INTRODUCTION TO DATA MINING: PROJECT PART 1

Ching-Han Kuo (r0911555)

1 Error Fixing

From the left graph below we can see that the format of storing the consumer's age is inconsistent, some are the actual age, and some seem to be the consumer's birth year. To fix this, I suggest transforming all data in this column into the format of the actual age instead of birth year. With the assumption that the age already shown in the desired format is based on the date when they arrived, I extract data in this column that have 4 digits, then use the year of booking date to deduct the extracted 4 digits and insert them back into the column.

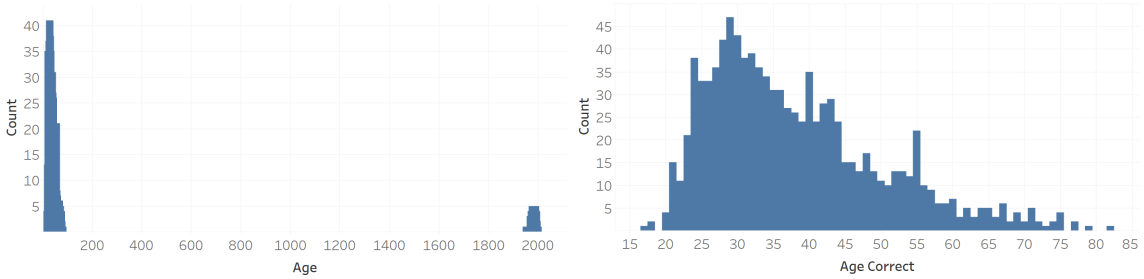

Figure 1: Histogram of Original (left) and Fixed (right) Age Column

2 Findings

2.1 Q3, the Tourist Season?

The graph below shows that Q3 (July, August, and September) can be the peak of travelling since it has the highest number of bookings (510).

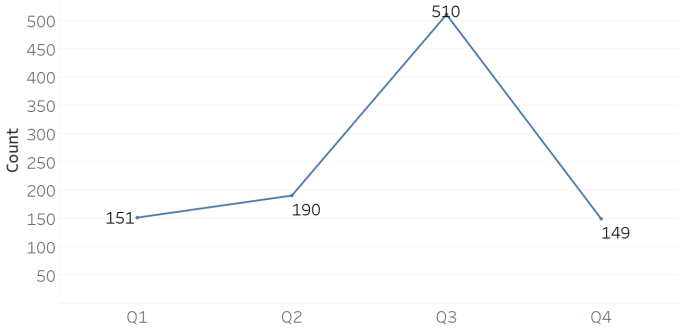

Figure 2: Number of Bookings in Different Quarters

We can also see that, for example, the spent nights and paid price per booking in Q3 are slightly higher than in other quarters. However, the difference is not statistical (from ANOVA), which may indicate that when people decide to travel, they will have almost the same behaviour no matter if they are travelling during the busy season. This can be an important message for travel service providers that you should maintain your quality even in the off-peak season because customers' anticipation is the same.

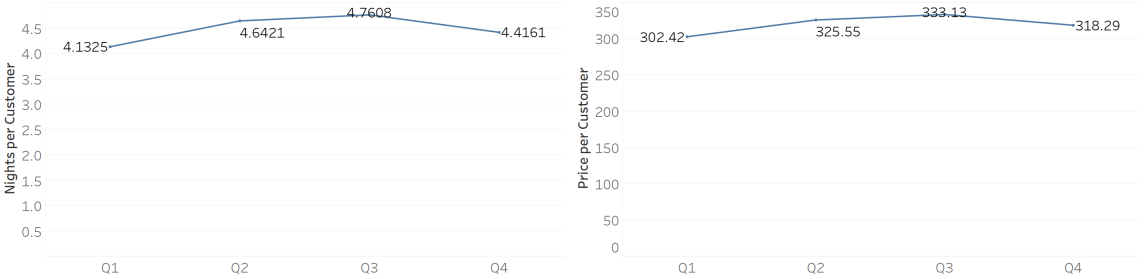

Figure 3: Spent Nights (left) and Paid Price (right) per Booking in Different Quarters

2.2 Short-Period Group Tour, A Popular Traveling Style?

From the graph below, it is obvious that the majority (89.4%) of bookings are for a double room, and almost half (47.2%) of the bookings are under 4 nights. A simple way to put this is that a short-period trip with friends is popular. However, there could be other possibilities such as 1) consumers tend to keep changing hotels during a long-period trip, or 2) consumers who travel alone would also prefer a double room. We don't have detailed information in this dataset to investigate further.

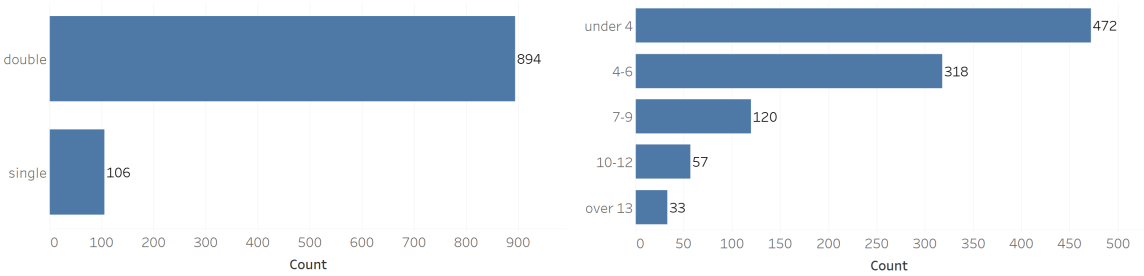

Figure 4: Number of Bookings of Room Type (left) and Duration (right)

2.3 France, An Ideal Target Country for Off-Peak Season?

From the graph in section 2-1, we may claim that Q1 is the off-peak travelling season because it has the lowest number of bookings (151), and its spent nights (4.13) and paid price (302.42) per booking are also the lowest. However, looking at the destination level, if we compute the proportion of quarterly bookings within each destination and compare the difference between destinations (proportional z-test), we can see that the proportion of Q1 bookings to France is significantly higher than in other countries. This may suggest that France is more attractive to customers travelling in the off-peak season, and travel agencies can promote travel packages in France more during this period.

Table 1: Cross Table (%) of Destination and Quarter

	Total	Belgium	France	Italy	Netherlands	Portugal	Spain
n	*1000*	*95*	*163*	*201*	*89*	*198*	*254*
Q1	15.10	13.68	22.09	11.44	10.11	17.17	14.17
Q2	19.00	22.11	14.11	22.39	20.22	17.17	19.29
Q3	51.00	46.32	53.37	54.23	50.56	50.51	49.21
Q4	14.90	17.89	10.43	11.94	19.10	15.15	17.32

Another interesting point of France being the destination is that customers tend to arrive there on weekdays more compare to other countries (significantly higher in proportional z-test). This may also be a demonstration of France being a good target in off-peak periods because weekdays are considered off-peak for travelling. However, there can be another explanation for this there are more business trips to France, and the on and off peaks of business trips can be different from tourist trips. We can't identify if it's a business trip or not from this dataset.

Table 2: Cross Table (%) of Destination and Arrival Day

	Total	Belgium	France	Italy	Netherlands	Portugal	Spain
n	*1000*	*95*	*163*	*201*	*89*	*198*	*254*
Weekday	72.70	65.26	84.66	73.63	73.03	68.18	70.47
Weekend	27.30	37.74	15.34	26.37	26.97	31.82	29.53

2.4 The Older the Customer, the Higher Their Value?

Although the most of bookings in this dataset are from customers under 40 (60.9 %), it does not necessarily mean that focusing on a younger target audience is more profitable. If we look at the conduct linear regression between age and paid price per booking, and between age and paid price per night, we can see that paid price gets higher when age is older. Both linear relations achieve statistical significance (p-values are 0.03 and 0.01, respectively).

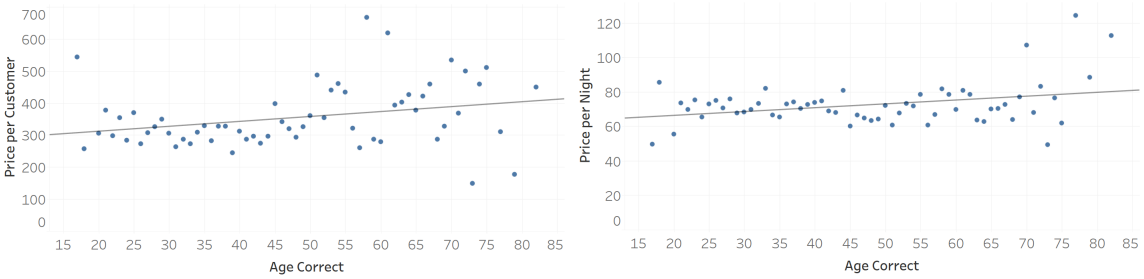

Figure 5: Relations of Age and Paid Price per Customer (left)/per Nights (right)

Furthermore, comparing the booking behaviour between age groups, age groups older than 50 have significantly higher proportions (z-test) of booking a resort and longer staying nights.

Table 3: Cross Table (%) of Age and Room Type/Stay Nights

	Total	Under 20	20s	30s	40s	50s	60s	70 Older
n	*1000*	*3**	*280*	*329*	*213*	*112*	*43*	*20**
City	74.4	66.67	79.29	83.59	75.59	44.64	51.16	60.00
Resort	25.6	33.33	20.71	16.41	24.41	55.36	48.84	40.00
1-3 Nights	47.2	33.33	47.14	53.50	46.01	38.39	30.23	45.00
4-6 Nights	31.8	33.33	31.79	30.09	35.68	26.79	39.53	30.00
7-9 Nights	12	0.00	12.50	11.25	11.27	13.39	16.28	10.00
10-12 Nights	5.7	33.33	5.36	3.65	5.63	10.71	11.63	0.00
12 Nights Longer	3.3	0.00	3.21	1.52	1.41	10.71	2.33	15.00

* Sample size is smaller than 30, the result may not be representative enough

The observations above can be evidence that older customers can be more willing to pay for a premium service which can be more profitable to travel agencies. However, since the dataset only stores the information of the booker him/herself, we are not sure if it is like, for instance, older customers tend to travel with families and what they book is for a larger group of people which makes it more expensive.

2.5 Different Target Audiences for Different Destinations?

This dataset has six destinations: Belgium, France, Italy, Netherlands, Portugal and Spain. We can categorise them into two groups, Southern Europe and Western Europe. Generally speaking, this dataset has more bookings travelling to Southern Europe. Nevertheless, it seems that different genders target different destinations. From the result of the proportional z-test, Southern Europe is significantly more attractive to females, and Western Europe is significantly more attractive to males.

Table 4: Cross Table (%) of Region and Gender

	Total	Southern EU	Western EU
n	*1000*	*653*	*347*
?	39.5	39.36	39.77
f	30.5	33.69	24.5
m	29.5	26.49	35.16
x	0.5	0.46	0.58

This may suggest that travel agencies can promote different destinations according to customers' gender. However, As mentioned in the previous section, this dataset only has the information of the person who is in charge of the booking. The situation may be different once we know the gender of people they are travelling with. In addition, most of the gender bookers were recorded as "?" (39.5%) and the result might change if we know the exact gender of the bookers recorded as gender "?".