2018

UNITED KINGDOM TOURISM – AN ANALYSIS ON SPEND BY OVERSEAS RESIDENTS

DIVYA BALASUBRAMANIAN
MSC DATA SCIENCE
CITY, UNIVERSITY OF LONDON

1. Analysis Domain, Questions, Plan

1.1. Domain Overview

According to visitbritain.org's Britain's economy facts, the UK tourism has seen a massive boom since 2010 and represented 9% of the UK's GDP in 2013. It is evident that a spend from overseas residents traveling to the UK ('visitors') has a direct impact on this industry. While the tourism data has been weighted using multiple criteria (*Office for National Statistics, 2016*), the comprehensive nature of the dataset encourages a deep dive which can potentially provide targeted promotion of tourism.

1.2. Domain Motivation

This project aims to investigate the causes that lead to the increase in visitor spend (in £) in the UK as well as the effects of this spend. One of the reasons for the former could be the currency exchange rate. A favourable currency exchange rate (*FxTop.com, 2018*) can encourage visitors to travel to the UK. For the latter, sectors such as accommodation, retail, etc., are those that are immediately impacted by tourism. It would be particularly interesting to see the impact on the retail sector (*Office for National Statistics, 2018*) due to its enormous coverage. The retail sector contributes to 5% of the UK's economic output (*House of Commons Library, 2018*). These perspectives contributed to the questions, listed below, this project is trying to answer.

1.3. Analysis strategy and plan

This project focuses on the data from 2016Q1 to 2018Q2 for running models due to the significant size of the dataset. A substantial amount of data wrangling (Section 1 of the ipynb) had to be performed to make the dataset suitable for analysis. Following this, extensive exploratory data analysis was performed, and the findings are detailed in Section 2 of the ipynb. Given the myriad of analytical possibilities within this dataset, the following questions have been derived to narrow down the approach. This can be broadly divided into three parts:

1. Regression Analyses specific to tourism data:
    - Using linear regression, can the spend for 2018Q3 be predicted?
    - Using multiple linear regression, what combination of the profile of the traveller, the visits they make and nights they stay determine the most accurate spend in the UK?
    - Using Logistic Regression, can the age of the traveller and the duration they're visiting for determine the sex of the traveller?
2. Regression Analysis using retail revenue data:
    - Using multiple linear regression on year and spend from travel, can the retail revenue be predicted?
3. Causal effect of Currency Exchange Data:
    - How does the currency exchange rate of each quarter in selected countries determine the spend by visitors from that country?

2. Findings and Evaluation

2.1. Regression Analyses specific to tourism data

To ensure the outliers do not skew the analysis, the spend, visits and nights were normalised.

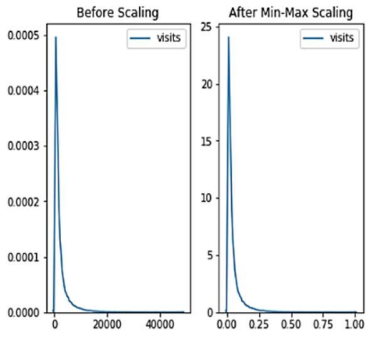

Figure 1: Normalisation of visits

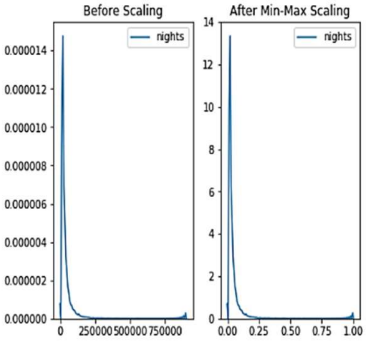

Figure 2: Normalisation of nights

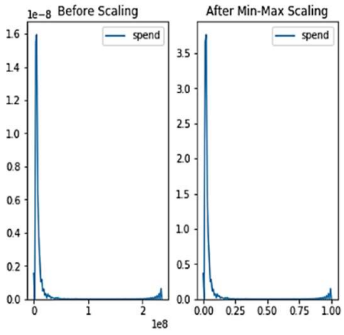

Figure 3: Normalisation of spend

2.1.1. Linear Regression Model

As seen in Figure 4, there is a spend pattern across the quarters with a spike in Q3 of each year.

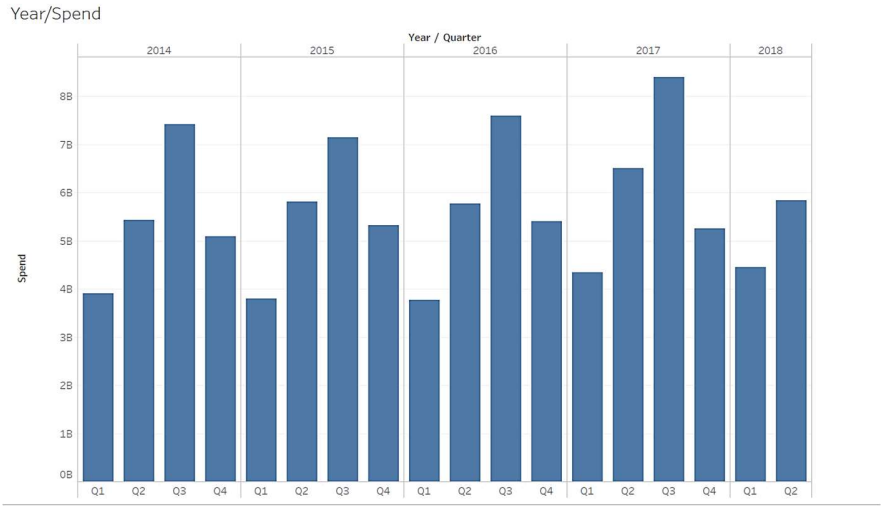

Figure 4: Spend (in billions) from visitors across the quarters between 2014 - 2018

To predict the 2018Q3 spend, a linear regression was run between period and spend (Model 1). The r2 value of the model was 100%. However, upon cross-validation the model generated an r2 of -223.9%. Due to the limited data used to fit the linear regression model, there is a possibility that the model was over-fitted. There are several factors influencing spend, it would hence be favourable to consider as many variables as possible for prediction.

2.1.2. Multiple Linear Regression Models

As seen in the correlation matrix (Figure 5), visits, nights and spend have high correlation with each other.

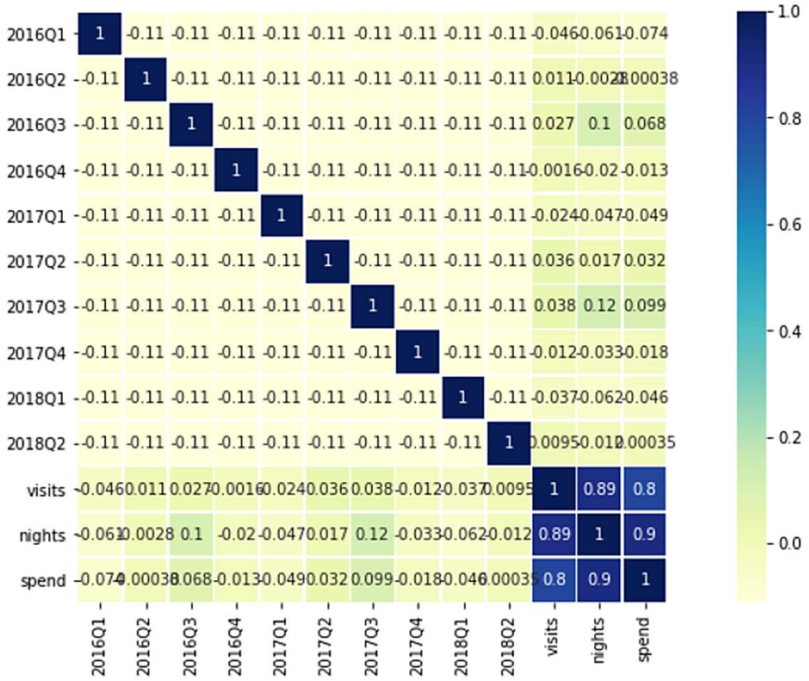

Figure 5: Correlation matrix for the data frame considered for Model 2

In Figure 6, considering USA in 2017, the spend was high in both visits/spend as well as nights/spend, however there were greater number of visits from France. This implies both visits and nights mutually contribute to spend.

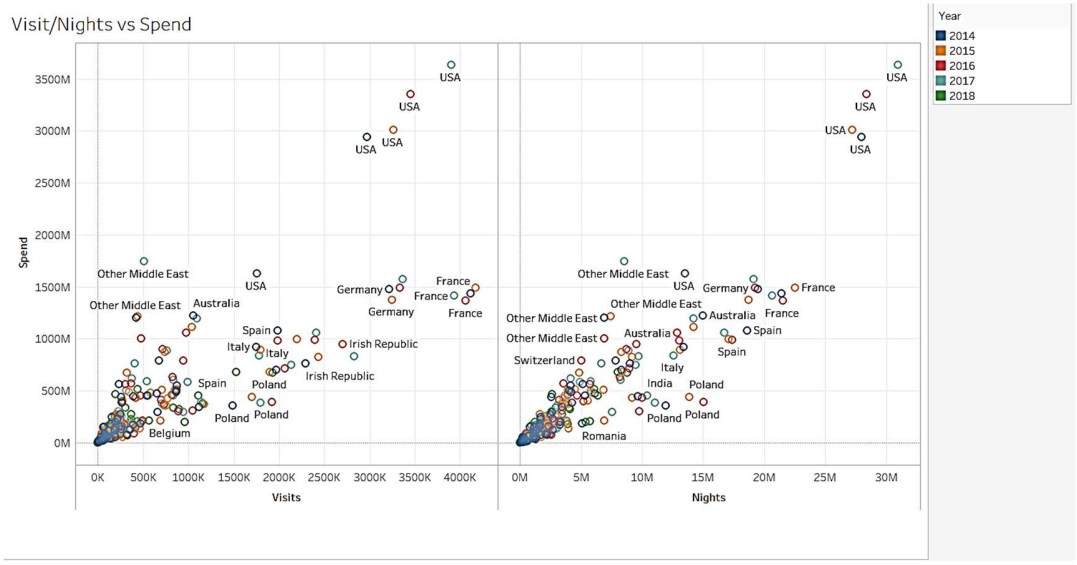

Figure 6: Scatter plots of visits/spend and nights/spend across the years detailed on country

The above finding led to a multiple linear regression considering period, visits and nights to predict spend (Model 2). The r2 value of the model was 81.9% with cross validation r2 at 81.1%. This model observes a larger amount of data thereby improving its reliability.

Since visitor spend is of primary interest, in order to narrow down the analysis, top 10 countries (from Figure 7) from where visitors traveling to the UK spend more were focused.

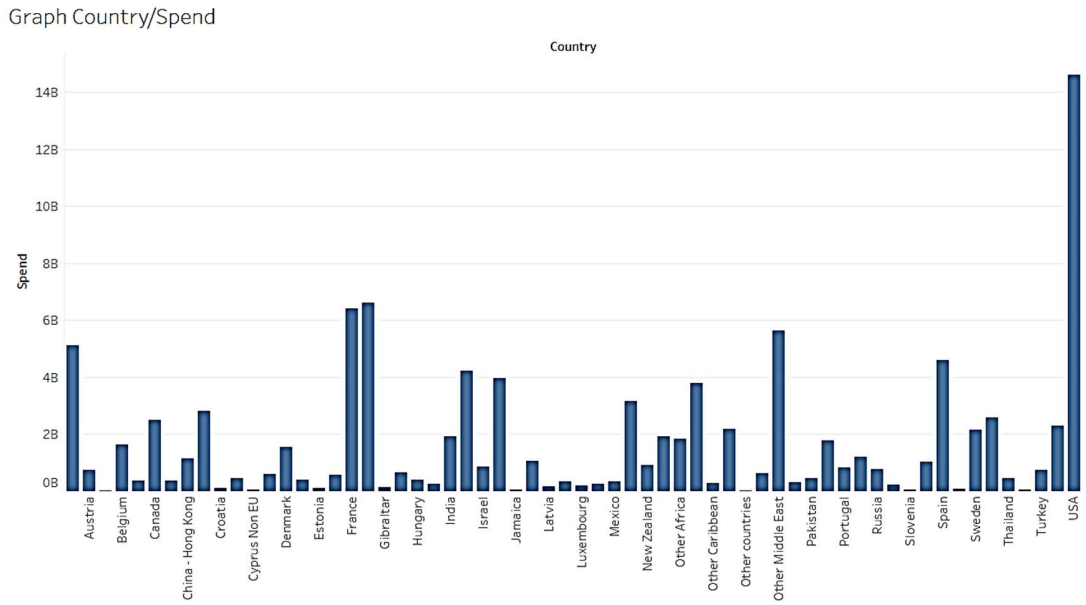

Figure 7: Countries plotted by spend

A multiple linear regression was run considering country, period, visits and nights to predict spend (Model 3). The r2 value of the model was 92% and cross validation r2 at 89.3%. It can be noticed that the r2 value gradually increased with increasing number of features considered in each of the models.

Error Metrics:

Due to the robust nature of such moving data with large number of outliers, Mean Absolute Error (MAE) was calculated for the models employed (*DATAQUEST, 2018*). The models generated the following MAE values:

Model	MAE (with normalisation)	r2 (without normalisation)	MAE (without normalisation)
Model 1	0.0000001	100%	0
Model 2	0.066	81.9%	5.84 million
Model 3	0.0607	92%	14.32 million

Given the scale of spend values, greater number of features and larger amount of data considered, the MAE values appear to be consistently low. The r2 values of the models without normalisation (provided in the supplementary ipynb) are almost the same as those with normalisation. It is thus proven that for this dataset, the feature values can be considered as is.

2.1.3. Logistic Regression Model

Identifying the gender of the traveller would enable the retail sector to position their products accordingly. To predict if the traveller is male or female, a logistic regression was run considering two independent variables, age and duration. As seen in Figure 8, the precision of the model was an unfavourable 0.51 which makes the model unfit for prediction. This model does not have the discrimination capacity between male or female.

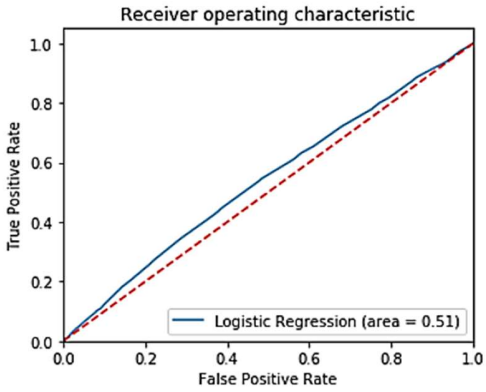

Figure 8: ROC curve for the Logistic Regression to predict gender of the traveller

2.2. Regression Analyses using retail revenue data

In order to consider substantial amount of data to build this model, historical data from the year 1996 for both retail revenue and tourism were collated to run a multiple linear regression between period and spend to predict retail revenue. Figure 9 shows that both spend as well as retail revenue have an imminent pattern.

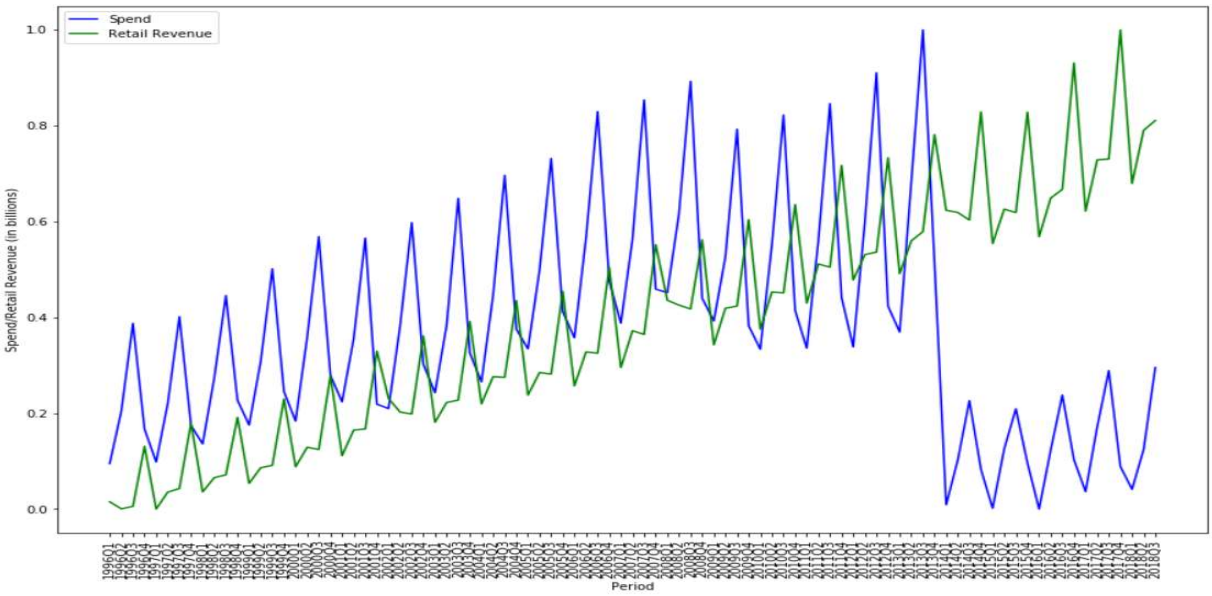

Figure 9: Period vs Spend/Retail Revenue

The r2 value of this model was 85.9% and cross validation r2 was 85.88%. The values suggest a reasonably high correlation between the tourism spend and retail revenue. The MAE was 0.026 which is low. However, gathering further information such as section of the retail sector visitors tend to spend on would assist building a more informed model.

2.3. Causal effect of Currency Exchange rate Data

As explained in Section 6 of the ipynb, visitor spend and currency exchange rate data were investigated for specific countries. The visitor spend from 2014Q1 – 2018Q2 are projected in Figure 10.

Figure 10: Variation in country-wise spend by quarter from 2014Q1-2018Q2

The comparison study between an increase in spend by visitors from specific currency zones and the change in currency exchange rates demonstrates a potential causal relationship between the two. A consistent favourable currency exchange rate could have encouraged travellers to visit the UK. For example, in 2017 Q3, the exchange rate was in favour of Euro (Figure 8). Given the high affordability of this currency at the time, school holiday period and short duration of travel from Europe to UK, there is a possibility that a greater number of travellers from Europe (using Euro currency) have travelled to the UK thus reflecting the increase in 2017 Q3 spend.

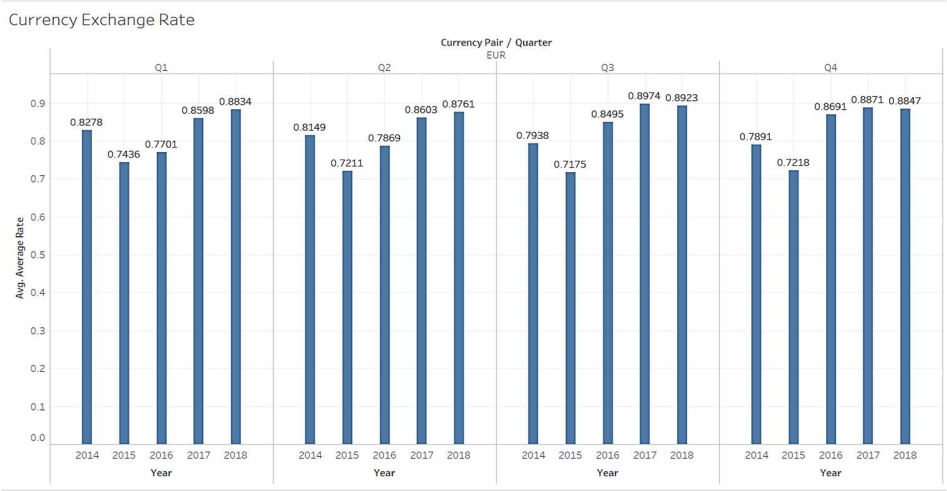

Figure 11: Variation in EUR/GBP Currency Exchange Rate across the quarters from 2014Q1-2018Q4

3. Reflection and Recommendations:

Based on the analyses performed, this project can support retailers in positioning their products to a targeted market (E.g.: USA) to increase their revenue. Also, the predicted visitor spend derived from the models can be utilized by currency traders to hold more stock of GBP to support increasing demands.

Due to the seasonal nature of tourism, the consideration of data on a quarterly basis is critical. Seasonality here refers to periodic fluctuation that occurs based on season (*Investopedia, 2018*).

The use of one hot encoding mandates the mention of curse of dimensionality. Were the scope of this project widened, principal component analysis could be performed for dimensionality reduction or grid search to identify and retain the ideal parameters.

A limitation to this project is that several assumptions (such as currency zones) were made in parts to perform a comprehensive analysis on a narrowed down scope. A detailed analysis considering all currency zones may provide better insights. For further work, I would be interested in utilising the other features of the dataset such as mode and purpose of visit to analyse the impact on spend. Another potential scope is to predict mode of travel using cluster based logistic regression as performed by Li et al. (2016) or using other machine learning models such as Random Forest or Naive Bayes.

Link to Public html:

https://smcse.city.ac.uk/student/aczd113/INM430/Divya_Balasubramanian-PODS.html

References:

Office for National Statistics (2016) *International passenger survey methodology*.

Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/methodologies/internationalpassengersurveymethodology

(Accessed: 25 October 2018).

FxTop.com (2018) *Historical rates*

Available at: http://fxtop.com/en/historical-exchange-rates.php?MA=1

(Accessed: 8 December 2018)

Office for National Statistics (2018) *Retail sales pounds data*

Available at: https://www.ons.gov.uk/businessindustryandtrade/retailindustry/datasets/poundsdatatotalretailsales/current

(Accessed: 29 October 2018)

House of Commons Library (2018) *Retail sector in the UK*

Available at: https://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN06186

(Accessed: 4 December 2018)

DATAQUEST (2018) *Understanding Regression Error Metrics*

Available at: https://www.dataquest.io/blog/understanding-regression-error-metrics/

(Accessed: 5 December 2018)

Investopedia (2018) *Seasonality*

Available at: https://www.investopedia.com/terms/s/seasonality.asp

(Accessed: 13 December 2018)

Li et al. (2016) 'Cluster-Based Logistic Regression Model for Holiday Travel Mode Choice', *Procedia Engineering*, Volume 137, pp. 729-737. doi: 10.1016/j.proeng.2016.01.310