**Mini Project (CS399)**

**Time Series Analysis of Human Birth Data**

Team Members:

Peluru Goutham 19bcs085

Kola Lokesh 19bcs056

Guttikonda Mohith Krishna 19bcs043

B S Sathwik Goud 19bcs015

Under the guidance of:

**Dr. Kavi Mahesh**

Ph.D. (Georgia Institute of Technology, Atlanta, USA)



INDIAN INSTITUTE OF INFORMATION TECHNOLOGY

**Abstract :**

Time-series analysis is a statistical method of analysing data from repeated observations on a single unit or individual at regular intervals over a large number of observations. A time series analysis can help us to understand the underlying naturalistic process, the pattern of change over time, or evaluate the effects of either a planned or unplanned intervention. Time series is performed in many domains like in prediction of stocks, Temperature forecasting, Air quality prediction and many other things. But our goal (or) objective is to analyze if there are any insights or patterns in humans birth data.

**Introduction:**

A time series is a chronological sequence of observations on a particular variable. Usually, the observations are taken at regular intervals (days, months, years), but the sampling could be irregular.

A time series analysis consists of three steps:

(1) building a model that represents a time series

(2) validating the model proposed

(3) using the model to predict (forecast) future values and/or impute missing values.

The goal of building a time series model is the same as the goal for other types of predictive models which is to create a model such that the error between the predicted value of the target variable and the actual value is as small as possible.

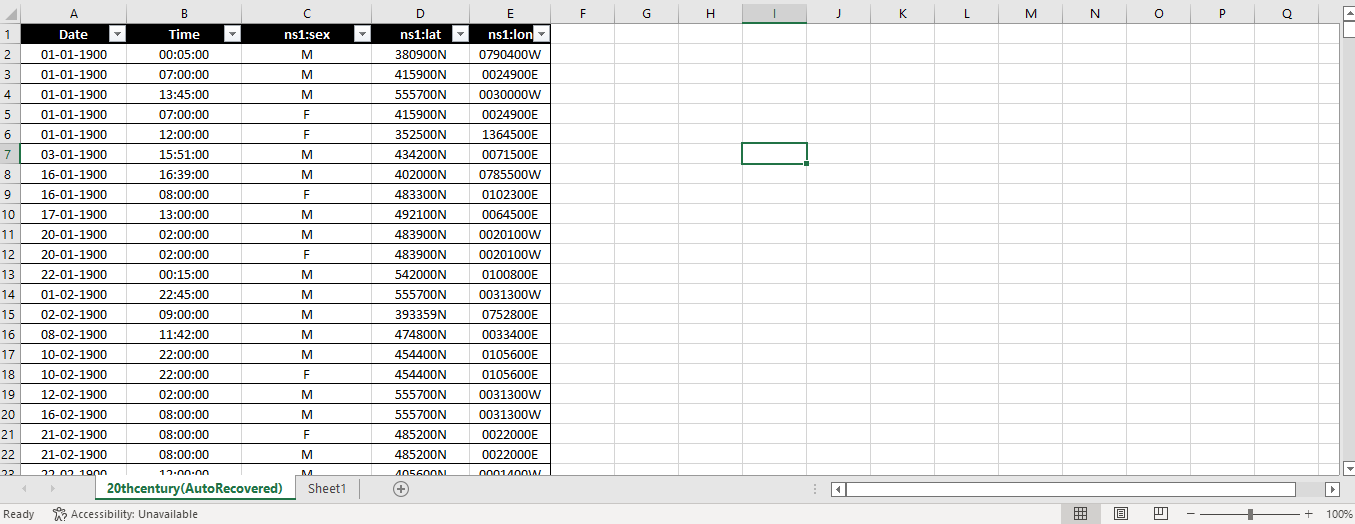
The primary difference between time series models and other types of models is that lag values of the target variable are used as predictor variables, whereas traditional models use other variables as predictors, and the concept of a lag value doesn’t apply because the observations don’t represent a chronological sequence.

From a statistical point of view, time series are regarded as recordings of stochastic process which vary over time.

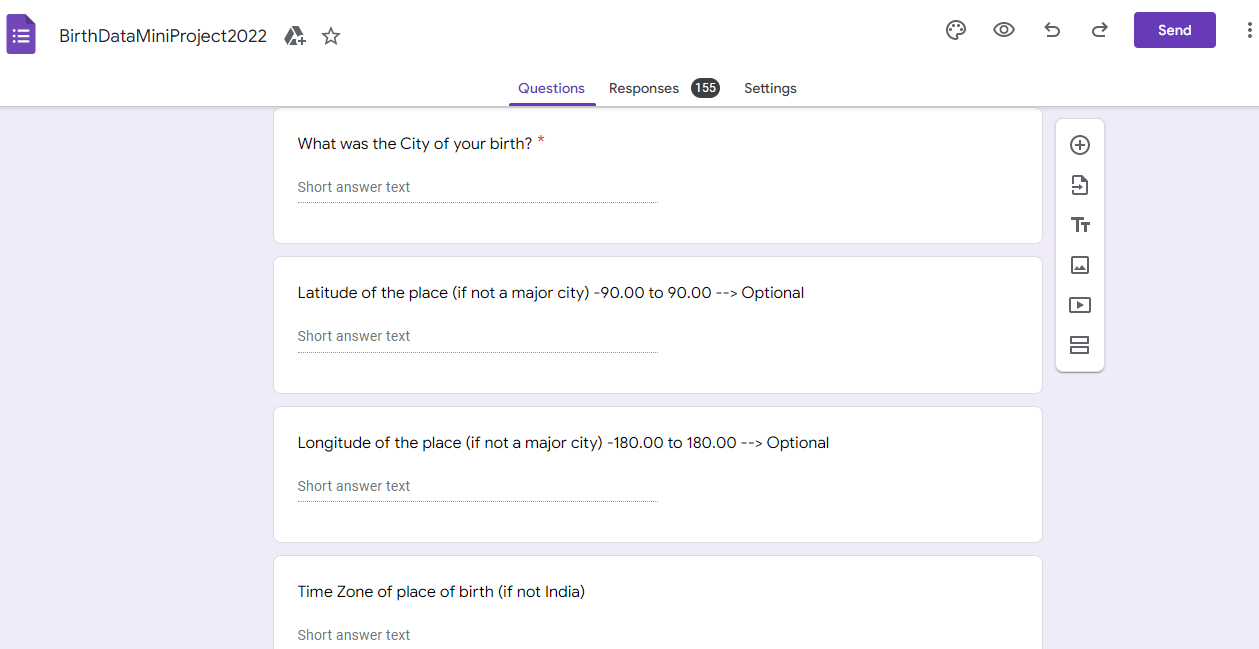
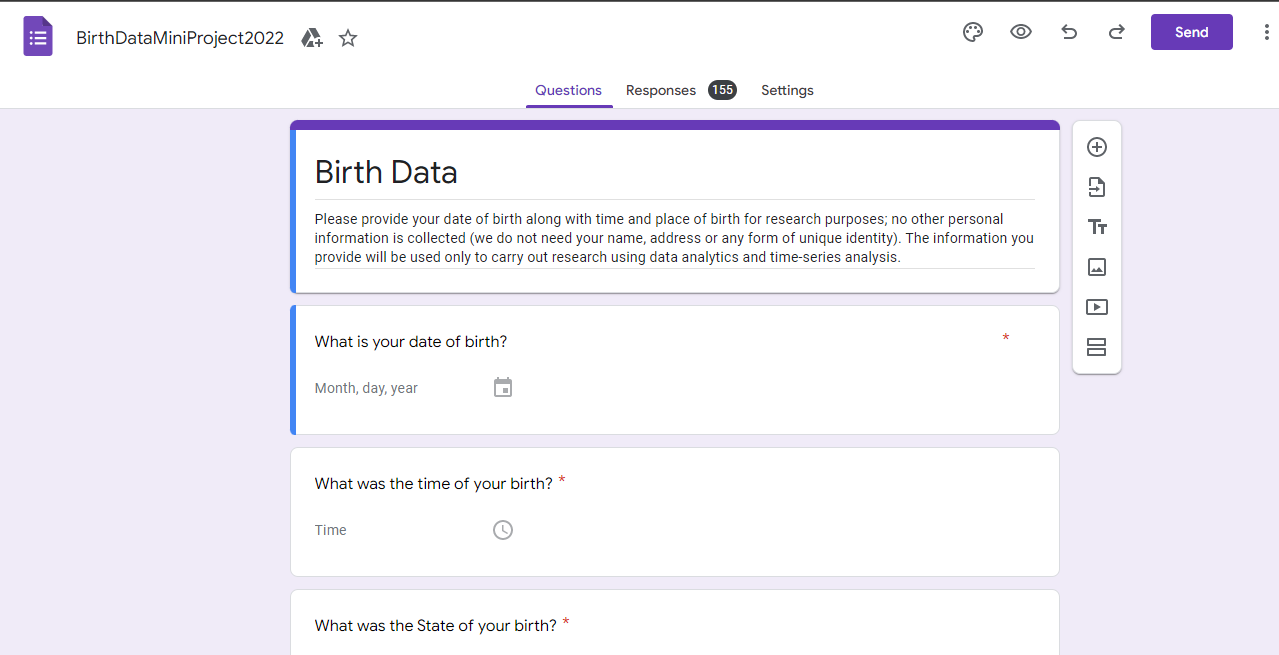
**Data Collection:**

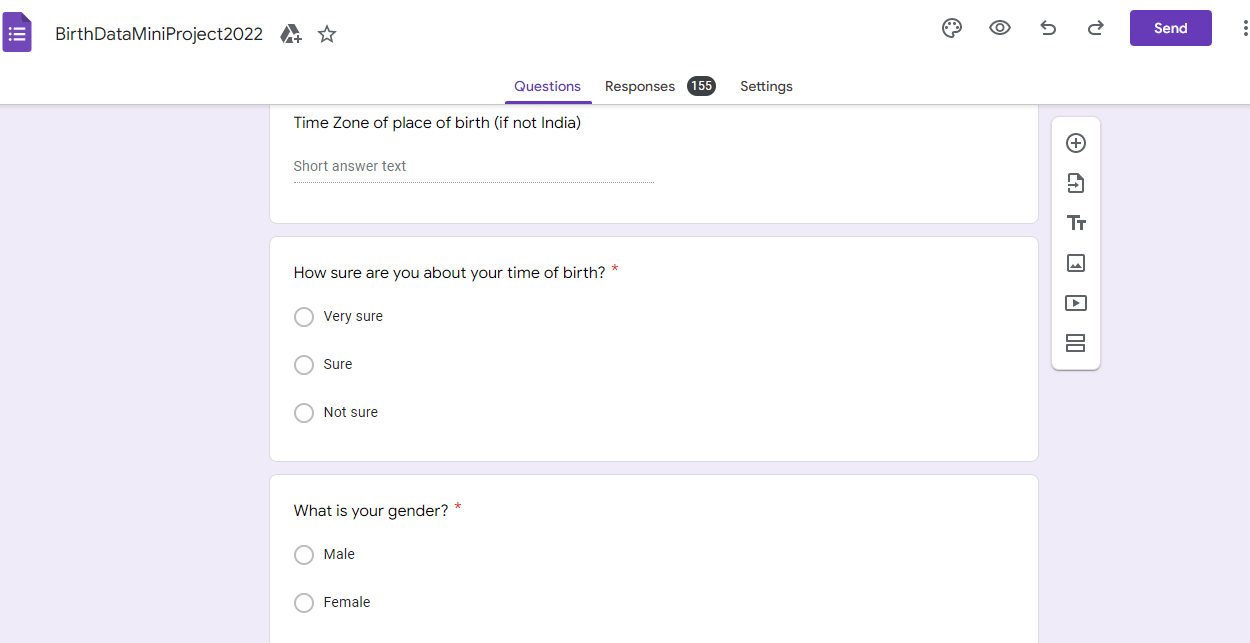
Coming to the second phase of our analysis i.e DATA COLLECTION. We created our own dataset. Our team collected data from multiple sources which include government hospitals and surveys, questionaries , form filling etc., Everything is entered into excel file. Our dataset has 5 attributes.

* Date of Birth
* Time of Birth
* Gender
* Latitudes
* Longitudes



Our team prepared a GOOGLE FORM to collect data through surveying.





Through this form we got quite a number of responses which was processed in further steps.

We got our hands on a large Russian dataset which are around 35000 data points

**Data Cleaning:**

Third phase of our Project is DATA CLEANING. Which is quite a work but we were able to manage and successfully clean it.

Data Cleaning: - Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

So, we proceeded in this phase with utmost care. As it is very crucial step, we can’t risk the data loss. So, the first step taken by our team is we replicated every raw source file and stored them in our local system.

Our most cleansing part is done using the best tool available out there that is MS EXCEL.

Initially our raw data contains around 35000 data points. Which consists of irregular data, inconsistent data, data columns with different formats within, wrong location co-ordinates, time recorded in different zones, duplicate values, data of births which makes no sense, etc.….

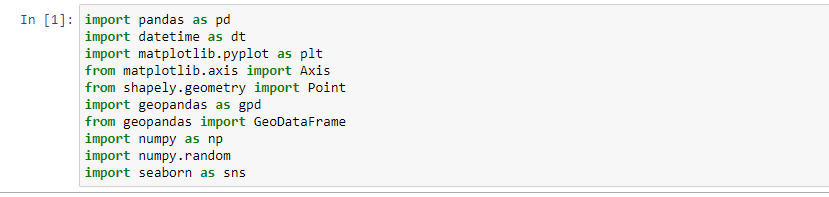
So, our team made a workflow for this phase and started cleaning them with MS EXCEL.

1. Firstly, we extracted the data column attributes which are more useful for our analysis and separated them.
2. Secondly, we converted every column to the same format which is quite understandable.
3. We removed the duplicates from the extracted dataset. Now we are with data which date back from 18th century to 21st century.
4. For to remove inaccurate data and perform analysis without bias or any error. We separated 20th century data and made a separate csv file.
5. Now to remove inconsistencies we used filters and sorting tools in Excel. We made sure that at every step we took care of bias factor.
6. Locations which were not mentioned with the guidance from our mentor removed them.
7. The data records were evenly spread across the 20th century so that the data doesn’t lean on a particular year.
8. At last, we were left with perfect 14898 data points which is good for performing analysis.
9. Latitudes and longitudes were given in different format for converting them into a universally acceptable format we used a function in excel FLASH FILL.
10. The cleaned dataset is subject to further phases.

**Data Analytics: -**

Data Analytics: - Analytics is the systematic computational analysis of data or statistics. It is used for the discovery, interpretation, and communication of meaningful patterns in data.

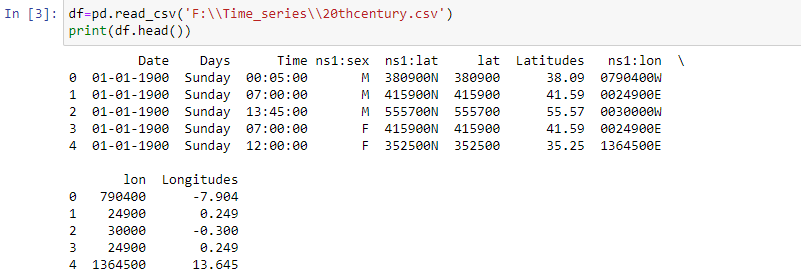
For this phase we used the most famous tool PYTHON.



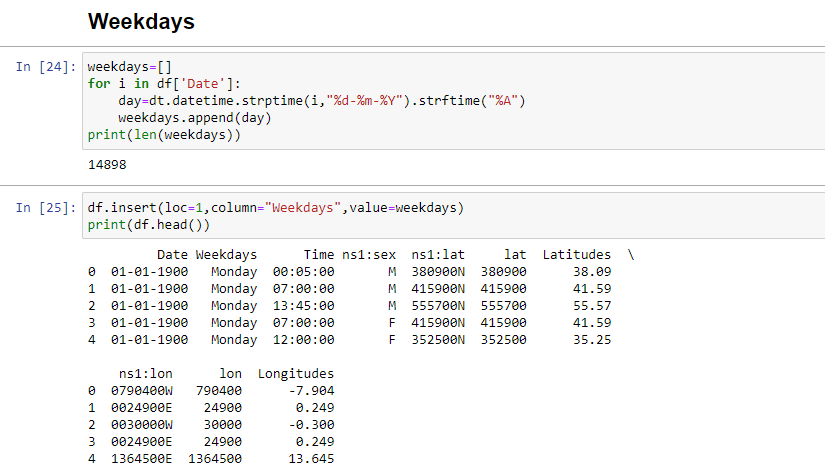
The packages that were imported so as to perform analytics and preprocessing.

From the directions by our mentor, we made our first task to be calculating the weekday of a particular day with the help of Julian date.

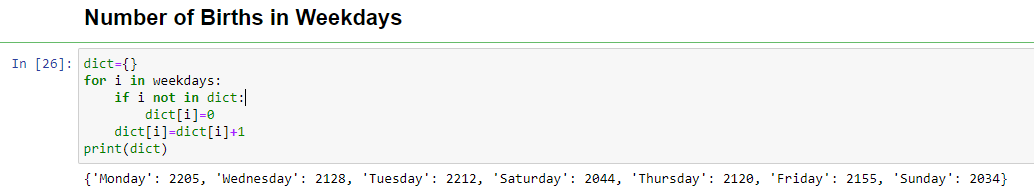
Julian Day: - The Julian day is the continuous count of days since the beginning of the Julian period, and is used primarily by astronomers, and in software for easily calculating elapsed days between two events.



The column above i.e. Days is calculated accurately in excel and the whole csv data is loaded to data frame. The code snippet for calculating weekdays is

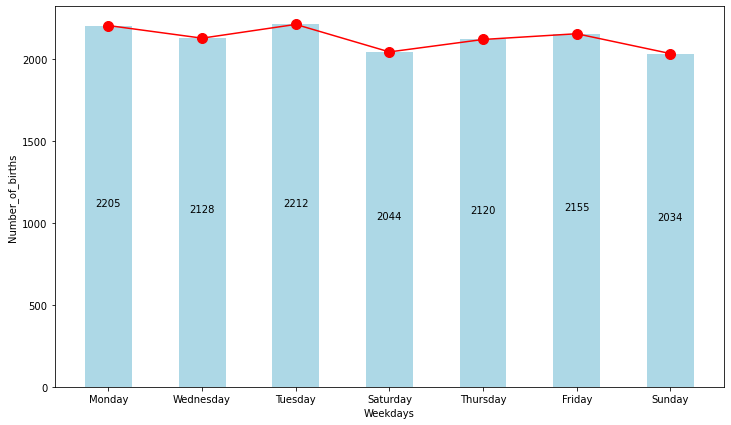


After calculation of weekdays column we computed the total number of births in weekdays for all 100 yr. span

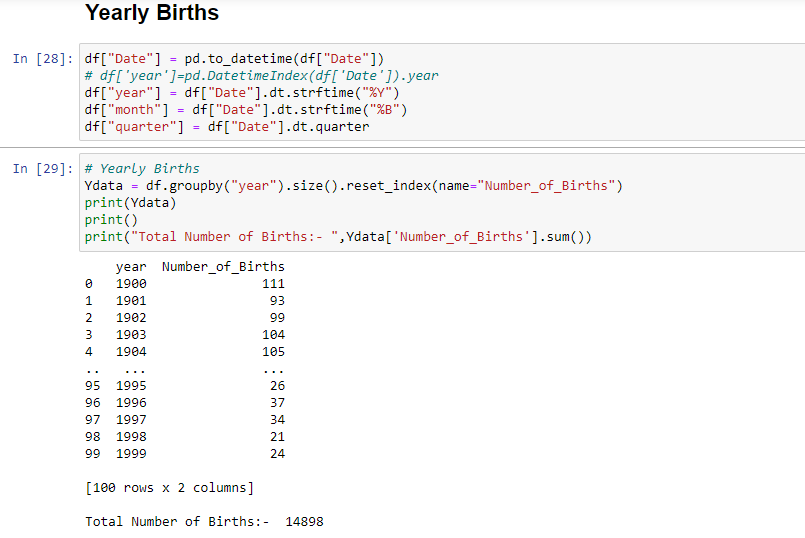


Plotting the number of births vs the weekdays the graph obtained was

**Scale: - x-axis: - Number\_of\_births, y-axis: - Weekdays**

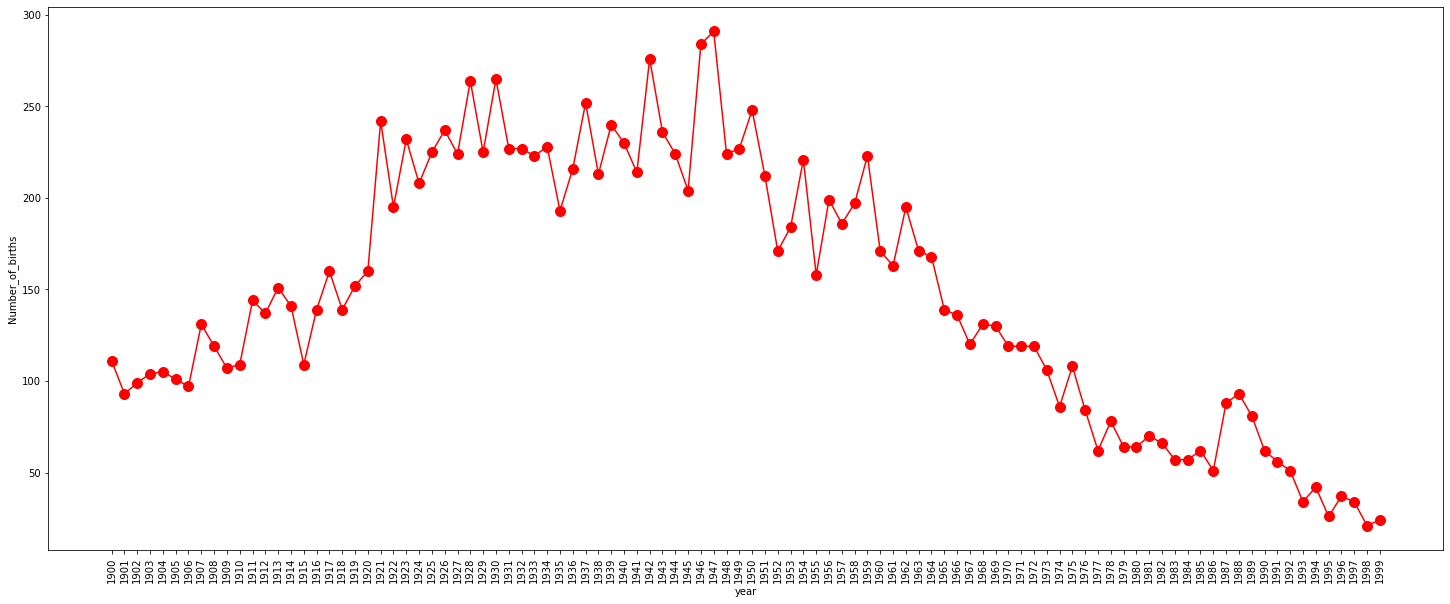


Our third task was calculating the number of people born in every year



After calculating the number of births, we plotted the graph between number of births and the respective years.

**Scale: - x-axis: - Number\_of\_births, y-axis: - Years**

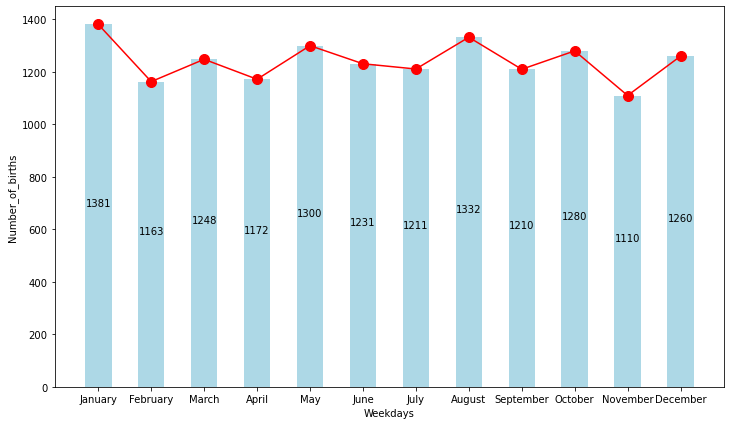


Next task we were assigned was calculate the number of months for all 100 yr. span



By plotting between the number of births and months the following graph is obtained

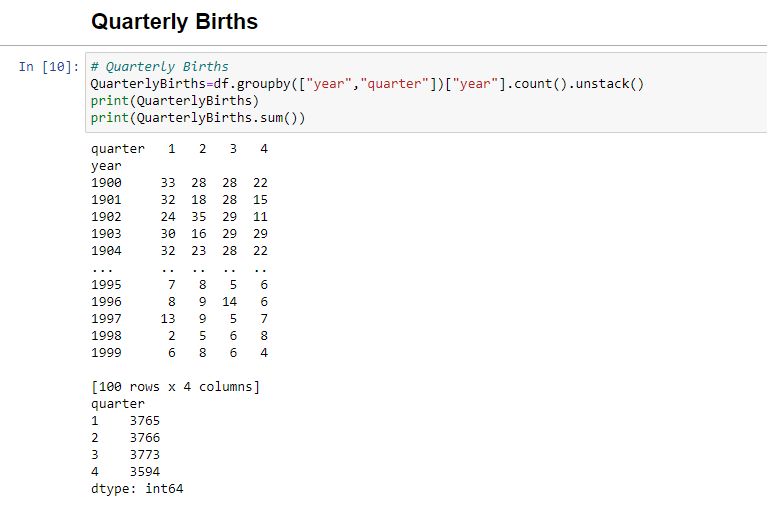
**Scale: - x-axis: - Number\_of\_births, y-axis: - Months**



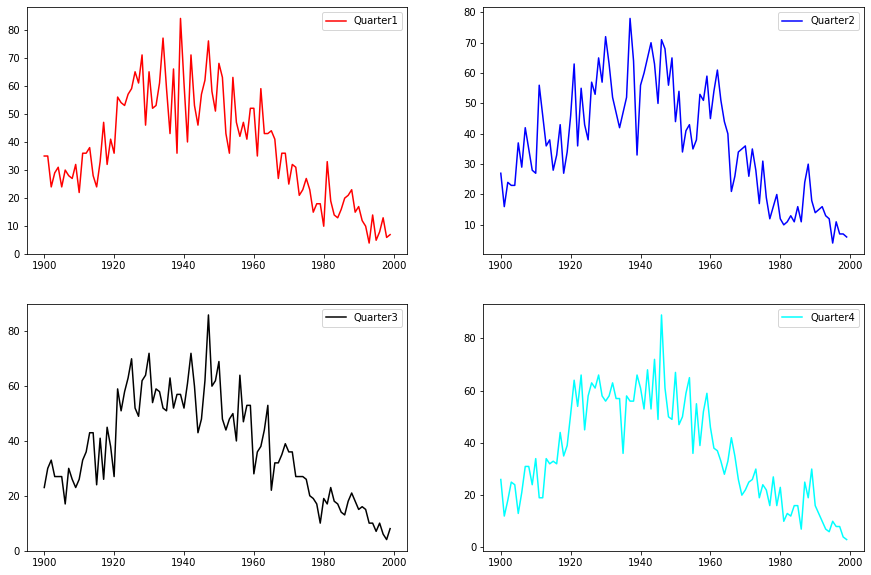
Next, we computed the quarter wise births.

The calendar year can be divided into four quarters, often abbreviated as Q1, Q2, Q3, and Q4.

* First quarter, Q1: 1 January – 31 March (90 days or 91 days in leap years)
* Second quarter, Q2: 1 April – 30 June (91 days)
* Third quarter, Q3: 1 July – 30 September (92 days)
* Fourth quarter, Q4: 1 October – 31 December (92 days)



We did plot a graph for each quarter vs years and the graph look like:

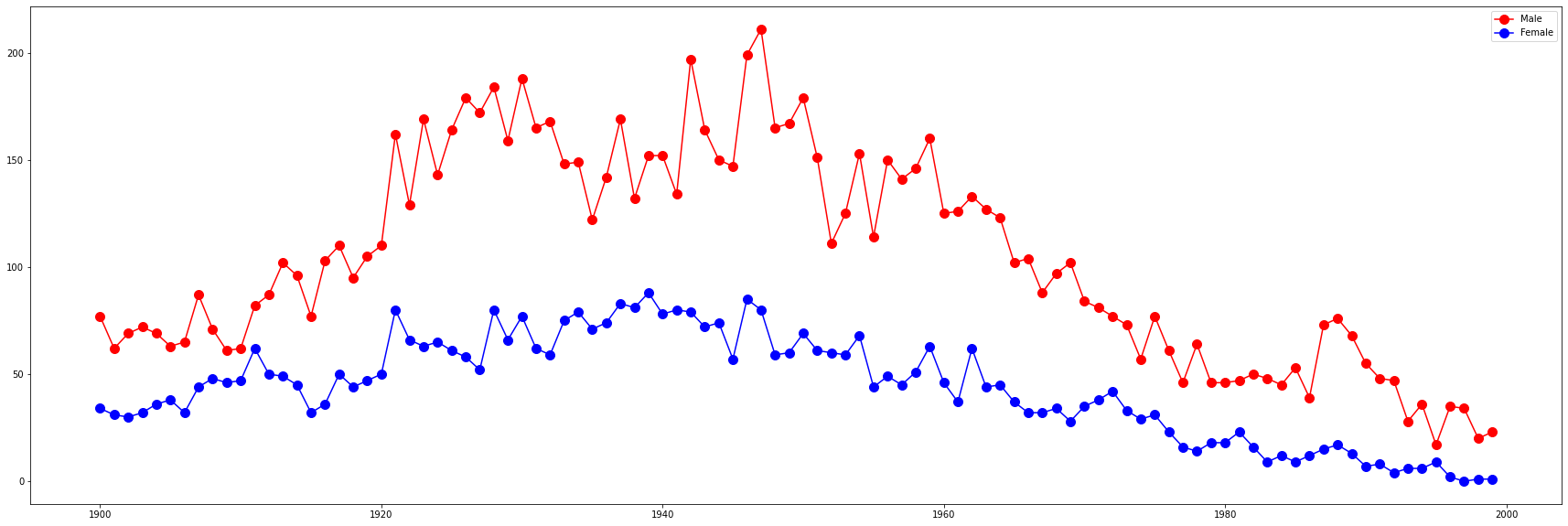


We also calculated the number of males and females born in every year for 100-year data.



After calculating of Gender births, we just plotted graph of two genders and compared them

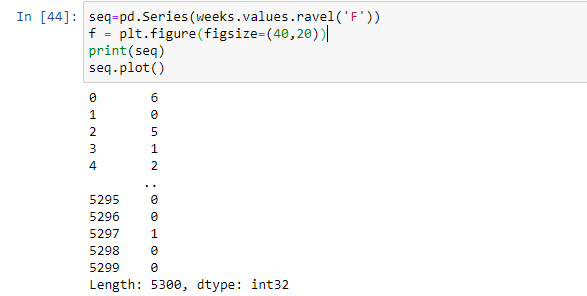
**Scale: - x-axis: - Number\_of\_births, y-axis: - Gender**



Till here statistical analysis, plots and formulas are done on processed data. After this we have made a time series dataset on weeks and months in the form of sequential data.

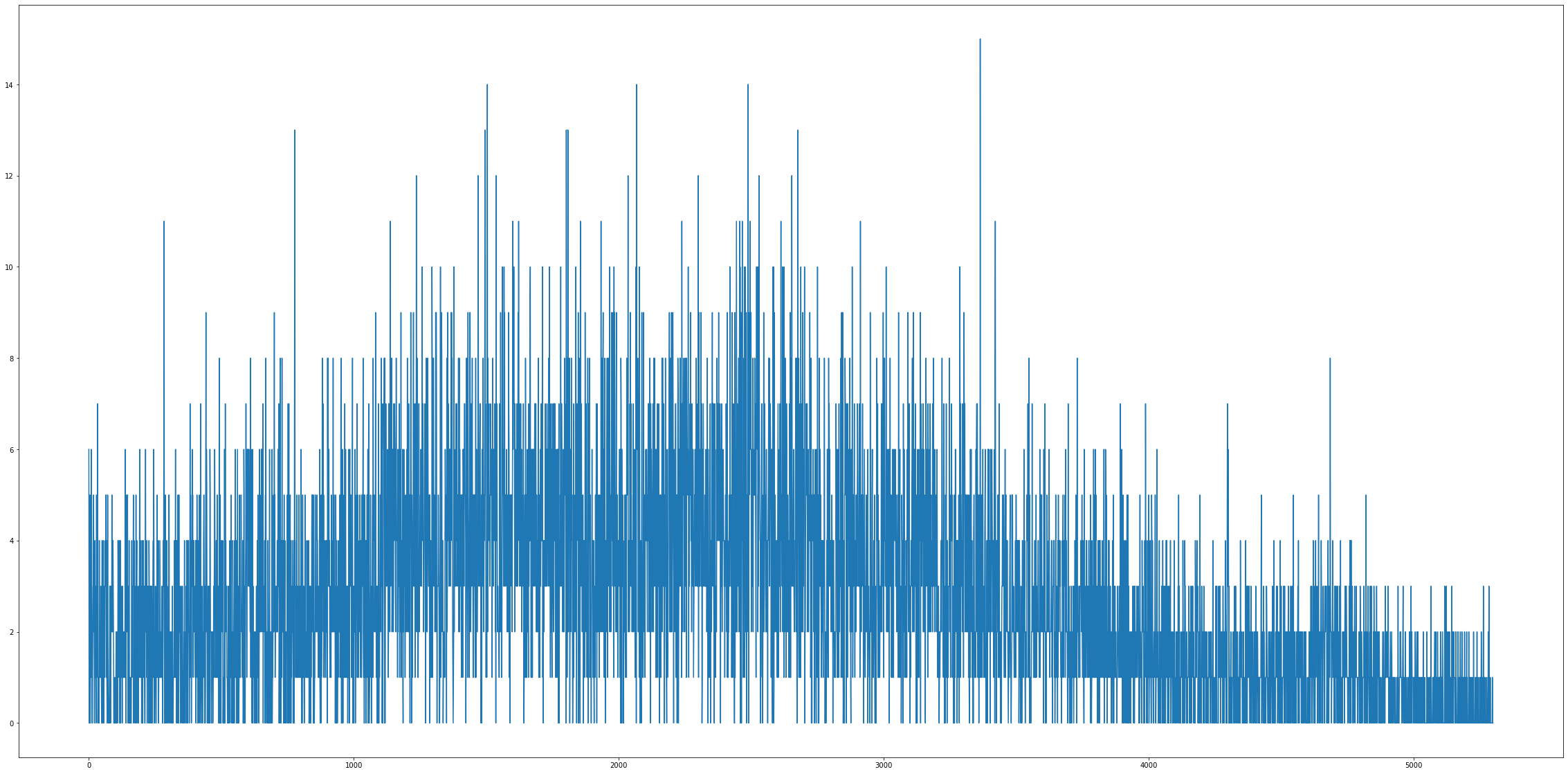
As you can see below, we calculated the number of births in every year for all 100 years we took into account of 53 weeks inconsideration for leap years after this we sequentially concatenated one year after another in terms of weeks like 1st yr. weeks + 2nd yr. weeks +…



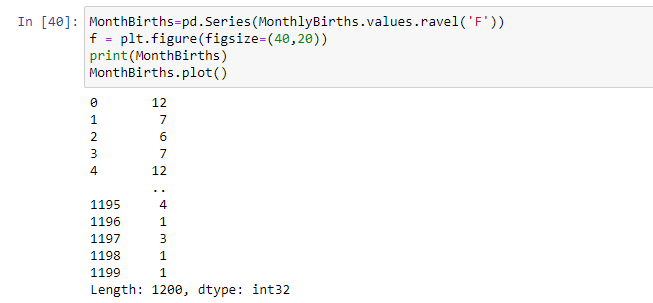


After arranging the values sequentially, we plotted the graph and some crazy insights from the graph and all 5300 points were plotted. The graph is below

**Scale: - x-axis: - Number\_of\_births, y-axis: - Week Births**

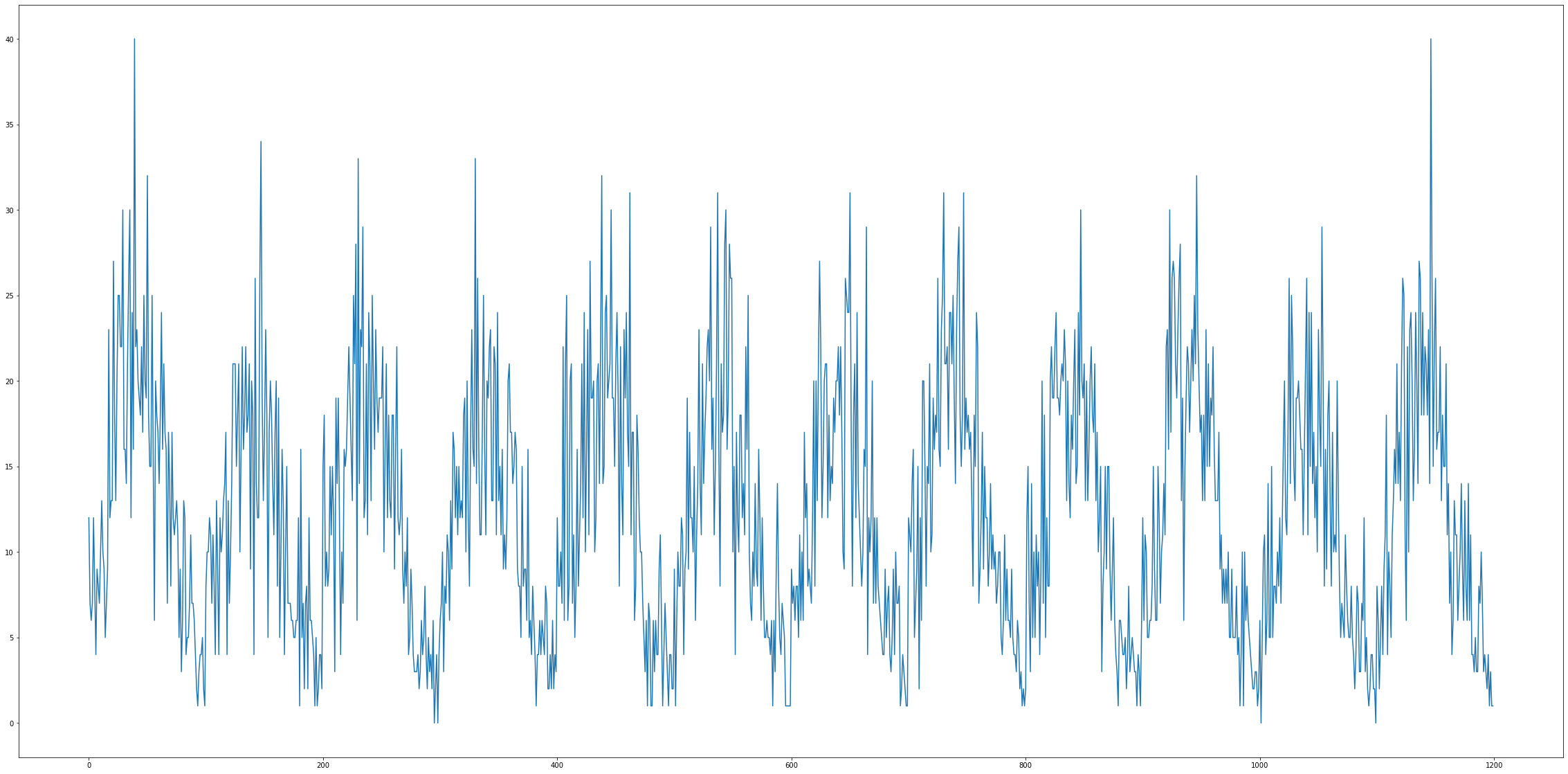


And next thing we have done is we made sequential time series dataset for months. Similar to previous dataset we concatenated 1st year months + 2nd year months +…..



We obtained 1200 datapoints and just plotted a graph with this data records.

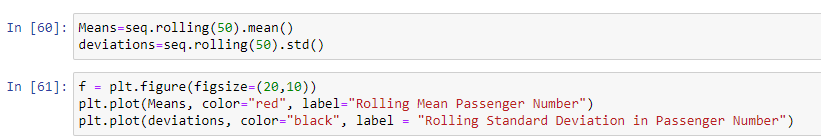
**Scale: - x-axis: - Number\_of\_births, y-axis: - Monthly Births**

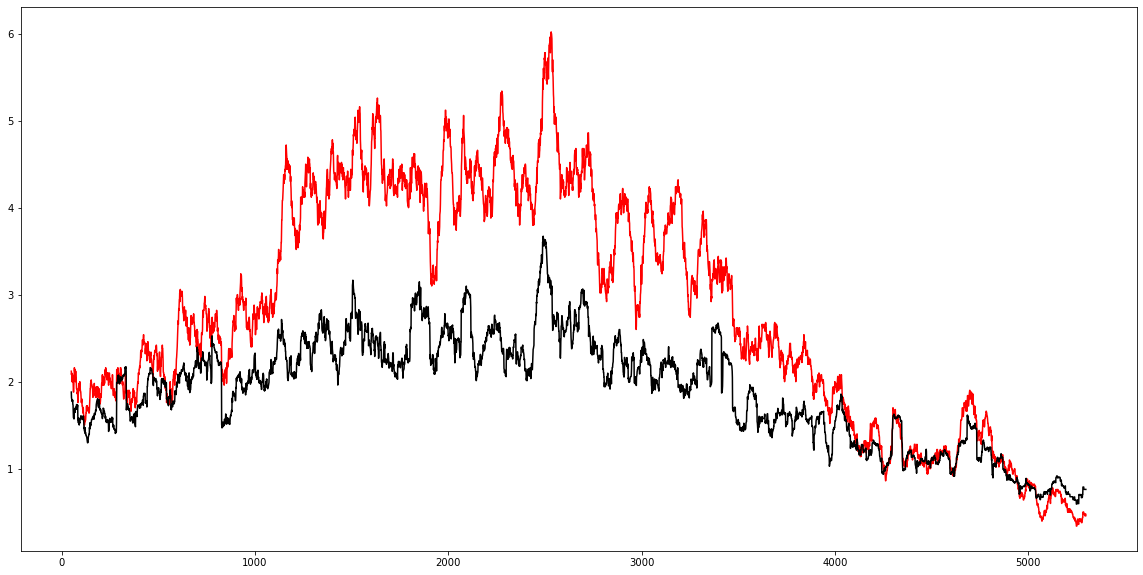


**Rolling Mean: -** In statistics, a moving average is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is also called a moving mean or rolling mean and is a type of finite impulse response filter.

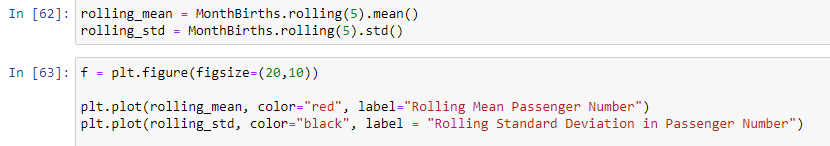
**Rolling Standard Deviation**: - It is derived by calculating an ‘n’ time period Simple Moving Average of the data item. It then sums the squares of the difference between the data item and its Moving Average over each of the preceding ‘n’ time periods. Finally, it divides this sum by ‘n’ and calculates the square root of this result.

We have done rolling mean and rolling standard deviation for weeks dataset.

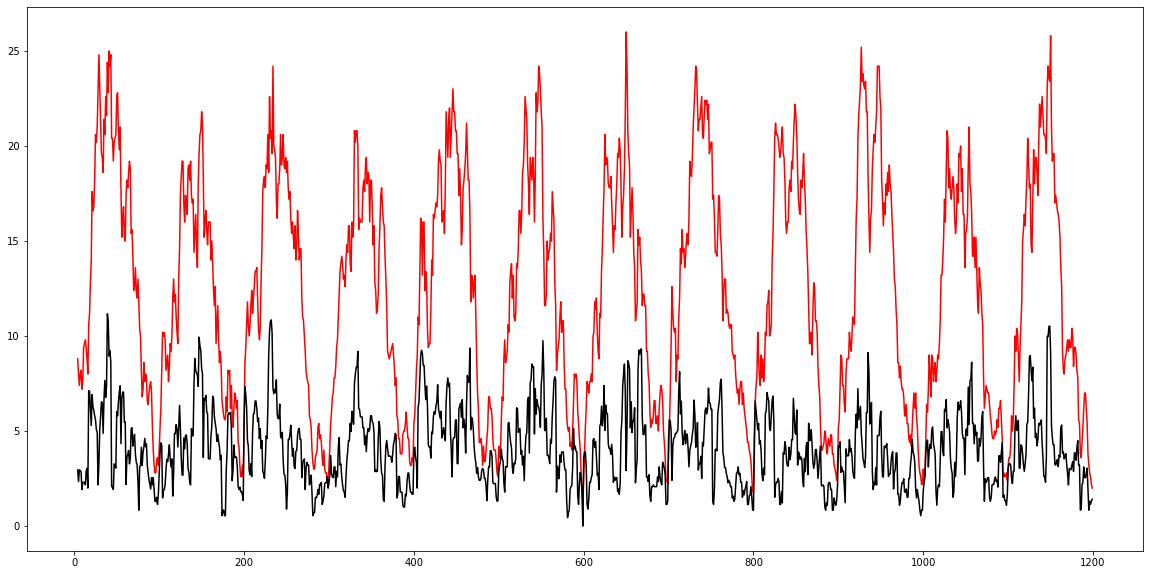
 **Scale: - Two plots in a graph Rolling Mean, Rolling Standard deviation for week dataset**



We have done rolling mean and rolling standard deviation for month dataset.

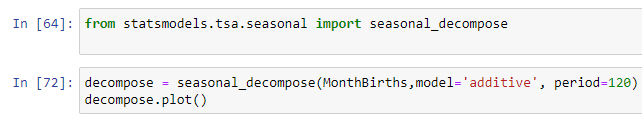


And the plot of the rolling mean and rolling standard deviation

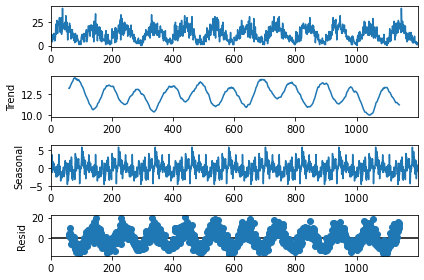
**Scale: - Two plots in a graph Rolling Mean, Rolling Standard deviation for Months dataset** 

Basically, a time series consists of four components. Variation of those components causes the change in the pattern of the time series. These components are:

* **Level:** It is the main value that goes on average with time.
* **Trend:** The trend is the value that causes increasing or decreasing patterns in a time series.
* **Seasonality**: This is a cyclic event that occurs in time series for a short time and causes the increasing or decreasing patterns for a short time in a time series.
* **Noise:** These are the random variations in the time series.



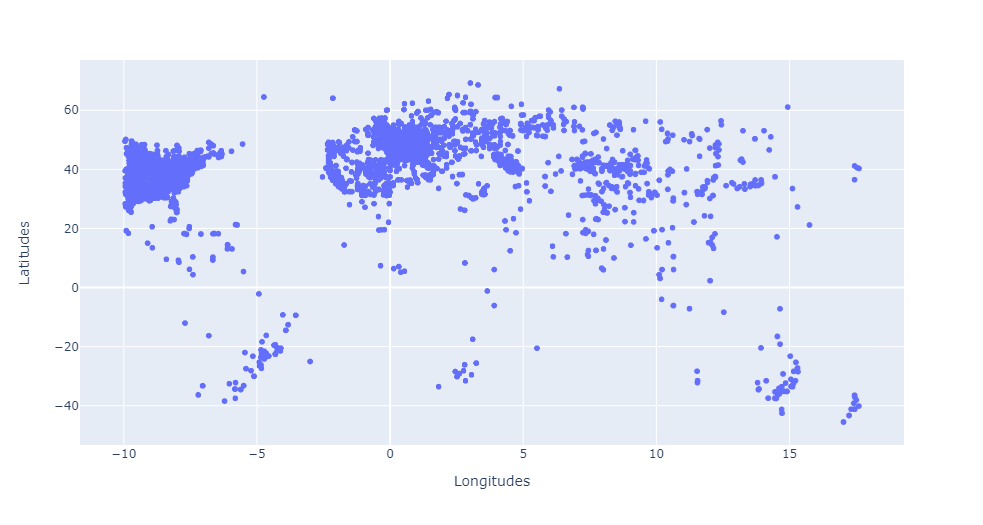
After running the code snippet, we were able to obtain component graphs



And our next assigned task was to work with geospatial data we made a scatter plot

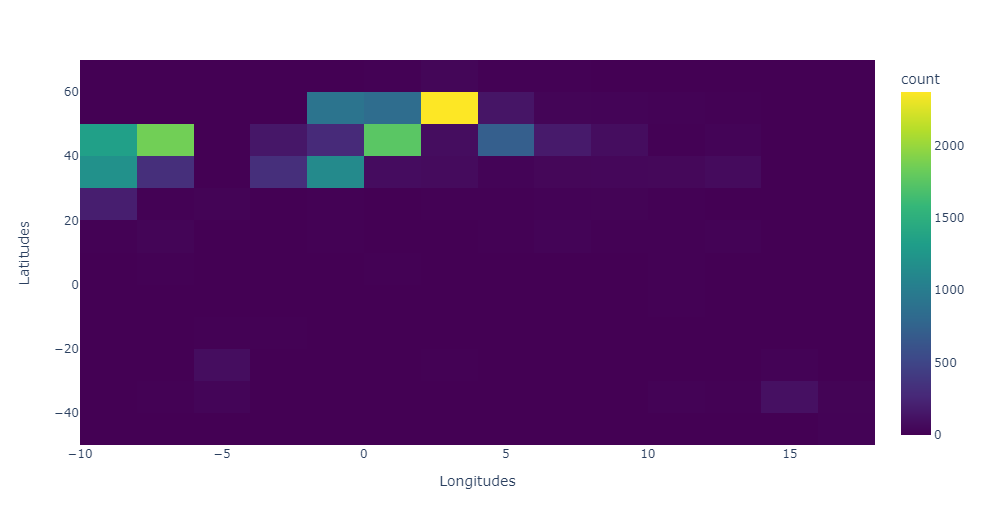


Scatter plot is produced as a result

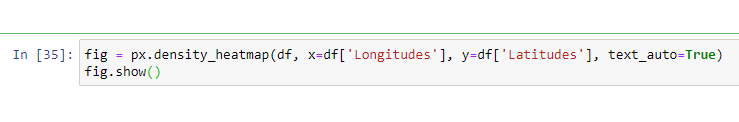


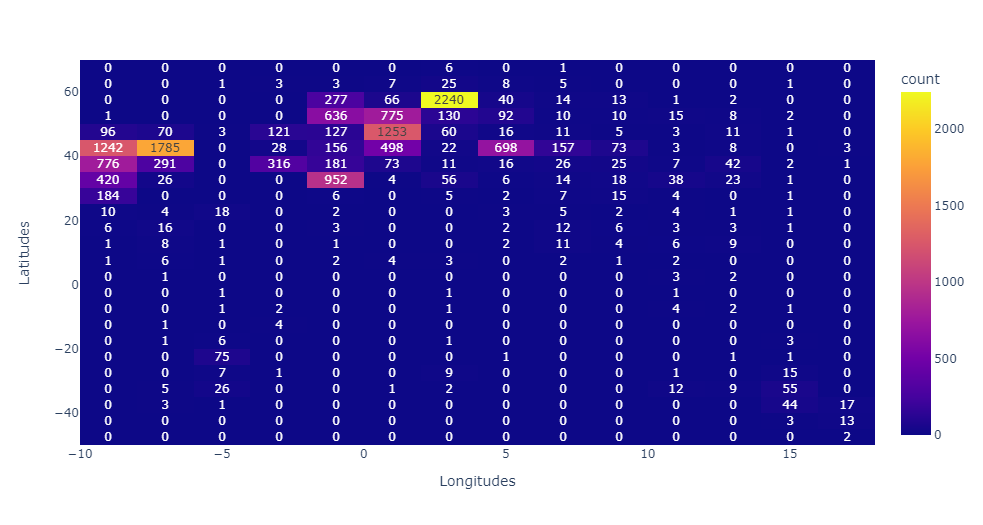


Heat density map is produced by running the above code snippet



We also divided the region into boxes and tried to produce the number of people born in it.





**Observation: -**

Many parameters and components were calculated and graphs were plotted.

1. Coming to number of births in weekdays we observed that a greater number of births were occurred On Tuesdays in this 100-year span.
2. We observed that in yearly births more births were in year 1947.
3. In monthly births people born in January were more.
4. In quarter births plots or yearly birth plots we didn’t find any significant observations but we can see roughly the population reached its peak in 1940’s and gradually the population decreased.
5. In gender births we can obviously observe that the number of female births is less than that of male births.
6. After plotting time series dataset, we observed a fascinating observation that there is actually a seasonality in the number of births.
7. From scatter plot there can be seen that some places the population density is high some places there is much scattering (low density).

**Conclusion: -**

1. Major conclusions we can conclude that there is a seasonality in the dataset we collected.
2. And also, the number of female births is very lower when compared to male births there can e many reasons for that some of them might female infanticide as parents then would have felt that girl is a burden.
3. More people were born on Tuesdays and Mondays maybe weekend hospitals were closed so they postponed the delivery -just a probabilistic conclusion.

**Acknowledgement: -**

We take this opportunity to express our profound gratitude and deep regards to my guide Dr. Kavi Mahesh for his exemplary guidance, monitoring and constant encouragement throughout the course of this Project. The blessing, help and guidance given by him time to time shall carry us a long way in the journey of life on which we are about to embark. His mentoring helped us to explore different concepts in the field of DATA SCIENCE. And the way of approaching the objective