In [1]:

```
library(tidyverse)
library(sparklyr)
library(tictoc)
library(fs)
```

```
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.1.0.9000     ✔ purrr   0.3.1
✔ tibble  2.0.1          ✔ dplyr   0.8.0.1
✔ tidyr   0.8.3          ✔ stringr 1.4.0
✔ readr   1.3.1          ✔ forcats 0.4.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()

Attaching package: 'sparklyr'

The following object is masked from 'package:purrr':

    invoke
```

In [2]:

```
sparklyr::spark_install(version = "2.4.0")
```

In [3]:

```
sc = spark_connect(master = "local[*]")
```

In [4]:

```
dir_info("../input/ontime")$path
```

'../input/ontime/1987.csv.gz' '../input/ontime/1988.csv.gz' '../input/ontime/1989.csv.gz' '../input/ontime/1990.csv.gz'
'../input/ontime/1991.csv.gz' '../input/ontime/1992.csv.gz' '../input/ontime/1993.csv.gz' '../input/ontime/1994.csv.gz'
'../input/ontime/1995.csv.gz' '../input/ontime/1996.csv.gz' '../input/ontime/1997.csv.gz' '../input/ontime/1998.csv.gz'
'../input/ontime/1999.csv.gz' '../input/ontime/2000.csv.gz' '../input/ontime/2001.csv.gz' '../input/ontime/2002.csv.gz'
'../input/ontime/2003.csv.gz' '../input/ontime/2004.csv.gz' '../input/ontime/2005.csv.gz' '../input/ontime/2006.csv.gz'
'../input/ontime/2007.csv.gz' '../input/ontime/2008.csv.gz' '../input/ontime/airports.csv' '../input/ontime/carriers.csv'

In [5]:

```
ontime_tbl =
spark_read_csv(sc,
               name = "ontime",
               path="../input/ontime/*.csv.gz",
               memory = F, null_value = "NA",
               infer_schema = F,
               columns = list(
                   Year="integer",
                   Month="integer",
                   DayofMonth ="integer",
                   DayOfWeek="integer",
                   DepTime  ="integer",
                   CRSDepTime="integer",
                   ArrTime="integer",
                   CRSArrTime ="integer",
                   UniqueCarrier ="character",
                   FlightNum ="integer",
                   TailNum ="character",
                   ActualElapsedTime ="integer",
                   CRSElapsedTime ="integer",
                   AirTime ="integer",
                   ArrDelay ="integer",
                   DepDelay ="integer",
```

```
                    Origin ="character",
                    Dest ="character",
                    Distance ="integer",
                    TaxiIn ="integer",
                    TaxiOut ="integer",
                    Cancelled ="integer",
                    CancellationCode ="character",
                    Diverted ="character",
                    CarrierDelay ="integer",
                    WeatherDelay ="integer",
                    NASDelay ="integer",
                    SecurityDelay ="integer",
                    LateAircraftDelay ="integer"
                )
            )
```

분석1. 항공 출발 지연 데이터 분석

1.SQL 처리

In [6]:

```
depcount_by_month_by_year_sql =
sdf_sql(sc, "
            select year, month, count(*) as n
            from ontime
            where depdelay is not null
            and depdelay > 0
            group by year, month
            order by year, month
") %>% collect()

head(depcount_by_month_by_year_sql,5)
```

year	month	n
1987	10	175568
1987	11	177218
1987	12	218858
1988	1	198610
1988	2	177939

2.dplyr처리

In [7]:

```
depcount_by_month_by_year_dqplyr =
ontime_tbl %>% filter(!is.na(DepDelay)) %>% filter(DepDelay>0) %>% count(Year, Month) %>% arrange(Y
ear,Month) %>% collect()
head(depcount_by_month_by_year_dqplyr,5)
```

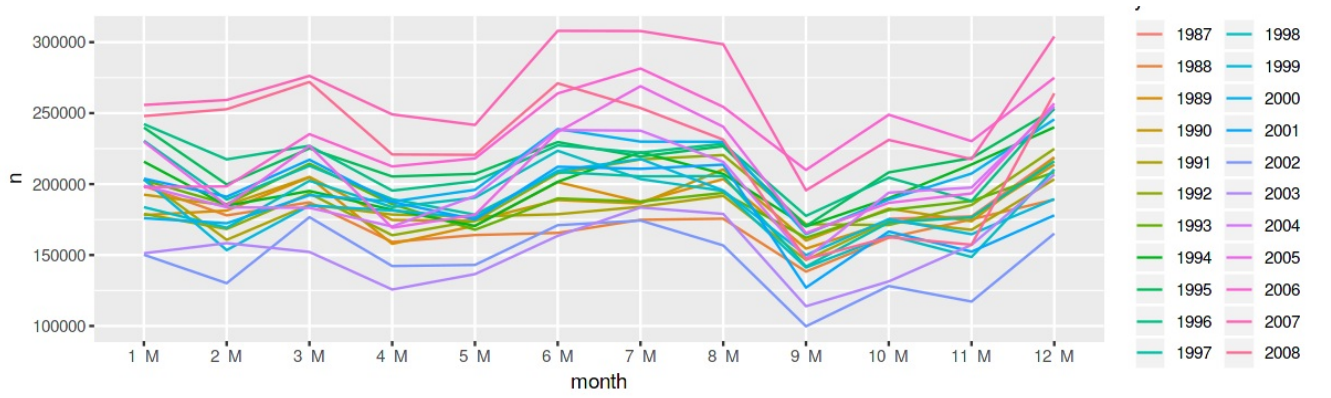
Year	Month	n
1987	10	175568
1987	11	177218
1987	12	218858
1988	1	198610
1988	2	177939

3.Line Plot 처리

In [8]:

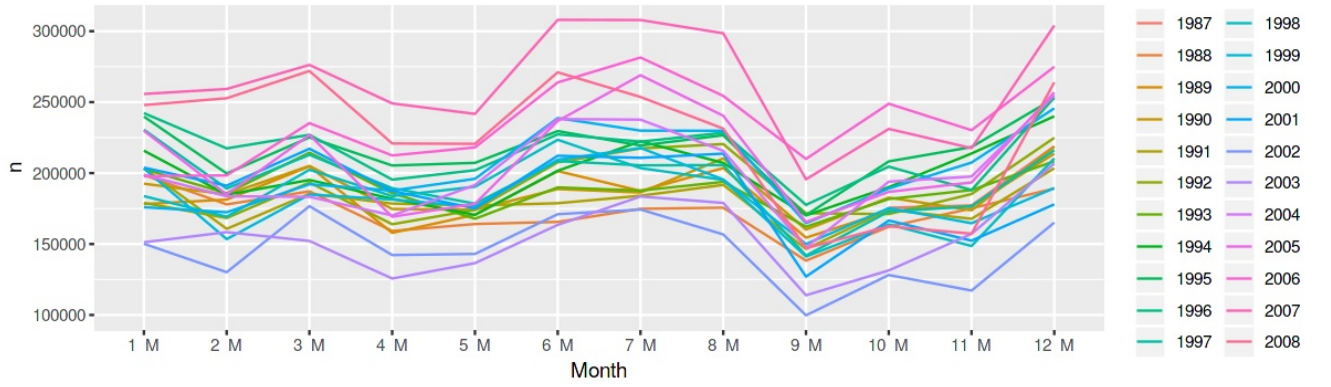
```
library(repr)
options(repr.plot.width=10,repr.plot.height=3) #비율
depcount_by_month_by_year_sql %>% mutate(year=factor(year)) -> plot_df1

plot_df1 %>% ggplot(aes(month,n,color=year))+geom_line()+scale_x_discrete(limits=1:12, labels=paste
(1:12, " M"))
```



In [9]:

```
depcount_by_month_by_year_dqplyr %>% mutate(Year=factor(Year)) -> plot_df2
plot_df2 %>% ggplot(aes(Month,n,color=Year))+geom_line()+scale_x_discrete(limits=1:12, labels=paste
(1:12, " M"))
```



분석2. 항공 도착 지연 데이터 분석

1. SQL 처리

In [10]:

```
arrcount_by_month_by_year_sql =
sdf_sql(sc, "
            select year, month, count(*) as n
            from ontime
            where arrdelay is not null
            and arrdelay > 0
            group by year, month
            order by year, month
") %>% collect()
head(arrcount_by_month_by_year_sql,5)
```

year	month	n
1987	10	265658
1987	11	255127
1987	12	287408
1988	1	261810
1988	2	242219

2. dplyr 처리

In [11]:

```
arrcount_by_month_by_year_dqplyr =
ontime_tbl %>% filter(!is.na(ArrDelay)) %>% filter(ArrDelay>0) %>% count(Year, Month) %>% arrange(Y
ear,Month) %>% collect()
head(arrcount_by_month_by_year_dqplyr,5)
```

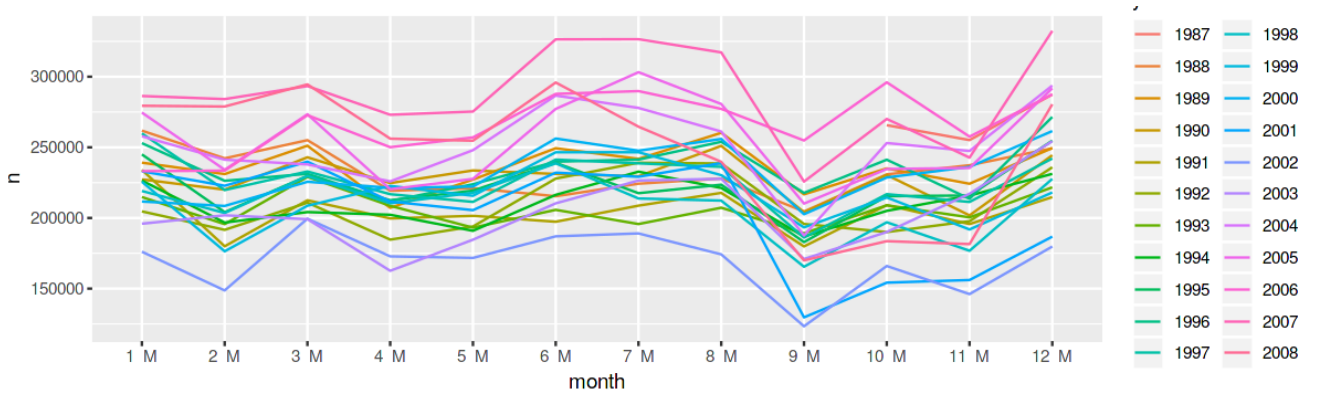
Year	Month	n
1987	10	265658
1987	11	255127
1987	12	287408
1988	1	261810
1988	2	242219

3.Line Plot 처리

In [12]:

```
arrcount_by_month_by_year_sql %>% mutate(year=factor(year)) -> plot_df3

plot_df3 %>% ggplot(aes(month,n,color=year))+geom_line()+scale_x_discrete(limits=1:12, labels=paste
(1:12, " M"))
```



In [13]:

```
arrcount_by_month_by_year_dqplyr %>% mutate(Year=factor(Year)) -> plot_df4
plot_df4 %>% ggplot(aes(Month,n,color=Year))+geom_line()+scale_x_discrete(limits=1:12, labels=paste
(1:12, " M"))
```