<https://www.kaggle.com/code/khansahil128/online-shopping-dataset> from her

Kate O Dwyer (sbs23025)

CA2 (***Integrated***)

**Data Visualisation and Machine Learning for Business**

**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Visualisation and Machine Learning for Business |
| **Assessment Title:** | CA2 (Integrated) |
| **Lecturer Name:** | James Garza and Sam Weiss |
| **Student Full Name:** | Kate O'Dwyer |
| **Student Number:** | sbs23025 |
| **Assessment Due Date:** | 26/11/2023 |
| **Date of Submission:** | 26/11/2023 |

**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**

Contents

[1. Introduction 4](#_Toc151915997)

[Preprocessing and Cleaning 4](#_Toc151915998)

[Wireframe 4](#_Toc151915999)

[Plots Created 6](#_Toc151916000)

[Map Plot 6](#_Toc151916001)

[Stacked Bar Chart 6](#_Toc151916002)

[Line Plot 7](#_Toc151916003)

[Pie Chart 8](#_Toc151916004)

[Other Panel Sections 9](#_Toc151916005)

[Heading 10](#_Toc151916006)

[Tables 11](#_Toc151916007)

[Text 12](#_Toc151916008)

# 

# Data Visualisation

## Preprocessing and Cleaning

The data was preprocessed and cleaned in the Machine Learning for Business section of this assignment. The dataset features (Table 1) contained missing and repeated values, as well as different data types.

**Table 1.** Data set prior to data preparation and cleaning.

|  |  |
| --- | --- |
| **Features (Columns)** | 18 |
| **Observations (Rows)** | 9880 |
| **NaN Values** | Yes |
| **Data types** | Object(15), int64(1), and float64(2) |

The data was cleaned and prepared to perform exploratory data analysis. This included:

* Removing ‘Row ID’ which contained the index information, and was not necessary
* Removing null values (NaNs)
* Removing one duplicate row.
* Order Date was transformed to datetime64 to have in a date series format
* The data set was sorted by order date
* Average sales was added
* Year was added

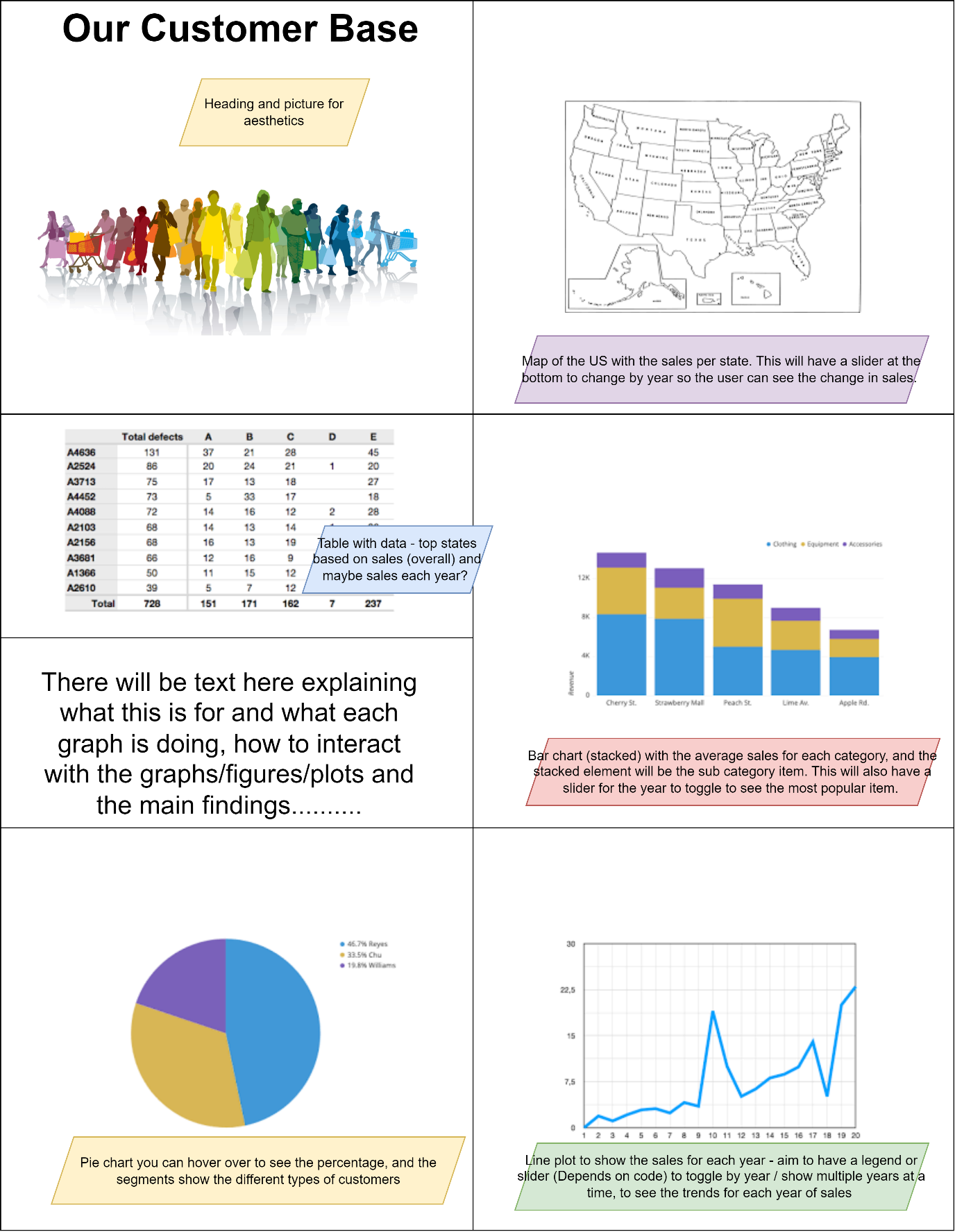
The new data set (Table 2) will be used for this report as the data preparation was deemed sufficient.

**Table 2.** Data set prior to data preparation, cleaning, and column addition.

|  |  |
| --- | --- |
| Features (Columns) | 18 |
| Observations (Rows) | 9,789 |
| NaN Values | No |
| Data types |  |

## Wireframe

Before creating a dashboard for this data set, a wireframe was created using draw.io to brainstorm ideas for how the dashboard should look (Figure 1)

******

***Figure 1.*** *WireFrame for dashboard precreation.*

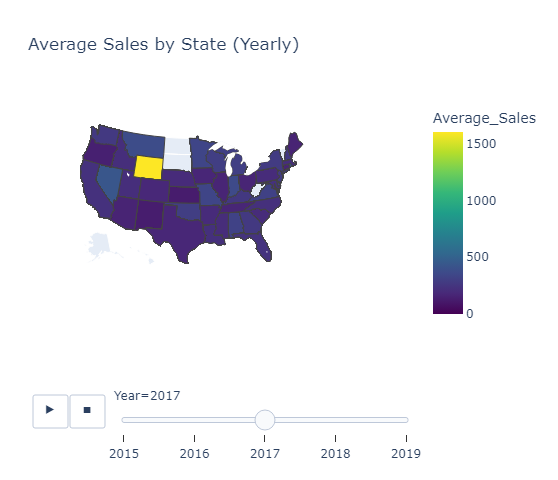
The wireframe shows the layout of the dashboard before creation, and has explanations on each segment, and why these would be beneficial for the user.

## Plots Created

All plots were created using px.plot, which creates interactive plots on Jupyter notebook. This data frame consisted of many categorical variables, and I wanted to show as much data as possible with the plots

### Map Plot

A map plot (Figure 2) was created to show the average sales per state and slider to change the year. This plot is interactive, and when the user hovers over a state it shows the state (as two letters, for example New York is NY) and the average sales. The slider can be used to change the year. This plot is beneficial at also showing what states did not have any sales for a given year, which are the greyed-out states.



***Figure 2.*** *Map Plot of the United States use in the top right hand segment of the Panel dashboard.*

This plot was created to understand the customers from a geographic perspective, i.e., which states spend the most and which states spend the least. From using this plot, it was clear that most of the states have similar sales, mostly less than 500. In 2017 however, Wyoming was the highest state with 1200, this was insightful and can be something the company can investigate to determine why Wyoming had such a big spend that year. A lot of states did not have purchases for 2019 this was because the data set does not go past January 2019.

The colour used is the default, however I think that it is very helpful in showing the differences as the yellow and blue are so different but there is a gradient to show slight differences.

### Stacked Bar Chart

A stacked bar chart (Figure 3) was created to visualise the top categories and the subcategories within these, based on average sales, with the year on a slider similar to Figure 2.

A screenshot of a graph

Description automatically generated

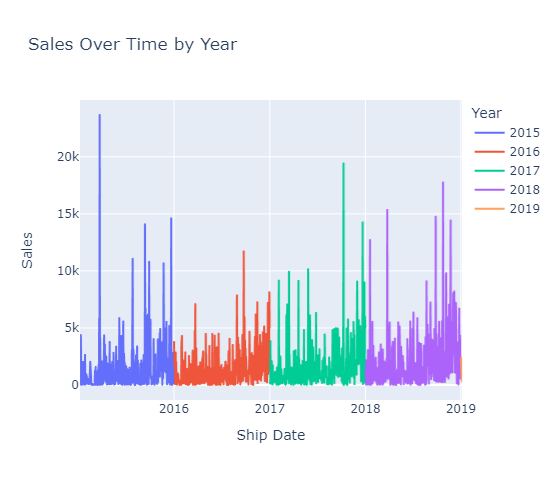
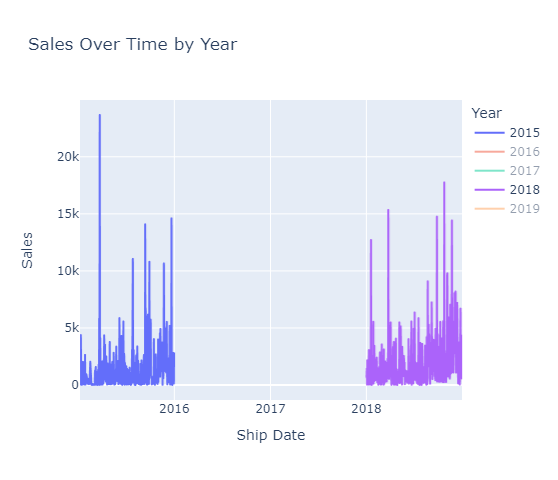
***Figure 3.*** *Stacked Bar Chart for Subcategories and Categories.*

There were three categories: Furniture, Office Supplies, and Technologies. There were many subcategories within these, and when the user hovers over the bar chart, the top five (there may be less than five subcategories within a category) subcategories are shown. This graph is very helpful in determining which category had the highest sales each year, and within that category, it is easy to determine which subcategory sold the most. Although the colours do overlap, (tables and copiers for example) when the use hovers over the bar chart, it is highlighted what the subcategory is.

The top category each year was technology, and depending on the year the top subcategory within technology was copiers and machines.

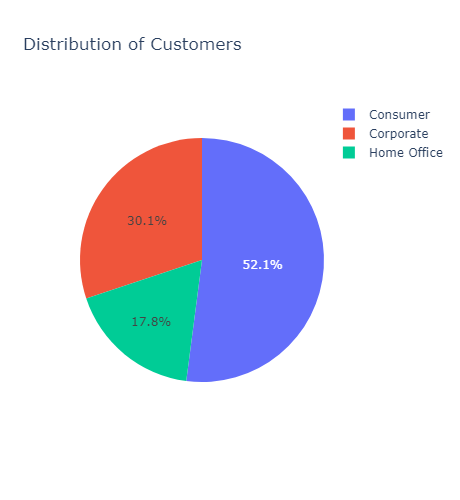
### Line Plot

A line plot (Figure 4) was created to visualise the sales per day for each year. A slider was not used here and instead a legend used where the user can pick which years to show.

***Figure 4.*** *Line plot for the sales each year. Left hand side (LHS) showcases all data. Right hand side (RHS) is an example of the interactive element of the plot, where the user can click the legend to remove or add years.*

The sales per year increased every year (this is in the table on the LHS of the dashboard) which isn’t necessarily evident from the graphs plotted in Figure 4. However it is interesting that each year there is the same rough trends in the sales visually. There was only a small amount of data for 2019 hence 2019 cannot be seen well in the images provided. Different colours were used by default and they are different enough that they did not need to be changed to enhance usability.

### Pie Chart

A pie chart (Figure 5) was created to visualise the percentage of customer types: consumer, corporate, and home office.

***Figure 5.*** *Pie Chart for Customer Types.*

This pie chart is interactive; the user can hover over the segments to see which customer type the segment represents and the associated percentage. The ‘consumer’ section is the highest percentage (52.1 %) and home office is the lowest (17.8 %). This suggests that the company could increase sales by targeted home office items in discount and could also reach out to companies to increase the corporate section.

## Other Panel Sections

Other panel sections (Figure 6) were created.

A group of people walking in a line

Description automatically generated

***Figure 6.*** *Heading, tables, and text section of the panel.*

### Heading

The heading (Figure 7) is the same as that in the wire frame (Figure 1) and the aim of this was to make the panel look attractive to the user, and to make it as colourful as possible.

A group of people walking with shopping cart

Description automatically generated

***Figure 7.*** *Heading and picture used in the panel dashboard.*

### Tables

Two tables outlining the sales per year and top sales per state were added using the tab command for the panel, where the user can click on the tab and show the desired table (Figure 8).

A screenshot of a computer

Description automatically generated

***Figure 8.*** *Tables on the Panel, activated by clicking on the tabs.*

This section is to give a brief overview of important information that the plots don’t capture in a clear and specific manner.

### Text

A group of people walking in a line

Description automatically generatedThe text section explains each section of the panel dashboard, and the key findings from the visualisations performed (Figure 9). These have been explained in the ‘Plots’ section above.

***Figure 9.*** *Text section of the Panel Dashboard.*

# Machine Learning for Business

**Assessment details**

Question 1:

Discuss the concept and application of Time series analysis using machine learning modelling by providing a real-world data set. What is the purpose of The Augmented Dickey-Fuller test in time series?

a) Apply an appropriate Box-Jenkins model to the chosen dataset (ARMA, ARIMA, SARIMA etc). Check for the model adequacy.

b) Make one-step-ahead forecasts of the last 10 observations. Determine the forecast errors.

c) Make a time series plot of the data, and further calculate and illustrate the sample autocorrelation and partial autocorrelation (PA). Is there significant autocorrelation in the chosen time series?

(40 Marks)

Question 2:

Discuss the concept and application of Text Analytics (eg text categorisation, topic modelling and document summarisation) using machine learning modelling by providing a real-world data set of social media posts.

Apply appropriate text analytics tools to the chosen dataset.

(40 marks)

Observations about the assessment:

In your report, please include:

* + **Introduction:** Briefly explanation of the topics/ questions and the steps that were followed to write the report.
  + **Argument:** Justification of process you followed and the machine learning models performed to gain insights about the topics you are working on. Variety, research, and rationality explanation is expected.
  + **Conclusions:** Here you should lead to some final comments regarding the topics you worked with and the relation between the raw data and machine learning.

(20 marks for report)