Investment Opportunities in Irish Tourism: A Geospatial and Market Forecast Analysis

Introduction

## Problem Domain and Motivation

The Irish tourism industry is very important element in the national economy. In 2019, the Irish tourism industry generated €10 billion, with 73% coming from inbound tourists. The industry also contributed 4.4% to the national GVA and directly employed 284,800 people, representing 13% of the national workforce (Irish Tourism Industry Confederation, 2019).

Considering these figures, we can see that the Irish tourism industry presents significant investment opportunities. However, stakeholders often lack localised insights into the current infrastructure and future demand trends.

Existing data provides a broad overview of the industry, but it does not take into account the regional variations in tourism potential across Ireland. This gap in understanding makes it difficult for investors and businesses to make informed decisions on where to allocate resources or how to adapt to fluctuations of demand patterns. Addressing this problem requires a comprehensive analysis of both geographic clusters and future demand forecasts to support strategic investment decisions in the tourism sector.

This research is created for providing actionable data for investors, stakeholders, managers and policymakers, enabling them to make informed decisions in this sector. In our first objective, geo clustering, we will analysis the geographic distribution of tourism-related sectors across Ireland. This will help businesses associate strategies to specific regions, identifying high-potential zones for investment. The second part, time series analysis, will forecast the financial trajectory of the Irish tourism industry over the next five years. By examining key financial indicators, we aim to provide a reliable estimate of future demand, helping businesses align their growth strategies with expected market conditions. Both part together will create a comprehensive report of current opportunities and future trends in Irish tourism, supporting strategic investment decisions.

## Objectives

**Major objective**

The goal of this project is to investigate the investment potential of the Irish tourism industry over the next five years, considering key sectors related to tourist services. As a result, we will develop a report that displays the current state of tourism infrastructure in each zone and predicts demand for the next five years.

Using this report, the reader will be able to identify investment opportunities, whether it's starting a new business or adjusting an existing one in a specific zone. Additionally, the report will provide insights into how the current infrastructure is expected to meet future demand.

To reach major objective we need to define 2 minor objective which align with our CA requirements:

**Spatial clustering**

Every business requires demographic research, competitive analysis, and demand assessment. In this section, we will work with a dataset that contains geospatial information about significant places related to the tourism industry in Ireland, along with descriptions of each place, such as restaurants, pubs, museums, and more.

The dataset provides sufficient information to form a general overview of the Irish tourism industry. However, it is quite likely that the overall insights drawn about Ireland as a whole may not be equally reflected in specific zones. In other words, we need to assess how accurately the general "ingredients" for tourism (akin to ingredients for a pizza) are represented in each "slice" or zone.

To understand how accurately the individual zone is reflecting the general ratios of Ireland, we will apply clustering models and compare the ratio of attraction types within each cluster to the general ratio of the whole Ireland. Continue using the pizza analogy, we will slice the pizza in the most effective way and analyse how accurately the ingredients in each slice matches the overall ratio of the entire pizza.

This approach can vary depending on different business requests and needs. Large businesses with widespread coverage, such as distribution companies or retail chains, might require large clusters such as county level. However, smaller cluster sizes, such as city or town levels, would be more beneficial for small companies, such as a single restaurant,a hotel, or a pub.

**Time series analysis**

In the second part of our report, we will forecast future demand. In other words, we will to estimate how much revenue the Irish tourism industry will generate over the next 5 years. To achieve this, we will forecast 4 indexes that are directly related to the Irish tourism industry and available on Yahoo Finance: Ryanair's share prices, Aer Lingus's share prices, Dalata Group's share prices, and the ISEQ Index.

I believe that the averaged forecast of those four time series will provide a reliable representation of the real value of the industry's future demand. Once we obtain the projected demand for the Irish tourism industry, we can apply that value to the results of our first section, allowing us to complete the report.

## Data Sources

For our CA we will use 6 databases:

**1. Attractions dataset**

This dataset was obtained from Ireland's open data portal (<https://data.gov.ie/dataset/attractions>). It contains almost 6k records about the most significant places and tours in Ireland. The dataset includes multiple features with data related to each attraction. However, for our purposes, we will only need the geospatial data (latitude and longitude) and the "Tags" column, which contains descriptions of the attractions.

**2. Cities names**

We will use this dataset only for associating clusters with specific cities or towns that are closest to the centroids of our clusters. The dataset was obtained from (<https://simplemaps.com/data/ie-cities>).

**3-6. Yahoo Finance**

Other 4 datasets we will get from Yahoo Finance website using library "yfinance".

* **Ryanair's Share Prices:** This dataset includes historical stock price data for Ryanair Holdings. It is one of the main low-coster in Europe. We can assosiete companie's performance with the overall health of the airline industry as well as how it contributes tourism.
* **Aer Lingus's Share Prices:** Similarly to Ryanair will will also include Aer Lingus. Aer Lingus is another important airline in Ireland. Its stock performance can reflect the demand for air travel to and from Ireland.
* **Dalata Group's Share Prices:** The Dalata Hotel Group is the largest hotel operator in Ireland. As the Dalata Group operates a significant portion of the hotel and accommodation sector, its stock performance can serve as a proxy for the state of the hotel industry and overall tourism demand in Ireland.
* **DISEQ Index:** The ISEQ Index is the benchmark stock market index of the Irish Stock Exchange. The ISEQ index reflects the broader performance of Ireland’s economy, which can include tourism-related businesses.

## Methods

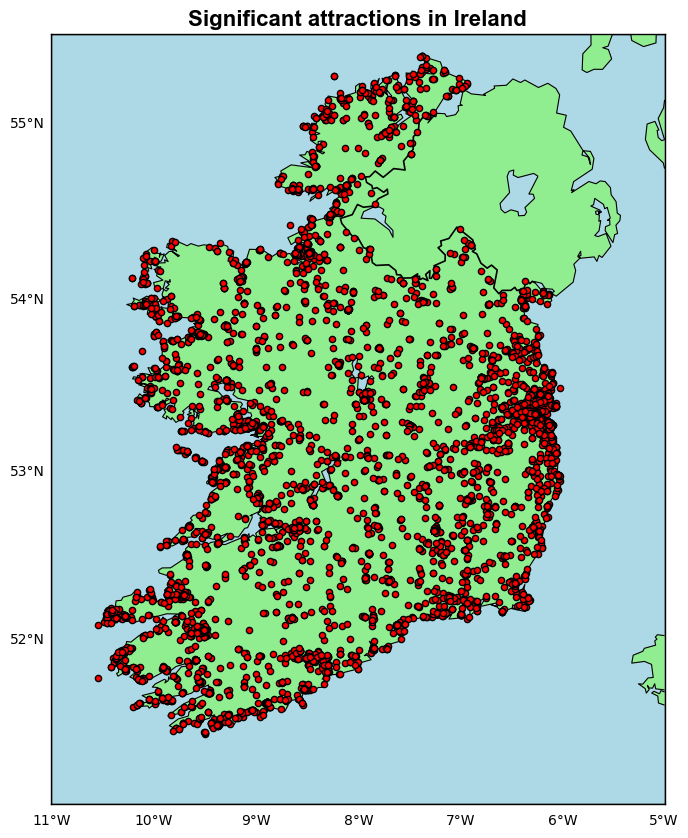
In this project, we will primarily focus on theoretical review and model selection. Although we will explore multiple models and methods, only a few will be applied to our data specifically those that demonstrate the best performance.

1. Clustering:
   * K-Mean
   * DBSCAN
   * Agglomerative
   * OPTICS
   * BIRCH
   * Affinity Propagation
   * Gaussian Mixture
2. Time-Series analysis:
   * Exponential Smoothing
     + Simple Exponential Smoothing
     + Holt's lineal
     + Holt's Winters
   * ARIMA
     + Autoregression
     + Rolling avarage
   * SARIMA
3. Evaluation methods:
   * Elbow method
   * Silhouette score
   * David-Bouldin index
   * Dendrogram
   * Calinski-Harabasz score
   * Ljung-Box testing
   * Autocorrelation function (ACF)
   * Partial Autocorrelation (PACF)

## Clustering

### Geodata overview

Geo position of all 5956 records on an Irish simplified map.



### NLP

Extract all separate words from feature "Tags"

Unique tags:  
['Activity' 'Food' 'and' 'Drink' 'Experience' 'Restaurant' 'Fast'  
 'Operator' 'Tour' 'Cafe' 'Shops' 'Shopping' 'Walking' 'Fishing' 'Angling'  
 'Nature' 'Wildlife' 'Golf' 'Course' 'Attraction' 'Historic' 'Houses'  
 'Castle' 'Church' 'Abbey' 'Monastery' 'Churches' 'Abbeys' 'Ruins'  
 'Centres' 'Department' 'Store' '' 'Transport' 'Coach' 'Road' 'Tracing'  
 'Your' 'Ancestors' 'Museums' 'Learning' 'Cycling' 'Bike' 'Rental' 'Zoos'  
 'Aquarium' 'Craft' 'Bird' 'Watching' 'Climbing' 'Horse' 'Riding'  
 'Equestrian' 'Kayaking' 'Kitesurfing' 'Windsurfing' 'Surfing' 'Swimming'  
 'Pools' 'Water' 'Park' 'Venue' 'Island' 'Offshore' 'Artisan' 'Sailing'  
 'Adventure' 'Boat' 'Pampering' 'Health' 'Farm' 'Spa' 'Wellness'  
 'Specialised' 'Retreat' 'Local' 'Produce' 'Movies' 'Cinema' 'Zip'  
 'Lining' 'Fitness' 'Leisure' 'Pool' 'Sports' 'Venues' 'Driving' 'Range'  
 'General' 'Art' 'Gallery' 'Music' 'Day' 'Embarkation' 'Point'  
 'Agriculture' 'Public' 'Sculpture' 'Seafood' 'Falconry' 'Natural'  
 'Landscape' 'Gardens' 'Garden' 'Visitor' 'Trails' 'Photography' 'Pubs'  
 'Bar' 'Cooking' 'Cookery' 'Fine' 'Dining' 'Comedy' 'Vegetarian'  
 'Traditionally' 'Irish' 'Beach' 'Literary' 'Ireland' 'Gardening' 'Vegan'  
 'Forest' 'Walk' 'Pitch' 'And' 'Putt' 'Marina' 'Cruising' 'River'  
 'Discovery' 'Race' 'Gaa' 'National' 'Covid' 'Safety' 'Charter' 'Stadium'  
 'Casinos' 'Banquet']

Tranform tags into “dummy-like” columns

Food & Drink Outdoor Activities & Adventure Culture & History Wellness & Leisure Shopping & Purchase  
0 1 1 0 0 0  
1 0 1 0 0 0   
2 1 1 0 0 1   
3 0 1 0 0 0   
4 0 1 0 0 0

## Define optimal number of clusters

### Theory overview

#### Elbow Method

Inertia measures how tight data points stick to the centroid in each cluster. The Elbow method defines optimal number between Interia and number of cluster. We are going to use function "KneeLocator" from "kneed" library, which will apply following equation.

Bholowalia P. and Kumar A. (2014) reviewed in details a use elbow method to define optimal number of clusters particularly for K-Mean model.

#### Silhouette Score

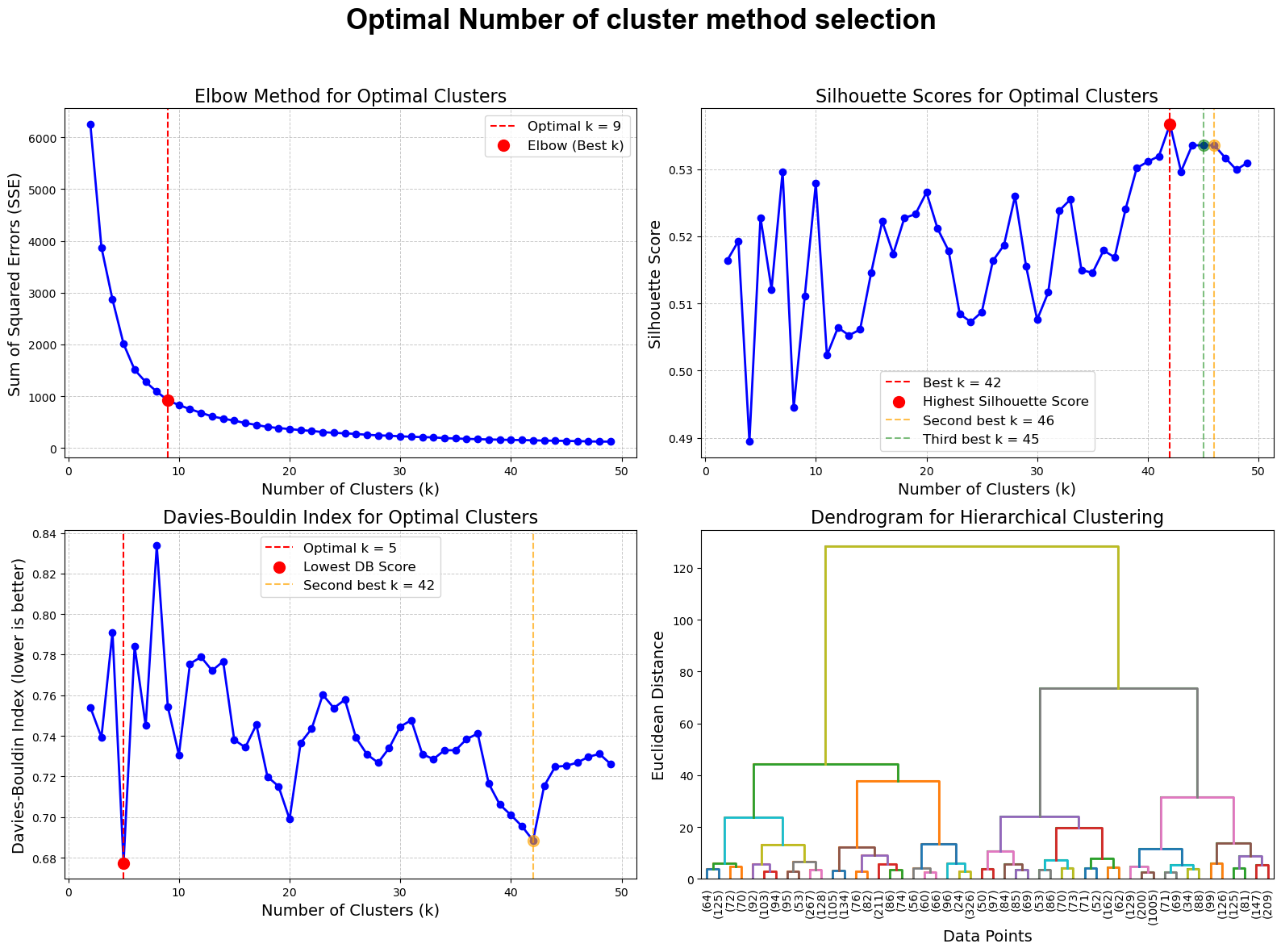
Silhouette Score measures each data point's similarity between its own cluster "cohesion" and other clusters "separation". Clear explanation how Silhouette Score can be applied for defining optimal number of clusters and evaluation is provided in study of Shutaywi M. and Kachouie N. (2021). In this CA we will use package "silhouette\_score" provided by "sklearn.metrics" library.

#### Davies-Bouldin Score

Davies-Bouldin Score measures similarity ratio between clusters, considaring intra-cluster similarity (distanse within each cluster) and inter-cluster difference (difference between clusters). Hidayat R. et. al. (2022) in their research applied evaluation by Davies-Bouldin Score with Agglomerative clustering for computer vision solution. Similarly, to previos method we are going to use "davies\_bouldin\_score" package provided by "sklearn".

#### Dendrogram

Dendogram is produced by hierarchical clusters and has tree-like structure. Briefly, it assigns 2 closest data points as a cluster and continues agglomeration until there are only 2 clusters left. We can use gap (largest vertical distance) to define the optimal number of clusters. A detailed overview of this method can be found in reaserch provided by Sisodia D. et al. (2012).



On some plots we displayed a few options for defining optimal number of clusters depending on business needs.

## Model selection

### Theory overview

In this section, we will briefly overview theory of some clustering algorithms and apply hyperparameters tunning combined with evaluation metrics to filter the best parameters for each model.

The reason why I have selected following model is the comparison between clustering algorithms based on different approaches such as density, centroids and hierarchy.

I think that each clustering model has its own application and benefits fir specific business needs. We will discuss those needs in the section with visual evaluation.

#### K-Means

K-Means clustering algorithm is centroid based used to group data into k clusters by minimising the variance inside of each cluster.

Firstly, model selects initial centoinds (k).

Secondly, each data point is assigned to the nearest centroid according to Euclidean distance, organising clusters.

Lastly, the model recalculates the centroids as the mean value of all points within each cluster. This process repeats until centroids are stable, meaning points no longer switch clusters, or a predefined number of iterations is reached. The goal is to minimise the total variance inside clusters.

The study by Sinaga K. and Yang M. (2020) provides a detailed explanation of the implementation of the K-Means clustering model.The study by Sinaga K. and Yang M. (2020) provides a detailed explanation of the implementation of the K-Means clustering model.

#### DBSCAN

DBSCAN is abbreviation for Density Based Spatial Clustering of Applications with Noise. This algorithm clusters data points based on density rather than distance to centroids. It defines two key parameters:

**Epsilon:** the neighborhood radius

**MinPts:** minimum number of points required to create a dense radius.

Firstly, DBSCAN selects a random point and checks if there are at least MinPts points within Epsilon. If so, DBSCAN assigns this point as "core point" and forms a new cluster. After that the radius expands to reach all points within Epsilon. DBSCAN does not require specifying the number of clusters.

Ahmed, K.N. and Razak, T.A. (2016) compared the efficiency of DBSCAN and other density-based models with spatial clustering analysis.”

#### Agglomerative

Agglomerative clustering algorithm belongs to hierarchical family. It uses an approach "bottom-up" to cluster data points.

It consuders each data point as it own cluster. Firtly, it finds out two the closest datapoint and assign them as a new cluster. Secondly, it merges all data point in the same manner until there only one cluster left or it reaches predefined number of clusters.

If a cluster contains only 1 data point starting poing for calculating distance is defined by this point, however if there are more than 1 point there may be several way to define starting point.

**Single linkage:** distance between closest point.

**Complete linkage:** distance between farthest points.

**Avarage linkage:** average distance between points.

This hierarchical approach creates a tree-like structure called a dendrogram, which we displayed earlier. For our or similar cases this model may be beneficial because it does not require predefined number of clusters and we can select any number we need.

In this research, Müllner, D. (2011) compares traditional and modern hierarchical algorithms, with a particular focus on the Agglomerative Clustering algorithm.

#### OPTICS

OPTICS stands for Ordering Points to Identify the Clustering Structure. It is another density based clustering algorithm similar to DBSCAN but designed to handle clusters of different density more effectively, which can be helpful considering uneven distrebution of our datapoint. It orders data points based on their density reachability.

**Epsilon:** maximum neighborhood radius.

**MinPts:** minimum number of points required to create a dense radius.

The difference between OPTICS and DBSCAN is that OPTICS detect clusters across different density levels. It calculates a "reachability distance" for each point, reflecting how accessible each point is from its neighbors.

Another study by Ahmed, M.A. et al. (2020) compares different clustering algorithms, including K-Means, DBSCAN, and OPTICS.

#### BIRCH

BIRCH stands for Balanced Iterative Reducing and Clustering using Hierarchies. This is a hierarchical clustering algorithm particularly designed for handling large datasets.

The main feature of this model is Clustering Feature Tree, where each node stores a summary of a certain cluster using "clustering features". It stores data about number of points within the cluster, linear sum, and square sum, which allows efficient calculations of cluster metrics.

A very detailed study by Lorbeer, B. et al. (2018) demonstrates the implementation of the BIRCH model, particularly in terms of parameter settings, which will assist us in defining parameters for model tuning.

#### Affinity Propagation

Affinity Propagation identifies clusters by relationship between data points which are representative points for each cluster.

Affinity Propagation automatically determines number of clusters this based on similarity between pairs of points.

There are two types of of relationship which the model checks:

**Responsibility:** how well-suited a point is as an exemplar for another

**Availability:** how appropriate it is for a point to choose another as its exemplar.

Model creates a cluster around pair of data points with the highest responsibility and availability.

Serdah, A.M. and Ashour, W.M. (2016), in their research, compared traditional Affinity Propagation with K-Affinity Propagation combined with K-Means.

#### Gaussian Mixture

Gaussian Mixture Model is a probabilistic clustering algorithm which assumes that data is generated from a mixture of several Gaussian distributions. The model estimates the probability that each data point belongs to each Gaussian component based on its location and the parameters of the Gaussians (mean and covariance).

In this study, Fung, G. (2001) provides a clear overview of basic clustering models, including the Gaussian Mixture model.

#### Calinski-Harabasz score

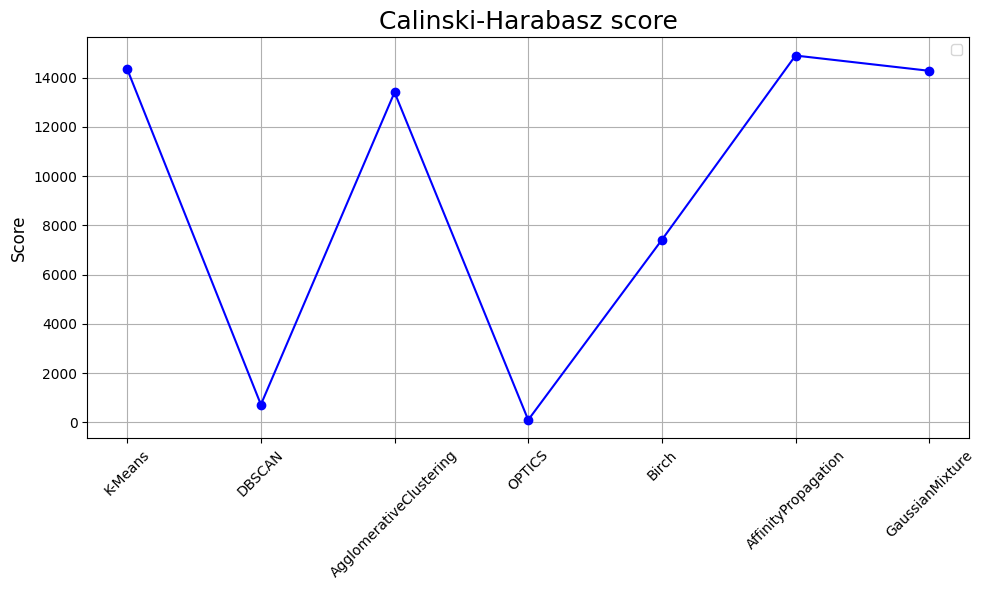
One of the evaluation techniques we are going to use is the Calinski-Harabasz score. In their research, Cengizler, C. and Kerem-Un, M. (2017) applied the Calinski-Harabasz score to evaluate a segmentation algorithm.

### Implementation

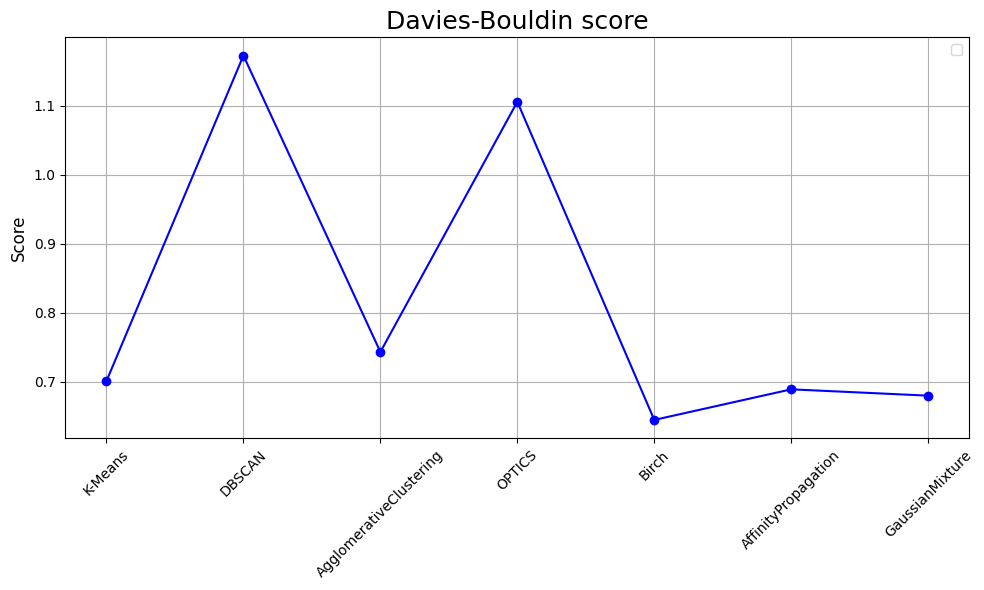
model params calinski\_harabasz\_score davies\_bouldin\_score   
0 K-Means {'init': 'k-means++', 'max\_iter': 100, 'n\_clus... 14347.735458 0.701011   
1 DBSCAN {'eps': 0.1, 'min\_samples': 20} 717.697397 1.172948   
2 AgglomerativeClustering {'linkage': 'ward', 'n\_clusters': 50} 13409.617245 0.743072   
3 OPTICS {'max\_eps': 0.1, 'min\_samples': 20} 95.135790 1.105909  
4 Birch {'branching\_factor': 20, 'threshold': 0.5} 7412.965956 0.644527  
5 AffinityPropagation {'damping': 0.9} 14902.685030 0.688882   
6 GaussianMixture {'covariance\_type': 'tied', 'n\_components': 50} 14283.210755. 0.679714

### Evaluation

#### Calinski-Harabasz score



#### Davies-Bouldin score



#### Visual evaluation



As we can see our models in general follows 3 main principles of segmentation:

1. **K-Mean, Agglomerative, Affinity Propagation and Gaussian Mixture**

These algorithms incorporate all data points into a segmentation and divide the data into smaller clusters. These clustering models are very suitable for general evaluation of specific zones, as each data point within a cluster is close to and reachable from the centroid. For this reason, we can associate certain characteristics of the entire cluster with each data point within it. For example, if we discover an overabundance of restaurants in Dublin, we can conclude that the competitiveness level in Dublin is quite high, and all restaurants within that cluster may be affected. In contrast, Galway has its own distinct conditions, which would not be influenced by the situation in Dublin.

1. **DBSCAN and OPTICS**

These two density-based algorithms provide interesting segmentation results. Unlike the previous models, they only assign data points with the highest density to clusters, leaving other points in the "noise" area. This approach can be beneficial for businesses that rely on a certain number of tourists arriving in specific areas, such as souvenir shops, pubs, or restaurants. Additionally, these algorithms can highlight areas with high competition. In some cases, investors may want to avoid such highly competitive areas for example, a new coffee shop brand that is still relatively unknown might struggle to thrive in a highly competitive environment.

1. **BIRCH**

The hierarchical BIRCH model produced three large clusters. Large-scale clustering may be beneficial in specific business or political scenarios. For instance, a large supplier for restaurants may decide to build its own warehouses and needs to determine the three best locations. In this case, the centroids of these clusters could provide a good solution. Another example is if the government wants to localize all significant locations for the Irish tourism industry into three main zones for more precise management and protection.

## Balance analysis

### Associate centroids with cities

#### Calculate centroids

K-Means\_cluster Centroid\_Latitude Centroid\_Longitude  
0 0 53.248447 -9.035009  
1 1 53.352834 -6.251831  
2 2 52.032597 -9.562302  
3 3 53.981241 -7.526154  
4 4 52.168645 -7.788499

#### Teory review: Haversine Formula

Chopde, N.R. and Nichat, M., (2013)

### Implementation: Balance Analysis

### Absolute values

| **Zone** | **Food & Drink** | **Outdoor & Adventure** | **Culture & History** | **Wellness & Leisure** | **Shopping & Purchase** |
| --- | --- | --- | --- | --- | --- |
| **Ardfert, Zone 14** | 66 | 152 | 68 | 39 | 30 |
| **Baile an Ghearlánaigh, Zone 22** | 63 | 126 | 53 | 34 | 20 |
| **Baile Átha Luain, Zone 12** | 44 | 91 | 30 | 20 | 15 |
| **Ballinrobe, Zone 17** | 34 | 85 | 32 | 19 | 17 |
| **Ballyshannon, Zone 35** | 39 | 89 | 43 | 26 | 9 |
| **Cahersiveen, Zone 34** | 19 | 67 | 37 | 17 | 15 |
| **Cappoquin, Zone 4** | 65 | 152 | 64 | 32 | 29 |
| **Carlow, Zone 28** | 49 | 118 | 55 | 31 | 27 |
| **Carndonagh, Zone 21** | 32 | 82 | 37 | 26 | 10 |

### Percentage

| **Zone** | **Food & Drink** | **Outdoor & Adventure** | **Culture & History** | **Wellness & Leisure** | **Shopping & Purchase** |
| --- | --- | --- | --- | --- | --- |
| **Ardfert, Zone 14** | 18 | 43 | 19 | 10 | 8 |
| **Baile an Ghearlánaigh, Zone 22** | 21 | 42 | 17 | 11 | 6 |
| **Baile Átha Luain, Zone 12** | 22 | 45 | 15 | 10 | 7 |
| **Ballinrobe, Zone 17** | 18 | 45 | 17 | 10 | 9 |
| **Ballyshannon, Zone 35** | 18 | 43 | 20 | 12 | 4 |
| **Cahersiveen, Zone 34** | 12 | 43 | 23 | 11 | 9 |
| **Cappoquin, Zone 4** | 19 | 44 | 18 | 9 | 8 |
| **Carlow, Zone 28** | 17 | 42 | 19 | 11 | 9 |
| **Carndonagh, Zone 21** | 17 | 43 | 19 | 13 | 5 |

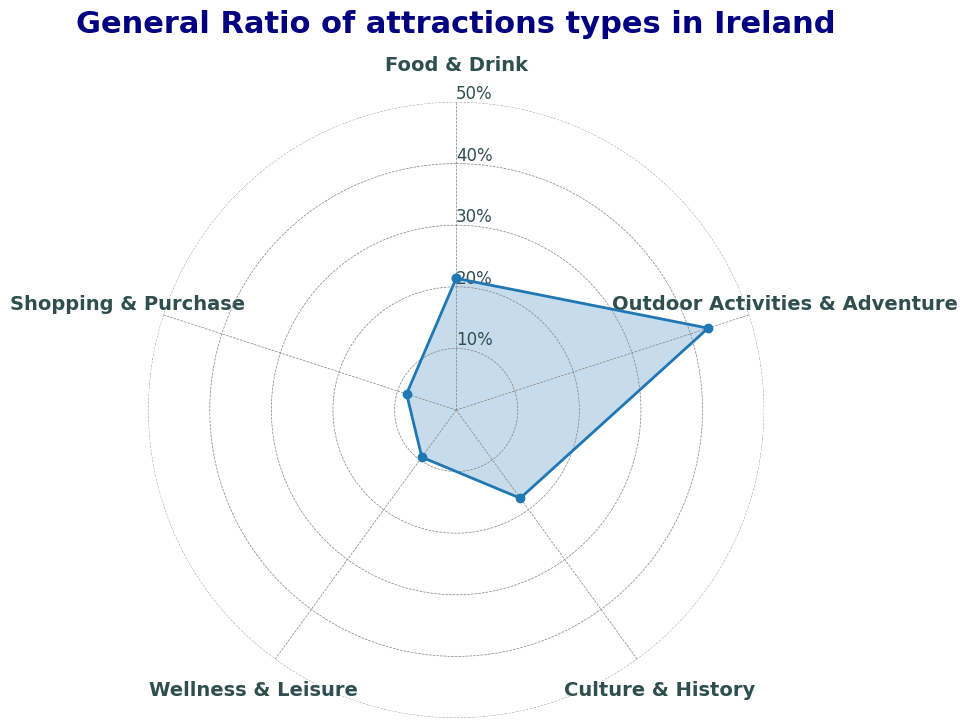
### General ratio in Ireland

| **Category** | **Total** | **Percentage** |
| --- | --- | --- |
| **Food & Drink** | 2559 | 21 |
| **Outdoor Activities & Adventure** | 5148 | 42 |
| **Culture & History** | 2119 | 17 |
| **Wellness & Leisure** | 1134 | 9 |
| **Shopping & Purchase** | 1013 | 8 |

### 

### Visualisation: General ratio of attractions in Ireland

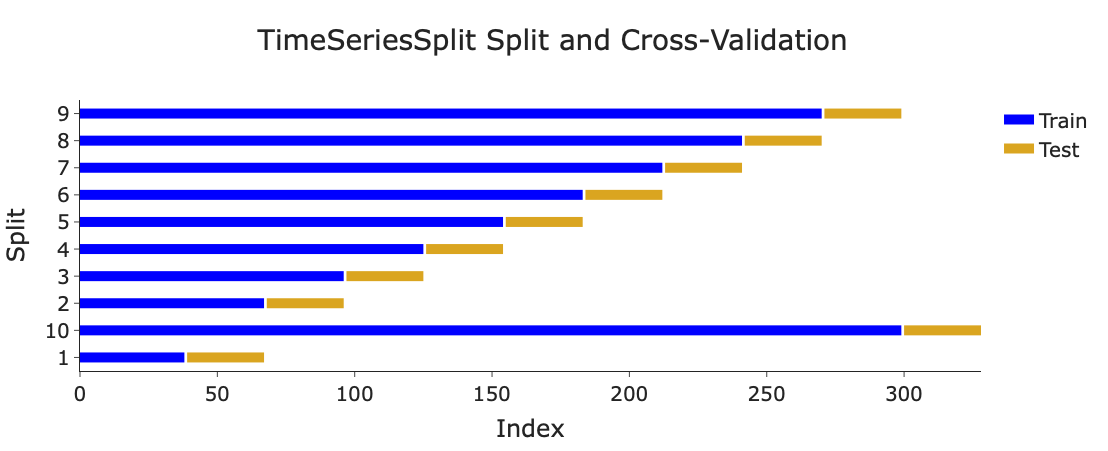
For creating a radar chart, we used a guide article provided by Peh (2020) on Medium.com.



Time-series modeling

## Time series cross validation

We are going to include cross validation in our hyperparameters tunning. Because train/test and cross validation have certain peculiarities. In contract with other ML models we use for classification or regression problems for time series we can not just select random records for split. Instead we need to select consecutive period of time as a test split. Moreover that period must the most recent one compered to train.



## Exponential Smoothing

### Simple Exponential Smoothing

This model is base component of Holt's Winters model we are going to use later. The simplicity of this model is that it does not cosider trend and seasonality. According to Ostertagová, E. and Ostertag, O., (2011) the general idea of all models belond to exponential smoothing family is they assign more weight for recent data and less for old data. Smoothing parametrs presented as (a) in the equation and varies between 0 and 1 depending on the time in series.

### Holt's linear trend model

**Level equation:**

**Trend equation:**

**Forecastequation:**

Holt's linear model, also known as double exponential smoothing, adds a linear component to the simple exponential smoothing model. This enhancement makes the model more accurate as it accounts for the trend in the data. Although the model does not include a seasonality component, it is widely applied in linear regression problems. For example, in the study conducted by Bujang, M.A., et al. (2009), Holt's linear model performed better in certain cases than Simple Linear Regression.

## Holt's winters

#### Additive Holt's Winters Model

**Level Equation:**

**T****rend Equation:**

**S****easonal Equation:**

**F****orecasting Equation:**

**M****u****ltiplicative Holt's Winters Model**

**Level Equation:**

**T****rend Equation:**

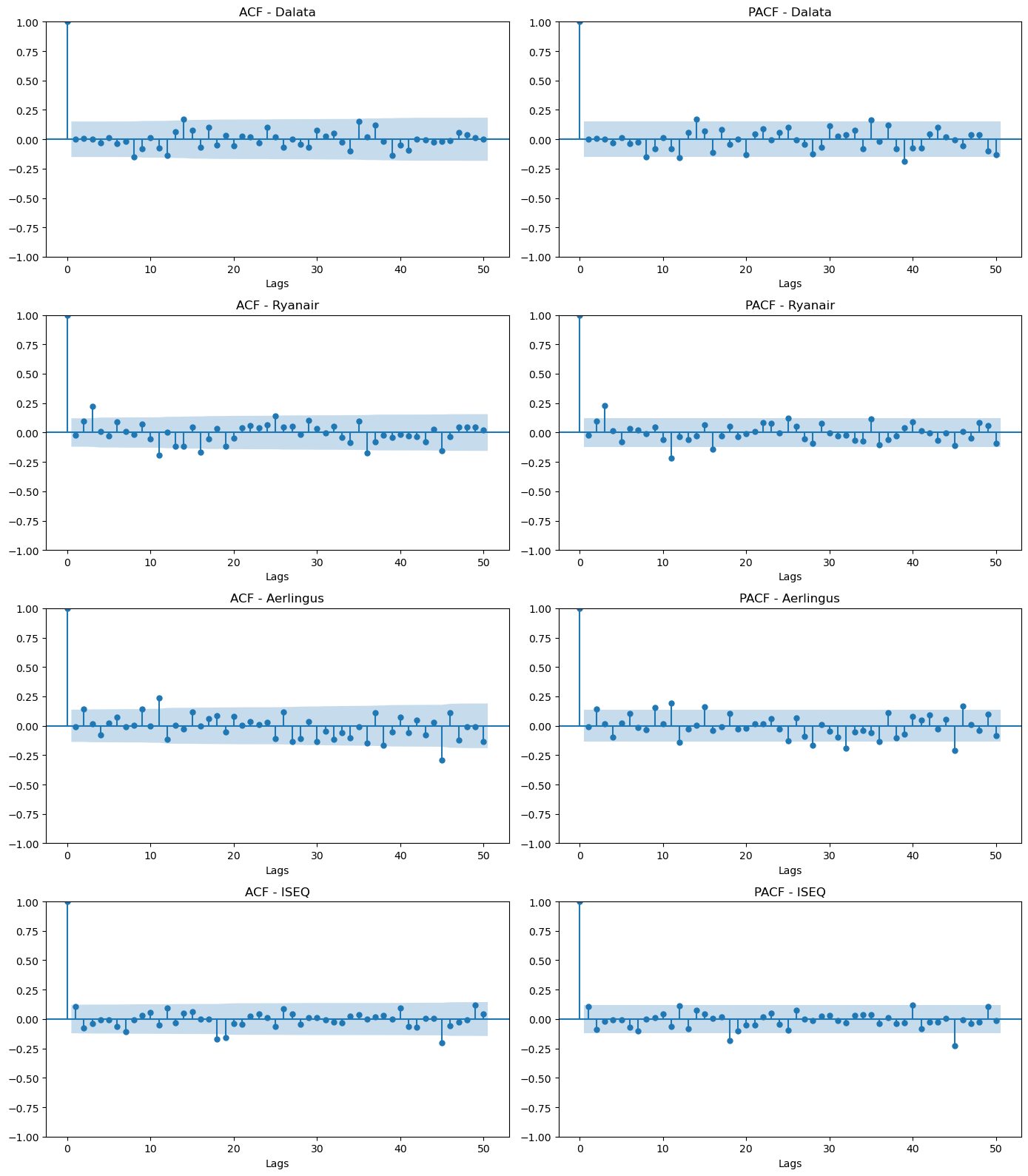
**S****easonal Equation:**

**F****orecasting Equation:**

Holt's Winters, also known as Triple Exponential Smoothing, as the name suggests, includes two components we discussed previously (level and trend) and adds a seasonality component. This method accounts for all three standard components found in most time series data. The study conducted by Almazrouee, A.I., et al. (2020) was referenced to guid for key stages of building our model.

### Evaluation: analysing residuals

#### Plot ACF and PACF



#### Ljung-Box Test

Ljung-Box is a basically hypothesis testing to underrstand if residuals are distributed independently with no pattern or there is a relationship in their distribution which model have not cover yet.

**H0:** Residuals are independently distributed.

**H1:** Residuals are not independently distributed.

Detailed procedure of resuduals testing using Ljung-Box method can be found in study by Safi, S.K. and Al-Reqep, A.A., (2014)

aerlingus

lb\_stat lb\_pvalue  
1 0.009256 0.923357  
2 4.195128 0.122755  
3 4.239690 0.236718  
4 5.442856 0.244798  
5 5.575430 0.349746  
6 6.779312 0.341739  
7 6.803568 0.449615  
8 6.805816 0.557721  
9 11.150676 0.265522  
10 11.154098 0.345639

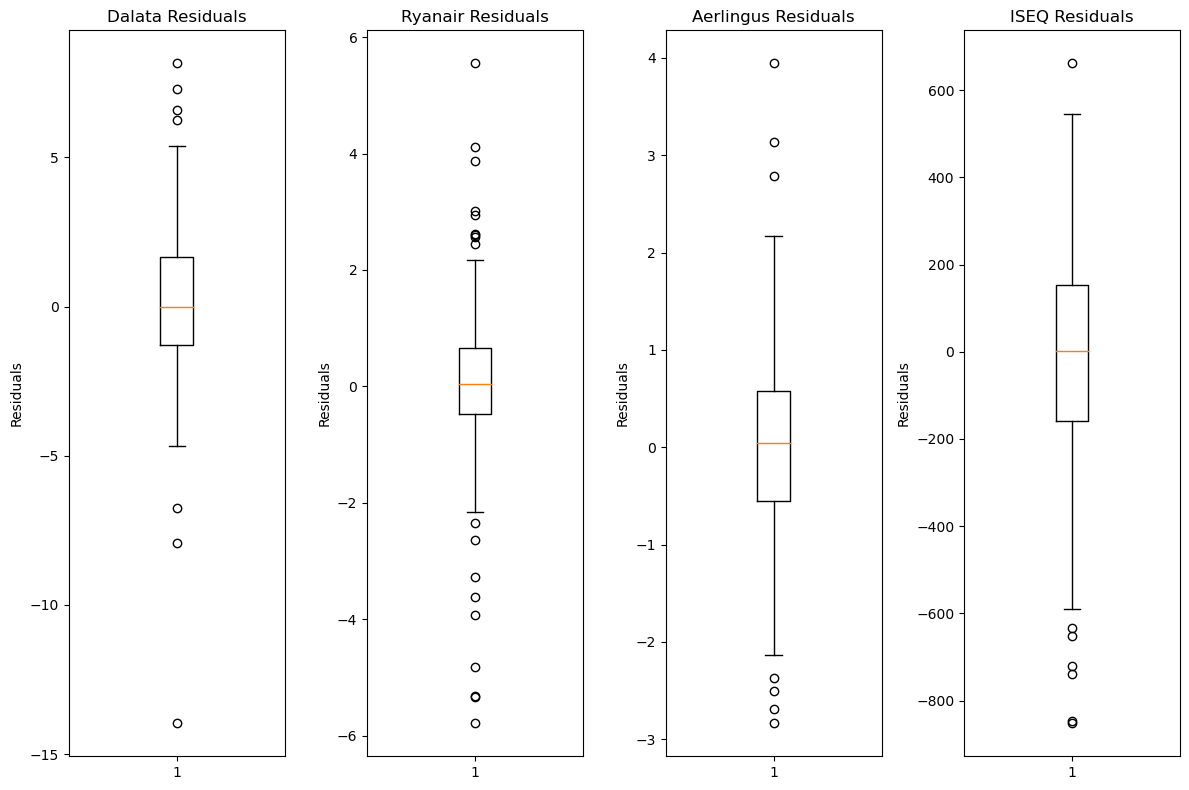
ryanair

lb\_stat lb\_pvalue  
1 0.131450 0.716933  
2 2.656226 0.264977  
3 15.830850 0.001228  
4 15.846722 0.003232  
5 16.066717 0.006656  
6 18.335125 0.005447  
7 18.361509 0.010441  
8 18.443274 0.018138  
9 19.946778 0.018244  
10 20.728654 0.023067

dalata

lb\_stat lb\_pvalue  
1 0.000039 0.994995  
2 0.012170 0.993933  
3 0.012669 0.999622  
4 0.144772 0.997503  
5 0.192335 0.999194  
6 0.455401 0.998340  
7 0.533775 0.999313  
8 4.620695 0.797240  
9 5.836826 0.756132  
10 5.869021 0.826145

#### Residuals distredution



As we can see, the distribution is generally normal. However, there is slight skewness, which indicates that the models could be tuned further for better performance. Given that we are analyzing real-world data with many unpredictable factors, a good outcome would be achieving a distribution without outliers.

## ARIMA

### Stationarity

Stationarity is a key factor that impacts a time-series analysis. Basically, Stationarity is the condition of time-series data when there are no any long term trends and no clear seasonality. In statistic it means that 2 key statistical properties remain constant:

1. Mean (Trend)
2. Varience (Seasonality)

Stationarity as a key object in time series analysis is reviewed in study by Petrică, et. al. (2017)

To check on time series if it is stationary or not we can apply Augmented Dickey-Fuller test.

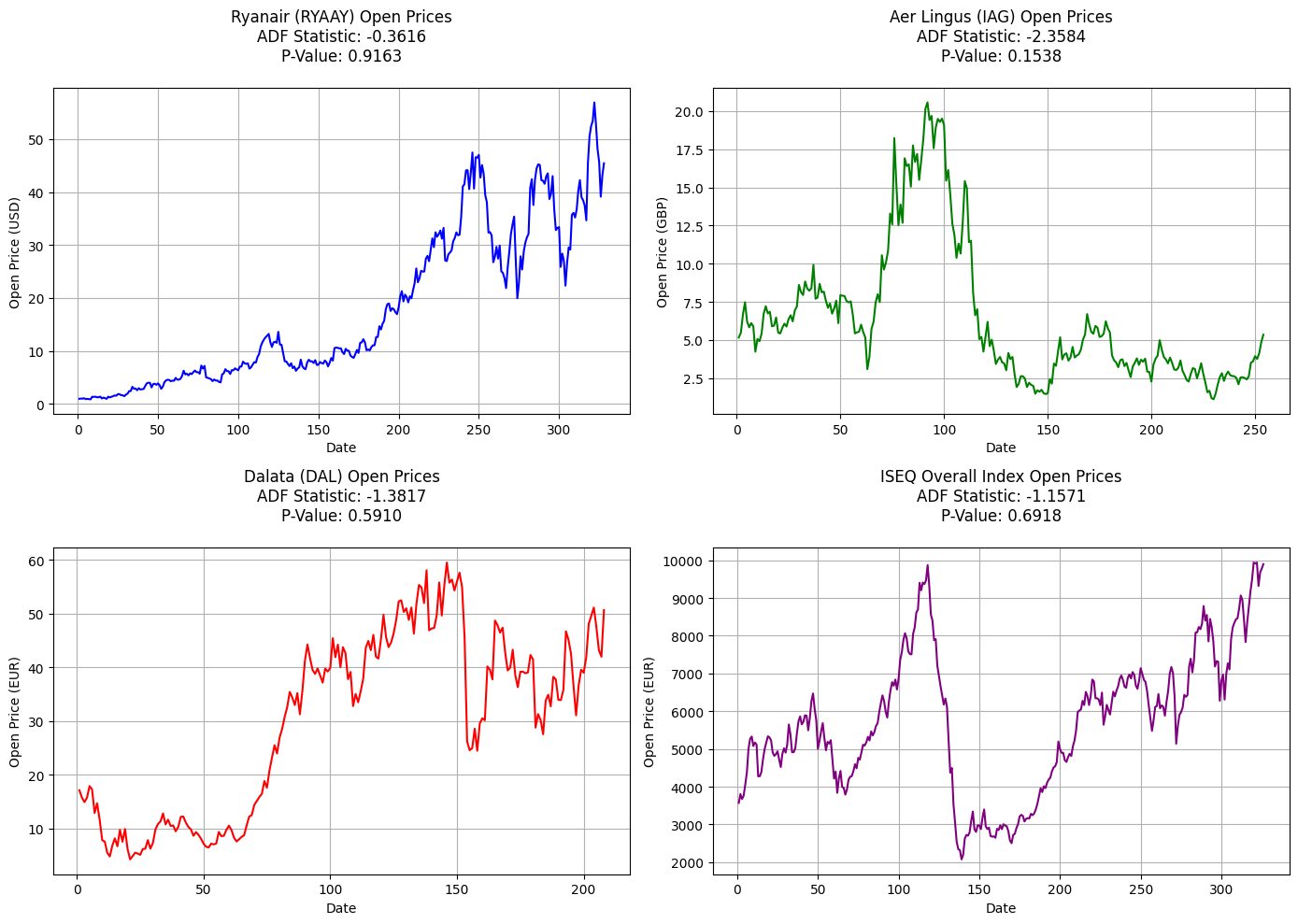
#### Dickey-Fuller testing for Stationarity

The most commont method to assess stationarity of the timeseries dataset is Dickey-Fuller testing which similarly to Ljung-Box test applies hypothesis testing.

**H0:** non-stationary.

**H1:** stationary.

Also we can display p-value to understand level of certainty stationarity has.



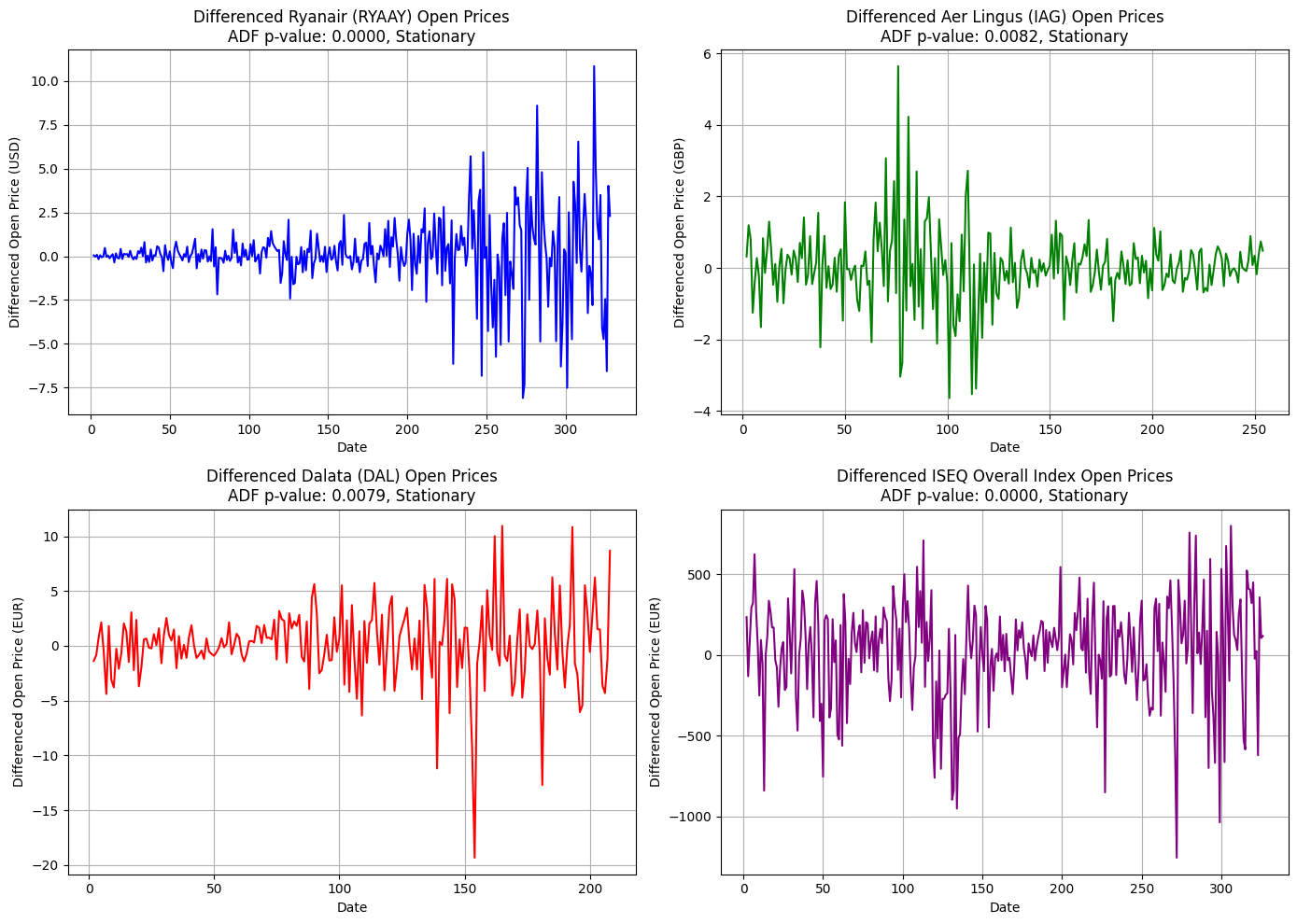
### Tranformation

#### Differencing

Differencing is one of the most popular and effective approaches for levelling mean value off. It has a simple equation and all we need to do is just take a difference between 2 data points with certain steps which we need to specify.

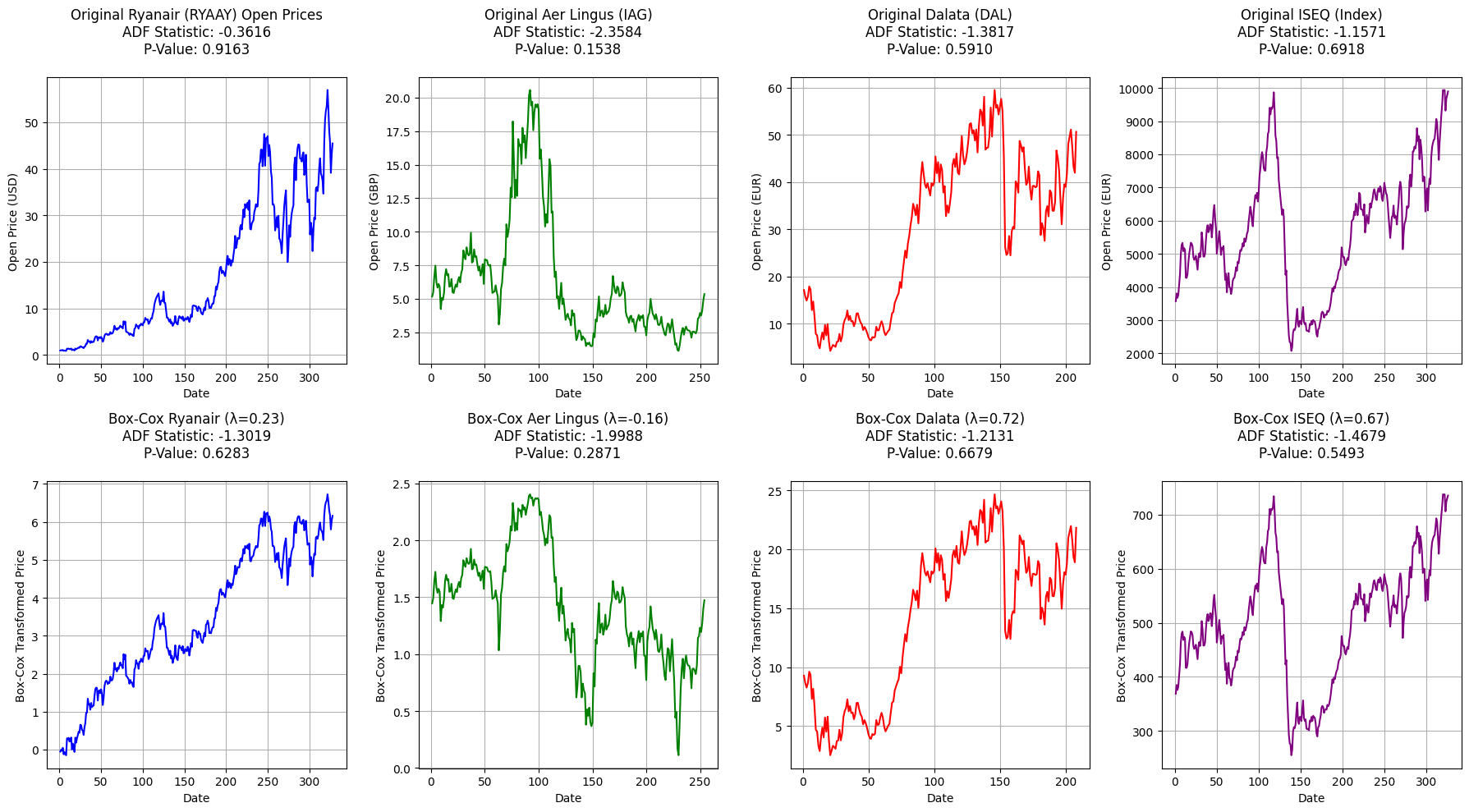
It deserves to be mentioned that after taking differences we end up with some NaN values in the beginning. It happens because the first values served as a starting point and they didn't have previous values. So, we need to drop those rows in the beginning which contain NaNs.

In Python Pandas library has a method (.diff) which applies a differencing formula which is displayed below. By default, .diff method takes a different which is 1 step behind, however, the number of steps can be modified depending on the results.



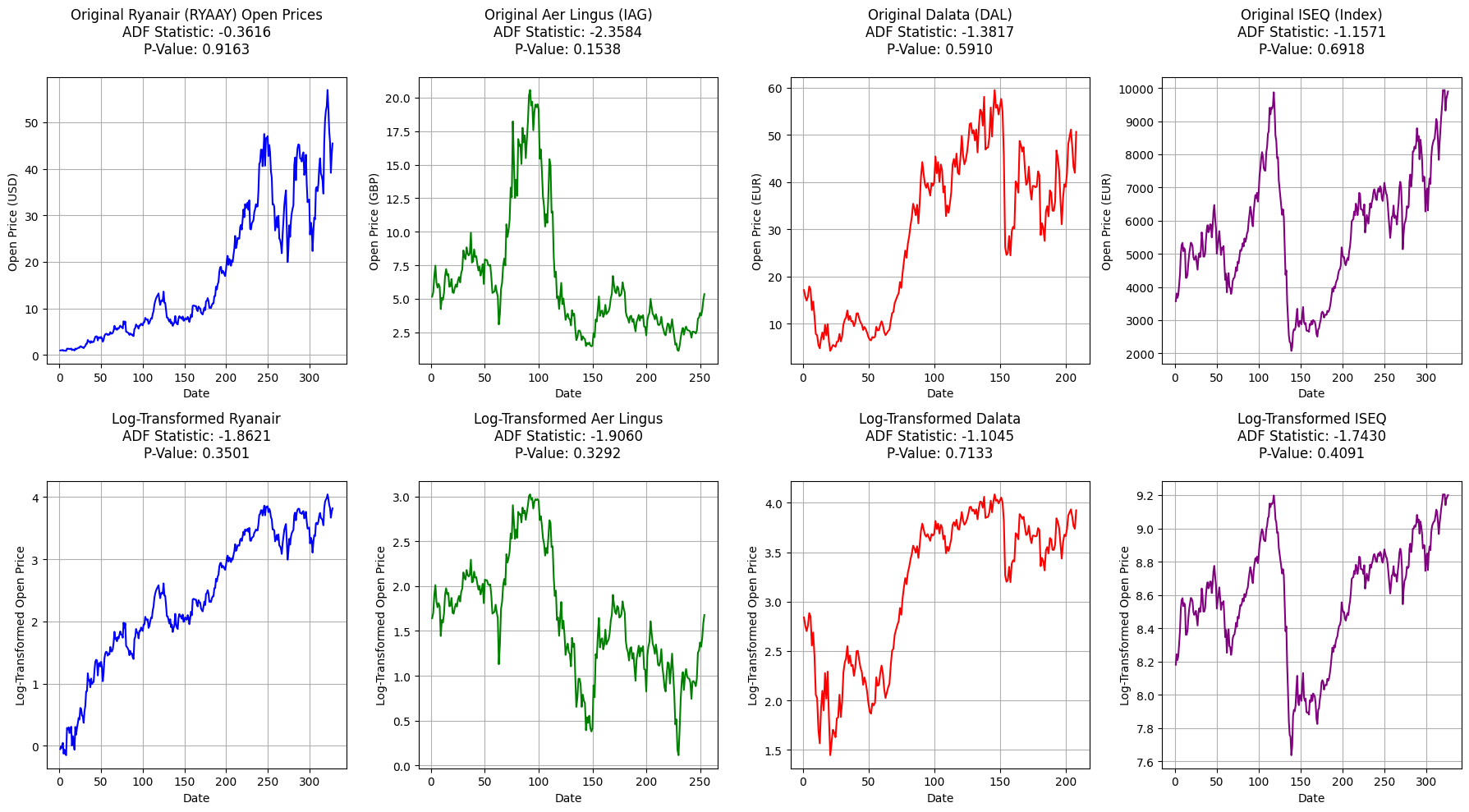
#### Box-Cox tranformation

Similarly to other tranformation methods Box-Cox transform normalise distribution. It is important because some models and ARIMA included apply the maximum likelihood estimation (MLE) to set parameters. And it is expected that data normally distributed.

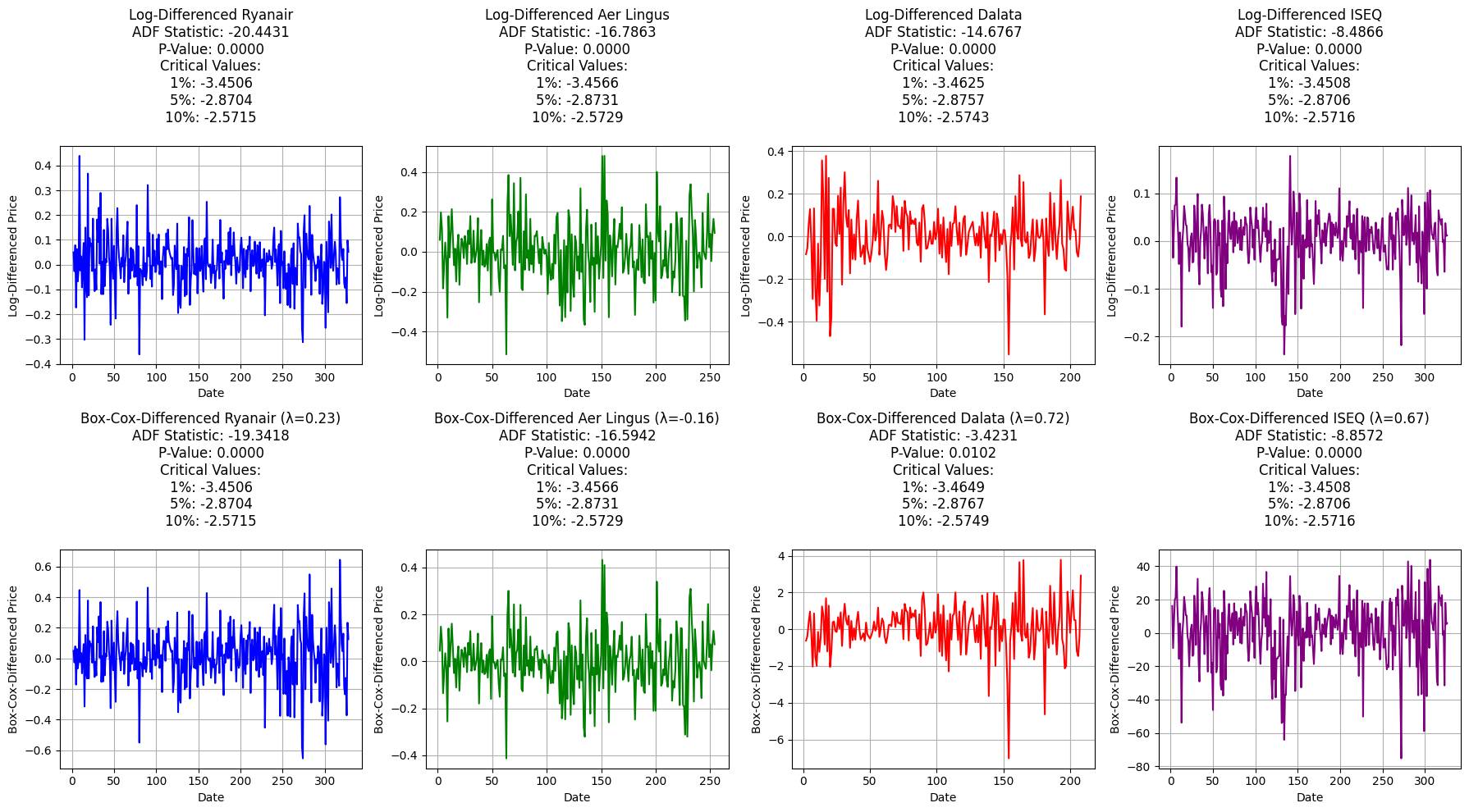


#### Log transformation

Another approach to handle variance (flactuation) is log transform. Which has similar but simplified principles compered with Box-Cox tranformation. It only takes natural log from each value. In other word it levels off the data by reducing the scale.



#### Log VS. Box-Cox



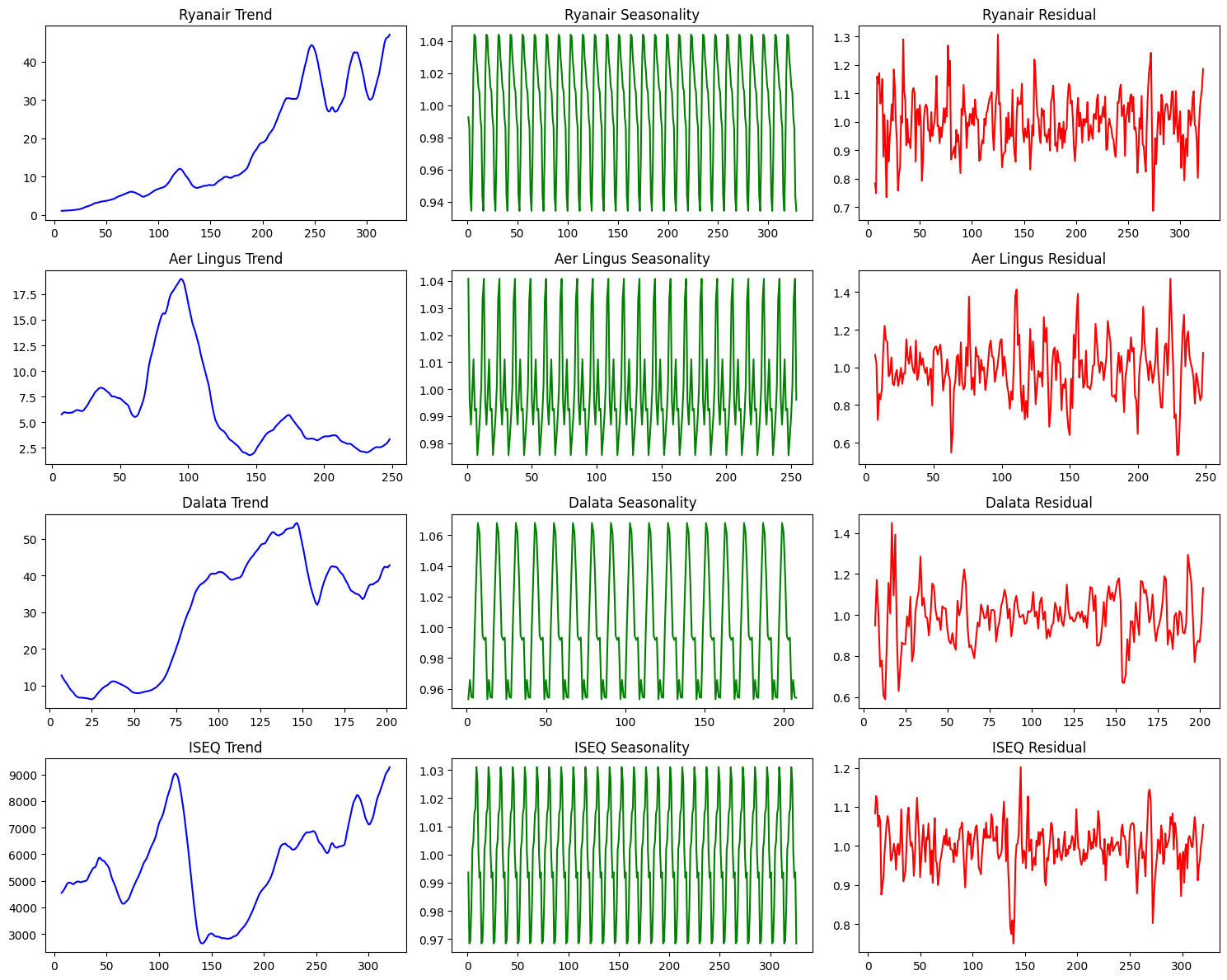
### Decomposition

Another important aspect of time-series data is seasonality. Some time-series datasets contain obvious seasonality components, it can be found in sales data. Depending on the dataset and the goal there may be different levels of seasonality considerations (daily, weekly, monthly, yearly ...). To visualise seasonality and understand how it may affect further modelling we can apply the decomposition model.

There are 2 types of decomposition which can be applied:

Additive - series with constant absolute variations in the trend and seasonality. Series = Trend + Seasonality + Residuals

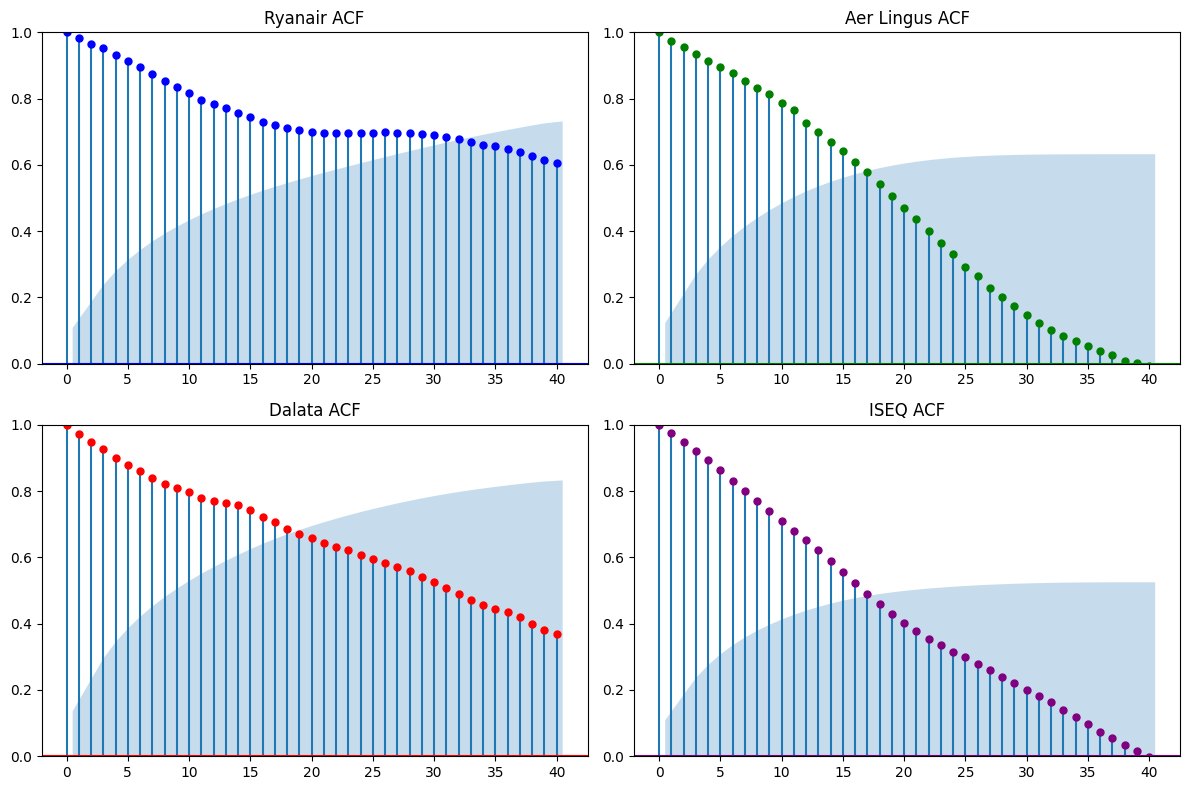
Multiplicative - series where variations increase or decrease proportionally to the level of the series. Series = Trend × Seasonality × Residuals



### Autocorrelation (ACF)

One of the parameter we need to define is a number of lags. This number is required for Autoregreesion part of ARIMA model (AR). In this section we will define optimal number of lags for each dataset using autocorrelation. However, we will also consider different approaches before fitting final model, such as partial autocorrelation and hyperparametrs tunning.

Autocorrelation displays correlation with the same feature but different time. In other words we can understand how historical data correlates with recent one.



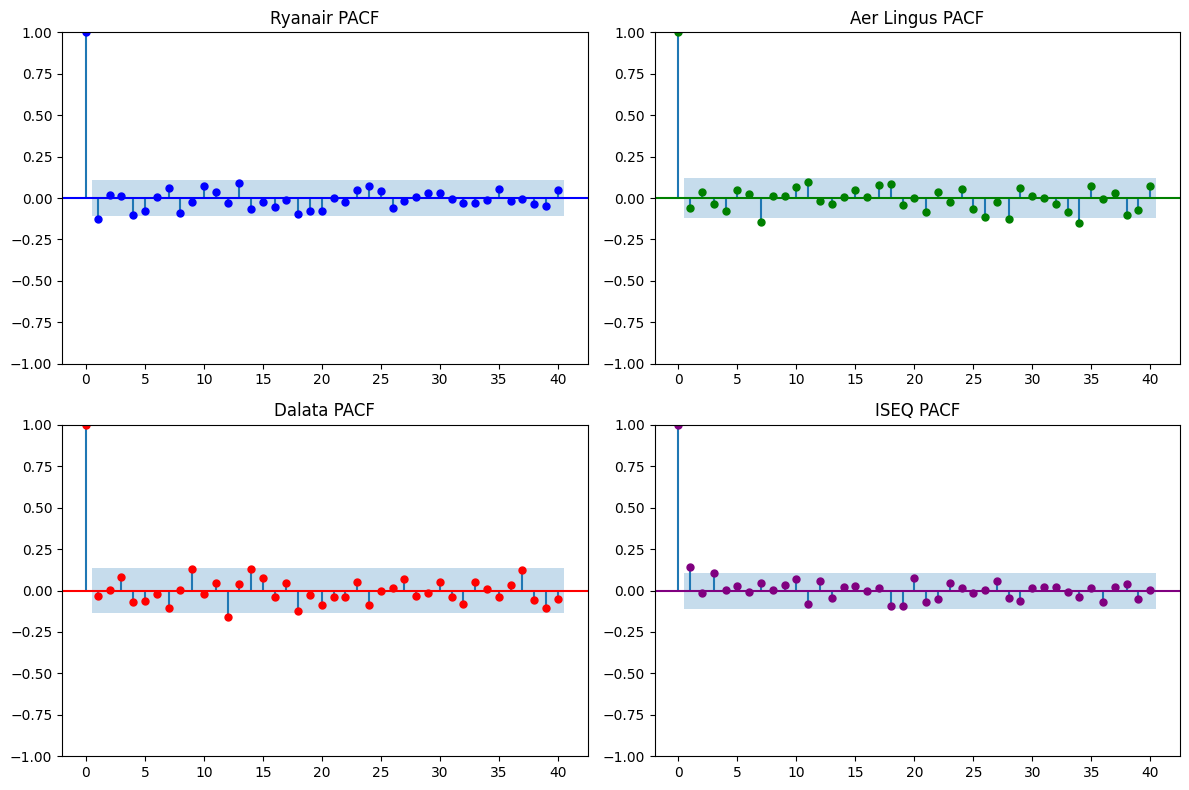
We can see different patterns of autocorrelation. For example, historical data of Ryanair has the most impact on recent data across all datasets. Up to 32-33 lags are in the significance zone. It means that current changes in stock prices are correlated with changes happened 32 lags (Months) with coeff. 0.65.

Even though, after 33rd lag correlation coefficient dives into "non-significant zone" we can see it still has solid correlation (0.6) with recent lags, comparing to other data sets.

Aer Lingus and ISEQ index lags correlation have sharply reduced. We can conclude that only data from last 16-18 lags (months) is statistically significant.

### Partial Autocorrelation (PACF)

Partial Autocorrelation measures the correlation between observations at a specific lag while removing the influence of the intermediate lags. The PACF helps us to identify the number of lags we should include in the model for the autoregressive (AR) part.



Considering PACF testing displayed we can make several conclusions:

**Seasonality pattern:**

If our datasets had a clear seasonality pattern, we would be able to see it on these plots. However, only Ryanair dataset has a straightforward pattern between 25th and 40th lags. Other datasets' do not have clear cyclic pattern.

**Noise and outliers:**

We can see that majority of lags are not significanly correlated. It may indicate that there were unpredicted factors impacted time series, which have an uncertain frequency.

**Significant events:**

Those lags which entered into a significance zone indicate a certain events or short trends which are correlating with current lags. For example an increase of fuel prices may negatevly impact share process. This event does not have certain seasonality pattern but it creates the same downward pattern which has multiple correlations across the time series.

### AutoRegression

Autoregression is one of the base blocks for ARIMA model.

The model applies regression to itself, in other words predicts next values based on a certain number of previos values.

The number of those previos lags which model uses for forecasting is value which we need to specify in ARIMA model (p).

Brifly, autoregression forecasts the next value by summing avarage level of the time series (c), previous values with certain cofficient and error term.

### Moving Avarage

The Moving Average model predicts future values based on past forecast errors (rather than past values as we discussed previously). In an MA model, each point in the time series is expressed as a linear combination of past error terms, typically with recent errors having the most significant impact.

The number of error terms considered is defined by the model's order as (q), where q is the number of past errors included. By incorporating past forecast errors, the MA model smooths out short-term fluctuations and shows underlying trends or patterns in the data.

### ARIMA

As we can see in an equation, ARIMA model combines **AutoRegressive (AR)**, **Integrated (I)**, and **Moving Average (MA)** components.

Similar study predicting demand condacted by Fattah, J., et. al. (2018) applies ARIMA model aligned with Box–Jenkins time series procedure, what gave us right direction for building our model.

#### Hyperparameters Tunning

dataset best MAPE best params  
0 Ryanair 22.716610 (0, 1, 2)  
1 Aerlingus 46.800227 (3, 1, 2)  
2 Dalata 21.217374 (2, 1, 2)  
3 ISEQ 16.545628 (4, 1, 8)

### Auto ARIMA

Similarly we can use Auto ARIMA package "auto\_arima" provided by "pmdarima" library to get the best parameters for ARIMA and SARIMA models.

## SARIMA

SARIMA or Seasonal ARIMA model extends ARIMA's by adding a seasonal components. In total we have 7 components we need to set or in our case include in tunning:

**p,d,q**

Non-seasonal parameters (Autoregression, Differencing and Moving Avarage).

**P,D,Q**

Seasonal parameters (Autoregression, Differencing and Moving Avarage).

**m**

The length of the seasonal period.

### Hyperparameters Tunning

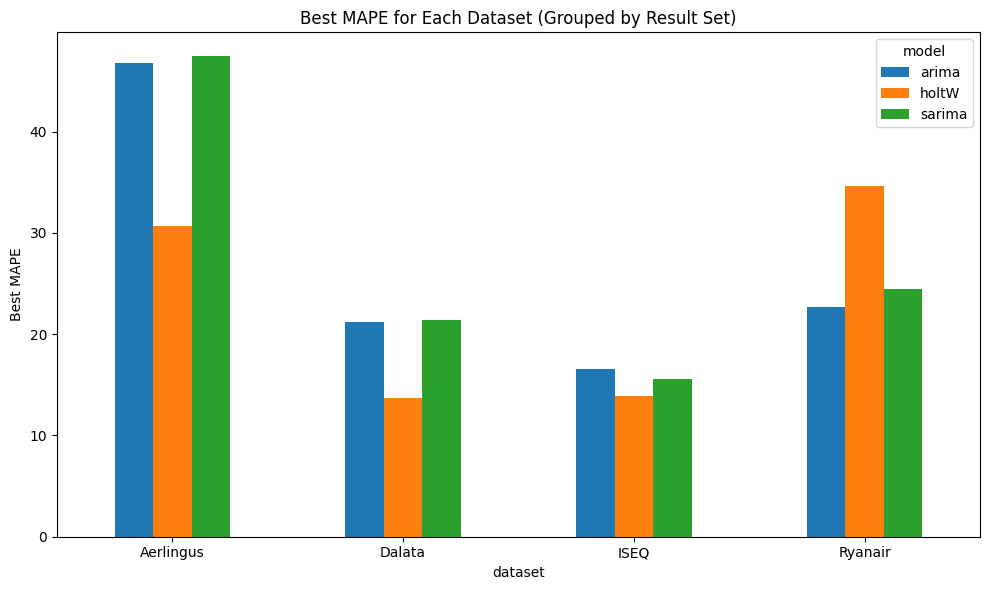
dataset best MAPE best params (p,d,q) best seasonal params (P,D,Q,m)  
0 Ryanair 24.488506 (0, 1, 2) (2, 0, 2, 12)  
1 Aerlingus 47.438843 (3, 1, 2) (0, 0, 0, 12)  
2 Dalata 21.442403 (4, 1, 4) (0, 1, 2, 12)  
3 ISEQ 15.577594 (2, 1, 3) (0, 1, 2, 12)

## Model selection

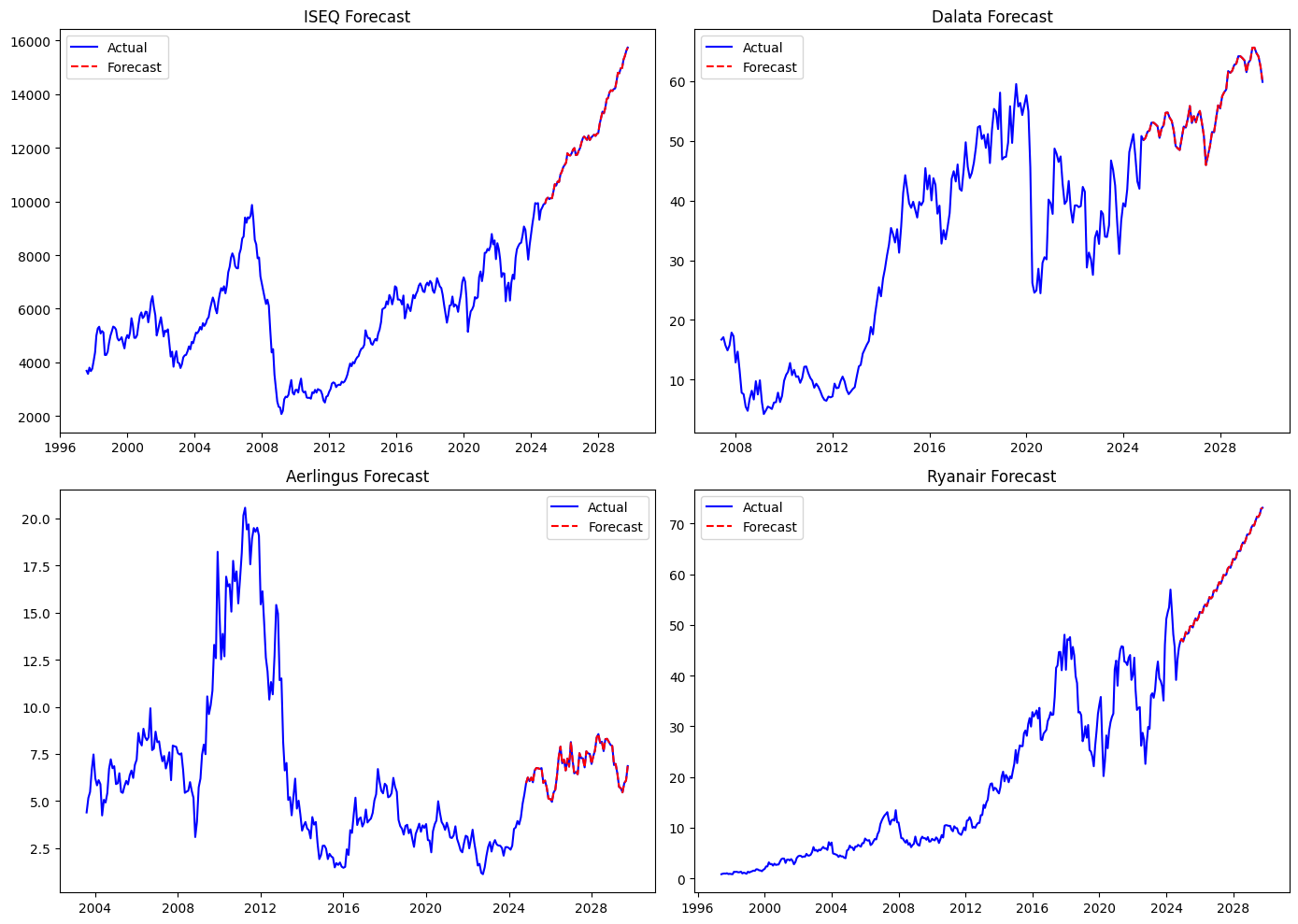
### Evaluation metric (MAPE)

For evaluation we are going to use MAPE (mean absolute percentage error). I think considering the variety of datasets we are analysing and models we test, MAPE may be good solution because it intuitive and uses the same units across all models and datasets.

### Compare model for each dataset



### Plot Forecast



### Summarise forecast

#### Get estimated growth in percentage

**Forecast growth ISEQ index**  
Value at 2024-10-01: 9895.8   
Estimated value at 2029-10-01: 15738.3  
Estimated growth in percentage: 59.04%

**Forecast growth Aerlingus**  
Value at 2024-10-01: 5.3   
Estimated value at 2029-10-01: 6.9  
Estimated growth in percentage: 28.20%

**Forecast growth Ryanair**Value at 2024-10-01: 45.5   
Estimated value at 2029-10-01: 72.2  
Estimated growth in percentage: 58.82%

**Forecast growth Dalata Group**  
Value at 2024-10-01: 50.8   
Estimated value at 2029-10-01: 59.9  
Estimated growth in percentage: 17.82%

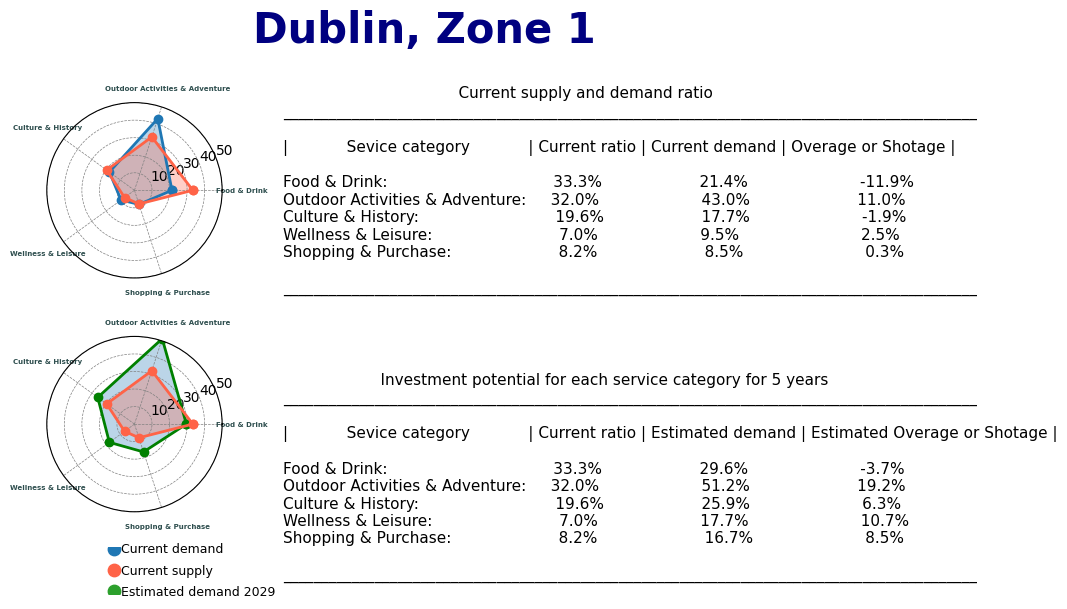
**Avarager growth**: 40.97094350075811

Investment potential 5-years forecast

The first question we need to answer how we will apply the percentage of growth we calculated to our attractions data. I can see two options:

1. Treat value as potential growth of number of tourists who are coming in Ireland. In this case overall growth (41%) will be represented equally for each service category we defined earlier (Food, Outdoor, Shopping ...) because those categories are complimentary customer wise.
2. Treat value as potential growth of income which Irish tourism industry will receive. In this case overall growth should be equally spread among categories of service. For 5 categories it is (8.2%) net growth for each service category.

We do not know exactly what is the growth of indexes we have done in time series related to.



References

Ahmed, K.N. and Razak, T.A., 2016. An overview of various improvements of DBSCAN algorithm in clustering spatial databases. International Journal of Advanced Research in Computer and Communication Engineering, 5(2), pp.360-363.

Ahmed, M.A., Baharin, H. and Nohuddin, P.N., 2020. Analysis of K-means, DBSCAN and OPTICS Cluster algorithms on Al-Quran verses. International Journal of Advanced Computer Science and Applications, 11(8), pp.248-254.

Almazrouee, A.I., Almeshal, A.M., Almutairi, A.S., Alenezi, M.R. and Alhajeri, S.N., 2020. Long-term forecasting of electrical loads in kuwait using prophet and holt–winters models. Applied Sciences, 10(16), p.5627.

Bholowalia, P. and Kumar, A., 2014. EBK-means: A clustering technique based on elbow method and k-means in WSN. International Journal of Computer Applications, 105(9).

Bujang, M.A., Adnan, T.H., Supramaniam, P., Abd Hamid, A.M. and Haniff, J., 2009. Prediction number of deaths by occurrence in Malaysia: a comparison between simple linear regression model and holt’s linear trend model. Statistics Malaysia—Journal of the Department of Statistics, Malaysia, 2, pp.25-37.

Cengizler, C. and Kerem-Un, M., 2017. Evaluation of Calinski-Harabasz criterion as fitness measure for genetic algorithm based segmentation of cervical cell nuclei. J. Adv. Math. Comput. Sci, 22, pp.1-13.

Chopde, N.R. and Nichat, M., 2013. Landmark based shortest path detection by using A\* and Haversine formula. International Journal of Innovative Research in Computer and Communication Engineering, 1(2), pp.298-302.

Fattah, J., Ezzine, L., Aman, Z., El Moussami, H. and Lachhab, A., 2018. Forecasting of demand using ARIMA model. International Journal of Engineering Business Management, 10, p.1847979018808673.

Fung, G., 2001. A comprehensive overview of basic clustering algorithms.

Hidayat, R., Harjoko, A. and Musdholifah, A., 2022. A Robust Image Retrieval Method Using Multi-Hierarchical Agglomerative Clustering and Davis-Bouldin Index. International Journal of Intelligent Engineering & Systems, 15(2).

Irish Tourism Industry Confederation, 2019. Tourism in the Irish Economy 2019. Available at: <https://www.itic.ie/cso-tourism-satellite-account-july-2023/> [Accessed 25 Oct. 2024]

K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

Lorbeer, B., Kosareva, A., Deva, B., Softić, D., Ruppel, P. and Küpper, A., 2018. Variations on the clustering algorithm BIRCH. Big data research, 11, pp.44-53.

Müllner, D., 2011. Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378.

Ostertagová, E. and Ostertag, O., 2011, September. The simple exponential smoothing model. In The 4th International Conference on modelling of mechanical and mechatronic systems, Technical University of Košice, Slovak Republic, Proceedings of Conference (pp. 380-384).

Peh, R. (2020) 'Creating a complex radar chart with Python', Medium, 26 February. Available at: <https://medium.com/@reinapeh/creating-a-complex-radar-chart-with-python-31c5cc4b3c5c> (Accessed: 24 October 2024).

Petrică, A.C., Stancu, S. and Ghițulescu, V., 2017. Stationarity–The Central Concept in Time Series Analysis. International Journal of Emerging Research in Management, 6(1), pp.6-16.

Safi, S.K. and Al-Reqep, A.A., 2014. Comparative study of portmanteau tests for the residuals autocorrelation in ARMA models. Science Journal of Applied Mathematics and Statistics, 2(1), pp.1-13.

Serdah, A.M. and Ashour, W.M., 2016. Clustering large-scale data based on modified affinity propagation algorithm. Journal of Artificial Intelligence and Soft Computing Research, 6(1), pp.23-33.

Shutaywi, M. and Kachouie, N.N., 2021. Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy, 23(6), p.759.

Sisodia, D., Singh, L., Sisodia, S. and Saxena, K., 2012. Clustering techniques: a brief survey of different clustering algorithms. International Journal of Latest Trends in Engineering and Technology (IJLTET), 1(3), pp.82-87.

### Author

Ilia Grishkin

<https://github.com/Ilia-Grishkin>

<https://github.com/replicant34>