STAT 112 PROJECT: ANALYSIS OF TOURIST DATA

Name and Surname: School ID:

1-Hüseyin Çağatay Bayrak 2561074

2-İsmet Eren 2561249

3-Bülent Coşkun 2561124

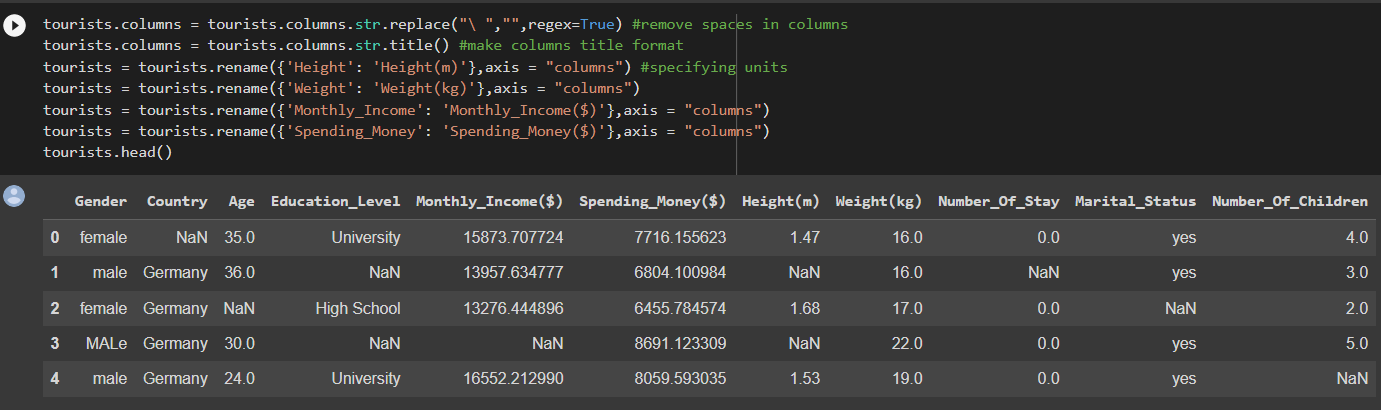
4-Elif Güvendik 2561280

ABSTRACT

We have information on the tourists to Antalya, Turkey, in the summer of 2022, including their gender, age, country of origin, level of education, monthly income, number of children, height, and weight, as well as their monthly income, duration of stay, and marital status.

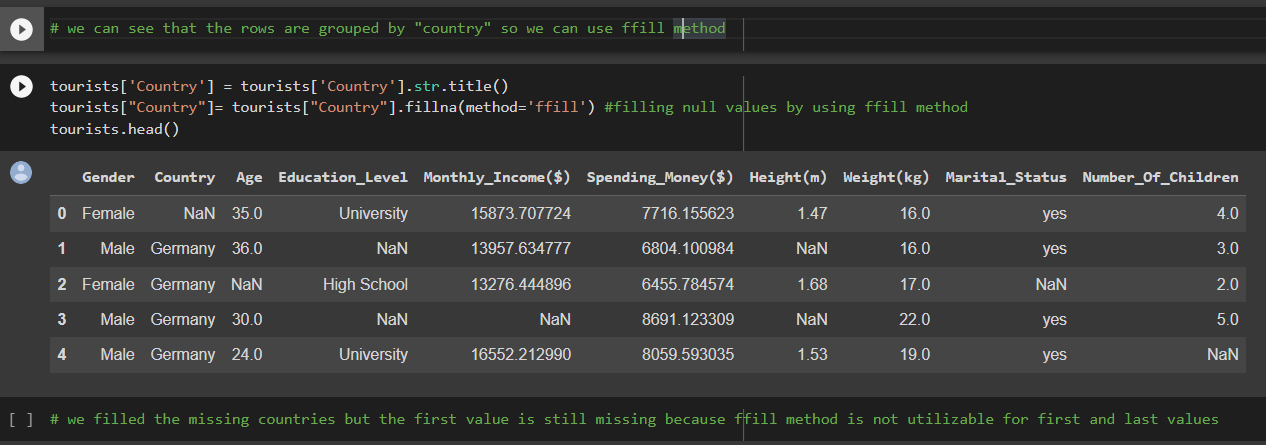
INTRODUCTION

We made an effort to ensure that the variables and responses in the dataset could be used before posing any queries to it. In contrast, none of the numerical variables in our sample had units. We safely assumed that the numerical variable's units were the same. Consequently, we determined weight in kilos, length in meters, monthly income, and spending in dollars. We eliminated the column during data cleaning because one of our variables in our dataset, period of stay, contained "0," "9999," and null values.

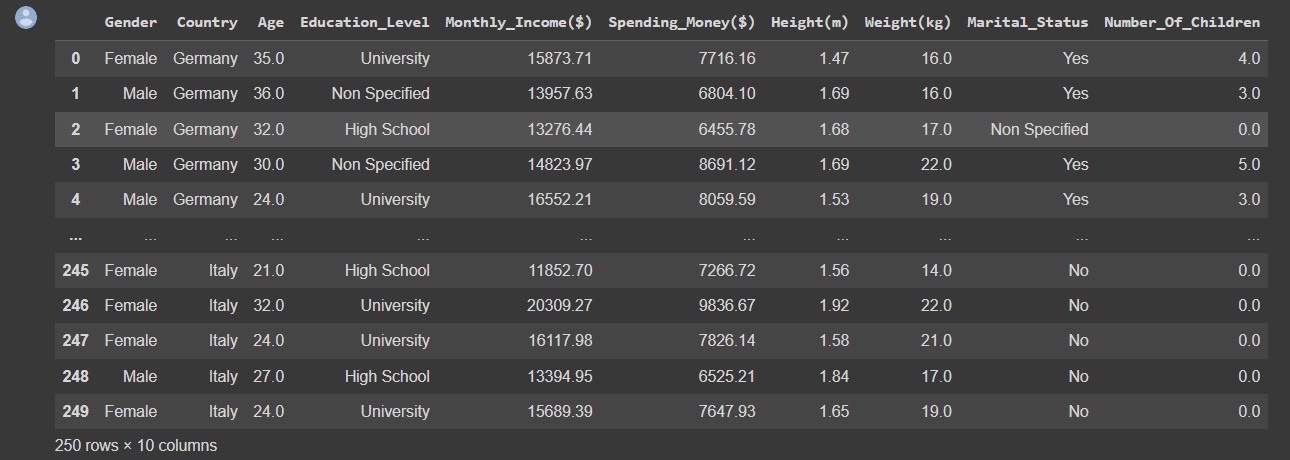


DATA TIDYING AND CLEANING STEPS

In order to adequately analyze data, we first focused on the data cleansing process. Pandas, Seaborn, NumPy, and Matplotlib were among the libraries we imported. Then, after importing our data (tourist data.xlsx), we created a piece of code that displayed any variables that were null. None of the values were duplicates. The period of stay was a variable we had, but we eliminated it because it had meaningless values. To verify irrelevant strings in the values, we built a frequency table. Columns were cleared of blank spaces. We specified units and made the column titles in that format. Our “gender” variable was transformed into a string. Using the ffill method, we grouped the rows in our data by "country."



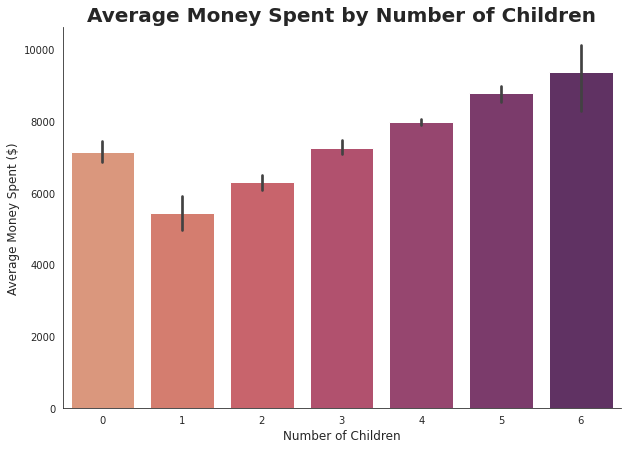
Although the missing nations have been added, the first value still needs to be entered because the first and last values cannot be filled using the ffill method. It is clear that only the first value, which was already mentioned, has to be supplied. We entered Germany as the first value and NA as the remaining two. For example, we filled the NA married status and NA education level status entries with non-specified, respectively, then converted them all to title format. We filled in the NA genders with "none specified" and formatted them all as titles. Separate averages for weight and height were entered for men and women in each nation. Finally, we used mean imputation to fill in the monthly income and spending money. We changed negative numbers to positive ones, filled NAs with mean imputation by nation, and set the number of unmarried people's children to "0."



Our data's rows are categorized by "country," thus we may apply the ffil method. Using the ffil method, we filled in the null values. The ffill method was unable to use the first and last values, hence the first value is absent. The information is arranged by the name of the nation. Because of this, we filled in NA country values with the nation group that they belong to. The NA education values also contained non-specified. Additionally, we replaced NA marital status data with non-specified and converted every value to title format. Non-specific genders (NA genders) were filled with non-specified, and we formatted them as titles. The average for men and women in each country was used to fill in the NA values for height and weight. We calculated an average value for each nation and used mean imputation to fill in the NA values for Monthly Income. Additionally, mean imputation was used to fill out the NA Spending Money field. Negative values were converted to positive ones, NAs were filled using mean imputation by nation, and the proportion of children born to unmarried couples was set at 0. Finally, in order to analyze the data correctly, we exported the data to Excel.

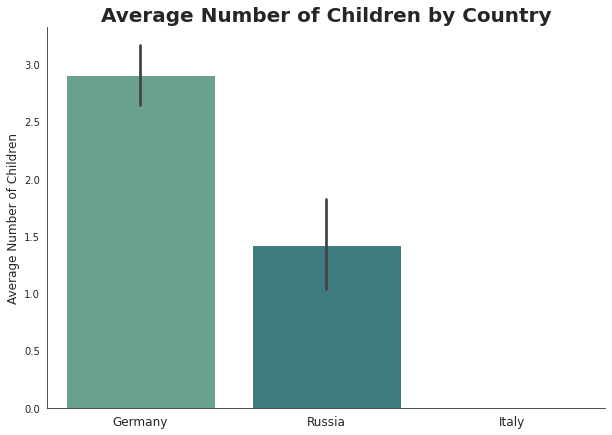
EXPLORATORY DATA ANALYSIS

1-What does the number of children brought on vacation have to do with the money we spend?



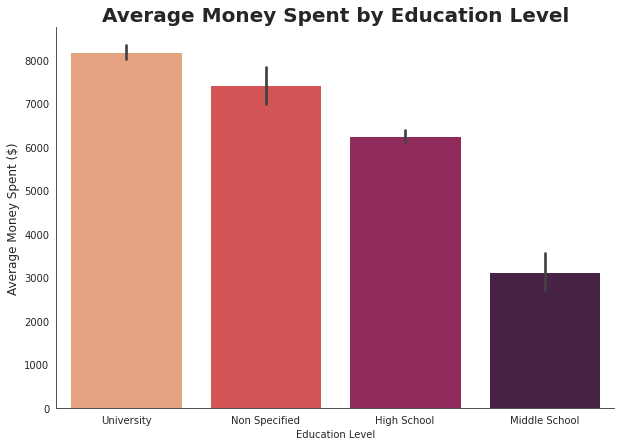
The visitors that visited Antalya in 2022 were categorized into a bar chart based on how many children they brought along for the trip in the data group we have. With the help of this graph, it is quite evident that as the number of tourists with children visiting Antalya rises, so does the amount of money they spend. By drawing this conclusion, we can see that although it is reasonable to assume that those without children will spend less money than those who do, they actually spend more than those who have one, two, or three children.

2- Does the nationality of tourists affect the number of children they bring on vacation?



Using the dataset we have, we created a bar chart to compare the average number of children visitors bring with them and their country of origin. As a result, the average number of children brought by German visitors to Antalya is 2.8; for Russian tourists, the average number of children brought is 1.5; and for Italian tourists, the average number of children brought is 0. In this instance, we can see that more German tourists brought children to Antalya than Russian and Italian tourists did. The Italians, on the other hand, do not bring any kids with them when they travel to Antalya.

3-Does the degree of education of tourists affect the money they spend on vacation?



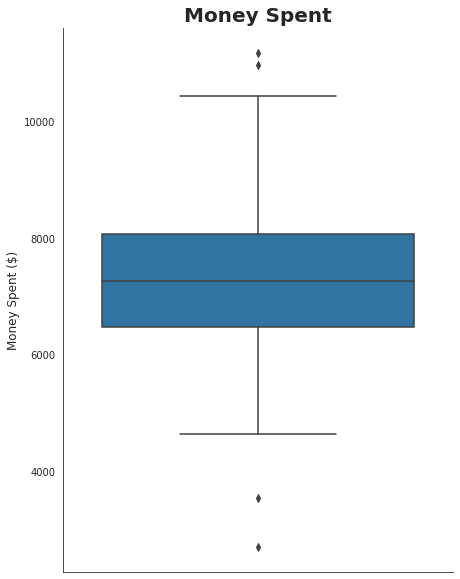
Using a bar chart and the data set, we compared the average amount of money spent and education level, two distinct variables. It is evident that higher education graduates spend more money than high school graduates, and middle school graduates spend more money than high school graduates. As a result, we can see that as education levels rise, so does the amount of money spent on vacations in Antalya.

4-Does the monthly income of tourists affect the money they spend on vacation?

Using the information we have, we created a scatter plot to compare the tourists' monthly income with their spending in Antalya. With the exception of the outliers, it is evident that spending on vacations in Antalya rises as monthly income does. We can infer from the graph that the two variables have a positive correlation with one another.

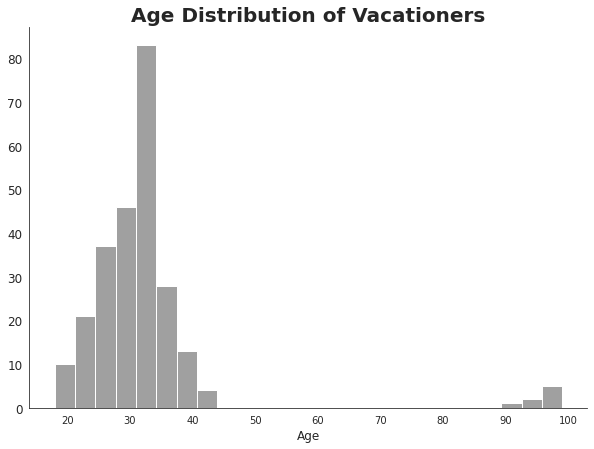
5-Does the weight of tourists affect the money they spend on vacation?

With the help of our data set, we created a scatter plot to visualize the correlation between the two variables of weight and spending. A positive correlation between the two variables can be seen in the graph. This correlation leads us to the conclusion that as weight rises, so do expenses in Antalya.

6-What is the distribution of money spent by tourists on vacation?

According to the data set we have, we created a box plot using the variable "money spent," and from the graph we have, we can see that the median amount of money spent is roughly 7200 dollars, and the range is between 4700 and 12000 dollars, however there are outliers in our data.

7-What is the age distribution of tourists coming to Antalya for vacation?



Our data indicates that the majority of visitors to Antalya are between the ages of 20 and 40. In addition, it is true that tourists between the ages of 90 and 100 visit Antalya.

CONCLUSİON  
  
We utilized various techniques for cleaning and organizing data related to individuals visiting Antalya during the summer of 2022. This step proved to be the most challenging aspect of the process. To make it manageable, we broke the data into smaller sections and formulated questions that could be effectively depicted through data visualization.