Business Analytics

Session 7

Ing. Juan José Franklin Uraga, PhD.

MODELO EDUCATIVO TEC 21

Tecnológico de Monterrey

Exploring and visualizing categorical data

R Studio®

Dataset

Tecnológico de Monterrey

```
library(visdat)
comics <- read.csv("/Users/jjfu/Downloads/comics.csv")
glimpse(comics)
```

```
## Rows: 23,272
## Columns: 11
## $ name         <chr> "Spider-Man (Peter Parker)", "Captain America (Steven Rog…
## $ id           <chr> "Secret", "Public", "Public", "Public", "No Dual", "Publi…
## $ align        <chr> "Good", "Good", "Neutral", "Good", "Good", "Good", "Good"…
## $ eye          <chr> "Hazel Eyes", "Blue Eyes", "Blue Eyes", "Blue Eyes", "Blu…
## $ hair         <chr> "Brown Hair", "White Hair", "Black Hair", "Black Hair", "…
## $ gender       <chr> "Male", "Male", "Male", "Male", "Male", "Male", "Male", "…
## $ gsm          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ alive        <chr> "Living Characters", "Living Characters", "Living Charact…
## $ appearances  <int> 4043, 3360, 3061, 2961, 2258, 2255, 2072, 2017, 1955, 193…
## $ first_appear <chr> "Aug-62", "mar-41", "oct-74", "mar-63", "nov-50", "nov-61…
## $ publisher    <chr> "marvel", "marvel", "marvel", "marvel", "marvel", "marvel…
```

RStudio®

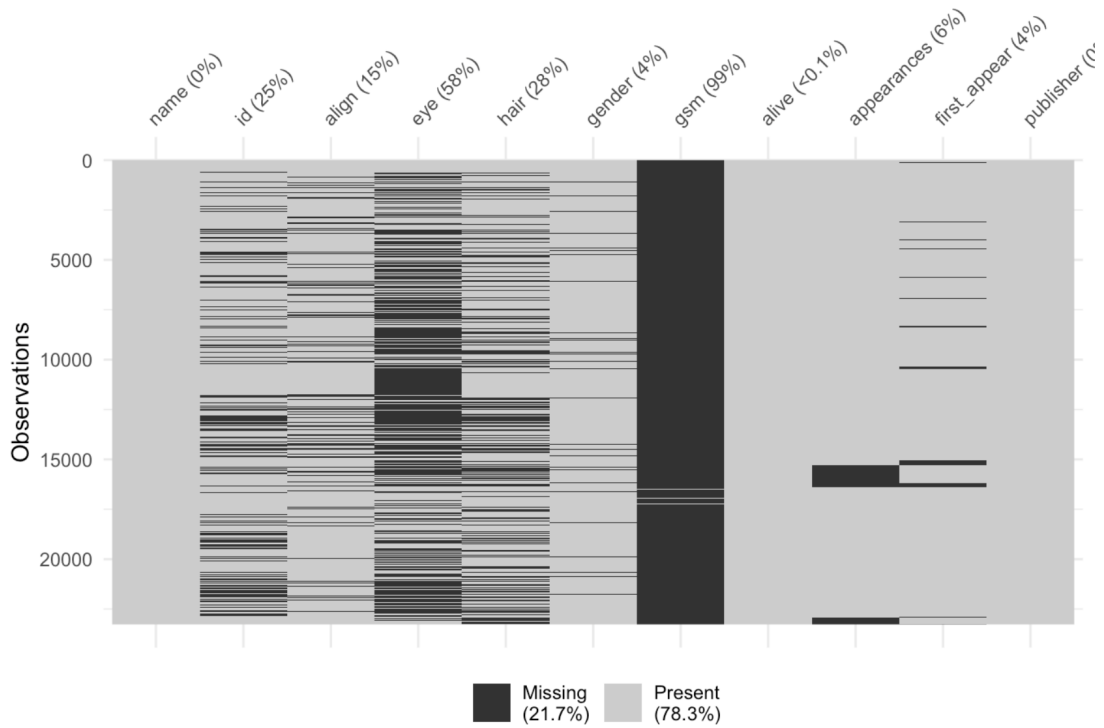
Exploring and visualizing categorical data

Tecnológico de Monterrey

Evidence Session # 07

1. Create a new dataset “comics2” that keeps complete observations (drop NA’s) for the variables “id” and “align”.

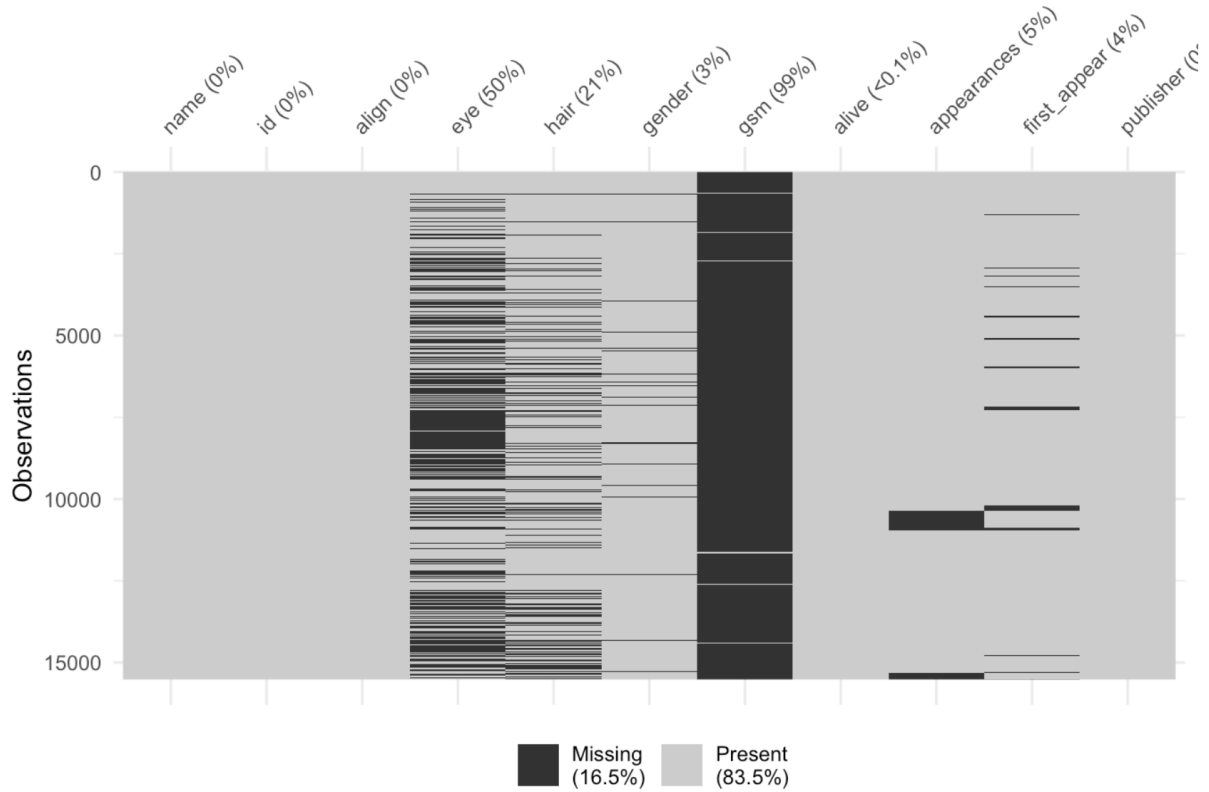
```
vis_miss(comics)
```



RStudio®

Exploring and visualizing categorical data

```
comics2 <- comics[complete.cases(comics$id,comics$align),]
vis_miss(comics2)
```



Exploring and visualizing categorical data

Tecnológico de Monterrey

2. Get the levels for “id”, “align” and then build a contingency table for both variables.

```
levels(as.factor(comics2$id))
```

```
## [1] "No Dual" "Public"  "Secret"  "Unknown"
```

```
levels(as.factor(comics2$align))
```

```
## [1] "Bad"                 "Good"                "Neutral"
## [4] "Reformed Criminals"
```

```
table(comics2$id)
```

```
##
## No Dual  Public  Secret Unknown
##    1511    6068    7928       9
```

```
table(comics2$align)
```

```
##
##                Bad               Good            Neutral Reformed Criminals
##               7146               6052               2316                  2
```

RStudio®

Exploring and visualizing categorical data

Tecnológico de Monterrey

```
cnt_tab <- table(comics2$id,comics2$align)
cnt_tab
```

```
##
##             Bad Good Neutral Reformed Criminals
##   No Dual   474  647     390                   0
##   Public   2172 2930     965                   1
##   Secret   4493 2475     959                   1
##   Unknown     7    0       2                   0
```

RStudio

Exploring and visualizing categorical data

Tecnológico de Monterrey

```
cnt_tab <- table(comics2$id,comics2$align)
cnt_tab
```

```
##
##             Bad Good Neutral Reformed Criminals
##   No Dual   474  647     390                   0
##   Public   2172 2930     965                   1
##   Secret   4493 2475     959                   1
##   Unknown     7    0       2                   0
```

RStudio

Exploring and visualizing categorical data

Tecnológico de Monterrey

3. The contingency table from the last exercise revealed that there are some levels that have very low counts. To simplify the analysis, it often helps to drop such levels. Overwrite comics2 and filter values that do not have "Reformed Criminals". Get the sum for the rows and columns.

```
comics2 <- comics2 %>%
  filter(align!= "Reformed Criminals")

cnt_tab <- addmargins(table(comics2$id,comics2$align))
cnt_tab
```

```
##
##              Bad  Good Neutral   Sum
##   No Dual    474   647     390  1511
##   Public    2172  2930     965  6067
##   Secret    4493  2475     959  7927
##   Unknown      7     0       2     9
##   Sum       7146  6052    2316 15514
```

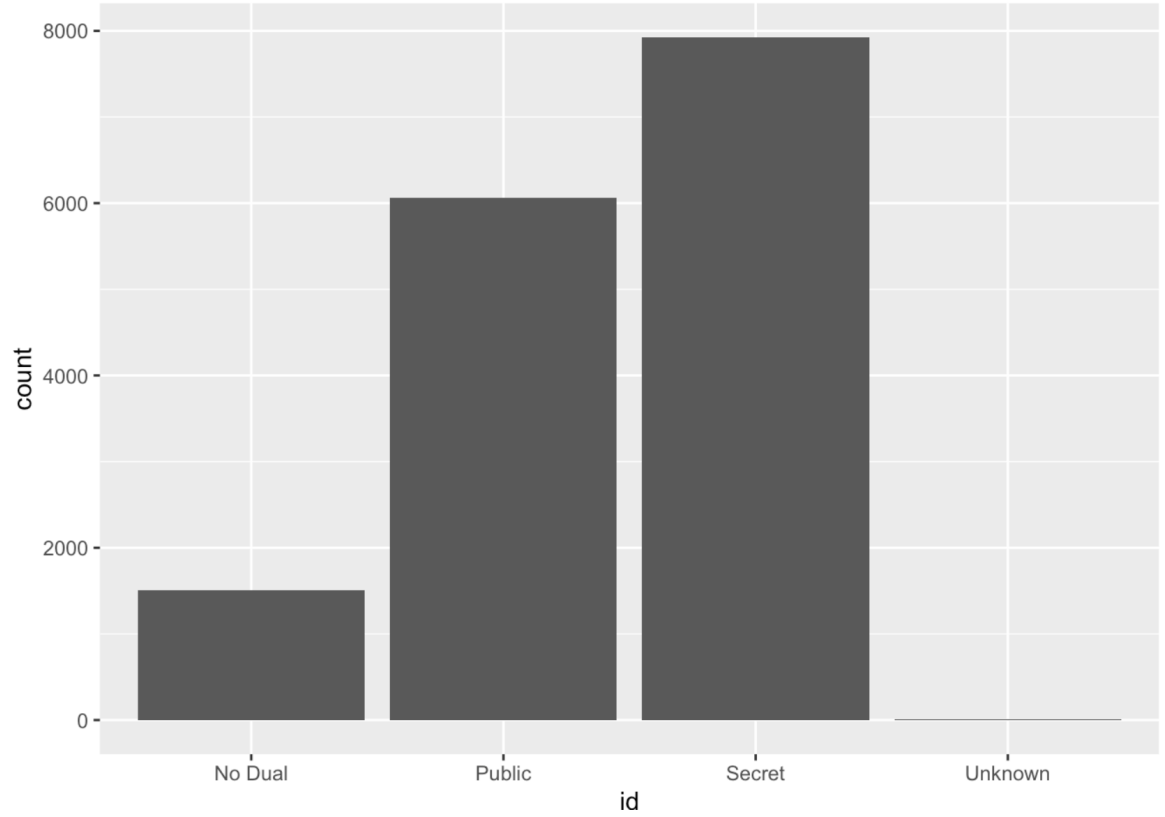
RStudio®

Exploring and visualizing categorical data

Tecnológico de Monterrey

4. Create a bar chart for the variable id.

```
comics2 %>%
  ggplot(aes(x=id))+geom_bar()
```



RStudio®

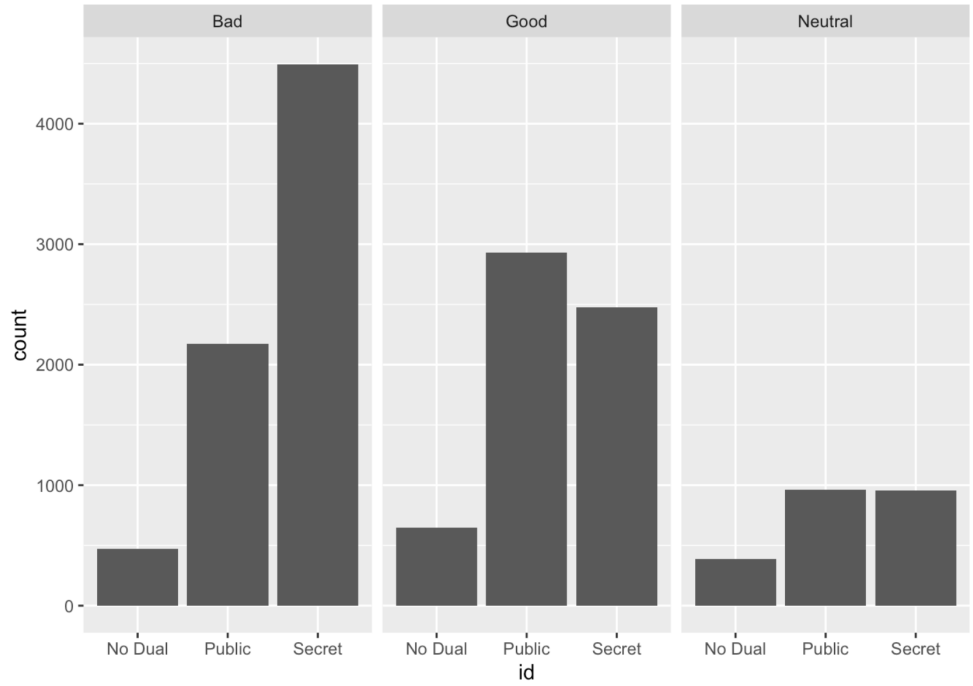
Exploring and visualizing categorical data

Tecnológico de Monterrey

5. We still have a level (“unkwown”) with a low count, remove it and now create a bar chart for the variable id and group by (facet) align.

```
comics2 <- comics2 %>%
  filter(id!="Unknown")

comics2 %>%
  ggplot(aes(x=id))+geom_bar()+facet_wrap(~align)
```



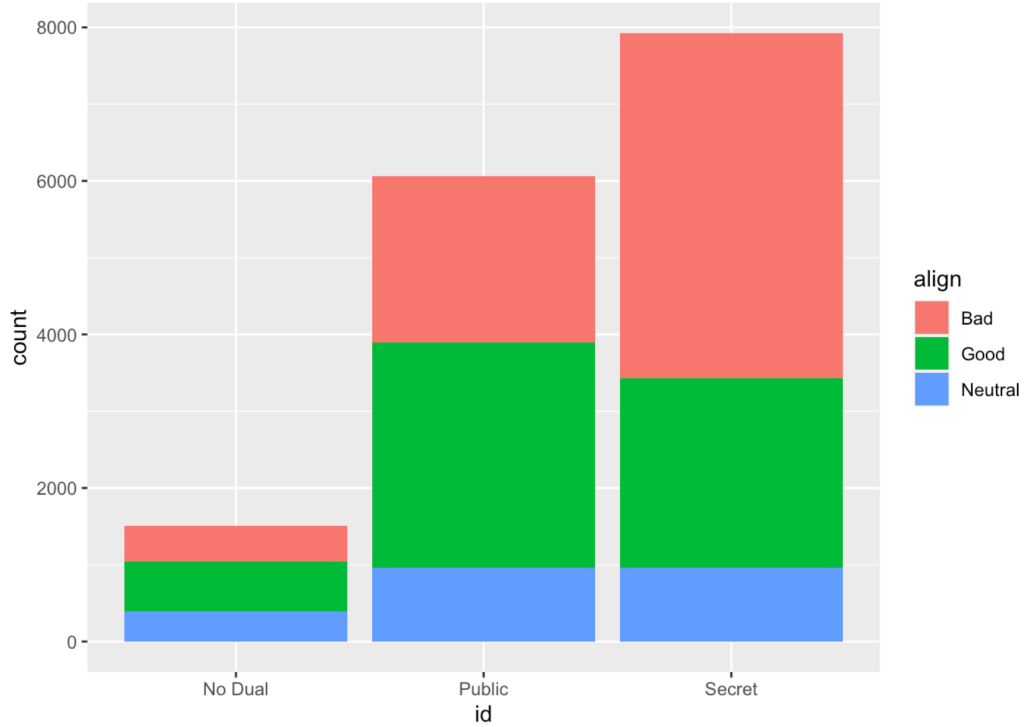
RStudio®

Exploring and visualizing categorical data

Tecnológico de Monterrey

6. While a contingency table represents the counts numerically, it's often more useful to represent them graphically. Build a side-by-side bar plot and a stacked bar plot for ID and Align variables.

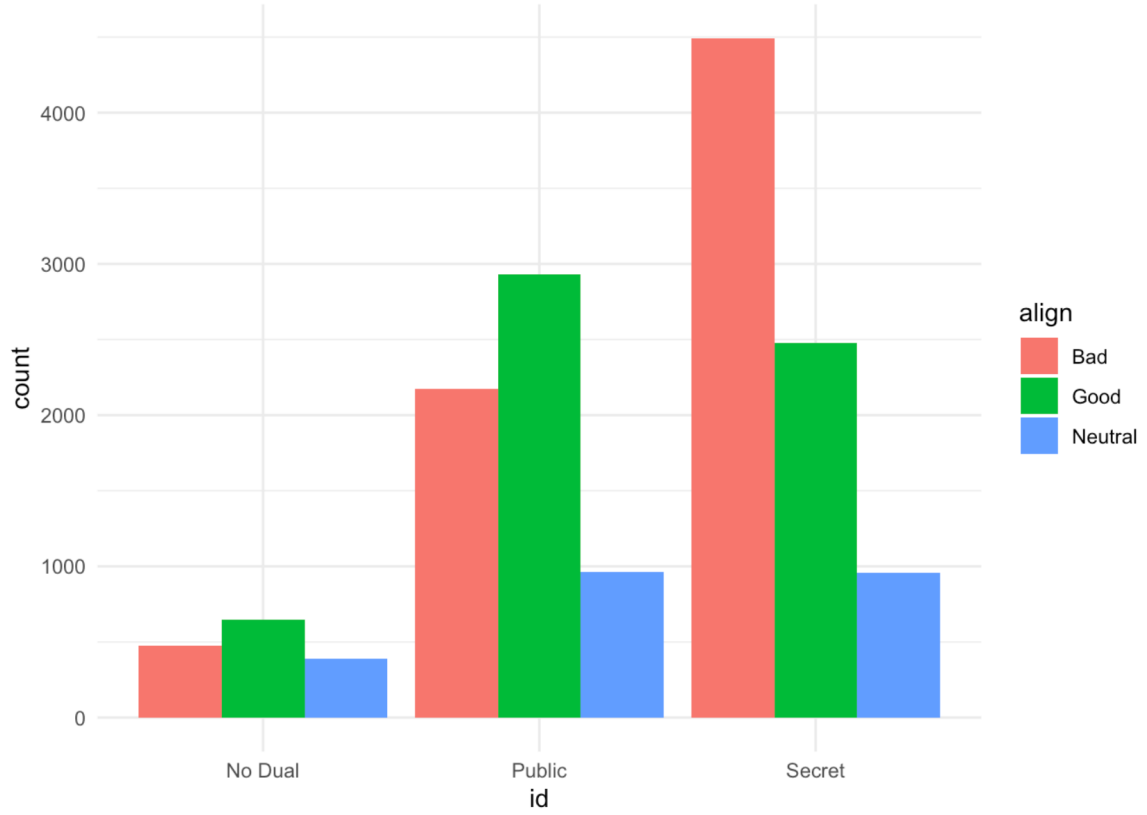
```
comics2 %>%
  ggplot(aes(x=id,fill=align))+geom_bar(position = "stack")
```



RStudio®

Exploring and visualizing categorical data

```
comics2 %>%
  ggplot(aes(x=id,fill=align))+geom_bar(position = "dodge")+theme_minimal()
```



Tecnológico de Monterrey

Exploring and visualizing quantitative data

RStudio®

Exploring and visualizing quantitative data

Tecnológico de Monterrey

2.1 For this section, we will use the cars dataset. Import it, then get a glimpse for it and identify the number of observations and variables.

```
cars <- read.csv("/Users/jjfranklin/Downloads/cars.csv")
glimpse(cars)
```

```
## Rows: 428
## Columns: 19
## $ name         <chr> "Chevrolet Aveo 4dr", "Chevrolet Aveo LS 4dr hatch", "Chev…
## $ sports_car   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
## $ suv          <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
## $ wagon        <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
## $ minivan      <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
## $ pickup       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
## $ all_wheel    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
## $ rear_wheel   <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FA…
## $ msrp         <int> 11690, 12585, 14610, 14810, 16385, 13670, 15040, 13270, 13…
## $ dealer_cost  <int> 10965, 11802, 13697, 13884, 15357, 12849, 14086, 12482, 12…
## $ eng_size     <dbl> 1.6, 1.6, 2.2, 2.2, 2.2, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 1.7…
## $ ncyl         <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4…
## $ horsepwr     <int> 103, 103, 140, 140, 140, 132, 132, 130, 110, 130, 130, 115…
## $ city_mpg     <int> 28, 28, 26, 26, 26, 29, 29, 26, 27, 26, 26, 32, 36, 32, 29…
## $ hwy_mpg      <int> 34, 34, 37, 37, 37, 36, 36, 33, 36, 33, 33, 38, 44, 38, 33…
## $ weight       <int> 2370, 2348, 2617, 2676, 2617, 2581, 2626, 2612, 2606, 2606…
## $ wheel_base   <int> 98, 98, 104, 104, 104, 105, 105, 103, 103, 103, 103, 103, …
## $ length       <int> 167, 153, 183, 183, 183, 174, 174, 168, 168, 168, 168, 175…
## $ width        <int> 66, 66, 69, 68, 69, 67, 67, 67, 67, 67, 67, 67, 67, 68, 66…
```

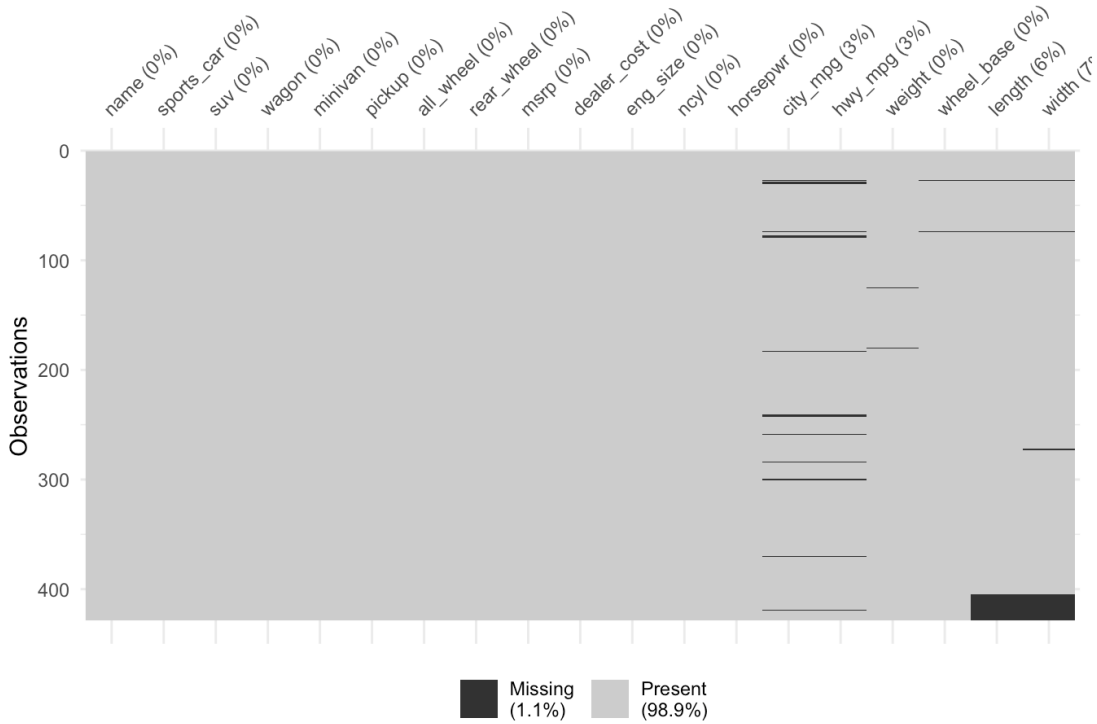
RStudio®

Exploring and visualizing quantitative data

Tecnológico de Monterrey

2.2 Let's start with some basic graphs, build a dotplot, horizontal-boxplot, histogram and density plot for the variable weight. Before making the plots, present a graphical tool to identify NA's, ggplot is robust to them, you don't need to exclude them. Keep the outliers for this time.

```
vis_miss(cars)
```



RStudio®

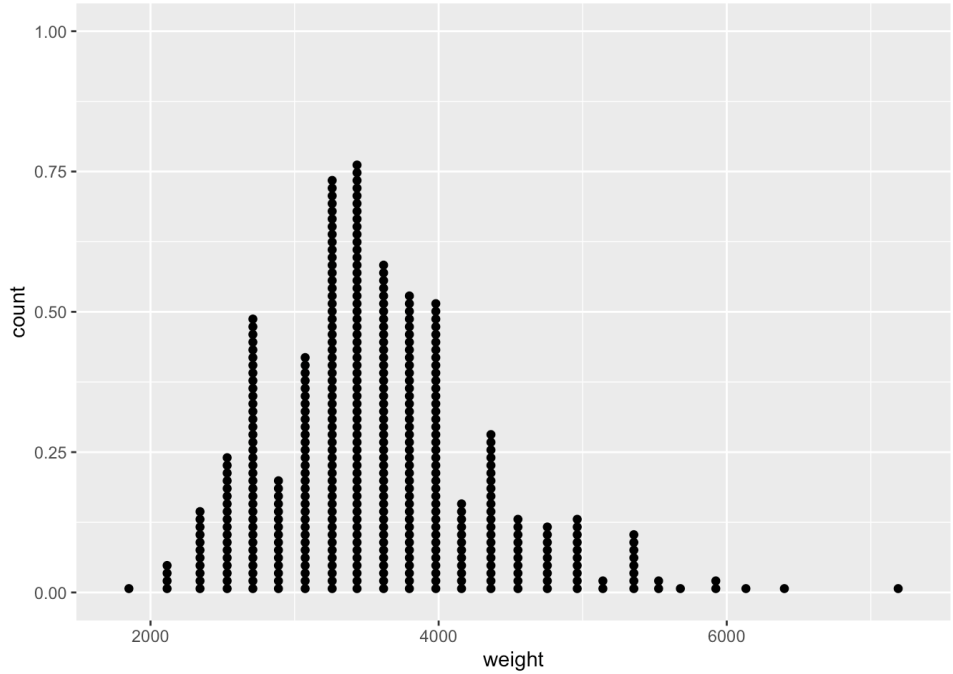
Exploring and visualizing quantitative data

Tecnológico de Monterrey

```
cars %>%
  ggplot(aes(x=weight))+geom_dotplot(dotsize = 0.3)
```

```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (`stat_bindot()`).
```

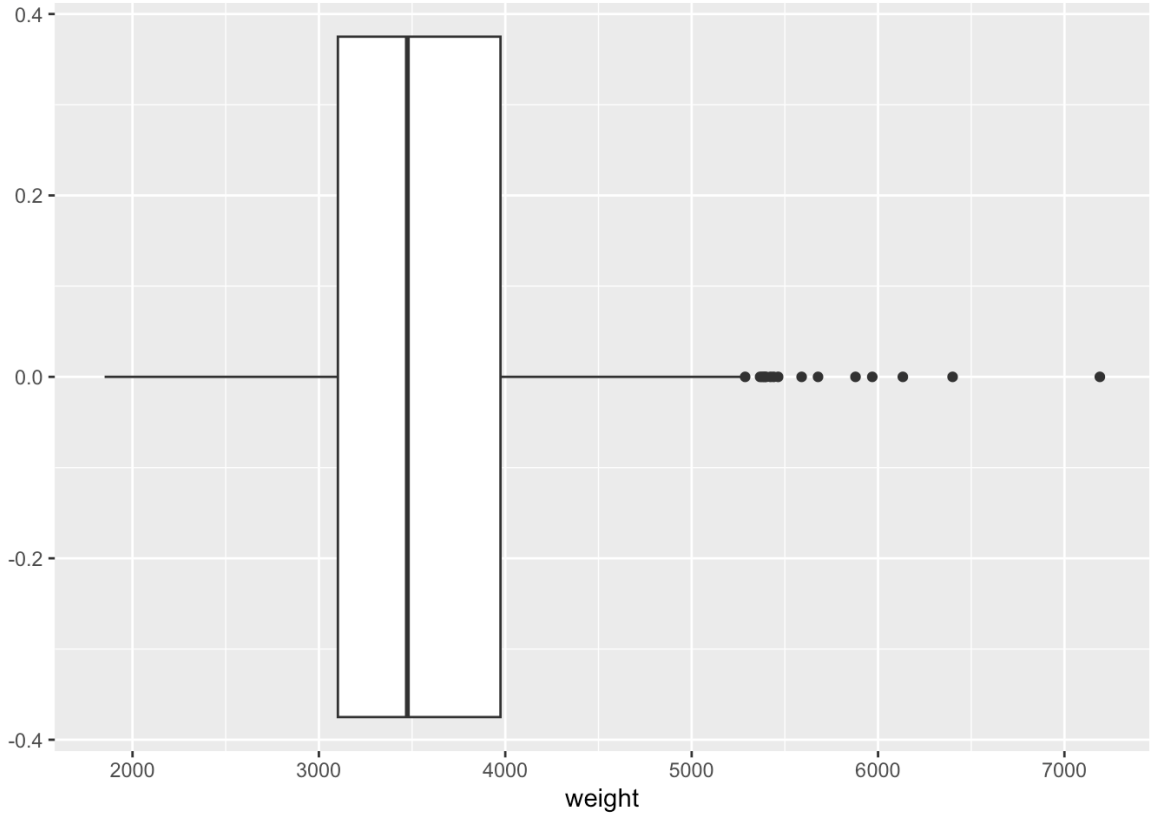


RStudio®

Exploring and visualizing quantitative data

```
cars %>%
  ggplot(aes(x=weight))+geom_boxplot()
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```

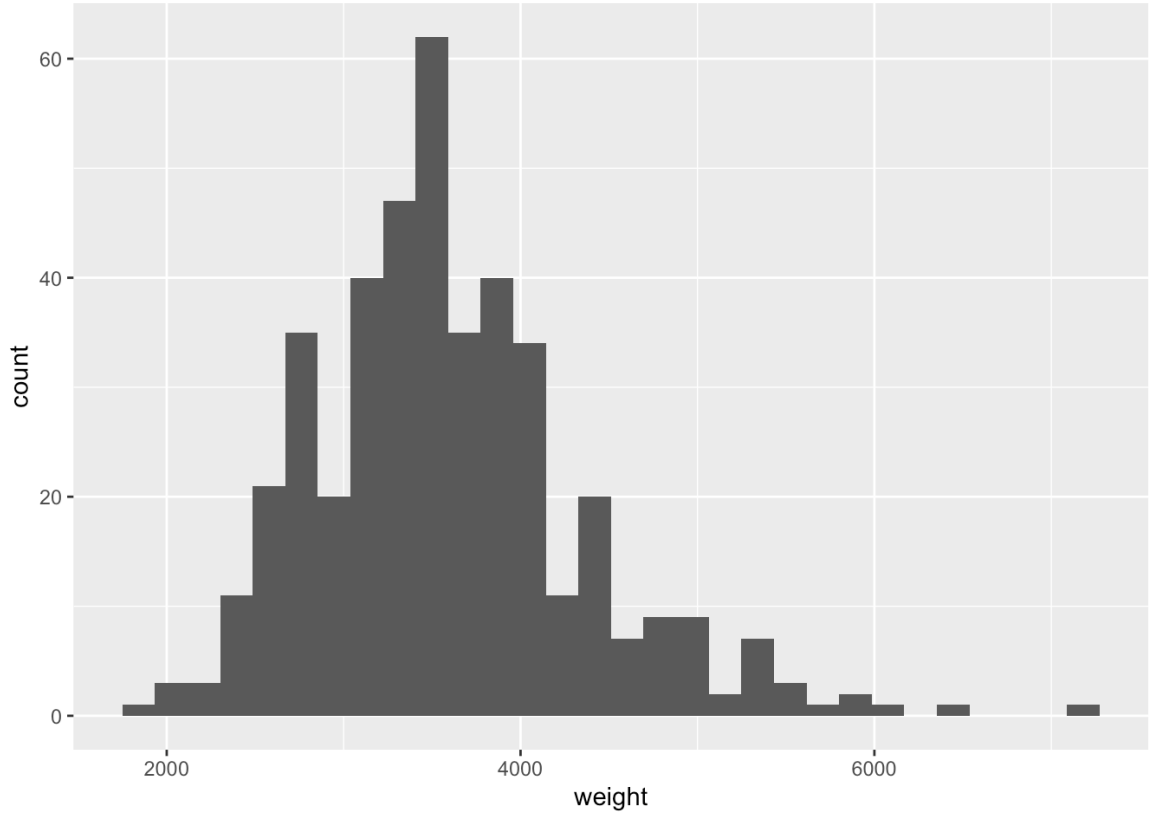


Exploring and visualizing quantitative data

```
cars %>%
  ggplot(aes(x=weight))+geom_histogram()
```

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

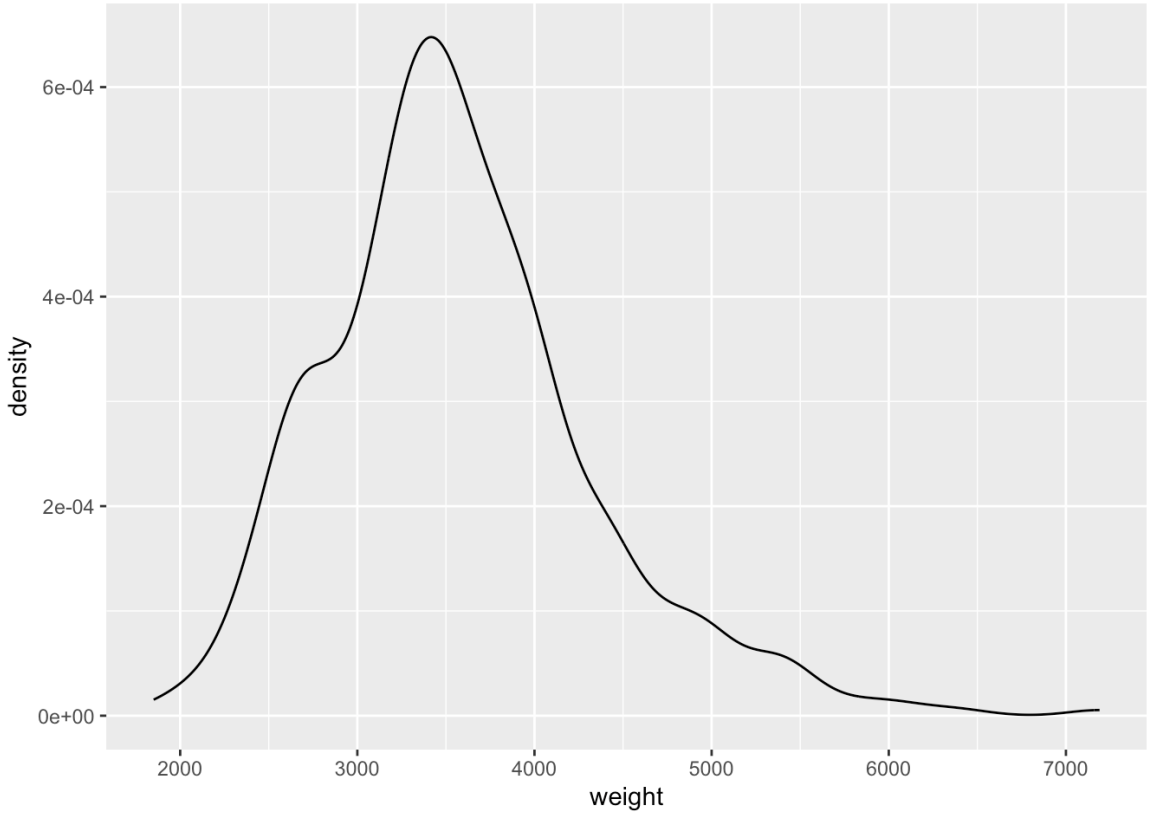
```
## Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```



Exploring and visualizing quantitative data

```
cars %>%
  ggplot(aes(x=weight))+geom_density()
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_density()`).
```



Exploring and visualizing quantitative data

Tecnológico de Monterrey

2.3 Is the city mileage per gallon different between SUV's and the other types of cars? Get the value for mean and median for city_mpg of SUV's and not-SUV's cars.

```
cars %>%
  group_by(suv) %>%
  summarize(mean_mpg=mean(city_mpg,na.rm=T),median_mpg=median(city_mpg,na.rm=T))
```

```
## # A tibble: 2 × 3
##   suv   mean_mpg median_mpg
##   <lgl>    <dbl>      <int>
## 1 FALSE     20.7         20
## 2 TRUE      16.2         16
```

RStudio®

Exploring and visualizing quantitative data

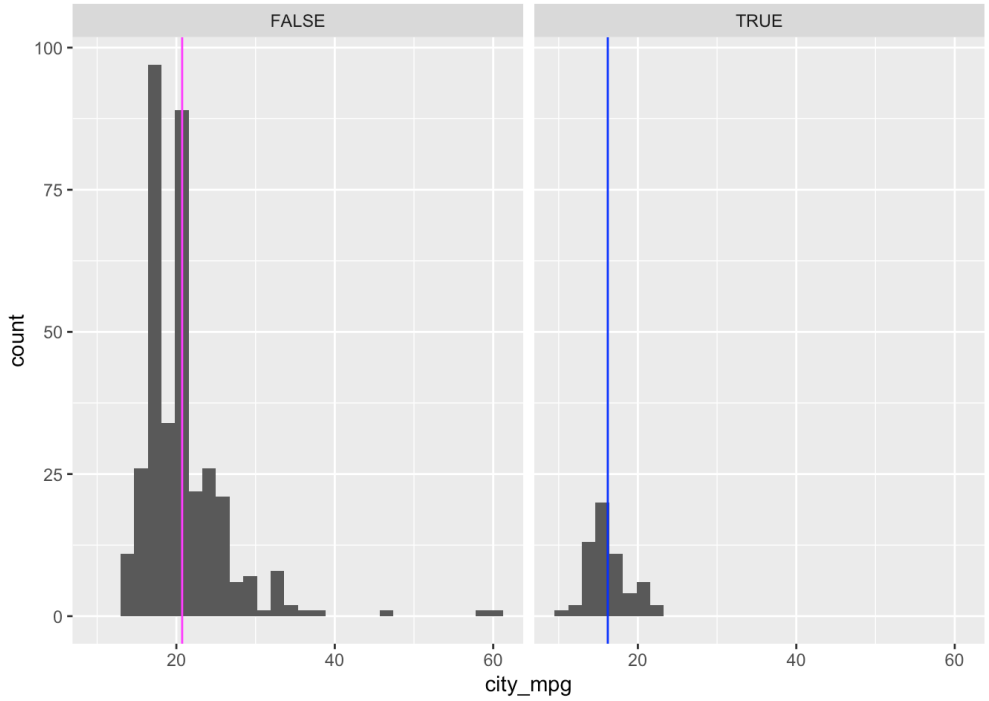
Tecnológico de Monterrey

2.4 Using a faceted histogram, describe the mileage per gallon (mpg) for the vehicles by grouping the suv's and not suv's vehicles.

```
ggplot(data=cars,aes(x=city_mpg))+geom_histogram()+facet_wrap(~suv) + geom_vline(data=subset(cars,suv==TRUE),aes(
xintercept=16.20),col="blue")+ geom_vline(data=subset(cars,suv==FALSE),aes(xintercept=20.73),col="magenta")
```

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

```
## Warning: Removed 14 rows containing non-finite values (`stat_bin()`).
```



RStudio®

Exploring and visualizing quantitative data

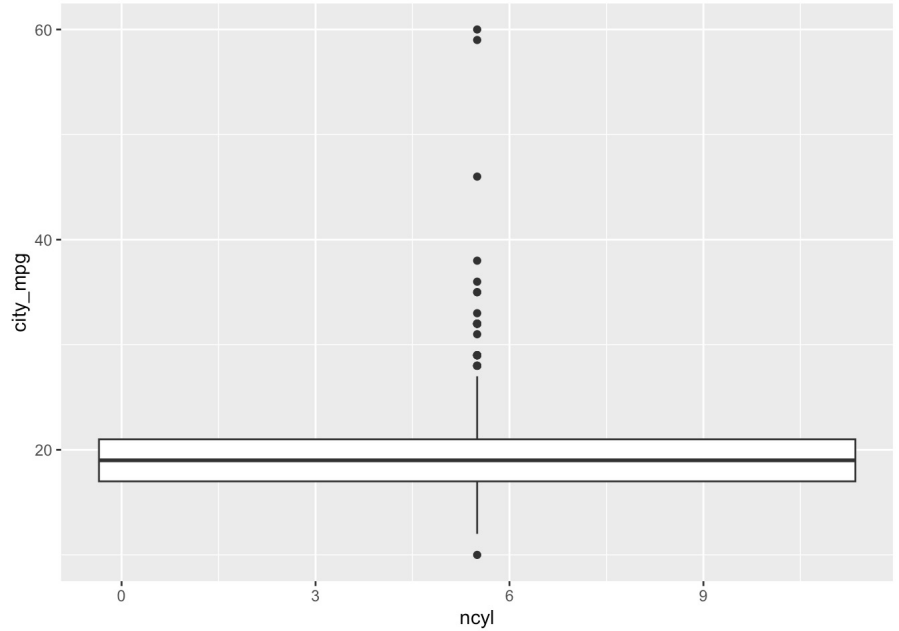
Tecnológico de Monterrey

2.5 Build a side-by-side boxplot to visualize the city miles per galone (city_mpg) depending on the numbers of cilynders of the vehicle (ncyil).

```
ggplot(data=cars,aes(x=ncyl,y=city_mpg)) + geom_boxplot()
```

```
## Warning: Continuous x aesthetic
## ℹ did you forget `aes(group = ...)`?
```

```
## Warning: Removed 14 rows containing non-finite values (`stat_boxplot()`).
```



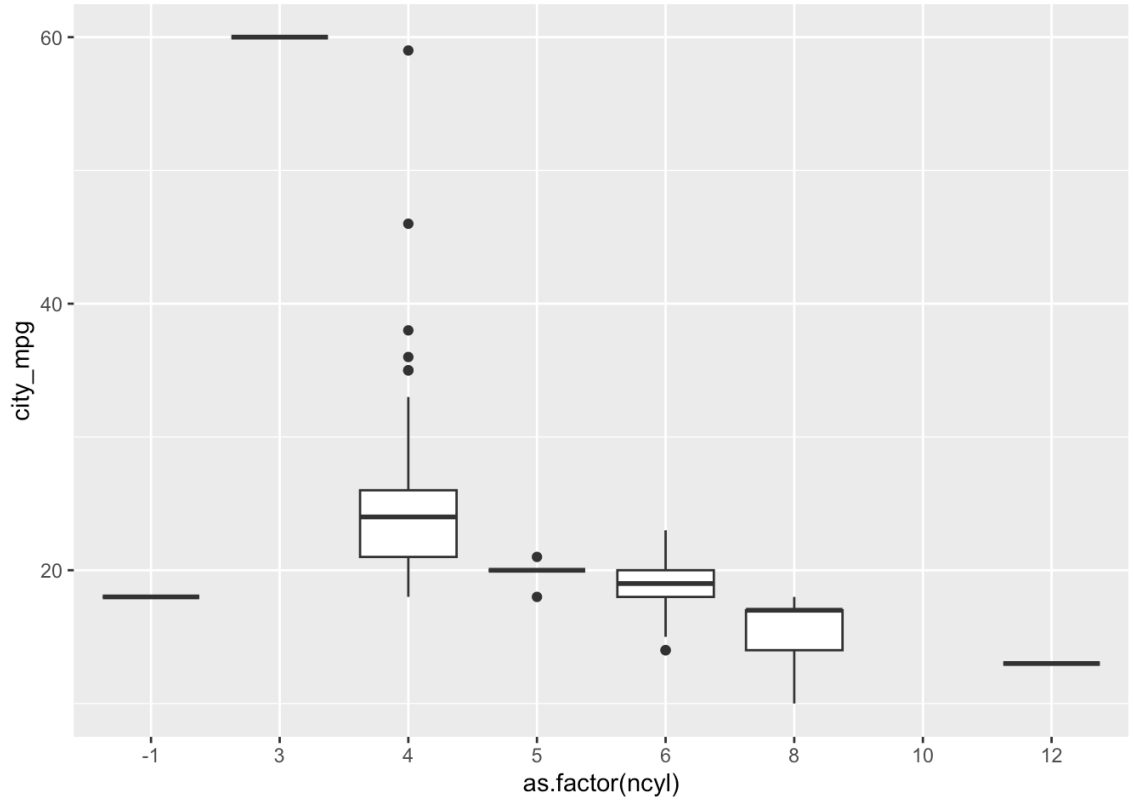

```
class(cars$ncyl)
```

RStudio®

Exploring and visualizing quantitative data

```
ggplot(data=cars,aes(x=as.factor(ncyl),y=city_mpg)) + geom_boxplot()
```

```
## Warning: Removed 14 rows containing non-finite values (`stat_boxplot()`).
```



Exploring and visualizing quantitative data

Tecnológico de Monterrey

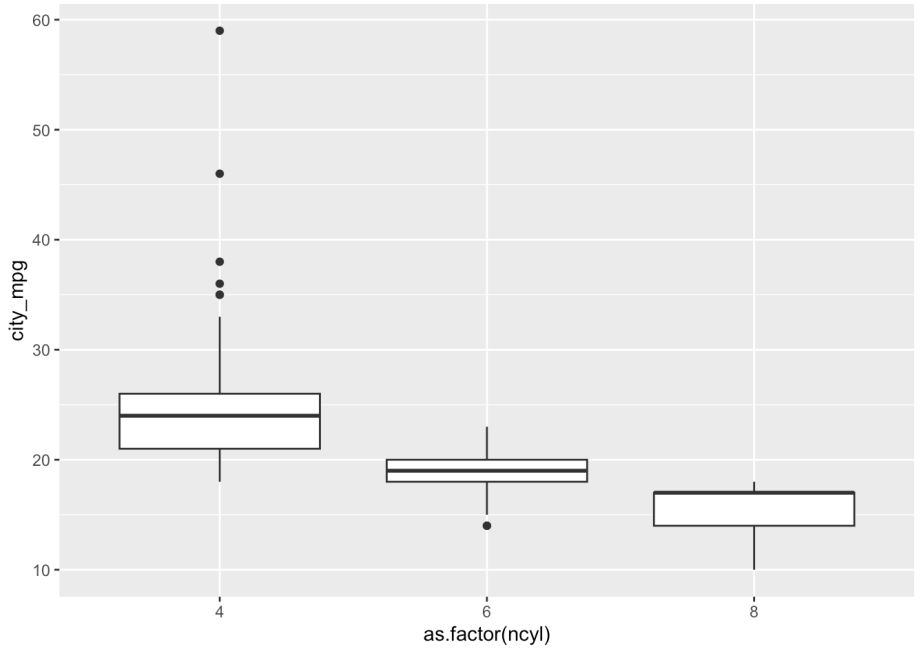
```
levels(as.factor(cars$ncyl))
```

```
## [1] "-1" "3"  "4"  "5"  "6"  "8"  "10" "12"
```

2.6 It seems that there are errors in the values for number of cylinders. Filter cars to include only cars with 4, 6, or 8 cylinders and save the result as common_cyl. The %in% operator may prove useful here.

```
common_cyl <- cars %>%
  filter(ncyl %in% c(4,6,8))
  ggplot(data=common_cyl,aes(x=as.factor(ncyl),y=city_mpg)) + geom_boxplot()
```

```
## Warning: Removed 11 rows containing non-finite values (`stat_boxplot()`).
```



RStudio®

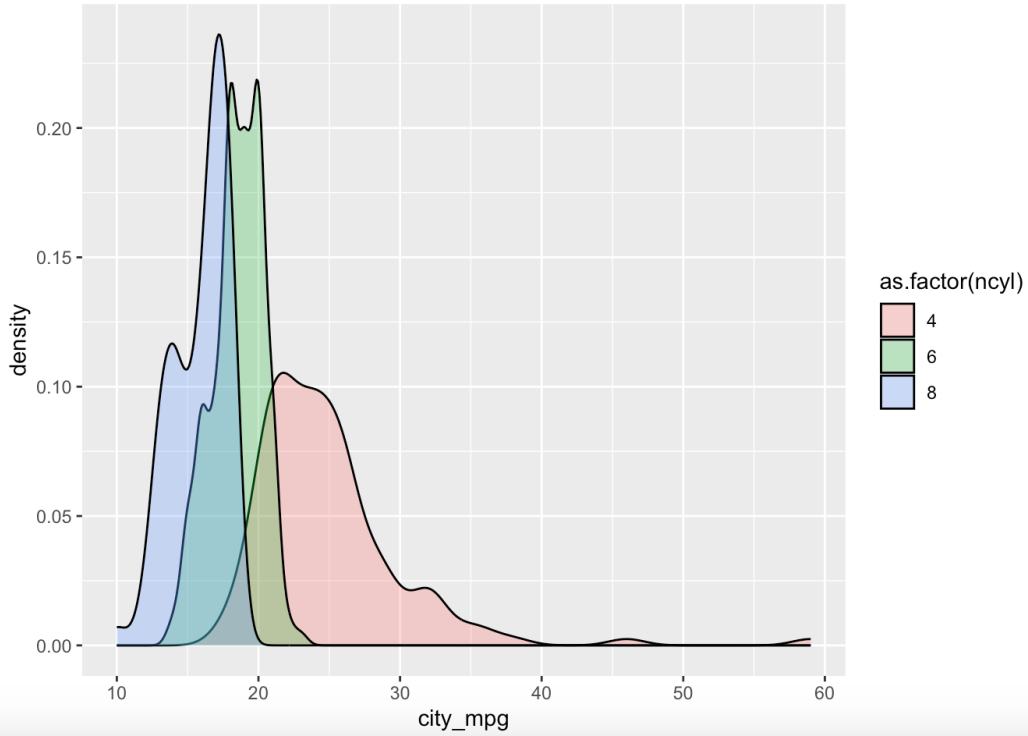
Exploring and visualizing quantitative data

Tecnológico de Monterrey

2.7 Buil overlapped density plots to show the distribution of miles per gallon depending on the number of cylinders.

```
ggplot(common_cyl, aes(x = city_mpg, fill = as.factor(ncyl))) +
  geom_density(alpha = .3)
```

```
## Warning: Removed 11 rows containing non-finite values (`stat_density()`).
```



RStudio®


Statistics
organization
scientific
descriptive
analysis
samples
collection
data
interpretation
statistical
Inference
theory
experiments
science
population
prediction
probability
mathematical
mathematics
called
design
discipline
pertaining
parameters
provides
inferential
consider
concerned
word
wide
also
well
statistic
total
including
often
branch
referring
used
experience
application
deduce
part
considered
social
working
useful
ways
improve
process
accounts
problems
soundness
terms
element
using
necessary
successful
summarize
empirical
logically
companies
usually
modeled
expertise
sciences
people
gained
thinking
planning
models direct
inferences
collecting
since
given
larger
roots
wrong
survey
comprise
applications
aspects
art
applied
difference
variety
studied
interpreting
randomness
based
starts
guiding
natural
someone
versed
vital
statistician
others
help
moves
organizations
inductive
grouped
provide
confused
one
observations
theoretical
prove
draw
presentation
methods
consultants
disciplines
set
hypothesis
mean
particularly
increase
forecasting
particular
sampling
Statisticians
explanation
closely
research
seems
may
focus
probabilities
misleading
new
academic
way
opposite
together
communicating
uncertainty
calculated
quality
plural
rather
fields
method
patterns
true
results
government
direction
number
deals
singular
distinct
whose
roughly
basis
holds
predictions
subject
surveys
tested
describe
tools
addition
relevant
predictive
quantity
median
access
business
applicable
related
available
without
advance
leads
a.k.a