**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Problem Solving for Industry |
| **Assessment Title:** | Problem Solving for Industry |
| **Lecturer Name:** | Dr Muhammad Iqbal (miqbal@cct.ie) |
| **Group Members:** | Syed Ali Jarrar Raza Rizvi **2018458**  Muhammad Irfan **2018426**  Waseem Ullah **2019465**  Muhammad Nazim **2018370**  Nouman Rafique **2018403** |
| **Due Date:** | Sunday, 7 August 2022, 11:59 PM |
| **Date of Submission:** | Sunday, 7 August 2022 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of Contents

[List of Figures 3](#_Toc110773496)

[List of Tables 4](#_Toc110773497)

[Introduction 5](#_Toc110773498)

[Group Members' Contributions and Work Distribution 6](#_Toc110773499)

[Business Understanding 7](#_Toc110773500)

[Data Understanding 9](#_Toc110773501)

[Motor Industry Dataset 9](#_Toc110773502)

[Tourism Industry Dataset 15](#_Toc110773503)

[Data Preparation 21](#_Toc110773504)

[Modeling 23](#_Toc110773505)

[Evaluation 23](#_Toc110773506)

[Deployment 24](#_Toc110773507)

## List of Figures

[Figure 1 Vehicles dataset by SIMI 6](#_Toc109727967)

[Figure 2 Dataset has no null entries 7](#_Toc109727968)

[Figure 3 Dataset has no duplicates 7](#_Toc109727969)

[Figure 4 Unique cars in the dataset 7](#_Toc109727970)

[Figure 5 Histogram of the dataset 8](#_Toc109727971)

[Figure 6 Boxplot of the cars dataset from 2012-2019 8](#_Toc109727972)

[Figure 7 Scatter plot of car sales from 2012 to 2019 9](#_Toc109727973)

[Figure 8 Individual sales of different car manufacturer from 2012 to 2019 11](#_Toc109727974)

[Figure 9 Yearly aggregated sales of all the cars from 2012 to 2019 11](#_Toc109727975)

[Figure 10 Tourism dataset 12](#_Toc109727976)

[Figure 11 Data features and their data types 13](#_Toc109727977)

[Figure 12 Tourism dataset duplicates 13](#_Toc109727978)

[Figure 13 Tourism dataset histogram 13](#_Toc109727979)

[Figure 14 Tourism dataset category-based boxplot stats 15](#_Toc109727980)

[Figure 15 Tourism dataset stacked bar plot category wise tourism revenue 15](#_Toc109727981)

[Figure 16 Tourism dataset quarterly based tourism revenue 16](#_Toc109727982)

[Figure 17 Tourism dataset quarterly tourism category revenue 18](#_Toc109727983)

[Figure 18 Yearly tourism revenue 18](#_Toc109727984)

[Figure 19 Master dataset after merging cars and tourism dataset 18](#_Toc109727985)

[Figure 20 Histogram of master dataset 19](#_Toc109727986)

[Figure 21 Heatmap of the correlation matrix 19](#_Toc109727987)

## List of Tables

[Table 1 Datatypes and decription of variables 7](#_Toc109727993)

[Table 2 Tourism dataset datatypes and column description 12](#_Toc109727994)

[Table 3 Correlation matrix 19](#_Toc109727995)

# Introduction

The idea under consideration is more of an opportunity than a solution to a problem. The idea is to provide a digital platform that will be made up of multiple digital apps, like a web app and a mobile app. This will be the same app with multiple interfaces on the web as well as on the mobile. The need to implement these multiple interfaces is to better facilitate the user's ability to easily use the platform on the go. These digital platforms will be facilitating users divided into two main groups with different needs and requirements.

The first type of user is the ride-hailing service providers for long tours to different tourist spots across different geographical locations. The second type of user is a ride-hailing consumer that will be booking the service providers to go for visiting the tourist spots.

It is a type of extension to the already existing idea of ride-hailing services that is already in operation in different parts of the world in different forms. But the common thing among all of them is that they are all limited within the bounds of a single city or geographical location. The currently existing ride-hailing services are mostly limited to short travels both in time and distance as they are within the limits of a city. You hire a ride-hailing service provider from some platform for a very short span and time to mainly move from one part of the city to the other. The other thing about this currently implemented model is that the variety and range of vehicles that people can use to provide services are limited. Our platform will also compete with already existing players in the field and will lead in this new dimension as a first mover, taking full advantage and capturing most of the market that will be created in this new domain.

Our idea will target or will be dependent on two major industries. These industries are the automotive industry and the tourism industry. Both of these industries are on the rise and creating more and more opportunities for people to grab on to. We will be supporting this argument and our idea with different datasets. We will be showing different analyses, visualizations, and projections from these datasets to further strengthen our stance, our estimations, and our proposals. In this document, we will be doing a business analysis of this idea using the CRISP-DM methodology.

# Group Members' Contributions and Work Distribution

This being a group project is a result of the collective effort of all the members of the group, which is mentioned above at the start. The project was thoroughly discussed and analyzed by all the group members. The project was planned and the workload was evenly distributed among the group members. We worked with concrete coordination and shared proper feedback throughout our work on the project to ensure that we were all are on the same page and are equally contributing. The project kicks off with the search for the right data set that serves the purpose of the research. This is the most critical part of the project as whole project ahead depends on the data set. And if the data set is not right or is not tailored for the purpose the whole effort will go in vain.

So, we as a group started our project by searching the data set. Different ideas were considered and discussed for this project. Many of them were discarded because of their huge scope which exceeds both the scope of this project and the time required and available to get this project done. Every group member came up with their findings in this regard which were discussed and examined by all the fellow members. After going through different potential ideas, we finalized our current subject matter. As our project was a group project, the next task was to split the workload among the group members. The project was related to CRISP-DM, so it was based on different steps. So, we decided to divide the responsibility based on the steps among us. Most of the steps were research based about the topic under consideration.

The very first step was Business Understanding. Each step in this project is related to this Business Understanding. So, every member of the group should be crystal clear about the purpose of this whole activity. The group members started researching the subject matter on their own and write their own version of their understanding about the business and then compiled them all together.

The next step was Data Understanding followed by Data Preparation. Both these steps of the process are very closely related to each other and extensive. So, Ali, Nouman and Waseem were mainly involved and responsible for in these two steps. So, the data understanding task was conducted mainly by these members and as all these steps are interrelated, so the findings of these steps were explained and discussed with other fellow members as well. These tasks were implemented by mutual discussion and research of these three members mainly.

The next step in the process which is Modeling followed by Evaluation and Deployment. These tasks are mainly implemented by two other members Nazim and Irfan. The modeling step was conducted considering the output from the work done in initial steps. So, they already knew a lot from their continuous involvement in discussions of the previous steps. This made their work to be in coordination with the previous work.

After all these steps we ended up with multiple outcomes in a scattered form. The scattered deliverables were needed to be unified into a single report as required in CA brief as well. So again, everybody chipped in wrote about their part, their contributions and their findings related to the project in a predefined structured and organized manner. This gave us a report in raw form that needs to be revied to get the final report. This report was iteratively reviewed by all the group members and then became the final report.

# Business Understanding

In reference to what was explained earlier in the introduction section, the business model that we are proposing here and will try to implement is to extend and capture an already existing and very well-implemented idea by different companies and startups. Different ride-hailing services are booming across the globe with tremendous success and huge gains in existing market capture and inclusion of new untouched sectors into the market. These services have raised the market cap to new heights. Our goal is to capture a new potential market that is untouched until now and we’ll try to explore this new market through our new venture under considerations. It is same old market but new in the way that we are providing services on a different new level.

Our idea is an extension to the existing ones as explained earlier and also aims at the same targets and that too on new highs. Our idea is to target the same existing user base with lucrative schemes and more attractive and fruitful services for both types of the users of our services. In addition to this, we also aim to capture the new user base from a different and broad perspective and thus increasing the market base. We will be more successful in this new dimension by making full advantage of being the first movers in this unexplored field.

We will be providing the existing and new customers a new alternative to the existing ride-hailing services bringing in more competition and quality to the services. The new and main addition that our platform will provide is taking this same model to a new high by providing the same services to the tourism industry. We will take the same services currently being provided by existing platforms to the level of tourism. The service providers can plan and publish their availability of services to the users. The service providers can come on our platform and search for these published service providers. They can plan their journeys by filtering based on different filters, getting the best deals for them at the best time that suits them the most.

The first mandatory deliverable that must be implemented to run the basic end-to-end business will include a mobile app and a web app. Both these apps will allow the following:

* Register the User
* Make an availability plan for the service
* Publish the above plan
* Search different plans
* Filter among different plans depending on different criteria
* Hire service provider after filtering and selecting one

The initial release with all necessary modules and functionalities can be released in a time span of 6 months after going through the complete research, design, and development lifecycle. The team required to achieve this milestone in the given time will be between 12-20 members. These members will be from different departments involved in the process like designers, architects, developers, and researchers.

The IT world is evolving continuously at a very rapid pace. To keep up with this pace is very critical for success. A good idea is not the only ingredient for the recipe to succeed. But a good well executed and implemented at the right time is definitely a key to success. So, missing the right time and deadline will have an impact on the proposed and estimated costs of the project and will affect the profits ultimately. Our idea with poor execution can even benefit others. A good idea but not on time to the market can make us loose the advantage of first mover. An on-time idea to the market but with loopholes can give other the advantage to rectify those and take the lead on us.

As explained earlier how crucial time management is and meeting the deadlines to the project's success. So, we must be very cautious and have a detailed thought before selecting the tools and technologies for the project. Keeping all of this in mind we would like to go with the MERN stack. Using this stack will keep us in both frontend and backend all in JavaScript and we can go for React Native to develop the mobile app which is a lot like that of ReactJs. This will bring consistency to all of our stack. This move will enable us to rotate our resources and make better use of resources depending on the need of the hour. The project will kick off with the research phase which will include different feasibility plans and an analysis of other key components that may affect our project. This will be followed by the design phase where the architecture of the product will be defined. This phase will eventually lead to the actual development of the product. The development phase is both accompanied and followed by the testing of the product which will ensure that a quality product meeting all the best quality standards is launched in the market.

# Data Understanding

As explained earlier the two industries involved are Auto Industry and Tourism Industry in Ireland. So, our focus for the research were these two industries and that too in Ireland. The dataset acquired for this cross industrials data mining technique are taken from various sources such as

* The Society of Irish Motor Industry (SIMI)
* Ireland Open Data Portal

There were many other sources and data sets that were examined and considered for this project initially

but the two sources mentioned above are the major sources of data for this project. No direct data sets in this regard were available so we based our research on multiple data sets and then worked on them in the subsequent steps of Data Preparation and Data Cleaning. So, some data was from one source and the rest of the data form some other source and got our resultant data set.

## Motor Industry Dataset

The motor industry datasets were gathered through the extensive datasets provided by The Society of Irish Motor Industry [reference] in its National Vehicle Statistics from year 2012 to 2019. The dataset initially contained statistics related to different category of vehicles which also included the private cars category which we have considered for our analysis. The dataset contains the total number of cars sold by each manufacturer in a year. It contains 296 rows and 3 columns which give us the make of the car along with the number of units manufactured in that year and the year it was made in.

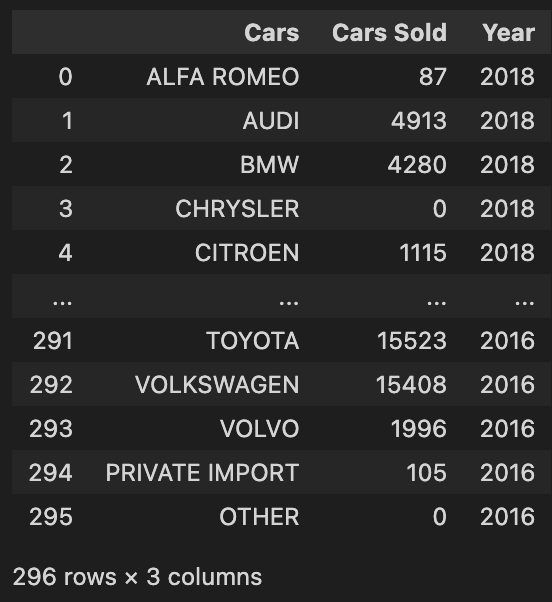


Figure 1 Vehicles dataset by SIMI

The dataset variables along with their data types are also given in the table below.

Table 1 Datatypes and description of variables

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Column Type** | **Column Description** |
| Cars | Object | The manufacturer’s name |
| Cars Sold | Int64 | Number of cars sold |
| Year | Int64 | The year in which the cars were sold |

Some statistics related to this dataset are also given below.

* Dataset has zero null entries.

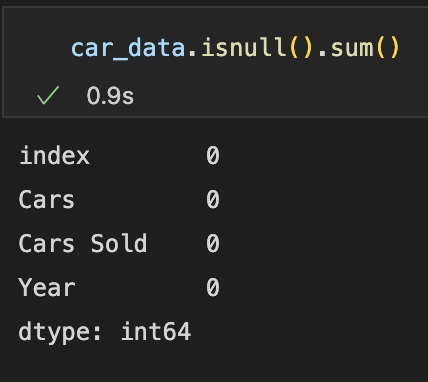


Figure 2 Dataset has no null entries

* Dataset has zero duplicate entries.

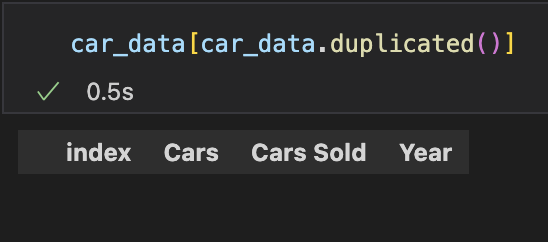


Figure 3 Dataset has no duplicates

* There are 47 unique cars in the dataset.

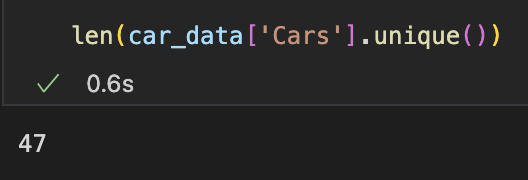


Figure 4 Unique cars in the dataset

The histogram of the dataset is also given below. Here we can clearly see that the histogram of the cars sold is heavily skewed towards the left. This means that most of the car units that are sold over the years are in small quantity and a few are of very large quantity.

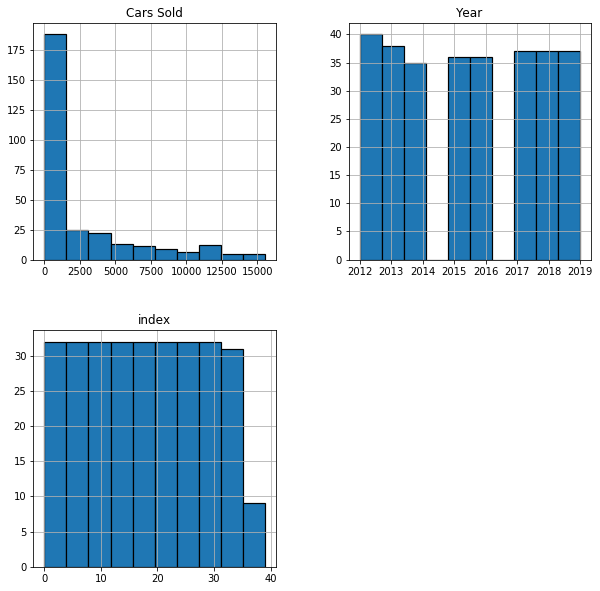


Figure 5 Histogram of the dataset

We can also see the individual cars sales in the boxplot which is given below. As there are 47 cars with varied sales over the years. Here we can see that “Ford”, “Hyundai”, “Toyota” and “Volkswagen” have highest sales distribution over these years. Whereas “Nissan” also lies in that spectrum. We can also see the cars sales distribution of other cars as well. In the scatter plot below, we can also see the year wise cars sales stats clearly.

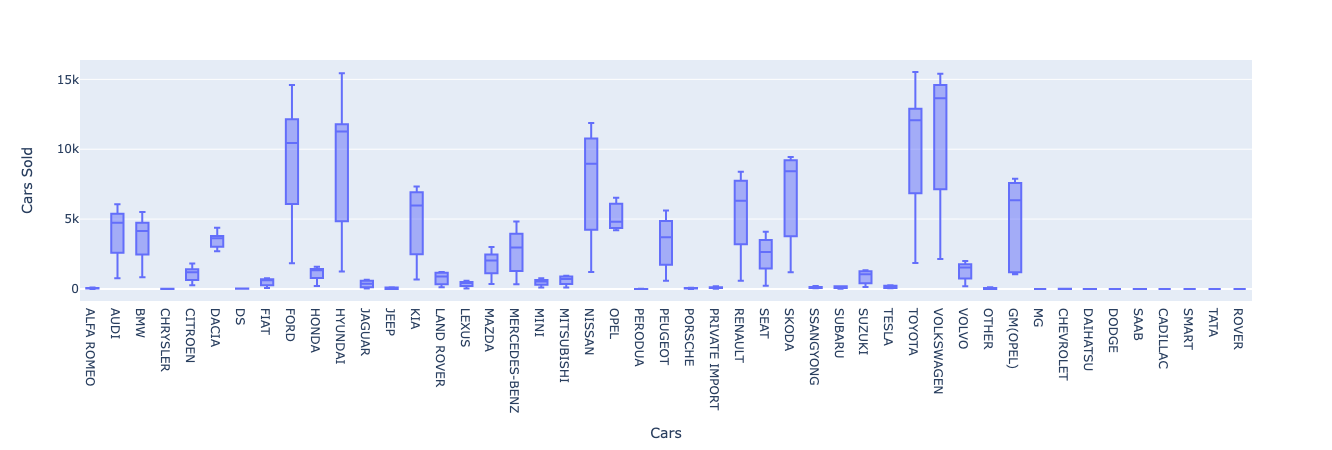


Figure 6 Boxplot of the cars dataset from 2012-2019

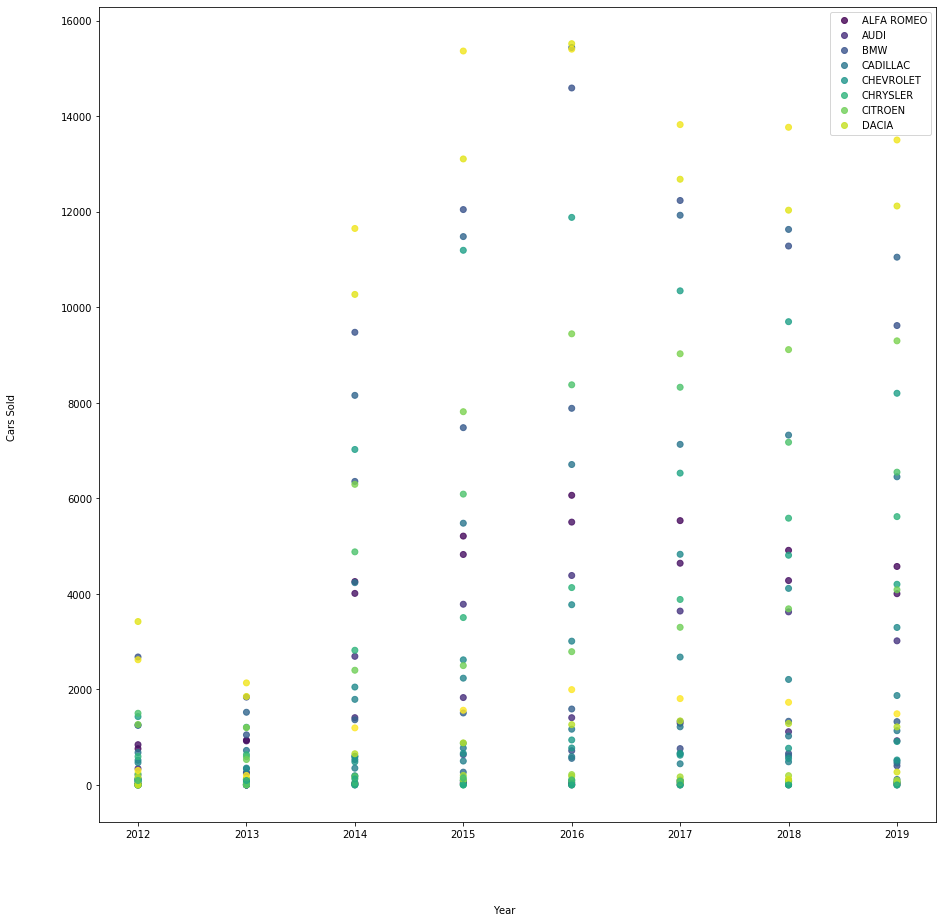


Figure 7 Scatter plot of car sales from 2012 to 2019

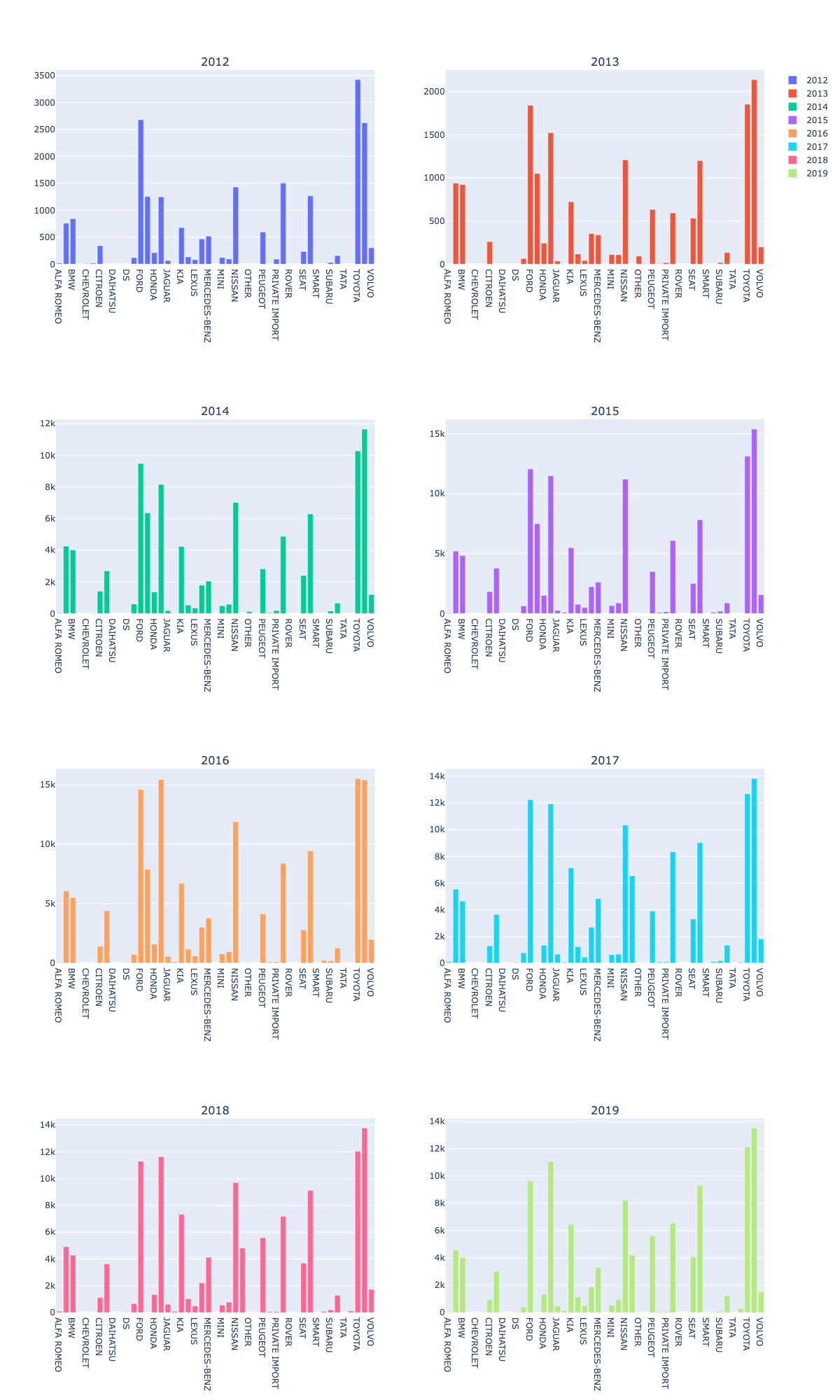


Figure 8 Individual sales of different car manufacturer from 2012 to 2019

In the figure above, the overall sales of each car manufacturer in each year is given. As mentioned above Ford”, “Hyundai”, “Toyota”, “Volkswagen” and “Nissan” show a greater sales ratio as comparison with other car manufacturers in the same year. This is because of the fact that these cars are durable and they are cheap as compared to other expensive car manufacturing brands.

In the figure below we have the overall sales of all the brand over these years. The eventual rise in the cars sales shows a positive growth in these industries and the customer demand which ultimately provides other horizons to these industries in the form of car ride hailing services which in turn imply the domestic demand as well as demand due to tourism which would be discussed in the later sections. The slight decline in 2019 onwards is because of the COVID and how it affected the economies all over the world.

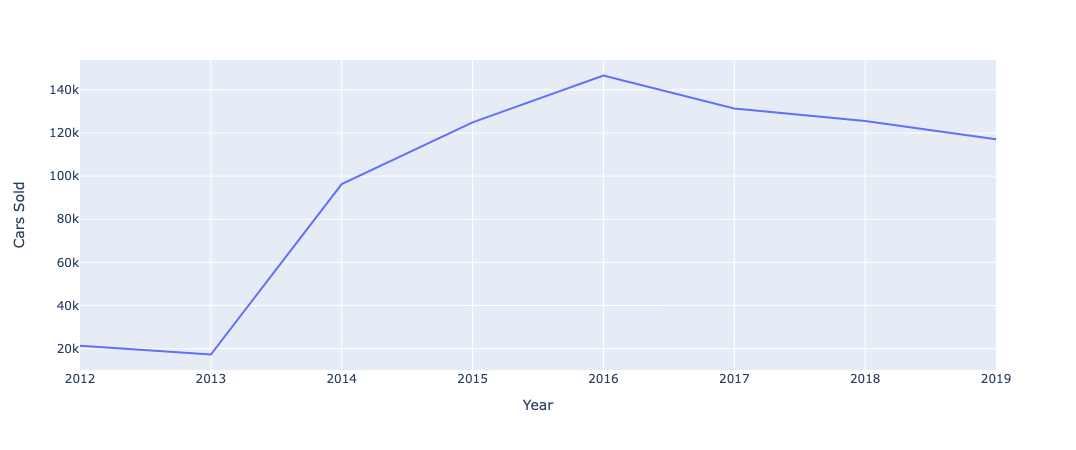


Figure 9 Yearly aggregated sales of all the cars from 2012 to 2019

## Tourism Industry Dataset

The dataset of tourism industry was gathered from the Ireland’s open data portal of government. The dataset “TMQ05 - Tourism and Travel Estimated Earnings and Expenditure” was published by Central Statistics Office and contains 8 columns and 97 rows. It is dated from 2012 to 2019.

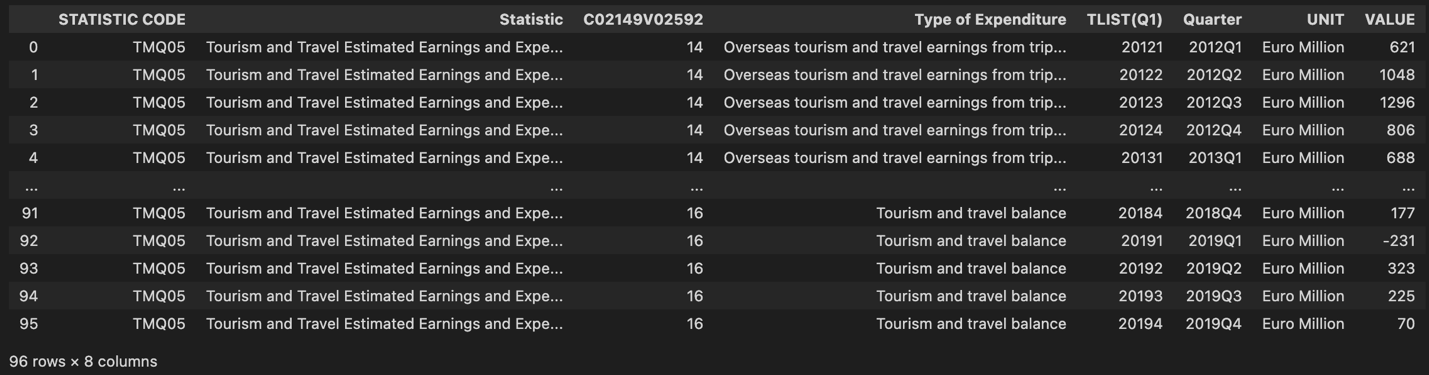


Figure 10 Tourism dataset

The dataset columns along with their metadata are given as follows.

Table 2 Tourism dataset datatypes and column description

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Column Type** | **Column Description** |
| STATISTIC CODE | Object | Contains a unique code for this statistic file. |
| Statistic | Object | Statistics displayed in the file. |
| C02149V02592 | Int64 | Category code. |
| Type of Expenditure | Object | Categorical variable with 'Overseas tourism and travel earnings from trips to Ireland', 'Overseas tourism and travel expenditure by Irish residents abroad', 'Tourism and travel balance' as the categories. |
| TLIST(Q1) | Int64 | Year and quarter in integer format. |
| Quarter | Object | Year and quarter in string format. |
| UNIT | Object | The unit in which the revenue is calculated (Euro Million) |
| VALUE | Int64 | The revenue which is generated from the tourism category. |

Some statistics related to this dataset are also given below.

* Dataset has zero null entries.

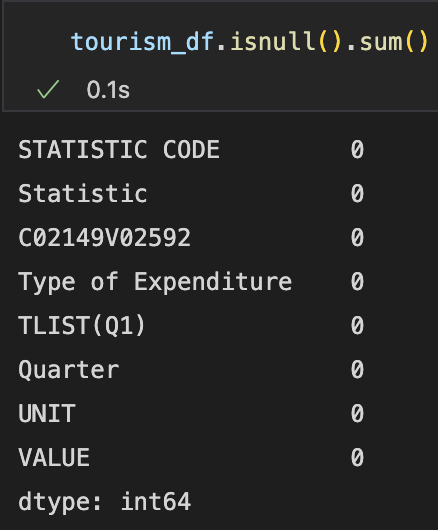


Figure 11 Data features and their data types

* Dataset has zero duplicate entries.

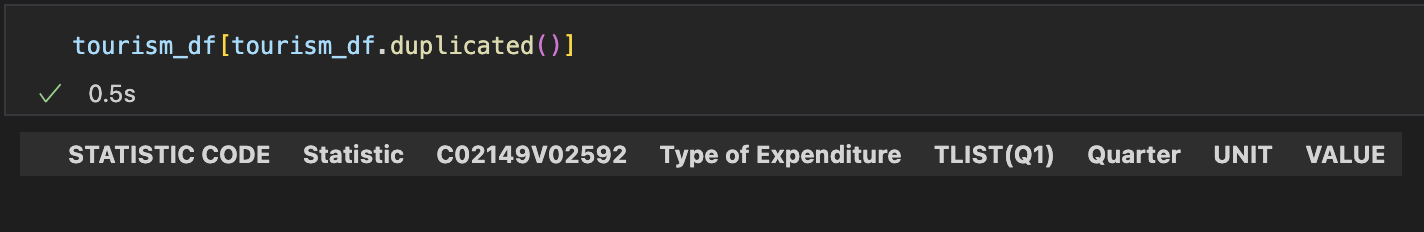


Figure 12 Tourism dataset duplicates

The histogram of the dataset is also given as follows. The revenue generated from tourism is somewhat uniform with a slightly left skewed irregular behavior. Whereas the years are evenly distributed.

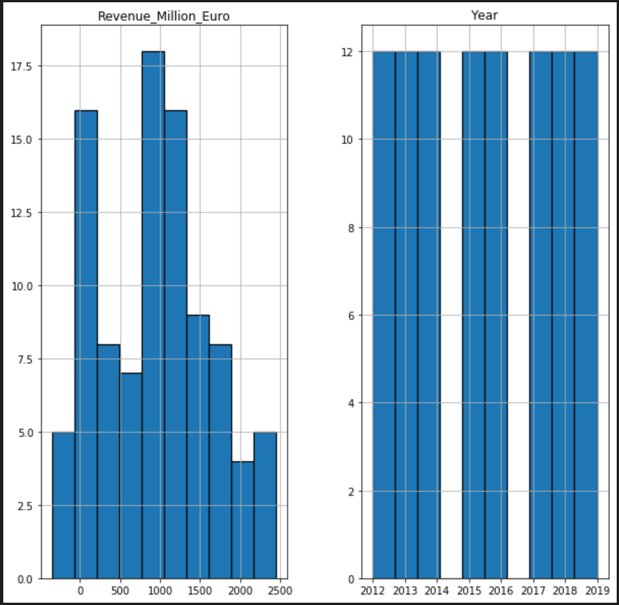


Figure 13 Tourism dataset histogram

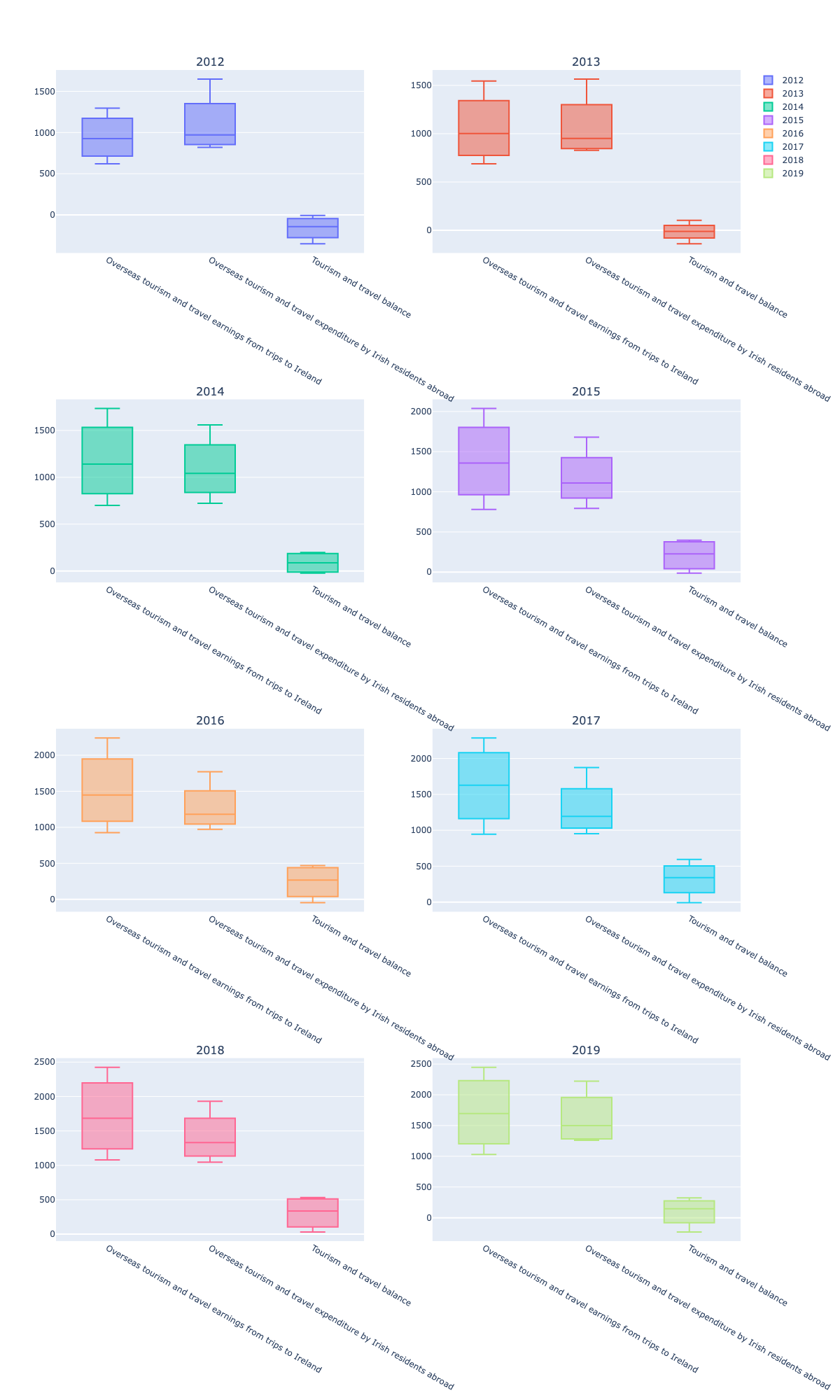


Figure 14 Tourism dataset category-based boxplot stats

Here we also do the outlier analysis by plotting the categorical data with numerical data in the form of boxplots. Given above is the boxplot which shows us the revenue generated by foreigners by coming to Ireland for the tourism purposes and we can clearly see that the revenue generated by foreign tourism has been growing in comparison to the money spent by Irish people who go to other countries for the tourism. It is a clear indicator that tourism industry has been boosting over the time and if given proper attention it could generate billions of euros.

This can also be seen through the graph which is given below. Here we can also see that the tourism revenue deficit which was created due to the fact the tourism was not as much prevalent in the earlier years as compared to the later years leading to 2019 where that deficit has turned to positive. This suggests an overall revenue surplus in monetary terms.

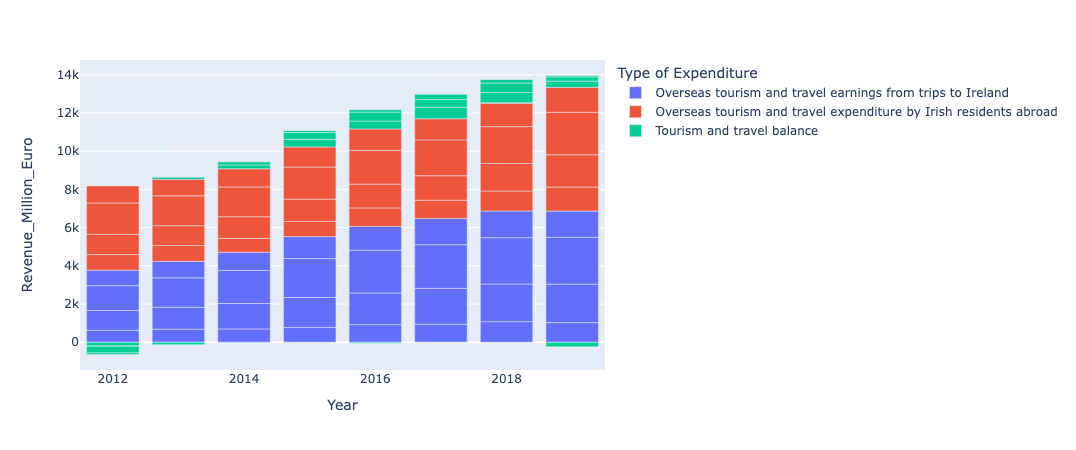


Figure 15 Tourism dataset stacked bar plot category wise tourism revenue

Now in the graph below we analyze the quarterly behavior of the revenue generated. We see that most of the tourism related revenue is generated in the 3rd quarter of the year in the months of July, August and September. It can be attributed to the fact that the weather in these months is very good.

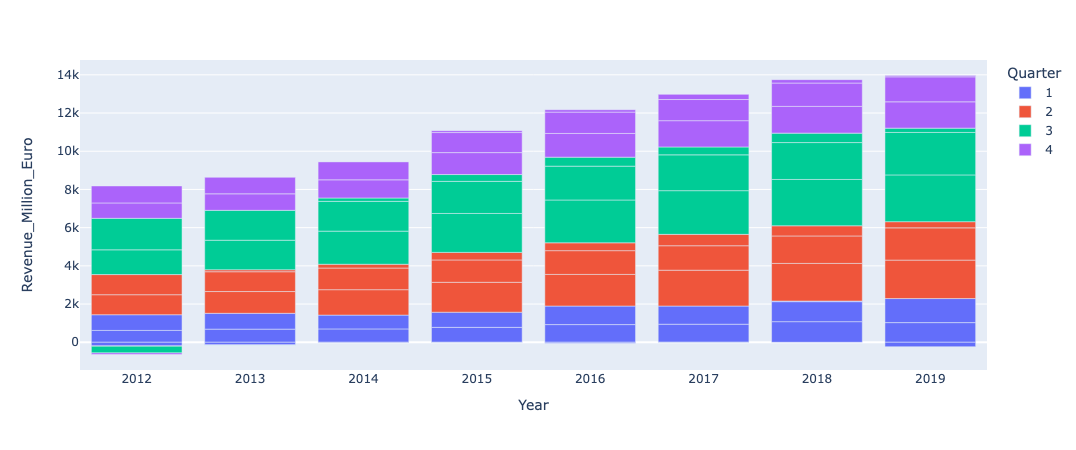


Figure 16 Tourism dataset quarterly based tourism revenue

The detailed boxplot of this behavior year wise can also be seen below.

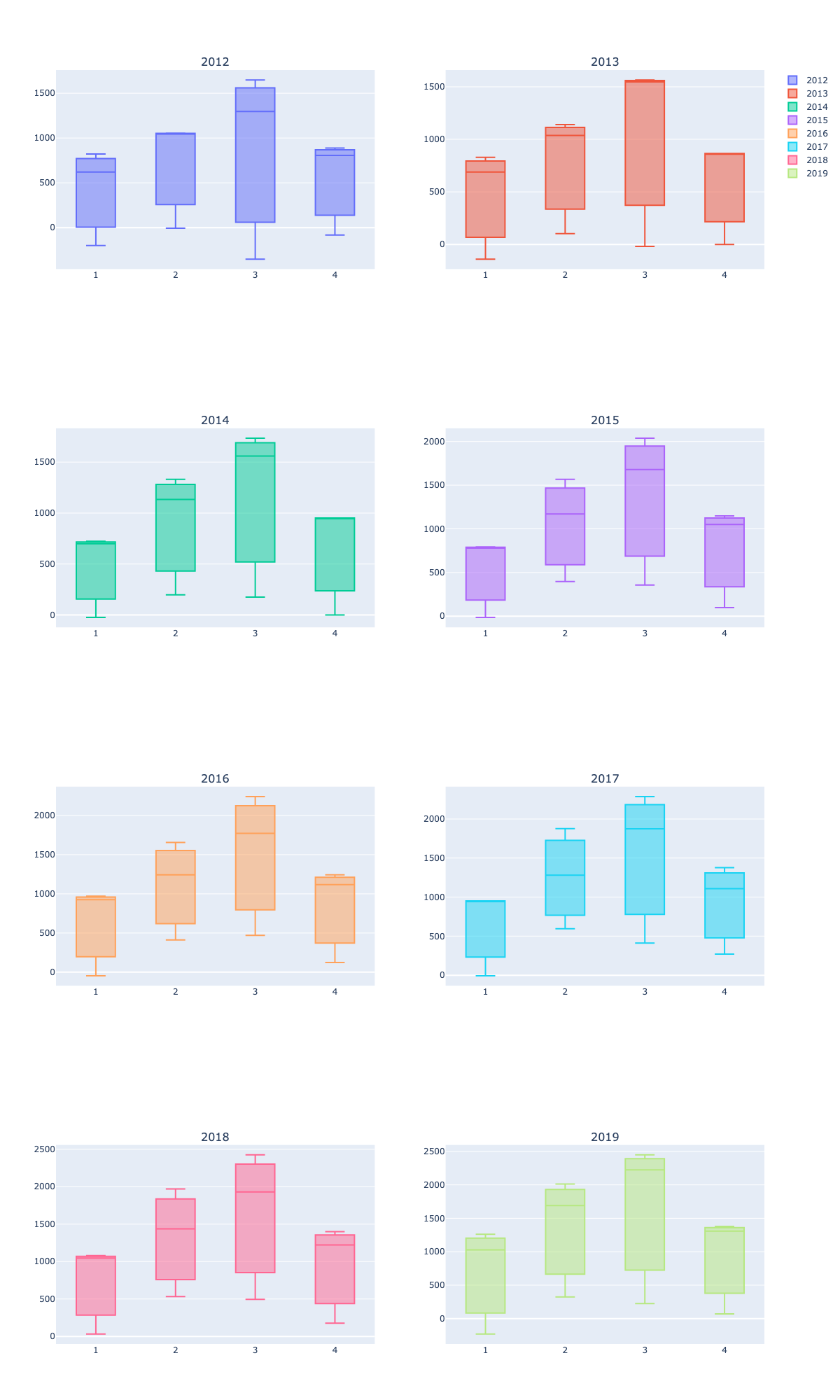


Figure 17 Tourism dataset quarterly tourism category revenue

Whereas the yearly aggregated tourism revenue (Overseas tourism and travel earnings from trips to Ireland) is also projected in the following graph. This shows a steady climb over the years suggesting that the tourism sector has been generating more and more revenue over the years.

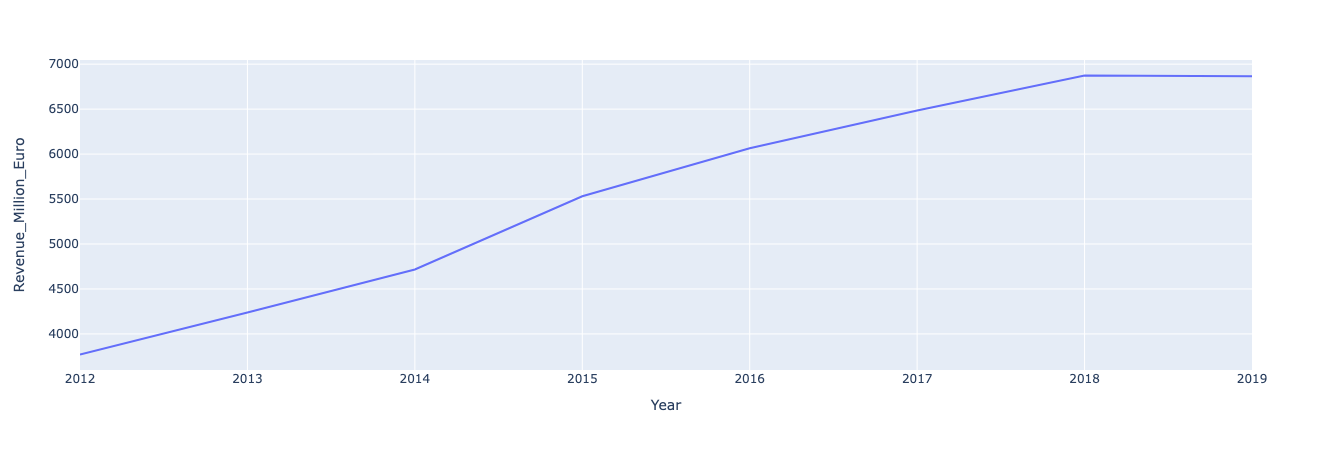


Figure 18 Yearly tourism revenue

# Data Preparation

The data preparation steps involving merging the two datasets into one so that one analysis and modelling can be done which would consist of all the relevant data and give us a clear picture of how the data conforms with our assumptions and how well the data can be modelled.

The data sets were already validated for null values or any other deficiency in the previous step in Data Cleaning and all such issues are resolved if present.

The dataset obtained by merging the yearly aggregated datasets of the tourism department and the car manufacturer industry is given as follows.

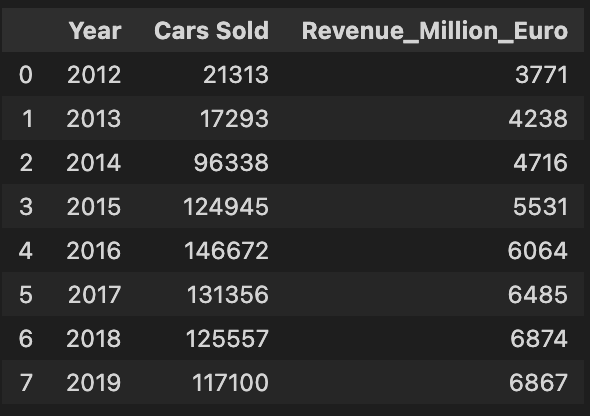


Figure 19 Master dataset after merging cars and tourism dataset

The histogram of the merged dataset is also given as follows. We can see that the number of cars sold each year has a rightly skewed distribution whereas the tourism revenue also has a distribution which is skewed to the right.

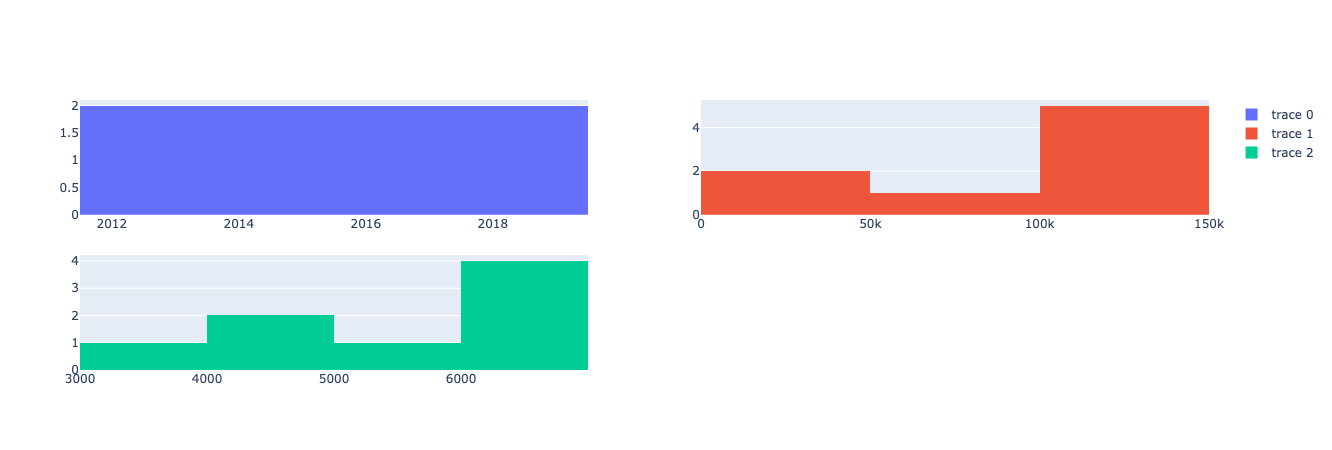


Figure 20 Histogram of master dataset

Whereas the correlation between the variables is given as follows. Here we can see that the variables are highly correlated with each other. The correlation matrix is given as follows.

Table 3 Correlation matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Year** | **Cars Sold** | **Revenue\_Million\_Euro** |
| **Year** | 1 | 0.775665 | 0.982745 |
| **Cars Sold** | 0.775665 | 1 | 0.848953 |
| **Revenue\_Million\_Euro** | 0.982745 | 0.848953 | 1 |

The heatmap generated from this correlation matrix is also given below.

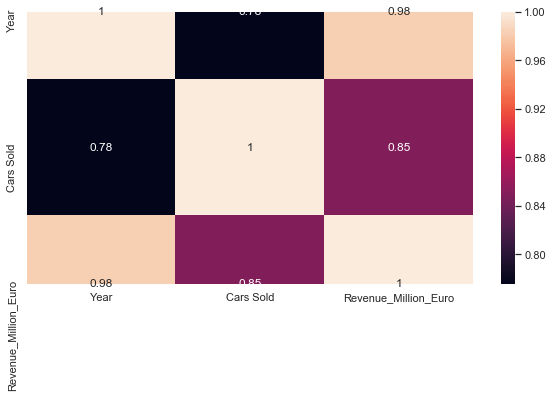
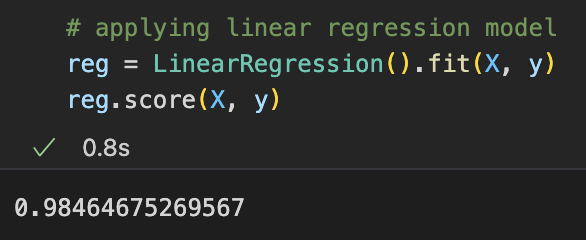


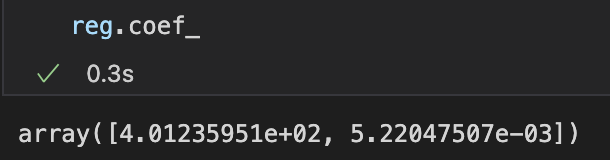
Figure 21 Heatmap of the correlation matrix

# Modeling

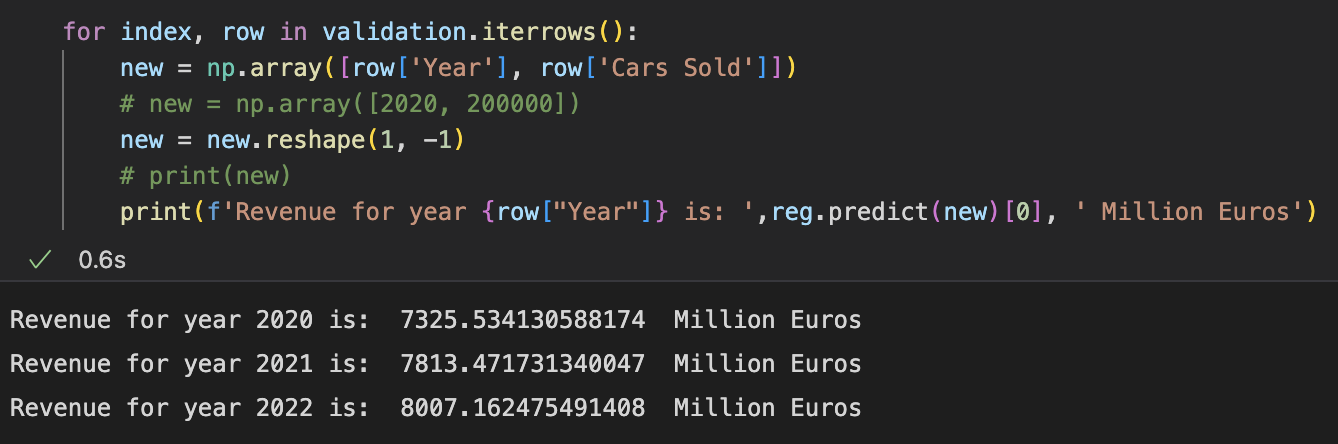
The data modelling part of this methodology involves modelling the data to produce the results which are in coherence with our hypothesis which is that with greater personal vehicles manufacturing which accounts for ride hailing services, the tourism also increases. We apply a linear regression model on the merged dataset.



The model gives us a pretty good score since there isn’t much data involved. But this also means that the model would not perform better on new unseen data. It would overfit on the existing data and would underfit on the new data. Therefore, to cater this a lot more data with much granularity is needed.



The estimated coefficients of this model are also given in the figure above. Whereas a validation dataset has also been created which consists of only 3-year data from 2019 to 2022. And the results obtained from it are somewhat confirmative to our hypothesis.



# Evaluation

The success criteria for the business under consideration are clearly mentioned and explained in the initial phase of business understanding. The question was to analyze the market trends based on the data available and forecast the performance of the tourism and auto industry in relation to each other in coming years. As both these industries were to be very critical in making the business idea a hit. The business idea was to be considered as a success if the data show growth in both these industries. After critical analysis of the data sets on hand, the analysis based on the modeling techniques such as regression modeling that shows a growth in these industries in coming years so depending on this fact the model meets the business success criteria.

It is an established fact that the data available for the modeling determines the performance, quality, accuracy and other such features of a model. So, the availability of the right data was a critical point in our project. The data available on hand was limited and was not that much extensive. It was also not in a consolidated form and was scattered. Therefore, data preparation and cleaning were the major tasks in this process.

Based on the data available although the model is performing very well as indicated from the score we are getting from the model. But we should also consider the facts that the amount of the data used for the modeling is not very large and the quality is also not the best. So, they may not perform very well for some unseen data. But under the given circumstances the results achieved are the best.

Further work can be done in doing some extensive work for looking data with more quality and data that is more meaningful. As most of direct sources for such data have already exhausted in this whole process new indirect sources can be identified and used for the extraction of data. This will add more to the accuracy and perfection of the model and hence much better results.

# Deployment

The deployment phase depends on the customer or the entity who initiated the whole process. As referenced in the guide for CRISP-DM, “the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise”. If the final deliverable is only a report then the process ends there and nothing else needs to be done The deployment phase is customer centric. Whatever is done in this phase depends on the requirements of the customer.

In this case, this report document will act as a deliverable of the project. This document is documenting and capturing all the key details of the project. This document will also serve as a reference for any further work that has to be done in this regard. This further work could be anything form the deployment in any form based on the requirement or any future work that has to be done with some more dataset. But for now this document will be the only deliverable.

# References

1. Data Science Process Alliance. 2022. *What is CRISP DM? - Data Science Process Alliance*. [online] Available at: <https://www.datascience-pm.com/crisp-dm-2/>
2. Statoo.com. 2022. [online] Available at: <http://www.statoo.com/CRISP-DM.pdf> .
3. Medium. 2022. *Using CRISP-DM to Predict Car Prices*. [online] Available at: <https://medium.com/@christophberns/using-crisp-dm-to-predict-car-prices-f15eb5b14025>.
4. Scirp.org. 2022. *Shearer, C. (2000) The CRISP-DM Model The New Blueprint for Data Mining. Journal of Data Warehousing, 5, 13-22. - References - Scientific Research Publishing*. [online] Available at: <https://www.scirp.org/(S(vtj3fa45qm1ean45vvffcz55))/reference/ReferencesPapers.aspx?ReferenceID=1592780>.