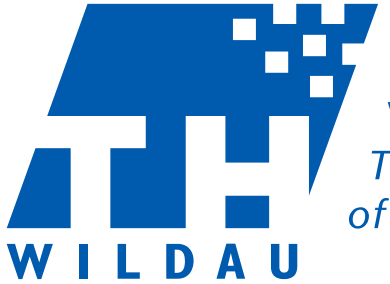


Technische
Hochschule
Wildau [FH]
Technical University
of Applied Sciences

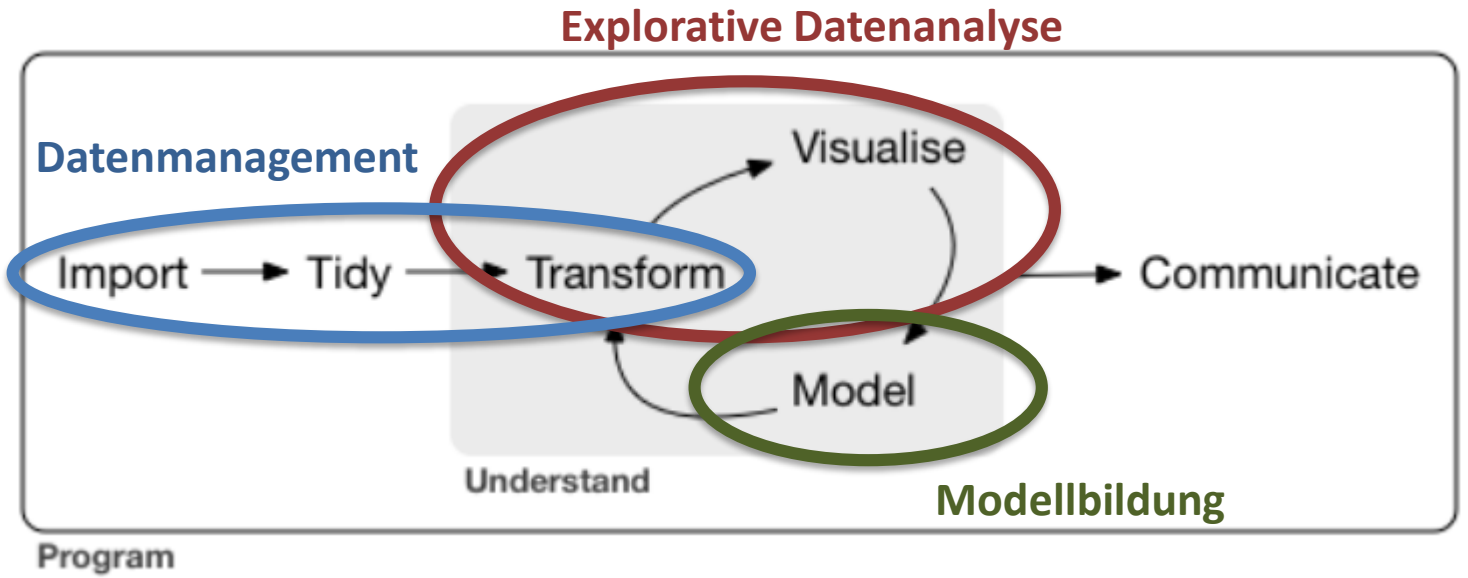
Explorative Datenanalyse

Prof. Dr. Rainer Stollhoff

Vgl.
R for Data Science, Grolemund & Wickham,
http://r4ds.had.co.nz/exploratory-data-analysis.html

Übersicht

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU



Explorative Datenanalyse

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

- Beschreibung des Datensatzes
- Verteilungen der Werte einzelner Variablen (Univariate Analysen)
- Zusammenhänge zwischen Variablen (Bivariate Analysen)

Explorative Datenanalyse

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

- **Beschreibung des Datensatzes**
  - Typische Fragen:
    - Wieviele Beobachtungen sind im Datensatz?
    - Wieviele Variablen sind im Datensatz
    - Welchen Datentyp haben die Variablen und was für Wertebereiche?
  - `View()` zur Anzeige des gesamten Datensatzes und `head()` für die ersten Zeilen
  - `dim()` zur Anzeige der Zeilen und Spalten
  - `str()` für Informationen zu den enthaltenen Variablen
  - `summary()` für einen Überblick über die Wertebereiche der Variablen
- Verteilungen der Werte einzelner Variablen (Univariate Analysen)
- Zusammenhänge zwischen Variablen (Bivariate Analysen)

Beschreibung des Datensatzes

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

```
> ## Zeigt die ersten paar Zeilen
> head(mpg)
# A tibble: 6 x 11
  manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
  <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compact
2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compact
3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compact
4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compact
5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compact
6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compact
> ## Größe des Datensatzes
> dim(mpg)
[1] 234  11
> ## Struktur der Variablen
> str(mpg)
tibble [234 x 11] (S3: tbl_df/tbl/data.frame)
 $ manufacturer: chr [1:234] "audi" "audi" "audi" "audi" ...
 $ model       : chr [1:234] "a4" "a4" "a4" "a4" ...
 $ displ       : num [1:234] 1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
 $ year        : int [1:234] 1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
 $ cyl         : int [1:234] 4 4 4 4 6 6 6 4 4 4 ...
 $ trans       : chr [1:234] "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
 $ drv         : chr [1:234] "f" "f" "f" "f" ...
 $ cty         : int [1:234] 18 21 20 21 16 18 18 18 16 20 ...
 $ hwy         : int [1:234] 29 29 31 30 26 26 27 26 25 28 ...
 $ fl          : chr [1:234] "p" "p" "p" "p" ...
 $ class       : chr [1:234] "compact" "compact" "compact" "compact" ...
> ## Statistische Zusammenfassung des Wertebereichs der Variablen
> summary(mpg)
 manufacturer          model               displ            year          cyl            trans
 Length:234         Length:234         Min.   :1.600   Min.   :1999   Min.   :4.000   Length:234
 Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999   1st Qu.:4.000   Class :character
 Mode  :character   Mode  :character   Median :3.300   Median :2004   Median :6.000   Mode  :character
                                       Mean   :3.472   Mean   :2004   Mean   :5.889
                                       3rd Qu.:4.600   3rd Qu.:2008   3rd Qu.:8.000
                                       Max.   :7.000   Max.   :2008   Max.   :8.000
     drv                 cty              hwy             fl                class
 Length:234         Min.   : 9.00   Min.   :12.00   Length:234         Length:234
 Class :character   1st Qu.:14.00   1st Qu.:18.00   Class :character   Class :character
 Mode  :character   Median :17.00   Median :24.00   Mode  :character   Mode  :character
                    Mean   :16.86   Mean   :23.44
                    3rd Qu.:19.00   3rd Qu.:27.00
                    Max.   :35.00   Max.   :44.00
> |
```

Zeilen
Beobachtg

Spalten
Variablen

Explorative Datenanalyse

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

- Beschreibung des Datensatzes
- Verteilungen der Werte einzelner Variablen (Univariate Analysen)
  - Typische Fragen:
    - Welche Werte nimmt die Variable an?
    - Was sind typische / untypische Werte?
    - Gibt es Häufungen?
  - `arrange()` zum Sortieren der Daten
  - `summarise()` zum Aggregieren der Daten z.B. `mean, max, min`
  - `count()` zum Berechnen von Häufigkeiten
  - `geom_bar()` bei kategorischen Variablen
  - `geom_histogram(), geom_freqpoly()` bei stetigen Variablen
- Zusammenhänge zwischen Variablen (Bivariate Analysen)

Univariate Analyse

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

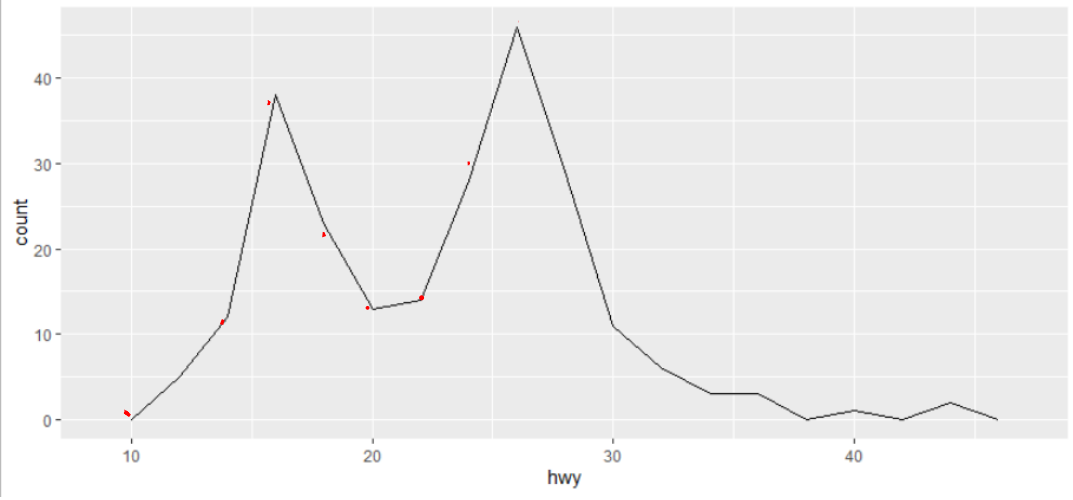
```
> ## Wer hat die niedrigste Reichweite?
> arrange(mpg, cty)
# A tibble: 234 x 11
   manufacturer model              displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>              <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 dodge        dakota pickup 4wd    4.7  2008     8 auto(l5)   4         9    12 e     pickup
 2 dodge        durango 4wd          4.7  2008     8 auto(l5)   4         9    12 e     suv
 3 dodge        ram 1500 pickup 4wd  4.7  2008     8 auto(l5)   4         9    12 e     pickup
 4 dodge        ram 1500 pickup 4wd  4.7  2008     8 manual(m6) 4         9    12 e     pickup
 5 jeep         grand cherokee 4wd   4.7  2008     8 auto(l5)   4         9    12 e     suv
 6 chevrolet    c1500 suburban 2wd   5.3  2008     8 auto(l4)   r        11    15 e     suv
 7 chevrolet    k1500 tahoe 4wd      5.3  2008     8 auto(l4)   4        11    14 e     suv
 8 chevrolet    k1500 tahoe 4wd      5.7  1999     8 auto(l4)   4        11    15 r     suv
 9 dodge        caravan 2wd          3.3  2008     6 auto(l4)   f        11    17 e     minivan
10 dodge        dakota pickup 4wd    5.2  1999     8 manual(m5) 4        11    17 r     pickup
# ... with 224 more rows
> ## Und wer die höchste?
> arrange(mpg, desc(cty))
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 volkswagen   new beetle   1.9  1999     4 manual(m5) f        35    44 d     subcompact
 2 volkswagen   jetta        1.9  1999     4 manual(m5) f        33    44 d     compact
 3 volkswagen   new beetle   1.9  1999     4 auto(l4)   f        29    41 d     subcompact
 4 honda        civic        1.6  1999     4 manual(m5) f        28    33 r     subcompact
 5 toyota       corolla      1.8  2008     4 manual(m5) f        28    37 r     compact
 6 honda        civic        1.8  2008     4 manual(m5) f        26    34 r     subcompact
 7 toyota       corolla      1.8  1999     4 manual(m5) f        26    35 r     compact
 8 toyota       corolla      1.8  2008     4 auto(l4)   f        26    35 r     compact
 9 honda        civic        1.6  1999     4 manual(m5) f        25    32 r     subcompact
10 honda        civic        1.8  2008     4 auto(l5)   f        25    36 r     subcompact
# ... with 224 more rows
```

```
> ## Was ist die durchschnittliche Reichweite?
> summarise(mpg,mean(cty))
# A tibble: 1 x 1
  `mean(cty)`
        <dbl>
1        16.9
> ## Was sind die Mittelwerte?
> summarise_if(mpg,is.numeric,funs(mean))
# A tibble: 1 x 5
  displ  year   cyl   cty   hwy
  <dbl> <dbl> <dbl> <dbl> <dbl>
1  3.47 2004.  5.89  16.9  23.4
> ## Wieviele Autos gibt es pro Hersteller?
> count(mpg, manufacturer)
# A tibble: 15 x 2
   manufacturer     n
   <chr>        <int>
 1 audi            18
 2 chevrolet       19
 3 dodge           37
 4 ford            25
 5 honda            9
 6 hyundai         14
 7 jeep             8
 8 land rover       4
 9 lincoln          3
10 mercury          4
11 nissan          13
12 pontiac          5
13 subaru          14
14 toyota          34
15 volkswagen      27
```

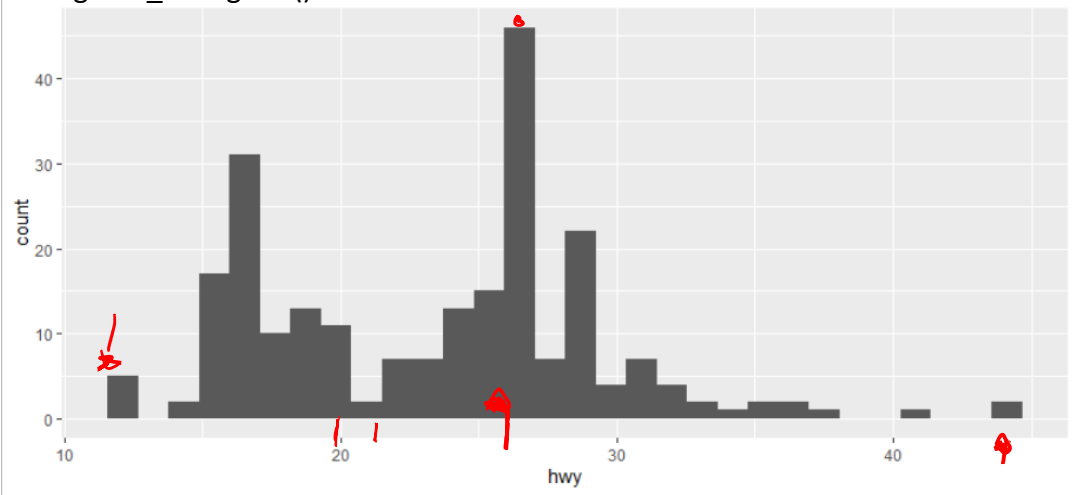
Univariate Analyse

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

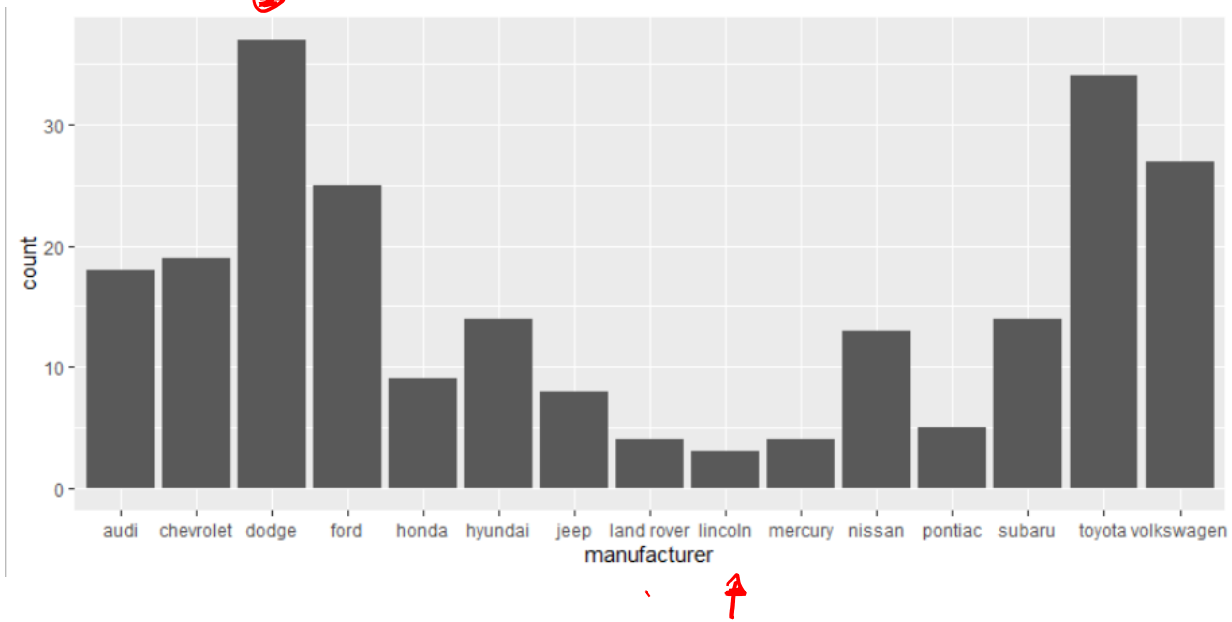
```
ggplot(mpg, mapping=aes(x=hwy)) +
  geom_freqpoly(binwidth=2)
```




```
ggplot(mpg, mapping=aes(x=hwy)) +
  geom_histogram()
```




```
ggplot(mpg, mapping=aes(x=manufacturer)) +
  geom_bar()
```



Explorative Datenanalyse

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

- Beschreibung des Datensatzes
- Verteilungen der Werte einzelner Variablen (Univariate Analysen)
- **Zusammenhänge zwischen Variablen (Bivariate Analysen)**
  - Typische Fragen:
    - Gibt es Zusammenhänge zwischen Variablen?
    - Wenn ja, sind diese positiv/negativ, sind diese stark oder schwach ausgeprägt?
    - Gibt es nichtlineare Zusammenhänge?
  - `summarise()` in Verbindung mit `group_by()`
  - `cor()` und `cov()` zum Berechnen statistischer Zusammenhangsmaße
  - `stat_bin(x=stet,color=kat)` oder `geom_boxplot(x=kat,y=stet)` für x stetig und y kategorisch
  - `geom_count(x=kat1,y=kat2)` für x und y kategorisch
  - `geom_point(x=stet1,y=stet2)` für x und y stetig

Bivariate Analyse

Technische Hochschule Wildau
Technical University of Applied Sciences
WILDAU

```
> ## Minimale, Maximale und durchschnittliche Reichweite je Hersteller
> summarise(
+   group_by(mpg,manufacturer),
+   min(cty),max(cty),mean(cty))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 15 x 4
   manufacturer `min(cty)` `max(cty)` `mean(cty)`
   <chr>             <int>      <int>        <dbl>
 1 audi                 15         21         17.6
 2 chevrolet            11         22         15
 3 dodge                 9         18         13.1
 4 ford                 11         18         14
 5 honda                21         28         24.4
 6 hyundai              16         21         18.6
 7 jeep                  9         17         13.5
 8 land rover           11         12         11.5
 9 lincoln              11         12         11.3
10 mercury              13         14         13.2
11 nissan               12         23         18.1
12 pontiac              16         18         17
13 subaru               18         21         19.3
14 toyota               11         28         18.5
15 volkswagen           16         35         20.9
```

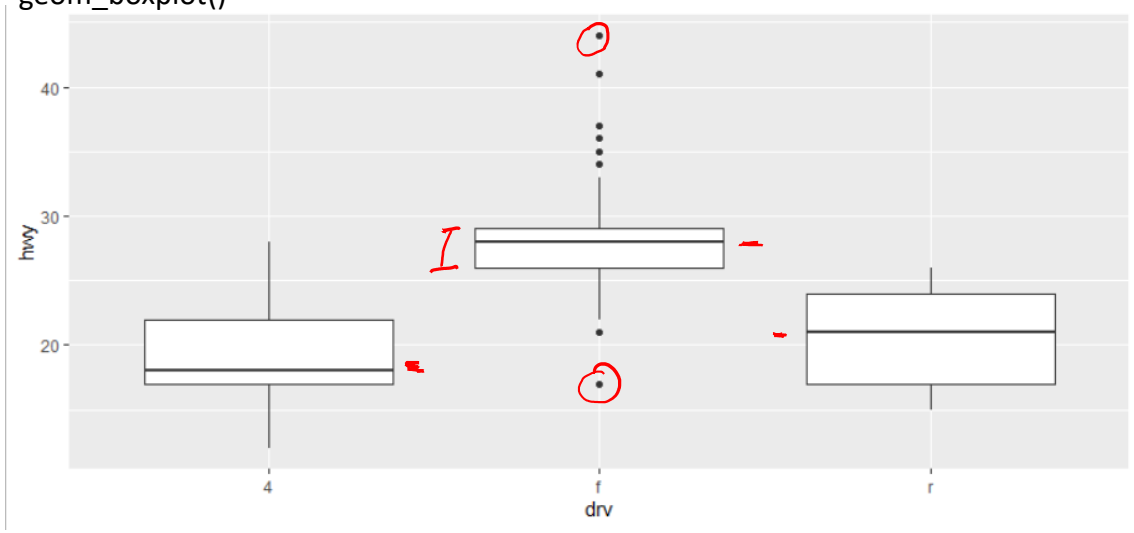
cor(mpg) → Fehler

```
> ## Korrelationen zwisschen den numerischen Werten
> cor(select_if(mpg,is.numeric))
            displ         year        cyl         cty         hwy
displ   1.0000000  0.147842816  0.9302271 -0.79852397 -0.76602002
year    0.1478428  1.000000000  0.1222453 -0.03723229  0.00215764
cyl     0.9302271  0.122245347  1.0000000 -0.80577141 -0.76191235
cty    -0.7985240 -0.037232291 -0.8057714  1.00000000  0.95591591
hwy    -0.7660200  0.002157643 -0.7619124  0.95591591  1.00000000
```

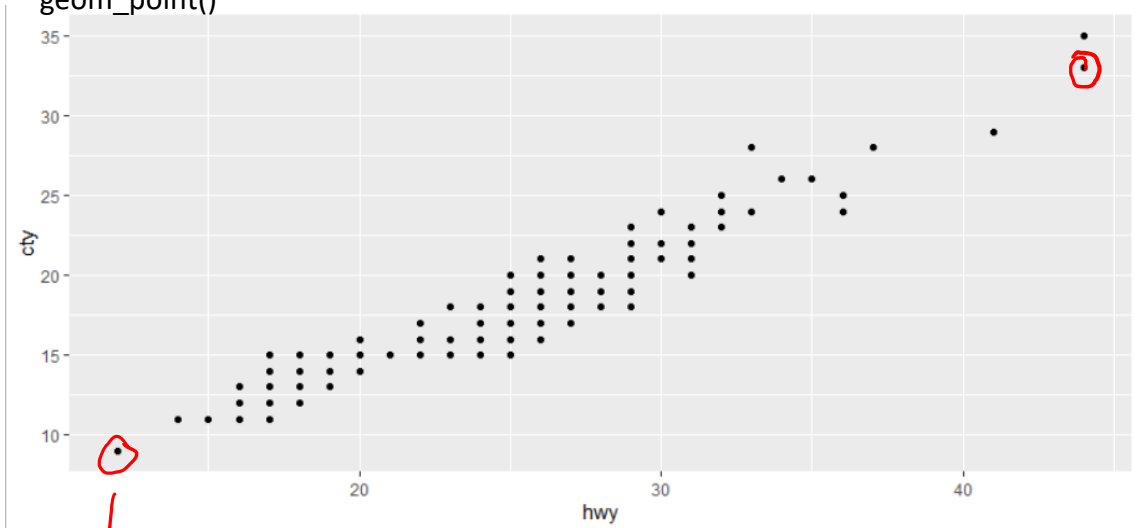
Bivariate Analyse

Technische Hochschule Wildau
Technical University of Applied Sciences

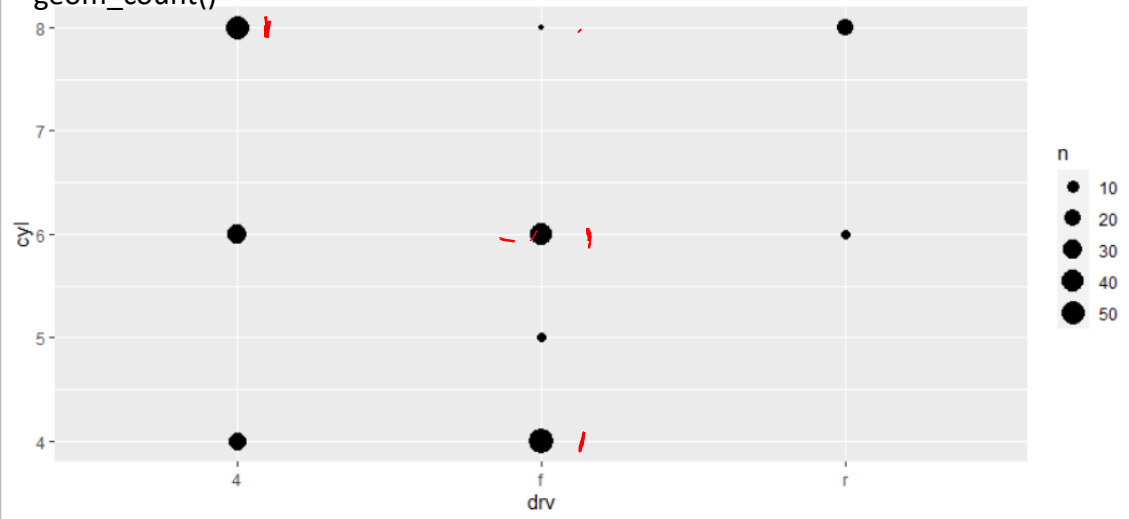
```
ggplot(mpg, mapping=aes(x=drv,y=hwy)) +
  geom_boxplot()
```



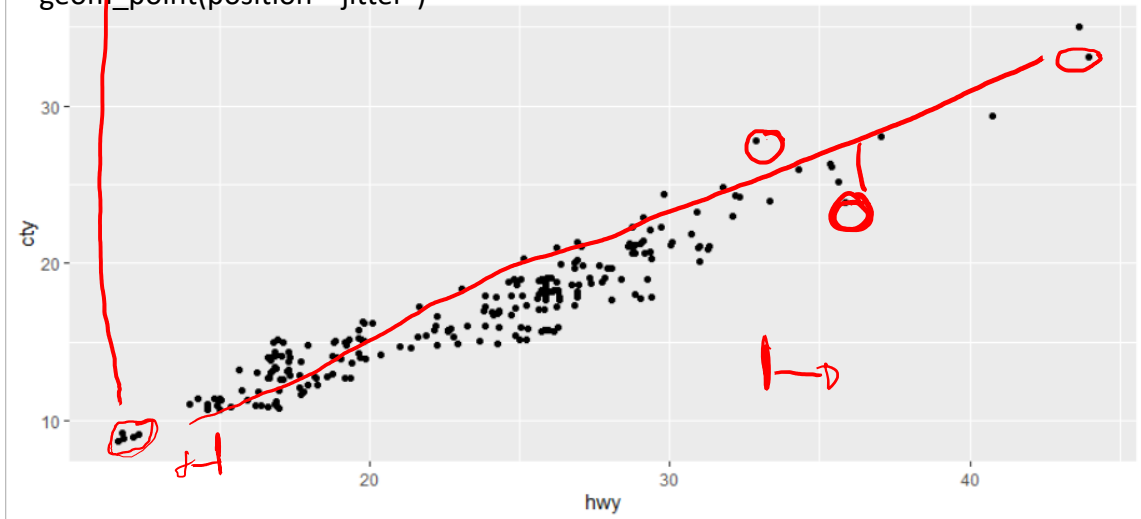

```
ggplot(mpg,mapping=aes(x=hwy,y=cty)) +
  geom_point()
```




```
ggplot(mpg,mapping=aes(x=drv,y=cyl)) +
  geom_count()
```




```
ggplot(mpg, mapping=aes(x=hwy,y=cty)) +
  geom_point(position="jitter")
```



Multivariate Analyse

Technische Hochschule Wildau
Technical University of Applied Sciences

```
ggplot(mpg, mapping=aes(x=hwy,y=displ,shape=drv,color=class)) +
  geom_point(position="jitter")
```