Homework 2

Mohamat Eirban Ali

September 7, 2019

```
# Homework 2
#load library
library('ggplot2')
library('tidyverse')
```

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --

## v tibble  2.1.3     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.8.3
## v readr   1.1.1     v stringr 1.3.1
## v tibble  2.1.3     v forcats 0.3.0

## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

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

```
library('naniar')
```

```
## Warning: package 'naniar' was built under R version 3.5.3
```

```
library('gridExtra')
```

```
## Warning: package 'gridExtra' was built under R version 3.5.3

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library('Amelia')
```

```
## Warning: package 'Amelia' was built under R version 3.5.3

## Loading required package: Rcpp

## Warning: package 'Rcpp' was built under R version 3.5.3
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2019 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library('VIM')
```

```
## Warning: package 'VIM' was built under R version 3.5.3
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library('mice')
```

```
## Warning: package 'mice' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

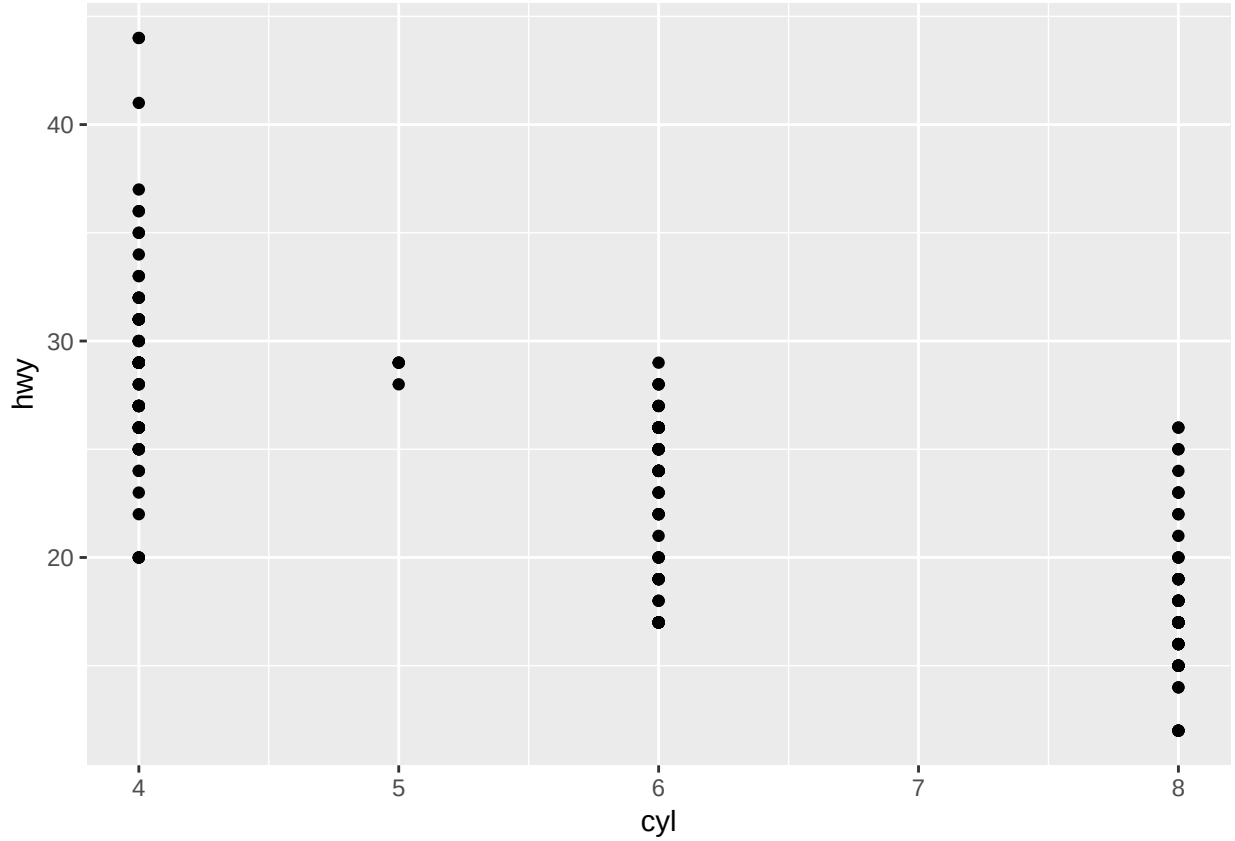
```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library('dplyr')
library('ggthemes')
```

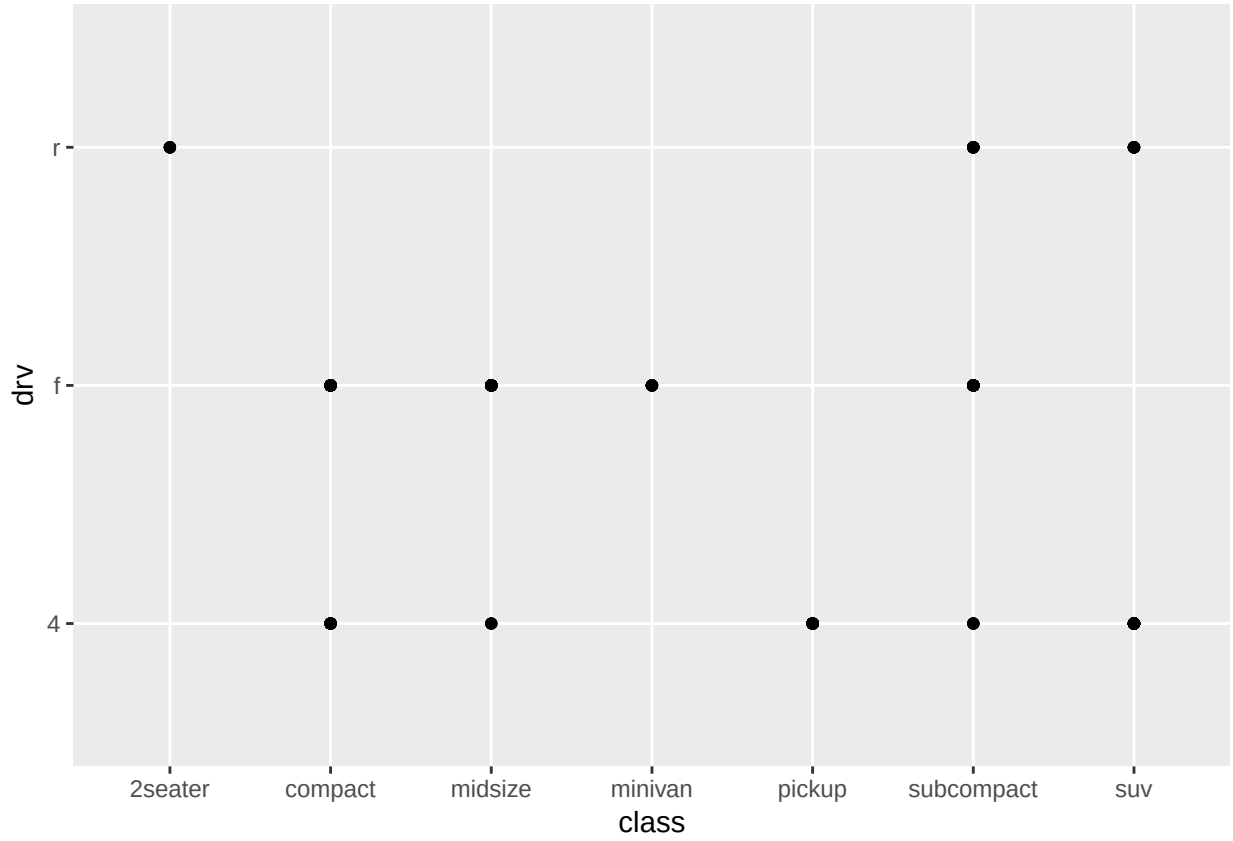
```
## Warning: package 'ggthemes' was built under R version 3.5.3
```

```
#Problem 1
#Problem 1a
# Answering questions from Chapter 3 'R for Data Science'
#3.2.5 Exercise 4
# Answer:
ggplot(mpg, aes(x=cyl, y=hwy)) +
  geom_point()
```

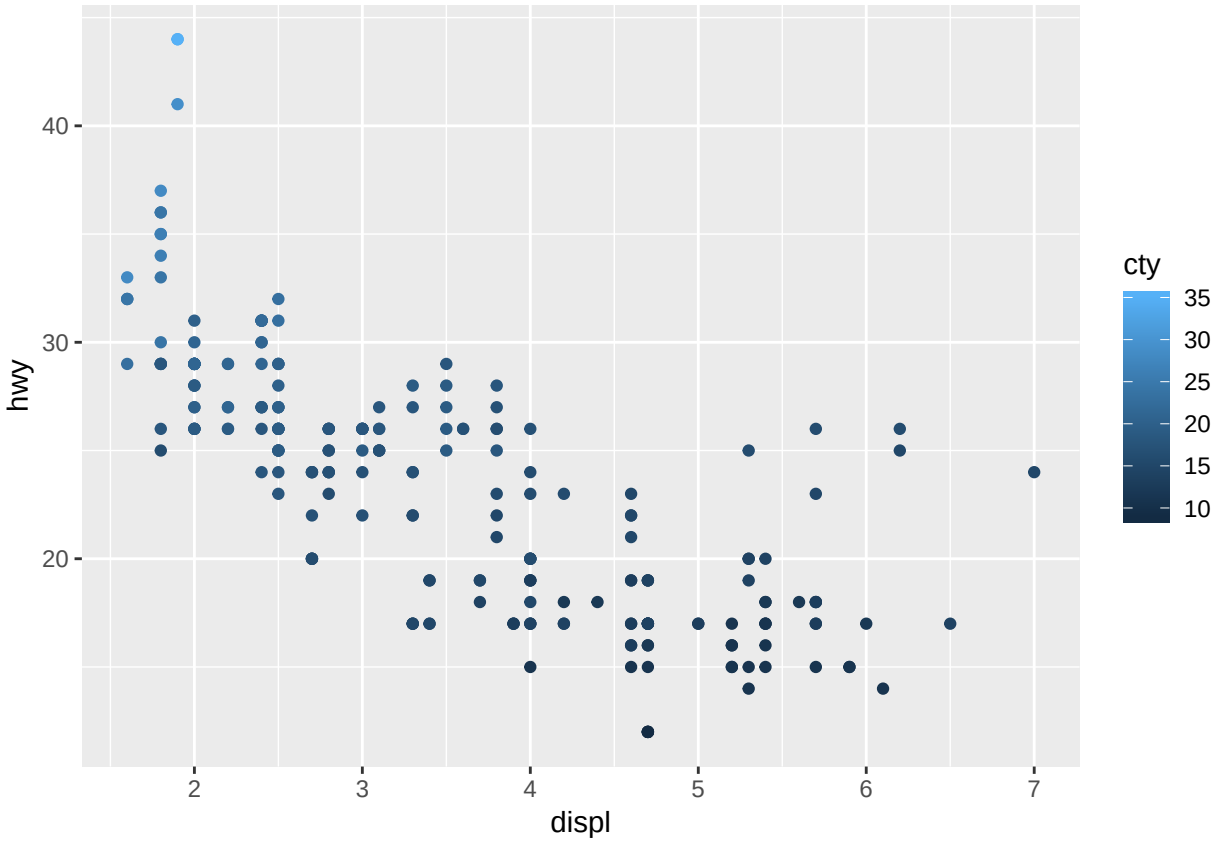



```
# Scatter plot for hwy vs cyl

#3.2.4 Exercise 5
```

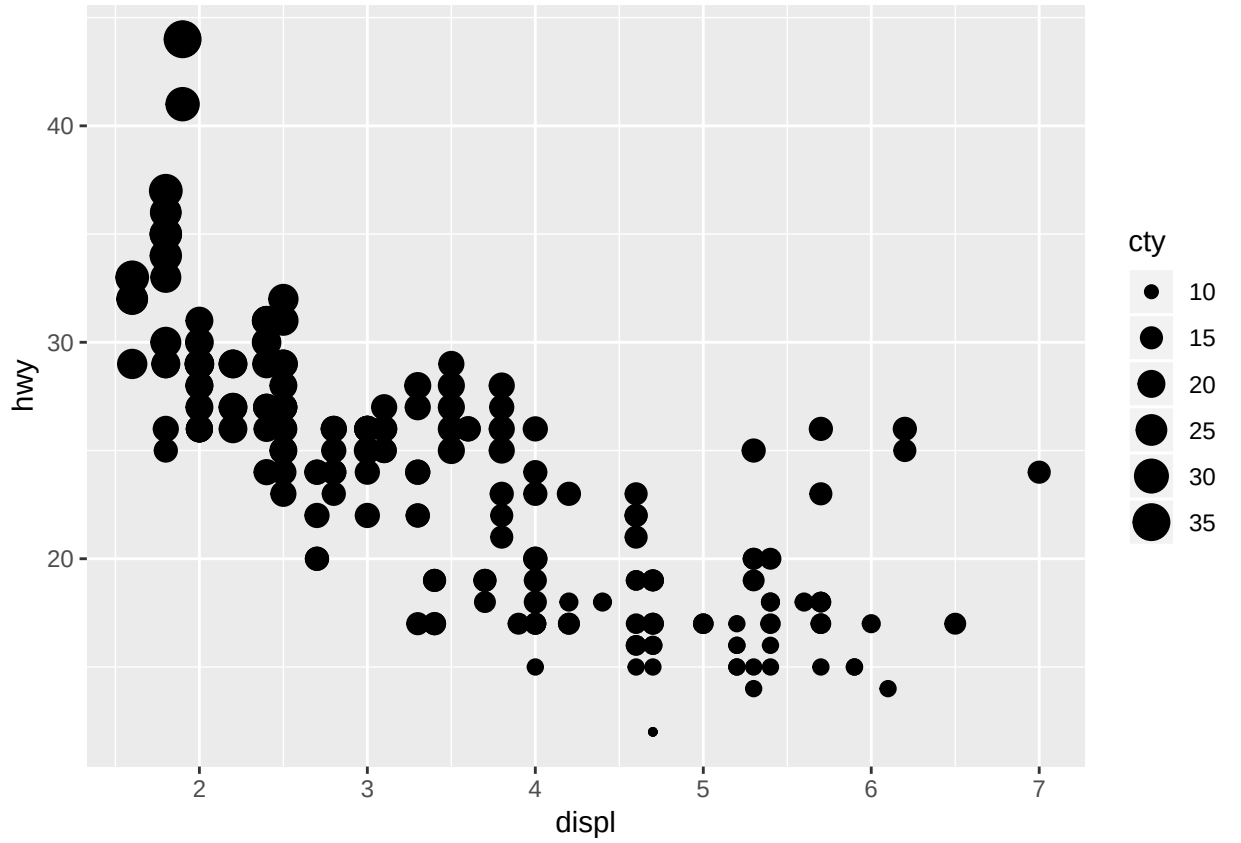
```r
# Answer:
ggplot(mpg, aes(x=class, y=drv)) +
  geom_point()
```



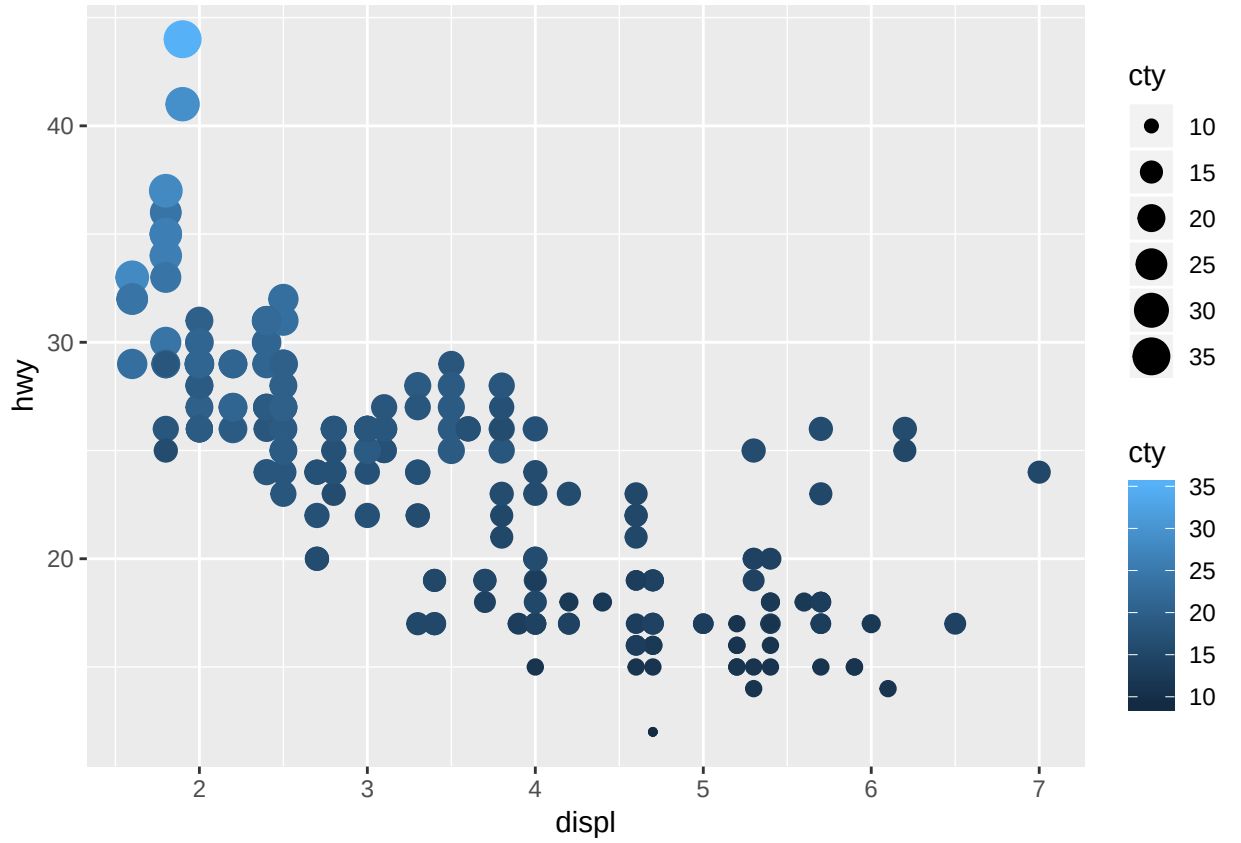

```r
# Scatter plot for drv vs class but it is kinda worthless to use this plot to
# analyze data as many of the data are overlapping because both class and drv
# variable are categorical in nature

#3.3.1 Exercise 3
# Answer:
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, colour = cty))
```




```
# ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, shape = cty)) ==Error: A continuous
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, size = cty))
```

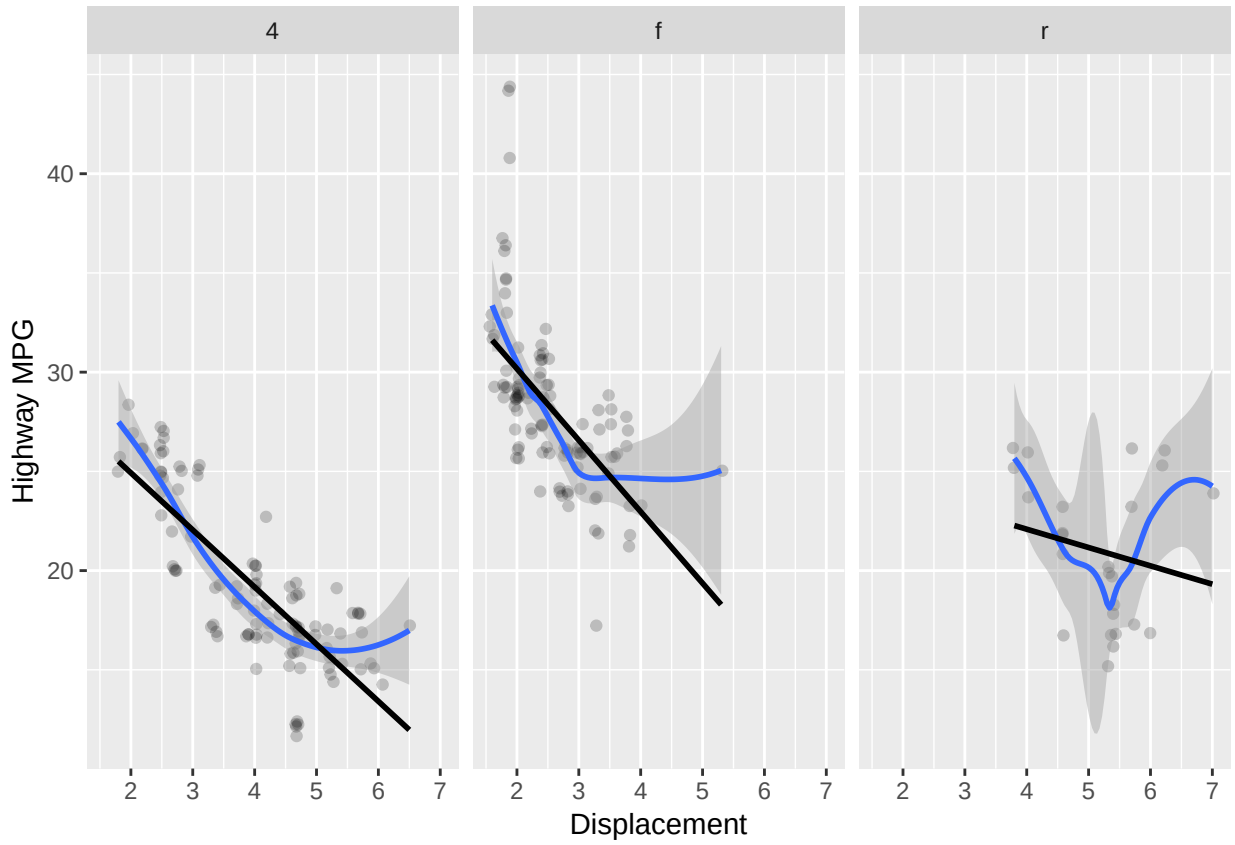



```
# For both colour and size, the continuous variable relates to it by
# saturation and area size respectively.Instead for categorical variable,
# it both colour and size will have a set outcome for each category instead of a scale.

#3.3.1 Exercise 4
# Answer:
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy, colour = cty, size = cty))
```




```
# Seems like it does work, as long as both aesthetic corresponds to each other
# to describe the variable.

#3.3.1 Exercise 5
# Answer: Stroke controls the width of the border of certain shapes that
# have border attribute.

#3.5.1 Exercise 4
# Answer:Faceting help us focus the trendline or pattern for each group
# instead of an overall distribution while the colour aesthetic gives us
# an overall pattern of the distribution. If we have a large number of groups,
# colours will not be able discretize them much as they are limited while facets
# will not do well trend comparison between the groups.

#Problem 1b
ggplot(data = mpg) +
  geom_point(alpha=0.2, position='jitter', mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ drv, nrow = 1) +
  geom_smooth(mapping = aes(x = displ, y = hwy)) +
  geom_smooth(mapping = aes(x = displ, y = hwy), method = 'lm', colour = 'black',se=F) +
  labs(x="Displacement", y="Highway MPG")
```

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

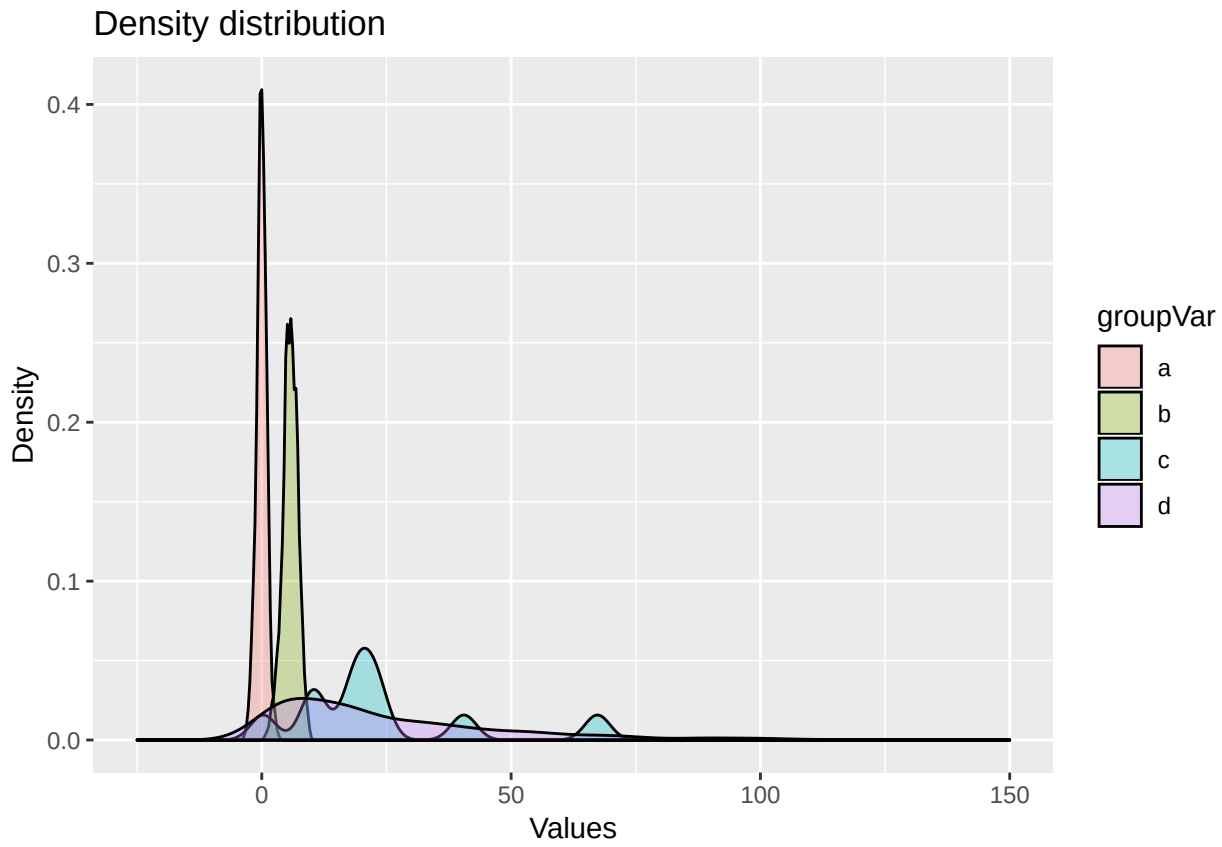



```
#Problem 2
#Problem 2a
set.seed(100)
#seed for random generator
df <- data.frame("a" = rnorm(1:500), "b" = rbinom(1:500, 9, 0.64), "c" = rexp(10,1/25), "d" = rexp(500,
df2 <- df %>% gather(groupVar, values, a, b, c, d)
head(df2)
```

```
##   groupVar      values
## 1        a -0.50219235
## 2        a  0.13153117
## 3        a -0.07891709
## 4        a  0.88678481
## 5        a  0.11697127
## 6        a  0.31863009
```

```
#Problem 2b
library(ggplot2)
ggplot(df2, aes(x=values, fill = groupVar )) +
  geom_density(alpha=.3) +
  xlim(-25,150) +
  labs(x='Values', y='Density', title = 'Density distribution')
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```

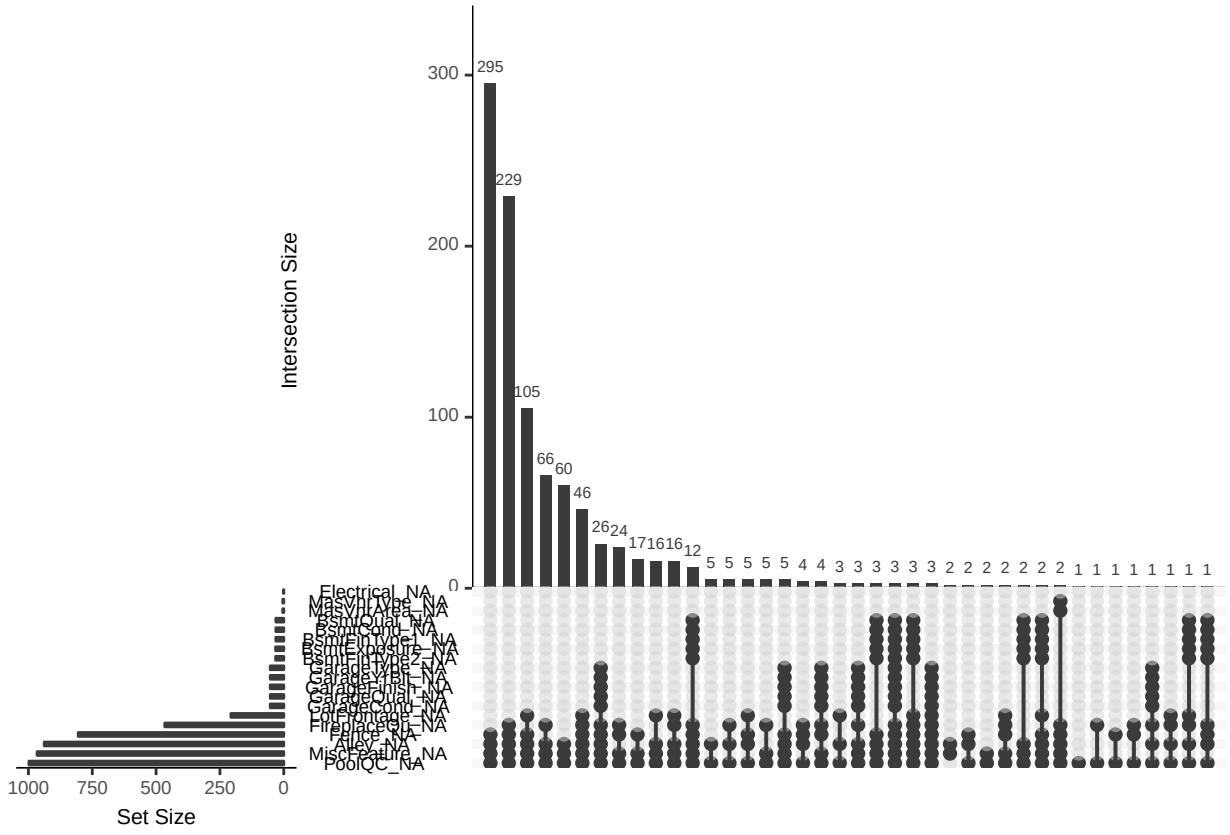



```
#Problem 3
# Load the housing data, allow for header, and remove the ID column(not relevant)
house <- read.csv("~/DSA -Homework 2/housingData.csv", header=TRUE)
rownames (house) <- house [,1]
house <- house [,-1]
# Then read through the pdf for the housingData variable explanation,
# used the summary function to the overall data

# Find the missing values for each variable to see if they need to be
# removed or kept/imputed
colSums(is.na(house))
```

```
##    MSSubClass      MSZoning  LotFrontage      LotArea        Alley
##             0             0          207            0          938
##      LotShape  LandContour    LotConfig    LandSlope Neighborhood
##             0             0            0            0            0
##    Condition1     BldgType   HouseStyle  OverallQual  OverallCond
##             0             0            0            0            0
##     YearBuilt  YearRemodAdd    RoofStyle  Exterior1st  Exterior2nd
##             0             0            0            0            0
##    MasVnrType   MasVnrArea    ExterQual    ExterCond   Foundation
##             4             4            0            0            0
##      BsmtQual     BsmtCond BsmtExposure BsmtFinType1   BsmtFinSF1
##            31            31           32           31            0
## BsmtFinType2   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF      Heating
##            32             0            0            0            0
```

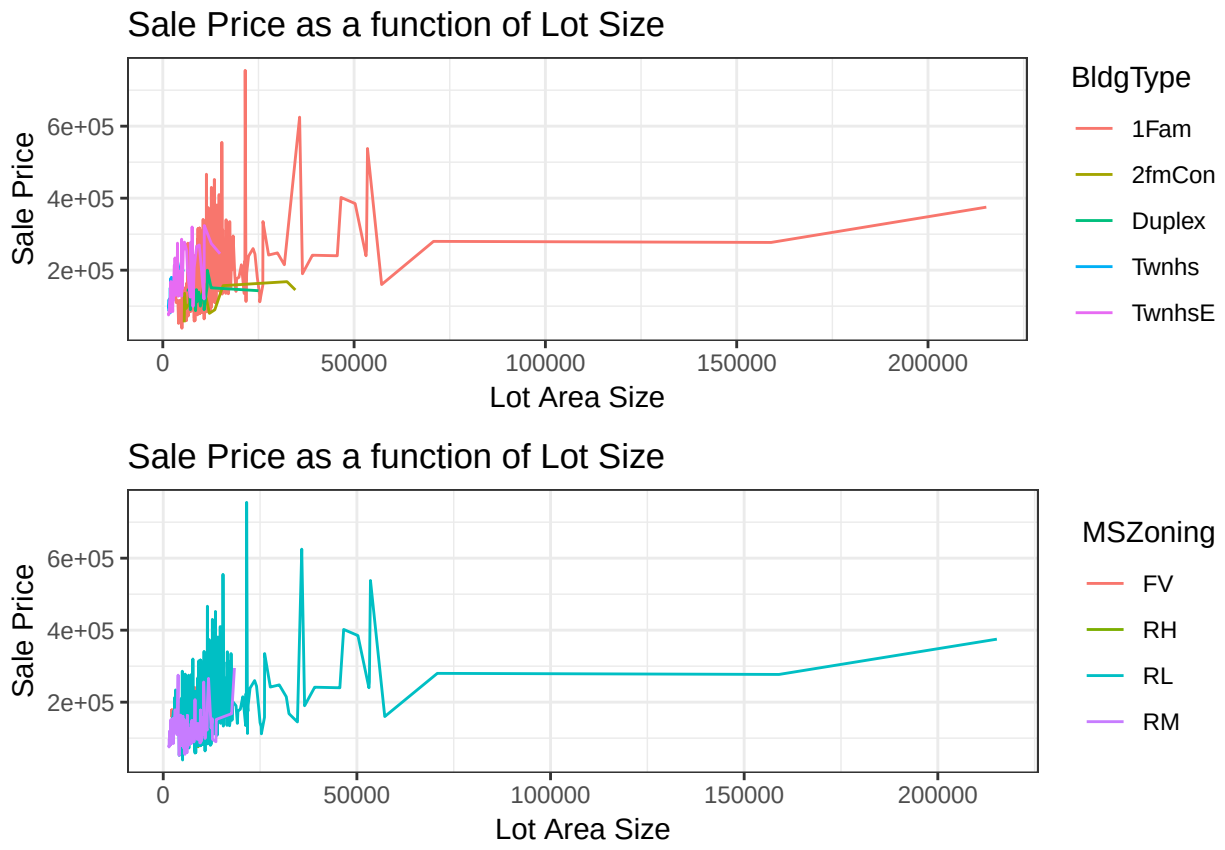
```
##    HeatingQC   CentralAir   Electrical    X1stFlrSF    X2ndFlrSF
##            0            0            1            0            0
## LowQualFinSF    GrLivArea BsmtFullBath BsmtHalfBath     FullBath
##            0            0            0            0            0
##     HalfBath BedroomAbvGr KitchenAbvGr  KitchenQual TotRmsAbvGrd
##            0            0            0            0            0
##   Functional   Fireplaces  FireplaceQu   GarageType  GarageYrBlt
##            0            0          466           53           53
## GarageFinish   GarageCars   GarageArea   GarageQual   GarageCond
##           53            0            0           53           53
##   PavedDrive   WoodDeckSF  OpenPorchSF   EncPorchSF     PoolArea
##            0            0            0            0            0
##       PoolQC        Fence  MiscFeature      MiscVal       MoSold
##          998          805          966            0            0
##       YrSold     SaleType    SalePrice
##            0            0            0
```

```
# Some of the missing variable seem to correlate to the missingness in other
# variable such as
# BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2
# GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond,
gg_miss_upset(house, nsets = n_var_miss(house))
```

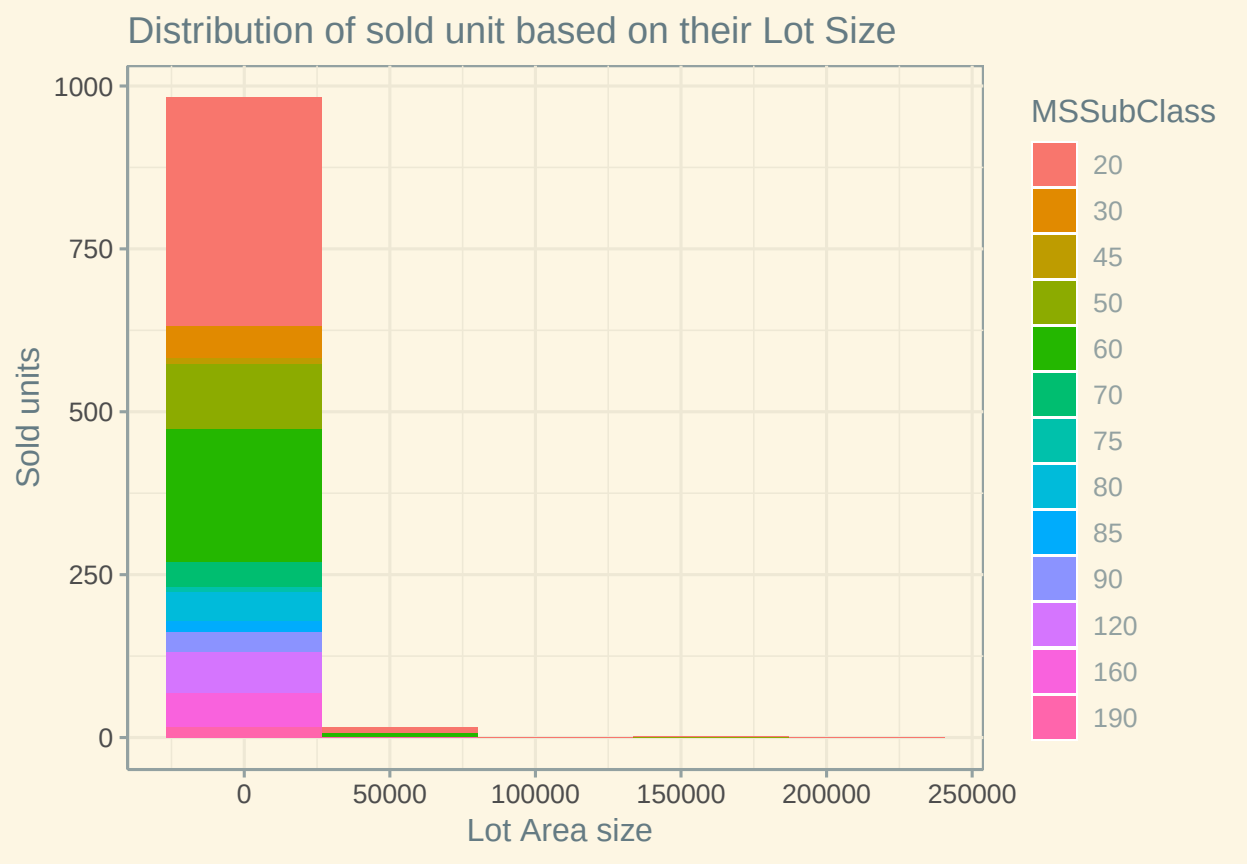



```
# We can see the link between the missing value from this graph
```

```r
house$MSSubClass <- factor(house$MSSubClass)
# We convert MSSubClass with the factor function as this is a categorical variable

# Data Visualization 1
p <- ggplot(house,aes(x=LotArea, y=SalePrice, colour = BldgType)) +
  geom_line() +
  theme_bw() +
  labs(x='Lot Area Size', y='Sale Price', title= 'Sale Price as a function of Lot Size')
q <- ggplot(house,aes(x=LotArea, y=SalePrice, colour = MSZoning)) +
  geom_line() +
  theme_bw() +
  labs(x='Lot Area Size', y='Sale Price', title= 'Sale Price as a function of Lot Size')
grid.arrange(p,q, nrow=2)
```

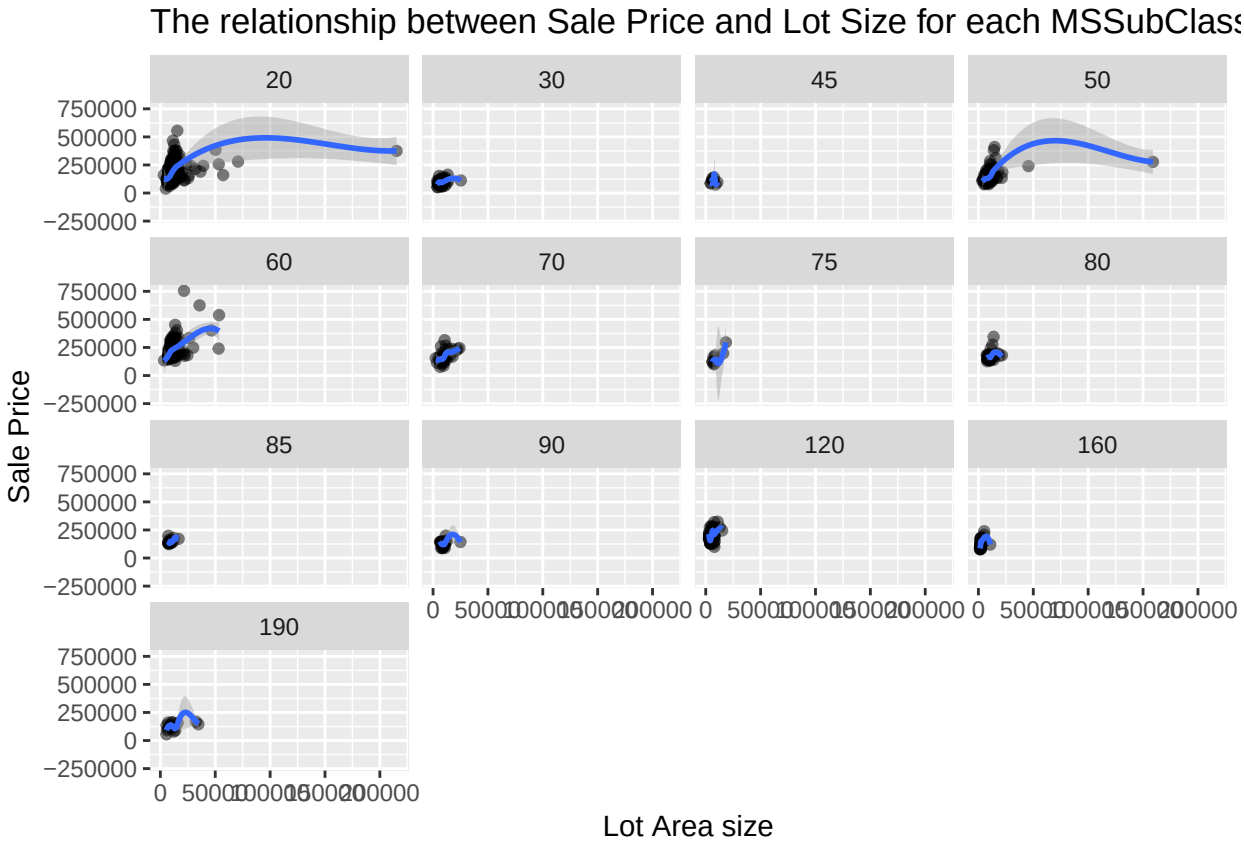



```r
# We can similarity in MSZoning and Building Type, a good hypothesis
# would be that the building type depends on the MS Zone

# Data Visualization 2
ggplot(house, aes(x=LotArea, fill=MSSubClass)) +
  geom_histogram(bins=5) +
  theme_solarized() +
  labs(x=' Lot Area size', y= 'Sold units', title = 'Distribution of sold unit based on their Lot Size')
```

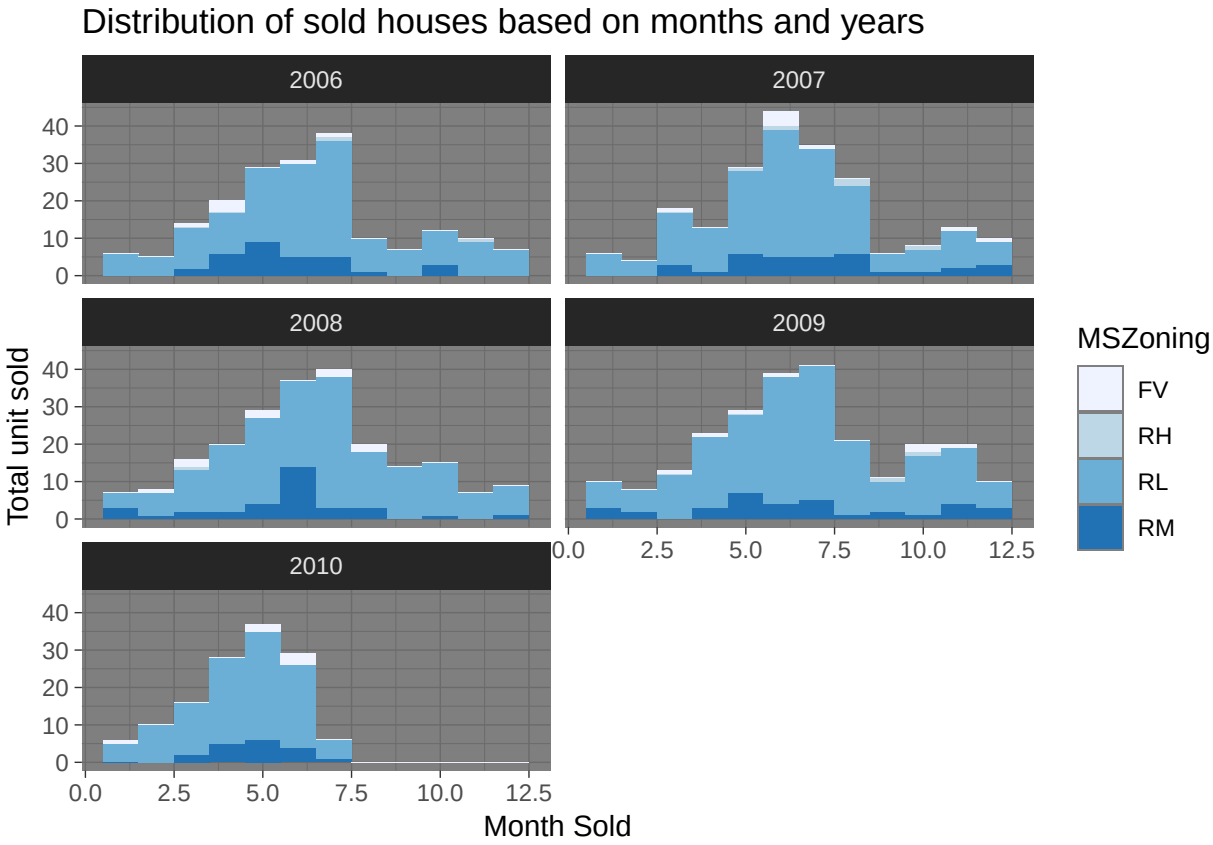



```
# To observe if there is any potential link to buying trend with the Lot Size

# Data Visualization 3
ggplot(house) +
  geom_point(alpha =0.5, position = 'jitter', mapping =aes(x=LotArea, y=SalePrice)) +
  facet_wrap(~ MSSubClass, nrow = 4) +
  geom_smooth(mapping= aes(x = LotArea, y = SalePrice)) +
  labs(x= 'Lot Area size', y='Sale Price', title='The relationship between Sale Price and Lot Size for
```

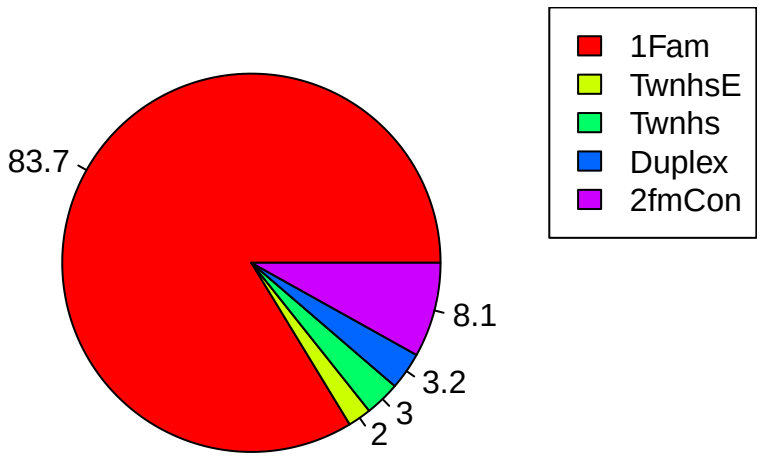
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



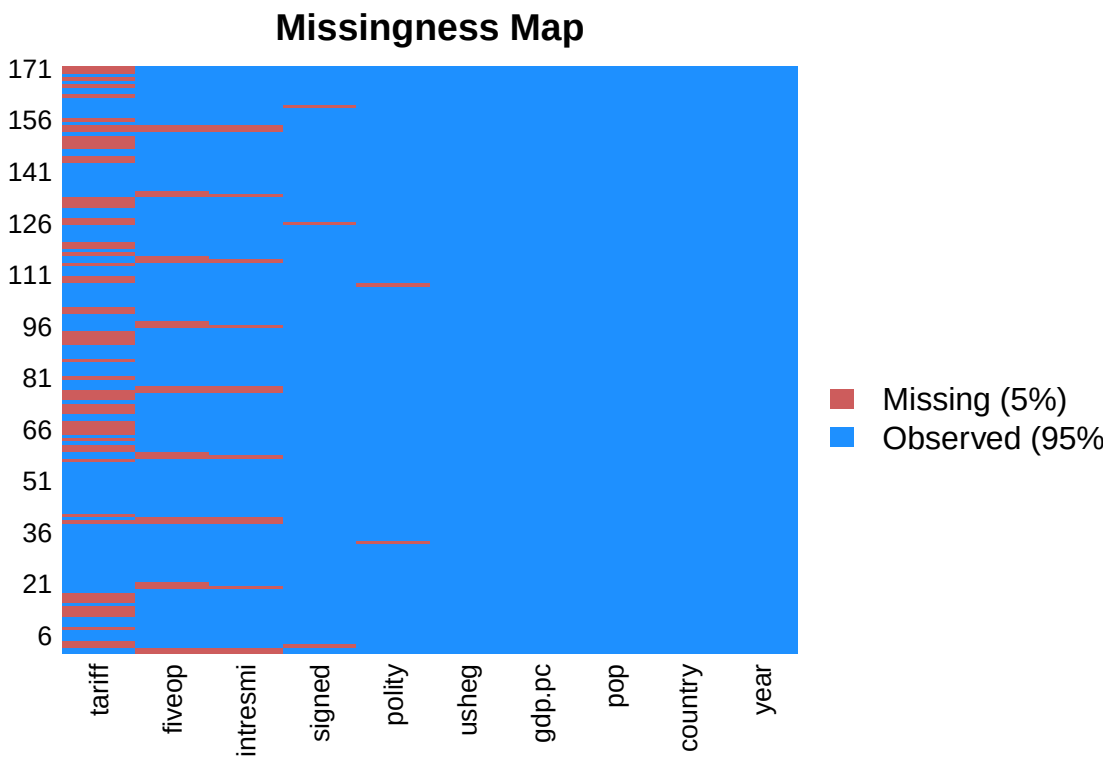

```
# Relationship between Lot Size and Selling Price

# Data Visualization 4
ggplot(house, aes(x=MoSold, fill=MSZoning)) +
  geom_histogram(bins=12) +
  facet_wrap(~ YrSold, nrow = 3) +
  labs(x = 'Month Sold', y = 'Total unit sold',title = 'Distribution of sold houses based on months and
  scale_fill_brewer(direction = -2) +
  theme_dark()
```




```
# Distribution of the sold units according to months and years

# Data Visualization 5
z <- data.frame(table(house$BldgType))
pie(z$Freq, labels = round(100*z$Freq/sum(z$Freq), 1), main = "Building Type sold",col = rainbow(length
legend("topright", c("1Fam","TwnhsE","Twnhs","Duplex","2fmCon"), cex = 1,
       fill = rainbow(length(z$Freq)))
```



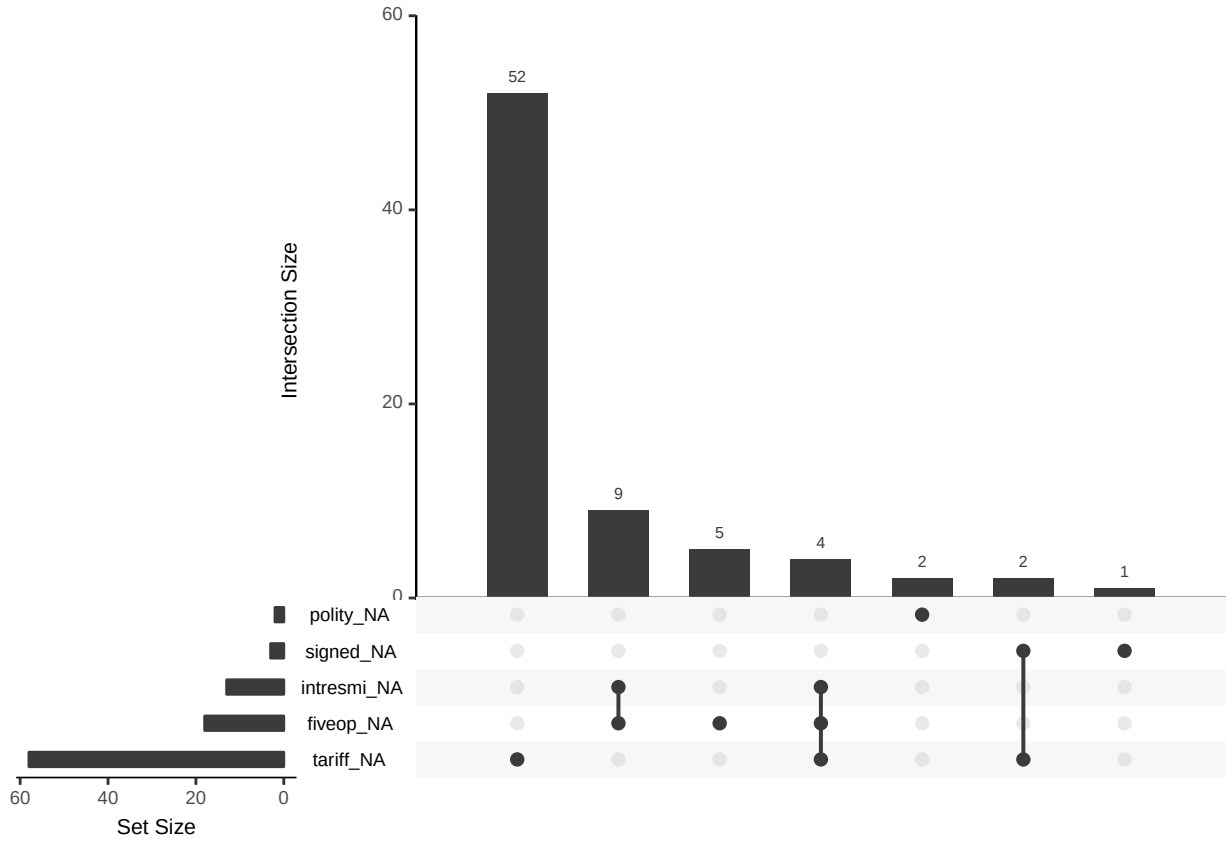

```
# Pie chart showing the distribution of sold houses based on their Building Type

#Problem 4
#Problem 4a
data("freetrade", package="Amelia") # load the data using data command
trade <- freetrade

missmap(trade)
```




```
# Missing data distribution in the data frame
#trade[!complete.cases(trade),] ...uncomment if you want to see all missing data rows

colSums(is.na(trade))
```

```
##     year  country   tariff   polity      pop   gdp.pc intresmi   signed
##        0        0       58        2        0        0       13        3
##   fiveop    usheg
##       18        0
```

```
# Summarize the missing data count for each variable

gg_miss_upset(trade, nsets = n_var_miss(trade))
```




```
# Gives a plot of missing variables and if the missing variables
# are related by observations

matrixplot(trade)
```

```
## Warning in data.matrix(x): NAs introduced by coercion
```

```
# Missing data visualization

j <- as.data.frame(abs(is.na(trade)))
o <- j[,sapply(j, sd) > 0]
cor(o)
```

```
##               tariff      polity    intresmi      signed      fiveop
## tariff     1.00000000 -0.07793749 -0.01907852  0.09243593 -0.08473587
## polity    -0.07793749  1.00000000 -0.03120432 -0.01453710 -0.03731317
## intresmi  -0.01907852 -0.03120432  1.00000000 -0.03833091  0.83628170
## signed     0.09243593 -0.01453710 -0.03833091  1.00000000 -0.04583492
## fiveop    -0.08473587 -0.03731317  0.83628170 -0.04583492  1.00000000
```

```
# Correlation matrix between all missing data

cor(trade$tariff, o, use = "pairwise.complete.obs")
```

```
## Warning in cor(trade$tariff, o, use = "pairwise.complete.obs"): the
## standard deviation is zero
```

```
##      tariff      polity  intresmi      signed     fiveop
## [1,]     NA -0.03973145 -0.177108 -0.02975753 -0.2480984
```

```
cor(trade$polity, o, use = "pairwise.complete.obs")
```

```
## Warning in cor(trade$polity, o, use = "pairwise.complete.obs"): the
## standard deviation is zero
```

```
##          tariff polity  intresmi      signed    fiveop
## [1,] -0.1318329     NA 0.1454525 -0.01390191 0.1584924
```

```
cor(trade$pop, o, use = "pairwise.complete.obs")
```

```
##           tariff      polity    intresmi     signed     fiveop
## [1,] -0.05703336 -0.04336596 -0.01135788 -0.0352636 0.03343194
```

```
cor(trade$gdp.pc, o, use = "pairwise.complete.obs")
```

```
##           tariff    polity   intresmi      signed     fiveop
## [1,] -0.08485325 0.0762219 0.06430122 -0.02368099 0.08859163
```

```
cor(trade$intresmi, o, use = "pairwise.complete.obs")
```

```
## Warning in cor(trade$intresmi, o, use = "pairwise.complete.obs"): the
## standard deviation is zero
```

```
##         tariff      polity intresmi      signed    fiveop
## [1,] 0.1065122 -0.07004969       NA 0.002939237 0.1834332
```

```
cor(trade$fiveop, o, use = "pairwise.complete.obs")
```

```
## Warning in cor(trade$fiveop, o, use = "pairwise.complete.obs"): the
## standard deviation is zero
```

```
##          tariff      polity intresmi     signed fiveop
## [1,] -0.1759771 -0.07895048       NA 0.09070606     NA
```

```
cor(trade$usheg, o, use = "pairwise.complete.obs")
```

```
##            tariff     polity  intresmi      signed    fiveop
## [1,] -0.007128887 0.03322148 0.5316772 -0.02035588 0.5805495
```

```r
# Correlation of missing values and observed variables
# (but only if the observed variables are numeric)

#Problem 4 b
Miss <- rep ("0", nrow(trade ))
Miss [is.na(trade$tariff) == TRUE] <- "1"
Miss <- as.factor (Miss)
trade <- data.frame(trade ,Miss)
# Created a logic variable for missing data in tariff variable

tradetest <- select(trade,c(country, Miss))
# a new dataframe consisting of only tariff and country
chisq.test(table(tradetest$Miss,tradetest$country))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(tradetest$Miss, tradetest$country)
## X-squared = 23.064, df = 8, p-value = 0.003283
```

```r
# The p-value is way small (if we were to take a 0.01 significance value),
# then the null hypothesis that the variable tariff and country are
# independent is rejected

tradetest1 <- select(filter(tradetest,country!='Nepal'), c(country,Miss))
#new dataframe similar to above except rows with 'Nepal' is removed
chisq.test(table(tradetest1$Miss,tradetest1$country))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(tradetest1$Miss, tradetest1$country)
## X-squared = 15.836, df = 7, p-value = 0.02666
```

```r
# obvious change in the p-value...in fact higher than 0.01 significance level;
# it is probable that the variable
# tariff and country is not related in this case

tradetest2 <- select(filter(tradetest,country!='Philipines'), c(country,Miss))
# new data frame with 'Philipines removed instead
chisq.test(table(tradetest2$Miss,tradetest2$country))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(tradetest2$Miss, tradetest2$country)
## X-squared = 23.064, df = 8, p-value = 0.003283
```

```r
# slight change in the p-value as compared to the original chi-square test...
# well smaller than the significane level of 0.01
# In conclusion, the statistical test shows that the variable tariff and country
```

```
# are related because mainly most of the missingvalues are populated in
# rows of country where it's Nepal
# Safe to assume the tariff is not able to be obtained for Nepal due to some challenges
```