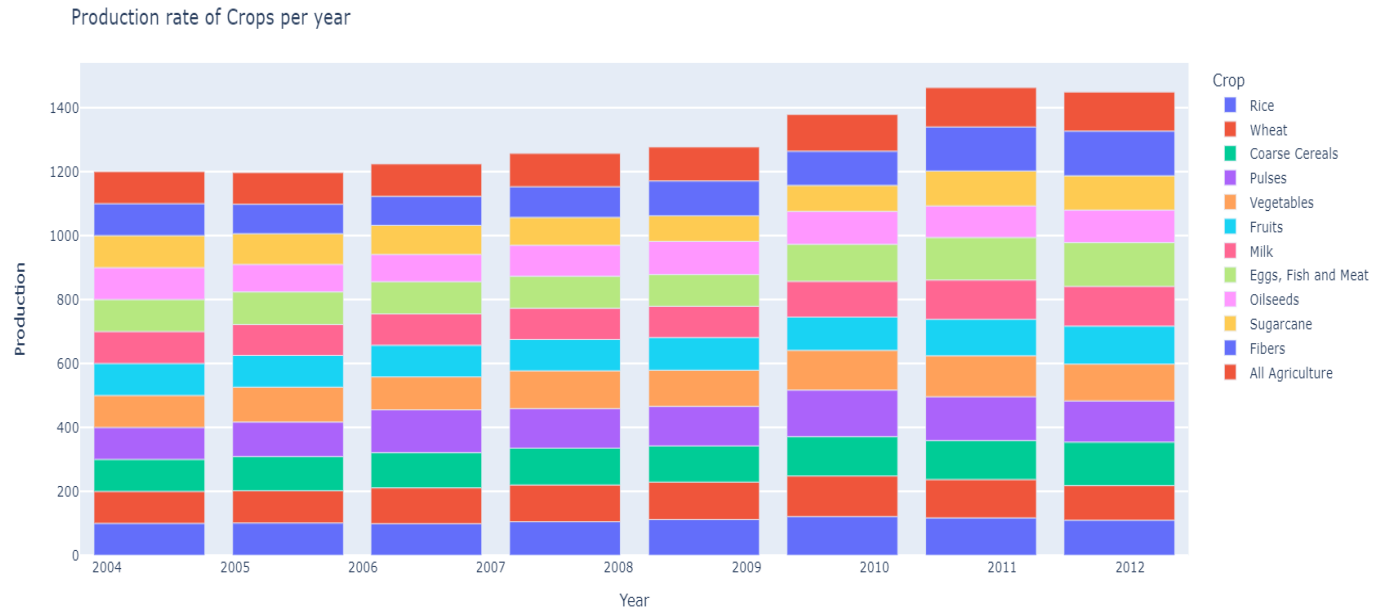
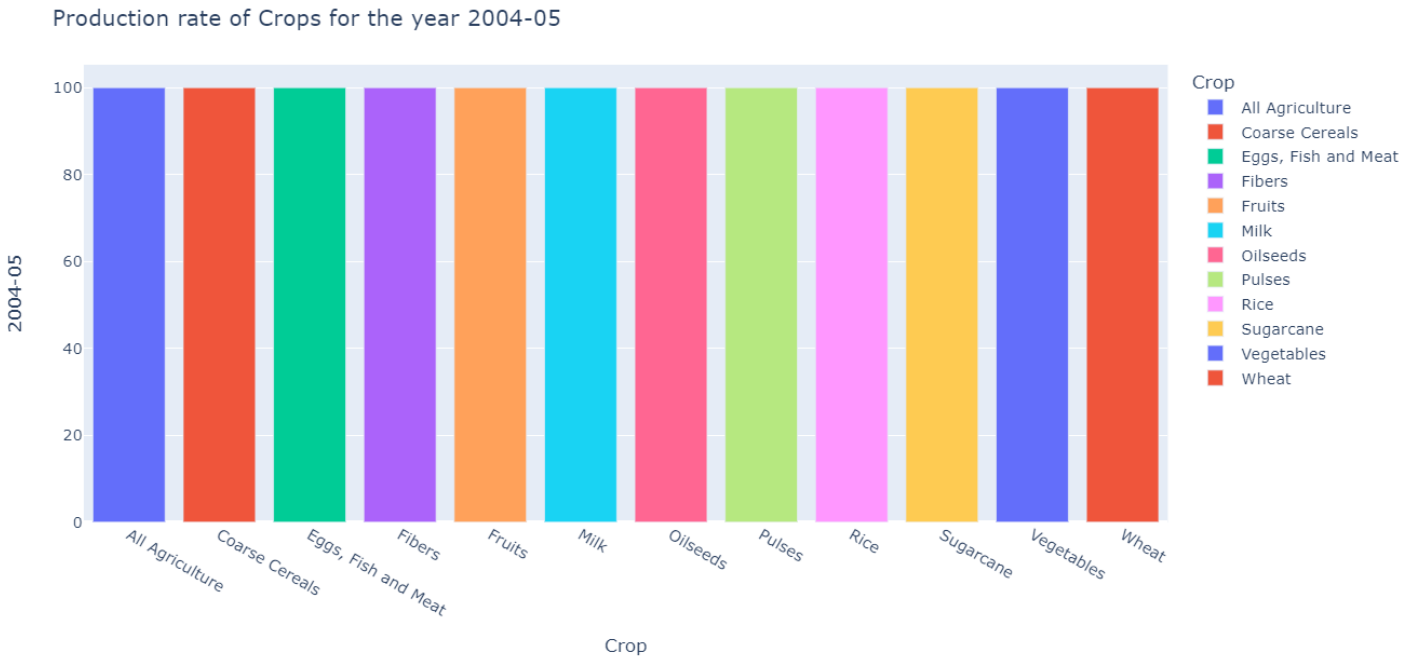
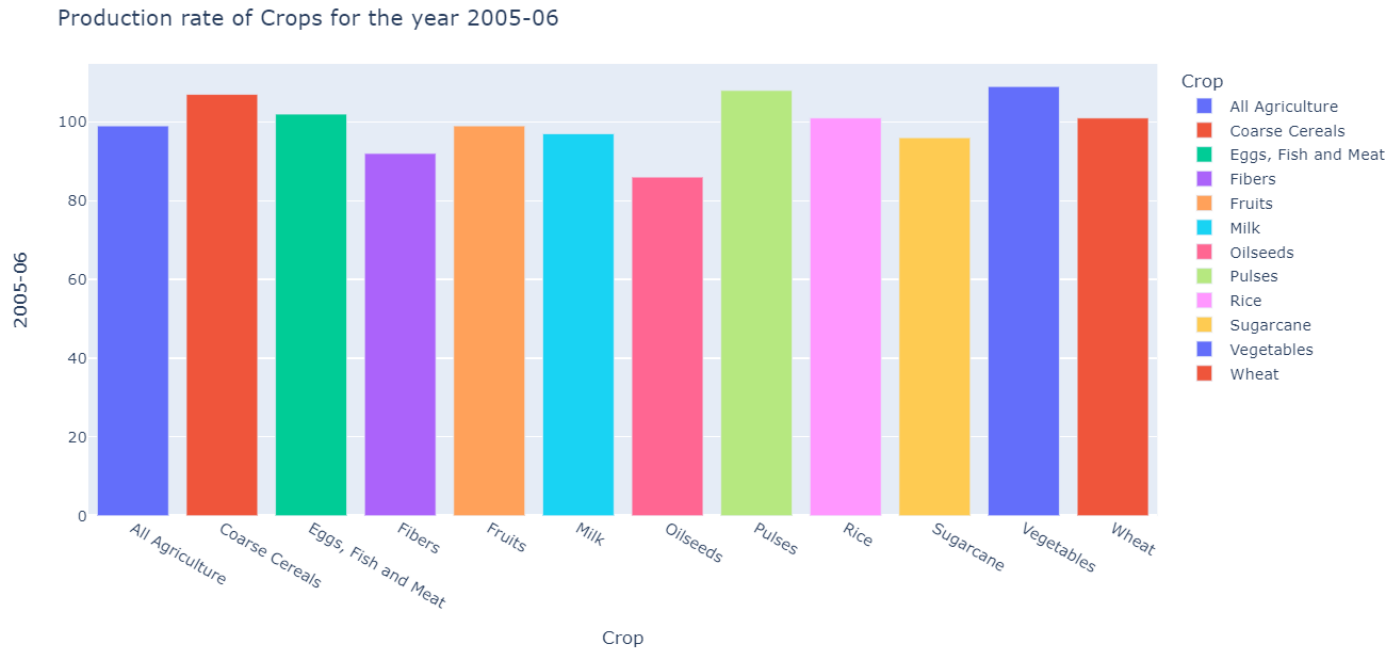
I have chosen a dataset containg the agricultural crops and their production , cost of cultivation, yield per year for the different states of India are given.The plot below shows the production rate of crops from the year 2004 to 2012.



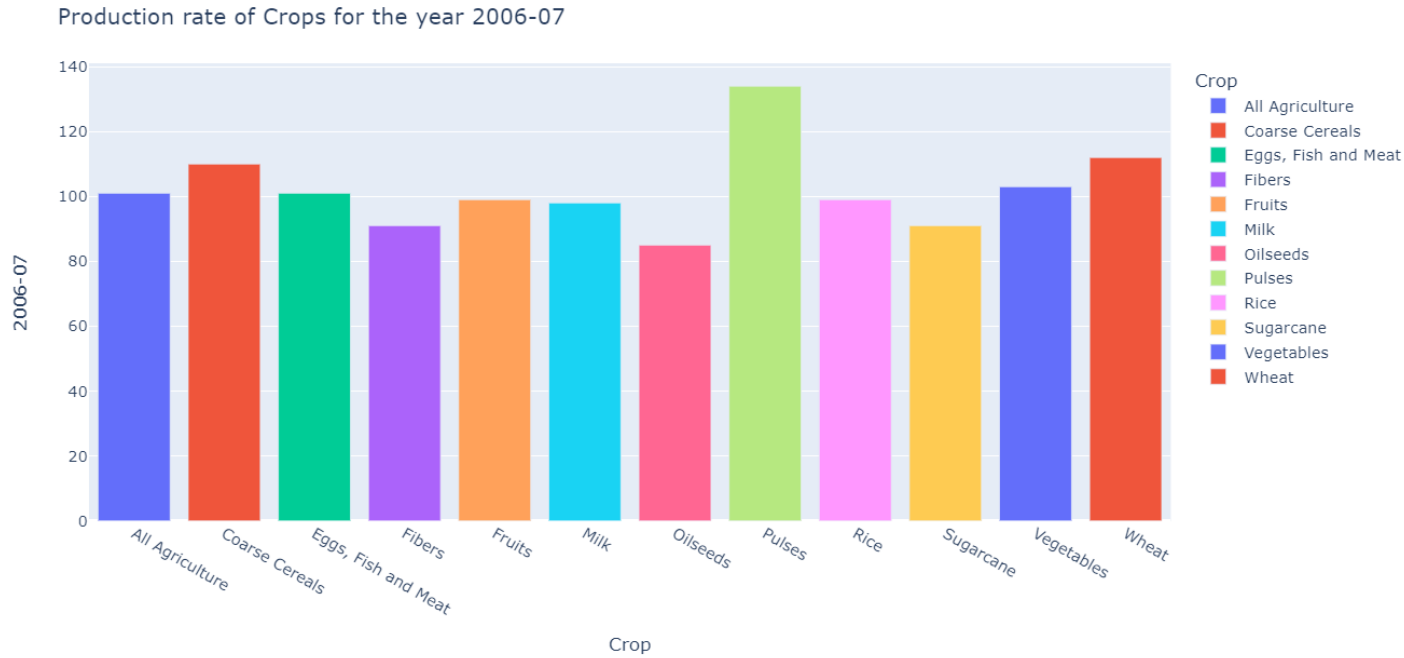
We can observe from the following graph that in the year 2004-05 all the crops had equal amount of production.



From the following graph for the year 2005-06 we can see that vegetables had the highest amount of production followed by pulses and coarse cereals.

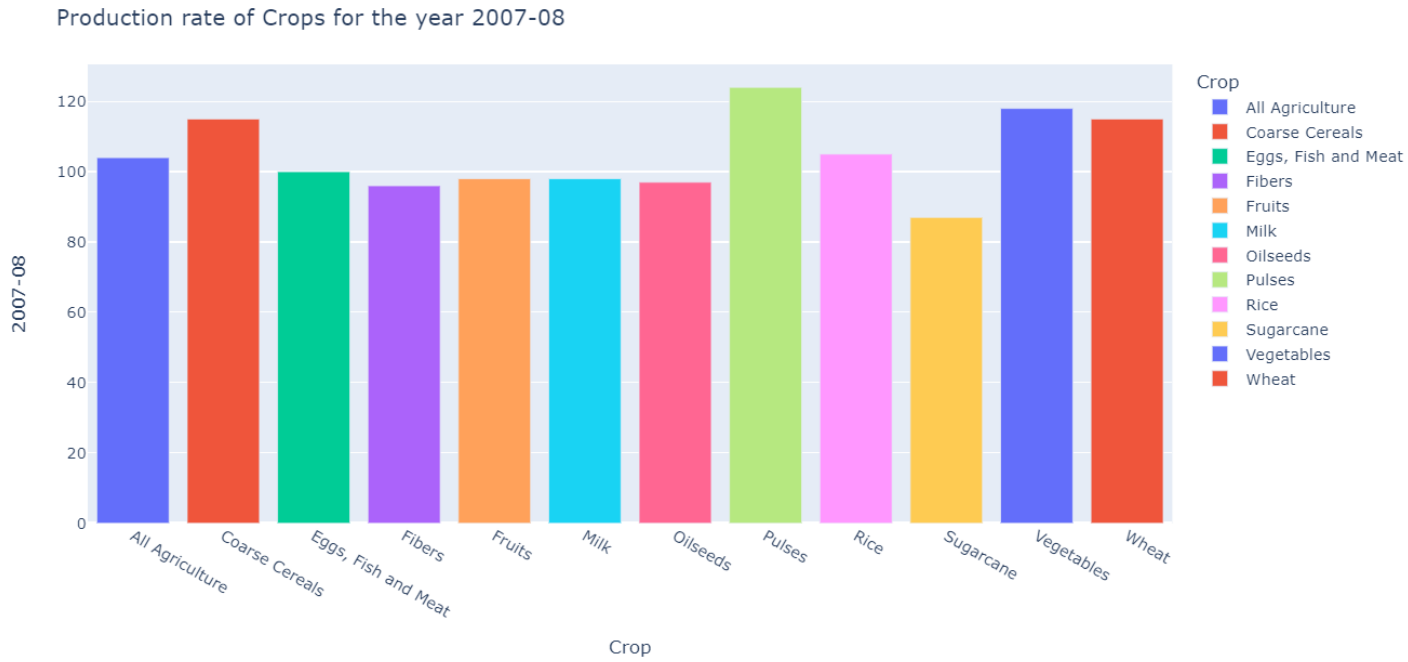


From the following graph for the year 2006-07 we can see that pulses had the highest amount of production rate.

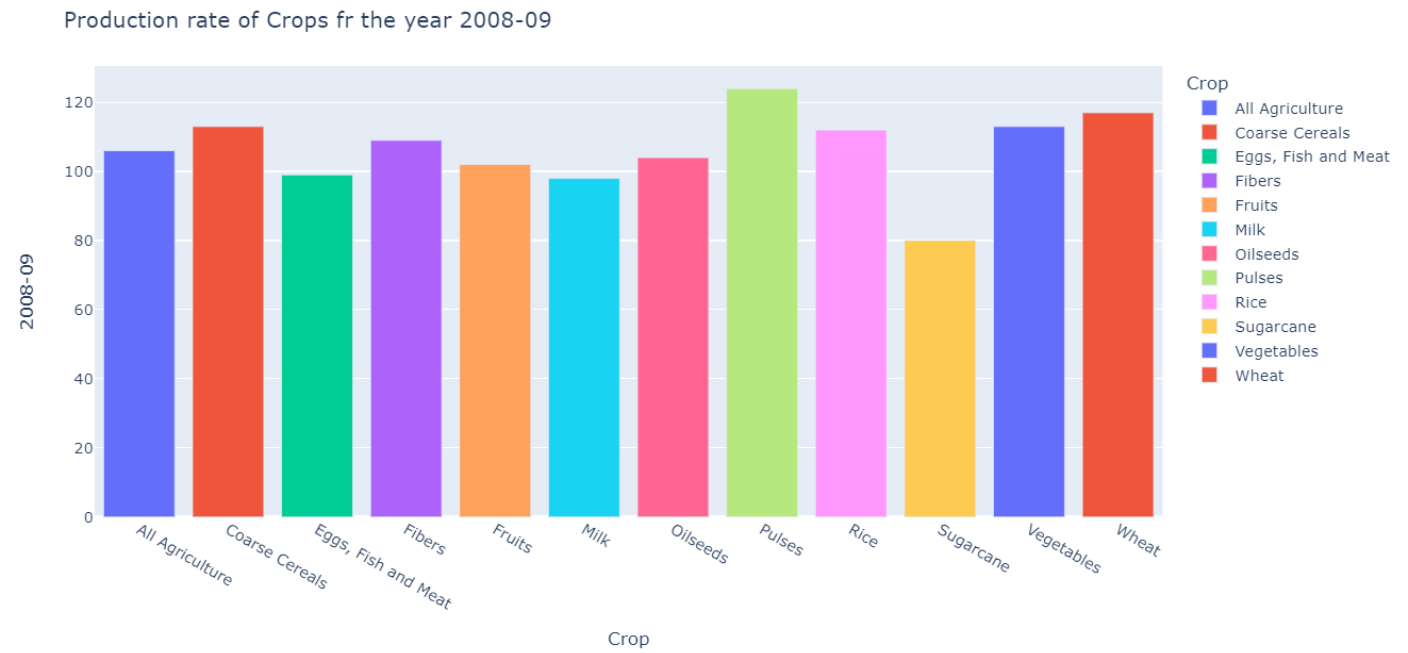


In 2007-08 vegetables had the highest amount of production rate followed by

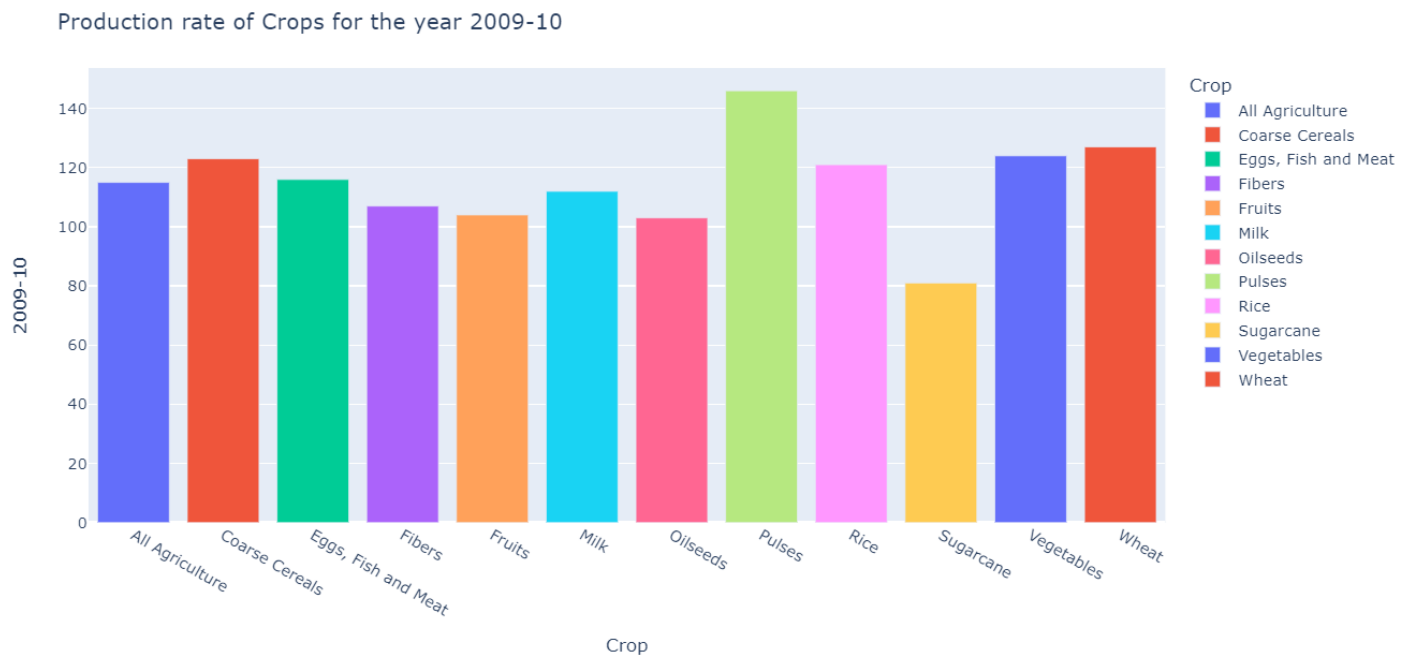
vegetables and wheat.



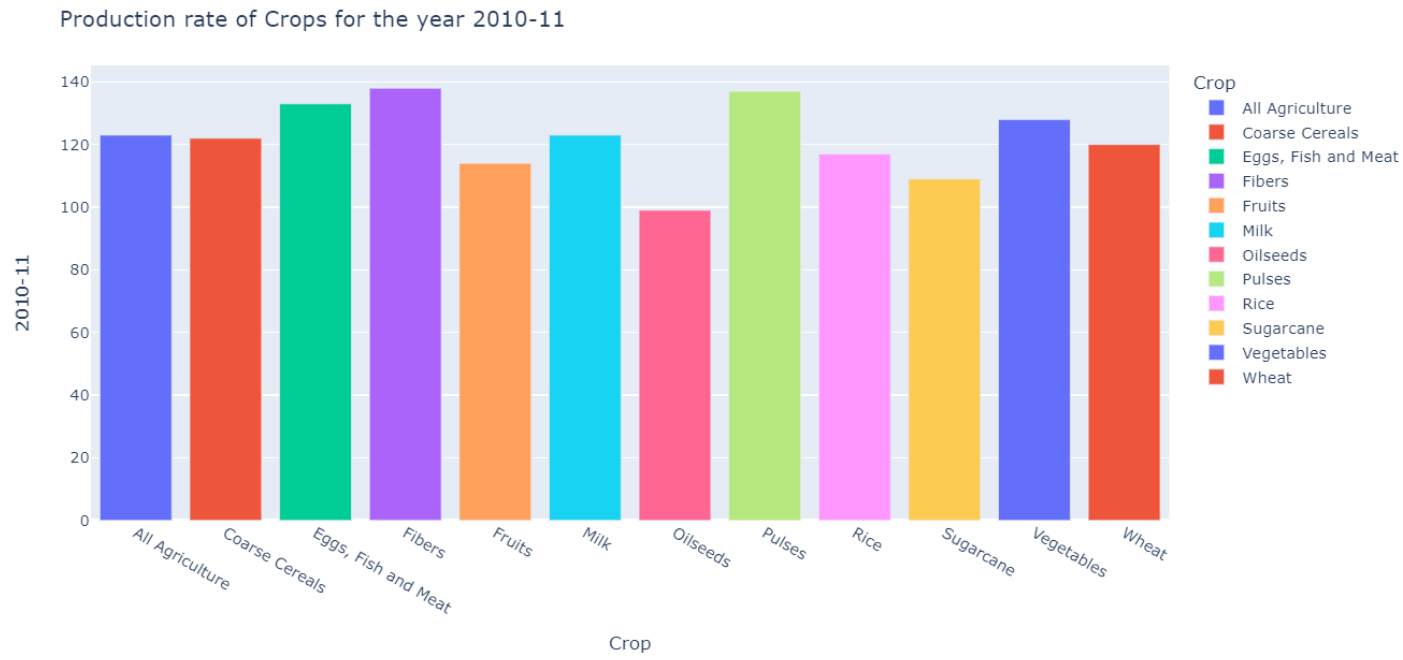
In 2008-09 pulses had the highest amount of production rate followed by wheat and coarse cereals.



Again in 2009-10 pulses had the highest amount of production rate.

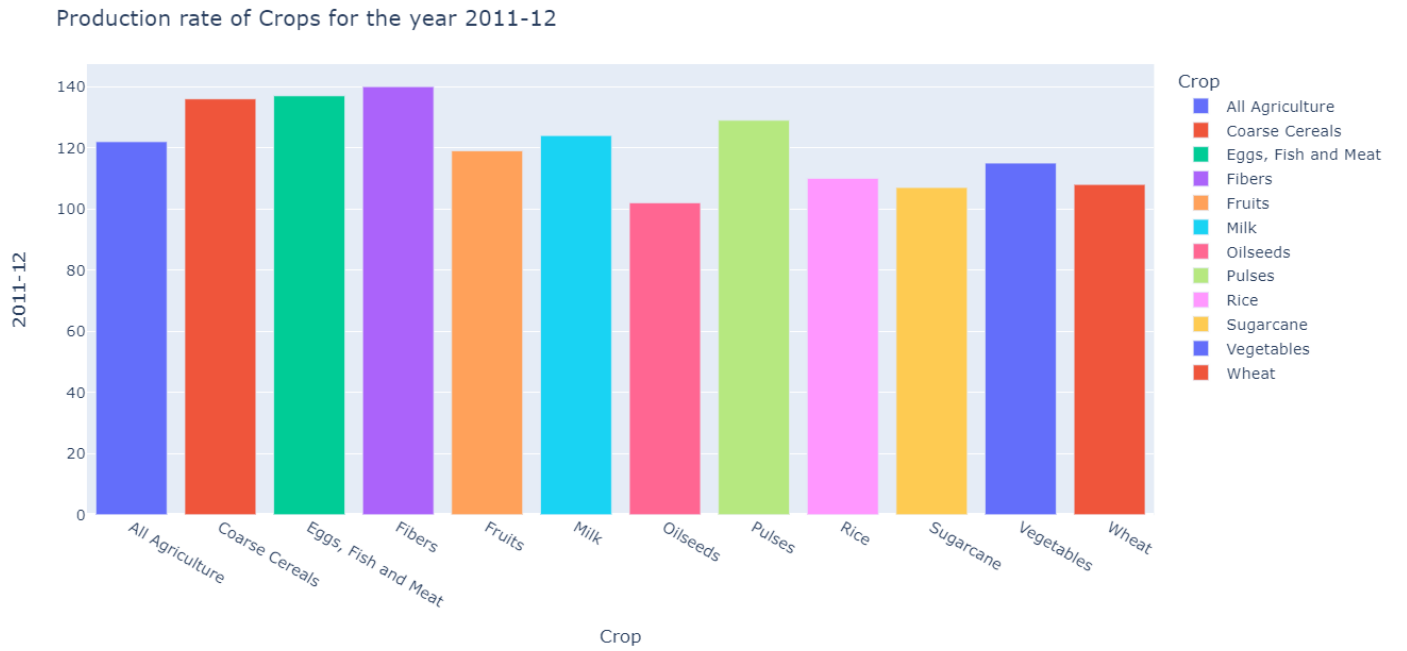


In 2010-11 Fibers had the highest amount of production rate followed by pulses.

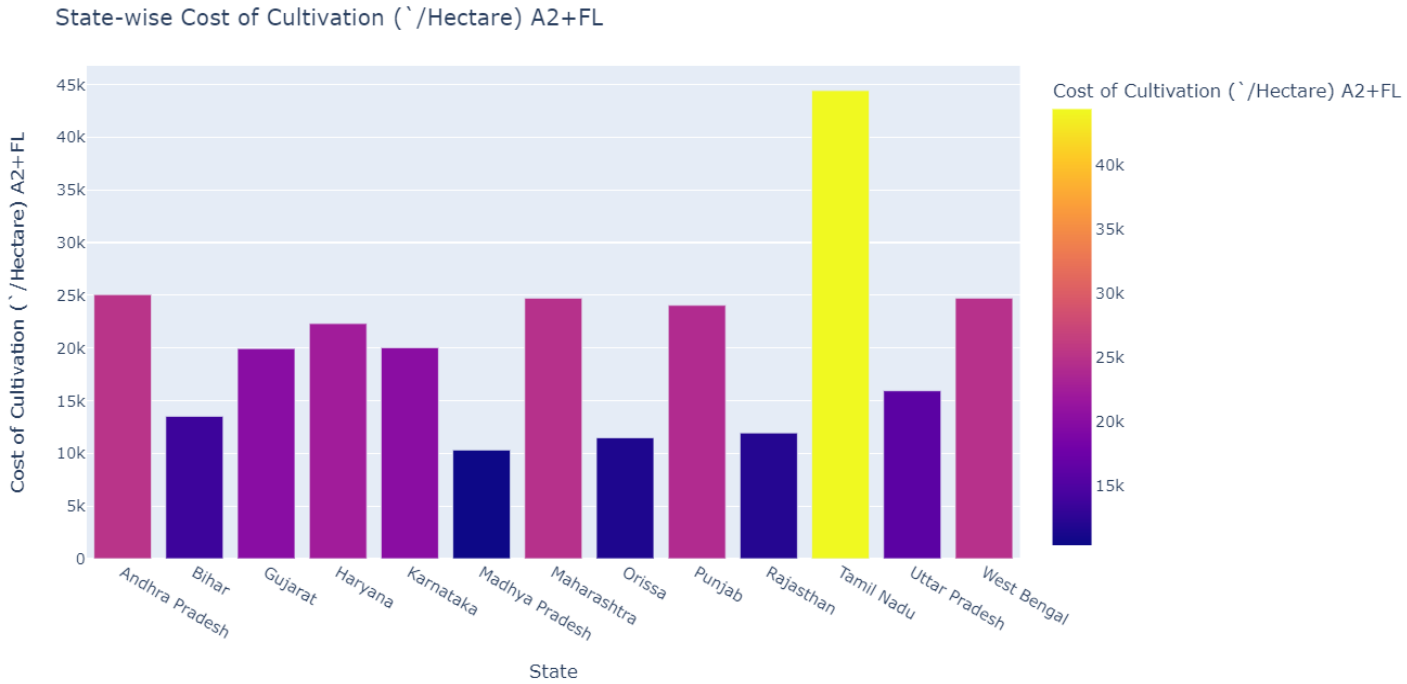


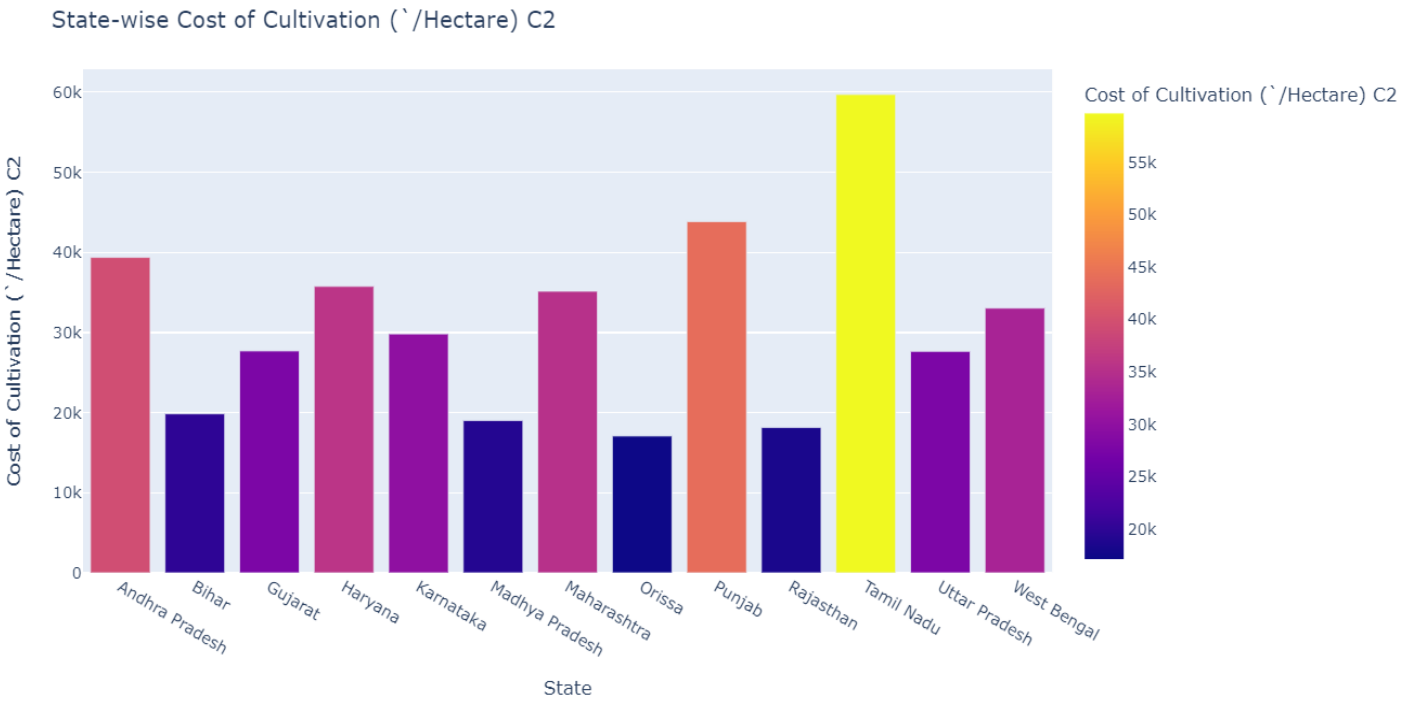
In the year of 2011-12 fibers had the highest amount of production rate followed by

eggs,fish and meat and coarse cereals whereas the production rate of vegetables went down.

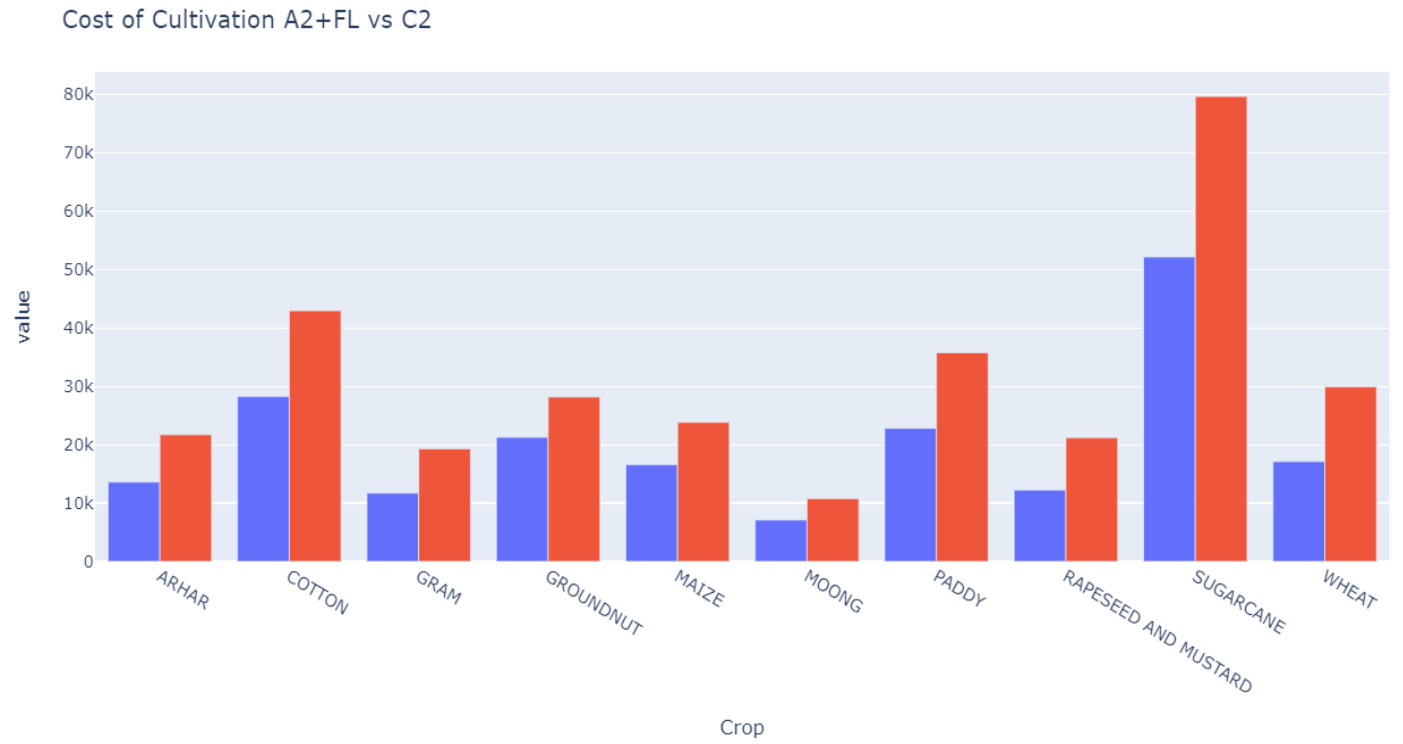


The following plot shows the state-wise cost of cultivation (/hectare). We can clearly see that Tamil Nadu had the highest cost of cultivation of more than 40000 followed by Andhra Pradesh.

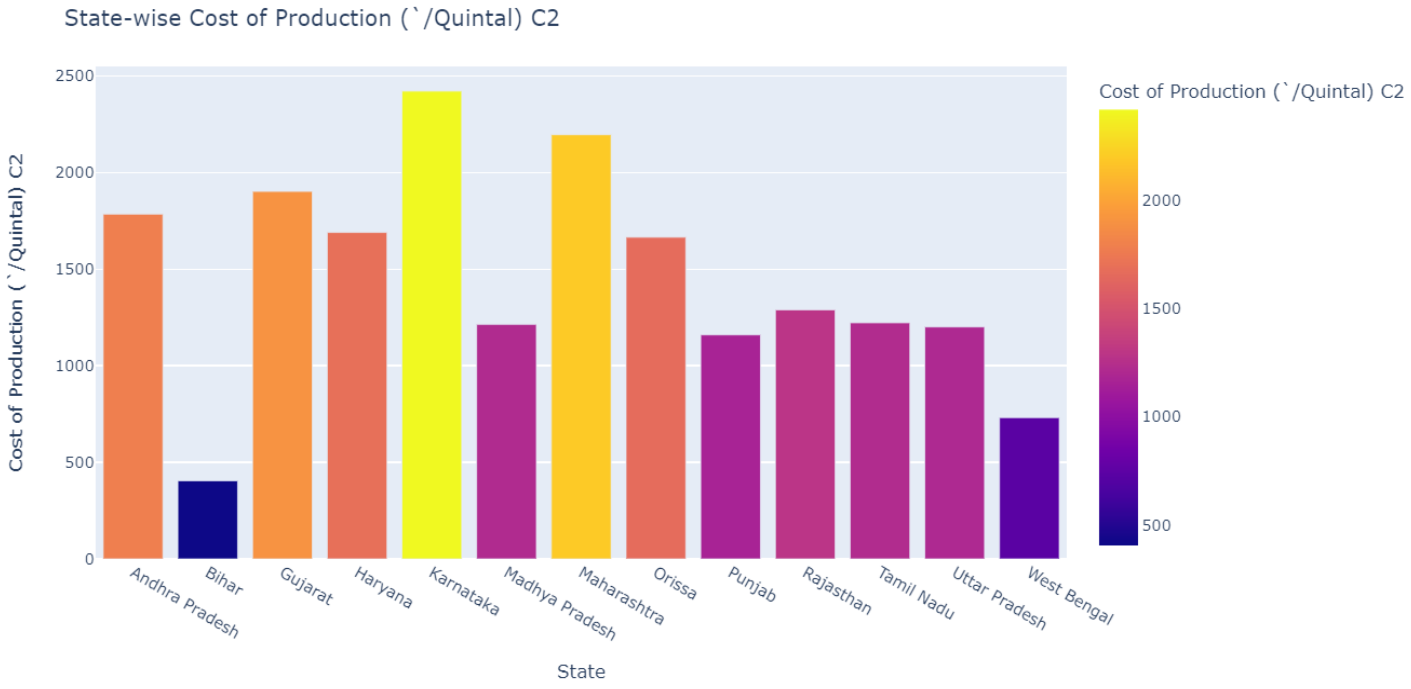




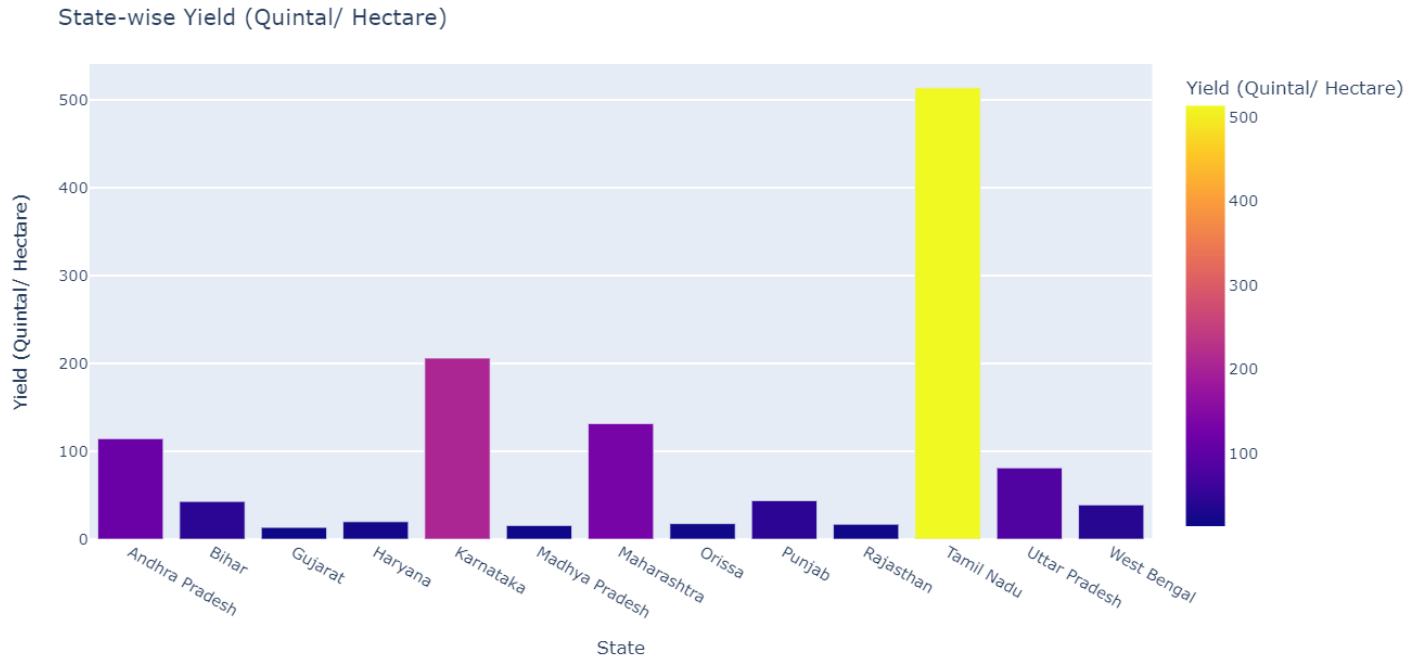
The following graph shows the cost of cultivation A2+FL vs C2. We can clearly see that for all the crops cost of cultivation C2 is more than cost of cultivation A2+FL.



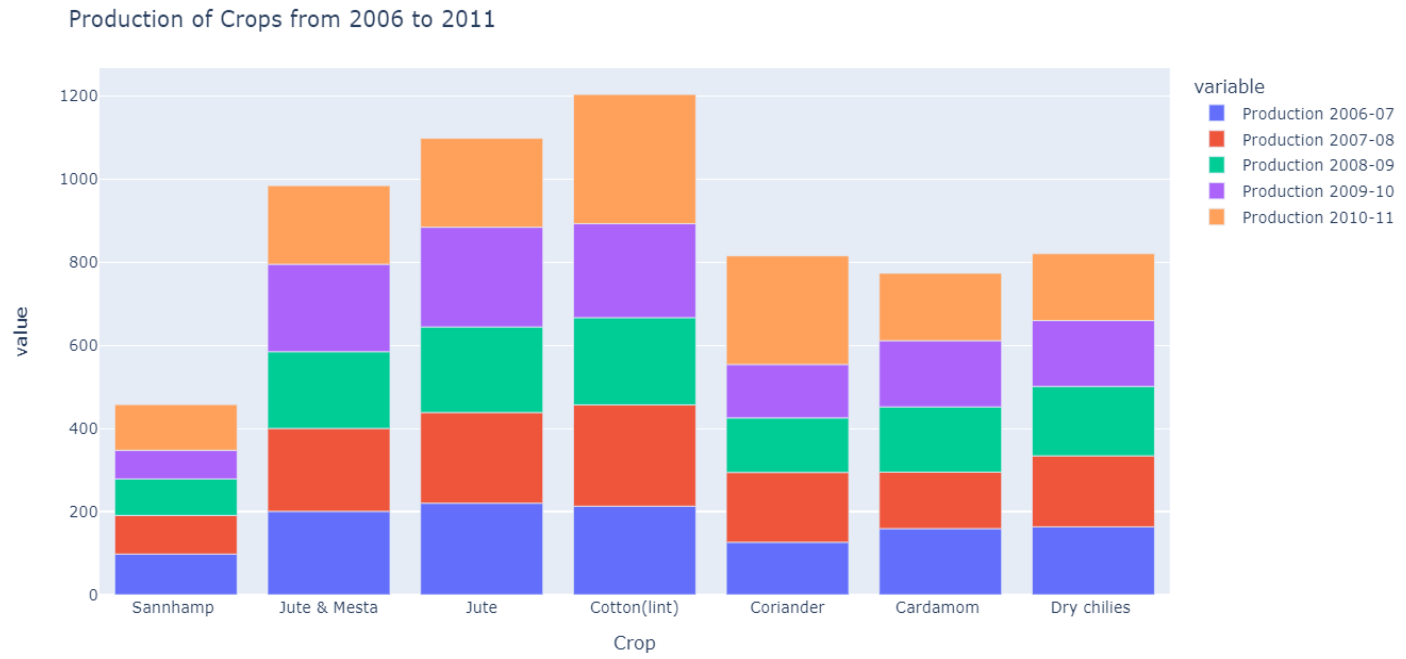
Next is the plot of state-wise cost of production (/ Quintal) C2. We can see that Karnataka had the highest cost of production followed by Maharashtra, Gujarat and Andhra Pradesh whereas Bihar had the lowest cost of production.



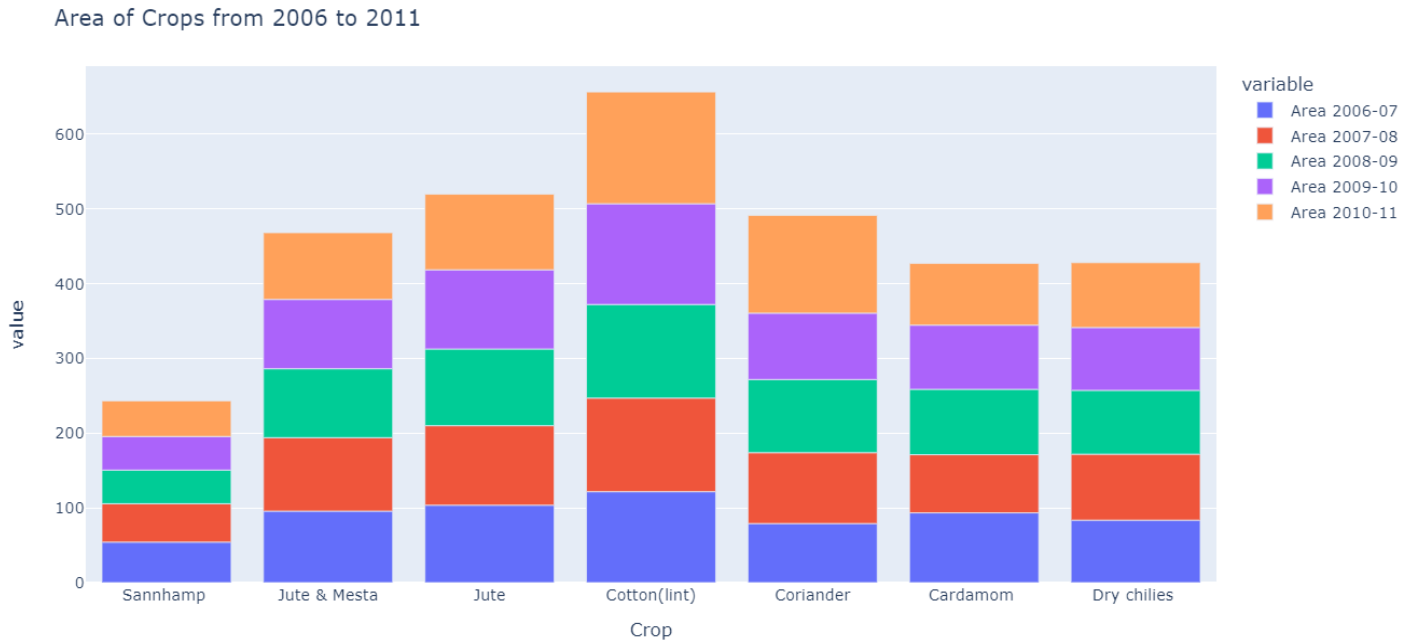
The following graph shows the state-wise Yield (Quintal/Hectare).Tamil Nadu shows a phenomenal amount of yield of crops. But one more thing to keep in notice is the yield is decided by Quintal/Hectare. So if the hectare amount is low compared to other states then also the figures will go high but in reality production won't be that high.



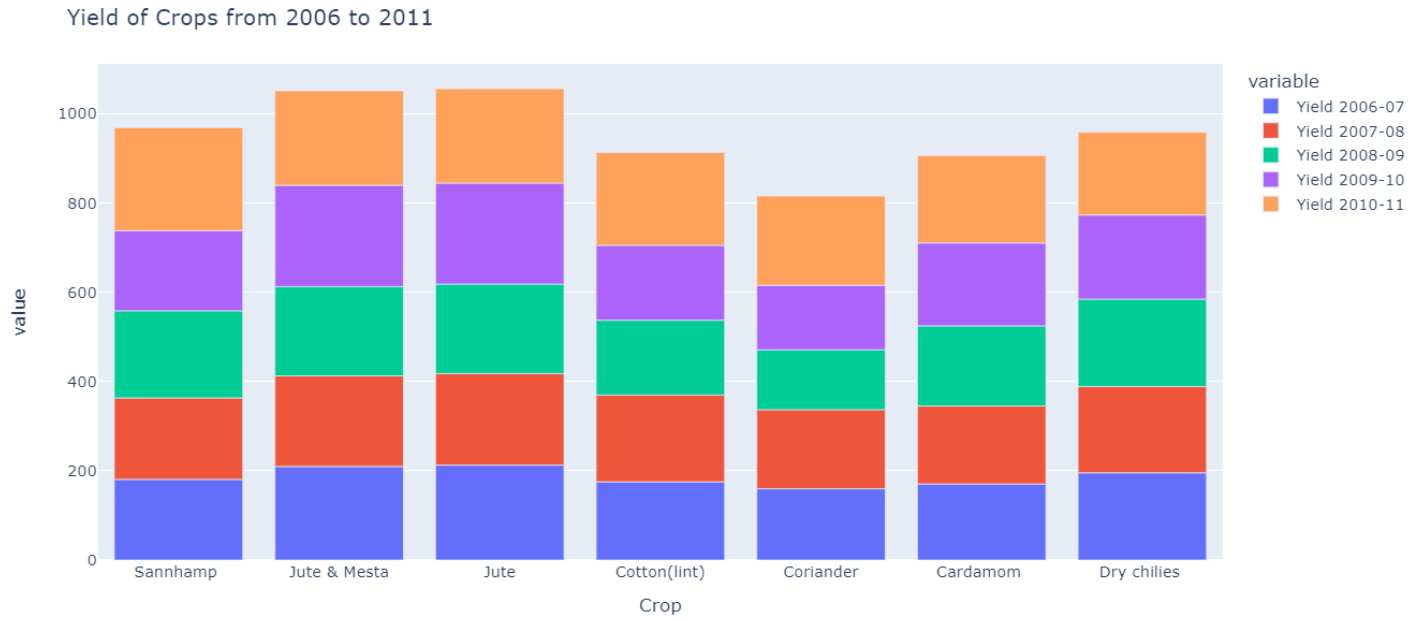
The following graph shows the production of crops from the year of 2006 to 2011. Cotton had the highest amount of production in all these years whereas sannhamp had the lowest production.



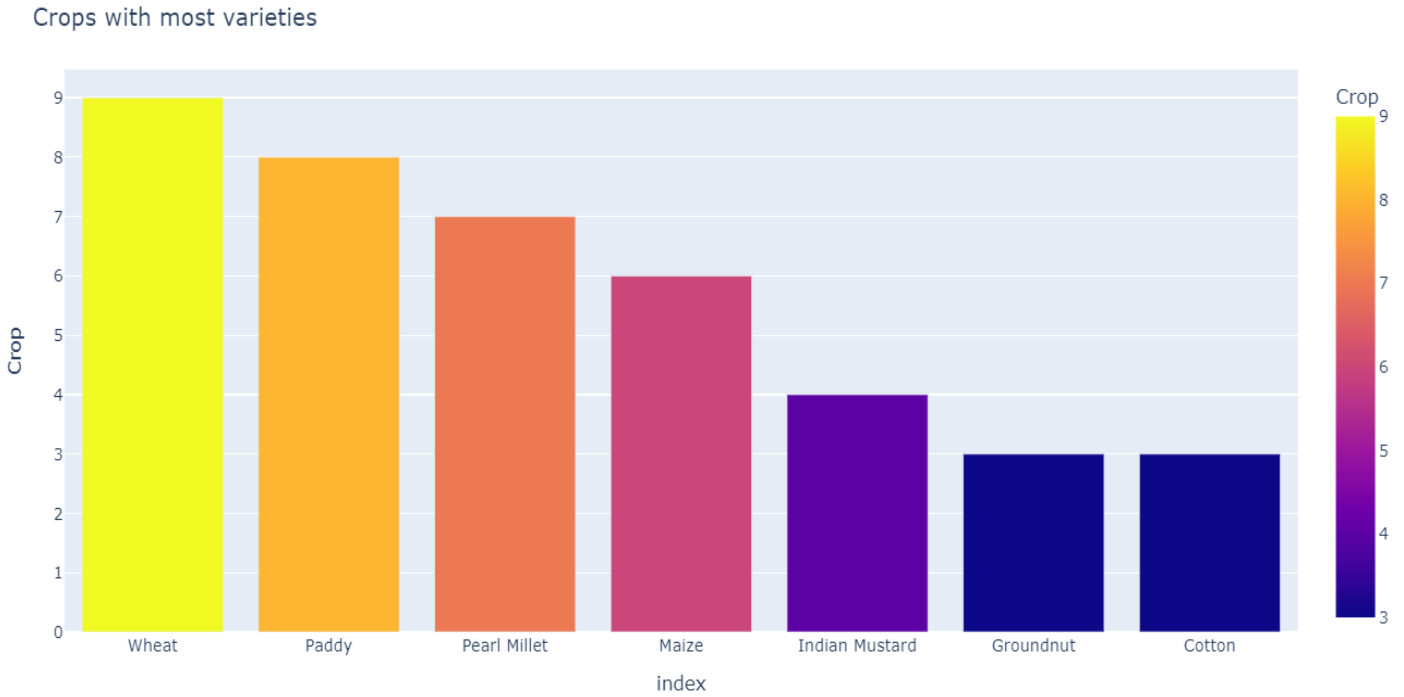
Next we have the area of land occupied by the crops from the year of 2006 to 2011. Again cotton occupied the largest area whereas sannhamp covered smallest area.



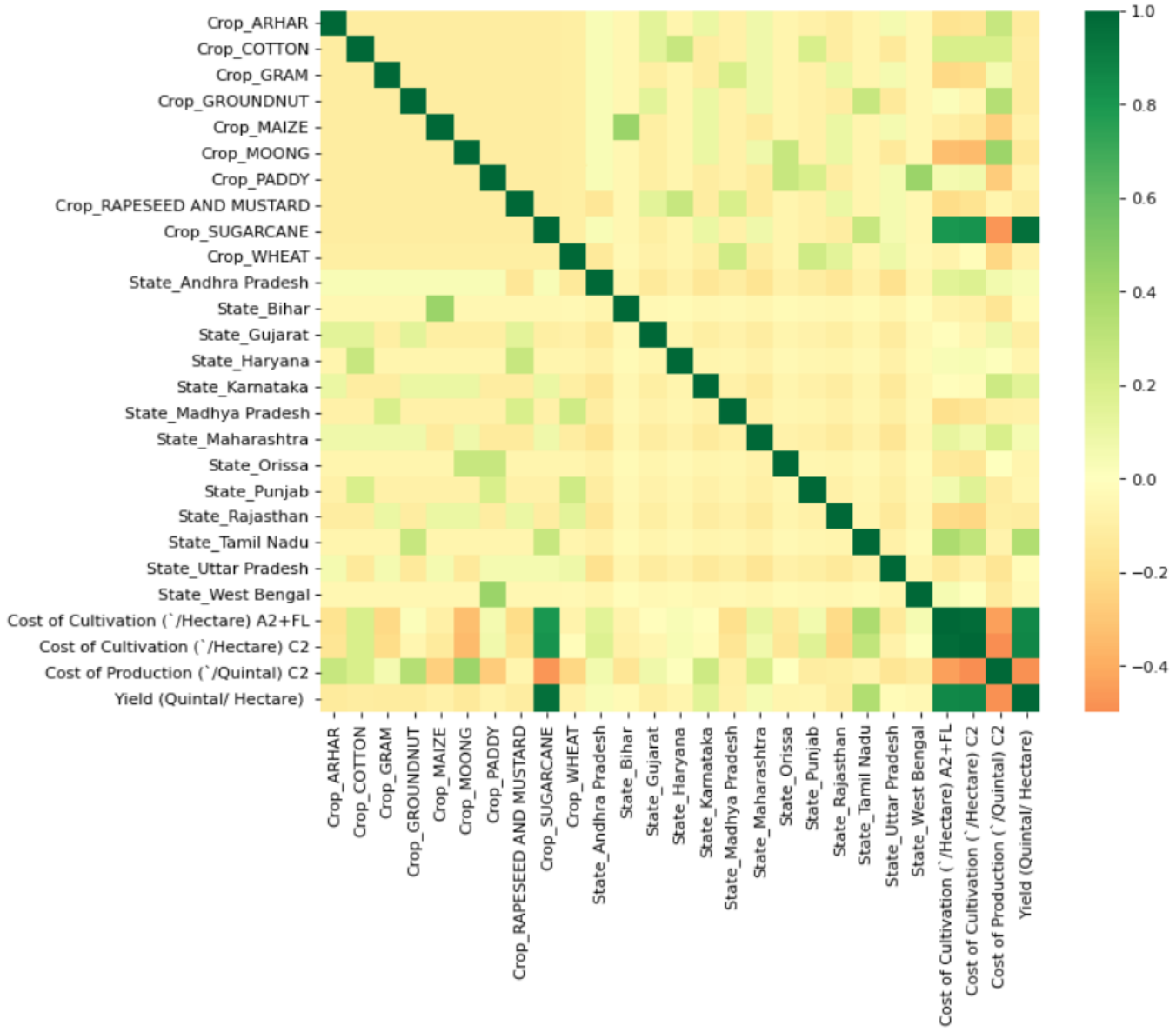
The following plot shows the yield of crops from the year 2006 to 2011.



The graph below shows the crops with most varieties. We can see that wheat had the largest amount of varieties.



For the prediction of the yield I have used datafile(1) dataset which I have uploaded in the github. It includes Crops, State, cost of cultivation, cost of production as the feature variables. I have considered the Yield feature as the target variable.



From the above correlation heatmap we can see that sugarcane has a high correlation with Yield and cost of cultivation whereas the same sugarcane negative correlation with cost of production. Similarly cost of production has a negative correlation with cost of cultivation.

So finally I used all these features and built a Machine learning model which predicts the yield of the crop based on the above features. I used Random forest regression model to predict the yield and achieved an r squared score of 0.95.