Introduction to ggplot

- **gg** in `ggplot` stands for the grammar of graphics.
- The recipe:
  - Tidy data
  - Mapping *aesthetics* to the plot
  - Geoms - geometric shapes on the plot
  - Coordinate system
  - Lables, guides (legends), and other annotations.
- Dataset
  - Data from the Center for Effective Lawmaking (CEL) from Volden and Wisemen
  - This dataset is great for visualization purposes.
  - Review the codebook and look at the .csv attached.
- Business Problem
  - How does the number of bills passed by a member in a Congress relate to seniority?
  - How long members have been in the chamber?
  - We'll focus on the 115th Congress.

Load the libraries

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

Load the Data

```
data <- read_csv("cel_dataset_coursera.csv")
```

```
## Rows: 10262 Columns: 38
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): thomas_name, st_name
## dbl (36): thomas_num, icpsr, congress, year, cd, dem, elected, female, votep...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# see the column names
names(data)
```

```
##  [1] "thomas_num"     "thomas_name"    "icpsr"          "congress"
##  [5] "year"           "st_name"        "cd"             "dem"
##  [9] "elected"        "female"         "votepct"        "dwnom1"
## [13] "deleg_size"     "speaker"        "subchr"         "afam"
## [17] "latino"         "votepct_sq"     "power"          "chair"
## [21] "state_leg"      "state_leg_prof" "majority"       "maj_leader"
## [25] "min_leader"     "meddist"        "majdist"        "all_bills"
## [29] "all_aic"        "all_abc"        "all_pass"       "all_law"
## [33] "les"            "seniority"      "benchmark"      "expectation"
## [37] "TotalInParty"   "RankInParty"
```

```
# see the dimension of data
dim(data)
```

```
## [1] 10262    38
```

```
# see the statistics summary
summary(data$all_bills)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00   12.00   16.78   21.00  258.00
```

```
table(data$year)
```

```
##
## 1973 1975 1977 1979 1981 1983 1985 1987 1989 1991 1993 1995 1997 1999 2001 2003
##  444  444  443  442  447  444  445  446  449  447  446  445  449  442  447  444
## 2005 2007 2009 2011 2013 2015 2017
##  445  452  451  449  450  443  448
```

Data Visualization

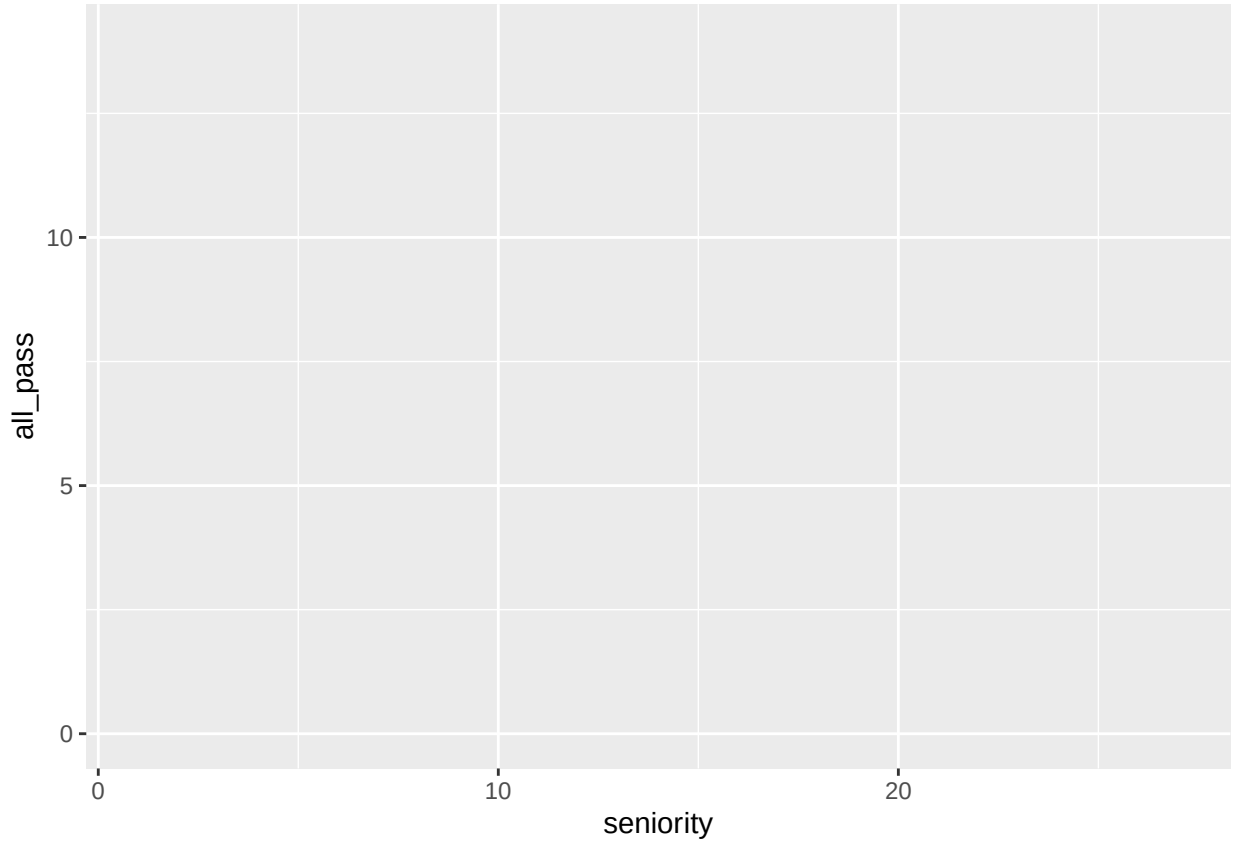
Basic ggplot

- To create a scatter plot, once we add the aesthetic to the `ggplot` function, we can add the geometric shape to it by adding a function called `geom_point`.
- Note that we don't need to pass any parameter inside the function because it inherits the data and aesthetic mapping from the original from the original ggplot command.

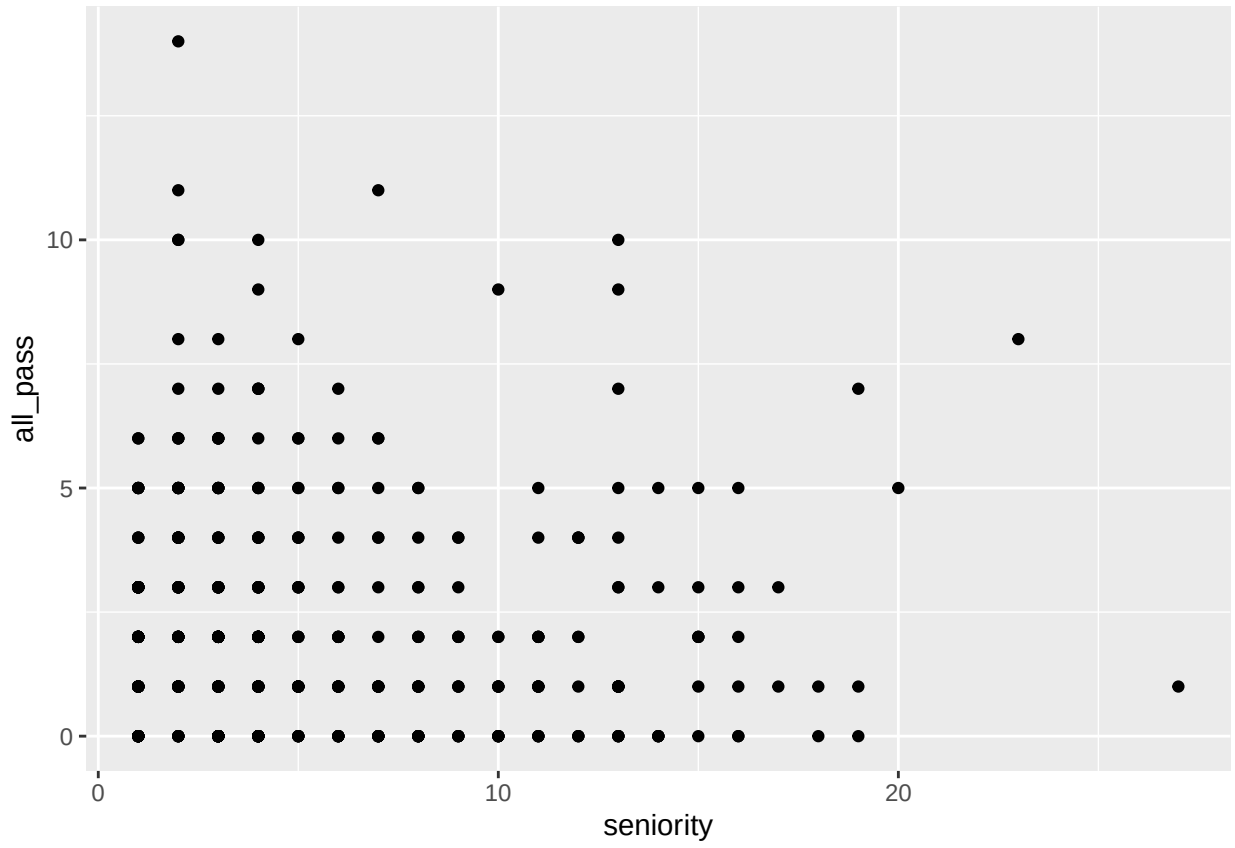
```r
# Filter the data we want
fig115 <- data %>% filter(congress == 115) %>% select("seniority", "all_pass")
head(fig115)
```

```
## # A tibble: 6 x 2
##   seniority all_pass
##       <dbl>    <dbl>
## 1         2        1
## 2         3        2
## 3        11        0
## 4         2        3
## 5         2        1
## 6         4        1
```

```r
# set up the data and aesthetic
ggplot(fig115, aes(x = seniority, y = all_pass)) # it will generate a blank plot with x and y axis
```



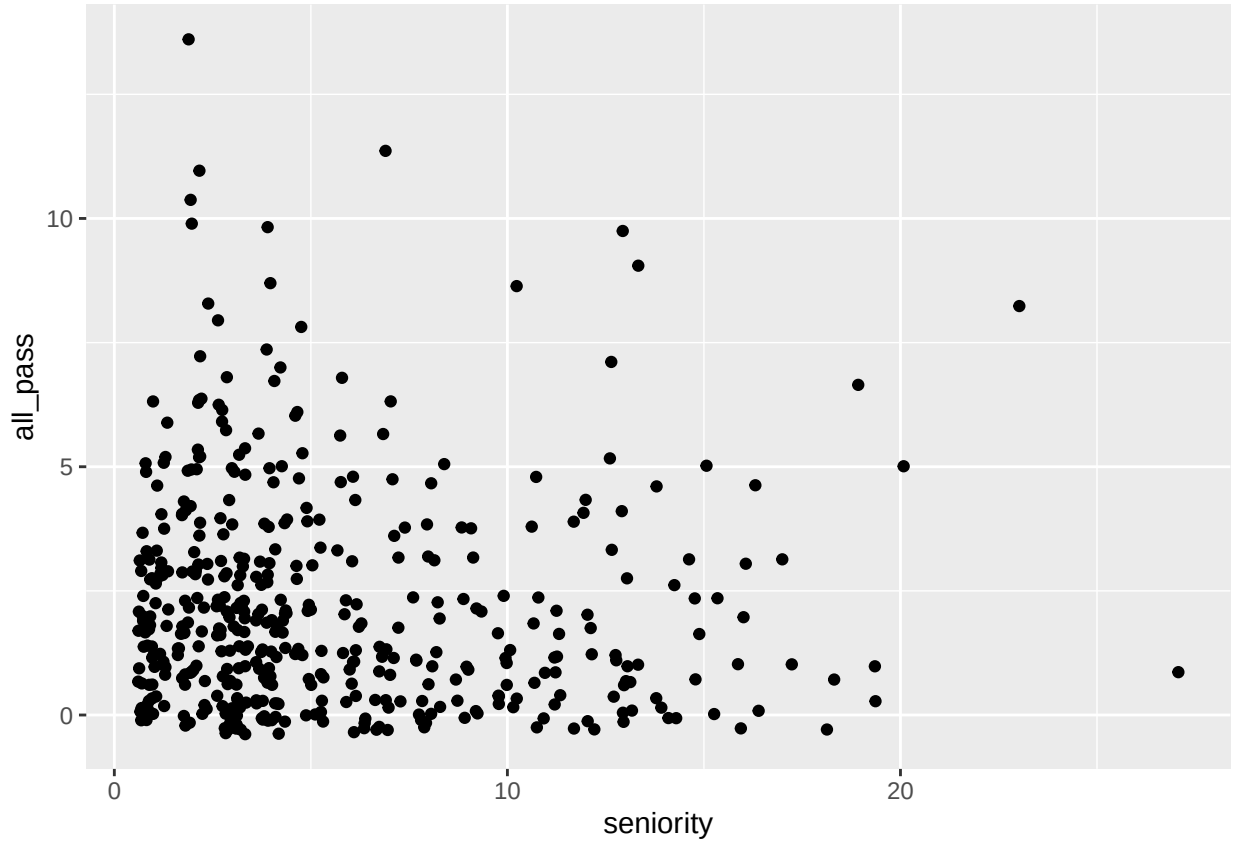

```r
# add the marks or points
ggplot(fig115, aes(x = seniority, y = all_pass))+
  geom_point()
```



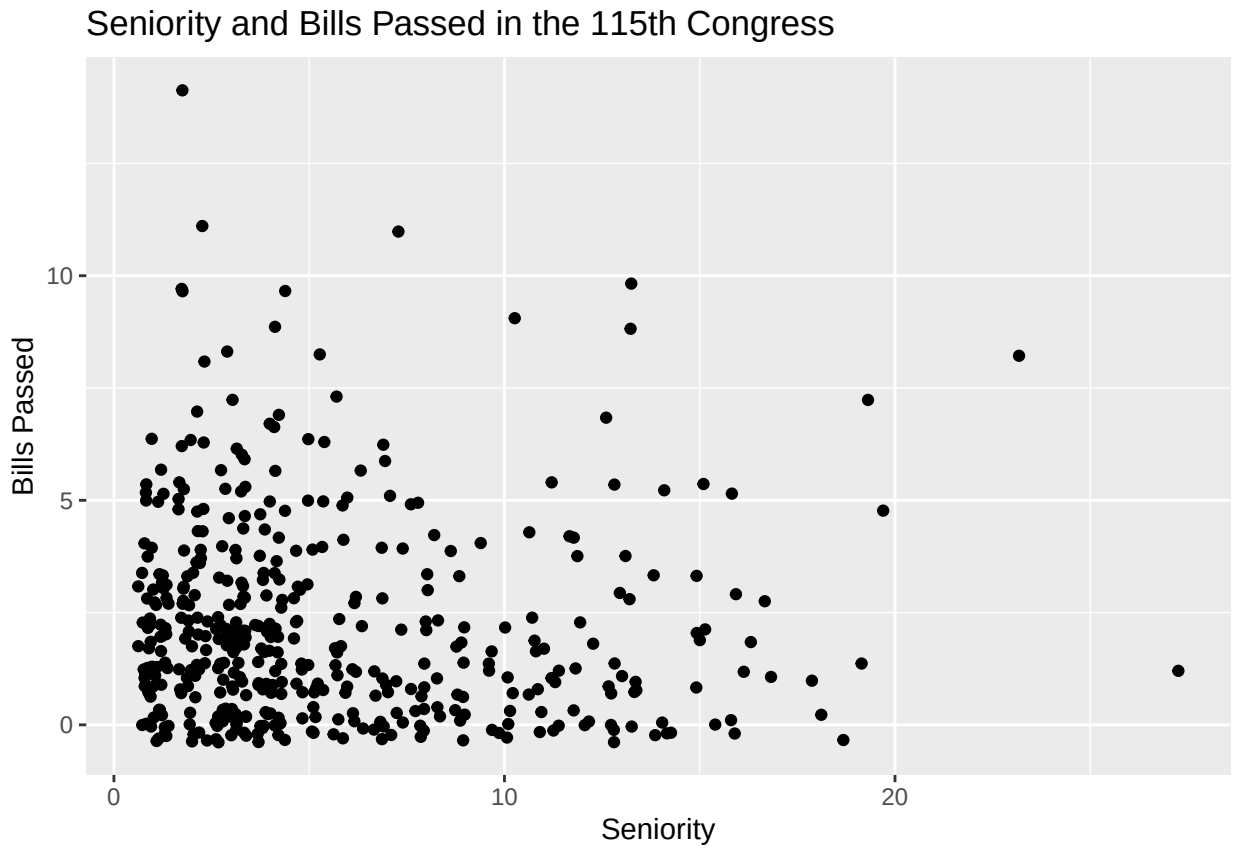
Add Jitter and Labels

- Notice that many points in the scatterplot are on top of each other.
- To solve this, we can use `geom_jitter` to add randoms noise to the data to avoid overplotting.
- This technique is useful to maximize the reader's ability to interpret the data while minimizing any kind of deceptive or manipulating practice.

```
# jitter adds random noise tot he ata to avoid overplotting
ggplot(fig115, aes(x = seniority, y = all_pass))+
  geom_jitter()
```

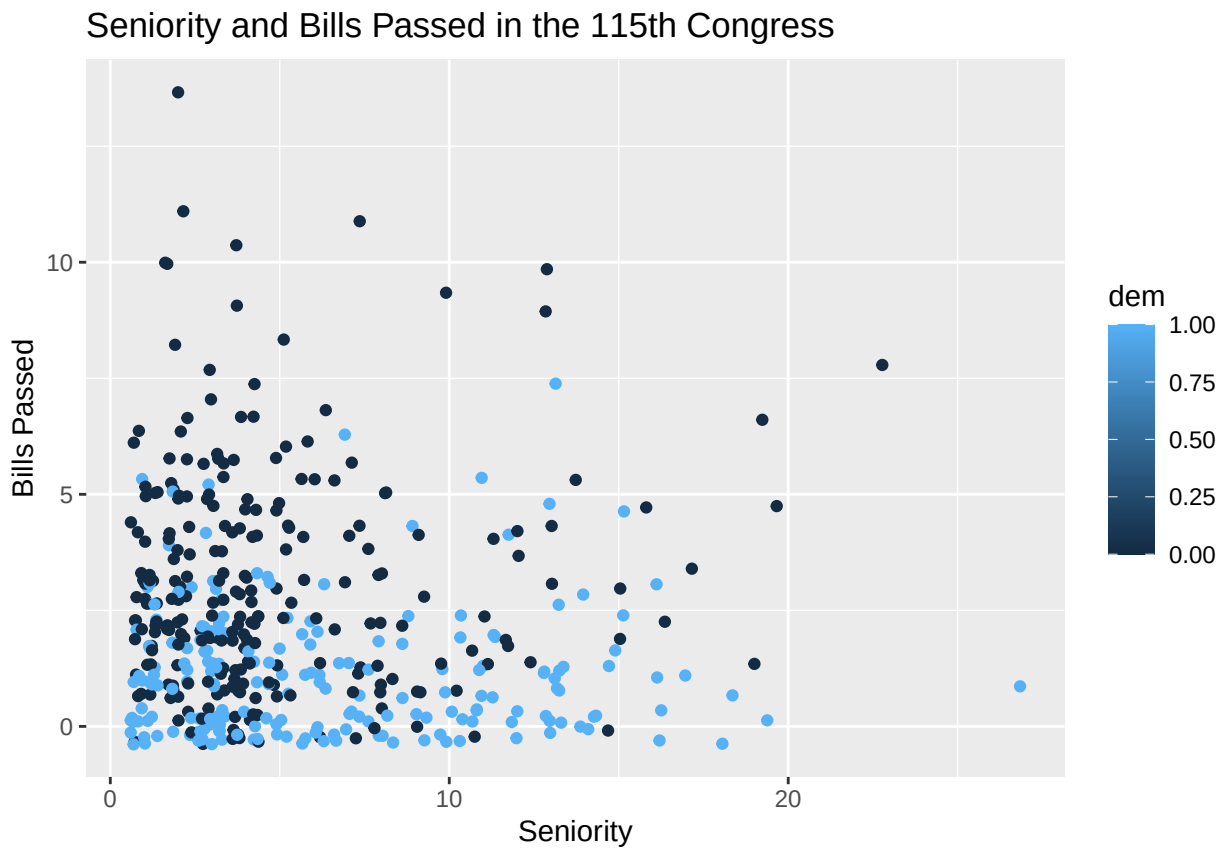



```
# add some labels and a title
ggplot(fig115, aes(x = seniority, y = all_pass))+
  geom_jitter()+
  labs(x = "Seniority", y = "Bills Passed", title = "Seniority and Bills Passed in the 115th Congress")
```



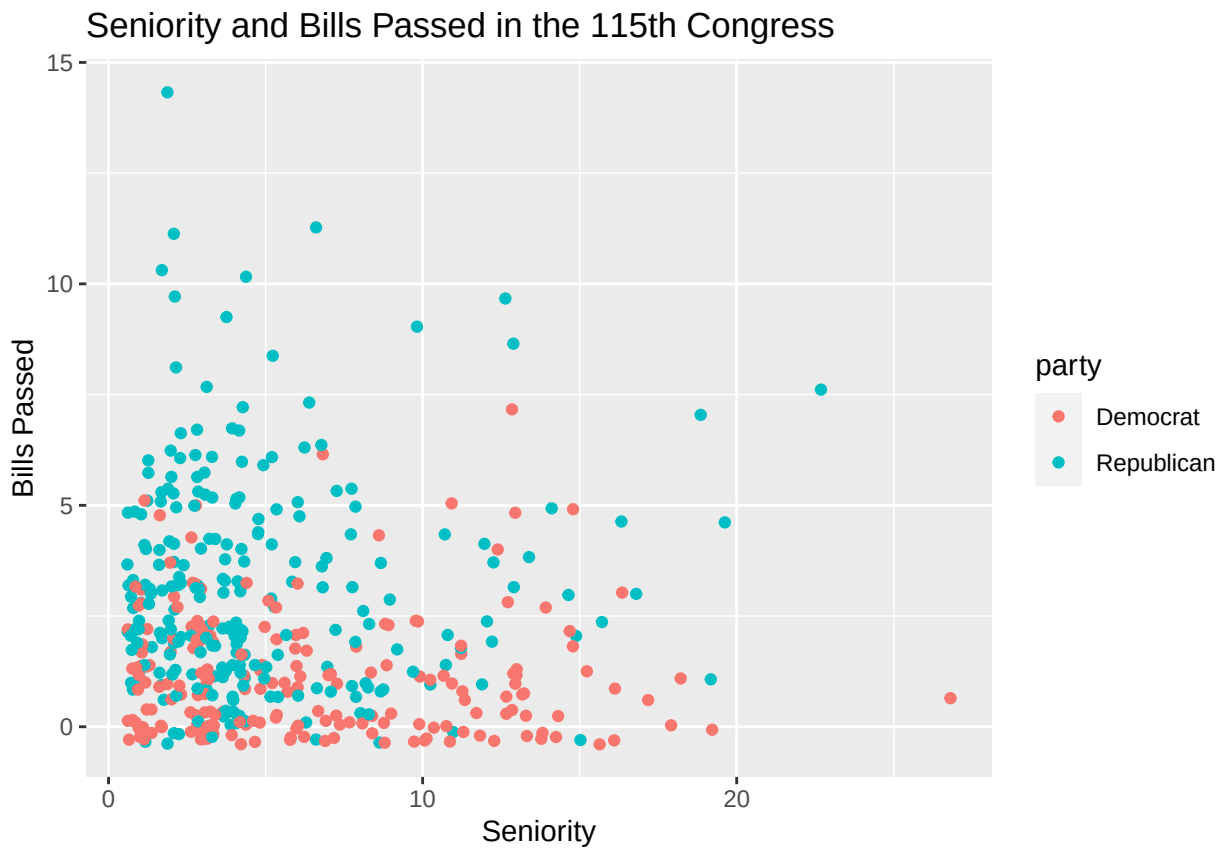
Add and Modify the Color

```
# modify the filter and select to grab "dem"
fig115 <- data %>%
  filter(congress == 115) %>%
  select("seniority", "all_pass", "dem")

ggplot(fig115, aes(x = seniority, y = all_pass, color = dem))+
  geom_jitter()+
  labs(x = "Seniority", y = "Bills Passed", title = "Seniority and Bills Passed in the 115th Congress")
```



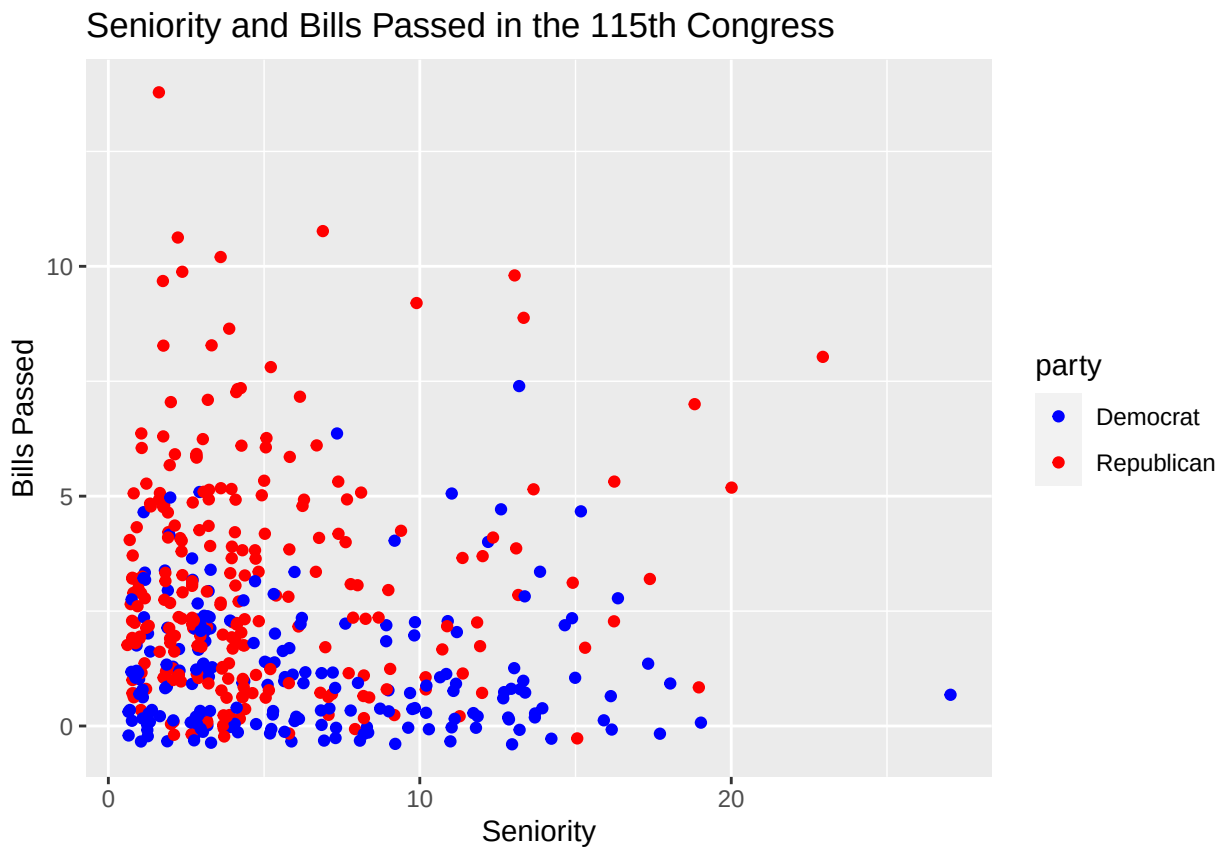
- Notice that the color legend is strange since it's not supposed to be a continuous one rather than a discrete one. Let's fix it!

```
party <- recode(fig115$dem, '1' = 'Democrat', '0' = 'Republican')

fig115 <- add_column(fig115, party)

ggplot(fig115, aes(x = seniority, y = all_pass, color = party))+
  geom_jitter()+
  labs(x = "Seniority", y = "Bills Passed", title = "Seniority and Bills Passed in the 115th Congress")
```



- The color that we just saw is set automatically by R. Let's modify it!
- let's make the colors match traditional blue democrats and red republicans.

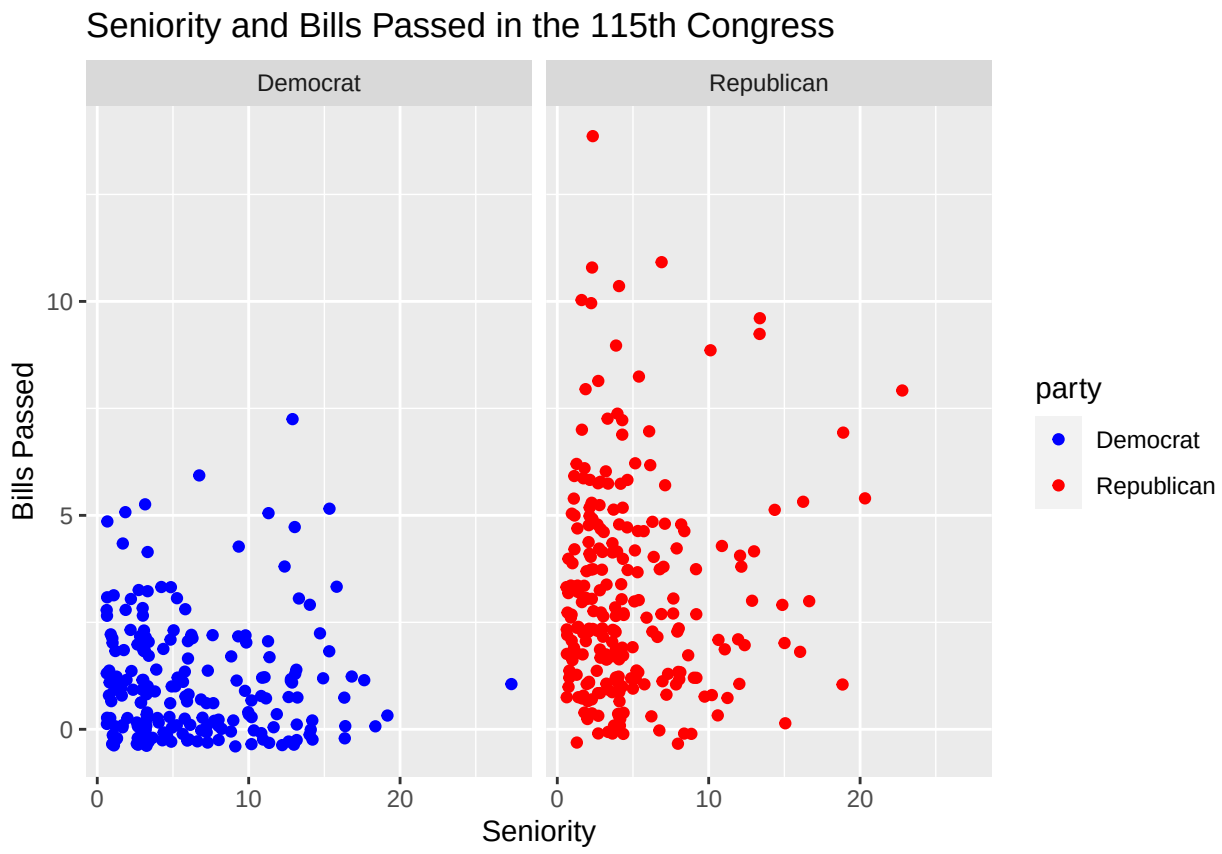
```
ggplot(fig115, aes(x = seniority, y = all_pass, color = party))+
  geom_jitter()+
  labs(x = "Seniority", y = "Bills Passed", title = "Seniority and Bills Passed in the 115th Congress")+
  scale_color_manual(values = c("blue", "red"))
```



Separate the Plot into Subplots

- Let's make the plot into two separate plots using `face_wrap`

```
ggplot(fig115, aes(x = seniority, y = all_pass, color = party))+
  geom_jitter()+
  labs(x = "Seniority", y = "Bills Passed", title = "Seniority and Bills Passed in the 115th Congress")+
  scale_color_manual(values = c("blue", "red"))+
  facet_wrap(~party)
```


Seniority and Bills Passed in the 115th Congress
Democrat
Republican
Bills Passed
10
5
0
0
10
20
Seniority
party
Democrat
Republican