# Temasek Polytechnic

## **School of Informatics & IT**

**DIPLOMA IN BIG DATA & ANALYTICS**

**Data Warehousing and Business Intelligence (CDA2C01)**

**AY 2020/2021 October SEMESTER**

**Individual Report**

Practical Class: P04

Tutor: Mr Surojit Dutta

|  |  |
| --- | --- |
| **Student name** | **Admission number** |
| Noor Mohamed Abul Kalam Asath | 1901711J |

**Data Warehousing and Business Intelligence (CDA2C01)**

**AY 2020/2021 October SEMESTER**

**Individual Project**

**Practical Class**: P04

**Submitted by**: <1901711J> <Noor Mohamed Abul Kalam Asath>

**Date:** 11/12/2020

**“By submitting this work, I declare that I the originator(s) of this work and that all other original sources used in this work has been appropriately acknowledged.**

**I understand that plagiarism is the act of taking and using the whole or any part of another person’s work and presenting it as our own without proper acknowledgement.**

**I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism.”**

**Name and Signature of student: Abul kalam Asath (asath)**

**Data Sources**

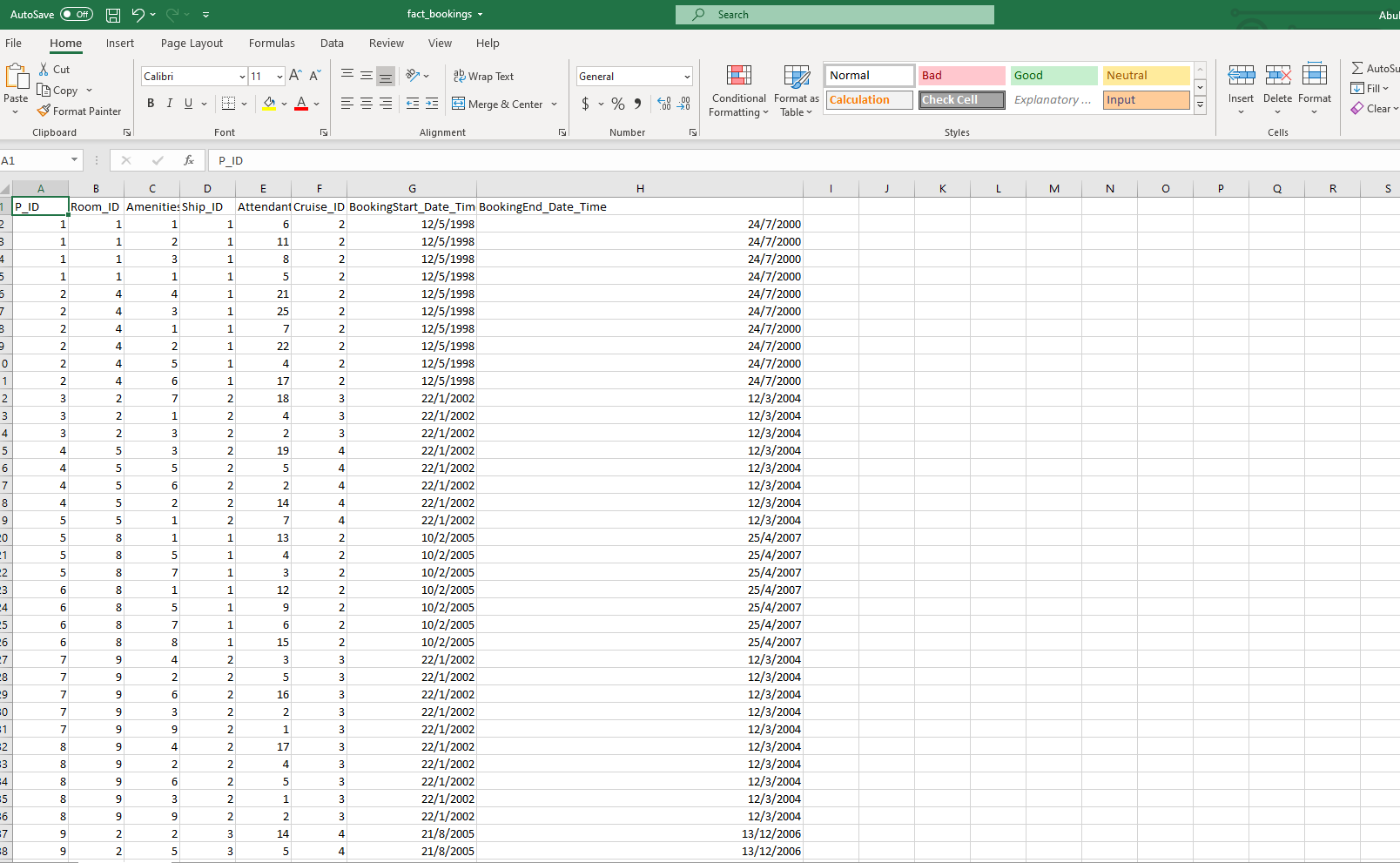
1. **Fact\_bookings:**

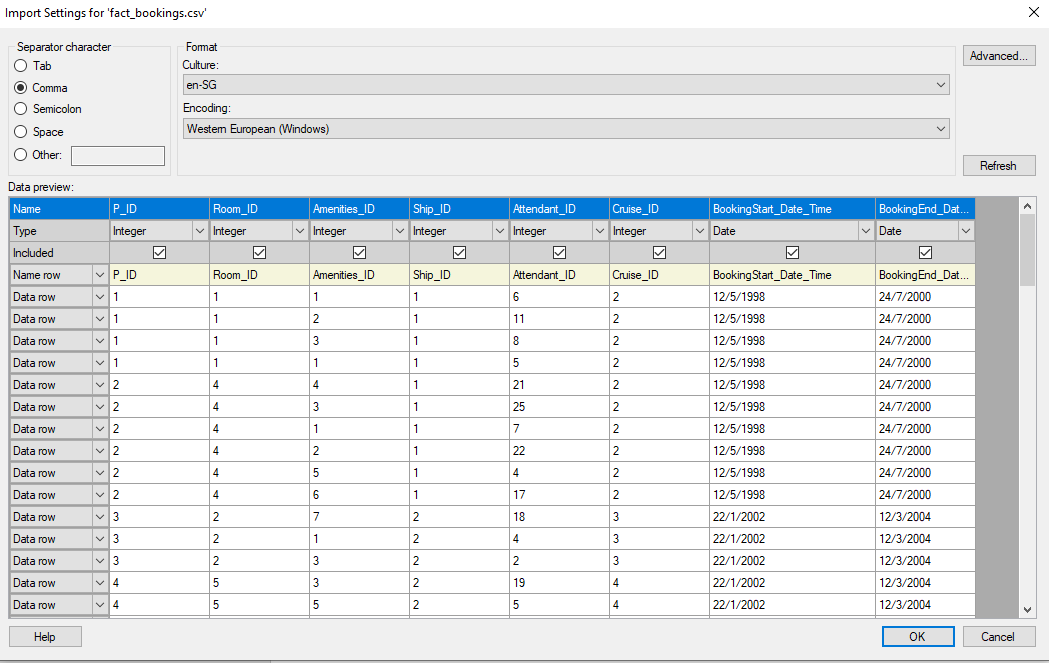
The dataset for the fact table is in Excel CSV format. It is comma delimited when loaded into Tibco Software.

The fact\_bookings table consist of 8 columns, 6 for integer and 2 for date. The 6 ID’s are Foreign key of 6 dimensions table. We have 99 records in this table. Each row contains the booking activity of passengers.

Fact table is in excel csv format because it was easier to add data in excel.

|  |  |
| --- | --- |
| **Column Name** | **Column Type** |
| P\_ID | Integer |
| Room\_ID | Integer |
| Amenities\_ID | Integer |
| Ship\_ID | Integer |
| Attendant\_ID | Integer |
| Cruise\_ID | Integer |
| BookingStart\_Date\_Time | Date |
| BookingStart\_Date\_Time | Date |

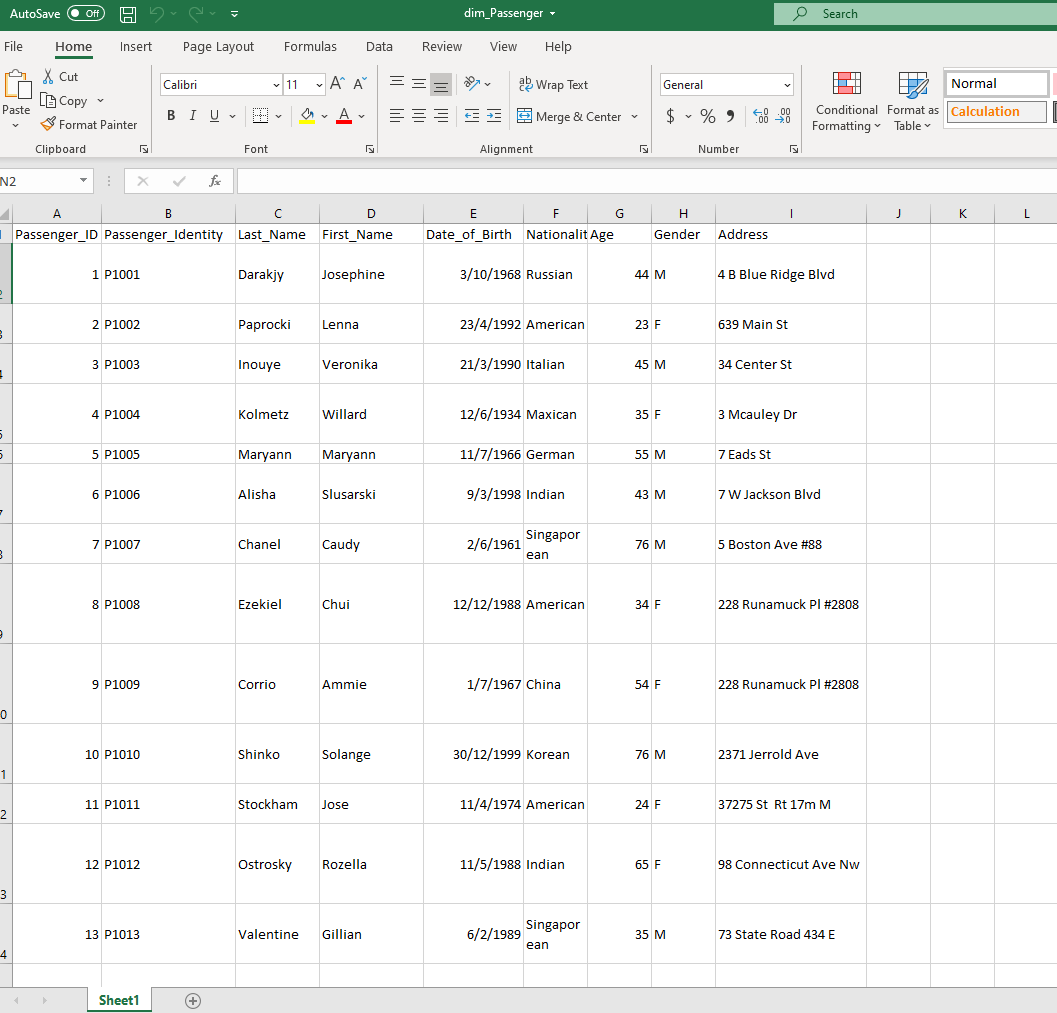


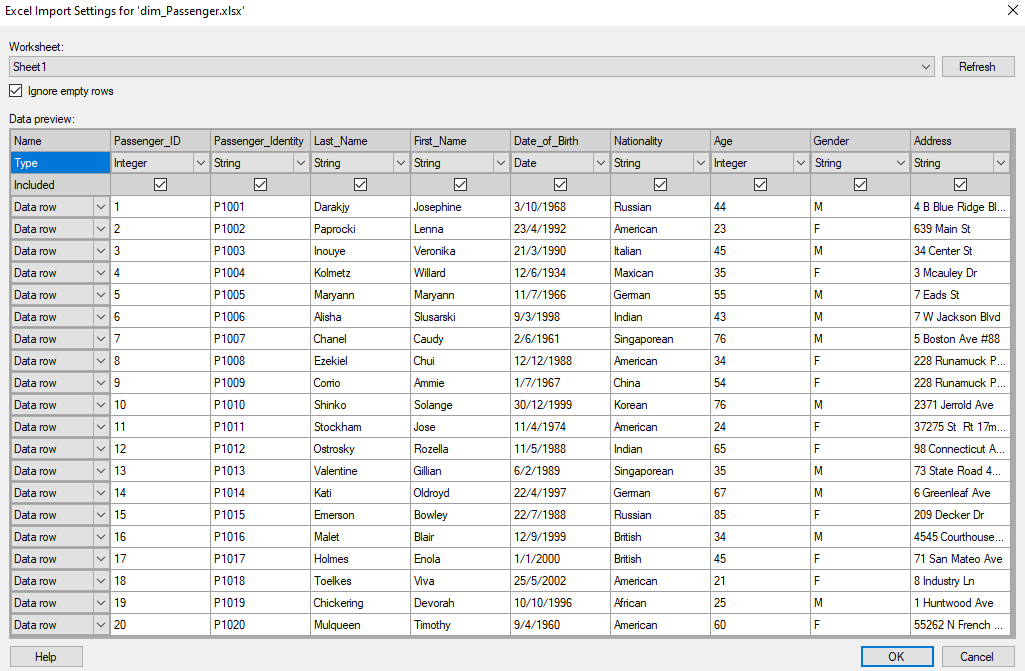


1. **dim\_Passenger:**

Dataset for the passenger dimension is in Excel workbook xlsx format. It has 20 rows and 9 columns which contains all the passenger’s details. It is useful for analysing passenger’s demographic.

|  |  |
| --- | --- |
| **Column Name** | **Column Type** |
| Passenger\_ID | Integer |
| Passenger\_Identity | String |
| Last\_Name | String |
| First\_Name | String |
| Date\_of\_Birth | Date |
| Nationality | String |
| Age | Integer |
| Gender | String |
| Address | String |

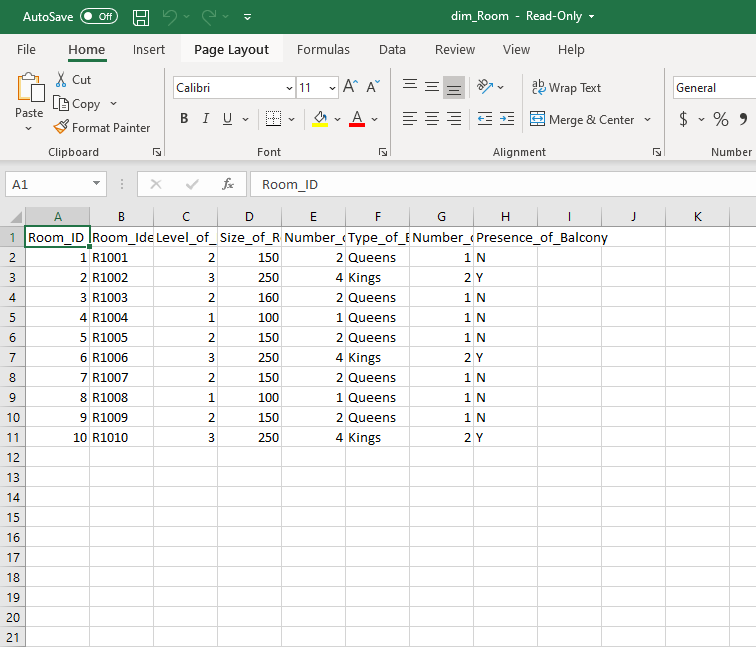


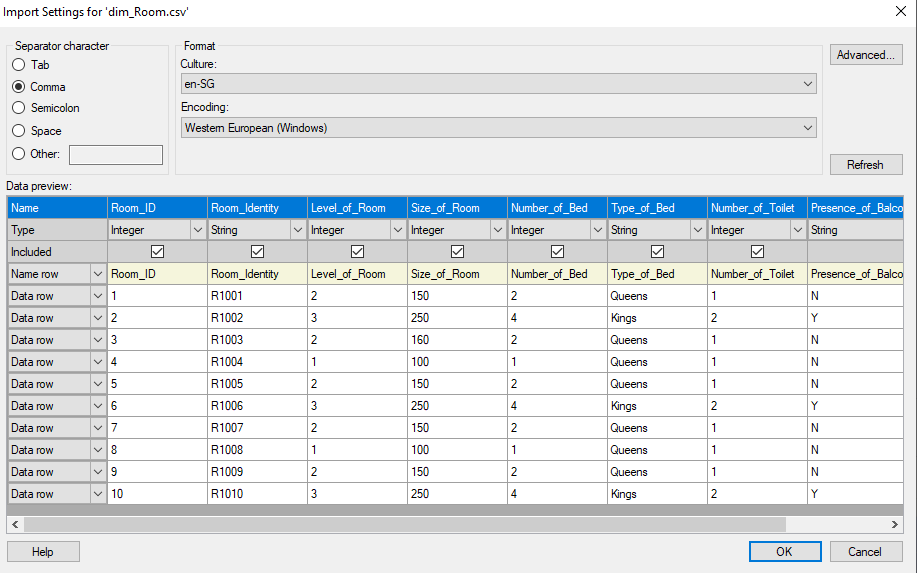


1. **dim\_Room:**

This dimension table contains information about rooms in the cruise. Each room comes with different type of comfort. Passengers can book a room based on their preferred price range. The dataset for room dimension is in Excel csv format and contains 10 rows and 8 columns.

|  |  |
| --- | --- |
| **Column Name** | **Column Type** |
| Room\_ID | Integer |
| Room\_Identity | String |
| Level\_of\_Room | Integer |
| Size\_of\_Room | Integer |
| Number\_of\_Bed | Integer |
| Type\_of\_Bed | String |
| Number\_of\_Toilet | Integer |
| Presence\_of\_Balcony | String |

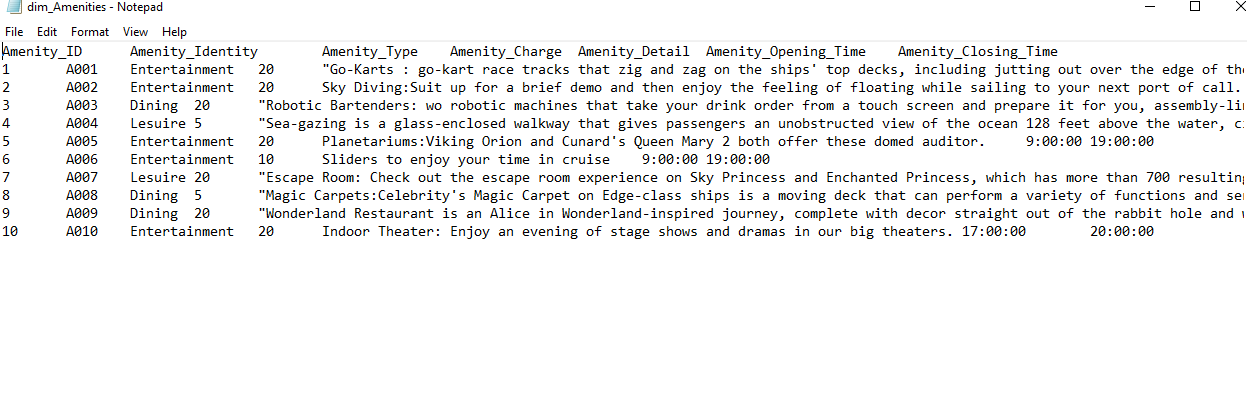


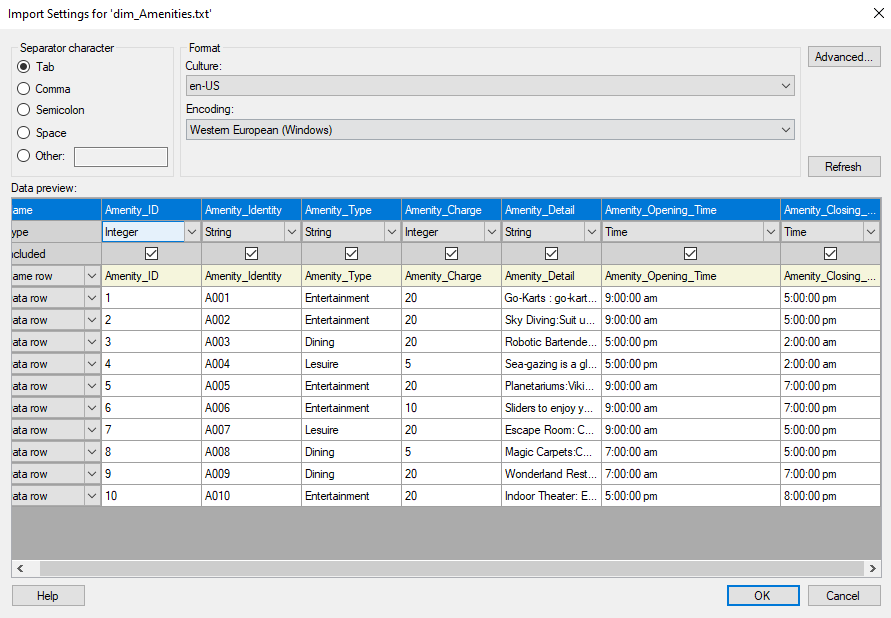


1. **dim\_Amenities:**

Data for dim\_amenities table is loaded in a text file. It contains details about all the amenities available in the cruise ship. Passenger could use all the amenities in the cruise, and they will be charged a small fee for using it. The fact table will record the amenity used by the passengers while on the cruise.

|  |  |
| --- | --- |
| **Column Name** | **Column Type** |
| Amenity\_ID | Integer |
| Amenity\_Identity | String |
| Amenity\_Type | String |
| Amenity\_Charge | Integer |
| Amenity\_Detail | String |
| Amenity\_Opening\_Time | Time |
| Amenity\_Closing\_Time | Time |

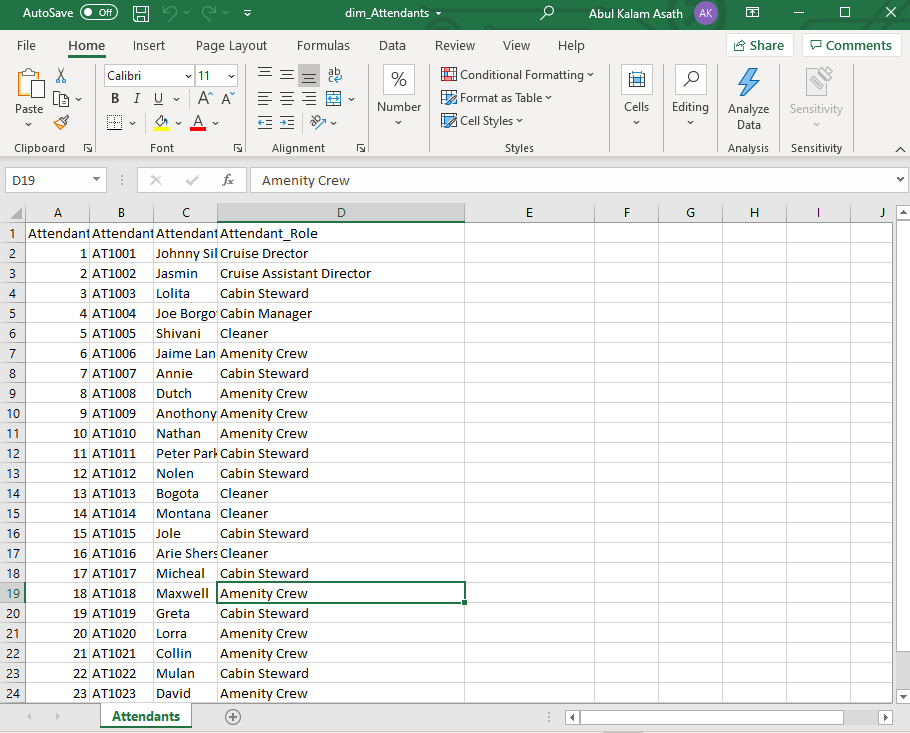


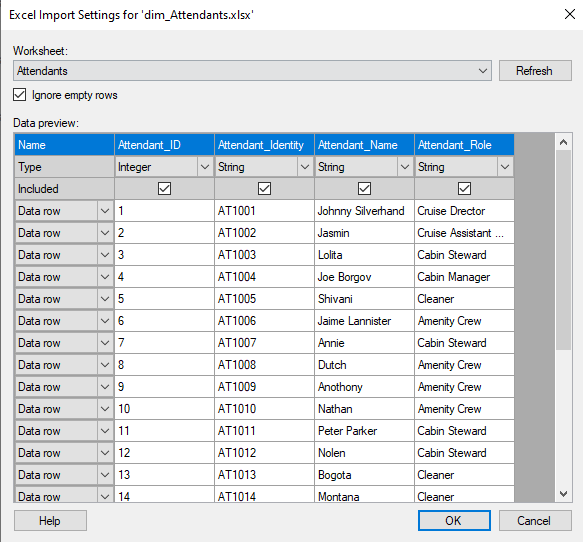
****

1. **dim\_Attendants:**

Attendants table contain information about all the workers who are working onboard the ship. It is stored in excel file in xlsx format and contains 25 rows and 4 columns.

|  |  |
| --- | --- |
| **Column Name** | **Column Type** |
| Attendant\_ID | Integer |
| Attendant\_Identity | String |
| Attendant\_Name | String |
| Attendant\_Role | String |

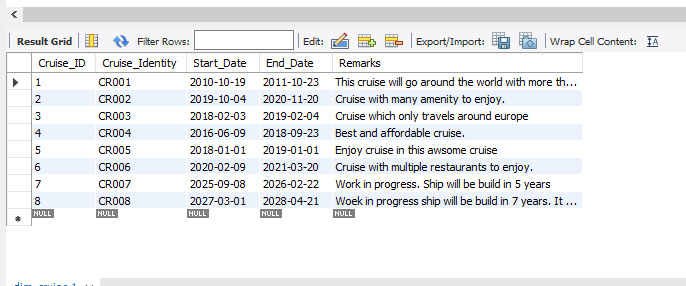


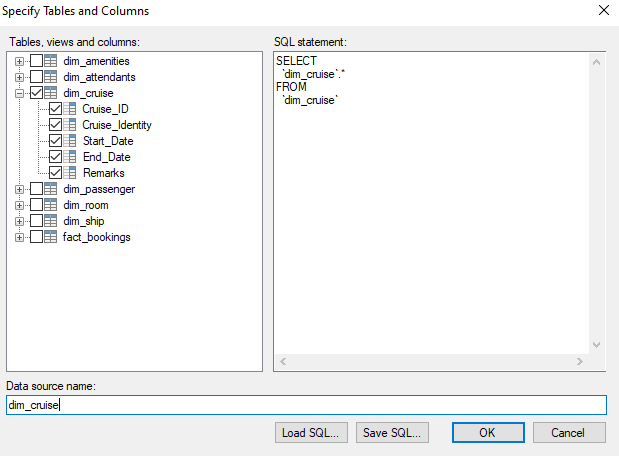


1. **dim\_Cruise:**

This table contains detail about all the Cruise available in le Dot company. The dataset for the table is located at MySQL server. It is imported into Tibco using ODBC.

|  |  |
| --- | --- |
| **Column Name** | **Column Type** |
| Cruise\_ID | Integer |
| Cruise\_Identity | String |
| Start\_Date | DateTime |
| End\_Date | DateTime |
| Remark | String |

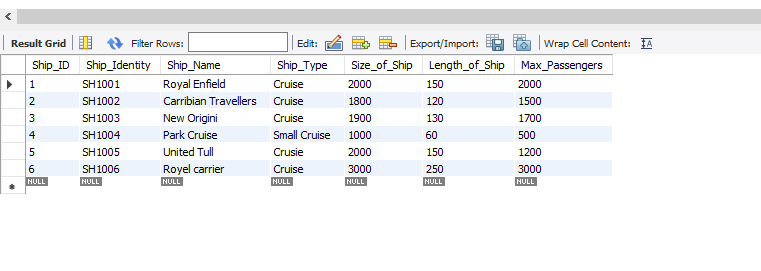


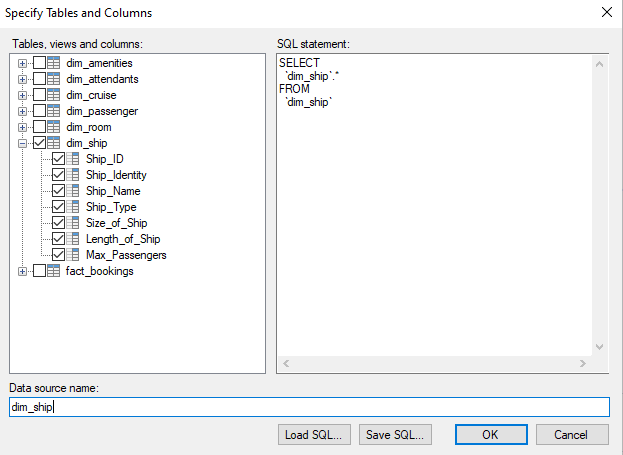


1. **dim\_Ship:**

Dimension ship contains information of all the ships details. Data for the table is located in MySQL server. we will import it into Tibco using ODBC and name it dim\_ship.

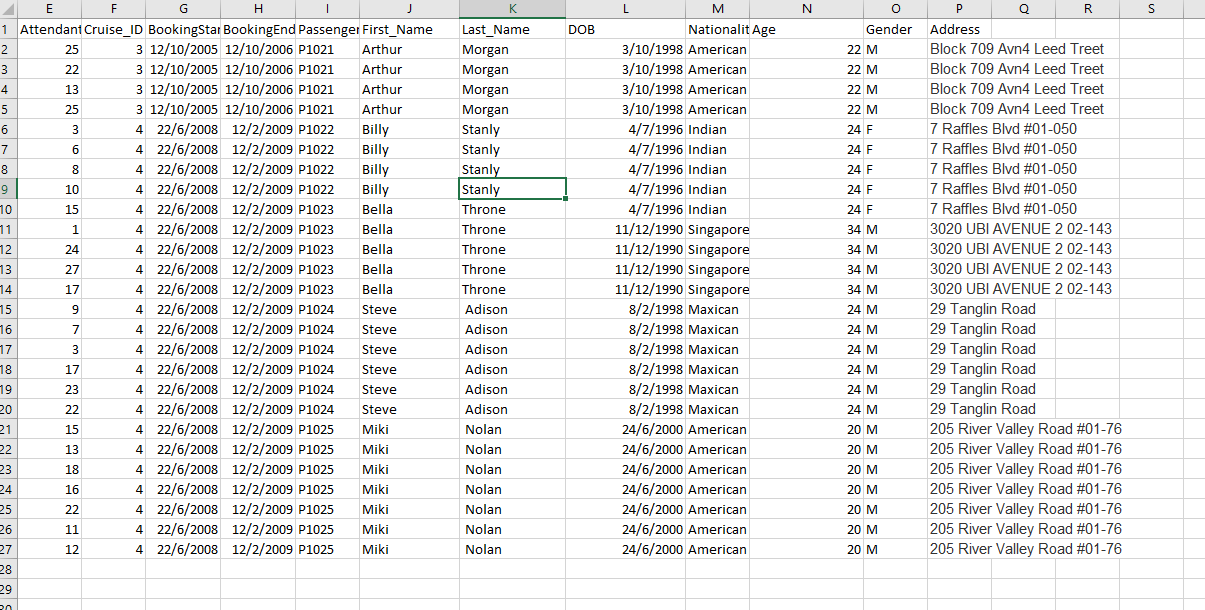
|  |  |
| --- | --- |
| **Column Name** | **Column Type** |
| Ship\_ID | Integer |
| Ship\_Identity | String |
| Ship\_Name | String |
| Ship\_Type | String |
| Size\_of\_Ship | Integer |
| Length\_of\_Ship | Integer |
| Max\_Passengers | Integer |



****

**Additional Dataset:**

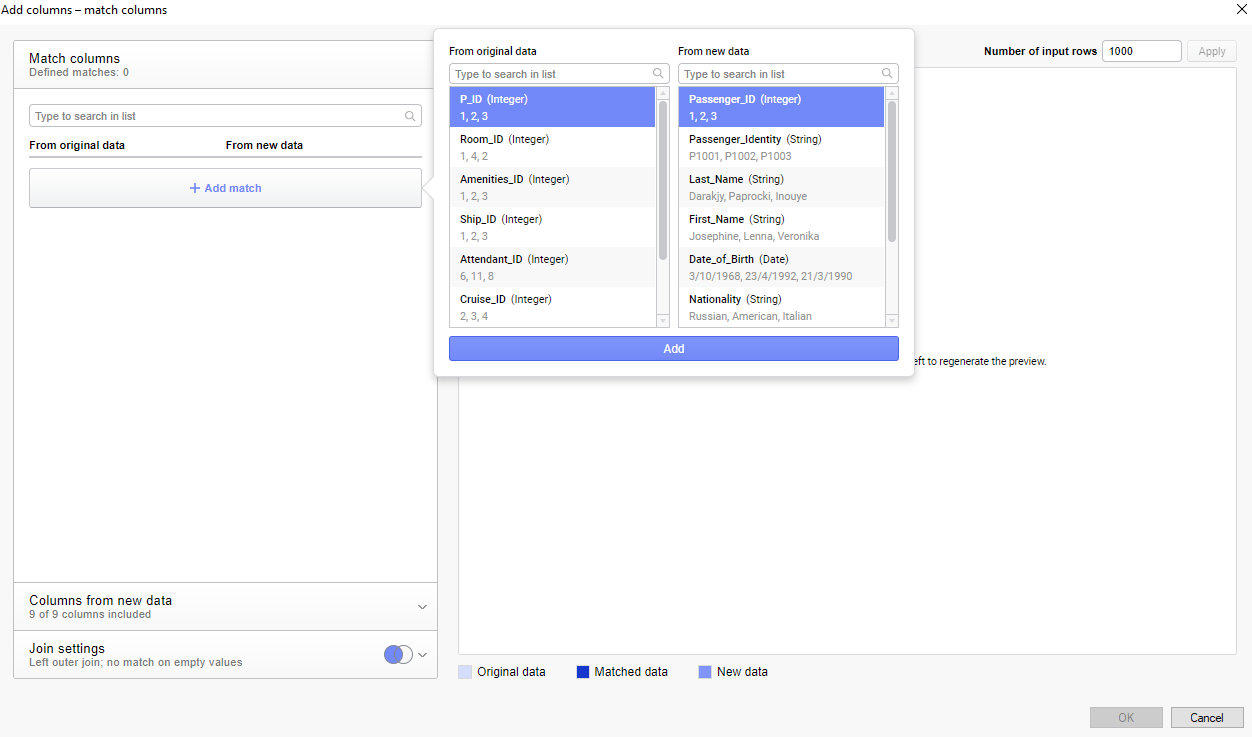
We also have an additional dataset in excel XLS format which contains both fact dimension data and passenger dimension data combined.

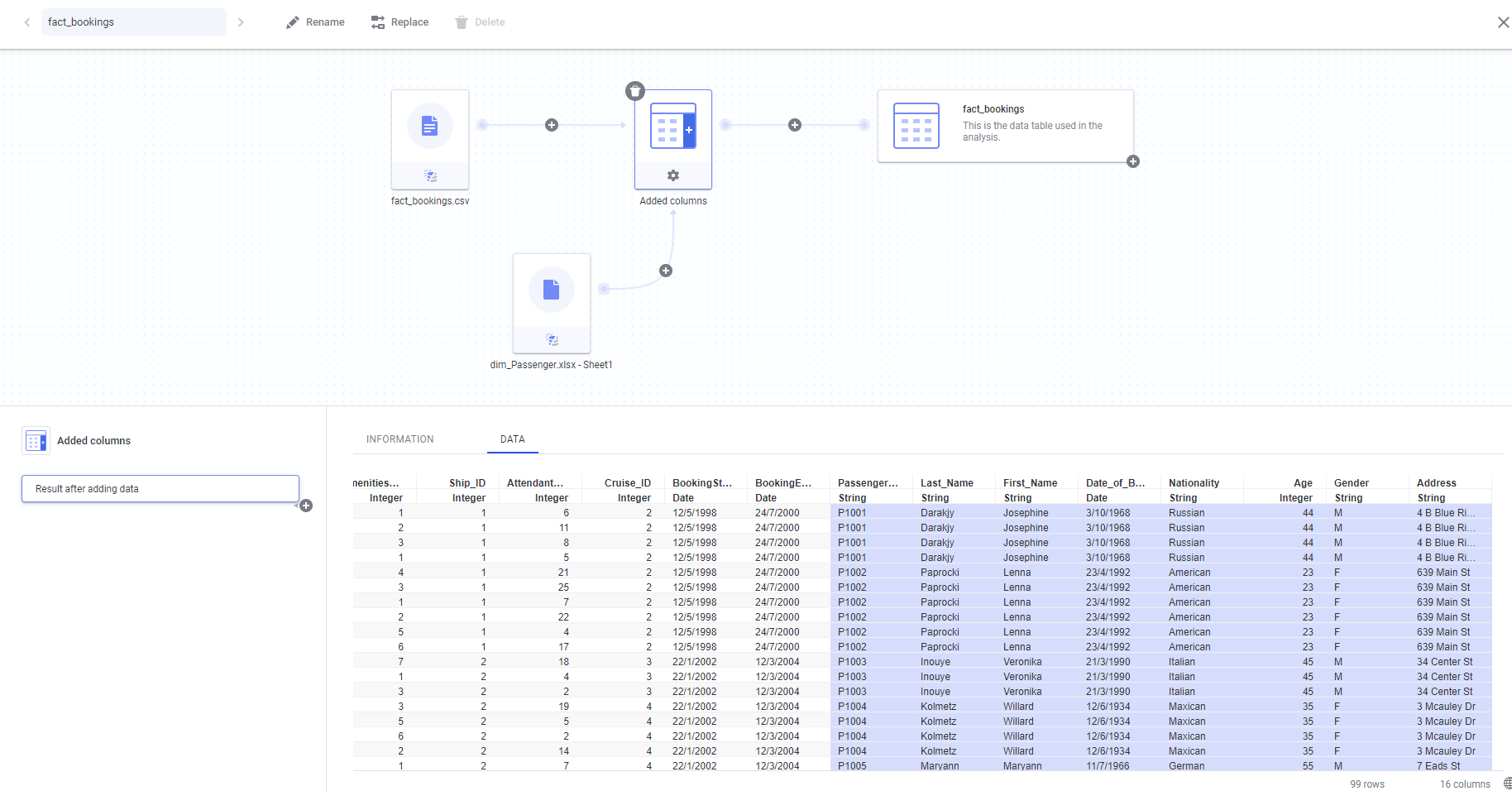


**ETL Data Blending**

**Matching Column 1:**

First data blending was made between Fact\_table and Passenger dimension. We can observe that Passenger\_ID column is not matched with any column. So, we choose the right column (P\_ID) and match it with Passenger\_ID in Tibco. We then did an inner join to join the tables

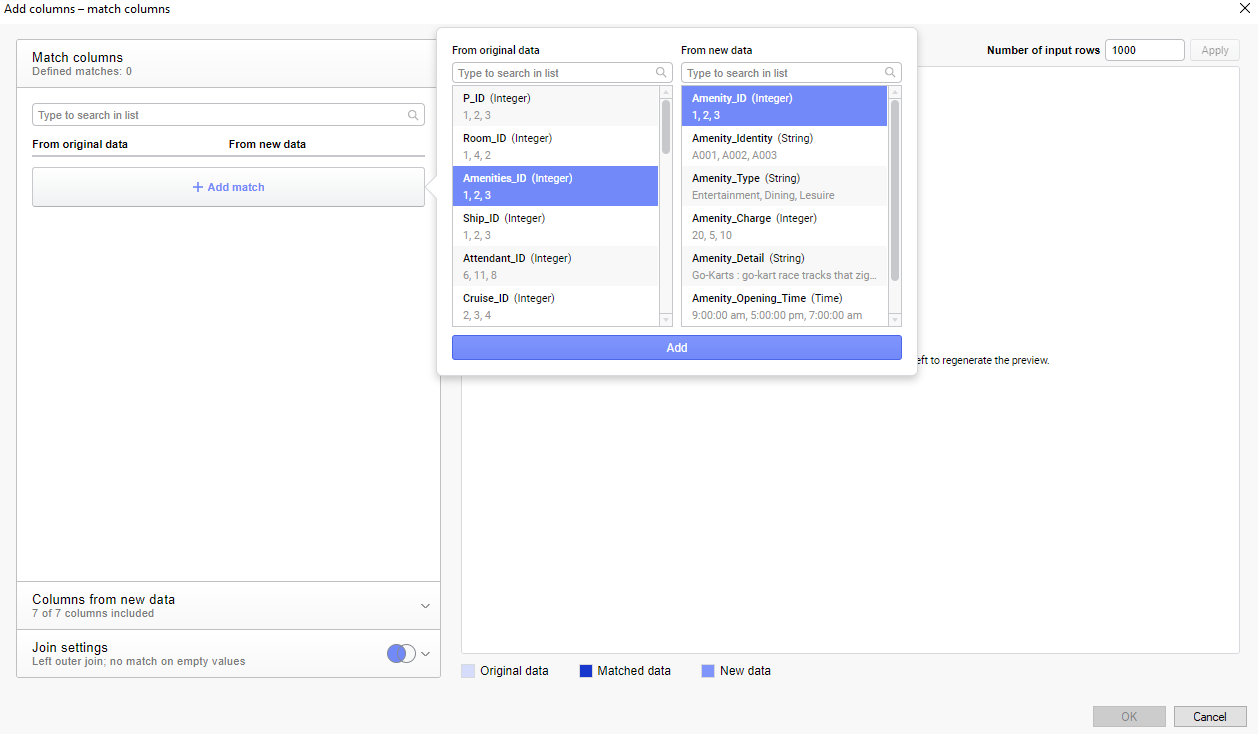




We can see that the data is blended properly with no missing values.

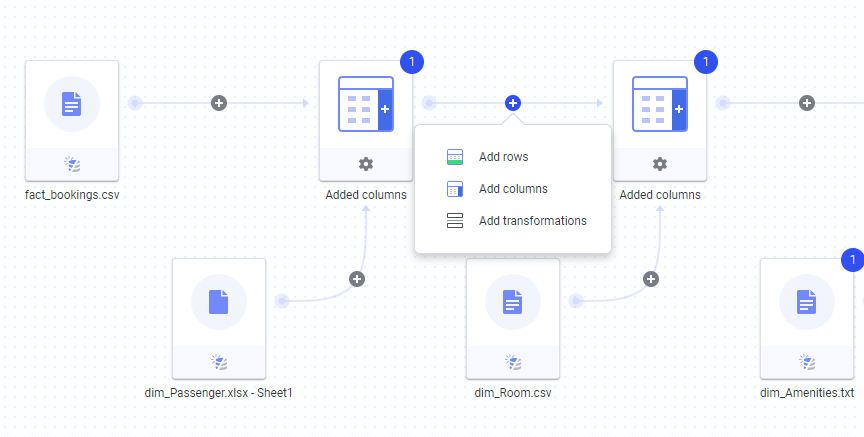
**Matching Column 2:**

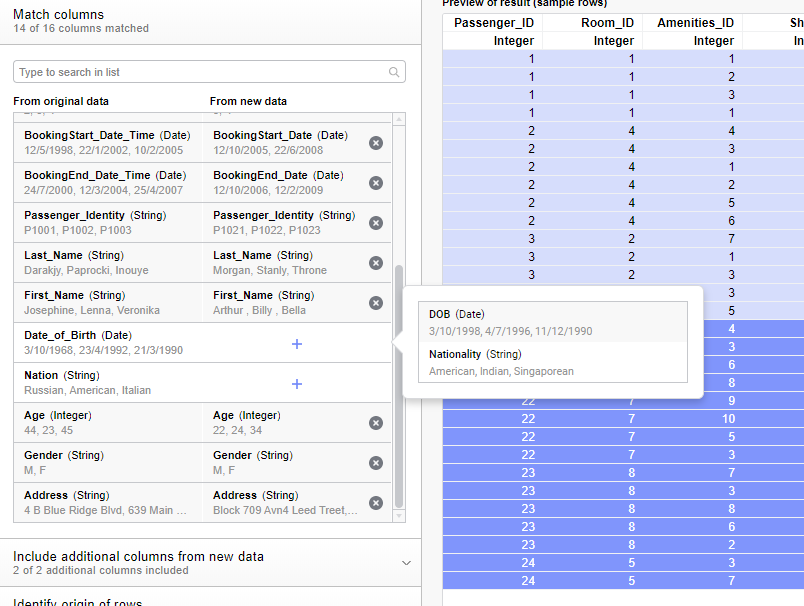
We can observe that Amenity column is misspelled in Amenity Dimension. Hence, we correctly match the column together using Add Match button.



**Matching Column 3**:

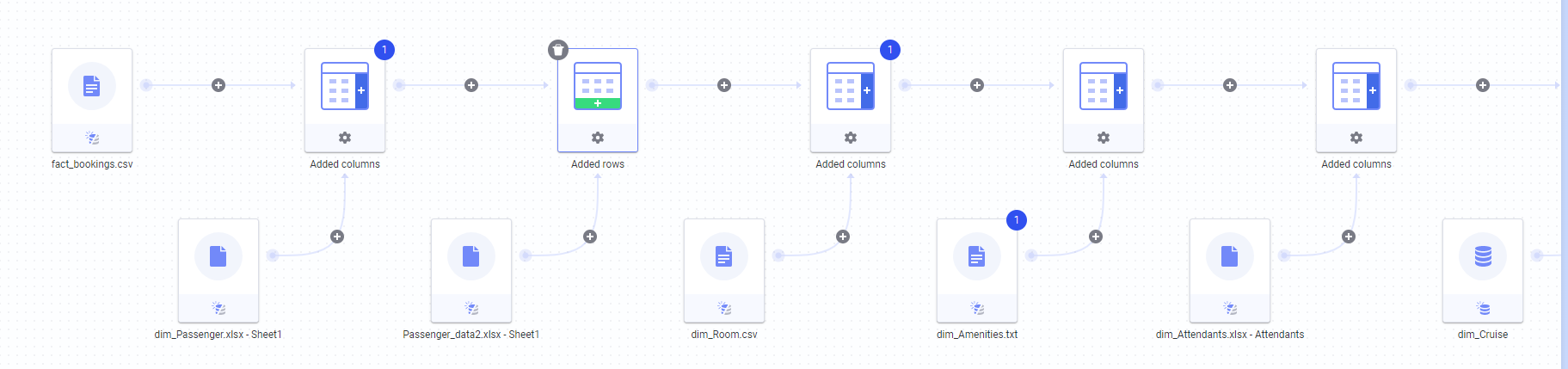
The additional data which contains both fact table data and passenger data is appended using Add rows dialog.

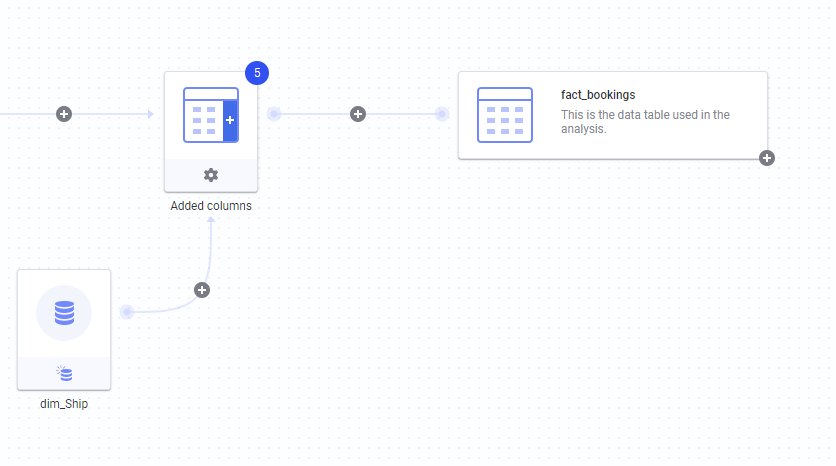




Later, i matched both Date\_of\_Birth and Nation column with matching columns.

Here are all the joins and append that I have made with my dataset so far.

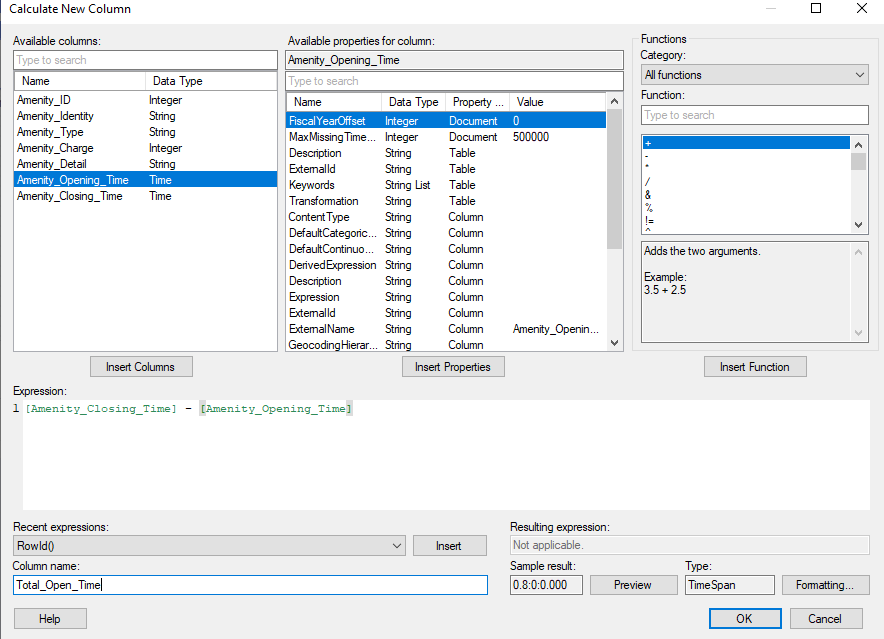


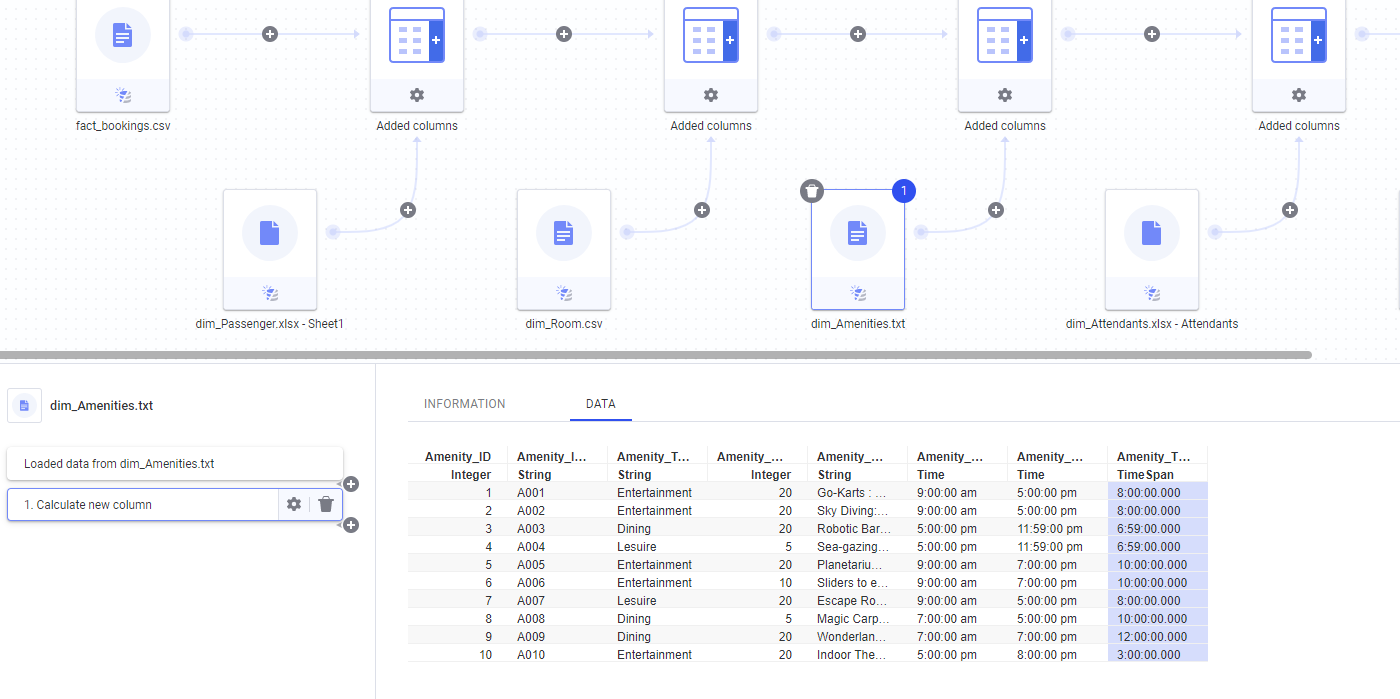


**ETL Data Transformation**

**Transformation 1:**

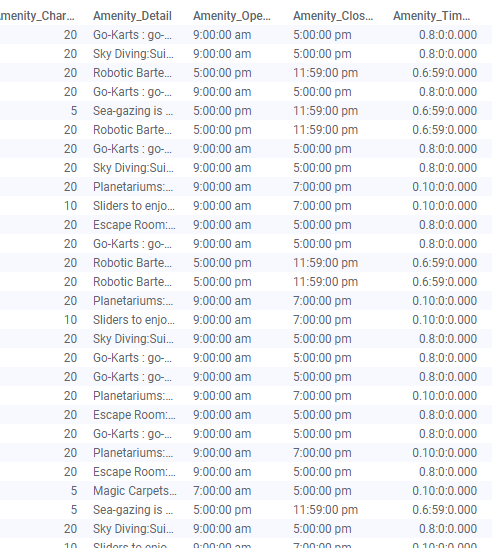
I Create a new column called Amenity\_Timespan table to calculate the total period of time an amenity stayed open. To calculate it, I used Amentiy\_Closing\_Time Column and Amenity\_Opening\_Time column from amenity table.

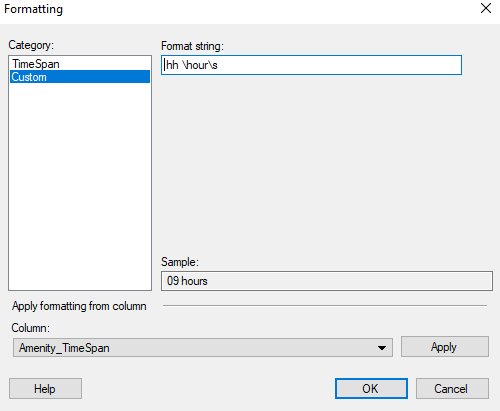




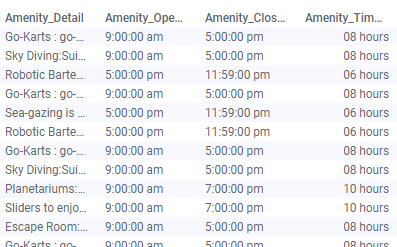
We can observe that TimeSpan column is a bit messy with multiple zeros with no indication if it’s hour or minute. Hence, I’m going to change the format.

BEFORE





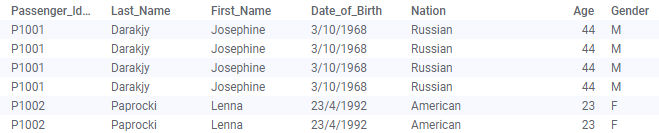
AFTER

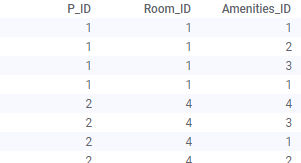


Now the column looks much cleaner for analysis.

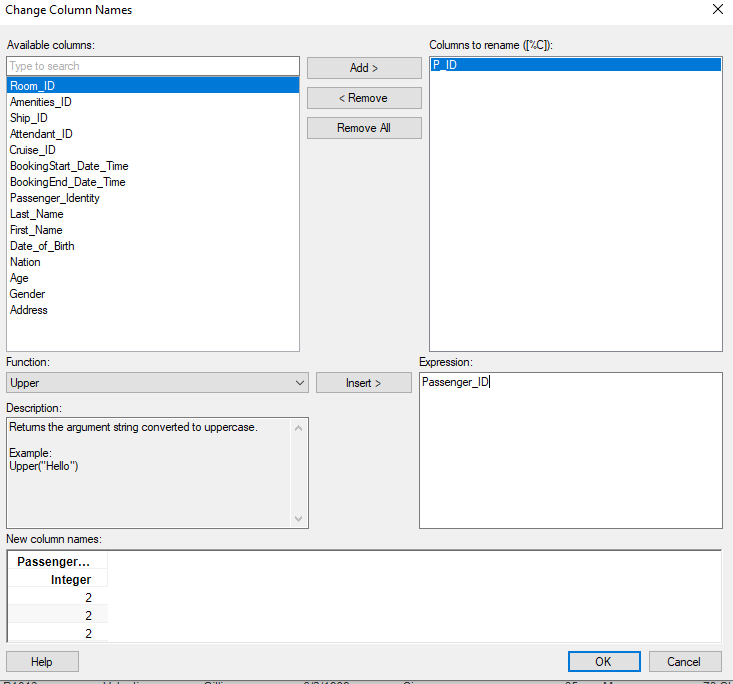
**Transformation 2:**

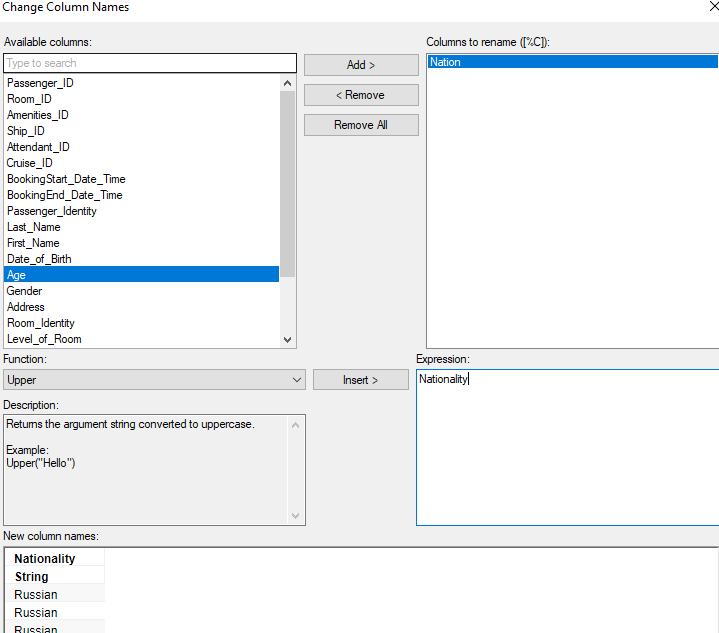
We can see from the image below that Nation column and P\_ID (Passenger \_ID) column is not spelled correctly. We will now change it.





Now, I will change Nation column name to Nationality and P\_ID column name to Passenger\_ID using **change column name** Transformation function.





**Transformation 3:**

Now, I will create a new column which contains calculated passengers ticket price under certain guidelines, For example;

* Rooms in level 1 cost $1000
* Rooms in level 2 cost $1400
* Rooms in level 3 cost $1800
* Rooms more then 150 to 200 square fit cost $200 extra
* Rooms more than 200 square fit cost $400 extra
* Rooms with 2 beds cost $100 extra
* Rooms with 4 beds cost $300 extra
* Presence of balcony cost $300

Writing IF, AND, OR , ELSE expression to calculate the ticket price can be a difficult process as we have to write many lines of code. Example:

(If level of room = 1 and Size of room < 150 and bed < 2 and presence of balcony = N , “1000”.)

(If level of room = 2 and Size of room < 150 and bed < 2 and presence of balcony = Y , “1000”.)

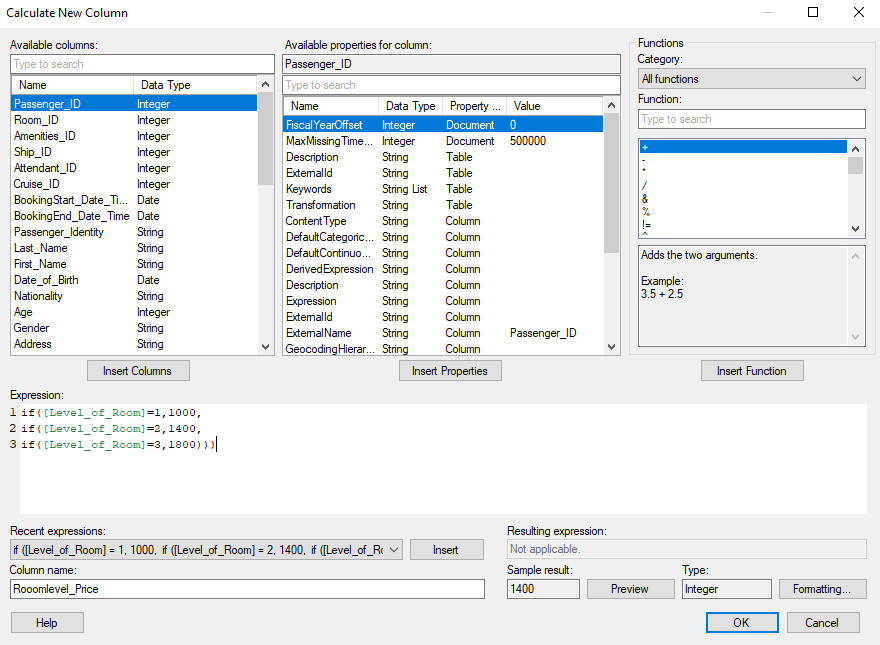
So, to make the process much easier, we will first create 4 columns detailing the price and then we add all of them in a 5th column called Ticket\_Price.

1. Column: Roomlevel\_price

If level of room = 1, “1000”,

If level of room = 2, “1400”,

If level of room = 3, “1800”

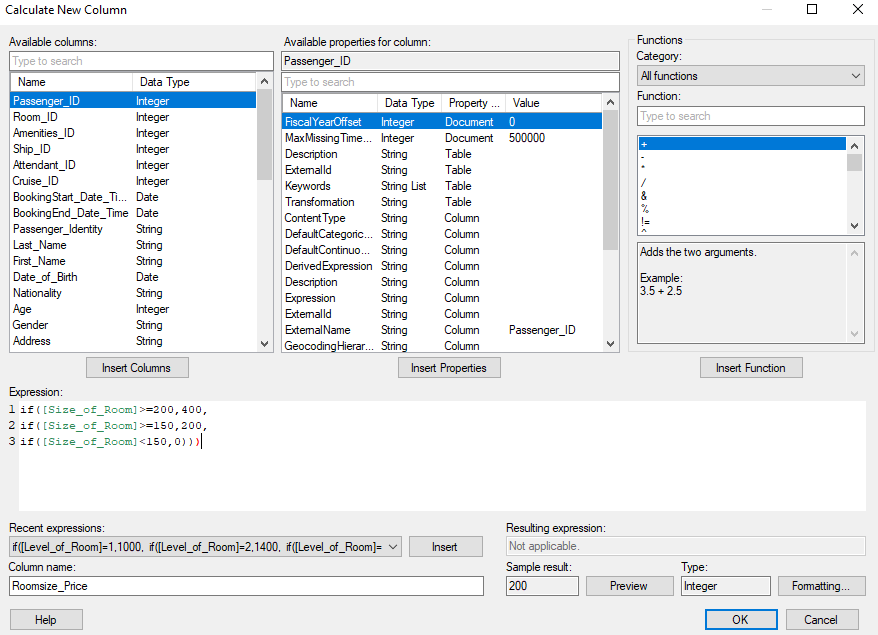


1. Column: Roomsize\_price

If Size of room < 150, 0,

If size of room >= 150, 200,

If size of room >= 200, 400

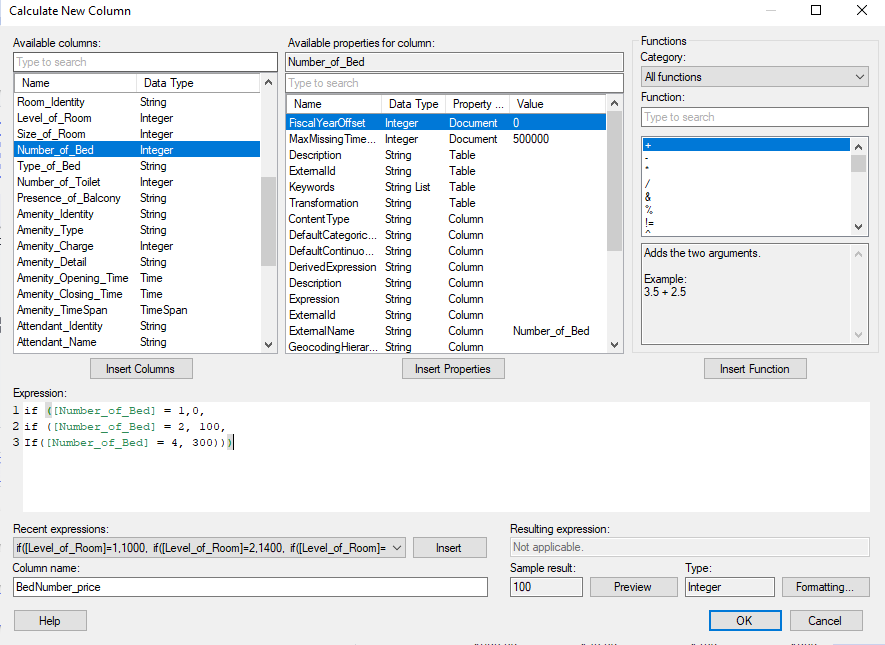


1. Column: BedNumber\_price

If Number of bed = 1, 0,

If Number of bed = 2, 100,

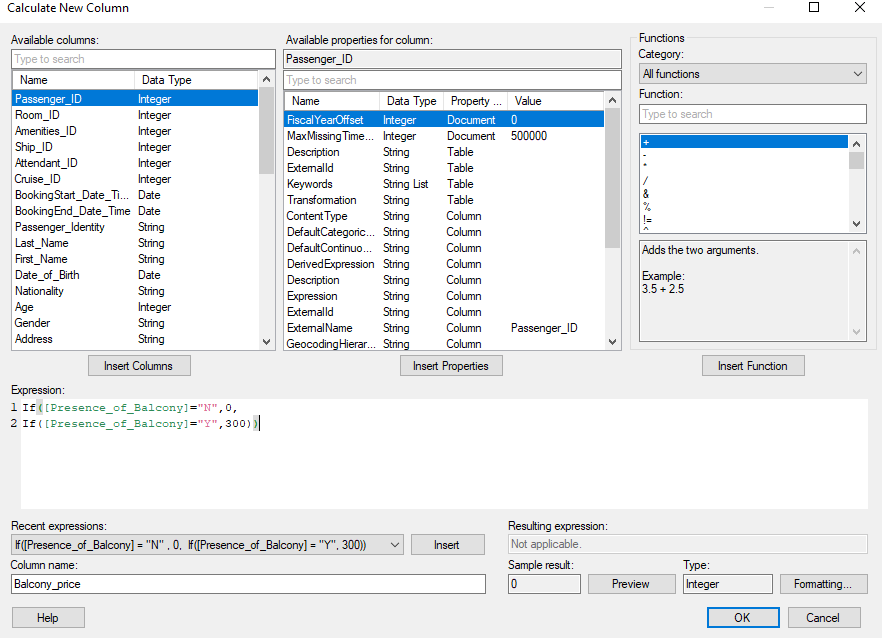
If Number of bed = 4, 300



1. Column: Balcony\_price

If presence of balcony = “N”, 0

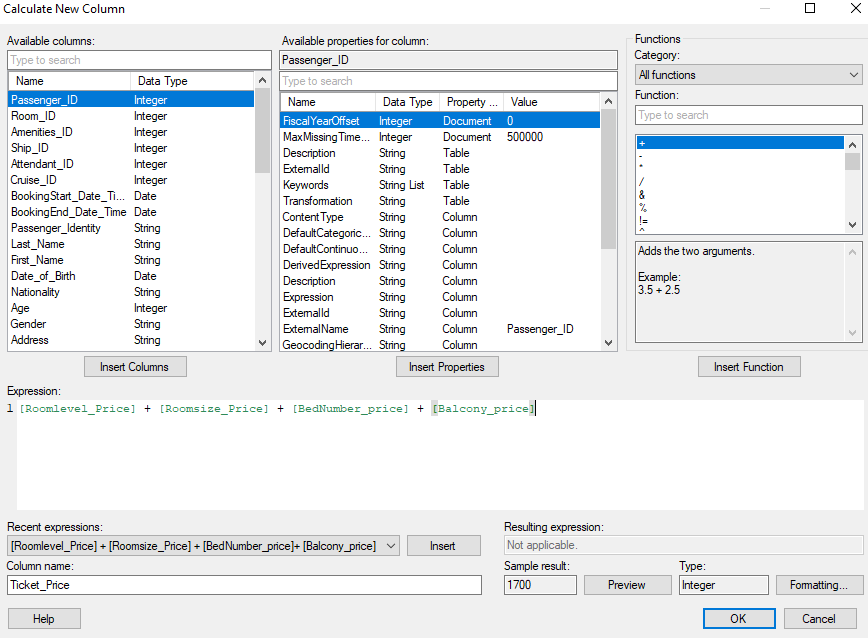
If presence of balcony = “Y”, 300

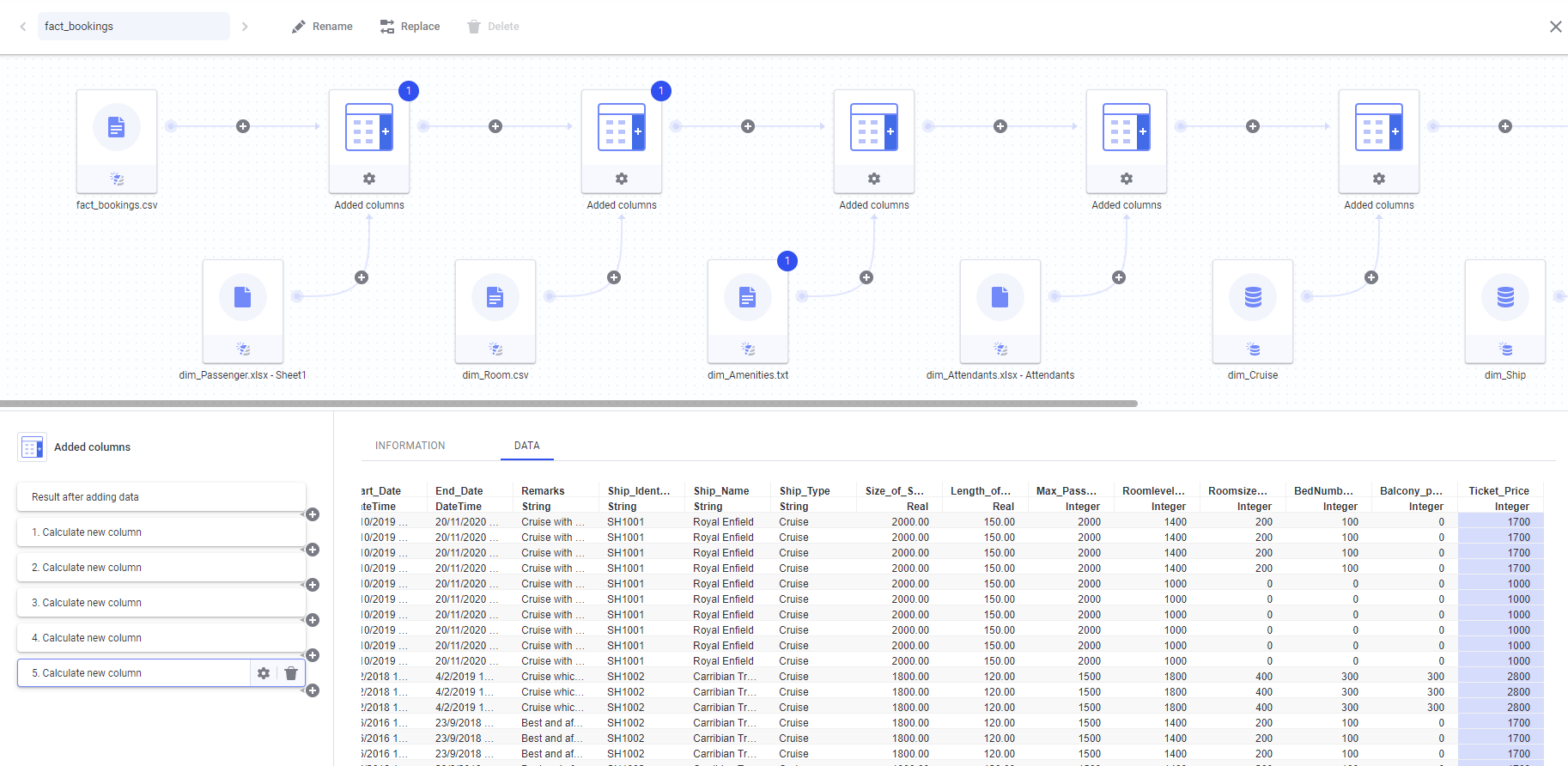


1. Column: Ticket\_Price

Finally, I will create Ticket\_Price column which adds all the price together to show us the cruise boarding ticket price of each passenger.

Roomlevel\_price + Roomsize\_price + BedNumber\_price + Balcony\_price



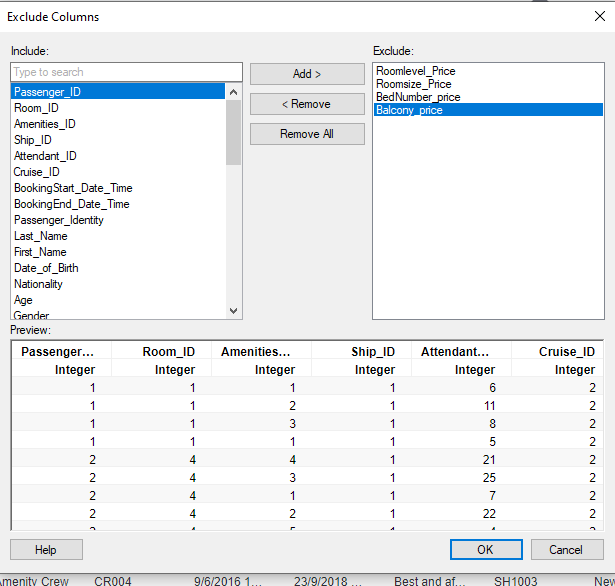


The above picture depicts all 5-transformation made to create the ticket price column.

**Transformation 4:**

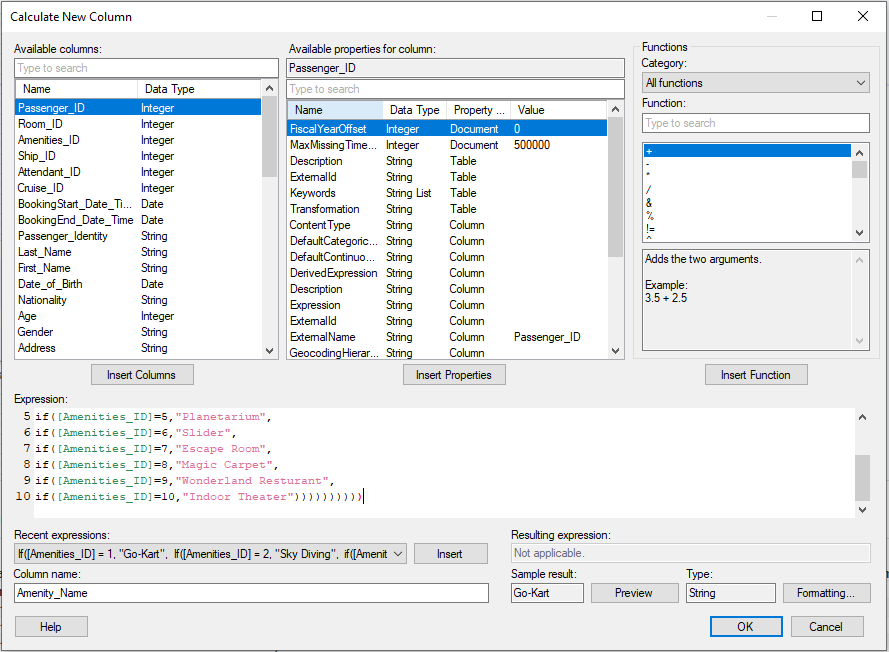
I will removeRoomlevel\_price , Roomsize\_price ,BedNumber\_price and Balcony\_price column from the table as we won’t need them later.

To execute it, I used **Exclude Columns** transformation.



**Transformation 5:**

I will create a new column called amenity\_names which contains names of all amenity on the cruise using **Calculate New Column** feature.



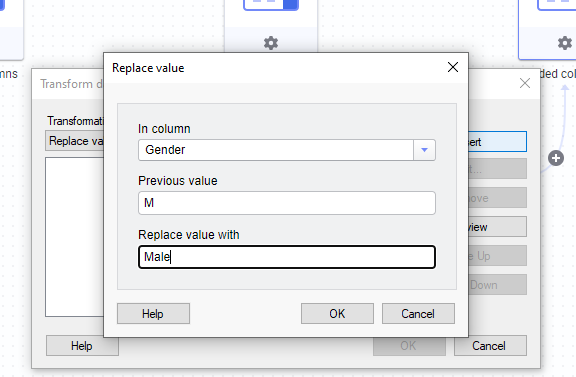


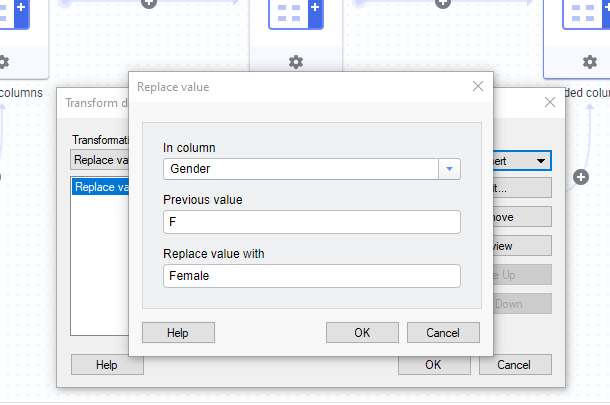
Here is our final table after transformation 5.

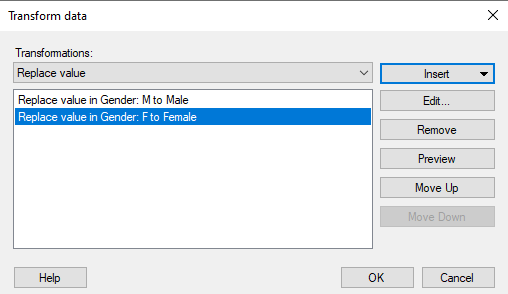
**Transformation 6:**

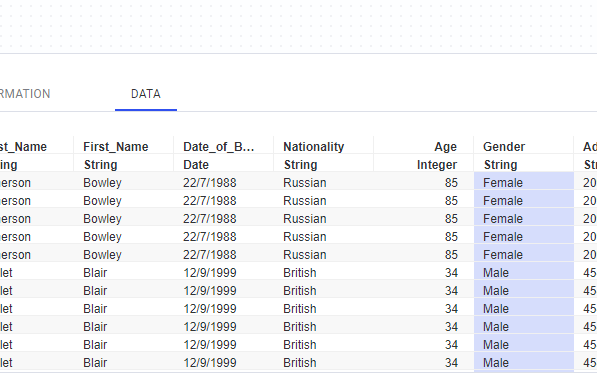
I will change values in gender column from M and F to Male and Female for better understand of my table.

I used **Replace Value** transformation in TIBCO to change the two values.









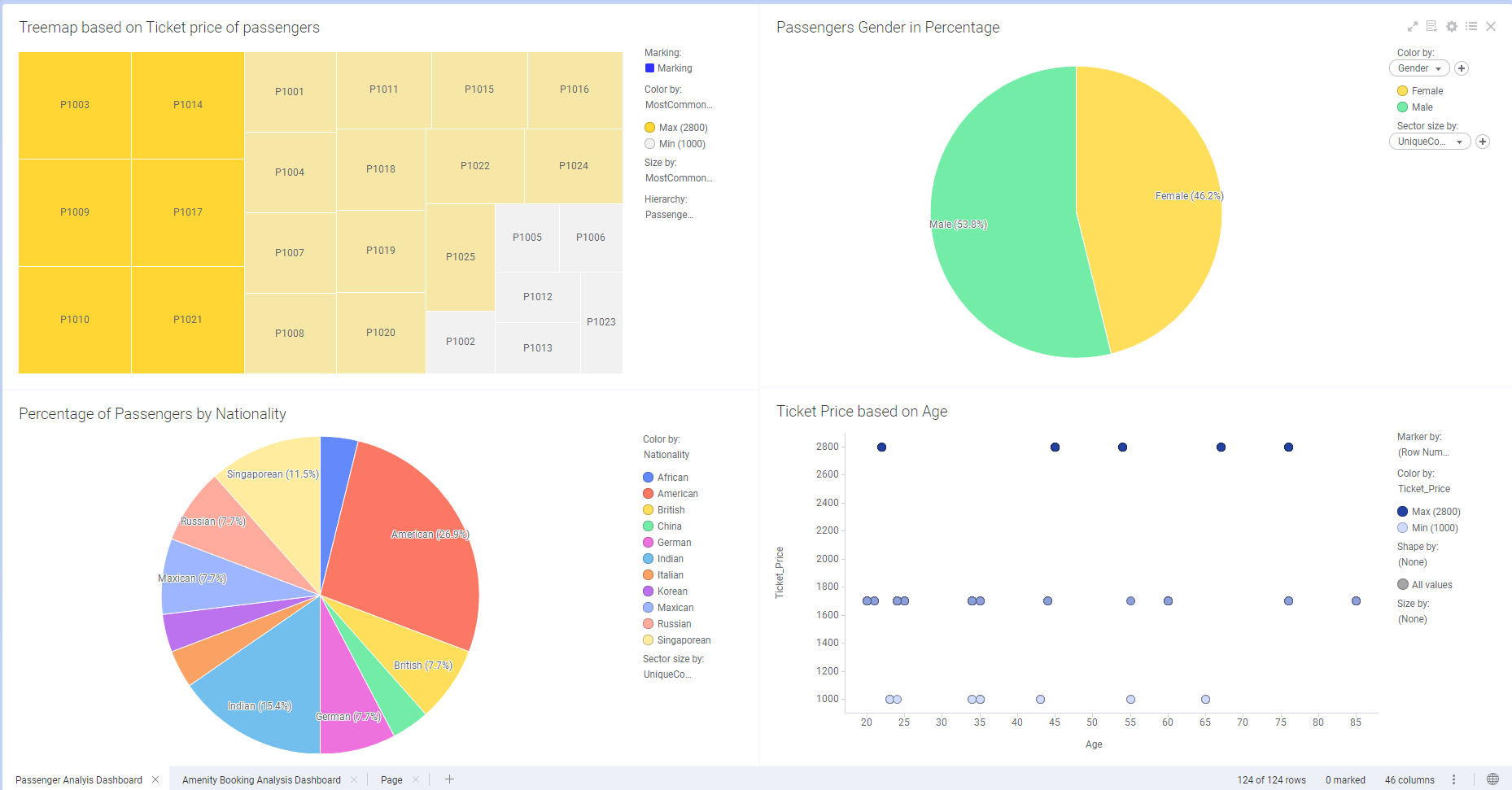
Picture above shows the final result after the transformation.

**Data Visualization:**

**Dashboard 1**

**About Dashboard:**

Dashboard below represents the booking activity and demography of passengers. Studying the dashboard will allow us to understand the passengers we serve and provide better service through data driven decision making.

****

**Chart 1:**

Using Passenger\_Identity and Ticket\_column, I have created a TreeMap chart. The size of the box represents the ticket price. We could observe that Passenger P1003, P1014, P1009, P1017, P1010 and P1021 has spent the most amount of money for the ticket. They are coloured in dark gold. We can also see that there are passengers in grey boxes who just spent $1000 on the ticket.

**Insight:**

This chart will allow Le Dot company to give better discount for the passengers based on their ticket price. Passengers who spend more then $2000 on the ticket could get discount on amenities and services so that they will spend more money.

**Chart 2:**

The second chart is a pie chart which shows the percentage of Male and Female on our cruise. I used gender and passenger\_ID column to derive at this chart.

**Chart 3:**

We also have another pie chart which shows percentage of passengers based on their nationality. It reveals that 26.9% of passenger onboard our cruise are Americans.

**Insight:**

Le Dot should put more resource for advertising and marketing in America as we have more Americans boarding our cruise.

**Chart 4:**

The fourth chart is a scatter plot. I used Ticket\_Price and Age column to observe if there is a correlation between ticket price and the age of passengers. The dots are coloured based on ticket price. Lower ticket price represents dimmer blue, higher ticket price represents brighter blue.

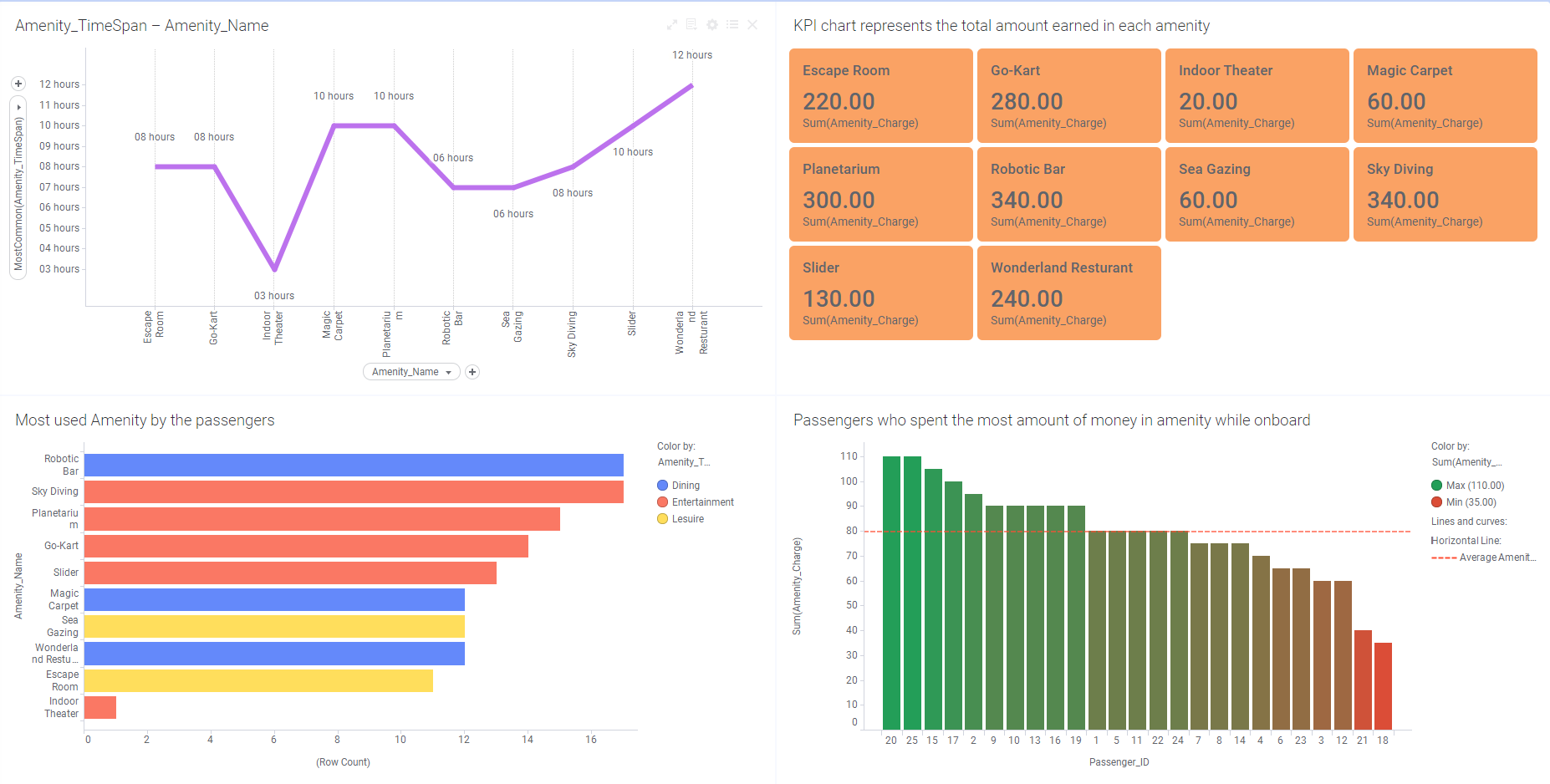
**Insight:**

The scatter plot clearly shows that older passengers tend to purchase expense tickets compared to younger passengers. Le Dot could provide enjoyable and exciting amenity and services for older passengers.

**Dashboard 2**

**About Dashboard:**

Dashboard below shows the amenity booking activity of all the passengers. It could be used to improve amenities available on cruise to make the long journey enjoyable.

****

**Chart 1:**

I have created a line chart to describe how long an amenity stayed open using Amenity\_Name and Amenity\_Timespan columns.

**Insight:**

In the line chart, we could see that Indoor theatre stayed open only for 3 hours. This could potentially be one of the reason why it earned the least amount of money.

**Chart 2:**

I have a KPI chart in my dashboard which shows total amount each amenity earned. I used Amenity\_charge and Amenity\_Name column to create this chart.

**Insight:**

From the KPI chart, we can see that Sky Diving and Robotic bar amenity earn the most amount of money. Le Dot should create amenities similar to this two on the cruise to earn more profit as passengers like them more compared to other amenities.Le Dot should also consider removing indoor theatres as it earns the least amount of money.

**Chart 3:**

Third chart in the dashboard shows which amenity does the passenger use the most and what type does it belong to. I used Amenity\_Name and Amenity\_Type column to create this chart. It is coloured based on 3 different amenity type.

**Insight:**

We could see that passengers used entertainment amenities more then leisure and dining. This tells us that there should be more amenities for entertainment in ships created in the future.

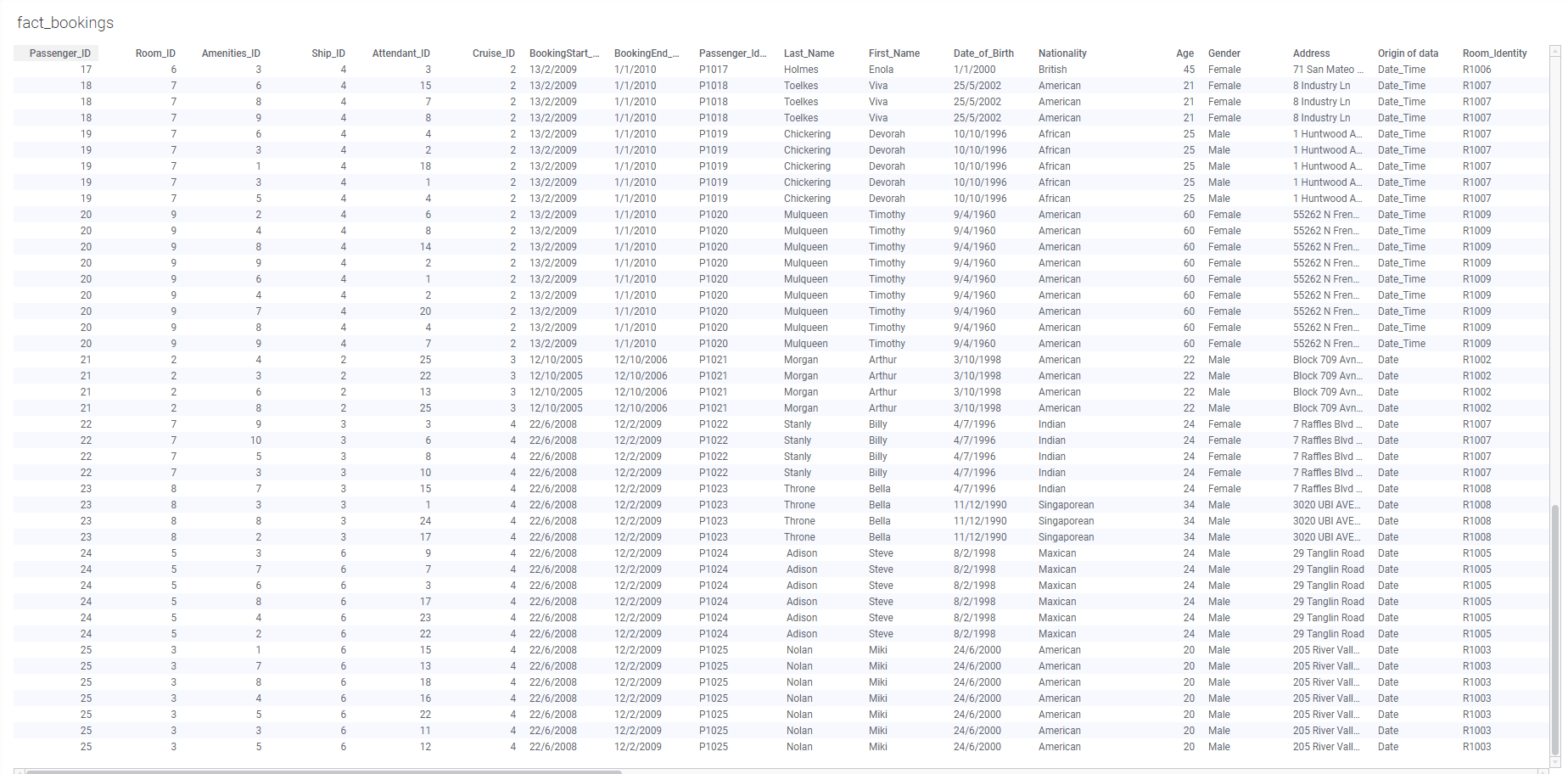
**Chart 4:**

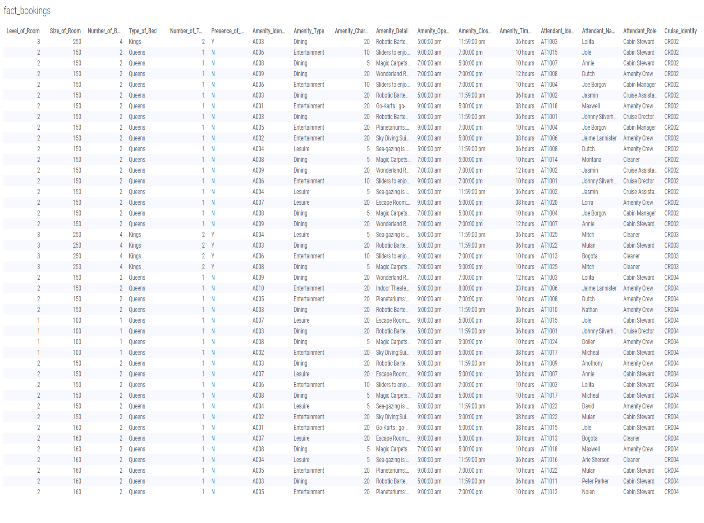
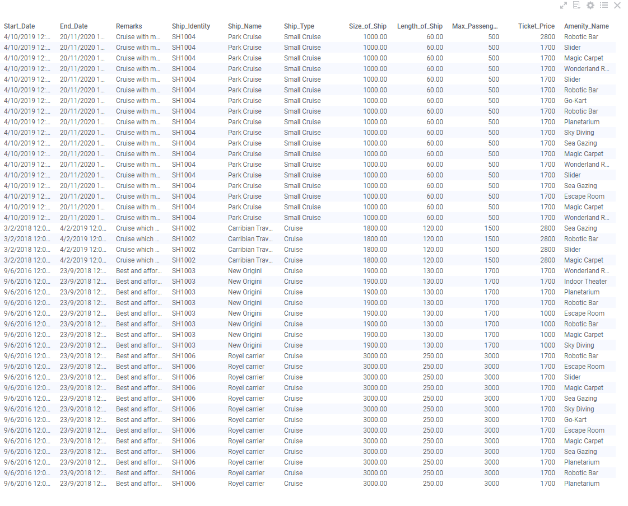
The final dashboard is a bar graph which shows the sum of money spent by each passenger on amenity. To create the chart, I used Amenity\_Charge and Passenger\_Name column. The orange line in the chart outline the overall average amount spent by all the passengers on amenity.

**Insight:**

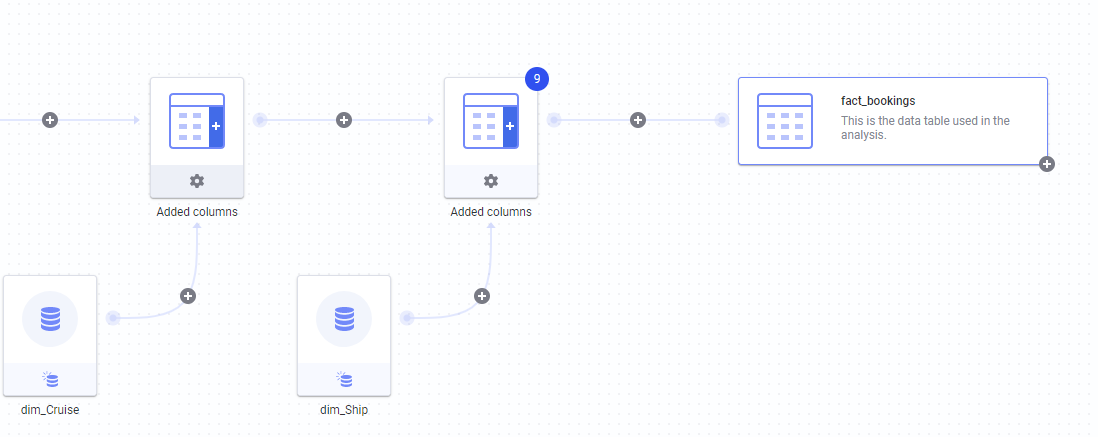
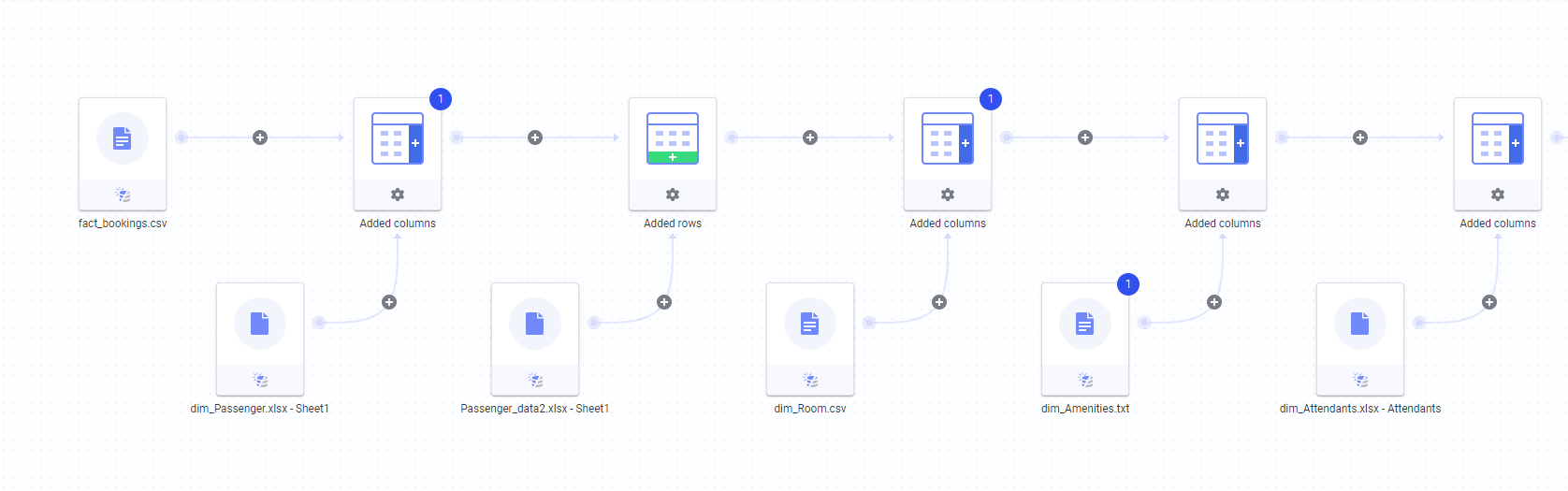
Maximum/largest amount of money on amenity was spent by passenger with id 20 and 25 and the minimum/lowest amount was spent by passenger 18. The overall average 80 tells us that amenities on the cruise are making a good profit.

**Final Table:**





**Final Data Canvas:**

****