DIAGNOSTIC ANALYTICS: BIRTH RATE IN INDONESIA BY MATERNAL AGE GROUP

Ayat Tulloh Rahulloh Khomeini

Information System, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia

e-mail: dkhomeini79@gmail.com

ABSTRACT

Analyzing birth rates in Indonesia by maternal age group revealed fascinating insights. Using Python, we investigated both total birth rates and age-specific rates. Our journey, starting with data cleaning and statistics, unfolded through captivating visualizations like line charts and correlation matrices. The 25-29 age group emerged as the most productive, averaging a 139% birth rate and showcasing a concentrated distribution peaking between 100%-150%. Remarkably, this peak nearly touched 250%, highlighting the significant contribution of this age group. However, outliers with extreme values added another layer of complexity. Meanwhile, the 35-39 age group stood out as the most influential on total birth rates. Its correlation of 0.737764 indicated a positive, albeit imperfect, relationship – when this group's birth rate rose, the overall rate followed suit. Additionally, a nearly normal distribution with less extreme outliers suggested some stability within this influential group. In conclusion, both the 25-29 and 35-39 age groups hold key roles in understanding and influencing birth rates in Indonesia. Future research delving deeper into factors impacting these key groups holds immense potential for further illuminating this complex landscape.

Keywords: Analyzing, Birth-Rate, Python

CHAPTER 1

INTRODUCTION

1. Background

Birth represents the actual productive ability of the population. The total fertility rate is the average number of children born to a woman during her childbearing years, aged 15-49 years. This total birth rate is an important and strategic indicator to determine the extent of success of a country or an entire country in controlling its population through family planning (KB) programs. Quoted from the results of a survey by the Central Statistics Agency in 2020, the total fertility rate was 2.10. This means that the average woman in Indonesia will give birth to two children in her reproductive years, and with the onset of Covid-19 also having an impact on the total birth rate, although it is said that there has been a decline, some regions still have high birth rates, including East Nusa Tenggara, Papua, West Papua, and Maluku. However, it has never been specifically mentioned at what age has the most influence and what age has the stability of distribution. This is worth paying attention to, because at the most fertile age can be the biggest modifying factor.

2. Research

This study used a comprehensive diagnostic analytical method to improve the accuracy of results in the level of influence of certain groups of maternal age in Indonesia. This study aims to determine the influence of maternal age group on the birth rate in 2020. It is hoped that the results of this study will be taken into consideration in deciding opinions to solve the problem of birth increase in certain regions

CHAPTER 2

RESEARCH PREPARATION

1. Dataset Preparation

Researchers used datasets from the survey results of the Central Statistics Agency (BPS) consisting of 2 datasets relevant to Indonesia's Birth Rate. The first dataset is "Total Fertility Rate (TFR) of SP2020 Long Form Results (LF) by Province/District/City, 2020", which provides information on the total birth rate of each region for a decade. And the second is "Birth Rate of Long Form SP2020 Results According to Maternal Age Group (Age Specific Fertility Rate / ASFR) and Province / Regency / City, 2020", containing information on the number of live births from women with certain age groups per 1000 women in the same age group. Women covered are women in the fertile period (15-49 years) which are grouped by age group 5 years. Both datasets are merged with Province/District/City as the index.

2. Tool Preparation

To support a better level of accuracy because you have to perform diagnostic analysis, then in this case, it is best to use Python with the DataSpell / Jupyter IDE and here Excel will also be used to rename the problematic dataset columns. The first step is to prepare the required libraries, such as pandas, numpy, matplotlib, seaborn and statsmodel.

CHAPTER 3

RESEARCH METHODS

A. Input

1. Data retrieval

Here the researcher imports the dataset with the pandas library, where the first dataset will be identified as a 'form', and the second dataset as a 'total'. Then run the merge function to merge both datasets.

```
form=pd.read_csv('datasets/Angka Kelahiran Hasil Long Form SP2020 Menurut Kelompok Umur Ibu (Age Spesific Fertility Rate_ASFR) da.csv')
total=pd.read_csv('datasets/Angka Kelahiran Total _ Total Fertility Rate (TFR) Hasil Long Form (LF) SP2020 Menurut Provinsi_Kabupa.csv')

df=pd.merge(form, total, on='Provinsi/Kabupaten/Kota')
```

This function makes 2 datasets into one data frame with "Province/District/City" as the merge key which will later be used as an index.

2. Data cleansing

Data cleansing is performed if that data has anomalies, null values, NaN. Datasets must avoid these problems for diagnostic analysis. The first step is to use the .isna().sum() function, which is useful for knowing if our dataset has a NaN value

```
df_indexed.isna().sum()
```

In this case, researchers found no NaN. So coupled with the column drop process, the goal is to delete columns that have nothing to do with research

```
drop_column=['idx','id','Unnamed: 3']
df=df.drop(columns=drop_column)

df_indexed=df.set_index('Provinsi/Kabupaten/Kota')
df_indexed
```

Here the researcher drops 3 columns and uses the column 'Province/Regency/City' as the index.

B. Process

1. Statistics

To find out the summary of descriptive statistics from a DataFrame, the .describe() function is performed to generate a summary of descriptive statistics from the DataFrame. The statistics involve the mean, standard deviation, minimum value, quartile (25%, 50%, 75%), and maximum value of each numeric column in the DataFrame.

```
df_indexed.describe().round(3)
```

.round(3) is used to round 3 numbers after the comma, this helps make the output easier to read and can minimize the number of numbers displayed. Furthermore, the mean is identified into variables to later be used as a visualization to find the average age of the largest number of live births.

```
df_mean=df_indexed.mean()
df_mean
```

2. Line Chart

In an effort to find out the trend, matplotlib is used to visualize the line chart, where the canvas size is 17x9, and the dataset used is 'df_mean' which means it will display the average trend of birth rate per age group. By details, x-axis = age group, y= average value.

```
plt.figure(figsize=(17,9))
plt.plot(df_mean)
plt.title("Nilai rata-rata Angka kelahiran")
```

3. Correlation

To find out the relationship between variables, proven by correlation. The first step is to calculate the correlation value between columns to produce a matrix that will later be visualized into a heatmap. Next touch on the visualization stage with seaborn heatmap, with matrix data.

```
corr_mat=df_indexed.corr()
plt.figure(figsize=(8,                                  6))
sns.heatmap(corr_mat,    cmap="crest",    annot=True)
plt.title('Correlation                              Plot')
plt.show()

corr_mat
```

4. Pairplot

In an effort to describe the relationship as well as distribution between variables, the first step is to create a pairplot using the Seaborn pairplot module. The process starts directly with the dataset (all columns) to distribute, the goal is to know which distribution is good.

```
sns.pairplot(df_indexed)
plt.show()
```

This pairplot will automatically describe a histogram where we can see the results of the distribution.

5. Boxplot

The use of boxplot is to look for outliers in the data, the process starts with the seaborn boxplot

```
sns.boxplot(df['25-29'])
plt.title("25-29")
plt.show()
```

```
sns.boxplot(df['35-39'])
plt.title("35-39")
plt.show()
```

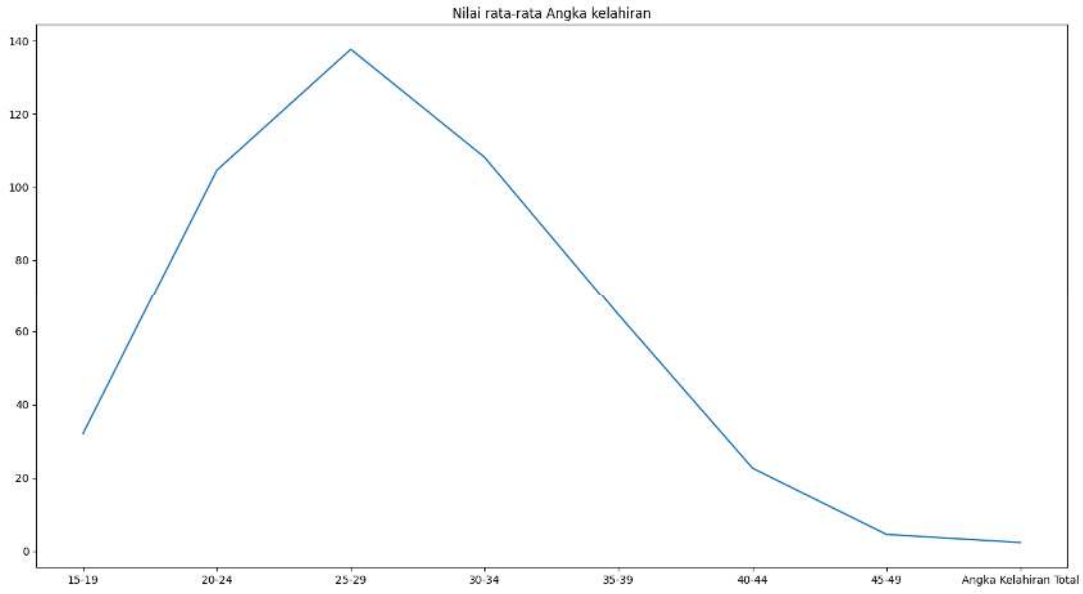
Here the researcher uses 2 variables of choice, for the reasons will be explained in the next chapter.

C. Output

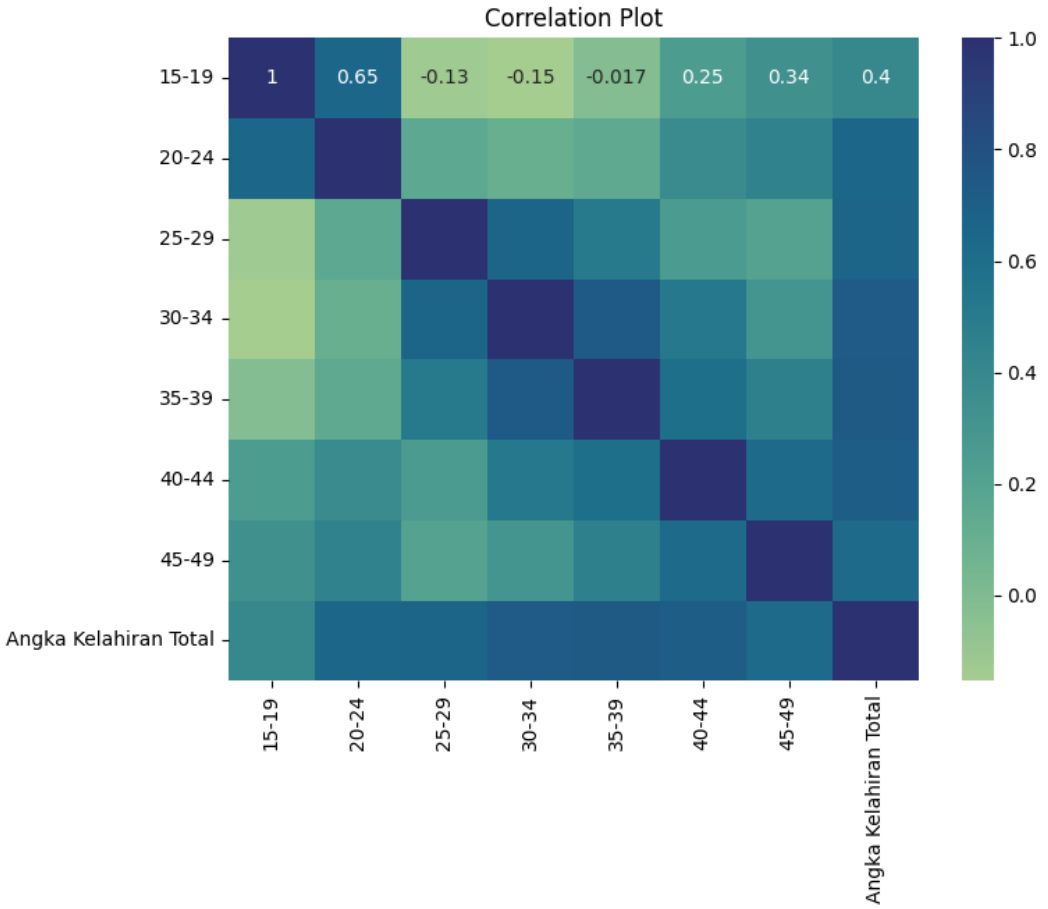
1. Tabel Statistik

	15-19	20-24	25-29	30-34	35-39	40-44	45-49	Total Birth Rate
Count	551.000	551.000	551.000	551.000	551.000	551.000	551.000	551.000
Mean	32.164	104.637	137.764	108.141	64.482	22.635	4.579	2.372
Std	17.215	22.706	22.202	20.858	16.132	11.189	5.439	0.375
Min	4.000	44.500	93.100	53.200	25.800	3.900	0.100	1.540
25%	18.500	92.200	123.250	95.050	54.150	15.850	1.800	2.150
50%	30.100	103.900	134.900	104.000	61.200	19.800	2.900	2.300
75%	43.750	116.500	147.500	117.700	74.050	25.800	5.000	2.490
Max	105.600	237.600	311.500	210.700	132.600	114.700	52.200	4.220

2. Line Chart

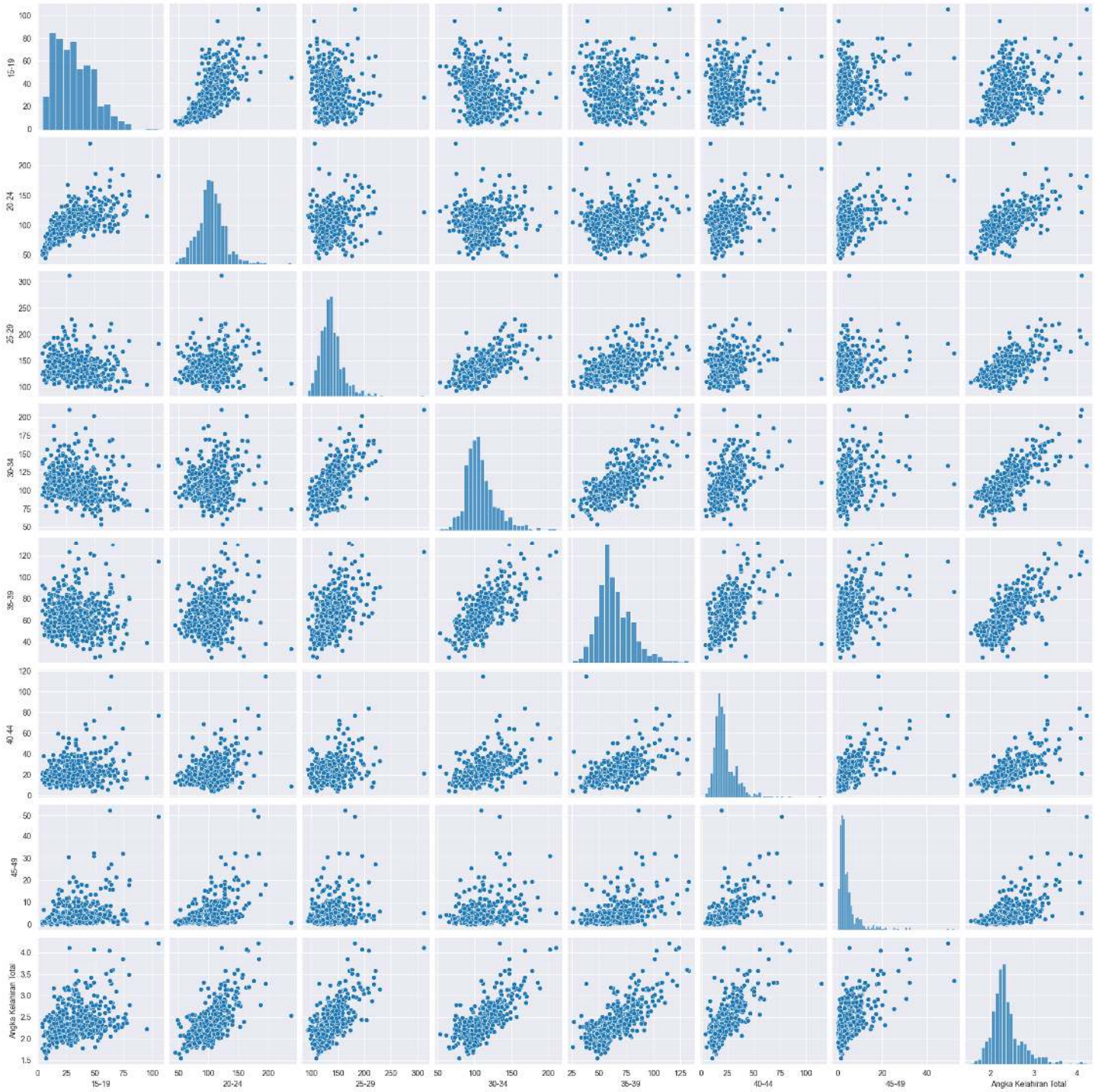


3. Correlation

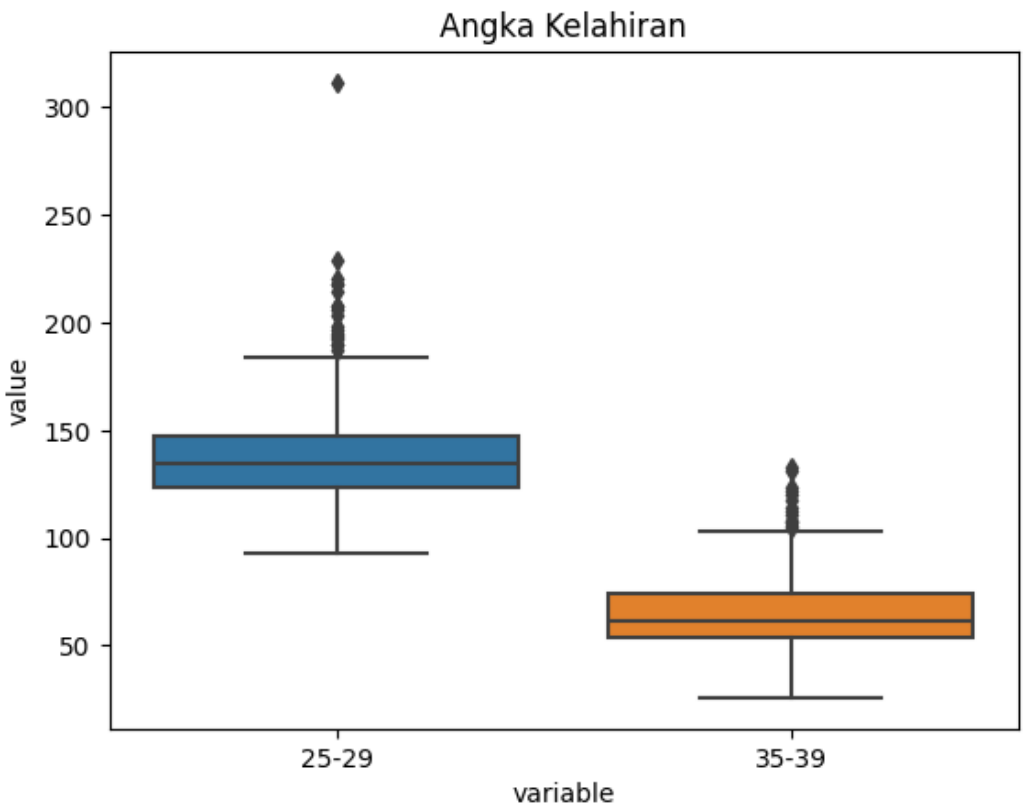


	15-19	20-24	25-29	30-34	35-39	40-44	45-49	Angka Kelahiran Total
15-19	1.000000	0.652937	-0.129266	-0.153082	-0.017089	0.249939	0.341239	0.404484
20-24	0.652937	1.000000	0.154378	0.102471	0.147155	0.377570	0.447702	0.646466
25-29	-0.129266	0.154378	1.000000	0.670381	0.517305	0.256708	0.201242	0.664622
30-34	-0.153082	0.102471	0.670381	1.000000	0.744160	0.520234	0.308669	0.733357
35-39	-0.017089	0.147155	0.517305	0.744160	1.000000	0.590752	0.459593	0.737764
40-44	0.249939	0.377570	0.256708	0.520234	0.590752	1.000000	0.621877	0.713653
45-49	0.341239	0.447702	0.201242	0.308669	0.459593	0.621877	1.000000	0.623297
Angka Kelahiran Total	0.404484	0.646466	0.664622	0.733357	0.737764	0.713653	0.623297	1.000000

4. Pairplot & Histogram



5. Boxplot


Angka Kelahiran
300
250
200
150
100
50
value
25-29
35-39
variable

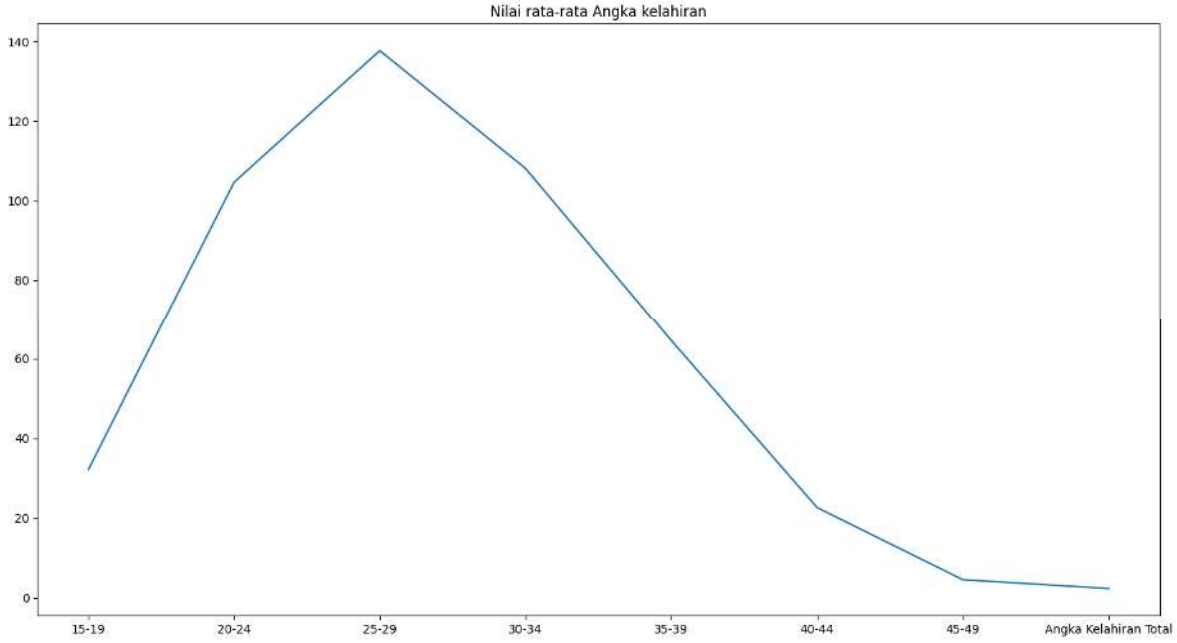
CHAPTER 4

RESULTS & ANALYSIS

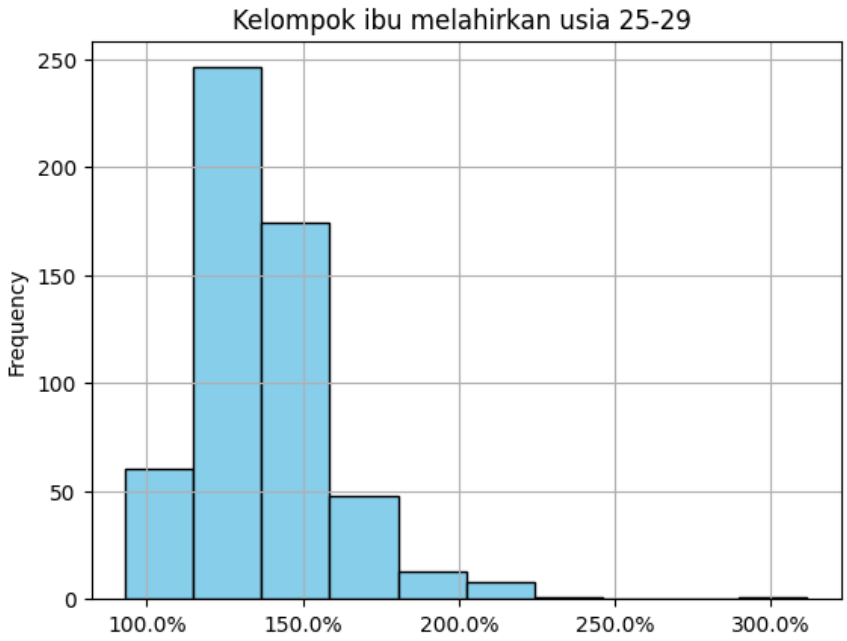
A. The most productive age groups

	15-19	20-24	25-29	30-34	35-39	40-44	45-49	Total Birth Rate
Count	551.000	551.000	551.000	551.000	551.000	551.000	551.000	551.000
Mean	32.164	104.637	137.764	108.141	64.482	22.635	4.579	2.372
Std	17.215	22.706	22.202	20.858	16.132	11.189	5.439	0.375
Min	4.000	44.500	93.100	53.200	25.800	3.900	0.100	1.540
25%	18.500	92.200	123.250	95.050	54.150	15.850	1.800	2.150
50%	30.100	103.900	134.900	104.000	61.200	19.800	2.900	2.300
75%	43.750	116.500	147.500	117.700	74.050	25.800	5.000	2.490
Max	105.600	237.600	311.500	210.700	132.600	114.700	52.200	4.220

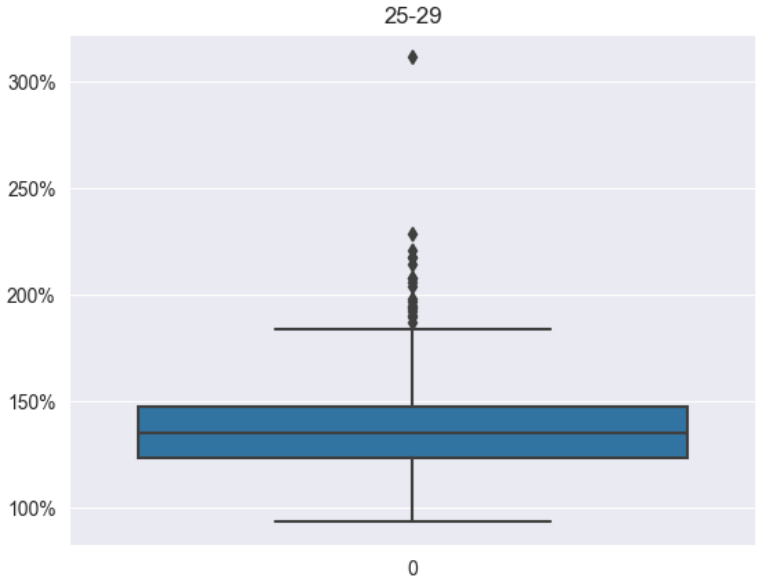
From the results described above, the largest average was achieved by the age group of 25-29 years with a value of 137,764 with a standard deviation value of 22,202. If you look at the Min and Max values, it has a difference of 218,400. This value is very large and needs to be monitored in the future. For distribution, it can be seen in quartiles with the most relevant value being 50% with a value of 134,900.



The evidence that at the age of 25-29 years for 1 decade is the highest. Judging from the graph above, that the **average** age has almost touched 140% in 1 decade. The decline is noticeable at the age of 30-49. Although in 2020 Indonesia was hit by the Covid-19 pandemic, and it experienced a significant decrease compared to the previous period.



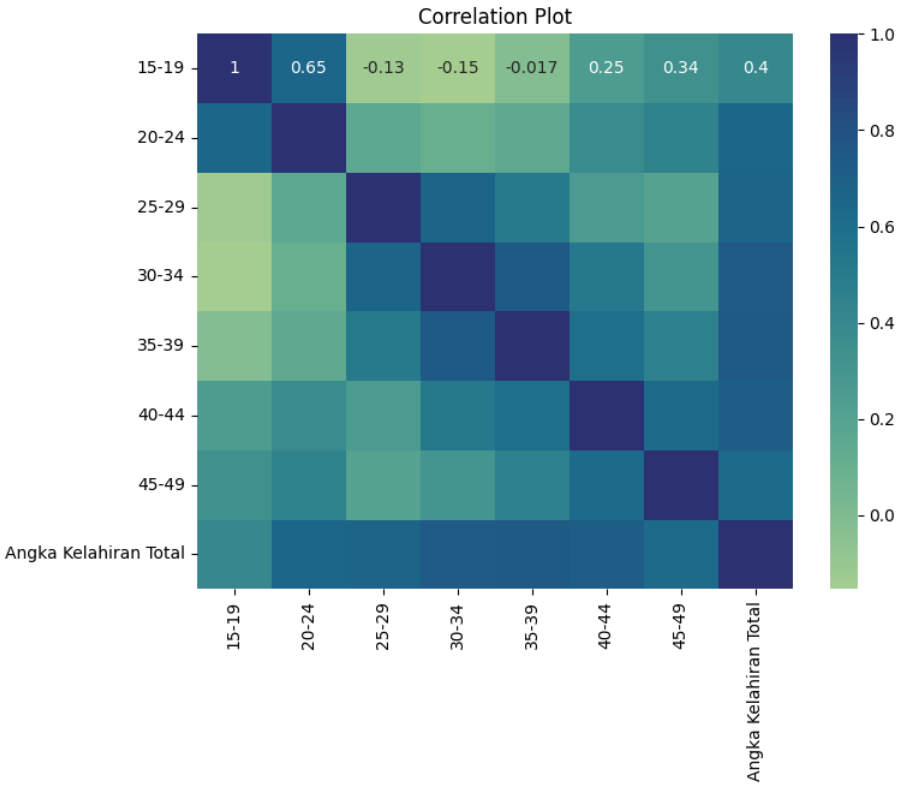
Here is a look at the **overall** distribution at the age of 25-29, it can be seen that the highest distribution is between 100%-150% which almost touches 250. This indicates a high concentration of birth rates on these values. The possibility for a decrease in the birth rate is only when the value of low concentration is strengthened.



The boxplot above shows the data outliers, indicating that there is extreme value in data distribution. The maximum possible value is <200% and the minimum value is 93.10% indicating a range of birth rate variations.
The presence of outliers in the data causes the mean and standard deviation to increase the value significantly.

B. Age groups that have an influence in the increase in the total birth rate

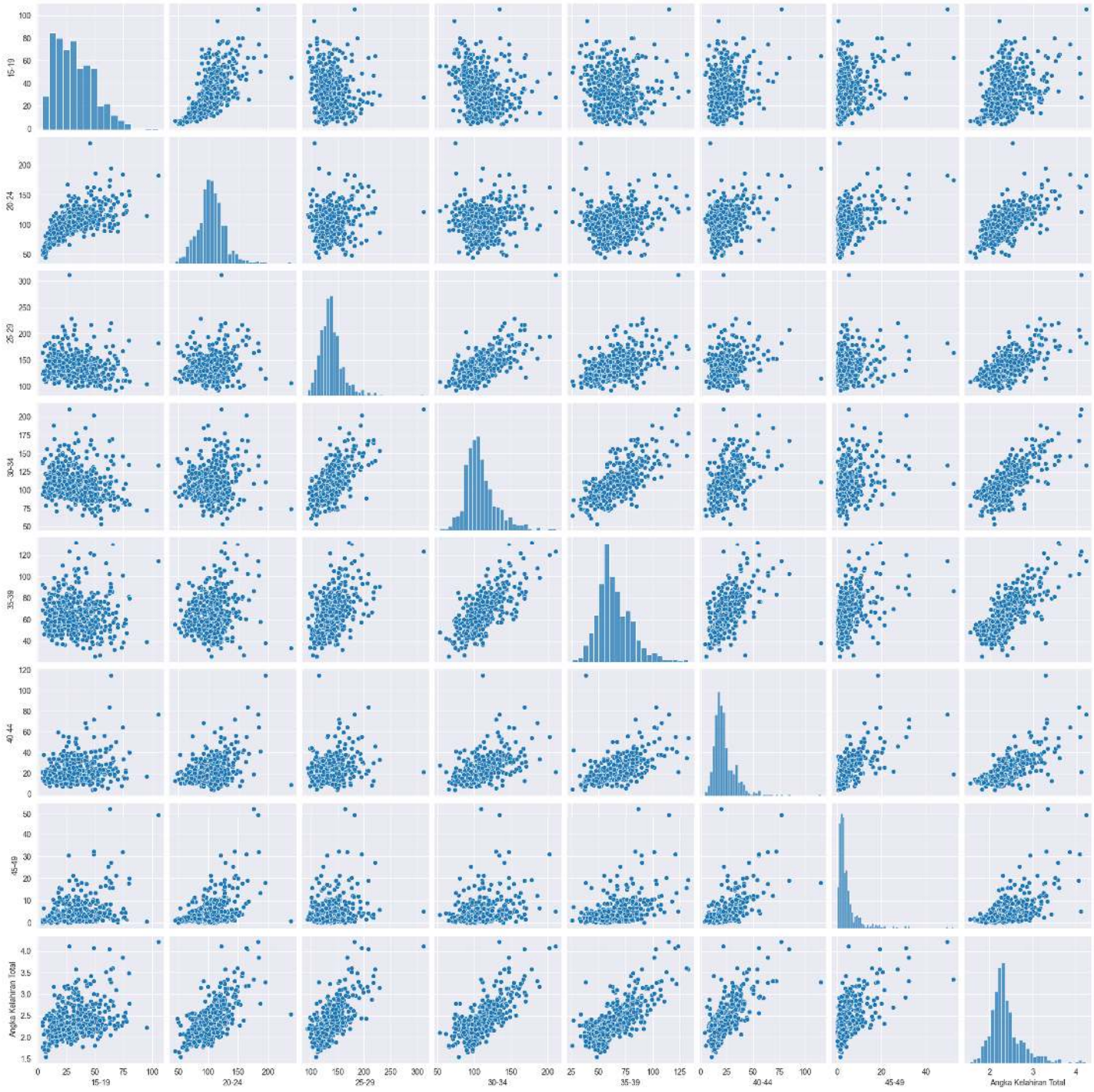
To find out the age group that has an influencing role is to review the correlation value. Here is the correlation heatmap:



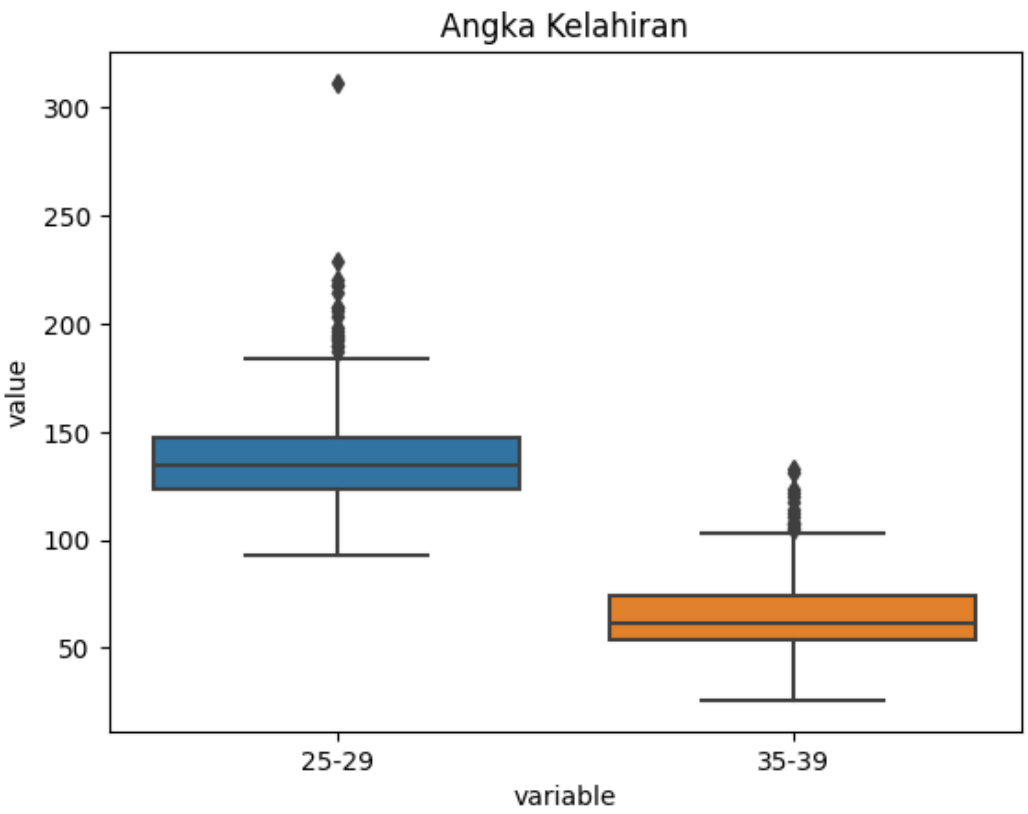
Diagnostic analysis of correlation plots related to age groups that have an influence on Total Birth Rate shows a positive pattern in the age group of 35-39 years. This is also evident in the table below:

	15-19	20-24	25-29	30-34	35-39	40-44	45-49	Total Birth Rate
15-19	1.000000	0.652937	-0.129266	-0.153082	-0.017089	0.249939	0.341239	0.404484
20-24	0.652937	1.000000	0.154378	0.102471	0.147155	0.377570	0.447702	0.646466
25-29	-0.129266	0.154378	1.000000	0.670381	0.517305	0.256708	0.201242	0.664622
30-34	-0.153082	0.102471	0.670381	1.000000	0.744160	0.520234	0.308669	0.733357
35-39	-0.017089	0.147155	0.517305	0.744160	1.000000	0.590752	0.459593	0.737764
40-44	0.249939	0.377570	0.256708	0.520234	0.590752	1.000000	0.621877	0.713653
45-49	0.341239	0.447702	0.201242	0.308669	0.459593	0.621877	1.000000	0.623297
Angka Kelahiran Total	0.404484	0.646466	0.664622	0.733357	0.737764	0.713653	0.623297	1.000000

It can be seen in the table that the value of the age group 35-39 is the highest than in other age groups with a correlation value of 0.737764. This number almost touches 1, which indicates a positive but imperfect relationship between the two variables. That is, when one variable rises, the other variable also rises proportionally, and vice versa.



It can also be seen in the histogram that groups 35-39 have a near-perfect distribution graph.



When compared to the age group 25-29, the age group 35-39 has less extreme values. Evidence that the age group 25-29 has a maximum value probability of <200% and the minimum value is 93.10% and has an extreme value of >300%, and, in the age group of 35-39 with a maximum value of around 130% and extreme outliers almost touching 150%. This indicates that the age group of 35-39 plays a major role in decreasing or increasing the Total Birth Rate in that period.

CHAPTER 5

CONCLUSION

Diagnostic analysis has been carried out on the Total Birth Rate dataset and the Birth Rate dataset Based on Maternal Age Group in Indonesia with Python as a programming language. With the process beginning with data import, data cleaning, data statistics, line charts, correlation matrix, pairplot, boxplot, so that all desired goals are met, The most productive age group in the period was completed by taking an average, which is 139%, and in the barcode view, **the** overall distribution at the age of 25-29, it can be seen that the highest distribution is between 100%-150% which almost touches 250. This indicates a high concentration of birth rates on these values. This also happens because there are outliers with extreme values that touch >300%. It can be concluded that the age group 25-29 is the age group that contributes the most Total Birth Rate in that period.

The age group 35-39 is the age group that has a major influence in decreasing or increasing the Total Birth Rate in that period. This is evidenced by the highest correlation value compared to other age groups, with a value of 0.737764, which shows a positive but imperfect relationship between the two variables. That is, when one variable rises, the other variable also rises proportionally, and vice versa. As well as having an almost normal distribution, the outlier is not too extreme, about almost touching 150% with a maximum value of 132.60%.