Column Types

- **Numerical** - Age,Fare,PassengerId
- **Categorical** - Survived, Pclass, Sex, SibSp, Parch,Embarked
- **Mixed** - Name, Ticket, Cabin

Univariate Analysis

Univariate analysis focuses on analyzing each feature in the dataset independently.

- **Distribution analysis**: The distribution of each feature is examined to identify its shape, central tendency, and dispersion.
- **Identifying potential issues**: Univariate analysis helps in identifying potential problems with the data such as outliers, skewness, and missing values

The shape of a data distribution refers to its overall pattern or form as it is represented on a graph. Some common shapes of data distributions include:

- **Normal Distribution**: A symmetrical and bell-shaped distribution where the mean, median, and mode are equal and the majority of the data falls in the middle of the distribution with gradually decreasing frequencies towards the tails.
- **Skewed Distribution**: A distribution that is not symmetrical, with one tail being longer than the other. It can be either positively skewed (right-skewed) or negatively skewed (left-skewed).
- **Bimodal Distribution**: A distribution with two peaks or modes.
- **Uniform Distribution**: A distribution where all values have an equal chance of occurring.

The shape of the data distribution is important in identifying the presence of outliers, skewness, and the type of statistical tests and models that can be used for further analysis.

Dispersion is a statistical term used to describe the spread or variability of a set of data. It measures how far the values in a data set are spread out from the central tendency (mean, median, or mode) of the data.

There are several measures of dispersion, including:

- **Range**: The difference between the largest and smallest values in a data set.
- **Variance**: The average of the squared deviations of each value from the mean of the data set.
- **Standard Deviation**: The square root of the variance. It provides a measure of the spread of the data that is in the same units as the original data.
- **Interquartile range (IQR)**: The range between the first quartile (25th percentile) and the third quartile (75th percentile) of the data.

Dispersion helps to describe the spread of the data, which can help to identify the presence of outliers and skewness in the data.

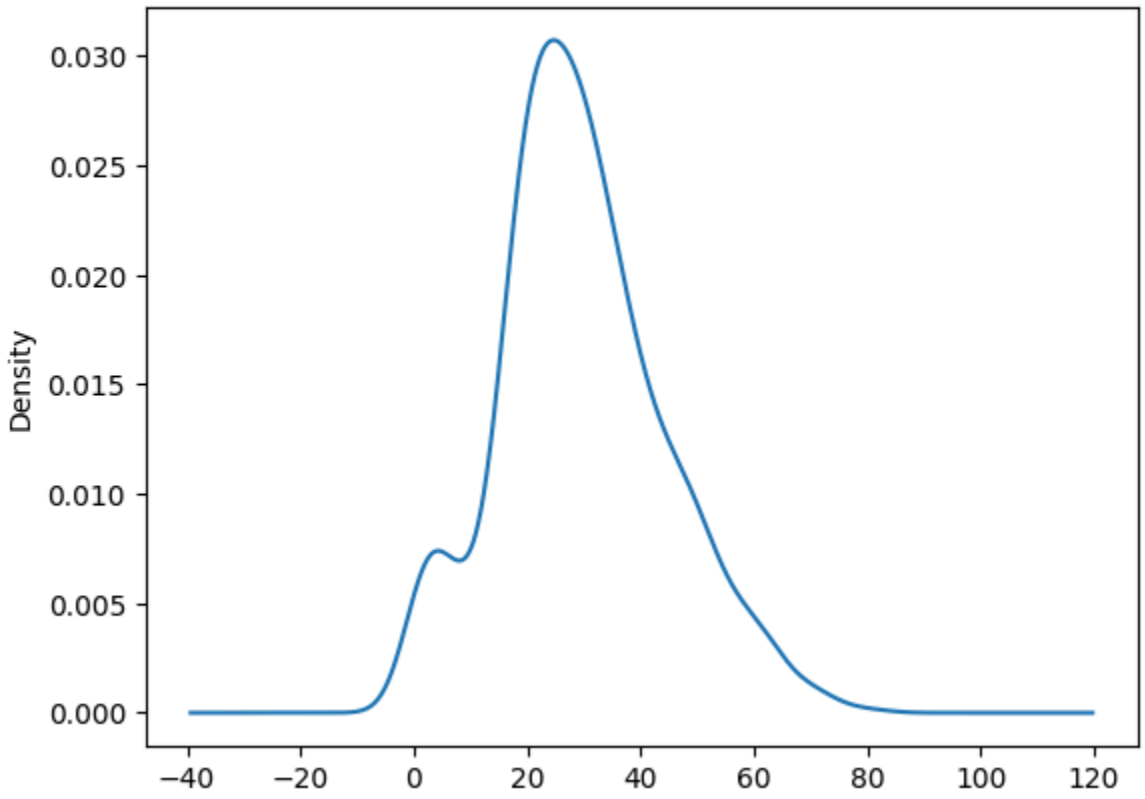
Steps of doing Univariate Analysis on Numerical columns

- **Descriptive Statistics**: Compute basic summary statistics for the column, such as mean, median, mode, standard deviation, range, and quartiles. These statistics give a general understanding of the distribution of the data and can help identify skewness or outliers.
- **Visualizations**: Create visualizations to explore the distribution of the data. Some common visualizations for numerical data include histograms, box plots, and density plots. These visualizations provide a visual representation of the distribution of the data and can help identify skewness an outliers.
- **Identifying Outliers**: Identify and examine any outliers in the data. Outliers can be identified using visualizations. It is important to determine whether the outliers are due to measurement errors, data entry errors, or legitimate differences in the data, and to decide whether to include or exclude them from the analysis.
- **Skewness**: Check for skewness in the data and consider transforming the data or using robust statistical methods that are less sensitive to skewness, if necessary.
- **Conclusion**: Summarize the findings of the EDA and make decisions about how to proceed with further analysis.

Kernel Density Estimate (KDE) plot for the 'Age' variable

```
df['Age'].plot(kind='kde')
```

```
<Axes: ylabel='Density'>
```



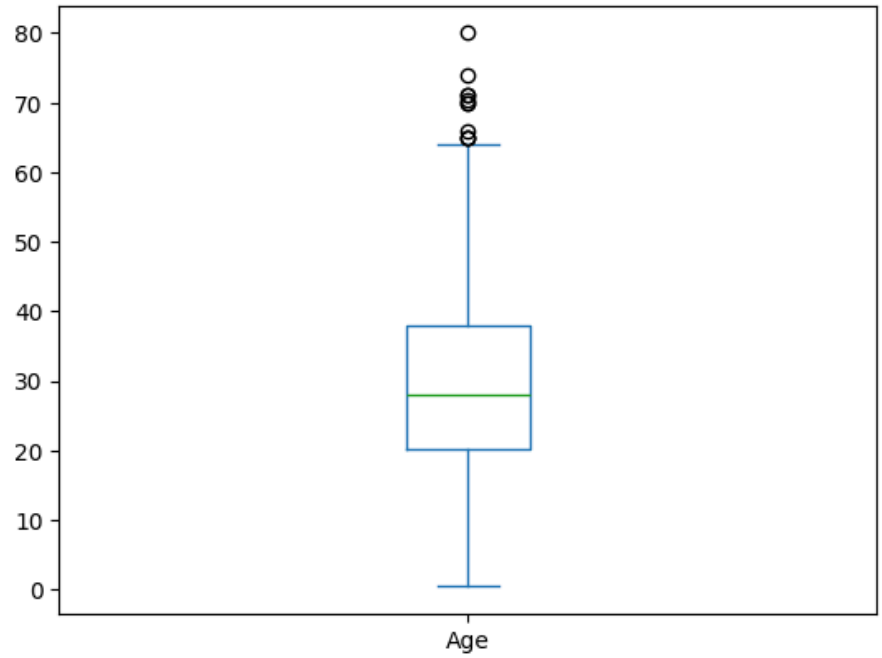
Key observations from the plot:

- **Primary Peak:** There's a prominent peak in the density around the age of 25-30. This indicates that the most frequent ages in the dataset are within this range.
- **Secondary Smaller Peak:** A smaller peak is visible around the age of 5-10, suggesting a smaller concentration of individuals in this younger age group.
- **Skewness:** The distribution appears to be slightly right-skewed. This means the tail of the distribution extends more towards the higher ages, indicating that there are relatively fewer individuals in the older age ranges compared to the younger and middle-aged groups.
- **Spread:** The curve shows that the ages are distributed over a considerable range, although the density is highest in the 20-40 age group.
- **Smoothness:** The KDE plot provides a smoother view of the distribution compared to a histogram, highlighting the underlying shape of the data.

Box plot for the 'Age' variable

```
df['Age'].plot(kind='box')
```

```
<Axes: >
```



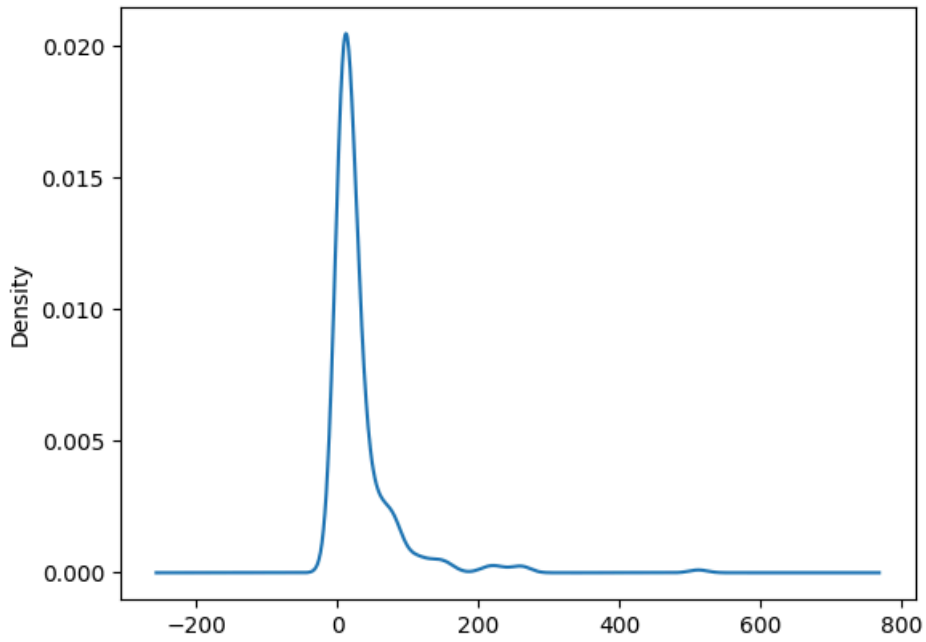
Insights from the Box Plot:

- **Central Tendency:** The median age (green line) is around 28 years. This gives us a sense of the typical age in the dataset.
- **Spread or Variability:** The height of the box (IQR) shows the spread of the middle 50% of the ages, which ranges from approximately 20 years (Q1) to 37 years (Q3). This IQR of around 17 years indicates the variability within the central portion of the age distribution.
- **Skewness:** The median line is positioned slightly below the center of the box. The upper whisker is also somewhat longer than the lower whisker. These observations suggest a slight positive skew in the age distribution, meaning there's a longer tail towards older ages. This aligns with the slight right skew observed in the KDE plot you shared earlier.
- **Range:** While not explicitly marked, we can infer that the majority of the data falls roughly between 0 and around 65 years (the upper whisker).
- **Outliers:** There are several data points identified as outliers above the upper whisker, with ages ranging from approximately 65 to 80 years. These individuals are significantly older than the majority of the dataset based on the 1.5 * IQR rule. These outliers warrant further investigation to understand if they are genuine extreme values or potential errors.
- **Concentration:** The box itself shows where the bulk of the data lies – between 20 and 37 years.

Kernel Density Estimate (KDE) plot for the 'Fare' variable

```
df['Fare'].plot(kind='kde')
```

```
<Axes: ylabel='Density'>
```



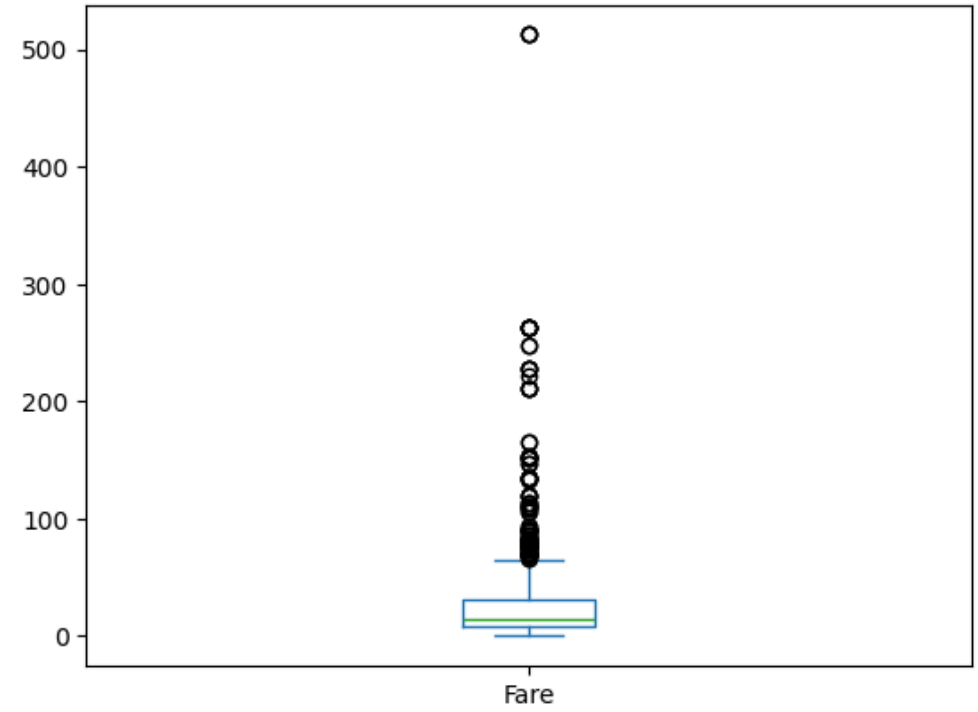
Insights from the KDE Plot (in points):

- **Strong Central Peak at Low Fare:** There is a very sharp and high peak in the density at a fare value close to zero (likely indicating a significant number of entries with very low or even zero fare). This suggests a high frequency of either free tickets, very discounted fares, or potentially a data entry artifact.
- **Rapid Decay After the Peak:** The density drops very sharply immediately after this initial peak. This implies that as the fare increases slightly from the very low values, the number of occurrences decreases dramatically.
- **Long Right Tail (Positive Skew):** The curve extends towards higher fare values, creating a long right tail. This indicates that while most fares are low, there are some instances of significantly higher fares. This distribution is strongly positively skewed.
- **Smaller Secondary Humps/Shoulders:** There appear to be smaller humps or shoulders in the density curve at higher fare values (perhaps around 50-100 and a smaller one around 200). These might suggest smaller groups of observations clustered around these specific fare ranges.
- **Very Few High Fares:** The density becomes very low for fares above 200, indicating that very high fare values are relatively rare in the dataset.
- **Non-Negative Fare:** While the x-axis extends to negative values due to the nature of KDE estimation, in reality, fare is likely a non-negative value. The peak near zero reinforces this idea.

Box plot for the 'Fare' variable

```
df['Fare'].plot(kind='box')
```

```
<Axes: >
```



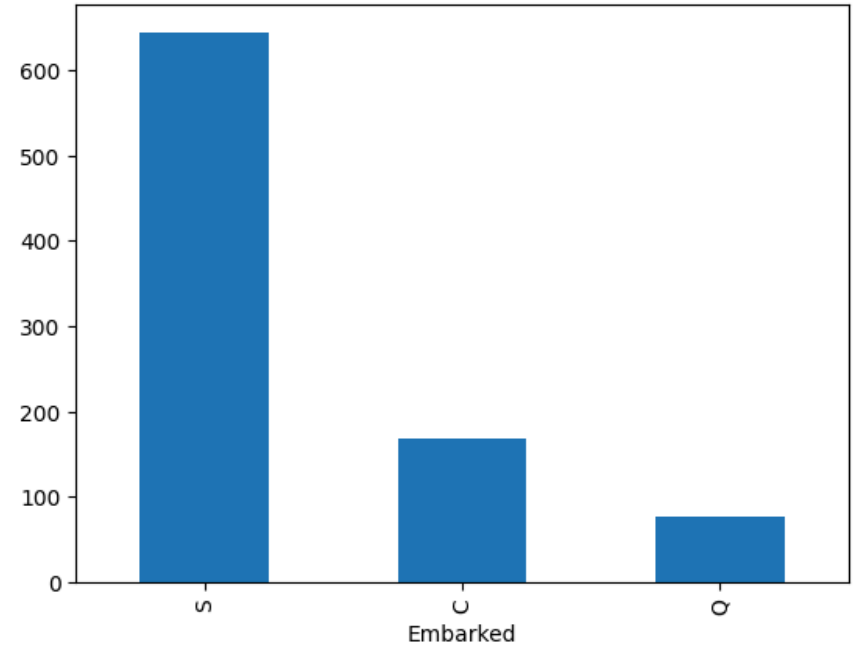
Insights from the Box Plot (in points):

- **Low Median Fare:** The median fare (around 14) is relatively low, indicating that half of the fares in the dataset are below this value.
- **Concentration of Lower Fares:** The box itself, spanning from approximately 7 to 31, shows that the middle 50% of the fares are concentrated in this lower range.
- **Positive Skew:** The upper whisker is significantly longer than the lower whisker, and there are numerous outliers on the higher end. This strongly suggests a positive skew in the fare distribution, meaning there are more observations with lower fares and a tail extending towards higher fares. This aligns with the KDE plot you showed previously.
- **Wide Range of Fares:** Despite the concentration of lower fares, the presence of outliers indicates a very wide range of fare values in the dataset, extending to over 500.
- **Numerous Outliers:** There is a considerable number of outliers with high fare values. These could represent premium ticket classes, special fares, or potentially even data errors that might warrant investigation.
- **Most Fares are Below 65:** The upper whisker suggests that the vast majority of the fares fall below approximately 65. The values above this are exceptional.

Bar Plot of the 'Embarked' variable

```
df['Embarked'].value_counts().plot(kind='bar')
```

```
<Axes: xlabel='Embarked'>
```



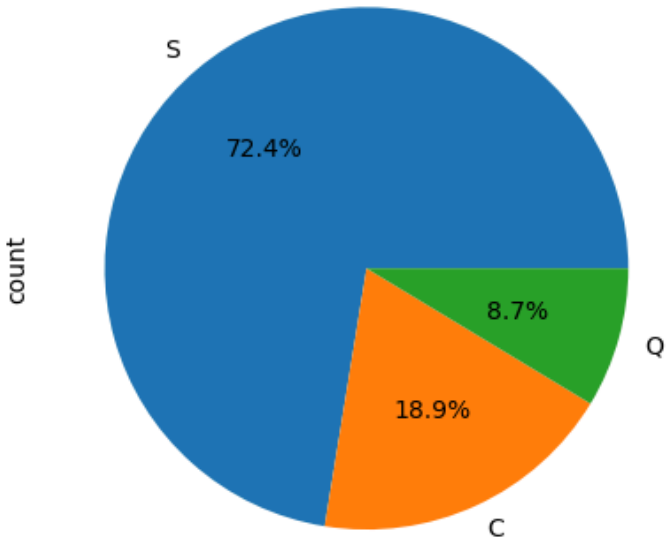
Insights from the Bar Plot (in points):

- **Dominant Embarkation Port 'S':** The bar corresponding to 'S' is significantly the tallest. This indicates that the vast majority of individuals in the dataset embarked from the port represented by 'S'. The count for 'S' is approximately 640.
- **Second Most Frequent Port 'C':** The bar for 'C' is the second tallest, showing that a considerably smaller, but still significant, number of individuals embarked from this port. The count for 'C' is around 170.
- **Least Frequent Port 'Q':** The bar for 'Q' is the shortest, revealing that the fewest individuals in the dataset embarked from the port represented by 'Q'. The count for 'Q' is approximately 80.
- **Uneven Distribution:** The distribution of embarkation ports is highly uneven. Port 'S' accounts for a much larger proportion of the individuals compared to ports 'C' and 'Q'.
- **Potential Implications:** This uneven distribution could have implications for other analyses. For example, survival rates or other characteristics might differ based on the port of embarkation. It might be worth investigating if there are any underlying reasons for this disparity (e.g., the location of the port relative to the origin of the passengers, the type of ship or route).

Pie chart representing the distribution of the 'Embarked' variable

```
df['Embarked'].value_counts().plot(kind='pie',autopct='%0.1f%%')
```

```
<Axes: ylabel='count'>
```



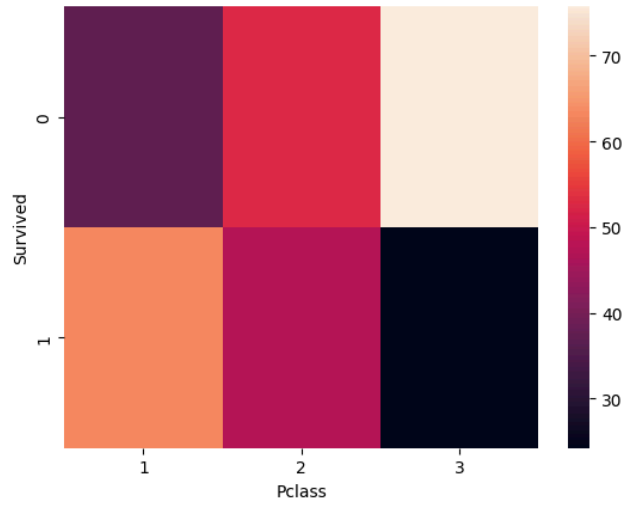
Insights from the Pie Chart (in points):

- **'S' is the Dominant Embarkation Port:** The largest slice, representing 72.4% of the pie, corresponds to the embarkation port 'S'. This confirms that a vast majority of the individuals in the dataset embarked from this port.
- **'C' is the Second Largest Group:** The second largest slice, accounting for 18.9% of the pie, represents the embarkation port 'C'. This indicates that 'C' was the second most common port of embarkation, though significantly less frequent than 'S'.
- **'Q' is the Smallest Group:** The smallest slice, making up only 8.7% of the pie, corresponds to the embarkation port 'Q'. This shows that 'Q' was the least common port of embarkation in the dataset.
- **Clear Proportionality:** The pie chart visually emphasizes the significant difference in the number of individuals embarking from each port. The slice for 'S' is much larger than the other two combined.
- **Reinforces Bar Plot Insights:** The insights gained from this pie chart align perfectly with those from the bar plot you shared earlier. Both visualizations clearly demonstrate the dominance of 'S', followed by 'C', and then 'Q' in terms of the number of individuals who embarked from each port.
- **Easy Comparison of Proportions:** The pie chart makes it easy to quickly compare the relative proportions of individuals from each embarkation port to the whole dataset.

Heatmap which visualises the relationship between 'Pclass' (Passenger Class) and 'Survived'

```
sns.heatmap(pd.crosstab(df['Survived'],df['Pclass'],normalize='columns')*100)
```

```
<Axes: xlabel='Pclass', ylabel='Survived'>
```



Insights from the Heatmap (in points):

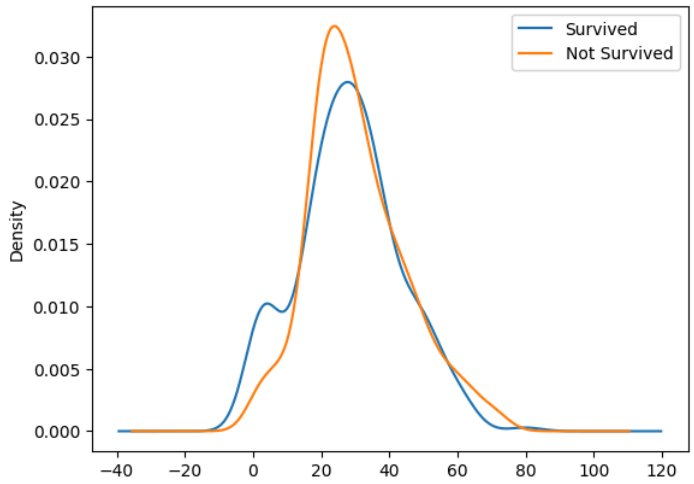
- **Higher Survival Rate in Higher Pclasses:** The top row (Survived = 0) shows darker colors for Pclass 1 and 2 compared to Pclass 3. Conversely, the bottom row (Survived = 1) shows lighter colors for Pclass 1 and 2 compared to Pclass 3. This strongly indicates that passengers in higher passenger classes (1st and 2nd) had a significantly higher survival rate than those in the lower passenger class (3rd).
- **Lowest Survival Rate in Pclass 3:** The darkest cell in the 'Survived = 1' row is under 'Pclass = 3', and the lightest cell in the 'Survived = 0' row is also under 'Pclass = 3'. This highlights that the 3rd class had the lowest percentage of survivors and the highest percentage of non-survivors.
- **Highest Survival Rate in Pclass 1:** The lightest cell in the 'Survived = 1' row is under 'Pclass = 1', and the darkest cell in the 'Survived = 0' row is also under 'Pclass = 1'. This clearly shows that the 1st class had the highest percentage of survivors and the lowest percentage of non-survivors.
- **Pclass 2 Intermediate:** The survival rate for the 2nd class falls between that of the 1st and 3rd classes. The color intensities for Pclass 2 are intermediate compared to the extremes seen in Pclass 1 and 3.
- **Quantifiable Survival Rates (Approximate):** By referencing the color scale, we can get approximate survival rates for each class:
    - **Pclass 1:** Survival rate appears to be around 60-70%.
    - **Pclass 2:** Survival rate seems to be around 40-50%.
    - **Pclass 3:** Survival rate looks to be around 20-30%.

Kernel Density Estimate (KDE) plot showing the age distribution of passengers who survived and those who did not

```
# survived and age

df[df['Survived'] == 1]['Age'].plot(kind='kde',label='Survived')
df[df['Survived'] == 0]['Age'].plot(kind='kde',label='Not Survived')

plt.legend()
plt.show()
```



Insights from the KDE Plot (in points):

- **Higher Survival for Infants and Young Children:** The blue curve ('Survived') shows a noticeable peak in the very young age range (around 0-10 years) that is higher than the corresponding peak in the orange curve ('Not Survived'). This suggests that infants and young children had a higher probability of survival compared to other age groups.
- **Lower Survival for Young Adults:** The orange curve ('Not Survived') has a higher peak in the young adult age range (roughly 20-30 years) compared to the blue curve. This indicates that young adults had a relatively lower survival rate.
- **Similar Distribution in Middle Ages:** The two curves are quite close in the middle age range (approximately 30-60 years), suggesting that survival rates in this age group were less strongly influenced by age compared to the younger and young adult categories.
- **Slightly Lower Survival for Older Passengers:** The orange curve ('Not Survived') is slightly higher than the blue curve for older ages (above 60), hinting at a slightly lower survival rate for elderly passengers. However, the density in this range is low for both groups, meaning fewer individuals were in these age brackets.
- **Shape of Distributions:**
    - The 'Survived' distribution shows a more pronounced peak in early childhood and a slightly broader distribution overall.
    - The 'Not Survived' distribution has a more prominent peak in the young adult years.