**Documentation**

**Data Collection:**

* Data was collected on Social media and factors that helped me analyze various effects of Social media on youth (One assumed bias).
* Data collection was both telephonic and survey based with high accuracy and clarity. A total of 38 members were there. *[Note: Data has small sampling biasing as participant members were less and location based. Do not use it for real world analysis as it will not be accurate. This project is only for practicing data analysis.]*
* Data contains 12 columns as follows:
  + Serial\_number: counts total rows
  + name: Use of random names to preserve personal identity and privacy.
  + gender: Has two possible value M and F for male and female respectively.
  + time\_spend\_on\_Social\_media\_per\_day: Time spent on social media per day by a user.
  + rating\_social\_media: Rating each user gave to social media based on his/her experience (0-5).
  + self\_estimated\_health\_condition: A user’s personal views on his/her physical or mental condition with respect effect of social media.
  + productive\_work\_hrs\_per\_day: Productive work done by user per day in hours.
  + outdoor\_activity\_time\_per\_day: Outdoor activity time such as walking, exercising, etc. per day in hour.
  + educational: If user likes to view educational content (True(1) or False(0)).
  + motivational: If user likes to view motivational content (True(1) or False(0)).
  + informative\_and\_current\_affairs: If user likes to view informational or current affairs content (True(1) or False(0)).
  + entertainment: If user likes to view entertainment content (True(1) or False(0)).
* Data is highly accurate as it was personally taken giving sufficient time to user.

**Chain log**:

* Done conditional formatting on gender to have count of females. Also highlighted none cells.
* For Tylla it was found that health rating was none. In data it was observered that 79% of sample gave rating as 3 and average rated value was 3.45. Hence I conclude that none should be replaced by 3.
* For Charles productive work hours was not set. I observed that maximum data lies below 4hrs (76%) for productive work and 63% lies below 3 hours. Hence I conclude that median will represent data more. So I replaced none with median 2.5.
* A csv file was created to visualize data on tableau.
* A summary table was created to analyze column F and G.
* A second summary table was created average, mean and median values of columns.
* A new sheet was created with sorted columns(First self\_estimated\_health\_condition and second rating\_social\_media) .
* A correlation table was made in sheet 2.

**Insights:**

* Male and Females are not different while analysis social media impact or analysis is independent of gender (no visualization was different for male or female).
* Around 81% of user rated bad health effect of social media >=3 and 55% rated it as 4 or 5.
* Most user rated social media as 3 with rare votes on 0 or 5. Nobody deeply loves or hates social media on average.
* Around 80% of users like to watch entertainment and informational affairs.
* There seems to be no correlation between time on social media and productive work.
* There seems to be no correlation between time on social media and outdoor activities.
* Average social media user spends around 3 hours on social media.
* Average productive work done by users is 2.5 hours each day.
* Outdoor activities lay between 0-1 hours for most of the people.
* User who rated there health good(2) tends to do more productive work, at the same they rated it 4. People doing nice productive work and health enjoys social media as well.
* Outdoor activity is also a factor that let people like social media as they tend to be healthy.

[**NOTE**: *Data may have sampling biases as it is based on a particular city and it is less. Analyze it for study based projects not on real life projects.]*