HILTON

https://www.hilton.com/en

BUSINESS INTELLIGENCE

Krish Doshi
Adrian Marroquin
Hemalatha Ramakrishna

TABLE OF CONTENTS

Executive Summary

The goal of our project consisted of providing aid to Hilton Hotels in determining the next country to develop and construct a new hotel or franchise its license to. To decide on a particular country, several variables were considered that were related to the hospitality and tourism industry. The country selected ultimately had to help increase Hilton's revenue as well as keep consistent on their mission on "building the premier global hospitality business". By acquiring and building their property portfolio, the company stands to increase its brand awareness and to increase and expand its' international market share. The team will analyze data acquired from several sources to gain insight by visualizing data and finding a recommendation that Hilton can consider based on the insights from the dataset.

As a result, the project goal could put Hilton on a path to continue increasing their property portfolio and annual revenues.

Hilton

HOTELS & RESORTS

Opportunities:

- Increase Revenue
- Further International Expansion.
- Increase Market Share.

Challenges:

- High Amount of Debt
- International Market Expansion
- Need Flexible due to its large size.

Business Understanding

Based on the Harvard Business Case, the Hilton Hotels has experienced exponential growth over the past century, becoming a recognizable global brand among hotel visitors. In 2007, the company property portfolio consisted of 2,906 properties in 87 countries (Harvard Business Case) and in 2020, it has double its size to 6,215 total properties in 118 countries (Wikipedia).

The company mission statement is to "To fill the earth with the light and warmth of hospitality by delivering exceptional experiences – every hotel, every guest, every time." In order to deliver on that promise, Hilton Hotels has invested heavily on information systems and in 2004 developed and deployed a proprietary technology platform called "OnQ" which utilizes technology to enhance guest recognition and efficiency. OnQ has been critical to the success of Hilton's aggressive expansion strategy allowing them to replicate customer experience to new locations in a shorter amount of time. Despite the company's success and the millions spent on the technology, the company has faced issues in standardizing the use of the technology due to the design of the system. Furthermore, Hilton Hotels has been experiencing high employee turnover, which in result, has increase cost to the company having to retrain employees to effectively utilize the system.

Despite its challenges, Hilton Hotels have been very successful domestically and has been slowly expanding internationally to increase its market share and revenues. The company's main competitor is the Marriot Hotels who implement a similar strategy as Hilton but lacking a strong presence in all hotel segments (luxury, economy, family, etc.). Hilton must be diligent in their process of selecting the country in which the next hotel will be located as many factors can determine its success or failure.

Business Objectives

- **Increase Revenue**: By acquiring more property portfolio, the company will increase its annual revenue.
- **Increasing International Market Share:** Selecting an international country to expand to
- **Reduce Employee Turnover**: Focusing on factors that affect employee retention

Business Success Criteria

The analysis of Hilton hotels can help the company in international expansion. This analysis helps them increase their revenue so that they can increase their international market expansion by analyzing data. Furthermore, this helps them Identifying specific countries to expand market and Focusing on Tourist attraction and pattern

- **Country Selection:** Identifying one country with very appealing factors to Hilton
- **Identifying Insights:** Discovering how countries compare to the world
- **Making Recommendations**: Being able to recommend a country with data driven decisions

Assess the Situation

Risks and contingencies

- **Covid 19 –** The 2020 pandemic has affected almost all industries including the tourism industry. Data provided for the year will not reflect on previous years, yielding results that may not be accurate if/once conditions stabilize
- **Lodging alternatives –** Businesses like Airbnb and similar local lodging alternatives can affect the data being collected. It is unknown if and how much these substitutes can affect the overall tourism industry

- **Data Source –** The data source located were incomplete and actions were taken to fill in any missing values to complete the dataset, it is possible that data sources have different values
- **Variable Determination –** It can only be assumed that the variables chosen will impact the Hilton Hotels decision to select a country for expansion. Other variables may hold stronger value that were not included in the dataset
- **Lack of Country Data –** Only a small subset of countries were found that contain complete information. Most of the countries were from Europe. It is possible that our recommendations would have changed if more country data were available.

Data Analysis Goals

For Data Analysis it would be by analyzing these datasets so that we can forecast where visitors will spend the most money and recommend where Hilton should expand next to increase their international revenue.

- ## Data Analysis.
    - ### Description

We will assess the number of tourist arrivals, population size, crime rate, literacy rate, female and male ratio, international and domestic revenue, etc. in different countries from the year 1995 – 2020 and by depending on that we can determine the highest-ranking and lowest-ranking country among the variables

- ### Dependency

We will analyze the correlation between crime rate and literacy rate, the correlation between international revenue and tourist arrivals, the correlation between average salary and average working hours, the correlation between population size and number of people employed in the tourism industry and the correlation between average salary and literacy rate.

- **Prediction**

We can predict the future crime rate of countries based on historical crime data. Our goal is to forecast to determine whether crime will increase, decrease or remain constant in a country. By doing so, we can select a country where visitors will feel safe and therefore, they are more likely to return

Business Process Model

To understand the process of Hilton's business process, we need to understand what is business process model. A business process is the logical arrangement of events that lead to a specific result relevant to your association. Furthermore, it demonstrates the practice of back-engineering each process in your business to comprehend the various parts expected to accomplish the objective and track down potential improvements for every part. Now, let's start with how Hilton's business process model achieves its objective.

To start with, Hilton will need seven teams, each having its tasks to follow:

1. **Development Team**

- **Identifying Goal and Requirements Initial Process:** The development team's responsibility is to create goals and requirements for the project that will align with the tasks of other teams
- **Analyzing Data Report from Data Analyst:** After the initial process, they will need a report containing data according to the goals and requirements of the data analysts to help them analyze what further steps are needed
- **Evaluation of Construction report:** The development team will select sites (potential countries) to build the branch and analyze the report. At this point comes the construction team that will evaluate and handle sites and builds the branch. After site evaluation according to the list, the construction team sends a report of potential sites back to the development team to align them with goals and requirements. Once the potential site satisfies goals and requirements, they will create a concept design. The process goes back and forth between the development team,

construction team, and engineering department for the resources needed to build Hilton's infrastructure

- **Creation of Initial Cost Plan:** The construction and engineering department will make a list of resources they need that will help in building the branch, and the development team will make an initial cost plan, which the finance team will analyze
- **Implementation of Business Plan:** A company's business plan defines its objectives and plans to achieve its goals. A business plan lays out a composed walkthrough for the firm from financial, marketing, and operational standpoints, then sent to the program committee for further implementation.

2. **Data Analyst Team**

- **Assessing the Situation:** The data analyst team will assess the situation keeping in mind all the risks and threats while preparing the data and tasks according to goals and requirements.
- **Creating Data Analysis Goals:** Their task is to prepare what type of data is needed to align and support the business objective. The type of data they have to decide on is descriptive, dependent, predictive, or all three
- **Understanding the Data:** Understanding data is the process where they will start with data exploration. The team will go through all attributes needed to visualize it so it's easy to understand and implement. Once data understanding is complete, they start with data gathering, describing data, checking the quality, and preparing the data for cleaning
- **Data Cleaning:** The next part is to remove all the data problems they identified in their gathered data by data cleaning techniques. Data problems like duplicate values, missing/null values, inaccurate data, unnecessary columns, date formatting, etc. After cleaning, visualization comes next
- **Data Visualization:** With the clean data in hand, the data analyst will frame questions according to the business objective and analyze which factors help them depict proper visualizations. For

example, the most likely factors are population, revenue (domestic and international), crime rate, literacy rate, etc.

- **Final Data Report:** With all the information they need, the data analyst team will create a detailed report on data and visualization with specific factors highlighted. Therefore, as a result, the report will be sent back to the development team to identify further the sites that show potential and one to the sales/marketing team to analyze the profit they can earn.

3. **Finance Team**

- **Checks Project Cost Plan:** The finance team's responsibility is to maintain and keep the project's budget in check by not going overboard or underboard with its spending. It checks for the resources needed by the construction, and engineering teams analyze whether it can be affordable or not depending on the team's budget allowance.
- **Correlates with Sales and Marketing Analysis**: In addition, the finance team also checks the report from the sales and marketing team that has highlighted areas indicating profit insights. Using these insights, the finance team will correlate them with the project cost plan and conclude whether to approve or deny it
- **Decision Making:** Approving and denying the project plan plays a significant part because if the project cost plan has some flaws, the whole cost plan needs to be revised, which should be according to the budget specification. In contrast, if accepted, the next step is sending the acknowledgment report to the development team to create a business plan
- **Create a Final Budget Plan:** The final process of the finance team is to make a final budget plan for the program committee to design the infrastructure and hire interior designers for work. In addition, the hotel program's opening with a schedule.

4. **Construction Team:**

- **Site Evaluation and Documentation Report:** The site evaluation is where Hilton contacts a construction team in each country to analyze the location and whether the site is suitable. Is

the field solid? Are there any legal bindings? The site area/square feet? Etc. After the evaluation, each construction team will send the report to the development team to analyze further

- **Concept Design Analysis Decision:** Concept Design indicates the design of interactions, experiences, processes, and strategies. It involves understanding people's needs and how to meet them with products, services, and processes. The team will review all aspects, and they will be sending a decision on whether to move forward with this concept design
- **Pre Construction Study:** It's the process where the construction team creates the list of resources required at the time of development of the branch on the site. They will create a brief overview of all machines, tools, and items required to construct the base design
- **Interior Designing and Documentation:** Interior design is essential because even if the outside of the building looks good, there's no guarantee the inside would be as great. For it to be appealing to the customers, designing the interior with the help of an interior designer goes into a process where the program committee will help with selecting a potential interior designer to work with the construction team
- **Start Construction:** After the final project plan comes in with all the approval and budget plan, the construction team will develop the building according to the pre-construction study for the building skeleton structure. Construction will take a couple of months or years, considering natural factors. Once the construction work is over, the program committee's job is to start with the pre-opening of the Hotel.

5. **Sales and Marketing Team:**

- **Check For Profit in Data Report:** The sales marketing team will review the report sent by the data analyst team and analyze where they gain much more profit and revenue. They will need to check what factors they need to do by marketing in the potential countries to increase their sales revenue

- **Highlighted Profit Report to Finance Department:** Once the potential country's factors research is complete, they will highlight all the sections and resources that will help them earn sales revenue. After completing the report, the finance team will further analyze the cost plan and final budget report.

6. **Engineering and Maintenance Department**

- **Create Resource List:** The engineering and maintenance department will create a resource list to make the internal workflow efficient. The resources will include pipes for water and gas flow, electricity wiring for lights and other amenities, making a garbage disposal shaft, etc
- **Review the Infrastructure Design Overview:** Once they have the infrastructure design, they will know where to place the pipes, electricity wire connections, etc. Therefore, they can plan and schedule what to start first and further.

7. **Program Committee**

- **Optimize Business Plan by Creating Detailed Hotel Program:** The program committee optimizes and adds further program details after the construction of the building. They create detailed programs containing the interior design plan, costing, hiring of the workforce, etc
- **Schematic Design Draft for Proposing Interior Designer:** Schematic design is a rough representation of the insides of the building will look. How much space is required? What type of design will be suitable? Is the space proper or not? Etc
- **Create Project Structure for Hotel Opening:** After finalizing everything, a project structure contains a framework where employees continuously function on inside projects to finalize the hotel opening.

Business Process Diagram

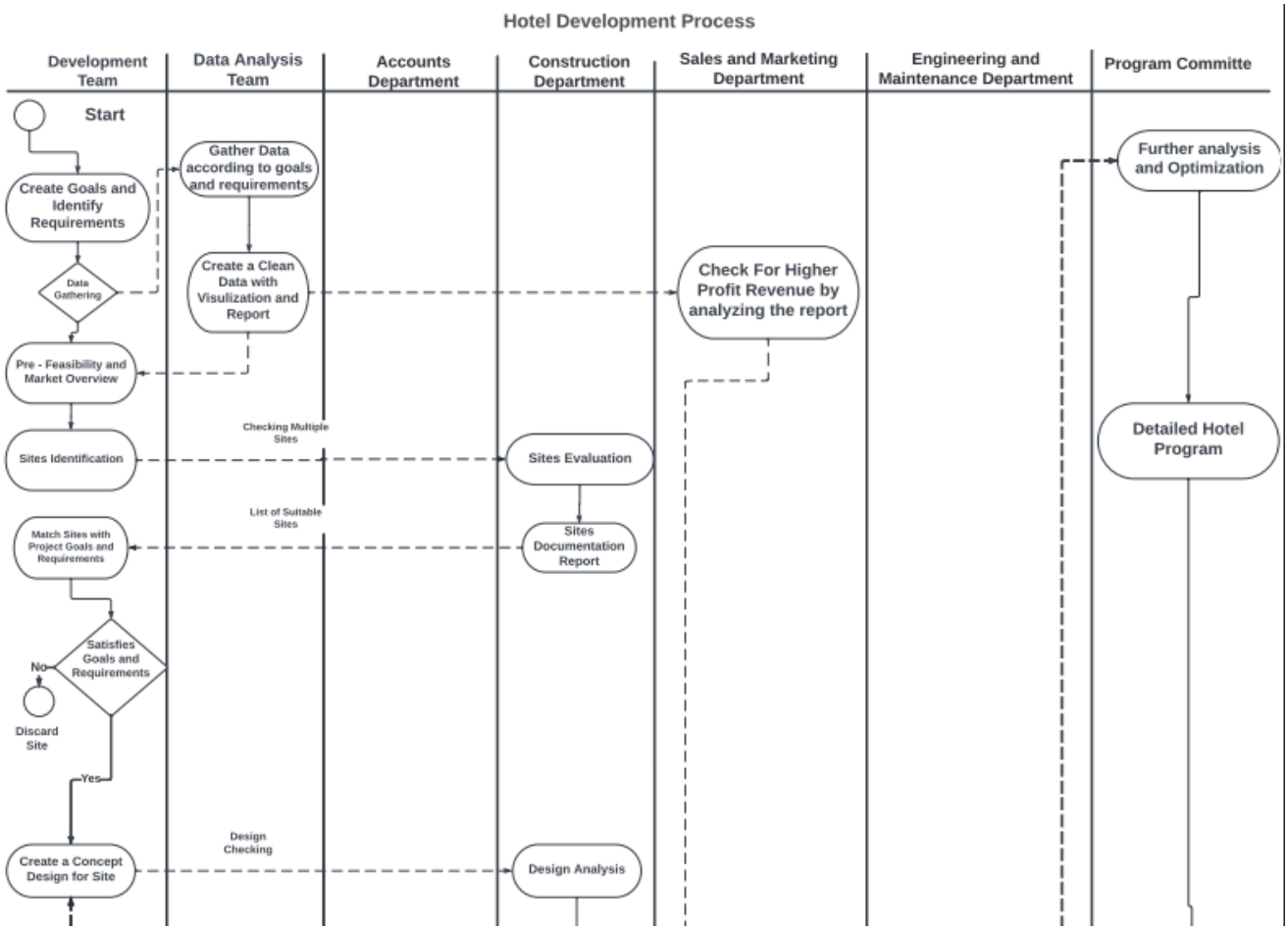


Figure 1: Business Process Model

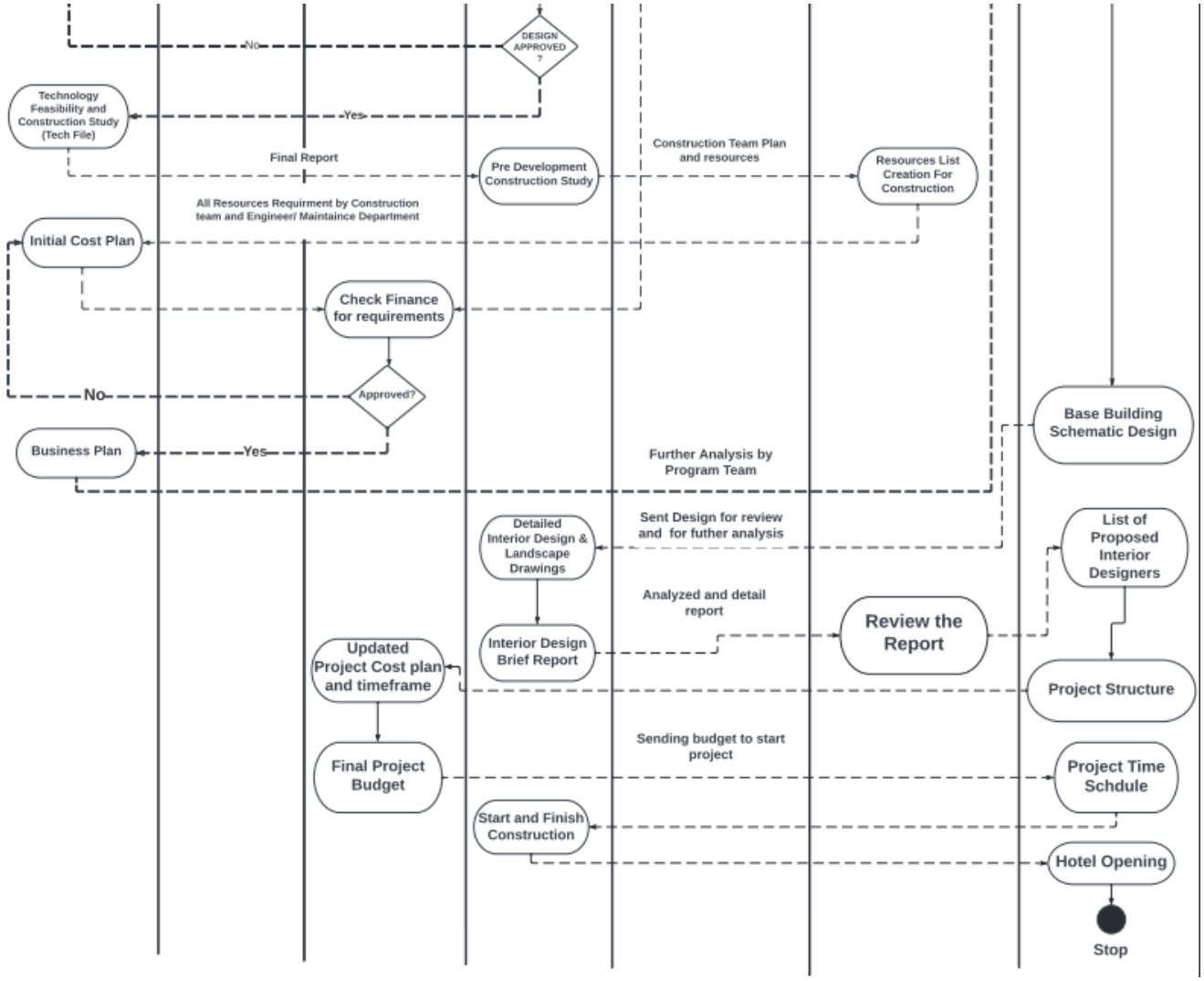


Figure 2: Business Process Model continued

Data Understanding

This data set is regarding Hilton hotel Case. The objective of the project is to help them increase their revenue so that they can increase their international market expansion by analyzing data. The data set was collected from three different source

- Data Source 1: OECD.Stat: https://stats.oecd.org/
- Data Source 2: Data World Bank: https://data.worldbank.org/
- Data Source 3: Macrotrends: https://www.macrotrends.net/

This dataset contains information like countries names, total population, male and female population, domestic revenue, international revenue, total country revenue, working hours, crime rate, literacy rate. All these variables are useful to solve our data analysis goals. It has all the details of 48 different countries from the year 2005 to 2020. The dataset contains 768 rows and 22 columns, shown below.

Variable Name	Data Type	Data description
Country	Categorical	It represents the It name of different country
Year	Numerical: Discrete	It shows year from 2005 to 2020
Male Population	Numerical: Discrete	It represents the total male population of different countries for different year
Female Population	Numerical: Discrete	It represents the total female population of different countries for different year
Country Population	Numerical: Discrete	It represents the total population of different countries for different year

Tourism Employment (%)	Numerical: Continuous	It represents the total percentage of employment in the tourism industry
Employed Tourist	Numerical: Discrete	It represents the total number of people employed in tourism industry
Tourist_Arrival	Numerical: Discrete	It represents the total number of tourist arrival for a particular country for different year
Domestic Revenue (Billions)	Numerical: Continuous	It represents the total domestic revenue in Billions for different countries and different years
Literacy	Categorical	It shows the literacy rate of each country in different years. It shows whether the literacy rate is high, medium or low
Crime Rate	Numerical: Continuous	It represents the crime rate of different countries in different year
Annual Avg WorkHrs	Numerical: Continuous	It represents the average working hours of employees annually
Upcoming_Year	Numerical: Discrete	It represents Year from 2021 to 2036
Projected Population	Numerical: Discrete	It represents the projected population in further years of different countries
Total Revenue (Billions)	Numerical: Continuous	It represents the total revenue in Billions for different countries and different years

International Revenue (Billions)	Numerical: Continuous	It represents the total international revenue in Billions for different countries and different years
CoCur_1DCon	Numerical: Continuous	It represents how much is one US dollar in other country currency
Currency Name	Categorical	It shows the currency **names** for different countries
Sal/Hr	Numerical: Continuous	It represents the salary per hour of employees in different countries

Table 1: Shows the data description of each variable

Collect Initial Data

- The following data set is from OECD. Stat for the Demography and population of the world. This dataset has information on Migration Statistics, and Population Statistics of different parts of the world from the year 2005 to 2020 and includes different age group from the age of 0 to more than 50 years of age. This data set gives us values of total population, Male Population, Female Population and Projected population which will be used in data analysis goals.

Variable Name	Included/Excluded	Rational
Immigrants by Citizenship and age	Excluded	Irrelevant to our Hilton case
Immigrants by Detailed Occupation	Excluded	Irrelevant to our Hilton case
Immigrants by duration of Stay	Excluded	Irrelevant to our Hilton case
Immigrants by field of study	Excluded	Irrelevant to our Hilton case

Immigrants by labor force status	Excluded	Irrelevant to our Hilton case
Immigrants by occupation	Excluded	Irrelevant to our Hilton case
Immigrants by sector	Excluded	Irrelevant to our Hilton case
Immigrants by sex and age	Excluded	Irrelevant to our Hilton case
International Migration Database	Excluded	Irrelevant to our Hilton case
Employment, Unemployment, Participation rates by sex and place of birth	Excluded	Irrelevant to our Hilton case
Employment rates by place of birth and educational attainment	Excluded	Irrelevant to our Hilton case
Historical population Data and Projection (1950-2060)	Excluded	Irrelevant to our Hilton case
Population	Excluded	Irrelevant to our Hilton case
Population and Vital Statistics	Excluded	Irrelevant to our Hilton case
Historical population Male	Included	It shows the male population in different countries from the year 2005 to 2022. This data will help us know the male population in the specific country for different years which will be used in data analysis
Historical population Female	Included	It shows the female population in different countries from the year 2005 to 2022. This data will help us know the female population in the specific country for different years which will be used in data analysis
Historical population Total	Included	It shows the total population which of different countries for different years. According to our data analysis goals, we will examine total population

Population Projections	Included	It shows the projected population value in different countries from the year 2005 to 2020. We will use this data to find the projected population of different countries. With this data, we can predict the population count in further years.

- The following data set is from OECD. Stat for the Industry and Services for Tourism of the world. This dataset has information on domestic tourism, inbound and outbound tourism, Enterprises and employment in tourism, Internal tourism consumption, key tourism indicator, receipts, and expenditure. This data set gives values of percent of total employment, inbound and outbound tourism, total expenditure which will be used in data analysis goals.

Variable Name	Included/Excluded	Rational
Receipts and Expenditure National Currency	Excluded	Not relevant for the data analysis goals
Receipts and Expenditure US Dollars	Excluded	Not relevant for the data analysis goals
Receipts and Expenditure Euros	Excluded	Not relevant for the data analysis goals
Total international receipts	Excluded	Not relevant for the data analysis goals
International travel receipts	Excluded	Not relevant for the data analysis goals
International passenger transport receipts	Excluded	Not useful for our data analysis
International sea passenger transport receipts	Excluded	Not useful for our data analysis
International air passenger transport receipts	Excluded	Not useful for our data analysis
International other passenger transport receipts	Excluded	Not useful for our data analysis

Total international expenditure	Excluded	Not useful for our data analysis
International travel expenditure	Excluded	Not useful for our data analysis
International passenger transport expenditure	Excluded	Not useful for our data analysis
International sea passenger transport expenditure	Excluded	Not useful for our data analysis
International air passenger transport expenditure	Excluded	Not useful for our data analysis
International other passenger transport expenditure	Excluded	Not useful for our data analysis
Total International Arrivals	Included	This gives the information on total international arrivals for different countries for different years. We use this data in our data analysis
Domestic Tourism	Included	This gives information on domestic tourism for different countries for different years. We use this data in our data analysis
Domestic Tourism Overnight visitors (tourists)	Excluded	Not useful for our data analysis
Domestic Tourism Hotels and similar establishments	Excluded	Not useful for our data analysis
Domestic Overnight visitors (tourists)	Excluded	Not useful for our data analysis
Inbound Tourism	Included	This gives information on Inbound tourism for different countries for different years. We use this data in our data analysis
Inbound Tourism Total international arrivals	Excluded	Not useful for our data analysis
Inbound Tourism Overnight visitors (tourists)	Excluded	Not useful for our data analysis
Inbound Tourism Hotels and similar establishments	Excluded	Not useful for our data analysis

Inbound Tourism Private accommodation	Excluded	Not useful for our data analysis
International Overnight visitors (tourists)	Excluded	Not useful for our data analysis
Outbound Tourism	Included	This gives information on outbound tourism for different countries for different years. We use this data in our data analysis
Outbound Tourism Total international departures	Excluded	Not useful for our data analysis
Outbound Overnight Visitors (Tourists)	Excluded	Not useful for our data analysis
Outbound Same day Visitors (Excursionists)	Excluded	Not useful for our data analysis
Total tourism Industry	Excluded	Not useful for our data analysis
Enterprise Tourism industries	Excluded	Irrelevant Data
Enterprise Accommodation services for visitors	Excluded	Irrelevant Data
Enterprise Hotels and similar establishments	Excluded	Irrelevant Data
Enterprise Food and beverage serving industry	Excluded	Irrelevant Data
Enterprise Passenger transport	Excluded	Irrelevant Data
Enterprise Air passenger transport	Excluded	Irrelevant Data
Enterprise Railways passenger transport	Excluded	Irrelevant Data
Enterprise Road passenger transport	Excluded	Irrelevant Data
Enterprise Water passenger transport	Excluded	Irrelevant Data
Enterprise Passenger transport supporting services	Excluded	Irrelevant Data
Enterprise Transport equipment rental	Excluded	Irrelevant Data

Enterprise Travel agencies and other reservation services industry	Excluded	Irrelevant Data
Enterprise Cultural industry	Excluded	Irrelevant Data
Enterprise Sports and recreation industry	Excluded	Irrelevant Data
Enterprise Retail trade of country-specific tourism characteristic goods	Excluded	Irrelevant Data
Enterprise Other country-specific tourism industries	Excluded	Irrelevant Data
Enterprise Other industries	Excluded	Irrelevant Data
Employment in Tourism industries	Included	It gives total employment in tourism which is useful in our data analysis
Employment in Accommodation services for visitors	Excluded	Not useful for our data analysis
Employment in Hotels and similar establishments	Excluded	Not useful for our data analysis
Employment in Food and beverage serving industry	Excluded	Not useful for our data analysis
Employment in Passenger transport	Excluded	Not useful for our data analysis
Employment in Air passenger transport	Excluded	Not useful for our data analysis
Employment in Railways passenger transport	Excluded	Not useful for our data analysis
Employment in Road passenger transport	Excluded	Not useful for our data analysis
Employment in Water passenger transport	Excluded	Not useful for our data analysis
Employment in Passenger transport supporting services	Excluded	Not useful for our data analysis
Employment in Transport equipment rental	Excluded	Not useful for our data analysis

Employment in Travel agencies and other reservation services industry	Excluded	Not useful for our data analysis
Employment in Cultural industry	Excluded	Not useful for our data analysis
Employment in Sports and recreation industry	Excluded	Not useful for our data analysis
Employment in Retail trade of country-specific tourism characteristic goods	Excluded	Not useful for our data analysis
Employment in Other country-specific tourism industries	Excluded	Not useful for our data analysis
Employment in Other industries	Excluded	Not useful for our data analysis
Accommodation services for visitors	Excluded	Not useful for our data analysis
Hotels and similar establishments	Excluded	Not useful for our data analysis
Total Consumption products	Excluded	Not useful for our data analysis
Tourism connected products	Excluded	Not useful for our data analysis
Non-tourism related consumption products	Excluded	Not useful for our data analysis
Non-consumption products	Excluded	Not useful for our data analysis
Domestic tourism consumption	Excluded	Not useful for our data analysis
Inbound tourism consumption	Excluded	Not useful for our data analysis
Internal tourism consumption	Excluded	Not useful for our data analysis
Other components of tourism consumption	Excluded	Not useful for our data analysis
Total Expenditure	Included	It gives total tourism expenditure which is useful in our data analysis
Internal tourism expenditure	Excluded	Not useful for our data analysis
Domestic tourism expenditure	Excluded	Not useful for our data analysis
Inbound tourism expenditure	Excluded	Not useful for our data analysis
Other components of tourism expenditure	Excluded	Not useful for our data analysis
Tourism GDP as % of Total GDP	Excluded	Not useful for our data analysis

Tourism employment as % of total employment	Included	This gives the percentage of employment in tourism which is useful in our data analysis
Tourism GVA as % of total GVA	Excluded	Irrelevant Data

- The following data set is from OECD. Stat for the Labor data of different countries. This dataset has information on Job tenure, Employment by job tenure interval by average tenure, Employment by job tenure interval by frequency and Employment by job tenure intervals by person of different countries from the year 1976 to 2020 and includes different employment status, gender, and different age group from 15 to 75+ years. This data set gives the details of employment by job tenure which will be used in data analysis goals.

Variable Name	Included/Excluded	Rational
Job Tenure <1 Month	Excluded	We are considering a job tenure frequency of five years. So, we can exclude this data
Job Tenure 1 to <6 months	Excluded	We are considering a job tenure frequency of five years. So, we can exclude this data
Job Tenure 6 to <12 months	Excluded	We are considering a job tenure frequency of five years. So, we can exclude this data
Job Tenure 1 to <3 years	Excluded	We are considering a job tenure frequency of five years. So, we can exclude this data
Job Tenure 3 to <5 years	Excluded	We are considering a job tenure frequency of five years. So, we can exclude this data
Job Tenure <5 years	Included	This data gives information about people with less than five years of experience; we will use this data in our data analysis goals

Job Tenure 5 to <10 years	Included	This data gives information about people with five years to ten years of experience; we will use this data in our data analysis goals
Job Tenure 10 years and over	Included	This data gives information about people with more than ten years of experience; we will use this data in our data analysis goals
Total Job Tenure	Excluded	This data is not necessary for our analysis. We are considering the data for job tenure greater than ten years
Male	Excluded	Data is not useful foe our data analysis goals
Female	Excluded	Data is not useful for our data analysis goals
All Person	Included	This data is combined with both male and female employees; we will use this data in our data analysis goals
Employment Status: Total Employment	Excluded	Irrelevant Data
Employment Status: Dependent Employment	Excluded	Irrelevant Data
Total age	Included	This data gives information of all the age groups combined; we will use this data in our data analysis goals
Age 15 to 19	Excluded	Since we are considering total age, we can exclude this data
Age 15 to 24	Excluded	Since we are considering total age, we can exclude this data
Age 15 to 64	Excluded	Since we are considering total age, we can exclude this data

Age 20 to 24	Excluded	Since we are considering total age, we can exclude this data
Age 25 to 29	Excluded	Since we are considering total age, we can exclude this data
Age 25 to 54	Excluded	Since we are considering total age, we can exclude this data
Age 30 to 34	Excluded	Since we are considering total age, we can exclude this data
Age 35 to 39	Excluded	Since we are considering total age, we can exclude this data
Age 40 to 44	Excluded	Since we are considering total age, we can exclude this data
Age 45 to 49	Excluded	Since we are considering total age, we can exclude this data
Age 50 to 54	Excluded	Since we are considering total age, we can exclude this data
Age 55 to 59	Excluded	Since we are considering total age, we can exclude this data
Age 55 to 64	Excluded	Since we are considering total age, we can exclude this data
Age 60 to 64	Excluded	Since we are considering total age, we can exclude this data
Age 65 to 69	Excluded	Since we are considering total age, we can exclude this data
Age 65+	Excluded	Since we are considering total age, we can exclude this data
Age 70 to 74	Excluded	Since we are considering total age, we can exclude this data
Age 70+	Excluded	Since we are considering total age, we can exclude this data
Age 75+	Excluded	Since we are considering total age, we can exclude this data
Employment by job tenure intervals - average tenure	Excluded	Not relevant for the data analysis goals

Employment by job tenure intervals - frequency	Excluded	Not relevant for the data analysis goals
Employment by job tenure intervals - person	Excluded	Not relevant for the data analysis goals

- The following data set is from The World Bank for International tourism, number of arrivals. This data set has information of number of tourism arrival of different countries from the year 2000 to 2020.

Variable Name	Included/Excluded	Rational
Country	Included	This data gives information about different countries of the world. This data is useful to solve data analysis goals
Year	Included	This data is useful to solve data analysis goals
Tourist arrival Value	Included	This data gives information about tourist arrival value of different countries. This data is useful to solve data analysis goals

- The following data set is from OECD. Stat for the Labor data of different countries. This dataset has information on Job tenure, Employment by job tenure interval by average tenure, Employment by job tenure interval by frequency and Employment by job tenure intervals by person of different countries from the year 1976 to 2020 and includes different employment status, gender, and more. This data set gives the details of total hours worked which will be used in data analysis goals.

Variable Name	Included/Excluded	Rational
Minimum wages at current prices in NCU	Excluded	Irrelevant Data
Real minimum wages	Excluded	Irrelevant Data

Wage gap by age	Excluded	Irrelevant Data
Average annual wages	Excluded	Irrelevant Data
Decile ratios of gross earnings	Excluded	Irrelevant Data
Gender wage gap (median)	Excluded	Irrelevant Data
Incidence of high pay	Excluded	Irrelevant Data
Incidence of low pay	Excluded	Irrelevant Data
Hourly earnings	Excluded	Irrelevant Data
Minimum relative to average wages of full-time workers	Excluded	Irrelevant Data
Strictness of employment protection – collective dismissals	Excluded	Irrelevant Data
Strictness of employment protection – individual and collective dismissals	Excluded	Irrelevant Data
Strictness of employment protection – individual dismissals	Excluded	Irrelevant Data
Strictness of employment protection – temporary dismissals	Excluded	Irrelevant Data
Annual labor force statistics tables	Excluded	Irrelevant Data
Employment by activities and status	Excluded	Irrelevant Data
LFS by sex and age	Excluded	Irrelevant Data
Short term Statistics	Excluded	Irrelevant Data
Full-time Employment	Excluded	Irrelevant Data
Part-time Employment	Excluded	Irrelevant Data
Involuntary Part time workers	Excluded	Irrelevant Data
Economics short time workers	Excluded	Irrelevant Data
Permanent temporary employment	Excluded	Irrelevant Data
Employment by job tenure intervals-average tenure	Excluded	Irrelevant Data

Employment by job tenure intervals-frequency	Excluded	Irrelevant Data
Employment by job tenure intervals-person	Excluded	Irrelevant Data
Hours worked	Included	It shows the total number of hours works which will be used in data analysis
Unemployment by duration	Excluded	Irrelevant Data
Marginal Labor force	Excluded	Irrelevant Data
Annual labor force statistics tables - Archives	Excluded	Irrelevant Data
Job Quality	Excluded	Irrelevant Data
Skills for jobs	Excluded	Irrelevant Data
Earnings	Excluded	Irrelevant Data
Employment Protection	Excluded	Irrelevant Data
Labor Market Programs	Excluded	Irrelevant Data

- The following data set is from macrotrends for crime rate, literacy, and population. This data set has information of crime rate, literacy of different countries from the year 2000 to 2020.

Variable Name	Included/Excluded	Rational
Population	Included	This data has information on the total population count of the world for different years; this data can be used for our data analysis goals
Economy	Excluded	Irrelevant Data
Trade	Excluded	Irrelevant Data
Health	Excluded	Irrelevant Data
Education	Included	This data has information on the literacy rate of different countries for different years; this data can be used for our data analysis goals
Development	Excluded	Irrelevant Data

Labor Force	Excluded	Irrelevant Data
Environment	Excluded	Irrelevant Data
Crime	Included	This data has information on the crime rate of different countries for different years; this data can be used for our data analysis goals
Immigration	Excluded	Irrelevant Data
Murder/Homicide Rate	Excluded	Irrelevant Data
Growth Rate	Excluded	Irrelevant Data
Density	Excluded	Irrelevant Data
Urban	Excluded	Irrelevant Data
Rural	Excluded	Irrelevant Data
Life Expectancy	Excluded	Irrelevant Data
Birth Rate	Excluded	Irrelevant Data
Death Rate	Excluded	Irrelevant Data
Infant Mortality Rate	Excluded	Irrelevant Data
Fertility Rate	Excluded	Irrelevant Data

- The following data set is from OECD. Stat for the revenue statistics of different countries. This dataset has information on Taxation, tax revenue, revenue statistics of different countries from the year 1965 to 2020 and includes different revenue for different sector. This data set gives the details of total revenue in USD of different countries which will be used in data analysis goals.

Variable Name	Included/Excluded	Rational
Supranational Revenue	Excluded	Irrelevant to Hilton Case
Federal or Central Government Revenue	Excluded	Irrelevant to Hilton Case
State/Regional Revenue	Excluded	Irrelevant to Hilton Case

Local governance Revenue	Excluded	Irrelevant to Hilton Case
Social security funds	Excluded	Irrelevant to Hilton Case
Total Tax revenue	Excluded	Irrelevant to Hilton Case
Total revenue in national currency	Excluded	We will use total revenue in USD, so we can exclude this data
Total revenue in USD	Included	This gives the total revenue in USD for different countries for different years which will be used in data analysis goals
Total revenue as % of GDP	Excluded	Data is not useful for our data analysis goals
Total revenue as % of total taxation	Excluded	Irrelevant to Hilton Case
Tax revenue of sub-sectors of general government as % of total tax revenue	Excluded	Irrelevant to Hilton Case

Data Gathering

Data were collected from three different sources that are from OECD. Stat, DataWorld bank, and macrotrends, these data were added to the spreadsheet. All the required data was collected to solve data analysis goals. All the data collected were added to a different sheet in the spreadsheet and were merged into one final dataset. Below is the data set that was collected from a different source.

Sample Data for the variable Country: It has data set of 53 different countries

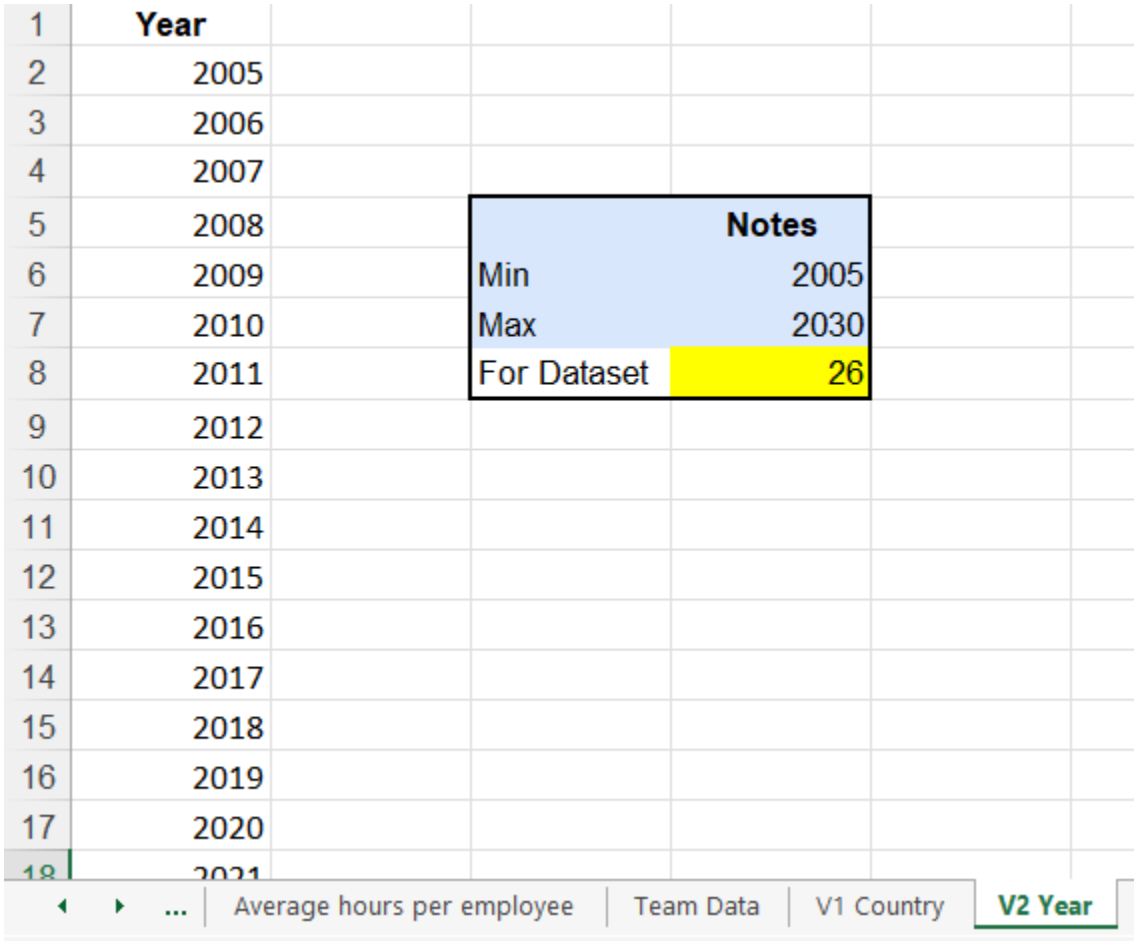
	Number	Countries
2	1	Argentina
3	2	Australia
4	3	Austria
5	4	Belgium
6	5	Brazil
7	6	Bulgaria
8	7	Canada
9	8	Chile
10	9	China
11	10	Colombia
12	11	Costa Rica
13	12	Croatia
14	13	Cyprus
15	14	Czech Republic
16	15	Denmark
17	16	Estonia

	Notes
Total	195 Countries
For Dataset	53

... | Average hours per employee | Team Data | V1 Country | V2 Year

Figure 3: Sample Data for the variable Country

Sample Data for variable year: It contains data of different years from 2005 to 2030

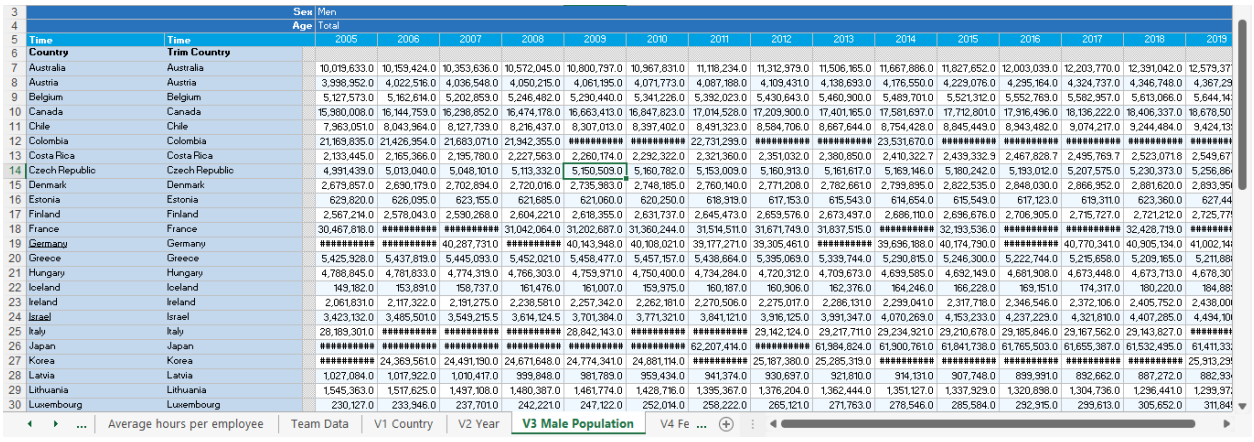


	Year
1	
2	2005
3	2006
4	2007
5	2008
6	2009
7	2010
8	2011
9	2012
10	2013
11	2014
12	2015
13	2016
14	2017
15	2018
16	2019
17	2020
18	2021

Notes	
Min	2005
Max	2030
For Dataset	26

Figure 4: Sample Data for the variable Year

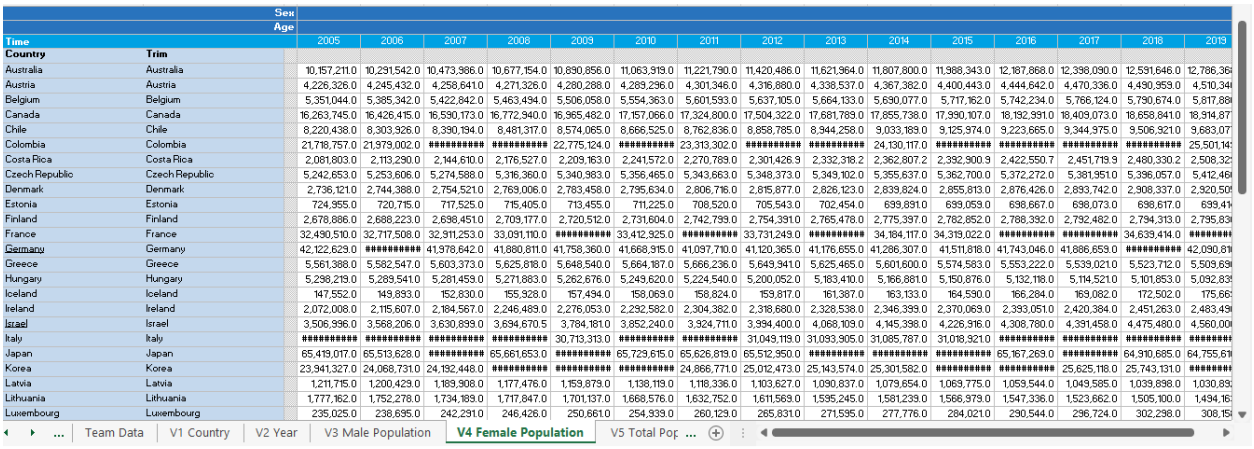
Sample Data for the variable Male Population



Sex	Men														
Age	Total														
Time	Time	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Country	Trim Country															
Australia	Australia	10,019,633.0	10,159,424.0	10,353,636.0	10,572,045.0	10,800,737.0	10,967,831.0	11,116,234.0	11,312,979.0	11,506,165.0	11,667,886.0	11,827,652.0	12,003,039.0	12,203,770.0	12,391,042.0	12,579,37
Austria	Austria	3,998,952.0	4,022,516.0	4,036,548.0	4,050,215.0	4,061,195.0	4,071,773.0	4,087,188.0	4,109,431.0	4,138,693.0	4,176,550.0	4,229,076.0	4,295,164.0	4,324,737.0	4,346,748.0	4,367,29
Belgium	Belgium	5,127,573.0	5,162,614.0	5,202,859.0	5,246,482.0	5,290,440.0	5,341,226.0	5,392,023.0	5,430,643.0	5,460,900.0	5,489,701.0	5,521,312.0	5,552,769.0	5,582,957.0	5,613,066.0	5,644,14
Canada	Canada	15,980,008.0	16,144,753.0	16,298,852.0	16,474,178.0	16,663,413.0	16,847,823.0	17,014,528.0	17,209,900.0	17,401,165.0	17,581,697.0	17,712,801.0	17,916,496.0	18,136,222.0	18,406,337.0	18,678,50
Chile	Chile	7,963,051.0	8,043,964.0	8,127,739.0	8,216,437.0	8,307,013.0	8,397,402.0	8,491,323.0	8,584,706.0	8,667,644.0	8,754,428.0	8,845,449.0	8,943,482.0	9,074,217.0	9,244,484.0	9,424,13
Colombia	Colombia	21,169,835.0	21,426,954.0	21,683,071.0	21,942,355.0	##########	##########	22,731,299.0	##########	##########	23,531,670.0	##########	##########	##########	##########	########
Costa Rica	Costa Rica	2,133,445.0	2,165,366.0	2,195,780.0	2,227,563.0	2,260,174.0	2,292,322.0	2,321,360.0	2,351,032.0	2,380,850.0	2,410,322.7	2,439,332.9	2,467,828.7	2,495,769.7	2,523,071.8	2,549,67
Czech Republic	Czech Republic	4,991,439.0	5,013,040.0	5,048,101.0	5,113,332.0	5,150,509.0	5,160,782.0	5,153,009.0	5,160,913.0	5,161,617.0	5,169,146.0	5,180,242.0	5,193,012.0	5,207,575.0	5,230,373.0	5,256,86
Denmark	Denmark	2,679,857.0	2,690,179.0	2,702,894.0	2,720,016.0	2,735,983.0	2,748,185.0	2,760,140.0	2,771,208.0	2,782,661.0	2,793,895.0	2,822,535.0	2,848,030.0	2,866,952.0	2,881,620.0	2,893,95
Estonia	Estonia	629,820.0	626,095.0	623,155.0	621,685.0	621,060.0	620,250.0	618,919.0	617,153.0	615,543.0	614,654.0	615,549.0	617,123.0	619,311.0	623,360.0	627,44
Finland	Finland	2,567,214.0	2,578,043.0	2,590,268.0	2,604,221.0	2,618,355.0	2,631,737.0	2,645,473.0	2,659,576.0	2,673,497.0	2,686,110.0	2,696,676.0	2,706,905.0	2,715,727.0	2,721,212.0	2,725,77
France	France	30,467,818.0	##########	##########	31,042,064.0	31,202,687.0	31,360,244.0	31,514,511.0	31,671,743.0	31,837,515.0	##########	32,193,536.0	##########	##########	32,428,719.0	########
Germany	Germany	##########	##########	40,287,731.0	##########	40,143,948.0	40,108,021.0	39,177,271.0	39,305,461.0	##########	39,696,188.0	40,174,790.0	##########	40,770,341.0	40,905,134.0	41,002,14
Greece	Greece	5,425,928.0	5,437,819.0	5,445,093.0	5,452,021.0	5,458,477.0	5,457,157.0	5,438,664.0	5,395,069.0	5,339,744.0	5,290,815.0	5,246,300.0	5,222,744.0	5,215,658.0	5,209,165.0	5,211,88
Hungary	Hungary	4,788,845.0	4,781,833.0	4,774,319.0	4,766,303.0	4,759,371.0	4,750,400.0	4,734,284.0	4,720,312.0	4,709,673.0	4,699,585.0	4,692,149.0	4,681,908.0	4,673,448.0	4,673,713.0	4,678,30
Iceland	Iceland	149,182.0	153,891.0	158,737.0	161,476.0	161,007.0	159,975.0	160,187.0	160,906.0	162,376.0	164,246.0	166,228.0	169,151.0	174,317.0	180,220.0	184,88
Ireland	Ireland	2,061,831.0	2,117,322.0	2,191,275.0	2,238,581.0	2,257,342.0	2,262,181.0	2,270,506.0	2,275,017.0	2,286,131.0	2,299,041.0	2,317,718.0	2,346,546.0	2,372,106.0	2,405,752.0	2,438,00
Israel	Israel	3,423,132.0	3,485,501.0	3,549,215.5	3,614,124.5	3,701,384.0	3,771,321.0	3,841,121.0	3,916,125.0	3,991,347.0	4,070,269.0	4,153,233.0	4,237,223.0	4,321,810.0	4,407,285.0	4,494,10
Italy	Italy	28,189,301.0	##########	##########	##########	28,842,143.0	##########	##########	29,142,124.0	29,217,711.0	29,234,321.0	29,210,678.0	29,185,846.0	29,167,562.0	29,143,827.0	########
Japan	Japan	##########	##########	##########	##########	##########	##########	62,207,414.0	##########	61,984,824.0	61,900,761.0	61,841,738.0	61,765,503.0	61,655,387.0	61,532,495.0	61,411,33
Korea	Korea	##########	24,369,561.0	24,431,190.0	24,671,648.0	24,774,341.0	24,881,114.0	##########	25,187,380.0	25,285,319.0	##########	##########	##########	##########	##########	25,313,29
Latvia	Latvia	1,027,084.0	1,017,922.0	1,010,417.0	999,848.0	981,789.0	959,434.0	941,374.0	930,697.0	921,810.0	914,131.0	907,748.0	899,991.0	892,662.0	887,272.0	882,93
Lithuania	Lithuania	1,545,363.0	1,517,625.0	1,497,108.0	1,480,387.0	1,461,774.0	1,428,716.0	1,395,367.0	1,376,204.0	1,362,444.0	1,351,127.0	1,337,923.0	1,320,896.0	1,304,736.0	1,296,441.0	1,293,97
Luxembourg	Luxembourg	230,127.0	233,946.0	237,701.0	242,221.0	247,122.0	252,014.0	256,222.0	265,121.0	271,763.0	278,546.0	285,584.0	292,915.0	299,613.0	305,652.0	311,84

Figure 5: Sample Data for the variable Male population

Sample Data for variable Female Population

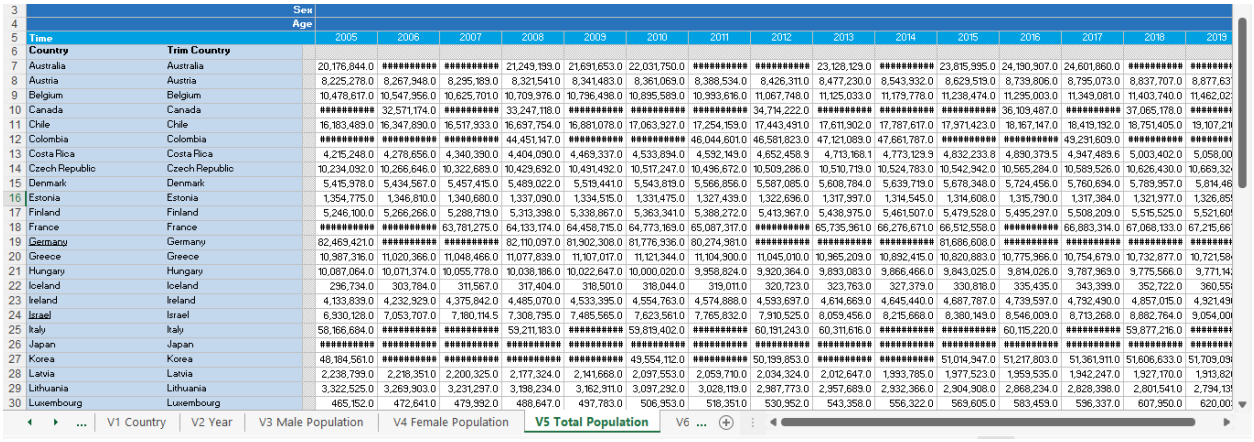


Sex																	
Age																	
Time		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	
Country	Trim																
Australia	Australia	10,157,211.0	10,291,542.0	10,473,986.0	10,677,154.0	10,890,856.0	11,063,919.0	11,221,790.0	11,420,486.0	11,621,964.0	11,807,800.0	11,988,343.0	12,187,868.0	12,398,090.0	12,591,646.0	12,786,36	
Austria	Austria	4,226,326.0	4,245,432.0	4,258,641.0	4,271,326.0	4,280,288.0	4,289,236.0	4,301,346.0	4,316,880.0	4,338,537.0	4,367,382.0	4,400,443.0	4,444,642.0	4,470,336.0	4,490,959.0	4,510,34	
Belgium	Belgium	5,351,044.0	5,385,342.0	5,422,842.0	5,463,494.0	5,506,058.0	5,554,363.0	5,601,593.0	5,637,105.0	5,664,133.0	5,690,077.0	5,717,162.0	5,742,234.0	5,766,124.0	5,790,674.0	5,817,88	
Canada	Canada	16,263,745.0	16,426,415.0	16,590,173.0	16,772,940.0	16,965,482.0	17,157,066.0	17,324,800.0	17,504,322.0	17,681,789.0	17,855,738.0	17,990,107.0	18,192,991.0	18,409,073.0	18,658,841.0	18,914,87	
Chile	Chile	8,220,438.0	8,303,926.0	8,390,194.0	8,481,317.0	8,574,065.0	8,666,525.0	8,762,836.0	8,858,785.0	8,944,258.0	9,033,189.0	9,125,974.0	9,223,665.0	9,344,975.0	9,506,921.0	9,683,07	
Colombia	Colombia	21,718,757.0	21,979,002.0	##########	##########	22,775,124.0	##########	23,313,302.0	##########	##########	24,130,117.0	##########	##########	##########	##########	25,501,14	
Costa Rica	Costa Rica	2,081,803.0	2,113,290.0	2,144,610.0	2,176,527.0	2,209,163.0	2,241,572.0	2,270,789.0	2,301,426.9	2,332,318.2	2,362,807.2	2,392,900.9	2,422,550.7	2,451,719.9	2,480,330.2	2,508,32	
Czech Republic	Czech Republic	5,242,653.0	5,253,606.0	5,274,588.0	5,316,360.0	5,340,983.0	5,356,465.0	5,343,663.0	5,348,373.0	5,349,102.0	5,355,637.0	5,362,700.0	5,372,272.0	5,381,951.0	5,396,057.0	5,412,46	
Denmark	Denmark	2,736,121.0	2,744,388.0	2,754,521.0	2,769,006.0	2,783,458.0	2,795,634.0	2,806,716.0	2,815,877.0	2,826,123.0	2,839,824.0	2,855,813.0	2,876,426.0	2,893,742.0	2,908,337.0	2,920,50	
Estonia	Estonia	724,955.0	720,715.0	717,525.0	715,405.0	713,455.0	711,225.0	708,520.0	705,543.0	702,454.0	699,891.0	699,059.0	698,667.0	698,073.0	698,617.0	699,41	
Finland	Finland	2,678,886.0	2,688,223.0	2,698,451.0	2,709,177.0	2,720,512.0	2,731,604.0	2,742,799.0	2,754,391.0	2,765,478.0	2,775,397.0	2,782,852.0	2,788,392.0	2,792,482.0	2,794,313.0	2,795,83	
France	France	32,490,510.0	32,717,508.0	32,911,253.0	33,091,110.0	##########	33,412,925.0	##########	33,731,249.0	##########	34,184,117.0	34,319,022.0	##########	##########	34,639,414.0	########	
Germany	Germany	42,122,629.0	##########	41,978,642.0	41,880,811.0	41,758,360.0	41,668,915.0	41,097,710.0	41,120,365.0	41,176,655.0	41,286,307.0	41,511,818.0	41,743,046.0	41,886,653.0	##########	42,090,81	
Greece	Greece	5,561,388.0	5,582,547.0	5,603,373.0	5,625,818.0	5,648,540.0	5,664,187.0	5,666,236.0	5,643,941.0	5,625,465.0	5,601,600.0	5,574,583.0	5,553,222.0	5,539,021.0	5,523,712.0	5,509,69	
Hungary	Hungary	5,298,219.0	5,289,541.0	5,281,459.0	5,271,883.0	5,262,676.0	5,249,620.0	5,224,540.0	5,200,052.0	5,183,410.0	5,166,881.0	5,150,876.0	5,132,118.0	5,114,521.0	5,101,853.0	5,092,83	
Iceland	Iceland	147,552.0	149,893.0	152,830.0	155,928.0	157,494.0	158,069.0	158,824.0	159,817.0	161,387.0	163,133.0	164,590.0	166,284.0	169,082.0	172,502.0	175,66	
Ireland	Ireland	2,072,008.0	2,115,607.0	2,164,567.0	2,246,489.0	2,276,053.0	2,232,582.0	2,304,382.0	2,318,680.0	2,328,538.0	2,346,399.0	2,370,069.0	2,393,051.0	2,420,384.0	2,451,263.0	2,483,49	
Israel	Israel	3,506,996.0	3,568,206.0	3,630,893.0	3,694,670.5	3,784,181.0	3,852,240.0	3,924,711.0	3,994,400.0	4,068,109.0	4,145,398.0	4,226,916.0	4,308,780.0	4,391,458.0	4,475,480.0	4,560,00	
Italy	Italy	##########	##########	##########	##########	30,713,313.0	##########	##########	31,049,119.0	31,093,905.0	31,085,787.0	31,018,921.0	##########	##########	##########	########	
Japan	Japan	65,419,017.0	65,513,628.0	##########	65,661,653.0	##########	65,729,615.0	65,626,819.0	65,512,950.0	##########	##########	##########	65,167,269.0	##########	64,910,685.0	64,755,61	
Korea	Korea	23,941,327.0	24,068,731.0	24,192,448.0	##########	##########	##########	24,866,771.0	25,012,473.0	25,143,574.0	25,301,582.0	##########	##########	25,625,118.0	25,743,131.0	########	
Latvia	Latvia	1,211,715.0	1,200,429.0	1,189,308.0	1,177,476.0	1,159,873.0	1,138,119.0	1,118,336.0	1,103,627.0	1,090,837.0	1,079,654.0	1,069,775.0	1,059,544.0	1,049,585.0	1,039,898.0	1,030,83	
Lithuania	Lithuania	1,777,162.0	1,752,278.0	1,734,189.0	1,717,847.0	1,701,137.0	1,668,576.0	1,632,752.0	1,611,569.0	1,595,245.0	1,581,239.0	1,566,973.0	1,547,336.0	1,523,662.0	1,505,100.0	1,494,16	
Luxembourg	Luxembourg	235,025.0	238,695.0	242,231.0	246,426.0	250,661.0	254,939.0	260,129.0	265,831.0	271,595.0	277,776.0	284,021.0	290,544.0	296,724.0	302,298.0	308,15	

... | Team Data | V1 Country | V2 Year | V3 Male Population | V4 Female Population | V5 Total Pop ...

Figure 6: Sample Data for the variable Female population

Sample data for the variable Total Population



3	Sex																
4	Age																
5	Time		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
6	Country	Trim Country															
7	Australia	Australia	20,176,844.0	##########	##########	21,243,193.0	21,691,653.0	22,031,750.0	##########	##########	23,128,129.0	##########	23,815,995.0	24,190,907.0	24,601,860.0	##########	########
8	Austria	Austria	8,225,278.0	8,267,948.0	8,295,189.0	8,321,541.0	8,341,483.0	8,361,069.0	8,388,534.0	8,426,311.0	8,477,230.0	8,543,932.0	8,629,519.0	8,739,806.0	8,795,073.0	8,837,707.0	8,877,63
9	Belgium	Belgium	10,478,617.0	10,547,956.0	10,625,701.0	10,709,976.0	10,796,498.0	10,895,589.0	10,993,616.0	11,067,748.0	11,125,033.0	11,179,778.0	11,238,474.0	11,295,003.0	11,349,081.0	11,403,740.0	11,462,02
10	Canada	Canada	##########	32,571,174.0	##########	33,247,118.0	##########	##########	##########	34,714,222.0	##########	##########	##########	36,109,487.0	##########	37,065,178.0	########
11	Chile	Chile	16,183,489.0	16,347,890.0	16,517,333.0	16,697,754.0	16,881,078.0	17,063,927.0	17,254,159.0	17,443,491.0	17,611,902.0	17,787,617.0	17,971,423.0	18,167,147.0	18,419,192.0	18,751,405.0	19,107,21
12	Colombia	Colombia	##########	##########	##########	44,451,147.0	##########	##########	46,044,601.0	46,581,823.0	47,121,089.0	47,661,787.0	##########	##########	49,291,609.0	##########	########
13	Costa Rica	Costa Rica	4,215,248.0	4,278,656.0	4,340,390.0	4,404,090.0	4,469,337.0	4,533,894.0	4,592,149.0	4,652,458.9	4,713,168.1	4,773,129.9	4,832,233.8	4,890,379.5	4,947,489.6	5,003,402.0	5,058,00
14	Czech Republic	Czech Republic	10,234,092.0	10,266,646.0	10,322,689.0	10,423,692.0	10,491,492.0	10,517,247.0	10,496,672.0	10,509,286.0	10,510,719.0	10,524,783.0	10,542,942.0	10,565,284.0	10,589,526.0	10,626,430.0	10,669,32
15	Denmark	Denmark	5,415,378.0	5,434,567.0	5,457,415.0	5,489,022.0	5,519,441.0	5,543,819.0	5,566,856.0	5,587,085.0	5,608,784.0	5,639,719.0	5,678,348.0	5,724,456.0	5,760,694.0	5,789,357.0	5,814,46
16	Estonia	Estonia	1,354,775.0	1,346,810.0	1,340,680.0	1,337,090.0	1,334,515.0	1,331,475.0	1,327,439.0	1,322,636.0	1,317,997.0	1,314,545.0	1,314,608.0	1,315,790.0	1,317,384.0	1,321,377.0	1,326,65
17	Finland	Finland	5,246,100.0	5,266,266.0	5,288,719.0	5,313,398.0	5,338,867.0	5,363,341.0	5,388,272.0	5,413,967.0	5,438,975.0	5,461,507.0	5,479,528.0	5,495,297.0	5,508,209.0	5,515,525.0	5,521,60
18	France	France	##########	##########	63,781,275.0	64,133,174.0	64,458,715.0	64,773,169.0	65,087,317.0	##########	65,735,961.0	66,276,671.0	66,512,558.0	##########	66,883,314.0	67,068,133.0	67,215,66
19	Germany	Germany	82,469,421.0	##########	##########	82,110,097.0	81,902,308.0	81,776,936.0	80,274,981.0	##########	##########	##########	81,686,608.0	##########	##########	##########	########
20	Greece	Greece	10,987,316.0	11,020,366.0	11,048,466.0	11,077,839.0	11,107,017.0	11,121,344.0	11,104,900.0	11,045,010.0	10,965,209.0	10,892,415.0	10,820,883.0	10,775,966.0	10,754,679.0	10,732,877.0	10,721,58
21	Hungary	Hungary	10,087,064.0	10,071,374.0	10,055,778.0	10,038,186.0	10,022,647.0	10,000,020.0	9,958,824.0	9,920,364.0	9,893,083.0	9,866,466.0	9,843,025.0	9,814,026.0	9,787,969.0	9,775,566.0	9,771,14
22	Iceland	Iceland	296,734.0	303,784.0	311,567.0	317,404.0	318,501.0	318,044.0	319,011.0	320,723.0	323,763.0	327,379.0	330,818.0	335,435.0	343,399.0	352,722.0	360,55
23	Ireland	Ireland	4,133,839.0	4,232,929.0	4,375,842.0	4,485,070.0	4,533,395.0	4,554,763.0	4,574,888.0	4,593,697.0	4,614,669.0	4,645,440.0	4,687,787.0	4,739,597.0	4,792,490.0	4,857,015.0	4,921,49
24	Israel	Israel	6,930,128.0	7,053,707.0	7,180,114.5	7,308,795.0	7,485,565.0	7,623,561.0	7,765,832.0	7,910,525.0	8,059,456.0	8,215,668.0	8,380,149.0	8,546,009.0	8,713,268.0	8,882,764.0	9,054,00
25	Italy	Italy	58,166,684.0	##########	##########	59,211,183.0	##########	59,819,402.0	##########	60,191,243.0	60,311,616.0	##########	##########	60,115,220.0	##########	59,877,216.0	########
26	Japan	Japan	##########	##########	##########	##########	##########	##########	##########	##########	##########	##########	##########	##########	##########	##########	########
27	Korea	Korea	48,184,561.0	##########	##########	##########	##########	49,554,112.0	##########	50,199,853.0	##########	##########	51,014,947.0	51,217,803.0	51,361,911.0	51,606,633.0	51,709,09
28	Latvia	Latvia	2,238,799.0	2,218,351.0	2,200,325.0	2,177,324.0	2,141,668.0	2,097,553.0	2,059,710.0	2,034,324.0	2,012,647.0	1,993,785.0	1,977,523.0	1,959,535.0	1,942,247.0	1,927,170.0	1,913,82
29	Lithuania	Lithuania	3,322,525.0	3,269,903.0	3,231,297.0	3,198,234.0	3,162,911.0	3,097,232.0	3,028,119.0	2,987,773.0	2,957,689.0	2,932,366.0	2,904,908.0	2,868,234.0	2,828,398.0	2,801,541.0	2,794,13
30	Luxembourg	Luxembourg	465,152.0	472,641.0	479,992.0	488,647.0	497,783.0	506,953.0	518,351.0	530,952.0	543,358.0	556,322.0	569,605.0	583,459.0	596,337.0	607,950.0	620,00

... | V1 Country | V2 Year | V3 Male Population | V4 Female Population | V5 Total Population | V6 ...

Figure7 : Sample Data for the variable Total population

Sample data for the variable projected population

Sex												
Age												
Time		2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
Country	Trim Country											
Australia	Australia	25873480	26301274	26727025	27147199	27562195	27970435	28372315	28765734	29157085	29545877	29931725
Austria	Austria	8921789	8960853	8996412	9028774	9058686	9087974	9116849	9145100	9172621	9199379	9225271
Belgium	Belgium	11496671	11515885	11557467	11608815	11657306	11703039	11746811	11787107	11824214	11860730	11896866
Canada	Canada	37873700	38284600	38694300	39102600	39509100	39913500	40315200	40713900	41109200	41500700	41888100
Chile	Chile	19458310	19678363	19828563	19960889	20086377	20206953	20322807	20433791	20539666	20640230	20735289
Colombia	Colombia	50911747	51207695	51302547	51310834	51390964	51651988	51875165	52091281	52339728	52672498	53125306
Costa Rica	Costa Rica	5111238.216	5163037.971	5213373.671	5262236.759	5309637.921	5355591.835	5400092.969	5443143.172	5484772.755	5525015.942	5563905.549
Czech Republic	Czech Republic	10712117	10744378	10767997	10782816	10788901	10789467	10787455	10782952	10776110	10767155	10756374
Denmark	Denmark	5825337	5848940	5869636	5892984	5916080	5939367	5963084	5986930	6010417	6033108	6054770
Estonia	Estonia	1329721	1329047	1327821	1326050	1323726	1321134	1318467	1315684	1312823	1309903	1306955
Finland	Finland	5527744	5533930	5539762	5545162	5550107	5554491	5558285	5561420	5563842	5565526	5566436
France	France	67283471	67454355	67622798	67789311	67954481	68112956	68262515	68406939	68546816	68682537	68814295
Germany	Germany	83193640	83299201	83381413	83439240	83472138	83487046	83492264	83489369	83479312	83463166	83442071
Greece	Greece	10681341	10649091	10613272	10574159	10531999	10489274	10447497	10405877	10364544	10323584	10283015
Hungary	Hungary	9765011	9750873	9736243	9721067	9705281	9689462	9674012	9658583	9642974	9627063	9610780
Iceland	Iceland	365936	370563	376961	384204	391681	397639	400936	402306	403171	404590	406219
Ireland	Ireland	5001392	5068447	5131568	5190794	5246185	5297122	5344915	5391635	5437365	5482191	5526174
Israel	Israel	9211817	9384714	9559547	9736520	9915715	10097068	10280369	10465622	10652995	10842602	11034494
Italy	Italy	60256664	60201461	60155851	60120854	60097323	60077987	60054689	60027172	59995796	59960852	59922532
Japan	Japan	125324842	124836189	124310239	123750589	123160804	122544102	121903097	121239693	120555142	119850168	119125137
Korea	Korea	51780579	51821669	51846339	51868100	51887623	51905126	51920462	51933215	51941946	51940598	51926953

V3 Male Population | V4 Female Population | V5 Total Population | **V6 Population Projection**

Figure 8: Sample Data for the variable Projected population

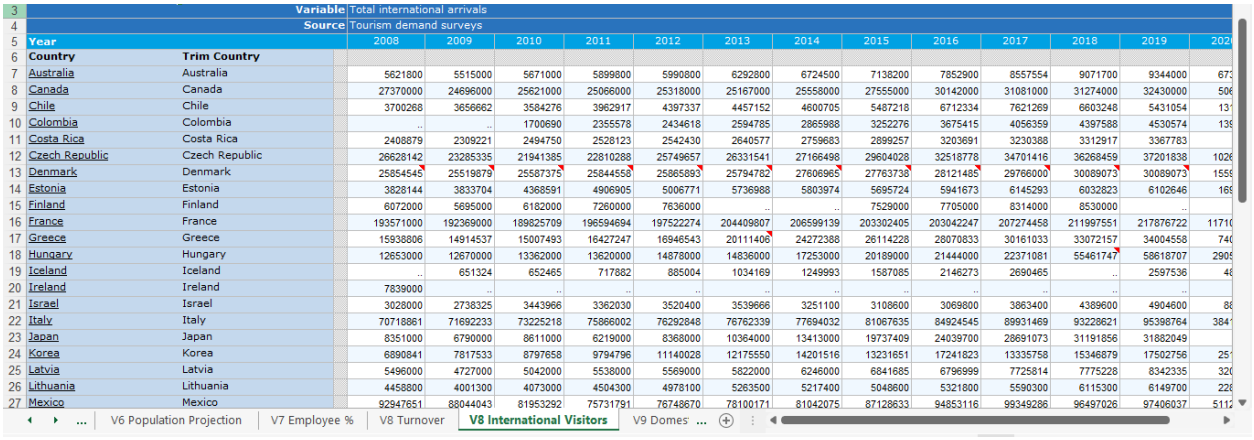
Sample data for Employee % to Total %

Key indicator	Total tourism employment (direct) as % of total employment												
Unit	Percentage												
Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Country	Trim Country													
Australia	Australia	4.7	4.6	4.6	4.6	4.7	4.7	4.9	4.9	5.1	5.2	5.3	5.0	
Austria	Austria	5.6	5.6	5.8	5.8	5.9	6.1	..	6.2	6.3	6.4	..	..	
Belgium	Belgium	..	..	..	..	..	..	..	..	6.7	..	..	..	
Canada	Canada	3.6	3.6	3.5	3.4	3.5	3.5	3.5	3.6	3.6	3.6	3.6	3.6	
Chile	Chile	6.0	6.3	6.3	6.2	6.1	6.1	6.0	6.2	6.4	6.5	7.2	7.2	
Colombia	Colombia	7.9	7.9	8.1	8.0	8.1	8.2	8.3	2.8	3.1	3.2	3.1	3.5	
Costa Rica	Costa Rica	..	..	..	6.6	6.6	6.7	6.5	6.7	6.8	6.8	6.6	..	
Czech Republic	Czech Republic	4.6	4.6	4.7	4.6	4.5	4.5	4.4	4.4	4.4	4.4	4.4	4.4	
Denmark	Denmark	7.6	7.7	7.9	8.0	8.2	8.3	8.3	8.5	8.8	8.9	9.0	9.1	
Estonia	Estonia	3.5	3.3	3.3	3.0	3.1	3.8	4.1	4.1	4.1	3.9	4.3	4.4	
Finland	Finland	..	..	..	5.2	5.3	5.5	5.5	5.6	5.5	5.5	5.6	5.8	
France	France	6.8	7.0	7.1	7.1	7.2	7.2	7.2	7.3	7.4	7.5	7.5	..	
Greece	Greece	11.6	11.4	11.2	11.6	11.7	11.9	13.1	13.7	14.4	14.3	14.1	14.4	1
Hungary	Hungary	8.7	8.6	8.6	9.0	9.2	9.1	9.2	10.0	10.0	9.6	9.4	9.5	
Iceland	Iceland	8.6	9.0	9.7	10.3	11.1	11.9	12.8	14.0	15.8	16.9	16.8	15.9	1
Ireland	Ireland	9.0	9.4	9.3	9.4	9.6	9.6	9.8	10.0	10.2	10.3	10.3	..	
Israel	Israel	3.3	3.2	3.4	3.4	3.4	3.6	3.7	3.7	3.6	3.7	3.6	3.8	
Italy	Italy	..	..	8.4	..	..	..	..	8.3	..	8.8	..	8.9	
Japan	Japan	6.9	6.9	6.9	6.9	6.9	6.9	9.8	9.6	9.6	9.3	9.5	9.6	
Latvia	Latvia	7.5	7.7	7.7	7.3	8.3	7.8	8.5	8.2	8.9	8.4	8.5	8.3	
Lithuania	Lithuania	4.0	4.1	4.3	4.5	4.4	4.6	4.8	4.8	4.8	4.9	4.9	5.1	

V4 Female Population | V5 Total Population | V6 Population Projection | **V7 Employee %** | V8 Tu

Figure 9: Sample Data for the variable Employee % to total %

Sample data for the variable International Visitors

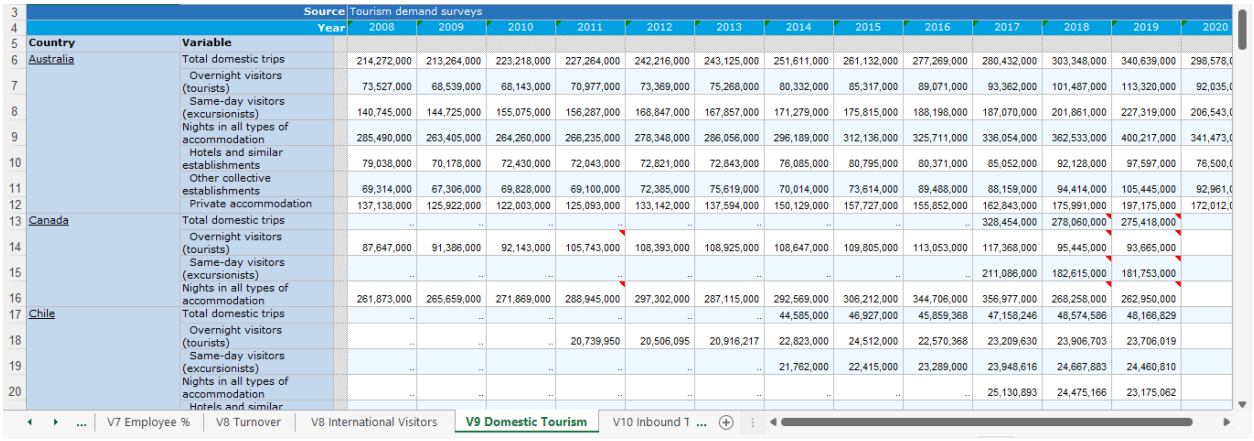


Variable	Total international arrivals												
Source	Tourism demand surveys												
Year		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	202
Country	Trim Country													
Australia	Australia	5621800	5515000	5671000	5899800	5990800	6292800	6724500	7138200	7852900	8557554	9071700	9344000	67
Canada	Canada	27370000	24696000	25621000	25066000	25318000	25167000	25558000	27555000	30142000	31081000	31274000	32430000	506
Chile	Chile	3700268	3656662	3584276	3962917	4397337	4457152	4600705	5487218	6712334	7621269	6603248	5431054	13
Colombia	Colombia	..	..	1700690	2355578	2434618	2594785	2865988	3252276	3675415	4056359	4397588	4530574	139
Costa Rica	Costa Rica	2408879	2309221	2494750	2528123	2542430	2640577	2759683	2899257	3203691	3230388	3312917	3367783	
Czech Republic	Czech Republic	26628142	23285335	21941385	22810288	25749657	26331541	27166498	29604028	32518778	34701416	36268459	37201838	1026
Denmark	Denmark	25854545	25519879	25587375	25844558	25865893	25794782	27606965	27763738	28121485	29766000	30089073	30089073	1559
Estonia	Estonia	3828144	3833704	4368591	4906905	5006771	5736988	5803974	5695724	5941673	6145293	6032823	6102646	169
Finland	Finland	6072000	5695000	6182000	7260000	7636000	..	..	7529000	7705000	8314000	8530000	..	
France	France	193571000	192369000	189825709	196594694	197522274	204409807	206599139	203302405	203042247	207274458	211997551	217876722	1171
Greece	Greece	15938806	14914537	15007493	16427247	16946543	20111406	24272388	26114228	28070833	30161033	33072157	34004558	740
Hungary	Hungary	12653000	12670000	13362000	13620000	14878000	14836000	17253000	20189000	21444000	22371081	55461747	58618707	2995
Iceland	Iceland	..	651324	652465	717882	885004	1034169	1249993	1587085	2146273	2690465	..	2597536	48
Ireland	Ireland	7839000	..	..	..	..	..	..	..	..	..	..	..	
Israel	Israel	3028000	2738325	3443966	3362030	3520400	3539666	3251100	3108600	3069800	3863400	4389600	4904600	88
Italy	Italy	70718861	71692233	73225218	75866002	76292848	76762339	77694032	81067635	84924545	89931469	93228621	95398764	384
Japan	Japan	8351000	6790000	8611000	6219000	8368000	10364000	13413000	19737409	24039700	28691073	31191856	31882049	
Korea	Korea	6890841	7817533	8797658	9794796	11140028	12175550	14201516	13231651	17241823	13335758	15346879	17502756	251
Latvia	Latvia	5496000	4727000	5042000	5538000	5569000	5822000	6246000	6841685	6796999	7725814	7775228	8342335	320
Lithuania	Lithuania	4458800	4001300	4073000	4504300	4978100	5263500	5217400	5048600	5321800	5590300	6115300	6149700	228
Mexico	Mexico	92947651	88044043	81953292	75731791	76748670	78100171	81042075	87128633	94853116	99349286	96497026	97406037	5112

V6 Population Projection | V7 Employee % | V8 Turnover | V8 International Visitors | V9 Domes ...

Figure 10: Sample Data for the variable international visitors

Sample data for the variable Domestic Tourism



Source	Tourism demand surveys												
Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Country	Variable													
Australia	Total domestic trips	214,272,000	213,264,000	223,218,000	227,264,000	242,216,000	243,125,000	251,611,000	261,132,000	277,269,000	280,432,000	303,348,000	340,639,000	298,578,(
	Overnight visitors (tourists)	73,527,000	68,539,000	68,143,000	70,977,000	73,369,000	75,268,000	80,332,000	85,317,000	89,071,000	93,362,000	101,487,000	113,320,000	92,035,(
	Same-day visitors (excursionists)	140,745,000	144,725,000	155,075,000	156,287,000	168,847,000	167,857,000	171,279,000	175,815,000	188,198,000	187,070,000	201,861,000	227,319,000	206,543,(
	Nights in all types of accommodation	285,490,000	263,405,000	264,260,000	266,235,000	278,348,000	286,056,000	296,189,000	312,136,000	325,711,000	336,054,000	362,533,000	400,217,000	341,473,(
	Hotels and similar establishments	79,038,000	70,178,000	72,430,000	72,043,000	72,821,000	72,843,000	76,085,000	80,795,000	80,371,000	85,052,000	92,128,000	97,597,000	76,500,(
	Other collective establishments	69,314,000	67,306,000	69,828,000	69,100,000	72,385,000	75,619,000	70,014,000	73,614,000	89,488,000	88,159,000	94,414,000	105,445,000	92,961,(
	Private accommodation	137,138,000	125,922,000	122,003,000	125,093,000	133,142,000	137,594,000	150,129,000	157,727,000	155,852,000	162,843,000	175,991,000	197,175,000	172,012,(
Canada	Total domestic trips	..	..	..	..	..	..	..	..	..	328,454,000	278,060,000	275,418,000	
	Overnight visitors (tourists)	87,647,000	91,386,000	92,143,000	105,743,000	108,393,000	108,925,000	108,647,000	109,805,000	113,053,000	117,368,000	95,445,000	93,665,000	
	Same-day visitors (excursionists)	..	..	..	..	..	..	..	..	..	211,086,000	182,615,000	181,753,000	
	Nights in all types of accommodation	261,873,000	265,659,000	271,869,000	288,945,000	297,302,000	287,115,000	292,569,000	306,212,000	344,706,000	356,977,000	268,258,000	262,950,000	
Chile	Total domestic trips	..	..	..	..	..	..	44,585,000	46,927,000	45,859,368	47,158,246	48,574,586	48,166,829	
	Overnight visitors (tourists)	..	..	..	20,739,950	20,506,095	20,916,217	22,823,000	24,512,000	22,570,368	23,209,630	23,906,703	23,706,019	
	Same-day visitors (excursionists)	..	..	..	..	..	..	21,762,000	22,415,000	23,289,000	23,948,616	24,667,883	24,460,810	
	Nights in all types of accommodation	..	..	..	..	..	..	..	..	..	25,130,893	24,475,166	23,175,062	
	Hotels and similar													

V7 Employee % | V8 Turnover | V8 International Visitors | V9 Domestic Tourism | V10 Inbound T ...

Figure 11: Sample Data for the variable domestic tourism

Sample Data for the variable Inbound Tourism

3			Source	Tourism demand surveys							
4			Year	2008	2009	2010	2011	2012	2013	2014	2015
5	Country	Variable									
6	Australia	Total international arrivals		5621800	5515000	5671000	5899800	5990800	6292800	6724500	7138200
7		Top markets	China	373200	355100	390400	498400	583800	689600	773700	936000
8			United Kingdom	709800	679700	678200	661200	626700	637900	671200	667100
9			Japan	519800	402300	361800	363000	344000	339100	330700	332700
10			New Zealand	1113600	1096500	1111100	1175500	1184100	1182500	1226500	1270800
11			United States	461700	455900	494200	475700	483900	509800	538200	584600
12			Singapore	231100	237100	248400	268300	265900	306900	357800	363300
13		Nights in all types of accommodation		158956372	172730578	182268932	190471255	200392859	217692167	217920861	235482508
14		Nights in all types of accommodation	Hotels and similar establishments	24164102	21586720	20734140	22139322	20949660	22029176	22560801	24481674
15			Other collective establishments	75853947	86892363	91945088	95189774	102858663	109134556	105163950	113670724
16			Private accommodation	58938322	64251494	69589704	73142159	76584535	86528436	90196109	97330110
17	Canada	Total international arrivals		27370000	24696000	25621000	25066000	25318000	25167000	25558000	27555000
18		Total international arrivals	Overnight visitors (tourists)	16997000	15737000	16219000	16016000	16344000	16059000	16537000	17977000
19			Same-day visitors (excursionists)	10373000	8959000	9402000	9050000	8974000	9108000	9021000	9578000
20		Top markets	Australia	219000	188000	202000	216000	219000	270000	283000	291000
21			China	159000	160000	193000	237000	273000	342000	448000	483000
22			Germany	315000	292000	316000	290000	277000	313000	324000	325000
23			France	405000	389000	408000	422000	423000	452000	465000	477000
24			United Kingdom	837000	686000	661000	623000	597000	609000	659000	686000

V8 Turnover | V8 International Visitors | V9 Domestic Tourism | **V10 Inbound Tourism** | V11 Exp ...

Figure12: Sample Data for the variable inbound tourism

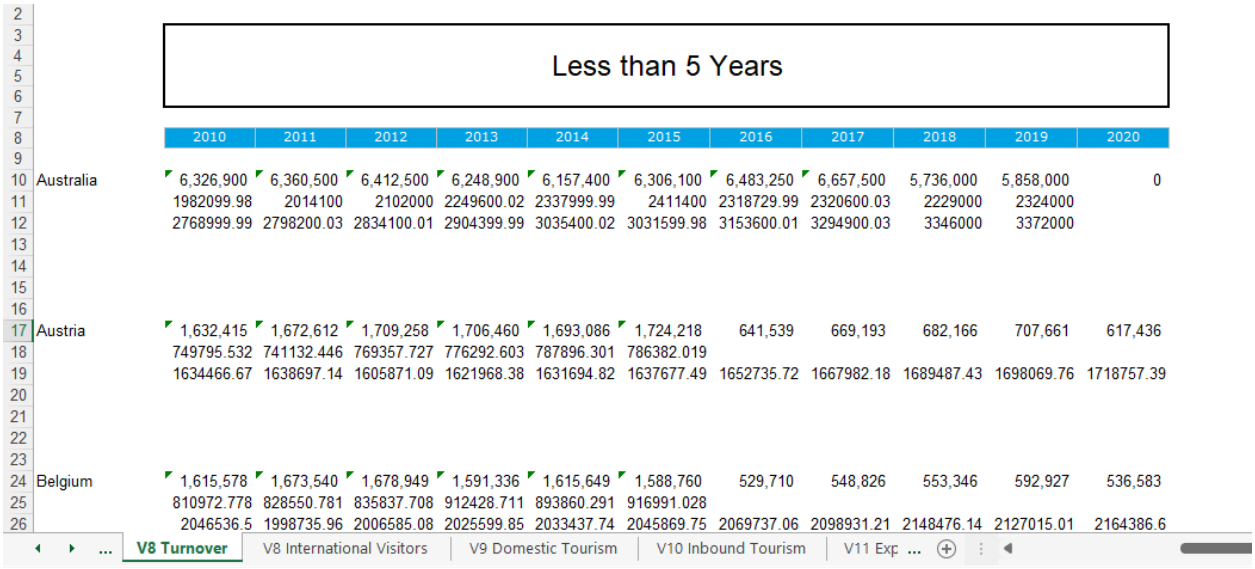
Sample data for the variable employment tenure

Employment status		Total employment										
Frequency		Annual										
Unit		Employed, Thousands										
Time		2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Country	Job tenure											
Australia	<1 month	409	382	413	371	353	362	393	418	..	..	..
	1 to <6 months	822	912	853	779	807	827	855	909	..	..	..
	6 to <12 months	1,016	1,078	1,031	982	1,007	1,087	1,127	1,127	1,238	1,284	..
	1 to <3 years	2,346	2,299	2,452	2,457	2,298	2,312	2,413	2,486	2,715	2,733	..
	3 to <5 years	1,734	1,690	1,662	1,660	1,692	1,720	1,696	1,717	1,783	1,841	..
	5 to <10 years	1,982	2,014	2,102	2,250	2,338	2,411	2,319	2,321	2,229	2,324	..
	10 years and over	2,769	2,798	2,834	2,904	3,035	3,032	3,154	3,295	3,346	3,372	..
Austria	<1 month	141	138	131	130	130	131	138	145	144	150	128
	1 to <6 months	221	233	222	217	209	223	228	242	246	249	217
	6 to <12 months	239	273	269	266	255	266	275	282	292	308	272
	1 to <3 years	535	513	579	582	583	563	..	..	..	..	..
	3 to <5 years	497	516	508	512	516	541	..	..	..	..	..
	5 to <10 years	750	741	769	776	788	786	..	..	..	..	..
	10 years and over	1,634	1,639	1,606	1,622	1,632	1,638	1,653	1,668	1,689	1,698	1,719
Belgium	<1 month	110	125	123	111	116	120	130	127	136	136	121
	1 to <6 months	183	207	191	172	182	176	194	192	191	206	187
	6 to <12 months	201	236	229	208	215	209	206	230	226	251	229
	1 to <3 years	571	560	603	576	548	553	..	..	..	..	..
	3 to <5 years	550	546	533	524	556	531	..	..	..	..	..
	5 to <10 years	811	829	836	912	894	917	..	..	..	..	..

V8 Turnover | V8 International Visitors | V9 Domestic Tourism | V10 Inbound Tourism | V11 Exp ...

Figure 13: Sample Data for the variable employment tenure

Sample data for the variable tenure < 5 years of experience



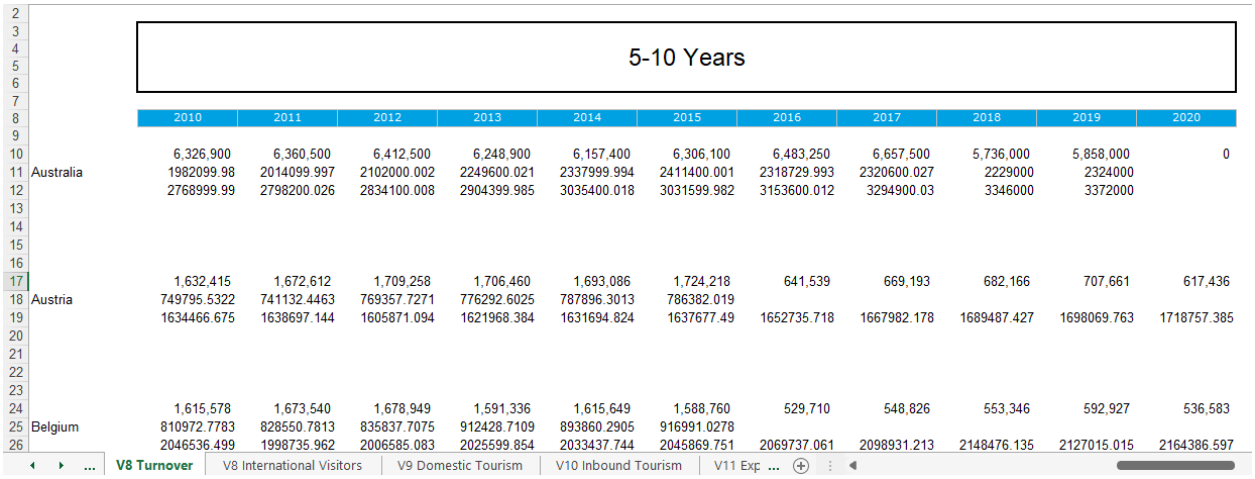
Less than 5 Years

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Australia	6,326,900	6,360,500	6,412,500	6,248,900	6,157,400	6,306,100	6,483,250	6,657,500	5,736,000	5,858,000	0
	1982099.98	2014100	2102000	2249600.02	2337999.99	2411400	2318729.99	2320600.03	2229000	2324000	
	2768999.99	2798200.03	2834100.01	2904399.99	3035400.02	3031599.98	3153600.01	3294900.03	3346000	3372000	
Austria	1,632,415	1,672,612	1,709,258	1,706,460	1,693,086	1,724,218	641,539	669,193	682,166	707,661	617,436
	749795.532	741132.446	769357.727	776292.603	787896.301	786382.019					
	1634466.67	1638697.14	1605871.09	1621968.38	1631694.82	1637677.49	1652735.72	1667982.18	1689487.43	1698069.76	1718757.39
Belgium	1,615,578	1,673,540	1,678,949	1,591,336	1,615,649	1,588,760	529,710	548,826	553,346	592,927	536,583
	810972.778	828550.781	835837.708	912428.711	893860.291	916991.028					
	2046536.5	1998735.96	2006585.08	2025599.85	2033437.74	2045869.75	2069737.06	2098931.21	2148476.14	2127015.01	2164386.6

V8 Turnover | V8 International Visitors | V9 Domestic Tourism | V10 Inbound Tourism | V11 Exp ...

Figure 14: Sample Data for the variable tenure < 5 years of experience

Sample data for the variable tenure 5-10 years of experience



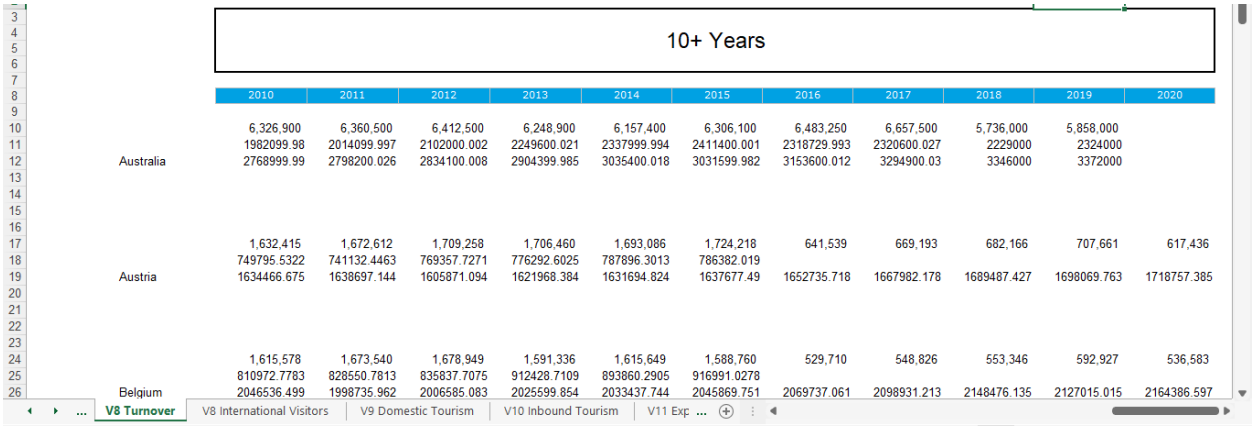
5-10 Years

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
	6,326,900	6,360,500	6,412,500	6,248,900	6,157,400	6,306,100	6,483,250	6,657,500	5,736,000	5,858,000	0
Australia	1982099.98	2014099.997	2102000.002	2249600.021	2337999.994	2411400.001	2318729.993	2320600.027	2229000	2324000	
	2768999.99	2798200.026	2834100.008	2904399.985	3035400.018	3031599.982	3153600.012	3294900.03	3346000	3372000	
	1,632,415	1,672,612	1,709,258	1,706,460	1,693,086	1,724,218	641,539	669,193	682,166	707,661	617,436
Austria	749795.5322	741132.4463	769357.7271	776292.6025	787896.3013	786382.019					
	1634466.675	1638697.144	1605871.094	1621968.384	1631694.824	1637677.49	1652735.718	1667982.178	1689487.427	1698069.763	1718757.385
	1,615,578	1,673,540	1,678,949	1,591,336	1,615,649	1,588,760	529,710	548,826	553,346	592,927	536,583
Belgium	810972.7783	828550.7813	835837.7075	912428.7109	893860.2905	916991.0278					
	2046536.499	1998735.962	2006585.083	2025599.854	2033437.744	2045869.751	2069737.061	2098931.213	2148476.135	2127015.015	2164386.597

V8 Turnover | V8 International Visitors | V9 Domestic Tourism | V10 Inbound Tourism | V11 Exp ...

Figure 15: Sample Data for the variable tenure 5-10 years of experience

Sample data for the variable tenure > 10 years of Experience

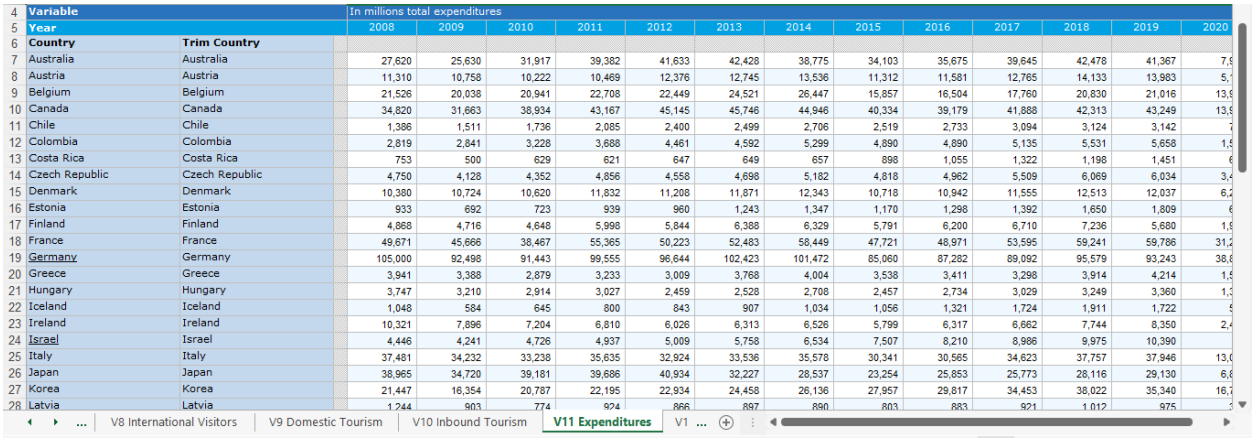


	10+ Years										
	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
	6,326,900	6,360,500	6,412,500	6,248,900	6,157,400	6,306,100	6,483,250	6,657,500	5,736,000	5,858,000	
	1982099.98	2014099.997	2102000.002	2249600.021	2337999.994	2411400.001	2318729.993	2320600.027	2229000	2324000	
Australia	2768999.99	2798200.026	2834100.008	2904399.985	3035400.018	3031599.982	3153600.012	3294900.03	3346000	3372000	
	1,632,415	1,672,612	1,709,258	1,706,460	1,693,086	1,724,218	641,539	669,193	682,166	707,661	617,436
	749795.5322	741132.4463	769357.7271	776292.6025	787896.3013	786382.019					
Austria	1634466.675	1638697.144	1605871.094	1621968.384	1631694.824	1637677.49	1652735.718	1667982.178	1689487.427	1698069.763	1718757.385
	1,615,578	1,673,540	1,678,949	1,591,336	1,615,649	1,588,760	529,710	548,826	553,346	592,927	536,583
	810972.7783	828550.7813	835837.7075	912428.7109	893860.2905	916991.0278					
Belgium	2046536.499	1998735.962	2006585.083	2025599.854	2033437.744	2045869.751	2069737.061	2098931.213	2148476.135	2127015.015	2164386.597

V8 Turnover | V8 International Visitors | V9 Domestic Tourism | V10 Inbound Tourism | V11 Exp ...

Figure 16: Sample Data for the variable tenure >10 years of experience

Sample data for the variable Expenditure



Variable		In millions total expenditures												
Year		2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Country	Trim Country													
Australia	Australia	27,620	25,630	31,917	39,382	41,633	42,428	38,775	34,103	35,675	39,645	42,478	41,367	7,9
Austria	Austria	11,310	10,758	10,222	10,469	12,376	12,745	13,536	11,312	11,581	12,765	14,133	13,983	5,1
Belgium	Belgium	21,526	20,038	20,941	22,708	22,449	24,521	26,447	15,857	16,504	17,760	20,830	21,016	13,9
Canada	Canada	34,820	31,663	38,934	43,167	45,145	45,746	44,946	40,334	39,179	41,888	42,313	43,249	13,9
Chile	Chile	1,386	1,511	1,736	2,085	2,400	2,499	2,706	2,519	2,733	3,094	3,124	3,142	7
Colombia	Colombia	2,819	2,841	3,228	3,688	4,461	4,592	5,299	4,890	4,890	5,135	5,531	5,658	1,5
Costa Rica	Costa Rica	753	500	629	621	647	649	657	898	1,055	1,322	1,198	1,451	6
Czech Republic	Czech Republic	4,750	4,128	4,352	4,856	4,558	4,698	5,182	4,818	4,962	5,509	6,069	6,034	3,4
Denmark	Denmark	10,380	10,724	10,620	11,832	11,208	11,871	12,343	10,718	10,942	11,555	12,513	12,037	6,2
Estonia	Estonia	933	692	723	939	960	1,243	1,347	1,170	1,298	1,392	1,650	1,809	6
Finland	Finland	4,868	4,716	4,648	5,998	5,844	6,388	6,329	5,791	6,200	6,710	7,236	5,680	1,9
France	France	49,671	45,666	38,467	55,365	50,223	52,483	58,449	47,721	48,971	53,595	59,241	59,786	31,2
Germany	Germany	105,000	92,498	91,443	99,555	96,644	102,423	101,472	85,060	87,282	89,092	95,579	93,243	38,6
Greece	Greece	3,941	3,388	2,879	3,233	3,009	3,768	4,004	3,538	3,411	3,298	3,914	4,214	1,5
Hungary	Hungary	3,747	3,210	2,914	3,027	2,459	2,528	2,708	2,457	2,734	3,029	3,249	3,360	1,3
Iceland	Iceland	1,048	584	645	800	843	907	1,034	1,056	1,321	1,724	1,911	1,722	5
Ireland	Ireland	10,321	7,896	7,204	6,810	6,026	6,313	6,526	5,799	6,317	6,662	7,744	8,350	2,4
Israel	Israel	4,446	4,241	4,726	4,937	5,009	5,758	6,534	7,507	8,210	8,986	9,975	10,390	
Italy	Italy	37,481	34,232	33,238	35,635	32,924	33,536	35,578	30,341	30,565	34,623	37,757	37,946	13,0
Japan	Japan	38,965	34,720	39,181	39,686	40,934	32,227	28,537	23,254	25,853	25,773	28,116	29,130	6,8
Korea	Korea	21,447	16,354	20,787	22,195	22,934	24,458	26,136	27,957	29,817	34,453	38,022	35,340	16,1
Latvia	Latvia	1,244	903	774	924	866	897	890	803	883	921	1,012	975	3

V8 International Visitors | V9 Domestic Tourism | V10 Inbound Tourism | V11 Expenditures | V1 ...

Figure 17: Sample Data for the variable expenditure

Sample Data for variable Capital

	Indicator	Growth capital and venture capital						
	Year		2007	2008	2009	2010	2011	2012
	Country	Unit						
6	Australia	Australian Dollar, Millions	6939	8315	7903	8912	8700	7652
7	Austria	Euro, Millions	83.79	73.03	113.1	75.29	208.67	64.55
8	Belgium	Euro, Millions	502.26	507.83	618.05	363.6	411.11	445.36
9	Canada	Euro, Billions	..	..	..	..	..	..
10	Chile	Chilean Peso, Billions	26.71	19.29	22.23	27.12	33.93	43.1
11	Colombia	Colombian Peso, Thousands	..	..	..	..	..	..
12	Czech Republic	Euro, Thousands	..	103986	219659	153843	18284	9492
13	Denmark	Euro, Millions	263.48	204.61	158.68	280.18	186.36	270.46
14	Estonia	Euro	..	..	..	..	..	..
15	Finland	Euro, Millions	189	218	146	351	148	185
16	France	Euro, Millions	1987	2411	2385	2915	3537	2389
17	Greece	Euro, Thousands	19027	32715	16679	25030	10097	0
18	Hungary	Forint, Millions	3949	13782	720	6982	11308	19361
19	Ireland	Euro, Millions	225.9	242.9	288.1	310.2	274.4	268.9
20	Israel	US Dollar, Billions	1.76	2.08	1.12	2.1	2.41	1.51
21	Italy	Euro, Millions	361	556	358	352	582	639
22	Japan	Yen, Billions	193	136	87	113	124	102
23	Korea	Won, Millions	0.99	0.72	0.87	1.09	1.26	1.23

... | V9 Domestic Tourism | V10 Inbound Tourism | V11 Expenditures | V12 Capital

Figure 18: Sample Data for the variable capital

Sample data for the variable annual average working hours per worker

	A	B	C	D	E
1	**Country**	**Year**	**Concat**	**Average annual working hours per worker**	
2	Argentina	2005	Argentina2005	1761.399	
3	Argentina	2006	Argentina2006	1765.601	
4	Argentina	2007	Argentina2007	1780.557	
5	Argentina	2008	Argentina2008	1781.425	
6	Argentina	2009	Argentina2009	1742.904	
7	Argentina	2010	Argentina2010	1751.766	
8	Argentina	2011	Argentina2011	1750.974	
9	Argentina	2012	Argentina2012	1726.258	
10	Argentina	2013	Argentina2013	1714.871	
11	Argentina	2014	Argentina2014	1695.364	
12	Argentina	2015	Argentina2015	1691.536	
13	Argentina	2016	Argentina2016	1691.536	
14	Argentina	2017	Argentina2017	1691.536	
15	Argentina	2018	Argentina2018	1703.343	
16	Argentina	2019	Argentina2019	1746.456	
17	Argentina	2020	Argentina2020	1800.231	
18	Australia	2005	Australia2005	1803.01	
19	Australia	2006	Australia2006	1791.993	
20	Australia	2007	Australia2007	1792.652	
21	Australia	2008	Australia2008	1790.54	
22	Australia	2009	Australia2009	1761.796	

Final Dataset | **Average hours per employee** | Team Data | V1 Country

Figure 19: Sample Data for the variable annual average working hours per worker

Sample data for literacy rate, crime rate

ID	Country	Year	Concat	Male	Female	Total Population	Total Revenue Gain (Billions in USD)	Tourist_arrival	Tourist total Spendings (millions in USD)	Literacy rate/year	Crime Rate/ 1000000 population
1	Argentina	2005	Argentina2005	1951014	1871587	3822601	179.543	3823000	17770	High	5.65
1	Argentina	2006	Argentina2006	1759837	1675605	3435442	142.906	4173000	20235	High	5.37
1	Argentina	2007	Argentina2007	1779247	1689717	3468964	236.516	4562000	25385	High	5.39
1	Argentina	2008	Argentina2008	1798871	1703947	3502818	317.578	4700000	29935	High	5.92
1	Argentina	2009	Argentina2009	1818712	1718298	3537010	360.951	4308000	29040	High	6.53
1	Argentina	2010	Argentina2010	1838771	1732769	3571540	483.014	6800000	32240	High	5.83
1	Argentina	2011	Argentina2011	1865789	1757173	3622962	638.429	6703000	38010	High	6.08
1	Argentina	2012	Argentina2012	1890649	1780362	3671011	808.131	6497000	42330	High	6.34
1	Argentina	2013	Argentina2013	1911781	1800830	3712611	1043.967	6510000	42105	High	7.28
1	Argentina	2014	Argentina2014	1927925	1817374	3745299	1425.148	7165000	39185	High	7.57
1	Argentina	2015	Argentina2015	1936908	1827828	3764736	1872.915	6816000	46740	High	6.59
1	Argentina	2016	Argentina2016	1933361	1824348	3757709	2522.854	6668000	61365	High	6.03
1	Argentina	2017	Argentina2017	1928846	1820000	3748846	3202.762	6711000	70900	High	5.21
1	Argentina	2018	Argentina2018	1923444	1814785	3738229	4203.177	6942000	65420	High	5.32
1	Argentina	2019	Argentina2019	1917283	1808879	3726162	6143.732	7399000	49225	High	5.68
1	Argentina	2020	Argentina2020	1910563	1802426	3712989	1.702	5705970	13730	High	6.21
2	Australia	2005	Australia2005	10019633	10157211	20176844	225.719	4852654	17235	High	1.28
2	Australia	2006	Australia2006	10159424	10291542	20450966	252.625	5221315	20374	High	1.37

Figure 20: Sample Data for the variable crime rate and literacy rate

Final Dataset : Merged to a single sheet

#	Countries	Year	Concat	Male	Female	Total Population DB	Real Population	How Off	Projected Population	Employee % to Total %	# of people employed in
1	Argentina	2005	Argentina2005	18951245	19782161	38733406	38733406	100.00%			
1	Argentina	2006	Argentina2006	19141314	19985925	39127239	39127239	100.00%			
1	Argentina	2007	Argentina2007	19335197	20194243	39529440	39529440	100.00%			
1	Argentina	2008	Argentina2008	19532999	20407235	39940234	39940234	100.00%		9.77%	390
1	Argentina	2009	Argentina2009	19734800	20625028	40359828	40359828	100.00%		9.65%	389
1	Argentina	2010	Argentina2010	19940704	20847749	40788453	40788453	100.00%		9.93%	405
1	Argentina	2011	Argentina2011	20180791	21080699	41261490	41261490	100.00%		10.00%	412
1	Argentina	2012	Argentina2012	20420391	21312880	41733271	41733271	100.00%		10.17%	424
1	Argentina	2013	Argentina2013	20659037	21543898	42202935	42202935	100.00%			
1	Argentina	2014	Argentina2014	20896203	21773297	42669500	42669500	100.00%			
1	Argentina	2015	Argentina2015	21131346	22000620	43131966	43131966	100.00%			
1	Argentina	2016	Argentina2016	21364470	22225898	43590368	43590368	100.00%			
1	Argentina	2017	Argentina2017	21595623	22449188	44044811	44044811	100.00%			
1	Argentina	2018	Argentina2018	21824372	22670130	44494502	44494502	100.00%			
1	Argentina	2019	Argentina2019	22050332	22888380	44938712	44938712	100.00%			
1	Argentina	2020	Argentina2020	22273132	23103631	45376763	45376763	100.00%	45195774		
1	Argentina	2021	Argentina2021						45605826		
1	Argentina	2022	Argentina2022						46010234		
1	Argentina	2023	Argentina2023						46409168		
1	Argentina	2024	Argentina2024						46803049		

Figure 21: Sample Data for the merged data set

Employee % to Total %	# of people employed in tour	Less than 5 Years	5-10 Years	10+ Years	International Visitor Tourist Arrival	Total Revenue Gain	Total Tourist Spending	Expenditures (millions)	Lite
					5221315	289.86	20374		Hig
					5854442	271.886	30476.2	$ 27,620.00	Hig
4.70%	998712				5621800	5621800	279.257	27620	$ 25,629.60	Hig
4.60%	997816				5515000	5515000	346.463	25629.6	$ 31,916.50	Hig
4.60%	1013461	6326900	1982100	2769000	5671000	5 671 000	399.797	31916.5	$ 39,381.50	Hig
4.60%	1027641	6360500	2014100	2798200	5899800	5 899 800	411.598	39381.5	$ 41,632.80	Hig
4.70%	1068473	6412500	2102000	2834100	5990800	5 990 800	401.892	41632.8	$ 42,428.30	Hig
4.70%	1087022	6248900	2249600	2904400	6292800	6 292 800	361.522	42428.3	$ 38,774.50	Hig
4.90%	1150309	6157400	2338000	3035400	6724500	6 724 500	344.314	38774.5	$ 34,102.90	Hig
4.90%	1166984	6306100	2411400	3031600	7138200	7 138 200	365.268	34102.9	$ 35,674.60	Hig
5.10%	1233736	6483250	2318730	3153600	7852900	7 852 900	397.953	35674.6	$ 39,644.90	Hig
5.20%	1279297	6657500	2320600	3294900	8557554	8 557 554	401.232	39644.9	$ 42,478.00	Hig
5.30%	1324082	5736000	2229000	3346000	9071700	9 071 700	380.012	42478	$ 41,367.20	Hig
5.00%	1268287	5858000	2324000	3372000	9344000	9 344 000	472.4	41452	$ 7,964.40	Hig
3.90%	1002226	0			6737200	6 737 200		30023	..	Hig

Figure 22: Sample Data for the merged data set

1	5-10 Years	10+ Years	International Visitor	Tourist Arrival	Total Revenue Gain	Total Tourist Spending	Expenditures (millions)	Literacy	Crime	Average annual working hours per w	Capital %
29				5221315	289.86	20374		High	1.37	1791.993	
30				5854442	271.886	30476.2	$ 27,620.00	High	1.22	1792.652	
31			5621800	5621800	279.257	27620	$ 25,629.60	High	1.22	1790.54	27 620
32			5515000	5515000	346.463	25629.6	$ 31,916.50	High	1.21	1761.796	25 630
33	1982100	2769000	5671000	5 671 000	399.797	31916.5	$ 39,381.50	High	1.04	1769.309	31 917
34	2014100	2798200	5899800	5 899 800	411.598	39381.5	$ 41,632.80	High	1.1	1768.649	39 382
35	2102000	2834100	5990800	5 990 800	401.892	41632.8	$ 42,428.30	High	1.06	1759.519	41 633
36	2249600	2904400	6292800	6 292 800	361.522	42428.3	$ 38,774.50	High	1.05	1755.827	42 428
37	2338000	3035400	6724500	6 724 500	344.314	38774.5	$ 34,102.90	High	1.03	1747.937	38 775
38	2411400	3031600	7138200	7 138 200	365.268	34102.9	$ 35,674.60	High	0.99	1747.009	34 103
39	2318730	3153600	7852900	7 852 900	397.953	35674.6	$ 39,644.90	High	0.94	1734.215	35 675
40	2320600	3294900	8557554	8 557 554	401.232	39644.9	$ 42,478.00	High	0.83	1731.494	39 645
41	2229000	3346000	9071700	9 071 700	380.012	42478	$ 41,367.20	High	0.89	1734.546	42 478
42	2324000	3372000	9344000	9 344 000	472.4	41452	$ 7,964.40	High	0.82	1736.426	41 452
43			6737200	6 737 200		30023	..	High	0.8	1752.489	

Figure 23: Sample Data for the merged data set

Describe, Explore and Verify Data Quality

One variable summary in Stat Tool report

StatTools Report

Analysis: One Variable Summary
Performed By: Hemalatha Ramakrishna
Date: 7 May, 2022
Updating: Live

One Variable Sum	Year Data Set #1	Male Data Set #1	Female Data Set #1	Country Population Data Set #1
Mean	2012.500	34839070.96	34510120.22	69349191.21
Variance	21.278	9389434162717300.00	8159724550810460.00	35039927423343200.00
Std. Dev.	4.613	96899092.68	90331193.68	187189549.45
Skewness	0.0000	5.7260	5.5124	5.6251
Kurtosis	1.7905	37.3053	35.3215	36.3676
Median	2012.000	5668635.00	5838303.00	11506938.00
Mean Abs. Dev.	4.000	41188405.74	40234269.15	81420732.81
Mode	2005.000	622794.63	2777118.29	1331278.63
Minimum	2005.000	149182.00	147552.00	296734.00
Maximum	2020.000	717100970.00	662903415.00	1380004385.00
Range	15.000	716951788.00	662755863.00	1379707651.00
Count	768	768	768	768
Sum	1545600.000	26756406498.00	26503772329.00	53260178846.00
1st Quartile	2008.000	2654099.00	2773713.00	5415978.00
3rd Quartile	2016.000	28302528.00	30050368.00	58399863.00
Interquartile Range	8.000	25648429.00	27276655.00	52983885.00
1.00%	2005.000	161476.00	159817.00	320723.00

Final Sheet | One Var Summary | Correlation and Covariance | Regression

Figure 24: Sample Screenshot of One variable summary in Stat Tool report

Variable Domestic Revenue (in Billion)

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name	Mean	400.39
Domestic Revenue (Billions)	Variance	328904.96
	Std. Dev.	573.50
Data Volume (number of observation/rows)	Skewness	3.07
768	Kurtosis	20.00
	Median	175.00

Meaning of the attribute It represents the total domestic revenue in Billions for different countries and different years	Mean Abs. Dev.	397.41
	Mode	8.00
	Minimum	4.00
	Maximum	6143.00
Meaning of the attribute in business terms Total domestic revenue in Billion	Range	6139.00
	Count	768.00
	Sum	307498.00
	1st Quartile	59.00
Attribute types (select from the list) Continuous	3rd Quartile	448.00
	Interquartile Range	389.00
	Missing / Blank	0.00

Explore Data

Data Type	Summary measures	Assessment
Numerical	Mean	The mean of domestic revenue is 400.39 Billion which indicates the industry is lucrative on average. Considering that people travel from country to country on a consistant basis, it is likely that the average will only increase as long as there is no heavy restrictions on traveling in the future.
	Variance	The variance of the domestic revenue is 328904.96 which indicate that there are some countries that are making a lot of money in the tourism industry and there are countries that are making very little. It can also indicate that the domestic population are not spending money for tourism related activities.
Domestic Revenue (In Billions)	Standard Deviation	The standard deviation is 573.5 which is larger than the mean. This indicates that the values are spread apart. This makes sense as we are looking at the world and some locations are very poor and other areas are not.
	Count	The count of rows is 768. Each country had multiple years which contribute to the high number. There were also many countries within the dataset.
	Median	The median is 175 Billion. This number is relatively low compare to the mean which indicates that the distribution of data is skewed to the right.
	Mode	The mode is 8 Billion indicating that this is the value that appear most frequently. This number is relatively low compare to the mean which indicates that the distribution of data is skewed to the right.
	Minimum	The minimum number is 4 Billion. Certain countries have very low domestic revenue which can help us remove them from the available choices for our country selection.
	Maximum	The maximum number is 6143 Billion. This is a high number indicating that the country with this number observes a very active tourism domestically.
	Range	The range is 6139. This shows that there is a large difference in the amount of domestic revenue between countries.
	Sum	The sum of all the years is 307498 Billion of dollars spent on domestic revenue for the years 2005-2020.

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Country	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	No			
	Duplicate	Duplicated records (observations)	Yes	There were few repeated countries which was removed	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	Yes	Spellings of few countries were wrong	Correct	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable International Revenue (in Billion)

Describe Data

Data description for International Revenue (in Billion)

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name International Revenue (Billions)	Mean	28.35
	Variance	38680.31
	Std. Dev.	196.67
Data Volume (number of observation/rows) 768	Skewness	24.58
	Kurtosis	646.93
	Median	8.44
Meaning of the attribute	Mean Abs. Dev.	32.22

It represents the total international revenue in Billions for different countries and different years	Mode	1.24
	Minimum	0.39
	Maximum	5234.24
Meaning of the attribute in business terms Total international revenue in Billion	Range	5233.85
	Count	768.00
	Sum	21769.33
	1st Quartile	3.01
Attribute types (select from the list) Continuous	3rd Quartile	21.86
	Interquartile Range	18.85
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
International Revenue (in Billion)	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is “noise” or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Total Revenue (in Billion)

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name Total Revenue (Billions)	Mean	428.73
	Variance	398425.52
	Std. Dev.	631.21
Data Volume (number of observation/rows) 768	Skewness	3.53
	Kurtosis	25.00
	Median	187.47
Meaning of the attribute It represents the total revenue in Billions for different countries and different years	Mean Abs. Dev.	424.55
	Mode	4.65
	Minimum	4.39
	Maximum	6444.24
Meaning of the attribute in business terms Total revenue in Billion	Range	6439.85
	Count	768.00
	Sum	329267.33
	1st Quartile	62.54
Attribute types (select from the list) Continuous	3rd Quartile	476.66
	Interquartile Range	414.13
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Total Revenue (in Billion)	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Tourism Employment (%)

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name Tourism Employment (%)	Mean	6.95
	Variance	6.35
	Std. Dev.	2.52
Data Volume (number of observation/rows) 768	Skewness	0.71
	Kurtosis	3.67
	Median	6.78
Meaning of the attribute	Mean Abs. Dev.	1.99

It represents the total percentage of employment in tourism industry	Mode	4.40
	Minimum	2.00
	Maximum	16.90
Meaning of the attribute in business terms Percentage of Employment in Tourism Industry	Range	14.90
	Count	768.00
	Sum	5334.76
	1st Quartile	5.00
Attribute types (select from the list)	3rd Quartile	8.40
Continuous	Interquartile Range	3.40
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Tourism Employment (%)	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	Yes	Many values were not in percentage, Data was converted to percentage	Correct	

Variable Annual Avg WorkHrs

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name Annual Avg WorkHrs	**Mean**	**1822.64**
	Variance	**57994.93**
	Std. Dev.	**240.82**
Data Volume (number of observation/rows) 768	**Skewness**	**0.36**
	Kurtosis	**2.72**
	Median	**1789.00**

Meaning of the attribute It represents the average working hours of employees annually	Mean Abs. Dev.	193.29
	Mode	1878.00
	Minimum	1353.00
	Maximum	2454.00
Meaning of the attribute in business terms Annual average working hours of employees	Range	1101.00
	Count	768.00
	Sum	1399787.00
	1st Quartile	1663.00
Attribute types (select from the list)	3rd Quartile	1982.00
Continuous	Interquartile Range	319.00
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Annual Avg WorkHrs	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	Yes	Few values has alphabets which was removed	Correct	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	Yes	Few values had different format	Correct	

Variable Crime Rate

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name Crime Rate	Mean	4.41
	Variance	63.46
	Std. Dev.	7.97
Data Volume (number of observation/rows) 768	Skewness	2.95
	Kurtosis	11.26
	Median	1.30

Meaning of the attribute It represents the crime rate of different countries in different year	Mean Abs. Dev.	4.84
	Mode	1.00
	Minimum	0.15
	Maximum	45.80
Meaning of the attribute in business terms Represents the crime rate	Range	45.65
	Count	768.00
	Sum	3389.91
	1st Quartile	0.86
Attribute types (select from the list) Continuous	3rd Quartile	3.61
	Interquartile Range	2.75
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Crime rate	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable CoCur_1DCon

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name CoCur_1DCon	Mean	454.75
	Variance	4414354.26
	Std. Dev.	2101.04
Data Volume (number of observation/rows) 768	Skewness	6.07
	Kurtosis	39.85
	Median	3.21

Meaning of the attribute It represents the how much is one US dollar in another country's currency	Mean Abs. Dev.	770.40
	Mode	0.92
	Minimum	0.77
	Maximum	14308.00
Meaning of the attribute in business terms One US dollar in another country's currency	Range	14307.23
	Count	768.00
	Sum	349246.99
	1st Quartile	0.92
Attribute types (select from the list)	3rd Quartile	22.48
Continuous	Interquartile Range	21.56
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
CoCur_1DCon	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	Yes	Few values had spelling errors		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	Yes	Few values had wrong currency value	Correct	
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	Yes	Few values had wrong currency value	Correct	
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Sal/Hr

Describe Data

Numerical: Continuous (Cross-sectional)	Summary Measures	Value
Attribute/Variable Name Sal/Hr	Mean	1224.32
	Variance	15980024.69
	Std. Dev.	3997.50
Data Volume (number of observation/rows) 768	Skewness	3.82
	Kurtosis	16.77

	Median	22.41
Meaning of the attribute It represents the salary per hour of employees in different countries	Mean Abs. Dev.	2042.18
	Mode	110.00
	Minimum	2.01
	Maximum	19619.00
Meaning of the attribute in business terms Shows salary per hour of employees	Range	19616.99
	Count	768.00
	Sum	940275.95
	1st Quartile	8.00
Attribute types (select from the list) Continuous	3rd Quartile	96.40
	Interquartile Range	88.40
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Sal/Hr	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	Yes	Few values had spelling errors		
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Country population

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Country Population	Mean	69349191.21
	Variance	35039927423343200.00
	Std. Dev.	187189549.45
Data Volume (number of observation/rows) 768	Skewness	5.6251
	Kurtosis	36.3676
	Median	11506938.00

Meaning of the attribute It represents the total population of different countries for different year	Mean Abs. Dev.	81420732.81
	Mode	1331278.63
	Minimum	296734.00
	Maximum	1380004385.00
Meaning of the attribute in business terms The total population count of different countries	Range	1379707651.00
	Count	768
	Sum	53260178846.00
	1st Quartile	5415978.00
Attribute types (select from the list) Numerical: Discrete	3rd Quartile	58399863.00
	Interquartile Range	52983885.00
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Country population	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few missing values which was filled manually	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values, All the duplicate value were deleted	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	Had a vague value which was corrected	Correct	
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Tourist_arrival

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Tourist_Arrival	Mean	25068659.19
	Variance	1650423208504890.00
	Std. Dev.	40625401.03
Data Volume (number of observation/rows) 768	Skewness	2.85
	Kurtosis	11.14
	Median	8628000.00

Meaning of the attribute It represents total number of tourist arrival for a particular country for different year	Mean Abs. Dev.	25665668.19
	Mode	2409000.00
	Minimum	429000.00
	Maximum	212094373.00
Meaning of the attribute in business terms Total number of tourist arrival	Range	211665373.00
	Count	768.00
	Sum	19252730257.00
	1st Quartile	4504000.00
Attribute types (select from the list) Numerical: Discrete	3rd Quartile	25318000.00
	Interquartile Range	20814000.00
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Tourist arrival	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	No			
	Duplicate	Duplicated records (observations)	No			
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Employed Tourists

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Employed Tourist	Mean	4297151.29
	Variance	115272483549522.00
	Std. Dev.	10736502.39
Data Volume (number of observation/rows) 768	Skewness	5.80
	Kurtosis	40.91
	Median	1021425.00
Meaning of the attribute It represents the total number of people employed the in tourism industry	Mean Abs. Dev.	4868789.51
	Mode	43586.63
	Minimum	24421.00
	Maximum	96600306.00
Meaning of the attribute in business terms Total number of people employed in the tourism industry	Range	96575885.00
	Count	768.00
	Sum	3300212190.00
	1st Quartile	326207.00
Attribute types (select from the list) Numerical: Discrete	3rd Quartile	4180408.00
	Interquartile Range	3854201.00
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Employed Tourists	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few N/A values	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Projected Population

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Projected Population	Mean	77908075.10
	Variance	47666050081581700.00
	Std. Dev.	218325559.84
Data Volume (number of observation/rows) 768	Skewness	5.70
	Kurtosis	36.89
	Median	12122414.79
Meaning of the attribute It represents projected population in further years of different countries	Mean Abs. Dev.	92751424.58
	Mode	10075605.00

	Minimum	370563.00
	Maximum	1578869824.45
Meaning of the attribute in business terms Shows the projected population in further years	Range	1578499261.45
	Count	768.00
	Sum	59833401673.43
	1st Quartile	5688498.78
Attribute types (select from the list) Numerical: Discrete	3rd Quartile	60155851.00
	Interquartile Range	54467352.22
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Projected Population	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few N/A values	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	Yes	There were few values with very low number	Correct	
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	There were few values with very low number	Correct	
	High Cardinality	A high number of values in a set	Yes		Correct	
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	Yes	Had few float values which was converted to integer	Correct	

Variable Male Population

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Male	Mean	34839070.96
	Variance	9389434162717300.00
	Std. Dev.	96899092.68
Data Volume (number of observation/rows) 768	Skewness	5.73
	Kurtosis	37.31
	Median	5668635.00
Meaning of the attribute It represents total male population of different countries for different year	Mean Abs. Dev.	41188405.74
	Mode	622794.63
	Minimum	149182.00
	Maximum	717100970.00
Meaning of the attribute in business terms Total male population count of different countries	Range	716951788.00
	Count	768.00
	Sum	26756406498.00
	1st Quartile	2654099.00
Attribute types (select from the list) Numerical: Discrete	3rd Quartile	28302528.00
	Interquartile Range	25648429.00
	Missing / Blank	0

Explore Data

Data Type	Summary measure	Assessment
Discrete	Mean	3.483907e+07 is the mean, the average value for male population to have in a country.
	Variance	9389434162717296.0 variance is high, that means we have larger variability in our dataset for that column. In the other way, we can say more values are spread out around our mean value.
Male	Standard Deviation	9.689909e+07 high standard deviation indicates data are more spread out therfore its relaible for calculation
	Count	768 is the row count
	Median	6815843.0 is the median theefore the data is skewed to the right indicating its positive. Positive skew refers to longer or fatter tail on right side
	Mode	Each rows has unique value therefore 768 values for male will be shown.
	Minimum	1.491820e+05 is minimum value we have in the dataset for visualization on later part.
	Maximum	7.171010e+08 is the maximum value we have in the dataset for visualization on later part.
	Range	The range of countries is from 0 to 768 rows
	Sum	26756406498 is the total population of male in world.

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Male Population	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few N/A values	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	Yes	Had few float values which was converted to integer	Correct	

Variable Female Population

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Female	Mean	34510120.22
	Variance	8159724550810460.00
	Std. Dev.	90331193.68
Data Volume (number of observation/rows) 768	Skewness	5.51
	Kurtosis	35.32
	Median	5838303.00
Meaning of the attribute	Mean Abs. Dev.	40234269.15

It represents total female population of different countries for different year	Mode	2777118.29
	Minimum	147552.00
	Maximum	662903415.00
Meaning of the attribute in business terms Total female population count of different countries	Range	662755863.00
	Count	768.00
	Sum	26503772329.00
	1st Quartile	2773713.00
Attribute types (select from the list) Numerical: Discrete	3rd Quartile	30050368.00
	Interquartile Range	27276655.00
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Female Population	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few N/A values	Correct	
	Duplicate	Duplicated records (observations)	Yes	There were few repeated values	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	Yes	Had few float values which was converted to integer	Correct	

Variable Year

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Year	Mean	2012.5
	Variance	21.27770535
	Std. Dev.	4.612776316
Data Volume (number of observation/rows) 768	Skewness	0
	Kurtosis	1.790530768
	Median	2012
Meaning of the attribute It represents Year from 2005 to 2020	Mean Abs. Dev.	4
	Mode	2005

	Minimum	2005
	Maximum	2020
Meaning of the attribute in business terms Year from 2005 to 2020	Range	15
	Count	768
	Sum	1545600
	1st Quartile	2008
Attribute types (select from the list) Numerical: Discrete	3rd Quartile	2016
	Interquartile Range	8
	Missing / Blank	0

Explore Data

Data Type	Summary measure	Assessment
Discrete	Mean	2012.5 is the mean, but for year it doesn't add any value to the calculation.
	Variance	21.2777053455019 is the variance but compared to other columns it doesn't add any value.
Year	Standard Deviation	4.612776 is the standard deviation which indicates that the the value of years are clustered around mean that is 2012.5
	Count	768 is the row count
	Median	2012.5 is the median which is same as mean, but for year the value isnt useful for calculation
	Mode	there 16 values for each countries therefore each year has 16 values. The mode will be all year.
	Minimum	2005 is early year we have in the dataset for visualization on later part.
	Maximum	2020 is the lasr year we have in the dataset for visualization on later part.
	Range	The range of countries is from 0 to 768 rows
	Sum	1545600 which isnt of use in the calculation for further evaluation.

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Year	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	No			
	Duplicate	Duplicated records (observations)	No			
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Upcoming_Year

Describe Data

Numerical: Discrete	Summary Measures	Value
Attribute/Variable Name Upcoming_Year	Mean	2028.50
	Variance	21.28
	Std. Dev.	4.61
Data Volume (number of observation/rows) 768	Skewness	0.00
	Kurtosis	1.79
	Median	2028.00
Meaning of the attribute It represents Year from 2021 to 2036	Mean Abs. Dev.	4.00
	Mode	2021.00
	Minimum	2021.00
	Maximum	2036.00

Meaning of the attribute in business terms		
Year from 2021 to 2036	Range	15.00
	Count	768.00
	Sum	1557888.00
	1st Quartile	2024.00
Attribute types (select from the list)	3rd Quartile	2032.00
Numerical: Discrete	Interquartile Range	8.00
	Missing / Blank	0

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Upcoming_Year	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	No			
	Duplicate	Duplicated records (observations)	No			
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Literacy

Describe Data

Categorical	Literacy rate (High, Medium, Low)	High	Medium	Low
Attribute/Variable Name	**Count / Frequency**	**604**	**25**	**139**

Literacy	**Percentage**	**78.64%**	**3.25%**	**18.09%**
Data Volume (number of observation/rows)				
768	**Mode**	**High**		
	Minimum	**Low**		
Meaning of the attribute	**Maximum**	**High**		
It shows the literacy rate of each country in different years				
Meaning of the attribute in business terms				
Literacy rate of each country				
Attribute types (select from the list)				
Categorical				

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Literacy	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	No			
	Duplicate	Duplicated records (observations)	No			
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	No			
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	No			
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	No			
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Country

Describe Data

Categorical	Country	Count / Frequency	Percentage
Attribute/Variable Name	Argentina	16	2.08%
Country	Australia	16	2.08%
	Austria	16	2.08%
Data Volume (number of observation/rows)	Belgium	16	2.08%
768	Brazil	16	2.08%
	Bulgaria	16	2.08%
Meaning of the attribute	Canada	16	2.08%
It shows the name of each country	Chile	16	2.08%
	Colombia	16	2.08%
	Costa Rica	16	2.08%

Meaning of the attribute in business terms Shows the name of each country			
	Czech Republic	16	2.08%
	Denmark	16	2.08%
	Estonia	16	2.08%
	Finland	16	2.08%
Attribute types (select from the list) Categorical	France	16	2.08%
	Germany	16	2.08%
	Greece	16	2.08%
	Hungary	16	2.08%
	Iceland	16	2.08%
	India	16	2.08%
	Indonesia	16	2.08%
	Ireland	16	2.08%
	Israel	16	2.08%
	Italy	16	2.08%
	Japan	16	2.08%
	Korea	16	2.08%
	Latvia	16	2.08%
	Lithuania	16	2.08%
	Luxembourg	16	2.08%
	Mexico	16	2.08%
	Netherlands	16	2.08%
	New Zealand	16	2.08%
	Norway	16	2.08%
	Poland	16	2.08%
	Portugal	16	2.08%
	Romania	16	2.08%
	Russia	16	2.08%
	Saudi Arabia	16	2.08%
	Singapore	16	2.08%
	Slovak Republic	16	2.08%
	Slovenia	16	2.08%
	South Africa	16	2.08%
	Spain	16	2.08%
	Sweden	16	2.08%
	Switzerland	16	2.08%
	Turkey	16	2.08%
	United Kingdom	16	2.08%
	United States	16	2.08%

Explore Data

Data Type	Summary measure	Assessment
Categorical	Mean	Mean cannot be calculated because it's a string value therefore no calculation needed.
	Variance	Variance cannot be calculated because it's a string value therefore no calculation needed.
Countries	Standard Deviation	Standard Deviation cannot be calculated because it's a string value therefore no calculation needed.
	Count	768
	Median	Median cannot be calculated because it's a string value therefore no calculation needed.
	Mode	Every attribute would have same 16 count therefore Mode would be all countries.
	Minimum	There will be no minimum value for string as all have count of 16
	Maximum	There will be no maximum value for string as all have count of 16
	Range	The range of countries is from 0 to 768 rows
	Sum	All the countries values are string so no sum will be calculated.

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Country	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	No			
	Duplicate	Duplicated records (observations)	Yes	There were repeated countries	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	Yes	There were few spelling errors	Correct	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	Few countries names were unknown	Correct	
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	Yes	There were few values with different country names	Correct	
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Variable Currency Name

Describe Data

Categorical	Country	Count / Frequency	Percentage
Attribute/Variable Name	Col Peso	16	2.08%
Currency Name	ARS	16	2.08%
	ATS	16	2.08%
Data Volume (number of observation/rows)	AUD	16	2.08%
768	BEF	16	2.08%
	BRL	16	2.08%
Meaning of the attribute	CD	16	2.08%
It shows the currency for different countries	Costa Rican	16	2.08%

	CPeso	16	2.08%
	Czech Koruna	16	2.08%
Meaning of the attribute in business terms It shows the currency for different countries	Danish Krone	16	2.08%
	Dollar	16	2.08%
	Euro	240	31.25%
	Forint	16	2.08%
Attribute types (select from the list) Categorical	Indo Rupiyah	16	2.08%
	Israeli	16	2.08%
	Krona	16	2.08%
	Leu	16	2.08%
	LEV	16	2.08%
	Lira	16	2.08%
	Mexi Peso	16	2.08%
	Nor Korne	16	2.08%
	NZ Dollar	16	2.08%
	Pound	16	2.08%
	Rand	16	2.08%
	Riyal	16	2.08%
	Ruble	16	2.08%
	Rupees	16	2.08%
	Sek	16	2.08%
	Singa Dollar	16	2.08%
	Swiss Franc	16	2.08%
	won	16	2.08%
	Yen	16	2.08%
	Zoty	16	2.08%

Verify Data Quality

Variable Name	Data Quality Issue	Description/ Example	Proble	Assessment	Data Cleaning	Construct Data
	Check coverage	All possible values are represented •Use metadata (e.g., domain, range, dependency, distribution)	No			
Currency Name	**Meaning Of Attributes**	Verify that the meanings of attributes and contained values fit together	No			
	Missing Attribute or blank fields	How will you address this?	Yes	There were few null values		
	Duplicate	Duplicated records (observations)	Yes	There were repeated values	Correct	
	Spelling and format	E.g., same value but sometimes beginning with a lower-case letter, sometimes with an upper-case letter	Yes	There were few spelling errors	Correct	
	Deviations	Decide whether a deviation is "noise" or may indicate an interesting phenomenon	No			
	Plausibility	E.g., all fields having the same or nearly the same values	No			
	Conflict with Common Sense	E.g., teenagers with high income levels.	Yes	There were few unknown currency	Correct	
	High Cardinality	A high number of values in a set	No			
	Outliers	An observation that lies well outside of the norm.	No			
	Redundant Input	Does not give any new information that was not already explained by other inputs	Yes	Few values had wrong currency	Correct	
	Sparseness	Any data which as very large zero value and very little no zero value	No			
	Irrelevant Input	Does not provide information about the target (dependent Variable)	No			
	Unstructured Data	Unstructured data is data that does not follow a specified format	No			

Data Preparation

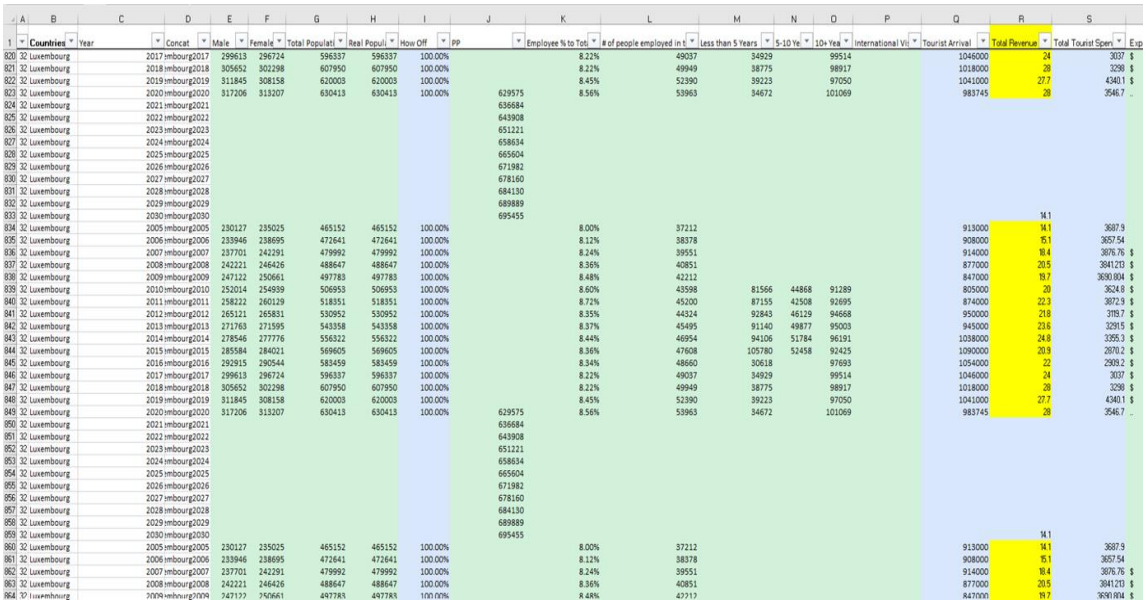
Data cleaning is the process of identifying and fixing problems in a dataset. The purpose of data cleansing is to correct data that are inaccurate, incomplete, malformed, duplicated, or irrelevant to the purpose of the dataset. This is typically achieved by replacing, modifying, or deleting data that falls into one of these categories.

When combining multiple data sources, it is likely that the data will be duplicated or mislabeled. If the data is wrong, the results and algorithms are unreliable, even if they look correct. Since the process is different for each dataset, there is no absolute way to indicate the exact steps of the data cleansing process. Our decisions are usually based on datasets, so if the quality of the data is poor, our results

will not be accurate. Therefore, data cleaning is very important because you can get high-quality data that leads to better quality decisions.

Not all data is good data in the data set. There were few junk data. This dataset used for this analysis contained some null values, and some datasets contained blank / missing values, so while focusing on the required data set, these data were deleted and filtered. Unnecessary columns were removed, and few columns were split. Below are the few steps taken to clean the dataset.

Sample Final Dataset combing all data set in one excel sheet



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1		Countries	Year	Concat	Male	Female	Total Populati	Real Popula	How Off	PP	Employee % to Tota	# of people employed in t	Less than 5 Years	5-10 Ye	10+ Yea	International Vi	Tourist Arrival	Total Revenue	Total Tourist Spen
820	32	Luxembourg	2017	mbourg2017	299613	296724	596337	596337	100.00%		8.22%	49037	34929		99514		1046000	24	3037 $
821	32	Luxembourg	2018	mbourg2018	305652	302298	607950	607950	100.00%		8.22%	49949	38775		98917		1018000	28	3298 $
822	32	Luxembourg	2019	mbourg2019	311845	308158	620003	620003	100.00%		8.45%	52390	39223		97050		1041000	27.7	4340.1 $
823	32	Luxembourg	2020	mbourg2020	317206	313207	630413	630413	100.00%	629575	8.56%	53963	34672		101069		983745	28	3546.7 ..
824	32	Luxembourg	2021	mbourg2021						636684									
825	32	Luxembourg	2022	mbourg2022						643908									
826	32	Luxembourg	2023	mbourg2023						651221									
827	32	Luxembourg	2024	mbourg2024						658634									
828	32	Luxembourg	2025	mbourg2025						665604									
829	32	Luxembourg	2026	mbourg2026						671982									
830	32	Luxembourg	2027	mbourg2027						678160									
831	32	Luxembourg	2028	mbourg2028						684130									
832	32	Luxembourg	2029	mbourg2029						689889									
833	32	Luxembourg	2030	mbourg2030						695455								14.1	
834	32	Luxembourg	2005	mbourg2005	230127	235025	465152	465152	100.00%		8.00%	37212					913000	14.1	3687.9
835	32	Luxembourg	2006	mbourg2006	233946	238695	472641	472641	100.00%		8.12%	38378					908000	15.1	3857.54
836	32	Luxembourg	2007	mbourg2007	237701	242291	479992	479992	100.00%		8.24%	39551					914000	18.4	3876.76 $
837	32	Luxembourg	2008	mbourg2008	242221	246426	488647	488647	100.00%		8.36%	40851					877000	20.5	3841.213 $
838	32	Luxembourg	2009	mbourg2009	247122	250661	497783	497783	100.00%		8.48%	42212					847000	19.7	3690.804 $
839	32	Luxembourg	2010	mbourg2010	252014	254939	506953	506953	100.00%		8.60%	43598	81566	44868	91289		805000	20	3624.8 $
840	32	Luxembourg	2011	mbourg2011	258222	260129	518351	518351	100.00%		8.72%	45200	87155	41508	92695		874000	22.3	3872.9 $
841	32	Luxembourg	2012	mbourg2012	265121	265831	530952	530952	100.00%		8.35%	44324	92843	46129	94668		950000	23.8	3719.7 $
842	32	Luxembourg	2013	mbourg2013	271763	271595	543358	543358	100.00%		8.37%	45495	91140	49877	95003		945000	23.6	3291.5 $
843	32	Luxembourg	2014	mbourg2014	278546	277776	556322	556322	100.00%		8.44%	46954	94106	51784	96191		1038000	24.8	3355.3 $
844	32	Luxembourg	2015	mbourg2015	285584	284021	569605	569605	100.00%		8.36%	47608	105780	52458	92425		1090000	20.9	2870.2 $
845	32	Luxembourg	2016	mbourg2016	292915	290544	583459	583459	100.00%		8.34%	48660	30618		97693		1054000	22	2903.2 $
846	32	Luxembourg	2017	mbourg2017	299613	296724	596337	596337	100.00%		8.22%	49037	34929		99514		1046000	24	3037 $
847	32	Luxembourg	2018	mbourg2018	305652	302298	607950	607950	100.00%		8.22%	49949	38775		98917		1018000	28	3298 $
848	32	Luxembourg	2019	mbourg2019	311845	308158	620003	620003	100.00%		8.45%	52390	39223		97050		1041000	27.7	4340.1 $
849	32	Luxembourg	2020	mbourg2020	317206	313207	630413	630413	100.00%	629575	8.56%	53963	34672		101069		983745	28	3546.7 ..
850	32	Luxembourg	2021	mbourg2021						636684									
851	32	Luxembourg	2022	mbourg2022						643908									
852	32	Luxembourg	2023	mbourg2023						651221									
853	32	Luxembourg	2024	mbourg2024						658634									
854	32	Luxembourg	2025	mbourg2025						665604									
855	32	Luxembourg	2026	mbourg2026						671982									
856	32	Luxembourg	2027	mbourg2027						678160									
857	32	Luxembourg	2028	mbourg2028						684130									
858	32	Luxembourg	2029	mbourg2029						689889									
859	32	Luxembourg	2030	mbourg2030						695455								14.1	
860	32	Luxembourg	2005	mbourg2005	230127	235025	465152	465152	100.00%		8.00%	37212					913000	14.1	3687.9
861	32	Luxembourg	2006	mbourg2006	233946	238695	472641	472641	100.00%		8.12%	38378					908000	15.1	3857.54
862	32	Luxembourg	2007	mbourg2007	237701	242291	479992	479992	100.00%		8.24%	39551					914000	18.4	3876.76 $
863	32	Luxembourg	2008	mbourg2008	242221	246426	488647	488647	100.00%		8.36%	40851					877000	20.5	3841.213 $
864	32	Luxembourg	2009	mbourg2009	247122	250661	497783	497783	100.00%		8.48%	42212					847000	19.7	3690.804 $

Figure 25: Image shows the Final Raw dataset

Our primary purpose of data preparation is to ensure that raw data being put together for processing and analysis for accurate and consistent results for BI and analytics applications will be valid. The most common anomalies we found were missing/null values, inaccuracies, duplicate values, and combining multiple data sets leads to different formats that must be balanced. Data errors, validating data quality, unnecessary columns, and consolidating data sets were big data preparation tasks. Here are some data anomalies we found:

Unexpected Amount of Data: The raw dataset contained a total of 38283 (rows and columns included), which concluded that this data has a lot more random values. To remove the values, we had to search for the values which weren't necessary.

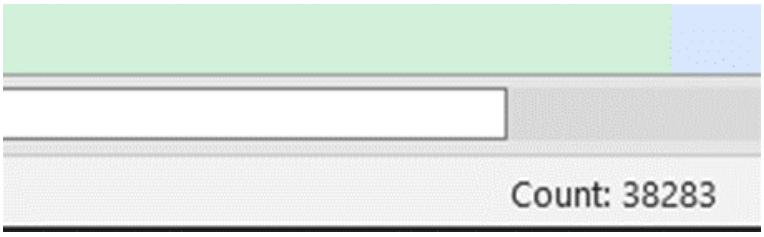


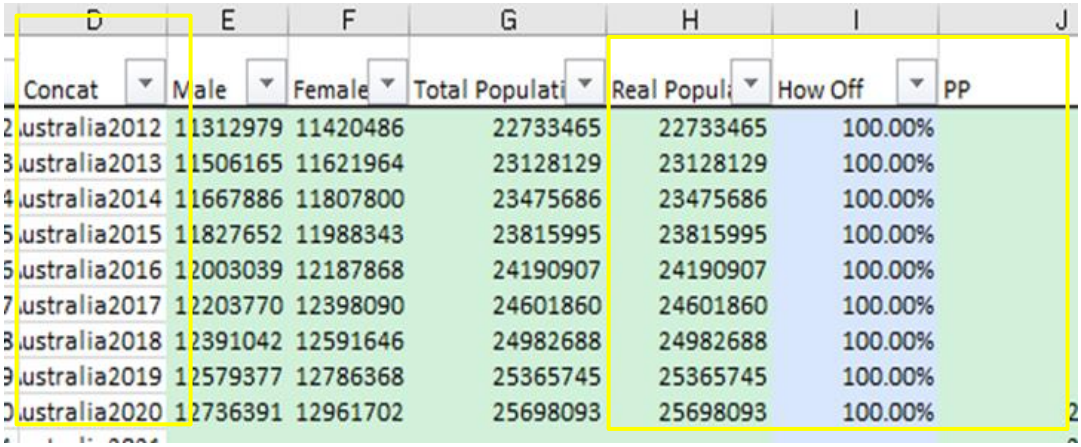
Figure 26: Image shows the unexpected amount of Data

Missing Values/Null Values Some of the country's population values are missing in the figure below. Having a missing or null value in a dataset is a problem that can lead to inaccurate data analysis and visualization. Giving wrong information to clients will lead the company downhill.

Male	Female	Total Populati	Real Popul
18951245	19782161	38733406	38733406
19141314	19985925	39127239	39127239
19335197	20194243		
19532999	20407235		
19734800	20625028	40359828	40359828
19940704	20847749	40788453	40788453
20180791	21080699	41261490	41261490
20420391	21312880	41733271	41733271
20659037	21543898		
20896203	21773297		
21131346	22000620		
21364470	22225898		
21595623	22449188		
21824372	22670130		
22050332	22888380		
22273132	23103631		

Figure 27: Image shows the Missing/Null Values

Unnecessary Columns: When combining multiple datasets or working on one dataset, it sometimes tends to have unnecessary columns for data analysis and visualization. Having unnecessary columns can hinder other important columns from analyzing factors needed to answer some company's goals and requirements questions.



	D	E	F	G	H	I	J
	Concat	Male	Female	Total Populati	Real Popul	How Off	PP
2	ustralia2012	11312979	11420486	22733465	22733465	100.00%	
3	ustralia2013	11506165	11621964	23128129	23128129	100.00%	
4	ustralia2014	11667886	11807800	23475686	23475686	100.00%	
5	ustralia2015	11827652	11988343	23815995	23815995	100.00%	
6	ustralia2016	12003039	12187868	24190907	24190907	100.00%	
7	ustralia2017	12203770	12398090	24601860	24601860	100.00%	
8	ustralia2018	12391042	12591646	24982688	24982688	100.00%	
9	ustralia2019	12579377	12786368	25365745	25365745	100.00%	
10	ustralia2020	12736391	12961702	25698093	25698093	100.00%	

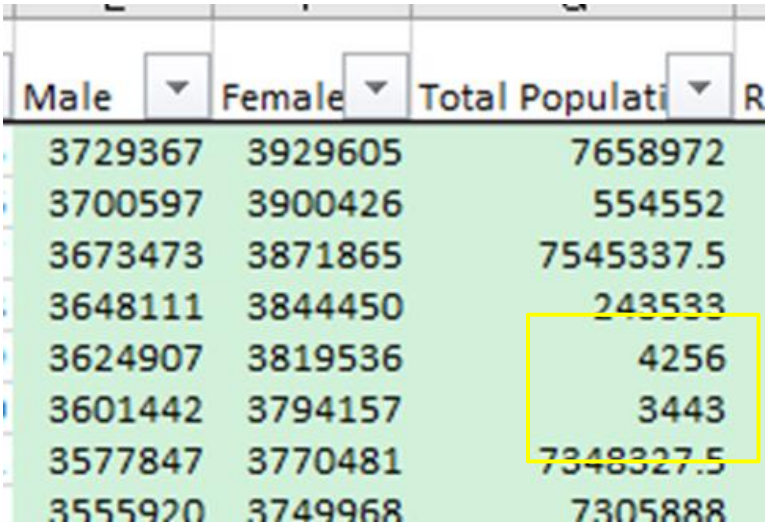
Figure 28: Image shows the Unnecessary Columns

Duplicate Values: While working on the raw dataset for analysis, we encountered an incredible amount of unrequired data known as duplicate values. Having duplicate values in the dataset will generate false analysis that will lead to unnecessary budget spending, hinder personalization, harm brand perception, inaccurate information on sales, etc.

32	Luxembourg		2007	embourg2007	237701	242291
32	Luxembourg		2008	embourg2008	242221	246426
32	Luxembourg		2009	embourg2009	247122	250661
32	Luxembourg		2010	embourg2010	252014	254939
32	Luxembourg		2011	embourg2011	258222	260129
32	Luxembourg		2012	embourg2012	265121	265831
32	Luxembourg		2013	embourg2013	271763	271595
32	Luxembourg		2014	embourg2014	278546	277776
32	Luxembourg		2015	embourg2015	285584	284021
32	Luxembourg		2016	embourg2016	292915	290544
32	Luxembourg		2017	embourg2017	299613	296724
32	Luxembourg		2018	embourg2018	305652	302298
32	Luxembourg		2019	embourg2019	311845	308158
32	Luxembourg		2020	embourg2020	317206	313207
32	Luxembourg		2021	embourg2021		
32	Luxembourg		2022	embourg2022		
32	Luxembourg		2023	embourg2023		
32	Luxembourg		2024	embourg2024		
32	Luxembourg		2025	embourg2025		
32	Luxembourg		2026	embourg2026		
32	Luxembourg		2027	embourg2027		
32	Luxembourg		2028	embourg2028		
32	Luxembourg		2029	embourg2029		
32	Luxembourg		2030	embourg2030		
32	Luxembourg		2005	embourg2005	230127	235025
32	Luxembourg		2006	embourg2006	233946	238695
32	Luxembourg		2007	embourg2007	237701	242291
32	Luxembourg		2008	embourg2008	242221	246426
32	Luxembourg		2009	embourg2009	247122	250661
32	Luxembourg		2010	embourg2010	252014	254939
32	Luxembourg		2011	embourg2011	258222	260129
32	Luxembourg		2012	embourg2012	265121	265831
32	Luxembourg		2013	embourg2013	271763	271595
32	Luxembourg		2014	embourg2014	278546	277776
32	Luxembourg		2015	embourg2015	285584	284021
32	Luxembourg		2016	embourg2016	292915	290544
32	Luxembourg		2017	embourg2017	299613	296724
32	Luxembourg		2018	embourg2018	305652	302298
32	Luxembourg		2019	embourg2019	311845	308158
32	Luxembourg		2020	embourg2020	317206	313207

Figure 28: Image shows the Duplicate Values

Inaccurate Data: If business data contains false information, the business purpose is defeated. We found the population for a country to be in only four digits which can't be quite right should a business wish to carry out the inaccurate information about a country towards business practices. The project's entire purpose has failed, and the business will not receive the information that they are seeking to finalize a country to make their branch.



Male	Female	Total Populati	R
3729367	3929605	7658972	
3700597	3900426	554552	
3673473	3871865	7545337.5	
3648111	3844450	243533	
3624907	3819536	4256	
3601442	3794157	3443	
3577847	3770481	7348327.5	
3555920	3749968	7305888	

Figure 29: Image shows the Inaccurate Data

Inaccurate Datatypes: While analyzing and visualizing the data, we need our attributes format appropriately. A wrong datatype can generate wrong values or errors while analyzing the data. If a string column value has integer/float as its datatype, the analysis won't calculate values, plus the visualization software will not be displaying proper visualizations.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1430 entries, 0 to 1429
Data columns (total 17 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Countries                             1430 non-null   object
 1   Year                                  1430 non-null   int64
 2   Male                                  864 non-null    float64
 3   Female                                864 non-null    float64
 4   Total Population DB                   864 non-null    float64
 5   Employee % to Total %                 828 non-null    float64
 6   # of people employed in tourism       828 non-null    float64
 7   Tourist Arrival                       816 non-null    object
 8   Total Revenue Gain                    861 non-null    float64
 9   Total Tourist Spending                816 non-null    float64
 10  Literacy                              816 non-null    object
 11  Crime                                 816 non-null    float64
 12  Average annual working hours per worker  816 non-null  float64
 13  1 dollar (USD)  = country currency     816 non-null    object
 14  Salary/ Hr                            816 non-null    object
 15  Upcoming Year                         810 non-null    float64
 16  Projected Population                  810 non-null    float64
dtypes: float64(11), int64(1), object(5)
memory usage: 190.0+ KB
```

Figure 29: Image shows the Inaccurate Datatye

Clean Data

Data cleaning is the most common way of fixing or eliminating incorrect, corrupted, inaccurately designed, duplicated, or inadequate information inside a dataset. While consolidating various data sources, there are numerous open doors for copying or mislabeling data. On the off chance that data is incorrect, results and calculations are inconsistent, although they might look correct. There is no absolute method for endorsing the specific strides in the data cleaning process because the cycles will fluctuate from dataset to dataset.

Now, when we started the process of data cleaning using the following techniques in jupyter python framework:

Dropping Unnecessary Columns: The first look at the raw dataset showed us unnecessary columns that are irrelevant or of any use to the company goals and requirements. So we decided to drop the columns using the function drop().

Dataset Before: By loading the data into the python framework, there were 41 columns. Of those 41 columns, only some were important, and else were occupying the space creating an incredible amount of data anomaly.

Code: To read file: HL = pd.read_excel("FinalDataset.xlsx")

To know info: HL.info()

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from IPython.display import display
from matplotlib import patheffects
import os
import seaborn as sb
import plotly.express as px

HL = pd.read_excel("FinalDataset.xlsx")

HL.info()
```

Figure 30: Loading data set code and knowing information code

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1430 entries, 0 to 1429
Data columns (total 41 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   #                                     1430 non-null   int64
 1   Countries                             1430 non-null   object
 2   Year                                  1430 non-null   int64
 3   Concat                                1430 non-null   object
 4   Male                                  864 non-null    float64
 5   Female                                864 non-null    float64
 6   Total Population DB                   864 non-null    float64
 7   Real Population                       864 non-null    float64
 8   How Off                               864 non-null    float64
 9   PP                                    594 non-null    float64
 10  Employee % to Total %                 828 non-null    float64
 11  # of people employed in tourism       828 non-null    float64
 12  Less than 5 Years                     473 non-null    float64
 13  5-10 Years                            320 non-null    float64
 14  10+ Years                             465 non-null    float64
 15  International Visitors                468 non-null    object
 16  Tourist Arrival                       816 non-null    object
 17  Total Revenue Gain                    861 non-null    float64
 18  Total Tourist Spending                816 non-null    float64
 19  Expenditures (millions)               700 non-null    object
 20  Literacy                              816 non-null    object
 21  Crime                                 816 non-null    float64
 22  Average annual working hours per worker  816 non-null  float64
 23  1 dollar (USD)  = country currency     816 non-null    object
 24  Salary/ Hr                            816 non-null    object
 25  Upcoming Year                         810 non-null    float64
```

Figure 31: Column from 0 to 25 before cleaning

```
26  Projected Country Name                  810 non-null    object
27  Projected Population                    810 non-null    float64
28  Unnamed: 28                             0 non-null      float64
29  Unnamed: 29                             3 non-null      object
30  Unnamed: 30                             9 non-null      object
31  Unnamed: 31                             0 non-null      float64
32  Unnamed: 32                             0 non-null      float64
33  Unnamed: 33                             0 non-null      float64
34  Unnamed: 34                             0 non-null      float64
35  Unnamed: 35                             0 non-null      float64
36  Unnamed: 36                             0 non-null      float64
37  Unnamed: 37                             0 non-null      float64
38  Unnamed: 38                             0 non-null      float64
39  Unnamed: 39                             0 non-null      float64
40  Unnamed: 40                             1 non-null      object
dtypes: float64(27), int64(2), object(12)
memory usage: 458.2+ KB
```

Figure 32: Column from 25 to 40 before cleaning

Dataset After: After dropping the table with the function drop (), we have 16 columns which will be easy for us to analyze data further. We removed 25 columns: real population, concat, how off, pp, less than five years, 5 - 10 years and 10+ years, International Visitors, Expenditures (millions), projected country name, and all unnamed from 28 to 40.

Code used: #Cleaning 1 - Dropping Unnecessary Columns.

HL.drop(['#', 'Concat', 'Real Population', 'How Off', 'International Visitors', '5-10 Years', '10+ Years', 'Unnamed: 28', 'Unnamed: 29', 'Unnamed: 30','Unnamed: 31','Unnamed: 32','Projected Country Name', 'Unnamed: 33', 'Unnamed: 34', 'Unnamed: 35', 'Unnamed: 36', 'Unnamed: 37', 'Unnamed: 38', 'Less than 5 Years', 'Unnamed: 39', 'Unnamed: 40', 'Expenditures (millions)', 'PP'], axis = 1, inplace = True)

display(HL.head())

HL.info()

```
# Cleaning 1 - Dropping Unecessary Columns

HL.drop(['#', 'Concat', 'Real Population', 'How Off', 'International Visitors', '5-10 Years', '10+ Years',
        'Unnamed: 28','Unnamed: 29','Unnamed: 30','Unnamed: 31','Unnamed: 32','Projected Country Name',
         'Unnamed: 33','Unnamed: 34','Unnamed: 35','Unnamed: 36','Unnamed: 37','Unnamed: 38', 'Less than 5 Years',
        'Unnamed: 39','Unnamed: 40','Expenditures (millions)','PP'], axis = 1, inplace = True)

display(HL.head())
HL.info()
```

Figure 32: Drop Column Code

Output

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1430 entries, 0 to 1429
Data columns (total 17 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Countries                             1430 non-null   object
 1   Year                                  1430 non-null   int64
 2   Male                                  864 non-null    float64
 3   Female                                864 non-null    float64
 4   Total Population DB                   864 non-null    float64
 5   Employee % to Total %                 828 non-null    float64
 6   # of people employed in tourism       828 non-null    float64
 7   Tourist Arrival                       816 non-null    object
 8   Total Revenue Gain                    861 non-null    float64
 9   Total Tourist Spending                816 non-null    float64
 10  Literacy                              816 non-null    object
 11  Crime                                 816 non-null    float64
 12  Average annual working hours per worker  816 non-null  float64
 13  1 dollar (USD)  = country currency     816 non-null    object
 14  Salary/ Hr                            816 non-null    object
 15  Upcoming Year                         810 non-null    float64
 16  Projected Population                  810 non-null    float64
dtypes: float64(11), int64(1), object(5)
memory usage: 190.0+ KB
```

Figure 33: Output of the code

Dropping/Filling Missing Null values: As we identified missing and null values, we had to drop some of the rows or do research. First, we tried to fill the missing values with the mean of the columns, but due to excessive missing/null values, the analysis would not have been accurate. Therefore, we found the important ones by researching on the internet to fill, and the ones we weren't able to find were dropped. As a result, we dropped some countries such as China, Cyprus, Croatia, Malta, and World, and for years we dropped columns from 2021 to 2030.

Before: With all the values we had, there were total rows of 1430, from which 630 rows were occupying unnecessary space in the dataset. We used drop () function with loc () function to identify the specific value.

Code: # Dropping Rows with Specific Value = "World."

```
HL.drop(HL.loc[HL['Countries']=='World'].index, inplace=True)

# Dropping Rows with Specific Value = "2021 to 2030"
HL.drop(HL.loc[HL['Year'] == 2021].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2022].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2023].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2024].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2025].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2026].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2027].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2028].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2029].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2030].index, inplace = True)
```

Dropping Rows with Specific Value = "China", "Croatia", "Cyprus", "Malta".

HL.drop(HL.loc[HL['Countries']=='China'].index, inplace=True)

HL.drop(HL.loc[HL['Countries']=='Croatia'].index, inplace=True)

HL.drop(HL.loc[HL['Countries']=='Cyprus'].index, inplace=True)

HL.drop(HL.loc[HL['Countries']=='Malta'].index, inplace=True)

display(HL)

```
# Dropping Rows with Specific Value = "World"
HL.drop(HL.loc[HL['Countries']=='World'].index, inplace=True)

# Dropping Rows with Specific Value = "2021 to 2030"
HL.drop(HL.loc[HL['Year'] == 2021].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2022].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2023].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2024].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2025].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2026].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2027].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2028].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2029].index, inplace = True)
HL.drop(HL.loc[HL['Year'] == 2030].index, inplace = True)

# Dropping Rows with Specific Value = "China", "Croatia", "Cyprus", "Malta".
HL.drop(HL.loc[HL['Countries']=='China'].index, inplace=True)
HL.drop(HL.loc[HL['Countries']=='Croatia'].index, inplace=True)
HL.drop(HL.loc[HL['Countries']=='Cyprus'].index, inplace=True)
HL.drop(HL.loc[HL['Countries']=='Malta'].index, inplace=True)
```

Figure 34: Dropping specific rows

After: After dropping the values, we had a row count of 800, which still leads to doubt there is still a data problem in the dataset. We have put 16 values for each country/ year, ranging from 2005 to 2020.

So there should be 16 multiplied by 48 countries (due to dropping some countries), that is 768. So this leads to finding duplication in the dataset.

Code for display:

display(HL.head()) #displaying dataset

HL.isnull().sum() #Checking Null values

```
display(HL.head())
HL.isnull().sum()
```

Figure 35: To display the dataset wthe ith top 5 rcheckchecking if their null value or not

Output:

ountries	Year	Male	Female	Country Population	Tourism Employment (%)	Employed Tourist	Tourist_Arrival	Domestic Revenue (Billions)	International Revenue (Millions)	Literacy	Crime Rate	Annual Avg WorkHrs/Worker	CountryCur = 1$ (USD)	Salary/ Hr	Upcoming_Year	Proje Popul
rgentina	2005	18951245.0	19782161.0	38733406.0	0.0977	3.784254e+06	3823000	179.543	17770.0	High	5.65	1761.399	112.344 ARS	110 Ars	2021.0	45605
rgentina	2006	19141314.0	19985925.0	39127239.0	0.0965	3.775779e+06	4173000	142.906	20235.0	High	5.37	1765.601	112.344 ARS	110 Ars	2022.0	46010
rgentina	2007	19335197.0	20194243.0	39529440.0	0.0993	3.925273e+06	4562000	236.516	25385.0	High	5.39	1780.557	112.344 ARS	110 Ars	2023.0	46409
rgentina	2008	19532999.0	20407235.0	39940234.0	0.1000	3.994023e+06	4700000	317.578	29935.0	High	5.92	1781.425	112.344 ARS	110 Ars	2024.0	46803
rgentina	2009	19734800.0	20625028.0	40359828.0	0.1017	4.104595e+06	4308000	360.951	29040.0	High	6.53	1742.904	112.344 ARS	110 Ars	2025.0	47192

Figure 35: Display the dataset

```
[11]: Countries                                0
      Year                                     0
      Male                                     0
      Female                                   0
      Country Population                       0
      Tourism Employment (%)                   0
      Employed Tourist                         0
      Tourist_Arrival                          0
      Domestic Revenue (Billions)              0
      International Revenue (Millions)         0
      Literacy                                 0
      Crime Rate                               0
      Annual Avg WorkHrs/Worker                0
      CountryCur = 1$ (USD)                    0
      Salary/ Hr                               0
      Upcoming_Year                            0
      Projected Population                     0
      dtype: int64
```

Figure 36: Checking null value count in each column

Identifying Duplicate Values:

We found duplicate values using a code that searches for duplicate values in a dataset. So for us to identify, we tried to count each country in the dataset as it can be a unique identifier to locate duplicate values. This process would help us locate and drop the row immediately.

Before: As we analyzed before, we had 800 rows in the dataset instead of 768 because every country has 16 values from 2005 to 2020. So now, we used the count code first to identify count and duplicates in each column.

Code:

```
# Count of Countries in Dataset
CO = HL['Countries'].groupby(HL['Countries'])
SV = CO.count()
print('Total Number of Countries:', SV)

#Checking Duplicates
duplicate = HL[HL.duplicated()]
duplicate.count()
```

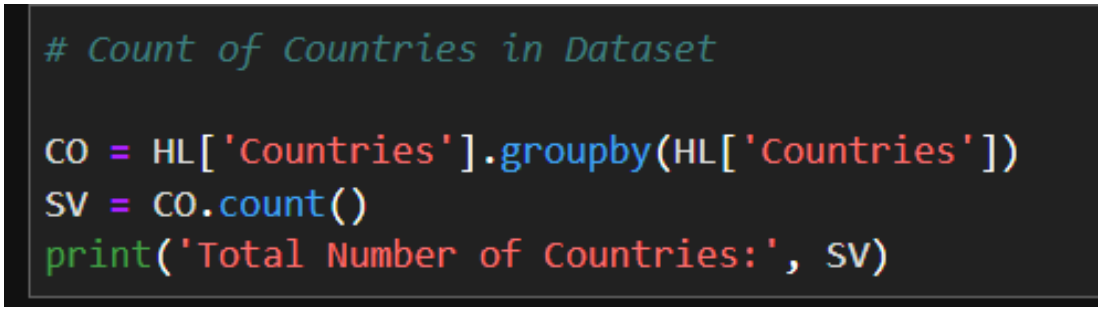


Figure 37: Checking Country Count

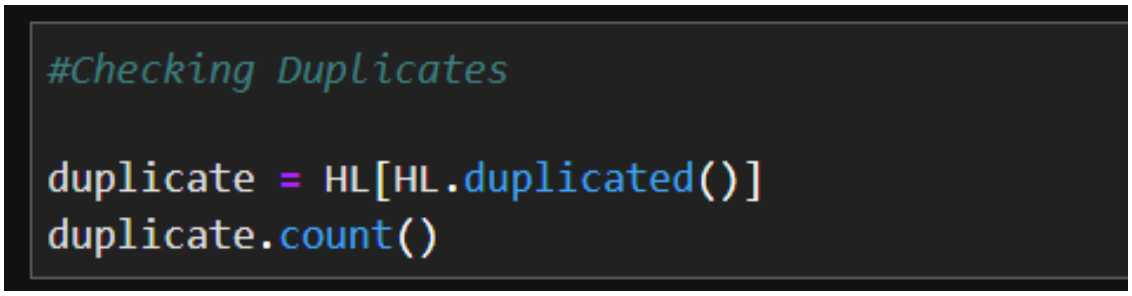


Figure 38: Checking Duplicates Count in Each column

Output Before Dropping Duplicates: Luxembourg was identified as a duplicate and the count of 32 duplicate values in each column

```
Korea                16
Latvia               16
Lithuania            16
Luxembourg           48
Mexico               16
Netherlands          16
New Zealand          16
Norway               16
Poland               16
```

Figure 39: Duplicates Count in Each column

```
Countries                           32
Year                                32
Male                                32
Female                              32
Country Population                  32
Tourism Employment (%)              32
Employed Tourist                    32
Tourist_Arrival                     32
Domestic Revenue (Billions)         32
International Revenue (Millions)    32
Literacy                            32
Crime Rate                          32
Annual Avg WorkHrs/Worker           32
CountryCur = 1$ (USD)               32
Salary/ Hr                          32
Upcoming_Year                       32
Projected Population                32
dtype: int64
```

Figure 40: Duplicates Count in Each column

After: As the dropping of duplicate values is completed, we have 768 rows and duplicate count as 0. Having data with no missing/null values and duplicates is a significant relief, but now the only thing left is to resolve the inaccurate data.

Code:

```
# Removing Duplicates
HL.drop_duplicates(keep = "last",inplace=True)
display(HL)
DP = HL[HL.duplicated()]
DP.count()
```

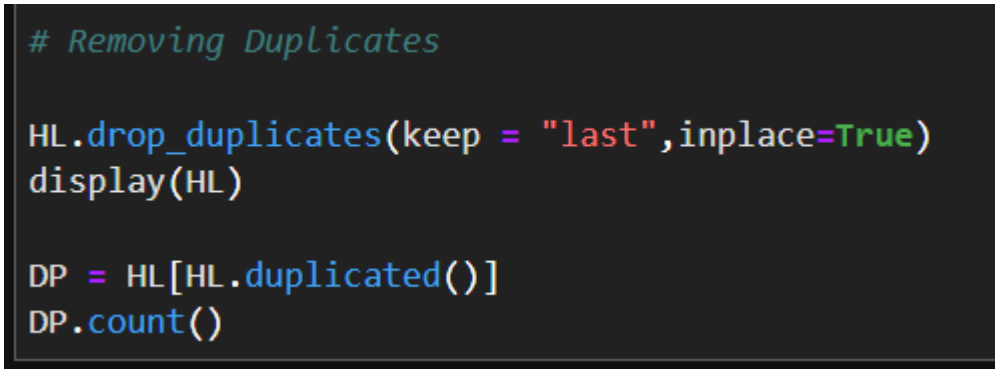


Figure 41: Dropping Duplicate values and Checking Count

Output:

```
768 rows × 17 columns
```

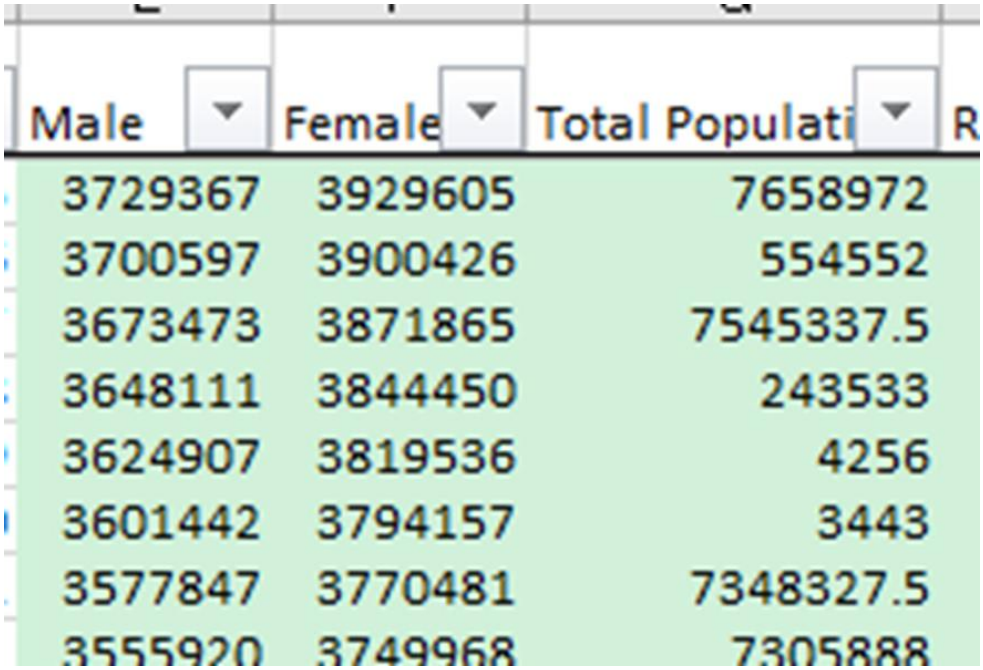
Countries	0
Year	0
Male	0
Female	0
Country Population	0
Tourism Employment (%)	0
Employed Tourist	0
Tourist_Arrival	0
Domestic Revenue (Billions)	0
International Revenue (Millions)	0
Literacy	0
Crime Rate	0
Annual Avg WorkHrs/Worker	0
CountryCur = 1$ (USD)	0
Salary/ Hr	0
Upcoming_Year	0
Projected Population	0

dtype: int64

Figure 42: Duplicate values Count

Filling Inaccurate Data: By resolving other anomalies working with an efficient amount of data is always accessible. Now, still, there are problems with the efficient amount of data that is inaccurate information. To resolve this, we searched for the accurate date on the internet to replace it with an inaccurate one.

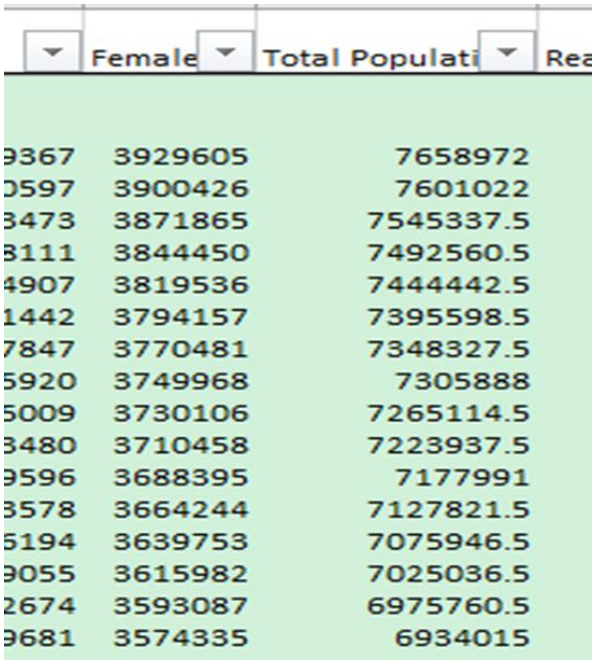
Before Cleaning:



Male	Female	Total Populati	R
3729367	3929605	7658972	
3700597	3900426	554552	
3673473	3871865	7545337.5	
3648111	3844450	243533	
3624907	3819536	4256	
3601442	3794157	3443	
3577847	3770481	7348327.5	
3555920	3749968	7305888	

Figure 42: Inaccurate Values for example 4256

After Cleaning:



	Female	Total Populati	Rea
9367	3929605	7658972	
0597	3900426	7601022	
3473	3871865	7545337.5	
3111	3844450	7492560.5	
4907	3819536	7444442.5	
1442	3794157	7395598.5	
7847	3770481	7348327.5	
5920	3749968	7305888	
5009	3730106	7265114.5	
3480	3710458	7223937.5	
9596	3688395	7177991	
3578	3664244	7127821.5	
5194	3639753	7075946.5	
9055	3615982	7025036.5	
2674	3593087	6975760.5	
9681	3574335	6934015	

Figure 43: Accurate Values by researching on google

Split Columns: We wanted to have a Salary per hour and country currency conversion in an integer format, but we couldn't as it had data type as an object because its value shows "7.98 ARS". We want to split them into Sal/hr and currency names. Therefore, having the proper value creates a calculated column further in the process.

Before: We can see the value for columns CountryCur and Sal/Hr as a mixed value which will solve a problem in the further analysis if not split. We will be splitting them by using a separator.

International Revenue (Millions)	Literacy	Crime Rate	Annual Avg WorkHrs/Worker	CountryCur = 1$ (USD)	Salary/ Hr	Upcoming_
17770.0	High	5.65	1761.399	112.344 ARS	110 Ars	202
20235.0	High	5.37	1765.601	112.344 ARS	110 Ars	202
25385.0	High	5.39	1780.557	112.344 ARS	110 Ars	202
29935.0	High	5.92	1781.425	112.344 ARS	110 Ars	202
29040.0	High	6.53	1742.904	112.344 ARS	110 Ars	202

Figure 44: Displaying the columns before splitting.

After: Using the separator as " " (blank space), we split the column into two having names as for first column sal/hr containing integer value and currency name containing string value. Therefore, as a result, we will be having new columns and the old column we just drop.

Code:

```
HFDS[['CoCur_1DCon', 'Currency Name']] =HFDS['CountryCur = 1$ (USD)'].str.split(' ',  1, expand=True)
HFDS[['Sal/Hr', 'B']] =HFDS['Salary/ Hr'].str.split(' ',  1, expand=True)
display(HFDS)
```

Output:

Projected Population	Total Revenue (Billions)	International Revenue (Billions)	CoCur_1DCon	Currency Name	Sal/Hr
45605826.0	196.770	17.770	112.344	ARS	110
46010234.0	162.235	20.235	112.344	ARS	110
46409168.0	261.385	25.385	112.344	ARS	110
46803049.0	346.935	29.935	112.344	ARS	110
47192100.0	389.040	29.040	112.344	ARS	110

Figure 45: Split Columns

Calculated Fields: We created a calculated field for analyzing country salaries according to the working hours per month and annually. In addition, we created total revenue earned by adding to columns international revenue and domestic revenue. We had to convert international revenue into millions in billion formats to add them correctly.

Before: All columns that sal/hr, average working hours, and international and domestic revenue are proper to calculate the new attributes.

Employed Tourist	Tourist_Arrival	Domestic Revenue (Billions)	International Revenue (Millions)	Literacy	Crime Rate	Wor
784254e+06	3823000.0	179.543	17770.0	High	5.65	
775779e+06	4173000.0	142.906	20235.0	High	5.37	
925273e+06	4562000.0	236.516	25385.0	High	5.39	
994023e+06	4700000.0	317.578	29935.0	High	5.92	
104595e+06	4308000.0	360.951	29040.0	High	6.53	

Figure 46: Before Calculated Columns.

After: After changing international revenue to billions, we added a new column named total revenue and added other calculated fields through tableau.

Code:

```
l1 = len(HL["Domestic Revenue (Billions)"])
l2 = len(HL["International Revenue (Millions)"])
l1 == l2
DR = (HL["Domestic Revenue (Billions)"])
IR = (HL["International Revenue (Millions)"] * 0.001)
HL['Total Revenue (Billions)'] = DR + IR
HL["International Revenue (Billions)"] =  (HL["International Revenue (Millions)"] * 0.001)
HL["Tourism Employment (%)"] = (HL["Tourism Employment (%)"] * 100)
display(HL)
```

```
l1 = len(HL["Domestic Revenue (Billions)"])
l2 = len(HL["International Revenue (Millions)"])
l1 == l2

DR = (HL["Domestic Revenue (Billions)"])
IR = (HL["International Revenue (Millions)"] * 0.001)

HL['Total Revenue (Billions)'] = DR + IR
HL["International Revenue (Billions)"] =  (HL["International Revenue (Millions)"] * 0.001)
HL["Tourism Employment (%)"] = (HL["Tourism Employment (%)"] * 100)
display(HL)
```

Figure 47: Calculated Field Code

Python Calculated Field:

rojected pulation	Total Revenue (Billions)	International Revenue (Billions)	Co
605826.0	196.770	17.770	
010234.0	162.235	20.235	
409168.0	261.385	25.385	
803049.0	346.935	29.935	
192100.0	389.040	29.040	

Figure 48: International revenue in Billions

Tableau Calculated field:

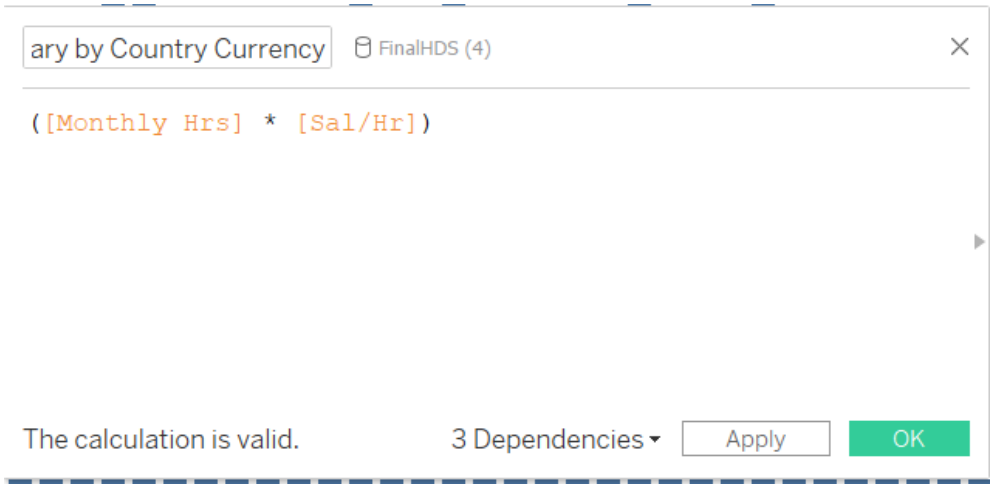


Figure 49: Monthly salary for the country in its currency

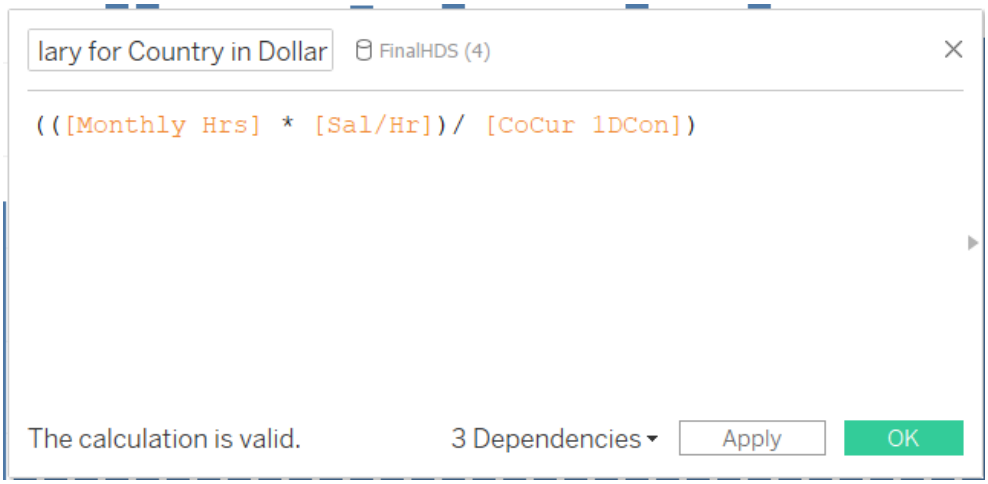


Figure 50: Monthly salary for the country in USD

Figure 51: Annual Salary for the country in its currency

Figure 52: Annual Salary for the country in USD

Figure 53: Monthly Hours

Modeling

Data visualization is a graphical representation of information and data. By using visual elements such as charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in your data. In the world of big data, data visualization tools and technologies are essential to analyze vast amounts of information and make informed decisions.

Once the dataset is cleaned, it is now possible to start loading the dataset into a data visualization software and model the data. The software's used was Microsoft Excel, Power BI, and Tableau to create visually appealing graphs and charts that can help discover insights not possible by simply viewing the dataset. With these charts, our goal is to pinpoint any insight that can help determine countries that would be best suited for Hilton Hotels to consider expanding to. Based on the visuals below, the team was able to produce descriptive models, dependency analysis models, and a predictive model to better understand the data variables to select the best country within the dataset.

Modeling Techniques

- **Descriptive Models**
    - Tree Maps

- Dual Axis Line Graph
- Histogram
- Pie Chart
- World Map
- 3D Map

Visualization represents the domestic revenue earned per country in the year 2020

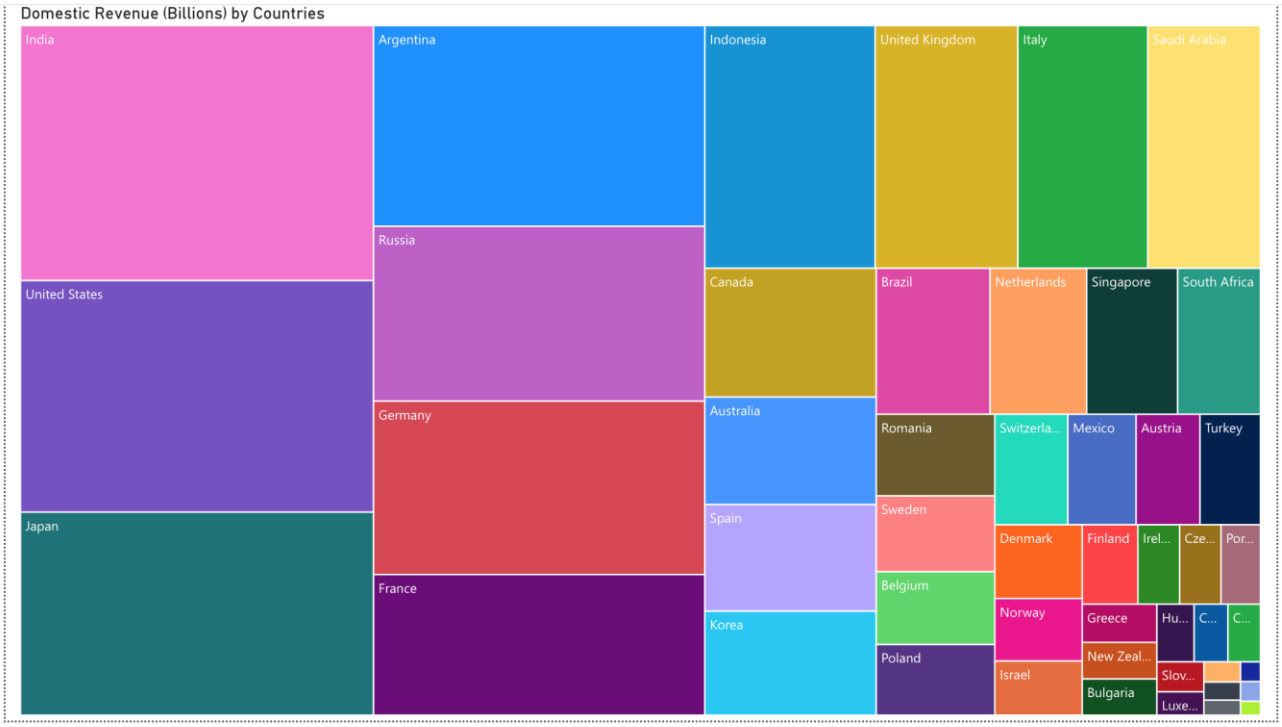


Figure 54: Tree map shows the domestic revenue by countries

This Tree Map shows the domestic revenue earned per country in the year 2020. The purpose for the graph is to easily identify the countries that bring in the highest amount of annual domestic revenue per year for the Tourism industry. The higher the amount, the more appealing the country is for Hilton to expand to. Based on the visual, it was determine that the top country was India followed by the United States.

Visualization represents the tree map for the international revenue earned per country in the year 2020

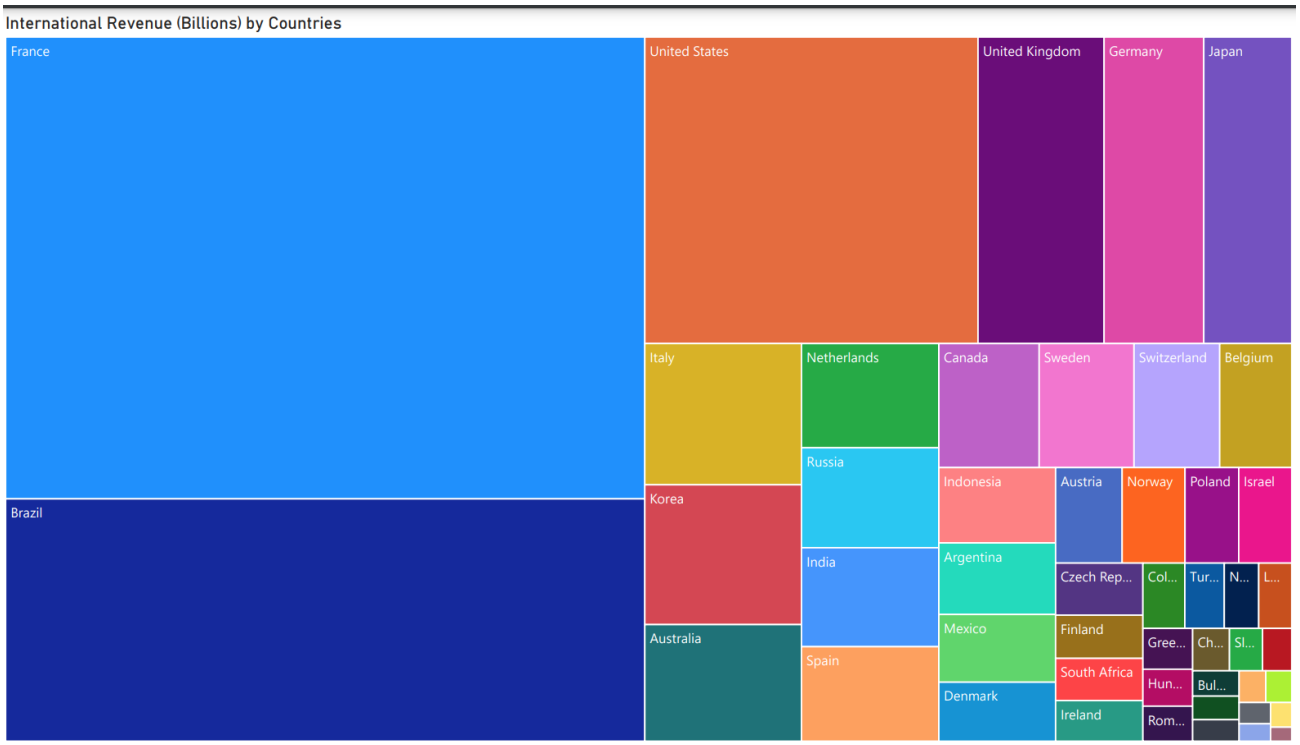


Figure 55: Tree map shows the international revenue by countries

This Tree Map shows the international revenue earned per country in the year 2020. The purpose for the graph is to easily identify the countries that bring in the highest amount of annual international revenue per year for the Tourism industry. The difference between the tree maps is that tourist from out of country should also be considered when selecting a country for Hilton to expand to. Based on the visual, it was determine that the top country was France followed by the Brazil.

Visualization represents the dual axis line graph for the domestic revenue and international revenue earned per country in the year 2020

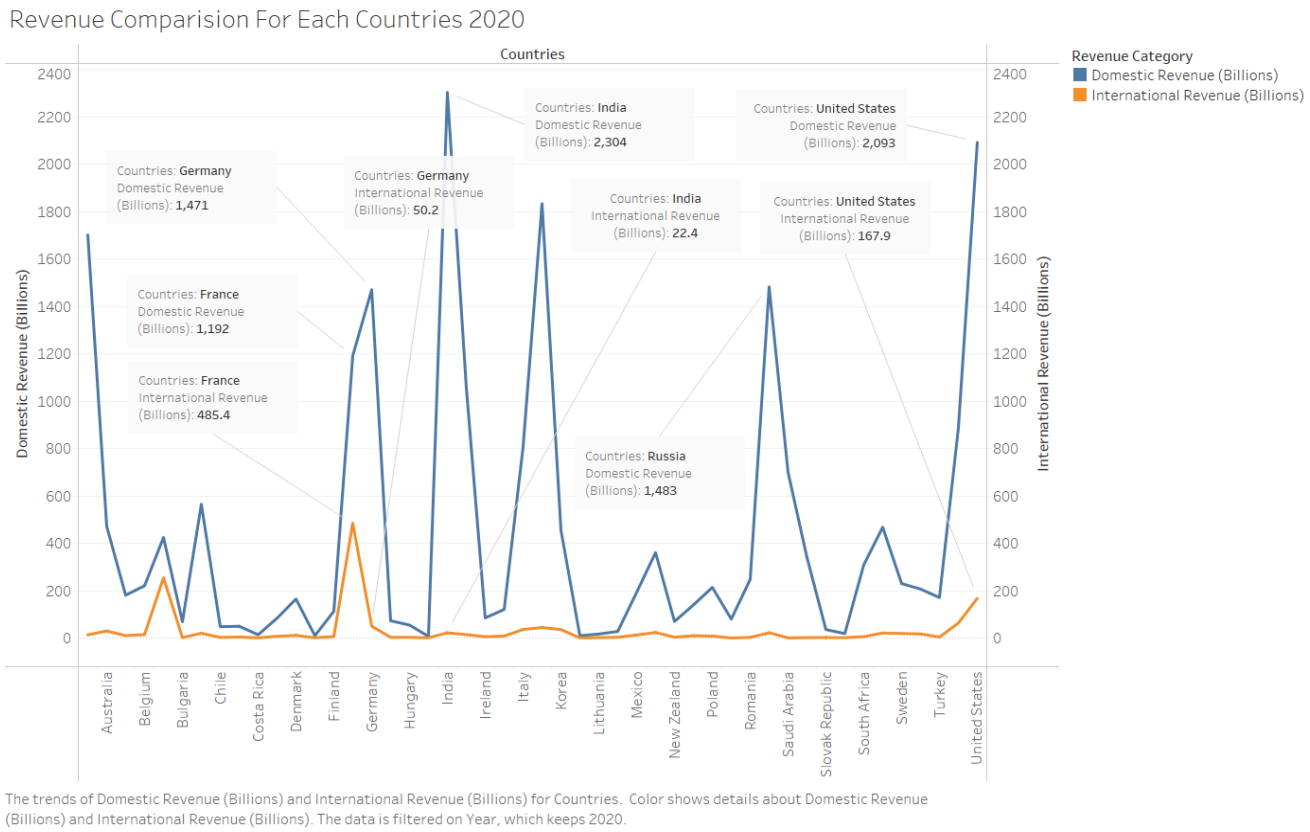


Figure 56: Dual Axis Line Graph shows both the domestic revenue (blue line) and international revenue (orange line) earned per country in the year 2020

This Dual Axis Line Graph shows both the domestic revenue (blue line) and international revenue (orange line) earned per country in the year 2020 in more detail. The purpose for the graph is to compare side by side, the revenue the countries bring in per year for the Tourism industry. Based on the visual, it can be observed that countries such as Russia, Argentina, Japan, and the United States, are among the highest domestic revenue coming in with India having the highest. In comparison, it can be observed that Germany, the United Kingdom, the United States, and Brazil are among the highest international revenue with France having the highest.

Visualization represents the histogram graph for Annual Average Working Hours by Country in the year 2020

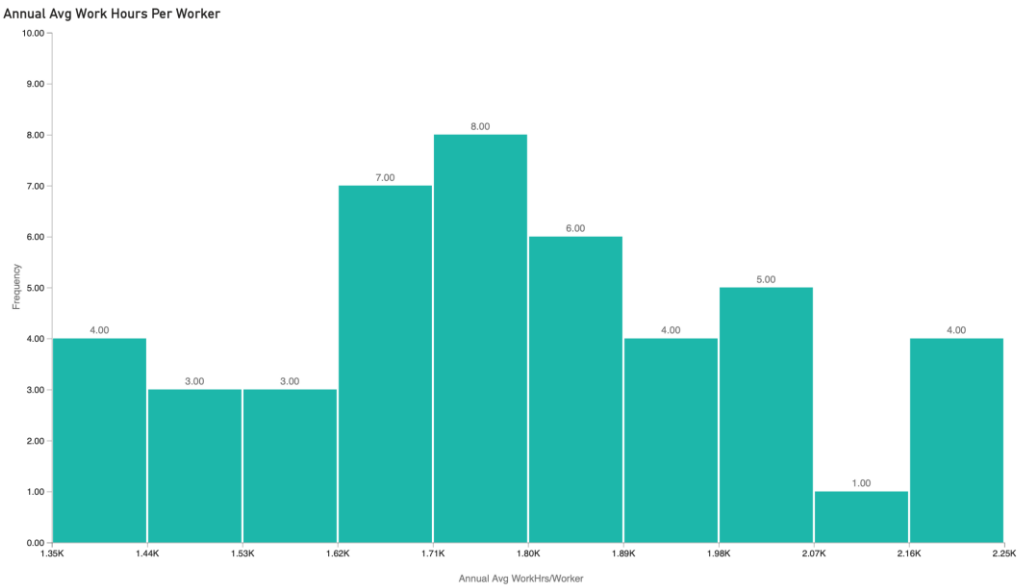


Figure 57: Histogram shows the annual average working hours per country in the year 2020

This Histogram shows the annual average working hours per country in the year 2020. The purpose for the graph is to display the frequency distribution between the countries. It can be observed that the range of hours worked per country varies significantly (about 1,000 hours). It can also be observed that there are four countries that have very high number of working hours. The higher the amount, the more appealing the country is for Hilton since Hilton would be able to utilize the hardworking labor force for their hotel chains. Based on the visual, it was determine that countries had enough variation that Hilton should consider going after the countries with higher working hours.

Visualization represents the pie chart for Annual Average Working Hours by Country

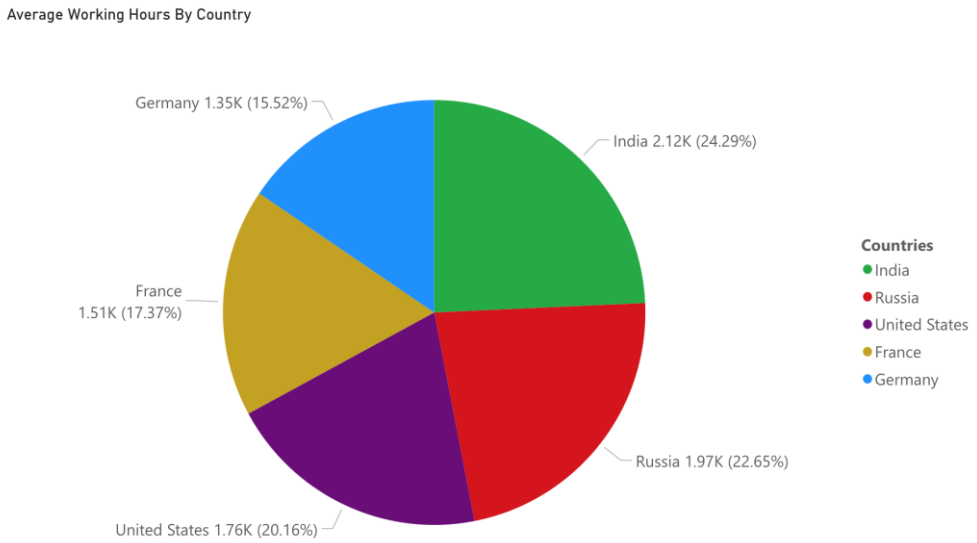


Figure 58: Pie Chart shows five different countries and comparing their annual average working hours

This Pie Chart shows five different countries and comparing their annual average working hours. As mentioned in the earlier graph, the higher the working hours, the more likely that Hilton will not have a labor shortage as workers will be use to the long hours familiar to a hotel that is open 24 hours. The countries selected were India, Russia, the United States, France, and Germany as they were likely candidates for our selection process. Based on the visual, it was determine that the top country was India who had an average higher than two thousand hours compared to a country like Germany where the people are used to only working on average about 1,350 hours per year.

Visualization represents the World Map that shows Country Details – Crime Rate, Literacy Rate, Population Size, Total Revenue

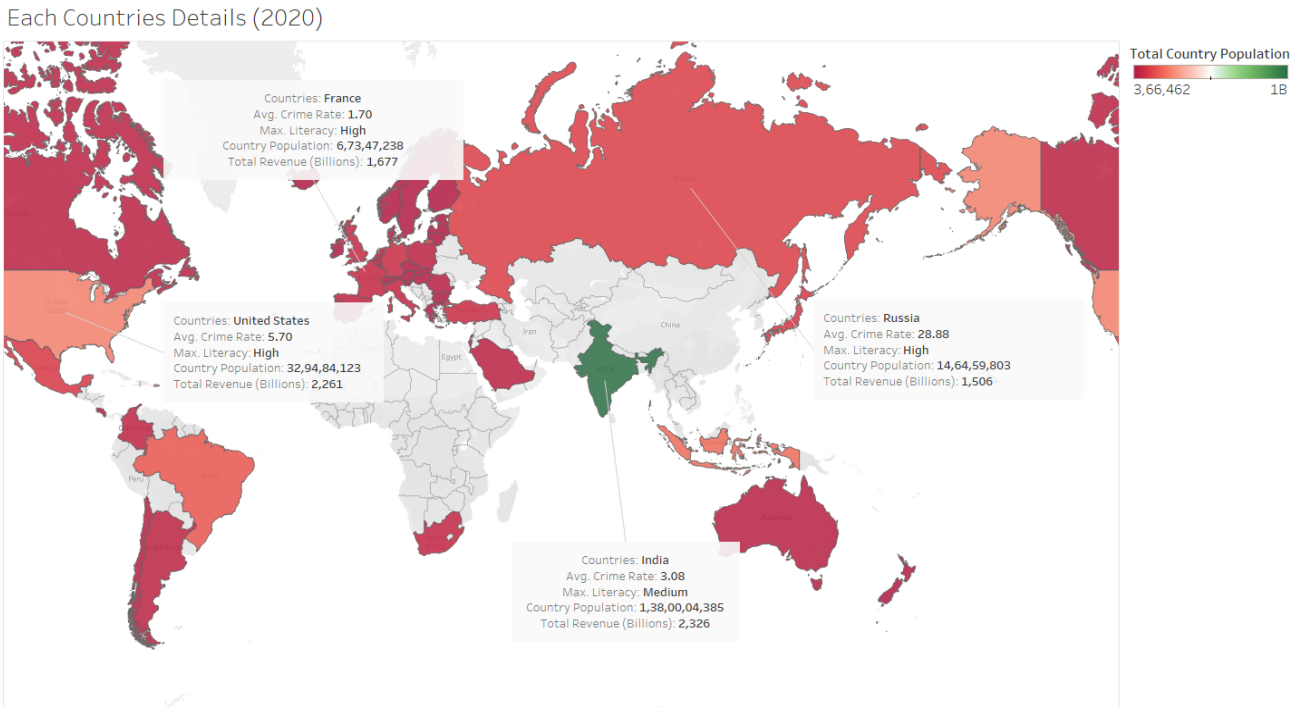


Figure 59: the World Map that shows Country Details – Crime Rate, Literacy Rate, Population Size, Total Revenue

This World Map shows numerous descriptive data about certain countries so that we may get a quick overview of some of the variables of the country. The purpose for the graph is to demonstrate these several variables in a concise view displaying the countries; population size, crime rate, literacy rate, and the total annual revenue. Based on the visual, it was be observed that India has the highest population, highest total revenue in the tourism industry as well as a medium literacy rate and a lower crime rate (compared to the United States and Russia but not France).

Visualization represents the 3D Maps that shows Population Size and Female/Male Ratio

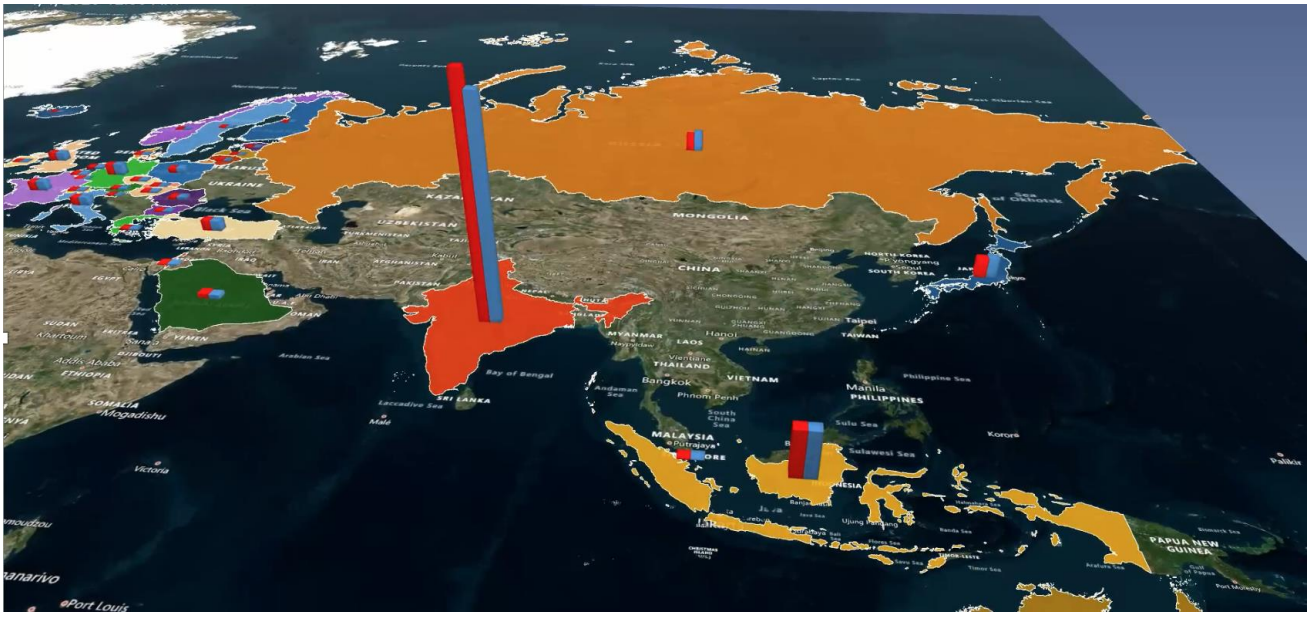


Figure 60: the 3D Maps that shows Population Size and Female/Male Ratio

This 3D Map shows a glimpse of the countries that were available within the dataset by color. As can be seen, most of the countries available were from Europe but there were a few countries available around the world. In addition, the double bars indicate the female(blue) and male(red) population and the height of the bars indicate the population size. The purpose for the graph is to identify which countries were available within the dataset, how different the female and male population were, and the size of the population. Based on the visual, it was determine that the country with the highest population was India and that it had a higher male population.

- **Dependency Analysis**
    - Scatter Plot(s)with Trend Line
    - Bubble Graph
    - Ranking Graph
    - World Map
    - Multi Variable Bar Graph

Visualization represents the Scatter Plot shows the correlation between international revenue and the number of tourist arrivals by country

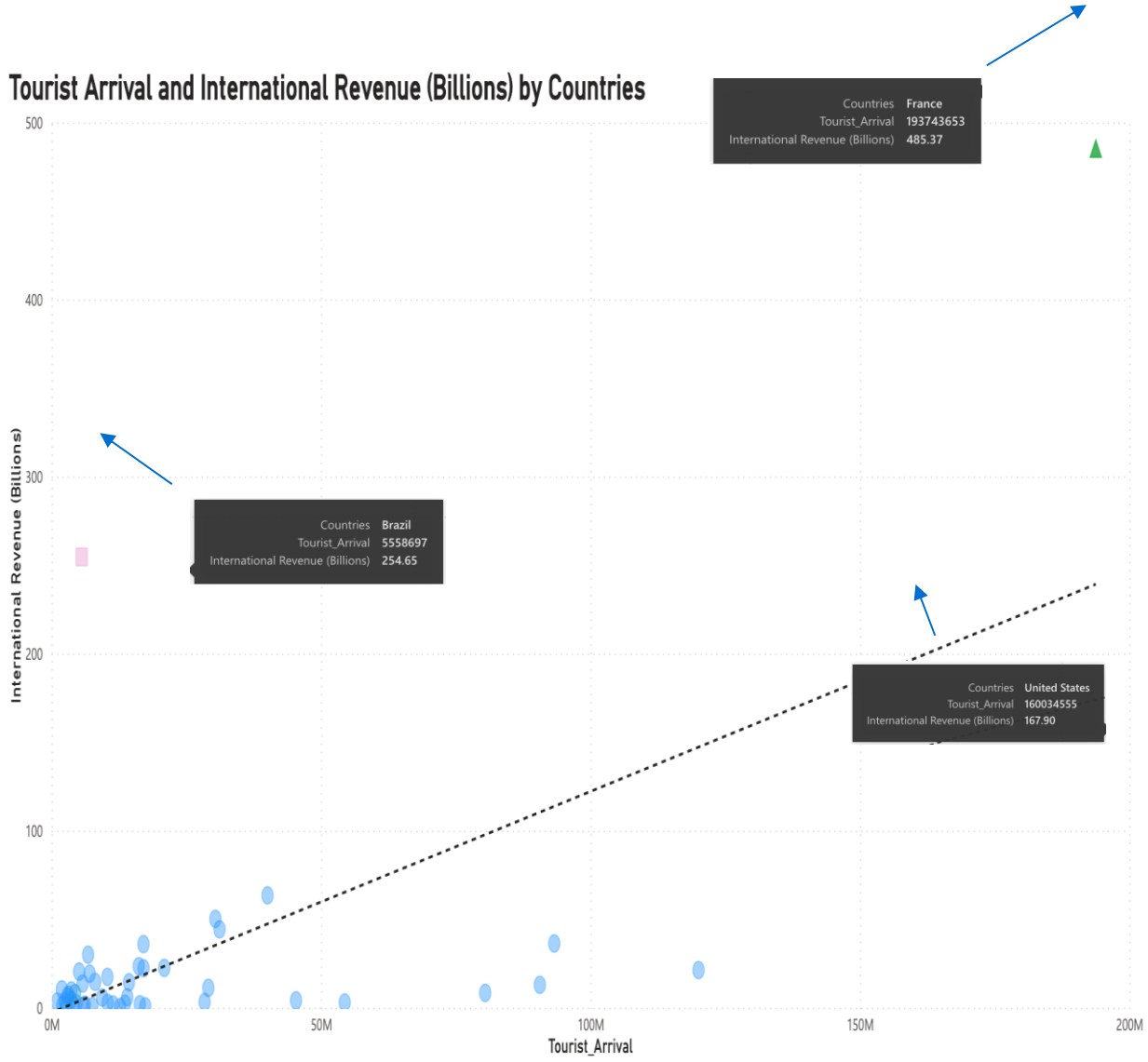


Figure 61: Scatter Plot shows the correlation between international revenue and the number of tourist arrivals by country

This Scatter Plot shows the correlation between international revenue and the number of tourist arrivals by country. The purpose for the graph is to determine if the number of tourists affect revenue. It can be observed from the graph that there are a few countries that are outliers and are highlighted

above. One outlier is Brazil, who has very low tourist arrivals but is second in international revenue indicating that tourist in general spend more in Brazil than other countries. France, the most visited country in the world, has the highest number of visitors and international revenue. Also included is a trend line that shows the overall direction of the data. Based on the trend line, it can be observed that there is indeed a correlation between the amount of tourist arrivals and international revenue gained from them.

Visualization represents the Bubble Chart shows the percentage of people employed in the tourism industry and the total number of people that work in tourism by country in the year 2020

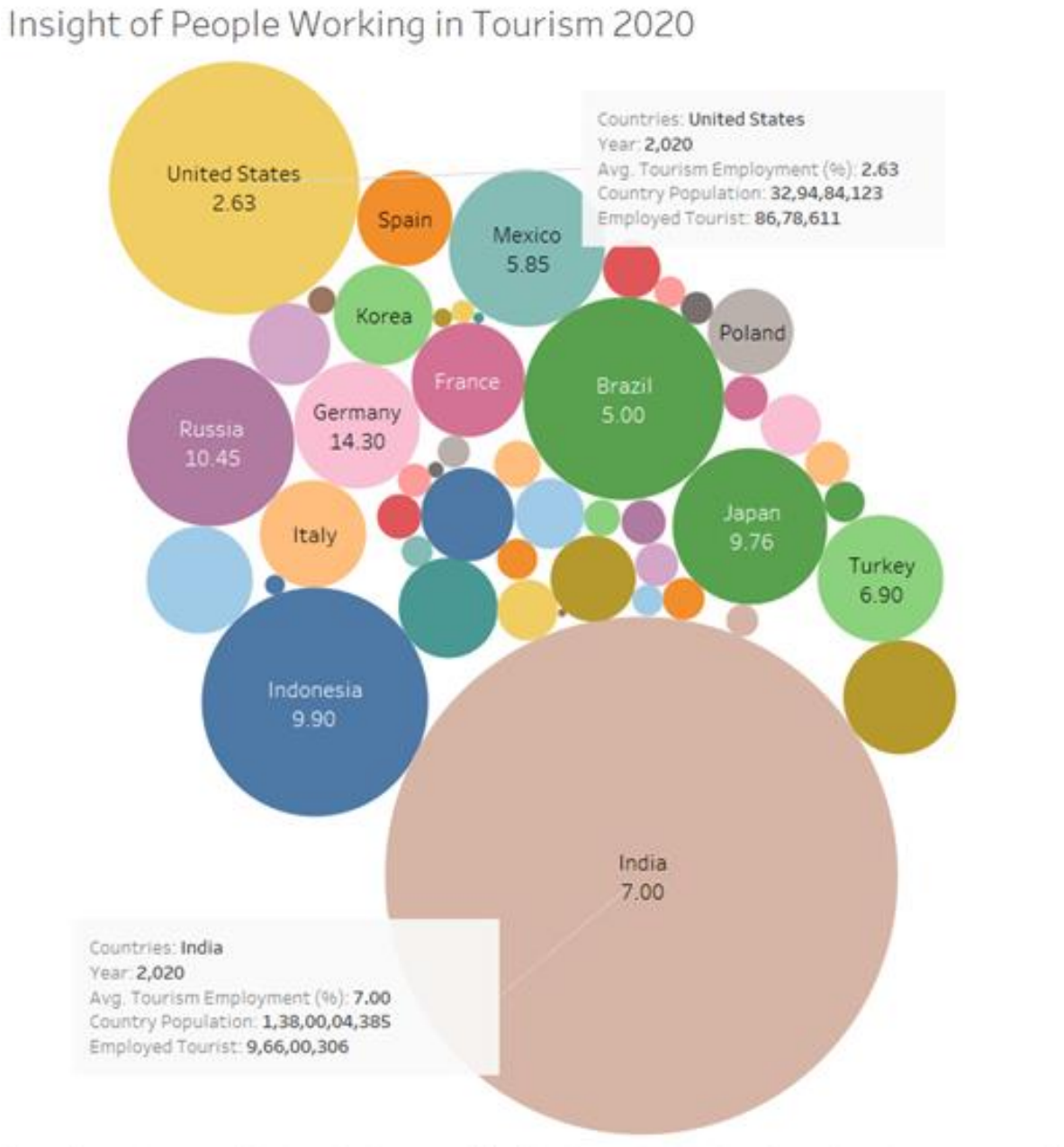


Figure 62: Bubble chart shows the percentage of people employed in the tourism industry and the total number of people that work in tourism by country in the year 2020.

This Bubble chart shows the percentage of people employed in the tourism industry and the total number of people that work in tourism by country in the year 2020. The purpose for the graph is

determine if there is a correlation between the size of the country and the percentage of people working in the tourism industry. Based on the visual ,it can be observed that despite its large size, Germany has the highest amount of people employed in tourism by percentage (14.3 %) than India (7%).

Visualization represents the Ranking Graph that shows monthly salary by monthly hours per country in the year 2020

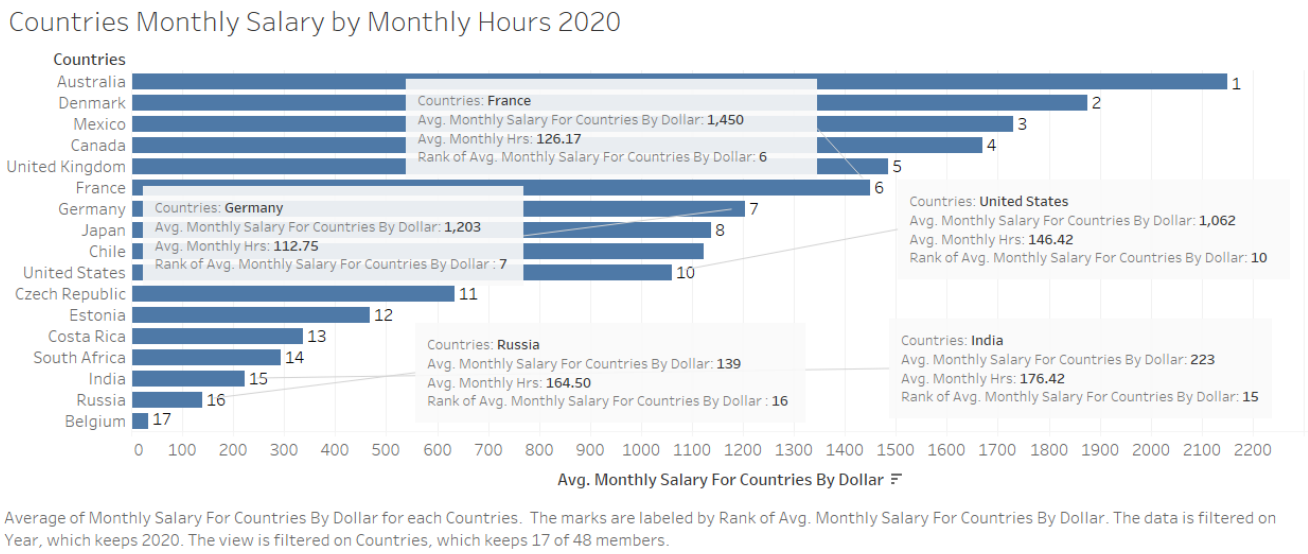


Figure 63: Visualization represents the Ranking Graph that shows monthly salary by monthly hours per country in the year 2020

This Ranking Graph shows monthly salary by monthly hours per country in the year 2020. The purpose for the graph is to determine if there is a correlation between the salary and the hours worked. It can be observed that there is no correlation since it can be seen that USA works longer hours than France but is paid less. Based on the visual, it was determine that the top country was Australia who had the best pay for the hours they work.

Visualization represents the World Map shows the literacy rate and average salary per country in the year 2020

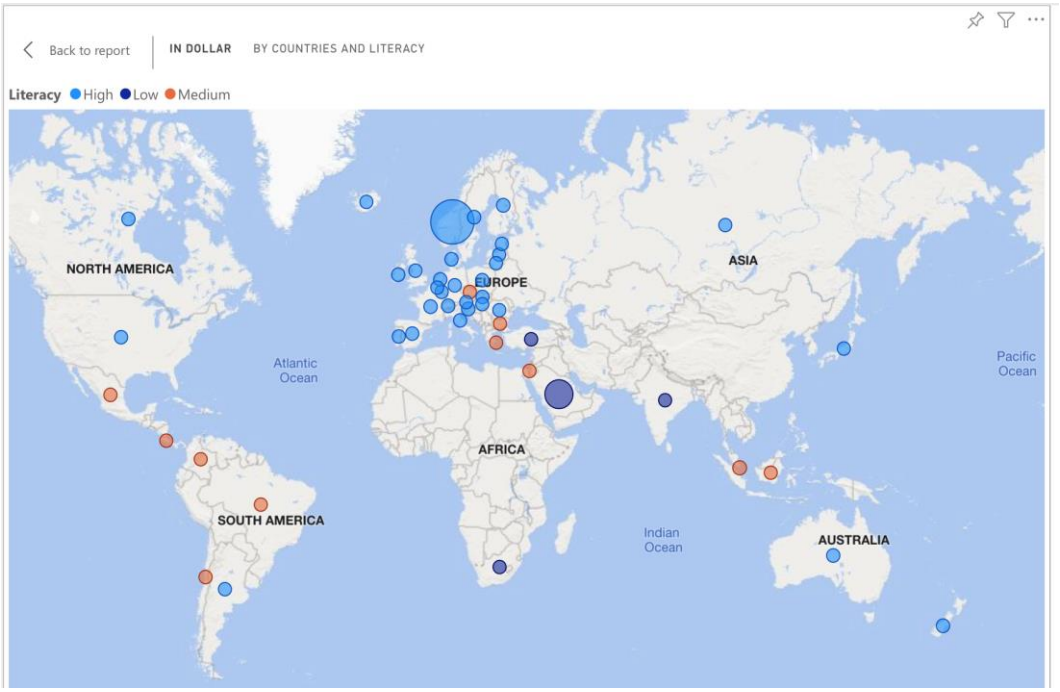


Figure 64: represents the World Map shows the literacy rate and average salary per country in the year 2020

This World Map shows the literacy rate and average salary per country in the year 2020. The purpose for the graph is to spot literacy rates and if they correlate with higher salaries in USD. Again, the higher the literacy rate, the more appealing the country is for Hilton to expand to. Based on the visual, it was determine that the higher paid countries are located in Europe which is also where they have very high literacy rate. India is a surprise as it has a medium literacy rate but has a high average salary.

Visualization represents the Scatter Plot that shows the correlation between monthly salary in country currency and USD for each country in the year 2020

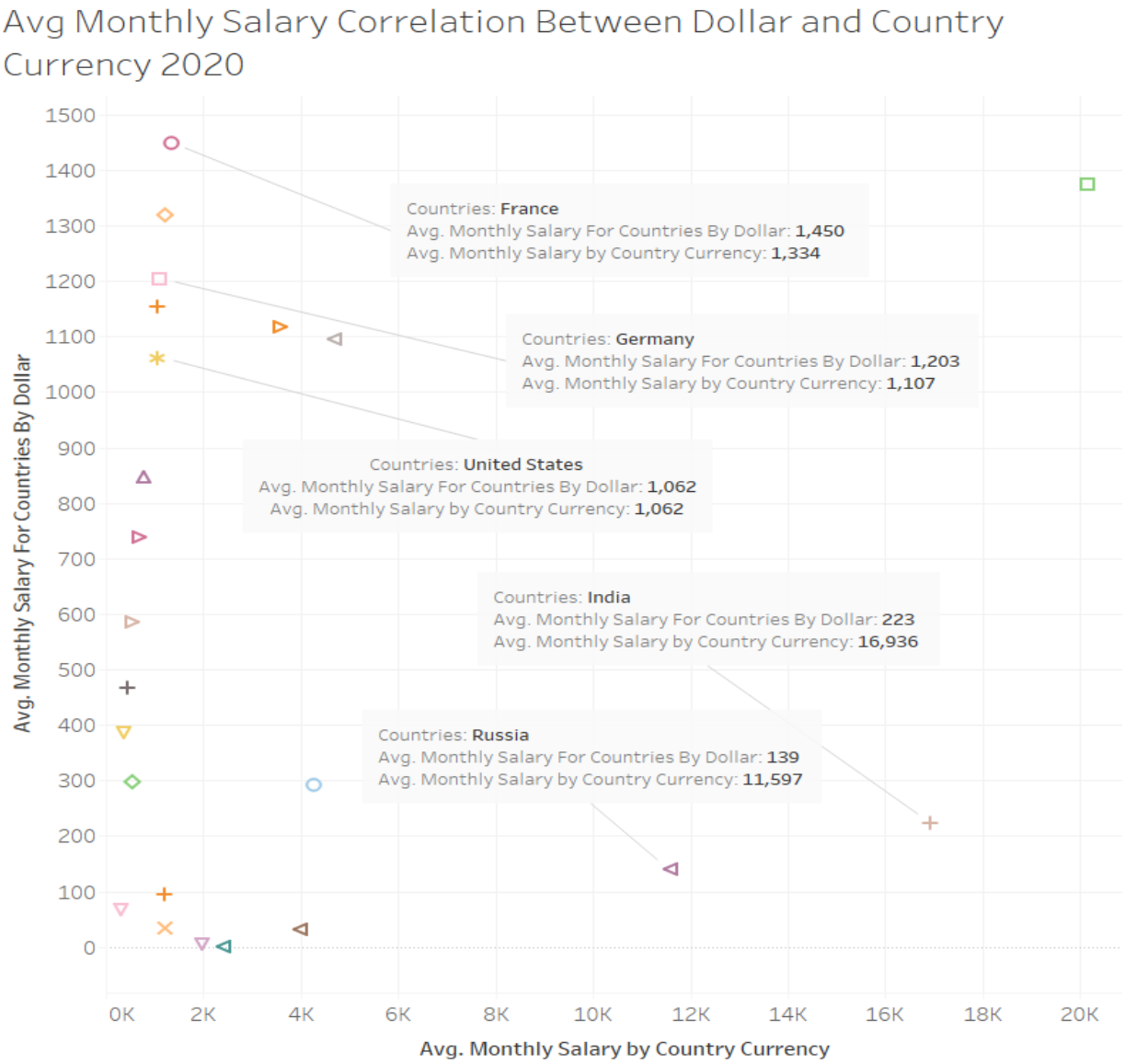


Figure 65: Scatter Plot that shows the correlation between monthly salary in country currency and USD for each country in the year 2020

This Scatter Plot shows the correlation between monthly salary in country currency and USD for each country in the year 2020. The purpose for the graph is to see which countries pay higher on average to their employees. This information is useful to Hilton since they would increase their revenue by finding

countries that on average pay lower in USD. Based on the visual, it was determine that the highest paid workers by month is France, with United States, India and Russia having lower monthly salary in USD.

Visualization that represents the Multi Variable Bar Graph shows the literacy rate, and crime rate per country in the year 2020

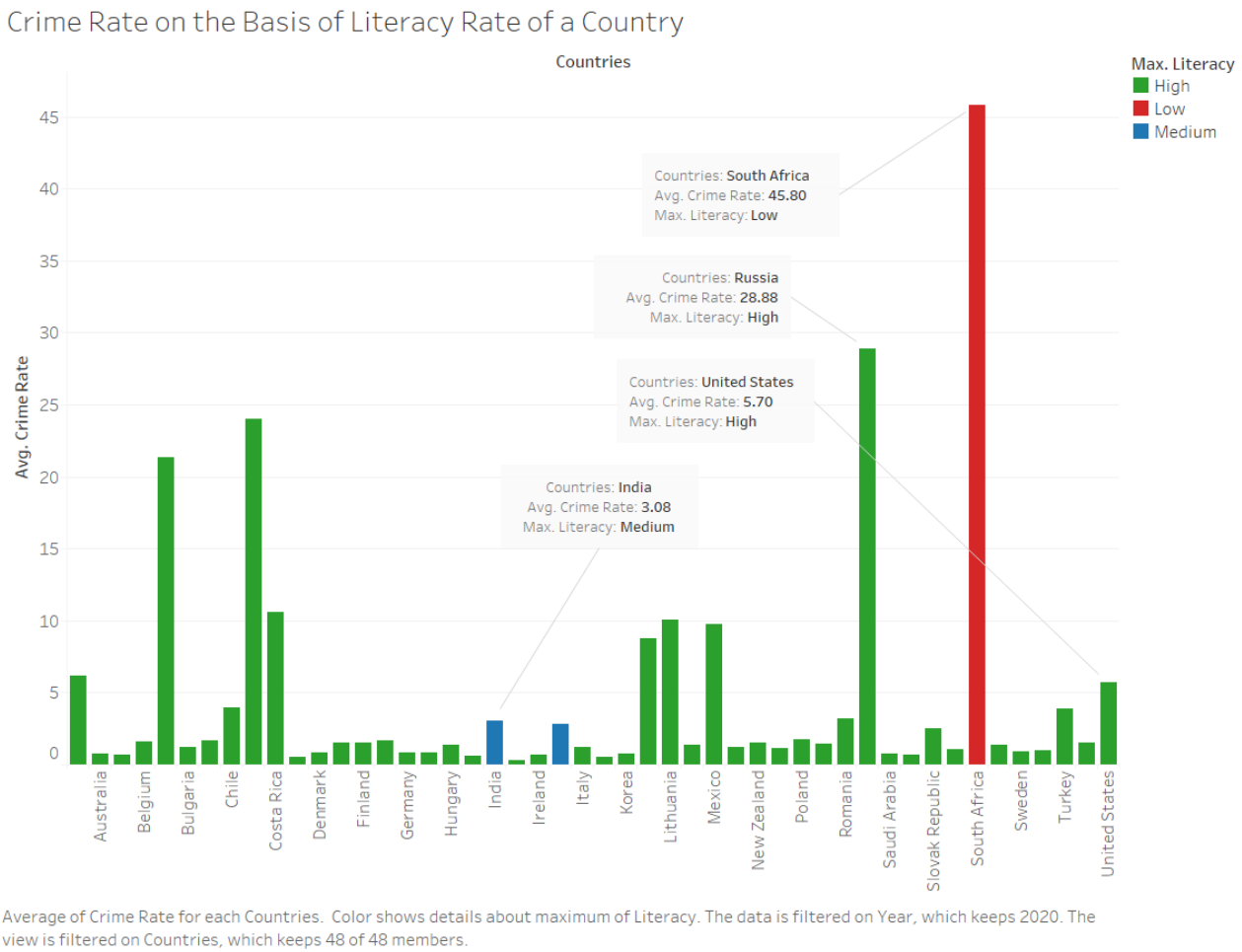


Figure 66: Multi Variable Bar Graph shows the literacy rate, and crime rate per country in the year 2020

This Multi Variable Bar Graph shows the literacy rate, and crime rate per country in the year 2020. The purpose for the graph is to determine if there is a correlation between the two variables based on the data found. As Hilton Hotels required employees who are literate, it is important to consider the

average literacy rate of each country. In addition, the crime rate is also important as a lower crime rate will make tourists feel safer and more likely to choose that country to visit. Based on the visual, it can be observed that South Africa has the highest crime rate overall and has low literacy. On the other hand, it can be observed that countries tend to have higher literacy rate and low crime rate. This indicates that there is a correlation between then, with less crime in higher literacy rate countries.

- **Predictive Models**
    - Forecast Model

Visualization represents the dotted graph of the current population and projected population

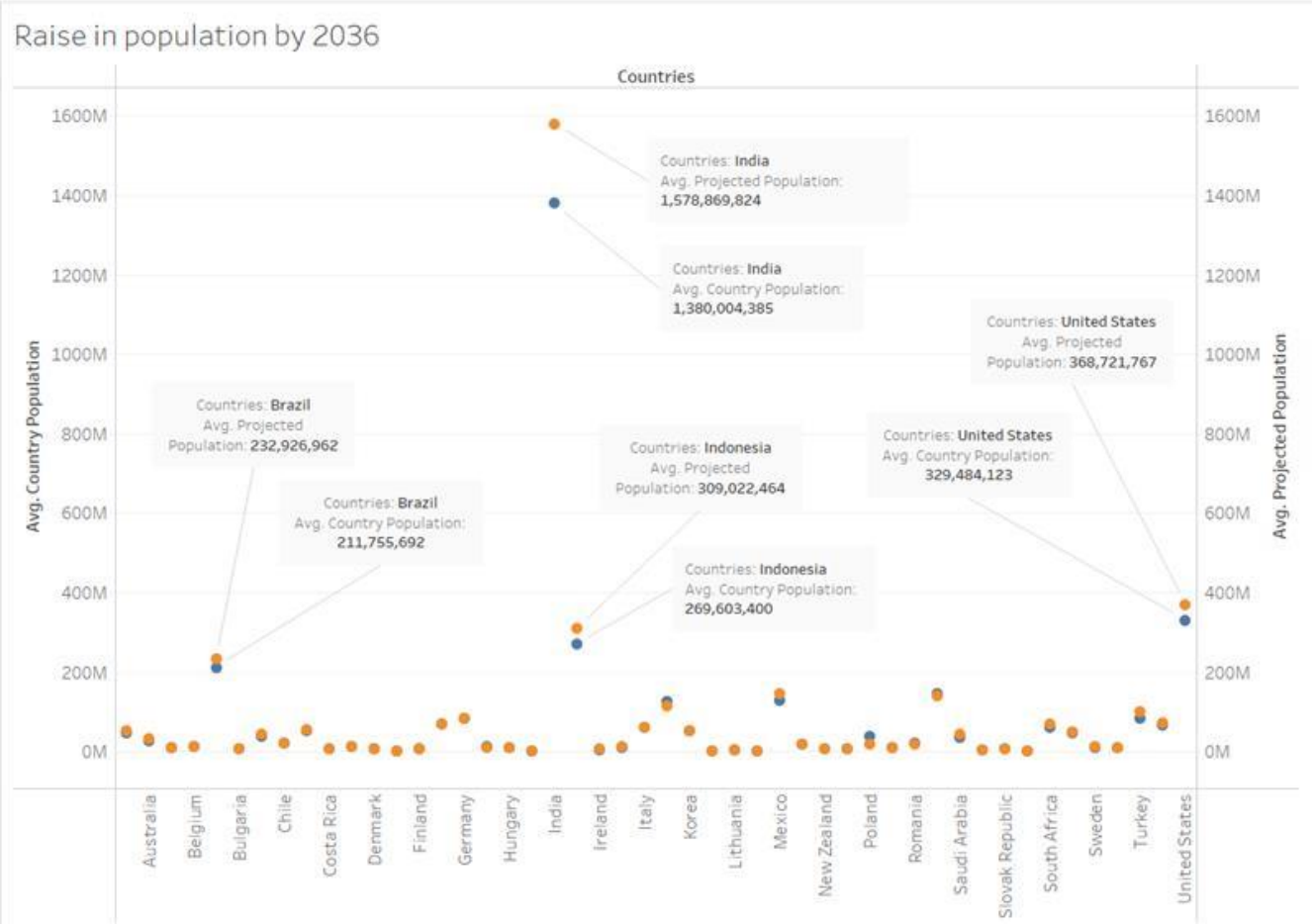


Figure 67: Dotted graph of the current population and projected population

This dotted graph shows the current population(blue) and the projected population(orange) per country for the year 2036. The purpose for the graph is to show the increase or decrease in population depending on the country. It can be observed that India will have the greatest increase in population. Brazil, the United States, and Indonesia will also see a slight increase. Mexico is actually projected to decrease in population. Based on the visual, it was determined that Hilton can capitalize on the growth of India by entering the market early on before the population skyrockets.

Visualization represents the forecast of the crime rate for five countries

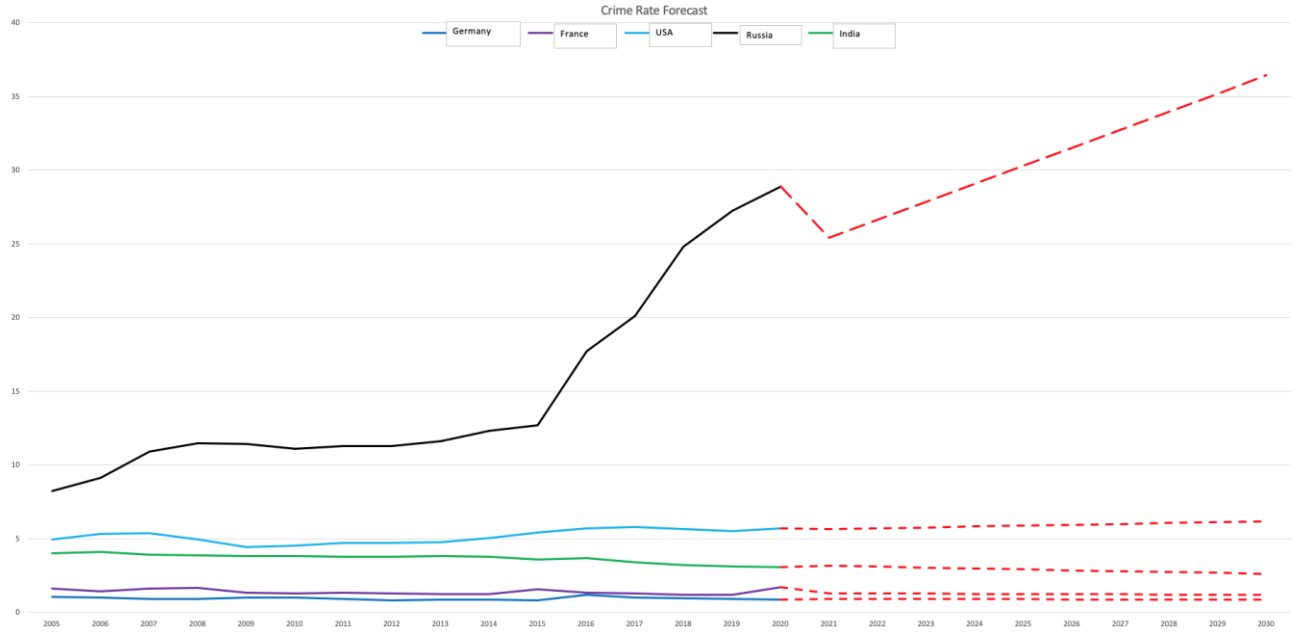


Figure 68: Shows the forecast of the crime rate for five countries

This Forecast model shows the forecast of the crime rate for five countries (Germany, France, USA, Russia, and India.) The purpose of this graph is to predict the level of crime in the years to come based on historical crime data. It can be observed that Russia has tremendous amount of crime and is expected to increase. The same can be observed for the United States. Germany and France have very

little crime and are forecast to remain low. But what is most interesting is that India is forecasted to decrease in crime rate, although not by much, but based on the size of India's population, it is a very appealing insight that Hilton can used in determining the next country they expand to.

Dashboard(s)

Dashboard for revenue, average monthly salary, crime rate and literacy rate for all the countries

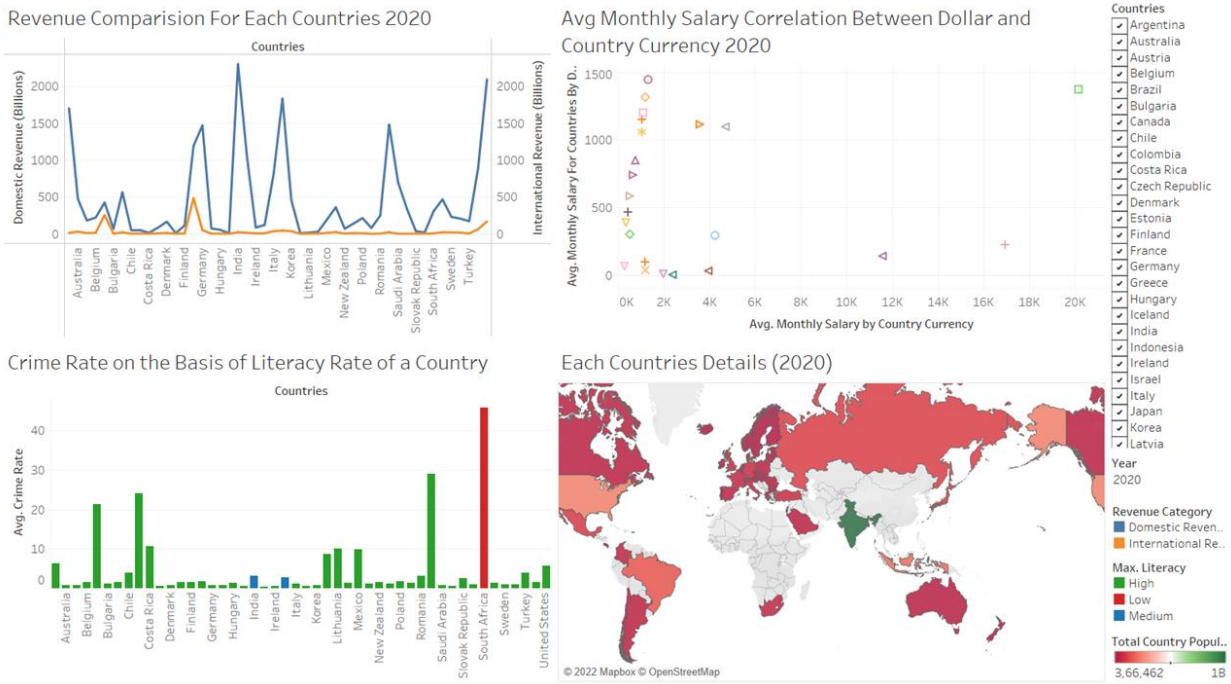


Figure 69: Shows the dashboard for revenue, average monthly salary, crime rate and literacy rate for all the countries

Dashboard for revenue, average monthly salary, crime rate and literacy rate for top five countries

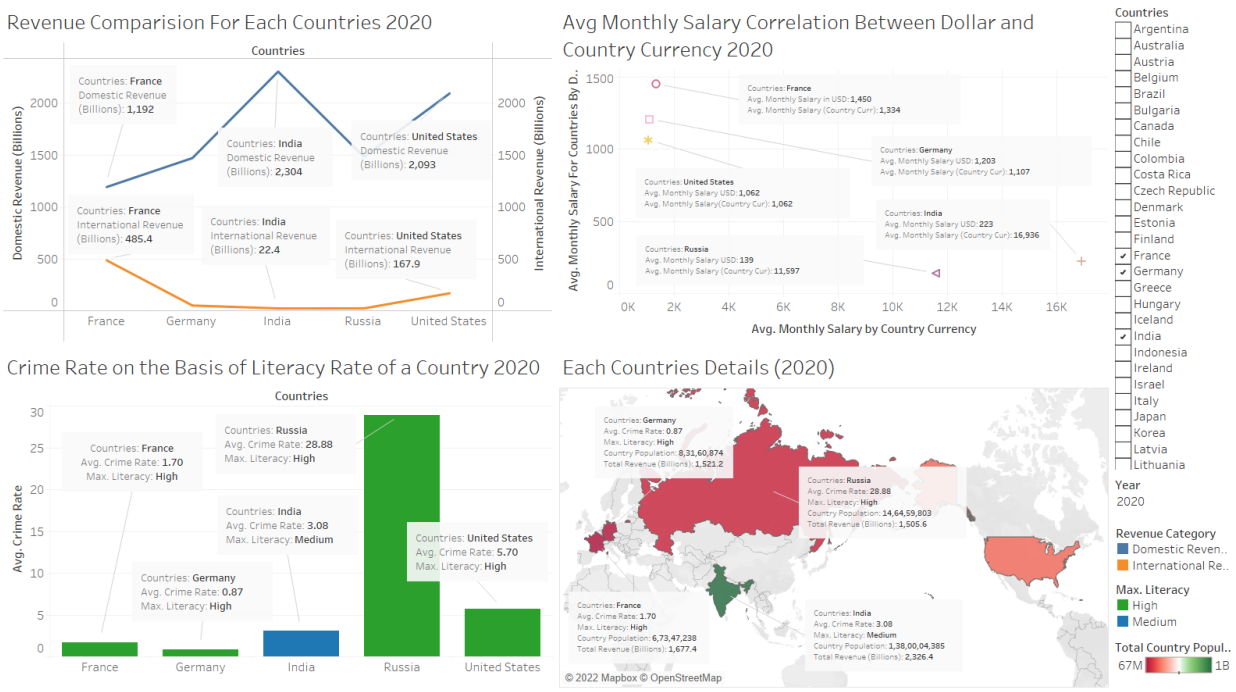


Figure 70: Shows the dashboard for revenue, average monthly salary, crime rate and literacy rate for top five countries

These dashboards above were created to show that there are endless possibilities to create visuals that can inform Hilton Hotels on the different country's details. In addition, variables can be filtered or modified to be included or excluded from the visual. This interactive ability is integrated in most data visualization software, allowing the user to explore every country in detail. The dashboard shows four different charts, first one is the domestic and international revenue comparison for different countries. From that graph we see that India has highest domestic revenue and France has the highest international revenue. Second graph is about the salary comparison between dollars and other country currencies. From the analysis we see that Russia has the lowest annual average salary in USD and India is the second-lowest country with less annual average salary in USD. Third chart is about crime rate and literacy rate in different countries. From the analysis we see that Russia has the highest crime rate and France has the second highest crime rate. Fourth graph is the geographical map which shows the all the details like revenue, crime rate, literacy rate and population. From this graph we see that India has

the highest revenue and high population and when comparing to India's population, its crime rate is low compared to other countries.

Evaluation

a. Description

The primary goal was to determine the highest-ranking and lowest-ranking country by considering factors like number of tourist arrivals, population size, crime rate, literacy rate, female and male ratio, international and domestic revenue, etc. in different countries from the year 2005 – 2020. From the analyses its shown that the top five countries that Hilton should consider expanding are France, Germany, India, United States and Russia.

b. Dependency

For dependency analysis, the objective was to find the dependent data and analyze how they are dependent on each other. From the dataset, we analyzed the correlation between international revenue and tourist arrivals, the correlation between average salary and average working hours, the correlation between population size and number of people employed in the tourism industry and the correlation between average salary and working hours. Based on the analysis we see that the literacy rate is directly proportional to crime rate. Countries with highest literacy rate has low crime rate and countries with low literacy rate has highest crime rate. From the correlation between average salary and working hours it was determine that the highest paid worker by month is France, with United States, India and Russia having lower monthly salary in USD. After evaluation, we can say that we have achieved the objective by analyzing the relationship between the dependent variables

c. Prediction

The purpose of the prediction model is to predict the level of crime in different countries based on historical crime data. It can be observed that Russia has tremendous amount of crime and is expected

to increase. The same can be observed for the United States. Germany and France have very little crime and are forecast to remain low. But what is most interesting is that India is forecasted to decrease in crime rate, although not by much, but based on the size of India's population, it is a very appealing insight that Hilton can used in determining the next country they expand to. With this analysis we could conclude that India feels safe, therefore visitors might feel safe visiting India

Deployment

Our business objective involved identifying potential countries that Hilton Hotel can expand to, basing our decisions on any available country data that we can find. Considering Hilton Hotels aggressive expansion strategy, the group decided on factors that would affect the company's decision such as crime rate, literacy rate, population size, average salary, average number of working hours, and others. Using Business Intelligence techniques, we explored and observe the data to identify any insight that we can derived from visualizing the data. We compared country data side by side to determine which country had the overall best appealing factors within the available countries.

- Based on our descriptive model analysis it is its shown that the top five countries that Hilton should consider expanding are France, Germany, India, United States and Russia. There were many factors that was considered for these analysis like population, revenue, tourist arrival, crime rate, literacy rate etc. We could determine the highest-ranking and lowest-ranking by country
- Based on our dependency analysis, we see the correlation between international revenue and tourist arrivals, the correlation between average salary and average working hours, the correlation between population size and number of people employed in the tourism industry and the correlation between average salary and working hours.

- Based on our predictive analysis we predict the level of crime in different countries based on historical crime data. With this analysis we could conclude that India feels safe, therefore visitors might feel safe visiting India

Ultimately, the group recommends India as the next country to expand considering its medium literacy rate, very low crime rate compared to population size, total revenue, the number of hours the average person works for annually, and the amount of people working in the tourism industry. Although other countries had better rankings in other areas, India shows potential to increase Hilton's revenue and international property portfolio and finding and retaining a skilled and knowledgeable workforce.

Glossary

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications.

Palisade StatTools - statistical analysis add-in for MS Excel.

References

- Applegate, L.M., Dev, C. & Piccoli, G.(2008) 'Hilton hotels brand differentiation through customer relationship management' [Case Study], *Harvard Business Review,* pp.1-18. Available at: https://www.hbs.edu/faculty/Pages/item.aspx?num=36237 [Accessed 8 February 2022]
- FAQs. (n.d.). Retrieved from https://ir.hilton.com/investor-resources/faqs
- Hilton Worldwide Holdings Inc (Tii:HLT). (n.d.). Retrieved from https://www.tiicker.com/brand/HLT

- Hilton worldwide revenue by segment 2021. (2022, February 16). Retrieved from https://www.statista.com/statistics/297761/revenue-of-hilton-worldwide-holdings-inc-by-business-segment/