Capstone Project Report

Exploratory Factor Analysis

Name: Polluri Kodanda Rama Durgarao
Course: AI & ML (Batch - 4)

Problem Statement
Use the Airline Passenger Satisfaction dataset to perform factor analysis. (Use only the columns that represent the ratings given by the passengers, only 14 columns). Choose the best features possible that helps in dimensionality reduction, without much loss in information.

Prerequisites
Along with Python below packages needed to be installed

Pandas
Seaborn
Sklearn

Dataset Used
Airline Passenger Satisfaction dataset
https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

Implementation
Import required libraries and load data

```
In [1]: import pandas as pd
        from sklearn.decomposition import FactorAnalysis
        import seaborn as sns
```

```
In [2]: # Load train and test datasets
        train = pd.read_csv('train.csv')
        test = pd.read_csv('test.csv')
```

Explore data

In [3]:
```
train.head()
```

Out[3]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service
0	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	...	5	4	3	4	4
1	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	...	1	1	5	3	1
2	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	...	5	4	3	4	4
3	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	...	2	2	5	3	1
4	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	...	3	3	4	4	3

5 rows × 25 columns

In [4]:
```
train.columns
```

Out[4]:
```
Index(['Unnamed: 0', 'id', 'Gender', 'Customer Type', 'Age', 'Type of Travel',
       'Class', 'Flight Distance', 'Inflight wifi service',
       'Departure/Arrival time convenient', 'Ease of Online booking',
       'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort',
       'Inflight entertainment', 'On-board service', 'Leg room service',
       'Baggage handling', 'Checkin service', 'Inflight service',
       'Cleanliness', 'Departure Delay in Minutes', 'Arrival Delay in Minutes',
       'satisfaction'],
      dtype='object')
```

Filter required fields for analysis

In [5]:
```
# Filtering out features other than rating features
filtered_train = train[train.columns[8:21]]
filtered_test = test[test.columns[8:21]]
```

In [6]:
```
filtered_train[1:5]
```

Out[6]:

	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service
1	3	2	3	3	1	3	1	1	1	5	3	1	4
2	2	2	2	2	5	5	5	5	4	3	4	4	4
3	2	5	5	5	2	2	2	2	2	5	3	1	4
4	3	3	3	3	4	5	5	3	3	4	4	3	3

In [7]:
```
#check number of dimensions of test and train datasets
print(filtered_train.shape)
print(filtered_test.shape)
```

```
(103904, 13)
(25976, 13)
```

Applying EFA on the data for getting 3 latent factor loadings

```
In [23]: # Fit and transform EFA on train data set
         efa = FactorAnalysis(n_components=3, random_state=0)
         filtered_train_efa = efa.fit(filtered_train)
```

```
In [25]: t(filtered_train_efa.components_)
         pd.DataFrame(filtered_train_efa.components_.T, index=filtered_train.columns, columns=['Factor1', 'Factor2', 'Factor3'])
```

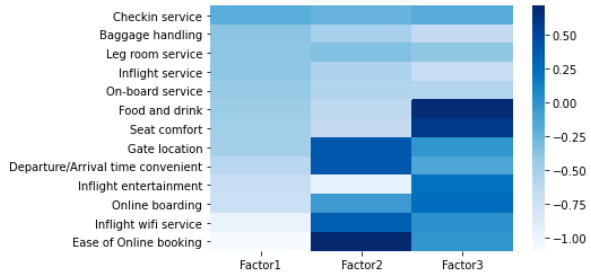
Out[25]:

	Factor1	Factor2	Factor3
Inflight wifi service	-0.986934	0.349598	0.021366
Departure/Arrival time convenient	-0.600355	0.394448	-0.123652
Ease of Online booking	-1.098992	0.712383	-0.005985
Gate location	-0.487074	0.386288	-0.006124
Food and drink	-0.459062	-0.627094	0.686846
Online boarding	-0.714322	-0.045327	0.242004
Seat comfort	-0.482956	-0.645300	0.600262
Inflight entertainment	-0.696703	-0.952546	0.223363
On-board service	-0.426483	-0.559507	-0.566329
Leg room service	-0.390836	-0.339461	-0.393837
Baggage handling	-0.382065	-0.502087	-0.650893
Checkin service	-0.193120	-0.246380	-0.181144
Inflight service	-0.391158	-0.531888	-0.682489

Factor loading 1 says that Check In, Baggage handling, Leg room service and Inflight service are correlated

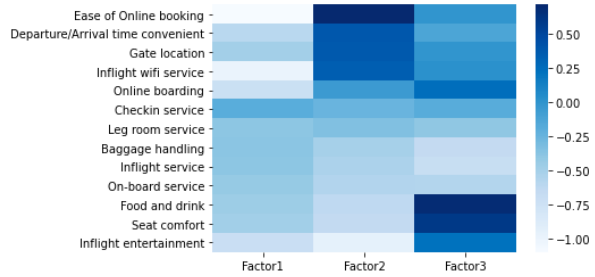
```
In [44]: sns.heatmap(df.sort_values('Factor1', axis=0, ascending=False, inplace=False), cmap="Blues")
```

Out[44]: <AxesSubplot:>



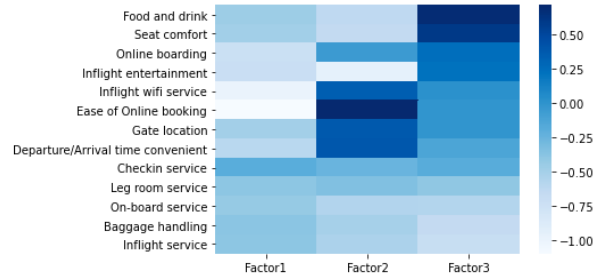
Factor loading 2 says that Ease of Online booking, Departure/Arrival time convenient, Gate location, Online boarding etc are correlated.

```
In [45]: sns.heatmap(df.sort_values('Factor2', axis=0, ascending=False, inplace=False), cmap="Blues")
Out[45]: <AxesSubplot:>
```



Factor loading 3 says that **Food and drink, Seat comfort, Online boarding, Inflight entertainment, Inflight wifi service** etc are correlated.

```
In [46]: sns.heatmap(df.sort_values('Factor3', axis=0, ascending=False, inplace=False), cmap="Blues")
Out[46]: <AxesSubplot:>
```



Conclusion

Factor 1 focus more on the external services of the flight.
Factor 2 focus more on the convenience journey of passenger to get a flight.
Factor 3 focus more on the internal services of the flight.