Problem#2

1) Based on their history, which is most likely to be successful?

Based on 12 months cumulative gas production, “LEWIS PETRO PROPERTIES, INC.” with highest gas production (530812874) is most likely to be successful.
Based on 12 months cumulative oil production, "EOG RESOURCES, INC." with highest production (138398725) is most likely to be successful.

2) Who is the best company by size?

First I group companies into 8 bins according to their size as follow,
[1, 1] [2, 2] [3, 3] [4, 5] [6, 8] [9, 14] [15, 32] [33, 2818]. The first bin contains about 2000 companies, second contains about 1000 companies, the rest each contains about 600 companies.
For each size bin, the best companies according to their initial oil production are

```
> oil_gmax1
         operator_name MAX(ipmo_oil_sum) well_num
1 PETROEDGE OPERATING LLC           37242       1
> oil_gmax2
              operator_name MAX(ipmo_oil_sum) well_num
1 "MARLIN ENERGY RESOURCES, LLC"           26608       2
> oil_gmax3
            operator_name MAX(ipmo_oil_sum) well_num
1 "KEYSTONE EXPLORATION, INC"           25260       3
> oil_gmax4
          operator_name MAX(ipmo_oil_sum) well_num
1 MW PETROLEUM CORPORATION           19640       5
> oil_gmax5
       operator_name MAX(ipmo_oil_sum) well_num
1 "TEXOZ E&P I, INC."           57797       8
> oil_gmax6
     operator_name MAX(ipmo_oil_sum) well_num
1 "NFR ENERGY, LLC"           179468      14
> oil_gmax7
             operator_name MAX(ipmo_oil_sum) well_num
1 "HALCON OPERATING CO., INC."           324494      21
> oil_gmax8
        operator_name MAX(ipmo_oil_sum) well_num
1 "EOG RESOURCES, INC."         34481036     2818
```

For each size bin, the best companies according to their initial gas production are

```
> gas_gmax1
        operator_name MAX(ipmo_gas_sum) well_num
1 "LAREDO ENERGY V, LP"           162055       1
> gas_gmax2
            operator_name MAX(ipmo_gas_sum) well_num
1 "CONTANGO OPERATORS, INC."           98985       2
> gas_gmax3
```

```
            operator_name MAX(ipmo_gas_sum) well_num
1 "TEXAS, L. PETROLEUM, INC."           375862       3
> gas_gmax4
          operator_name MAX(ipmo_gas_sum) well_num
1 "EEX E&P COMPANY, L.P."           181099       4
> gas_gmax5
     operator_name MAX(ipmo_gas_sum) well_num
1 GENIE CORPORATION           644955       8
> gas_gmax6
                operator_name MAX(ipmo_gas_sum) well_num
1 "MICHAEL PETR. & HURD ENT., LTD."           989202       14
> gas_gmax7
             operator_name MAX(ipmo_gas_sum) well_num
1 "SQUARE MILE ENERGY, L.L.C."          2171057      15
> gas_gmax8
               operator_name MAX(ipmo_gas_sum) well_num
1 "LEWIS PETRO PROPERTIES, INC."        128499929     2029
```

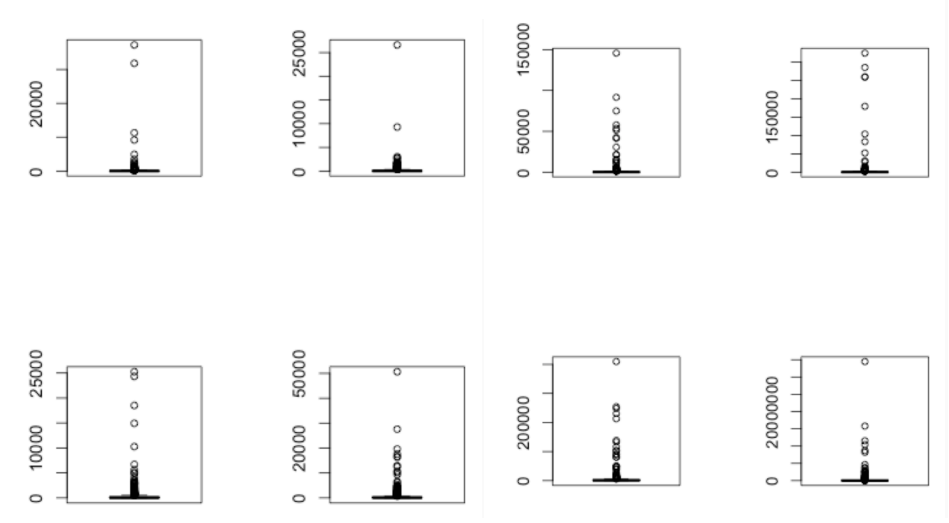
OUTLIERS Detection:

i) Find outliers in each bin.

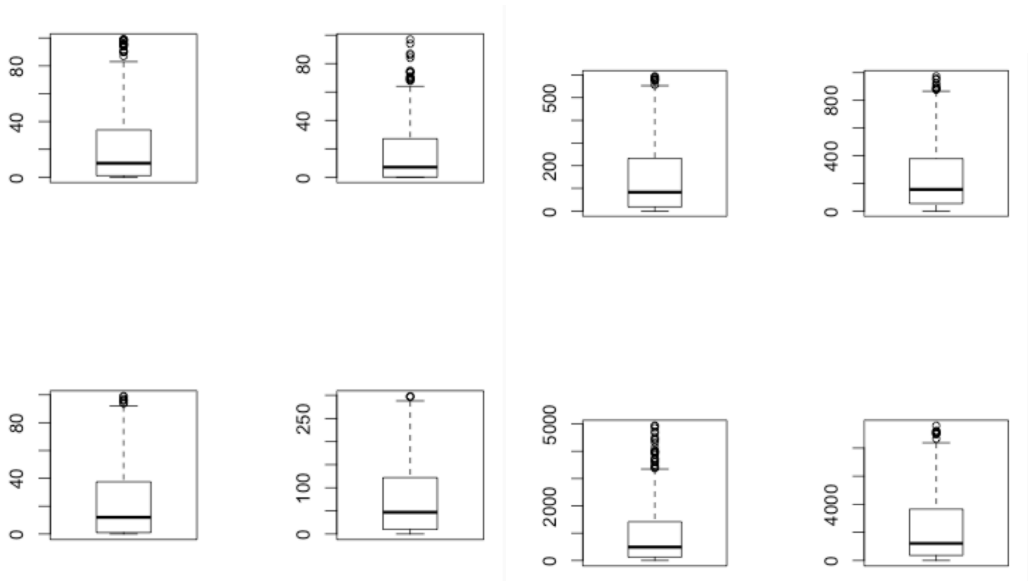
I make boxplots of initial production in each size bin to find outliers.

Oil:

At first all boxplots looks really bad like:



Then I subset each bin and get:



In this way we could get outliers in each bin.
However, it seems there are too many outliers, so maybe we should divide data into more bins.

3) Who is most likely to succeed in the future when drilling wells?
Assume first 6 month's production is related to next 6 month's production. So I regressed first 6 month's production on next 6 month's production for each company. And use the second 6 month's data to predict future production. In this way, I get "PETROHAWK OPERATING COMPANY" is most likely to succeed in the future for gas production. "BURLINGTON RESOURCES O & G CO LP" is most likely to succeed in the future for oil production.

Besides, using monthly production data for wells in EagleFord_Production could make timeseries forecasting model to predict production of wells in future. However, EagleFord_Production don't contain all well's that listed in EagleFord_Well_Meta. So it's hard to forecast future production for each company.

4) Who has the best performing production where more production is better?
First I sum up all wells' decline of each company and then find the companies with best 12 month oil and gas production.
According to oil production, "OAK VALLEY OPERATING, LLC" has no decline of wells and has best production (1854245).
According to gas production, "LAREDO ENERGY OPERATING, LLC" has no decline of wells and has best production (7975543).

Comparison of new and old wells:
Here I divided wells into two groups according to their initial productions time. The group represents new wells are made up of well's initial production date later than 2004-01-01, the old wells are made up of well's earlier than 2004-01-01. Then I did two-sample t-test (H0: old wells perform better. Ha: Not H0) on these two groups.

According to their first year decline number:
> t.test(old_dec,new_dec,alternative = c("greater"),var.equal=F)

Welch Two Sample t-test

data: old_dec and new_dec
t = 4.1391000000000000014, df = 35483.877000000000407, p-value = 0.00001747
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.0072194644641346969113 Inf
sample estimates:
mean of x mean of y
0.48181337881789099153 0.46983284223094995413

According to cum_12_oil:
> t.test(old_oil,new_oil,alternative = c("less"),var.equal=F)

Welch Two Sample t-test

data: old_oil and new_oil
t = -110.40760000000000218, df = 22498.13799999999901, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -35307.008787501057668
sample estimates:
mean of x mean of y
2255.7590955153295909 38096.7491016308849794

According to cum_12_gas:
> t.test(old_gas,new_gas,alternative = c("less"),var.equal=F)

Welch Two Sample t-test

data: old_gas and new_gas
t = -53.566899999999996851, df = 39779.713999999999942, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -151638.41076348078786
sample estimates:
mean of x mean of y
69457.852802670531673 225900.175435363489669

Conclusion:

The small p-value shows that we have enough evidence to reject H0 “old-wells perform better”. And I also did two side t-test, which H0 is “old and new wells perform equally”. In this test, I also got small p-values, which means old and new wells are not performing equally. So I conclude that new wells are performing better than old wells.