# Group ID - MSc in Data Analytics FT/SB+ (Feb24 start)

Author: Madina Sagatova

e-mail: [2021255@student.cct.ie](mailto:2021255@student.cct.ie)

Student ID: 2021255

**Domestic Tourism in Ireland by Irish Residents**

**Domestic Tourism in Ireland by Irish Residents****0**

[Acknowledgements](#_27g3duvfofbc) 2

[Abstract](#) 3

[Introduction](#) 3

[**Data Preparation and Visualisation**](#_v1ha3eyenlx8) **4**

[1. Data Acquisition](#_rbc1tvls7ttv) 4

[2. Data Loading](#_4p0egvjf8qp) 4

[3. Data Inspection](#_64154zhmba13) 4

[4. Data Cleaning](#_9i6tc0wd9btw) 6

[5. Data Transformation](#_yo2arfur5149) 7

[6. Data Visualisation](#_on2eexyn8tih) 9

[1.1 Total Number of Trips DataFrame](#_gt56wlvo68o4) 9

1.[2 All counties DataFrame](#_wd83e9cml7p0) 11

1.3 [Top 5 visited counties DataFrame](#_jlz051xwznt6) 13

[**Statistics**](#_i2k0oyhhtdpk) **16**

[1. Descriptive Statistics](#_uxihe9961jj3) 16

[1.1 Measures of Central Tendency](#_3uo84xvqlp4v) 16

[1.2 Measures of dispersion or variability](#_95xrrpip6uzp) 18

[1.3 Measures of shape](#_dtbeqj8e4xr3) 19

[2. Normal Distribution](#_hgqazpu8u5uq) 21

[3. Binomial Distribution](#_o50ukoskddx) 24

[4. Poisson Distribution](#_o9tbgnmrhokh) 26

[**Machine Learning for Data Analytics**](#_sljl2rtf80ga) **28**

[1. CRISP-DM Framework](#_oiljvmazi966) 28

[2. Linear Regression](#_fhvrduk30ei7) 28

[3. Lasso Regression](#_6vgrldybxng) 30

[4. Ridge Regression](#_laetq2o5uoa4) 31

[**Programming**](#_kk7m44awgqgv) **32**

[1. Imperative Programming](#_vac199nb9l06) 32

[2. Declarative Programming](#_5aqevi9aczqp) 33

[3. Functional Programming (FP)](#_czf7mje4x7pl) 33

[4. Procedural Programming](#_wynwnol8a31d) 34

[5. OOP (Object Oriented Programming)](#_yhwnnwqt6jkr) 34

[References](#) 35

## 

## 

## 

## 

## 

## 

## 

## 

## 

## Acknowledgements

# I would like to express my sincere gratitude to the CCT College for giving me an opportunity to pursue the Masters programme in Data Analytics.

I am particularly grateful for excellent and individual teaching methods, support and invaluable experience I received from following professors:

David McQuaid, Sam Weiss, Muhammad Iqbal and Taufique Ahmed/Marina Iantorno.

I also want to extend my thanks to the Central Statistics Office (CSO), their creators and maintainers for their invaluable contribution to this project. The resources provided by the CSO have been instrumental for gathering and exploring the data.

Sincerely,

Madina Sagatova.

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

## Abstract

*Local tourism plays a crucial role in the domestic economy, sustainability of countries and income distribution. This study investigates the trend of domestic tourism in Ireland by Irish residents, impact of the COVID-19 on local tourism and analyzes the most popular counties within the tourism context since 2010.*

*Using different approaches, such as Exploratory Data Analysis (EDA), Statistics, Machine Learning (ML) and Python Programming language, we examined the domestic tourism patterns, including travel destinations, top five popular counties in Ireland by residents and results of COVID-19 in 2019 - 2021. Our findings reveal that Ireland has experienced steady growth in domestic tourism in the period from 2010 to 2022 (based on the data from 2010 to 2022 of domestic tourism by Irish Residents from the Central Statistic Office). However the COVID-19 led to a significant decrease in the number of trips taken within the country from 2019 to 2022. Applying Machine Learning techniques helped us to construct predictive models in order to forecast potential changes in local tourism by the end of 2030.*

## Introduction

Attractive domestic tourism by locals helps to promote a positive image of Ireland overseas, and contributes to employment across the country. Compared to international tourism, local travels usually involve lower distances and reduce long-hours flight which minimizes the environmental impact associated with transportation and makes it environmentally sustainable. In this research project, we investigated local tourism in Ireland by residents from 2010 to 2022, applying data visualization and preparation techniques, statistical analysis, ML approaches and programming skills to gain apprehension about travel behavior, trends and preferences.

We followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) project management framework to achieve success with research objectives. The CRISP-DM framework consists of 6 stages in a cycle. The first phase of our research is Research Understanding (Business Understanding in the origin) which involves determining research objectives and goals, research success criteria and producing a project plan.

In the next phase called Data Understanding we performed the following procedures: collecting initial data about domestic tourism in Ireland from the CSO website, describing the data, exploring the data (performing EDA) by using Python libraries such as Pandas and performing summary statistics and verifying data quality.

The third phase is Data Preparation. In this phase we selected appropriate features, cleaned the data (handle missing values and duplicates), constructed the data by using Normal, Binomial and Poisson Distributions in order to explain or identify some information about our dataset, integrated and formatted the data.

Modelling is the next phase and during this stage we chose modeling techniques such as Linear, Redge and Lasso Regressions for forecasting changes in local tourism, generated test design, built and assessed models.

The fifth phase is Evaluation. The stage included evaluating the results, reviewing the results and determining the next actions and decisions.

The last phase is Deployment. We skipped the deployment, monitoring and maintenance (not a business project) parts. We produced the final report and reviewed the project.

The main objectives of this research project were:

* How has the domestic tourism in Ireland by Irish residents changed in the period from 2010 to 2022?
* What are the most visited counties in Ireland?
* Forecast tourism in Ireland until 2030
* Is there a correlation between tourism and fuel cost?

We divided the report into four parts where we explain all approaches in detail:

1. Data Preparation and Visualization
2. Statistics
3. Machine Learning for Data Analytics
4. Programming

The project on GitHub: <https://github.com/CCT-Dublin/ca1-madinasagatova>

# Data Preparation and Visualisation

## Data Acquisition

We obtained and ensured that the data is complete and relevant to the analysis objectives. All data analysis manipulations were conducted using the Jupiter Notebook.

Dataset in the form of .csv file:

* IrishResidentsDomesticTravel.csv
* Database resource: <https://data.cso.ie/table/HTA11>

## Data Loading

Loaded the data into the appropriate data environment. We used appropriate functions and libraries:

* Pandas library
* .read\_csv (“fileName.csv”) function

df = pd.read\_csv("IrishResidentsDomesticTravel.csv")

## Data Inspection

Included examination of the structure and contents of the dataset, understanding the features, data types and any potential issues such as missing values, outliers or duplicates.

* .head() function checks the first 5 observations

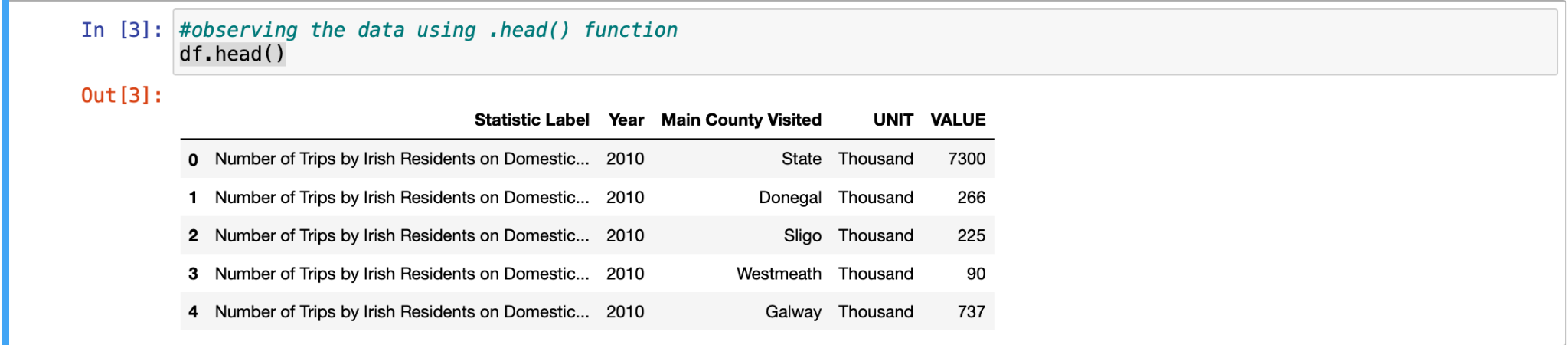


Figure 1. Usage of .head() function

* .shape - shows the shape of the dataset(numbers of rows, number of columns)

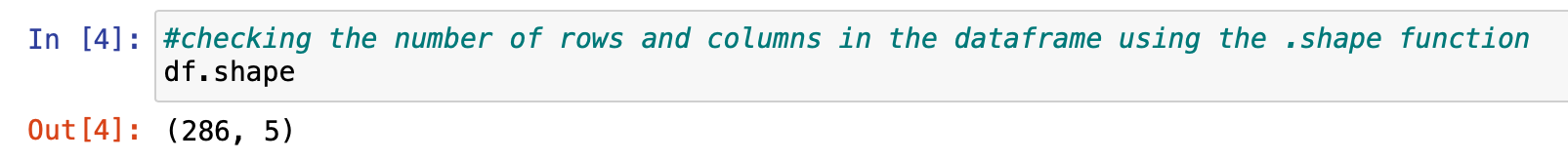


Figure 2. Usage of .shape attribute

* .info() function checks the data types of the features, checks non-null values and shows the memory usage

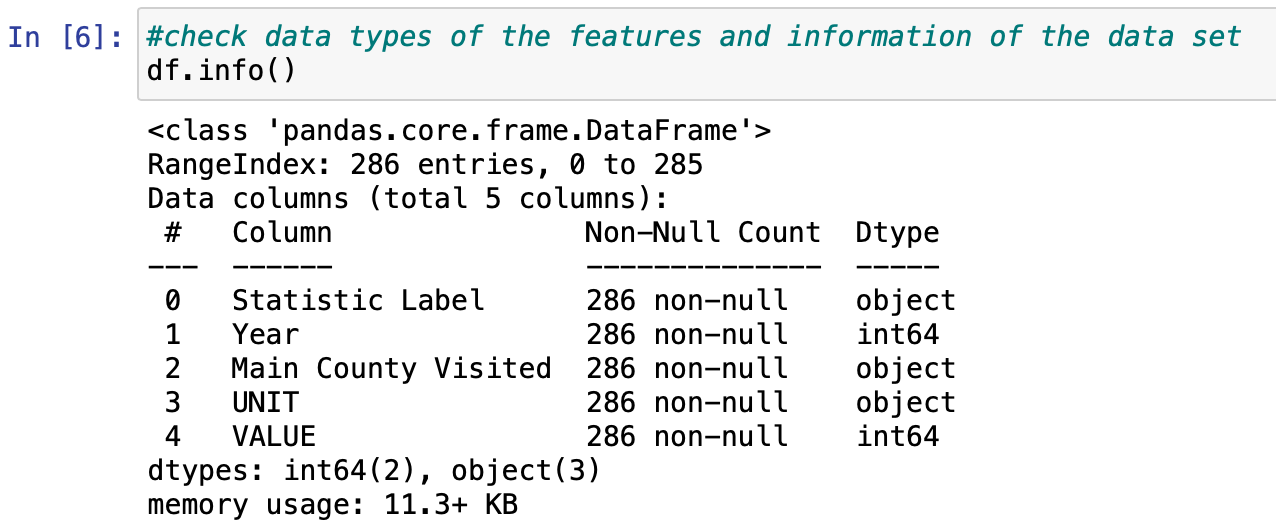


Figure 3. Usage of .info() function

* .duplicated() - check if there are any duplicates

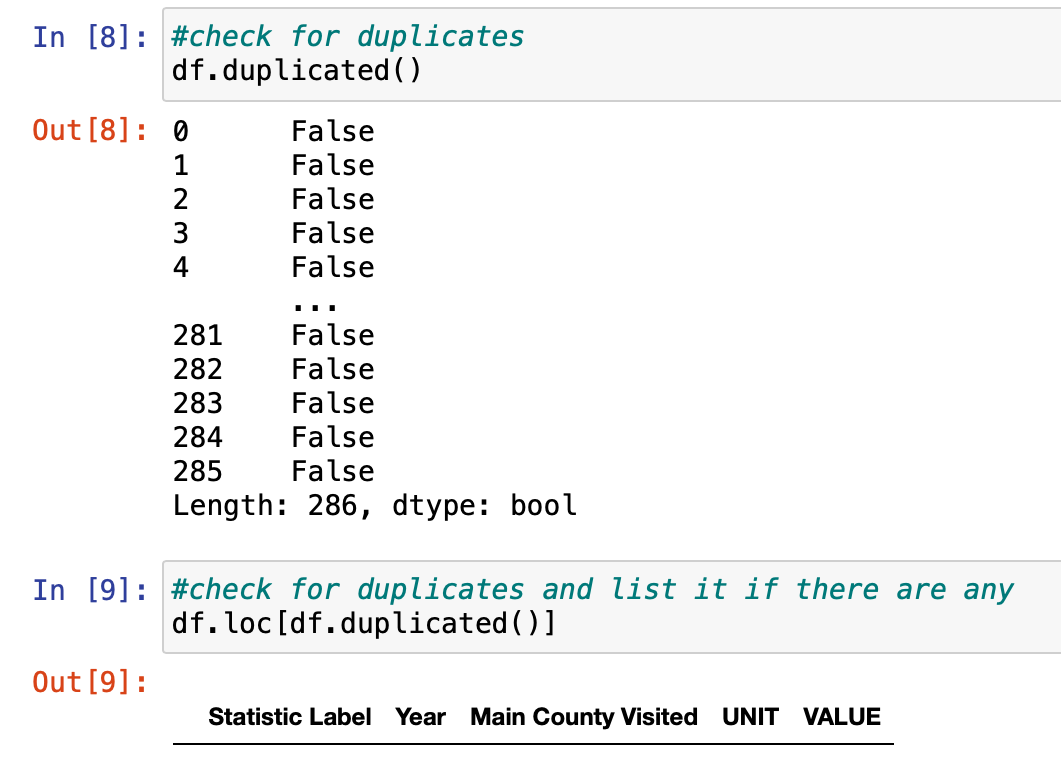


Figure 4. Duplicates using of .duplicated() function

## Data Cleaning

During the data cleaning we handle missing values, outliers or inconsistency in the dataset. It involves imputing missing values, deleting duplicates and standardizing variables as needed.

* Create a function get\_unique\_values(df, column\_name) - function returns a column name an array with unique values in the column

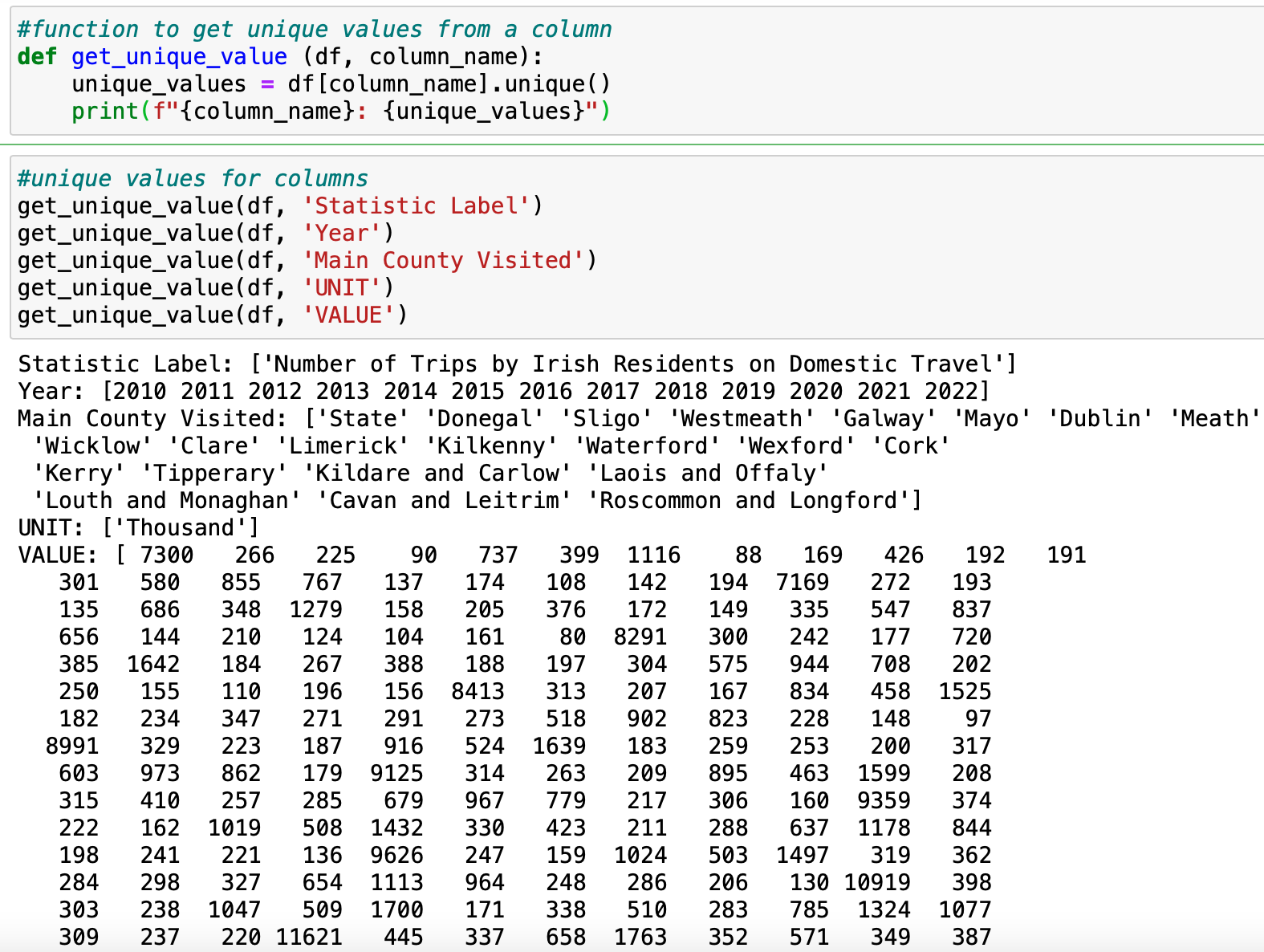


Figure 5. get\_unique\_values(df, column\_name)

We dropped two columns: “*Statistic Label*” and “*UNIT*”, because they obtained only one unique value “*Number of Trips by Irish Residents on Domestic Travel*” and “*Thousand*” respectively.

* .drop(columns =[“column\_name”], axis = 0 )

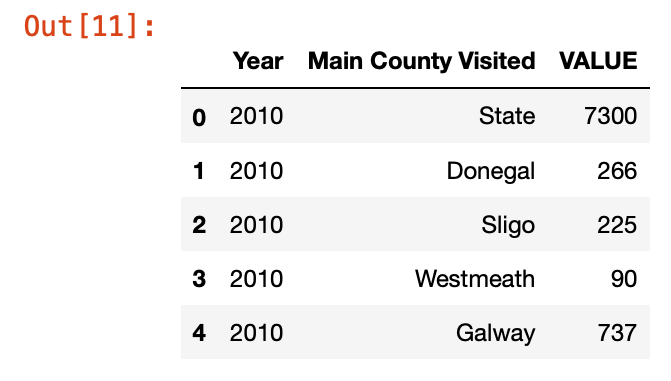


Figure 6. DataFrame after dropping “Statistic Label” and “UNIT” columns

We renamed the column '*Main County Visited'* to '*Counties'* and the column ‘*VALUE’* to

*‘Number\_Of\_Trips’*

* .rename(columns={'oldColumName:'newColumnName'})

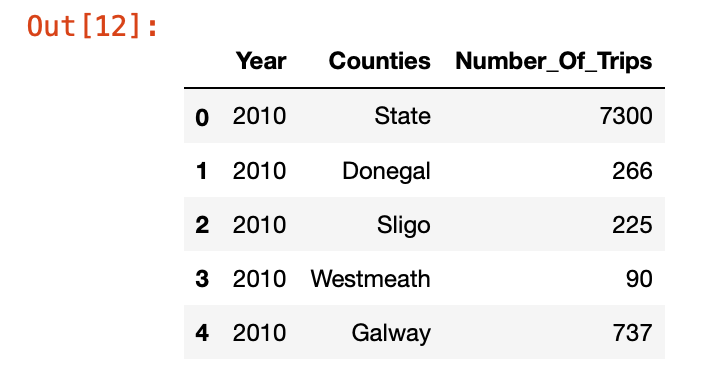


Figure 7. DataFrame after renaming columns

We converted the column “*Number\_Of\_Trips*” into thousands using.

* A function was created in order to convert numbers into thousands

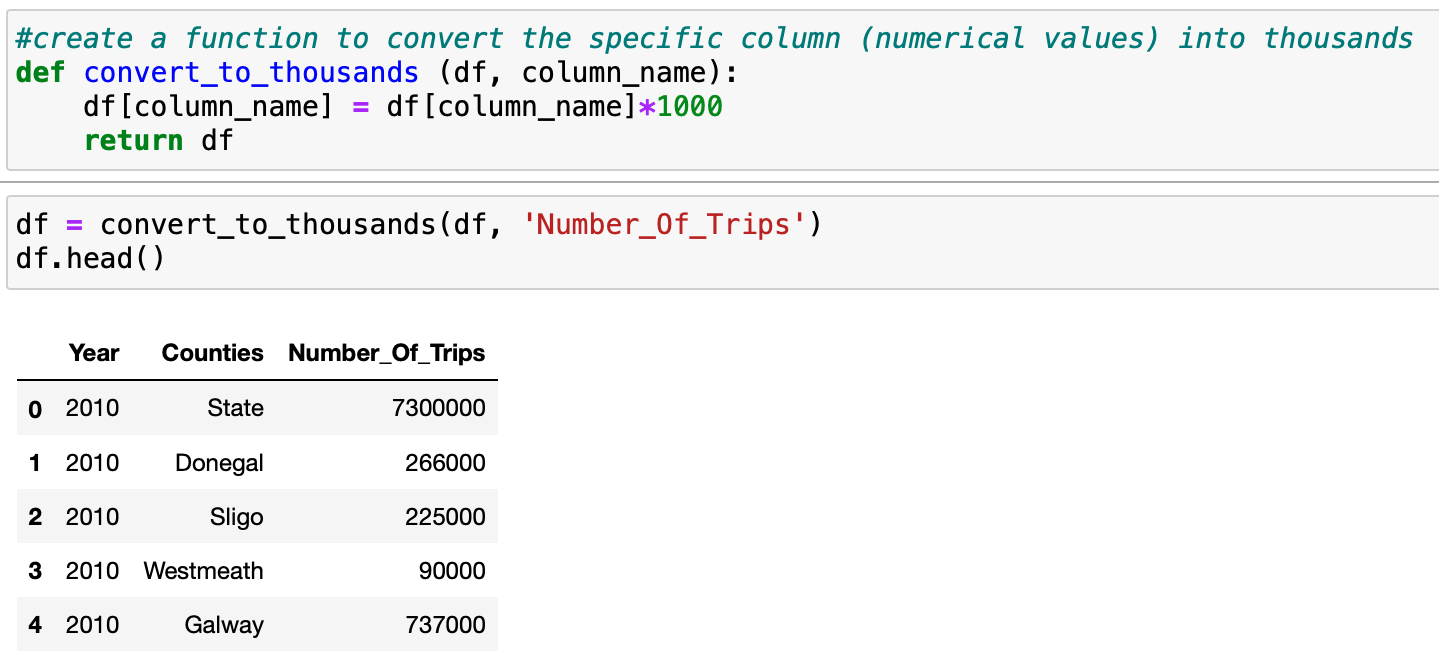


Figure 8. Function convert\_to\_thousands(df, column\_name)

We renamed the value “*State*” in the column “*Counties*” into “*Total*” using the .replace() function.

* df['Counties'] = df['Counties'].replace(['State'], 'Total')

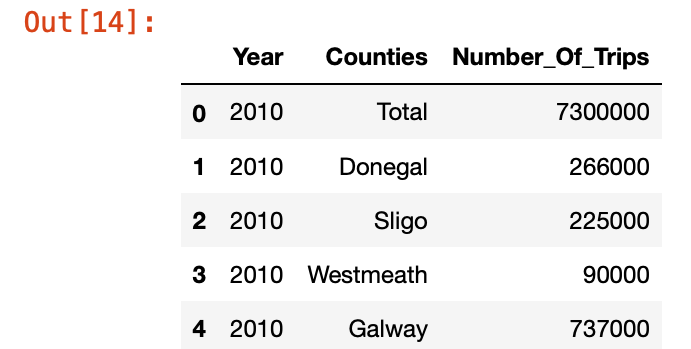


Figure 9. DataFrame after replacing “State” to “Total”

## Data Transformation

We decided to split the dataset into three datasets:

1. total\_df (contain only total values of trips) using .isin() function

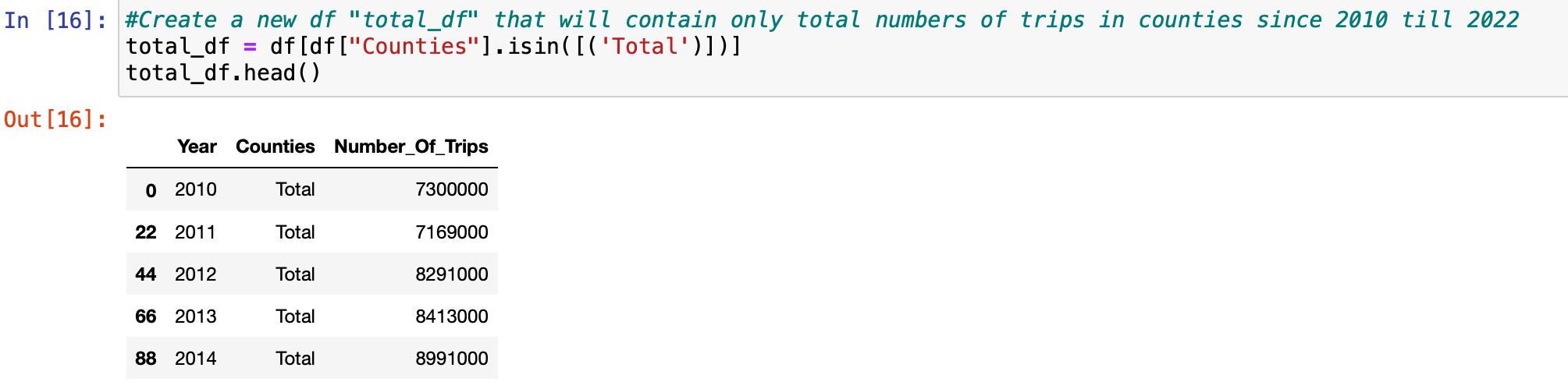


Figure 10. DataFrame “total\_df “

1. counties\_trips\_df (all counties without the total values) using str.contains(“Total”) == False

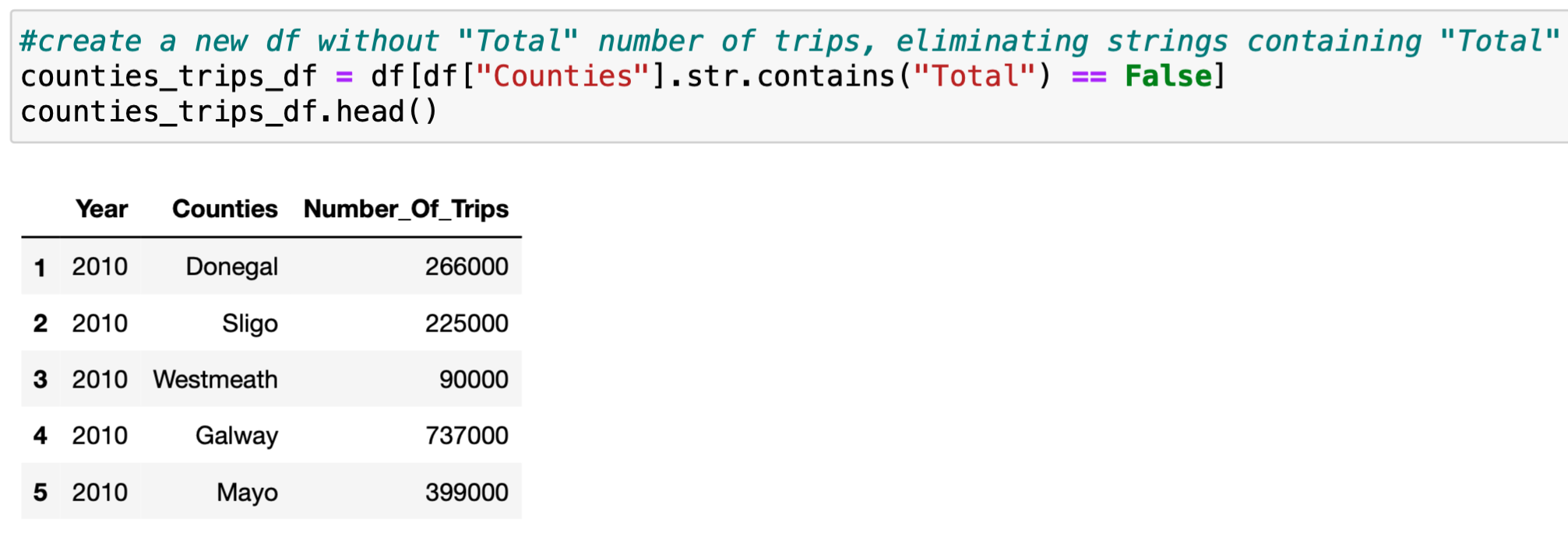


Figure 11. DataFrame “counties\_trips\_df “

1. top\_5\_counties\_df (only five top visited counties )

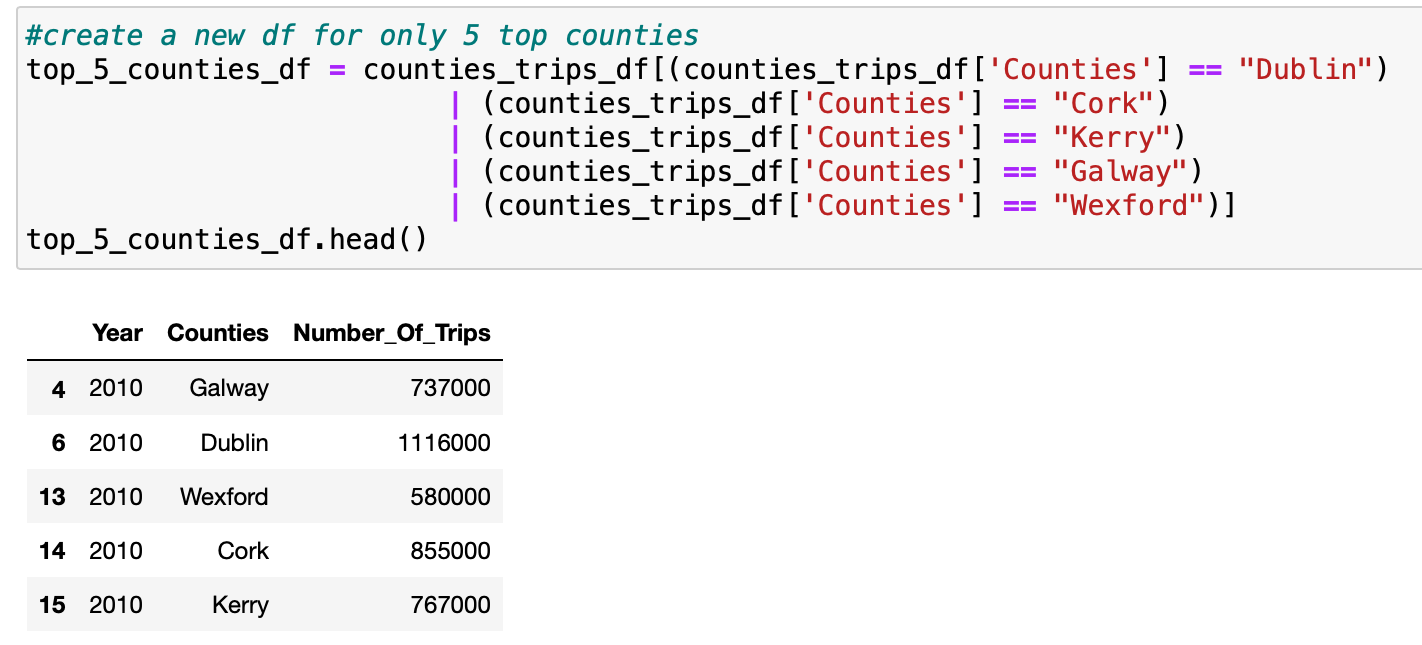


Figure 12. DataFrame “top\_5\_counties\_df“

We used pivot() in order to reshape our DataFrame. We rearranged the unique values of the column "Counties" to the five columns.

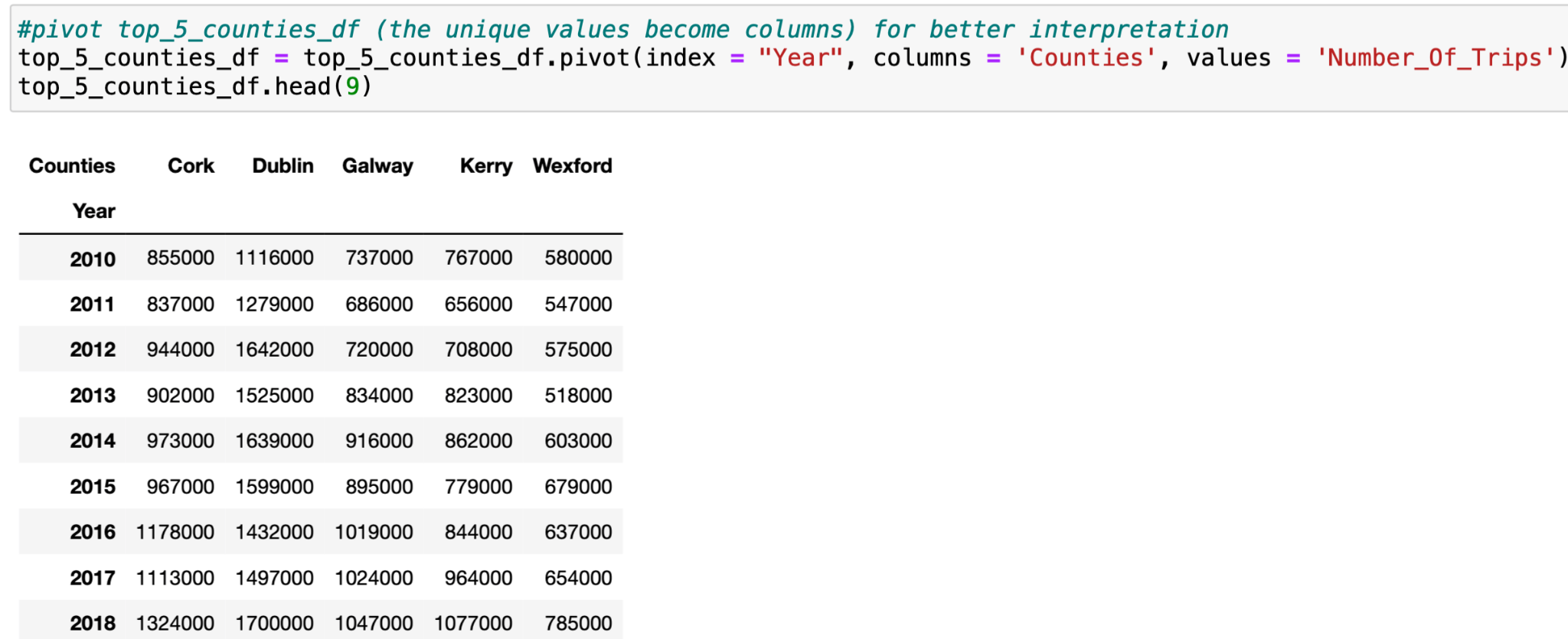


Figure 13. DataFrame “top\_5\_counties\_df“ after applying .pivot()

## Data Visualisation

#### 1.1 Total Number of Trips DataFrame

Once the data is cleaned and prepared we can start to visualize key aspects of the dataset to gain some insight. We used next techniques such as line graphs, histograms, bar plots, box plots and heatmaps.

* A line graph to visualize how the total number of trips changes in the period 2010 - 2022.

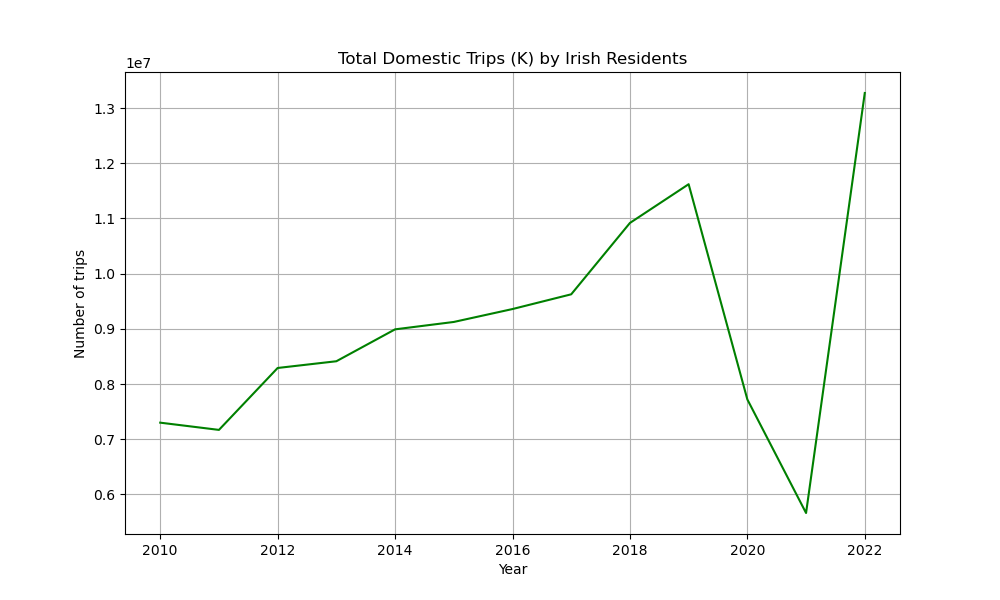


Figure 14. Linear graph of Total Domestic Trips(K) by Irish Residents

The total number of trips in counties by Irish Residents were increasing steadily. However they significantly dropped down during the COVID-19. And we can see how the graph reached its peak in 2022.

* A histogram of how distributed the values in the column “Number\_Of\_Trips”.

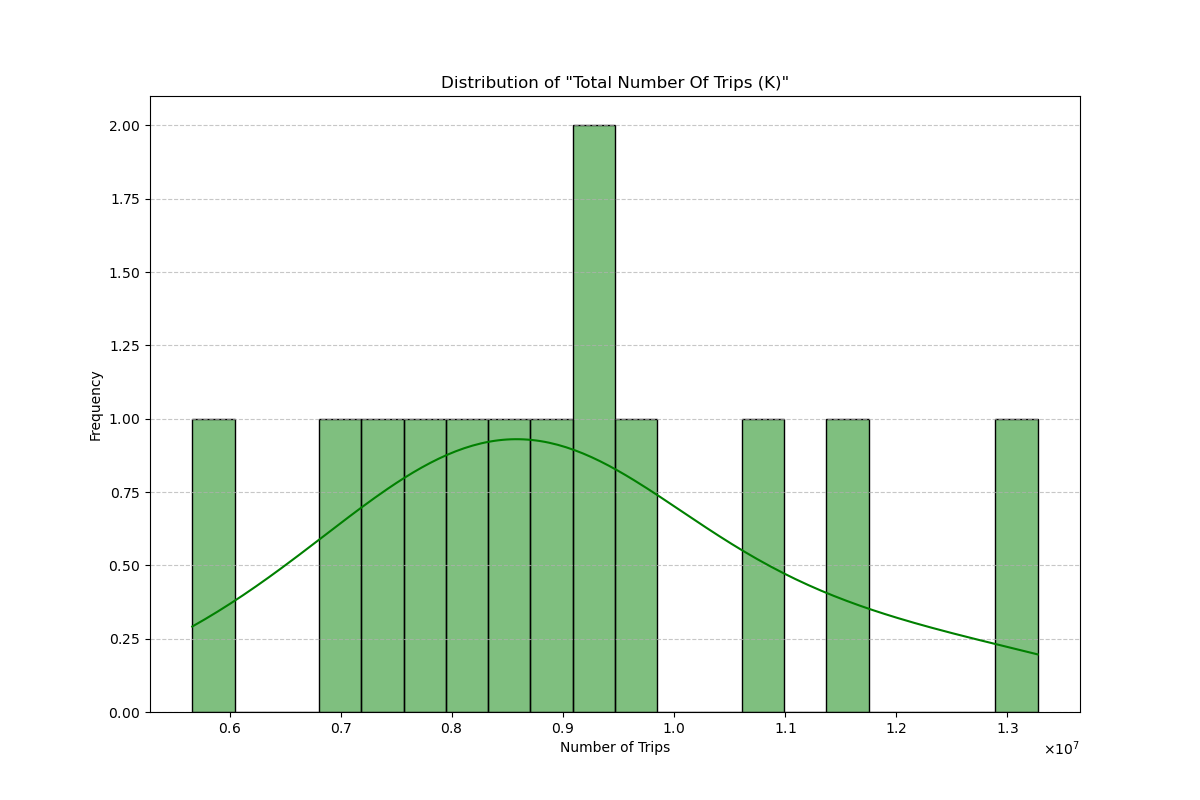


Figure 15. Histogram of Distribution of “Total Number Of Trips(K)”

The histogram shows that the data in the "Number\_Of\_Trips" column is almost uniformly distributed with one outlier.

* Two box plots for identifying potential outliers and displaying the distribution.

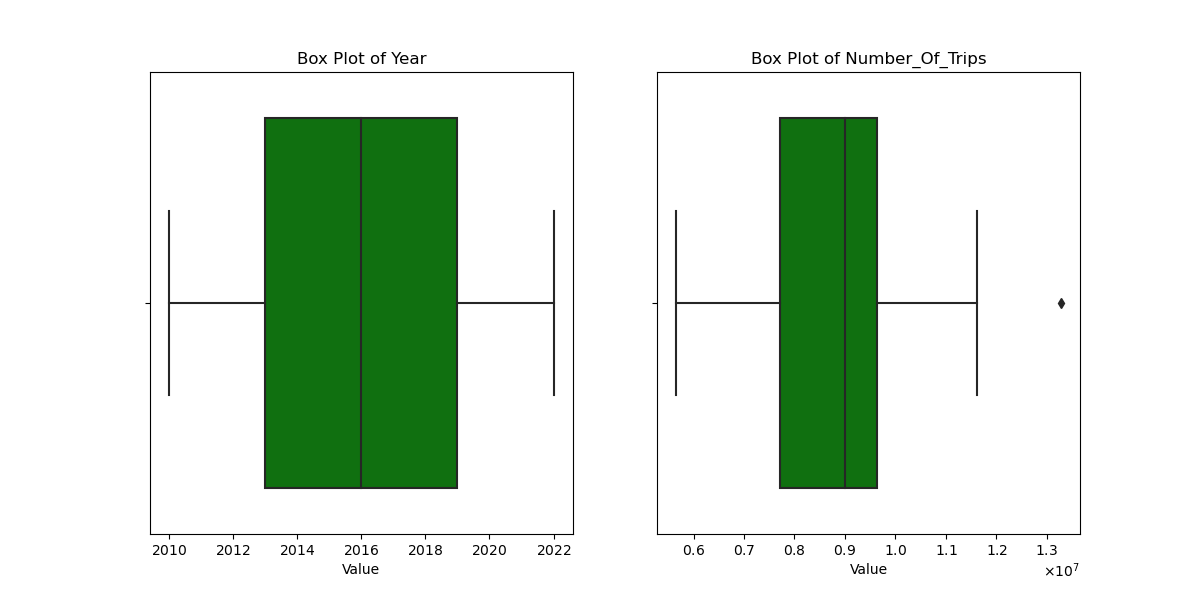


Figure 16. Box Plots of “Year” and “Number\_Of\_Trips” in “total\_df” DataFrame

The box plot of the "Year" column shows a normal distribution with no outliers. However, the Box plot of the "Number\_Of\_Trips" has negative skewness (the median is closer to the top) and outliers.

### 

#### 1.2 All counties DataFrame

* The Bar Plot of average domestic trips for all counties since 2010.

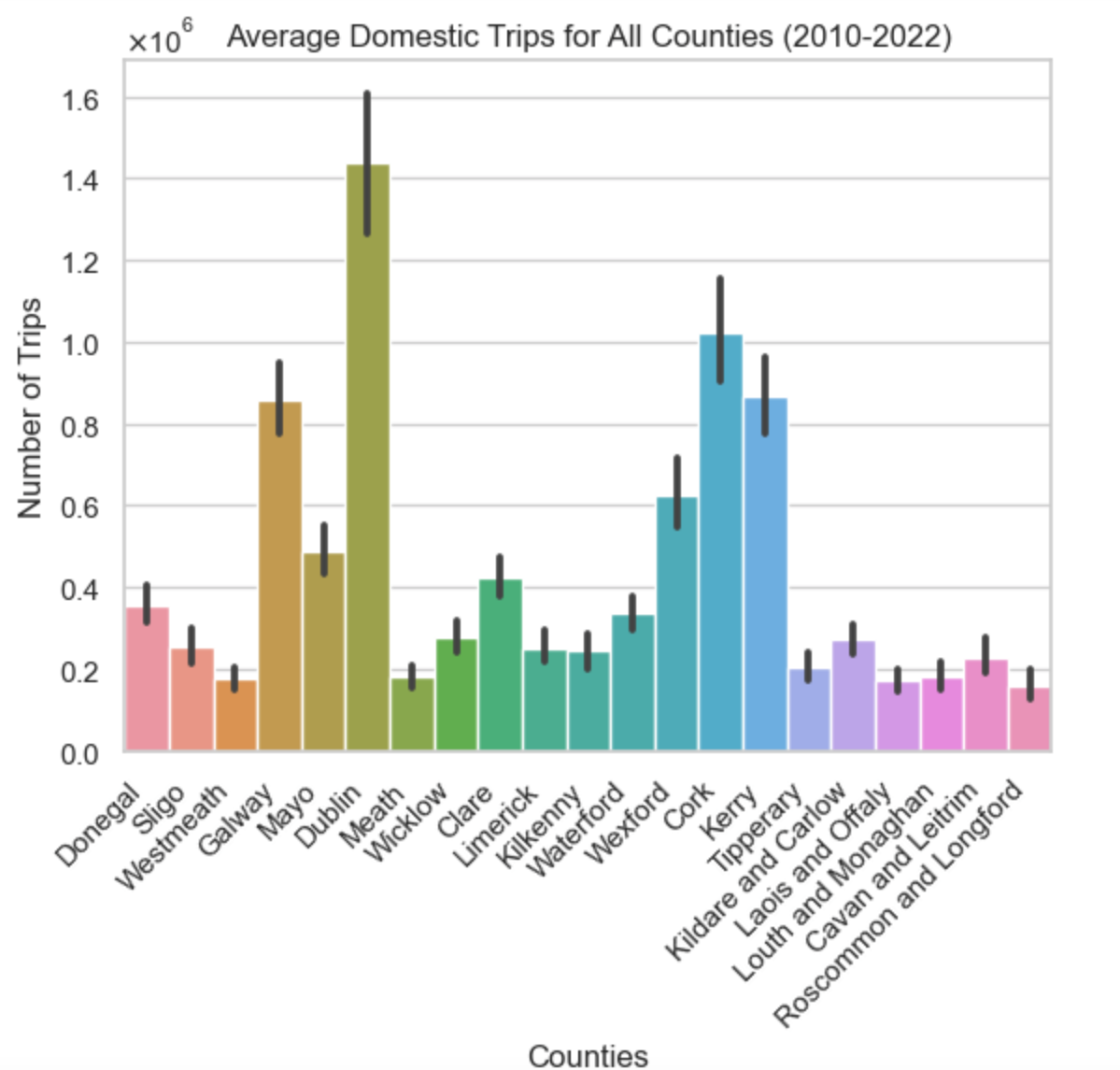


Figure 17. Bar Plot of average domestic trips for all counties

There are 5 counties that were most visited since 2010: Dublin, Cork, Galway, Wexford and Kerry.

* A histogram of how distributed the values in the column “Number\_Of\_Trips” for all counties in the period 2010-2022.

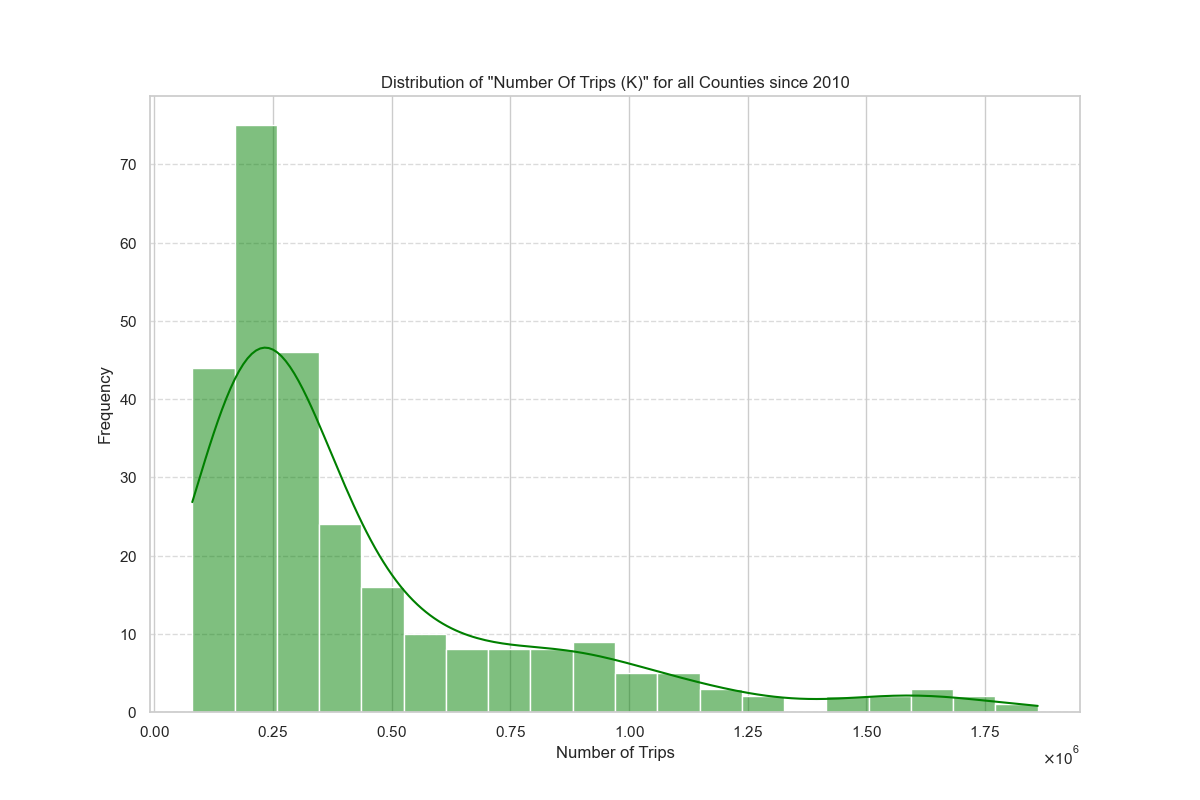


Figure 18. Histogram of Distribution of “Number\_Of\_Trips” in “counties\_trips\_df ” DataFrame

The histogram shows that values in the column “Number\_Of\_Trips” are positively skewed (skewed right) with many outliers.

* Two box plots for identifying potential outliers and displaying the distribution.

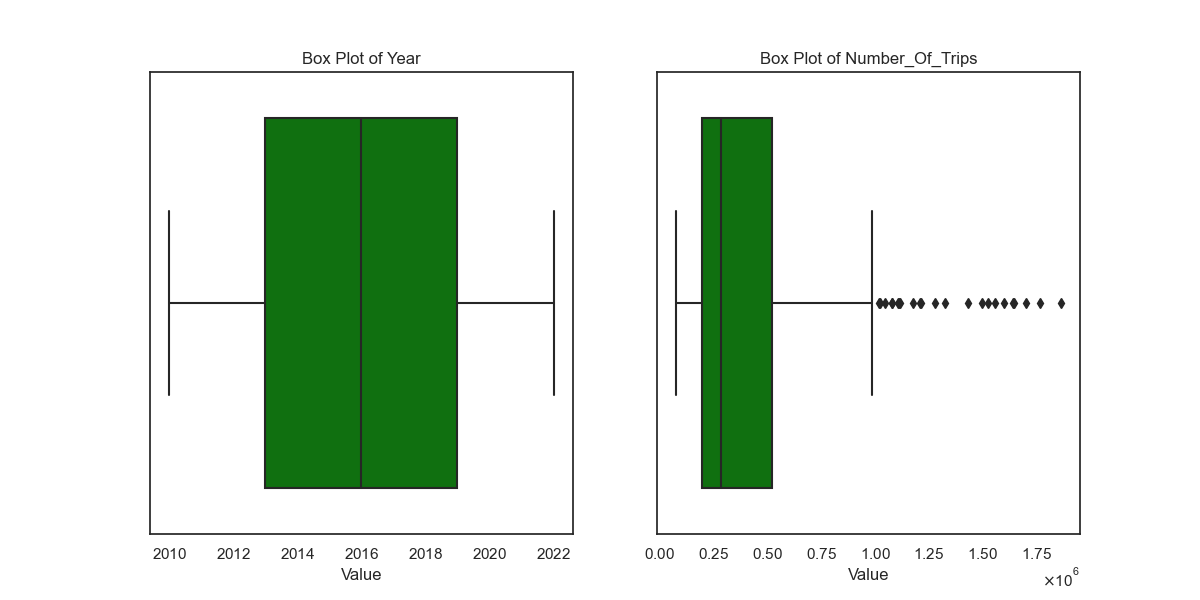


Figure 19. Box Plots of “Year” and “Number\_Of\_Trips” in “counties\_trips\_df ” DataFrame

As we said above the box plot of “Number\_Of\_Trips” has many outliers and is positively skewed to the right. The “Year” values are normally distributed.

#### 1.3 Top 5 visited counties DataFrame

* The Bar Plot of domestic trips for the top five counties since 2010.

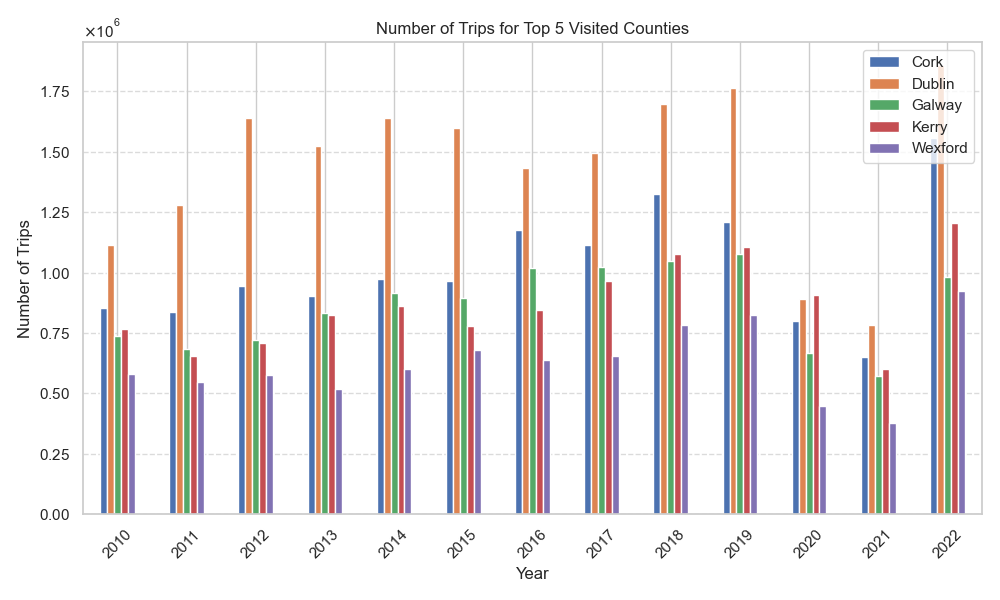


Figure 20. Bar Plot of domestic trips for the top 5 counties

We can see the impact of COVID-19 in 2020 and 2021. The most visited county was Dublin. However in 2020 during the lockdown Kerry was the most visited county.

* Histograms of how distributed the values of each county in the top\_5\_counties\_df since 2010.

How distributed the values for each county based on histograms:

Dublin: Left-skewed distribution(Negatively-skewed)

Cork: Right-skewed distribution(Positively skewed)

Galway: Left-skewed distribution(Negatively-skewed)

Wexford: Slightly right-skewed distribution(Positively skewed)

Kerry: Slightly right-skewed distribution(Positively skewed)

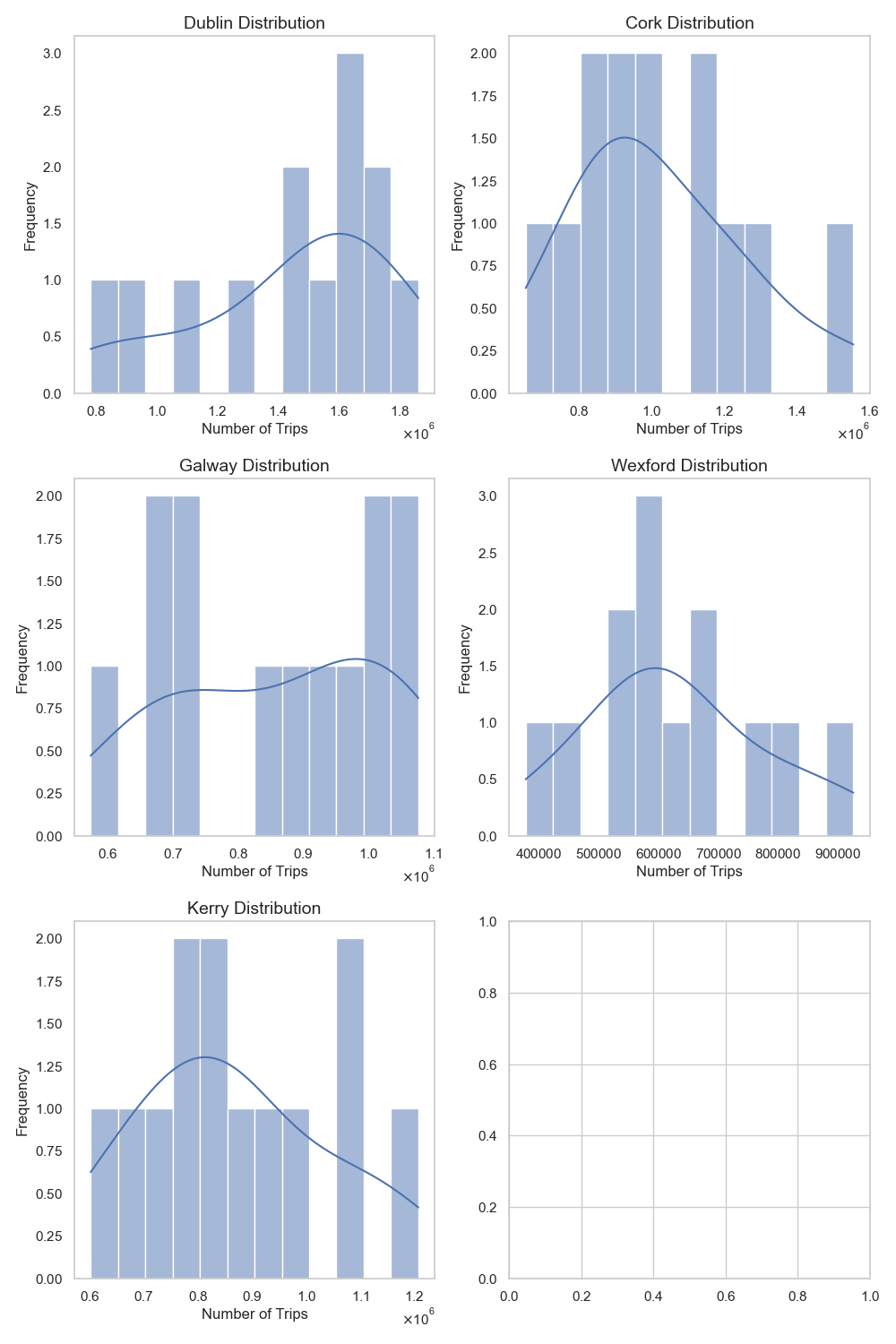


Figure 21. Histogram of Distribution of “Number\_Of\_Trips” for the top 5 counties

* Box plots for the top 5 visited Counties: Dublin, Cork, Galway, Wexford and Kerry.

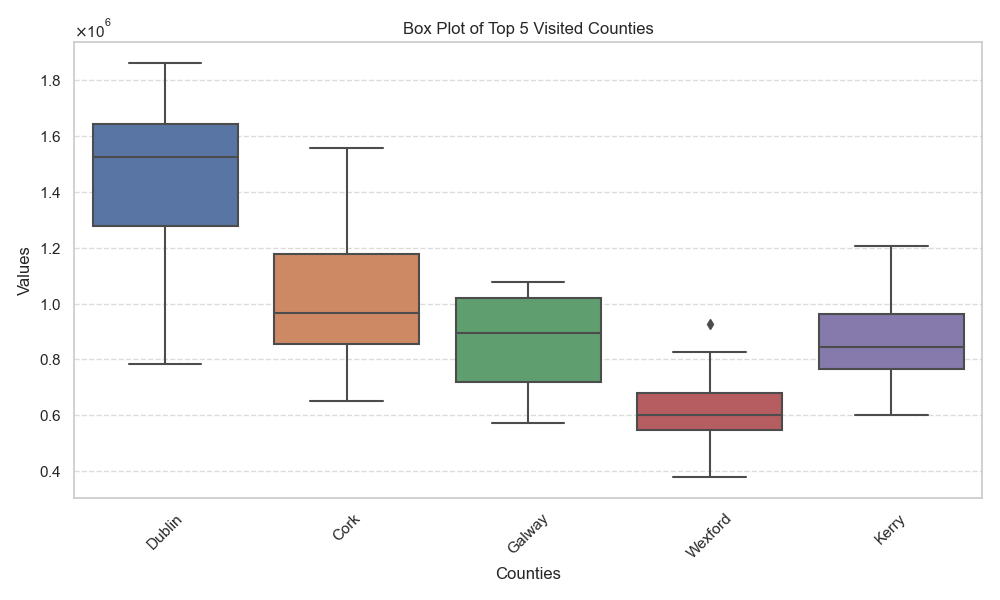


Figure 22. Box Plots of “Number\_Of\_Trips” for the top 5 counties

The Box Plot above shows that:

Dublin has negatively skewed distribution (skewed to the left, median is closer to the top of the box).

Cork has positively skewed distribution (skewed to the right, median is closer to the bottom of the box).

Galway has almost normal distribution, but we can see that the median is a little bit closer to the top and skewed to the right - negatively skewed distribution.

Wexford has positively skewed distribution and some outliers. Nevertheless the distribution of Wexford values appears to be nearly normal.

Kerry’s distribution is skewed to the right - positively skewed.

* The heatmap shows us a correlation between counties.

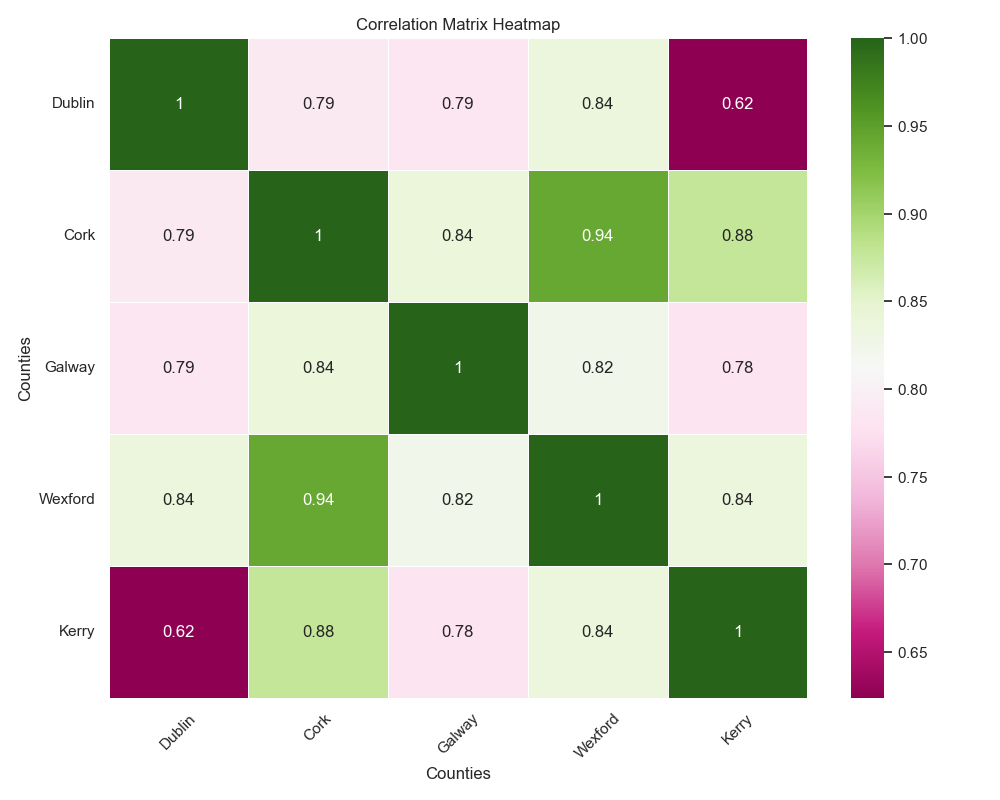


Figure 23. Heatmap correlation

# Statistics

## Descriptive Statistics

Descriptive statistics play a crucial role in data analysis. It gives us a basic summary measure of our dataset and allows researchers and analysts to quickly understand patterns and variability in the data. Summary measures are:

* Measures of central tendency such as mean, median and mode
* Measures of variability such as variance, standard deviation, minimum and maximum values, Interquartile Range (IQR)
* Measures of shape such as skewness and kurtosis.

Based on the Visualisation part, it was decided to observe the top 5 visited counties in Ireland using Descriptive Statistics.

#### 1.1 Measures of Central Tendency

Measures of Central Tendency are also known as statistical average. There are the most common measures of central tendency: mean, median and mode.

* **Mean**

Mean is the most common measure and known as a simple average. It is the value that is derived by summing all the values and dividing it by the number of observations[https://ncert.nic.in/textbook/pdf/legy302.pdf].

We calculated mean values for each county using .mean()

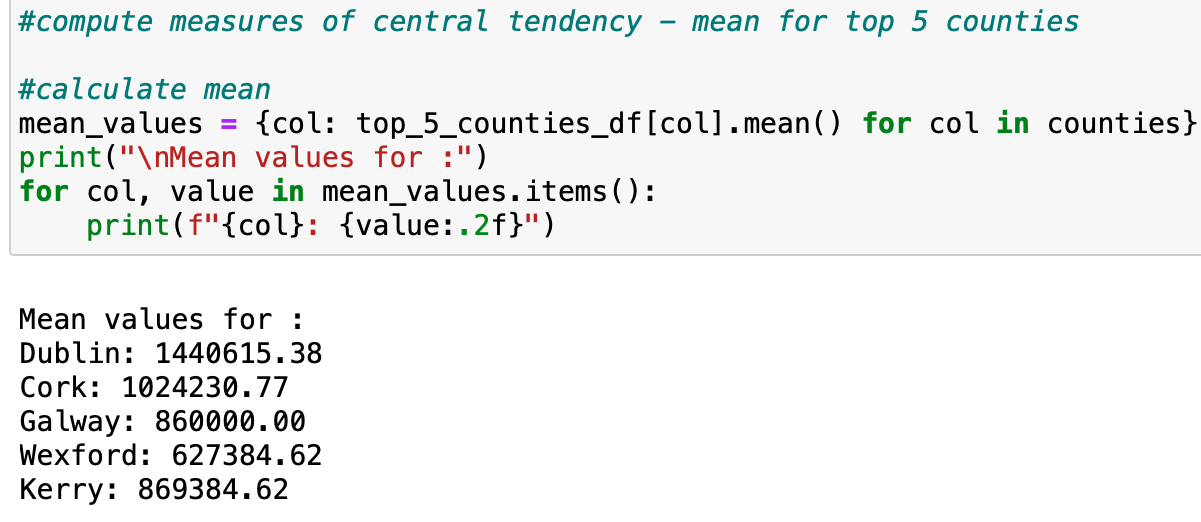


Figure 24. Mean values for the top 5 counties

* **Median**

Median is the value of the rank, which divides the arranged series into two equal numbers. To calculate the median, we have to arrange our dataset of n numbers in ascending or descending order. Median is preferred for the skewed distribution or there are concerns about outliers.

We calculated median values for each county using .median()

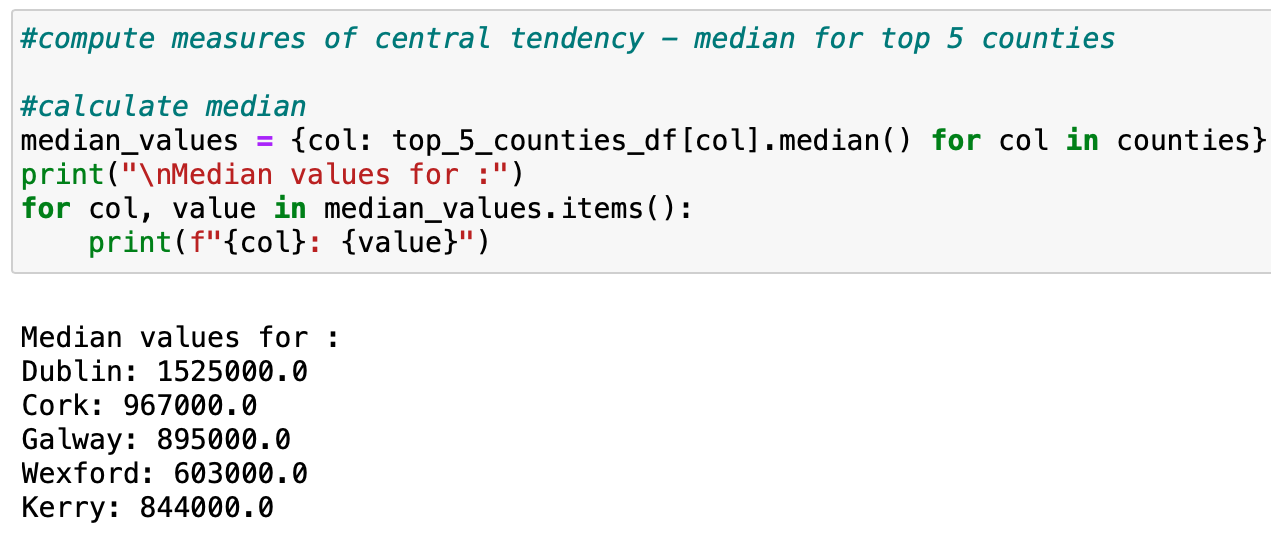


Figure 25. Median values for the top 5 counties

* **Mode**

Mode is the maximum frequency or occurrence at a particular point or value.

We did not calculate mode values for each county due to the values being numerical not categorical.

***Insights****:*

After obtaining the mean and median values we compared them for each county:

Dublin: mean < median - skewed distribution to the left.

Cork: mean > median - means positively skewed distribution to the right.

Galway: mean < median - skewed distribution to the left.

Wexford: mean > median - means positively skewed distribution to the right.

Kerry: mean > median - means positively skewed distribution to the right.

The results align with what we observed earlier visualising the box plots for each county.

#### 1.2 Measures of dispersion or variability

They provide information about how spread out the values are from the central tendency. Common measures of dispersion:

* **Range**

It is a difference between the maximum and the minimum values given in a dataset.

We calculated the Range by subtracting the minimum values (found using . min()) from the maximum values (using .max()).

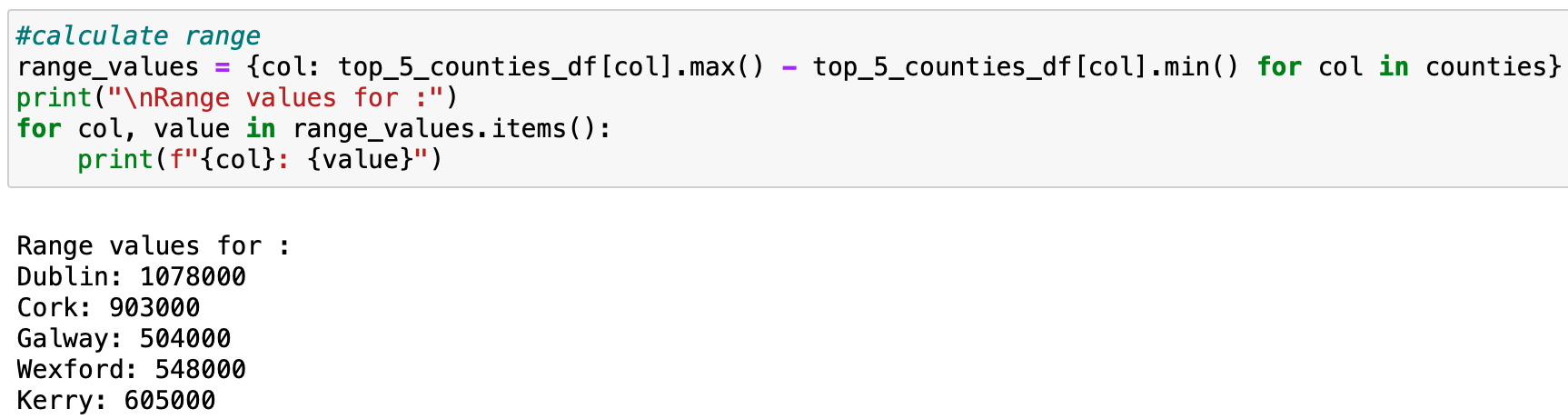


Figure 26. Range values for the top 5 counties

* **Variance**

It is a mean of the squares of the individual deviations. Gives the result in the original units squared.

We obtained the Variance values by applying the .var()

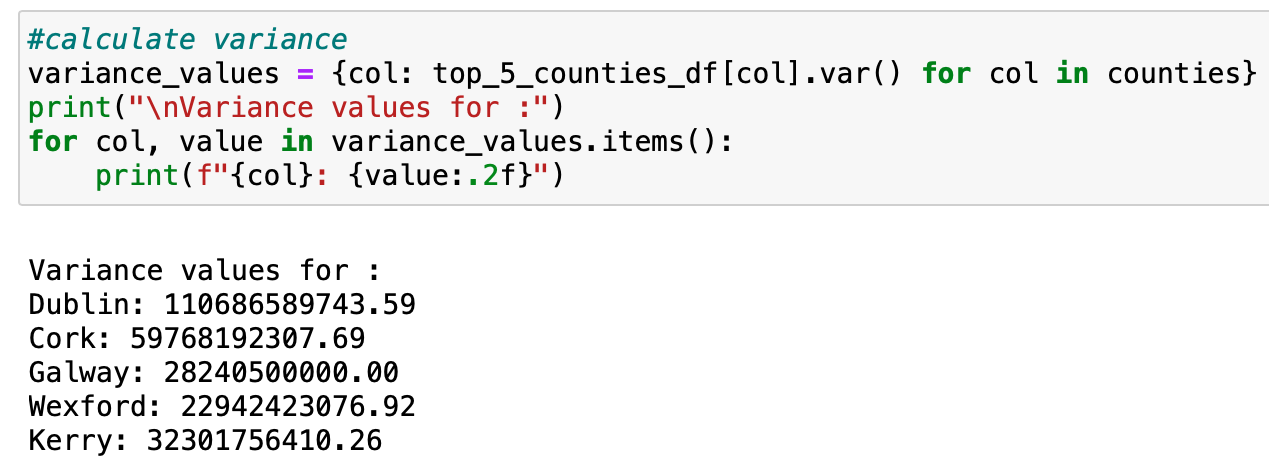


Figure 27. Variance values for the top 5 counties

* **Standard Deviation**

Standard Deviation is the most commonly used measure of dispersion. The square root of variance is known as the Standard Deviation.

The Standard Deviation results using .std()

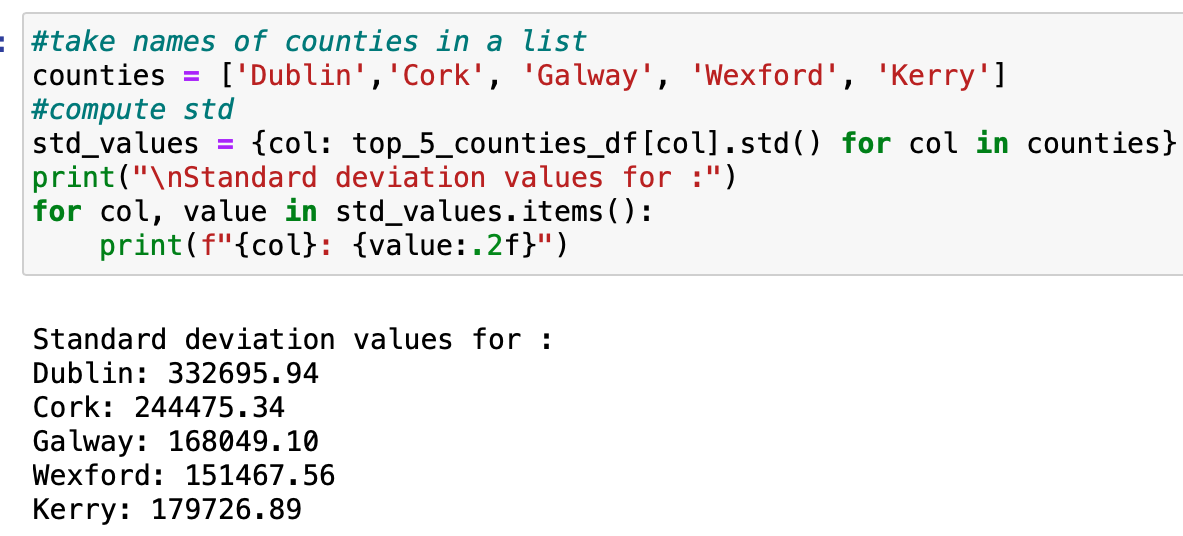


Figure 28. Standard Deviation values for the top 5 counties

* **Interquartile Range(IQR)**

Interquartile Range(IQR) is the difference between upper and lower quartiles. It is a more robust measure of spread than the variance and standard deviation.

IQR = Q3-Q1

The first step was to calculate the Q1 and Q3 quartiles using .quantile(0.25) and .quantile(0.75) respectively. The following step was subtraction.

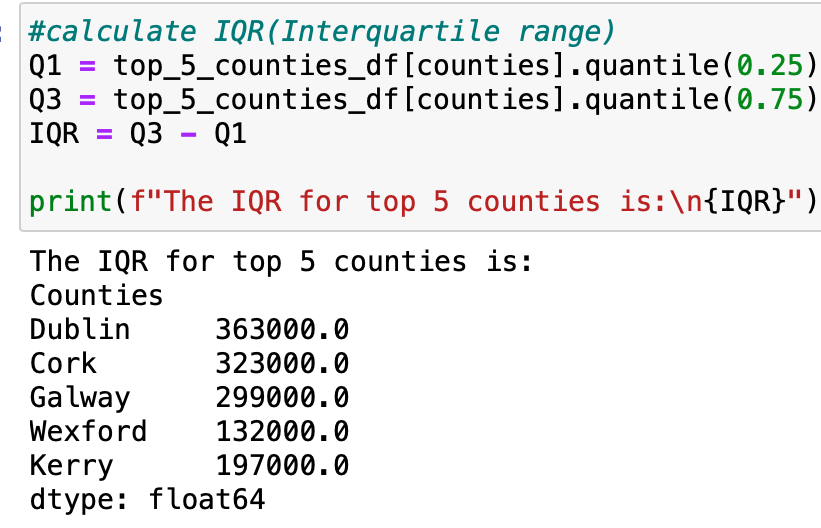


Figure 29. Interquartile Range(IQR) for the top 5 counties

#### 1.3 Measures of shape

Skewness and Kurtosis are two statistical measures that can be used to describe the shape of distribution.

* **Skewness**

Skewness means lack of symmetry. In statistics, a distribution is called symmetric if:

Mean = Median = Mode

If the right tail is longer - the distribution is positively skewed:

Mean > Median > Mode

If the left tail is longer - the distribution is negatively skewed:

Mean < Median < Mode

We obtained the values of skewness for each county using .skew(). After we used the rule of thumb for skewness values:

If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.

If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed.

If the skewness is less than -1 or greater than 1, the data are highly skewed.

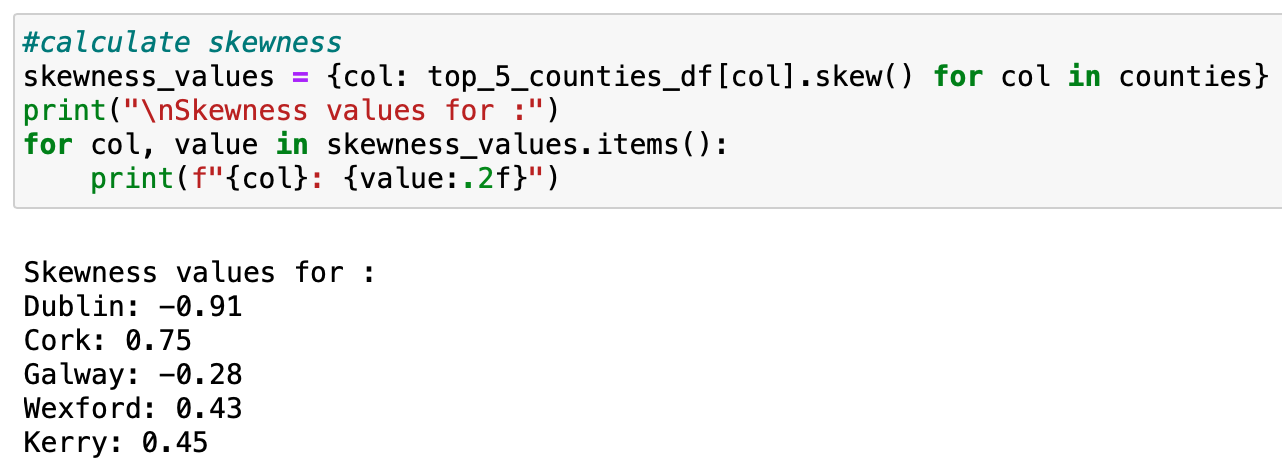


Figure 30. Skewness values for the top 5 counties

***Insights:***

Dublin: skewness = -0.91 the data is moderately skewed.

Cork: skewness = 0.75 the data is moderately skewed.

Galway: skewness = -0.28 the data is fairly skewed.

Wexford: skewness = 0.43 the data is fairly skewed.

Kerry: skewness = 0.45 the data is fairly skewed.

* **Kurtosis**

Kurtosis is a measure of the peakedness of a distribution. There are 3 types of the distribution:

**Leptokurtic**: Sharply peaked with fat tails, and less variable.

Kurtosis > 3 (Excess kurtosis > 0)

**Mesokurtic**: Medium peaked

Kurtosis =3 (Excess kurtosis = 0)

**Platykurtic**: Flattest peak and highly dispersed.

Kurtosis < 3 (Excess kurtosis < 0)

The kurtosis values were obtained by using kurtosis() from the statistical package of Scipy.

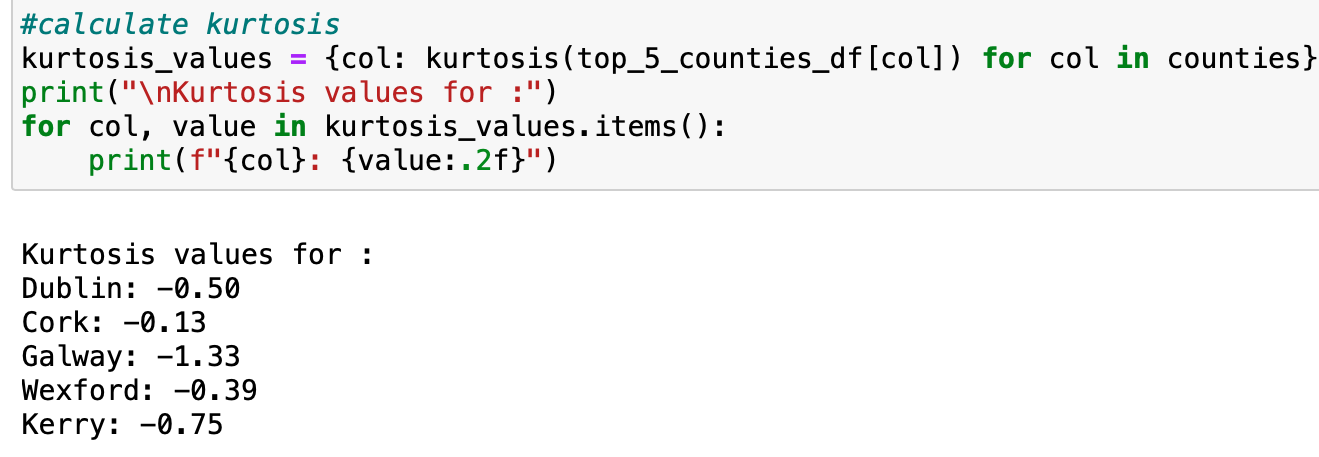


Figure 31. Kurtosis values for the top 5 counties

***Insights*:**

Kurtosis values for all five counties < 3 which means they all have Platykurtic curves. Compared to normal distribution their central peaks are broader and lower, and their tails are shorter and thinner.

The bar plot of kurtosis values:

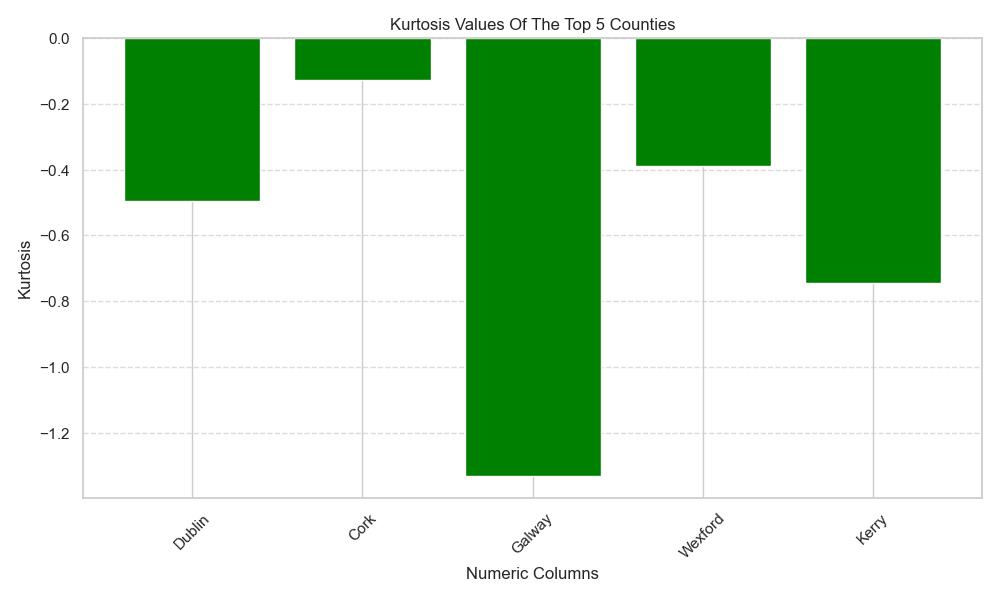


Figure 32. Bar plot of kurtosis values for the top 5 counties

## Normal Distribution

Normal Distribution or Gaussian Distribution is known as a bell-shaped symmetric probability distribution. Many things are actually normally distributed, or very close to it such as height or IQ score. The Empirical Rule also called 66-95-99.7 Rule says: In a normal distribution, approximately 66% of the data falls within one standard deviation of the mean, 95% falls within two standard deviations, and 99.7% falls within three standard deviations.

There are many tests available to test the normality of continuous data. The most popular are the Shapiro-Wilk test and the Kolmogorov–Smirnov test. We decided to perform the Shapiro-Wilk test due to the fact that it is a more appropriate method for small sample sizes < 50. While the Kolmogorov–Smirnov test is usually used for n >= 50. When P > 0.05, we accept the null hypothesis and classify data as normally distributed.

The test was performed for five counties and the total number of trips:



Figure 33. Shapiro-Wilk test for the top 5 counties

After performing the Shapiro-Wilk test we can see that the most bigger P values have “wexford\_test” and “kerry\_test”. The P values are 0.94 and 0.86 respectively, which is more likely normally distributed. It was decided to take wexford-df with the biggest P value.

We obtained the mean and the standard deviation values in order to calculate z-score. Z-score is also known as a standard score. It is a measure of how many standard deviations a particular data point is from the mean of a dataset.

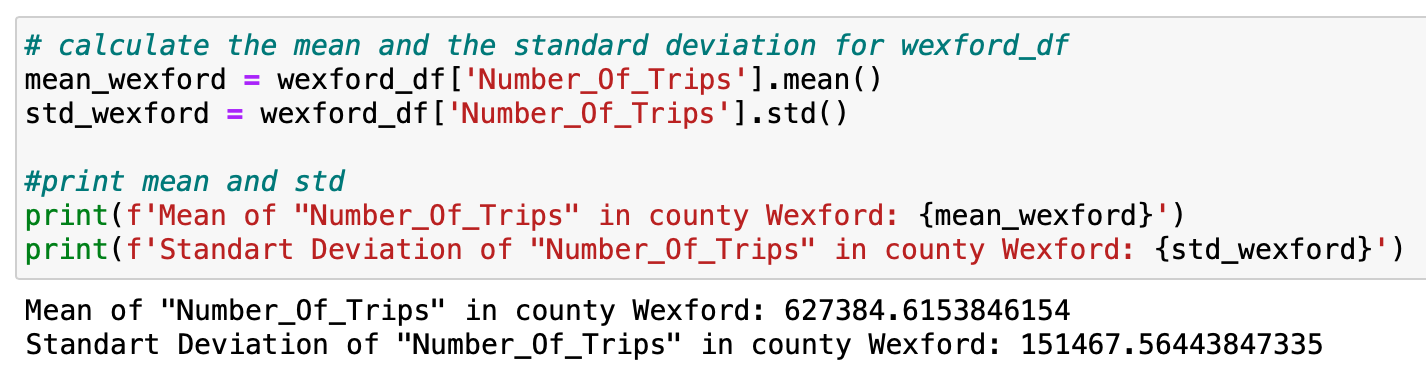


Figure 34. Mean and Standard Deviation for county Wexford

After that we calculated the z-score and what is the probability of Number of Trips higher than 700.000 in Wexford.

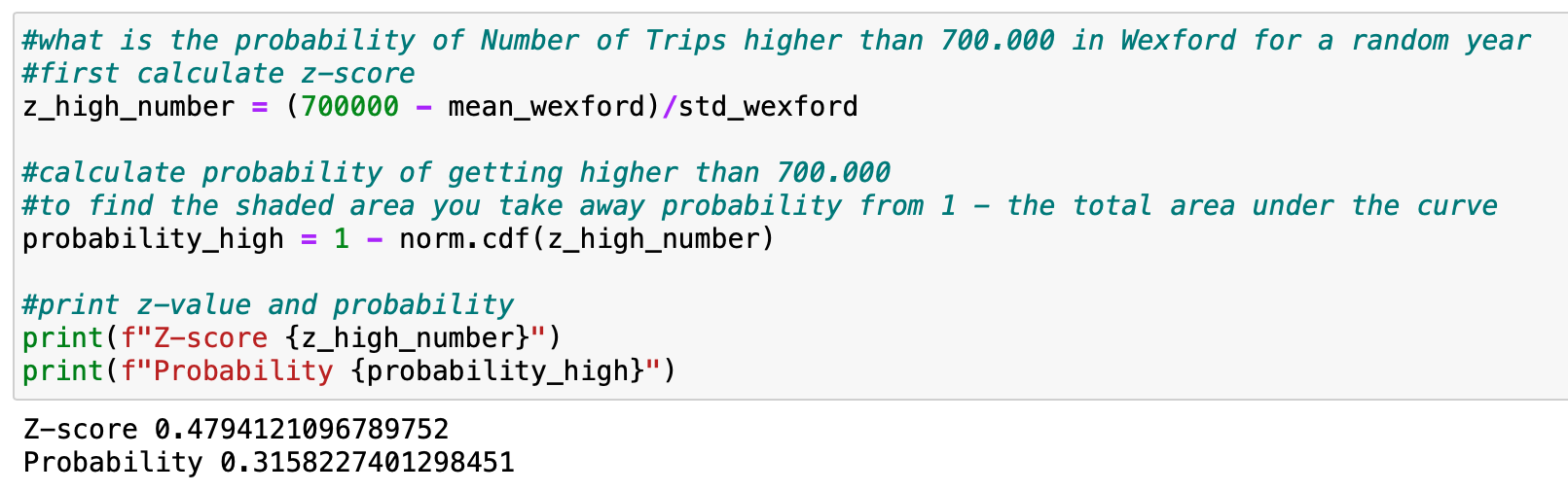


Figure 35. Z-score and Probability (Number of Trips >700.000) for county Wexford

The probability of observing the Number of Trips more than 700.000 in Wexford is approximately 0.3158.

The next step was to calculate the z-score and what is the probability of Number of Trips lower than 500.000 in Wexford.

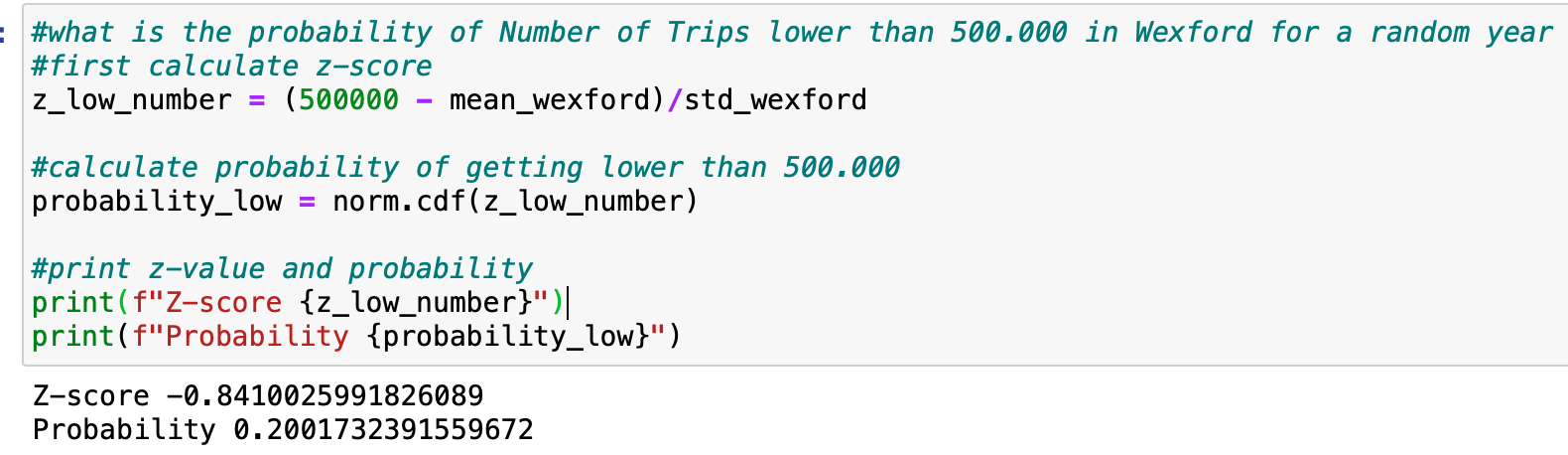


Figure 36. Z-score and Probability (Number of Trips < 500.000) for county Wexford

The probability of observing the Number of Trips less than 500.000 in Wexford is approximately 0.2001.

And we checked the probability of Number of Trips between 500.000 and 700.000 in Wexford.

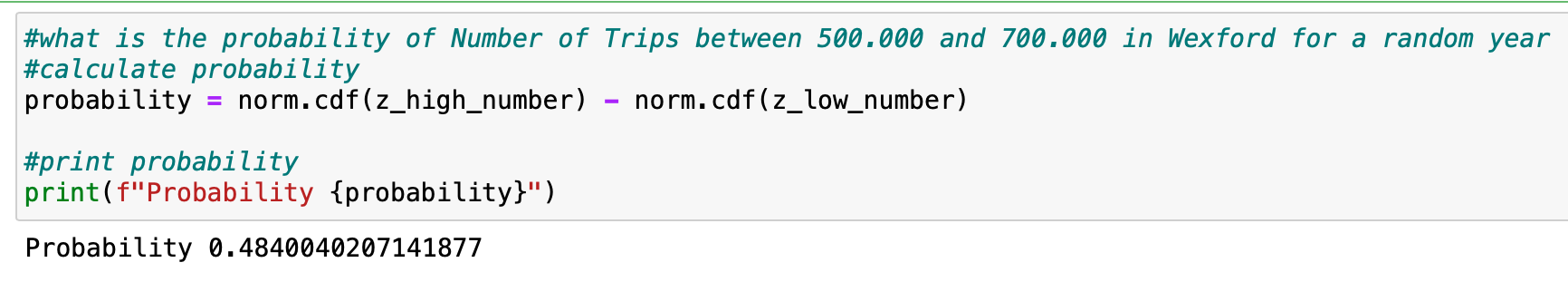


Figure 37. Probability (700.000 < Number of Trips < 500.000) for county Wexford

The probability of observing the Number of Trips between 500.000 and 700.000 in Wexford is approximately 0.48.

***Insights:***

Based on the result the number of trips tends to be more likely to exceed 700.000 than fall below 500.000. And there is a relatively high probability (0.48) of the number of trips falling between 500.000 and 700.000. We can conclude that the distribution of the number of trips in county Wexford is more likely to follow a normal distribution.

We also performed a normal distribution fitting to data in the “Number\_Of\_Trips” column of the “wexford\_df” DataFrame and plotted it using a bar chart.

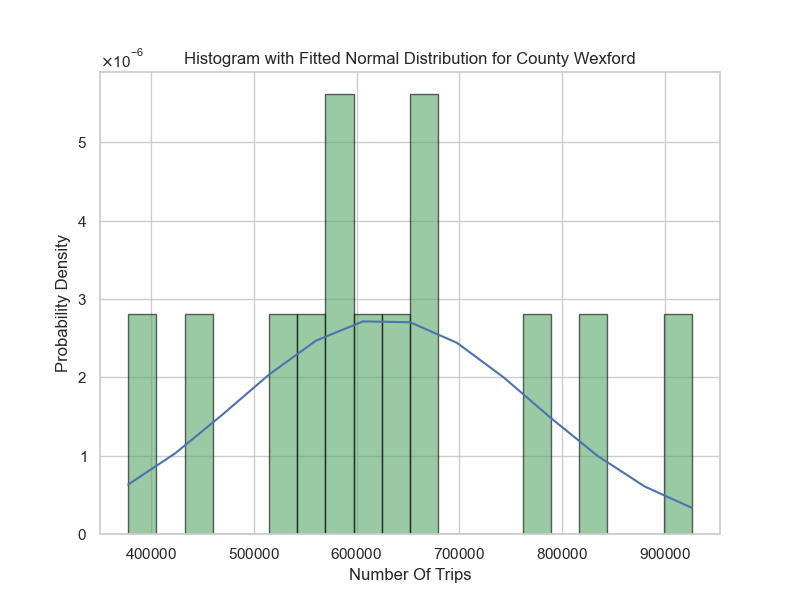


Figure 38. Fitted Normal Distribution for county Wexford

## Binomial Distribution

The Binomial Distribution is used when there are exactly two possible outcomes of a trial: success or failure. It is a discrete probability distribution. The Binomial Distribution characterized by two parameters:

n - the number of trials or experiments

p - the probability of success on each trial

We applied the binomial probability distribution on the “Number\_Of\_Trips” column of the “counties\_trips\_df” DataFrame that contains all trips for all counties in 2010-2022.

The question was “What is the probability of getting a number of trips > 200.000 if we pick a random number from "Number\_Of\_Trips" in all years”.

First we obtained how many numbers in our dataset that are bigger than 200.000

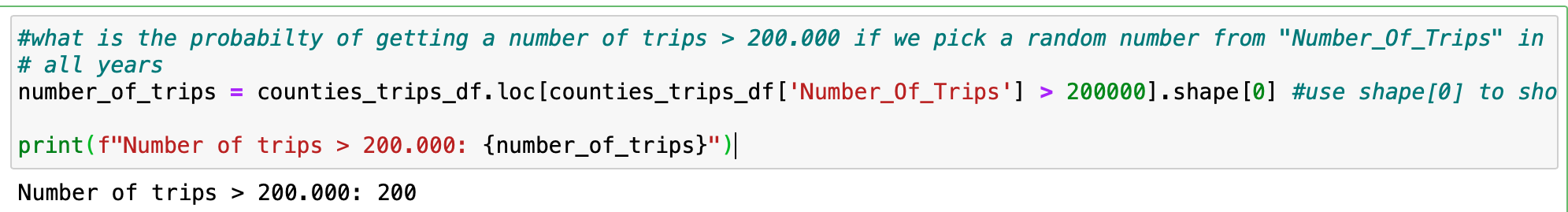


Figure 39. Number of trips >200.000 in “counties\_trips\_df” DataFrame

After we calculated the probability:

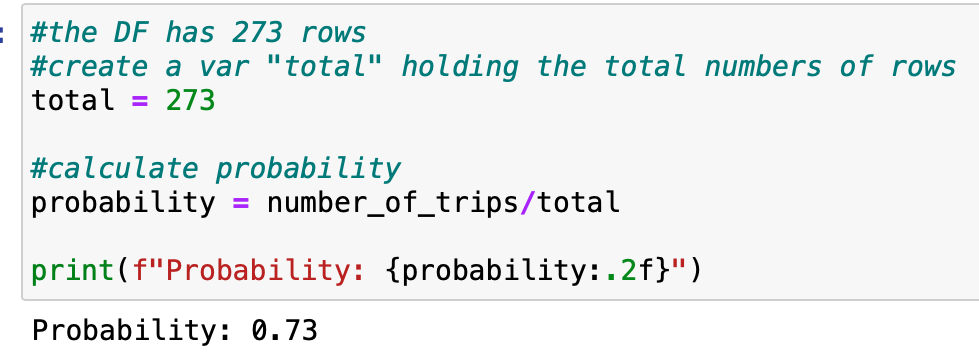


Figure 40. Probability that Number of trips >200.000 in “counties\_trips\_df” DataFrame

A Probability Mass Function was used because it gives the probability of observing each possible number of successes in a fixed number of Bernoulli trials.

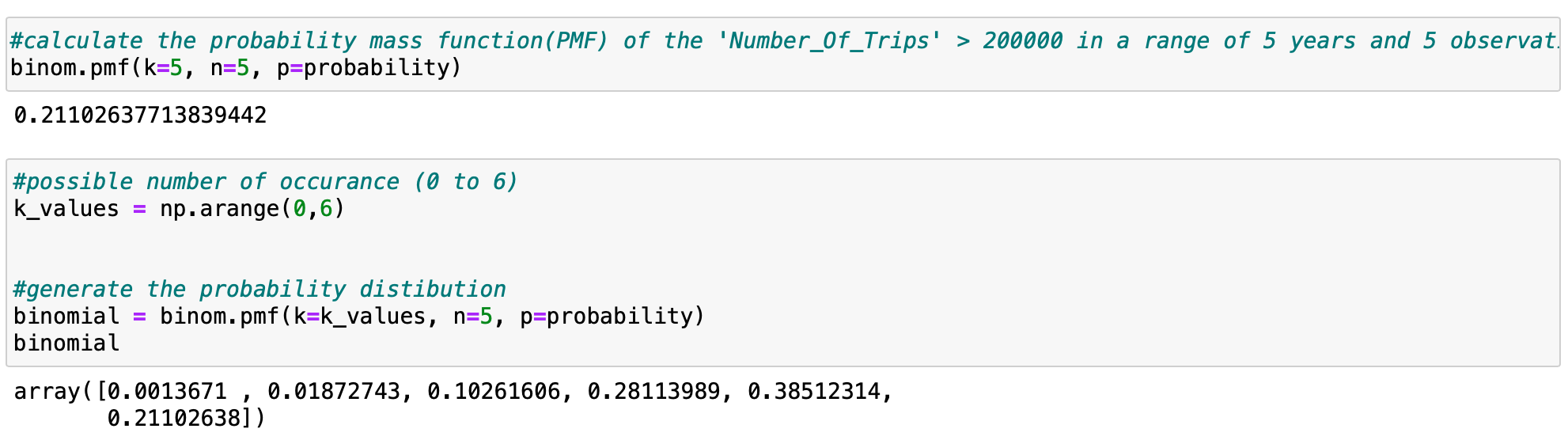


Figure 41. Probability Mass Function

We visualized the Binomial Distribution using a line graph and a bar plot

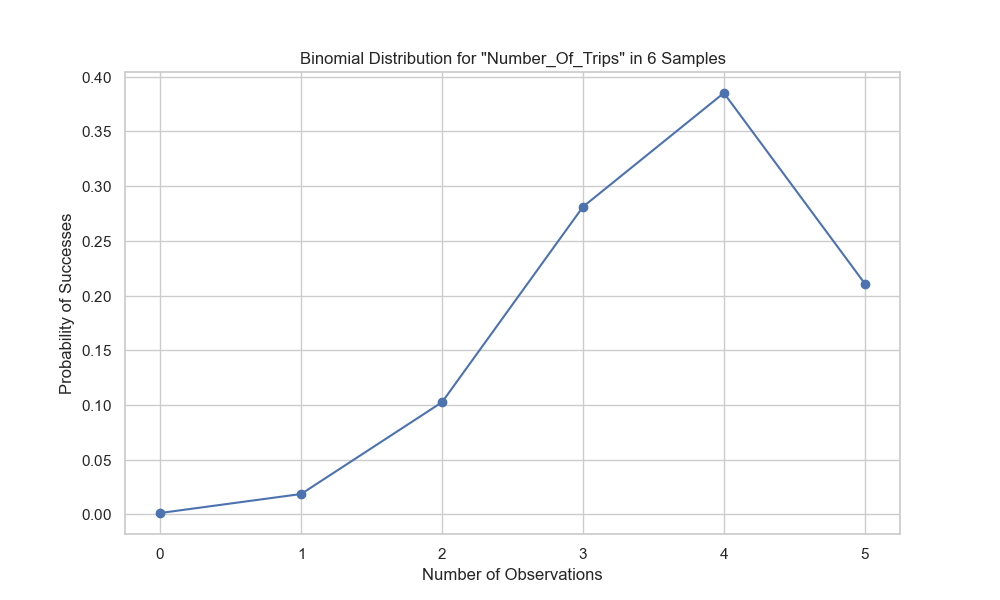


Figure 42. Line graph of Binomial Distribution

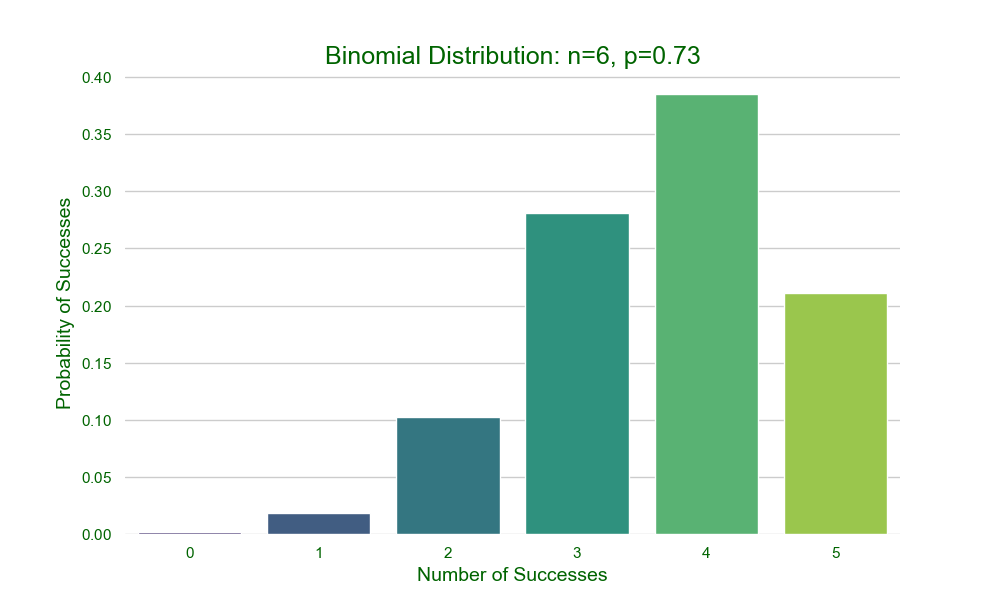


Figure 43. Bar plot of Binomial Distribution

## Poisson Distribution

The Poisson Distribution is a discrete probability distribution. It describes the number of events that occur in a fixed interval of time or space, given a known average rate of occurrence (assuming the events happened independently of each other). λ (lambda) is a parameter of the Poisson Distribution. It is the average rate of occurrence of the events within an interval.

We applied the Poisson Distribution to model the number of occurrences of every county in the fixed number of samples.

First we calculated the λ - average rate:

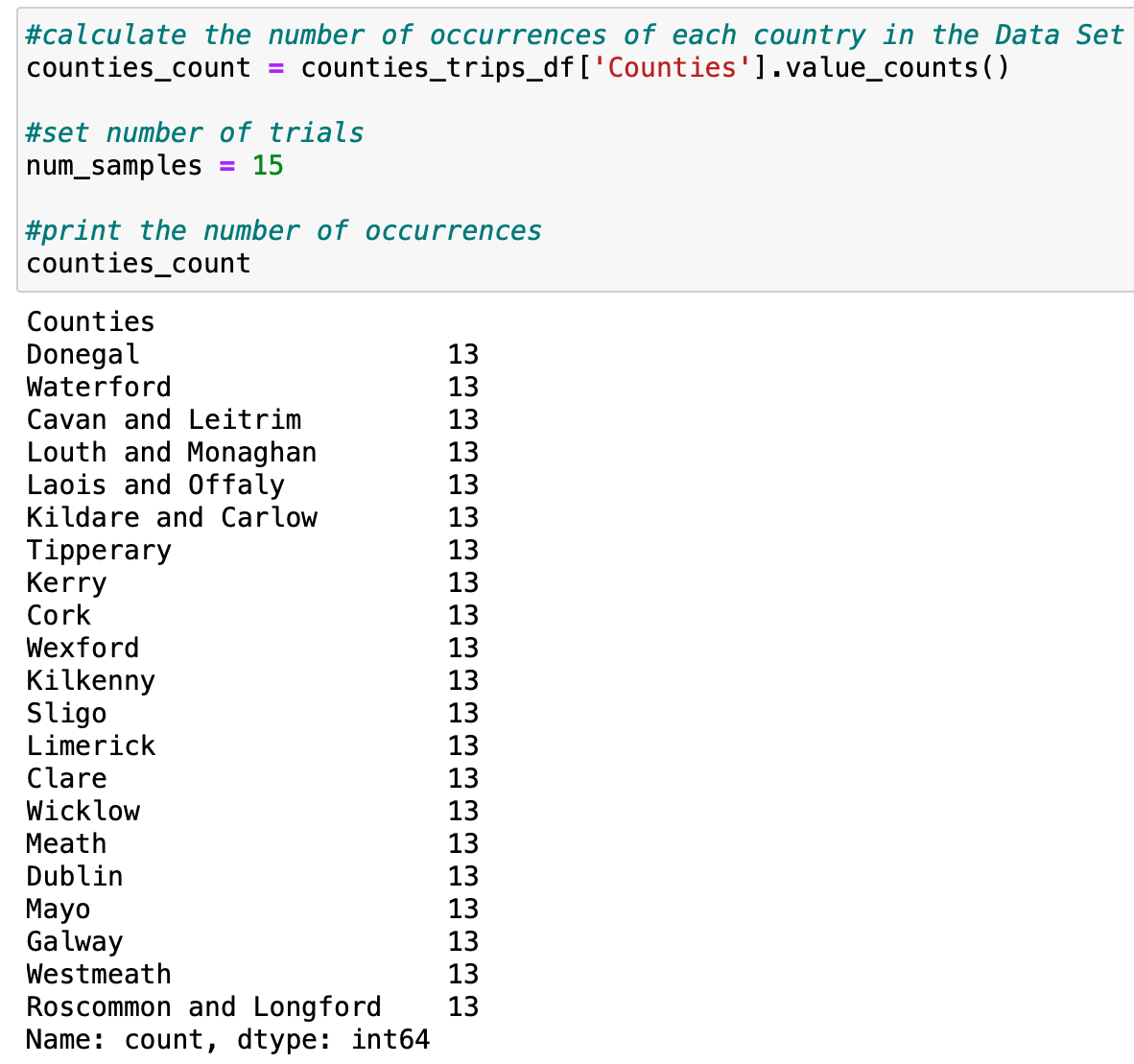


Figure 44. Average rate (λ) for Poisson Distribution

The Poisson Distribution was applied:

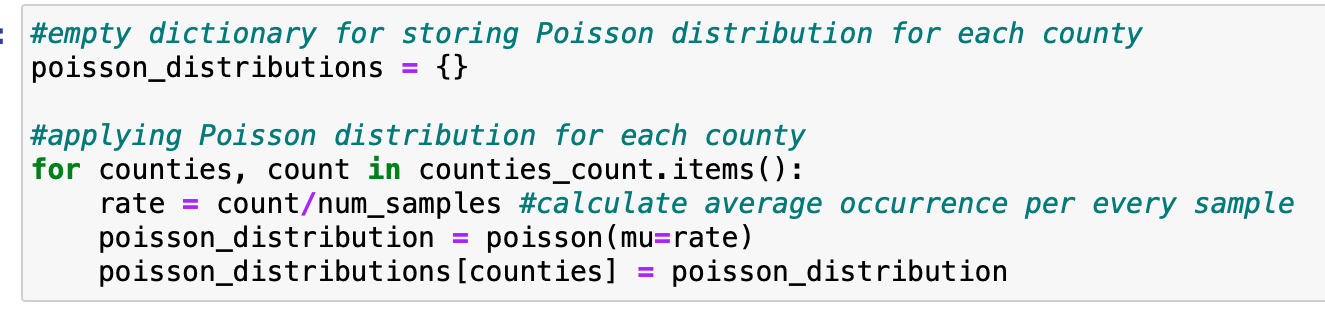


Figure 45. Applying Poisson Distribution

We plotted the PMF(Probability Mass Function) of the Poisson Distribution for each county to visualize the probability of observing different numbers of occurrence:

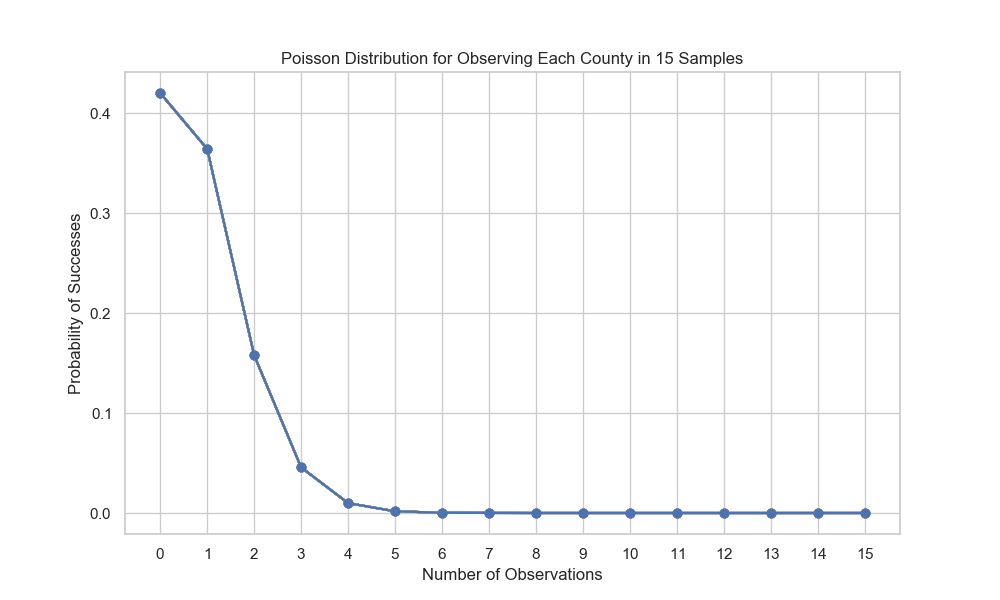


Figure 46. Poisson Distribution for Observing Each County in 15 samples

# Machine Learning for Data Analytics

## CRISP-DM Framework

It was decided to follow the CRISP-DM project management framework. There are six stages in a cycle. The main difference in the structure is that the transitions between stages can be reversed. It means that if specialists found the data not sufficient to resolve the goal of the project they can return to the data preparation stage and select different target variables, generate features, etc, without returning all the way to the start of the cycle. It’s been found that CRISP-DM useful for almost all the projects.[1]

## Linear Regression

Liner Regression is a supervised ML algorithm. We chose a supervised ML algorithm, because the algorithm learns from labeled data, where both the input features and corresponding output are known.

We applied Linear Regression for the purpose of forecasting local tourism by Irish Residents. The DataFrame with the total number of trips was chosen for forecasting tourism in Ireland overall.

Our first step was to remove the potential outliers which were caused by COVID-19. We applied .drop() to remove values for the years 2021 and 2022 years from the DataFrame respectively. After removing the outliers we imported “LinearRegression” from Sklearn and fit the model. The results we plot using a line graph.

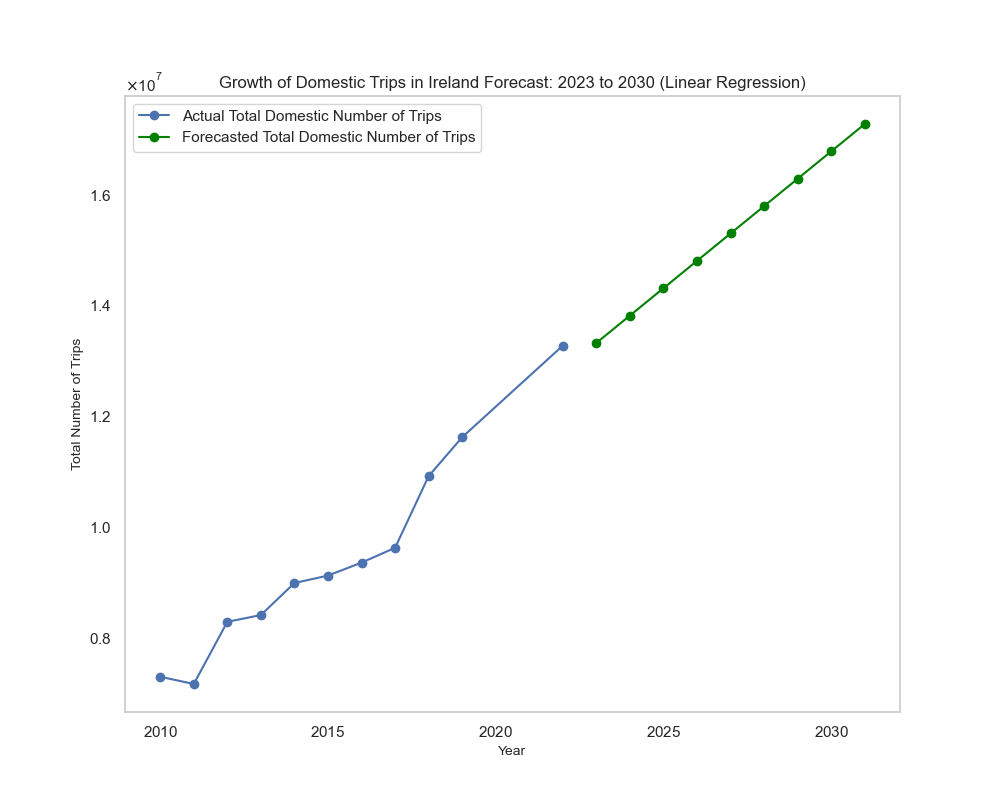


Figure 47. Linear Regression - Forecasting Local Tourism in Ireland

We also split the DataFrame into testing data (40%) and training data (60%) and fit the LinearRegression. After fitting the model we used a graphic residual plot to assess the goodness of fit of the Linear Regression.

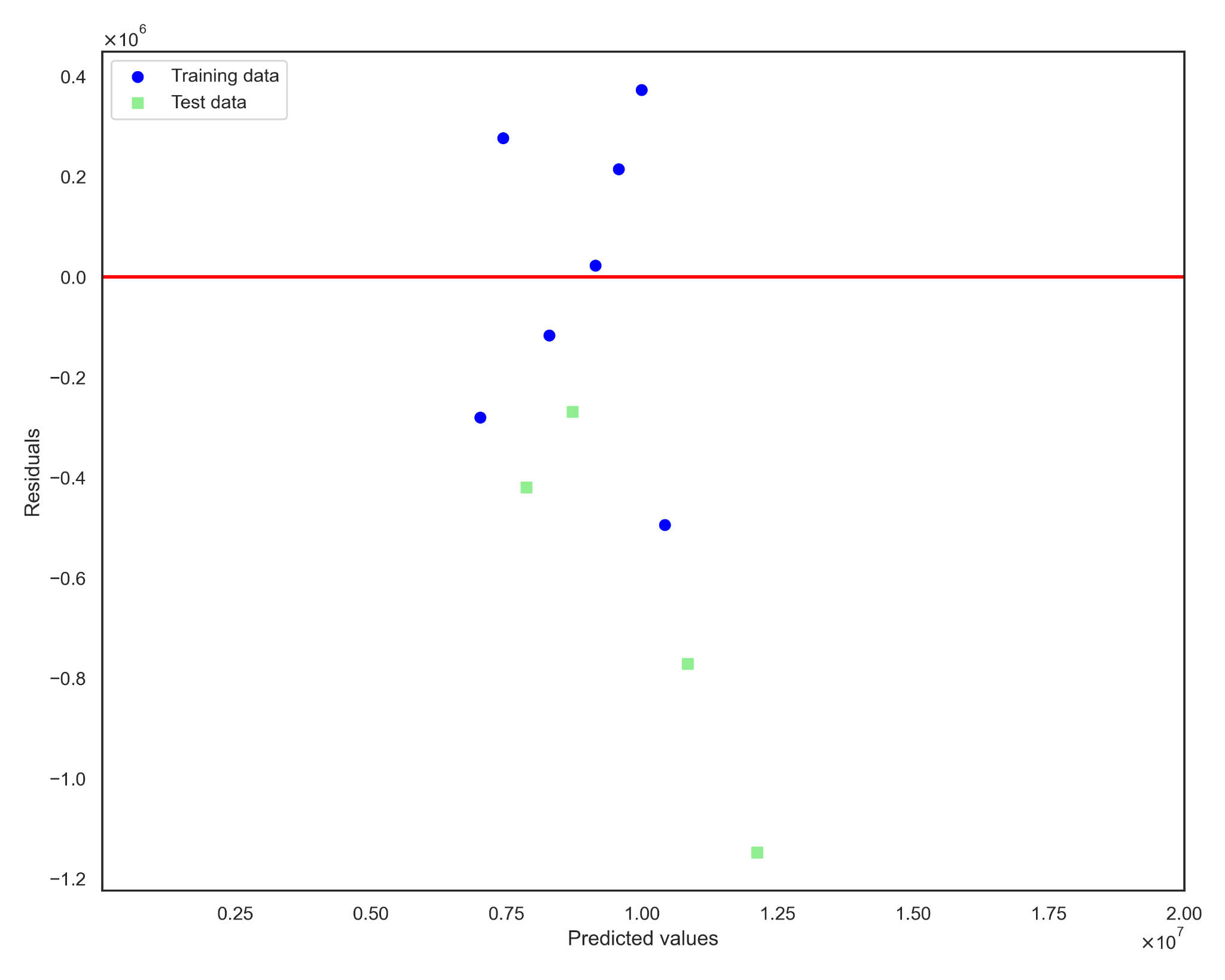


Figure 48. Residual plot of Linear Regression

We calculated two parameters MSE (Mean Squared Error ) and R^2 (R-squared score) of our train and test data to see how well the model performed the prediction. R^2 is a measure of how well the model explains the proportion of the variance in the target variable. MSE is a measure of closeness predicted values to the actual values.

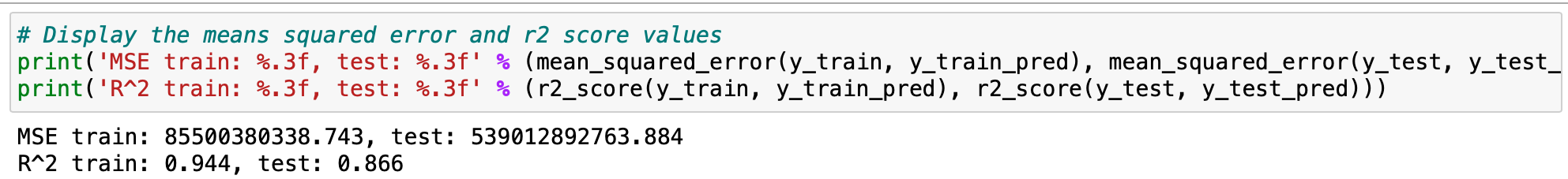


Figure 49. MSE and R^2 of Linear Regression

## Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) Regression is a regression analysis method that performs both variance selection and regularization in order to improve prediction accuracy and interpretability of the model.

We trained the Lasso model by calling the .fit() method, plotted the forecast and calculated R^2 score for the training and testing data.

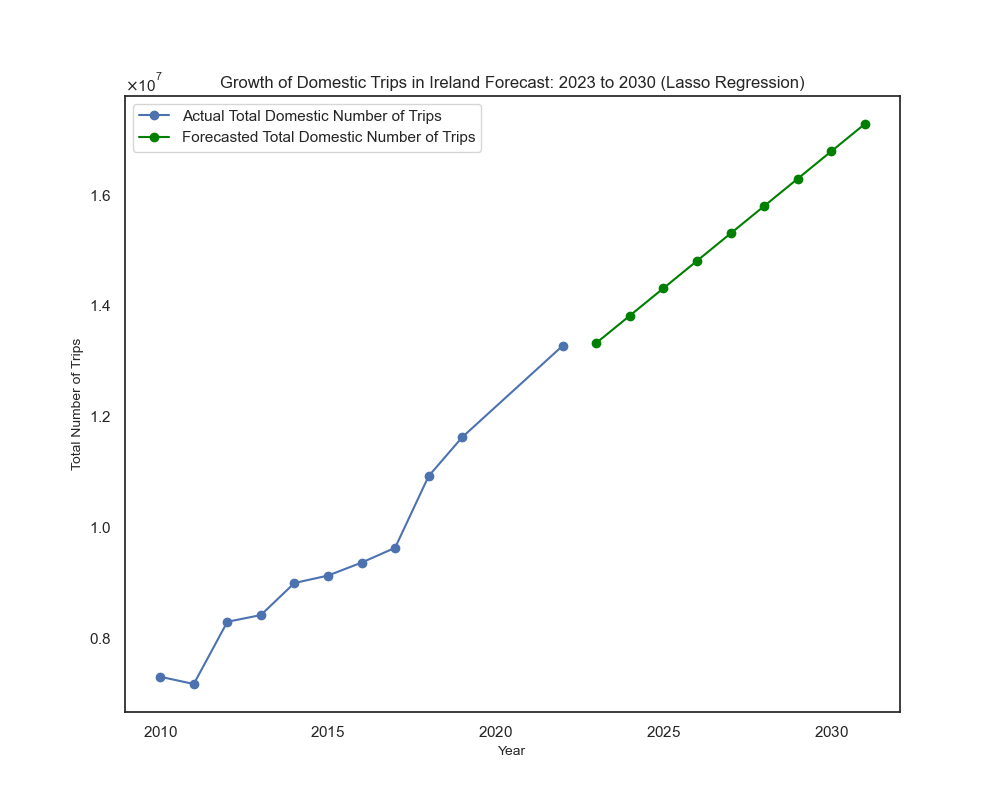


Figure 50. Lasso Regression - Forecasting Local Tourism in Ireland

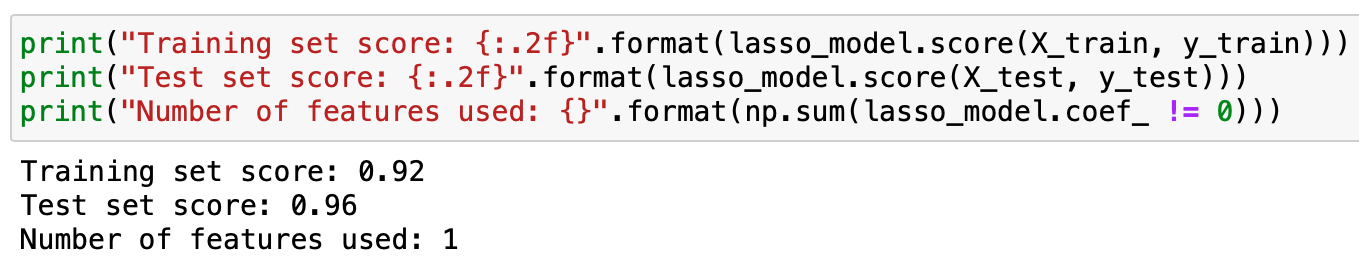


Figure 51. R^2 training and testing scores of Lasso Regression

## Ridge Regression

Ridge Regression also known as Tikhonov or L2 regularization. Ridge regression specifically corrects for multicollinearity and overfitting in regression analysis.

We trained the Ridge model by calling the .fit() method, plotted the forecast and calculated R^2 score for the training and testing data.

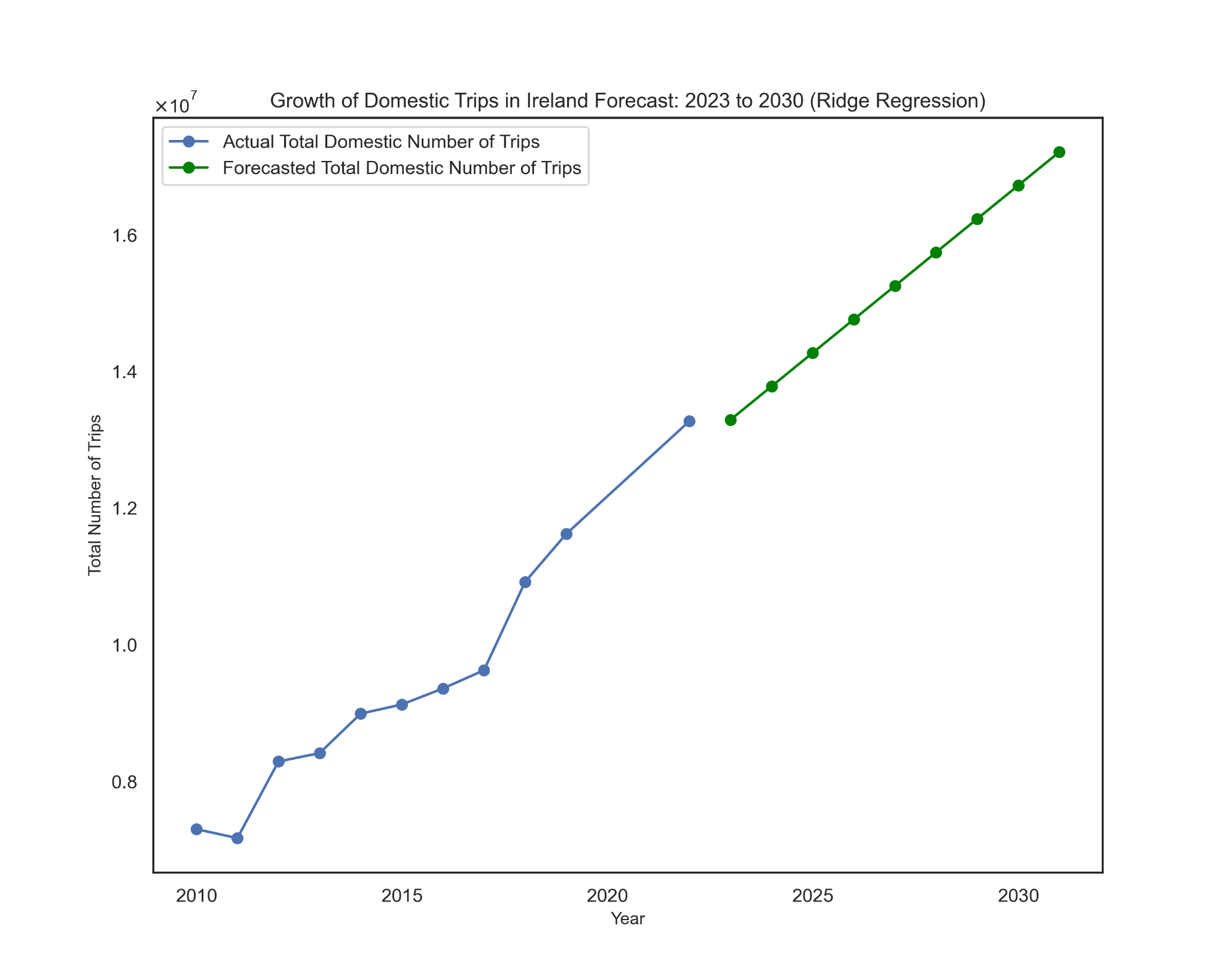


Figure 52. Ridge Regression - Forecasting Local Tourism in Ireland

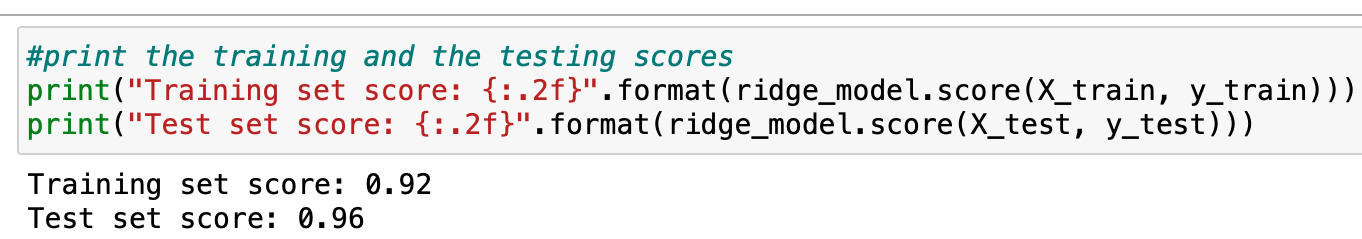


Figure 51. R^2 training and testing scores of Ridge Regression

We compared the results of 3 models: Linear, Lasso and Ridge Regressions.

Table 1. Comparison of Regressions

| **R^2** | **Linear Regression** | **Lasso Regression** | **Ridge Regression** |
| --- | --- | --- | --- |
| Train | 0.94 | 0.92 | 0.92 |
| Test | 0.86 | 0.96 | 0.96 |

In summary:

1. Lasso and Ridge Regression models performed similarly in terms of R^2 on both training and testing data.
2. Linear Regression shows signs of overfitting, with a higher R^2 on the training data compared to the testing data.

# Programming

In the development of our research project, we made use of various programming paradigms to solve problem efficiency and enhance our design decisions.

## Imperative Programming

Imperative programming sees a program as a structured sequence of statements. Imperative programming paradigms in our project: print(), .head(), read\_csv(), .drop(), rename(), etc.

Example of Imperative programming (step-by-step performance):

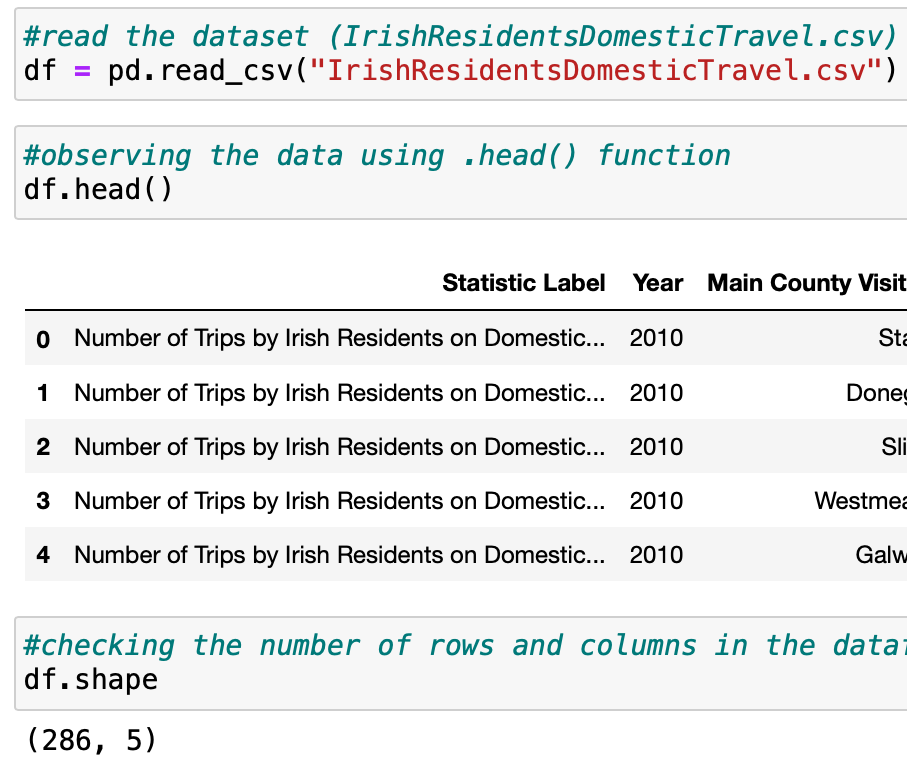


Figure 52. Example of Imperative programming

## Declarative Programming

Declarative Programming allows us to focus on specifying what we want to achieve rather than instruction on how to achieve it. Leads to more maintainable code.

Example of Declarative programming in the project - the dictionary comprehension {col: top\_5\_counties\_df[col].std() for col in counties} focuses on what need to be computed:



Figure 53. Example of Declarative programming

## Functional Programming (FP)

Functional Programming principles such as pure-function and higher-order function helps to write cleaner and more efficient code.

In the project we created a function called “*convert\_to\_thousands(df, column\_name)*” which o convert numbers in column into thousands

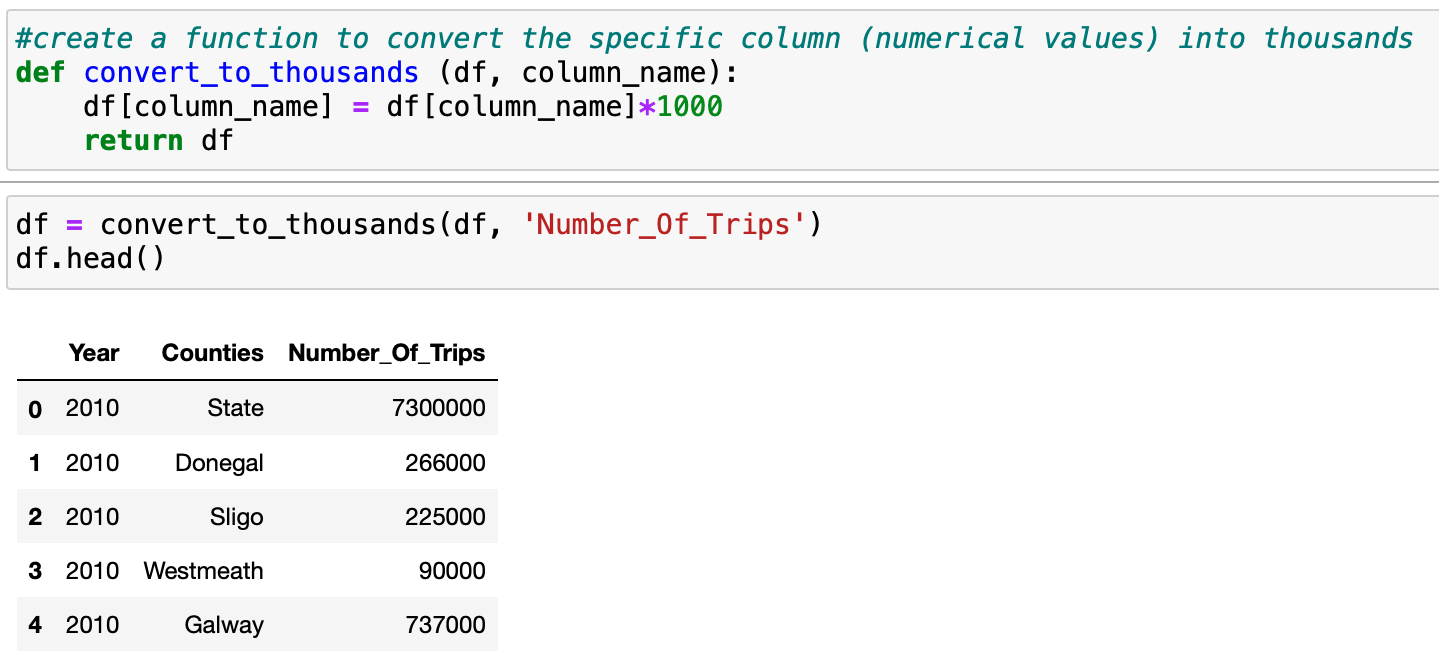


Figure 54. Example of Functional programming

## Procedural Programming

Procedural Programming focuses on procedures or routines that are executed to achieve a specific task.

For example .merge() function from Pandas library - is a procedural approach which merges two DataFrames based on a specific key and a specific method.

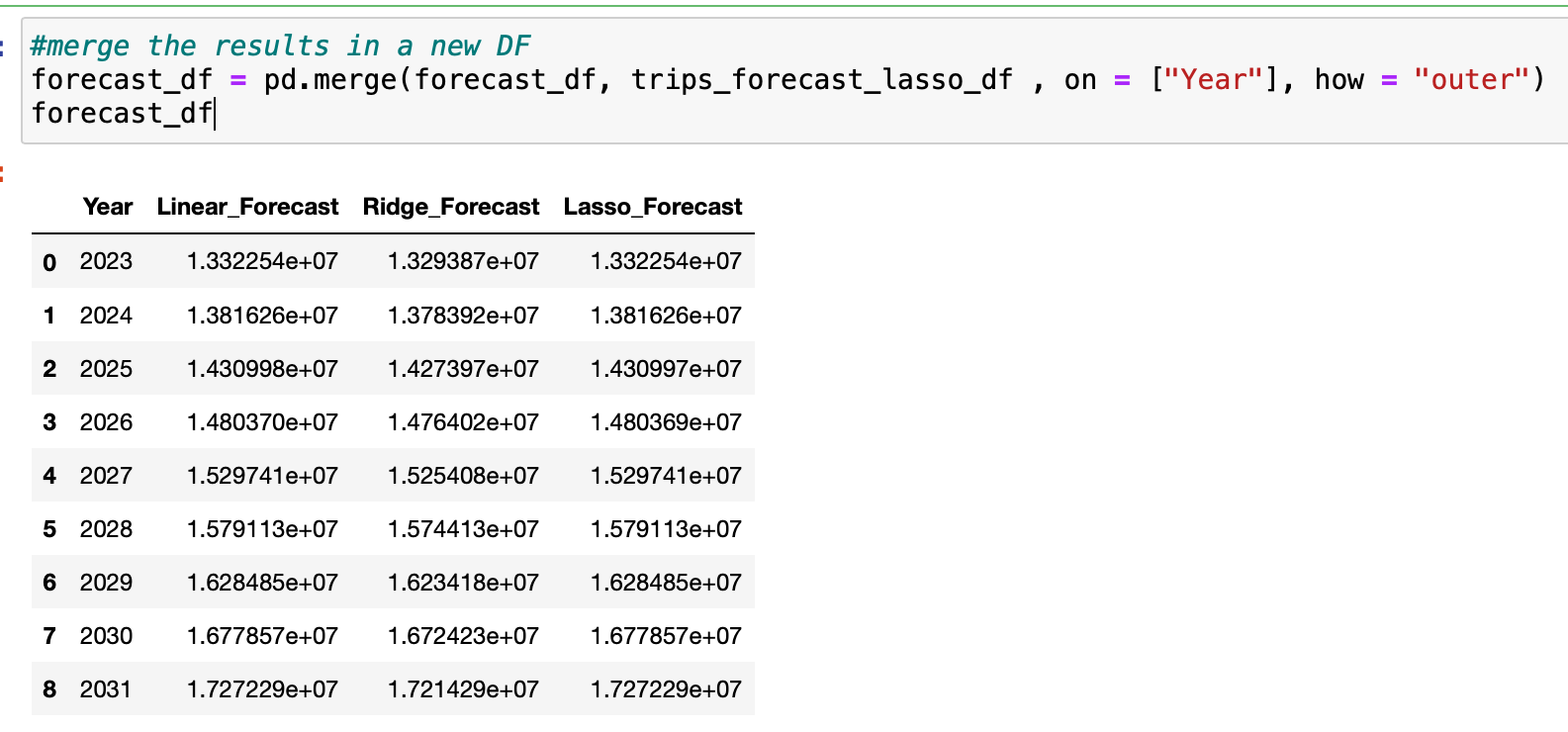


Figure 55. Example of Procedural Programming

## OOP (Object Oriented Programming)

OOP is widely used in software development. It promotes reusable, modular and maintainable code by organizing it into objects with well-defined interfaces.

We reused our function called “get\_unique\_values” to print the unique names of columns.

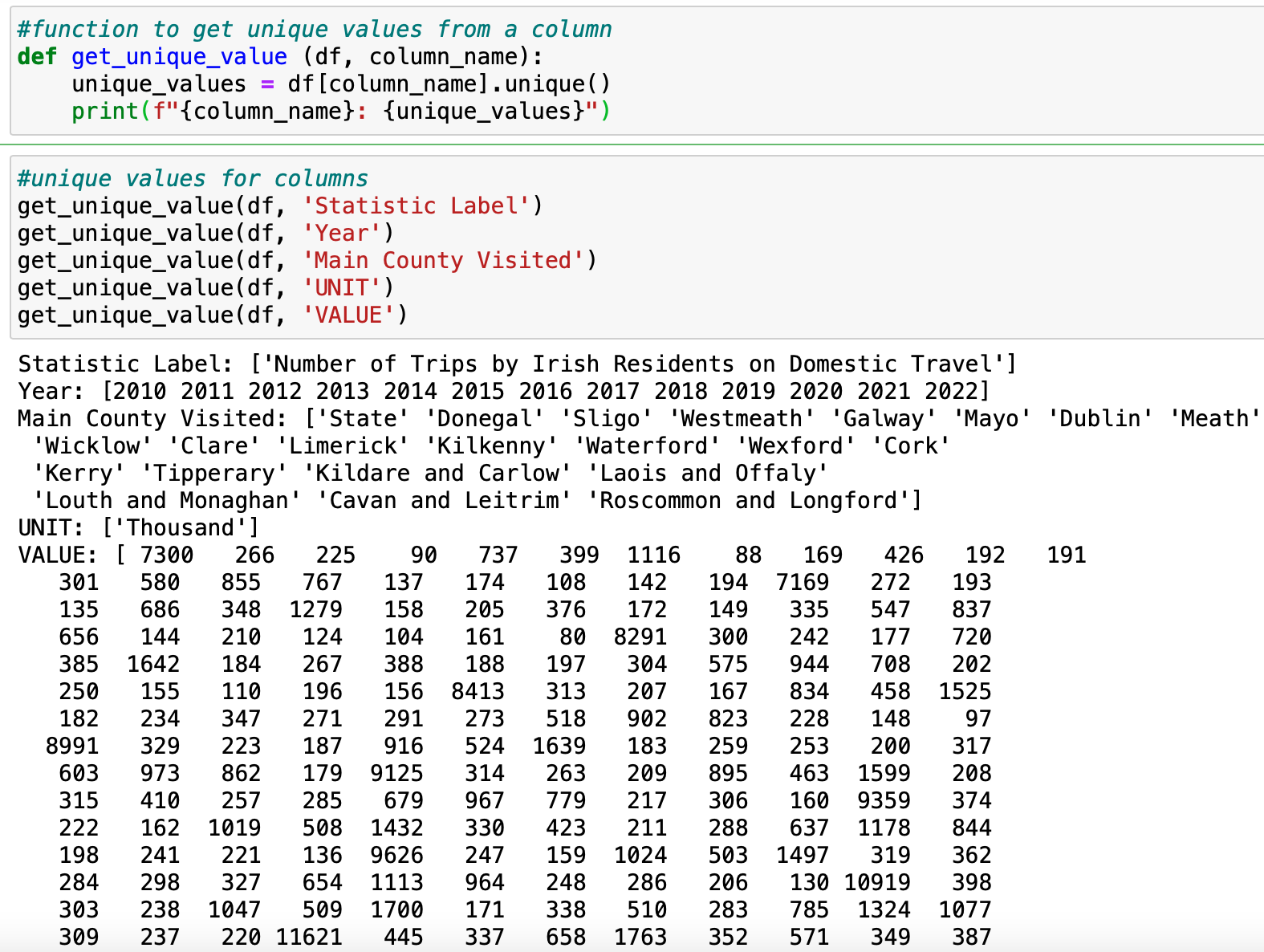


Figure 56. Example 1 of Object Oriented Programming

Also we used OOP concepts in the context of Scikit-learn regression implementation such as (LinerRegression(), Ridge() and Lasso()).

For example we imported the Ridge class from Scikit-learn’s linear\_model module. Initialized an object ridge\_model = Ridge(). ridge\_model is an instance of the class Ridge(). We called the .fit() method and the predict() method.

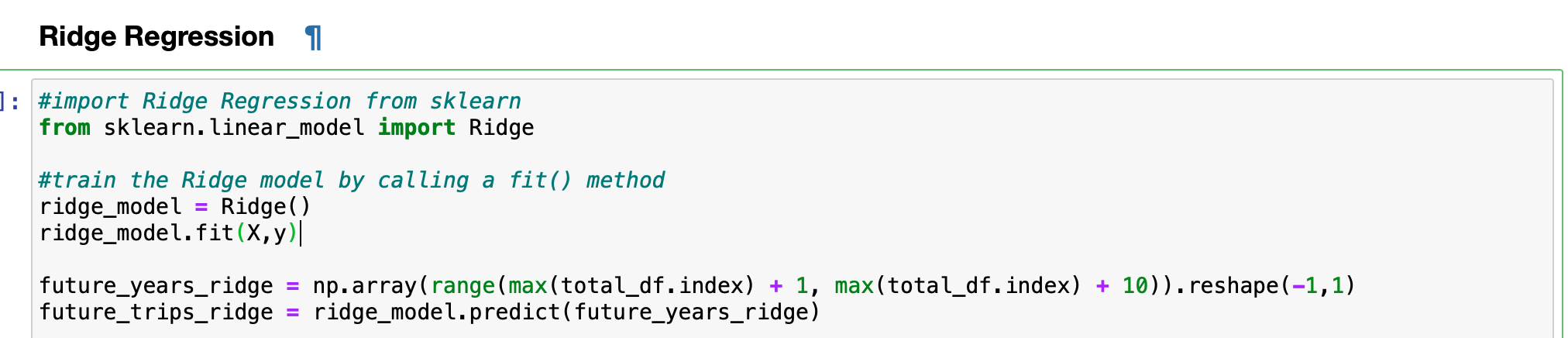


Figure 57. Example 2 of Object Oriented Programming

## Conclusion

In conclusion, this project provided valuable insights into Domestic Tourism in Ireland by the residents from 2010 to 2022, focusing on two primary objectives: understanding changes in tourism over time and identifying the most popular destinations. The research utilized forecasting techniques to predict tourism trends in Ireland up to the year 2030. While the correlation with fuel cost is not achieved, the project still offers significant contribution.

## References

1. Kakkar, Neil. “68-95-99 Rule – Normal Distribution Explained in Plain English.” *freeCodeCamp*, 7 August 2020, Available from:

*https://www.freecodecamp.org/news/normal-distribution-explained/*. [Viewed 30 March 2024.]

1. Li, Timothy. “Kurtosis Definition, Types, and Importance.” *Investopedia*, Available from: https://www.investopedia.com/terms/k/kurtosis.asp. [Viewed 25 March 2024.]
2. “SKEWNESS AND KURTOSIS.” *ignou*, Available from:

*https://www.igntu.ac.in/eContent/IGNTU-eContent-467281593500-B.Com-4-Prof.ShailendraSinghBhadouriaDean&-BUSINESSSTATISTICS-All.pdf*. [Viewed 4 April 2024.]

1. Safarli, Javad. “Binomial Distribution: Probability Mass Function with Python.” *Medium*, 7 January 2024, Available from:

*https://medium.com/@c\_safarli/binomial-distribution-probability-mass-function-cumulative-distribution-function-calculation-5ae74fd3a340*. [Viewed 20 March 2024.]

1. Jain, Sandeep. “Python - Normal Distribution in Statistics.” *GeeksforGeeks*, 11 March 2024, Available from:

[*https://www.geeksforgeeks.org/python-normal-distribution-in-statistics/*](https://www.geeksforgeeks.org/python-normal-distribution-in-statistics/). [Viewed 23 March 2024.]

1. “Descriptive Statistics and Normality Tests for Statistical Data.” *NCBI*, Available from:

*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6350423/*. [Viewed 23 March 2024.]