Inc.5000-2019

Zouari Adem

Introduction

Mon projet consiste à nettoyer et analyser un ensemble de données sur les 5000 meilleures entreprises en croissance en Amérique en 2019. L'ensemble de données comprend 14 variables et 5013 entreprises. L'objectif est de déterminer quelles industries et quels États ont les meilleurs revenus et taux de croissance. Mon objectif est également de déterminer s'il existe une corrélation entre : "revenu - travailleurs / revenu - années sur la liste / croissance - travailleurs / croissance - années sur la liste" Et mon objectif aussi est de visualiser mes données et finalement prédire le revenue en fonction des années sur la liste et le nombre des travailleurs.

Source de Dataset: INC 5000 - 2019 - dataset by aurielle | data.world

1.Chargement des données

```
library(readr)
library(ggplot2)
library(dplyr)
library(e1071)
library(ggfortify)
library(plot3D)
library(randomForest)
```

```
inc5000 <- read_csv("C:/Users/Administrator/Downloads/inc5000_2019.csv")
head(inc5000)
```

```
# A tibble: 6 x 14
   rank profile        name  url   state revenue growth industry workers founded
  <dbl> <chr>          <chr> <chr> <chr> <chr>    <dbl> <chr>      <dbl>   <dbl>
1     1 https://www.i~ Free~ http~ AZ    36.9 M~ 36680. Adverti~      40    2015
2     2 https://www.i~ Frei~ http~ TN    33.6 M~ 30548. Logisti~      39    2015
3     3 https://www.i~ Cece~ http~ TX    24.9 M~ 23880. Food & ~     190    2015
```

```
4     4 https://www.i~ Lady~ http~ NM    32.4 M~ 21850. Consume~      57    2014
5     5 https://www.i~ Perp~ http~ PA    22.5 M~ 18166. Retail        25    2014
6     6 https://www.i~ Cano~ http~ FL    271.8 ~ 14183. Health       742    2009
# i 4 more variables: yrs_on_list <dbl>, previous_workers <dbl>, metro <chr>,
#   city <chr>
```

2.Nettoyage

2.1 Vérifier le type de données de chaque colonne

```
print(sapply(inc5000, class))
```

```
          rank        profile           name            url
     "numeric"    "character"    "character"    "character"
         state        revenue         growth       industry
 "character"    "character"      "numeric"    "character"
       workers        founded   yrs_on_list previous_workers
     "numeric"      "numeric"      "numeric"      "numeric"
         metro           city
 "character"    "character"
```

2.2 Suppression des caractères non numériques et conversion en numérique

```
inc5000$revenue <- as.numeric(gsub("[^0-9.]+", "", inc5000$revenue))
```

2.3 Vérification du type de la colonne “revenue”

```
class(inc5000$revenue)
```

```
[1] "numeric"
```

2.4 Vérifier s’il y a des valeurs NA dans le dataframe

```
if (any(is.na(inc5000))) {
  na_rows <- which(apply(is.na(inc5000), 1, any))
  print("Rows containing NA values:")
  count <- 0
  for (row in na_rows) {
    na_columns <- colnames(inc5000)[which(is.na(inc5000[row, ]))]
    print(paste("Row with rank", inc5000[row, "rank"],
```

```
    "contains NA values in columns:", paste(na_columns, collapse = ", ")))
    count <- count + 1
    if (count == 10) {
      break
    }
  }
} else {
  print("There are no NA values in the dataframe.")
}
```

```
[1] "Rows containing NA values:"
[1] "Row with rank 4 contains NA values in columns: metro"
[1] "Row with rank 21 contains NA values in columns: metro"
[1] "Row with rank 22 contains NA values in columns: metro"
[1] "Row with rank 26 contains NA values in columns: metro"
[1] "Row with rank 29 contains NA values in columns: metro"
[1] "Row with rank 33 contains NA values in columns: metro"
[1] "Row with rank 49 contains NA values in columns: metro"
[1] "Row with rank 62 contains NA values in columns: metro"
[1] "Row with rank 64 contains NA values in columns: metro"
[1] "Row with rank 76 contains NA values in columns: metro"
```

J'ai remarqué que la plupart des valeurs NA sont dans la colonne "metro".

Étape 1 : Vérifier le pourcentage de valeurs NA dans la colonne "metro"

```
na_count <- sum(is.na(inc5000$metro))
total_rows <- nrow(inc5000)
na_percentage <- na_count / total_rows * 100
print(na_percentage)
```

```
[1] 16.22107
```

```
if (na_percentage > 50) {
  print(na_percentage)
  inc5000 <- inc5000[, !names(inc5000) %in% "metro"]
  print("The 'metro' column has been removed due to exceeding 50% NA values.")
} else {
  print("The 'metro' column does not exceed 50% NA values.")
}
```

```
[1] "The 'metro' column does not exceed 50% NA values."
```

Le pourcentage est de 16,22 %. Cela signifie que nous pouvons conserver la colonne.

Nous devons la supprimer pour vérifier d'autres valeurs NA :

```
inc5000 <- inc5000[, !names(inc5000) %in% "metro"]
```

Ensuite, nous exécutons à nouveau l'étape 2.4.

```
if (any(is.na(inc5000))) {
  na_rows <- which(apply(is.na(inc5000), 1, any))
  print("Rows containing NA values:")
  for (row in na_rows) {
    na_columns <- colnames(inc5000)[which(is.na(inc5000[row, ]))]
    print(paste("Row with rank", inc5000[row, "rank"],
    "contains NA values in columns:", paste(na_columns, collapse = ", ")))
  }
} else {
  print("There are no NA values in the dataframe.")
}
```

```
[1] "Rows containing NA values:"
[1] "Row with rank 736 contains NA values in columns: profile"
[1] "Row with rank 3746 contains NA values in columns: workers"
```

Il y a deux lignes : une qui a une valeur NA dans "workers" et l'autre dans "profile". Nous n'utiliserons pas la colonne "profile" donc pas besoin de la supprimer, mais nous utiliserons "workers". Cependant, nous supprimerons la ligne lorsque nous l'utiliserons.

3 Analyse générale de la variable "revenue"

```
inc5000 <- read_csv("C:/Users/Administrator/Downloads/inc5000_2019.csv")
inc5000$revenue <- as.numeric(gsub("[^0-9.]+", "", inc5000$revenue))
```

3.1 Afficher les premières lignes avec les revenus les plus élevés

```
head(inc5000[order(inc5000$revenue, decreasing = TRUE), ])
```

```
# A tibble: 6 x 14
   rank profile        name  url   state revenue growth industry workers founded
  <dbl> <chr>          <chr> <chr> <chr>   <dbl>  <dbl> <chr>      <dbl>   <dbl>
1  3260 https://www.i~ Conv~ conv~ IL       991.  111.  Security    4112    2001
2  3496 https://www.i~ PURE~ pure~ NY       983.  100.  Insuran~     650    2006
3  3400 https://www.i~ Youn~ youn~ UT       886.  104.  Retail       864    1925
4  4493 https://www.i~ Pro ~ prom~ KY       872.   65.6 Manufac~    3087    1998
5  1314 https://www.i~ Nola~ ntgf~ GA       812.  314.  Logisti~     858    2005
6  4513 https://www.i~ Guar~ rate~ IL       803.   65.2 Financi~    4642    2000
# i 4 more variables: yrs_on_list <dbl>, previous_workers <dbl>, metro <chr>,
#   city <chr>
```

3.2 Analyse descriptive : analyse univariée de la variable "revenue"

```r
min(inc5000$revenue)
```

```
[1] 1
```

```r
max(inc5000$revenue)
```

```
[1] 990.6
```

```r
Q1 <- quantile(inc5000$revenue, 0.25)
Q2 <- quantile(inc5000$revenue, 0.5)
Q3 <- quantile(inc5000$revenue, 0.75)


print(paste("Q1 (25th percentile):", Q1))
```

```
[1] "Q1 (25th percentile): 4.7"
```

```r
print(paste("Q2 (50th percentile, median):", Q2))
```

```
[1] "Q2 (50th percentile, median): 10.3"
```

```r
print(paste("Q3 (75th percentile):", Q3))
```

```
[1] "Q3 (75th percentile): 26.5"
```

```r
mean(inc5000$revenue)
```

```
[1] 31.08659
```

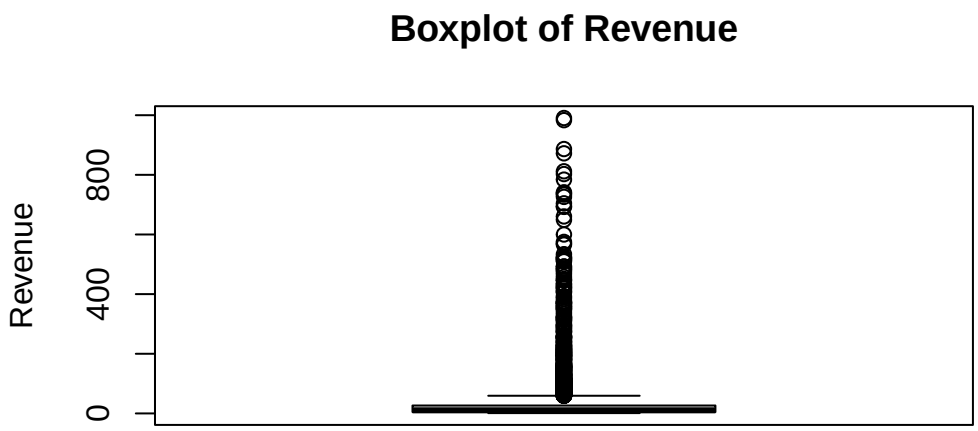
```r
var(inc5000$revenue)
```

```
[1] 5090.564
```

```r
sd(inc5000$revenue)
```

```
[1] 71.34818
```

```r
boxplot(inc5000$revenue, main = "Boxplot of Revenue", ylab = "Revenue")
```



Interpretation :

-Le maximum 990.6 et le minimum 1 sont très éloignés l'un de l'autre.

-La médiane est de 10,3, ce qui signifie que 50 % des revenus sont inférieurs à 10,3.

-25 % des revenus sont inférieurs à 4,7.

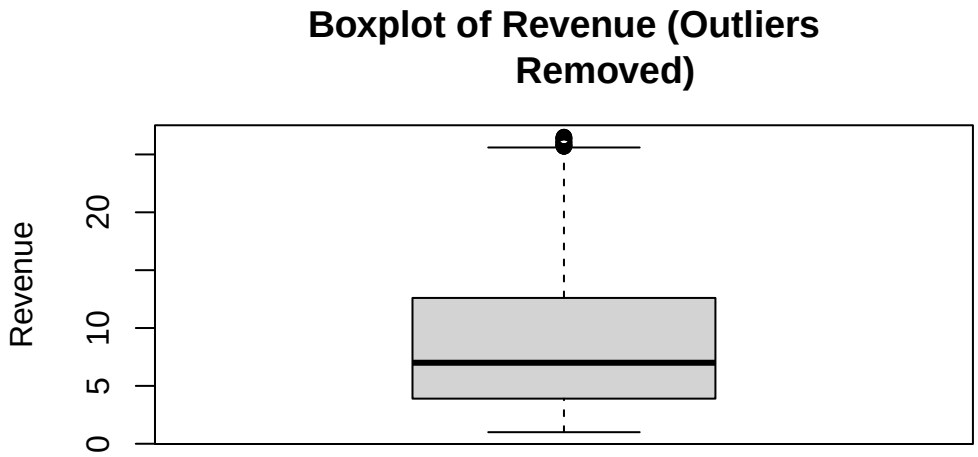
-75 % des revenus sont inférieurs à 26,5.

-L'écart type est de 71,34, ce qui est très élevé et cela montre que les valeurs des revenus sont très dispersées.

-Le diagramme en boîte montre qu'il y a trop de valeurs aberrantes

3.4 Nettoyage des valeurs aberrantes

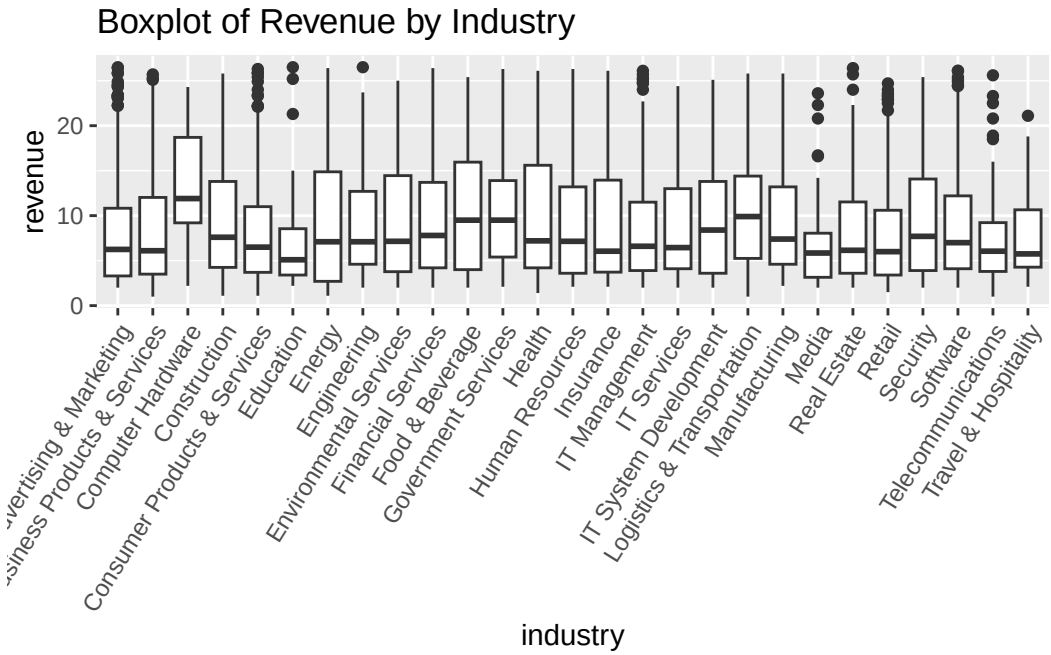
```
inc5000_revenue <- inc5000[inc5000$revenue <= Q3, ]

boxplot(inc5000_revenue$revenue, main = "Boxplot of Revenue (Outliers
        Removed)", ylab = "Revenue")
```



4.Revenue par industrie.

4.1 Créer un diagramme en boîte groupé du revenu par industrie

```
ggplot(inc5000_revenue, aes(x = industry, y = revenue)) +
  geom_boxplot() +
  labs(title = "Boxplot of Revenue by Industry") +
  theme(axis.text.x = element_text(angle = 58, hjust = 1))
```



-J'ai remarqué qu'il y a quelques valeurs aberrantes qui devraient être supprimées car elles ne nous permettent pas de faire des statistiques correctes. Pour ce faire, nous devons vérifier l'écart-type du revenu par industrie et les valeurs correspondantes.

4.2 Nettoyage des valeurs aberrantes

```
inc5000_bound <- inc5000_revenue %>%
  group_by(industry) %>%
  mutate(
    upper_bound = quantile(revenue, 0.75)
  ) %>%
  ungroup()


inc5000_clean <- inc5000_bound %>%
  filter(revenue <= upper_bound)

inc5000_clean <- inc5000_clean %>%
  filter(!(rank %in% c(219,4322,2292,1344,1878,3923)))
```

4.3 Calculer les statistiques sommaires pour le revenu par industrie.

```r
summary_stats_clean <- inc5000_clean %>%
  group_by(industry) %>%
  summarize(min = min(revenue),
            max = max(revenue),
            mean = mean(revenue),
            q1 = quantile(revenue, probs = 0.25),
            median = median(revenue),
            q3 = quantile(revenue, probs = 0.75),
            variance = var(revenue),
            sd = sd(revenue)) %>%
  arrange(desc(mean))


print(summary_stats_clean, n = 27)
```

```
# A tibble: 27 x 9
   industry                   min   max  mean    q1 median    q3 variance    sd
   <chr>                    <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>    <dbl> <dbl>
 1 Computer Hardware          2.2  18.2 10.1   7.5   10.9  12.0     16.8   4.10
 2 Logistics & Transportati~  1    14.3  7.64  4.2    8.1  10.6     14.5   3.80
 3 Government Services        2.1  13.9  7.50  4.5    7.2  10.2     12.0   3.47
 4 Food & Beverage            2    15.9  7.39  3.3    6.4  11.1     19.8   4.44
 5 IT System Development      2    13.8  6.69  3.4    6.8   9.5     12.7   3.56
 6 Health                     1.4  15.6  6.65  3.3    5.35  8.88    15.9   3.98
 7 Environmental Services     2    14.4  6.56  3.53   5.85  9.8     15.0   3.87
 8 Financial Services         2    13.7  6.35  3.5    5.45  9       10.8   3.29
 9 Construction               1.1  13.7  6.34  3.6    5.6   9       10.4   3.22
10 Manufacturing              2.2  13.2  6.34  4      5.75  7.9      7.76  2.78
11 Security                   2    14    6.21  3.68   4.9   8.6     12.7   3.56
12 Engineering                2    12.7  6.14  3.4    5.7   8.5      9.39  3.06
13 Human Resources            2.1  13.1  6.07  3.2    5.6   8.3     10.6   3.26
14 Software                   2    12.2  5.88  3.7    5.1   7.9      7.70  2.78
15 Energy                     1.1  13.3  5.70  2.4    5.25  9.02    12.0   3.47
16 IT Management              2    11.4  5.53  3.2    5.3   7.25     6.65  2.58
17 Real Estate                2    11.5  5.48  3.1    5     7.3      7.39  2.72
18 Consumer Products & Serv~  1.1  11    5.46  3      4.8   7.43     7.23  2.69
19 IT Services                2    12.3  5.44  3.45   4.8   7.3      6.94  2.64
20 Business Products & Serv~  1    12    5.39  3.1    4.55  7.2      7.83  2.80
21 Advertising & Marketing    2    10.8  5.10  2.9    4.7   7        6.07  2.46
22 Insurance                  2.1  11.1  5.07  2.7    4.9   6.3      6.26  2.50
23 Retail                     1.5  10.6  5.00  3.1    4.5   6.57     5.45  2.33
24 Telecommunications         1     8.9  4.91  3.1    4.5   7        4.78  2.19
```

```
25 Travel & Hospitality        2.1   7.8  4.85  3.92   4.75  5.88     2.42  1.56
26 Education                   2.2   8.4  4.45  2.85   4     5.4      3.39  1.84
27 Media                       2     7.9  4.43  2.45   4     6.05     4.11  2.03
```

4.4 Trouver le groupe de revenus le plus élevé par industrie

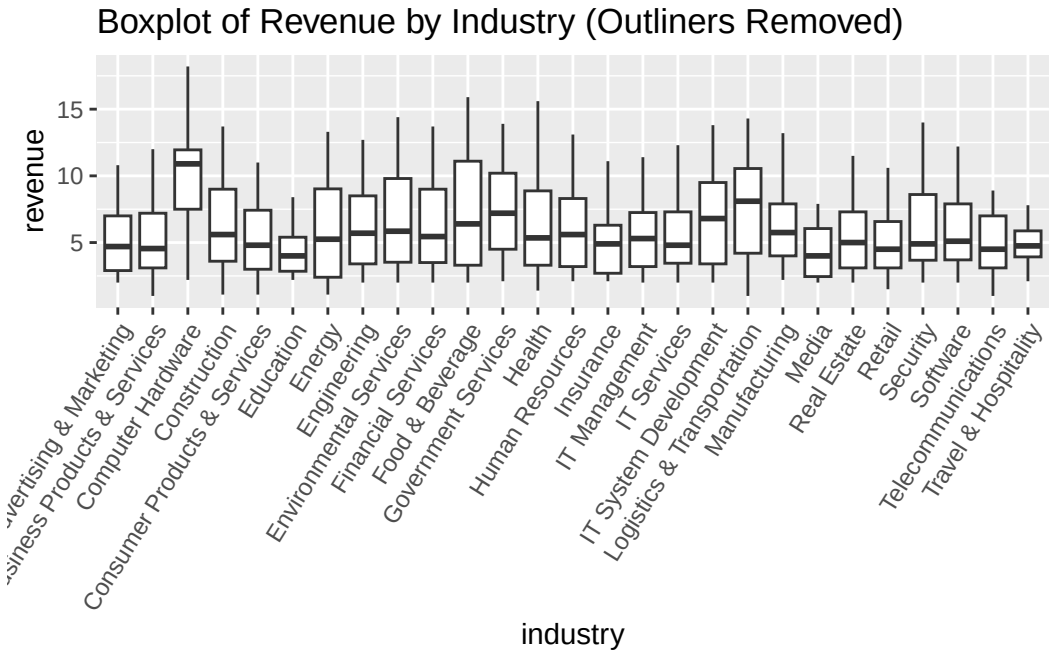
```
top_revenue <- inc5000_clean %>%
  group_by(industry) %>%
  slice_max(order_by = revenue, n = 2)

print(top_revenue,n=66)
```

```
# A tibble: 66 x 15
# Groups:   industry [27]
    rank profile       name  url   state revenue growth industry workers founded
   <dbl> <chr>         <chr> <chr> <chr>   <dbl>  <dbl> <chr>      <dbl>   <dbl>
 1  2036 https://www.~ Mara~ http~ NY       10.8  198.  Adverti~      41    2008
 2  3374 https://www.~ eROI  eroi~ OR       10.8  106.  Adverti~      83    2002
 3  2885 https://www.~ Boom~ boom~ UT       12    130.  Busines~     645    2007
 4  3854 https://www.~ memo~ memo~ VA       12     86.3 Busines~     140    2002
 5  3188 https://www.~ King~ king~ OH       18.2  115.  Compute~      58    2009
 6  4761 https://www.~ Mobi~ rugg~ IA       15.9   59.0 Compute~      40    2003
 7  2302 https://www.~ D&B ~ dand~ PA       13.7  175.  Constru~      25    2010
 8  2284 https://www.~ Emer~ erx2~ NC       12.8  176.  Constru~      90    2013
 9   995 https://www.~ Noma~ noma~ UT       11    425.  Consume~      20    2014
10   771 https://www.~ Skar~ skar~ FL       10.9  561.  Consume~      11    2012
11  2654 https://www.~ Sell~ sell~ OH       10.9  146.  Consume~      24    2009
12  3431 https://www.~ Grow~ grow~ AZ        8.4  103.  Educati~      79    2014
13  2909 https://www.~ Deve~ http~ CO        8.1  129.  Educati~      20    2003
14  4553 https://www.~ Educ~ educ~ PA        8.1   64.1 Educati~      31    2002
15   162 https://www.~ Kine~ http~ TX       13.3 2358.  Energy         8    2015
16  1949 https://www.~ EnTo~ ento~ TX       11.7  206.  Energy        42    2008
17  4454 https://www.~ Crit~ crit~ NY       12.7   66.6 Enginee~      43    1997
18   880 https://www.~ Parr~ parr~ SC       12.1  485.  Enginee~      63    2013
19  1948 https://www.~ SRP ~ srpe~ LA       14.4  206.  Environ~      56    1996
20  2825 https://www.~ The ~ junk~ CT       13    133.  Environ~      62    2016
21    97 https://www.~ Paya~ http~ NY       13.7 3350.  Financi~      54    2014
22  2777 https://www.~ Rain~ rain~ WA       13.7  136.  Financi~      33    2009
23   442 https://www.~ Upti~ http~ CA       15.9 1031.  Food & ~      49    1985
24  1048 https://www.~ High~ high~ GA       15.3  399.  Food & ~      45    2010
25  2901 https://www.~ Sust~ sust~ OR       15.3  129.  Food & ~     305    2008
26  4540 https://www.~ Doub~ doub~ IL       15.3   64.5 Food & ~      54    1998
```

```
27  1305 https://www.~ G2S   g2sc~ TX       13.9  315.  Governm~      84    2009
28  1380 https://www.~ Lent~ lent~ NM       13.9  300.  Governm~      90    2008
29  2044 https://www.~ Oasys oasy~ DC       13.9  197.  Governm~      84    2011
30  4466 https://www.~ Inte~ inte~ VA       13.9   66.3 Governm~      85    2002
31  2535 https://www.~ Nuve~ nuve~ NC       15.6  155.  Health        56    2008
32   167 https://www.~ Tami~ http~ FL       15.5 2314.  Health        40    2013
33  3448 https://www.~ Spar~ spar~ PA       15.5  102.  Health        88    2010
34  1328 https://www.~ Veri~ http~ ID       13.1  311.  Human R~     108    2013
35  4026 https://www.~ Lewi~ lewi~ VA       12.8   80.5 Human R~      99    2003
36   756 https://www.~ Cira~ cira~ GA       11.4  577.  IT Mana~      92    2005
37  3382 https://www.~ Cerd~ cerd~ OH       11.3  105.  IT Mana~      49    2002
38  4118 https://www.~ Inte~ inte~ GA       11.3   77.4 IT Mana~      85    1980
39  4660 https://www.~ The ~ copl~ MA       11.3   61.3 IT Mana~      45    1989
40  4495 https://www.~ Kenw~ http~ IL       12.3   65.6 IT Serv~      52    2004
41  3946 https://www.~ 2020~ 2020~ VA        9.8   83.1 IT Serv~      24    2012
42   769 https://www.~ Inte~ inte~ CA       13.8  563.  IT Syst~     255    2007
43  1873 https://www.~ Agil~ http~ CA       13.8  216.  IT Syst~     250    2013
44  2265 https://www.~ ITI ~ itic~ NC       13.8  177.  IT Syst~      12    2010
45  3892 https://www.~ Lega~ lega~ TX       11.1   85.0 Insuran~      60    2003
46  1021 https://www.~ Reli~ reli~ TN       10.1  413.  Insuran~     118    2009
47  1644 https://www.~ Ecog~ ecog~ IL       14.3  251.  Logisti~      25    2009
48   465 https://www.~ Vonl~ vonl~ TX       14.2  967.  Logisti~     103    2013
49  3403 https://www.~ Hern~ hern~ FL       13.2  104.  Manufac~      77    1978
50  1650 https://www.~ Floo~ floo~ TX       12.5  249.  Manufac~      12    2001
51  4730 https://www.~ Dciny http~ NY        7.9   59.6 Media         22    2007
52  4383 https://www.~ Cana~ cana~ CA        7.6   68.6 Media         24    2010
53  4576 https://www.~ OPAV  op-a~ FL        7.6   63.5 Media         49    2008
54  1605 https://www.~ Fult~ fult~ IL       11.5  256.  Real Es~      49    2008
55  3101 https://www.~ Lama~ lama~ MA       11.1  119.  Real Es~      27    2005
56  4098 https://www.~ Corv~ corv~ TX       10.6   77.8 Retail        35    2010
57  3385 https://www.~ Maxt~ maxt~ CA       10.4  105.  Retail         7    2011
58  3803 https://www.~ Rapi~ thes~ FL       14     88.1 Security      35    2009
59   434 https://www.~ Kenn~ kenn~ CA       13.2 1045.  Security     130    2011
60  1121 https://www.~ Nuvo~ nuvo~ NJ       12.2  371.  Software    131    2014
61  1464 https://www.~ Dock~ http~ MA       12.2  283.  Software     70    2010
62  2990 https://www.~ Infr~ infr~ IN       12.2  125.  Software    117    2003
63  1526 https://www.~ TekC~ tekc~ PA        8.9  271.  Telecom~      13    2005
64  4999 https://www.~ HNM ~ hnms~ CA        8.8   52.2 Telecom~     132    2011
65  2384 https://www.~ Abod~ abod~ UT        7.8  168.  Travel ~      17    2010
66  1722 https://www.~ Trav~ trav~ FL        7.5  236.  Travel ~      35    2013
# i 5 more variables: yrs_on_list <dbl>, previous_workers <dbl>, metro <chr>,
#   city <chr>, upper_bound <dbl>
```

-Si nous exécutons à nouveau, nous remarquerons qu'il y a moins de valeurs aberrantes et que les statistiques sont meilleures.

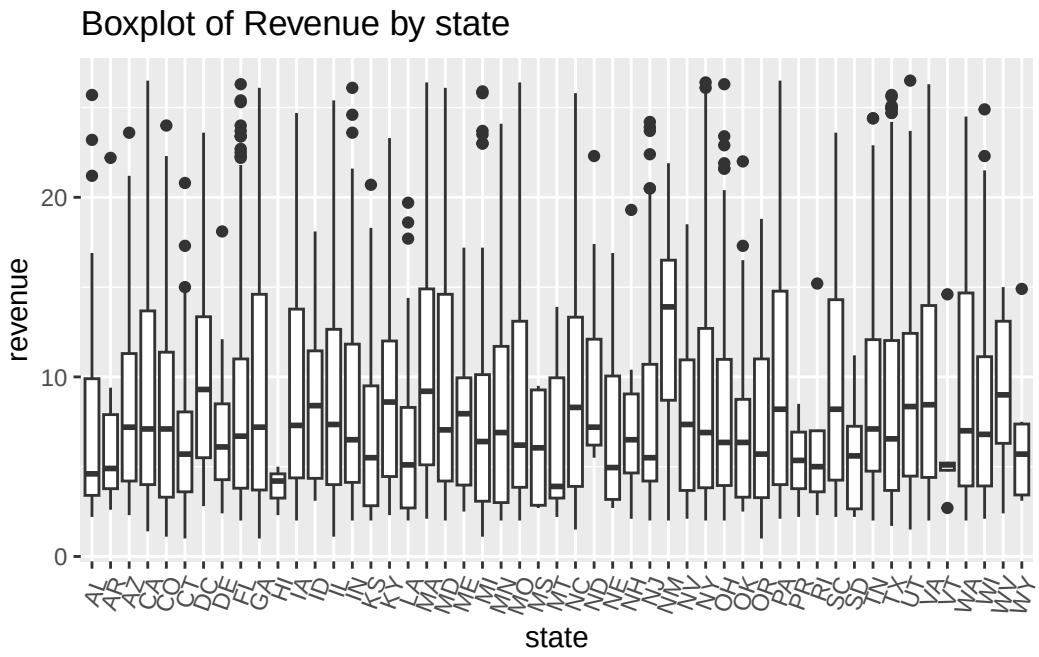
```
ggplot(inc5000_clean, aes(x = industry, y = revenue)) +
  geom_boxplot() +
  labs(title = "Boxplot of Revenue by Industry (Outliners Removed)") +
  theme(axis.text.x = element_text(angle = 58, hjust = 1))
```



5.Revenue par état

5.1 Créer un diagramme en boîte groupé du revenu par état

```
ggplot(inc5000_revenue, aes(x = state, y = revenue)) +
  geom_boxplot() +
  labs(title = "Boxplot of Revenue by state") +
  theme(axis.text.x = element_text(angle = 70, hjust = 1))
```



-J'ai remarqué qu'il y a quelques valeurs aberrantes qui devraient être supprimées car elles ne nous permettent pas de faire des statistiques correctes. Pour ce faire, nous devons vérifier l'écart-type du revenu par industrie et les valeurs correspondantes.

5.2Nettoyage des valeurs aberrantes

```
inc5000_bound <- inc5000_revenue %>%
  group_by(state) %>%
  mutate(
    upper_bound = quantile(revenue, 0.75)
  ) %>%
  ungroup()

inc5000_clean <- inc5000_bound %>%
  filter(revenue <= upper_bound)

inc5000_clean <- inc5000_clean %>%
  filter(!(rank %in% c(2170,2533,3488,1057,2748,2487,189)))
```

5.3 Calculer les statistiques sommaires pour le revenu par état.

```
summary_stats <- inc5000_clean %>%
  group_by(state) %>%
```

```
  summarize(min = min(revenue),
            max = max(revenue),
            mean = mean(revenue),
            q1 = quantile(revenue, probs = 0.25),
            median = median(revenue),
            q3 = quantile(revenue, probs = 0.75),
            variance = var(revenue),
            sd = sd(revenue)) %>%
  arrange(desc(mean))

print(summary_stats,n=51)
```

```
# A tibble: 51 x 9
   state   min   max  mean    q1 median    q3 variance     sd
   <chr> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>    <dbl>  <dbl>
 1 NM      2    16.5 10.3   7.02  11.3  14.6    40.9    6.40
 2 MA      2.1  14.9  7.73  4.18   7.2  11.1    15.2    3.90
 3 WV      2.4  13.1  7.7   5.32   7.65 10.0    20.3    4.51
 4 DC      2.8  12.8  7.42  4.68   7.4  10.2    10.6    3.25
 5 PA      2.1  14.4  6.77  3.7    6.3   9.8    13.5    3.67
 6 VA      2    13.9  6.75  3.85   5.7   9.85   12.1    3.48
 7 KY      2.3  11.7  6.68  3.8    7.3   9.2     9.14   3.02
 8 SC      2.2  13.7  6.63  3      5.9   9.18   13.8    3.72
 9 UT      1.5  12.3  6.61  3.8    6     9      10.2    3.19
10 WA      2    14    6.32  3.4    4.8   8.95   13.4    3.66
11 NC      1.5  12.8  6.21  3      5.45  8.85   11.4    3.38
12 ID      3.1   9.8  6.21  3.4    6.6   9       6.90   2.63
13 MD      2    14.5  6.11  3.55   5     8.8    11.2    3.35
14 TN      2    11.7  6.11  3.6    6     8.3     7.87   2.80
15 GA      1    14.6  6.10  3.3    5.3   7.9    11.2    3.35
16 CA      1.4  13.6  6.02  3.3    5.5   8.1     9.95   3.15
17 IL      1.1  12.5  5.98  3.4    5.35  8.25    9.41   3.07
18 ND      5.5   6.2  5.92  5.72   6     6.2     0.116  0.340
19 AZ      2.3  11.3  5.92  3.5    5.4   8.17    7.24   2.69
20 IA      2    11    5.91  2.88   5.35  7.9    10.1    3.18
21 NY      2    12.7  5.85  3.3    5.2   7.7     8.52   2.92
22 ME      2.5   9.8  5.76  3.65   4.8   7.95    7.74   2.78
23 MO      2    12.7  5.64  3.2    5     7.42    8.30   2.88
24 WI      2.1  11.1  5.63  3.55   5.4   7.05    6.81   2.61
25 CO      1.1  11.3  5.61  2.85   5     8.2     8.85   2.97
26 TX      1.7  12    5.61  3.2    5     7.52    7.89   2.81
27 IN      2    11.6  5.59  3.8    4.8   7.5     8.43   2.90
```

```
28 NV      2.1  10.5  5.56  3.1   4.2   8.7    8.83   2.97
29 FL      2    11    5.45  3.1   5.05  7.5    6.67   2.58
30 NH      2.1   9    5.45  3.9   5.9   6.7    5.04   2.25
31 OH      2    10.9  5.43  3.45  4.7   7.1    6.57   2.56
32 NJ      2    10.7  5.22  3.5   4.9   6.5    4.87   2.21
33 MN      2    11.7  5.18  2.7   4.1   7.62   8.35   2.89
34 OK      2.5   8.7  5.06  3.15  5     7      4.33   2.08
35 DE      2.4   7.3  5     3.82  4.9   6.5    3.59   1.90
36 MS      2.7   9.2  4.93  2.8   2.9   6.05  13.7    3.70
37 MI      1.1  10    4.79  2.8   3.6   7.2    6.75   2.60
38 OR      1    10.8  4.60  2.4   4.25  6.15   5.24   2.29
39 CT      1     7.8  4.60  3.15  4.55  6      4.02   2.00
40 AR      2.6   7.4  4.55  3.73  4     5.25   2.89   1.70
41 RI      2.3   7    4.47  3.28  4.3   5.5    4.05   2.01
42 WY      3.1   7    4.4   3.1   3.75  5.05   3.38   1.84
43 KS      2     9.2  4.38  2.3   4     5.5    5.24   2.29
44 NE      2.7   7.5  4.26  3     3.6   5.4    2.69   1.64
45 LA      2     8.3  4.23  2.62  3.8   5.62   3.79   1.95
46 VT      2.7   5.1  4.2   3.75  4.8   4.95   1.71   1.31
47 SD      2.2   6.3  4.15  2.55  3.9   5.85   3.59   1.89
48 AL      2.2   7.7  4.14  3.1   3.6   5      2.25   1.50
49 MT      2.2   5.3  3.49  2.8   3.5   3.9    1.05   1.03
50 HI      2.3   4.2  3.25  2.78  3.25  3.73   1.81   1.34
51 PR      2.2   2.2  2.2   2.2   2.2   2.2    NA     NA
```

5.4 Trouver le groupe de revenus le plus élevé par état.

```r
top_revenue <- inc5000_clean %>%
  group_by(state) %>%
  slice_max(order_by = revenue, n = 2)

print(top_revenue,n=510)
```

```
# A tibble: 110 x 15
# Groups:   state [51]
     rank profile      name  url   state revenue growth industry workers founded
    <dbl> <chr>        <chr> <chr> <chr>   <dbl>  <dbl> <chr>      <dbl>   <dbl>
  1  1802 https://www~ Alta~ alta~ AL        7.7  226.  Telecom~      19    2001
  2  1538 https://www~ Wave~ wave~ AL        7.6  269.  Telecom~      40    2005
  3  3299 https://www~ Elit~ elit~ AR        7.4  109.  Manufac~      42    2014
  4  2600 https://www~ Trin~ trin~ AR        5.6  150.  Real Es~     419    2004
  5  2738 https://www~ Nort~ apsp~ AZ       11.3  140.  Financi~      48    2006
```

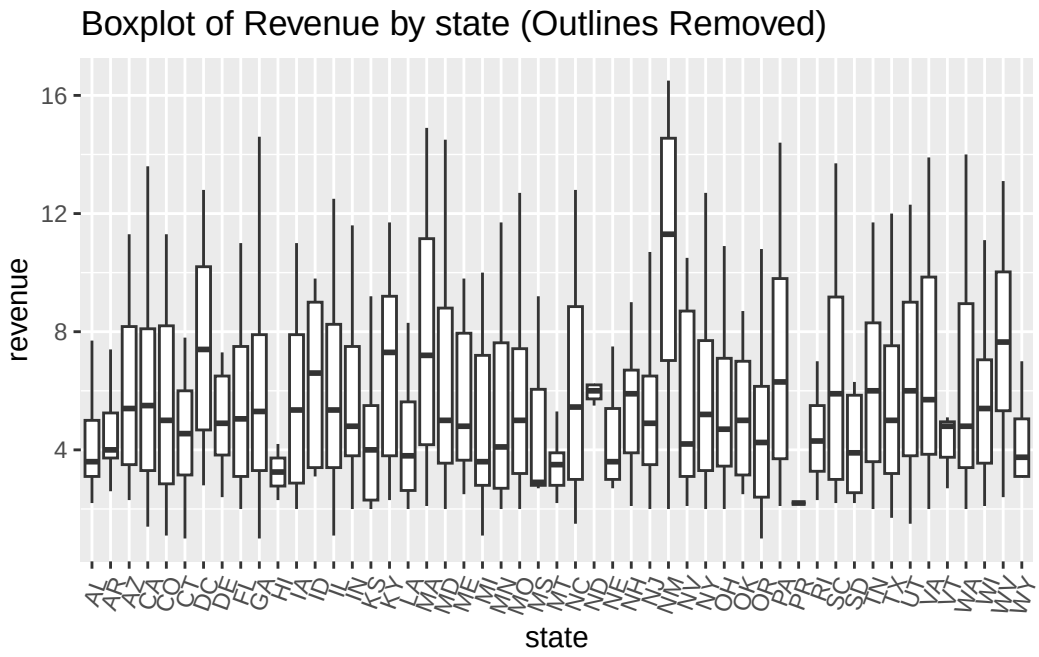
6	4199	https://www~	B2Gn~	b2gn~	AZ	11.3	74.2	Software	60	1999
7	537	https://www~	Icon~	http~	CA	13.6	816.	Logisti~	39	2013
8	642	https://www~	Bott~	http~	CA	13.6	690.	Retail	8	2013
9	1060	https://www~	Bree~	bree~	CA	13.6	395.	IT Mana~	18	2010
10	3440	https://www~	Alo ~	alor~	CA	13.6	102.	Health	90	2013
11	2043	https://www~	Bigg~	bigg~	CO	11.3	197.	Educati~	28	2004
12	2091	https://www~	Orth~	orth~	CO	11.1	193.	Software	214	2013
13	3997	https://www~	Tend~	tend~	CO	11.1	81.2	Food & ~	22	2010
14	2744	https://www~	Choi~	choi~	CT	7.8	139.	Financi~	18	2008
15	4905	https://www~	In S~	inst~	CT	7.4	55.2	Adverti~	25	2006
16	385	https://www~	Urba~	urba~	DC	12.8	1192.	Consume~	90	2014
17	1071	https://www~	Blue~	blue~	DC	12.2	392.	Governm~	70	2011
18	3546	https://www~	Kari~	kari~	DE	7.3	98.8	Enginee~	52	1973
19	1301	https://www~	Home~	home~	DE	6.9	317.	Constru~	50	2013
20	1847	https://www~	Base4	base~	FL	11	220.	Enginee~	275	2009
21	2130	https://www~	Lodg~	lodg~	FL	11	189.	Busines~	19	2002
22	2895	https://www~	How ~	http~	FL	11	129.	Busines~	32	2009
23	1397	https://www~	Stru~	stru~	GA	14.6	296.	Manufac~	51	2011
24	4219	https://www~	Arke~	arke~	GA	14.5	73.4	Adverti~	70	2005
25	2766	https://www~	Maui~	coco~	HI	4.2	137.	Travel ~	9	2011
26	2042	https://www~	The ~	gole~	HI	2.3	197.	Educati~	4	2012
27	2189	https://www~	Eagl~	eagl~	IA	11	184.	Energy	56	2010
28	2949	https://www~	Scha~	call~	IA	10.6	127.	Constru~	47	1956
29	2800	https://www~	Prec~	ppe-~	ID	9.8	135.	Constru~	20	2008
30	297	https://www~	Doll~	doll~	ID	9.1	1508.	Logisti~	15	2008
31	4744	https://www~	Impe~	impe~	IL	12.5	59.4	Busines~	52	1973
32	342	https://www~	Wavi~	wavi~	IL	12.4	1347.	IT Syst~	171	2013
33	3505	https://www~	Clea~	clea~	IN	11.6	100.	Environ~	98	2001
34	2220	https://www~	Nix ~	nixc~	IN	11.3	181.	Manufac~	63	1902
35	147	https://www~	Seth~	seth~	KS	9.2	2540.	Logisti~	18	1999
36	3355	https://www~	Chel~	chel~	KS	6.9	107.	IT Mana~	75	2010
37	4691	https://www~	LSS ~	life~	KY	11.7	60.5	Busines~	100	2004
38	1913	https://www~	Unit~	unit~	KY	10.6	209.	Constru~	200	2012
39	686	https://www~	Emer~	emer~	LA	8.3	635.	Busines~	29	2012
40	2322	https://www~	Leve~	http~	LA	7.4	173.	Software	143	2007
41	2485	https://www~	Koya~	koya~	MA	14.9	159.	Human R~	67	2004
42	3422	https://www~	Clin~	clin~	MA	14.9	103.	Human R~	170	2002
43	4812	https://www~	Cons~	cse-~	MD	14.5	57.8	Governm~	85	2002
44	2160	https://www~	Ripp~	ripp~	MD	13.5	186.	Governm~	136	2003
45	3549	https://www~	Sea ~	seab~	ME	9.8	98.7	Busines~	110	1999
46	2264	https://www~	Veri~	http~	ME	8.1	177.	Software	42	2013
47	3605	https://www~	Supe~	gosl~	MI	10	97.0	Logisti~	20	2009
48	4530	https://www~	ZZPe~	zzpe~	MI	9.8	64.8	Consume~	51	2001

49	1033	https://www~	Nimb~	nimb~	MN	11.7	407.	Telecom~	18	2013
50	4838	https://www~	Soft~	soft~	MN	11.6	57.0	IT Mana~	210	2000
51	1566	https://www~	Paci~	gemi~	MO	12.7	264.	IT Syst~	400	2012
52	3814	https://www~	Gade~	gade~	MO	11.8	87.8	IT Mana~	85	2003
53	3734	https://www~	Verg~	verg~	MS	9.2	91.4	Software	49	2006
54	4024	https://www~	SMG	http~	MS	2.9	80.6	Health	9	2010
55	553	https://www~	XY P~	xypl~	MT	5.3	804.	Financi~	33	2014
56	4587	https://www~	Adva~	aed.~	MT	3.9	63.2	Enginee~	24	1994
57	4665	https://www~	Gold~	gold~	MT	3.9	61.1	Software	18	1998
58	2284	https://www~	Emer~	erx2~	NC	12.8	176.	Constru~	90	2013
59	1756	https://www~	Neo ~	neop~	NC	12.6	231.	IT Syst~	125	2011
60	1480	https://www~	Infi~	infi~	ND	6.2	279.	IT Mana~	31	2011
61	2515	https://www~	Myri~	myri~	ND	6.2	157.	Software	87	2011
62	2700	https://www~	Berr~	jsbe~	NE	7.5	143.	Consume~	60	1965
63	3124	https://www~	Bulu	bulu~	NE	7.1	118.	Consume~	189	2012
64	3626	https://www~	Orio~	orio~	NH	9	96.3	Manufac~	38	2009
65	1789	https://www~	Halo~	http~	NH	8.4	228.	Security	14	2008
66	4626	https://www~	The ~	thec~	NJ	10.7	62.1	Consume~	65	2009
67	2432	https://www~	JSK ~	http~	NJ	10.6	164.	Logisti~	14	2012
68	2979	https://www~	APIC~	apic~	NM	16.5	126.	Constru~	132	2012
69	1380	https://www~	Lent~	lent~	NM	13.9	300.	Governm~	90	2008
70	2018	https://www~	Nort~	nort~	NV	10.5	199.	Educati~	100	1997
71	2312	https://www~	Real~	real~	NV	10.1	174.	Consume~	70	2002
72	4440	https://www~	MyJo~	myjo~	NY	12.7	66.9	Busines~	13	2012
73	4454	https://www~	Crit~	crit~	NY	12.7	66.6	Enginee~	43	1997
74	2654	https://www~	Sell~	sell~	OH	10.9	146.	Consume~	24	2009
75	4643	https://www~	TruT~	trut~	OH	10.9	61.7	Busines~	12	2007
76	4780	https://www~	Prio~	prio~	OH	10.9	58.6	Busines~	65	1990
77	2997	https://www~	Lind~	lind~	OK	8.7	124.	Adverti~	13	1999
78	4030	https://www~	Host~	host~	OK	7.5	80.5	IT Mana~	46	1998
79	3374	https://www~	eROI	eroi~	OR	10.8	106.	Adverti~	83	2002
80	580	https://www~	King~	king~	OR	10	764.	Financi~	23	2008
81	1418	https://www~	Rese~	rese~	PA	14.4	291.	Adverti~	300	2009
82	2531	https://www~	Clut~	clut~	PA	14.2	155.	Software	50	2012
83	341	https://www~	Bidw~	http~	PR	2.2	1354.	Adverti~	7	2015
84	3903	https://www~	Trib~	trib~	RI	7	84.7	Adverti~	63	2010
85	3011	https://www~	Prot~	prot~	RI	5	123.	IT Mana~	28	2000
86	283	https://www~	Madd~	http~	SC	13.7	1553.	Manufac~	23	2015
87	1885	https://www~	Nati~	nati~	SC	12.6	213.	Real Es~	23	2007
88	1298	https://www~	Prai~	prai~	SD	6.3	317.	Retail	3	2010
89	2462	https://www~	Mark~	mark~	SD	6.1	161.	Media	6	2008
90	2073	https://www~	Prov~	prov~	TN	11.7	194.	Software	68	2010
91	4792	https://www~	Clar~	clar~	TN	11.4	58.2	Security	465	2009

```
 92   234 https://www~ Simp~ simp~ TX       12   1790.  Compute~      44    2015
 93  3007 https://www~ Firs~ firs~ TX       12    124.  Real Es~      36    2000
 94  3716 https://www~ Tach~ tach~ TX       12     92.4 IT Syst~     120    2011
 95  4225 https://www~ Appl~ appl~ UT       12.3   73.2 Human R~     120    2006
 96  3679 https://www~ Utop~ utop~ UT       12.2   94.3 Travel ~      40    2011
 97  4466 https://www~ Inte~ inte~ VA       13.9   66.3 Governm~      85    2002
 98   962 https://www~ Sava~ sava~ VA       13.8  441.  Governm~      67    2006
 99  4414 https://www~ Wher~ wher~ VA       13.8   67.6 Adverti~     103    1999
100  3405 https://www~ Jama~ jama~ VT        5.1  104.  Manufac~      50    1995
101  4987 https://www~ iMar~ imar~ VT        4.8   52.6 Adverti~      40    2010
102  2763 https://www~ Vita~ vita~ WA       14    137.  Health       105    2011
103  2777 https://www~ Rain~ rain~ WA       13.7  136.  Financi~      33    2009
104  3446 https://www~ RedC~ redc~ WA       13.7  102.  IT Mana~      71    1995
105  4875 https://www~ Exte~ exte~ WI       11.1   56.1 Human R~      34    2002
106  3744 https://www~ Chan~ chan~ WI       10.4   90.9 Governm~       8    2002
107  2443 https://www~ Adva~ adva~ WV       13.1  163.  IT Mana~      83    2004
108  1516 https://www~ Moun~ ms3-~ WV        9    273.  IT Syst~      51    2009
109   579 https://www~ Brea~ brea~ WY        7    765.  Adverti~       8    2014
110   211 https://www~ Book~ lear~ WY        4.4 1940.  Educati~       8    2015
# i 5 more variables: yrs_on_list <dbl>, previous_workers <dbl>, metro <chr>,
#   city <chr>, upper_bound <dbl>
```

-Si nous exécutons à nouveau, nous remarquerons qu'il y a moins de valeurs aberrantes et que les statistiques sont meilleures.

```r
ggplot(inc5000_clean, aes(x = state, y = revenue)) +
  geom_boxplot() +
  labs(title = "Boxplot of Revenue by state (Outlines Removed)") +
  theme(axis.text.x = element_text(angle = 70, hjust = 1))
```



6 Analyse générale de la variable "growth"

```
inc5000 <- read_csv("C:/Users/Administrator/Downloads/inc5000_2019.csv")
inc5000$revenue <- as.numeric(gsub("[^0-9.]+", "", inc5000$revenue))
```

6.1 Afficher les premières lignes avec la croissance la plus élevée

```
head(inc5000)
```

```
# A tibble: 6 x 14
   rank profile        name  url   state revenue growth industry workers founded
  <dbl> <chr>          <chr> <chr> <chr>   <dbl>  <dbl> <chr>       <dbl>   <dbl>
1     1 https://www.i~ Free~ http~ AZ       36.9 36680. Adverti~      40    2015
2     2 https://www.i~ Frei~ http~ TN       33.6 30548. Logisti~      39    2015
3     3 https://www.i~ Cece~ http~ TX       24.9 23880. Food & ~     190    2015
4     4 https://www.i~ Lady~ http~ NM       32.4 21850. Consume~      57    2014
5     5 https://www.i~ Perp~ http~ PA       22.5 18166. Retail        25    2014
6     6 https://www.i~ Cano~ http~ FL      272.  14183. Health       742    2009
# i 4 more variables: yrs_on_list <dbl>, previous_workers <dbl>, metro <chr>,
#   city <chr>
```

6.2 Analyse descriptive : analyse univariée de la variable "growth"

```r
min(inc5000$growth)
```

```
[1] 52.1691
```

```r
max(inc5000$growth)
```

```
[1] 36680.39
```

```r
mean(inc5000$growth)
```

```
[1] 454.6801
```

```r
var(inc5000$growth)
```

```
[1] 1649397
```

```r
sd(inc5000$growth)
```

```
[1] 1284.289
```

```r
Q1 <- quantile(inc5000$growth, 0.25)
Q2 <- quantile(inc5000$growth, 0.5)
Q3 <- quantile(inc5000$growth, 0.75)

print(paste("Q1 (25th percentile):", Q1))
```

```
[1] "Q1 (25th percentile): 90.5625"
```

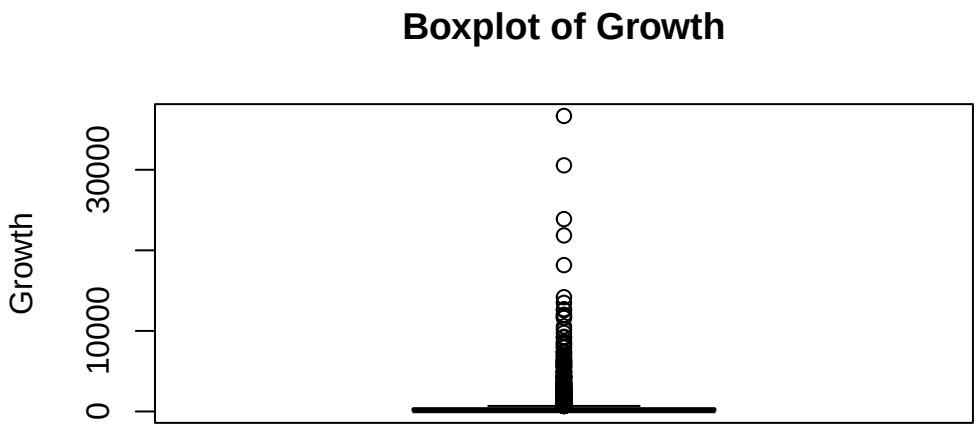
```r
print(paste("Q2 (50th percentile, median):", Q2))
```

```
[1] "Q2 (50th percentile, median): 157.53065"
```

```r
print(paste("Q3 (75th percentile):", Q3))
```

```
[1] "Q3 (75th percentile): 330.42725"
```

```
boxplot(inc5000$growth, main = "Boxplot of Growth", ylab = "Growth")
```



Interprétation:

-Le maximum 36680.39 et le minimum 52.1691 sont très éloignés l'un de l'autre.

-La médiane est de 157.5306, ce qui signifie que 50 % des revenus sont inférieurs à 157.5306.

-25 % des revenus sont inférieurs à 90.5625.

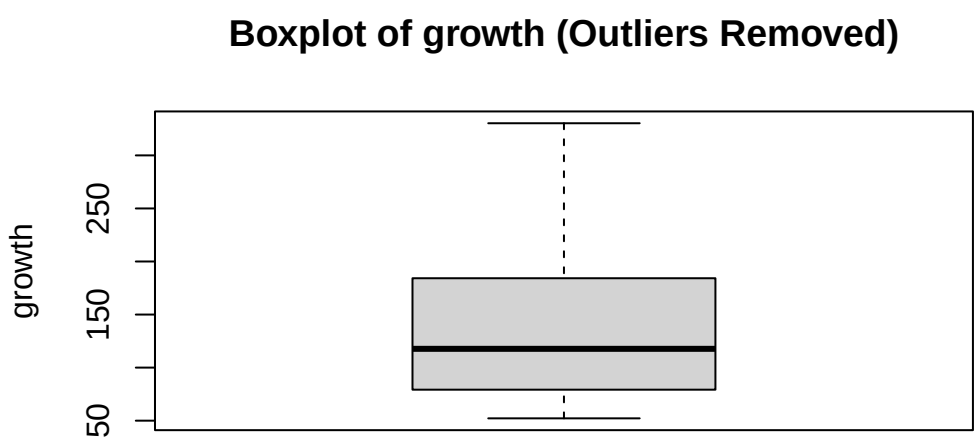
-75 % des revenus sont inférieurs à 330.42725.

-L'écart type est de 1284.289, ce qui est très élevé et cela montre que les valeurs des revenus sont très dispersées.

-Le diagramme en boîte montre qu'il y a trop de valeurs aberrantes

6.3 Nettoyage des valeurs aberrantes

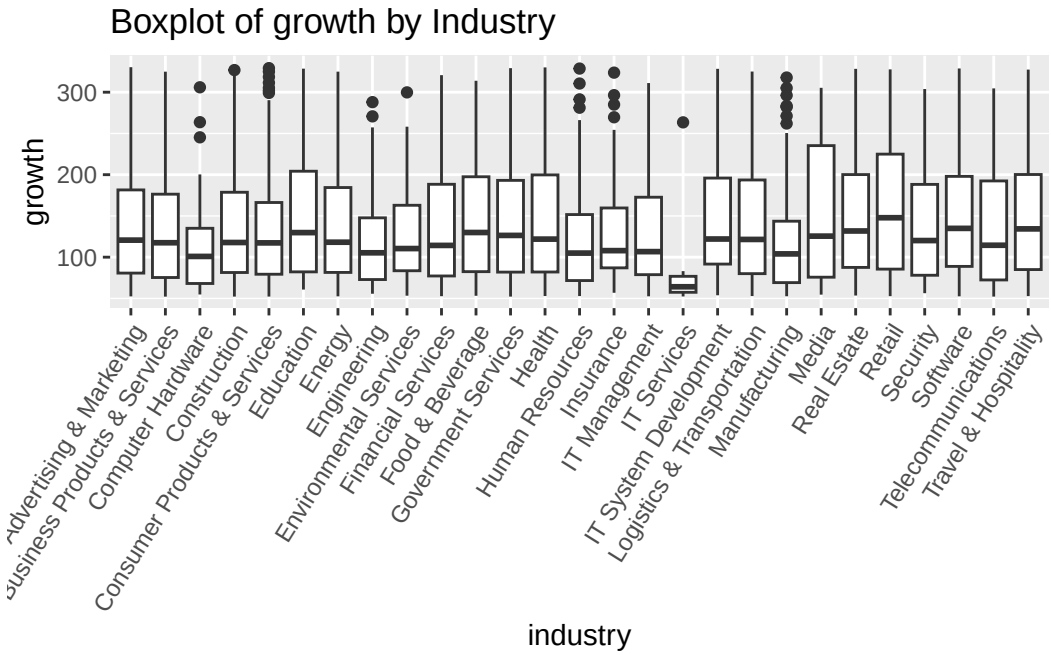
```
inc5000_growth <- inc5000[inc5000$growth <= Q3, ]

boxplot(inc5000_growth$growth, main = "Boxplot of growth (Outliers Removed)",
        ylab = "growth")
```



7.Croissance par industrie

7.1 Créer un diagramme en boîte groupé de croissance par industrie

```
ggplot(inc5000_growth, aes(x = industry, y = growth)) +
  geom_boxplot() +
  labs(title = "Boxplot of growth by Industry") +
  theme(axis.text.x = element_text(angle = 58, hjust = 1))
```



-J'ai remarqué qu'il y a quelques valeurs aberrantes qui devraient être supprimées car elles ne nous permettent pas de faire des statistiques correctes. Pour ce faire, nous devons vérifier l'écart-type du revenu par industrie et les valeurs correspondantes.

7.2 Nettoyage des valeurs aberrantes

```
inc5000_bound <- inc5000_growth %>%
  group_by(industry) %>%
  mutate(
    upper_bound = quantile(growth, 0.75)
  ) %>%
  ungroup()

inc5000_clean <- inc5000_bound %>%
  filter(growth <= upper_bound)
```

7.3 Calculer les statistiques sommaires pour de croissance par industrie.

```
summary_stats <- inc5000_clean %>%
  group_by(industry) %>%
  summarize(min = min(growth),
            max = max(growth),
            mean = mean(growth),
```

```
            q1 = quantile(growth, probs = 0.25),
            median = median(growth),
            q3 = quantile(growth, probs = 0.75),
            variance = var(growth),
            sd = sd(growth)) %>%
  arrange(desc(mean))

print(summary_stats,n=27)
```

```
# A tibble: 27 x 9
   industry                    min   max  mean    q1 median    q3 variance    sd
   <chr>                     <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>    <dbl> <dbl>
 1 Retail                     52.7 225.  120.   77.0  105.  164.    2631.  51.3
 2 Education                  60.7 199.  118.   76.9  109.  167.    2055.  45.3
 3 Software                   52.5 198.  115.   79.4  109.  149.    1732.  41.6
 4 Real Estate                53.5 198.  114.   80.6  111.  146.    1662.  40.8
 5 Travel & Hospitality       52.8 198.  114.   83.6  103.  160.    1923.  43.9
 6 Food & Beverage            53.2 195.  113.   73.5   96.2 156.    2058.  45.4
 7 Media                      54.6 235.  113.   66.0   86.5 147.    3249.  57.0
 8 IT System Development      53.8 196.  112.   84.8  104.  135.    1502.  38.8
 9 Health                     52.9 200.  109.   71.1  102.  136.    1671.  40.9
10 Government Services        52.2 191.  107.   73.1   97.9 134.    1610.  40.1
11 Logistics & Transportati~  52.8 194.  107.   74.8   97.3 133.    1708.  41.3
12 Energy                     52.5 184.  106.   78.9   97.8 126.    1552.  39.4
13 Advertising & Marketing    52.4 182.  104.   71.2  101.  129.    1317.  36.3
14 Construction               52.2 178.  103.   72.8   99.8 127.    1174.  34.3
15 Security                   56.4 188.  103.   69.5   91.9 131.    1568.  39.6
16 Telecommunications         52.2 189.  102.   70.1   97.9 126.    1523.  39.0
17 Financial Services         52.7 188.  102.   69.8   96.4 125.    1373.  37.0
18 Business Products & Serv~  52.2 176.  101.   69.2   93.2 133.    1341.  36.6
19 Insurance                  56.9 160.  100.   78.3   94.4 121.     752.  27.4
20 Consumer Products & Serv~  52.2 165.   99.8  71.0   98.8 126.    1009.  31.8
21 IT Management              52.6 173.   97.9  72.5   90.5 117.     973.  31.2
22 Environmental Services     53.3 161.   96.7  74.8   96.8 115.     838.  29.0
23 Engineering                55.7 147.   94.1  66.0   94.0 118.     834.  28.9
24 Human Resources            52.9 152.   92.0  66.6   88.8 112.     765.  27.7
25 Manufacturing              52.6 142.   88.2  64.1   82.1 111.     674.  26.0
26 Computer Hardware          54.8 130.   84.8  63.8   79.3 111.     649.  25.5
27 IT Services                52.3  76.0  61.6  56.0   60.3  65.8     48.5  6.96
```

7.4 Trouver le groupe de croissance la plus élevée par industrie

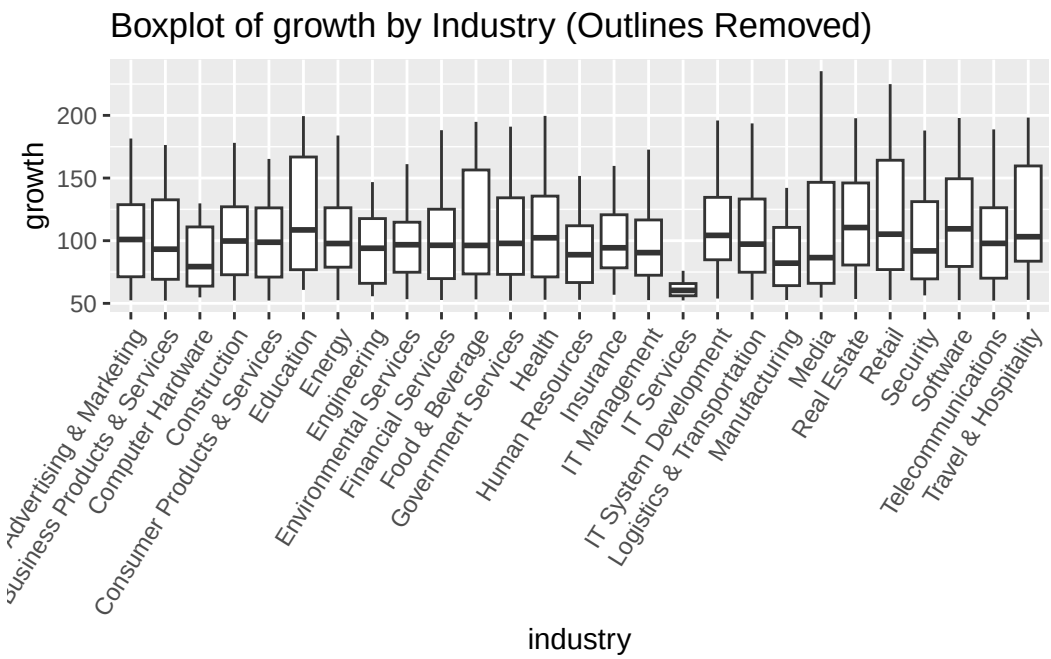
```r
top_growth <- inc5000_clean %>%
  group_by(industry) %>%
  slice_max(order_by = growth, n = 2)

print(top_growth,n=54)
```

```
# A tibble: 54 x 15
# Groups:   industry [27]
    rank profile       name  url   state revenue growth industry workers founded
   <dbl> <chr>         <chr> <chr> <chr>   <dbl>  <dbl> <chr>      <dbl>   <dbl>
 1  2217 https://www.~ Stra~ stra~ CO        9.7  182.  Adverti~      18    2013
 2  2228 https://www.~ beMa~ bema~ PA        2.1  181.  Adverti~      14    2010
 3  2272 https://www.~ Clou~ clou~ CO        2.1  176.  Busines~       3    2012
 4  2296 https://www.~ BigS~ bigs~ CA       29.7  175.  Busines~      31    1995
 5  2888 https://www.~ thom~ thom~ OH        7.2  130.  Compute~      21    2007
 6  3065 https://www.~ Busi~ bits~ IL      129.   121.  Compute~      50    2008
 7  2260 https://www.~ SkyL~ skyl~ CA       15.2  178.  Constru~      80    2005
 8  2261 https://www.~ Mill~ mill~ MA        8    178.  Constru~      50    1979
 9  2409 https://www.~ Stor~ byda~ NY       61.3  165.  Consume~      20    2010
10  2454 https://www.~ Dood~ hugy~ FL        2.4  162.  Consume~      14    2010
11  2018 https://www.~ Nort~ nort~ NV       10.5  199.  Educati~     100    1997
12  2042 https://www.~ The ~ gole~ HI        2.3  197.  Educati~       4    2012
13  2189 https://www.~ Eagl~ eagl~ IA       11    184.  Energy        56    2010
14  2235 https://www.~ Jagu~ jagu~ TX       25    180.  Energy        33    2014
15  2648 https://www.~ RGD   rgde~ FL        6.9  147.  Enginee~      50    1988
16  2695 https://www.~ True~ true~ OH        7.1  143.  Enginee~      28    2007
17  2464 https://www.~ Envi~ envi~ IL       15    161.  Environ~      25    1993
18  2641 https://www.~ Ecos~ epru~ TX        5.9  147.  Environ~      26    2012
19  2135 https://www.~ Resc~ resc~ CA       57.1  188.  Financi~     179    2010
20  2164 https://www.~ Fund~ fund~ WY        7.5  186.  Financi~      29    2007
21  2070 https://www.~ Huss~ huss~ AZ        5    195.  Food & ~      26    2013
22  2079 https://www.~ Chic~ chic~ AL       29.1  194.  Food & ~    1000    2008
23  2109 https://www.~ F. H~ fhca~ MA       36.4  191.  Governm~     375    1999
24  2111 https://www.~ SRT ~ srtg~ FL       57.2  191.  Governm~     228    1999
25  2012 https://www.~ ProS~ pros~ TX       23.1  200.  Health      1733    2010
26  2030 https://www.~ MFI ~ mfim~ CA       15.2  198.  Health        53    1980
27  2580 https://www.~ Barr~ barr~ OK        7.3  152.  Human R~      14    2009
28  2643 https://www.~ Tale~ tale~ NY       80.7  147.  Human R~    1919    1985
29  2330 https://www.~ Fort~ fort~ SC        5.7  173.  IT Mana~      20    2014
30  2332 https://www.~ SADA~ sada~ CA      127.   173.  IT Mana~     225    2000
31  4154 https://www.~ Inte~ inte~ OH        6.4   76.0 IT Serv~      25    2011
32  4190 https://www.~ Nexc~ gone~ NC        7.6   74.6 IT Serv~      50    2001
```

```
33  2053 https://www.~ Data~ http~ MD        2    196.  IT Syst~       8    2009
34  2145 https://www.~ Tx3 ~ tx3s~ PA        7.3  187.  IT Syst~      33    2014
35  2475 https://www.~ The ~ west~ MA       15.2  160.  Insuran~      18    2007
36  2602 https://www.~ Lamb~ lamb~ NY       17.7  150.  Insuran~      96    2008
37  2086 https://www.~ MacG~ macg~ NC        9.7  194.  Logisti~      64    2012
38  2087 https://www.~ Takt~ http~ CA        7.3  193.  Logisti~       5    2015
39  2710 https://www.~ Home~ home~ PA        3.2  142.  Manufac~      25    2011
40  2711 https://www.~ inTe~ inte~ IN       35.3  142.  Manufac~     138    2010
41  1727 https://www.~ Lemo~ lemo~ CA        3.3  235.  Media         40    2014
42  1740 https://www.~ Scie~ scie~ VA       12.7  233.  Media         88    1994
43  2035 https://www.~ CapG~ capg~ IL        8.7  198.  Real Es~      10    2005
44  2041 https://www.~ Peak~ sold~ CO        2.9  197.  Real Es~      23    2010
45  1813 https://www.~ Spar~ spar~ UT       15.2  225.  Retail        28    2011
46  1849 https://www.~ Josh~ josh~ MI        6.8  219.  Retail        55    2004
47  2139 https://www.~ Reta~ reta~ NY       11.9  188.  Security      38    2014
48  2232 https://www.~ OPS ~ opss~ PA        9.2  180.  Security     291    2012
49  2032 https://www.~ Futu~ futu~ OH       20.1  198.  Software      55    2009
50  2073 https://www.~ Prov~ prov~ TN       11.7  194.  Software      68    2010
51  2132 https://www.~ CTI ~ ctit~ IL        4.1  189.  Telecom~      41    1998
52  2290 https://www.~ eSqu~ e2cc~ AZ       28.8  176.  Telecom~     130    2001
53  2028 https://www.~ Maui~ maui~ HI        5    198.  Travel ~      35    2010
54  2216 https://www.~ TRAV~ pres~ NY        3.3  182.  Travel ~      15    2012
# i 5 more variables: yrs_on_list <dbl>, previous_workers <dbl>, metro <chr>,
#   city <chr>, upper_bound <dbl>
```

-Si nous exécutons à nouveau, nous remarquerons qu'il y a moins de valeurs aberrantes et que les statistiques sont meilleures.

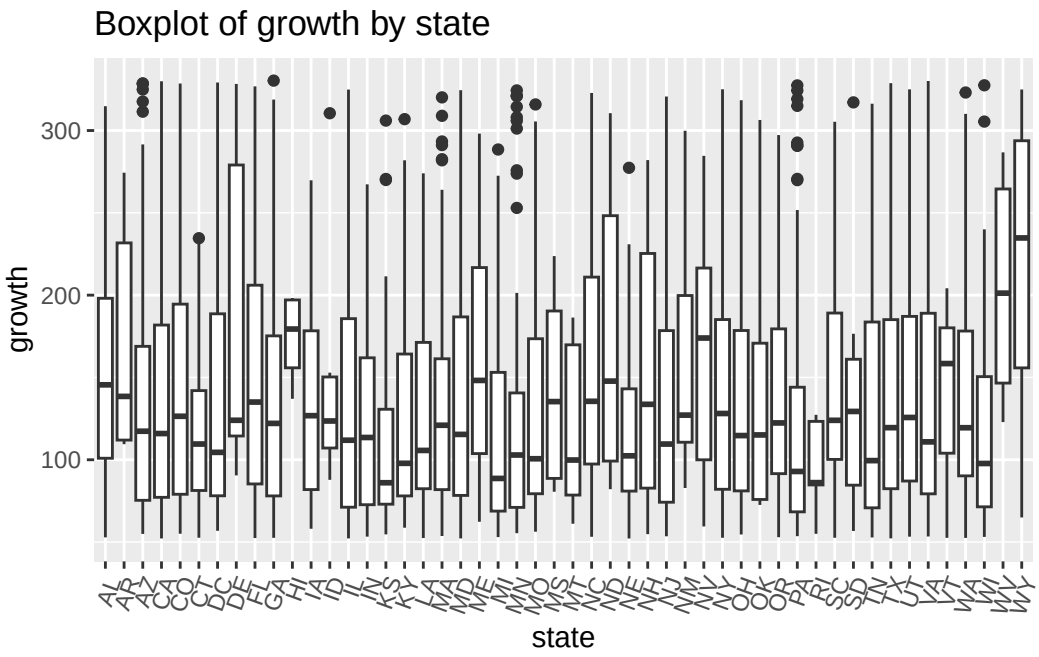
```r
ggplot(inc5000_clean, aes(x = industry, y = growth)) +
  geom_boxplot() +
  labs(title = "Boxplot of growth by Industry (Outlines Removed)") +
  theme(axis.text.x = element_text(angle = 58, hjust = 1))
```



8.Croissance par état

8.1 Créer un diagramme en boîte groupé de croisssance par état.

```
ggplot(inc5000_growth, aes(x = state, y = growth)) +
  geom_boxplot() +
  labs(title = "Boxplot of growth by state") +
  theme(axis.text.x = element_text(angle = 70, hjust = 1))
```



-J'ai remarqué qu'il y a quelques valeurs aberrantes qui devraient être supprimées car elles ne nous permettent pas de faire des statistiques correctes. Pour ce faire, nous devons vérifier l'écart-type du revenu par industrie et les valeurs correspondantes.

8.2 Nettoyage des valeurs aberrantes.

```
inc5000_bound <- inc5000_growth %>%
  group_by(state) %>%
  mutate(
    upper_bound = quantile(growth, 0.75)
  ) %>%
  ungroup()


inc5000_clean <- inc5000_bound %>%
  filter(growth <= upper_bound)

inc5000_clean <- inc5000_clean %>%
  filter(!(rank %in% c(1796,2600,2951,1479,3069,2559,2424,3011,2146,2202,2455,2563)))
```

8.3 Calculer les statistiques sommaires pour la croissance par état.

```r
summary_stats <- inc5000_clean %>%
  group_by(state) %>%
  summarize(min = min(growth),
            max = max(growth),
            mean = mean(growth),
            q1 = quantile(growth, probs = 0.25),
            median = median(growth),
            q3 = quantile(growth, probs = 0.75),
            variance = var(growth),
            sd = sd(growth)) %>%
  arrange(desc(mean))

print(summary_stats,n=51)
```

```
# A tibble: 50 x 9
   state   min   max  mean    q1 median    q3 variance     sd
   <chr> <dbl> <dbl> <dbl> <dbl>  <dbl> <dbl>    <dbl>  <dbl>
 1 WY     64.9 283.  178.  126.   186.  235.  11978.   109.
 2 WV    123.  240.  167.  137.   152.  182.   2638.    51.4
 3 HI    137.  197.  165.  150.   162.  179.    898.    30.0
 4 NV     59.5 216.  127.   85.9  105.  178.   2953.    54.3
 5 AL     52.9 194.  124.   82.6  126.  160.   2024.    45.0
 6 VT     52.6 180.  124.   91.1  131.  164.   3280.    57.3
 7 NH     54.8 225.  118.   69.5  107.  169.   2994.    54.7
 8 NC     53.2 211.  118.   84.2  112.  146.   1729.    41.6
 9 MS     80.6 179.  117.   86.0   91.4 135.   2930.    54.1
10 ND     82.1 157.  116.   85.0  112.  143.   1410.    37.5
11 ID     87.8 149.  115.   98.6  112.  129.    551.    23.5
12 FL     52.4 206.  115.   73.0  103.  154.   2147.    46.3
13 ME     62.3 177.  114.   89.6  109.  134.   2313.    48.1
14 AR    109.  112.  111.  111.   112.  112.      1.99   1.41
15 SC     52.6 173.  111.   92.9  109.  127.   1051.    32.4
16 NM     82.8 129.  111.  100.   116.  126.    448.    21.2
17 CO     55.1 194.  110.   71.9   99.0 139.   1938.    44.0
18 NY     52.6 185.  110.   76.9  101.  144.   1493.    38.6
19 DE     90.5 124.  109.   98.8  114.  120.    203.    14.3
20 UT     53.3 182.  109.   76.4  107.  131.   1336.    36.6
21 OR     53.0 175.  106.   78.7  106.  125.   1137.    33.7
22 TX     52.2 185.  106.   74.9   99.0 131.   1408.    37.5
23 GA     52.6 175.  105.   71.3   97.7 134.   1387.    37.2
24 IA     58.1 176.  105.   74.9   87.1 131.   1627.    40.3
25 MD     52.2 186.  104.   72.5   97.8 132.   1463.    38.3
```

```
26 MA     53.7 161.  103.   73.8  105.  128.    988.    31.4
27 SD     56.7 161.  103.   68.9  102.  133.   1533.    39.1
28 CA     52.2 182.  102.   68.7   98.1 126.   1323.    36.4
29 WA     52.5 178.  102.   70.5  103.  124.   1072.    32.7
30 MO     56.3 173.  101.   70.7   89.6 121.   1369.    37.0
31 OH     54.6 176.  100.   72.4   95.0 125.   1044.    32.3
32 VA     53.4 189.   99.9  71.4   91.5 123.   1223.    35.0
33 IL     52.2 186.   99.8  66.0   89.2 128.   1470.    38.3
34 AZ     55.0 168.   99.4  69.1   99.4 123.    987.    31.4
35 OK     72.6 152.   99.3  73.2   92.9 119.    793.    28.2
36 NJ     53.5 176.   98.7  69.6   83.6 132.   1250.    35.4
37 DC     56.8 165.   97.7  76.9   89.2 113.    953.    30.9
38 LA     52.4 165.   96.4  70.0   87.3 119.   1152.    33.9
39 KY     58.7 164.   95.7  68.6   90.8 109.    977.    31.3
40 CT     52.6 139.   95.0  74.0   94.3 123.    796.    28.2
41 IN     53.3 159.   94.7  62.2   94.2 118.   1041.    32.3
42 NE     52.2 143.   93.5  70.3   89.1 118.    824.    28.7
43 TN     52.8 175.   92.9  60.8   78.8 126.   1319.    36.3
44 MN     55.4 141.   89.0  63.9   77.8 115.    754.    27.5
45 PA     53.6 144.   87.8  66.2   79.8 109.    741.    27.2
46 WI     53.1 148.   86.8  65.3   83.5 102.    633.    25.2
47 MT     61.1 101.   81.5  63.2   83.7  99.0   357.    18.9
48 MI     53.0 124.   79.8  60.8   75.6  94.8   454.    21.3
49 KS     54.7 109.   77.3  64.1   77.8  86.0   295.    17.2
50 RI     55.0  86.1  75.3  69.9   84.7  85.4   308.    17.6
```

8.4 Trouver le groupe de croissance la plus élevéé par état.

```r
top_growth <- inc5000_clean %>%
  group_by(state) %>%
  slice_max(order_by = growth, n = 2)

print(top_growth,n=104)
```

```
# A tibble: 100 x 15
# Groups:   state [50]
     rank profile      name  url   state revenue growth industry workers founded
    <dbl> <chr>        <chr> <chr> <chr>   <dbl>  <dbl> <chr>      <dbl>   <dbl>
  1  2077 https://www~ Spar~ spar~ AL       25.7  194.  Real Es~      19    2011
  2  2079 https://www~ Chic~ chic~ AL       29.1  194.  Food & ~    1000    2008
  3  3243 https://www~ B-Un~ b-un~ AR        9.4  112.  Adverti~      86    1994
  4  3249 https://www~ Gree~ gree~ AR        3.8  112.  Constru~      40    2004
```

5	2376	https://www~	P.R.~ pros~	AZ	5.5	168.	Travel ~	10	2013
6	2506	https://www~	WebPT webp~	AZ	80.5	157.	Software	513	2008
7	2213	https://www~	Issq~ http~	CA	13.9	182.	IT Mana~	140	2010
8	2214	https://www~	Bran~ http~	CA	25.3	182.	Busines~	130	2010
9	2085	https://www~	Madw~ madw~	CO	90.6	194.	Adverti~	489	2009
10	2091	https://www~	Orth~ orth~	CO	11.1	193.	Software	214	2013
11	2744	https://www~	Choi~ choi~	CT	7.8	139.	Financi~	18	2008
12	2825	https://www~	The ~ junk~	CT	13	133.	Environ~	62	2016
13	2413	https://www~	Atla~ atla~	DC	56.1	165.	Governm~	202	2008
14	2479	https://www~	vTec~ vtec~	DC	10.2	159.	Governm~	82	2006
15	3000	https://www~	AR S~ arso~	DE	4.5	124.	IT Mana~	30	2009
16	3091	https://www~	Ches~ ches~	DE	29.2	120.	Constru~	133	2003
17	1946	https://www~	Benz~ benz~	FL	352.	206.	Health	1200	2009
18	1958	https://www~	Scan~ scan~	FL	10.1	205.	Software	62	1989
19	2300	https://www~	Clie~ clie~	GA	21.7	175.	Adverti~	83	1999
20	2305	https://www~	Achi~ achi~	GA	3.9	175.	Software	40	2010
21	2042	https://www~	The ~ gole~	HI	2.3	197.	Educati~	4	2012
22	2451	https://www~	Elit~ elit~	HI	40.5	162.	Real Es~	37	2005
23	2287	https://www~	Medi~ medi~	IA	56.1	176.	Health	660	2007
24	2379	https://www~	Conv~ conv~	IA	5.7	168.	Adverti~	47	2011
25	2612	https://www~	Cutt~ cutt~	ID	18.1	149.	Busines~	188	2015
26	2800	https://www~	Prec~ ppe-~	ID	9.8	135.	Constru~	20	2008
27	2168	https://www~	Espe~ espe~	IL	42.1	186.	Health	8	2009
28	2176	https://www~	Glob~ gwt-~	IL	6.8	185.	Environ~	48	1990
29	2481	https://www~	Crea~ thew~	IN	19.2	159.	Busines~	65	1997
30	2500	https://www~	The ~ theb~	IN	11.1	157.	Adverti~	35	2007
31	3312	https://www~	Prop~ prop~	KS	12.5	109.	Busines~	34	1998
32	3355	https://www~	Chel~ chel~	KS	6.9	107.	IT Mana~	75	2010
33	2428	https://www~	Dent~ dent~	KY	46.4	164.	Real Es~	79	2008
34	2576	https://www~	Sent~ sent~	KY	97	152.	Governm~	1960	2003
35	2418	https://www~	Mode~ mode~	LA	67.8	165.	Environ~	23	1979
36	2583	https://www~	Anyt~ anyt~	LA	4.4	151.	Constru~	7	2009
37	2460	https://www~	AMC ~ amcb~	MA	19.6	161.	Software	536	2008
38	2475	https://www~	The ~ west~	MA	15.2	160.	Insuran~	18	2007
39	2160	https://www~	Ripp~ ripp~	MD	13.5	186.	Governm~	136	2003
40	2200	https://www~	Mary~ mary~	MD	6.3	183.	Health	50	2013
41	2264	https://www~	Veri~ http~	ME	8.1	177.	Software	42	2013
42	3100	https://www~	Drea~ drea~	ME	3.6	119.	Adverti~	35	2009
43	2998	https://www~	Dobe~ dobe~	MI	2	124.	IT Mana~	11	2005
44	3004	https://www~	365 ~ 365r~	MI	52	124.	Software	184	2009
45	2729	https://www~	Aspe~ aspe~	MN	17.9	141.	Constru~	21	2002
46	2756	https://www~	Grea~ grea~	MN	9.3	138.	Financi~	36	2012
47	2324	https://www~	Agil~ agil~	MO	35.6	173.	Software	104	2004

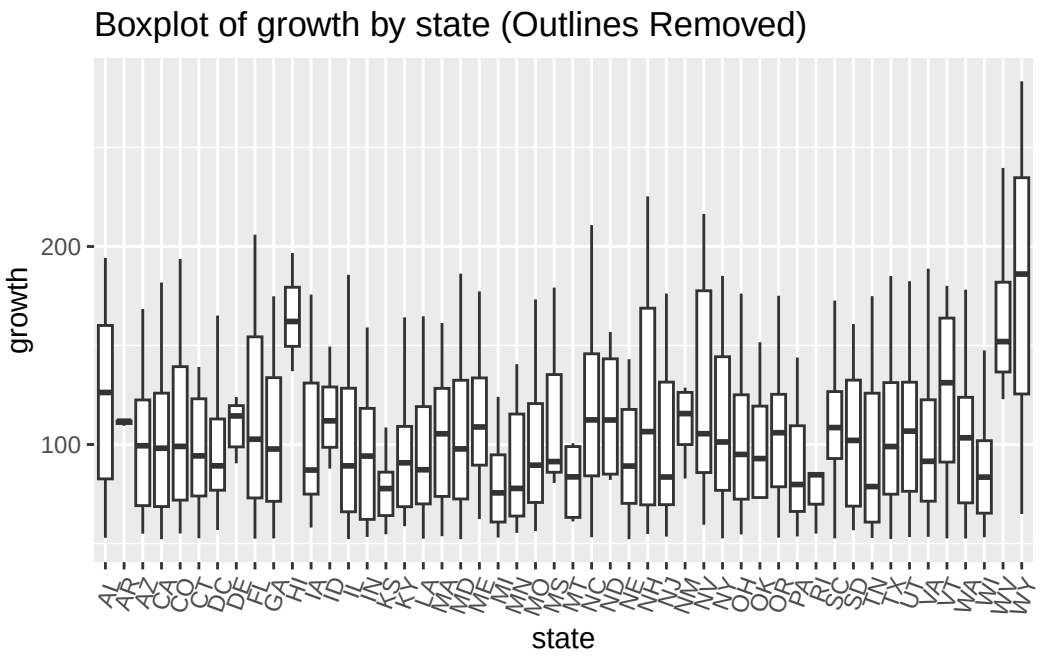
48	2381	https://www~	Abst~	abst~	MO	23.3	168.	Adverti~	260	2009
49	2246	https://www~	Plum~	plum~	MS	9.5	179.	Environ~	64	2009
50	3734	https://www~	Verg~	verg~	MS	9.2	91.4	Software	49	2006
51	3488	https://www~	Foun~	foun~	MT	8.1	101.	Software	96	2007
52	3541	https://www~	Heal~	http~	MT	3	99.0	Health	28	2012
53	1905	https://www~	Done~	done~	NC	9.5	211.	Retail	230	2006
54	1942	https://www~	Stak~	stak~	NC	71.1	206.	Constru~	611	1997
55	2515	https://www~	Myri~	myri~	ND	6.2	157.	Software	87	2011
56	2748	https://www~	BNG ~	bngt~	ND	9	139.	Busines~	70	2007
57	2700	https://www~	Berr~	jsbe~	NE	7.5	143.	Consume~	60	1965
58	2732	https://www~	Metr~	metr~	NE	3.4	140.	Educati~	97	2006
59	1809	https://www~	Silv~	silv~	NH	6.9	225.	Software	54	2005
60	2182	https://www~	Opti~	opti~	NH	35.2	185.	Constru~	180	2006
61	2276	https://www~	Capa~	capa~	NJ	76.9	176.	Logisti~	740	1999
62	2285	https://www~	MedE~	mede~	NJ	5.1	176.	Health	17	1999
63	2910	https://www~	Roof~	roof~	NM	8.7	129.	Constru~	82	2008
64	2979	https://www~	APIC~	apic~	NM	16.5	126.	Constru~	132	2012
65	1869	https://www~	Blue~	blue~	NV	3.9	216.	IT Mana~	5	2008
66	1947	https://www~	Cult~	cult~	NV	3	206.	Manufac~	11	2011
67	2177	https://www~	Rema~	rema~	NY	21.2	185.	Food & ~	80	2012
68	2179	https://www~	Inno~	inno~	NY	19.8	185.	Software	600	2011
69	2278	https://www~	LSP ~	lspt~	OH	11	176.	Manufac~	49	1995
70	2315	https://www~	Trip~	trip~	OH	14.2	174.	Manufac~	45	2003
71	2580	https://www~	Barr~	barr~	OK	7.3	152.	Human R~	14	2009
72	2623	https://www~	Coll~	coll~	OK	39.6	148.	Consume~	300	1996
73	2294	https://www~	Risi~	risi~	OR	13.7	175.	Food & ~	35	1985
74	2374	https://www~	Prop~	prop~	OR	17.8	169.	Busines~	100	2012
75	2686	https://www~	Mans~	mans~	PA	9.5	144.	Constru~	35	1998
76	2701	https://www~	Deal~	myde~	PA	2.9	143.	Adverti~	23	2009
77	3859	https://www~	Best~	best~	RI	3.6	86.1	Energy	24	2010
78	3903	https://www~	Trib~	trib~	RI	7	84.7	Adverti~	63	2010
79	2330	https://www~	Fort~	fort~	SC	5.7	173.	IT Mana~	20	2014
80	2338	https://www~	Inte~	inte~	SC	15.4	172.	Security	50	2009
81	2465	https://www~	Inde~	inde~	SD	5.1	161.	Financi~	28	1997
82	2713	https://www~	Fit ~	http~	SD	2.2	142.	Retail	22	2013
83	2298	https://www~	Rede~	rede~	TN	7	175.	Constru~	55	2007
84	2441	https://www~	REN ~	rend~	TN	2.8	163.	Health	20	2013
85	2178	https://www~	Caps~	caps~	TX	3	185.	Real Es~	20	2013
86	2184	https://www~	Thre~	thre~	TX	7.3	185.	Adverti~	29	2013
87	2205	https://www~	DFPG~	dfpg~	UT	27.5	182.	Financi~	19	2010
88	2248	https://www~	The ~	arbi~	UT	13.1	179.	Busines~	50	1979
89	2131	https://www~	Bril~	bril~	VA	31.8	189.	Governm~	603	2006
90	2138	https://www~	Ston~	ston~	VA	9	188.	Food & ~	175	2007

```
 91  2236 https://www~ Reso~ reso~ VT       14.6  180.  Busines~      68    2005
 92  2487 https://www~ New ~ newb~ VT        5.2  158.  Adverti~      49    2002
 93  2258 https://www~ Leno~ leno~ WA       15.8  178.  IT Syst~     100    2012
 94  2540 https://www~ Surd~ pure~ WA        3.8  154.  Consume~      25    2008
 95  2639 https://www~ Beva~ beva~ WI       14.5  148.  Busines~     108    2014
 96  2764 https://www~ Tund~ tund~ WI       42.3  137.  Constru~     200    2009
 97  1703 https://www~ Next~ next~ WV       15    240.  IT Syst~     140    2012
 98  2443 https://www~ Adva~ adva~ WV       13.1  163.  IT Mana~      83    2004
 99  1461 https://www~ Red ~ redr~ WY       14.9  283.  Manufac~      71    2013
100  2164 https://www~ Fund~ fund~ WY        7.5  186.  Financi~      29    2007
# i 5 more variables: yrs_on_list <dbl>, previous_workers <dbl>, metro <chr>,
#   city <chr>, upper_bound <dbl>
```

-En utilisant le boxplot, summary_stats et top_growth, nous pouvons identifier les entreprises qui doivent être supprimées pour garantir de meilleures statistiques.

-Si nous exécutons à nouveau, nous remarquerons qu'il y a moins de valeurs aberrantes et que les statistiques sont meilleures.

```
ggplot(inc5000_clean, aes(x = state, y = growth)) +
  geom_boxplot() +
  labs(title = "Boxplot of growth by state (Outlines Removed)") +
  theme(axis.text.x = element_text(angle = 70, hjust = 1))
```



9.Corrélation

9.1 Corrélation entre revenue et workers

```
cor(inc5000$revenue, inc5000$workers, use = "complete.obs")
```

```
[1] 0.1026319
```

-La corrélation est de 0.1, ce qui indique que la relation entre le nombre de travailleurs et le revenue est très faible.

9.2 Corrélation entre revenue et yrs_on_list

```
cor(inc5000$revenue, inc5000$yrs_on_list)
```

```
[1] 0.3000241
```

-La corrélation est de 0.3, ce qui indique qu'il existe une relation bien que faible entre le revenue et les années sur la liste.

9.3 Corrélation entre growth et workers

```
cor(inc5000$growth, inc5000$workers, use = "complete.obs")
```

```
[1] -0.01108796
```

-La corrélation est de -0.01, ce qui indique qu'il n'y a pas de relation entre les travailleurs et growth.

9.4 Corrélation entre growth et yrs_on_list

```
cor(inc5000$growth, inc5000$yrs_on_list)
```

```
[1] -0.1611326
```

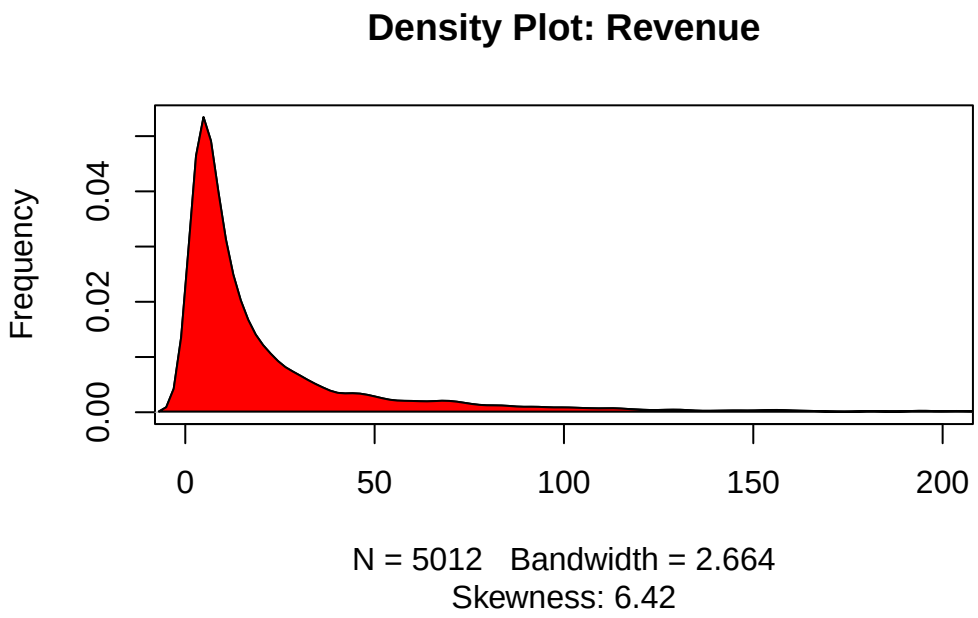
-La corrélation est de -0.16, ce qui indique qu'il y a une relation faible entre growth et les années sur la liste.

10.Density Plots

```r
par(mfrow = c(1, 2))
```

10.1 Density plot pour revenue

```r
plot(density(inc5000$revenue), main = "Density Plot: Revenue", ylab = "Frequency",
     sub = paste("Skewness:", round(e1071::skewness(inc5000$revenue), 2)),
     xlim = c(0, 200))
polygon(density(inc5000$revenue), col = "red")
```

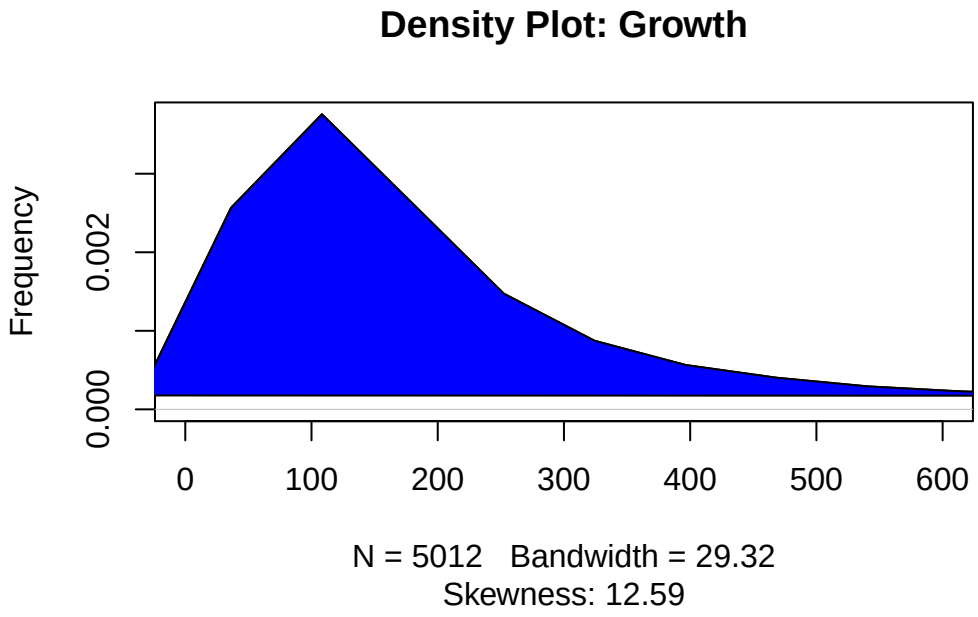


Conclusion :

-la variable revenue suit une distribution “Skewed left”.

10.2 Density plot pour growth

```r
plot(density(inc5000$growth), main = "Density Plot: Growth", ylab = "Frequency",
     sub = paste("Skewness:", round(e1071::skewness(inc5000$growth), 2)),
     xlim = c(0, 600))
polygon(density(inc5000$growth), col = "blue")
```



Conclusion :

-la variable growth suit une distribution “Skewed left”.

11.Supprimer la ligne avec la valeur nulle de travailleur.

```
inc5000 <- inc5000[inc5000$rank != 3746, ]
```

12.PCA

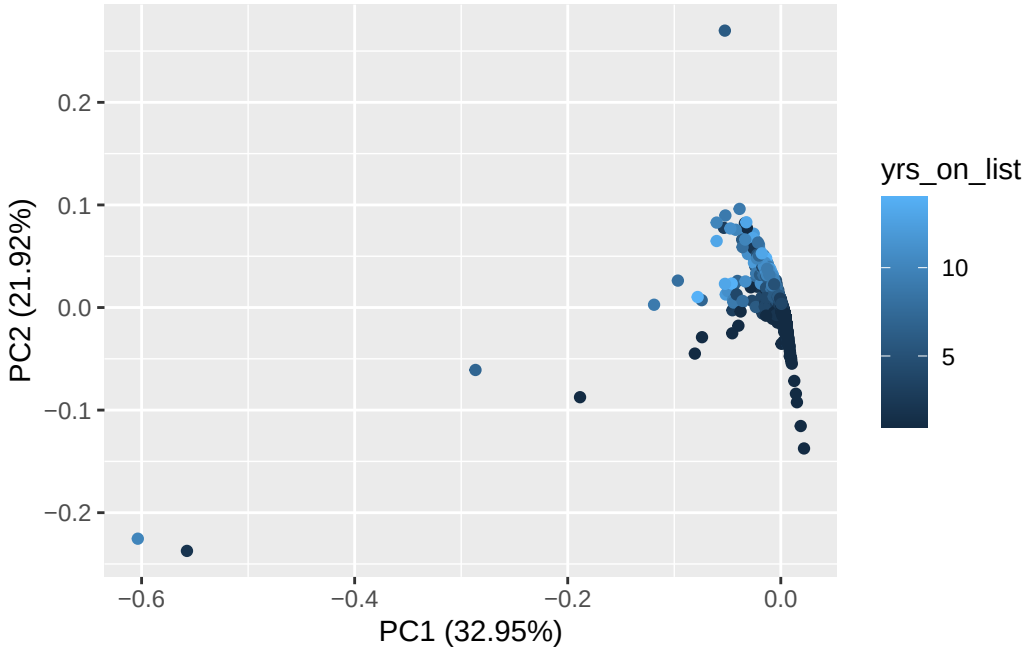
```
inc5000_numeric <- inc5000[, sapply(inc5000, is.numeric)]

inc5000_standardize <- as.data.frame(scale(inc5000_numeric[2:7]))
head(inc5000_standardize)
```

```
      revenue   growth     workers   founded yrs_on_list previous_workers
1  0.08139263 28.20418 -0.07241825 0.3152040  -0.8005098      -0.09935473
2  0.03514445 23.42964 -0.07277537 0.3152040  -0.8005098      -0.09656140
3 -0.08678256 18.23856 -0.01884998 0.3152040  -0.8005098      -0.09469917
4  0.01832693 16.65761 -0.06634718 0.2822155  -0.8005098      -0.10214807
5 -0.12041760 13.78976 -0.07777507 0.2822155  -0.8005098      -0.09842362
6  3.37342204 10.68872  0.17828122 0.1172733  -0.8005098      -0.08725027
```

```r
inc5000.mat <- as.matrix(inc5000_standardize)
cov.mat <- cor(inc5000_standardize)
pca <- prcomp(inc5000_standardize,center = T,scale. = T)
summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6
Standard deviation     1.4060 1.1469 0.9971 0.9745 0.8150 0.31579
Proportion of Variance 0.3295 0.2192 0.1657 0.1583 0.1107 0.01662
Cumulative Proportion  0.3295 0.5487 0.7144 0.8727 0.9834 1.00000
```

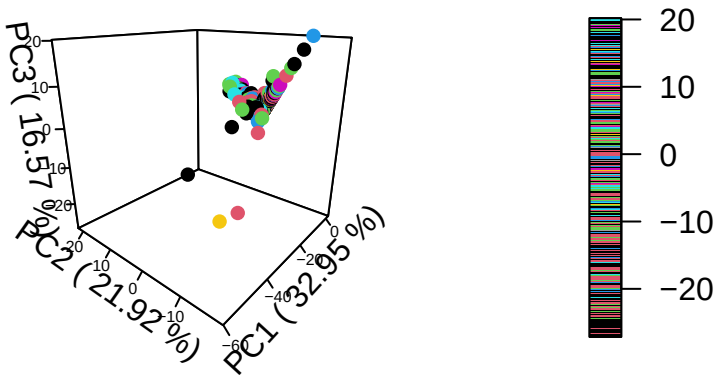
```r
autoplot(pca, data = inc5000, colour = 'yrs_on_list')
```



-On peut remarquer que les contributions de PC1 et PC2 sont inférieures à 80 %. Par conséquent, nous devons ajouter PC3.

```r
PC_scores <- pca$x

xlim <- c(min(PC_scores[,1]), max(PC_scores[,1]))
ylim <- c(min(PC_scores[,2]), max(PC_scores[,2]))
zlim <- c(min(PC_scores[,3]), max(PC_scores[,3]))
```

```r
variance_explained <- pca$sdev^2 / sum(pca$sdev^2) * 100

scatter3D(PC_scores[,1], PC_scores[,2], PC_scores[,3],
          pch = 16, col = inc5000$yrs_on_list,
          xlab = paste("PC1 (", round(variance_explained[1], 2), "%)"),
          ylab = paste("PC2 (", round(variance_explained[2], 2), "%)"),
          zlab = paste("PC3 (", round(variance_explained[3], 2), "%)"),
          xlim = xlim, ylim = ylim, zlim = zlim,
          ticktype = "detailed",
          theta = -50, phi = 20,
          cex.axis = 0.5)
```



13.Prédiction

13.1 Régréssion linéaire

13.1.1 Relation entre Revenue et years on list

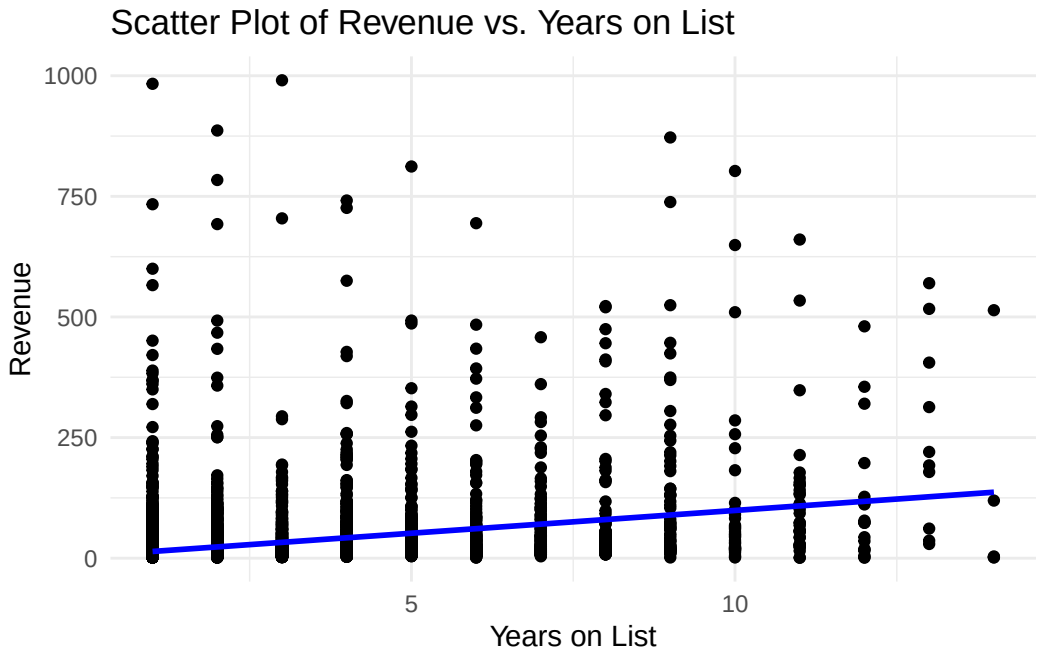
```r
linearMod <- lm(inc5000$yrs_on_list ~ inc5000$revenue, inc5000=inc5000)
print(linearMod)
```

```
Call:
```

```
lm(formula = inc5000$yrs_on_list ~ inc5000$revenue, inc5000 = inc5000)

Coefficients:
    (Intercept)  inc5000$revenue
       2.517796         0.009527
```

```
revenue = 2.517435+9.529e-03*inc5000['yrs_on_list']
```

```
ggplot(inc5000, aes(x = yrs_on_list, y = revenue)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot of Revenue vs. Years on List",
       x = "Years on List",
       y = "Revenue") +
  theme_minimal()
```



Interprétation :

-Dans la régression linéaire avec les années sur la liste comme entrée et revenue comme sortie, nous constatons que : lorsque la valeur des années sur la liste augmente d'une unité, la valeur de revenue augmente de 9,529e-03 unités. Bien que la relation entre ces deux variables soit très faible, le coefficient est statistiquement significatif. Avec un nombres des années nul, nous nous attendons à ce que la valeur de revenue soit égale à 2.517435. Le modèle n'est pas une bonne représentation de la variable revenue.

```r
summary(linearMod)$r.squared
```

```
[1] 0.08999047
```

-Elle explique seulement 9 % du jeu de données.

13.1.2Relation entre Revenue et Workers

```r
linearMod <- lm(inc5000$workers ~ inc5000$revenue, inc5000=inc5000)
print(linearMod)
```

```
Call:
lm(formula = inc5000$workers ~ inc5000$revenue, inc5000 = inc5000)

Coefficients:
    (Intercept)  inc5000$revenue
        117.556            4.028
```

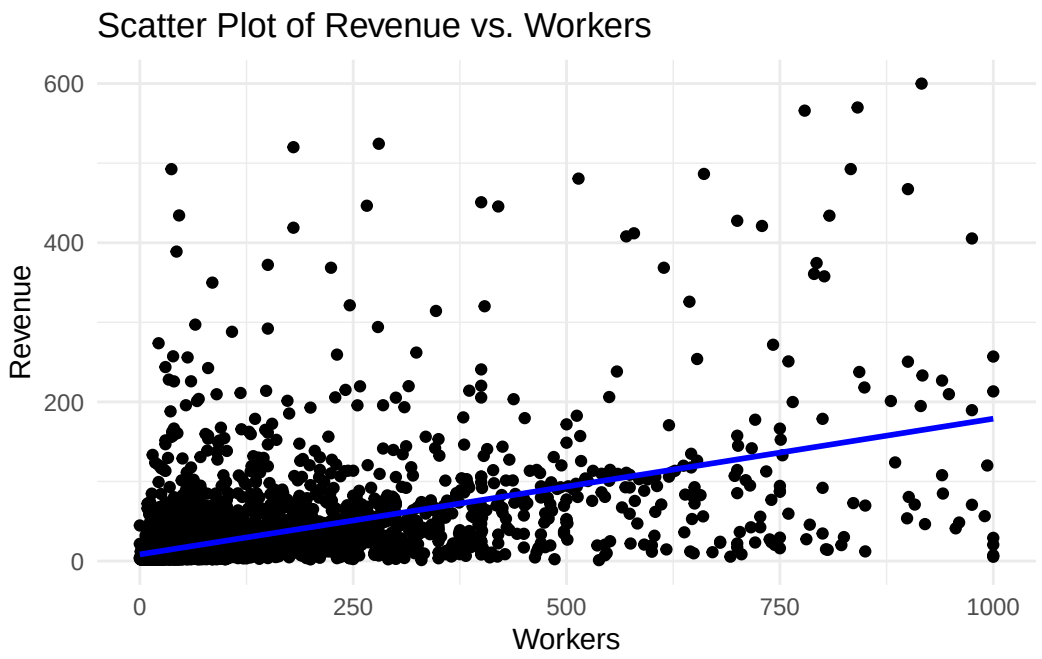
```r
revenue = 117.556+4.028*inc5000['workers']
```

```r
summary(linearMod)$r.squared
```

```
[1] 0.01053331
```

-Elle explique seulement 1 % du jeu de données.

```r
ggplot(inc5000, aes(x = workers, y = revenue)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot of Revenue vs. Workers",
       x = "Workers",
       y = "Revenue") +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 600)) +
  scale_x_continuous(limits = c(0, 1000))
```



Interprétation :

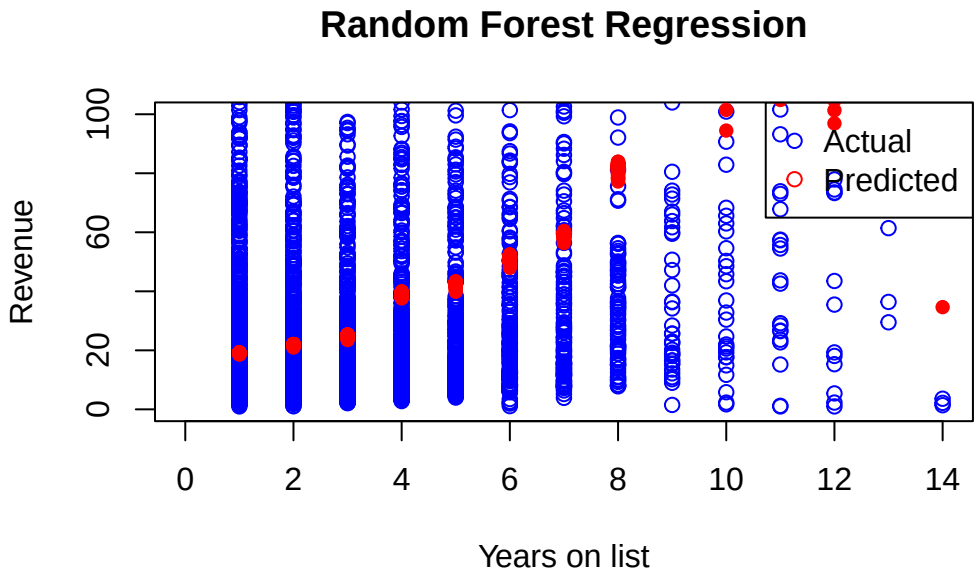
-Dans la régression linéaire avec les travailleurs comme entrée et revenue comme sortie, nous constatons que : lorsque la valeur des travailleurs augmente d'une unité, la valeur de revenue augmente de 4.028 unités. Bien que la relation entre ces deux variables soit très faible, le coefficient est statistiquement significatif. Avec un nombres des travailleurs nul, nous nous attendons à ce que la valeur de revenue soit égale à 117.556. Le modèle n'est pas une bonne représentation de la variable revenue.

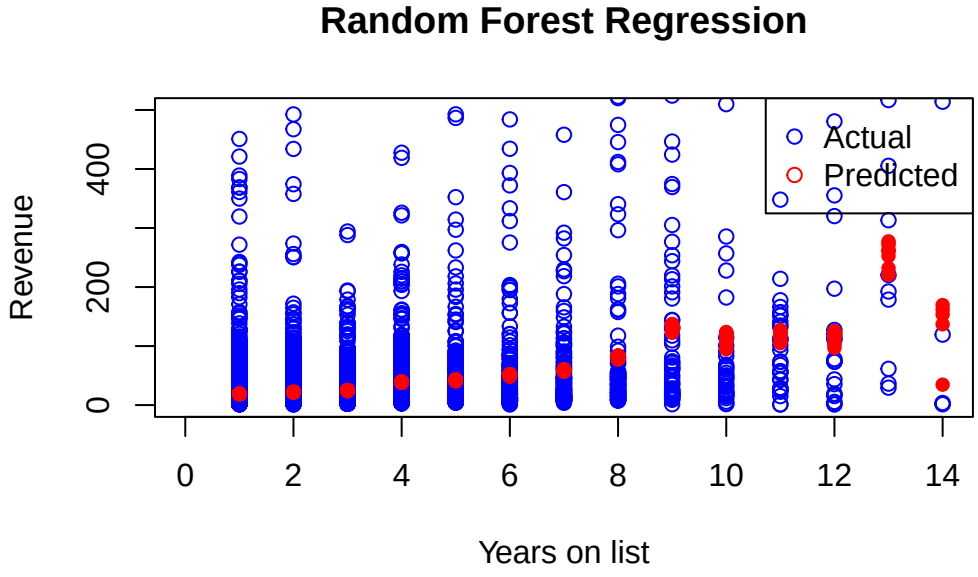
13.2 Prédiction RandomForest

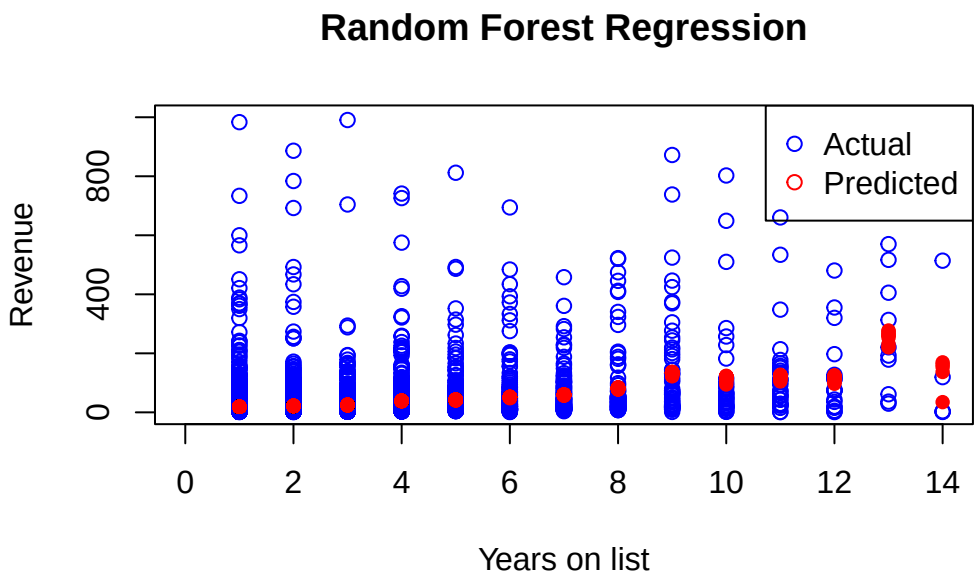
13.2.1 Relation entre Revenue et years on list

```
rf_model <- randomForest(revenue ~ yrs_on_list, data = inc5000)
```

```
plot(inc5000$yrs_on_list, inc5000$revenue, col = "blue",
     main = "Random Forest Regression",
     xlab = "Years on list", ylab = "Revenue", xlim = c(0, 14),
     ylim = c(0, 100))
points(inc5000$yrs_on_list, predict(rf_model), col = "red", pch = 16)
legend("topright", legend = c("Actual", "Predicted"), col = c("blue", "red"),
 pch = c(1, 1))
```


Random Forest Regression
Revenue
0
20
60
100
0
2
4
6
8
10
12
14
Years on list
Actual
Predicted

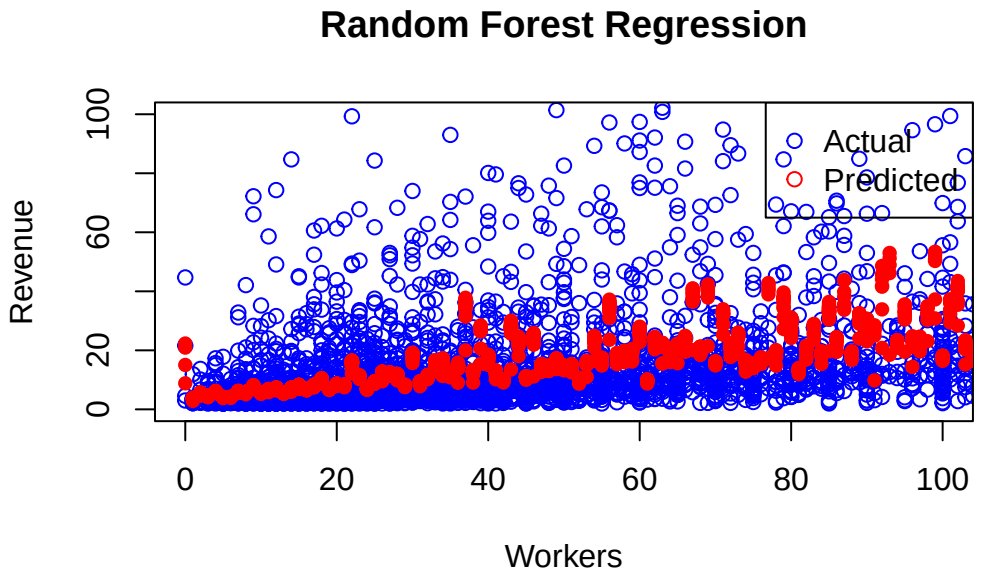

Random Forest Regression
Revenue
0
200
400
0
2
4
6
8
10
12
14
Years on list
Actual
Predicted

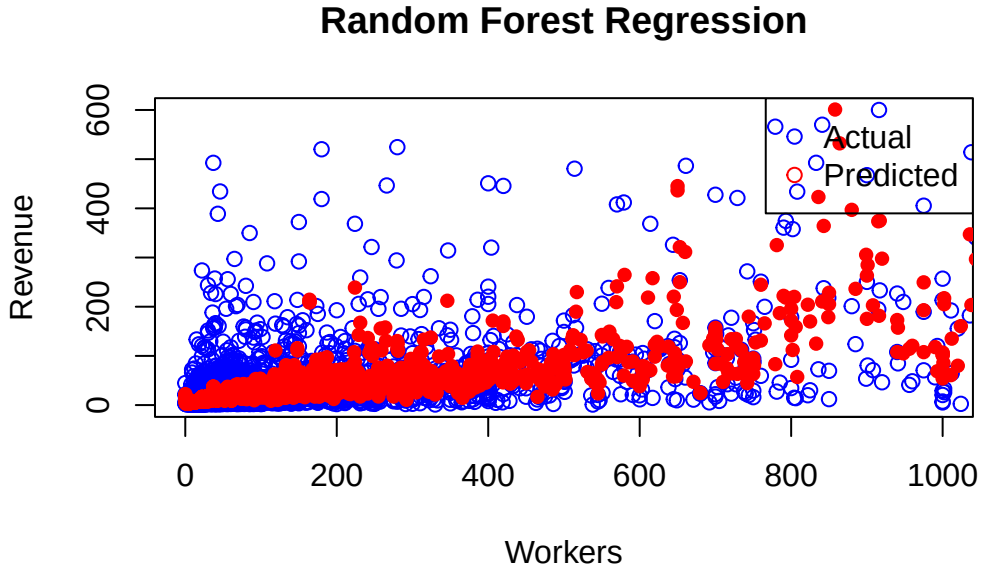


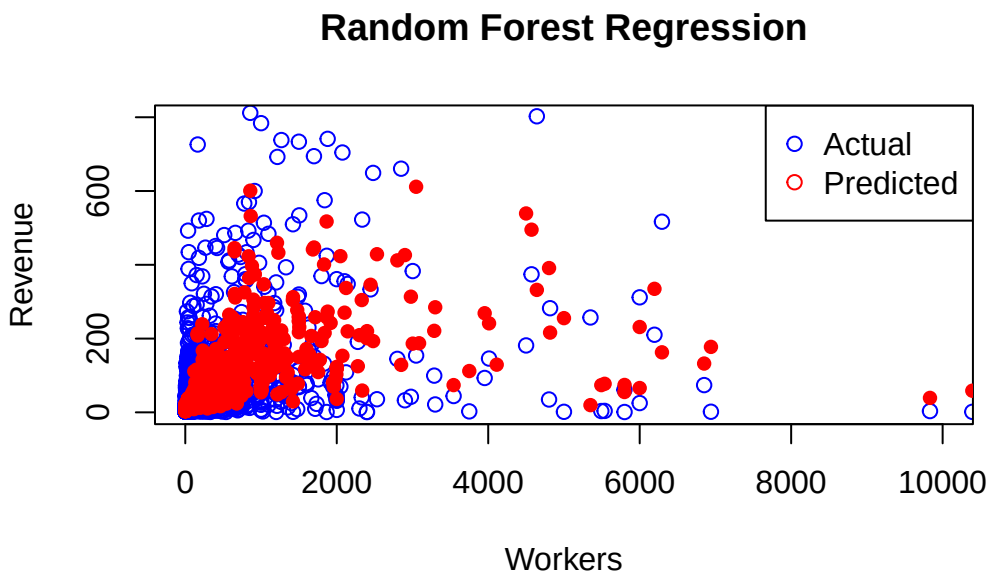
13.2.2 Relation entre Revenue et Workers

```
rf_model <- randomForest(revenue ~ workers, data = inc5000)
```

```
plot(inc5000$workers, inc5000$revenue, col = "blue", main = "Random Forest Regression",
     xlab = "Workers", ylab = "Revenue", xlim = c(0, 100),
     ylim = c(0, 100))
points(inc5000$workers, predict(rf_model), col = "red", pch = 16)
legend("topright", legend = c("Actual", "Predicted"), col = c("blue", "red"),
pch = c(1, 1))
```


Random Forest Regression
Revenue
Workers
Actual
Predicted
0
20
40
60
80
100


Random Forest Regression
Revenue
Workers
Actual
Predicted
0
200
400
600
800
1000



Interprétation

Nous remarquons que dans les deux cas, la prédiction RandomForest est meilleure que la régression linéaire.