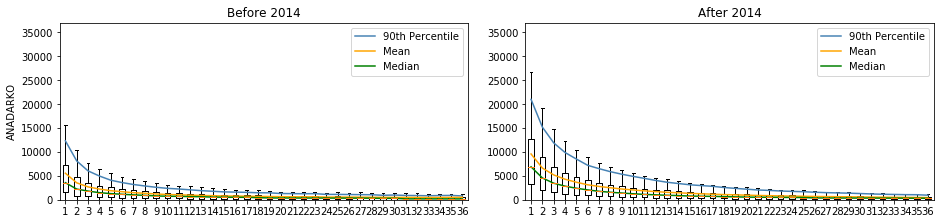
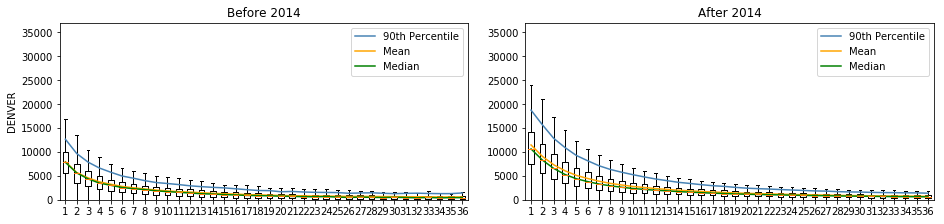
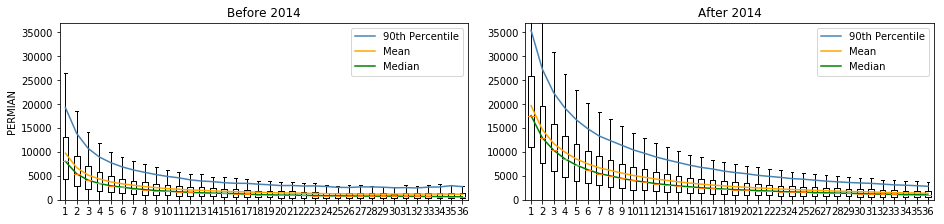
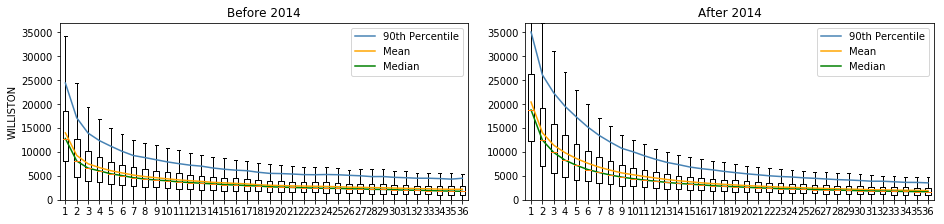
Insights:

**Dataset:**

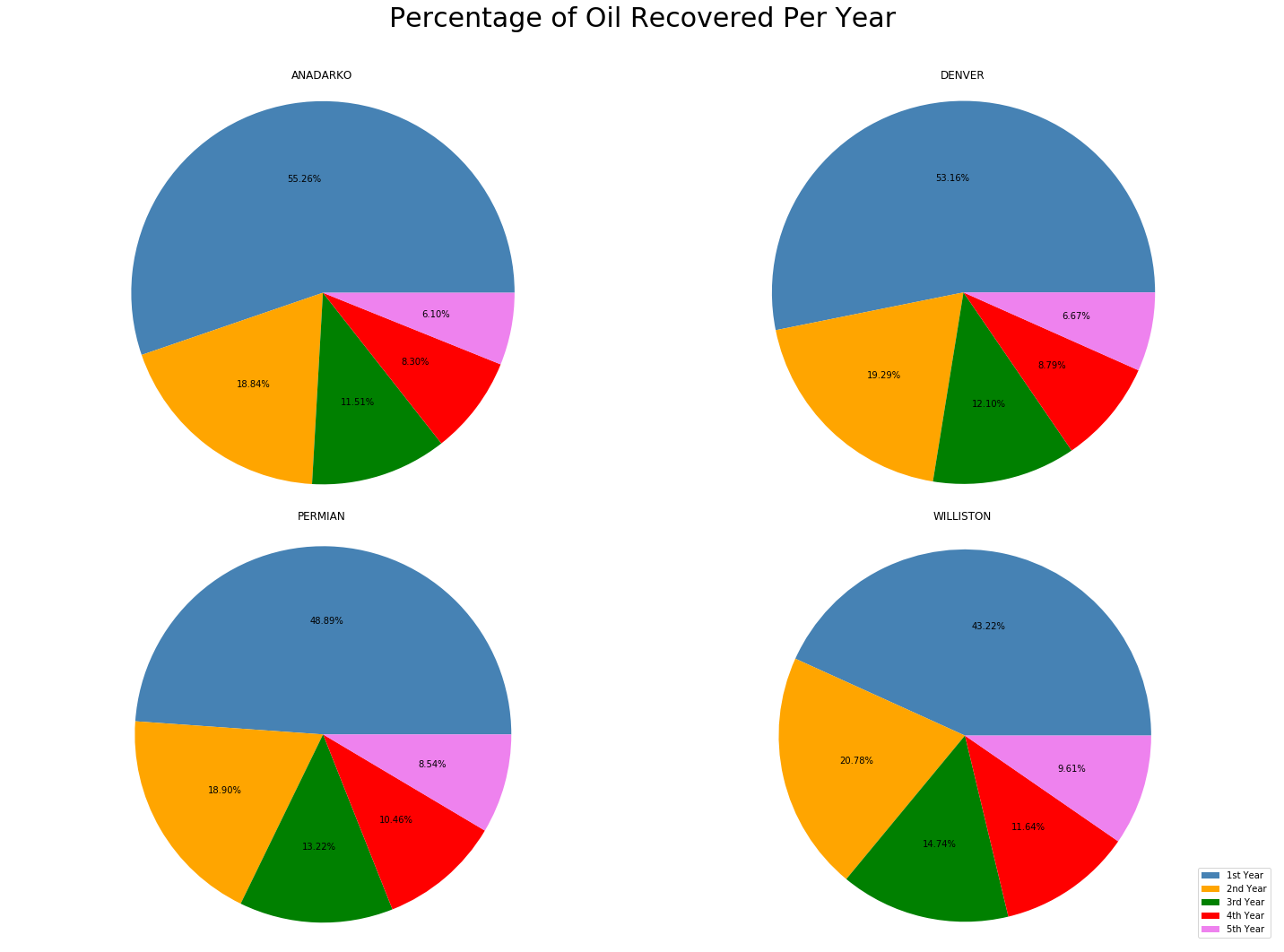
* Two groups of APIs in the production\_data file are from the same basin and have the same operator. In addition to that they produced the exact amount of gas, oil, water at the same dates for all their life time. That’s why we believe that these 2 groups APIs refer to the same wells
  + - 35073253760000, 35073254060000, 35073254740000, 35073255250000, 35073255350000, 35073255360000, 35073256100000, 35073250310000
    - 35073256530000, 35073256910000
* ShortName and Alias are the same thing in operator\_aliasing
* Different production records for the same well in the same month, year
* Some records were duplicated in the dataset.

**Plot 1:**

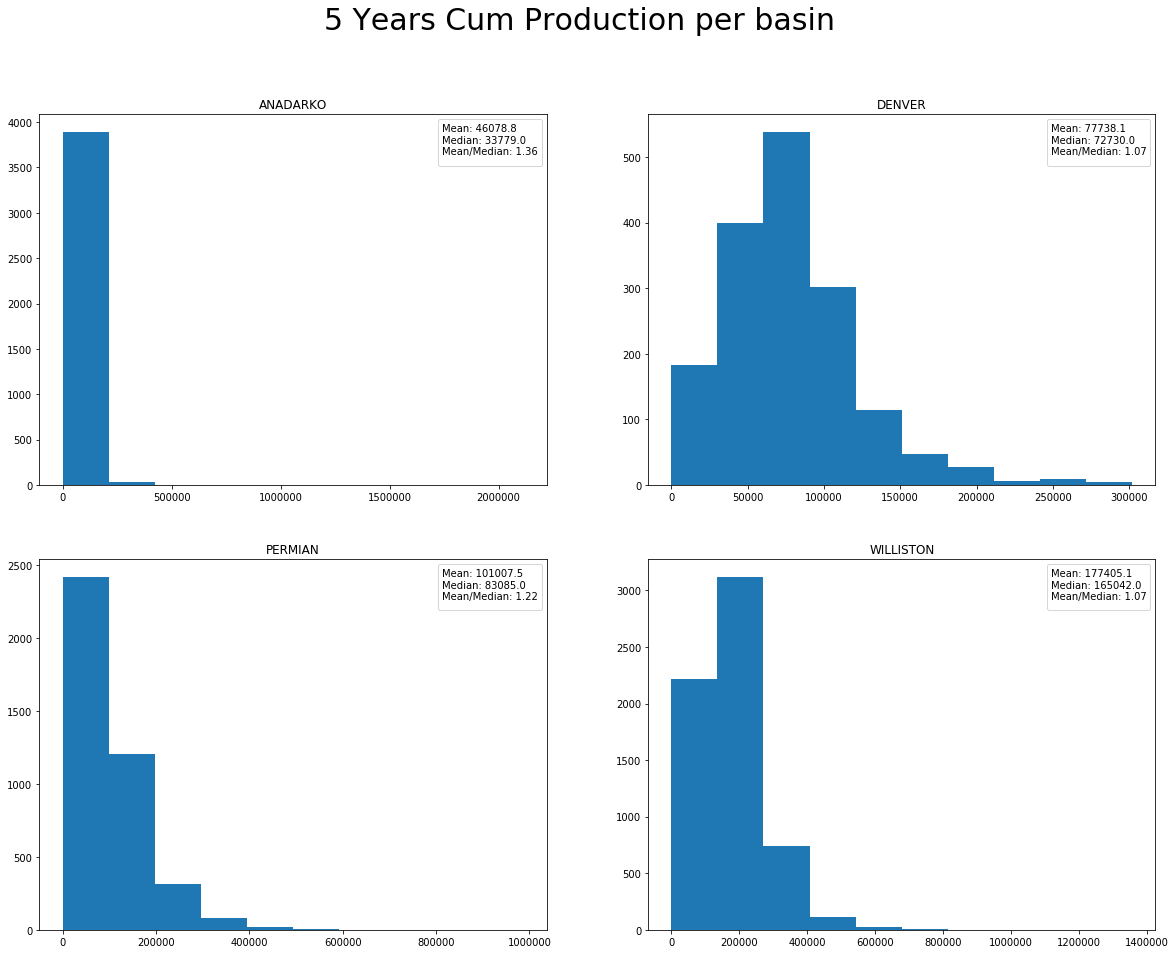
* The months at which the basins production attain a form of low production monotonic horizontal lines increased by a rough average of 11 months after applying the new drilling techniques in 2014.  
  **i.e.** around one year of higher levels of production is added to each basin with the new applied techniques.
* The production amount of wells in each basin acquires similar behavior resulting in close mean and median values. However, the difference between the mean and the median in “Anadarku” indicates the presence of outliers.

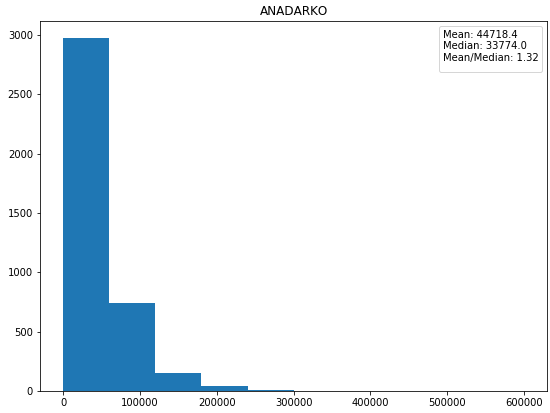
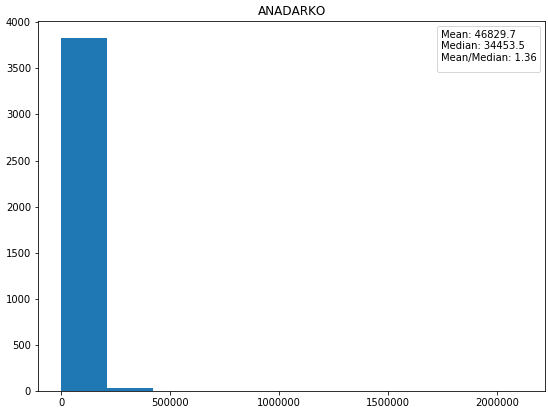
**Plot 2:**



* With plot one taken into consideration, we can infer that basins starting with higher production volume in their peaks, decline with higher percentage rates than those starting with lower levels. **i.e.** peak production value is directly related to the production-declination percentage.

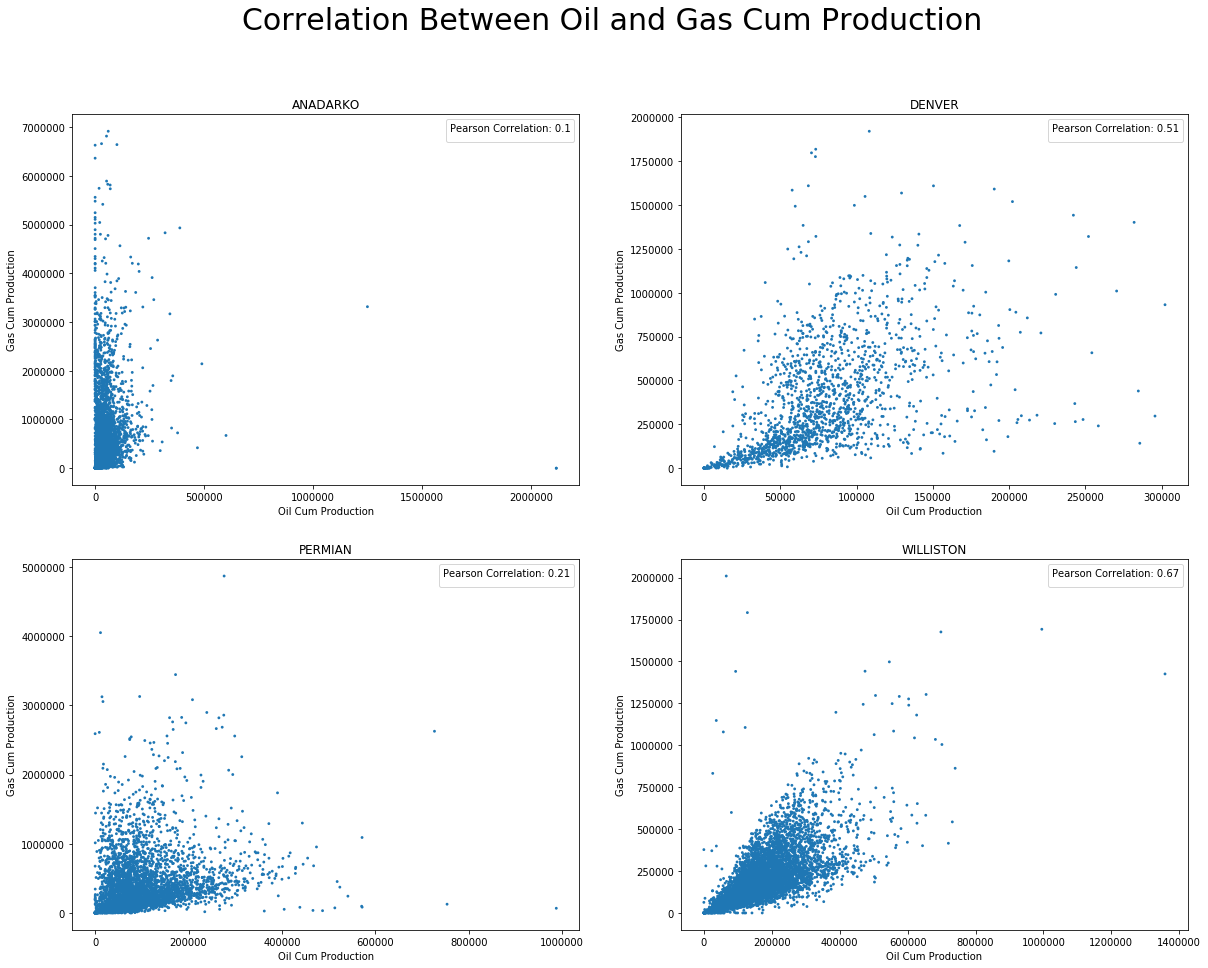
**Plot 3:**





* As the weight of the data gets centralized, the difference between the mean and the median values thus, it is obvious that “Anadarku Basin” has the highest number of outliers followed by “Permian Basin”.
* Wells with highest cumulative production were found in the “Anadarko basin” yet we believe that these values can be wrong and might be considered as outliers, since one of the wells achieved a 5 years cumulative production of more than 2M.
  + We included the two figures after removing what we considered outliers
* Basins have high number of wells in the first column [0-X], since we did not filter out wells that have not yet produced oil.
* The average value in the Williston basin is the highest.
* Most of Denver Basin’s wells produced relatively close amounts of liquid

**Plot 4:**



* Oil and gas production is always positively correlated. Moreover, the basins with outliers have smaller correlation coefficients. However, correlation doesn’t imply causation.
* 3 outliers in plot for “Anadarko Basin” were spotted from this plot as seen in the first sub figure, Our curiosity got us to remove them and produce the 2nd figure.
* The highest correlation between oil and gas was found in the “Williston Basin”, yet it is still low.
* In the “Andarko Basin” there is very low correlation between the 2 variables, our guess that this area does not have much oil to produce while the gas production rates are high (check when oil production is zero).

**Plot 5:**