From In-Depth Analysis to 3D Visualization:

Understanding Tourism in China and South Korea Through Data

Table of contents:

Introduction:

This documentation presents the development and outcomes of an interactive 3D data visualization and data analysis centred on visitor arrivals to China and South Korea.

The project's primary objective was to create an engaging and intuitive visualization that enables users to explore complex tourism datasets. It highlights various demographic factors, such as age, gender, country of origin, and crew numbers, all presented in a 3D interactive format.

The project began with an analysis of existing data to understand tourism's economic impact on China and South Korea, including developing machine learning models to predict future trends.

This documentation details the research, development, technical solutions, and reflections on the project, showcasing its creation and the insights gained.

Research:

The research phase of this project involved extensive exploration of data visualization techniques and tools, with a focus on identifying effective methods for representing multi-dimensional data interactively.

A significant aspect of the research was understanding the demographics and trends of visitor arrivals in China and South Korea. This involved analyzing how different demographic factors such as age, gender, country of origin, and crew numbers of impact tourism patterns.

Additionally, I delved into the economic impact of tourism on China and South Korea. The tourism industry is a crucial component of national economic growth, contributing significantly to GDP, creating jobs, and driving development in related sectors. Understanding the economic implications of tourism helped in contextualizing the data and highlighting its importance. This research aimed to provide insights into how fluctuations in visitor numbers can affect the economies of these countries and underscore the importance of the tourism industry in national economic strategies.

Through this comprehensive research, I aimed to ensure that the final visualization not only accurately represents the data but also conveys the broader economic context and significance of tourism for China and South Korea.

Ideas:

The research phase of this project began with the idea of analyzing existing data in various ways to understand how tourism has impacted China and South Korea so far. This initial analysis aimed to identify trends and patterns in visitor arrivals and to assess the economic implications of tourism on these countries. Recognizing the significant role that the tourism industry plays in national economic growth, the research focused on how different demographic factors such as age, gender, country of origin, and job category (crew) influence tourism patterns.

In addition to the retrospective analysis, the project included the development of machine learning models to predict future tourism trends. This involved using historical data to train predictive models, providing insights into how tourism numbers might change in the future and helping policymakers and businesses prepare for these changes.

Following this comprehensive analysis, the idea of creating a final artwork emerged. The objective was to share the findings in a dynamic and interactive 3D visualization, allowing users to explore tourism statistics and appreciate the importance of the tourism industry. This visualization aimed to represent different dimensions of the data in an intuitive and visually appealing manner, enhancing user engagement and providing clear insights into the data and its economic significance.

Objective:

The primary objective of this project was to create an interactive 3D data visualization that allows users to explore and understand tourism data for China and South Korea. The visualization aimed to make complex datasets accessible and engaging, highlighting various demographic factors such as age, gender, country of origin, and crew numbers.

The project sought to achieve the following goals:

- Engagement: Develop a visually appealing and interactive 3D visualization that captures users' interest and encourages exploration of the data.
- Intuitiveness: Design an interface that is easy to navigate, allowing users to intuitively interact with and understand the data.
- Insightfulness: Provide meaningful insights into the tourism trends and patterns for China and South Korea, demonstrating the economic impact and importance of the tourism industry.
- Predictive Analysis: Incorporate machine learning models to predict future tourism trends, offering valuable foresight for policymakers and businesses.
- Accessibility: Ensure the visualization is accessible to a broad audience, including those with colour vision deficiencies, by using a colourblind-friendly palette.

Through these objectives, the project aimed to deliver a comprehensive tool for exploring tourism data, enhancing user understanding of the significance and implications of tourism for China and South Korea.

Data Analysis and Cleaning:

The data analysis and cleaning phase was a crucial step for understanding the datasets used in various analyses, including visits by country/region, gender, age, crew visits, and predictive modeling.

This phase involved several key steps to ensure the accuracy and relevance of the data:

- Data Collection: Datasets related to visitor arrivals in China and South Korea were sourced from various reliable websites. These datasets included information on visitor numbers segmented by age, gender, country of origin, and crew numbers.
- Initial Analysis: An initial analysis was conducted to understand the structure of the collected data, identify any missing values or inconsistencies, and gain an overview of tourism trends and patterns.
- Data Cleaning: The datasets underwent a thorough cleaning process to remove any irrelevant or erroneous entries. This involved handling missing values, standardizing data formats, and ensuring consistency across different datasets. The aim was to retain only the necessary and accurate data for further analysis.
- Data Transformation: The cleaned data was transformed into new CSV files, organized in a way that facilitated subsequent analyses. This included restructuring the data to be compatible with the analytical tools and processes used in the Jupyter Notebooks.
- Validation: The final step involved validating the cleaned and transformed data to ensure it accurately reflected the original datasets. This validation was essential to maintain the integrity and reliability of the data used in the analyses.

Furthermore, after these datasets were used for detailed analyses in the Jupyter Notebooks, the same data was utilized for the final artwork outcome - the 3D visualization. However, additional data cleaning was required to ensure the data was perfectly suited for the 3D visualization, emphasizing the importance of precision and accuracy in the final representation.

The meticulous data analysis and cleaning process ensured that the datasets were accurate, relevant, and structured appropriately for in-depth analyses. This foundational work enabled detailed examinations of visitor demographics and trends, including visits by country/region, gender, age, and crew, as well as predictive modeling to forecast future tourism trends.

Development - Jupyter Notebook analysis:

For the analysis of visits by country/region, gender, age, and crew visits, data was collected from various reliable sources. The data sources provided comprehensive information necessary for the detailed analysis and visualizations. Here are the websites where the data was sourced:

South Korea - https://kosis.kr/statHtml/statHtml.do?orgId=150&tblId=TX_15002_A015&vw_cd=MT_ETITLE&list_id=H2_10&scrId=&language=en&seqNo=&lang_mode=en&obj_var_id=&itm_id=&conn_path=MT_ETITLE&path=%252Feng%252FstatisticsList%252FstatisticsListIndex.do

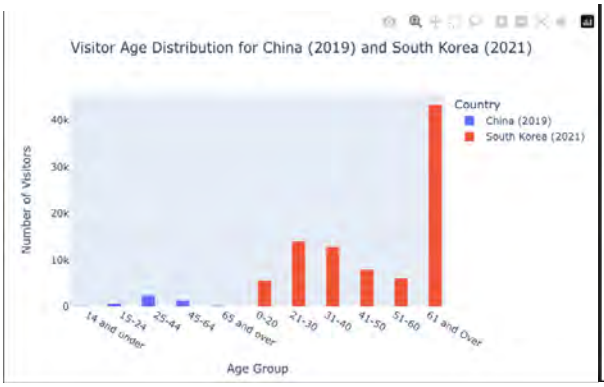
China Yearbooks - https://www.stats.gov.cn/english/Statisticaldata/yearbook/

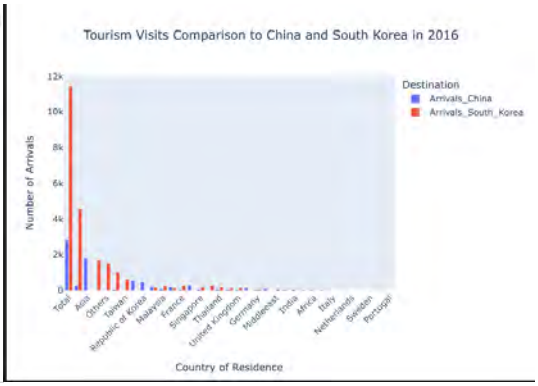
1. Visits by Country / Region

The objective was to visualize and compare the number of tourist visits to China and South Korea in 2016 based on different regions or countries of origin. The development process involved loading, inspecting, and cleaning the datasets, merging them for comparison, and creating a dynamic bar chart using Plotly Express and pandas for data manipulation. The visualization provided clear insights into regional visit patterns, and future work could include trend analysis over multiple years.

Failures and Experiments:

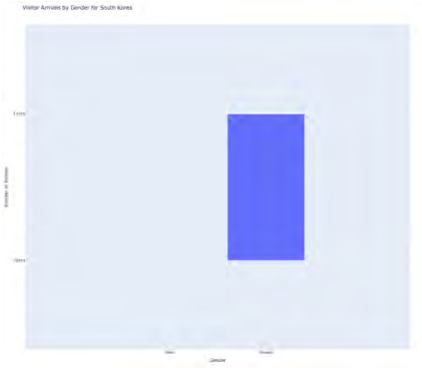
- Initial issues with data column mismatches were resolved by standardizing column names.
- Some regions had missing data, which required data cleaning steps.





2. Visits by gender

This analysis aimed to visualize the number of tourist visits to China and South Korea by gender. The development steps included loading, inspecting, and summarizing the data by gender, followed by creating a bar chart using Plotly Express and pandas for data manipulation. The gender-based analysis revealed interesting patterns in tourist demographics, highlighting the differences between the two countries.



Failures and Experiments:

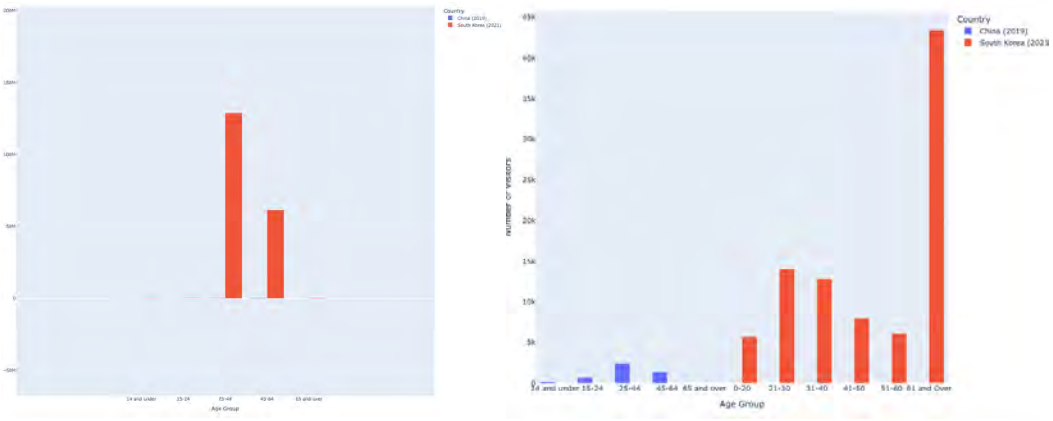
- Initial data inspection revealed missing gender data in some entries, which required cleaning.

3. Visits by age

The objective was to analyze and visualize the number of tourist visits to China and South Korea by age group. The process involved loading and inspecting the datasets, summarizing the data by age group, and creating a bar chart using Plotly Express and pandas for data manipulation. The age-based analysis provided insights into the demographic trends in tourist visits, although aligning age categories required careful mapping.

Failures and Experiments:

- Mismatched age group categories required careful mapping and alignment.



4. Crew visits

This analysis focused on comparing the number of crew visits to China and South Korea. The development steps included loading and inspecting the datasets, extracting relevant data for crew visits, and creating a pie chart using Plotly Express and pandas for data manipulation. The analysis revealed significant differences in crew visits between the two countries.

```
China Visitors Data Columns:
Index(['Item', '2019'], dtype='object')

China Visitors Data Sample:
           Item     2019
0         Total  4911.36
1          Male  2881.29
2        Female  2030.07
3  14 and under   184.92
4         15-24   686.20

South Korea Visitors by Age Data Columns:
Index(['Year', '2021', '2021.1', '2021.2', '2021.3', '2021.4', '2021.5',
       '2021.6', '2021.7'],
      dtype='object')

South Korea Visitors by Age Data Sample:
                Year   2021 2021.1 2021.2 2021.3 2021.4 2021.5       2021.6  \
0  By nationality(2)  Total   0-20  21-30  31-40  41-50  51-60  61 and Over
1              Total  90150   5677  14043  12866   8002   6135         3456
2              China  11691    341   1437   1902   1184   1323          669
3              Japan   1007     86    235    173    184    123           58
4             Taiwan    290      7     55     47     28     12           10

  2021.7
...
 8   2021.7  62 non-null     object
dtypes: object(9)
memory usage: 4.5+ KB
None
```

Failures and Experiments:

Initial extraction of crew data required careful inspection and validation.

5. Predictive Modeling

The goal of this project was to develop a predictive model to forecast international tourism receipts for China and South Korea.

The process involved data cleaning, preprocessing, model training, and evaluation using a dataset from the World Bank, focusing on the series "International tourism, receipts (current US$)". The dataset contains annual tourism receipts for China and South Korea from 1995 to 2020.

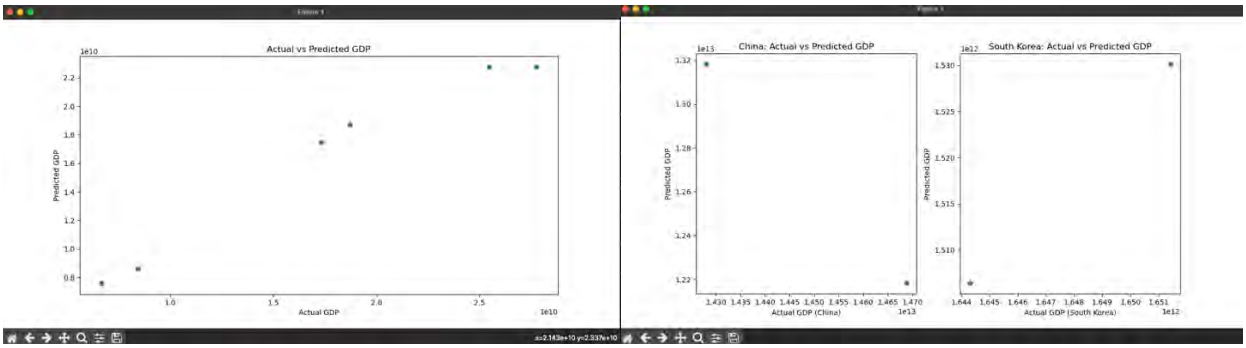
Here is a link:
https://databank.worldbank.org/reports.aspx?source=2&series=ST.INT.RCPT.CD&country=KOR#

The project utilized tools like Pandas, Numpy, Matplotlib, and Scikit-Learn in Python. Key steps included loading the data, transforming it, selecting relevant features, training a Linear Regression model, and evaluating its performance.

Failures and Experiments:

1. Initial Data Issues: Encountered issues with missing data and incorrect formats, which required extensive cleaning.

Given the constraints and missing features in the dataset, the final model was simplified to use only the available "International tourism, receipts (current US$)" feature for China. Other potential predictive features like "Tourist Arrivals" and "Inflation Rate" were excluded due to data unavailability.

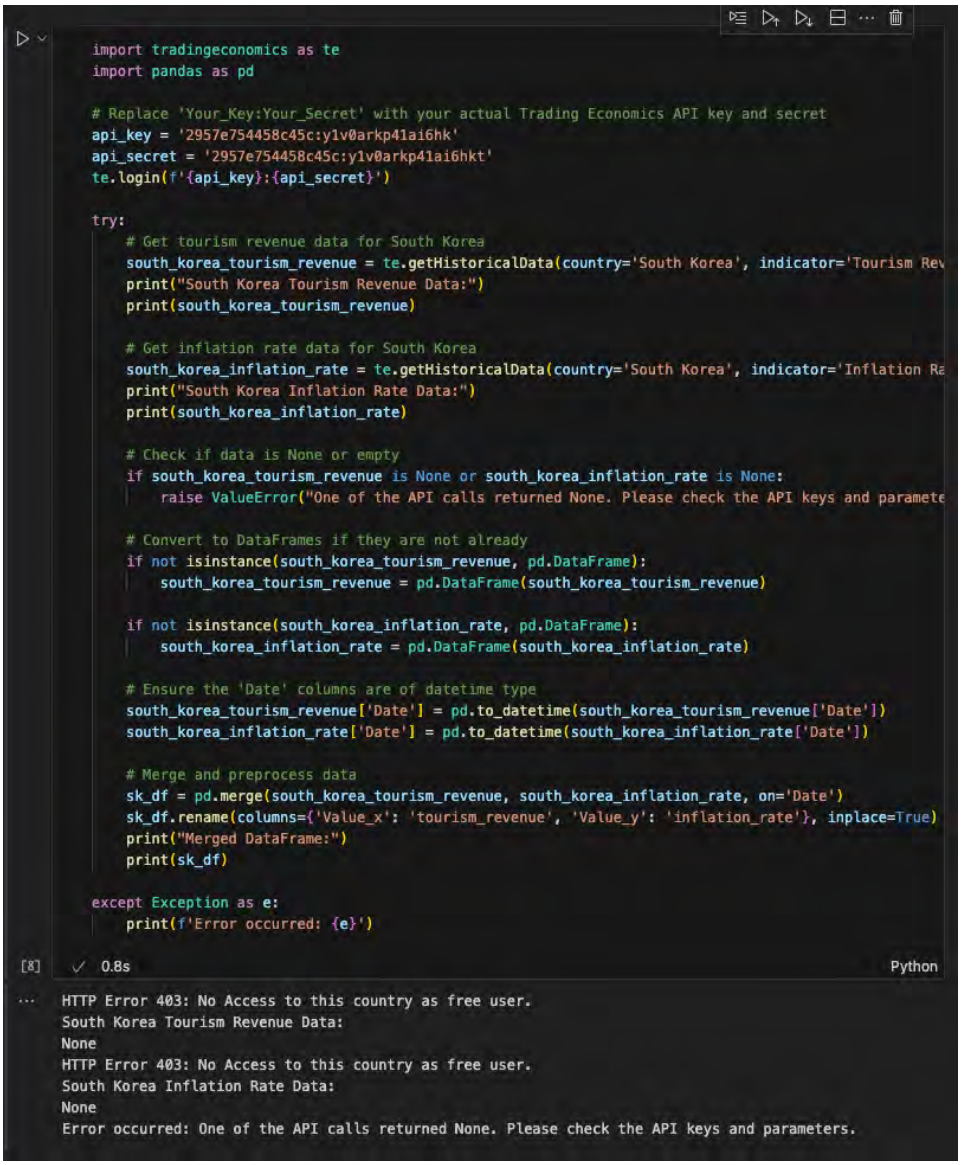


2. API Access Issues:

Initially, I attempted to use the Trading Economics API to gather data on tourism revenue and inflation rates for South Korea. However, due to access restrictions and the API rejecting my free user credentials, I was unable to obtain the necessary data. Despite contacting the website, I request for access was denied, leading me to source the data from an alternative CSV file.

https://tradingeconomics.com/south-korea/tourism-revenues
https://tradingeconomics.com/south-korea/inflation-cpi
https://tradingeconomics.com/china/inflation-cpi

Symbol	Ticker	Name	Country	Date	State	Last	Close	CloseDate	Group	URL	Importance	DailyChange
MEXBOL:IND	MEXBOL	IPC	Mexico	6/12/2024 8:26:00 PM	CLOSED	52975.890000000000	52975.890000000000	6/12/2024 12:00:00 AM	America G20 Major	/mexico/stock-market	170	-158.170000000
S30:IND	S30	Stockholm	Sweden	6/12/2024 3:30:59 PM	CLOSED	2627.115570000000	2627.115570000000	6/12/2024 12:00:00 AM	Europe	/sweden/stock-market	230	40.15300000000
SET50:IND	SET	SET 50	Thailand	6/12/2024 10:03:00 AM	CLOSED	812.740000000000	812.739990000000	6/12/2024 12:00:00 AM	Asia	/thailand/stock-market	350	2.820010000000
NZSE50FG:IND	NZSE50FG	NZX 50	New Zealand	6/13/2024 2:22:56 AM	OPEN	11817.140000000000	11817.140000000000	6/13/2024 2:23:00 AM	Australia Major	/new-zealand/stock-market	570	75.26000000000
			"Free accounts have access to the following countries: Mexico, New Zealand, Sweden, Thailand. For more, contact us at support@tradingeconomics.com."									



```
import tradingeconomics as te
import pandas as pd

# Replace 'Your_Key:Your_Secret' with your actual Trading Economics API key and secret
api_key = '2957e754458c45c:y1v0arkp41ai6hk'
api_secret = '2957e754458c45c:y1v0arkp41ai6hkt'
te.login(f'{api_key}:{api_secret}')

try:
    # Get tourism revenue data for South Korea
    south_korea_tourism_revenue = te.getHistoricalData(country='South Korea', indicator='Tourism Rev
    print("South Korea Tourism Revenue Data:")
    print(south_korea_tourism_revenue)

    # Get inflation rate data for South Korea
    south_korea_inflation_rate = te.getHistoricalData(country='South Korea', indicator='Inflation Ra
    print("South Korea Inflation Rate Data:")
    print(south_korea_inflation_rate)

    # Check if data is None or empty
    if south_korea_tourism_revenue is None or south_korea_inflation_rate is None:
        raise ValueError("One of the API calls returned None. Please check the API keys and paramete

    # Convert to DataFrames if they are not already
    if not isinstance(south_korea_tourism_revenue, pd.DataFrame):
        south_korea_tourism_revenue = pd.DataFrame(south_korea_tourism_revenue)

    if not isinstance(south_korea_inflation_rate, pd.DataFrame):
        south_korea_inflation_rate = pd.DataFrame(south_korea_inflation_rate)

    # Ensure the 'Date' columns are of datetime type
    south_korea_tourism_revenue['Date'] = pd.to_datetime(south_korea_tourism_revenue['Date'])
    south_korea_inflation_rate['Date'] = pd.to_datetime(south_korea_inflation_rate['Date'])

    # Merge and preprocess data
    sk_df = pd.merge(south_korea_tourism_revenue, south_korea_inflation_rate, on='Date')
    sk_df.rename(columns={'Value_x': 'tourism_revenue', 'Value_y': 'inflation_rate'}, inplace=True)
    print("Merged DataFrame:")
    print(sk_df)

except Exception as e:
    print(f'Error occurred: {e}')
```

[8] ✓ 0.8s Python

```
HTTP Error 403: No Access to this country as free user.
South Korea Tourism Revenue Data:
None
HTTP Error 403: No Access to this country as free user.
South Korea Inflation Rate Data:
None
Error occurred: One of the API calls returned None. Please check the API keys and parameters.
```

Future Work:

1. Enhanced Feature Set: Future iterations of this project would benefit from a richer set of features, including socio-economic factors, more detailed tourism data, and macroeconomic indicators.
2. Model Comparison: Exploring different modeling techniques and comparing their performance could yield more insights and potentially better predictions.
3. Access to Premium APIs: Securing access to comprehensive datasets through premium APIs or institutional partnerships could significantly enhance data quality and model accuracy.

Development – Artwork:

For this 3D artwork web visualization, I utilized the data sources identified during the analysis phase. To ensure the accuracy and relevance of the data for the 3D visualization, I created new .csv files and performed data cleaning to retain only the necessary data.

The final outcome of this project is an interactive 3D data visualization artwork created using p5.js, HTML, CSS, and JavaScript. This artwork visualizes visitor arrival data for China and South Korea, providing a comparative analysis based on age, gender, country of origin, and crew number. The goal was to provide an intuitive and visually appealing way for users to explore visitor data by different demographics, enhancing their understanding of the trends and patterns in international tourism.

Why p5.js? For example, why Processing was not chosen instead?

p5.js and Processing are both excellent tools for creating interactive visualizations, but they have distinct differences that can influence the choice of one over the other based on project needs.

p5.js was chosen for this project primarily because it is web-based and runs directly in the browser. This makes it easy to share the project online, as the visualizations can be embedded directly into a documentation and website. p5.js uses JavaScript, a widely used language that integrates well with web development, further facilitating ease of sharing and enhancing interactivity for web-based projects.

However, p5.js does have some limitations. It may not perform as well as Processing for very complex or heavy visualizations, and its 3D capabilities are not as advanced as those offered by Processing.

Processing, on the other hand, is known for its superior performance, particularly for complex visualizations and computations. It offers strong support for 3D graphics and more advanced rendering features and is often easier for beginners to get started with visual programming. However, Processing runs as a desktop application, which makes it less straightforward to share projects online. Additionally, it uses Java, which might be less familiar to those more comfortable with web technologies.

In summary, p5.js was chosen for its web-based nature, ease of sharing, and seamless integration with web development, despite its performance limitations compared to Processing.

The `index.html` file sets up the basic structure of the web page. It includes the p5.js library and links to the `sketch.js` file, where the main visualization code is implemented. Basic styling is applied to ensure the web page is formatted correctly.

The `sketch.js` file contains the p5.js code that creates the 3D visualization. The key functions and processes in the file include:

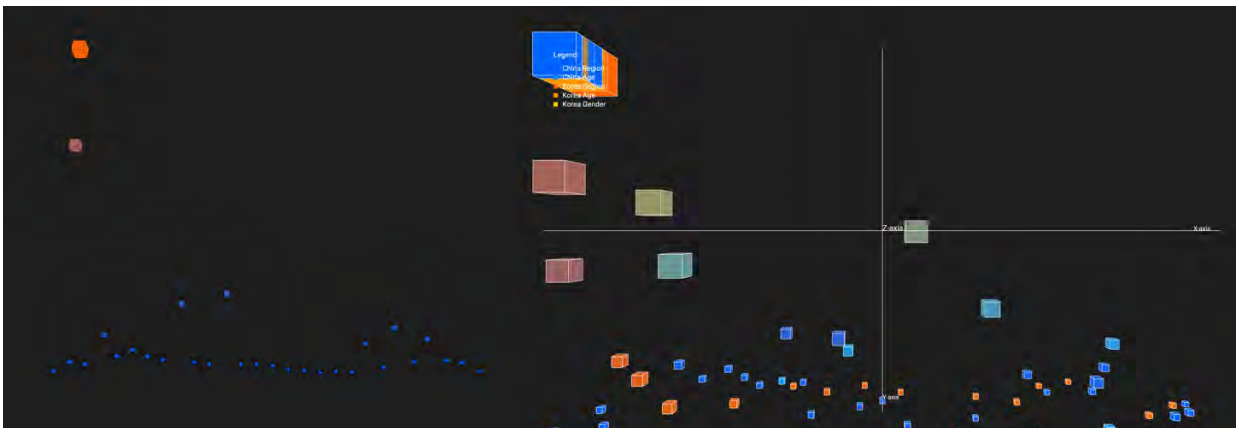
- Preloading Data: The `preload()` function loads the necessary CSV files and font before the sketch starts, ensuring all data and resources are available for use.
- Processing Data: Separate functions process each dataset, parsing the data, mapping values to appropriate ranges, and creating data points with specific positions, sizes, and colors.
- Interactivity: The `mouseDragged()` and `mouseWheel()` functions enable user interaction, allowing users to rotate and zoom the 3D view.

Accessibility:

A colorblind-friendly palette was chosen to ensure the visualization is accessible to colorblind viewers. Colors were carefully selected to be distinguishable, making the visualization usable by a broader audience.

Failures and Experiments:

1. Font Loading Issues: Encountered issues with loading the font initially, which were resolved by ensuring the correct file paths and formats.
2. Data Parsing: Some datasets had unexpected formats, leading to NaN values. This was handled by adding checks and validations.
3. Legend Visibility: Initially, the legend was not clearly visible, and its placement was adjusted for better visibility.



Final outcome:

Section 1: Analysis

The analysis phase involved using Jupyter Notebooks to examine various aspects of visitor arrivals to China and South Korea.

Here are the key analyses conducted:

- Visits by Country/Region (region_country_visits.ipynb):

This notebook analyzed the number of tourist visits to China and South Korea based on different regions or countries of origin. The data was visualized using dynamic bar charts, allowing for a clear comparison of visits from various regions.

- Visits by Gender (visitors_by_gender.ipynb):

This notebook analyzed the gender distribution of tourists visiting China and South Korea. The data was summarized and visualized to reveal interesting patterns and differences in visitor demographics based on gender.

- Visits by Age (visitors_by_age.ipynb):

This analysis examined the number of tourist visits categorized by age groups. By summarizing and visualizing the data, the analysis provided insights into the age demographics of visitors, highlighting trends and patterns in tourism behavior.

- Crew Visits (visitors_by_crew.ipynb):

This analysis focused on the number of crew visits to China and South Korea. The data was extracted, summarized, and visualized using pie charts, providing a clear comparison of crew visits between the two countries.

- Predictive Modeling (Predictive_Modeling.ipynb):

This analysis focused on developing a predictive model to forecast international tourism receipts for China. By using historical data and applying machine learning techniques, the model aimed to predict future trends in tourism, providing valuable insights for policymakers and businesses.

These analyses were essential for understanding the various factors influencing tourism in China and South Korea. The insights gained from these analyses informed the development of the final 3D visualization artwork.

Section 2: Artwork

The final artwork is an interactive 3D data visualization that is designed to showcase the impact of tourism on China and South Korea, allowing users to explore the data dynamically.

Users can interact with the visualization by rotating and zooming (use your mouse or trackpad), making it an engaging and intuitive way to explore the data.

The 3D visualization highlights various demographic factors, such as age, gender, country of origin, and crew numbers. Each data point is represented in a 3D space, with different colors and sizes indicating different values. This visual representation helps users understand the significance of tourism data and its impact on the economies of China and South Korea.

By interacting with the 3D visualization, users can gain insights into tourism patterns and trends, appreciating the importance of the tourism industry. The use of a colorblind-friendly palette ensures that the visualization is accessible to a broad audience, making the data exploration experience inclusive and informative.

Overall, the final artwork successfully translates complex datasets into an engaging and interactive visual format, effectively communicating the impact of tourism on China and South Korea.

You can access the GitHub repository to view the analysis section, which includes all the code and data related to my project.

Additionally, the repository features a web-based 3D visualization artwork and the rest necessary features being used.

Here is a link: https://github.com/Aki120900/big_data_torusm_China_and_S.Korea

Check README file (there will be PNG images of each analysis conducted using Jupyter Notebooks)

Here is a link to 3D visualization artwork website:
https://aki120900.github.io/big_data_torusm_China_and_S.Korea/

Reflections:

From a research perspective, the project emphasized the critical role of tourism in national economic growth. The ability to visualize and predict tourism trends provided valuable insights for policymakers and businesses, highlighting the broader economic impact of the tourism industry. The project demonstrated how data-driven approaches could inform strategic decisions and promote sustainable tourism development.

Overall, this project showcased the power of combining data analysis with advanced visualization techniques to communicate complex information effectively. It illustrated the importance of accessible, interactive tools in enhancing understanding and decision-making in the context of tourism economics.

Conclusion:

This project successfully created an interactive 3D data visualization that provides valuable insights into tourism data for China and South Korea. The final artwork allows users to explore visitor demographics and trends dynamically.

The project highlighted the importance of thorough data preparation, including cleaning and pre-processing, to ensure accurate and relevant datasets. Additionally, developing machine learning models offered useful predictions for future tourism trends.

Overall, the project underscored the critical role of data visualization in effectively communicating complex information. The engaging and intuitive design of the 3D visualization makes it a powerful tool for understanding tourism's impact, offering valuable perspectives for stakeholders in the tourism industry.

References:

Guo, China. (2017). *BRICS joint statistical publication. 2017 / BRICS joint statistical publication. 2017*. 中国统计出版社, Beijing Shi: Zhongguo Tong Ji Chu Ban She.

Oosterhaven, J. and Fan, T. (2006). Impact of international tourism on the Chinese economy. *International Journal of Tourism Research*, 8(5), pp.347–354.

Simpson, J. (n.d.). *Travel & Tourism Economic Impact 2023 South Korea*. [online] World Travel & Tourism Council. Available at: https://assets-global.website-files.com/6329bc97af73223b575983ac/64874f04facb1e1073df3c1b_EIR2023-SouthKorea.pdf To download reports or data, please visit: ResearchHub.WTTC.org .

Songhong, G. (2002). Measuring the Economic Impact of Tourism in China. *Forum of International Development Studies /『国際開発研究フォーラム』*. [online] Available at: https://www.gsid.nagoya-u.ac.jp/bpub/research/public/forum/21/04.pdf.

Surugiu, C. (2009). The Economic Impact of Tourism. An Input-Output Analysis. *DOAJ (DOAJ: Directory of Open Access Journals)*.

UN.ESCAP (1993). The economic impact of tourism in Republic of Korea. *repository.unescap.org*. [online] Available at: https://hdl.handle.net/20.500.12870/2839 [Accessed 14 Jun. 2024].