Task 3: Data Analysis

2011_Transactions.csv is a database of transactions in the travel industry. This task requires you to provide summary statistics and run some basic analysis on this file

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

In [2]:
```
# read dataset as pandas dataframe

trans = pd.read_csv('2011_Transactions.csv', encoding= 'unicode_escape')
```

In [3]:
```
# check column information in dataset

trans.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8437 entries, 0 to 8436
Data columns (total 21 columns):
machine_id            8437 non-null int64
site_session_id       8437 non-null int64
domain_id             8437 non-null uint64
event_date            8437 non-null int64
event_time            8344 non-null object
prod_category_id      8437 non-null int64
prod_name             8437 non-null object
prod_qty              8437 non-null int64
prod_totprice         8437 non-null float64
basket_tot            8437 non-null float64
hoh_most_education    8437 non-null int64
census_region         8437 non-null object
household_size        8437 non-null object
hoh_oldest_age        8437 non-null int64
household_income      8437 non-null int64
children              8437 non-null int64
racial_background     8437 non-null int64
connection_speed      8437 non-null int64
country_of_origin     8437 non-null int64
zip_code              8437 non-null int64
domain_name           8437 non-null object
dtypes: float64(2), int64(13), object(5), uint64(1)
memory usage: 1.4+ MB
```

In [4]:
```
# check the first 5 lines in dataset

trans.head(5)
```

Out[4]:

	machine_id	site_session_id	domain_id	event_date	event_time	prod_category_id
0	13512886	3669584318587	9530952911301729568	20110720	2:03:02	45
1	49645796	3355708493805	17475197073474272331	20110228	21:16:51	43
2	58622574	3360275042310	1875457313788268580	20110325	15:58:34	43
3	59850911	3815456641153	1875457313788268580	20110726	21:31:36	43
4	59850911	71716842443028	1875457313788268580	20111220	23:29:16	43

5 rows × 21 columns

1. What is the mean number of transactions made per individual (machine_id)? Restricting to the travel categories (prod_category_id = 43, 44 and 45) what is the mean number of transactions?

In [7]:
```
trans['machine_id'].value_counts().reset_index().mean(axis=0)
```

Out[7]:
```
index         8.607783e+07
machine_id    2.015047e+00
dtype: float64
```

The mean number of transactions made per individual (machine_id) is 2.015.

In [8]:
```
# Restricting to the travel categories (prod_category_id = 43, 44 and 45)

travCate = trans[trans['prod_category_id'].isin([43, 44, 45])]
```

In [9]:
```
travCate['machine_id'].value_counts().reset_index().mean(axis=0)
```

Out[9]:
```
index         8.607268e+07
machine_id    1.958835e+00
dtype: float64
```

The mean number of transactions made per individual (machine_id) with travel categories

restriction is 1.959.

2. Restricting attention again to the travel categories, for those who had multiple transactions:
   - How far apart are the transactions? Using this, categorize transactions that are likely related to each other, e.g. car rented and flight booked for the same trip.
   - For transactions related to each other, document the sequence of bookings (e.g. car booked first, then air travel, then hotel) in a visual manner.

In [7]:
```
travCate
```

Out[7]:

	machine_id	site_session_id	domain_id	event_date	event_time	prod_category_id
0	13512886	3669584318587	9530952911301729568	20110720	2:03:02	45
1	49645796	3355708493805	17475197073474272331	20110228	21:16:51	43
2	58622574	3360275042310	18754573137882685 80	20110325	15:58:34	43
3	59850911	3815456641153	18754573137882685 80	20110726	21:31:36	43
4	59850911	71716842443028	18754573137882685 80	20111220	23:29:16	43
...	...	...	...	...	...	...
8432	95319234	7671937175476	14244539005989555924	20110102	4:16:06	43
8433	95319894	5598578479186	5127767531286671227	20110609	0:21:18	43
8434	95319894	5598578479186	5127767531286671227	20110609	0:21:18	43
8435	95319894	68167160172789	14244539005989555924	20111119	20:40:14	43
8436	95320067	5474873184413	3010609366849421442	20110823	0:07:21	45

7804 rows × 21 columns

In [26]:
```
# The transactions with same machine_id appearing at least twice in the dat

travCate['machine_id'].value_counts().loc[lambda x : x > 1]
```

Out[26]:
```
93855448    34
85432156    29
82495406    28
91433731    26
81202618    24
            ..
83749430     2
89076285     2
90022106     2
80261697     2
87107610     2
Name: machine_id, Length: 1632, dtype: int64
```

In [9]:
```
# check the transactions with a specific machine_id with sorted dates

travCate[travCate['machine_id'] == 82495406].sort_values(by=['event_date',
```

Out[9]:

	machine_id	site_session_id	domain_id	event_date	event_time	prod_category_id
2355	82495406	4307935170488	1037677886457106237	20110106	17:55:12	43
2358	82495406	4340104368144	1037677886457106237	20110404	16:50:50	43
2359	82495406	4340104368144	1037677886457106237	20110404	16:50:50	43
2360	82495406	4340104368144	1037677886457106237	20110404	16:50:50	45
2361	82495406	4340104368144	1037677886457106237	20110404	16:50:50	43
2349	82495406	4222094741522	1037677886457106237	20110406	17:48:44	43
2350	82495406	4222094741522	1037677886457106237	20110406	17:48:44	45
2351	82495406	4222094741522	1037677886457106237	20110406	17:48:44	43
2354	82495406	4295779160166	10326615452578885078	20110629	14:23:12	43
2348	82495406	4216806969452	1037677886457106237	20110705	15:56:38	43
2353	82495406	4284431208557	10326615452578885078	20110706	14:56:57	43
2357	82495406	4315771048053	1037677886457106237	20110714	18:25:41	43
2356	82495406	4314067374201	1037677886457106237	20110718	17:03:00	43
2352	82495406	4273007235202	7101213156062330967	20110727	15:34:30	43
2362	82495406	4452757016732	10326615452578885078	20110822	16:45:04	43
2364	82495406	69766386684102	1037677886457106237	20111003	13:28:54	43

	machine_id	site_session_id	domain_id	event_date	event_time	prod_category_id
2365	82495406	69766386684102	1037677886457106237	20111003	13:28:54	45
2366	82495406	69766386684102	1037677886457106237	20111003	13:28:54	43
2367	82495406	69911374860488	1037677886457106237	20111005	14:28:43	43
2368	82495406	69911374860488	1037677886457106237	20111005	14:28:43	43
2369	82495406	69911374860488	1037677886457106237	20111005	14:28:43	45
2371	82495406	71805965832400	10326615452578885078	20111013	15:00:27	43
2370	82495406	70951134695652	10326615452578885078	20111102	16:50:27	43
2375	82495406	73582287655154	1037677886457106237	20111116	21:24:32	43
2363	82495406	69682984784130	10326615452578885078	20111202	22:14:09	43
2372	82495406	72042780168468	1037677886457106237	20111220	17:32:31	43
2373	82495406	72042780168468	1037677886457106237	20111220	17:32:31	45
2374	82495406	72042780168468	1037677886457106237	20111220	17:32:31	43

28 rows × 21 columns

In [30]:
```
travCate[(travCate['machine_id'] == 93855448) & (travCate['event_date'] ==
```

Out[30]:

	machine_id	site_session_id	domain_id	event_date	event_time	prod_category_id
7344	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7345	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7346	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7347	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7348	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7349	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7350	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7351	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7352	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7353	93855448	5833288192182	1037677886457106237	20110917	12:26:20	43
7354	93855448	5833288257718	1037677886457106237	20110917	13:17:39	43
7355	93855448	5833288257718	1037677886457106237	20110917	13:17:39	43

12 rows × 21 columns

In [ ]:
```

```