Introduction to Data Science

Aug-Dec 2018

DATA ANALYSIS OF VARIOUS SOCIAL ISSUES

Project Report

Jigyasa Yadav(16UCS084)
Mansi Vijay(16UCS101)
Mani Goyal(16UCC052)

Acknowledgement

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.We are highly indebted to Dr Sakthi Balan and Dr. Subrat Dash for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. Our thanks and appreciations also go to our colleagues and other batchmates in developing the project and people who have willingly helped us out with their abilities.

Abstract

This project addresses various social indicators like population,infant mortality rates,sex ratio in rural and urban areas,literacy rates among women in different age groups,children who received and not received vaccinations etc of our country.It is very important to analyze this data because it will help the government to make policies and take decisions to improve them and increase the living standards of people in our country.

The data set has been collected from Kaggle https://www.kaggle.com/rajanand/key-indicators-of-annual-health-survey . The data is of Annual Health Survey 2012-2013 which was conducted in 9 states which are Uttarakhand, Rajasthan, Uttar Pradesh, Bihar, Jharkhand, Odisha, Chhattisgarh & Madhya Pradesh and Assam.The population in these listed states has been analysed on parameters including:

1. Total Population in respective districts
2. Number of Married women in the age group 15-49 yrs in both rural and urban areas
3. Dependency Ratio
4. Married Illiterate Women in the age group 15-49 yrs in rural and urban areas
5. Sex Ratio at birth
6. Marriages among females below 18 yrs in rural and urban areas
7. Number of differently abled per 100,000 people
8. Crude Birth Rate
9. Crude Death Rate
10. Natural Growth Rate
11. Children aged between 6-14 yrs attending school
12. Children who have/ haven't received Polio dose at birth in rural and urban areas
13. Infant Mortality Rate

The data was collected from all the districts of these states.About 21 million people and 4.32 households were covered including both rural and urban were covered in this survey. It originally contains 284 rows and 644 columns but for our project we have considered 284 rows and 29 columns containing major important indicators only.

Some specifications of the data are as follows:

Column Name	Unit in which it is measured
Currently Married Illiterate Women aged 15-49 years	%
Children who have received Polio dose at birth	%
Children who did not receive any vaccination	%
Children Currently Attending School Age 6 17 Years	%
Marriages Among Females Below Legal Age 18 Years	%
Currently Married Women	%

NEED FOR PREPROCESSING

Sometimes we encounter data which is incomplete,has missing attribute values , has data fields inconsistent with rest of the data or contains many errors. In order to resolve these issues data preprocessing is used.

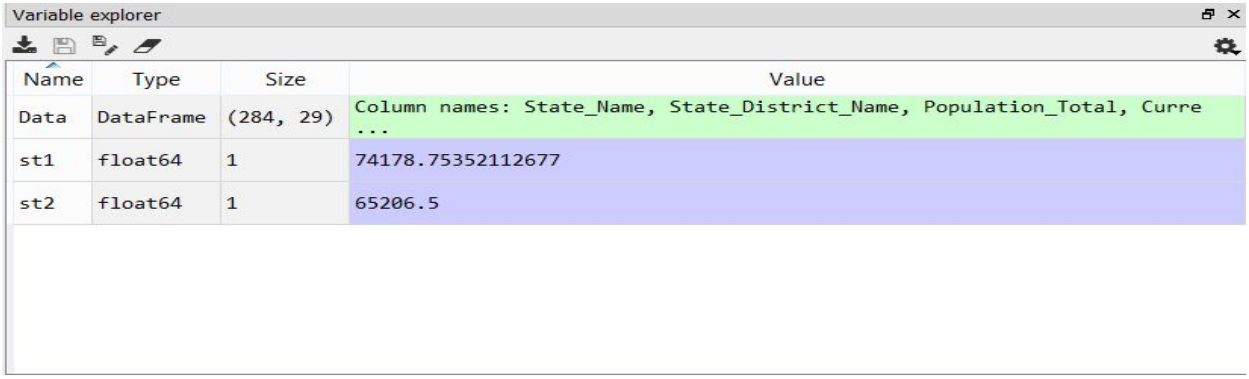
- Incomplete:When some of the attributes are missing. In order to resolve this issue we in our project have replaced these missing values by mean of that attributes. Some other ways could be replacing by median or just simply ignoring these spaces while doing the analysis.
- Noisy:These are the outliers or the errors in the data.
- Inconsistent:containing discrepancies in codes or names.For ex:for a field which is text a numerical data has been entered.

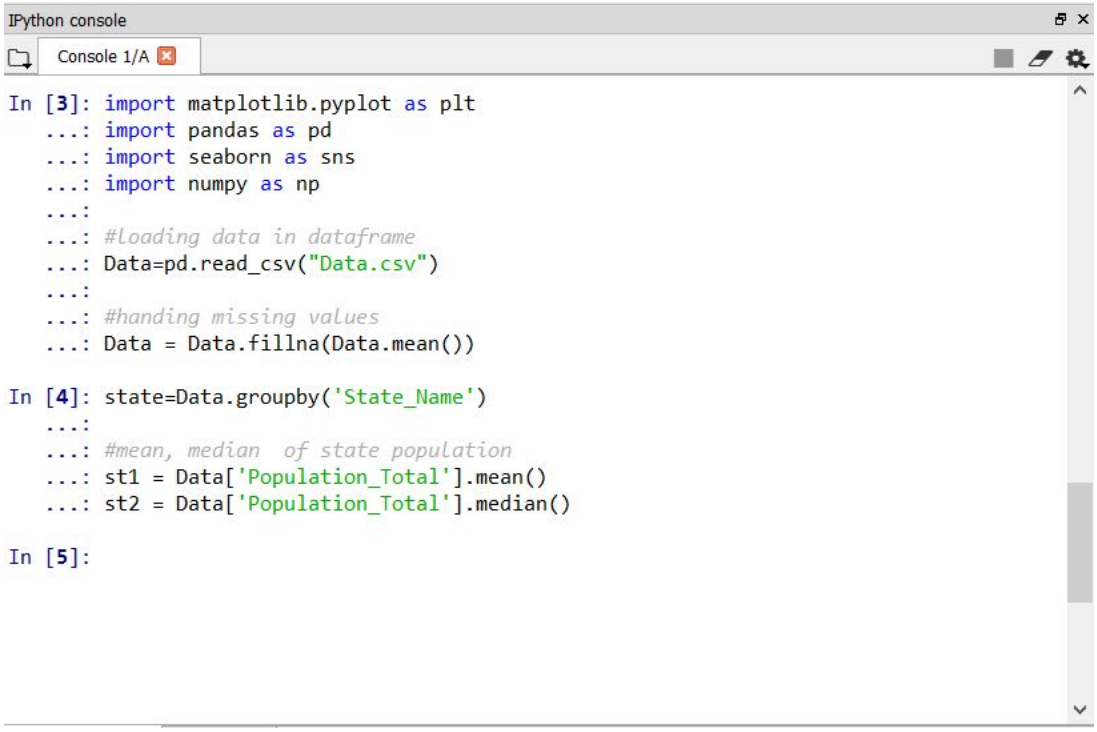
After preprocessing in our data all of the missing values have been replaced by the respective means.

RESULTS AND ANALYSIS

Question 1: What is the mean population of all the 9 states?

Question 2: What is the median of the all the 9 of states?

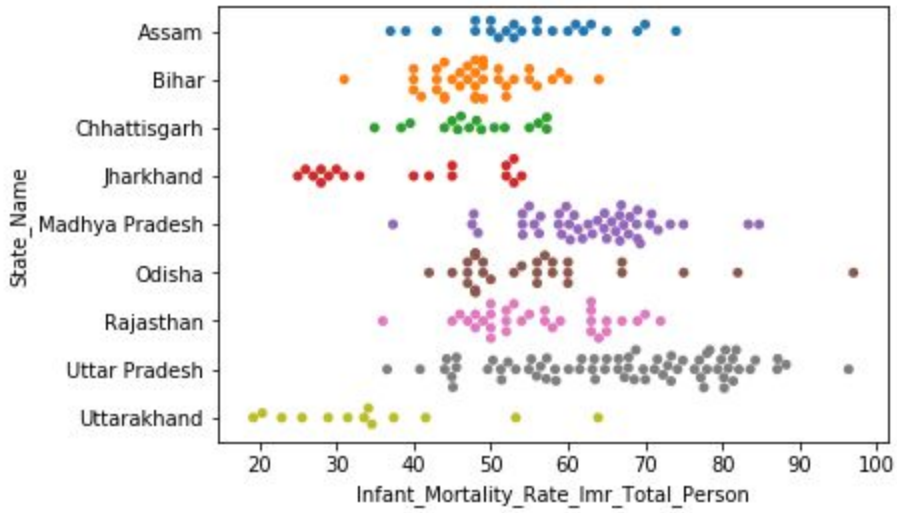




Here we have used mean and median function which directly gives us the value of mean and median of the population.

Question 3: What is the infant mortality rate in each state?

To analyze this we have used swarm plot as it gives a clear distribution of values and we can easily analyze and tell how infant mortality rates vary within various districts in each state.

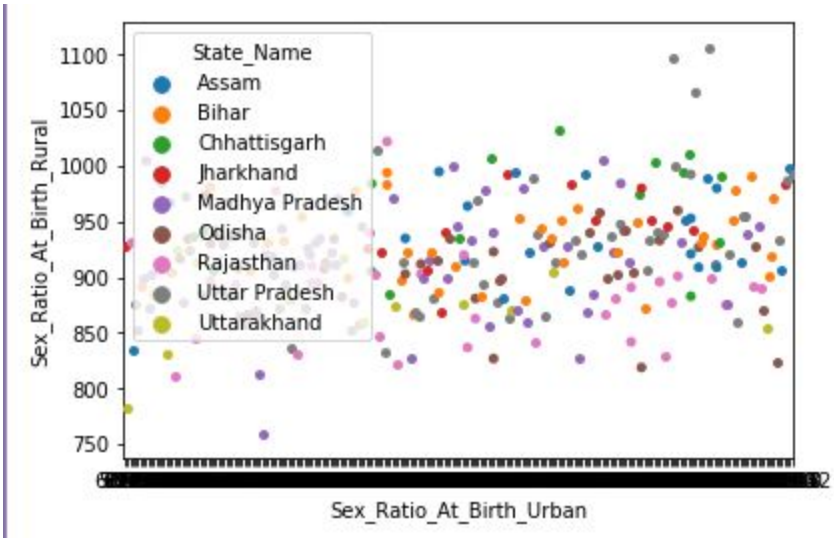


Code:

```
sns.swarmplot(y='State_Name',x='Infant_Mortality_Rate_Imr_Total_Person',data=Data)
```

From the swarm plot we could easily find the outlier points and in which state they lie i.e in Odisha and Uttar Pradesh also how the rates are clustered about certain values in some states like Madhya Pradesh and Bihar while in some the values are uniformly distributed over some range like in Assam and Chattisgarh.

Question 4:What is the sex ratio at birth in urban regions and in rural regions for all states?
We can easily compare the sex ratio in rural and urban regions of a district of same state at same time using swarm plot.

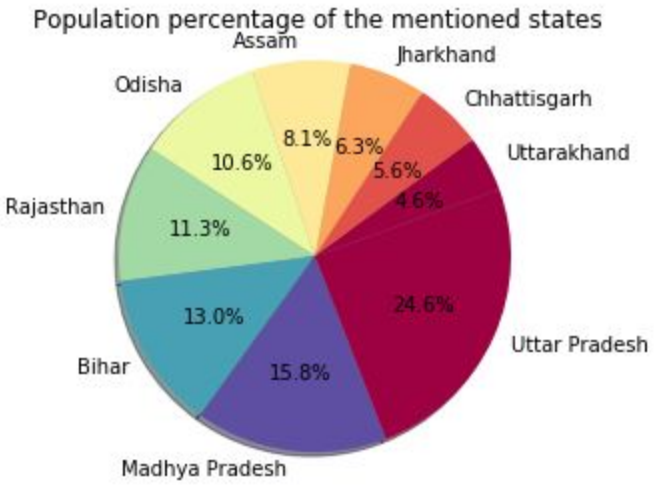


Code:

```
In [7]: sns.swarmplot(x='Sex_Ratio_At_Birth_Urban',y='Sex_Ratio_At_Birth_Rural',data=Data,hue='State_Name')
```

From the plot through the color codes we could easily analyze the data.We see that sex ratio in urban regions is greater as compared to that in rural. Also we could see the whole data is clustered about the range 900-1000. Also we could see see the 3 outliers belonging to Uttar Pradesh.

Question 5: What percentage of total population is present in each state? This can be easily visualized in pie chart.

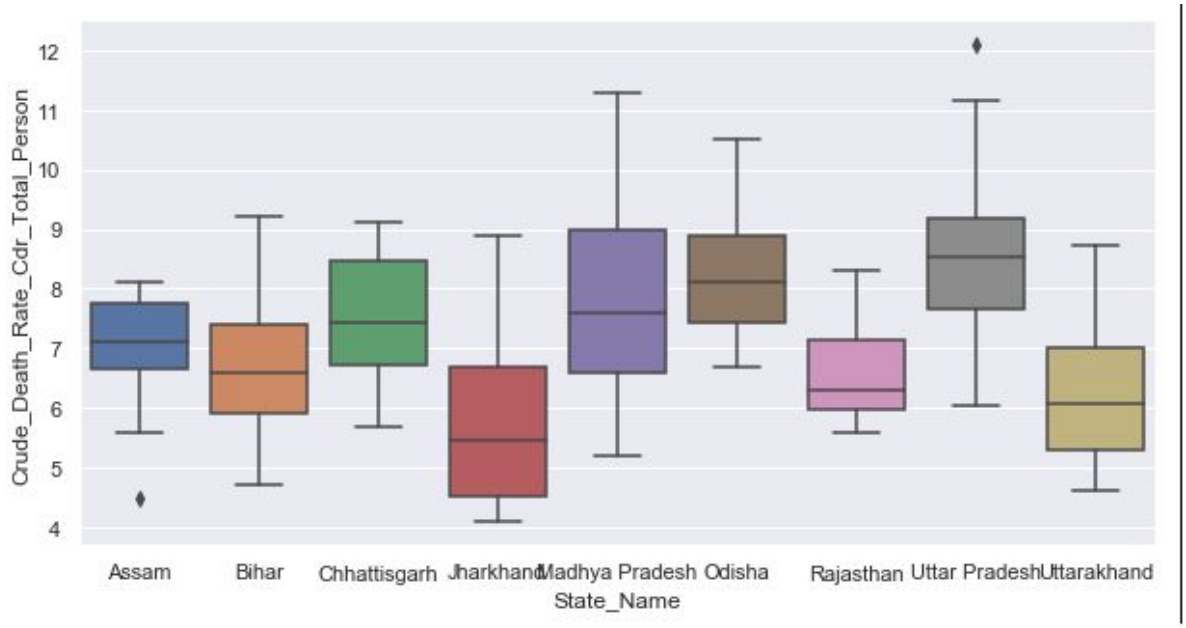


‘

```
In [8]: cmap = plt.get_cmap('Spectral')
   ...: colors = [cmap(i) for i in np.linspace(0, 1, 8)]
   ...: pie_sources = Data.groupby('State_Name').agg('count')
   ...: source_labels = pie_sources.Population_Total.sort_values().index
   ...: source_counts = pie_sources.Population_Total.sort_values()
   ...: plt.pie(source_counts, labels=source_labels, colors=colors,autopct='%1.1f%%',
shadow=True, startangle=20)
   ...: plt.axis('equal')
   ...: fig=plt.title('Population percentage of the mentioned states')
```

We could see that most of the population lies in the state of Uttar Pradesh and minimum lies in Uttarakhand.

Question 6: What is the death rate in each state? What is the interquartile range and median of the death rate?
Median and interquartile range can be best analyzed by using box plots.

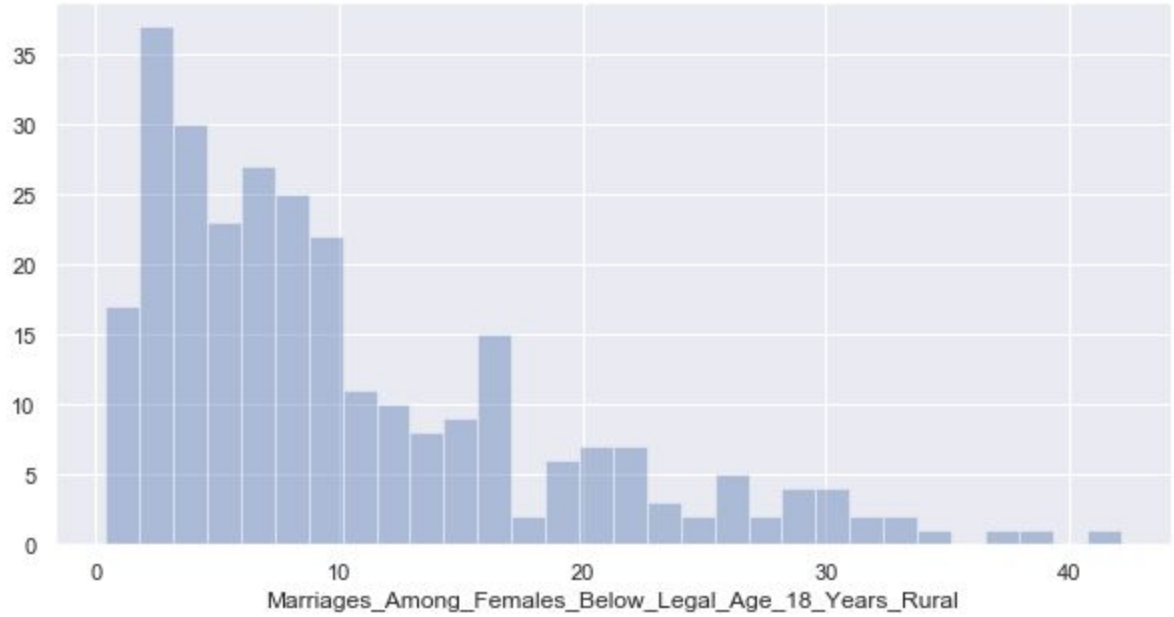



```
In [9]: sns.set(rc={'figure.figsize':(10,5)})
   ...: sns.boxplot(x="State_Name", y="Crude_Death_Rate_Cdr_Total_Person", data=Data)
```

We could see that the highest median is in the state of Uttar Pradesh which implies that they have highest mortality rate.

Question 7: How many females have been married before they reached their legal age of 18 years?

This could be easily analyzed through histogram. It gives number in terms of range for example 0-5,5-10 and so on.

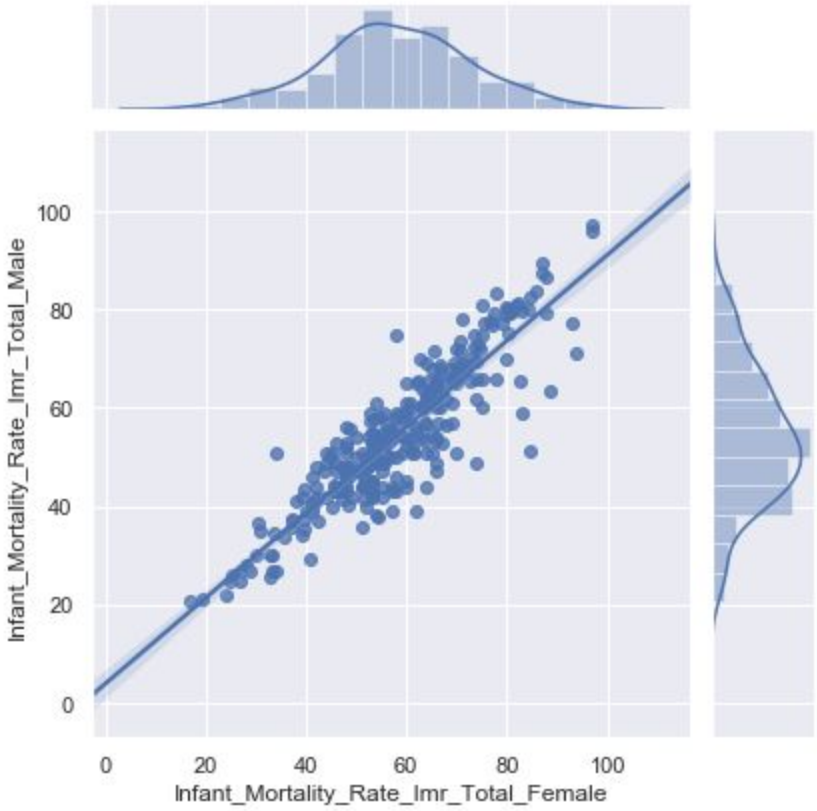



```
sns.distplot(Data['Marriages_Among_Females_Below_Legal_Age_18_Years_Rural'], bins=30,kde=False)
```

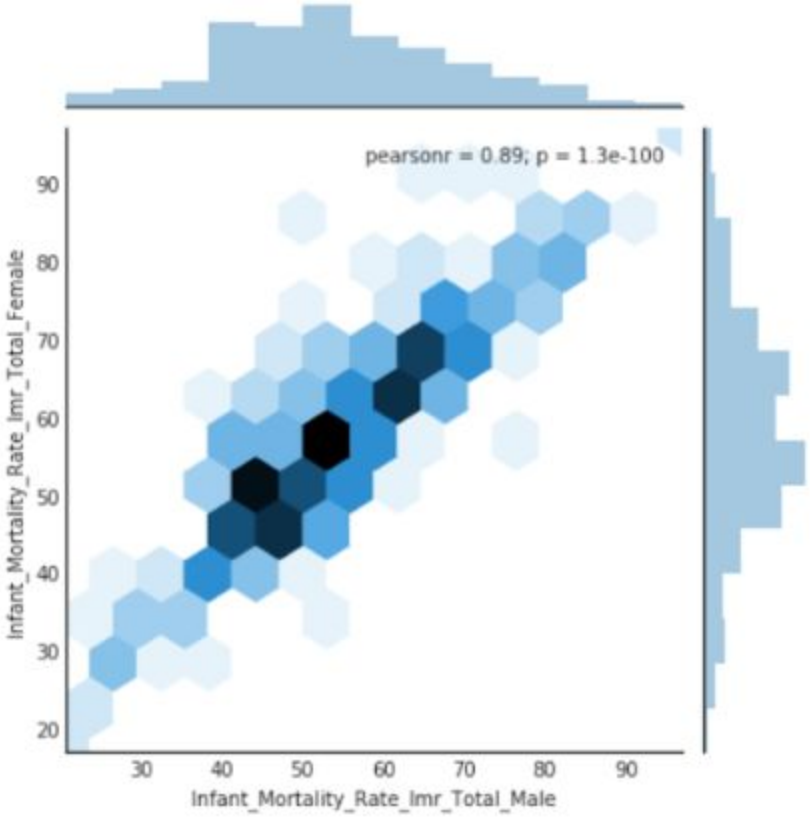
We see that the graph is right skewed.

Question 8: What is the difference in the infant mortality rate of males and females ?

We have used joint plot to analyzed this data.Using this we can analyze both the columns together.




```
sns.jointplot(x='Infant_Mortality_Rate_Imr_Total_Female',y='Infant_Mortality_Rate_Imr_Total_Male',data=Data,
kind='reg')
```

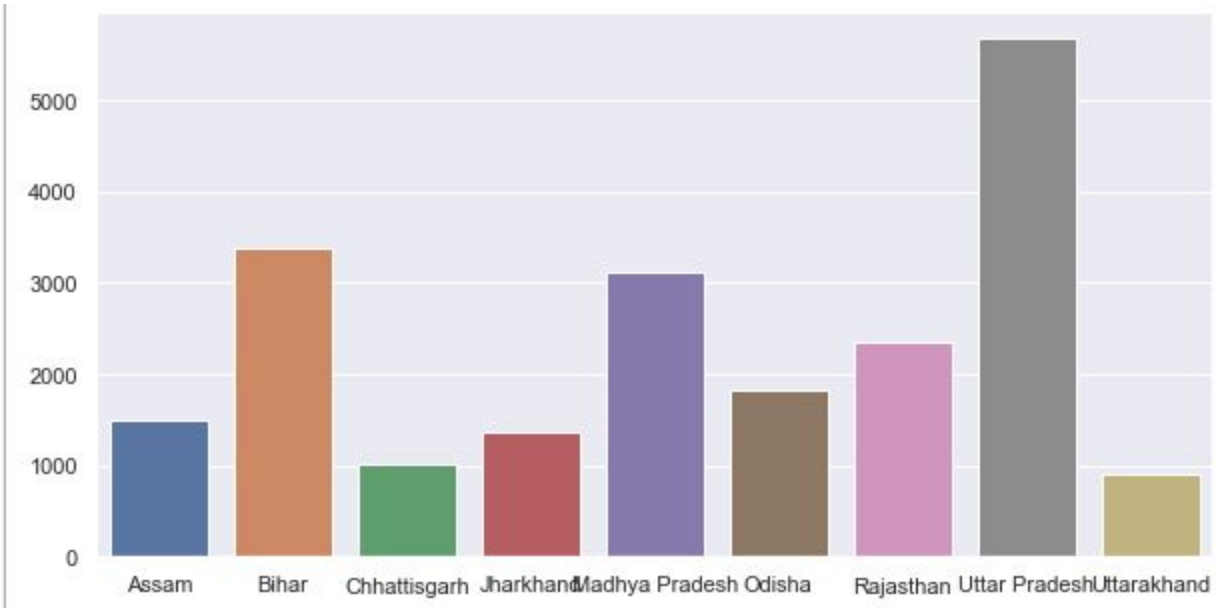



```
with sns.axes_style('white'):
    sns.jointplot("Infant_Mortality_Rate_Imr_Total_Male", "Infant_Mortality_Rate_Imr_Total_Female", Data, kind='hex')
```

Here we could see the distribution curves of infant mortality rates for both male and female is approximately normal.
Their mean lies where the peaks of the normal distribution curves lie.
We observe that in females infant mortality rate is bit higher than males.
In females the graph is slightly right skewed.

Question 9: What is the dependency ratios in urban regions in each state?

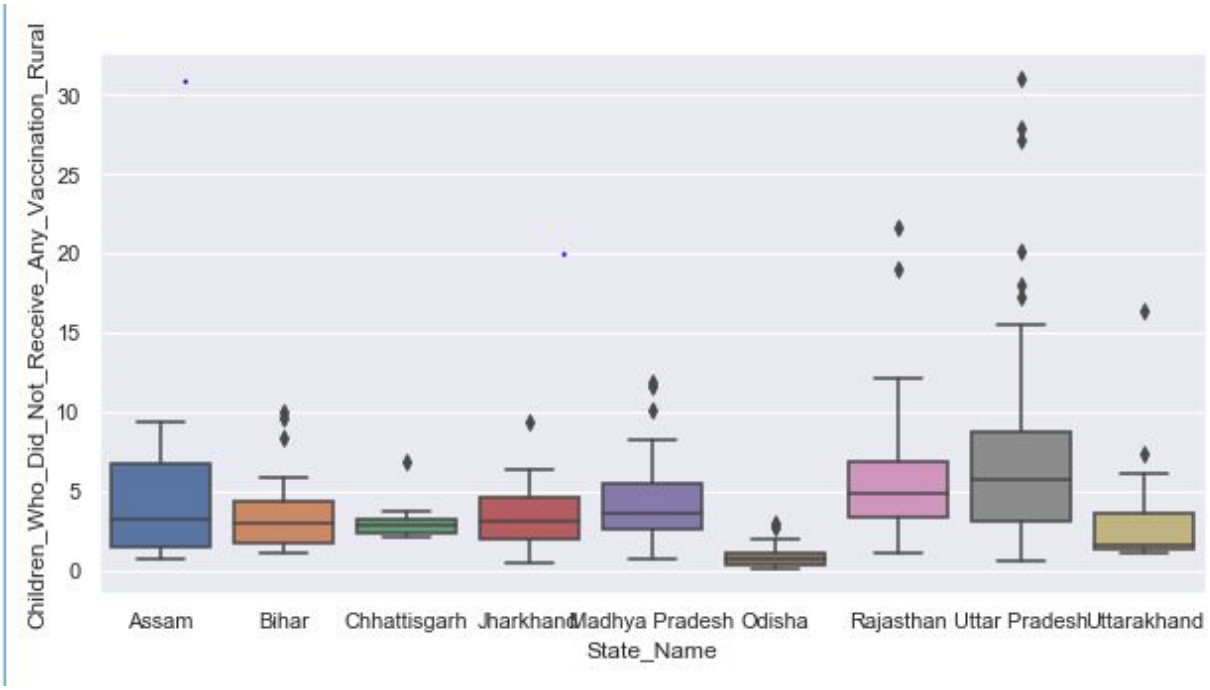
This can be easily shown by bar graph.




```
In [12]: Factor=['Assam','Bihar','Chhattisgarh','Jharkhand','Madhya Pradesh','Odisha','Rajasthan','Uttar
Pradesh','Uttarakhand']
    ...: y=states_no['Dependency_Ratio_Urban']
    ...: sns.barplot(x=Factor, y=y.values,  data=Data)
    ...:
    ...: x=states_no['Dependency_Ratio_Rural']
    ...: sns.barplot(x=Factor, y=x.values,  data=Data)
Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x113705f0>
```

We can easily see here which state has how much dependency ratio.Also we could figure out that Uttar Pradesh has highest dependency ratio and Uttarakhand has the minimum.

Question 10: How many children in rural regions did not receive any vaccinations in rural regions and also what is the median in each state and interquartile range?

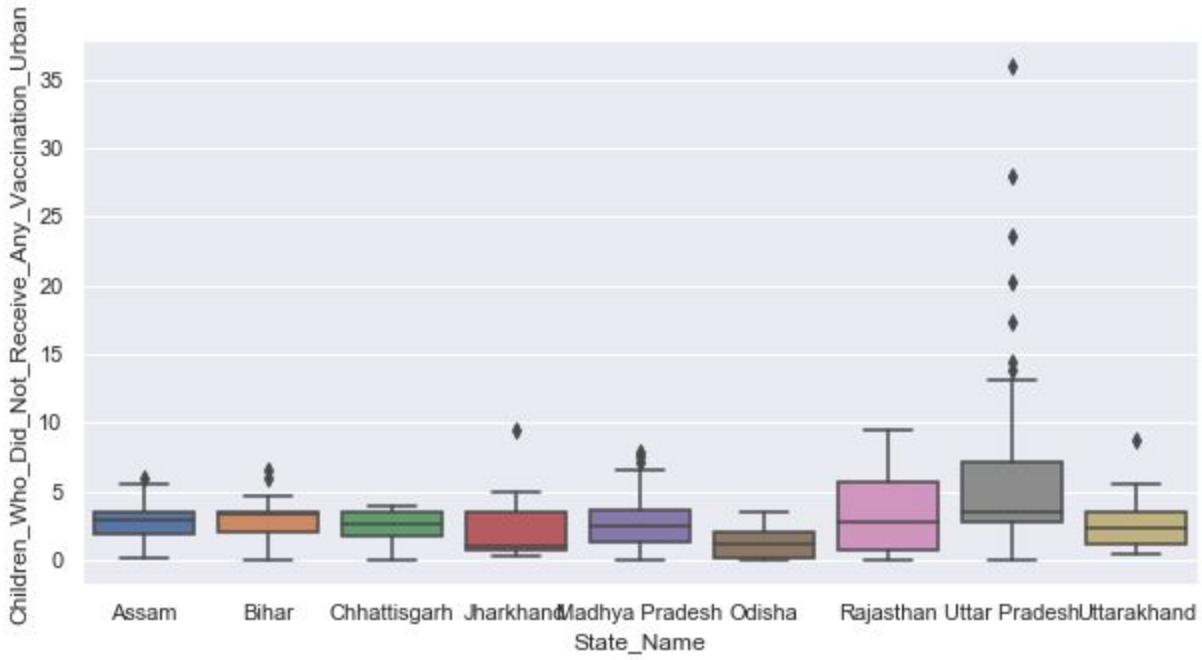



```
In [15]: sns.set(rc={'figure.figsize':(10,5)})
    ...: sns.boxplot(x="State_Name", y="Children_Who_Did_Not_Receive_Any_Vaccination_Rural", data=Data)
```

Here we observe that Chhattisgarh and Odisha have very small interquartile range and the no of children receiving vaccination is minimum in the state of Odisha.Also maximum is from Uttar Pradesh.

Question 11: How many children in rural regions did not receive any vaccinations in rural regions and also what is the median in each state and interquartile range?

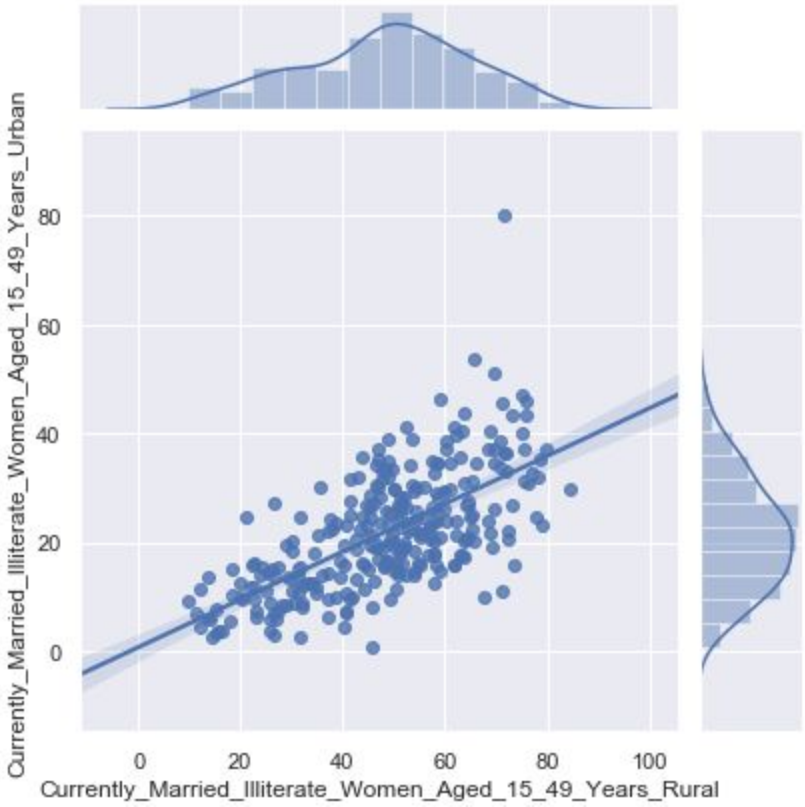
We are going to use box plots here as they clearly depict what is median,interquartile range maximum and minimum values etc.




```
sns.set(rc={'figure.figsize':(10,5)})
sns.boxplot(x="State_Name", y="Children_Who_Did_Not_Receive_Any_Vaccination_Urban", data=Data)
```

Here we observe that Chhattisgarh and Odisha have very small interquartile range also maximum number of children not getting vaccination is from Uttar Pradesh.Also we can see that in Bihar median coincides with upper quartile and in Jharkhand it coincides with lower quartile.

Question 12: How many women are married and illiterate in both urban and rural regions?
Whenever we have to compare 2 sets of data is always preferable to use jointplot.



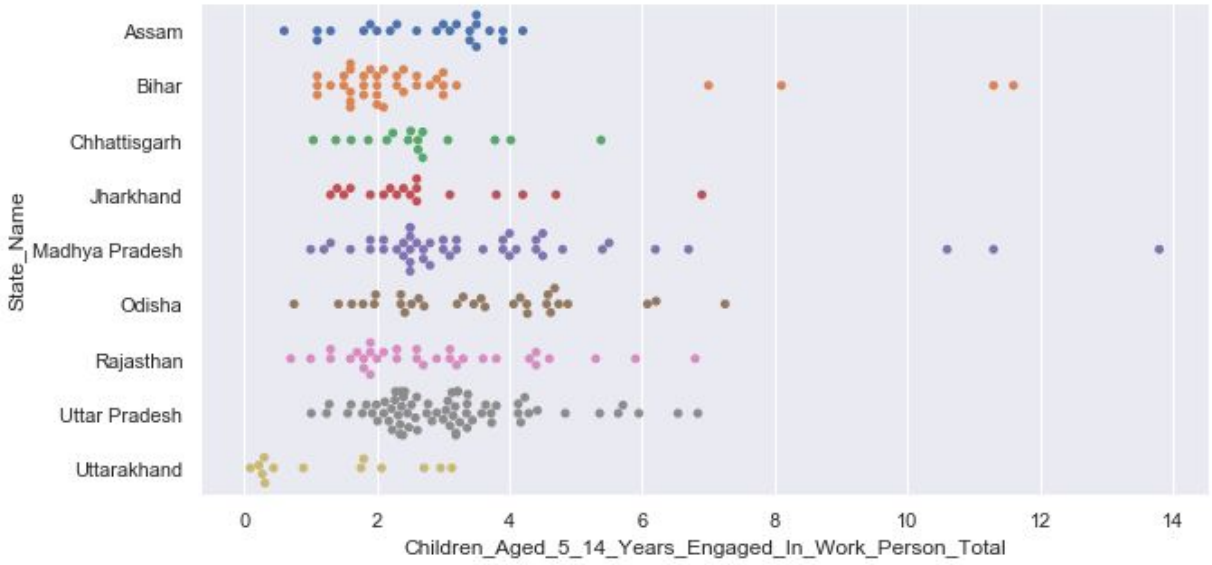
Code:

```
In [13]:
sns.jointplot(x='Currently_Married_Illiterate_Women_Aged_15_49_Years_Rural',y='Currently_Married_Illiterate_Women_Aged_15_49_Years_Urban',data=Data,kind='reg')
```

Here we could easily see that there less illiterate women in urban regions than in rural regions.Also the distribution curves of both the regions is approximately normal.Also we observe an outlier in urban region.

Question 13: How many children in the age group of 5yrs to 14 yrs are engaged in work in each state?

To analyze this we have used swarm plot as it gives a clear distribution of values and we can easily analyze the data.



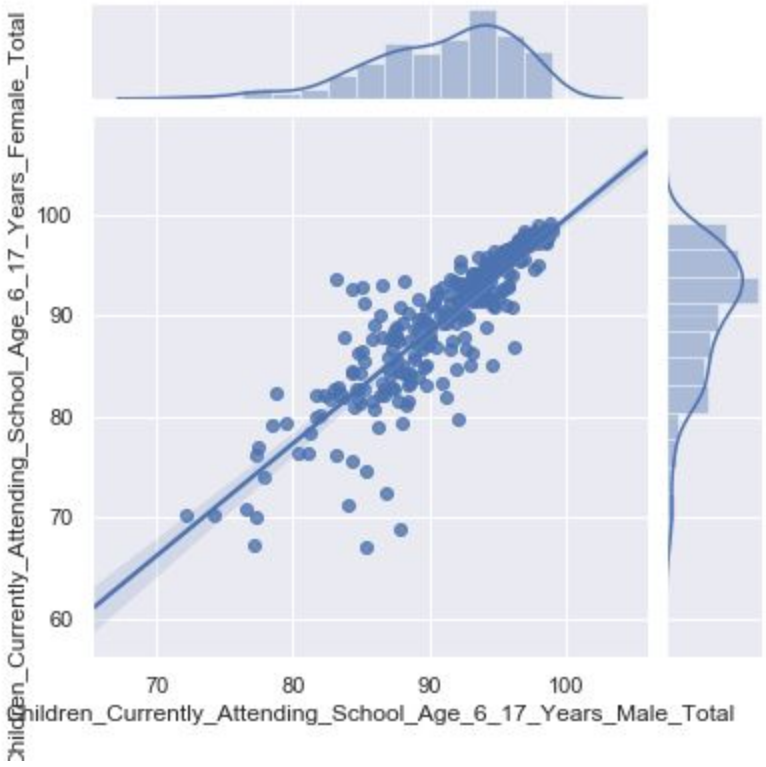
Code:

```
In [19]: sns.swarmplot(y='State_Name',x='Children_Aged_5_14_Years_Engaged_In_Work_Person_Total',data=Data)
```

Here observe various outliers in the states of Madhya Pradesh and Bihar. Maximum % of children work in Madhya Pradesh. Also in most of the states the data is clustered about 2%-4% which means that mean % lies in this range.

Question 14: How many girls and how many boys attend the school in age group of 6-17 years?

This can be easily analyzed by joint plot.




```
sns.jointplot(x='Children_Currently_Attending_School_Age_6_17_Years_Male_Total',y='Children_Currently_Attending
School_Age_6_17_Years_Female_Total',data=Data,kind='reg')
```

We observe that graph for female is right skewed while graph for males is left skewed.We can see that the number of girls attending school decreases drastically as compared to males.

Question 15: What is the difference in mean and standard deviation in sex ratio of urban and rural regions ?

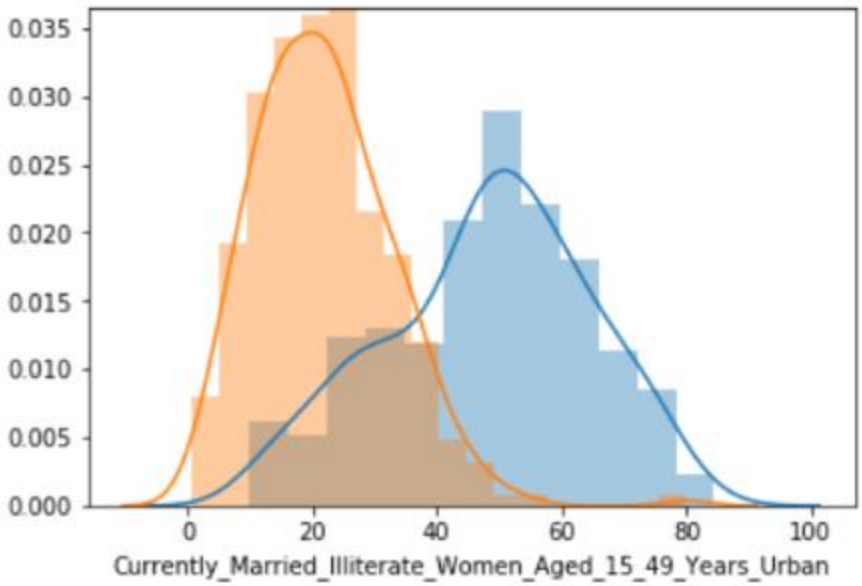
mean1	float64	1	908.7374733096084
mean2	float64	1	919.2452112676057

std1	float64	1	100.27759220587593
std2	float64	1	50.08896424188132

```
In [30]: mean1 = Data['Sex_Ratio_At_Birth_Urban'].mean()
    ...: mean2 = Data['Sex_Ratio_At_Birth_Rural'].mean()

In [31]: std1 = Data['Sex_Ratio_At_Birth_Urban'].std()
    ...: std2 = Data['Sex_Ratio_At_Birth_Rural'].std()
```

Question 16 How the number of illiterate married women differ in rural and urban region?
We have used distplot for this.




```
sns.distplot(Data['Currently_Married_Illiterate_Women_Aged_15_49_Years_Rural'])
sns.distplot(Data['Currently_Married_Illiterate_Women_Aged_15_49_Years_Urban']);
```

From the graph it is visible that there are more number of illiterate women in rural regions as compared to urban region. Orange region represents

represents rural region while blue corresponds to urban.The peak of the orange is higher as compared to the blue. The rural curve is normal while the urban curve is slightly left skewed.