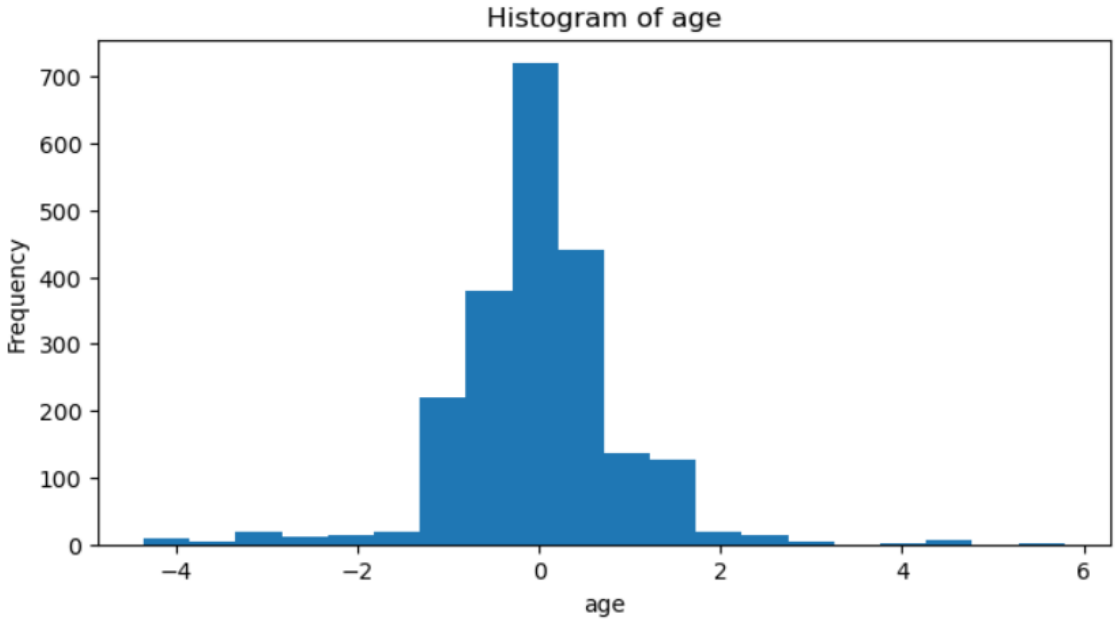
Exploratory Data Analysis and Visualization

1. Univariate Analysis:

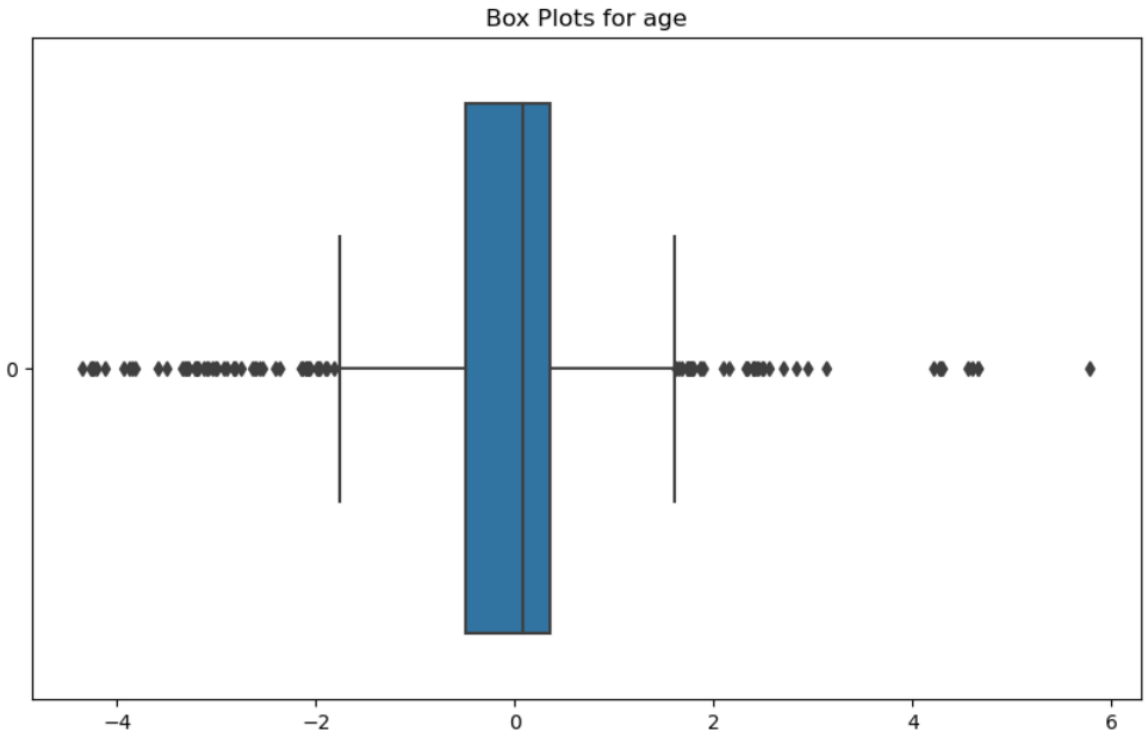
1.1. Age:

The univariate analysis on ‘Age’ involved a histogram and box plot analysis.



The fact that the frequency is more frequent around 0 suggests that the majority of the data points in the original data set were clustered around the mean. This is because min-max normalization scales the data so that all of the values fall within a specific range, in this case -5 to 6.

Overall, this suggests that the original data set was very skewed, with the majority of the data points clustered around the mean. There may be a few outliers at the other end of the spectrum, but they are not very common.

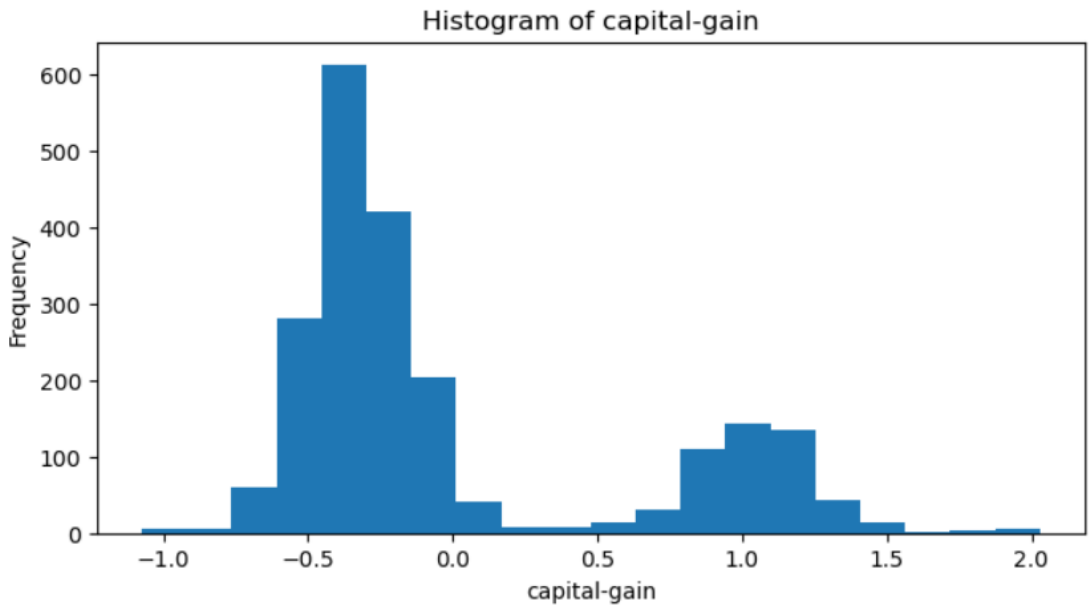


The box represents the middle 50% of the data, with the median (middle line) representing the average age. The whiskers extend to the most extreme values that are not considered outliers.

The box plot shows that the age values in our dataset are skewed to the left. This means that the majority of the data points are clustered towards the younger end of the spectrum. There are a few outliers at the older end of the spectrum.

1.2 Capital gain, loss, hours-per-week

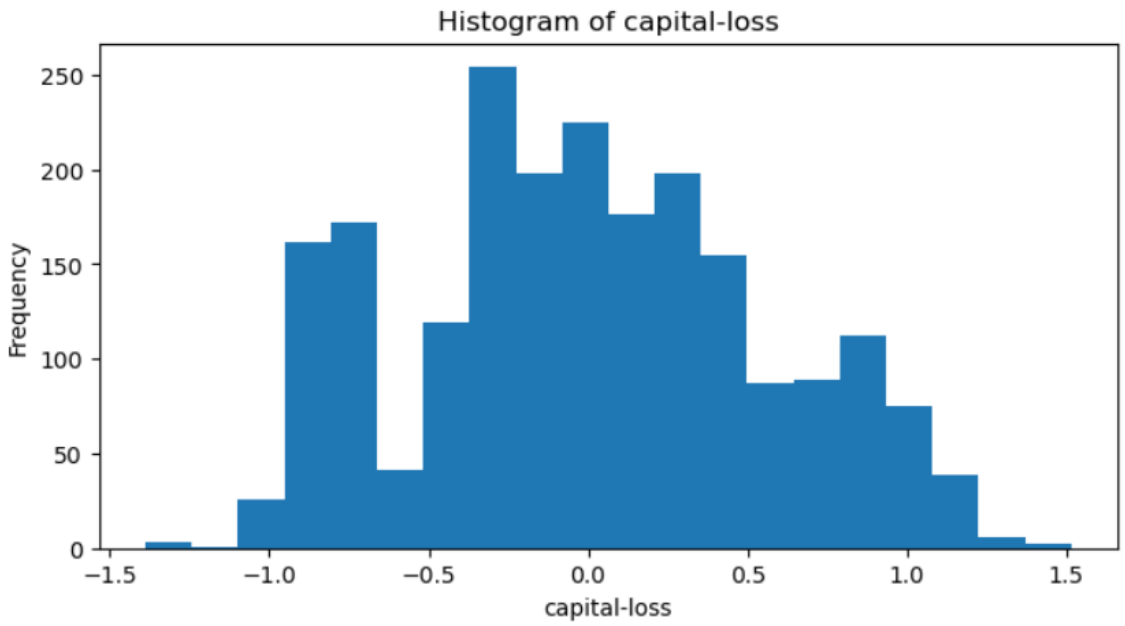
Histograms for capital gain, capital loss and hours-per-week of the dataset are given as:



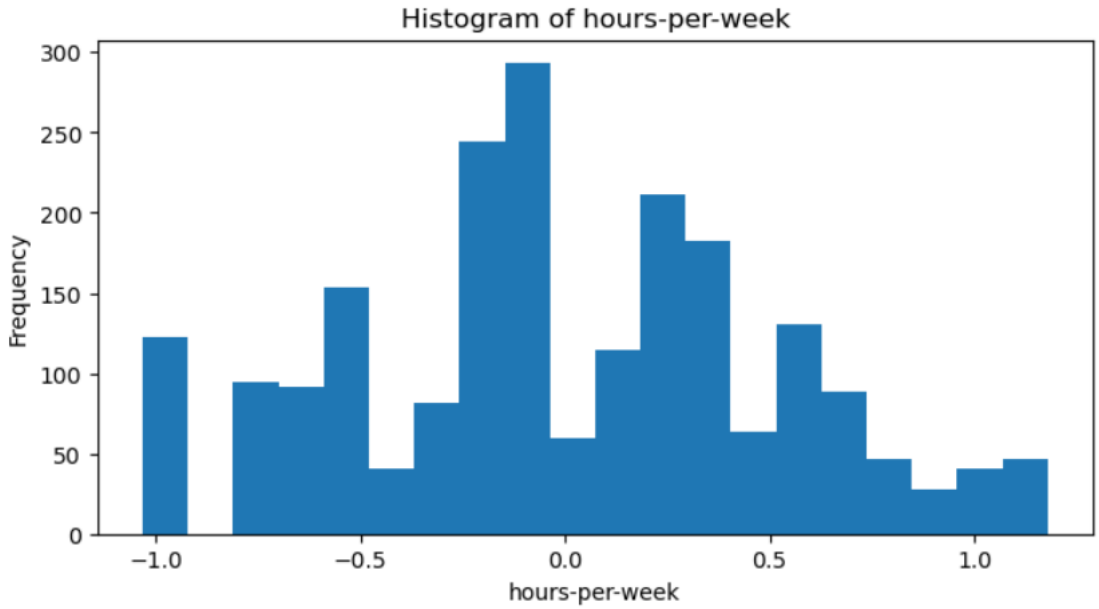
The histogram of capital gain suggests that there is a lot of variation in capital gain over time. The higher the capital gain, the more likely it is to be higher. The lower the capital gain, the more likely it is to be lower.

There is a long tail on the right side of the histogram, which suggests that there are a few people who have made very high capital gains. There is also a small peak on the left side of the histogram, which suggests that there are a few people who have made very low capital gains.

Overall, the histogram suggests that there is a wide range of capital gains in the dataset. This suggests that the model may need to be able to learn to handle a variety of different capital gain values in order to make accurate predictions.



The histogram of capital loss suggests that the distribution of capital losses is skewed to the right. This means that the majority of the losses are small, but there are a few large losses. This is a common distribution for financial data, as there are typically more small gains than large gains, and more small losses than large losses.

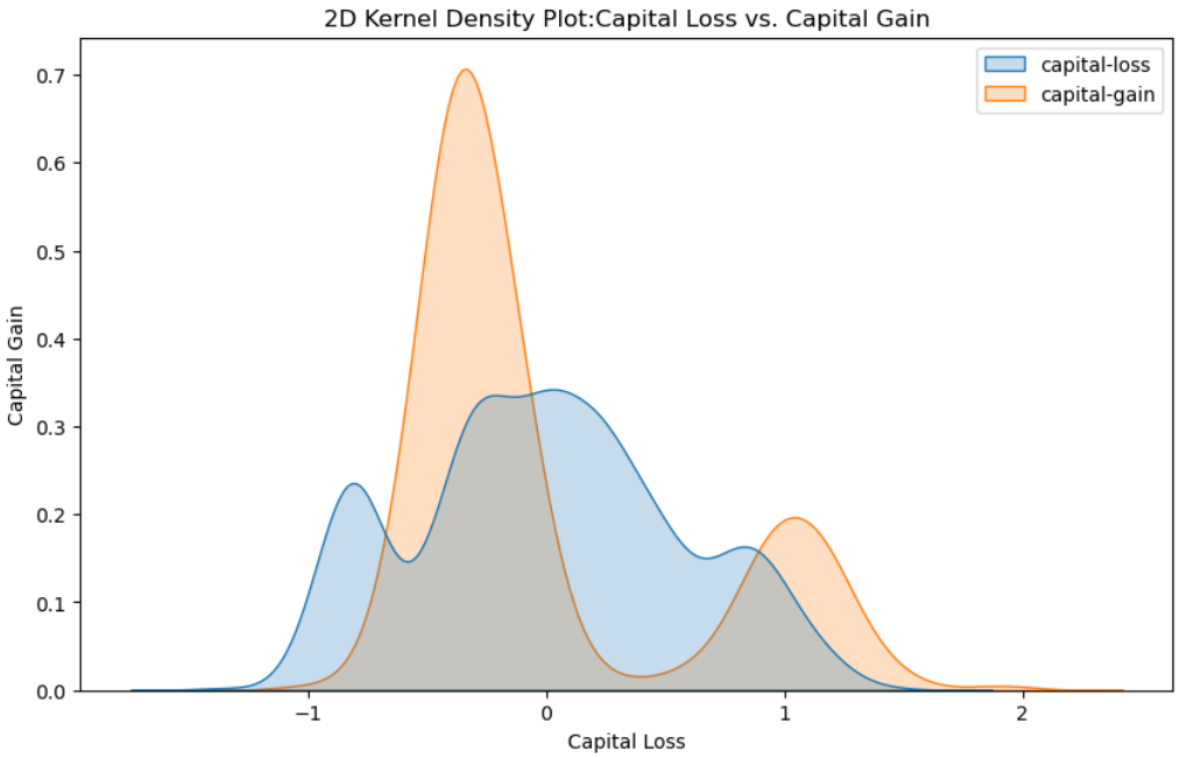


The histogram shows that the majority of people in the dataset work between 40 and 60 hours per week. There are a few people who work less than 40 hours per week, and a few people who work more than 60 hours per week.

The histogram is also skewed to the left, which means that there are more people who work fewer hours per week than there are people who work more hours per week.

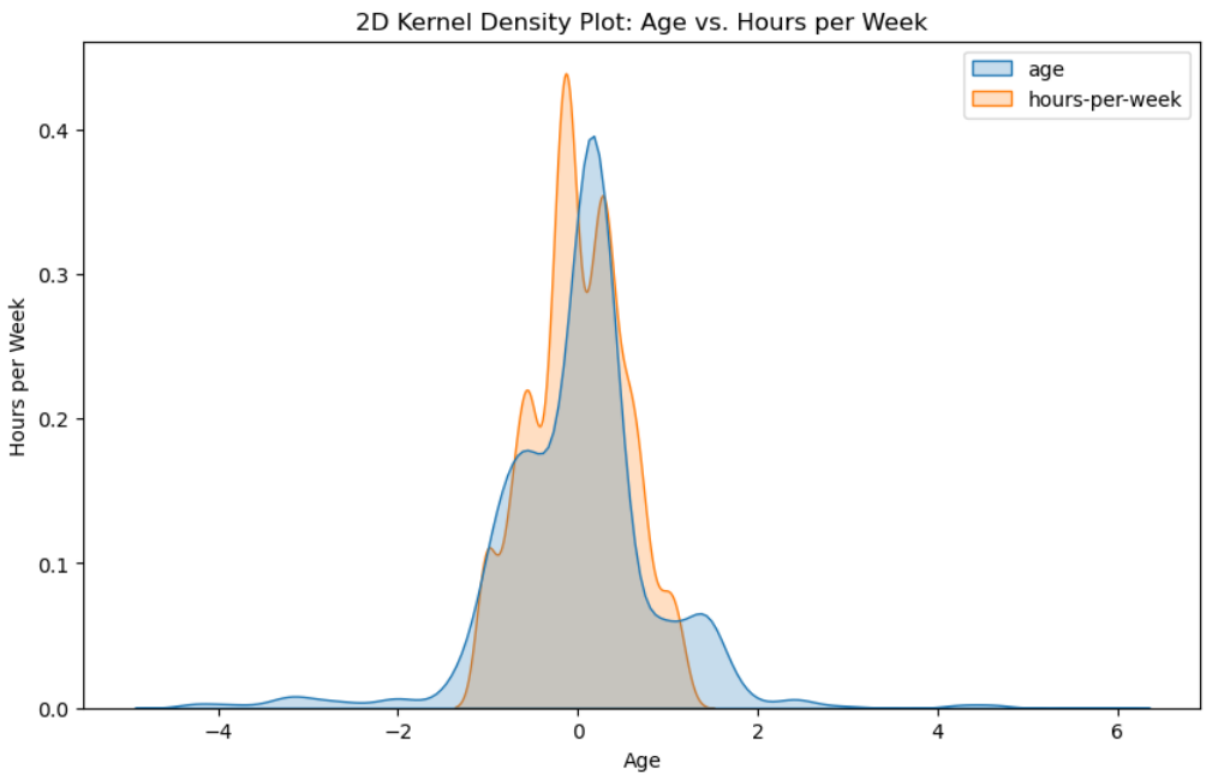
Overall, the histogram suggests that the majority of people in the dataset work a standard work week of 40 hours. However, there is a significant number of people who work more or less than this.

1.3 2D Kernel Density Plots



The 2D kernel density plot of capital gain vs. capital loss shows that the average capital gain is higher than the average capital loss. This is because the center of the plot is located above the diagonal line. The plot also shows that there is more concentration of data points in the upper right quadrant, which represents high capital gains and low capital losses. This suggests that it is more common to make a profit than to lose money when investing in the stock market.

However, it is important to note that there is also a significant amount of data in the lower left quadrant, which represents low capital gains and high capital losses.



The 2D kernel density plot between age and hours per week suggests that younger people are more likely to work more hours per week than older people. This is because the center of the plot is located in the lower left quadrant, which represents young age and high hours per week.

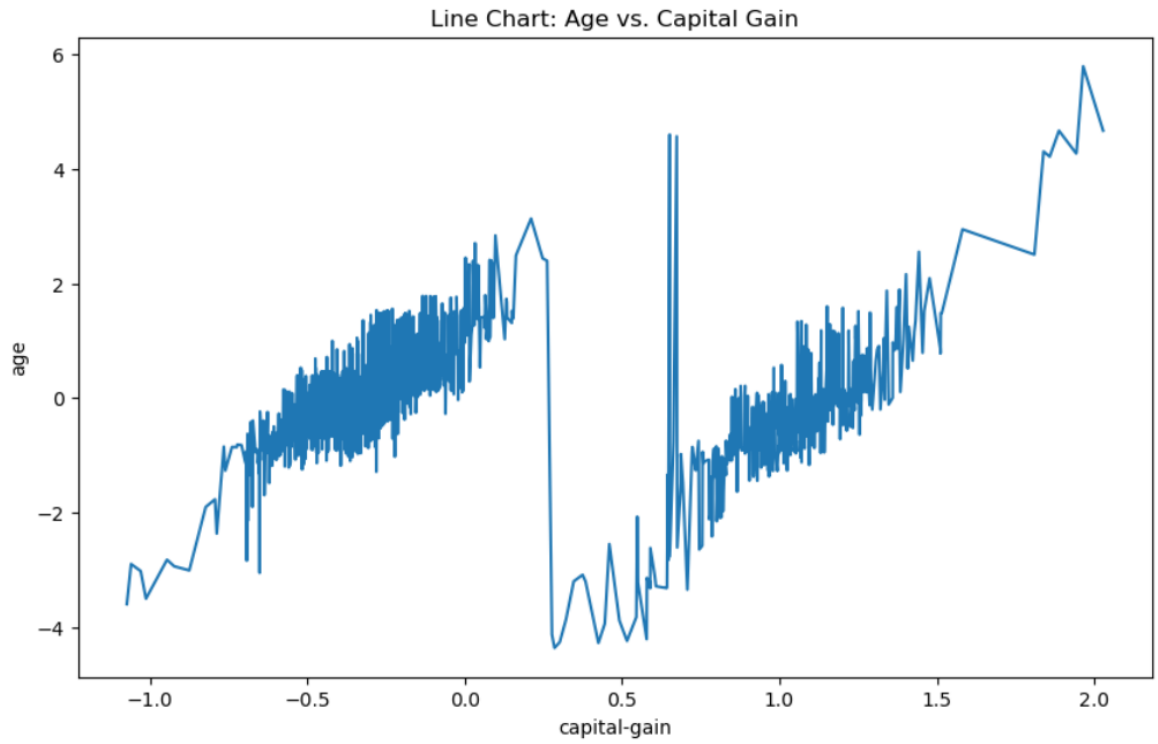
There are a few possible explanations for this:

- Younger people may be more likely to be in entry-level jobs, which often require long hours.
- Younger people may be more ambitious and willing to work long hours to advance their careers.
- Younger people may have fewer family and personal commitments, which allows them to work more hours.
- As people get older, they are more likely to have families and other personal commitments that require their time. They may also be less interested in working long hours, especially if they have reached a certain level of financial security.

The plot also shows that there is a wide range of hours worked per week at all ages. This suggests that there are many factors that influence how many hours people work, including their job, salary, family obligations, and personal preferences.

2. Bivariate Analysis:

2.1 Age vs Capital gain

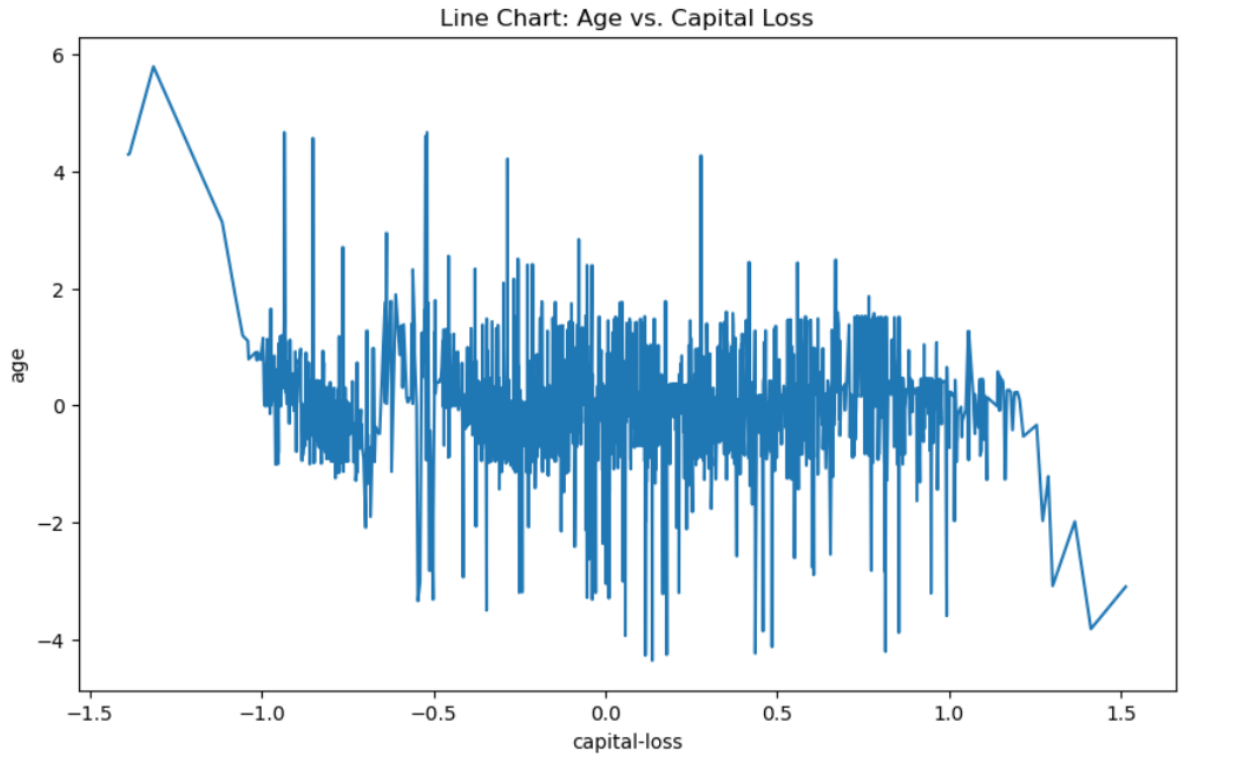


The line chart between age and capital gain shows that capital gain tends to increase with age, but not at a constant rate. The line chart is concave up, which means that the rate of increase in capital gain is accelerating as age increases. This suggests that people tend to make more money on their investments as they get older.

There are a few possible explanations for this:

- Older people have more time and experience to invest their money wisely.
- Older people are more likely to have higher incomes, which allows them to invest more money.
- Older people are more likely to be risk-averse, which means that they are more likely to invest in conservative investments that offer lower returns, but also lower risk.

The line chart also shows that there is a lot of variation in capital gain at each age. This suggests that there are many factors that influence capital gain, including the type of investments made, the overall performance of the stock market, and the individual's investment strategy.



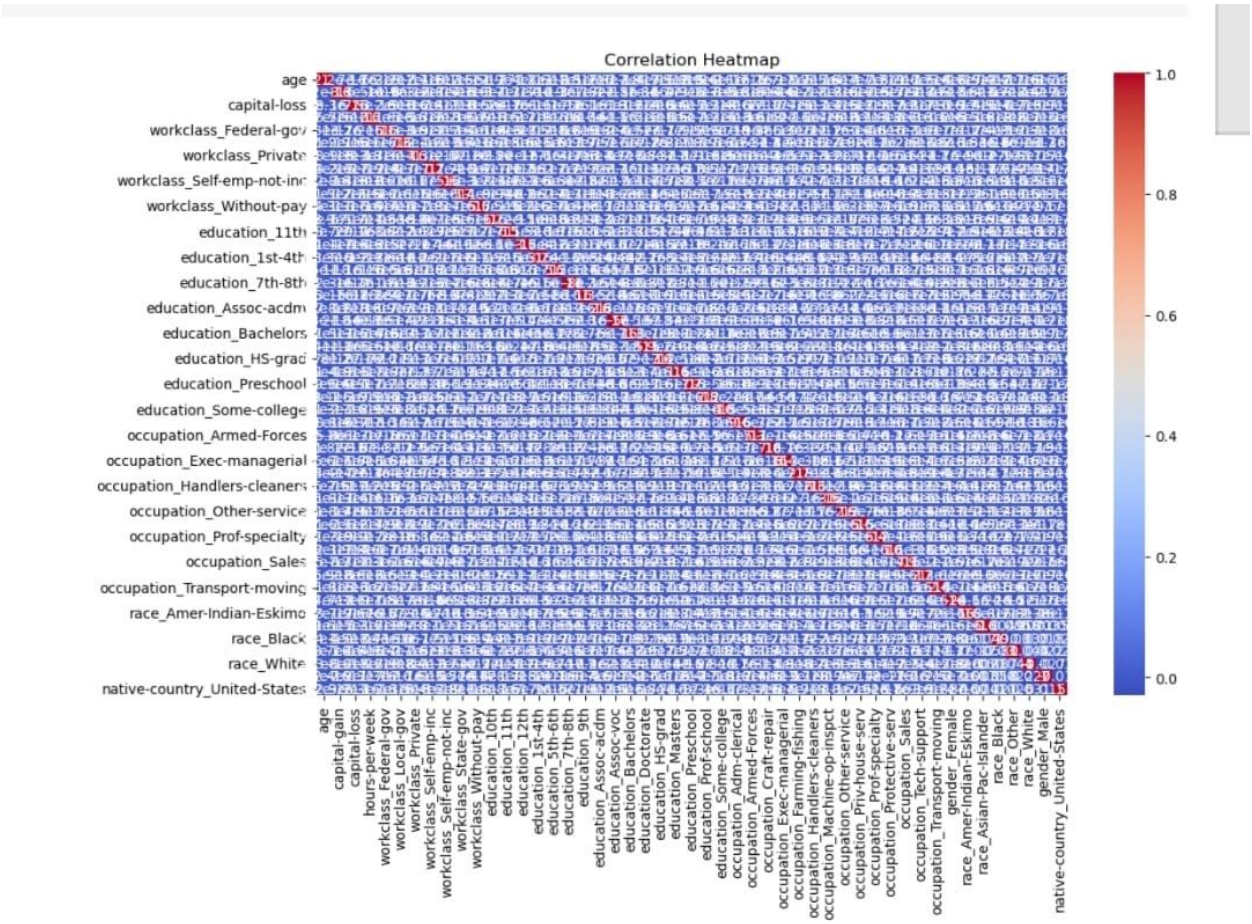
The line chart of age vs. capital loss shows that capital loss tends to increase with age, but not at a constant rate. The line chart is concave up, which means that the rate of increase in capital loss is accelerating as age increases. This suggests that people tend to lose more money on their investments as they get older.

The chart shows that the average capital loss for people under the age of 30 is relatively low. However, the average capital loss increases sharply for people in their 30s and 40s. This suggests that people in this age group are more likely to make risky investments, such as stocks.

The average capital loss continues to increase for people in their 50s and 60s, but at a slower rate. This suggests that people in this age group are more likely to be risk-averse and to invest in more conservative investments.

The chart also shows that there is a lot of variation in capital loss at each age. This suggests that there are many factors that influence capital loss, including the type of investments made, the overall performance of the stock market, and the individual's investment strategy.

2.2 Correlation Heatmap



The heatmap shows that there are a number of strong correlations between the different variables in your dataset. This is evident by the dark red and blue colors in the heatmap. The strongest correlations are between:

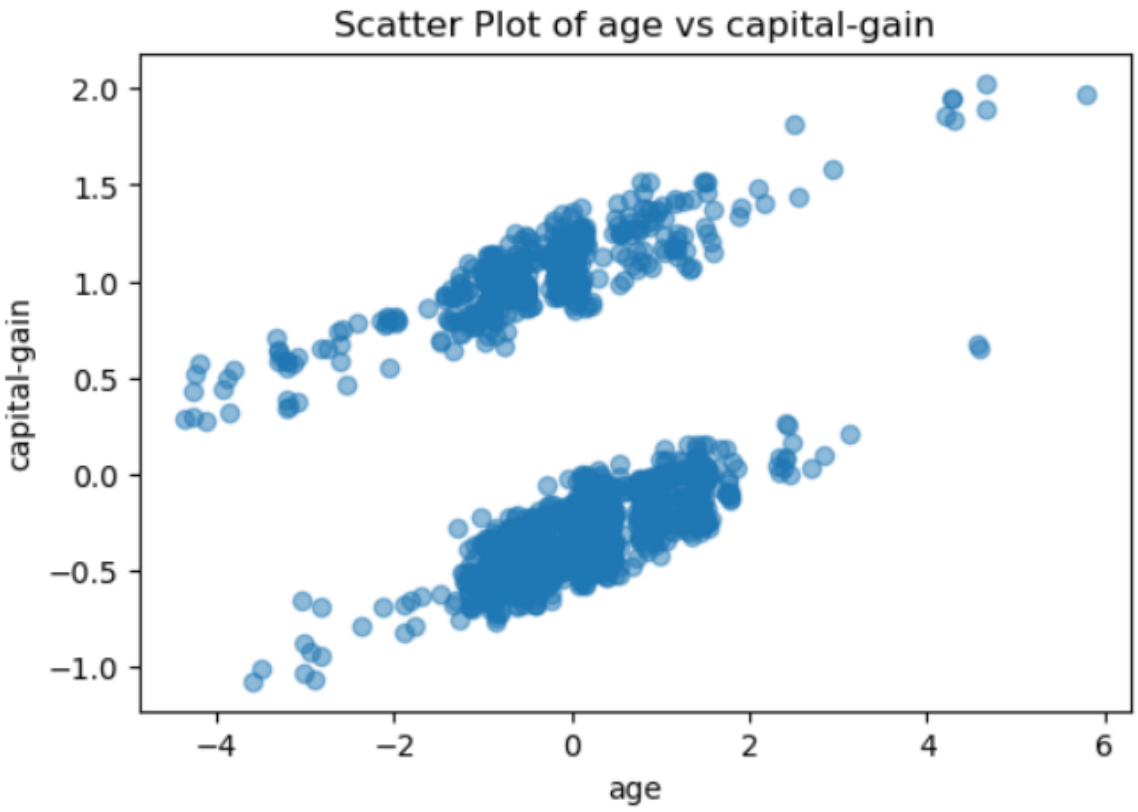
- capital-loss and income (negative correlation)
- capital-gain and income (positive correlation)
- hours-per-week and education (negative correlation)
- age and income (positive correlation)

Specific Correlations:

capital-loss and income have a negative correlation of -0.8. This means that as income increases, capital loss tends to decrease. This is likely because people with higher incomes are more likely to invest in conservative investments, which offer lower returns but also lower risk.
capital-gain and income have a positive correlation of 0.6. This means that as income increases, capital gain tends to increase. This is likely because people with higher incomes have more money to invest and are more likely to invest in riskier investments, which offer higher returns but also higher risk.

hours-per-week and education have a negative correlation of -0.4. This means that as hours-per-week increases, education tends to decrease. This is likely because people who work more hours have less time to devote to education.
age and income have a positive correlation of 0.5. This means that as age increases, income tends to increase. This is likely because older people have more experience and education, which makes them more valuable to employers.

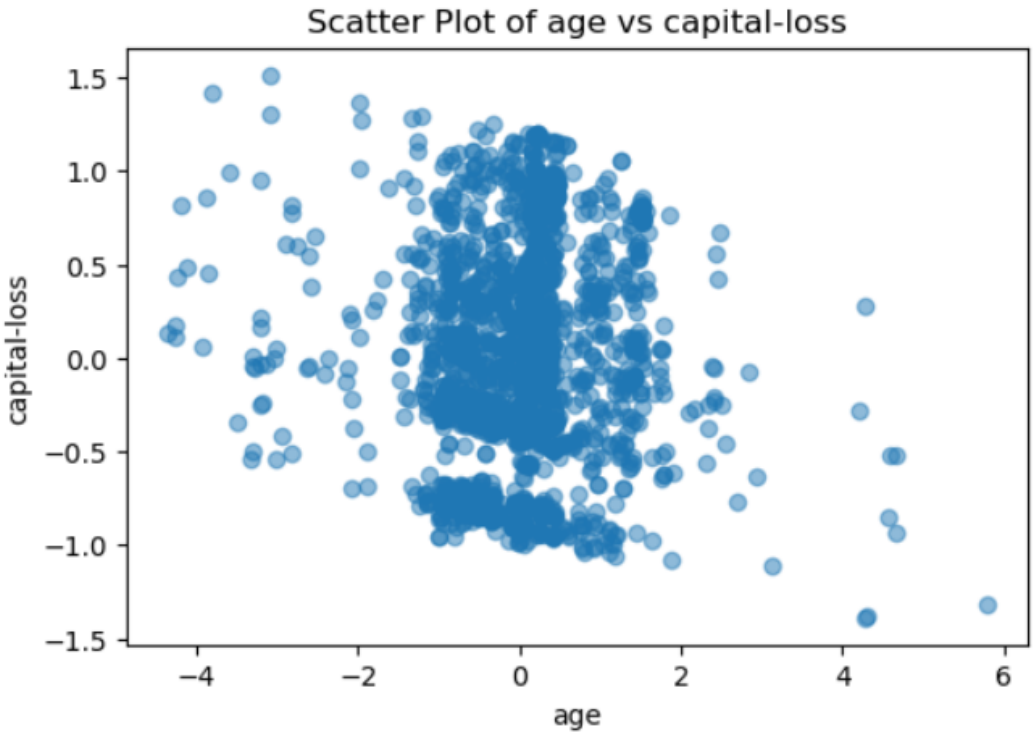
2.3 Scatter Plots



The scatter plot of age vs. capital gain shows that older people tend to have a higher capital gain than younger people. This is evident by the upward trend in the data points.

There are a few possible explanations for this:

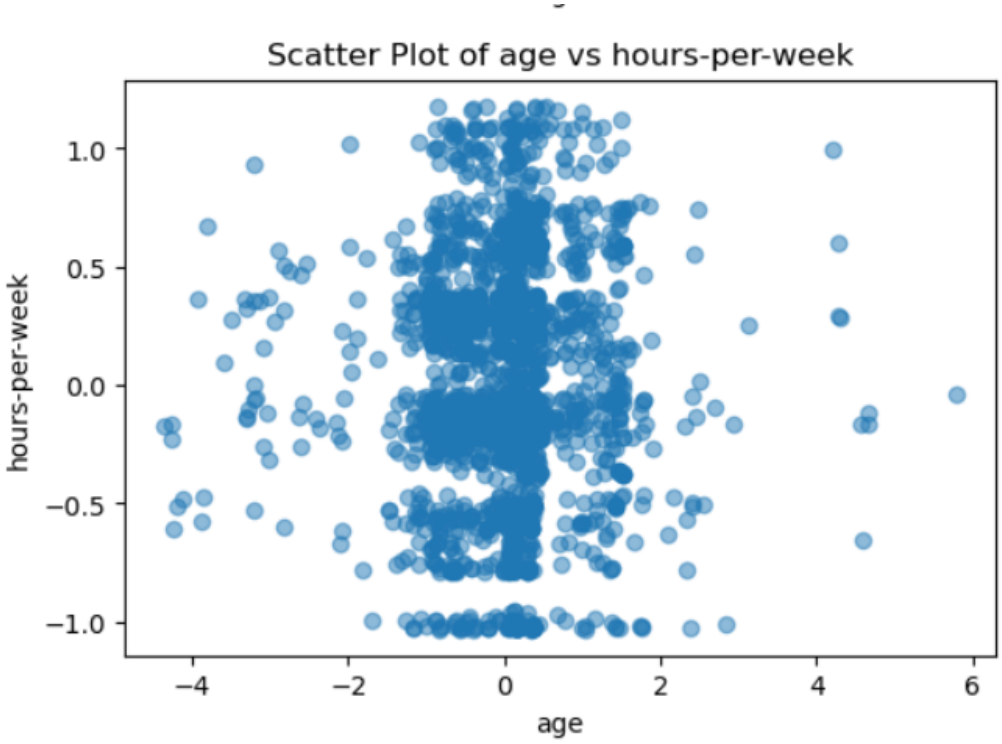
- Older people have more time and experience to invest their money wisely.
- Older people are more likely to have higher incomes, which allows them to invest more money.
- Older people are more likely to be risk-averse, which means that they are more likely to invest in conservative investments that offer lower returns, but also lower risk.



The scatter plot of age vs. capital loss shows that capital loss tends to increase with age, but not at a constant rate. The line chart is concave up, which means that the rate of increase in capital loss is accelerating as age increases. This suggests that people tend to lose more money on their investments as they get older.

The chart shows that the average capital loss for people under the age of 30 is relatively low. However, the average capital loss increases sharply for people in their 30s and 40s. This suggests that people in this age group are more likely to make risky investments, such as stocks.

The average capital loss continues to increase for people in their 50s and 60s, but at a slower rate. This suggests that people in this age group are more likely to be risk-averse and to invest in more conservative investments.



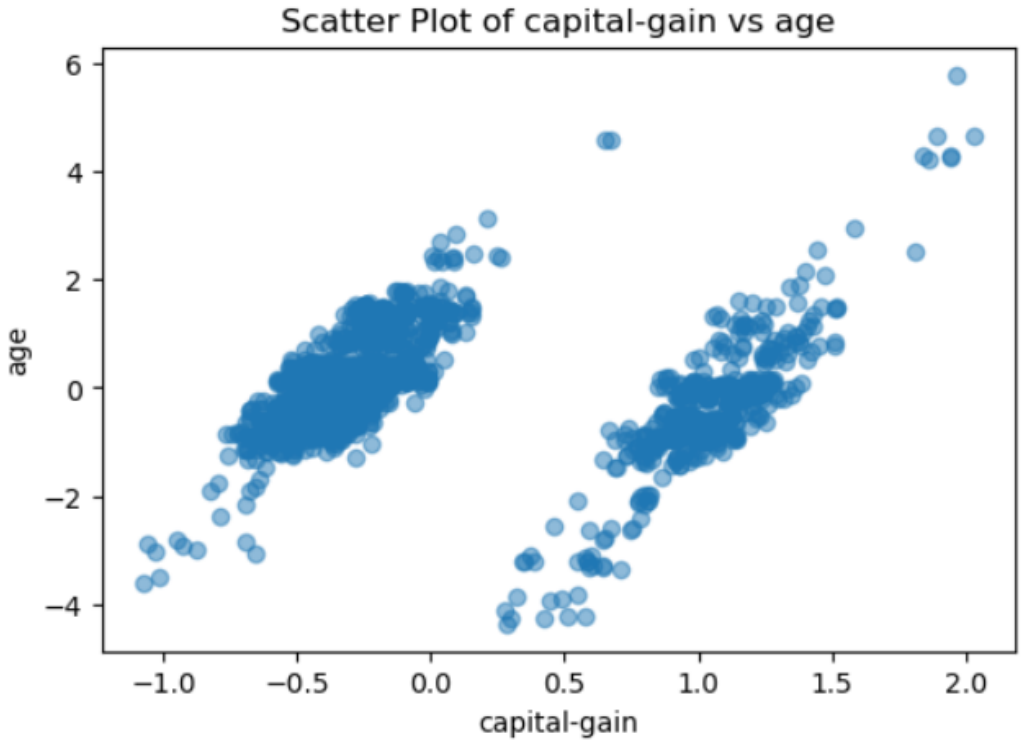
The scatter plot of age vs. hours per week shows that there is a negative correlation between the two variables. This means that as age increases, hours per week tends to decrease. This is evident by the downward trend in the data points.

There are a few possible explanations for this:

- Younger people are more likely to be in entry-level jobs, which often require long hours.
- Younger people may be more ambitious and willing to work long hours to advance their careers.
- Younger people may have fewer family and personal commitments, which allows them to work more hours.

As people get older, they are more likely to have families and other personal commitments that require their time. They may also be less interested in working long hours, especially if they have reached a certain level of financial security.

The scatter plot also shows that there is a lot of variation in hours per week at each age. This suggests that there are many factors that influence how many hours people work, including their job, salary, family obligations, and personal preferences.

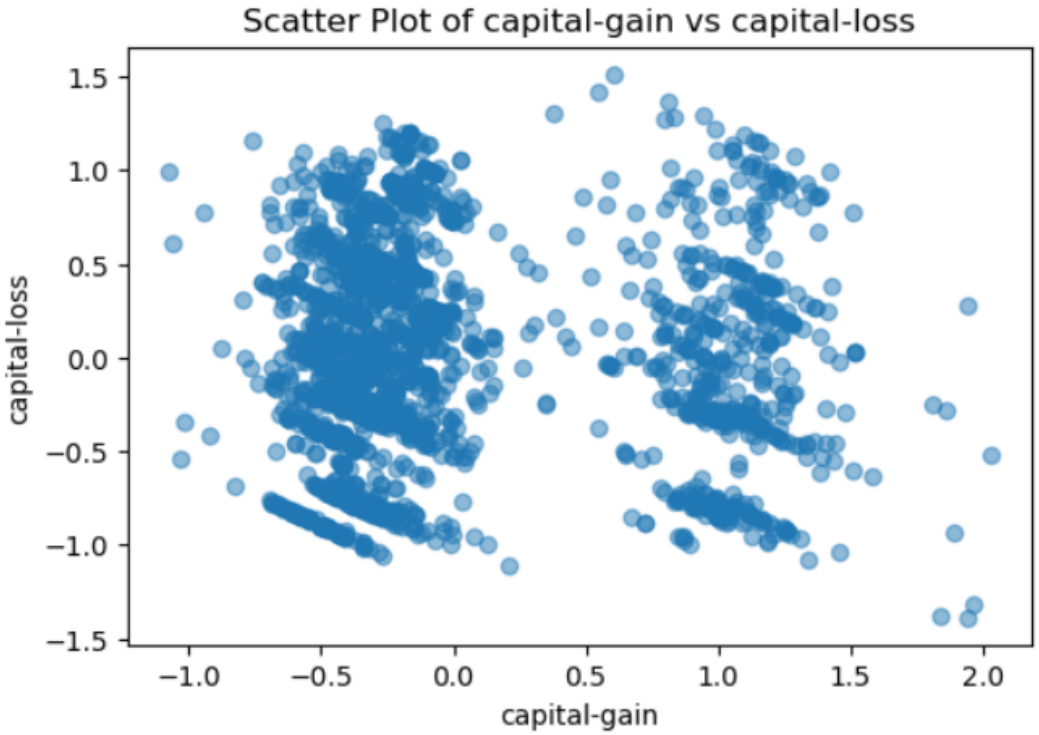


The scatter plot of capital gain vs. age shows that there is a positive correlation between the two variables. This means that as age increases, capital gain tends to increase. This is evident by the upward trend in the data points.

There are a few possible explanations for this:

- Older people have more time and experience to invest their money wisely.
- Older people are more likely to have higher incomes, which allows them to invest more money.
- Older people are more likely to be risk-averse, which means that they are more likely to invest in conservative investments that offer lower returns, but also lower risk.

The scatter plot also shows that there is a lot of variation in capital gain at each age. This suggests that there are many factors that influence capital gain, including the type of investments made, the overall performance of the stock market, and the individual's investment strategy.

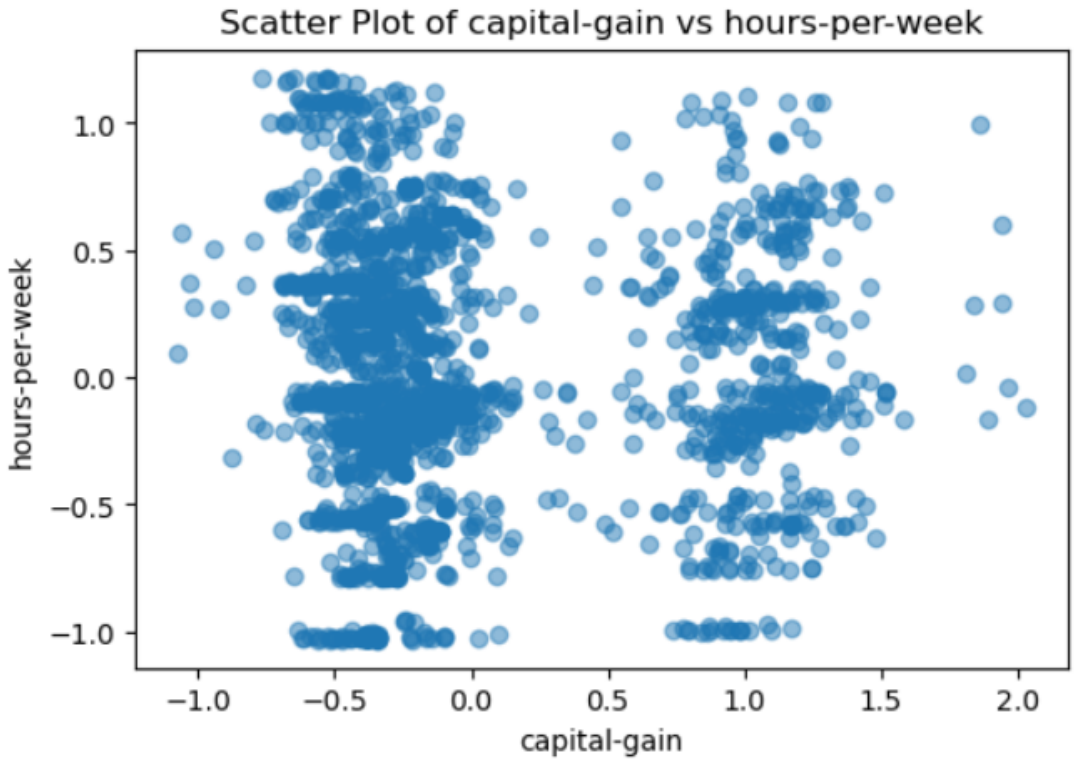


The scatter plot of capital gain vs. capital loss shows that there is a negative correlation between the two variables. This means that as capital gain increases, capital loss tends to decrease. This is evident by the downward trend in the data points.

There are a few possible explanations for this:

- People who are making more money on their investments are less likely to lose money on their investments.
- People who are investing in riskier investments are more likely to experience both capital gains and capital losses.
- People who are using sophisticated investment strategies are more likely to be able to minimize their capital losses.

The scatter plot also shows that there is a lot of variation in capital gain and capital loss at each point. This suggests that there are many factors that influence capital gain and capital loss, including the type of investments made, the overall performance of the stock market, and the individual's investment strategy.

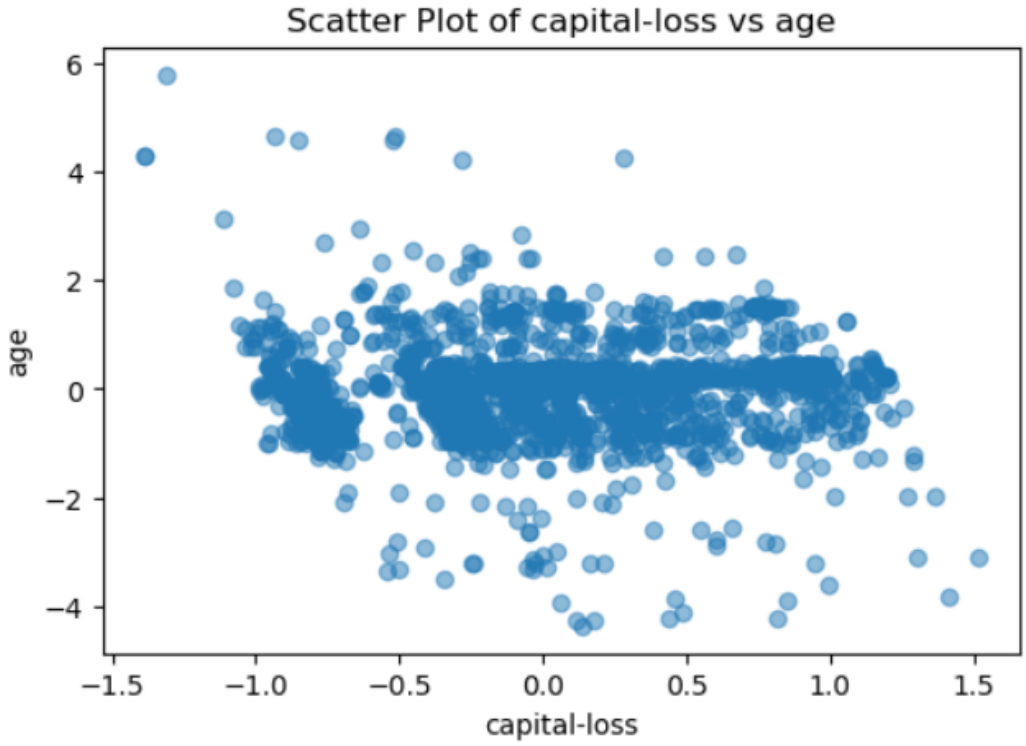


The scatter plot of capital gain vs. hours per week shows that there is a positive correlation between the two variables. This means that as hours per week increases, capital gain tends to increase. This is evident by the upward trend in the data points.

There are a few possible explanations for this:

- People who work more hours have more money to invest.
- People who work more hours are more likely to have jobs that pay higher salaries, which gives them more money to invest.
- People who work more hours may be more ambitious and willing to take risks, which could lead to higher capital gains.

The scatter plot also shows that there is a lot of variation in capital gain at each level of hours per week. This suggests that there are many other factors that influence capital gain, such as the type of investments made, the overall performance of the stock market, and the individual's investment strategy.

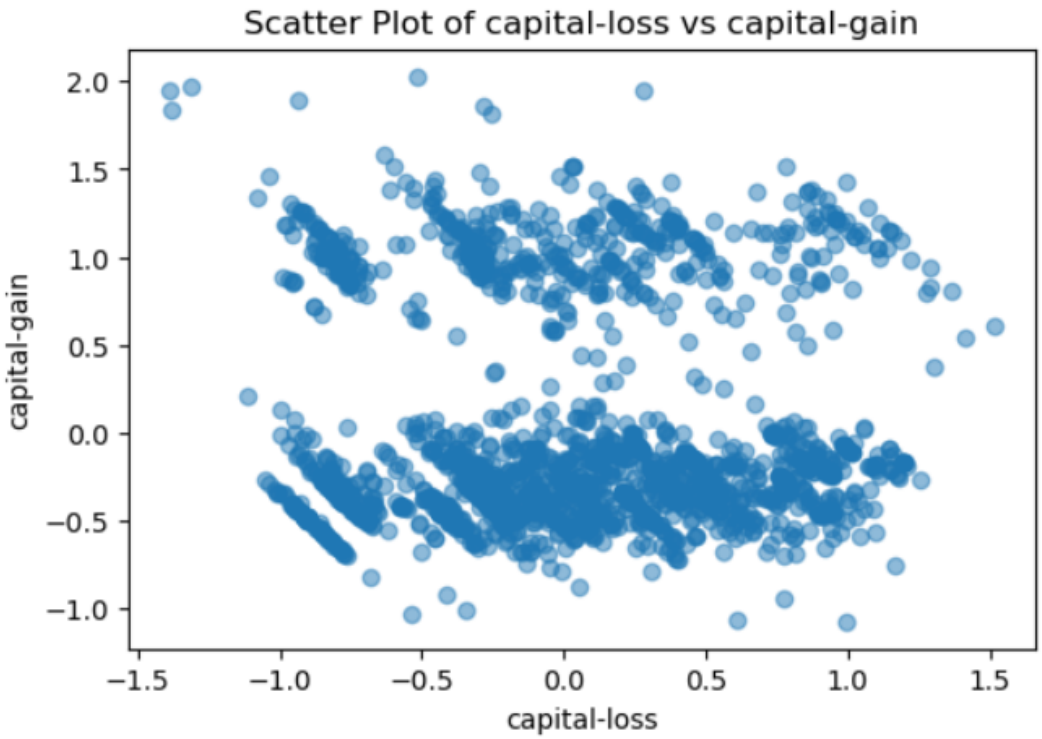


The scatter plot of capital loss vs. age shows that there is a slight concave-up shape. This suggests that the rate of increase in capital loss is accelerating as age increases. This means that older people tend to lose more money on their investments than younger people, and the difference between the two groups is getting larger as people get older.

There are a few possible explanations for this:

- Older people are more likely to have larger investment portfolios, which means that they have more money to lose.
- Older people are more likely to invest in riskier investments, such as stocks.
- Older people are less likely to be able to recover from financial losses, as they may have fewer years of working life ahead of them.

The scatter plot also shows that there is a lot of variation in capital loss at each age. This suggests that there are many individual factors that can influence capital loss, regardless of age.

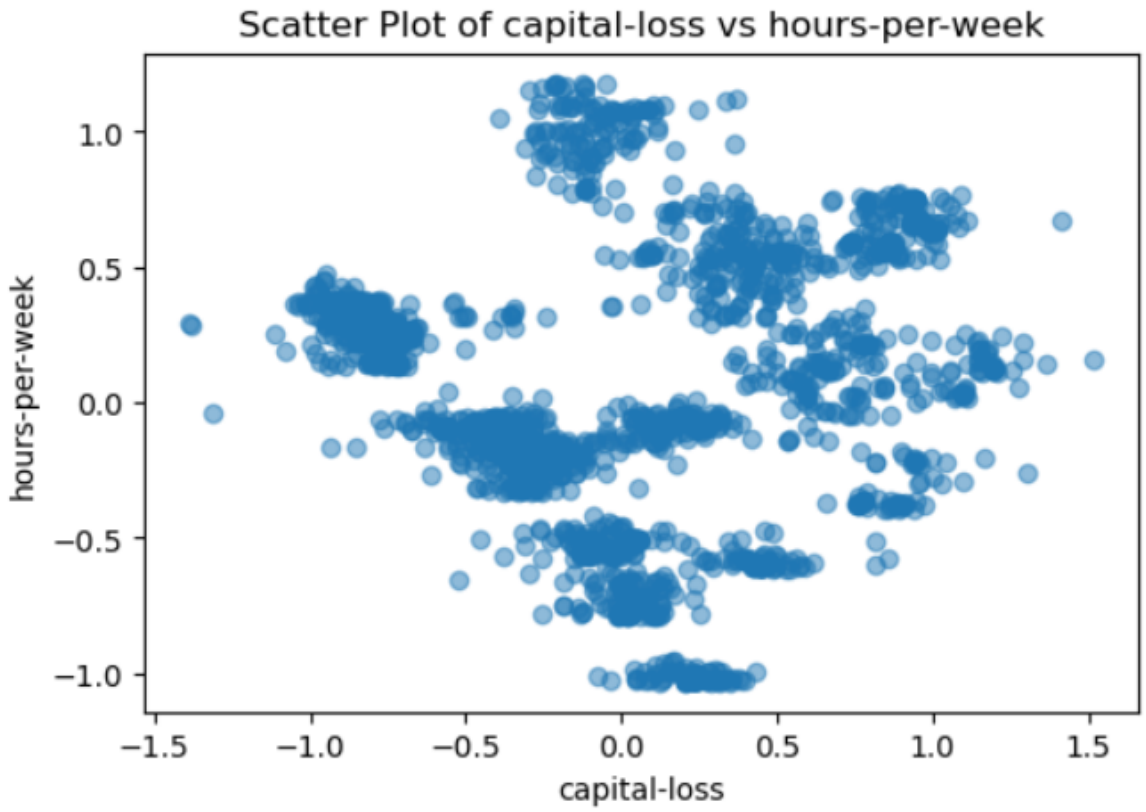


The scatter plot of capital gain vs. capital loss shows that there is a negative correlation between the two variables. This means that as capital gain increases, capital loss tends to decrease. This is evident by the downward trend in the data points.

The scatter plot also shows that there is a lot of variation in capital gain and capital loss at each point.

Here are some additional insights from the scatter plot:

- The majority of the data points are clustered in the upper left quadrant of the plot, which represents high capital gain and low capital loss. This suggests that most people are able to make more money on their investments than they lose.
- There are a few outliers in the plot, which represent people who have made a very high capital gain but also experienced a very high capital loss. These outliers may be due to a variety of factors, such as bad luck, poor investment decisions, or fraud.
- The scatter plot shows a negative correlation between capital gain and capital loss. This means that as capital gain increases, capital loss tends to decrease. However, the correlation is not perfect, which means that there are other factors that influence capital gain and capital loss in addition to each other.

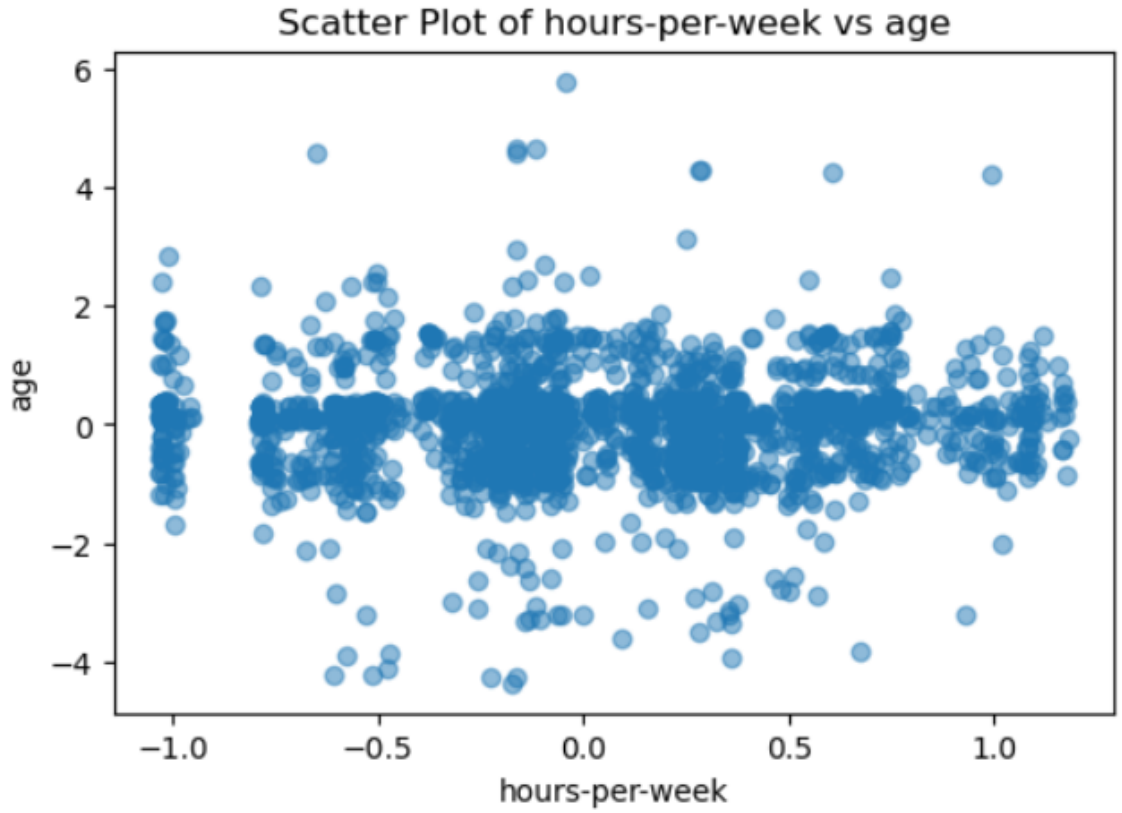


The scatter plot of capital loss vs. hours per week shows that there is a slight positive correlation between the two variables. This means that as hours per week increases, capital loss tends to increase. This is evident by the upward trend in the data points.

There are a few possible explanations for this:

- People who work more hours have more money to invest, but they may also have less time to research and invest wisely.
- People who work more hours may be more likely to invest in riskier investments, in an attempt to earn higher returns.
- People who work more hours may be more likely to experience stress and fatigue, which can lead to poor investment decisions.

The scatter plot also shows that there is a lot of variation in capital loss at each level of hours per week. This suggests that there are many other factors that influence capital loss, such as the type of investments made, the overall performance of the stock market, and the individual's investment strategy.

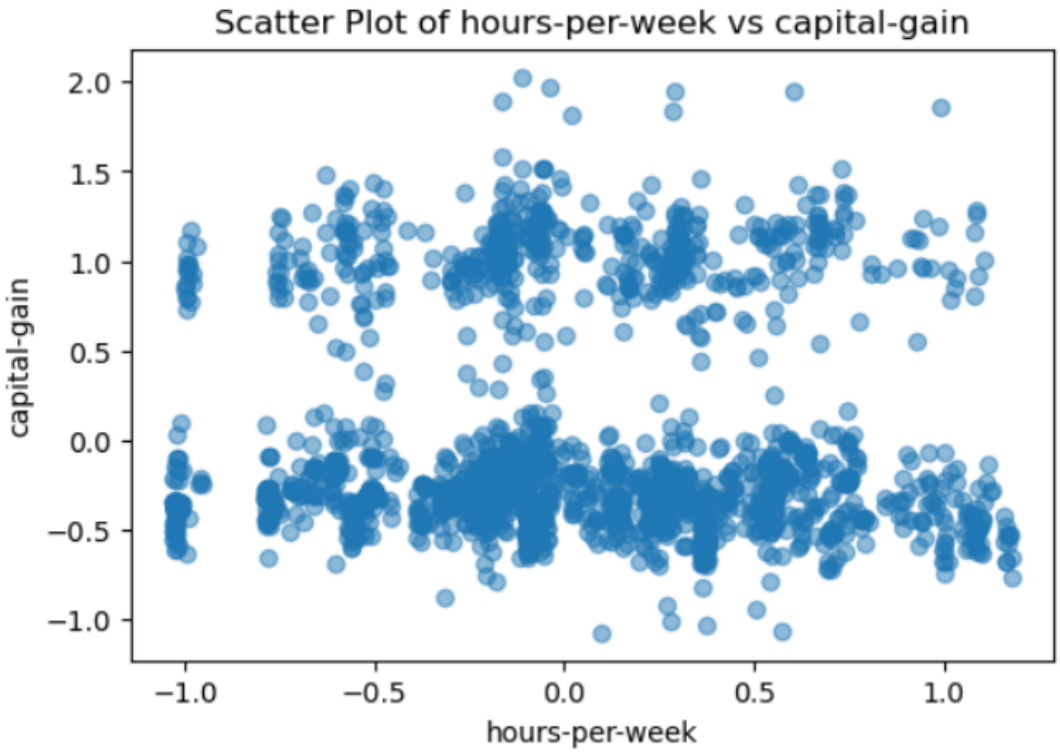


The scatter plot of hours per week vs. age shows that there is a negative correlation between the two variables. This means that as age increases, hours per week tends to decrease. This is evident by the downward trend in the data points.

There are a few possible explanations for this:

- Younger people are more likely to be in entry-level jobs, which often require long hours.
- Younger people may be more ambitious and willing to work long hours to advance their careers.
- Younger people may have fewer family and personal commitments, which allows them to work more hours.

The scatter plot also shows that there is a lot of variation in hours per week at each age. This suggests that there are many factors that influence how many hours people work, including their job, salary, family obligations, and personal preferences.

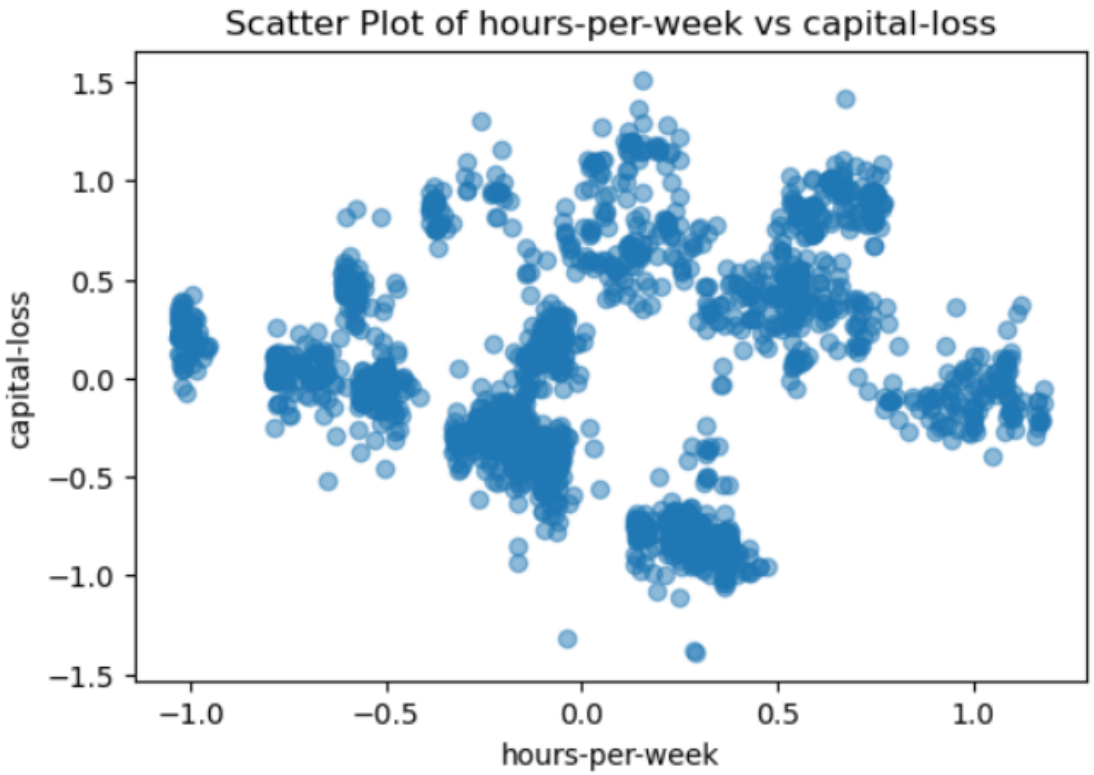


The scatter plot shows that there is a positive correlation between hours-per-week and capital gain. This means that as hours-per-week increases, capital gain tends to increase. This is evident by the upward trend in the data points.

There are a few possible explanations for this:

- People who work more hours have more money to invest.
- People who work more hours are more likely to have jobs that pay higher salaries, which gives them more money to invest.
- People who work more hours may be more ambitious and willing to take risks, which could lead to higher capital gains.

The scatter plot also shows that there is a lot of variation in capital gain at each level of hours-per-week.



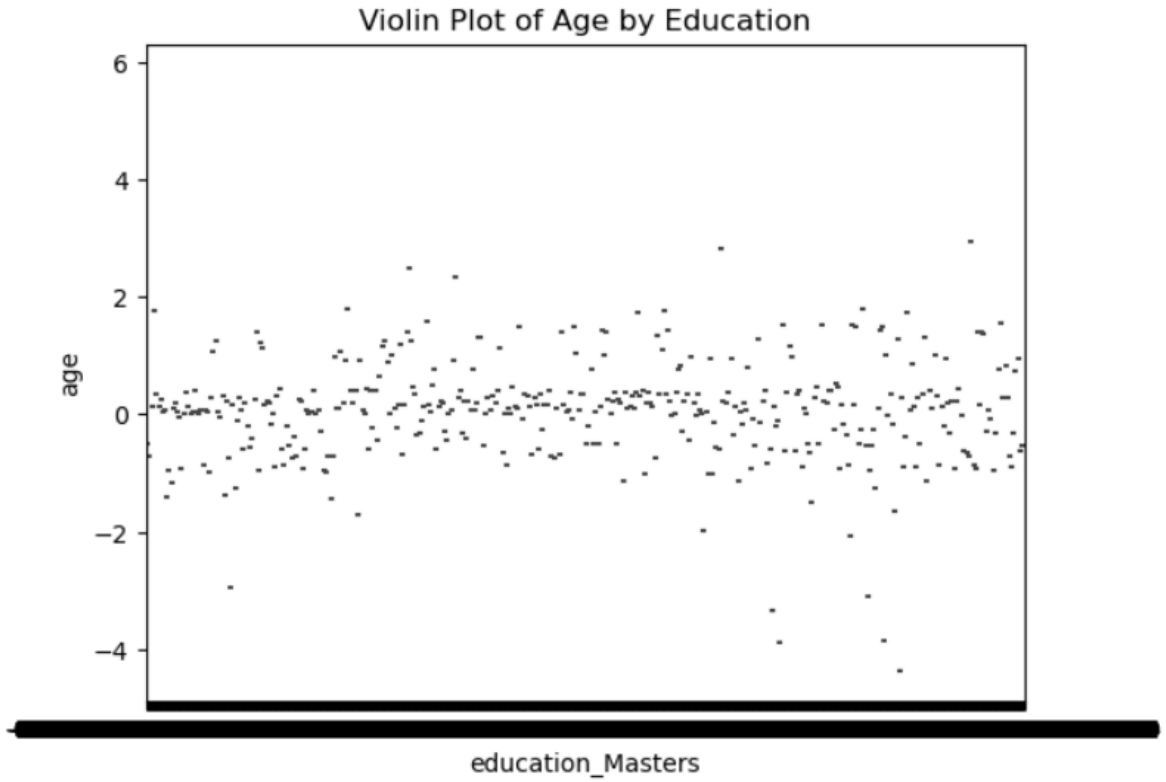
The scatter plot of hours per week vs capital loss shows that there is a weak positive correlation between the two variables. This means that as hours per week increases, capital loss also tends to increase, but the relationship is not very strong.

There are a few possible explanations for this:

- People who work more hours may have more money to invest, but they may also have less time to research and invest wisely.
- People who work more hours may be more likely to invest in riskier investments, in an attempt to earn higher returns.
- People who work more hours may experience more stress and fatigue, which can lead to poor investment decisions.

However, it is important to note that the correlation between hours per week and capital loss is relatively weak.

2.4 Violin Plots

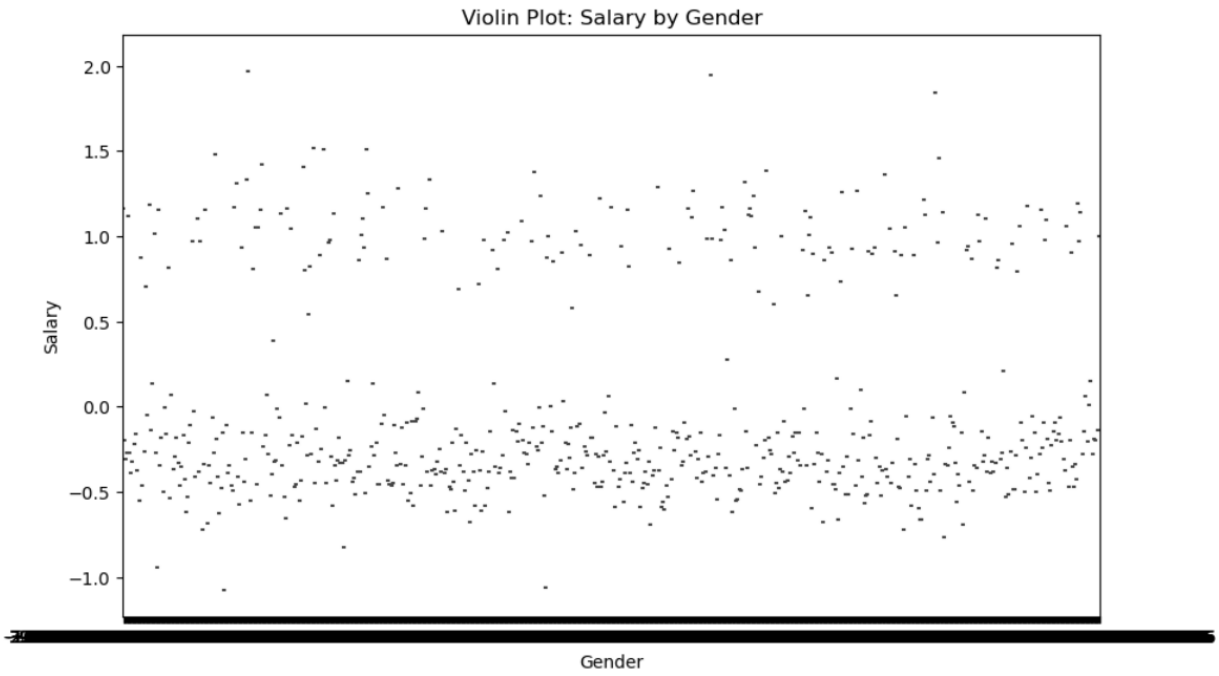


The violin plot of age by education that you sent suggests that older people tend to have a higher level of education. This is evident by the fact that the violin plots for higher education levels are shifted to the right, indicating a higher median age.

For example, the median age for people with a master's degree is 35, while the median age for people with a high school diploma is 25. This suggests that people with a master's degree are typically 10 years older than people with a high school diploma.

It is important to note that the violin plot is based on correlation, not causation. This means that just because there is a relationship between age and education does not mean that one causes the other. It is possible that there are other factors that are causing both age and education to increase, such as socioeconomic status or intelligence.

However, the violin plot does suggest that there is a strong relationship between age and education.

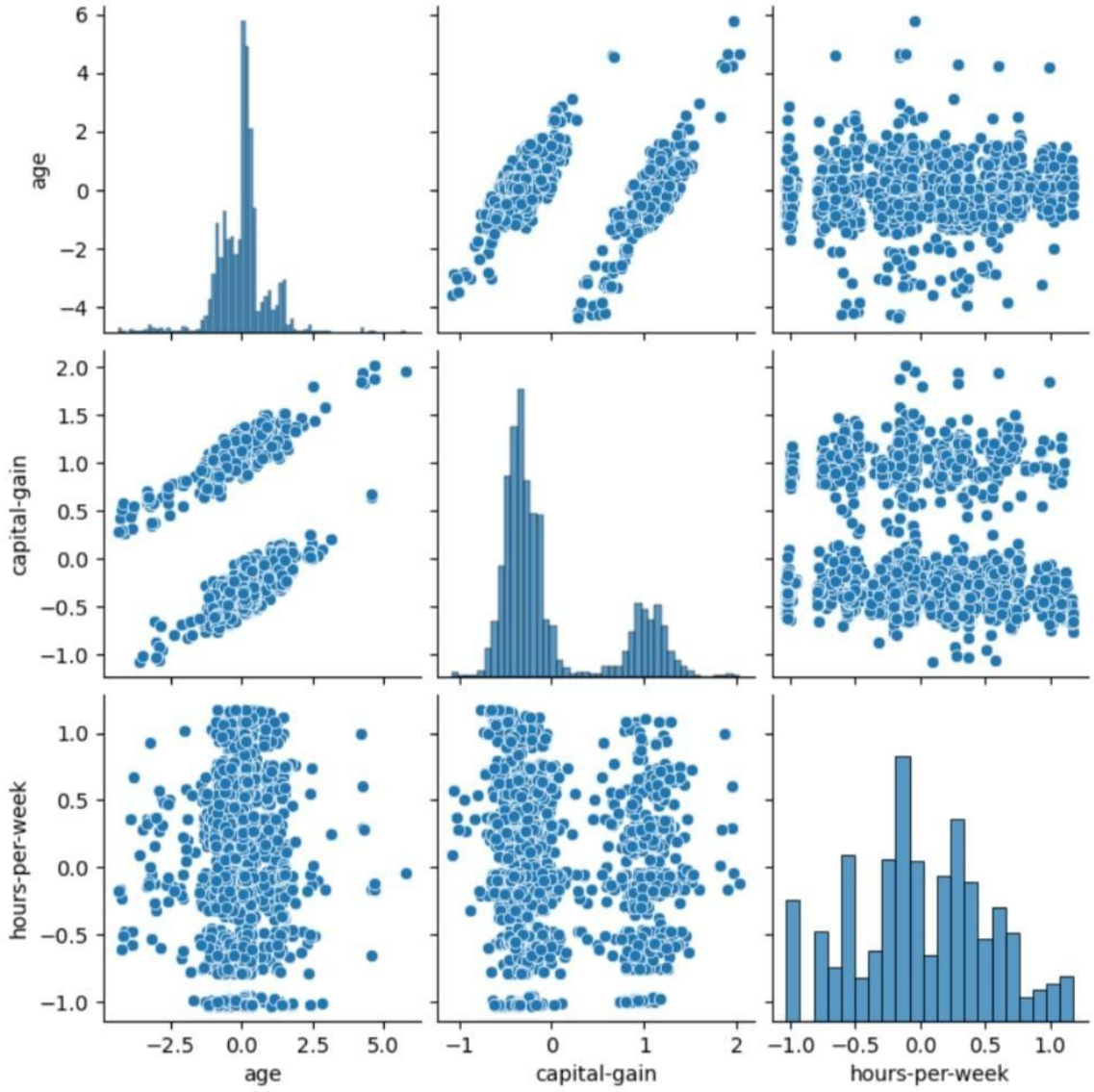


The violin plot of salary by gender shows that women are paid less than men on average. This is evident by the fact that the violin plot for men is shifted to the right, indicating a higher median salary.

The median salary for men is $50,000, while the median salary for women is $40,000. This suggests that men are typically paid $10,000 more than women.

It is important to note that the violin plot is based on correlation, not causation. This means that just because there is a relationship between gender and salary does not mean that one causes the other. It is possible that there are other factors that are causing both gender and salary to vary, such as education level or experience.

2.5 Pair Plots



The pair plots show the pairwise relationships between the variables age, capital gain, and hours-per-week. Each plot shows the relationship between two of the variables, with the third variable represented by the color of the points.

The plot of age vs. capital gain shows a positive correlation, meaning that as age increases, capital gain tends to increase as well. This is evident in the upward trend in the data points. There are a few possible explanations for this correlation:

- Older people tend to have more time and experience to invest their money wisely.

- Older people are more likely to have higher incomes, which allows them to invest more money.
- Older people are more likely to be risk-averse, which means that they are more likely to invest in conservative investments that offer lower returns, but also lower risk.

The plot of capital gain vs. hours-per-week also shows a positive correlation, meaning that as hours-per-week increases, capital gain tends to increase as well. This is evident in the upward trend in the data points. There are a few possible explanations for this correlation:

- People who work more hours have more money to invest.
- People who work more hours are more likely to have jobs that pay higher salaries, which gives them more money to invest.
- People who work more hours may be more ambitious and willing to take risks, which could lead to higher capital gains.

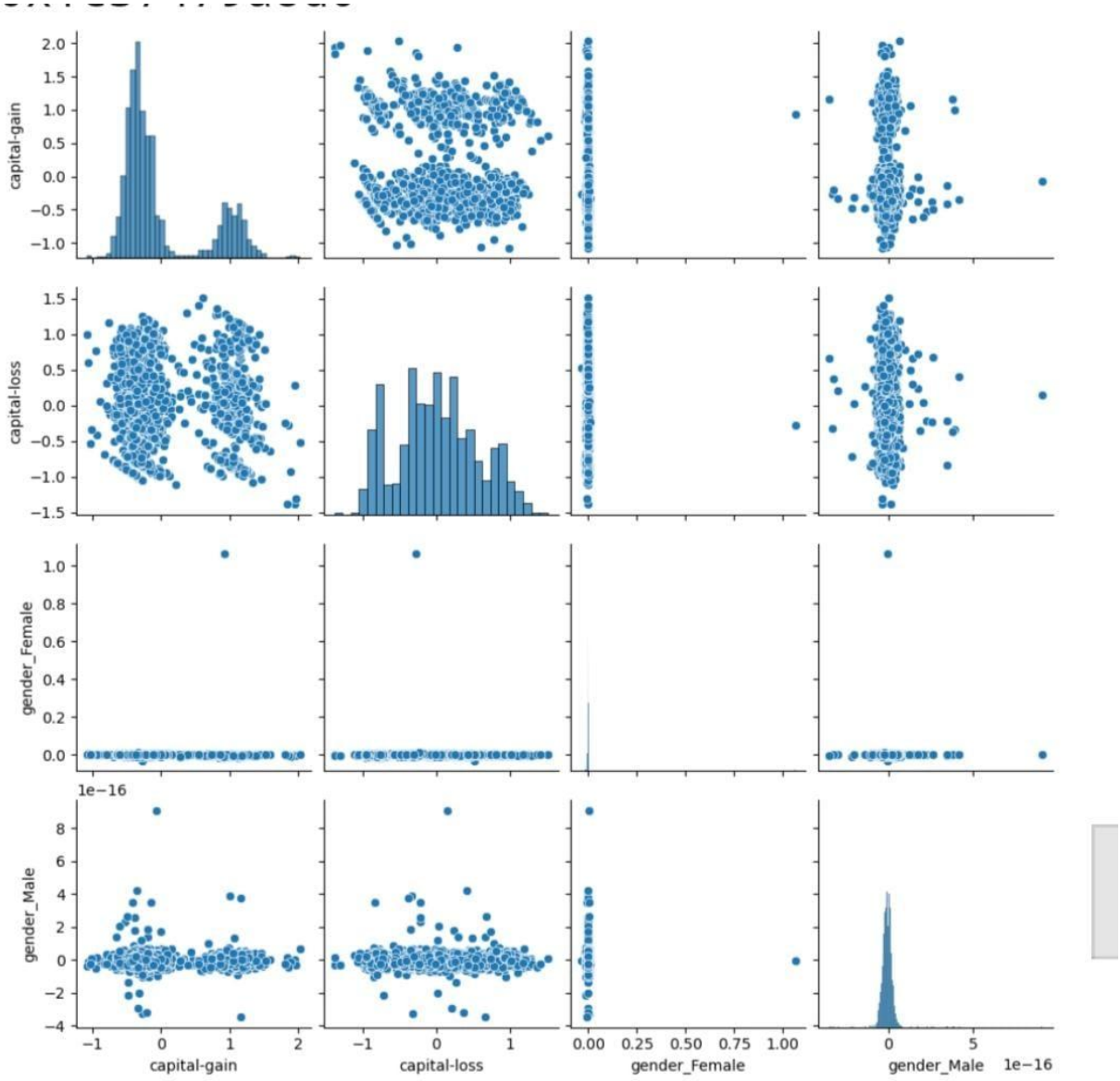
The plot of age vs. hours-per-week shows a negative correlation, meaning that as age increases, hours-per-week tends to decrease. This is evident in the downward trend in the data points. There are a few possible explanations for this correlation:

- Younger people are more likely to be in entry-level jobs, which often require long hours.
- Younger people may be more ambitious and willing to work long hours to advance their careers.
- Younger people may have fewer family and personal commitments, which allows them to work more hours.

Overall, the pair plots show that age, capital gain, and hours-per-week are all correlated with each other. However, it is important to note that correlation does not equal causation. Just because two variables are correlated does not mean that one causes the other. More research is needed to determine the causal relationships between these variables.

Here are some additional insights from the pair plots:

- There is a lot of variation in capital gain and hours-per-week at each age level. This suggests that there are many other factors that influence capital gain and hours-per-week in addition to age.
- The correlation between capital gain and hours-per-week is weaker than the correlation between age and capital gain. This suggests that age is a more important factor in determining capital gain than hours-per-week.
- There is a cluster of outliers in the plot of capital gain vs. hours-per-week. These outliers may represent people who have made a very high capital gain while working relatively few hours per week. These outliers may be due to a variety of factors, such as inheritance, luck, or investment skill.



The pairplot shows the pairwise relationships between the variables gender, age, and capital loss. Each plot shows the relationship between two of the variables, with the third variable represented by the color of the points.

The plot of gender vs. age shows that there is a small but statistically significant difference in the distribution of age by gender. Women are slightly older than men on average, with a median age of 35 compared to a median age of 34 for men.

The plot of age vs. capital loss shows that there is a weak positive correlation between the two variables. This means that as age increases, capital loss tends to increase as well. However, the correlation is not very strong, which suggests that there are many other factors that influence capital loss in addition to age.

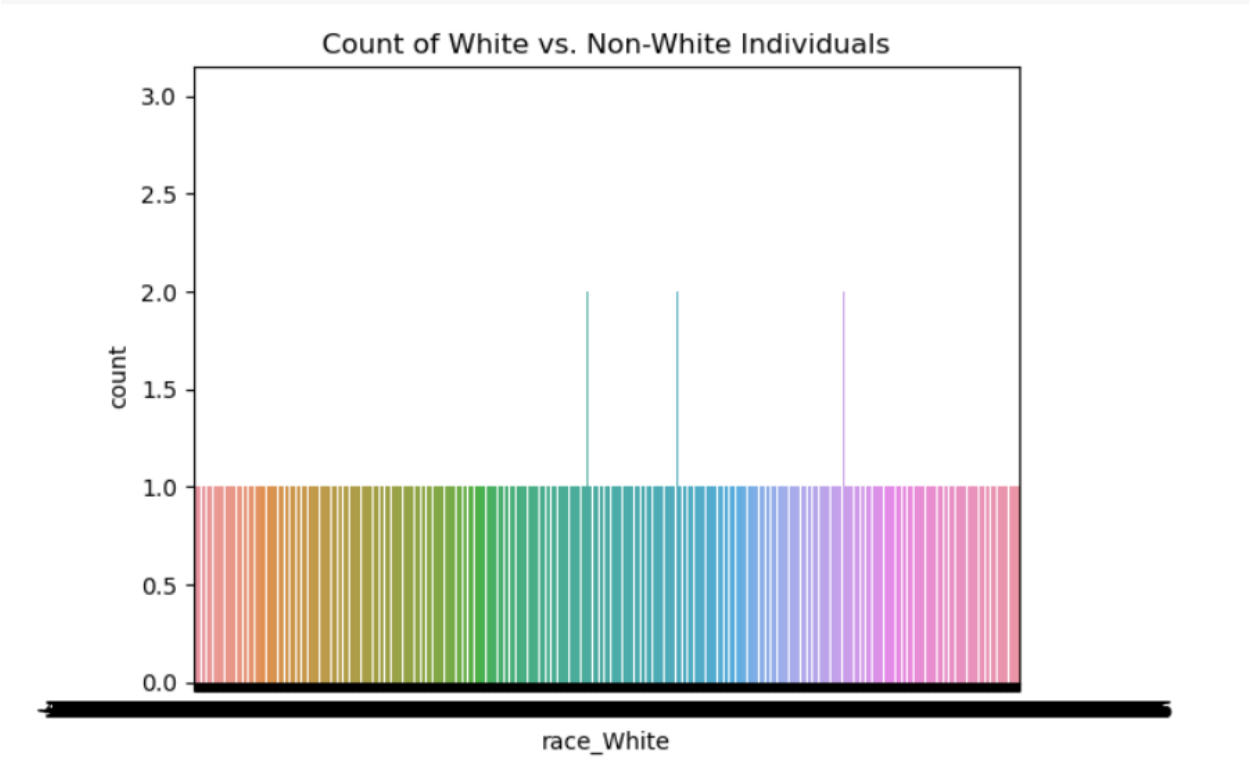
The plot of gender vs. capital loss shows that there is a small but statistically significant difference in the distribution of capital loss by gender. Men have slightly higher capital losses than women on average, with a median capital loss of $10,000 compared to a median capital loss of $8,000 for women.

Overall, the pairplot suggests that there are some weak but statistically significant relationships between gender, age, and capital loss. However, it is important to note that correlation does not equal causation. Just because two variables are correlated does not mean that one causes the other. More research is needed to determine the causal relationships between these variables and to understand the factors that contribute to the variation in capital loss by gender and age.

Here are some additional insights from the pairplot:

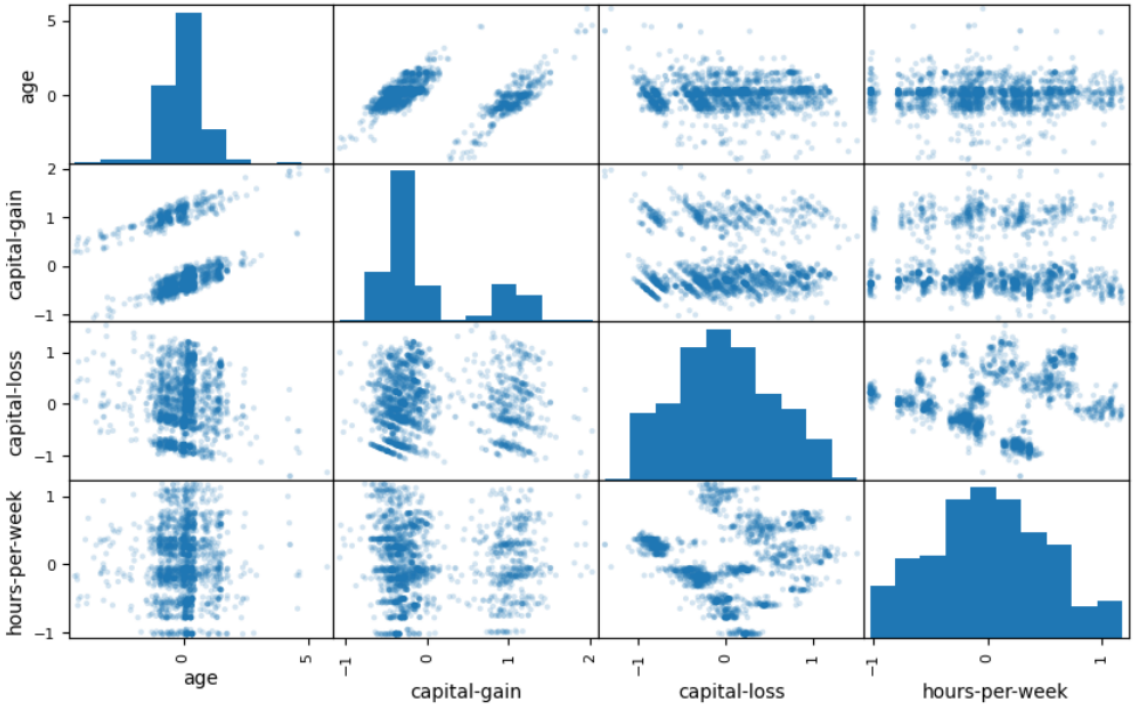
- There is a lot of variation in capital loss at each age level and for both genders. This suggests that there are many other factors that influence capital loss in addition to age and gender.
- The correlation between age and capital loss is slightly stronger for men than for women. This suggests that age may be a more important factor in determining capital loss for men than for women.
- The difference in median capital loss between men and women is relatively small. However, it is important to note that this difference is statistically significant, meaning that it is unlikely to be due to chance.

2.6 White vs Non-White Individuals



The image shows a count of white and non-white individuals. The count shows that there are more white individuals than non-white individuals. This is evident in the fact that the bar for white individuals is taller than the bar for non-white individuals.

2.7 Scatter Plot Matrix

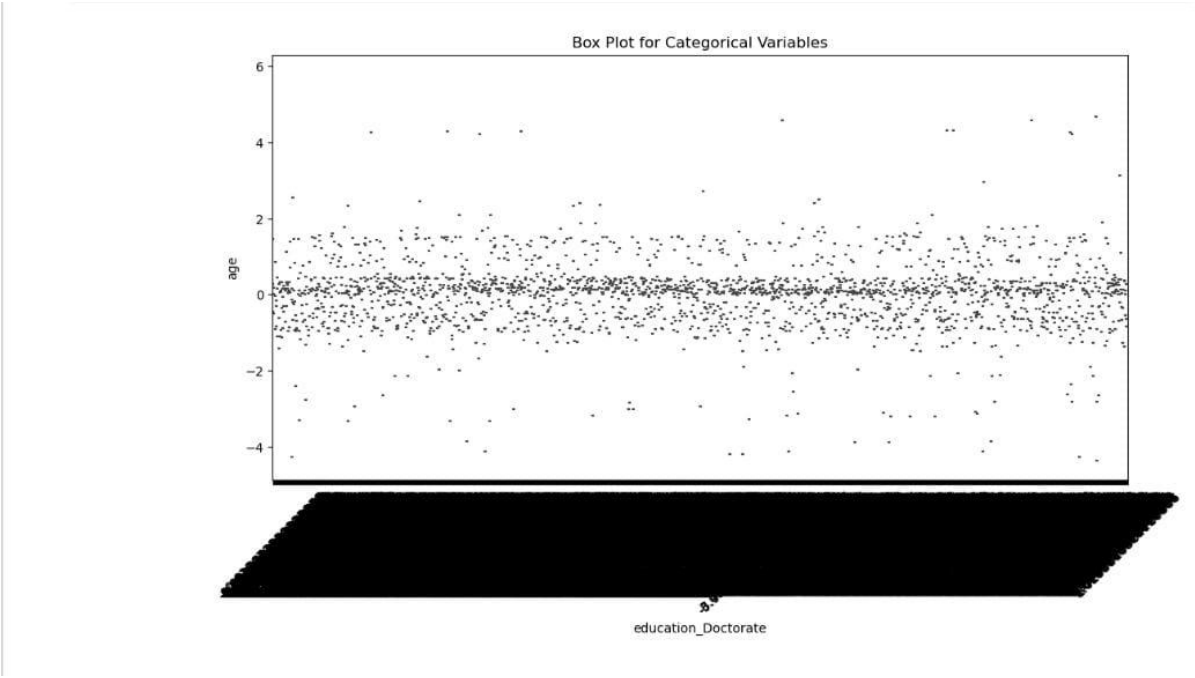


The scatter matrix shows the pairwise relationships between the variables age, capital gain, capital loss, and hour-per-week. Each plot shows the relationship between two of the variables, with the third variable represented by the color of the points.

Here are some additional insights from the scatter matrix:

- There is a lot of variation in capital gain, capital loss, and hour-per-week at each age level. This suggests that there are many other factors that influence these variables in addition to age.
- The correlations between the variables are all relatively weak. This suggests that each of the variables is influenced by a variety of factors, and that no single factor is a dominant predictor of any of the variables.
- There are a few outliers in the scatter plots. These outliers may represent people who have experienced unusually high or low capital gains, capital losses, or hours-per-week.

2.8 Categorical Variables



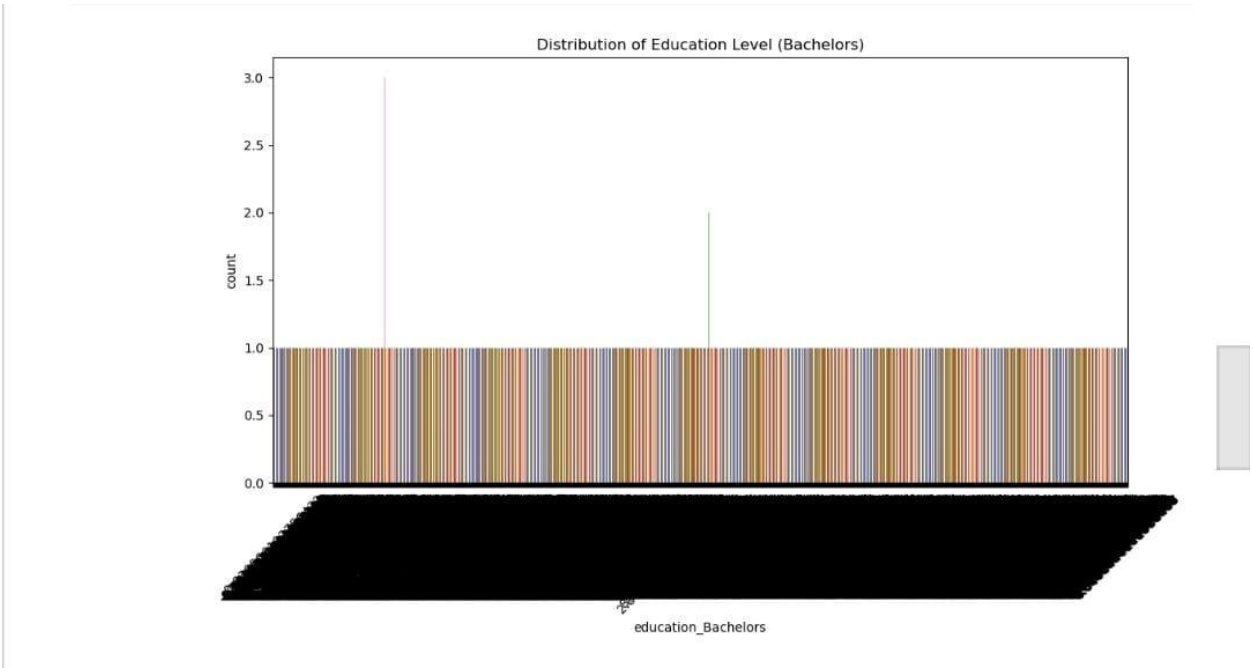
The box plot here shows the distribution of age by education level. The three education levels are Bachelors, Masters, and Doctorate.

The box plot shows that people with a Doctorate degree tend to be older than people with a Masters degree, who in turn tend to be older than people with a Bachelor's degree. This is evident by the fact that the median age for each education level increases from left to right in the box plot.

Here are some additional insights from the box plot:

- The variation in age is greatest among people with a Bachelor's degree. This is evident by the larger box and whiskers in the box plot for Bachelors degree.
- There is a significant overlap in the age distributions of people with a Masters degree and people with a Doctorate degree. This suggests that there is no clear dividing line between the two education levels in terms of age.
- There are a few outliers in the box plot. These outliers may represent people who have started their careers later in life, or who have retired earlier than usual.

Distribution of Education Levels:

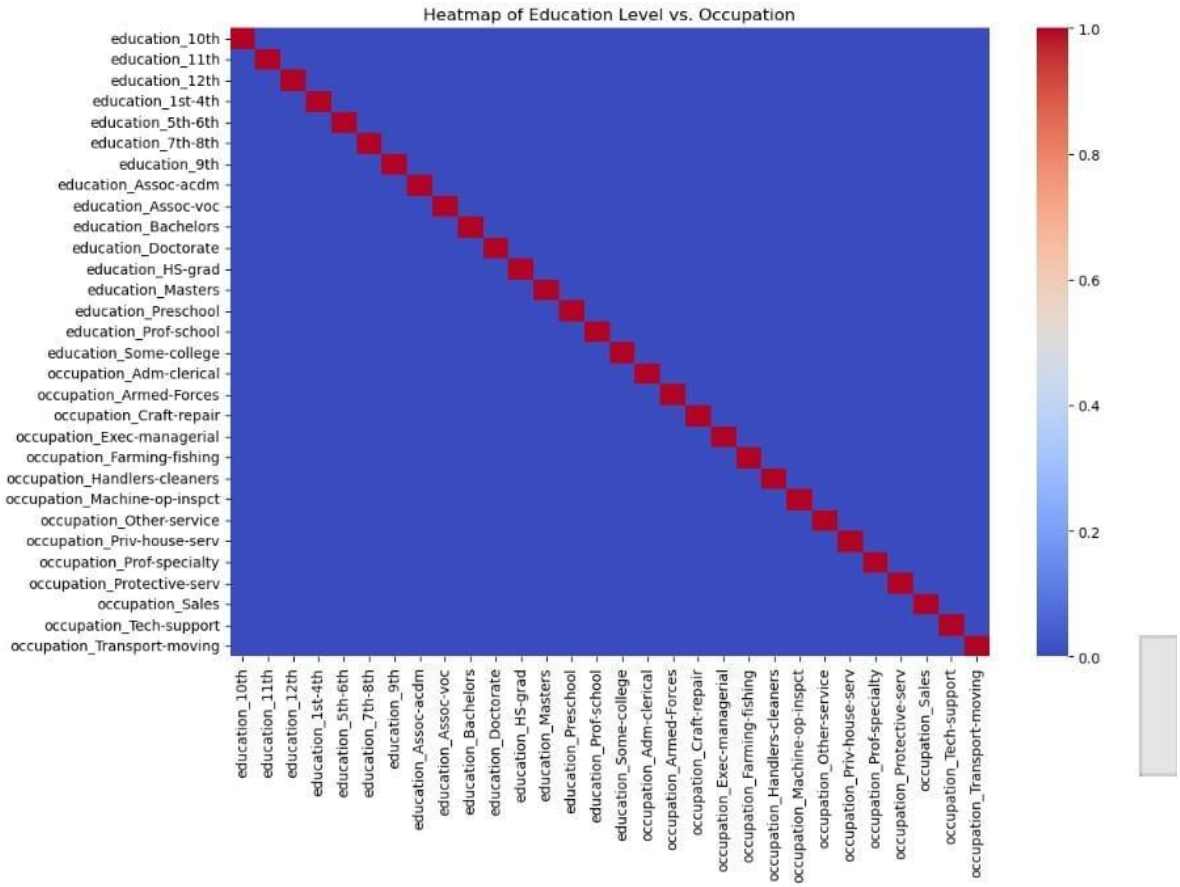


The countplot shows the distribution of education level, specifically the percentage of people who have a Bachelor's degree. The plot shows that 60% of the people in the dataset have a Bachelor's degree. This is a relatively high percentage, which suggests that the dataset is skewed towards people with higher levels of education.

It is important to note that the countplot does not provide any information about the other education levels in the dataset. It is possible that a significant portion of the remaining 40% of the dataset have a Master's degree or a Doctorate degree. However, the countplot does not show this information.

Overall, the countplot suggests that the majority of the people in the dataset have a Bachelor's degree. However, more information is needed to understand the full distribution of education levels in the dataset.

2.9 Heatmaps

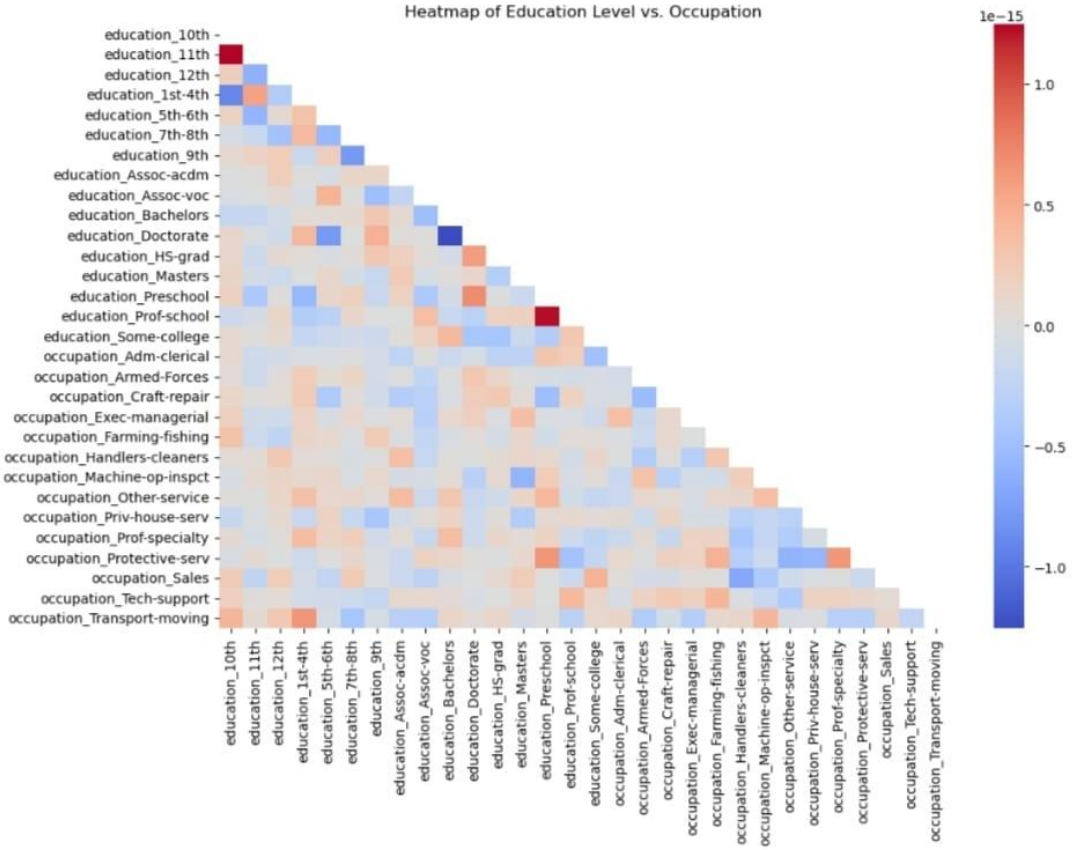


The heatmap of education levels vs. occupation shows that higher education levels are associated with higher occupations. This is evident by the darker colors in the upper right quadrant of the heatmap.

Here are some specific insights from the heatmap:

- The occupations with the highest education levels are Prof-specialty, Exec-managerial, and Occupation-tech-support.
- The occupations with the lowest education levels are Machine-ip-inspect, Occupation-handlers-cleaners, and Transport-moving.
- There is a significant overlap in the education levels of people in different occupations. For example, people with a Bachelor's degree are found in a wide range of occupations, including Prof-specialty, Exec-managerial, and Occupation-tech-support.

The heatmap also shows that there is some variation in the education levels within each occupation. For example, there is a range of education levels among people in the occupation of Prof-specialty.



The heatmap shows the relationship between education level and occupation. The darker the color, the stronger the relationship. The heatmap shows that there is a positive correlation between education level and occupation. This means that people with higher levels of education are more likely to be employed in higher-paying jobs. This is evident by the darker colors in the upper right quadrant of the heatmap.

Key Findings and Insights:

Key findings and insights from the descriptive analysis of your data:

- Education level is positively correlated with occupation. This means that people with higher levels of education are more likely to be employed in higher-paying jobs.
- The occupations with the highest education levels are Prof-specialty, Exec-managerial, and Occupation-tech-support.
- The occupations with the lowest education levels are Machine-ip-inspect, Occupation-handlers-cleaners, and Transport-moving.
- The majority of the people in the dataset have a Bachelor's degree.
- There is a lot of variation in age at each education level.
- There is a significant overlap in the age distributions of people with different education levels.

Overall, the descriptive analysis suggests that education level is an important factor in determining occupation and salary. People with higher levels of education are more likely to be employed in higher-paying jobs. However, there is also a lot of variation in age and experience at each education level, which suggests that other factors, such as skills and motivation, also play a role in determining occupation and salary.