Homework 1.1 Report

Zhihao Pei , Shixiu Wang , Chen Xu

March 2, 2025

Group Member			
Name	**ID**	**Main Work**	**Proportion**
Zhihao Pei	A20563413	Data Processing, Visualization	33.3%
Shixiu Wang	A20563415	Data Processing, Visualization	33.3%
Chen Xu	A20563433	Report Writing	33.3%

Dataset 1: heart_attack_prediction_India

Introduction

We are working with a heart risk prediction dataset, and our goal is to use data mining techniques to identify correlations between different variables and uncover potential factors that contribute to heart risk.

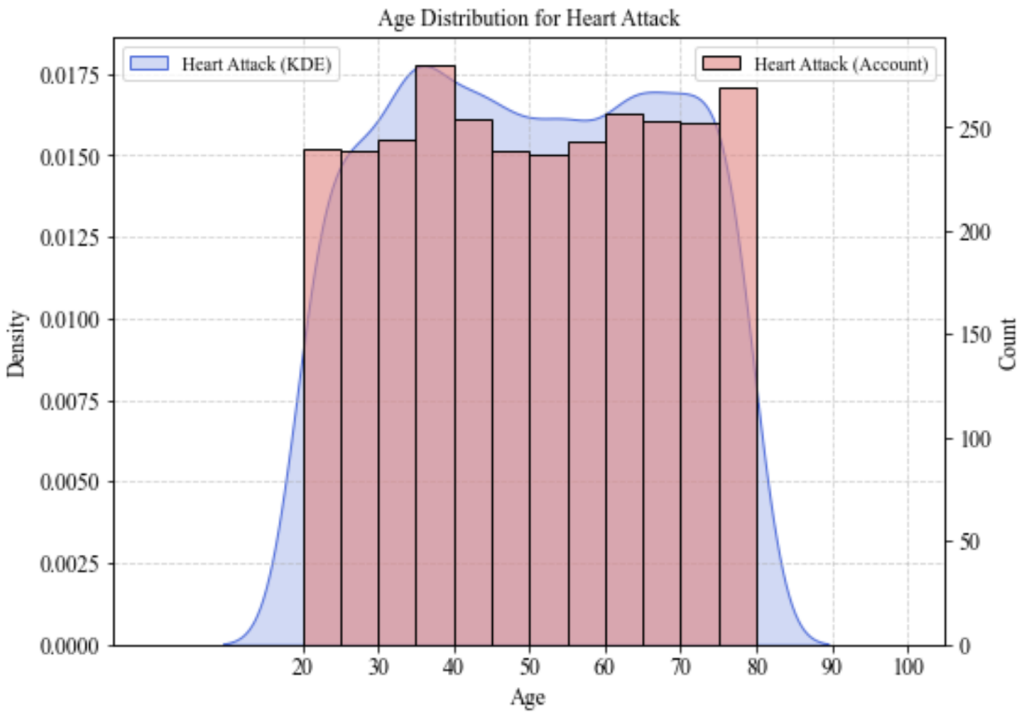
Dataset description

The dataset I used focuses on cardiovascular diseases (CVD) in India. It includes key medical and lifestyle risk factors such as diabetes, high blood pressure, obesity, smoking, exposure to air pollution, and access to healthcare services. I selected data from several columns, including: Patient_ID, State_Name, Gender, Age, Diabetes, Obesity, Smoking, Family_History, Stress_Level, and Heart_Attack_Risk.

I selected data from several columns, including: Patient_ID, State_Name, Gender, Age, Diabetes, Obesity, Smoking, Family_History, Stress_Level, and Heart_Attack_Risk. Specifically, Patient_ID represents the unique ID of the individual, State_Name indicates the region the individual is from, Gender refers to the individual's sex, Age represents the individual's age, Diabetes indicates whether the individual has diabetes, Obesity shows whether the individual is obese, Smoking indicates whether the individual smokes, Family_History refers to whether the individual has a family history of heart disease, Stress_Level represents the individual's stress level, and Heart_Attack_Risk indicates whether the individual is at risk of heart disease.

Method and Visualization

Visualization 1:



Methods used

First, discretize the continuous age range into discrete age groups (using the pandas.cut() function).

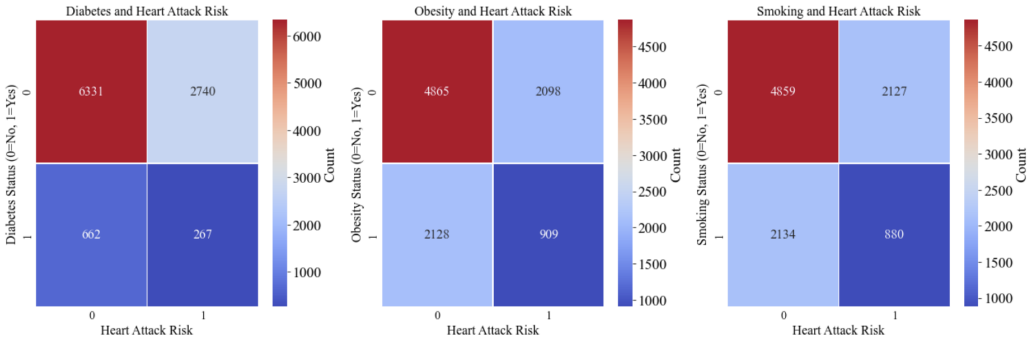
Next, use the plotting functions in the seaborn library (`sns.kdeplot`) and (`sns.histplot`) functions to specify some parameters and plot the KDE plot and histogram on different y-axes corresponding to the x-axis.

Finally, use matplotlib functions to customize the title, axis labels, ticks, grid lines, and legend.

Result analysis

From the above chart, it can be seen that the number of people with a risk of heart disease is roughly the same across all age groups, with the highest number in the 35-40 age range.

Visualization 2:



Methods used

First, use the `pd.crosstab` function to compute the cross-tabulation, which counts the frequency of each combination of disease status and heart disease risk.

Next, create a subplot layout to generate a row of multiple subplots (the number of columns equals the number of disease types), where each subplot is used to plot a heatmap.

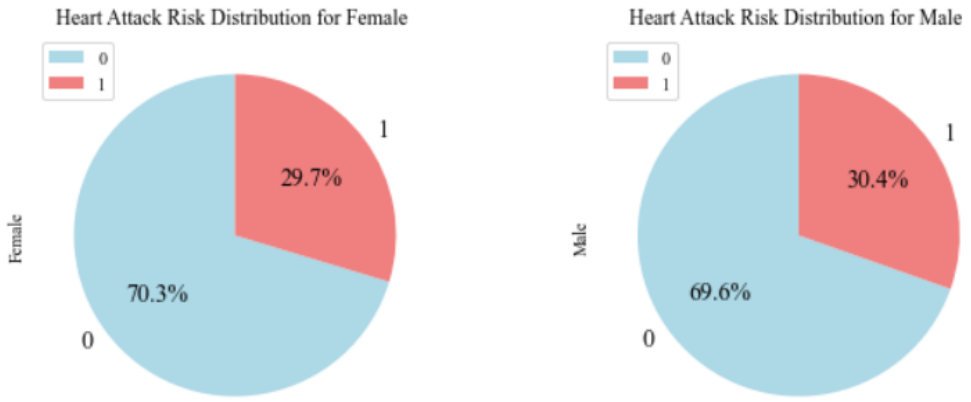
Then, use the `sns.heatmap` function to plot the heatmaps.

Finally, adjust the style to achieve a visually appealing result.

Result Analysis

From the chart, it can be seen that people with diabetes have the highest probability of heart disease risk, with a probability of $\frac{662}{662+267}$. This is followed by those with obesity, and then those with smoking habits.

Visualization 3:



Methods used

First, use the pd.crosstab function to compute the cross-tabulation, which counts the frequency of each combination of gender and heart disease risk.

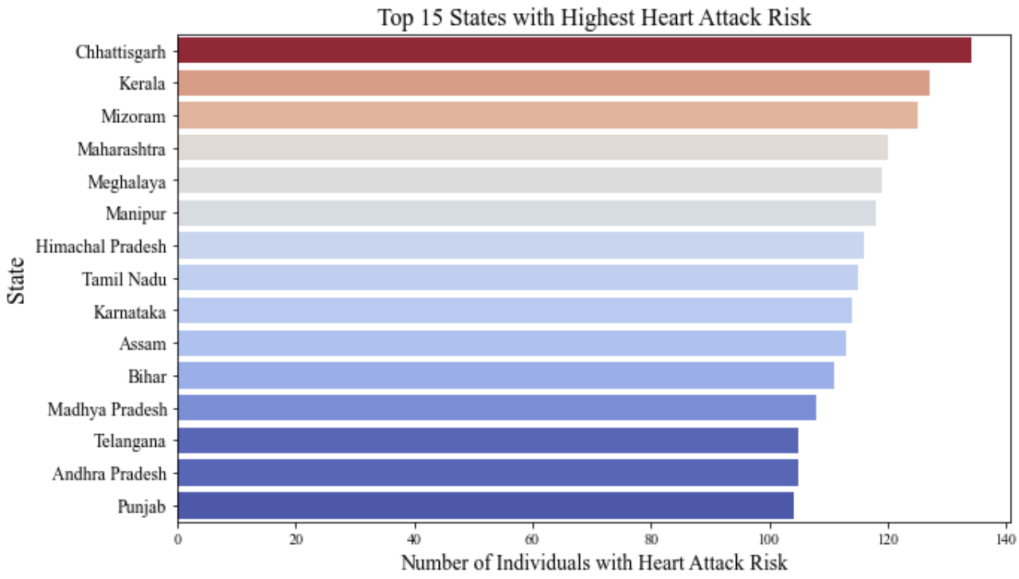
Next, create two pie charts: one for females (`gender_0_risk.plot`) and one for males (`gender_1_risk.plot`), labeling the distribution proportions of heart attack risk for each gender respectively.

Finally, adjust the wspace parameter to achieve a visually appealing result.

Result Analysis

From the above chart,it can be seen that there is no significant relationship between the probability of heart disease risk and gender.

Visualization 4:



Methods used

First, count the number of heart disease patients in each state and sort them in descending order.

Then, obtain the color mapping, assign colors to each state, and use sns.barplot to plot the bar chart.

Finally, add a title and labels to the bar chart for better understanding.

Result Analysis

From the chart, it can be seen that there are notable regional disparities in the risk of heart disease across various states in India. The high-risk states are dispersed across different regions, and there is no direct correlation between economic levels and heart disease risk. This suggests that heart disease risk is influenced by multiple factors rather than being solely determined by economic conditions.

Dataset 2: Mobiles Dataset (2025)

Introduction

This dataset mainly involves data on mobile devices from various companies. We selected this data to uncover the relationships within it.

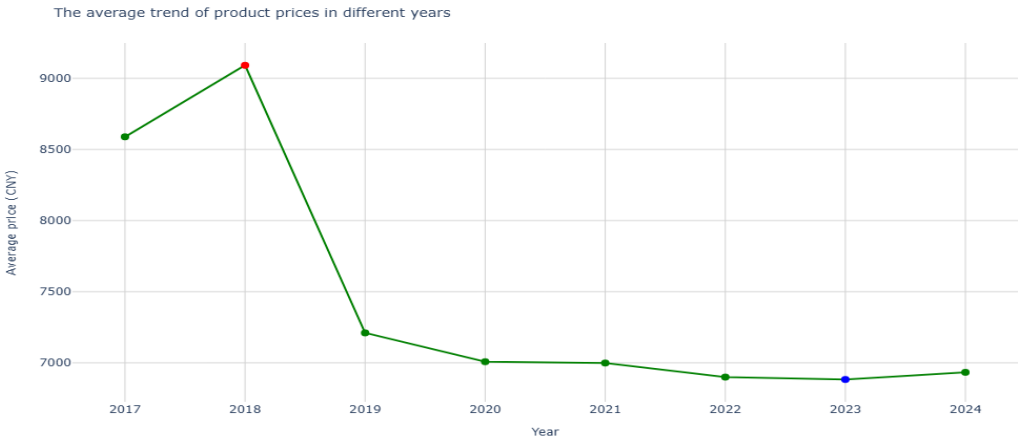
Dataset description

This dataset contains detailed specifications and official launch prices of various mobile phone models from different manufacturers. It provides valuable data for analyzing smartphone hardware, pricing trends, and brand competitiveness across multiple countries. Key features in the dataset include RAM, camera specifications, battery capacity, processor details, and screen size.

The pricing information in this dataset is particularly significant. The recorded prices represent the official launch prices of the phones at the time of their initial release. Since prices vary by country and launch period, older models reflect their original launch prices, while newer models show their most recent launch prices. As such, this dataset is highly valuable for studying price trends over time and comparing smartphone affordability across different regions.

Method and Visualization

Visualization 5:



Methods used

Firstly, group by year and calculate the average price, with the default color being green.

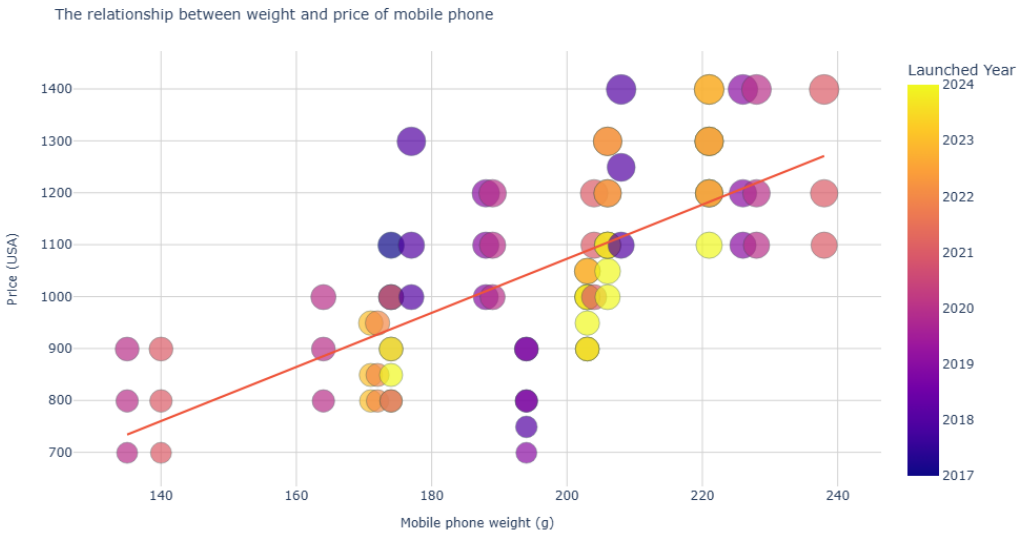
Then, use px.line to create a line chart, setting the X-axis as the year and the Y-axis as the average price. The line will be in green, while the year with the maximum average price will be marked in red and the year with the minimum average price will be marked in blue.

Finally, add a title and configure the background color and other settings to make the chart more visually appealing.

Result Analysis

From the chart, it can be seen that from 2017 to 2018, there was an increase, and then from 2018 to 2019, there was a rapid decrease, which proves that the business strategy has changed. And after 2019, the fluctuation is small, which proves that the business strategy has stabilized.

Visualization 6:



Methods used

First, perform data preprocessing by converting the phone weight, price in the USA, and price in China into floating-point numbers.

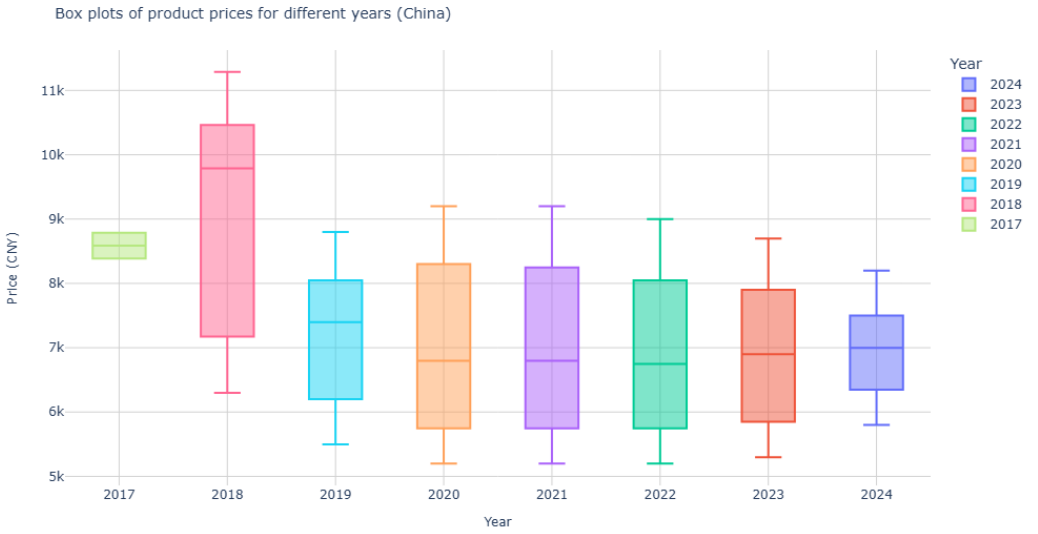
Next, use Plotly Express to plot a scatter plot, where the x-axis represents phone weight and the y-axis represents the price of the product in China. Add an Ordinary Least Squares (OLS) regression line to show the relationship between weight and price.

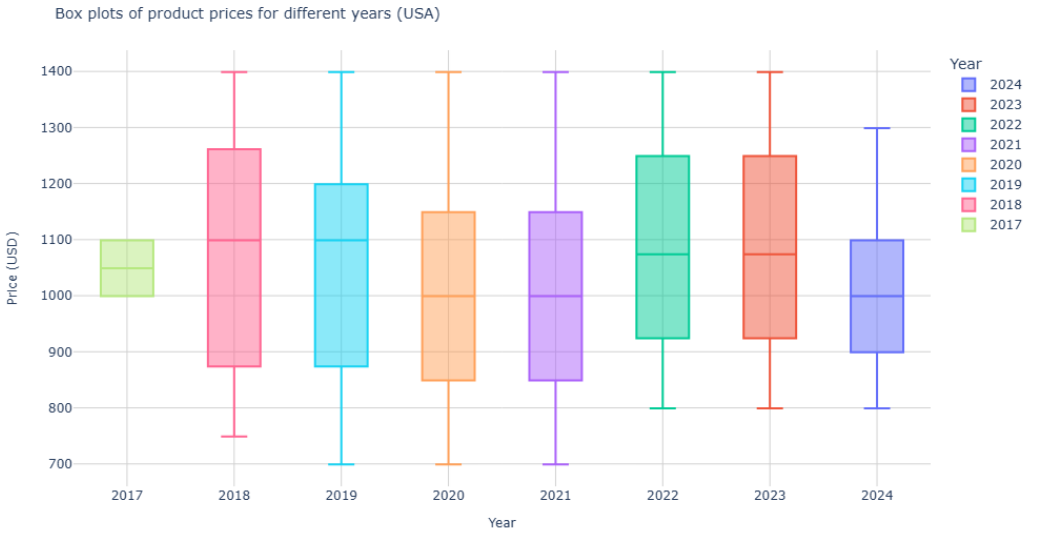
Finally, adjust the layout to achieve a visually appealing result.

Result Analysis

From the chart, we can see that there is a positive correlation between the phone's weight and its price.

Visualization 7:





Methods used

For the first one,first, use the px.box function to create box plots of product prices for different launch years in China, with 'Launched Year' as the x-axis, 'Launched Price (China)' as the y-axis, and each year represented by a different color. Next, set the title and labels for the plot, and specify the order of the categories in 'Launched Year' by sorting the unique values in reverse order

Then, use the `update_layout` function to customize the layout, including setting the background color, grid color, font sizes, and margin sizes.

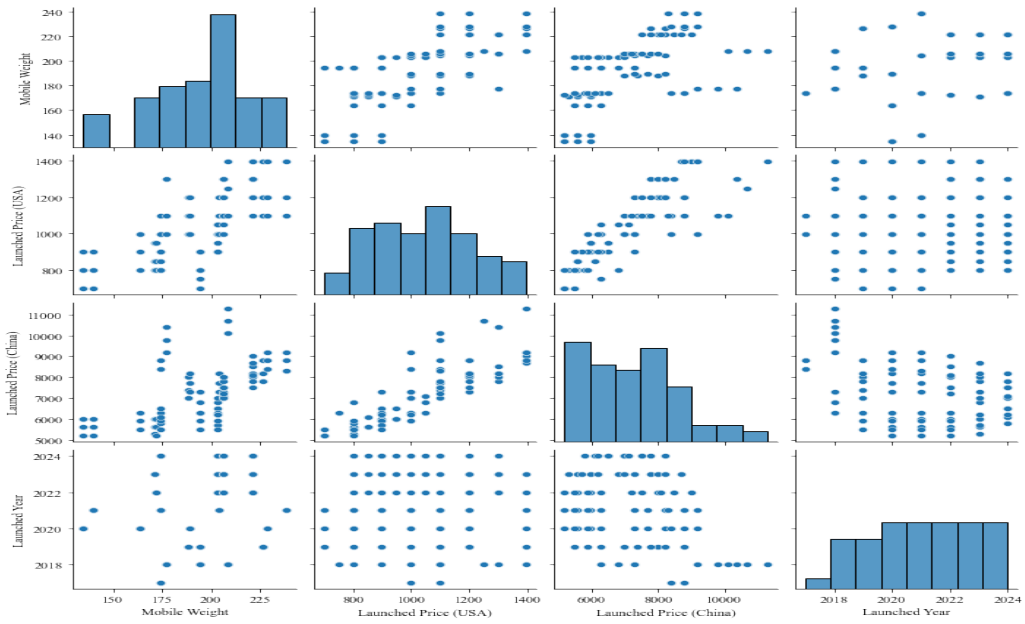
Finally, use the show function to display the plot.

And for the second one,the 'Launched Price (USA)' column is used as the y-axis.

Result Analysis

From the chart,we can see that the prices of products released in different years generally showed an upward trend, with the median increasing year by year, and the price distribution was relatively concentrated. And the price trends and distribution characteristics of products in both markets are similar over roughly the same time period

Visualization 8:



Methods used

First, use the `sns.pairplot` function from the Seaborn library to create a pair of graphs for the dataset `selected_data_2`.

Then, use `plt.show()` function to display the resulting plot.

Result Analysis

From the chart,we can see the consistency of the price of the iPhone in both markets and the difference in price and weight of the products released in different years.

Dataset 3: marriage_data_india

Introduction

This dataset explores marriage trends in India, comparing the differences between love marriages and arranged marriages across demographic, social, and economic factors.

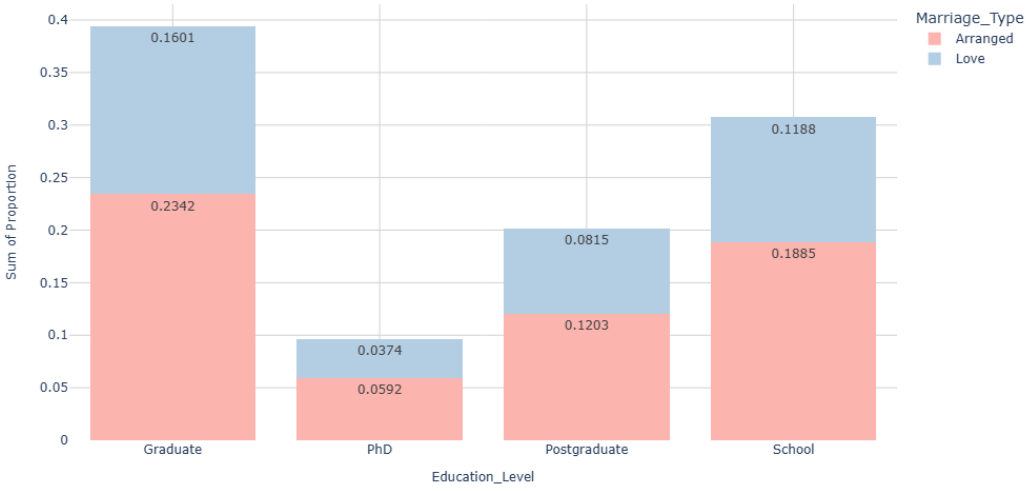
Dataset description

This dataset explores marriage trends in India, comparing love marriages and arranged marriages across various demographic, social, and economic factors. I, capturing key aspects such as age at marriage, caste and religion dynamics, parental approval, dowry exchange, marital satisfaction, divorce rates, income levels, and urban-rural differences.

The dataset aims to provide valuable insights into changing marriage patterns, the role of tradition vs. modernity, and their impact on marital outcomes. Researchers, sociologists, and data analysts can use this dataset to study relationship trends, predict marriage success, and analyze social influences on marriage in India.

Method and Visualization

Visualization 9:



Methods used

First, use the `groupby` function to group the dataset by `'Education_Level'` and `'Marriage_Type'`, then calculate the size of each group and reset the index to create a new column `'Count'`.

Next, calculate the total count of `'Marriage_Type'` and compute the proportion for each combination by dividing the count of each group by the total count. Then, create a histogram using the `px.histogram` function, setting `'Education_Level'` as the x-axis, `'Proportion'` as the y-axis, and `'Marriage_Type'` as the color.

Finally, customize the layout and style of the plot to achieve a visually appealing result.

Result Analysis

From the chart, it can be seen that the proportion of arranged marriages is higher than that of love marriages, regardless of the level of education. However, with the increase in the level of education, the proportion of arranged marriages and love marriages has decreased, although the proportion of arranged marriages is always higher than that of love marriages.

Libraries and functions used

1.Pandas (pd)
pd.cut(): Discretize continuous age range into discrete age groups
pd.crosstab(): Compute cross-tabulation
2.Matplotlib (plt)
plt.title(): Customize title
plt.grid(): Customize grid lines
plt.legend(): Customize legend
plt.show(): Display plot
3.Seaborn (sns)
sns.kdeplot(): Plot kernel density estimate
sns.histplot(): Plot histogram
sns.heatmap(): Plot heatmap
sns.barplot(): Plot bar chart
sns.pairplot(): Plot pair grid
4.Matplotlib.colors(mcolors)
Used for custom color mapping
5.Numpy (np)
Used for data processing and calculations

6.Plotly Express (px)
px.line(): Plot line chart
px.box(): Plot box plot
px.scatter(): Plot scatter plot
px.histogram(): Plot histogram
7.Plotly Graph Objects (go)
Used for creating and customizing graph objects
8.Plotly Colors (pc)
Used for custom colors

Summary of Analysis and Conclusions

In this project, we created several visualizations to analyze three datasets. The visualizations include KDE plots and histograms for heart attack risk by age group, heatmaps for heart attack risk by disease status, pie charts for heart attack risk by gender, bar charts for heart disease patients by state, line charts for average mobile phone prices by year, scatter plots for phone weight vs. price in China, and histograms for marriage trends in India. From these visualizations, we concluded that heart attack risk is consistent across age groups, diabetes is a significant risk factor, and regional disparities exist in India. Mobile phone pricing strategies have changed over the years, and there is a positive correlation between phone weight and price. In India, arranged marriages are more common than love marriages, with both types decreasing as education levels increase.