United Airlines SkyHack Hackathon

Importing Basic Libraries

In [1]:
```
import numpy as np
import pandas as pd
```

Importing Datasets

In [2]:
```
df_satisfaction = pd.read_csv('Datasets/Survey data_Inflight Satisfaction Score.csv')
df_satisfaction.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 47074 entries, 0 to 47073
Data columns (total 31 columns):
 #   Column                           Non-Null Count  Dtype 
---  ------                           --------------  ----- 
 0   flight_number                    47074 non-null  int64 
 1   origin_station_code              47074 non-null  object
 2   destination_station_code         47074 non-null  object
 3   record_locator                   47074 non-null  object
 4   scheduled_departure_date         47074 non-null  object
 5   question_text                    47074 non-null  object
 6   score                            47074 non-null  object
 7   satisfaction_type                34963 non-null  object
 8   driver_sub_group1                47074 non-null  object
 9   driver_sub_group2                47074 non-null  object
 10  arrival_delay_minutes            47074 non-null  int64 
 11  arrival_delay_group              47074 non-null  object
 12  cabin_code_desc                  47074 non-null  object
 13  cabin_name                       27094 non-null  object
 14  entity                           47071 non-null  object
 15  number_of_legs                   47074 non-null  int64 
 16  seat_factor_band                 47074 non-null  object
 17  loyalty_program_level            35458 non-null  object
 18  generation                       47074 non-null  object
 19  fleet_type_description           47074 non-null  object
 20  fleet_usage                      47074 non-null  object
 21  equipment_type_code              47074 non-null  object
 22  ua_uax                           47074 non-null  object
 23  actual_flown_miles               47074 non-null  int64 
 24  haul_type                        47074 non-null  object
 25  departure_gate                   46977 non-null  object
 26  arrival_gate                     46547 non-null  object
 27  international_domestic_indicator 47074 non-null  object
 28  response_group                   47074 non-null  object
 29  media_provider                   45535 non-null  object
 30  hub_spoke                        47074 non-null  object
dtypes: int64(4), object(27)
memory usage: 11.1+ MB
```

In [3]:
```
df_satisfaction.iloc[0]
```

Out[3]:

```
flight_number                                                    3802
origin_station_code                                               MKX
destination_station_code                                          ORX
record_locator                                                 CYXXJJ
scheduled_departure_date                                     9/1/2022
question_text                 How satisfied were you with the food & beverag...
score                                                               2
satisfaction_type                                         Dissatisfied
driver_sub_group1                                          food & beverage
driver_sub_group2                        food and beverage satisfaction
arrival_delay_minutes                                             -24
arrival_delay_group                                   Early & Ontime
cabin_code_desc                                                Economy
cabin_name                                                     Economy
entity                                                        Domestic
number_of_legs                                                       2
seat_factor_band                                                   80+
loyalty_program_level                                              NaN
generation                                                       Gen X
fleet_type_description                                         CRJ-200
fleet_usage                                                    Express
equipment_type_code                                                CRZ
ua_uax                                                             UAX
actual_flown_miles                                                  67
haul_type                                                       Short
departure_gate                                                    C12
arrival_gate                                                      F10
international_domestic_indicator                              Domestic
response_group                                              non-member
media_provider                                                     NaN
hub_spoke                                             spoke departure
Name: 0, dtype: object
```

In [4]:

```
columnHeaders = df_satisfaction.columns

for header in columnHeaders:
    temp = f'{header}'
    df_satisfaction[temp].info()
    print(df_satisfaction[temp].value_counts(),"\n\n")
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: flight_number
Non-Null Count  Dtype
--------------  -----
47074 non-null  int64
dtypes: int64(1)
memory usage: 367.9 KB
219     212
42      181
985     173
86      158
363     156
       ...
4696      1
4663      1
4409      1
3658      1
3595      1
Name: flight_number, Length: 4058, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: origin_station_code
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
IAX    5758
EWX    5589
DEX    4152
ORX    3981
SFX    3495
       ...
DDX       1
CGX       1
AGX       1
TKX       1
DVX       1
Name: origin_station_code, Length: 213, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: destination_station_code
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
EWX    7058
IAX    6821
ORX    4578
DEX    4027
SFX    4003
       ...
OAX       1
ECX       1
```

```
WYX       1
PAX       1
TKX       1
Name: destination_station_code, Length: 212, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: record_locator
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
JNXXVS    6
JXXXMS    6
EDXXY1    6
J3XXCM    6
N4XXT6    6
         ..
N0XX6S    1
A6XXGG    1
CDXX0R    1
B2XX8D    1
G9XX2V    1
Name: record_locator, Length: 33095, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: scheduled_departure_date
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
9/10/2022    1872
9/17/2022    1844
9/24/2022    1826
9/27/2022    1700
9/3/2022     1660
9/6/2022     1657
9/13/2022    1642
9/23/2022    1630
9/19/2022    1605
9/9/2022     1599
9/20/2022    1597
9/16/2022    1586
9/11/2022    1584
9/15/2022    1584
9/2/2022     1575
9/4/2022     1572
9/26/2022    1569
9/18/2022    1559
9/14/2022    1556
9/7/2022     1549
9/12/2022    1547
9/21/2022    1519
9/22/2022    1510
9/1/2022     1505
```

```
9/5/2022     1495
9/29/2022    1474
9/28/2022    1469
9/25/2022    1428
9/8/2022     1370
9/30/2022     991
Name: scheduled_departure_date, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: question_text
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
How satisfied were you with the food & beverage served on your flight from [CITY] to
[CITY]?    34963
What item did you choose?
12111
Name: question_text, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: score
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
2                          8250
1                          7713
5                          6733
3                          6390
4                          5877
chicken entrée             4936
other (specify)            2982
vegetarian entrée          2010
beef entrée                 896
sandwich/burger/wrap        666
seafood entrée              403
snack basket selection      218
Name: score, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: satisfaction_type
Non-Null Count  Dtype
--------------  -----
34963 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Dissatisfied    22353
Satisfied       12610
Name: satisfaction_type, dtype: int64
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: driver_sub_group1
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
food & beverage    47074
Name: driver_sub_group1, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: driver_sub_group2
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
food and beverage satisfaction    34963
comp                              12111
Name: driver_sub_group2, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: arrival_delay_minutes
Non-Null Count  Dtype
--------------  -----
47074 non-null  int64
dtypes: int64(1)
memory usage: 367.9 KB
-14     1288
-15     1280
-10     1207
-13     1198
-12     1195
        ...
 206       1
 276       1
 259       1
 226       1
 301       1
Name: arrival_delay_minutes, Length: 355, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: arrival_delay_group
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Early & Ontime    32868
Delayed           14206
Name: arrival_delay_group, dtype: int64
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: cabin_code_desc
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Economy                27094
Business               18018
United Premium Plus     1962
Name: cabin_code_desc, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: cabin_name
Non-Null Count  Dtype
--------------  -----
27094 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Economy         18438
Economy Plus     8656
Name: cabin_name, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: entity
Non-Null Count  Dtype
--------------  -----
47071 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Domestic    29242
Atlantic    12906
Latin        3378
Pacific      1545
Name: entity, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: number_of_legs
Non-Null Count  Dtype
--------------  -----
47074 non-null  int64
dtypes: int64(1)
memory usage: 367.9 KB
1    31504
2    14941
3      629
Name: number_of_legs, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: seat_factor_band
Non-Null Count  Dtype
```

```
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
90+        31718
80+         8359
70+         3626
0 to 70     3371
Name: seat_factor_band, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: loyalty_program_level
Non-Null Count  Dtype
--------------  -----
35458 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
non-elite           19331
premier silver       4857
premier 1k           4334
premier gold         3362
premier platinum     2637
global services       934
NBK                     3
Name: loyalty_program_level, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: generation
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Boomer        22282
Gen X         14889
Millennial     6559
Silent         2302
Gen Z          1036
Greatest          3
NBK               3
Name: generation, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: fleet_type_description
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
B737-900     6838
B777-200     5677
B737-800     5161
B787-9       3224
ERJ-175      2923
```

```
A320-200     2787
B767-300     2509
B777-300     2435
A319-100     2425
B737-MAX9    1918
B787-10      1671
B757-200     1643
B767-400     1414
B787-8       1302
B737-700     1293
CRJ-200      1078
B737-MAX8     805
B757-300      526
ERJ-170       486
ERJ-145       426
CRJ-550       293
CRJ-700       240
Name: fleet_type_description, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: fleet_usage
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Mainline    41628
Express      5446
Name: fleet_usage, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: equipment_type_code
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
37K    6366
73Y    3702
78P    3026
20S    2707
77X    2435
37X    1918
77E    1880
76L    1690
78J    1671
19F    1539
76S    1414
77U    1396
78H    1302
E75    1295
73G    1293
73Q    1226
75B     928
19G     886
77N     850
```

```
76A     819
E7A     807
37E     805
77G     728
75S     715
CRJ     697
77M     539
75E     526
E7F     498
E7R     486
73C     472
XMJ     426
CRZ     381
E7Q     322
C5G     293
77O     284
CR7     240
73U     233
78Z     198
20C      80
E7M       1
Name: equipment_type_code, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: ua_uax
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
UA     41628
UAX     5446
Name: ua_uax, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: actual_flown_miles
Non-Null Count  Dtype
--------------  -----
47074 non-null  int64
dtypes: int64(1)
memory usage: 367.9 KB
2565    709
3466    664
5368    581
2454    562
4292    484
       ...
784       1
1540      1
1648      1
1351      1
45        1
Name: actual_flown_miles, Length: 731, dtype: int64


<class 'pandas.core.series.Series'>
```

```
RangeIndex: 47074 entries, 0 to 47073
Series name: haul_type
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Medium    26423
Long      16364
Short      4287
Name: haul_type, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: departure_gate
Non-Null Count  Dtype
--------------  -----
46977 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
A8     589
3      469
2      450
4      405
B8     367
      ...
S1       1
78       1
A29      1
D09      1
A3E      1
Name: departure_gate, Length: 656, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: arrival_gate
Non-Null Count  Dtype
--------------  -----
46547 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
HOLD    965
B55     480
A8      443
C5      435
C3      434
       ...
89        1
A5F       1
A3E       1
410       1
-2-       1
Name: arrival_gate, Length: 692, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: international_domestic_indicator
```

```
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
Domestic         29242
International    17832
Name: international_domestic_indicator, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: response_group
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
member        35382
non-member    11689
NBK               3
Name: response_group, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: media_provider
Non-Null Count  Dtype
--------------  -----
45535 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
PANASONIC    24979
THALES       12765
GOGO          3907
VIASAT        3884
Name: media_provider, dtype: int64


<class 'pandas.core.series.Series'>
RangeIndex: 47074 entries, 0 to 47073
Series name: hub_spoke
Non-Null Count  Dtype
--------------  -----
47074 non-null  object
dtypes: object(1)
memory usage: 367.9+ KB
hub departure      24343
spoke departure    22731
Name: hub_spoke, dtype: int64
```

Cleaning the table

In [5]:

```
# FILLING NULL VALUES WITH MOST RELEVANT VALUES

df_satisfaction.entity[df_satisfaction["entity"].isnull() == True] = 'Domestic'
df_satisfaction.loyalty_program_level[df_satisfaction["loyalty_program_level"].isnull(
```

```
df_satisfaction.media_provider[df_satisfaction["media_provider"].isnull() == True] = '
df_satisfaction.cabin_code_desc[df_satisfaction.cabin_name == 'Economy Plus'] = 'Econo


# REPLACING VALUES TO CREATE MORE MEANINGFULL CATEGORIES

df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes >= 120] = 'L
df_satisfaction.arrival_delay_minutes[df_satisfaction.arrival_delay_minutes >= 120] =

df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes >= 60] = 'Me
df_satisfaction.arrival_delay_minutes[df_satisfaction.arrival_delay_minutes >= 60] = M

df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes >= 0] = 'Sma
df_satisfaction.arrival_delay_minutes[df_satisfaction.arrival_delay_minutes >= 0] = No

df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes <= -40] = 'V
df_satisfaction.arrival_delay_minutes[df_satisfaction.arrival_delay_minutes <= -40] =

df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes <= -20] = 'M
df_satisfaction.arrival_delay_minutes[df_satisfaction.arrival_delay_minutes <= -20] =

df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes < 0] = 'Smal
df_satisfaction.arrival_delay_minutes[df_satisfaction.arrival_delay_minutes < 0] = Nor
```

```
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:3: SettingWithCopyWar
ning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.entity[df_satisfaction["entity"].isnull() == True] = 'Domestic'
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:4: SettingWithCopyWar
ning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.loyalty_program_level[df_satisfaction["loyalty_program_level"].isnu
ll() == True] = 'not-member'
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:5: SettingWithCopyWar
ning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.media_provider[df_satisfaction["media_provider"].isnull() == True]
= 'PANASONIC'
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:6: SettingWithCopyWar
ning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.cabin_code_desc[df_satisfaction.cabin_name == 'Economy Plus'] = 'Ec
onomy Plus'
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:11: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes >= 120] =
'Long_Delay'          # 2+ hours
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:12: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.arrival_delay_minutes[df_satisfaction.arrival_delay_minutes >= 120]
= None
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:14: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes >= 60] =
'Medium_Delay'         # 1+ hours
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:17: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
```

```
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes >= 0] =
'Small_Delay'           # 0-1 hours
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:20: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes <= -40] =
'Very_Early'          # 30 min + hours
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:23: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes <= -20] =
'Medium_Early'          # 30 min + hours
C:\Users\jaing\AppData\Local\Temp\ipykernel_22084\4259318331.py:26: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/us
er_guide/indexing.html#returning-a-view-versus-a-copy
  df_satisfaction.arrival_delay_group[df_satisfaction.arrival_delay_minutes < 0] = 'S
mall_Early'              # 0-30min early
```

In [6]:
```
# REMOVING OUTLYERS

indexes1 = np.where(df_satisfaction.loyalty_program_level == 'NBK')
indexes2 = np.where(df_satisfaction.generation == 'Greatest')
indexes3 = np.where(df_satisfaction.equipment_type_code == 'E7M')

arr = indexes1[0].tolist() + indexes2[0].tolist() + indexes3[0].tolist()

print(arr)

for i in reversed(range(0,len(arr))):
    df_satisfaction.drop(arr[i],axis=0,inplace=True)
```

```
[2390, 22231, 27848, 562, 43835, 43836, 43235]
```

In [7]:
```
# REMOVING THE COLUMNS
# actual_flown_miles         CAN BE USED FOR A MORE SPECIFIC EXAMINATION LATER
remove_columns = ["record_locator","driver_sub_group1", "arrival_delay_minutes", "cabi
                  "departure_gate", "arrival_gate"]
df_satisfaction = df_satisfaction.drop(columns=remove_columns)
```

Splitting tables by question type

In [8]:
```
a = df_satisfaction.driver_sub_group2 == 'food and beverage satisfaction'
df_q1 = df_satisfaction[a]
df_q2 = df_satisfaction[~a]
```

In [9]:
```
df_q1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 34957 entries, 0 to 47071
Data columns (total 24 columns):
 #   Column                            Non-Null Count  Dtype 
---  ------                            --------------  ----- 
 0   flight_number                     34957 non-null  int64 
 1   origin_station_code               34957 non-null  object
 2   destination_station_code          34957 non-null  object
 3   scheduled_departure_date          34957 non-null  object
 4   question_text                     34957 non-null  object
 5   score                             34957 non-null  object
 6   satisfaction_type                 34957 non-null  object
 7   driver_sub_group2                 34957 non-null  object
 8   arrival_delay_group               34957 non-null  object
 9   cabin_code_desc                   34957 non-null  object
 10  entity                            34957 non-null  object
 11  number_of_legs                    34957 non-null  int64 
 12  seat_factor_band                  34957 non-null  object
 13  loyalty_program_level             34957 non-null  object
 14  generation                        34957 non-null  object
 15  fleet_type_description            34957 non-null  object
 16  fleet_usage                       34957 non-null  object
 17  equipment_type_code               34957 non-null  object
 18  ua_uax                            34957 non-null  object
 19  haul_type                         34957 non-null  object
 20  international_domestic_indicator  34957 non-null  object
 21  response_group                    34957 non-null  object
 22  media_provider                    34957 non-null  object
 23  hub_spoke                         34957 non-null  object
dtypes: int64(2), object(22)
memory usage: 6.7+ MB
```

In [10]:
```
remove_columns = ["question_text", "driver_sub_group2"]
df_q1 = df_q1.drop(columns=remove_columns)
```

In [11]:
```
df_q1['score'].value_counts()
```

Out[11]:
```
2    8250
1    7710
5    6733
3    6388
4    5876
Name: score, dtype: int64
```

Splitting question 1 table by satisfaction vs disatisfaction

In [12]:
```
a = df_q1.satisfaction_type == 'Satisfied'
df_satisfied = df_q1[a]
df_dissatisfied = df_q1[~a]
```

In [13]:
```
remove_columns = ["satisfaction_type", "score"]
df_q1 = df_q1.drop(columns=remove_columns)
df_satisfied = df_satisfied.drop(columns=remove_columns)
df_dissatisfied = df_dissatisfied.drop(columns=remove_columns)
```

In [14]:
```
columnHeaders = df_satisfied.columns

for header in columnHeaders:
```

```
    temp = f'{header}'
    df_satisfaction[temp].info()
    print(df_satisfaction[temp].value_counts(),"\n\n")
```

```
<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: flight_number
Non-Null Count  Dtype
--------------  -----
47067 non-null  int64
dtypes: int64(1)
memory usage: 735.4 KB
219     212
42      181
985     173
86      158
363     156
       ...
4663      1
4409      1
3658      1
4188      1
2219      1
Name: flight_number, Length: 4058, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: origin_station_code
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
IAX    5755
EWX    5589
DEX    4152
ORX    3981
SFX    3495
       ...
DDX       1
CGX       1
AGX       1
TKX       1
DVX       1
Name: origin_station_code, Length: 213, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: destination_station_code
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
EWX    7057
IAX    6819
ORX    4578
DEX    4026
SFX    4003
       ...
OAX       1
ECX       1
```

```
WYX       1
PAX       1
TKX       1
Name: destination_station_code, Length: 212, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: scheduled_departure_date
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
9/10/2022    1872
9/17/2022    1844
9/24/2022    1826
9/27/2022    1700
9/3/2022     1660
9/6/2022     1657
9/13/2022    1642
9/23/2022    1630
9/19/2022    1605
9/9/2022     1599
9/20/2022    1597
9/16/2022    1586
9/11/2022    1584
9/15/2022    1583
9/2/2022     1574
9/4/2022     1572
9/26/2022    1569
9/18/2022    1558
9/14/2022    1556
9/7/2022     1549
9/12/2022    1547
9/21/2022    1519
9/22/2022    1510
9/1/2022     1504
9/5/2022     1495
9/29/2022    1474
9/28/2022    1466
9/25/2022    1428
9/8/2022     1370
9/30/2022     991
Name: scheduled_departure_date, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: arrival_delay_group
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
Small_Early     20934
Small_Delay     12968
Medium_Early     9811
Medium_Delay     1416
Very_Early       1273
```

```
Long_Delay        665
Name: arrival_delay_group, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: cabin_code_desc
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
Economy                18435
Business               18015
Economy Plus            8655
United Premium Plus     1962
Name: cabin_code_desc, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: entity
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
Domestic    29239
Atlantic    12906
Latin        3377
Pacific      1545
Name: entity, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: number_of_legs
Non-Null Count  Dtype
--------------  -----
47067 non-null  int64
dtypes: int64(1)
memory usage: 735.4 KB
1    31500
2    14938
3      629
Name: number_of_legs, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: seat_factor_band
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
90+        31714
80+         8358
70+         3625
0 to 70     3370
```

```
Name: seat_factor_band, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: loyalty_program_level
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
non-elite           19331
not-member          11615
premier silver       4855
premier 1k           4334
premier gold         3362
premier platinum     2636
global services       934
Name: loyalty_program_level, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: generation
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
Boomer        22281
Gen X         14889
Millennial     6559
Silent         2302
Gen Z          1036
Name: generation, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: fleet_type_description
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
B737-900     6835
B777-200     5677
B737-800     5161
B787-9       3224
ERJ-175      2920
A320-200     2787
B767-300     2509
B777-300     2435
A319-100     2425
B737-MAX9    1918
B787-10      1671
B757-200     1643
B767-400     1414
B787-8       1302
B737-700     1292
```

```
CRJ-200      1078
B737-MAX8     805
B757-300      526
ERJ-170       486
ERJ-145       426
CRJ-550       293
CRJ-700       240
Name: fleet_type_description, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: fleet_usage
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
Mainline    41624
Express      5443
Name: fleet_usage, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: equipment_type_code
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
37K    6363
73Y    3702
78P    3026
20S    2707
77X    2435
37X    1918
77E    1880
76L    1690
78J    1671
19F    1539
76S    1414
77U    1396
78H    1302
E75    1295
73G    1292
73Q    1226
75B     928
19G     886
77N     850
76A     819
E7A     805
37E     805
77G     728
75S     715
CRJ     697
77M     539
75E     526
E7F     498
E7R     486
```

```
73C     472
XMJ     426
CRZ     381
E7Q     322
C5G     293
77O     284
CR7     240
73U     233
78Z     198
20C      80
Name: equipment_type_code, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: ua_uax
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
UA     41624
UAX     5443
Name: ua_uax, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: haul_type
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
Medium    26418
Long      16364
Short      4285
Name: haul_type, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: international_domestic_indicator
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
Domestic         29236
International    17831
Name: international_domestic_indicator, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: response_group
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
```

```
memory usage: 735.4+ KB
member        35379
non-member    11688
Name: response_group, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: media_provider
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
PANASONIC    26517
THALES       12761
GOGO          3905
VIASAT        3884
Name: media_provider, dtype: int64


<class 'pandas.core.series.Series'>
Int64Index: 47067 entries, 0 to 47073
Series name: hub_spoke
Non-Null Count  Dtype
--------------  -----
47067 non-null  object
dtypes: object(1)
memory usage: 735.4+ KB
hub departure      24340
spoke departure    22727
Name: hub_spoke, dtype: int64
```

Creating new dataframe to calculate factor-effect

In [15]:

```
imbalance_pos = df_satisfied.shape[0]/df_q1.shape[0]*100

head = []
factors = []
total_counts = []
pos_count = []
neg_count = []
factor_percent = []
pos_effect = []

columnHeaders = df_q1.columns
for header in columnHeaders:
    temp = f'{header}'
    arr = df_q1[temp].unique()
    for i in range(0,len(arr)):
        head.append(temp)

        factors.append(arr[i])

        len_factor0 = np.where(df_q1[temp] == arr[i])
        total_counts.append(len(len_factor0[0]))
```

```
        len_factor1 = np.where(df_satisfied[temp] == arr[i])
        pos_count.append(len(len_factor1[0]))

        len_factor2 = np.where(df_dissatisfied[temp] == arr[i])
        neg_count.append(len(len_factor2[0]))

        a = round((len(len_factor0[0])/df_q1.shape[0])*100,2)
        factor_percent.append(a)

        b = round((len(len_factor1[0])/len(len_factor0[0]))*100-imbalance_pos,2)
        pos_effect.append(round(a*b,2))



df_factor_effect = pd.DataFrame(np.column_stack([head,factors,total_counts,pos_count,r
                                columns=['Head', 'Factors','Total Counts', 'Satisfied C
                                         'Factor Contribution Percent', 'Factor Affect'
```

In [16]:
```
# REMOVING LOW AFFECTING FACTORS

indexes = []
low_affect = df_factor_effect['Factor Affect'].to_list()
for i in range(0,len(low_affect)):
    low_affect[i] = abs(float(low_affect[i]))
    if low_affect[i] <= 10:
        indexes.append(i)
for i in reversed(range(0,len(indexes))):
    df_factor_effect.drop(indexes[i],axis=0,inplace=True)
```

In [17]:
```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', 1000)
pd.set_option('display.colheader_justify', 'center')
pd.set_option('display.precision', 3)

display(df_factor_effect)
```

	Head	Factors	Total Counts	Satisfied Count	Disatisfied Count	Factor Contribution Percent	Facto Affec
4060	origin_station_code	DEX	3215	1285	1930	9.2	35.8
4062	origin_station_code	EWX	4011	1278	2733	11.47	-48.2
4063	origin_station_code	LAX	1582	532	1050	4.53	-11.0
4067	origin_station_code	ORX	3131	1228	1903	8.96	28.2
4072	origin_station_code	SFX	2553	846	1707	7.3	-21.3
4085	origin_station_code	CLX	241	123	118	0.69	10.3
4160	origin_station_code	HNX	610	151	459	1.75	-19.8
4186	origin_station_code	TLX	304	58	246	0.87	-14.7
4272	destination_station_code	DEX	3136	1198	1938	8.97	19.1
4280	destination_station_code	IAX	5005	2027	2978	14.32	63.4
4281	destination_station_code	EWX	5018	1632	3386	14.35	-50.9
4284	destination_station_code	SFX	2915	972	1943	8.34	-22.7
4312	destination_station_code	MCX	444	209	235	1.27	13.9
4372	destination_station_code	LHX	769	221	548	2.2	-16.1
4485	scheduled_departure_date	9/3/2022	1218	491	727	3.48	14.7
4486	scheduled_departure_date	9/4/2022	1142	447	695	3.27	10.0
4492	scheduled_departure_date	9/10/2022	1362	527	835	3.9	10.2
4493	scheduled_departure_date	9/11/2022	1177	383	794	3.37	-11.9
4495	scheduled_departure_date	9/13/2022	1214	487	727	3.47	14.0
4513	arrival_delay_group	Medium_Early	7191	2712	4479	20.57	33.7
4514	arrival_delay_group	Small_Delay	9476	3221	6255	27.11	-56.3
4515	arrival_delay_group	Small_Early	15964	6017	9947	45.67	73.9
4516	arrival_delay_group	Long_Delay	499	109	390	1.43	-20.3
4518	arrival_delay_group	Medium_Delay	1031	247	784	2.95	-35.7
4519	cabin_code_desc	Economy	15842	5793	10049	45.32	22.6
4520	cabin_code_desc	Business	10741	4012	6729	30.73	39.3
4521	cabin_code_desc	Economy Plus	7270	2400	4870	20.8	-63.6
4523	entity	Domestic	23851	8813	15038	68.23	60.0
4526	entity	Atlantic	7627	2552	5075	21.82	-56.9
4527	number_of_legs	2	11571	4275	7296	33.1	29.1
4529	number_of_legs	1	22881	8139	14742	65.45	-32.7
4531	seat_factor_band	90+	23723	8394	15329	67.86	-46.8

	Head	Factors	Total Counts	Satisfied Count	Disatisfied Count	Factor Contribution Percent	Facto Affec
4533	seat_factor_band	0 to 70	2439	994	1445	6.98	32.6
4534	loyalty_program_level	not-member	9380	3555	5825	26.83	49.
4535	loyalty_program_level	premier platinum	1773	562	1211	5.07	-22.1
4536	loyalty_program_level	non-elite	14705	5343	9362	42.07	10.9
4538	loyalty_program_level	premier gold	2337	790	1547	6.69	-15.1
4539	loyalty_program_level	premier 1k	2816	936	1880	8.06	-22.8
4541	generation	Gen X	11088	3962	7126	31.72	-10.7
4542	generation	Boomer	16366	5959	10407	46.82	15.9
4545	generation	Millennial	4979	1735	3244	14.24	-17.3
4548	fleet_type_description	A319-100	2064	850	1214	5.9	30.1
4551	fleet_type_description	B737-800	4146	1577	2569	11.86	23.3
4555	fleet_type_description	B737-900	5342	2019	3323	15.28	26.2
4561	fleet_type_description	B777-200	3663	1182	2481	10.48	-39.8
4565	fleet_type_description	B787-10	1043	332	711	2.98	-12.6
4567	fleet_type_description	B787-9	1944	606	1338	5.56	-27.2
4573	equipment_type_code	19F	1312	540	772	3.75	19.0
4577	equipment_type_code	73Y	2996	1159	1837	8.57	22.3
4584	equipment_type_code	37K	4982	1868	3114	14.25	20.2
4585	equipment_type_code	19G	752	310	442	2.15	11.0
4601	equipment_type_code	77E	1156	364	792	3.31	-15.1
4603	equipment_type_code	77U	854	267	587	2.44	-11.7
4604	equipment_type_code	78J	1043	332	711	2.98	-12.6
4607	equipment_type_code	78P	1822	567	1255	5.21	-25.7
4611	haul_type	Short	4208	1415	2793	12.04	-29.3
4612	haul_type	Medium	21038	8002	13036	60.18	118.5
4613	haul_type	Long	9711	3192	6519	27.78	-88.5
4614	international_domestic_indicator	Domestic	23848	8813	15035	68.22	60.0
4615	international_domestic_indicator	International	11109	3796	7313	31.78	-60.3
4616	response_group	non-member	9433	3576	5857	26.98	49.6
4617	response_group	member	25524	9033	16491	73.02	-49.6
4618	media_provider	PANASONIC	18148	6385	11763	51.92	-46.2
4620	media_provider	THALES	10157	3815	6342	29.06	43.3

	Head	Factors	Total Counts	Satisfied Count	Disatisfied Count	Factor Contribution Percent	Facto Affec
4621	media_provider	VIASAT	3093	1160	1933	8.85	12.6
4622	hub_spoke	spoke departure	16657	6053	10604	47.65	12.8

In [18]:
```
df_factor_effect.to_csv(r'Datasets/Factors_Effect_Dataset.csv')
```

In [ ]:
```

```

From this point onwards, the analysis was done in excel

G-Drive link:
https://docs.google.com/spreadsheets/d/1KYWEmVz-TJAaF1vNqoRVMNemXV0HHkJpveUgmSQmKm8/edit#gid=0