ProcessData

March 12, 2023

[33]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import chi2_contingency, ttest_ind, pearsonr
```

[8]:
```
cab_data = pd.read_csv("Cab_Data.csv")
customer_id = pd.read_csv("Customer_ID.csv")
transaction_id = pd.read_csv("Transaction_ID.csv")
city = pd.read_csv("City.csv")
```

[10]:
```
cab_data.head()
cab_data.dtypes

customer_id.head()
customer_id.dtypes

transaction_id.head()
transaction_id.dtypes

city.head()
city.dtypes
```

[10]:
```
City          object
Population    object
Users         object
dtype: object
```

[11]:
```
# Merge transaction_id and customer_id dataframes on customer ID column
transaction_customer = pd.merge(transaction_id, customer_id, on='Customer ID')

# Merge cab_data and city dataframes on City column
cab_city = pd.merge(cab_data, city, on='City')
```

[14]:
```
# Create a new column in cab_city dataframe for profit
cab_city['Profit'] = cab_city['Price Charged'] - cab_city['Cost of Trip']
```

[15]:
```
# Merge transaction_customer and cab_city dataframes on Transaction ID column
master_data = pd.merge(transaction_customer, cab_city, on='Transaction ID')
# The master_data dataframe contains all the relevant information from the␣
 ↪original datasets, including details of transactions, customer demographics,␣
 ↪and cab company details.
```

[21]:
```
# Check for duplicates in master_data dataframe
master_data.duplicated().sum()
```

[21]:
```
0
```

[22]:
```
# Drop any duplicate rows from master_data dataframe
master_data.drop_duplicates(inplace=True)
```

[19]:
```
# Check for missing values in the master_data dataframe
master_data.isna().sum()
```

[19]:
```
Transaction ID        0
Customer ID           0
Payment_Mode          0
Gender                0
Age                   0
Income (USD/Month)    0
Date of Travel        0
Company               0
City                  0
KM Travelled          0
Price Charged         0
Cost of Trip          0
Population            0
Users                 0
Profit                0
dtype: int64
```

[20]:
```
# Detect and remove outliers in the Profit column using Z-score
from scipy import stats
z_scores = stats.zscore(master_data['Profit'])
abs_z_scores = np.abs(z_scores)
filtered_entries = (abs_z_scores < 3)
master_data = master_data[filtered_entries]
```

1 EDA

[26]:
```
# Merge the datasets
merged_data = pd.merge(cab_data, transaction_id, on='Transaction ID')
merged_data = pd.merge(merged_data, customer_id, on='Customer ID')
merged_data = pd.merge(merged_data, city, on='City')
```

[29]:
```
merged_data.head()
```

[29]:
```
   Transaction ID  Date of Travel     Company        City  KM Travelled  \
0        10000011           42377    Pink Cab  ATLANTA GA         30.45
1        10351127           43302  Yellow Cab  ATLANTA GA         26.19
2        10412921           43427  Yellow Cab  ATLANTA GA         42.55
3        10000012           42375    Pink Cab  ATLANTA GA         28.62
4        10320494           43211  Yellow Cab  ATLANTA GA         36.38

   Price Charged  Cost of Trip  Customer ID Payment_Mode Gender  Age  \
0         370.95      313.6350        29290         Card   Male   28
1         598.70      317.4228        29290         Cash   Male   28
2         792.05      597.4020        29290         Card   Male   28
3         358.52      334.8540        27703         Card   Male   27
4         721.10      467.1192        27703         Card   Male   27

   Income (USD/Month) Population     Users
0               10813   814,885    24,701
1               10813   814,885    24,701
2               10813   814,885    24,701
3                9237   814,885    24,701
4                9237   814,885    24,701
```

[27]:
```
# Explore the data
print(merged_data.head())
print(merged_data.describe())
print(merged_data.info())
```

```
   Transaction ID  Date of Travel     Company        City  KM Travelled  \
0        10000011           42377    Pink Cab  ATLANTA GA         30.45
1        10351127           43302  Yellow Cab  ATLANTA GA         26.19
2        10412921           43427  Yellow Cab  ATLANTA GA         42.55
3        10000012           42375    Pink Cab  ATLANTA GA         28.62
4        10320494           43211  Yellow Cab  ATLANTA GA         36.38

   Price Charged  Cost of Trip  Customer ID Payment_Mode Gender  Age  \
0         370.95      313.6350        29290         Card   Male   28
1         598.70      317.4228        29290         Cash   Male   28
2         792.05      597.4020        29290         Card   Male   28
3         358.52      334.8540        27703         Card   Male   27
4         721.10      467.1192        27703         Card   Male   27
```

```
   Income (USD/Month) Population     Users
0               10813   814,885    24,701
1               10813   814,885    24,701
2               10813   814,885    24,701
3                9237   814,885    24,701
4                9237   814,885    24,701
       Transaction ID  Date of Travel   KM Travelled  Price Charged  \
count    3.593920e+05   359392.000000  359392.000000  359392.000000
mean     1.022076e+07    42964.067998      22.567254     423.443311
std      1.268058e+05      307.467197      12.233526     274.378911
min      1.000001e+07    42371.000000       1.900000      15.600000
25%      1.011081e+07    42697.000000      12.000000     206.437500
50%      1.022104e+07    42988.000000      22.440000     386.360000
75%      1.033094e+07    43232.000000      32.960000     583.660000
max      1.044011e+07    43465.000000      48.000000    2048.030000

        Cost of Trip    Customer ID            Age  Income (USD/Month)
count  359392.000000  359392.000000  359392.000000       359392.000000
mean      286.190113   19191.652115      35.336705        15048.822937
std       157.993661   21012.412463      12.594234         7969.409482
min        19.000000       1.000000      18.000000         2000.000000
25%       151.200000    2705.000000      25.000000         8424.000000
50%       282.480000    7459.000000      33.000000        14685.000000
75%       413.683200   36078.000000      42.000000        21035.000000
max       691.200000   60000.000000      65.000000        35000.000000
<class 'pandas.core.frame.DataFrame'>
Int64Index: 359392 entries, 0 to 359391
Data columns (total 14 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   Transaction ID      359392 non-null  int64
 1   Date of Travel      359392 non-null  int64
 2   Company             359392 non-null  object
 3   City                359392 non-null  object
 4   KM Travelled        359392 non-null  float64
 5   Price Charged       359392 non-null  float64
 6   Cost of Trip        359392 non-null  float64
 7   Customer ID         359392 non-null  int64
 8   Payment_Mode        359392 non-null  object
 9   Gender              359392 non-null  object
 10  Age                 359392 non-null  int64
 11  Income (USD/Month)  359392 non-null  int64
 12  Population          359392 non-null  object
 13  Users               359392 non-null  object
dtypes: float64(3), int64(5), object(6)
memory usage: 41.1+ MB
None
```

1.1 Create a list of hypotheses to investigate

[31]:
```
# Hypothesis 1: Male customers prefer Yellow Cab over Pink Cab
# Create a contingency table of cab company and gender
cont_table = pd.crosstab(merged_data['Company'], merged_data['Gender'])
```

[34]:
```
# Conduct a chi-squared test to determine if there is a significant difference
# in the proportion of male customers between the two cab companies
chi2_stat, p_value, dof, expected = chi2_contingency(cont_table)
alpha = 0.05
if p_value < alpha:
    print("Hypothesis 1: Reject null hypothesis. There is a significant␣
 ↪difference in the proportion of male customers between the two cab companies.
 ↪")
else:
    print("Hypothesis 1: Fail to reject null hypothesis. There is no␣
 ↪significant difference in the proportion of male customers between the two␣
 ↪cab companies.")
```

```
Hypothesis 1: Reject null hypothesis. There is a significant difference in the
proportion of male customers between the two cab companies.
```

[36]:
```
# Hypothesis 2: The cost of the trip is correlated with the distance traveled
# Calculate the correlation coefficient and p-value between distance traveled␣
 ↪and price charged
corr_coef, p_value = pearsonr(merged_data['KM Travelled'], merged_data['Price␣
 ↪Charged'])
if p_value < alpha:
    print("Hypothesis 2: Reject null hypothesis. There is a significant␣
 ↪correlation between distance traveled and price charged.")
else:
    print("Hypothesis 2: Fail to reject null hypothesis. There is no␣
 ↪significant correlation between distance traveled and price charged.")
```

```
Hypothesis 2: Reject null hypothesis. There is a significant correlation between
distance traveled and price charged.
```

[37]:
```
# Hypothesis 3: Customers paying with cash have a higher income than customers␣
 ↪paying with a card
# Calculate the mean income of customers paying with cash and those paying with␣
 ↪a card
cash_mean = merged_data.loc[merged_data['Payment_Mode'] == 'Cash', 'Income (USD/
 ↪Month)'].mean()
card_mean = merged_data.loc[merged_data['Payment_Mode'] == 'Card', 'Income (USD/
 ↪Month)'].mean()
```

```
# Conduct a t-test to determine if there is a significant difference in the␣
 ↪mean income between the two payment modes
t_stat, p_value = ttest_ind(merged_data.loc[merged_data['Payment_Mode'] ==␣
 ↪'Cash', 'Income (USD/Month)'],
                            merged_data.loc[merged_data['Payment_Mode'] ==␣
 ↪'Card', 'Income (USD/Month)'])
if p_value < alpha:
    print("Hypothesis 3: Reject null hypothesis. Customers paying with cash␣
 ↪have a higher income than customers paying with a card.")
else:
    print("Hypothesis 3: Fail to reject null hypothesis. There is no␣
 ↪significant difference in the mean income between the two payment modes.")
```

```
Hypothesis 3: Fail to reject null hypothesis. There is no significant difference
in the mean income between the two payment modes.
```

[ ]: