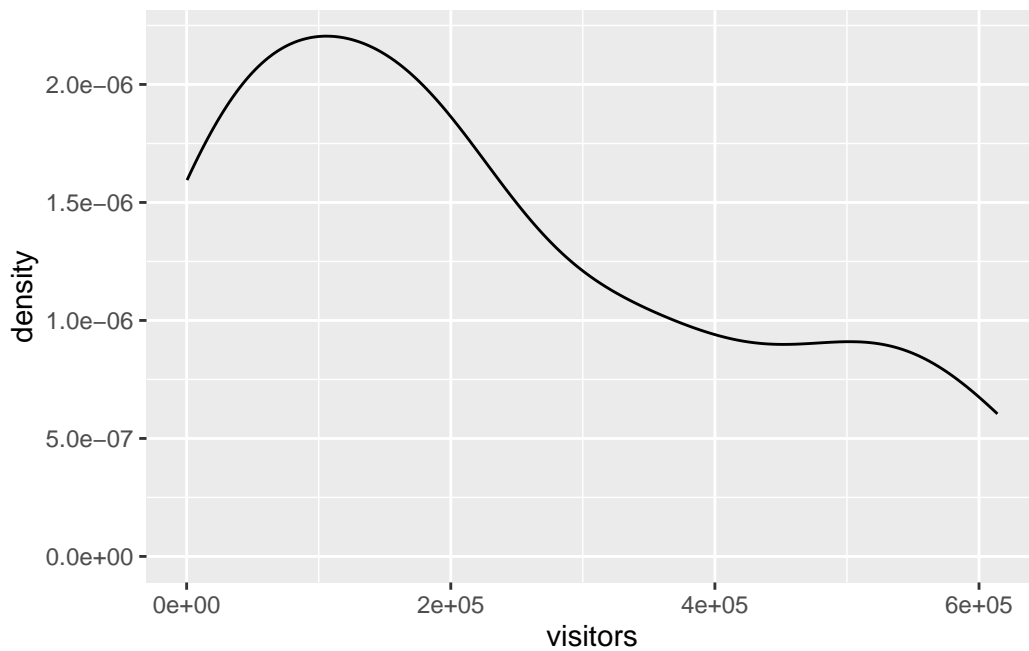


```
# created histogram
ggplot(data = ci_np, mapping = aes(x = visitors)) +
  geom_histogram()
```

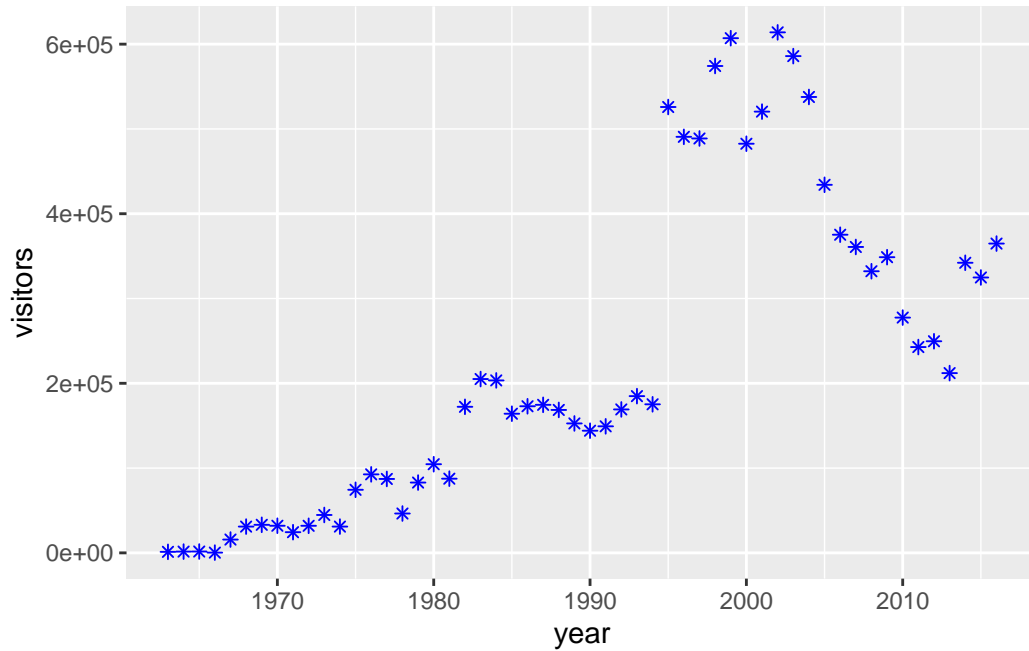
``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



```
# created density line instead of histogram  
ggplot(data = ci_np, mapping = aes(x = visitors)) +  
  geom_density()
```



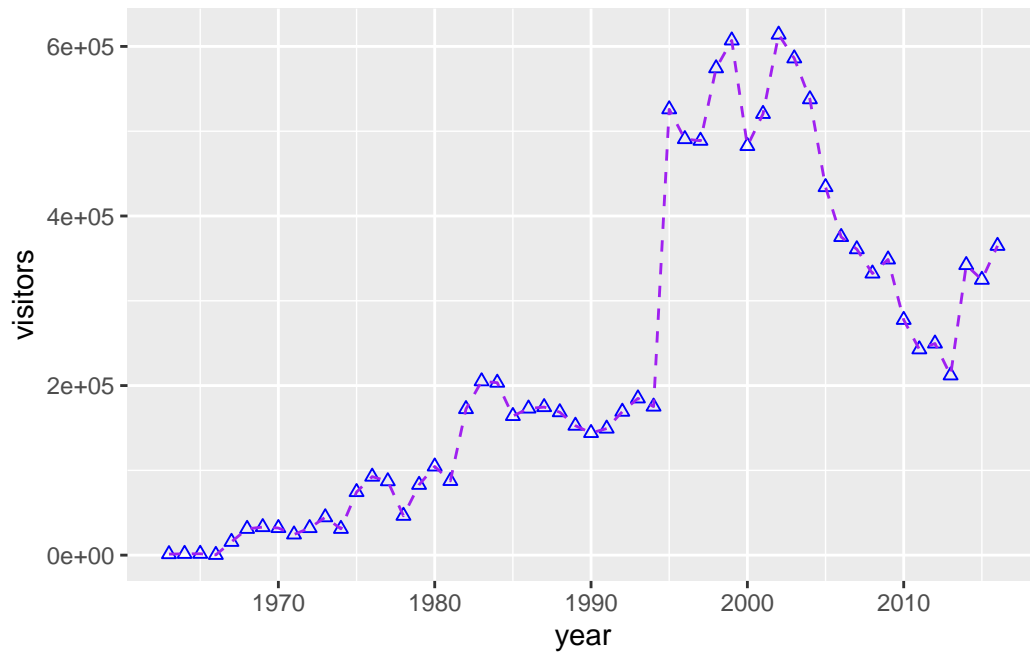
```
# made points on graph blue and differently shaped
ggplot(data = ci_np, mapping = aes(x = year, y = visitors)) +
  geom_point(shape = 8, color = "blue")
```



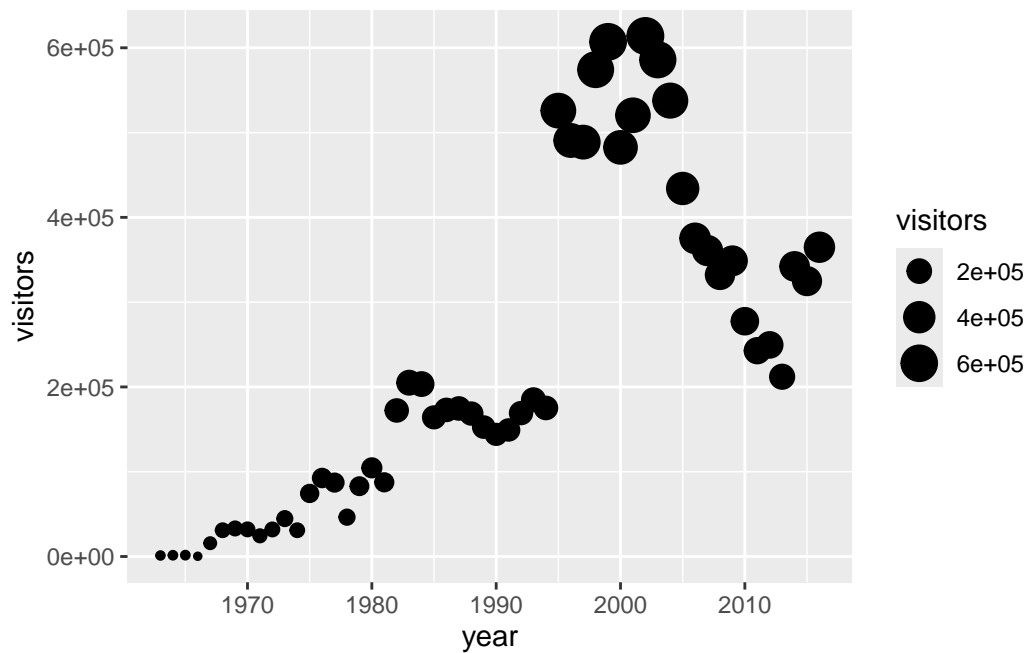
**Q3)**

```
# made line plot but changed line color and type
ggplot(data = ci_np, mapping = aes(x = year, y = visitors)) +
  geom_point(color = "blue",
            shape = 2) +
  geom_line(color = "purple",
            linetype = "dashed")
```

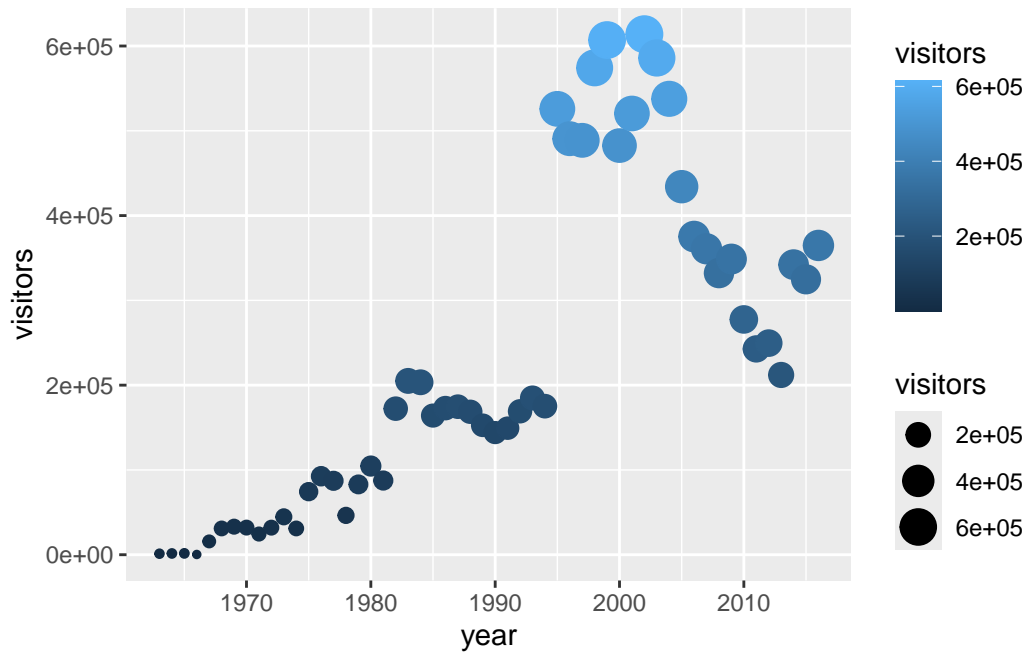




```
# changed point size based on visitor number
ggplot(ci_np, aes(x = year, y = visitors)) +
  geom_point(aes(size = visitors))
```



```
# also color coded the number of visitors
ggplot(ci_np, aes(x = year, y = visitors)) +
  geom_point(aes(size = visitors,
                 color = visitors))
```



Q4)

```
# read new csv file
ca_np <- read.csv("ca_np.csv")

# Fetch the column only - not informative! Too many values!
ca_np$park_name
```

```
[1] "Channel Islands National Park" "Channel Islands National Park"
[3] "Channel Islands National Park" "Channel Islands National Park"
[5] "Channel Islands National Park" "Channel Islands National Park"
[7] "Channel Islands National Park" "Channel Islands National Park"
[9] "Channel Islands National Park" "Channel Islands National Park"
[11] "Channel Islands National Park" "Channel Islands National Park"
[13] "Channel Islands National Park" "Channel Islands National Park"
```



[illegible]

13

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

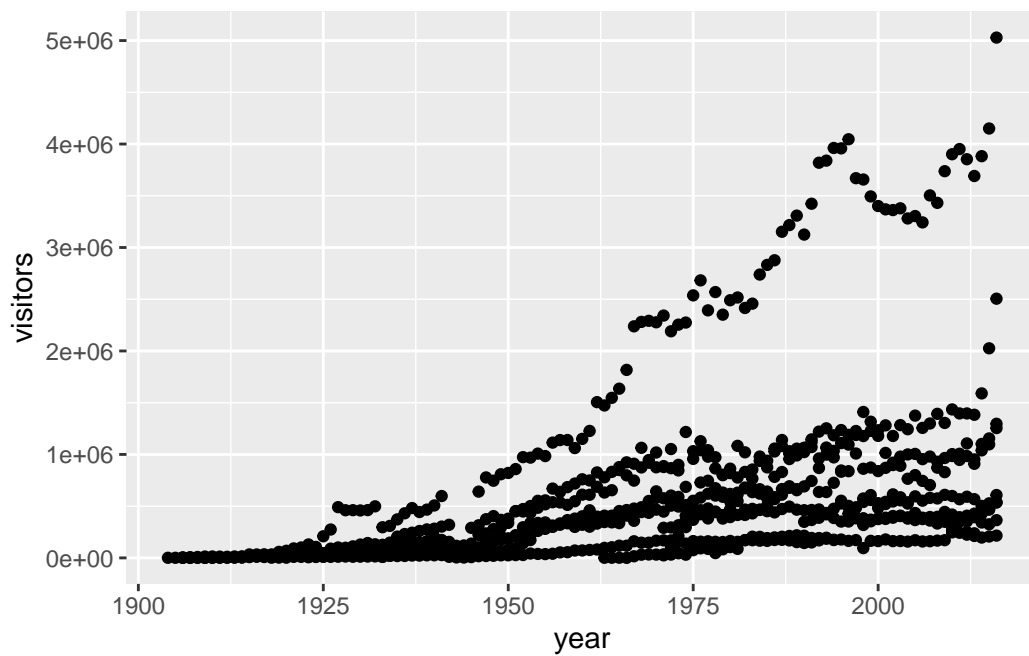


```
[789] "Yosemite National Park"
```

```
# Fetch the column but with the output run through the unique() function  
unique(ca_np$park_name)
```

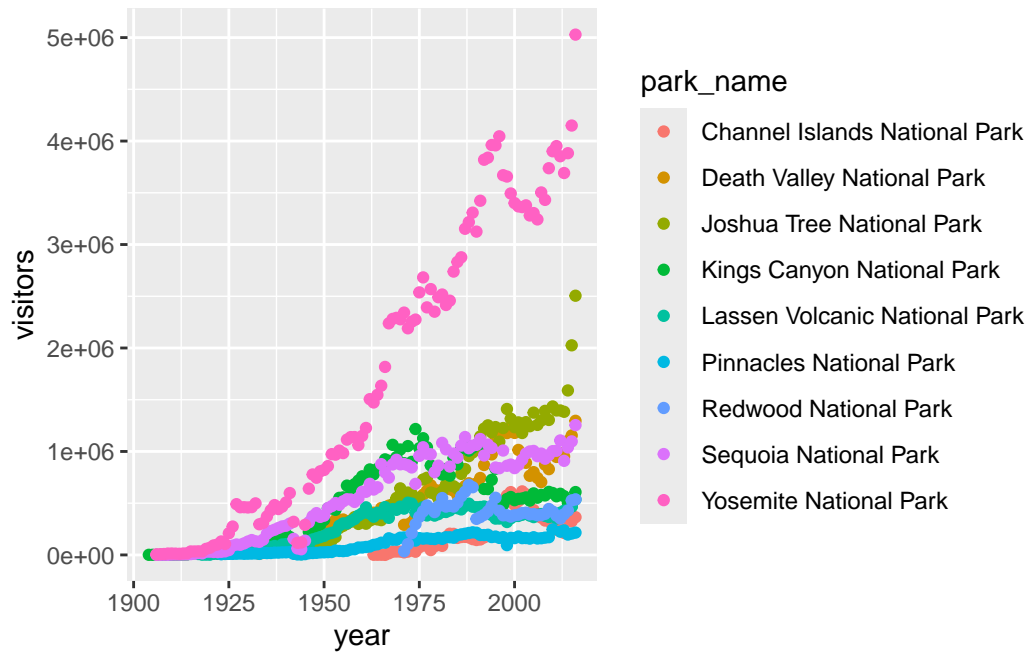
```
[1] "Channel Islands National Park" "Death Valley National Park"  
[3] "Joshua Tree National Park"     "Kings Canyon National Park"  
[5] "Lassen Volcanic National Park" "Pinnacles National Park"  
[7] "Redwood National Park"         "Sequoia National Park"  
[9] "Yosemite National Park"
```

```
# Scatterplot for all parks  
ggplot(data = ca_np, mapping = aes(x = year, y = visitors)) +  
  geom_point()
```

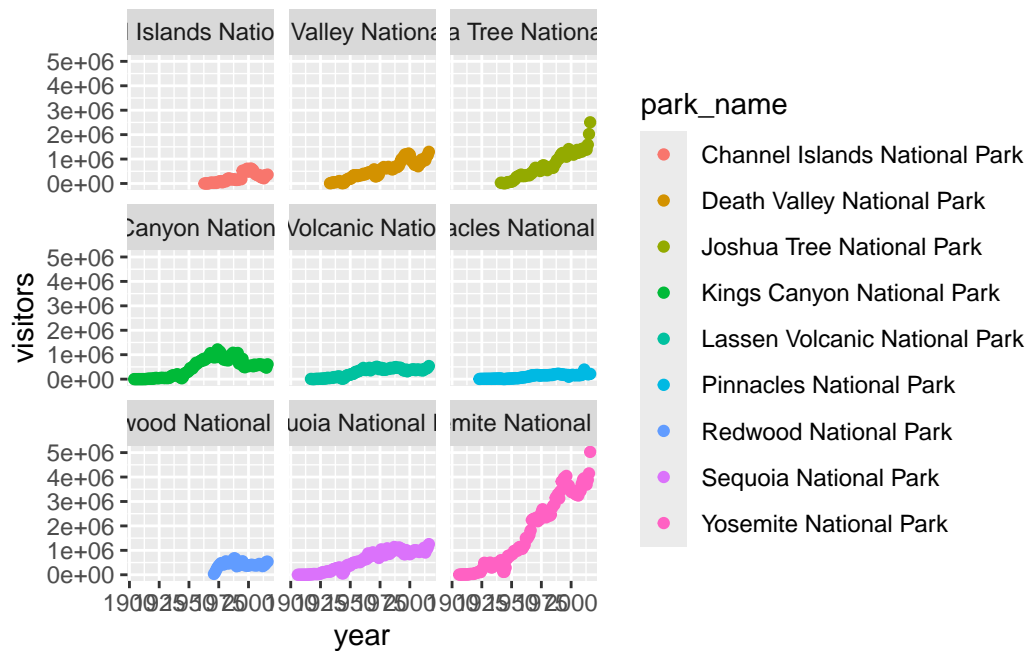


Q5)

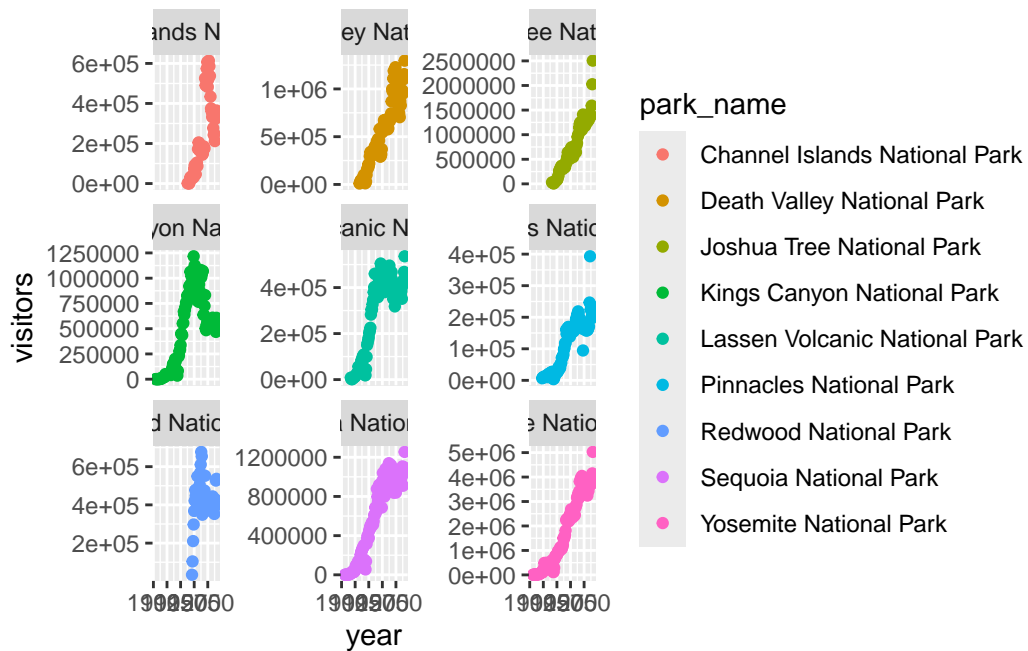
```
# Color coded the different parks  
ggplot(data = ca_np, mapping = aes(x = year, y = visitors)) +  
  geom_point(aes(color = park_name))
```



```
# Added facet_wrap() to separate the different parks
ggplot(data = ca_np, mapping = aes(x = year, y = visitors)) +
  geom_point(aes(color = park_name)) +
  facet_wrap(park_name ~ .)
```



```
# changed the y-axis so each graph goes according to itself
ggplot(data = ca_np, mapping = aes(x = year, y = visitors)) +
  geom_point(aes(color = park_name)) +
  facet_wrap(park_name ~ ., scales = "free_y")
```



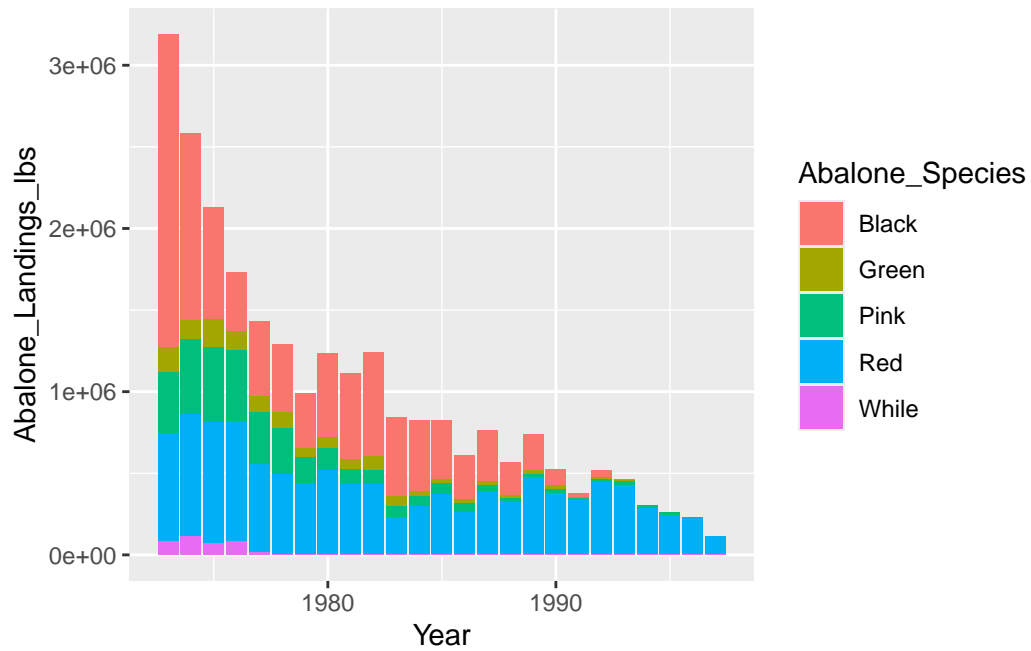
**Q6)**

How does abalone landings differ based on species over a 25 year period?

**Q7)**

```
abalone_landings <- read.csv("abalone_landings.csv")

abalone_landings <- read.csv("abalone_landings.csv")
ggplot(abalone_landings, aes(fill = Abalone_Species , x = Year, y = Abalone_Landings_lbs)) +
  geom_bar(position = "stack", stat = "identity")
```



- 1) I have learned from the graph that in general all of the abalone species surveyed are decreasing. However, some are decreasing at more dramatic rates than others.
- 2) How does abalone landing change based on location? This would require data on where the counts were conducted.