Jupyter Notebook Homework 1.1:

Creating Data Visualizations

Due: 04th March 2025

Individual Contribution			
CWID	**Name**	**Contribution (description)**	**Percent Contribution**
A20563440	Yanxi Pu	Collect dataset1 data, draw dataset1 images, and write a report for dataset1	33.3%
A20563462	Hengyi Li	Collect dataset2 data, draw dataset2 images, and write a report for dataset2	33.3%
A20563447	Hanshen Yu	Collect dataset3 data, draw dataset3 images, and write a report for dataset3	33.3%

I. Description of the dataset

The dataset we use comes from the website — Kaggle.com/datasetsLinks to an external site.
In order to meet the requirements of this assignment, we used 3 dataset. The links are as follow:

1.kaggle.com/datasets/himelsarder/cinema-hall-ticket-sales-and-customer-behavior
2.kaggle.com/datasets/tejas14/student-final-grade-prediction-multi-lin-reg
3.kaggle.com/datasets/marcelobatalhah/quality-of-life-index-by-country

The first dataset records detailed information about ticket sales and customer behavior at a cinema hall,offering data about demographics, movie genre, ticket price, and the number of customers travelling together. From this data, we can infer the viewing needs and preferences of customers of different age groups within a certain range.

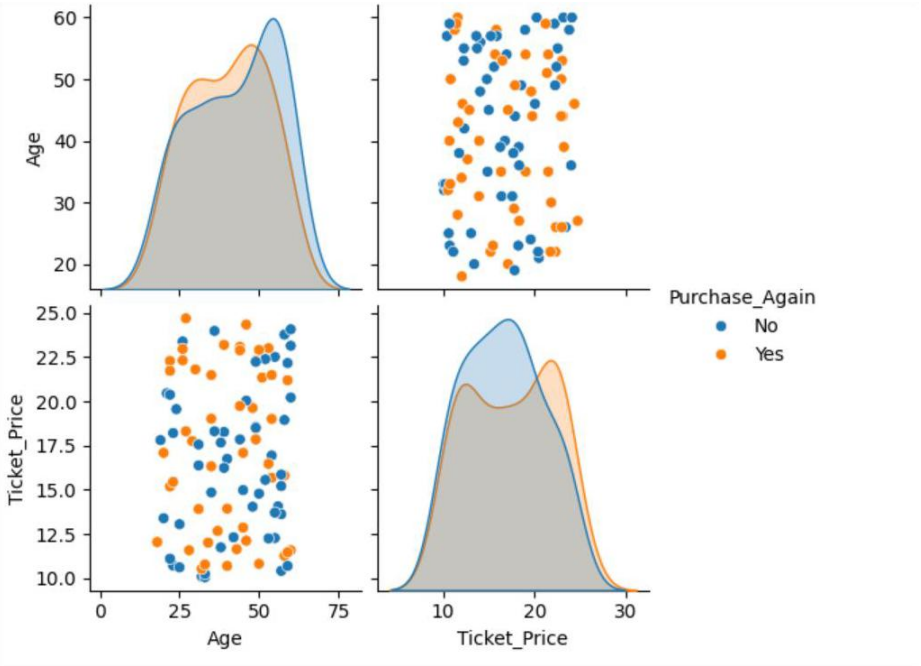
The second dataset records detailed information about students' personal information as well as their learning, providing data on basic information, school information, and learning. From this data, we can infer the learning of students within a certain range.

The third dataset records the details of the quality of life index for each country, providing data such as quality of life index, purchasing power index, safety index and health index. This data helps make decisions such as choosing a place to live or analyzing global trends in quality of life.

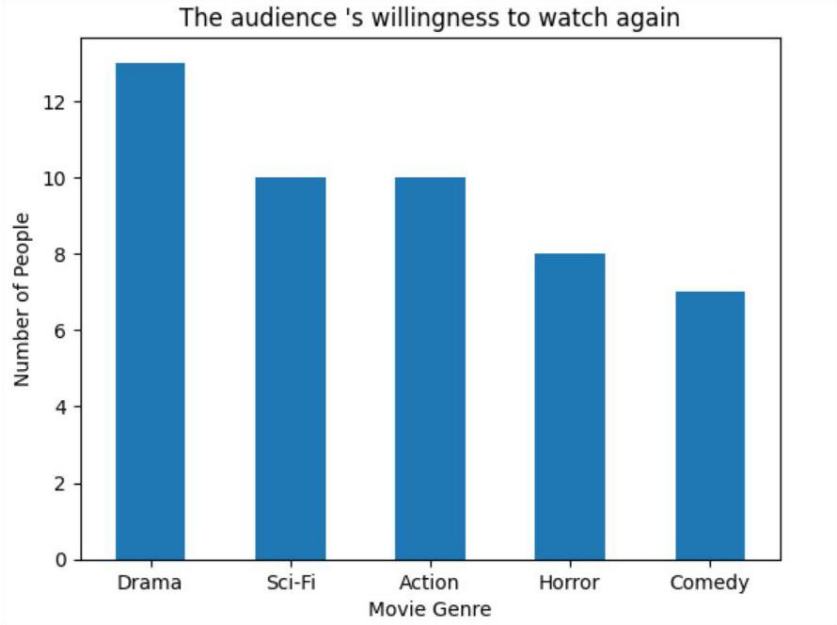
II. Visualizations

1. Cinema Hall Ticket Sales and Customer Behavior

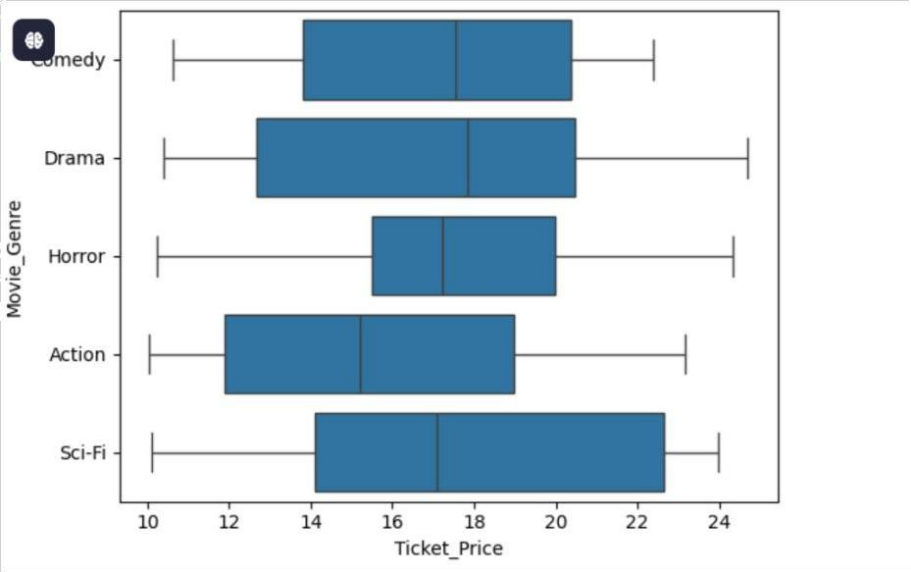
i. Analysis of the relationship between a customer's willingness to watch a movie again and customer's age and the ticket price



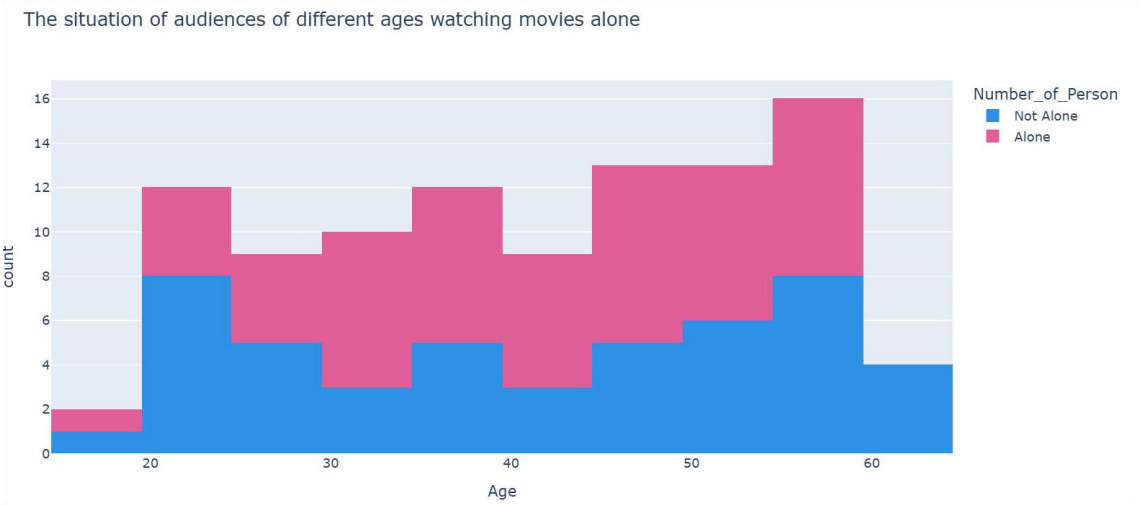
ii. Analysis of the relationship between a customer's willingness to watch a movie again and movie genre



iii. Analysis of the relationship between ticket price and movie genre

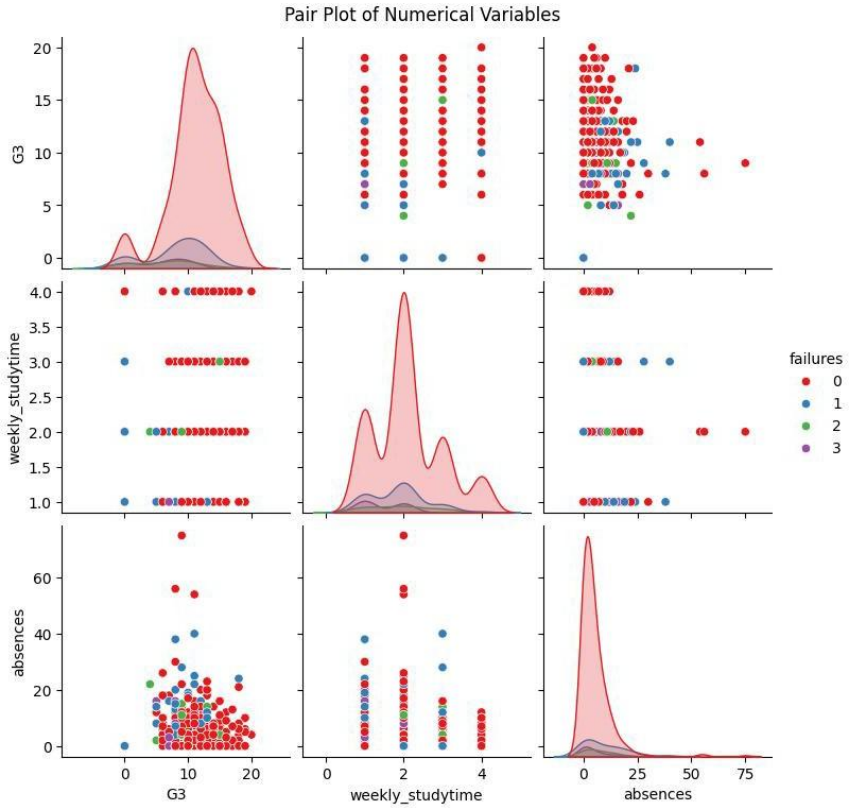


iv. The situation of audiences of different ages watching movies alone

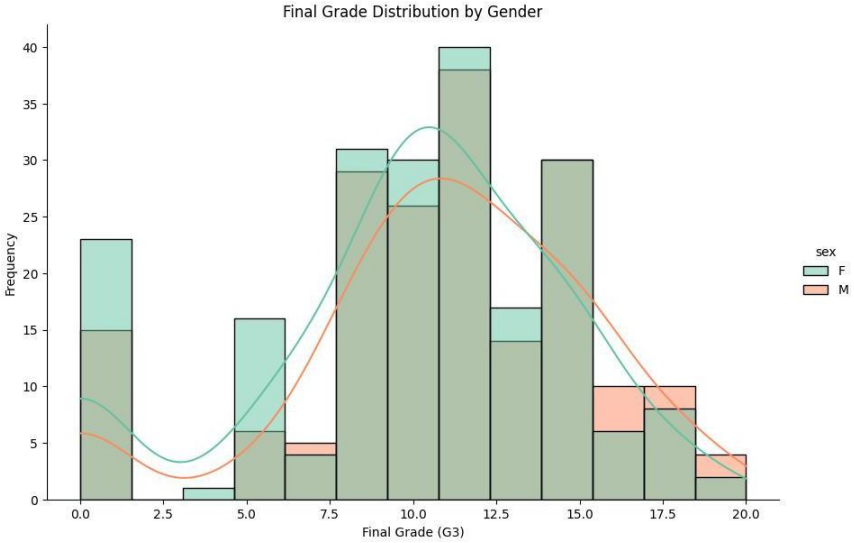


2. Student Final Grade Prediction

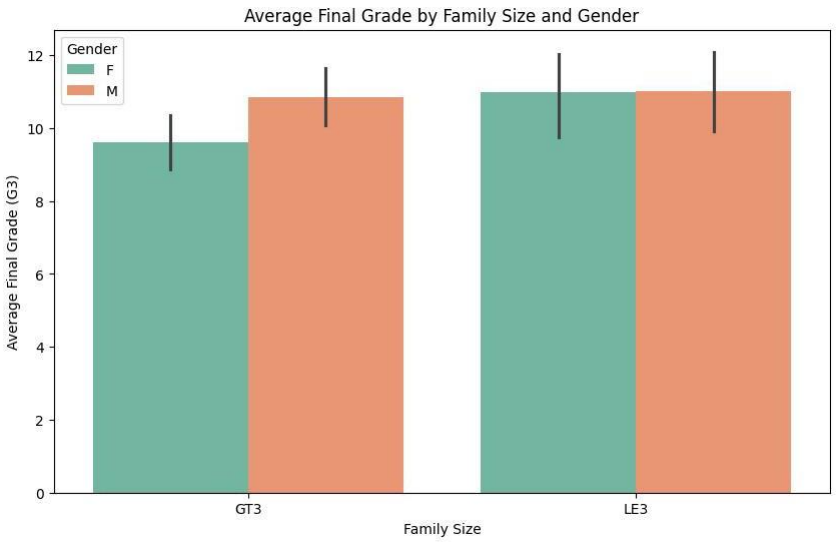
i. The analysis shows the relationship between the different numerical variables in the dataset



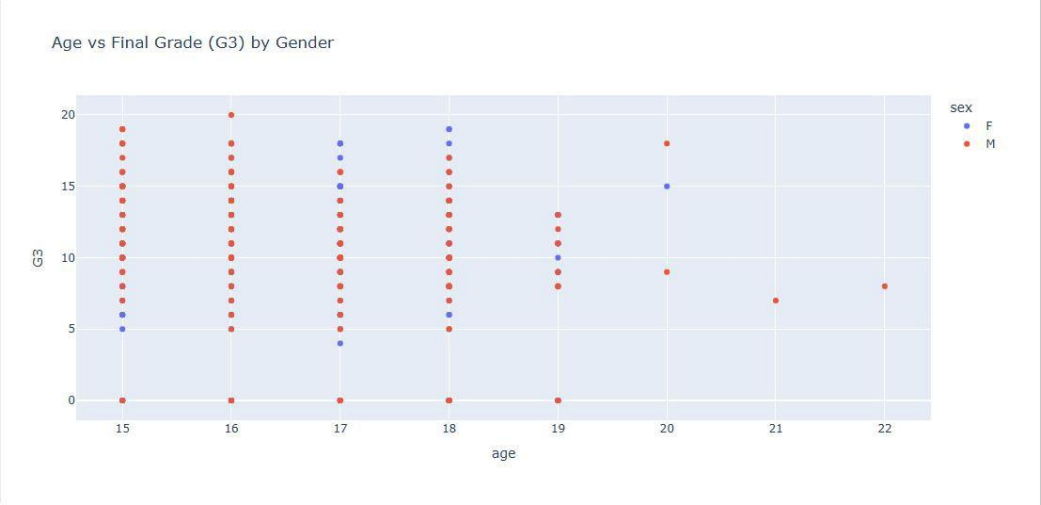
ii. The distribution of Final Grades (G3) grouped by gender is shown



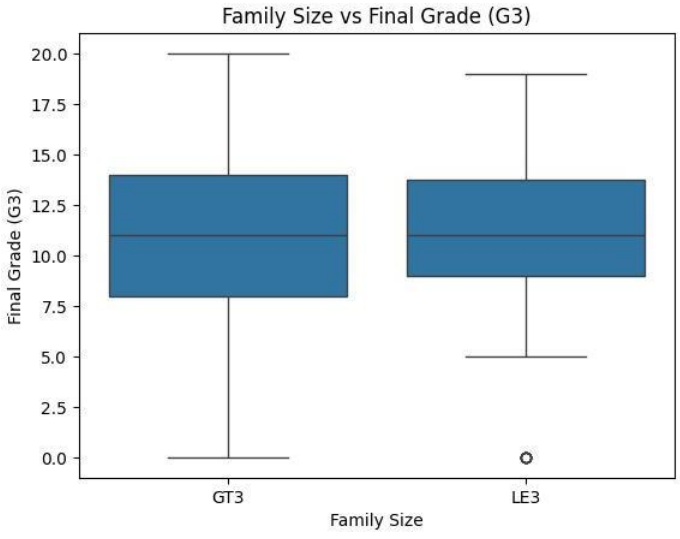
iii. A comparison of students' average final grade (G3) based on family size and gender is shown



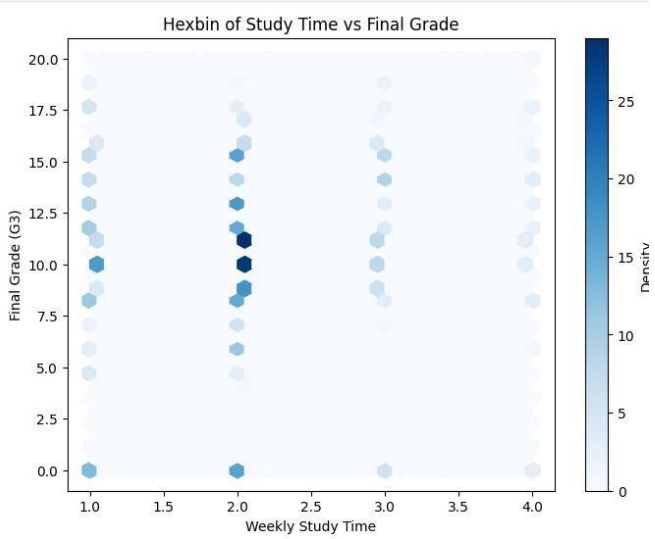
iv. The relationship between age and final grade (G3), separated by gender



v. The relationship between household size and final grade (G3), represented by a box plot

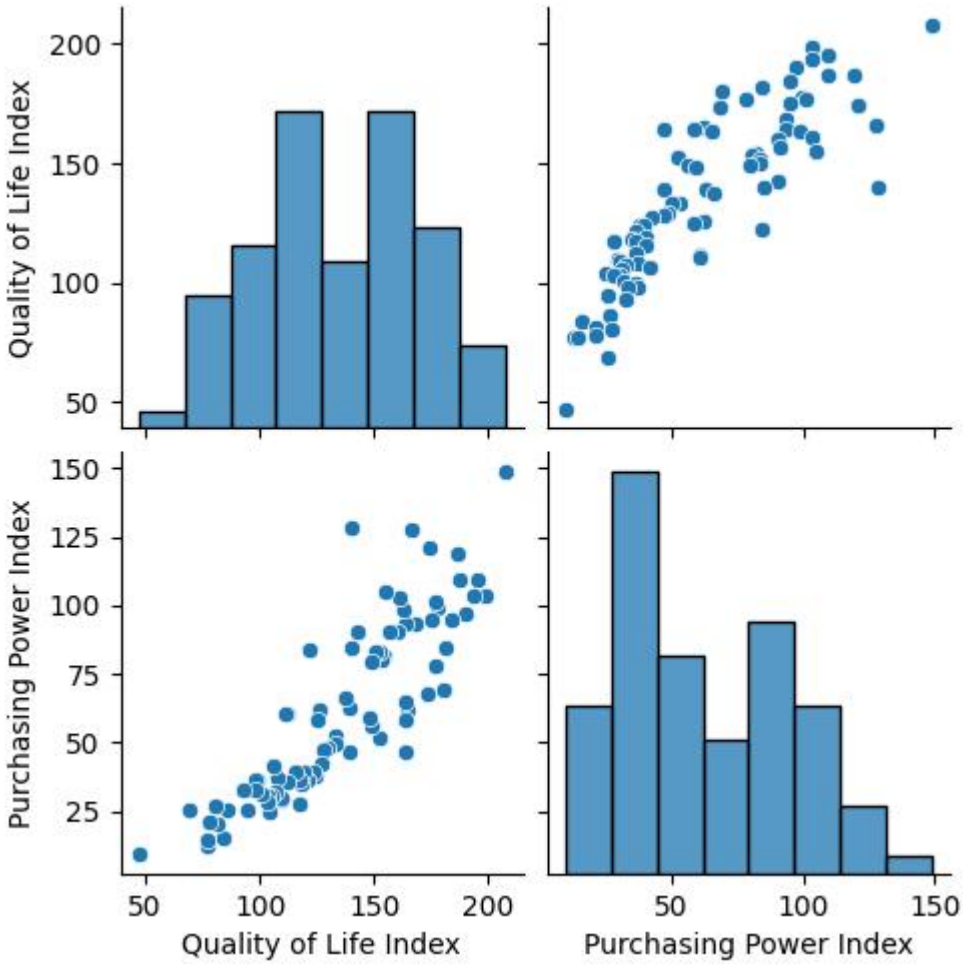


vi. The relationship between the number of hours of study per week and the final grade (G3) is used in a hexagonal box plot

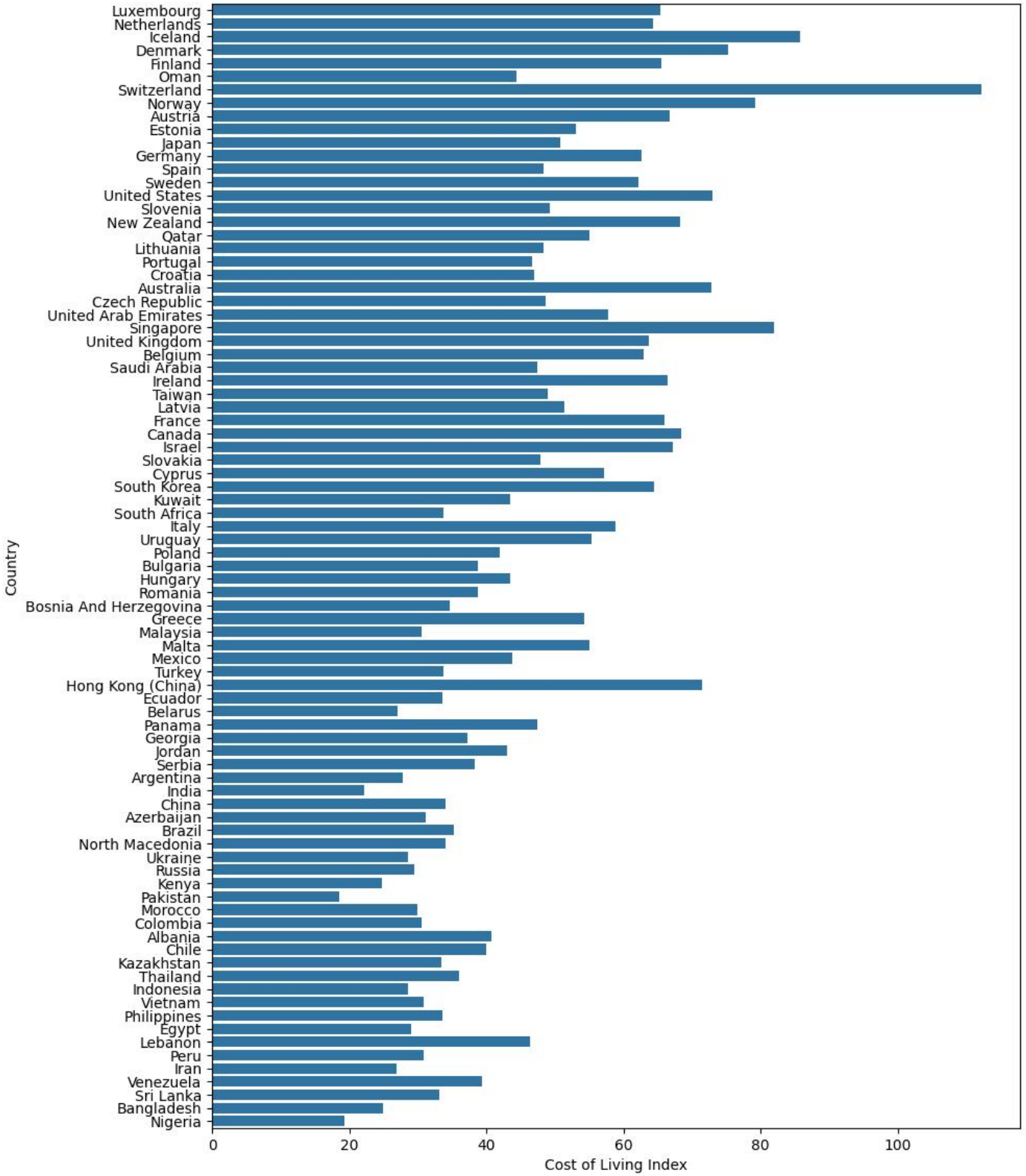


3. Quality of life Index by country

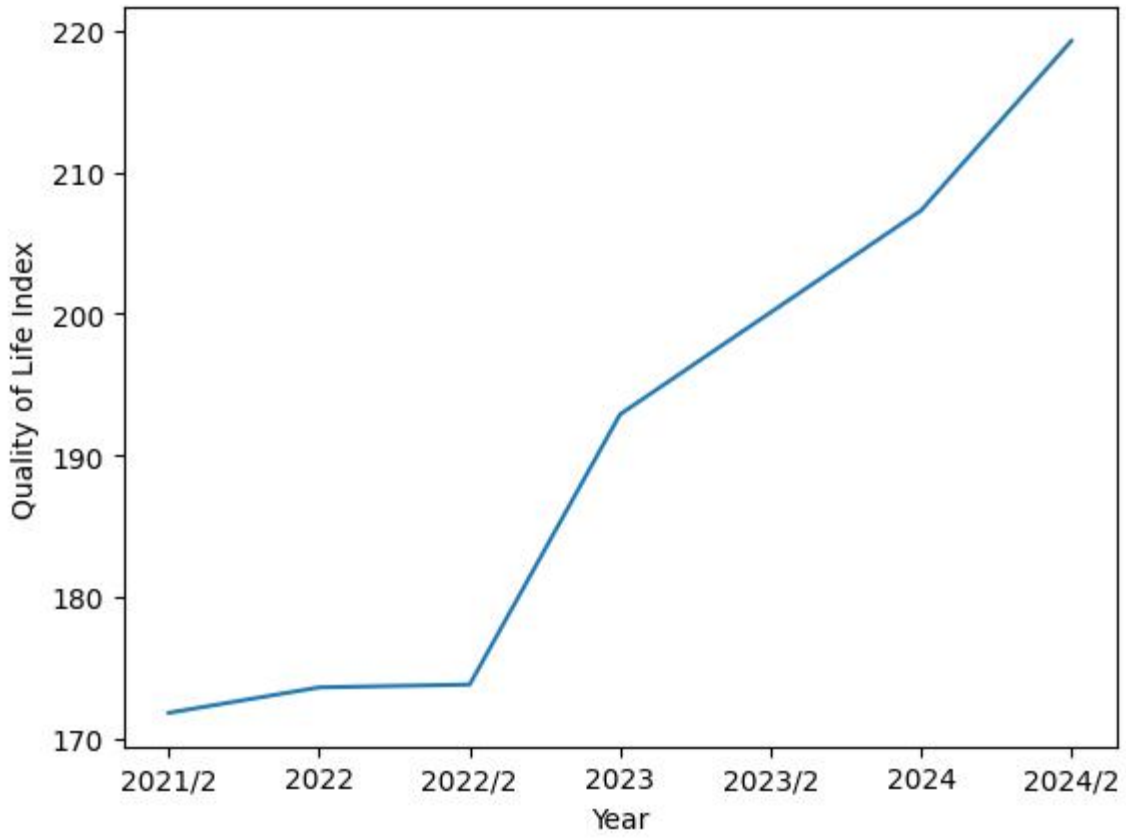
i. Relationship between quality of life index and purchasing power index



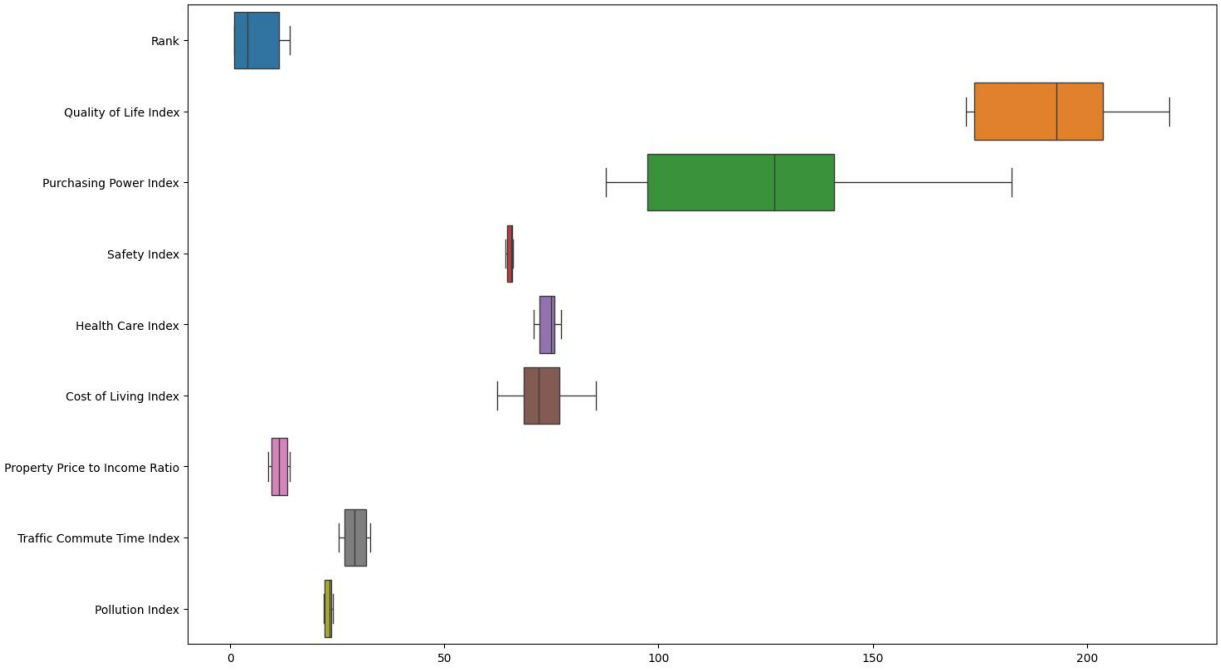
ii. The cost of living index for each country



iii. Changes in the Quality of life index in Luxembourg



iv.Box plots of the indices for Luxembourg



III. Method

1.Cinema Hall Ticket Sales and Customer Behavior

To analyze the data in this dataset, we used the Seaborn library to plot pairwise plots and box plots(by using sns.pairplot() and sns.boxplot()), the pandas library to plot histograms (by usindf.plot(kind='bar')), and the plotly library to plot two-color histograms(by using px.histogram()).

2.Student Final Grade Prediction

To analyze the data in this dataset, we plotted pairwise plots, two-color histograms and box plots using the Seaborn library (by using sns.pairplot()sns.barplot() and sns.boxplot()), scatter plots using the plotly library (plt.hexbin()), and two-color histograms using the plotly.express library (by using px.scatter()).

3.Quality of life Index by country

To analyze the data in this dataset, we used the Seaborn library to create curve pairs, bar plots, line plots, and box plots (by using sns.pairplot (), sns.barplot(), sns.lineplot (), and sns.boxplot ()).

IV. Libraries

pandas, os, plotly, matplotlib, seaborn

We use pandas library to get data, clean data and plot histograms, use plotly library to plot two-color histograms, use os library to find the csv file, use matplotlib library to draw charts and modify labels, use seaborn library to plot pairwise plots.

V. The results

1. Cinema Hall Ticket Sales and Customer Behavior

i. Analysis of the relationship between a customer's willingness to watch a movie again and customer’s age and the ticket price

From the curve in the upper left corner of this diagram, we can obviously find that Customers between the ages of 20 and 30 are less likely to watch a movie again, while customers between the ages of 40 and 50 are more willing to watch a movie again.

From the scatter chart in the lower left corner, we can see that customers are more willing to buy movie tickets in the lowest price range (10~12.5) or the highest price range (21~25) again.

From the curve in the lower right corner, we find that even many people who buy lowest price range (10~12.5) is willing to buy their movie ticket again, more people who buy ticket in this price do not want to watch again.

In the scatter plot in the upper right corner, we can see that the price of the movie ticket that the customer chooses to buy is independent of the customer's age

ii. Analysis of the relationship between a customer's willingness to watch a movie again and movie genre

In the histogram, the drama and sci-fi genres are more recognized among

audiences, with customers who are willing to buy tickets again than those who are willing to buy tickets again, while horror and comedy are relatively low.

iii. Analysis of the relationship between ticket price and movie genre

In this box chart, we can see that the median ticket price for action movies is the lowest, around 15, and the box is smaller, indicating that the range of ticket price changes is smaller. The median ticket price for plays is slightly higher, between 17 and 18. The wider box indicates that the ticket price fluctuates greatly. Horror movies have the smallest box, indicating that ticket prices are the least likely to fluctuate. Medians for comedy, horror, drama, and sci-fi movies are all similar, but sci-fi movies have significantly larger boxes and more fluctuating ticket prices.

iv. The situation of audiences of different ages watching movies alone

In this two-color bar chart, we can see that customers in the age group of 20 to 30 and over 60 are more likely to choose to watch movies alone, while customers in other age groups may be more likely to watch movies with family and friends.

2. Student Final Grade Prediction

i. The analysis shows the relationship between the different numerical variables in the dataset

This pairwise plot shows the correlation between the number of failures, learning time, the number of absences, and the final grade, and color-codes the distribution of students with different failures across these variables, revealing some interesting trends and correlations.
Observations:
G3 (final grade): Students with higher G3 values tend to fail less often (predominantly red). As the number of failures increases (from blue to green to purple), the value of G3 usually decreases gradually.
Learning time per week: This variable has a strong correlation with the number of failures. Students who spend less time studying tend to fail more.
Number of absences: There is a clear relationship between the number of absences and failures, and students who miss more classes tend to have more failures, especially with some outliers with higher absences.
This graph shows that students who spend less time studying and miss more classes are more likely to fail, while students with better grades tend to study longer and miss fewer classes.

ii. The distribution of Final Grades (G3) grouped by gender is shown

The number of students in the lower band (close to 0) is relatively small, with a similar distribution of male and female students.
In the middle band (close to 10 points), the number of female students is higher, while the number of male students is gradually increasing.
In the higher band (15 points and above), the distribution of females and males gradually widens, with a higher number of male students.
As can be seen from this graph, there are some differences in the distribution of female and male students in the final grade distribution, especially in the middle to high score range, where the number of male students is higher.

iii. A comparison of students' average final grade (G3) based on family size and gender is shown

GT3 (more than 3 family members):
The average grade point of female students is slightly higher than that of male students.
LE3 (number of family members less than or equal to 3):
There is little difference in the average grade between female and male students.
As can be seen from the graph, when there are more family members (greater than 3), the average final grade of female students is slightly higher than that of male students. When the number of family members is small (less than or equal to 3), the average score of male and female students is almost equal.

iv. The relationship between age and final grade (G3), separated by gender

As can be seen from the figure, there are female and male students in each age group (15 to 22 years old), and there is no obvious age dependence on the grade (G3), and the students' grades do not show a clear pattern with age.
The distribution of female students (blue dots) and male students (red dots) was staggered in age, and there was no obvious tendency to cluster.
The distribution of age and achievement shown in the graph is relatively discrete and does not show a strong correlation.

This graph shows that there does not appear to be a clear linear relationship between age and final grade (G3), and that male and female students are evenly distributed across age groups.

v. The relationship between household size and final grade (G3), represented by a box plot

The median final score of a GT3 (with more than 3 family members) was approximately 12.5 points, and the distribution was relatively wide, indicating a large difference in achievement. Students with LE3 (3 or less family members) also had a median final score of about 12.5, but the distribution of their grades was narrower, indicating that the grades were more concentrated.

As can be seen from the box plot, family size has less effect on final grade (G3), and the median grades of students in the two categories are not much different and have a similar distribution.

vi. The relationship between the number of hours of study per week and the final grade (G3) is used in a hexagonal box plot

In the graph, students who study less time per week (about 1 hour and 2 hours) have lower grades (G3) and have relatively more points in these areas.
The results of students who study 2 to 3 hours per week are more scattered, but there is still a certain density, showing a correlation between study time and performance.
For students who studied 4 hours per week, the distribution of grades (G3) was more concentrated in the lower grade range, indicating that there may be other factors affecting the grades.

As you can see from the graph, there is a certain correlation between the weekly study time and the final grade. Most students study less per week and have lower grades, while students with average study time have a larger distribution of grades.

3. Quality of life Index by country

i. Relationship between quality of life index and purchasing power index

Through this figure, we can see that the quality of life index has a strong correlation with the purchasing power index, and when the purchasing power index shows an upward trend, the quality of life index also shows an upward trend, that is, the quality of life index is positively correlated with the purchasing power index

Ii.The cost of living index for each country

Through this figure, we can intuitively see the difference between the living cost index of different countries, and also see whether the ranking of the quality of life index and the living cost index are related. It can be seen from the figure that the ranking of the quality of life index has no obvious relationship with the living cost index

Iii.Changes in the Quality of life index in Luxembourg

Through this graph, we can see that the quality of life index of Luxembourg has been on the rise, from which we can see the development of Luxembourg in recent years

Iv.Box plots of the indices for Luxembourg

Through this graph, we can see the distribution of the quality of life index, purchasing power index, safety index, health care index, cost of living index, house price to income ratio, transportation commute time index, pollution index and climate index of Luxembourg, and see the changes of each index in Luxembourg

END