**Project Report**

**GitHub URL**

My project is in the following GitHub location.

<https://github.com/Maryrig2/UCDPA-Mary-Rig>

**Abstract**

**Short overview of the entire project and features**

This project gave me the opportunity to work with APIs and get data from Bord Fáilte Ireland’s Open Data API. The data available to download, provides information on various accommodations, activities and attractions, in the Republic of Ireland.

The data is available at the following location <https://failteireland.developer.azure-api.net/apis>

This data is available under a Creative Commons license. Thank you to Bord Fáilte for making this data available, so that I could practice taking data from the web.

In the project, I import the data from API and save to CSV files.

The files are then analysed, and any data issues addressed. Missing values replaced and inconsistent columns are made consistent. There are checks for duplicate values.

Once the data is cleaned, an extra column is added to each resulting DataFrame. The three resulting DataFrames are then merged into one combined DataFrame.

The cleaned data is then ready for reporting on. There are five different bar charts depicting information gleaned from the datasets.

I had to use a different dataset for the machine learning demonstrations. This was sourced from Kaggle at

[https://www.kaggle.com/datasets/carlmcbrideellis/house-prices-advanced-regression-solution-file.](https://www.kaggle.com/datasets/carlmcbrideellis/house-prices-advanced-regression-solution-file. )

The license is CCO Public Domain.

**Introduction**

**Explain why you chose this project use case**

I chose this project use case, as I am interested in Tourism and curious about the data available. I wanted to be able to see from the data which counties had the most to offer and which ones could be developed further! I also wanted to be able to use API data for future projects in work, so this was good practice.

The project takes in API data from the Bord Fáilte website and saves the data into three separate CSV files.

Each file has the same columns, but the data lists the number of various accommodation types in the accommodation.csv. The types of accommodation available consist of hotels, guesthouses, self-catering, camping, B&Bs, activity holiday accommodation, glamping, historic houses at various locations throughout the country. There is also a welcome standard which offers exceptional accommodation – see link for details <https://www.failteireland.ie/welcomestandard.aspx>

The attractions.csv file is a list of all attractions such as art galleries, zoos, beaches, museums etc.

The activities.csv lists activities such as walking, activity operator, food and drink, cycling, bird watching, shopping etc.

The dataset I used for machine learning was a list of house prices in the US. I had planned to use datasets from the house price register in Ireland but this one was very suitable for regression.

**Dataset**

**Provide a description of your dataset and source. Also justify why you chose this source**

The datasets were sources from the Bord Fáilte website – see link below and details on each file sourced.

[APIs: Details - Microsoft Azure API Management - developer portal (azure-api.net)](https://failteireland.developer.azure-api.net/api-details#api=opendata-api-v1&operation=accommodation-csv)

Accommodation/Csv

Gets all accommodation listings, hotels, campground, bed and breakfasts etc. in CSV format

Request

https://failteireland.azure-api.net/opendata-api/v1/accommodation/csv

Activities/Csv

Gets all activity listings, tourism businesses, restaurants, cafes, golf courses, shopping centres etc. in CSV format

Request

https://failteireland.azure-api.net/opendata-api/v1/activities/csv

Attractions/Csv

Gets all attraction, art galleries, parks, zoos, beaches etc. in CSV format

Request

https://failteireland.azure-api.net/opendata-api/v1/attractions/csv

I chose this source as I am interested in Tourism and wanted to find out more about Tourism in Ireland.

[House Prices: Advanced Regression 'solution' file | Kaggle](https://www.kaggle.com/datasets/carlmcbrideellis/house-prices-advanced-regression-solution-file)

Solution.csv

I chose this source as it was ideal for regression.

**Implementation Process**

The tools I used were, PyCharm to test my python code and then Jupyter Notebook to create a Python notebook of the project. I also used Excel in the early stages to check out the CSV data files.

The data was downloaded from the Bord Fáilte website and stored in three separate CSV files for further analysis. I created a function called get\_api\_to\_csv to get the API data in to the CSV files. This function takes in two parameters, the Urlname and the CSVName.

See below for the function and the parameter used to create each separate CSV file.

get\_api\_to\_csv('https://failteireland.azure-api.net/opendata-api/v1/accommodation/csv','accomodation.csv')

get\_api\_to\_csv('https://failteireland.azure-api.net/opendata-api/v1/attractions/csv','attractions.csv')

get\_api\_to\_csv('https://failteireland.azure-api.net/opendata-api/v1/activities/csv','activities.csv')

These CSV files were then analysed to see how complete the data was.

I wrote the function check\_out\_data to enable me to decide what needed to be cleaned in each of the CSV files. The function takes in one parameter – CSVName which is the name of the CSV file to be analysed.

This function reads a CSV file, looks at the first 5 lines of the data. It displays the rows and columns using the shape attribute.

The datatypes are listed by displaying data types.

The data is checked for null values.

The data is check for duplicate rows.

The function was called per CSV file as follows:

check\_out\_data('accomodation.csv')

check\_out\_data('attractions.csv')

check\_out\_data('activities.csv')

When I knew what data needed to be cleaned, I wrote another function to clean the data.

This is called clean\_the\_data and takes in one parameter, CSVName, the name of the CSV file to be cleaned.

There were three files to be cleaned and each one had the same columns.

This function reads the CSV file and checks for missing values.

The following columns in the files had missing values or nulls:

Url

Telephone

AddressRegion

AddressLocality

AddressCountry

These are replaced with various text fillers.

Missing Url is replace by ‘No Url Available’, Telephone by ‘No phone number’, AddressRegion and AddressLocality with ‘Unknown’ and missing Tag values with ‘General Activities’.

The AddressCounty column had inconsistent data – some ‘Ireland’, more ‘Republic of Ireland’ and some missing. So all missing values were replaced with ‘Republic of Ireland’ and all values of ‘Ireland’ were replaced with ‘Republic of Ireland’.

Finally, the function returns a DataFrame with clean data.

The function was called as follows:

df1 = clean\_the\_data('accomodation.csv')

df2 = clean\_the\_data('attractions.csv')

df3 = clean\_the\_data('activities.csv')

Three DataFrames were created at then end of the process.

The next step was to add a new column called Type to each of the dataframes. Df1 has a type of ‘Accommodation’, df2 has a type of Attraction’ and df3 has a type of ‘Activity’. This was done so that each type of data could be identified before the three DataFrames were merged into one combined DataFrame.

The three DataFrames, df1, df2 and df3 were then merged into one DataFrame called combined\_df.

The data was then represented in a number of bar charts and each tells it’s own story!

See the results section for more details on the charts.

**Machine Learning**

The purpose of machine learning is to predict the future based on past evidence.

The method I am using is Linear Regression. This is a statistical model for modelling the relationship between a set of variables. The dataset I am using is called solution.csv and is sourced from Kaggle.

It lists house prices in the US.

The csv file is imported into a DataFrame called housingdf. The data is checked using the head() attribute and the distribution is analysed using describe().transpose().

The model is then prepared into X and y variables with X having the item column and y having the SalesPrice column. The test and training split variable are then set up with 70% for training and 30% for testing.

The coefficient for the Id column is then calculated – this is -10.479594797547671.

Then the intercept is checked – this is 203414.01274057492

.

The regression model score is based on the X\_test and y\_test variables. This is

-0.0036285614387627785

The plot is then output – see below for details.

**Boosting**

In boosting the models are trained sequentially, with future models trying to improve on the performance of the previous model.

In order to test my data, the solution.csv file was imported into the housingdf2 DataFrame.

I used iteration to loop over the columns in the DataFrame to get the value\_counts of the 2 columns, Id and SalesPrice.

Unique value count of Id is 1459

Unique value count of SalePrice is 677

I then had a look at the data using the head() attribute.

I used a DecisionTreeClassifier to guage the accuracy of my data. The accuracy was 100% so I don’t think this accuracy could be improved on by using boosting.

However, if I was going to use boosting, it would be AdaBoost, as it seems to be easier to

learn for a beginner.

**Results**

**Include the charts and describe them**

This first chart is a bar chart with the total number of accommodations, attractions and activities shown in a horizontal bar chart with each type in a different colour.

By far activity is the largest category with attraction being the smallest. I think perhaps the attractions and activities could be merged as there was some overlap. However, the three separate categories are representative of each file.

Chart, bar chart, waterfall chart

Description automatically generated

The second bar chart shows the number of accommodations available per county in Ireland. This is grouped by value\_count and sorted by index in ascending order, i.e. the list of counties in Ireland. Kerry has the most accommodation available by far, with Cork ahead of Dublin.

Chart, bar chart, histogram

Description automatically generated

The third bar chart show the data for Attractions, again sorted by index on value\_count.

Kerry has the most attractions, followed by Cork with Dublin in last place!

Chart, bar chart, histogram

Description automatically generated

The fourth bar chart shows the data for activities by county. Here Dublin scores way above everyone else!

Chart, bar chart

Description automatically generated

I then wanted to see if I could create one bar chart combining the data in the three separate bar charts in one. I sorted all the counties in ascending order and used the same colours to represent each category. I then created a legend depicting the three categories and colour coded! Here’s the result:

Chart, bar chart

Description automatically generated

The first plot I created in the Machine learning demonstration is a simple scatter plot showing the distribution of the data in the X and y variables created from the housingdf DataFrame.

Chart, scatter chart

Description automatically generated

The last plot produced is a Seaborn scatter plot. The diagonal subplots is KDE or Kernal Density Estimate.

Chart

Description automatically generated

**Insights**

(Point out at least 5 insights in bullet points)

* I had never used Python before this course, and I’m impressed with the number of options there are for cleaning data.
* The data cleaning process took the longest time. However, it was worth it as the charts look much better as a result.
* It took me a while to get the bar charts to work as I wanted them to! However, the effort was worth it. The pictures are so much more descriptive than text.
* It was interesting to see the difference in the counties between accommodation available, activities and attractions. I would have liked to have had more time to dig deeper to find more correlations in the data.
* The individual charts were interesting. Cork and Kerry had the most attractions, which probably makes them the most visited for tourism. However, Dublin has by far the most activities, followed by Cork, Kerry and Galway.
* I had to source a different dataset for the machine learning experiments, as my original datasets were too text based.
* Unfortunately, the dataset I picked was not suitable to demonstrate boosting, and I just ran out of time to source another one!

**References**

**Include any references if required**

**Data source** [APIs: List - Microsoft Azure API Management - developer portal (azure-api.net)](https://failteireland.developer.azure-api.net/apis)

The data is sourced from Bord Fáilte

The data for the Machine Learning is sourced from Kaggle - <https://www.kaggle.com/datasets/carlmcbrideellis/house-prices-advanced-regression-solution-file>

The license is Creative Commons.