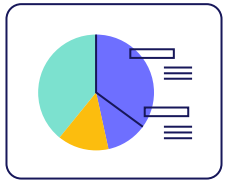
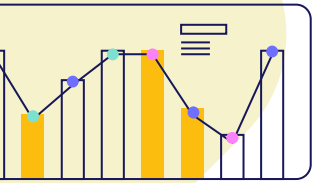
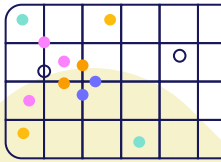


Introduction to Data Analysis

Descriptive Statistics

Fathima Khadija Ramzi -





Dataset we will be working with:

The dataset has **47** observations and **6** variables:

Independent Variables:

Fertility – Ig common standard fertility measure
Agriculture - % of males involved in agriculture as occupation
Examination - % of draftees receiving highest mark on army examination
Education - % education beyond primary school for draftees
Catholic - % “catholic” (as opposed to “protestant”)

Dependent Variable:

InfantMortality – live births who live less than 1 year

These are the Key Statistics of the dataset in review:

	count	mean	std	min	25%	50%	75%	max
Fertility	47.0	70.142553	12.491697	35.00	64.700	70.40	78.450	92.5
Agriculture	47.0	50.659574	22.711218	1.20	35.900	54.10	67.650	89.7
Examination	47.0	16.489362	7.977883	3.00	12.000	16.00	22.000	37.0
Education	47.0	10.978723	9.615407	1.00	6.000	8.00	12.000	53.0
Catholic	47.0	41.143830	41.704850	2.15	5.195	15.14	93.125	100.0
InfantMortality	47.0	19.942553	2.912697	10.80	18.150	20.00	21.700	26.6

What will we cover in this analysis?

01.

Independent Variable 01 - Fertility

Type of distribution, Central Tendency Metrics, Standard Deviation Spread Graphs

02.

Independent Variable 02 - Agriculture

Type of distribution, Central Tendency Metrics, Standard Deviation Spread Graphs

03.

Independent Variable 03 - Education

Type of distribution, Central Tendency Metrics, Standard Deviation Spread Graphs

04.

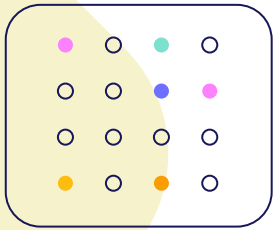
Key Statistics - General Insights

3 Identified general insights

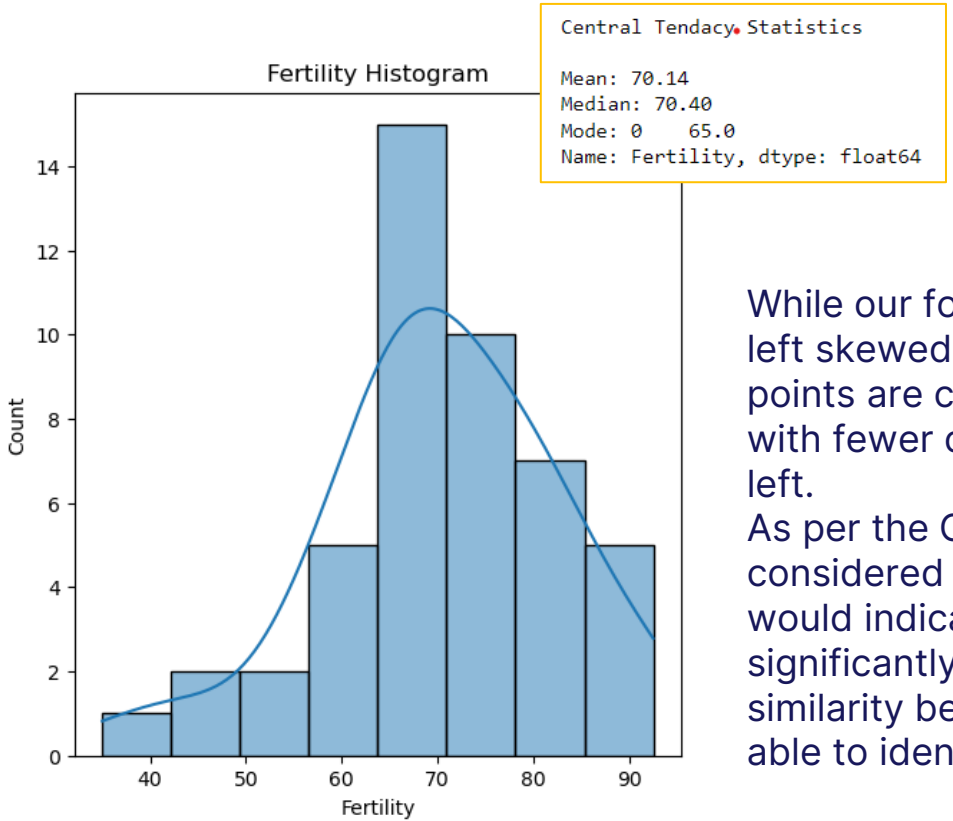
01.

Independent Variable 01 - Fertility

Let’s look at the Type of Distribution, Central Tendency Metrics, Standard Deviation Spread Graphs



Fertility – Type of Distribution



Left Skewness (Negative Skew)

Looking at just the histogram, we can identify a slight left skewness (negative skew). But, as the histogram looks approximately bell-shaped, we can possibly justify the distribution as approximately normal.

While our following findings will prove that Fertility data is left skewed, we should also note that most of the data points are concentrated on the right side of the distribution, with fewer data points on the left side, creating a tail on the left.

As per the Central Tendency Statistics, even if we considered that the Mean is almost equal to Median, which would indicate a normal distribution type, the mode is significantly less than Mean or Median (in comparison to the similarity between Mean and Median). Therefore, we will be able to identify a left skewness.

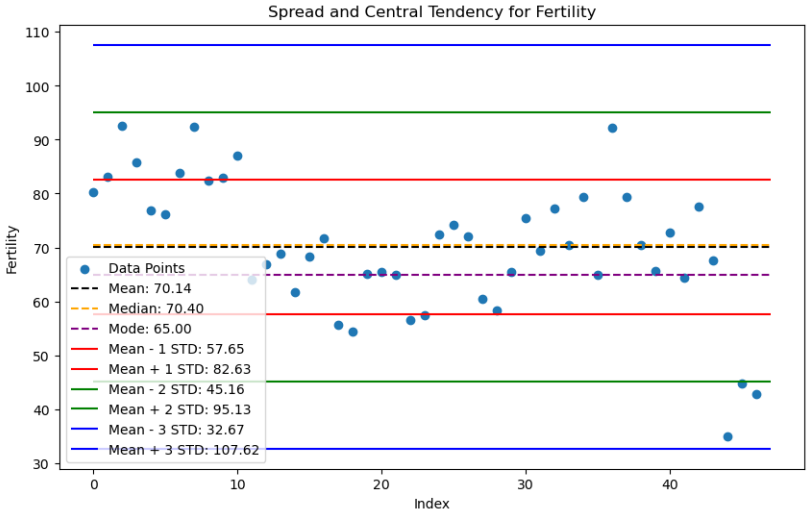
Fertility – Ideal Central Tendency Metric

Mean Vs. Median Vs. Mode

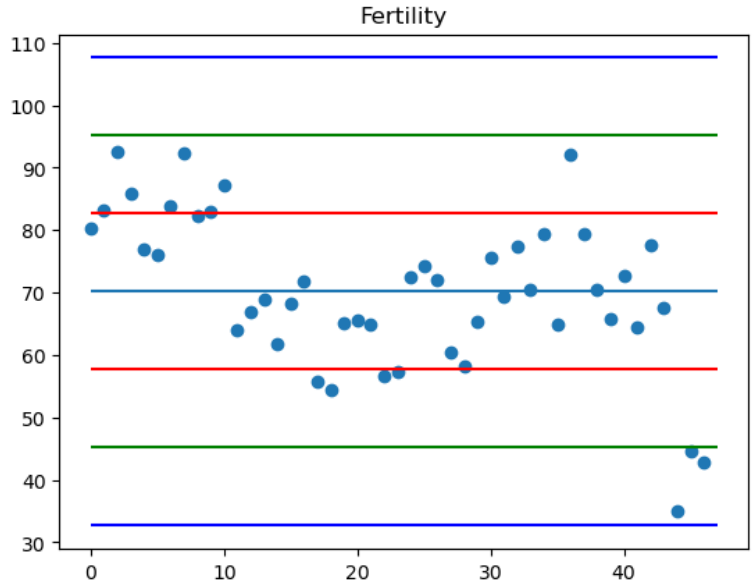
As we see even a slight left skewness, Median would be the best central tendency metric to describe the Fertility independent variable.
While the Median is not affected by skewness, it also is not affected by outliers in the data, considering there are a few notable low values.
The Median being the middle value remains unaffected, providing a more accurate central point for most of the Fertility dataset.
As mean is sensitive to outliers, the low values in the Fertility dataset will drag it down.
Furthermore, there are multiple Modes in the Fertility dataset, making it a less likely candidate to measure Central Tendency.

```
Central Tendacy Statistics

Mean: 70.14
Median: 70.40
Mode: 0    65.0
Name: Fertility, dtype: float64
```



Fertility – Standard Deviation Spread of Graph

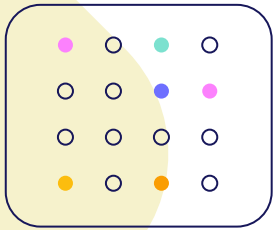


In reference to the following Standard Deviation Spread of Graph, there is a significant variability in the observations/data points of fertility rates from being as low as 35 and high as 92.5. We see this in the previous graph where the mean line was further away from the standard deviation line, whereas if it had been closer, it would've indicated less variability in the dataset.
There are a considerable number of outliers present as well.

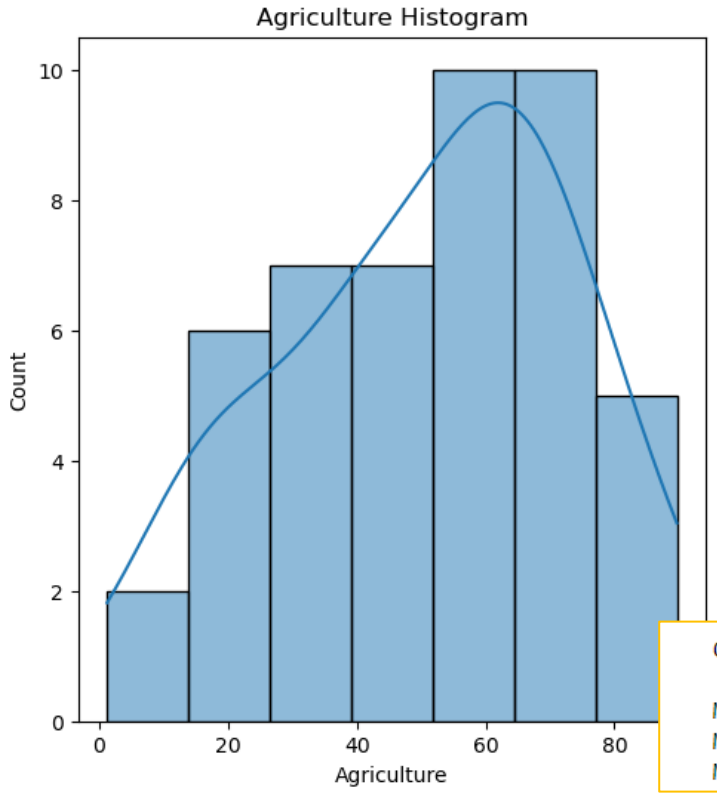
02.

Independent Variable 02 - Agriculture

Let’s look at the Type of Distribution, Central Tendency Metrics, Standard Deviation Spread Graphs



Agriculture – Type of Distribution



Central Tendacy Statistics

Mean: 50.66
Median: 54.10
Mode: 0 1.2

Left Skewness (Negative Skew)

The histogram indicates a left skewness (Negative Skew), as seen by the appearance of a left tail. Note that the mean is less than the median, while both are greater than the mode, which confirms negative skew.

The Mean being less than the Median indicates that the lower values are pulling the average down.

The mode being lower than both the mean and median, further supports the left skew.

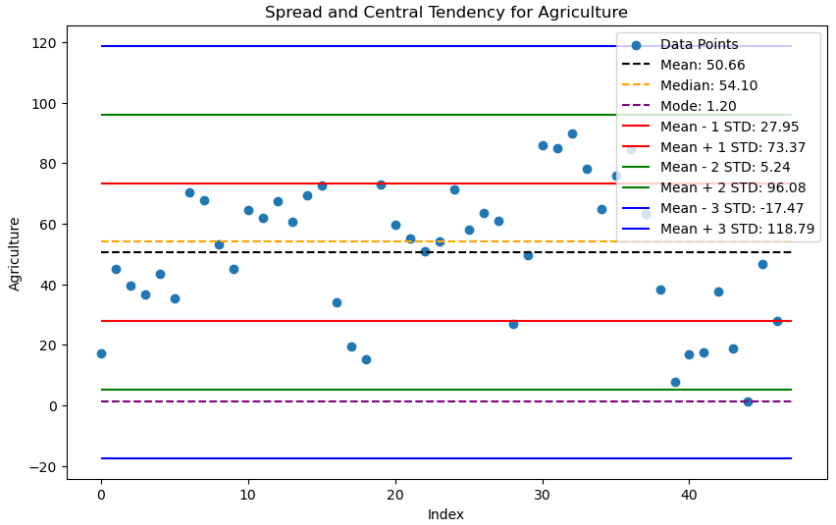
Agriculture – Ideal Central Tendency Metric

Mean Vs. Median Vs. Mode

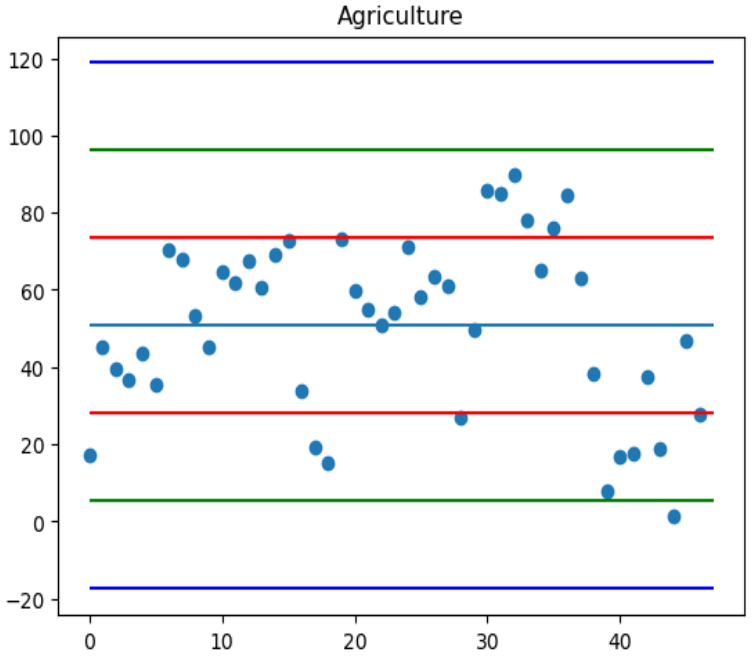
The Median is less affected by outliers and Skewness. Therefore, this would be the best central tendency metric for Agriculture.
As the dataset shows several low values which can influence the Mean but not the Median.
Further, due to the number of low values in the dataset, the Median will not be dragged towards the lower (unlike the Mean) and will remain close to the center of majority of the data, providing a more accurate centre point.
The Mode here is a range of lower values, making it comparatively less useful than Mean or Median to be a measure of Central Tendency

```
Central Tendacy Statistics

Mean: 50.66
Median: 54.10
Mode: 0      1.2
```



Agriculture – Standard Deviation Spread of Graph

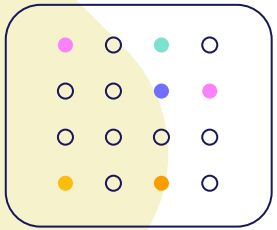


The following graph showcases data points ranging from 1.2 to 89.7, showcasing a wide range of values across the dataset.
The standard deviation being approximately at 22.71 suggests that there is a moderate variability in the data points clustering around the mean.
A few outliers are observed. Indicating extreme values in comparison to the rest of the data points.

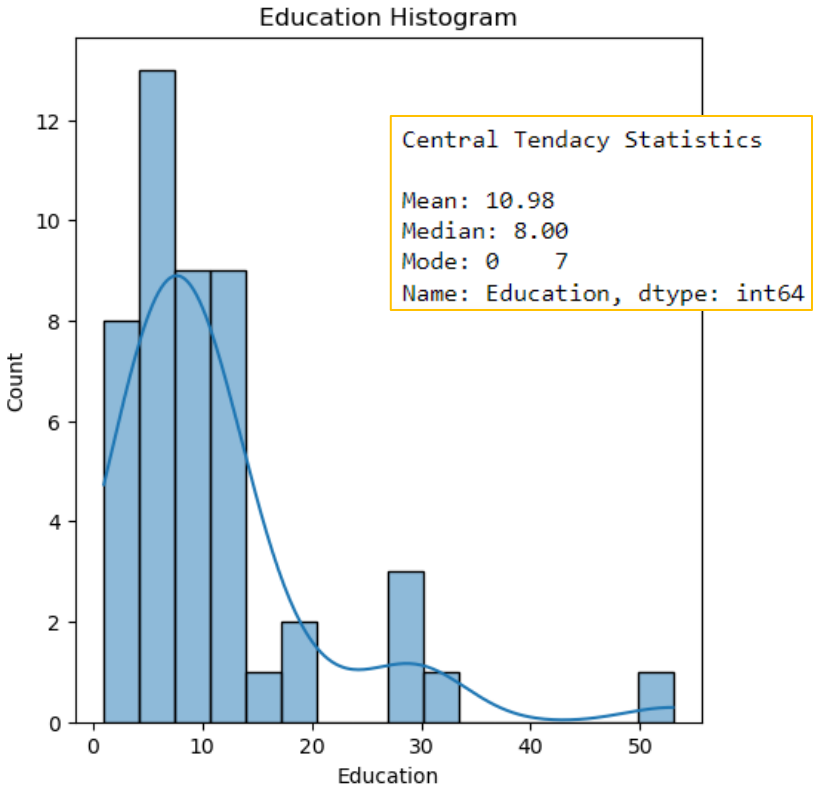
03.

Independent Variable 03 - Education

Let’s look at the Type of Distribution, Central Tendency Metrics, Standard Deviation Spread Graphs



Education – Type of Distribution



Right Skewness (Positive Skew)

As we see in the histogram, there is an appearance of a right tail, making it a distribution with right skewness (positive skew). Majority of the data points are focused on the left side. Further, the Mean is greater than the Median, which is greater than the Mode, which also explains that it is Right Skewed.
The Mean is being pulled to the right by the higher values, while the Mean and Median remain lower.

Education – Ideal Central Tendency Metric

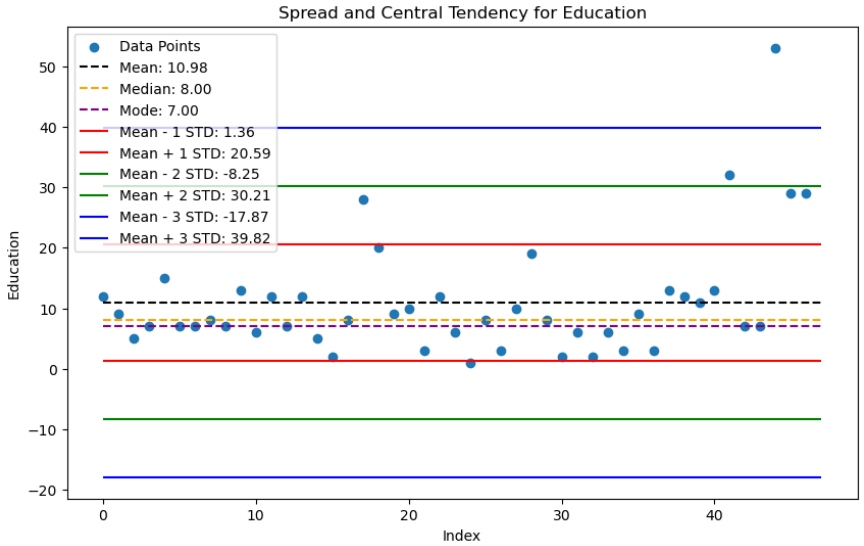
Mean Vs. Median Vs. Mode

As we determined that the Education dataset shows right-skewness, the Median will be the best central tendency metric. The median is less affected by outliers and skewness, compared to Mean. Here, the high values in the right will drag the mean towards the high end, making it a less appropriate measure of central tendency.

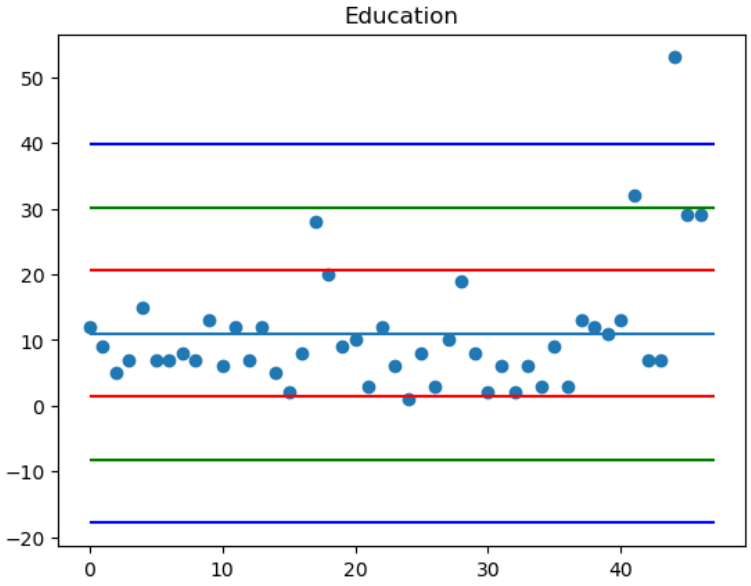
Mode is also not an ideal metric, as datasets with skewed distributions that has outliers on the higher end (positive skew) affect it disproportionately, while there are multiple modes present in the dataset as well.

```
Central Tendacy Statistics

Mean: 10.98
Median: 8.00
Mode: 0    7
Name: Education, dtype: int64
```



Education – Standard Deviation Spread of Graphs

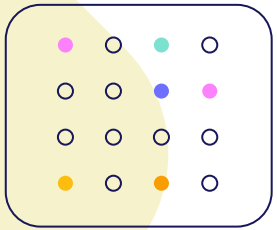


A large number of data points are clustered around the mean, indicating less variability among majority of the data points.
However, there are a few outliers observed indicating extreme values in comparison to the rest of the data points.
While minimum value of the dataset is close to the first standard deviation line, the highest value is an outlier indicating a deviatin from the Mean.

04.

Key Statistics
- General Insights

Let’s look at 3 identified general insights



- Fertility with a mean of 70.14 and moderate standard deviation, could mean that fertility rates are considerably high, which would in turn be a strain on the healthcare sector. This could relate to Infant Mortality.
- The percentage of male population being involved in agriculture indicates that a significant number of males live in rural areas, with minimal access to healthcare and education. This could relate to infant mortality.
- There is a significant variability in the Education beyond primary school among draftees. Higher Education percentage could have better healthcare, knowledge and awareness of Infant Mortality. Therefore, this could relate to Infant Mortality.

	count	mean	std	min	25%	50%	75%	max
Fertility	47.0	70.142553	12.491697	35.00	64.700	70.40	78.450	92.5
Agriculture	47.0	50.659574	22.711218	1.20	35.900	54.10	67.650	89.7
Examination	47.0	16.489362	7.977883	3.00	12.000	16.00	22.000	37.0
Education	47.0	10.978723	9.615407	1.00	6.000	8.00	12.000	53.0
Catholic	47.0	41.143830	41.704850	2.15	5.195	15.14	93.125	100.0
InfantMortality	47.0	19.942553	2.912697	10.80	18.150	20.00	21.700	26.6

Thank you!

Fathima Khadija Ramzi - 100948193

fathimakhadija.ramzi@dcmail.ca

+1 647 450 2118