**Detailed Analytics Report**

Given dataset is extracted from “The Panel Study of Income Dynamics (PSID)” which is the longest running longitudinal household survey in the world. In this dataset there are 4,856 data points with nine different fields. In our study of this data set, we have calculated the mean, standard deviation, minimum, 25% value, 50% value, 75% value and max value for the fields of age, education level, earnings, working hours and number of kids in a family. These values are tabulated in table 01 and visualized in different graphs in attached jupyter notebook file.

**Visualization**

In the data visualization stage, we first tried to remove the outliers in the dataset. According to the box plots, we identified that there are outliers in the education level distribution, earnings distribution and in number of kids’ distribution. In order to carry out further analytics we removed these outliers from the extracted data. During this process following assumptions are taken

* Removing records having earnings higher than 55000 will not affect the mean of the dataset
* A family cannot have number of kids higher than 6. Removing those outliers will not affect the dataset’s mean
* Education level lied between 9 and 20

In order to identify correlation between the parameters, we used spearman correlation coefficient on filtered data set. According to the observations there was a significant positive correlation between earnings and working hours.

**Hypothesis Testing**

In the hypothesis analysis, as per the first hypothesis study, we have taken the effect of income to the number of kids in a family. Therefore, as per the null hypothesis, we assumed that there is no effect on income to the number of kids in a family. This is visualized as the average income of families with having 2 kids or less than 2 kids is equal to average income of general population. And as per the alternative hypothesis, if there is an effect, sample mean income is unequal to the population mean. Z value for the average income of the sample is calculated as 0.1324. Therefore, as conclusion, we cannot reject null hypothesis with 95% of significance level.

As per the second hypothesis study, we’ve considered the effect of age to the income. To formulate this, a sample is taken with data points with having age higher than 35. As for the null hypothesis, we assume that the average income of the sample is same to population average income and vice versa in alternative hypothesis. From the given dataset we’ve calculated the Z value of sample income as 0.092 and in here also we couldn’t reject the null hypothesis with 95% significant level.