Name: 林成鍇 Julius Ling

Student ID: D84085360

Department: Psychology 112

Homework 2 – Exploratory Factor Analysis (EFA)

*** This assignment is performed with *Python* coding. For detailed display in IPython Notebook, click the link below.

https://colab.research.google.com/drive/1cQ3wOmp9zWgWaF2XUKM71qfREz8npTry?usp=sharing

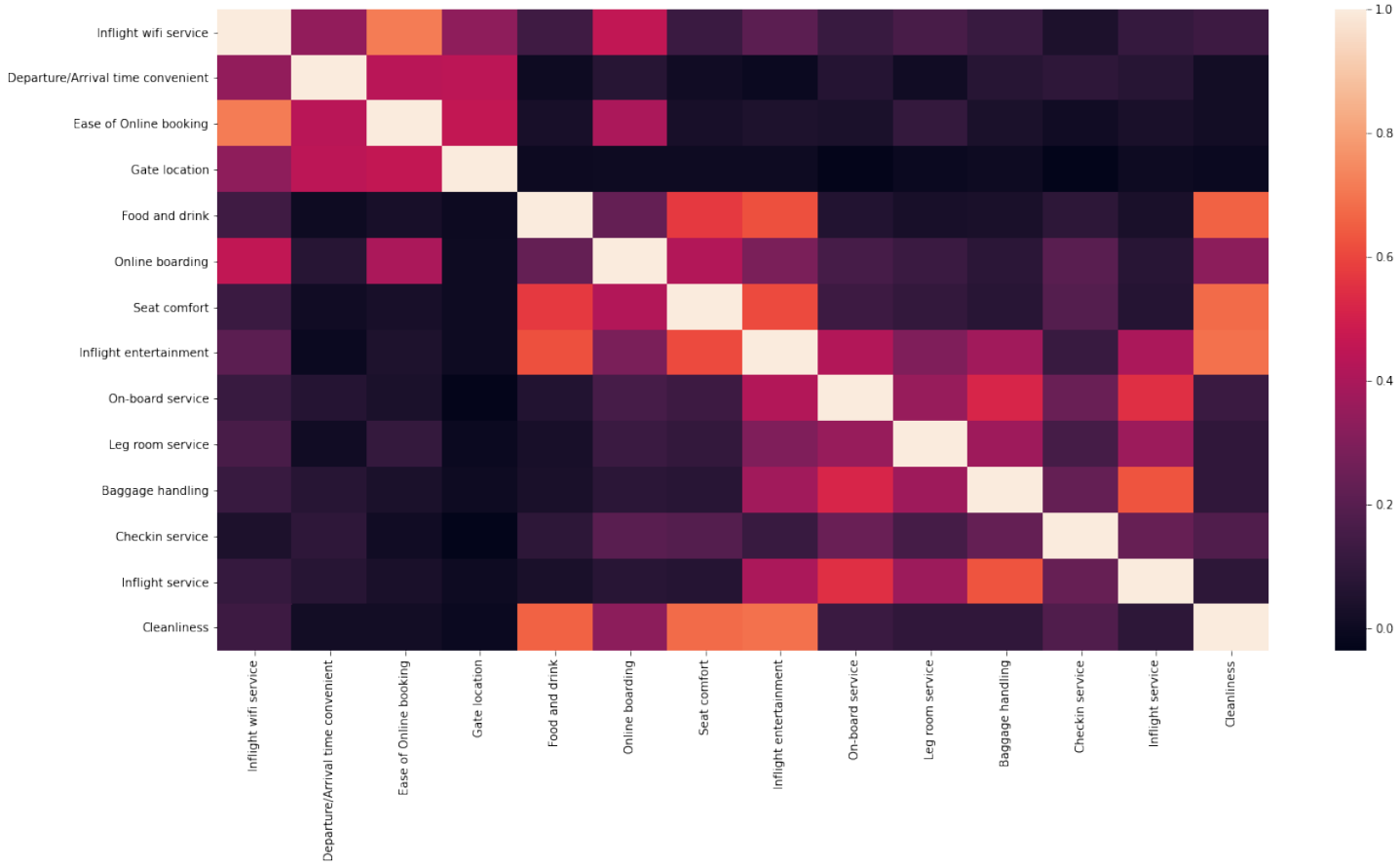
Dataset: Airline Passenger Satisfaction

This dataset contains an airline passenger satisfaction survey with the variables of interest shown as below:

- Gender: Gender of the passengers (Female, Male)
- Customer Type: The customer type (Loyal customer, disloyal customer)
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- Flight distance: The flight distance of this journey
- Inflight wifi service: Satisfaction level of the inflight wifi service (0: Not Applicable;1-5)
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of Gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of Seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment
- On-board service: Satisfaction level of On-board service
- Leg room service: Satisfaction level of Leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of Check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure
- Arrival Delay in Minutes: Minutes delayed when Arrival
- Satisfaction: Airline satisfaction level (Satisfaction, neutral or dissatisfaction)

It has 103,904 observations and 25 columns, with 14 of those representing customers responses, on a scale of 1 to 5, to a survey evaluating different aspects of the flights (Inflight wifi service, food and drink, online boarding, seat comfort, etc). These 14 columns will be very important for our upcoming factor analysis.

The Correlation Heatmap of the 14 variables is plotted to skim through the inner correlation structure:



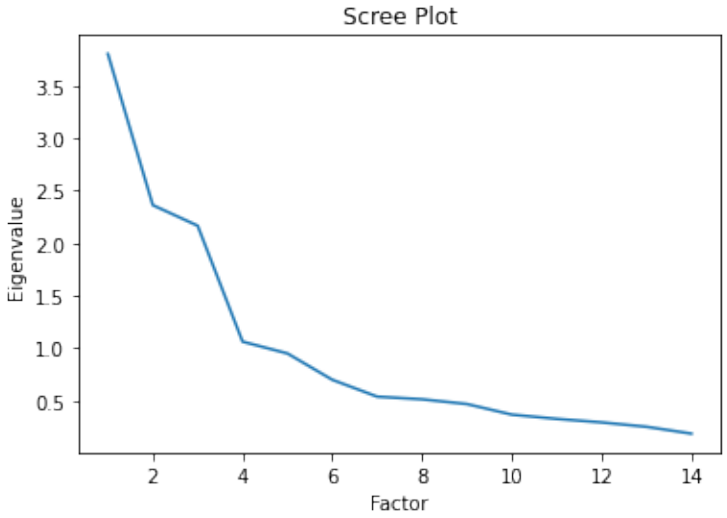
Visually, we can see certain "clusters" within the correlation heatmap, implying the existence of the inner structure.

The Kaiser-Meyer-Olkin Test is conducted and the sum of all KMO statistics is

`0.7812294257164448`

revealing a strong multicollinearity between the variables consistent to the result of the Bartlett's test of sphericity:

`(601690.8930479775, 0.0)` with respect to Chi-square value & P-value



The Scree Plot shows that the eigenvalue drops drastically after the 3rd factor. Till the 4th factor the eigenvalue are bigger than 1. Hence 3 or 4 factors would be a better choice.

Here's the Factor Loading and the Total Communalities with respect to each variable:

4 factors under varimax (orthogonal) rotation (loading = 0.5 as a cutoff point):

	FACTOR_1	FACTOR_2	FACTOR_3	FACTOR_4	TOTAL_COMMUNALITIES
INFLIGHT WIFI SERVICE	-0.01794	0.041272	0.49975	0.464001	0.467072
DEPARTURE/ARRIVAL TIME CONVENIENT	0.016234	0.054087	0.606817	-0.06873	0.37614
EASE OF ONLINE BOOKING	-0.12867	-0.04902	0.665939	0.450739	0.6656
GATE LOCATION	0.08261	-0.04143	0.730367	-0.18767	0.577198
FOOD AND DRINK	0.810338	-0.12042	0.05792	-0.03273	0.675576
ONLINE BOARDING	0.088286	-0.0413	-0.09248	0.823187	0.695689
SEAT COMFORT	0.732636	-0.06542	-0.05332	0.169248	0.572522
INFLIGHT ENTERTAINMENT	0.747024	0.367021	0.058879	-0.08036	0.702672
ON-BOARD SERVICE	-0.0134	0.715014	-0.01318	-0.0086	0.511672
LEG ROOM SERVICE	-0.02766	0.483435	0.008883	0.061796	0.238372
BAGGAGE HANDLING	-0.05189	0.801769	0.043494	-0.10795	0.659072
CHECK IN SERVICE	0.048957	0.262046	-0.06176	0.119487	0.089156
INFLIGHT SERVICE	-0.05183	0.843393	0.048264	-0.13509	0.734578
CLEANLINESS	0.870436	-0.05918	0.009305	0.028748	0.762073
SS_LOADING	2.549586	2.33521	1.621478	1.22112	
PROPORTIONAL_VARIANCE	0.182113	0.166801	0.11582	0.087223	
CUMULATIVE_VARIANCE	0.182113	0.348914	0.464734	0.551957	

It is notable that Factor 3 & 4 possesses certain "ambiguity" (in terms of cross-loading). It's quite intriguing as it seems reasonable to me that *Inflight wifi service, Ease of Online booking, Online boarding* tell their own story as well (in terms of "electrical convenience").

Note that the cumulative variance goes up to 0.55, not bad.

The overall total communalities are good, except the *Leg room service* and *Check in service.*

I'm trying out 3-factor analysis for further exploration.

3 factors under varimax (orthogonal) rotation (loading = 0.5 as a cutoff point):

	FACTOR_1	FACTOR_2	FACTOR_3	TOTAL_COMMUNALITIES
INFLIGHT WIFI SERVICE	0.16827	0.128271	0.758091	0.619471
DEPARTURE/ARRIVAL TIME CONVENIENT	-0.02951	0.059681	0.501384	0.255818
EASE OF ONLINE BOOKING	0.030231	0.020914	0.932775	0.871421
GATE LOCATION	-0.03383	-0.03231	0.504044	0.256249
FOOD AND DRINK	0.752639	0.010946	0.006167	0.566623
ONLINE BOARDING	0.395453	0.113811	0.359065	0.298264
SEAT COMFORT	0.78999	0.081463	0.027258	0.631464
INFLIGHT ENTERTAINMENT	0.745693	0.46675	0.012034	0.774059
ON-BOARD SERVICE	0.093881	0.701154	0.029009	0.501272
LEG ROOM SERVICE	0.074455	0.481442	0.08065	0.243834
BAGGAGE HANDLING	0.023463	0.764748	0.027693	0.586157
CHECK IN SERVICE	0.143512	0.284182	0.028882	0.102189
INFLIGHT SERVICE	0.018131	0.799771	0.018252	0.640295
CLEANLINESS	0.85842	0.088148	-0.00171	0.744659
SS_LOADING	2.706954	2.295467	2.089355	
PROPORTIONAL_VARIANCE	0.193354	0.163962	0.14924	
CUMULATIVE_VARIANCE	0.193354	0.357316	0.506555	

3-factor analysis makes things clearer as each variable (except *Online boarding, Leg room service, Check in service*) has only a relative large loading on respective factors, it's quite nicely cut. It seems that I'm able to endow a concept to respective factors:

1) Comfort
2) Service
3) Convenience

However, it's still worth regurgitating to me that the ambiguity of *Inflight entertainment* and *Leg room service* which seems to have an effect marginally on the 2nd factor. The *Inflight entertainment* may be assumed as a kind of services but more to a comfort issue.

The cumulative variance in this case has dropped to about 0.51. However, each factor has more than 10% of variance "explained". I would assume 3-factor analysis as a better choice in this case.

3 factors under promax (oblique) rotation (loading = 0.5 as a cutoff point):

	FACTOR_1	FACTOR_2	FACTOR_3	TOTAL_COMMUNALITIES
INFLIGHT WIFI SERVICE	0.096712	0.055555	0.750116	0.575114
DEPARTURE/ARRIVAL TIME CONVENIENT	-0.08108	0.037162	0.509249	0.26729
EASE OF ONLINE BOOKING	-0.04458	-0.04223	0.946036	0.898755
GATE LOCATION	-0.06966	-0.05923	0.518053	0.276739
FOOD AND DRINK	0.796638	-0.13293	-0.05625	0.655467
ONLINE BOARDING	0.371677	0.021834	0.326031	0.244917
SEAT COMFORT	0.82221	-0.06701	-0.04224	0.682304
INFLIGHT ENTERTAINMENT	0.708443	0.348884	-0.07815	0.629719
ON-BOARD SERVICE	-0.02622	0.719897	-0.02193	0.519421
LEG ROOM SERVICE	-0.01215	0.48859	0.045962	0.24098
BAGGAGE HANDLING	-0.11209	0.800551	-0.02145	0.653907
CHECKIN SERVICE	0.099999	0.270514	-0.00012	0.083177
INFLIGHT SERVICE	-0.12318	0.839112	-0.03281	0.72036
CLEANLINESS	0.895947	-0.07125	-0.07778	0.813847
SS_LOADING	2.814757	2.334335	2.112904	
PROPORTIONAL_VARIANCE	0.201054	0.166738	0.150922	
CUMULATIVE_VARIANCE	0.201054	0.367792	0.518714	

At the end the oblique rotation method is applied to optimize the result.

With Loading = 0.5 as a cutoff point, we have the three factors nicely created:

1) Comfort: *Food and Drink, Seat Comfort, Inflight Entertainment, Cleanliness*

2) Service: *On-board Service, Baggage Handling, Inflight Service*

3) Convenience: *Inflight Wifi Service, Departure/Arrival Time Convenient, Ease of Online Booking, Gate Location*

3 variables are left out: *Online Boarding, Leg Room Service, Check In Service* (Note that their respective communalities are minute as well)

And the final results is presented,

3 factors under promax (oblique) rotation (loading = 0.5 as a cutoff point):

	COMFORT	CONVENIENCE	SERVICE	TOTAL_COMMUNALITIES
INFLIGHT WIFI SERVICE	0.116614	0.715414	0.060798	0.529112
DEPARTURE/ARRIVAL TIME CONVENIENT	-0.04504	0.543998	0.035553	0.299226
EASE OF ONLINE BOOKING	-0.01174	0.89807	-0.03511	0.8079
GATE LOCATION	-0.02918	0.562047	-0.05571	0.319852
FOOD AND DRINK	0.792072	0.007425	-0.11396	0.640419
SEAT COMFORT	0.774135	0.000262	-0.05337	0.602132
INFLIGHT ENTERTAINMENT	0.747417	-0.00681	0.364857	0.691799
ON-BOARD SERVICE	0.022285	-0.0019	0.686404	0.471651
BAGGAGE HANDLING	-0.05069	0.0049	0.772885	0.599945
INFLIGHT SERVICE	-0.06249	-0.00506	0.826092	0.68636
CLEANLINESS	0.876884	-0.01212	-0.05935	0.772595
SS_LOADING	2.577808	1.930478	1.912705	
PROPORTIONAL_VARIANCE	0.234346	0.175498	0.173882	
CUMULATIVE_VARIANCE	0.234346	0.409844	0.583726	

Though the total communalities of *Departure/Arrival time convenient* and *Gate location* are below 0.3, I think they are still meant to be preserved as their loading on the 3rd factor are quite big, big enough to lay my eyes on them. I think it's reasonable to cluster them under Factor 3.

Cronbach's alpha of Factor 1 = `0.8762877916624099`

Cronbach's alpha of Factor 2 = `0.7942916933090214`

Cronbach's alpha of Factor 3 = `0.7679754211110683`

These results confirm the reliability of these factors. I'm prone to adopt that these underlying factors could have dictating the overall satisfaction of a passenger.

With that, I rest my case.

**** With the prior knowledge of such factor structure, I conduct a confirmatory factor analysis in *Homework 5 – Confirmatory Factor Analysis*. The solution of EFA doesn't necessarily converge with the CFA's.