Assignment 2

Nicholas Jacob

2024-08-28

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```
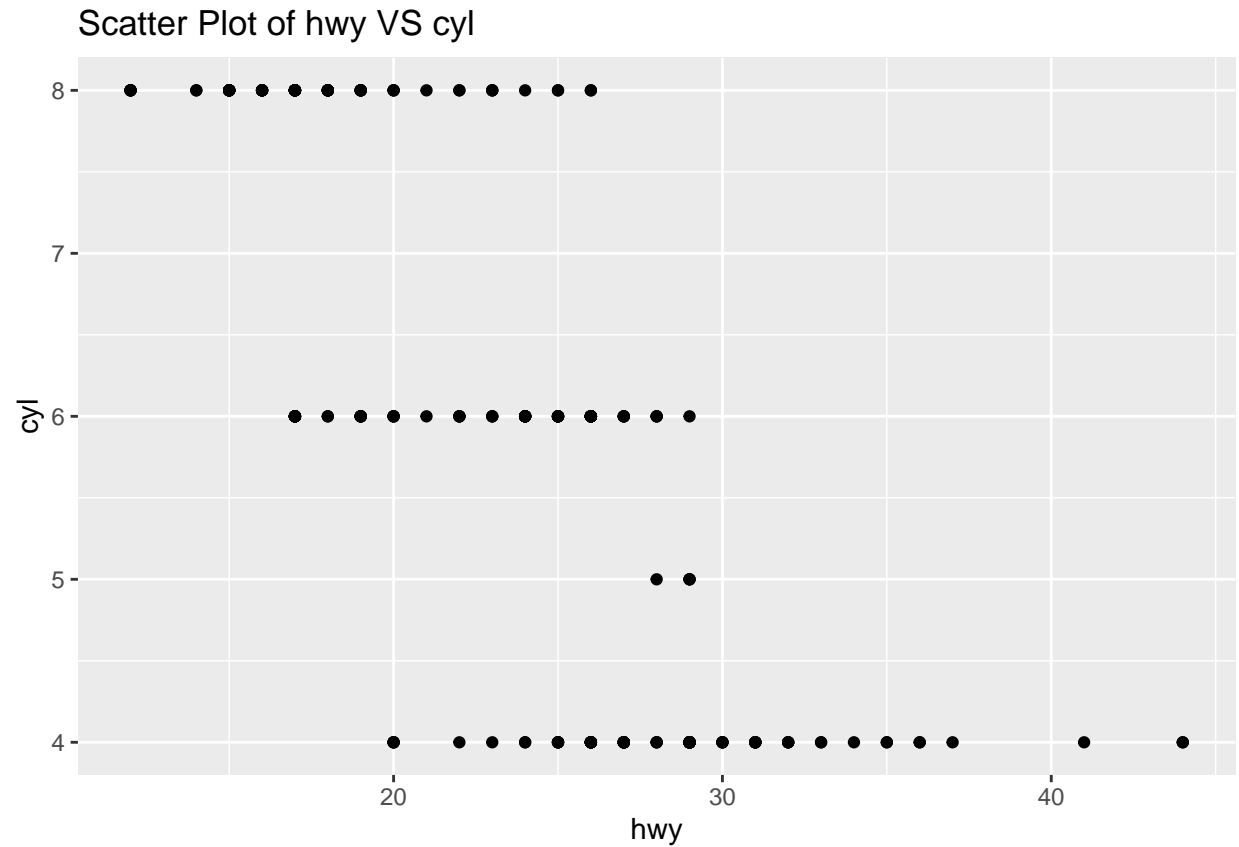
1 Learning `ggplot1`

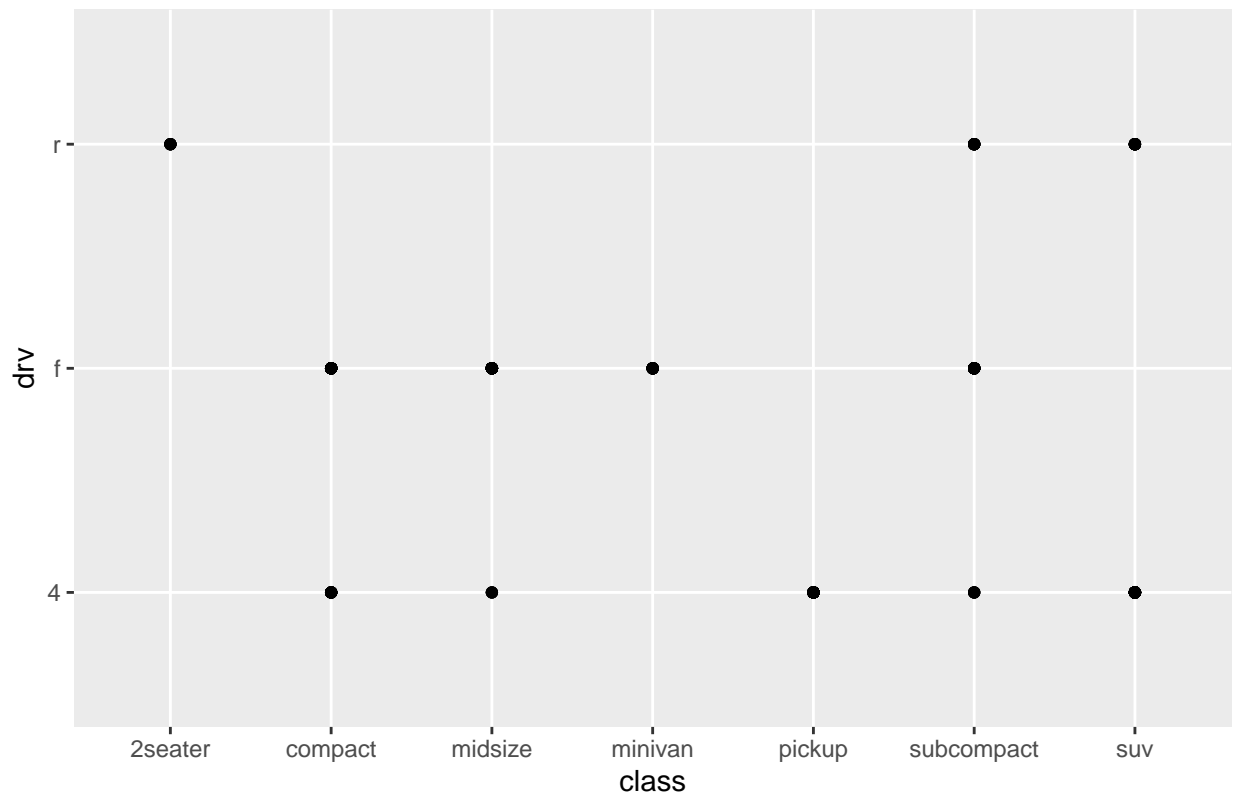   a. §3.2.4 #4 Make a scatterplot of hwy vs cyl.

```r
ggplot(data = mpg, aes(x = hwy,y = cyl)) +
  geom_point() +
  ggtitle("Scatter Plot of hwy VS cyl")
```

Scatter Plot of hwy VS cyl

a. §3.2.4 #5 What happens if you make a scatterplot of class vs drv? Why is the plot not useful?

```
ggplot(data = mpg, aes(x = class,y = drv)) +
  geom_point() +
  ggtitle("Scatter Plot of class VS drv")
```
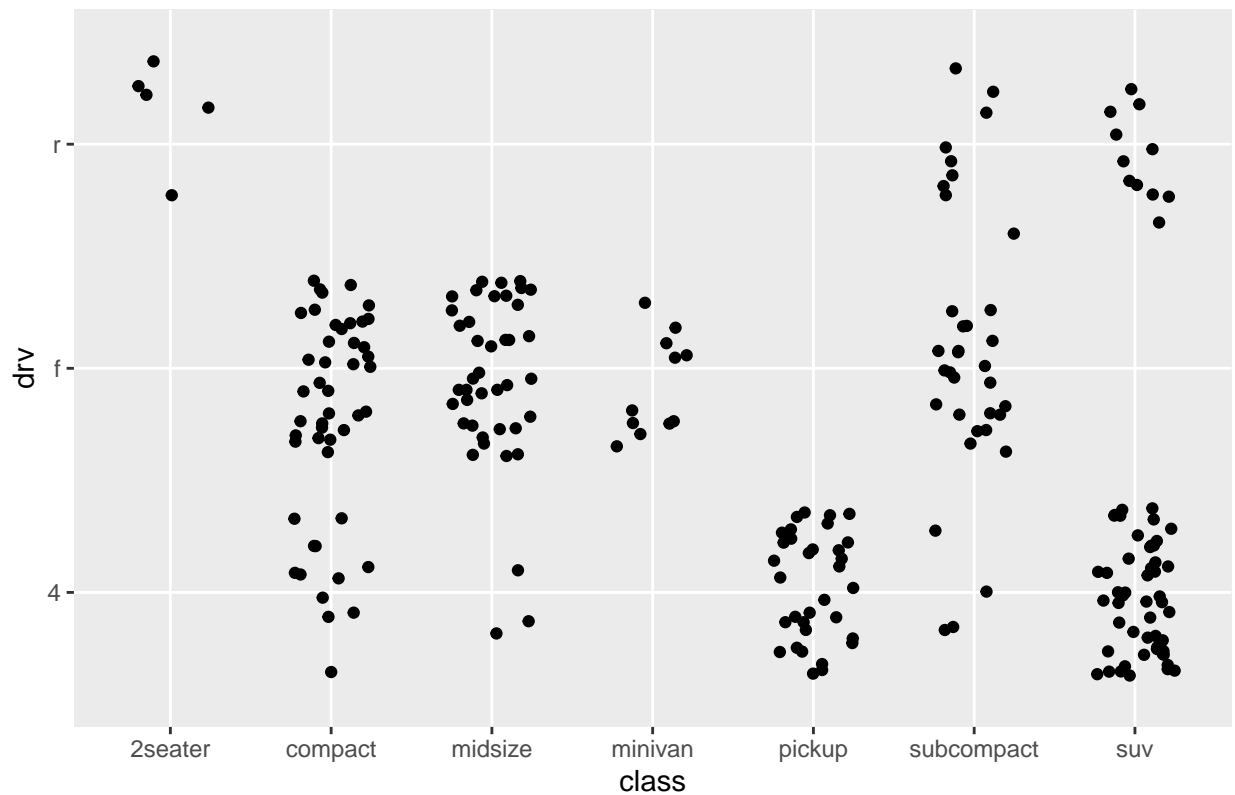
Scatter Plot of class VS drv

Since both are categorical variables, all data points are printed on top of one another. You can use `jitter` to improve the graphic.

```
ggplot(data = mpg, aes(x = class,y = drv)) +
  geom_jitter(width = 0.25) +
  ggtitle("Scatter Plot of class VS drv Now with Jitter")
```

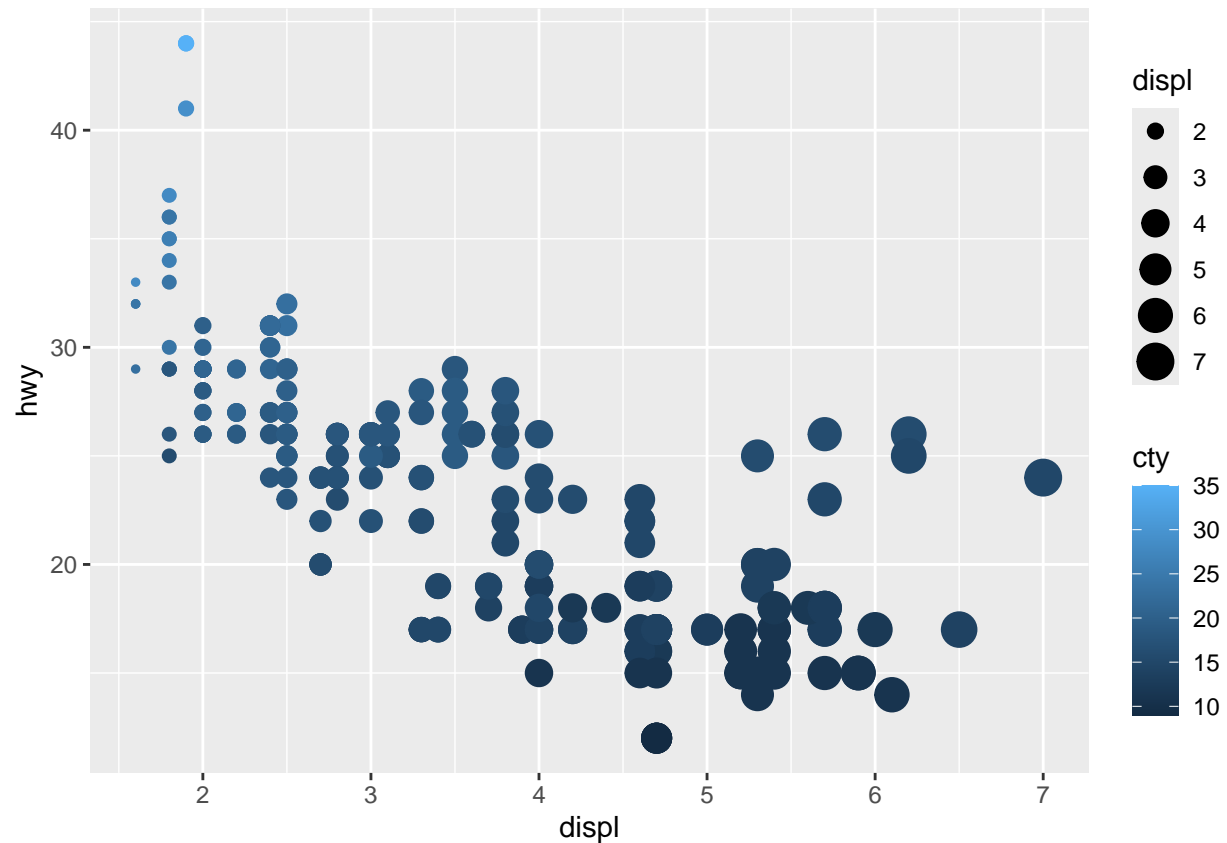## Scatter Plot of class VS drv Now with Jitter

I did relax the jitter a bit to make certain you could tell where the data belonged.

a. §3.3.1 #3 Map a continuous variable to color, size, and shape. How do these aesthetics behave differently for categorical vs. continuous variables?

```r
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = cty, size = displ))
```

Color gave a gradient and size made some differentiation of the size in the point. `shape` through an error so I removed it so that I could have a picture. The error specifically says that shape cannot be a continuous variable.

a. §3.3.1 #4 What happens if you map the same variable to multiple aesthetics?

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = displ))
```

displ

2

3

4

5

6

7

hwy

40

30

20

2   3   4   5   6   7
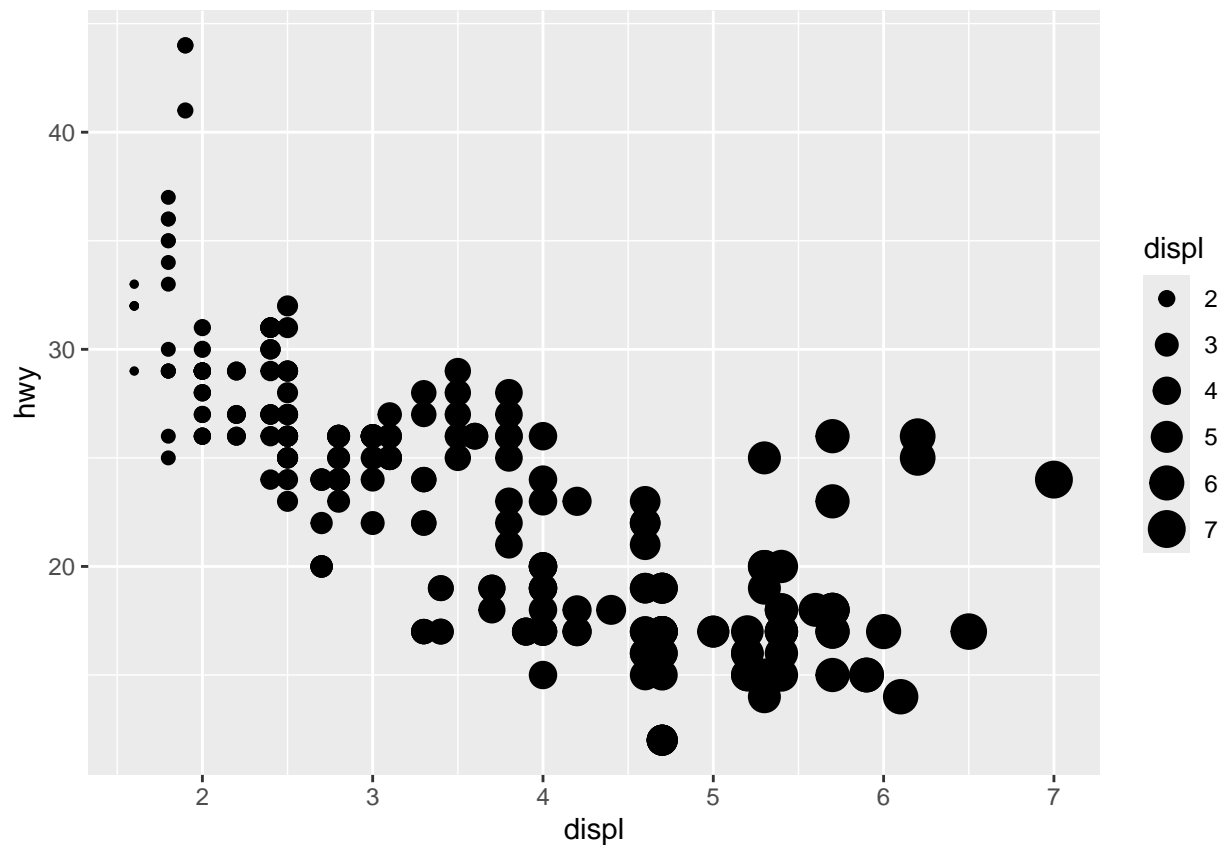
displ

Nothing, it makes the graphic that you asked for. . .

a. §3.3.1 #6 What happens if you map an aesthetic to something other than a variable name, like aes(colour = displ < 5)? Note, you'll also need to specify x and y.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = displ<5))
```

Warning: Using size for a discrete variable is not advised.
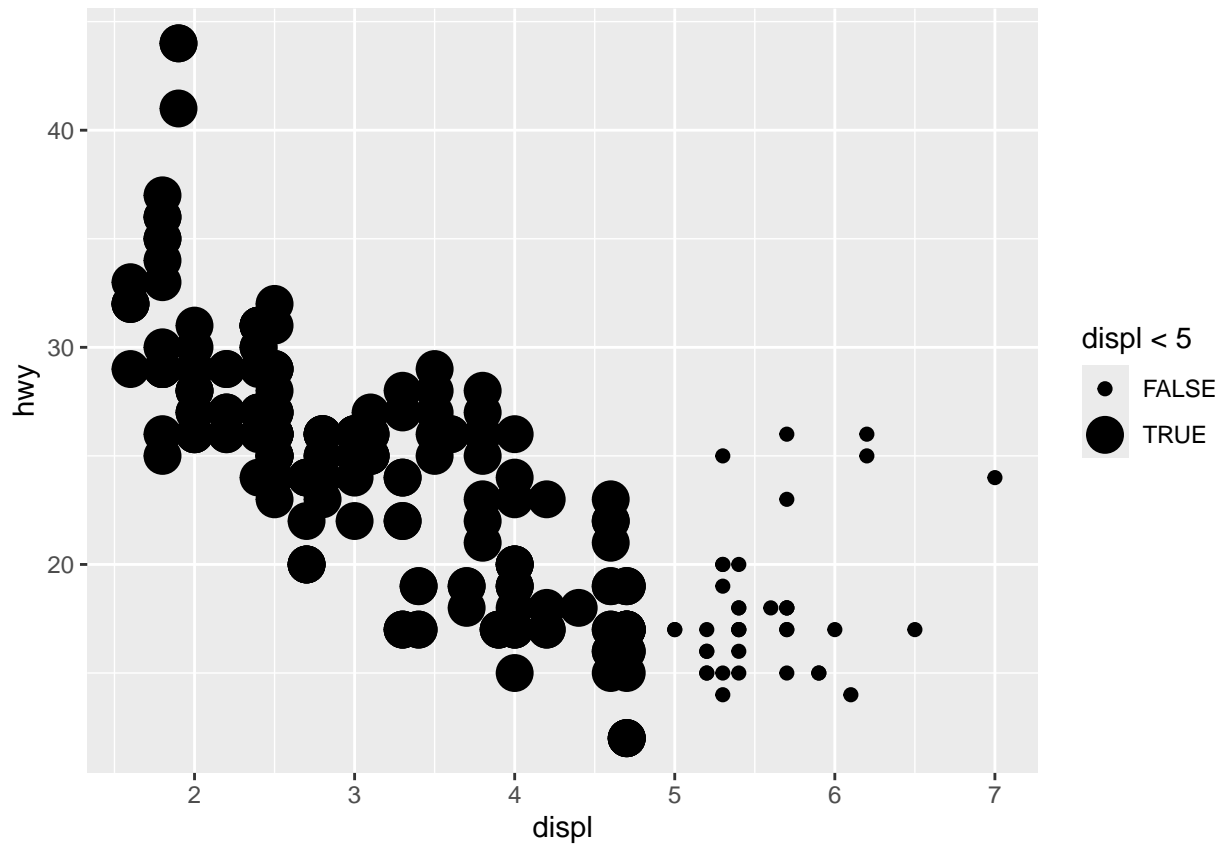
It converted that variable from the dataset into a boolean and then graphed it.

   a. §3.5.1 #4 Take the first faceted plot in this section:

```r
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```
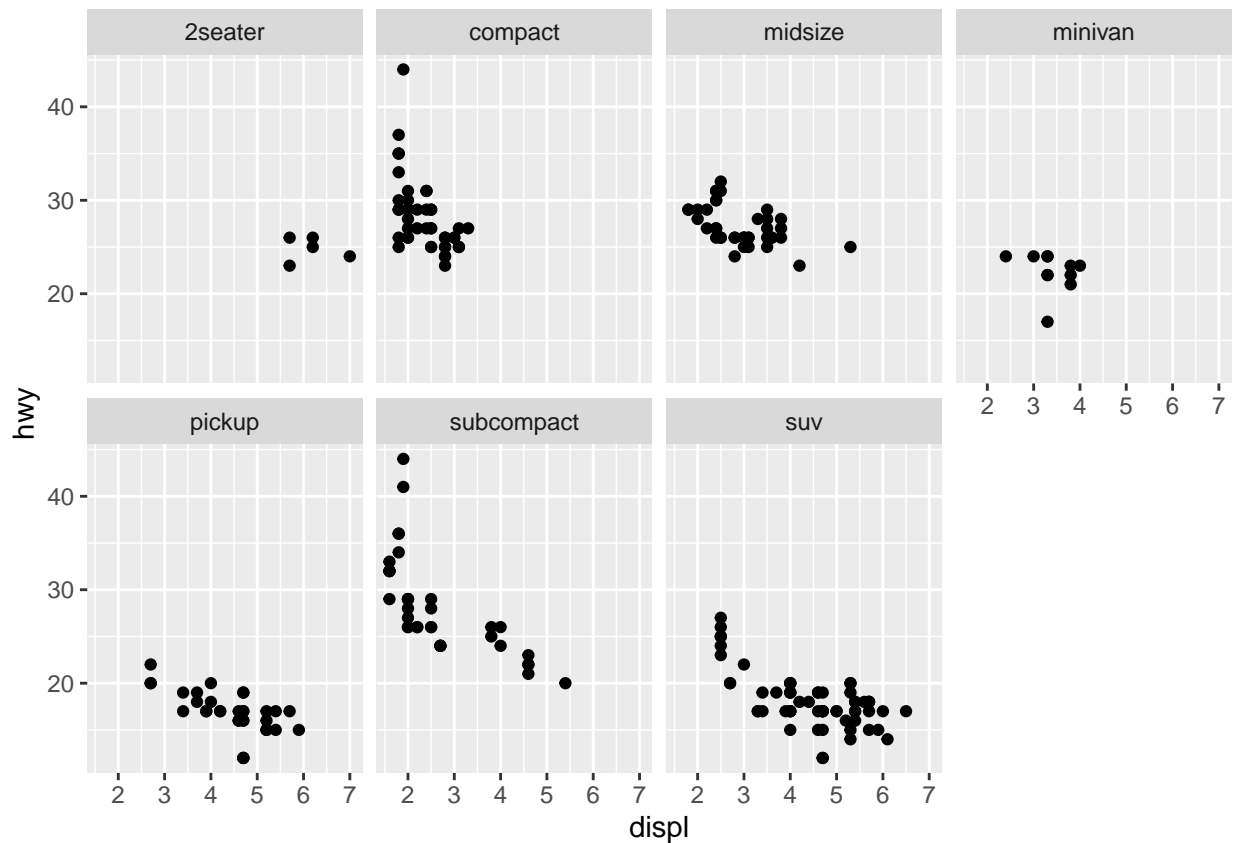
What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

Facet is going to give each `class` by itself. You can quickly see each class and recognize where it congregates in the data. If we had done this with the color aesthetic as below,

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = class))
```
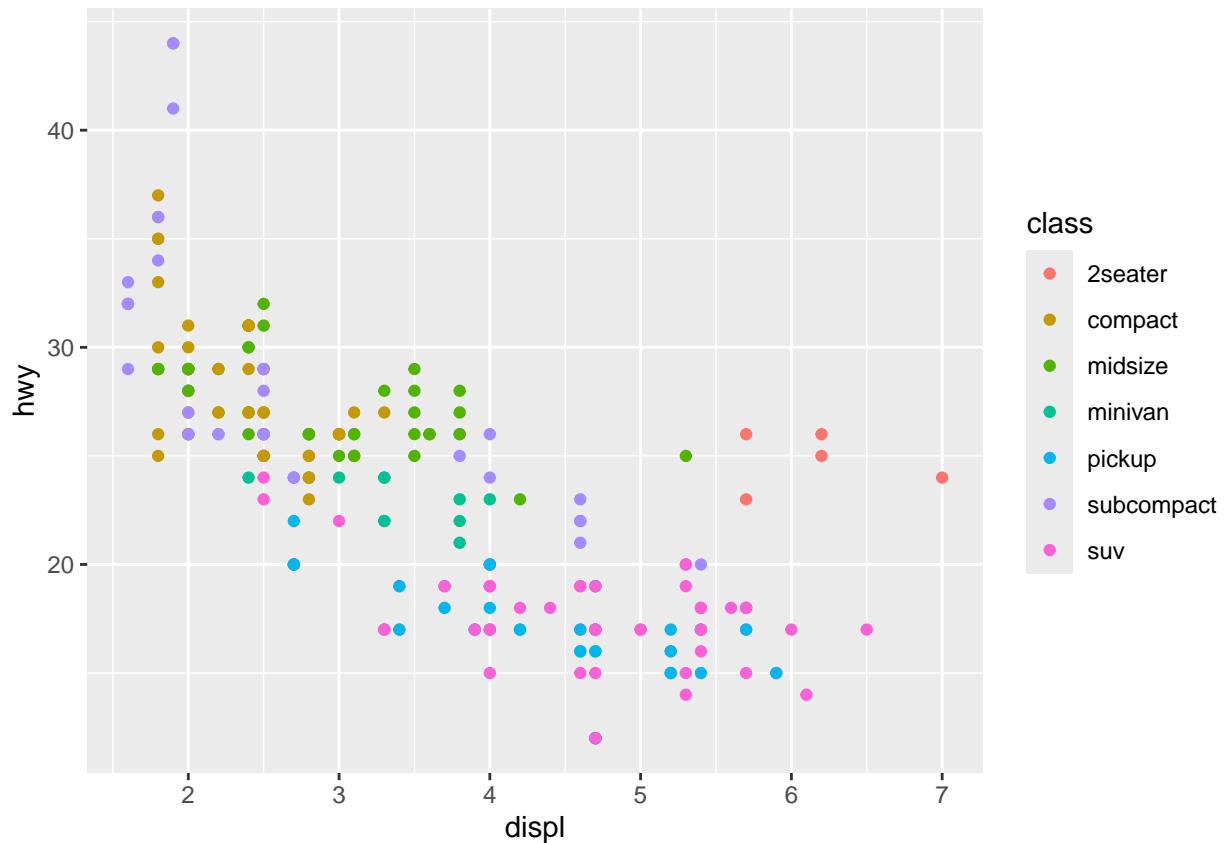
We could observe something similar in the clustering. I am color blind so if there are too many colors, I cannot distinguish them all. The defaults are normally fine for me but the fancier folks try to get the harder it is for me to distinguish the colors. In a large dataset, faceting would be necessary as the color would blob over each other with some entries right on top of each other.

b.

```
ggplot(data = mpg, aes(x = displ, y = hwy ))+
  geom_jitter(alpha = 0.5) + #jitter but not too much
  facet_wrap(~drv) + #facit into three graphs
  geom_smooth(method = "lm", se = FALSE, color = 'black' ) + #straight line of best fit
  geom_smooth(method = "loess") +  #curvy one
  ylab("Highway MPG") + #labels
  xlab("Displacement") + #labels
  ggtitle("Recreating the Master Hadley Wickham") #credit where credit is due
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

Recreating the Master Hadley Wickham
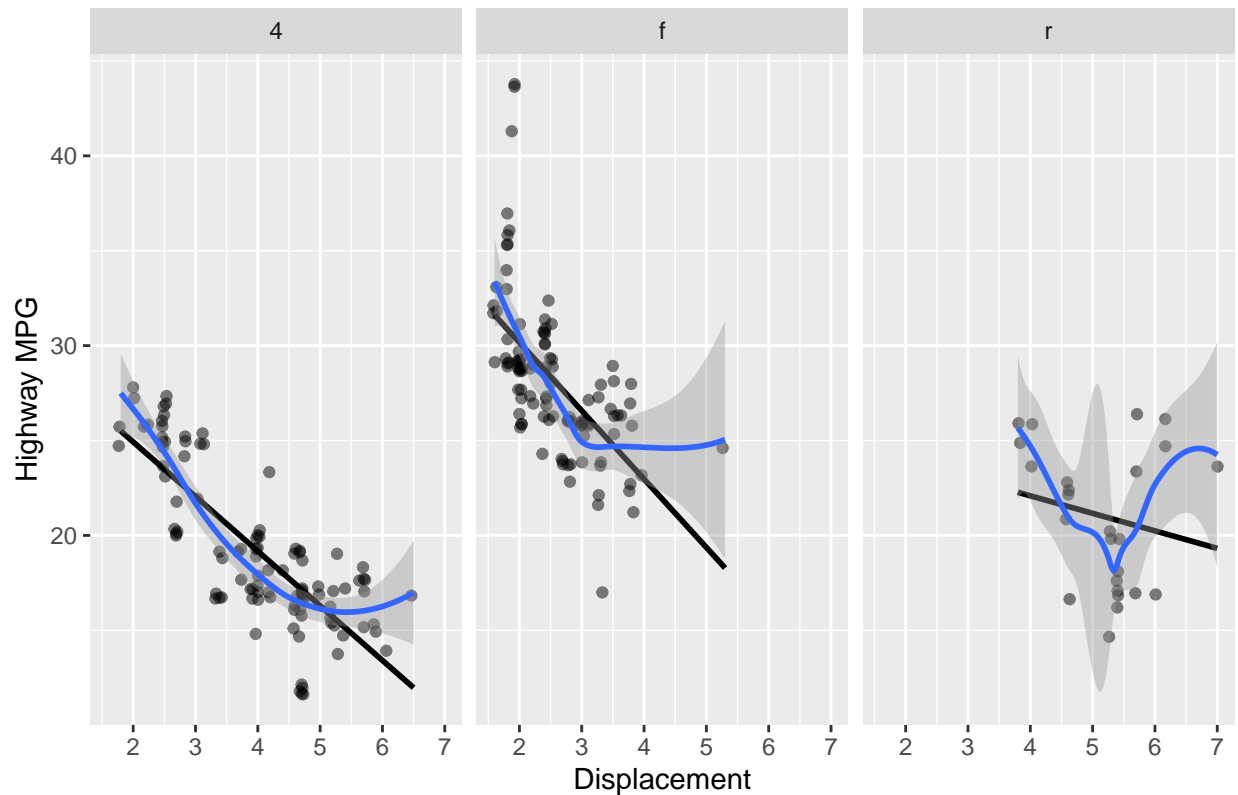
2. House Price Data: EDA and Viz

```r
house <- read.csv('housingData-1.csv')
head(house)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Alley LotShape LandContour
## 1  1         20       RL          NA   11000  <NA>      IR1         Lvl
## 2  2         20       RL          NA   36500  <NA>      IR1         Low
## 3  3         20       RL          57    9764  <NA>      IR1         Lvl
## 4  4         70       RL          NA    7500  <NA>      IR1         Bnk
## 5  5         20       RL          80    9200  <NA>      Reg         Lvl
## 6  6         60       RL          72   11317  <NA>      Reg         Lvl
##   LotConfig LandSlope Neighborhood Condition1 BldgType HouseStyle OverallQual
## 1   CulDSac       Gtl        NAmes       Norm     1Fam     1Story           5
## 2    Inside       Mod      ClearCr       Norm     1Fam     1Story           5
## 3     other       Gtl       Sawyer      Feedr     1Fam     1Story           5
## 4    Inside       Gtl      Crawfor       Norm     1Fam     2Story           6
## 5    Inside       Gtl        NAmes       Norm     1Fam     1Story           6
## 6    Inside       Gtl      CollgCr       Norm     1Fam     2Story           7
##   OverallCond YearBuilt YearRemodAdd RoofStyle Exterior1st Exterior2nd
## 1           6      1966         1966     Gable     Plywood     Plywood
## 2           5      1964         1964     Gable     Wd Sdng     Wd Sdng
## 3           7      1967         2003     Gable     VinylSd     VinylSd
## 4           7      1942         1950     Gable     Wd Sdng     Wd Sdng
## 5           6      1965         1965     Gable     HdBoard     HdBoard
## 6           5      2003         2003     Gable     VinylSd     VinylSd
##   MasVnrType MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond
```

```
## 1     BrkFace      200      Avg      Avg     CBlock      Avg      Avg
## 2     BrkCmn       621      Avg  AboveAvg    CBlock      Avg      Avg
## 3      None          0      Avg      Avg     CBlock      Avg      Avg
## 4      None          0      Avg      Avg     CBlock      Avg      Avg
## 5      None          0      Avg      Avg     CBlock      Avg      Avg
## 6     BrkFace      101  AboveAvg     Avg      PConc  AboveAvg     Avg
##    BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF
## 1           Mn          BLQ        740          Rec        230       184
## 2           Av          Rec        812          Unf          0       812
## 3           No          BLQ        702          Unf          0       192
## 4           No          BLQ        547          Unf          0       224
## 5           No          Rec        892          Unf          0       244
## 6           No          Unf          0          Unf          0       840
##    TotalBsmtSF Heating HeatingQC CentralAir Electrical X1stFlrSF X2ndFlrSF
## 1         1154    GasA  AboveAvg          Y      SBrkr      1154         0
## 2         1624    GasA  BelowAvg          Y      SBrkr      1582         0
## 3          894    GasA  AboveAvg          Y      SBrkr       894         0
## 4          771    GasA  BelowAvg          Y      SBrkr       753       741
## 5         1136    GasA       Avg          Y      SBrkr      1136         0
## 6          840    GasA  AboveAvg          Y      SBrkr       840       828
##    LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
## 1             0      1154            0            0        1        1
## 2             0      1582            0            1        2        0
## 3             0       894            1            0        1        0
## 4             0      1494            0            0        1        0
## 5             0      1136            1            0        1        0
## 6             0      1668            0            0        2        1
##    BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional Fireplaces
## 1             3            1         Avg            6        Typ          1
## 2             4            1         Avg            7        Typ          0
## 3             3            1    AboveAvg            5        Typ          0
## 4             3            1    AboveAvg            7        Typ          2
## 5             3            1         Avg            5        Typ          1
## 6             3            1    AboveAvg            8        Typ          0
##    FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars GarageArea
## 1     BelowAvg     Attchd        1966          RFn          2        480
## 2         <NA>     Attchd        1964          Unf          2        390
## 3         <NA>     Attchd        1967          RFn          2        450
## 4     AboveAvg     Attchd        1942          Unf          1        213
## 5     AboveAvg     Attchd        1965          RFn          1        384
## 6         <NA>     Attchd        2003          RFn          2        500
##    GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF EncPorchSF PoolArea
## 1         Avg        Avg          Y          0          58          0        0
## 2         Avg        Avg          N        168         198          0        0
## 3         Avg        Avg          Y          0           0          0        0
## 4         Avg        Avg          P          0           0        224        0
## 5         Avg        Avg          Y        426           0          0        0
## 6         Avg        Avg          Y        144          68          0        0
##    PoolQC Fence MiscFeature MiscVal MoSold YrSold SaleType SalePrice
## 1    <NA> MnPrv        <NA>       0     11   2009       WD    154000
## 2    <NA>  <NA>        <NA>       0      6   2006       WD    190000
## 3    <NA>  <NA>        <NA>       0      5   2008       WD    130000
## 4    <NA>  <NA>        <NA>       0     11   2009       WD    177500
## 5    <NA>  <NA>        <NA>       0      7   2008       WD    140000
```

```
## 6    <NA>  <NA>       <NA>       0      9    2007       WD     180000
```

First visualization I want to create is the kitchen sink. I've seen this data set before so I am just going to try and get everything I can into one visualization. I am thinking I can get 6 parts of the data...

```
ggplot(data = house, aes(y = SalePrice, x = GrLivArea, color = OverallQual, size = Exterior1st, shape =
  facet_wrap(~Neighborhood) +
  geom_point()+
  scale_x_log10() +
  scale_y_log10()
```

```
## Warning: Using size for a discrete variable is not advised.
```
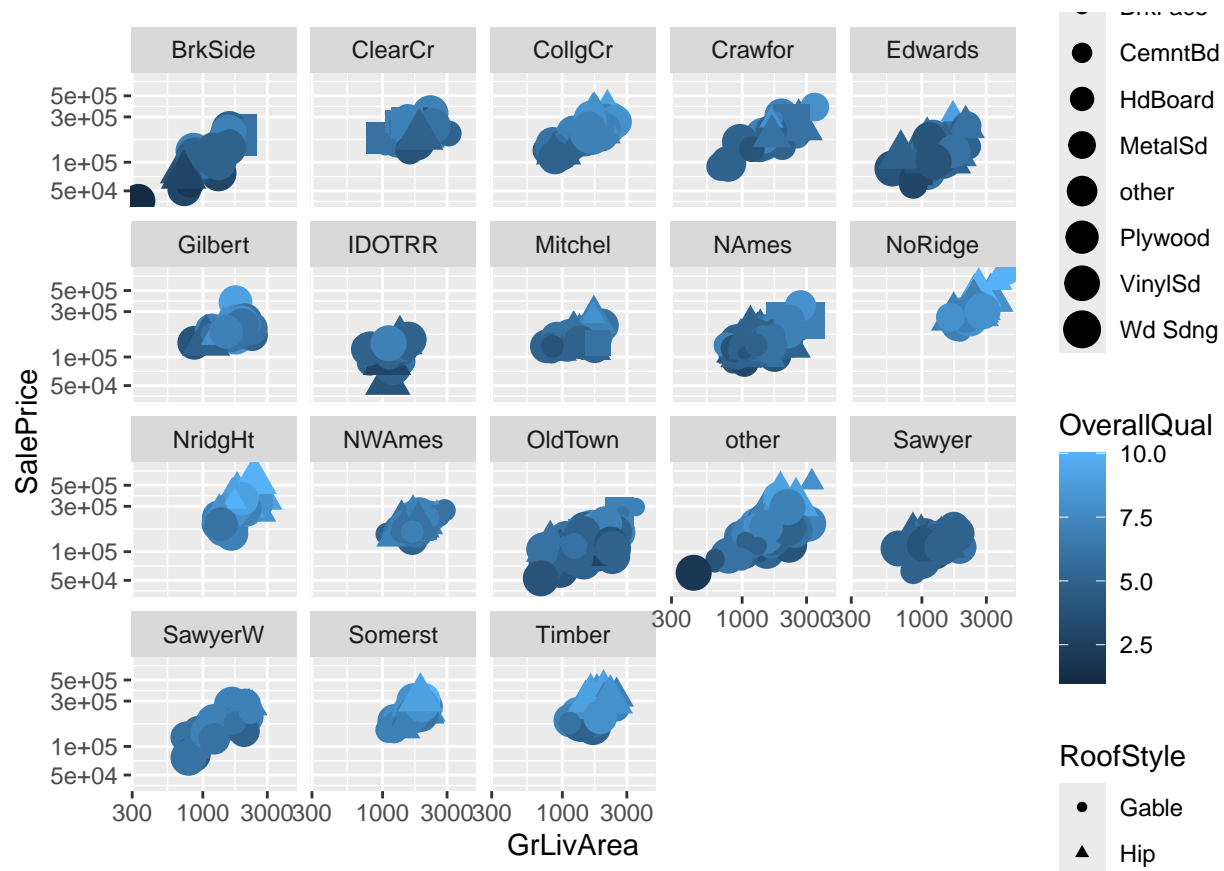

Not as good as I hoped for the kitchen sink method. Did a bit of playing around to try to make it usable but not sure you can glean much from it. I did get 6 peices of info represented in one graph so there is somthing to say for that. I do like the different roof styles as those shapes too.

Let's try another! I did not know what was meant by sploms, but I had seen these before and even had used the GGally package before. I picked the CentralAir for color because I am sweating in my office right now...

```
names(house)[c(3,4,5,6,7,13,14,15,16,17,18,38,74,73,66,67,68)]
```

```
##  [1] "MSZoning"     "LotFrontage"  "LotArea"      "Alley"        "LotShape"
##  [6] "BldgType"     "HouseStyle"   "OverallQual"  "OverallCond"  "YearBuilt"
## [11] "YearRemodAdd" "CentralAir"   "SalePrice"    "SaleType"     "PoolArea"
## [16] "PoolQC"       "Fence"
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
```

```
##    method from
##    +.gg   ggplot2
```

```r
ggpairs(house[,names(house)[c(3,4,5,6,7,13,14,15,16,17,18,38,74,73,66,67,68)]], aes(colour = CentralAir
```

```
## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_density()`).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 207 rows containing missing values

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
## Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
## Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
## Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 207 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 207 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 207 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 207 rows containing missing values

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 207 rows containing missing values

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 207 rows containing missing values

## Warning in cor(x, y): the standard deviation is zero

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
## Removed 207 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

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

```
## Warning: Removed 207 rows containing missing values or values outside the scale range
## (`geom_point()`).

## Warning in cor(x, y): the standard deviation is zero

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

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

## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).

## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).
## Removed 938 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).

## Warning: Removed 938 rows containing missing values or values outside the scale range
## (`stat_boxplot()`).
## Removed 938 rows containing missing values or values outside the scale range
## (`stat_boxplot()`).
## Removed 938 rows containing missing values or values outside the scale range
## (`stat_boxplot()`).
## Removed 938 rows containing missing values or values outside the scale range
## (`stat_boxplot()`).

## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).

## Warning: Removed 938 rows containing missing values or values outside the scale range
## (`stat_boxplot()`).

## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).

## Warning: Removed 938 rows containing missing values or values outside the scale range
## (`stat_boxplot()`).

## Warning: Removed 938 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).
## Removed 938 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

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

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

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

## Warning: Removed 207 rows containing missing values or values outside the scale range
## (`geom_point()`).

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

## Warning in cor(x, y): the standard deviation is zero

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

## Warning: Removed 207 rows containing missing values or values outside the scale range
## (`geom_point()`).

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

## Warning in cor(x, y): the standard deviation is zero

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

## Warning: Removed 207 rows containing missing values or values outside the scale range
## (`geom_point()`).

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

## Warning in cor(x, y): the standard deviation is zero

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

## Warning: Removed 207 rows containing missing values or values outside the scale range
## (`geom_point()`).

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

## Warning in cor(x, y): the standard deviation is zero

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

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

## Warning: Removed 207 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

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

## Warning in cor(x, y): the standard deviation is zero

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

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

## Warning: Removed 207 rows containing missing values or values outside the scale range
## (`geom_point()`).

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

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

## Warning: Removed 998 rows containing non-finite outside the scale range
## (`stat_g_gally_count()`).

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

## Warning: Removed 207 rows containing non-finite outside the scale range
## (`stat_bin()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
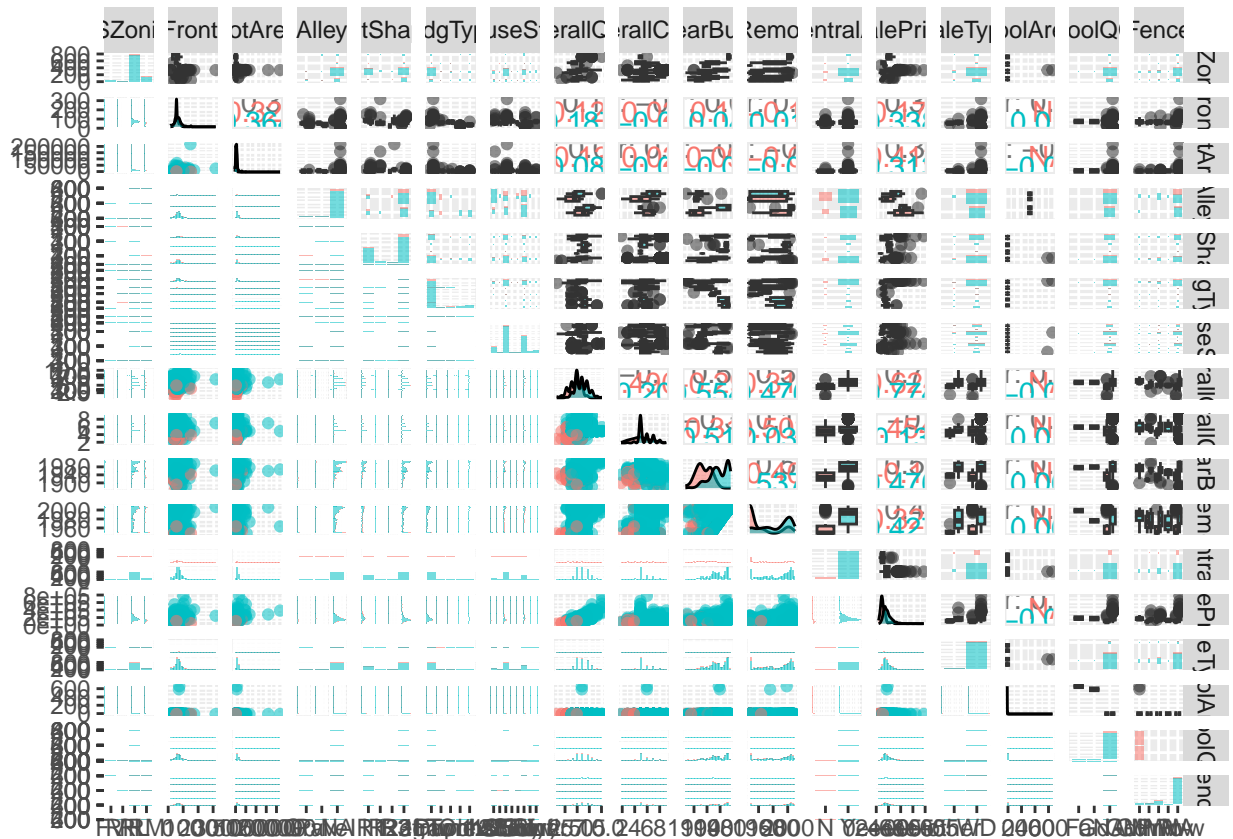
After a bunch of tries to get this code to compile, I limited the data available to it in the hopes of improving runtime. I think the graphic is niffty and perhaps reveals some correlations.

I'm going to keep moving through the suggested plots and try parallel histograms. I am most interested in location, location, location. So I think that effects house price the most.

```
ggplot(data = house, aes(x = SalePrice, color = CentralAir))+
  geom_histogram(fill = "white") +
  facet_wrap(~Neighborhood )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```
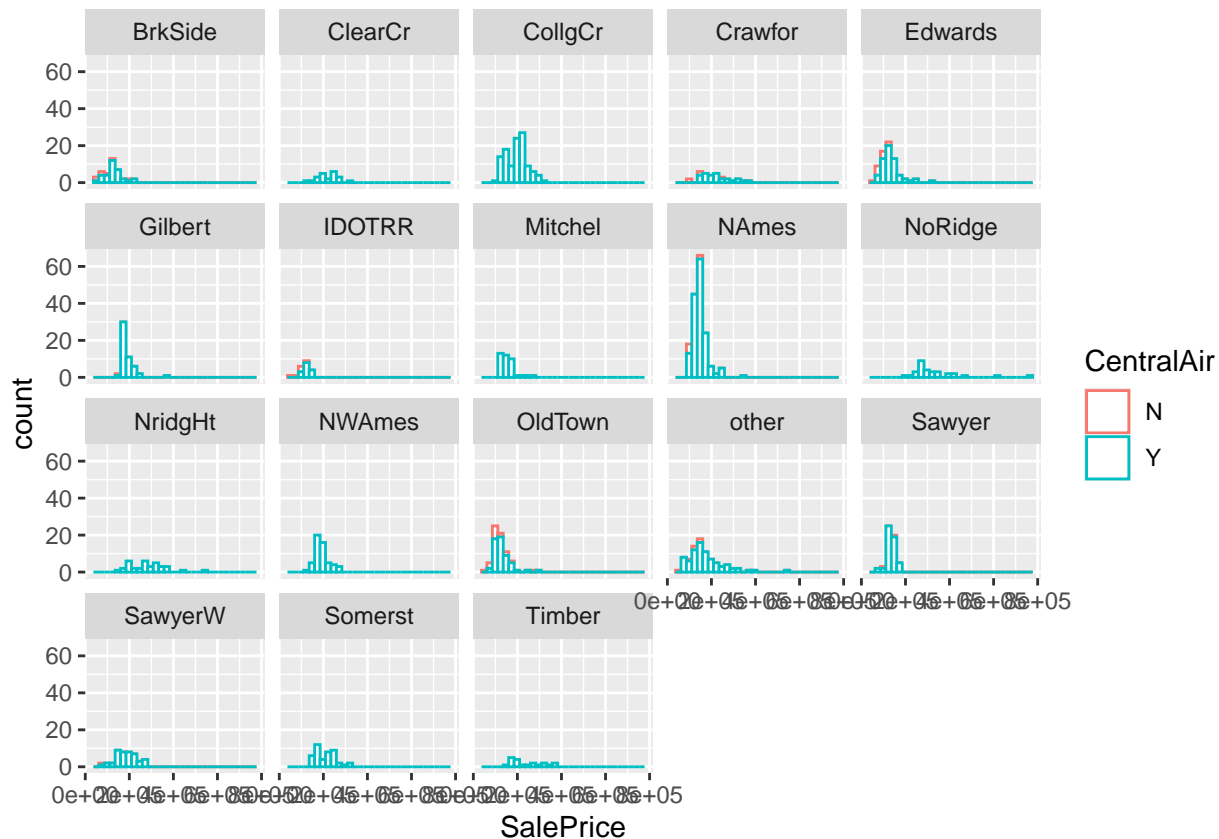
Still a bit fixated on the central air question. Cannot really see many houses without it but where you can they mostly fall in the cheaper range.

I'll try the ridge plot next because I think those are cool

```
library(ggridges)

ggplot(data = house, aes(x = SalePrice, y = Neighborhood, fill = Neighborhood))+
  geom_density_ridges()+
  theme_ridges()+
  theme(legend.position = "none")
```
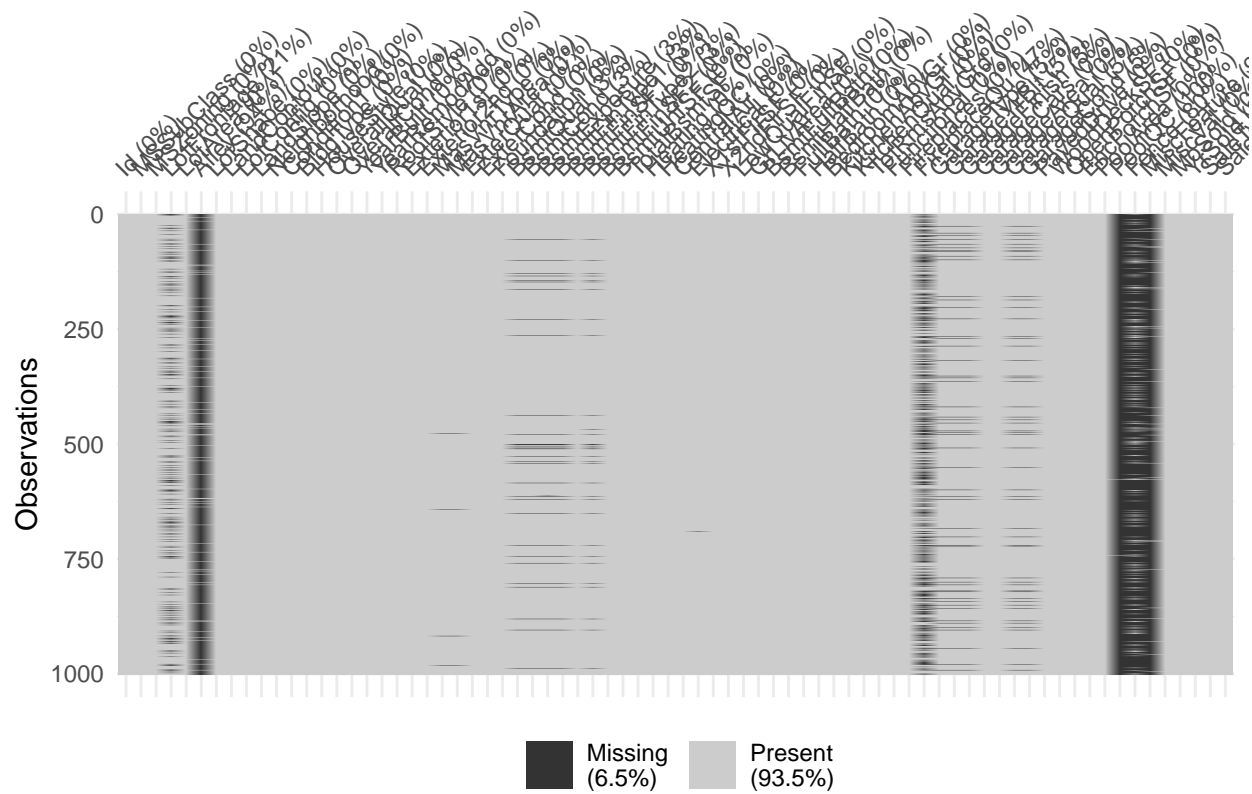
```
## Picking joint bandwidth of 17000
```
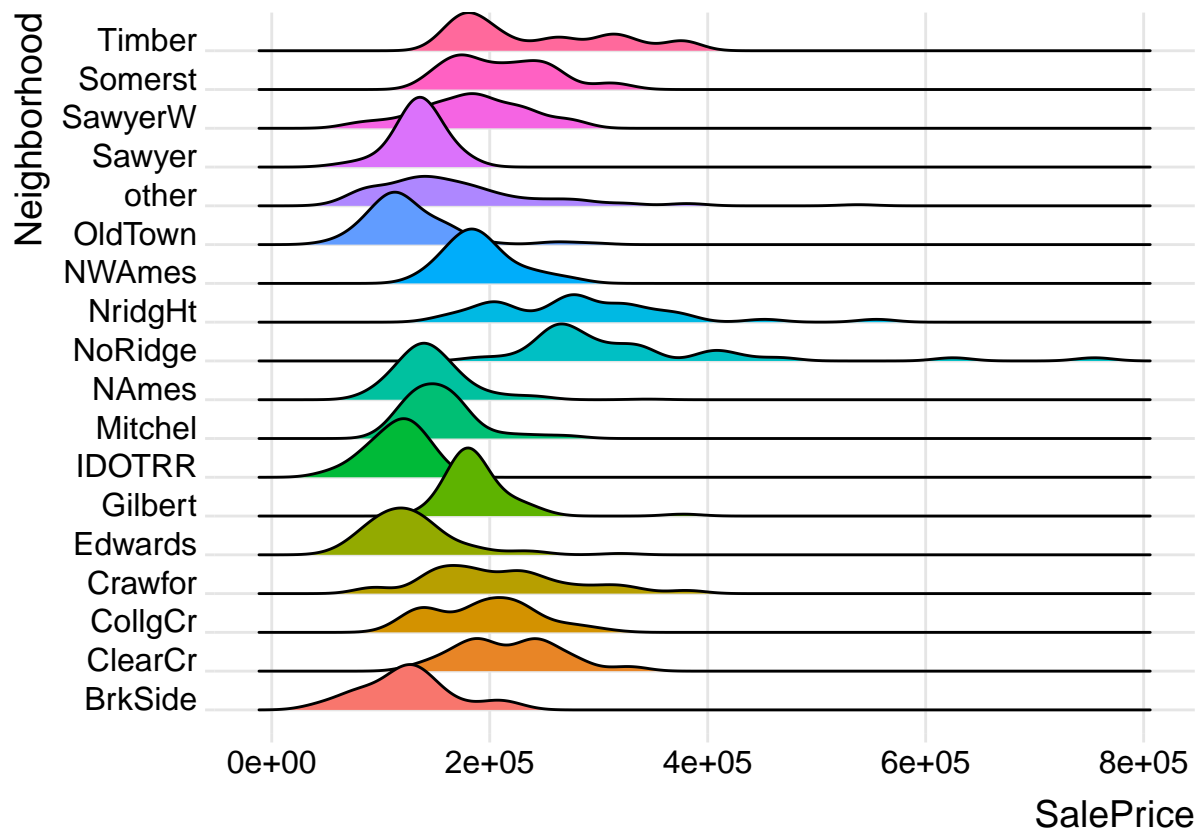
I had never seen a missing value visualization so I'll try a few here for my last exercise. Of course you need another library for it.
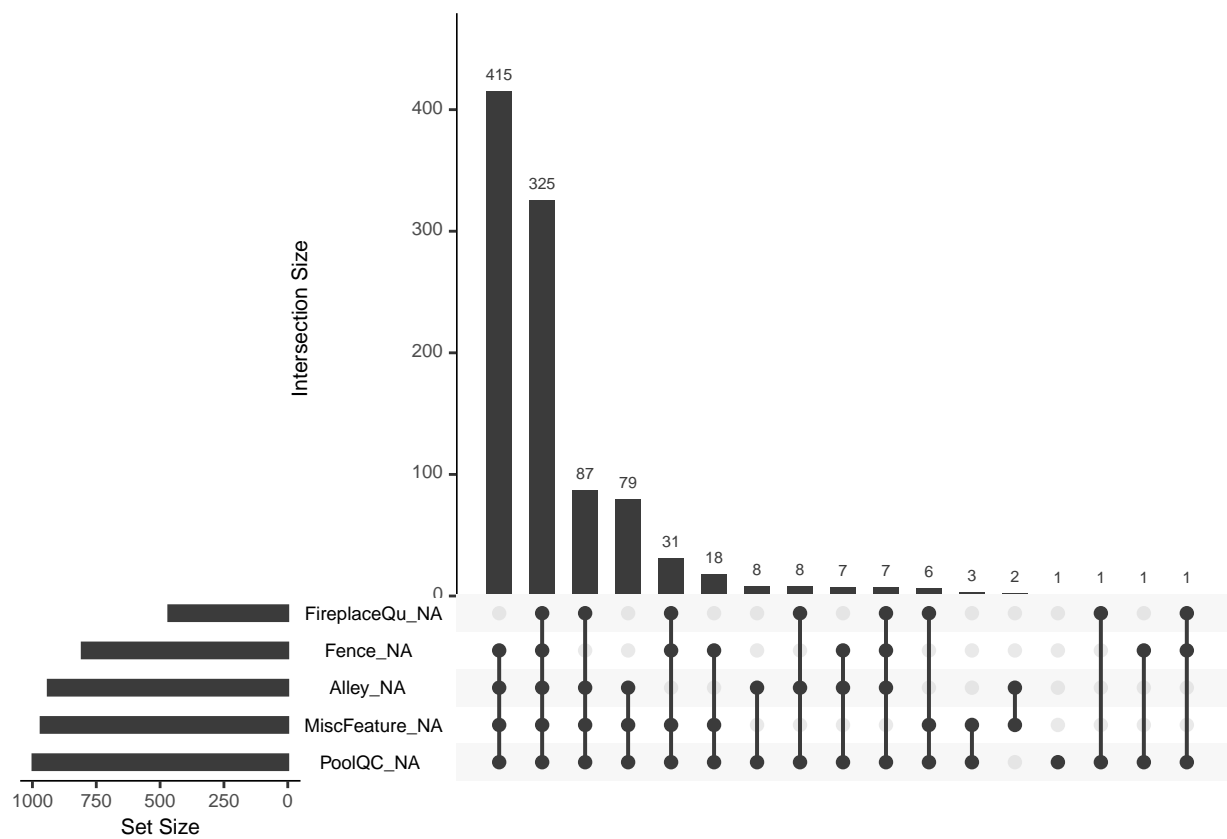
```
library(naniar)
```

```
vis_miss(house)
```

This graphic feels overwhelming.Let's look for patterns in this data.

```
gg_miss_upset(house)
```

This one is very informative, it shows some of the counts and the connections with missed data. Very cool!