Using Advanced Analytics to Predict Vehicle Registrations and Help in Business Continuity Planning

Khalid Ibrahim AlZarouni
ka1623@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Master's Project is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
Using Advanced Analytics to Predict Vehicle Registrations and Help in Business Continuity Planning

by

Khalid Ibrahim AlZarouni

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
May 2021
RIT

Master of Science in Professional Studies:
Data Analytics

Graduate Capstone Approval

Student Name: Khalid Ibrahim AlZarouni
Student UID: 356006403

Graduate Capstone Title: Using Advanced Analytics to Predict Vehicle Registrations and Help in Business Continuity Planning

Graduate Capstone Committee:

Name: Dr. Sanjay Modak                    Date:
Chair of committee

Name: Dr. Ehsan Warriach                  Date:
Member of committee/Mentor
Acknowledgments

First and foremost, prayers of praise and thankfulness to Allah for his blessings and for making me achieve this master’s degree in data analytics.

Secondly, I would like to express my deep and sincere gratitude to my mentor Dr. Ehsan Warriach as well as the chair of graduate programs and research department, Dr. Sanjay Modak who gave me the golden opportunity to do this amazing project and for their supervision and guidance throughout the program. Also, I would like to thank all the Rochester Institute of Technology (Dubai Branch) staff for their cooperation and support during this global pandemic which showed their true capabilities.

Thirdly, I am highly obliged in taking the opportunity to thank Dubai Government, Dubai Pulse, and Road and Transport Authority for giving me the opportunity to explore their valuable data and make a great use of it.

Finally, I am extremely grateful to my family for their love, prayers, caring and sacrifices for educating me and preparing me for my future. My thanks and appreciations also go to my colleagues Thani AlMansoori and Abdelaziz Zandaki who were always very cooperative and motivational throughout the program.
# Table of Contents

Acknowledgments .................................................................................................................. iii

List of figures .......................................................................................................................... vi

List of Tables ........................................................................................................................... vi

Abstract ........................................................................................................................................ 7

Statement of the Problem ........................................................................................................... 8

Background of the Problem ........................................................................................................ 9

Project Definition and Goals ..................................................................................................... 10

Limitations of the study ............................................................................................................... 10

Project Budget ........................................................................................................................... 10

Literature Review ....................................................................................................................... 11

  Queue Management: ................................................................................................................. 11

  Business Continuity Plan: .......................................................................................................... 12

  Using Analytics for Business Improvement: ............................................................................ 14

    Forecasting the Future: ........................................................................................................... 14

    Identifying Opportunities: ...................................................................................................... 14

  Using AI and Data Science for Solutions: .............................................................................. 14

  Web Scarping: ........................................................................................................................... 15

Methodology .............................................................................................................................. 17

  Data Collection ......................................................................................................................... 17

    Data Extraction Using Web Scraping ...................................................................................... 17

  Data Staging, Combining, and Feature Engineering ................................................................. 18

  Data Description ....................................................................................................................... 19

  Data Pre-Processing .................................................................................................................. 20

  Exploratory Data Analysis Using Python and Tableau ............................................................. 21

  Inferential Analysis Using Python .......................................................................................... 21

    Arima Model .......................................................................................................................... 22

Analysis ......................................................................................................................................... 23

  Data Collection ......................................................................................................................... 23

    Web Scarping Python Code ..................................................................................................... 23

  Data Staging, Combining, and Feature Engineering ................................................................. 24
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importing Libraries</td>
<td>25</td>
</tr>
<tr>
<td>Importing Data</td>
<td>25</td>
</tr>
<tr>
<td>Data Exploration</td>
<td>26</td>
</tr>
<tr>
<td>Data Pre-Processing</td>
<td>28</td>
</tr>
<tr>
<td>Identifying and Removing Null Values</td>
<td>28</td>
</tr>
<tr>
<td>Build boxplots and Identify Outliers</td>
<td>29</td>
</tr>
<tr>
<td>Exploratory Data Analysis Using Python and Tableau</td>
<td>30</td>
</tr>
<tr>
<td>Univariate Summary Plots:</td>
<td>30</td>
</tr>
<tr>
<td>Bi-Variate Analysis Using Tableau</td>
<td>32</td>
</tr>
<tr>
<td>Inferential Analysis Using Python</td>
<td>40</td>
</tr>
<tr>
<td>Ordinary Least Squares Regression (OLS)</td>
<td>40</td>
</tr>
<tr>
<td>Arima Model</td>
<td>41</td>
</tr>
<tr>
<td>Results</td>
<td>49</td>
</tr>
<tr>
<td>Enable tracking of current utilization and productivity of employee:</td>
<td>50</td>
</tr>
<tr>
<td>Analyze the collected data from tools and take appropriate actions:</td>
<td>50</td>
</tr>
<tr>
<td>Creating a productive environment for employee:</td>
<td>50</td>
</tr>
<tr>
<td>Better planning and scheduling of employee:</td>
<td>50</td>
</tr>
<tr>
<td>Challenges Faced During the Project</td>
<td>51</td>
</tr>
<tr>
<td>Conclusion and Future Work</td>
<td>52</td>
</tr>
<tr>
<td>Bibliography</td>
<td>53</td>
</tr>
</tbody>
</table>
List of figures
Figure 1. Queue waiting systems of multiple variations..........................................................11
Figure 2. Basic structure for web scraping..............................................................................15
Figure 3. Web scraping flow diagram ....................................................................................17
Figure 4. Number of files downloaded ..................................................................................18
Figure 5. Number of files and its structure ..........................................................................24
Figure 6. Python imported libraries .......................................................................................25
Figure 7. Correction of the date format .................................................................................25
Figure 8. Data frame .............................................................................................................26
Figure 9. Data Information......................................................................................................26
Figure 10. Data Describe ......................................................................................................27
Figure 11. Checking for null values in the data ...................................................................28
Figure 12. Box plot for variables .........................................................................................29
Figure 13. Univariate data visualizations ..............................................................................31
Figure 14. Skewness of data frame .......................................................................................31
Figure 15. Registered vehicle manufacturers in the battery electric motor segment ..........32
Figure 16. Registered vehicle manufacturers according to benzene fuel type ....................33
Figure 17. Registered vehicle manufacturers according to benzene-electric fuel type .........34
Figure 18. Registered vehicle manufacturers according to fuel type ................................35
Figure 19. Yearly registration per manufacturer ................................................................35
Figure 20. Number of registrations per age group ...............................................................36
Figure 21. Number of registrations per vehicle type ..............................................................37
Figure 22. Number of weekly registrations ..........................................................................38
Figure 23. Registration as per nationality ............................................................................38
Figure 24. Monthly registration for vehicle .........................................................................39
Figure 25. OLS regression results ........................................................................................40
Figure 26. Daily registrations of vehicles against date ..........................................................41
Figure 27. Rolling mean & standard deviation of daily registrations ...................................42
Figure 28. Rolling mean & standard deviation for first difference of daily registration data ..........................................................43
Figure 29. Autocorrelation plot of the registrations ...............................................................44
Figure 30. Autocorrelation plot for seasonal first difference data .....................................44
Figure 31. Partial autocorrelation plot for seasonal first difference data .............................45
Figure 32. Results from ARIMA model ..............................................................................46
Figure 33. Predicted time series for the future ....................................................................47
Figure 34. Actual vs predicted times series ........................................................................48

List of Tables
Table 1. Brief description of the data ..................................................................................19
Table 2. Augmented Dicky Fuller test p-value before and after transformation ....................49
Table 3. Predicted values vs actual values results ................................................................49
Abstract
Dramatic changes happen in the world. People and organizations face problems with these changes. The ones who have everything planned in case of emergencies; survive, and the ones who do not plan it right, have a chance of falling. Today, a lot of organizations suffer because of the Covid-19 global pandemic, even the largest organizations in Dubai like Roads and Transport Authority. Roads and Transport Authority (RTA) is one of the largest government entities and the sole planner and executer of all transport, road, and traffic projects in Dubai. It has happiness centers which register all the cars that are in Dubai. Those centers capacity is reduced because of this disease which led to making the long queues for customers. Customer’s satisfaction is one the main priorities of these centers. Not having this problem solved; will increase the number of unsatisfied customers. In this capstone project, I propose to explore the data of the registered cars then predict the upcoming customers. By predicting customers, we will solve all the problems that the organization is facing in this pandemic and plan for future dramatic scenarios.

Keywords: Roads and Transport Authority, vehicle registration, predict, business continuity planning, forecast, ordinary least squares regression, ARIMA model.
Statement of the Problem
In wake of the Global covid-19 pandemic every organization is suffering through mass reduction in efficiency. As per the customer reviews attained from Road and Transport Authority of Dubai, it was revealed that there is certain reduction in customer satisfaction. At the customer happiness centers established by RTA, lengthened queues as well as waiting hours were experienced by customers. The centers specialize in customer satisfaction; however, such circumstances have led to unsatisfied customers. Employee Optimization problems being faced by the RTA is the prime threat to the long-standing efficiency of the centers. Business continuity plan is required to be formed to re-pace the operations for future purposes.
Background of the Problem

A lot of dramatic scenarios can happen throughout the year and this year, organizations went through several struggles. One of them is the Covid-19. Having social distancing because of the disease made the customer happiness centres decrease the total capacity. Reducing the capacity and employees led to crowded registration centres, long queues, and long waiting hours. By having said that, customers became unsatisfied of the service.

This problem can be solved by collecting all the previous registration data and predicting the customers, then, optimizing the employee’s efficiency. Also, help plan for the future and have a clear strategy on how to react to different scenarios and events.

Covid-19 crisis had affected many companies around the world. Data and analytic leaders can prevent and lessen the impact of COVID-19 in their companies by using some steps and tools which will help in continuation of the business continuity plan these steps include: integrating organizational data with a comprehensive view of COVID-19 from external sources like WHO and CDC, identifying business scenarios, identifying data needed in order to manage and monitor areas of business that are impacted for example supply chain, finance, human resources etc., promoting the use of data to guide business practices and enhance well-being of the company and employees, developing core KPIs, designing an organizational dashboard that helps in enabling business impact measuring and tracking and considering how data and analytics platforms can help with staff communications (IT Convergence, 2020).
Project Definition and Goals
Using artificial intelligence and machine learning depending on the data we have we can:

- Optimize employee’s capacity efficiently.
- Reduce long queues and long waiting hours.
- Predict upcoming customers to help business continuity planning.
- Increase customer satisfaction.

Limitations of the study
- Limited data available.
- Limited coding experience.

Project Budget
This project does not require any budget since the data is from Dubai Pulse.
Literature Review

Queue Management:
Researchers have carried out studies to understand the main points associated with long queues at point of sales. As a retail supermarket, restaurants or a car dealer, researches have revealed that the long queues at any store leads to bad customer experience and even they lose most of the customers due to long waiting time. They have also studied and found what are the major costs associated with queuing. Stevenson and Hojati have identified majorly two kinds of cost associated with long queue at stores. One is the cost related to waiting customer in queue and the other is cost associated with capacity management of the long queues. Some of the examples of cost associated with queues are the salaries paid to an idle employee as there is long queue at previous counter, the cost associated with any customer leaving without any purchases. When does a queue size increases at any point of goods/service delivery? When there is a frequency mismatch between number of customers and number of servers in any situation, there is formation of long queue at those points. This mismatch might be due to shortage of serving staff or shortage of infrastructure to support the customers and servers. So, managing a queue is often thought of as an optimization problem where the goal is to reduce the cost associated with queue and increase the serving capacity as much as possible (William J. Stevenson, 2007). Queue waiting systems can be of multiple variations. Below few of the waiting systems are shown:

![Queue waiting systems of multiple variations](image)

*Figure 1. Queue waiting systems of multiple variations*

*Source: (Vidyarthi, 2013)*
In other researches, they have suggested three categories of models for managing queue. These models for queue managing are descriptive model for queuing, analytical model for queuing and prescriptive model for queuing. In descriptive model for queuing, they have suggested to analyze the effect of 2 counters instead of 3 counters on the queue (Howard J. Weiss, 1989). In prescriptive model for queuing, the emphasize on some modelling technique which focuses on improving the waiting situation. They suggested that the analytical model for queuing can be used effectively in certain scenarios especially in cases when customers arrive in random fashion and prediction is difficult for their arrival (Slack & al, 1995).

Business Continuity Plan:
Business continuity plan is a strategy to guarantee that an organization will adapt to a certain situation and continue its operations in all possible emergencies and disasters or to get back to its operations as soon as possible (Merrimack College, 2020). According to a study done by Mercer back in late February, “27.2% of companies did not have a business continuity plan or pandemic preparedness plan in place, and an additional 23.8% did not have a plan but were in the process of developing one.” (Bracey, 2020).

A business continuity plan ensures backup of data and maintains continuation of operations in a company in case of any occurrence of disaster event. On the hand, a disaster recovery planning must also be included to restore business operation and prevent any data loss that might be critical when an event of disaster happens. Moreover, companies with effective business continuity planning can get lots of benefits such as lessening the impact upon staff, organizational supple chain, service delivery and IT infrastructure, defending the corporate reputation, lessening financial impact and enabling the organization to return to normalcy sooner (IT Convergence, 2020).

When planning to start a business continuity plan it is important to include these four elements which are threat analysis, role assignment, communications and backups. In addition, a company must consider having a load balancing business continuity and a disaster recovery plan that has two factors which are: recovery time objective (RTO) and recovery point objective (RPO) to maintain its business continuity plan and enable worst troubles with less damage (Imperva, 2020).
Having a business continuity plan in every company it very important to avoid failures that can exist in real-life business and increase their risk of disaster, here are some examples that might occur with poor business continuity planning: no business continuity plan, no risk assessment, no business impact analysis, no prevention, and no recovery plan. To avoid these disasters a company must have a business continuity plan with a framework that has its own protocol for prevention and recovery to outline its threat, a list that defines risks of disasters scenarios that can happen, a list of analysis that can analyze how this threat can affect the business, protocol, and technologies to prevent disruptive events from happening and a clear path of protocol and systems for recovery. Not to mention, data loss, cyberattacks, malware and viruses, network and internet disruptions, hardware or software failure, fire etc. are examples of threats a company can experience and face if business community plan was poor and not taken seriously (Rock, 2018).

A company must train their individuals on an on-going basis specially if they are going to get involved in the development, implementation, evaluation, maintenance, exercises and executes of business continuity plan and they should not be trained only once. In order to train these individuals, they should go through four training phases including: pre-planning, planning, post-plan development and pre-exercise. Individuals can gain and develop lots of knowledge necessary for business continuity plan from this training, as an example, training individuals for planning methodology can be beneficial as they can adapt proper project management methodologies and tools, have a clear understanding about business continuity plan terminologies, apply an accepted methodology, implement documentation standards and software training. Furthermore, each individual must also be trained in order to know their roles and responsibilities towards several perspectives to understand the plan. Besides that, in post-plan development and pre-exercise phases elements where included. The elements that are included in post-plan development includes: role of team leaders and team members that consist of alternates, business continuity plan provision, notification, event logging and maintenance procedures, on-going evaluation, and post-execution review procedures. Coming to the elements included in the pre-exercise training phase, these elements are testing methodology and scheduling, developing objectives of exercises and scenarios, plan modification and update procedures subsequent to and based on the result of the exercises and auditing and evaluation the business continuity plan (San Jose State University, 2020).
Using Analytics for Business Improvement:
The data the organization has is crucial to how to predict for the future and to know how the business is running. The more the data, the more it allows the organizations to know how to run the business operations efficiently. Having big amounts of data can let the organization predict how the business is doing in certain months or in different occasions. Moreover, applying analytics can help the business to know the customer and how each customer is participating (Deoras, 2020). Furthermore, there are several things that companies can do to make use of big data:

Forecasting the Future:
During Covid-19 pandemic this year, companies got more cautious about the changes it made and took a small hint of how disruptions can happen to business unexpectedly. Having this kind of data in the current situation made businesses decide what is the best way to predict future incidents, analyze the risk, and develop the best way to react to it. They can depend on the data of this historical event and then predict for future unfortunate times (Chawla, How Sindhu Gangadharan Is Prioritising Business Continuity For SAP Labs, 2020).

Identifying Opportunities:
Exploring the data can help in identifying the gaps within the company and fill it will great opportunities. Understanding how your customers and employees engage can help you get better ways to solve their problems. It can help employees by giving them flexible timings depending on customers coming each day (Das, 2020).

Using AI and Data Science for Solutions:
Many industries have worked on being creative in which they used the data they had to find AI powered solutions. For example, having an AI chatbots to detect the problem from the certain words that has been put in the specific sentence then find the solution based on it. Furthermore, studies indicate that chatbot sales will increase around 50 times the sales that were pre Covid-19 (Chawla, Ensuring Business Continuity With AI During The Recession, 2020).
**Web Scrapping:**
Web scraping is a technique to extract or download desired from any website or web applications. In today’s digital world where the data is being collected at a rapid scale, these web scraping technologies and techniques help the individual or organizations to have access to the relevant data very quickly and efficiently (Yesi Novaria Kunang, 2018). Generally, in web scraping user writes a program which automatically runs commands on web pages (mostly using APIs) requesting to extract data. Then the program subsequently parses the data to extract it in desired format (Mitchell, 2018).

Web scraping is a means for data mining. The end goal of any web scraping technique is to extract the desired data set in required format like csv (comma separated values) format, spreadsheets or even directly into database (Mitchell, 2018).

As shown above researchers explains that any web scraping process will have these three things. First, there will be data stored in websites in their HTML web pages. These data can be in tabular format in HTML web pages as on embedded entity or it can be as an attachment in the web pages. Second, there will be a web scraping technology which can interact with these websites and HTML web pages. These technologies might be any program (python programs are widely used for web data scraping) or it can be any web scraping tool in market. Third, the output from all these processes. The output will be in terms of desired data which will be in structured format so that is can be used in data analysis or any other purpose (Sirisuriya, 2015).
There are various web scraping techniques suggested by researchers and industry experts to carry out the data extraction tasks. These techniques involve text wrapping using Unix commands, data extraction using HTTP requests, using semi structured query languages like HTQL and XQuery to parse web content and retrieve the data in desired format. Apart from this there are other techniques like customized web scraping software which reduces the effort to write any code manually, there are aggregation platforms which creates bots for extracting the data, there are computer vision techniques like OCR which can help in extracting data from web pages (Sirisuriya, 2015).

Most of the web scraping programs or software are written are quite generic and are written for general data extracts. This web scraping software might not be very flexible, and we must write some customized code on top of it use it for our use case. All the web scraping software available in market try to cover all different types of data extractions. But then also researchers have observed that some scraping software is more suitable for certain scenario.

With so much advancement in data analytics and machine learning, organizations and researchers are keen on collecting various kinds of data in desired structured format from online platforms. Also, there is a digital shift of using APIs for all interactions between systems. Thus, all the websites now a days have an exposed website which anyone can access to data. This is the reason why scraping code in python are becoming very popular. These programs connect with the web page APIs and extract structured data from the website in the desired format (Pratiba, 2018).
Methodology

Data Collection
For the purpose of this research, we have collected data from Dubai Pulse. The data uploaded on Dubai Pulse is from Roads and Transport Authority (RTA). It is the daily registration details of all the cars registered in Dubai and it is updated weekly. The available data is from 22-September-2019 and ongoing. The data collected was secondary data. The data has few missing values (NaN). These will be corrected in data preprocessing.

Data Extraction Using Web Scraping

The above representation shows the methodology used for data extraction. The source of data was Dubai Pulse government website. We created python script to automatically scrap the data from website. We have used the inbuilt BeautifulSoup library in python. BeautifulSoup helps us extract data from HTML or XML files using our desired parser. The parse tree can be navigated, searched, and modified in a variety of ways with the help of BeautifulSoup. It is popular for programmers as it saves hours or even days of work. The web scraper gave us the csv files in desired format. We used these extracted files in our statistical analysis.
The extracted data was in 463 different files. Each file corresponding to different dates from last one and half year. As a part of first step, we have used the file names to create a new variable which is date of registration. Then, we accumulated all 463 files into single file. We used the inbuilt glob library in python for it. Glob is a generic term that refers to techniques for matching specified patterns using Unix shell rules. The glob module in Python is used to find files/pathnames that fit a template. The glob pattern rules are the same as the Unix route expansion rules.

For feature engineering, we will not be using any variables from the dataset to do it. We will be using the titles of each csv file and add it to the data as a variable named date.

Now we combined all the datasets together into a main data file and have 18 variables ready, including the date.
Data Description

For our project we are provided with secondary data and we will be doing all the analysis on that secondary data. The secondary data collected is in CSV format and it contains 17 columns. We have different files for each of the different dates from last one and half year. Thus, we have a total of 463 files. Later we add the date column in each file so that the number of columns increase to 18. Each file contains different number of rows as the rows corresponds to number of sales made on that particular day. A quick data overview is given below:

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>nationality_e</td>
<td>Nationality of the customer</td>
</tr>
<tr>
<td>gender_e</td>
<td>Gender of the customer</td>
</tr>
<tr>
<td>person_birth_year</td>
<td>Birth Year of the customer</td>
</tr>
<tr>
<td>vechile_type</td>
<td>Type of Vehicle purchased by the customer</td>
</tr>
<tr>
<td>manufacture_year</td>
<td>Manufacture year of the vehicle</td>
</tr>
<tr>
<td>cylinders_num</td>
<td>Number of Cylinders in the vehicle</td>
</tr>
<tr>
<td>cylinger_capacity</td>
<td>Capacity of the Cylinder in the vehicle</td>
</tr>
<tr>
<td>color</td>
<td>Colour of the vehicle</td>
</tr>
<tr>
<td>business_individual</td>
<td>Business type of the customer</td>
</tr>
<tr>
<td>fuel_type</td>
<td>Fuel type of the vehicle</td>
</tr>
<tr>
<td>transmission_manual_automatic</td>
<td>Transmission type of the vehicle</td>
</tr>
<tr>
<td>number_of_passenger_seating_capacity</td>
<td>Passenger Seating Capacity of the vehicle</td>
</tr>
<tr>
<td>number_of_doors</td>
<td>Number of Doors in the vehicle</td>
</tr>
<tr>
<td>brand_model</td>
<td>Brand model of the vehicle</td>
</tr>
<tr>
<td>empty_weight</td>
<td>Weight of the vehicle</td>
</tr>
<tr>
<td>origin_country_the_place_where_assembled</td>
<td>Country of origin for the vehicle</td>
</tr>
<tr>
<td>manufacturer_brand_name</td>
<td>Brand Name of the vehicle</td>
</tr>
</tbody>
</table>

Table 1. Brief description of the data
Data Pre-Processing
Following set of steps were performed for data cleaning and preprocessing:

Step 1: Import all the necessary python libraries in Jupyter notebook.

As a result, I used the import command to import libraries from Python, and this is the most widely used library by data scientists. NumPy is a fundamental Python package for scientific computing.

Step 2: Import data set in Jupyter notebook for preprocessing

Then the accumulated file was imported to python Jupyter notebook for further processing.

Step 3: Clean the data by detecting the null and missing values and then either eliminate them or replace them with appropriate values.

We have used various methodologies to detect and eliminate/replace null values.

- For numerical variables, we have replaced null values with mean/median of column.
- For categorical variables, we can replace null by most frequent category or create a new category.

Step 4: Identify the outliers and remove them.

We have built boxplots using seaborn library. There are multiple methods to remove Outliers. They are:

- Using z-score
- Using mean and standard deviation
- Using Inter Quartile Range.

We have used 3rd method to remove outliers. IQR is practically the middle 50% of data lying between lower and upper quartile ranges. It is also called mid-spread. We first calculated the IQR range and then any value lying outside IQR was considered outlier and was removed.
**Exploratory Data Analysis Using Python and Tableau**

Python programming language was used for data cleaning, data exploring, and data interpreting. Tableau was mainly be used for presenting the data that is being cleaned and explored in the other tools since it is very user friendly and has great visualizations. Exploratory analysis was done in tableau on the data collected. Moreover, we have performed the descriptive analysis on the data to have an enhanced understanding of the it. The data have been explored to find some good insights about the factors which have significant effect on the vehicle registrations. The exploratory data analysis was intended for making us more prepared with good information of the data and have good understanding of the vehicle sales. By exploring the data, visualizations are produced to view the trends.

**Inferential Analysis Using Python**

We have decided to use ARIMA model to predict the sales of vehicle registrations since it is the most suitable. After exploring the data, it has been revealed that a positive association exists between the future and the past daily registrations and these registrations differ drastically over a period. Therefore, we must select a prediction model which predicts the registrations on daily basis. By stating that, this shows us that a non-liner approach is a great way to go through. This is the reason why we choose ARIMA model for prediction of daily sales of vehicle. Moreover, if we observe the daily registration of vehicles, we can see that it is dependent on time and any variable that is dependent on time fails to satisfy the theory of simple linear regression models. This strengthens our choice for choosing ARIMA model for predicting the future daily registration of vehicles.
Arima Model
ARIMA (Auto Regressive Integrated Moving Average) is a statistical examination model that is applied to understand a set of data then predict future patterns. It identifies the qualities of one variable and comparing it to other factors. This model has two parts which is auto regressive (AR) and moving average (MA) and the (I) is the differencing error between them (Prabhakaran, 2021). The AR-I-MA are often used as:

- Auto regressive (AR) = (p) value
- Integrated (I) = (d) value
- Moving average (MA) = (q) value

A Unit root test is performed on the data to check if the data is stationary or non-stationary. The most used test is the Augmented Dick Fuller test, and it is a statistical examination test that help us find out how significant a time series data can be explained (Prabhakaran, 2021).

We can have the below mentioned hypothesis of the test in mind:

1. Null Hypothesis (H0): It can be given by the statement that any time series data can be characterized by a unit root that is non-stationary.
2. Alternative Hypothesis (H1): Alternative Hypothesis of the test is that the time series data is stationary.

Interpretation of p value

1. p value > 0.05: We fail to reject the Null Hypothesis (H0) which means the data has a unit root and is non-stationary.
2. p value ≤ 0.05: We reject the Null Hypothesis (H0) which means the data is stationary.

When applying this test, we will be able to identify whether the data needs any transformation or not. For example, logarithmic transformation.
Analysis

Data Collection

Web Scraping Python Code

```python
import urllib.request
import requests
from bs4 import BeautifulSoup

def getAllLinks(url):
    allLinks = []
    res = requests.get(url)
    soup = BeautifulSoup(res.content, features='lxml')
    filesItem = soup.find_all('div', {'class': ['file-item']})
    for item in filesItem:
        link = item.find('a', {'class': ['action-icon-anchor']}).get('href')
        allLinks.append(link)
    return allLinks

def downloadFile(link, date):
    urllib.request.urlretrieve(link, 'input/Car_Registration_{0}_00-00-00.csv')

url = 'https://www.dubaipulse.gov.ae/data/allresource/roads-and-cars/rta_car_registration-open?count=459&result_per_page=100&as_sfid=AAAAAXoVF5Rr8fg56_QbFhQzUVP1qG-n_Skeq8602-inTp6wrka_gFiN5gbM9Cfx4ojIVYsjo186TqGpnoi3EoEKZ3x2cwkGxmyfoCB21ibA88SmSF8wDQDTFIMSILjTU%3D&as_fid=962439aeed1b9814a385c4d1a41e7653b829c0dc'
links = getAllLinks(url)
# print(links)
    for link in links:
        file = link
        print(file)
        date = file.split('car_registration_')
        date = date[1].split('_')
        date = date[0]
        downloadFile(link, date)
print('Process has been finished')
```
We can observe that we have scraped 463 csv files. Each file contains different no of records i.e. details of all the cars registered on that particular date. As we can see from the screenshot of csv file of 22-September-2019, we observe that it contains 394 instances of 17 variables. In other words, it gives various details of 394 vehicles registered on 22-September-2019 like buyer details and vehicle details.

**Data Staging, Combining, and Feature Engineering**

Since we will combine all the data in one main file, and we will not be using any variables in the data to create the 18th variable. We will be doing the feature engineering now to have the 18th variable ready for us when we explore the data. Furthermore, we will take the title of each csv file and add it to the data as a variable named date.
Importing Libraries

The libraries used are shown below.

```python
import pandas as pd
import numpy as np
import warnings
import math
warnings.filterwarnings('ignore')
import statsmodels.api as sm
import statsmodels.formula.api as smf
import matplotlib.pyplot as plt
from sklearn import datasets, linear_model
from sklearn.model_selection import train_test_split
import seaborn as sns
%matplotlib inline
```

*Figure 6. Python imported libraries*

Importing Data

The accumulated file was imported into Python Jupyter notebook to process then we converted the date into the proper date format.

data = pd.read_csv("CarRegistrationData.csv")
data["date"] = pd.to_datetime(data["date"], format="%Y/%m/%d")
data.head()

<table>
<thead>
<tr>
<th>nationality</th>
<th>gender</th>
<th>person_birth_year</th>
<th>vehicle_type</th>
<th>manufacture_year</th>
<th>cylinders_num</th>
<th>cylinder_capacity</th>
<th>color</th>
<th>business</th>
<th>individual</th>
<th>fuel_type</th>
<th>transmI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Armenia</td>
<td>MALE</td>
<td>1989</td>
<td>Light Vehicle</td>
<td>2002</td>
<td>6.0</td>
<td>3200.0</td>
<td>Silver</td>
<td>Private</td>
<td>Individual</td>
<td>Benzene</td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>FEMALE</td>
<td>1974</td>
<td>Light Vehicle</td>
<td>2016</td>
<td>6.0</td>
<td>2500.0</td>
<td>Black</td>
<td>Private</td>
<td>Individual</td>
<td>Benzene</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 7. Correction of the date format*
Data Exploration

Understanding data – We can observe that our merged data consists of 156668 instances of 18 variables.

![Data frame](image)

Figure 8. Data frame

Also, we will have a quick overview of the data by looking at the data information to determine how many quantitative variables and how many categorical variables we have.

![Data Information](image)

Figure 9. Data Information

We can observe from the following table that we have 7 quantitative variables and 11 categorical variables (including date variable of type datetime).
We can observe from the following table that maximum value of no. of cylinders is 560 which is impossible scenario and indicates erroneous data. Similarly, we can identify anomalies for number of doors also where minimum value = 0 and maximum value = 56. Both are erroneous values and needs to be treated with. We can remove them by identifying outliers and removing them.

In whole, all the variables other than date have minimum value as 0 which is not possible in real life scenarios. Similarly, exorbitantly high max values are also known as dirty data.

Lastly, if we observe mean and median values of data, they are fairly close indicating normal distribution. We can further confirm it by plotting histograms.
Data Pre-Processing
Identifying and Removing Null Values

We can see that only 4 variables have null values more than 1%. We found that cylinder capacity variable has a lot of missing values (around 30% of the values were missing). So, we replaced these missing values with most frequent observation in data set. We did the same thing for other variables of cylinder_num, number_of_passengers_seating_capacity and number_of_doors.

- For numerical variables, we have replaced null values with mean/median of column.
- For categorical variables, we can replace null by most frequent category or create a new category.
Build boxplots and Identify Outliers

We have built boxplots. Now, we can observe that for variables Number of Cylinders, CylCapacity, Seating Capacity, Doors, there are high no of outliers.

We will be using the Inter Quartile Range method to remove outliers. We first calculated the IQR range and then any value lying outside IQR was considered outlier and was removed.

Figure 12. Box plot for variables
Exploratory Data Analysis Using Python and Tableau

Univariate Summary Plots:
Univariate data analysis helps us to understand dispersion and distribution of data. There are various plots which can be used to give us column level summary of our data. It helps us to understand distribution of data whether it is normally distributed, right skewed or left skewed etc.

Categorical Variables- We have built bar charts for all categorical variables to understand the most frequent categories for each variable. We observed that we have 5 unique categories in fuel_type with Benzene being the most prominent one. Further, we have 9 unique categories in vehicle_type with Light vehicle being the most prominent one. As we observe gender of the users who did car registrations, Male individuals are more than 5 times the female individuals.

By looking at the following bar chart of Nationality, we can see that maximum registrations were made by the people of Indian Nationality.

Also, we can observe there were totally different nationalities of people who did car registrations.
Numerical Variables- Histograms are like bar charts which display the counts or relative frequencies of values falling in different class intervals or ranges. We have plotted histogram for numerical variables. By looking at the individual histograms we can say that the data is normalized. But there are few outliers which in each column leading to expansion of right tail.

Figure 13. Univariate data visualizations

By looking at the following bar chart of bar model, we can see that maximum registrations were made for Patrol brand model. Also, we can observe that 2200 unique brand models got registered.

Figure 14. Skewness of data frame
Bi-Variate Analysis Using Tableau

Registrations as per Fuel Type

Battery Electric Motor

The bubble charts generated here is depicting the prime registered vehicle manufacturers in the Battery electric motor segment. The chart clearly reveals that Tesla’s vehicles were dominant in the electric segment with maximum registered electric vehicles carrying its brand name in the 3-year analysis.
The bubble charts generated here is depicting the prime registered vehicle manufacturers according to benzene fuel type. As visible from the chart, Nissan’s vehicles seem to be the most registered vehicles over the year in mentioned fuel type. Though Toyota and Mercedes vehicles were dominant in the 2019, Mercedes vehicles registration declines the following years with Toyota dominating the market only second to Nissan in benzene fuel type section.
The bubble charts generated here is depicting the prime registered vehicle manufacturers according to benzene-electric fuel type segment. Though Nissan was a prominent name in vehicle registrations according to benzene fuel type, the chart prepared filtering benzene-electric fuel type reveals that higher amount of Toyota vehicles were registered consistently in pertaining three years. However, vehicles belonging to brand name Lexus were also registered heavily in previous years being only second to Toyota under this fuel type.
This chart depicts the registration of vehicles as per fuel type allowing the user to draw a comparison between vehicles of different brand name featuring diesel fuel type. The bubble chart formed above reveals that there was significant comparison in between brand names Mercedes, Land Rover and, Mitsubishi in the year 2019, however, during the year 2020 Mercedes vehicles were dominant in diesel vehicles registration followed by Land Rover and Ford.

The chart depicts yearly summary of vehicle registration as per manufacturer brand name. From the chart it can clearly noticed that the most vehicles registered by RTA featured the brand name Nissan, however Toyota was second highest followed by Mercedes, Honda, and Ford. Moreover, Nissan and Toyota have comparable registered vehicles in 2019.
The chart draws up a comparison between Genders taking into account the birth year or age of a person who registered their vehicle with RTA. From the chart it can be clearly seen in either genders people born between the period of 1986 to 1990 were prominent with registration of vehicle stating a certain age group more likely to visit a center. However, the drawn chart also reveals that there is a significant difference in male and female registration holders with the list clearly male dominant. From this it can be drawn that the work force should be more equipped for dealing with male adult customers prominently of a working age class.
The given chart is descriptive of the total registrations classified on the basis of the type of vehicle and further categorized over the year of registration. From the chart it is clearly visible that in Dubai, light vehicle is the most popular vehicle type among the people with highest registrations being conducted every year. Moreover, the chart also reveals that though light vehicle type is unrivalled in the count of registrations, the second highest vehicle registered by the residents of Dubai is motorcycles. This indicates that future registration requirements are also to circle around the same categories.
The given chart describes the weekly trend of registrations concerning a study of year registration data. The trend line on the chart clears points towards Thursday being the busiest day of the week with highest number of customers visiting for registration of their vehicles on the concerned day. However, there is no data of registration service rendered on Friday as it a holiday in Dubai. So, RTA should focus on employee optimization on Thursday to converge improvement in customer satisfaction.

The above charts categorize the registrations made by residents of Dubai according to their nationality over the course of 3 years. The population of Dubai comprises of people of numerous ethnicities belonging to variant nations. According to the chart generated from the data, Indian nations dominate the list of vehicle registration followed by the people of Emirates and then Pakistan.
This line charts depicts the overall registrations of vehicles spread over the year. It can be easily understood that there was an apparent dip in the number of registrations during the months March-April in the visible year. It should be noted that such witnessed due to the apparent lockdown declared during the period. Moreover, a filter for the exclusion of Fridays from the chart was implemented as it is a holiday in the nation.

Manufacturing as per vehicle type and cylinder capacity

The data received from the RTA was utilized for the formation of charts assistant in drawing comparison between various variables. The chart generated from the data draws a comparison of vehicles manufactured on the basis of vehicle type and their cylinder capacity. The charts revealed that light vehicles with undesignated cylinder capacity were manufactured in the highest number followed by light vehicles of cylinder capacity pertaining 2000. It can be clearly comprehended from the chart that light vehicles clearly dominate the vehicles manufactured.
Inferential Analysis Using Python

Ordinary Least Squares Regression (OLS)

We have done OLS to predict Cars registration on a particular day of week in a particular month. As we had seen from our exploratory analysis that Registrations are higher on Thursday and almost minimum on Fridays. This data is from the pandemic time, so it has some unusual trends and patterns for months giving April as lowest Registration month due to lockdown. With this model however applied on data over the normal working years can give significant results.

For this model we have use random split of training and test sets. We have training and test sets in the ratio of 75:25 and predicted results with 30% accuracy.

![Figure 25. OLS regression results](image_url)
Arima Model

We plotted the daily sale of vehicle against time which in our case is day. If we look at the graph, we can spot some trend i.e. We can identify where the sales are high and where the sales are low. Till April 2020 we can see that the data has seen similar highs and lows in sales. But post that the sale have been slightly less fluctuating and slightly increasing.

Figure 26. Daily registrations of vehicles against date
As we mention earlier in the report, we should do a ADF test to check if the data is stationary or not then decide whether we need to do a transformation. In the figure below, we applied the unit root test (ADF) and here is the result of rolling mean and standard deviation that is obtained for the test.

ADF Test Statistic: -2.280564407060426
p-value: 0.17831297164944132
#Lags Used: 15
Number of Observations: 444
weak evidence against null hypothesis, indicating it is non-stationary

After getting the result, we can see that the value of p is 0.178, which is greater than 0.05. So in this case, we know that the given time series data is non-stationary. Now we will have to do a transformation to the data and try to make it stationary.
We will do our transformation by getting the first difference and seasonal first difference of the daily sales data. On the figure below is the plot transformed data and its rolling mean and standard deviation.

![Rolling Mean & Standard Deviation for first difference of daily registration data](image)

Figure 28. Rolling mean & standard deviation for first difference of daily registration data

ADF Test Statistic: -4.0025350506857125  
p-value: 0.0013972950014337889  
#Lags Used: 18  
Number of Observations: 429  
strong evidence against the null hypothesis (Ho), reject the null hypothesis. Data is stationary

We can observe that this time we had a better p value which is 0.0013 and it is less than 0.05. By stating that, we know now that the data we have used is stationary.

Now we can go ahead with the Autocorrelation plot function (ACF) to identify the pattern of the ARIMA model that we want to use.
The Figure below shows the plot of autocorrelation of the sales variable and the autocorrelation plot of the seasonal first difference data.

![Autocorrelation plot of the registrations](image1)

*Figure 29. Autocorrelation plot of the registrations*

The autocorrelation plot of the sales variable shows that this data is not stationary.

![Autocorrelation plot for seasonal first difference data](image2)

*Figure 30. Autocorrelation plot for seasonal first difference data*

Whereas the autocorrelation plot for seasonal first difference data shows that the seasonal first difference data is now stationary.

Since we have found that the first level of difference (seasonal first difference data) gives us a stationary data variable, our desired value for order of differencing (d) in ARIMA model is equal to 1.
Next, we need to find the order of the AR term (p). The necessary number of AR terms can be determined by looking at the Partial Autocorrelation (PACF) map. After removing the contributions from the intermediate lags, partial autocorrelation can be thought of as the connection between the sequence and its lag. As a result, PACF encapsulates the pure connection between a lag and a string. That way, we will know whether that lag is needed in the AR word or not.

![Partial Autocorrelation](image)

*Figure 31. Partial autocorrelation plot for seasonal first difference data*

We can see that the PCAF for first order of lag is very much above the significant line, the order of the AR term (p) is equal to 1 for our ARIMA model.

The ACF also indicates the number of MA terms needed to eliminate any autocorrelation from the stationary sequence. So, we see that the order of the MA term (q) is also equal to 1.

Now that we have got all the values for our ARIMA model i.e. the desired value for order of differencing (d) equal to 1, the order of the AR term (p) equal to 1, the order of the MA term (q) equal to 1. After that, we produce the Auto Regressive Integrated Moving Average Model (ARIMA) for the time series forecasting and with the formation values of (1, 1, 1). ARIMA model makes the time series stationary through the process of differencing.
We get the below shown results from our ARIMA model:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>ARIMA(1, 1, 1)</td>
<td>Log Likelihood</td>
<td>-2862.786</td>
</tr>
<tr>
<td>Method:</td>
<td>css-mle</td>
<td>S.D. of innovations</td>
<td>123.516</td>
</tr>
<tr>
<td>Date:</td>
<td>Tue, 27 Apr 2021</td>
<td>AIC</td>
<td>5733.572</td>
</tr>
<tr>
<td>Time:</td>
<td>04:05:08</td>
<td>BIC</td>
<td>5750.088</td>
</tr>
<tr>
<td>Sample:</td>
<td>1</td>
<td>HQIC</td>
<td>5740.076</td>
</tr>
</tbody>
</table>

| coef | std err | z   | P>|z| | [0.025] | [0.975] |
|------|---------|-----|-----|---------|---------|
| const | 0.1221 | 0.697 | 0.175 | 0.861 | -1.244 | 1.488 |
| ar.L1.D.Sales | -0.0559 | 0.051 | -1.085 | 0.278 | -0.167 | 0.045 |
| ma.L1.D.Sales | -0.8744 | 0.022 | -39.826 | 0.000 | -0.917 | -0.831 |

The coefficients table is in the center, and the values under ‘coef’ are the weights of the various terms. We see that ma.L1.D.Sales is has a higher negative coefficient with significance value of 0.000 which makes it very significant as p < 0.05.

The Residual sum of squares (RSS) value is minimum for the ARIMA model and it shows that it is the best model for time series forecasting. Hence, the data is trained and tested and for each observation of the data there is standard error that is obtained. After obtaining the predicted values the mean error is calculated.
We have divided the entire data into train and test sets to predict the efficiency of our model. All the data before 2021 has been used to train the model and 2021 data is used to test the model. As we can see from the graph, the model performs good enough in predicting future registrations. We have incorporated SARIMA to incorporate the seasonality component.
Figure 34 is a plot that shows the actual and the predicted time series of the data:

*Figure 34. Actual vs predicted times series*
Results

Augusted Dicky Fuller test

<table>
<thead>
<tr>
<th>Transformation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before transformation</td>
<td>0.17831297164944132</td>
</tr>
<tr>
<td>After rolling mean &amp; std deviation</td>
<td>0.0013972950014337889</td>
</tr>
</tbody>
</table>

*Table 2. Augusted Dicky Fuller test p-value before and after transformation*

<table>
<thead>
<tr>
<th>Predicted values vs actual values</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Error</td>
<td>62.140431</td>
</tr>
<tr>
<td>Mean Squared Error</td>
<td>74.31568045758182</td>
</tr>
<tr>
<td>Mean Absolute Percentage Error</td>
<td>14.640310</td>
</tr>
</tbody>
</table>

*Table 3. Predicted values vs actual values results*

From our collected data, we were able to derive the number of vehicle sales daily for last 463 days. We created an ARMIA model to predict the number of sales daily. With the ARIMA model we were able to predict the number of sales for any given day. This was one of our goals to predict the number of sales using machine learning algorithms.

It is observed that during the month of March, the highest number of vehicle registrations recorded for a single day is around 520. This will help us determine how many employees we need for this specific day. Furthermore, if we are assuming that each employee registers 20 vehicles; we will need 26 employees ready for this day. Accordingly, we have two options for the management decision making:

1) Increase the number of employees in the center to 26.
2) Increase the number of KPI based on how many employees are available.

These number of sales can be assumed to be proportional to the number of customers visiting the store/business for any given day. So, with approximate number of customers and approximate number of sales know to the management they can plan efficiently their staffing and infrastructure for better customer satisfaction. These results can be implemented to various other customer happiness centers.
Based on our learning from this project and other surveys, we have suggested few steps any business can take to optimize their employee’s capacity and increase customer satisfaction. Data analytics is the core of these suggestions.

Ways in which we can optimize employee’s capacity efficiently are discussed below:

**Enable tracking of current utilization and productivity of employee:**
We must implement some form of tool to capture time for employee login activities during whole day. The first step toward managing employee timing is to start measuring the time put in by the employees for different works. There are number of tools in the market like resource scheduling software, project planning software which helps the organization in scheduling task for employees in system and can track these and help the management in reviewing the correct KPI’s of employees.

**Analyze the collected data from tools and take appropriate actions:**
In our analysis it is important that we capture the correct variable data during our tracking of employee KPI’s. We must use the KPI’s against which we can measure the performance and effectiveness of the employees. Like for a car sales organization we can analyze the number of sales leads created, number of sales executed by employee etc. can be used in analyzing data and then make decisions based on those.

**Creating a productive environment for employee:**
It is all about fostering a collaborative workplace where workers can exchange ideas, learn from one another, and mitigate the effect of any absent resources. This is where project and team collaboration are critical, as it is widely acknowledged that working together yields greater results. Creating a productive environment can surely bring out the best in the employee.

**Better planning and scheduling of employee:**
The organization now have a clear idea of what work their workers have done and how it translates into Productivity and Utilization thanks to their timesheet framework. What this information does not tell us, though, is how this relates to the jobs that workers were scheduled to do. So, we need a better scheduling tool which can help in effective planning and scheduling of employee.
Challenges Faced During the Project
The major challenges we faced during the project was with the dimensions of our data frame for time series forecasting. Since, we have tried to analyze the customer waiting time and employee efficiency during covid times, we do not have sufficiently large data sets for time series modeling. We did not find much literature on how to reduce queue wait time for customers at store. We only discovered few resources which focused on similar solutions like parallel processes, identifying bottleneck and employee more workforce there. Moreover, these researches were limited to retail industry. Furthermore, we struggled to find data sets which captures the employee time spends during the whole process. This could give us insights into their efficiency and hence can assist in better planning and staffing. So, we are suggesting it as the future work where we can collect the employee time spending on various activities and then analyze this data to facilitate better employee planning and staffing.
Conclusion and Future Work

This study was conducted keeping a focus on finding ways to improve employee utilization, decrease customer wait time and increase customer satisfaction. Since we see that in any business, waiting queue for customers can result in bad customer experience. In most of the cases these kinds of bad customer experience result in loss of customer for business as well. So, in our project we have emphasized how businesses can leverage data analytics capabilities to assist them in better decision making for all business processes. We collected data from Dubai Pulse where data is uploaded stored from Roads and Transport Authority (RTA). After data cleaning and preprocessing we created an ARIMA model to predict the number of daily sales. This prediction about the number of daily sales can give us great insights which can be used in better decision making for resource allocation by the management team. This better resource allocation by management will ultimately result in better customer service.

As future work, we can even collect the employee related data. We can track and capture time spent by their employees on various activities. Subsequently, we can analyze this data collected for all the employees and make strategies to increase efficiency of the employees. We can also collect the list of activities performed by them and then group similar activities together and remove the non-value-added activities. This exercise can aid us in better allocation and planning of their employees and offer better services to their customers.
Bibliography


