**Descriptive Analysis on Drinking Dataset**

**Questions for this analysis :**

A researcher is interested in observing the role of education, gender, and drinking habit on income. The aim is to help him by addressing the following questions from the obtained data:

**Education : 0 : No Education , 1 : High School , 2 : Class 12 , 3 : Graduation , 4 : Masters , 5 : PHD**

**Drinking Habits : 0 : Not at all , 1 : Drinks Rarely , 2 : Drinks Occasionally , 3 : Not Sure , 4 : Social Drinker , 5 : Regular Drinker , 6 : Heavy Drinker , 7 : Very Heavy Drinker**

**Gender : 1 : Male , 2 : Female**

**1. Which variable has not been assigned the right measurement scale?**

Variables namely -

i. Education

ii. Drinking Habits

were not assigned the right measurement scale.

Scale given : Nominal (for both)

Correct scale : Ordinal (for both)

**2. What is the median salary of people studied in sample? Is it different from the average salary? Is it same as 50th percentile.**

|  |  |  |
| --- | --- | --- |
| **Statistics** | | |
| income | | |
| N | Valid | 35 |
| Missing | 0 |
| Mean | | 8336453.8571 |
| Median | | 1265000.0000 |

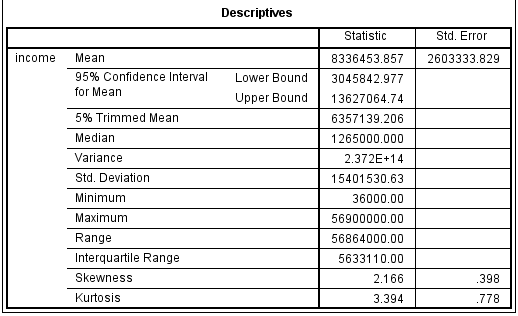
The median salary of people studied in the sample is 1265000.00 .

The mean salary of people is 8336453.86 which is quite different and also higher than the median salary of 1265000.00 .

Reason : The difference in the mean and median salaries is due to the outliers in the dataset. The mean is affected by the outliers whereas the median is not.

The median salary is same as the 50th percentile.

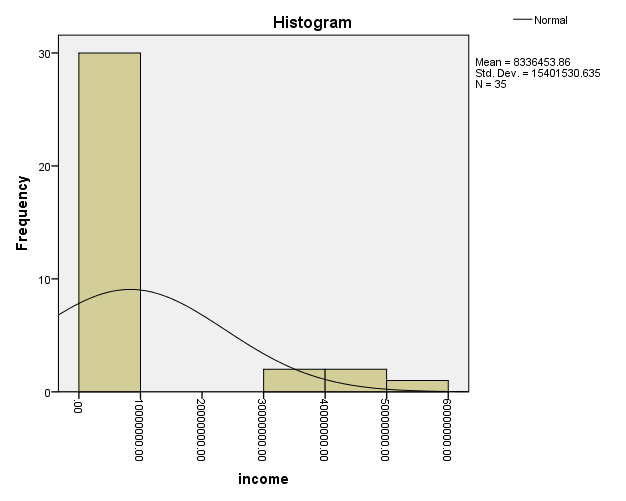
**3. Is income a normally distributed variable? Does distribution of income for PhDs follow a normal probability pattern as compared to High School pass subjects? Use Kurtosis, Skewness and Shapiro-Wilk & Kolmogorov-Smirnov tests.**



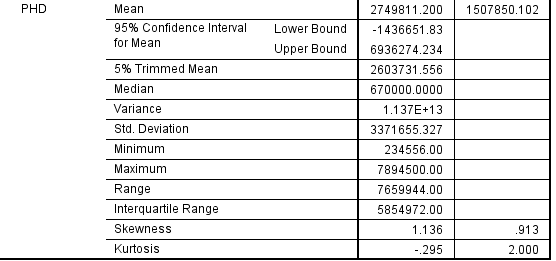
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | df | Sig. |
| Income | .340 | 35 | .000 | .569 | 35 | .000 |
| a. Lilliefors Significance Correction | | | | | | |

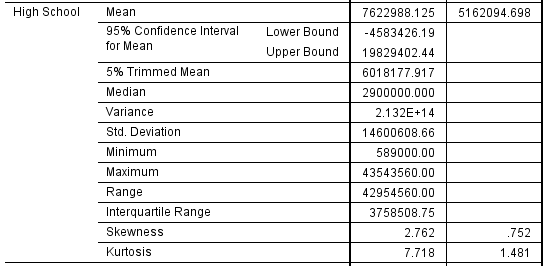
The income variable has a value of skewness and kurtosis of **5.44** (2.166/0.398) and **4.36** (3.394/0.778) respectively. For a variable to be normally distributed, the value of skewness and kurtosis should lie within the interval of **+1.96 to -1.96**. Looking at the significance of Kolmogorov-Smirnov and Shapiro-Wik test, we get a value of 0.00 which is less than 0.05. If the test significance value is less than 0.05 at 95% confidence interval, then we reject the null hypothesis(in our case H0 = Income is normally distributed) and accept the alternative hypothesis(H1=Income is not normally distributed).

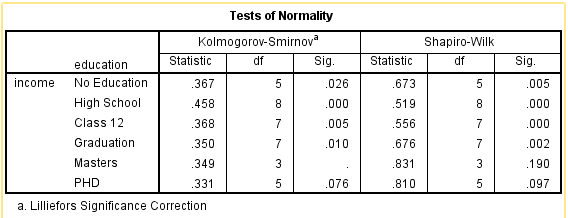
Also the graph of normal curve is not equally distributed



Hence, income is not normally distributed in our dataset.







Skewness and Kurtosis for PHD students is **1.24**(1.136/0.913) and **-0.15**(-0.295/2.00) respectively, which is in the range of **+-1.96**. Skewness and Kurtosis of High school students is **3.67**(2.762/0.752) and **5.21**(7.718/1.481) respectively, which is **NOT**in the range of **+-1.96**.

From the table of Tests of Normality Sig. of PHD students from Kolmogorov-Smirnov and Shapiro-Wik test is 0.076 and 0.097 which is greater than 0.05.So, we accept the null hypothesis of normality. Sig. of High school students from Kolmogorov-Smirnov and Shapiro-Wik test is 0.000 which is lesser than 0.05.So, we reject the null hypothesis of normality.

Hence, we can say that income for PHDs follow a normal probability pattern as compared to High school pass subjects.

**4. Find out and eliminate outliers from the income variable and retest its normality.**

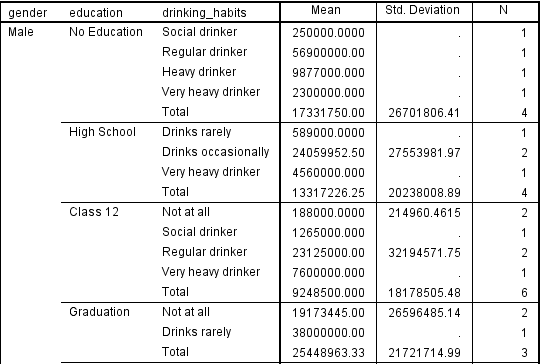
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptives** | | | | |
|  | | | Statistic | Std. Error |
| income | Mean | | 2315410.8333 | 491959.75116 |
| 95% Confidence Interval for Mean | Lower Bound | 1309240.1675 |  |
| Upper Bound | 3321581.4992 |  |
| 5% Trimmed Mean | | 2055076.8519 |  |
| Median | | 835000.0000 |  |
| Variance | | 7260731902961.937 |  |
| Std. Deviation | | 2694574.53097 |  |
| Minimum | | 36000.00 |  |
| Maximum | | 9877000.00 |  |
| Range | | 9841000.00 |  |
| Interquartile Range | | 4160000.00 |  |
| Skewness | | 1.389 | .427 |
| Kurtosis | | 1.109 | .833 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tests of Normality** | | | | | | |
|  | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
| Statistic | Df | Sig. | Statistic | df | Sig. |
| income | .252 | 30 | .000 | .787 | 30 | .000 |
| a. Lilliefors Significance Correction | | | | | | |

Income variable, after removing outliers, has Skewness and Kurtosis of 3.25(1.389/0.427) and 1.33(1.109/0.833) respectively. The value of kurtosis lies in the range of +-1.96 but not skewness. The Sig. from Kolmogorov-Smirnov and Shapiro-Wik test is 0.000 which is less than 0.05. So, we reject Null hypothesis of normality.

Hence, we infer that income variable is not normally distributed even after we remove the outliers.

**5. Find out the number of males and females who are graduate and social drinkers.**

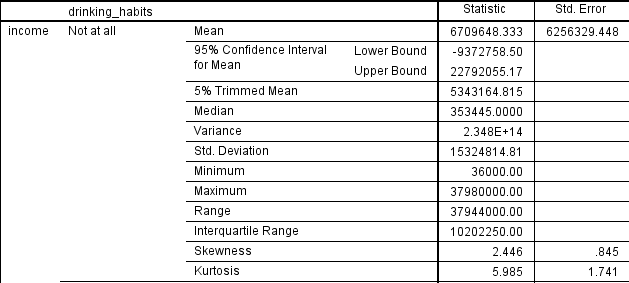




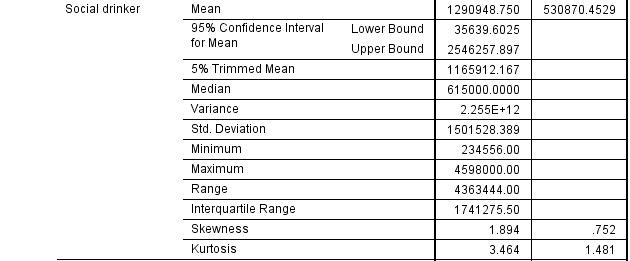
Number of males who are graduate and social drinkers = **0**

Number of females who are graduate and social drinkers = **1**

**6. What is 95% confidence interval of mean income for social drinkers and non-drinkers.**



95% Confidence Interval for Mean of income for **Non Drinkers** is =( **-9372758.50 , 22792055.17)**



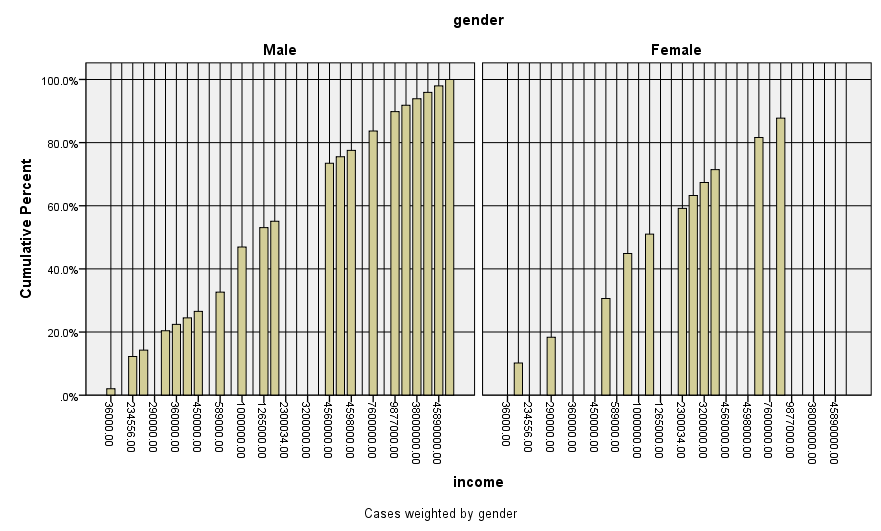
95% Confidence Interval for Mean of income for **Social drinkers** is = (**35639.60 , 2546257.90)**

**7. Find out the 5% trimmed mean for those males who are heavy drinker. Compare it with those females who are heavy drinker.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptives** | | | | | |
|  | Gender | | | Statistic | Std. Error |
| drinking\_habits = 6.00 (FILTER) | Male | Mean | | .05 | .048 |
| 95% Confidence Interval for Mean | Lower Bound | -.05 |  |
| Upper Bound | .15 |  |
| 5% Trimmed Mean | | .00 |  |
| Median | | .00 |  |
| Variance | | .048 |  |
| Std. Deviation | | .218 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 4.583 | .501 |
| Kurtosis | | 21.000 | .972 |
| Female | Mean | | .07 | .071 |
| 95% Confidence Interval for Mean | Lower Bound | -.08 |  |
| Upper Bound | .23 |  |
| 5% Trimmed Mean | | .02 |  |
| Median | | .00 |  |
| Variance | | .071 |  |
| Std. Deviation | | .267 |  |
| Minimum | | 0 |  |
| Maximum | | 1 |  |
| Range | | 1 |  |
| Interquartile Range | | 0 |  |
| Skewness | | 3.742 | .597 |
| Kurtosis | | 14.000 | 1.154 |

5% trimmed mean for female heavy drinkers is 0.02 and male heavy drinkers is 0.00. When compared we can say that the difference in trimmed mean is not very high. It might we because we have only 1 case of heavy drinker for each category.

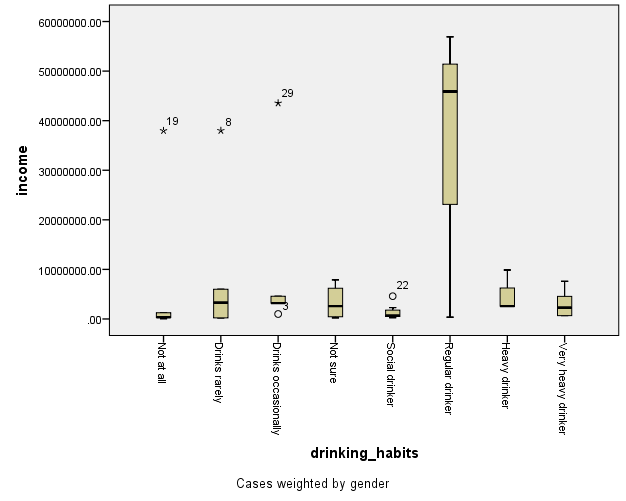
**8. What is the income below which 80% of females earn? Do 80% females earn a higher income than 80% males?**



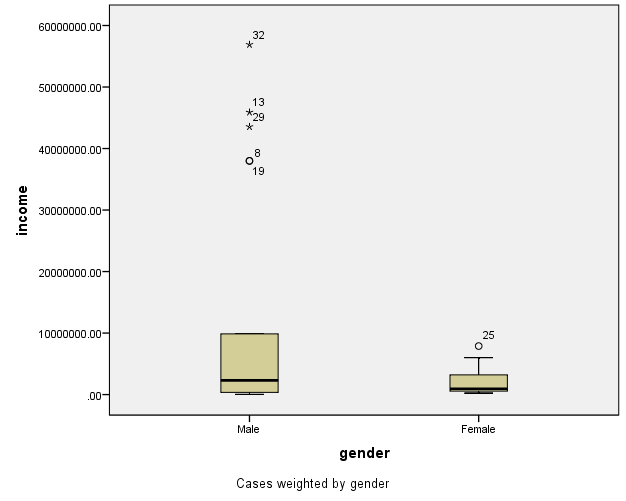
Income below which 80% of Females earn is between 38,80,000 to 60,99,000.

Income below which 80% of Males earn is between 45,98,000 to76,00,000.

So, we can say that 80% females do not earn a higher income than 80% males.

**9. Whose income is more stable: those who never drink or those who are heavy drinker? Also what is the scenario for male vs female? What insights you get from 95% confidence interval and measures of dispersion?**

Income of people who never drink is more stable than income of heavy drinkers if we ignore the outlier.



Income of Females is more stable than income of Males as the variability of box is small. We have many outliers for male income.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptives** | | | | | |
|  | Gender | | | Statistic | Std. Error |
| income | Male | Mean | | 12415064.3333 | 4116270.09283 |
| 95% Confidence Interval for Mean | Lower Bound | 3828675.3806 |  |
| Upper Bound | 21001453.2861 |  |
| 5% Trimmed Mean | | 10660006.5185 |  |
| Median | | 2300000.0000 |  |
| Variance | | 355817269020247.300 |  |
| Std. Deviation | | 18863119.28129 |  |
| Minimum | | 36000.00 |  |
| Maximum | | 56900000.00 |  |
| Range | | 56864000.00 |  |
| Interquartile Range | | 23565055.00 |  |
| Skewness | | 1.396 | .501 |
| Kurtosis | | .328 | .972 |
| Female | Mean | | 2218538.1429 | 643465.67987 |
| 95% Confidence Interval for Mean | Lower Bound | 828415.0566 |  |
| Upper Bound | 3608661.2292 |  |
| 5% Trimmed Mean | | 2013681.2698 |  |
| Median | | 957500.0000 |  |
| Variance | | 5796673136368.286 |  |
| Std. Deviation | | 2407628.11422 |  |
| Minimum | | 230000.00 |  |
| Maximum | | 7894500.00 |  |
| Range | | 7664500.00 |  |
| Interquartile Range | | 3032500.00 |  |
| Skewness | | 1.352 | .597 |
| Kurtosis | | 1.044 | 1.154 |

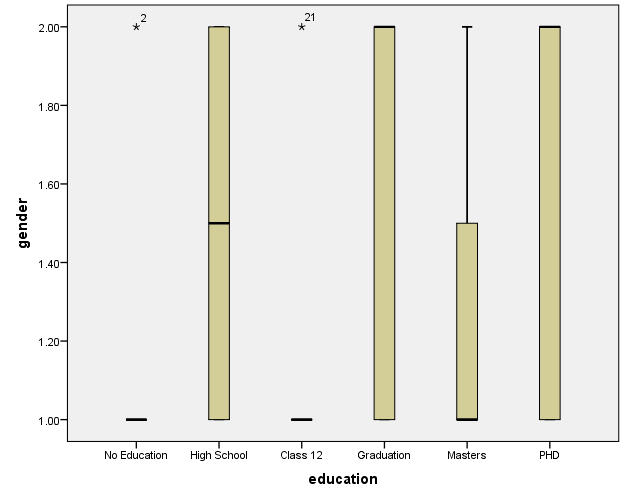
For male income mean is 12415064.33, median is 2300000.00, Std. Deviation is 18863119.28 and Range is 56864000.00.

For female income mean is 2218538.14, median is 957500.00, Std. Deviation is 2407628.11 and Range is 7664500.00.

The range of male income is more than that of female income. So, we can surely say that female income is stable than male income.

**10. Draw a boxplot to compare the normality in scores of people across educational category.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | | |
|  | education | Cases | | | | | |
|  | Valid | | Missing | | Total | |
|  | N | Percent | N | Percent | N | Percent |
| gender | No Education | 5 | 100.0% | 0 | 0.0% | 5 | 100.0% |
| High School | 8 | 100.0% | 0 | 0.0% | 8 | 100.0% |
| Class 12 | 7 | 100.0% | 0 | 0.0% | 7 | 100.0% |
| Graduation | 7 | 100.0% | 0 | 0.0% | 7 | 100.0% |
| Masters | 3 | 100.0% | 0 | 0.0% | 3 | 100.0% |
| PHD | 5 | 100.0% | 0 | 0.0% | 5 | 100.0% |



No Education category are all males except one female out of 5 cases.

Total Number of high school educated cases are 8 and it is normally distributed among both male and female.

Class 12 Education has 7 cases out of which all are males and 1 is female.

Graduation category is negatively skewed which means that we have more female graduates than male out of 7 total cases.

Masters category is positively skewed which means that we have more male Masters than female out of 3 cases.

PHD is again negatively skewed which means we have more female PHD cases than male out of 5 total cases.