Projet R - Analyse des données

Table of Contents

Imports

In [ ]:

```
# install.packages("corrplot")
```

In [2]:

```
library(tidyr)
library(ggplot2)
library(dplyr)
library(nycflights13)
library(knitr)
library(corrplot)
citation("corrplot")
#conda install -c conda-forge r-performanceanalytics
library("PerformanceAnalytics")
```

```
To cite corrplot in publications use:

  Taiyun Wei and Viliam Simko (2017). R package "corrplot":
  Visualization of a Correlation Matrix (Version 0.84). Available from
  https://github.com/taiyun/corrplot

A BibTeX entry for LaTeX users is

  @Manual{corrplot2017,
    title = {R package "corrplot": Visualization of a Correlation Matrix},
    author = {Taiyun Wei and Viliam Simko},
    year = {2017},
    note = {(Version 0.84)},
    url = {https://github.com/taiyun/corrplot},
  }
```

five datasets saved as “data frames

Premiers traitements

Parcourez la base flights en affichant les noms et les types des variables présentes.

Vous pouvez aussi accéder à un dictionnaire des variables en tapant ?nycflights13::flights dans votre console.

In [4]:

```
str(flights)
```

```
Classes 'tbl_df', 'tbl' and 'data.frame':       336776 obs. of  19 variables:
 $ year          : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
 $ month         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ day           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ dep_time      : int  517 533 542 544 554 554 555 557 557 558 ...
 $ sched_dep_time: int  515 529 540 545 600 558 600 600 600 600 ...
 $ dep_delay     : num  2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
 $ arr_time      : int  830 850 923 1004 812 740 913 709 838 753 ...
 $ sched_arr_time: int  819 830 850 1022 837 728 854 723 846 745 ...
 $ arr_delay     : num  11 20 33 -18 -25 12 19 -14 -8 8 ...
 $ carrier       : chr  "UA" "UA" "AA" "B6" ...
 $ flight        : int  1545 1714 1141 725 461 1696 507 5708 79 301 ...
 $ tailnum       : chr  "N14228" "N24211" "N619AA" "N804JB" ...
 $ origin        : chr  "EWR" "LGA" "JFK" "JFK" ...
 $ dest          : chr  "IAH" "IAH" "MIA" "BQN" ...
 $ air_time      : num  227 227 160 183 116 150 158 53 140 138 ...
 $ distance      : num  1400 1416 1089 1576 762 ...
 $ hour          : num  5 5 5 5 6 5 6 6 6 6 ...
 $ minute        : num  15 29 40 45 0 58 0 0 0 0 ...
 $ time_hour     : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00"
...
```

In [3]:

```
mode(flights)
```

'list'

In [4]:

```
class(flights)
```

'tbl_df' 'tbl' 'data.frame'

In [5]:

```
introduce(flights)
```

```
Error in introduce(flights): impossible de trouver la fonction "introduce"
Traceback:
```

In [5]:

```
head(flights)
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight
2013	1	1	517	515	2	830	819	11	UA	1545
2013	1	1	533	529	4	850	830	20	UA	1714
2013	1	1	542	540	2	923	850	33	AA	1141
2013	1	1	544	545	-1	1004	1022	-18	B6	725
2013	1	1	554	600	-6	812	837	-25	DL	461
2013	1	1	554	558	-4	740	728	12	UA	1696

In [6]:

```
nycflights13::flights
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight
2013	1	1	517	515	2	830	819	11	UA	1545
2013	1	1	533	529	4	850	830	20	UA	1714
2013	1	1	542	540	2	923	850	33	AA	1141
2013	1	1	544	545	-1	1004	1022	-18	B6	725
2013	1	1	554	600	-6	812	837	-25	DL	461
2013	1	1	554	558	-4	740	728	12	UA	1696
2013	1	1	555	600	-5	913	854	19	B6	507
2013	1	1	557	600	-3	709	723	-14	EV	5708
2013	1	1	557	600	-3	838	846	-8	B6	79
2013	1	1	558	600	-2	753	745	8	AA	301
2013	1	1	558	600	-2	849	851	-2	B6	49
2013	1	1	558	600	-2	853	856	-3	B6	71
2013	1	1	558	600	-2	924	917	7	UA	194
2013	1	1	558	600	-2	923	937	-14	UA	1124
2013	1	1	559	600	-1	941	910	31	AA	707
2013	1	1	559	559	0	702	706	-4	B6	1806
2013	1	1	559	600	-1	854	902	-8	UA	1187
2013	1	1	600	600	0	851	858	-7	B6	371
2013	1	1	600	600	0	837	825	12	MQ	4650
2013	1	1	601	600	1	844	850	-6	B6	343
2013	1	1	602	610	-8	812	820	-8	DL	1919
2013	1	1	602	605	-3	821	805	16	MQ	4401
2013	1	1	606	610	-4	858	910	-12	AA	1895
2013	1	1	606	610	-4	837	845	-8	DL	1743
2013	1	1	607	607	0	858	915	-17	UA	1077
2013	1	1	608	600	8	807	735	32	MQ	3768
2013	1	1	611	600	11	945	931	14	UA	303

In [20]:
```
?flights
```

In [22]:
```
?nycflights13::flights
```

In [7]:
```
print(nycflights13::flights)
```

```
# A tibble: 336,776 x 19
    year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
   <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
 1  2013     1     1      517            515         2      830            819
 2  2013     1     1      533            529         4      850            830
 3  2013     1     1      542            540         2      923            850
 4  2013     1     1      544            545        -1     1004           1022
 5  2013     1     1      554            600        -6      812            837
 6  2013     1     1      554            558        -4      740            728
 7  2013     1     1      555            600        -5      913            854
 8  2013     1     1      557            600        -3      709            723
 9  2013     1     1      557            600        -3      838            846
10  2013     1     1      558            600        -2      753            745
# ... with 336,766 more rows, and 11 more variables: arr_delay <dbl>,
#   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
#   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
```

In [3]:
```
View(flights)
```

```
Error in View(flights): 'View()' not yet supported in the Jupyter R kernel
Traceback:

1. View(flights)
2. stop(sQuote("View()"), " not yet supported in the Jupyter R kernel")
```

In [10]:
```
glimpse(flights)
```

```
Observations: 336,776
Variables: 19
$ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013...
$ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
$ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 55...
$ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 60...
$ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2,...
$ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 8...
$ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 8...
$ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7,...
$ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6"...
$ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301...
$ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N...
$ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LG...
$ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IA...
$ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149...
$ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 73...
$ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6...
$ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59...
$ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-0...
```

In [12]:
```
colnames(flights)
```

'year' 'month' 'day' 'dep_time' 'sched_dep_time' 'dep_delay' 'arr_time' 'sched_arr_time'
'arr_delay' 'carrier' 'flight' 'tailnum' 'origin' 'dest' 'air_time' 'distance' 'hour' 'minute' 'time_hour'

In [24]:
```
summary(flights)
```

```
      year          month             day           dep_time     sched_dep_time
 Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106  
 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906  
 Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359  
 Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344  
 3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729  
 Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359  
                                                 NA's   :8255                 
   dep_delay          arr_time    sched_arr_time   arr_delay       
 Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000  
 1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000  
 Median :  -2.00   Median :1535   Median :1556   Median :  -5.000  
 Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895  
 3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000  
 Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000  
 NA's   :8255      NA's   :8713                  NA's   :9430      
   carrier              flight       tailnum             origin         
 Length:336776      Min.   :   1   Length:336776      Length:336776     
 Class :character   1st Qu.: 553   Class :character   Class :character  
 Mode  :character   Median :1496   Mode  :character   Mode  :character  
                    Mean   :1972                                        
                    3rd Qu.:3465                                        
                    Max.   :8500                                        
                                                                        
     dest              air_time        distance         hour      
 Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00  
 Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00  
 Mode  :character   Median :129.0   Median : 872   Median :13.00  
                    Mean   :150.7   Mean   :1040   Mean   :13.18  
                    3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00  
                    Max.   :695.0   Max.   :4983   Max.   :23.00  
                    NA's   :9430                                  
     minute        time_hour                  
 Min.   : 0.00   Min.   :2013-01-01 05:00:00  
 1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00  
 Median :29.00   Median :2013-07-03 10:00:00  
 Mean   :26.23   Mean   :2013-07-03 05:22:54  
 3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00  
 Max.   :59.00   Max.   :2013-12-31 23:00:00  
```

les vols

On s'intéresse plus spécifiquement à la distance du vol, sa durée, son retard au départ et à l'arrivée. En utilisant la fonction is.na, regardez si ces variables ont des valeurs manquantes. Si oui, créez un dataframe où vous les regroupez. À quoi sont dûs ces valeurs manquantes? Comment les traiter (les sortir de la table ou remplacer la valeur manquante par une autre valeur)?

distance du vol

In [11]:
```
distance_na <- filter(flights, is.na(flights$distance) == TRUE)
```

In [12]:
```
distance_na
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	t

In [13]:
```
count(distance_na)
```

n
0

sa durée

In [14]:
```
air_time_na <- filter(flights, is.na(flights$air_time) == TRUE)
```

In [15]:

```
air_time_na
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight
2013	1	1	1525	1530	-5	1934	1805	NA	MQ	4525
2013	1	1	1528	1459	29	2002	1647	NA	EV	3806
2013	1	1	1740	1745	-5	2158	2020	NA	MQ	4413
2013	1	1	1807	1738	29	2251	2103	NA	UA	1228
2013	1	1	1939	1840	59	29	2151	NA	9E	3325
2013	1	1	1952	1930	22	2358	2207	NA	EV	4333
2013	1	1	2016	1930	46	NA	2220	NA	EV	4204
2013	1	1	NA	1630	NA	NA	1815	NA	EV	4308
2013	1	1	NA	1935	NA	NA	2240	NA	AA	791
2013	1	1	NA	1500	NA	NA	1825	NA	AA	1925
2013	1	1	NA	600	NA	NA	901	NA	B6	125
2013	1	2	905	822	43	1313	1045	NA	EV	4140
2013	1	2	1125	925	120	1445	1146	NA	9E	3658
2013	1	2	1848	1840	8	2333	2151	NA	9E	3325
2013	1	2	1849	1724	85	2235	1938	NA	EV	4321
2013	1	2	1927	1930	-3	2359	2306	NA	9E	3401
2013	1	2	2041	2045	-4	NA	2359	NA	B6	147
2013	1	2	2145	2129	16	NA	33	NA	UA	1299
2013	1	2	NA	1540	NA	NA	1747	NA	EV	4352
2013	1	2	NA	1620	NA	NA	1746	NA	EV	4406
2013	1	2	NA	1355	NA	NA	1459	NA	EV	4434
2013	1	2	NA	1420	NA	NA	1644	NA	EV	4935
2013	1	2	NA	1321	NA	NA	1536	NA	EV	3849
2013	1	2	NA	1545	NA	NA	1910	NA	AA	133
2013	1	2	NA	1330	NA	NA	1640	NA	AA	753
2013	1	2	NA	1601	NA	NA	1735	NA	UA	623
2013	1	3	1025	1032	-7	1521	1240	NA	EV	4255

```
In [16]: count(air_time_na)
```

n
9430

retard au départ

```
In [17]: dep_delay_na <- filter(flights, is.na(flights$dep_delay) == TRUE)
```

```
In [18]: count(dep_delay_na)
```

n
8255

retard à l’arrivée

```
In [19]: arr_delay_na <- filter(flights, is.na(flights$arr_delay) == TRUE)
```

```
In [20]: count(arr_delay_na)
```

n
9430

Dataframe

```
In [21]: fl_na <- rbind(distance_na, air_time_na, dep_delay_na, arr_delay_na)
```

```
In [22]: count(fl_na)
```

n
27115

```
In [23]: fl_na <- unique(fl_na)
```

```
In [24]: count(fl_na)
```

n
9430

```
In [25]: head(fl_na)
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight
2013	1	1	1525	1530	-5	1934	1805	NA	MQ	4525
2013	1	1	1528	1459	29	2002	1647	NA	EV	3806
2013	1	1	1740	1745	-5	2158	2020	NA	MQ	4413
2013	1	1	1807	1738	29	2251	2103	NA	UA	1228
2013	1	1	1939	1840	59	29	2151	NA	9E	3325
2013	1	1	1952	1930	22	2358	2207	NA	EV	4333

```
In [26]: flights_NA <- sapply(flights, function(x) sum(is.na(x)))
         flights_NA
```

year	0
month	0
day	0
dep_time	8255
sched_dep_time	0
dep_delay	8255
arr_time	8713
sched_arr_time	0
arr_delay	9430
carrier	0
flight	0
tailnum	2512
origin	0
dest	0
air_time	9430
distance	0
hour	0
minute	0
time_hour	0

À quoi sont dûs ces valeurs manquantes?

Les colonnes dep_delay et arr_delay n'ont pas été calculées les valeurs de la colonne air_time ne sont pas fournis

Comment les traiter

(les sortir de la table ou remplacer la valeur manquante par une autre valeur)?

- On peut calculer dep_delay et arr_delay
- On ne peut pas calculer air_time car cette durée ne correspond pas à la différence entre l'heure de départ et l'heure d'arrivée

On supprime toutes les lignes avec la fonction

```
In [27]: na.rm = TRUE
```

Pour info les fonctions pour calculer les valeurs manquantes

```
In [ ]: ! les fonctions ci-dessous sont fausses car il faut compter en minutes (sur 60)
```

```
In [35]: fl_na$dep_delay <- fl_na$dep_time - fl_na$sched_dep_time
```

```
In [36]: fl_na$arr_delay <- fl_na$arr_time - fl_na$sched_arr_time
```

```
In [39]: head(fl_na)
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	
2013	1	1	1525	1530	-5	1934	1805	NA	MQ	4525	I
2013	1	1	1528	1459	29	2002	1647	NA	EV	3806	
2013	1	1	1740	1745	-5	2158	2020	NA	MQ	4413	I
2013	1	1	1807	1738	29	2251	2103	NA	UA	1228	
2013	1	1	1939	1840	59	29	2151	NA	9E	3325	
2013	1	1	1952	1930	22	2358	2207	NA	EV	4333	

```
In [28]: fl_na <- filter(fl_na, is.na(x) == TRUE)
```

```
Error: objet 'x' introuvable
Traceback:

1. filter(fl_na, is.na(x) == TRUE)
2. filter.tbl_df(fl_na, is.na(x) == TRUE)
3. filter_impl(.data, quo)
```

```
In [41]: head(is.na(fl_na))
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	fl
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAI
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAI
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAI
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAI
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAI
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FAI

Statistiques

Sur ces variables, présentez des statistiques (moyenne, écart-type, min, max). Observez-vous des différences de ces statistiques selon l'aéroport d'où est parti l'avion?

Moyenne

- variables : distance du vol, sa durée, son retard au départ et à l'arrivée
- aéroport départ (origin)

In [28]:

```
#moyenne globale
summarize(flights, delay=mean(dep_delay, na.rm=TRUE))
```

delay
12.63907

In [29]:

```
mean_by_origin <- flights %>%
    group_by(origin) %>%
    summarize(dist=mean(distance, na.rm=TRUE),
              air_time=mean(air_time, na.rm=TRUE),
              dep_delay=mean(dep_delay, na.rm=TRUE),
              arr_delay=mean(arr_delay, na.rm=TRUE))
```

In [30]:

```
mean_by_origin
```

origin	dist	air_time	dep_delay	arr_delay
EWR	1056.7428	153.3000	15.10795	9.107055
JFK	1266.2491	178.3490	12.11216	5.551481
LGA	779.8357	117.8258	10.34688	5.783488

écart-type (sd)

In [31]:

```
sd_by_origin <- flights %>%
    group_by(origin) %>%
    summarize(dist=sd(distance, na.rm=TRUE),
              air_time=sd(air_time, na.rm=TRUE),
              dep_delay=sd(dep_delay, na.rm=TRUE),
              arr_delay=sd(arr_delay, na.rm=TRUE))
```

In [32]:

```
sd_by_origin
```

origin	dist	air_time	dep_delay	arr_delay
EWR	730.2239	93.34380	41.32370	45.52918
JFK	896.1084	113.79430	39.03507	44.27745
LGA	371.6615	49.39791	39.99302	43.86227

min

In [33]:

```
min_by_origin <- flights %>%
    group_by(origin) %>%
    summarize(dist=min(distance, na.rm=TRUE),
              air_time=min(air_time, na.rm=TRUE),
              dep_delay=min(dep_delay, na.rm=TRUE),
              arr_delay=min(arr_delay, na.rm=TRUE))
```

In [34]:

```
min_by_origin
```

origin	dist	air_time	dep_delay	arr_delay
EWR	17	20	-25	-86
JFK	94	21	-43	-79
LGA	96	21	-33	-68

max

In [35]:

```
max_by_origin <- flights  %>%
    group_by(origin)  %>%
    summarize(dist=max(distance, na.rm=TRUE),
              air_time=max(air_time, na.rm=TRUE),
              dep_delay=max(dep_delay, na.rm=TRUE),
              arr_delay=max(arr_delay, na.rm=TRUE))
```

In [36]:

```
max_by_origin
```

origin	dist	air_time	dep_delay	arr_delay
EWR	4963	695	1126	1109
JFK	4983	691	1301	1272
LGA	1620	331	911	915

Rapprochement avec des données météo

base weather

De la même manière, parcourez la base weather et proposez un traitement des valeurs manquantes.

In [5]:

```
colnames(weather)
```

'origin' 'year' 'month' 'day' 'hour' 'temp' 'dewp' 'humid' 'wind_dir' 'wind_speed' 'wind_gust' 'precip' 'pressure' 'visib' 'time_hour'

In [37]:

```
head(nycflights13::weather)
```

origin	year	month	day	hour	temp	dewp	humid	wind_dir	wind_speed	wind_gust	precip	pressure	visi
EWR	2013	1	1	1	39.02	26.06	59.37	270	10.35702	NA	0	1012.0	1
EWR	2013	1	1	2	39.02	26.96	61.63	250	8.05546	NA	0	1012.3	1
EWR	2013	1	1	3	39.02	28.04	64.43	240	11.50780	NA	0	1012.5	1
EWR	2013	1	1	4	39.92	28.04	62.21	250	12.65858	NA	0	1012.2	1
EWR	2013	1	1	5	39.02	28.04	64.43	260	12.65858	NA	0	1011.9	1
EWR	2013	1	1	6	37.94	28.04	67.21	240	11.50780	NA	0	1012.4	1

In [38]:

```
glimpse(weather)
```

```
Observations: 26,115
Variables: 15
$ origin     <chr> "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", "EWR", ...
$ year       <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 20...
$ month      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
$ day        <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
$ hour       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 1...
$ temp       <dbl> 39.02, 39.02, 39.02, 39.92, 39.02, 37.94, 39.02, 39.92, ...
$ dewp       <dbl> 26.06, 26.96, 28.04, 28.04, 28.04, 28.04, 28.04, 28.04, ...
$ humid      <dbl> 59.37, 61.63, 64.43, 62.21, 64.43, 67.21, 64.43, 62.21, ...
$ wind_dir   <dbl> 270, 250, 240, 250, 260, 240, 240, 250, 260, 260, 260, 3...
$ wind_speed <dbl> 10.35702, 8.05546, 11.50780, 12.65858, 12.65858, 11.5078...
$ wind_gust  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
$ pressure   <dbl> 1012.0, 1012.3, 1012.5, 1012.2, 1011.9, 1012.4, 1012.2, ...
$ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, ...
$ time_hour  <dttm> 2013-01-01 01:00:00, 2013-01-01 02:00:00, 2013-01-01 03...
```

In [39]:

```
weather_NA <- sapply(weather, function(x) sum(is.na(x)))
weather_NA
```

origin	0
year	0
month	0
day	0
hour	0
temp	1
dewp	1
humid	1
wind_dir	460
wind_speed	4
wind_gust	20778
precip	0
pressure	2729
visib	0
time_hour	0

In [ ]:

```
na.rm = TRUE
```

```
In [84]: summary(weather)
```

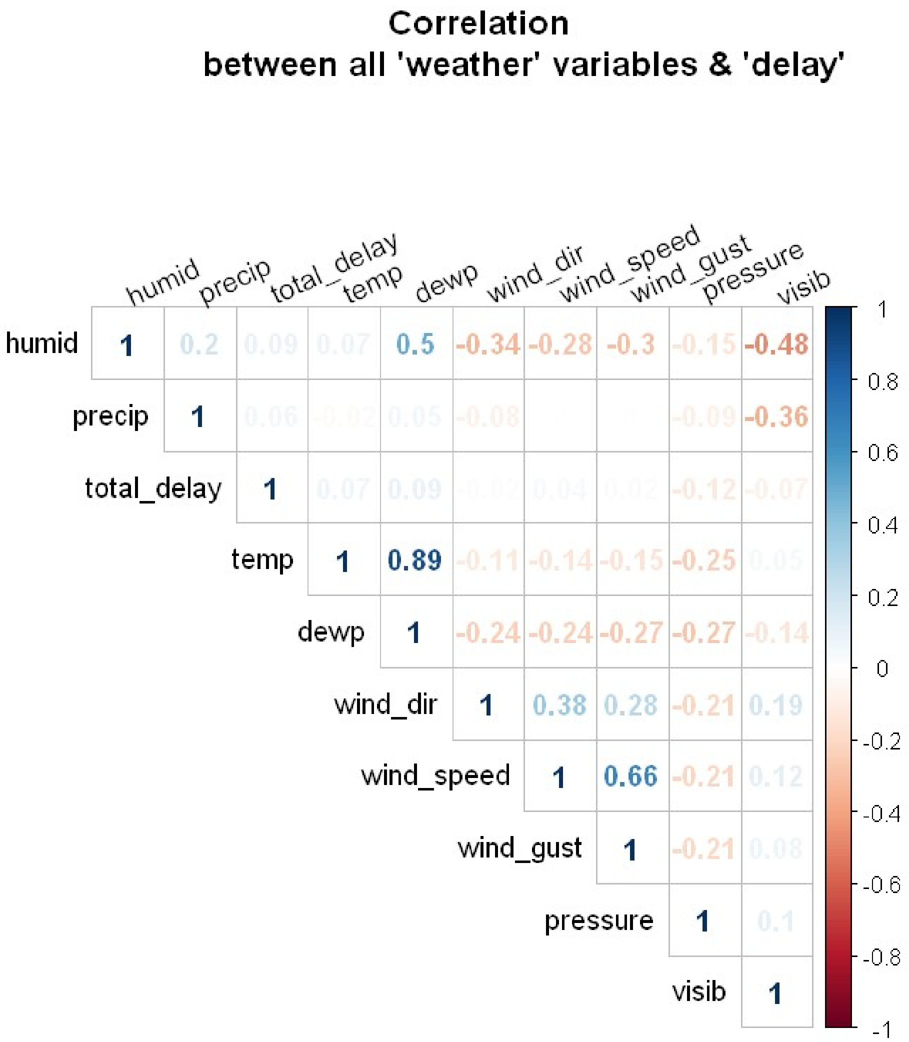
```
   origin               year          month             day       
Length:26115       Min.   :2013   Min.   : 1.000   Min.   : 1.00  
Class :character   1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00  
Mode  :character   Median :2013   Median : 7.000   Median :16.00  
                   Mean   :2013   Mean   : 6.504   Mean   :15.68  
                   3rd Qu.:2013   3rd Qu.: 9.000   3rd Qu.:23.00  
                   Max.   :2013   Max.   :12.000   Max.   :31.00  

     hour           temp              dewp            humid       
Min.   : 0.00   Min.   : 10.94   Min.   :-9.94   Min.   : 12.74  
1st Qu.: 6.00   1st Qu.: 39.92   1st Qu.:26.06   1st Qu.: 47.05  
Median :11.00   Median : 55.40   Median :42.08   Median : 61.79  
Mean   :11.49   Mean   : 55.26   Mean   :41.44   Mean   : 62.53  
3rd Qu.:17.00   3rd Qu.: 69.98   3rd Qu.:57.92   3rd Qu.: 78.79  
Max.   :23.00   Max.   :100.04   Max.   :78.08   Max.   :100.00  
                NA's   :1        NA's   :1       NA's   :1        
   wind_dir       wind_speed         wind_gust          precip        
Min.   :  0.0   Min.   :   0.000   Min.   : 0.000   Min.   :0.000000  
1st Qu.:120.0   1st Qu.:   6.905   1st Qu.: 0.000   1st Qu.:0.000000  
Median :220.0   Median :  10.357   Median : 0.000   Median :0.000000  
Mean   :199.8   Mean   :  10.518   Mean   : 5.209   Mean   :0.004469  
3rd Qu.:290.0   3rd Qu.:  13.809   3rd Qu.: 0.000   3rd Qu.:0.000000  
Max.   :360.0   Max.   :1048.361   Max.   :66.745   Max.   :1.210000  
NA's   :460     NA's   :4                                             
   pressure           visib         time_hour                  
Min.   : 983.8   Min.   : 0.000   Min.   :2013-01-01 01:00:00  
1st Qu.:1012.9   1st Qu.:10.000   1st Qu.:2013-04-01 21:30:00  
Median :1017.6   Median :10.000   Median :2013-07-01 14:00:00  
Mean   :1017.9   Mean   : 9.255   Mean   :2013-07-01 18:26:37  
3rd Qu.:1023.0   3rd Qu.:10.000   3rd Qu.:2013-09-30 13:00:00  
Max.   :1042.1   Max.   :10.000   Max.   :2013-12-30 18:00:00  
NA's   :2729                                                   
```

Sortez des statistiques sur les variables qui vous semblent pouvoir impacter le retard des avions, sur toute la base puis selon l’aéroport.

Un corrélogramme représente le graphique d’une matrice de corrélation. Le corrélogramme est très important pour mettre en évidence les variables les plus corrélées. Dans cet type de graphique, les coefficients de corrélation sont colorés en fonction de leur valeur. La matrice de corrélation peut être aussi réordonnée en fonction du degré de corrélation entre les variables. Le package corrplot de R est utilisé dans ce document.

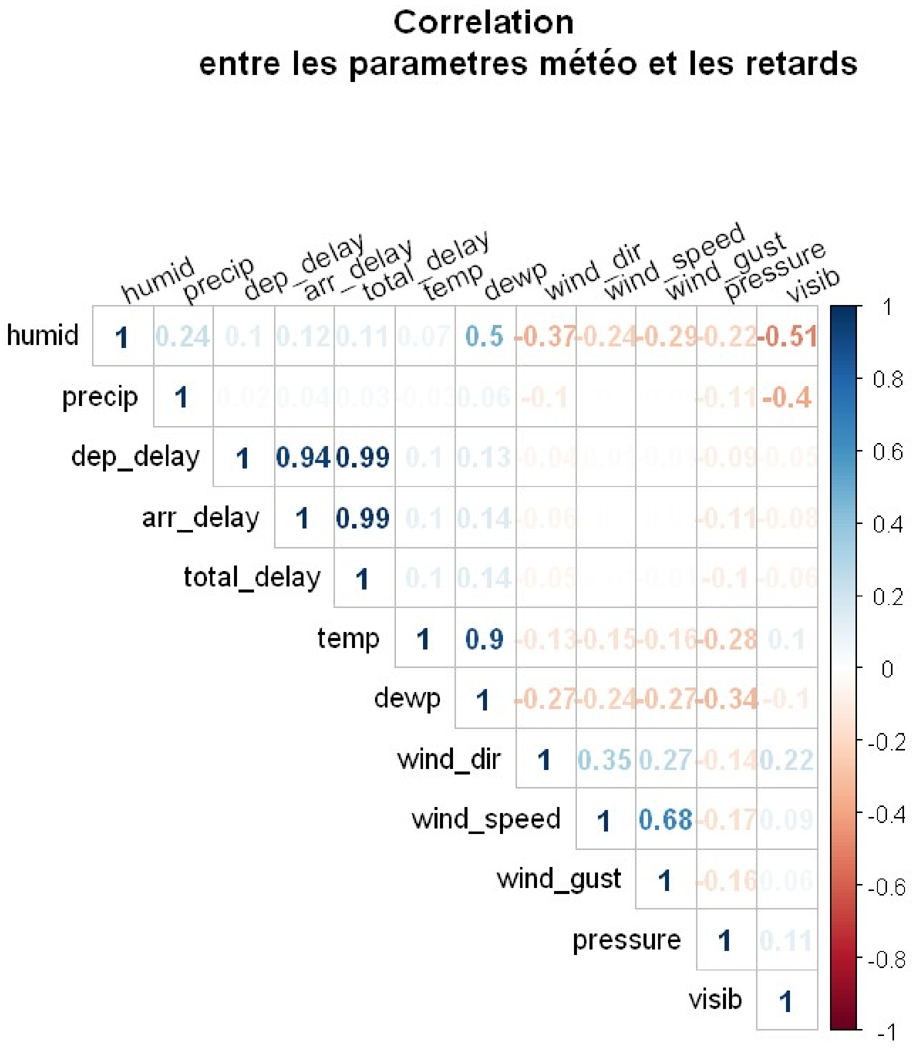
http://www.sthda.com/french/wiki/visualiser-une-matrice-de-correlation-par-un-correlogramme (http://www.sthda.com/french/wiki/visualiser-une-matrice-de-correlation-par-un-correlogramme)

In [52]:
```
flights$hour <- ifelse(flights$hour == 24, 0, flights$hour)
flights_weather <- inner_join(flights, weather, by = c("origin" = "origin", "time_h
our" = "time_hour"))
flights_weather$arr_delay <- ifelse(flights_weather$arr_delay >= 0,
                                    flights_weather$arr_delay, 0)
flights_weather$dep_delay <- ifelse(flights_weather$dep_delay >= 0,
                                    flights_weather$dep_delay, 0)

flights_weather$total_delay <- flights_weather$arr_delay + flights_weather$dep_dela
y
flights_weather$wind_gust[is.na(weather$wind_gust)] <- 0
cor_data <- select(flights_weather, total_delay, temp, dewp, humid,
                   wind_dir, wind_speed, wind_gust, precip, pressure, visib)
corrplot(cor(na.omit(cor_data)), method = "number", type = "upper", order="hclust",
         tl.srt = 25, tl.col = "Black", tl.cex = 1, title = "Correlation
         between all 'weather' variables & 'delay'", mar =c(0, 0, 4, 0) + 0.1)
```



In [83]:

```
flights$hour <- ifelse(flights$hour == 24, 0, flights$hour)
flights_weather <- inner_join(flights, weather, by = c("origin" = "origin", "time_hour" = "time_hour"))
flights_weather$arr_delay <- ifelse(flights_weather$arr_delay >= 0,
                                    flights_weather$arr_delay, 0)
flights_weather$dep_delay <- ifelse(flights_weather$dep_delay >= 0,
                                    flights_weather$dep_delay, 0)



flights_weather$total_delay <- flights_weather$arr_delay + flights_weather$dep_delay
weather$wind_gust[is.na(weather$wind_gust)] <- 0
cor_data <- select(flights_weather,arr_delay, dep_delay,total_delay, temp, dewp, humid,
                   wind_dir, wind_speed, wind_gust, precip, pressure, visib)
cor_data <- cor_data[!cor_data$dep_delay <= 10,]
cor_data <- cor_data[!cor_data$arr_delay <= 10,]
corrplot(cor(na.omit(cor_data)), method = "number", type = "upper", order="hclust",
         tl.srt = 25, tl.col = "Black", tl.cex = 1, title = "Correlation
         entre les parametres météo et les retards", mar =c(0, 0, 4, 0) + 0.1)
```



In [27]:

```
cor_data$dep_delay
```

```
Warning message:
"Unknown or uninitialised column: 'dep_delay'."

NULL
```

In [22]:

```
flights_weather$arr_delay
```

```
11  20  33  0  0  12  19  0  0  8  0  0  7  0  31  0  0  0  12  0  0  16  0  0  0  32  14  4
0  0  3  5  1  29  10  0  0  29  14  0  0  12  48  0  0  0  0  0  5  0  0  0  11  0  0  0  0
4  0  27  0  0  5  2  0  1  0  0  0  44  20  0  0  21  10  5  0  31  0  12  7  3  0  0  2
30  4  0  0  26  2  7  0  3  10  0  49  0  0  0  0  0  33  0  0  8  0  0  0  0  0  10  12  0
0  0  11  0  0  137  10  23  7  17  17  0  0  0  24  0  0  0  16  2  0  51  0  15  7  0  0
3  0  0  0  32  20  0  0  0  0  851  19  0  0  9  26  2  0  6  0  0  0  0  0  0  0  0  0  8
3  50  0  40  9  0  0  9  16  21  0  0  4  15  28  0  0  14  27  10  0  0  2  13  0  0  30
9  0  13  0  0  42  15  0  0  3  7  0  6  2  0  0  13  12  39  11  0  123  0  0  0  39  5  7
0  0  11  5  0  2  0  0  26  0  11  11  32  0  0  0  15  0  11  21  0  19  1  0  0  0  12  4
0  0  0  0  0  0  0  13  0  23  0  0  0  2  3  145  78  5  34  4  0  0  0  16  0  0  0  3  0
3  0  0  0  5  11  8  0  9  29  22  0  1  0  0  10  7  0  38  0  0  43  3  5  0  0  34  4  5
81  0  0  4  0  19  93  0  3  12  5  2  0  0  0  0  0  0  43  1  0  0  37  0  8  0  23  4
73  18  21  26  11  0  0  24  0  0  46  0  22  27  11  19  27  103  0  84  6  66  14  0  7
0  26  0  18  0  0  65  15  83  26  0  5  11  0  20  0  15  0  0  0  19  0  4  11  0  0  0
22  0  0  0  0  1  10  0  17  17  0  0  3  0  0  3  0  127  0  18  0  0  3  26  0  11  0  0
11  5  52  5  60  0  1  7  6  3  3  27  125  <NA>  24  0  27  0  44  <NA>  6  46  0  0  34
47  19  20  19  0  0  0  35  115  0  8  0  5  8  91  136  0  13  12  6  0  18  0  16  10  0
0  4  65  123  7  0  0  15  0  12  12  9  28  6  67  4  20  78  72  28  0  0  0  18  29  20
17  23  34  3  4  15  2  80  0  59  0  67  0  13  0  0  23  14  21  20  0  39  96  9  0  25
16  26  56  0  35  0  0  0  40  3  0  0  0  41  22  0  11  16  16  44  9  0  6  32  14  33
68  0  6  0  0  0  61  16  10  0  3  16  3  24  4  0  3  107  0  14  16  0  123  46  0  3  0
68  <NA>  0  66  0  27  0  10  34  0  4  65  21  17  51  0  0  0  37  8  0  0  0  0  37
138  20  0  17  <NA>  0  0  14  116  0  338  8  18  11  3  26  0  0  3  14  2  0  9  32  6
8  0  0  7  32  94  12  25  0  263  8  2  0  0  52  0  78  2  0  1  9  0  40  0  3  127  0
0  0  0  0  0  0  0  10  0  2  0  40  19  15  0  0  0  0  0  0  18  0  54  5  0  21  43  0
0  81  151  40  25  166  <NA>  0  12  0  174  16  14  44  <NA>  0  6  0  0  0  11  25  3
75  0  0  0  222  2  83  0  123  0  0  5  <NA>  50  47  0  0  0  10  34  91  0  7  24  50
45  0  0  12  10  9  0  0  34  3  0  0  45  0  0  4  0  44  61  0  5  17  25  5  9  17  0  0
0  0  0  0  55  17  250  0  142  0  43  22  0  3  0  28  0  0  2  0  246  21  23  28  0  73
127  8  49  0  0  49  23  35  9  191  69  73  33  456  0  0  0  <NA>  <NA>  <NA>  <NA>
36  154  13  0  0  12  0  0  0  11  21  0  0  1  0  9  23  0  0  0  25  0  4  0  12  0  12
0  0  0  0  1  5  0  15  19  10  31  11  0  44  20  0  0  43  3  1  8  17  3  15  11  17  0
27  6  0  0  1  0  17  0  23  0  6  0  0  9  33  0  0  13  0  0  15  0  0  0  8  0  0  0  0
0  6  34  0  0  37  1  2  18  46  26  0  0  29  0  0  0  4  12  0  1  0  0  90  0  0  0  0
9  6  8  0  0  5  24  7  4  0  20  0  6  0  0  2  10  12  0  1  10  32  0  0  28  0  0  0
13  75  0  6  0  0  0  19  0  0  0  56  0  0  0  4  4  4  2  0  4  0  0  0  0  0  142  8  6
39  1  24  12  0  31  0  23  0  0  0  0  48  18  0  0  0  0  9  0  1  0  171  21  41  58  38
13  0  21  1  70  0  5  12  0  3  15  0  0  65  0  5  22  0  0  0  10  0  0  0  48  27  0
11  0  0  2  0  0  0  <NA>  21  0  9  0  16  37  8  9  0  57  0  8  173  10  16  0  0  84
62  29  99  0  3  15  2  2  34  2  0  52  23  22  97  12  0  0  72  0  0  0  0  0  23  0  0
9  0  46  85  0  0  14  14  0  5  0  0  14  10  0  12  0  0  4  0  3  0  0  0  63  2  17  19
7  7  4  7  0  0  73  0  0  28  0  0  0  98  2  0  0  0  26  0  15  0  0  5  0  0  3  21  17
0  0  54  10  8  46  <NA>  0  0  8  0  0  0  0  24  0  0  3  0  11  0  0  0  0  0  0  15  0
0  0  0  1  33  0  12  43  8  0  10  1  37  4  0  5  46  11  6  0  0  0  0  15  102  6  31
0  0  6  4  0  0  47  47  8  0  16  0  25  207  0  5  0  0  0  0  8  0  5  8  24  0  2  6  2
0  0  0  7  0  2  0  13  4  21  0  18  8  9  0  0  0  0  83  0  288  0  45  0  5  35  3  12
9  96  44  0  107  0  16  0  47  3  0  0  38  0  25  8  0  6  30  7  0  0  0  0  323  0  4
60  39  19  20  18  14  0  0  51  1  6  0  85  11  8  0  0  0  0  45  0  0  0  32  136  3
15  4  34  28  1  23  14  0  0  14  1  2  33  0  40  0  0  0  0  32  14  0  0  0  6  3  11
41  27  13  3  0  32  65  5  4  0  3  20  13  0  0  0  0  14  26  46  119  0  9  0  0  59  2
41  43  9  17  38  0  8  6  17  30  15  0  9  35  130  25  45  0  1  8  2  0  0  109  24  0
0  20  13  13  26  63  3  69  12  0  29  35  0  8  81  27  20  4  10  23  1  368  67  15
14  16  16  9  18  15  0  20  48  9  0  14  20  102  0  0  6  0  23  0  0  0  3  0  47  0  1
22  96  4  10  24  0  0  9  4  0  0  0  0  0  7  36  25  18  22  5  70  16  0  5  7  122  0
7  12  82  0  36  38  28  6  2  0  7  21  0  11  0  2  0  82  131  0  6  0  19  0  0  69  4
```

In [17]:

```
flights_weather
```

year.x	month.x	day.x	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	...
2013	1	1	517	515	2	830	819	11	UA	...
2013	1	1	533	529	4	850	830	20	UA	...
2013	1	1	542	540	2	923	850	33	AA	...
2013	1	1	544	545	0	1004	1022	0	B6	...
2013	1	1	554	600	0	812	837	0	DL	...
2013	1	1	554	558	0	740	728	12	UA	...
2013	1	1	555	600	0	913	854	19	B6	...
2013	1	1	557	600	0	709	723	0	EV	...
2013	1	1	557	600	0	838	846	0	B6	...
2013	1	1	558	600	0	753	745	8	AA	...
2013	1	1	558	600	0	849	851	0	B6	...
2013	1	1	558	600	0	853	856	0	B6	...
2013	1	1	558	600	0	924	917	7	UA	...
2013	1	1	558	600	0	923	937	0	UA	...
2013	1	1	559	600	0	941	910	31	AA	...
2013	1	1	559	559	0	702	706	0	B6	...
2013	1	1	559	600	0	854	902	0	UA	...
2013	1	1	600	600	0	851	858	0	B6	...
2013	1	1	600	600	0	837	825	12	MQ	...
2013	1	1	601	600	1	844	850	0	B6	...
2013	1	1	602	610	0	812	820	0	DL	...
2013	1	1	602	605	0	821	805	16	MQ	...
2013	1	1	606	610	0	858	910	0	AA	...
2013	1	1	606	610	0	837	845	0	DL	...
2013	1	1	607	607	0	858	915	0	UA	...
2013	1	1	608	600	8	807	735	32	MQ	...
2013	1	1	611	600	11	945	931	14	UA	...
2013	1	1	613	610	3	925	921	4	B6	...
2013	1	1	615	615	0	1039	1100	0	B6	...
2013	1	1	615	615	0	833	842	0	DL	...
...	...	...	...	...	...	...	...	...	...	
2013	9	30	2123	2125	0	2223	2247	0	EV	...
2013	9	30	2127	2129	0	2314	2323	0	EV	...
2013	9	30	2128	2130	0	2328	2359	0	B6	...
2013	9	30	2129	2059	30	2230	2232	0	EV	...
2013	9	30	2131	2140	0	2225	2255	0	MQ	...
2013	9	30	2140	2140	0	10	40	0	AA	...
2013	9	30	2142	2129	13	2250	2239	11	EV	...
2013	9	30	2145	2145	0	115	140	0	B6	...
2013	9	30	2147	2137	10	30	27	3	B6	...
2013	9	30	2149	2156	0	2245	2308	0	UA	...

In [15]:

```
weather$wind_gust
```

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 20.71404 0 25.31716 0 0 26.46794 25.31716 0
25.31716 24.16638 0 0 0 0 0 0 0 0 0 25.31716 23.0156 25.31716 26.46794
20.71404 0 18.41248 0 0 16.11092 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 19.56326 27.61872 27.61872 25.31716
31.07106 28.7695 27.61872 25.31716 0 0 19.56326 0 0 0 0 0 0 0 23.0156 0 0 0
0 20.71404 0 21.86482 26.46794 0 19.56326 16.11092 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 19.56326 18.41248 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 23.0156 21.86482 20.71404 24.16638 0 20.71404 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 25.31716 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 20.71404 20.71404 0 0 0 0 0 0 0 0 0 0 0 0 20.71404
17.2617 0 0 26.46794 0 23.0156 20.71404 18.41248 19.56326 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 18.41248 0 0 0 0 0 0 17.2617 20.71404
0 27.61872 0 18.41248 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 17.2617 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 32.22184 24.16638 26.46794 0 21.86482 0 24.16638 16.11092 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 24.16638 0 28.7695 0 27.61872 27.61872 25.31716 25.31716 0
0 23.0156 24.16638 0 26.46794 23.0156 25.31716 0 0 0 0 0 0 0 0 0 24.16638
29.92028 41.42808 39.12652 34.5234 37.97574 31.07106 32.22184 34.5234 29.92028
23.0156 26.46794 0 0 18.41248 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 19.56326 18.41248 26.46794 25.31716 0 21.86482 0 0 29.92028 35.67418
28.7695 32.22184 33.37262 26.46794 27.61872 29.92028 26.46794 0 24.16638 26.46794
27.61872 21.86482 23.0156 21.86482 21.86482 0 18.41248 0 0 0 0 0 0 0 20.71404
17.2617 0 0 21.86482 20.71404 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
21.86482 32.22184 34.5234 40.2773 35.67418 29.92028 28.7695 26.46794 32.22184
27.61872 35.67418 27.61872 35.67418 20.71404 23.0156 20.71404 0 23.0156 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 18.41248 0
24.16638 0 24.16638 21.86482 28.7695 27.61872 0 24.16638 0 0 0 0 0 0 0 0 0 0
0 0 0 0 20.71404 0 0 0 0 17.2617 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 19.56326 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 18.41248 0 0 0 0 0
0 0 0 33.37262 36.82496 39.12652 47.18198 44.88042 58.68978 27.61872 55.23744
37.97574 48.33276 51.7851 41.42808 43.72964 42.57886 42.57886 42.57886 44.88042
40.2773 43.72964 29.92028 31.07106 26.46794 36.82496 35.67418 0 23.0156 0 0 0 0
20.71404 0 0 0 23.0156 27.61872 32.22184 27.61872 31.07106 33.37262 0 0 26.46794
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 19.56326 0 19.56326 0 20.71404 0
19.56326 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17.2617 0 0 0
0 0 0 0 0 0 26.46794 20.71404 0 0 0 0 23.0156 25.31716 26.46794 28.7695
26.46794 32.22184 26.46794 26.46794 28.7695 17.2617 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 19.56326
21.86482 21.86482 0 21.86482 24.16638 25.31716 18.41248 0 0 21.86482 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17.2617 0 21.86482 0
0 0 0 0 0 0 27.61872 29.92028 32.22184 33.37262 0 0 0 28.7695 0 27.61872
25.31716 26.46794 0 0 0 0 0 0 24.16638 0 0 27.61872 28.7695 28.7695 35.67418
26.46794 23.0156 33.37262 29.92028 29.92028 26.46794 19.56326 26.46794 23.0156 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 25.31716 0 0 0 0 20.71404 0
0 23.0156 0 31.07106 31.07106 24.16638 19.56326 0 26.46794 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 17.2617
19.56326 25.31716 0 0 0 0 0 0 0 0 28.7695 0 0 0 0 0 0 0 0 0 0 0 0
18.41248 18.41248 0 0 0 0 0 0 0 27.61872 26.46794 24.16638 27.61872 21.86482
25.31716 26.46794 0 25.31716 27.61872 27.61872 32.22184 34.5234 36.82496 37.97574
```

In [ ]:

```
# PerformanceAnalytics
# ! ça plante
chart.Correlation(cor_data, histogram=TRUE, pch=19)
```

In [7]:

```
flights_jfk <-
  nycflights13::flights %>%
  filter(origin == "JFK") %>%
  mutate(hh = round(sched_dep_time / 100, 0) - 1) %>%
  mutate(ymd = lubridate::ymd(sprintf("%04.0f-%02.0f-%02.0f", year, month, day))) %
>%
  mutate(wd = lubridate::wday(ymd, label = TRUE))
```

In [8]:

```
flights_jfk
```

year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	...	orig
2013	1	1	542	540	2	923	850	33	AA	...	JF
2013	1	1	544	545	-1	1004	1022	-18	B6	...	JF
2013	1	1	557	600	-3	838	846	-8	B6	...	JF
2013	1	1	558	600	-2	849	851	-2	B6	...	JF
2013	1	1	558	600	-2	853	856	-3	B6	...	JF
2013	1	1	558	600	-2	924	917	7	UA	...	JF
2013	1	1	559	559	0	702	706	-4	B6	...	JF
2013	1	1	606	610	-4	837	845	-8	DL	...	JF
2013	1	1	611	600	11	945	931	14	UA	...	JF
2013	1	1	613	610	3	925	921	4	B6	...	JF
2013	1	1	615	615	0	1039	1100	-21	B6	...	JF
2013	1	1	627	630	-3	1018	1018	0	US	...	JF
2013	1	1	628	630	-2	1137	1140	-3	AA	...	JF
2013	1	1	639	640	-1	739	749	-10	B6	...	JF
2013	1	1	645	647	-2	815	810	5	B6	...	JF
2013	1	1	651	655	-4	936	942	-6	B6	...	JF
2013	1	1	652	655	-3	932	921	11	B6	...	JF
2013	1	1	655	655	0	1021	1030	-9	DL	...	JF
2013	1	1	655	700	-5	1037	1045	-8	DL	...	JF
2013	1	1	656	659	-3	949	959	-10	AA	...	JF
2013	1	1	658	700	-2	1027	1025	2	VX	...	JF
2013	1	1	659	700	-1	1008	1007	1	B6	...	JF
2013	1	1	702	700	2	1058	1014	44	B6	...	JF
2013	1	1	711	715	-4	1151	1206	-15	B6	...	JF
2013	1	1	712	715	-3	1023	1035	-12	AA	...	JF
2013	1	1	719	721	-2	1017	1012	5	B6	...	JF
2013	1	1	729	730	-1	1049	1115	-26	VX	...	JF

In [ ]:

day

In [ ]:

In [ ]:

In [76]:
```
flights %>%
  select(day, arr_delay, dep_delay) %>%
  group_by(day) %>%
  summarise(avg_delay =  mean(arr_delay, na.rm = TRUE) +
                mean(dep_delay, na.rm = TRUE)) %>%
  arrange(-avg_delay)
```

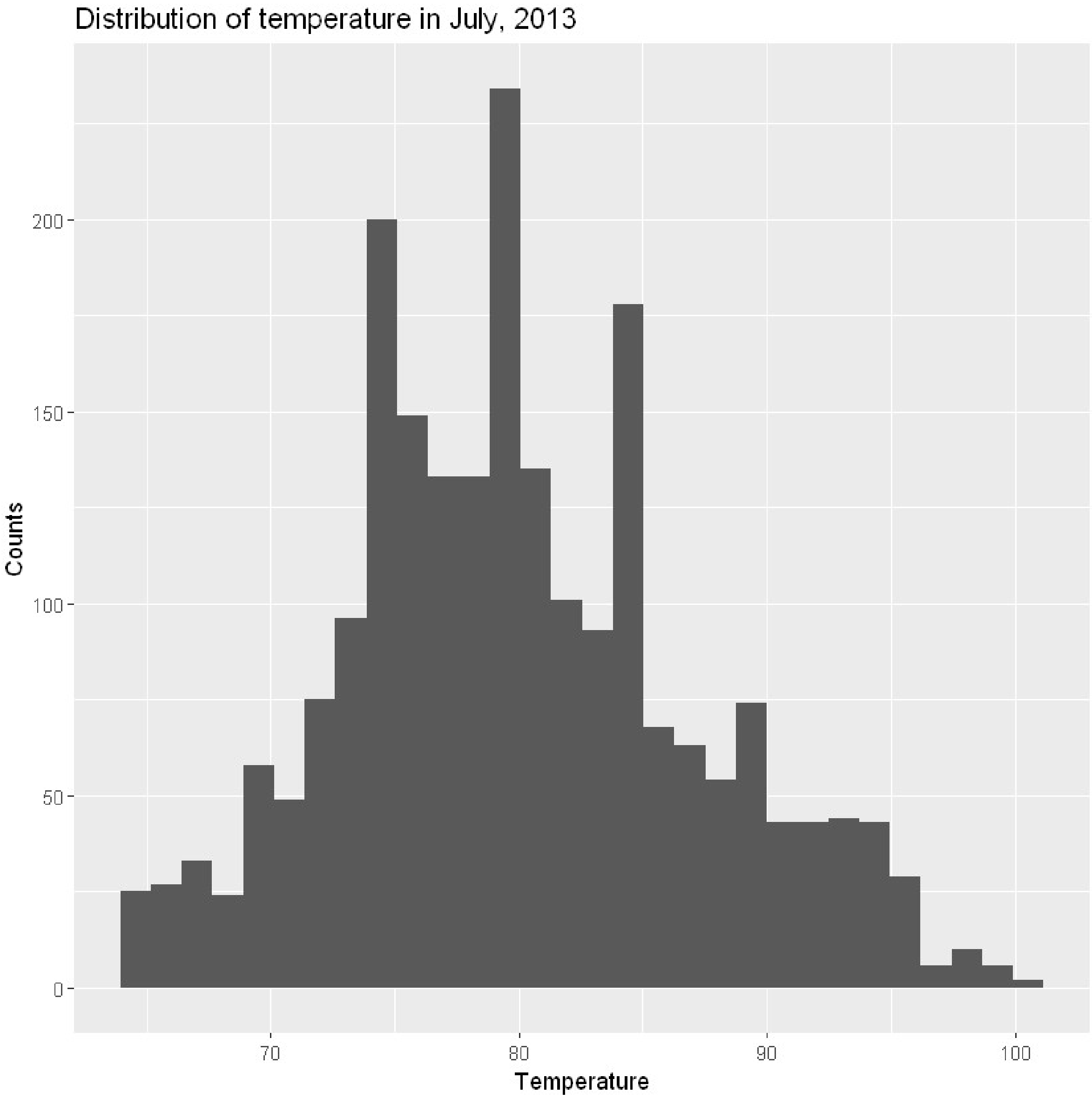
day	avg_delay
8	40.832950
22	36.116989
23	34.138017
10	33.036514
12	26.316738
11	26.309286
24	25.737473
19	25.403336
18	24.965716
25	24.524466
28	23.926781
7	23.896900
17	23.605112
9	23.592193
13	22.901398
1	21.536355
2	20.883606
27	15.415182
3	15.281214
26	13.404101
31	12.865746
16	12.475657
14	12.194987
21	11.896045
30	10.716131
20	9.940030
29	8.328306
5	8.313330
15	5.542129
6	5.241340
4	4.006931

temp

In [21]:

```
weather %>%
      filter(month == 7) %>%
      ggplot(aes(x = temp)) +
      geom_histogram() +
      labs(x = "Temperature", y = "Counts", title = "Distribution of temperature in
July, 2013")
```

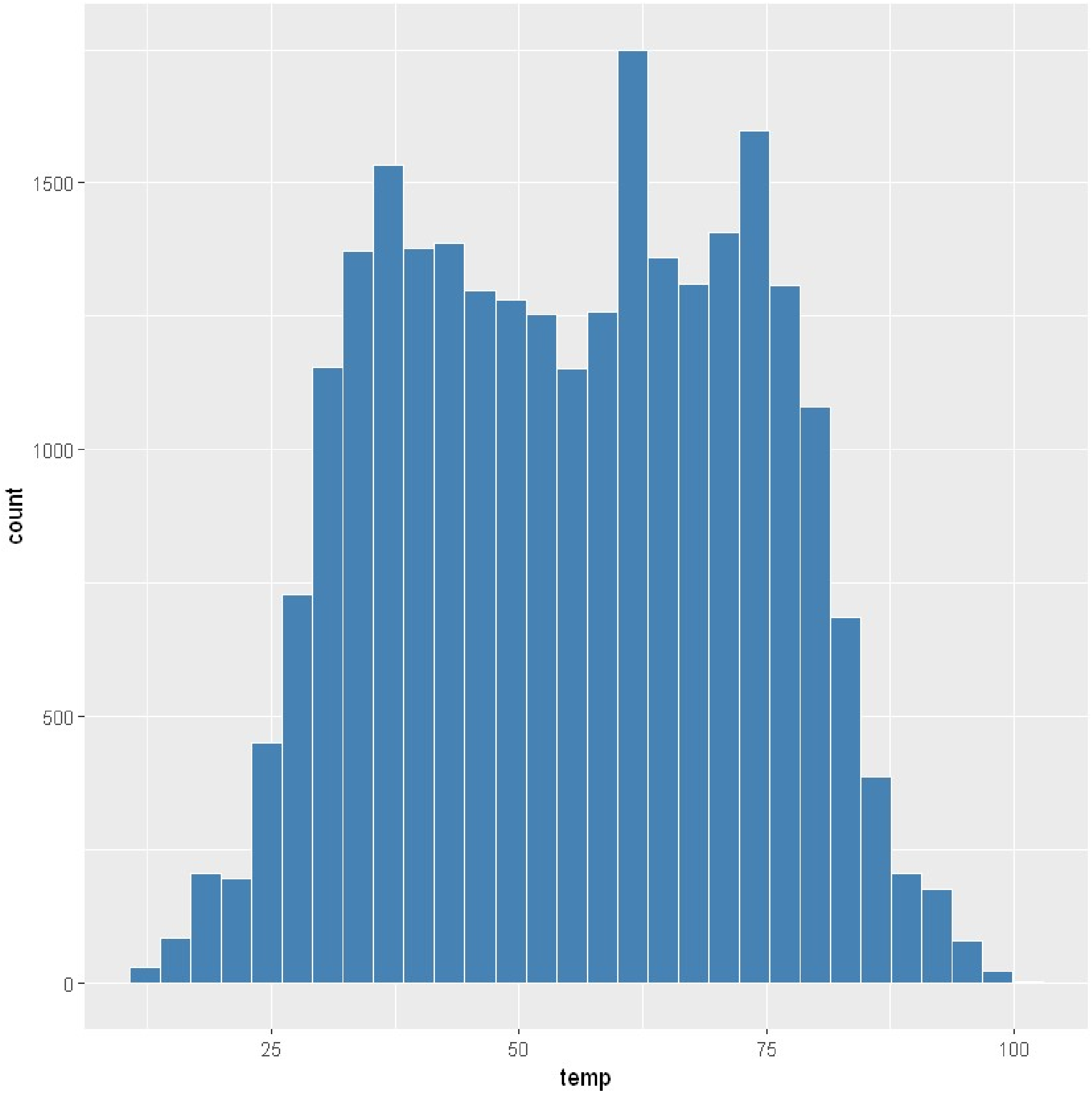
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In [9]:

```
ggplot(data = weather, mapping = aes(x = temp)) +
  geom_histogram(color = "white", fill = "steelblue")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
Warning message:
"Removed 1 rows containing non-finite values (stat_bin)."
```



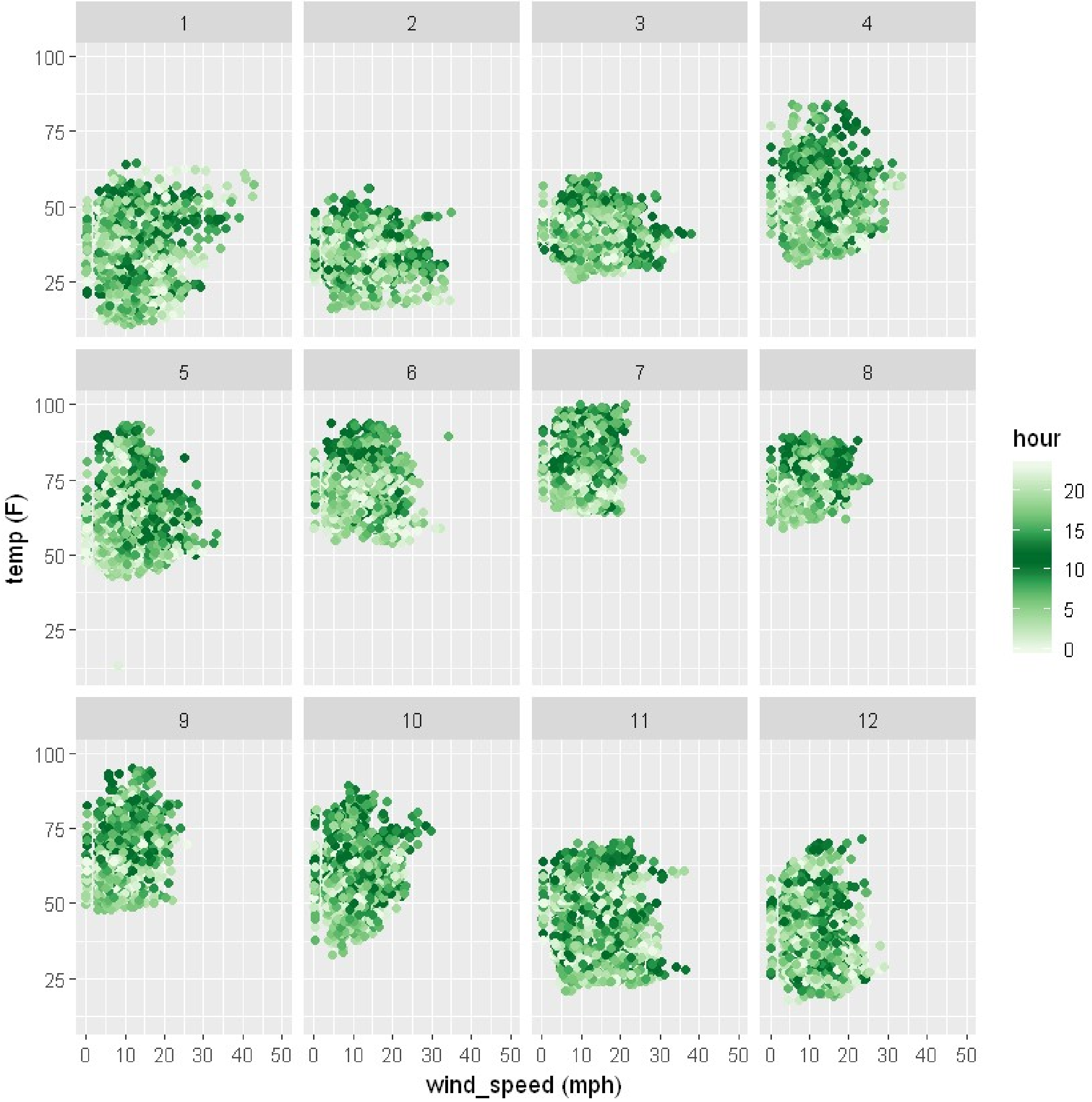
In [60]:

```
weather %>%
  group_by(month) %>%
  summarize(IQR = IQR(temp, na.rm=TRUE)) %>%
  arrange(desc(IQR))
```

month	IQR
11	16.02
12	14.04
1	13.77
9	12.06
4	12.06
5	11.88
6	10.98
10	10.98
2	10.08
7	9.18
3	9.00
8	7.02

In [83]:
```
mypal <- RColorBrewer::brewer.pal(6, "Greens")
mypal <- c(mypal,  rev(mypal))

ggplot(weather, aes(x=wind_speed, y=temp,  col=hour)) + geom_jitter() +xlim(0, 50)+
facet_wrap(~ month)+scale_color_gradientn(colors=mypal)+xlab("wind_speed (mph)")+yl
ab("temp (F)")
```

```
Warning message:
"Removed 660 rows containing missing values (geom_point)."
```



In [ ]:

Point de rosée : dewp (dewpoint in F)

[https://fr.wikipedia.org/wiki/Point_de_ros%C3%A9e (https://fr.wikipedia.org/wiki/Point_de_ros%C3%A9e)](https://fr.wikipedia.org/wiki/Point_de_ros%C3%A9e)

In [11]:
```
sfweather_hr2 <- sfweather_hr %>%
  select(FL_DATE, DEWP, HUMID) %>%
  filter(!is.na(DEWP), !is.na(HUMID)) %>%
  gather(key = grp, val = statistics, DEWP, HUMID)

ggplot(sfweather_hr2, aes(x = FL_DATE, y = statistics, color = grp)) +
  geom_line() +
  geom_smooth()
```

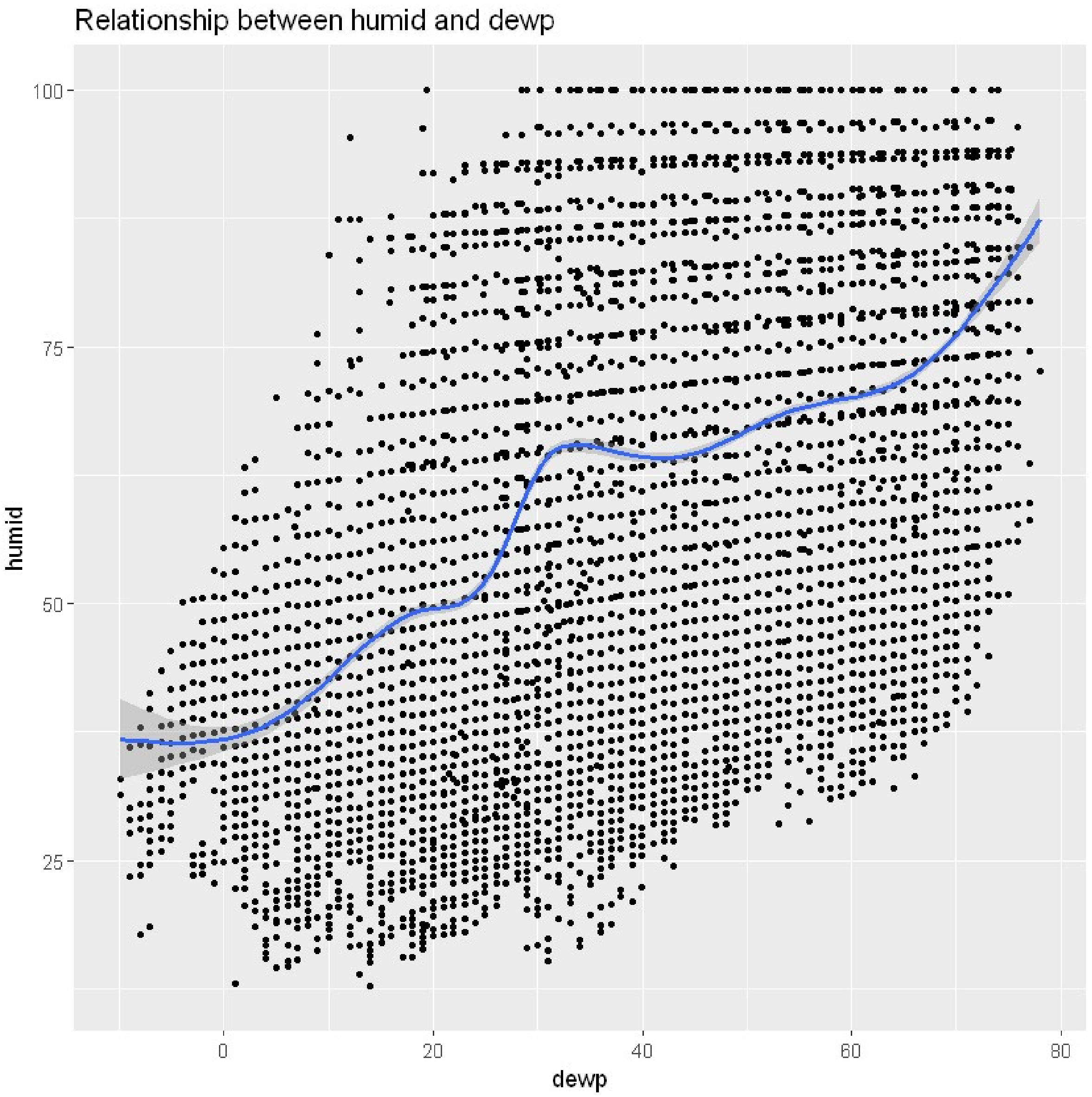
```
Error in eval(lhs, parent, parent): objet 'sfweather_hr' introuvable
Traceback:

1. sfweather_hr %>% select(FL_DATE, DEWP, HUMID) %>% filter(!is.na(DEWP),
 .     !is.na(HUMID)) %>% gather(key = grp, val = statistics, DEWP,
 .     HUMID)
2. eval(lhs, parent, parent)
3. eval(lhs, parent, parent)
```

humid

In [22]:

```
weather%>%
      ggplot(aes(x = dewp, y = humid)) +
      geom_point(size = 1) +
      geom_smooth() +
      labs(title = "Relationship between humid and dewp")
```

```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
Warning message:
"Removed 1 rows containing non-finite values (stat_smooth)."Warning message:
"Removed 1 rows containing missing values (geom_point)."
```



wind_dir

In [ ]:

```
wind_data <- select(flights_weather, total_delay, wind_speed)
#wind_data <- wind_data[is.na(wind_data)] <- 0
wind_data <- wind_data[!wind_data$wind_speed <= 0,]
wind_data <- wind_data[!wind_data$total_delay <= 0,]
wind_data <- wind_data[!is.na(wind_data$total_delay),]
```

In [ ]:

```
w3 %>%
ggplot(aes(x = wind_mean, y = total_delay)) +
geom_point(color = "blue") +
geom_smooth()
```

In [ ]:

```
```

wind_speed

In [5]:

```
wind_data <- select(flights_weather, total_delay, wind_speed)
#wind_data <- wind_data[is.na(wind_data)] <- 0
wind_data <- wind_data[!wind_data$wind_speed <= 0,]
wind_data <- wind_data[!wind_data$total_delay <= 0,]
wind_data <- wind_data[!is.na(wind_data$total_delay),]
#wind_data$total_delay <- wind_data$total_delay[is.na(wind_data$total_delay)] <- 0
```

In [81]:

```
wind_data <- wind_data %>%
    group_by(total_delay)
```

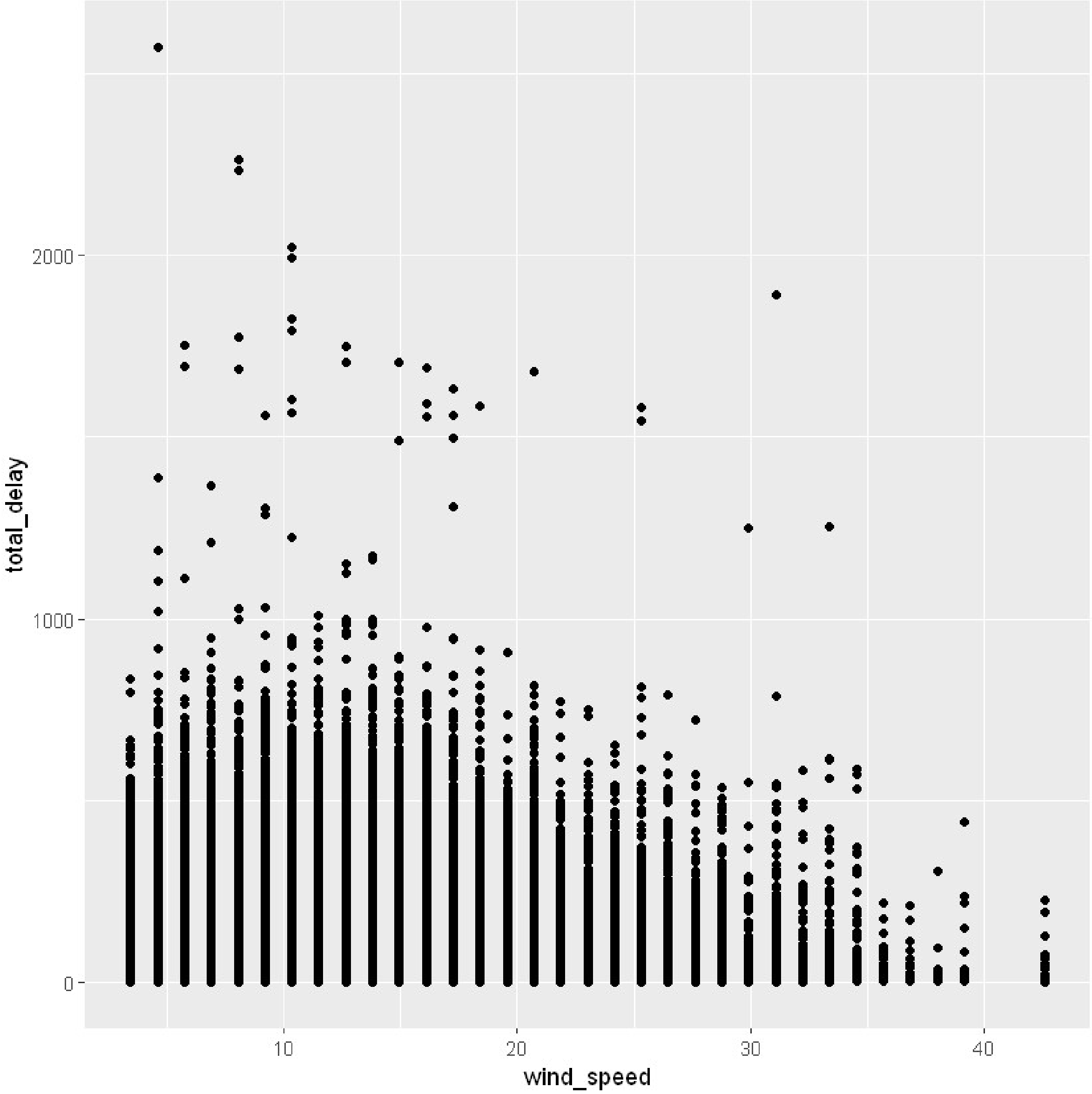
In [82]:

```
wind_data
```

total_delay	wind_speed
13	12.65858
24	14.96014
35	14.96014
12	12.65858
19	11.50780
8	16.11092
7	13.80936
31	16.11092
12	16.11092
1	11.50780
16	16.11092
40	11.50780
25	13.80936
7	13.80936
3	11.50780
18	16.11092
1	16.11092
29	11.50780
10	16.11092
29	11.50780
14	16.11092
36	11.50780
48	16.11092
8	11.50780
5	13.80936
1	16.11092
1	11.50780
11	13.80936
4	14.96014
27	14.96014
...	...
35	5.75390
73	4.60312
9	8.05546
26	6.90468
118	5.75390
56	5.75390
3	3.45234
7	6.90468
6	3.45234
5	8.05546

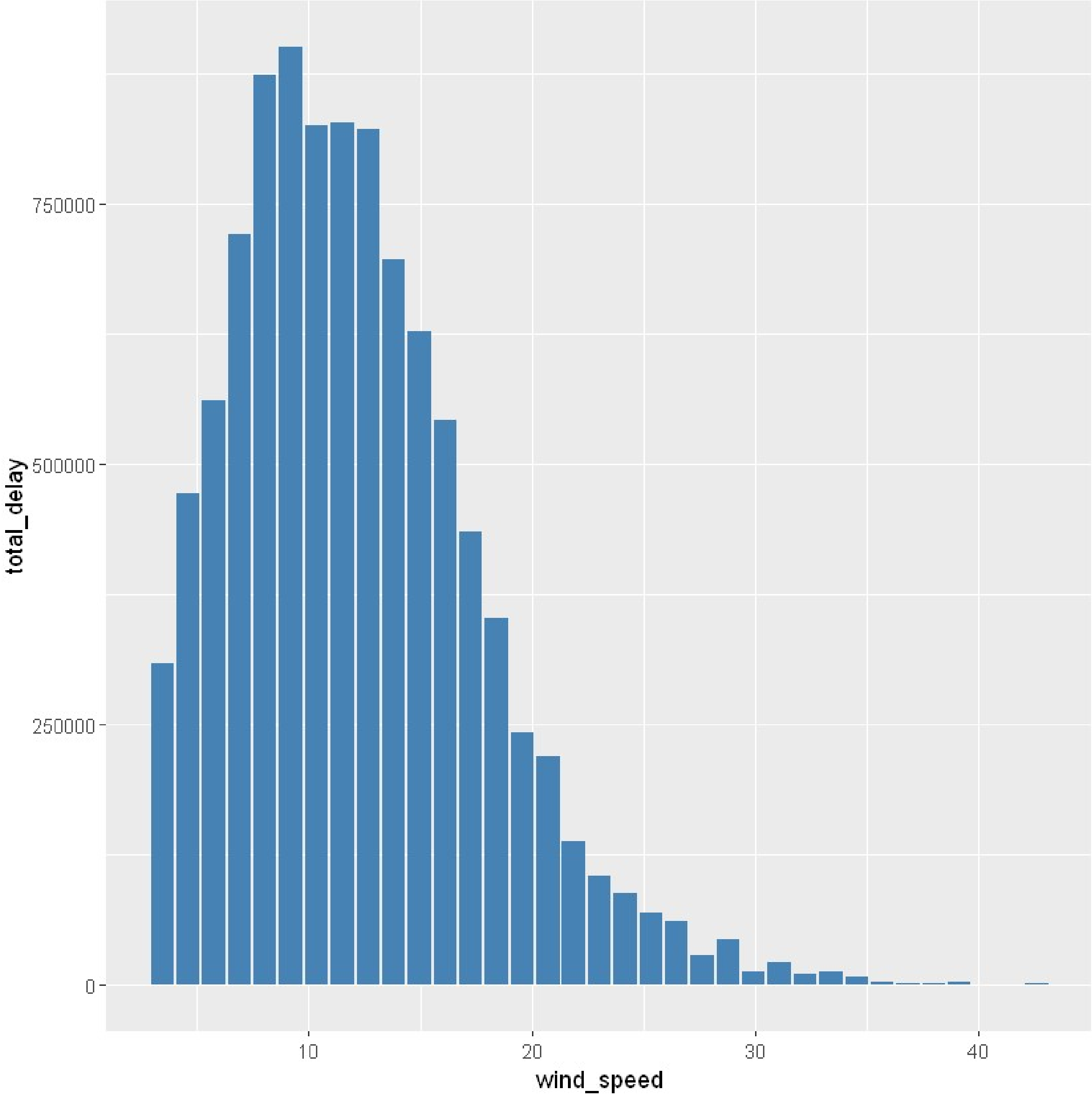
In [77]:

```
wind_data %>%
    group_by(total_delay)  %>%
ggplot(aes(x = wind_speed, y = total_delay)) +
geom_point()
```



In [90]:

```
wind_data %>%
    group_by(total_delay)  %>%
ggplot(aes(x = wind_speed, y = total_delay)) +
geom_bar(stat="identity", fill="steelblue")
```



In [18]:

```
w3 <- wind_data %>%
    group_by(total_delay)  %>%
    summarise(wind_mean = mean(wind_speed, na.rm = TRUE))
```

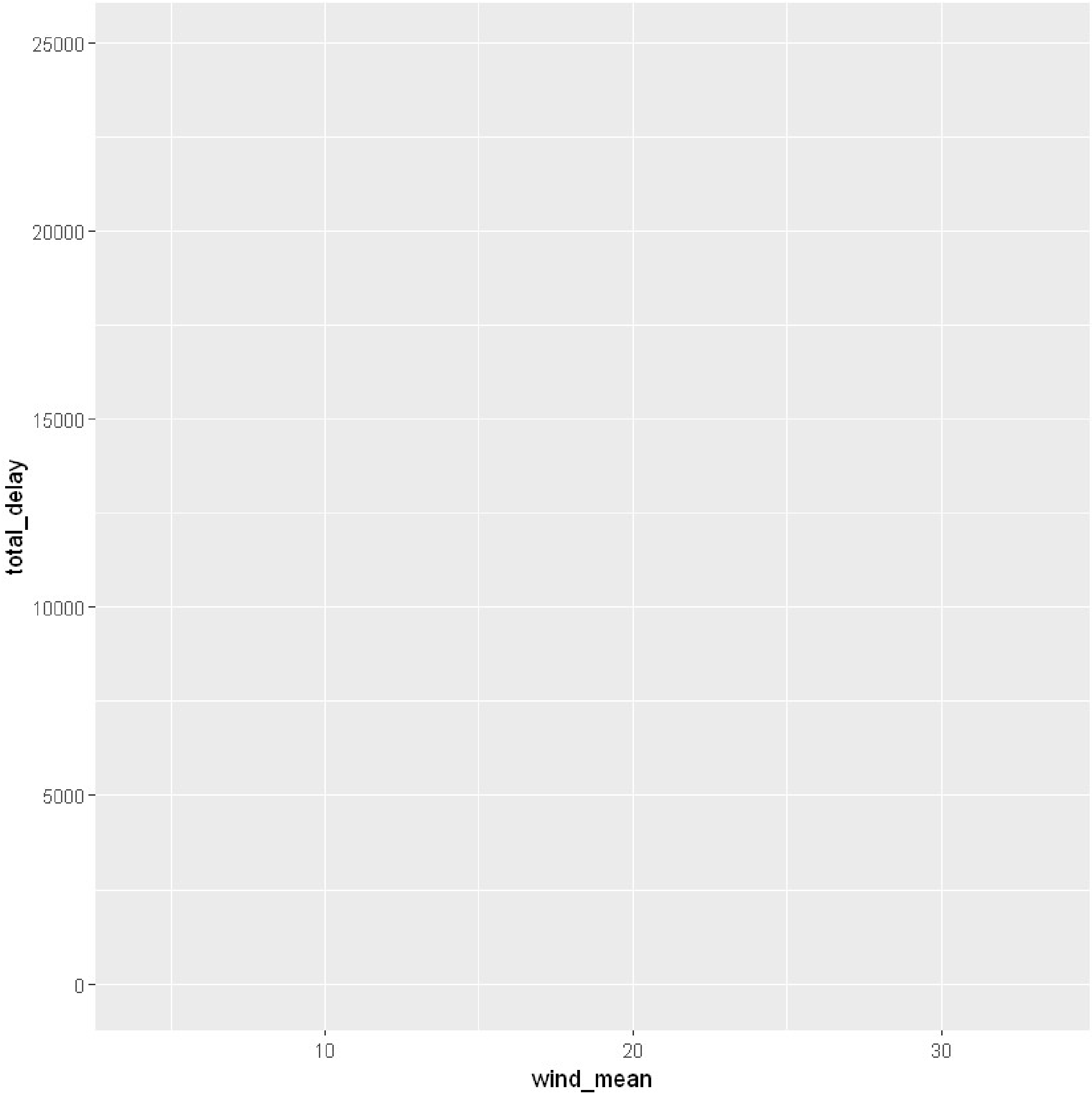
In [19]:

```
w3
```

total_delay	wind_mean
1	11.36362
2	11.45151
3	11.49513
4	11.39506
5	11.38491
6	11.49432
7	11.46044
8	11.66297
9	11.59318
10	11.54151
11	11.54235
12	11.62102
13	11.70702
14	11.64574
15	11.72257
16	11.78286
17	11.59848
18	11.76501
19	11.68409
20	11.68067
21	11.84066
22	11.76579
23	11.75868
24	11.79269
25	11.82421
26	11.98312
27	11.75278
28	11.75574
29	11.78883
30	11.89908
...	...
1368	6.90468
1390	4.60312
1491	14.96014
1497	17.26170
1544	25.31716
1555	16.11092
1559	13.23397
1567	10.35702
1580	25.31716
1584	18.41248

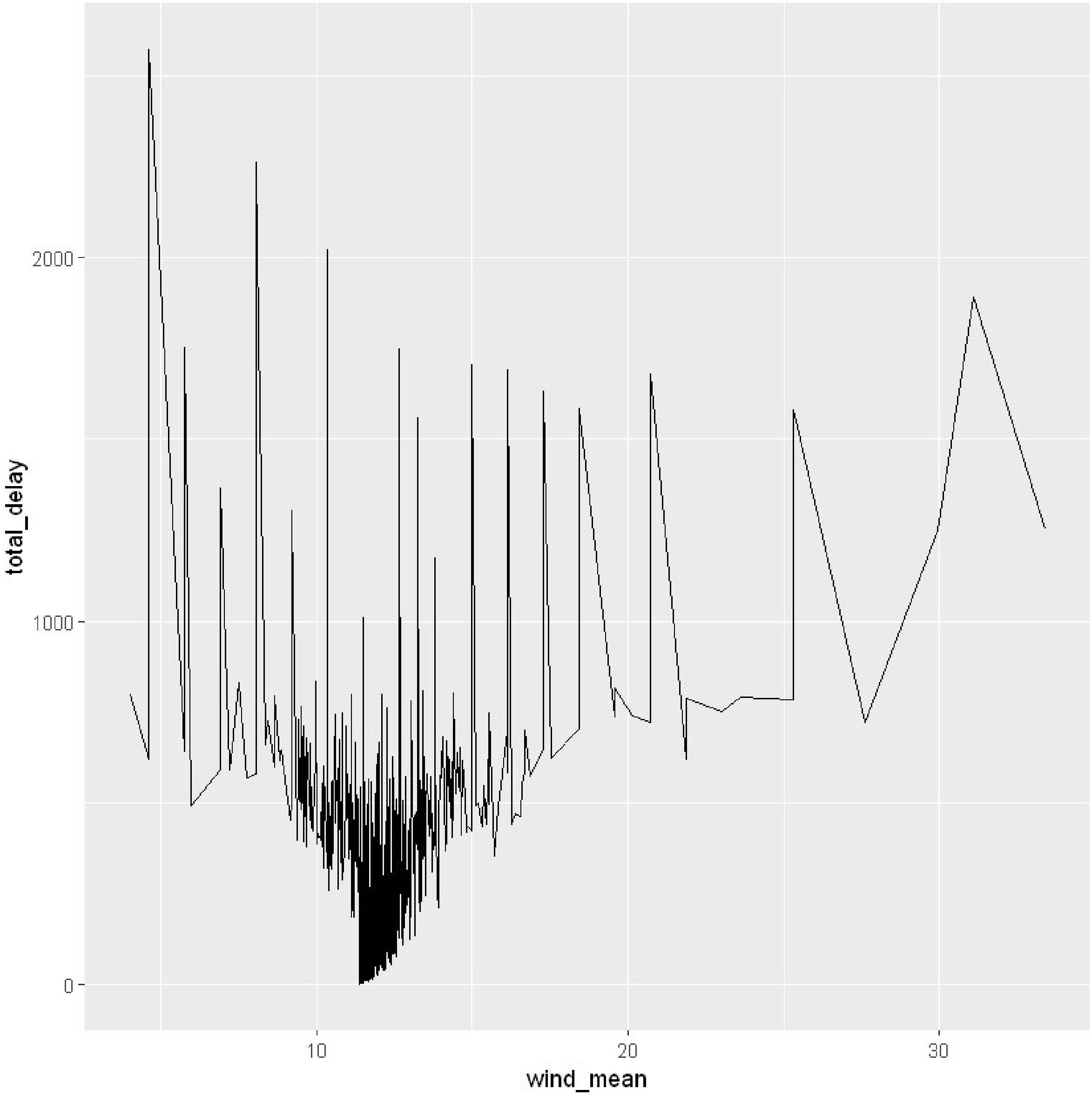
In [21]:

```
w3 %>%
ggplot(aes(x = wind_mean, y = total_delay)) +
geom_bar(stat="identity", fill="steelblue")
```



In [22]:

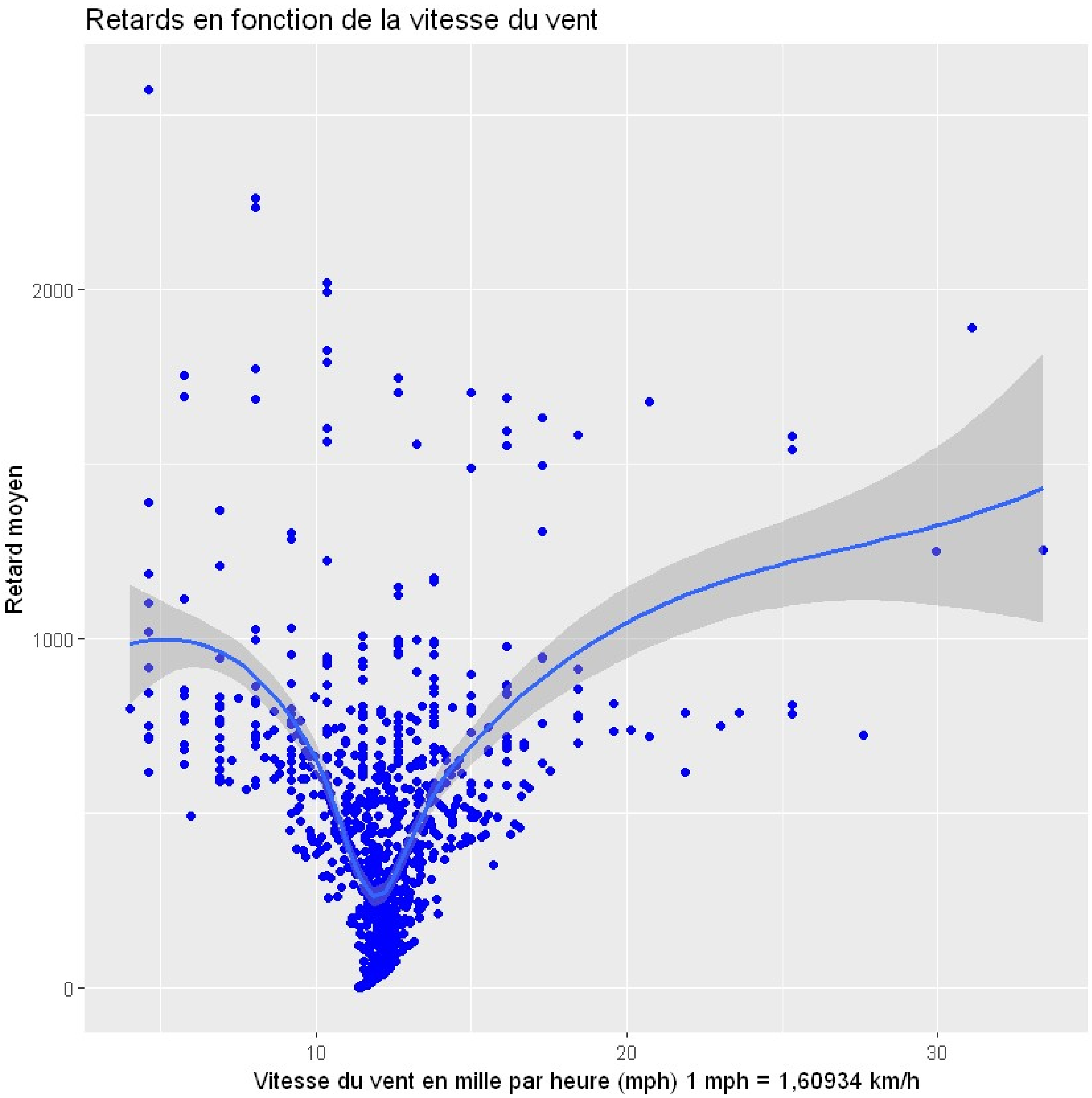
```
w3 %>%
ggplot(aes(x = wind_mean, y = total_delay)) +
geom_line()
```



In [82]:

```
w3 %>%
ggplot(aes(x = wind_mean, y = total_delay)) +
ggtitle("Retards en fonction de la vitesse du vent") +
labs(y="Retard moyen", x = "Vitesse du vent en mille par heure (mph) 1 mph = 1,60934 km/h") +
geom_point(color = "blue") +
geom_smooth()
```

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



In [9]:

```
w2 <- wind_data %>%
    summarise(mean(total_delay))
```

In [8]:

```
summarise(mean(arr_delay, na.rm=TRUE))
```

```
Error in mean(arr_delay, na.rm = TRUE): objet 'arr_delay' introuvable
Traceback:

1. summarise(mean(arr_delay, na.rm = TRUE))
2. mean(arr_delay, na.rm = TRUE)
```

In [10]:

```
w2
```

mean(total_delay)
61.76023

```
In [92]: wind_data
```

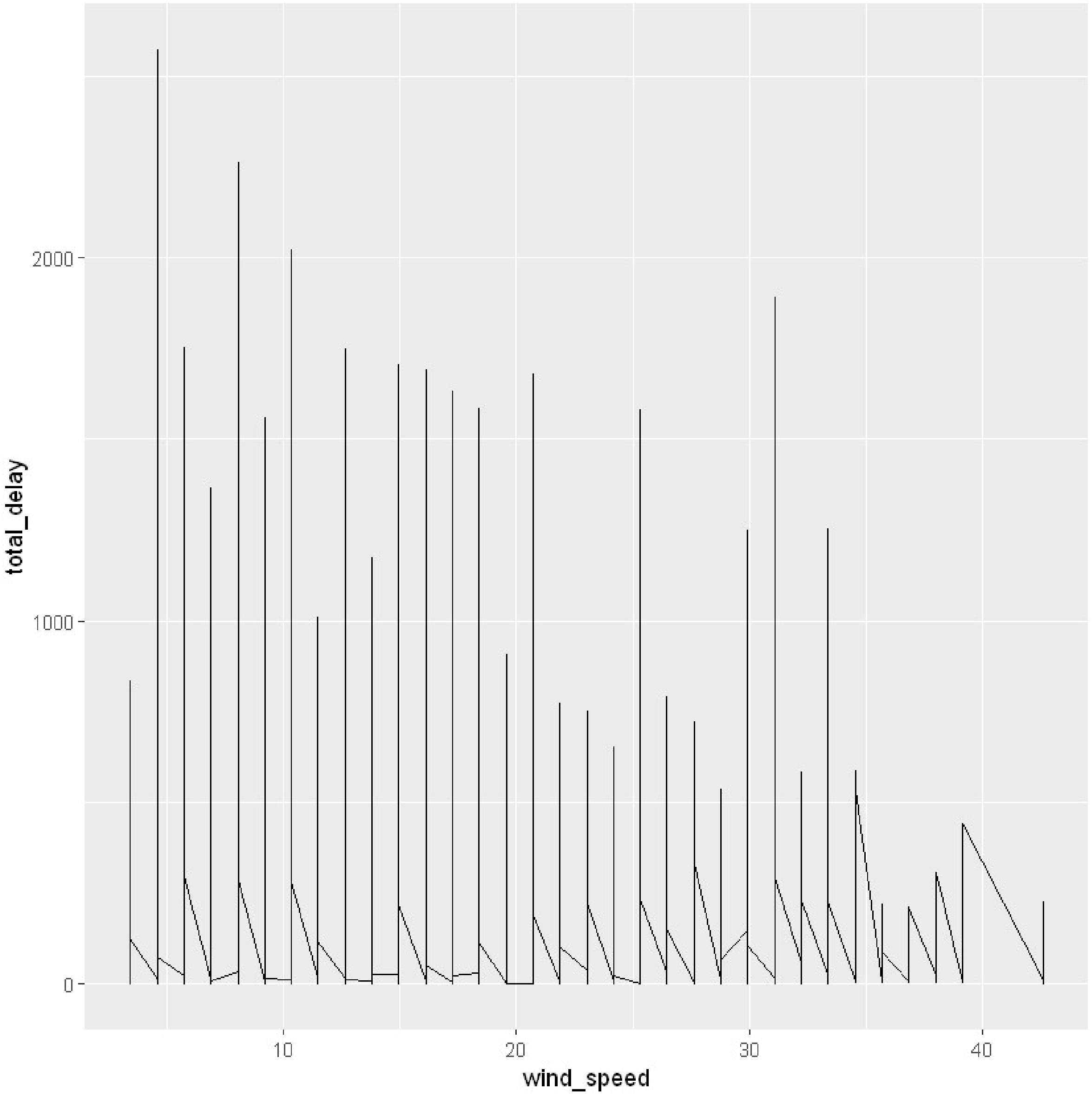
total_delay	wind_speed
13	12.65858
24	14.96014
35	14.96014
12	12.65858
19	11.50780
8	16.11092
7	13.80936
31	16.11092
12	16.11092
1	11.50780
16	16.11092
40	11.50780
25	13.80936
7	13.80936
3	11.50780
18	16.11092
1	16.11092
29	11.50780
10	16.11092
29	11.50780
14	16.11092
36	11.50780
48	16.11092
8	11.50780
5	13.80936
1	16.11092
1	11.50780
11	13.80936
4	14.96014
27	14.96014
...	...
35	5.75390
73	4.60312
9	8.05546
26	6.90468
118	5.75390
56	5.75390
3	3.45234
7	6.90468
6	3.45234
5	8.05546

In [88]:

```
wind_data %>%
    group_by(total_delay)  %>%
ggplot(aes(x = wind_speed, y = total_delay)) +
geom_histogram()
```

```
ERROR while rich displaying an object: Error: stat_bin() must not be used with a
y aesthetic.

Traceback:
1. FUN(X[[i]], ...)
2. tryCatch(withCallingHandlers({
 .     if (!mime %in% names(repr::mime2repr))
 .         stop("No repr_* for mimetype ", mime, " in repr::mime2repr")
 .     rpr <- repr::mime2repr[[mime]](obj)
 .     if (is.null(rpr))
 .         return(NULL)
 .     prepare_content(is.raw(rpr), rpr)
 . }, error = error_handler), error = outer_handler)
3. tryCatchList(expr, classes, parentenv, handlers)
4. tryCatchOne(expr, names, parentenv, handlers[[1L]])
5. doTryCatch(return(expr), name, parentenv, handler)
6. withCallingHandlers({
 .     if (!mime %in% names(repr::mime2repr))
 .         stop("No repr_* for mimetype ", mime, " in repr::mime2repr")
 .     rpr <- repr::mime2repr[[mime]](obj)
 .     if (is.null(rpr))
 .         return(NULL)
 .     prepare_content(is.raw(rpr), rpr)
 . }, error = error_handler)
7. repr::mime2repr[[mime]](obj)
8. repr_text.default(obj)
9. paste(capture.output(print(obj)), collapse = "\n")
10. capture.output(print(obj))
11. evalVis(expr)
12. withVisible(eval(expr, pf))
13. eval(expr, pf)
14. eval(expr, pf)
15. print(obj)
16. print.ggplot(obj)
17. ggplot_build(x)
18. ggplot_build.ggplot(x)
19. by_layer(function(l, d) l$compute_statistic(d, layout))
20. f(l = layers[[i]], d = data[[i]])
21. l$compute_statistic(d, layout)
22. f(..., self = self)
23. self$stat$setup_params(data, self$stat_params)
24. f(...)
25. stop("stat_bin() must not be used with a y aesthetic.", call. = FALSE)
```

```
In [78]: wind_data %>%
             group_by(total_delay)  %>%
         ggplot(aes(x = wind_speed, y = total_delay)) +
         geom_line()
```



```
In [48]: class(wind_data)
```

'numeric'

In [35]:

```
ggplot(aes(x = wind_speed, y = total_delay)) +
geom_line() + geom_point()
```

```
Error: `data` must be a data frame, or other object coercible by `fortify()`, no
t an S3 object with class uneval
Did you accidentally pass `aes()` to the `data` argument?
Traceback:

1. ggplot(aes(x = wind_speed, y = total_delay))
2. ggplot.default(aes(x = wind_speed, y = total_delay))
3. fortify(data, ...)
4. fortify.default(data, ...)
5. stop(msg, call. = FALSE)
```

In [77]:

```
avgdelay <- flights %>%
  group_by(month, day) %>%
  filter(month < 13) %>%
  summarise(avgdelay =mean(arr_delay, na.rm=TRUE))
precip <- weather %>%
  group_by(month, day) %>%
  filter(month < 13) %>%
  summarise(totprecip =sum(precip), maxwind =max(wind_speed))
precip <-mutate(precip, anyprecip =ifelse(totprecip==0, "No", "Yes"))
merged <-left_join(avgdelay, precip, by=c("day", "month"))
head(merged)
```

month	day	avgdelay	totprecip	maxwind	anyprecip
1	1	12.651023	0	24.16638	No
1	2	12.692888	0	20.71404	No
1	3	5.733333	0	17.26170	No
1	4	-1.932819	0	24.16638	No
1	5	-1.525802	0	20.71404	No
1	6	4.236429	0	16.11092	No

In [ ]:

wind_gust

In [31]:

```
wind_gust <- select(flights_weather, total_delay, wind_gust)
wind_gust <- wind_gust[!wind_gust$wind_gust <= 0,]
wind_gust <- wind_gust[!wind_gust$total_delay <= 0,]
wind_gust <- wind_gust[!is.na(wind_gust$total_delay),]
wind_gust <- wind_gust %>%
    group_by(total_delay)  %>%
    summarise(wind_gust_mean = mean(wind_gust, na.rm = TRUE))
```

In [32]:

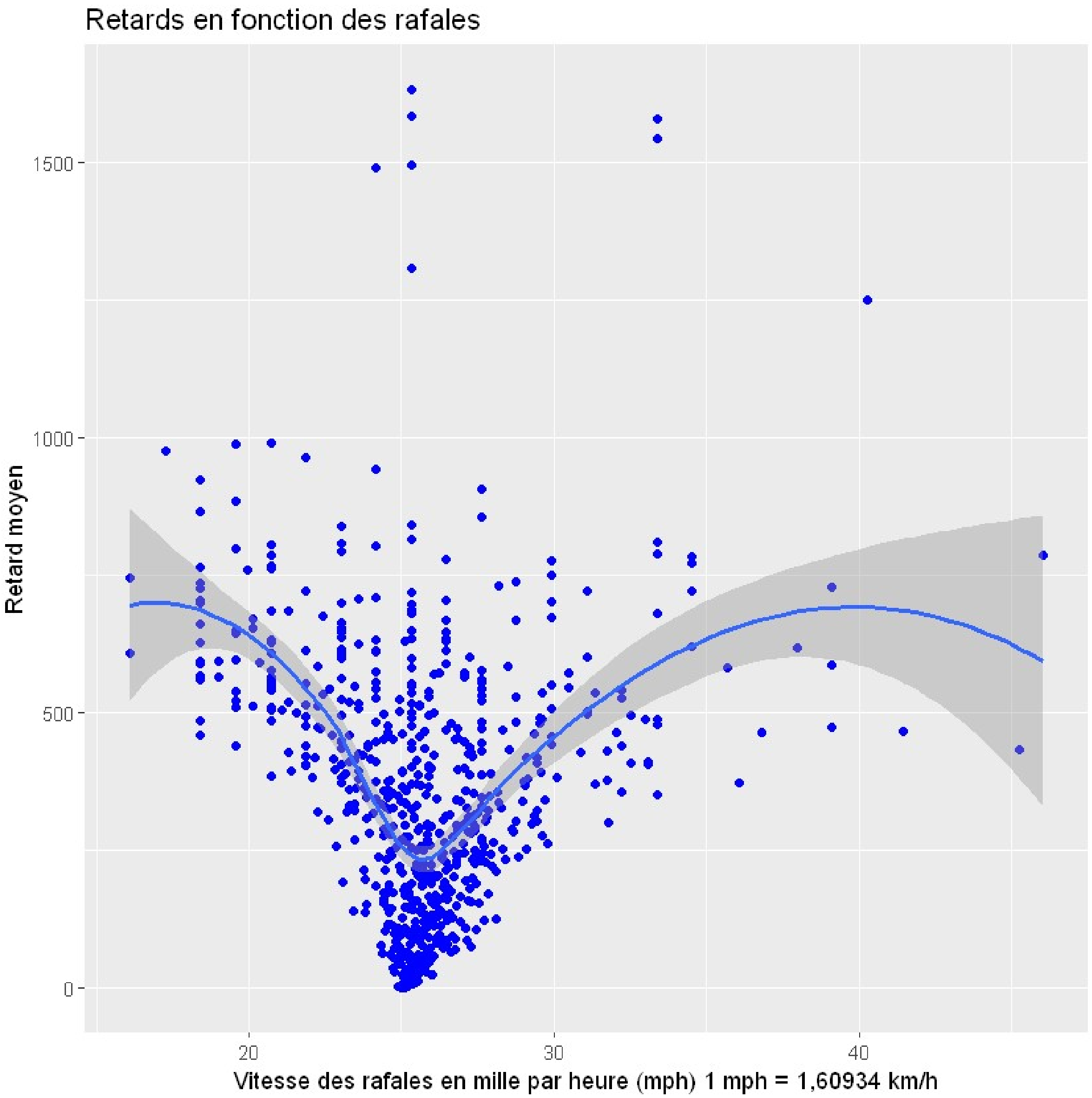
```
wind_gust
```

total_delay	wind_gust_mean
1	24.96079
2	25.12801
3	25.17795
4	24.90588
5	25.04525
6	24.98092
7	25.16848
8	25.37388
9	25.20208
10	25.17300
11	25.44884
12	25.14055
13	25.22647
14	25.55995
15	25.31355
16	25.33225
17	25.33474
18	25.52201
19	25.28474
20	25.08903
21	25.50598
22	25.43282
23	25.28120
24	25.00955
25	26.03235
26	25.97217
27	25.60112
28	25.38392
29	24.75133
30	25.81735
...	...
786	20.71404
788	46.03120
790	33.37262
795	23.01560
798	19.56326
803	24.16638
806	20.71404
809	23.01560
811	33.37262
817	25.31716

In [81]:

```
wind_gust %>%
ggplot(aes(x = wind_gust_mean, y = total_delay)) +
geom_point(color = "blue") +
ggtitle("Retards en fonction des rafales") +
labs(y="Retard moyen", x = "Vitesse des rafales en mille par heure (mph) 1 mph = 1,60934 km/h") +
geom_smooth()
```

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



pluie : precip

In [49]:

```
df_precip <- select(flights_weather, total_delay, precip)
df_precip <- df_precip[!df_precip$precip <= 0,]
df_precip <- df_precip[!df_precip$total_delay <= 0,]
df_precip <- df_precip[!is.na(df_precip$total_delay),]
df_precip <- df_precip %>%
    group_by(total_delay)  %>%
    summarise(precip_mean = mean(precip, na.rm = TRUE))
#df_visib <- df_visib[df_visib$visib_mean < 10,]
```

In [50]:

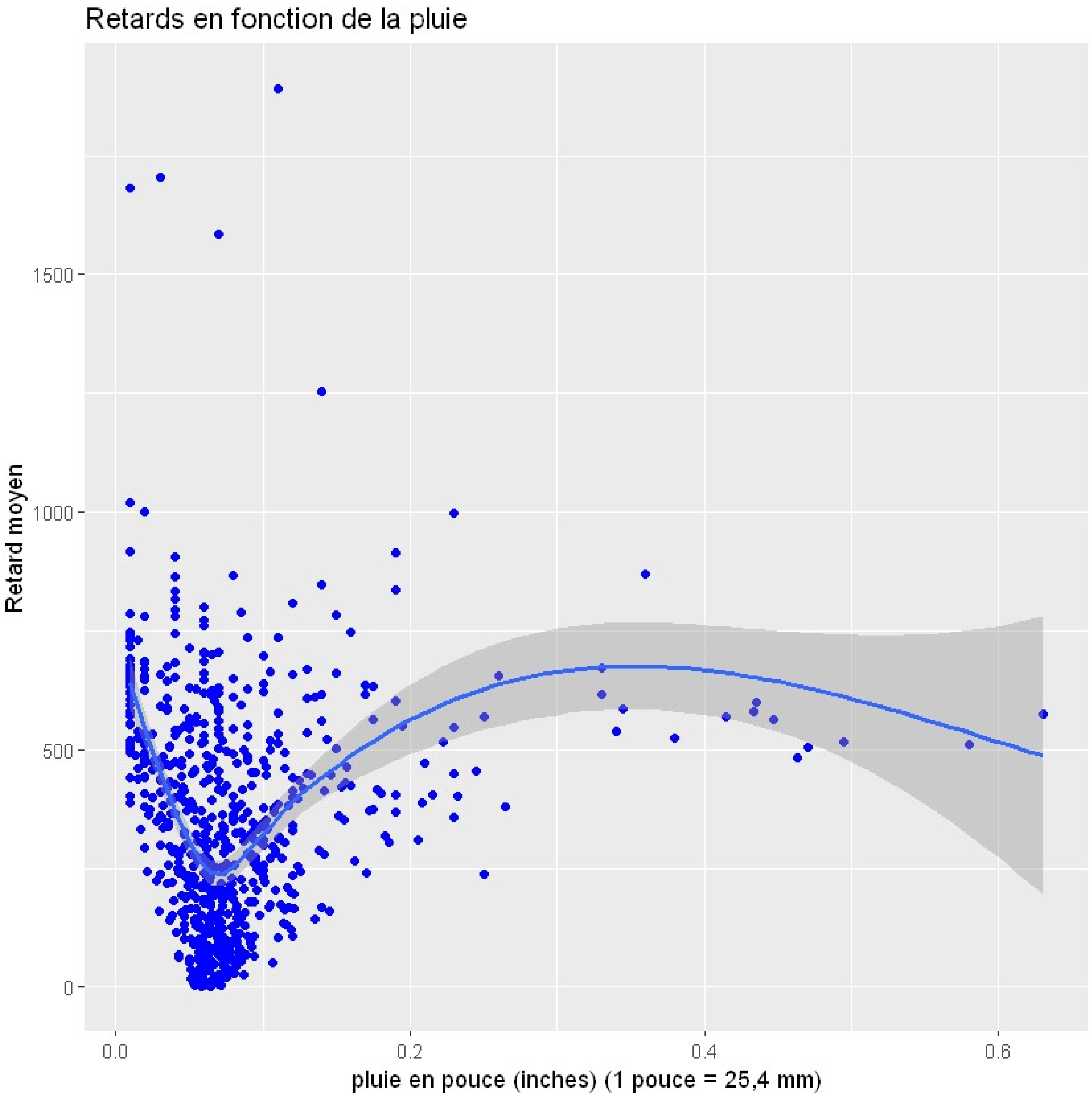
```
df_precip
```

total_delay	precip_mean
1	0.05827251
2	0.06426593
3	0.05344928
4	0.07196375
5	0.05979730
6	0.06046823
7	0.07033613
8	0.06207031
9	0.05262295
10	0.05534884
11	0.06431818
12	0.05543103
13	0.06050000
14	0.06995050
15	0.06558974
16	0.05605442
17	0.07098901
18	0.05111801
19	0.05917808
20	0.07335443
21	0.07564286
22	0.06750000
23	0.07000000
24	0.05767296
25	0.06983740
26	0.08685714
27	0.07082759
28	0.08043103
29	0.05490909
30	0.05984962
...	...
747	0.160
748	0.010
762	0.060
773	0.060
779	0.040
781	0.020
782	0.150
786	0.010
788	0.085
793	0.040

In [70]:

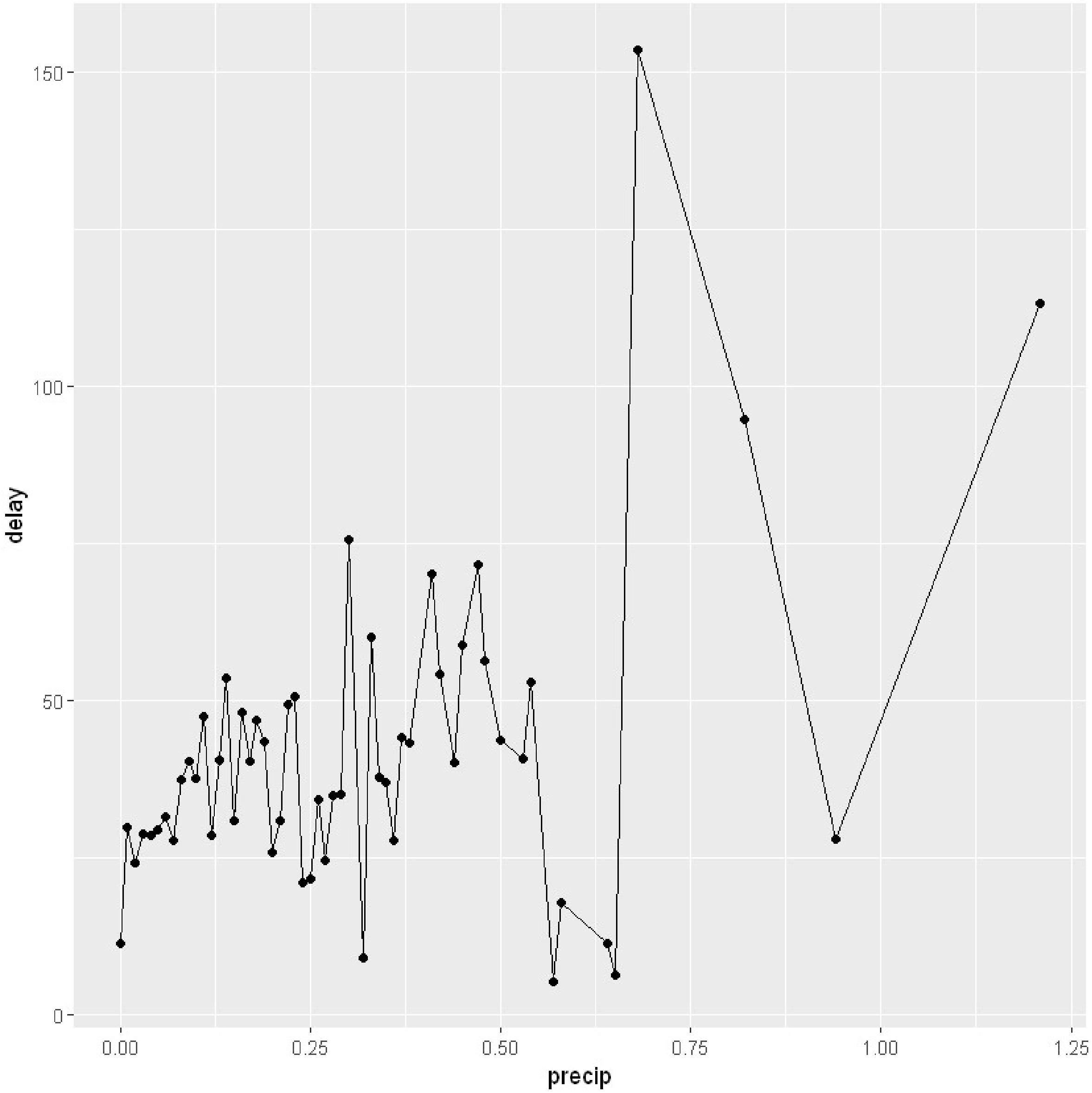
```
df_precip %>%
ggplot(aes(x = precip_mean, y = total_delay)) +
geom_point(color = "blue") +
ggtitle("Retards en fonction de la pluie") +
labs(y="Retard moyen", x = "pluie en pouce (inches) (1 pouce = 25,4 mm)") +
geom_smooth()
```

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



In [63]:

```
flight_weather <-
  flights %>%
  inner_join(weather, by = c(
    "origin" = "origin",
    "year" = "year",
    "month" = "month",
    "day" = "day",
    "hour" = "hour"
  ))

flight_weather %>%
  group_by(precip) %>%
  summarise(delay = mean(dep_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = precip, y = delay)) +
  geom_line() + geom_point()
```



test

In [64]:

```r
precip <- weather %>%
  group_by(month, day) %>%
  filter(month < 13) %>%
 mutate(totprecip = sum(precip), maxwind = max(wind_speed))
```

In [65]:

```
precip
```

origin	year	month	day	hour	temp	dewp	humid	wind_dir	wind_speed	wind_gust	precip	pressure	vis
EWR	2013	1	1	1	39.02	26.06	59.37	270	10.35702	NA	0	1012.0	
EWR	2013	1	1	2	39.02	26.96	61.63	250	8.05546	NA	0	1012.3	
EWR	2013	1	1	3	39.02	28.04	64.43	240	11.50780	NA	0	1012.5	
EWR	2013	1	1	4	39.92	28.04	62.21	250	12.65858	NA	0	1012.2	
EWR	2013	1	1	5	39.02	28.04	64.43	260	12.65858	NA	0	1011.9	
EWR	2013	1	1	6	37.94	28.04	67.21	240	11.50780	NA	0	1012.4	
EWR	2013	1	1	7	39.02	28.04	64.43	240	14.96014	NA	0	1012.2	
EWR	2013	1	1	8	39.92	28.04	62.21	250	10.35702	NA	0	1012.2	
EWR	2013	1	1	9	39.92	28.04	62.21	260	14.96014	NA	0	1012.7	
EWR	2013	1	1	10	41.00	28.04	59.65	260	13.80936	NA	0	1012.4	
EWR	2013	1	1	11	41.00	26.96	57.06	260	14.96014	NA	0	1011.4	
EWR	2013	1	1	13	39.20	28.40	69.67	330	16.11092	NA	0	NA	
EWR	2013	1	1	14	39.02	24.08	54.68	280	13.80936	NA	0	1010.8	
EWR	2013	1	1	15	37.94	24.08	57.04	290	9.20624	NA	0	1011.9	
EWR	2013	1	1	16	37.04	19.94	49.62	300	13.80936	20.71404	0	1012.1	
EWR	2013	1	1	17	35.96	19.04	49.83	330	11.50780	NA	0	1013.2	
EWR	2013	1	1	18	33.98	15.08	45.43	310	12.65858	25.31716	0	1014.1	
EWR	2013	1	1	19	33.08	12.92	42.84	320	10.35702	NA	0	1014.4	
EWR	2013	1	1	20	32.00	15.08	49.19	310	14.96014	NA	0	1015.2	
EWR	2013	1	1	21	30.02	12.92	48.48	320	18.41248	26.46794	0	1016.0	
EWR	2013	1	1	22	28.94	12.02	48.69	320	18.41248	25.31716	0	1016.5	
EWR	2013	1	1	23	28.04	10.94	48.15	310	16.11092	NA	0	1016.4	
EWR	2013	1	2	0	26.96	10.94	50.34	310	14.96014	25.31716	0	1016.3	
EWR	2013	1	2	1	26.06	10.94	52.25	330	12.65858	24.16638	0	1016.3	
EWR	2013	1	2	2	24.98	10.94	54.65	330	13.80936	NA	0	1017.0	
EWR	2013	1	2	3	24.08	8.96	51.93	320	14.96014	NA	0	1016.6	
EWR	2013	1	2	4	24.08	8.96	51.93	330	12.65858	NA	0	1016.9	

pressure

In [ ]:

visib

In [46]:

```
df_visib <- select(flights_weather, total_delay, visib)
df_visib <- df_visib[!df_visib$visib <= 0,]
df_visib <- df_visib[!df_visib$total_delay <= 0,]
df_visib <- df_visib[!is.na(df_visib$total_delay),]
df_visib <- df_visib %>%
    group_by(total_delay)  %>%
    summarise(visib_mean = mean(visib, na.rm = TRUE))
df_visib <- df_visib[df_visib$visib_mean < 10,]
```

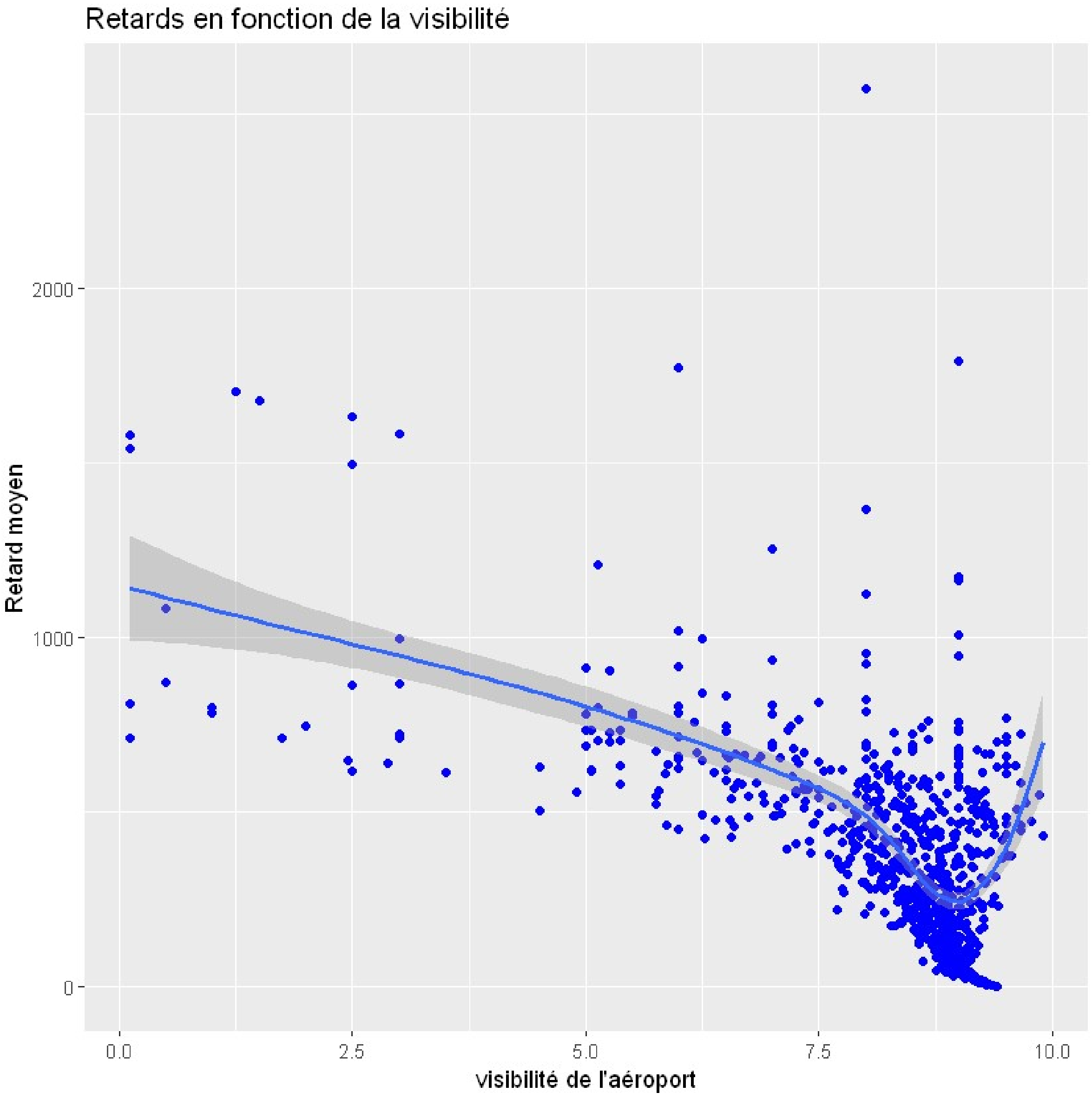
In [47]:

```
df_visib
```

total_delay	visib_mean
1	9.405004
2	9.399421
3	9.371053
4	9.352495
5	9.358589
6	9.292350
7	9.328019
8	9.326277
9	9.323222
10	9.311045
11	9.259924
12	9.296507
13	9.236995
14	9.214494
15	9.268127
16	9.274436
17	9.184680
18	9.199477
19	9.239882
20	9.145822
21	9.202586
22	9.199527
23	9.063173
24	9.152714
25	9.138356
26	9.058327
27	9.070717
28	9.161089
29	9.155214
30	8.940541
...	...
870	3.000
874	0.500
906	5.250
914	5.000
917	6.000
924	8.000
936	7.000
949	9.000
954	8.000
997	6.250

```
In [68]: df_visib %>%
         ggplot(aes(x = visib_mean, y = total_delay)) +
         ggtitle("Retards en fonction de la visibilité") +
         labs(y="Retard moyen", x = "visibilité de l'aéroport") +
         geom_point(color = "blue") +
         geom_smooth()
```

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



besoin d'utiliser la fonction unite dans tidyr et ajout de la fonction parse_date conda install -c conda-forge r-parsedate

[https://lokhc.wordpress.com/r-for-data-science-solutions/chapter-13-relational-data/ (https://lokhc.wordpress.com/r-for-data-science-solutions/chapter-13-relational-data/)](https://lokhc.wordpress.com/r-for-data-science-solutions/chapter-13-relational-data/)

In [20]:

```
flights_2day <- flights %>% group_by(year, month, day) %>%
  summarize(avg_dep_delay = mean(dep_delay, na.rm = TRUE),
            avg_arr_delay = mean(arr_delay, na.rm = TRUE)) %>%
  unite(date, year, month, day, sep = '-') %>%
  mutate(date = parse_date(date, "%Y-%m-%d")) %>%
  gather(key = 'mode', value = 'delay', 2:3) %>%
  mutate(mode = factor(mode, labels = c('Average arrival delay',
                                        'Average departure delay')))

weather_2day <- weather %>% group_by(year, month, day) %>%
  summarize(avg_wind_speed = mean(wind_speed, na.rm = TRUE),
            avg_wind_gust = mean(wind_gust, na.rm = TRUE),
            avg_precip = mean(precip, na.rm = TRUE),
            avg_visib = mean(visib, na.rm = TRUE)) %>%
  unite(date, year, month, day, sep = '-') %>%
  mutate(date = parse_date(date, "%Y-%m-%d"))

flights_2day %>% ggplot() +
  geom_point(mapping = aes(x = date, y = delay, color = mode)) +
  geom_line(mapping = aes(x = date, y = delay, color = mode)) +
  geom_line(data = weather_2day,
            mapping = aes(x = date, y = (avg_visib-10)*5, color = 'Average visibili
ty')) +
  scale_y_continuous(sec.axis = sec_axis(~./5 + 10,
                                         name = "Average visibility (km)")) +
  facet_wrap(~mode, ncol = 1) +
  labs(x = "Date",
       y = "Average delay (minutes)",
       color = 'Legend',
       title = "Average delay and average visibility")
```

```
Error in parse_date(date, "%Y-%m-%d"): impossible de trouver la fonction "parse_
date"
Traceback:

1. flights %>% group_by(year, month, day) %>% summarize(avg_dep_delay = mean(dep
_delay,
 .     na.rm = TRUE), avg_arr_delay = mean(arr_delay, na.rm = TRUE)) %>%
 .     unite(date, year, month, day, sep = "-") %>% mutate(date = parse_date(dat
e,
 .     "%Y-%m-%d")) %>% gather(key = "mode", value = "delay", 2:3) %>%
 .     mutate(mode = factor(mode, labels = c("Average arrival delay",
 .         "Average departure delay")))
2. withVisible(eval(quote(`_fseq`(`_lhs`)), env, env))
3. eval(quote(`_fseq`(`_lhs`)), env, env)
4. eval(quote(`_fseq`(`_lhs`)), env, env)
5. `_fseq`(`_lhs`)
6. freduce(value, `_function_list`)
7. function_list[[i]](value)
8. mutate(., date = parse_date(date, "%Y-%m-%d"))
9. mutate.tbl_df(., date = parse_date(date, "%Y-%m-%d"))
10. mutate_impl(.data, dots, caller_env())
```

time_hour

In [ ]:

Quel traitement reste-t-il à faire sur la base de données flights pour pouvoir la rapprocher des données météo?

```
In [ ]: flights %>% group_by(origin)
```

```
In [ ]: # grouper les vols par heure (moyenne)  ajouter au tableau
        # merger sur time_hour (grouper ou merger) sumerize
        # et origin
```

Fusion de tables

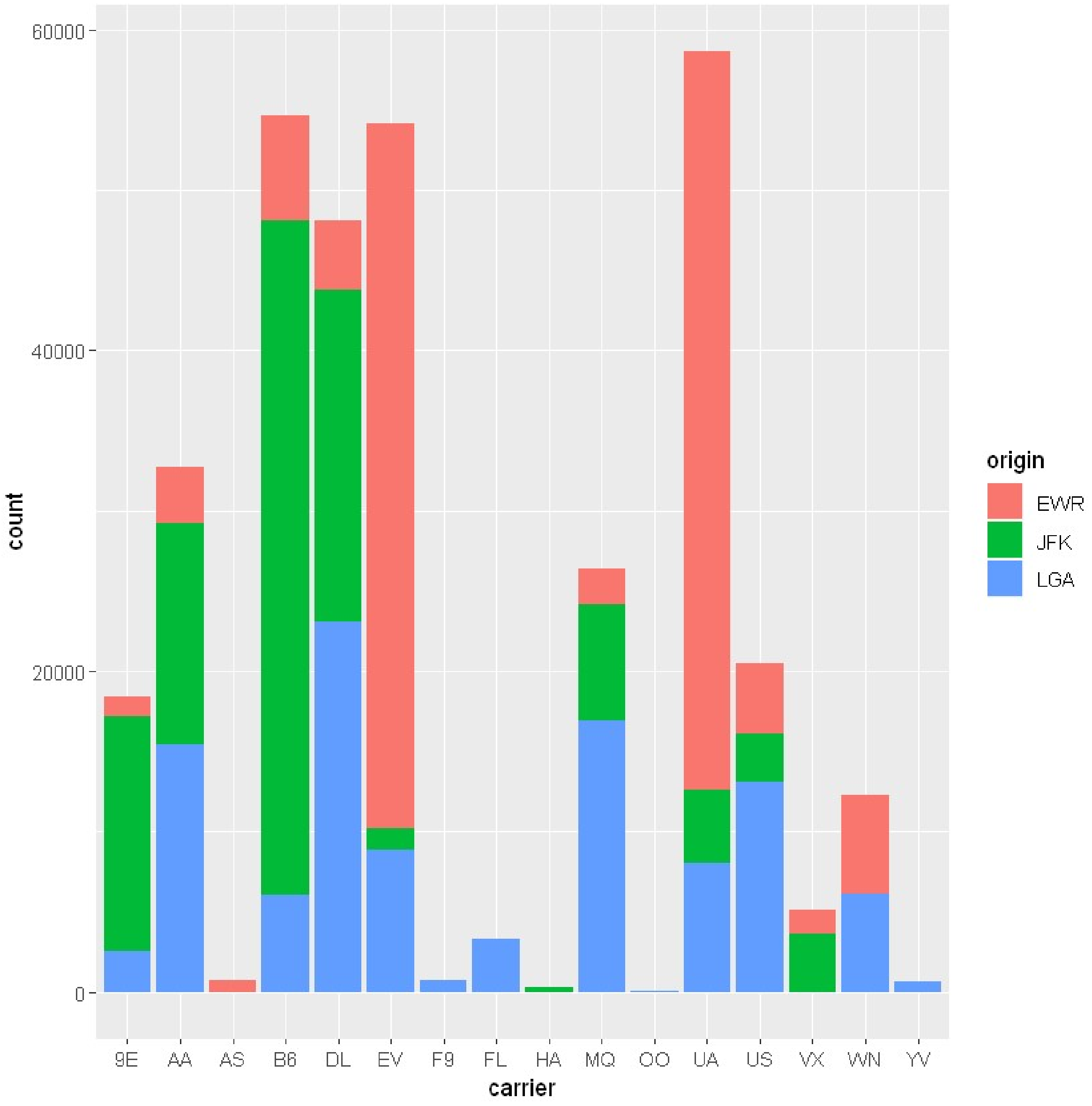
Fusionnez la table flights ainsi transformée et la table weather en utilisant la fonction de merge de dplyr qui vous semble la plus appropriée entre inner_join, left_join, right_join et outer_join.

Vérifiez que de nouvelles valeurs manquantes ne sont pas apparues dans cette nouvelle table, si oui traitez-les.

Analyse

En vous appuyant sur la comparaison des statistiques déjà réalisées et sur au moins 4 représentations graphiques bien choisies, proposez une analyse de l'effet des conditions météorologiques sur les retards des avions. Pensez à définir une problématique en amont, que vous êtes libres de choisir : vous n'êtes pas obligés d'utiliser toutes les variables à disposition!

```
In [10]: ggplot(data = flights, mapping = aes(x = carrier, fill = origin)) +
           geom_bar()
```



Export RDS

```
In [54]: delays_weather <- select(flights_weather,arr_delay, dep_delay,total_delay, temp, de
         wp, humid,
                             wind_dir, wind_speed, wind_gust, precip, pressure, visib)
```

```
In [ ]: delays_weather <-  delays_weather is.na(delays_weather)
        wind_data <- wind_data[!wind_data$total_delay <= 0,]
        wind_data <- wind_data[!is.na(wind_data$total_delay),]
```

```
In [65]: delays_weather <-  delays_weather (, na.rm=TRUE)
```

```
Error in delays_weather(, na.rm = TRUE): impossible de trouver la fonction "dela
ys_weather"
Traceback:
```

```
In [61]: test <- c(1,2,3,NA) is.na(test)
```

```
Error in parse(text = x, srcfile = src): <text>:1:21: unexpected symbol
1: test <- c(1,2,3,NA) is.na
                        ^

Traceback:
```

```
In [60]: na_list <- l(flights_weather,arr_delay, dep_delay,total_delay, temp, dewp, humid,
                            wind_dir, wind_speed, wind_gust, precip, pressure, visib) is.na
         (na_list)
```

```
Error in parse(text = x, srcfile = src): <text>:2:78: unexpected symbol
1: na_list <- delays_weather(flights_weather,arr_delay, dep_delay,total_delay, t
emp, dewp, humid,
2:                      wind_dir, wind_speed, wind_gust, precip, pressure, visib)
is.na

^
Traceback:
```

```
In [55]: saveRDS(delays_weather, file = "delays_weather.rds")
In [56]: readRDS(file = "delays_weather.rds")
```

arr_delay	dep_delay	total_delay	temp	dewp	humid	wind_dir	wind_speed	wind_gust	precip	pressure	vis
11	2	13	39.02	28.04	64.43	260	12.65858	0.00000	0	1011.9	
20	4	24	39.92	24.98	54.81	250	14.96014	21.86482	0	1011.4	
33	2	35	39.02	26.96	61.63	260	14.96014	0.00000	0	1012.1	
0	0	0	39.02	26.96	61.63	260	14.96014	0.00000	0	1012.1	
0	0	0	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
12	0	12	39.02	28.04	64.43	260	12.65858	0.00000	0	1011.9	
19	0	19	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
0	0	0	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
0	0	0	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
8	0	8	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
0	0	0	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
0	0	0	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
7	0	7	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
0	0	0	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
31	0	31	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
0	0	0	39.02	26.96	61.63	260	14.96014	0.00000	0	1012.1	
0	0	0	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
0	0	0	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
12	0	12	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
0	1	1	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
0	0	0	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
16	0	16	39.92	24.98	54.81	260	16.11092	23.01560	0	1011.7	
0	0	0	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
0	0	0	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
0	0	0	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
32	8	40	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
14	11	25	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
4	3	7	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
0	0	0	37.94	26.96	64.29	260	13.80936	0.00000	0	1012.6	
0	0	0	37.94	28.04	67.21	240	11.50780	0.00000	0	1012.4	
...	...	...	...	...	...	...	...	...	...	...	
0	0	0	64.94	53.06	65.37	210	8.05546	0	0	1015.8	
0	0	0	62.96	55.04	75.33	190	3.45234	0	0	1016.1	
0	0	0	62.06	57.02	83.54	230	9.20624	0	0	1016.4	
0	30	30	64.94	53.96	67.57	190	6.90468	0	0	1015.7	
0	0	0	62.06	57.02	83.54	230	9.20624	0	0	1016.4	
0	0	0	62.06	57.02	83.54	230	9.20624	0	0	1016.4	
11	13	24	62.96	55.04	75.33	190	3.45234	0	0	1016.1	
0	0	0	62.06	57.02	83.54	230	9.20624	0	0	1016.4	
3	10	13	64.94	53.06	65.37	210	8.05546	0	0	1015.8	
0	0	0	62.96	55.04	75.33	190	3.45234	0	0	1016.1	

In [ ]: