**EXPLORTATORY DATA ANALYSIS**

**Predicting Oil Production for Oil & Gas Fields**

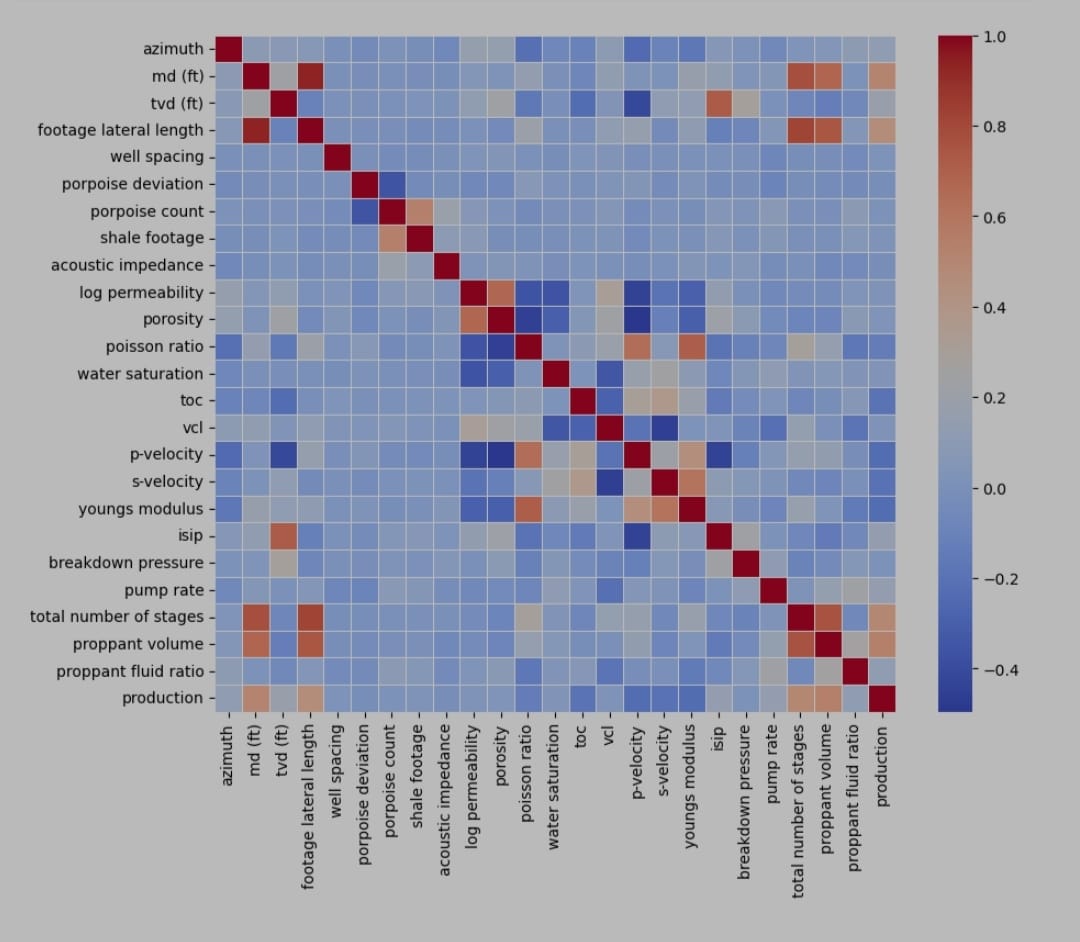
**ASSIGMENT 4**

**By-**

**IPBLGCET TEAM -03**

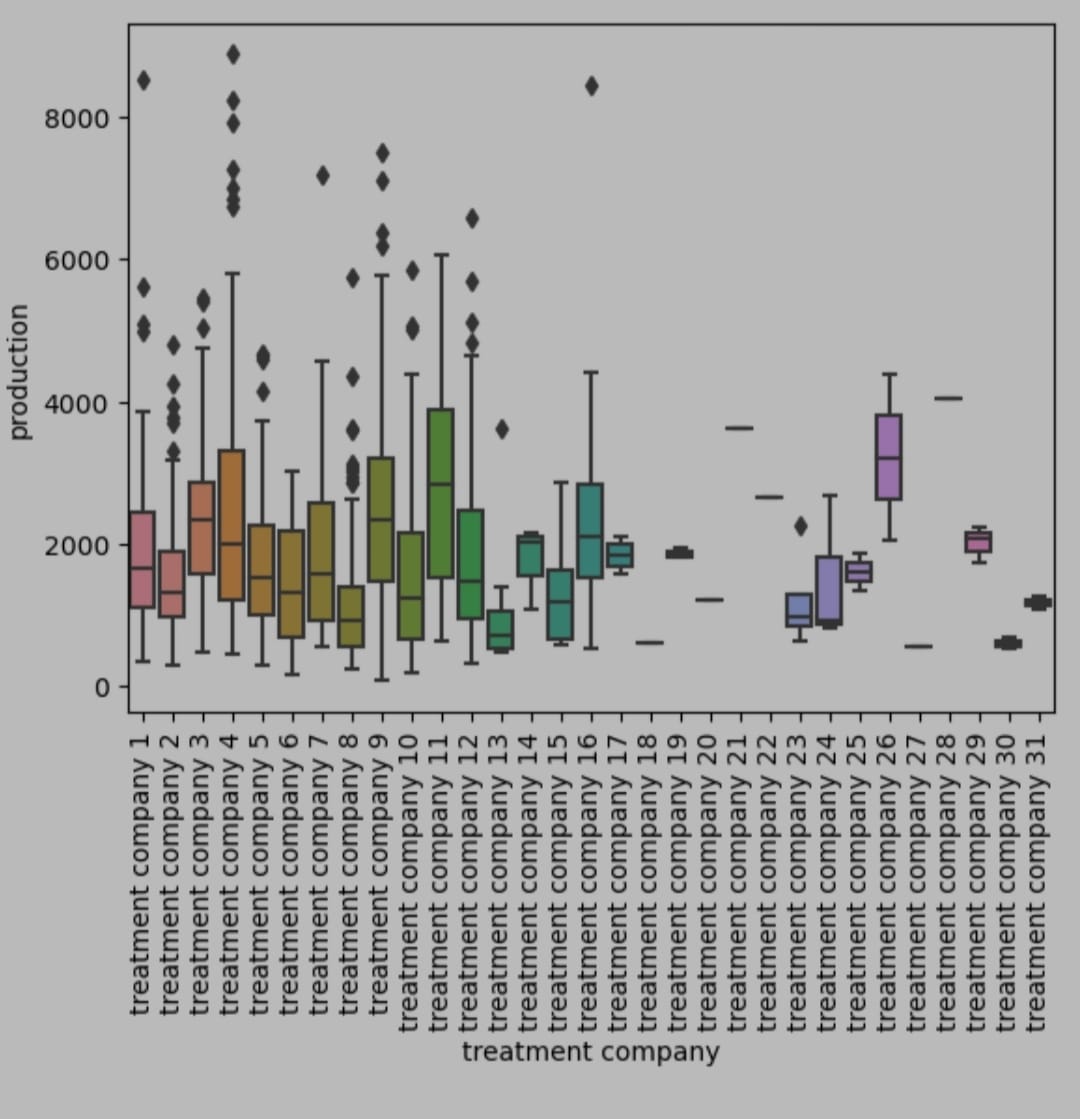
**Team Members:**

* A.Harshith Sai (Team Lead)
* P.Rohith
* G.Harshit
* G.Pranava sai
* **The methodology should include the following operations for regression modelling**
* **Import the dataset**
* **Clean the dataset removing outliers, NA values and unnecessary features.**
* **Explore the data to create hypothesis, think about a few insights and validate them.**
* **Prepare the data to be used by the modeling algorithms encoding variables, splitting train and test dataset and other necessary operations.**
* **Create the models using machine learning algorithms.**
* **Evaluate the created models to find the one that best fits to the problem.**
* **Tune the model to achieve a better performance.**
* **Use the dataset oilgas.csv for the Project**
  1. **HEAT MAP**

****

**From the above heat map we can observe the relationship between different parameters and there is nearest relationship between footage lateral length and M.D and both of them are directly proportional to each other**

**1.2 BOX PLOT**

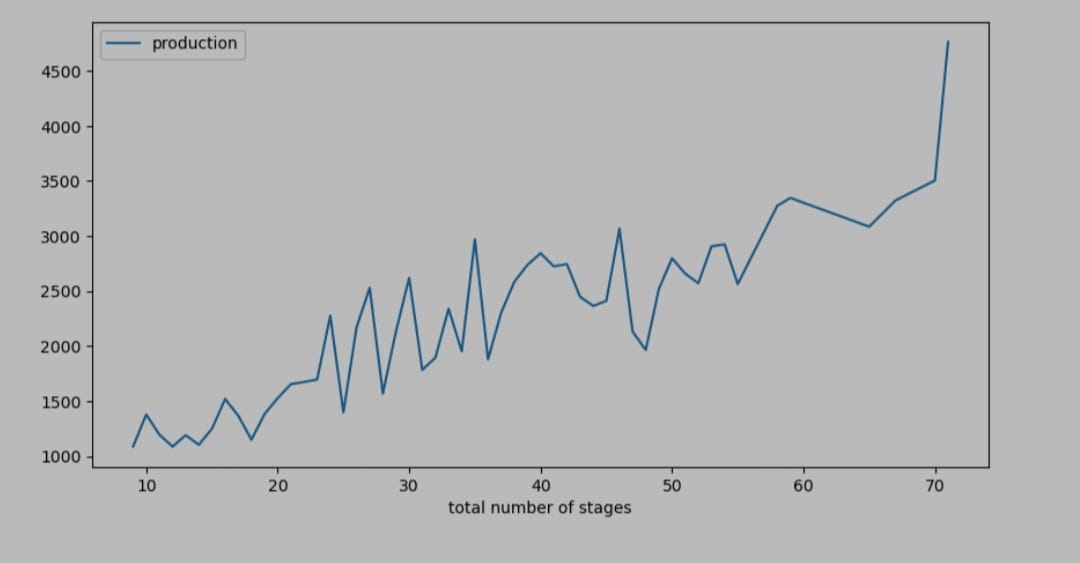
****

**The above box plot describes about relationship between treatment company and total production.By observing the above plot we can conclude that treatment company 11 has highest production and followed by treatment company 26.**

**Also we can observe that treatment company 26 has the highest mean value.**

**So in my opinion I can conclude that for production treatment company 26 is the best choice.**

**1.3 line graph**

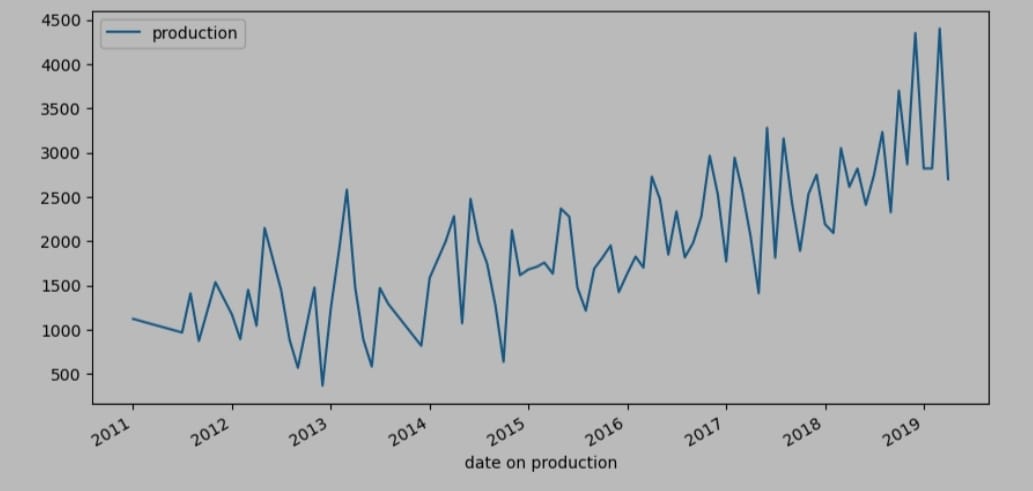
****

**The above graph is line graph between total number of stages and production.**

**So from the graph we can observe that as the total number of stages are increasing the production values increases**

**As the total number of stages values increases more than 70 there is drastic increase of production line**

**1.4 line graph**

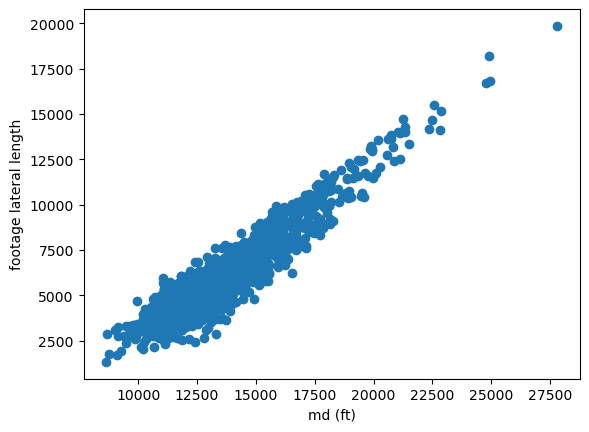
****

**The above graph is line graph between date of production and production.**

**So from the graph we can observe that there is a high volatility observed in the graph but as a final picture while the date of production increases automatically production also increases.**

**From this we can conclude that as time is being on due to new technologies production is increasing gradually year by year**

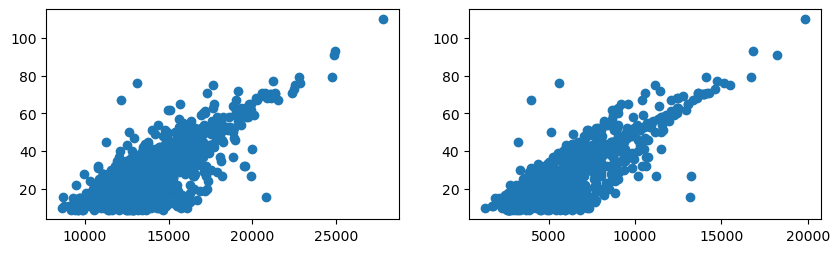
**1.5 count plot**



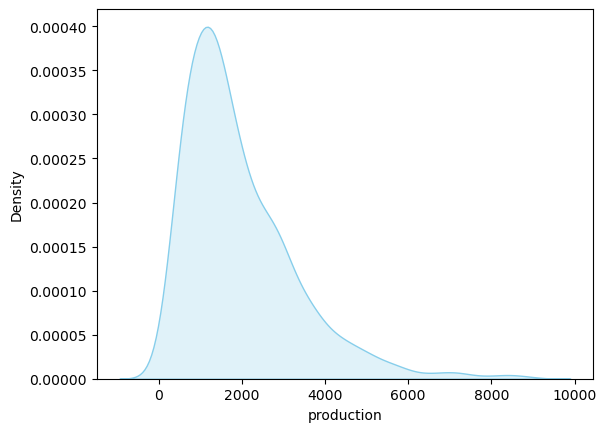
As we can see from the above plot the footage lateral length is increasing along with the measured depth.

Other correlations that can be considered are measured depth,total number of stages and footage lateral length,total number of stages which have the correlation values as 0.773874 and 0.822597respectively.

Scatter plots for above relations are as follow



1.6 **distribution graph and box plot**



The production of gas from the year 2011 to 2019 is ranging from 77 mmcf

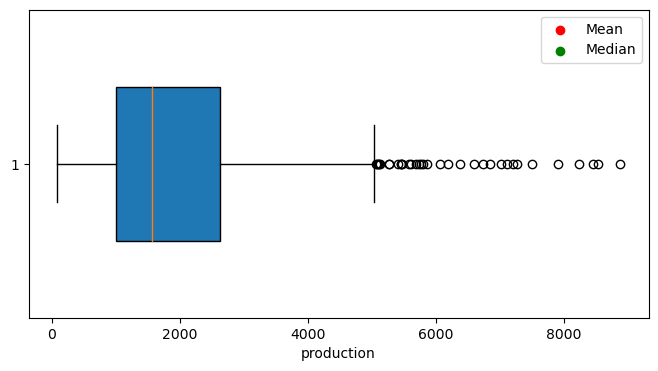
to 8880 mmcf.

In the most of the time the production is in between 77 mmcf to 4000mmcf

The distribution of production is right skewed .Median of the production is

less than the mean.

This means that there are a few relatively large values that are pulling the mean higher than the median.This can be observed in the following box plot,



**SOLUTIONS**

* Treatment company 26 is the best choice for production due to high production mean value than other treatment companies
* The production should be initiated whenever there is less s-velocity and p- velocity
* It is preferable to increase total number of stages as it is increase in production
* Azimuth angle should be maintained nearer to 90 degrees such that drilling becomes simple and efficient
* Oil drilling should be done at a place where is less water saturation and well spacing