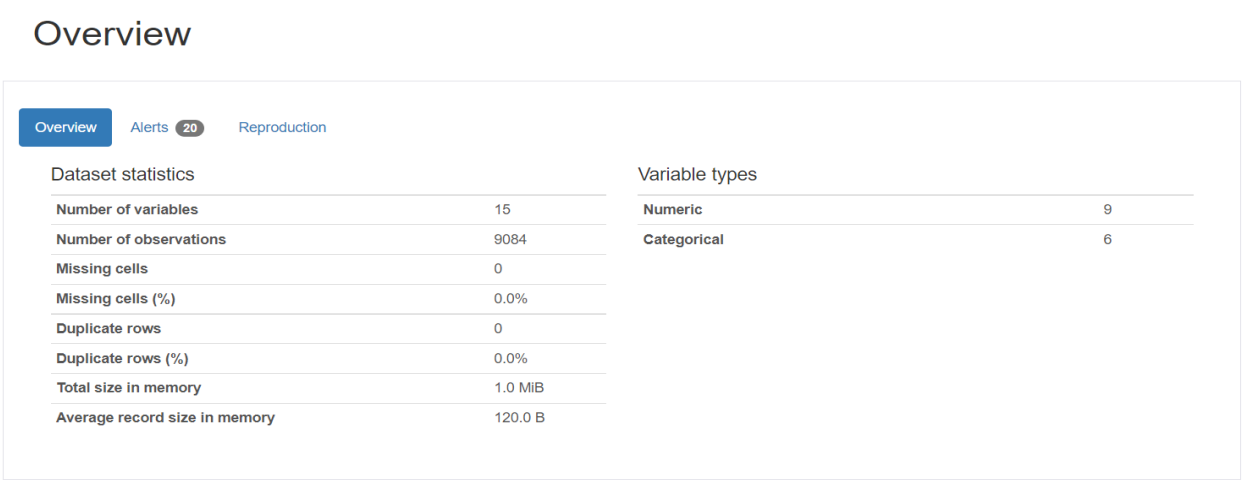
Travel Data Analysis for Business Growth

1. Introduction

The travel sector is a leading field in different parts of the world, with the proper structure and insight great potential can be achieved from this multiplying industry. The travel sector elevated after the pandemic with a drastic increase in income with more and more people involved. This increases the economy of the medium in which they travel. Here in this, we have taken the Travel Pac data from "The Office for National Statistics (ONS)" which is the managerial office of the UK Statistics Authority that provides insights for allocating and decision-making in the United Kingdom. The "TravelPac" data is recorded from the Compact datasets from the "International Passenger Survey (IPS)**"** and is made up of several data files that let us estimate and examine international travel and tourism among particular groupings**.** With this, we analyze proper insights by considering the problem statements, questions, and situations. The IPS is a survey conducted by ONS. The outcomes are taken from in-person interviews with a sample of passengers as they enter or exit the UK via major air, sea, and channel tunnel routes. Approximately 95% of passengers have a chance of being chosen. This can be examined and used to collect insights for Business and other technical purposes.

2. Descriptive Analysis



2.1 Overview of Data

In this component, we can look at the number of rows and columns present in the dataset. It shows the statistics of missing cells, duplicate rows, and the total size of memory for this process. I used pandas profiling to export the descriptive analysis report (pandas-profiling dev documentation. (n.d.).).

Overview of variables in the Dataset:

```
In [23]: 1 df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9084 entries, 0 to 9083
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Year      9084 non-null   int64
 1   quarter   9084 non-null   int64
 2   ukos      9084 non-null   int64
 3   mode      9084 non-null   int64
 4   country   9084 non-null   int64
 5   purpose   9084 non-null   int64
 6   package   9084 non-null   int64
 7   Age       9084 non-null   int64
 8   Sex       9084 non-null   object
 9   duration  9084 non-null   int64
 10  visits    9084 non-null   object
 11  nights    9084 non-null   object
 12  expend    9084 non-null   float64
 13  sample    9084 non-null   object
dtypes: float64(1), int64(9), object(4)
memory usage: 993.7+ KB
```

The data set contains 13 variables which are in the form of rows and columns and has 14 columns and 9084 rows. It contains variables of different data types (Travelpac 2022) .

Variable Overview:

```
In [17]: 1 list(df.columns)
Out[17]: ['Year',
          'quarter',
          'ukos',
          'mode',
          'country',
          'purpose',
          'package',
          'Age',
          'Sex',
          'duration',
          'visits',
          'nights',
          'expend',
          'sample']
```

1.	Year	Representing with (YYYY format) when the interview was taken.
2.	quarter	The four quarters of the calendar year are depicted by a single digit.
3.	ukos	Says whether the contact is a UK resident or an overseas resident.
4.	mode	Represents the mode of travel either as air, sea, or tunnel.
5.	country	The destination country is present in this variable
6.	purpose	The main cause of the visit is noted for all contacts.
7.	package	This indicates whether the contact traveled as part of an all-inclusive tour package or on their own.
8.	Age	Age group where the contact belongs.
9.	Sex	Gender is recorded for all contacts.
10.	duration	The length of the visit is measured in nights. This service is only offered to international residents, departures, and UK residents.
11.	visits	A visit is defined as a round-trip journey. It depicts a departure from and returns to the UK for a UK resident. It indicates an entry into and exit from the UK for an international resident. Those who visited the UK or traveled overseas on many occasions are recorded at each visit.
12.	nights	Nights refers to the total number of nights spent on a journey.
13.	expend	It indicates the total amount spent abroad (for UK residents) or in the UK (for visitors from overseas) throughout the visit.

14.	sample	The sample is the number of contacts from the primary IPS that were utilized to support each row of data in the Travelpac dataset. This could be used as a measure of the dependability of the data being evaluated.

2.2 Data Cleaning

While performing the overview of the data it was seen the duplications of rows and columns weren't present. But that isn't enough to confirm that the data doesn't have noise.

In [7]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import time as time
import seaborn as sns


#Loading Data Frame
df=pd.read_csv("TravelPac Q2 2022.csv")
print(df)
print(df.shape)
```

```
      Year  quarter  ukos  mode  country  purpose  package  Age Sex  duration  \
0     2022        2     1     1       10        1        1    1   2         2
1     2022        2     1     1       10        1        1    2   1         1
2     2022        2     1     1       10        1        1    2   1         2
3     2022        2     1     1       10        1        1    2   2         1
4     2022        2     1     1       10        1        1    2   2         2
...    ...      ...   ...   ...      ...      ...      ...  ...  ..       ...
9079  2022        2     2     3       90        4        1    4   2         2
9080  2022        2     2     3       90        4        1    6   1         2
9081  2022        2     2     3       90        4        1    6   1         3
9082  2022        2     2     3       90        4        1    6   2         4
9083  2022        2     2     3       91        1        1    4   1         2

         visits     nights       expend sample
0      1216.922   4867.687   578037.819      1
1       940.625   2821.874   188124.928      1
2      6044.895  36510.253  4621624.177      5
3      1102.325   3306.974   220464.930      1
4      4353.722  17414.889  1158992.280      3
...         ...        ...          ...    ...
9079    925.756   9257.564   370302.568      1
9080    925.756   5554.539   138863.463      1
9081   1061.371  16981.942  1040143.941      1
9082   1299.207  48070.667   129920.721      1
9083   1326.673   5306.693  2207584.088      1

[9084 rows x 14 columns]
(9084, 14)
```

In the above figure, we have loaded the data frame from that we can see that it contains 9084 rows and 14 columns. (Pandas 2022)

In [28]:
```
df.head(9035)
```

Out[7]:

	Year	quarter	ukos	mode	country	purpose	package	Age	Sex	duration	visits	nights	expend	sample
0	2022	2	1	1	10	1	1	1	2	2	1216.922	4867.687	578037.819	1
1	2022	2	1	1	10	1	1	2	1	1	940.625	2821.874	188124.928	1
2	2022	2	1	1	10	1	1	2	1	2	6044.895	36510.253	4621624.177	5
3	2022	2	1	1	10	1	1	2	2	1	1102.325	3306.974	220464.930	1
4	2022	2	1	1	10	1	1	2	2	2	4353.722	17414.889	1158992.280	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9030	2022	2	2	3	81	5	1	5	2	1	1341.448	1341.448	1397789.310	1
9031	2022	2	2	3	81	5	1	9	1	1	1341.448	1341.448	268289.695	1
9032	2022	2	2	3	81	6	1	5	2	9	#NULL!	#NULL!	49633.594	2
9033	2022	2	2	3	81	6	1	7	2	9	#NULL!	#NULL!	33536.212	1
9034	2022	2	2	3	82	1	1	3	1	1	1289.741	3869.224	1934611.797	1

9035 rows × 14 columns

Later on, I saw that there were Null strings in the visits and nights column which I decided to replace with zero (0) to continue with the process.

In [10]:
```
#Boolean run for null sets
df.isnull().head(9084)
```

Out[10]:

	Year	quarter	ukos	mode	country	purpose	package	Age	Sex	duration	visits	nights	expend	sample
0	False	False	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9079	False	False	False	False	False	False	False	False	False	False	False	False	False	False
9080	False	False	False	False	False	False	False	False	False	False	False	False	False	False
9081	False	False	False	False	False	False	False	False	False	False	False	False	False	False
9082	False	False	False	False	False	False	False	False	False	False	False	False	False	False
9083	False	False	False	False	False	False	False	False	False	False	False	False	False	False

9084 rows × 14 columns

In [11]:
```
#Look for null values in the data frame
df.isnull().sum()
```

Out[11]:
```
Year        0
quarter     0
ukos        0
mode        0
country     0
purpose     0
package     0
Age         0
Sex         0
duration    0
visits      0
nights      0
expend      0
sample      0
dtype: int64
```

Then I performed the Boolean run to check the null sets in the data frame. To double-check null sets examination was also done over the data frame.

In [12]:
```
#Replace Null String with number 0
df[colu] = df[colu].replace(['#NULL!'],(0))

```

In [13]:
```
df.head(9035)
```

Out[13]:

	Year	quarter	ukos	mode	country	purpose	package	Age	Sex	duration	visits	nights	expend	sample
0	2022	2	1	1	10	1	1	1	2	2	1216.922	4867.687	578037.819	1
1	2022	2	1	1	10	1	1	2	1	1	940.625	2821.874	188124.928	1
2	2022	2	1	1	10	1	1	2	1	2	6044.895	36510.253	4621624.177	5
3	2022	2	1	1	10	1	1	2	2	1	1102.325	3306.974	220464.930	1
4	2022	2	1	1	10	1	1	2	2	2	4353.722	17414.889	1158992.280	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9030	2022	2	2	3	81	5	1	5	2	1	1341.448	1341.448	1397789.310	1
9031	2022	2	2	3	81	5	1	9	1	1	1341.448	1341.448	268289.695	1
9032	2022	2	2	3	81	6	1	5	2	9	0	0	49633.594	2
9033	2022	2	2	3	81	6	1	7	2	9	0	0	33536.212	1
9034	2022	2	2	3	82	1	1	3	1	1	1289.741	3869.224	1934611.797	1

9035 rows × 14 columns

In [25]:
```
#df.to_csv("../data/[CLEANED]TravelPac Q2 2022.csv")
df.to_csv('[CLEANED]TravelPac Q2 2022.csv',sep='\t')
```

The null string was replaced with zero and was saved as cleaned data for further operations. After this, I observed the data and found a few problem statements by observing the data and saw the potential insights for business development.

2.3 Business Insights

In this to get efficient insights we must inspect the data and consider the variables by conducting a descriptive analysis of the data, and the noise and null variables must be cleaned from the data. In the later part, by considering the problem statements for a potential business in the travel sector must be analyzed by conducting exploratory data analytics (EDA) which can be used as a possible business insight for a larger scale.

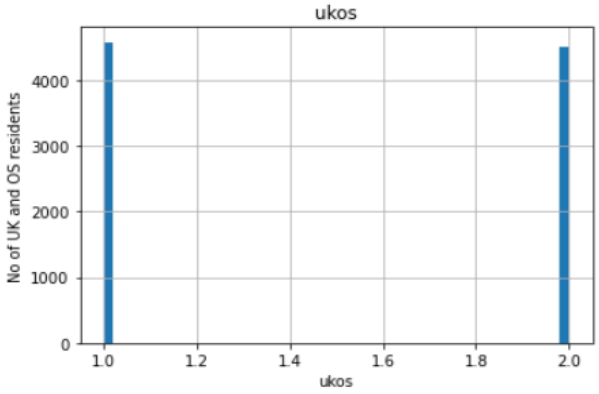
The problem statements which I considered to apply EDA are:

1. Keeping expenditure as the independent variable, compare the expenditure of UK residents and non-UK residents who travel in and out of the UK.
2. Calculate the average expenditure for people who travel independently.
3. Calculate the average expenditure for people who travel non-independently.
4. Find the expenditure of different age groups.
5. Purpose and expenditure required for that specific purpose.

After dealing with this problem state, we can collect insights and make business recommendations for growth and other technical problems. For those, I will be using models like Pearson's rho for finding the positive correlation, Kendall's Tau which usually deals with compact values, and Spearman's rho correlation. Additionally, Random Forest regression and classification learning have been used to find the accuracy of the model.

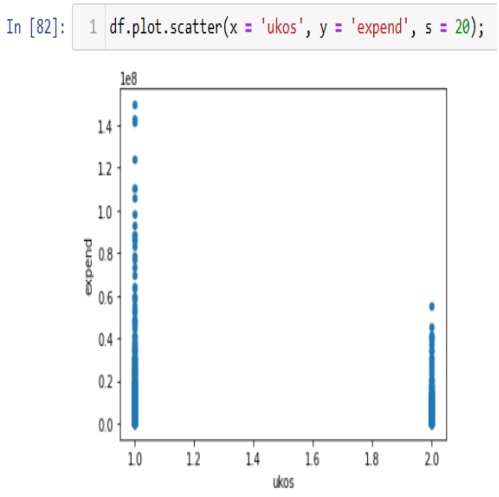
3. Explorative Data Analytics (EDA)

3.1 Expenditure for UK residents vs non-UK residents

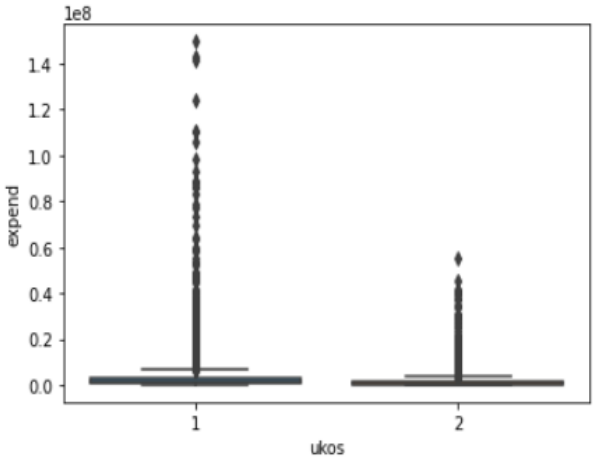


Hist plot representing the ukos count

Expenditure is the key to travel, which can be shown as a fundamental term for a client or a customer which can improve the business, here we have considered both UK residents represented by 1 and overseas residents represented by 2. Additionally, we have plotted a histogram with an x-axis representing the nationality and a y-axis representing the count by this we can see that the number of people in the UK has more travel history which can be seen as a potential market.



Scatter plot indicating expenditure spent



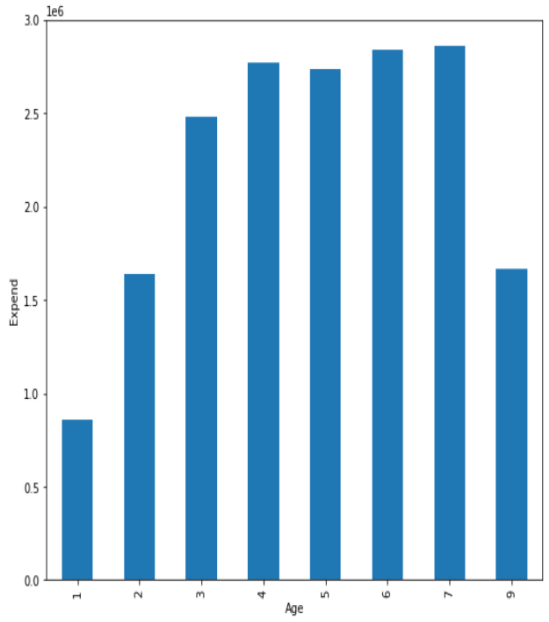
Box plot indicating the expenditure

This shows that the travel sector in the UK is in a good phase and people in the UK travel more to other countries and the expenditure spent by people in the UK is exceptionally higher than the expenditure spent by overseas residents this can be utilized as a business insight to boost the

trend by focusing marketing campaigns on them which can multiply the revenue in travel and tourism industry (Matplotlib Scatter. (n.d.).

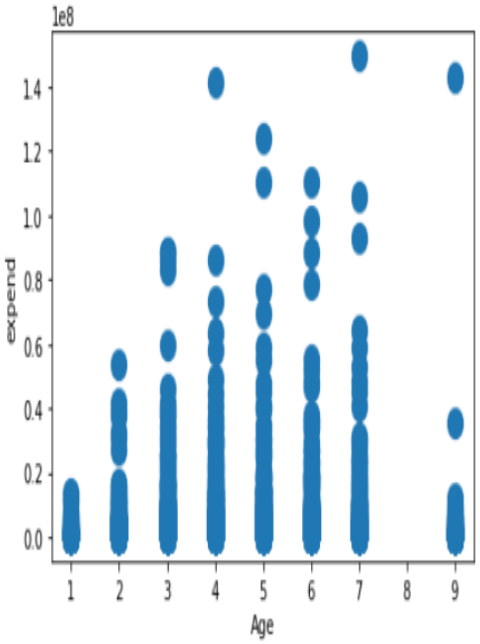
3.2 Expenditure for different age groups

```
import matplotlib.pyplot as mp
index2.plot(x="Age",y="expend",kind="bar",figsize=(9,8))
plt.xlabel("Age")
plt.ylabel("Expend")
mp.show()
```



Bar Graph representing age with expenditure as dependency

```
df.plot.scatter(x = 'Age', y = 'expend', s = 150);
```

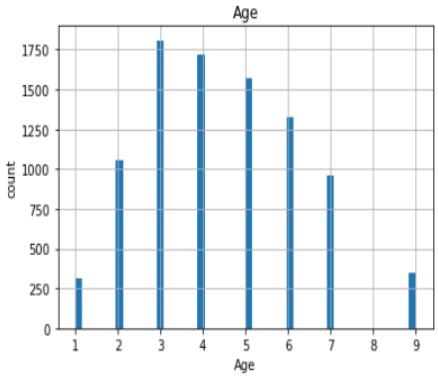


Scatter plot for age vs expenditure

Age is an important variable while analyzing the travel dataset and targeting the prospective set of admirers and it is represented by variable 1 which represents age group who are between the age group 1-15 years, 2 represents the 16-24 years age group, 3 represents 25-34 years age group, 4 represents 35-44 years group,5 represents 45-54 years, 6 represents 55-64 years,7 represents 65 and above age group, lastly, 9 represents the unknown age group. The people around the age group from variable 3 to 7 and the data was recorded during the summer with this analysis an opportunity can be seen to introduce schemes and exclusive packages to benefit the clients who come under the respective age group. Whereas even the expenditure spent by the age groups can also be seen in the scatter plot high returns come from the age group between 3 to 7.

```
df.hist(column='Age', bins=50);
plt.xlabel("Age")
plt.ylabel("count")
print("Age group who travel very often")
```

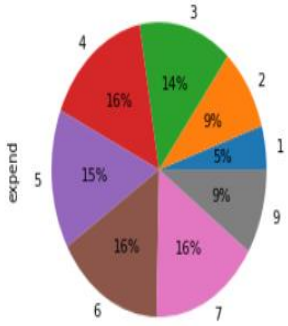
Age group who travel very often



Histogram representing age count

```
In [33]: index2.groupby(['Age']).sum().plot(
             kind='pie', y='expend', autopct='%1.0f%%')
```

Out[33]: <AxesSubplot:ylabel='expend'>



Pie Chart representing the age %

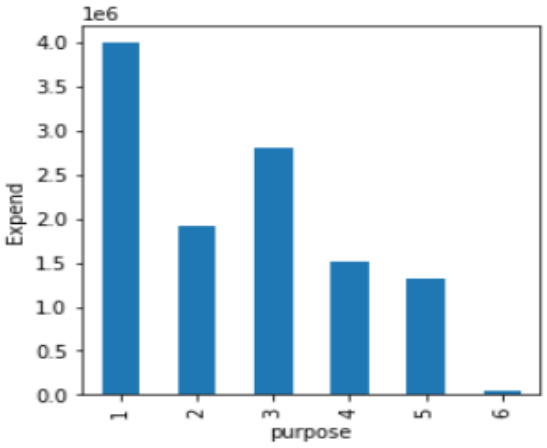
It can be seen that the age group with variable 3 are in a prominent and eye-catching end user of the services followed by variable 4 to 7 The age group with variables 1 and 9 are less targeted compared to the rest of the groups. And the percentage of the expenditure spent for the purpose can be observed which is an important insight for business development.

3.3 Purpose and expenditure required for that specific purpose.

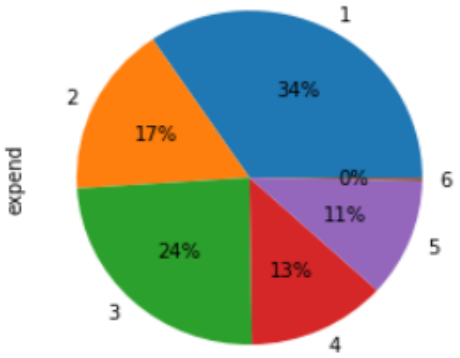
```
In [75]: index3=df.iloc[ : , [6,13]]
         index3=index3.groupby("purpose")['expend'].mean().round()
         index3
```

```
Out[75]: purpose
         1    3989935.0
         2    1917279.0
         3    2800079.0
         4    1504298.0
         5    1322950.0
         6      40664.0
         Name: expend, dtype: float64
```

Travel purpose is also a factor to consider while analysing the travel data. Here the variable is represented by the number 1 which represents a holiday, 2 represents the business, 3 represents study,4 represents VFR (Visiting friends and relatives), represents Miscellaneous and 6 represents transit. These are the average expenditure for the respective purposes.



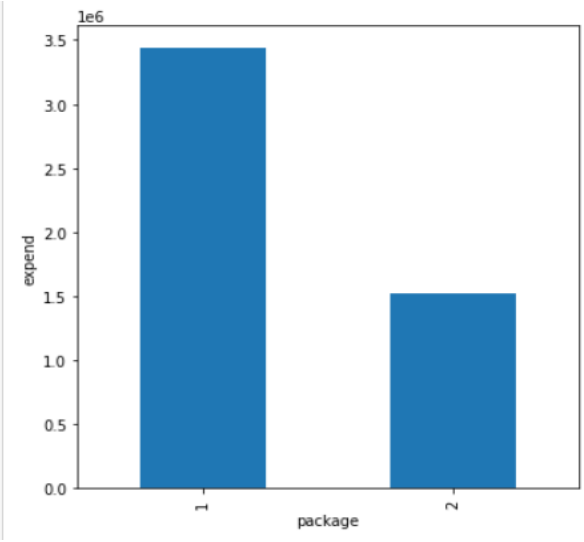
Bar graph for purpose vs expend



Pie Chart representing purpose

While looking at the bar graph we can see that the highest expenditure was spent by variable 1 which is for holidays with 34% and expense for study purposes is also an important potential with 24%. This specific analysis has shown various insights and business potential ideas which can be improvised for greater prominence. Using these insights and recommendations can be done for business purposes.

3.4 Estimate the average expenditure required for travel Independently and non-Independently



Bar plot representing the package

This variable clarifies the major part of the problem statement which can be used to resolve a greater problem here in the x-axis of the purpose 1 represents traveling independently and 2 represents traveling non-independently. By these statistics, the number of people who travel independently is lower so with that if the target is towards the higher client number the revenue of the travel sector will increase which will be a major perk to the business.

```
In [157]:  1 index6=j.iloc[ : , [7,13]].mean().round()
           2 index6

Out[157]: package          1.0
          expend     3441179.0
          dtype: float64
```

Average expenditure while traveling independently

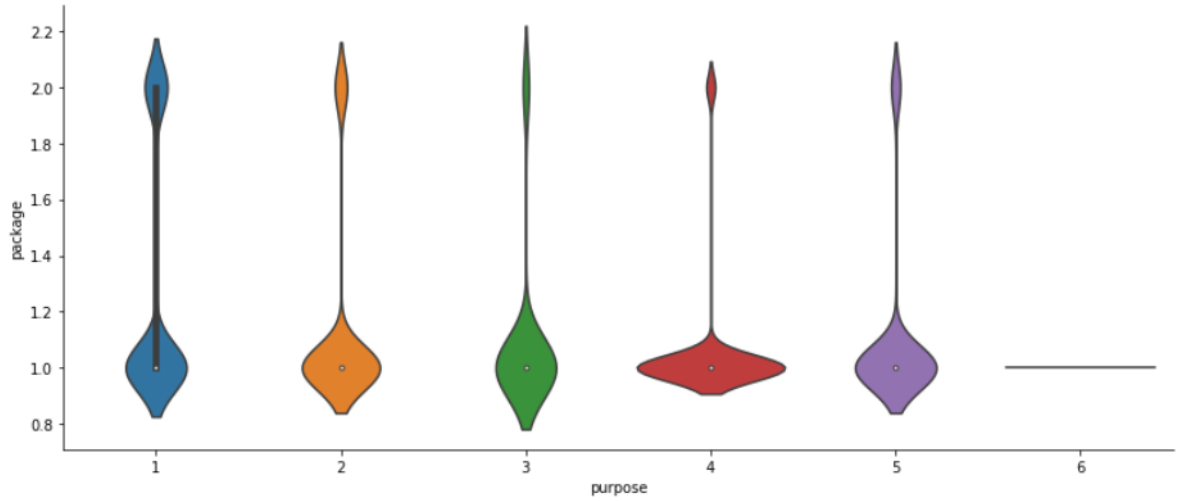
```
In [159]:  1 index7.mean().round()

Out[159]: package          2.0
          expend     3886246.0
          dtype: float64
```

Average expenditure while traveling non-independently

Here as per the problem statement, we found the average expenditure while traveling independently and non-independently, and as shown in the above figures we took 7684 rows that had independent package variables and 1400 rows that had non-independent with that we saw that people who travel non-independently are very minor so with these insights we can target those people with the accurate proposal.

3.5 Compare purpose vs package

As per the previous insights we have analysed the purpose and expenditure required for that we have also discussed the packages and expenditures for traveling with and without a package. So, with all that insights, a new problem statement was raised Do the purpose and package correlate?



In the above figure, the y-axis represents the package that 1 means traveling independently and 2 means traveling non-independently. In the y-axis 1 represents a holiday, 2 represents the business, 3 represents study,4 represents VFR (Visiting friends and relatives), represents Miscellaneous and 6 represents transit. With all these, we have plotted a violin graph which hybrid of a box plot and a density plot where the white dot represents the median. We can see that purpose 1 primarily uses packages more than the other purposes so we can see that even if the major package is independent if we target the proper audience, we can increase the turnovers(seaborn.violinplot).

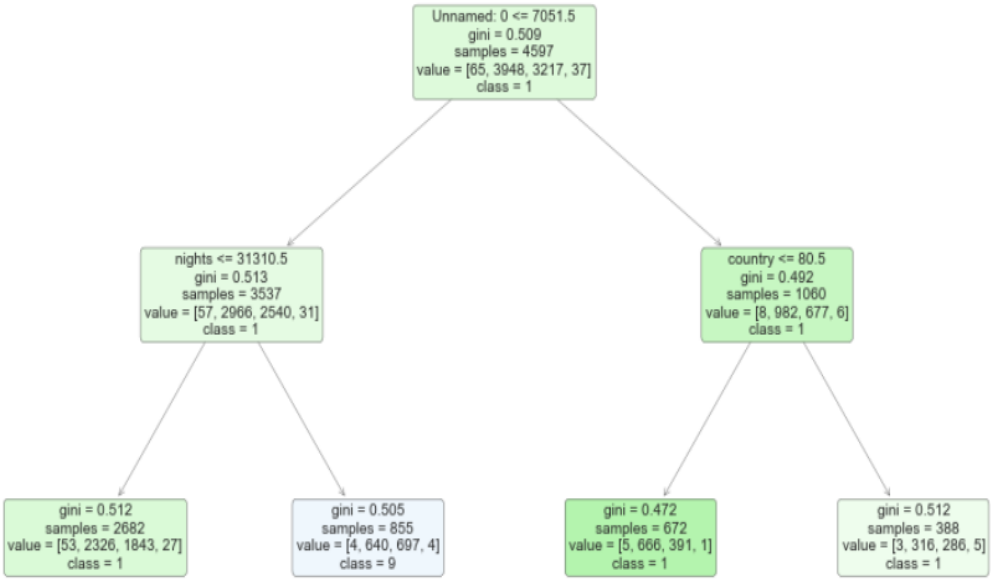
4. Model Prediction

4.1 Random Forest Model

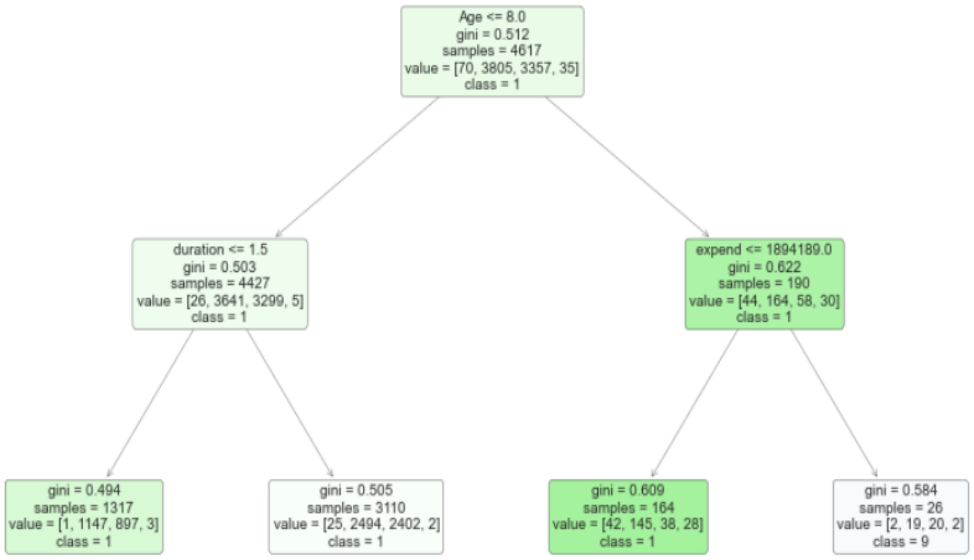
Now I am using a random forest algorithm to predict the accuracy of the model A random forest is a type of supervised machine learning algorithm built using decision tree algorithms. This algorithm is used to anticipate results in a variety of sectors, like statistics, and retail. It forecasts by averaging or averaging the output of several trees. Increasing the number of trees improves the outcome's accuracy. As this is a decision tree algorithm it has three layers root node, sub-node also called decision node, and leaf node.

In our model, we have primarily used the model to construct a decision tree with depth two which means that process will take place in two levels level 1 and level 2(Introduction to Random Forest in Machine Learning. (n.d.).

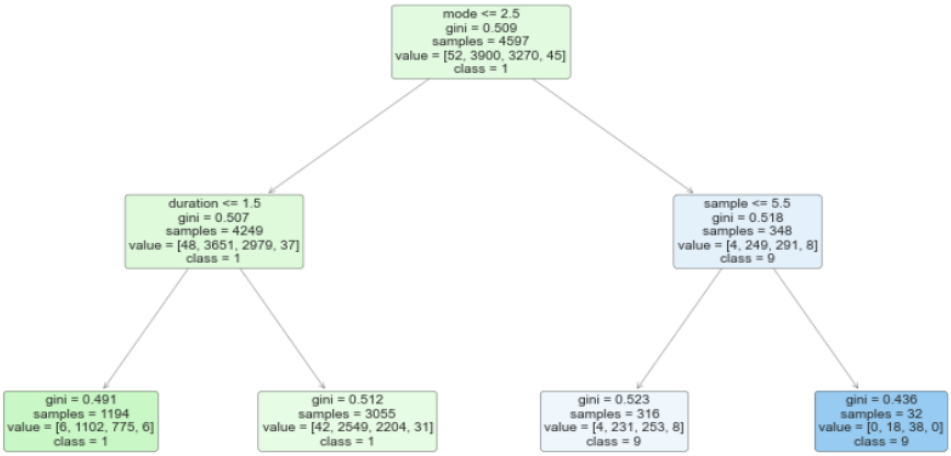
```
DecisionTreeClassifier(max_depth=2, max_features='auto',
                       random_state=1608637542)
```




```
DecisionTreeClassifier(max_depth=2, max_features='auto',
                       random_state=1273642419)
```



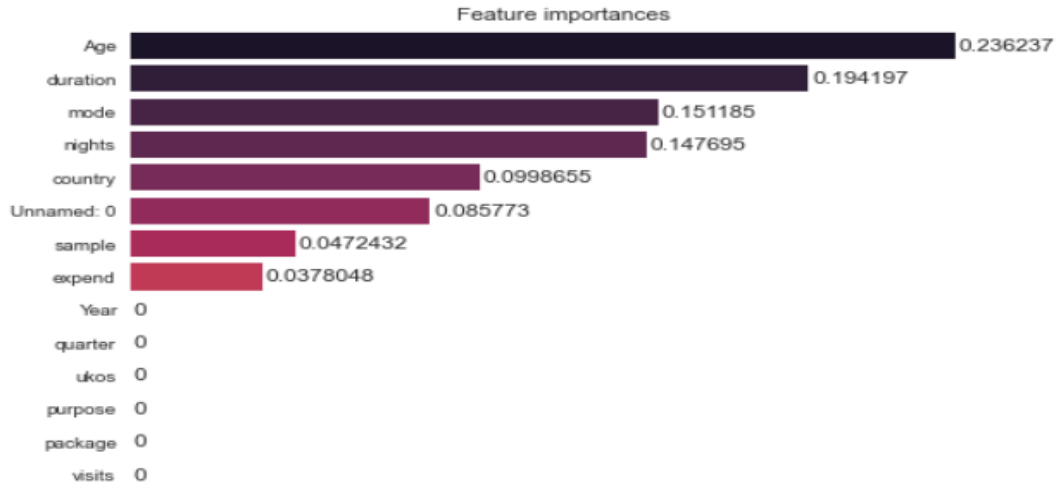
```
DecisionTreeClassifier(max_depth=2, max_features='auto',
                       random_state=1935803228)
```



In [263]:
```
from sklearn.metrics import classification_report, confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(classification_report(y_test,y_pred))
```

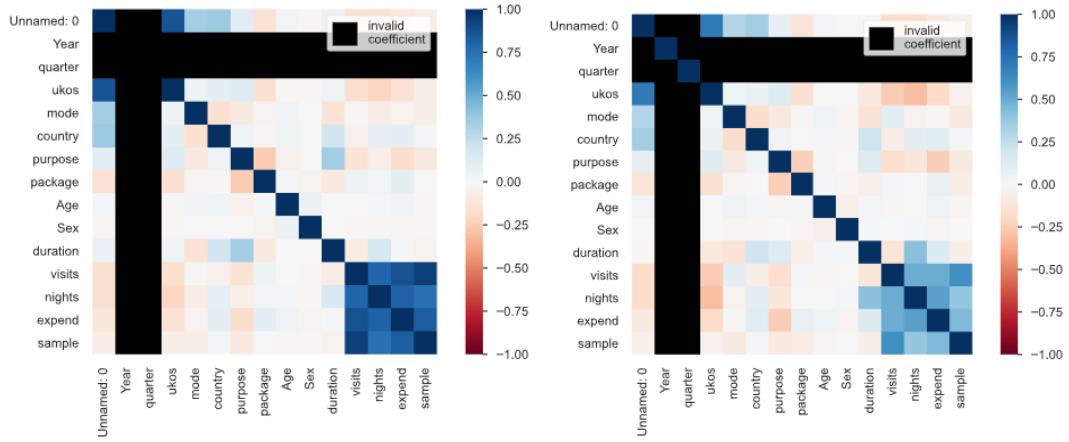
```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        15
           1       0.54      0.99      0.70       983
           2       0.53      0.01      0.02       806
           9       0.00      0.00      0.00        13

    accuracy                           0.54      1817
   macro avg       0.27      0.25      0.18      1817
weighted avg       0.53      0.54      0.39      1817
```

This model has got an accuracy score of 54% in predicting travel with gender as the factor for the prediction of the model. Then I computed the feature importance for the model to get the proper insight (. sklearn.model_selection.train_test_split. (n.d.).



Feature Importance will be important when addressing a classification problem, it will be beneficial in feature selection by locating the most significant characteristics. By observations, we have seen that this feature acts differently for diverse datasets (Rogers.et.al(2006)).

4.2 Correlation Model



Pearson's r correlation **Kendall's T**

In this, we have used Pearson's r and Kendall's T correlation models where Pearson's shows the linear correlation, and this is parametric. In the above figure, dark blue indicates a positive correlation, and the light color indicated a negative coefficient. Similarly, Kendall's T model is a non-parametric statistics model. The indicators are like Pearson's model. With this, we have analysed and noticed the variables which are correlated with each other. The advantage is that correlational research can be undertaken on factors that can be assessed but not modified, such as when doing an experiment that would be impractical and non-ethical (Shibboleth Authentication Request. (Peter, et.al(2011 Jan)).

5. Ethics

While working with the data, specific criteria must be observed, with ethics being one of the most crucial factors to take into account. Since the data set I utilized is connected to the tourism industry, ethics was a crucial factor to take into account. I did so by upholding moral principles both before and after planning. Age, resident's nationality, gender, and trip expenses are the variables I manage in my project. Although an individual wouldn't be impacted by the data I manage because it is open source, I nevertheless upheld ethical standards when processing and using the data. The UK data ethics framework, which adheres to the three principles of transparency, accountability, and fairness, is where I found the standards I used to work with this data. (Data Ethics Framework. (2020)).

5.1 Transparency

The data I worked with is a public source and working on it does not compromise an individual's privacy. The consent was obtained as the data was being recorded. Manipulation such as addition and subtraction were not performed on the data; the only adjustments done

were cleaning and replacing null strings in the sex column. This was noted when discussing the data.

5.2 Accountability

The issue statement I picked created a clear and relevant result, and I have enough evidence to support the change I made for my study. The data I utilized was obtained from the ONS website and did not include any noise or null values. The visualization output was responsive after compilation. This demonstrates that I ensured and followed the tailored requirements specified while handling the data.

5.3 Fairness

I've made certain that I haven't altered the existing data; instead, I've gone over the dataset painstakingly and deliberately picked the columns required for future study. The variables were treated using the bias factor for analysis when processing the data. The data was solely utilized for this research, and its integrity was maintained throughout the process.

6. Recommendations for Travel Sector

- There are a lot of people who like to travel individually, therefore it is necessary to offer programs that encourage them to view booking a package as a preferable option.
- Since students aiming for global exposure are more in number, study consultancies should collaborate with travel consultancies.
- The age range of 15 to 45 should be the focus of advertisements because they travel at the highest rate. Social media marketing can be focused on increasing returns. The travel industry in the UK market has enormous potential since UK people spend more money on it.

7. Conclusion

From this analysis where I discovered problem statements that were piqued while looking at the data then I did the analysis of the data and plotted visualization figures for the respective problem statements. Here from this, we cleaned and removed the noise from the data later we calculated the expenditure for UK residents and non-UK residents where we saw that number of UK residents was more which has a potential for business development, the expenditure spent by UK residents is also huge. Analysis was done with respect to an age group where we found the age groups to consider doing the marketing and making potential consumers. The purpose for which people usually travel is for holidays with the visualization and figures we can say that the holiday sector is the field to consider. With this, I conclude that this is a prominent business insight that can be used for purposeful development.

8. References

1. Travelpac: travel to and from the UK - Office for National Statistics. (2022, Nov). *https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/travelpac*

2. pandas. (2022, Nov). PyPI.

https://pypi.org/project/pandas/

3. Introduction to Random Forest in Machine Learning. (n.d.). Engineering Education (EngEd) Program | Section.

Https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

4. Peter Y. Chen & Paula M. Popovich. (2011,Jan). Login.uoelibrary.idm.oclc.org. Retrieved December 18, 2022 https://login.uoelibrary.idm.oclc.org/login?qurl=https://methods.sagepub.com%2fbook%2fcorrelation

5. Data Ethics Framework. (2020). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923108/Data_Ethics_Framework_2020.pdf

6. Rogers, J., & Gunn, S. (2006). Identifying Feature Relevance Using a Random Forest. SpringerLink. https://link.springer.com/chapter/10.1007/11752790_12?error=cookies_not_supported&code=1b2052dd-b305-4467-b186-770189fb5436

7. seaborn.violinplot — seaborn 0.12.1 documentation. (n.d.). https://seaborn.pydata.org/generated/seaborn.violinplot.html

8. pandas-profiling dev documentation. (n.d.). https://pandas-profiling.ydata.ai/docs/master/pages/getting_started/quickstart.html

9. sklearn.model_selection.train_test_split. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

10. Matplotlib Scatter. (n.d.). https://www.w3schools.com/python/matplotlib_scatter.asp